Light matters

Optical Interconnects in high-performance computing systems

Rick Lytel
Distinguished Engineer
Sun Microsystems Laboratories

May 2000

Agenda

- Trends in microelectronic systems
- System scaling
- Requirement for optical technologies
- Ten year roadmap

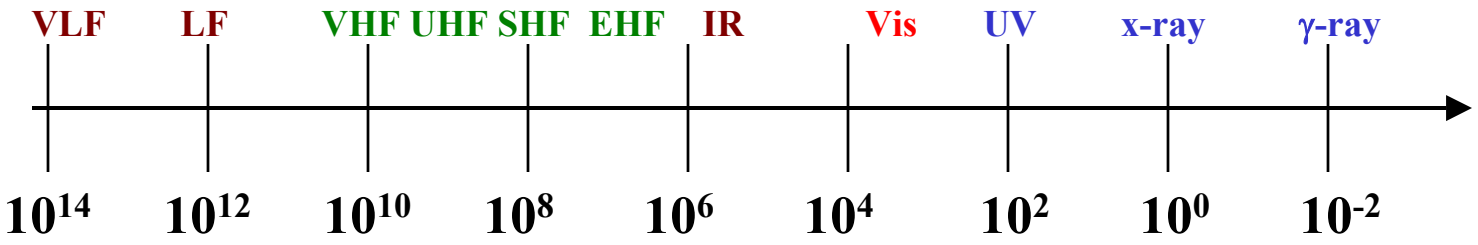
Why light matters…

- Can transmit 10 Gbps over 100's km w/no repeaters
- Can multiplex 100's of modulated wavelengths into a fiber
- One Sprint fiber carries $1M traffic revenue/hr
- One 125 μm thick fiber has ~ Terahertz bandwidth
- Optical amplifiers exist and are even used undersea
- It's cheaper now to install an undersea cable than to install a terrestrial system
- WW telecom market > $500B per year

Preliminaries

- What is light?
- Light sources
- Light carriers and media
- Light detectors
- An optical interconnect

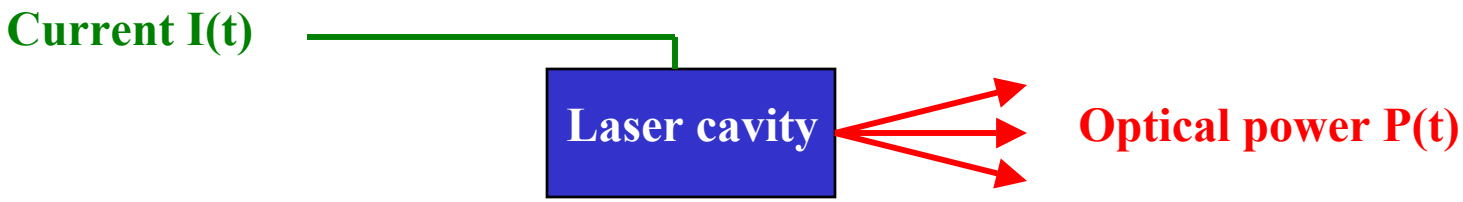
What is light?



- An electromagnetic wave with 0.4 μm < λ < 0.7 μm (visible)
- Ditto, but with 0.15 μm < λ < 0.4 μm (ultraviolet)
- Ditto, but with 0.7 μm < λ < 10.0 μm (infrared)
- A spin one, zero mass particle (photon)
- x-rays, visible rays, radio waves...

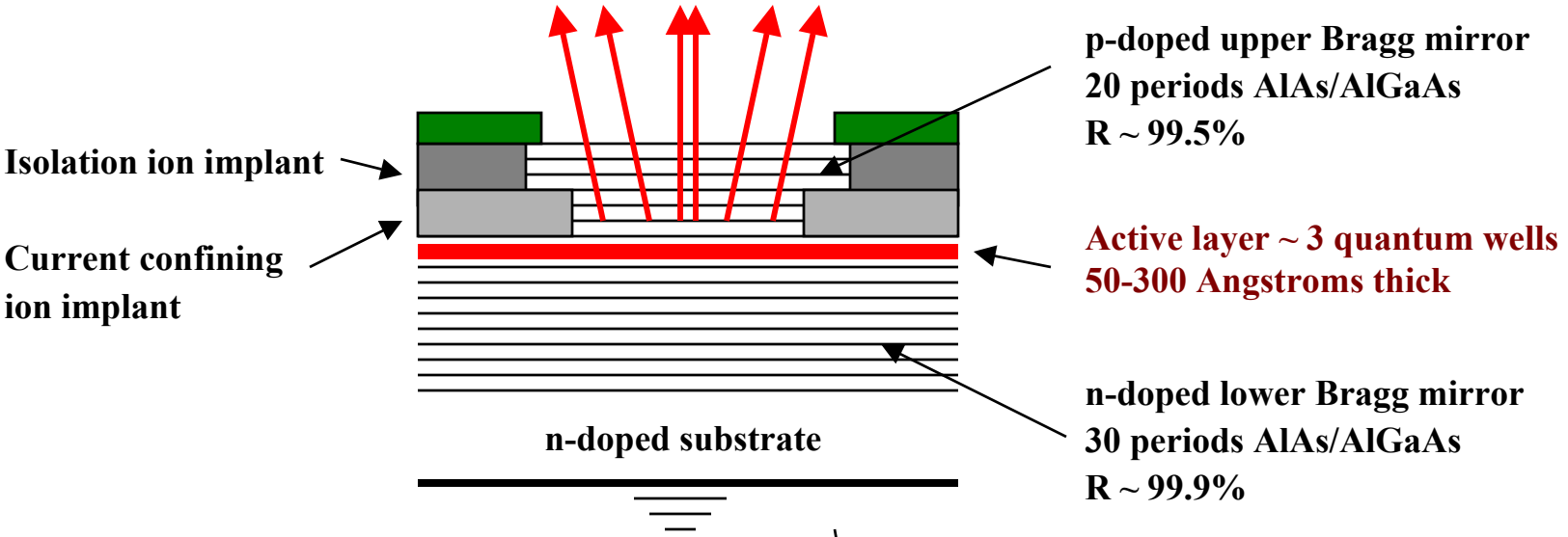
1 Angstrom = 10^{-10} meters

Coherent light sources

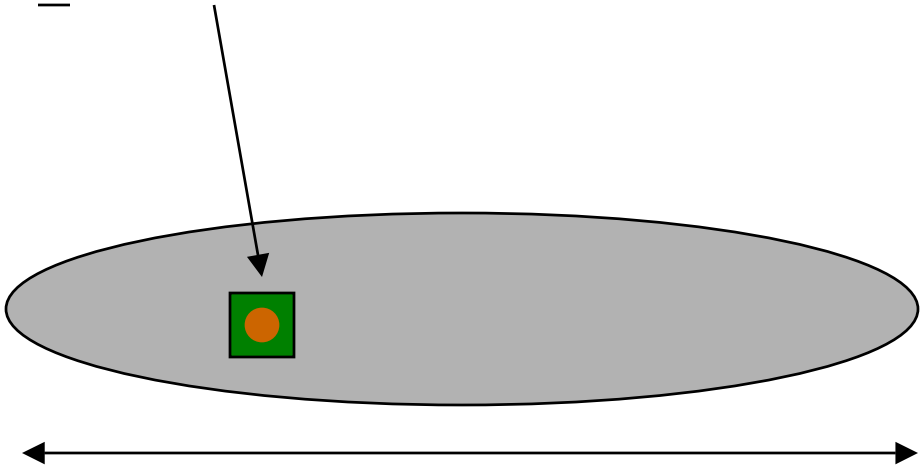


- Solid state, liquid, or gas cavities (lasing material)
- Deep UV through far IR
- Power outputs - 30 dBm to + 120 dBm
- Cavity lengths 1 μm to 10's m
- Aperture sizes 1's μm to 1's meters
- Pulse widths continuous to femtoseconds
- Repetition rates DC - 40 GHz
- Erbium-doped fiber: Coherent amplifier

Vertical cavity surface emitting laser (VCSEL)



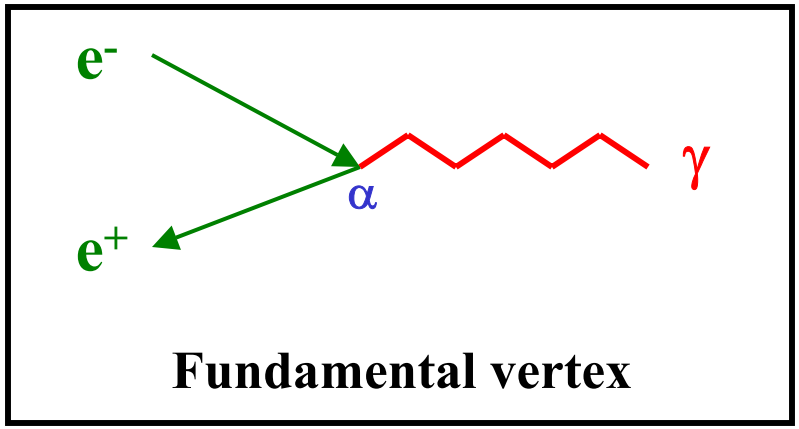
- FIT rate ~ 10
- 10 μm aperture
- few mA @ few volts
- 850 nm emitter



3” GaAs wafer with 20K devices

Quantum Electrodynamics:
The Theory of Photons and Electrons

- Relativistic quantum field theory
- Basic interaction has strength $\alpha \sim e^2$
- Perturbative in $\alpha \sim 1/137$
- No γ- γ interactions so scattering = 0 to order α^2

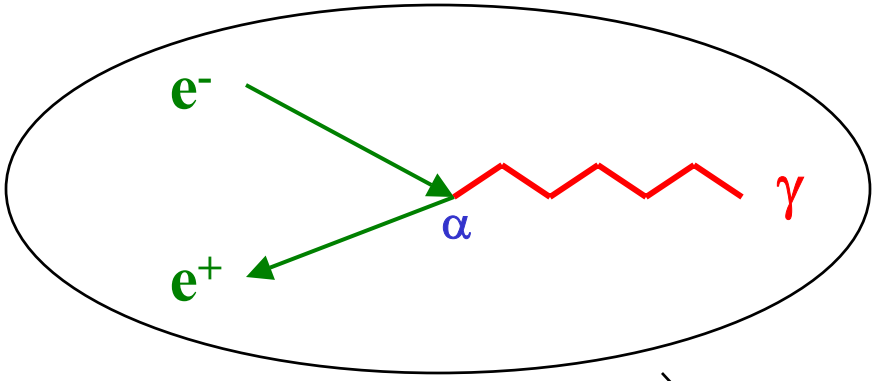


Fundamental vertex

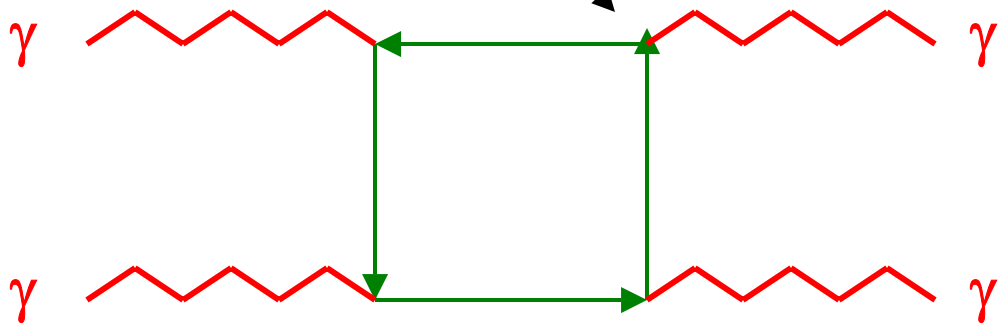
Anomalous magnetic moment of electron

- a_e (theory) = 1 159 652 216.0 (68.) x 10^{-12}
- a_e (exper't) = 1 159 652 188.4 (4.3) x 10^{-12}

Light-light scattering is effectively zero



- Use vertex to construct Feynman diagram
- Convert to mathematics
- Evaluate to get result

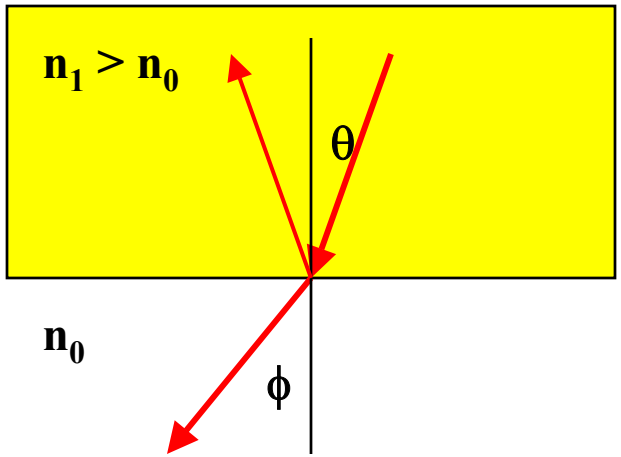


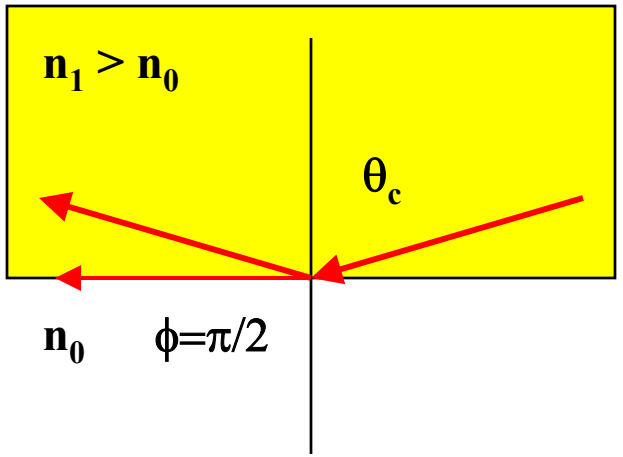
= 0

Vanishes by gauge invariance so γ- γ scattering = 0 to order α^4 !!

Light-material interactions (linear)

- Maxwell equations (classical)
  - $P = \chi^{(1)} E$ (linear susceptibility)
- Maxwell + quantum mechanics (semi-classical)
- Quantum electrodynamics (quantum field theory)





- Snell's law: $n_0 \sin\phi = n_1 \sin\theta$
- Frequency unchanged
- Reflection & transmission

- Critical angle: $n_0 = n_1 \sin\theta_c$
- Frequency unchanged
- Waveguiding

Light-material interactions (nonlinear)

- Maxwell equations (classical)
  - $P = \chi^{(1)} E + \chi^{(2)} E^2 + \chi^{(3)} E^3 + \ldots$

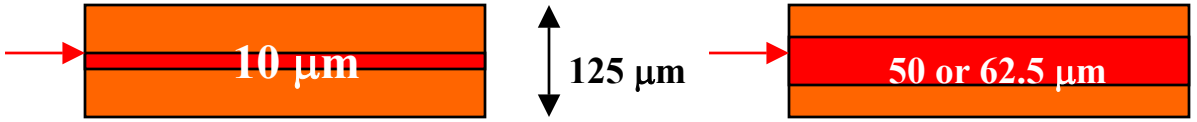
    (nonlinear susceptibility)
- Linear optics, plus some novel effects
  - frequency doubling, tripling
  - optical bistability, switching
  - phase conjugation

- *make new frequencies*
- *undo the twinkle in the stars*
- *femtosecond optical switching*

Light-carrying media

optical fiber

- glass
- polymer

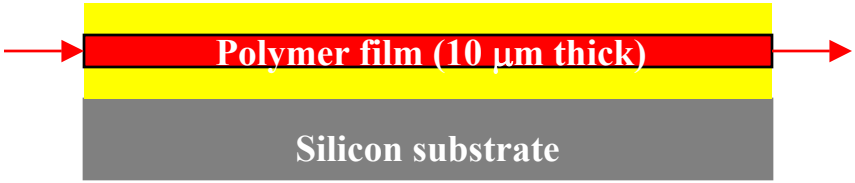


Single-mode
100 GHz-km
few $$/meter

Multi-mode
1 GHz-km
few $$/meter

optical films

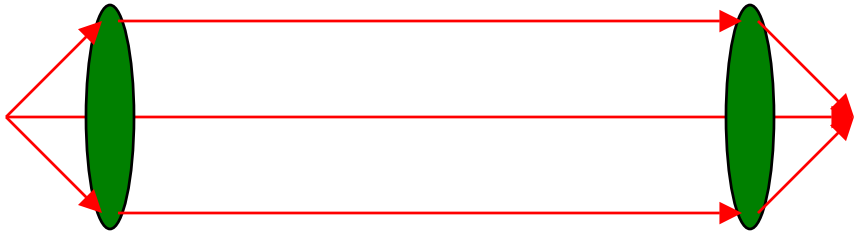
- glass, polymer
- crystal, liquid



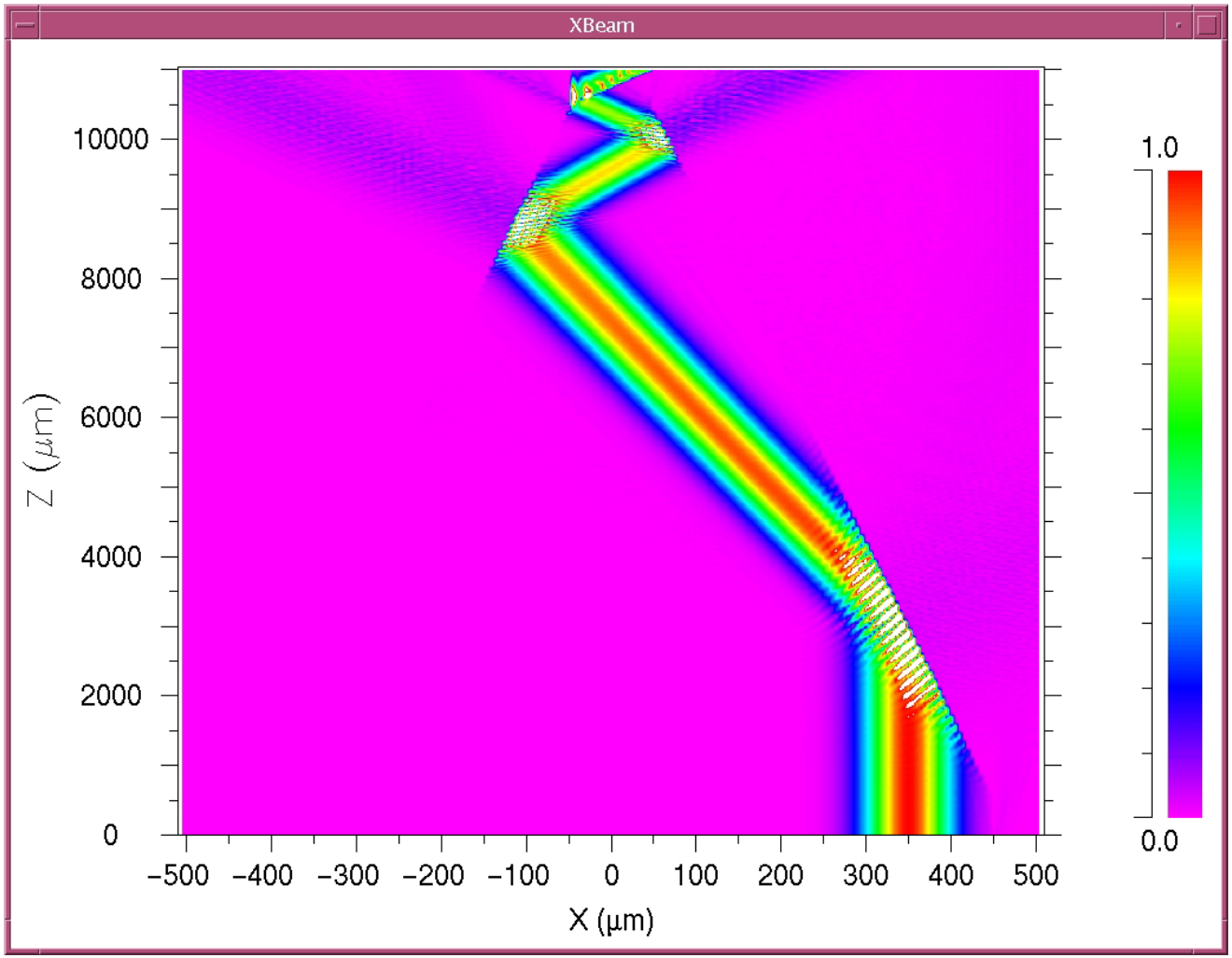
Single or multi-mode
15 cm max extension

free-space

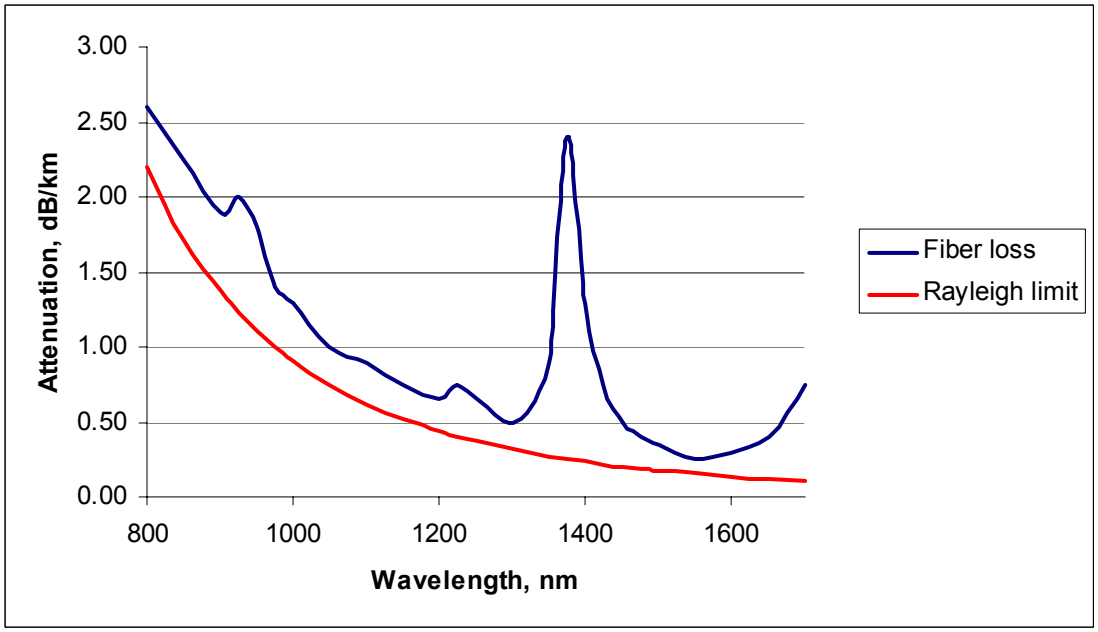
- air
- vacuum



You can bend light with glass



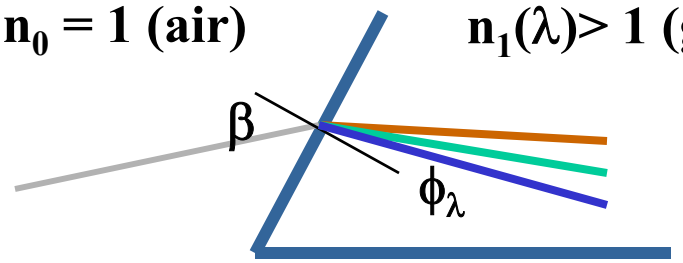
Amazing properties of silica fiber



- Three wavelength windows with a semiconductor laser for each!
  - AlGaAs-GaAs 0.85 μm, InGaAsP-InP 1.3, 1.55 μm (and InGaAs-GaAs 0.98 μm)
- Terahertz optical bandwidth
- $1 per meter, Lucent and Corning are dominant suppliers
- Easily pulled and spliced into 1000 km lengths, and connectorized

Wavelength division multiplexing (WDM)

- "Prism-like" device
- Collimate λ's (mux)
- Separate λ's (demux)



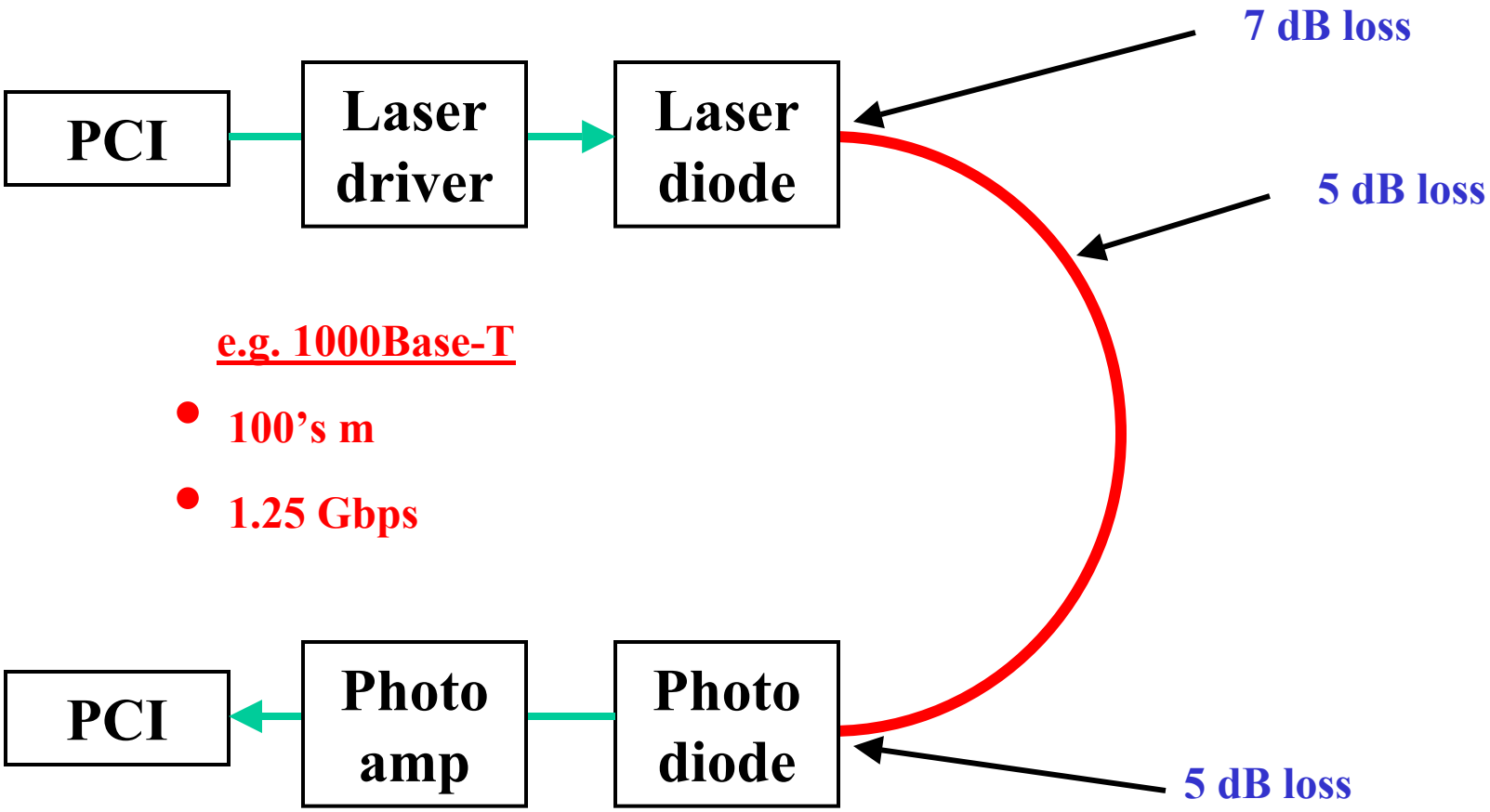
$$\sin\phi_\lambda = (n_0 / n_1)\sin\beta$$

- Telecom uses 1550 nm window w/amplifiers @ 0.8 nm channel spacing
  - "dense WDM" with 80-200 channels @ 2.5 Gbps per channel
  - temperature stabilized, edge-emitting, distributed feedback lasers
  - dense glass or fiber bragg grating mux/demux devices @ $700 per channel
- Datacom uses 850 nm window w/o amplifiers @ 10 nm channel spacing
  - "coarse or wide WDM" with 4-8 channels @ 2.5 Gbps per channel
  - air-cooled VCSELs, static tuning through materials processing
  - glass waveguide mux/demux @ $150 per channel

Light detectors

- pn, pin, avalanche diodes
- metal-semiconductor-metal (MSM) detectors
- apply photons, get currents
- Silicon 0.6-0.9 microns
- InGaAs 0.8 - 1.6 microns
- efficiencies range from 0.1% to near unity

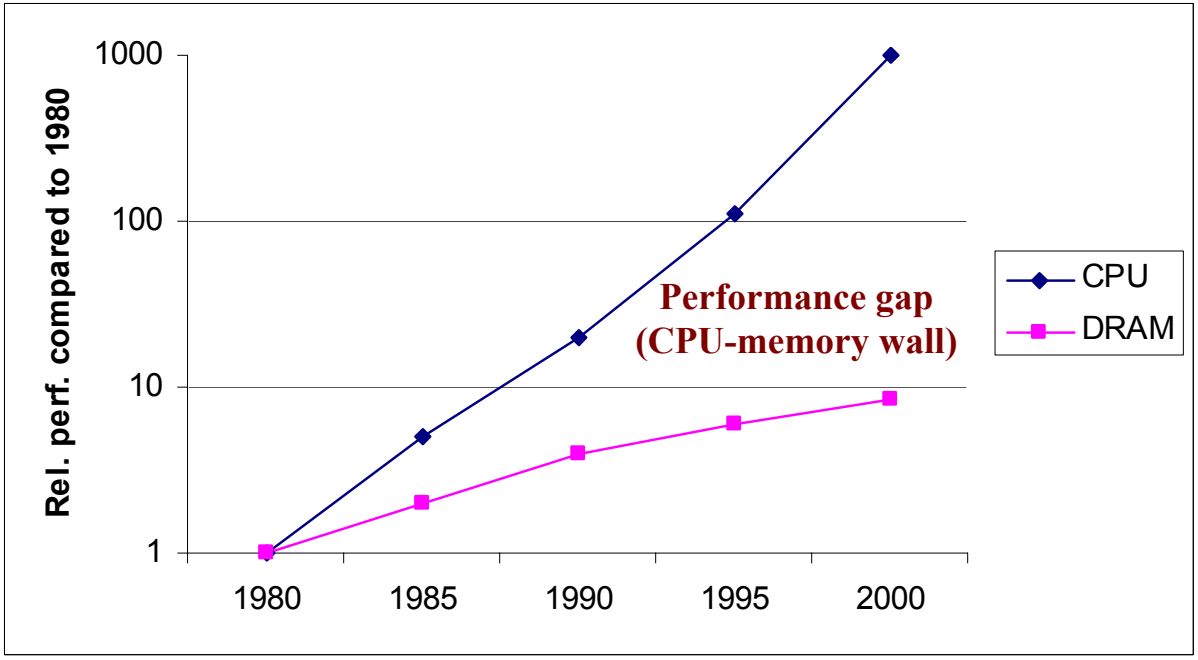
A fiber-optic interconnect



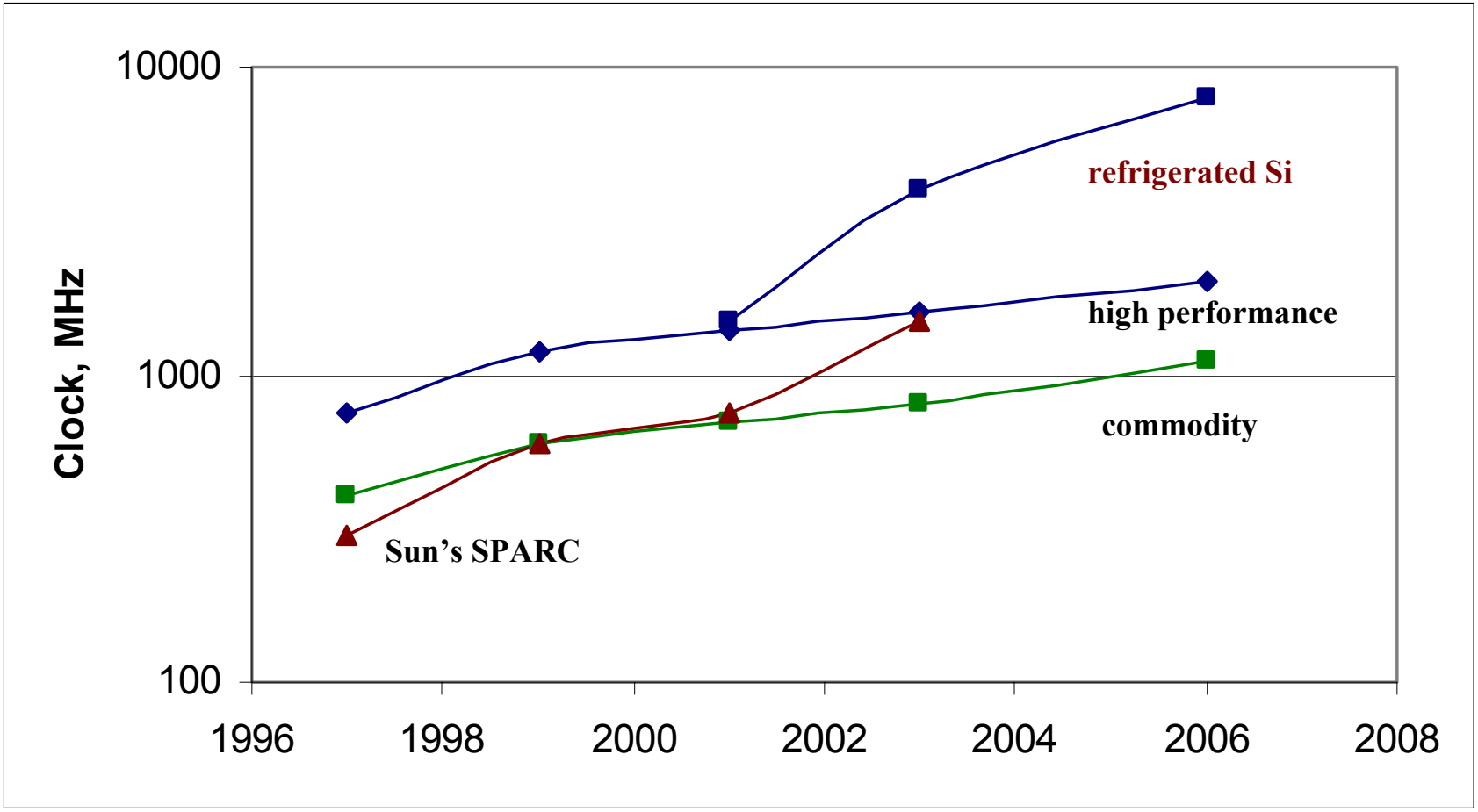
Typical link has -3 dBm in, -20 dBm out, 10^{-12} BER

Microelectronics: Year 2000 looking backward

- Si industry has process metrics that double every t_c months
  - e.g., clock frequencies $f_c \sim 2^{t/tc}$ ("Moore's law")
  - DRAM limited by 10% growth in access time
  - CPU-memory system performance does not scale

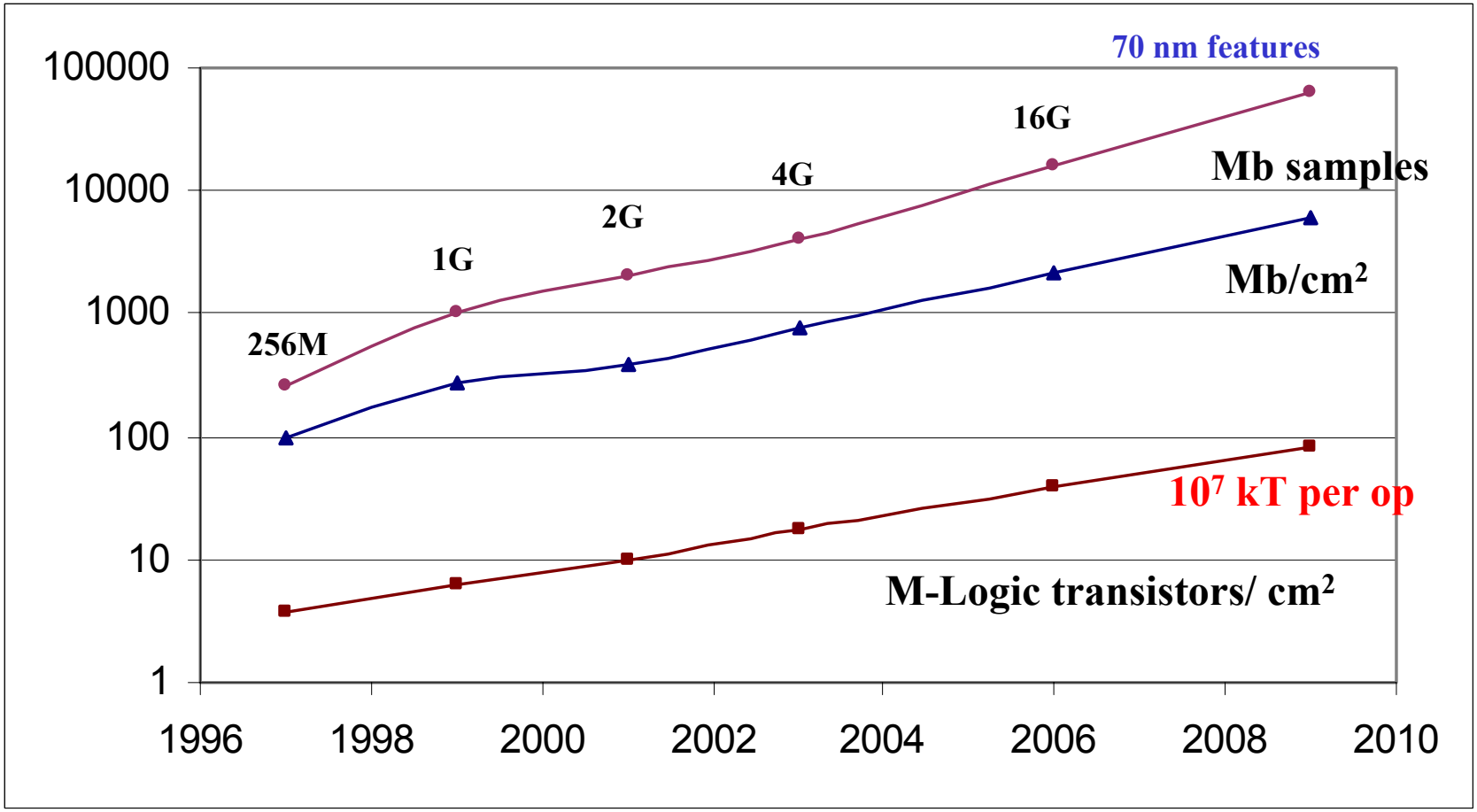


Processors:
Year 2000 looking forward*



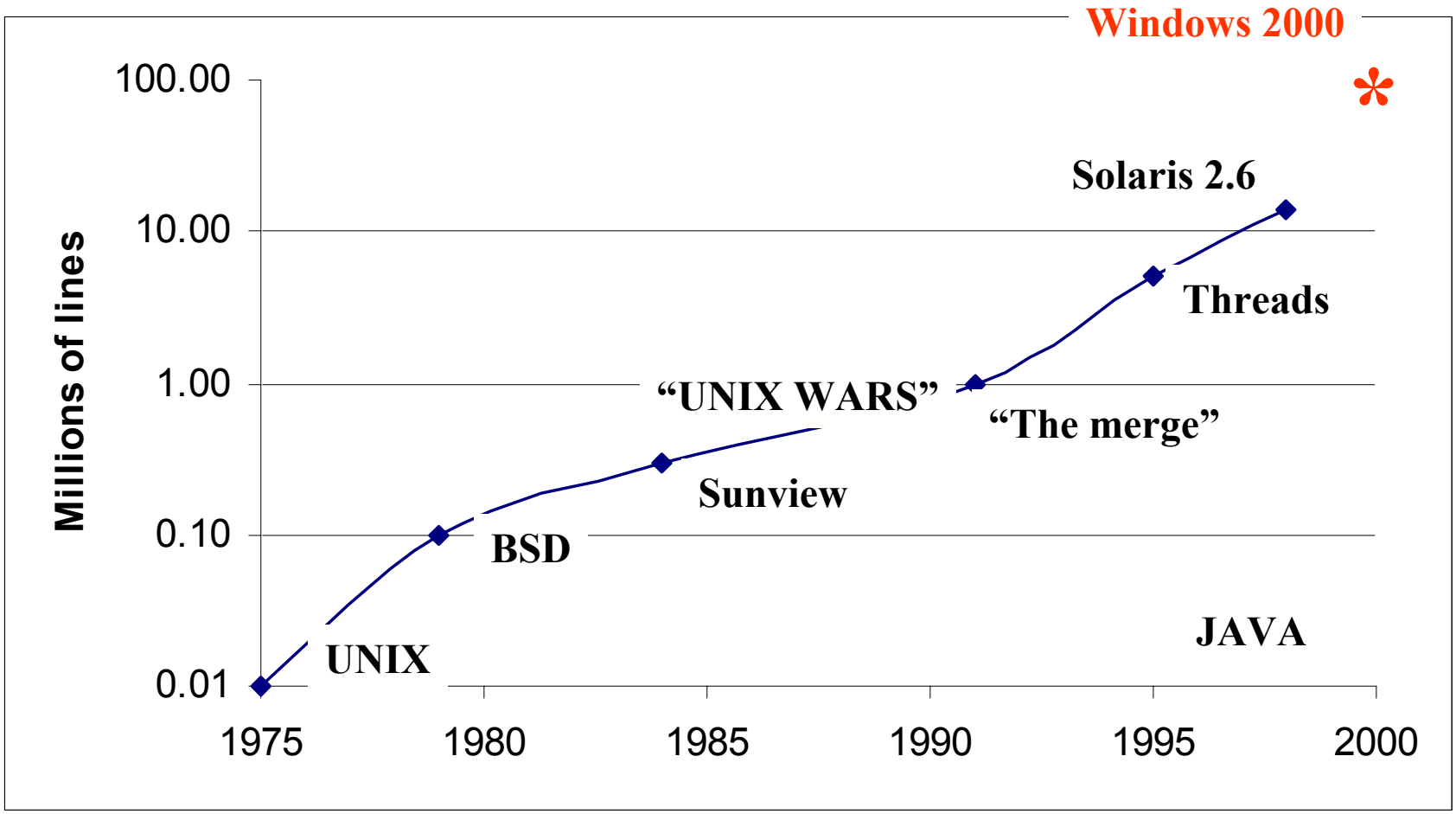
**1997 Semiconductor Industry Association (SIA) Roadmap*

VLSI:
Year 2000 looking forward*



**1997 Semiconductor Industry Association (SIA) Roadmap*

Lines of code



Scaling implications 1: engineering

- System design is stressed at all scales (μm-km)
  - > 5 metal layers on dice
  - 1000's pin packages
  - > 20 layered processor boards and centerplanes
- Novel subsystems implemented
  - Rambus
  - multithreaded architectures
  - differential signaling
- How much headroom remains at each scale?

Scaling implications 2: industry

- Computer hardware engineers get scared
  - known methods nearly broken
  - new methods aren't known and familiar
  - 'rock and a hard place'
- Researchers gets excited about the opportunities
  - *New technologies are required*
  - *These must be commercial*
  - *Do a startup--everyone else is...*
- But can things really change inside the box?

Microelectronics: Like biological evolution

- Early successes form (industry) substructure
- New developments evolve from and overlay prior success
  - architecture (memory, IO, CPU)
  - instruction sets (coded lines exponentiate)
  - process (phenotype development)
- Abrupt mutations usually die (multilevel logic)
- Broad adaptation in niche environments (MCM, GaAs)
- Reliable and tested methods persist forever (HOX genes)
- Change, when it comes, is usually dramatic (species)

A new challenge: Optics in computers

- Optics is in the telco network and in LAN
- Internet traffic is driving up server sales
- Processors run faster and hotter
- Opportunities abound
- Start a company!

The $64,000 (or better, 64,000 share) question:

Should you quit your day job?

Scaling gives clear messages...

- Silicon CPUs to 2010
- Processor-memory gap gets wider
- System growth will change to smaller OS's
- Networking, not computing, is market driver

Scale Silicon systems until you hit the brick wall*

** e.g., 5 Si atom thick oxide, D.A. Muller et. al, Nature 399, 758 (1999).*

Computer scaling according to…

EE	***Shrink Silicon process and lower voltage***
ME	***Refrigerate computer, then do what EE does***
Optics	***Photons, not electrons, in interconnects***
Chemist	***Organic molecular wires & logic***
Biologist	***Nucleic acid logic & processor, PCR chips***
SS physicist	***Carbon nanotube gates, HT superconductors***
Theorist	***Quantum computing: it's been demo'd, QED***
Grad student	***Can I get a job?***
Marketing	***"The network is the computer"***
Customer	***More for less money***
SysAdmin	***More for less work***
Al Gore	***"When I invented the internet…"***

Hardware prediction

- Hotter processors...
- Faster and fatter interconnects...
- … at lower cost per GB/s

MEANS

- New “superscale” solutions will appear

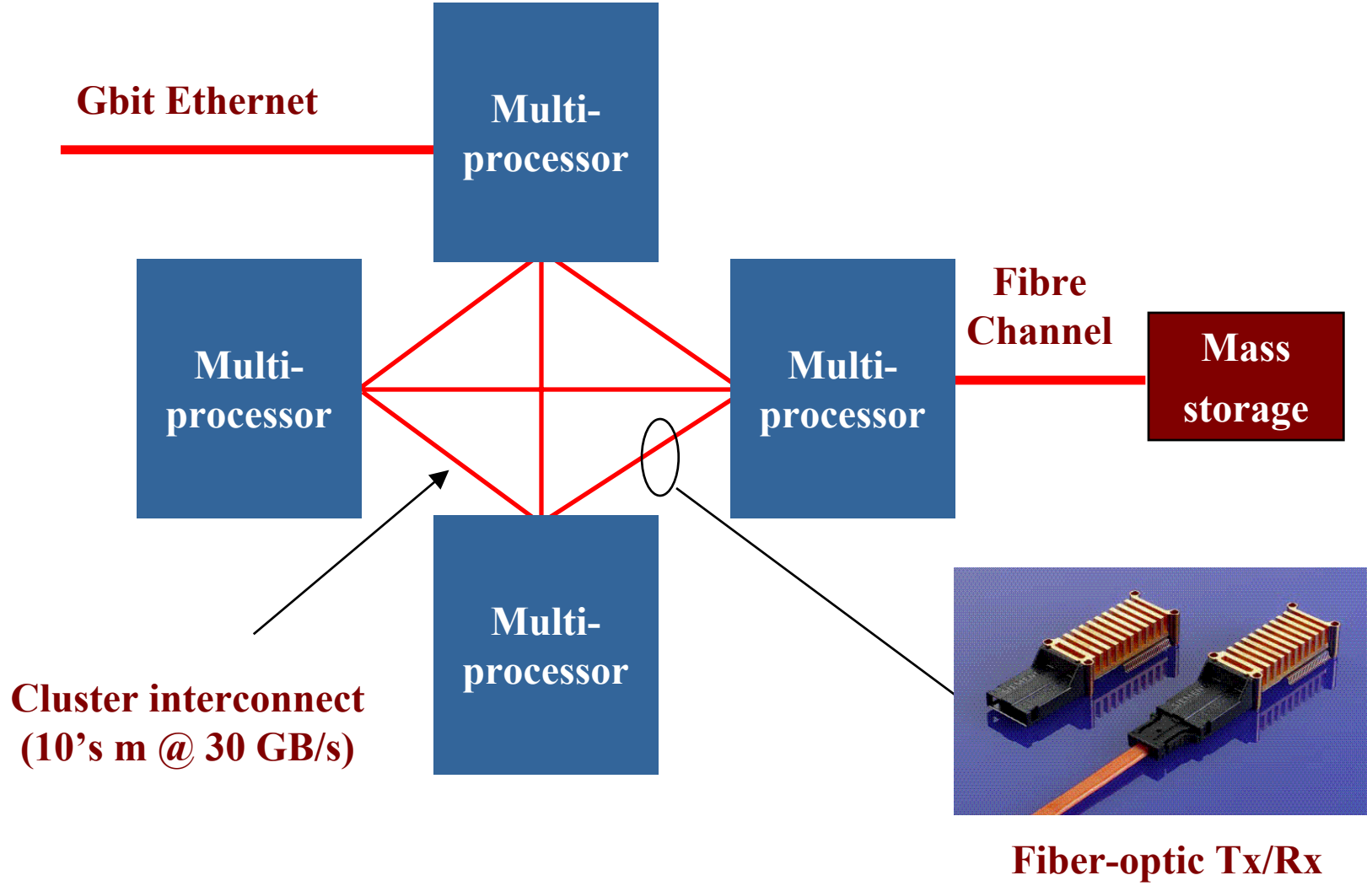
What will they be?

The interconnect hierarchy

(The network is not really the computer)

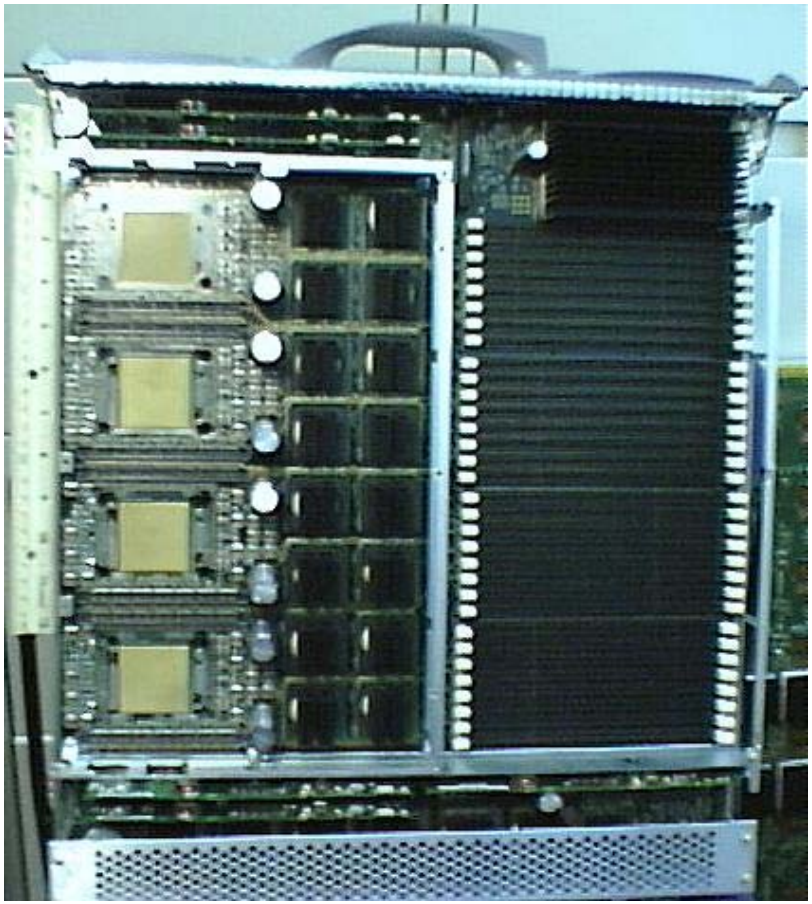
Circuit	Distance	Speed	Width	Link BW	Carrier	
Gate-gate	1-100 μm	1 GHz	100's	100 GHz	e^-	"the computer"
Chip-chip	1 cm	500 MHz	100's	50 GHz	e^-	
Board-board	10-100 cm	500 MHz	100's	50 GHz	e^-	
Cabinet-cabinet	1-10 m	2.5 GHz	10's	25 GHz	$h\nu$	
Floor-floor	10-100 m	100 MHz	1's	0.1 GHz	$h\nu$	"the network"
Campus	100-1000 m	1 GHz	1's	1 GHz	$h\nu$	
Intracity	1-10 km	2.5 GHz	1's	2.5 GHz	$h\nu$	
Intercity	10-100 km	2.5 GHz	10's	25 GHz	$h\nu_k$	
Continental	100-1000 km	10 GHz	100's	1 THz	$h\nu_k$	
Intercontinental	1000-10000 km	10 GHz	100's	1 THz	$h\nu_k$	

Optics in large systems today



Modern multiprocessor machine

centerplane

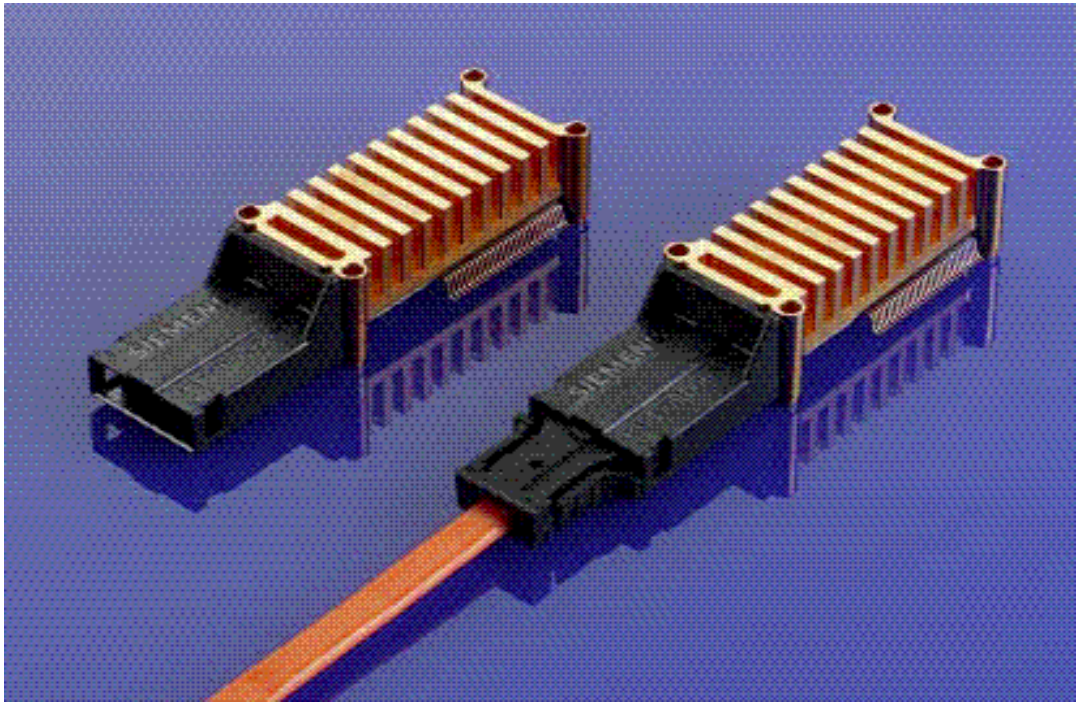


blade (1 of 18)

Copper and Fiber

- These two cables have the same bandwidth
- The ruler is 6”
- Big cable is 160 pair, 83 MHz LVDS, up to 10 m
- Small cable is 12 fiber, 1.25 Gbit/sec/fiber multimode ribbon, up to 100 m
- Latest versions of fiber are 2.5 Gbit/sec/fiber

Parallel optical transceivers



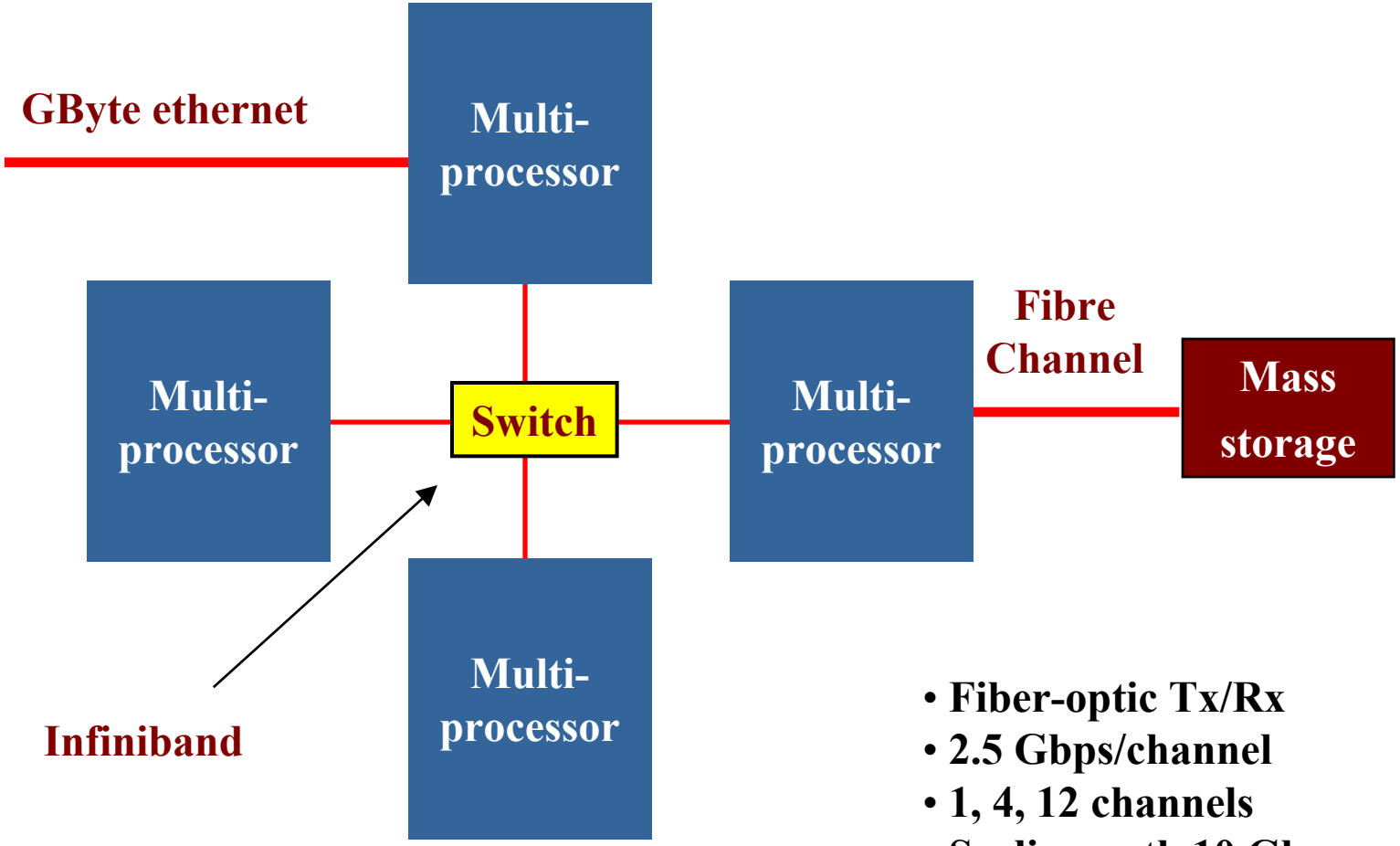
12 x 1.25 Gbit/sec Transceiver

Infineon Paroli Modules

Parallel optical transceivers

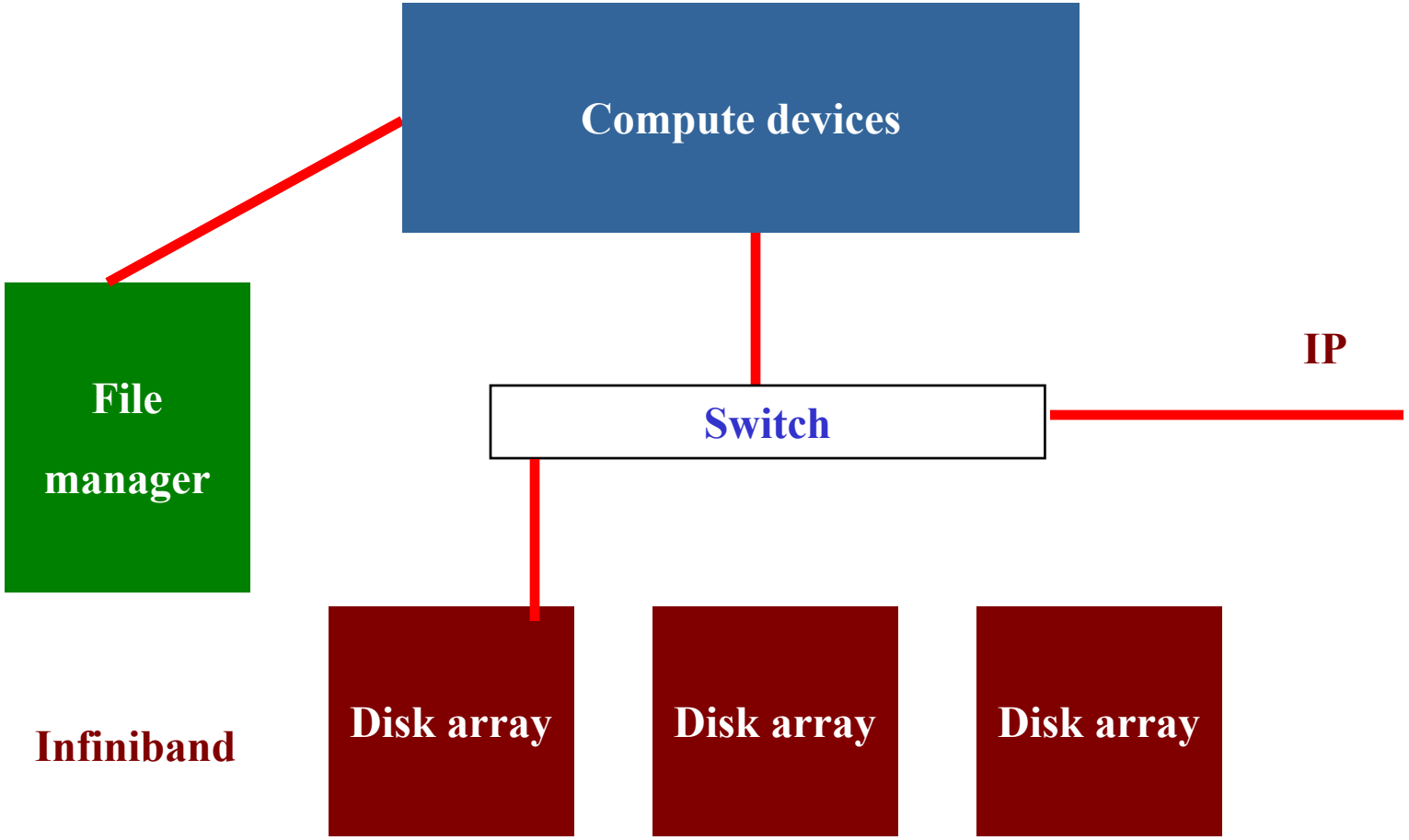
10 and 12 Channel, 2.5 Gbit/sec/fiber transceivers
Optobahn

Scaling of multiprocessor machines



- Fiber-optic Tx/Rx
- 2.5 Gbps/channel
- 1, 4, 12 channels
- Scaling path 10 Gbps
- Limited (or no) WDM

Scaling of computer systems



VLSI subsystem trends

- < 0.1 μm lithography
- > 2000 pads per die
- < 1 volt
- > 100 W/cm^2 die

Opportunity for optics?

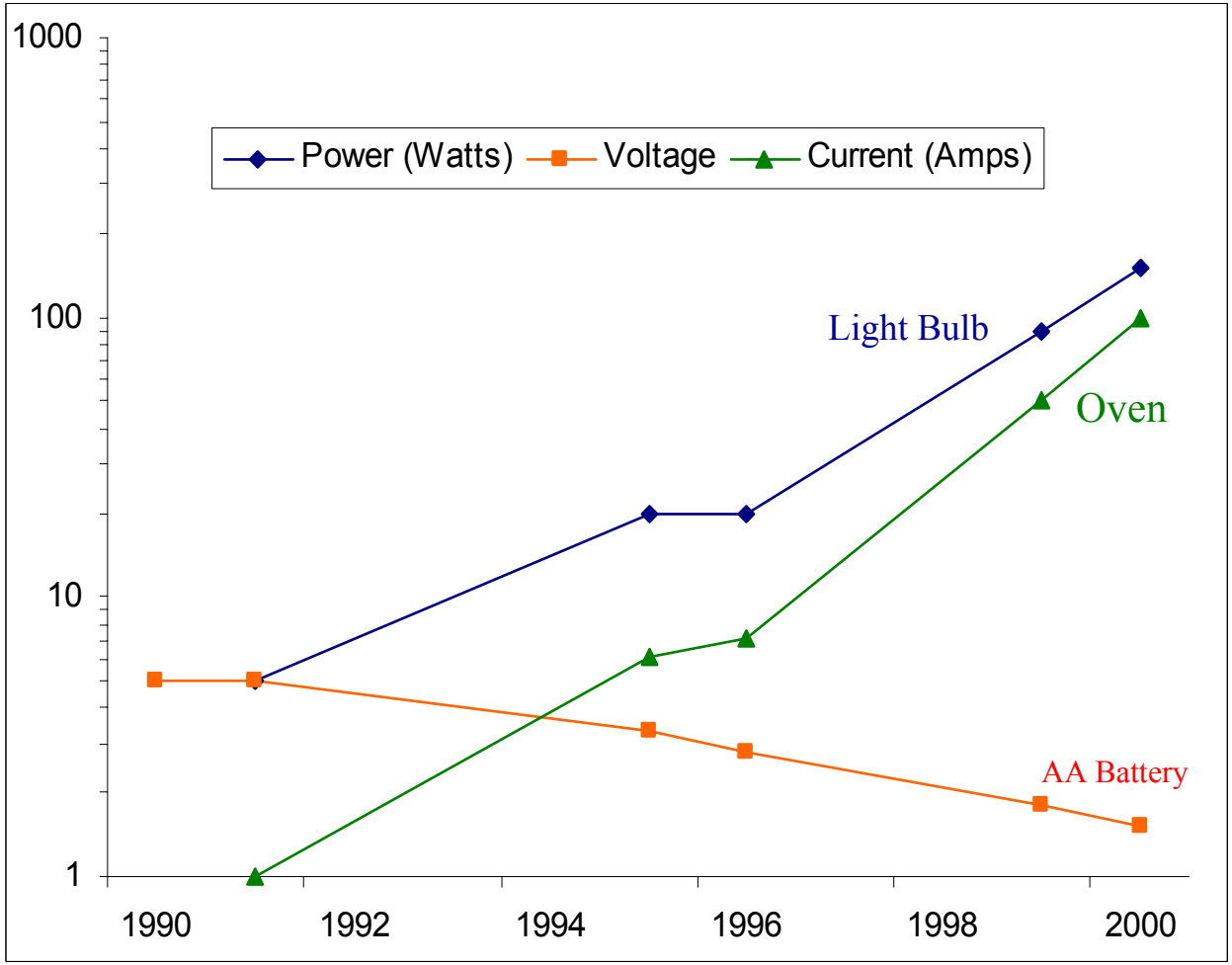
Optics won't integrate on chip with Silicon

- Optical elements are much larger than VLSI elements
  - 2-10 μm VCSEL and 20-50 μm PD apertures
  - 5-10 micron optical waveguides
- VLSI device elements are getting smaller
  - 0.18 micron features, micron-sized gates
  - 70 nm features are only a few years away
- Utilization of third dimension requires known good die
- Hybrids are way too strange for this industry

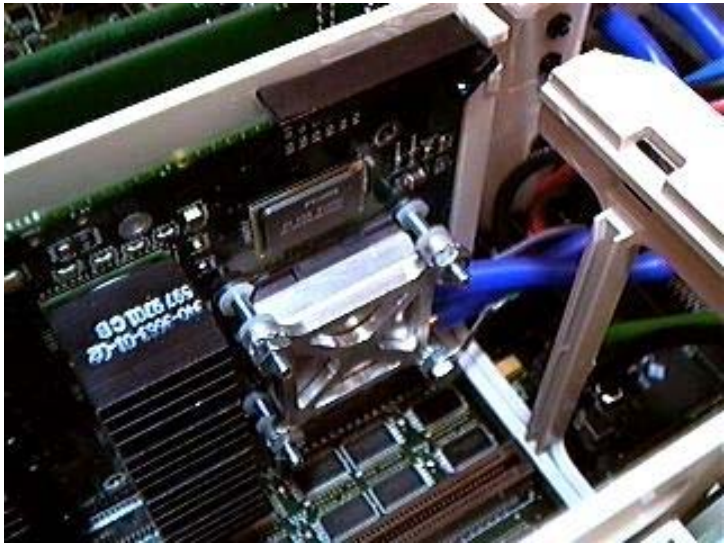
This could change if...

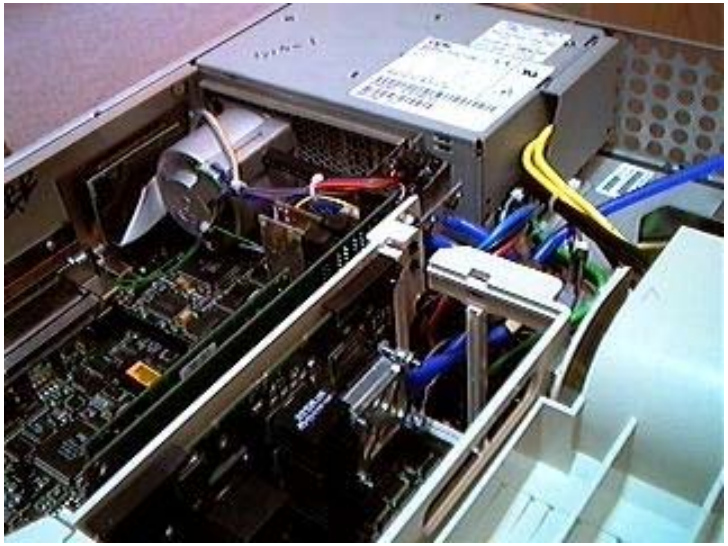
- Intel and IBM say it can, and develop own hybrids
- Standards evolve (very hard)
- Ultralow threshold devices (UCSB) are commercialized
- VCSEL suppliers provide devices @ $0.20 per pin
- CAD tools incorporate VCSEL/PD cells
- Electrical engineers learn optics

And even though CPUs get hotter...



…Cooling, not optics, will be implemented





3 years accumulated run-time

on bare, flipped US-II CPU

in Sun workstation

"Superscale" technologies

Inside the box (packaging & process)

- Cooling
- Low voltage CMOS
- Merged logic and memory
- Asynchronous circuits & systems
- SiGe HBT @ 30K gates, f_T ~ 75 GHz
- Free-space or fiber board-board links

Outside the box (transport and switching)

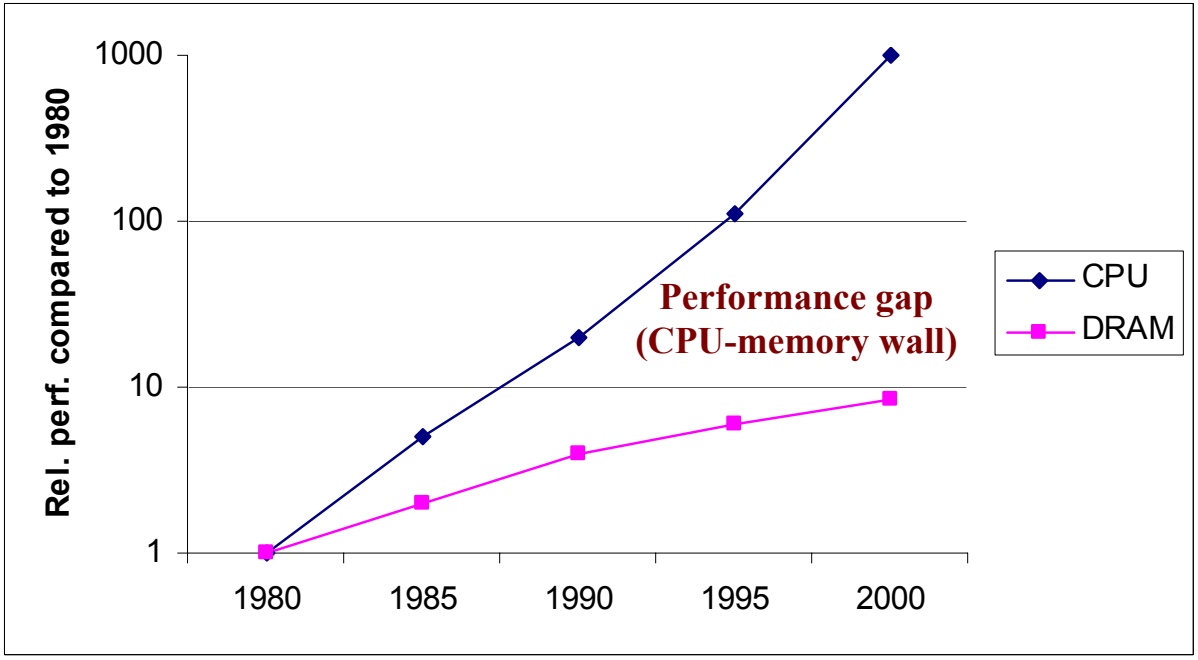
- TB/s optical links
- Cluster optical switches
- System area network

Optical switch applications

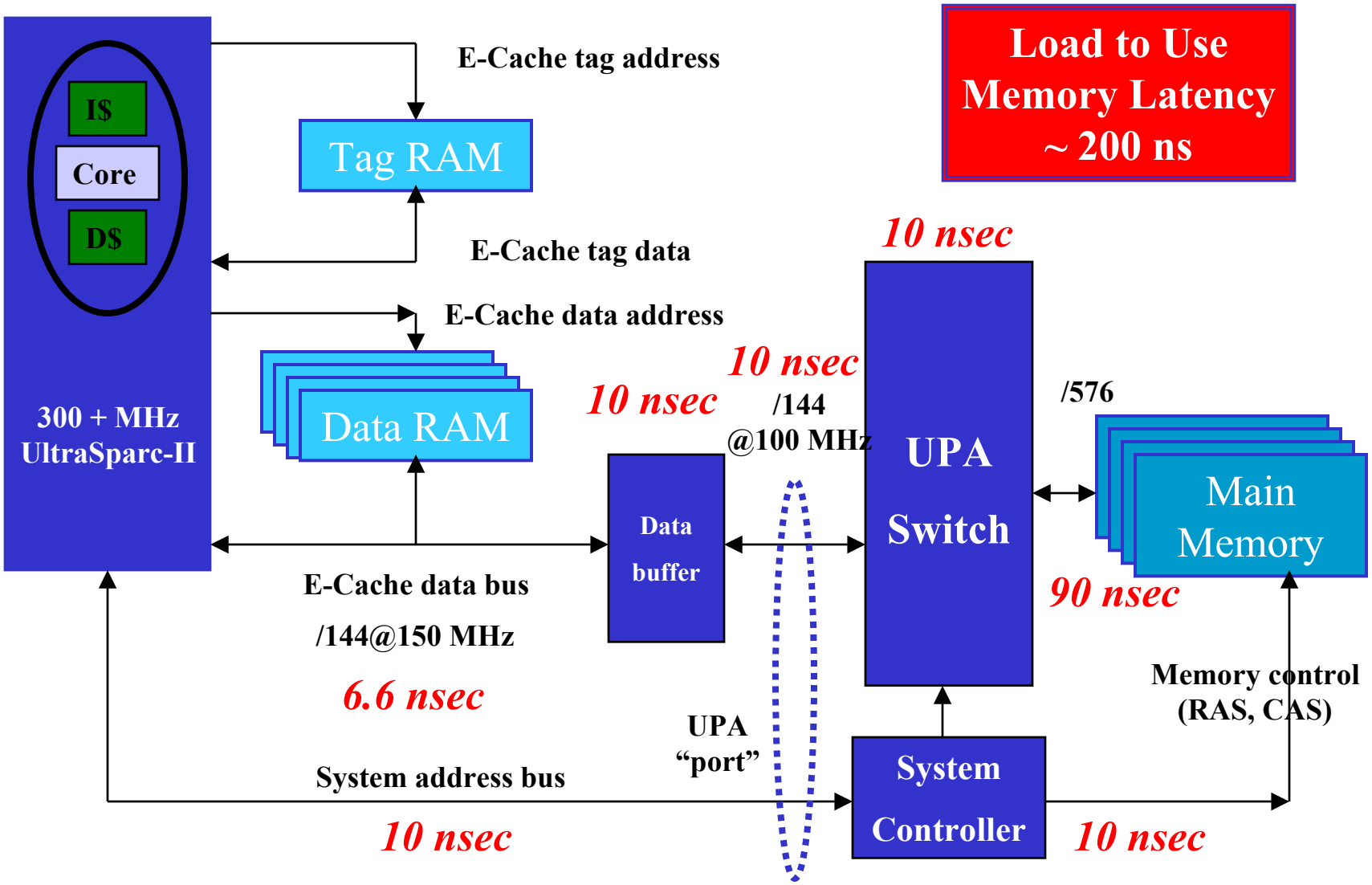
- Shared memory system domain reconfiguration
  - non-blocking matrix, N ~ 16, “slow and cheap”
- Shared memory packet switch
  - non-blocking matrix, N ~ 16, subμsec
- High-resolution, 3-D video graphics distribution
  - 1x100, 2x100, “slow and cheap”
  - 6 Gbps over 100’s m (short wavelength MM)

The CPU-memory wall: *The* serious performance limitation

- Cache model uses SRAM (L1, L2, L3) with 2-10 clock tick latency
- SRAM limited to < 1 MB in 2000
- Applications branch frequently, generating cache misses (& latency)



UltraSparc cache memory system



Closing the CPU-memory gap

- Merged logic and memory
  - fastest memory next to the CPU (practical)
- Hide latency
  - multithreading, prefetch, out-of-order execution
- Larger SRAM
  - 20 MB/die in 5-7 years
  - molecular electronics (10-20 years)
- Faster DRAM
  - custom memory modules (way $$$)
- Optical interconnects? No...
  - *Like real estate: location, location, location*
  - *Latency due to cache model, not wires*
  - *Increase, not decrease power, latency, noise*

Trends in CPU board interconnects

- Large boards with lots of processors, ASICS, and memory
- Cabinets with multiple boards plugged into centerplanes
- Very stressing board-level interconnects
  - Lower voltages
  - Wide data paths
  - Differential signaling
- Long (> 10”) traces of above to board edge

Next step is to use free-space optics between boards and blades, and eliminate the board edge

Sun's optics program

- Outsource parts, modules, links
- Internal demonstration of new, hard things
- Drive university IP toward industrial manufacturing
- Cause suppliers to expand and improve their efforts
- Drive components to standards
- Guide our product groups

Collaborators

UC San Diego	free-space link
UC Santa Barbara	980 nm monolithic VCSEL/PD/lens arrays
U. Pittsburgh	Chatoyant photonic CAD
U. Delaware	32 channel, 1 Gbps CMOS VCSEL drivers
UCLA	CDMA optical transmission
UC Santa Barbara	100 GHz Si circuits
U. Illinois	VCSEL drivers and CAD
Princeton	Terahertz optical switch (TOAD)
Infineon	12x1.25 Gbps, 12x10 Gbps links; SiGe
Honeywell	850 nm VCSEL arrays, MSM's
Applied Photonics	free-space interconnect hardware
LLNL	8 channel wide WDM
DARPA	we consult to their programs

1-3 year picture

- **Optical transport "outside the box"**
  - 2.5 Gbps lasers, probably 850 nm VCSELs
  - point-point fiber arrays (transport, transport, transport…)
  - 2.5 Gbps circuits (blaze, blaze, blaze…)
- **VLSI, packaging, and wire scaling "inside the box"**
  - lower voltage, higher density, cooling
  - differential signaling
  - no optics

4-7 year picture

- **Optical transport and switching "outside the box"**
  - **Optical links to LAN, storage, cluster, IO (system area net?)**
  - **10 Gbps transport links with some WDM (video, SAN)**
- **Free-space parallel optics "inside the box"**
  - **Array modules, blade-blade (1 Gbps x 256) interfaces**
  - **Drive to standard**
  - **Use with telecommunications switch & router boxes**

8-10 year picture

- **Free-space, multi-blade, multi-processor systems**
  - high optical bandwidth wherever it's needed in the box
  - smaller subsystems (CPU) but many more of them
  - fast, wide area interfaces
- **The network is the computer**
  - Everything's a router and a computer except termination devices
  - Everything's optically interconnected outside the Si chip
- **First explorations into nano-memory and system design**

Sun's "nanotechnology" program

- Collaborate with leading organizations
- Sponsor research to spur developments
- Guide researchers toward "practicality"
- Identify & solve physics & architecture issues
- Low level-of-effort, but has CTO support

Collaborators

- **UC Santa Barbara** — **quantum dot qubits**
- **Wash. U. St. Louis** — **carbon nanotube processing**
- **Yale University** — **conjugated oligomers & devices**
- **Penn State Univ.** — **molecular self-assembly**
- **Rice Univ.** — **“Tour wires”**
- **Others in government programs**
- **MITRE** — **device designs and architectures**
- **DARPA** — **we consult to Molectronics program**

Interesting nano-scale devices

- Ultra-dense (fast SRAM-like) memories
  - use as L2 cache for Si processors
  - eliminate main memory and bottleneck
- Ultra-low power, ultra-dense CPU logic
- Scalable manufacturing with a new cost model
  - device cost ~ size ~ time to run nano-factory
  - ‘wafer cost’ is a non sequitur

Potential SRAM cache densities

- Six transistors per SRAM cell
- Silicon transistors
  - 10^8 logic transistors/cm^2 in 2008 (SIA)
  - 10^9 SRAM transistors/cm^2 in 2008 (SIA)
  - 20 MByte SRAM L2 (1 cm^2) cache chip
- Nano-transistors (fast, < 1 nsec)
  - 1 nm x 10 nm, so 10^{13} SRAM transistors/cm^2
  - ~ 0.2 TByte SRAM L2 (1 cm^2) cache chip, but…
  - power/bit must scale down accordingly

How much is a mole of memory?

- 1 mole = 6.022×10^{23}
- A single processor generating addresses at 10 GHz will take 2×10^{6} years to touch every word
- Current large servers may have up to 1 sec of DRAM
- Would need ~ 10^{13}, 10 GHz processors to balance one mole of memory with present architectures

How large is a mole of memory?

- If we spread out a mole of memory on a 50 A grid, it would cover 1.5 x 10^7 m^2, a square 3.9 km on a side
- Assuming 10^4 kT per fetch, 10^{13} processors fetching at 10 GHz dissipate 4 MW
  - comparable to one of the ASCI machines
  - at 10^9 kT/op, this is 4 x 10^5 MW
- If we pack it in a cube we get 0.075 m^3, about 42 cm on a side
  - ~ 5 MW/m^3 at 10^4 kT/op

Assumptions

- Synthesis of elementary functional blocks is mostly atomically accurate
- Self- and directed-assembly of large systems will be error-prone
- Systems will not necessarily preserve topology over time, temperature fluctuations, and other disturbances

Applications to backing store

- Write once
  - slow is acceptable
  - behaving like tape is acceptable
- Archive the whole file state of machine forever
  - infinite undo
  - should be non-volatile at zero power
- Trees more probable than meshes…
- Fat trees might be fault-tolerant

Logic gates

- Fan-in and fan-out are required for existing designs
  - require level-restoration and I/O isolation...
- It is possible to obtain gain from a tunnel diode
  - discrete component tunnel diode logic attempted in the 1960's
  - it was intractable
  - twitchy

Power distribution

- All electrical logic families require a reference voltage or current to set threshold and a return path
- Immersion in a conductive liquid might provide global return
  - common ground return impedances are a noise source
- Assuming probabilistic assembly it is very important that any power distribution follow the actual logic structure
  - it might be easier to power the gates from an energetic compound dissolved in the ground return path
  - pump the liquid to clear decomposition products and move heat

Architecture

- Cellular automata designs might be evolved to work with ‘almost correct’ assembly
- Neural net architectures can almost certainly be made to work with given instances of assembly
- Test, route, and delete methods could be promising
- These are likely to be slow
- It’s possible that effective clock rates will be limited to very low frequencies by parasitics and gate drive
- Useful processing speed may require biological scale parallelism

Essential physics to solve

- Nano-transistor with gain
- Speed < 1 nsec desirable, 1 msec usable
- Low impedance power rails
- Long data buses for multilple SRAMs
- Energy per SRAM bit ~ 10^2 - 10^4 kT