

Digital Integrated Circuits (83-313)

Lecture 6: Interconnect



Emerging Nanoscaled
Integrated Circuits and Systems Labs

Dr. Adam Teman
4 June 2020



Bar-Ilan University
אוניברסיטת בר-אילן

Disclaimer: This course was prepared, in its entirety, by Adam Teman. Many materials were copied from sources freely available on the internet. When possible, these sources have been cited; however, some references may have been cited incorrectly or overlooked. If you feel that a picture, graph, or code example has been copied from you and either needs to be cited or removed, please feel free to email adam.teman@biu.ac.il and I will address this as soon as possible.

Lecture Content

A First Glance at Interconnect

Capacitance

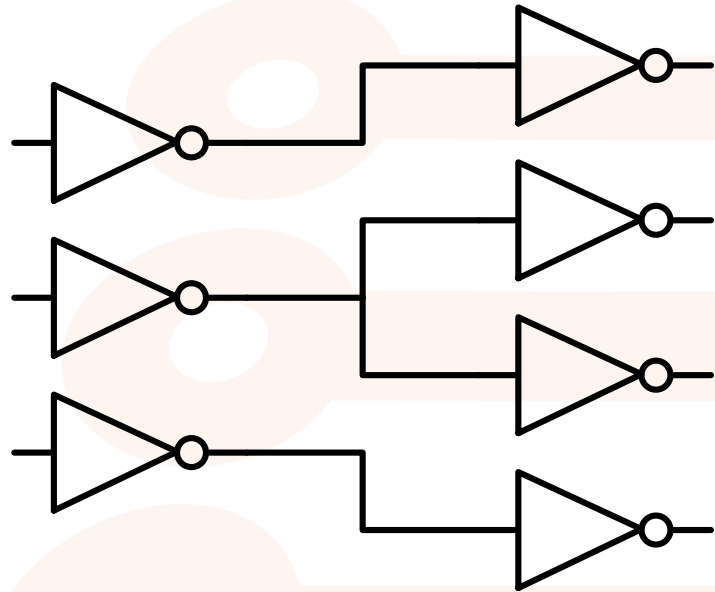
Resistance

Interconnect Modeling

Wire Scaling

A First Glance at Interconnect

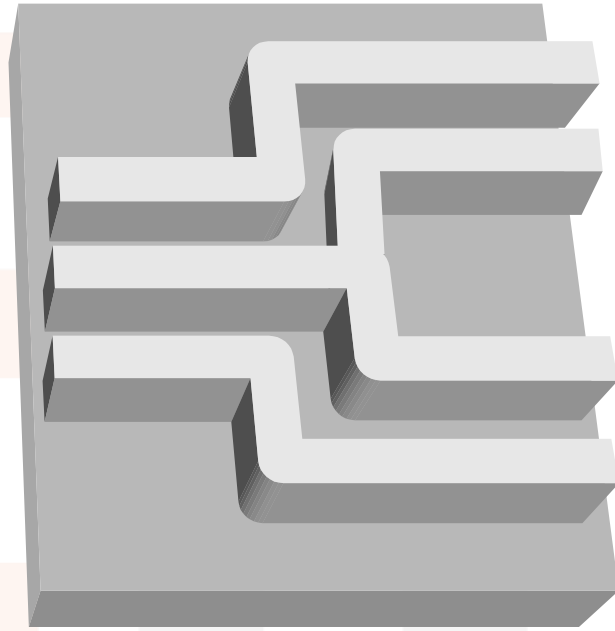
The Wire



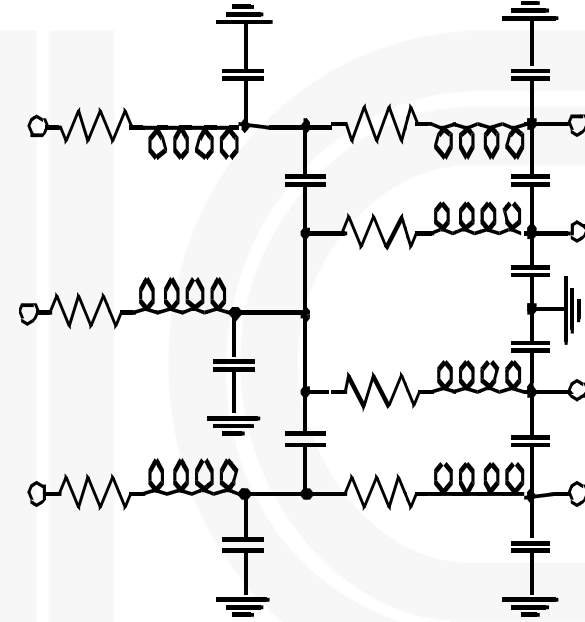
Transmitters

Receivers

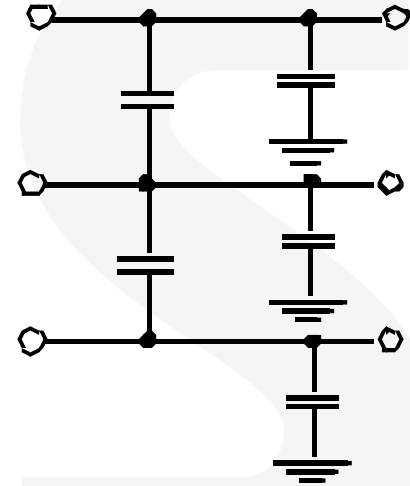
schematic view



physical realization

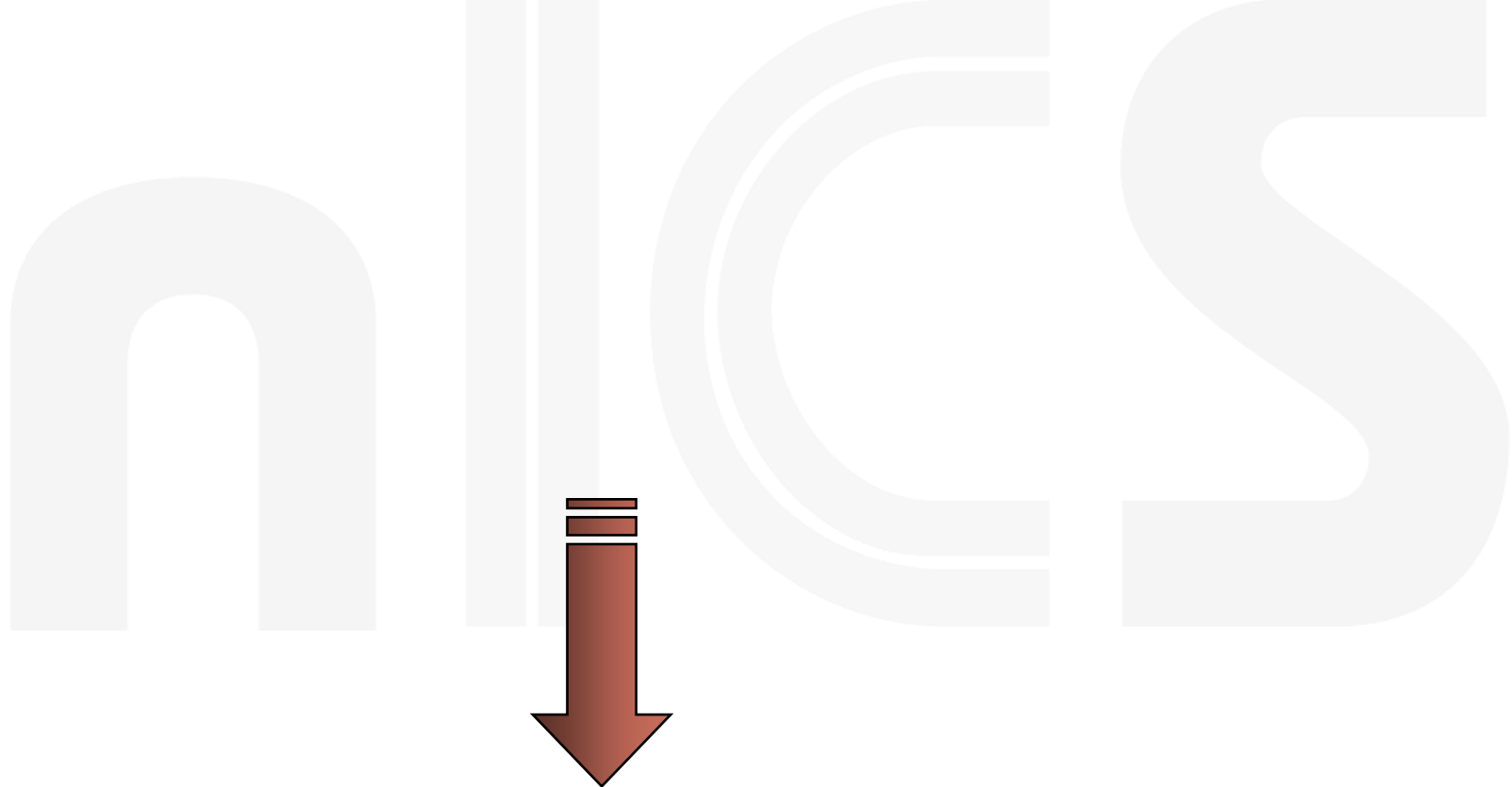


All-inclusive model

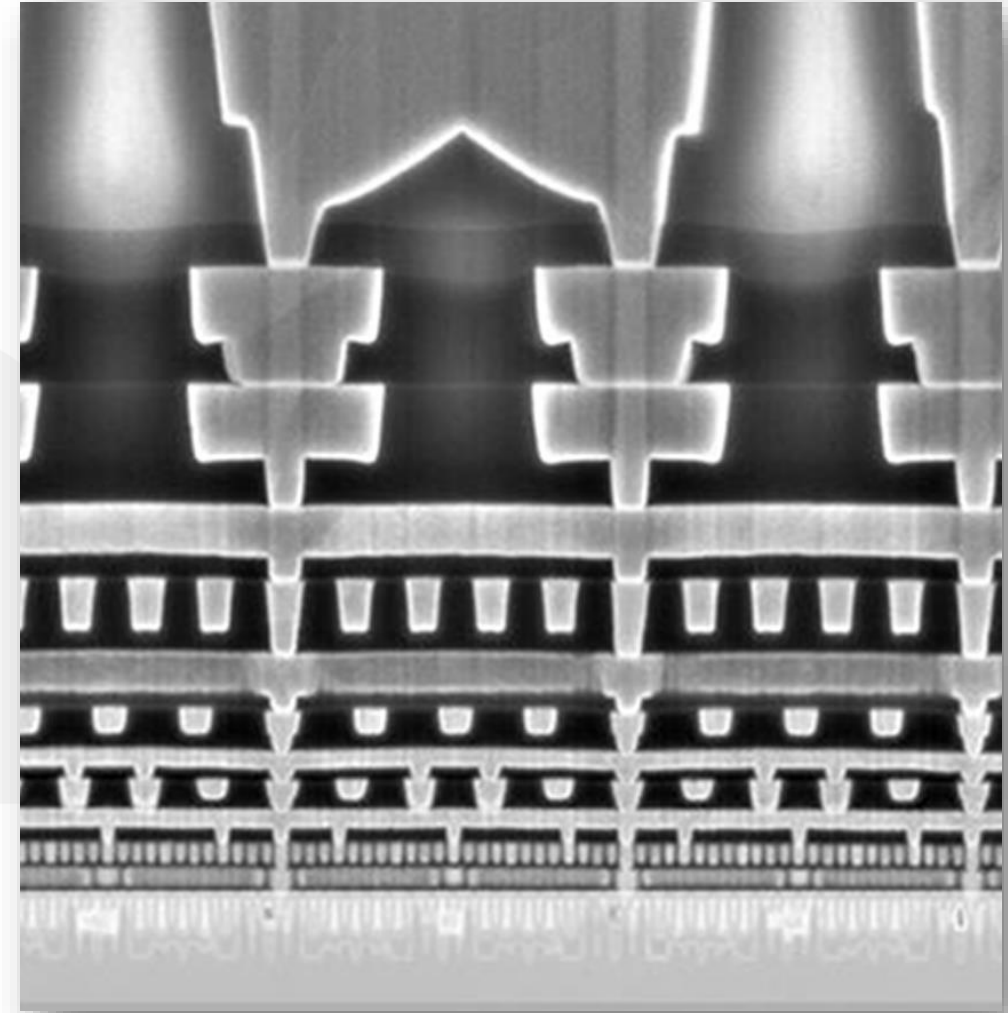
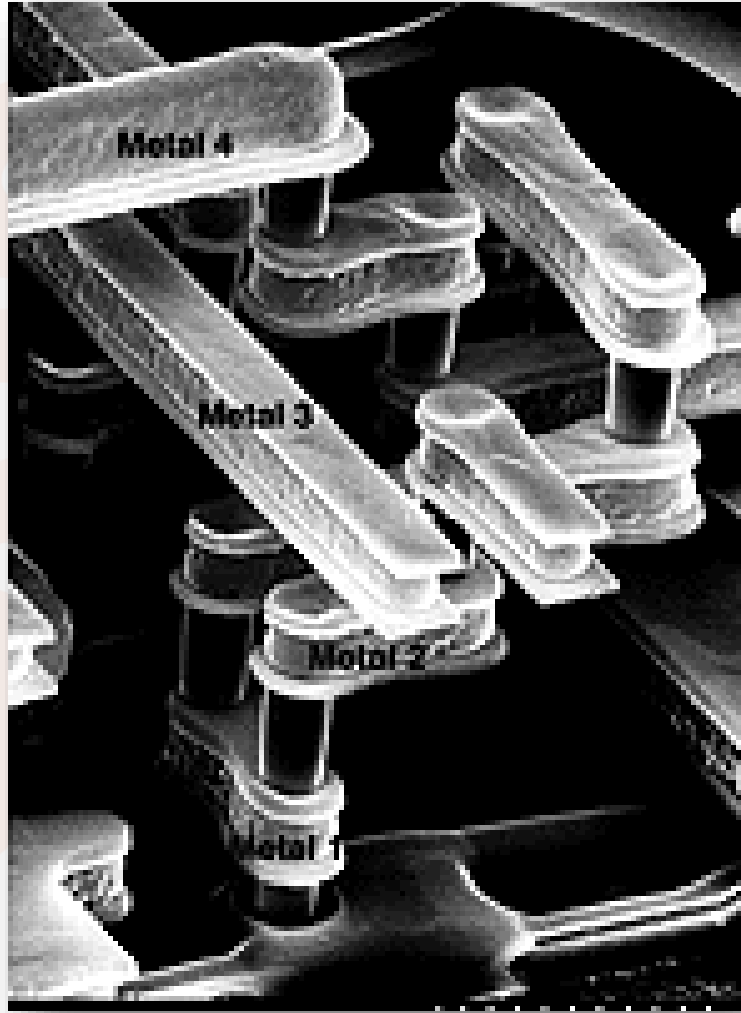


Capacitance-only

Impact of Interconnect Parasitics

- **Interconnect parasitics affect all the metrics we care about**
 - Reliability
 - Performance
 - Power Consumption
 - Cost
 - **Classes of parasitics**
 - Capacitive
 - Resistive
 - Inductive
- 

Modern Interconnect



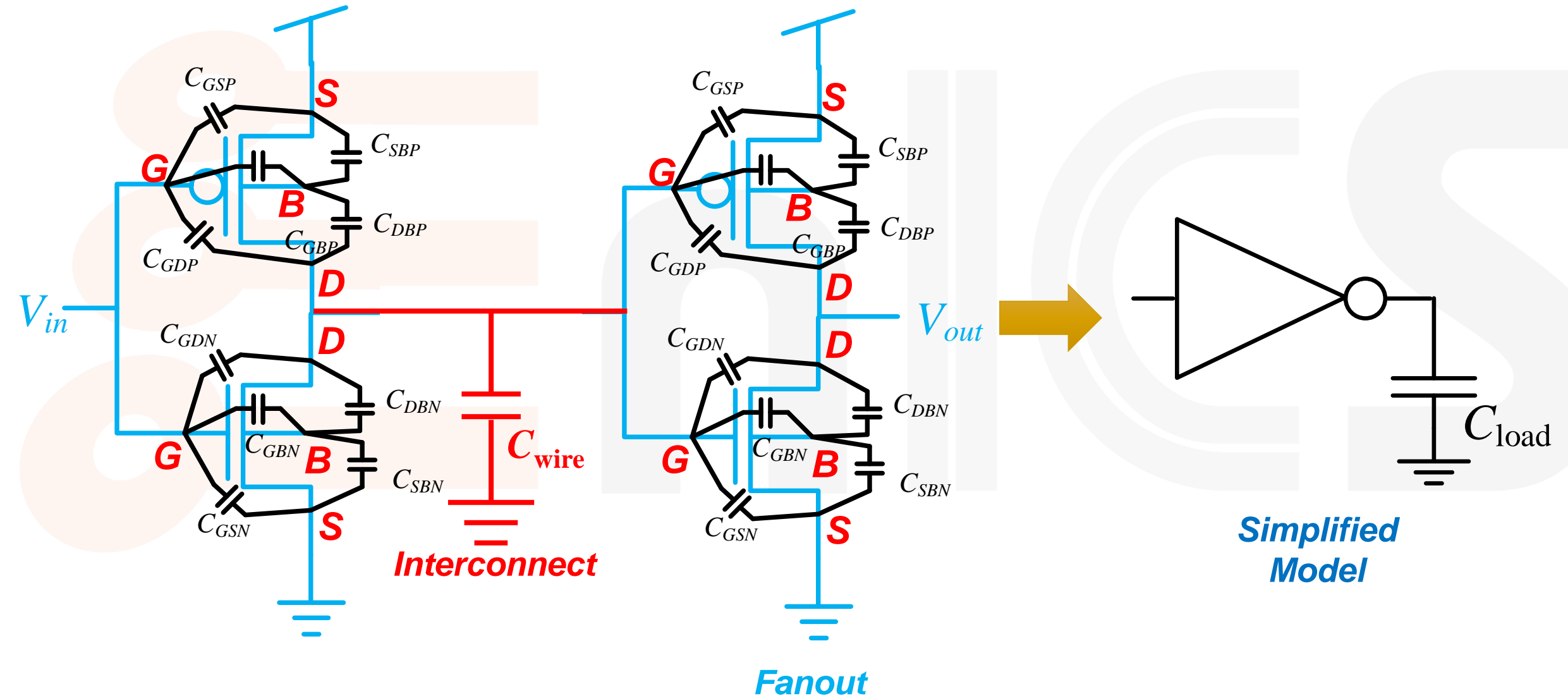
Intel 10nm Interconnect Stack

Source: Intel, IEDM 2017

© Adam Teman, 2020

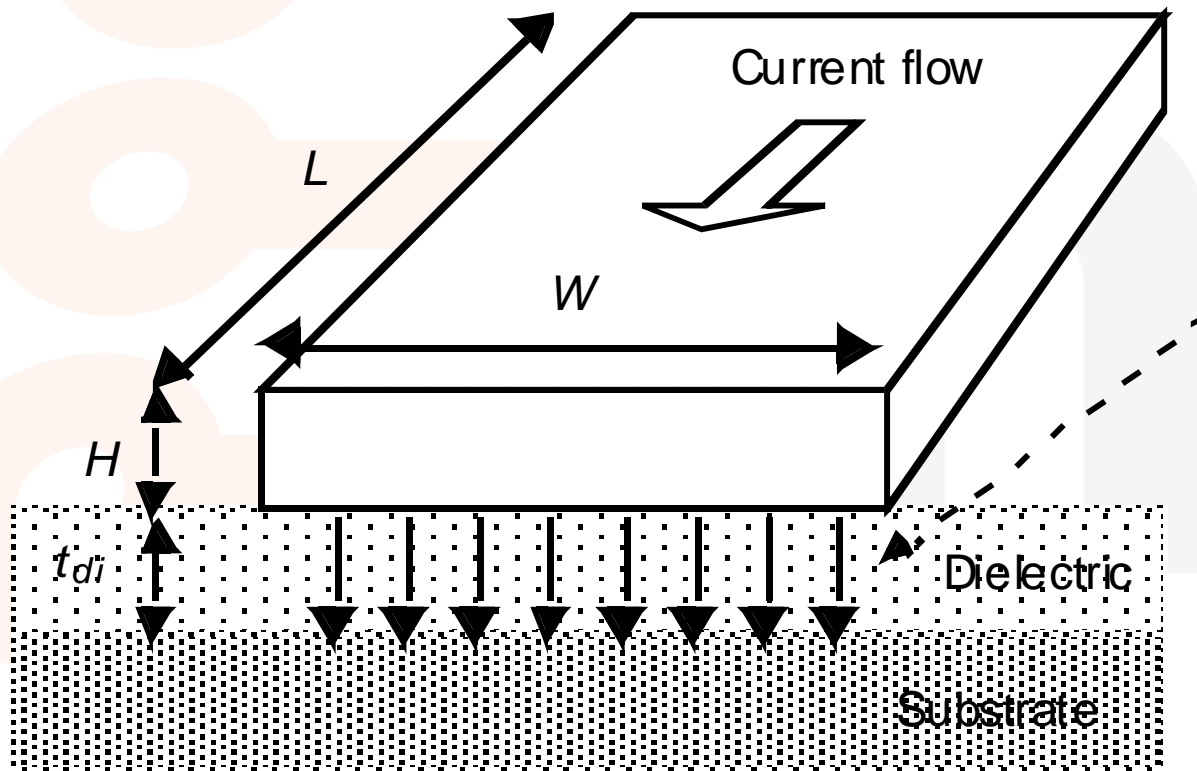
Capacitance

Capacitance of Wire Interconnect



Capacitance: The Parallel Plate Model

- How can we reduce this capacitance?



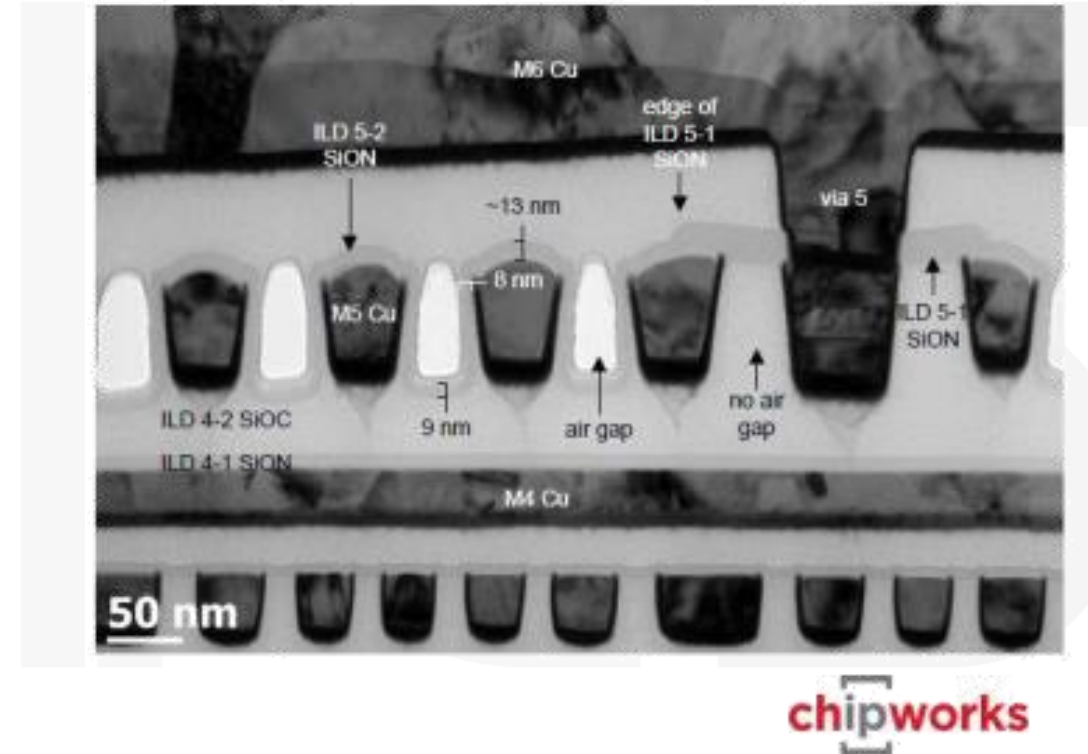
Typical numbers:

- Wire cap ~ 0.2 fF/ μm
- Gate cap ~ 2 fF/ μm
- Diffusion cap ~ 2 fF/ μm

$$C_{pp} = \frac{\epsilon_{di}}{t_{di}} WL$$

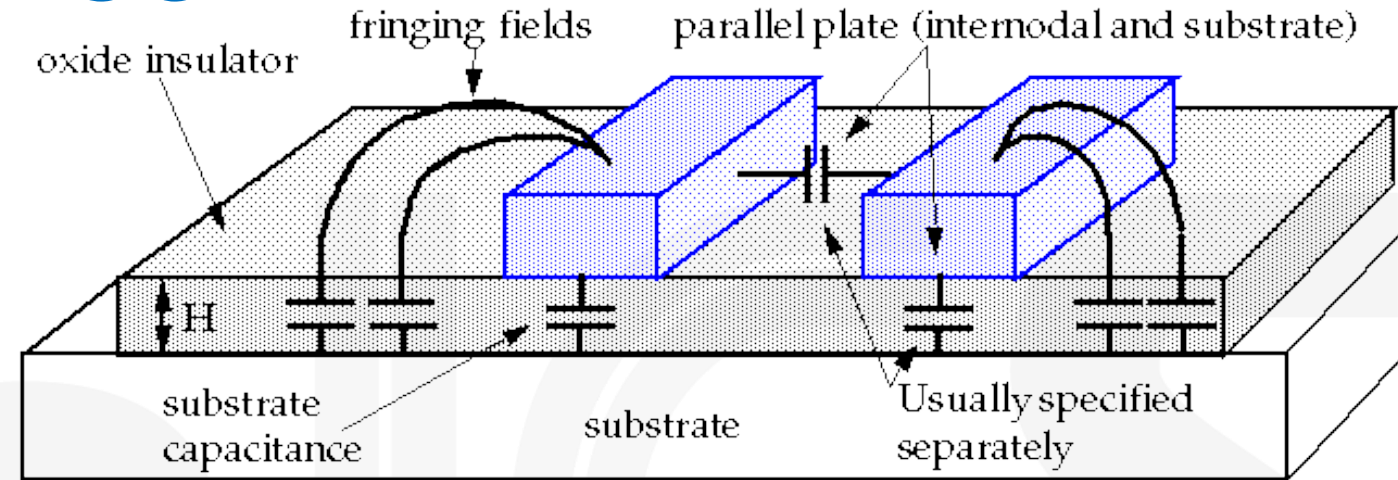
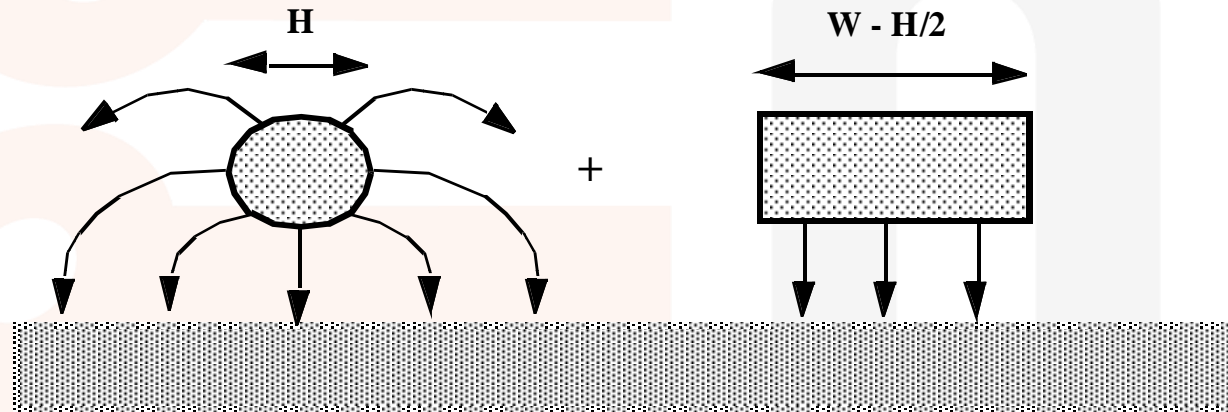
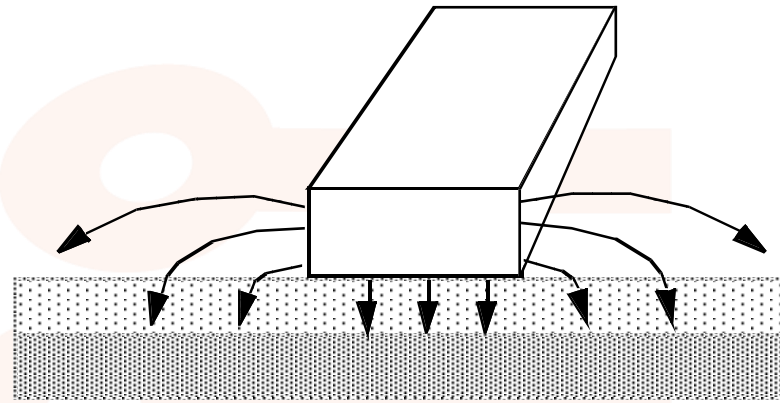
Permittivity

Material	ϵ_r
Free space	1
Aerogels	~ 1.5
Polyimides (organic)	3-4
Silicon dioxide	3.9
Glass-epoxy (PC board)	5
Silicon Nitride (Si_3N_4)	7.5
Alumina (package)	9.5
Silicon	11.7



Air Gaps (implemented in Intel 14nm)

Fringing Capacitance



$$w = W - H / 2$$

$$C [\text{F/mm}] = c_{pp} + c_{fringe} = \frac{(W - H/2) \epsilon_{di}}{t_{di}} + \frac{2\pi \epsilon_{di}}{\log(t_{di}/H)}$$

Fringing versus Parallel Plate



$$C_{fringe}/edge \approx 0.05 \text{ fF}/\mu\text{m}$$



$$C_{fringe} \propto L$$

$$C_{PP} \propto W \cdot L$$

A simple model for deriving wire cap

- Wiring capacitances in 0.25μm

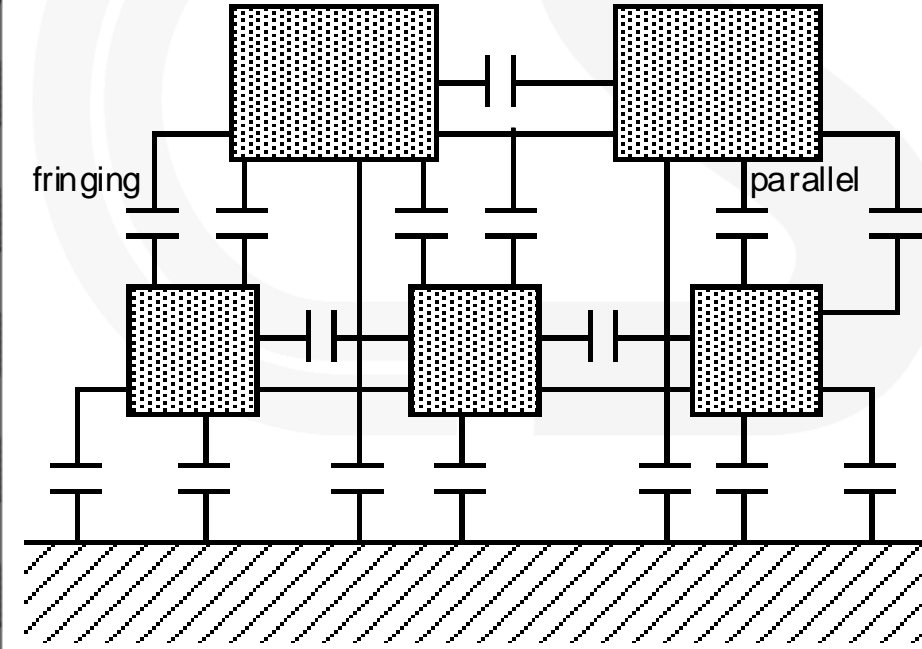
$$C_{wire} = C_{parallel_plate} \cdot W \cdot L + 2 \cdot C_{fringe} \cdot L$$

Bottom Plate

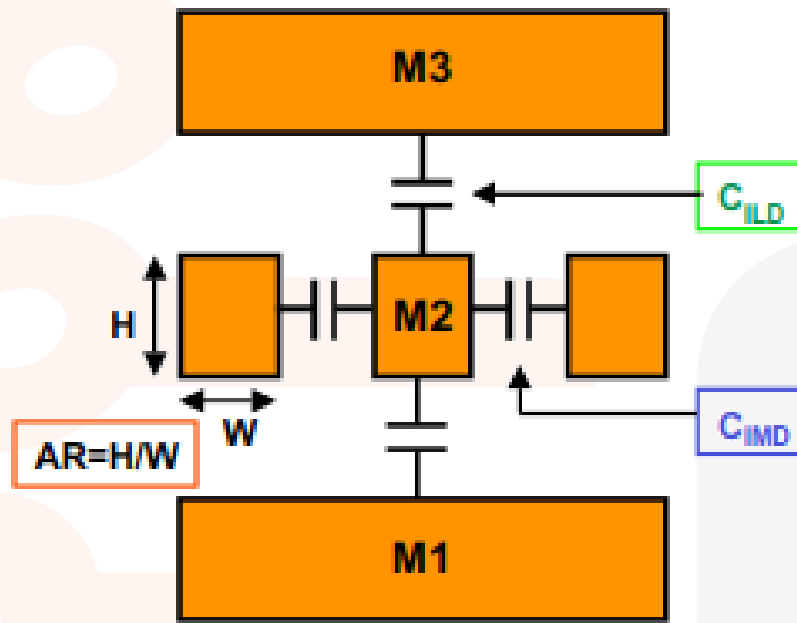
Top Plate

	Field	Active	Poly	Al1	Al2	Al3	Al4
Poly	88						
Al1	30	41	57				
Al2	13	15	17	36			
Al3	8.9	9.4	10	15	41		
Al4	6.5	6.8	7	8.9	15	35	
Al5	5.2	5.4	5.4	6.6	9.1	14	38

$aF/\mu m^2$ (pointing to 88)
 $aF/\mu m$ (pointing to 54)

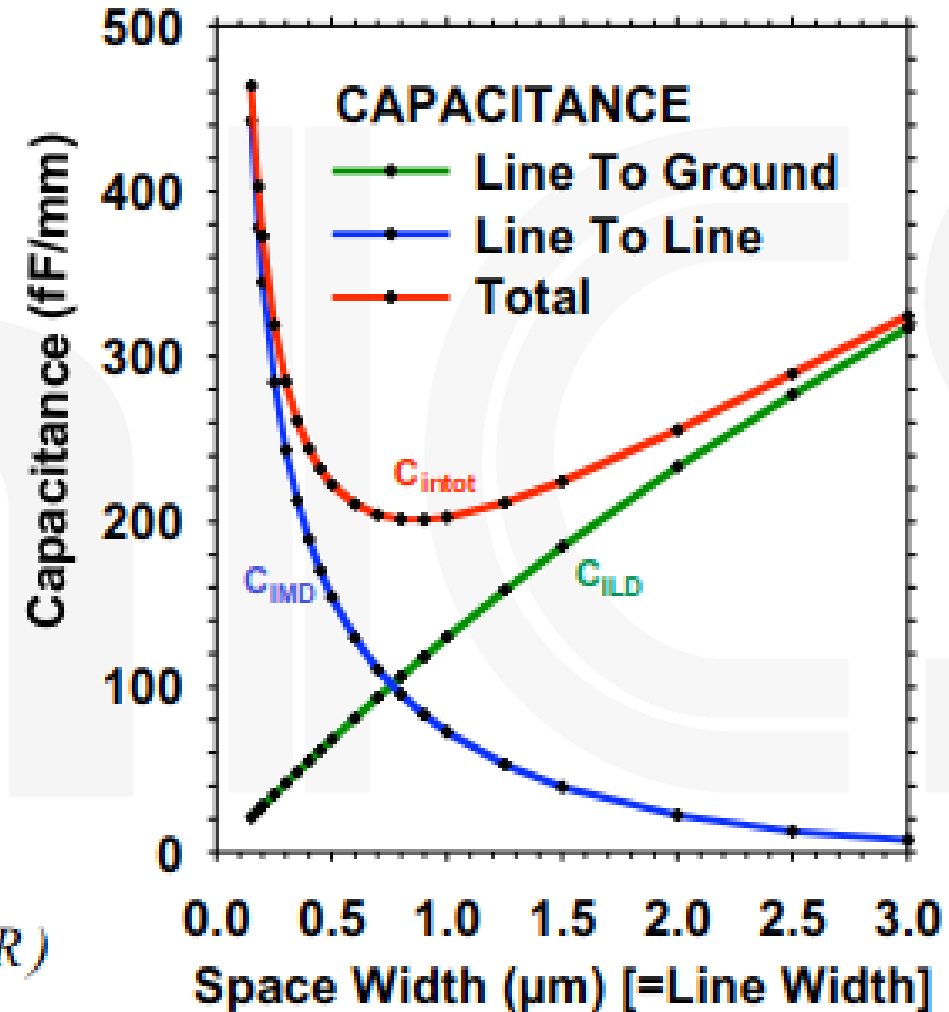


Impact of Interwire Capacitance

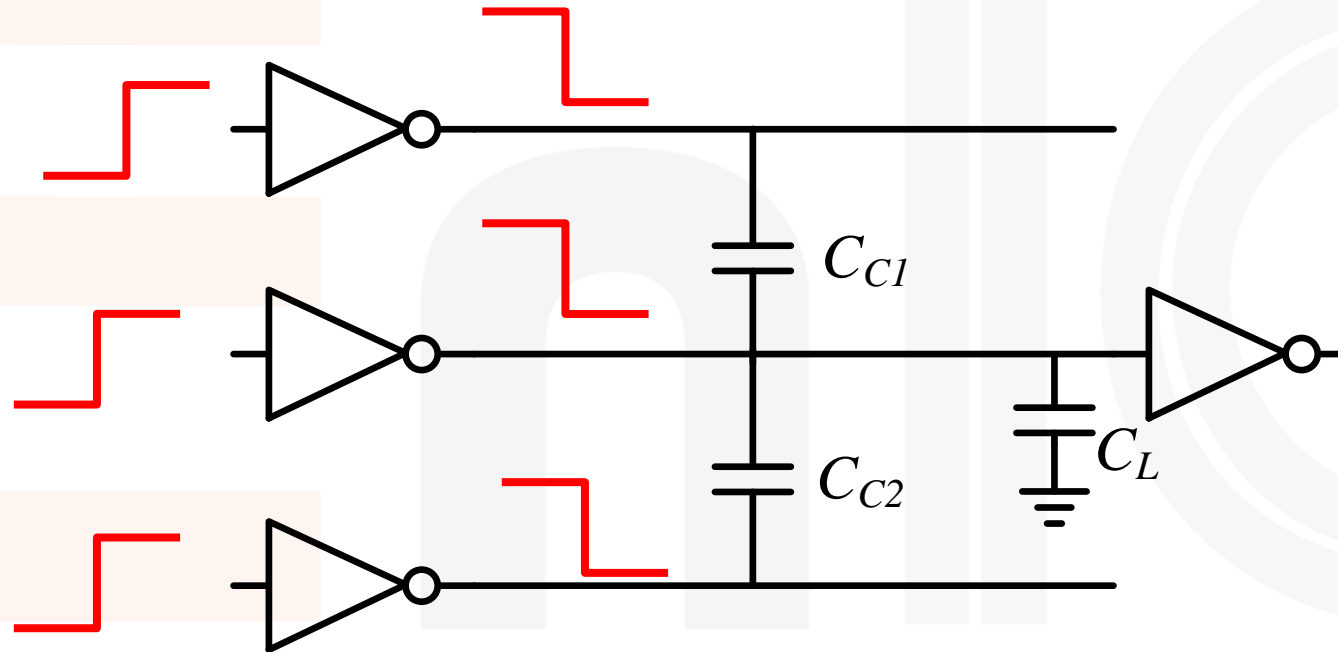


In general

$$C_{inttot} = C_{ILD} + C_{IMD} = 2l \left(\frac{\epsilon_{ILD}}{AR} + \epsilon_{IMD} AR \right)$$

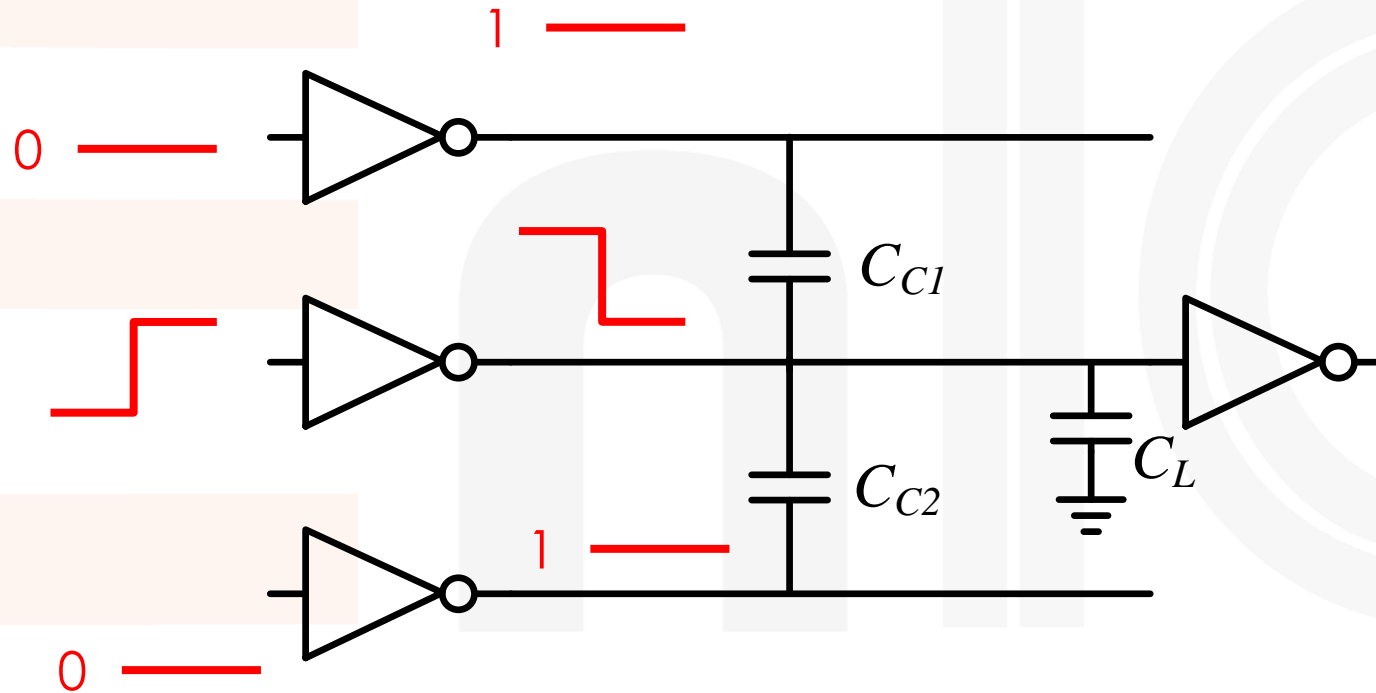


Coupling Capacitance and Delay



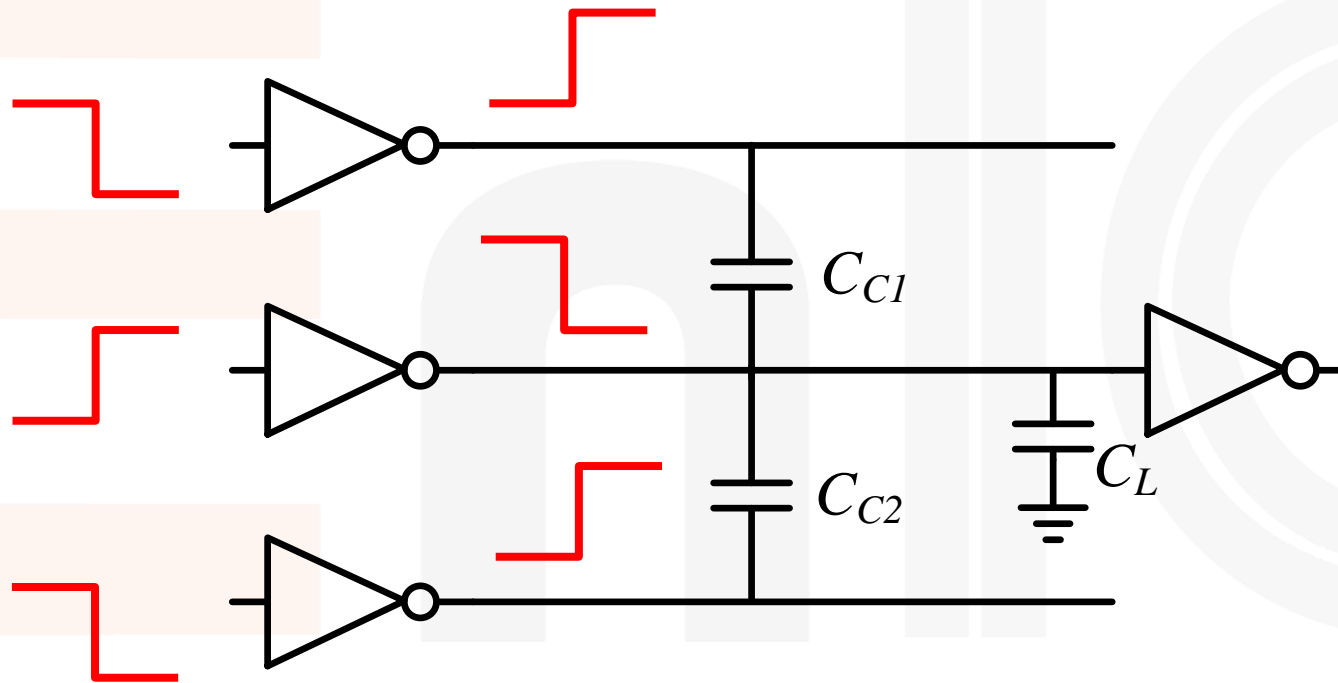
$$C_{tot} = C_L$$

Coupling Capacitance and Delay



$$C_{tot} = C_L + C_{C1} + C_{C2}$$

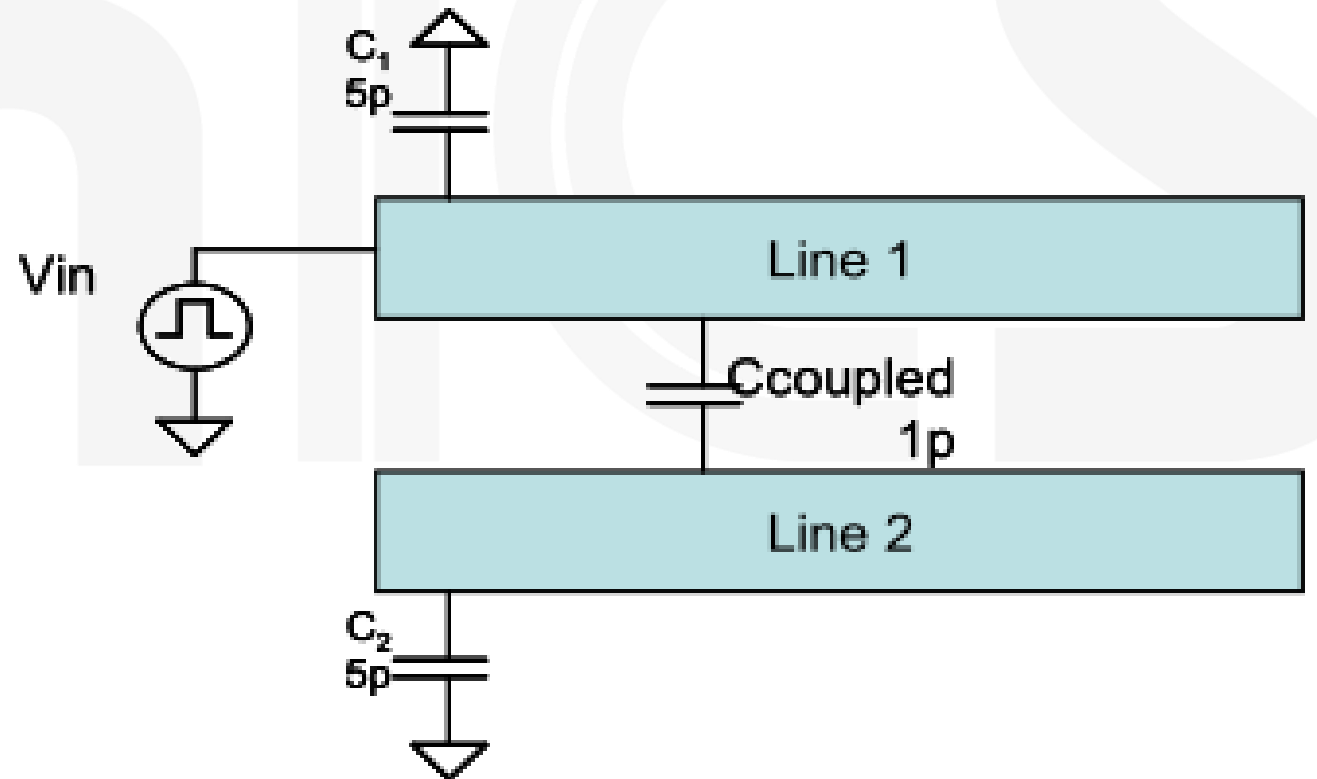
Coupling Capacitance and Delay



$$C_{tot} = C_L + 2(C_{C1} + C_{C2})$$

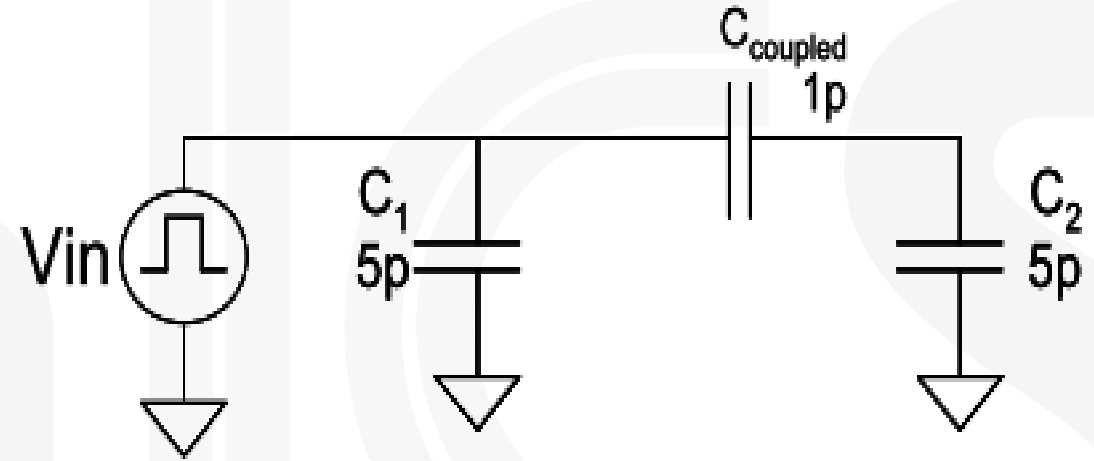
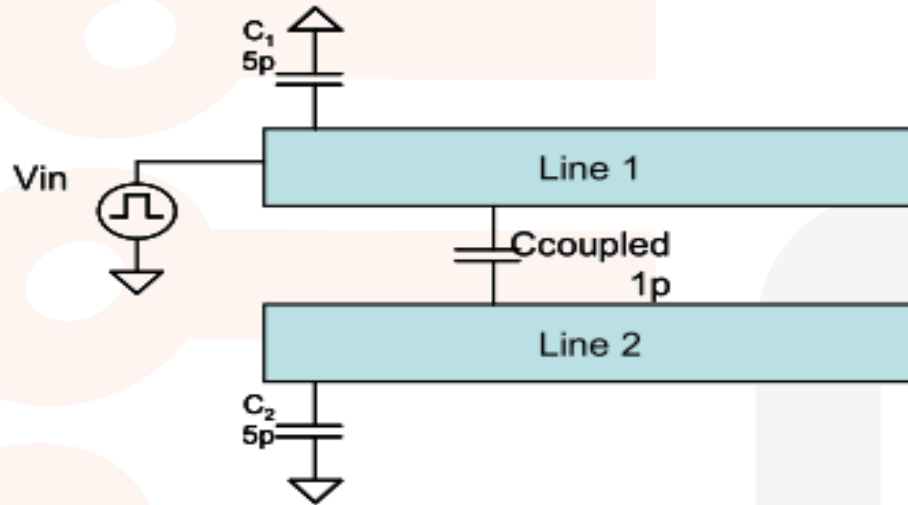
Example – Coupling Cap

- A pair of wires, each with a **capacitance to ground** of **5pF**, have a **1pF coupling capacitance** between them.
- A **square pulse** of **1.8V** (relative to ground) is connected to one of the wires.
- **How high** will the **noise pulse** be on the other wire?



Example – Coupling Cap

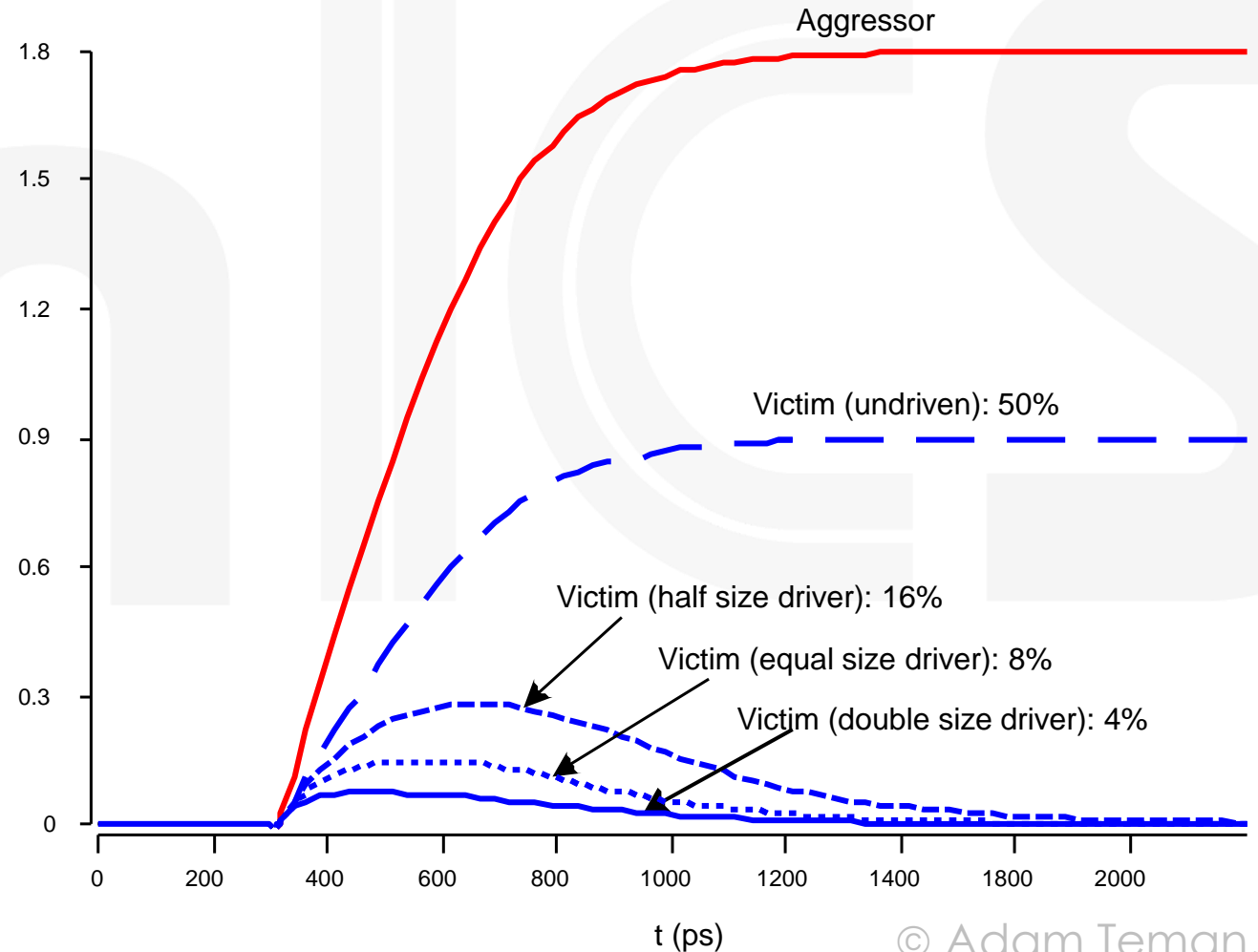
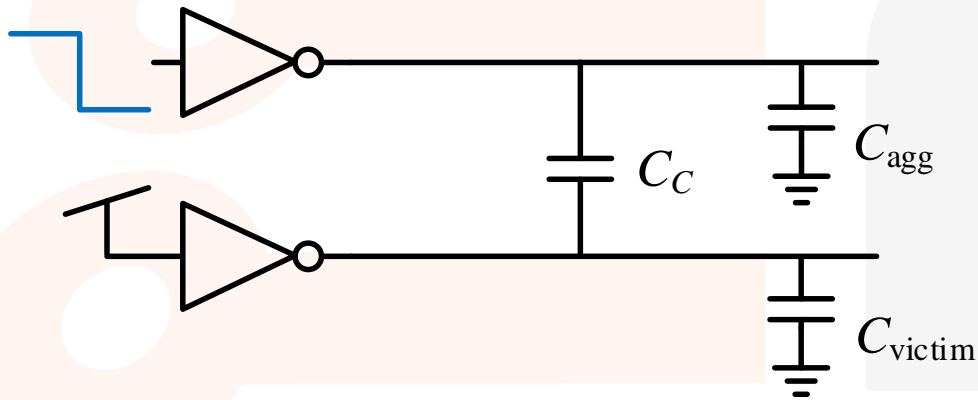
- Draw an Equivalent Circuit:



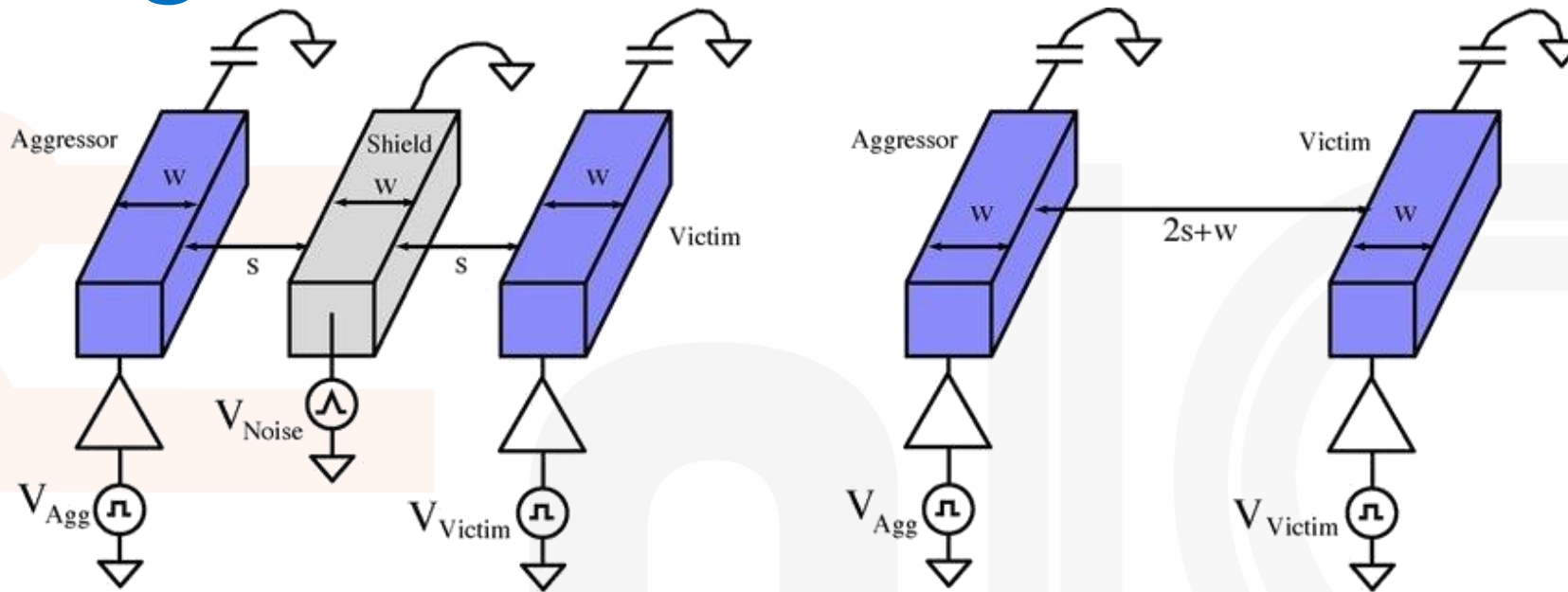
$$V_{C2} = \frac{V_{in} \cdot C_{coupled}}{C_{coupled} + C_2} = \frac{1.8 \cdot 1p}{1p + 5p} = 0.3V$$

Coupling Waveforms

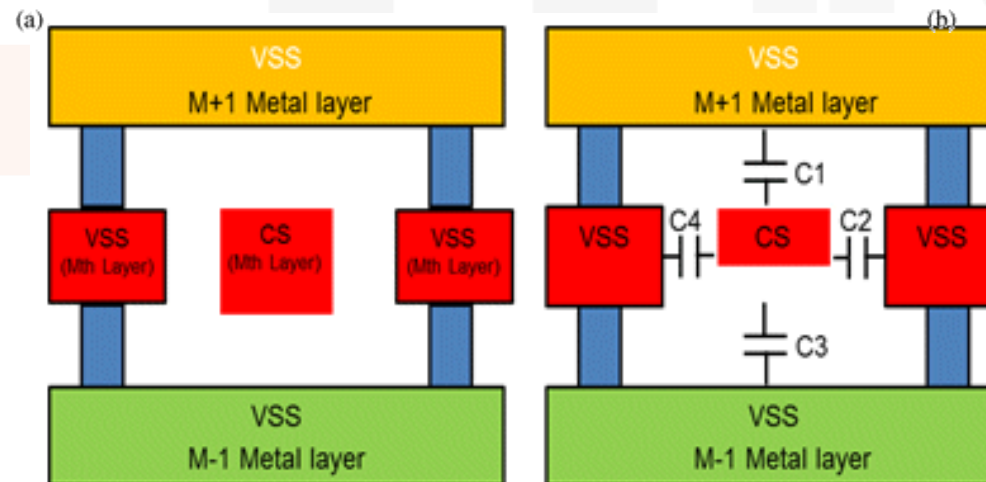
- Simulated coupling for $C_{\text{agg}} = C_{\text{victim}}$



Shielding



Source: Kose, et al



Source: Design-reuse.com

Feedthrough Cap



Measuring Capacitance

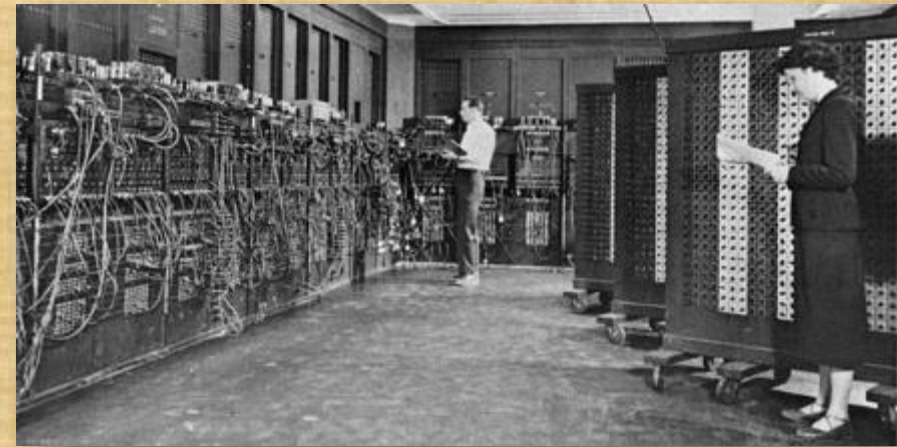


The Computer Hall of Fame

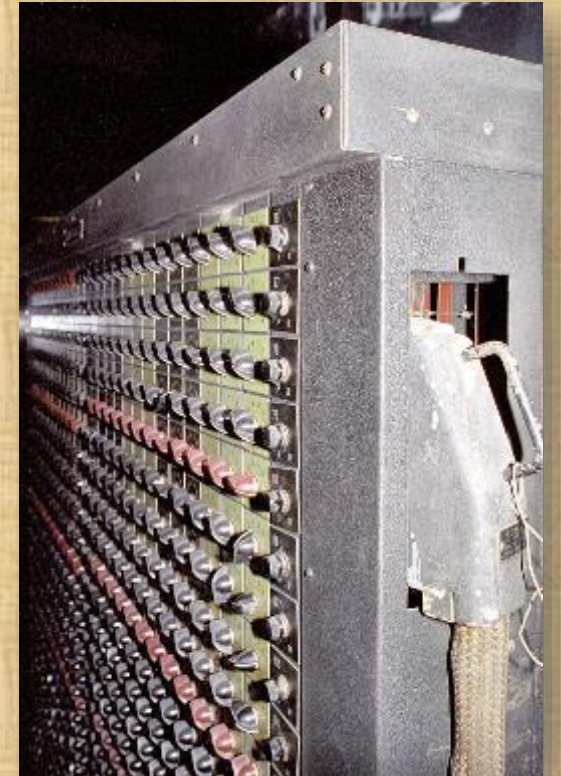
- The first computer or not the first computer?

ENIAC

- Electronic Numerical Integrator and Calculator
- 20,000 vacuum tubes, 1500 relays, 10,000 capacitors, 70,000 resistors
- 200kW, 30 tons, cost almost \$500K
- Completed in 1946 at the University of Pennsylvania by John Eckert and John Mauchly
- Used to calculate artillery firing tables for the US Army's Ballistic Research Laboratory
- Lost the patent for being the first computer in 1973, but it was the first general purpose programmable computer.



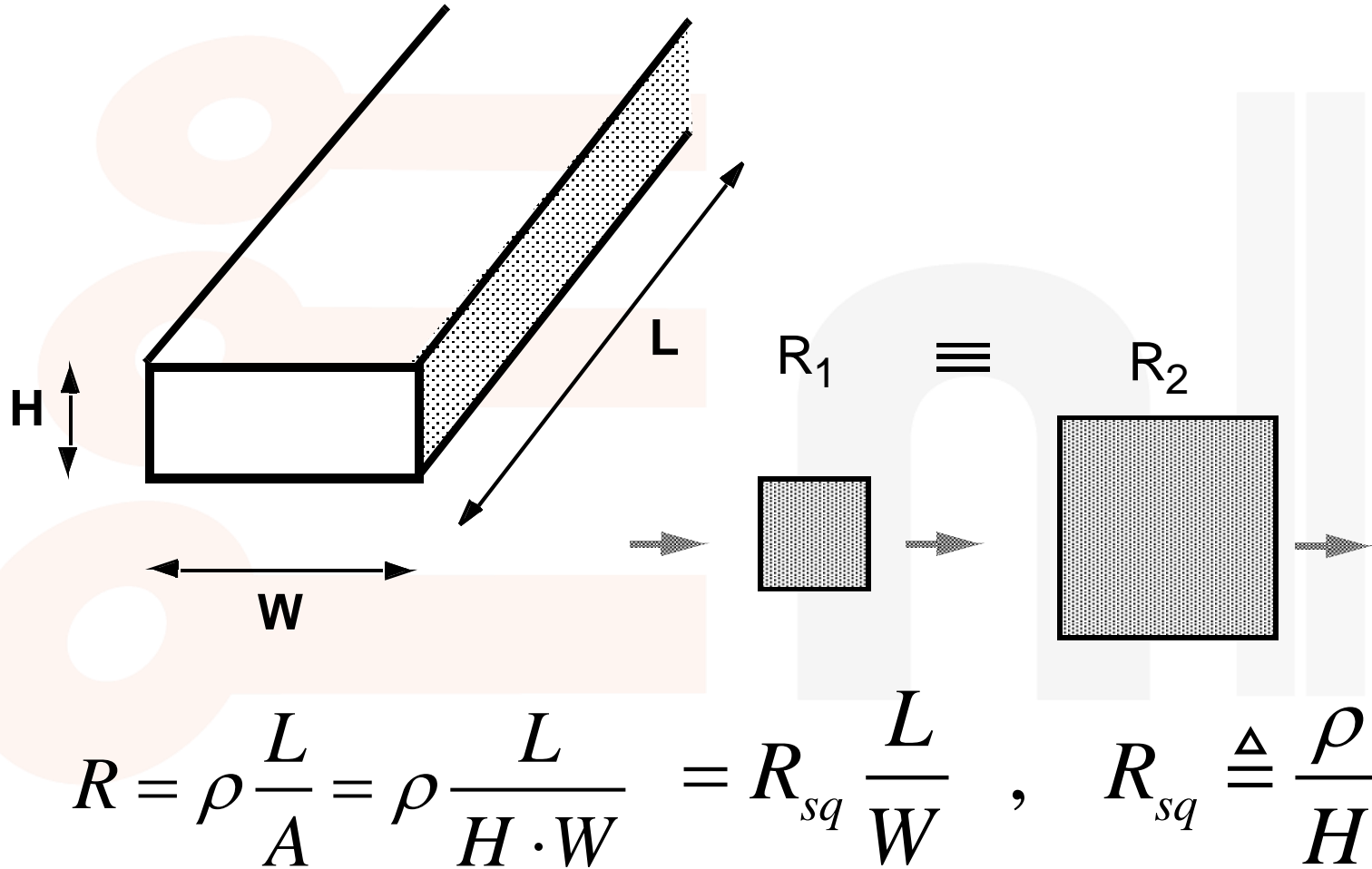
Source: US Army.



Source: Wikipedia

Resistance

Wire Resistance

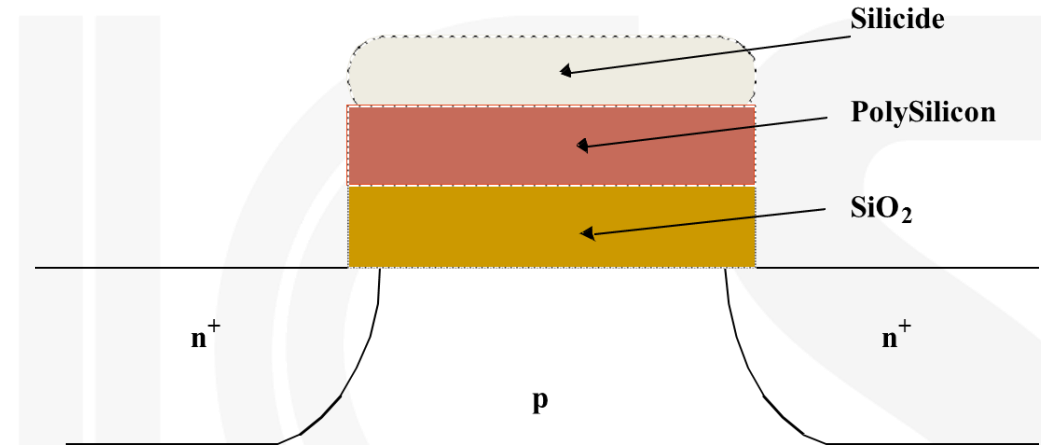


Metal	Bulk resistivity ($\mu\Omega \cdot \text{cm}$)
Silver (Ag)	1.6
Copper (Cu)	1.7
Gold (Au)	2.2
Aluminum (Al)	2.8
Tungsten (W)	5.3
Molybdenum (Mo)	5.3

Sheet Resistance

- Typical sheet resistances for 180nm process

Layer	Sheet Resistance (Ω/\square)
N-Well/P-Well	1000-1500
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



Silicides: WSi₂, TiSi₂, PtSi₂ and TaSi

Conductivity: 8-10 times better than Poly

$$R_{\square} \approx 100 \frac{m\Omega}{square}$$

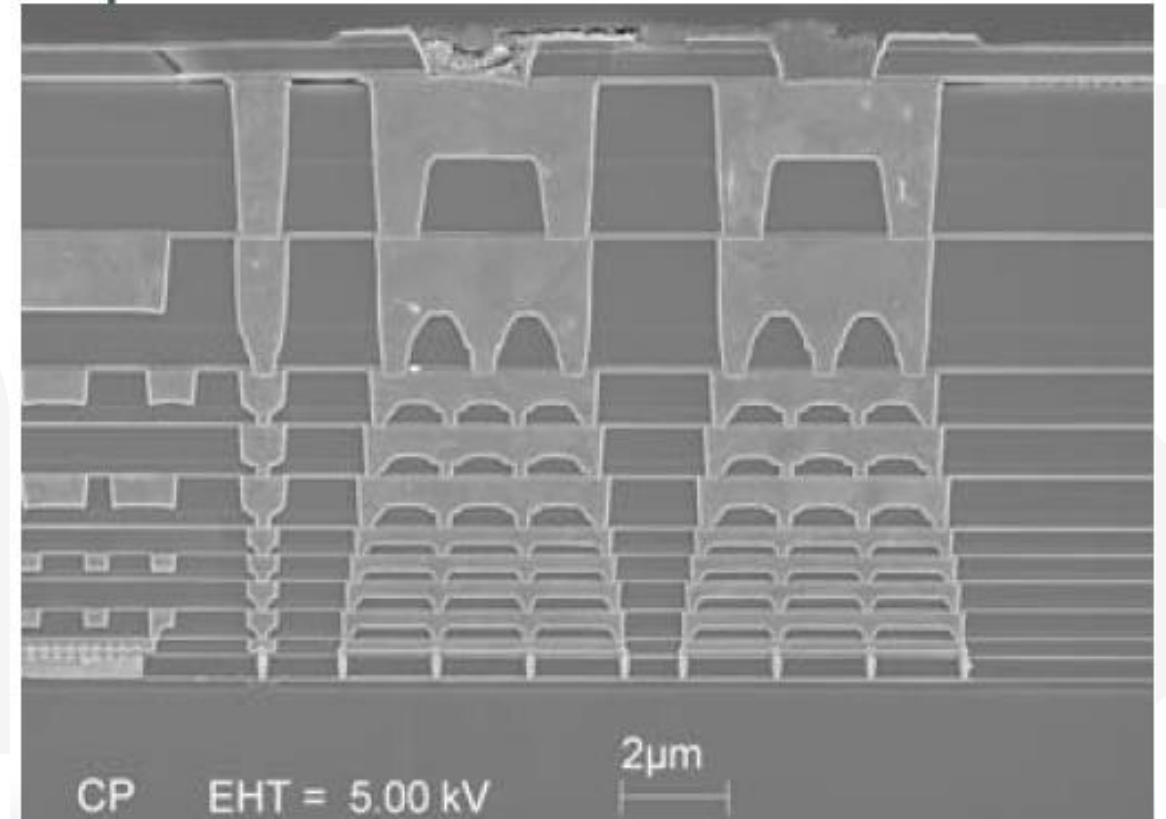
Contact Resistance

- **Contact/Vias add extra resistance**
 - Similar to changing between roads on the way to a destination...
 - Contact resistance is generally 2-20 Ω
- **Make contacts bigger**
 - BUT... current “crowds” around the perimeter of a contact.
 - There are also problems in deposition...
 - Contacts/Vias have a maximum practical size.
 - Use multiple contacts
 - But does this add overlap capacitance?



Dealing with Resistance

- **Selective Technology Scaling**
 - Don't scale the H
- **Use Better Interconnect Materials**
 - reduce average wire-length
 - e.g. copper, silicides
- **More Interconnect Layers**
 - reduce average wire-length
- **Minimize Contact Resistance**
 - Use single layer routing
 - When changing layers, use lots of contacts.



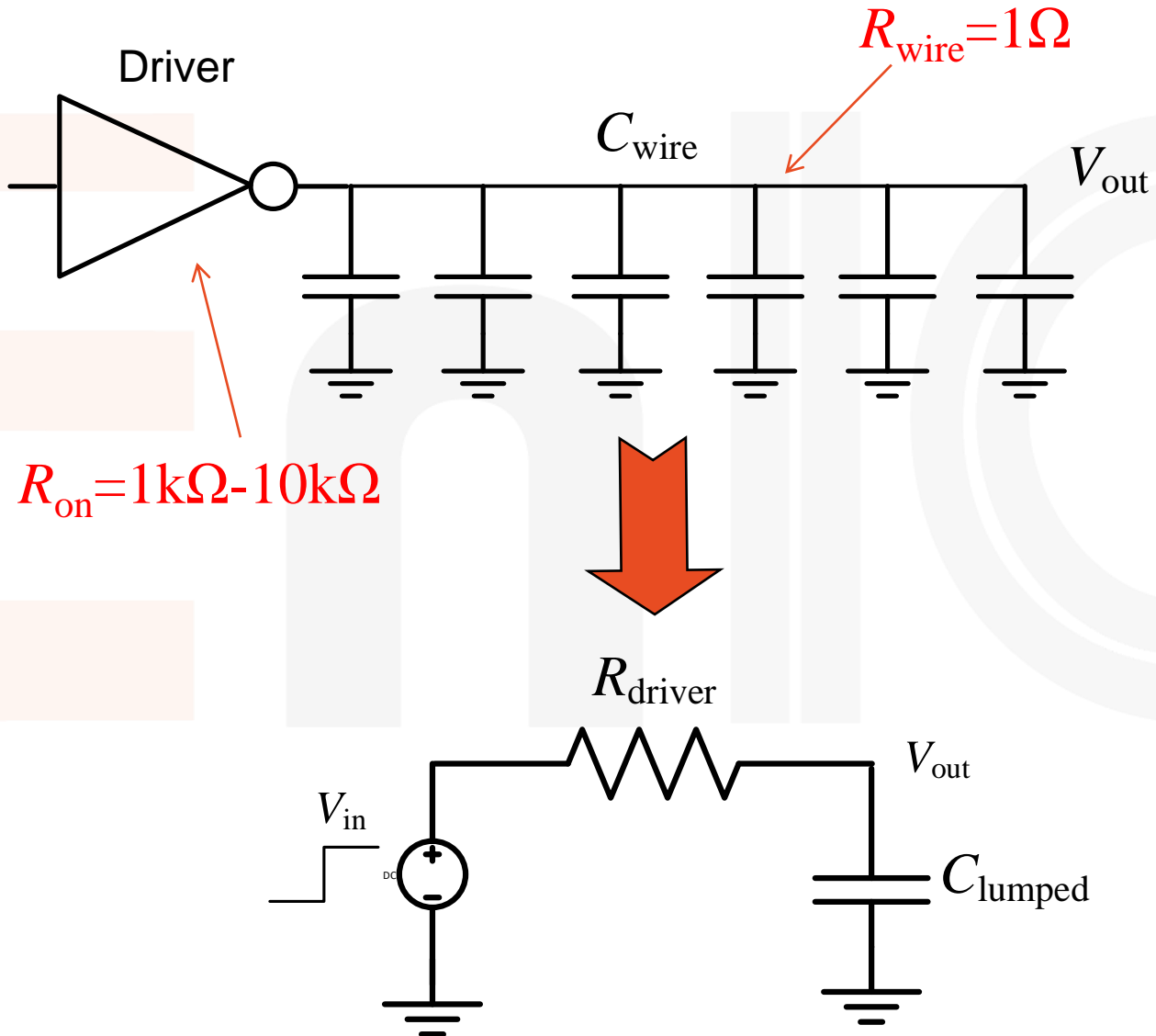
90nm Process

Interconnect Modeling

The Ideal Model

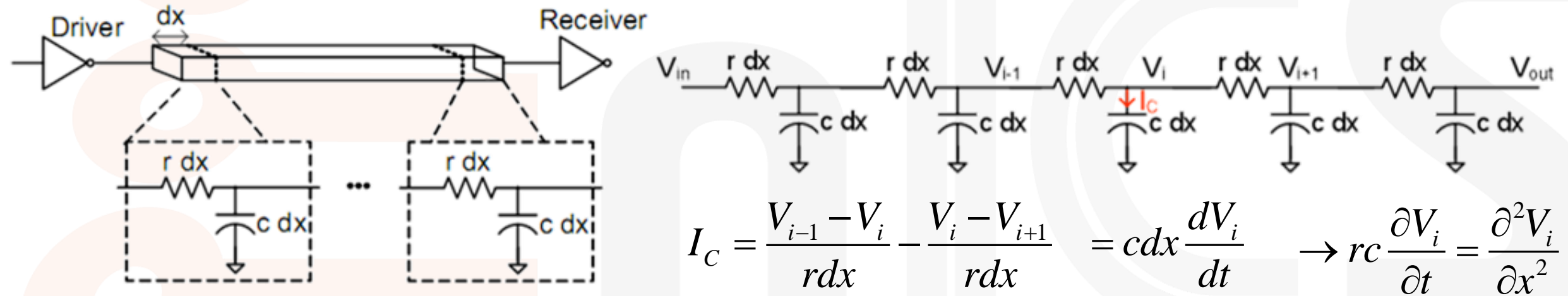
- In schematics, a wire has no parasitics:
 - The wire is a single equipotential region.
 - No effect on circuit behavior.
 - Effective in first stages of design and for very short wires.

The Lumped Model



The Distributed RC-line

- But actually, our wire is a distributed entity.
 - We can find its behavior by breaking it up into small RC segments.



$$\lim_{dx \rightarrow 0} \frac{f(x+dx) - f(x)}{dx} = f'(x)$$

$$I_c = \frac{V_{i-1} - V_i}{r dx} - \frac{V_i - V_{i+1}}{r dx} = c dx \frac{dV_i}{dt} \rightarrow rc \frac{\partial V_i}{\partial t} = \frac{\partial^2 V_i}{\partial x^2}$$

$$\tau = \frac{rc}{2} L^2$$

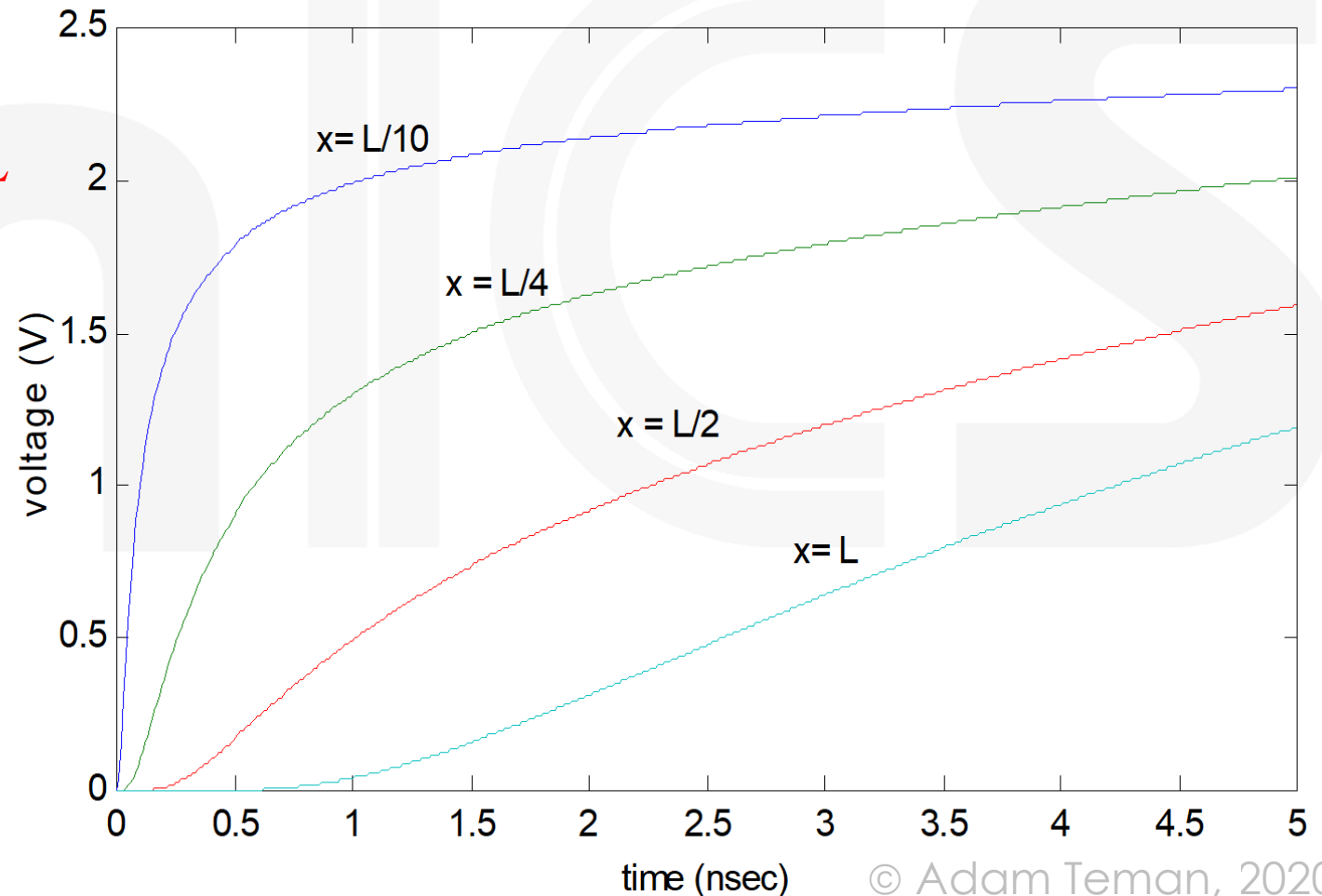
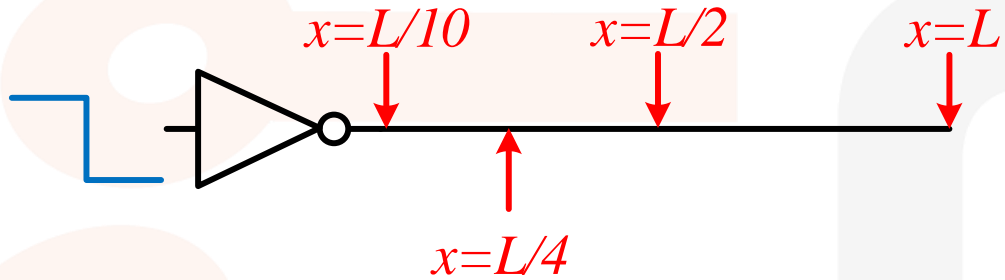
$$t_{pd} = 0.38RC$$

Quadratic dependence
on wire length

The lumped model is
pessimistic

Step-response of RC wire

- Step-response of RC wire as a function of time and space

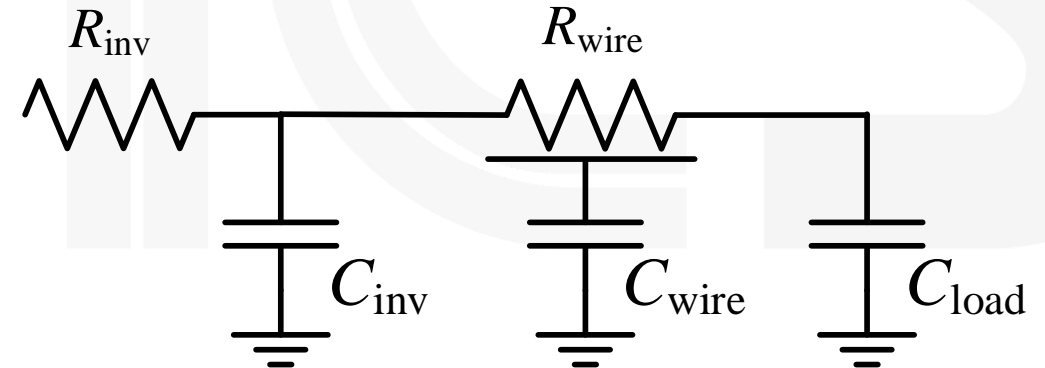
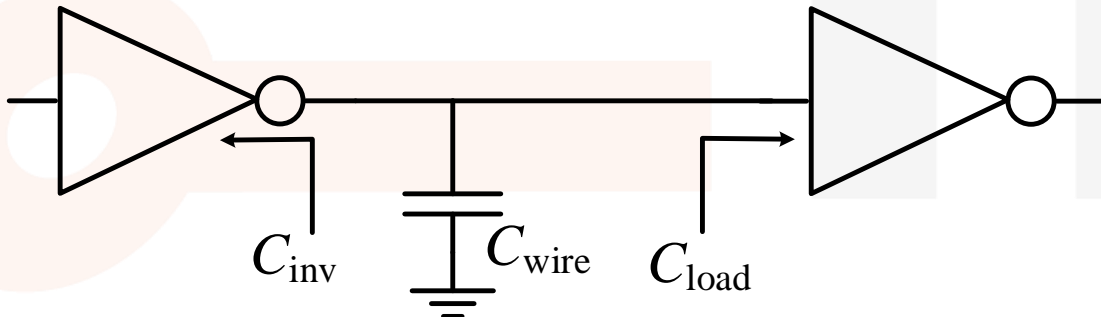


Modeling the distributed Cap

- Let's look at a driver with a distributed wire and an output load.
 - For the driver resistance, we can lump the output load as a capacitor.
 - For the wire resistance, we will use the distributed time constant.
 - For the load capacitance, we can lump the wire and driver resistance.

$$C_w \approx 0.2 \frac{\text{fF}}{\mu\text{m}}$$

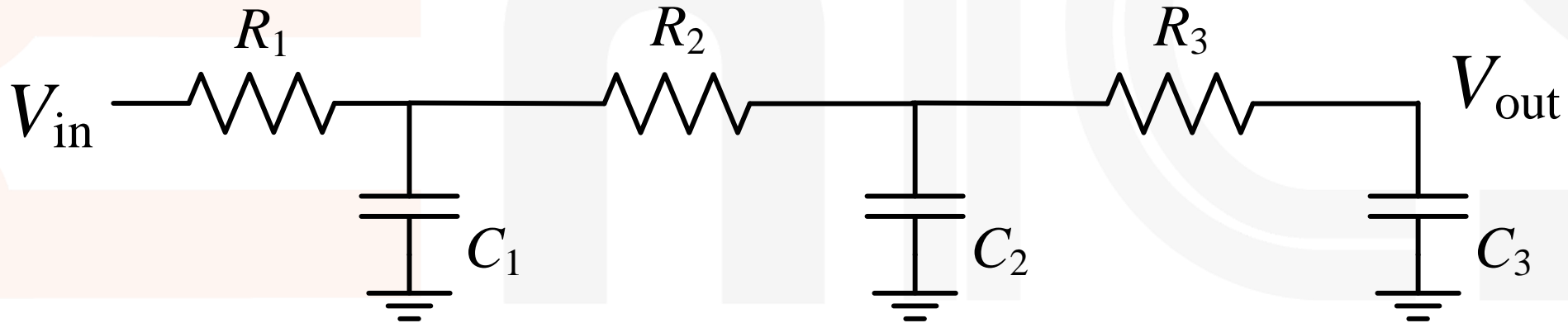
$$R_{\square} \approx 0.1 \frac{\Omega}{\square}$$



$$\tau_D = 0.69 R_{\text{inv}} (C_{\text{inv}} + C_{\text{wire}}) + 0.38 R_{\text{wire}} C_{\text{wire}} + 0.69 (R_{\text{inv}} + R_{\text{wire}}) C_{\text{load}}$$

Elmore Delay Approximation

- Solving the diffusion equation for a given network is complex.
- Elmore proposed a reasonably accurate method to achieve an approximation of the dominate pole.



$$\tau_{elmore} = R_1 C_1 + (R_1 + R_2) C_2 + (R_1 + R_2 + R_3) C_3$$

Elmore Generalized Ladder Chain

- Lets apply the Elmore approximation for our original distributed wire.
 - Divide the wire into N equal segments of $dx=L/N$ length with capacitance cdx and resistance rdx .

$$\begin{aligned}\tau_N &= c \left(\frac{L}{N} \right) \left(r \frac{L}{N} + 2r \frac{L}{N} + \dots + Nr \frac{L}{N} \right) \\ &= \left(\frac{L}{N} \right)^2 (rc + 2rc + \dots + Nrc) = rcL^2 \left(\frac{N(N+1)}{2N^2} \right)\end{aligned}$$

$$\lim_{N \rightarrow \infty} \tau_D = \frac{rcL^2}{2} = \frac{RC}{2}$$

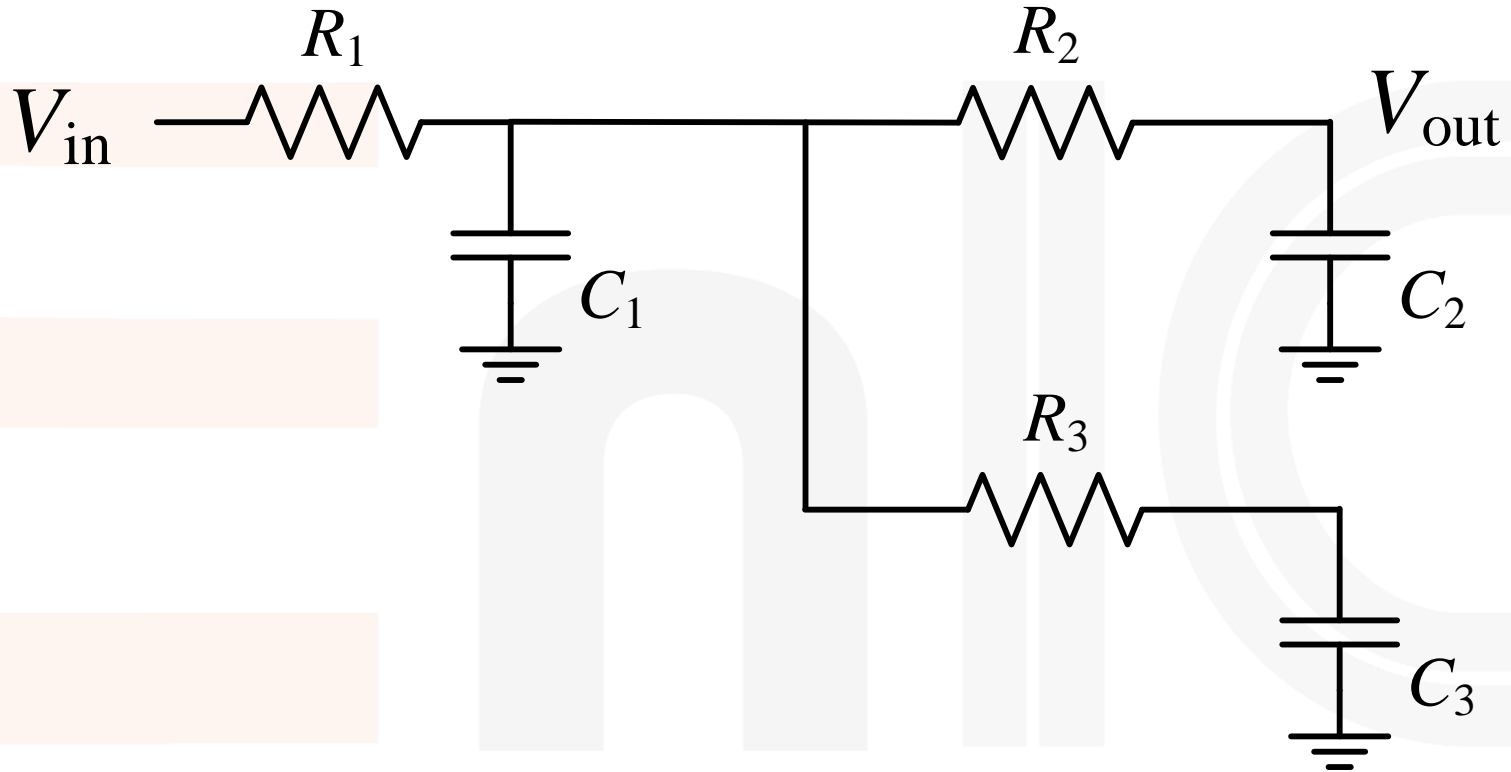
Elmore Delay Approximation

For a complex network use the following method:

- Find all the resistors on the path from in to out.
- For every capacitor:
 - Find all the resistors on the path from the input to the capacitor.
 - Multiply the capacitance by the resistors that are also on the path to out.
- The dominant pole is approximately the sum of all these time constants.

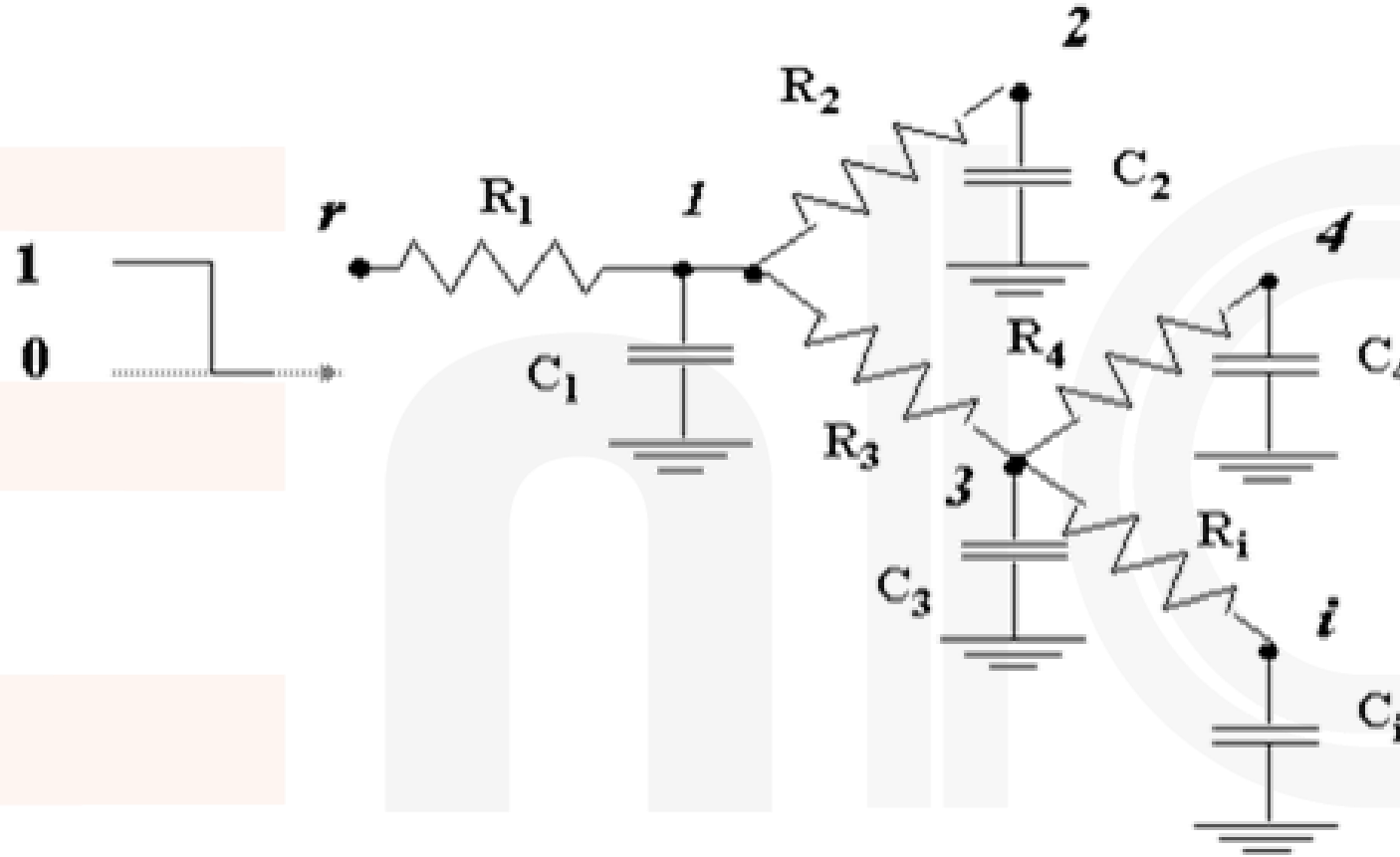
$$R_{ik} = \sum R_j \Rightarrow (R_j \in [path(s \rightarrow i) \cap path(s \rightarrow k)]) \quad \tau_{Di} = \sum_{k=1}^N C_k R_{ik}$$

Simple Elmore Delay Example



$$\tau_{elmore} = R_1 C_1 + (R_1 + R_2) C_2 + (R_1) C_3$$

General Elmore Delay Example



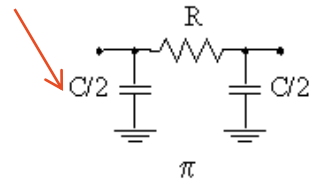
$$\tau_{elmore} = R_1 C_1 + R_1 C_2 + (R_1 + R_3) C_3 + (R_1 + R_3) C_4 + (R_1 + R_3 + R_i) C_i$$

RC-Models

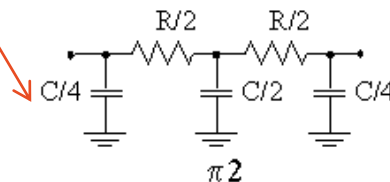
Voltage Range	Lumped RC-network	Distributed RC-network
0→50% (t_p)	0.69 RC	0.38 RC
0→63% (τ)	RC	0.5 RC
10%→90% (t_T)	2.2 RC	0.9 RC

Step Response of Lumped and Distributed RC Networks:
Points of Interest.

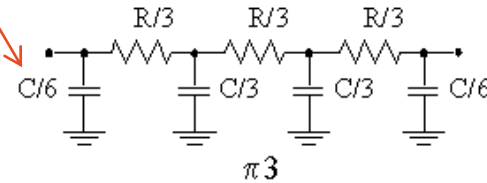
Pie Model



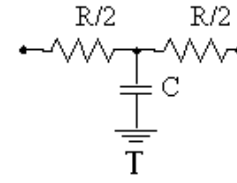
Pie-2 Model



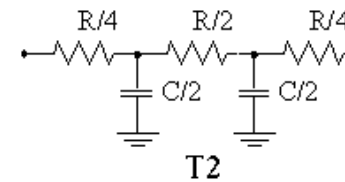
Pie-3 Model



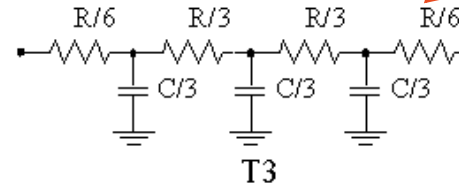
T-Model



T-2 Model

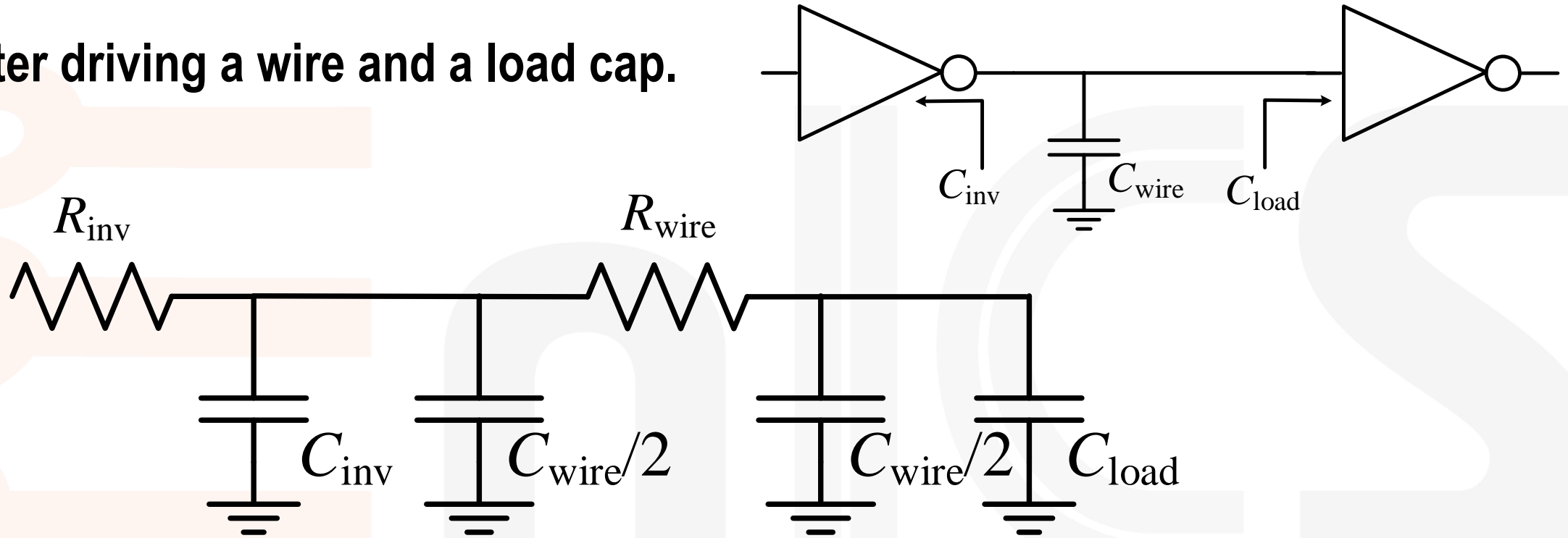


T-3 Model



Using the Elmore Delay and Pi Model

- Inverter driving a wire and a load cap.



$$\tau_{driver} = \left(C_{inv} + \frac{C_{wire}}{2} \right) R_{inv} + \left(C_{load} + \frac{C_{wire}}{2} \right) (R_{inv} + R_{wire})$$

Dealing with long wires

- Repeater Insertion



Dealing with long wires

- Buffer Tree Insertion



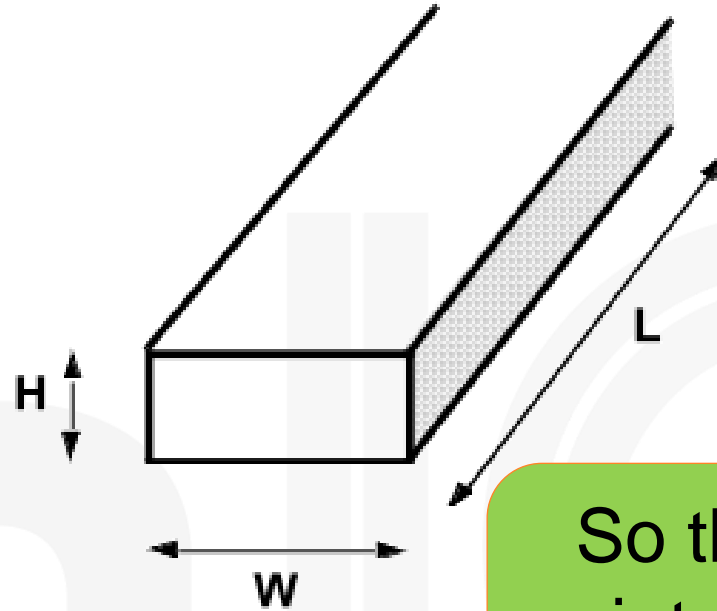
Wire Scaling

Wire Scaling

- We could try to scale interconnect at the same rate (S) as device dimensions.
 - This makes sense for *local wires* that connect smaller devices/gates.
 - But *global interconnections*, such as clock signals, buses, etc., won't scale in length.
- Length of *global interconnect* is proportional to *die size* or *system complexity*.
 - *Die Size* increased by 6% per year (X2 @10 years) *for a while*.
 - Devices have scaled, but complexity has grown!

Local Wire Scaling

- Looking at local interconnect:
 - W, H, t, L all scale at $1/S$
 - $C = LW/t \rightarrow 1/S$
 - $R = L/WH \rightarrow S$
 - $RC = 1$
- Reminder – Full Scaling of Transistors
 - $R_{on} = V_{DD}/I_{on} \propto 1$
 - $t_{pd} = R_{on} C_g \propto 1/S$

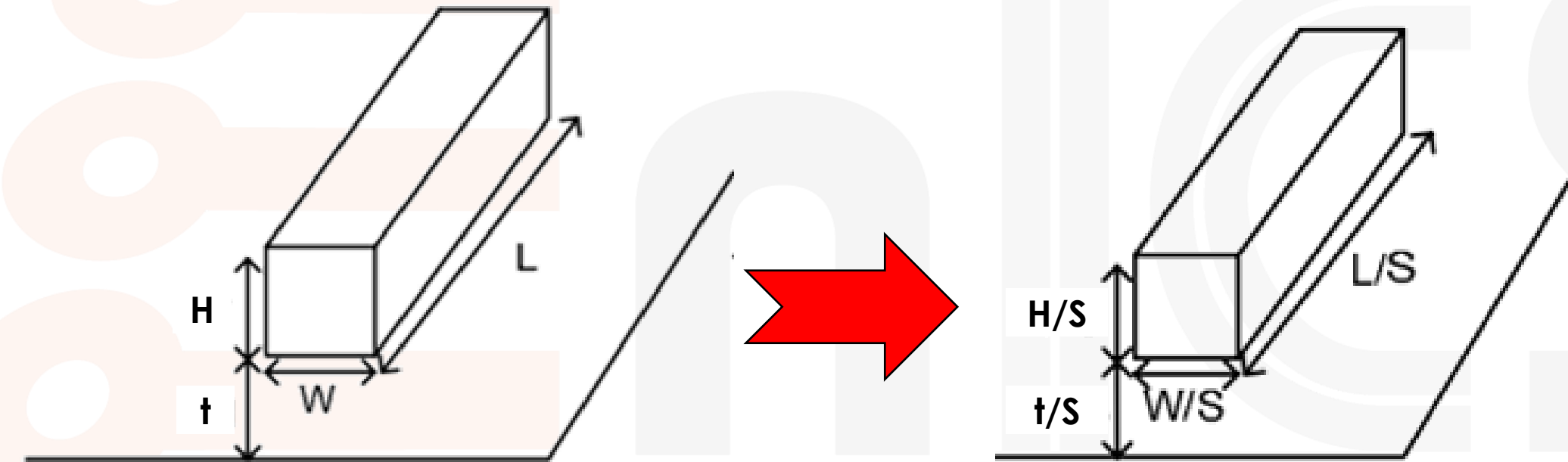


So the delay of local interconnect stays constant.

But the delay of local interconnect increases relative to transistors!

Local Wire Scaling – Full Scaling

- What about fringe cap?



$$C_{pp} \propto WL/t \quad C_{fringe} \propto L$$

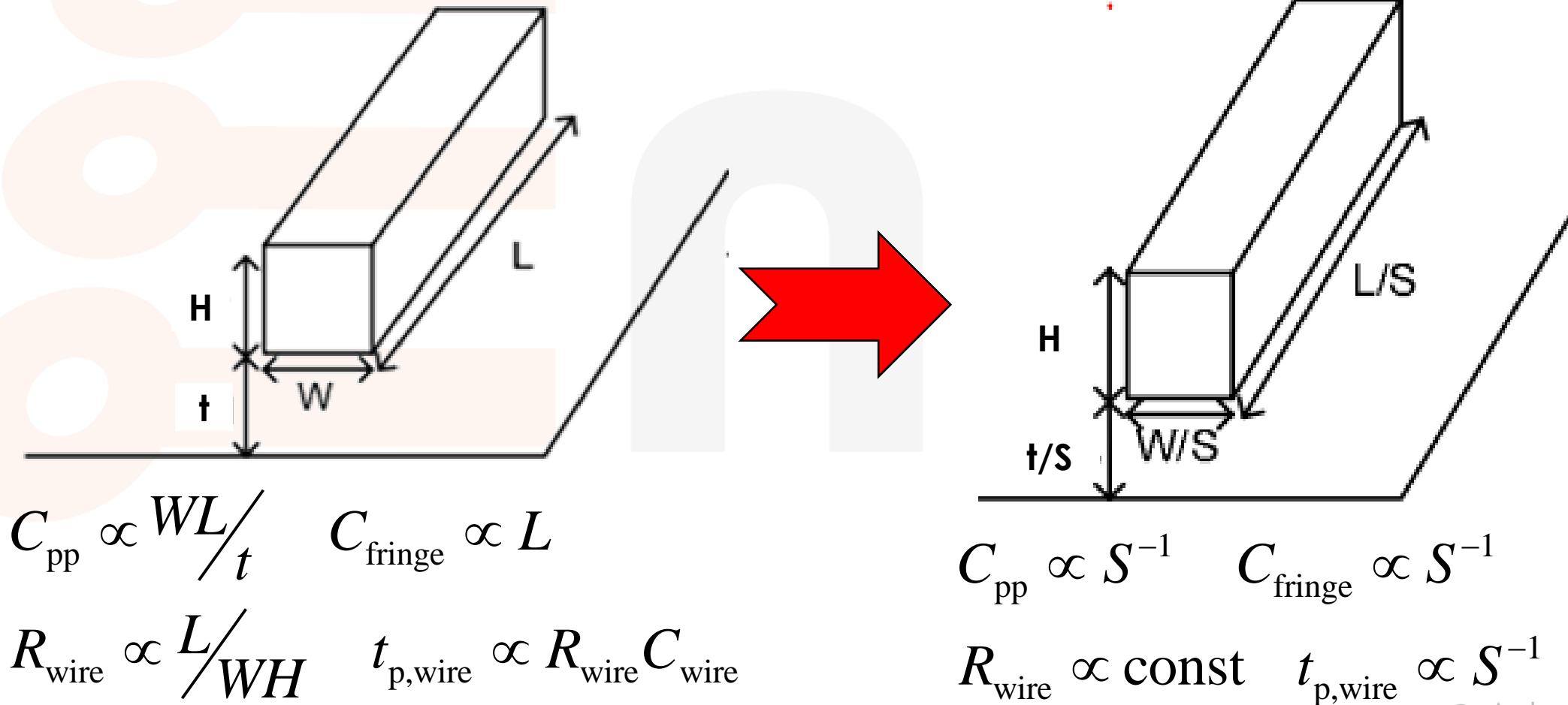
$$R_{wire} \propto L/WH \quad t_{p,wire} \propto R_{wire} C_{wire}$$

$$C_{pp} \propto S^{-1} \quad C_{fringe} \propto S^{-1}$$

$$R_{wire} \propto S \quad t_{p,wire} \propto \text{const}$$

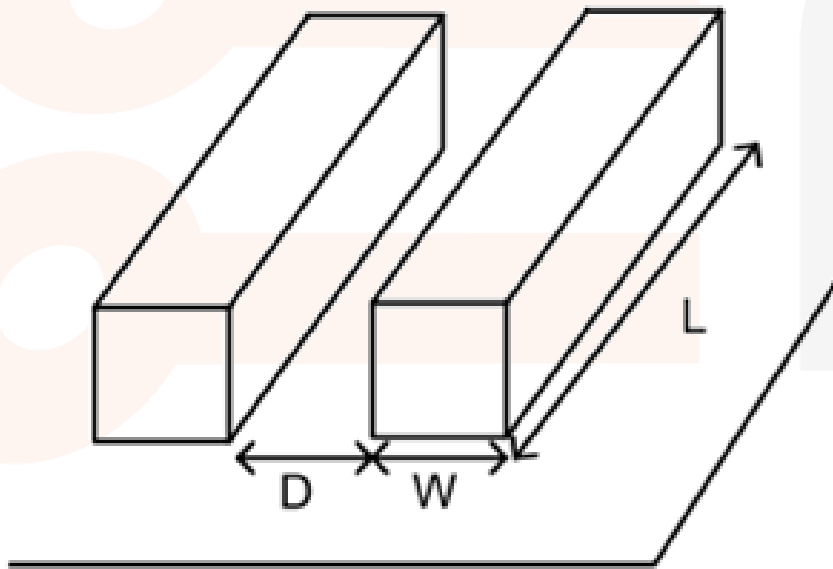
Local Wire Scaling - Constant Thickness

- Wire thickness (height) wasn't scaled!

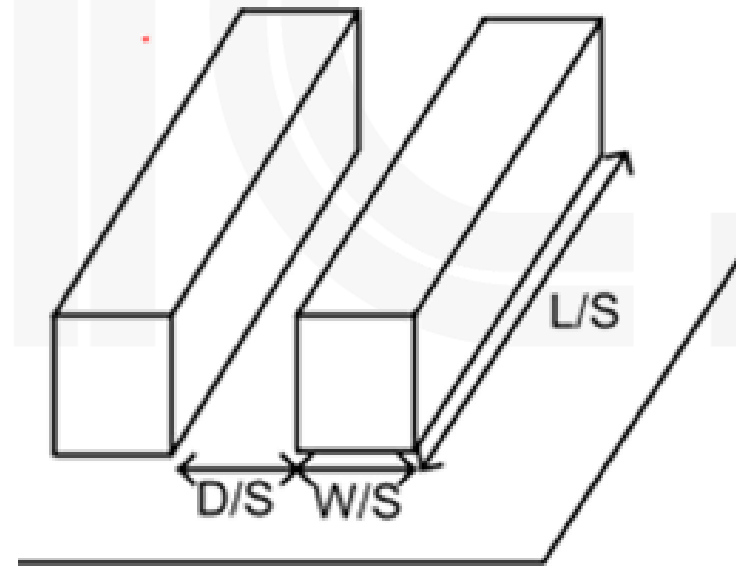
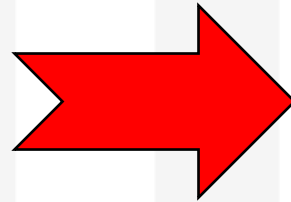


Local Wire Scaling – Interwire Capacitance

- Without scaling height, **coupling** gets much worse.
- **Aspect ratio** is limited and we eventually have to scale the height.

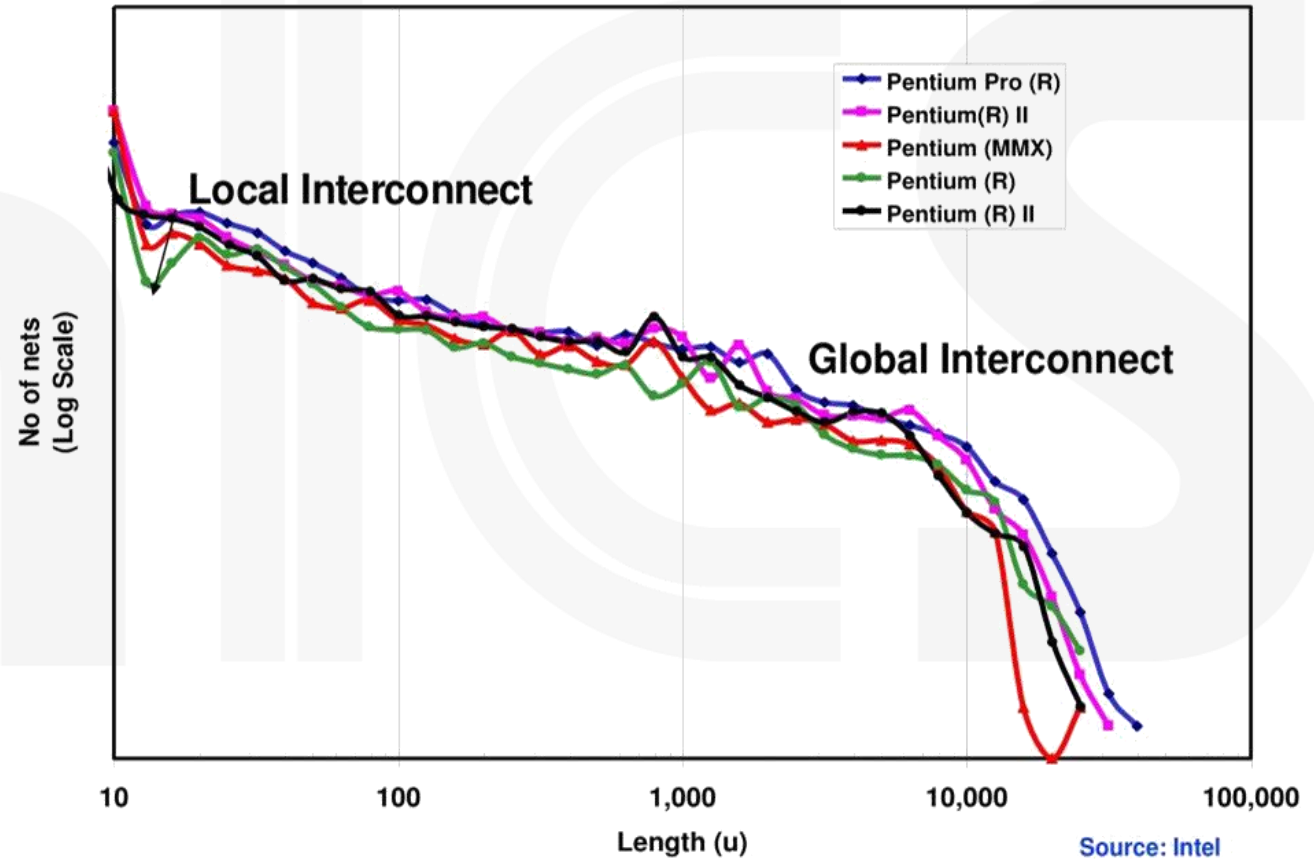
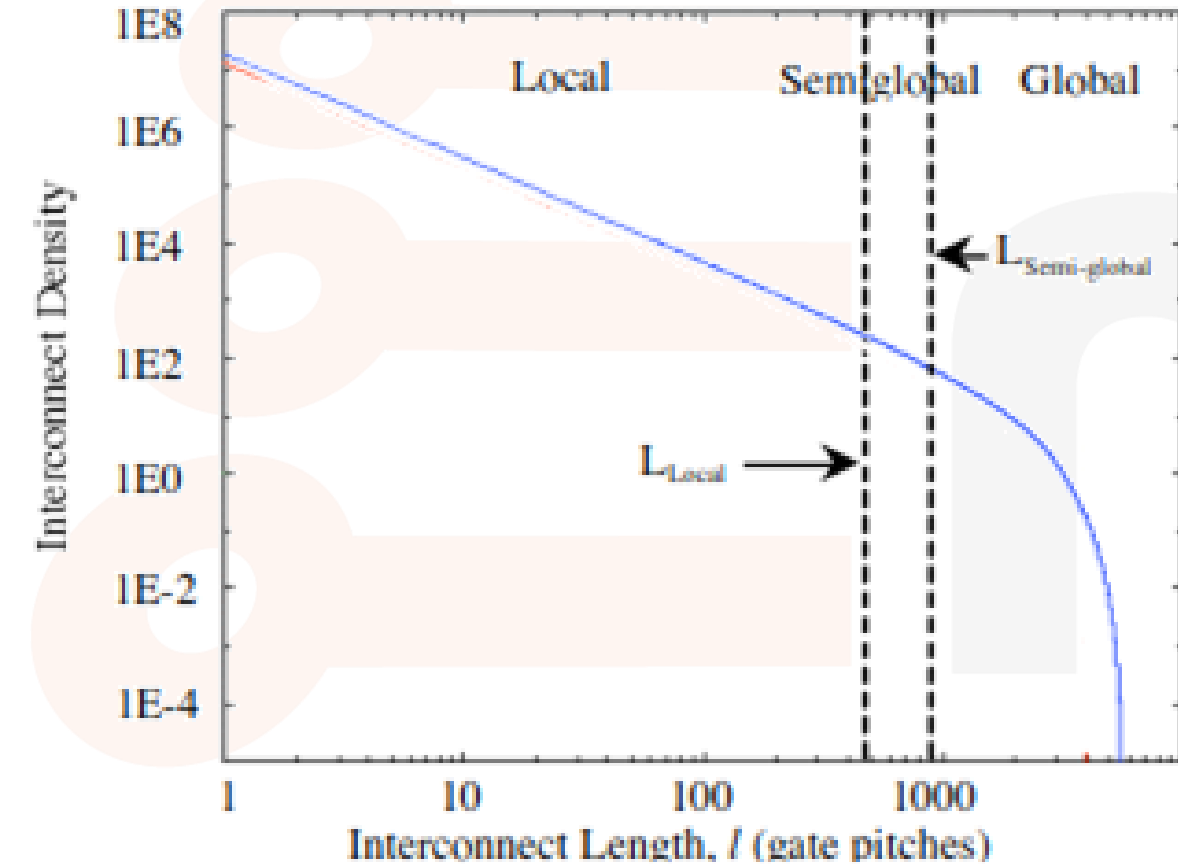


$$C_{pp, \text{side}} \propto LH/D$$



$$C_{pp, \text{side}} \propto \text{const}$$

Nature of Interconnect

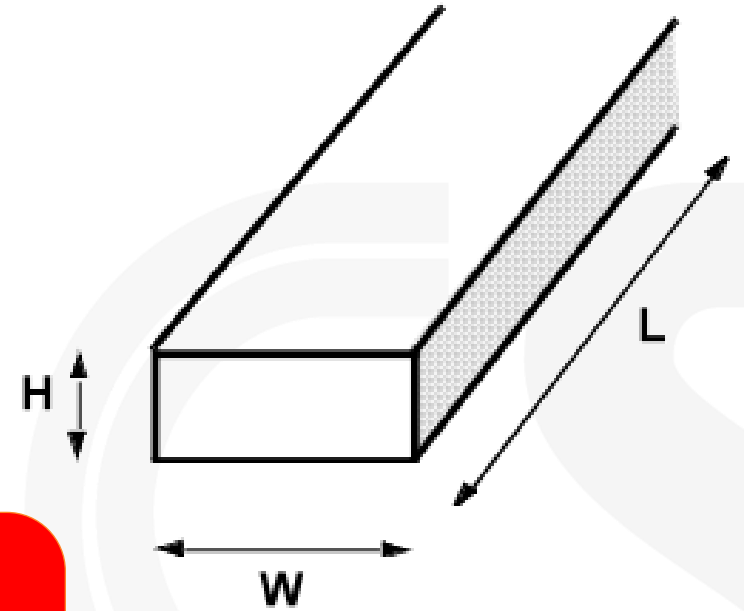


Global Wire Scaling

- Looking at **global interconnect**:

- W, H, t scale at $1/S$
- L doesn't scale!
- $C = LW/t \rightarrow 1$
- $R = L/WH \rightarrow S^2$
- $RC = S^2$!!!

Long wire
delay
increases



- And if chip size grows, L actually increases!

Global Wire Scaling – Constant Thickness

- Leave thickness constant for **global wires**
- But wire delay still gets quadratically worse than gate delay...

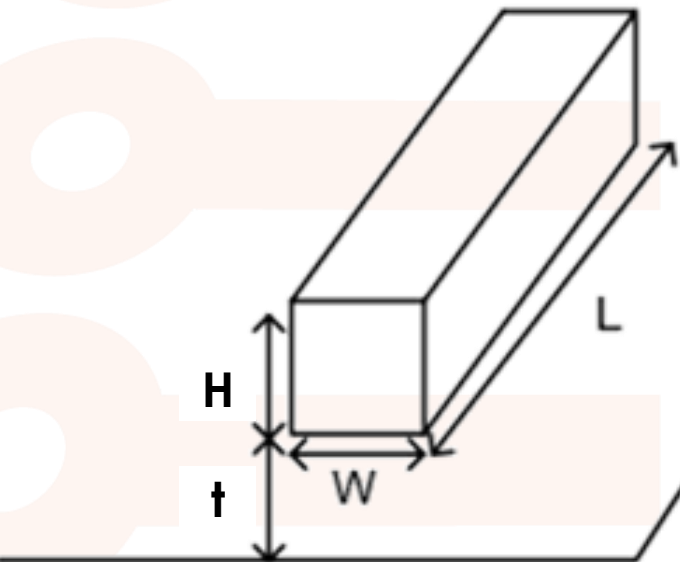
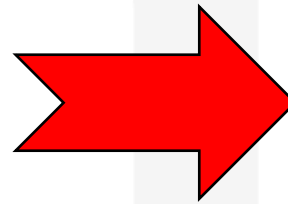


Diagram of a wire cross-section with dimensions L (length), W (width), H (height), and thickness t . The wire is shown as a rectangular prism on a substrate.

$$C_{pp} \propto WL/t \quad C_{fringe} \propto L$$
$$R_{wire} \propto L/WH \quad t_{p,wire} \propto R_{wire} C_{wire}$$



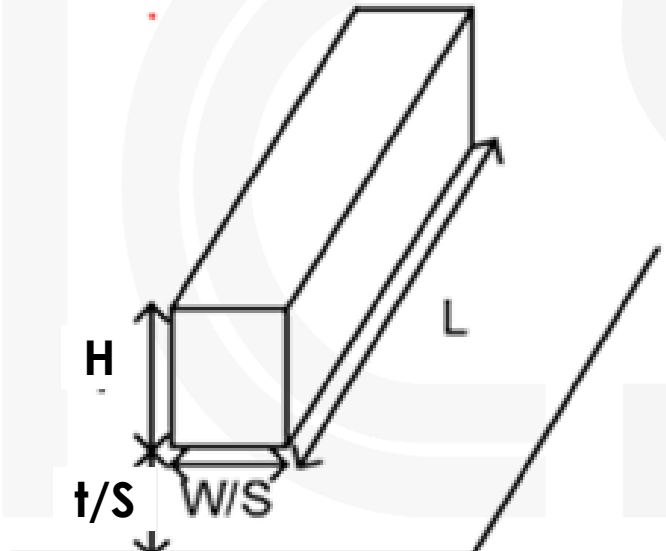


Diagram of a scaled wire cross-section with dimensions L (length), W/S (width), H (height), and thickness t/S . The wire is shown as a rectangular prism on a substrate.

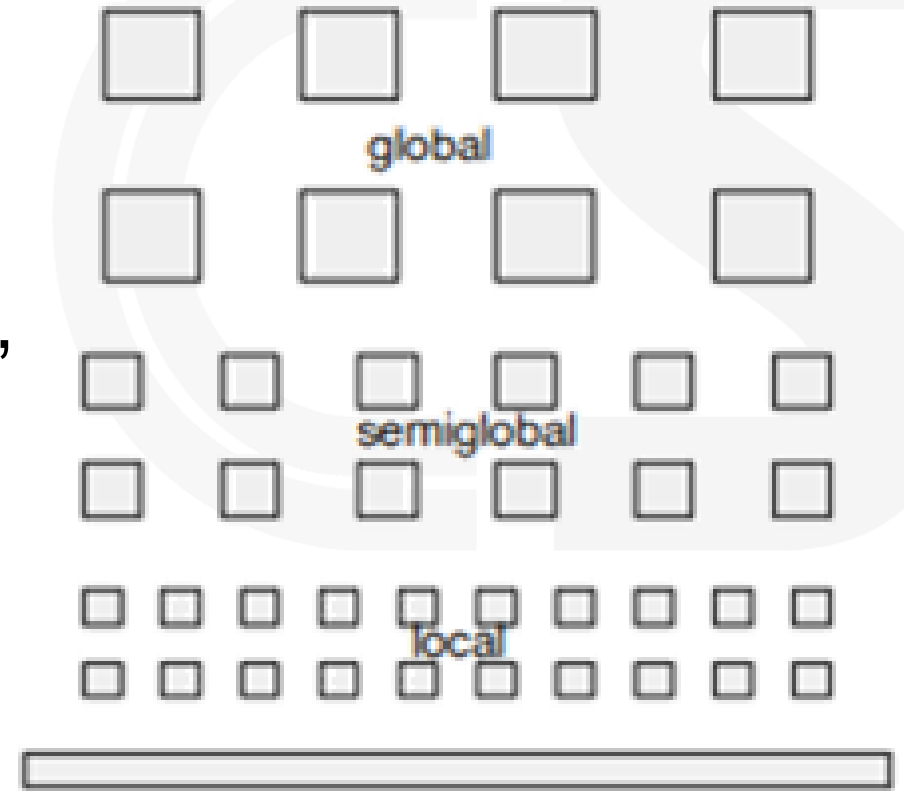
$$C_{pp} \propto \text{const} \quad C_{fringe} \propto \text{const}$$
$$R_{wire} \propto S \quad t_{p,wire} \propto S$$

Wire Scaling

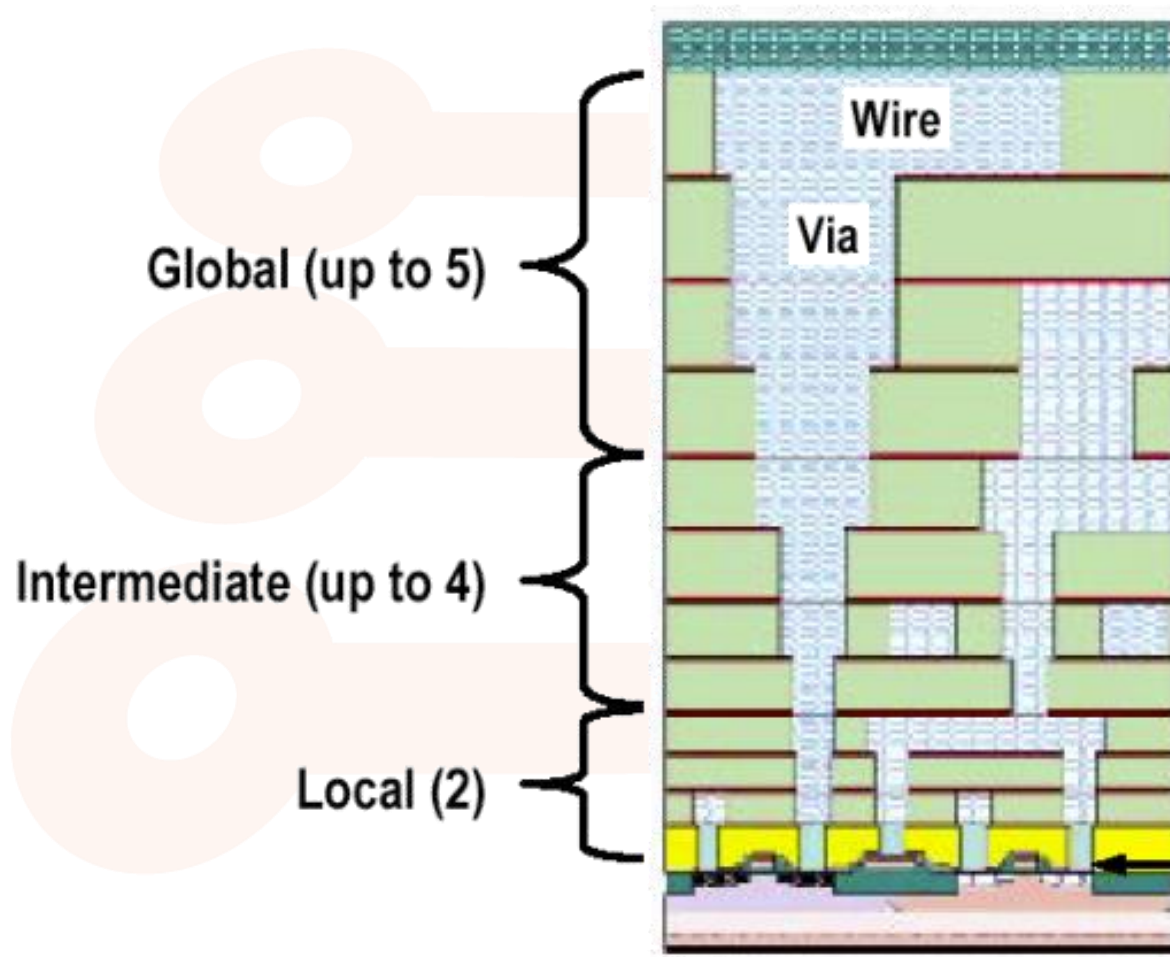
- So whereas device speed increases with scaling:
 - Local interconnect speed **stays constant**.
 - Global interconnect delays **increase quadratically**.
- Therefore:
 - Interconnect delay is often the **limiting factor** for speed.
- What can we do?
 - Keep the wire thickness (H) fixed.
 - This would provide $1/S$ for local wire delays and S for constant length global wires.
 - But fringing/coupling capacitance increase, so this is optimistic.

Wire Scaling

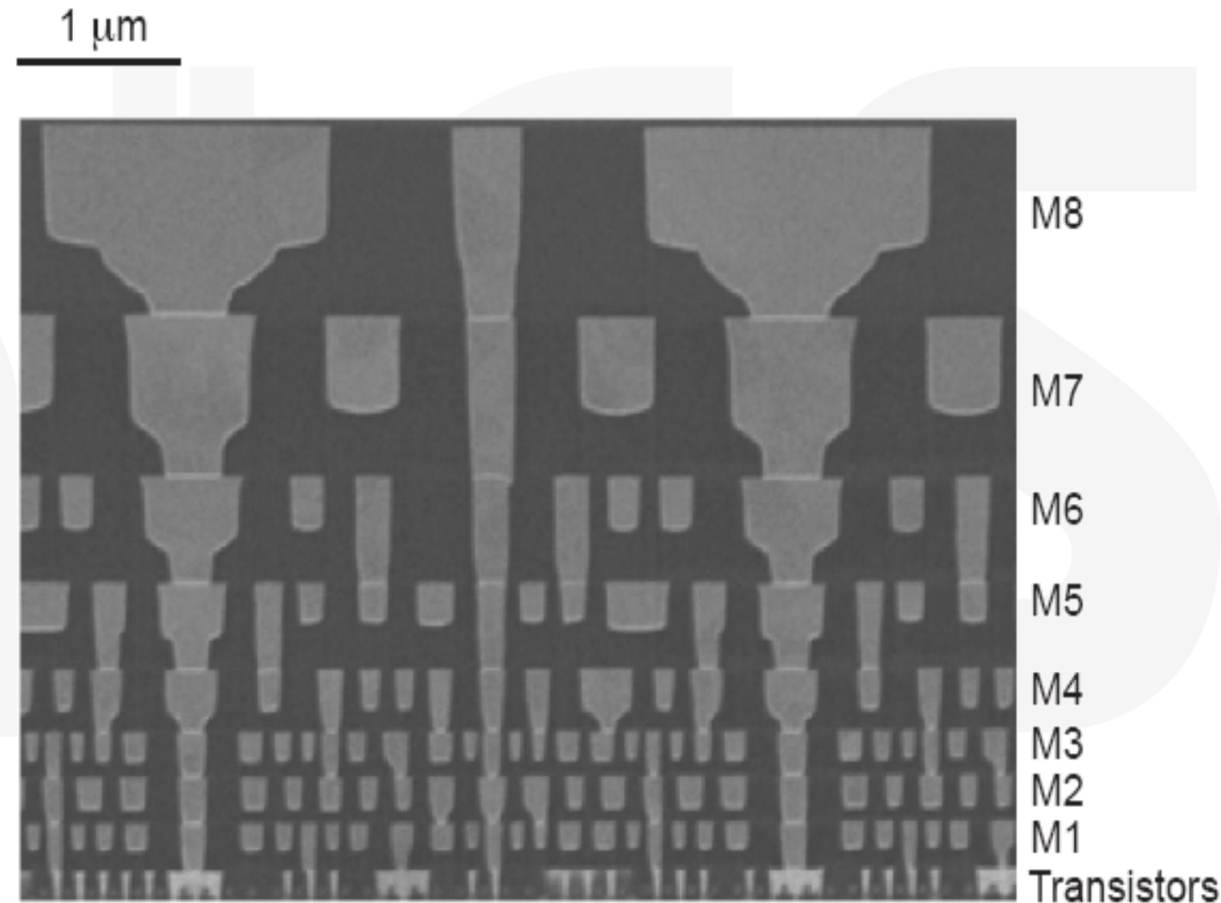
- What is done today?
 - Low resistance metals.
 - Low- K insulation.
 - Low metals ($M1$, $M2$) are used for local interconnect, so they are thin and dense.
 - Higher metals are used for global routing, so they are thicker, wider and spaced farther apart.



Modern Interconnect



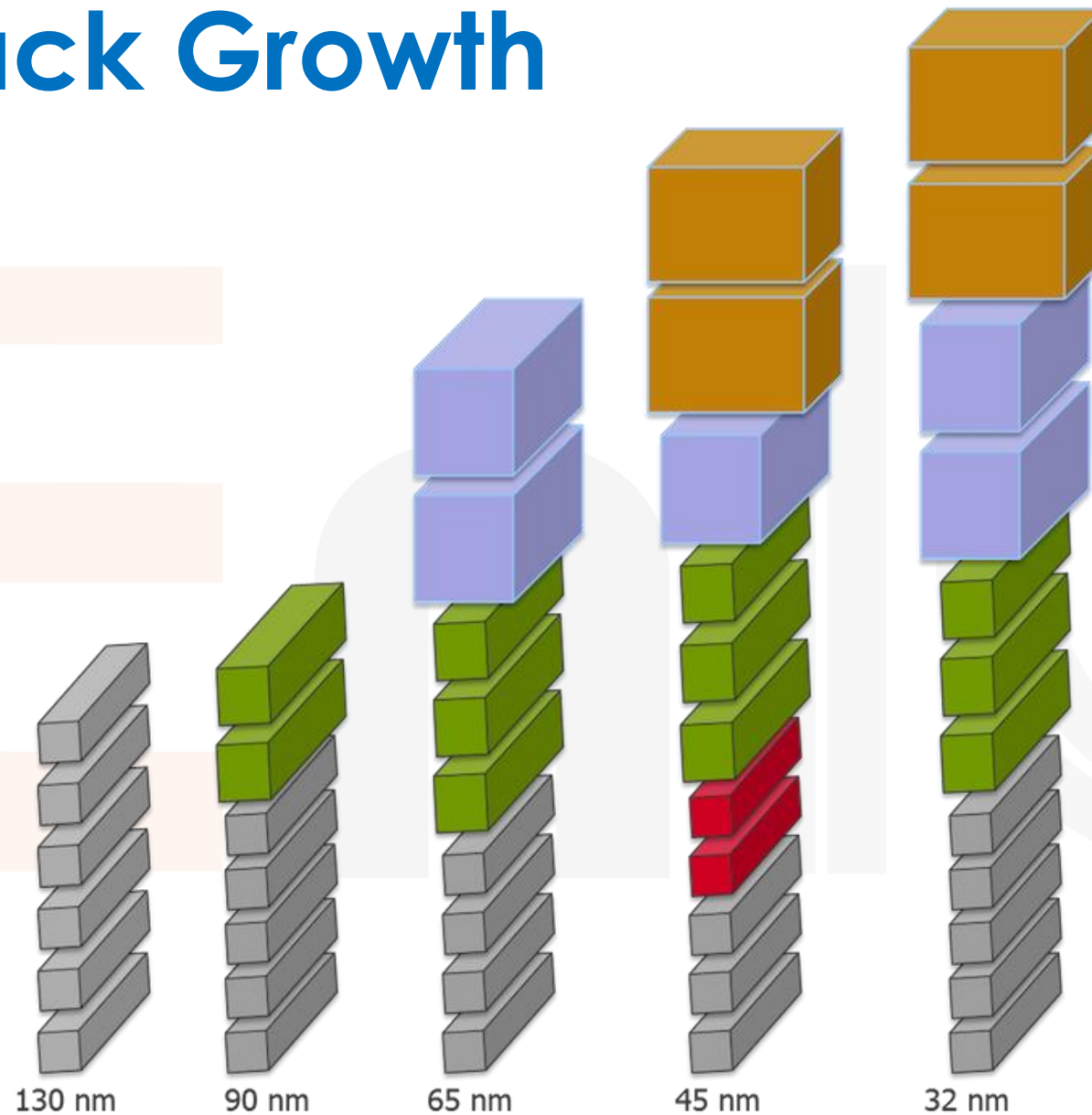
**Cross Sectional View:
For Height, Width and Spacing**



Intel 45 nm Stack

[Moon08]

Metal Stack Growth



Further Reading

- J. Rabaey, “*Digital Integrated Circuits*” 2003, Chapter 4
- E. Alon, Berkeley EE-141, Lectures 15,16 (Fall 2009)
http://bwrc.eecs.berkeley.edu/classes/icdesign/ee141_f09/
- B. Nikolic, Berkeley EE-241, Lecture 3 (Spring 2011)
http://bwrc.eecs.berkeley.edu/classes/icdesign/ee241_s11/
- Stanford EE311