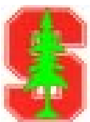
Scaling Induced Performance Limitations of Metal Interconnects

Prof. Krishna Saraswat
and
Students: P. Kapur, G. Chandra, T.-Y. Chiang, S. Souri

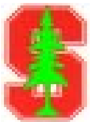
Department of Electrical Engineering
Stanford University
Stanford, CA 94305
saraswat@stanford.edu

Funding sources: DARPA, MARCO



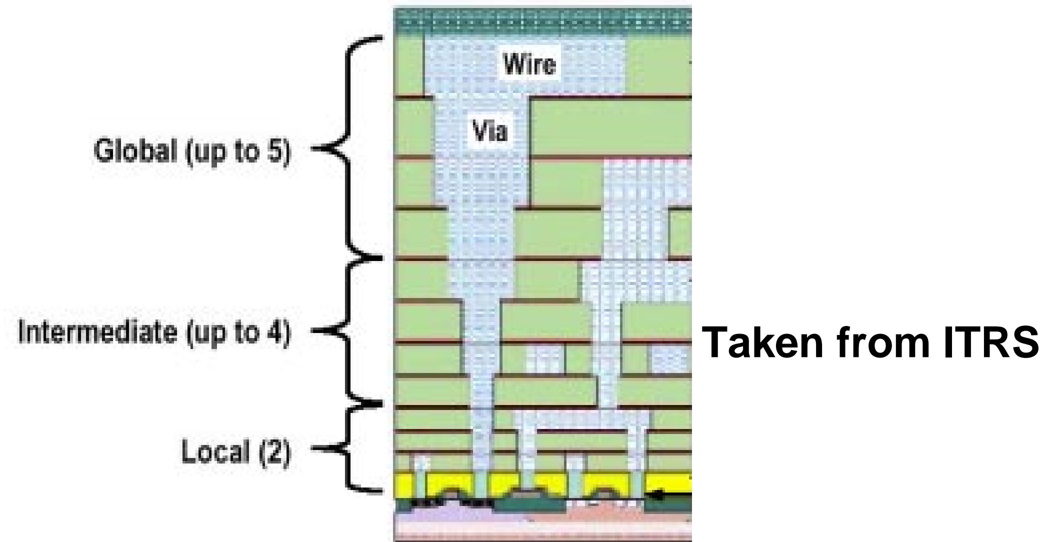
Outline

- Realistic metal resistivity modeling with technology constraints
  - Cu diffusion Barrier
  - Electron Scattering
- Performance assessment with realistic parameters
  - Delay
  - Repeaters
  - Power
  - Comparison of Cu with Al
- Novel communication mechanisms
  - Optical interconnects
  - 3-D technologies
  - RF wireless interconnects



Introduction: Types of Interconnects and Metrics

- Dimension based
  - Local
  - Intermediate/semiglobal
  - Global
- Function based
  - Signaling
  - Clocking
  - Power/Gnd distribution



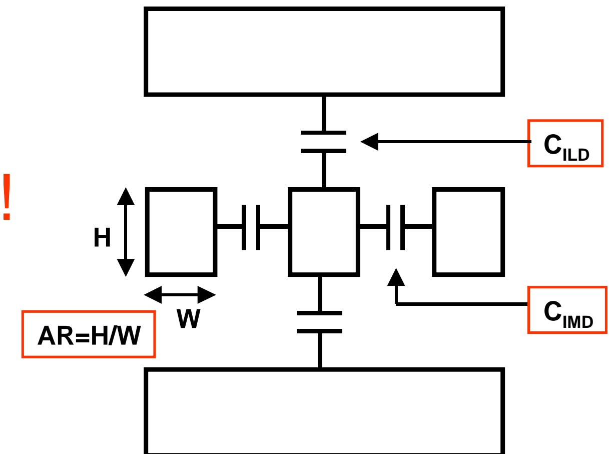
Performance

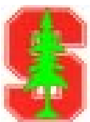
- Delay
- Power
- Bandwidth
- Area
- Self Heating
- Data Reliability (Noise)
  - Cross talk
  - ISI: impedance mismatch

Reliability

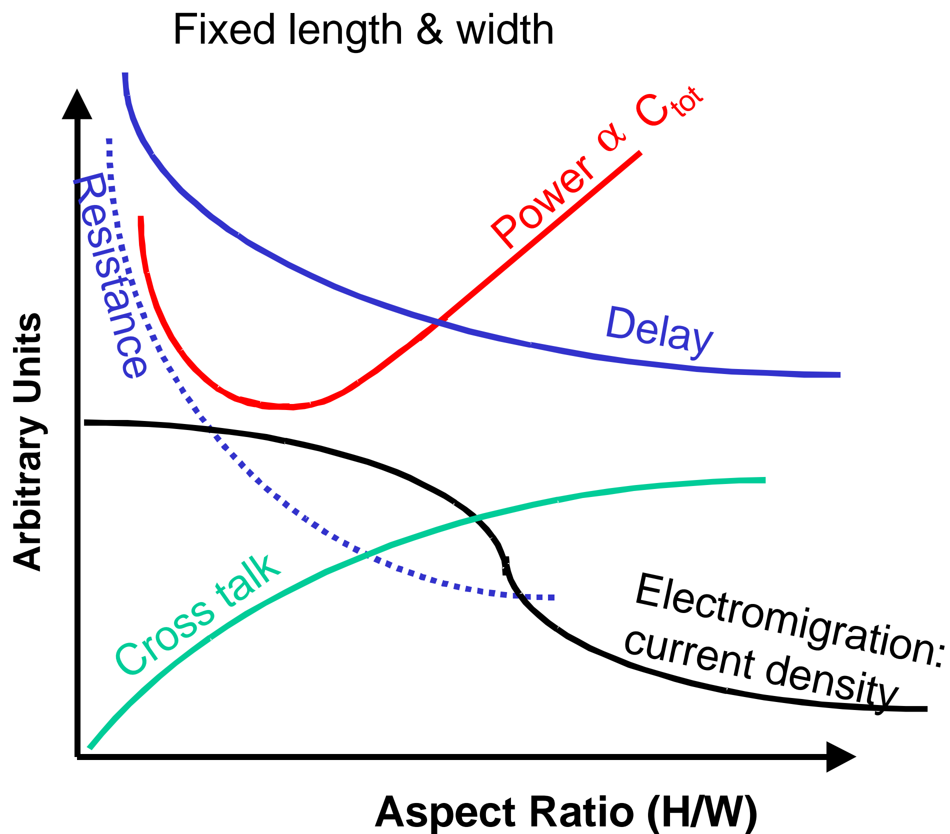
- Electromigration

Depend on R and C !





Motivation (I): Future Problems



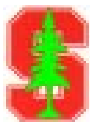
Aspect ratio increase (tradeoffs)=>

- **Better delay and electromigration**
- **Worse power and cross talk**

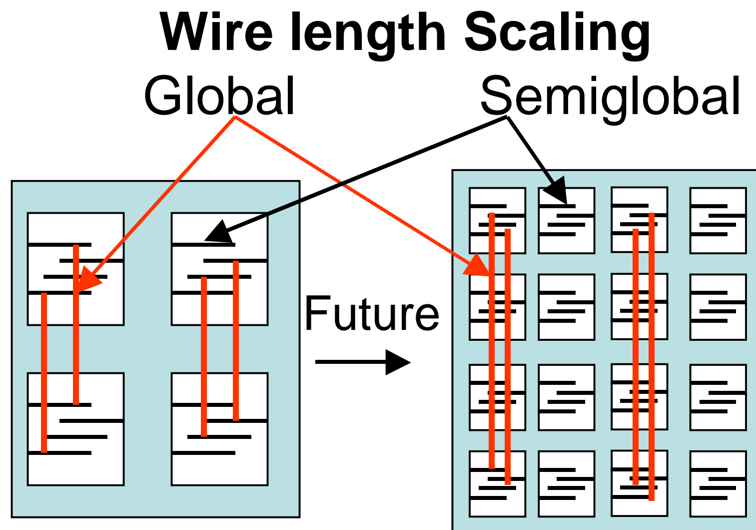
In future even with new materials

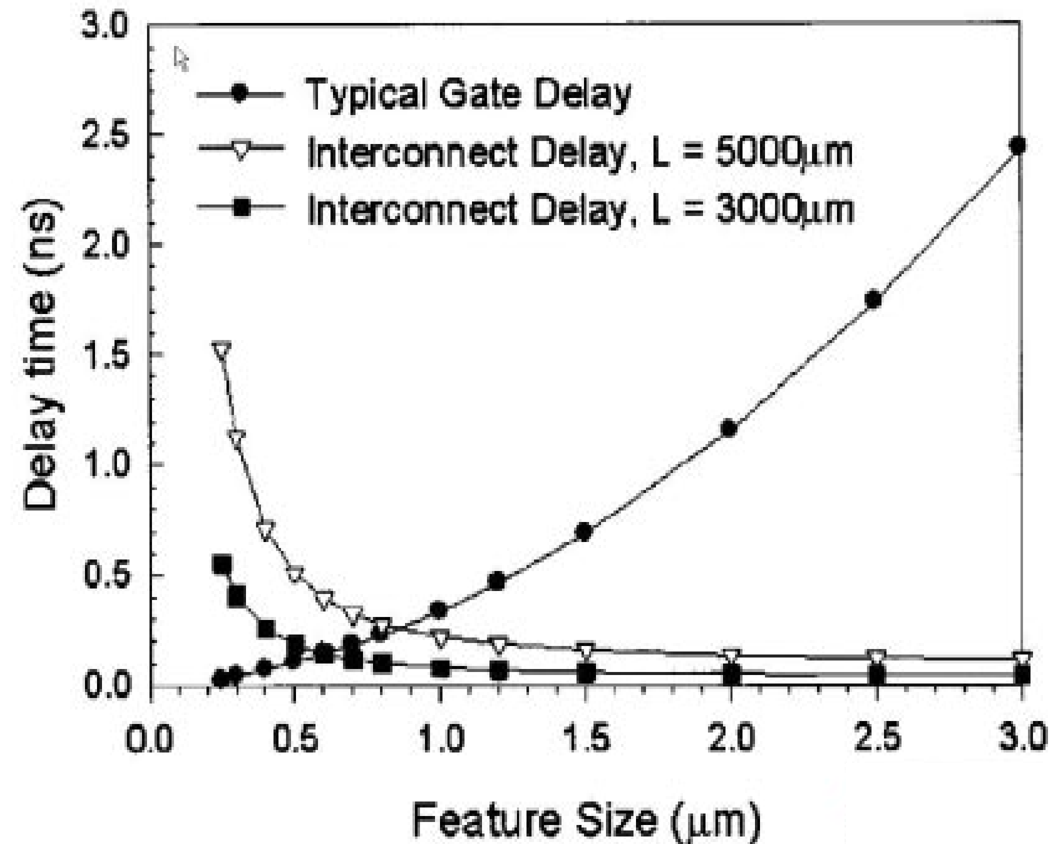
- Delay curve moves up
- Power curve moves down
  (total may go up)
- Cross talk curve same
  - If dielectric ratio is same
- Current density curve flatter

Shows design window complexity !



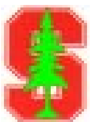
Motivation (II): Future Problems (Delay)



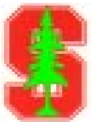


All types of signal wires delays are deteriorating wrt gate delay with scaling even with new low-k materials !

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

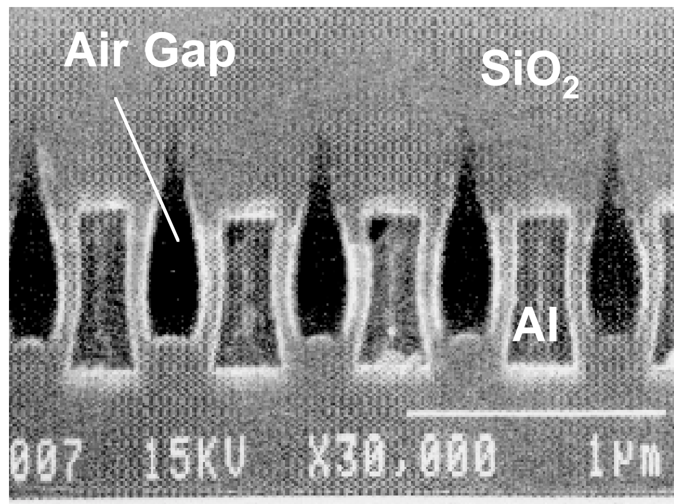


Will better materials like copper and low-k dielectrics solve the interconnect problem?



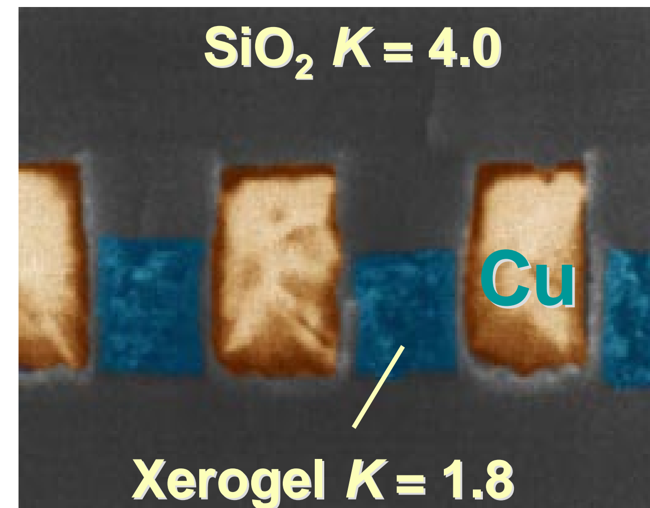
Limit of Low-k Dielectrics

Air-Gap/ Al

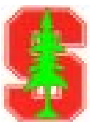


Stanford

Cu/xerogel

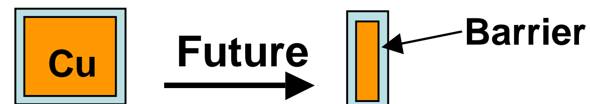


- **Old dielectric SiO_2 K = 4**
- **Polymers or air-gaps K = 2 - 3**
- **Ultimate limit is air with K = 1**

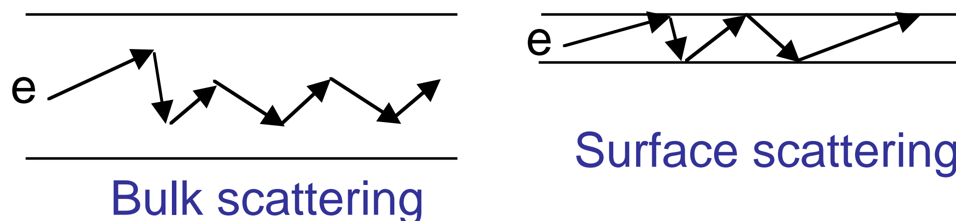


Cu Resistivity: Effect of Line Width Scaling

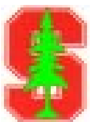
- Effect of Cu diffusion Barrier
  - Barriers have higher resistivity
  - Barriers can't be scaled below a minimum thickness
  - Consumes larger area as dimensions decrease



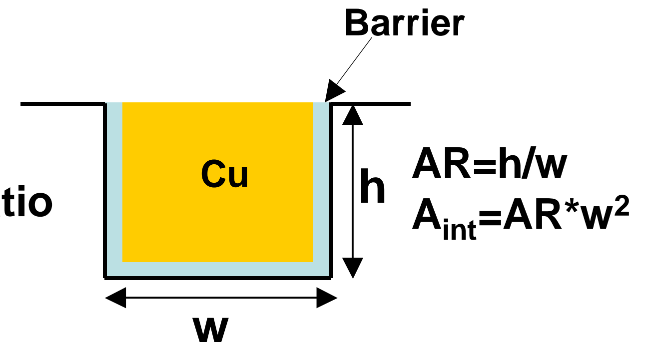
- Effect of Electron Scattering
  - Reduced mobility as dimensions decrease
  - Reduced mobility as chip temperature increases



➢ **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

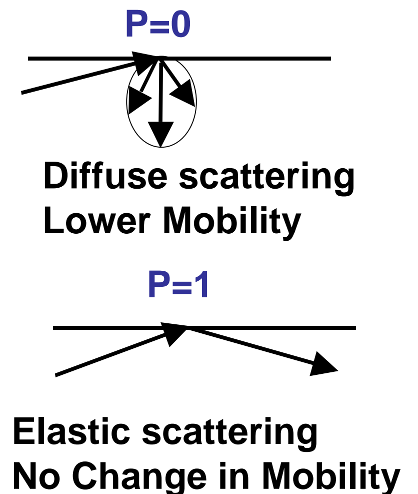
$$\frac{\rho_b}{\rho_o} = \frac{1}{1 - \frac{A_b}{AR * w^2}}$$

- **Important parameter: A_b to A_{int} ratio**
- **ρ_b increase with A_b to A_{int} ratio**
- **Future: ratio may increase**

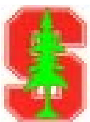
Electron Surface Scattering Effect

$$\frac{\rho_s}{\rho_o} = \frac{1}{1 - \frac{3(1-p)\lambda_{mfp}}{2d} \int_1^{\infty} \left(\frac{1}{T^3} - \frac{1}{T^5} \right) \frac{1 - e^{-kT}}{1 - pe^{-kT}} dT}$$

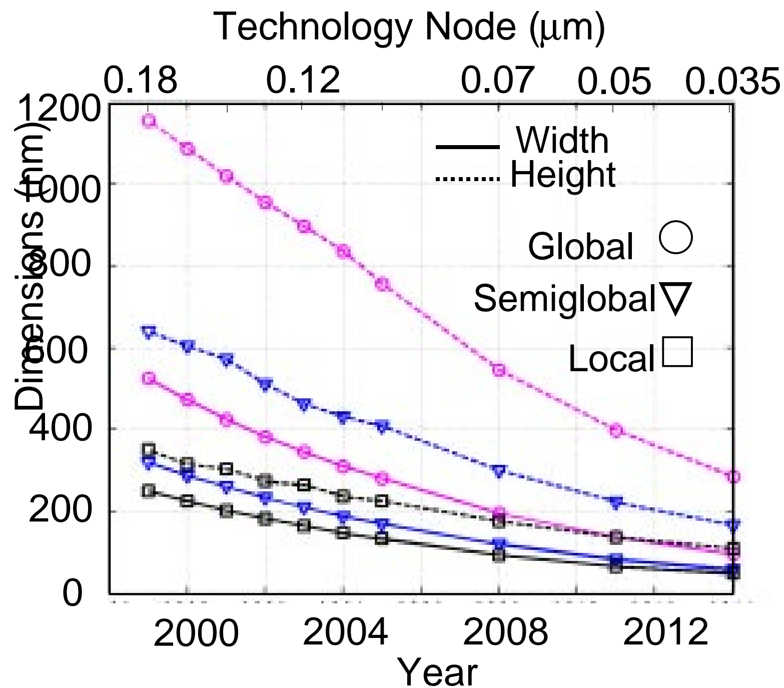
P: Fraction of electrons scattered elastically from the interface
k= d/ λ_{mfp}
λ_{mfp}: Bulk mean free path for electrons
d: Smallest dimension of the interconnect



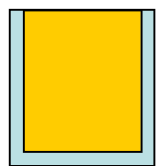
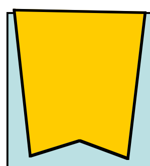
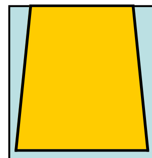
- **Reduced electron mobility**
  - **Operational temperature**
  - **Copper/barrier interface quality**
  - **Dimensions decrease in tiers: local, semiglobal, global**



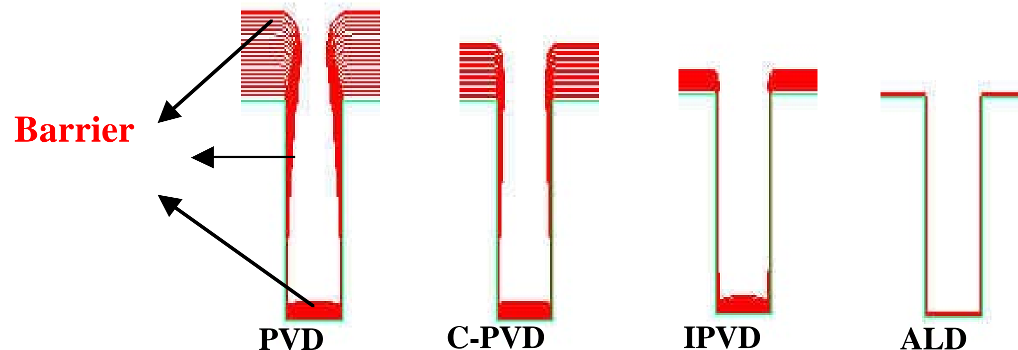
Methodology for Resistivity Calculations



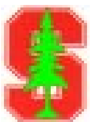
- SPEEDIE used to simulate barrier profiles
  - Different technologies
  - Different geometries: ITRS '99,
    - 180 nm to 35 nm technology node
    - Local, semi-global, global
  - Two barrier thicknesses: 5 and 10 nm
  - Surface scattering effect
    - P from 0 to 1 in step of 0.25
    - Temperature: 0ºC and 100ºC

ALD IPVD C-PVD

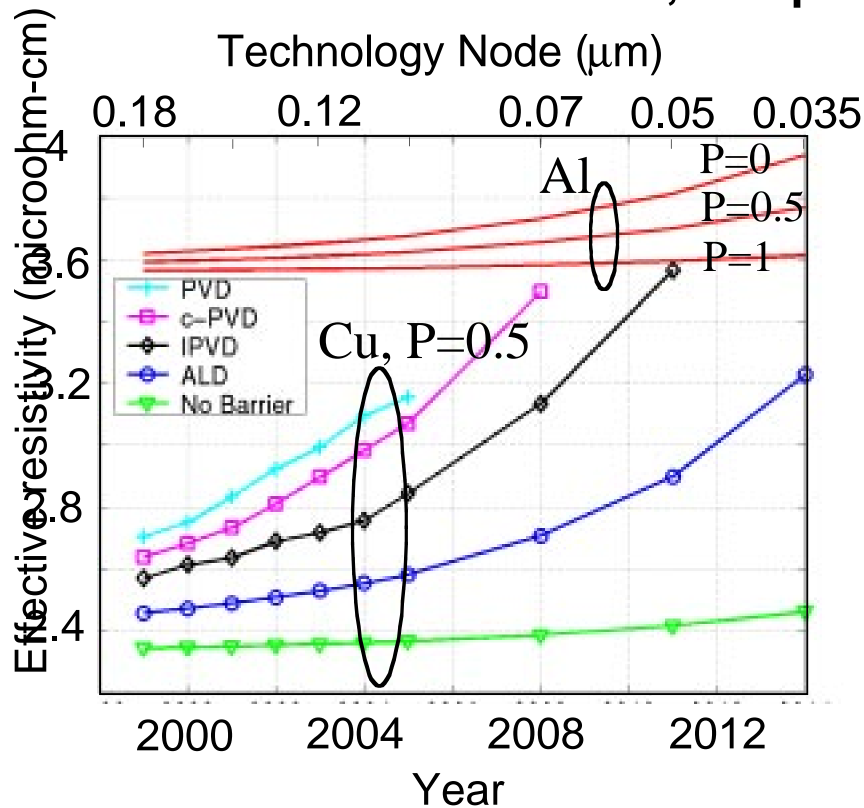


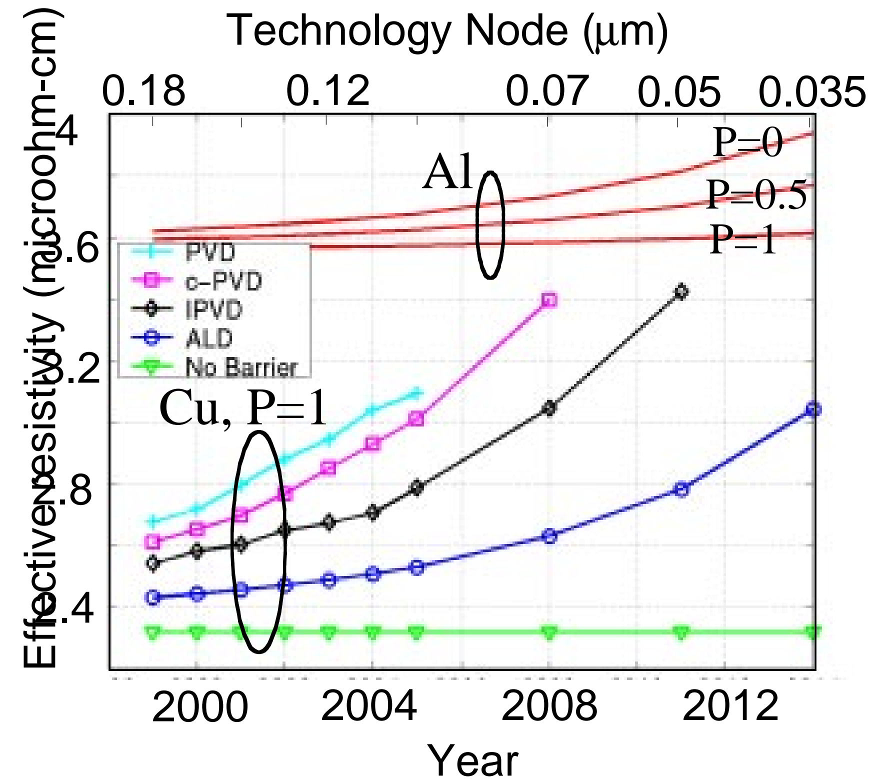
ALD most conformal => least barrier area => least resistivity



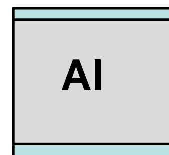
Cu Effective Resistivity: Effect of barrier deposition technology

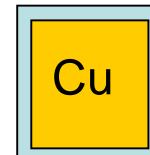
Global Wires, Temp.= 100⁰C, BT = 10nm, P = 0.5

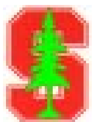




- With ALD least resistivity rise
- With best barrier (ALD) and reasonable P = 0.5, resistivity = 3.05 µΩ-cm in 2014
- Al resistivity rises slower than Cu. Cross over with Cu resistivity possible
  - no 4 sided barrier
  - smaller λ_{mfp} => smaller k



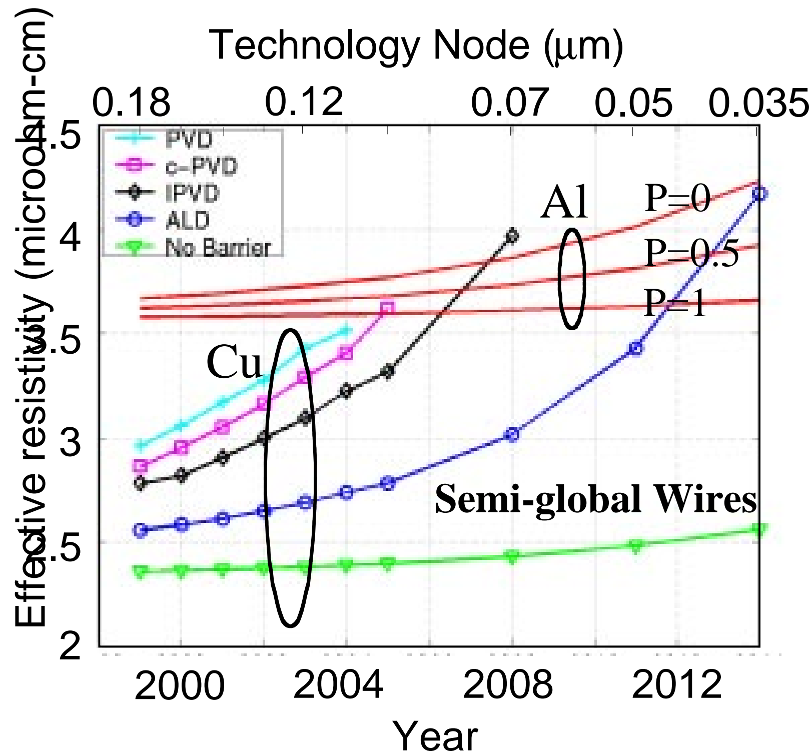


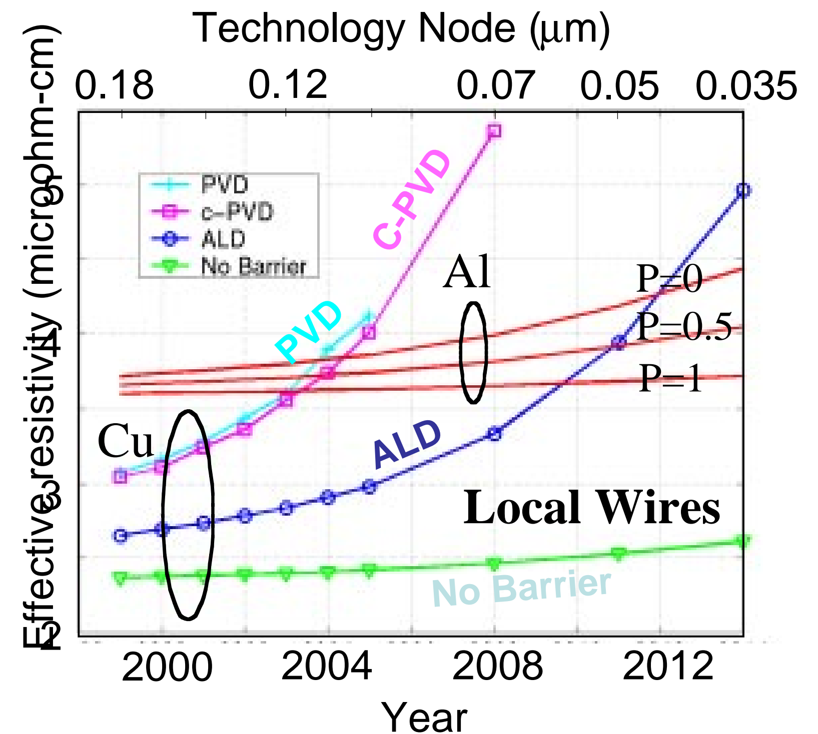


Semi-global & Local Interconnects

Effect of Barrier Deposition Technologies

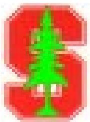
Temp.=100 ^{0}C, P=0.5, Barrier thickn. 10 nm





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

- **35 nm node: even with ALD resistivities=4.2 (semi-global) 5 µΩ-cm (local)**



Effect of Barrier Thickness: Global Wires

10 nm P=0.5, T=100 ⁰C **5 nm**

Technology Node (µm): 0.18, 0.12, 0.07, 0.05, 0.035

Effective resistivity (microohm-cm): 2.4, 2.8, 3.2, 3.6, 4

Year: 2000, 2004, 2008, 2012

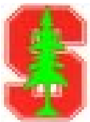
PVD, c-PVD, IPVD, ALD, No Barrier

Al: P=0, P=0.5, P=1

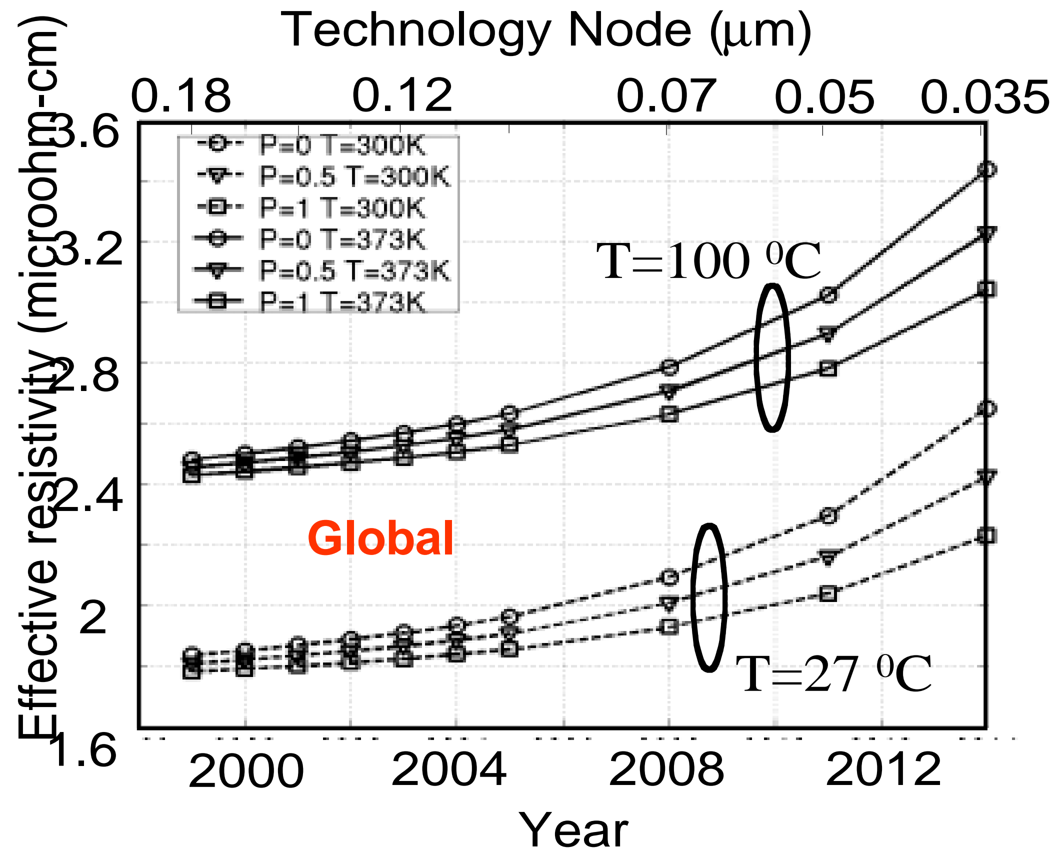
Cu, P=0.5

Cu

- Resistivity rises much faster with 10 nm
  - ➢**A barrierless Cu technology is desirable**

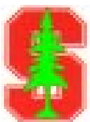


Cu Resistivity: Effect of Chip Temperature



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

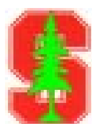
➢ Low power circuits and better packaging technology needed



Summary of resistance per unit length at 35 nm node

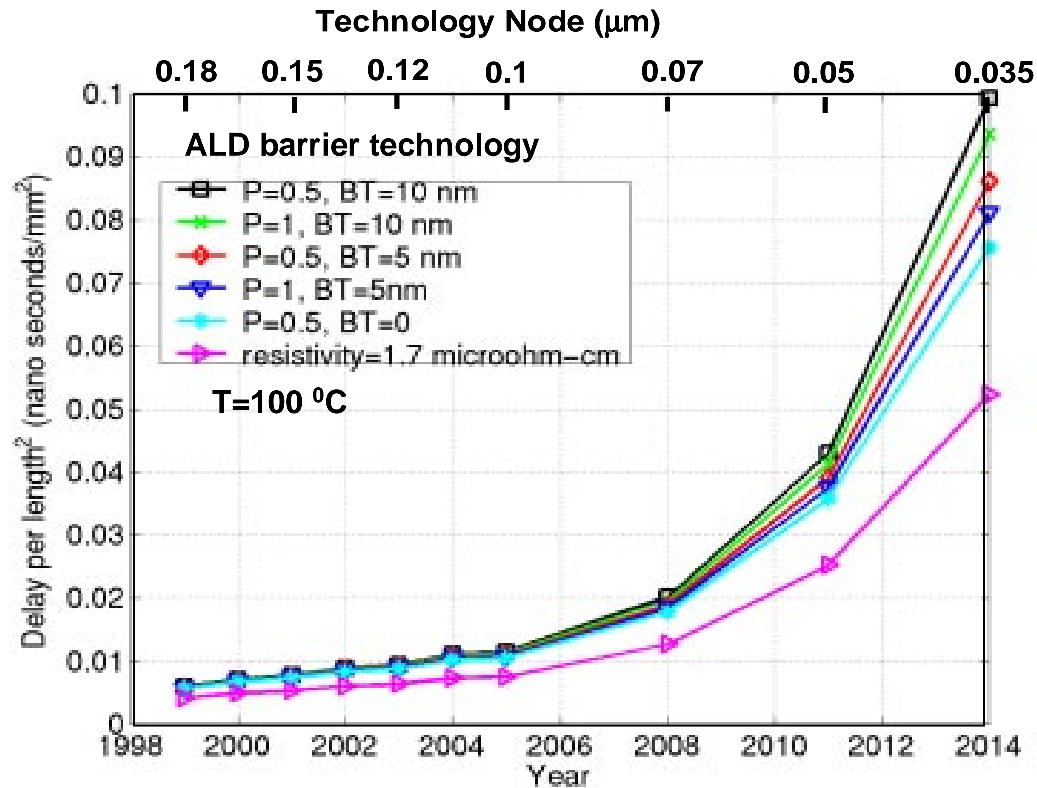
Practical Constraint	Global Resist. (Ω/mm)	Semi-global Resist. (Ω/mm)	Local Resist. (Ω/mm)
	Year 2014	Year 2014	Year 2014
None: ideal ρ=1.7$\mu\Omega$-cm	628	1773	3275
P=0.5, BT=10nm	1192 **(190%)**	4351 **(245%)**	9564 **(292%)**
P=1, BT=10nm	1123 **(179%)**	3942 **(222%)**	8490 **(259%)**
P=0.5, BT=0	908 **(145%)**	2668 **(151%)**	5030 **(154%)**

- Realistic Cu resistivity with technology constraints is much higher than the bulk value



Cu Interconnect Delay

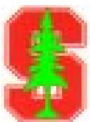
Global wire delay per unit length



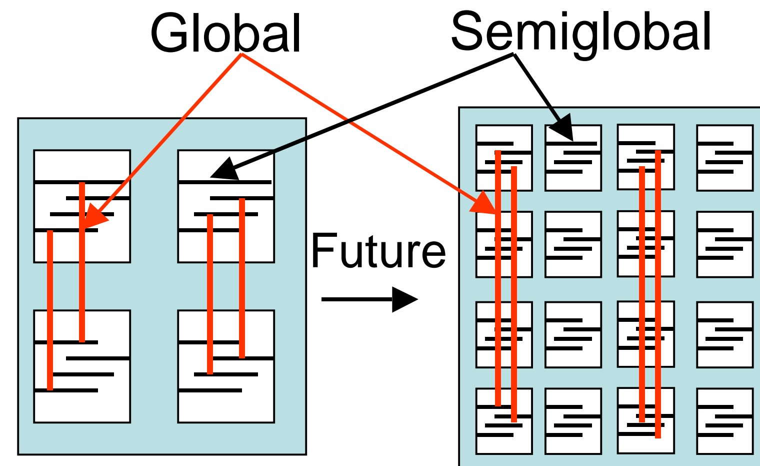
- Global Wires

Current (180nm)	Ideal (50 nm)	Most Realistic (50 nm)	Barr. Thick.=0, P=0.5
30 Ω/mm	310 Ω/mm	525 Ω/mm	400 Ω/mm

- Semi-global and local: much worse



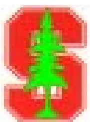
Delay of Signal Wires



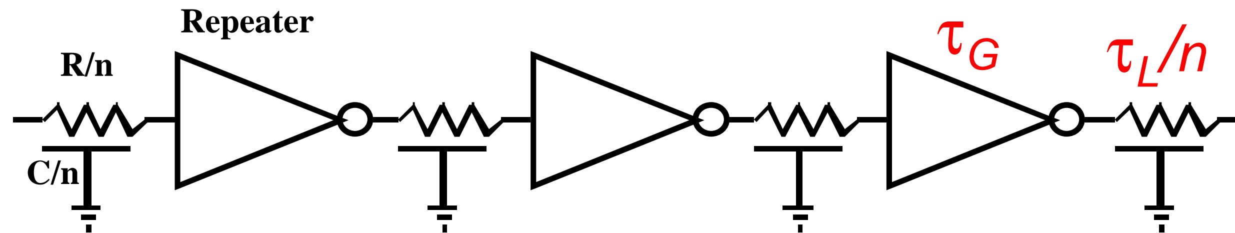
Three types of signal wires

- Wires whose length shrinks **(local)**
  - Local wire delay is also going up slowly wrt gate delay
- Wires whose length remains about the same **(semiglobal)**
  - Worse than local
- Wires whose length increase with scaling **(Global)**
  - Very bad

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

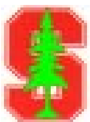


Can we solve the problem by using more repeaters?

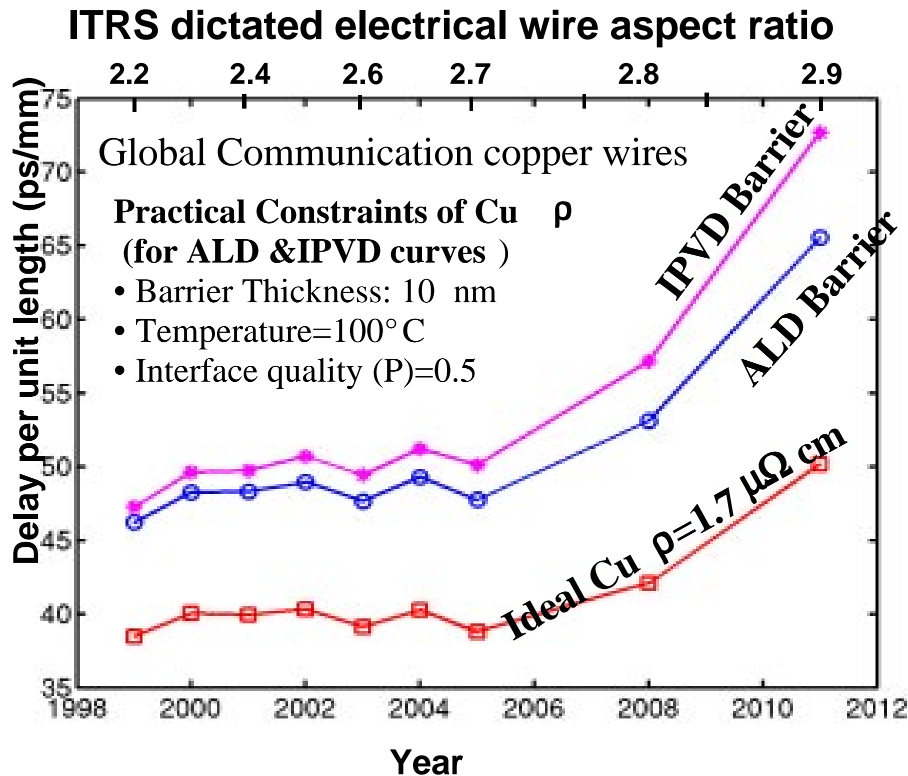


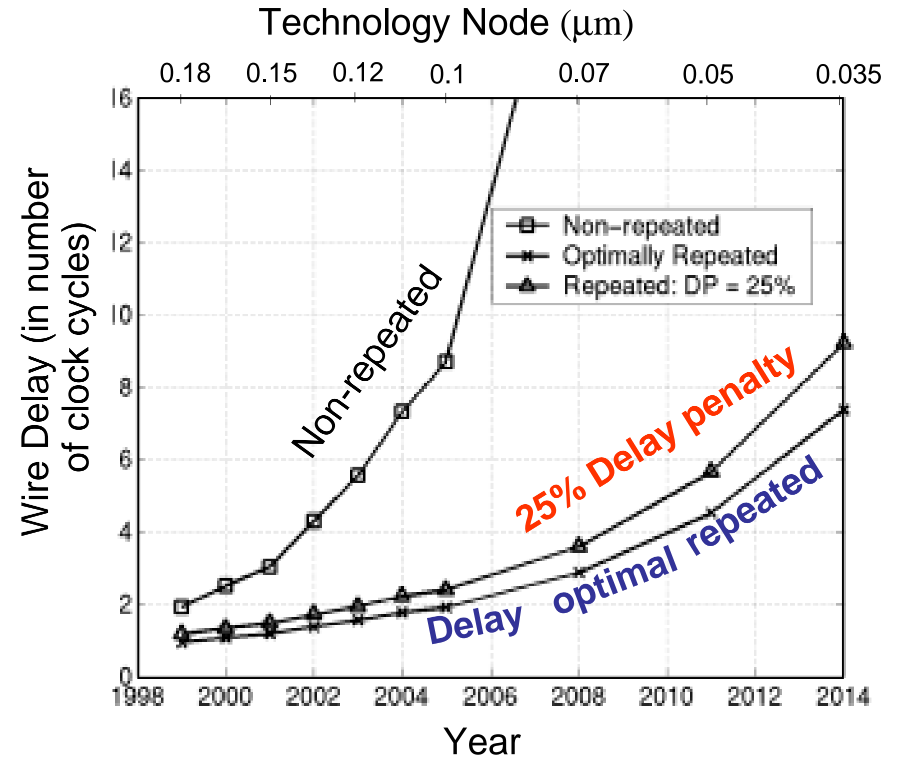
Delay of a line without repeaters $\tau_L = \frac{3.56 \cdot K_{ox}\varepsilon_o\rho}{\lambda^2} L^2$

Delay of a line with n repeaters $(\tau_{L/n} + \tau_G)n = \frac{3.56 \cdot K_{ox}\varepsilon_o\rho}{\lambda^2}\left(\frac{L^2}{n}\right) + n\tau_G$



Signaling wire delay modeling with repeaters

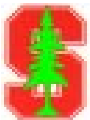




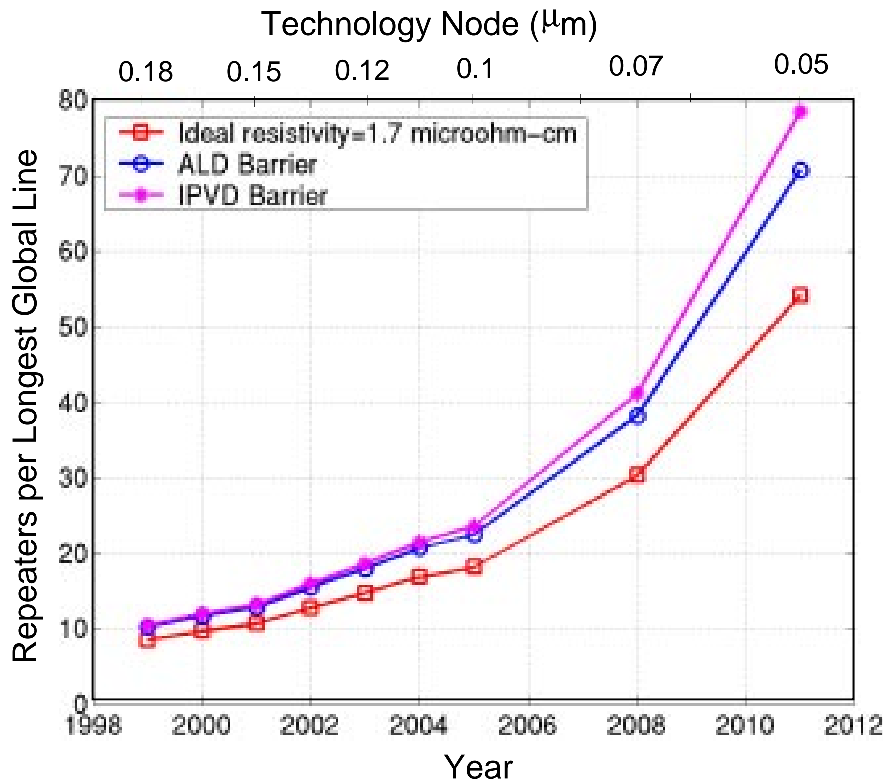
- ALD Barrier likely to be used in the future
  - 66 ps/mm at 50 nm;
  - 93 ps/mm at 35 nm node
  - 1/28 times C at 35 nm node
  - **30% more than with ideal Cu ρ at 50nm node**

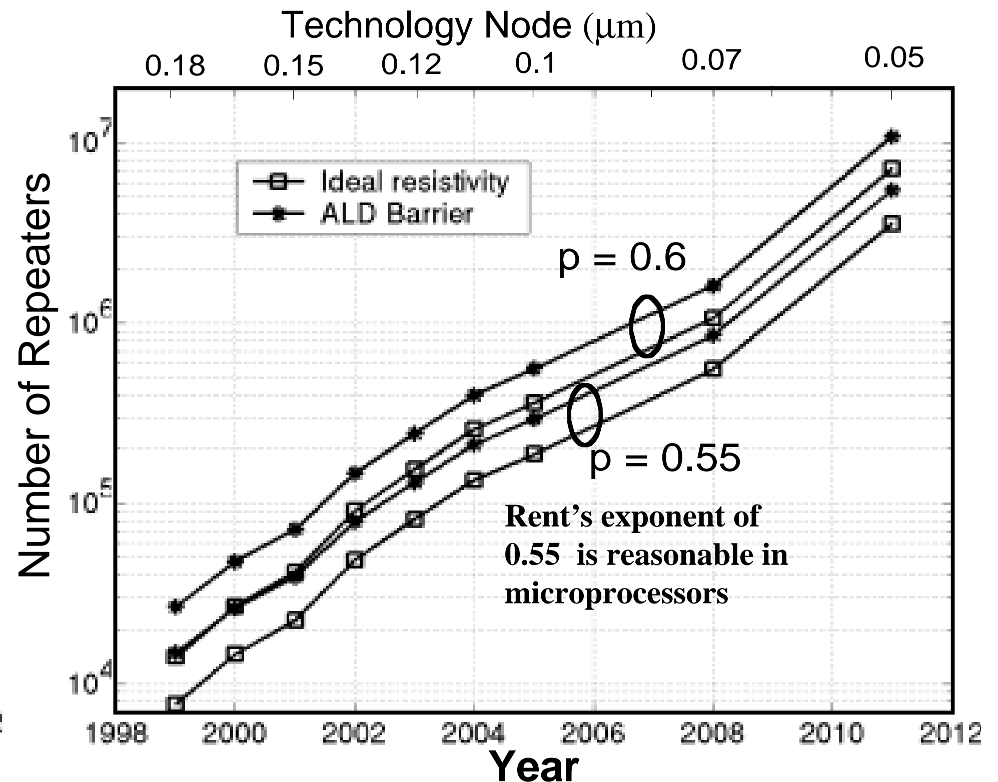
Also have Power and Area penalties
- **Pushing bottleneck to power**

- **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



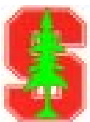
Number of Repeaters Required





- 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

  - A big fraction of the chip area would be occupied by repeaters
  - Additional power will be consumed by repeaters

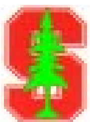


Chip Power: Breakdown

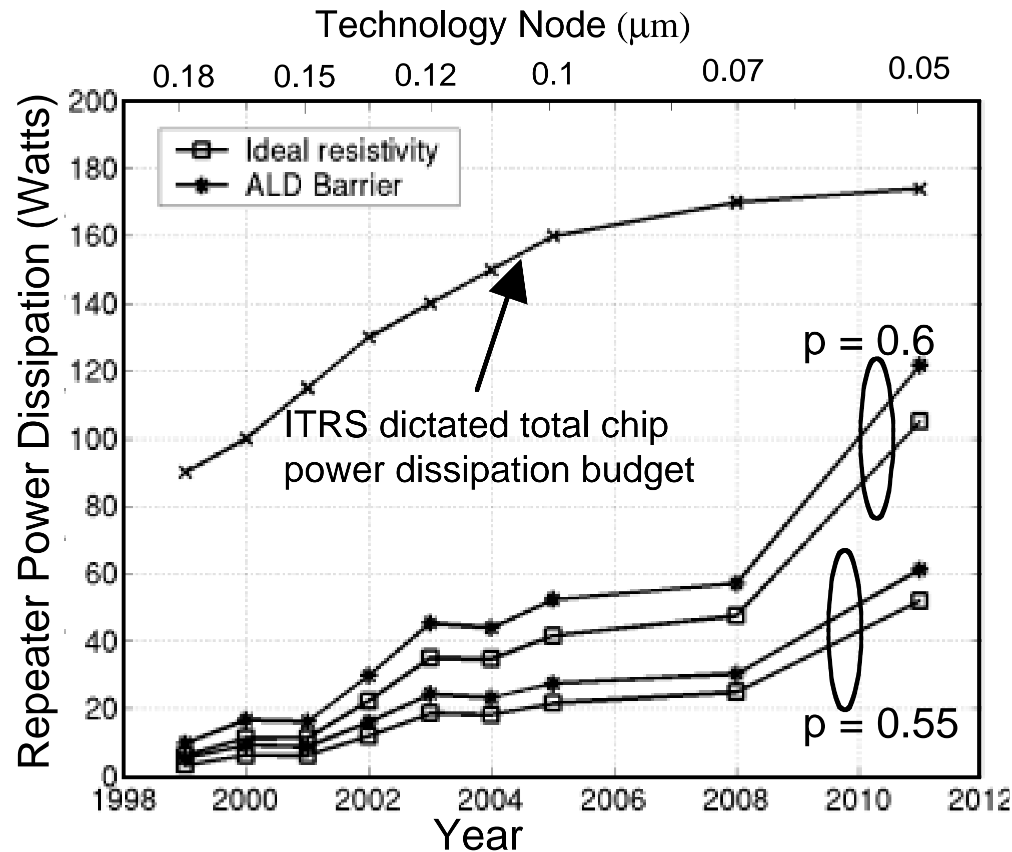
- Dynamic Power: CV^2f
- Leakage power: devices
- Short circuit power during switching
- Analog components (sense amps etc.):static power

Dynamic Power
- **clocking**
  - **Latches**
  - **Clocking Interconnects**
- **Signaling**
  - **Devices**
    - **Logic**
    - **memory**
  - **Signaling Interconnects**
- **I/O**
  - **Buffers**
  - **Off-chip load**

- Interconnect power
  - Due to Cint: dissipated in devices
  - Due to Rint: Joule heating (makes things worse)

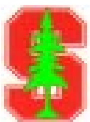


Global Signaling Wire: Repeater Power Penalty

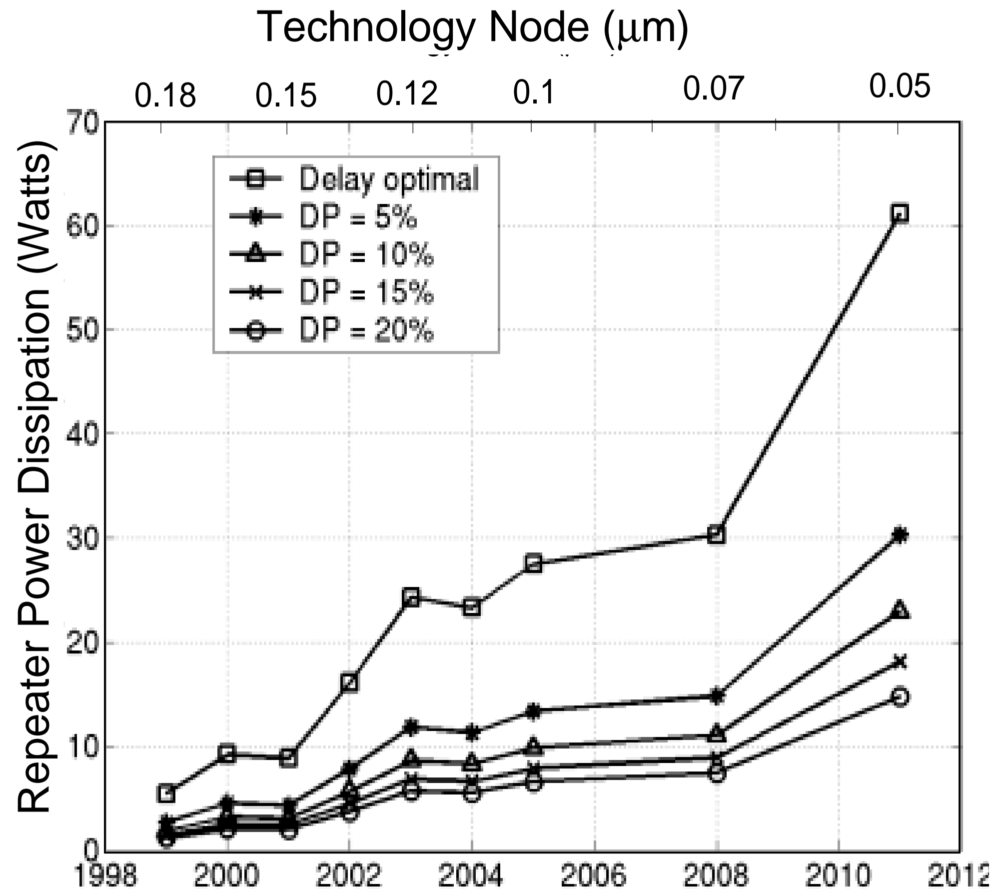


- **Exorbitant power signaling wires at future nodes (50nm)**
  - **Global Wires= 60 Watts (p=0.55)**
  - **Repeaters = 60 Watts (p=0.55)**
  - **120W for just global signaling wires**

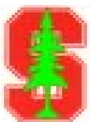
Delay optimal repeaters ~ double power consumption of the wire
- **Global wire power same as above**



Global Signaling Wire: Repeater Power minimization With Delay Tradeoff



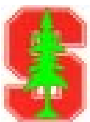
- Tolerable delay penalty depends on architecture
- Still 20W of power dissipation due to repeaters at 50nm node
- With about 20% more delay power dissipation by global wires with repeaters on them is now ~ **60+20=80W** at 50nm node



Electrical Wire latency not a problem: true or false???

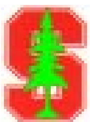
- Can stack repeaters
  - Even with repeaters delay rises
    - in absolute terms (2.7X 180 nm to 35 nm node)
    - compared to clock period (8X 180 nm to 35 nm node)
  - Power and area penalties
- Can pipeline deeper
  - More power penalty (especially with worsening delays with respect to clock period)
- Can exploit locality in communication: Techniques yet to be developed

➢ **Electrical wire delay is a problem because every solution pushes the bottleneck to the power problem and power is becoming EXTREMELY CRITICAL**



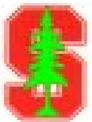
New techniques to minimize the communication distance/time will be needed to continue the evolution in integrated electronics

- **Minimize wire length**
  - **Better circuit design**
  - **3-D ICs**
- **Novel communication mechanisms**
  - **Optical interconnects**
  - **RF wireless interconnects**

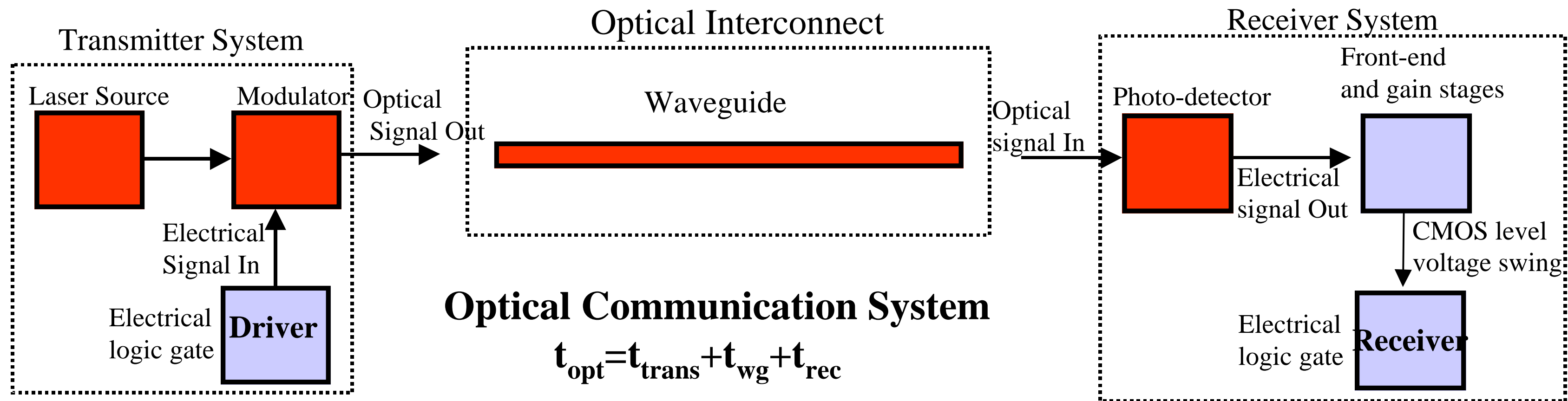


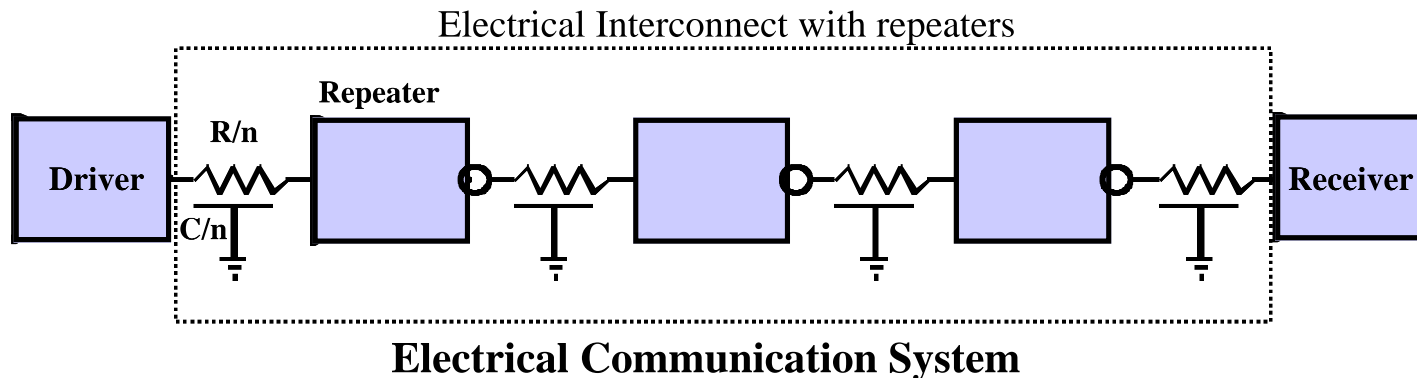
Can Optical Interconnects help?

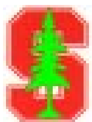
- Signal wires:
  - ✓Reduce delay
  - ✓Increase bandwidth
- Clock distribution
  - ✓Reduce jitter
  - ✓Reduce skew
  - ✓Reduce clock distribution power (50-60% of total power on chip)



Optical Vs. Electrical Wires: Signaling Delay

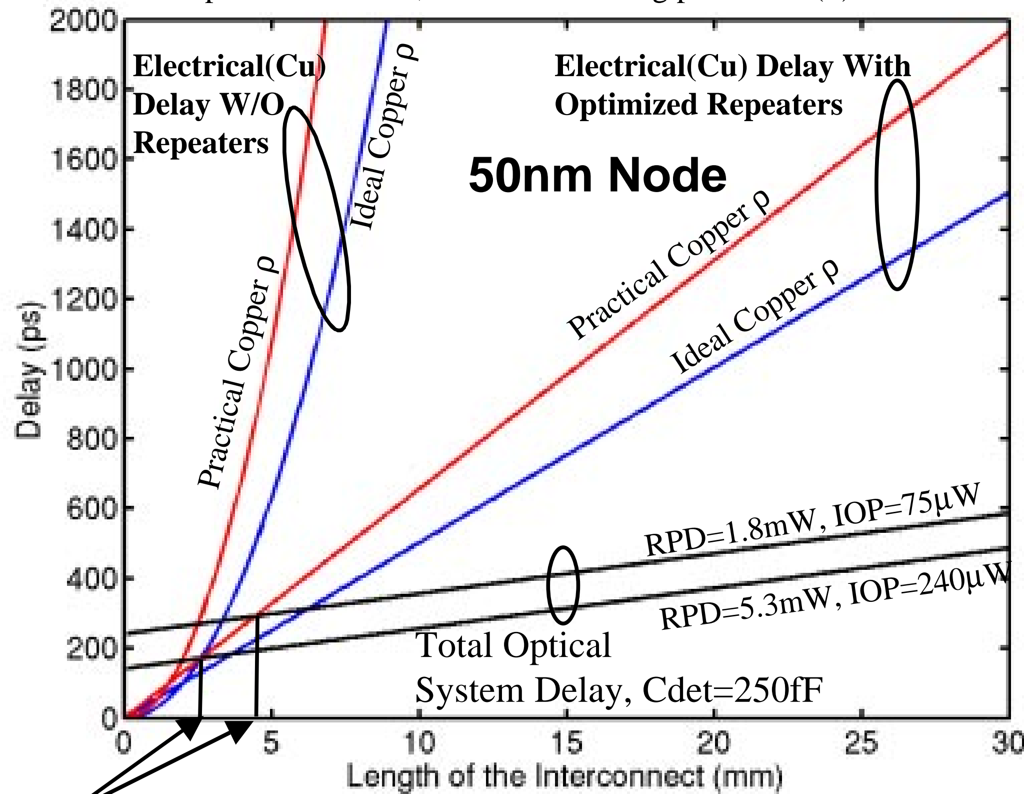






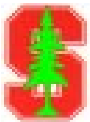
Optical Vs. Electrical Wires: Delay

IOP: Incident Optical Power at the receiver
Practical Cu ρ: ALD Barrier, Barrier Thickness=10nm, temperature=100 ^{0}C, Surface Scattering parameter (P)=0.5

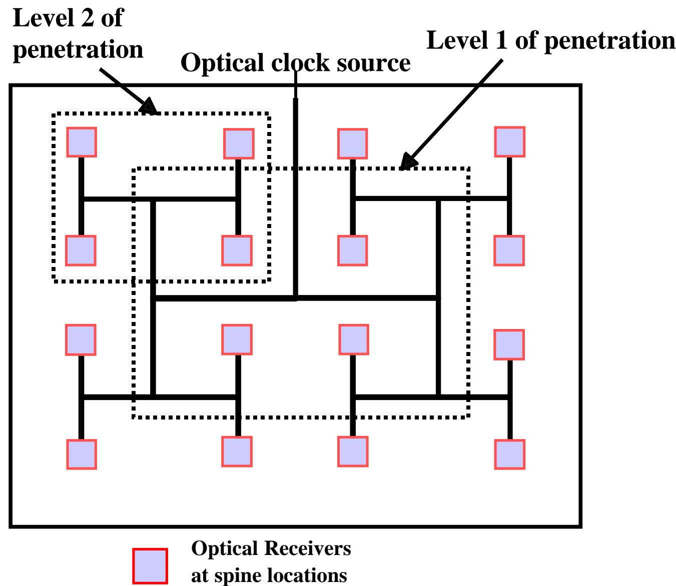


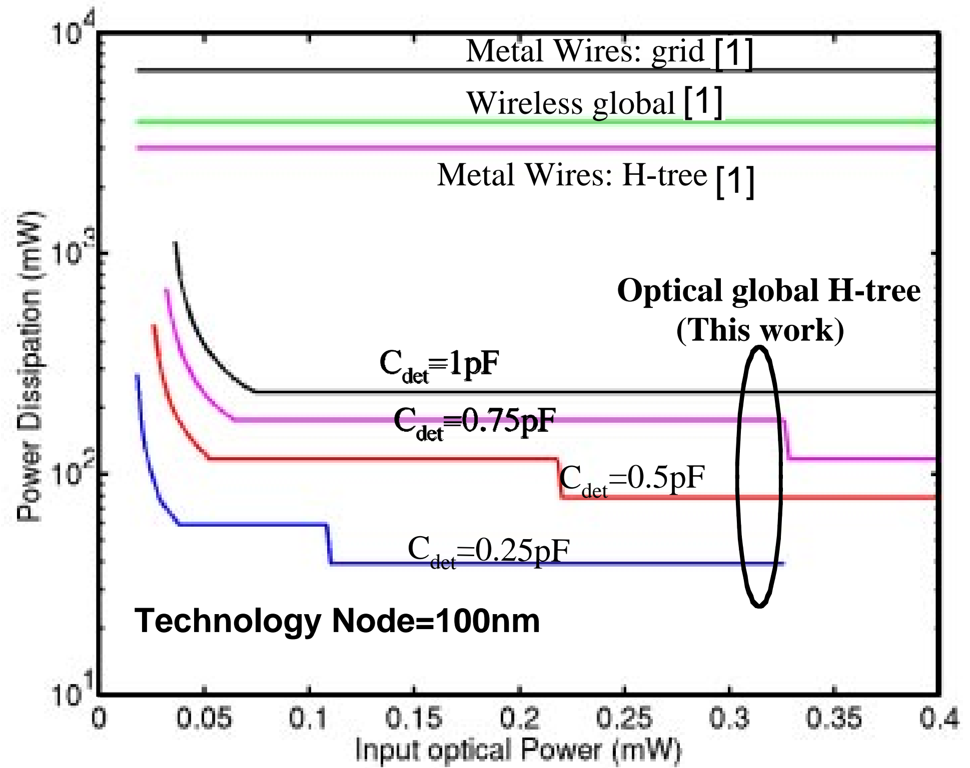
Critical length
above which optical System is
faster than even the electrical (Cu) repeated wires

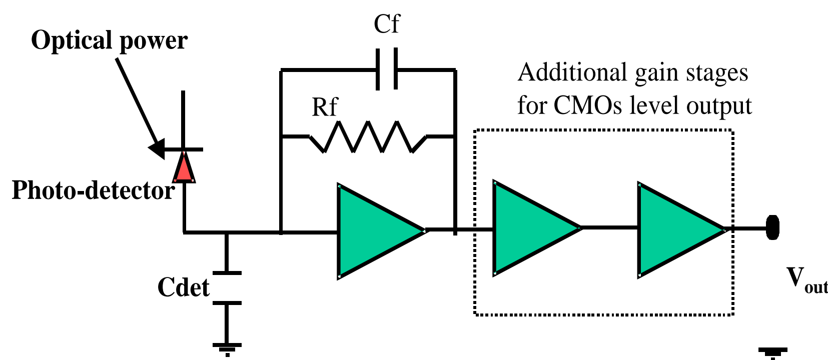
- Optical Interconnects are faster than repeated wires beyond a length well within chip size
- However for Signaling both delay and power are important
- 1.8 mW is approximately power dissipated by a repeated chip edge long wire



Power Dissipation Comparisons Between Metal, Optical and Wireless Clock Distribution

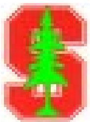




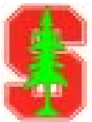


Lower Detector Capacitance and higher IOP for low Receiver power Dissipation

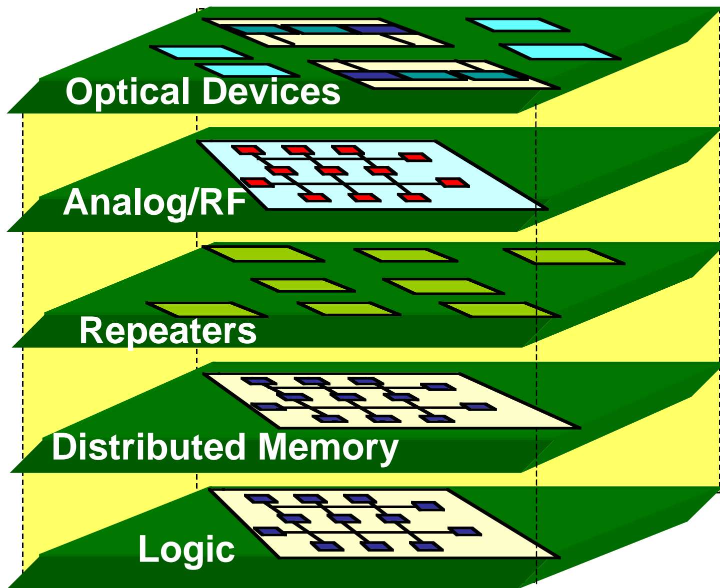
[1] Floyd et al, *Proc.IITC*, pp. 248-251, 1999.



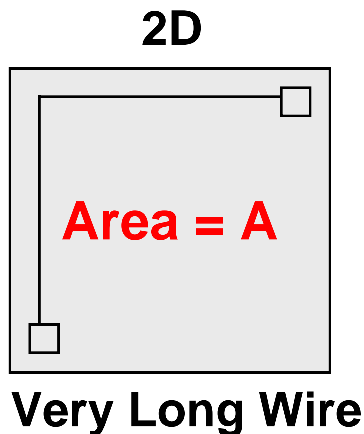
Can we solve the problem by using 3-D Integration?

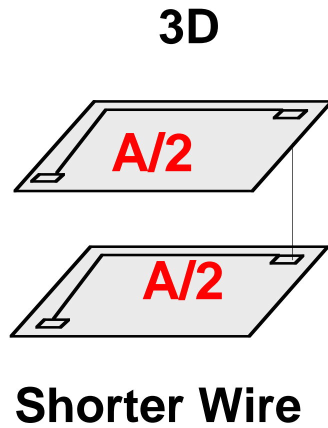


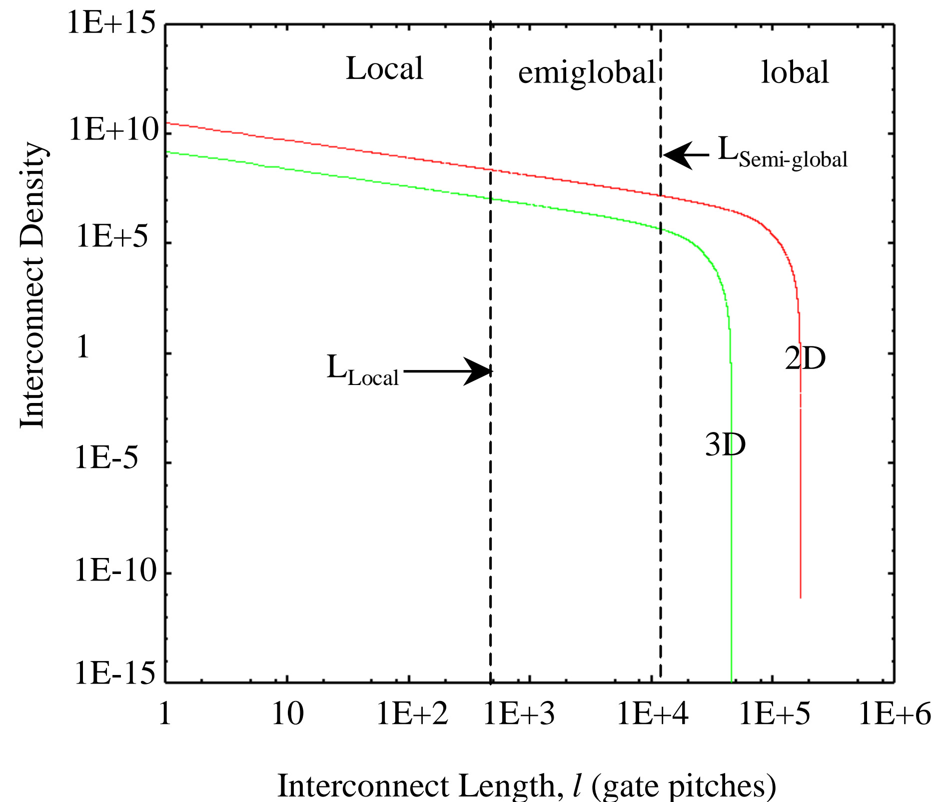
3D ICs: Motivation

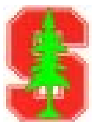


- Reduce Chip footprint
- Interconnect length and therefore R, L, C can be minimized
- Integration of heterogeneous technologies possible, e.g., memory & logic, optical I/O, etc.





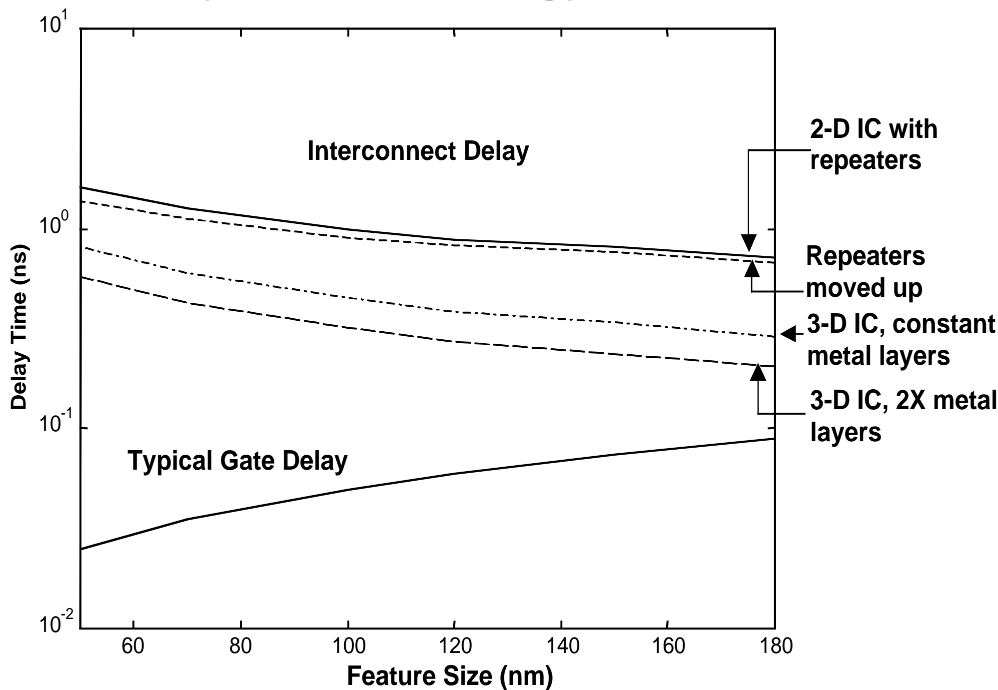




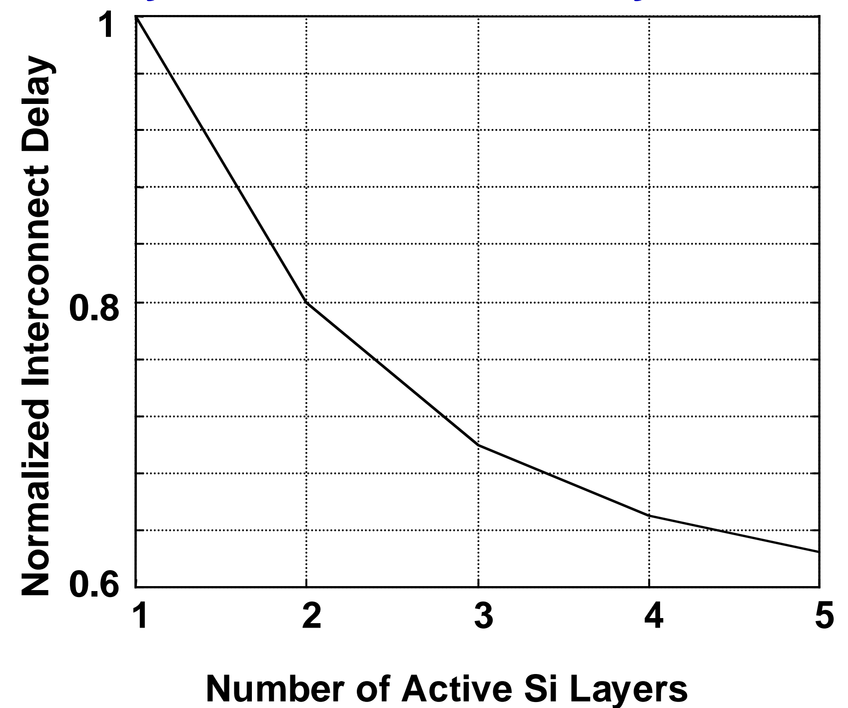
Banerjee, Souri, Kapur & Saraswat Proc IEEE May 2001

Delay of Scaled 3D ICs

3-D ICs with 2 Si Layers Delay vs. Technology Node

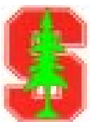


3-D IC with 270 M gates at 50 nm Node Delay vs. Number of Si Layers



Simulations assumed:

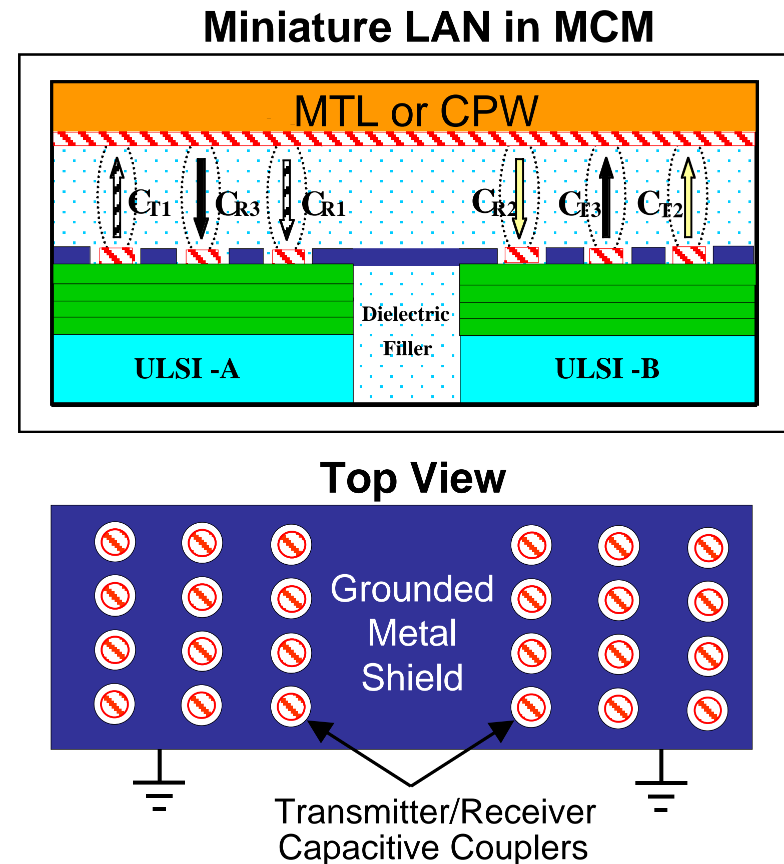
- State-of-the-art chip at a technology node with data from ITRS
- Entire area dedicated to logic
- Delay is due to the longest wire on the chip



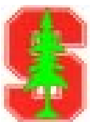
Banerjee, Souri, Kapur & Saraswat, Proc IEEE May 2001

RF-Interconnect System Concept and Payoff

- **Low loss, dispersion-free, ultra-high data rate (100Gbps/channel & 20Tbps/chip)**
- **Multi-I/Os per channel, simultaneous communications via shared MTL or CPW using FDMA/CDMA multiple access algorithms**
- **Reconfigurable network for on-line system-level rewiring (Architecture reconfigurable on-the-fly)**
- **Coherent chip-module combined interconnect scheme, compatible with mainstream ULSI, MCM or surface-mount PCB**



Source: Frank Chang, UCLA



Conclusions

- **Barrier and surface scattering effects vital in dictating Cu effective resistivity**

- **Cu effective resistivity will rise to prohibitively high values even with best barrier technology: Atomic Layer Deposition (ALD)**

- **Performance Parameters (ALD barrier 10 nm T=100 ^{0}C, P=0.5, global Interconnects)**
  - **Resistance per unit length: 1200 Ω/mm (35 nm node); 30 Ω/mm (180 nm node)**
  - **Delay even with repeaters: 6 times clock period (35 nm node); underestimate to about 4 times with ideal ρ**
  - **Number of repeaters per line: 70-80 (ALD & IPVD) vs. 55 (ideal)**
  - **Repeaters increase area and power**

- **Future Recommendations and identification of some key technological concentration**
  - **ITRS may need to be revised in light of above results**
  - **Need for barrierless technology, new ultra cooling mechanism (lower wire temperature)**
    and interface technology yielding P values close to 1
  - **Limitations of Copper Technology: May need alternate technologies such as optical interconnects, 3D or wireless**