VLSI Interconnects

Instructed by Shmuel Wimer
Eng. School, Bar-Ilan University

Credits: David Harris
Harvey Mudd College

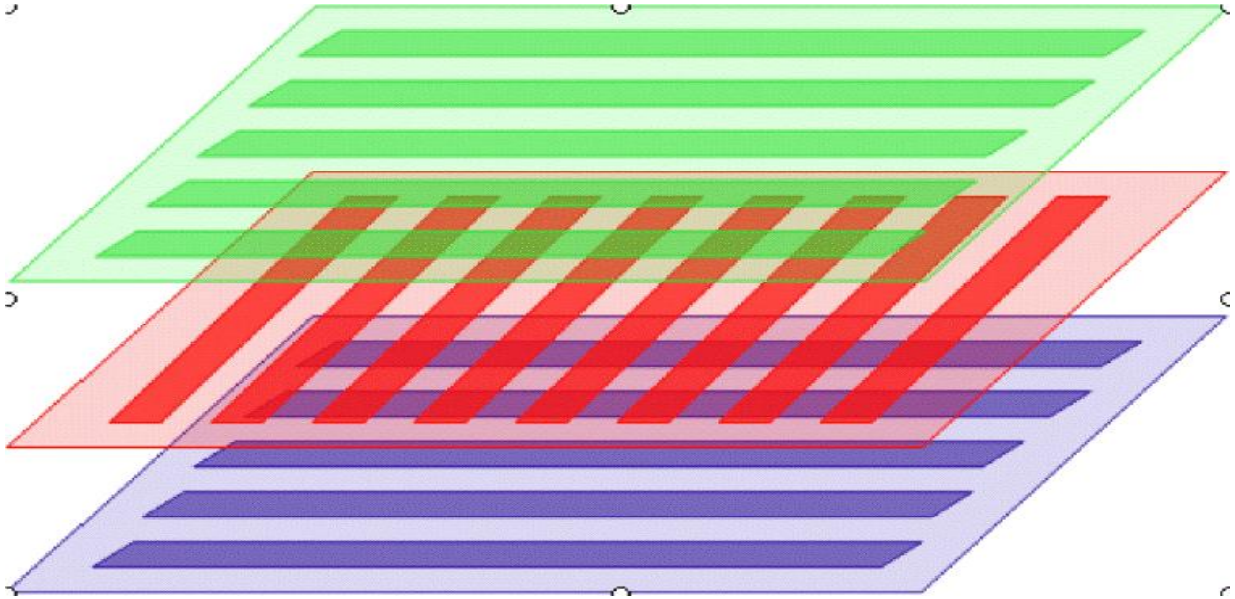
(Some material copied/taken/adapted from Harris' lecture notes)

Outline

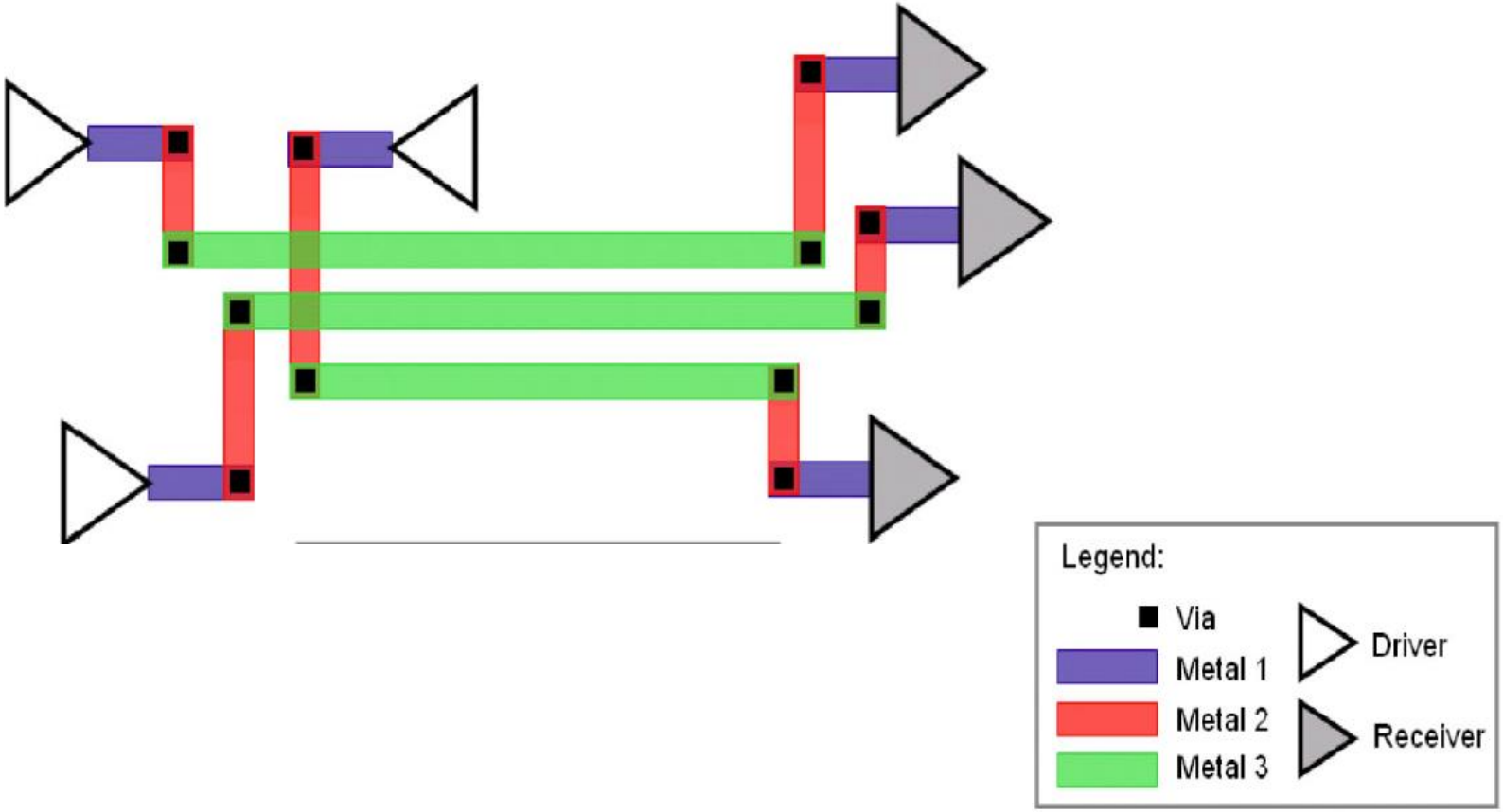
- Introduction
- Wire Resistance
- Wire Capacitance
- Wire RC Delay
- Crosstalk
- Wire Engineering
- Repeaters
- Scaling

Introduction

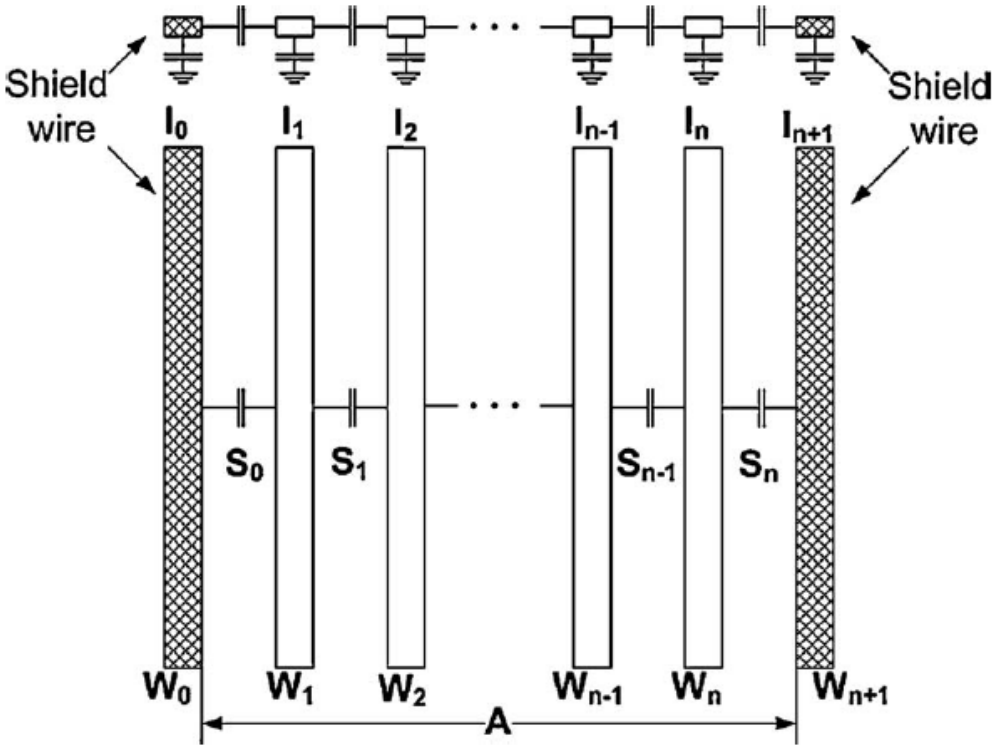
- Chips are mostly made of wires called *interconnect*
- Alternating layers run orthogonally



- Transistors are little things under the wires
- Many layers of wires

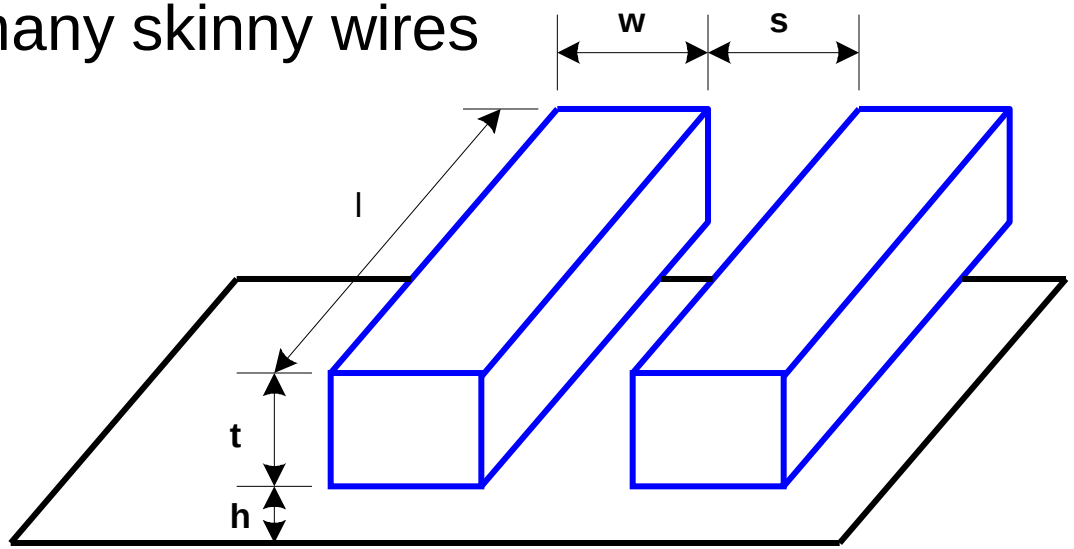


- Wires are as important as transistors
  - Speed
  - Power
  - Noise



Wire Geometry

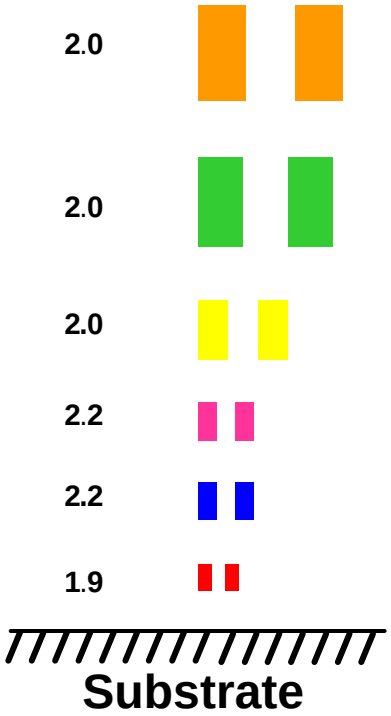
- Pitch = w + s
- Aspect ratio: AR = t/w
  - Old processes had AR << 1
  - Modern processes have AR ≈ 2
    - Pack in many skinny wires



Layer Stack

- 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
5	1600	800	800	2.0
4	1080	540	540	2.0
3	700	320	320	2.2
2	700	320	320	2.2
1	480	250	250	1.9

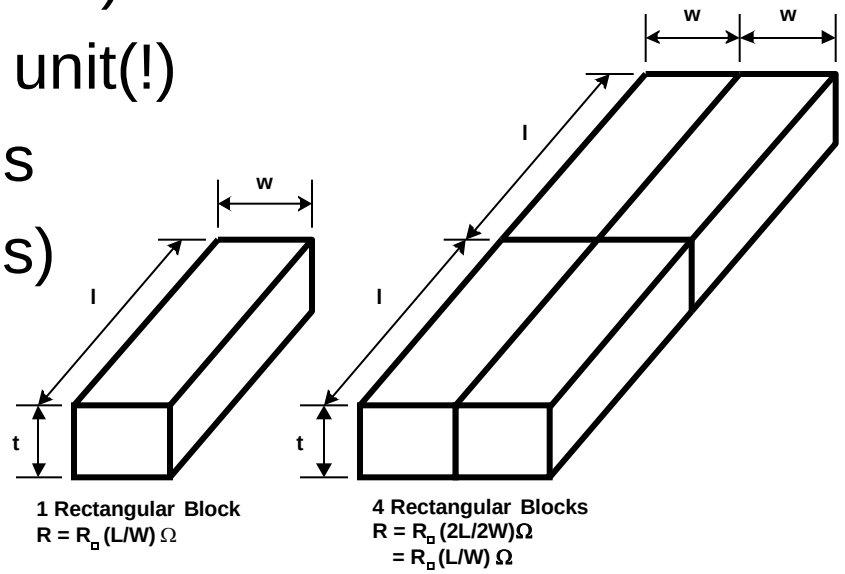


Wire Resistance

- ρ = *resistivity* (Ω*m)

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

- $R_{\square}$ = *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
- Modern processes 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

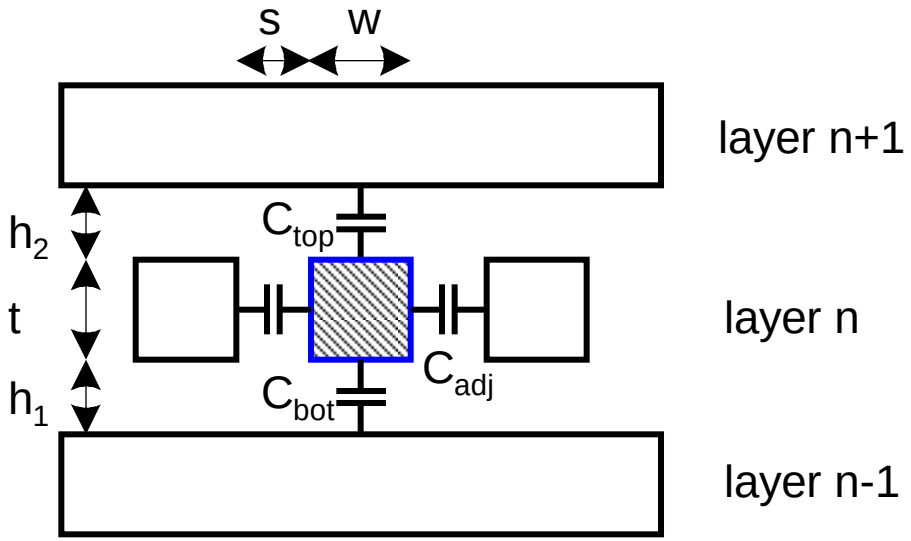
Layer	Sheet Resistance (Ω/□)
Diffusion (silicided)	3-10
Diffusion (no silicide)	50-200
Polysilicon (silicided)	3-10
Polysilicon (no silicide)	50-400
Metal1	0.08
Metal2	0.05
Metal3	0.05
Metal4	0.03
Metal5	0.02
Metal6	0.02

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

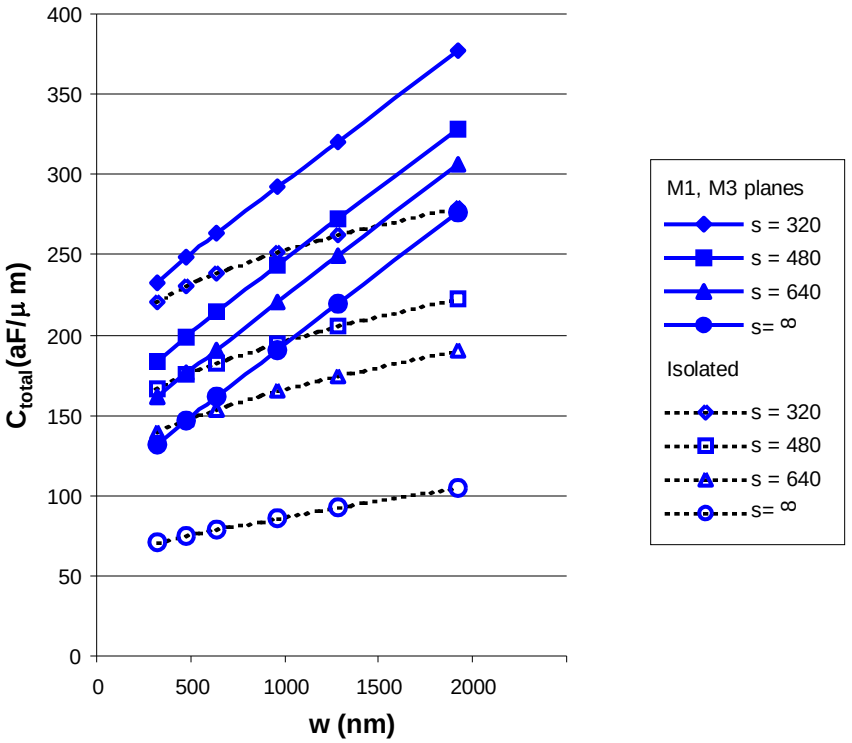
- Parallel plate equation: $C = \varepsilon A/d$
  - Wires are not parallel plates, but obey trends
  - Increasing area (W, t) increases capacitance
  - Increasing distance (s, h) decreases capacitance
- Dielectric constant
  - $\varepsilon = k\varepsilon_0$
- $\varepsilon_0 = 8.85 \times 10^{-14}$ F/cm
- 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

- Capacitance increases with metal planes above and below

- Capacitance decreases with spacing

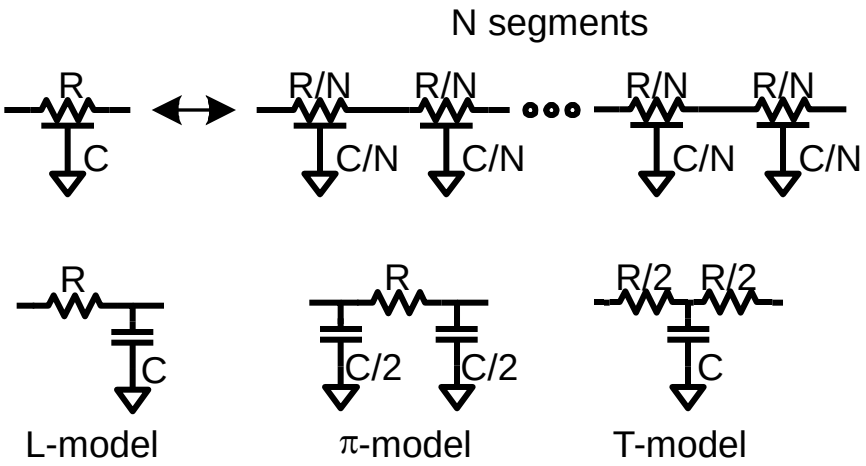


Diffusion & Polysilicon

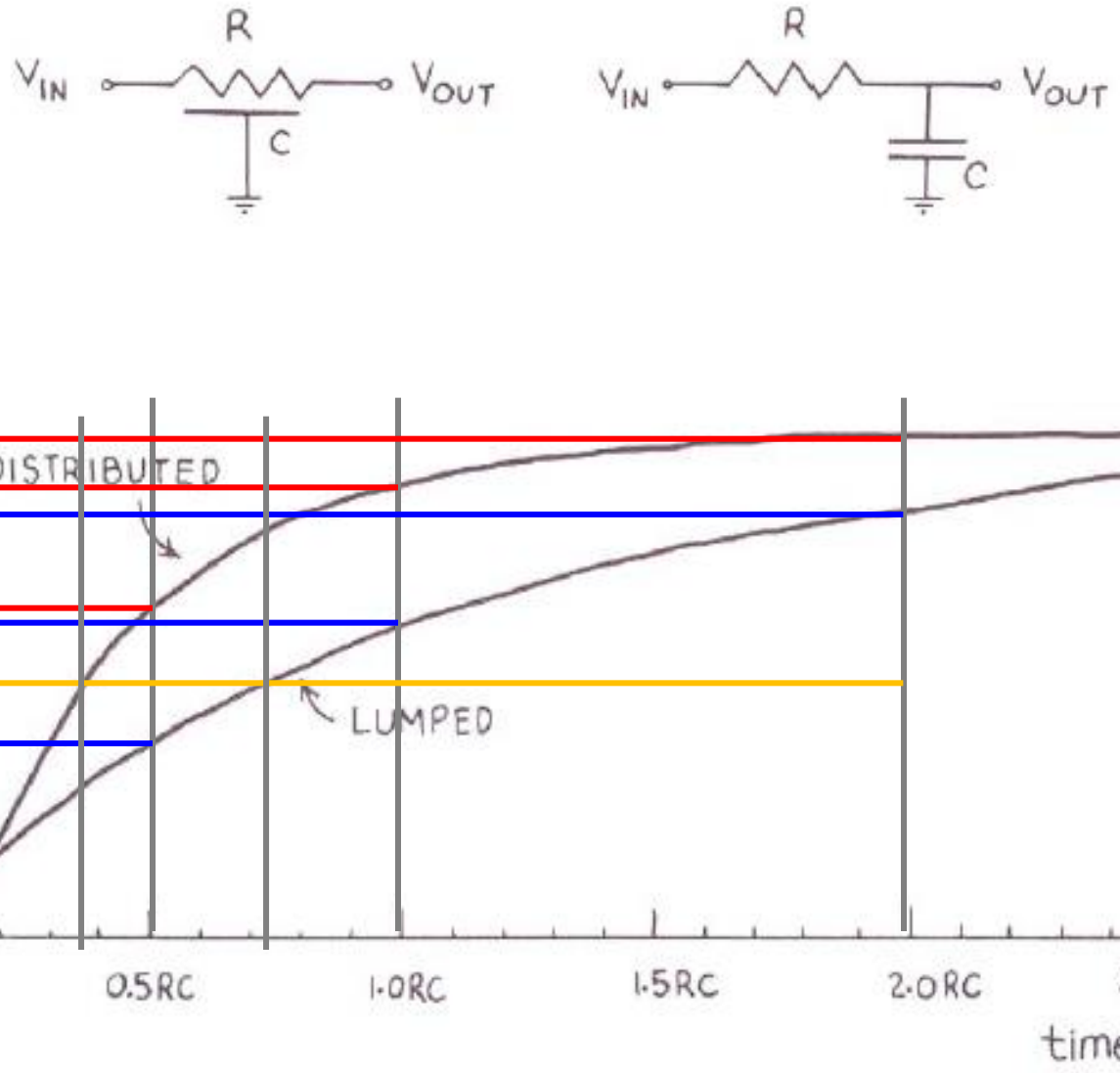
- Diffusion capacitance is very high (about 2 fF/μm)
  - Comparable to gate capacitance
  - Diffusion also has high resistance
  - Avoid using diffusion (and polysilicon) *runners* for wires!
- Polysilicon has lower C but high R
  - Use for transistor gates
  - Occasionally for very short wires between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

Lumped Element Models

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



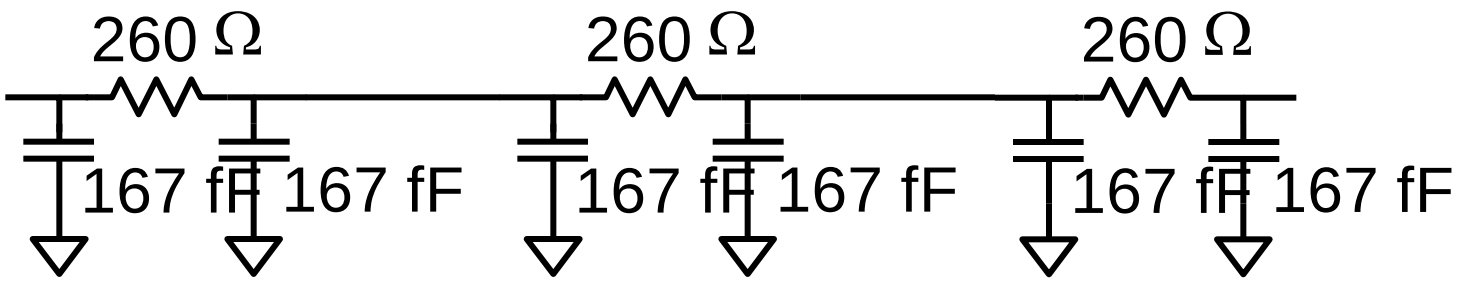
- 3-segment π-model is accurate to 3% in simulation

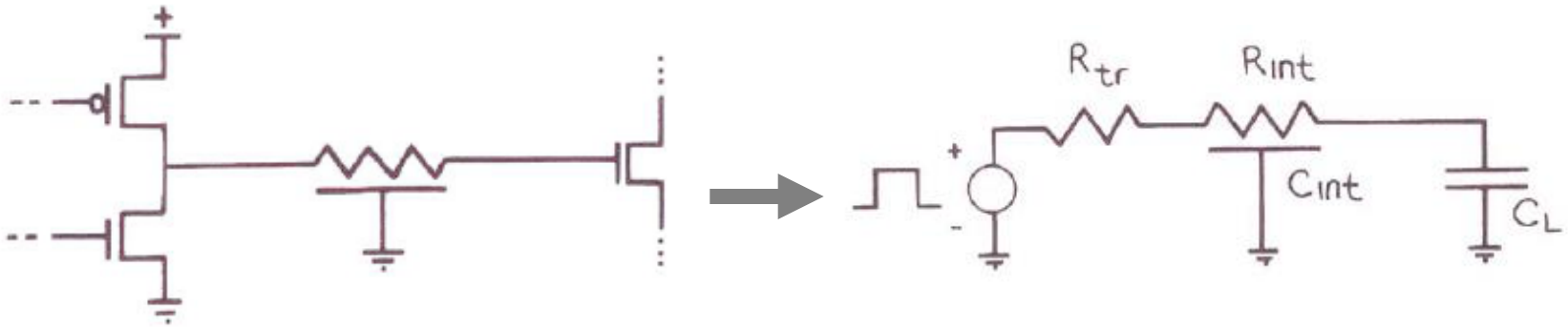

R
V_{IN}
V_{OUT}
C
R
V_{IN}
V_{OUT}
C
V_{OUT}
1.0
0.5
DISTRIBUTED
LUMPED
0.5RC
1.0RC
1.5RC
2.0RC
2.5RC
time

Output potential range	Time elapsed (Distributed RC Network)	Time elapsed (Lumped RC Network)
0 to 90%	1.0 RC	2.3 RC
10% to 90% (rise time)	0.9 RC	2.2 RC
0 to 63%	0.5 RC	1.0 RC
0 to 50% (delay)	0.4 RC	0.7 RC
0 to 10%	0.1 RC	0.1 RC

RC Example

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





distributed lumped

$$T_{90\%} = \overbrace{1.0R_{int}C_{int}}^{\text{distributed}} + \overbrace{2.3(R_{tr}C_{int} + R_{tr}C_L + R_{int}C_L)}^{\text{lumped}}$$
$$\approx (2.3R_{tr} + R_{int})C_{int} \quad \text{for} \quad C_L \ll C_{int}.$$

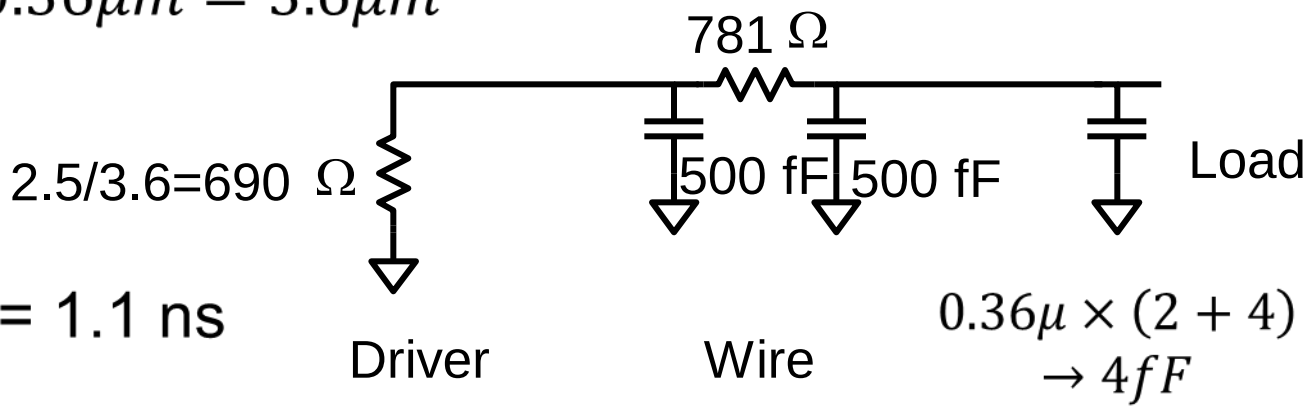
distributed lumped

$$T_{50\%} = \overbrace{0.4R_{int}C_{int}}^{\text{distributed}} + \overbrace{0.7(R_{tr}C_{int} + R_{tr}C_L + R_{int}C_L)}^{\text{lumped}}$$
$$\approx (0.7R_{tr} + 0.4R_{int})C_{int} \quad \text{for} \quad C_L \ll C_{int}.$$

Wire RC Delay Example

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

$10 \times 0.36\mu m = 3.6\mu m$



  - t_{pd} = 1.1 ns

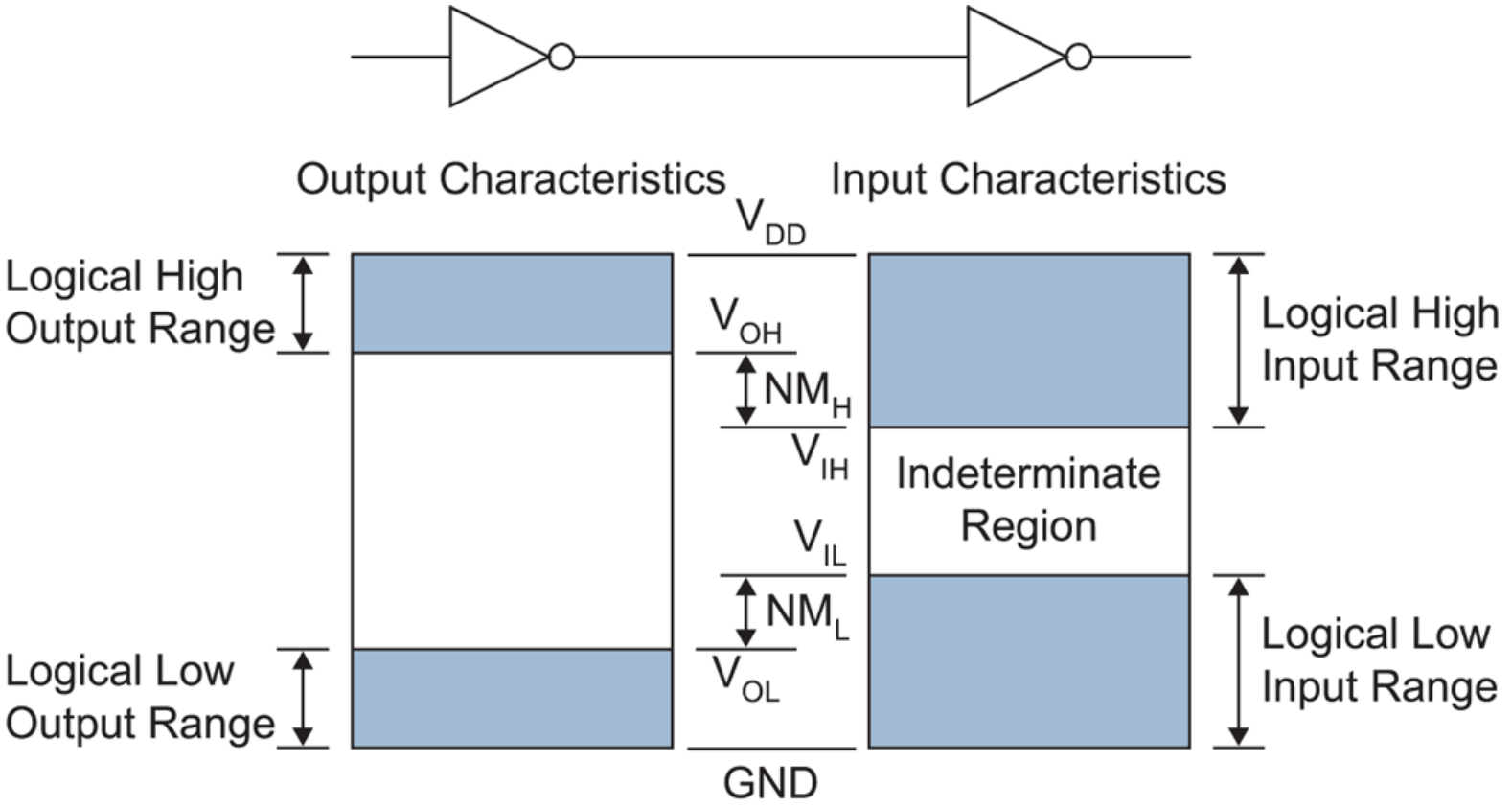


FIG 2.27 Noise margin definitions

Crosstalk

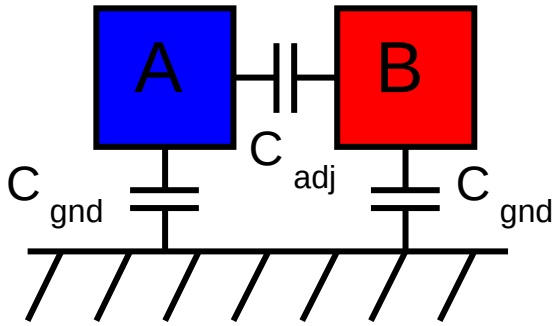
- A capacitor does not like to change its voltage instantaneously
- A wire has high capacitance to its neighbor.
  - When the neighbor switches from 1→ 0 or 0→ 1, the wire tends to switch too.
  - Called capacitive *coupling* or *crosstalk*
- Crosstalk effects
  - Noise on non switching wires
  - Increased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

Miller Effect

- 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 effect*

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

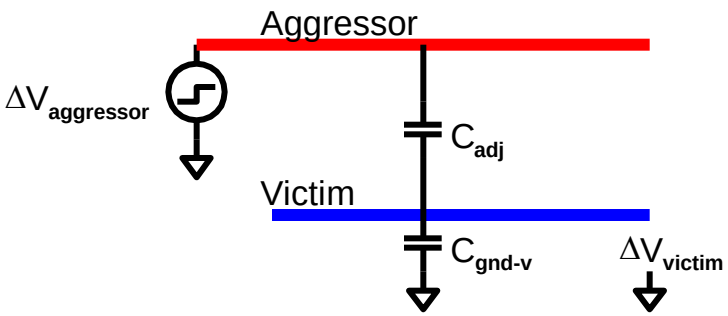
Delay and power increase



Crosstalk Noise

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

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

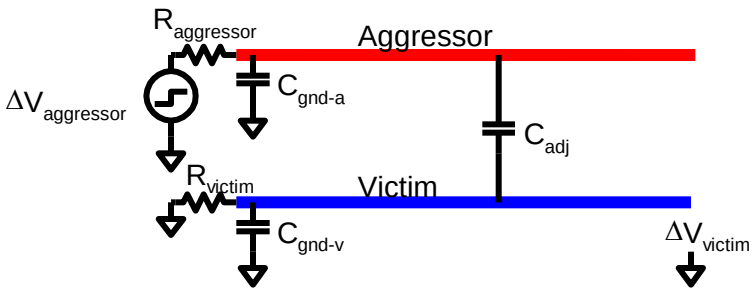


Driven Victims

- 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}

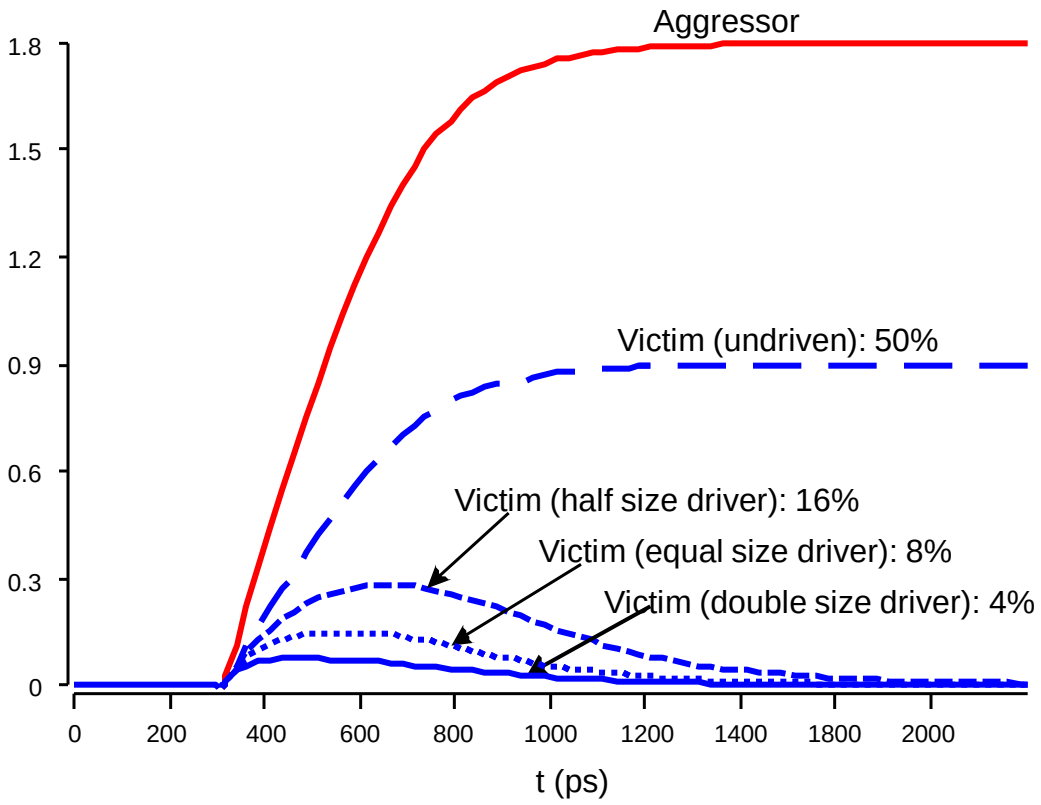
$$\Delta V_{victim} = \frac{C_{adj}}{C_{gnd\text{-}v} + C_{adj}} \frac{1}{1+k} \Delta V_{aggressor}$$

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



Coupling Waveforms

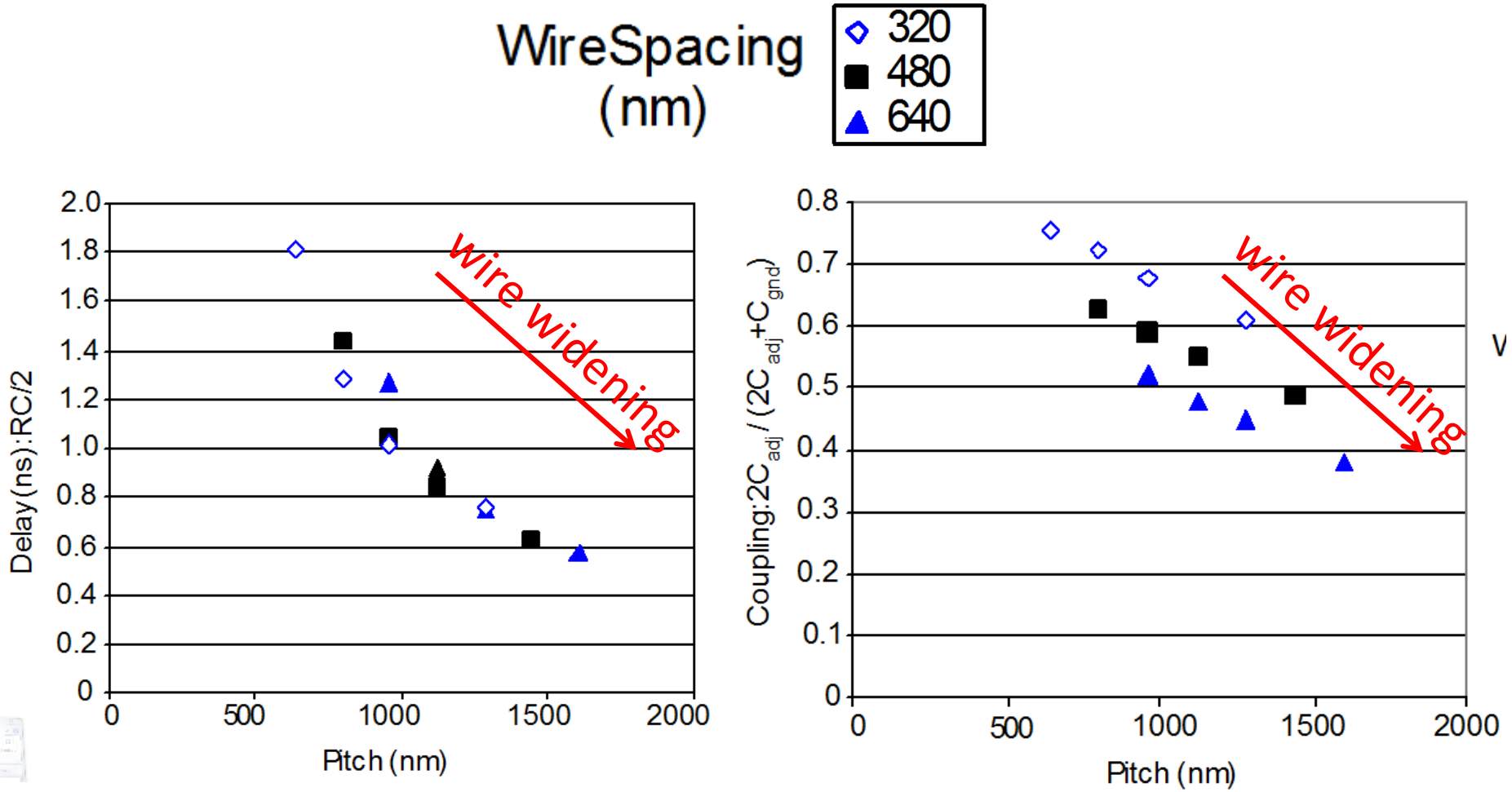
- Simulated coupling for $C_{adj} = C_{victim}$



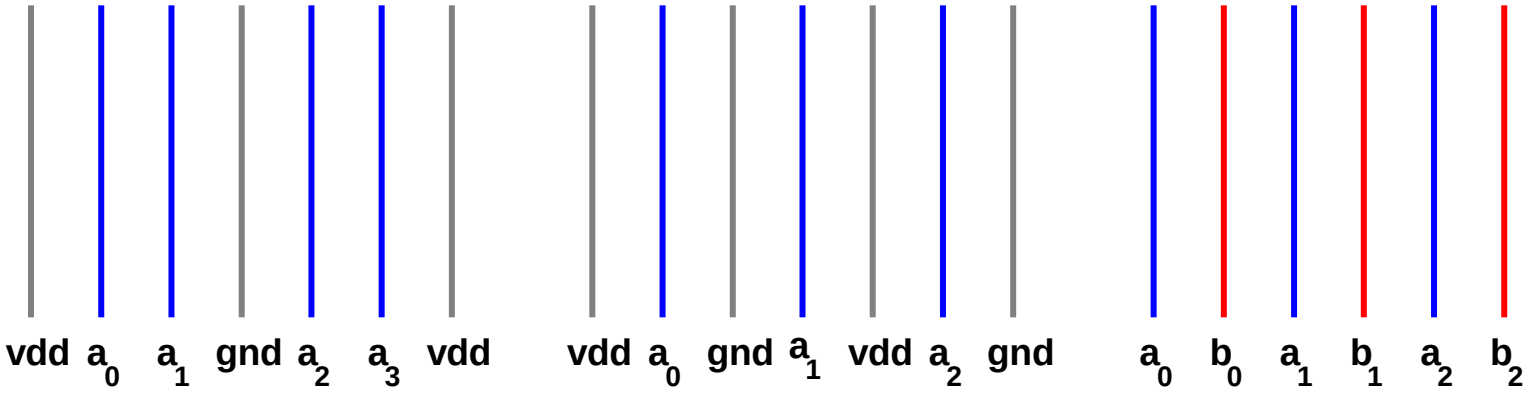
Noise Implications

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

Wire Engineering

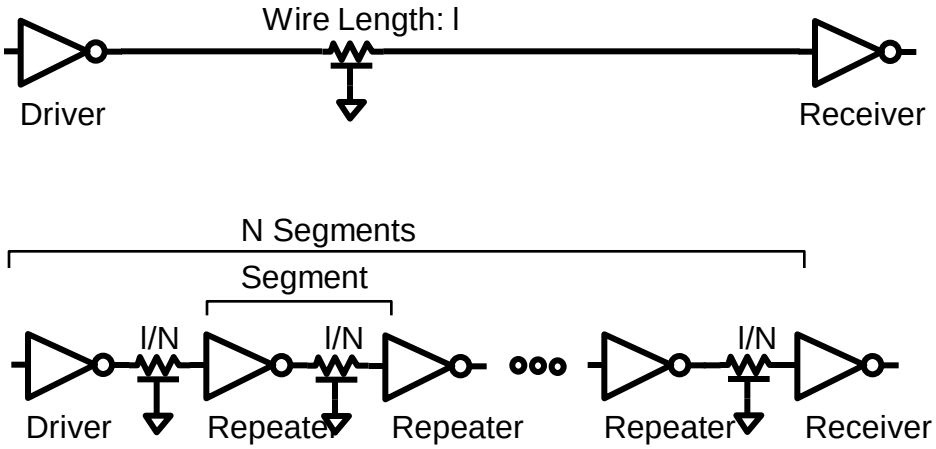


- Goal: achieve delay, area, power with acceptable noise
- Degrees of freedom:
  - Width
  - Spacing
  - Layer
  - Shielding



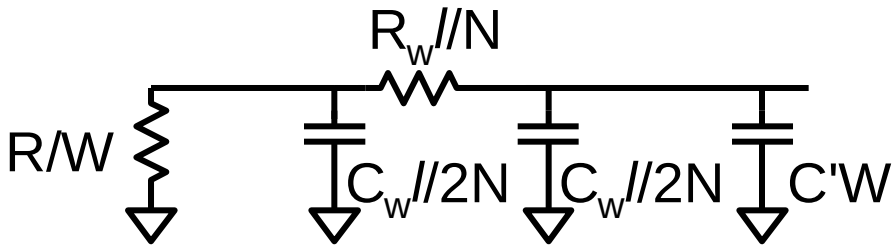
Repeaters

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

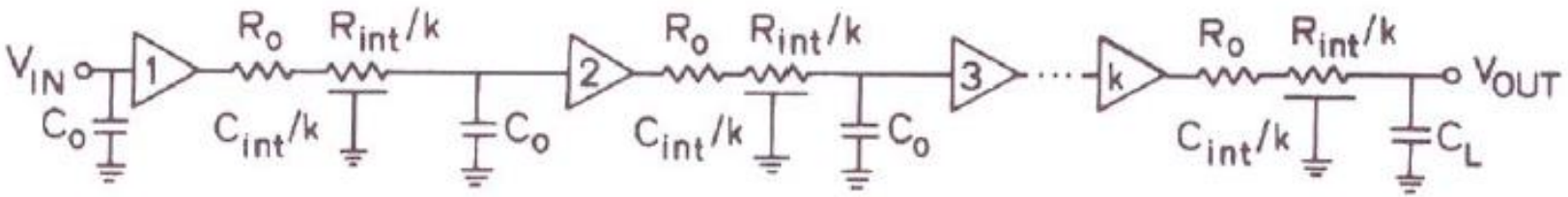


Repeater Design

- 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



(a) minimum-size repeaters,



single chain

$$T_{50\%} = k\left[0.7R_o\left(\frac{C_{int}}{k} + C_o\right) + \frac{R_{int}}{k}\left(0.4\frac{C_{int}}{k} + 0.7C_o\right)\right]$$

$$dT/dk = 0, \qquad 0.4\frac{R_{int}C_{int}}{k^2} = 0.7R_oC_o.$$

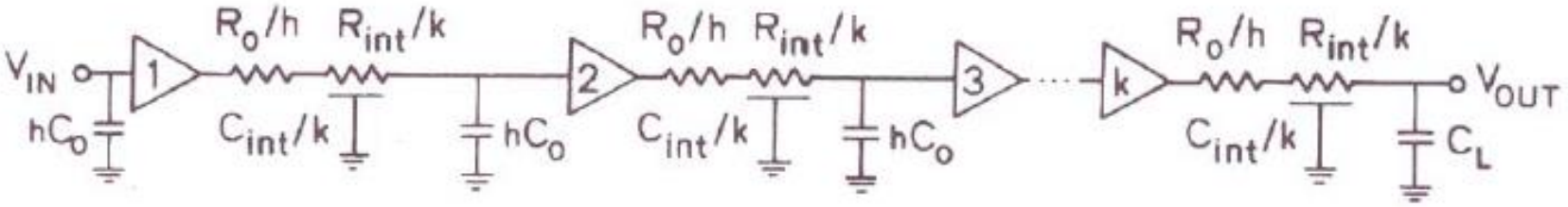
$$k = \sqrt{\frac{0.4R_{int}C_{int}}{0.7R_oC_o}}.$$

$$k = \sqrt{\frac{0.4 R_{int} C_{int}}{0.7 R_o C_o}}$$

Under what condition repeater insertion should take place?

$$\sqrt{\frac{0.4\, C_{int} R_{int}}{0.7\, C_o R_o}} \geq 2 \quad \frac{C_{int} R_{int}}{C_o R_o} \geq 7$$

(b) optimal repeaters

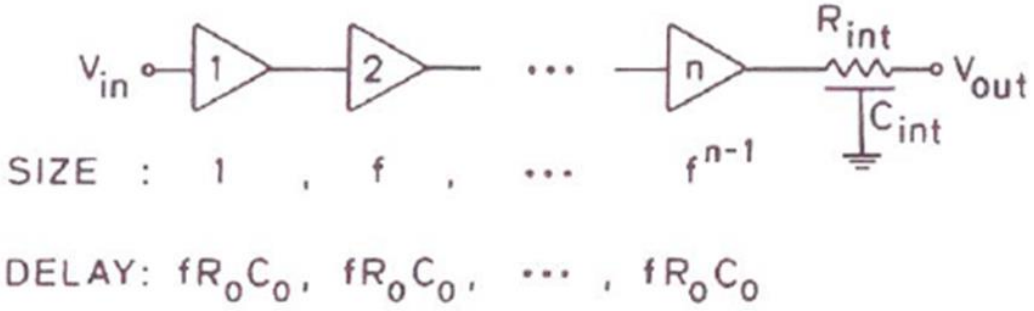


$$T_{50\%} = k\left[0.7\frac{R_o}{h}\left(\frac{C_{int}}{k} + hC_o\right) + \frac{R_{int}}{k}\left(0.4\frac{C_{int}}{k} + 0.7hC_o\right)\right]$$

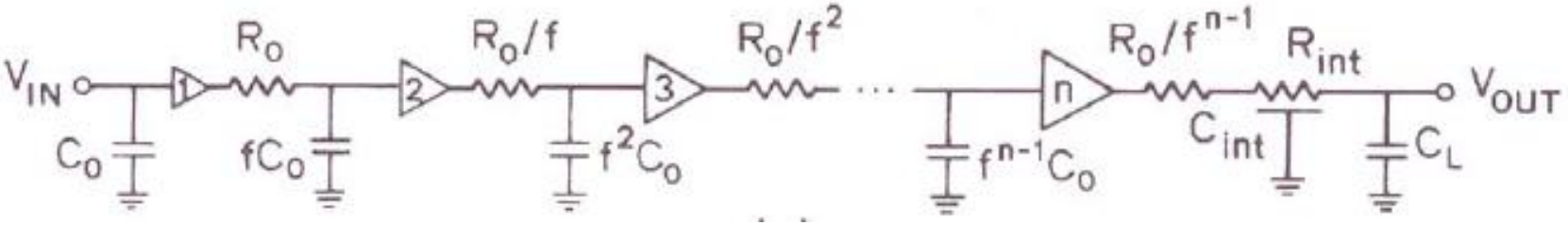
setting dT/dk and dT/dh to zero.

$$k = \sqrt{\frac{0.4R_{int}C_{int}}{0.7R_oC_o}} \qquad h = \sqrt{\frac{R_oC_{int}}{R_{int}C_o}}$$

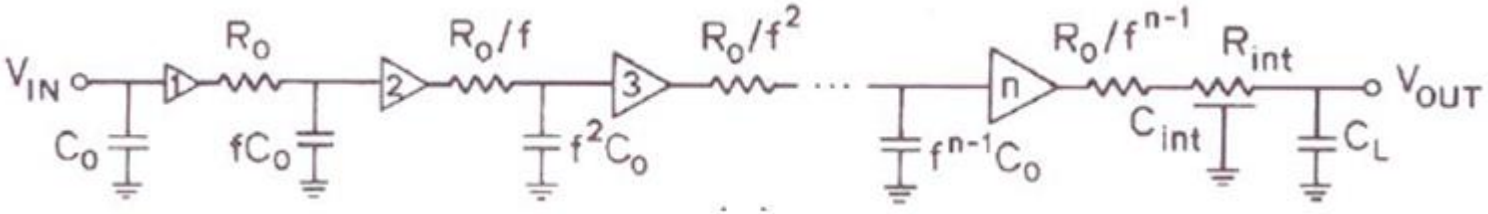
$$T_{50\%} = 2.5\sqrt{R_oC_oR_{int}C_{int}}$$



(c) cascaded drivers



The method is useful when R_{tr} is dominant and C_{int} is large

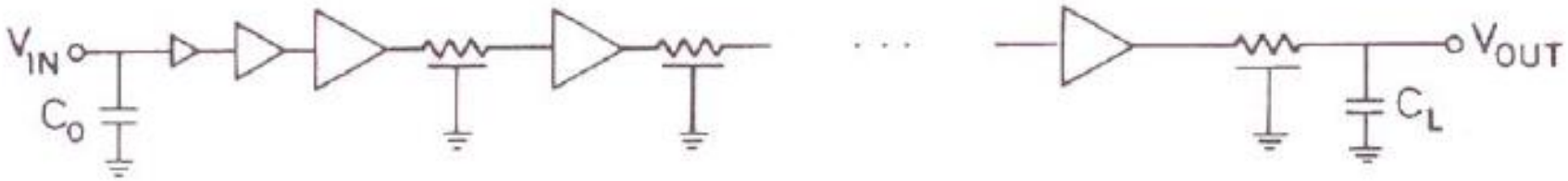


$$T_{50\%} = 0.7(n-1)fR_oC_o + \left(\frac{0.7R_o}{f^{n-1}} + 0.4R_{int}\right)C_{int} + \left(\frac{0.7R_o}{f^{n-1}} + 0.7R_{int}\right)C_L$$

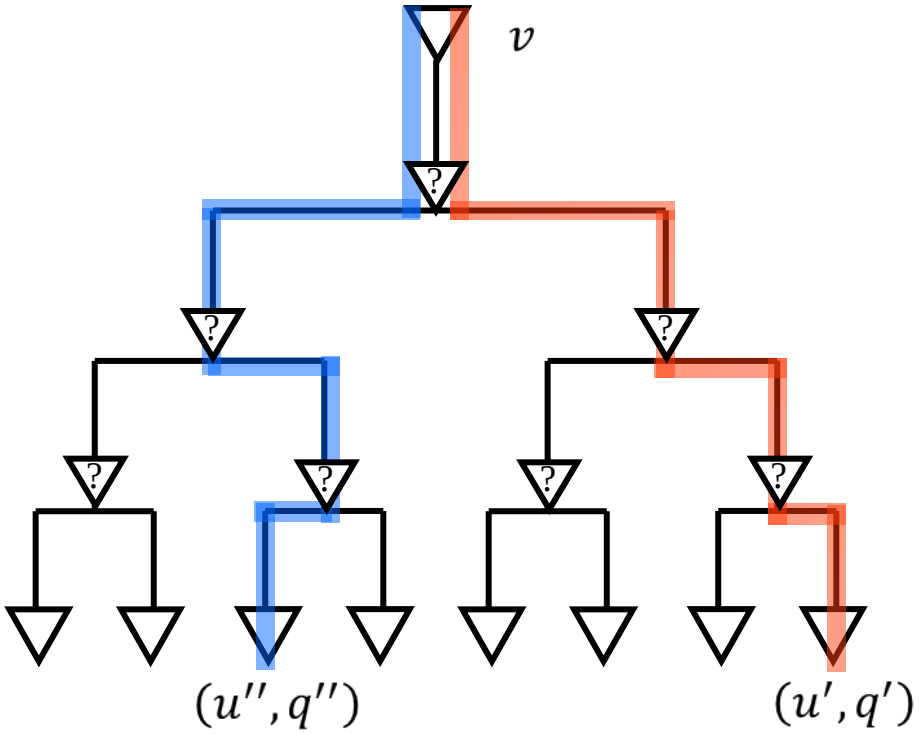
setting dT/dn and dT/df to zero,

$$f = e \qquad n = \ln\left(\frac{C_{int} + C_L}{C_o}\right)$$

(d) optimal repeaters with a cascaded first stage



Optimal Buffer Insertion



$d(v, u_i)$: driver to receiver delay.

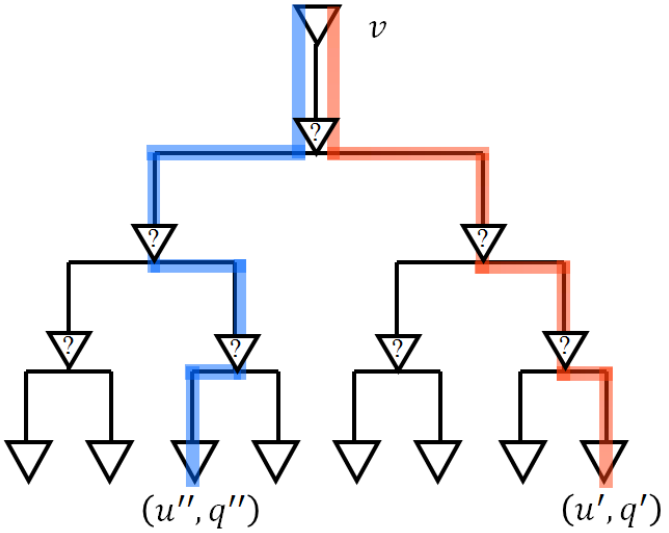
Root required time: $T = \min_i \{q_i - d(v, u_i)\}$.

Optimal Buffer Insertion

Buffer reduces load delay

but adds internal delay,

power and area.



Problem 1:

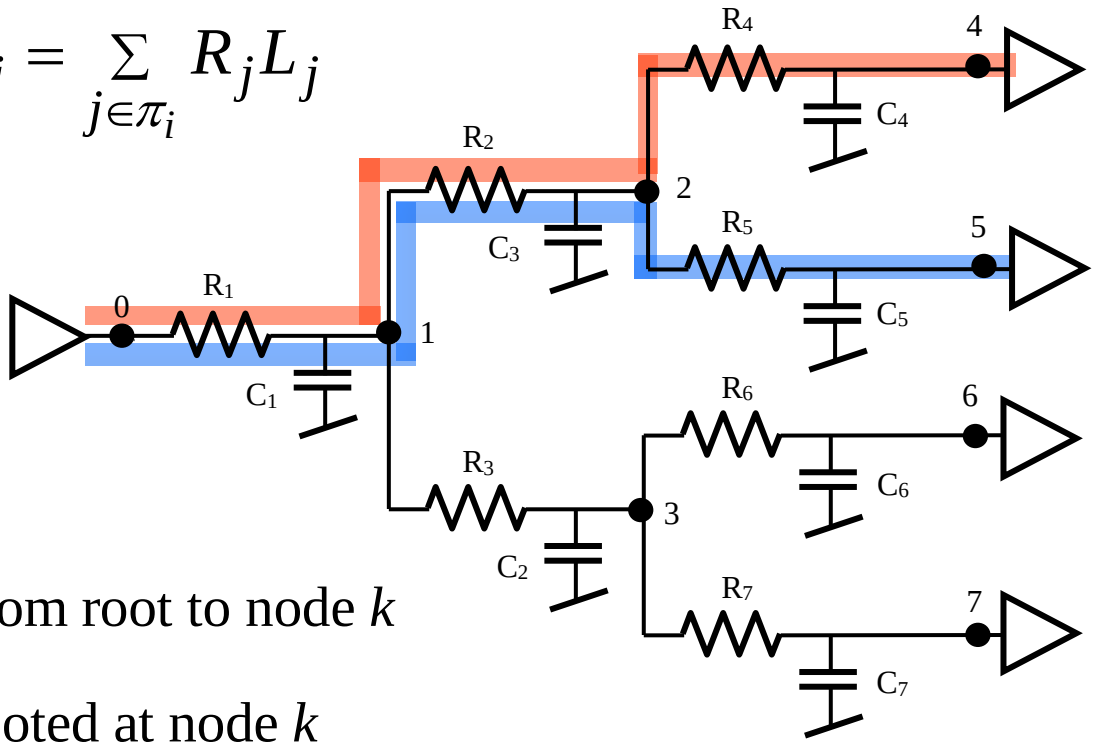
$$\max_{\text{buffer insertions}} \left\{ \min_i \left\{ q_i - d(v, u_i) \right\} \right\} \text{ by buffer insertions.}$$

Problem 2:

$$\max_{\text{buffer insertions}} \left\{ \min_i \left\{ q_i - d(v, u_i) \right\} \right\}, \text{ s.t. power and area constraints.}$$

Delay Model

$$d(v,u_i) = \sum_j R_{ji} C_j = \sum_{j \in \pi_i} R_j L_j$$



π_k - nodes along path from root to node k

T_k - nodes of sub-tree rooted at node k

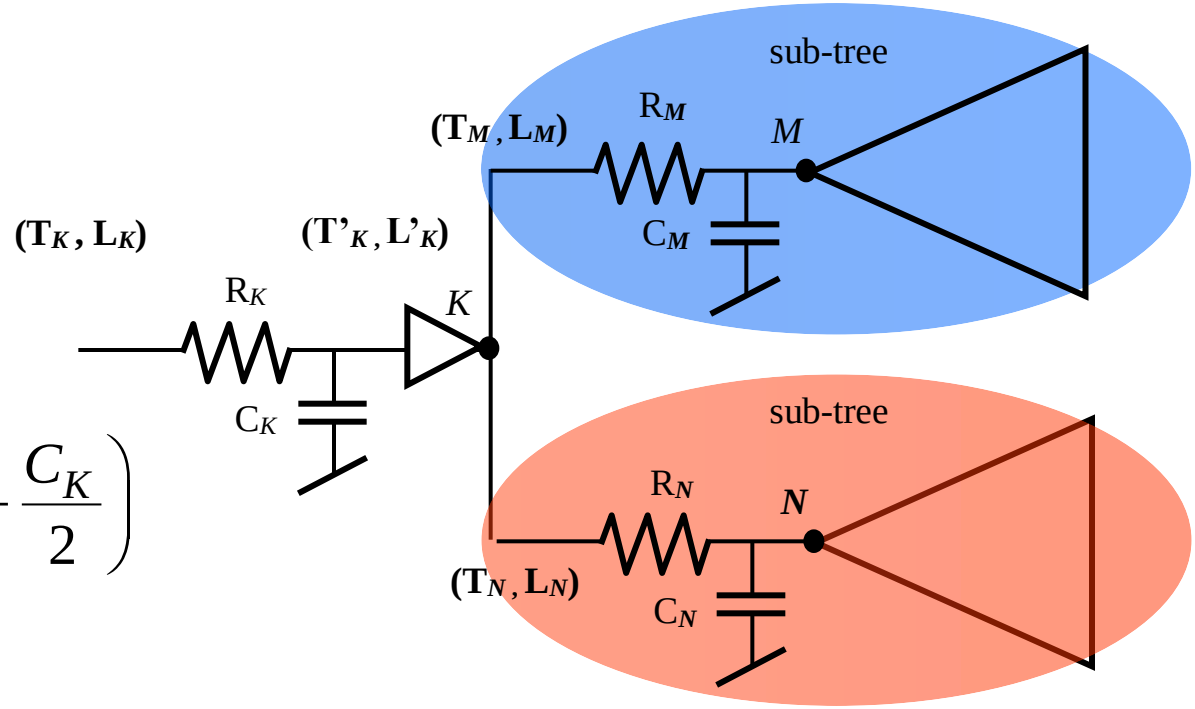
$R_{kl} = \sum_{j \in \pi_k \cap \pi_l} R_j$ - resistance along common paths

$L_k = \sum_{j \in T_k} C_j$ - capacitance of sub-tree

Bottom-Up Solution

$$T'_K = \min\{T_M, T_N\}$$

$$L'_K = L_M + L_N$$



without buffer

$$T_K = T'_K - R_K\left(L'_K + \frac{C_K}{2}\right)$$

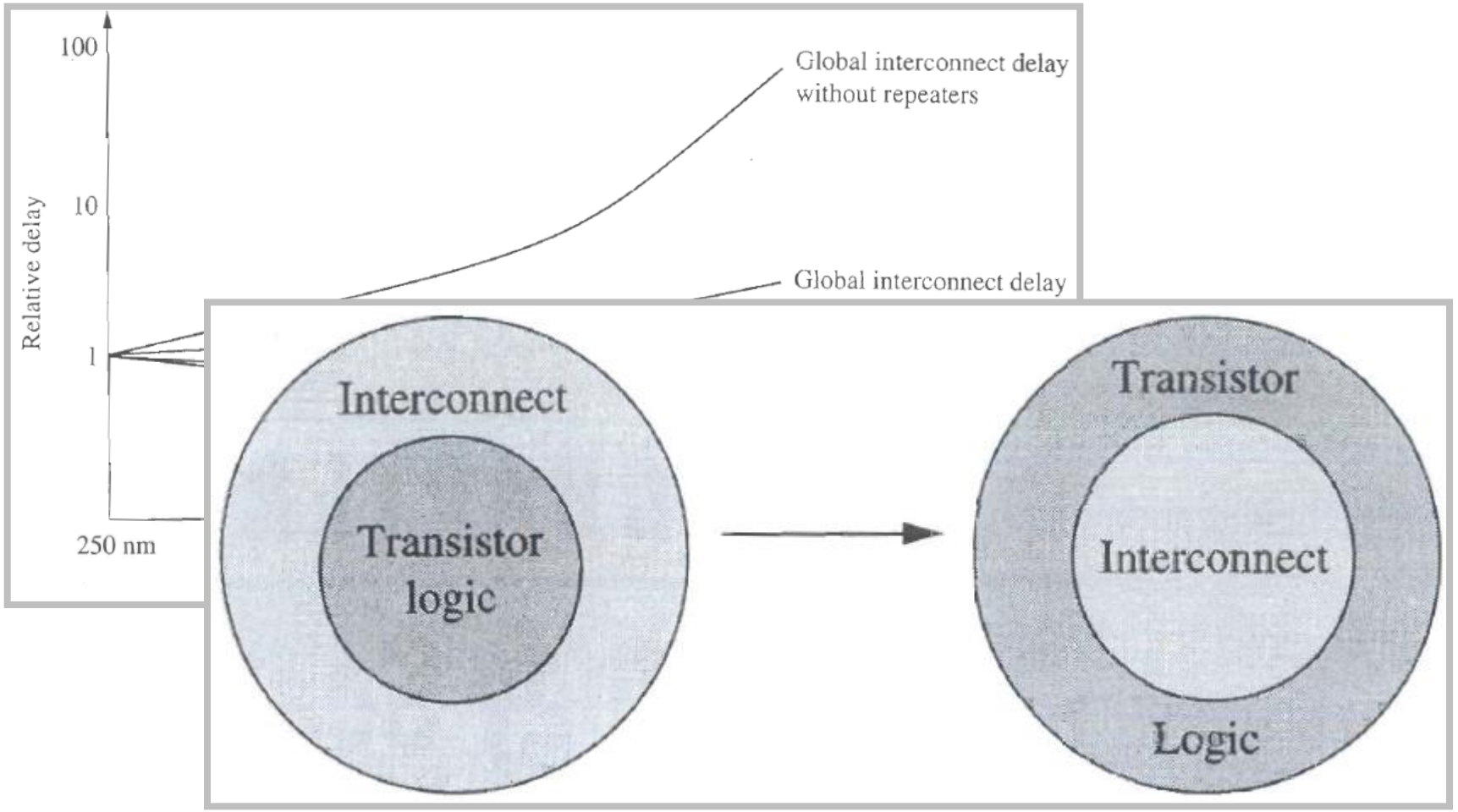
$$L_K = L'_K + C_K$$

with buffer

$$T_K = T'_K - D_{\text{buffer}} - R_{\text{buffer}}L'_K - R_K\left(C_{\text{buffer}} + \frac{C_K}{2}\right)$$

$$L_K = C_{\text{buffer}} + C_K$$

Scaling Trends



Device Scaling

Performance parameter	Constant electric field	Constant voltage
Threshold voltage	$1/S$	1
Current	$1/S$	S
Subthreshold slope	1	1
Capacitance	$1/S$	$1/S$
On-resistance	1	$1/S$
Delay	$1/S$	$1/S^2$
Power	$1/S^2$	S
Power-delay product	$1/S^3$	$1/S$
Power density	1	S^3

Interconnect Scaling

	Ideal scaling		Quasi-ideal scaling		Constant resistance		Constant thickness	
Interconnect parameters	**Local**	**Global**	**Local**	**Global**	**Local**	**Global**	**Local**	**Global**
Length (L_{int})	$1/S$	S_c						
Width (W_{int})	$1/S$	$1/S$						
Thickness (T_{int})	$1/S$	$1/S$						
Height (H)	$1/S$	$1/S$						
Spacing (W_{spa})	$1/S$	$1/S$						
Aspect ratio (AR)	1	1						
Resistance	S	$S_c S^2$						
Coupling capacitance (C^c)	$1/S$	S_c						
Ground capacitance (C^g)	$1/S$	S_c						
C^c/C^g	1	1						
RC delay	1	$S_c^2 S^2$						