Lecture 7
Circuit Delay, Area and Power

lecture notes from S. Mitra Intro VLSI System course (EE271)

Circuits and Delay

There are two properties of gates that really matter to a design: delay and power. Both of these are set by the resistances and capacitances associated with a circuit. This lecture starts by talking about R, C, and then uses them to estimate delay and power. It then looks a little about the R, C issues with wires.

After looking at the basic gates, we turn to the question of how to choose the best gate implementation for your circuit. We will start with some logic optimization rules, then realize that something is wrong, and then try again with logic minimization.

Power, Delay and Area: Gate's Metrics

- When we use a gate we care about its logic function
  - That is the desired output

- It also consumes "resources" that we care about
  - Delay
    - The output becomes valid some time after inputs settle
  - Power
    - The gate also consumes some energy from the power supply
  - Area
    - This is often less important in today's chips (it's not really free)
    - The wires needed to connect the gates takes the most space

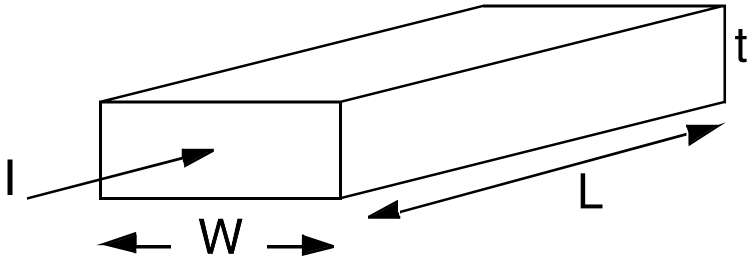
- Both of these "charges" are easy to estimate
  - But that means you will need to estimate them

R and C is All You Need

- Delay can be estimated by simple RC models
- Power can be estimated (mostly) from C alone
- But to do either, we need to have a R, C model of gates
- Need to work with both transistors and wires

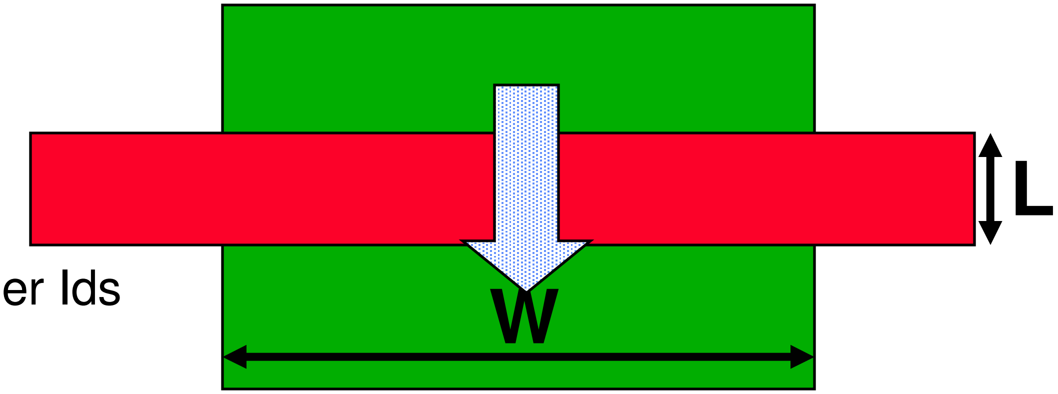
Resistance

- Resistance
  - Resistivity ρ * Length/Area
  - Designer does not control ρ, t
  - Generally deal with ρ /t
  - Called ohm/square

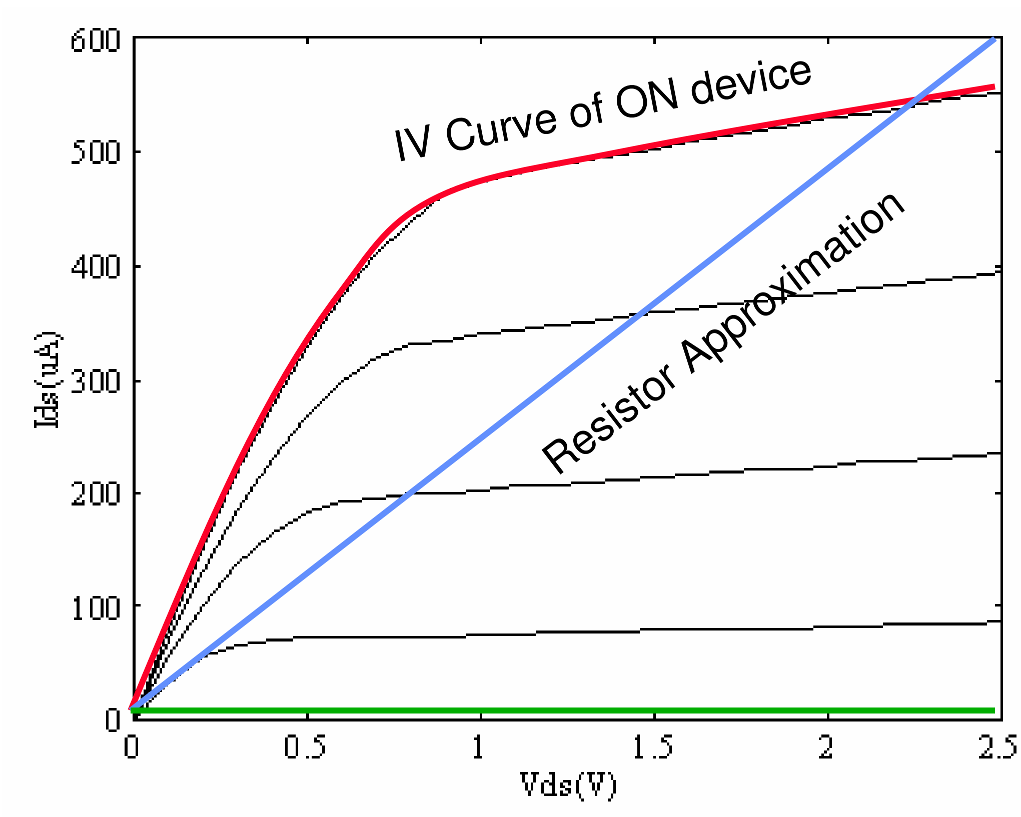


$$R = \frac{\rho L}{tW} = \frac{\rho}{t}\frac{L}{W}$$

- For transistors
  - Vgs controls R/sq
  - Designer chooses
    - W and L
  - Wider transistor
    - More current (Remember Ids expression?)
    - Lower R



Our Switched Resistor Model



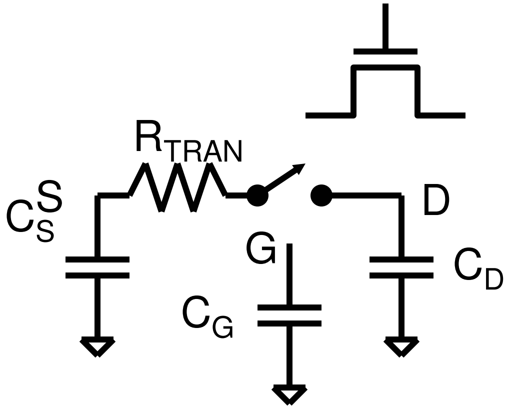
With digital input on gate, device is either ON or OFF

Approximate ON device with resistor (blue line)

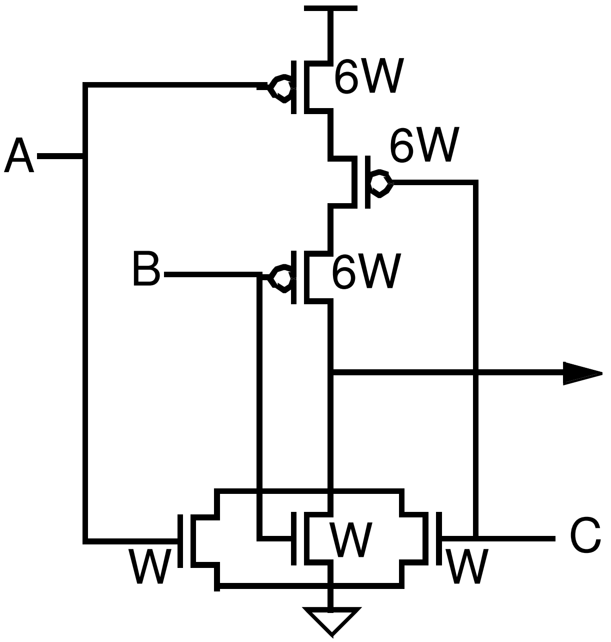
R = L/**W** x **Constant dependent on technology**

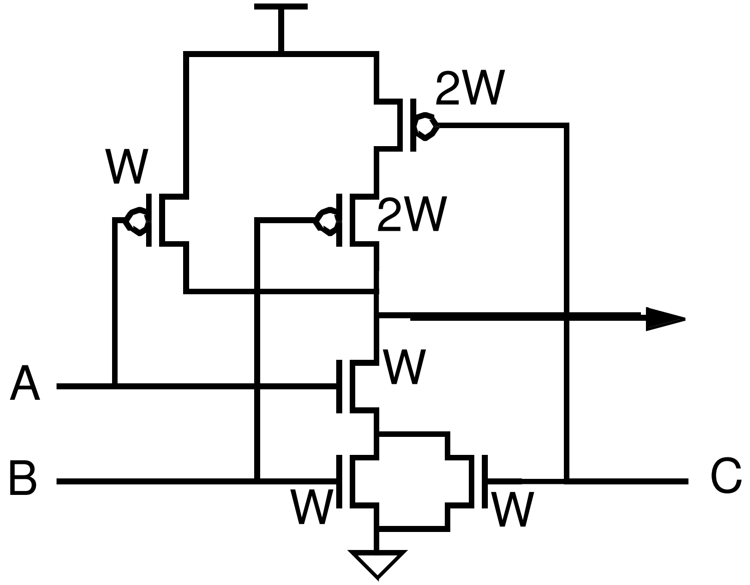
Since L is generally min, just give W

You can think in terms of current as well (proportional to W)



Resistance: NMOS vs. PMOS





- PMOS and NMOS Ids values differ: hole vs. electron mobility
  - We will go by: electron mobility = 2x hole mobility
- What about the pullup vs. pulldown resistances of the above gates ?
- PMOS with width W: R ∝ 2/W, NMOS with width W: R ∝ 1/W

Capacitance and Delay

- Capacitors store charge

  Q = CV <- charge is proportional to the voltage on a node

- This equation can be put in a more useful form

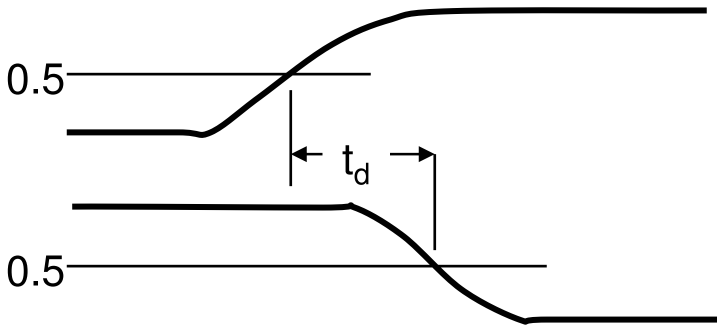
$$i = \frac{dQ}{dt} \Rightarrow i = C\frac{dV}{dt} \Rightarrow \frac{C\Delta V}{i} = \Delta t$$

- So to change the value of node (from 0 to 1 for example), the transistor or gate that is driving that node must charge (up, in our example) the capacitance associated with that node. The larger the capacitance, the larger the required charge, and the longer it will take to switch the node.
- Define R_{trans} so that the current (i) is approximately V/R_{trans}

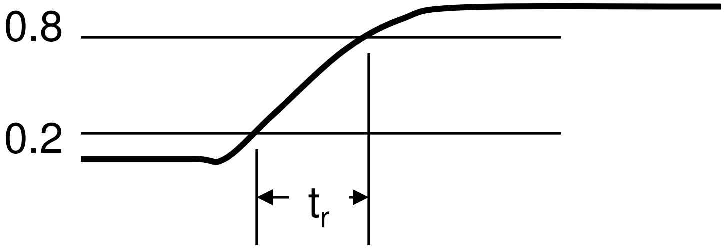
$$\Delta t = \frac{C\Delta V}{i} = \frac{C\Delta V}{V/R} = R_{trans}C$$

This is pretty high level, we are not worrying about small constant prefactors

Delay and Rise/Fall Time



Delay, t_d, is measured from 50% point to 50% point. Gives a measure that is composable. Define Rsq so RC product is roughly equal to the delay t_d



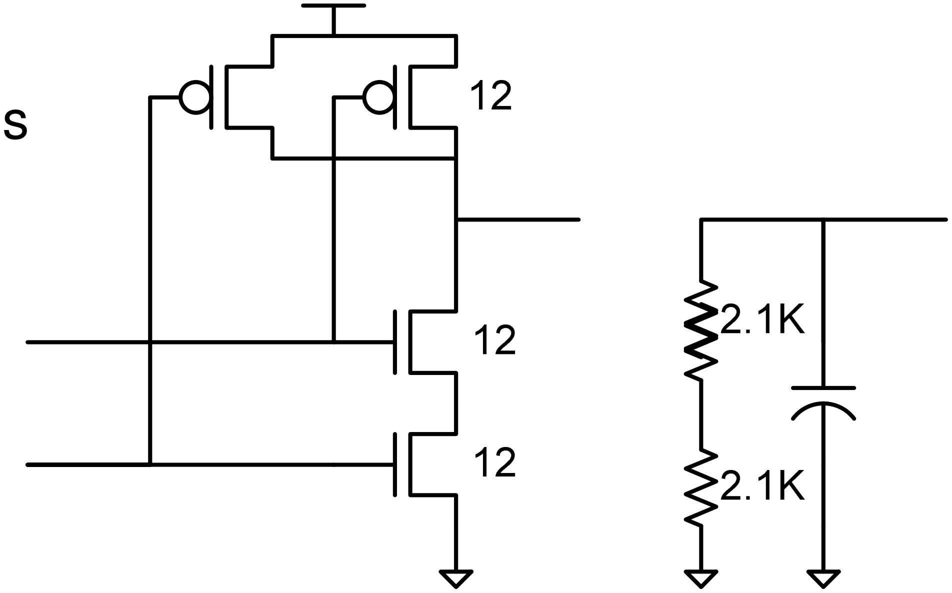
Rise time and fall time, t_r and t_f, are measured from 20% point to 80% point. But they are not generally used in digital design.

Why Fanin is Bad

- Pullup and pulldown are duals of each other
  - Implies there will be a series chain somewhere
    - Resistance will be the sum of resistances

- Delay is R Cload
  - Two input gates 2Rtrans
  - Three input gates 3Rtrans

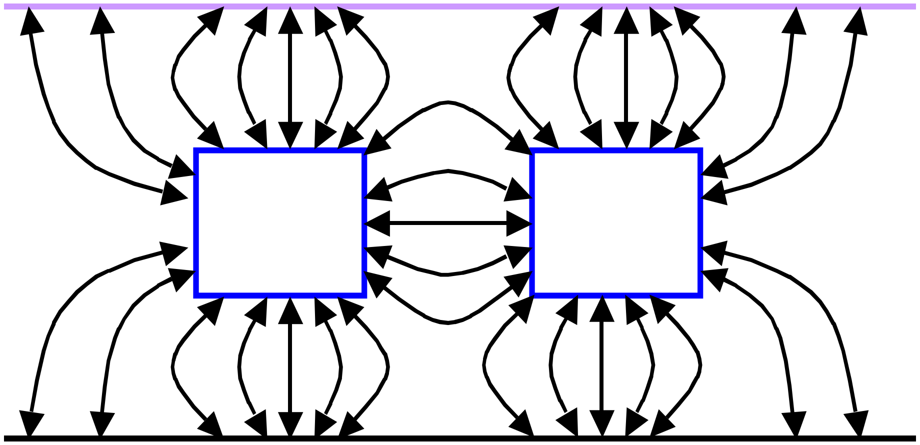
- Higher resistance
  - Slower gates



Load Capacitance

- C_{load} comes from three factors:
  1. Gate capacitance of driven transistors.
  2. Diffusion capacitance of source/drain (contacted source / drain to the body)
  3. Wire capacitance
- Today, a 1-2µ technology is the really cheap technology that students use, and advanced processes are running at 0.13µ to 0.09µ.
- I use metrics that don't change much with technology scaling (Rsq, and Cap/µ), but you should always find the correct numbers for the technology that you will use before starting a design. And, since you don't want to extract the C_{load} numbers by hand, make sure that the CAD tools have the right numbers too..

Real Wires



- Are not parallel plate capacitors.
  - Closest conductor is the neighboring wires
  - But capacitance still will be proportional to length

- Capacitance to neighboring wires is called coupling capacitance
  - Can inject noise (neighbor switches when you are quiet)
  - Can increase delay (neighbor is switching in opposite direction)

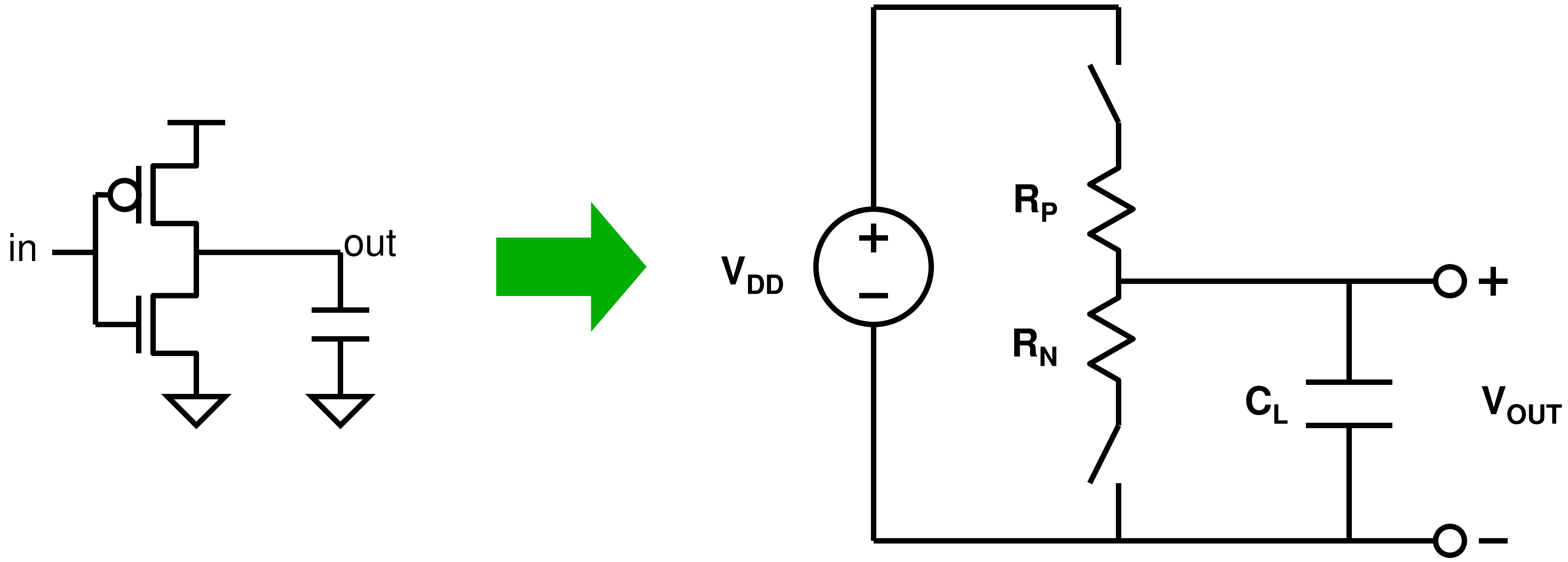
Rules of Thumb for Capacitance

Transistor Cap	Capacitance per μ of W	
	1μ	90nm
Cg - gate	2.0 fF	1.2fF
Cd - ndiff	2.0 fF	1.2fF
Cd - pdiff	2.0 fF	1.2fF

Wire Cap	Capacitance per μ		Length when C=Cinv	
	1μ	90nm	1μ	90nm
Poly wiring	0.2fF	0.15fF	40μ	3μ
Metal 1	0.3fF	0.25fF	27μ	2μ

Capacitive Charge and Discharge

- Major form of power consumption in CMOS
- Resistive networks of transistors charge and discharge capacitors
- Power is only dissipated when the output changes state



The Hard Way To Solve The Problem

- Current flows from supply though PMOS to charge

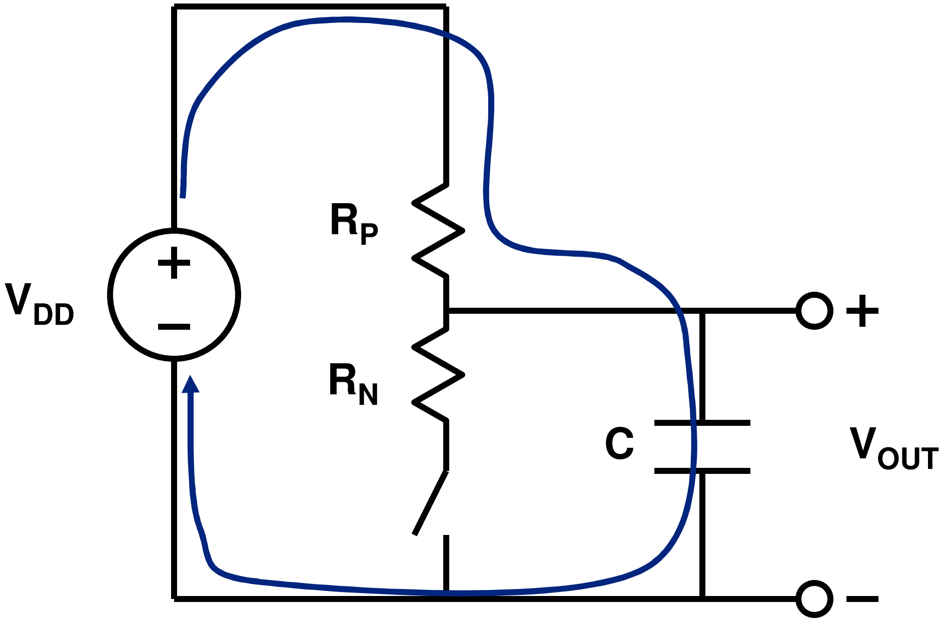
$V_{OUT}=V_{DD}(1 - e^{-t/RpC})$

$I_{DD}=(V_{DD}-V_{OUT})/Rp$

$I_{DD}=V_{DD}/R_P(e^{-t/RpC})$

$\text{Power} = I_{DD}{}^{*}V_{DD}$

$$\text{Avg Power} = \frac{\int_0^{t_{CYC}} V_{DD}^2/R_P(e^{-t/RpC})dt}{t_{CYC}}$$



$\text{Avg Power} = C\ V_{DD}^2\ f\ (1 - e^{-t_{CYC}/R_pC})$

$\text{Avg Power} \approx C\ V_{DD}^2\ f \quad \text{if } t_{CYC} >> R_PC$

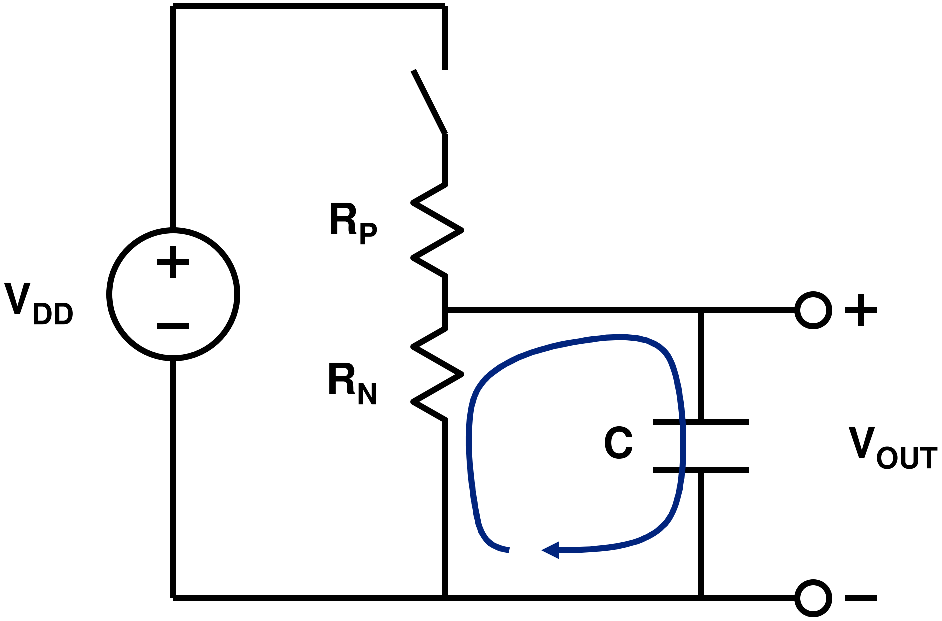
Not a function of R_P

Capacitive Charge: 1->0 Transition

- Current flows from supply though NMOS to discharge
- Current path is entirely on chip
- No current drawn from supply

$V_{OUT}=V_{DD}(\ e^{-t/RnC})$

$I_{DD}=0$



The Better Way

- Power = Energy/time
- Energy = Q * V
  - Say it takes energy to add charge to a voltage source, and the amount of energy needed is proportional to V, Q
- For a gate
  - When you charge up the output (0->1), take a charge C* Vdd out of the Vdd supply (since it is now on the output capacitance)
  - When you discharge the output (1->0) that charge is returned to Gnd
  - Gate dissipated $CVdd^2$ energy

CMOS Dynamic Power

- For each 0->1->0 cycle of a node
  - Takes $CVdd^2$ energy
- Let α = # transitions / clock cycle
  - α is generally less than one for most circuits
- Power = ½ α $CVdd^2$ F
- WHAT ABOUT THE A CLOCK NODE?
  - α is ? for a clock node (Pls. fill out)