EE3230 Lecture 5: Circuit Characterization and Performance Estimation II

Ping-Hsuan Hsieh (謝秉璇)

Delta Building R908

EXT 42590

phsieh@ee.nthu.edu.tw

Outline

- Delay estimation
- Logical effort and transistor sizing
- **Power dissipation**
- Interconnect
- Wire engineering
- Design margin
- Reliability
- Scaling

Power and Energy

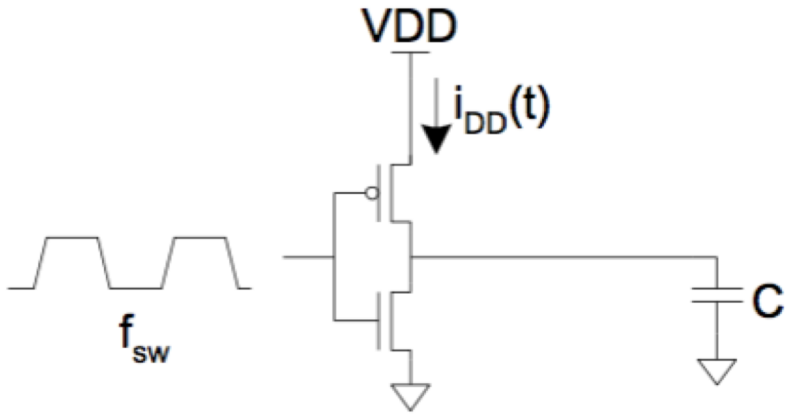
- Power is drawn from a voltage source attached to the V_{DD} pin(s) of a chip
- Instantaneous Power:
- Energy:
- Average Power:

Static and Dynamic Power Dissipation

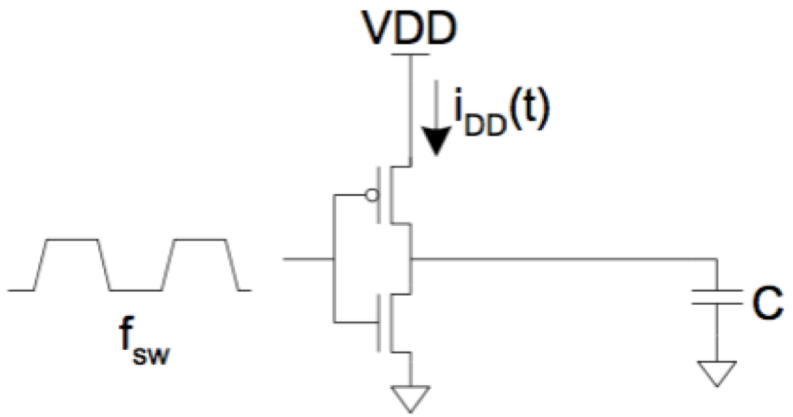
- Static power dissipation
  - Sub-threshold conduction through OFF transistors
  - Tunneling current through gate oxide
  - Leakage through reverse-biased diodes
  - Contention current in ratioed circuits
- Dynamic power dissipation
  - Charging and discharging of load capacitance
  - “Short-circuit” current while both PMOS and NMOS networks are partially ON

Dynamic Power (I)

- 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 f_{sw} times per second



Dynamic Power (II)



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 1 times per cycle, a = ½
  - **Static gates:** Depends on design, typically $\alpha = 0.1$

- Dynamic power:

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 (well-controlled)

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 *f*F/μm

Dynamic Power Consumption

- 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

Static Power Consumption

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

Ratioed Example

- The chip contains a 32 word x 48 bit ROM
  - Uses 1:32 pseudo-nMOS decoder and bit-line pull-ups
  - In average, one wordline and 24 bitlines are high
- Find static power drawn by the ROM
  - $\beta = 75\ \mu A/V^2$, $V_{DD} = 1.8\ V$
  - $V_{tp} = -0.4V$

Leakage Example (I)

- The process has two threshold voltages and two oxide thicknesses.
- Subthreshold leakage:
  - 20 nA/μm for low V_{th} devices
  - 0.02 nA/μm for high V_{th} devices
- 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 (II)

- Estimate static power:
  - High leakage:
  - Low leakage:

- Withow leakage devices, P_{static} = 749 mW **(!)**

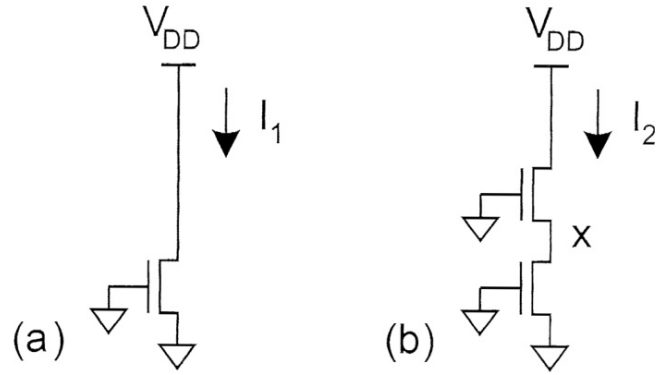
Low Power Design

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

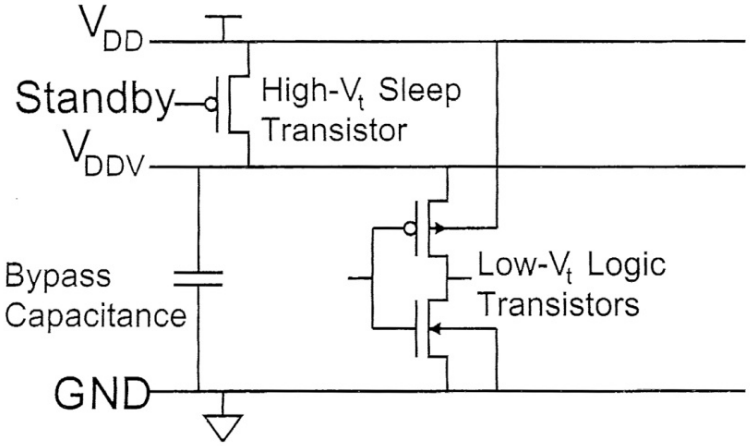
    stacked devices, body bias, low temperature

Reduce Static Power

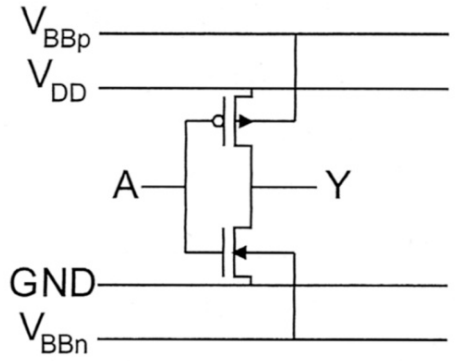
- Leakage stack effect

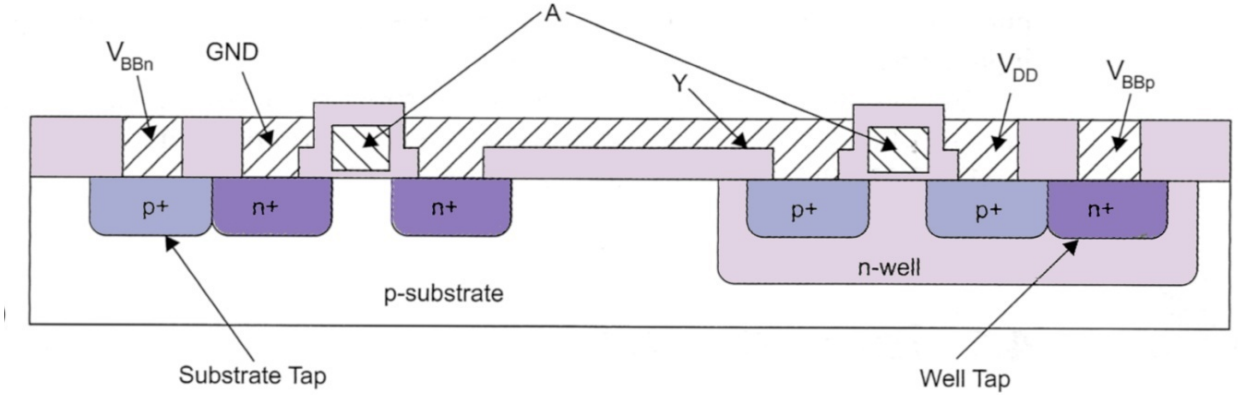


- MTCMOS: multiple threshold CMOS



- Body bias





Outline

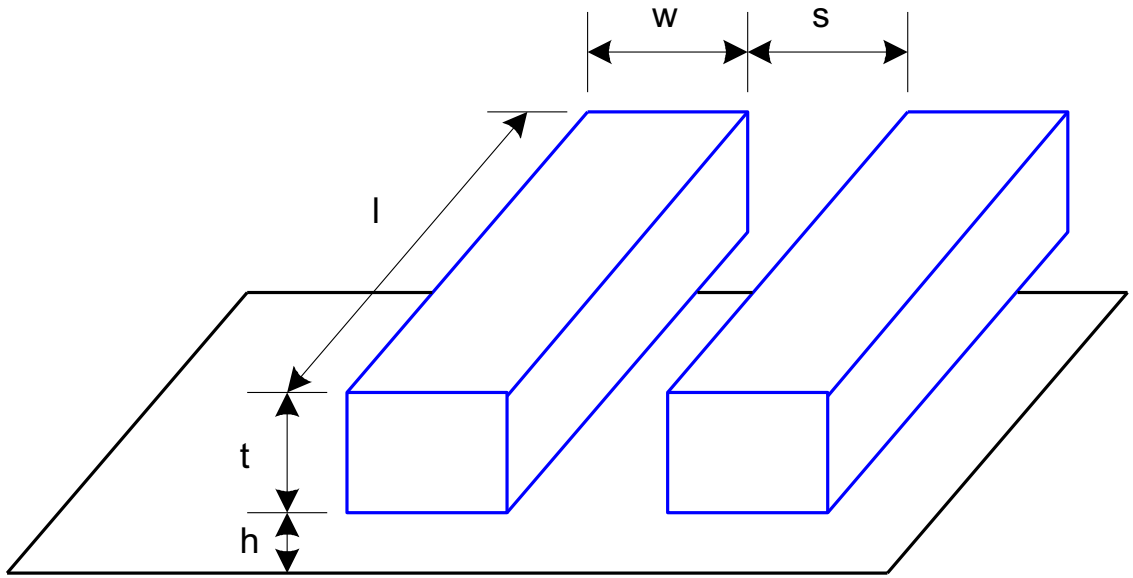
- Delay estimation
- Logical effort and transistor sizing
- Power dissipation
- **Interconnect**
- Wire engineering
- Design margin
- Reliability
- Scaling

Interconnect

- Chips are mostly made of wires called *interconnect*
  - In stick diagrams, wires determine 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

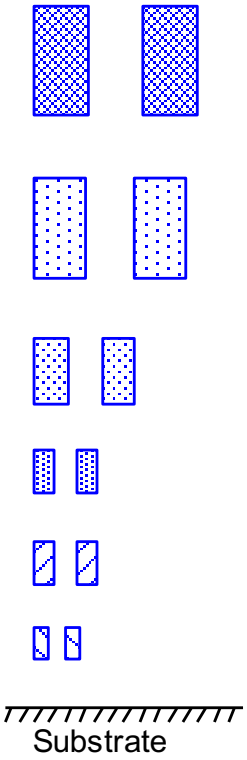
- Pitch = w + s
- Aspect ratio: AR = t/w
  - Old processes had AR << 1
  - Modern processes have AR ≈ 2
    - Pack in many skinny wires



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

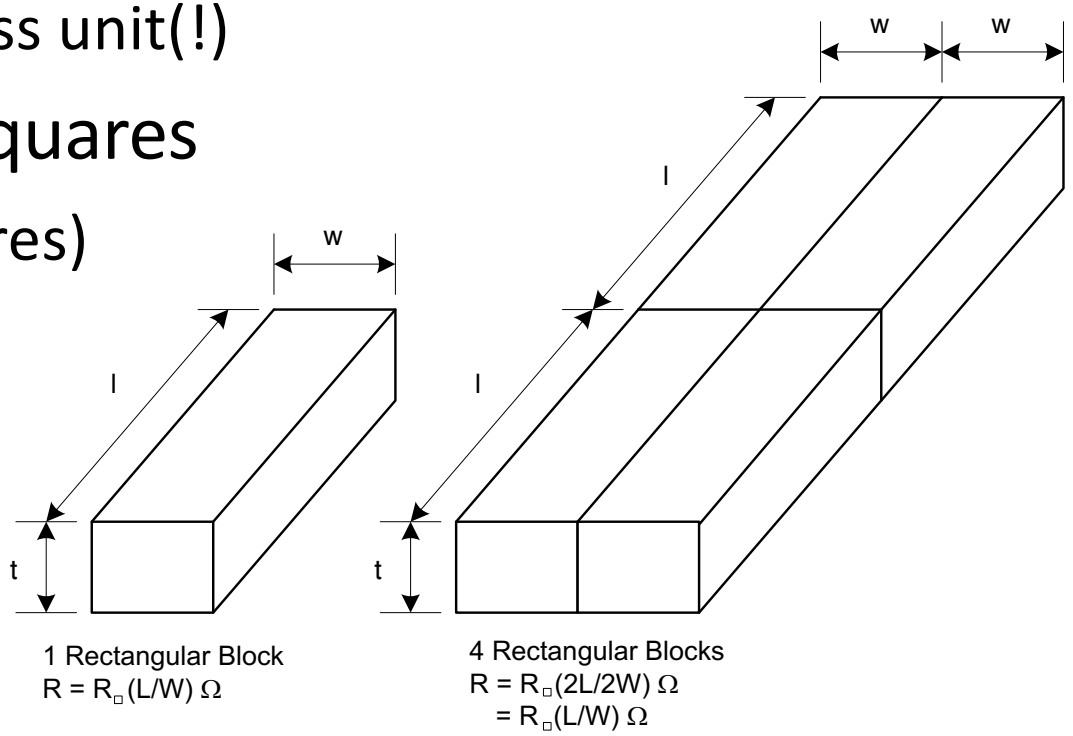
Layer	T (nm)	W (nm)	S (nm)	AR
6	1720	860	860	2.0
	1000			
5	1600	800	800	2.0
	1000			
4	1080	540	540	2.0
	700			
3	700	320	320	2.2
	700			
2	700	320	320	2.2
	700			
1	480	250	250	1.9
	800			



Wire Resistance

- ρ = *resistivity* (Ω*m)

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



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

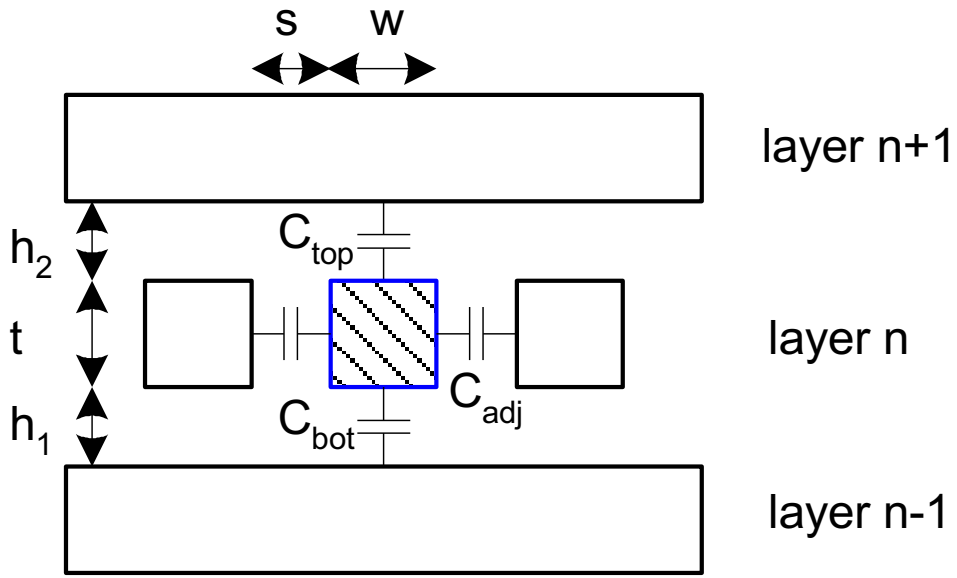
Layer	Sheet Resistance (Ω/□)
Diffusion (silicided)	3-10
Diffusion (no silicide)	50-200
Polysilicon (silicided)	3-10
Polysilicon (no silicide)	50-400
Metal1	0.08
Metal2	0.05
Metal3	0.05
Metal4	0.03
Metal5	0.02
Metal6	0.02

Contact 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
- $C_{total} = C_{top} + C_{bot} + 2C_{adj}$

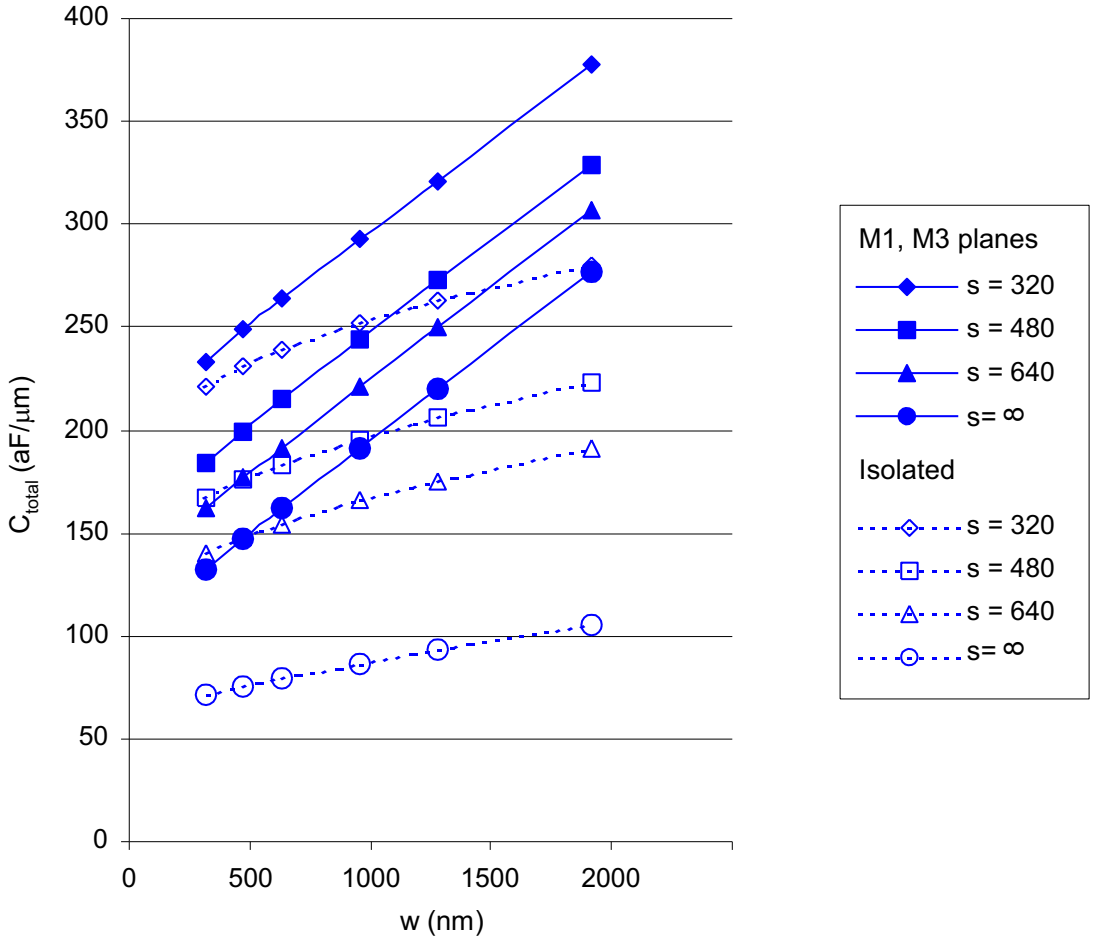


Capacitance Trend

- 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 *f*F/µm
  - Compare to 2 *f*F/µm for gate capacitance

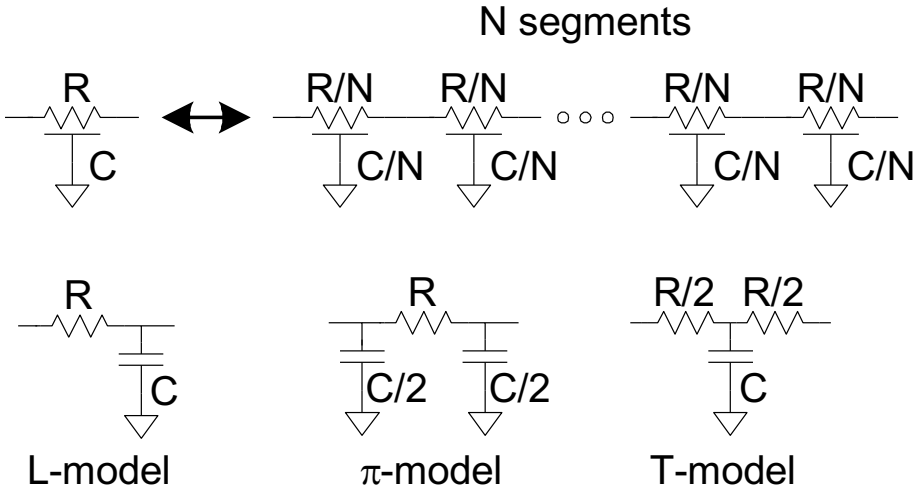


Diffusion and Polysilicon

- Diffusion capacitance is very high (about 2 *f*F/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

- M2 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 =
  - $C_{permicron} = 0.2\ f\text{F}/\mu\text{m}$ → C =

Wire RC Delay

- Estimate the delay of a 10x inverter driving a 2x inverter at the end of the 5-mm wire from previous example
  - Effective R = 2.5 kΩ/μm for gates, C = 2 *f*F/μm
  - Unit inverter: 4λ = 0.36 μm nMOS, 8λ = 0.72 μm pMOS

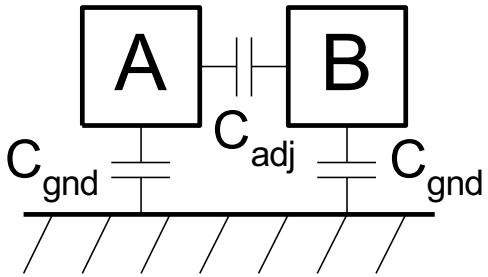
Crosstalk

- Capacitor do not change voltage instantaneously
- A wire has high capacitance to its neighbor
  - When the neighbor (aggressor) switches from 1→0 or 0→1, the wire (victim) tends to switch as well
  - Called *capacitive coupling* or *crosstalk*
- Impacts
  - Cause noise on non-switching wires
  - Increase delay on switching wires

Crosstalk Delay

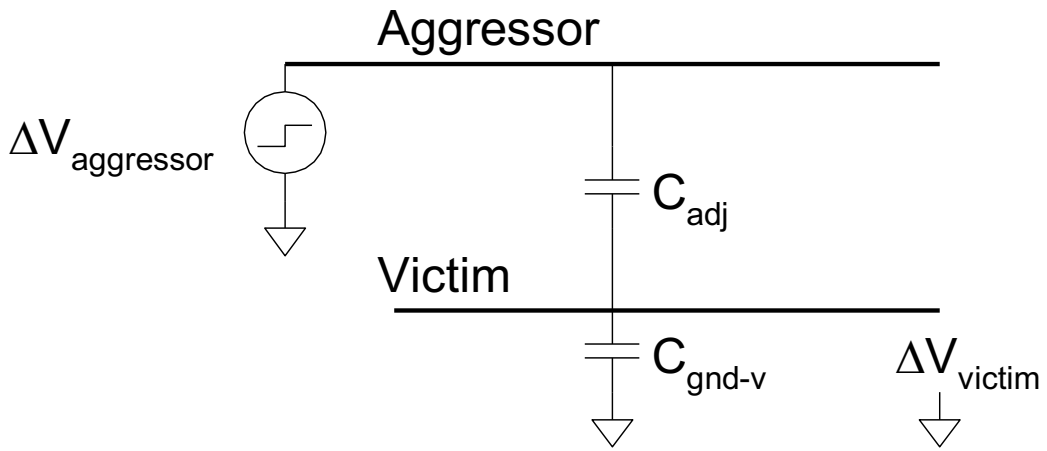
- Assume layers above and below in average are quiet
  - Second terminal of capacitor can be ignored
  - Modeled as $C_{gnd} = C_{top} + C_{bot}$
- Effective C_{adj} depends on behavior of neighbors
  - *Miller Coupling Factor (MCF)*

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



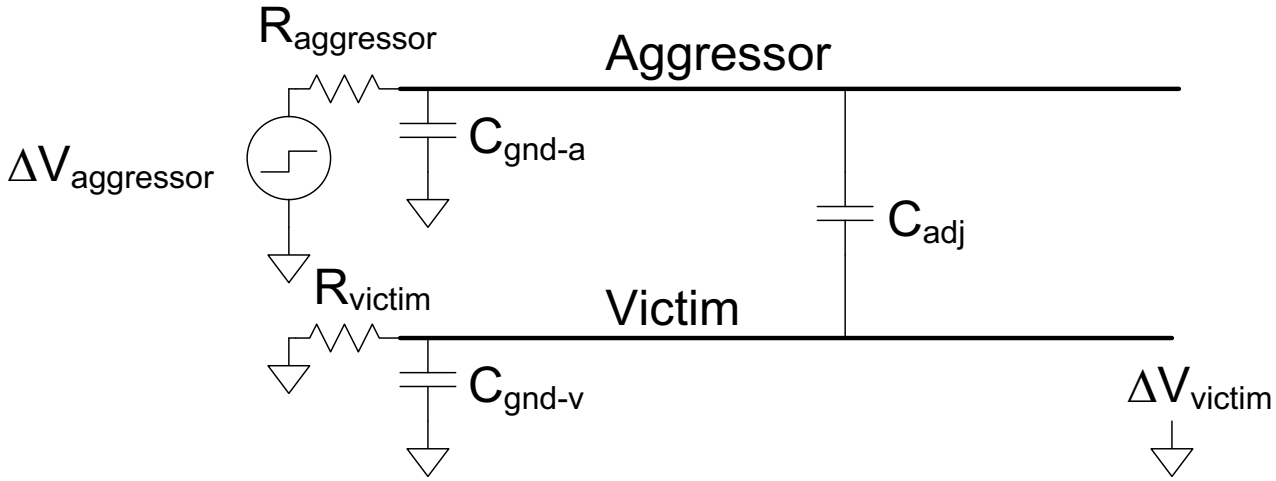
Crosstalk Noise (Floating Victims)

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



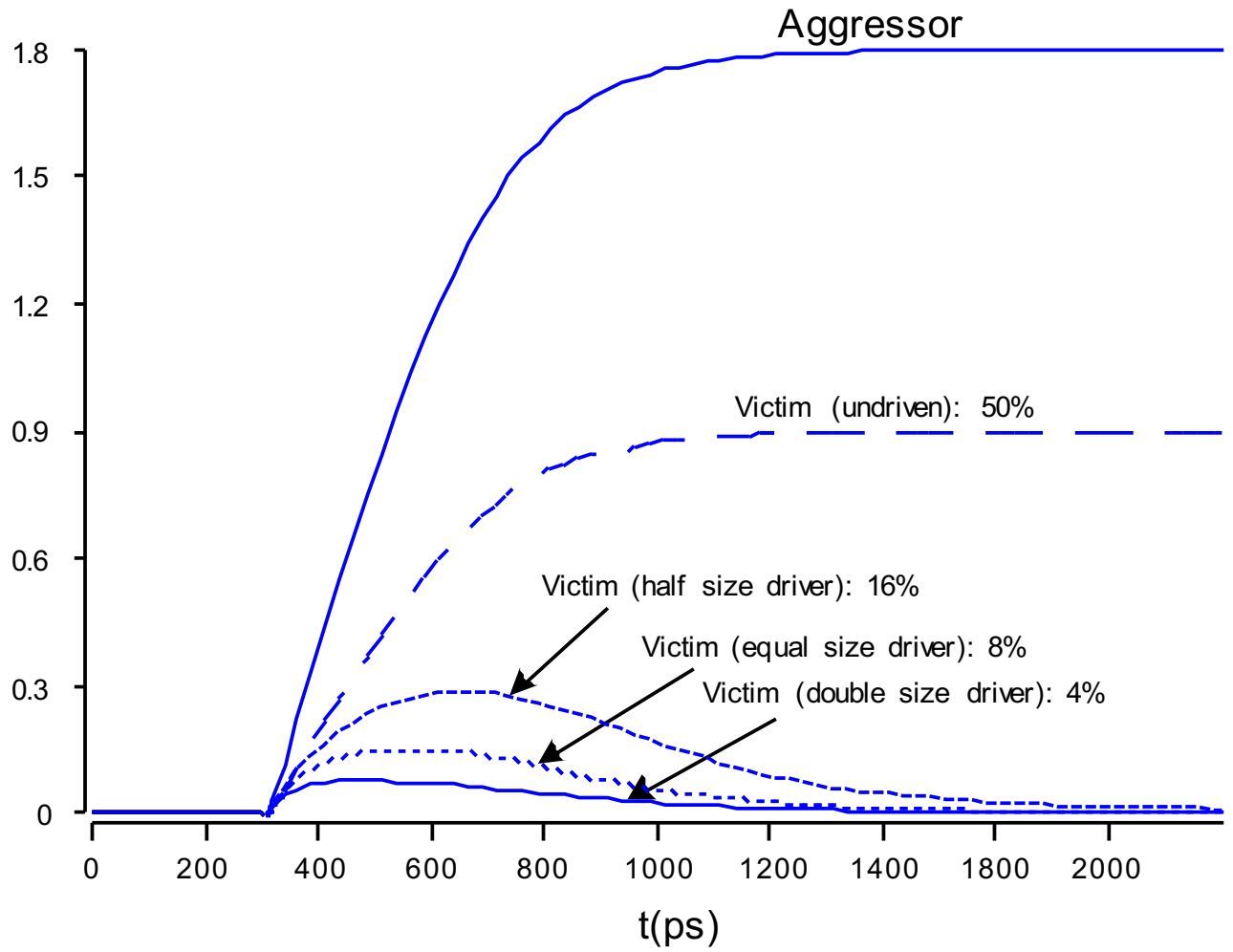
Crosstalk Noise (Driven Victims)

- Usually victim is driven by a gate that fights noise
  - Noise depends on relative resistances
  - Assume victim driver in linear region and aggressor driver in saturation (considering inverter operation)
  - With equal sizes, $R_{aggressor}$ = 2-4 x R_{victim}



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 wrong outputs

Outline

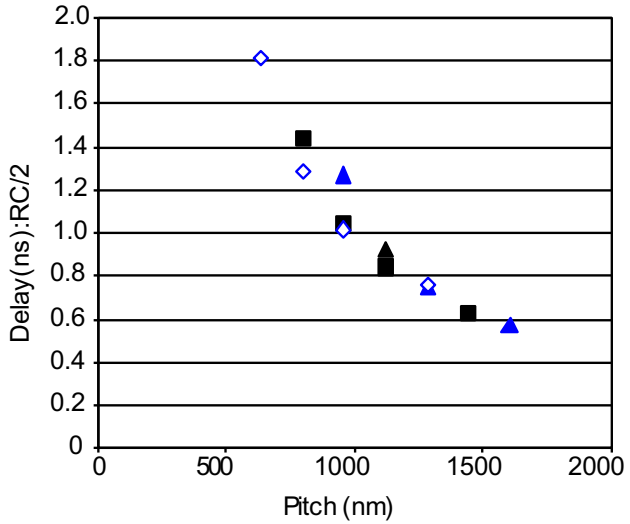
- Delay estimation
- Logical effort and transistor sizing
- Power dissipation
- Interconnect
- **Wire engineering**
- Design margin
- Reliability
- Scaling

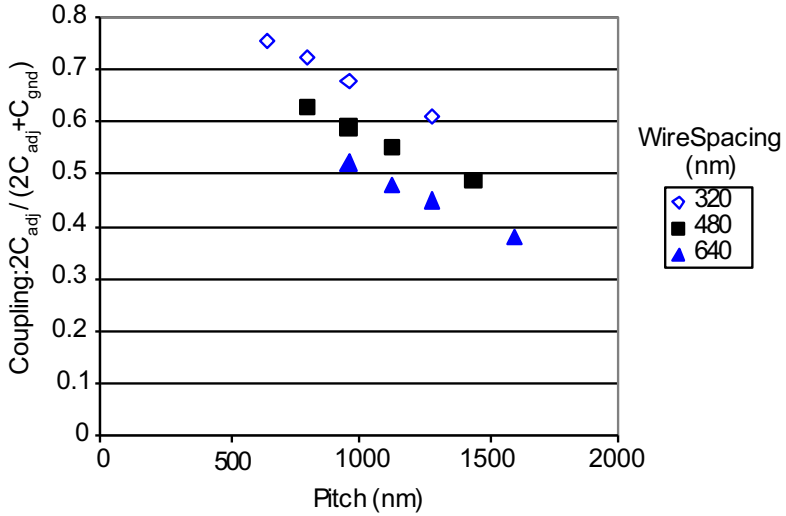
Wire Engineering

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

Wire Engineering

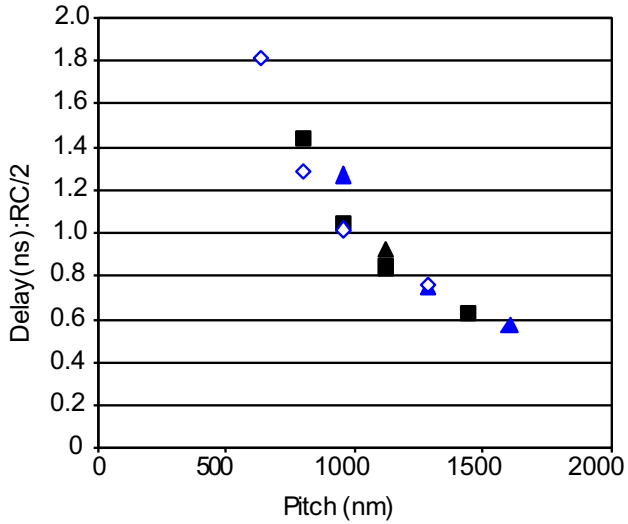
- Goal: to achieve delay, area, and power goals with acceptable noise
- Degrees of freedom:
  - Width
  - Spacing

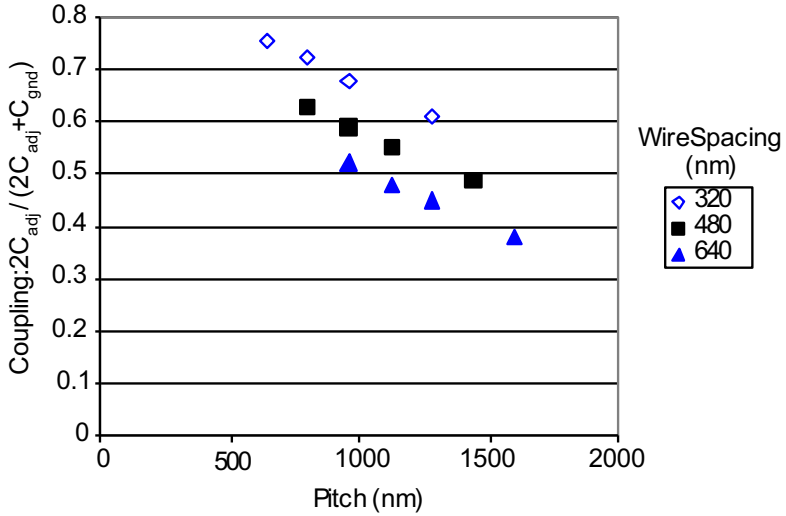


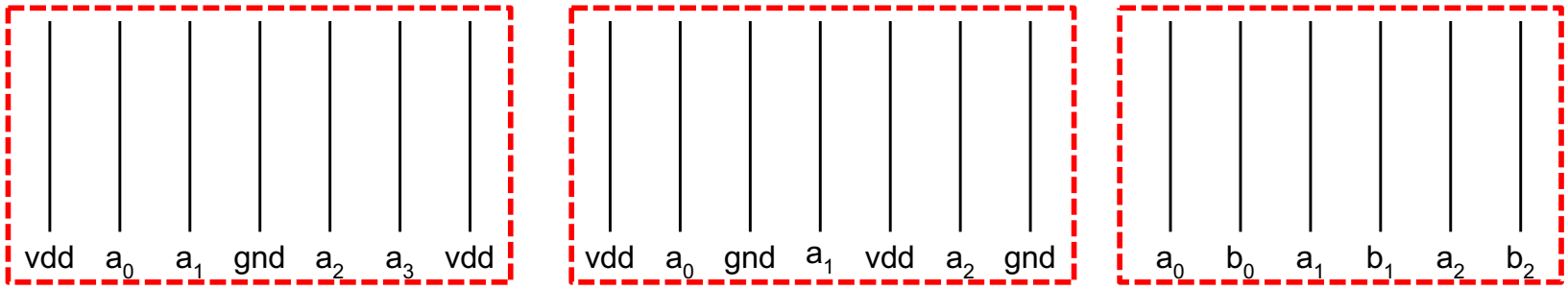


Wire Engineering

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

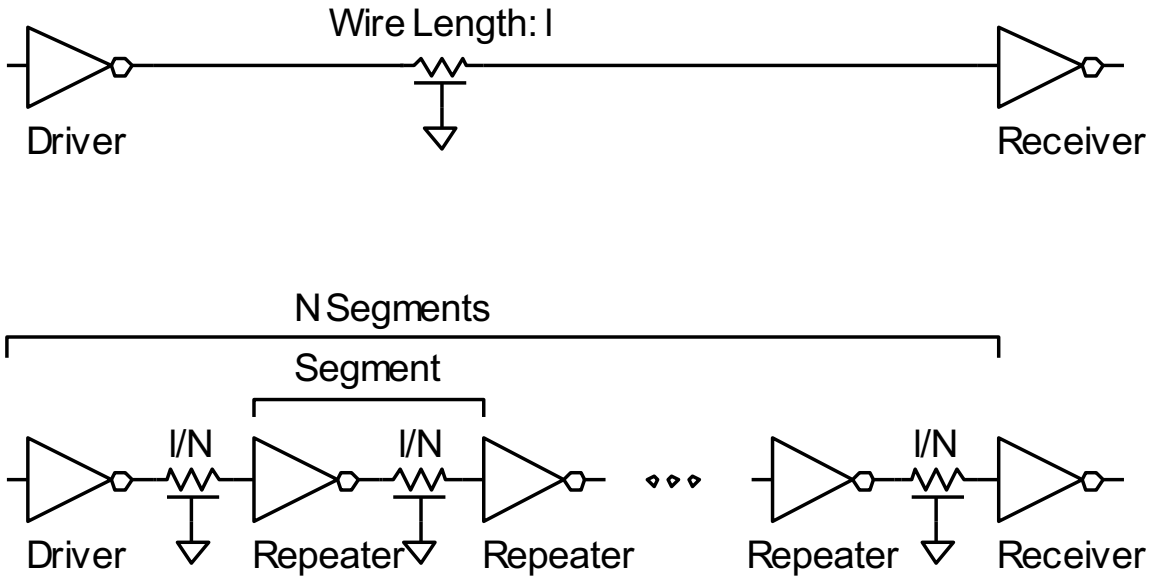






Repeaters

- R and C are proportional to *l* (length)
- RC delay is proportional to l^2
  - Unacceptably long delays 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 Capacitance 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