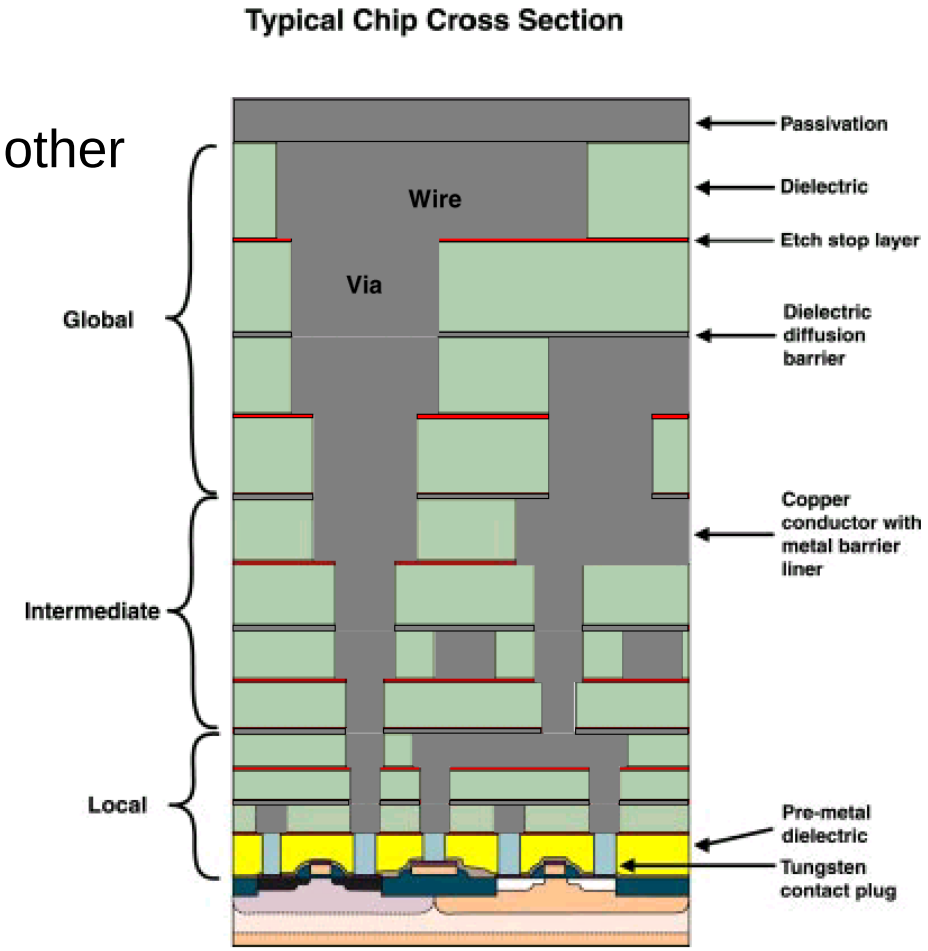
EE-429
Fundamentals of VLSI Design

Interconnect

Andreas Burg

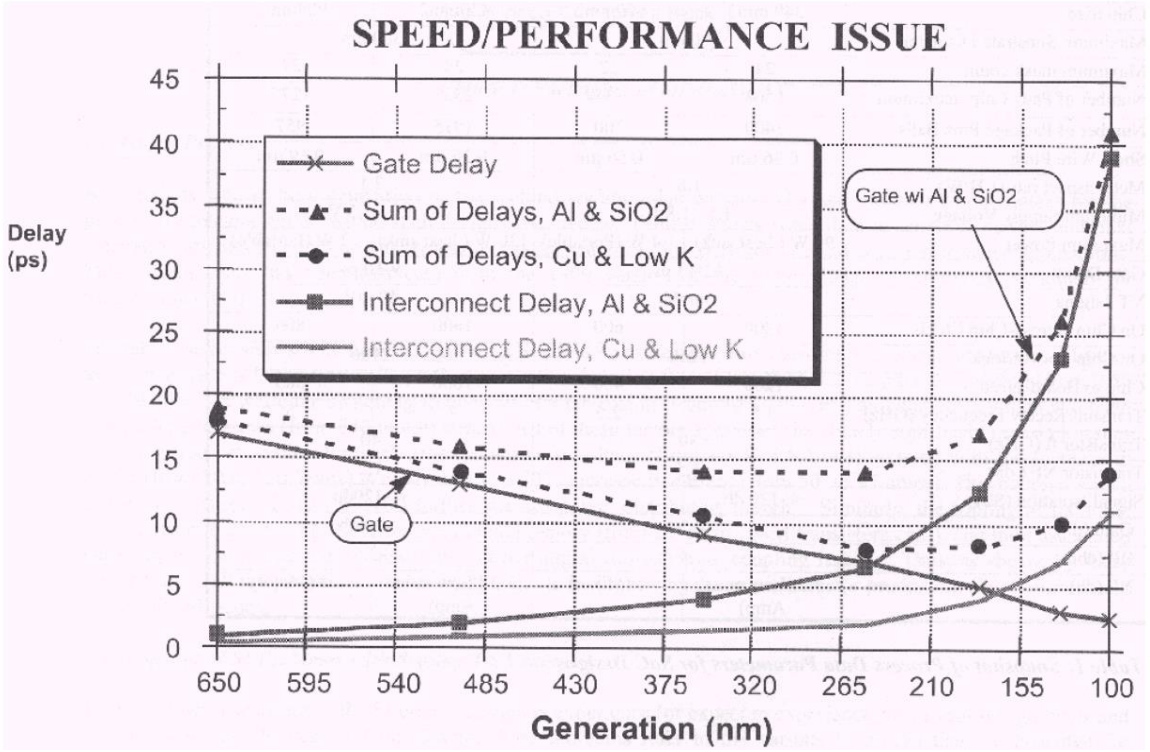
Interconnect Structures

- Integrated circuits offer **multiple levels of interconnect**
  - Wires on multiple metal layers used for routing
  - Vias connect the different layers
  - Dielectric field oxide separates/isolates layers from each other
- **Interconnect feature size and dimensions increase with each layer**
  - Different layers typically serve different purposes
  - Reaching to higher layers requires “punching through” lower layers
- **Active components are small compared to wire stack**
- **Fabrication of the interconnect is known as the Back End of Line (BEOL)**



Importance of Interconnect

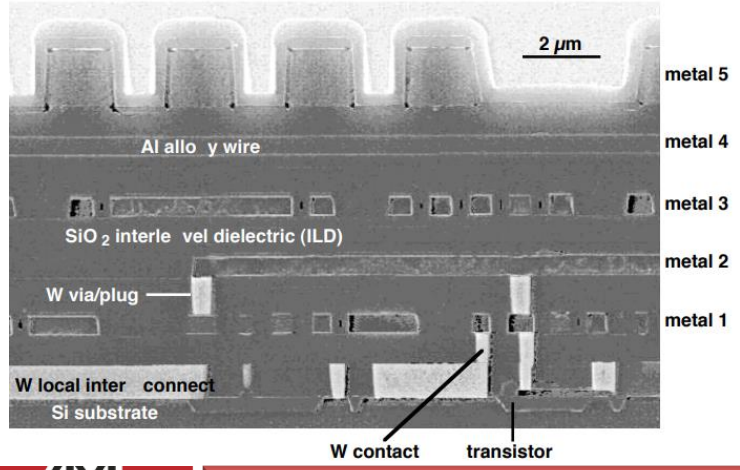
- **Parasetics** from wires **introduce delay** and **consume power**
- **Interconnect does not scale well** as feature size shrinks, while gate delay decreases
- **Interconnect becomes a dominant factor in modern technologies**
  - Important to
    - avoid long interconnects if possible
    - consider interconnect delay duting design
  - Need for accurate models



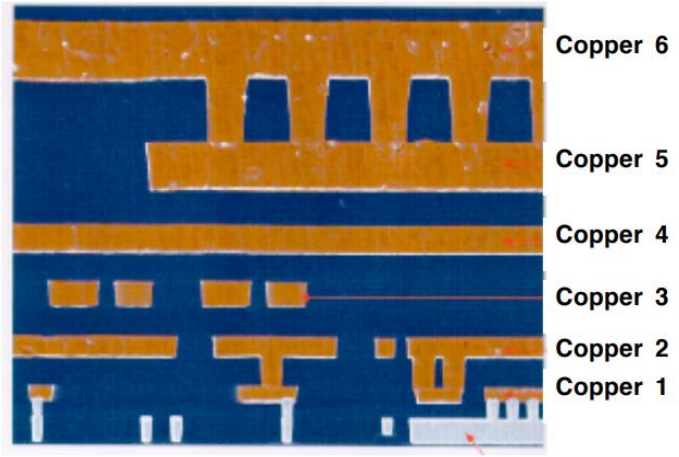
Interconnect Evolution

- **Established technologies up to 180nm:** focus on devices not on interconnect
  - 3-5 layers of interconnect made from aluminum
- **Technology nodes <90nm:** interconnect becomes more relevant
  - More **layers (>6-10 layers**), many **made from copper**: 40% less resistance
  - **Low-k dielectrics** (available below 90nm) reduce wire capacitance by 30-50%

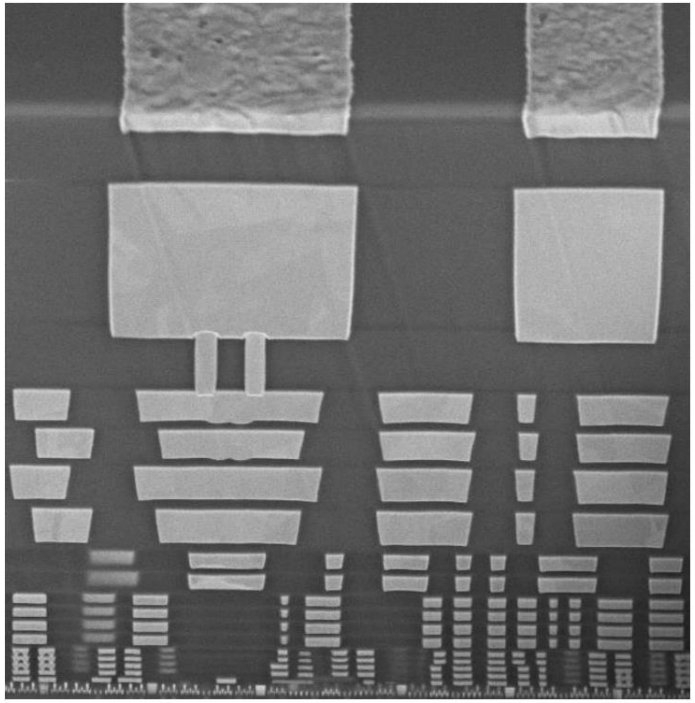
A Motorola μprocessor with 5 Al layers



IBMs 7 layer Copper Metallization Technology (1998)

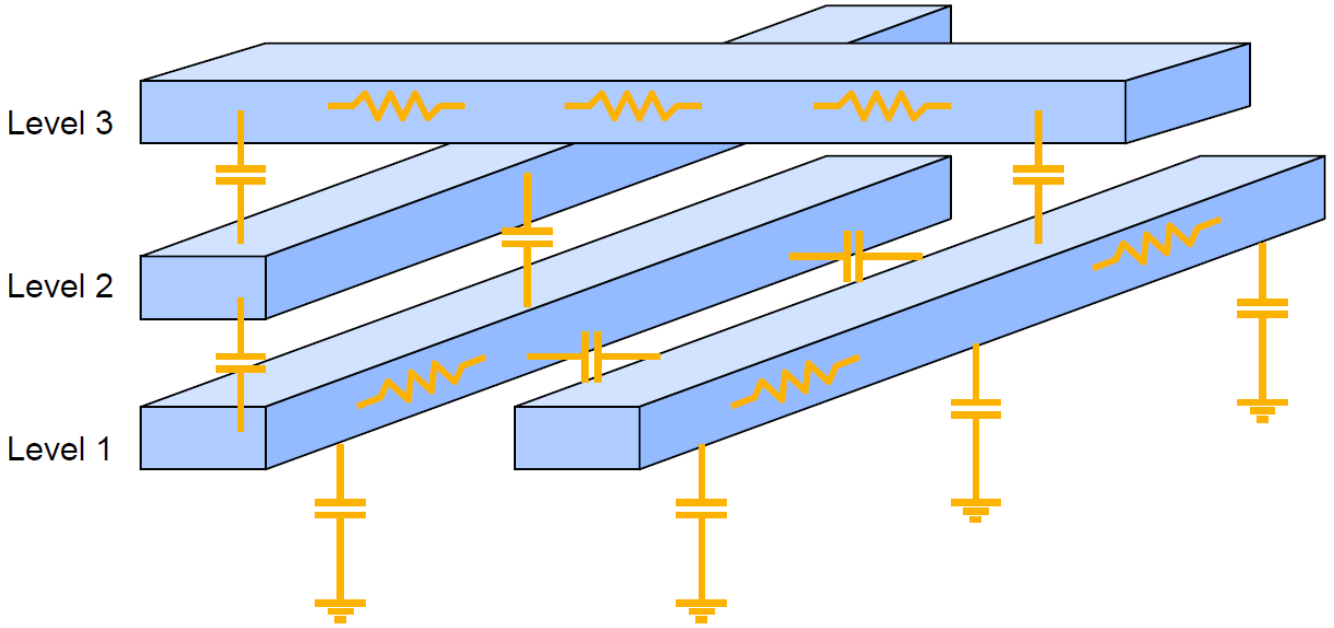


17 layers metal stack of a modern process



Multi-Layer Interconnect Model

- **Wires** are **modeled as their parasetic elements**



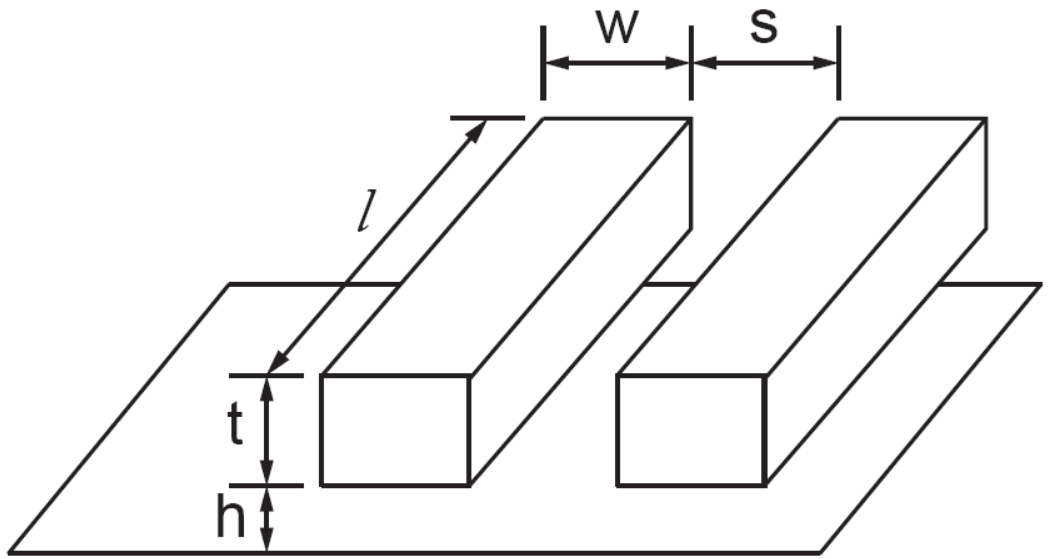
Capacitance to the bulk (GND)

Coupling capacitcnce between wires

Series resistance of wire segments

Basic Wire Model

- **Wire is defined by its geometrical properties (partially under control of the designer) and its electrical properties**



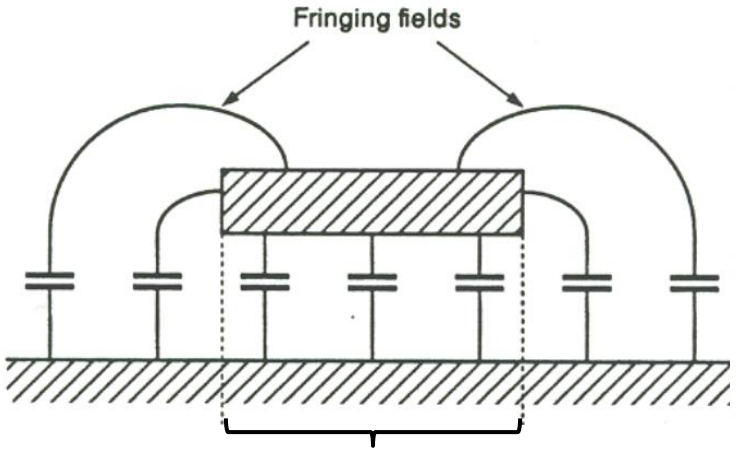
Key geometry parameters:

W:	Width
l:	Length
t:	Thickness
h:	Height above substrate (GND)
S:	Spacing between wires

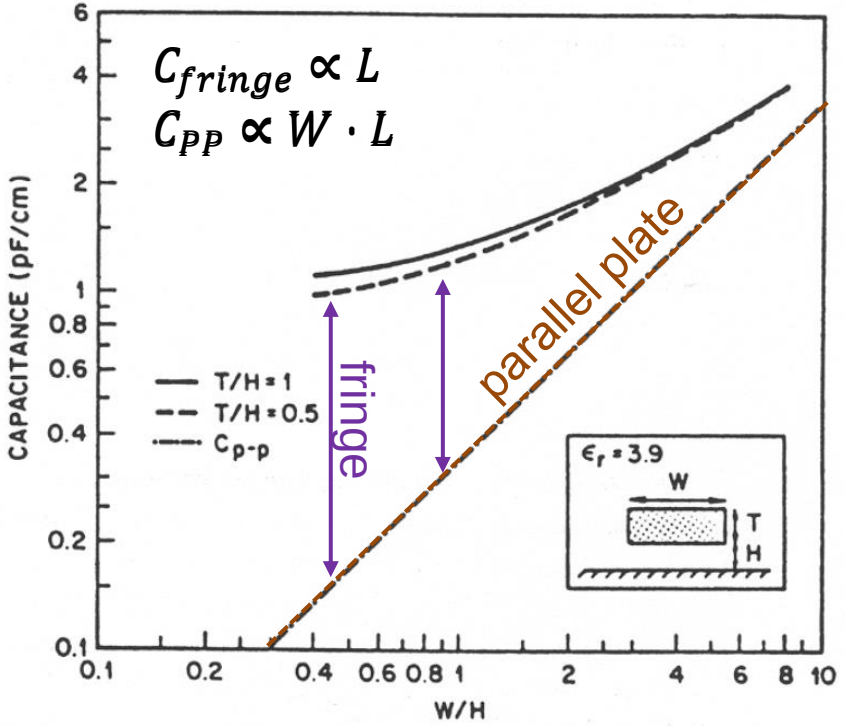
- Properties depend on the technology, the chosen layer, and the drawn layout
  - Assist the designer in making optimizing the layout
  - Allow to assess the impact of technology scaling

Inter-Layer Capacitance

- **Parallel plate with fringe capacitance:** wire over ground plane (substrate)
  - Ignores other wires
  - **Two types of capacitance**
    - **Parallel plate capacitance**
    - **Fringe capacitance:** can easily dominate, especially for narrow wires (very typical in nm-technologies)



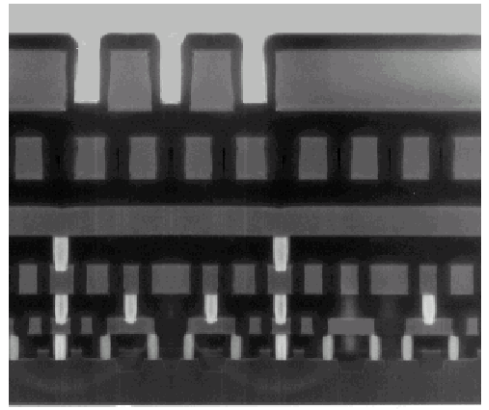
$$C_{pp} = \frac{\kappa\epsilon_0}{H} WL$$



- **Estimating substrate capacitance from wire geometries**
  - Based on **constants provided in the PDK**

Parallel plate	Fringe
$C_{PP} = C^0_{pp} \cdot W \cdot L$	$C_{fringe} = C^0_{fringe} \cdot L$

$$C_{\text{wire}} = C_{PP} + C_{fringe} = C^0(W^0) \cdot \frac{W}{W^0} \cdot L$$



Permittivity ($\kappa\epsilon_0$)

- **The specific** parallel plate and fringe **capacitances depend on** an important process parameter: **the relative permittivity** κ
  - Different materials have different relative permittivity κ

$$C \propto \kappa\epsilon_0$$

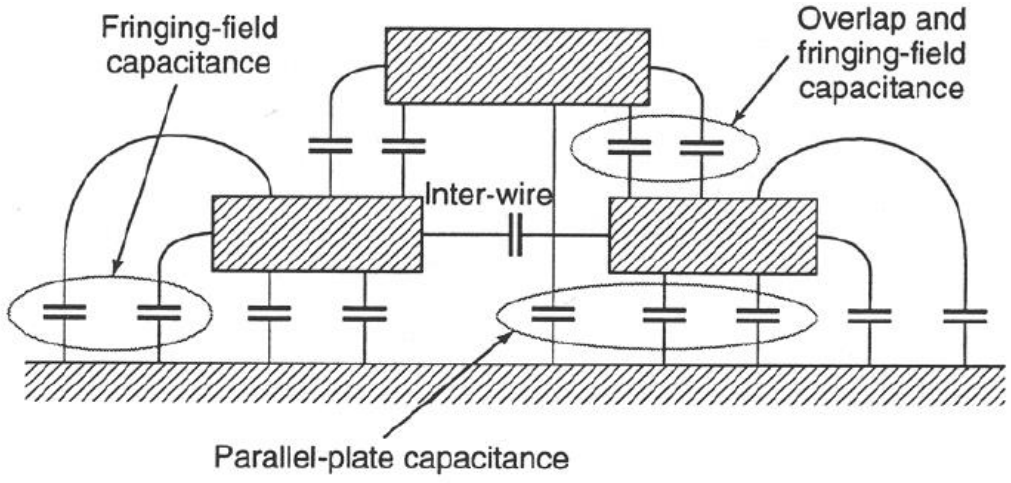
Material	κ
Vacuum	1
Aerogels	1.5
Polyimides (organic)	3-4
Silicon dioxide	3.9
Glass-epoxy (PCB)	5
Silicon Nitride	7.5
Aluminum	9.5
Silicon	11.7

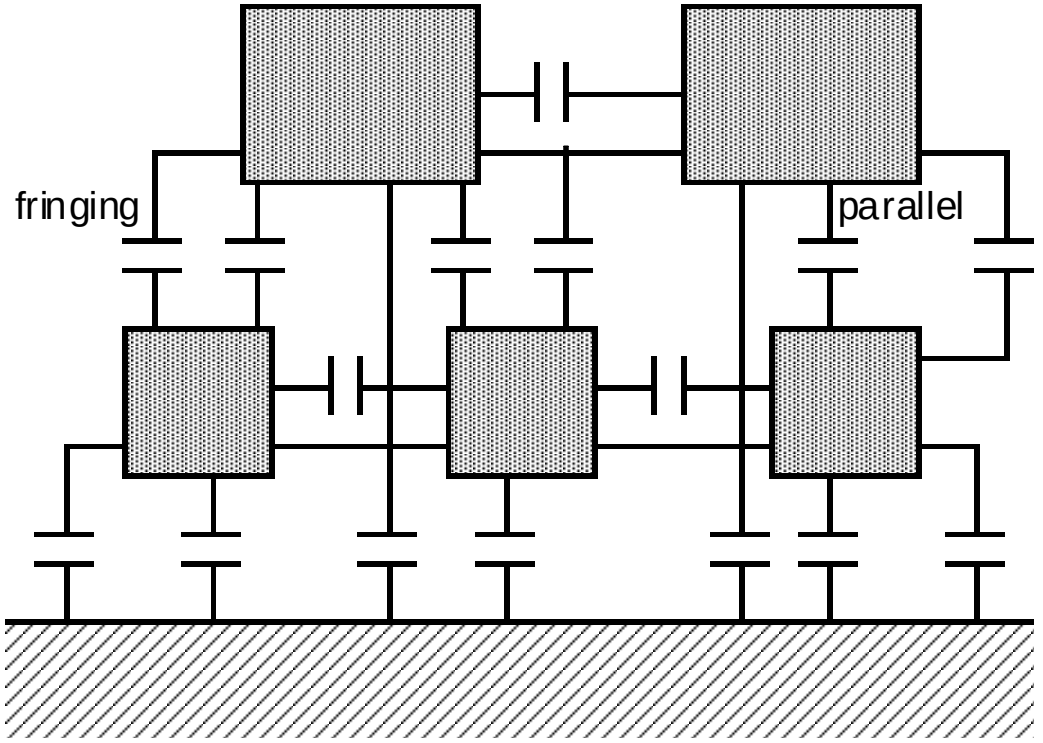
Low-κ dielectrics : less capacitance for a given distance -> good **for interconnect**

High-κ dielectrics : more capacitance for a given distance -> **good for transistors**

Generic Capacitance Model

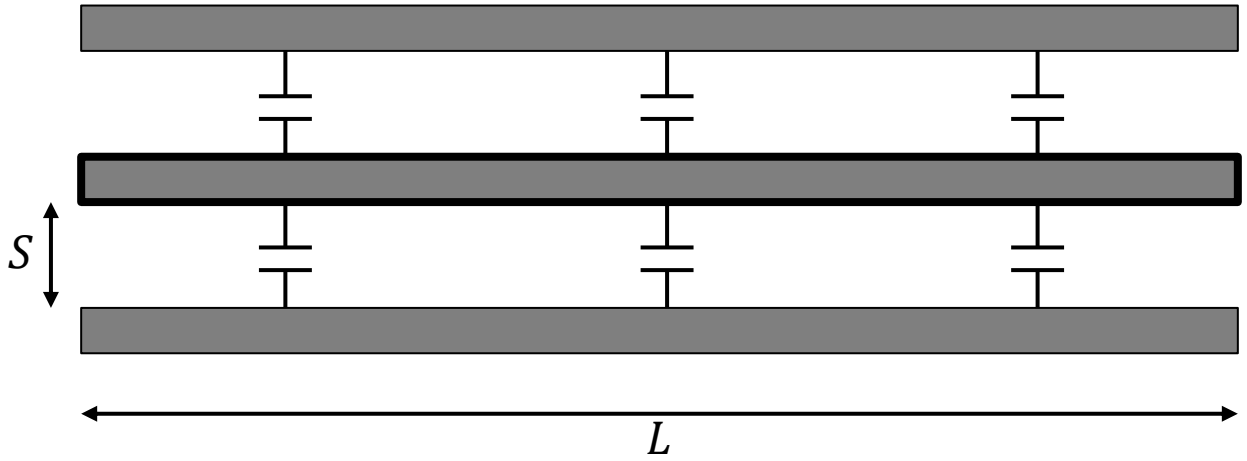
- **Adjacent wires also add parasitic capacitances**
- **Modelling of parasitic capacitance with many layers becomes complex**
  - Field lines end in various points
  - Parallel plate and fringe capacitances are no longer clearly defined
- **Accurate results only feasible with field solvers**





Intra-Layer (Adjacent Wire) Capacitance

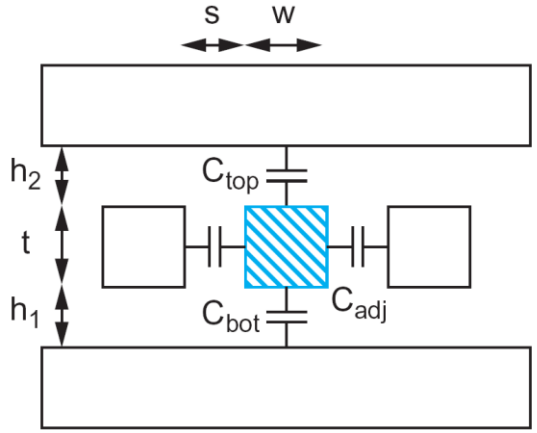
- Nevertheless, two important basic scenarios allow for first-order estimates
  - **Coupling between adjacent wires** can be described with inter-wire capacitance
    - Consider only nearest neighbours (one wire on each side)



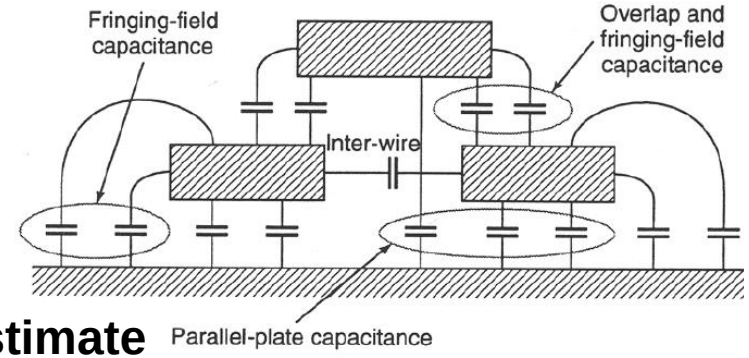
$$C^0_{adj}(S^0) \cdot L \cdot \frac{S^0}{S}$$

$$C^0_{adj}(S^0) \cdot L \cdot \frac{S^0}{S}$$

$$C^0_{adj}(S^0) \sim t$$

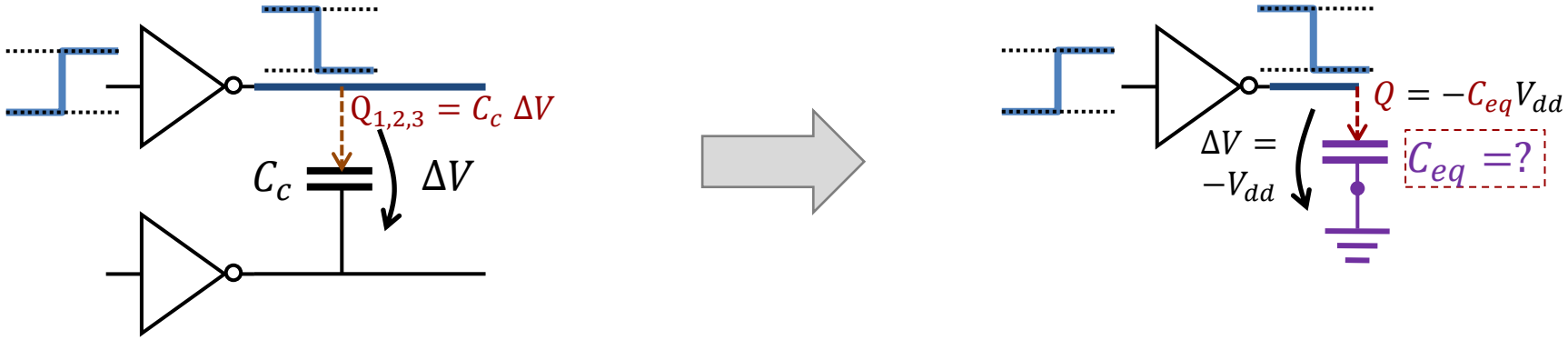


  - **Coupling to** overlapping wires or planes on **adjacent layers**
    - Consider only the nearest occupied layer (layers act as shield for other layers where they overlap)
    - Consider parallel plate for overlap area and fringe capacitance
    - **Coupling between layers is in general more difficult to** accurately **estimate**



Coupling (Inter-Wire) Capacitance and Delay

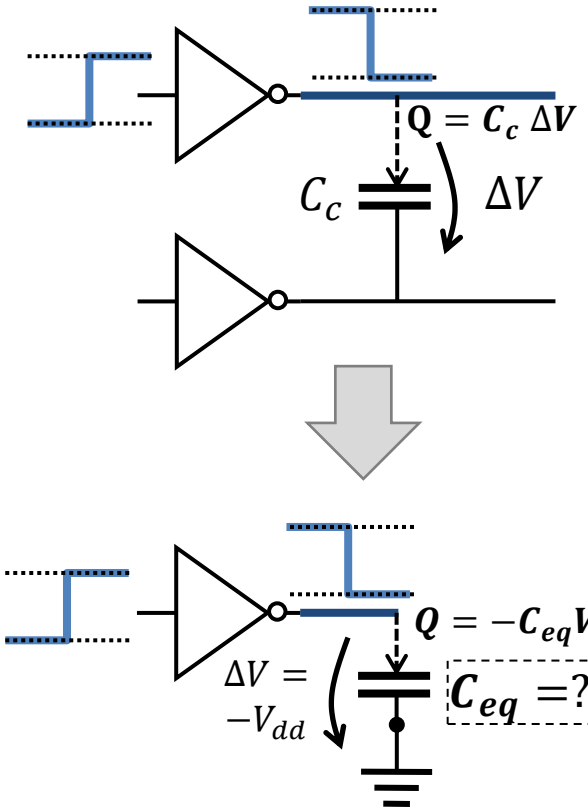
- **Inter-wire coupling is typically between signals (not to GND)**, that are not necessarily at GND and can have transitions as well

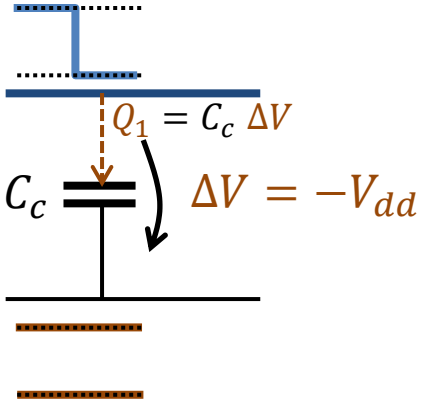


- **Model coupling capacitance as equivalent load to ground**
  - To obtain same impact on delay, **current / charge provided by the driver should be equivalent:** $Q = Q_{1,2,3}$
  - **Select equivalent load C_{eq} accordingly**

Coupling (Inter-Wire) Capacitance and Delay

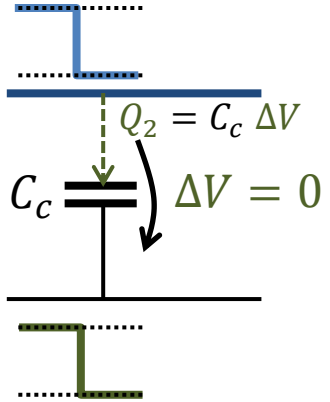
- **Inter-wire coupling is typically between signals (not to GND)**, that are not necessarily at GND and can have transitions as well
  - **Transition on coupled wires impacts the currents during switching**, i.e., effective load





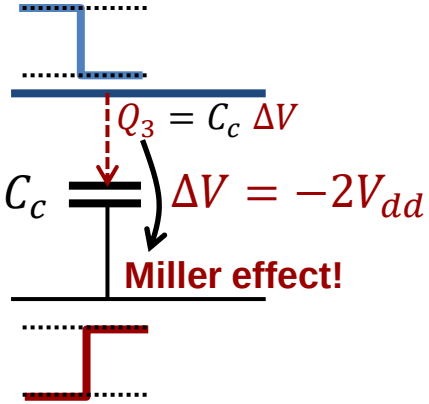
$Q_1 = C_c\,\Delta V$
$= -C_{eq}V_{dd}$

$C_{eq} = C_c$



$Q_2 = C_c\,\Delta V$
$= -C_{eq}V_{dd}$

$C_{eq} = 0$



Best-case

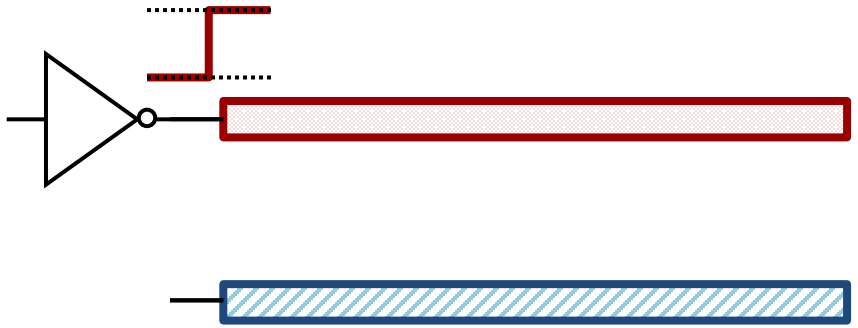
$Q_3 = C_c\,\Delta V$
$= -C_{eq}V_{dd}$

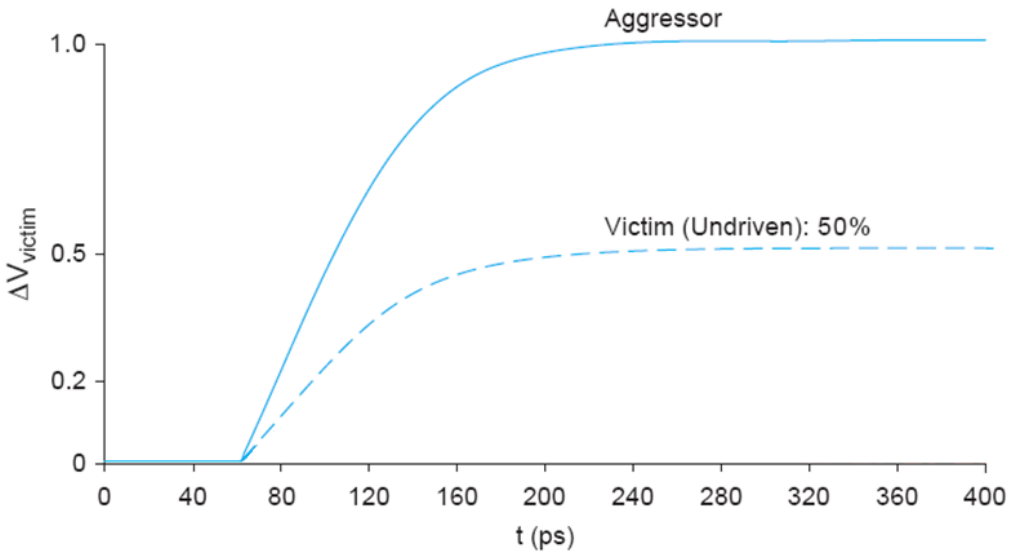
$C_{eq} = 2C_c$

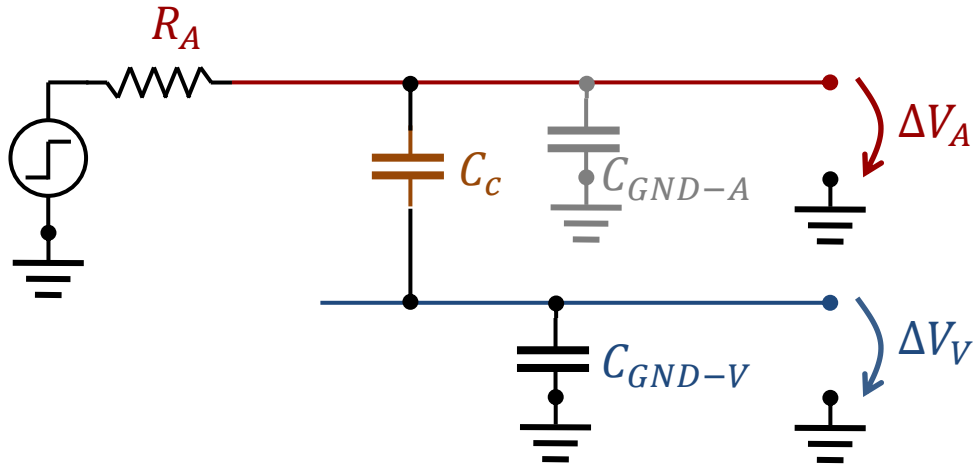
- **Worst-case equivalent load: up to 2x the coupling capacitance**
  - Coupling leads to significant uncertainties

Interaction Between Wires – Crosstalk

- **Coupling** between wires also **affects coupled wires: Crosstalk**
  - Consider **coupling from** an **aggressor A to** an **undriven victim V**







Impact on Floating Victim (worst-case)

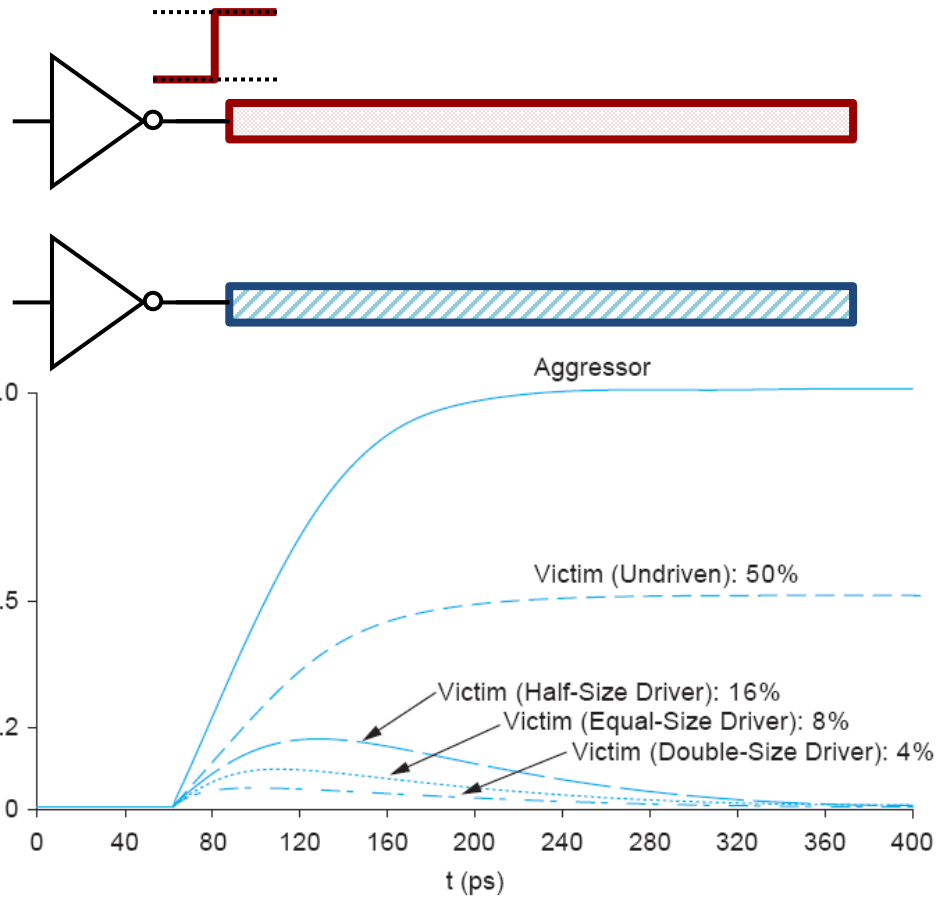
- Permanent impact on victim potential

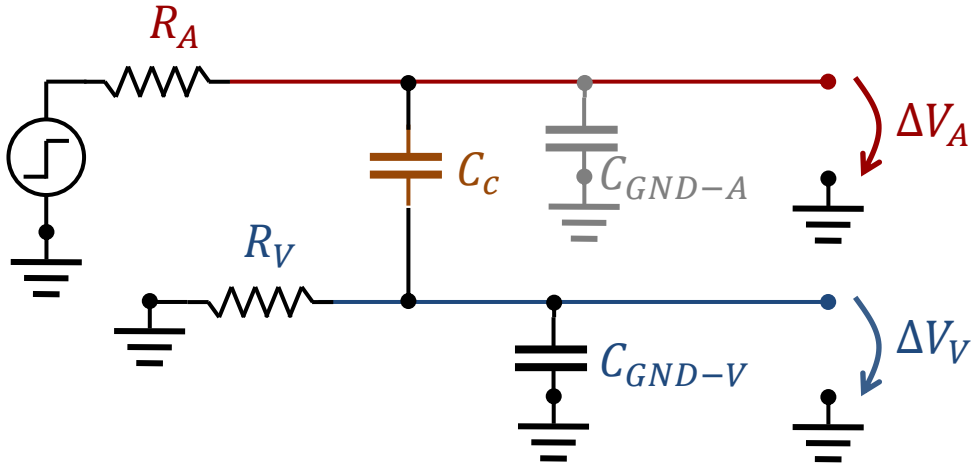
$$\Delta V_V = \frac{C_C}{C_C + C_{GND-V}} \Delta V_A$$

- Capacitive voltage divider between C_C and C_{GND-V}
- Large GND-load reduces impact of coupling

Interaction Between Wires – Crosstalk

- **Coupling** between wires also **affects coupled wires: Crosstalk**
  - Consider **coupling from** an **aggressor A to** a **driven victim V**





Impact on Driven Victim

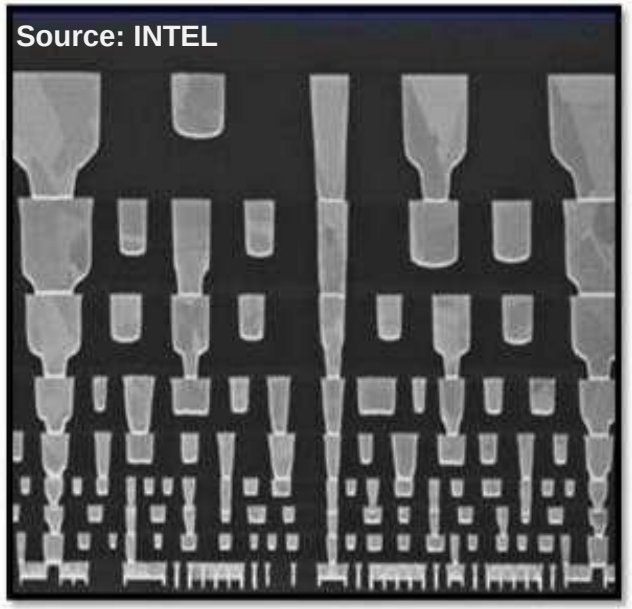
- **Impact** on victim is **not permanent**
- Time constants determine **peak impact**

$$\Delta V_{V-peak} = \frac{C_C}{C_C + C_{GND-V}} \frac{1}{1+k} \Delta V_A$$

$$k = \frac{\tau_A}{\tau_V} = \frac{R_A(C_C + C_{GND-V})}{R_V(C_C + C_{GND-V})}$$

Significance of Different Wire Capacitances

- **Often few coupling capacitances dominate, which simplifies calculations**

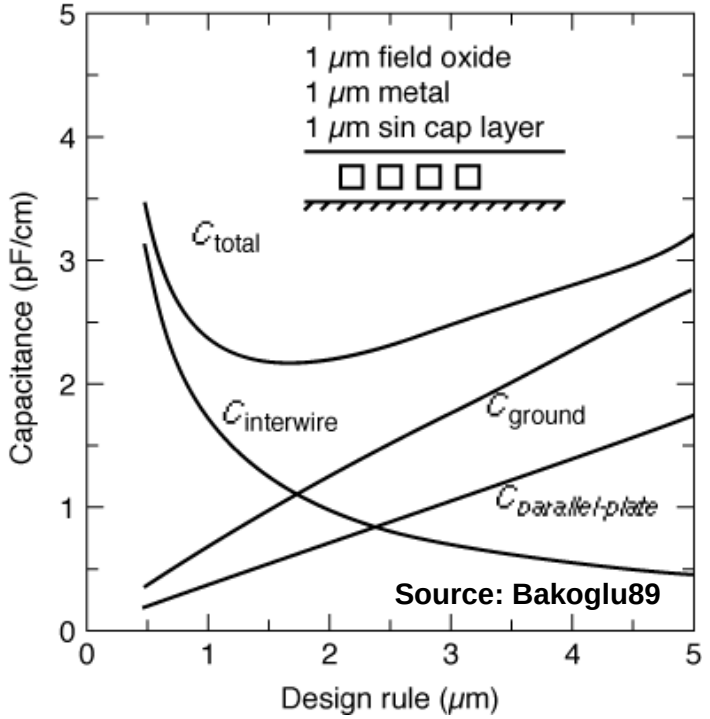


0.25um process

	Field	Active	Poly	Al1	Al2	Al3	Al4
Poly	88						
	54				pp in aF/µm²		
Al1	30	41	57		fringe in aF/µm		
	40	47	54				
Al2	13	15	17	36			
	25	27	29	41			
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

	Poly	Al1	Al2	Al3	Al4	Al5
Interwire Cap	40	95	85	85	85	115

per unit wire length in aF/µm for minimally-spaced wires



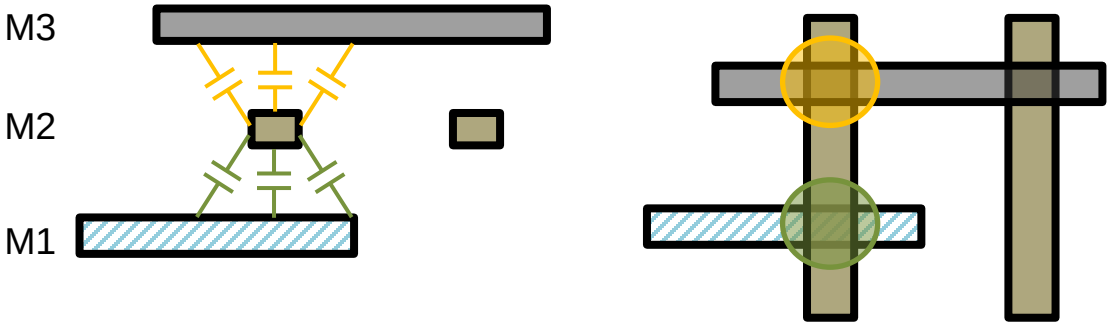
- Some observations
  - **Fringe capacitances are at least as important as parallel plate capacitance**
  - **Capacitance between layers drops rapidly for non-adjacent layers** (e.g., Al1-Al2 vs Al1-Al3)
  - **Interwire capacitance often dominates** especially for upper metal layers
  - More **dense routing** improves capacitance to GND and adjacent layers, but **worsens coupling**

Layout Techniques for Controlling Parasitics

- Some **basic layout measures** help to **reduce and better predict parasitics**

Alternating Directions

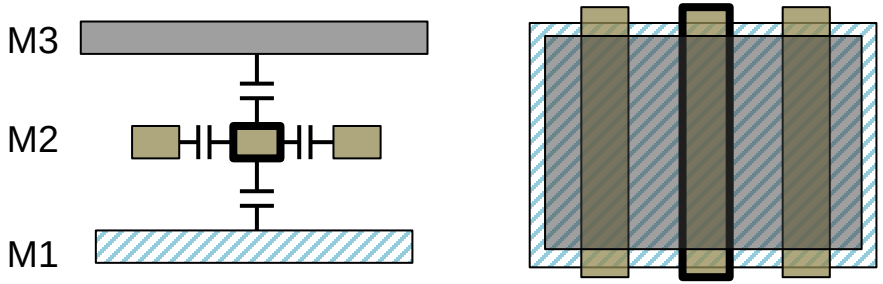
- Preferred routing direction changes from one layer to the next



- Minimizes overlap and coupling between wires on adjacent layers
- Well defined geometries for overlap regions simplifies inter-layer capacitance calculation

Shielding

- Surround critical wires completely with grounded parallel wires and planes



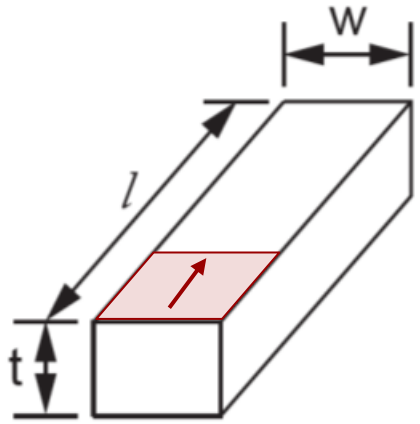
- Mostly restores parallel plate capacitance model to adjacent layers (without fringe cap)
- Coupling to parallel shielding wires replaces the fringe capacitance

Wire Resistance

- **For longer wires, series wire resistance becomes relevant**

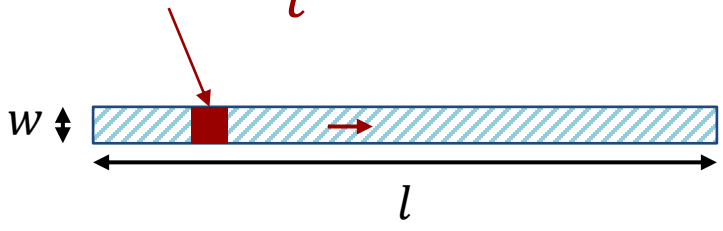
- **Resistance of a conductor:** described by
  - its geometry: width, length, and thickness
  - The specific resistance of the material ρ

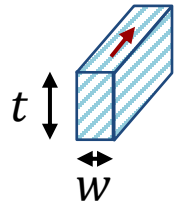
$$R = \frac{\rho}{t}\frac{l}{w}$$

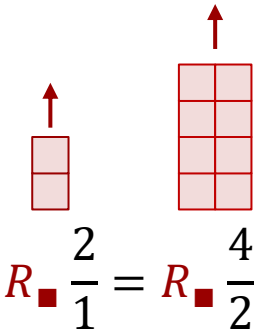


- The **designer** usually **has no control over the thickness** (for chosen layer)
  - Define the **sheet resistance as** the **resistance of a square** (equal length and width)

PDK specifies $R_{\blacksquare} = \frac{\rho}{t}$ ➔ $R = R_{\blacksquare}\frac{l}{w}$







  - Only width/length ration of a wire matters

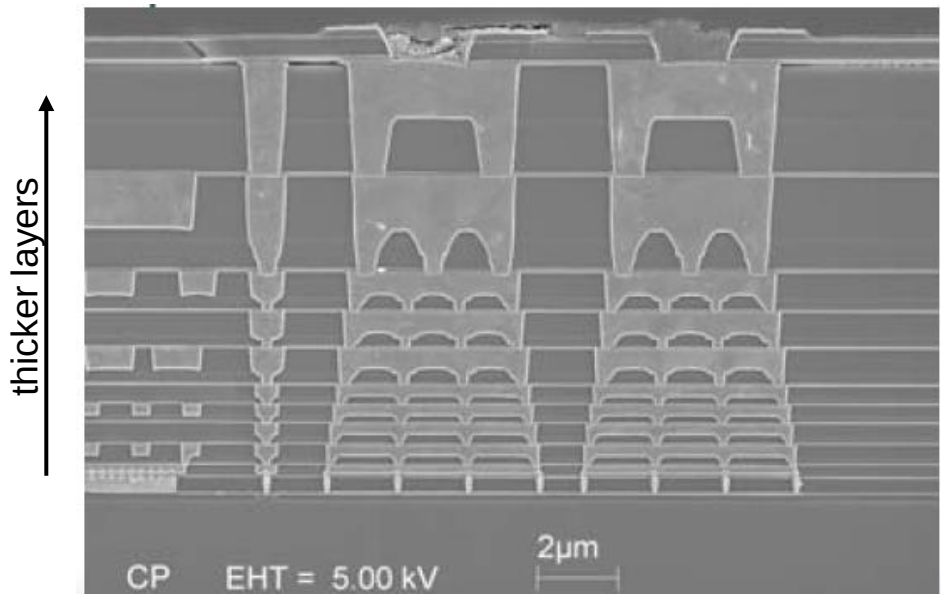
Impact of Layer and Technology on $R_{\blacksquare}$

- Two parameters influence sheet resistance: **material and layer thickness**

180nm process

Layer	Material	Sheer Res. (Ω/■)
Active	N, P well diffusion	1'000 – 1'500
	N+, P+ diffusion … with silicide	50 – 200 3 – 10
Poly	Polysilicon … with silicide	50 – 400 3 – 10
M1	Aluminum	0.08
M2	Aluminum	0.05
M6	Aluminum	0.02

thicker layers

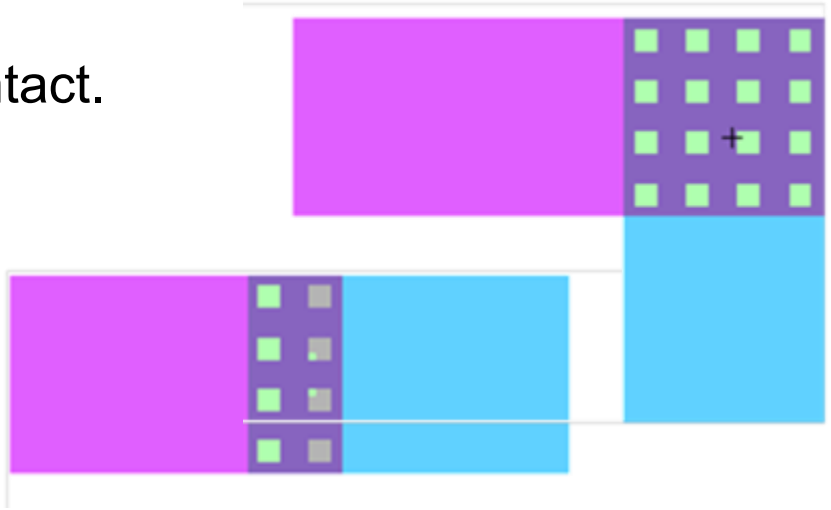


- **Both parameters depend on technology** and the **layer**
  - Active layers have high sheet resistance compared to dedicated routing layers
  - Upper layers have lower resistance due to increase in layer thickness
- **Careful layer selection is important** to optimize resistance

Contact Resistance is Important

- **Contact/Vias** also **add resistance**
  - Contact resistance is generally **2-20 ohms**
  - **Impact depends on** the **length of the** connected **wire**
    - Via resistance (2 Ohms) is similar to a wire that is 20-200 times longer than wide

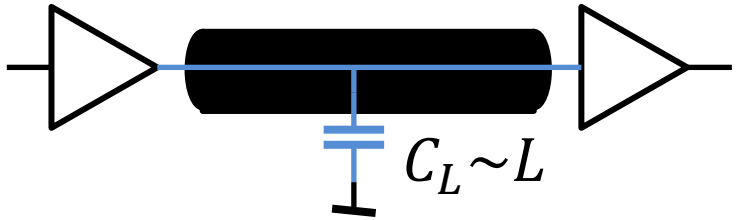
- **Reducing contact resistance:** enlarge vias
  - BUT… current “crowds” around the perimeter of a contact.
  - There are also problems in deposition…
  - Contacts/Vias have a maximum practical size.
  - **Use arrays of multiple small vias**



Delay Models

- **Lumped capacitor model:** ignores wire resistance

- **Wire presents only an additional load to the driving gate**



$t_{pd} \sim L$

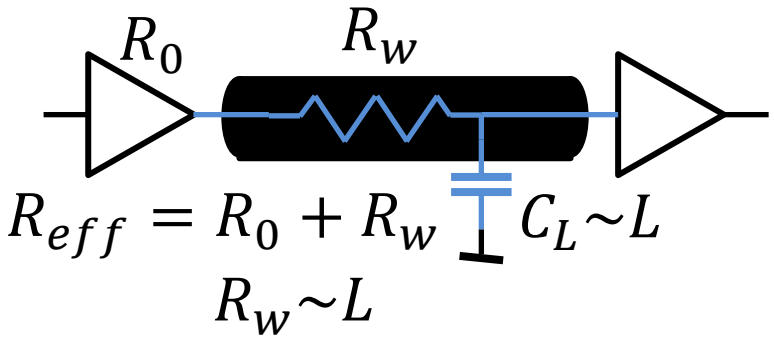
  - **Delay and slope, increase proportional to the wire length**

- **Use if wire-resistance is low compared to on-resistance of the driver**
  - OK **for short wires and** for **old technologies**
  - Too optimistic for new technologies and long wires

Delay Models

- **Lumped RC model:** considers also the wire resistance

- **Capacitance and resistance lumped into two elements (pessimistic!)**
  - Still ignore the distributed nature of the resistance and capacitance
  - Wire resistance R_w in series with on-resistance of the driver R_0 forms effective resistance R_{eff}



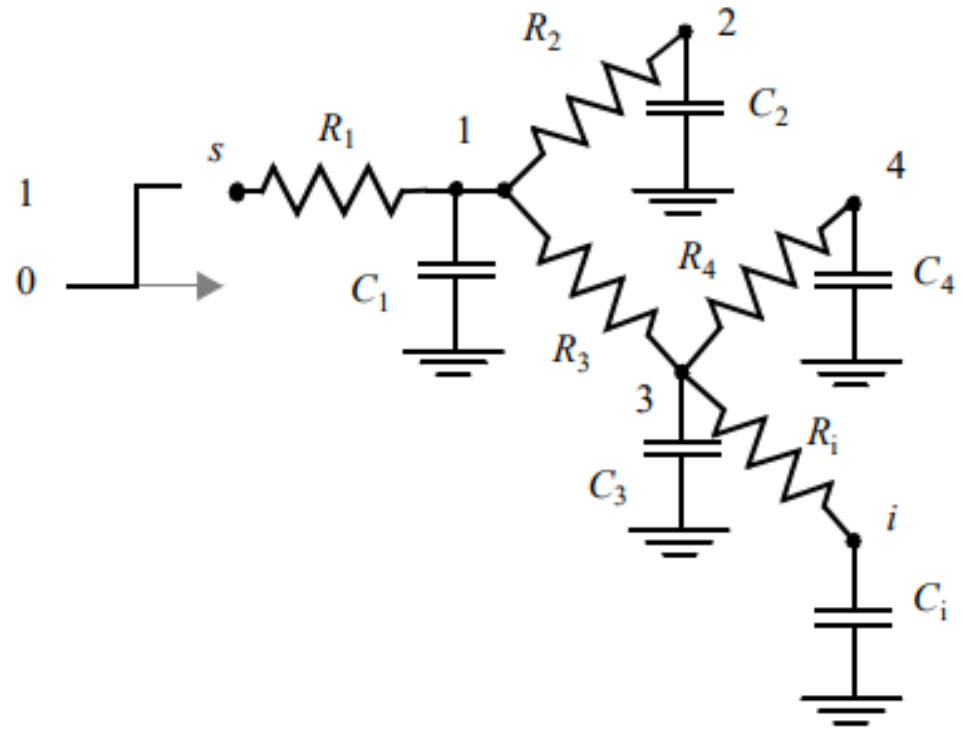
$$t_{pd} \sim L^2$$

  - **Delay and slope, increase quadratic to the wire length** once $R_w > R_0$

- **Use if wire-resistance is comparable to on-resistance of the driver AND a pessimistic estimate is required**

Elmore Delay Model

- **Consider a general RC network**
  - Accurate closed form solution for delay is not available



Shared path resistance: R_{ik} consider only the resistance that the path from root (s) to node k and the path from root to node of interest (i) have in common

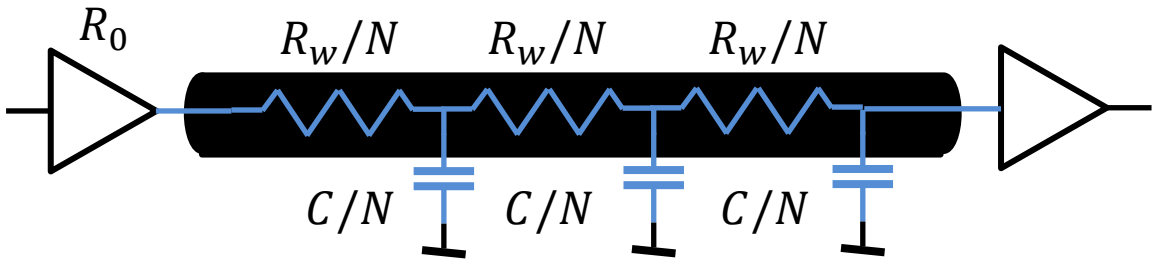
$$R_{ik} = \sum R_j \Rightarrow (R_j \in [path(s \to i) \cap path(s \to k)])$$

$$t_{pd} \sim 0.69 \sum_k R_{ik} C_k$$

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

Delay Models (Dominated by PP Capacitance)

- **Distributed RC model: special case of Elmor**
  - Approximate delay as sequence of RC pairs



$$t_{pd} = 0.69\left[\left(\frac{R_w}{N}\frac{C}{N} + \frac{2R_w}{N}\frac{C}{N} + \frac{3R_w}{N}\frac{C}{N} + \cdots + \frac{NR_w}{N}\frac{C}{N}\right) + R_0C\right]$$

$$= 0.69\left[\frac{R_wC}{N^2}(1+2+3+\cdots+N) + R_0C\right] = 0.69C\left[R_w\frac{N(N+1)}{2N^2} + R_0\right] \approx 0.69C\left(\frac{R_w}{2} + R_0\right)$$

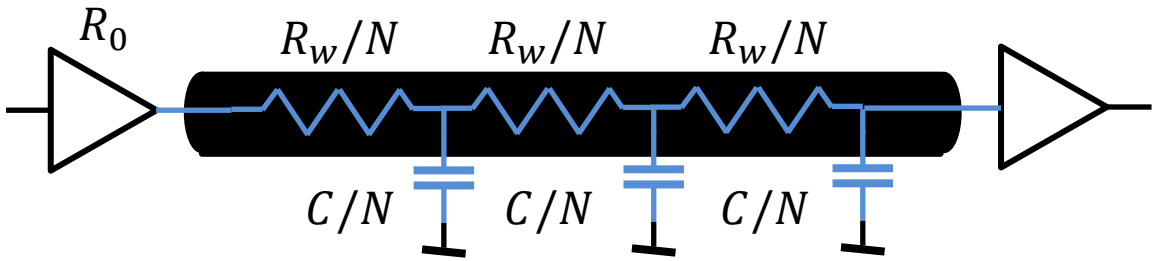
  - With $R \sim R_{\blacksquare}\frac{L}{W}$ and $C \sim C_0^{PP}LW$:

$$t_{pd}^{wire} = \frac{0.69}{2}R_{\blacksquare}C_0^{PP}L^2$$

Delay increases quadratically with wire length independent of *W*

Delay Models (Dominated by Fringe Capacitance)

- **Distributed RC model: special case of Elmor**
  - Approximate delay as sequence of RC pairs



$$t_{pd} = 0.69\left[\left(\frac{R_w}{N}\frac{C}{N} + \frac{2R_w}{N}\frac{C}{N} + \frac{3R_w}{N}\frac{C}{N} + \cdots + \frac{NR_w}{N}\frac{C}{N}\right) + R_0C\right]$$

$$= 0.69\left[\frac{R_wC}{N^2}(1+2+3+\cdots+N) + R_0C\right] = 0.69C\left[R_w\frac{N(N+1)}{2N^2} + R_0\right] \approx 0.69C\left(\frac{R_w}{2} + R_0\right)$$

  - With $R \sim R_{\blacksquare}\frac{L}{W}$ and $C \sim C_0^{fringe}L$:

Dominated by fringe capacitance

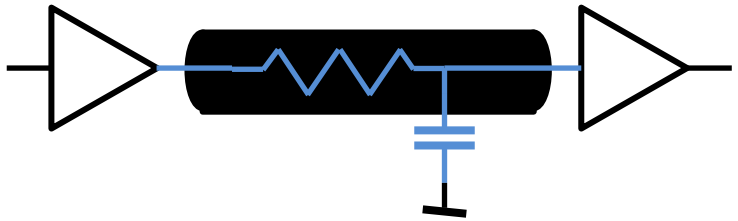
$$t_{pd}^{wire} = \frac{0.69}{2}R_{\blacksquare}C_0^{fringe}\frac{L^2}{W}$$

Delay increases quadratically with wire length **and** inverse proportional to wire width

Lumped vs. Distributed RC Model

- **Comparing the lumped RC model to the distributed RC model we find that**

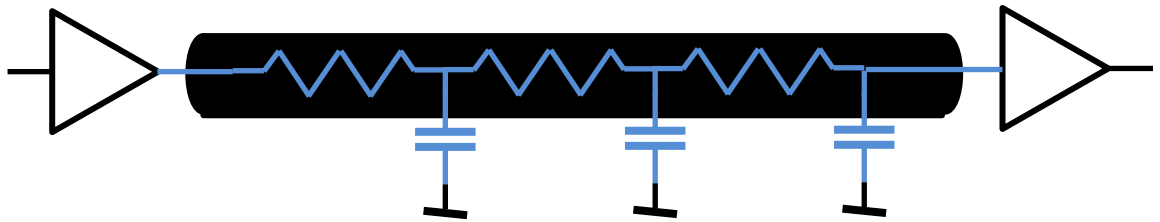
Lumped



$$t_{pd}^{wire} = 0.69\, R_{\blacksquare} C_0^{fringe} \frac{L^2}{W}$$

$$t_{pd}^{wire} = 0.69\, R_{\blacksquare} C_0^{PP} L^2$$

Distributed



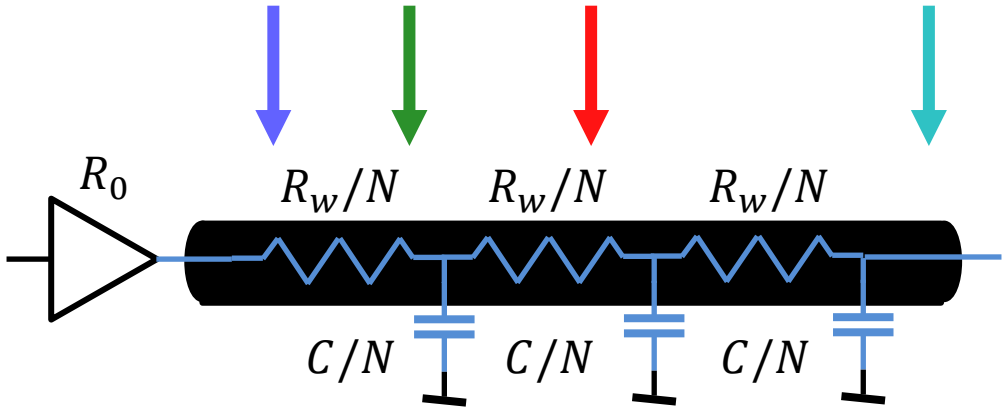
$$t_{pd}^{wire} = \frac{0.69}{2} R_{\blacksquare} C_0^{fringe} \frac{L^2}{W}$$

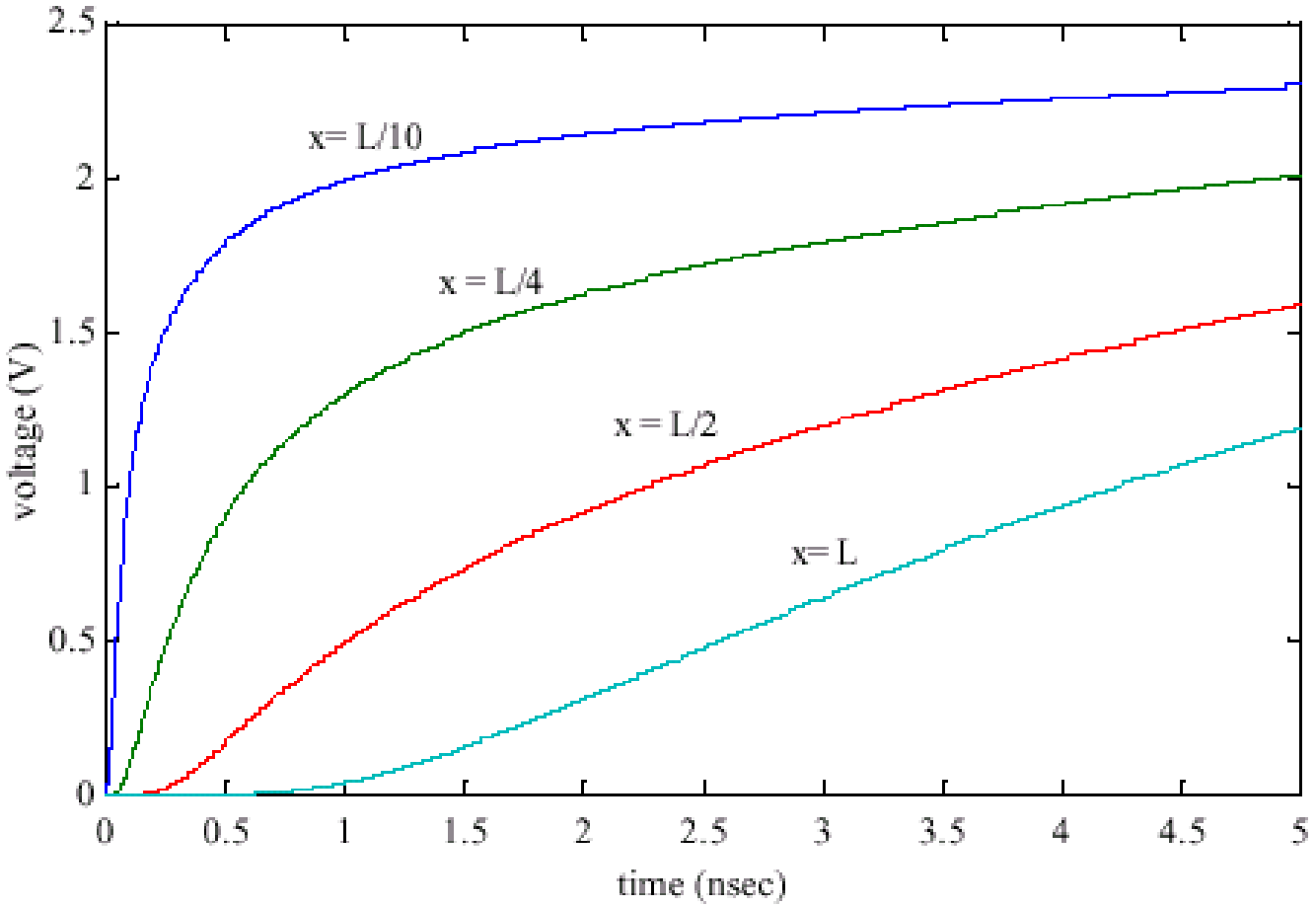
$$t_{pd}^{wire} = \frac{0.69}{2} R_{\blacksquare} C_0^{PP} L^2$$

- Both show the same scaling behaviour (for PP and for fringe dominated delay)
- The **lumped RC model is pessimistic**: delay is 2x that of the more accurate distributed model

Propagating Wavefront Along a Wire

- **Step response of a distributed RC wire as a function of location along wire and time**



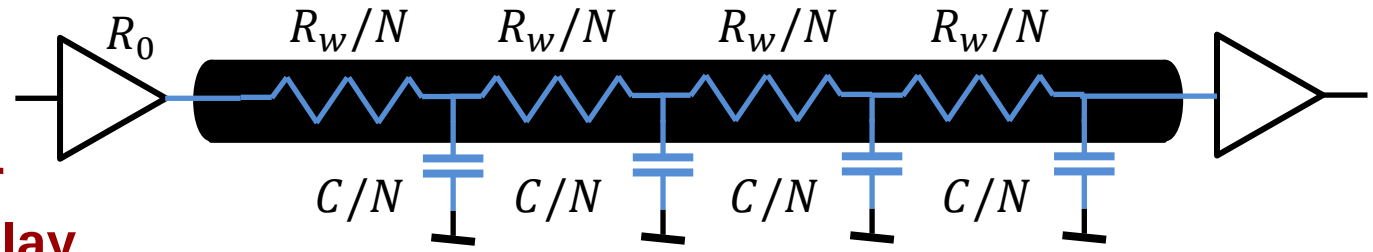


Buffering/Repeaters to Reduce Wire RC Delay

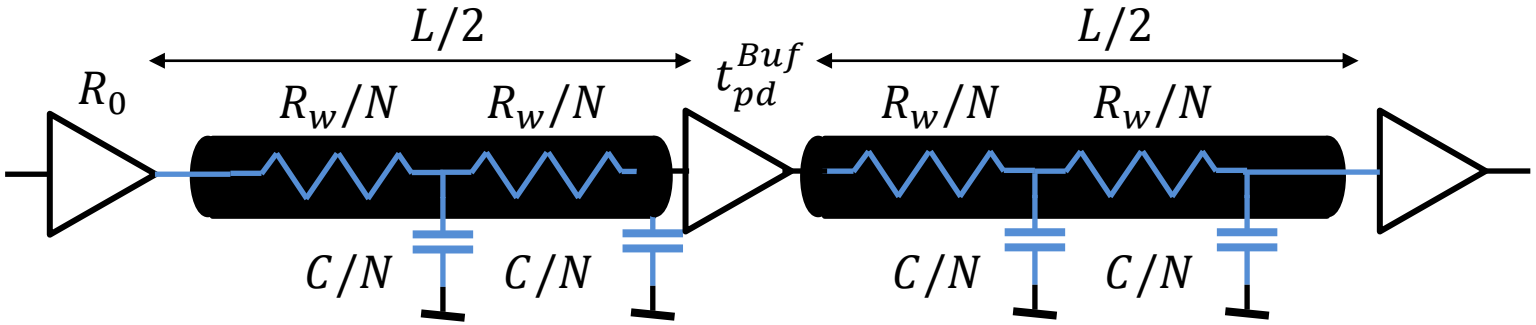
- **Consider a long wire $R_w > R_0$:**

$$t_{pd} = 0.69 R_0 C_w + 0.35 R_w C_w$$

  - **Increasing the strength of the driver does not significantly reduce the delay**



- **Splitting the wire into two with a buffer in between:**
  - Adds the delay of a buffer, but also cuts wire length in half



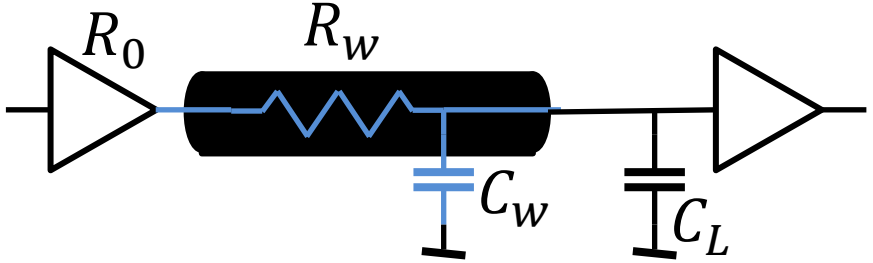
$$t_{pd} \sim 0.69 R_{\blacksquare} C_0 2 \frac{(L/2)^2}{W} + t_{pd}^{Buf} = 0.69 R_{\blacksquare} C_0 \frac{L^2}{2W} + t_{pd}^{Buf}$$

Relevance of Routing Parasitics

- **Routing parasitics must be considered when they have a notable impact**
  - Load is caused by fanout and routing parasitics
  - Resistance is caused by the driver and routing parasitics

- **Lumped-C model is more easy to handle than a distributed RC model**

- **Which parasetics are when relevant?**
  - Check the contributions to C_w: parallel plate, fringe capacitance, coupling capacitance
  - Check the time constants of the delay components



$$\tau_1 = R_0 C_L \qquad \tau_3 = R_w C_L$$

$$\tau_2 = R_0 C_w \qquad \tau_4 = \frac{1}{2} R_w C_w$$

Layout Guidelines

- **Poly:**
  - High resistance, but can directly contact to a gate
  - Use for very short local interconnect to multiple gates
- **Metal 1:**
  - Only layer that can connect directly to diffusion and Poly
  - Routing within a standard cell, often VDD/GND
- **Metal 2 to N-1:**
  - General global routing
- **Metal N:**
  - VDD/GND
  - Global clock distribution

EE-429
Fundamentals of VLSI Design

Tecnology Scaling (Wires)

Andreas Burg

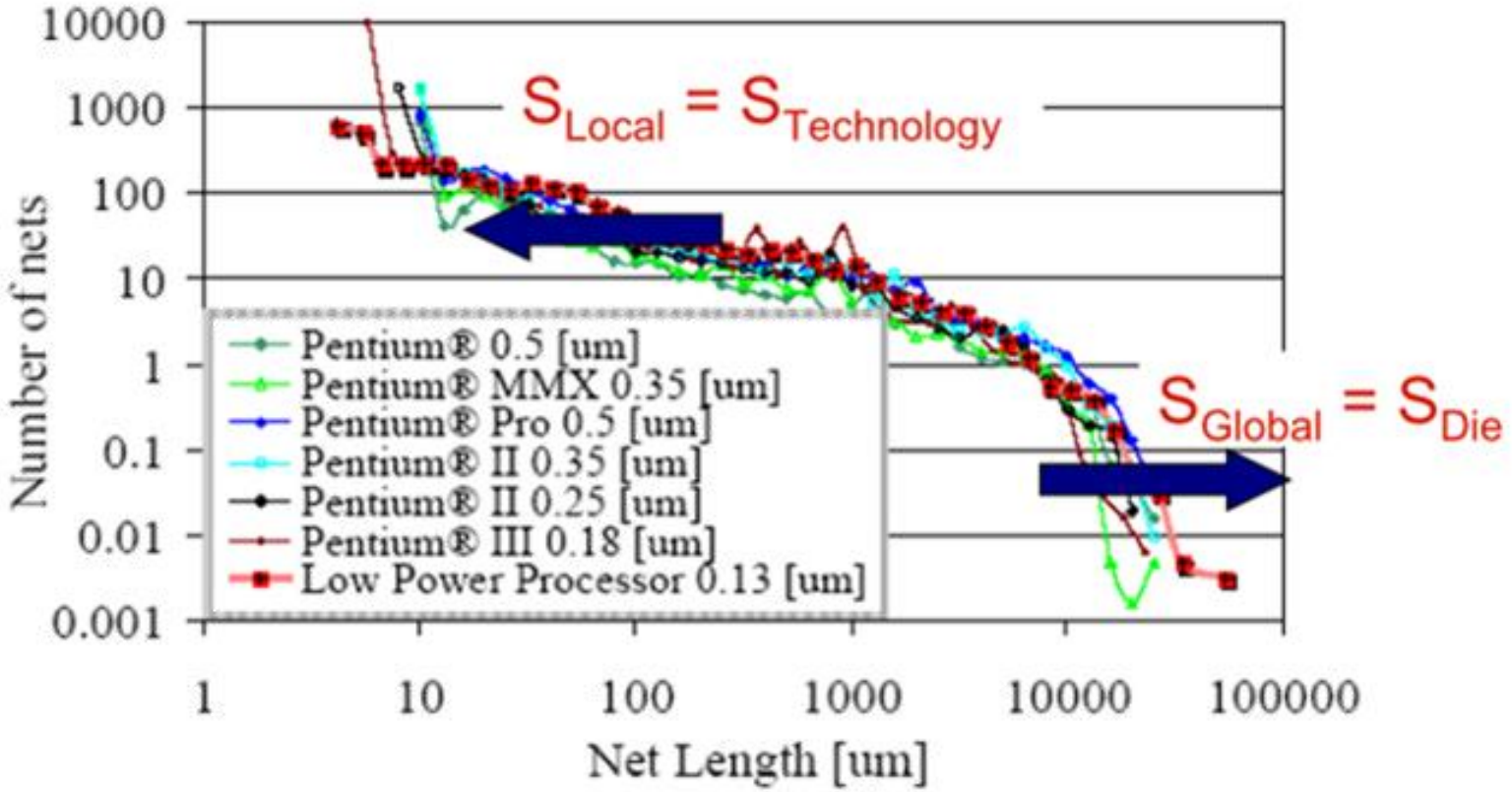
Wire Scaling

- **We could try to scale interconnect at the same rate (*S*) as device dimensions.**
  - This makes sense for *local interconnect* that connects 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

- **With device dimensions shrinking, we also expect shorter connections**
- **However, chip-size has remained constant or has even increased**
- **Wire-length distribution has remained constant across the full range**
  - Some extra short wires with some small increase in local interconnect



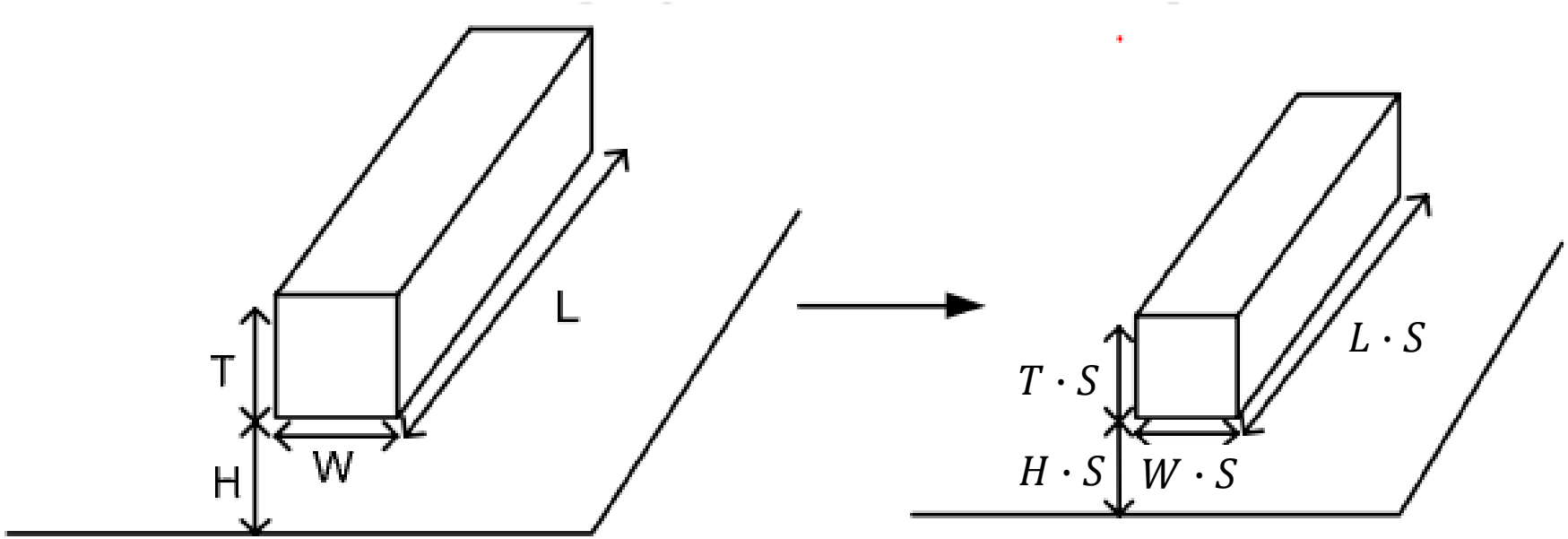
From Magen et al., "Interconnect Power Dissipation in a Microprocessor"

Need to check scaling behaviour for both local and global wires

Scaling of Local Interconnect ($L \propto S$)

- **Scaling of all wire dimensions**

$W \propto S$ $H \propto S$ $T \propto S$



$$C_{pp} \propto W \cdot L/H$$
$$C_f \propto L$$
$$R_w \propto L/(W \cdot T)$$

$$C_{pp} \propto S \cdot S/S \propto S$$
$$C_f \propto S \propto S$$
$$R_w \propto S/(S^2) \propto 1/S$$

$$\tau_w \propto S \cdot (1/S) \propto 1$$

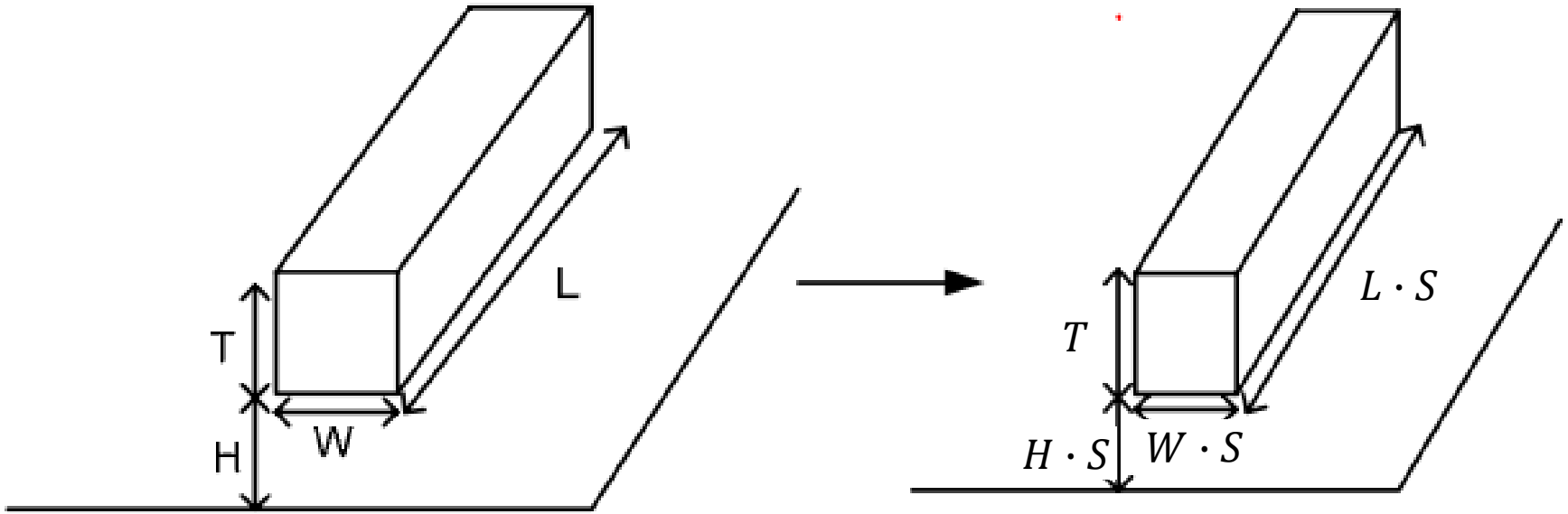
Compare to gate delay scaling

$$\tau_g \propto S - S^2$$

Scaling of Local Interconnect ($L \propto S$)

- **Scaling of wire dimensions with CONSTANT thickness**

$W \propto S$ $H \propto S$ $T \propto 1$



$$C_{pp} \propto W \cdot L/H$$
$$C_f \propto L$$
$$R_w \propto L/(W \cdot T)$$

$$C_{pp} \propto S \cdot S/S \propto S$$
$$C_f \propto S \propto S$$
$$R_w \propto S/(S \cdot 1) \propto 1$$

Compare to gate delay scaling

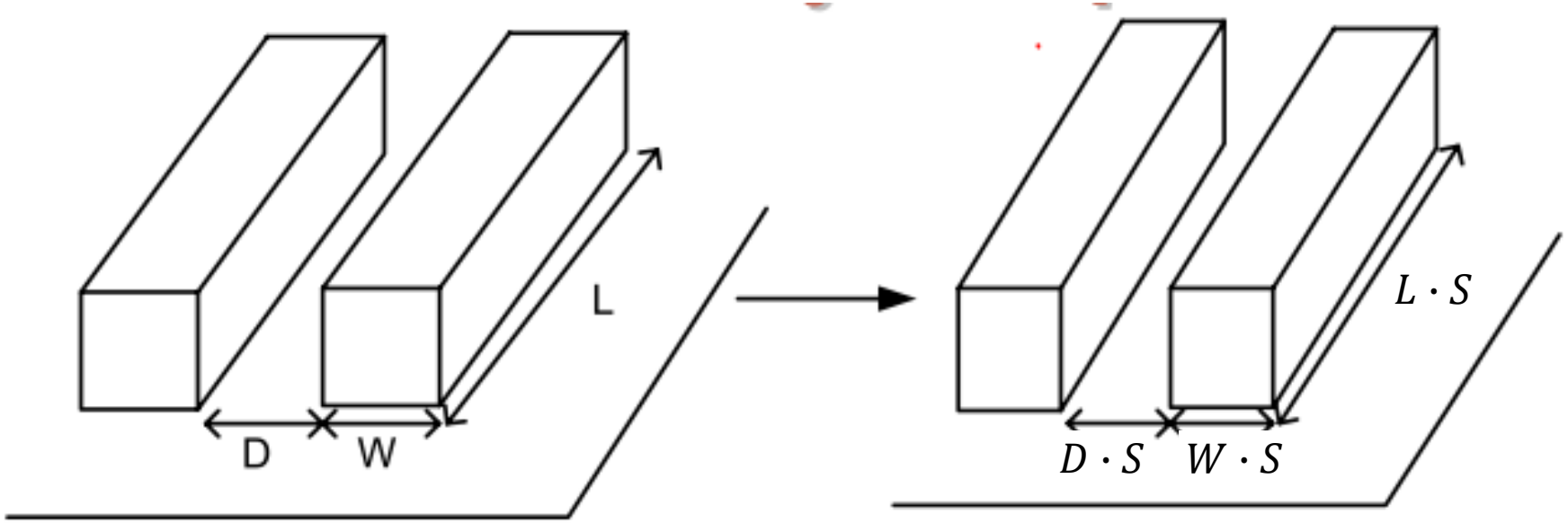
$$\tau_g \propto S - S^2$$

$$\tau_w \propto S \cdot 1 \propto S$$

Scaling of Local Interconnect ($L \propto S$) with Coupling

- **Scaling of wire dimensions with CONSTANT thickness and Coupling**

$W \propto S$ $H \propto S$ $T \propto 1$ $D \propto S$



$$C_C \propto T \cdot L/D$$
$$R_w \propto L/(W \cdot T)$$

$$C_C \propto 1 \cdot S/S \propto 1$$
$$R_w \propto S/(S \cdot 1) \propto 1$$

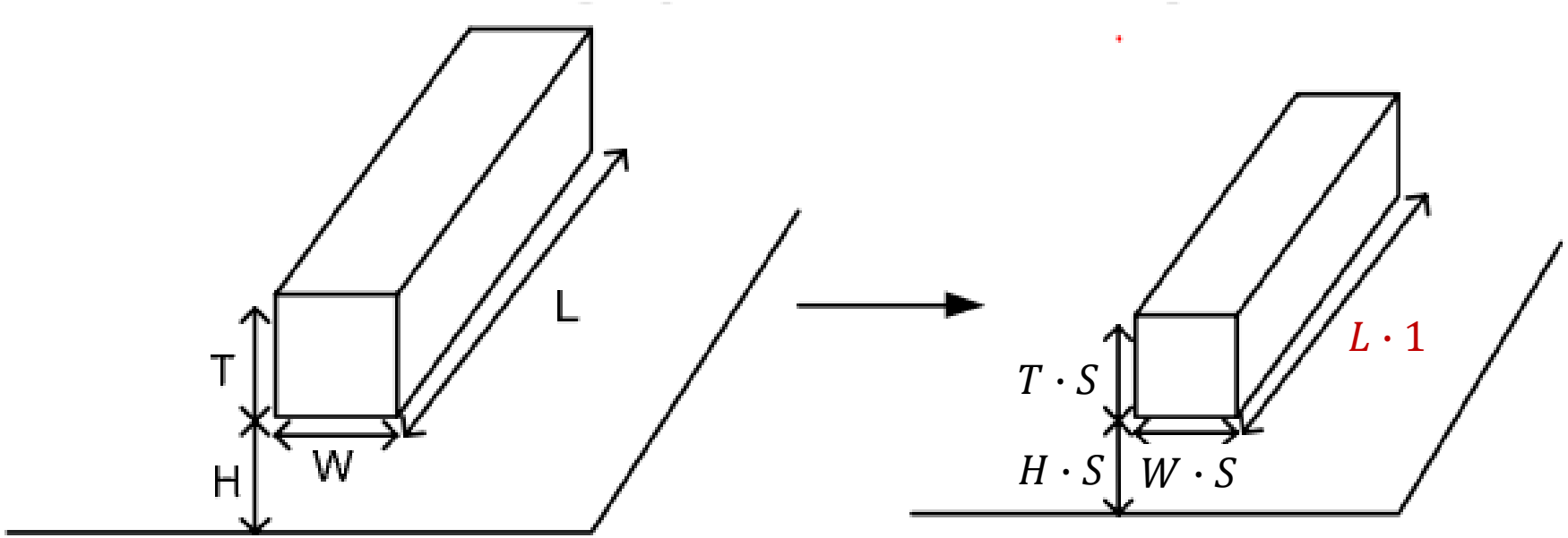
Approaching wires eliminate local wire length scaling advantage

$$\tau_w \propto 1 \cdot 1 \propto 1$$

Scaling of Global Interconnect ($L \propto 1$)

- **Scaling of all wire dimensions**

$W \propto S$ $H \propto S$ $T \propto S$



$$C_{pp} \propto W \cdot L/H$$
$$C_f \propto L$$
$$R_w \propto L/(W \cdot T)$$

$$C_{pp} \propto S \cdot 1/S \propto 1$$
$$C_f \propto 1 \propto 1$$
$$R_w \propto 1/(S \cdot S) \propto 1/S^2$$

Compare to gate delay scaling

$\tau_g \propto S - S^2$

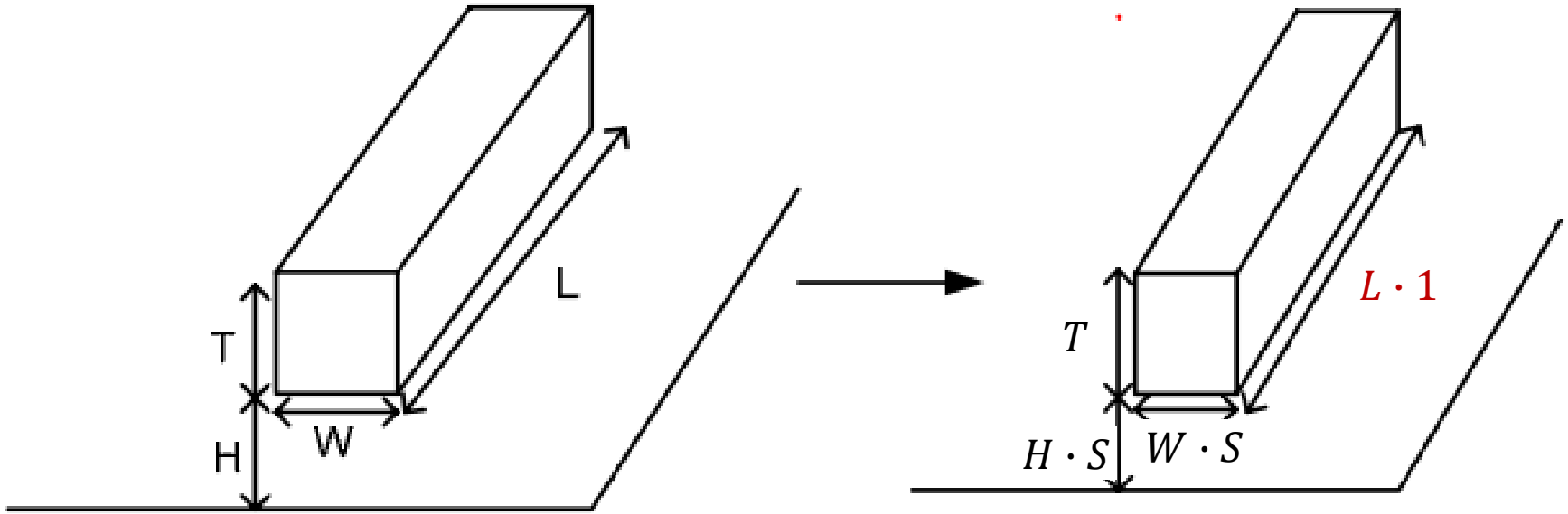
$\tau_w \propto 1/S^2$

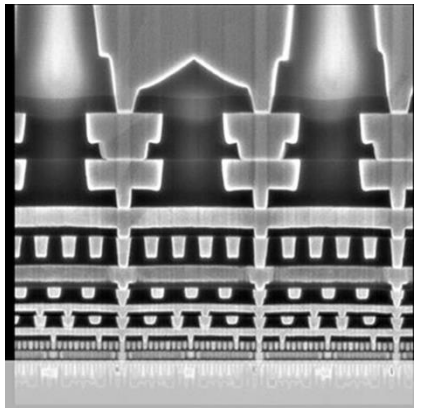
UNACCEPTABLE

Scaling of Global Interconnect ($L \propto 1$)

- **Global wires will use thick (upper metal) wires: constant thickness**

$W \propto S$ $H \propto S$ $T \propto 1$





Compare to gate delay scaling

$\tau_g \propto S - S^2$

$$C_{pp} \propto W \cdot L/H$$
$$C_f \propto L$$
$$R_w \propto L/(W \cdot T)$$

$$C_{pp} \propto S \cdot 1/S \propto 1$$
$$C_f \propto 1 \propto 1$$
$$R_w \propto 1/(S \cdot 1) \propto 1/S$$

$\tau_w \propto 1/S$

Slightly better, but still…

Wire Scaling

- **Whereas device speed increases with scaling:**
  - Local interconnect speed stays constant.
  - Global interconnect delays increase quadratically or at least linear.
  - Keep the wire thickness (*H*) fixed helps (provides *S* for local wires and *1/S* for global wires).
  - However, coupling increases, which demands for more careful routing

- **Interconnect delay is often the limiting factor for speed.**

- **Several measures can be applied to alleviate the issue:**
  - **Low resistance metals**
  - **Low-K insulation**
  - **Low metals (M1, M2) are used for local interconnect, so they are thin and dense**
  - **Higher metals used for global routing, so they are thicker, wider and spaced farther apart**