Delay

Song Chen

Nov. 12, 2015

http://staff.ustc.edu.cn/~songch/da-ug.htm

Outline

- Delay
  - Gate
  - Wire (Interconnect)

Timing Optimization

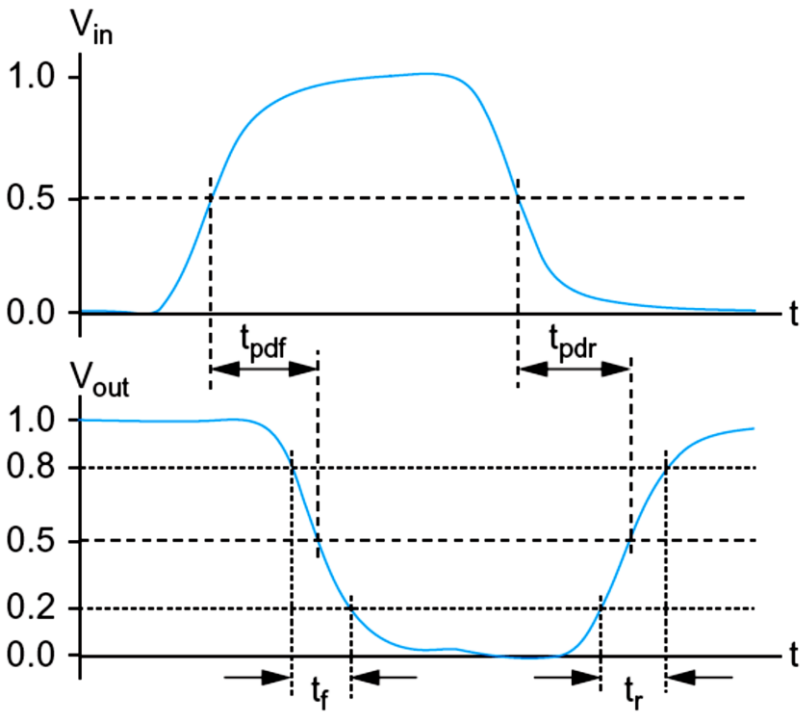
- Architecture/Micro-architecture
  - Pipeline stages, number of execution units (parallelism), size of memories
    - Selected Algorithms, Technology, Memory access speed, Wire delay
- Logic level
  - Types of functional blocks (e.g., ripple carry vs. look-ahead adders), number of gate stages in the clock cycle, fan-in/fan-out of the gates
- Circuit-level
  - Transistor sizes, CMOS logic styles
- Layout-level
  - Floorplan, cell layouts, routing, etc.

Common Design Practice

- To **write** RTL code, **Simulate**
- **Synthesize** to check if the results are fast enough
  - Timing optimization at Logic, circuit, placement level.
  - If not: recode the RTL with more parallelism or pipelining, or changes the algorithm and repeats until the timing constraints are satisfied.
  - Timing analyzers are used to check the ***Timing closure***.
  - Without an understanding of the lower levels of abstraction where the synthesizer is working, a designer may have a difficult time achieving timing closure on a challenging system.

Delay Definitions

- $\mathbf{t_{pdr}}/\mathbf{t_{pdf}}$: *rising/falling* ***propagation*** *(maximum) delay*
  - From input to rising/falling output crossing $V_{DD}/2$
- $\mathbf{t_{cdr}}/\mathbf{t_{cdf}}$: *rising/falling* ***contamination*** *(minimum) delay*
  - From input to falling output crossing $V_{DD}/2$
- $\mathbf{t_{pd}}/\mathbf{t_{cd}}$: *average propagation delay*
  - $t_{pd} = (t_{pdr} + t_{pdf})/2$
- $\mathbf{t_r}$: *rise time*
  - From output crossing 0.2 V_{DD} to 0.8 V_{DD}
- $\mathbf{t_f}$: *fall time*
  - From output crossing 0.8 V_{DD} to 0.2 V_{DD}

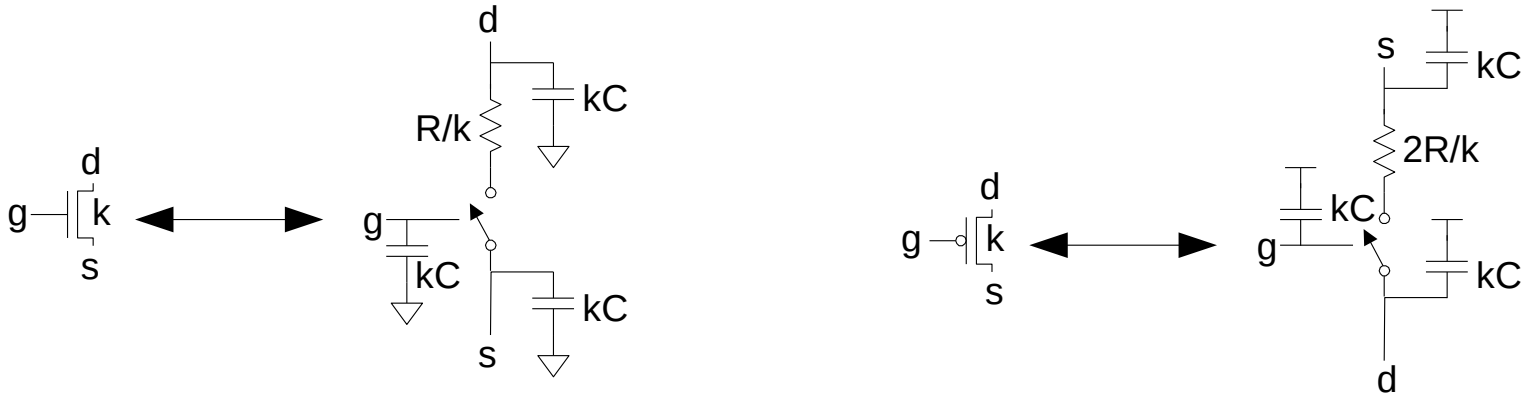


Delay Estimation

- We would like to be able to easily estimate delay
  - Not as accurate as simulation
  - But easier to ask “What if?”
- Use RC delay models to estimate delay
  - C = total capacitance on output node
  - Use *effective resistance* R
  - So that t_{pd} = RC
- Characterize transistors by finding their effective R
  - Depends on average current as gate switches

RC Delay Model

- Use equivalent circuits for MOS transistors
  - Ideal switch + capacitance and ON resistance
  - Unit nMOS has resistance R, capacitance C
  - Unit pMOS has resistance 2R, capacitance C
- Capacitance proportional to width
- Resistance inversely proportional to width

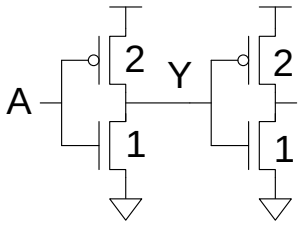


RC Values

- Capacitance
  - $C = C_g = C_s = C_d = 2$ fF/μm of gate width in 0.6 μm
  - Gradually decline to 1 fF/μm in nanometer techs.
- Resistance
  - $R \approx 6$ KΩ*μm in 0.6 μm process
  - Improves with shorter channel lengths
- Unit transistors
  - May refer to minimum contacted device (4/2 λ)
  - Or maybe 1 μm wide device
  - Doesn't matter as long as you are consistent

Inverter Delay Estimate

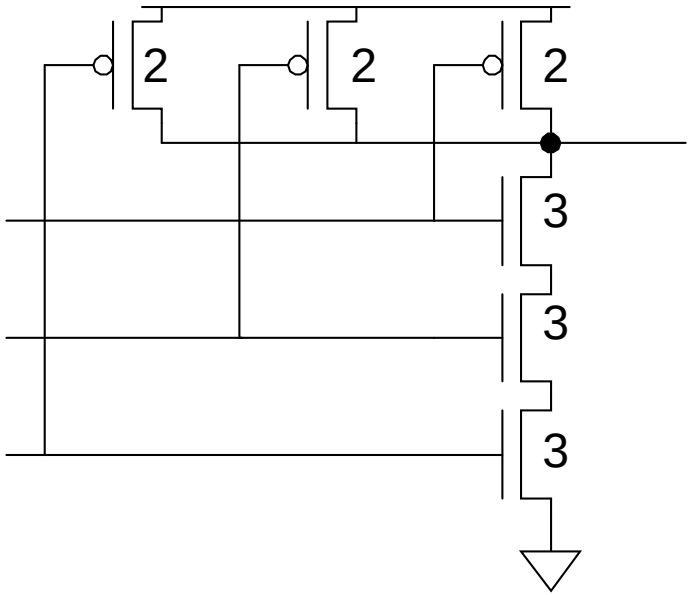
- Estimate the delay of a fanout-of-1 inverter



d = 6RC

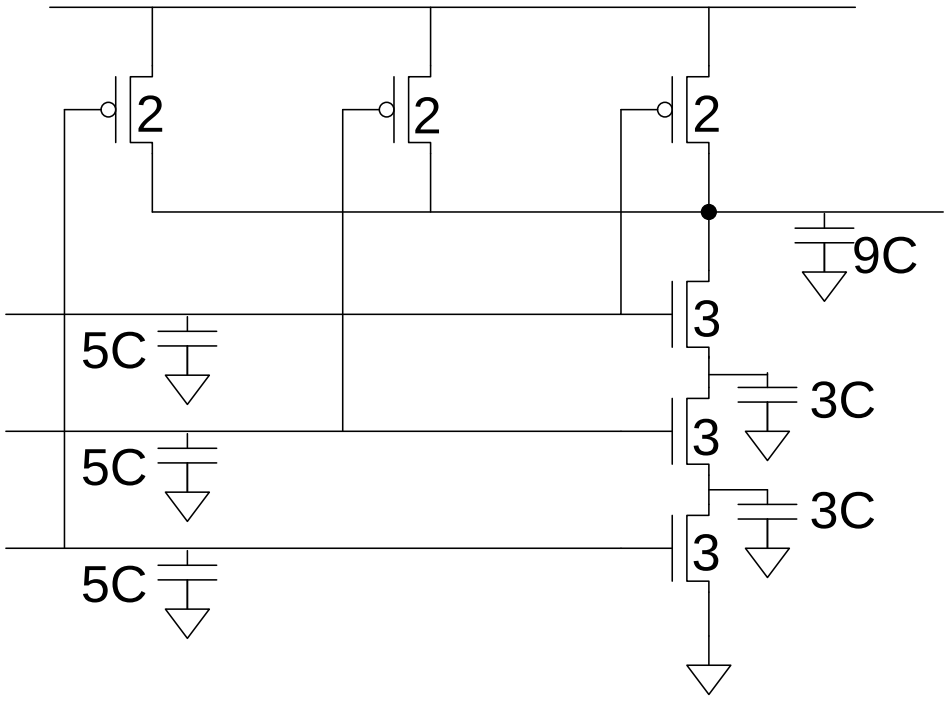
Example: 3-input NAND

- Sketch a 3-input NAND with transistor widths chosen to achieve effective rise and fall resistances equal to a unit inverter (R).



3-input NAND Caps

- Annotate the 3-input NAND gate with gate and diffusion capacitance.

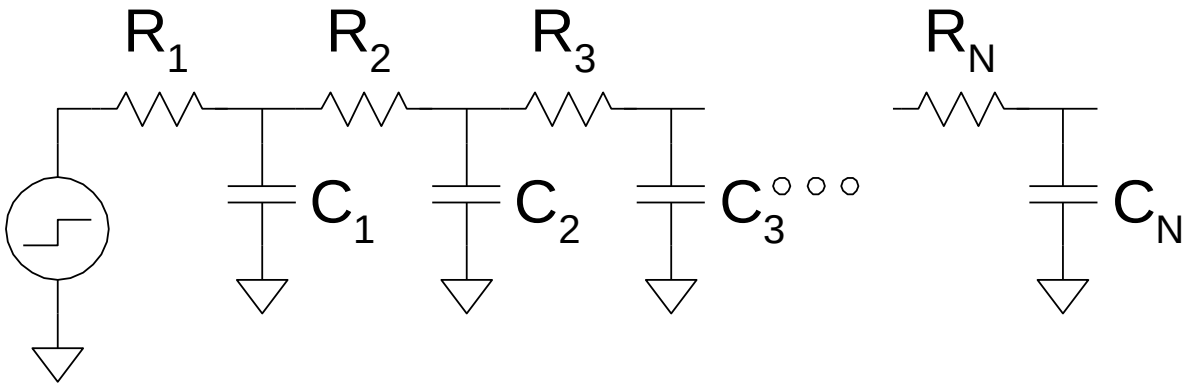


Elmore Delay

- ON transistors look like resistors
- Pullup or pulldown network modeled as *RC ladder*
- Elmore delay of RC ladder

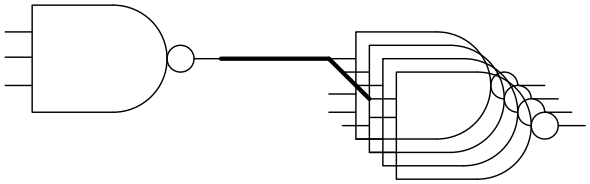
$$t_{pd} \approx \sum_{\text{nodes } i} R_{i-to-source} C_i$$

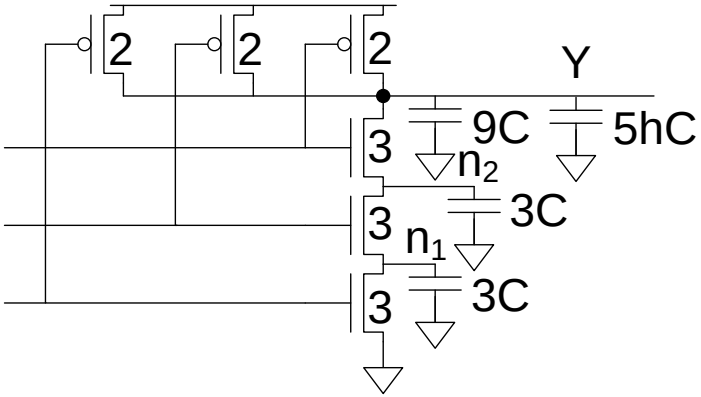
$$= R_1 C_1 + \left(R_1 + R_2\right) C_2 + \ldots + \left(R_1 + R_2 + \ldots + R_N\right) C_N$$

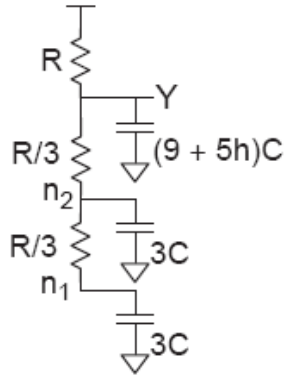


Example: 3-input NAND

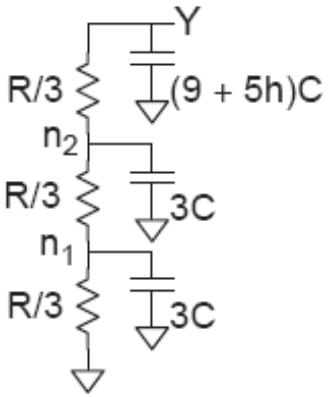
- Estimate worst-case rising and falling delay of 3-input NAND driving *h* identical gates.







$t_{pdr} = (15 + 5h)RC$



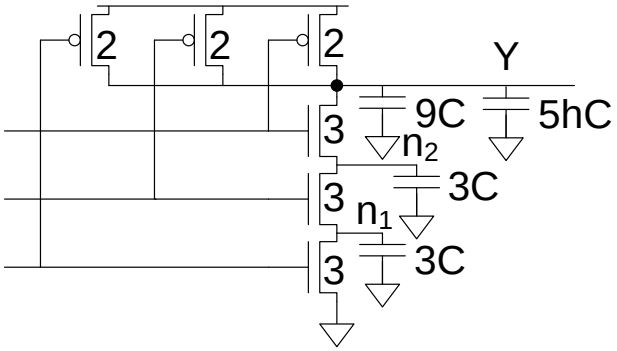
$$t_{pdf} = (3C)(R/3) + (3C)(R/3 + R/3) + [(9 + 5h)C](R/3 + R/3 + R/3)$$
$$= (12 + 5h)RC$$

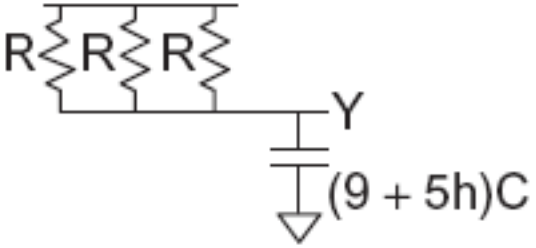
Delay Components

- Delay has two parts
  - *Parasitic delay*
    - 9 or 11 RC
    - Independent of load
  - *Effort delay*
    - 5h RC
    - Proportional to load capacitance

Contamination Delay

- Best-case (contamination) delay can be substantially less than propagation delay.
- Ex: If all three inputs fall simultaneously

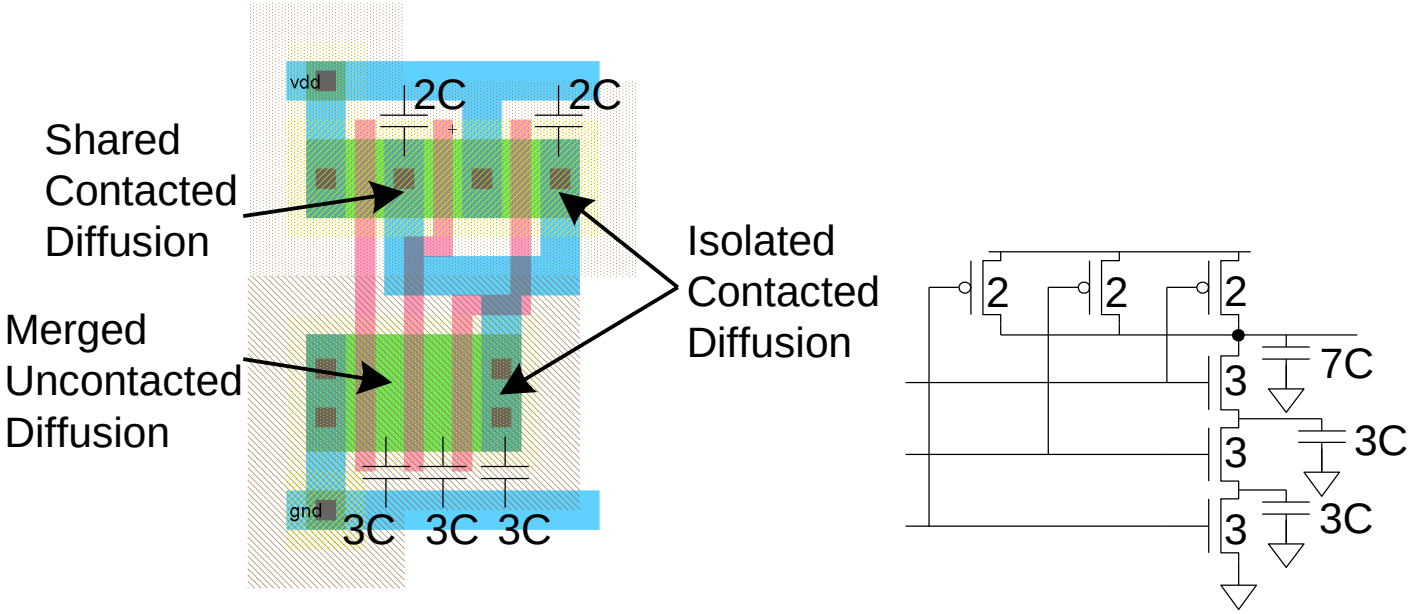




$$t_{cdr} = \left[\left(9+5h\right)C\right]\left(\frac{R}{3}\right) = \left(3+\frac{5}{3}h\right)RC$$

Diffusion Capacitance

- We assumed contacted diffusion on every s / d.
- Good layout minimizes diffusion area
- Ex: NAND3 layout shares one diffusion contact
  - Reduces output capacitance by 2C
  - Merged uncontacted diffusion might help too



Timing Analysis Delay Models

- Slope-Based Linear Model
  - delay_rise = intrinsic_rise + rise_resistance x capacitance + slope_rise x delay_previous
  - delay_fall = intrinsic_fall + fall_resistance x capacitance + slope_fall x delay_previous
- Nonlinear Delay Model: Two-dimensional interpolation
  - Look up the delay from a table based on the **load capacitance** and the **input slope**
- Current Source Model
  - Express the **output DC current** as a **nonlinear function** of the input and output **voltages** of the cell
  - Analyzer numerically integrates the output current to find the **voltage** as a **function of time** into an arbitrary RC network and to solve for the **propagation delay**.

Wire/Interconnect

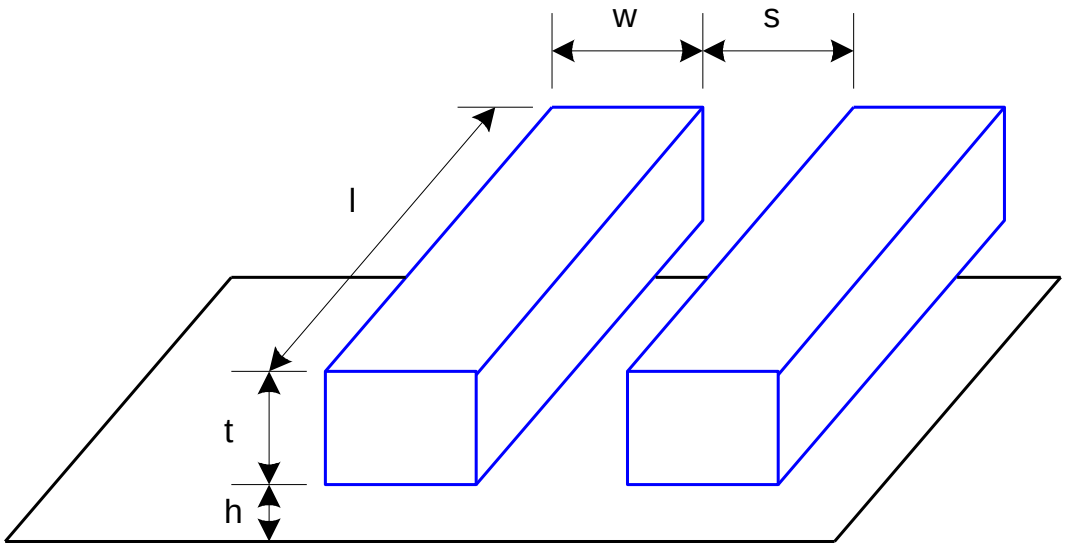
- Introduction
- Interconnect Modeling
  - Wire Resistance
  - Wire Capacitance
- Wire RC Delay
- Crosstalk
- Wire Engineering
- Repeaters

Introduction

- Chips are mostly made of wires called *interconnect*
  - Transistors are little things under the wires
  - Many layers of wires
- Wires are as important as transistors
  - Speed
  - Power
  - Noise
- Alternating layers run orthogonally

Wire Geometry

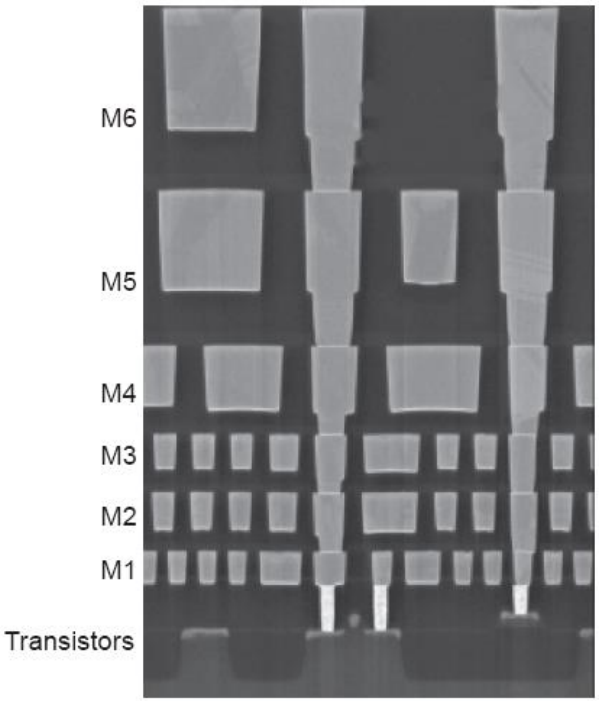
- Pitch = w + s
- Aspect ratio: AR = t/w
  - Old processes had AR << 1
  - Modern processes have AR ≈ 2
    - Pack in many skinny wires



Layer Stack

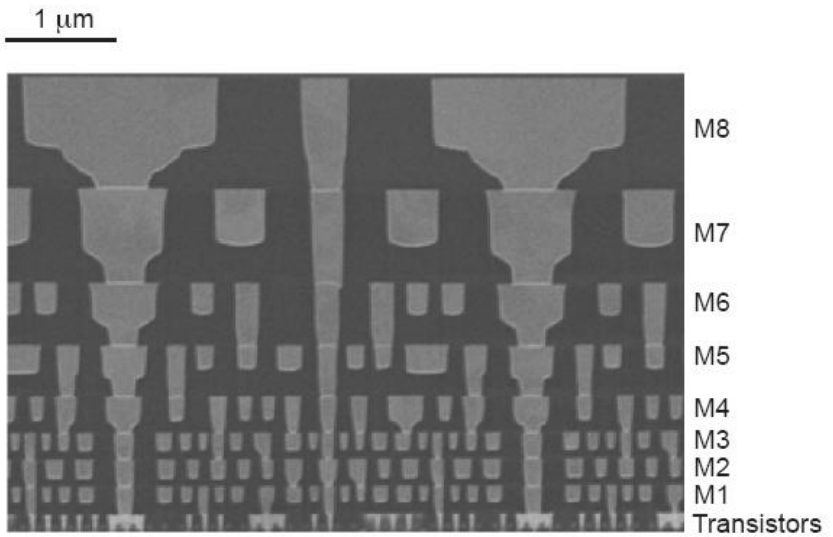
- AMI 0.6 μm process has 3 metal layers
  - M1 for within-cell routing
  - M2 for vertical routing between cells
  - M3 for horizontal routing between cells
- Modern processes use 6-10+ metal layers
  - M1: thin, narrow (< 3λ)
    - High density cells
  - Mid layers
    - Thicker and wider, (density vs. speed)
  - Top layers: thickest
    - For V_{DD}, GND, clk

Example



Intel 90 nm Stack

[Thompson02]

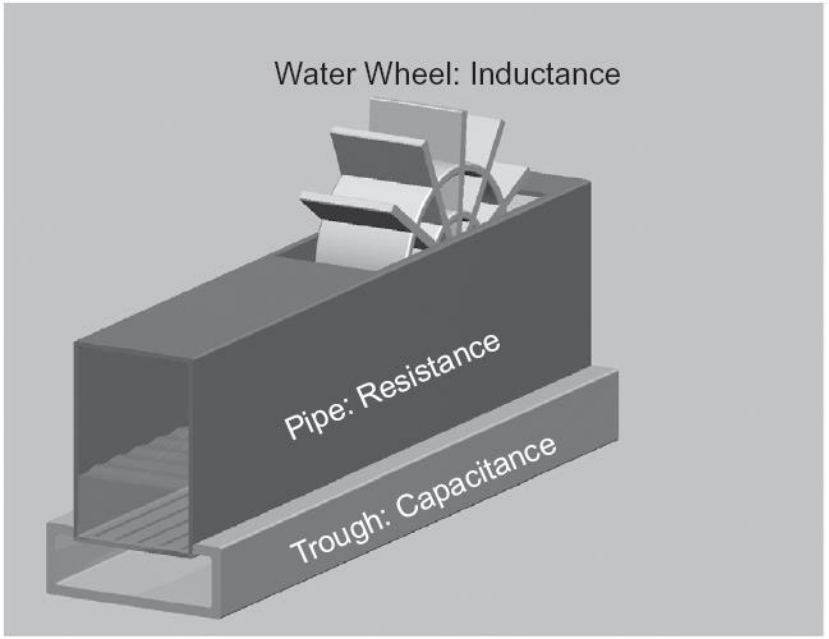


Intel 45 nm Stack

[Moon08]

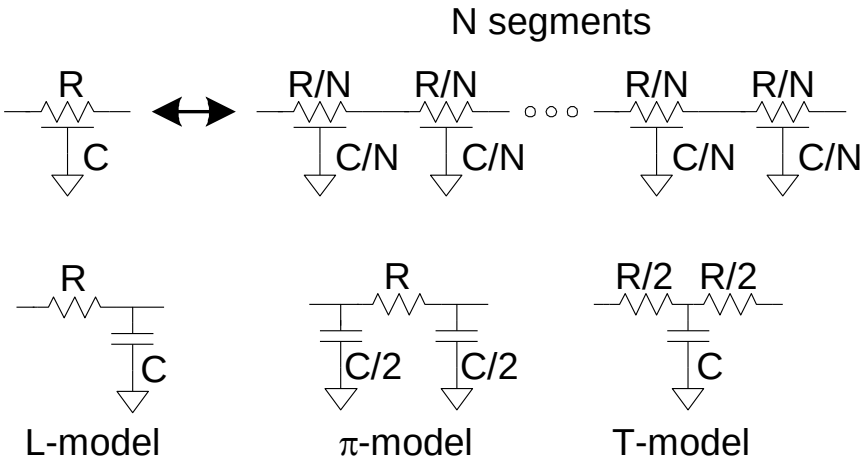
Interconnect Modeling

- Current in a wire is analogous to current in a pipe
  - Resistance: narrow size impedes flow
  - Capacitance: trough under the leaky pipe must fill first
  - Inductance: paddle wheel inertia opposes changes in flow rate
    - Negligible for most wires



Lumped Element Models

- Wires are a distributed system
  - Approximate with lumped element models



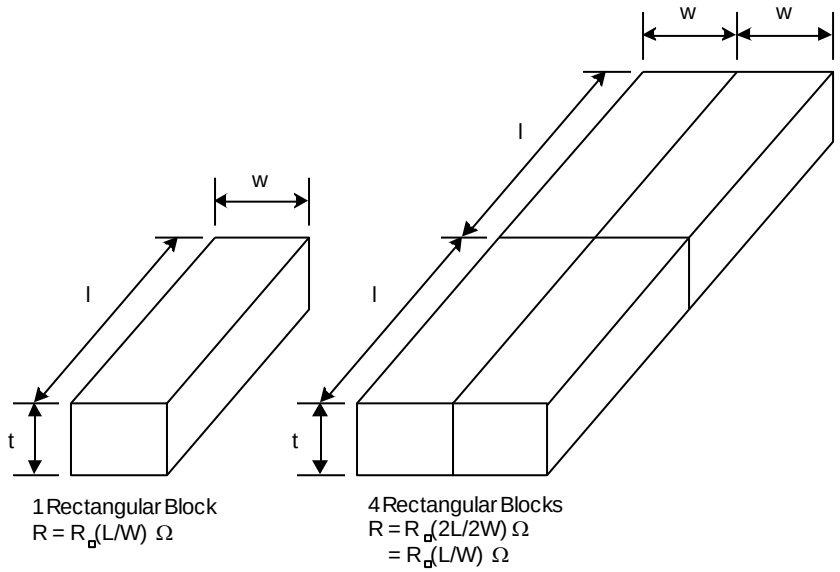
- 3-segment π-model is accurate to 3% in simulation
- L-model needs 100 segments for same accuracy!
- Use single segment π-model for Elmore delay

Wire Resistance

- ρ = *resistivity* (Ω*m)

$$R =$$

- $R_{\square}$ = *sheet resistance* ($\Omega/\square$)
  - $\square$ is a dimensionless unit(!)
- Count number of squares
  - $R = R_{\square} * (\#\ \text{of squares})$



Choice of Metals

- Until 180 nm generation, most wires were aluminum
- Contemporary processes normally use copper
  - Cu atoms diffuse into silicon and damage FETs
  - Must be surrounded by a diffusion barrier

Metal	**Bulk resistivity ($\mu\Omega \cdot$ cm)**
Silver (Ag)	1.6
Copper (Cu)	1.7
Gold (Au)	2.2
Aluminum (Al)	2.8
Tungsten (W)	5.3
Titanium (Ti)	43.0

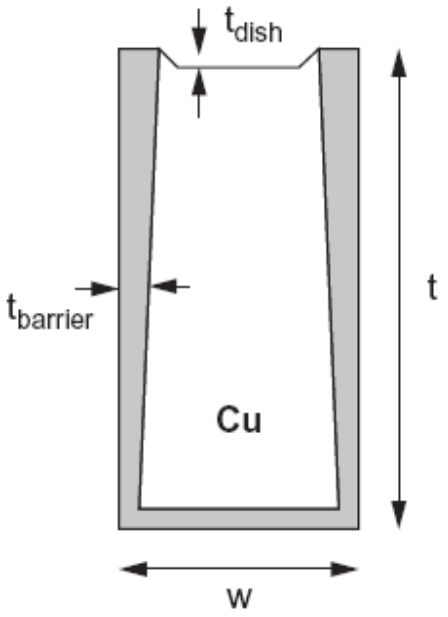
Contacts Resistance

- Contacts and vias also have 2-20 Ω
  - Depending on the contacted materials and size of the contact
- Use many contacts for lower R
  - Many small contacts for current crowding around periphery

Copper Issues

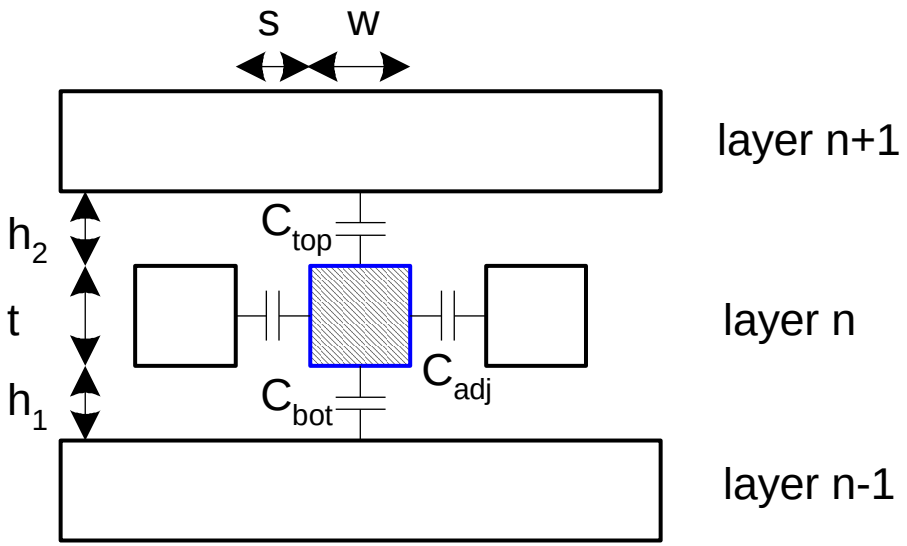
- Copper wires diffusion barrier has high resistance
- Copper is also prone to *dishing* polishing
- Effective resistance is higher

$$R = \frac{\rho}{\left(t - t_{\text{dish}} - t_{\text{barrier}}\right)} \frac{l}{\left(w - 2t_{\text{barrier}}\right)}$$



Wire Capacitance

- Wire has capacitance per unit length
  - To neighbors
  - To layers above and below
- $C_{total} = C_{top} + C_{bot} + 2C_{adj}$

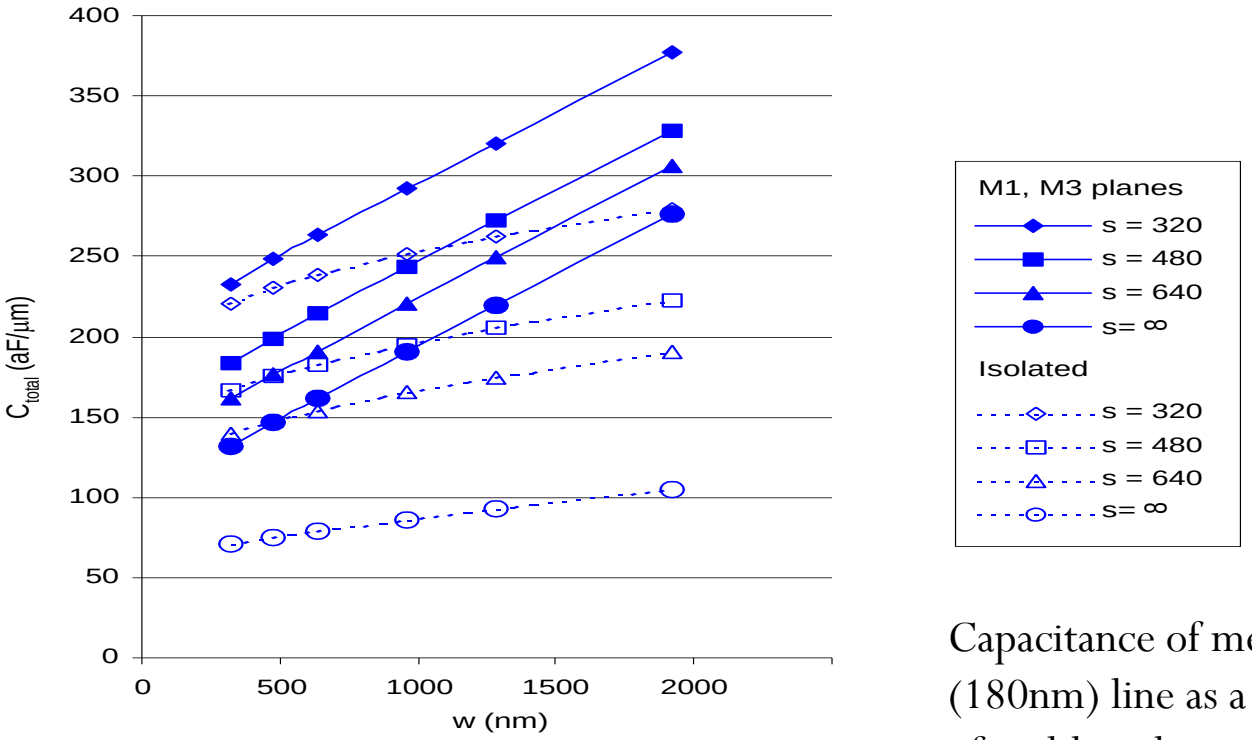


Capacitance Trends

- Parallel plate equation: $C = \varepsilon_{ox}A/d$
  - Wires are not parallel plates, but obey trends
  - Increasing area (w, t) increases capacitance
  - Increasing distance (s, h) decreases capacitance
- Dielectric constant
  - $\varepsilon_{ox} = k\varepsilon_0$
    - $\varepsilon_0 = 8.85 \times 10^{-14}$ F/cm
    - k = 3.9 for SiO_2
- Processes are starting to use low-k dielectrics
  - $k \approx 3$ (or less) as dielectrics use air pockets

M2 Capacitance Data

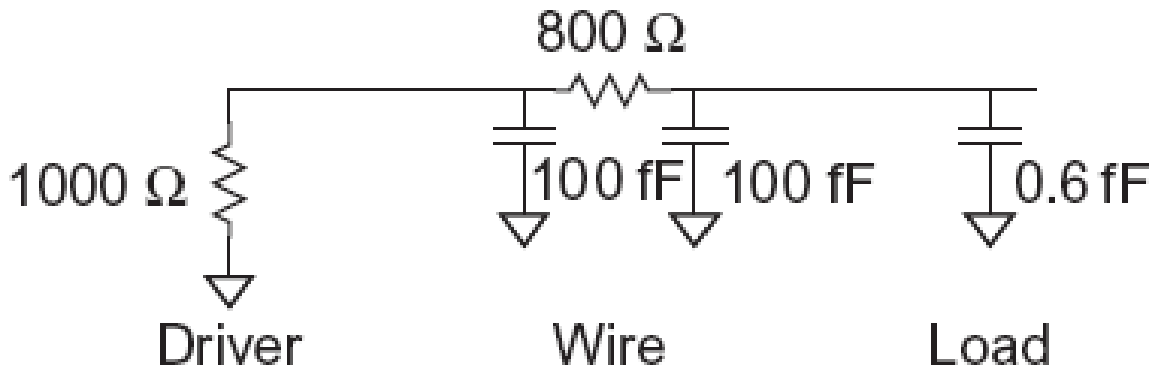
- Typical dense wires have ~ 0.2 fF/μm
  - *Compare to 1-2 fF/μm for gate capacitance*



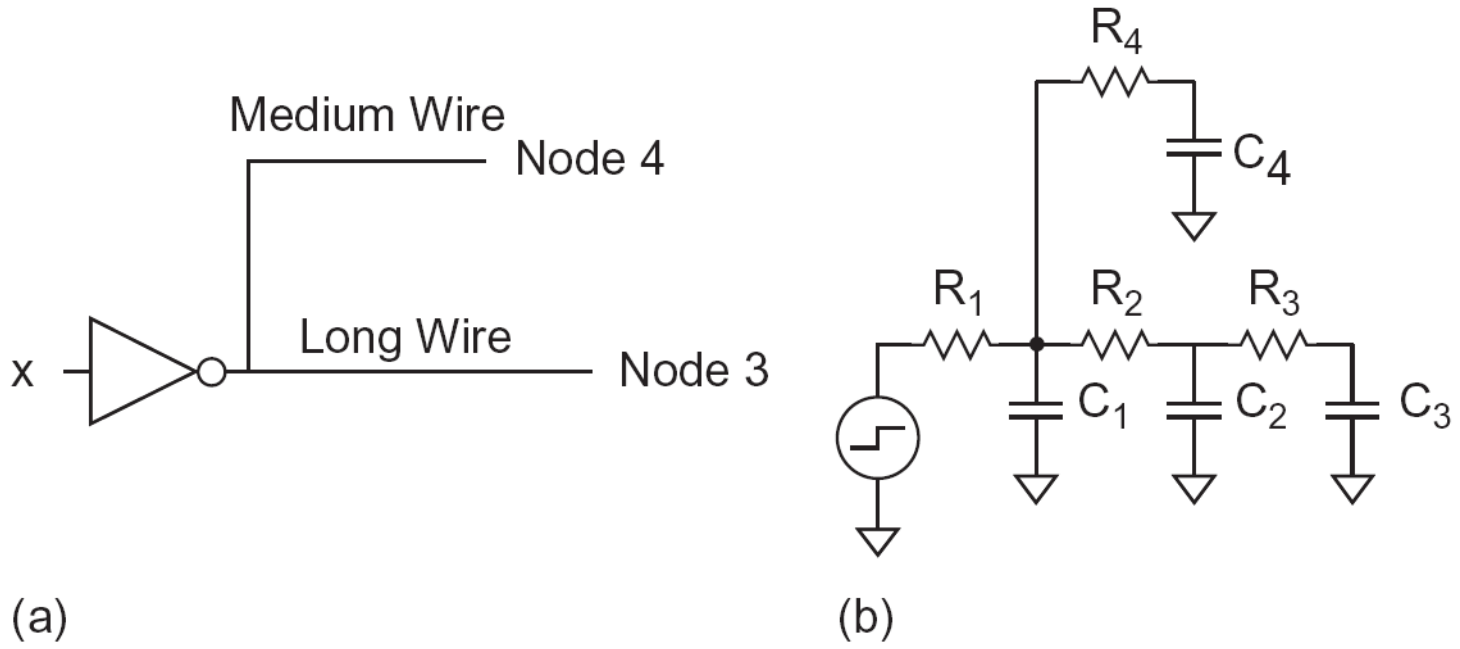
Capacitance of metal2 (180nm) line as a function of width and spacing

Wire RC Delay

- Estimate the delay of a 10x inverter driving a 2x inverter at the end of the 1 mm wire. Assume wire capacitance is 0.2 fF/μm and that a unit-sized nMOS transistor has R = 10 KΩ and C = 0.1 fF.
  - t_{pd} =



Interconnect Modeling with RC tree (Elmore Delay)



$$T_{D_3} = R_1C_1 + (R_1 + R_2)C_2 + (R_1 + R_2 + R_3)C_3 + R_1C_4$$

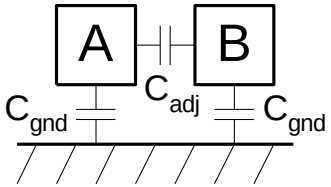
$$T_{D_4} = R_1C_1 + R_1(C_2 + C_3) + (R_1 + R_4)C_4$$

Crosstalk

- A capacitor does not like to change its voltage instantaneously.
- A wire has high capacitance to its neighbor.
  - When the neighbor switches from 1-> 0 or 0->1, the wire tends to switch too.
  - Called capacitive *coupling* or *crosstalk*.
- Crosstalk effects
  - Noise on nonswitching wires
  - Increased delay on switching wires

Crosstalk Delay

- Assume layers above and below on average are quiet
  - Second terminal of capacitor can be ignored
  - Model as $C_{gnd} = C_{top} + C_{bot}$
- A Switches: Effective C_{adj} depends on behavior of neighbors
  - *Miller effect*

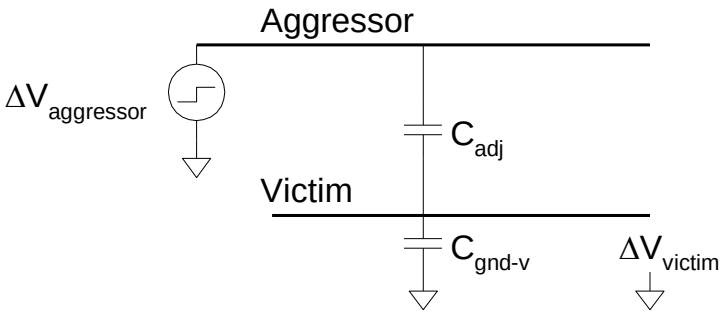


B	ΔV	$C_{eff(A)}$	MCF
Constant			
Switching with A			
Switching opposite A			

Crosstalk Noise

- Crosstalk causes noise on non-switching wires
- If victim is floating:
  - model as capacitive voltage divider

$$\Delta V_{victim} = \frac{C_{adj}}{C_{gnd-v} + C_{adj}} \Delta V_{aggressor}$$

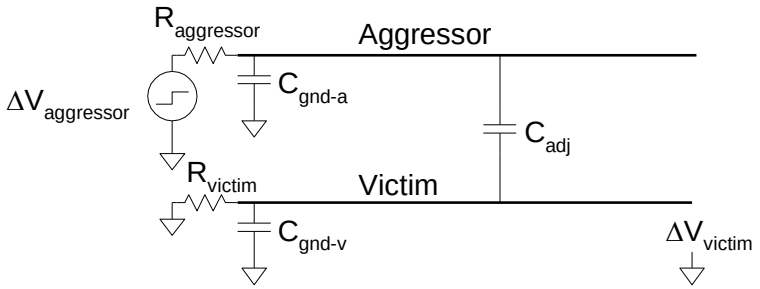


Driven Victims

- Usually victim is driven by a gate that fights noise
  - Noise depends on relative resistances
  - Victim driver is in linear region, agg. in saturation
  - If sizes are same, $R_{aggressor}$ = 2-4 x R_{victim}

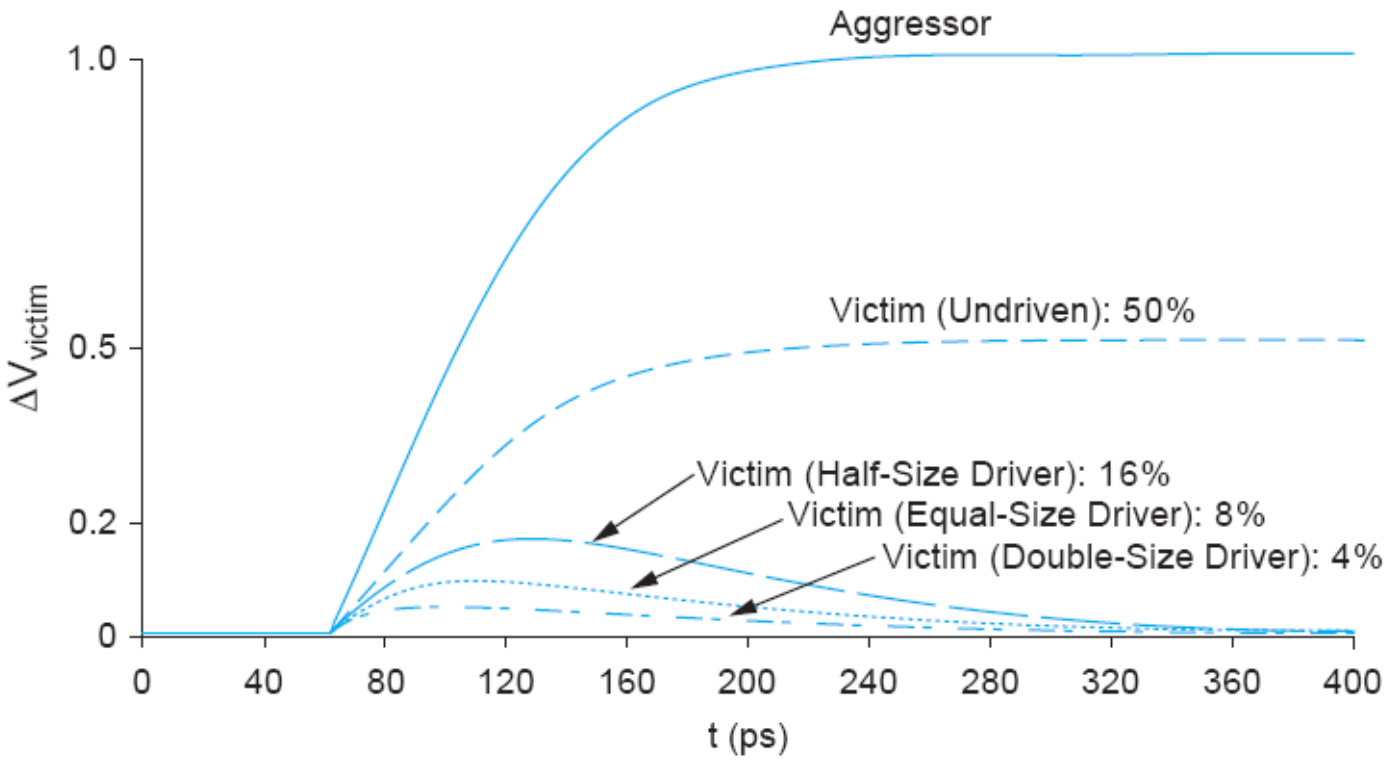
$$\Delta V_{victim} = \frac{C_{adj}}{C_{gnd-v} + C_{adj}} \frac{1}{1+k} \Delta V_{aggressor}$$

$$k = \frac{\tau_{aggressor}}{\tau_{victim}} = \frac{R_{aggressor}\left(C_{gnd-a} + C_{adj}\right)}{R_{victim}\left(C_{gnd-v} + C_{adj}\right)}$$



Coupling Waveforms

- Simulated coupling for $C_{adj} = C_{victim}$

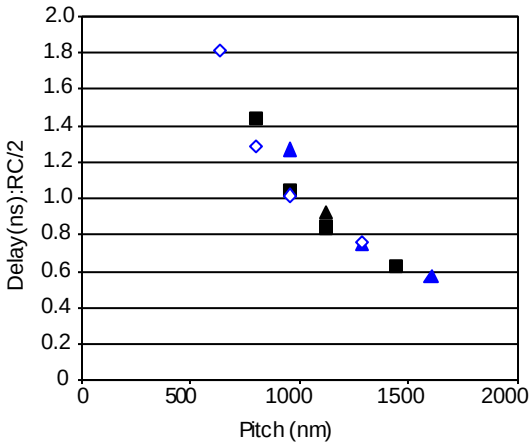


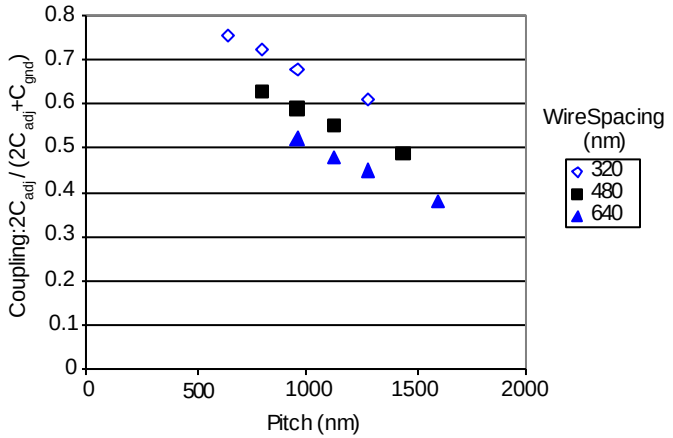
Noise Implications

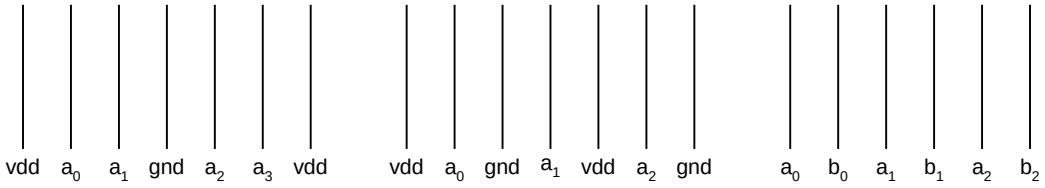
- *So what* if we have noise?
- If the noise is less than the noise margin, nothing happens
- Static CMOS logic will eventually settle to correct output even if disturbed by large noise spikes
  - But glitches cause extra delay
  - Also cause extra power from false transitions
- Memories and other sensitive circuits also can produce the wrong answer

Wire Engineering

- Goal: achieve delay, area, power goals with acceptable noise
- Degrees of freedom:
  - Width
  - Spacing
  - Layer
  - Shielding

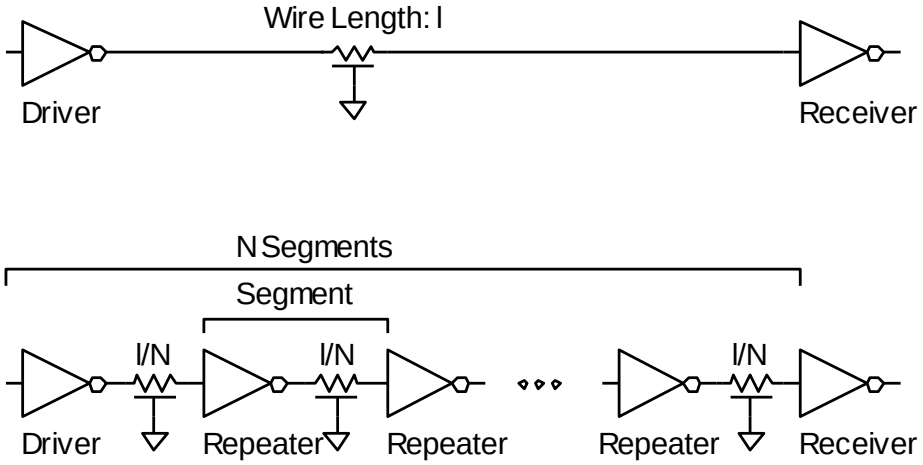






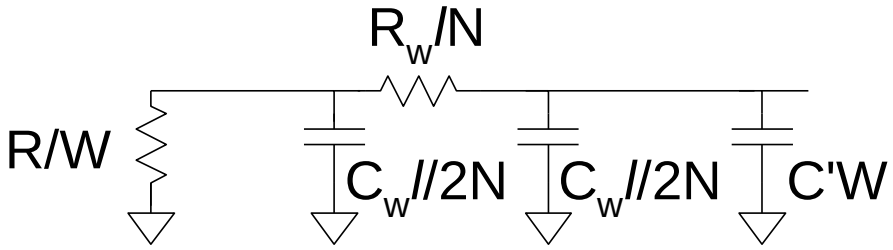
Repeaters

- R and C are proportional to *l*
- RC delay is proportional to l^2
  - Unacceptably great for long wires
- Break long wires into N shorter segments
  - Drive each one with an inverter or buffer



Repeater Design

- How many repeaters should we use?
- How large should each one be?
- Equivalent Circuit
  - Wire length *l/N*
    - Wire Capacitance C_w*l/N, Resistance R_w*l/N
  - Inverter width W (nMOS = W, pMOS = 2W)
    - Gate Capacitance C'*W, Resistance R/W



Repeater Results

- Write equation for Elmore Delay

$$t_{pd} = N\left[\frac{R}{W}\left(C_w\frac{l}{N} + CW\left(1+p_{\text{inv}}\right)\right) + R_w\frac{l}{N}\left(\frac{C_w}{2}\frac{l}{N} + CW\right)\right]$$

  - Differentiate with respect to W and N
  - Set equal to 0, solve

$$\frac{l}{N} = \sqrt{\frac{2RC'}{R_wC_w}}$$

$$\frac{t_{pd}}{l} = \left(2+\sqrt{2}\right)\sqrt{RC'R_wC_w}$$

~40 ps/mm

in 65 nm process

$$W = \sqrt{\frac{RC_w}{R_wC'}}$$

Repeater Energy

- Energy / length ≈ $1.87C_wV_{DD}^2$
  - 87% premium over unrepeated wires
  - The extra power is consumed in the large repeaters
- If the repeaters are downsized for minimum EDP:
  - Energy premium is only 30%
  - Delay increases by 14% from min delay

Reference

- Reference [4]. Chapter 4.3, Chapter 6