CHAPTER 4

Circuit Characterization and Performance Estimation II

Outline

1. Delay Estimation
2. Logical Effort and Transistor Sizing
3. **Power Dissipation**
4. Interconnect
5. Wire Engineering
6. Design Margin
7. Reliability
8. Scaling

Power and Energy

- Power is drawn from a voltage source attached to the V_{DD} pin(s) of a chip.

- Instantaneous Power: $P(t) = i_{DD}(t)V_{DD}$

- Energy: $E = \int_0^T P(t)dt = \int_0^T i_{DD}(t)V_{DD}dt$

- Average Power: $P_{\text{avg}} = \frac{E}{T} = \frac{1}{T}\int_0^T i_{DD}(t)V_{DD}dt$

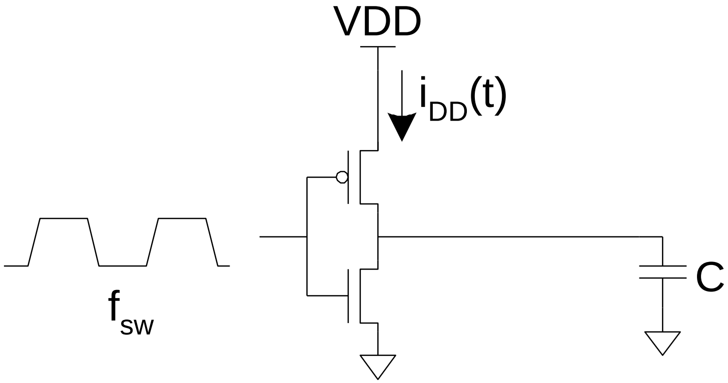
Static and Dynamic Dissipation

$$P_{\text{total}} = P_{\text{static}} + P_{\text{dynamic}}$$

- Static dissipation
  - Subthreshold conduction through OFF transistors
  - Tunneling current through gate oxide
  - Leakage through reverse-biased diodes
  - Contention current in ratioed circuits
- Dynamic dissipation
  - Charging and discharging of load capacitance
  - "Short-circuit" current while both pMOS and nMOS networks are partially ON

Dynamic Power

- Dynamic power is required to charge and discharge load capacitances when transistors switch.
- One cycle involves a rising and falling output.
- On rising output, charge $Q = CV_{DD}$ is required
- On falling output, charge is dumped to GND
- This repeats Tf_{sw} times over an interval of T



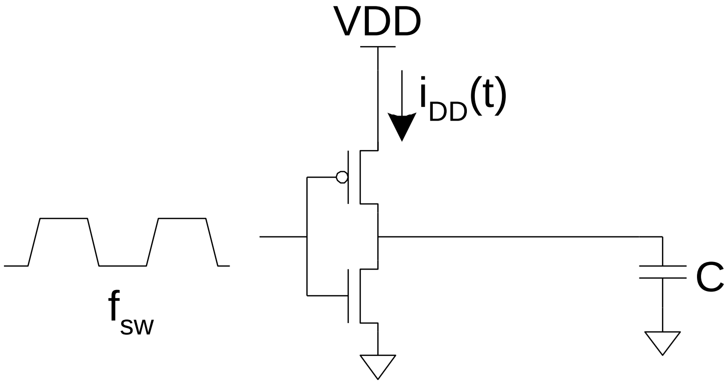
Dynamic Power Cont.

$$P_{\text{dynamic}} = \frac{1}{T}\int_0^T i_{DD}(t)V_{DD}dt$$

$$= \frac{V_{DD}}{T}\int_0^T i_{DD}(t)dt$$

$$= \frac{V_{DD}}{T}\left[Tf_{\text{sw}}CV_{DD}\right]$$

$$= CV_{DD}{}^2 f_{\text{sw}}$$



Activity Factor

- Suppose the system clock frequency = f
- Let $f_{sw} = \alpha f$, where α = activity factor
  - If the signal is a clock, $\alpha = 1$
  - If the signal switches once per cycle, $\alpha = ½$
  - Dynamic gates:
    - Switch either 0 or 2 times per cycle, $\alpha = ½$
  - Static gates:
    - Depends on design, but typically $\alpha = 0.1$
- Dynamic power:

$$P_{\text{dynamic}} = \alpha C V_{DD}^{\;2} f$$

Short Circuit Current

- When transistors switch, both nMOS and pMOS networks may be momentarily ON at once
- Leads to a blip of “short circuit” current.
- < 10% of dynamic power if rise/fall times are comparable for input and output

Example

- 200M transistor chip
  - 20M logic transistors
    - Average width: 12 λ
  - 180M memory transistors
    - Average width: 4 λ
  - 1.2 V 100 nm process (λ = 0.5* feature size = 50nm)
  - C_g = 2 fF/μm

Dynamic Example

- Static CMOS logic gates: activity factor = 0.1
- Memory arrays: activity factor = 0.05 (many banks and partially activated at a time!)
- Estimate dynamic power consumption per MHz.
  - Neglect wire capacitance and short-circuit current.

Dynamic Example

- Static CMOS logic gates: activity factor = 0.1
- Memory arrays: activity factor = 0.05 (many banks and partially activated at a time!)
- Estimate dynamic power consumption per MHz.
  - Neglect wire capacitance.

$$C_{\text{logic}} = \left(20\times10^{6}\right)\left(12\lambda\right)\left(0.05\mu m/\lambda\right)\left(2\,fF/\mu m\right) = 24nF$$

$$C_{\text{mem}} = \left(180\times10^{6}\right)\left(4\lambda\right)\left(0.05\mu m/\lambda\right)\left(2\,fF/\mu m\right) = 72nF$$

$$P_{\text{dynamic}} = \left[0.1C_{\text{logic}} + 0.05C_{\text{mem}}\right]\left(1.2\right)^{2}\ f = 8.6\ \text{mW/MHz}$$

$$= 8.6\ \ \text{W}\,@1\ \text{GHz}$$

Static Power

- Static power is consumed even when chip is quiescent.
  - Ratioed circuits burn power in fight between ON transistors
  - Leakage draws power from nominally OFF devices

$$I_{ds} = I_{ds0} e^{\frac{V_{gs} - V_t}{n v_T}} \left[1 - e^{\frac{-V_{ds}}{v_T}} \right]$$

$$V_t = V_{t0} - \eta V_{ds} + \gamma \left(\sqrt{\phi_s + V_{sb}} - \sqrt{\phi_s} \right)$$

Ratio Example

- The chip contains a 32 word x 48 bit ROM
  - Uses 1:32 pseudo-nMOS decoder and bitline pullups
  - On average, one wordline and 24 bitlines are high
- Find static power drawn by the ROM
  - β = 75 μA/V^2, V_{DD} = 1.8V
  - V_{tp} = -0.4V
- Solution:

$$I_{\text{pull-up}} = \beta \frac{\left(V_{DD} - \left|V_{tp}\right|\right)^2}{2} = 73\mu\text{A}$$

$$P_{\text{pull-up}} = V_{DD} I_{\text{pull-up}} = 130\mu\text{W}$$

$$P_{\text{static}} = (31+24) P_{\text{pull-up}} = 7.2\text{ mW}$$

Leakage Example

- The process has two threshold voltages and two oxide thicknesses.
- Subthreshold leakage:
  - 20 nA/μm for low V_t
  - 0.02 nA/μm for high V_t
- Gate leakage:
  - 3 nA/μm for thin oxide
  - 0.002 nA/μm for thick oxide
- Memories use low-leakage transistors everywhere
- Gates use low-leakage transistors on 80% of logic

Leakage Example Cont.

- Estimate static power:
  - High leakage: $\left(20\times10^{6}\right)\left(0.2\right)\left(12\lambda\right)\left(0.05\mu m/\lambda\right)=2.4\times10^{6}\,\mu m$
  - Low leakage: $\left(20\times10^{6}\right)\left(0.8\right)\left(12\lambda\right)\left(0.05\mu m/\lambda\right)+$
    $\left(180\times10^{6}\right)\left(4\lambda\right)\left(0.05\mu m/\lambda\right)=45.6\times10^{6}\,\mu m$

$$
\begin{aligned}
I_{static} &= \left(2.4\times10^{6}\,\mu m\right)\left[\left(20nA/\mu m\right)/2+\left(3nA/\mu m\right)\right]+\\
&\quad \left(45.6\times10^{6}\,\mu m\right)\left[\left(0.02nA/\mu m\right)/2+\left(0.002nA/\mu m\right)\right]\\
&= 32mA\\
P_{static} &= I_{static}V_{DD}=38mW
\end{aligned}
$$

- If no low leakage devices, P_{static} = 749 mW (!)

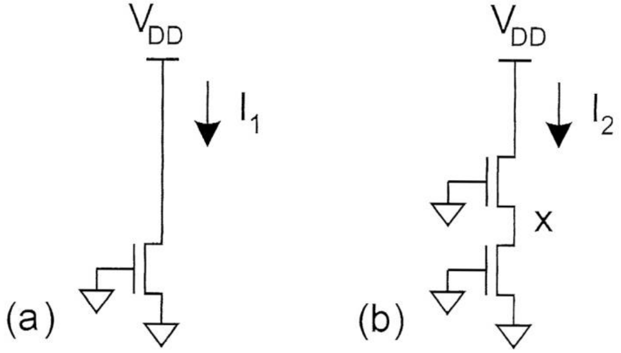
Low Power Design

- Reduce dynamic power
  - α: clock gating, sleep mode
  - C: small transistors (esp. on clock), short wires
  - V_{DD}: lowest suitable voltage
  - f: lowest suitable frequency
- Reduce static power
  - Selectively use ratioed circuits
  - Selectively use low V_t devices
  - Leakage reduction:

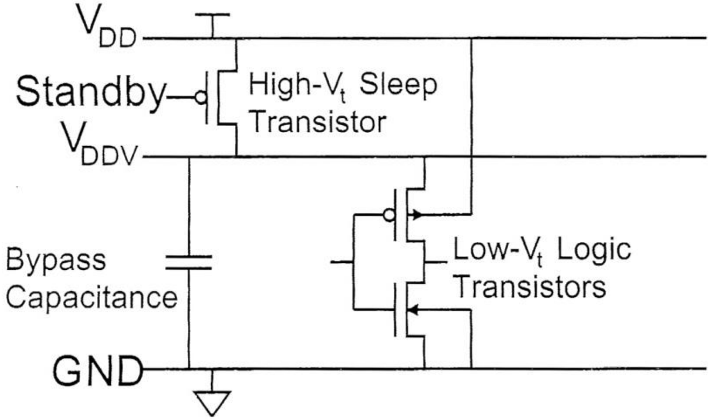
    stacked devices, body bias, low temperature

Reduce Static Power

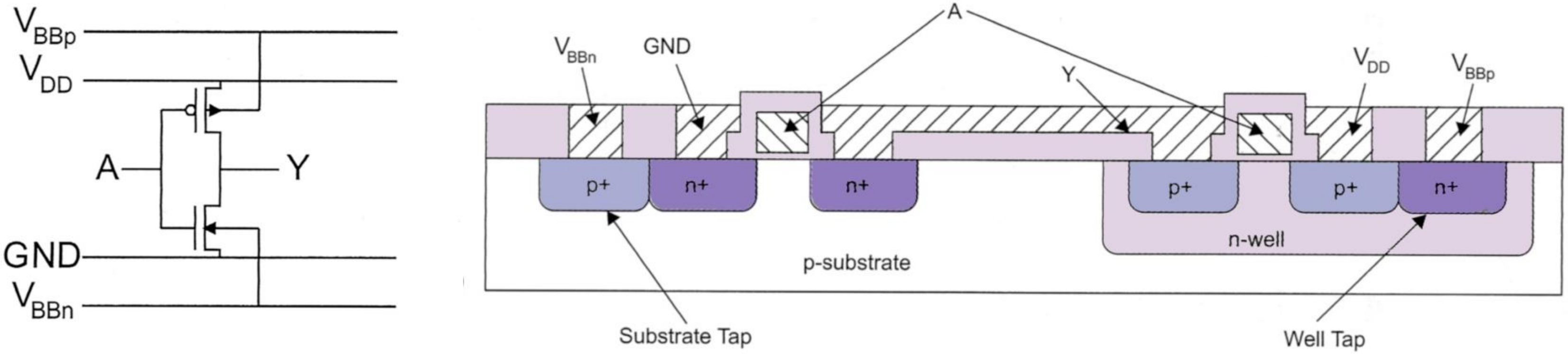
- Leakage stack effect



- MTCMOS : Multiple Threshold CMOS



- Body bias



Outline

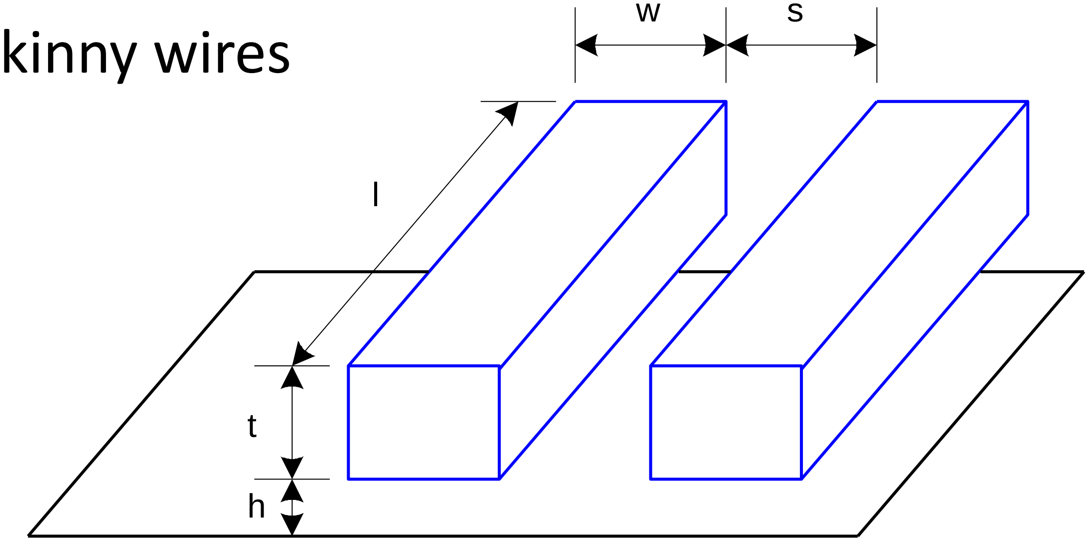
1. Delay Estimation
2. Logical Effort and Transistor Sizing
3. Power Dissipation
4. **Interconnect**
5. Wire Engineering
6. Design Margin
7. Reliability
8. Scaling

Interconnect

- Chips are mostly made of wires called *interconnect*
  - In stick diagram, wires set size
  - 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

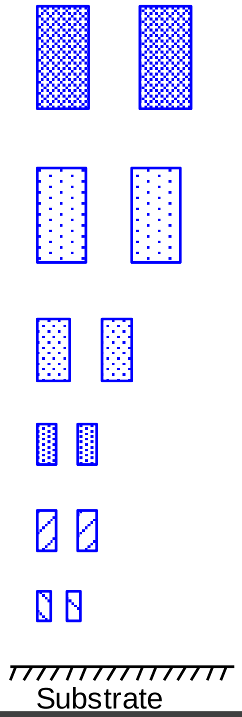
- Pitch = w + s
- Aspect ratio: AR = t/w
  - Old processes had AR << 1
  - Modern processes have AR ≈ 2
    - Pack in many skinny wires



Layer Stack

- AMI 0.6 μm process has 3 metal layers
- Modern processes use 6-10+ metal layers
- Example: Intel 180 nm process
- M1: thin, narrow (< 3λ)
  - High density cells
- M2-M4: thicker
  - For longer wires
- M5-M6: thickest
  - For V_{DD}, GND, clk

Layer	**T** (nm)	**W** (nm)	**S** (nm)	**AR**
6	1720	860	860	2.0
	1000			
5	1600	800	800	2.0
	1000			
4	1080	540	540	2.0
	700			
3	700	320	320	2.2
	700			
2	700	320	320	2.2
	700			
1	480	250	250	1.9
	800			

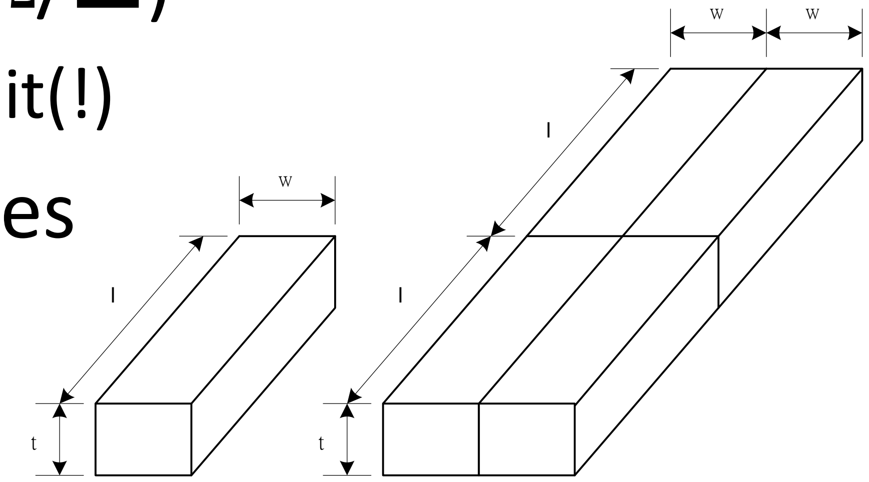


Wire Resistance

- ρ = *resistivity* (Ω*m)

$$R = \frac{\rho}{t}\frac{l}{w} = R_{\square}\frac{l}{w}$$

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



1 Rectangular Block
R = R$_{\square}$(L/W) Ω

4 Rectangular Blocks
R = R$_{\square}$(2L/2W) Ω
= R$_{\square}$(L/W) Ω

Choice of Metals

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

Metal	**Bulk resistivity (μΩ*cm)**
Silver (Ag)	1.6
Copper (Cu)	1.7
Gold (Au)	2.2
Aluminum (Al)	2.8
Tungsten (W)	5.3
Molybdenum (Mo)	5.3

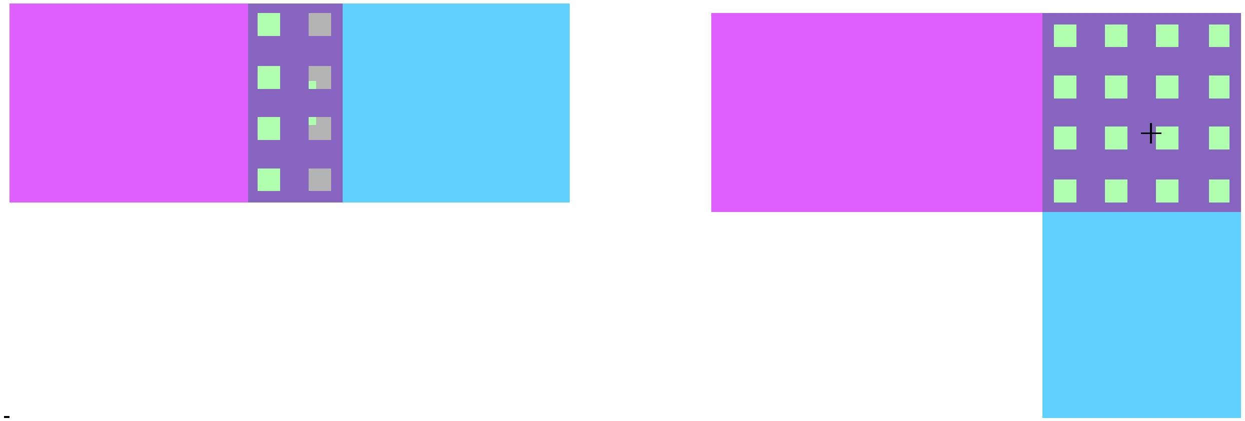
Sheet Resistance

- Typical sheet resistances in 180 nm process

Layer	Sheet Resistance (Ω/□)
Diffusion (silicided)	3-10
Diffusion (no silicide)	50-200
Polysilicon (silicided)	3-10
Polysilicon (no silicide)	50-400
Metal1	0.08
Metal2	0.05
Metal3	0.05
Metal4	0.03
Metal5	0.02
Metal6	0.02

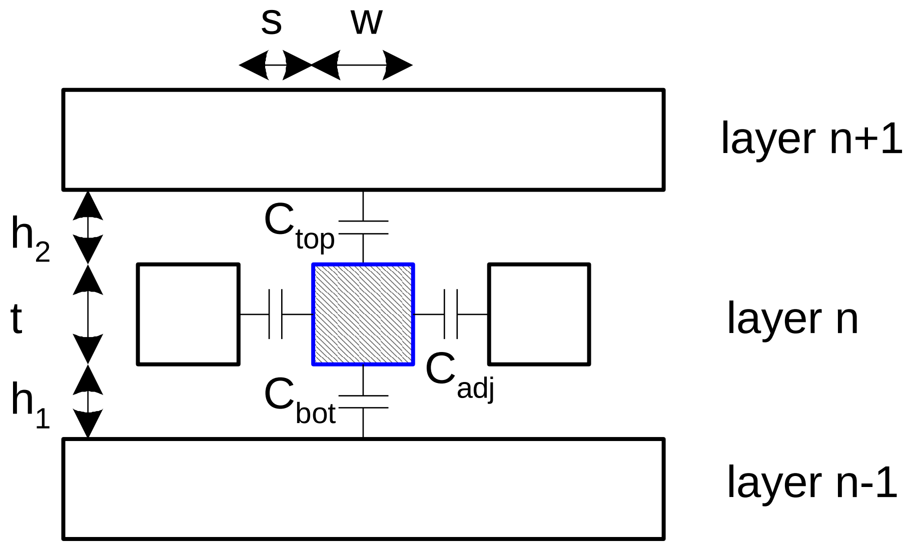
Contacts Resistance

- Contacts and vias also have 2-20 Ω
- Use many contacts for lower R
  - Many small contacts for current crowding around periphery



Wire Capacitance

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

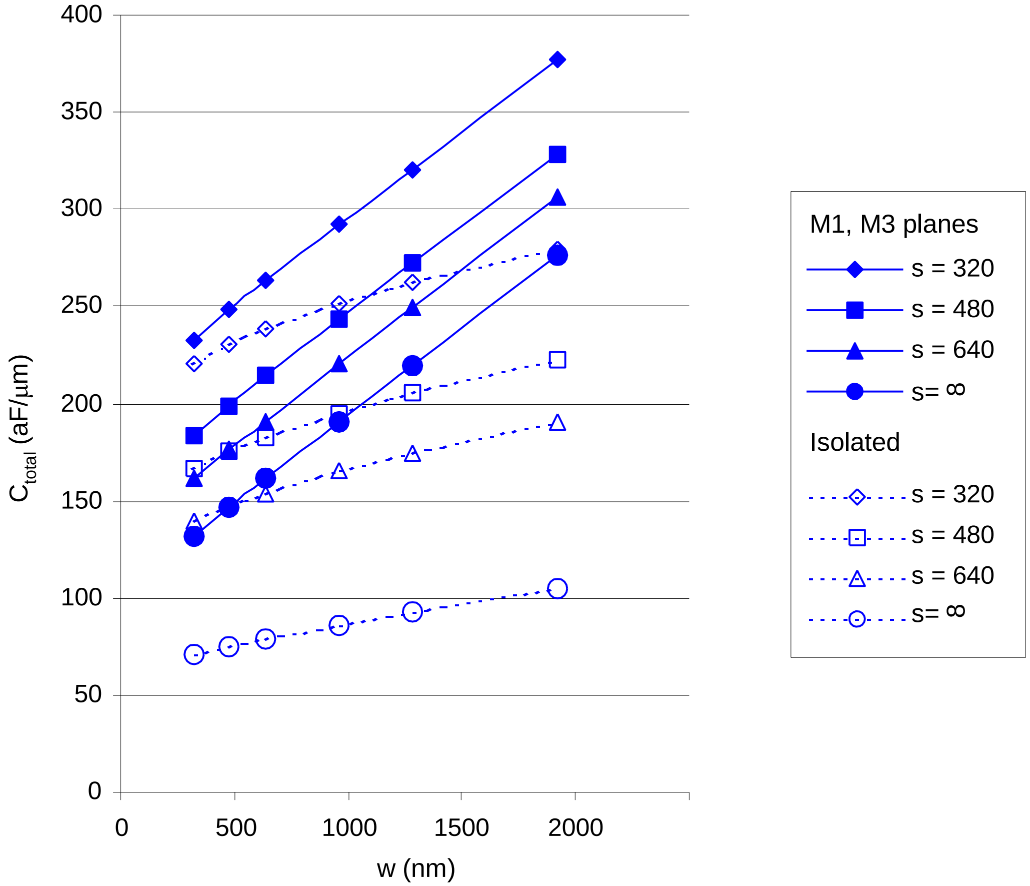


Capacitance Trends

- Parallel plate equation: $C = \varepsilon A/d$
  - Wires are not parallel plates, but obey trends
  - Increasing area (W, t) increases capacitance
  - Increasing distance (s, h) decreases capacitance
- Dielectric constant
  - $\varepsilon = k\varepsilon_0$
  - $\varepsilon_0 = 8.85 \times 10^{-14}$ 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 wires have ~ 0.2 fF/μm
  - Compare to 2 fF/μm for gate capacitance

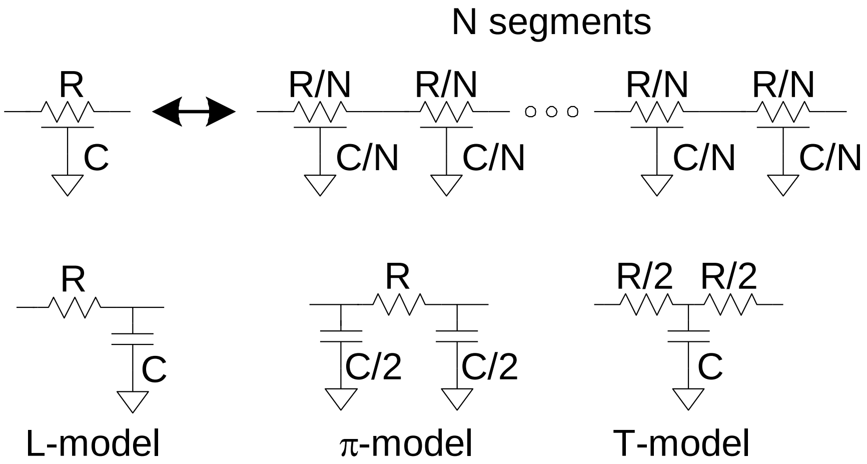


Diffusion & Polysilicon

- Diffusion capacitance is very high (about 2 fF/μm)
  - Comparable to gate capacitance
  - Diffusion also has high resistance
  - Avoid using diffusion *runners* for wires!
- Polysilicon has lower C but high R
  - Use for transistor gates
  - Occasionally for very short wires between gates

Lumped Element Models

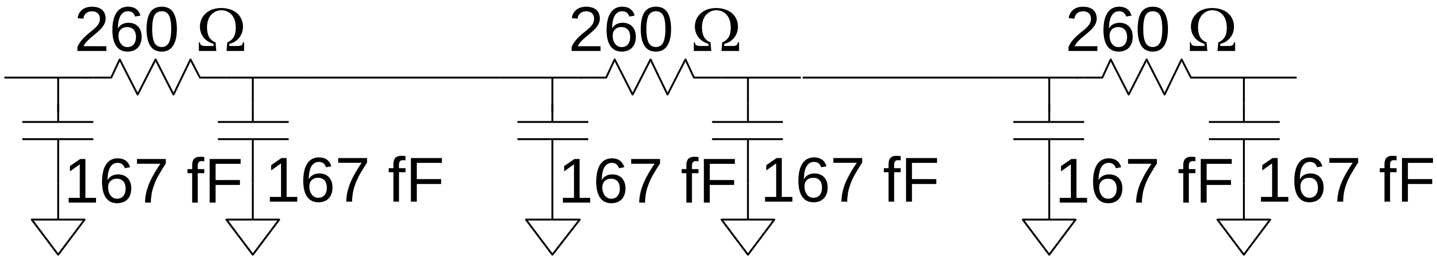
- Wires are a distributed system
  - Approximate with lumped element models



- 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

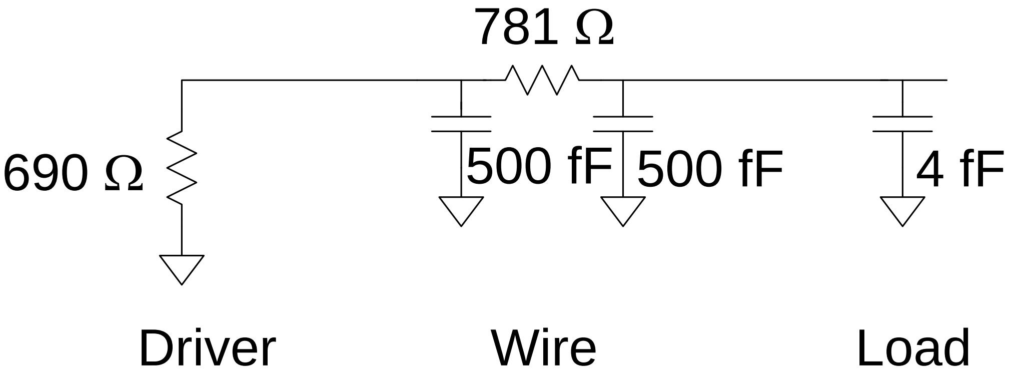
Example

- Metal2 wire in 180 nm process
  - 5 mm long
  - 0.32 μm wide
- Construct a 3-segment π-model
  - $R_\square = 0.05\ \Omega/\square$ => $R = 781\ \Omega$
  - $C_{permicron} = 0.2$ fF/μm => C = 1 pF



Wire RC Delay

- Estimate the delay of a 10x inverter driving a 2x inverter at the end of the 5mm wire from the previous example.
  - Effective R = 2.5 kΩ/μm for gates, C = 2 fF/μm
  - Unit inverter: 4λ = 0.36 μm nMOS, 8λ = 0.72 μm pMOS
    - R(10x) = 2.5kΩ/(0.36x10)=690, C(2x) = (0.36+0.72)x2=2 fF.
    - t_{pd} = (690Ω)*(500fF)+(690Ω+781Ω)*(5000fF+4fF)=1.1 ns.



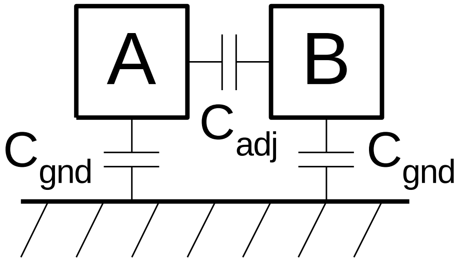
Crosstalk

- 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 $C_{gnd} = C_{top} + C_{bot}$
- Effective C_{adj} depends on behavior of neighbors
  - *Miller Coupling Factor (MCF)*

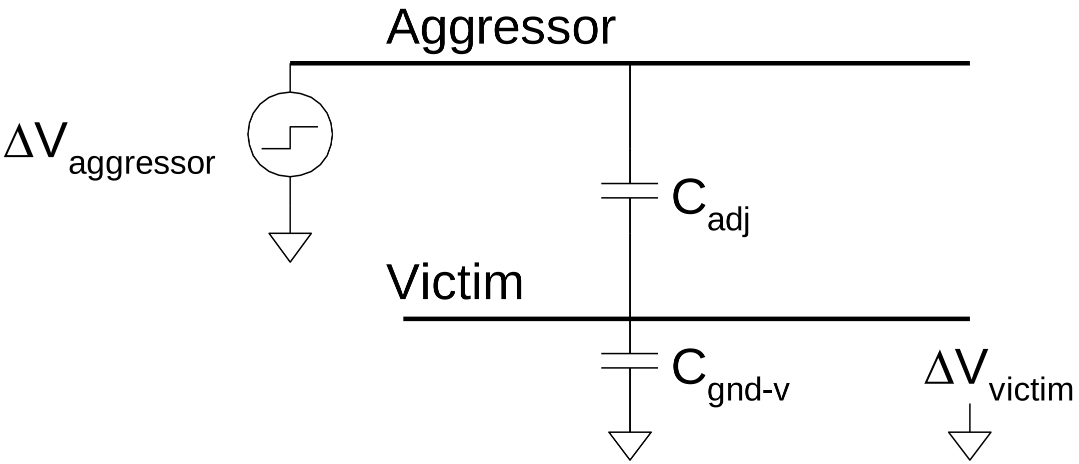
B	**ΔV**	**$C_{eff(A)}$**	**MCF**
Constant	V_{DD}	$C_{gnd} + C_{adj}$	1
Switching with A	0	C_{gnd}	0
Switching opposite A	$2V_{DD}$	$C_{gnd} + 2\ C_{adj}$	2



Crosstalk Noise

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

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

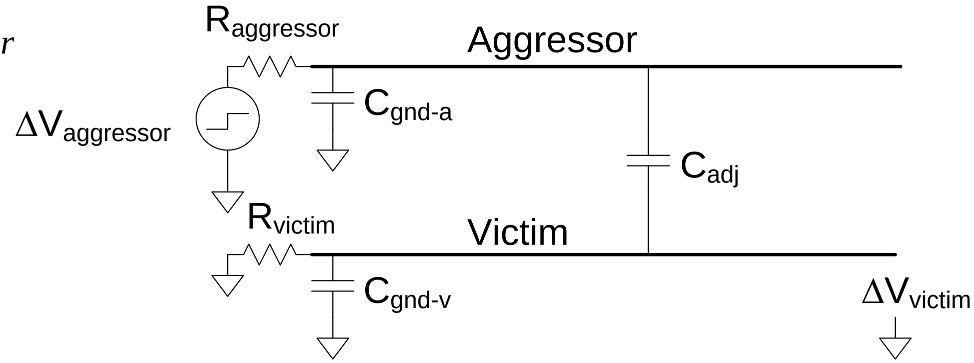


Driven Victims

- Usually victim is driven by a gate that fights noise
  - Noise depends on relative resistances
  - Victim driver is in linear region, and aggressor driver is in saturation. (p3-53)
  - If sizes are same, $R_{aggressor}$ = 2-4 x R_{victim}

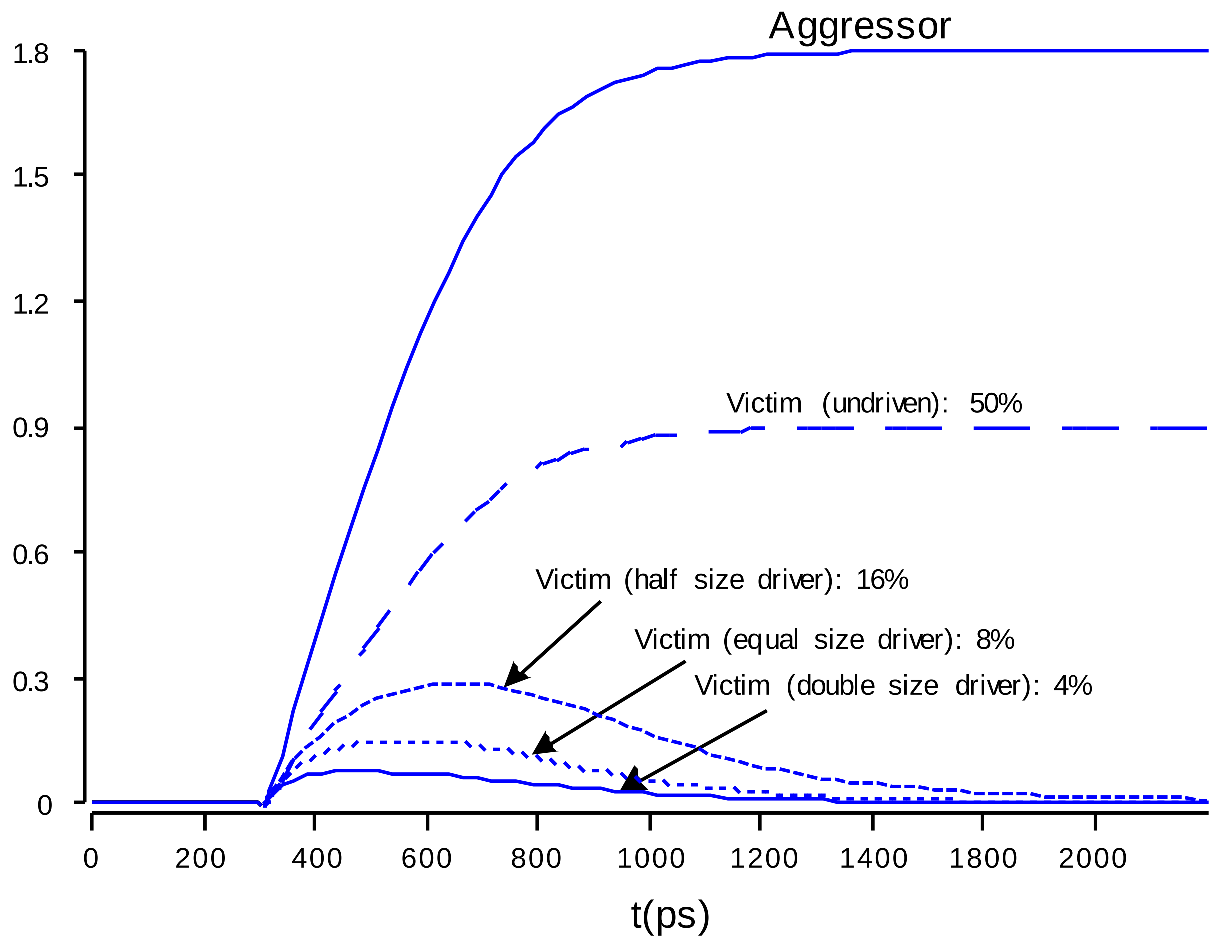
$$\Delta V_{victim} = \frac{C_{adj}}{C_{gnd-v} + C_{adj}} \frac{1}{1+k} \Delta V_{aggressor}$$

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



Coupling Waveforms

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



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
- Dynamic logic never recovers from glitches
- Memories and other sensitive circuits also can produce the wrong answer

Outline

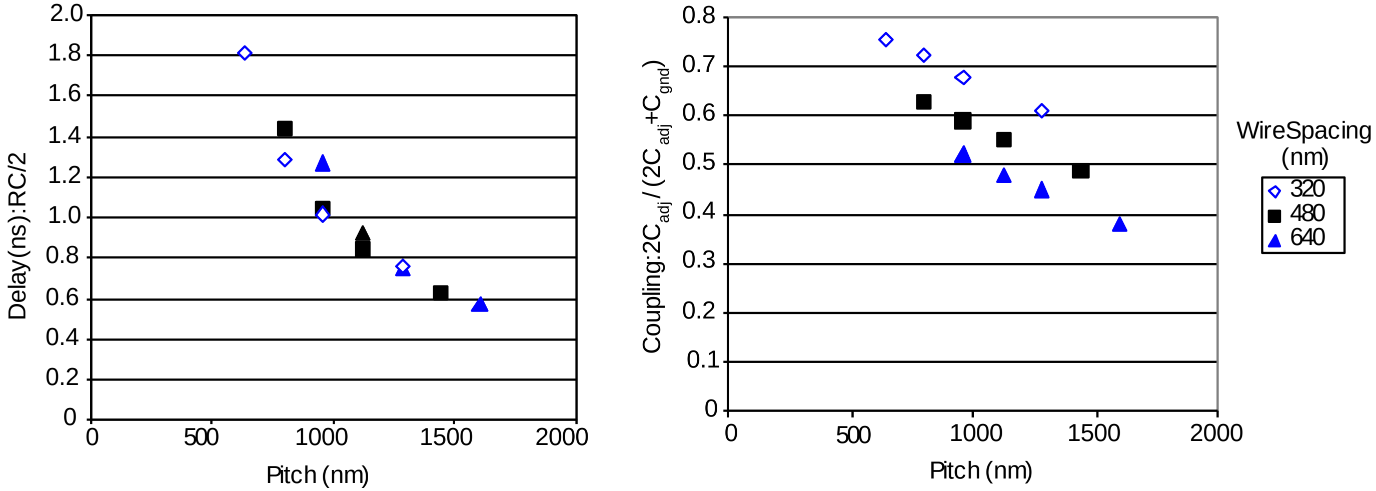
1. Delay Estimation
2. Logical Effort and Transistor Sizing
3. Power Dissipation
4. Interconnect
5. **Wire Engineering**
6. Design Margin
7. Reliability
8. Scaling

Wire Engineering

- Goal: achieve delay, area, power goals with acceptable noise
- Degrees of freedom:

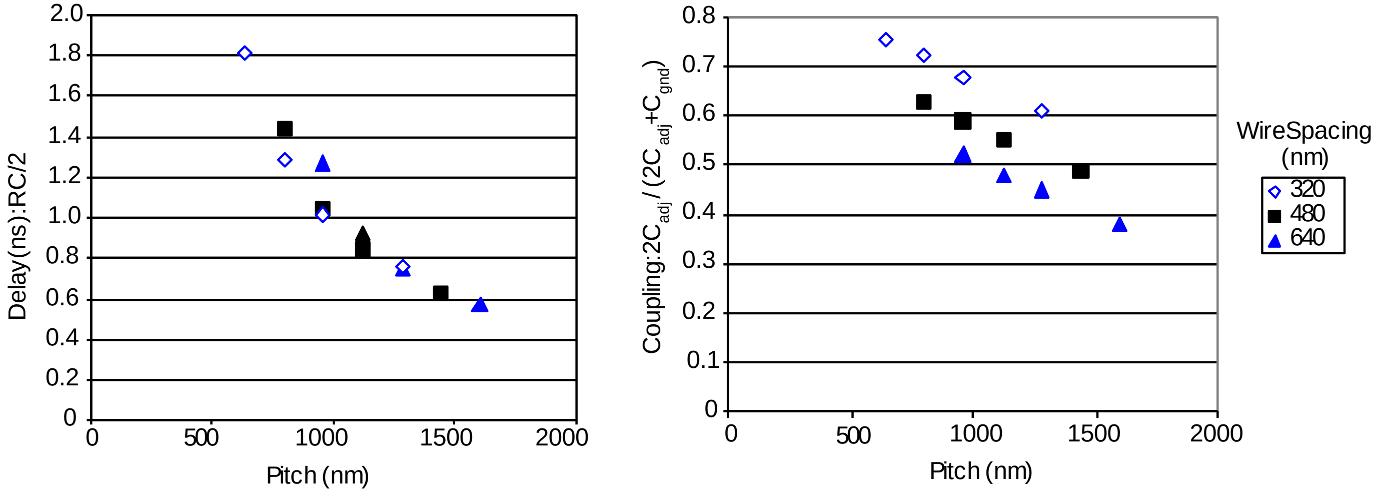
Wire Engineering

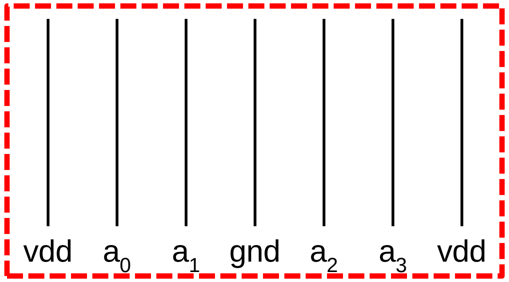
- Goal: achieve delay, area, power goals with acceptable noise
- Degrees of freedom:
  - Width
  - Spacing

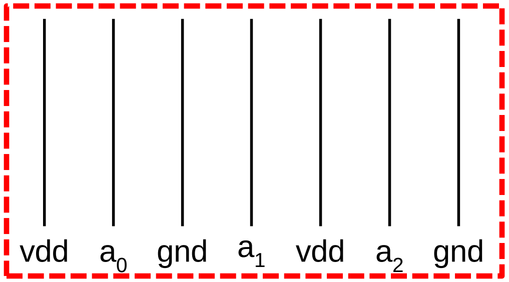


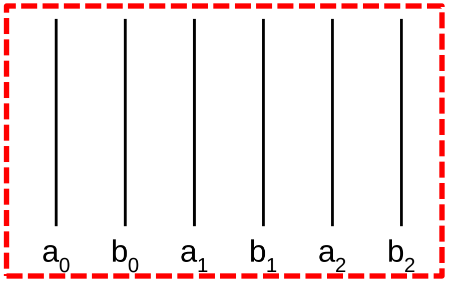
Wire Engineering

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



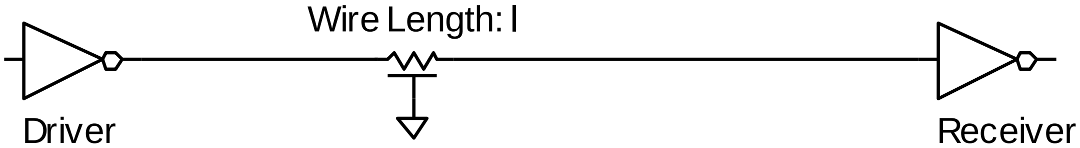


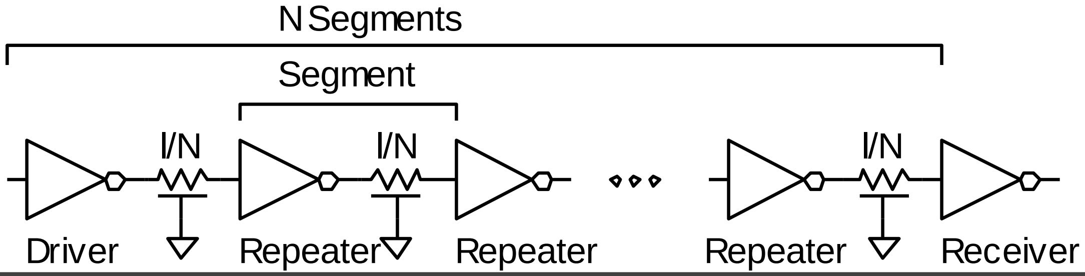




Repeaters

- R and C are proportional to *l (length)*
- 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



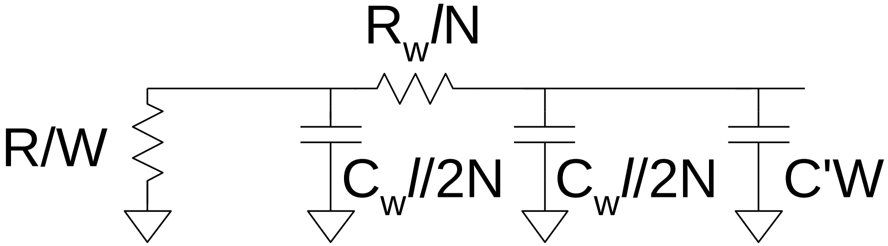


Repeater Design

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

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

$$t_{pd} = N\left[\frac{R}{W}\left(C_w\frac{l}{N} + C'W\right) + R_w\frac{l}{N}\left(\frac{C_w}{2}\frac{l}{N} + CW\right)\right]$$

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

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

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

~60-80 ps/mm

in 180 nm process

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