# Digital Integrated Circuits
(83-313)

## Lecture 9:

# Interconnect

Semester B, 2016-17


Lecturer: Dr. Adam Teman

TAs: Itamar Levi,
Robert Giterman


23 May 2017

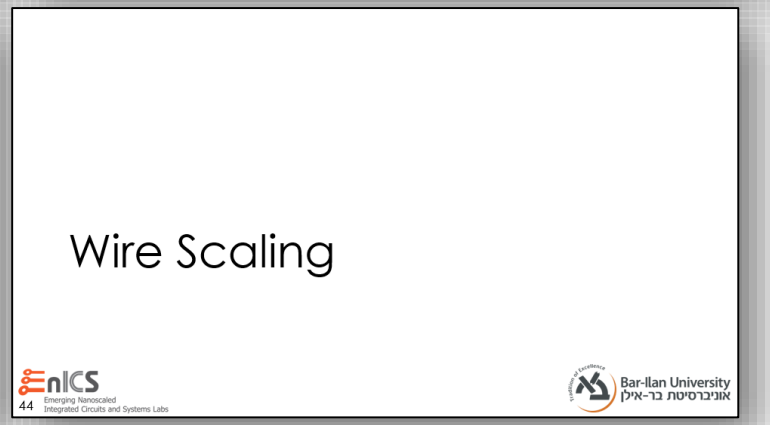

Emerging Nanoscaled
Integrated Circuits and Systems Labs

Bar-Ilan University
אוניברסיטת בר-אילן

# Lecture Content

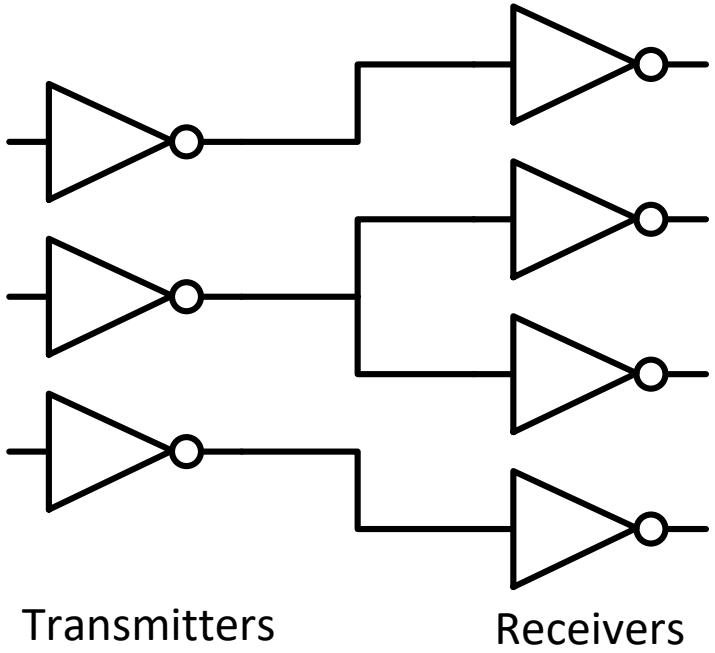




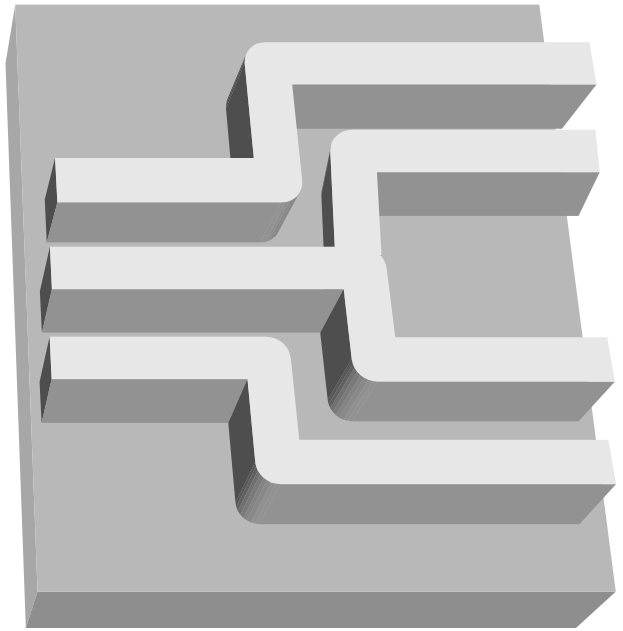


# A First Glance at Interconnect

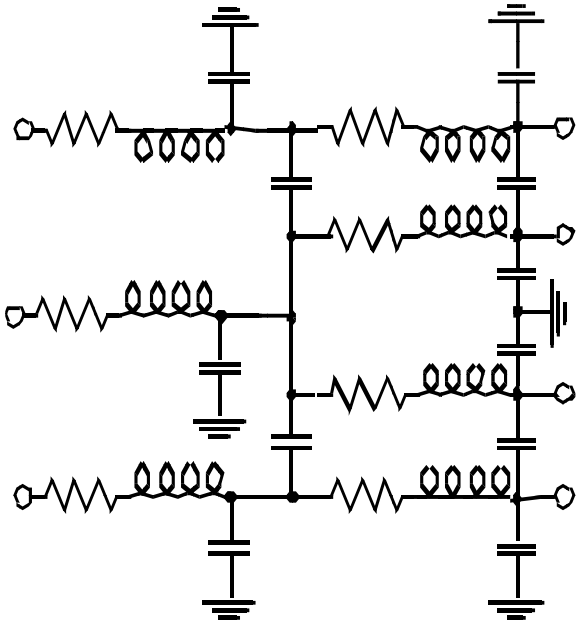
# The Wire



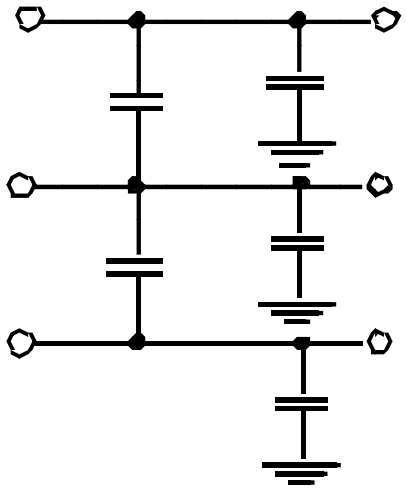

schematic view



physical realization



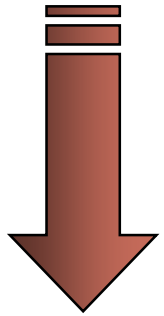
All-inclusive model



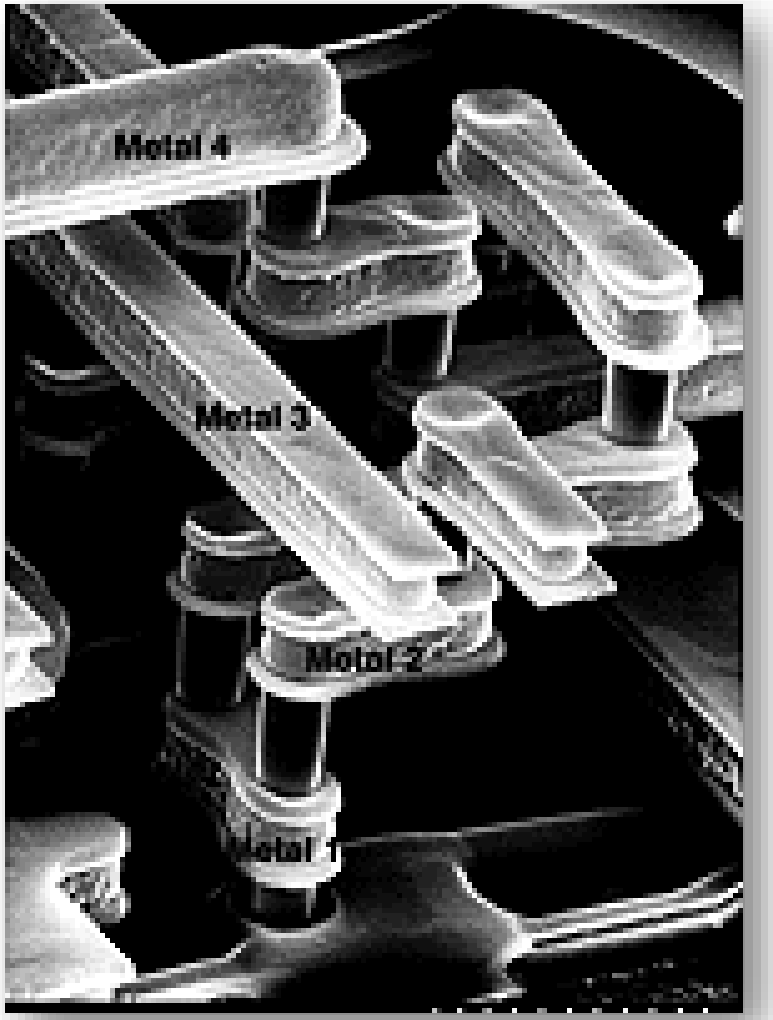
Capacitance-only

# Impact of Interconnect Parasitics

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

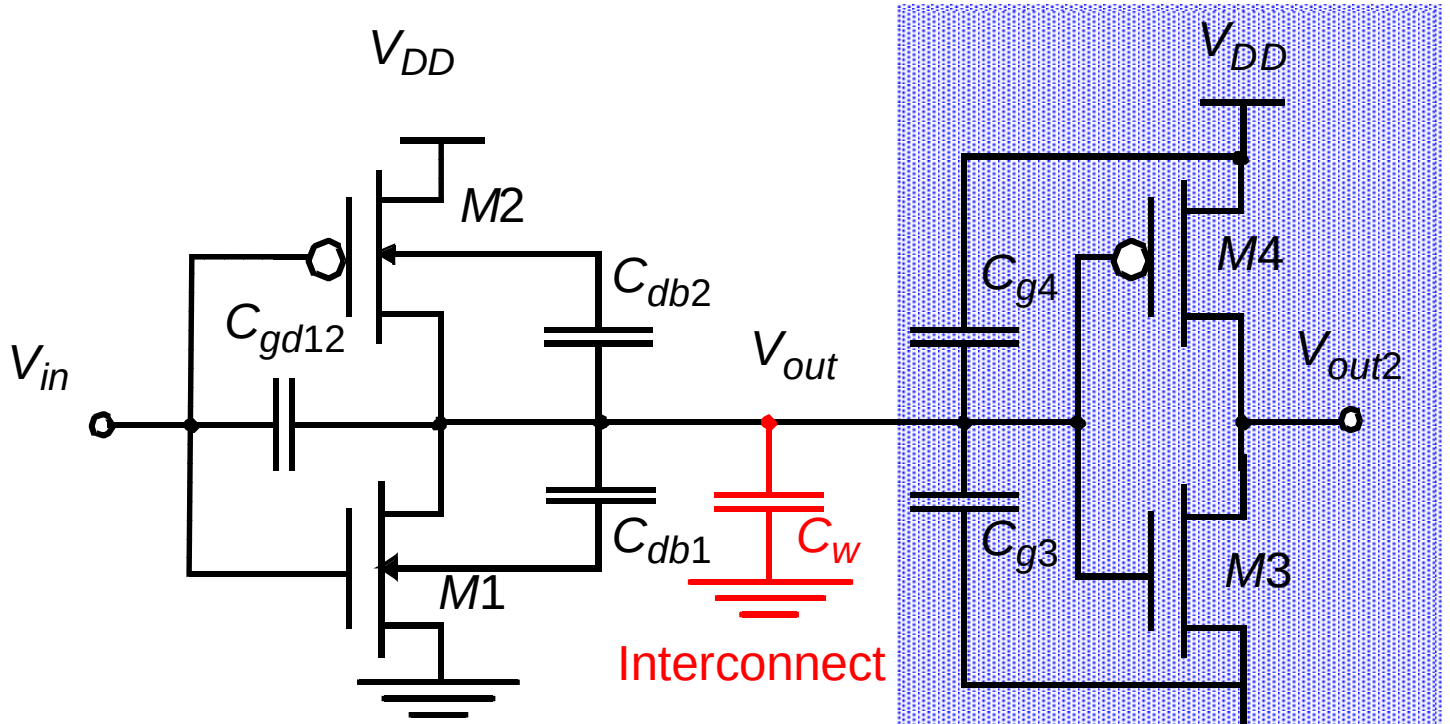


# Modern Interconnect

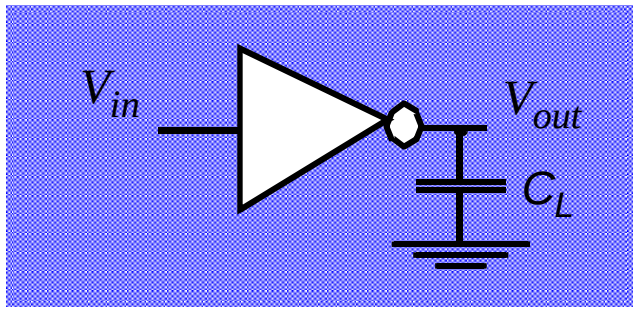


# Capacitance

# Capacitance of Wire Interconnect

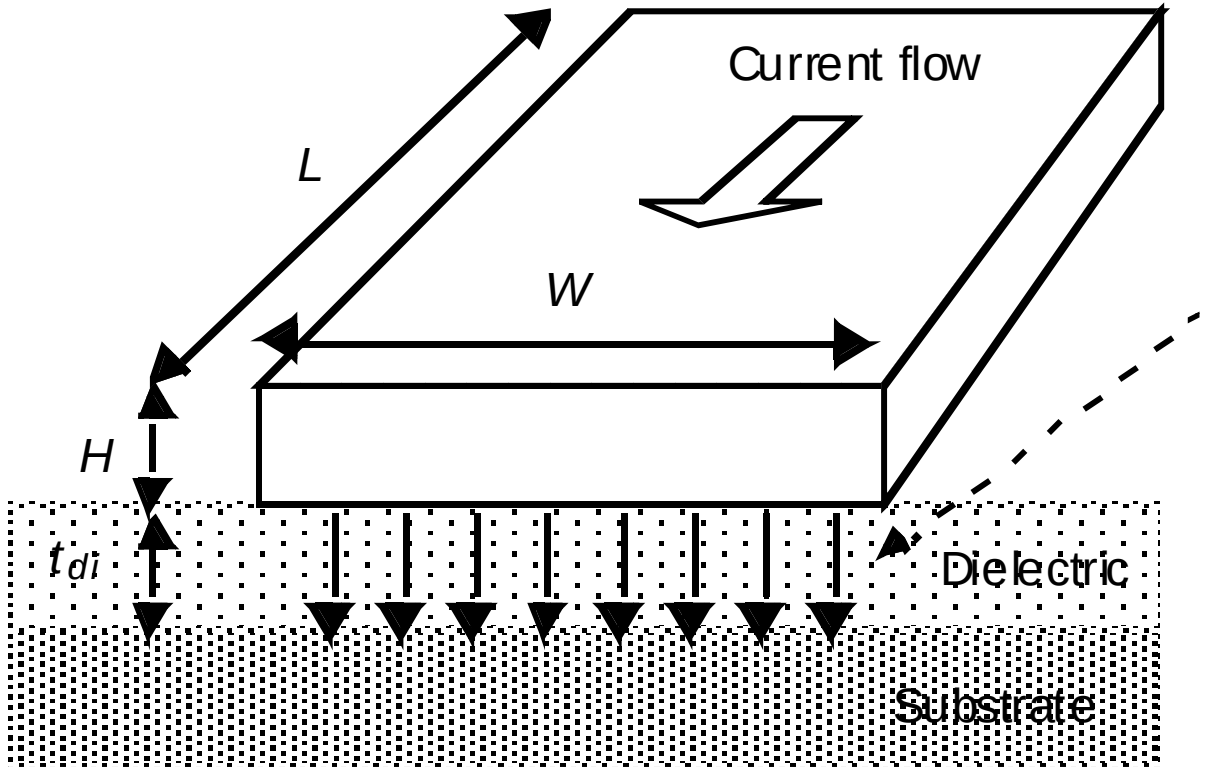



**Simplified Model**



# Capacitance: The Parallel Plate Model

- **How can we reduce this capacitance?**



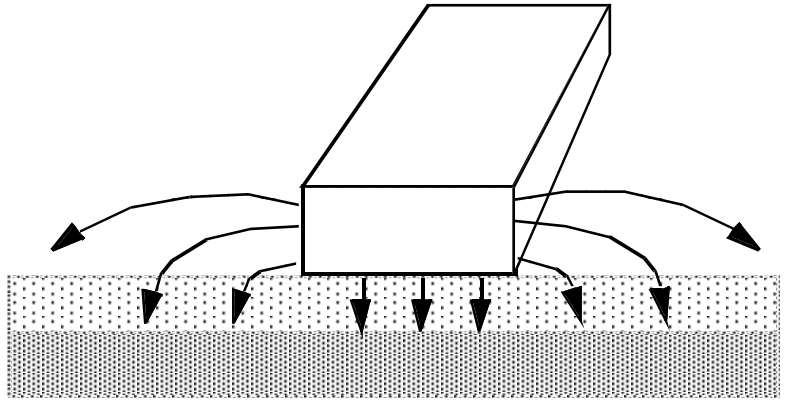

Typical numbers:
- Wire cap ~0.2 fF/um
- Gate cap ~2 fF/um
- Diffusion cap ~2 fF/um

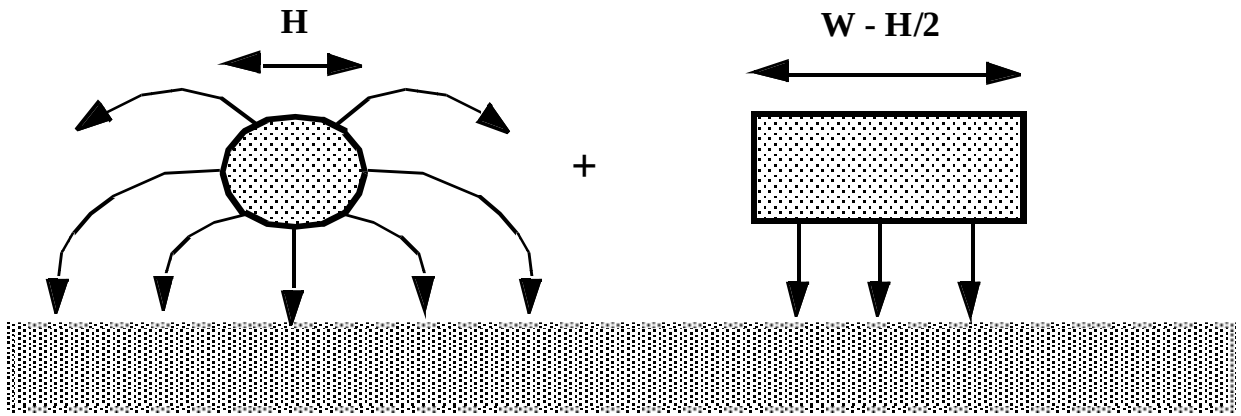
$$c_{pp} = \frac{\varepsilon_{di}}{t_{di}} WL$$

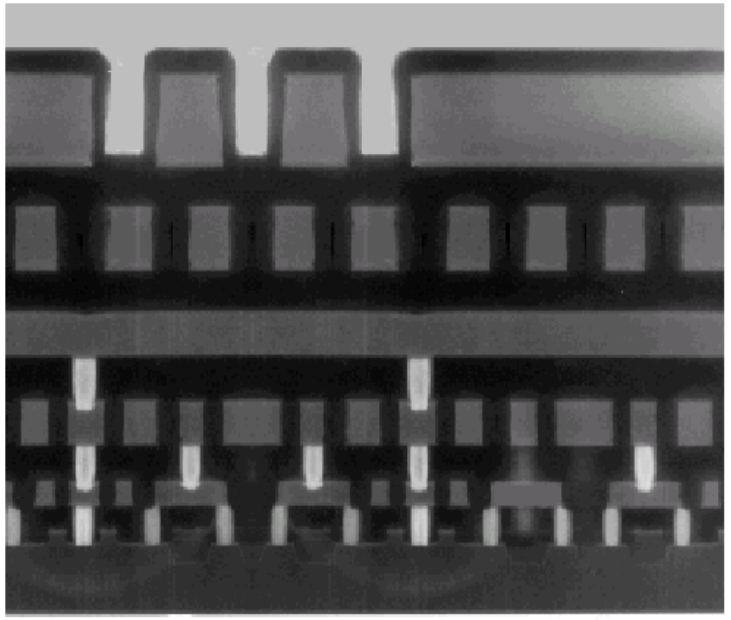
# Permittivity

| Material | $\varepsilon_r$ |
|---|---|
| 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 |

# Fringing Capacitance



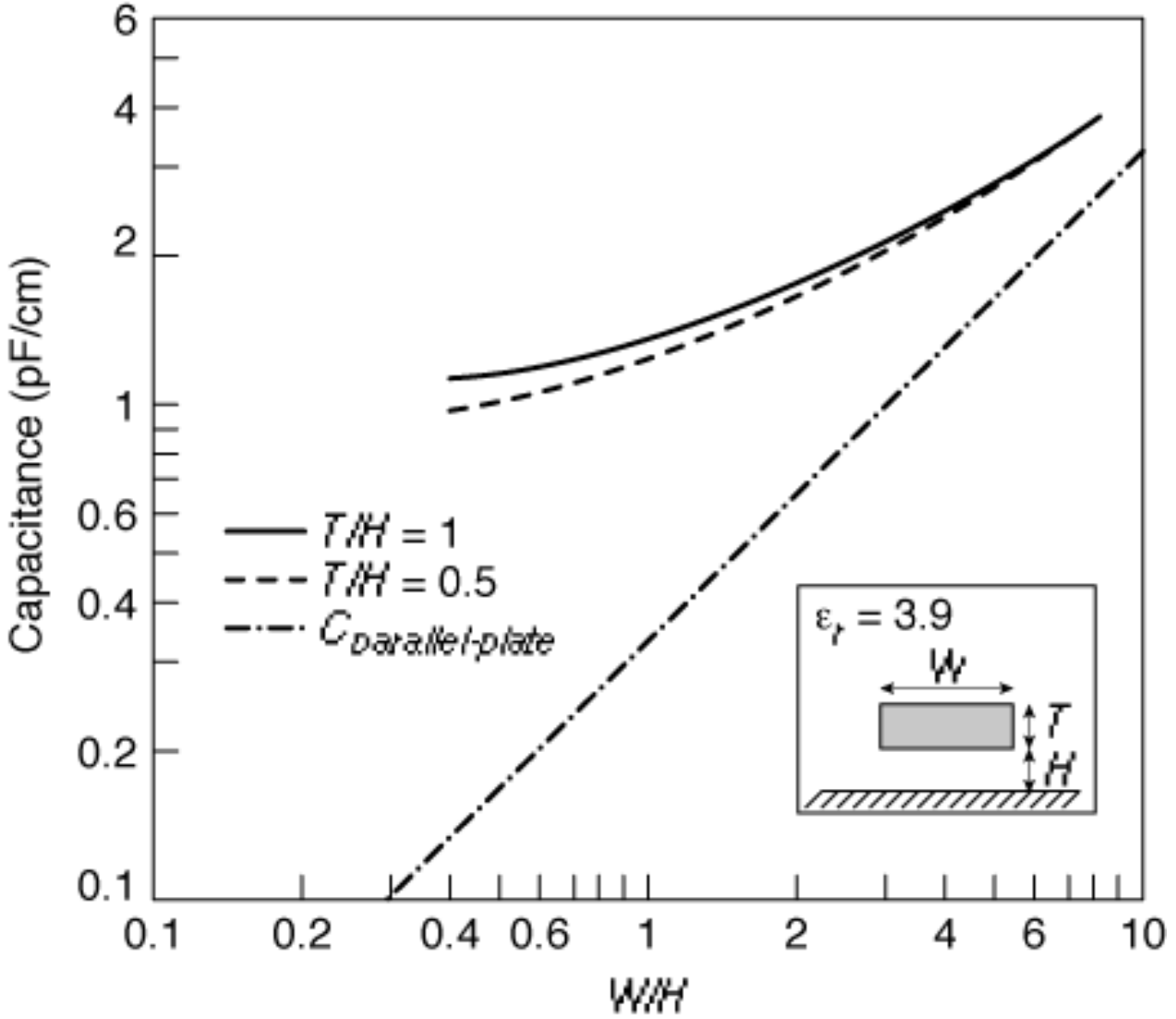


$$w = W - H/2$$

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

# Fringing versus Parallel Plate




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

$$C_{fringe} \propto L$$

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

(from [Bakoglu89])

# A simple model for deriving wire cap

- **Wiring capacitances in 0.25μm**

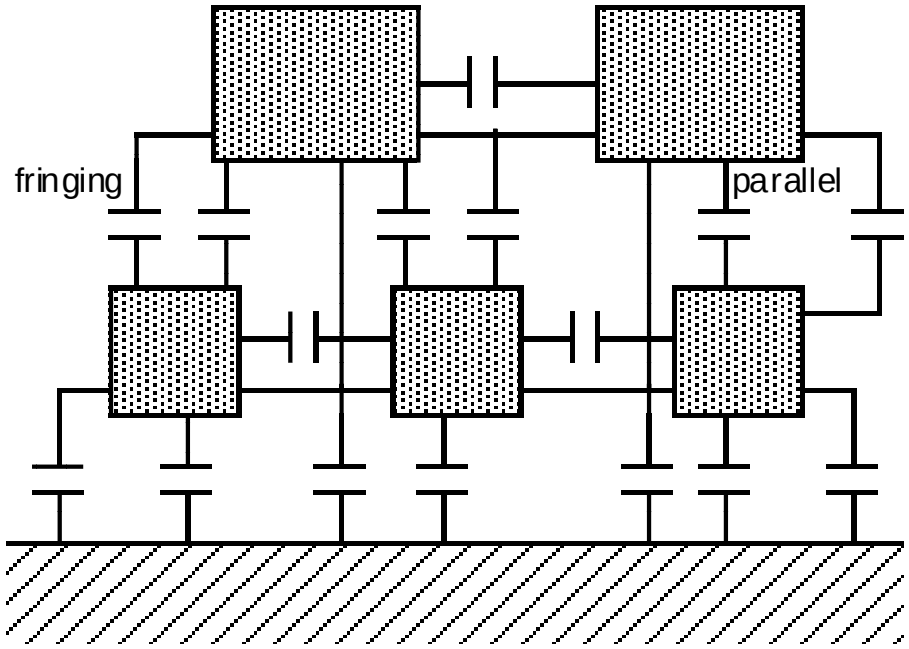
$$C_{wire} = C_{parallel\_plate} \cdot W \cdot L + 2 \cdot C_{fringe} \cdot L$$

Bottom Plate

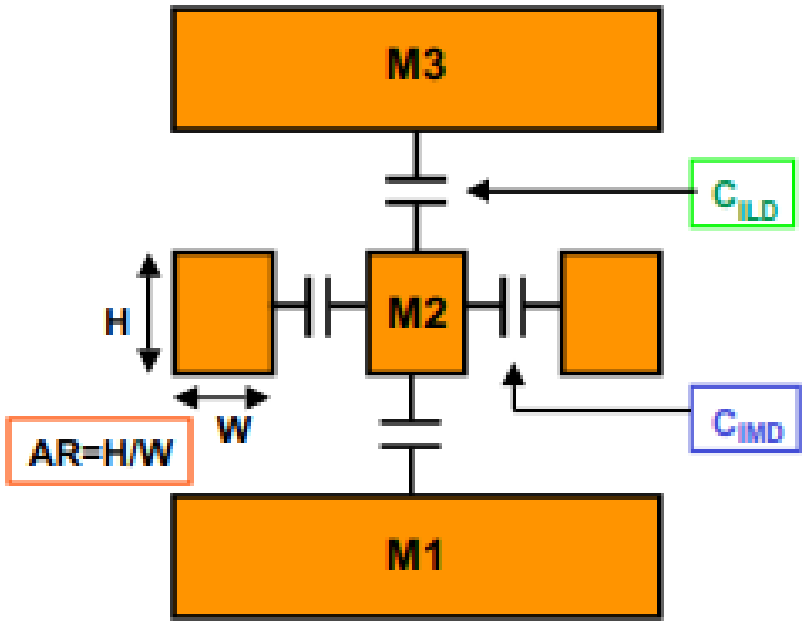
| Top Plate | Field | Active | Poly | Al1 | Al2 | Al3 | Al4 |
|---|---|---|---|---|---|---|---|
| Poly | 88 | | | | | | |
| | 54 | | | | | | |
| Al1 | 30 | 41 | 57 | | | | |
| | 40 | 47 | 54 | | | | |
| Al2 | 13 | 15 | 17 | 36 | | | |
| | 25 | 27 | 29 | 45 | | | |
| Al3 | 8.9 | 9.4 | 10 | 15 | 41 | | |
| | 18 | 19 | 20 | 27 | 49 | | |
| Al4 | 6.5 | 6.8 | 7 | 8.9 | 15 | 35 | |
| | 14 | 15 | 15 | 18 | 27 | 45 | |
| Al5 | 5.2 | 5.4 | 5.4 | 6.6 | 9.1 | 14 | 38 |
| | 12 | 12 | 12 | 14 | 19 | 27 | 52 |

aF/μm²

aF/μm

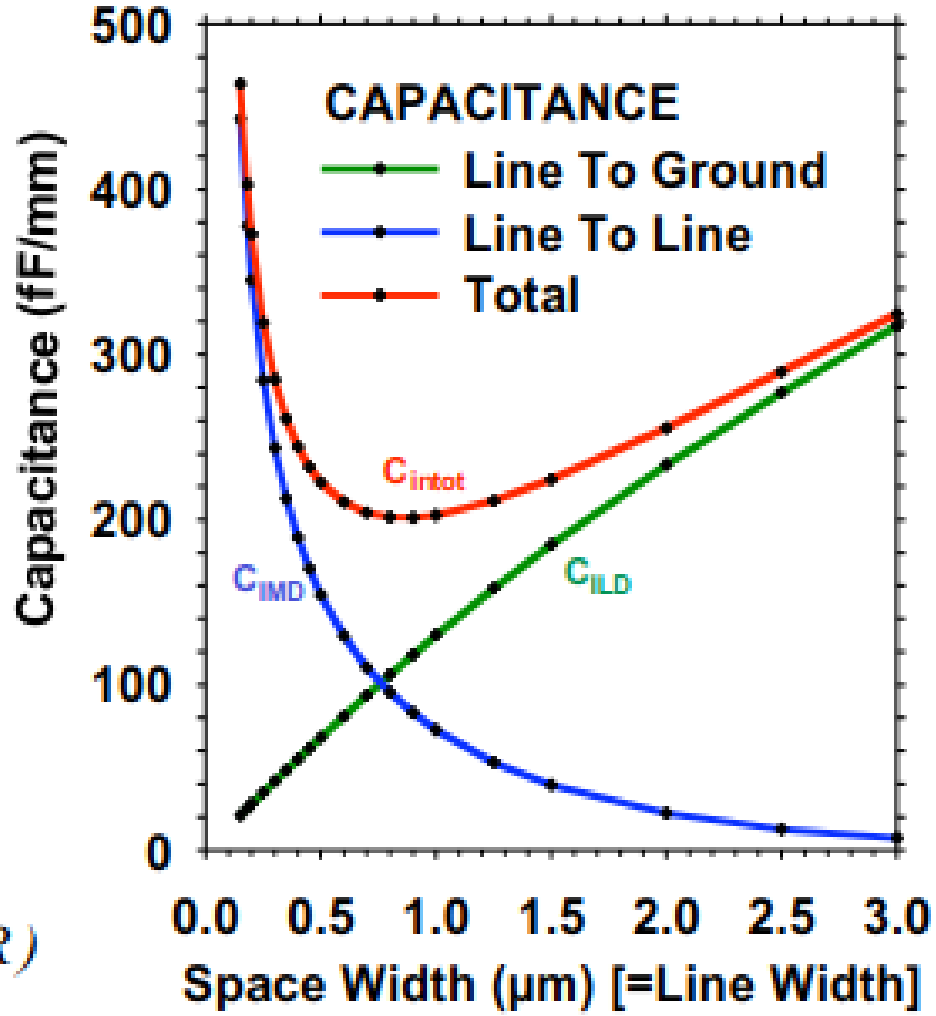


# Impact of Interwire Capacitance



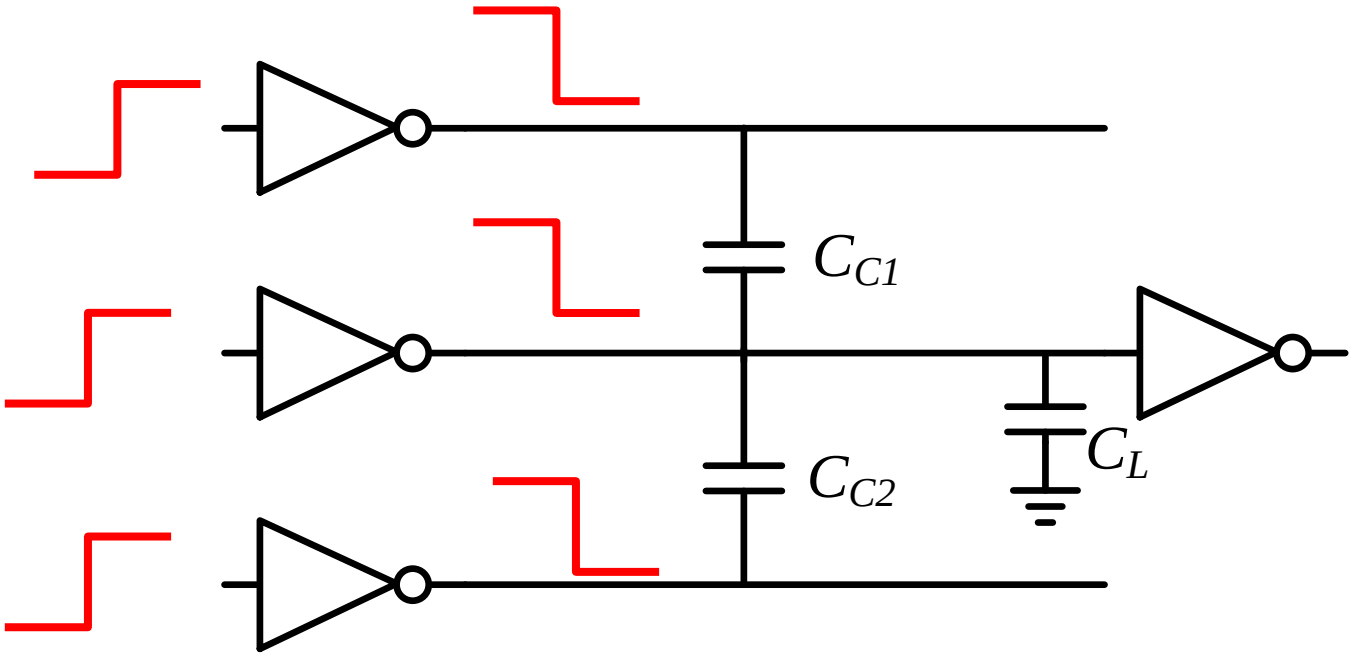

**In general**

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



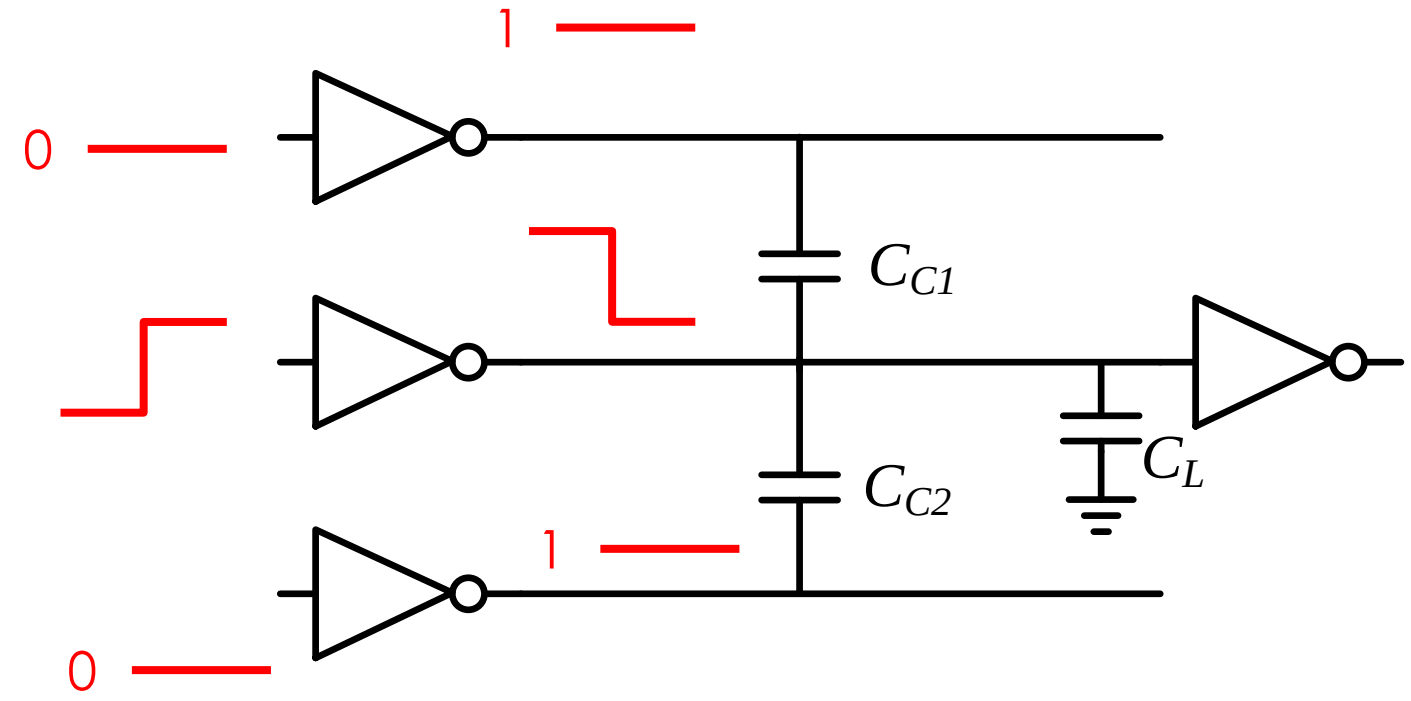

*Stanford: EE311*

# Coupling Capacitance and Delay



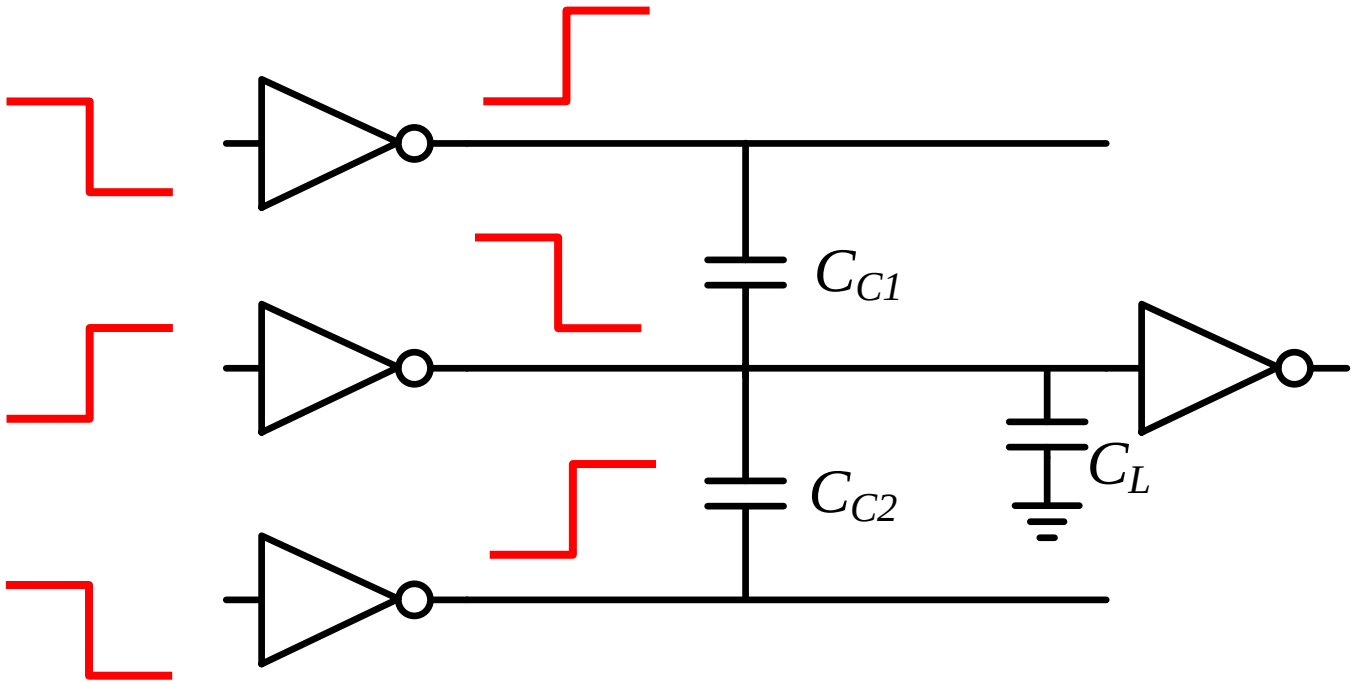

$$C_{tot} = C_L$$

# Coupling Capacitance and Delay




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

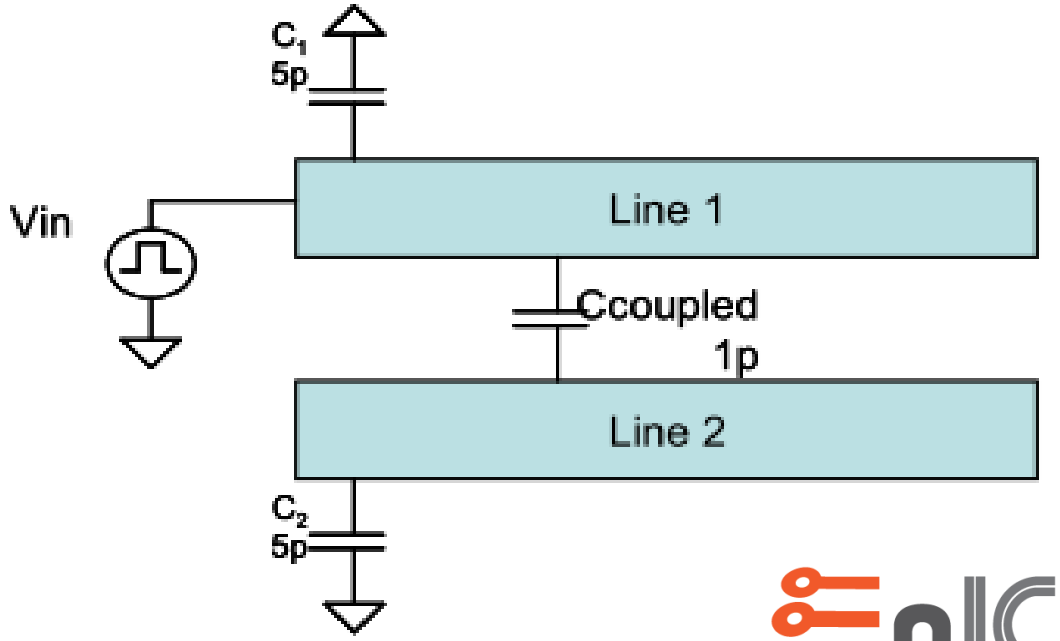
# Coupling Capacitance and Delay




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

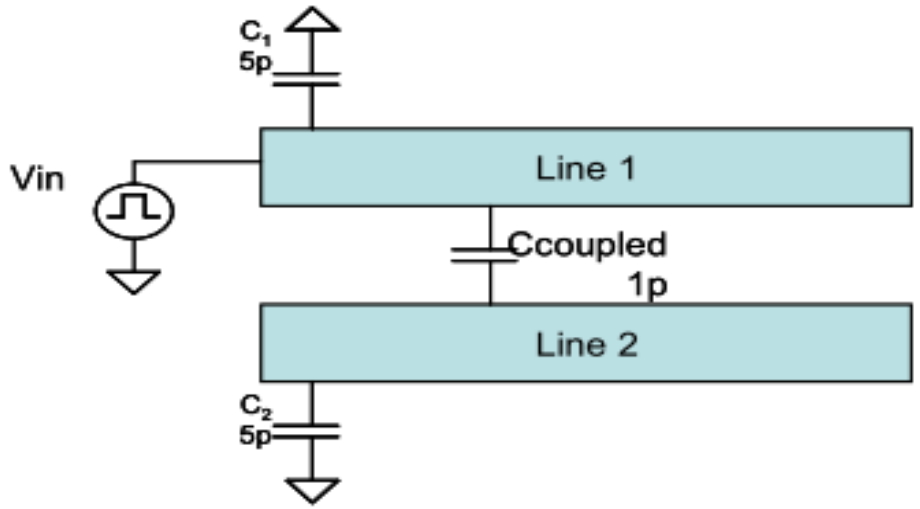
# Example – Coupling Cap

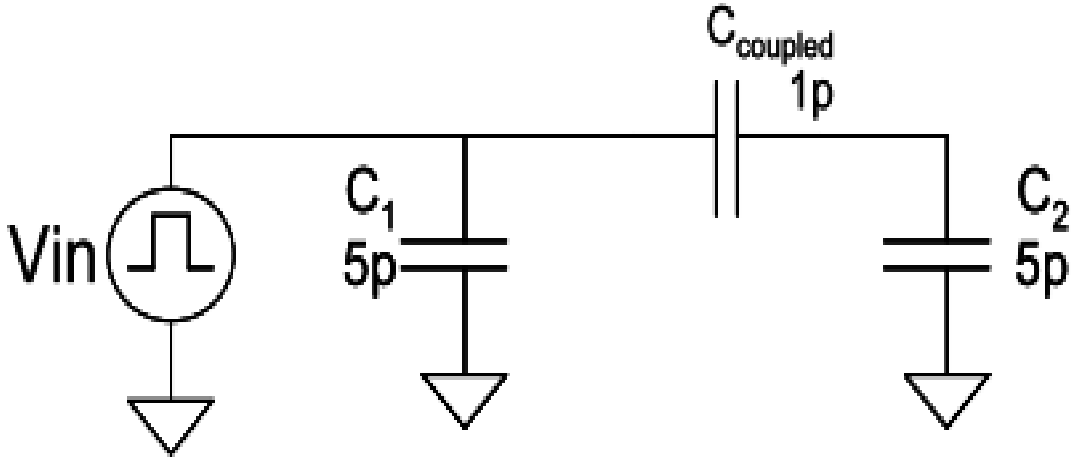
- A pair of wires, each with a capacitance to ground of $5\text{pF}$, have a $1\text{pF}$ coupling capacitance between them.
- A square pulse of $1.8\text{V}$ (relative to ground) is connected to one of the wires.
- How high will the noise pulse be on the other wire?



# Example – Coupling Cap

- **Draw an Equivalent Circuit:**

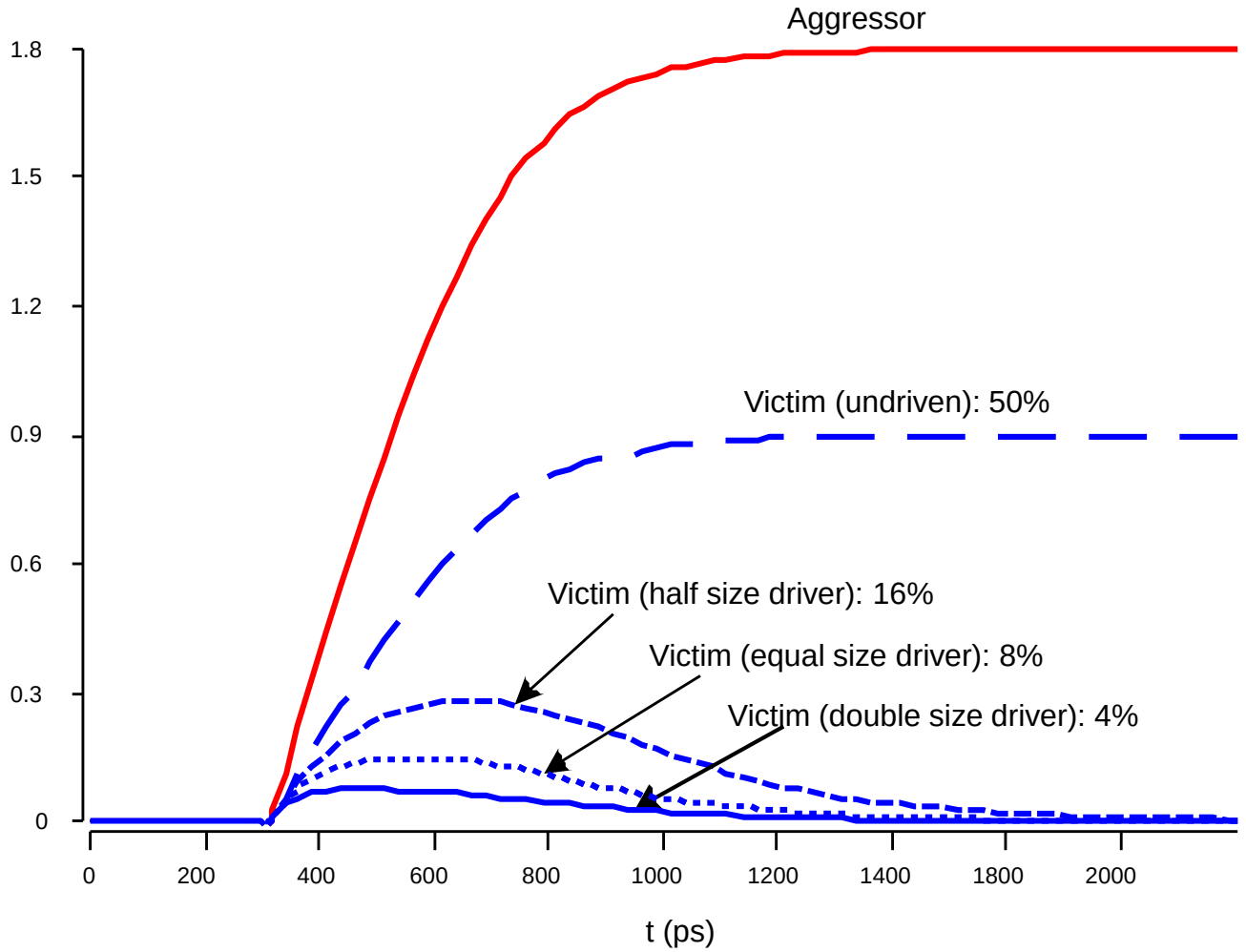



$$V_{C2} = \frac{V_{in} \cdot C_{coupled}}{C_{coupled} + C_2} = \frac{1.8 \cdot 1p}{1p + 5p} = 0.3V$$

# Coupling Waveforms

- **Simulated coupling for $C_{agg}=C_{victim}$**



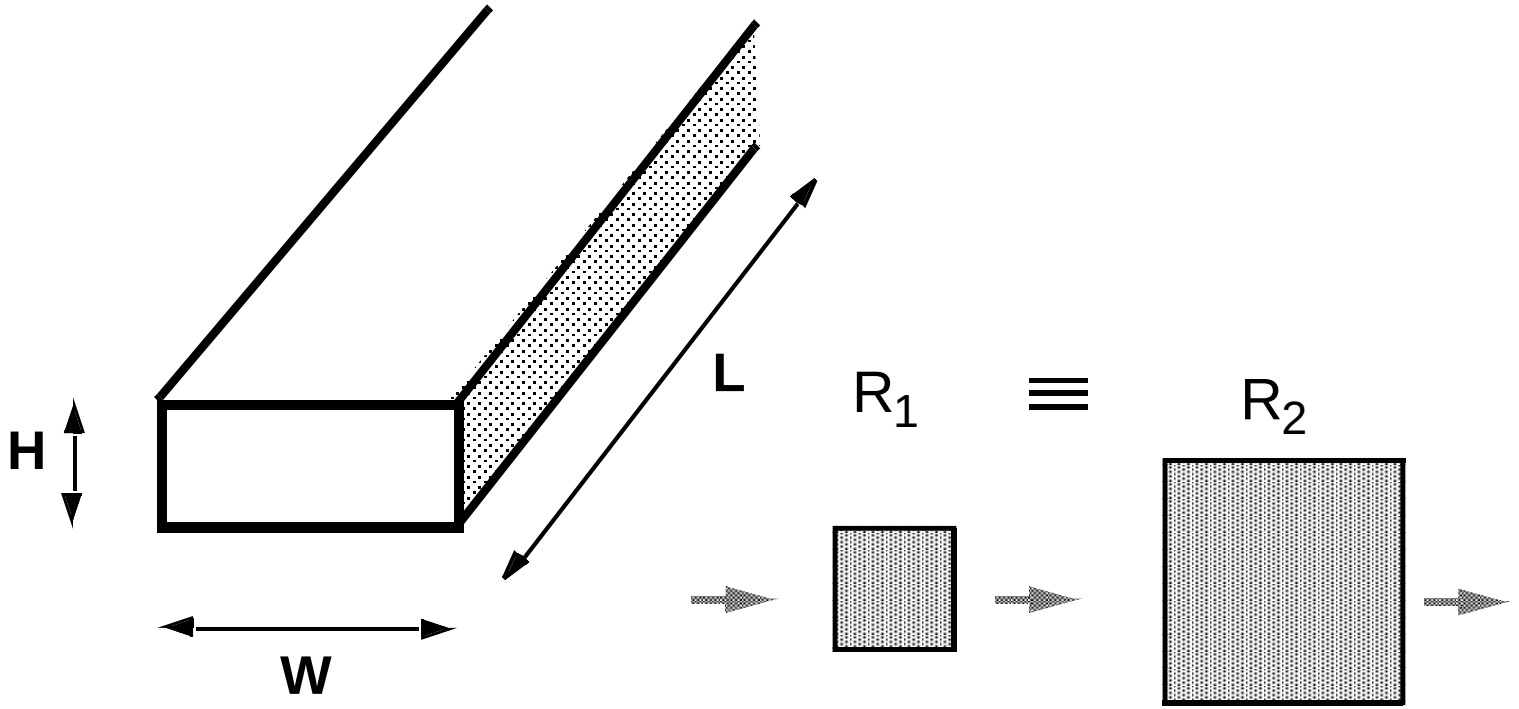
# Shielding

# Feedthrough Cap

# Measuring Capacitance

# Resistance

# Wire Resistance



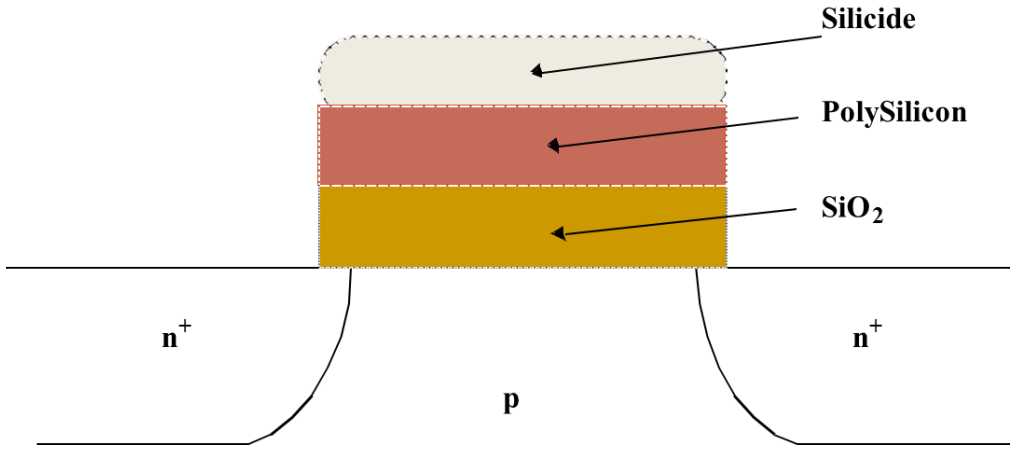

$$R = \rho \frac{L}{A} = \rho \frac{L}{H \cdot W} = R_{sq} \frac{L}{W} \quad , \quad R_{sq} \triangleq \frac{\rho}{H}$$

| 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 for 180nm process**

| Layer | Sheet Resistance (Ω/□) |
|---|---|
| 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 |



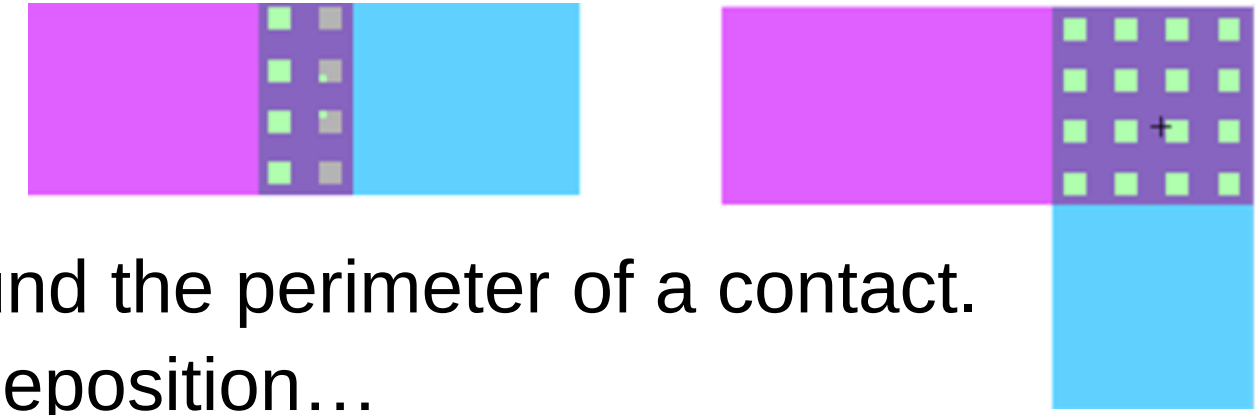

Silicides: $WSi_2$, $TiSi_2$, $PtSi_2$ and TaSi

Conductivity: 8-10 times better than Poly

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

# Contact Resistance

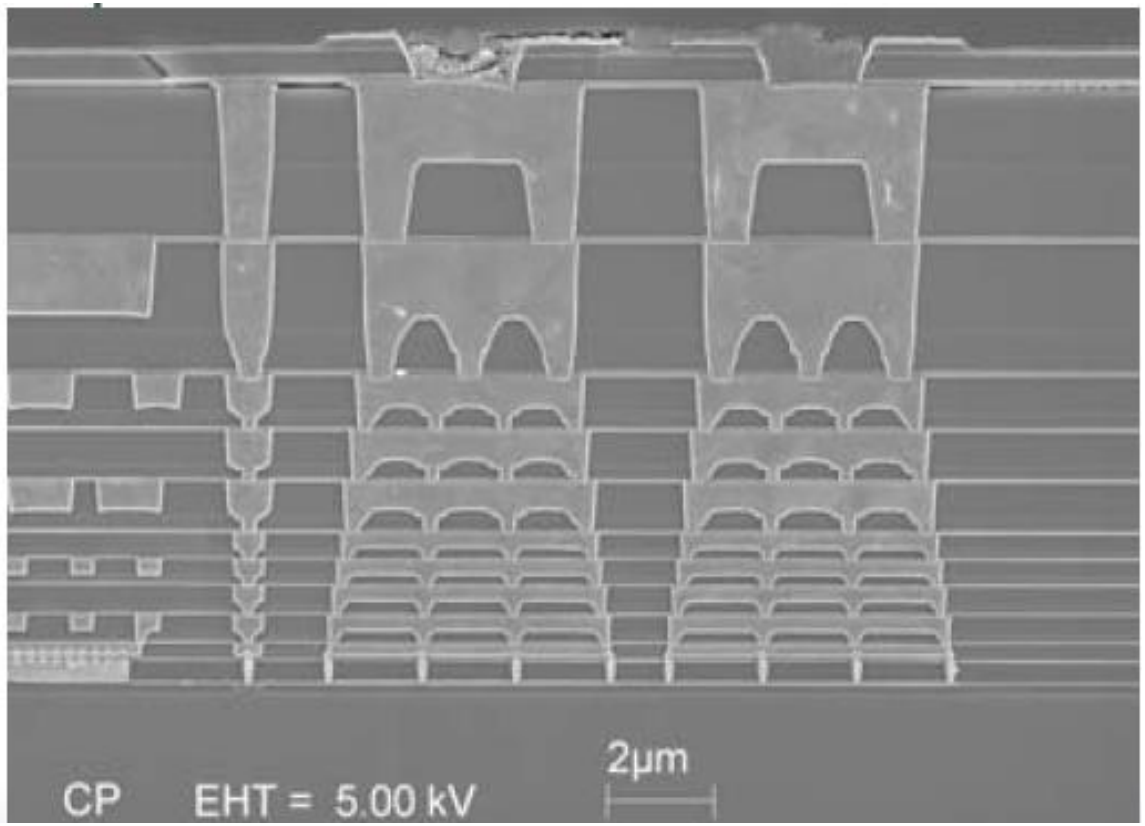
- **Contact/Vias add extra resistance**
  - Similar to changing between roads on the way to a destination…
  - Contact resistance is generally 2-20 Ω



- **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.



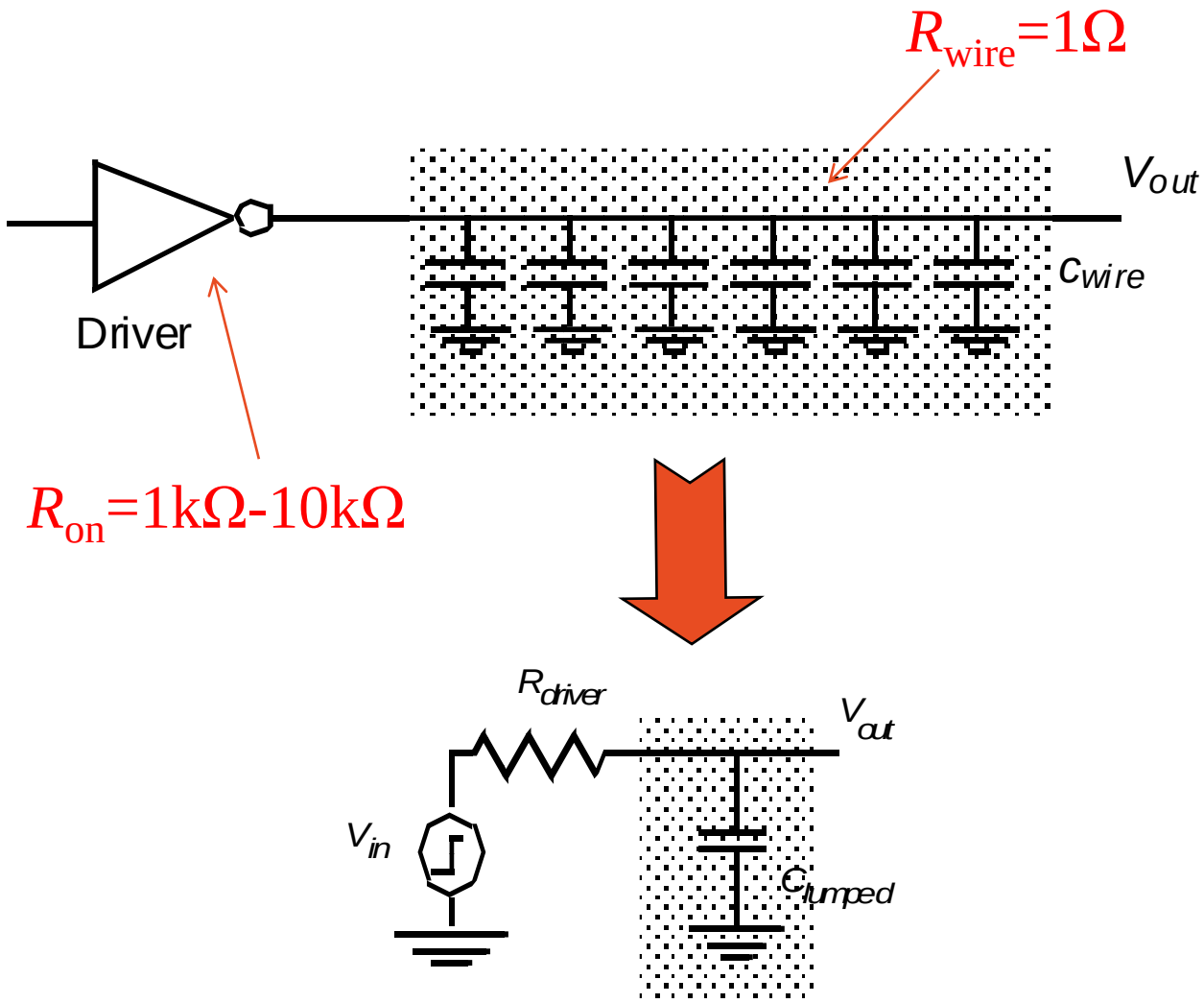

90nm Process

# Interconnect Modeling

# The Ideal Model

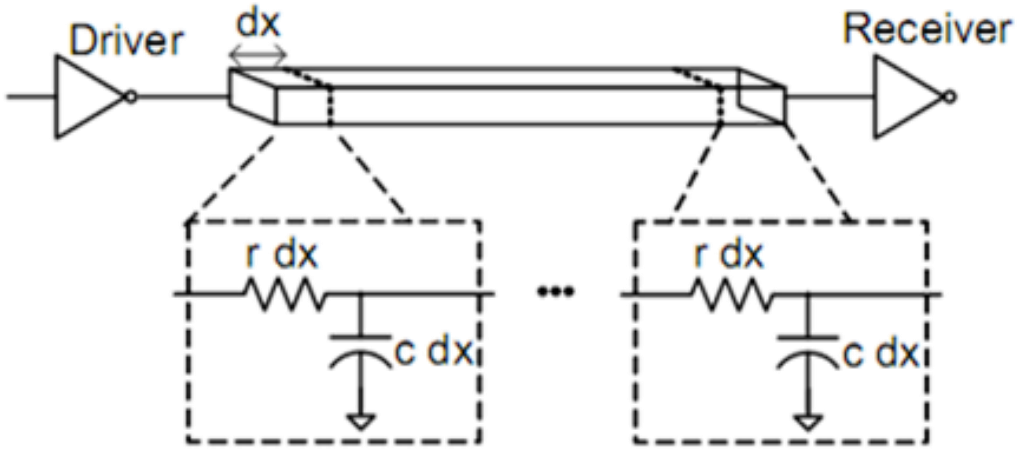
- **In schematics, a wire has no parasitics:**
  - The wire is a single equipotential region.
  - No effect on circuit behavior.
  - Effective in first stages of design and for very short wires.

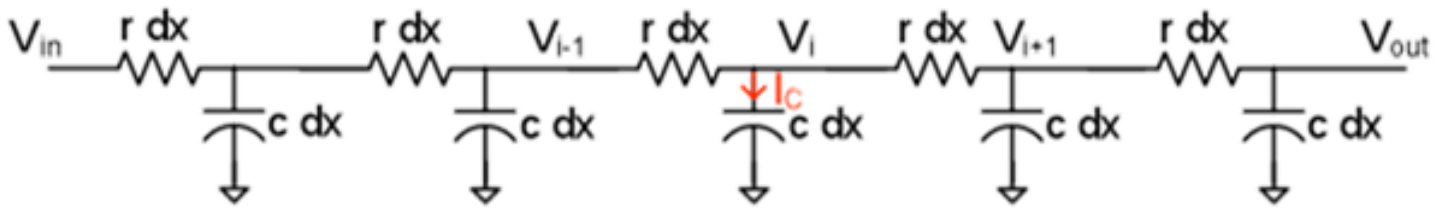
# The Lumped Model



# 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.




$$I_C = \frac{V_{i-1} - V_i}{rdx} - \frac{V_i - V_{i+1}}{rdx} = cdx\frac{dV_i}{dt} \rightarrow rc\frac{\partial V_i}{\partial t} = \frac{\partial^2 V_i}{\partial x^2}$$

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

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

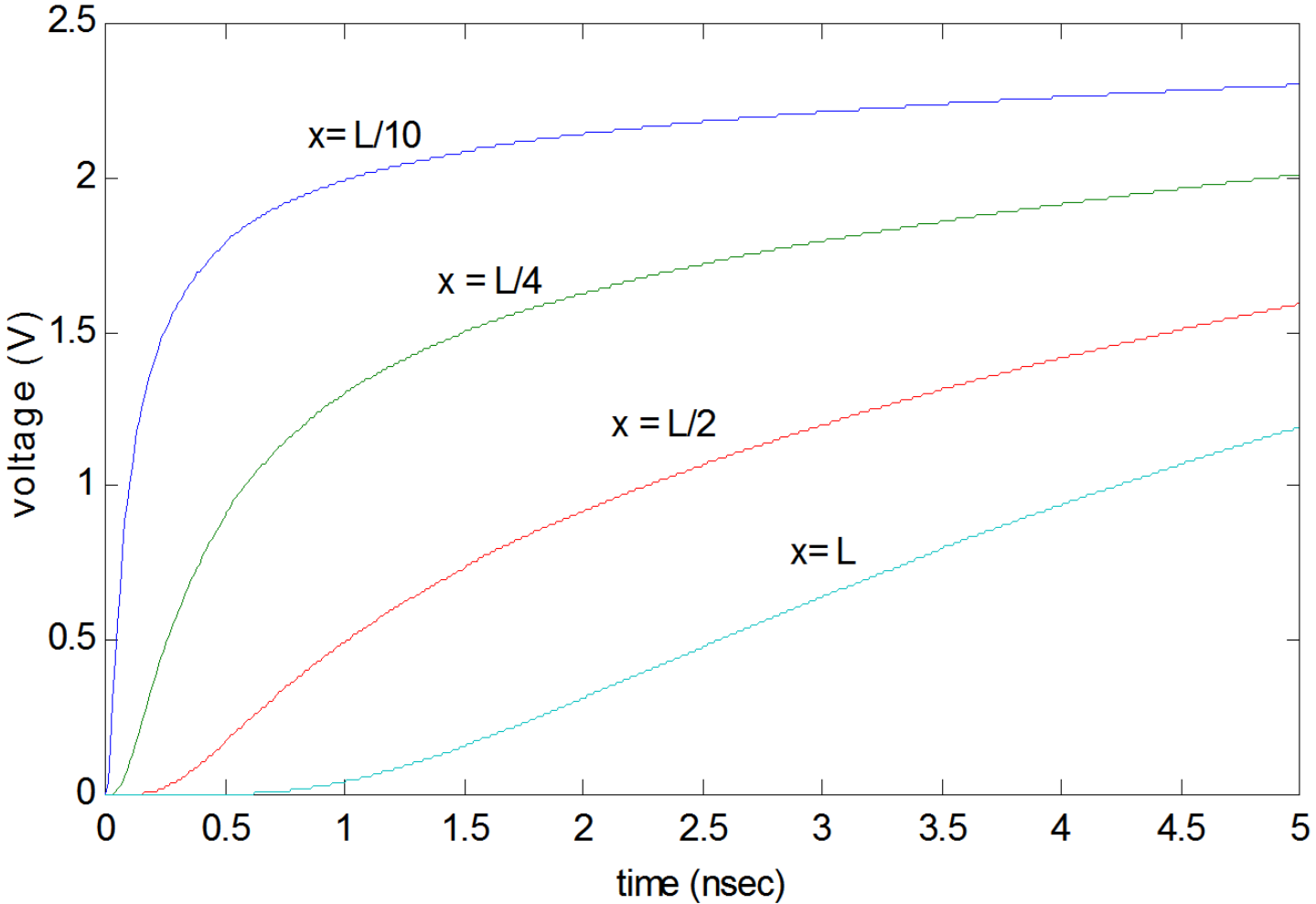
Quadratic dependence on wire length

$$t_{pd} = 0.38RC$$

The lumped model is pessimistic

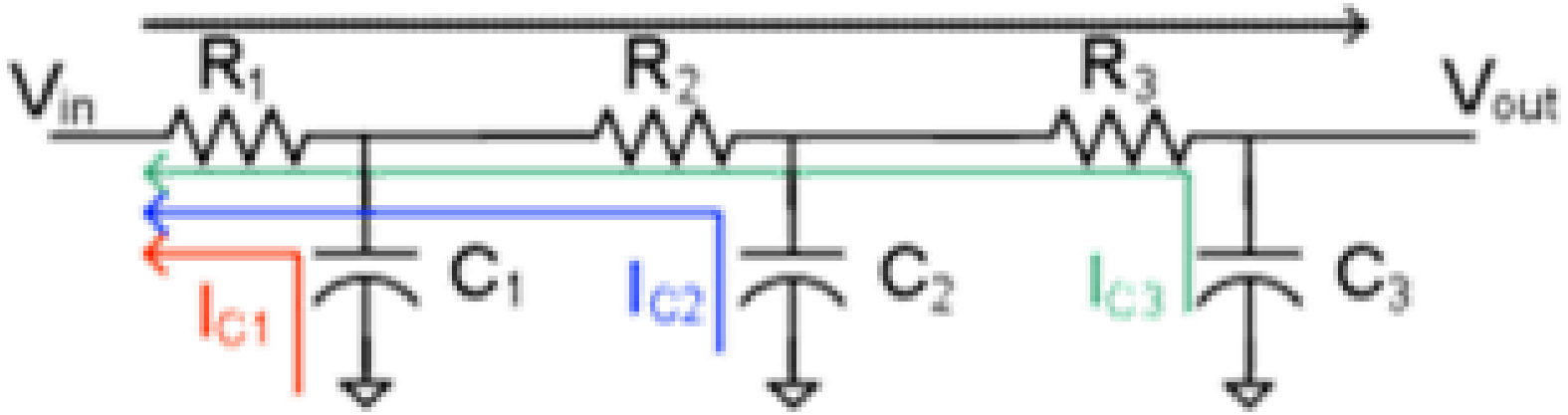
# Step-response of RC wire

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



# 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.




$$\tau_{elmore} = R_1C_1 + (R_1 + R_2)C_2 + (R_1 + R_2 + R_3)C_3$$

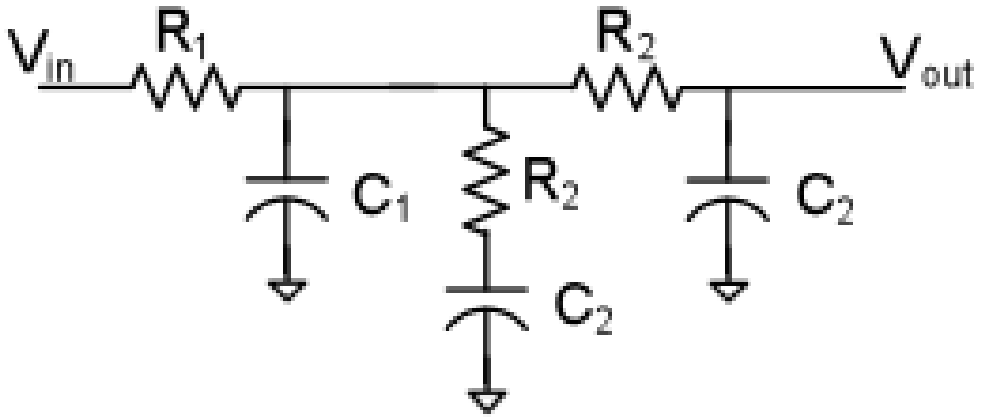
# Elmore Delay Approximation

**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.**

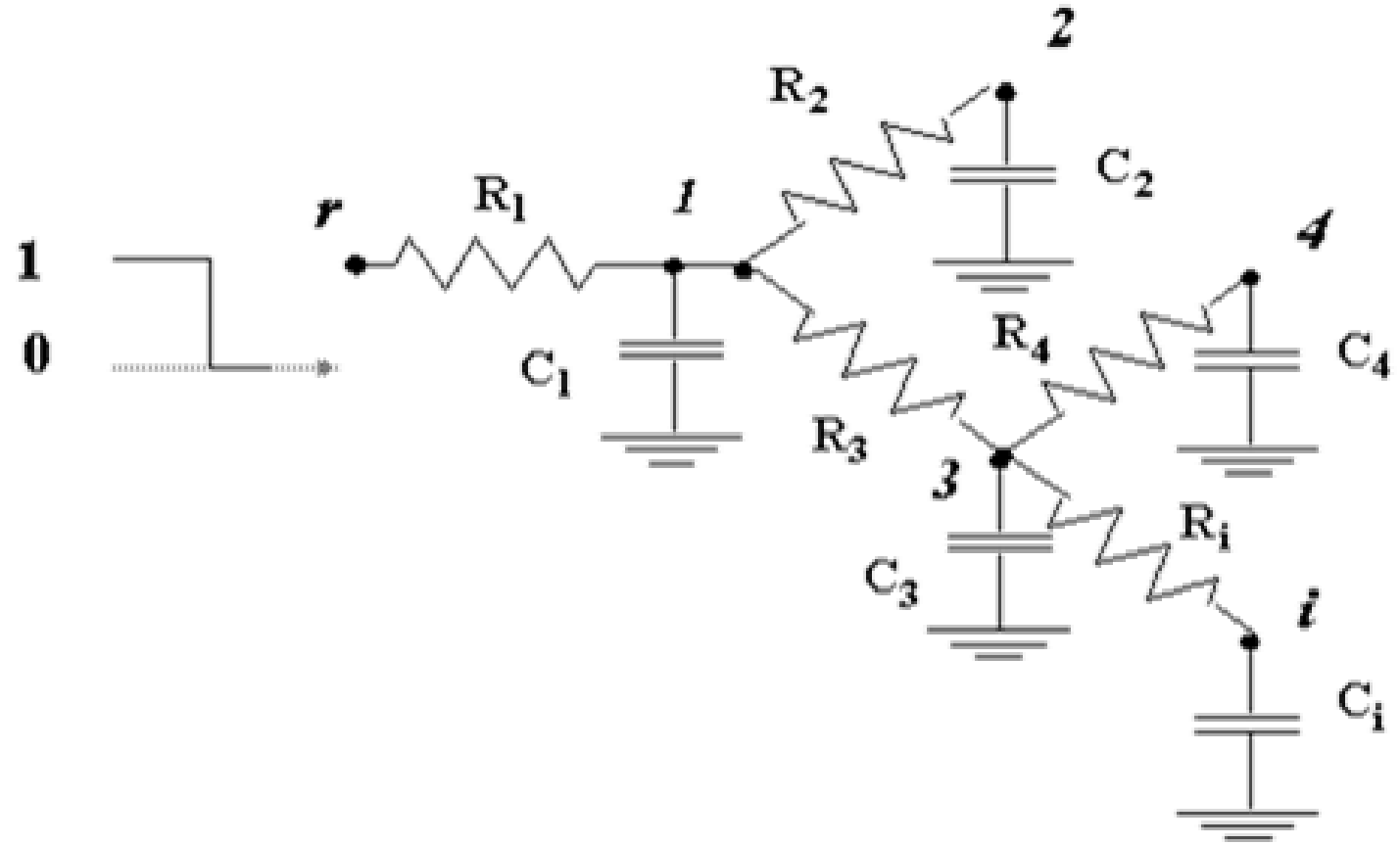
$$R_{ik} = \sum R_j \Rightarrow (R_j \in [path(s \rightarrow i) \cap path(s \rightarrow k)]) \qquad \tau_{Di} = \sum_{k=1}^{N} C_k R_{ik}$$

# Simple Elmore Delay Example




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

# General Elmore Delay Example




$$\tau_{elmore} = R_1C_1 + R_1C_2 + (R_1 + R_3)C_3 + (R_1 + R_3)C_4 + (R_1 + R_3 + R_i)C_i$$

# 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 $rdx$.

$$\tau_N = c\left(\frac{L}{N}\right)\left(r\frac{L}{N}+2r\frac{L}{N}+..+Nr\frac{L}{N}\right)$$

$$=\left(\frac{L}{N}\right)^2\left(rc+2rc+..+Nrc\right)=rcL^2\left(\frac{N\left(N+1\right)}{2N^2}\right)$$

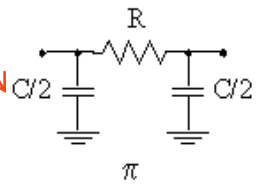
$$\lim_{N\to\infty}\tau_D=\frac{rcL^2}{2}=\frac{RC}{2}$$

# RC-Models

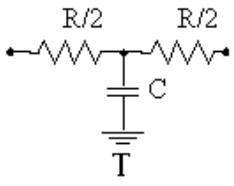
| Voltage Range | Lumped RC-network | Distributed RC-network |
|---|---|---|
| 0→50% ($t_p$) | 0.69 RC | 0.38 RC |
| 0→63% ($\tau$) | RC | 0.5 RC |
| 10%→90% ($t_r$) | 2.2 RC | 0.9 RC |

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

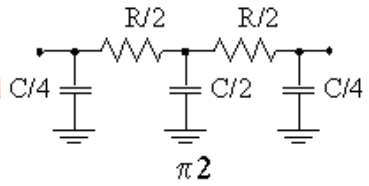
Pie Model



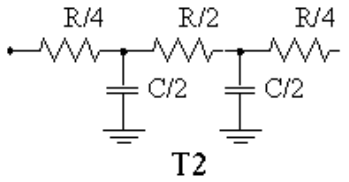

T-Model



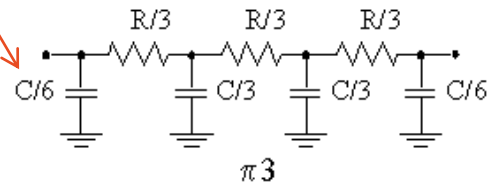

Pie-2 Model



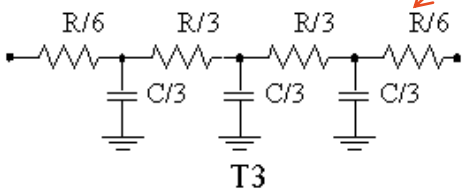

T-2 Model




Pie-3 Model




T-3 Model



# Wire Delay Example

- **Inverter driving a wire and a load cap.**

$$\tau_{driver} = \left( C_d + C_W/2 \right) R_{inv} + \left( C_{ext} + C_W/2 \right) \left( R_{inv} + R_w \right)$$

# A different look…

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

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

- **Again we'll look at our driver with a distributed wire.**
  - 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.

$$\tau_D = 0.69 R_{inv}\left(C_d + C_W\right) + 0.38 R_W C_W + 0.69\left(R_{inv} + R_w\right) C_L$$

# Dealing with long wires

- **Repeater Insertion**

# Dealing with long wires

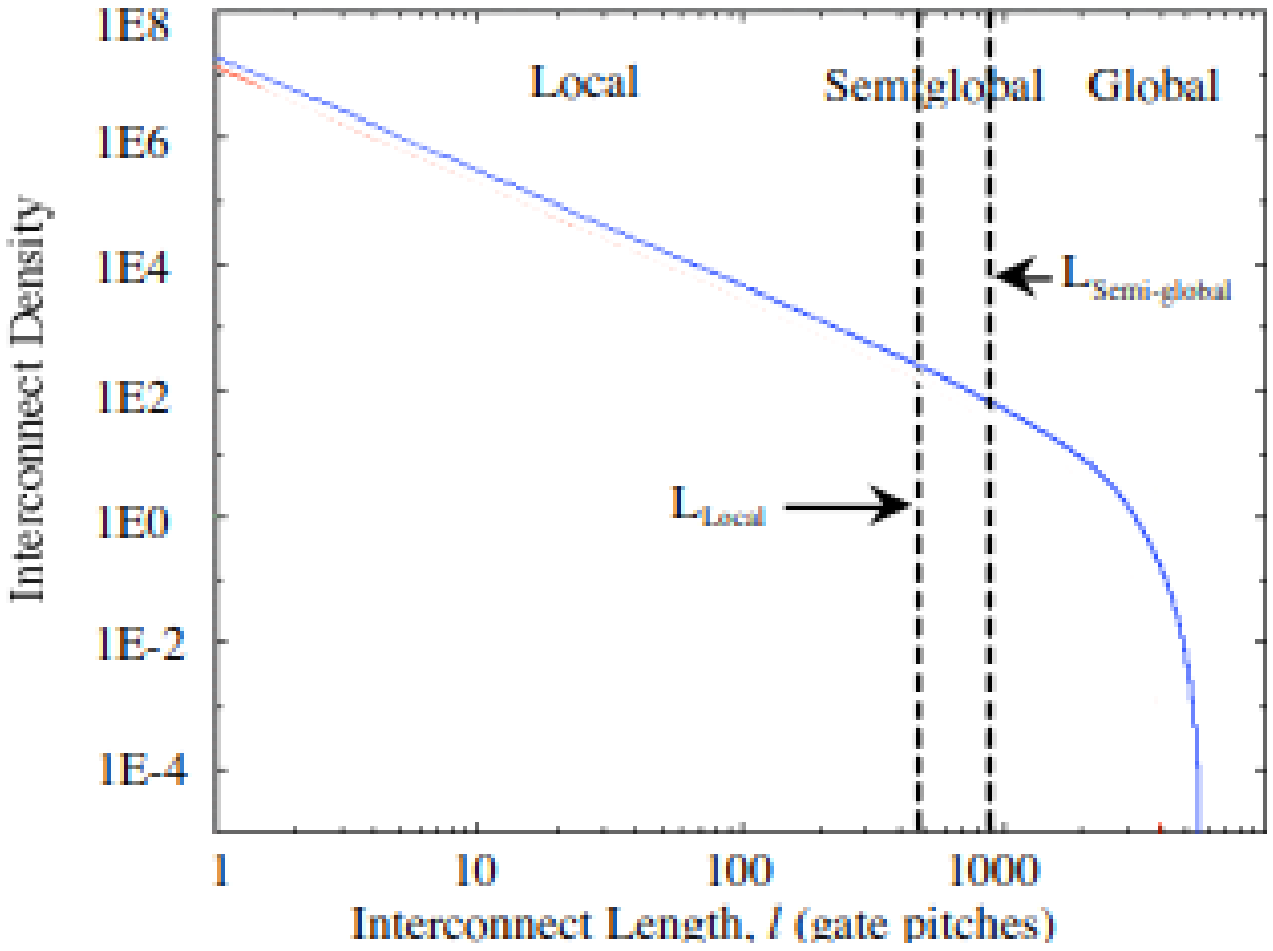
- **Buffer Tree Insertion**

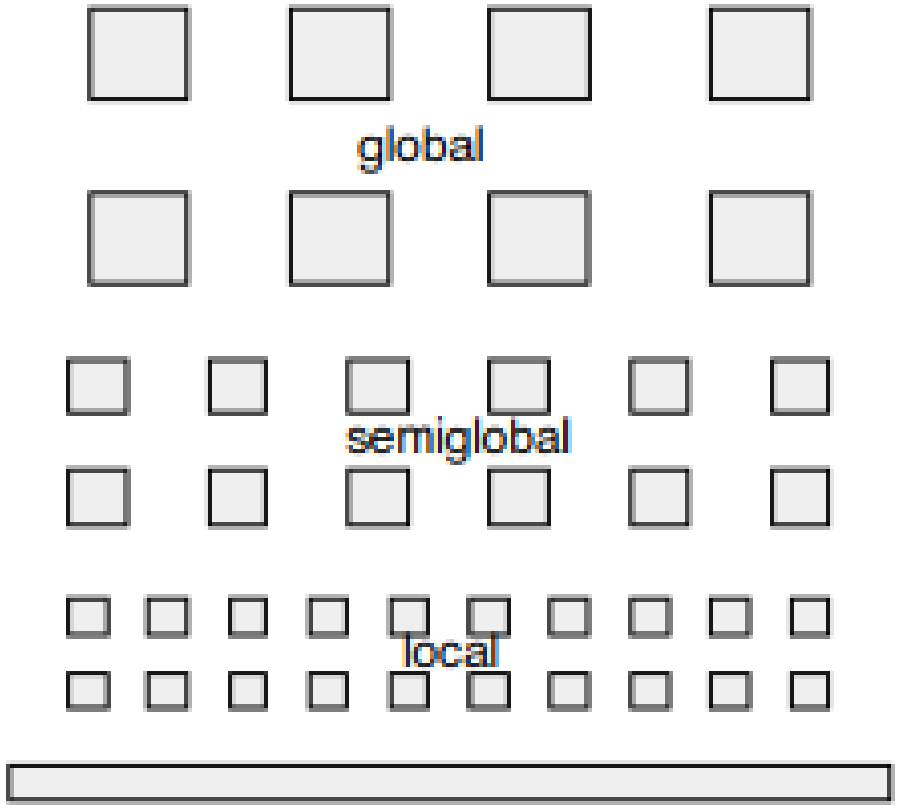
# Wire Scaling

# 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 has increased by 6% per year (X2 @10 years)
  - Devices have scaled, but complexity has grown!

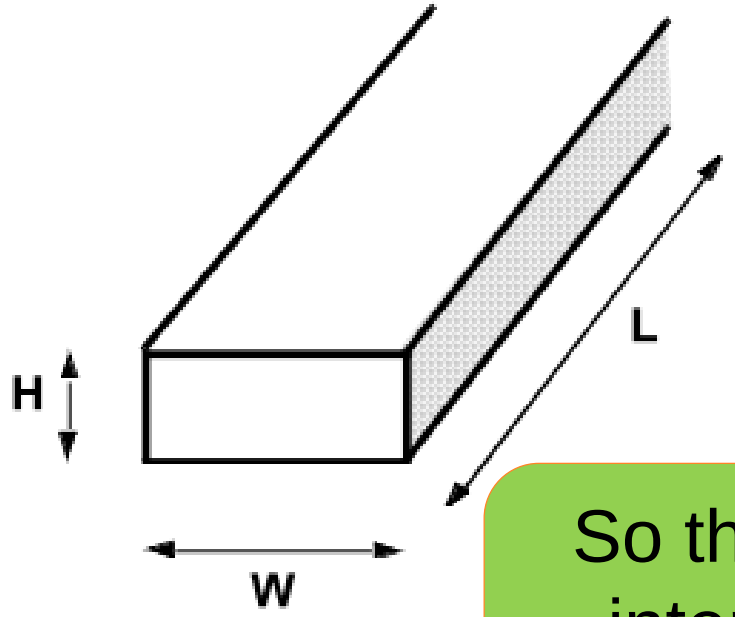
# Nature of Interconnect



# Local Wire Scaling

- **Looking at local interconnect:**
  - $W$, $H$, $t$, $L$ all scale at $1/S$
  - $C=LW/t \rightarrow 1/S$
  - $R=L/WH \rightarrow S$
  - $RC=1$



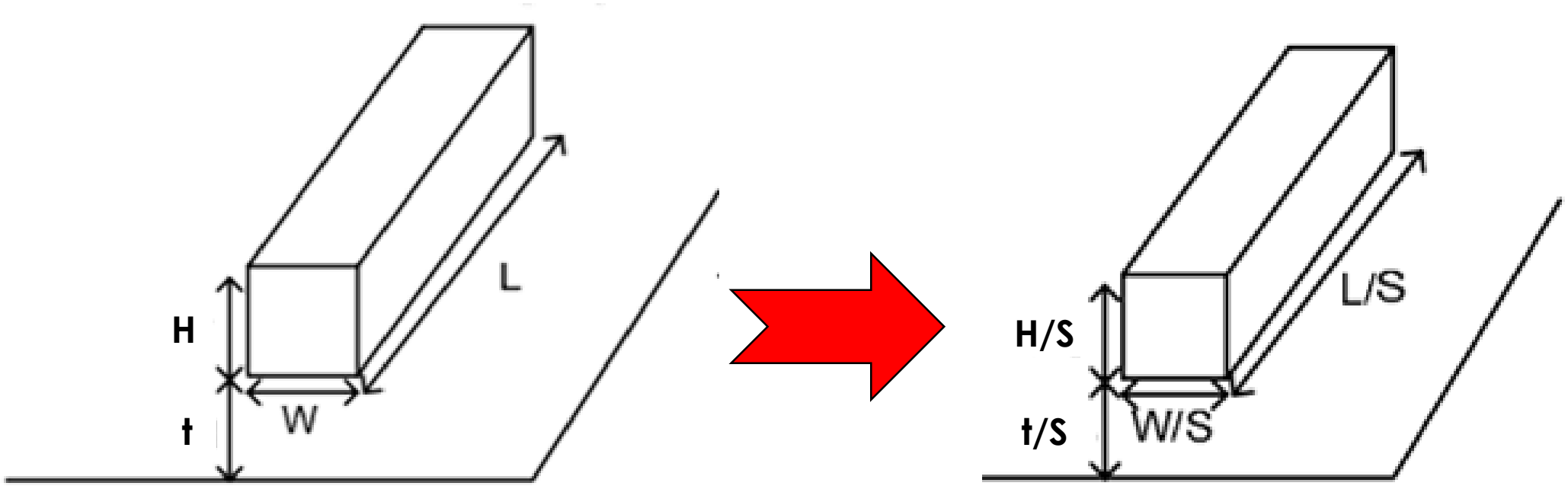

So the delay of local interconnect stays constant.

- **Reminder – Full Scaling of Transistors**
  - $R_{on}=V_{DD}/I_{on}$ α 1
  - $t_{pd}=R_{on}C_g$ α $1/S$

But the delay of local interconnect increases relative to transistors!

# Local Wire Scaling – Full Scaling

- **What about fringe cap?**




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

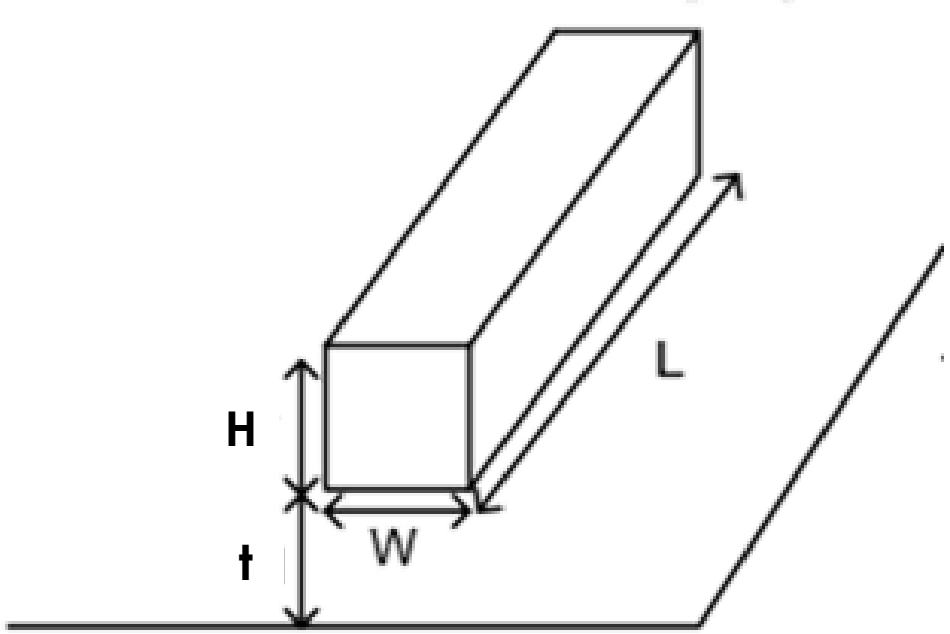
$$R_{wire} \propto L/WH \qquad t_{p,wire} \propto R_{wire}C_{wire}$$

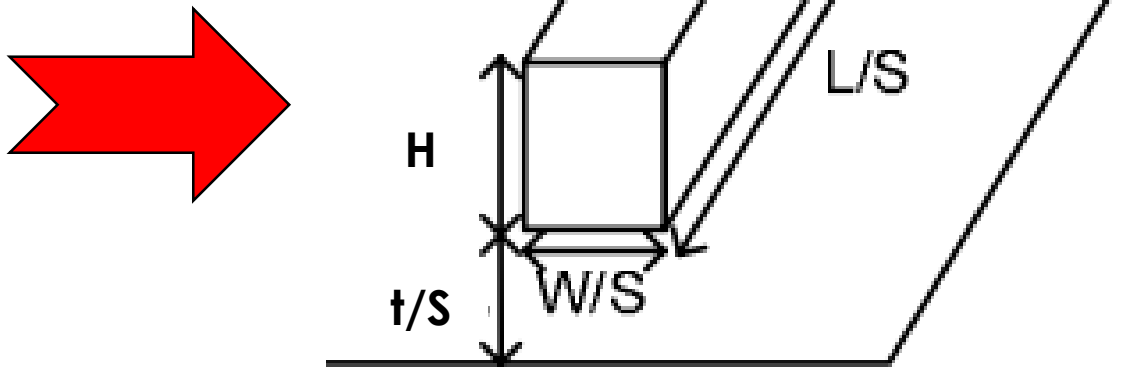
$$C_{pp} \propto S^{-1} \qquad C_{fringe} \propto S^{-1}$$

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

# Local Wire Scaling - Constant Thickness

- **Wire thickness (height) wasn't scaled!**






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

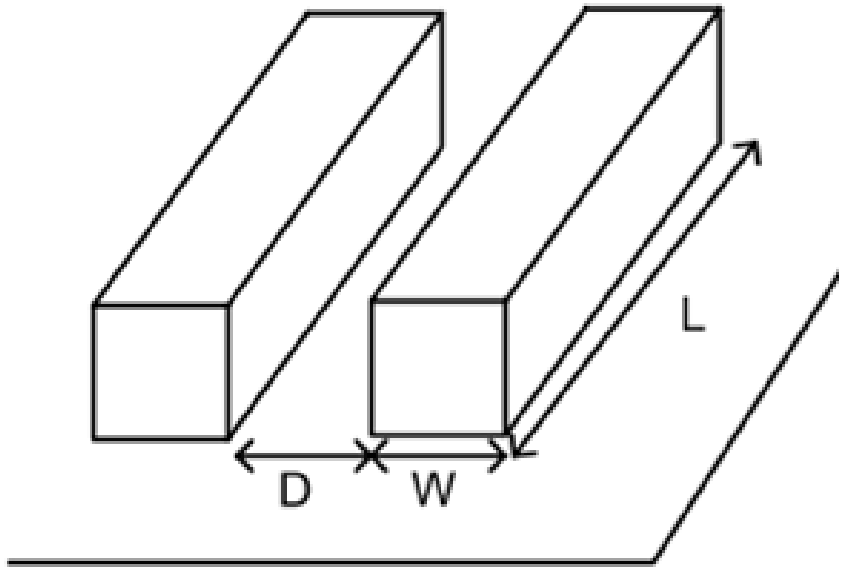
$$R_{wire} \propto L/WH \qquad t_{p,wire} \propto R_{wire}C_{wire}$$

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

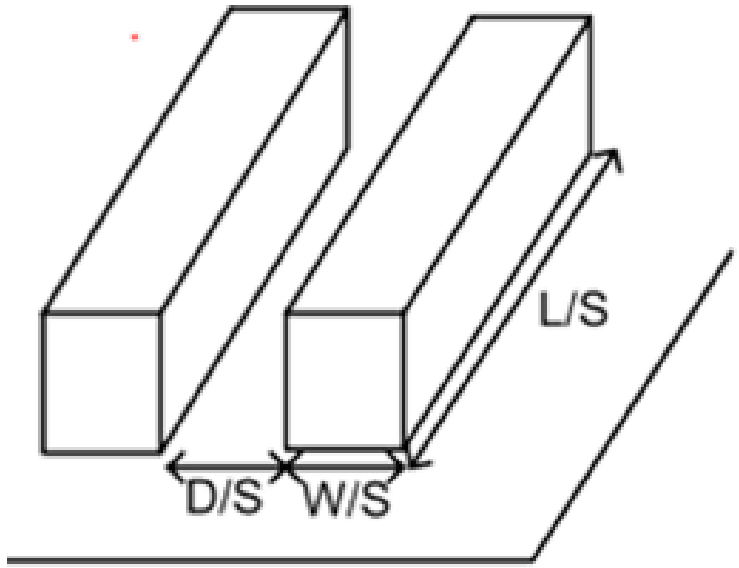
$$R_{wire} \propto \text{const} \qquad t_{p,wire} \propto S^{-1}$$

# Local Wire Scaling – Interwire Capacitance

- Without scaling height, coupling gets much worse.
- Aspect ratio is limited and we eventually have to scale the height.
- Therefore, different metal layers have different heights.



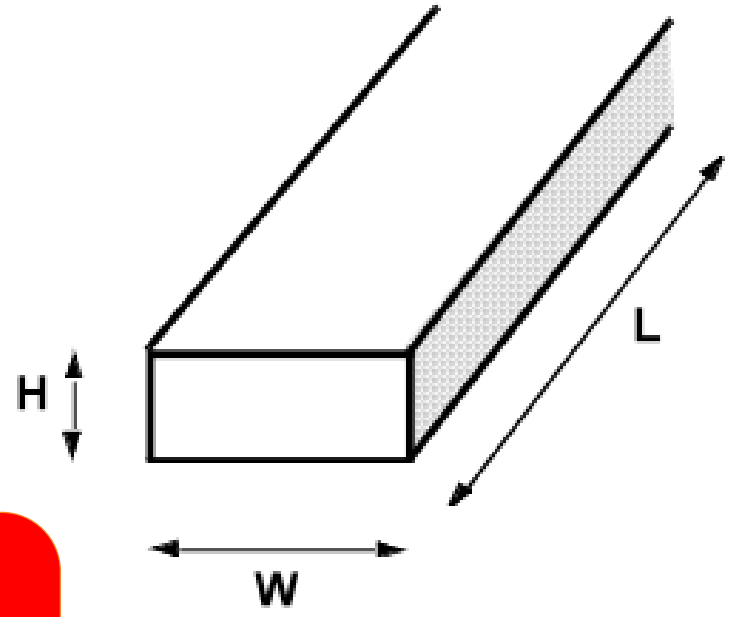

$$C_{\text{pp, side}} \propto LH/D$$




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

# Global Wire Scaling

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

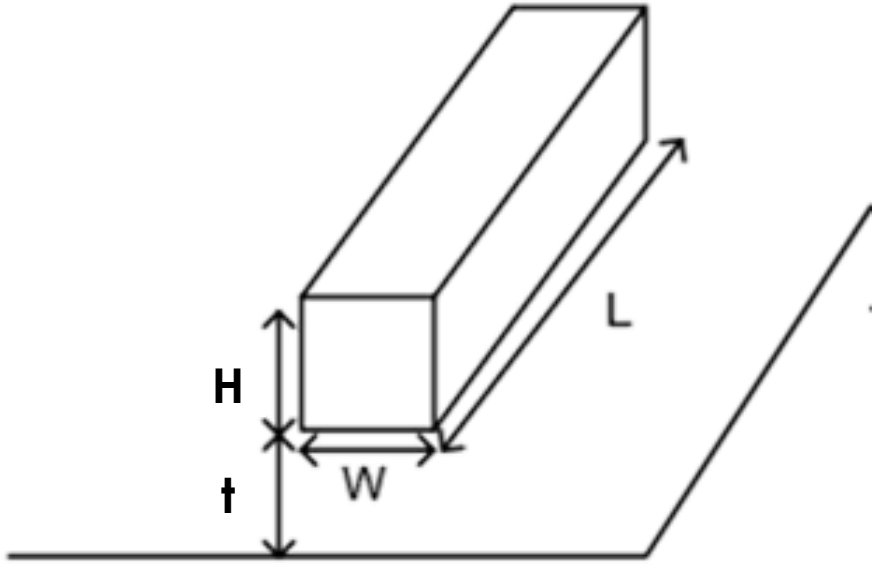


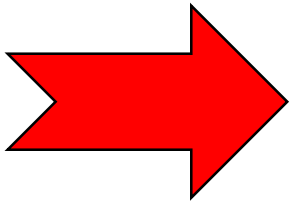

Long wire delay increases

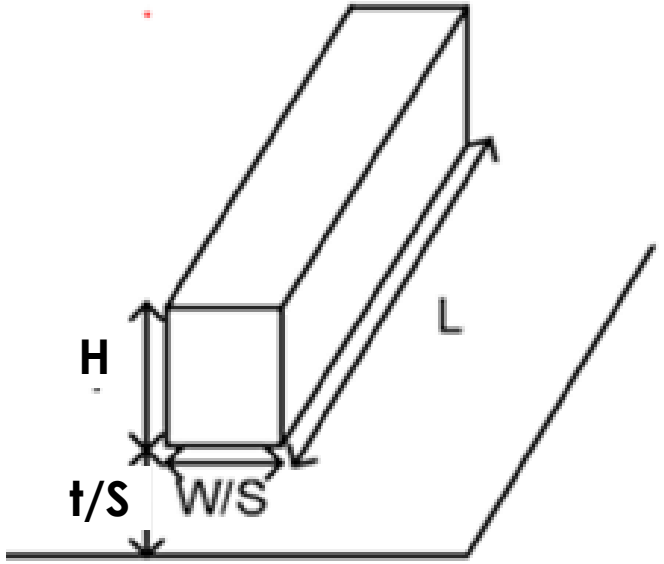
- **And if chip size grows, $L$ actually increases!**

# Global Wire Scaling – Constant Thickness

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






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

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

$$C_{pp} \propto \text{const} \qquad C_{fringe} \propto \text{const}$$

$$R_{wire} \propto S \qquad t_{p,wire} \propto S$$

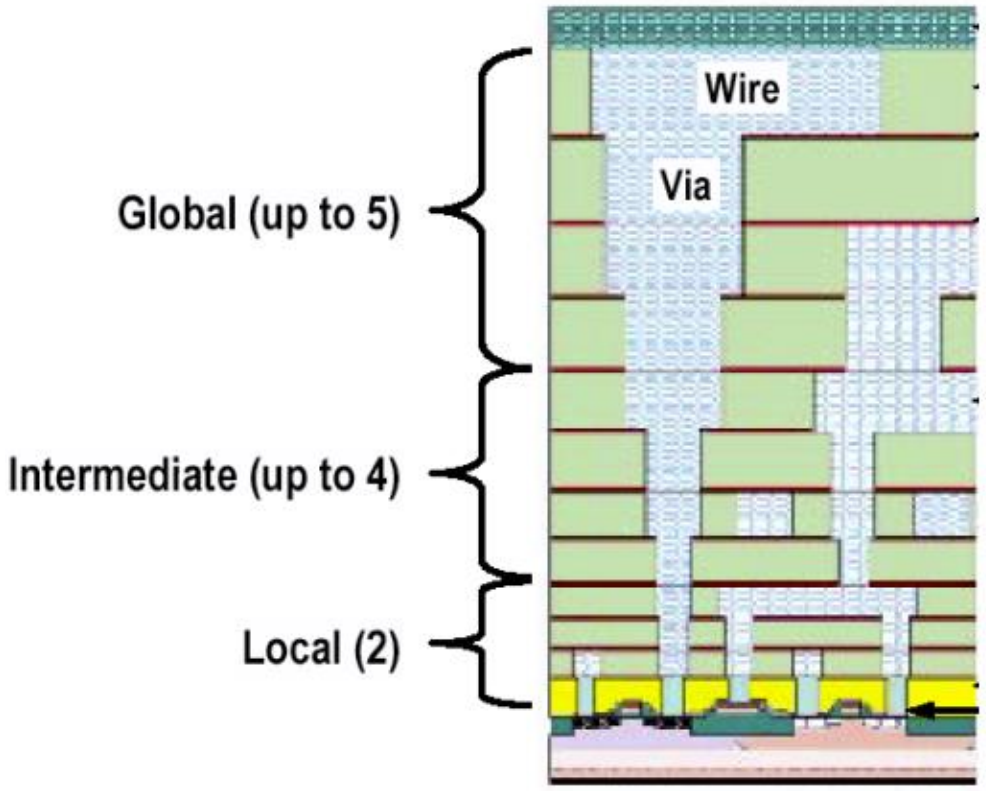
# Wire Scaling

- **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 capacitance increases, so this is optimistic.

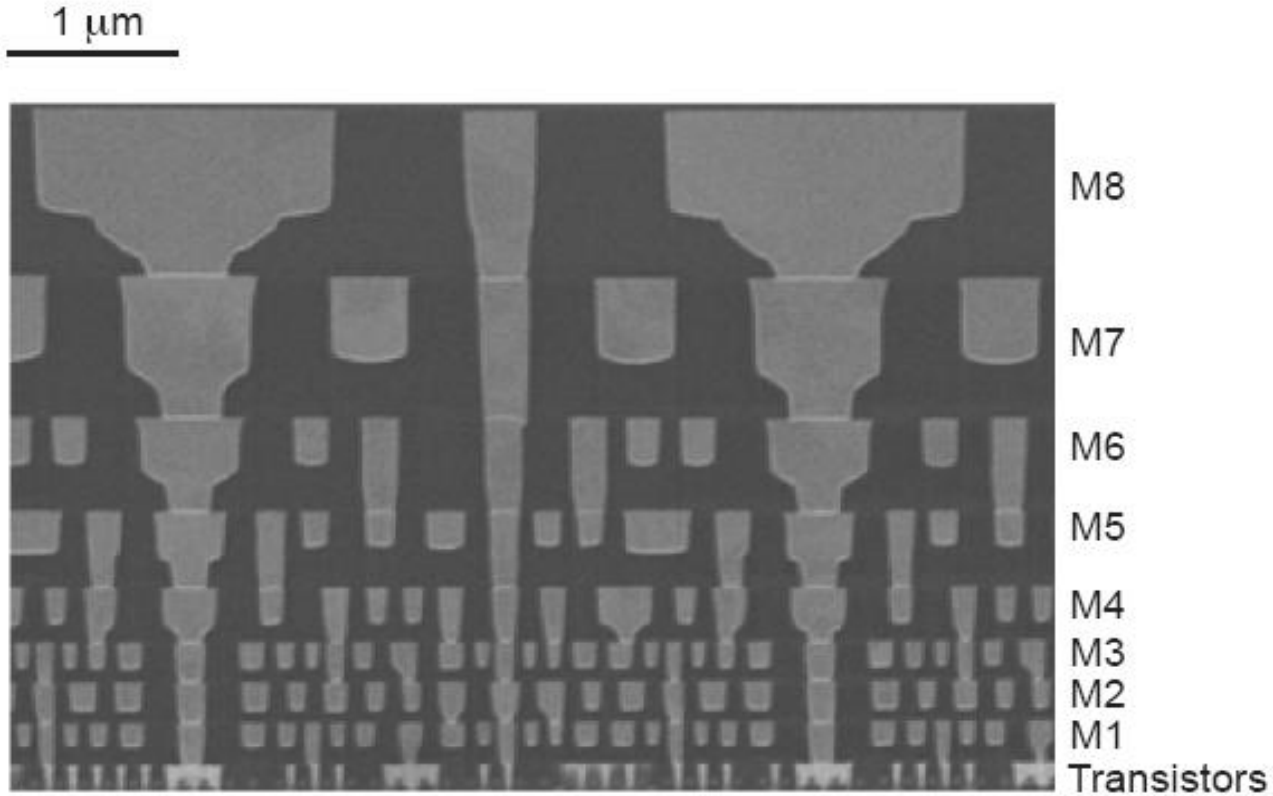
# Wire Scaling

- **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]

# 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. Nicolic, Berkeley EE-241, Lecture 3 (Spring 2011)
  http://bwrc.eecs.berkeley.edu/classes/icdesign/ee241_s11
- Stanford EE311