Wires

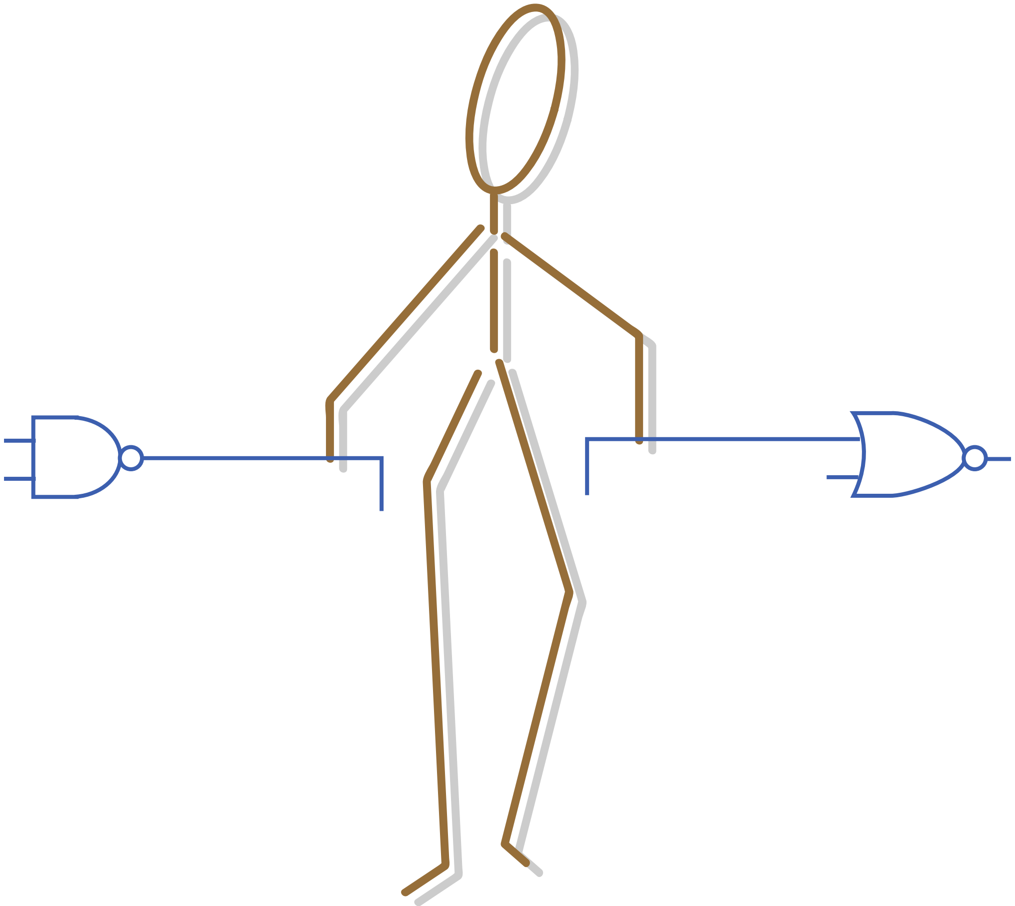
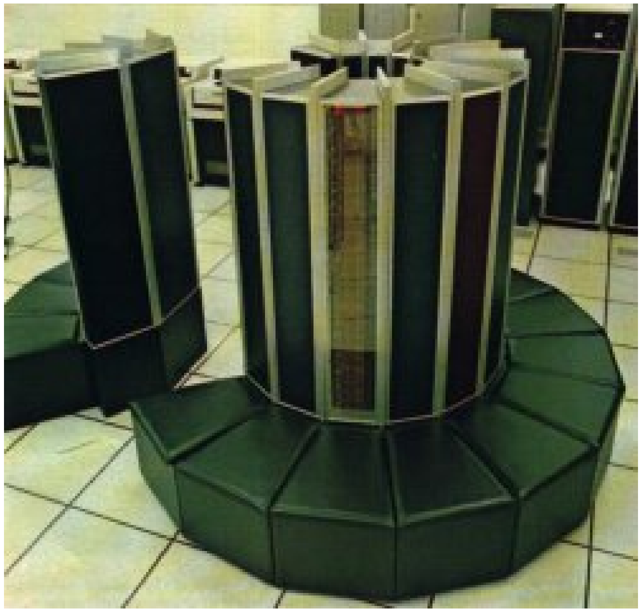
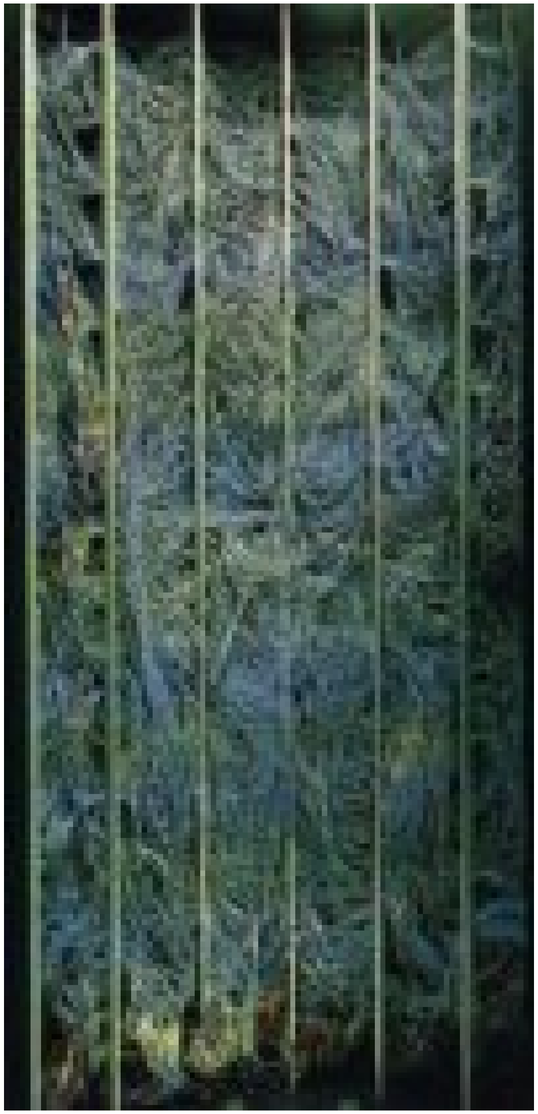


Figure by MIT OCW.

Wires are an Old Problem

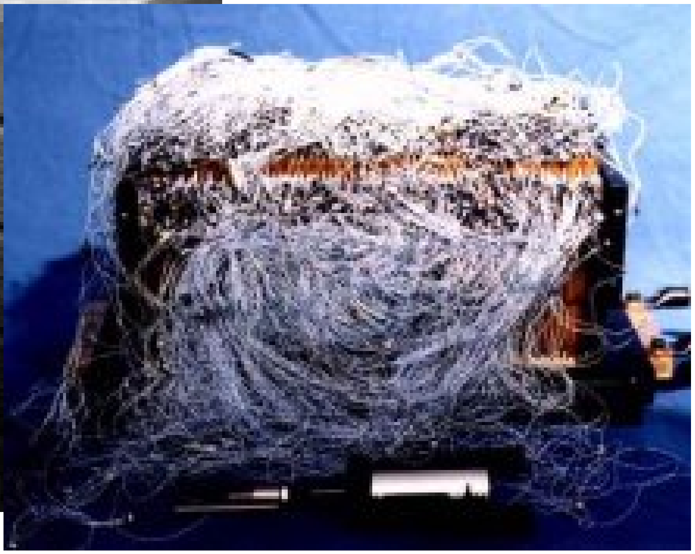
Cray-1, 1976





Cray-1 Wiring

Cray-3, 1993



Cray-3 wiring

Courtesy of Cray Company. Used with permission.

Interconnect Problems

- A lot of circuit designers are very worried about what's happening with wires in CMOS technology
- Device technology has been scaling well, with gate performance increasing linearly with decreasing feature size
- Wires scale differently, and long wires have been getting relatively slower over time
- Wire delay is a function of wire resistance and capacitance

Modern Interconnect Stack

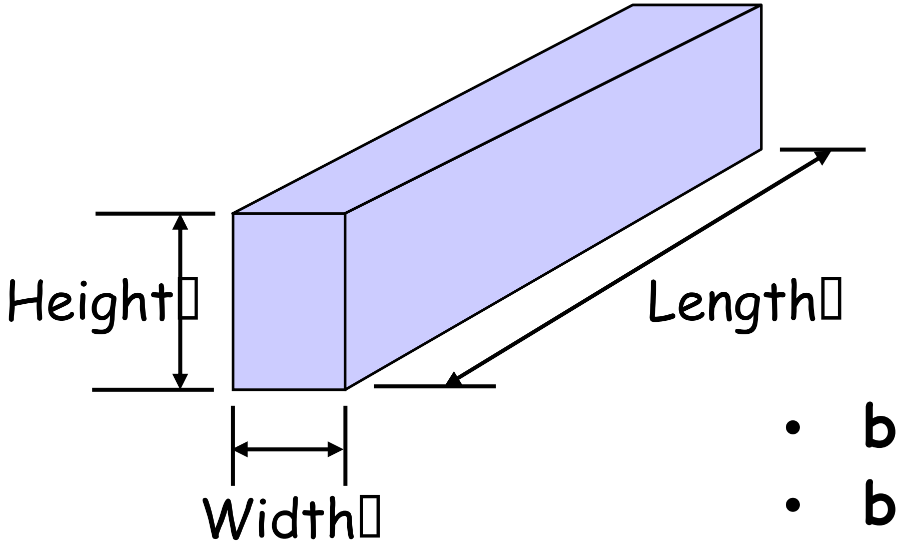
Images removed due to copyright restrictions.

IBM CMOS7 process

6 layers of copper wiring

1 layer of tungsten local interconnect

Wire Resistance



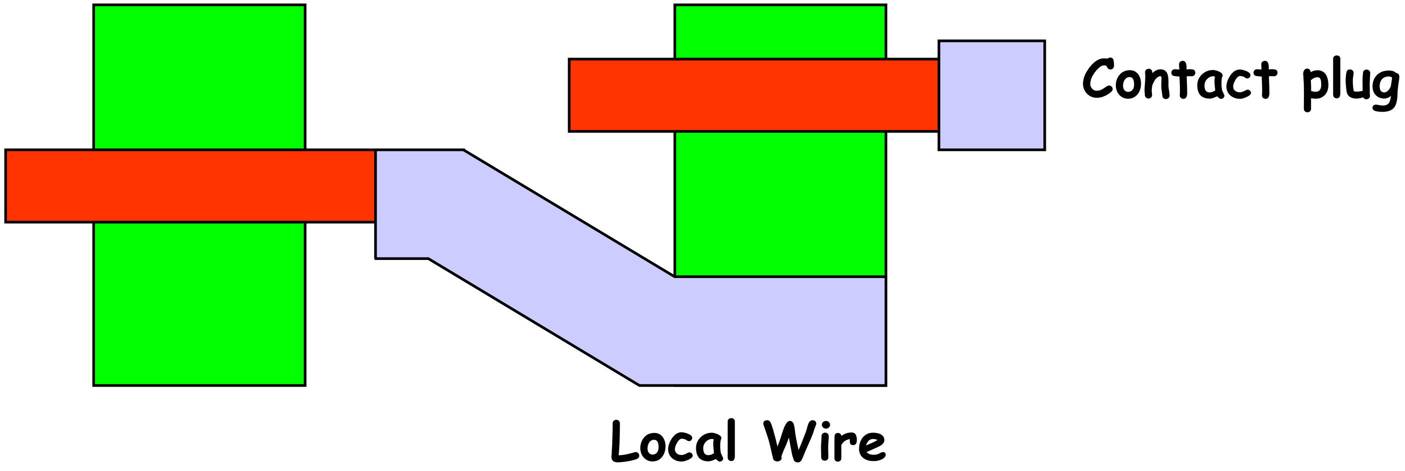
$$\text{resistance} = \frac{(\text{length} \times \text{resistivity})}{(\text{height} \times \text{width})}$$

• bulk aluminum	$2.8\text{x}10^{-8}$ Ω-m
• bulk copper	$1.7\text{x}10^{-8}$ Ω-m
• bulk silver	$1.6\text{x}10^{-8}$ Ω-m

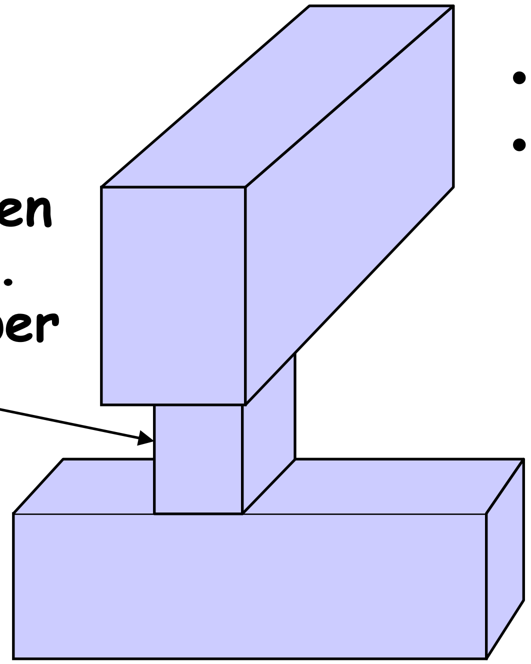
- Height (Thickness) fixed in given manufacturing process
- Resistances quoted as Ω/square
- TSMC 0.18μm 6 Aluminum metal layers
  - M1-5 0.08 Ω/square (0.5 μm x 1mm wire = 160 Ω)
  - M6 0.03 Ω/square (0.5 μm x 1mm wire = 60 Ω)

Local Interconnect

- Use contact material (tungsten) to provide extra layer of connectivity below metal 1
- Can also play same trick with silicided poly to connect gates to diffusion directly in RAMs
- Typically used to shrink memory cells or standard cells
- Contacts directly to poly gate or diffusion



Via Resistance



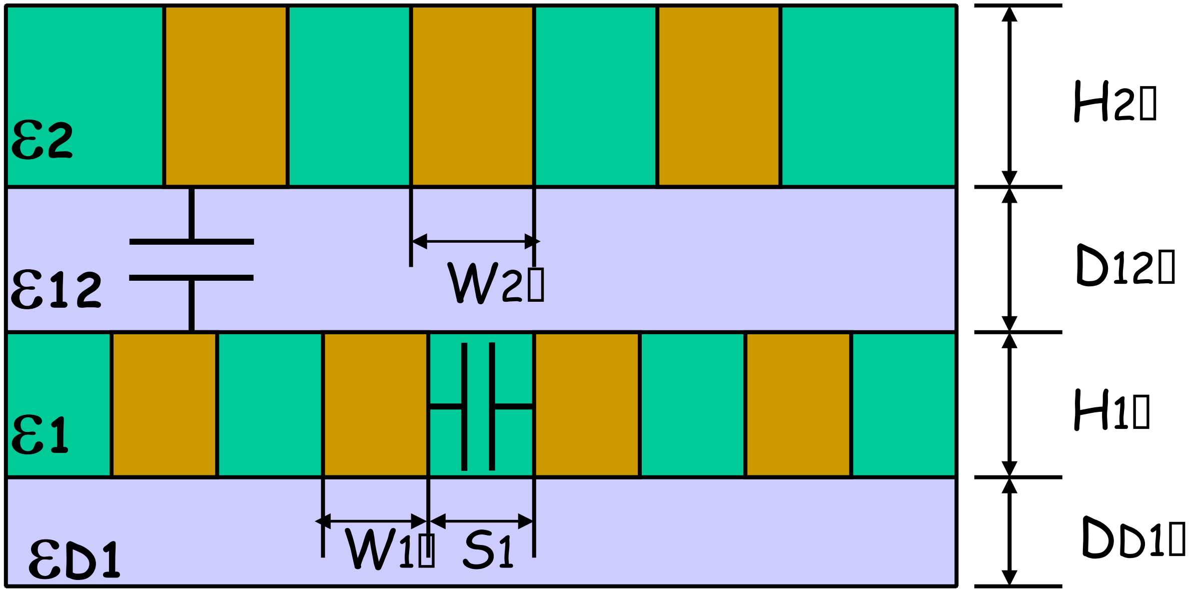
Vias made from Tungsten in Aluminum processes. Vias are Copper in Copper processes

- Via resistance significant
- TSMC 0.18μm 6-Al

Diff-M1	11.0 Ω
Poly-M1	10.4 Ω
M2-M1	4.5 Ω
M3-M1	9.5 Ω
M4-M1	15.0 Ω
M5-M1	19.6 Ω
M6-M1	21.8 Ω

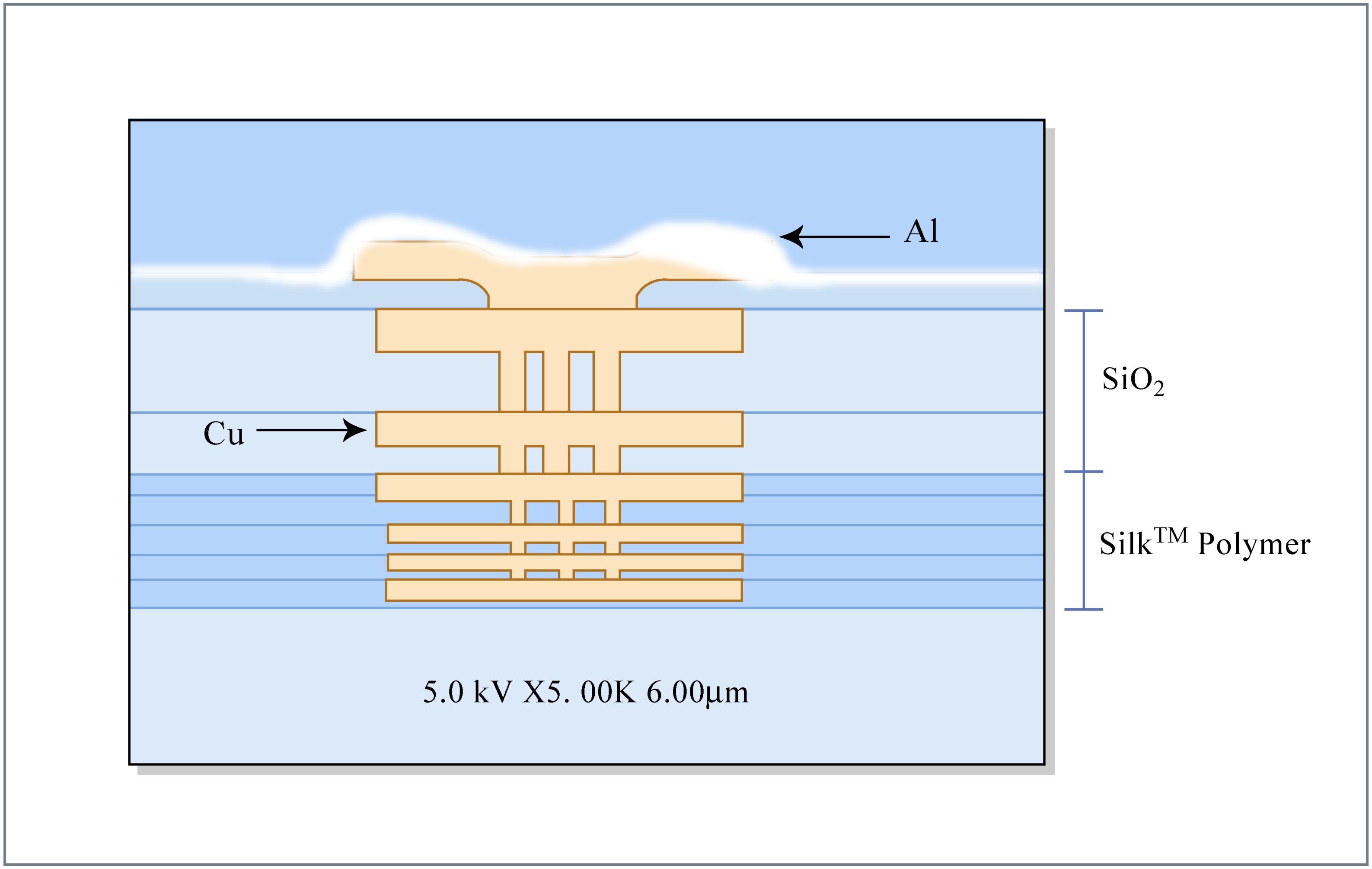
- Resistance of two via stacks at each end of M1 wire equivalent to about 0.1 mm wire (~20 Ω)
- Resistance of two via stacks at each end of M6 wire about the same as 1 mm narrow M6 wire (~60 Ω)!!!
- Use multiple vias in parallel to reduce effective contact resistance
- Copper processes have lower via resistance

Wire Capacitance



- 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, SiOF $\varepsilon_r = 3.1$
  - $SiLK^{TM}$ polymer, $\varepsilon_r = 2.6$
- Can have different materials between wires and between layers, and also different materials on higher layers

IBM Experimental 130nm Process

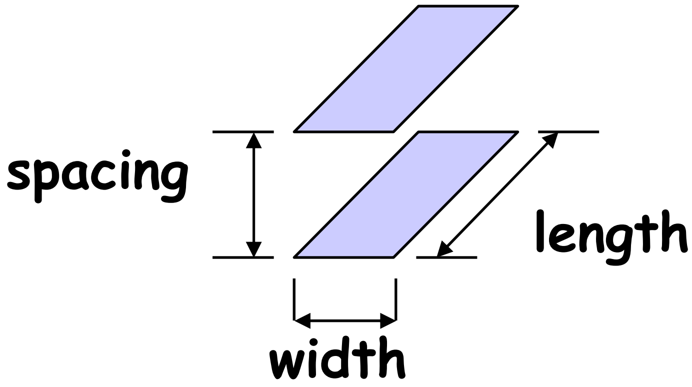


E. Barth, IBM Microelectronics

Figure by MIT OCW.

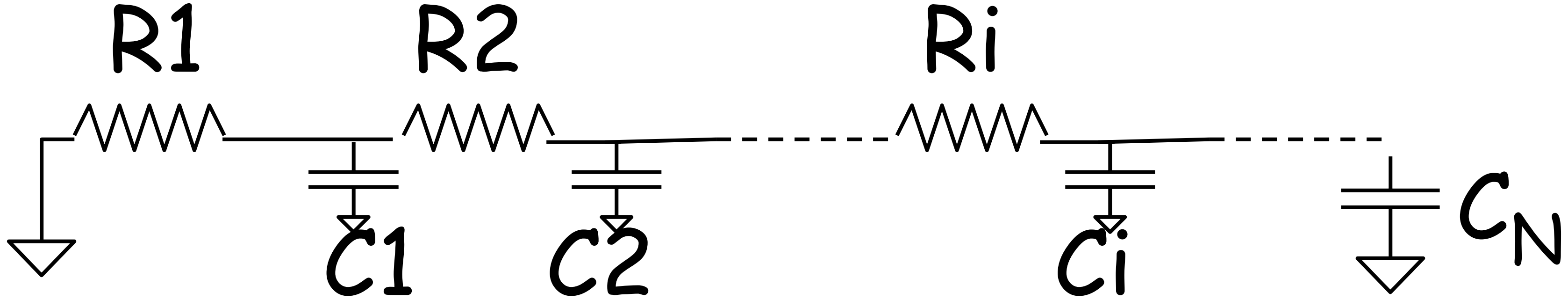
Capacitance Scaling

$$\text{parallel plate capacitance} \propto \frac{\text{width}}{\text{spacing}} \times \text{length}$$



- Capacitance/unit length ~constant with feature size scaling (width and spacing scale together)
  - Isolated wire sees approx. 100 fF/mm
  - With close neighbors about 160 fF/mm
- Need to use capacitance extractor to get accurate values

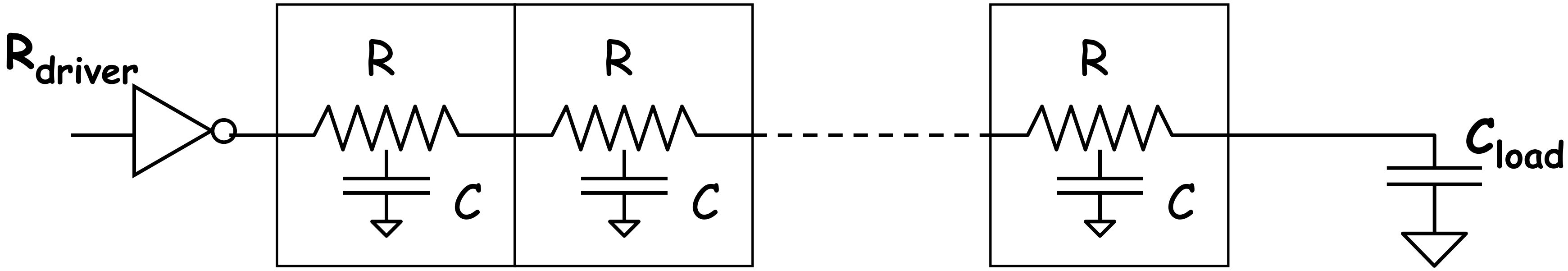
RC Delay Estimates



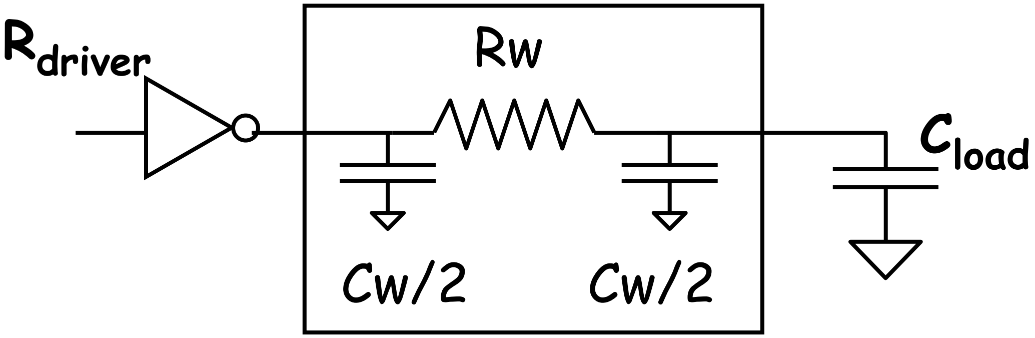
Penfield-Rubenstein model estimates:

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

Wire Delay Models



- Wire has distributed R and C per unit length
  - wire delay increases quadratically with length
  - edge rate also degrades quadratically with length
- Simple lumped Π model gives reasonable approximation
  - Rw is lumped resistance of wire
  - Cw is lumped capacitance (put half at each end)



$$\text{Delay} = \text{Rdriver} \times \frac{\text{Cw}}{2} + (\text{Rdriver} + \text{Rw}) \times \left(\frac{\text{Cw}}{2} + \text{Cload}\right)$$

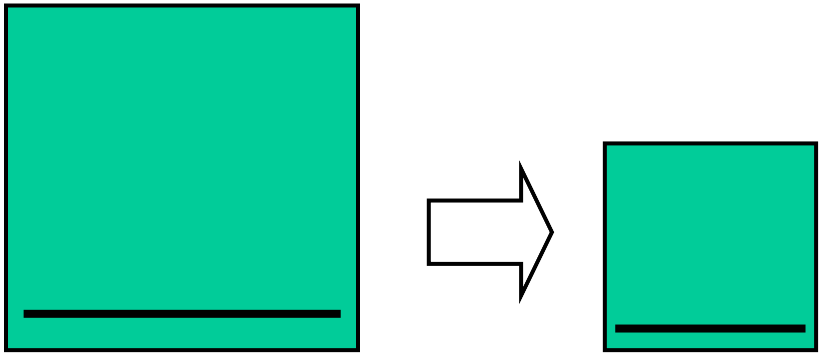
Wire Delay Example

- In 0.18μm TSMC, 5x minimum inverter with effective resistance of 3 kΩ, driving FO4 load (25fF)
- Delay = Rdriver x Cload = 75 ps

- Now add 1mm M1 wire, 0.25μm wide
  - Rw = 320Ω wire + 22Ω vias = 344Ω
  - Cw = 160fF

$$\begin{aligned}\text{Delay} &= \text{Rdriver} \times \frac{\text{Cw}}{2} + (\text{Rdriver} + \text{Rw}) \times \left(\frac{\text{Cw}}{2} + \text{Cload}\right) \\ &= 3\text{k}\Omega \times 80\text{fF} + (3\text{k}\Omega + 344\Omega) \times (80\text{fF} + 25\text{fF}) \\ &= 591\text{ps}\end{aligned}$$

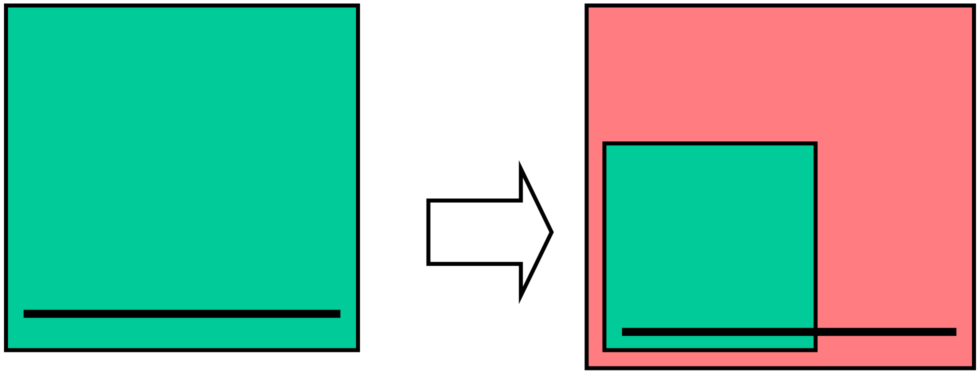
Wire Delay Scaling, Local Wires



- For wire crossing same amount of circuitry
  - Resistance stays roughly constant
    - length decreases by same amount as width, height stays large and/or change material to copper
  - Capacitance decreases by scaling factor
    - cap/unit length constant, length decreases
- Wire delay tracks improvement in gate delay

[From Mark Horowitz, DAC 2000]

Wire Delay Scaling, Global Wires



- For wire crossing whole chip
  - Resistance grows linearly
  - Capacitance stays fixed
- Wire delay increases relative to gate delay

[From Mark Horowitz, DAC 2000]

Fewer Gates per Clock Cycle

- Processors in Intel 386 generation, around 50 FO4 gate delays per clock cycle

- Pentium-4 around 16 FO4 in normal clock, around 8 FO4 delays in fast ALU section

- Fastest 64-bit adder around 7 FO4 delays

- *As measured in distance per clock cycle, wires are getting much slower*

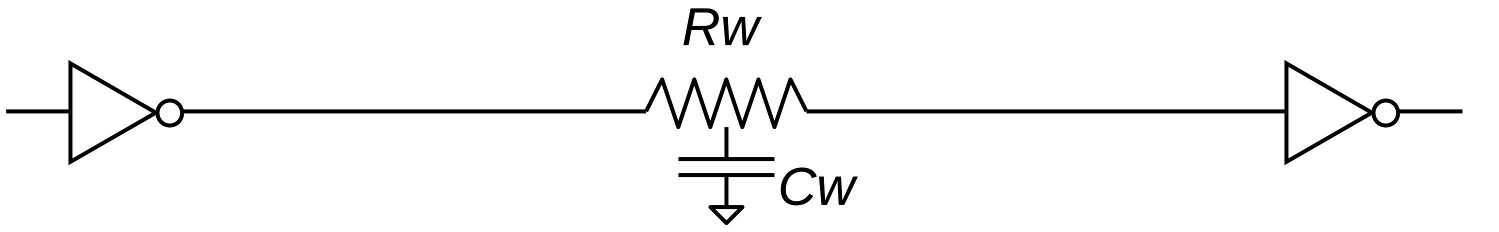
Process Technology Fixes

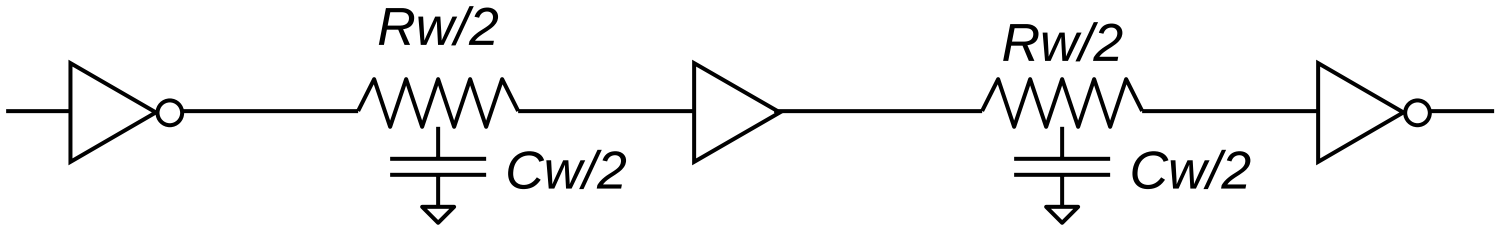
- Reduce R
  - use copper instead of aluminum, 40% reduction
  - provide upper layers with thicker metal for long range wires
  - provide more layers to improve density, makes wires shorter
- Reduce C
  - use low-k dielectric, >2x reduction possible
  - increase inter-layer spacing (limited effect, problems with via formation)
  - provide more layers to improve density, makes wires shorter

Layout Fixes

- Reduce R
  - make wires wider, increase in C is less than increase in C because of fringing fields
  - use parallel vias at contacts
  - floorplanning to keep wires short
  - careful routing to avoid unnecessary layer changes (vias)
- Reduce C
  - space wires further apart than minimum
  - avoid parallel wiring

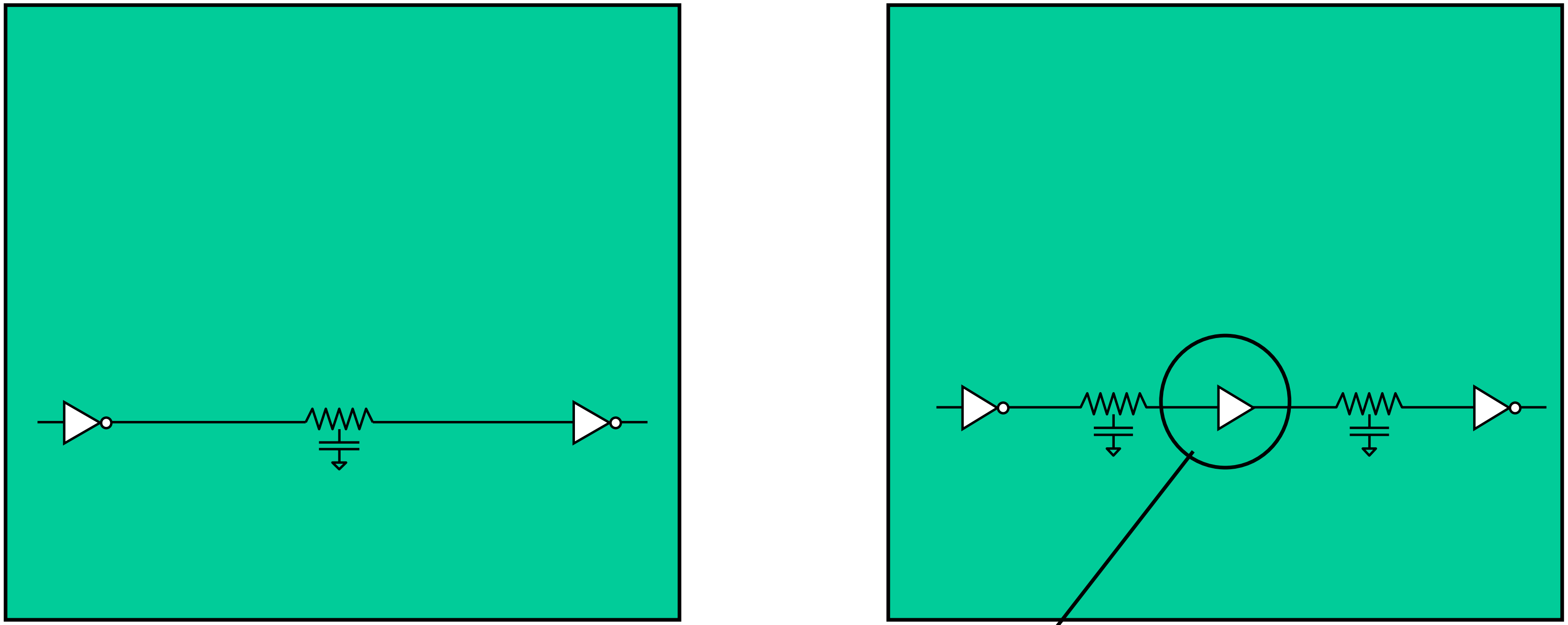
Circuit Fixes - Repeaters





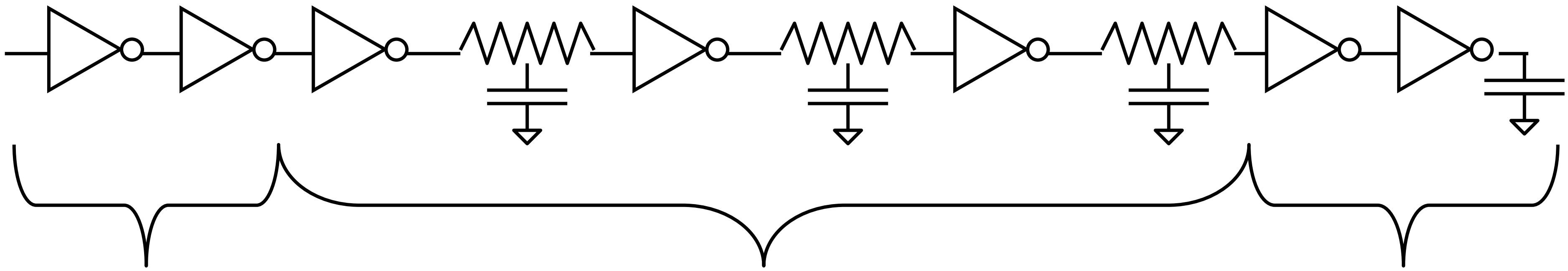
- **Use repeaters**
- **Converts quadratic dependence into linear dependence on length (but watch the constants)**
- **Can determine optimal repeater sizing for minimum delay**

Repeater Issues



- Repeater must connect to transistor layers
- Blocks other routes with vias that connect down
- Requires space on active layers for buffer transistors and power connections
- Repeaters often grouped in preallocated repeater boxes spread around chip
  - repeater location might not give ideal spacing

Repeater Staging In and Out



Stage up to optimal repeater size

Cross chip using optimal repeater size

Stage up to drive final load

- For minimum delay, must stage up to repeater size at start of wire, and stage up to load at end of wire

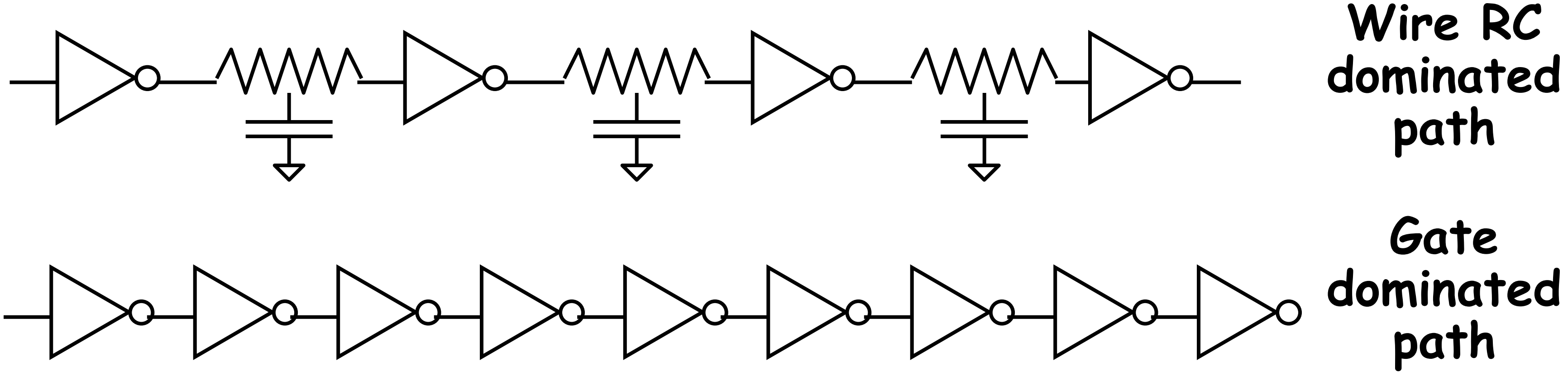
Architectural Fixes: Pentium-4

1	TC Next IP
2	
3	TC Fetch
4	
5	Drive
6	Alloc
7	Rename
8	
9	Queue
10	Schedule 1
11	Schedule 2
12	Schedule 3
13	Dispatch 1
14	Dispatch 2
15	Register File 1
16	Register File 2
17	Execute
18	Flags
19	Branch Check
20	Drive

Pipeline stages dedicated to driving signals across chip

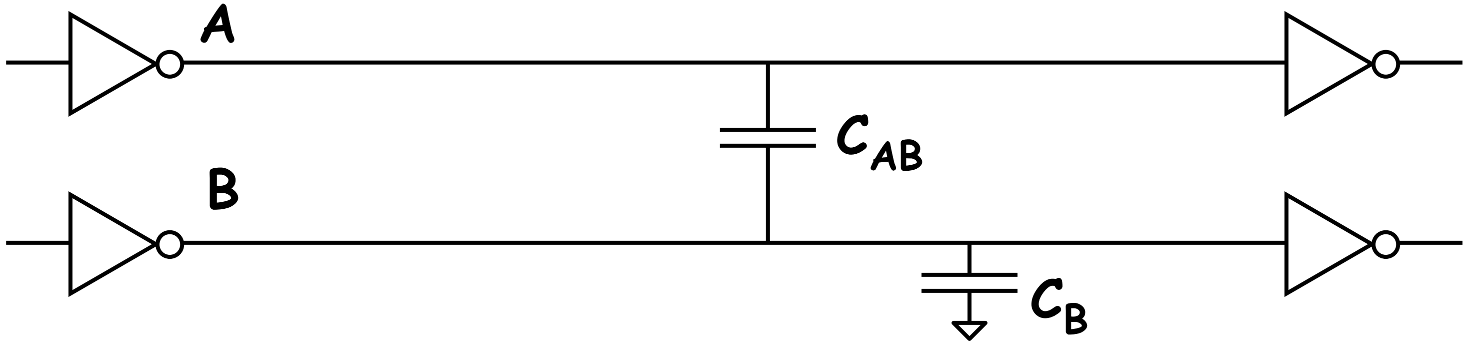
Image removed due to copyright restrictions.

Scalable Layout



- Design such that critical paths are gate dominated not interconnect dominated on first target process
- This helps ensure feature size shrinks give frequency improvements
- If critical path contains too much interconnect RC, then shrink won't see much frequency gain

Coupling Capacitances

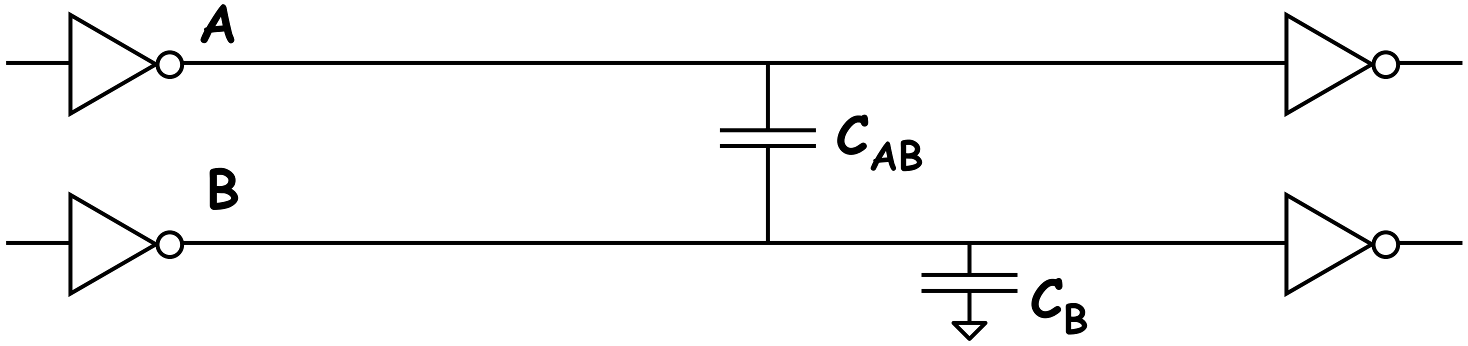


- Most capacitance is to neighboring wires
- If A switches, injects voltage noise on B
  - magnitude depends on capacitive divider formed:

$$C_{AB}/(C_{AB}+C_B)$$

- If A switches in opposite direction while B switches, coupling capacitance effectively doubles – Miller effect
- If A switches in same direction while B switches, coupling capacitance disappears
- These effects can lead to large variance in possible delay of B driver, possibly factor of 5 or 6 between best and worst case

Fixing Coupling Problems



- Avoid placing simultaneously switching signals next to each other for long parallel runs
- Reroute signals which will be quiet during switching inbetween simultaneous switching signals
- Route signals close to power rails to provide capacitance ballast
- Tough problem to solve – moving one wire can introduce new problems
  - "timing closure" causes many real-world schedule slips

Electromigration

Image removed due to copyright restrictions.

- The electrons from a DC current flow will tend to push metal atoms out of place (AC current flow OK)
- Main problem is power lines, but some signal lines have unidirectional current
- Manufacturers place a current density limit for metal to guarantee small resistance increase after ~10 years operation
- TSMC 0.18μm
  - 1mA/μm (metal wires 0.4μm thick)
  - 0.28 mA/via