Introduction to CMOS VLSI Design
Delay Part A

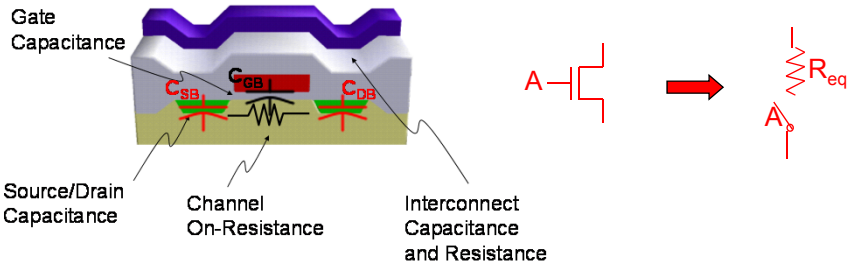
Lecture by Jay Brockman
University of Notre Dame Fall 2008
Modified by Peter Kogge Fall 2010,2011, 2015, 2018
Based on lecture slides by David Harris, Harvey Mudd College
http://www.cmosvlsi.com/coursematerials.html

Outline

- **Delay Part A**
  - **Capacitance**
  - **Effective Resistance**
  - **RC Delay**
- Delay Part B
  - Review
  - Inverter Delay
  - The Elmore Model
  - Effect of Load Capacitance

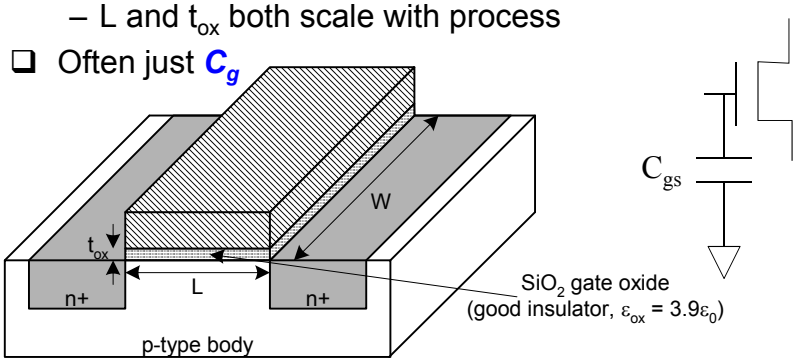
Capacitance

- Conductors separated by 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"
- $C_{gs} = \varepsilon_{ox}WL/t_{ox} = C_{ox}WL = C_{permicron}W$
  - *proportional to "width" ONLY*
- $C_{permicron} = \varepsilon_{ox}(L/t_{ox})$ typically ~2 fF/μm of gate width
  - L and t_{ox} both scale with process
- Often just C_g

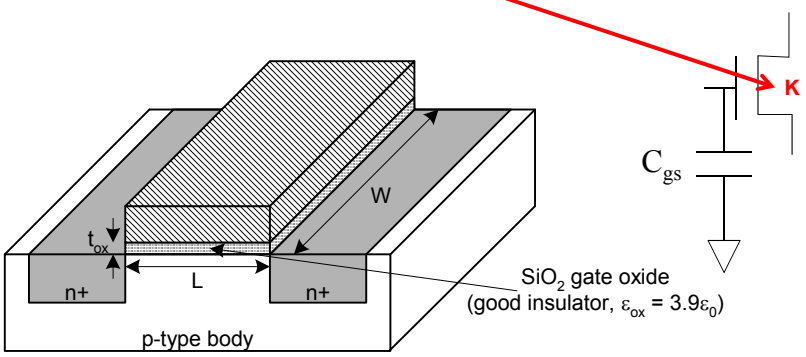


Caveat

- Statement that C_{gs} is proportional ONLY to width is true ONLY as long as Length and t_{ox} scale together
- We will assume this for rest of course
- BUT: ***NOT TRUE*** for really small feature sizes!

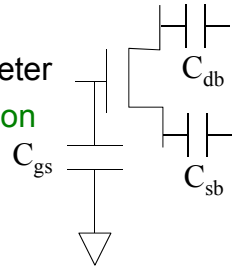
Denoting "Width"

- Assume width of smallest transistor is W_s
- If actual width of a transistor is "W"
- Then relative width "k" is W/W_s
- ***Write "k" inside transistor***



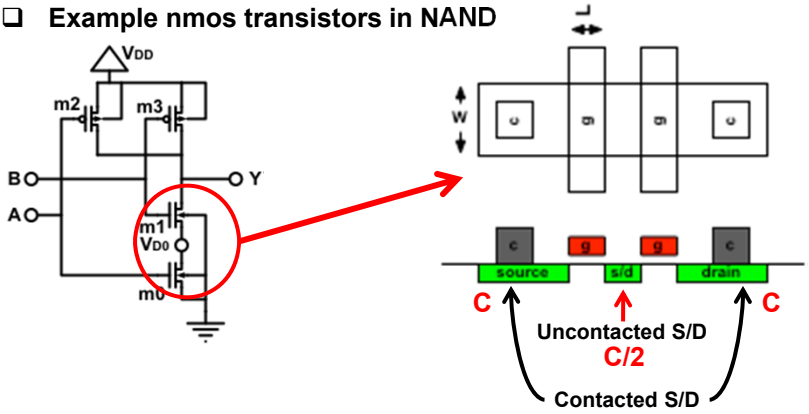
Diffusion Capacitance

- C_{sb}, C_{db} = "source/drain to bulk"
- Undesirable, called *parasitic* capacitance
- Capacitance depends on area and perimeter
  - Comparable to C_g for contacted diffusion
  - ½ C_g for uncontacted
  - Varies with process
  - Often just C_d
- "Contacted diffusion" occurs when there is a metal contact "touching" the diffusion
  - i.e. there's a "wire" on the source or drain
- Appears on both source & drain
  - But ignore when connected to V_{dd} or Gnd
    - Capacitance is "shorted out"

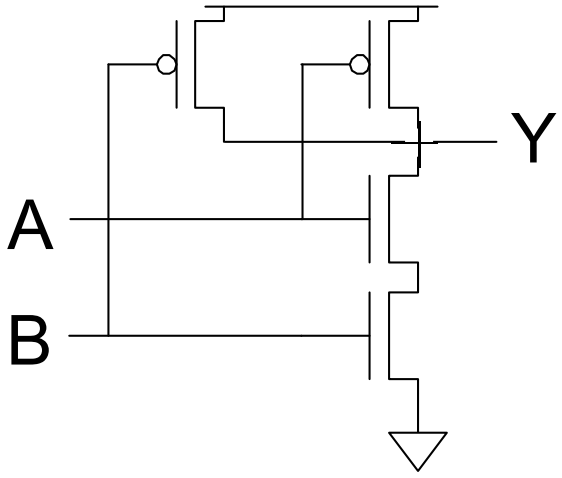


Contacted vs. Uncontacted Capacitance

- **Transistor Layout**
  - **Contacted Source/Drain – Physically Larger => more cap**
  - **Uncontacted Source/Drain – Smaller**
- **Example nmos transistors in NAND**

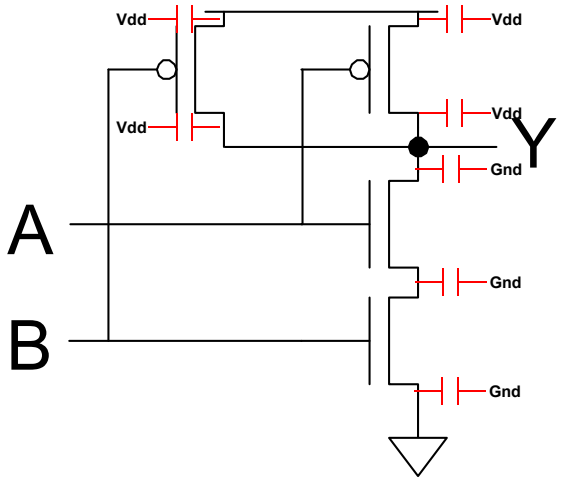


Let's Do An Example: How Many Diffusion Caps Are There? And Where?

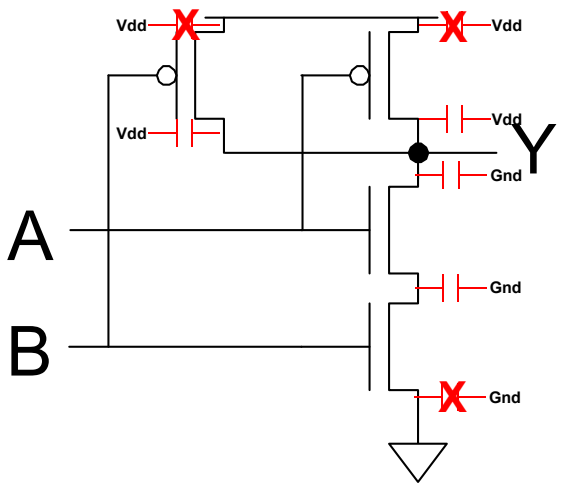


Answer: ____

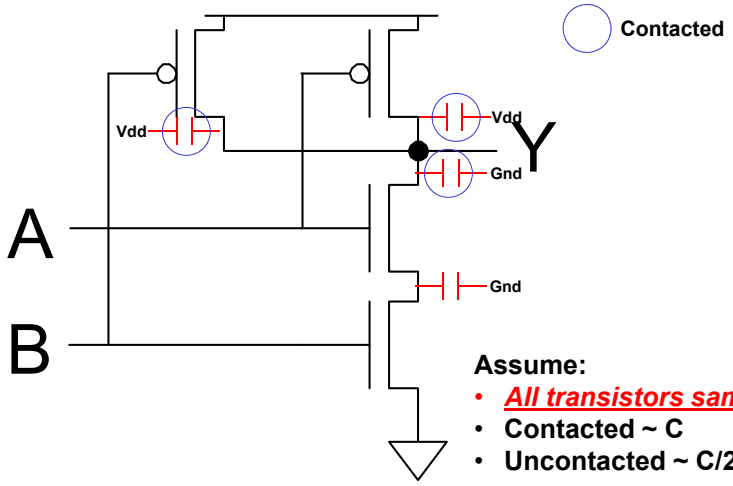
OK- How Many "Don't Count"?



OK- Which are Contacted?



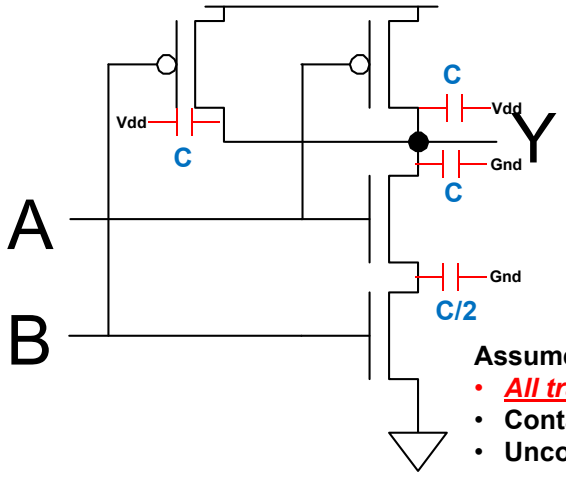
OK- What are Their Values?



Assume:

- ***All transistors same size***
- Contacted ~ C
- Uncontacted ~ C/2

OK – What's the Final Model

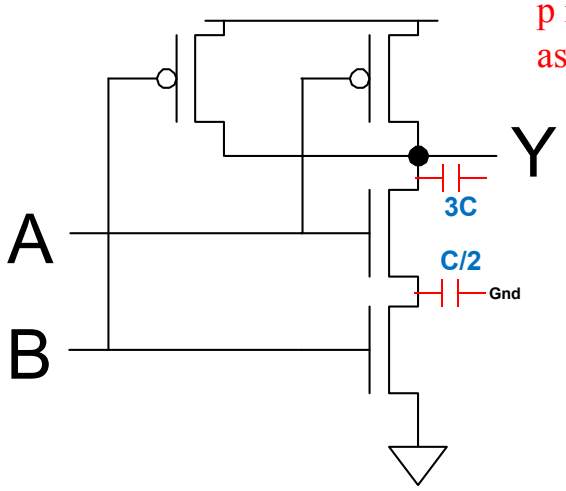


Assume:

- ***All transistors same size***
- Contacted ~ C
- Uncontacted ~ C/2

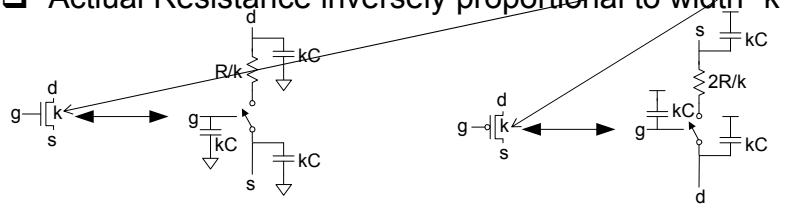
Final Model

Again this assumes p nypes same size as ntypes

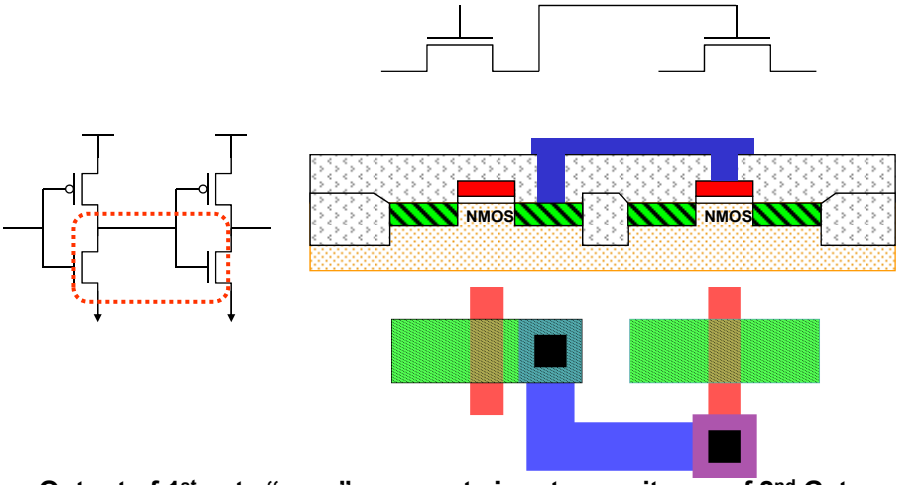


RC Delay Model

- 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
    - Why?
- Actual Capacitance proportional to minimum width "k"
- Actiual Resistance inversely proportional to width "k"

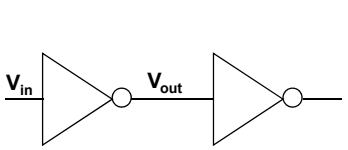


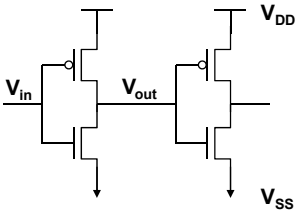
Driving Another Gate



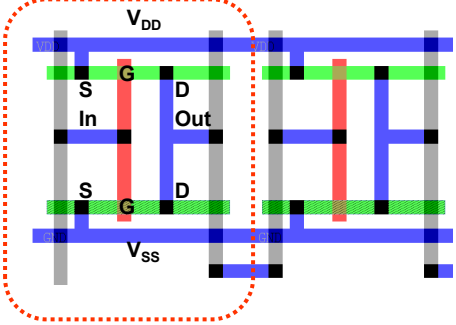
Output of 1st gate "sees" aggregate input capacitance of 2nd Gate
Termed *Load Capacitance C_L*

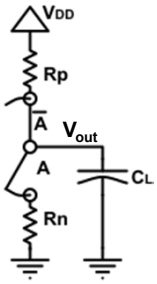
Between Cells





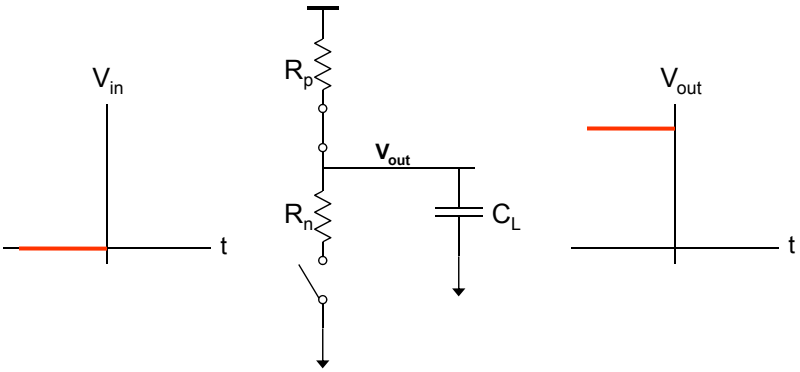
Inverter "Stick Diagram"





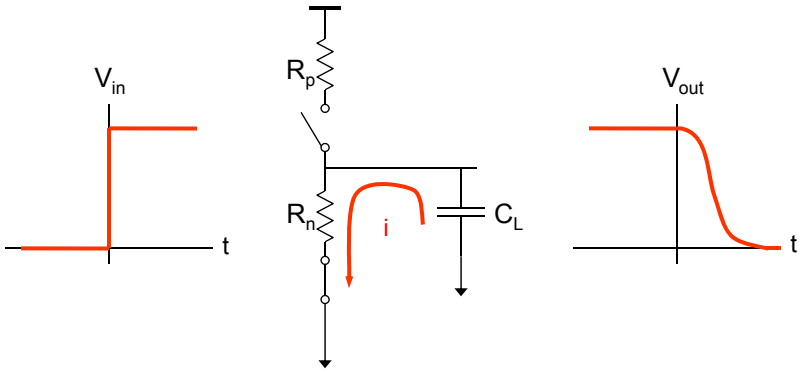
Green: diffusion; Red: poly; Blue/gray: metal wire; black: contact

1st Inv Falling Output (1)



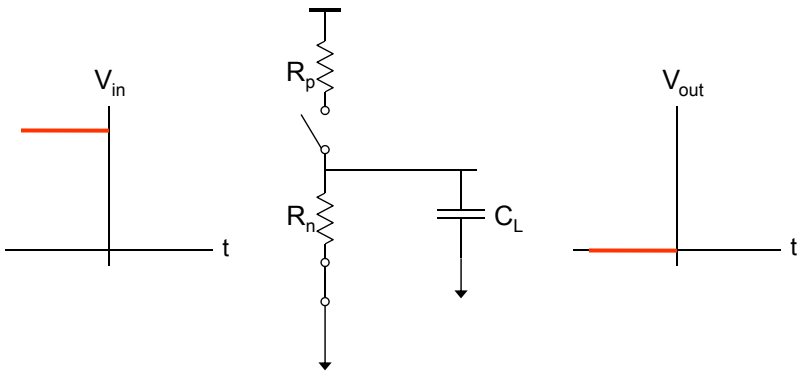
Before t=0, input stable & capacitor fully charged

Falling Output (2)



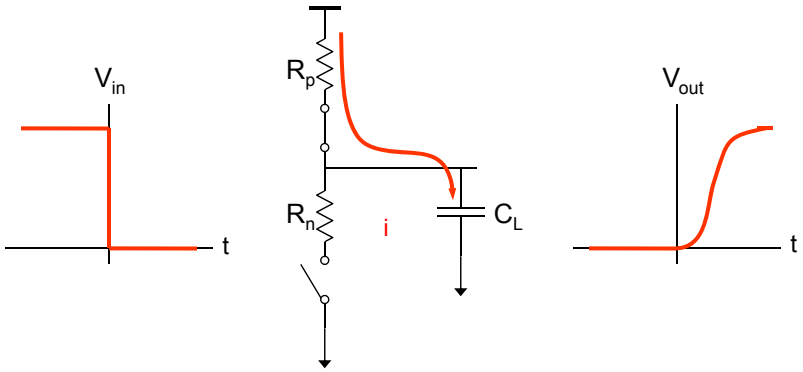
After t-0, capacitor discharges through NMOS

Rising Output (1)



Before t=0, input stable & capacitor fully discharged

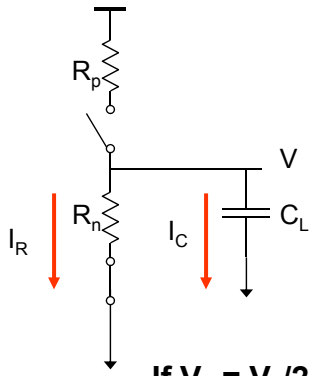
Rising Output (2)



After t=0, capacitor charges through PMOS

Solving

- How long does it take to discharge the output from starting voltage V_0 to voltage V_1?



Kirchhoff's current law at output

$$I_R = -I_C$$
$$\frac{V}{R} = -C\frac{dV}{dt}$$
$$\int_{t_0}^{t_1} dt = -RC\int_{V_0}^{V_1} \frac{1}{V} dV$$
$$t_1 - t_0 = -RC(\ln V_1 - \ln V_0)$$
$$= RC \ln \frac{V_0}{V_1}$$

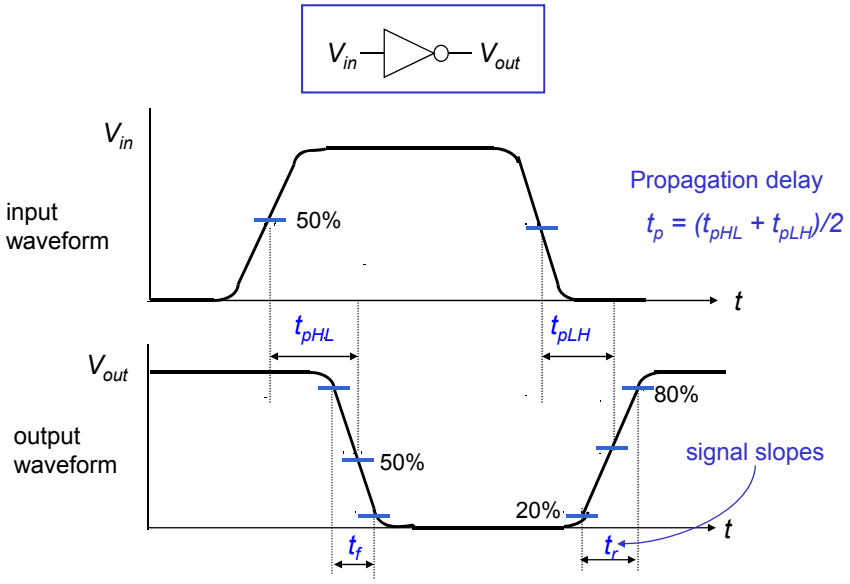
If $V_1 = V_0/2$, Time = RC ln(2) = 0.69*RC

REMEMBER THIS #!

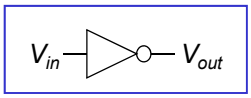
Definitions

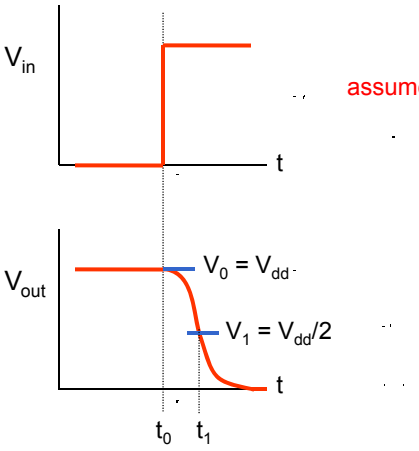
- Waveform
  - ***Rise time*** t_r = time to rise from 20% of Vdd to crossing 80% of Vdd
  - ***Fall time*** t_f = time fall from 80% of Vdd to crossing 20% of Vdd
  - ***Edge Rate*** t_{rf} = $(t_r + t_f)/2$
- Logic gate input to output
  - ***Propagation delay*** t_{pd} = max time from input crossing 50% of Vdd to output crossing 50% of Vdd
    - t_{pdr} = delay when input is rising
      - t_{pHL} = delay when output goes from High to Low
    - t_{pdf} = delay when input is falling
      - t_{pLH} = delay when output goes from Low to High
    - ***Delay*** t_p = $(t_{pHL} + t_{pLH})/2$
  - ***Contamination delay*** t_{pd} = min time from input crossing 50% of Vdd to output crossing 50% of Vdd
    - i.e **with no load on output, in either direction**

Delay Definitions



RC Propagation Delay Estimation





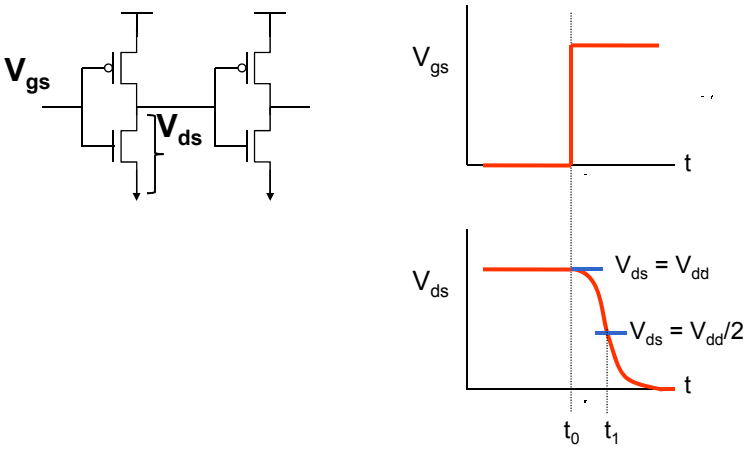
$$\Delta t = RC \ln \frac{V_0}{V_1}$$

$$t_{\mathrm{PHL}} = RC \ln \frac{V_{\mathrm{dd}}}{V_{\mathrm{dd}}/2} = RC \ln(2) = .69RC$$

Effective Resistance

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

Switching Voltages



Approximating Output Resistance (4.3.7)

- Want Effective Resistance that has ~ same current as transistor when $V_{DS} = V_{dd}/2$
- No single R value matches transistor through switching event
- Approach:
  - Look at IV trajectory as input Vgs rises high enough to start transition
    - While still being in saturation
  - Compute average V/I on this trajectory

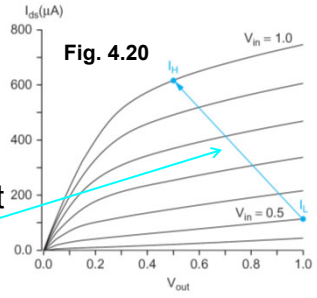


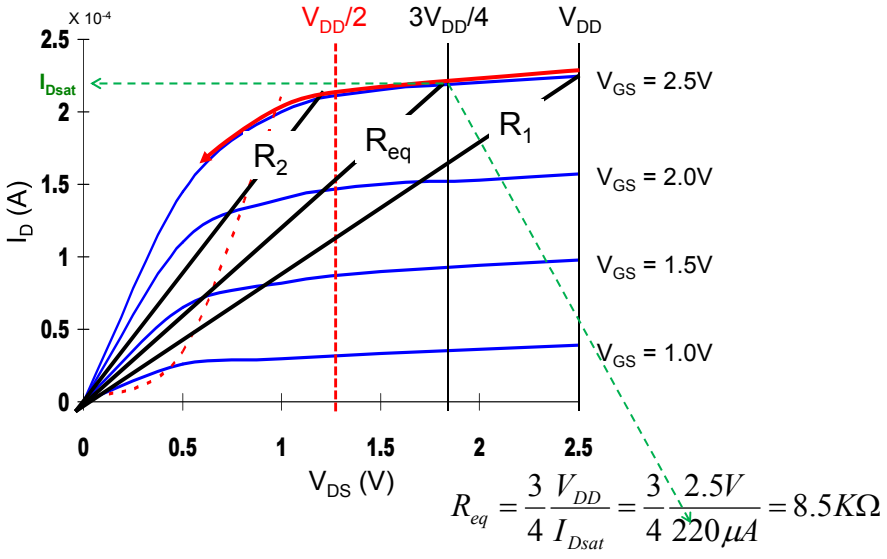
Fig. 4.20

$R = \ln(2) * (3/4) * V_{dd}/Idsat$
$\sim (1/2)*(V_{dd}/Idsat)$

Relating Back to Delay

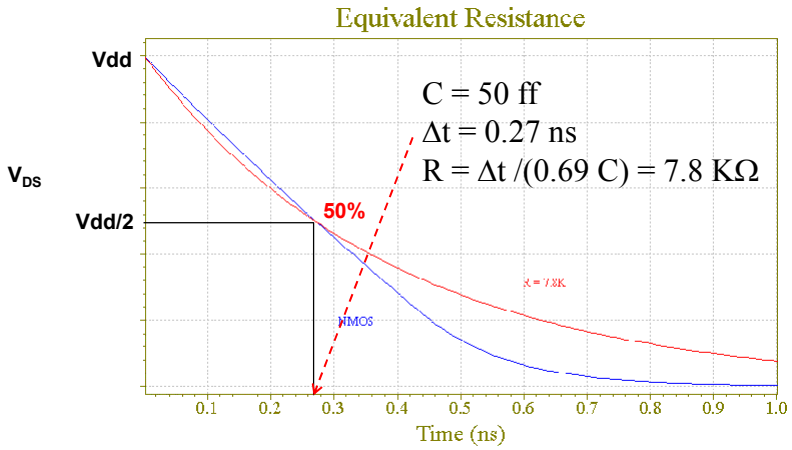
- From before: R = ln(2) * (3/4) * V_{dd}/I_{dsat}
- If we compute delay as ln(2) RC
  - then R ~ (3/4) * V_{dd}/I_{dsat}
- If we compute delay as "R'C" then use above
  - I.e. choose effective R to include ln(2)
  - OR: t_{pd} = RC

Let's Look at a Real Curve



$$R_{eq} = \frac{3}{4}\frac{V_{DD}}{I_{Dsat}} = \frac{3}{4}\frac{2.5V}{220\mu A} = 8.5K\Omega$$

Approximating R_{ON}



RC Values

- Capacitance
  - $C \approx C_g \approx C_{sb} \approx C_{db} \approx 2$ fF/μm of gate width
  - Values similar across many processes for minimal gate length.
- Resistance
  - $I_{Dsat} \approx 550$ μA/μm
  - $V_{DD} \propto \lambda$
  - $R_{eq} \approx 0.75\ V_{DD}/I_{dsat} \propto \lambda$
  - $R_{eq} \approx 7$ KΩ*μm in 0.6 μm process
  - $R_{eq} \approx 2$ KΩ*μm in 90 nm process
- 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

Question: Why?

	V_{DD}
0.6 μm:	5 V
0.35 μm:	3.3 V
0.25 μm:	2.5 V
0.18 μm:	1.8 V
130 nm:	1.5 V
90 nm:	1.2 V

Drawing Equivalent Circuits for Delay

- **Model rise and fall as separate circuits**
- Look at only the output of logic gate being modeled
  - But consider capacitance from the gate it is driving
- Model transistors that turn "OFF" by an open
- Ignore all capacitors with both ends to a rail
- Model transistors that turn "ON" by their resistance
  - If source is V_{dd} or V_{gnd}, model voltage on that terminal as a "STEP Voltage"
- Ignore all capacitors with one end "floating"
- Lump ALL capacitors tied to same wire together
  - Ignore which rail other side goes to
- Draw as equivalent RC "Ladder" driven by step voltage