VLSI Interconnects
Spring 2017

Gian Gerosa
Jerry Moench
Jacob Abraham
Mark McDermott

Outline

- INTRODUCTION
- DUAL DAMASCENE METAL
- QUICK REFRESHER on WIRE GEOMETRY
- WHY DO WE CARE ABOUT INTERCONNECT SO MUCH?
- SCALING
- CALCULATING RC DELAY
- REPEATERS
- NETWORKS ON CHIP
- CONCLUSIONS

INTRODUCTION

- VLSI chips are just not transistors; they need to be connected; they also need to be supplied with power and ground
- Contemporary designs have 8 to 11 layers of metals; alternating layers run orthogonally
- Wires are as important as transistors
  - They affect speed
  - They burn power
  - They can be subject to NOISE leading to functional failure
  - They can wear out too.
- Wires scale as well, but not very nicely (in DSM CMOS).

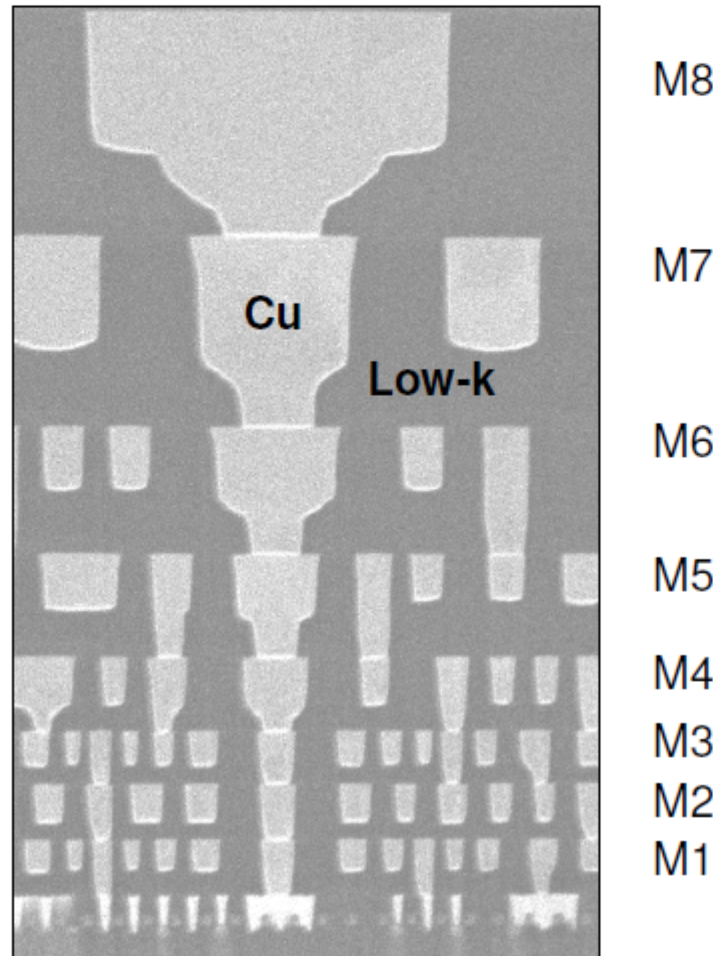
45nm Interconnect

Loose pitch + thick metal on upper layers:

- **High speed global wires**
- **Low resistance power grid**

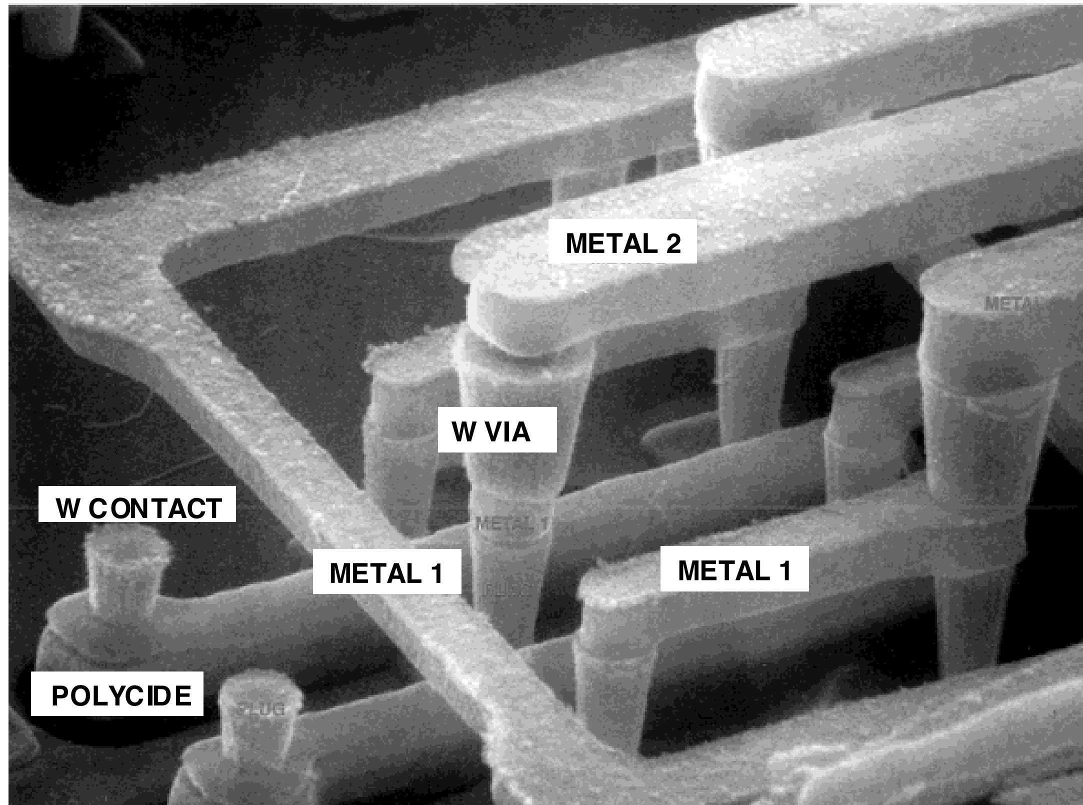
Tight pitch on lower layers:

- **Maximum density for local interconnects**



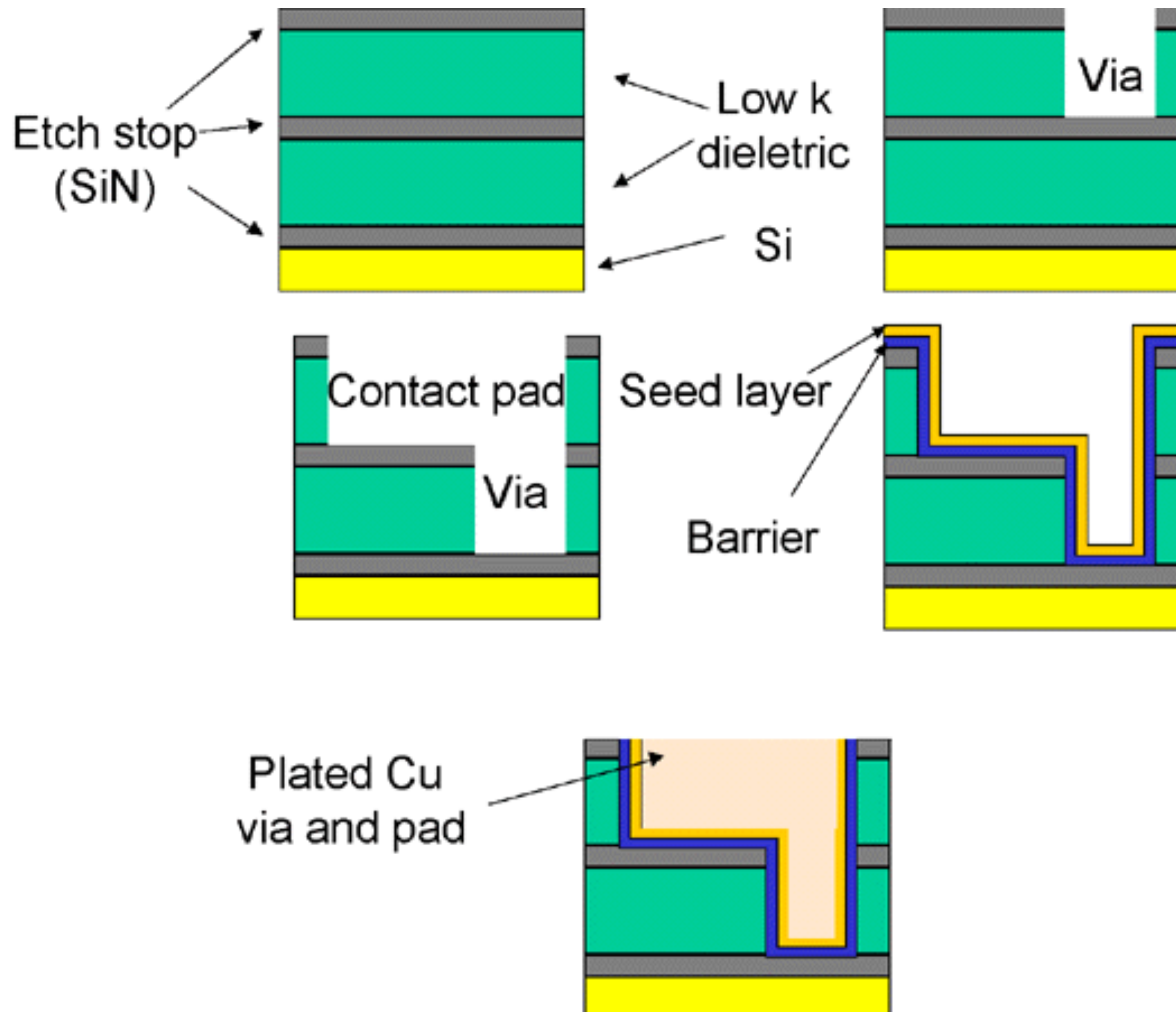
Source: Mark Bohr, Intel Corporation

SEM MICRO-GRAPH (ILM DIELECTRIC REMOVED)



Courtesy: IBM

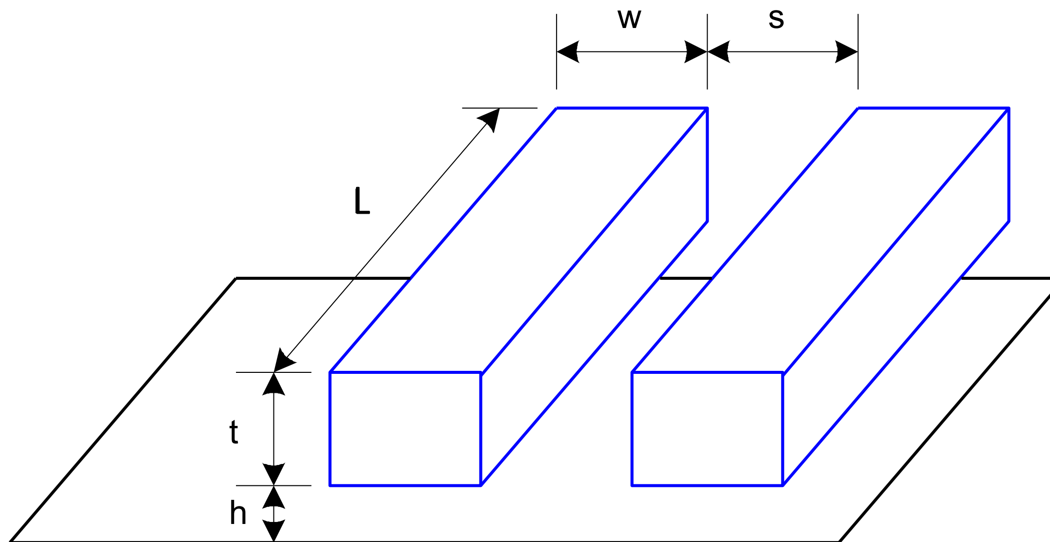
Interconnect Process Dual Damascene



C.-K. Hu and J.M.E. Harper, *Mater. Chem. Phys.*, 52 (1998), p. 5.

WIRE GEOMETRY

- PITCH = width + space
- Height (h) = distance to top/bottom routes
- ASPECT RATIO (AR) = thickness / width
  - Deep submicron processes have AR < 2 to maintain sheet resistances at a reasonable level
  - Coupling to neighboring routes dominates



$$R=\rho L/tw$$

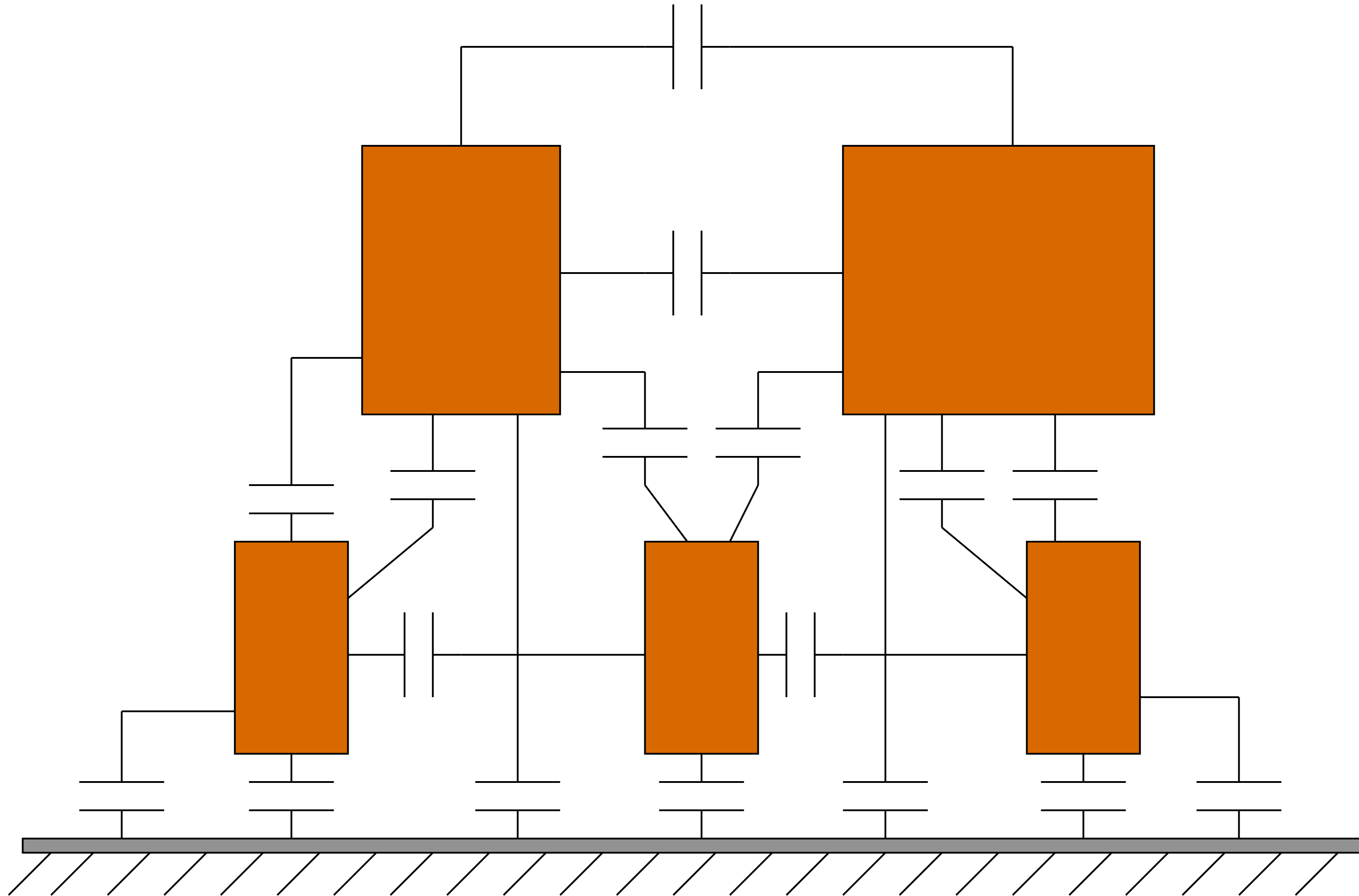
ρ=resistivity

METALS

- **Before 250nm, all interconnect used Aluminum**
- **Modern processes use Copper**
  - **Copper diffusion liner required to protect underlying transistors**

Metal	Bulk Resistivity (μohm-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

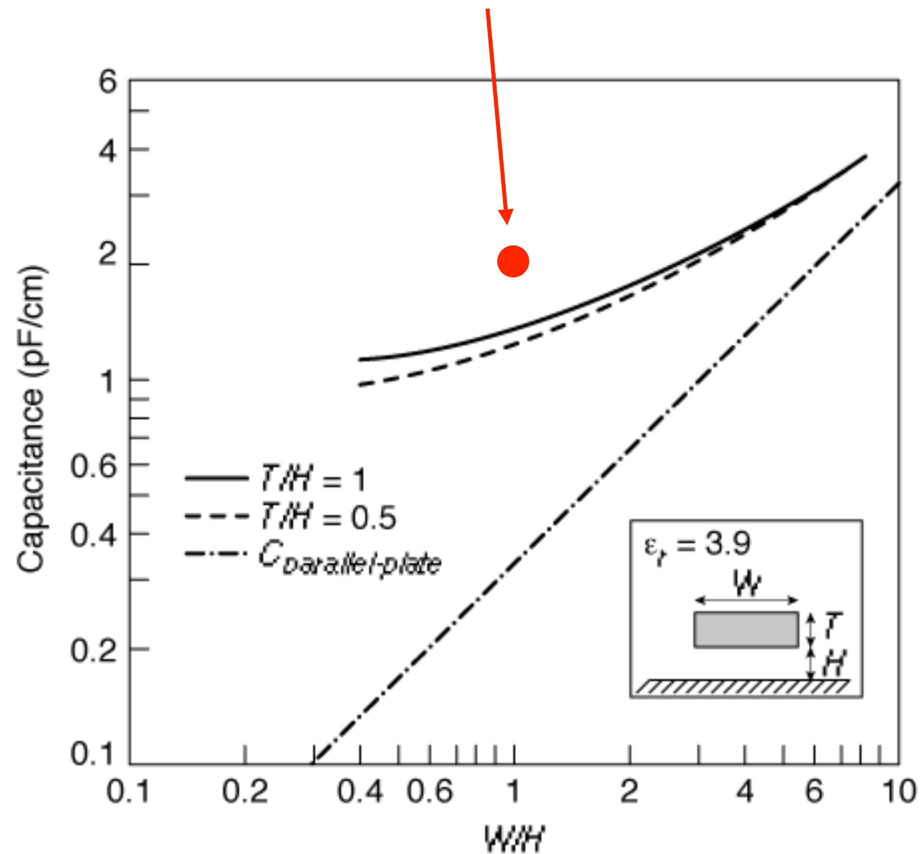
INTERCONNECT CAPACITANCE



FRINGING .vs. PARALLEL PLATE CAPACITANCE

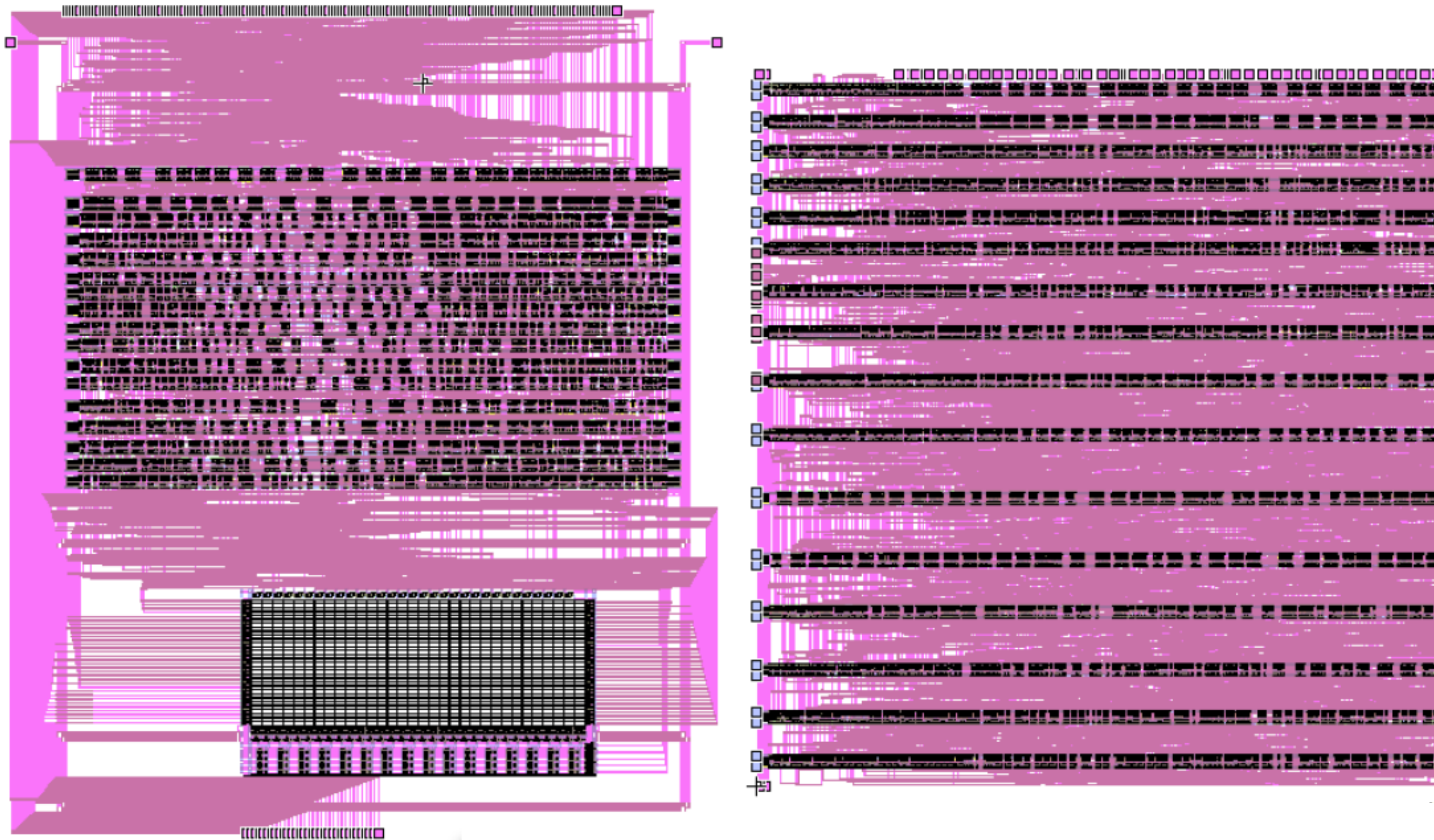
UT ECE

For T/H~2 and W/H~1, capacitance/distance ~ 2pF/cm or ~0.2 fF/micron.



[from http://infopad.eecs.berkeley.edu/~icdesign/. Copyright 1996 UCB and Prentice hall 1995]

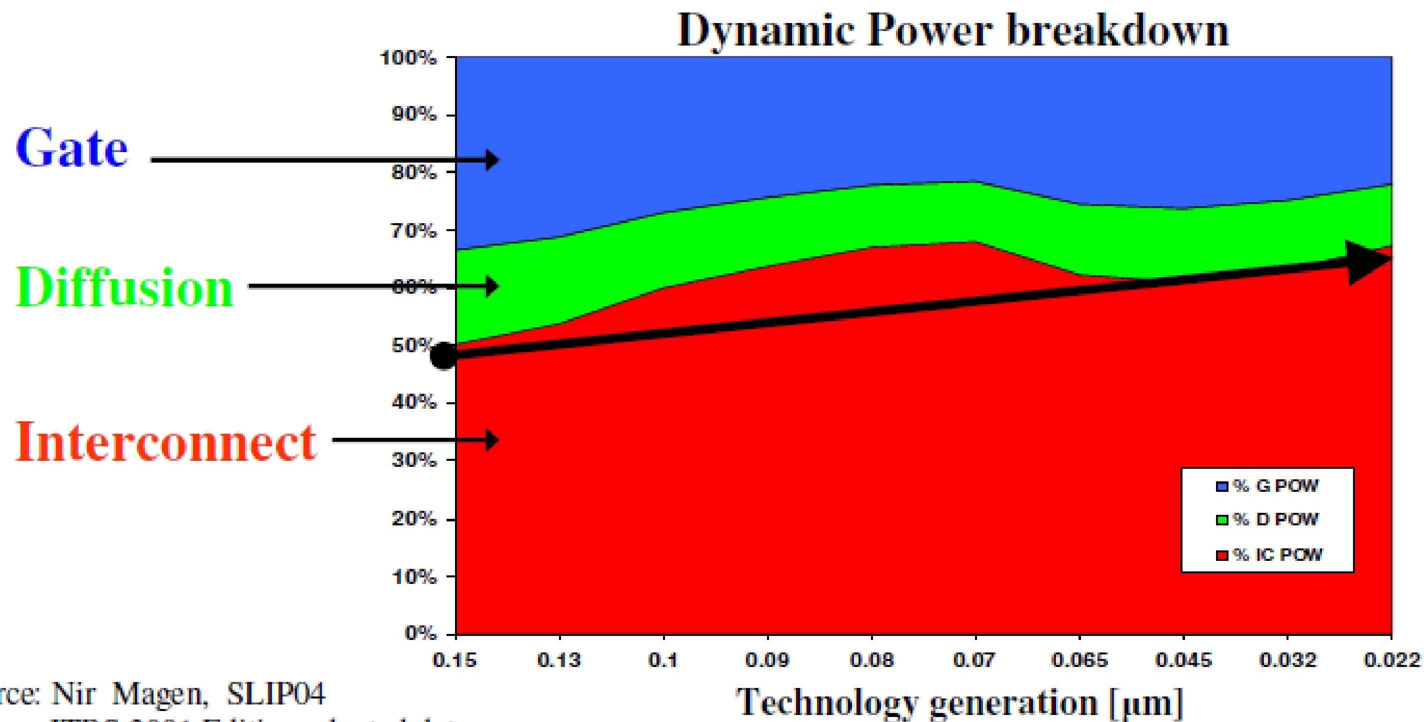
Impact of Having a 2nd Metal



CONTRIBUTION OF WIRES (they are not free)

- DELAY
  - Function of R, L and C, VIA resistance, local .vs. global
- POWER
  - Charging/discharging of C is a function of CV^2F
- NOISE
  - Attackers/victims impact on functionality and delay
- POWER SUPPLY IR DROPS and GROUND BOUNCE
  - Affects delay leading to timing failures
- RELIABILITY
  - Electro-migration, self heat, maximum current
- COST
  - Number of layers .vs. area/performance targets, yield

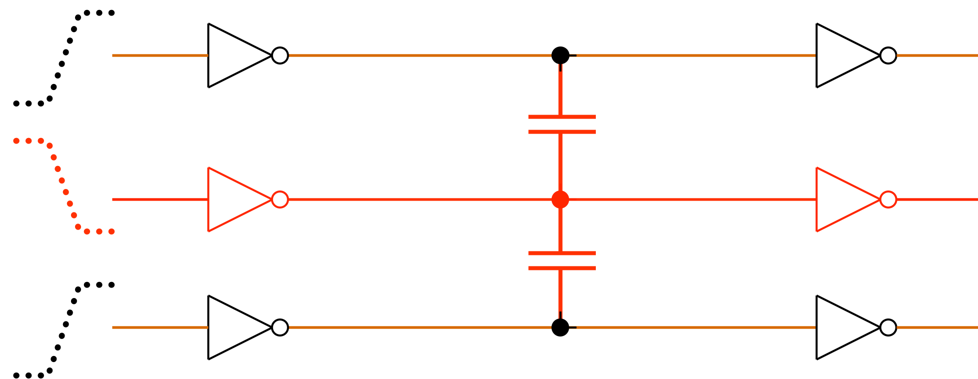
The Future of Interconnect Power



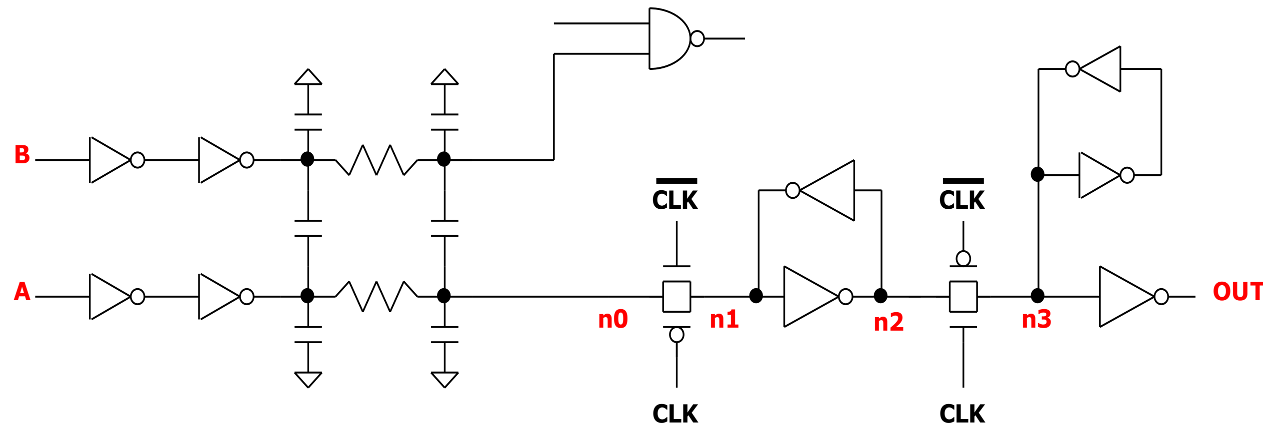
Interconnect power grows to 65%-80% within 5 years

INTERCONNECT NOISE

- *If* **neighboring signals switch in opposite directions:**
  - Both capacitors switch in opposite direction
  - Effective voltage is doubled leading to more charge required
  - Influences victim's delay leading to timing uncertainty
  - Victim's slope (or slew rate) is also affected

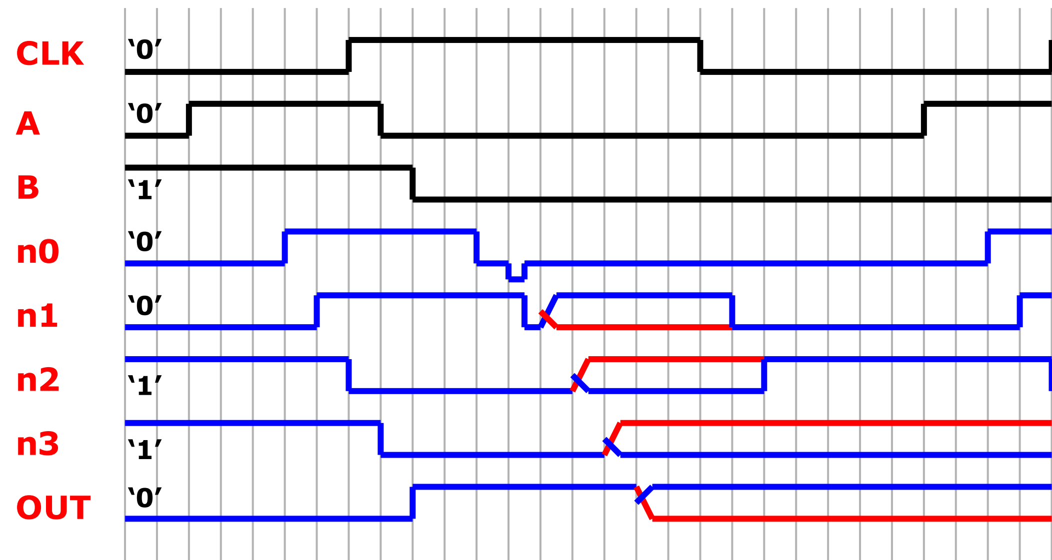


FUNCTIONAL FAILURE DUE to NOISE COUPLING



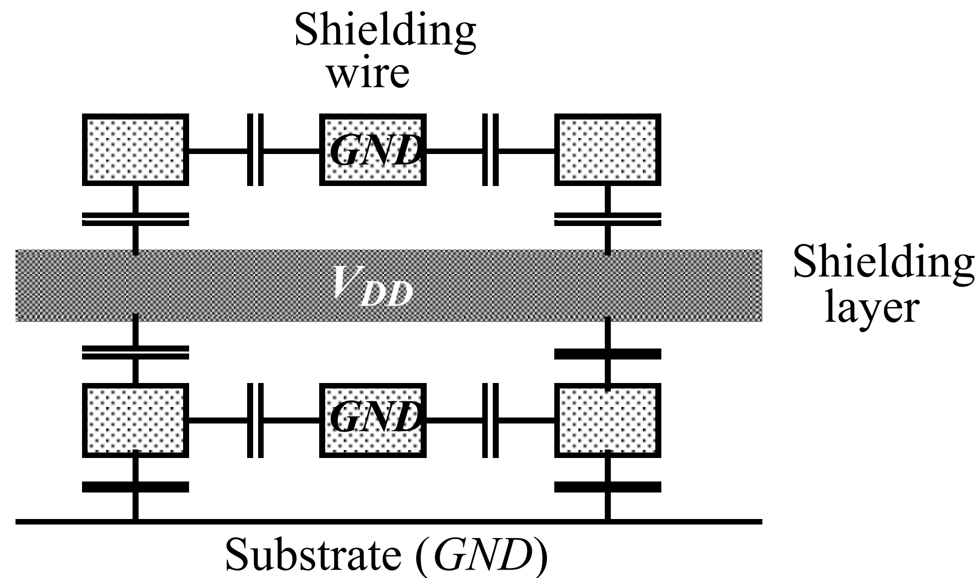
n0 can be coupled below ground (by a high-to-low transition on B), turning on the pass gate, discharging the state-node n1; if the master latch cannot recuperate, then a functional failure will occur.

How would you fix this?



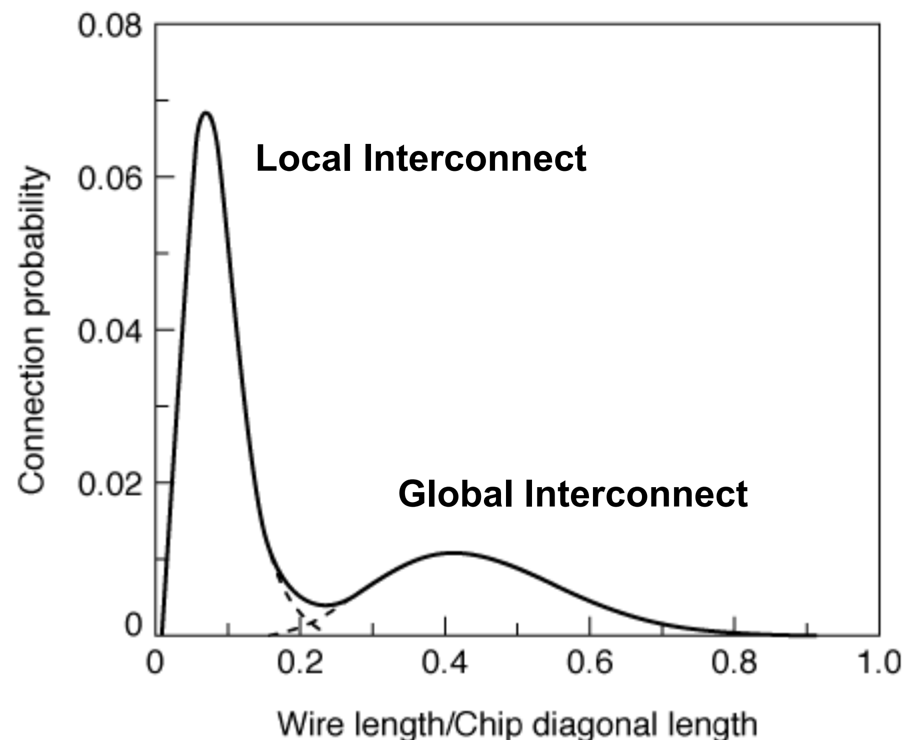
PHYSICAL SHIELDING

- **Best way to mitigate capacitive coupling is by inserting other metal layers between an attacker and a victim at a cost of using additional routing resources.**
- **Temporal and logical shielding will also work; see NOISE lecture.**



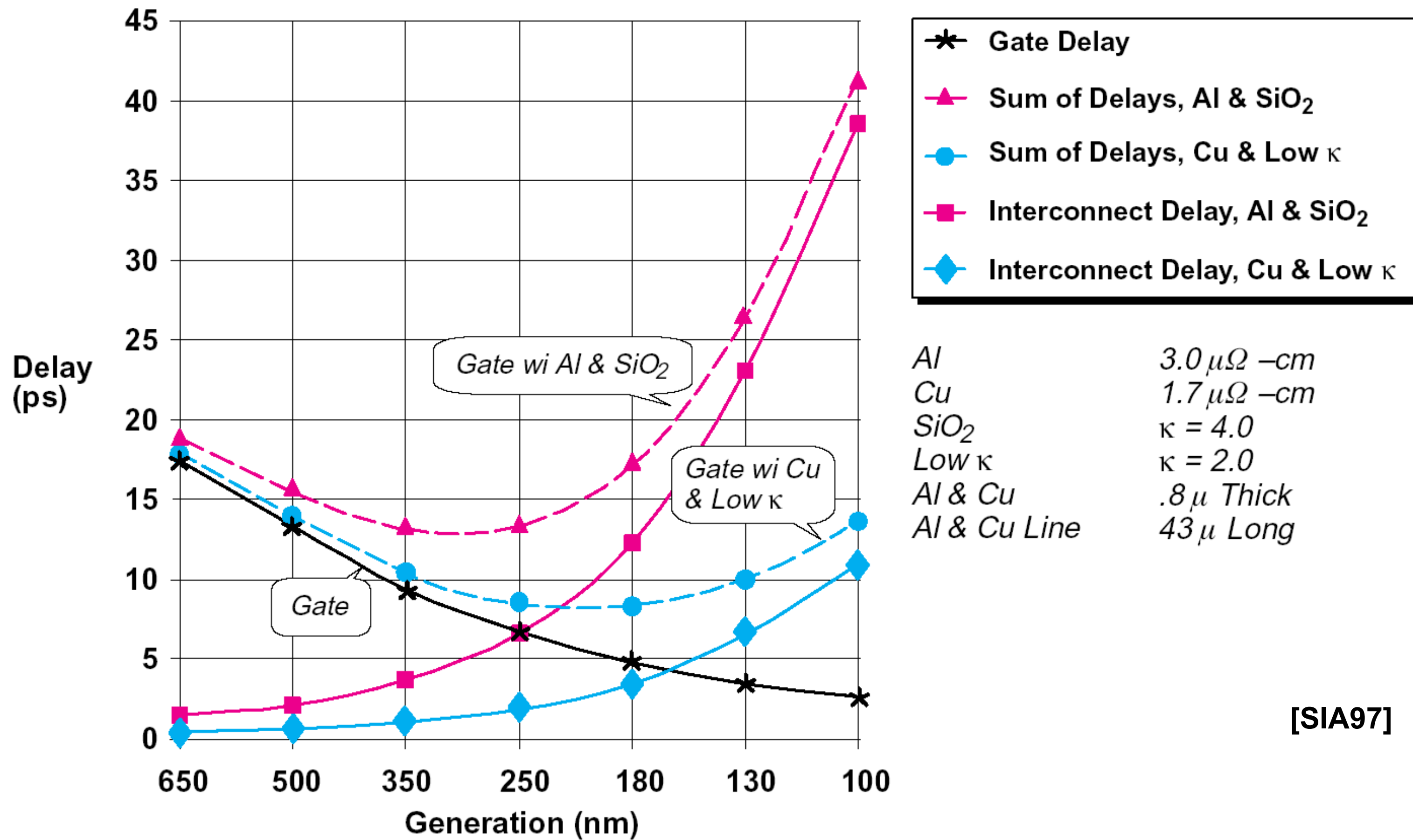
LOCAL AND GLOBAL WIRES

- **LOCAL WIRE:**
  - Scales in length
  - While transistors become faster, local wire delay remains unchanged
- **GLOBAL WIRE:**
  - Goes across the whole chip – does not scale.

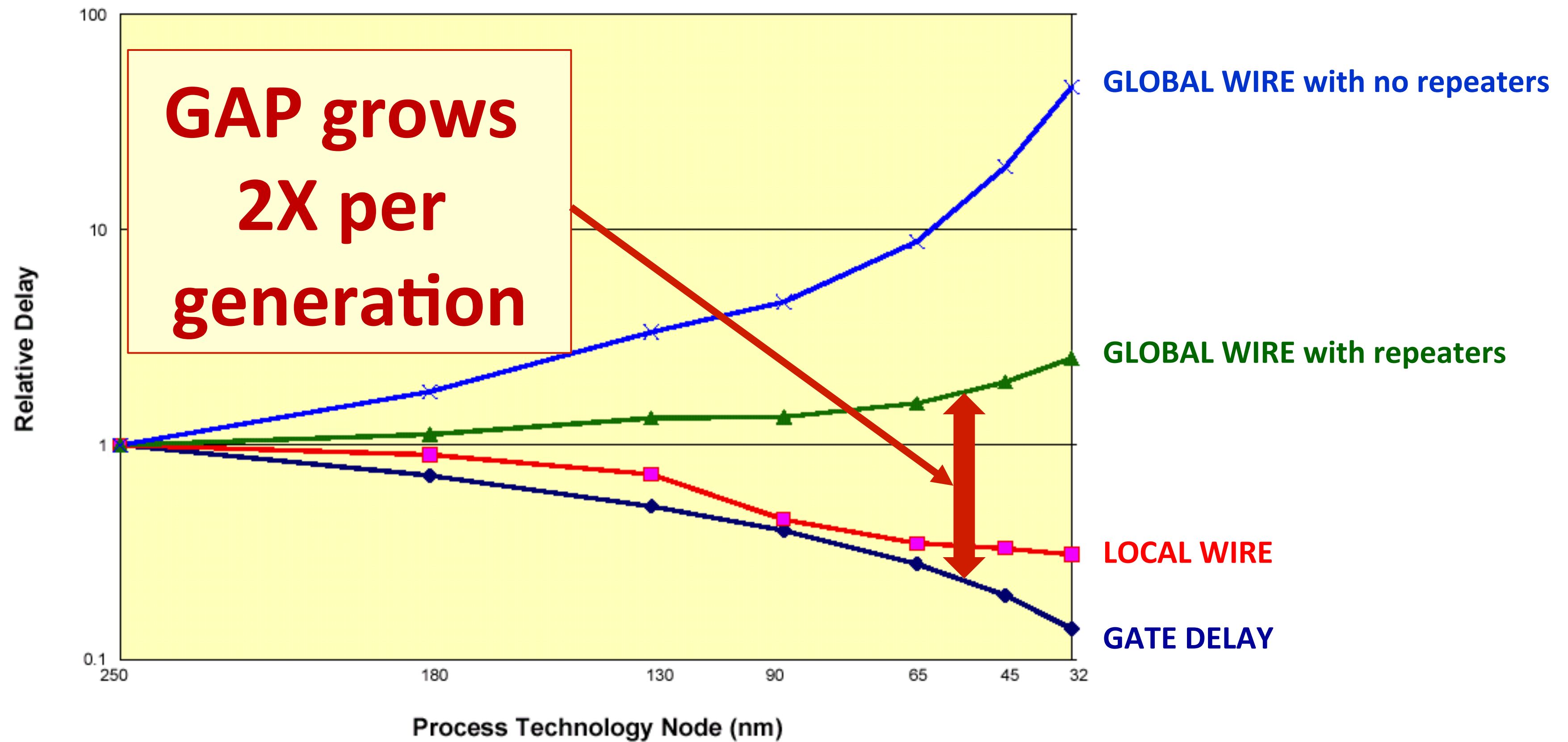


***Digital integrated circuit – a design perspective*, J. Rabaey Prentice Hall and a tutorial in SLIP by Dirk Stroobandt respectively**

GLOBAL WIRE SCALING PROBLEM



GLOBAL and LOCAL WIRE SCALING PROBLEM



source: ITRS 2003

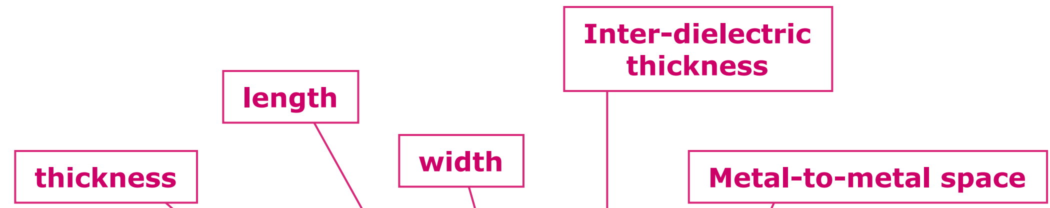
SCALING IMPACT on RC DELAY and POWER

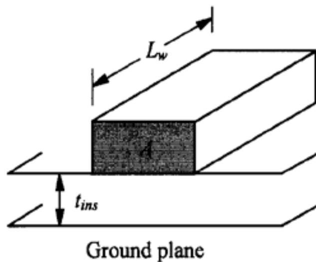
UT ECE

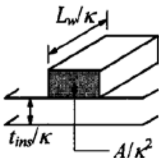
- **Resistance of the line increases as the square of the shrink but the line distance shrinks as the linear shrink.**
- **Resistance of the line thus is first order ~R_{old} /0.7**
- **Capacitance of the line goes down with the shrink factor so C new = C_{old}*0.7**

  therefore RC_{new} ~= RC_{old}
- **In practice the R does not scale perfectly due to edge effects.**
- **Interconnect CV^2F power *decreases* with the shrink because capacitance drops.**

INTERCONNECT SCALING







	Interconnect Parameters	Scaling Factor ($\kappa \geq 1$)
Scaling assumptions	Interconnect dimensions (t_w, L_w, W_w, t_{ins}, W_{sp})	$1/\kappa$
	Resistivity of conductor (ρ_w)	1
	Insulator permittivity (ε_{ins})	1
Derived wire scaling behavior	Wire capacitance per unit length (C_w)	1
	Wire resistance per unit length (R_w)	κ^2
	Wire RC delay (τ_w)	1
	Wire current density ($I / W_w t_w$)	κ

Assuming K = 1/0.7 ~ 1.43

RC delay = [(1/0.7)^2 * Rw * 0.7] * [Cw * 0.7] = RwCw

Current Density → **I / WwTw = [0.7 * I] / [(0.7 * Ww) * (0.7 * Tw)] = I / (0.7 * WwTw)**

Figures from: Y. Taur, T.H. Ning, Fundamentals of Modern VLSI Devices, Cambridge University Press, UK, 1998.

WIRE CAPACITANCE and TOTAL DYNAMIC POWER

UT ECE

- **EXAMPLE: Take a 10-metal layer 10mmx10mm micro processor; how much metal could there be with 5 layers at an average of 150nm pitch plus 5 layers at 300nm pitch?**
  - **4m**
  - **40m**
  - **400m**
  - **4000m**
  - **40000m**

- **Assuming 30% of this length is used for signals = 1200m**
- **And only 10% of them switch at any given time = 120m**
- **Assuming a capacitance per unit length of ~0.2fF/micron (or 0.2nF/m) and an operating frequency of 1GHz.**
- **The dynamic power due to metal interconnect switching could easily reach *12 Watts* @ 1 Volts (0.5*C*V^2*F).**

VIA and CONTACT SCALING

UT ECE

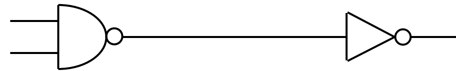
- VIAs and CONTACTs are similar to metal lines in that the cross sectional area goes down with the square and their height goes down linearly. As a result, VIA resistance increases linearly.
- Reliability may require dual VIAs when possible at the expense of increasing blockage for routing.
- Scaling will also result in fewer contacts per transistor layout; as the transistor current per unit width is increased, the contact resistance is becoming a larger issue.
- Contact resistance can be lowered with copper plugs (tungsten is currently used for contacts to diffusions) with improved barriers to meet reliability issues.

SCALING PROBLEMS with INTERCONNECT

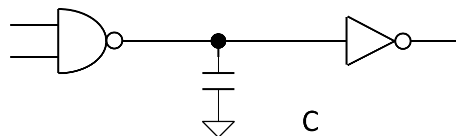
- Barrier metal is not as conductive as copper therefore the electrical cross section is smaller and resistance is higher.
  - If barrier metal is 4nm, the resistance of the line is ~8% greater than bulk Cu
- Edge effects cause carriers to have more collisions near the edges which increases resistance.
  - Edge effects are due to carriers interacting with the edges based on edge roughness (3-5nm) and random movement of carriers.
  - Impact is 5-10% on top of the barrier metal issue.
- Grain boundaries tend to be approximately the same as the width of the conductor so the number of grain boundaries does not decrease; the resistance of the line due to grain boundaries would go up 1/scale factor.
- Effective interconnect resistance can be 30-60% higher than bulk Copper.

EVOLUTION OF WIRE MODELING

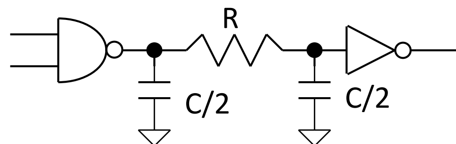
- **Ideal (R=0, R=0, L=0)**



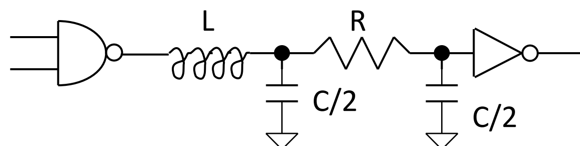
- **Capacitive only (C>0)**

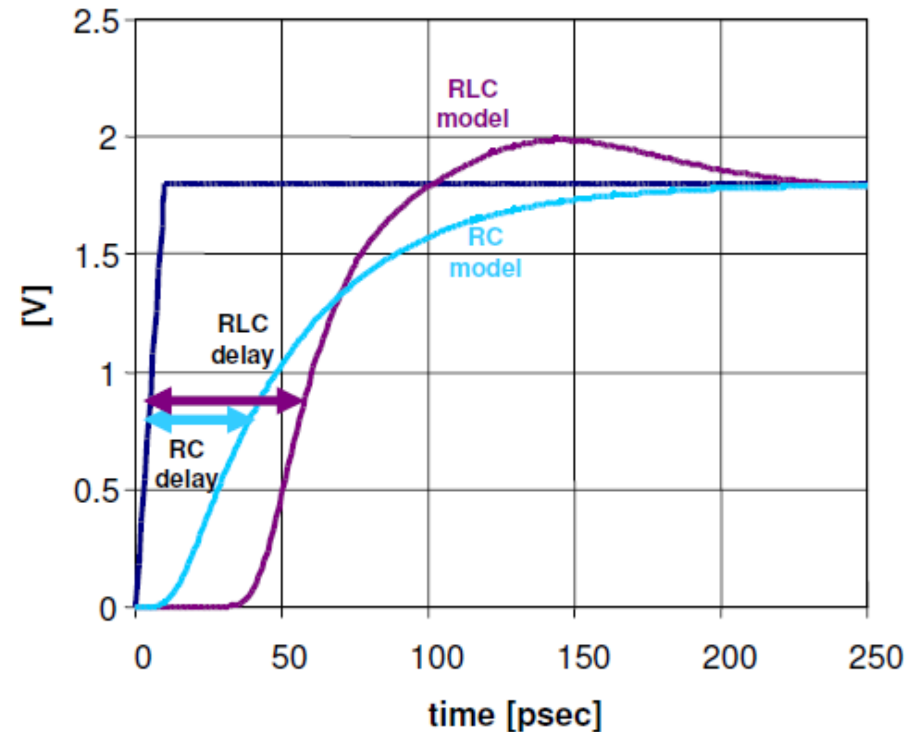


- **Resistive (C>0, R>0)**



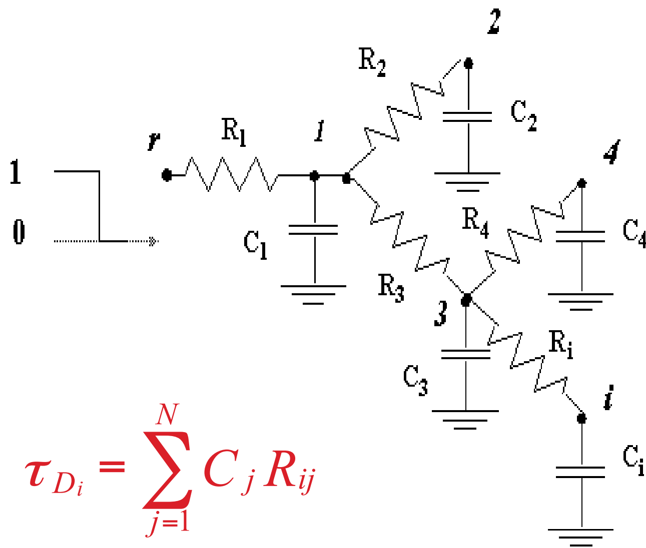
- **Inductive (C>0, R>0, L>0)**
  Longer delay, steeper slope, overshoot





ELMORE DELAY

- Consider a general RC tree network with ***no resistor loops*** and all of the capacitances are connected between a node and ground driven by one input node:
- First order time constant at node is a sum of RC components.
- All the upstream resistances are taken into account; all nodes contributes to the delay.
- Amount of contribution is the product of the cap at the node and the amount of resistance from source to the node.



$$\tau_{D_i} = \sum_{j=1}^{N} C_j R_{ij}$$

For ***N*** equal RC segments:

$$\tau_{DN} = \sum_{j=1}^{N} \frac{C}{N} \sum_{k=1}^{j} \frac{R}{N} = \frac{R}{N^2}\left(R + 2R + \cdots + NR\right) = RC\left(\frac{N+1}{2N}\right)$$

For large ***N*** where ***R*** and ***C*** are the total lumped resistance and capacitance of the wire:

τ_{DN} = RC/2

INTERCONNECT DELAY CALCULATION

- Delay without the wire:

  Delay = $R_{trans}C_{load}$

- Added delay from wire:

  The intrinsic wire delay (1/2 R_W C)

  Added delay from the wire cap (R_{trans} C)

  Added delay from the wire resistance (R_W C_{load})

  Delay = $R_{trans}(C+C_{load})$ + $R_W(C/2+C_{load})$

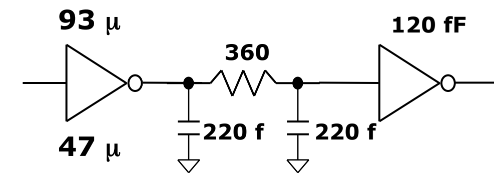
applying superposition

Intrinsic wire delay ~ 0.5*L^2*(R/um)*(C/um)

WIRE DELAY EXAMPLE

- In 0.18 um CMOS, assume a 2 mm M2 wire, minimum width (0.34um), and a 120fF load at the end.

- What is the intrinsic wire delay?
- What size driver should you use?



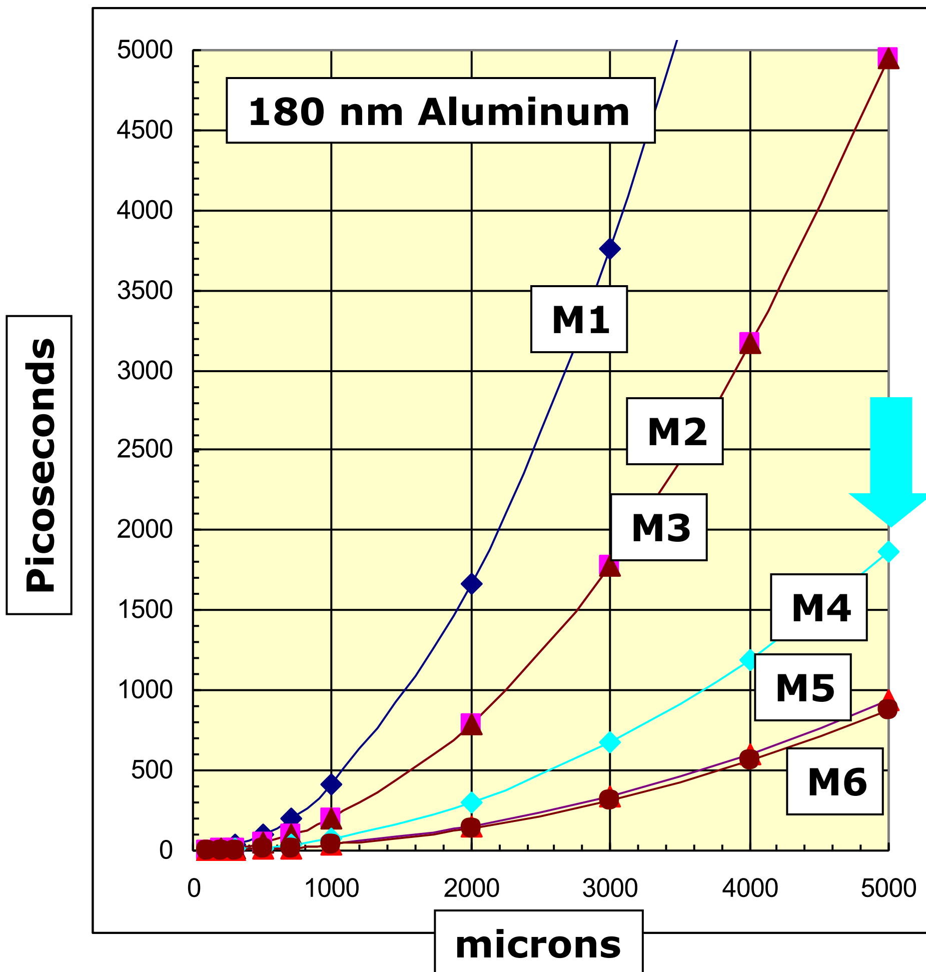
- Intrinsic Delay:
  - Using a worst case 0.18 ohms/um and 0.22 fF/um:
  - C = 2000*0.22 = 440 fF and R = 360 ohms.
  - Intrinsic wire delay = 0.5*RC ~ 79ps

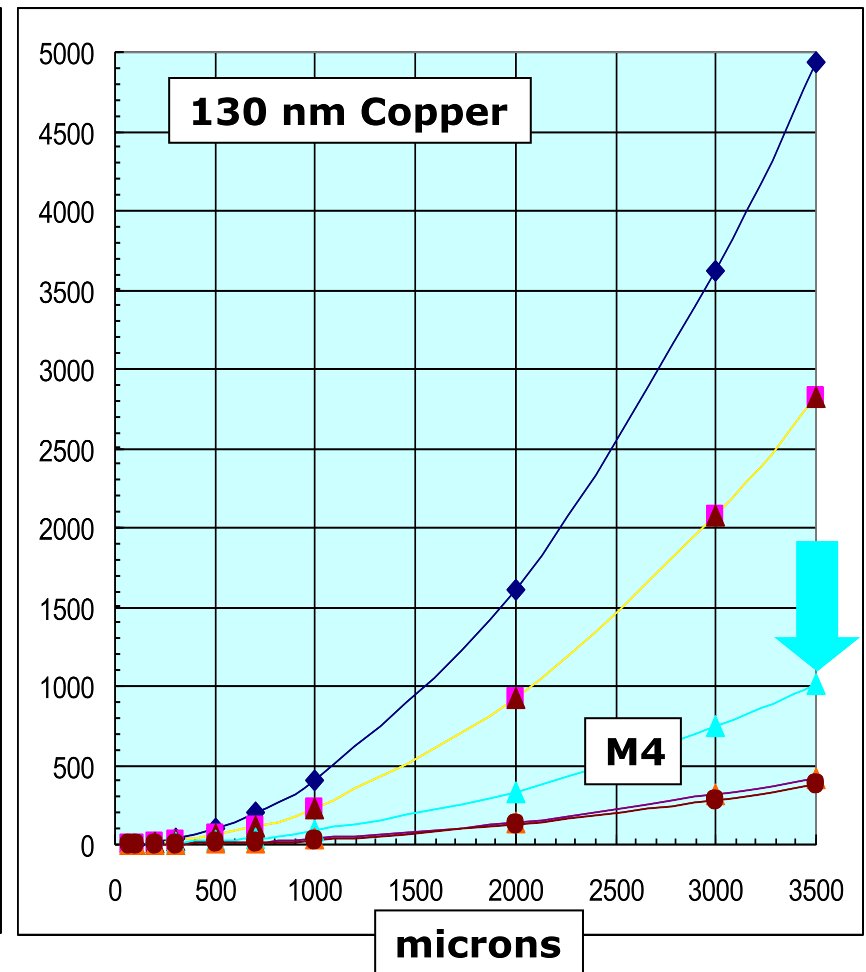
- Driver size (use FO=4 and P:N ratio ~2):
  - Input cap ~ (Cwire+Cload)/4 ~ 560fF/4 = 140fF
  - Using Cox ~1fF/um, PFET is 93um and NFET is 47um.

FO=4

- If M2 wire is 4mm long, intrinsic delay is ~ 317ps ….. 4X longer ….

INTERCONNECT DELAY CURVES

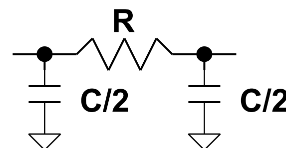




An M4 5mm 180 nm line (1.8ns un-repeated) would scale to 3.5mm in 130 nm; assuming fF/um remains constant, but ohms/um doubles, then the same wire would take 3.6ns. Copper interconnects (with their lower sheet resistance) takes this to 1.0ns.

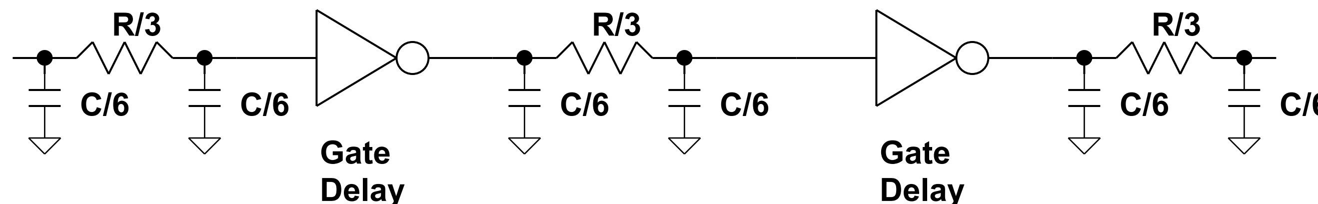
REPEATERS

- Repeaters are inverters or buffers inserted on signal paths to reduce over all delay.
- Inverter sizing to minimize delay results in FANOUTs in the order of 4 to 8.
- Using thicker (2x) higher metal layers can increase the distance between repeaters reducing the RC delay by ~ 4x or more depending on width and space of metal used.
- What is the distance that can be run before the RC is the same?

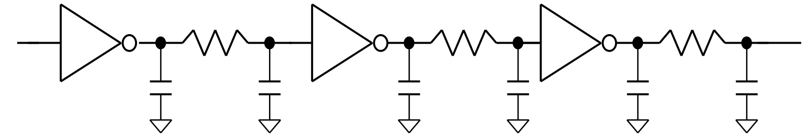


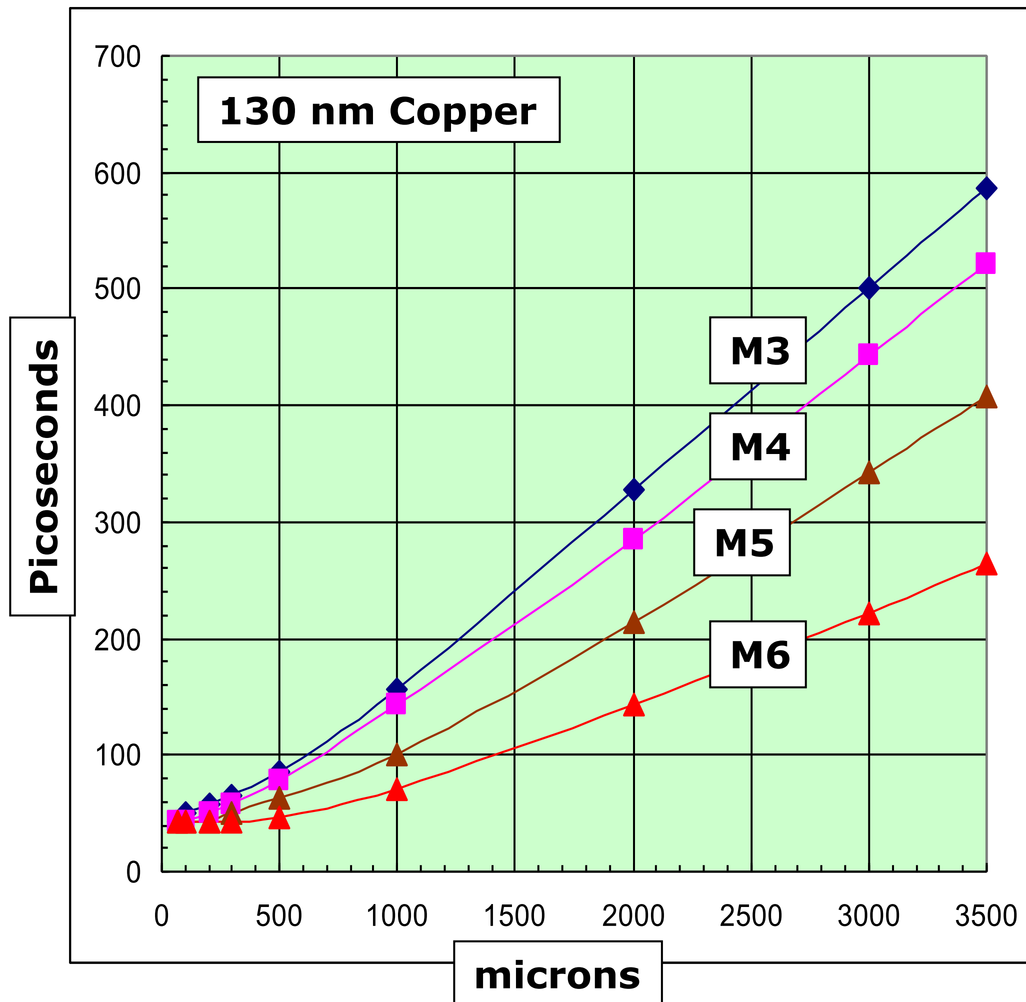
delay ~ RC/2

delay ~ RC/6 + 2* Gate Delay

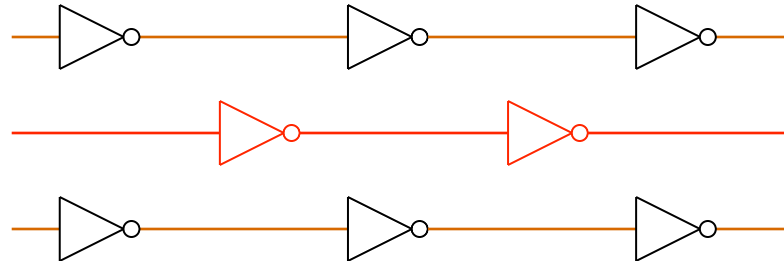


REPEATED INTERCONNECT

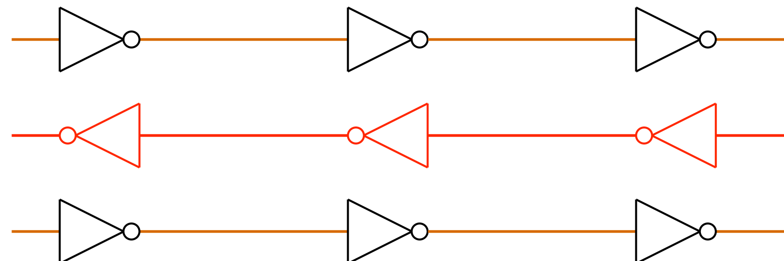




REPEATER PLACEMENTS



Staggering the inverters



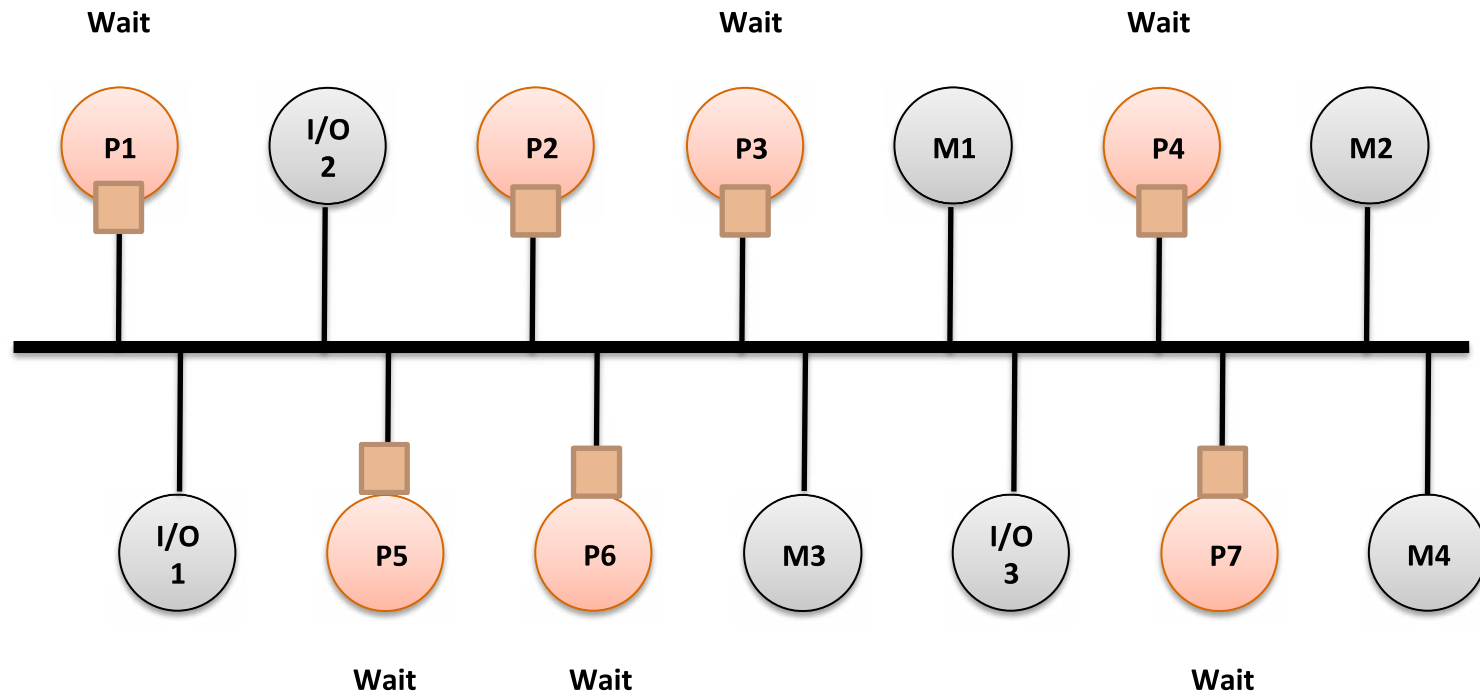
Avoiding the Miller cap by opposite going signals

Wire Delay vs. Logic Delay

Operation	Delay (.13μ)	Delay (.05μ)
32-bit ALU Operation	650ps	250ps
32-bit Register read	325ps	125ps
Read 32-bit from 8KB RAM	780ps	300ps
Transfer 32-bit across chip (10mm)	**1400ps**	**2300ps**
Transfer 32-bit across chip (200mm)	2800ps	4600ps

Source: W. J. Dally, HPCA Panel, 2002

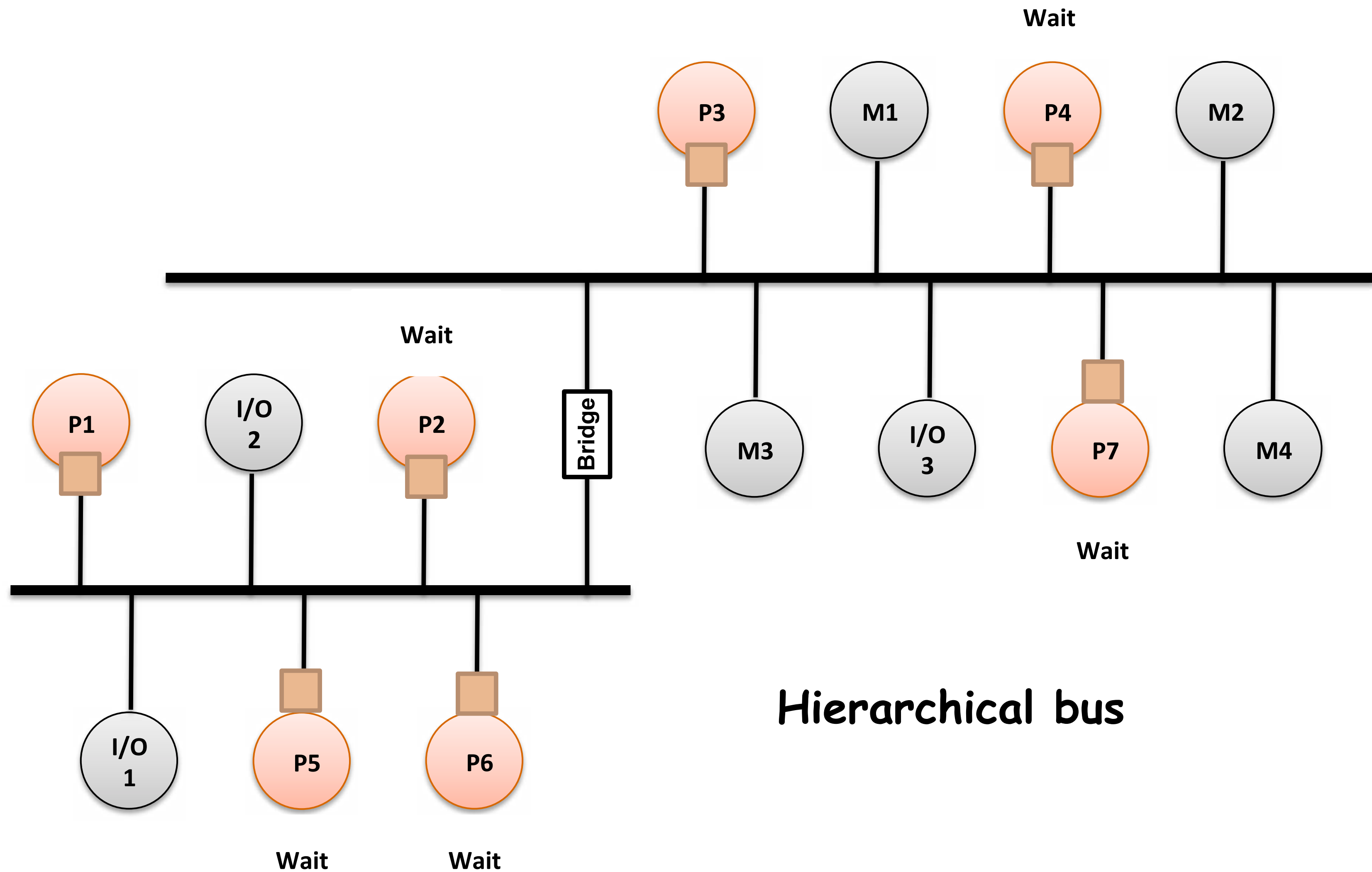
On-Chip Interconnections



Shared bus

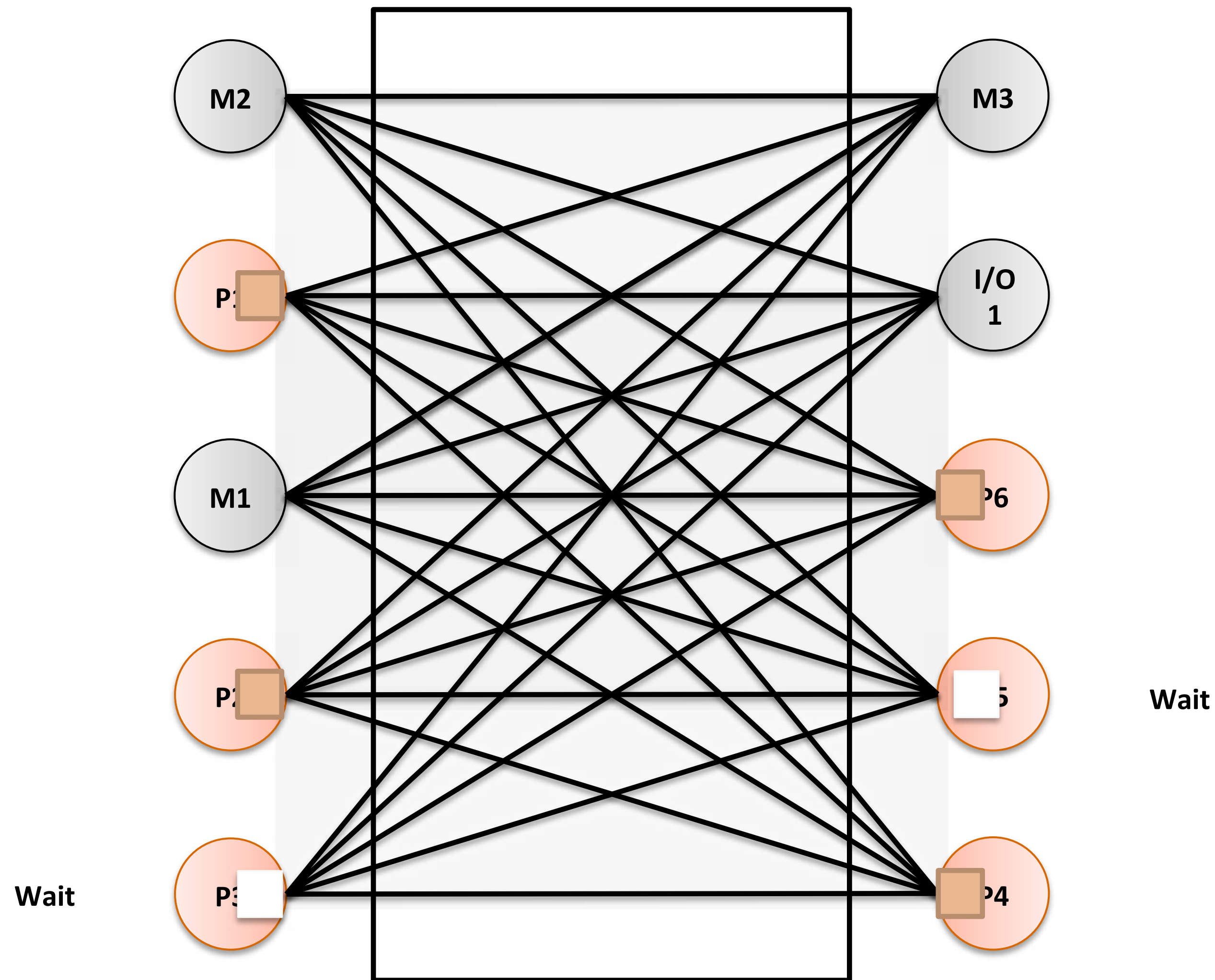
Source: Abderazek, 2013

On-Chip Interconnection



Source: Abderazek, 2013

On-Chip Interconnection, Cont'd



Interconnection Network

Source: Abderazek, 2013

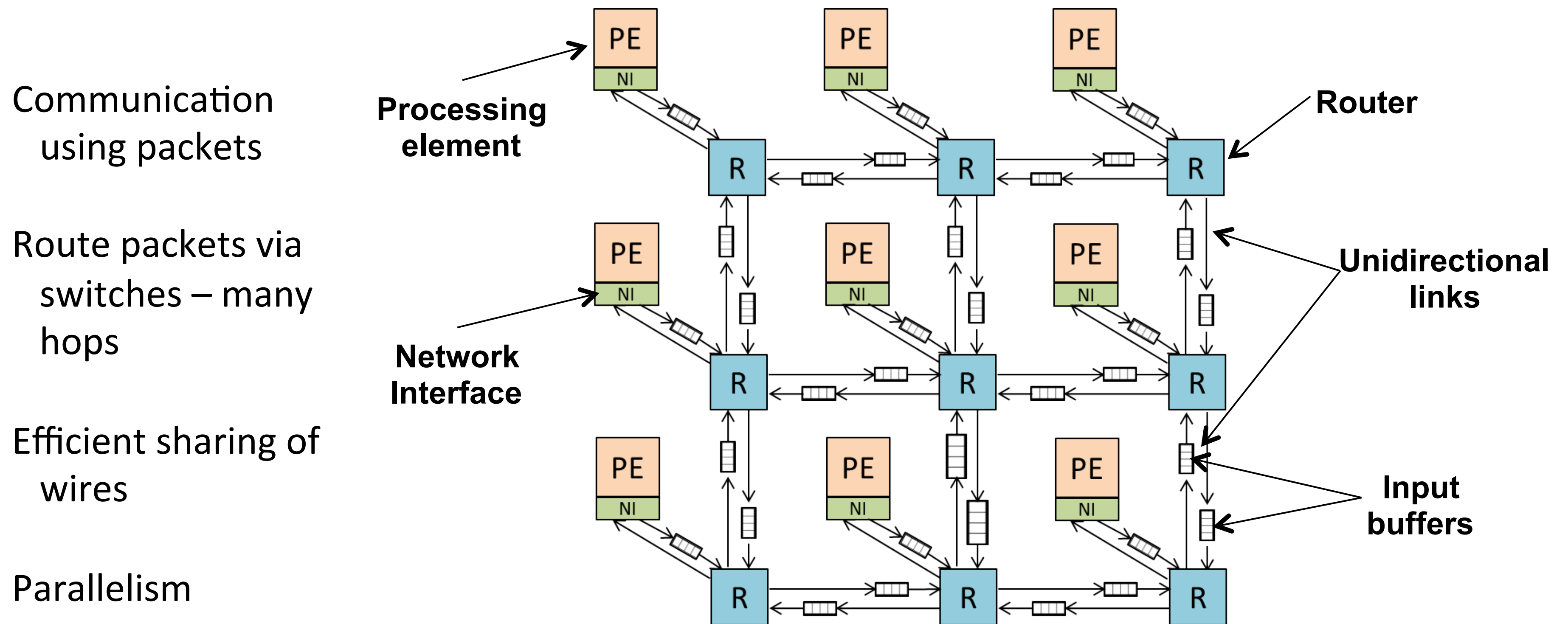
Network on Chip

Communication using packets

Route packets via switches – many hops

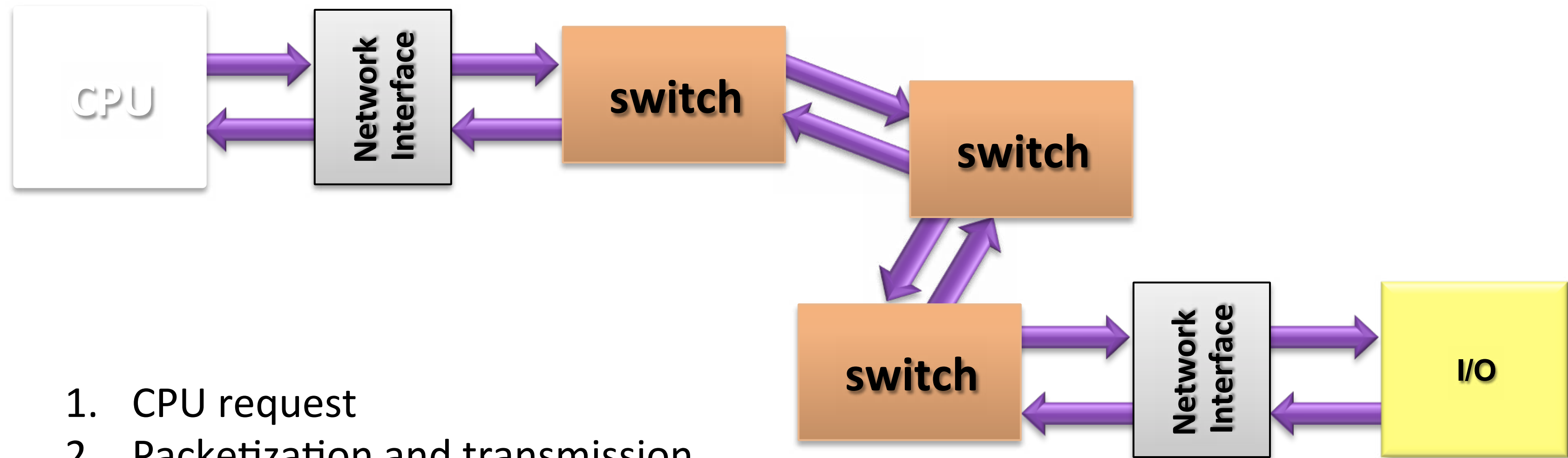
Efficient sharing of wires

Parallelism



Source: Abderazek, 2013

Example of NoC Operation



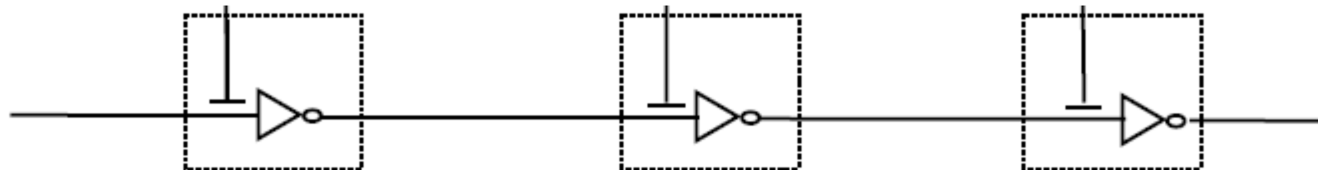
1. CPU request
2. Packetization and transmission
3. Routing
4. Receipt and de-packetization
5. Device response
6. Packetization and transmission
7. Routing
8. Receipt and de-packetization

Source: Abderazek, 2013

NoC Deals with Global Wire Delay

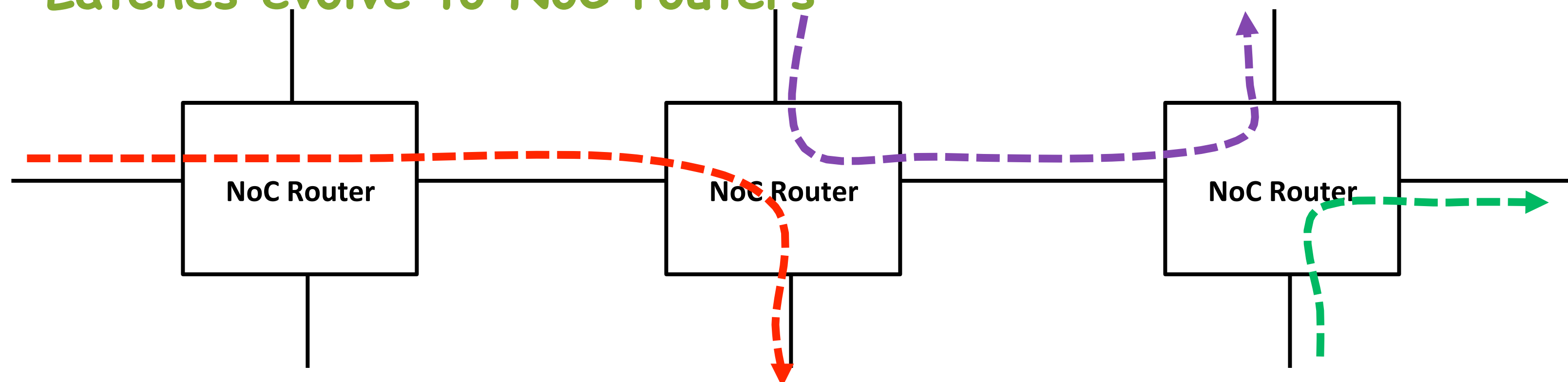
Long wire delay is dominated by Resistance

Add repeaters



Pipeline using latches

Latches evolve to NoC routers



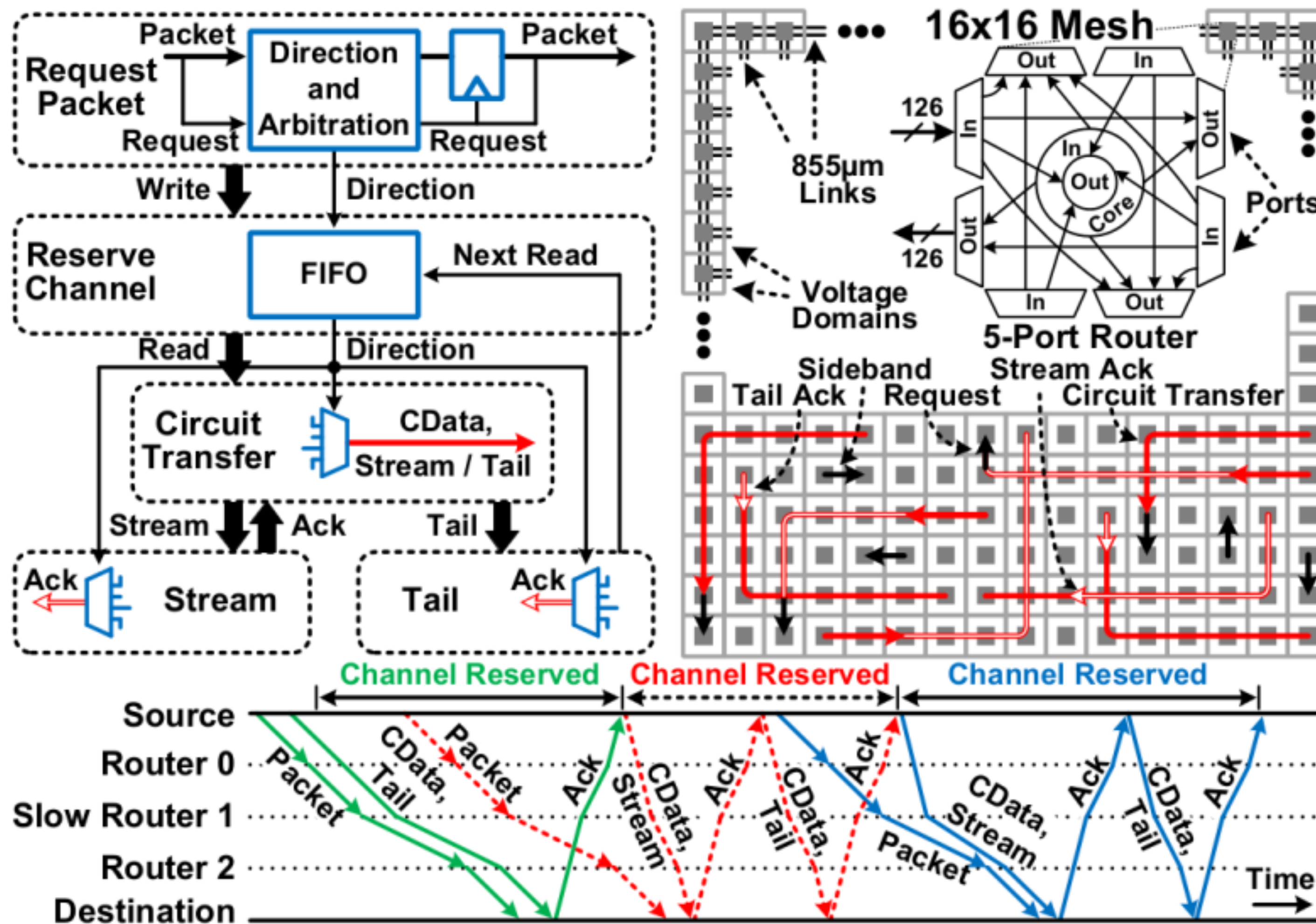
Source: Abderazek, 2013

Cisco CRS-1 Router

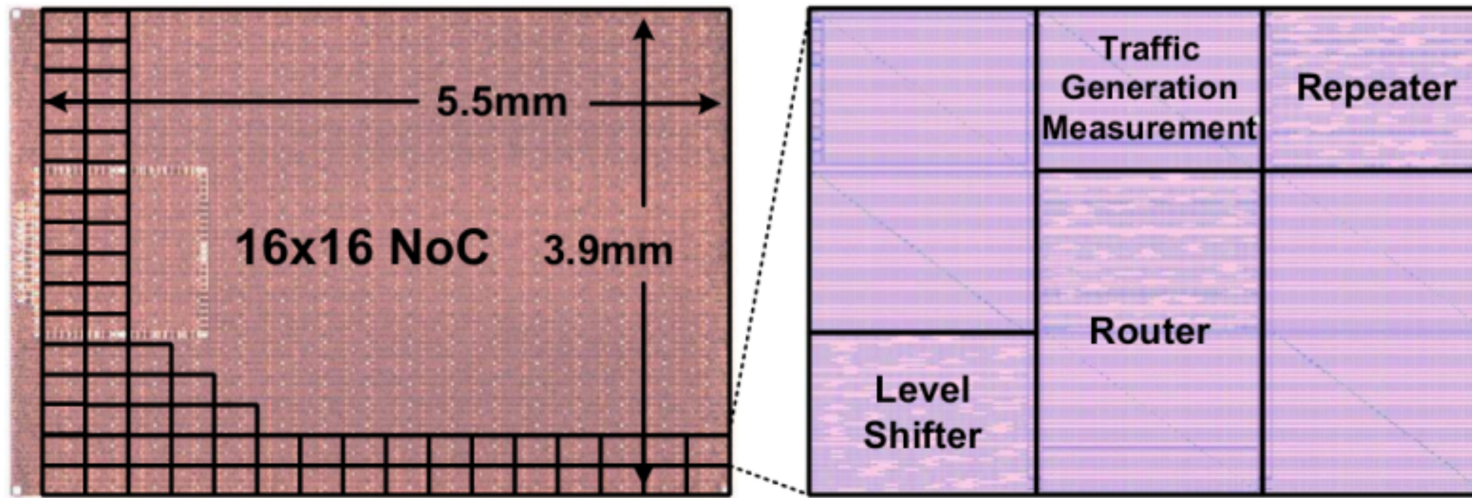
The CRS-1 Router is a “Packet Processor” on a chip
Chip has 188 extensible network processors

Intel NoC Overview (ISSCC 2014)

16X16 NoC, Source-Synchronous Hybrid Packet/Circuit Switching
340mV to 0.9V, 20.2 Tb/s
22nm Tri-Gate CMOS



Intel NoC Measurements



Process	22nm Tri-gate CMOS	Nominal Operation	0.9V, 25°C
Topology	16x16 Mesh	Throughput	20.2Tb/s
V_{DD}/Clock Domains	256	Energy Efficiency	7.0Tb/s/W
Data Bus Width	112b	Bisection Bandwidth	2.8Tb/s
Packet Size	11b Control + 32b Data	Circuit-switched Latency/Hop	407ps
Circuit Size	3b Control + 80b Data	Near-threshold Operation	430mV, 25°C
Link Length	855µm	Throughput	3.4Tb/s
Die Area	23mm^2	Peak Energy Efficiency	18.3Tb/s/W
Equivalent NoC Area	167mm^2	Ultra-low-voltage Operation	340mV, 25°C
Router Area	138µm x 109µm	Throughput	946Gb/s
Transistor Count	150M	Router Power	363µW

CONCLUSIONS

UT ECE

- VLSI interconnect is dominating design issues in deep-submicron designs. It is not just about transistors anymore.
- Designers need to learn how to design around the increasing wire resistance and coupling as technology scales; we need to also architect around it.
- Insertion of repeaters is a good solution, but they cost as well (power, delay, area, etc.).
- Interconnecting many cores in a System-on-Chip is facilitated by an on-chip network fabric

BACKUP

DESIGN RULES OF THUMB

- *RC* delays should be considered when $\tau_{RC} > \tau_{gate}$ of the driving gate

$$L_{crit} > \sqrt{\tau_{pgate} / 0.38 R_{unit} C_{unit}}$$

  - actual L_{crit} depends upon the size of the driving gate and the interconnect material

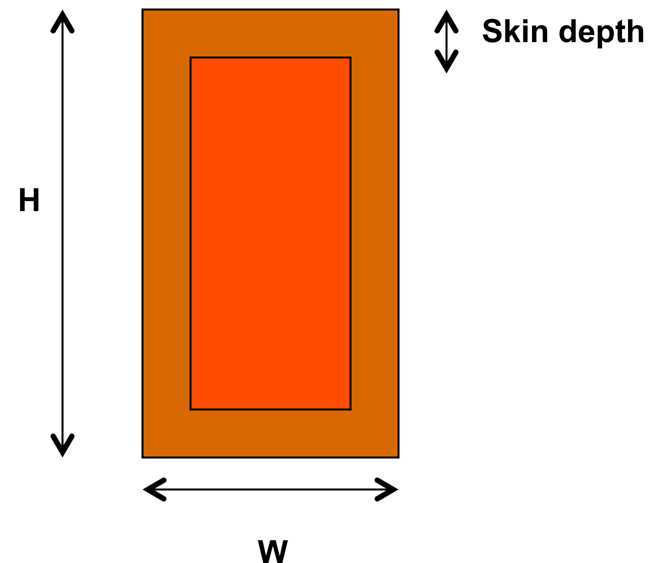
- *RC* delays should be considered when the rise (fall) time at the line input is smaller than *RC*, the rise (fall) time of the line

$$\tau_{rise} < RC$$

  - when not met, the change in the signal is slower than the propagation delay of the wire so a lumped *C* model suffices

Skin Effect

- At high frequencies current flows on the surface of the wire. This phenomenon is known as Skin Effect.
- The depth at which the current flows inside the conductor is known as skin depth (δ) and is function of the frequency of the signal.
- Due to the skin effect, the wire will have increased effective resistance.
- Increased current density along the surface can also cause reliability issues.



ROUTING/ROUTER ISSUES

- VIA resistance and current density through the VIA are important, if multiple VIAs are required better routers are required.
- Relaxation of metal spaces when possible helps improve RC and power.
- Route lines together that switch at different times in the cycle to improve speed (temporal shielding).
- Make every other track long route lines and the others as short route lines when possible (statistically reduces coupling).
- Coupling issues will create significant noise issues if not properly addressed (refer to NOISE lecture).
- Solving noise issues can delay product introduction or cause recalls on existing products.

Interconnect Issues =< 32nm

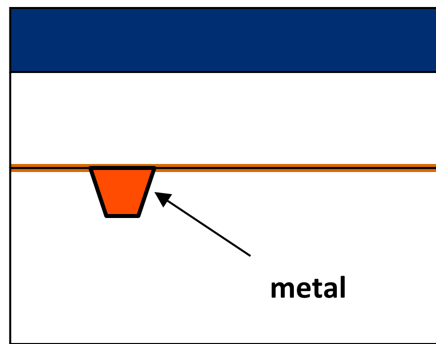
- **Material**
  - Integration and characterization challenges with the rapid introduction of new materials and processes to reduce dielectric permittivity and decrease interconnect power and delay.
- **Reliability**
  - Scaling of interconnect width, barrier and capping thicknesses.
  - Introduction of new materials and processes exacerbate the already challenging electrical (EM), and mechanical reliability.
  - Improved failure detection techniques, testing methodologies, modeling and identification of potentially new failure mechanisms will be required.
- **Metrology**
  - Line edge roughness, trench depth and profile, via shape, etch bias, thinning due to cleaning, planarization effects. The multiplicity of levels combined with new materials, reduced feature size, and pattern dependent processes create this challenge.

Interconnect Challenges =< 32 nm (cont.)

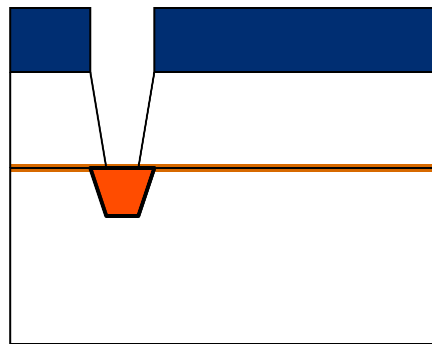
- **Manufacturability**
  - Implementation of increasingly complex integration schemes into high volume manufacturing demands close alignment of tools, material and process development. Compatibility of integration options with manufacturing requirements, such as high yields and process robustness.
- **Cost**
  - As feature sizes shrink, interconnect processes must be compatible with device roadmaps and meet manufacturing targets at the specified wafer size.
  - Plasma damage, contamination, thermal budgets, cleaning of high A/R features, defect tolerant processes, elimination/reduction of control wafers are key concerns.

Interconnect Process Dual Damascene

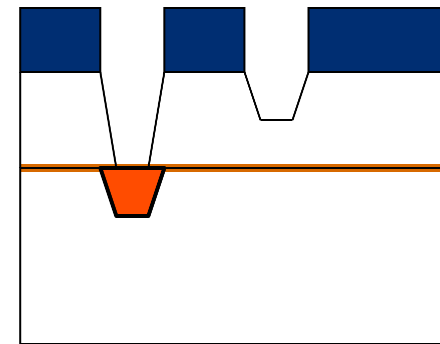
Dielectric Etch Stop



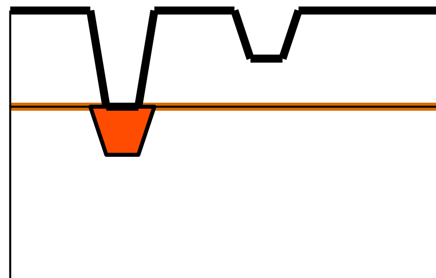
Pattern Via First



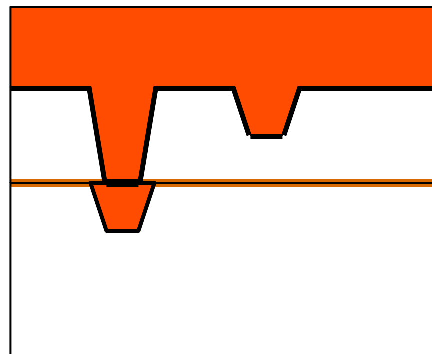
Pattern Metal Second



Deposit Barrier Metal



Copper Layer



Planarize & Cap Copper