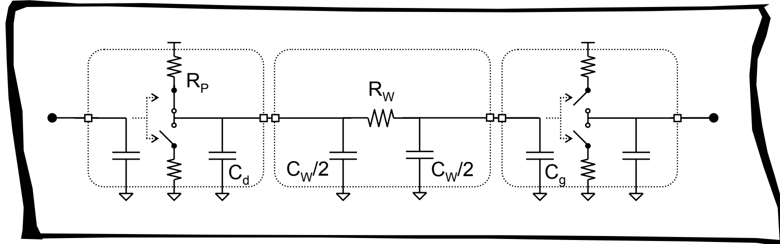
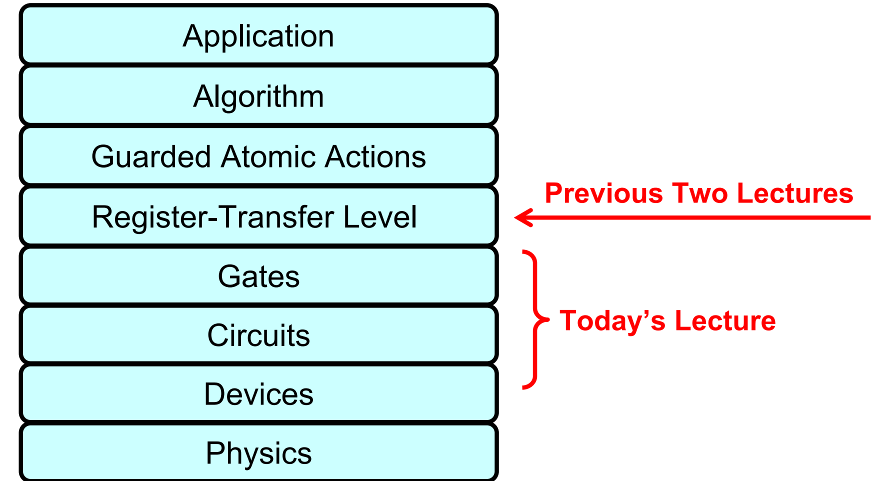
CMOS Transistors, Gates, and Wires

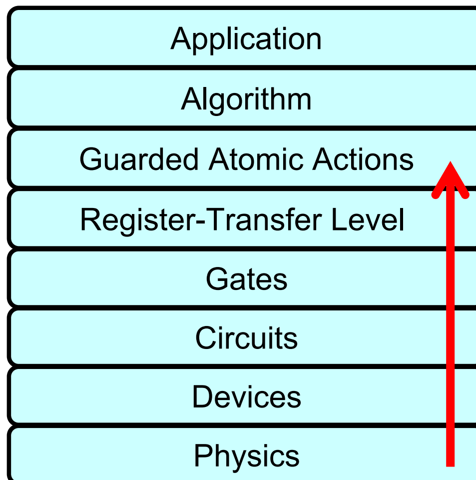


6.375 Complex Digital Systems
Christopher Batten
February 15, 2006

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



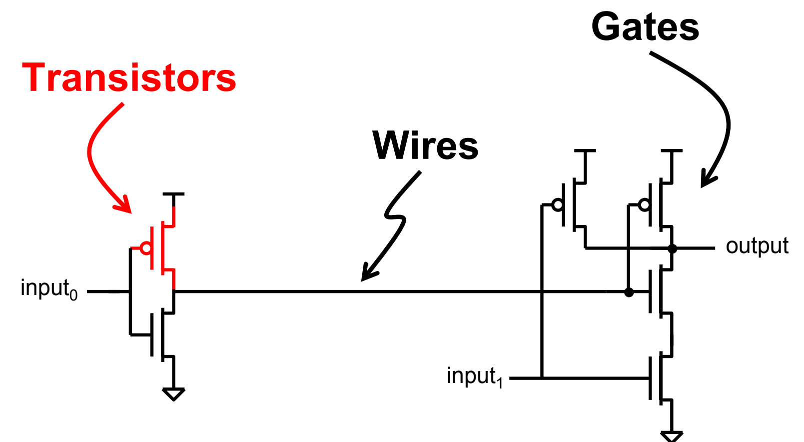
Should the hardware abstraction layers make today's lecture irrelevant?



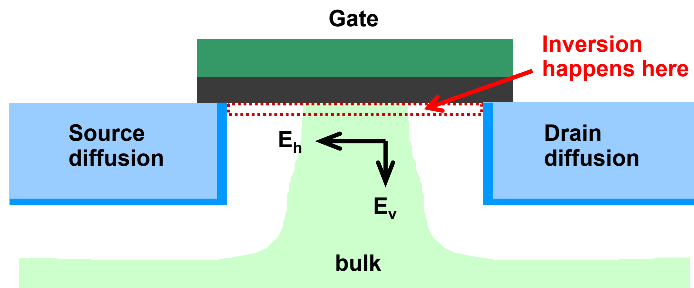
Physical design issues are increasingly pushing their way up the abstraction layers

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

CMOS Transistors, Gates, and Wires



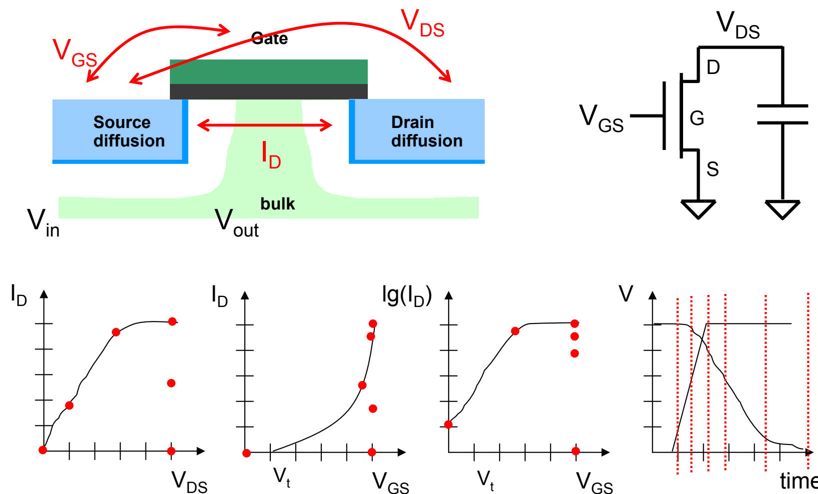
Metal Oxide-Semiconductor Field-Effect (MOSFET) Transistor



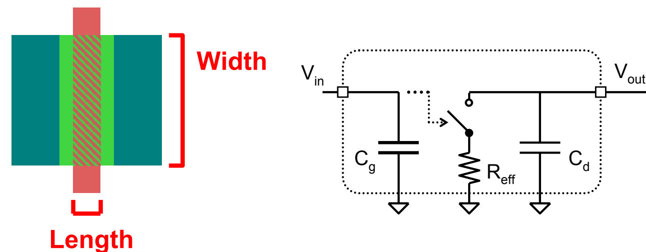
INVERSION: A sufficiently strong vertical field will attract enough electrons to the surface to create a conducting n-type channel between the source and drain.

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

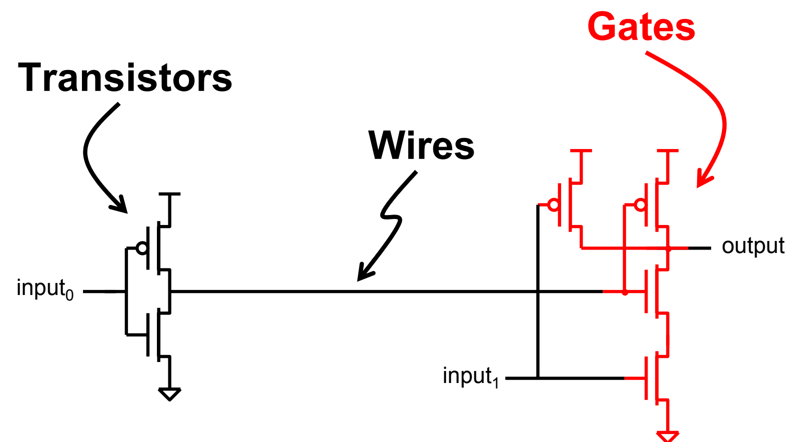


Key qualitative characteristics of MOSFET transistors

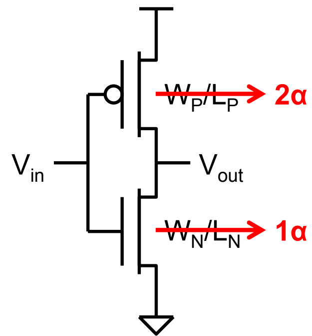


- Threshold voltage sets when transistor turns on – also impacts leakage
- I_{DS} is proportional to mobility x (W/L)
- NMOS mobility > PMOS mobility => $R_{effN} < R_{effP}$ (assume mobility ratio is 2)
- Increase W = Increase I = Decrease R_{eff}
- Increase L = Decrease I = Increase R_{eff}
- C_g proportional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

CMOS Transistors, Gates, and Wires



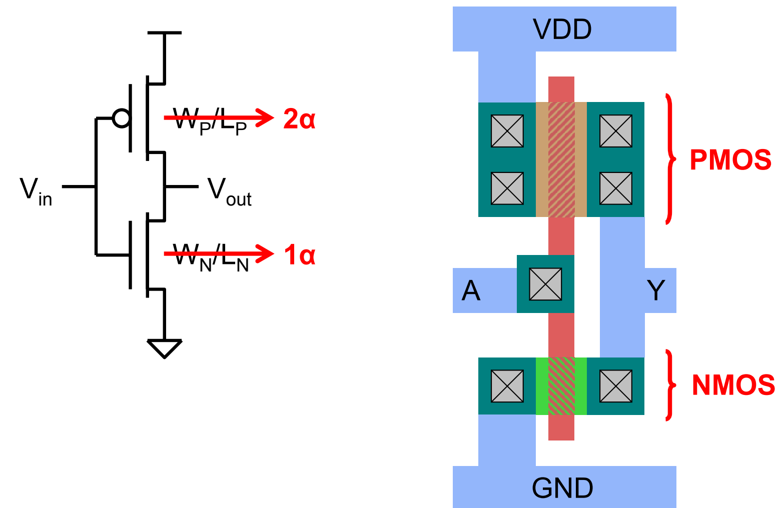
The most basic CMOS gate is an inverter



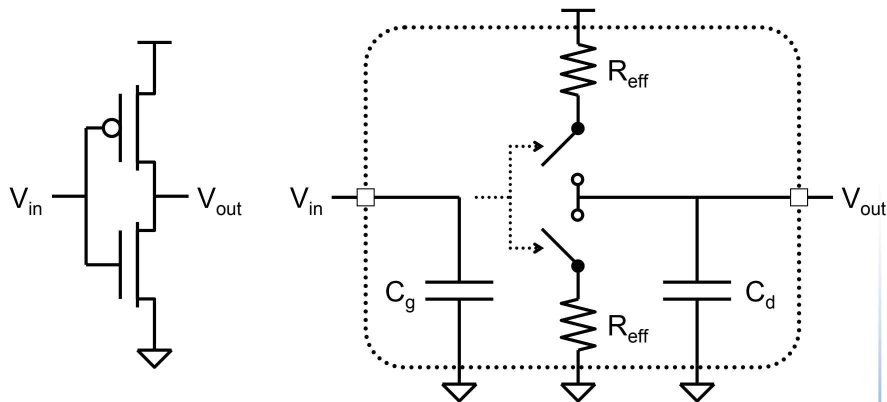
Let's make the following assumptions

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

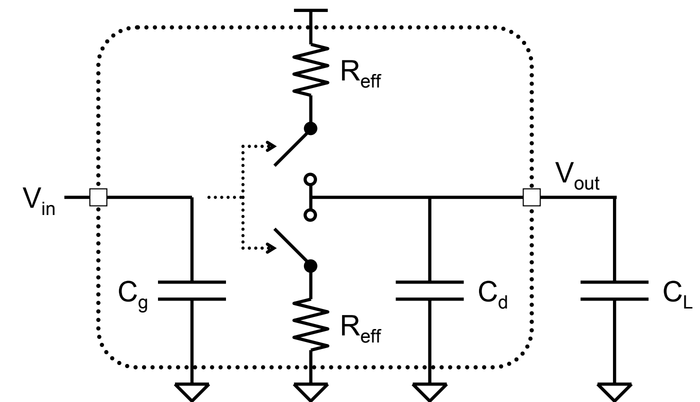


A simple RC model for the inverter can provide significant insight

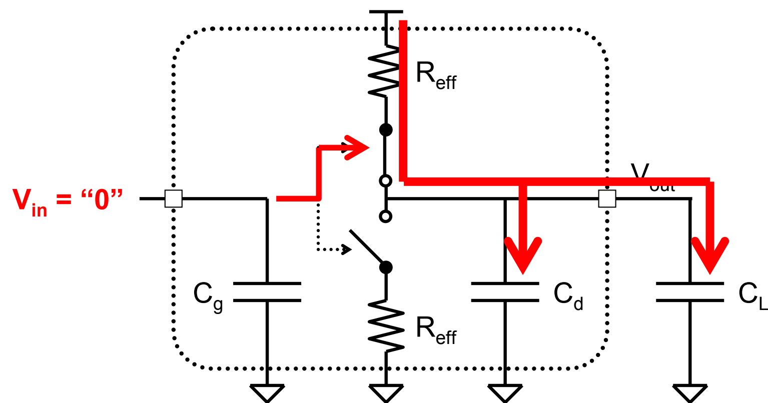


$R_{eff} = R_{eff,N} = R_{eff,P}$
$C_g = C_{g,N} + C_{g,P}$
$C_d = C_{d,N} + C_{d,P}$

A simple RC model for the inverter can provide significant insight

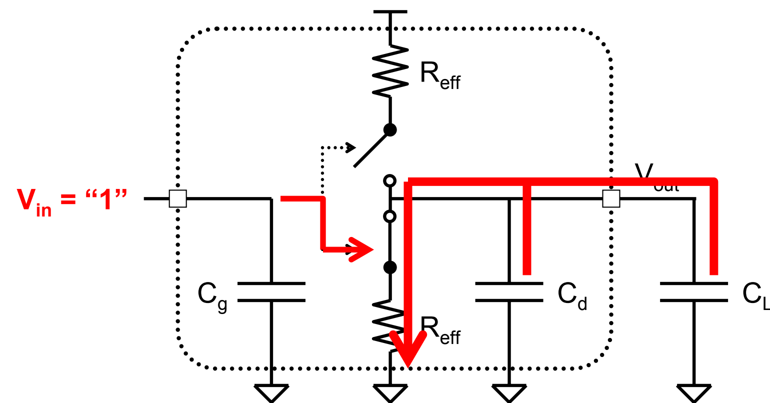


A simple RC model for the inverter can provide significant insight



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

The most basic CMOS gate is an inverter

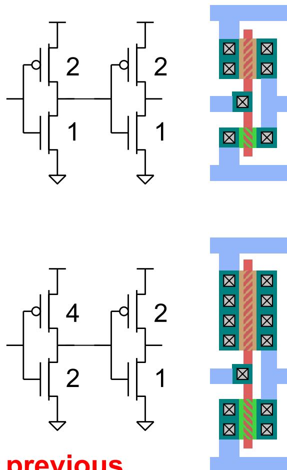


Discharge RC Time Constant = $R_{eff} \times (C_d + C_L)$

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

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

Param	Value	Units
$C_{d,N}$/μm	1.42	fF/μm
$C_{d,P}$/μm	2.40	fF/μm
$C_{g,N}$/μm	1.55	fF/μm
$C_{g,P}$/μm	1.48	fF/μm
$R_{eff,N}$ x μm	4.93	kΩ/μm
$R_{eff,P}$ x μm	10.83	kΩ/μm

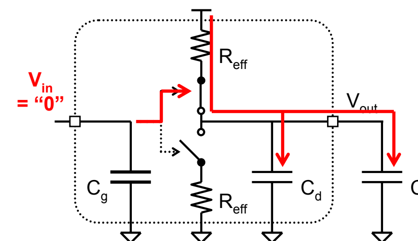


C_d = (0.5x1.42) + (1x2.40) = 3.11 fF
C_L = (0.5x1.55) + (1x1.48) = 2.26 fF
C_d+C_L = 5.37 fF
T_{PLH} = 2.2 x (10.83/1) x 5.37 = 128ps
T_{PHL} = 2.2 x (4.93/0.5) x 5.37 = 116ps

C_d = (1x1.42) + (2x2.40) = 3.66 fF
C_L = (0.5x1.55) + (1x1.48) = 2.26 fF
C_d+C_L = 5.92 fF
T_{PLH} = 2.2 x (10.83/2) x 5.92 = 70.5ps
T_{PHL} = 2.2 x (4.93/1) x 5.92 = 64.2ps

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



$$E_{0\to1} = \int_0^T P(t)\,dt = V_{DD}\int_0^T I(t)\,dt = V_{DD}\int_0^T \frac{dQ}{dt}\,dt$$

$$= V_{DD}\int_0^T C\frac{dV}{dt}\,dt = V_{DD}\int_0^{V_{DD}} (C_d + C_L)\,dV_{out}$$

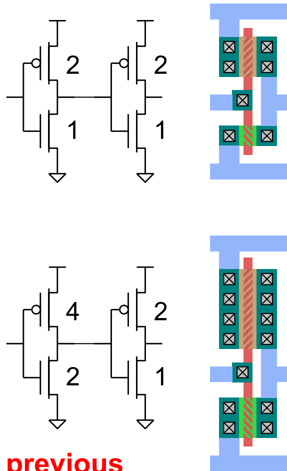
$$= (C_d + C_L)V_{DD}^2 = CV_{DD}^2$$

- 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
$C_{d,N}$/µm	1.42	fF/µm
$C_{d,P}$/µm	2.40	fF/µm
$C_{g,N}$/µm	1.55	fF/µm
$C_{g,P}$/µm	1.48	fF/µm
$R_{eff,N}$ x µm	4.93	kΩ/µm
$R_{eff,P}$ x µm	10.83	kΩ/µm

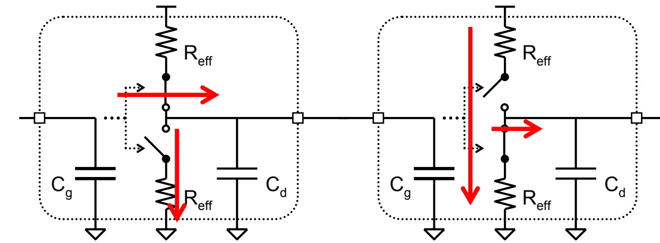


C_d = (0.5x1.42) + (1x2.40) = 3.11 fF
C_L = (0.5x1.55) + (1x1.48) = 2.26 fF
C_d+C_L = 5.37 fF
$E_{0>1}$ = 5.37 x 5^2 = 134fJ

C_d = (1x1.42) + (2x2.40) = 3.66 fF
C_L = (0.5x1.55) + (1x1.48) = 2.26 fF
C_d+C_L = 5.92 fF
$E_{0>1}$ = 5.92 x 5^2 = 148fJ

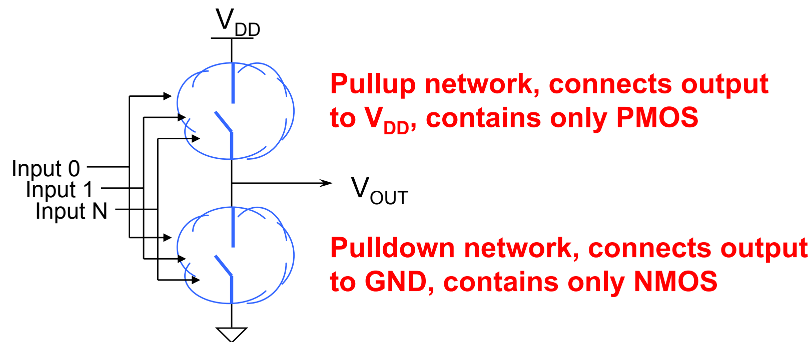
Ignores the fact that previous gate now must drive a bigger gate capacitance!

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



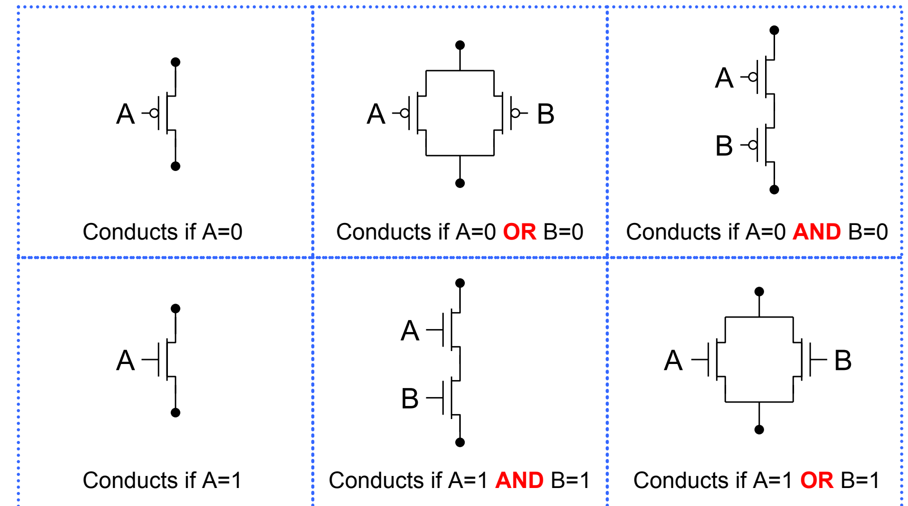
Pullup network, connects output to V_{DD}, contains only PMOS

Pulldown network, connects output to GND, contains only NMOS

For every set of input logic values, either pullup or pulldown network makes connection to VDD or GND

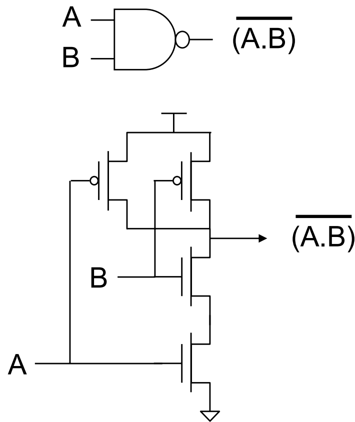
- If both connected, power rails would be shorted together
- If neither connected, output would float (tristate logic)

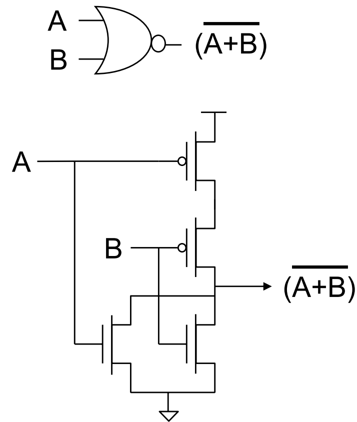
Series and parallel MOSFET networks provide natural duals of each other



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

NAND Gate **NOR Gate**

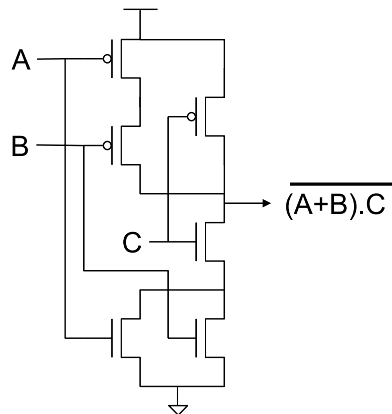




Designers can use a methodical approach to build more complex gates

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



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

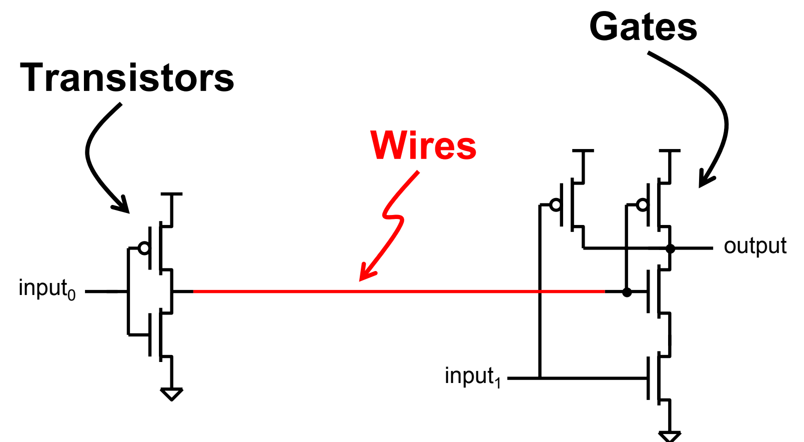
$$PD = (A + B) \cdot C$$

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

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

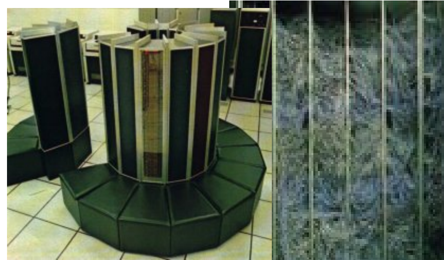
$$= (\overline{A} \cdot \overline{B}) + \overline{C}$$

CMOS Transistors, Gates, and Wires



Wires are an old problem

Cray-3
1993

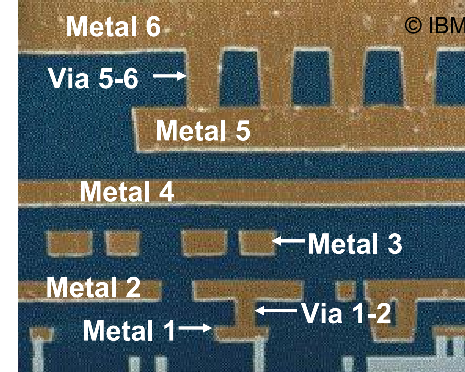


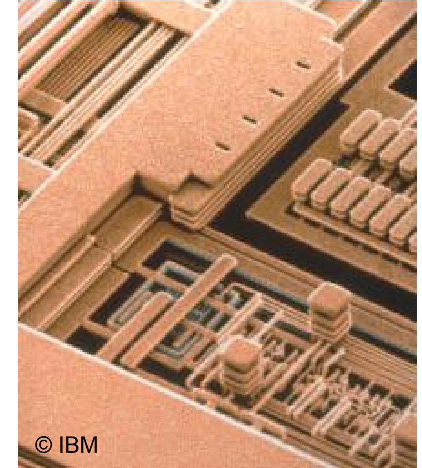
Cray-1
1976

Cray-3
wiring

Cray-1 Wiring

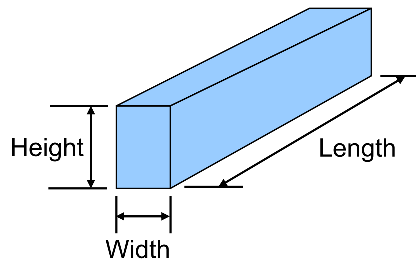
Modern interconnect stacks have six to nine or more metal layers





IBM CMOS7 process
6 layers of copper wiring
1 layer of tungsten local interconnect

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

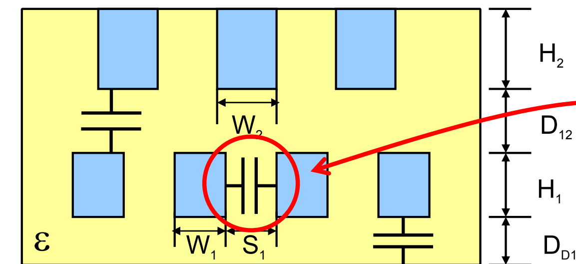


$$\text{resistance} = \frac{(\text{length} \times \text{resistivity})}{(\text{height} \times \text{width})}$$

bulk aluminum	2.8×10^{-8} Ω-m
bulk copper	1.7×10^{-8} Ω-m
bulk silver	1.6×10^{-8} Ω-m

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

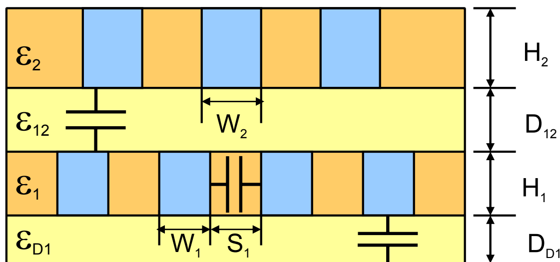
Wire capacitance is relative to the substrate and to neighboring wires



- Capacitance depends on geometry of surrounding wires and relative permittivity (ε_r) of insulating dielectric
  - silicon dioxide (SiO_2) $\varepsilon_r = 3.9$
  - silicon flouride (SiF_4) $\varepsilon_r = 3.1$
  - SiLK™ polymer $\varepsilon_r = 2.6$

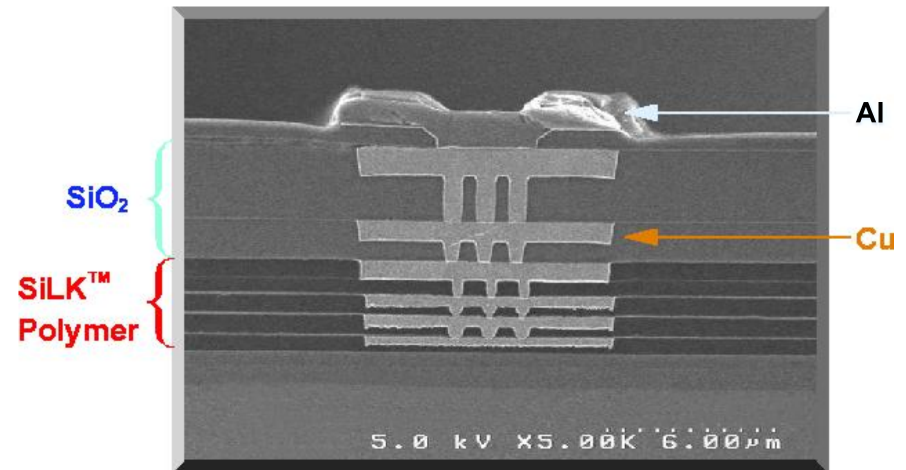
Capacitive coupling to neighbors is becoming a serious problem!

Wire capacitance is relative to the substrate and to neighboring wires



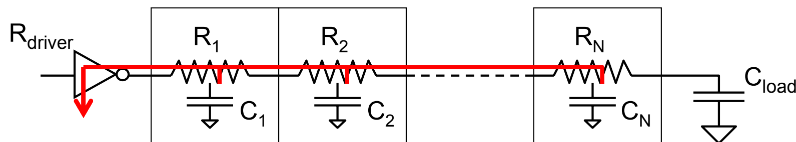
- Capacitance depends on geometry of surrounding wires and relative permittivity (ε_r) of insulating dielectric
  - silicon dioxide (SiO_2) $\varepsilon_r = 3.9$
  - silicon flouride (SiF_4) $\varepsilon_r = 3.1$
  - SiLK^TM polymer $\varepsilon_r = 2.6$
- Can have different materials between wires and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



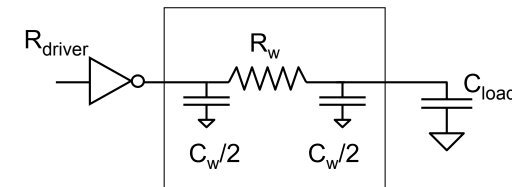
Use Penfield-Rubenstein equation to find delay

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

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



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

R_w is lumped resistance of the wire

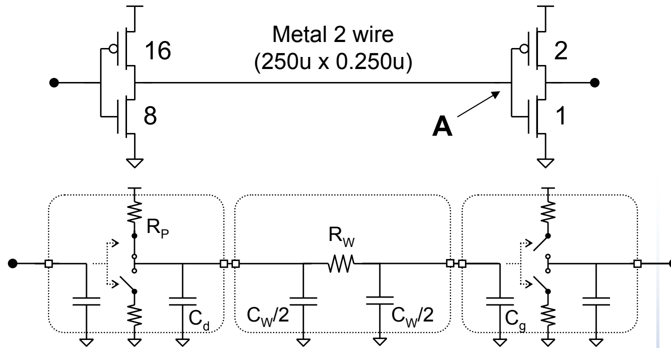
C_w is lumped capacitance

Partition half of C_w at each end

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

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

Param	Value	Units
$C_{d,N}$ / µm	1.42	fF/µm
$C_{d,P}$ / µm	2.40	fF/µm
$C_{g,N}$ / µm	1.55	fF/µm
$C_{g,P}$ / µm	1.48	fF/µm
$C_{A,M2}$ / µm²	0.016	fF/µm²
$C_{L,M2}$ / µm	0.084	fF/µm
$R_{eff,N}$ x µm	4.93	kΩ/µm
$R_{eff,P}$ x µm	10.83	kΩ/µm
R_{M2} / sq	0.07	Ω/sq

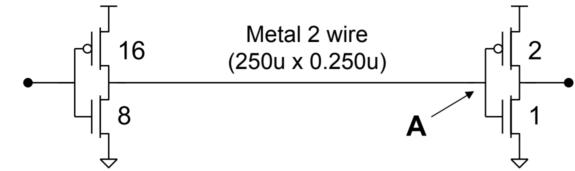


C_g = (0.5 x 1.55) + (1 x 1.48) = 2.26 fF
C_d = (4 x 1.42) + (8 x 2.40) = 24.88 fF
R_p = 10.83/8 = 1.35 kΩ
R_w = (250 / 0.25) x 0.07 = 70 Ω
C_w = ((250 x 0.25) x 0.0016) + (250 x 0.084) = 21.14 fF
T_{PLH} = 2.2 x (1350 x (21.14/2 + 24.88) + (1350 + 70) x (21.14/2 + 2.26)) = 66ps

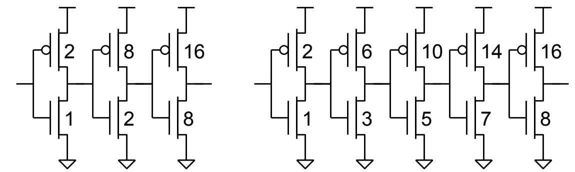
Estimate the rise time of node A using an RC delay model

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

Param	Value	Units
$C_{d,N}$ / µm	1.42	fF/µm
$C_{d,P}$ / µm	2.40	fF/µm
$C_{g,N}$ / µm	1.55	fF/µm
$C_{g,P}$ / µm	1.48	fF/µm
$C_{A,M2}$ / µm²	0.016	fF/µm²
$C_{L,M2}$ / µm	0.084	fF/µm
$R_{eff,N}$ x µm	4.93	kΩ/µm
$R_{eff,P}$ x µm	10.83	kΩ/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

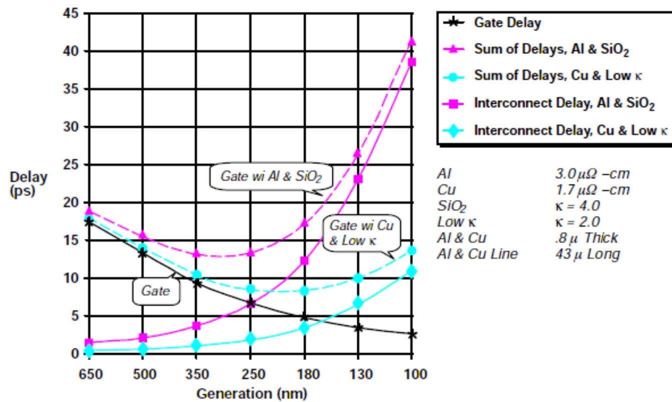
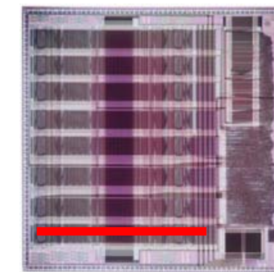


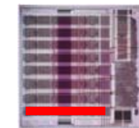
Figure 3 Calculated Gate and Interconnect Delay versus Technology Generation

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]

National Technology Roadmap for Semiconductors, SIA, 1997

Is there really an interconnect doomsday looming?



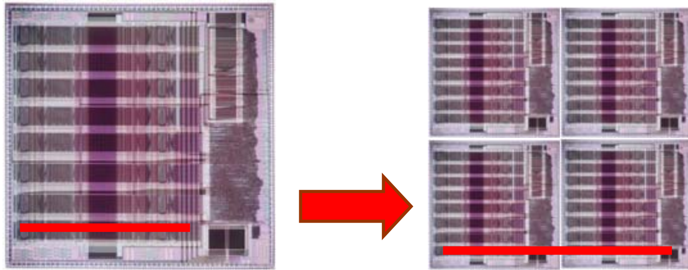


Local wire delay tracks improvement in gate delay

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

R. Ho, K. Mai, M. Horowitz, Proc. of the IEEE, Apr 2001

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

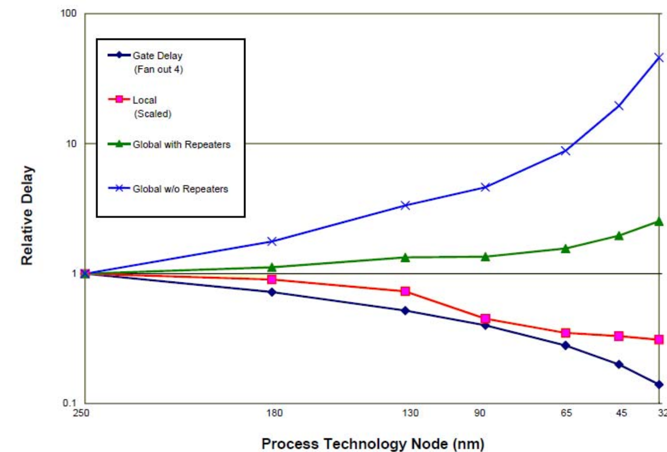


Figure 69 Delay for Metal 1 and Global Wiring versus Feature Size

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: Srini Devadas will be discussing algorithms and issues in synthesis and place+route