"Very short reach optical interconnects in systems: The case for fiber-to-the-processor"

A. F. J. Levi

The University of Southern California
University Park, DRB 118
Los Angeles, California 90089-1111

http://www.usc.edu/alevi

voice (213) 740 -7318
email alevi@usc.edu

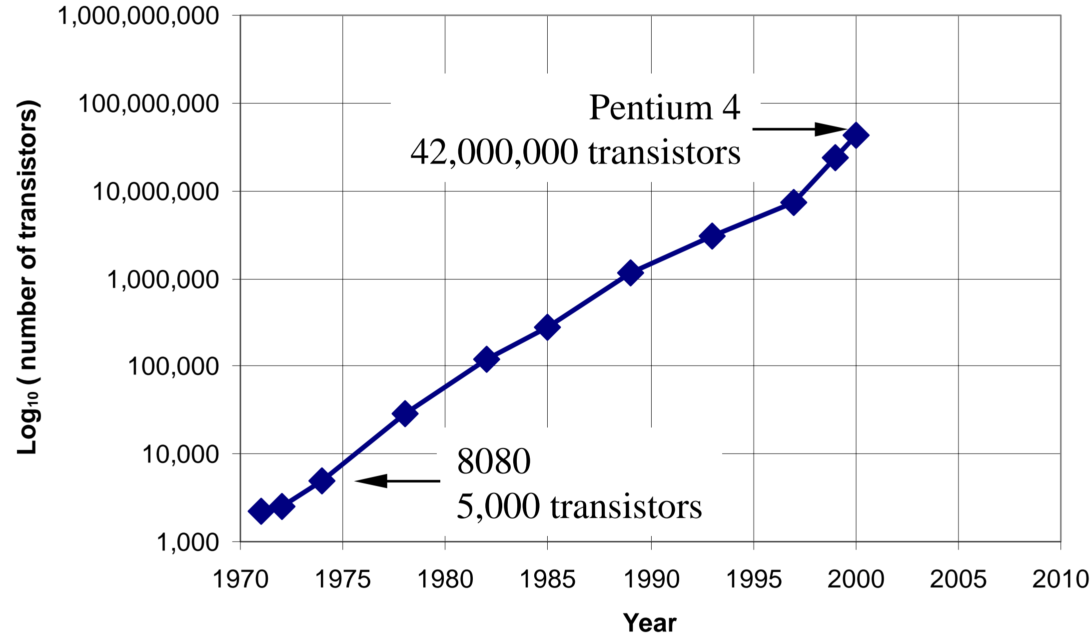
Presented at the Ratheon *Optical Digital Communication Seminar*, 10:30 a.m. to 11:00 a.m. PDST May 21, 2001.
Organizer: Irwin L. Newberg, vooice: (310) 647-3531, email: inewberg@west.raytheon.com

"Very short reach optical interconnects in systems: The case for fiber-to-the-processor"

- **In this presentation the focus is to evaluate the impact and benefit of advanced optical interconnect technologies in systems. It is likely that future system performance gains from simple scaling of transistor device dimensions will not contribute as much as they have in the past. Performance improvements will increasingly come from new architectures, better operating systems, and introduction of new technologies such as fiber-optics. The use of very short reach optical interconnects in systems will inevitably lead to direct optical connection to the processor.**

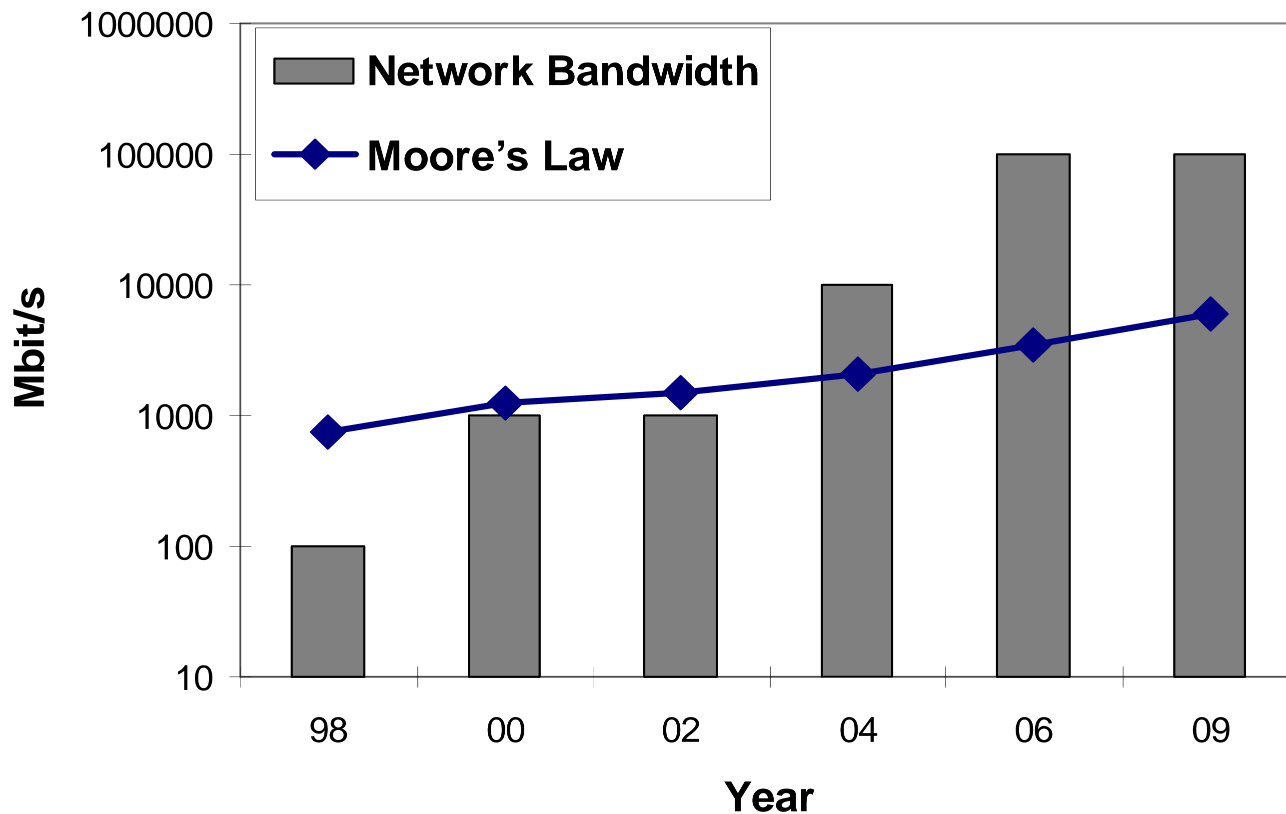
Impact of Moore's Law

- Increased number of transistors, increased clock rate and improved architecture has dramatically enhanced total microprocessor performance over past decade
- By 2010 architectures with a billion transistors per chip, clock rates > 10 *GHz* and nearly a trillion instructions per second will be considered. Multi-threading and multiple-processor architectures on a single-chip will be adopted to increase throughput and number of operations per second
- Dramatically increased power consumption and significant bandwidth performance demands imposed at the platform level



Optical interconnect bandwidth is increasing faster than CPU clock

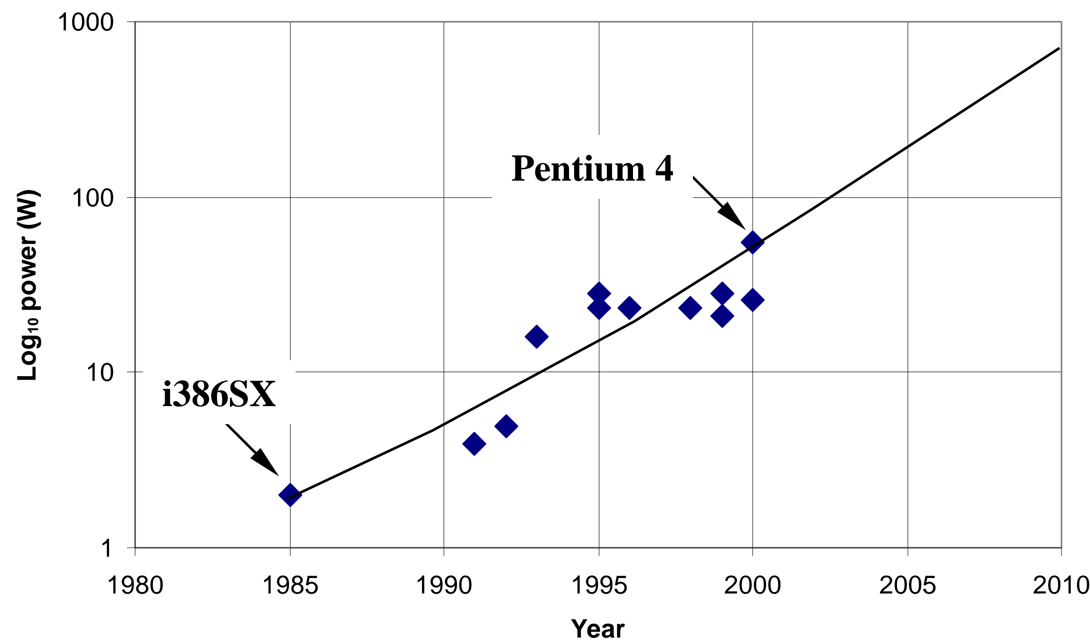
- **Network bandwidth increasing faster than Moore's Law**
  - **Presents an opportunity for new architectures**
- **Example: prospect of 100 Gb-Ethernet implementation**



Moore's Law: On-chip high-performance local clock (SIA 97)

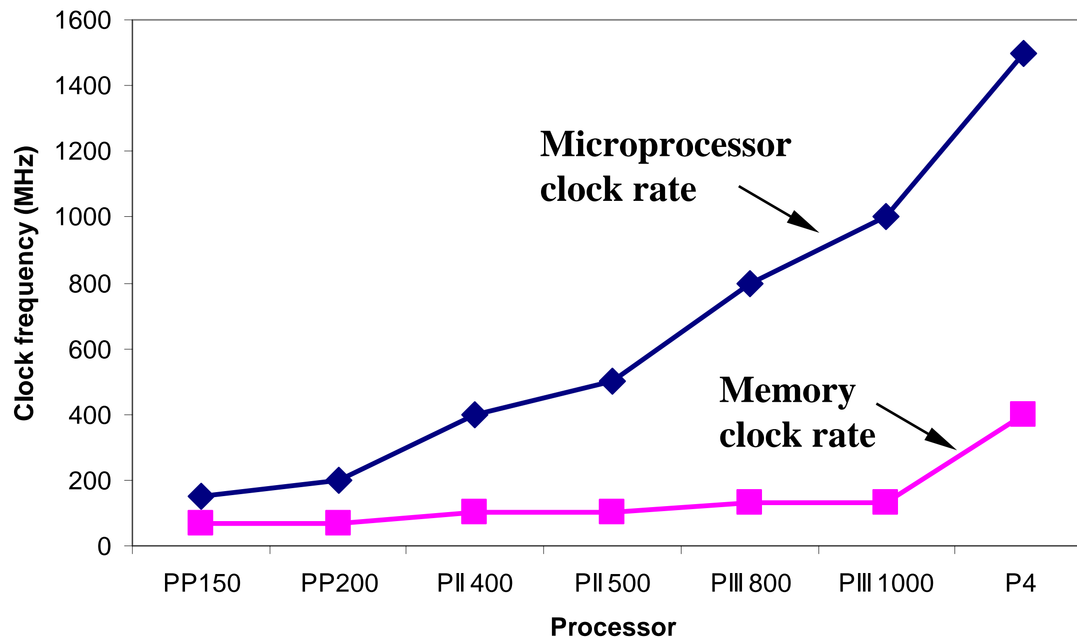
The power crisis

- Microprocessors with power dissipation approaching 1000 *W* by 2010 is not practical
- Power management will become a dominant aspect of design
- System architects will seek solutions from new and emerging technologies to find better system design points
  - Emphasize high-speed IO and de-emphasize increase in number of transistors per chip
  - Break Moore's Law by adopting new architectures and embracing new technologies
    - Fiber-optic interconnect enables distributed architectures with reduced power density



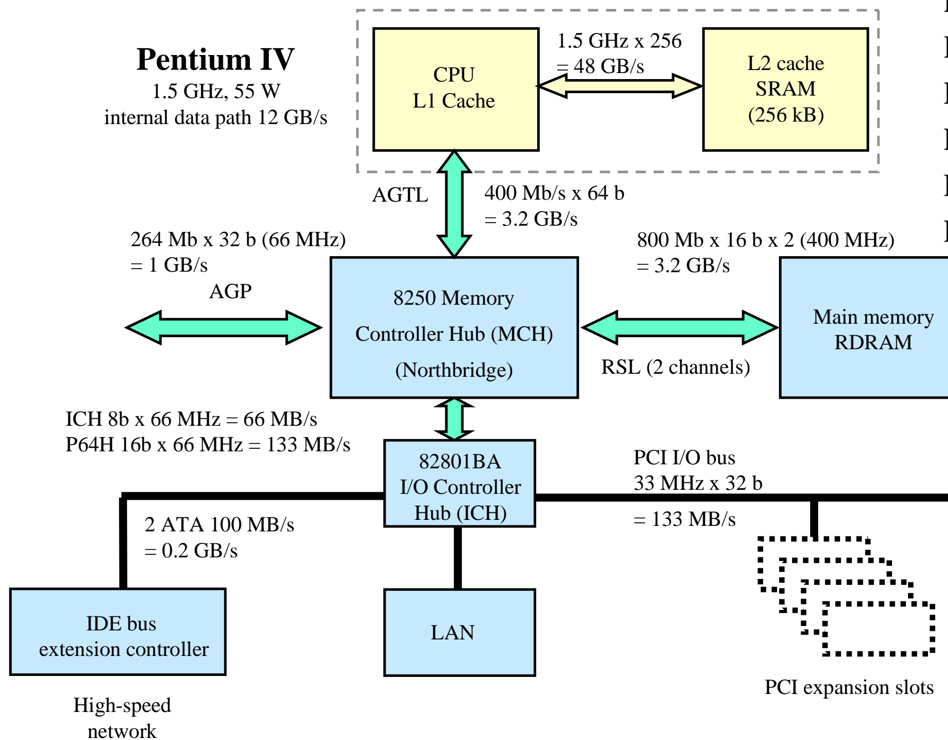
The memory access imbalance

- Long before 2010, the *imbalance between microprocessor performance and memory access* will be driven to a crisis point
- Microprocessor will lose hundreds of process cycles waiting for a single read from main memory
  - Limited improvement pushing conventional electrical bus rates into the *GHz* range and increasing bus widths



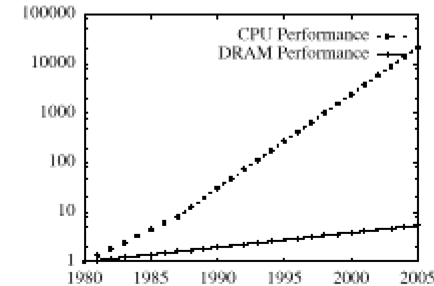
Microprocessor - DRAM performance gap

- **Average CPU clock rate doubles every 18 months**
- **Main memory data transfer speeds increase 10% every 18 months**
- **Conventional interconnects cannot deliver performance that matches improvement in CPU**



1997 - 2Q AMD K6 MMX (233 - 300 MHz)
1997 - 2Q Intel Pentium II (233 - 300 MHz)
1998 - Intel Deschutes (333 - 450 MHz)
1999 - Intel Pentium III (450 - 550 MHz)
2000 - Intel Pentium III / AMD Athlon (1 GHz)
2000 - Intel Pentium IV (1.5 GHz)
2001 - 2Q Intel Pentium IV (1.7 GHz)

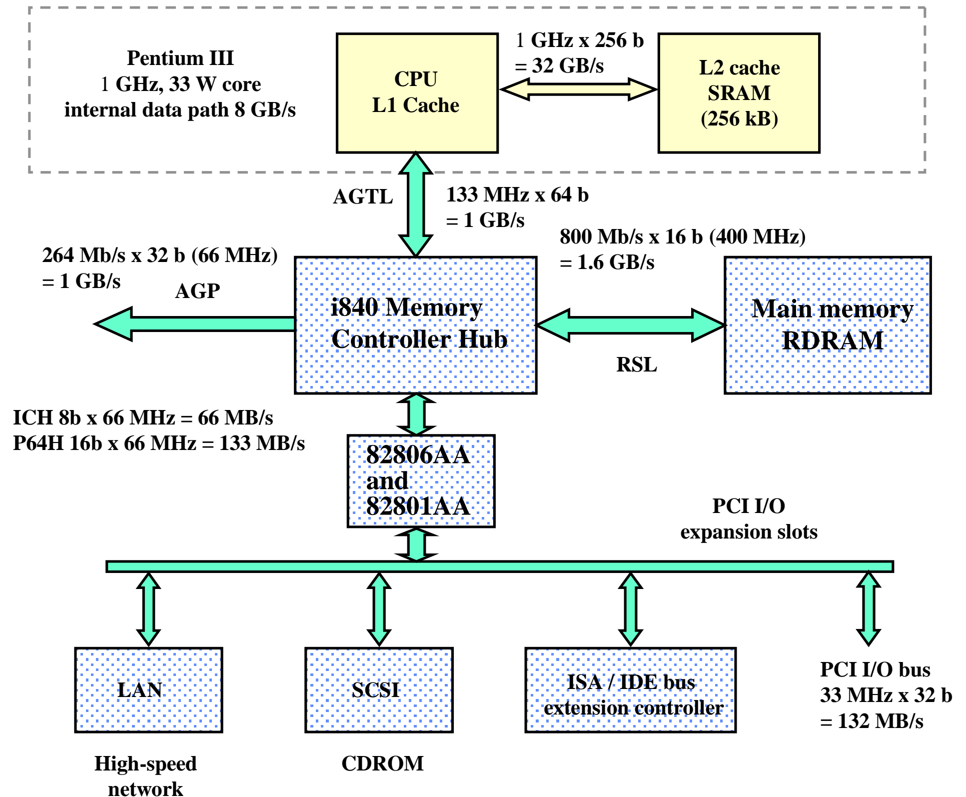
Processor	*Clock (MHz)*	*SPECint95/2000*	*SPECfp95/2000*
Pentium II	266	10.8	6.9
DEC Alpha	266	7.9	11.8
DEC Alpha	667	35/413	69.4/500
Pentium III	800	38.9/386	32.1/286
Pentium III	1000	---/448	---/318
Pentium IV	1700	---/1586	---/608

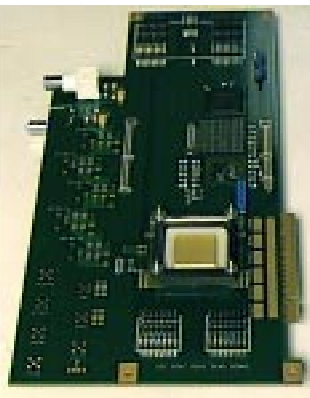


Source: Hennesy and Patterson "Computer architecture", Morgan Kaufmann (1996)

Nearing end of bus-based architecture today

- **One more generation of today's bus-based architecture**
  - multi-level signaling
  - 1000+ chip IO
  - exotic packaging and PCB technology
  - IB and LVDS at 2.5 Gb/s
- **New approach needed for products in 2004**
  - signal integrity focus for electronics
  - switched-based architecture
  - seamless integration of optical technology

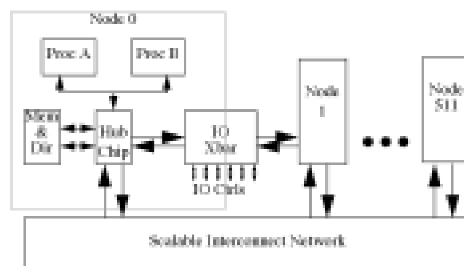




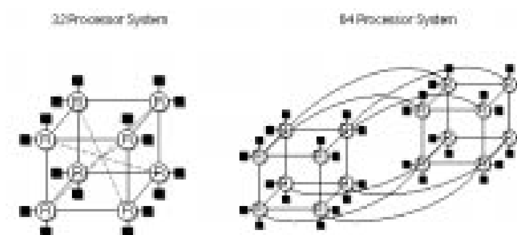
Example:
USC PONI-ROPE PCB area < 50% ICs, surface mount packages,
12-metal layers, Gb/s per signal line,
de-skewed signal lines to +/- 10 ps,
5 mil lines, 7 mil spaces

Example: SGI Origin switch-based system architecture

Electrical interconnect:
- 44 signal pins per direction
- up to 5 m electrical cable

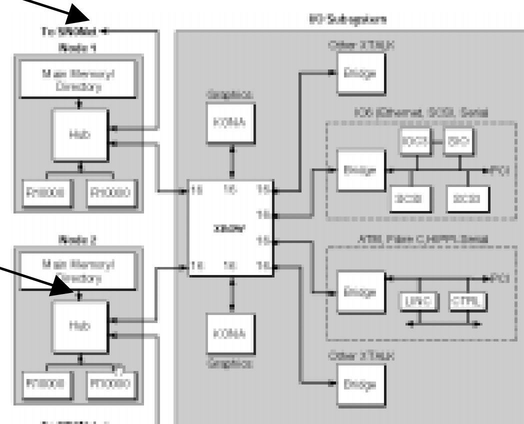


Origin block diagram

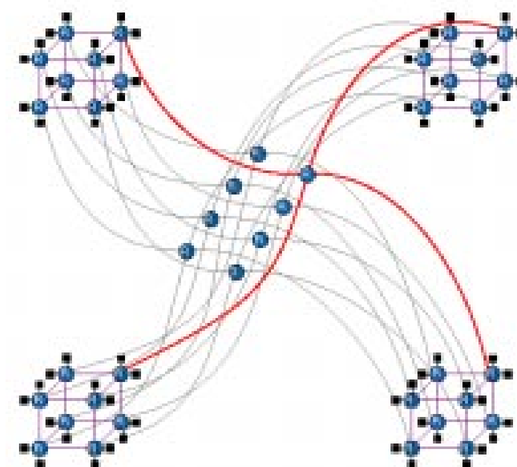


32P and 64P Bristled Hypercubes

Node-to-node access
0.73 GByte/s peak per direction
0.625 GByte/s sustained per direction



Example IO subsystem block diagram



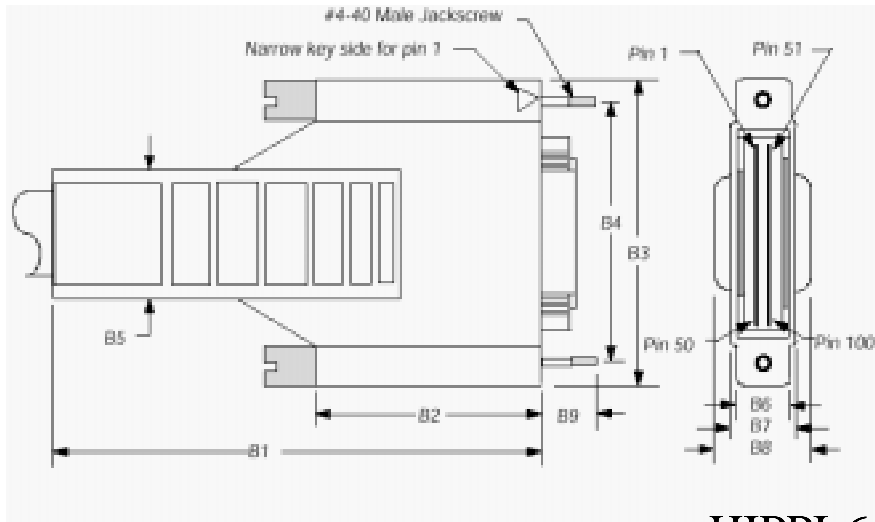
128P Heirarchical Fat Bristled Hypercube

Memory access
0.78 GByte/s peak total
0.78 GByte/s sustained total

Latency
Pin to pin hub 41 ns
Local memory 310 ns
4P remote memory 540 ns
8P average remote memory 707 ns
16P average remote memory 726 ns
32P average remote memory 773 ns
64P average remote memory 867 ns
128P average remote memory 945 ns

Node-to-node 16 kB page block transfer < 30 µs

HIPPI-6400 electrical interconnect



Dimension	mm	inches
B1	96.28 Max	3.80 Max
B2	43.18	1.70
B3	58.67 Max	2.31 Max
B4	50.80	2.00
B5	25.40	1.00
B6	10.92	0.43
B7	12.70	0.50
B8	19.05 Max	0.750 Max
B9	10.77	0.42

Copper interconnect:
- Poor form factor (< 1 GB/s/inch)
- Limited bandwidth (< 1 Gb/s)
- Limited distance (< 50 m)

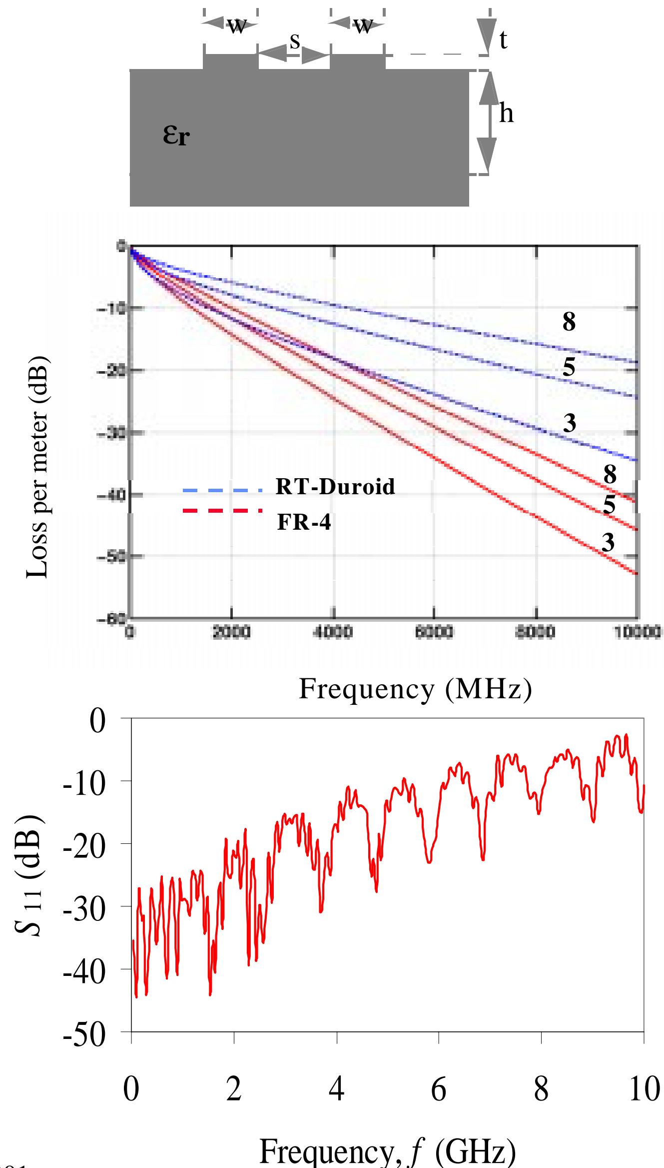
HIPPI-6400 electrical connector
- 2.31” × 0.75” edge dimension
- 2 row, 100 pin connector (23 × 2 × 2 = 92 signals)
- 0.664” diameter cable up to 50 m
- 6” bend radius for cable
- 2 Byte data 500 Mb/s per direction, 4b/5b coding
- 0.8 GB/s peak bandwidth per direction, 1.6 GB/s bisection bandwidth (bisection bandwidth density 0.7 GB/s/inch or 2.2 Gb/s/cm)

Bandwidth limitation of electrical PCB interconnect

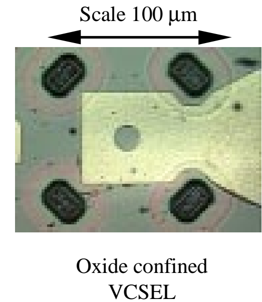
- Calculated loss per meter for 50 Ω microstrip PCB including skin-effect and dielectric losses for trace widths 8, 5 and 3 mils in 1 *oz* copper.
  - FR-4 ($\varepsilon_r = 4.5$, $\tan\delta = 0.02$)
  - RT-Duroid ($\varepsilon_r = 2.35$, $\tan\delta = 0.005$)

- -10 dB per 20 cm at 10 GHz for w = 3 mil and $\delta = 0.02$

- Measured S_{11} for a standard 50 Ω high-speed electrical test-fixture (Tektronix #671-3273-00).
  - Poor (SMA) launch onto a 3”-long, 60 mil wide microstrip trace results in severe reflections at 10 GHz.

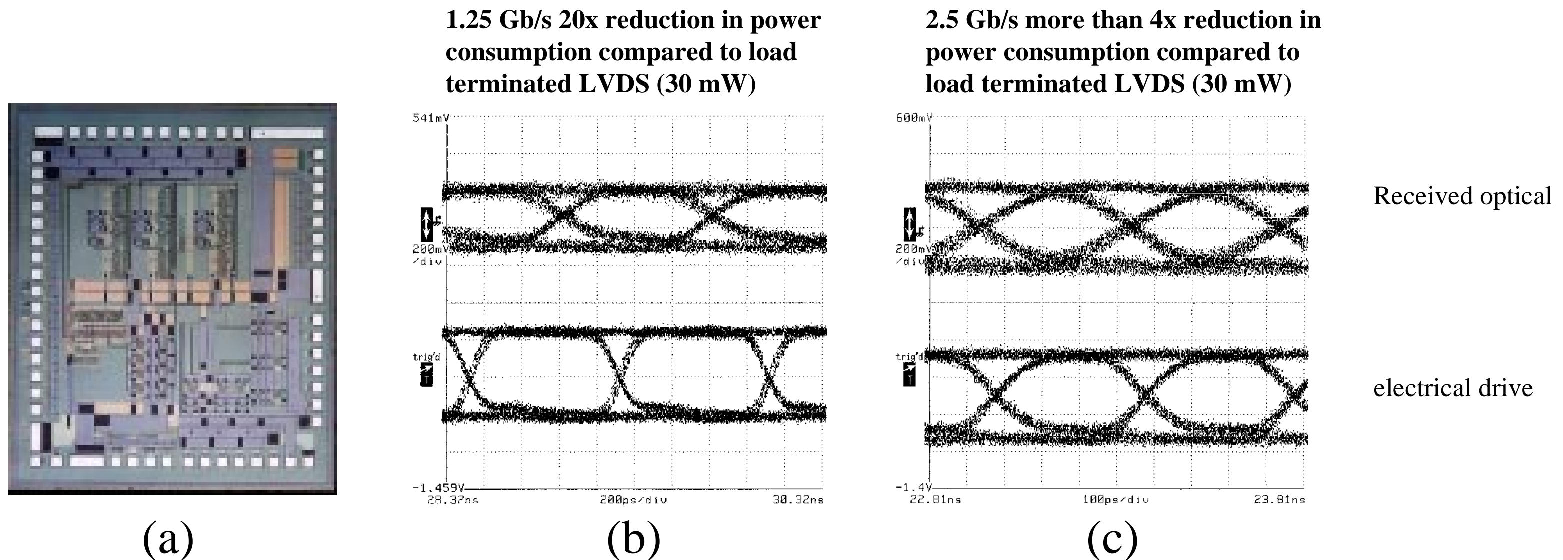


Power dissipation advantage of optics: Experimental results



- **Design point: Low-power 2.5 Gb/s 0.5 µm CMOS / VCSEL opto-electronic Tx**
  - (a) VCSEL driver in 0.5 µm CMOS. 2.1 x 2.5 mm^2 T8 test-die received at USC 5.8.97.
  - (b) Upper trace is output of oxide-VCSEL (I_{th} = 0.2 mA, V_{th} = 1.5 V, I_b = 0.5 mA, V_b = 1.6 V, L_{pp} = 0.078 mW) driven by +Tx circuit. **1.5 mW power consumption at 1.25 Gb/s**. Lower trace is -Tx electrical output. Input data is 50 mVpp 1.25 Gb/s NRZ 2^{31} - 1 PRBS.
  - (c) Upper trace detected output of oxide-VCSEL (I_{th} = 0.5 mA, V_{th} = 1.5 V, I_b = 1.9 mA, V_b = 1.7 V, L_{pp} = 0.744 mW) driven by + Tx circuit. **7 mW power consumption at 2.5 Gb/s**. Lower trace is -Tx electrical output. Input data is 50 mVpp 2.5 Gb/s NRZ 2^{31} - 1 PRBS.

B. Madhavan and A. F. J. Levi, *Low-power 2.5 Gbps VCSEL driver in 0.5 µm CMOS technology,* Electron. Lett. **34**, 178-179 (1998) and http://www.usc.edu/alevi



(a) (b) (c)

Electrical compared to optical at 10 Gb/s in 0.1 μm CMOS

- **Optics can have near 3 × improvement in total link power for *same* link loss because optics has no 50 Ω load and can have lower Tx driver swing**

- **Projected 10 Gb/s transceiver power dissipation using 0.1 μm CMOS (V_{dd} 1.1 V, t_{ox} = 3 nm, V_{th} = 0.3 V). Assume 850 nm VCSEL I_d = 2 mA, 300 μW, C_d = 300 fF, CE = 0.5. Tx pre-driver and output stage. 400 mV LVDS (bias at V_{dd} / 2) Tx electrical power includes 50 Ω parallel load termination. Rx first-stage designed to drive FF.**

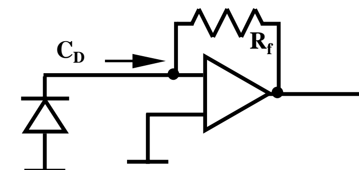
Link loss (power)	total electrical power dissipation (*Tx* + *Rx* + *Load*)	total optical link power dissipation
-5 dB	**4.90 + 0.11 + 3.2 = 8.21 mW**	**2.45 + 0.11 = 2.56 mW**
-10 dB	**4.90 + 1.11+ 3.2 = 9.21 mW**	**2.45 + 1.10 = 3.55 mW**

0.1 μm CMOS projections based on experience with 0.8, 0.5, 0.35, 0.25 μm CMOS, PECL, LVDS, and VCSEL TX and pin RX circuits operating at greater than 2.5 Gb/s.

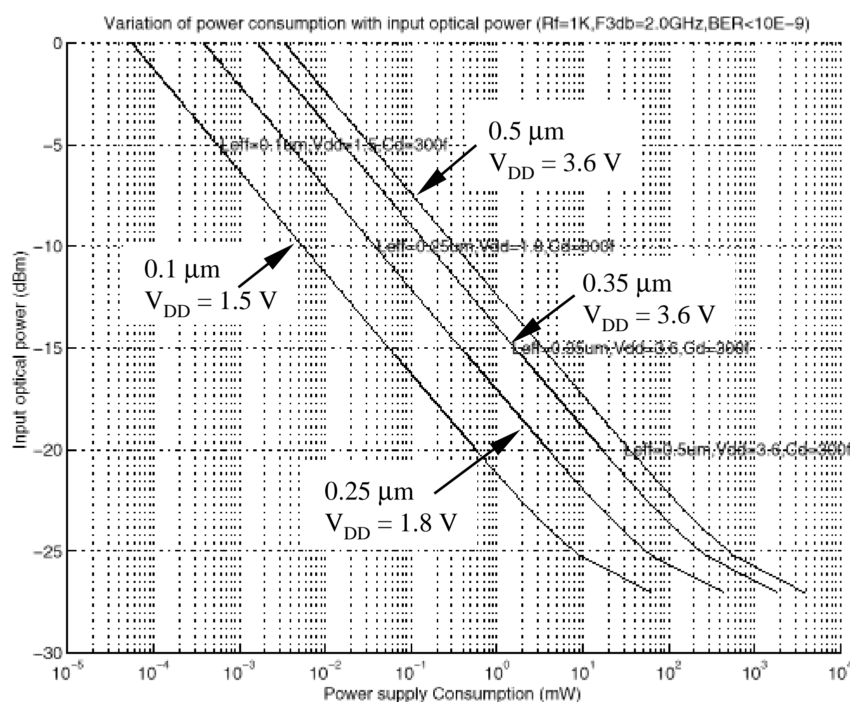
See, e.g., B. Madhavan and A. F. J. Levi, *Low-power 2.5 Gbps VCSEL driver in 0.5 μm CMOS technology,* Electron. Lett. **34**, 178-179 (1998) and http://www.usc.edu/alevi

Optical-to-electrical conversion using CMOS TIA

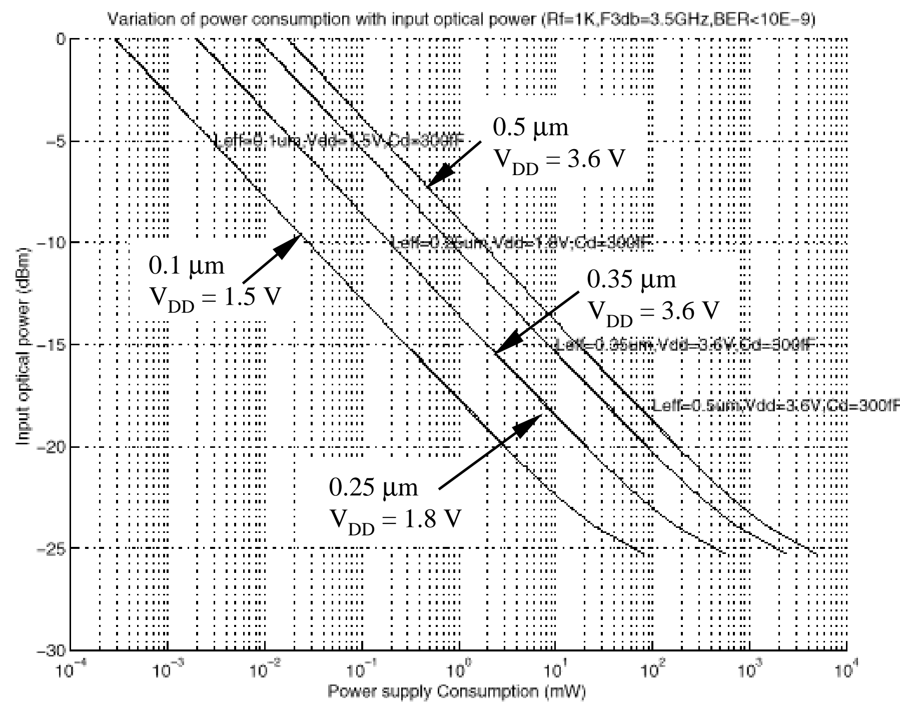
- **Minimum power dissipation depends on optical input power, CMOS technology, V_{DD}, and photo-diode capacitance**



1.25 Gb/s, R_f = 1000 Ω, C_D = 300 fF

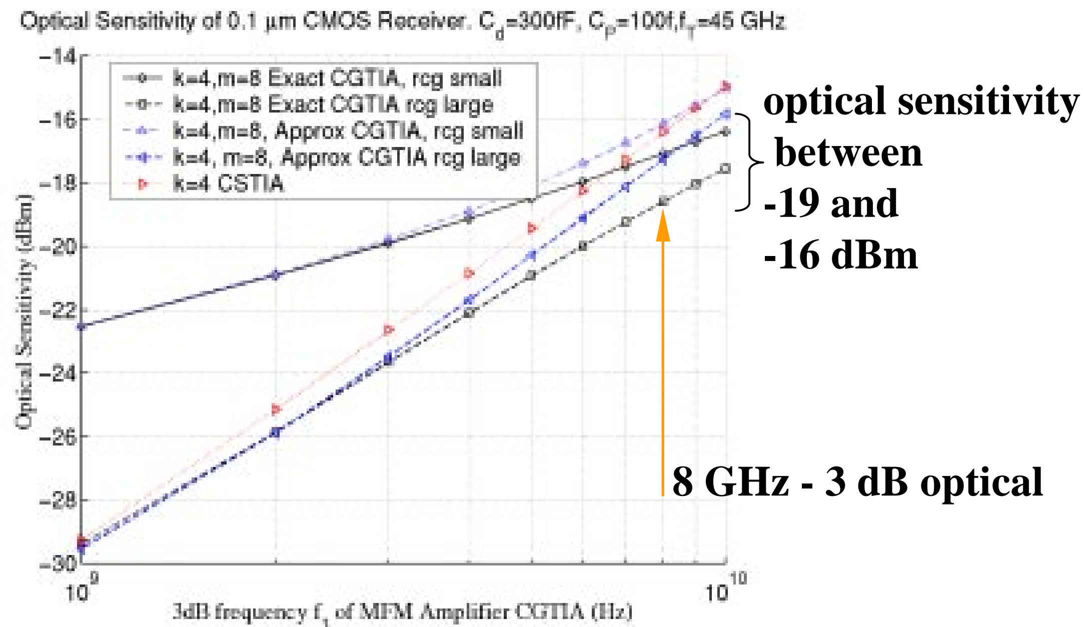


2.5 Gb/s, R_f = 1000 Ω, C_D = 300 fF



10 Gb/s in 0.1 μm CMOS receiver circuit sensitivity

- **-16 dBm optical sensitivity possible**
  - future integration of detector and TIA circuitry in CMOS



Calculated 0.1 μm CMOS optical receiver sensitivity

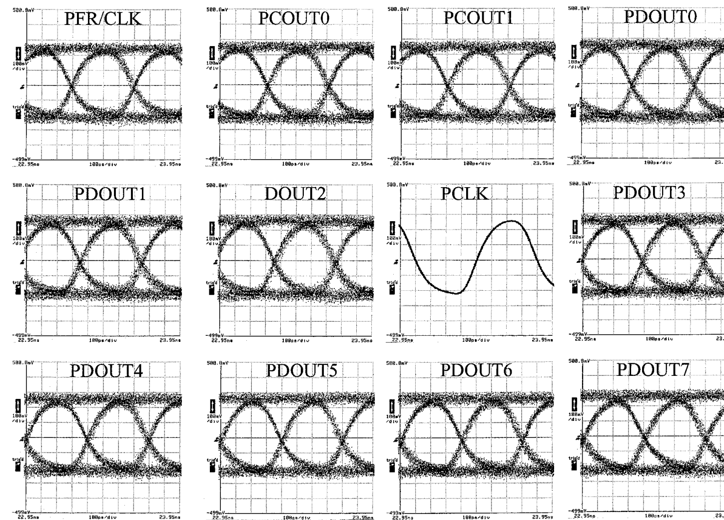
High-performance opto-electronic interface to CMOS

- **Potential power savings using all-optical signal processing**

USC PONI MUX IC (v1) 0.5 μm CMOS
10 × 2 mm² die
submitted 2/18/98
received 5/2/98
V_{DD} = 3.6 V
5.7 W

Power estimates

V_{DD} (V)	Power (W)	Tech. (μm)
3.6	5.7	0.5
3.3	4.7	0.35
2.5	2.9	0.25



11 × 2.5 Gb/s + CLK
1:2 / 2:1 MUX

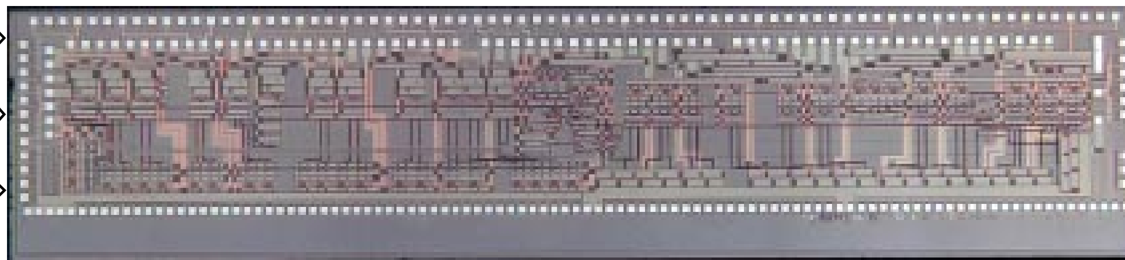
55 Gb/s bi-section data bandwidth per cm

2.7 ns Tx/Rx mux/demux end-to-end latency

Tx / Rx 50 Ω terminated
1.2 V < V_{TT} < 2.0 V
(LVDS compliant).

Power consumption using 0.5 μm CMOS technology

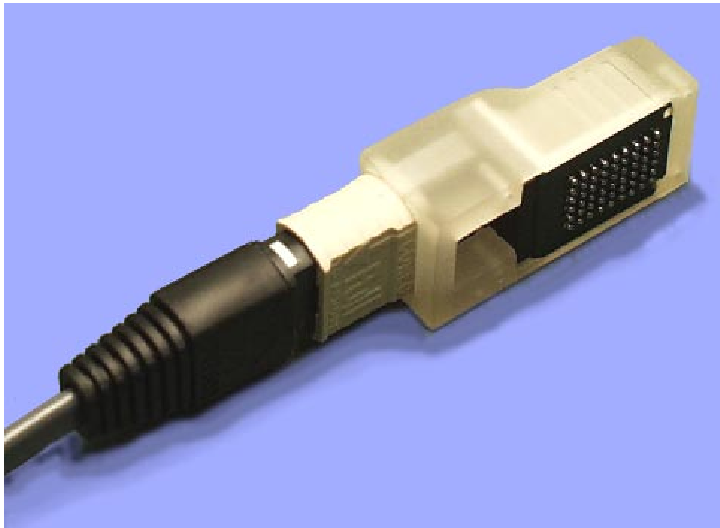
0.72 W 2.5 Gb/s I/O
3.74 W Core
1.24 W 1.25 Gb/s I/O
5.7 W Total



"*A 55 Gb/s/cm data bandwidth density interface in 0.5 μm CMOS ...*",
B. Madhavan and A. F. J. Levi, Electron. Lett. **34**, 1846-1847 (1998)

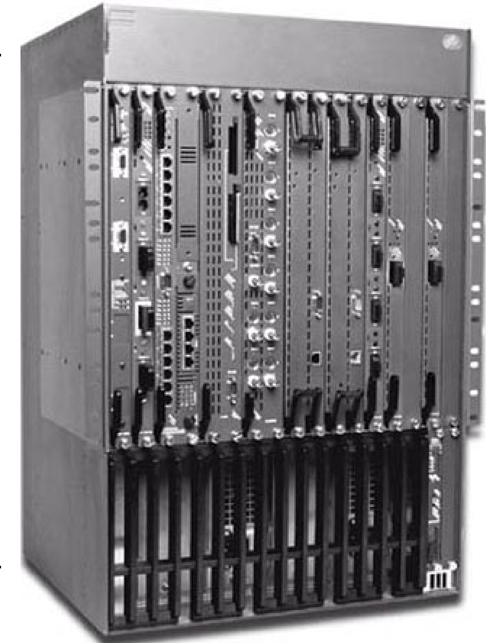
System impact of > 50 Gb/s/cm parallel fiber-optic bandwidth density

- **Multi-GByte/s data rate per linear cm at low cost**



Agilent PONI Tx-module

HP PONI module + USC CMOS interface IC supports > 4× capacity IBM ATM switch in 1 cm form-factor

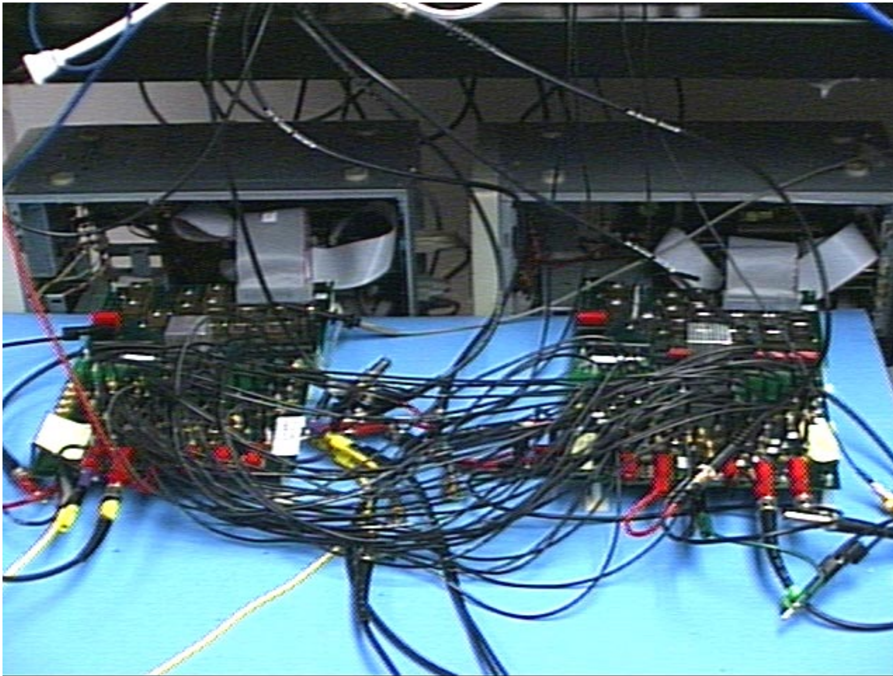


IBM 8265 Nways ATM switch
12.8 Gb/s capacity, ~ 1 m back-plane

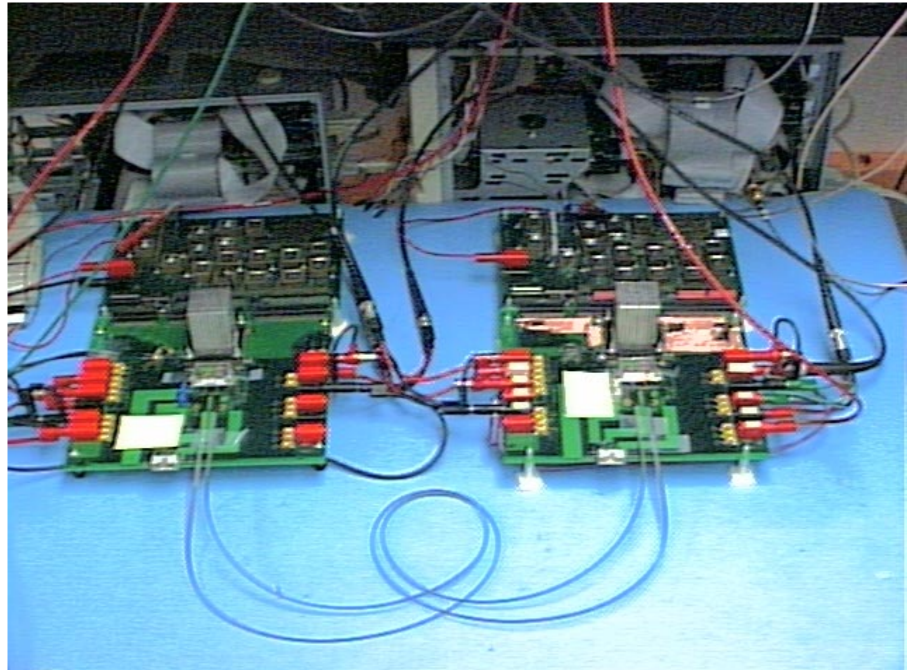
- **USC - Agilent PONI DARPA program can provide > 50 Gb/s per cm using**
  - USC CMOS interface IC
  - 12-wide fiber ribbon, 2.5 Gb/s per fiber signaling, and MTP connector
  - BGA surface mount to PCB

Fiber form factor advantage for GB/s interconnect

USC electrical and optical system test between two Pentium hosts
2.5 GB/s (20Gb/s) data transfer rate in each direction



Electrical test fixture for LA chip requires 40 coaxial cables



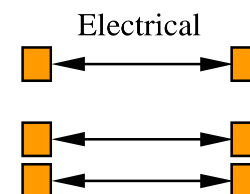
HP POLO-2 module and LA chip requires 2 ribbon fibers

Physical performance advantages of fiber-to-the-processor

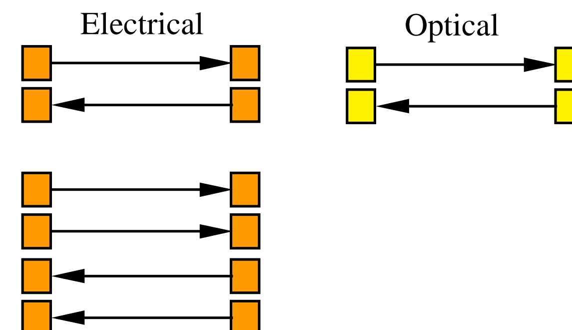
- **40 × form-factor bandwidth advantage**
  - 100 Gb/s in 125 μm (5 mil) diameter glass fiber using 100GbE technology compared to electrical signals in 10 differential micro-strip lines on 10 mil centers (200 mil total) and each signaling at 10 Gb/s
- **50,000 × interconnect distance advantage**
  - up to 10 km in glass fiber using 100GbE technology compared to 20 cm in micro-strip lines on 10 mil centers signaling at 10 Gb/s
- **infinite EMI-in-path advantage**
  - zero EMI in glass fiber using 100GbE technology compared to finite cross-talk and radiative emission in micro-strip lines signaling at 10 Gb/s
- **3 - 10 × power-dissipation advantage depending on exact implementation**
- **Future integration of optical components and CMOS electronics**
  - encapsulated processor

Fiber-to-the-processor requires controlled technology transition

- **Electronics driven to design-point (*slow-wide*) which does *not* provide natural transition to fiber-optic interconnects (*fast-narrow*)**
- **0.1 µm CMOS design trade-off**
  - Bi-directional saves I/O pins and is favored by electrical interconnects
    - **Single-ended electrical**
      - **1-pin**
        - multi-level signaling can give simultaneous Tx and Rx
        - noise, reference circuitry and reflections limit signaling rate
    - **Differential electrical**
      - **2-pins**
        - improved immunity to noise

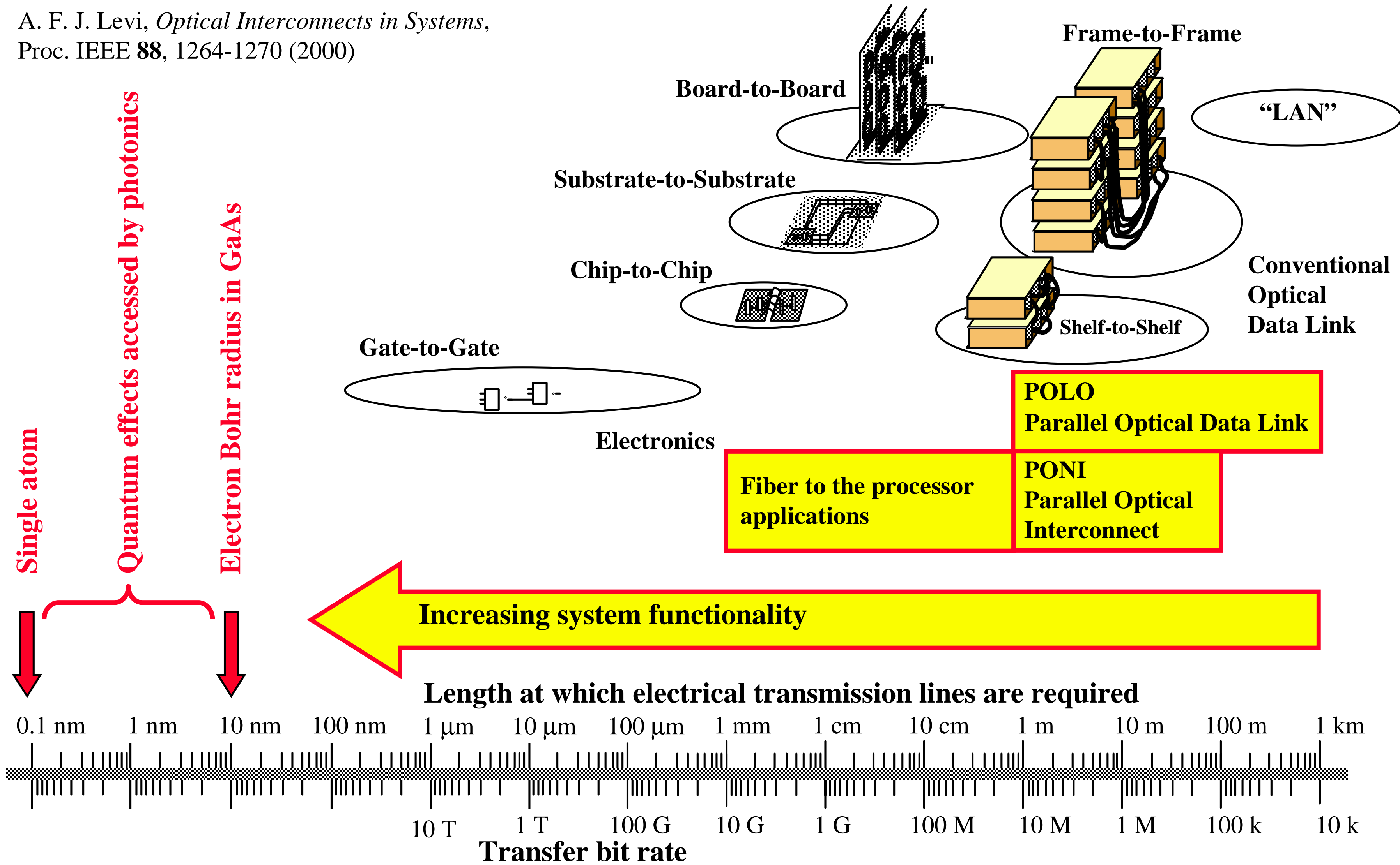


  - Challenge is 5 Gb/s/pin circuit design for bi-directional electrical
      - **Design of broad-band reference circuitry**
      - **Impedance matching**
      - **End-to-end signal integrity (IC, package, PCB)**
  - Uni-directional is natural choice for fiber-optic interface
    - **Single-ended electrical**
      - **2-pins**
        - noise
    - **Differential electrical**
      - **4-pins**
        - robust signaling at 10 Gb/s
        - minimum latency
        - minimum logic and control signals
        - no buffering
        - compatibility with optical interconnect solutions



System interconnect hierarchy and advanced optical solutions

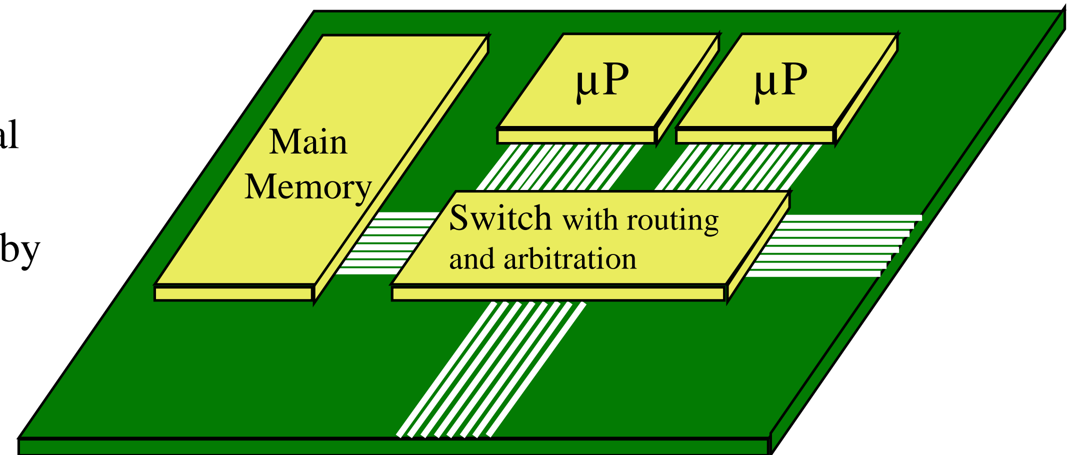
A. F. J. Levi, *Optical Interconnects in Systems*,
Proc. IEEE **88**, 1264-1270 (2000)



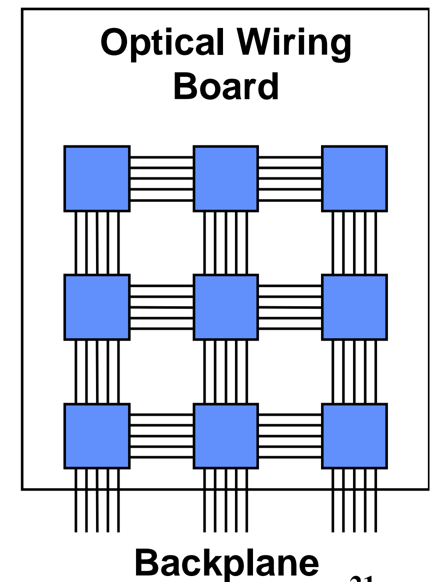
WDM for inter- and intra-PCB interconnects

- **Concept**
  - Topologically simple chip-to-chip optical interconnects
  - WDM to compete with Cu for spatial density
  - Complex logical topologies created by switching in silicon
- **WDM provides**
  - Improved data-bandwidth density compared to TDM or spatial multiplexing
  - Lower-cost connector compared to parallel fiber-optic solution
  - Future use of integrated nano-photonic components for all-optical functions

Multi-processor node 10x10 cm^2



- **Ideal network router**
  - Non-blocking connectivity
  - Speed-of-light latency
  - Infinite-bandwidth

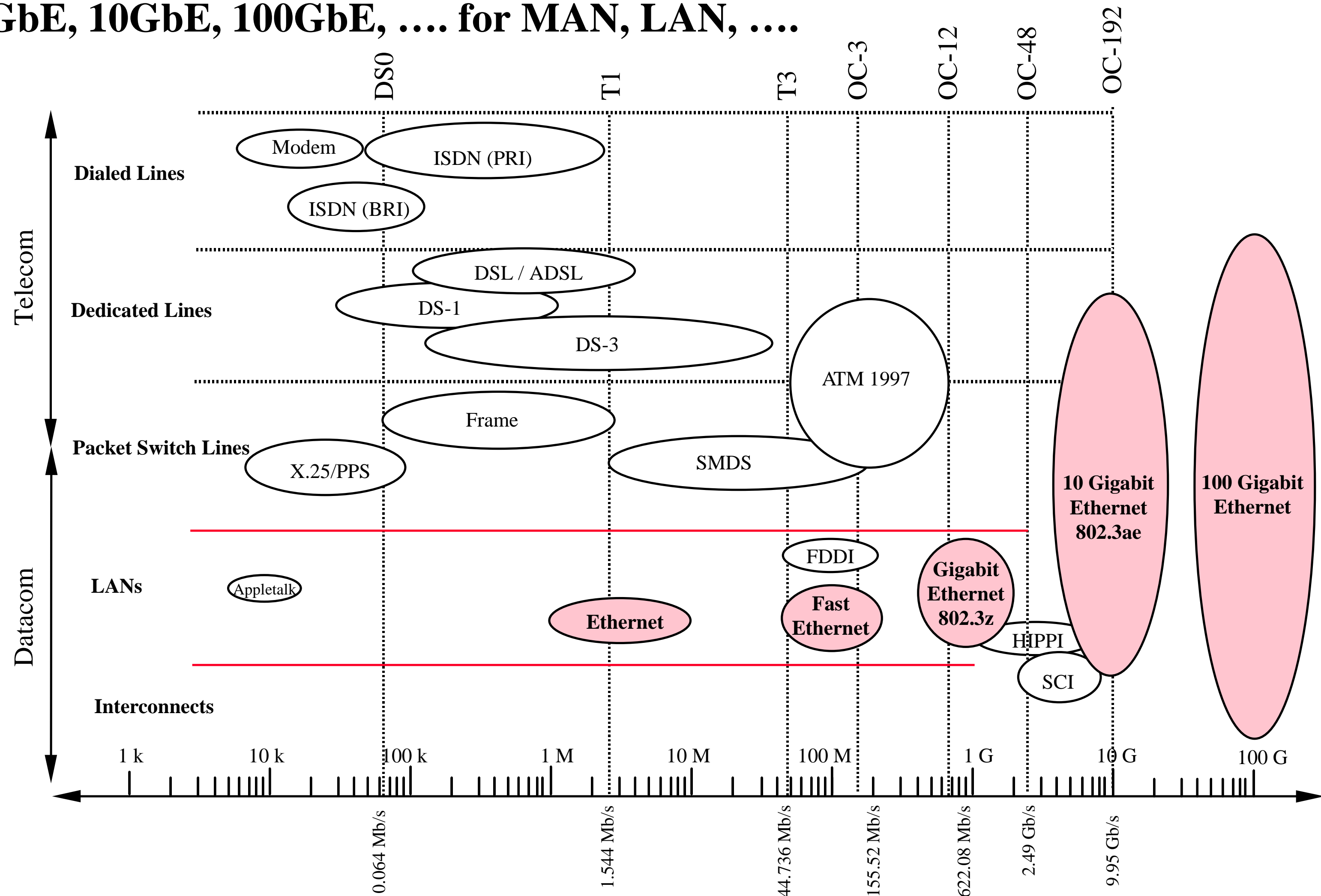


Fiber-to-the-processor bandwidth scaling

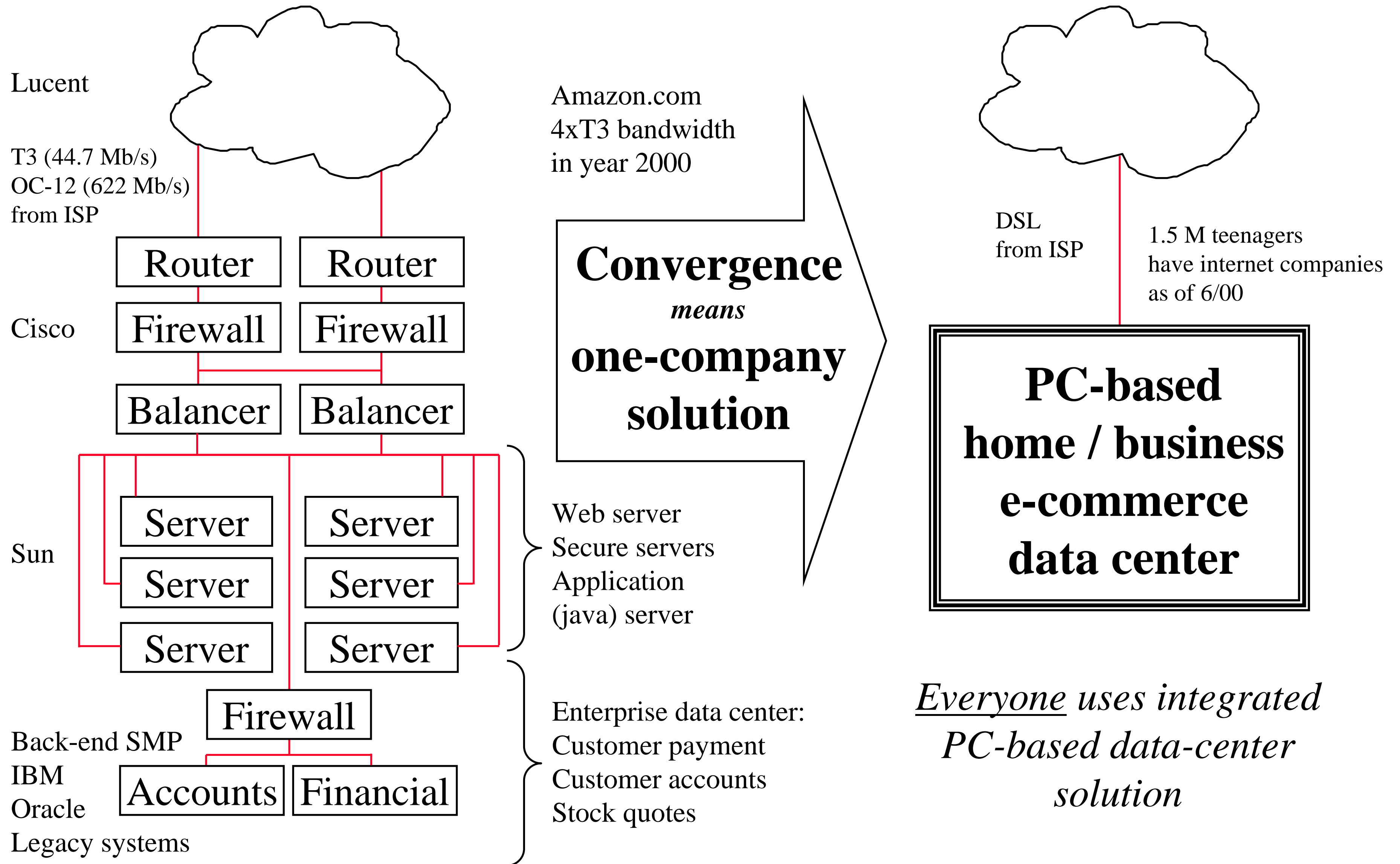
- **Willamette (PIV) bus**
  - 400 MHz × 64b = 25.6 Gb/s (3.2 GB/s) peak data
    - *one quarter* of the data bandwidth of a *single* 100GbE WDM optical fiber link
- **Future 10x bus frequency**
  - 4 GHz × 64b = 256 Gb/s (32 GB/s) data
    - less than three 100GbE WDM optical fibers
  - Assume *total* 64 data and 56 address and status lines with 4 GHz clock gives 4 GHz × 120b = 480 Gb/s (60 GB/s)
    - five 100GbE WDM optical fibers
- **Future scaled ring-based on 100GbE WDM interconnect**
  - 16 × 8-bit data-path implementation has network data-bandwidth of 16 × 8 × 10 Gb/s = 1280 Gb/s (160 GB/s)
    - 13 100GbE WDM optical fiber links
  - **4 rings use 64 100GbE WDM optical fiber links for a total data bandwidth of 5120 Gb/s (640 GB/s) which is *200 times greater* than the PIV-bus data bandwidth**

Physical interconnect technology convergence to IEEE 802.3xx Ethernet standard at 10 Gb/s and 100 Gb/s data rates

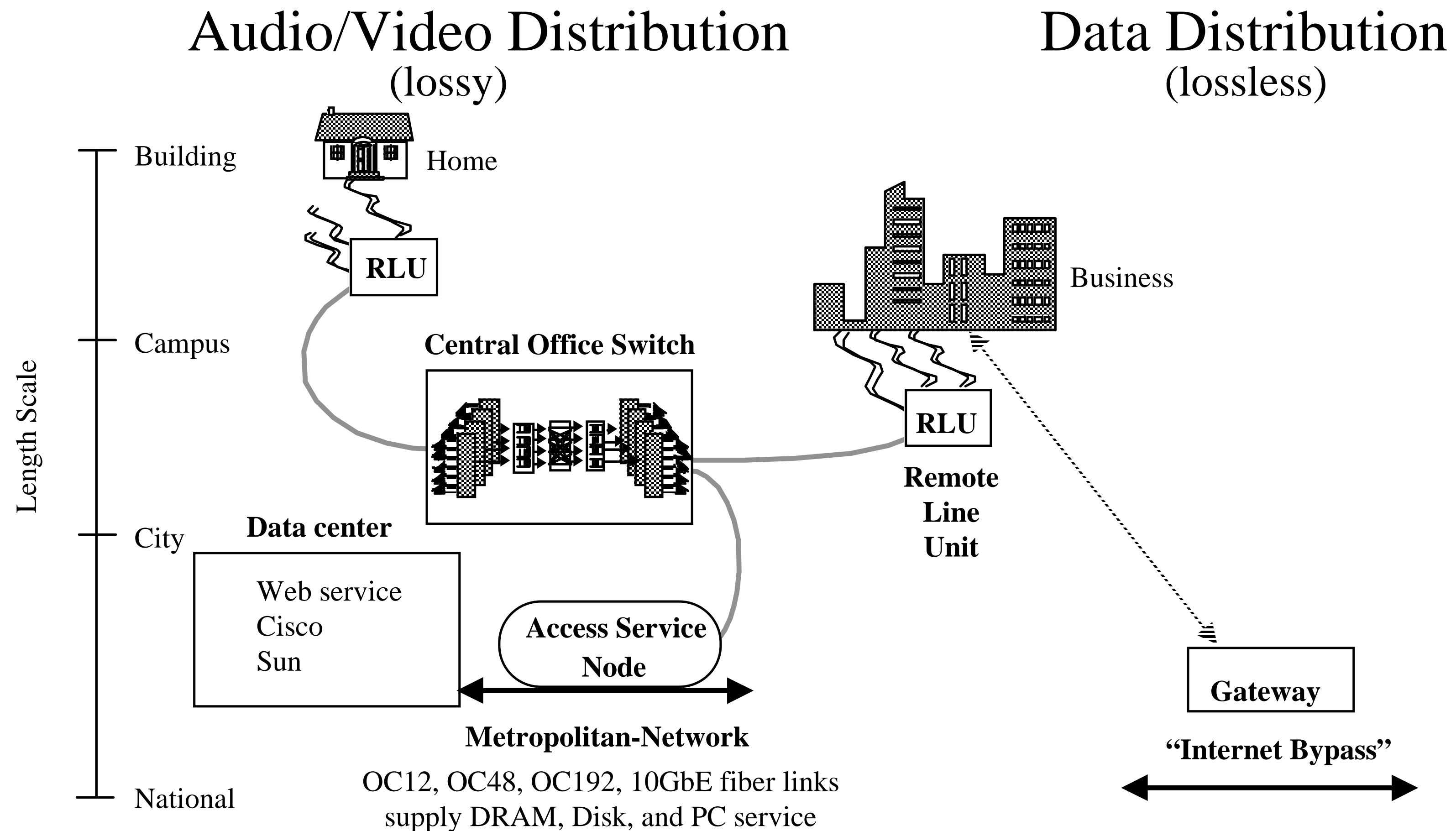
- **SONET for WAN**
- **GbE, 10GbE, 100GbE, …. for MAN, LAN, ….**



Possible future of the e-commerce data-center: Convergence

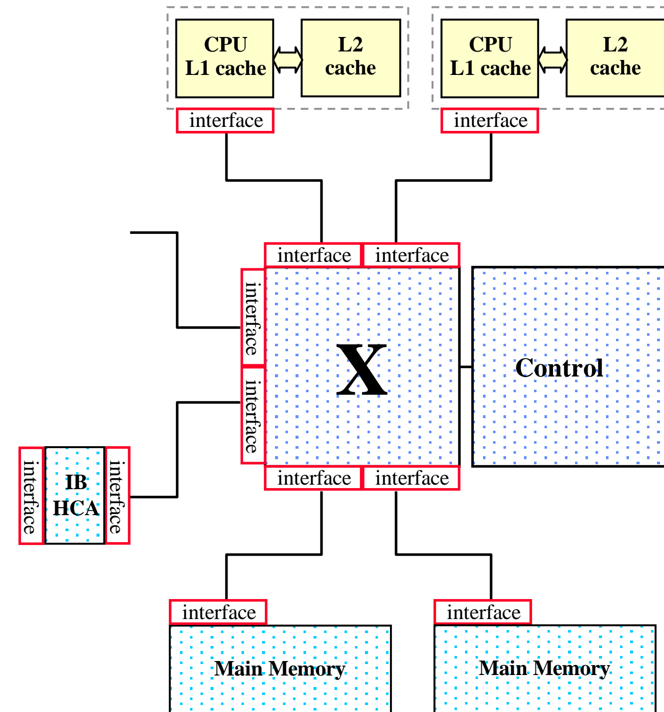


Service convergence



Technology enablers for a fiber-ready system

- **Signal integrity focus for electronics**
  - high-speed interface for optics
- **Switched-based architecture**
  - design for high-speed fiber-optic ports
  - Intel controller
- **Seamless integration of optical technology**
  - availability of interface electronics
    - electrical-to-optical conversion modules
    - electrical signaling conversion modules
      - multi-level bi-directional electrical to uni-directional optical interface
    - board-level integration modules
  - adaption of 10GbE and 100GbE fiber-optic components into micro-OSA



The promise of optics: "Almost-free, infinite-bandwidth, anywhere, anytime !"

- **New system design optimization and functionality**
  - **any distance for a given bandwidth**
  - **less heat *density* because of distributed nature of system enabled by optics**
  - **reduce I/O count**
  - **lower power than copper**
  - **low EMI**
  - **low cost**
  - **ps node latency, deadlock protection, adaptive routing**

- **Getting technology from here to there**
  - **Adoption helped by one-stop *technology shopping* for**
    - **standard packaging and board-level integration**
    - **standard CMOS library cells**
    - **standard OSA footprint**
    - **proven reliability as good or better than copper**
    - **demonstration systems and applications**

Supporting fiber-to-the-processor

- **Fiber-to-the-processor is coming**

- **Opto-electronics in CMOS-based systems**
  - **need data-com to keep component cost low**
  - **need availability**
  - **need standards that *help* the designer**
    - **complete design support**
      - library cells, evaluation boards, mechanical, system testing, software
  - **need compelling system demonstrations**
    - **new architectures, new functions, higher performance, reduced cost**
    - **integration with software**