Spiral 2-7

Capacitance, Delay and Sizing

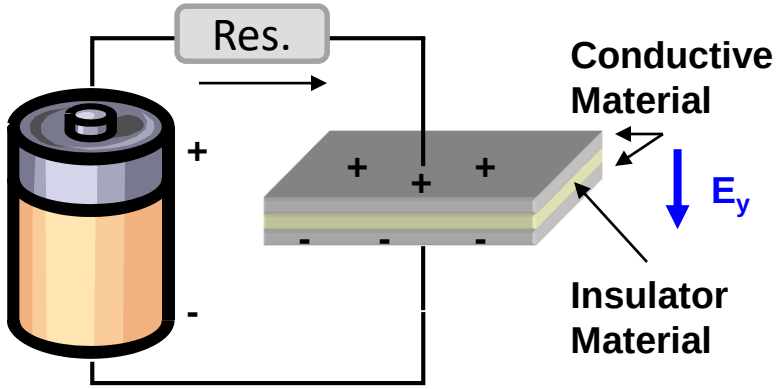
Learning Outcomes

- I understand the sources of capacitance in CMOS circuits
- I understand how delay scales with resistance, capacitance and voltage
- I can determine appropriate width of PMOS and NMOS transistors based on the configuration of the transistors and given current conduction parameters
- I understand how fan-in and fan-out affect the delay of a circuit
  - I understand how to use sizing to drive larger fan-out loads
- I understand the sources of static and dynamic power consumption and how they are affect by changes in various parameters

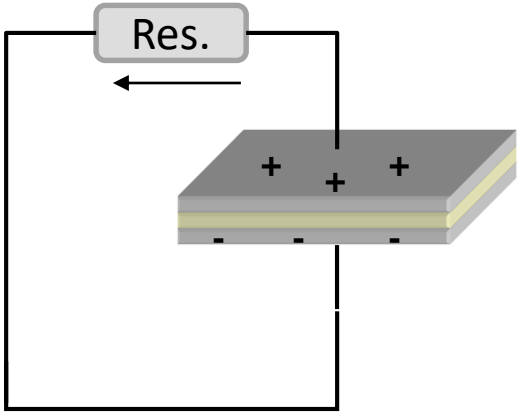
WHAT IS CAPACITANCE?

Capacitance

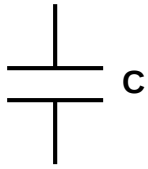
- Capacitors are formed by separating two conductive substances with an insulator
- Capacitors "store" charge
- Capacitance measures how much charge is needed to achieve a certain voltage (electric potential)
  - C = Charge (Q) / Voltage (V)
- Capacitance measured in Farads (F)



Connected to a source, charge will be stored on the conductive plates creating a positive voltage between the conductive plates



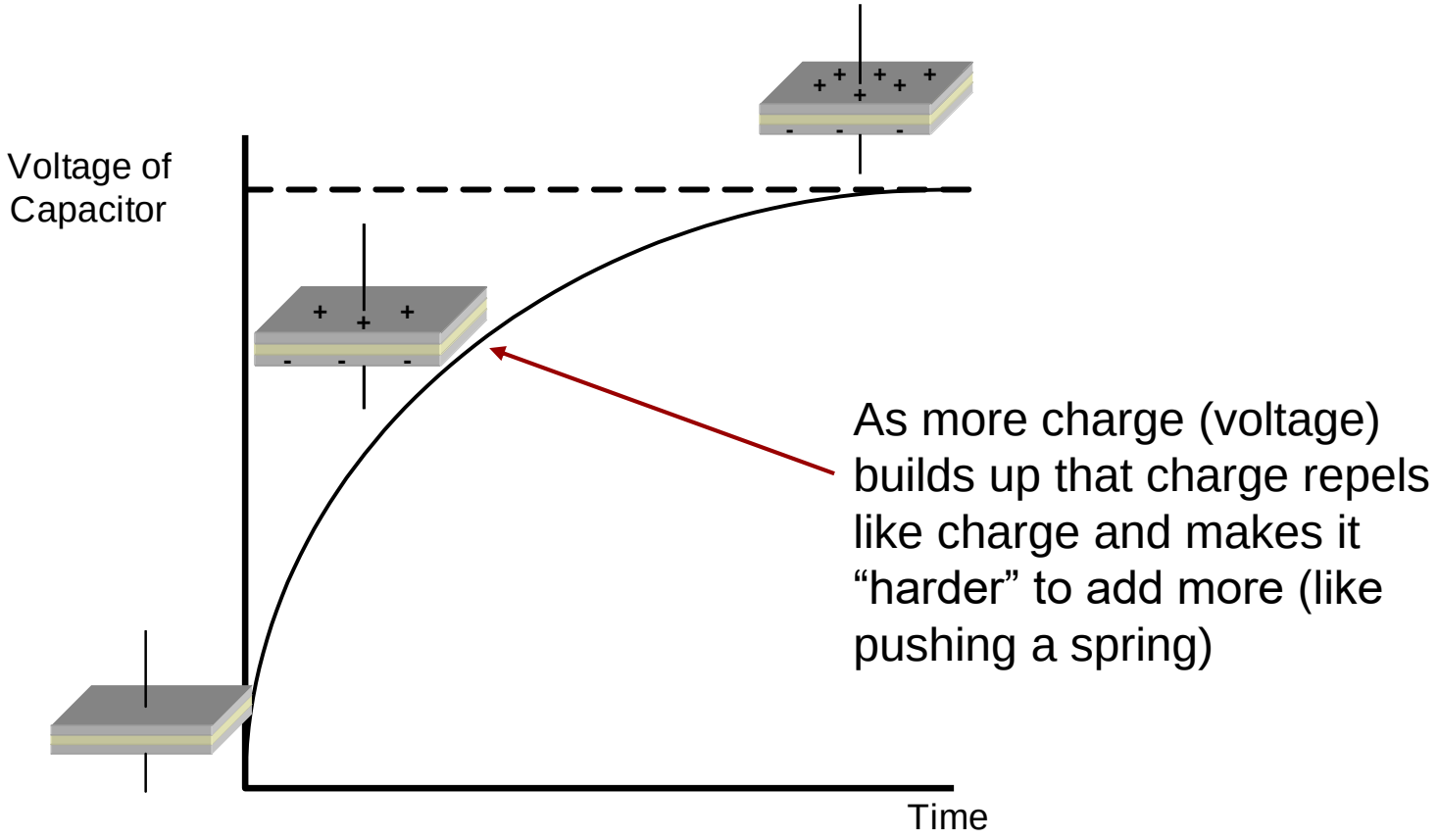
To change the voltage at the capacitor we must change the voltage (if we turn off the voltage source charge will drain off the capacitor)



Capacitor Schematic Symbol

Charging/Discharging Capacitors

- Charging a capacitor gets more “difficult” as more charge is added
- Eventually, no more charge can be added and the capacitor acts like an open circuit



Capacitor I-V Relationship

- Fact $C = \frac{Q}{V}$, or $Q = CV$
- Also recall $i = \frac{Q}{t} = \frac{dQ}{dt}$
- Thus, substituting $i = \frac{dQ}{dt} = C\frac{dV}{dt}$
- Current is linearly related (slope = C) to the ***change*** in voltage (not the absolute voltage)
  - No voltage change (constant voltage) means no current will flow

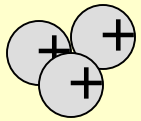
Measures of Capacitance

- $C = \frac{A\varepsilon}{d}$
  - ε is the permittivity of the insulator substance (intrinsic material property)
    - ε defined as 1 for a vacuum
    - Silicon dioxide (separates gate from silicon) = 3.9
    - Pure silicon = 11.68
  - A is the area of the conductive materials
  - d is the separation distance (or thickness of the capacitor)

First-order RC circuit step response

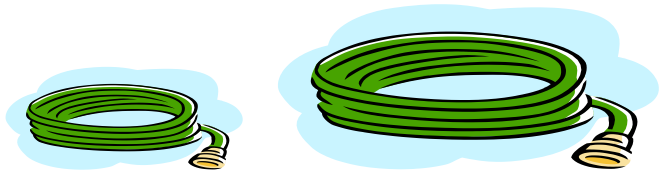
RC CIRCUIT ANALYSIS

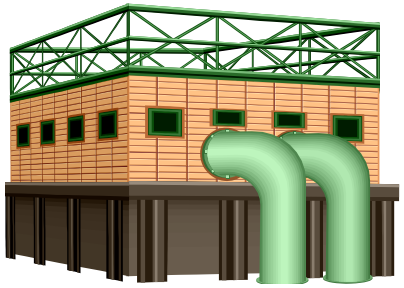
Resistance / Capacitance Analogy



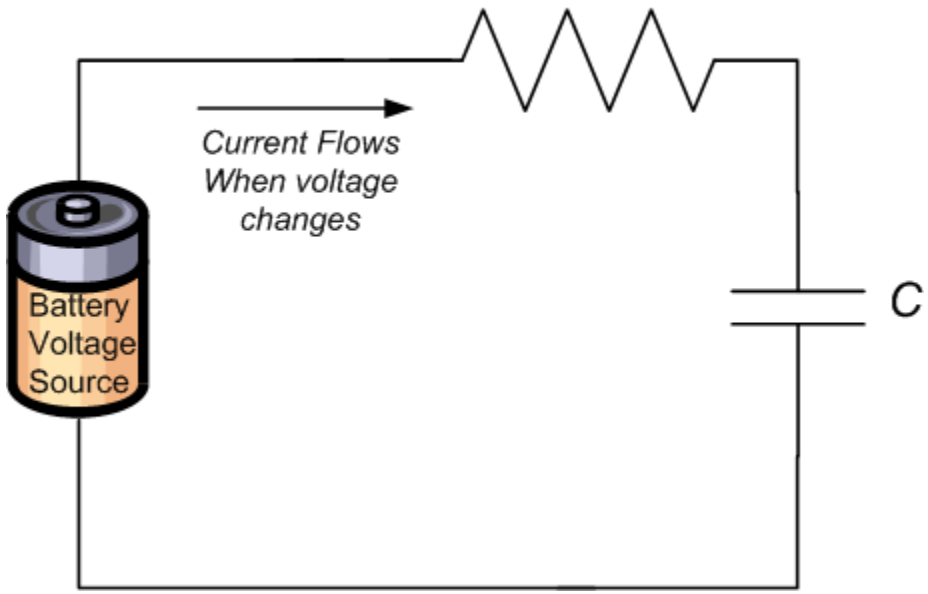
Charge = Water

Resistance = Limit of water flow





Voltage Source = Water Pressure



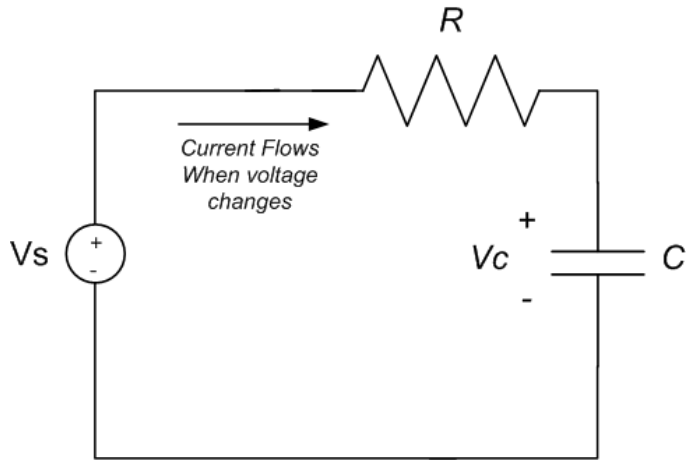
Capacitance = Total Water Needed

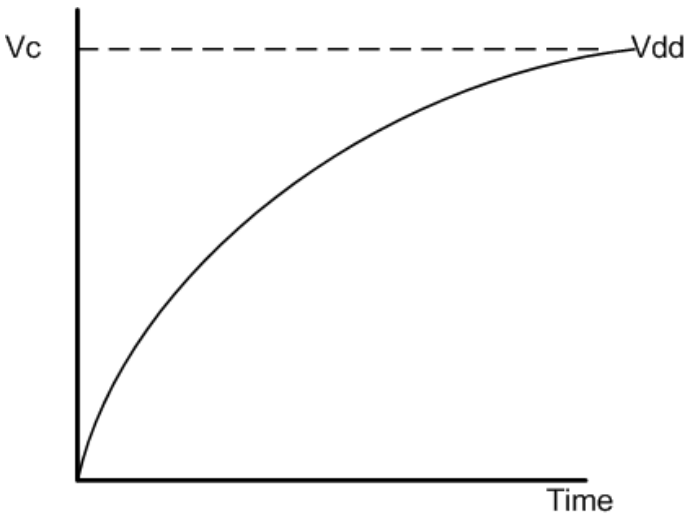
Switching Time = Time to fill or drain the capacitor (“bucket”) of charge

Thus, increase the voltage or decrease the resistance/capacitance.

Voltage, Resistance & Capacitance

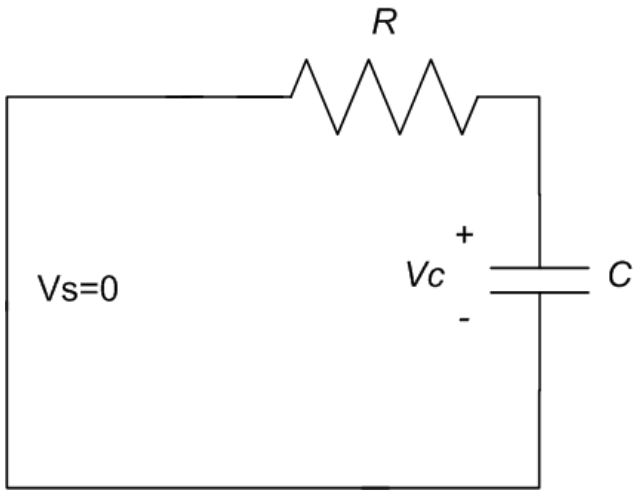
- Let's analyze a simple circuit
  - Known as an RC circuit (resistor & capacitor in series)
  - Assume t < 0, Vs = 0 then Vc=0
  - For t > 0, Vs = Vdd (voltage source turns on)
  - Current through R must be same as current "through" C
    - $i = \frac{V_R}{R} = C\frac{dV_c}{dt} \quad \Rightarrow \quad i = \frac{V_{dd}-V_c}{R} = C\frac{dV_c}{dt}$
  - Now let's solve for dVc/dt
    - $\frac{dV_c}{dt} = \frac{V_{dd}-V_c}{RC}$
  - We can solve this differential equation
    - $V_c(t) = V_{dd} + [V_c(0) - V_{dd}]e^{-\frac{t}{RC}}$
    - For Vc(0)=0 we have $V_c(t) = V_{dd}\left[1 - e^{-\frac{t}{RC}}\right]$

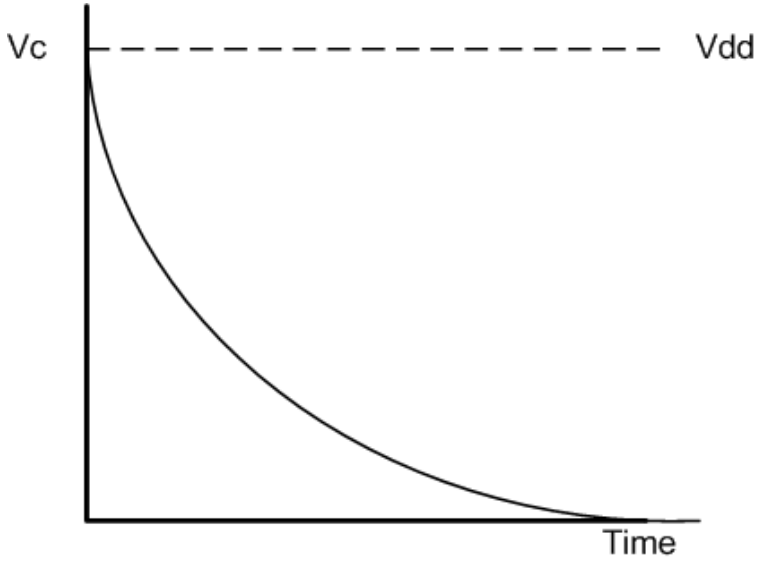




Voltage, Resistance & Capacitance

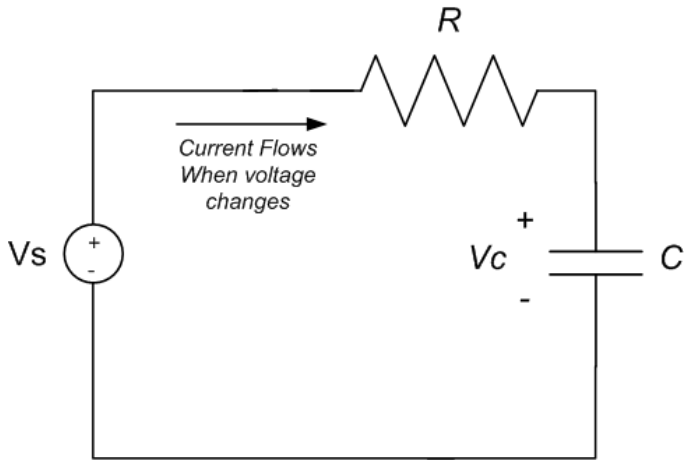
- Let's analyze a simple circuit
  - Known as an RC circuit (resistor & capacitor in series)
  - Assume t < 0, Vs = Vdd then Vc=Vdd
  - For t > 0, Vs = 0 (voltage source = GND)
  - Current through R must be same as current "through" C
    - $\frac{V_C}{R} = -C\frac{dV_c}{dt} \quad \Rightarrow \quad i = \frac{V_{dd}-V_c}{R} = C\frac{dV_c}{dt}$
  - Now let's solve for dVc/dt
    - $\frac{dV_c}{dt} = -\frac{V_c}{RC} \quad \Rightarrow \quad \frac{dV_c}{dt} + \frac{V_c}{RC} = 0$
  - We can solve this differential equation
    - $V_c(t) = V_c(0)e^{-\frac{t}{RC}}$
    - For Vc(0)=Vdd we have $V_c(t) = V_{dd} \cdot e^{-\frac{t}{RC}}$

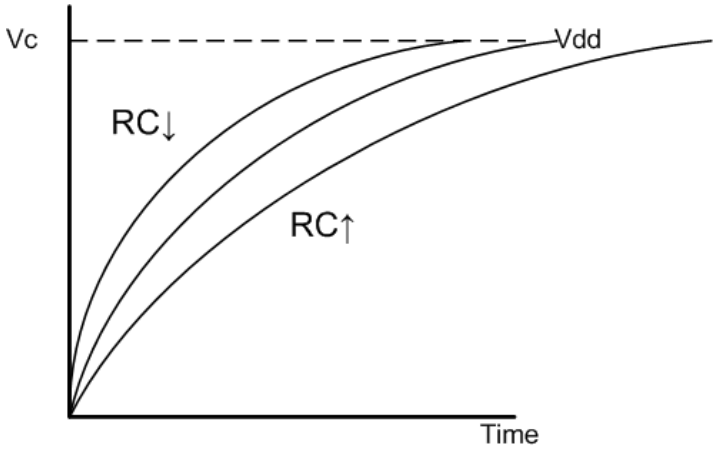




Time Constant

- Notice the charging (discharging) time is determined by product of R*C
- We refer to this as the time constant, τ
  - τ = RC
- As the product of RC increases we get slower switching times
- We can show that the time it takes to charge/discharge a capacitor to a fraction of Vdd is given in the table below



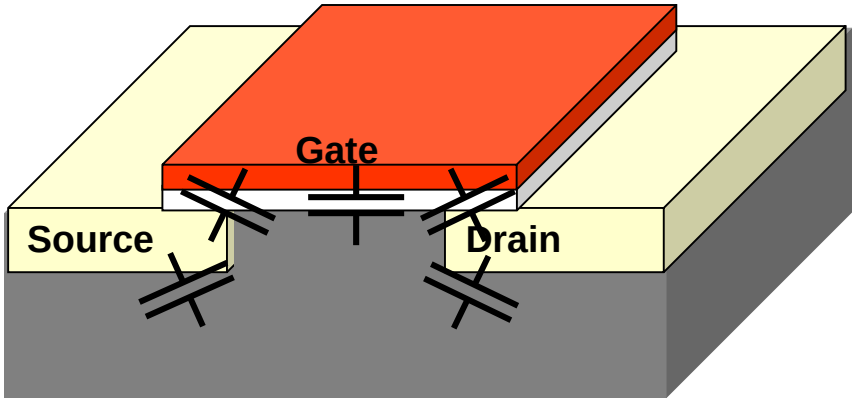


Voltage Range	Time
0 to 50% (t_p = prop. delay)	0.69*RC
0 to 63% (τ)	RC
10% to 90% (t_r=rise time/delay)	2.2*RC

DELAY

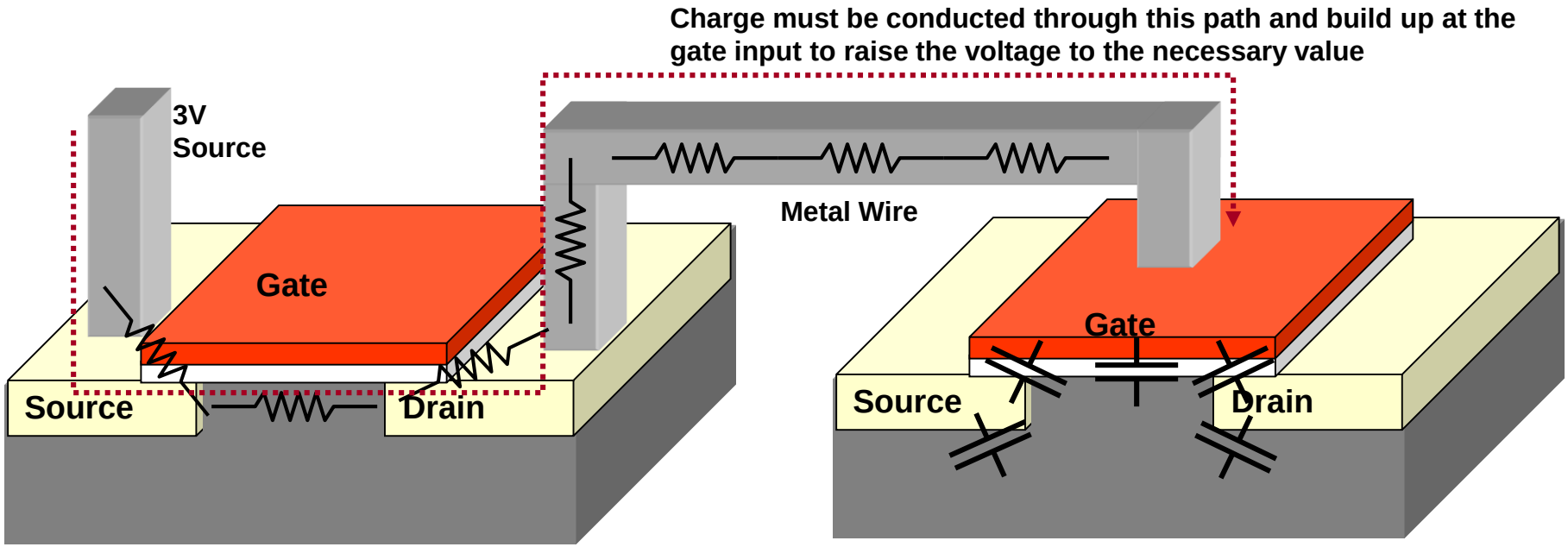
MOSFET Parasitic Capacitance

- In order to examine the delay of a MOSFET, we have to determine nature and amount of parasitic capacitance associated with MOS transistor
  - Parasitic: Unintentional, naturally occurring capacitors
  - It's not that we want caps, we're stuck with them due to the structure of the MOSFET.
- The oxide layer separating gate and substrate is a more obvious capacitor
- However the depletion regions around the source and drain also form capacitors



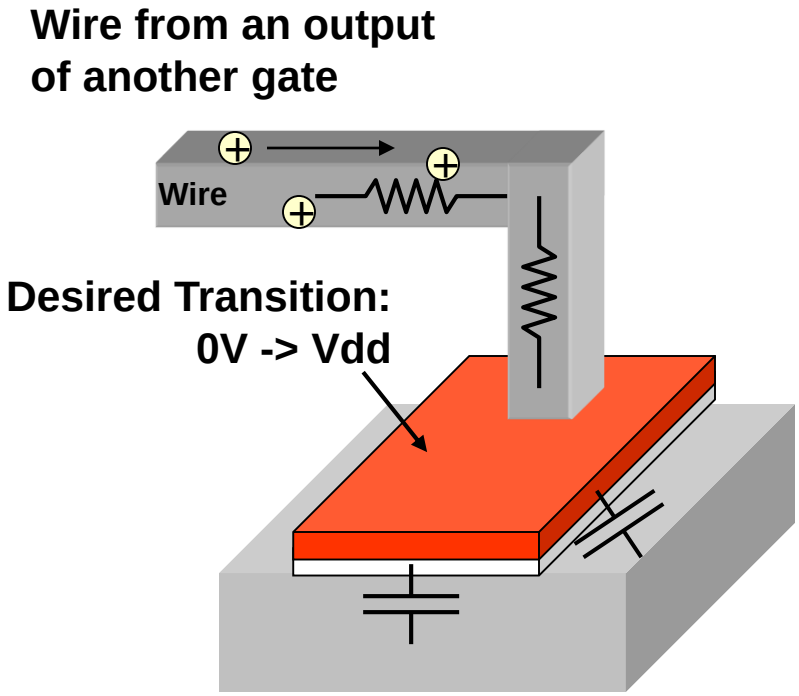
Transistor R and C values

- Observation: Output of one transistor usually drives input of another
  - Sources of resistance
    - Channel between source and drain
    - Wire connecting drain to gate of next transistor
  - Sources of capacitance
    - Gate input of the next transistor
    - Other small capacitances



Resistor and Capacitor Delay

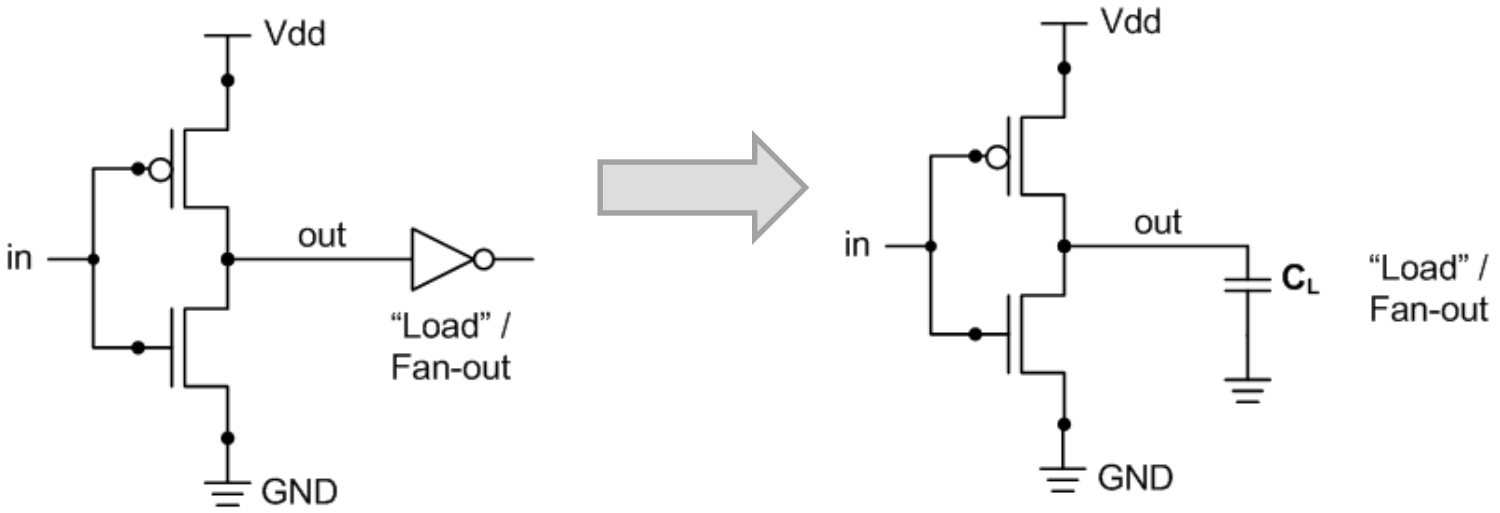
- Outputs connect to other inputs
- To change the output voltage (really the input of the next gate), we must conduct enough charge to raise or lower its present voltage
- Resistance limits the amount of charge that can be transferred per unit time
- Capacitance determines how much charge must be present to attain a certain voltage
- Time it takes to attain a certain voltage is proportional to R*C



Switching Time ~ Resistance*Capacitance

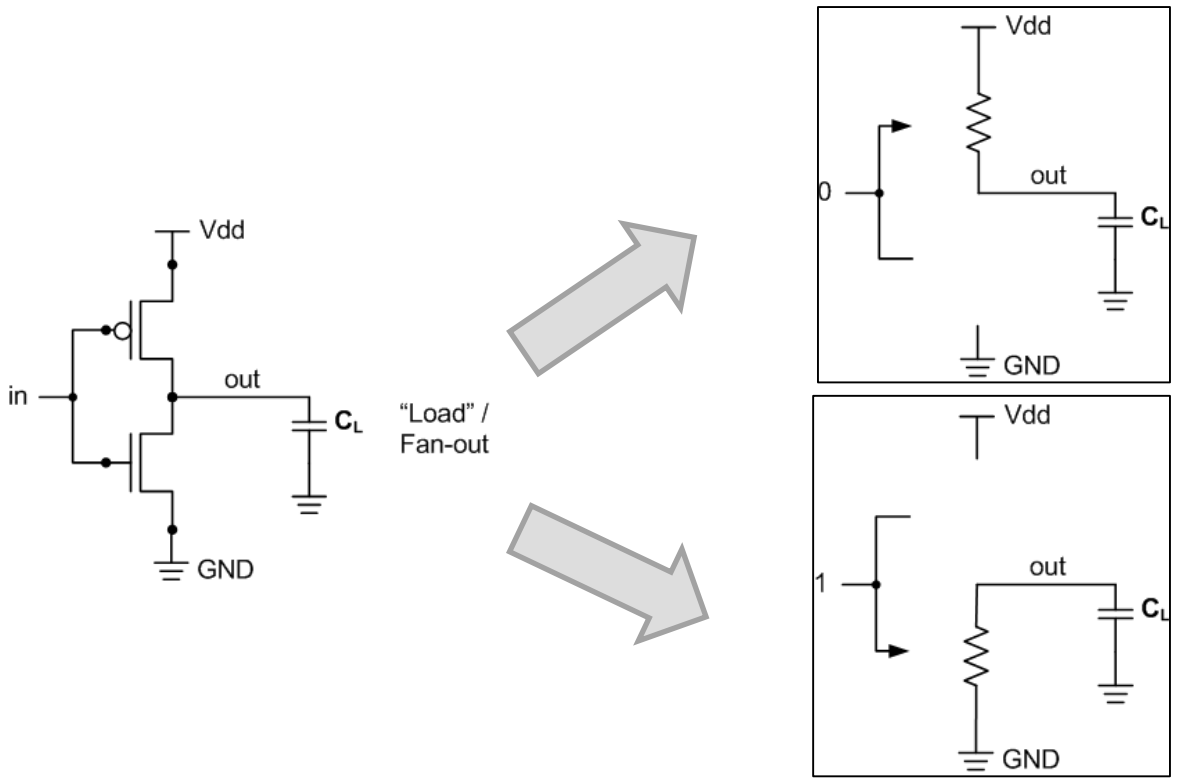
Where RC Circuits Occur

- Consider a CMOS gate driving others (loads)
  - The output connects to the gate inputs of the loads (fan-out) and thus can be modeled as a capacitive load (C_L)



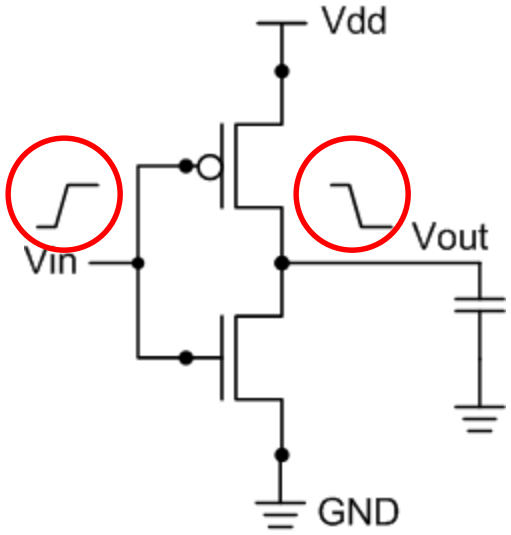
Where RC Circuits Occur

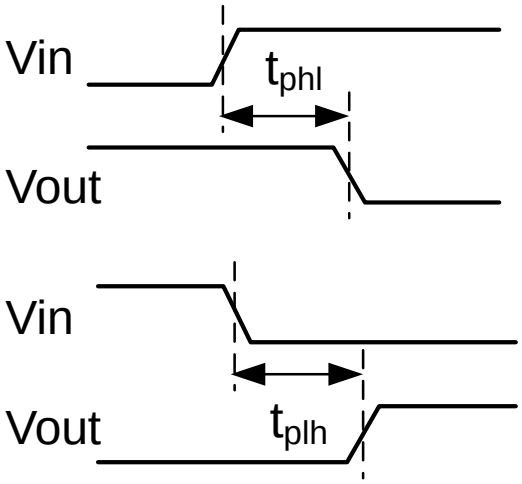
- Depending on the inverter input the PMOS or NMOS will be in resistive mode and can be modeled as resistors
  - Thus we have an RC circuit either charging or discharging the output (C_L)



Defining Delays

- We model
  - The next gate(s) and other parasitic capacitances as a lumped capacitance
  - The PDN or PUN transistors as resistors (since they are operating in linear mode when the input is near VDD or GND)
- t_{PLH} and t_{PHL} refer to the propagation delay of a circuit when the output changes from low to high (0→1) [t_{PLH}] or high to low (1→0) [t_{PHL}]
  - $t_{PLH} \approx 0.69 R_P C_L = \frac{C_L V_{DD}}{k_p [V_{DD} - V_{Tp}]^2}$
  - $t_{PHL} \approx 0.69 R_N C_L = \frac{C_L V_{DD}}{k_n [V_{DD} - V_{Tn}]^2}$
- We say the delay of the gate is then:
  - $t_{PD} \approx (t_{PLH} + t_{PHL}) / 2$
- Important: What reduces delay?
  - $C_L \downarrow$, W/L ↑, Vdd ↑

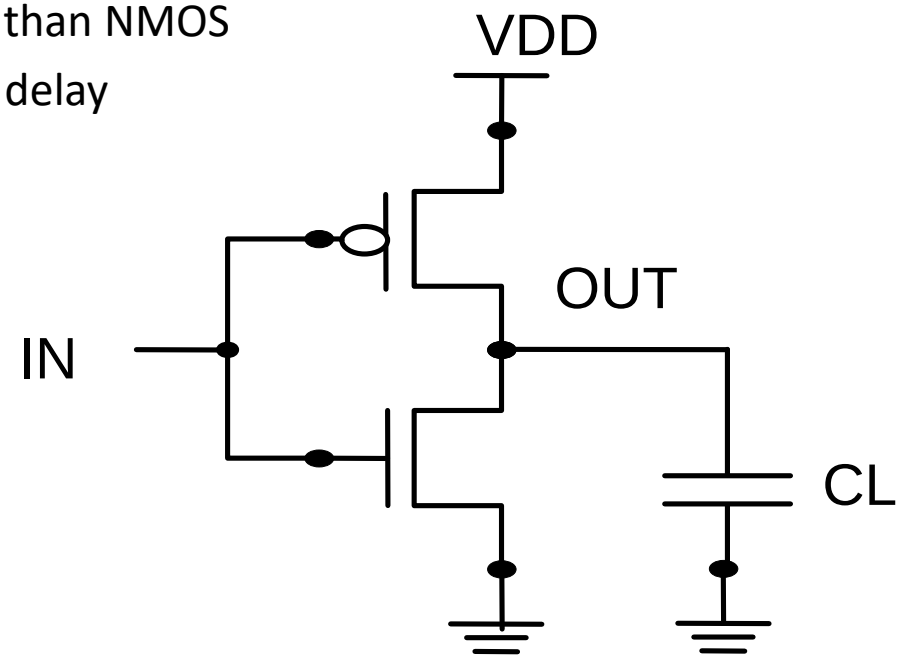




CMOS SIZING

Sizing – PMOS is Slower than NMOS!

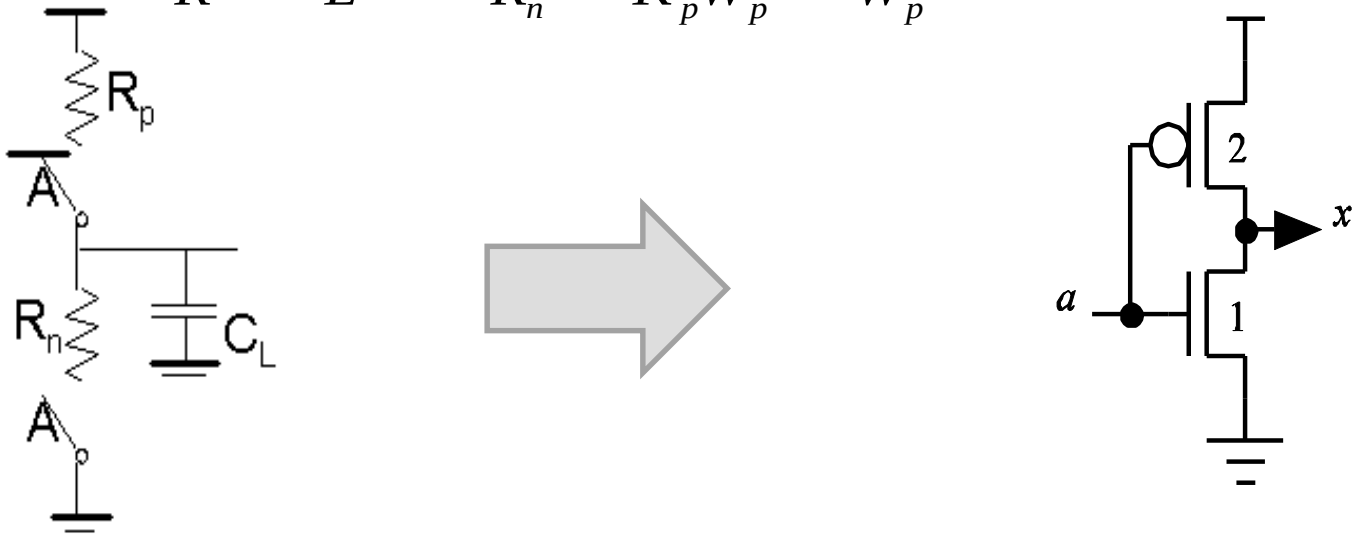
- Recall the equations for current through a transistor in linear mode
  - $|I_{ds}| = \frac{1}{2}K'\left(\frac{W}{L}\right)\left[2\left(|v_{gs}| - |V_T|\right)|V_{ds}| - |V_{ds}|^2\right]$
  - Ohm's law says I = V/R so in the equation above $\frac{1}{R} \propto K'\frac{W}{L}$
- Problem $K_N > K_P$ ($K_N \approx 2.5K_P$)
  - PMOS are worse at conducting than NMOS
  - This will leave to imbalances in delay (i.e. $t_{PLH} > t_{PHL}$)
  - To balance the delay when pulling up vs. pulling down we can play with Width
- Solution: Make $W_P > W_N$ by about a factor of 2 or 2.5



Sizing – Inverter

- Assume $K_p \approx \frac{K_n}{2}$
- Find the ratio and W_p and W_n that balances the delay of output during falling and rising transitions

$$\frac{1}{R} \propto \frac{KW}{L} \Rightarrow \frac{R_p}{R_n} \propto \frac{K_n W_n}{K_p W_p} = \frac{2W_n}{W_p} \Rightarrow W_p = 2W_n$$



INV

Sizing – Simple CMOS Gates

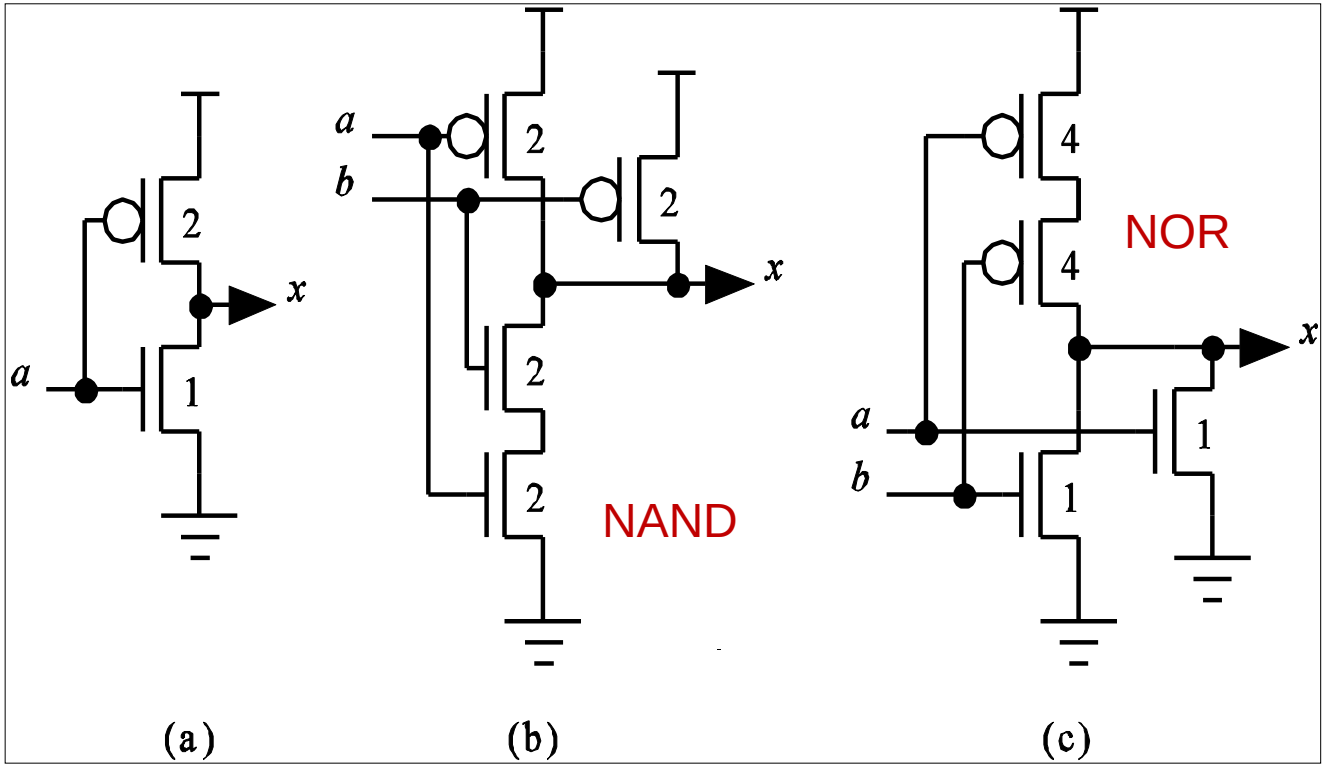
- Goal: Make any gate have the same worst case resistance as an inverter
- The ratio of the $\{W/L\}_{PUN}$/ $\{W/L\}_{PDN}$ should be about two (or higher) to make up for slow PMOS

Important Notes:

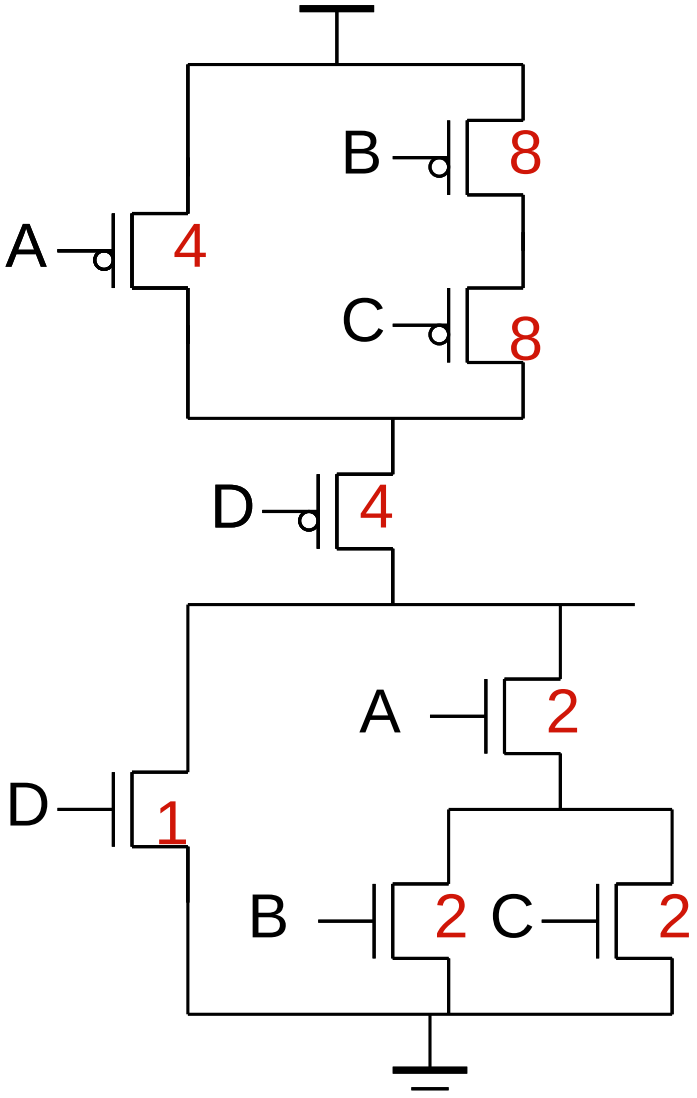
For parallel transistors consider only the case if 1 is on (Remember if R||R then Reff=R/2 which is a better case so we assume only one is on)

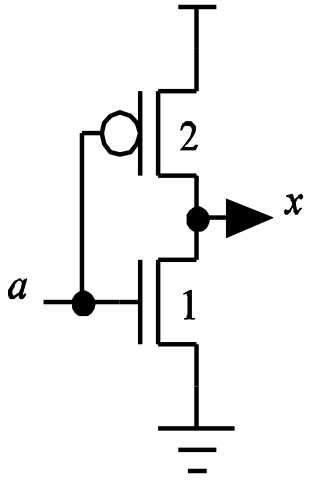
Series transistors in series add lengths/ resistance

All paths in a PxN should have same resistance



Sizing – Complex CMOS Gates

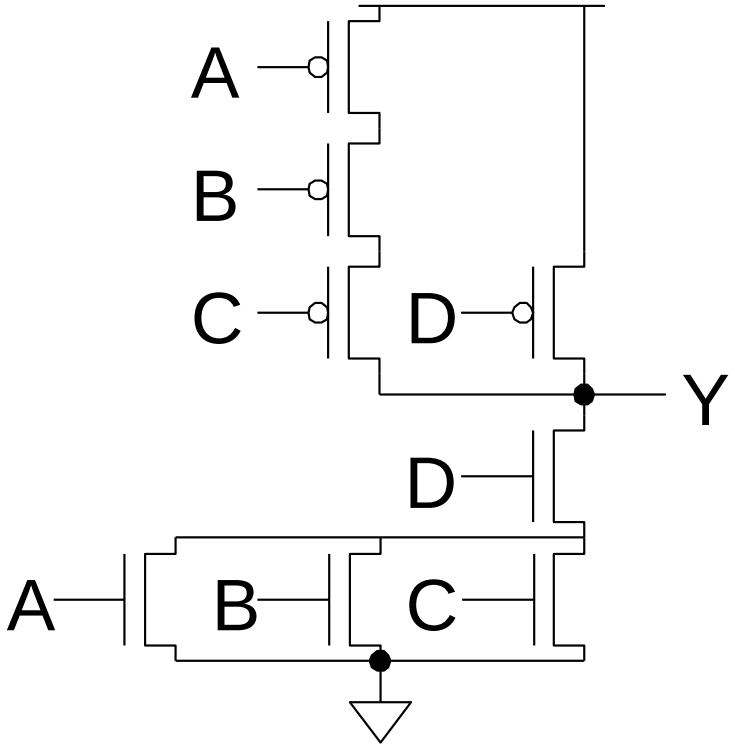




$$OUT = \overline{D + A \bullet (B + C)}$$

Compound Gate Example

$$Y = \overline{D \bullet (A + B + C)}$$



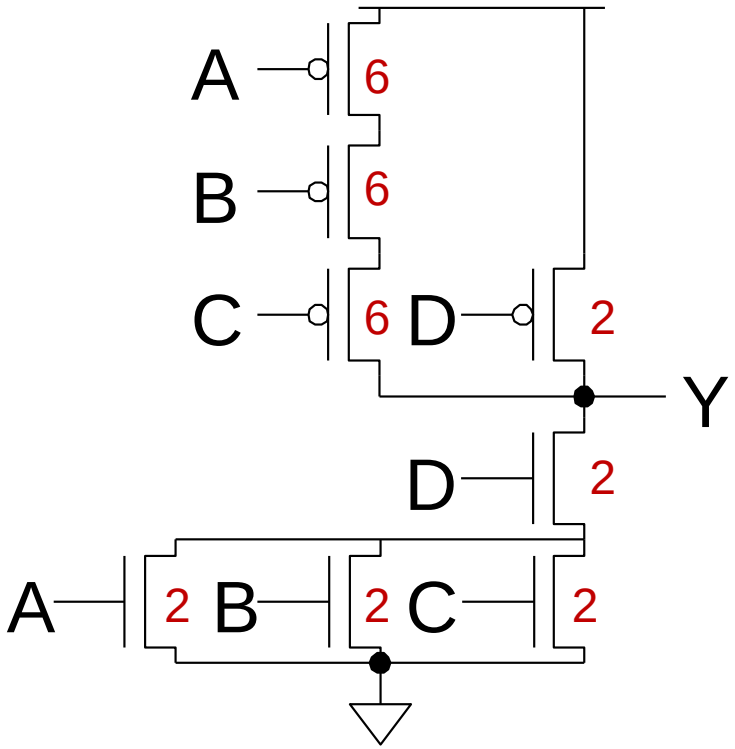
Compound Gate Example

$$Y = \overline{D \bullet (A + B + C)}$$

PUN one length from Vdd to output is L=3 so W=6 (to make W/L = 2).

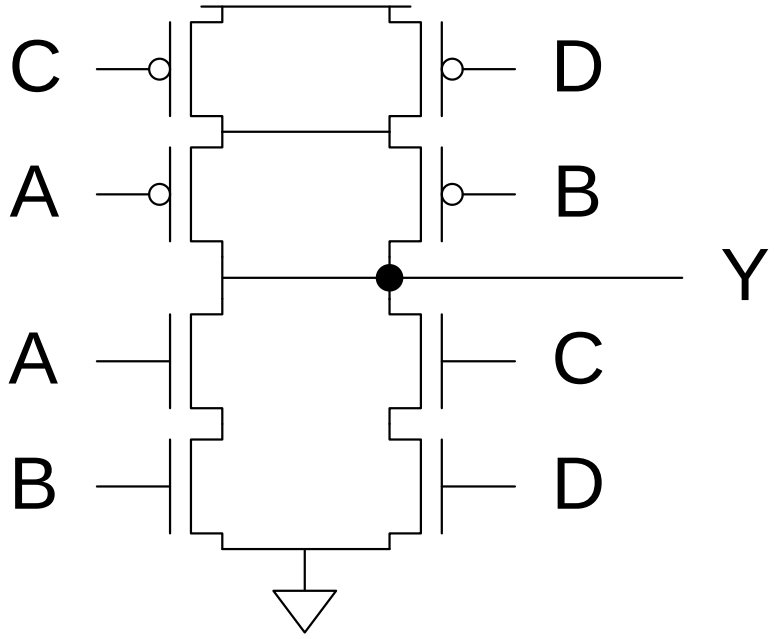
The other branch is L=1 so W=2

PDN worst case channel length from output to GND is L=2, so W=2 (to make W/L = 1)



Compound Gate Example

$$Y = \overline{AB+CD}$$

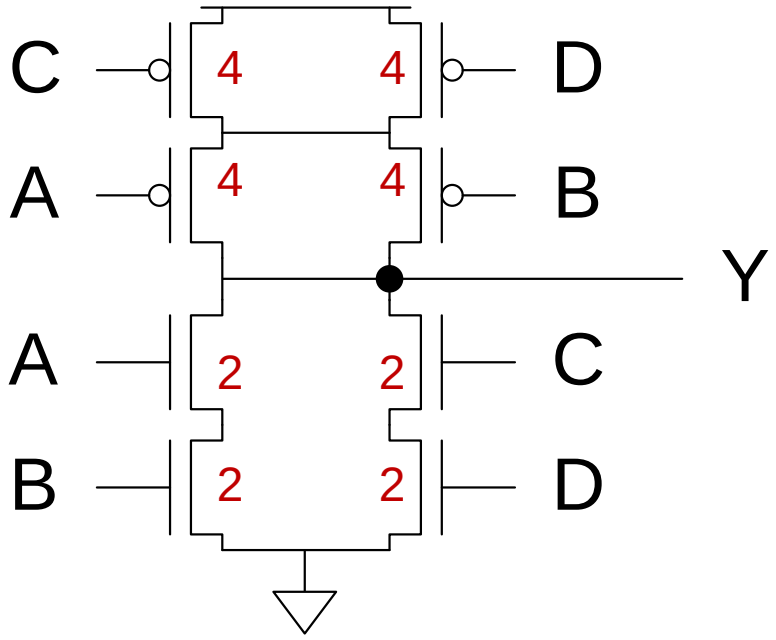


Compound Gate Example

$$Y = \overline{AB+CD}$$

PUN worst case channel length from Vdd to output is L=2, so W=4 (to make W/L = 2)

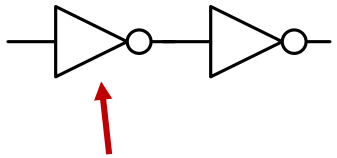
PDN worst case channel length from output to GND is L=2, so W=2 (to make W/L = 1)



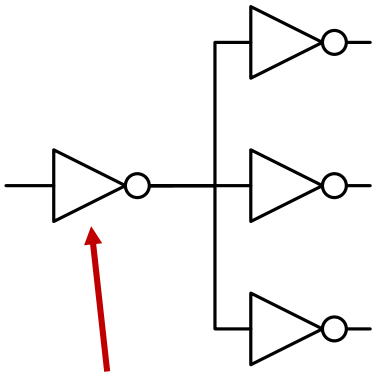
FANIN & FANOUT

Fanout

- Fanout refers the number of gates an output connects to
- As the fanout increases C_L goes up proportionally and means the delay increases



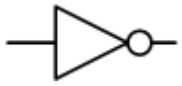
This inverter has a fanout (# of loads) = 1



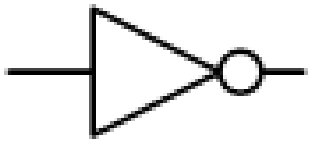
This inverter has a fanout (# of loads) = 3

Increasing Drive Strength

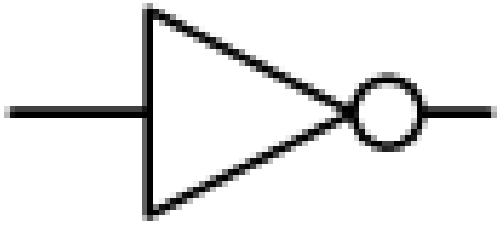
- So far we've always modeled an idea inverter for our transistor sizes
  - Ideal inverter => NMOS: 1/1 and PMOS 2/1 or 3/1
- We can counteract the increase in CL by reducing R through increasing transistor widths
  - NMOS and PMOS widths increase by 2x, 3x, 4x, 6x, 8x, etc.



1x INV ($W_n/L_n = 1$)



2x INV ($W_n/L_n = 2$)

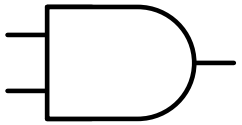


4x INV ($W_n/L_n = 4$)

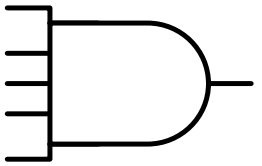
Fan-in

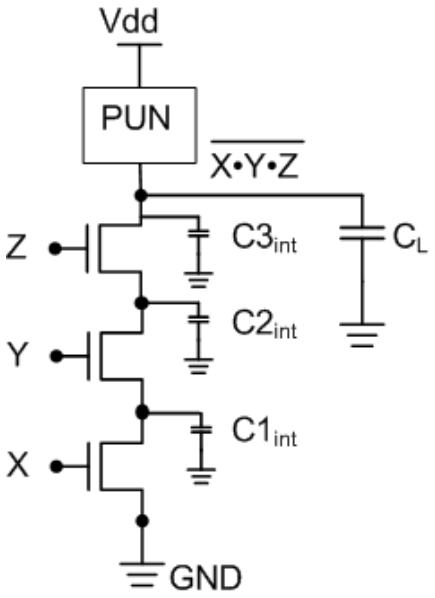
- Fan-in refers to the number of inputs to a gate
- Each input adds intrinsic, parasitic capacitance
  - $t_{phl}=0.69*R_n(C_1+2C_2+3C_3+C_L)$
    - To discharge C2 requires going 2L (i.e. $2R_N$)
    - To discharge C3 requires 3L (i.e. $3R_N$)
- This means delay grows quadratically with fan-in but linearly with fanout
  - $t_{pd} \sim a_1FI + a_2FI^2 + a_3FO$
- Important: Rarely want FI > 4

Fanin = 2



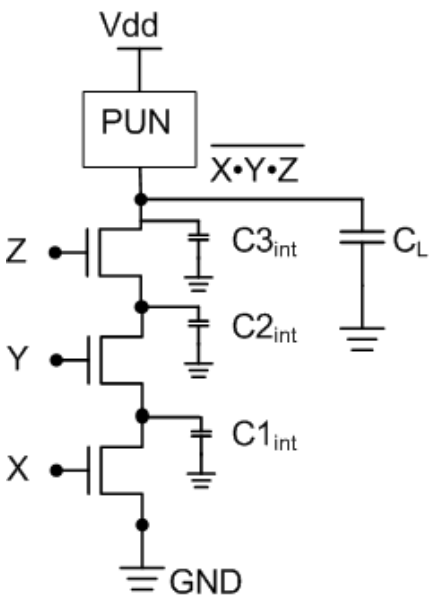
Fanin = 5





Mitigating Delay

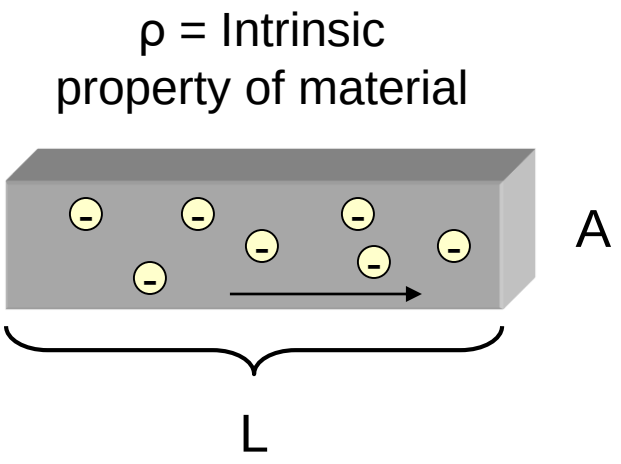
- May increase widths progressively
  - Bottom transistor has higher width
- Order transistors to allow the latest arriving input control the transistor closest to the output
  - If Z is the latest arriving input we want to put it closest to the output



INTERCONNECT DELAY

Ideal vs. Realistic Wire

- Ideal wire should have R=0 and little capacitance
- In real life it has some small R and C
- R = ρL/A
  - ρ = resistivity of material (some intrinsic property)
  - L = length or wire
  - A = cross-section area of the material
- As technology scales (we build smaller devices) L↑ and A↓ means resistance goes up a lot

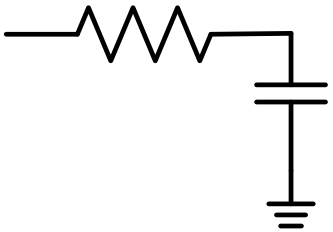


Modeling Interconnect Delay

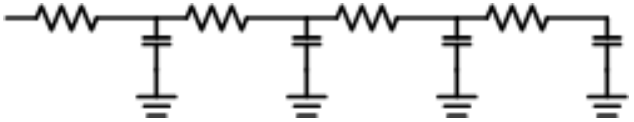
- Interconnect delay is starting to (already is) rival switching delay
- Important design considerations
  - Long wire traces slow a signal down, thus global signals on a chip require special attention
  - Clock, reset, and other signals must be routed carefully and a whole tree of buffers inserted to decrease the delay

A real wire can be modeled as…

Ideal wire



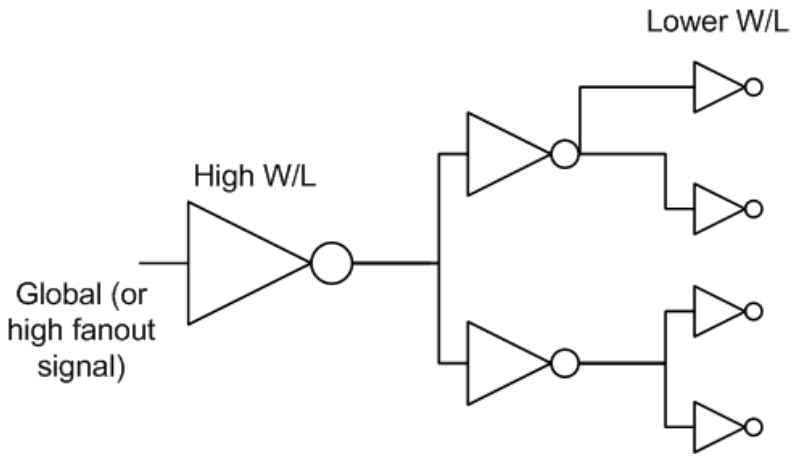
Lumped Model (overestimates delay)



Distributed Model (better estimate)

Dealing With Interconnect

- Interconnect delay is starting to (already is) rival switching delay
- Important design considerations
  - Long wire traces slow a signal down, thus global signals on a chip require special attention
  - Clock, reset, and other signals must be routed carefully and a whole tree of buffers inserted to decrease the delay



DYNAMIC POWER

Power

- Power consumption decomposed into:
  - Static: Power constantly being dissipated (grows with # of transistors)
  - Dynamic: Power consumed for switching a bit (1 to 0)
- $P_{DYN} = I_{DYN} * V_{DD} \approx \frac{1}{2} C_{TOT} V_{DD}^2 f$
  - Recall, I = C dV/dt
  - V_{DD} is the logic '1' voltage, f = clock frequency
- Dynamic power favors parallel processing vs. higher clock rates
  - V_{DD} value is tied to f, so a reduction/increase in f leads to similar change in Vdd
  - Implies power is proportional to f^3 (a cubic savings in power if we can reduce f)
  - Take a core and replicate it 4x => 4x performance and 4x power
  - Take a core and increase clock rate 4x => 4x performance and 64x power
- Static power
  - Leakage occurs no matter what the frequency is

Temperature

- Temperature is related to power consumption
  - Locations on the chip that burn more power will usually run hotter
    - Locations where bits toggle (register file, etc.) often will become quite hot especially if toggling continues for a long period of time
  - Too much heat can destroy a chip
  - Can use sensors to dynamically sense temperature
- Techniques for controlling temperature
  - External measures: Remove and spread the heat
    - Heat sinks, fans, even liquid cooled machines
  - Architectural measures
    - Throttle performance (run at slower frequencies / lower voltages)
    - Global clock gating (pause..turn off the clock)
    - None…results can be catastrophic
- Fun video: http://www.tomshardware.com/2001/09/17/hot_spot/