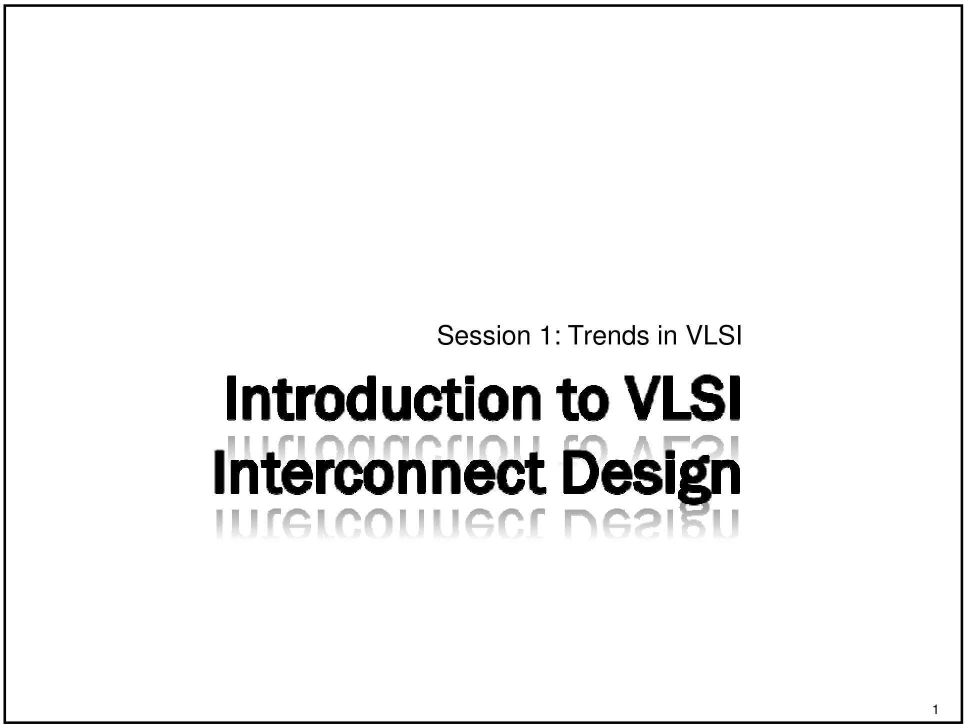
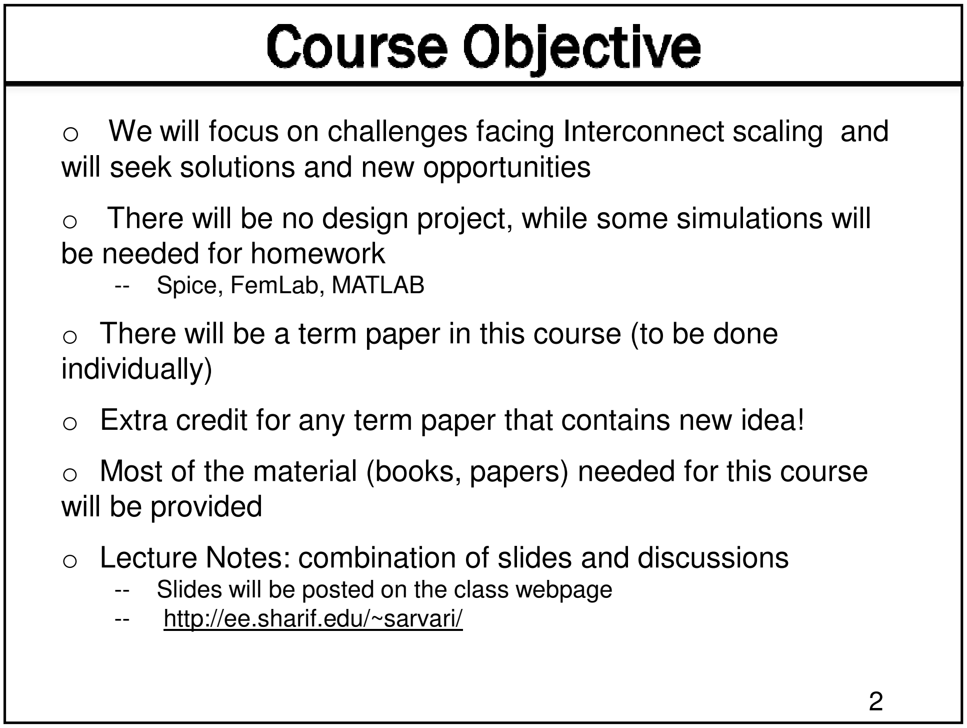


Session 1: Trends in VLSI

Introduction to VLSI Interconnect Design



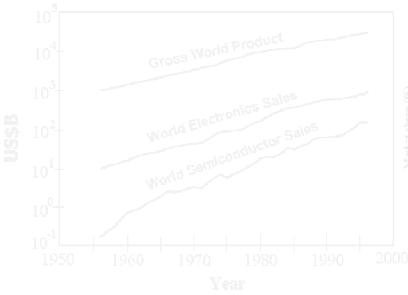
Course Objective

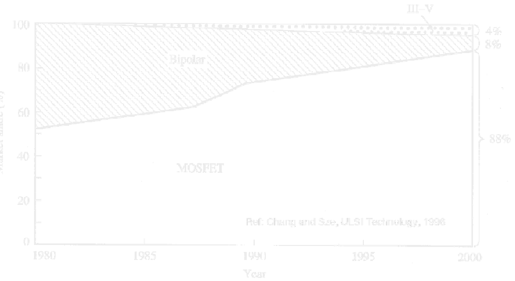
- We will focus on challenges facing Interconnect scaling and will seek solutions and new opportunities
- There will be no design project, while some simulations will be needed for homework
  - Spice, FemLab, MATLAB
- There will be a term paper in this course (to be done individually)
- Extra credit for any term paper that contains new idea!
- Most of the material (books, papers) needed for this course will be provided
- Lecture Notes: combination of slides and discussions
  - Slides will be posted on the class webpage
  - http://ee.sharif.edu/~sarvari/

Required Text/Reference Material

- H. B. Bakoglu, Circuits, Interconnections, and Packaging for VLSI, Addison-Wesley Publishing Company.
- J. A. Davis, J. D. Meindl, Interconnect technology and design for gigascale integration, Kluwer. Academic Publishers.
- Nurmi, J.; Tenhunen, H.; Isoaho, J.; Jantsch, A., Interconnect-Centric Design for Advanced SOC and NOC, Springer.
- C.-K. Cheng, J. Lillis, S. Lin, N. Chang, Interconnect Analysis and Synthesis, Wiley Inter-Science.
- Hall, S.H., G. W. Hall and J. McCall, High-Speed Digital System Design, Wiley-Interscience.
- Selected research papers from the literature

Trends in Semiconductor/CMOS Market





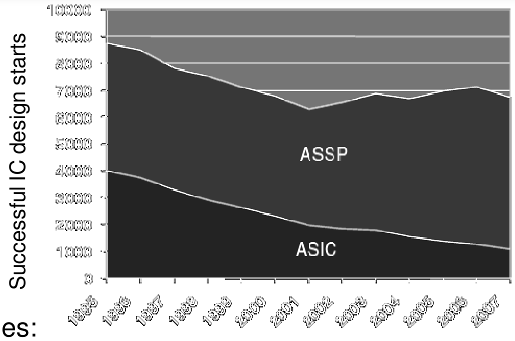
Ref: Chang and Sze, ULSI Technology, 1996

Semiconductors have become increasingly more important part of world economy

CMOS has become the pervasive technology

In 2000: 0.7% of GWP
Today: 5% of GWP

Intriguing … but Challenging



Challenges:

- The NRE cost of IC manufacturing (about 2M$ for mask)
- Deep-submicron effects
- Complexity (100 million transistors)
- Power and Energy
- Reliability and Robustness
- Beyond Silicon!

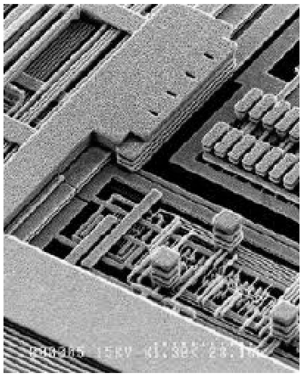
ASSP is an integrated circuit that implements a specific function that appeals to a wide market. As opposed to ASICs that combine a collection of functions and designed by or for one customer

Interconnect?!

2 Major problems facing Moore's law:

- **Power dissipation**
- **Interconnects**

IBM Cu technology



Connectivity and Complexity

Challenge of System Complexity

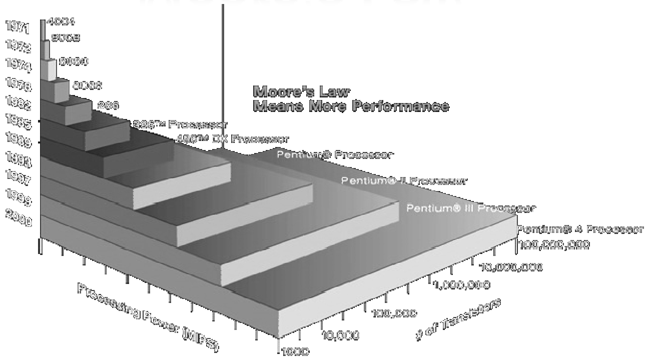
NoC Network-on-a-Chip

Traditional communication techniques:
point-to-point connection, busses

The wires occupy much of the area of the chip, and in nanometer CMOS technology, interconnects dominate both performance and dynamic power dissipation, as signal propagation in wires across the chip requires multiple clock cycles

NoC is similar to a modern telecommunications network, using digital bit-packet switching over multiplexed links. An NoC is constructed from multiple point-to-point data links interconnected by switches (a.k.a. routers).
NoC links can reduce the complexity of designing wires for predictable speed, power, noise, reliability, etc

Moore’s Law



Moore's Law, the empirical observation that the transistor density of integrated circuits doubles every 2 years.

Moore: Moore's law has been the name given to everything that changes exponentially. I say, if Gore invented the Internet, I invented the exponential.

9

Moore’s Law in Perspective

The number of transistors shipped in 2003 had reached about 10^{18}. That’s about 100 times the number of ants estimated to be in the world.

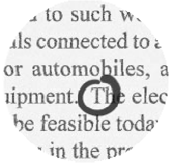
A chip-making tool levitated images within a tolerance of 1/10,000 the thickness of a human hair — a feat equivalent to driving a car straight for 1000 km while deviating less than one 3.8cm.

It would take you about 25,000 years to turn a light switch on and off 1.5 trillion times, but Intel has developed transistors that can switch on and off that many times each second..

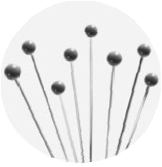
10

Moore’s Law in Perspective

In 1978, a flight between New York and Paris cost around $900 and took 7 hours. If the principles of Moore’s Law had been applied to the airline industry the way they have to the semiconductor industry, that flight would now cost about a penny and take less than 1 sec.

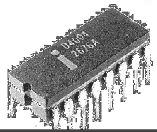


The price of a transistor is now about the same as that of one printed newspaper character.

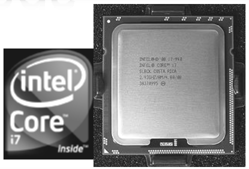


Intel has developed transistors so small that about 200 million of them could fit on the head of each of these pins.

Intel µP Trends

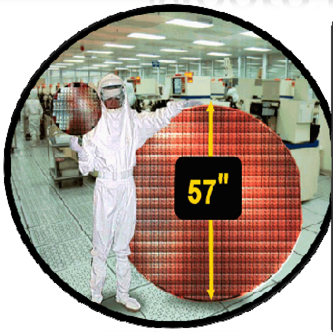


- Intel 4004: first single-chip microprocessor
- November 15, 1971
- Clock rate 740 kHz
- Bus Width 4 bits (multiplexed address/data due to limited pins)
- PMOS
- 2,300 Transistors at **10 µm**
- Addressable Memory 640 bytes
- Program Memory 4 KB (4 KB)

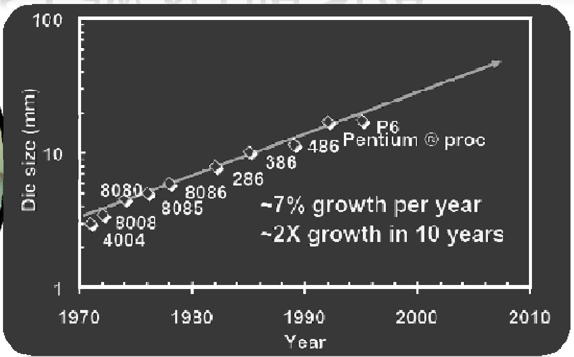


- Intel Core i7
- Today
- Clock rate 2.66GHz-3.33GHz
- 64 bit processor
- 4 cores
- 731M Transistors at **45 nm**
- Oregon 32 nm plant
- Price 273-562 $
- 263 mm2 die size

Moore’s Law & Die Size



Wafer size!



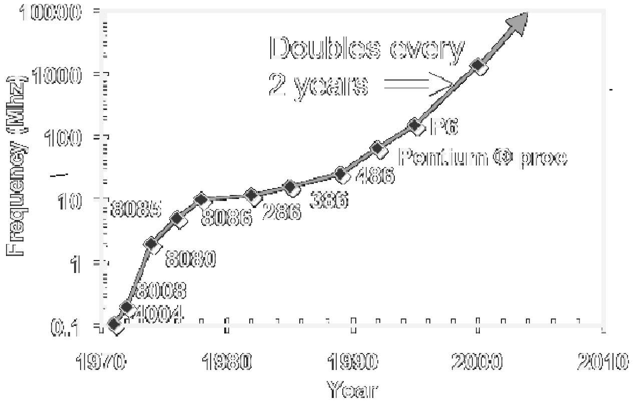
Moore was not always accurate

Projected Wafer in 2000, circa 1975

Die size has grown by 14% to satisfy Moor’s law, BUT the growth is almost stopped because of manufacturing and cost issues

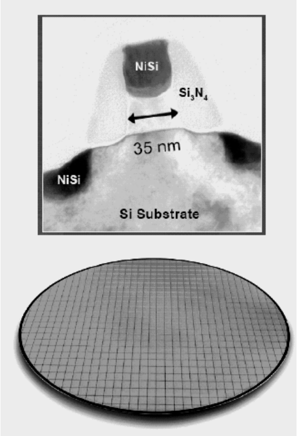
The die size of the processor refers to its physical surface area size on the wafer, the first generation Pentium used a 0.8 micron circuit size, and required 296 mm^2 per chip. The second generation chip had the circuit size reduced to 0.6 microns, and the die size dropped by a full 50% to 148 mm^2!!!

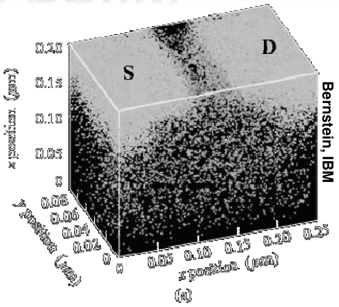
Trends in Clock Frequency



Lead microprocessors frequency doubles every 2 year, BUT the growth is slower because of power dissipation issue

MOS in 65nm





Bernstein, IBM

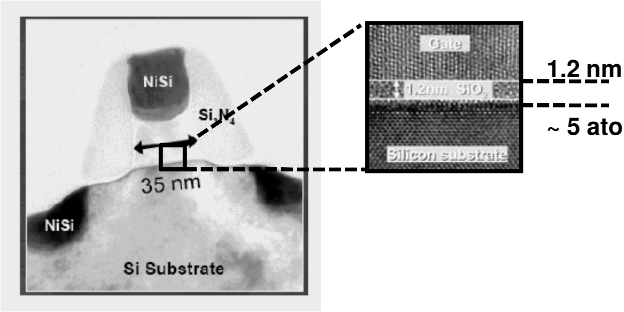
Distance between Si atoms = 5.43 ºA

# of atoms in channel =
35 nm / 0.543 nm = 64 Atoms!

Problem: Uncertainty in transistor behavior and difficult to control variation!
Randomly placed dopants in channel

15

Gate Insulator Thickness in 65nm



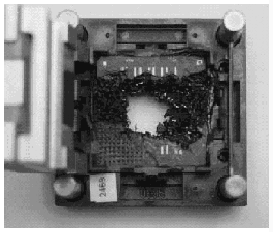
1.2 nm SiO2

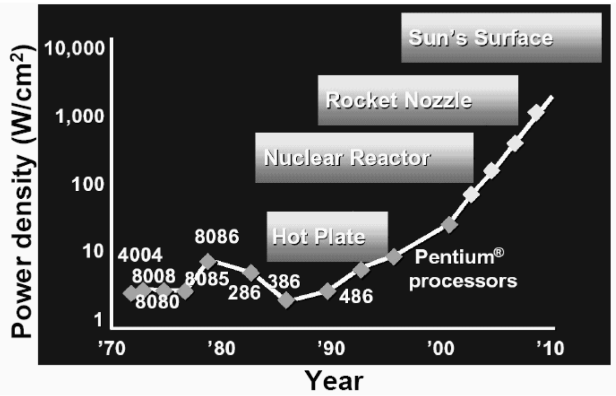
~ 5 atomic layers!

Problem: Electrons can easily jump over
the 5 atomic layers!
This is known as leakage current

16

Power Density Problem





Power density too high to keep junction
at low temperature.
Power reaching limits of air cooling.

Power Density Problem

Power = 115 Watts
Supply Voltage = 1.2 V
Supply Current = 115 W / 1.2 V
= 96 Amps!

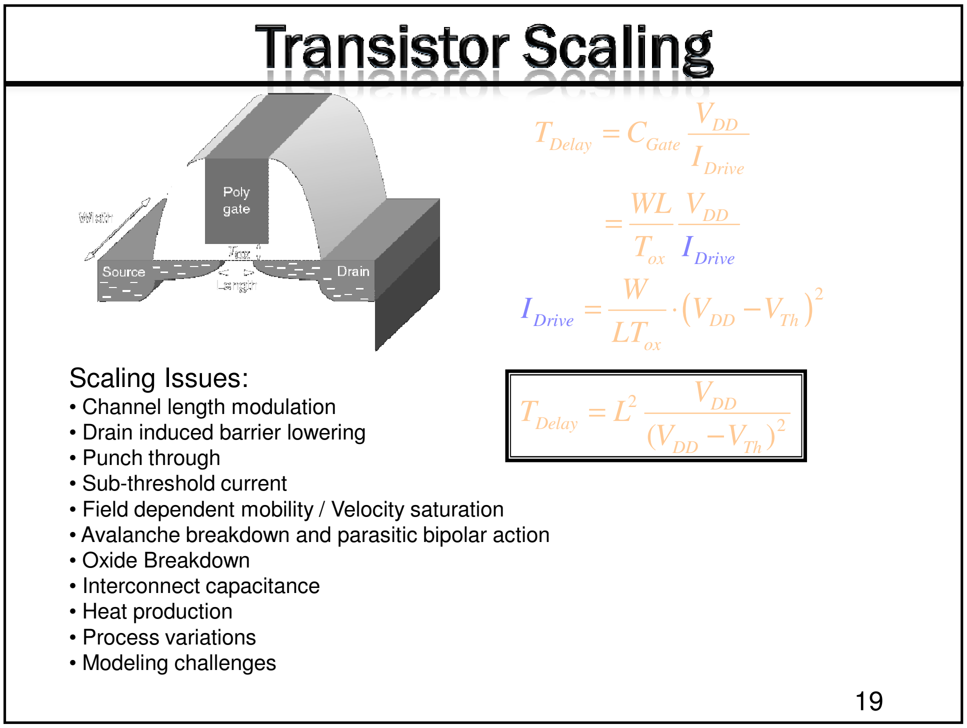
Note:
Fuses used for household appliances = 15 to 40 Amps

Problem:
Current density becomes a serious problem!
This is known as electromigration

Power = 115 Watts
Chip Area = 2.2 cm^2
Heat Flux = 115 W / 2.2 cm^2
= 50 W/cm^2 !

Notes:
Heat flux in iron = 0.2 W/cm^2
Heat flux in frying pan = 10 W/cm^2

Problem:
Heat flux is another serious issue!



Transistor Scaling

$$T_{Delay} = C_{Gate} \frac{V_{DD}}{I_{Drive}}$$

$$= \frac{WL}{T_{ox}} \frac{V_{DD}}{I_{Drive}}$$

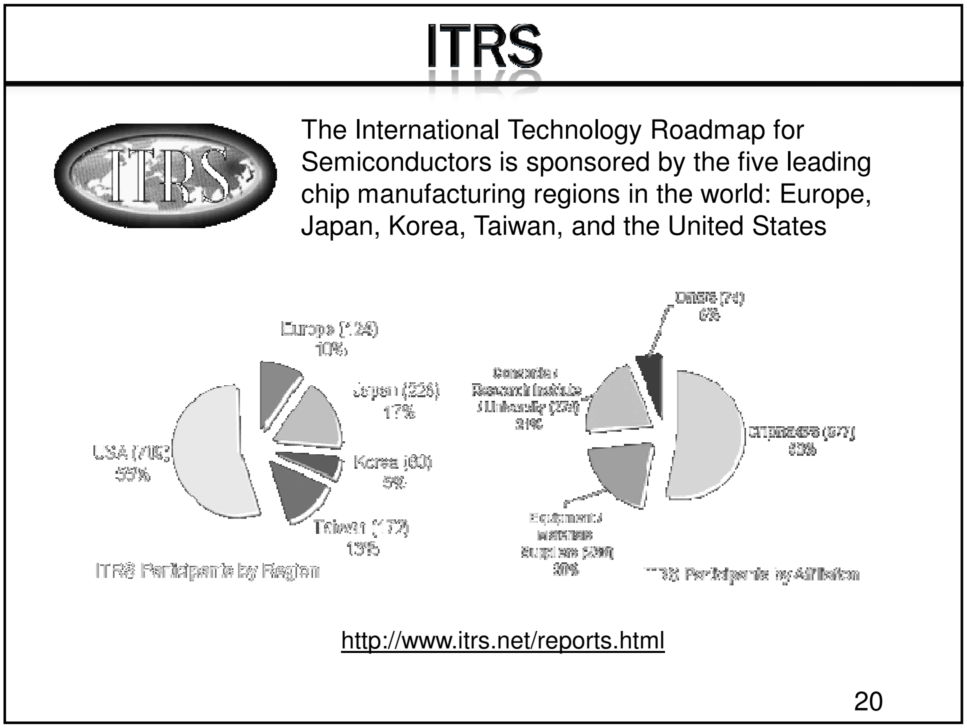
$$I_{Drive} = \frac{W}{LT_{ox}} \cdot (V_{DD} - V_{Th})^2$$

$$\boxed{T_{Delay} = L^2 \frac{V_{DD}}{(V_{DD} - V_{Th})^2}}$$

Scaling Issues:

- Channel length modulation
- Drain induced barrier lowering
- Punch through
- Sub-threshold current
- Field dependent mobility / Velocity saturation
- Avalanche breakdown and parasitic bipolar action
- Oxide Breakdown
- Interconnect capacitance
- Heat production
- Process variations
- Modeling challenges

19



ITRS

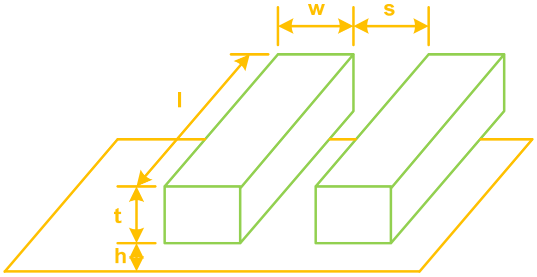
The International Technology Roadmap for Semiconductors is sponsored by the five leading chip manufacturing regions in the world: Europe, Japan, Korea, Taiwan, and the United States

http://www.itrs.net/reports.html

20

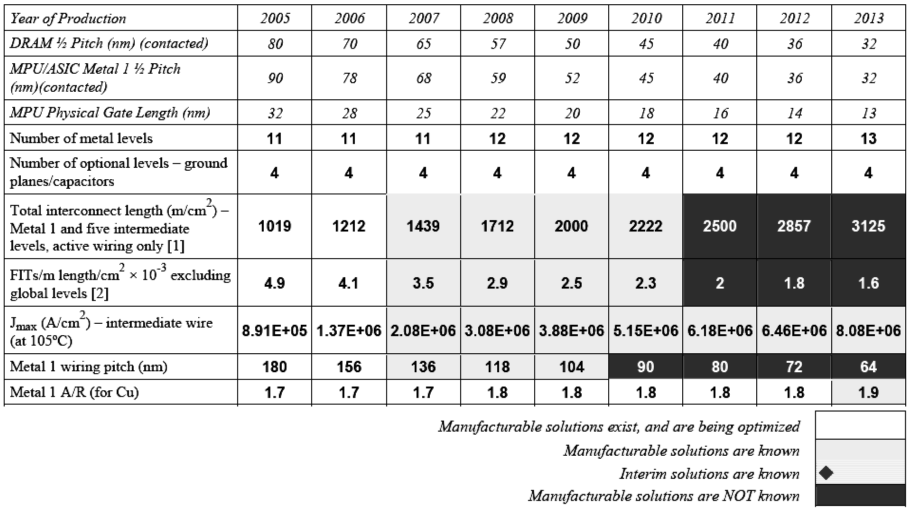
Wire Geometry

- ○ Pitch = w + s
- ○ Aspect ratio: AR = t/w
  - Old processes had AR << 1
  - Modern processes have AR ≈ 2
  - Pack in many skinny wires



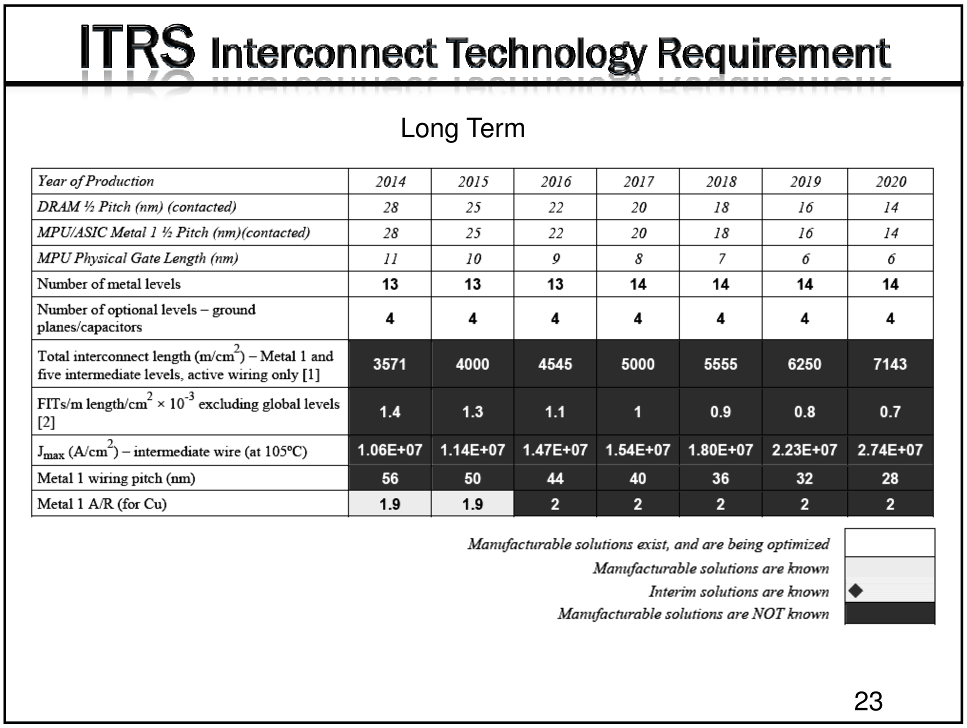
ITRS Interconnect Technology Requirement

Short Term



Year of Production	*2005*	*2006*	*2007*	*2008*	*2009*	*2010*	*2011*	*2012*	*2013*
DRAM ½ Pitch (nm) (contacted)	*80*	*70*	*65*	*57*	*50*	*45*	*40*	*36*	*32*
MPU/ASIC Metal 1 ½ Pitch (nm)(contacted)	*90*	*78*	*68*	*59*	*52*	*45*	*40*	*36*	*32*
MPU Physical Gate Length (nm)	*32*	*28*	*25*	*22*	*20*	*18*	*16*	*14*	*13*
Number of metal levels	11	11	11	12	12	12	12	12	13
Number of optional levels – ground planes/capacitors	4	4	4	4	4	4	4	4	4
Total interconnect length (m/cm^2) – Metal 1 and five intermediate levels, active wiring only [1]	1019	1212	1439	1712	2000	2222	2500	2857	3125
FITs/m length/cm^2 × 10^{-3} excluding global levels [2]	4.9	4.1	3.5	2.9	2.5	2.3	2	1.8	1.6
J_{max} (A/cm^2) – intermediate wire (at 105°C)	8.91E+05	1.37E+06	2.08E+06	3.08E+06	3.88E+06	5.15E+06	6.18E+06	6.46E+06	8.08E+06
Metal 1 wiring pitch (nm)	180	156	136	118	104	90	80	72	64
Metal 1 A/R (for Cu)	1.7	1.7	1.7	1.8	1.8	1.8	1.8	1.8	1.9

Manufacturable solutions exist, and are being optimized	□
Manufacturable solutions are known	▨
Interim solutions are known	◆
Manufacturable solutions are NOT known	■



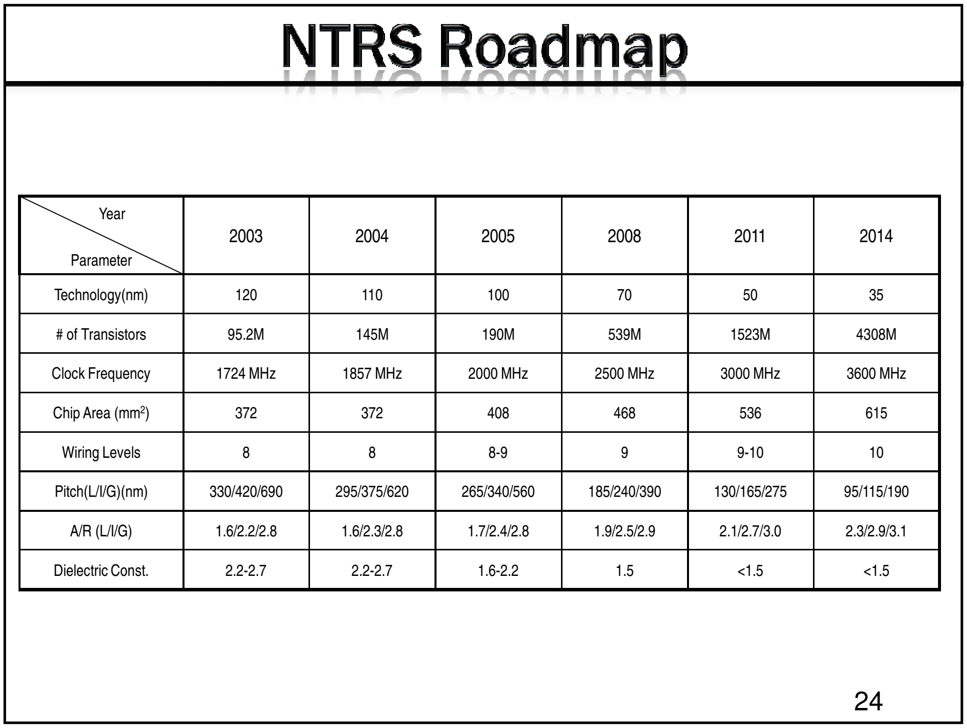
ITRS Interconnect Technology Requirement

Long Term

Year of Production	*2014*	*2015*	*2016*	*2017*	*2018*	*2019*	*2020*
DRAM ½ Pitch (nm) (contacted)	*28*	*25*	*22*	*20*	*18*	*16*	*14*
MPU/ASIC Metal 1 ½ Pitch (nm)(contacted)	*28*	*25*	*22*	*20*	*18*	*16*	*14*
MPU Physical Gate Length (nm)	*11*	*10*	*9*	*8*	*7*	*6*	*6*
Number of metal levels	**13**	**13**	**13**	**14**	**14**	**14**	**14**
Number of optional levels – ground planes/capacitors	**4**	**4**	**4**	**4**	**4**	**4**	**4**
Total interconnect length (m/cm^2) – Metal 1 and five intermediate levels, active wiring only [1]	**3571**	**4000**	**4545**	**5000**	**5555**	**6250**	**7143**
FITs/m length/cm^2 × 10^{-3} excluding global levels [2]	**1.4**	**1.3**	**1.1**	**1**	**0.9**	**0.8**	**0.7**
J_{max} (A/cm^2) – intermediate wire (at 105°C)	**1.06E+07**	**1.14E+07**	**1.47E+07**	**1.54E+07**	**1.80E+07**	**2.23E+07**	**2.74E+07**
Metal 1 wiring pitch (nm)	**56**	**50**	**44**	**40**	**36**	**32**	**28**
Metal 1 A/R (for Cu)	**1.9**	**1.9**	**2**	**2**	**2**	**2**	**2**

Manufacturable solutions exist, and are being optimized	□
Manufacturable solutions are known	▨
Interim solutions are known	◆
Manufacturable solutions are NOT known	■

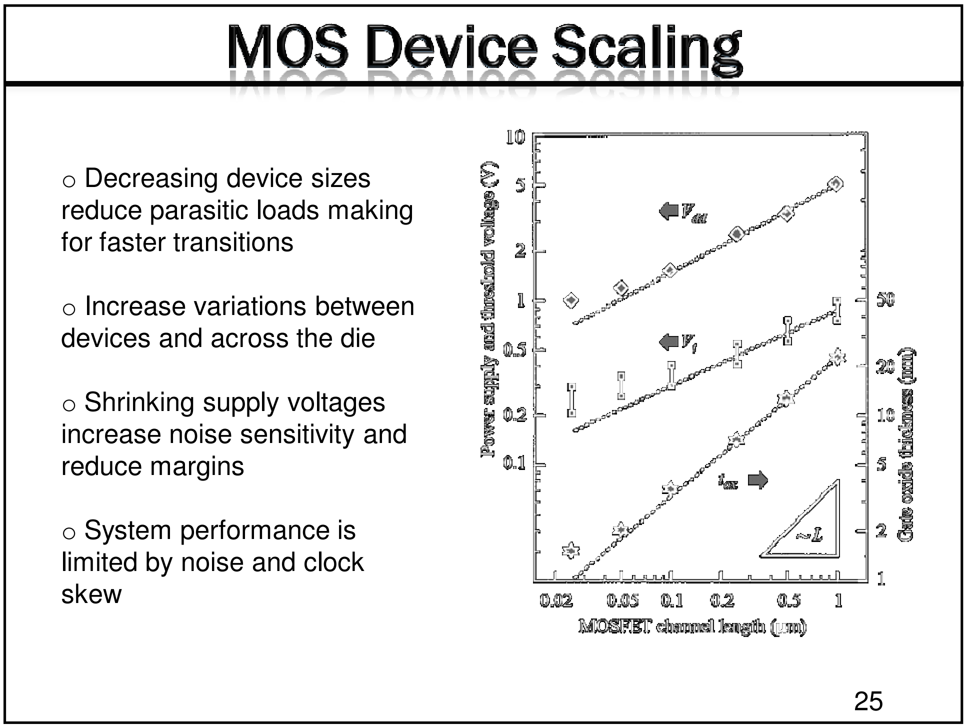
23



NTRS Roadmap

Year / Parameter	2003	2004	2005	2008	2011	2014
Technology(nm)	120	110	100	70	50	35
# of Transistors	95.2M	145M	190M	539M	1523M	4308M
Clock Frequency	1724 MHz	1857 MHz	2000 MHz	2500 MHz	3000 MHz	3600 MHz
Chip Area (mm^2)	372	372	408	468	536	615
Wiring Levels	8	8	8-9	9	9-10	10
Pitch(L/I/G)(nm)	330/420/690	295/375/620	265/340/560	185/240/390	130/165/275	95/115/190
A/R (L/I/G)	1.6/2.2/2.8	1.6/2.3/2.8	1.7/2.4/2.8	1.9/2.5/2.9	2.1/2.7/3.0	2.3/2.9/3.1
Dielectric Const.	2.2-2.7	2.2-2.7	1.6-2.2	1.5	<1.5	<1.5

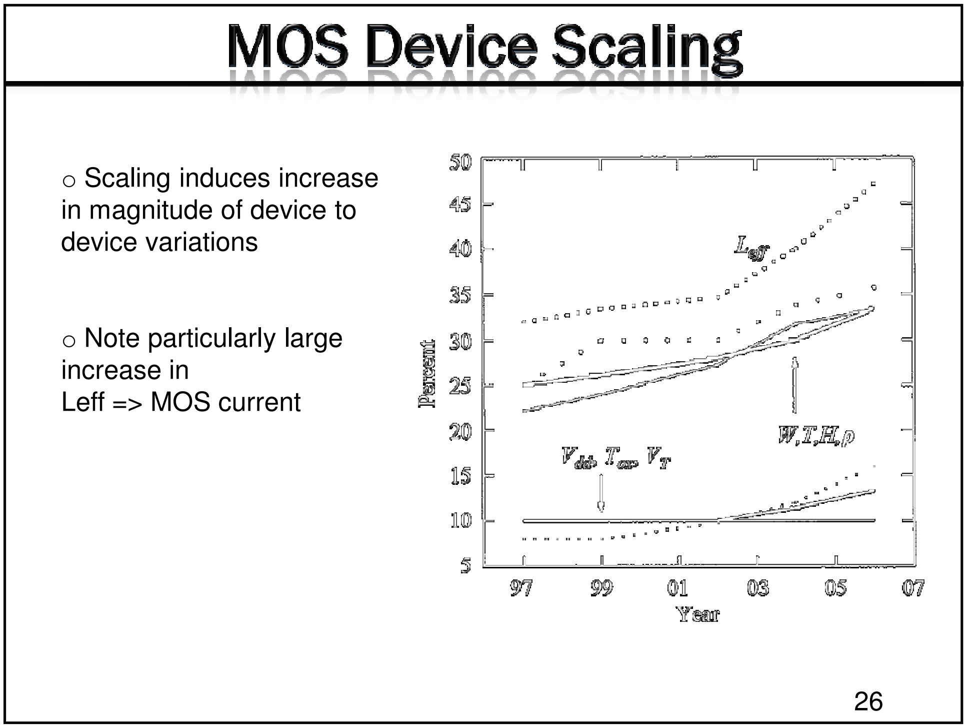
24



MOS Device Scaling

- o Decreasing device sizes reduce parasitic loads making for faster transitions
- o Increase variations between devices and across the die
- o Shrinking supply voltages increase noise sensitivity and reduce margins
- o System performance is limited by noise and clock skew

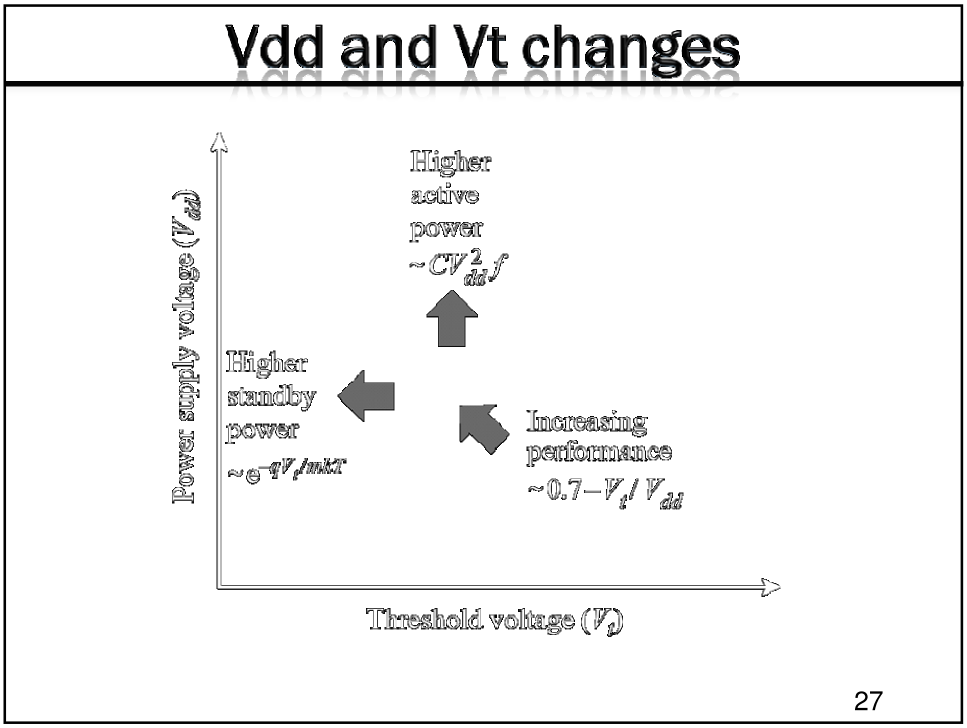
25



MOS Device Scaling

- o Scaling induces increase in magnitude of device to device variations
- o Note particularly large increase in
Leff => MOS current

26



Vdd and Vt changes

Higher active power

~$CV_{dd}^2 f$

Higher standby power

~$e^{-qV_t/mkT}$

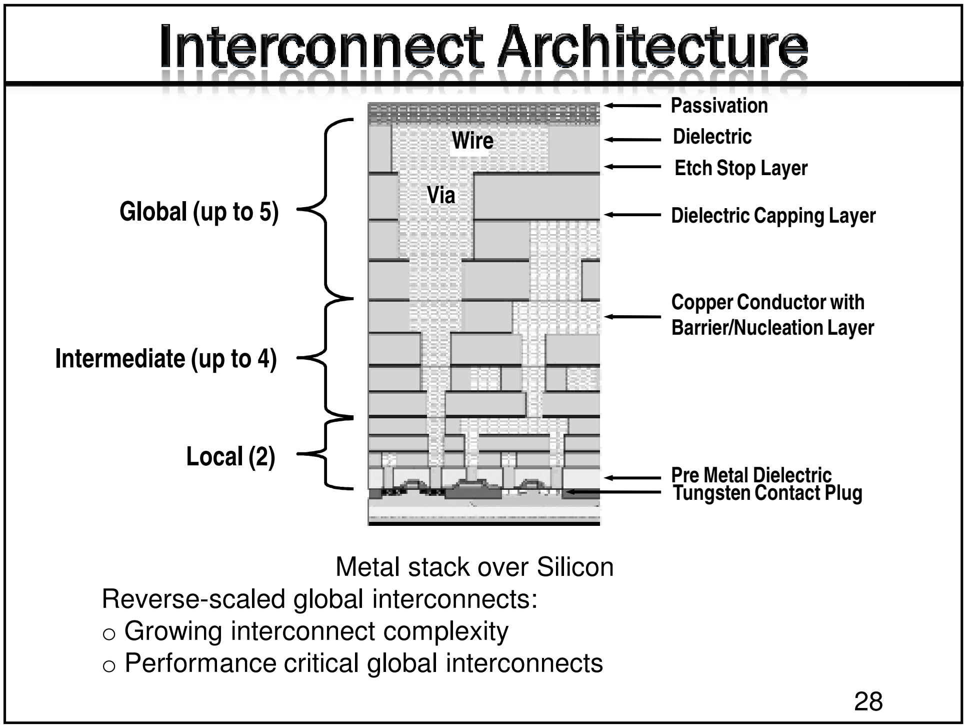
Increasing performance

~$0.7 - V_t / V_{dd}$

Power supply voltage (V_{dd})

Threshold voltage (V_t)

27



Interconnect Architecture

Global (up to 5)

Intermediate (up to 4)

Local (2)

Wire

Via

Passivation

Dielectric

Etch Stop Layer

Dielectric Capping Layer

Copper Conductor with Barrier/Nucleation Layer

Pre Metal Dielectric

Tungsten Contact Plug

Metal stack over Silicon

Reverse-scaled global interconnects:

o Growing interconnect complexity

o Performance critical global interconnects

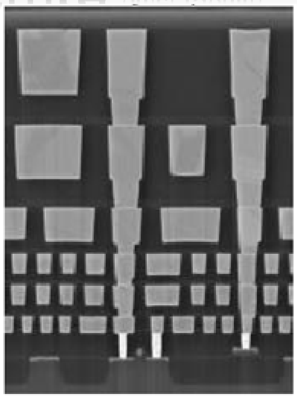
28

Interconnect Architecture (Intel 130nm)

Intel 6LM 130nm process with vias shown (connecting layers)

LAYER PITCH, THICKNESS AND ASPECT RATIO

Layer	Pitch (nm)	Thick (nm)	Aspect Ratio
Isolation	364	450	
Poly-silicon	336	160	
Metal 1	350	280	1.6
Metal 2, 3	448	360	1.6
Metal 4	756	570	1.5
Metal 5	1120	900	1.6
Metal 6	1204	1200	2.0

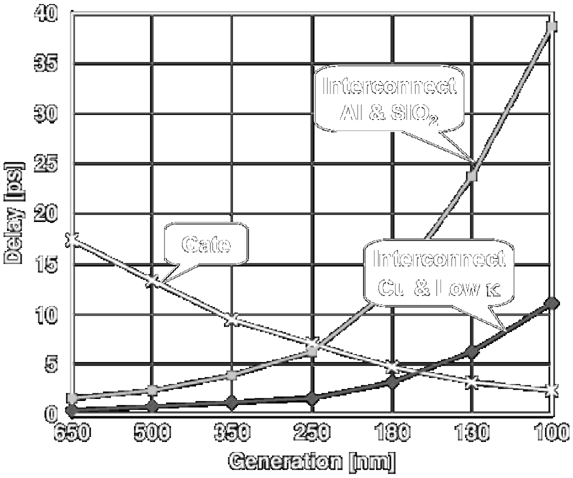


P. Bai et al, IEDM 2004

Real wiring cross section photograph

29

Interconnect vs. Gate Delay



[Bohr, IEDM '95]

Al	3.0 μΩ cm
Cu	1.7 μΩ cm
SiO2	κ = 4.0
Low κ	κ= 2.0
Al & Cu	0.8 μ Thick
Al & CU	Line 43 μ Long

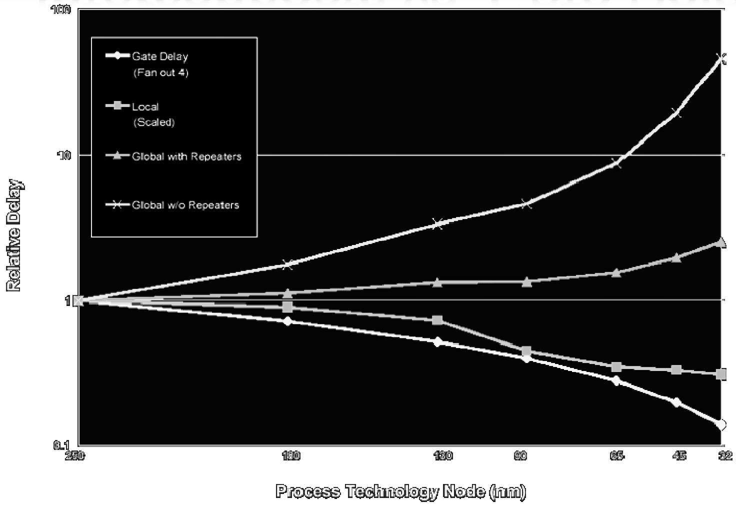
30

Choice of Metal

- Until 180 nm generation, most wires were aluminum
- Modern processes often use copper
  - Cu atoms diffuse into silicon and damage FETs
  - Must be surrounded by a diffusion barrier

Metal	Bulk resistivity (mΩ*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

Interconnects vs. Gate Delay



Delay for Metal 1 and global wiring vs feature size

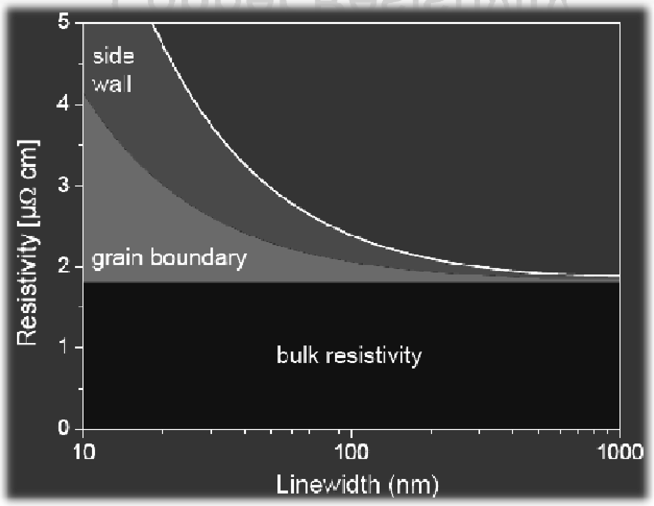
Interconnect Scaling Scenario

Problem with Interconnects?

	Technology generation		
	1um	100nm	35nm
MOSFET switching delay (ps)	~20	~5	~2.5
Interconnect *RC* response time, L=1mm (ps)	~1	~30	**~250**
MOSFET switching energy (fJ)	~30	~2	~0.1
Interconnect switching energy (fJ)	~40	~10	**~3**

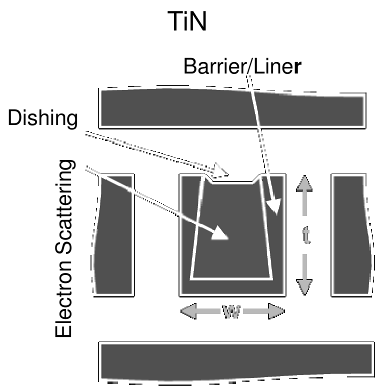
Calculations made by considering
bulk resistivity of Cu

Copper Resistivity



Resistivity of Cu increases with scaling

Interconnect Resistance



Barrier/Liner is usually another metal preventing Copper to diffuse into Si or SiO_2

- Diffusion barrier reduces wire's cross-section
- Cu over polish (dishing) reduces it's thickness

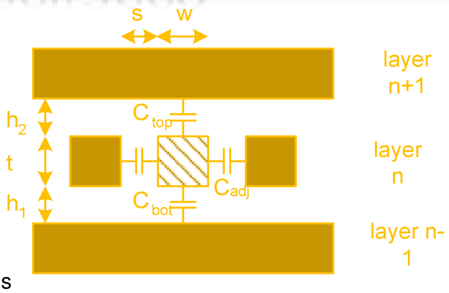
Wire Capacitance

○ Wire has capacitance per unit length

To neighbors

To layers above and below

○ $C_{total} = C_{top} + C_{bot} + 2C_{adj}$



○ Parallel plate equation: C = eA/d

Wires are not parallel plates, but obey trends

Increasing area (W, t) increases capacitance

Increasing distance (s, h) decreases capacitance

plus a fringe term

○ Dielectric constant

$e = ke_0$

$e_0 = 8.85 \times 10^{-14}$ F/cm

k = 3.9 for SiO_2

Processes are starting to use low-k dielectrics

$k \approx 3$ (or less) as dielectrics use air pockets

Capacitance Extraction

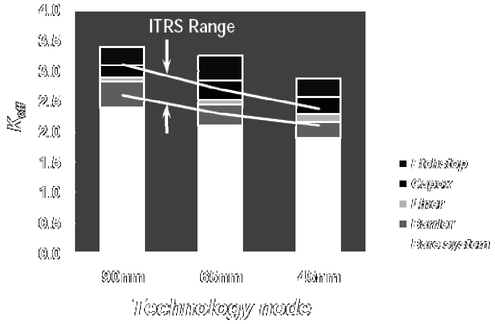
○ Extraction of interconnect capacitance in modern VLSI technology is complicated because of

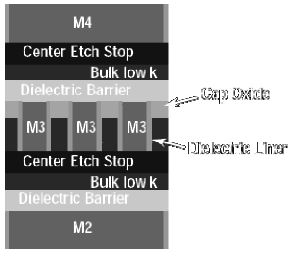
Non-homogenous dielectric (etch stop, barrier liner, etc.)

Complex pattern of neighboring interconnects (need 3D modeling)

○ Sometimes, the overhead layers increases the effective K value

Overhead layers are hard to scale but needs to be controlled





Inductance Figure of Merit

- Should we model wires as full transmission line? (no)

Unless we intentionally make inductance important: very wide wires

Or we are designing the clock grid

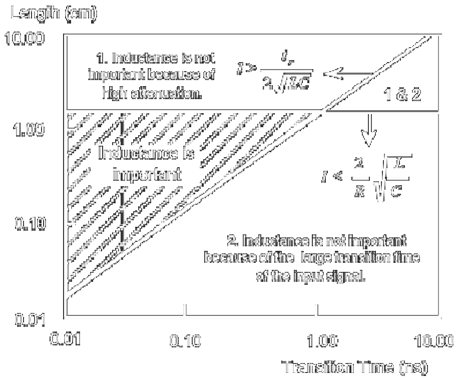
- Transmission line effects can be ignored if the wire is:

Very short, when signal transition is slower than the roundtrip delay

$$t_r > 2L\sqrt{lc}$$

Very long, when it becomes too lossy (resistance is more than 2Zo)

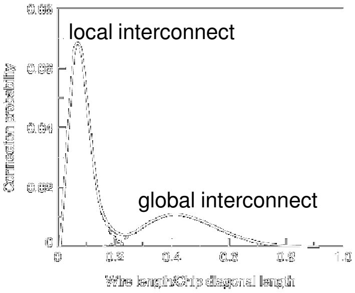
$$rL > 2\sqrt{l/c}$$

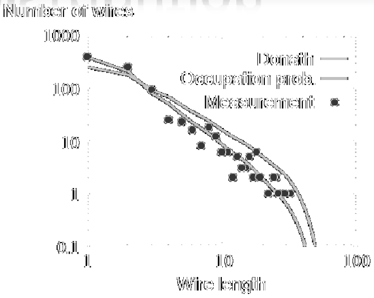


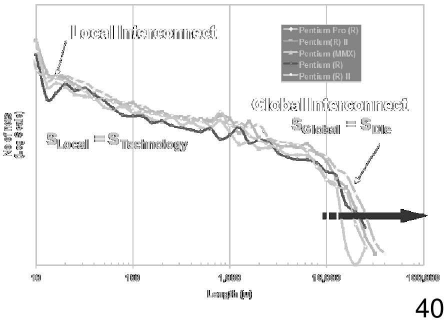
Problems of Inductance Modeling

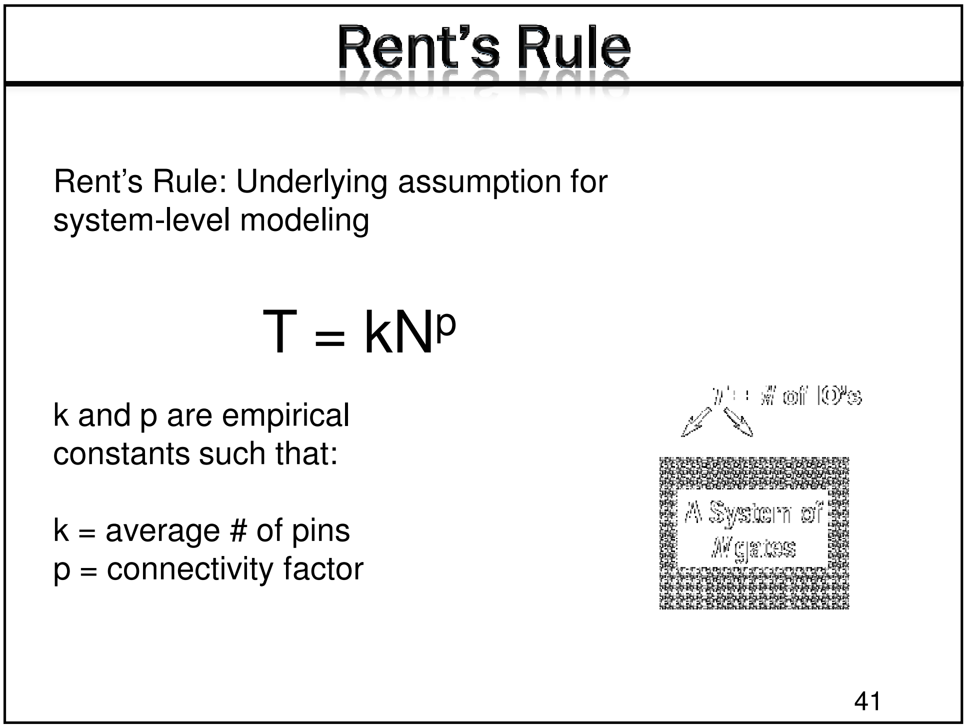
- Extraction of on-chip inductance is very challenging
  - Hard to define return path (unless we use partial inductance technique)
  - Require a huge amount of netlist data (10x more than *RC* netlist data size)
- Simulation of on-chip inductance is also challenging
  - Requires a lot more computation for delay calculation
  - Available techniques have limited accuracy for large circuit structures
- Fortunately, it is not required to include inductance for whole chip analysis

Wire Distribution









Rent’s Rule

Rent’s Rule: Underlying assumption for system-level modeling

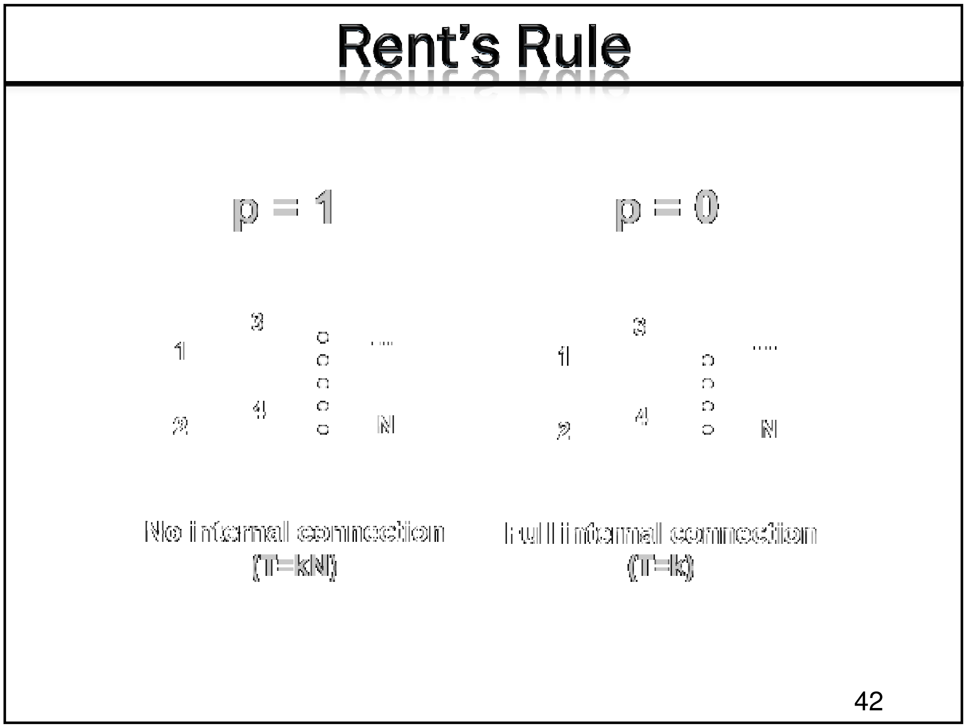
$$T = kN^p$$

k and p are empirical constants such that:

k = average # of pins
p = connectivity factor

T = # of IO's
A System of
N gates

41



Rent’s Rule

p = 1
p = 0
1
2
3
4
N
1
2
3
4
N

No internal connection
(T=kN)

Full internal connection
(T=k)

42

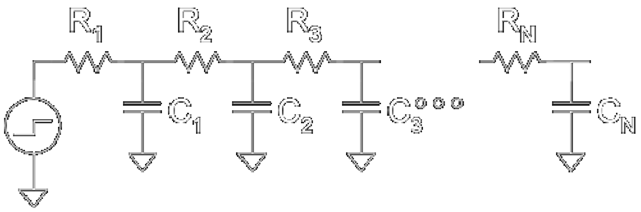
Delay Estimation Techniques

□ SPICE Simulation
- Very slow – not practical for chip level analysis
- Good for specific nets such as clocks or critical path

□ Asymptotic Waveform Evaluation (AWE)
- Is an industry standard for delay estimation
- uses moment matching to determine a set of low frequency dominant poles that approximate the transient response

□ Elmore Delay Analysis
- Uses only the first moment (dominant pole)
- Can be used for first order approximation in a complicated *RC* tree

Elmore Delay in RC Ladder

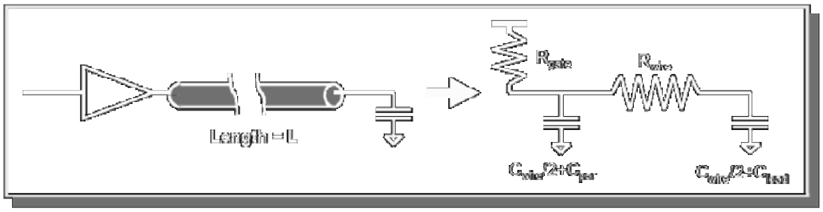


$$\tau_{Di} = \sum_{k=1}^{N} R_{ki} C_k$$

$$= R_1 C_1 + (R_1 + R_2) C_2 + \cdots + (R_1 + \cdots + R_N) C_N$$

Delay of Long Interconnect

- Delay of gate driving a long wire governed by RC time constants

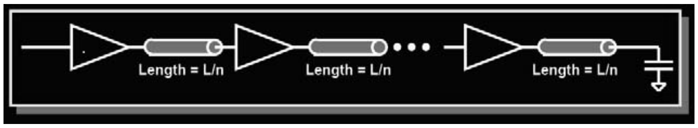


  - Elmore delay: $D = R_{gate}(C_{wire} + C_{par} + C_{load}) + R_{wire}(0.5C_{wire} + C_{load})$
  - Quadratic in total wire length

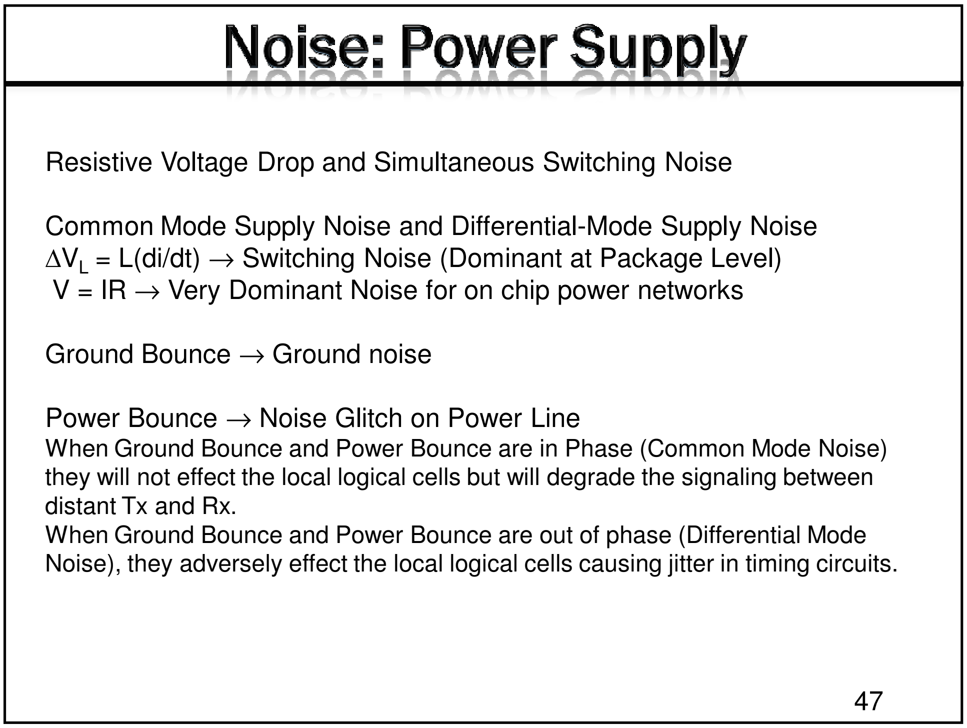
- For long wires, this delay quickly becomes untenable
  In a 65nm process, wire delay looks like 2-3*(gate delay)/mm2

Delay Reduction in Long Wires

- For slow long wires we use repeaters
  Gain stages that break up the wire and "refresh" the signal
  Inverters are the simplest gain stage



- Delay of a repeated line is linear in total length, not quadratic
  Delay is the geometric mean of the wire delay and the gate delay
  D = constant * sqrt(gate_delay * R_WC_W)



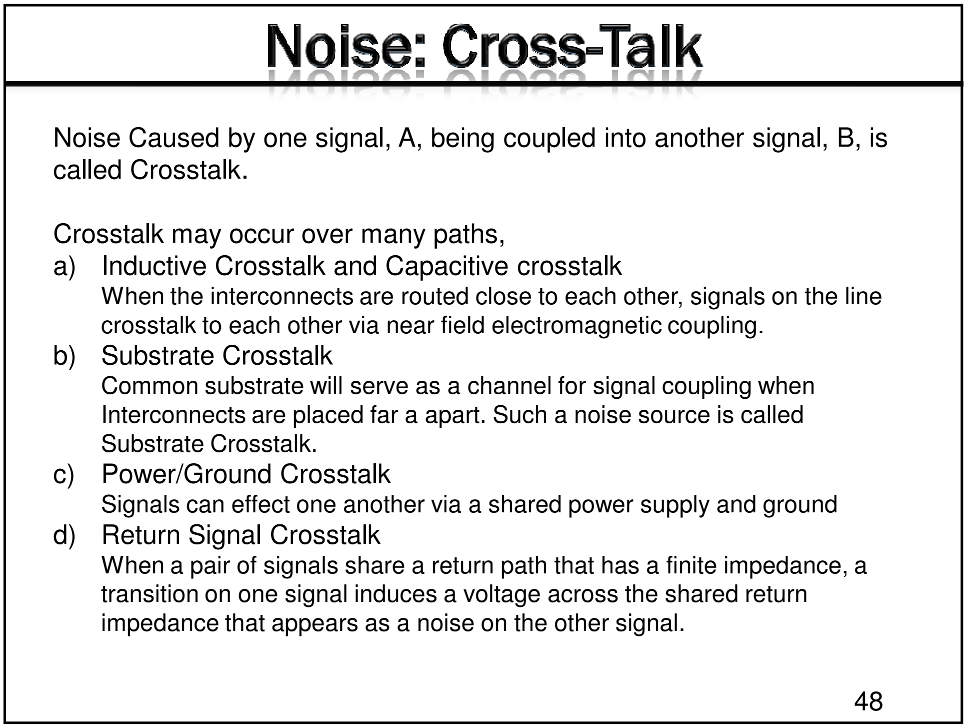
Noise: Power Supply

Resistive Voltage Drop and Simultaneous Switching Noise

Common Mode Supply Noise and Differential-Mode Supply Noise
$\Delta V_L = L(di/dt) \rightarrow$ Switching Noise (Dominant at Package Level)
$V = IR \rightarrow$ Very Dominant Noise for on chip power networks

Ground Bounce → Ground noise

Power Bounce → Noise Glitch on Power Line
When Ground Bounce and Power Bounce are in Phase (Common Mode Noise) they will not effect the local logical cells but will degrade the signaling between distant Tx and Rx.
When Ground Bounce and Power Bounce are out of phase (Differential Mode Noise), they adversely effect the local logical cells causing jitter in timing circuits.



Noise: Cross-Talk

Noise Caused by one signal, A, being coupled into another signal, B, is called Crosstalk.

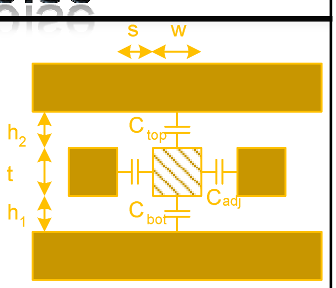
Crosstalk may occur over many paths,

a) Inductive Crosstalk and Capacitive crosstalk
   When the interconnects are routed close to each other, signals on the line crosstalk to each other via near field electromagnetic coupling.
b) Substrate Crosstalk
   Common substrate will serve as a channel for signal coupling when Interconnects are placed far a apart. Such a noise source is called Substrate Crosstalk.
c) Power/Ground Crosstalk
   Signals can effect one another via a shared power supply and ground
d) Return Signal Crosstalk
   When a pair of signals share a return path that has a finite impedance, a transition on one signal induces a voltage across the shared return impedance that appears as a noise on the other signal.

Interconnect Noise

- Wires are skinny and tall and have lots of sidewall capacitance

Aspect ratios at 2.2 now and are projected to scale up to 3-3.5
We will have to live with some coupled noise



- Traditional estimates use a simple capacitive divider

$$V_{noise} = \frac{C_{adj}}{C_{adj} + C_{top} + C_{bot}}$$

But this is pessimistic, because the "victim" is usually driven, too

- In reality, you must account for both victim and attacker drivers

$$V_{noise} = \underbrace{\frac{C_{adj}}{C_{adj} + C_{top} + C_{bot}}}_{\text{about 75\%}} \frac{1}{1 + \underbrace{\tau_a / \tau_v}_{2\sim4}}$$

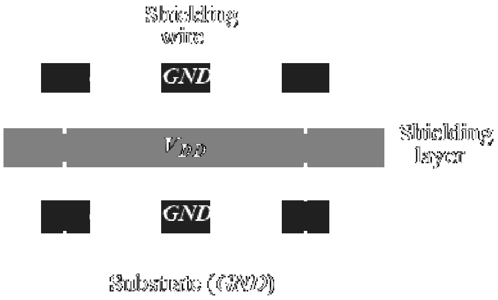
49

Solution ?

To avoid cross talk noise

Prevent parallel lines

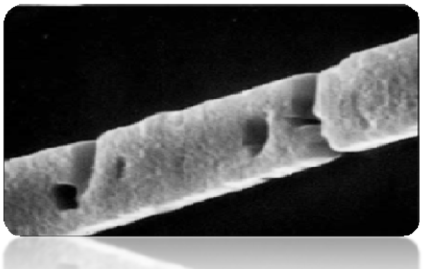
Shielding



50

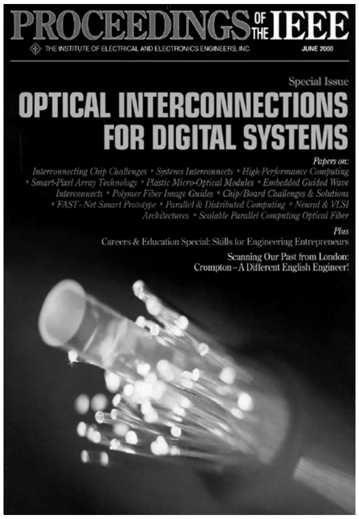
Electromigration

- Electromigration: Electrons smack into lattice, displacing atoms
  Caused by unidirectional current flow
  Wires with bidirectional current is “selfhealing”
  Copper’s MTTF is 5x better than Aluminum’s
- Highest at vias, where the current crowds from the vias
- Calculate max DC current, which depends on total capacitance
  Rule is “max current per wire cross section” (e.g., 1mA/μm^2)



51

Replacing Wires?



PROCEEDINGS OF THE IEEE

THE INSTITUTE OF ELECTRICAL AND ELECTRONICS ENGINEERS, INC. JUNE 2000

Special Issue

OPTICAL INTERCONNECTIONS FOR DIGITAL SYSTEMS

Papers on:
Interconnecting Chip Challenges • Systems Interconnects • High-Performance Computing • Smart-Pixel Array Technology • Plastic Micro-Optical Modules • Embedded Guided Wave Interconnects • Polymer Fiber Image Guides • Chip/Board Challenges & Solutions • FAST-Net Smart Prototype • Parallel & Distributed Computing • Neural & VLSI Architectures • Scalable Parallel Computing Optical Fiber

Plus
Careers & Education Special: Skills for Engineering Entrepreneurs

Scanning Our Past from London:
Crompton–A Different English Engineer!

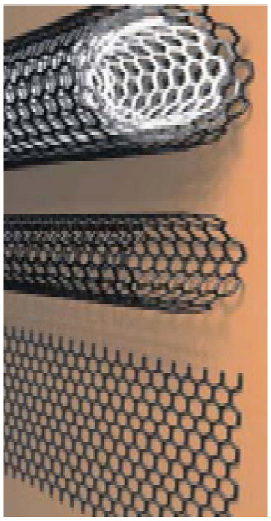
Optical interconnect

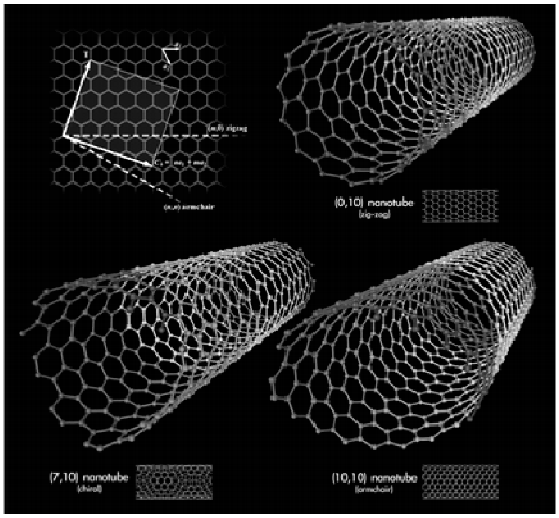
- Bandwidth
- Power
- Delay

52

Replacing Wire?

○ Sumio Iijima of NEC in 1991





CNT properties

Electrical:

- Ballistic conduction over distances of order 1 micron (~10^{-4} Ω·cm). ‘Metals’ with low resistivities, Semiconductors with high mobilities
- Conductivity a strong function of adsorbates or reactants.

Mechanical:

- High elastic modulus (high stiffness) (~1 to 5 TPa vs. ~0.2 for steel).
- Very high tensile strength (~10 to 100 GPa vs. ~1 for steel).

Thermal:

- High room temperature thermal conductivity (~2000W/mK vs. ~400W/mK for copper).

Electrical Stability:

- Maximum current density (10^9 A/cm^2 vs. <10^7 A/cm^2 for Cu).

Chemical Stability:

- C binding energy in graphene ~12 eV vs. Cu at a Cu surface ~ 4eV