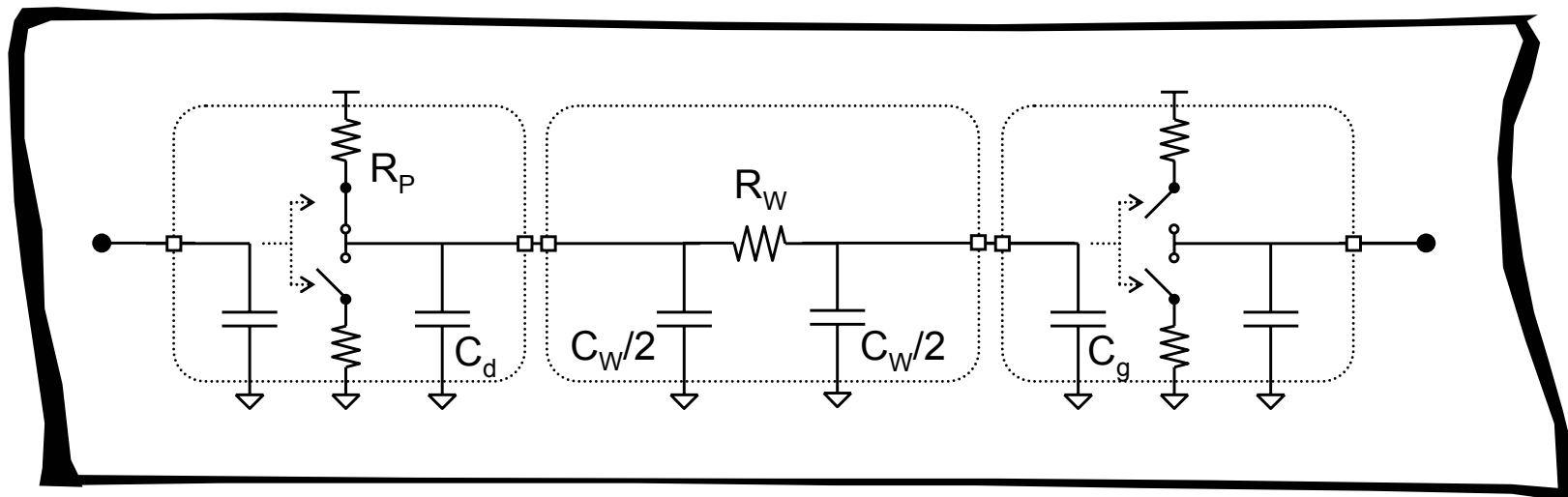
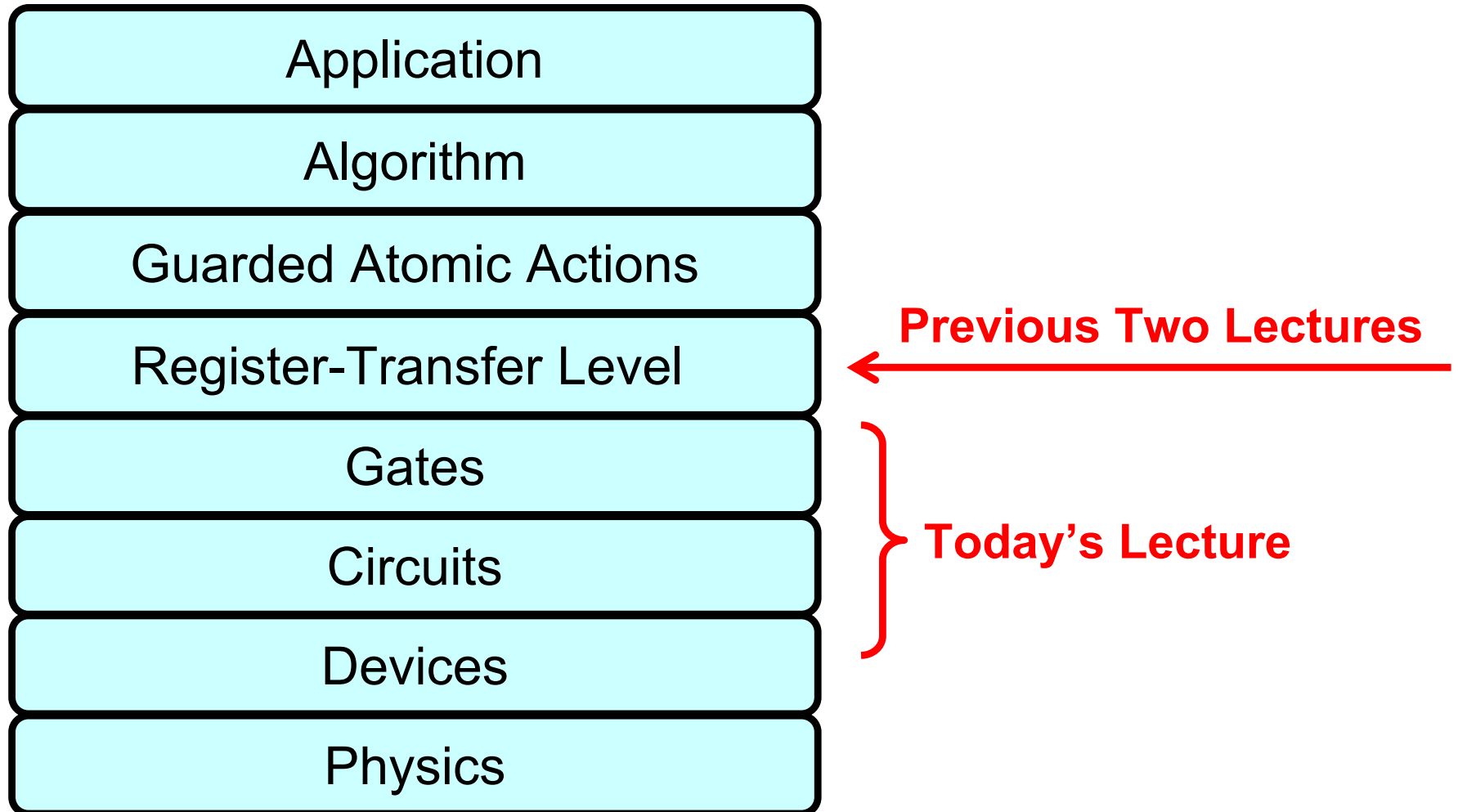


CMOS Transistors, Gates, and Wires

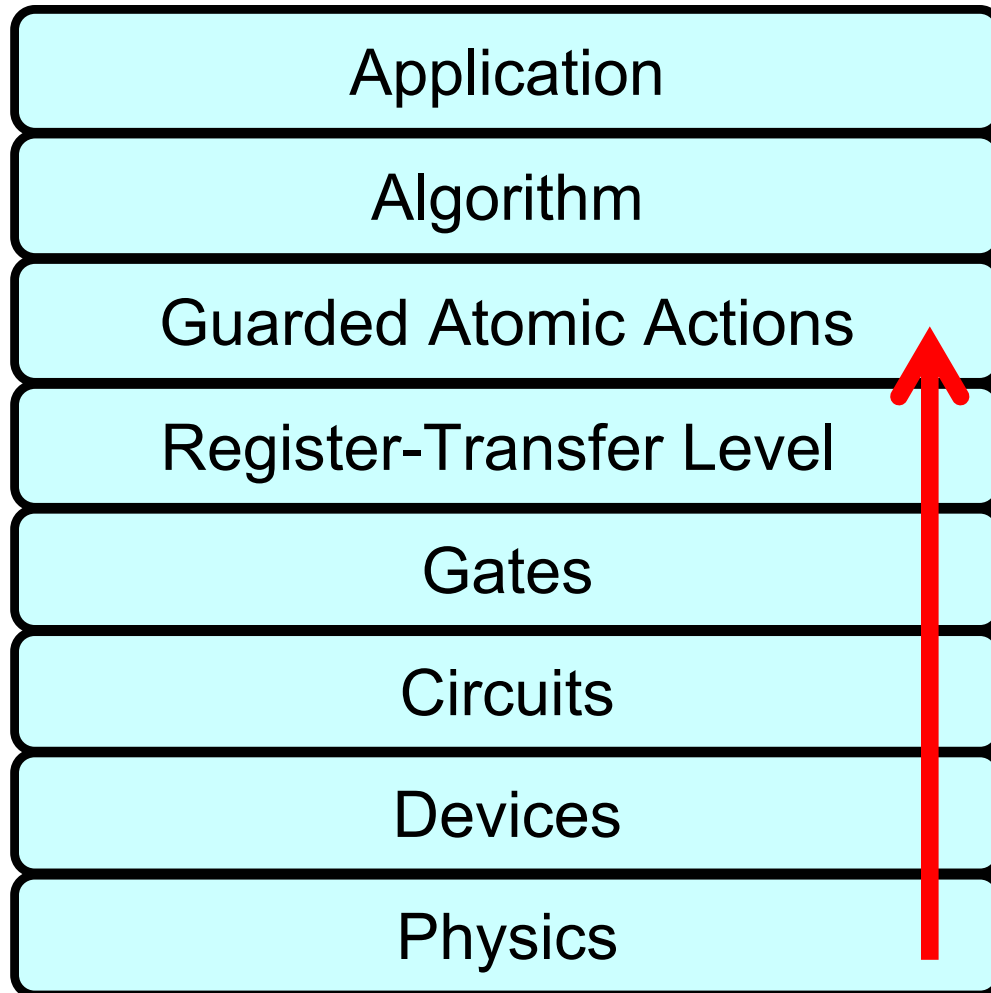


6.375 Complex Digital Systems
Christopher Batten
February 15, 2006

Should the hardware abstraction layers make today's lecture irrelevant?



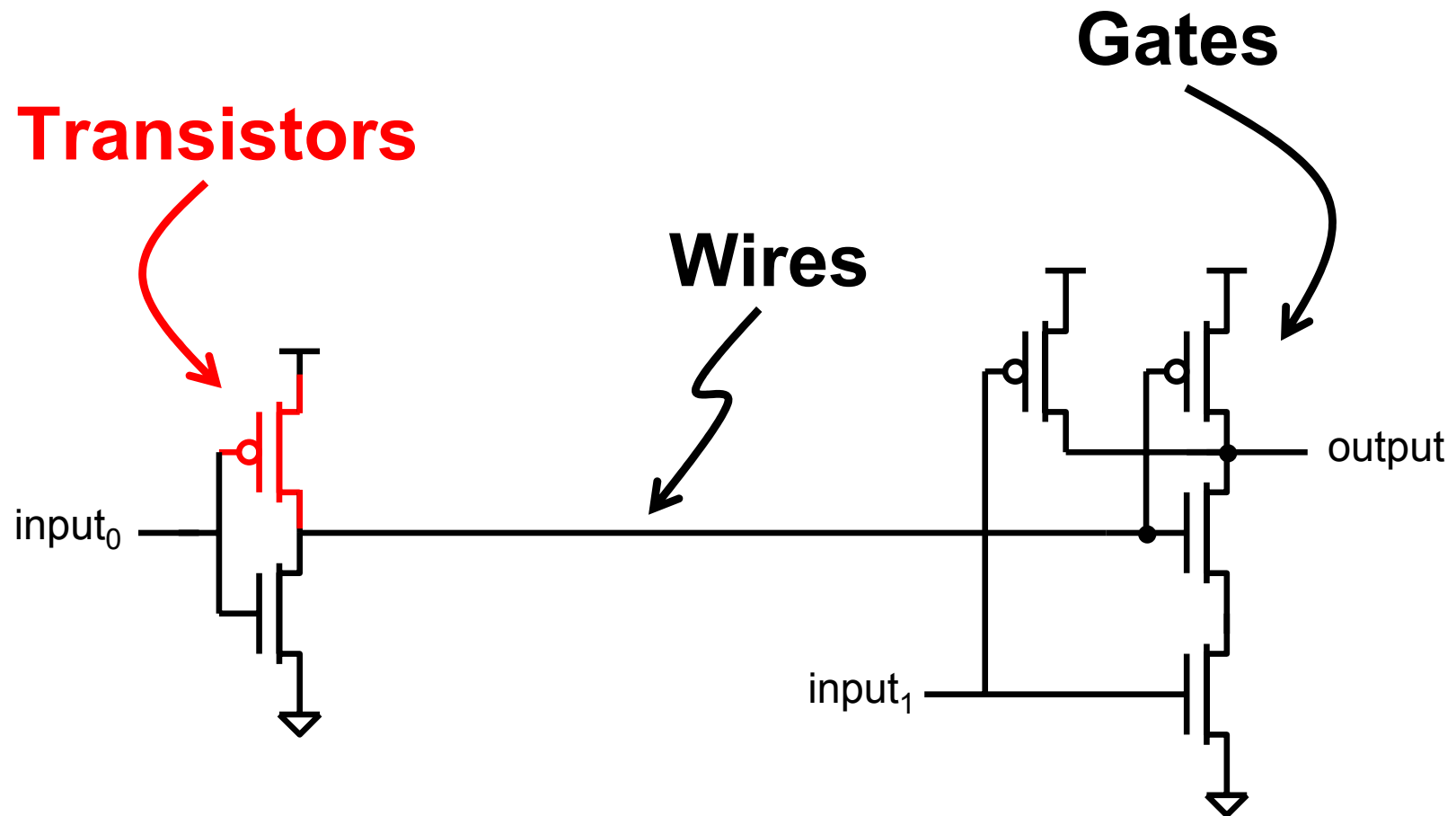
Should the hardware abstraction layers make today's lecture irrelevant?



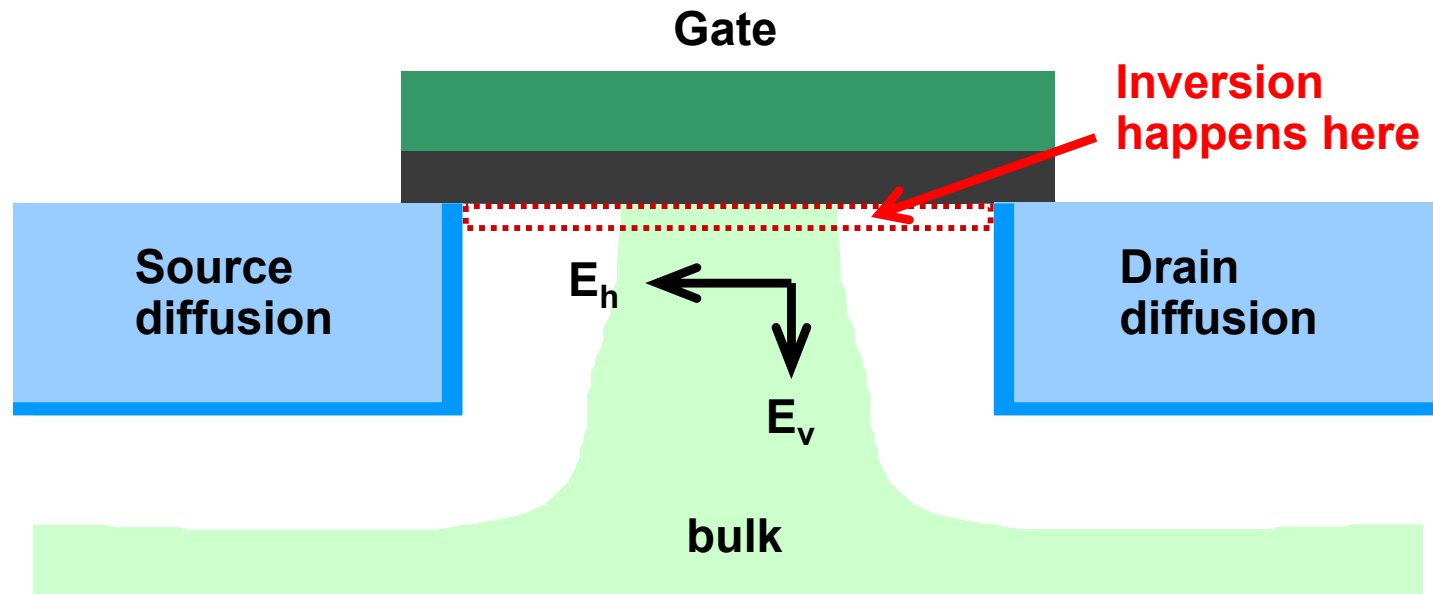
Physical design issues are increasingly pushing their way up the abstraction layers

It is essential for modern digital designers to have some intuition about the lower level physical design issues

CMOS Transistors, Gates, and Wires



Metal Oxide-Semiconductor Field-Effect (MOSFET) Transistor



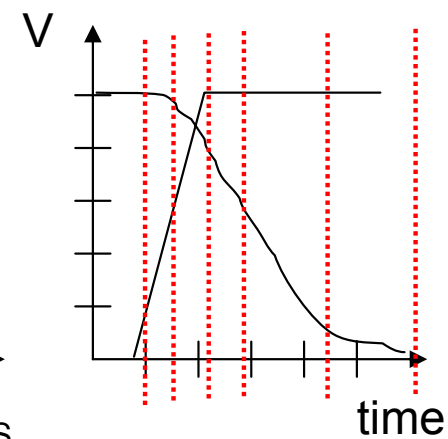
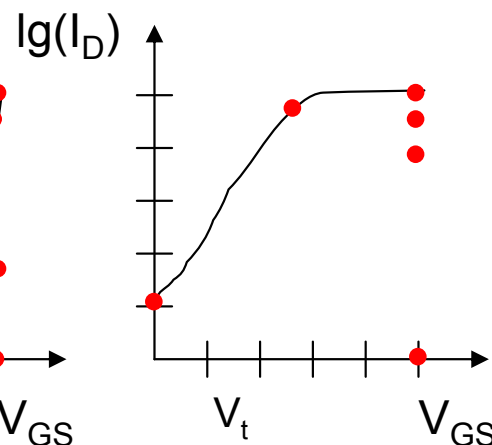
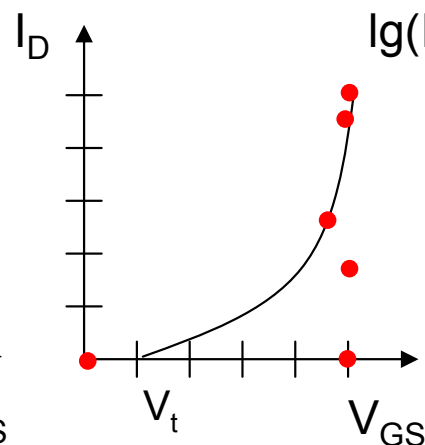
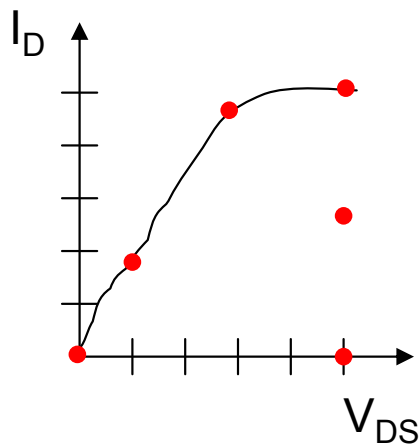
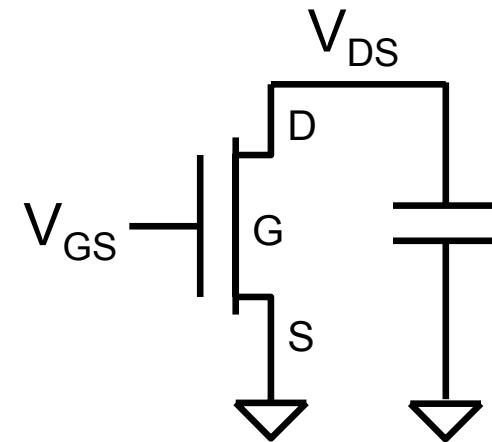
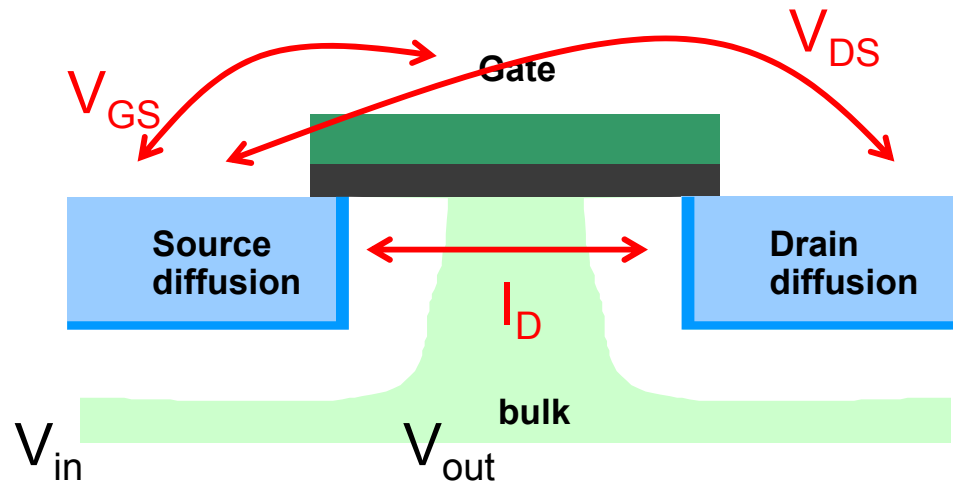
INVERSION:

A sufficiently strong vertical field will attract enough electrons to the surface to create a conducting n-type channel between the source and drain.

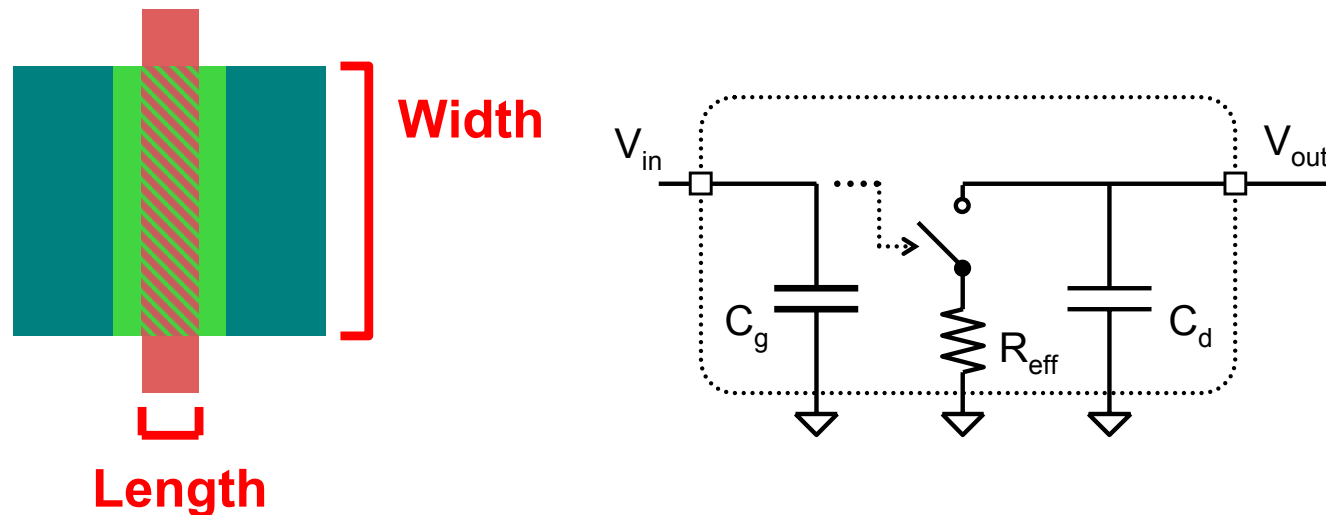
CONDUCTION:

If a channel exists, a horizontal field will cause a drift current from the drain to the source.

Overview of operation of a NMOS transistor

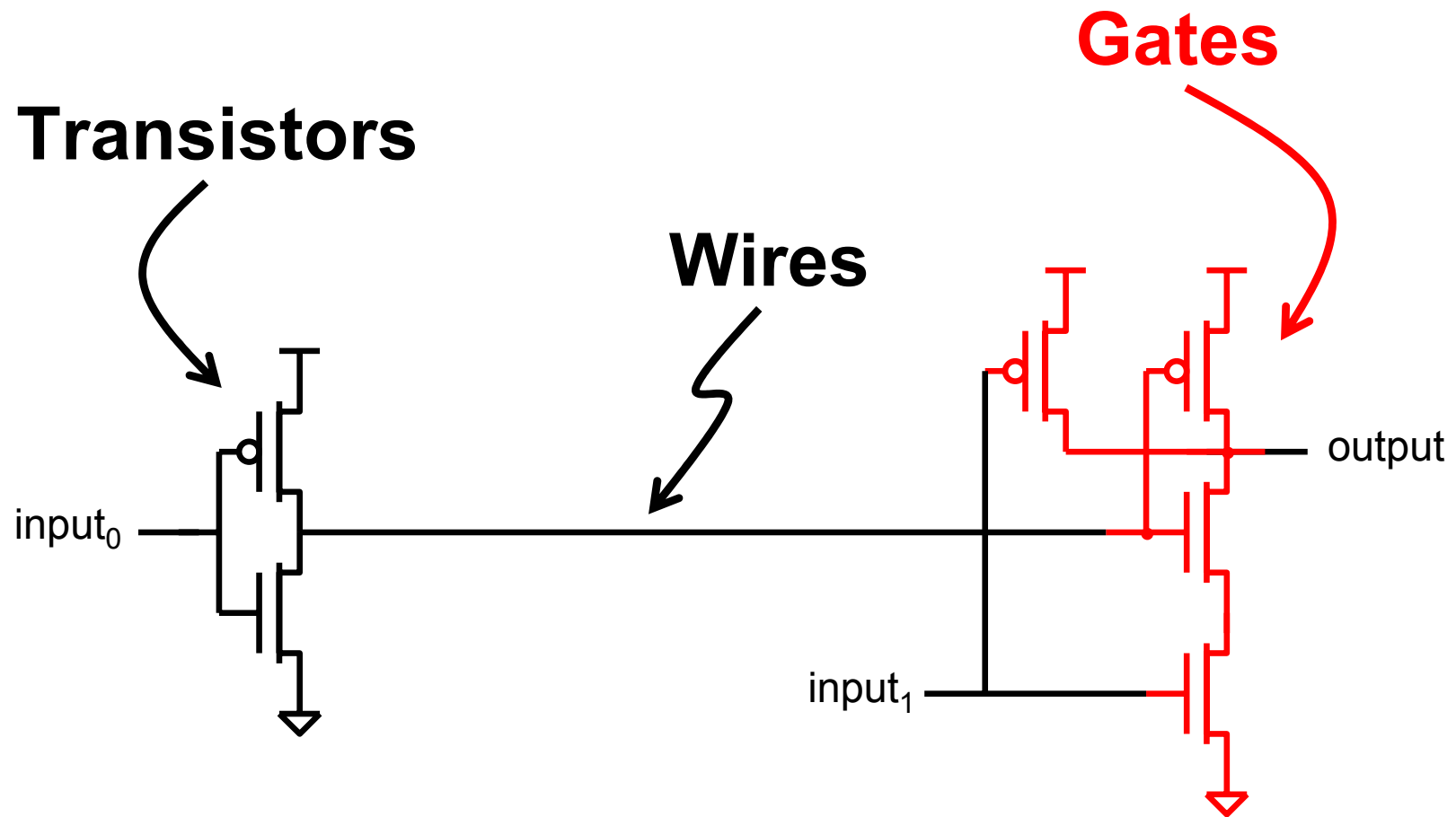


Key qualitative characteristics of MOSFET transistors

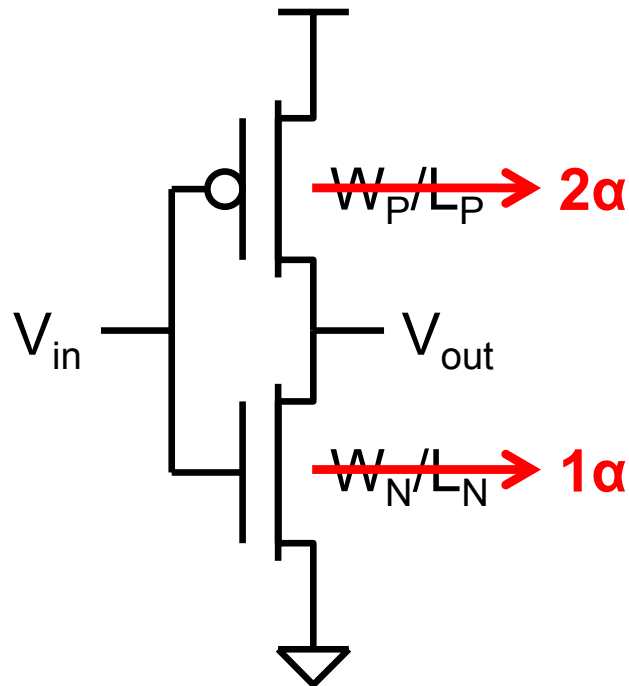


- Threshold voltage sets when transistor turns on – also impacts leakage
- I_{DS} is proportional to mobility \times (W/L)
- NMOS mobility $>$ PMOS mobility $\Rightarrow R_{effN} < R_{effP}$ (assume mobility ratio is 2)
- Increase W = Increase I = Decrease R_{eff}
- Increase L = Decrease I = Increase R_{eff}
- C_g proportional to (W \times L) and C_d proportional to W

CMOS Transistors, Gates, and Wires



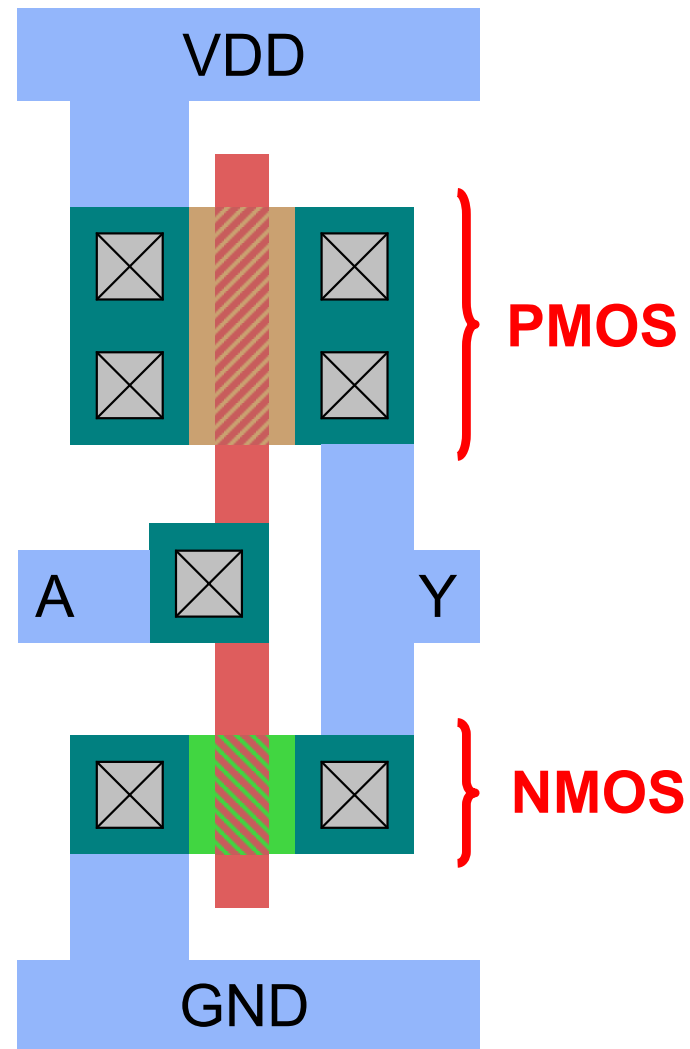
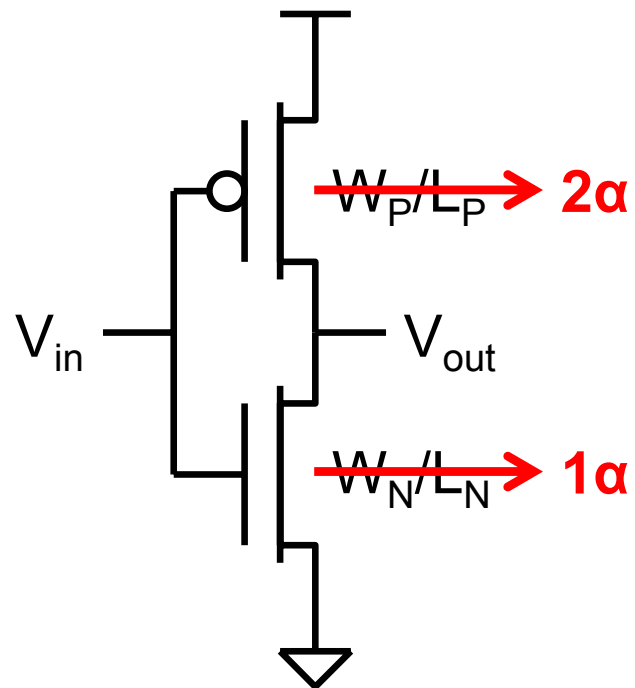
The most basic CMOS gate is an inverter



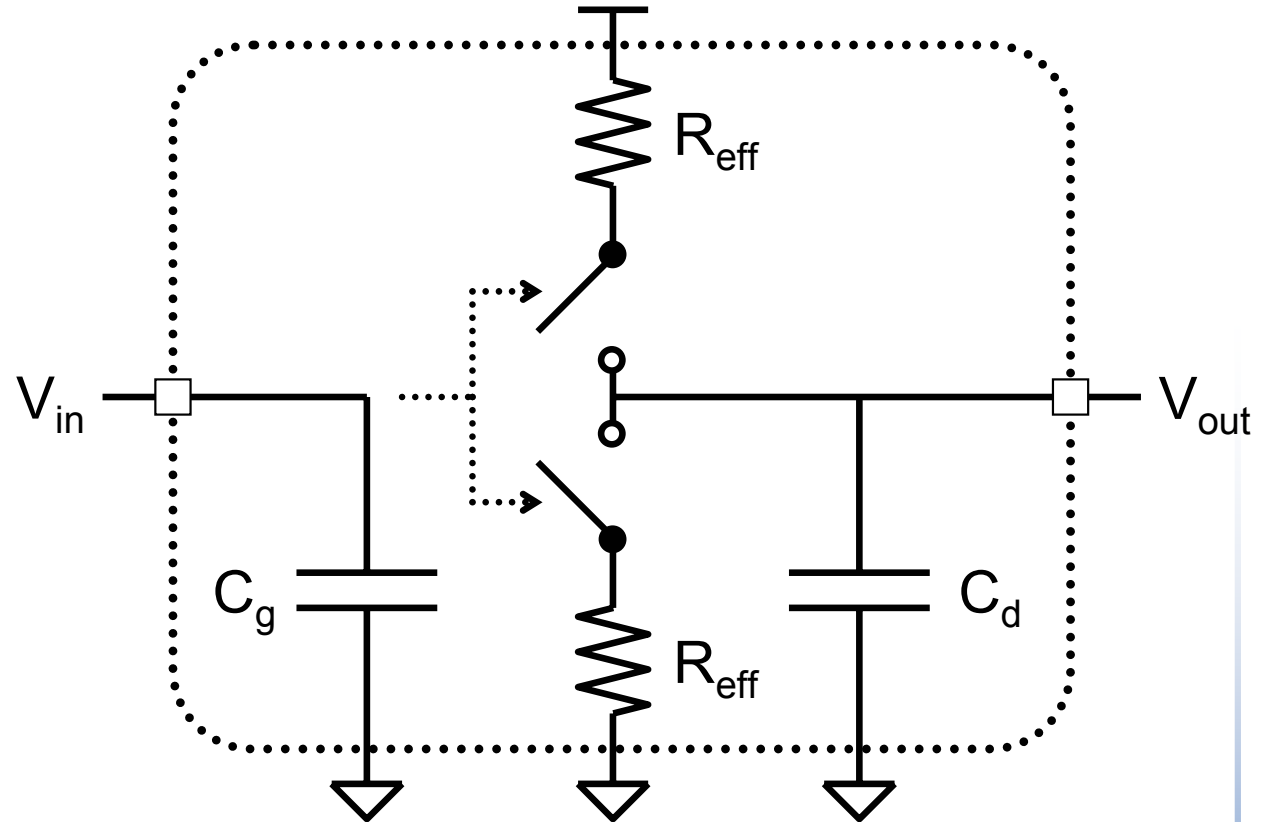
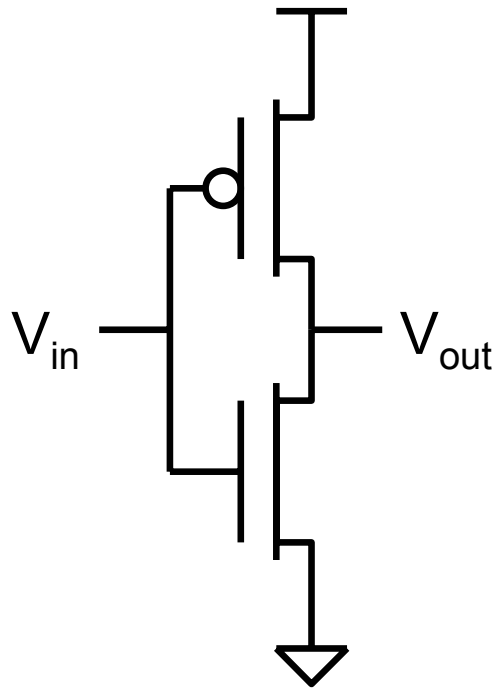
Let's make the following assumptions

1. All transistors are minimum length
2. All gates should have equal rise/fall times. Since PMOS are twice as slow as NMOS they must be twice as wide to have the same effective resistance
3. Normalize all transistor widths to minimum width NMOS

The most basic CMOS gate is an inverter

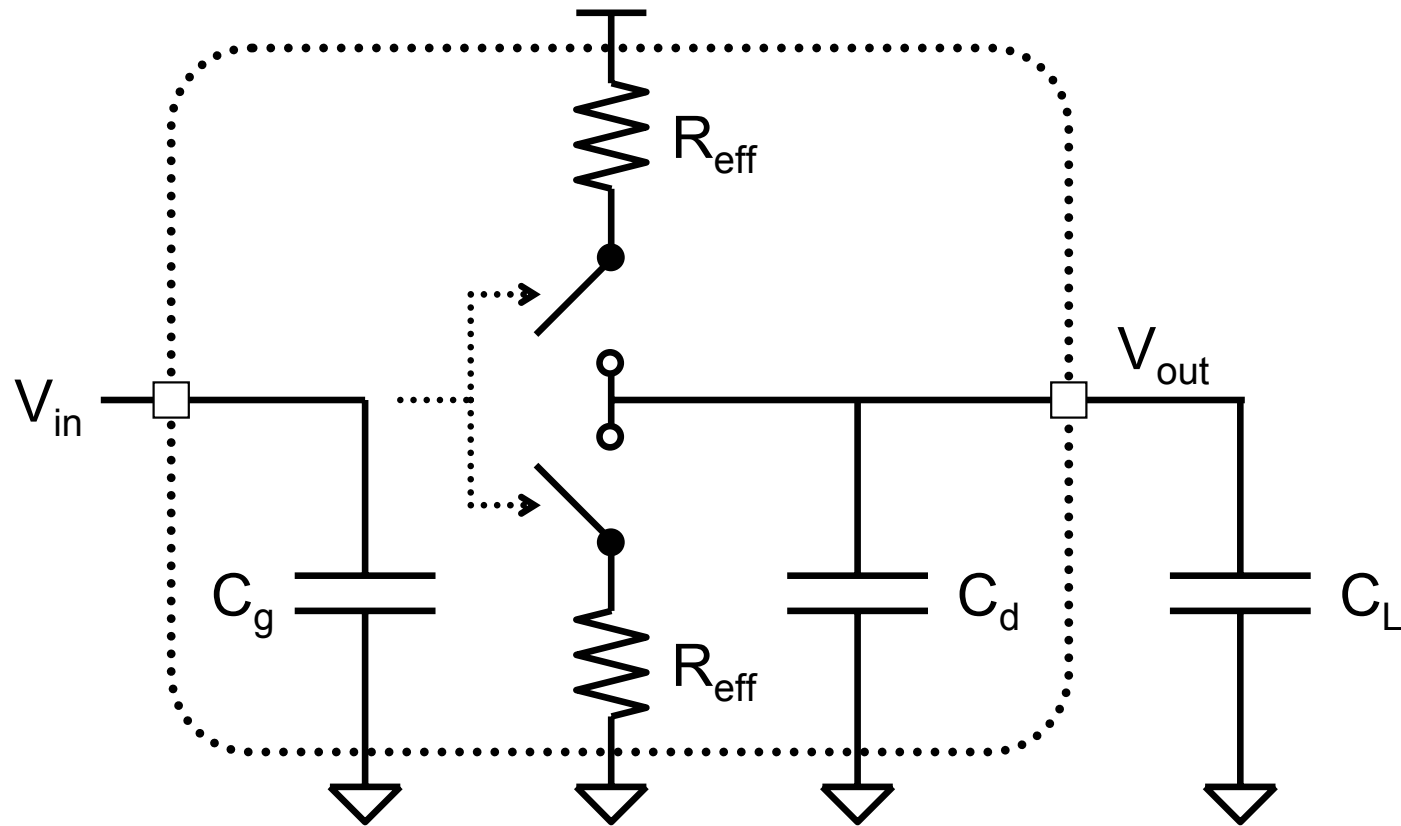


A simple RC model for the inverter can provide significant insight

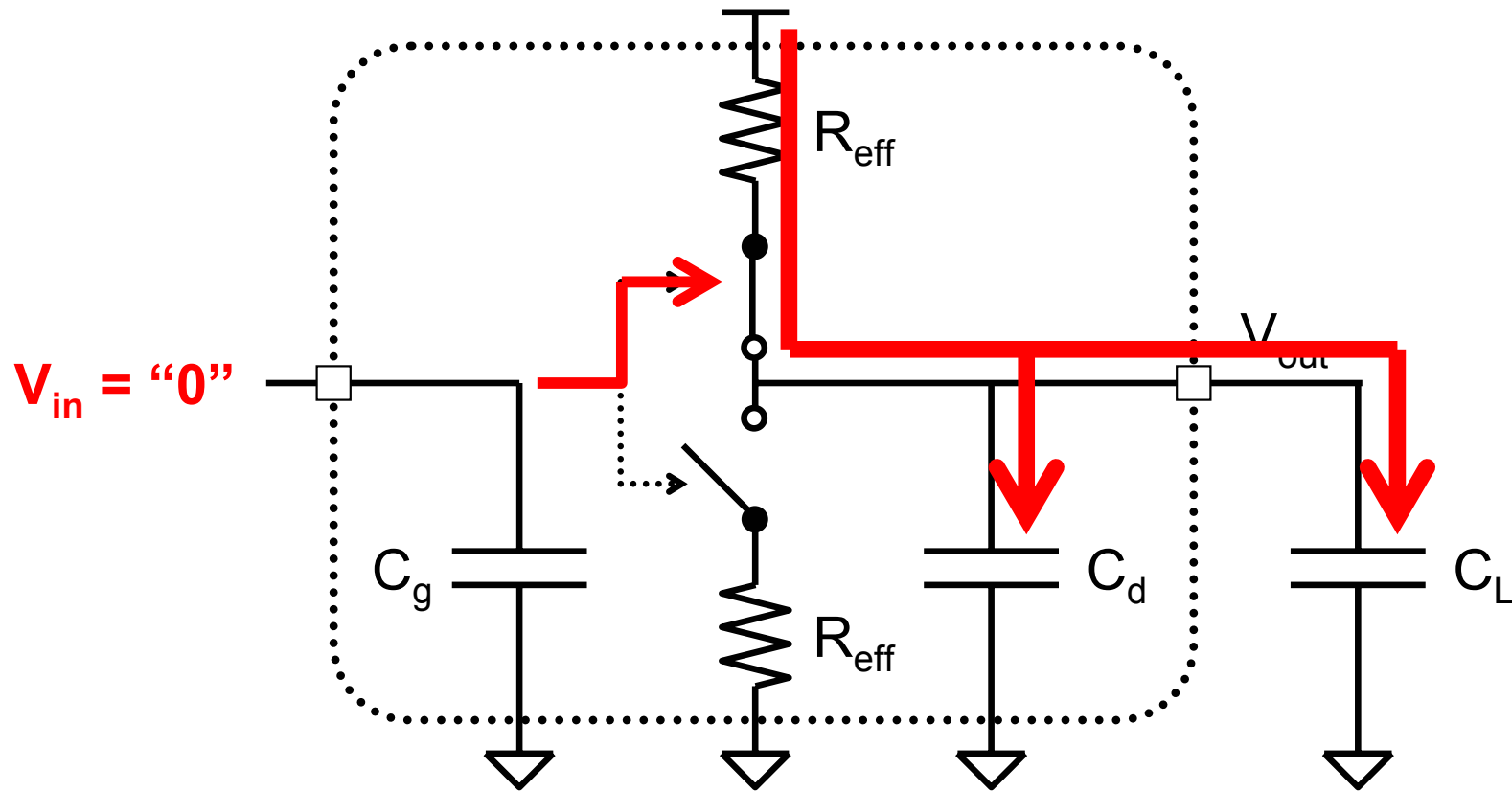


$$R_{eff} = R_{eff,N} = R_{eff,P}$$
$$C_g = C_{g,N} + C_{g,P}$$
$$C_d = C_{d,N} + C_{d,P}$$

A simple RC model for the inverter can provide significant insight

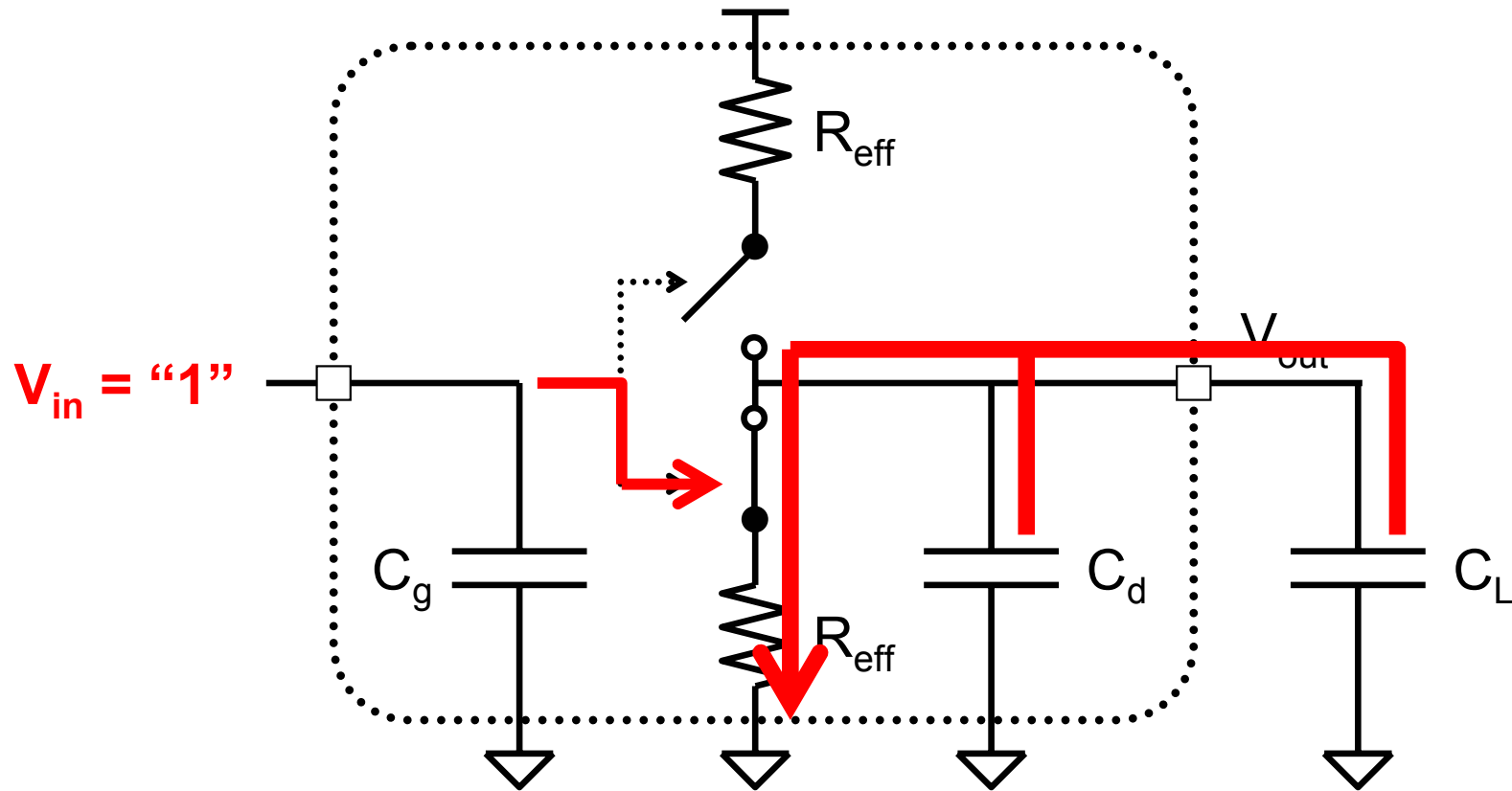


A simple RC model for the inverter can provide significant insight



$$\text{Charge RC Time Constant} = R_{eff} \times (C_d + C_L)$$

The most basic CMOS gate is an inverter

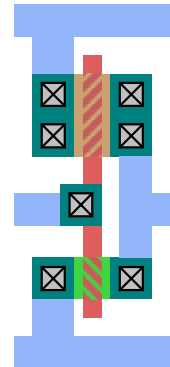
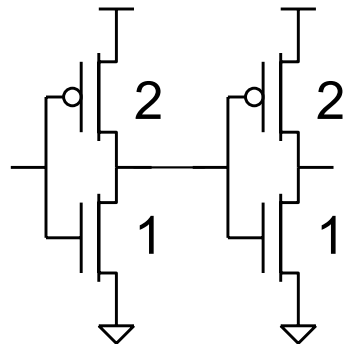


$$\text{Discharge RC Time Constant} = R_{eff} \times (C_d + C_L)$$

Larger gates are faster since they decrease R_{eff} (but they also increase C_d !)

Process gen = $0.25\mu\text{m}$
 Supply voltage = 5V
 Min width NMOS = $0.5\mu\text{m}$

| Param | Value | Units |
|---------------------------------------|-------|------------------------|
| $C_{d,N}/\mu\text{m}$ | 1.42 | fF/ μm |
| $C_{d,P}/\mu\text{m}$ | 2.40 | fF/ μm |
| $C_{g,N}/\mu\text{m}$ | 1.55 | fF/ μm |
| $C_{g,P}/\mu\text{m}$ | 1.48 | fF/ μm |
| $R_{\text{eff},N} \times \mu\text{m}$ | 4.93 | k $\Omega/\mu\text{m}$ |
| $R_{\text{eff},P} \times \mu\text{m}$ | 10.83 | k $\Omega/\mu\text{m}$ |



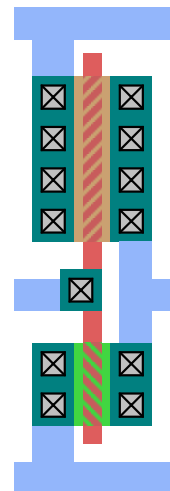
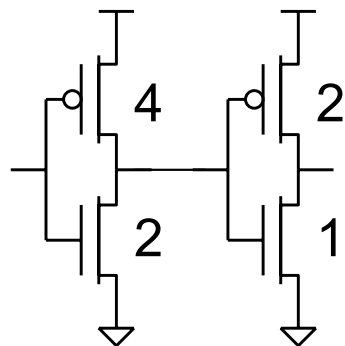
$$C_d = (0.5 \times 1.42) + (1 \times 2.40) = 3.11 \text{ fF}$$

$$C_L = (0.5 \times 1.55) + (1 \times 1.48) = 2.26 \text{ fF}$$

$$C_d + C_L = 5.37 \text{ fF}$$

$$T_{\text{PLH}} = 2.2 \times (10.83/1) \times 5.37 = 128 \text{ ps}$$

$$T_{\text{PHL}} = 2.2 \times (4.93/0.5) \times 5.37 = 116 \text{ ps}$$



$$C_d = (1 \times 1.42) + (2 \times 2.40) = 3.66 \text{ fF}$$

$$C_L = (0.5 \times 1.55) + (1 \times 1.48) = 2.26 \text{ fF}$$

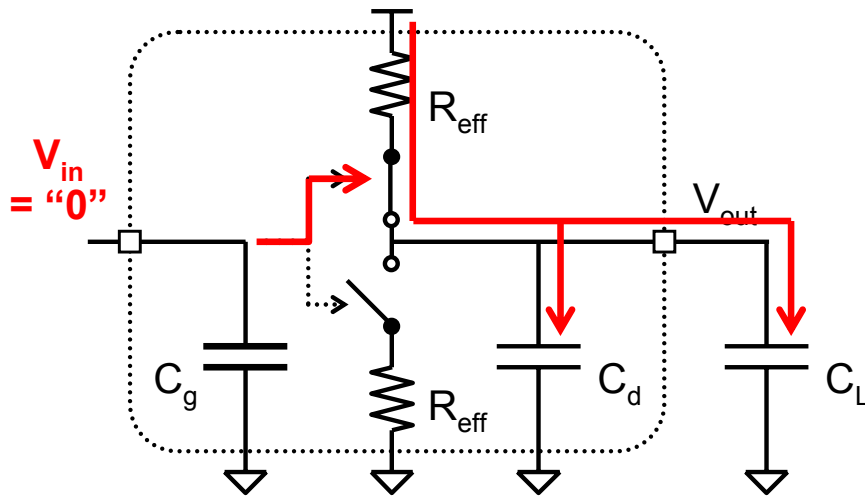
$$C_d + C_L = 5.92 \text{ fF}$$

$$T_{\text{PLH}} = 2.2 \times (10.83/2) \times 5.92 = 70.5 \text{ ps}$$

$$T_{\text{PHL}} = 2.2 \times (4.93/1) \times 5.92 = 64.2 \text{ ps}$$

Ignores the fact that previous gate now must drive a bigger gate capacitance!

Simple RC model can also yield intuition on energy consumption of inverter



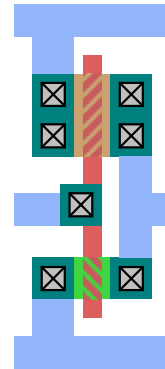
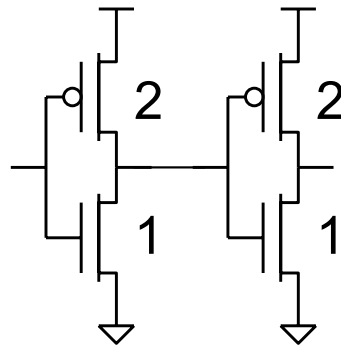
$$\begin{aligned} E_{0 \rightarrow 1} &= \int_0^T P(t) dt = V_{DD} \int_0^T I(t) dt = V_{DD} \int_0^T \frac{dQ}{dt} dt \\ &= V_{DD} \int_0^T C \frac{dV}{dt} dt = V_{DD} \int_0^{V_{DD}} (C_d + C_L) dV_{out} \\ &= (C_d + C_L) V_{DD}^2 = C V_{DD}^2 \end{aligned}$$

- During 0→1 transition, energy CV_{DD}^2 removed from power supply
- After transition, $1/2 CV_{DD}^2$ stored in capacitor, the other $1/2 CV_{DD}^2$ was dissipated as heat in pullup resistance
- The $1/2 CV_{DD}^2$ energy stored in capacitor is dissipated in the pulldown resistance on next 1→0 transition

Larger gates use slightly more energy, real concern is affect on **previous gate**

Process gen = 0.25μm
 Supply voltage = 5V
 Min width NMOS = 0.5μm

| Param | Value | Units |
|---------------------------------------|-------|-------|
| $C_{d,N}/\mu\text{m}$ | 1.42 | fF/μm |
| $C_{d,P}/\mu\text{m}$ | 2.40 | fF/μm |
| $C_{g,N}/\mu\text{m}$ | 1.55 | fF/μm |
| $C_{g,P}/\mu\text{m}$ | 1.48 | fF/μm |
| $R_{\text{eff},N} \times \mu\text{m}$ | 4.93 | kΩ/μm |
| $R_{\text{eff},P} \times \mu\text{m}$ | 10.83 | kΩ/μm |

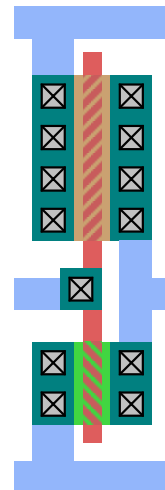
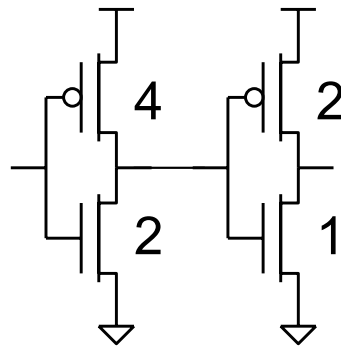


$$C_d = (0.5 \times 1.42) + (1 \times 2.40) = 3.11 \text{ fF}$$

$$C_L = (0.5 \times 1.55) + (1 \times 1.48) = 2.26 \text{ fF}$$

$$C_d + C_L = 5.37 \text{ fF}$$

$$E_{0>1} = 5.37 \times 5^2 = 134 \text{ fJ}$$



$$C_d = (1 \times 1.42) + (2 \times 2.40) = 3.66 \text{ fF}$$

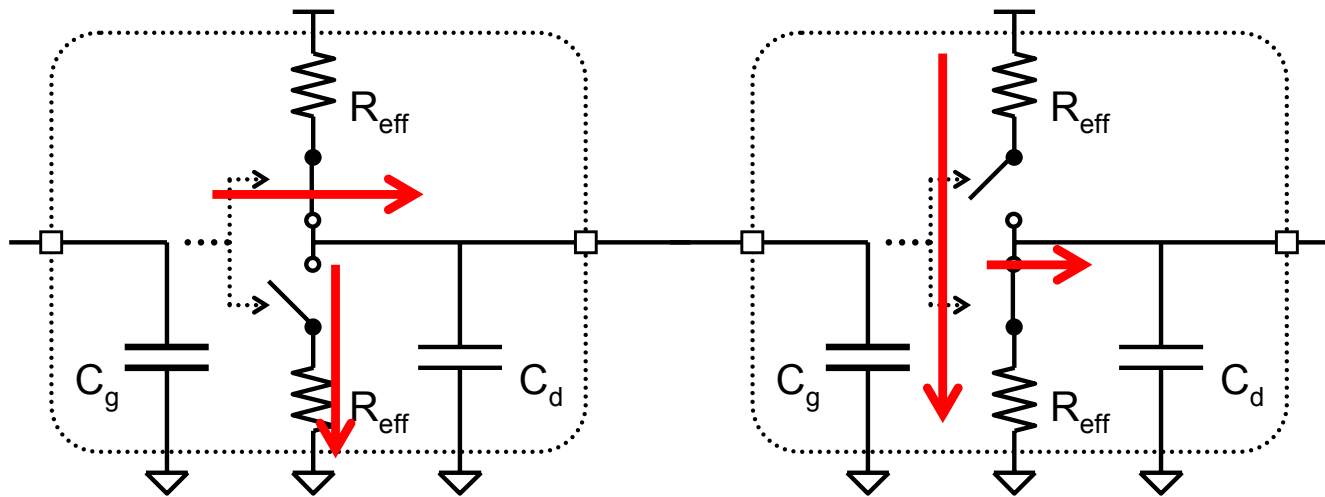
$$C_L = (0.5 \times 1.55) + (1 \times 1.48) = 2.26 \text{ fF}$$

$$C_d + C_L = 5.92 \text{ fF}$$

$$E_{0>1} = 5.92 \times 5^2 = 148 \text{ fJ}$$

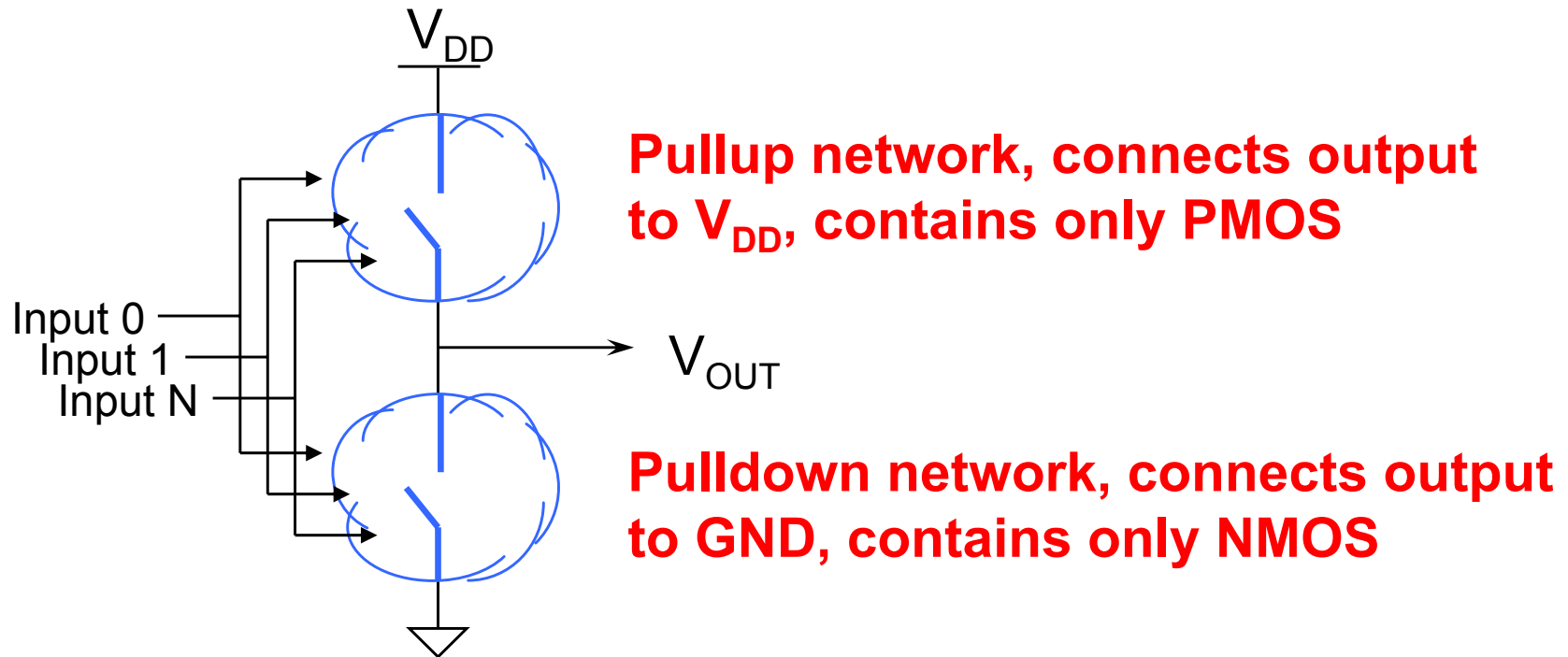
Ignores the fact that previous gate now must drive a bigger gate capacitance!

Many other types of power consumption in addition to dynamic power



| | |
|------------------------------|--|
| Short Circuit Current | Fast edges keep to <10% of cap charging current |
| Subthreshold Leakage | Approaching 10-40% of active power |
| Diode Leakage | Usually negligible |
| Gate Leakage | Was negligible, increasing due to thin gate oxides |

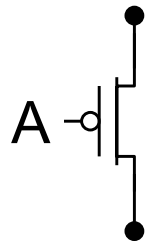
More complicated gates use more transistors in pullup/pulldown networks



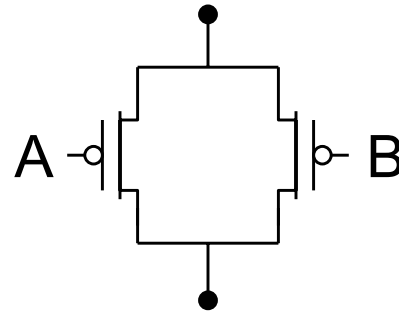
For every set of input logic values, either pullup or pulldown network makes connection to V_{DD} or GND

- If both connected, power rails would be shorted together
- If neither connected, output would float (tristate logic)

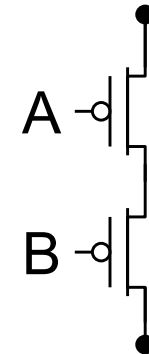
Series and parallel MOSFET networks provide natural duals of each other



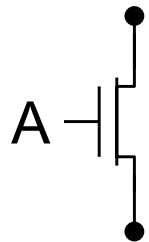
Conducts if $A=0$



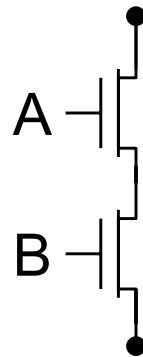
Conducts if $A=0$ **OR** $B=0$



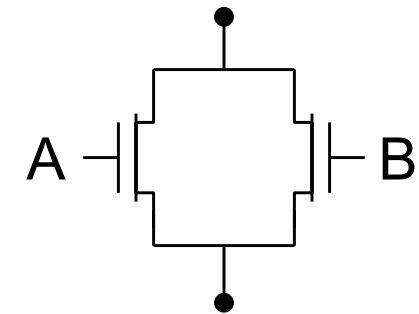
Conducts if $A=0$ **AND** $B=0$



Conducts if $A=1$



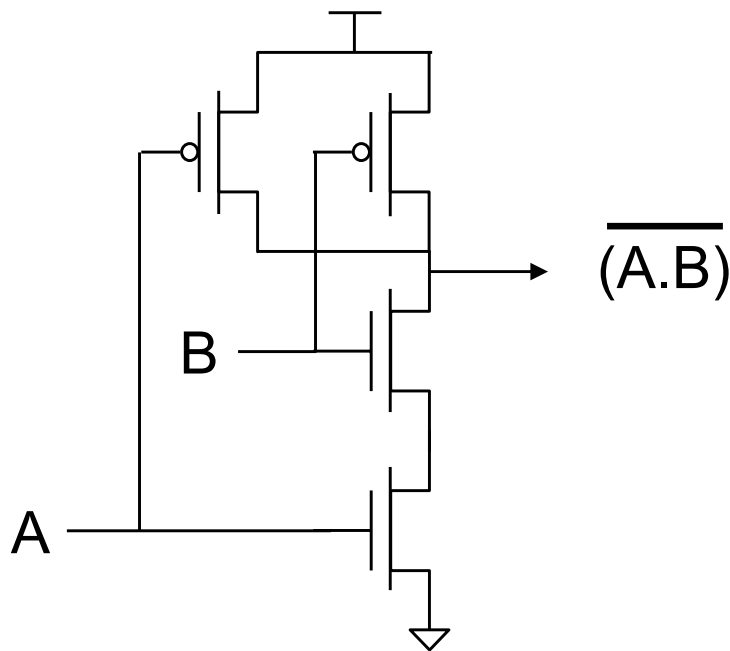
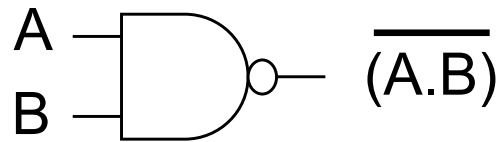
Conducts if $A=1$ **AND** $B=1$



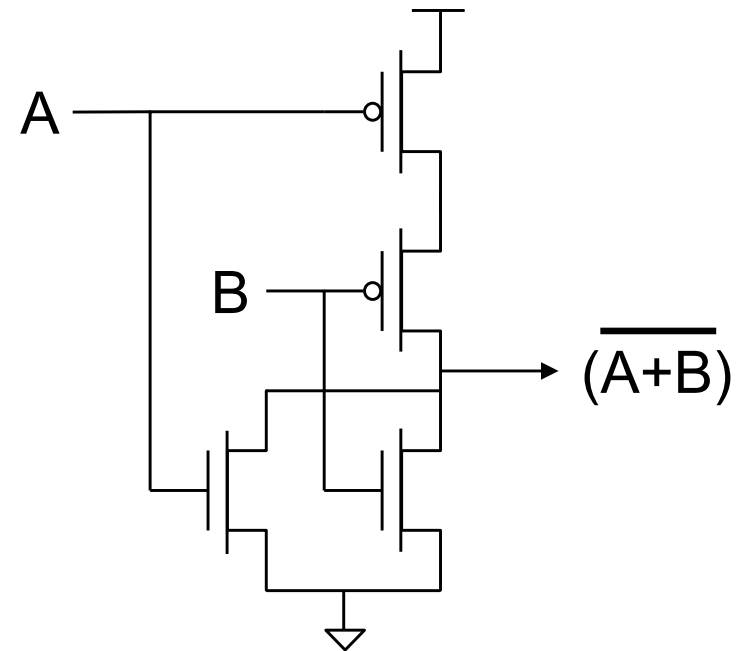
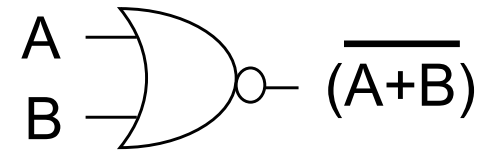
Conducts if $A=1$ **OR** $B=1$

NAND and NOR gates illustrate the dual nature of the pullup/pulldown networks

NAND Gate



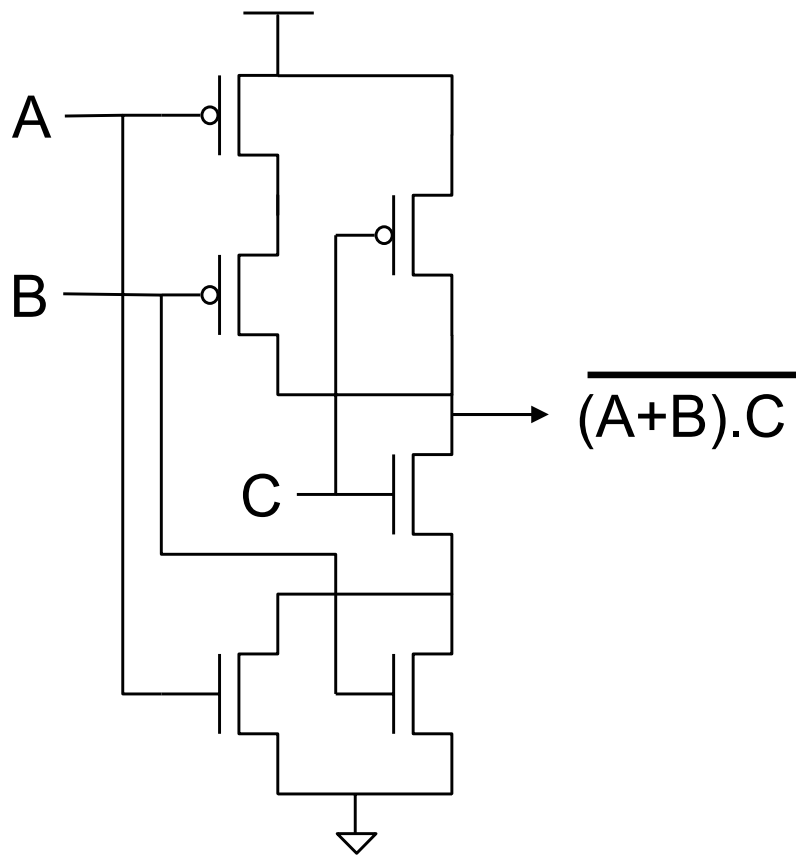
NOR Gate



Designers can use a methodical approach to build more complex gates

- Goal is to create an logic function $f(x_1, x_2, \dots)$
 - We can only implement inverting logic with one CMOS stage
- Implement pulldown network
 - Write PD = $\bar{f}(x_1, x_2, \dots)$
 - Use parallel NMOS for OR of inputs
 - Use series NMOS for AND of inputs
- Implement pullup network
 - Write pullup network PU = $f(x_1, x_2, \dots) = g(\bar{x}_1, \bar{x}_2, \dots)$
 - Use parallel PMOS for OR of complemented inputs)
 - Use series PMOS for AND of complemented inputs)

Designers can use a methodical approach to build more complex gates

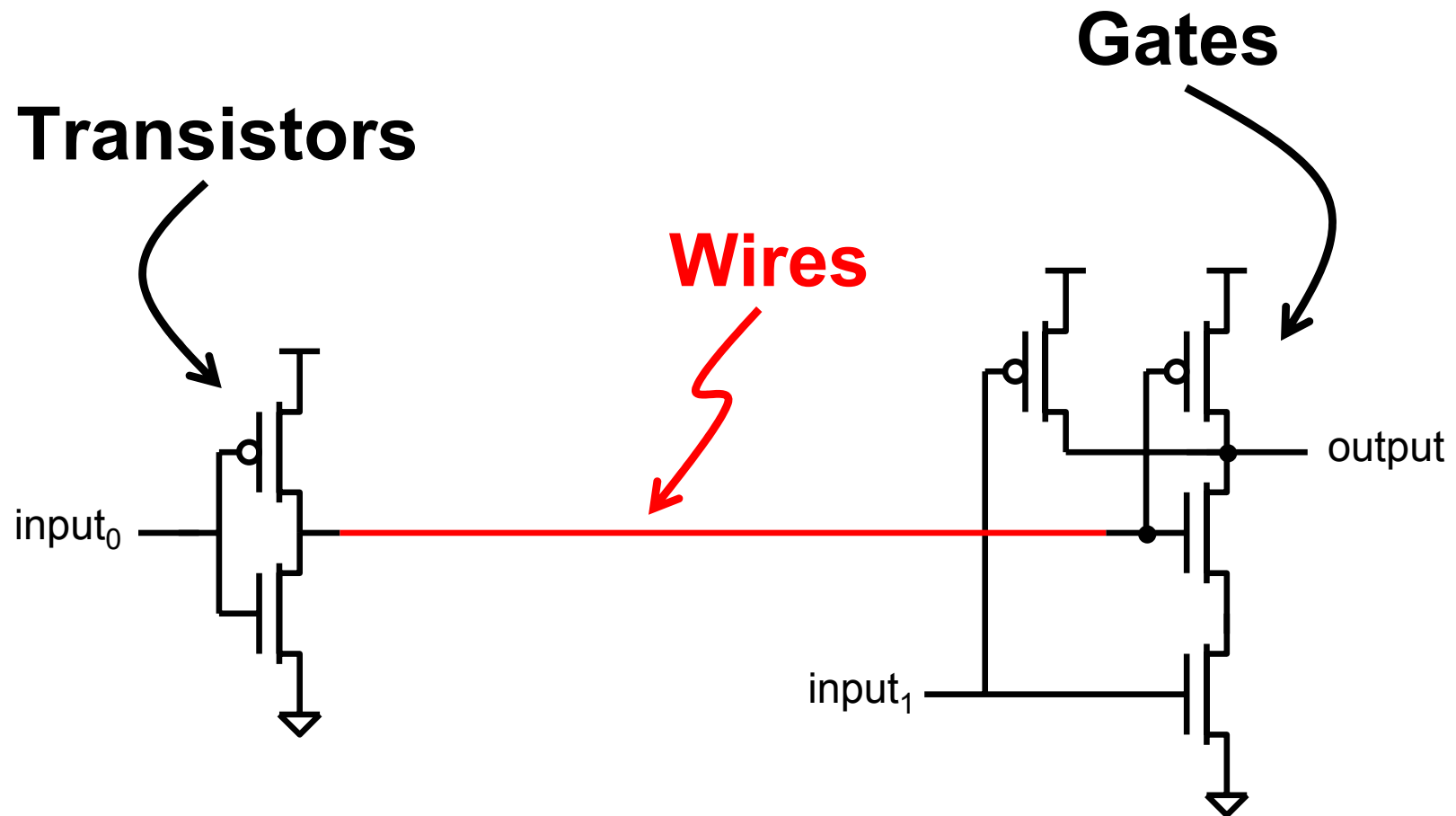


$$f = \overline{(A + B) \cdot C}$$

$$PD = (A + B) \cdot C$$

$$\begin{aligned} PU &= \overline{(A + B) \cdot C} \\ &= \overline{(A + B)} + \overline{C} \\ &= (\overline{A} \cdot \overline{B}) + \overline{C} \end{aligned}$$

CMOS Transistors, Gates, and Wires

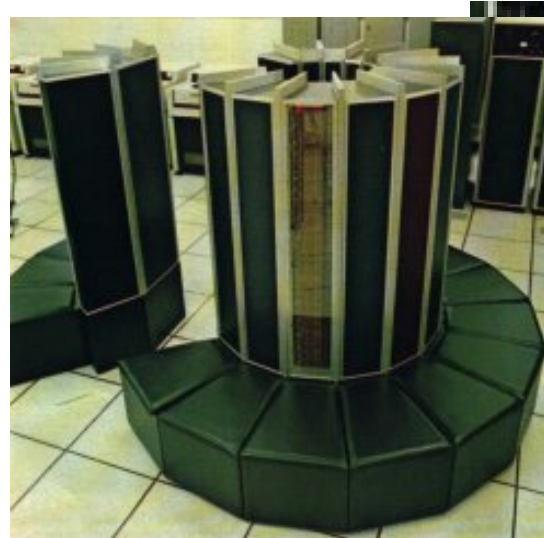


Wires are an old problem



**Cray-3
1993**

**Cray-1
1976**

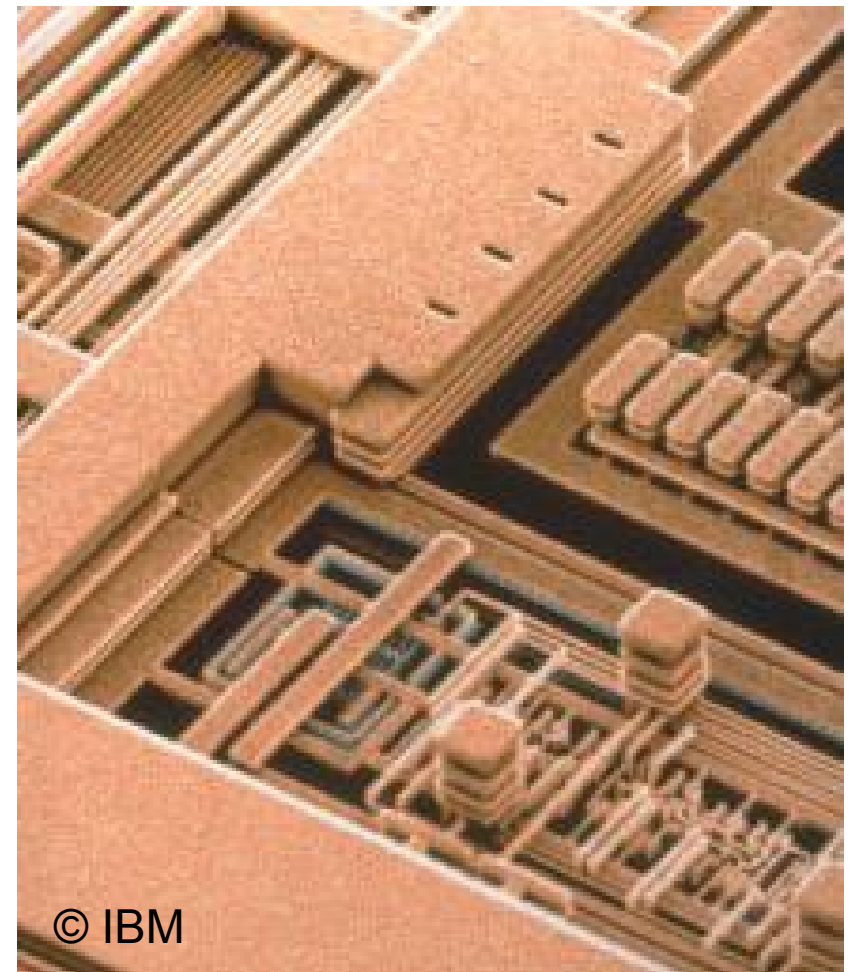
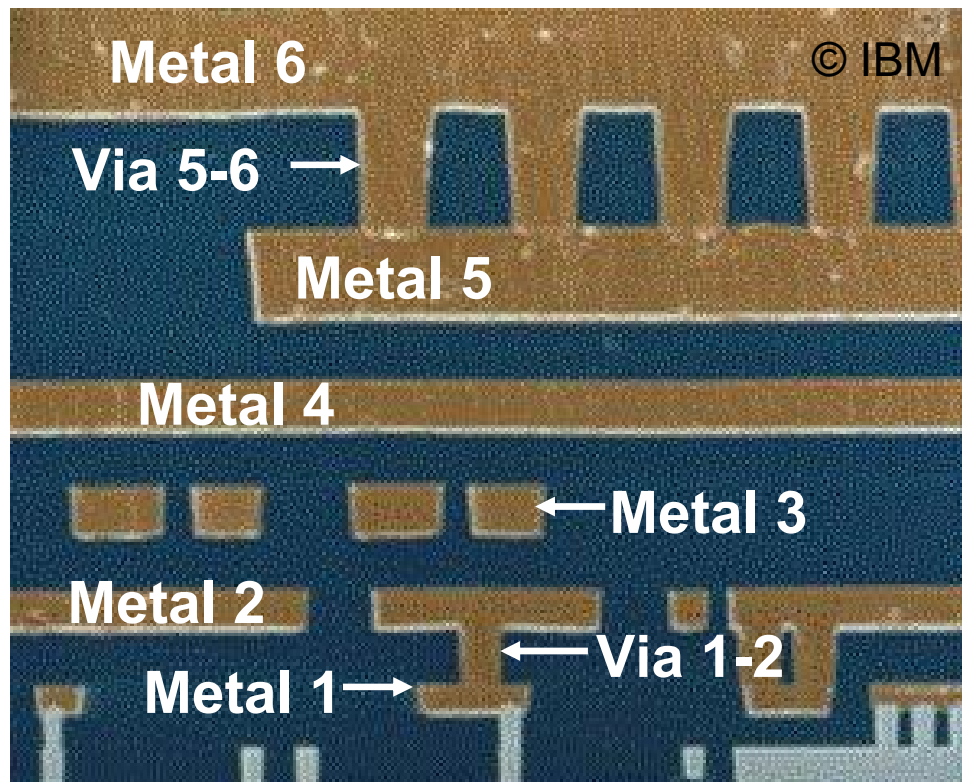


**Cray-3
wiring**



Cray-1 Wiring

Modern interconnect stacks have six to nine or more metal layers

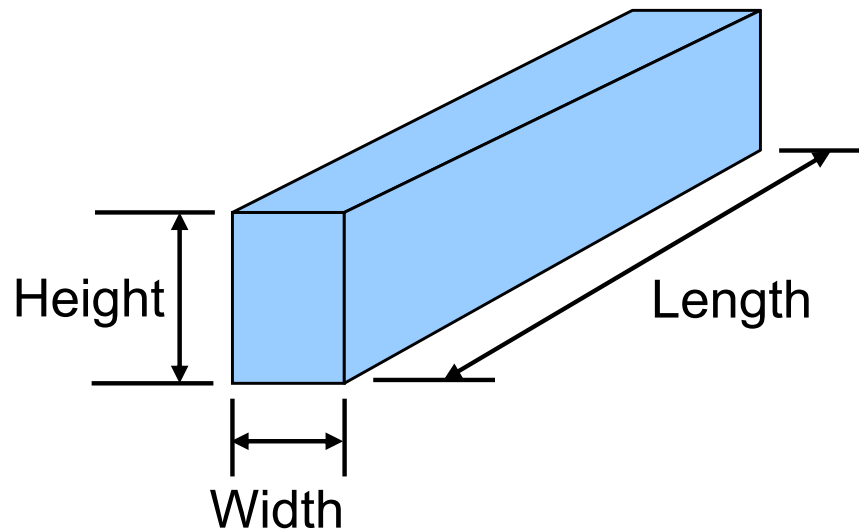


IBM CMOS7 process

6 layers of copper wiring

1 layer of tungsten local interconnect

Wire resistance is a function of height, width, and length

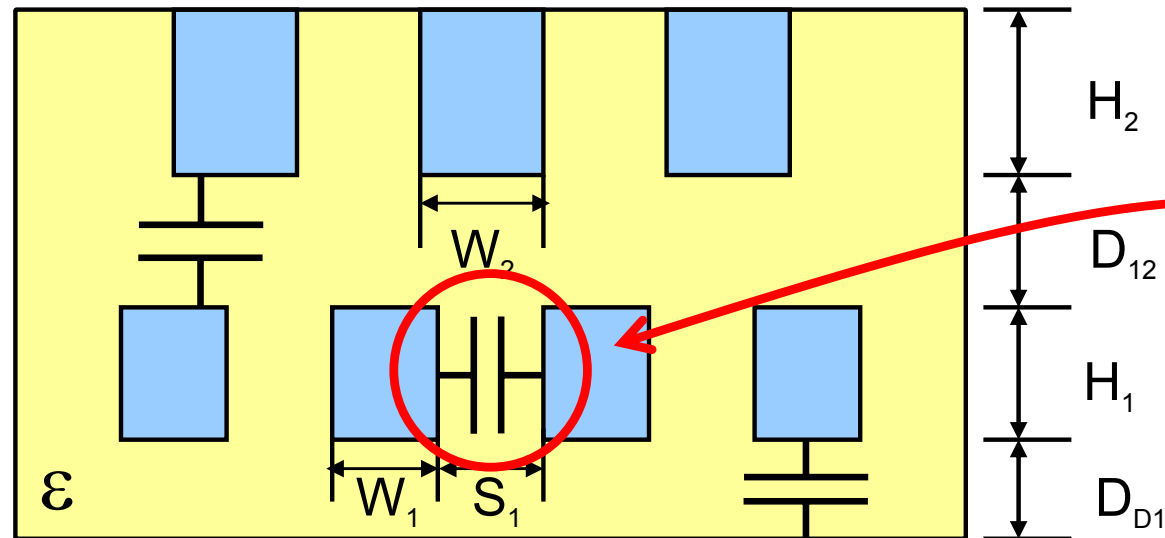


$$\text{resistance} = \frac{(\text{length} \times \text{resistivity})}{(\text{height} \times \text{width})}$$

| | |
|---------------|--------------------------------------|
| bulk aluminum | $2.8 \times 10^{-8} \Omega\text{-m}$ |
| bulk copper | $1.7 \times 10^{-8} \Omega\text{-m}$ |
| bulk silver | $1.6 \times 10^{-8} \Omega\text{-m}$ |

- Height (Thickness) fixed in given manufacturing process
- Resistances quoted as Ω/square
- TSMC 0.18mm 6 Aluminum metal layers
 - M1-5 $0.08 \Omega/\text{square}$ ($0.5 \text{ mm} \times 1 \text{ mm}$ wire = 160Ω)
 - M6 $0.03 \Omega/\text{square}$ ($0.5 \text{ mm} \times 1 \text{ mm}$ wire = 60Ω)

Wire capacitance is relative to the substrate and to neighboring wires

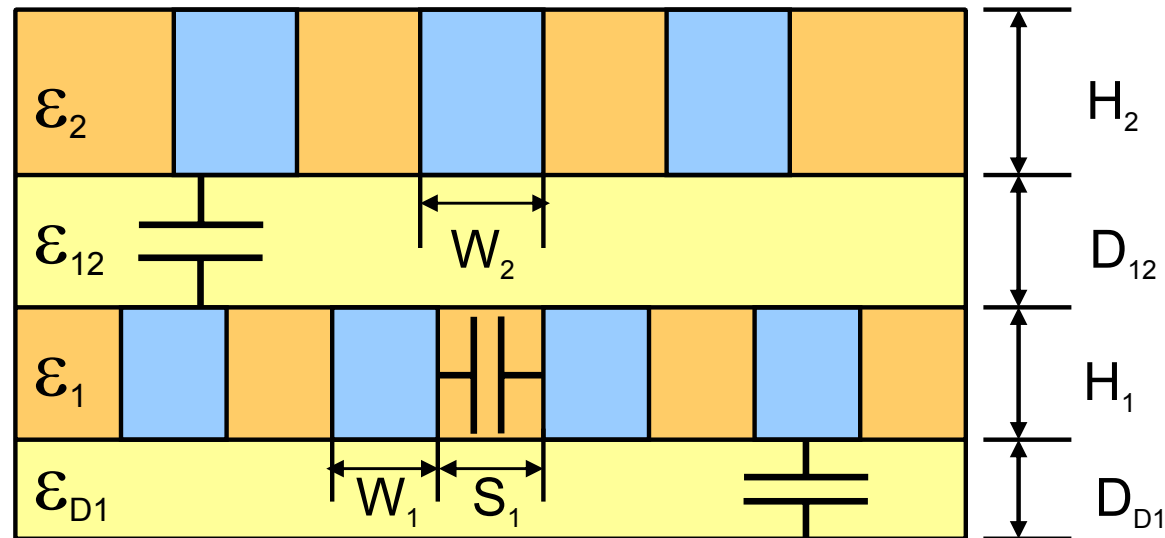


- Capacitance depends on geometry of surrounding wires and relative permittivity (ϵ_r) of insulating dielectric

- silicon dioxide (SiO_2) $\epsilon_r = 3.9$
- silicon flouride (SiF_4) $\epsilon_r = 3.1$
- SiLK™ polymer $\epsilon_r = 2.6$

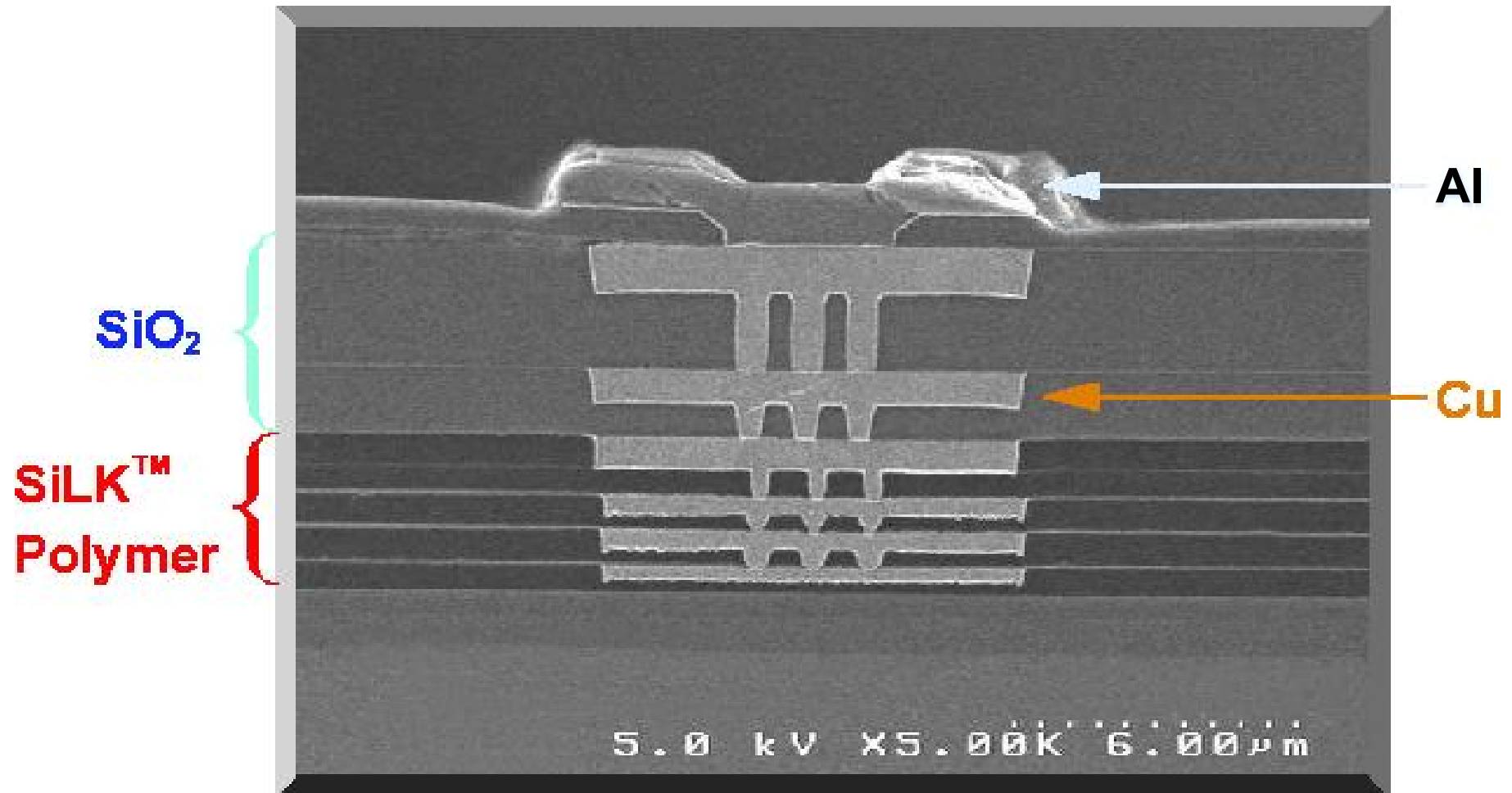
Capacitive coupling to neighbors is becoming a serious problem!

Wire capacitance is relative to the substrate and to neighboring wires



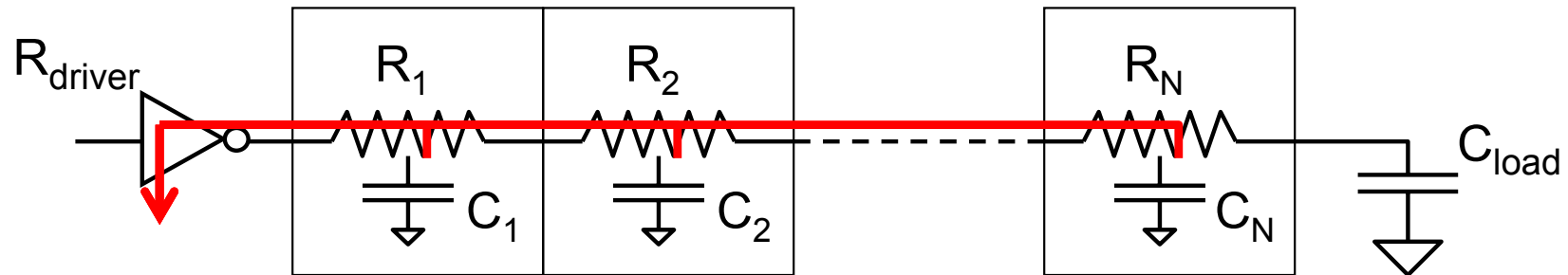
- Capacitance depends on geometry of surrounding wires and relative permittivity (ϵ_r) of insulating dielectric
 - silicon dioxide (SiO_2) $\epsilon_r = 3.9$
 - silicon flouride (SiF_4) $\epsilon_r = 3.1$
 - SiLK™ polymer $\epsilon_r = 2.6$
- Can have different materials between wires and between layers, and also different materials on higher layers

This IBM experimental 130nm process includes two metals and two dielectrics



E. Barth, IBM Microelectronics

Distributed RC wire model gives accurate results but is computationally expensive



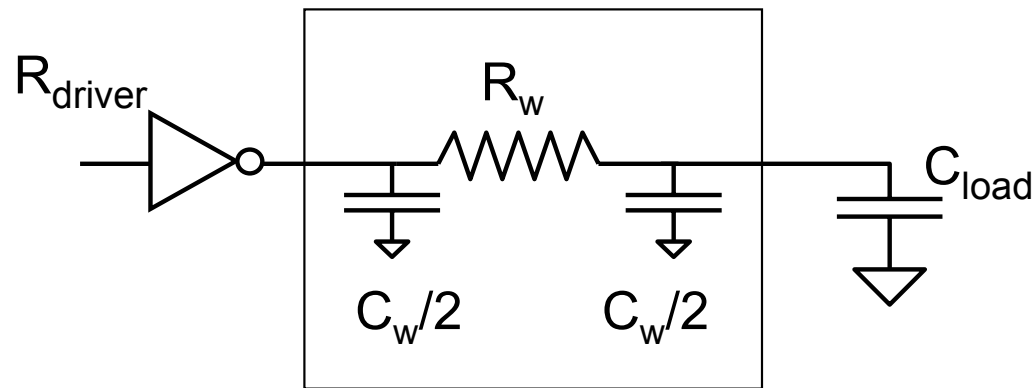
Use Penfield-Rubenstein equation to find delay

$$\text{Delay} \propto \sum_i^N \left(\sum_{j=1}^{j=i} R_j \right) C_i$$

How does the delay scale with longer wires?

- Wire delay increases quadratically
- Edge rate also degrades quadratically

Lumped Π model can provide a quick reasonable approximation



$$\text{Delay} \propto R_{\text{driver}} \times \frac{C_w}{2} + (R_{\text{driver}} + R_w) \times \left(\frac{C_w}{2} + C_{\text{load}} \right)$$

R_w is lumped resistance of the wire

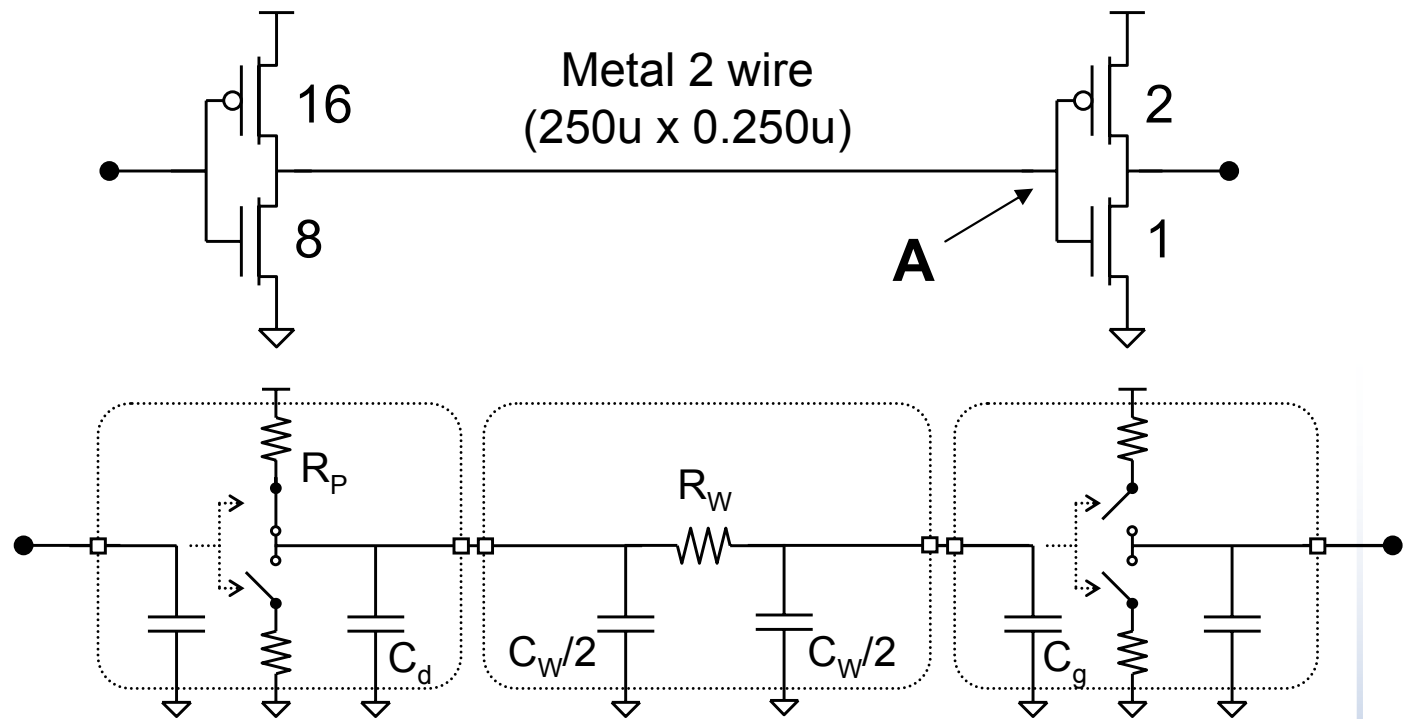
C_w is lumped capacitance

Partition half of C_w at each end

Estimate the rise time of node A using an RC delay model

Process gen = 0.25 μ m
 Supply voltage = 5V
 Min width NMOS = 0.5 μ m

| Param | Value | Units |
|---------------------------------------|-------|----------------------------|
| $C_{d,N} / \mu\text{m}$ | 1.42 | fF/ μm |
| $C_{d,P} / \mu\text{m}$ | 2.40 | fF/ μm |
| $C_{g,N} / \mu\text{m}$ | 1.55 | fF/ μm |
| $C_{g,P} / \mu\text{m}$ | 1.48 | fF/ μm |
| $C_{A,M2} / \mu\text{m}^2$ | 0.016 | fF/ μm^2 |
| $C_{L,M2} / \mu\text{m}$ | 0.084 | fF/ μm |
| $R_{\text{eff},N} \times \mu\text{m}$ | 4.93 | k Ω / μm |
| $R_{\text{eff},P} \times \mu\text{m}$ | 10.83 | k Ω / μm |
| R_{M2} / sq | 0.07 | Ω /sq |



$$C_g = (0.5 \times 1.55) + (1 \times 1.48) = 2.26 \text{ fF}$$

$$C_d = (4 \times 1.42) + (8 \times 2.40) = 24.88 \text{ fF}$$

$$R_p = 10.83/8 = 1.35 \text{ k}\Omega$$

$$R_w = (250 / 0.25) \times 0.07 = 70 \Omega$$

$$C_w = ((250 \times 0.25) \times 0.0016) + (250 \times 0.084) = 21.14 \text{ fF}$$

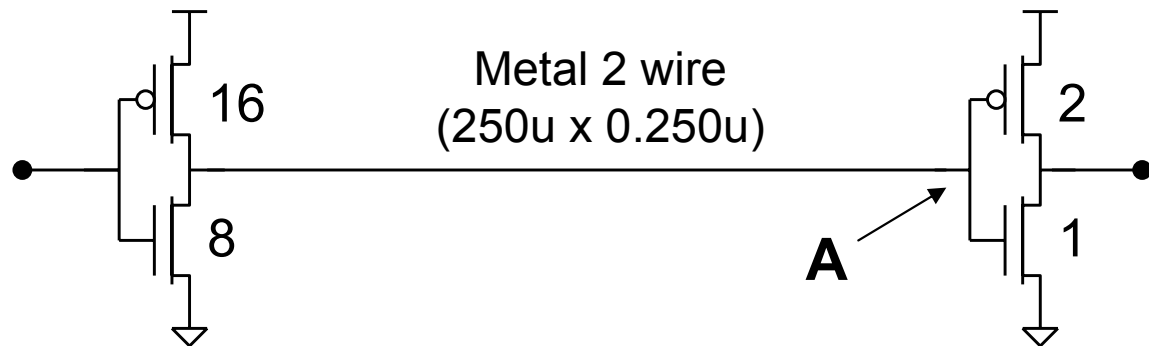
$$T_{\text{PLH}} = 2.2 \times (1350 \times (21.14/2 + 24.88)$$

$$+ (1350 + 70) \times (21.14/2 + 2.26)) = 66\text{ps}$$

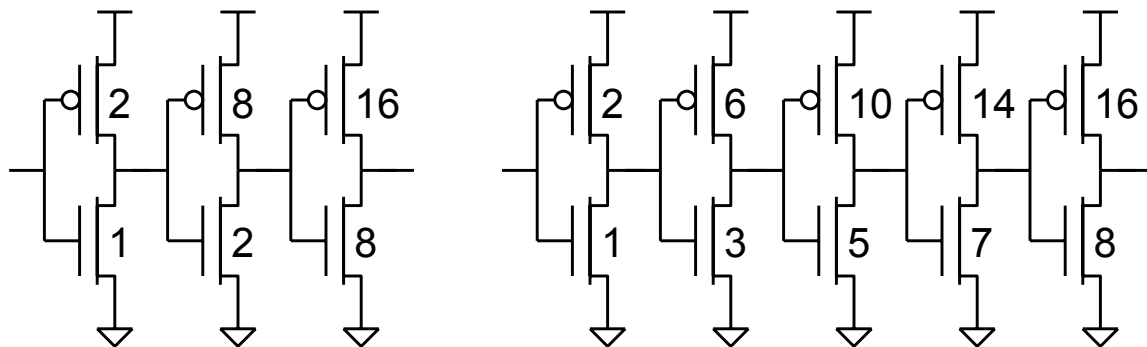
Estimate the rise time of node A using an RC delay model

Process gen = 0.25 μ m
 Supply voltage = 5V
 Min width NMOS = 0.5 μ m

| Param | Value | Units |
|---------------------------------------|-------|----------------------------|
| $C_{d,N} / \mu\text{m}$ | 1.42 | fF/ μm |
| $C_{d,P} / \mu\text{m}$ | 2.40 | fF/ μm |
| $C_{g,N} / \mu\text{m}$ | 1.55 | fF/ μm |
| $C_{g,P} / \mu\text{m}$ | 1.48 | fF/ μm |
| $C_{A,M2} / \mu\text{m}^2$ | 0.016 | fF/ μm^2 |
| $C_{L,M2} / \mu\text{m}$ | 0.084 | fF/ μm |
| $R_{\text{eff},N} \times \mu\text{m}$ | 4.93 | k Ω / μm |
| $R_{\text{eff},P} \times \mu\text{m}$ | 10.83 | k Ω / μm |
| R_{M2} / sq | 0.07 | Ω /sq |



**How should we buffer up this signal?
 Should we have a few big stages or
 many small stages?**



In deep submicron technologies many predicted an interconnect doomsday

SPEED / PERFORMANCE ISSUE *The Technical Problem*

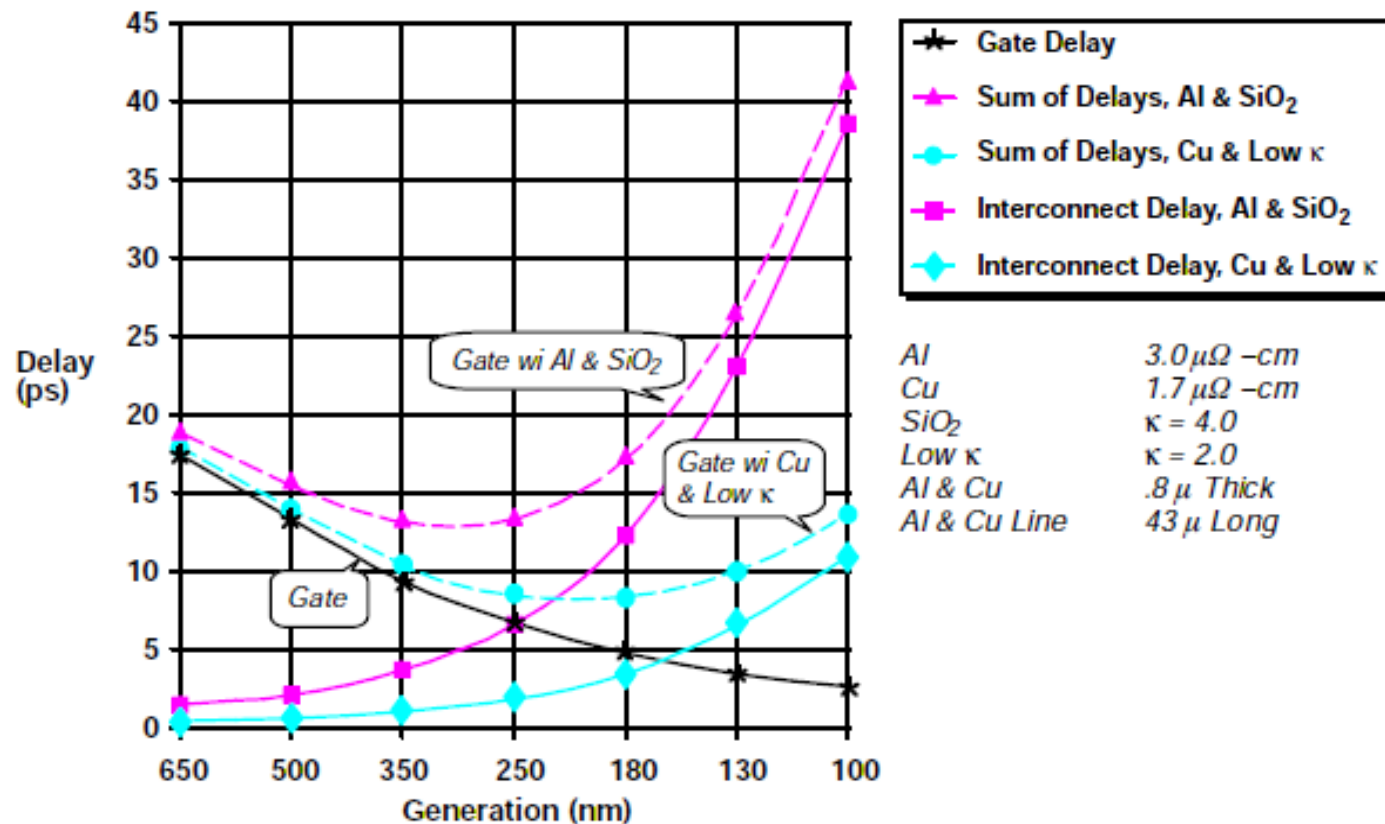
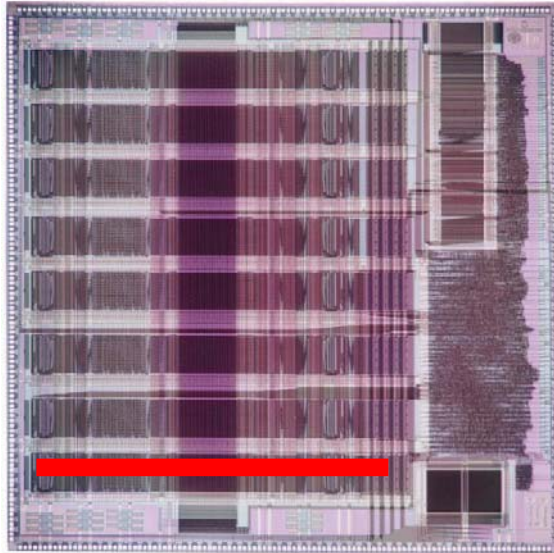


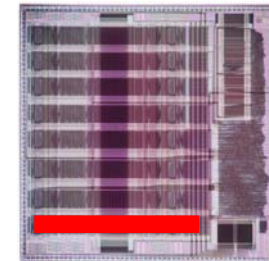
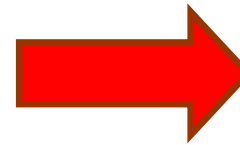
Figure 3 Calculated Gate and Interconnect Delay versus Technology Generation

Calculated gate and interconnect delay versus technology generation illustrating the dominance of interconnect delay over gate delay for aluminum metallization and silicon dioxide dielectrics as feature sizes approach 100 nm. Also shown is the decrease in interconnect delay and improved overall performance expected for copper and low κ dielectric constant insulators.¹

Is there really an interconnect doomsday looming?

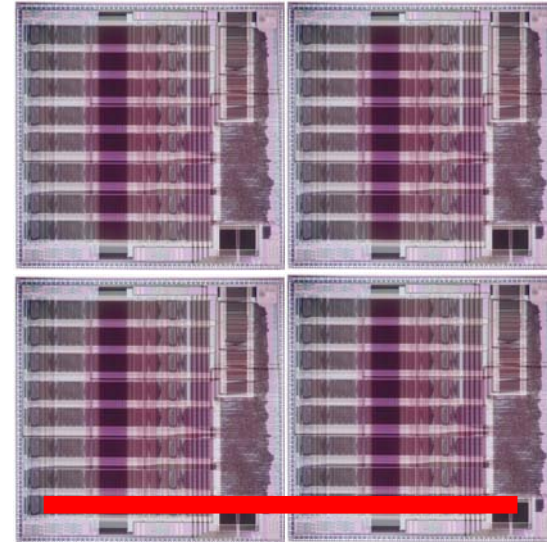
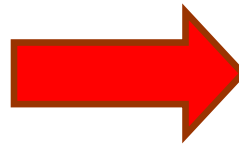
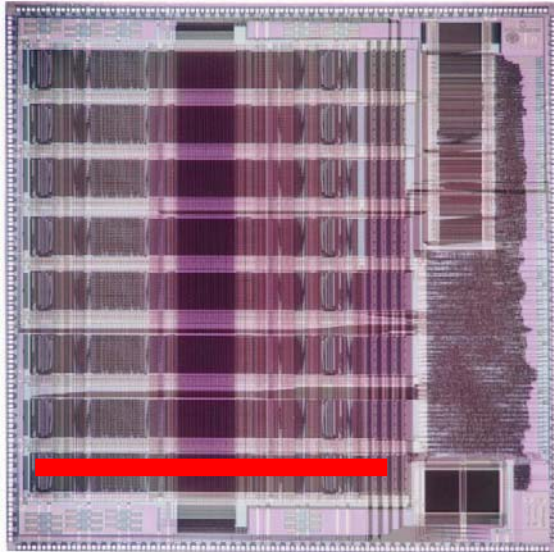


Local wire delay tracks improvement in gate delay



| | Scaling Impact | Affect on Resistance | Affect on Capacitance |
|---------------|-----------------------|-----------------------------|------------------------------|
| Length | Decreases | Decreases | Decrease |
| Width | Decreases | Increases | Decrease |
| Height | ~ Constant | -- | -- |

Is there really an interconnect doomsday looming?



| | Scaling Impact | Affect on Resistance | Affect on Capacitance |
|--------|----------------|----------------------|-----------------------|
| Length | ~ Constant | -- | -- |
| Width | Decreases | Increases | Decrease |
| Height | ~ Constant | -- | -- |

Global wire delay increases relative to wire delay!

No doomsday, just one more physical design issue to carefully manage

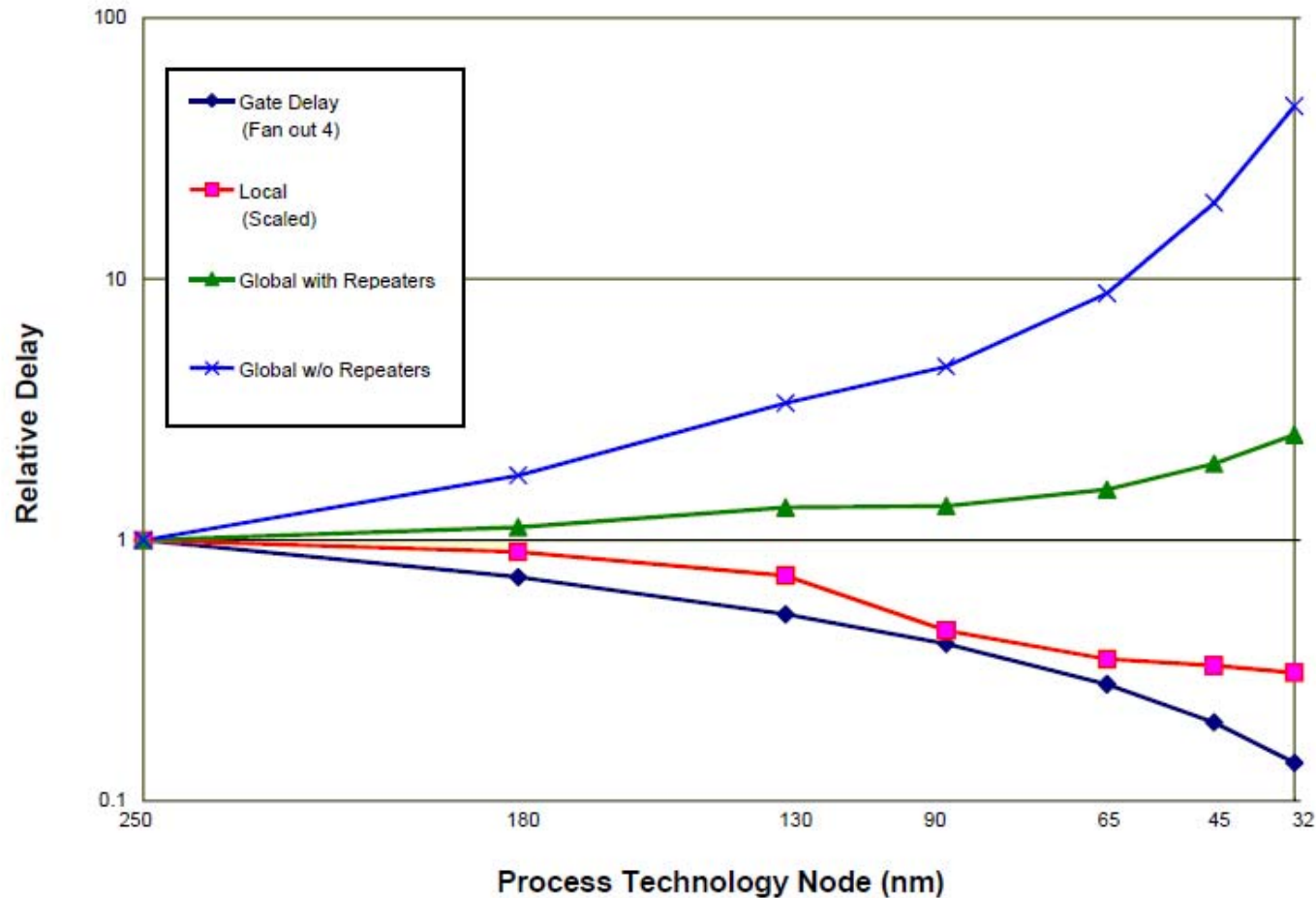


Figure 69 Delay for Metal 1 and Global Wiring versus Feature Size

Take away points

- Simple **RC models** of CMOS transistors, gates, and wires can provide reasonable insight into the power and delay of circuits
- A **methodological approach** enables creating static CMOS gates relatively straightforward
- Although **global wire delay** is getting worse relative to gate delay, **local wire delay** is scaling with gate delay which forces designers to better manage global wires early in the design process

Next Lecture: Srinivasa Devadas will be discussing algorithms and issues in synthesis and place+route