Problems with Scaling Cu Interconnects and Near-term Alleviation with ALD Barrier

Pawan Kapur

Department of Electrical Engineering
Stanford University

November 21st, 2002

kapurp@stanford.edu

This work funded by MARCO Interconnect Focus Center

Outline

- Cu Interconnect Scaling Induced Problems
  - Interconnect metrics
- Technology impact on interconnects

  (Near-term alleviation with ALD barrier)
  - Realistic Resistance
    - Cu diffusion Barrier
    - Electron Scattering
    - Comparison of Cu with Al
  - Capacitance: low-k may not be adequate
- Performance assessment with realistic parameters
  - Delay
  - Repeaters
  - Power
- Long-term solutions: novel communication mechanisms
  - Optical interconnects

Performance Metrics

- **Signaling**
  - Delay
  - Power dissipation
  - Bandwidth
  - Area
  - Self heating
  - Data reliability (Noise)
    - Cross talk
    - ISI: impedance mismatch

Reliability
- Electromigration

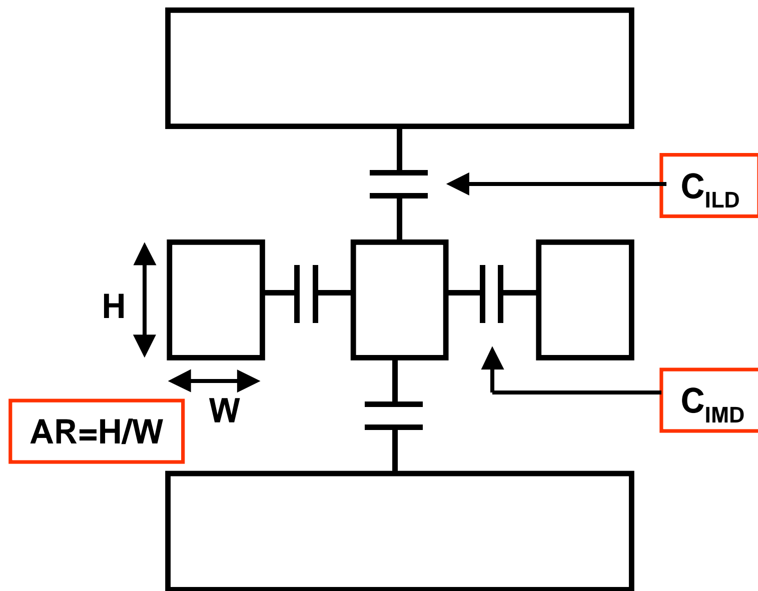
- **Clocking**
  - Timing uncertainty (skew and jitter)
  - Power dissipation
  - Slew rate
  - Area

- **Power Distribution**
  - Supply reliability

Depend on R and C and L !

Interplay Between Signaling Metrics (I)

Simplistic formulae to see trends



$$\tau \propto RC_{inttot}$$

$$P = \alpha C_{inttot} V^2 f \propto C_{inttot}$$

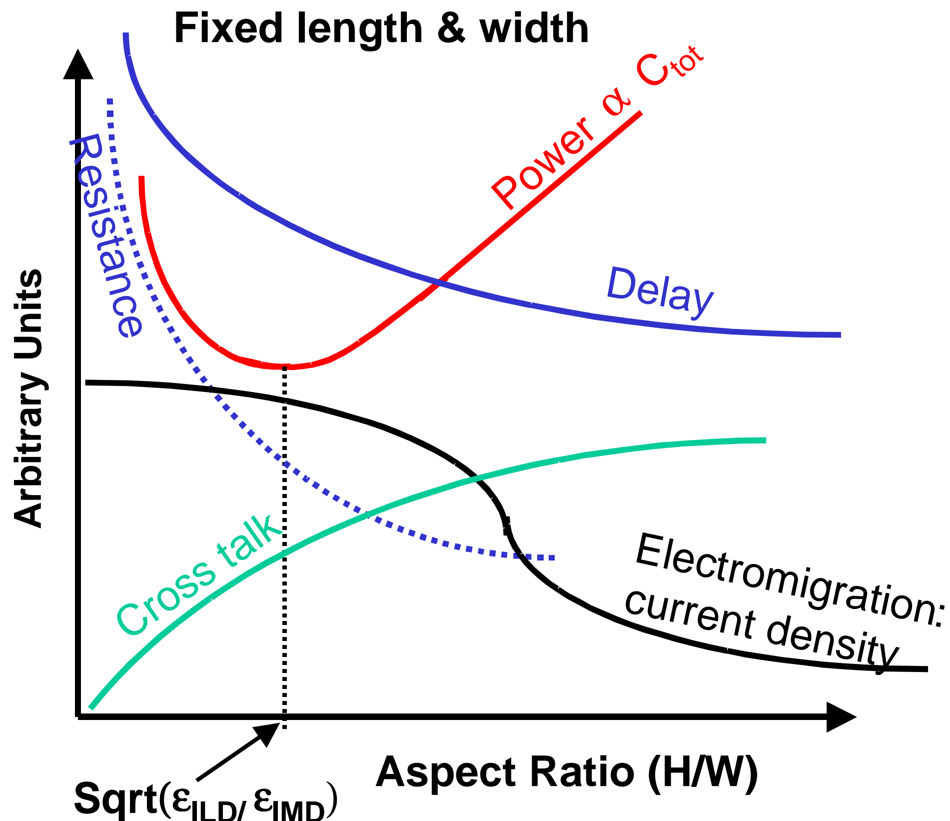
$$X_{talk} \propto \frac{C_{IMD}}{C_{inttot}} = \frac{1}{1 + \frac{\left(\varepsilon_{ILD} / \varepsilon_{IMD} \right)}{AR^2}}$$

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

$$R = \frac{\rho L}{(AR)W^2}$$

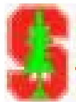
Minimum in power exists wrt AR

Interplay Between Signaling Metrics (II)



- AR increase (tradeoffs)=>
  - Better delay and electromigration
  - Worse power and cross talk
- In future increasing aspect ratio may not help
- Explains why AR dropped when Al to Cu switch

- **Pay attention to different metrics simultaneously rather than just delay**
- **Design window quite complex**

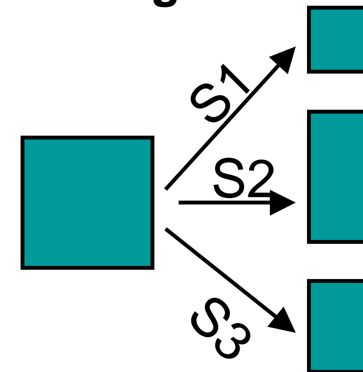


Motivation (I): Future Problems (Delay)

Simple Scaling Scenarios

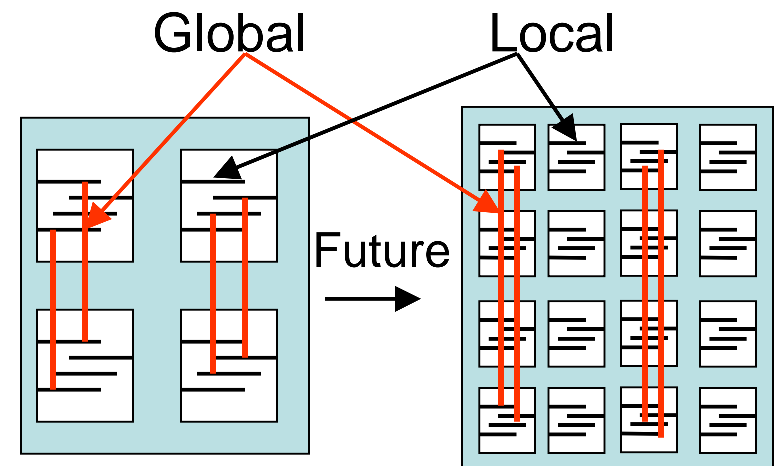
- **Local:** Wires whose length shrinks
  - S1: AR maintained (3D shrink)
    - R up by α **(worse)**
    - C down by α (geometrical effect)
    - C down by low-k material
    - RC delay down as low-k
    - **Delay going up compare to gate delay**

Wire Cross section Scaling Scenarios



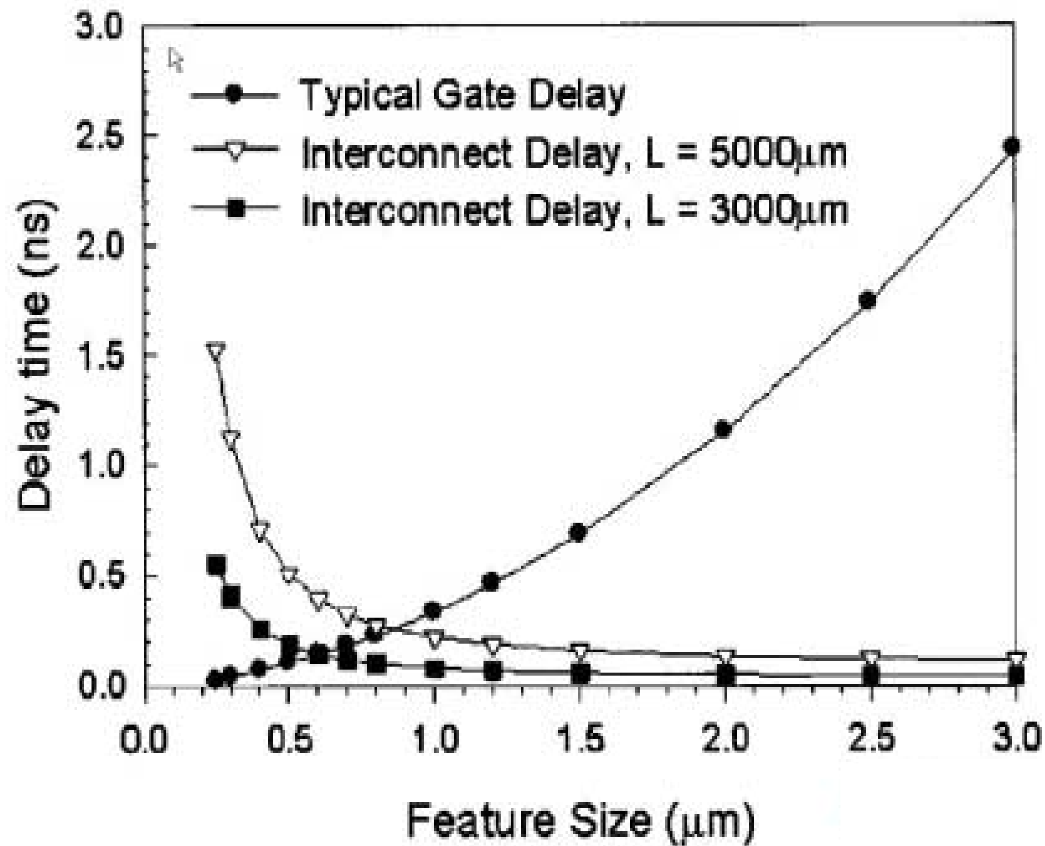
- **Semiglobal/Global:** Length does not shrink
  - Much worse than local (Will focus on global)

Wire length Scaling



All types of signal wires delays are deteriorating wrt gate delay with scaling even with new low-k materials !

Motivation (II): Future Problems (Delay)

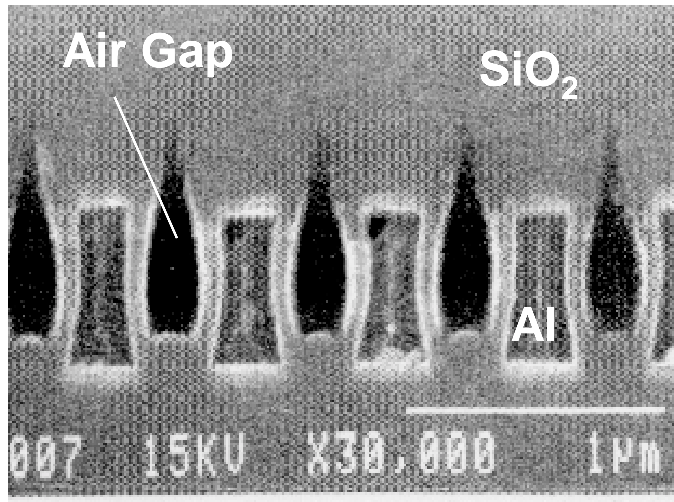


R. Havemann et. al., Proc. of IEEE,vol. 89, No.5, 2001

Careful about gate delay comparisons!

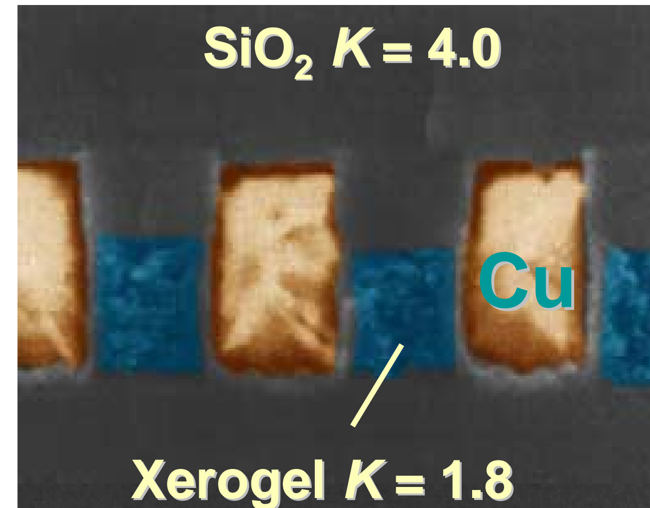
Is Copper/low-k Enough?: Long Term

Air-Gap/ Al



Stanford

Cu/xerogel



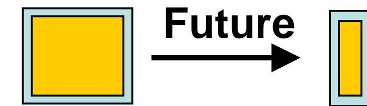
- **Old dielectric SiO_2 K = 4**
- **Polymers or air-gaps K = 2 - 3**
- **Ultimate limit is air with K = 1**

Interconnect DC Resistance: Technology Effects with Scaling

Cu Resistivity: Effect of Line Width Scaling

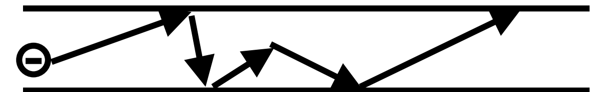
Diffusion barrier

- Consumes progressively larger fractional **area**
  - Barrier thickness (BT) doesn't scale
  - Higher AR => larger barrier area
- Technology dictates
  - Minimum thickness: reliability constraints
  - Profile: deposition technology



Electron surface scattering

- Reduced electron mobility with scaling
- Depends on
  - Ratio of λ_{mfp} to thickness
  - Interface quality: **Roughness (P)**

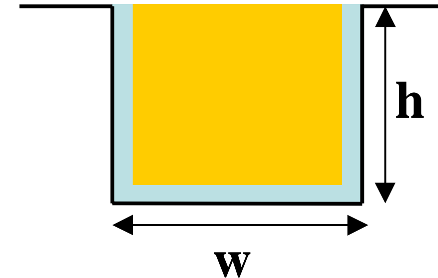


➢**Resistivity of metal wires could be much higher than bulk value**
➢**Problem is worse than anticipated in the ITRS roadmap**

Cu Resistivity: Theoretical Background

- **Barrier Effect**

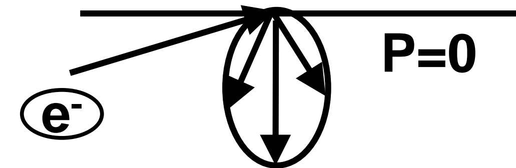
$$\frac{\rho_b}{\rho_o} = \frac{1}{1 - \frac{A_b}{AR \quad w^2}}$$



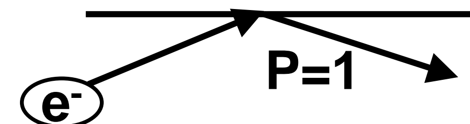
- **Electron Surface Scattering Effect**

$$\frac{\rho_s}{\rho_o} = \frac{1}{1 - \frac{3(1-P)\lambda_{mfp}}{2d} \int_1^{\infty} \left(\frac{1}{X^3} - \frac{1}{X^5} \right) \frac{1-e^{-kX}}{1-Pe^{-kX}} dX}$$

- $k = d/\lambda_{mfp}$
- P (phenomenological parameter)
  - Surface properties
    - Rms roughness (asperity): temp., thickn.,
    - Surface potentials: film types
  - Incidence angles



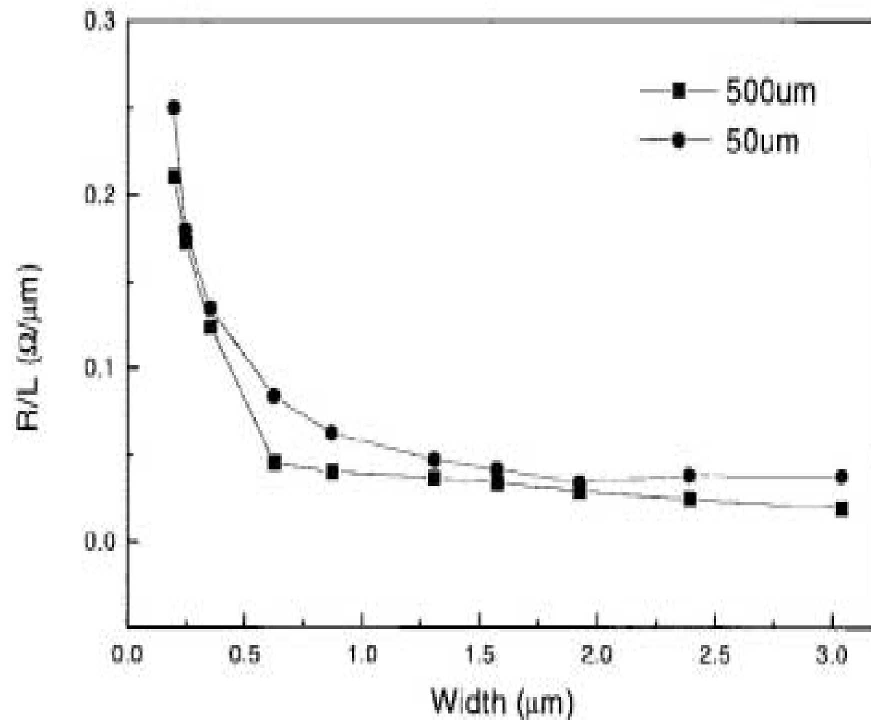
Diffuse scattering: lower mobility



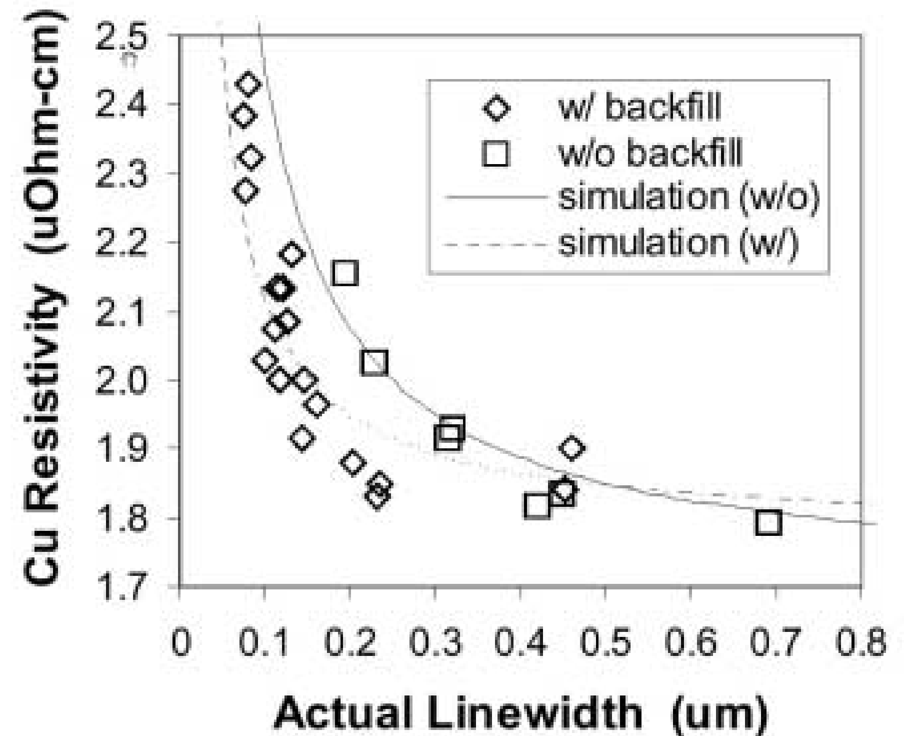
Elastic scattering: no mobility change

K. Fuchs, *Proc. Cambridge Phil. Soc.*, 1938
E. H. Sondheimer, *Advan. Phys*, 1952.

Cu Resistivity: Experimental Results

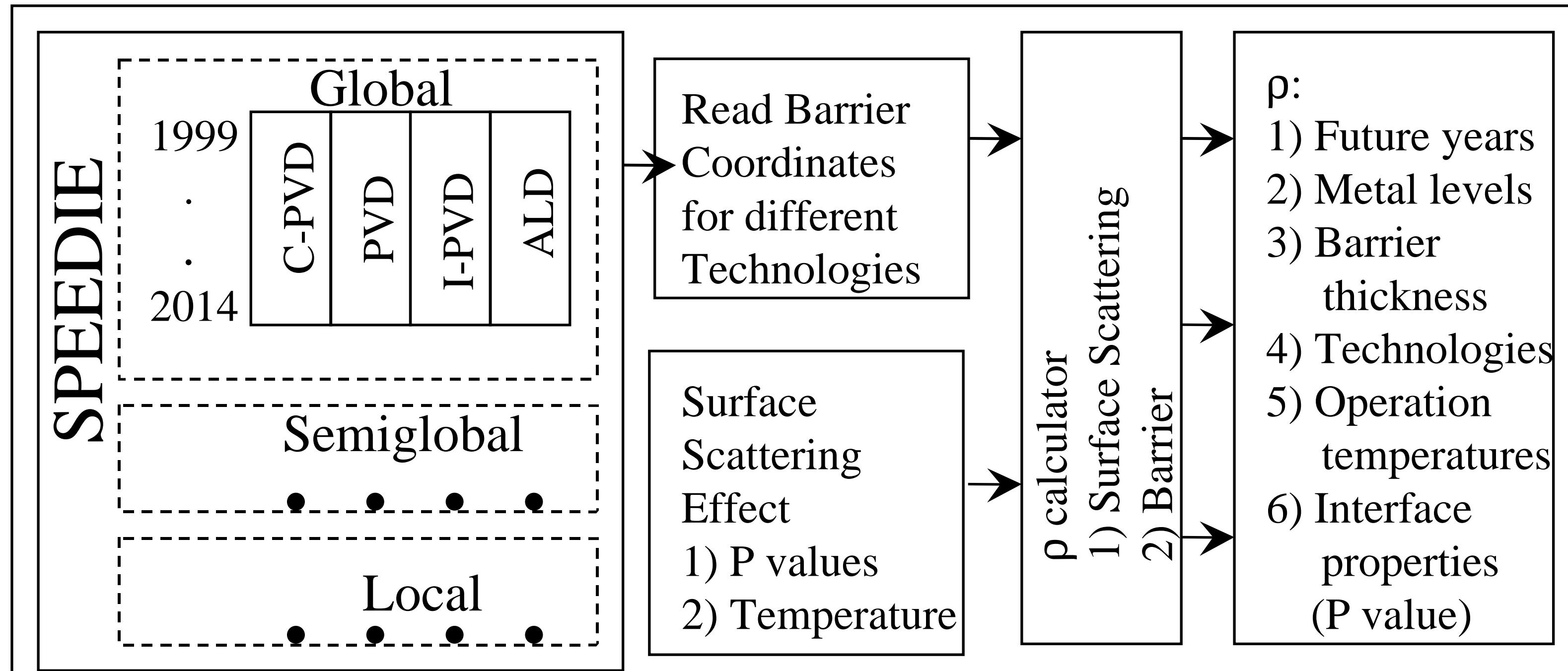


F. Chen and D. Gardner, *IEEE Electron Device Letters*, December 1998



Q. T. Jiang et. al., *Proc. IITC*, 2001, pp. 227-229

Methodology for Resistivity Calculations

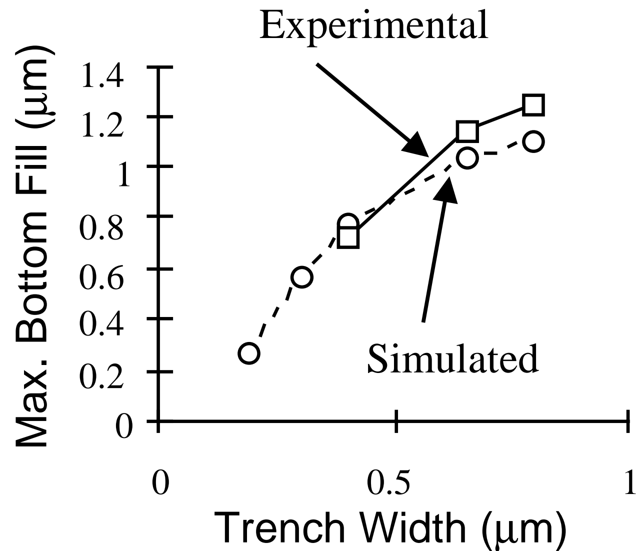


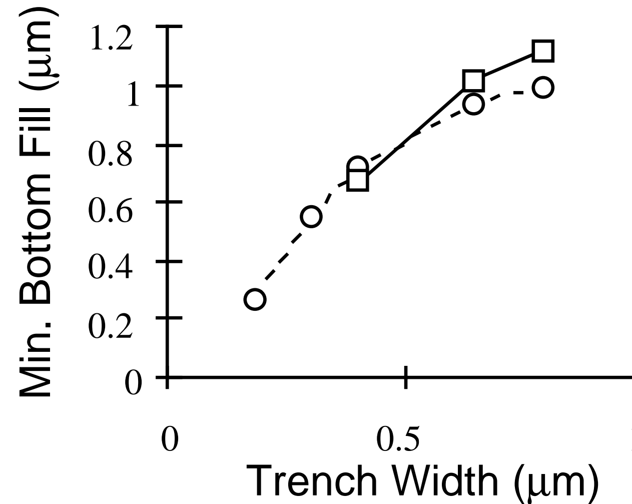
- **Surface scattering effect**
  - P from 0 to 1 in step of 0.25
  - Temperatures: 27^{0}C and 100^{0}C

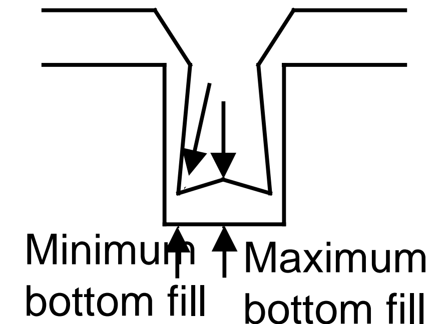
- **Barrier profiles**
  - SPEEDIE
    - Different technologies
    - 180 to 35nm node geometry
    - Tiers
  - Two barrier thicknesses: 5 and 10nm

IPVD Profile Modeling Using SPEEDIE

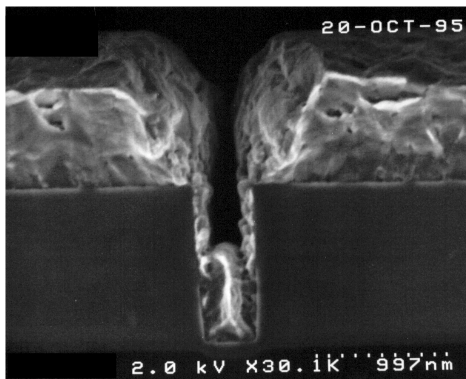
Comparisons between SPEEDIE and experiments for Al deposition



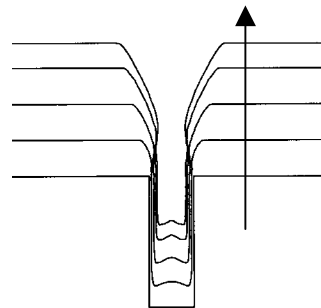




0.4µm trench

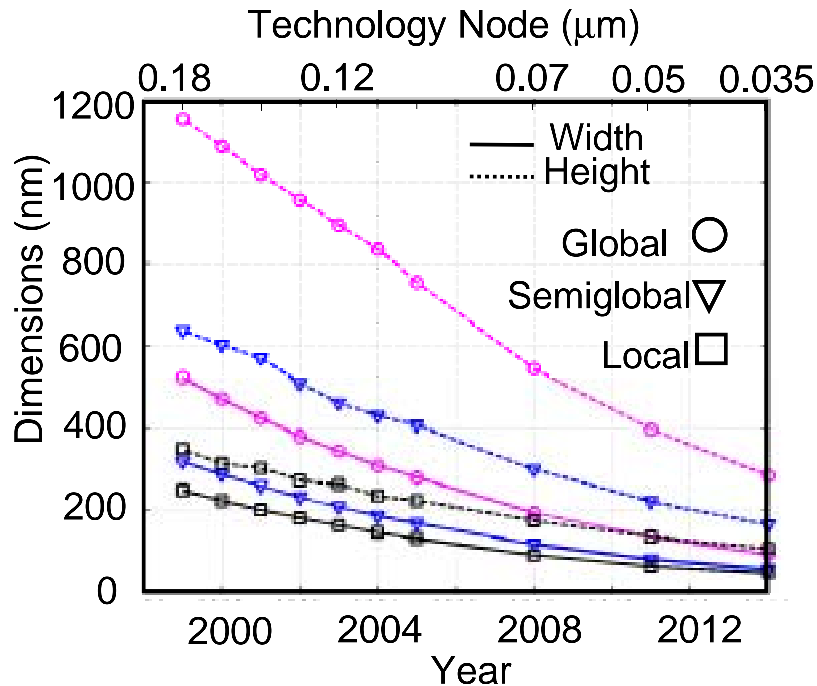


Time Evolution



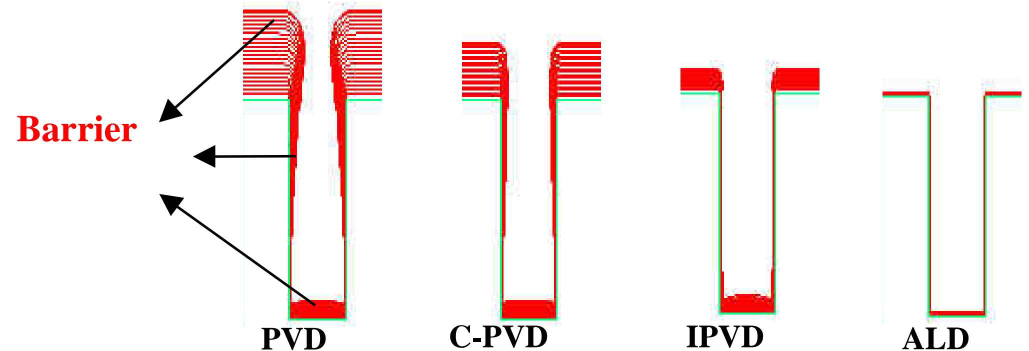
Establishes SPEEDIE's credibility for metal deposition profile simulation using IPVD

Methodology for Resistivity Calculations



SPEEDIE used to simulate barrier profiles

- **Different technologies**
- **Different geometries: ITRS**
  - **180 nm to 35 nm technology node**
  - **Local, semi-global, global**
- **Two barrier thicknesses: 5 and 10 nm**
- **Surface scattering effect**

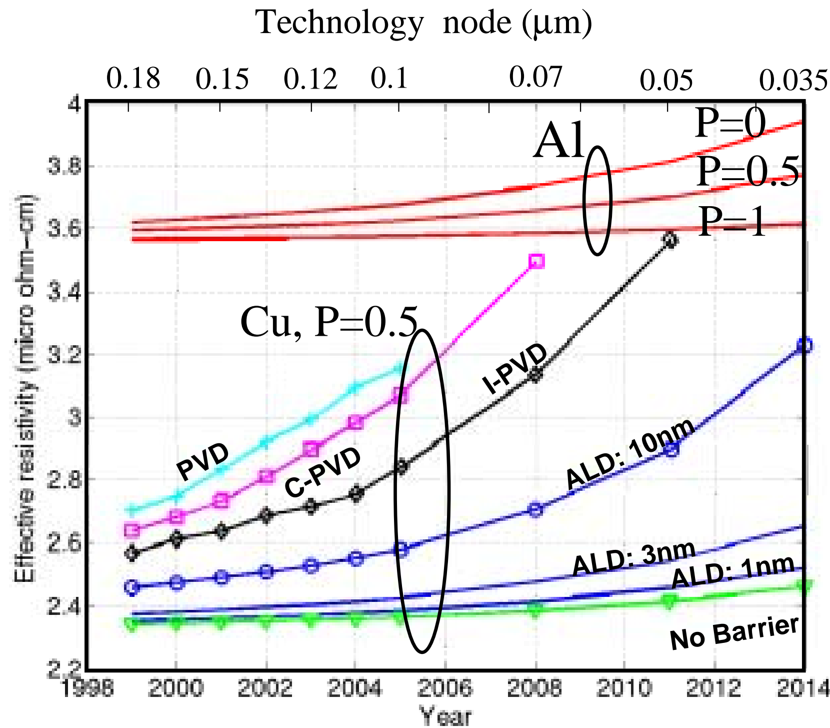


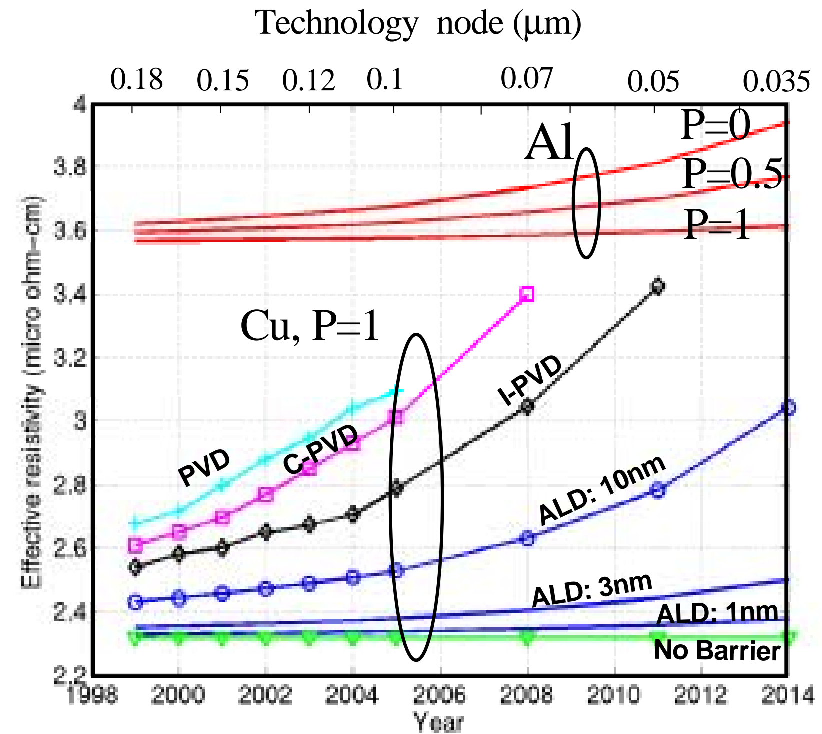
Most recently 1 and 3nm ALD barrier simulations

ALD most conformal => least barrier area => least resistivity

Cu Resistivity: Effect of Barrier Technology

Global Wires, Temp.= 100^0C, P = 0.5, BT=10nm->1nm

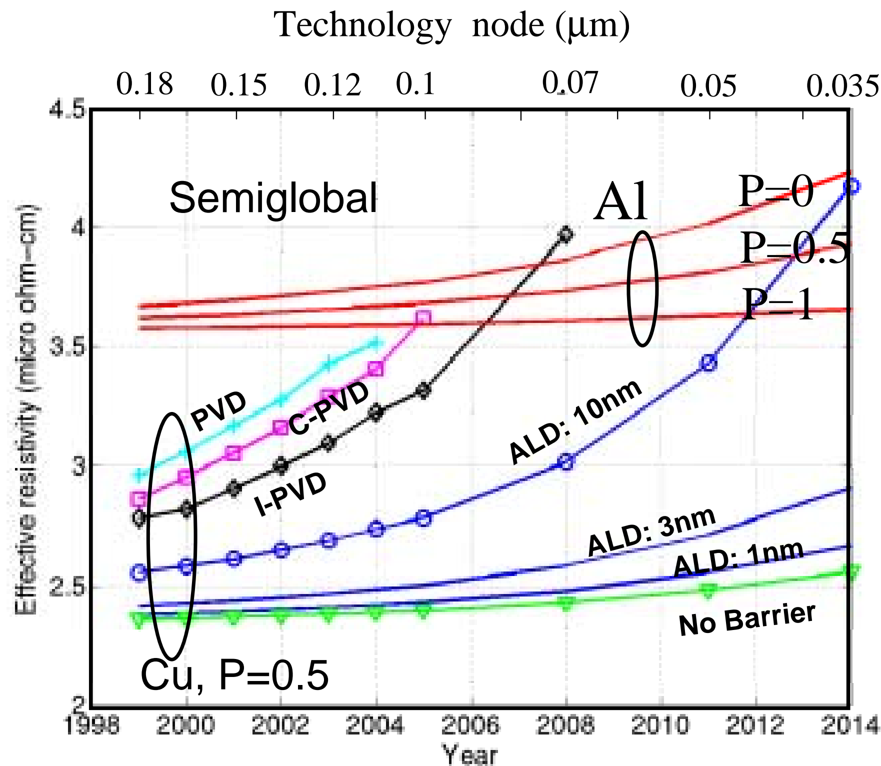


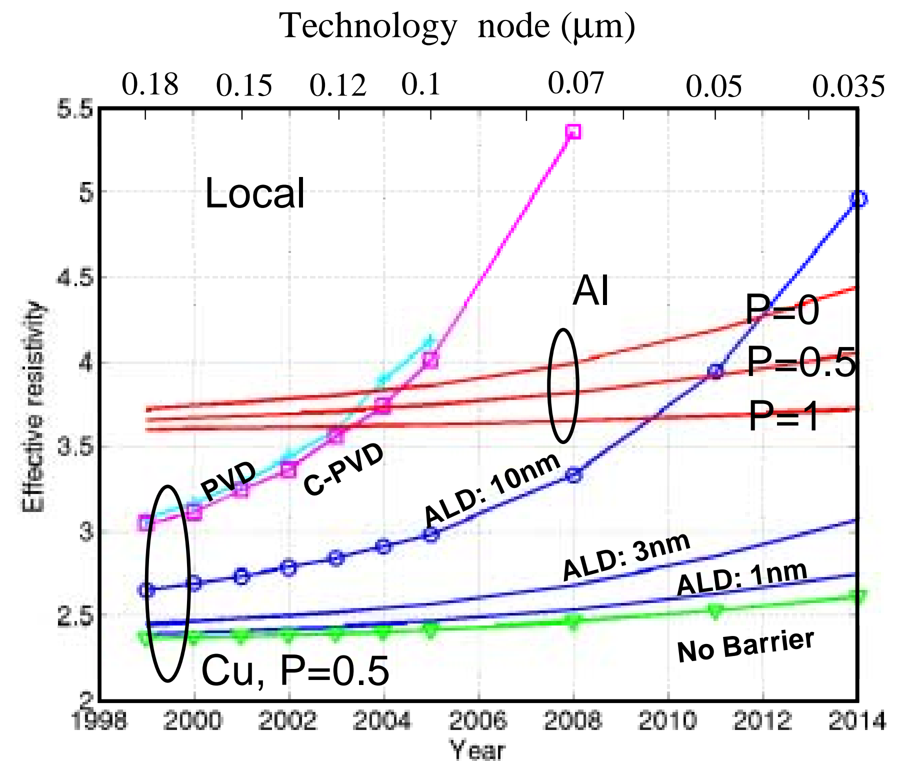


- ALD least resistivity rise
- ALD (10nm) and reasonable P = 0.5, resistivity = 3.2 µΩ-cm at 35nm
- **3nm ALD: 2.7 and 1nm ALD: 2.5 µΩ-cm**
- Al resistivity rises slower than Cu. Cross over with Cu resistivity possible
- Increasing P, reduces resistivity only slightly

Semi-global & Local Interconnects

Temp.=100 ^{0}C, P=0.5, Barrier thickn. 10 nm->1nm





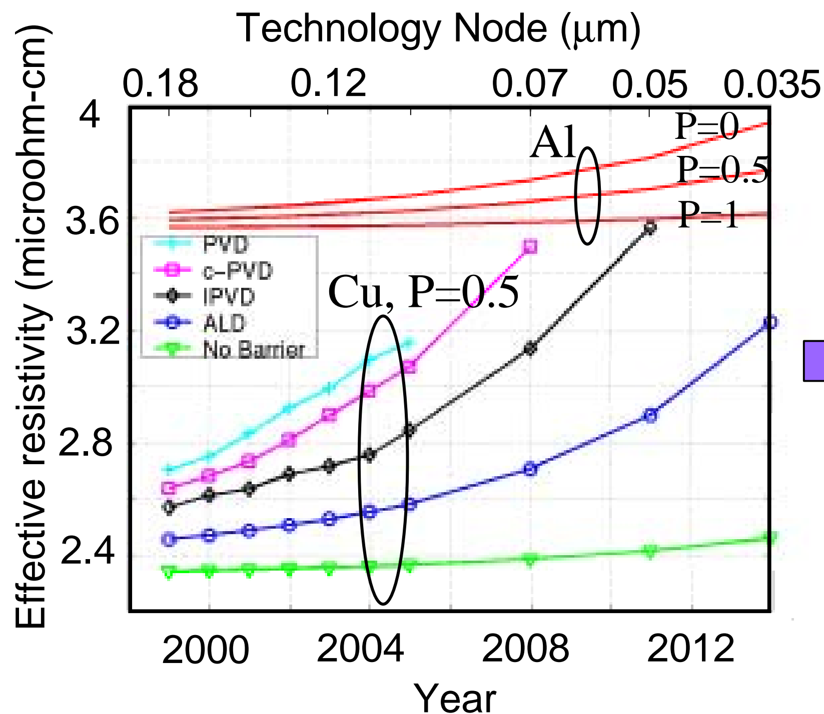
- **Resistivity rises faster for local**
- **Cu exceeds Al resistivity**

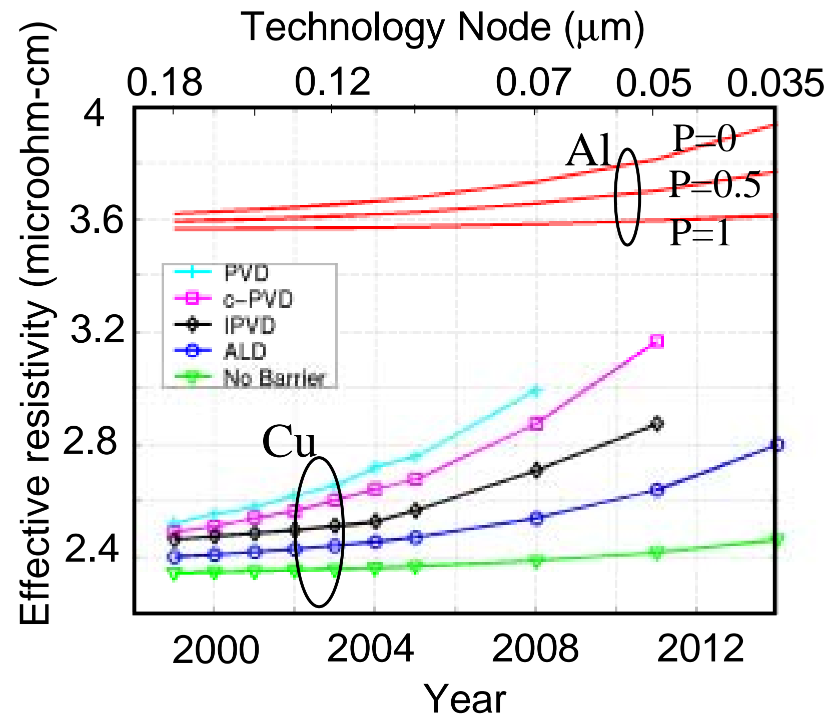
- **35 nm node: 10nm ALD 4.2 (semi-global), 5 µΩ-cm (local)**
- **3nm ALD: 2.9 (semi-global) 3.1µΩ-cm (local) at 35nm node**

Big advantage with ALD 3nm or less !

Effect of Barrier Thickness: Global Wires

10 nm P=0.5, T=100 ^{0}C **5 nm**

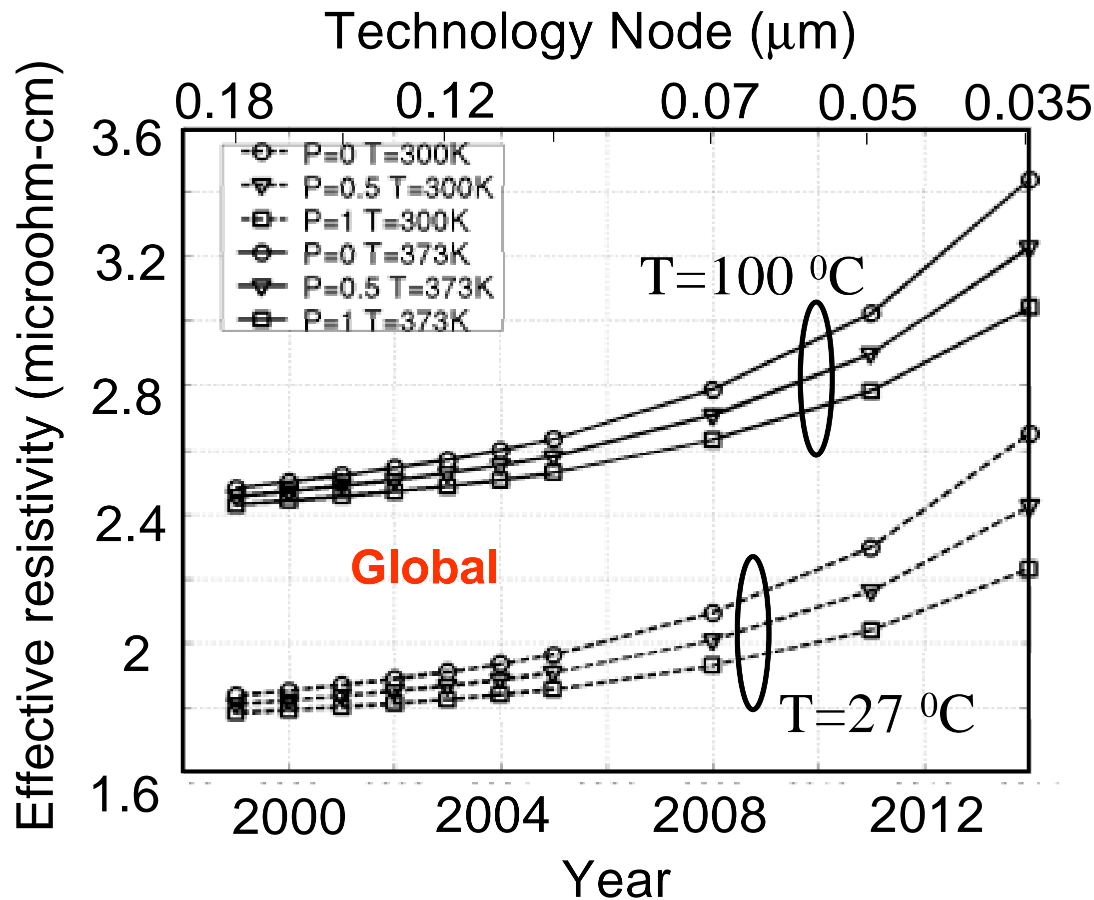




- Resistivity rises much faster with 10 nm

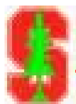
➢ **A barrierless Cu technology is desirable**

Cu Resistivity: Effect of Chip Temperature and P



- **Higher temperature ⇒ lower mobility ⇒ higher resistivity**
- **Realistic Values at 35nm node: P=0.5, temp=100 ⁰C**
  - **- local ~ 5 μΩ-cm**
  - **- semi-global ~ 4.2 μΩ-cm**
  - **- global ~ 3.2 μΩ-cm**

➢ **Low power circuits and better packaging technology needed**



Summary of resistance per unit length at 35nm node

Practical Constraint	Global Resist. (Ω/mm)	Semi-global Resist. (Ω/mm)	Local Resist. (Ω/mm)
	35nm node	35nm node	35nm node
None: ideal ρ=1.7μΩ-cm	628	1773	3275
P=0.5, BT=10nm	1192 **(190%)**	4351 **(245%)**	9564 **(292%)**
P=1, BT=10nm	1123 **(179%)**	3942 **(222%)**	8490 **(259%)**
P=0.5, BT=0	908 **(145%)**	2668 **(151%)**	5030 **(154%)**

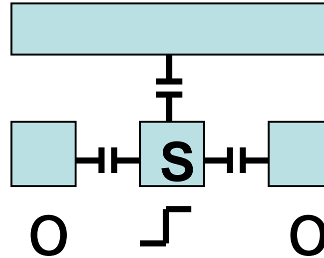
- Realistic Cu resistivity with technology constraints is much higher than the bulk value
- With 1 to 3nm ALD Barrier: significant reduction in resistivity

Interconnect Performance: In Light of Technology Effects

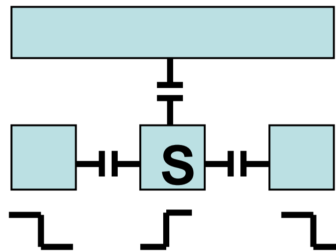
Delay: Nominal vs. Worst Case

Depends on switching condition on adjacent wires

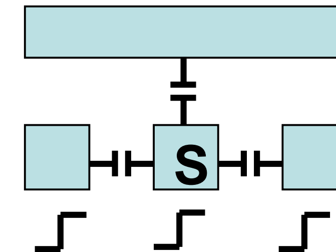
- Nominal
  $C_{inttot}=C_{IMD}+C_{ILD}$



- Worst Case
  $C_{inttot}=2C_{IMD}+C_{ILD}$



- Best Case
  $C_{inttot}=C_{ILD}$

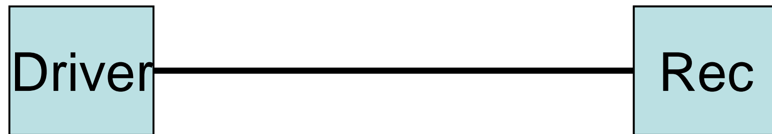


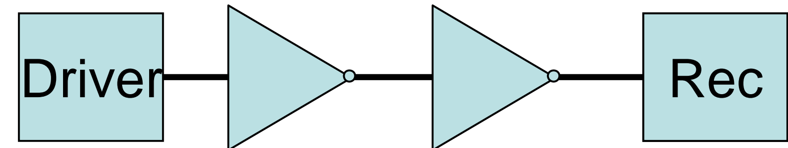
Not only total capacitance plays a role in **delay**, IMD plays a very import. Role

C_{IMD} ~ 70% of C_{inttot}

Cu Interconnect Delay: With and Without Repeaters

Repeaters Reduce delay enormously for long global link





A long global link w/o Repeaters

$t_{total}=0.4R_wC_wl^2$

With Repeaters

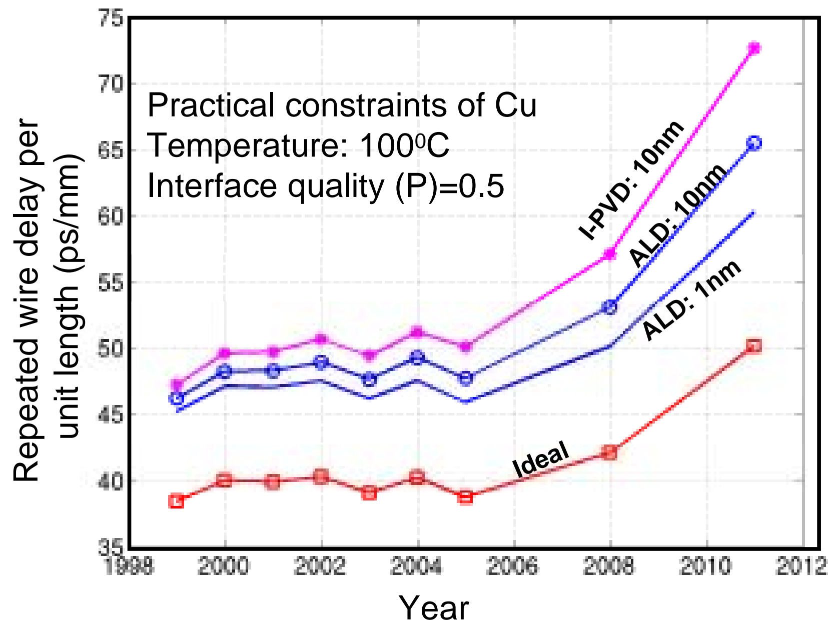
$$t_{total}=5l\sqrt{r_oC_{nmos}R_wC_w}=2l\sqrt{(0.4R_wC_w)(t_{FO4})}$$

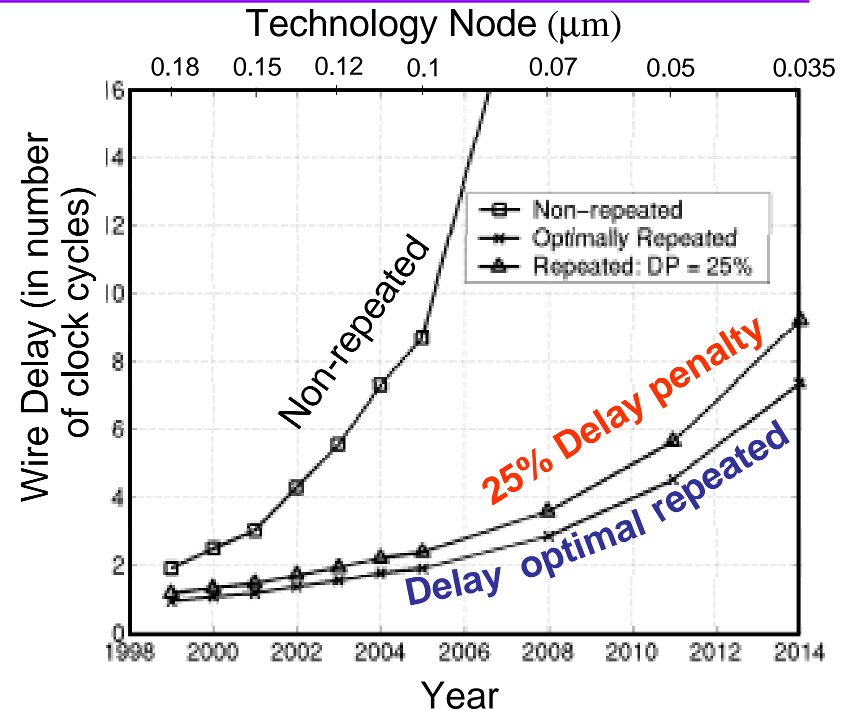
$$l_{opt}=3.24\sqrt{\frac{r_oC_{nmos}}{R_wC_w}}$$

$$S_{opt}=0.58\sqrt{\frac{r_oC_w}{R_wC_{nmos}}}$$

- Repeaters give best possible interconnect delay
- Delay linear with length (quadratic without them)
- Delay scales much better
  - only sqrt depend. on deteriorating R_w
  - dependence on t_{FO4}
- But have power and area penalty

Signaling Wire Delay Modeling With Repeaters





- ALD Barrier likely to be used in the future
  - 66 and 93ps/mm at 50 and 35nm, resp.
  - 30% more than with ideal Cu ρ at 50nm node

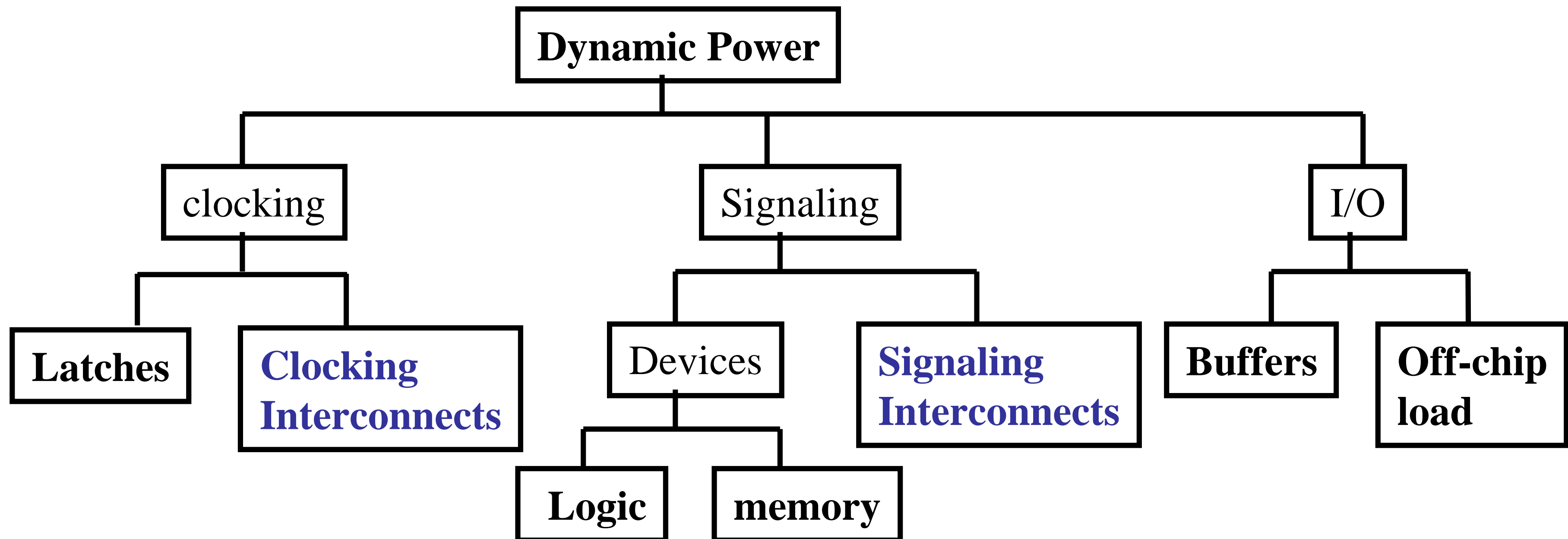
Also have Power and Area penalties
Pushing bottleneck to power!

- **Even with repeaters, 7.5X Clock at 35nm node 8X increase compared to 180nm node**
  - 3X from clock speed
  - 1.85X from delay per mm
  - 1.45X from length increase
- **Worst case delay**
  - 11 times clock period at 35 nm

Chip Power Breakdown & Future Power Problems

- **Dynamic Power: αCV^2f**
- **Leakage power: devices**
- Short circuit power during switching
- Analog components: static power

- **Interconnect power**
  - C_{int}: dissipated in devices
  - R_{int}: Joule heating (makes things worse)

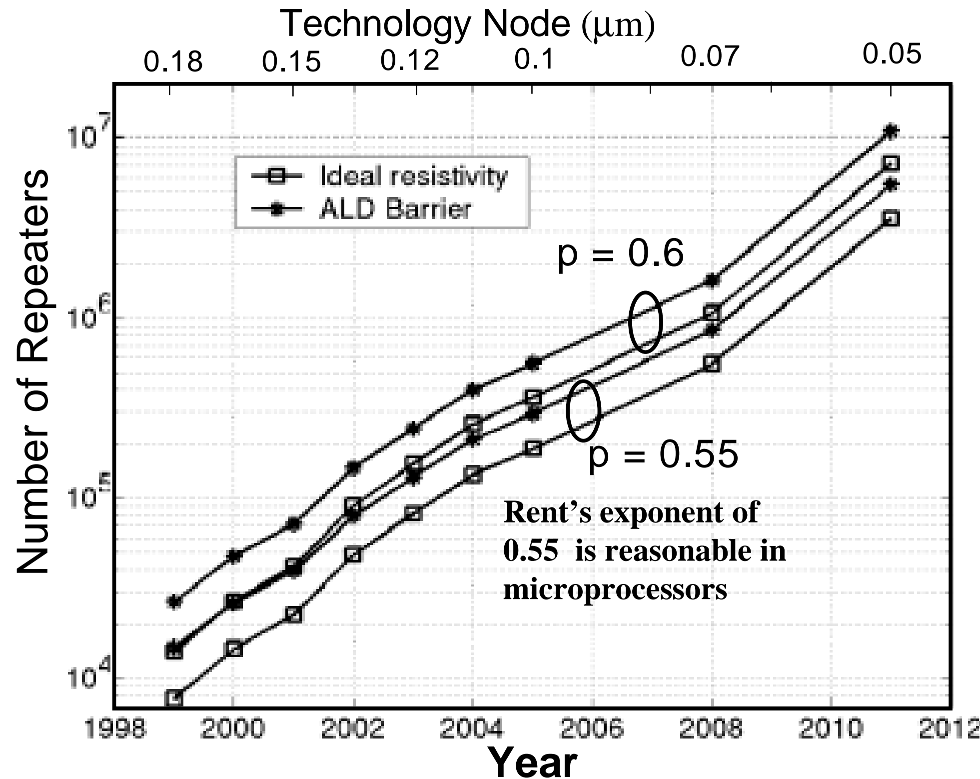


- Transistor leakage and capacitance
  - 20 Watts/cm^2 -> 200 watts/cm^2
- Interconnect and clock power further adds to this problem

- Clock frequency limited not by delay but by power?
  - Clock frequency ~ 16FO4 delays
  - $(CV^2f+P_{static}) < (\Delta T)(\text{Area/thermal resistance})$

V. Swerdlov et. al., *IEEE Intern. SOI Conf.*, 2001

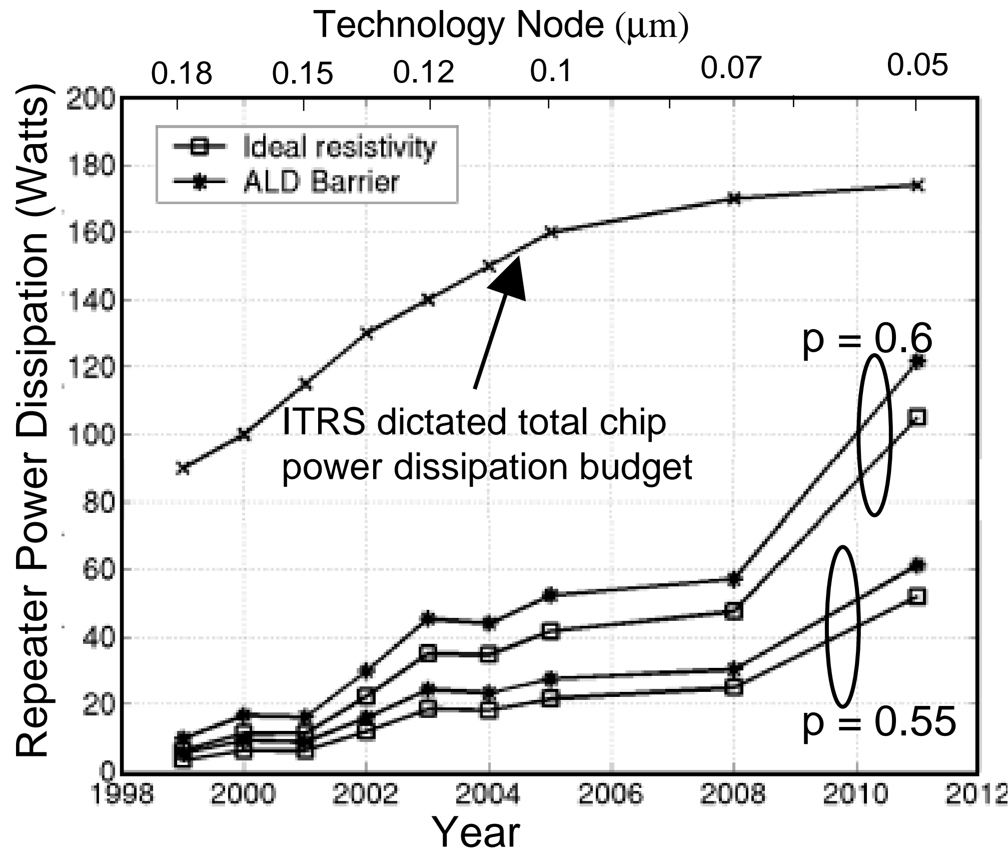
Number of Repeaters Required



- **ITRS wire dimensions: justified based on barely enough metal levels to fit the wires**
- **Separation of memory and logic area because different wire length distributions**
- **Rent's rule based distribution for logic area**

➢ **Additional power will be consumed by repeaters**

Global Signaling Wire: Repeater Power Penalty



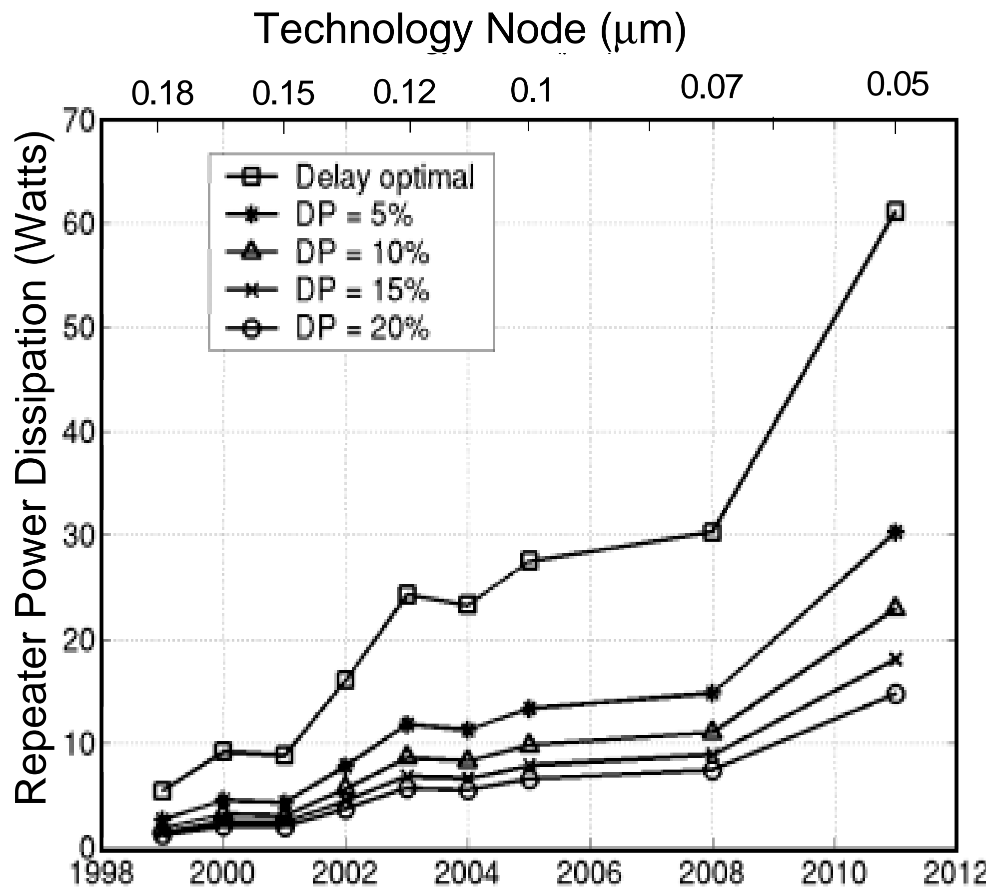
Exorbitant power signaling wires at future nodes (50nm)

- **Global Wires= 60 Watts (p=0.55)**
- **Repeaters = 60 Watts (p=0.55)**
- **120W for just global signaling wires**

Delay optimal repeaters ~ double power consumption of the wire

- **Global wire power same as above**

Global Signaling Wire: Repeater Power minimization With Delay Tradeoff



- Tolerable delay penalty depends on architecture
- Still 20W of power dissipation due to repeaters at 50nm node
- With about 20% more delay power dissipation by global wires with repeaters on them is now ~ **60+20=80W** at 50nm node

An Interesting Point about Interconnect Performance

Are inhomogeneous dielectrics better than homogeneous low-k?

- Cross talk better

- Delay could be better
  - Yes total cap would be lower with homogeneous but… improvement small cuz ILD is small fraction of total

  - Heat dissipation would be poorer and rise in resistance due to a higher temperature could offset above cap. advantage

Summary:Signaling Metal Wire Performance with Scaling

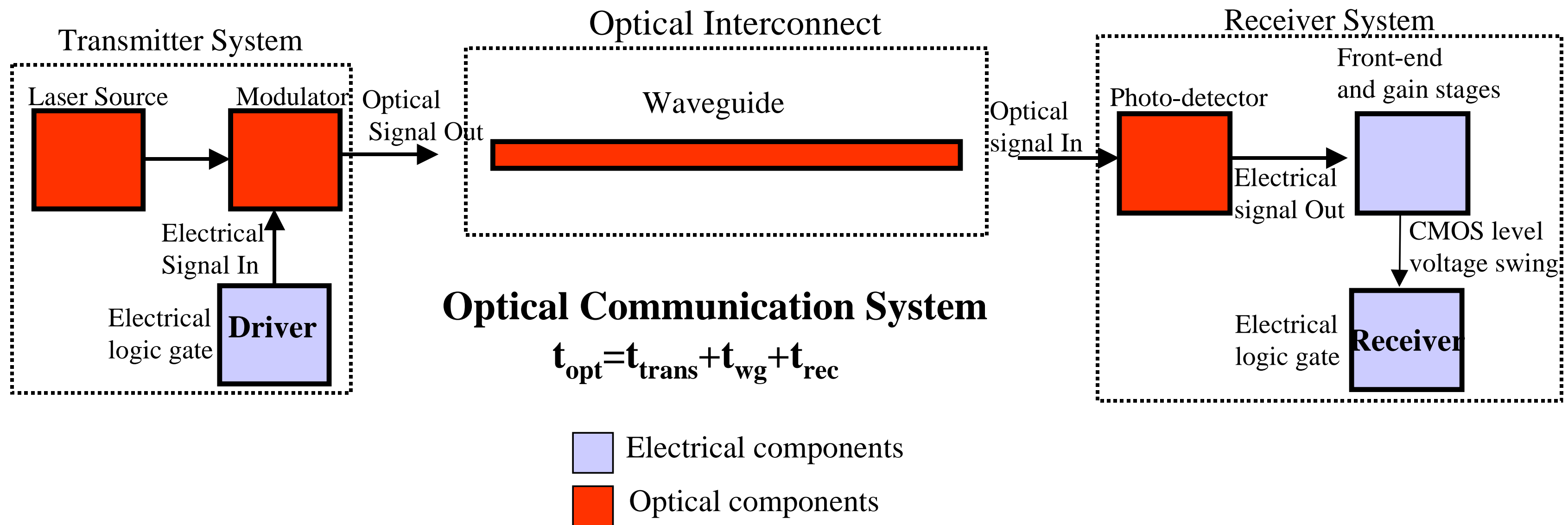
- Latency of metal based interconnects rises
- Power also rises
- Niche for Other Technologies?

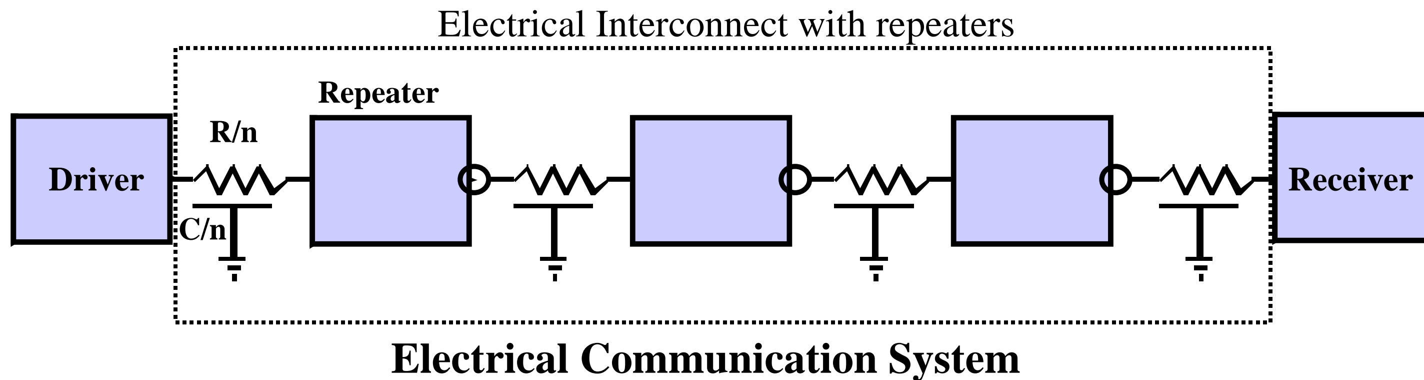
$\Rightarrow$ **Can we do better delay and power with optics (we will see)**

Long-term Alternatives: Optics?

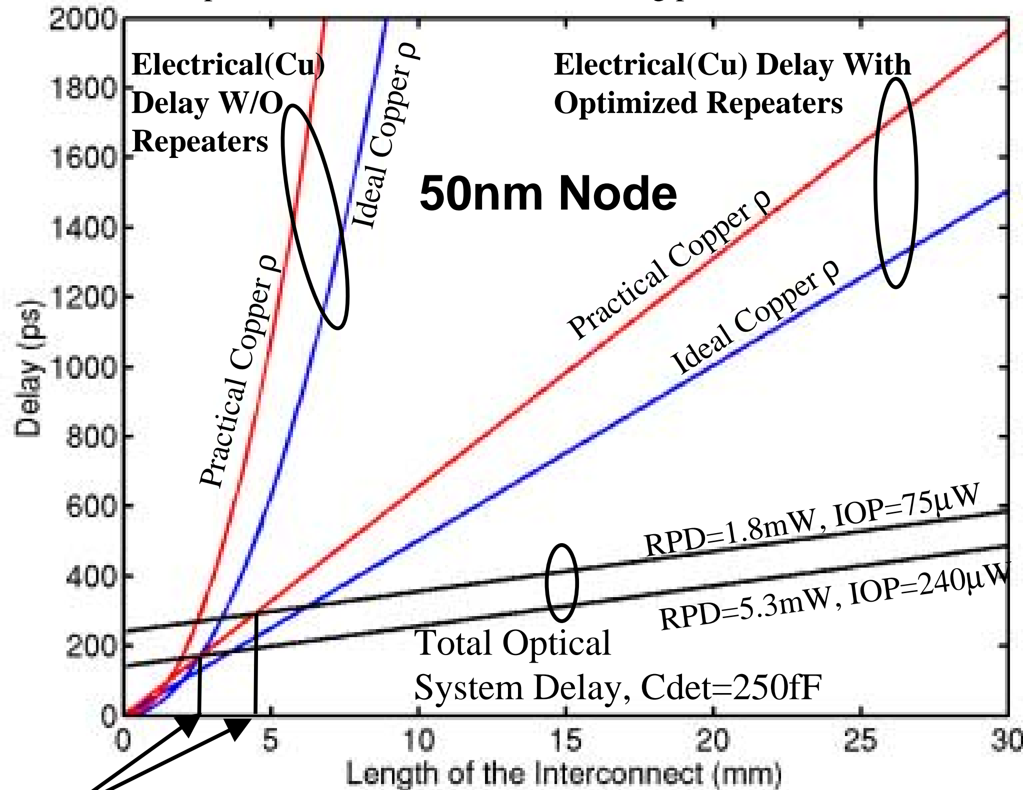
- **Signal wires:**
  - ✓**Reduce delay?**
  - ✓ **Power?**
- **Clock distribution**
  - ✓**Reduce jitter?**
  - ✓**Reduce skew?**
  - ✓**Reduce clock distribution power (50-60% of total power on chip)**

Signaling Application





Optical Vs. Electrical Wires: Delay

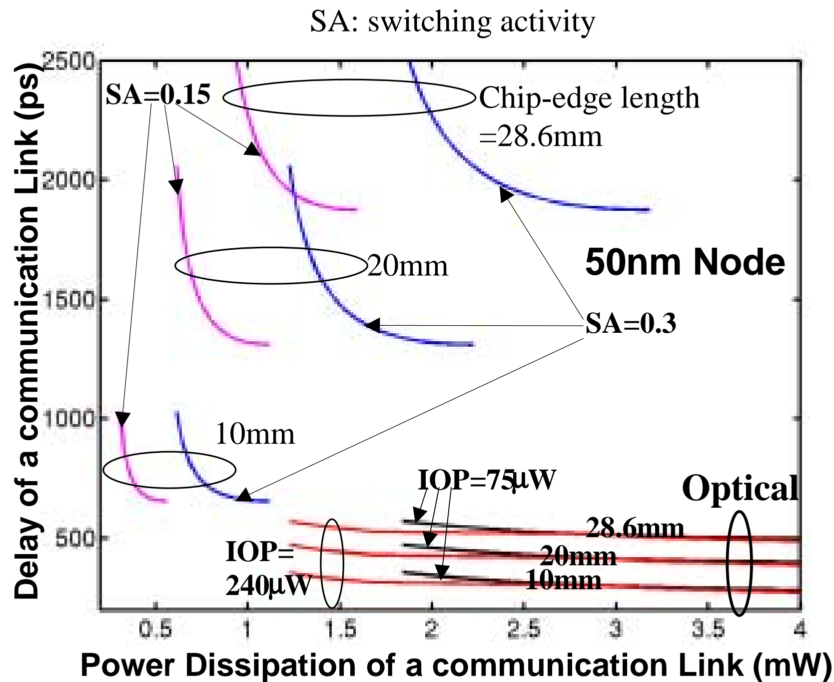


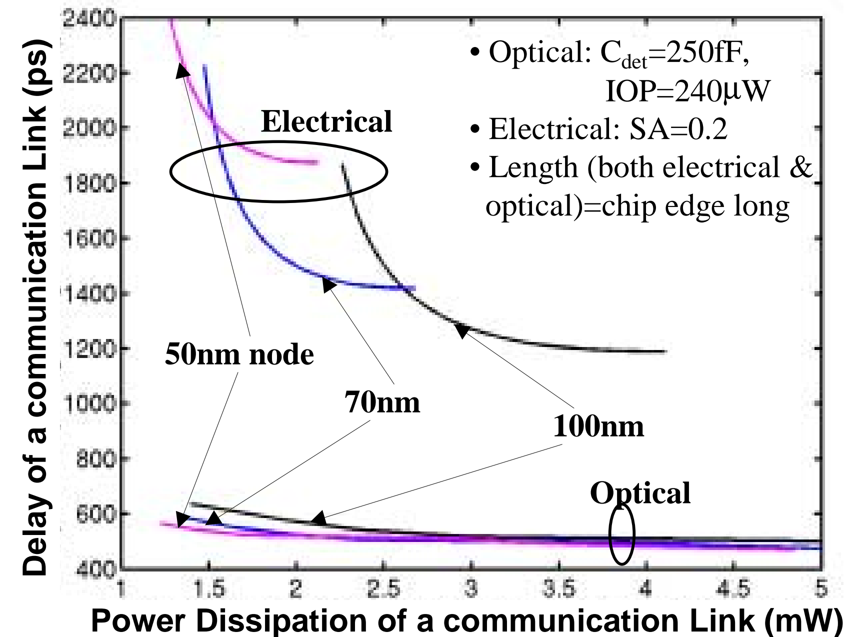
Critical length above which optical System is faster than even the electrical (Cu) repeated wires

- Optical Interconnects are faster than repeated wires beyond a length well within chip size
- However for Signaling both delay and power are important
- 1.8 mW is approximately power dissipated by a repeated chip edge long wire

Optical Vs. Electrical Wires: Delay & Power

SA: switching activity

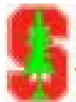




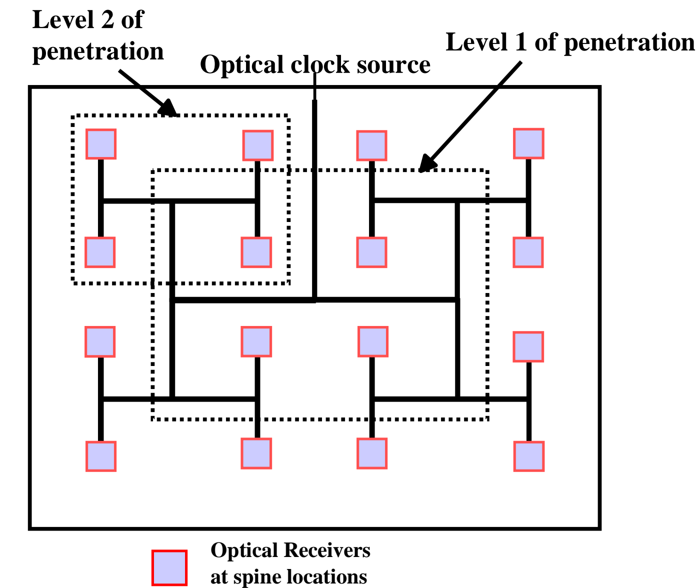
- **Longer lengths:** optics both power and delay advantage
- **Shorter lengths**: diminishing delay advantage and power disadvantage

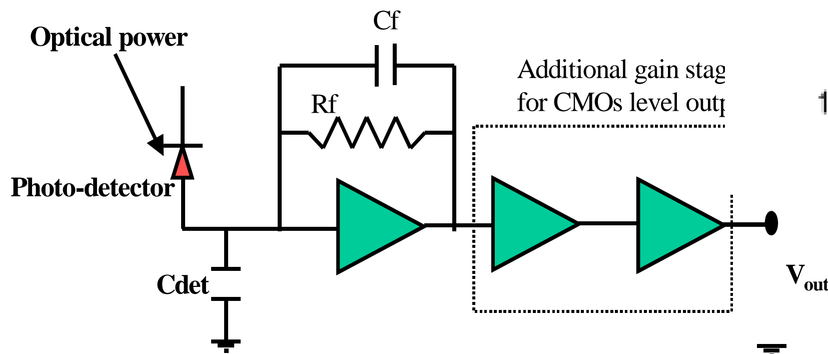
- With tech. node power advantage diminishes but delay advantage increases
- Still good for long global wires whose number is not large

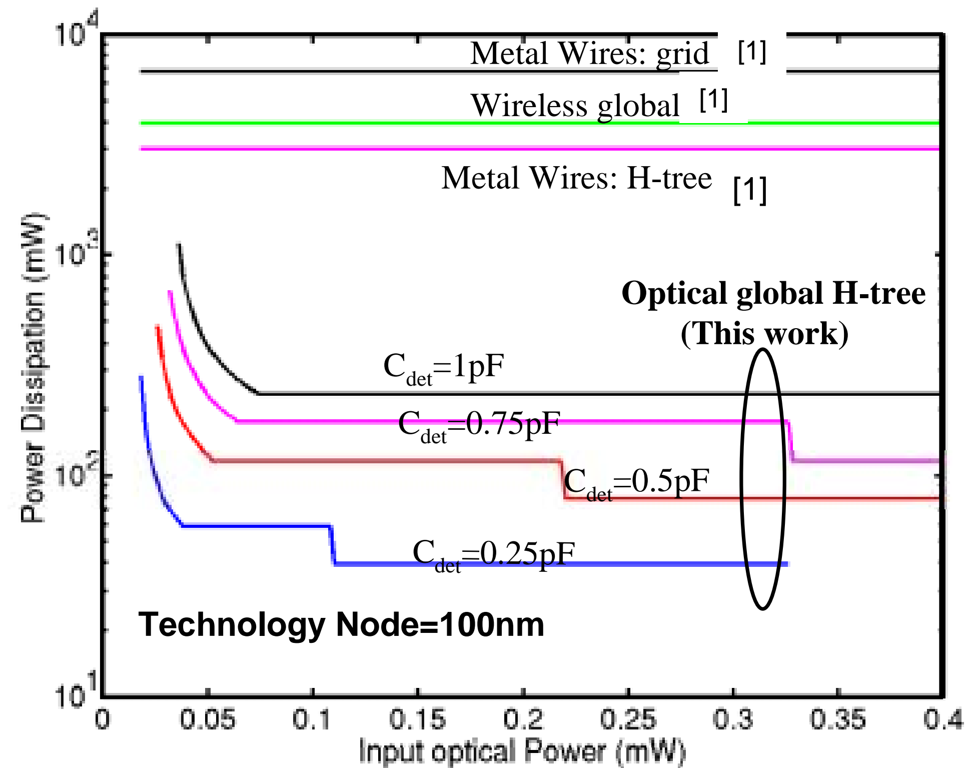
Alternate architecture using wires more efficiently (higher SA) can give huge power as well as delay advantages with optics



Clock Application: Incremental Approach







Lower Detector Capacitance and higher IOP for low Receiver power Dissipation

[1] Floyd et al, *Proc.IITC*, pp. 248-251, 1999.

Summary

- **Conventional Interconnects: Challenges and Limitations**
  - Realistic resistance modeling at future nodes
    - Barrier & surface scattering effects vital in dictating Cu effective resistivity
    - Cu effective ρ rises dramatically at all tiers: technology effects
      - ALD 3nm or less helps alleviate some problems but only near-term
    - A barrierless tech. as well as low temperature very beneficial
  - Realistic Interconnect Delay modeling in future
    - Delay rises significantly compared to clock period even with repeaters
  - Interconnect Power also rises in future
    - Delay optimized repeaters double the wire power
- **Future Recommendations and identification of key technological concentration**
  - Need for barrierless technology, new ultra cooling mechanism (lower wire temperature) and interface technology yielding P values close to 1
  - Optical Interconnects promising for longer links: Delay and Power