Fiber-to-the-processor and other challenges for photonics in future systems

A.F.J. Levi

http://www.usc.edu/alevi

with contributions from
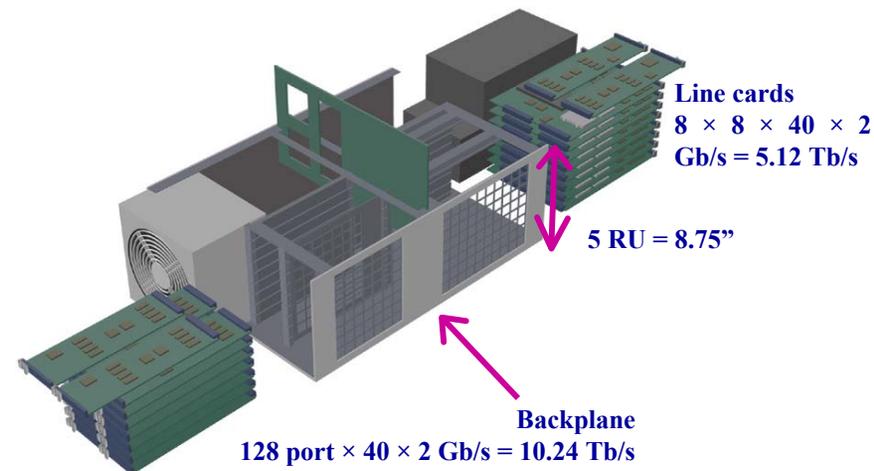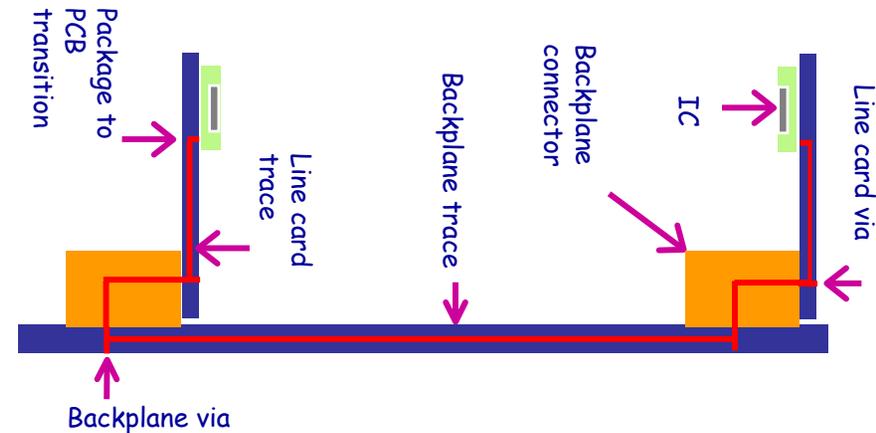
**Bindu Madhavan – USC
and
Agilent Technologies**

Stanford, April 21, 2005

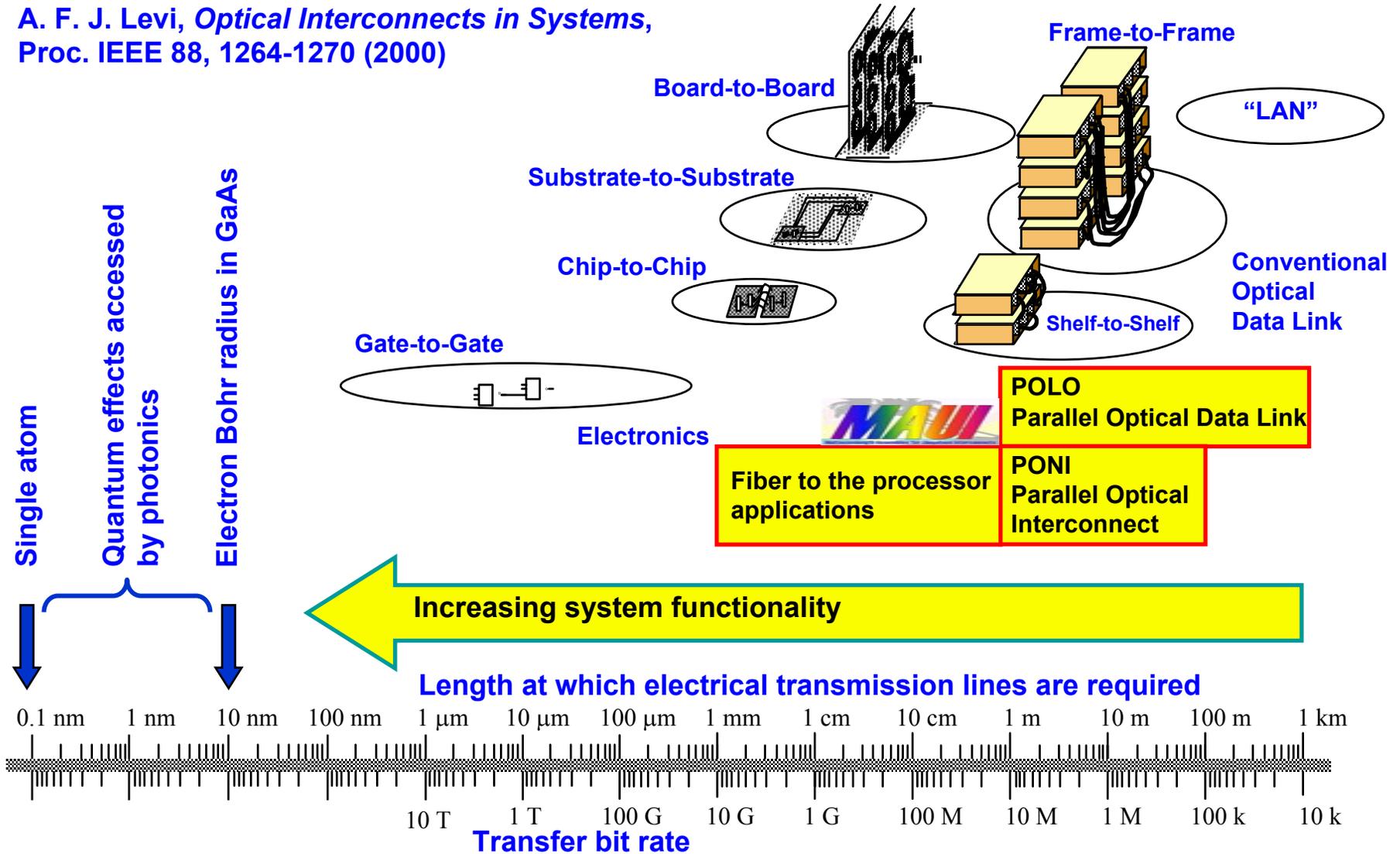# What is a system ?

- ➢ **Understand electronics in systems**
  - – **Definition of system**
    - • **Complex enough to require system area network**
      - – **Multi-processor rack-based system, router, data center, telephone switch, automobile etc., *are* systems**
      - – **Cell-phone, telephone handset, camera, pocket calculator, etc., are not complex enough to be systems**
  - – **Chip IO performance**
  - – **Backplane performance**
- ➢ **Chassis systems composed of passive backplane with connectors for linecards**
  - – **Backplane supplies power to linecards**
  - – **Connectors are interconnected by traces in backplane**
- ➢ **Chassis systems have slots for linecards that plug into backplane at connectors**
- ➢ **Total chip-to-chip interconnect length up to 1meter.**
- ➢ **Interconnect loss is a tradeoff between**
  - – **Cost – improved line-characteristic using costlier dielectric materials, blind-via techniques,counterboring of backplane press-fit connector vias.**
  - – **Density – reduced signal density at linecard-backplane interface allows for cheaper PCB manufacturing options**

Package to PCB transition

Line card trace

Line card trace

Backplane trace

Backplane connector

Backplane connector

IC

Line card via

Backplane via

Line cards
8 × 8 × 40 × 2 Gb/s = 5.12 Tb/s

5 RU = 8.75"

Backplane
128 port × 40 × 2 Gb/s = 10.24 Tb/s

System interconnect hierarchy and advanced optical solutions

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

Frame-to-Frame

Board-to-Board

"LAN"

Substrate-to-Substrate

Chip-to-Chip

Conventional
Optical
Data Link

Shelf-to-Shelf

Gate-to-Gate

POLO
Parallel Optical Data Link

Electronics

MIAUI

Fiber to the processor
applications

PONI
Parallel Optical
Interconnect

Single atom

Quantum effects accessed
by photonics

Electron Bohr radius in GaAs

Increasing system functionality

Length at which electrical transmission lines are required

| 0.1 nm | 1 nm | 10 nm | 100 nm | 1 µm | 10 µm | 100 µm | 1 mm | 1 cm | 10 cm | 1 m | 10 m | 100 m | 1 km |

| 10 T | 1 T | 100 G | 10 G | 1 G | 100 M | 10 M | 1 M | 100 k | 10 k |

Transfer bit rate

# Parallel optical interconnect products emerge from DARPA funded POLO – PONI – MAUI programs

POLO (1994 – 1997)

PONI (1997 – 2000) - inspired products for 10 m – 600 m interconnect lengths: **Agilent, Zarlink, Picolight, Gore, Emcore, Paracer, E20, Silicon Light Machines, Cielo**
Agilent announced 12 x 3.3 Gb/s = 40 Gb/s November 2000
Full production November 2001, customers: Nortel, Cisco, IBM
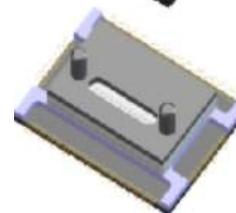**12 x 10 Gb/s = 120 Gb/s demonstrated 2003**

1995 ——————→ 2000 ——————→ 2004 ——→ *time*
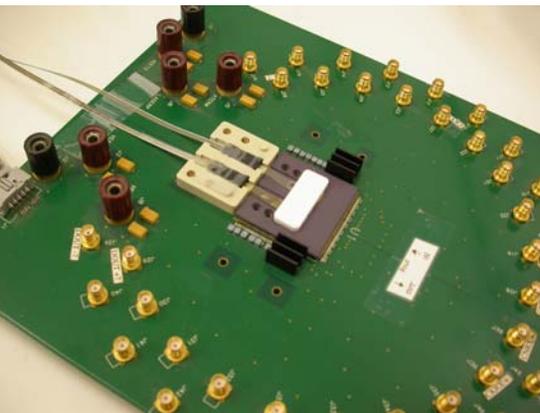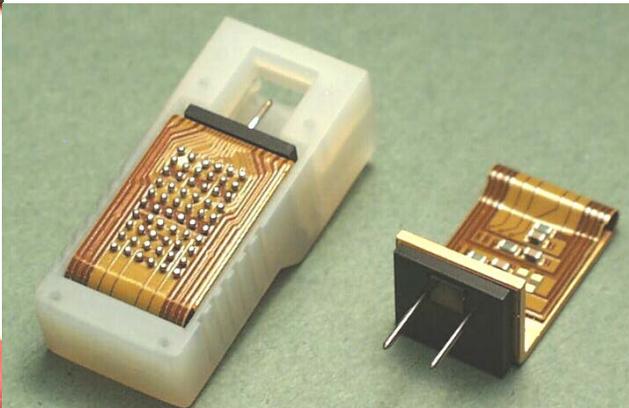
VCSELs / PINs   Guide pin   Passives
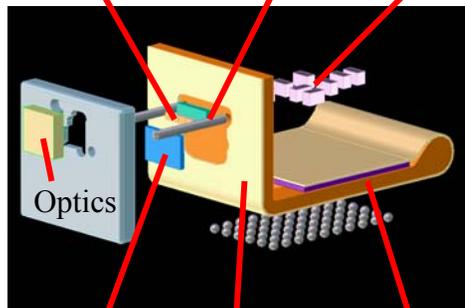
Optics

Silicon IC   Flex circuit   Metal base
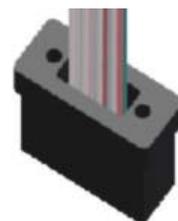
MAUI (2002 – present)
Combination of VCSEL WDM and parallel fiber optic technology for FTTP
1 m – 100 m interconnect length applications
240 Gb/s < 1 W demonstrated 2004

8 mm x 6 mm PMOSA
240 – 1000 Gb/s, < 1W

USC UNIVERSITY OF SOUTHERN CALIFORNIA

MAUI
Multiwavelength Assemblies for Ubiquitous Interconnects
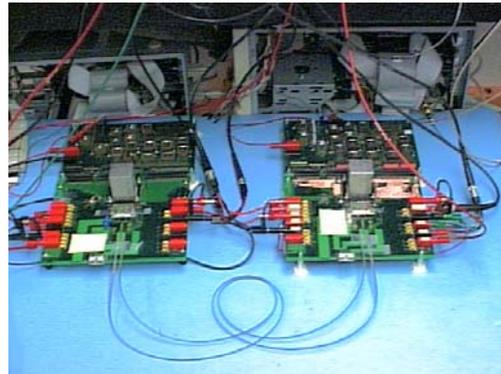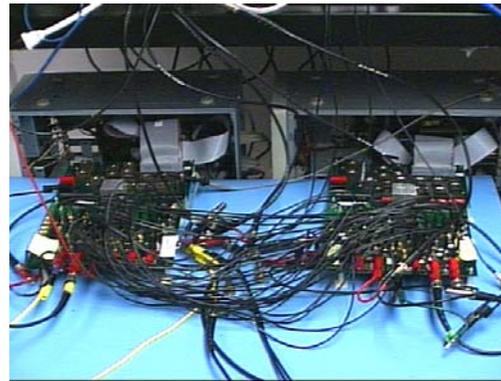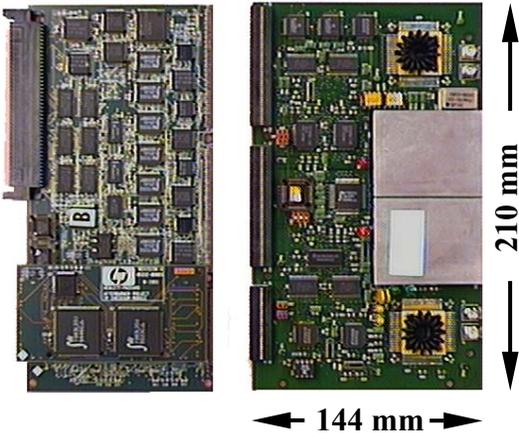
Parallel optics and CMOS integration

**HP experimental
JetStream ring network
1 Gb/s Tx
1 Gb/s Rx**

**Point-to-point host
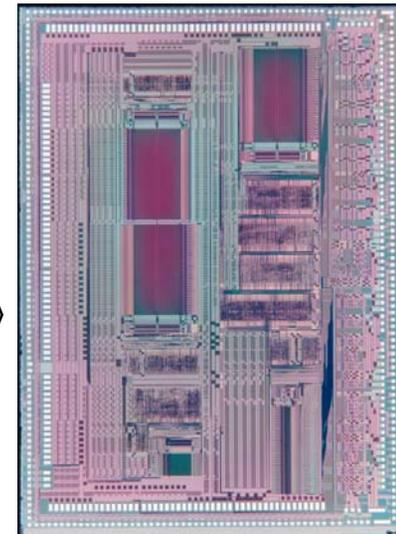interface for parallel optics
16 Gb/s Tx
16 Gb/s Rx**

**Ring network for parallel optics
integrated in single CMOS IC
20 Gb/s Tx
20 Gb/s Rx
20× JetStream on a chip**

Afterburner   **JetStream**

210 mm

← 144 mm →

Host

High-speed parallel fiber-optic interface

**Link Adapter Chip for parallel fiber-optic ring
network**

– 400,000 transistors includes ring MAC

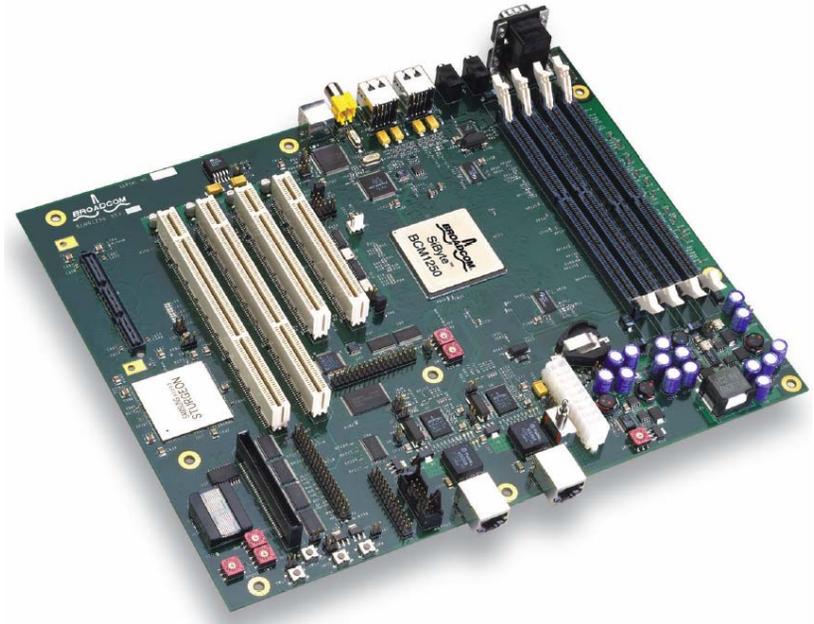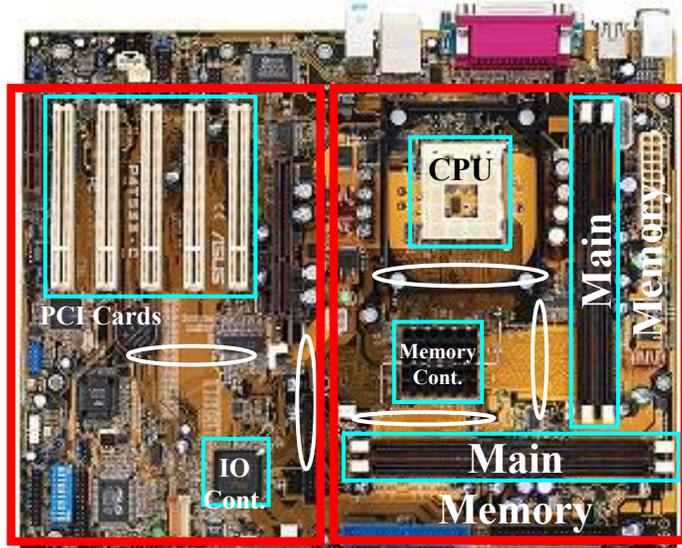– 10.2 mm x 7.3 mm in 0.5 μm CMOS

– tape-out 8.17.00, received 11.10.00

July 1995   **October 1997**   **December 2000**

New markets for optical interconnects: Solving the electronics interconnect and packaging mess!

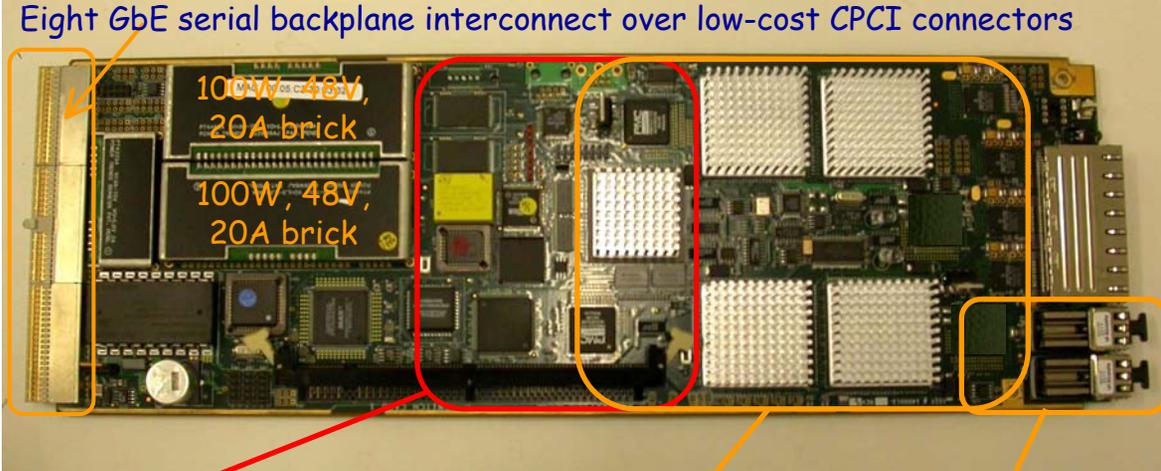**The SAN**   **The memory access bottleneck**

- ➤ **Integration trend places multi-processors on single chip**
  - – **Chip multi-processor (CMP) from Broadcom (SiByte BCM1250)**
- ➤ **Main memory likely to remain separate in most systems**
  - – **10nm CMOS circuits have 100M transistors/mm^2**
    - • **6 transistors per bit in SRAM $\rightarrow$ 16 Mb = 2MB/mm^2 or 200MB/cm^2**
    - • **1 transistor per bit in DRAM $\rightarrow$ 100 Mb = 12MB/mm^2 or 1.2GB/cm^2**
      - – **Might be useful for single-chip notebook computer or make an interesting L2 cache for a CMP**
- ➤ **Multiple processor boards in chassis systems are connected by switches**
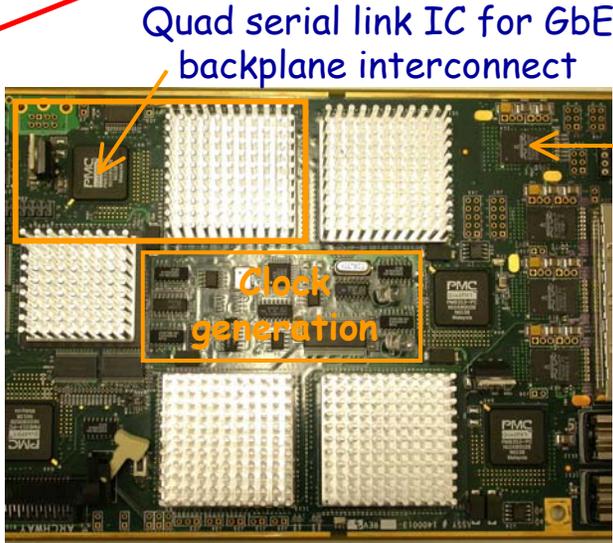
1U (1.75") thick 20-port GbE switch/router for chassis servers (2001)

System example

- **96W, hot-swappable 20-port GbE router**
  - **15.5" x 5.35"**
  - **~2300 components**
  - **~7000 nets, ~11000 pins**
  - **Electrical and optical GbE IO**
    - **8 GbE optical links**
    - **8 GbE backplane links**
    - **4 GbE Cat-5 links**

Eight GbE serial backplane interconnect over low-cost CPCI connectors

Quad serial link IC for GbE backplane interconnect

GbE PHY IC

Clock generation

Management Microprocessor and support circuitry

Quad 8-port, mesh-connected GbE Switch ICs with 20 external ports

SERDES + dual quad-channel MMF optical modules
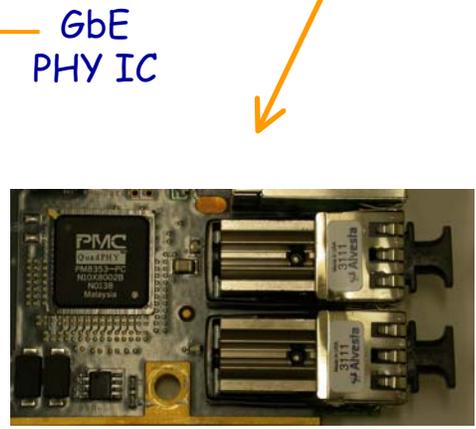
Integration and packing driven processor crisis: The case for fiber-to-the-processor (FTTP)

- ➢ **Electronics fails to deliver**
  - ➢ **Power crisis - projected kW CPU not viable**
  - ➢ ***Processor crisis*** driving multi-core processor design with increased IO demand and only a fraction of transistors being active at any one time
    - ➢ Intel moves to CMP and Pentium IV uni-processor development terminated - 2005
  - ➢ **Bandwidth density and latency crisis**
    - ➢ **increasing mismatch between memory bus bandwidth and CPU**
    - ➢ **many CPU cycles wasted after cache miss**
  - ➢ **Signal integrity crisis**
    - ➢ **EMI, reflections, crosstalk, device noise may lead the way to optical interconnects**
    - ➢ **high-speed electrical signaling not reliable**
      - ➢ $400M i820 memory translator hub recall because of electrical noise - 5.10.00
      - ➢ 1.13 GHz PIII recall because of electrical noise in circuit element - 8.28.00
- ➢ **Fiber-to-the-processor is a new design point**
  - ➢ **Less power, less power *density* in distributed system using WDM SAN**
  - ➢ **Better signal integrity, optical isolation**
  - ➢ **More bandwidth density gives reduced latency in node and SAN**
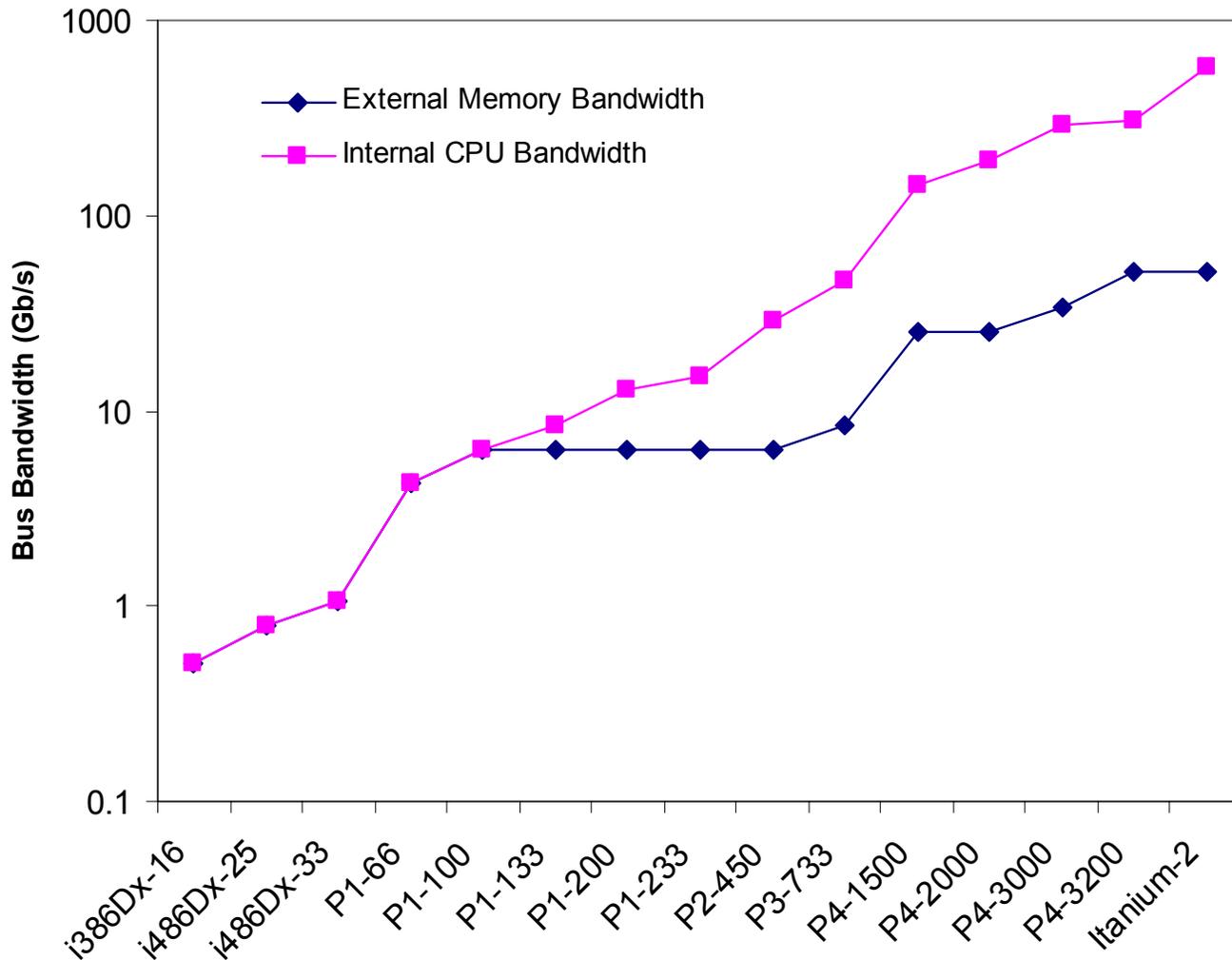 - ➢ **Removes electrical backplane bottleneck for future multi-processor systems**
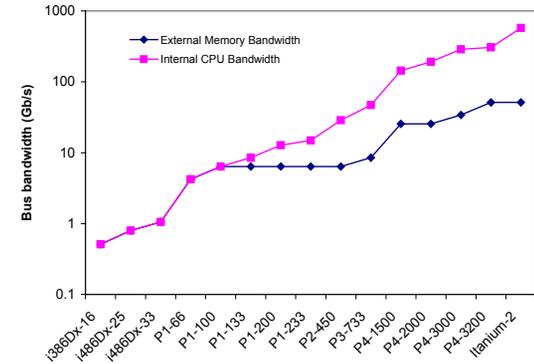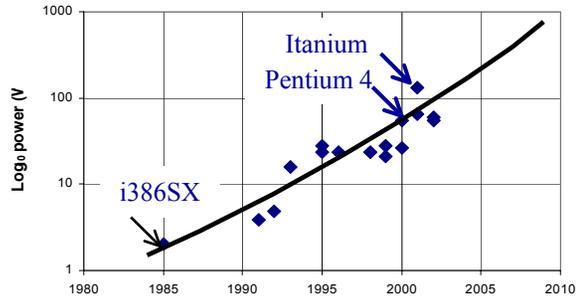
accounts for superscalar microprocessor architecture by multiplying internal datapath width by the number of instructions that can be issued simultaneously.

Ethernet switch-port deployment

Ethernet data-rate deployment

Moore's Law 2× every 2 years
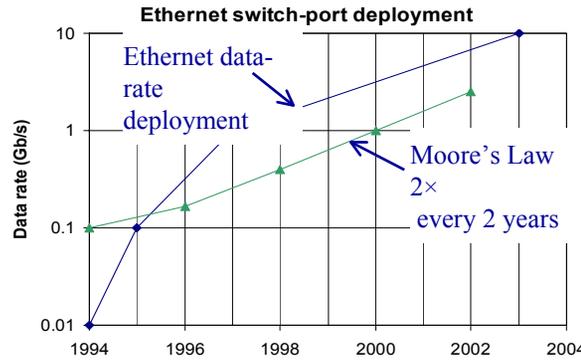
Moore's Law: On-chip high-performance local clock (SIA 97)

Optical interconnects and the memory access bottleneck

FTTP

Bus Bandwidth (Gb/s) vs processor generations (i386Dx-16, i486Dx-25, i486Dx-33, P1-66, P1-100, P1-133, P1-200, P1-233, P2-450, P3-733, P4-1500, P4-2000, P4-3000, P4-3200, Itanium-2)

- External Memory Bandwidth
- Internal CPU Bandwidth

Optical interconnect can fill the memory-access performance gap with bandwidth edge density of 60 – 600 Gb/s/mm

FTTP: A new architecture enabled by optical interconnects and high-performance CMOS integration

System level issues

- ➢ **New technology**
  - **Optical interconnect**
    - **Ultra-high bandwidth**
    - **Low power**
    - **Low latency**
- ➢ **Integration**
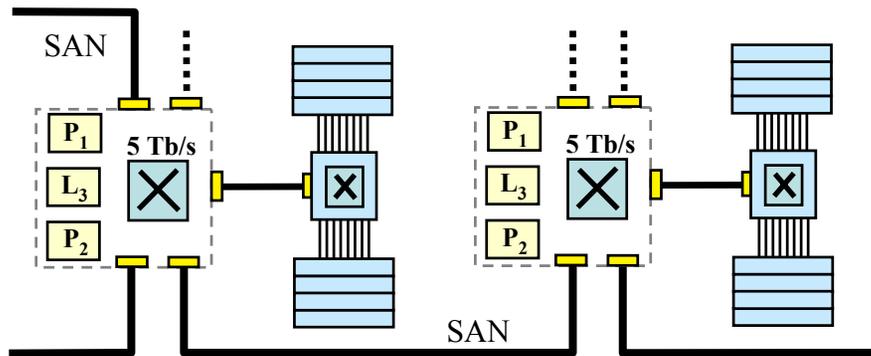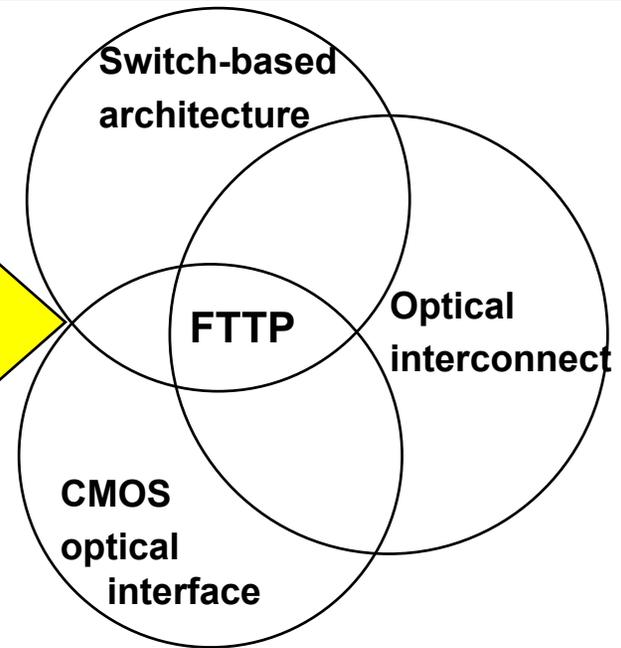 - **CMOS interface to optics**
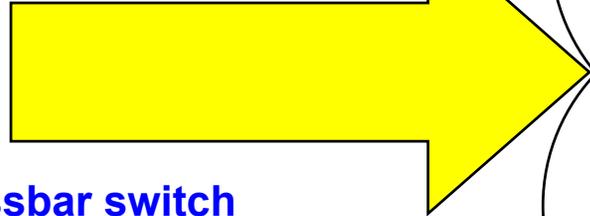   - **High-performance crossbar switch**
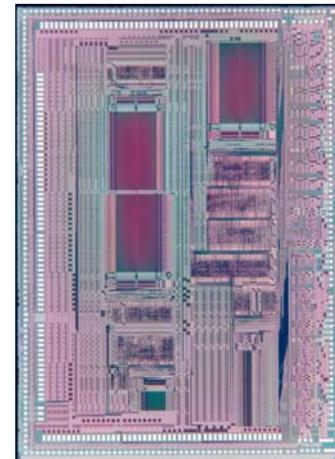- ➢ **New switch-based architecture**
  - **Next generation scalable NUMA**
    - **Switch integrated in processor and memory**

Driving to a "*technology convergence point*"

Switch-based architecture

Optical interconnect

FTTP

CMOS optical interface
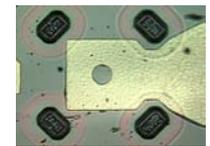
Multi-processor switched-based network

High-performance CMOS interface

Parallel optics and WDM

VCSEL

Example latency estimate

10 Cy at 125 MHz (80 ns)

5 Cy at 500 MHz (10 ns)

4+4 Cy at 500 MHz (16 ns)

15 Cy at 500 MHz (30 ns)

4+4 Cy at 500 MHz (16 ns)

15 Cy at 500 MHz (30 ns)

P   P   P   P

Ctl — Memory   Ctl — Memory

Cross Bar ⟷ Cross Bar

16 ns   30 ns

Cross Bar ⟷ Cross Bar   16 ns

20 ns

Ctl — Memory   Ctl — Memory

10ns   50ns

P   P   P   P

Round-trip time

80 ns
+ 10 ns
+ 16 ns
+ 30 ns
+ 16 ns
+ 30 ns
+ 16 ns
+ 30 ns
+ 16 ns
+ 10 ns
+ 20 ns
+ 50 ns

Round-trip time per segment

= 324 ns
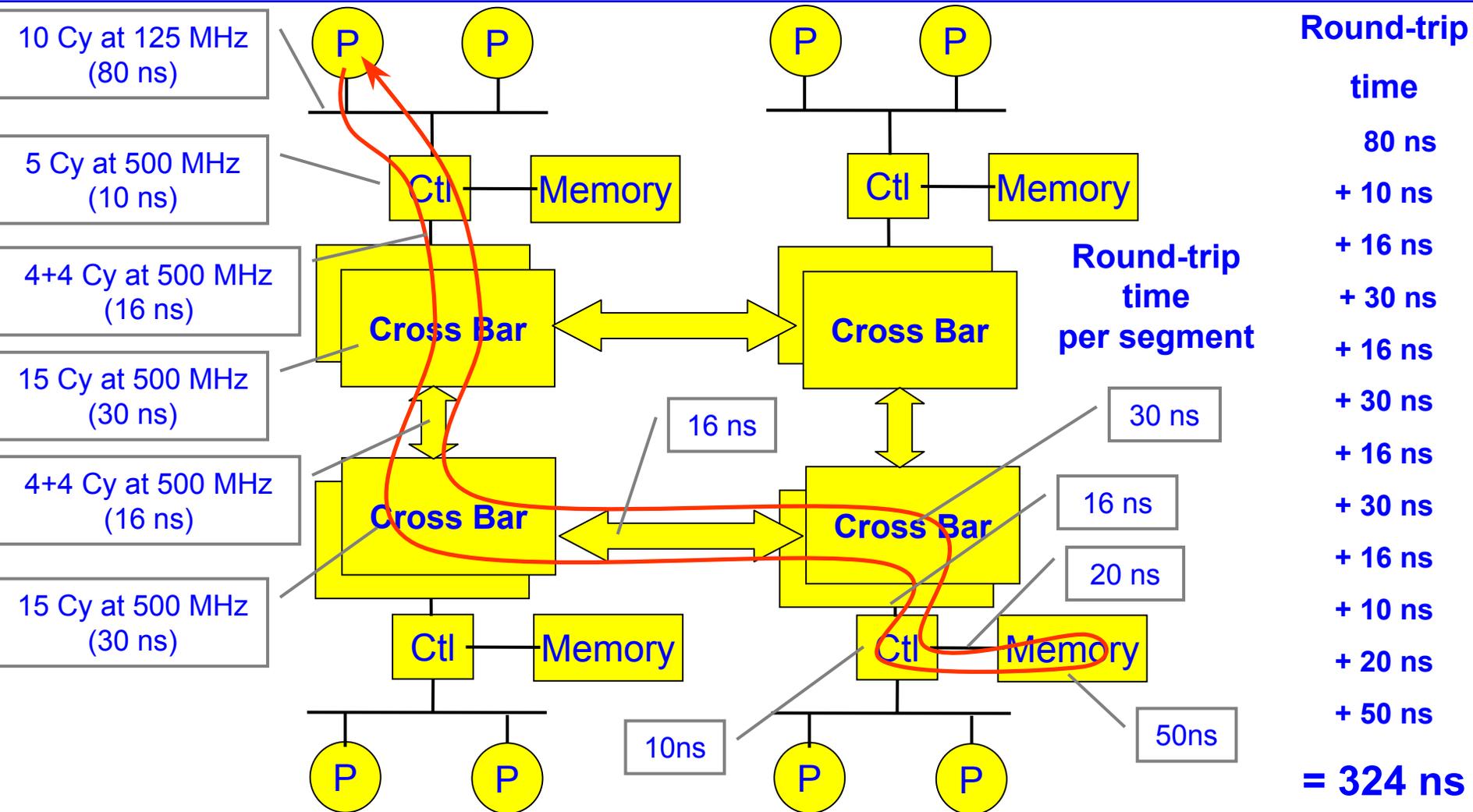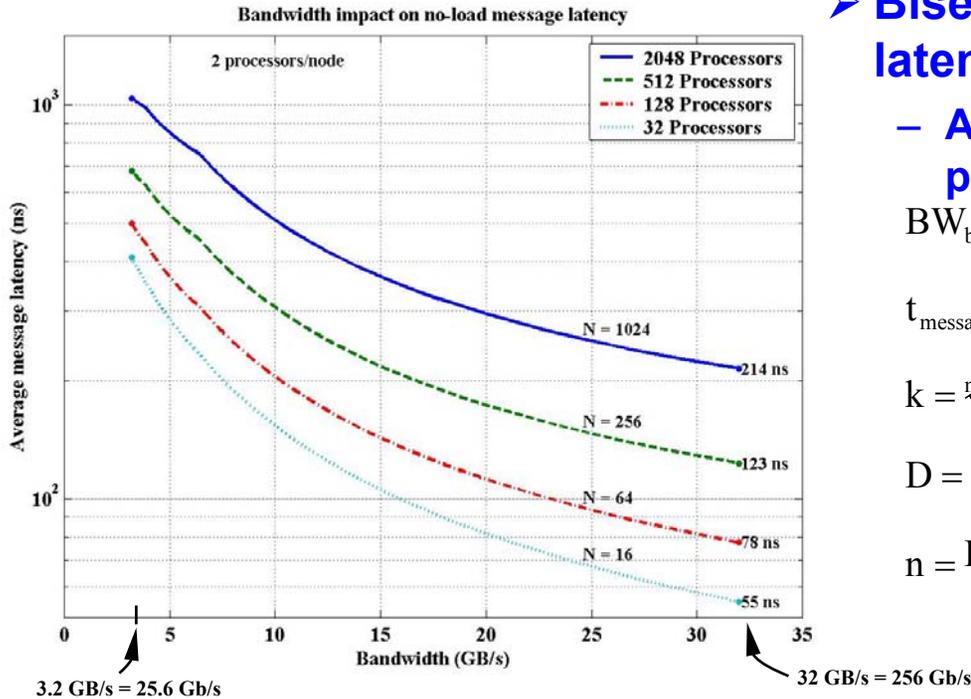
➤**10× increase in clock rate reduces round-trip time ~10×**

➤**Assume time-of-flight ~ 0 ns**

System impact of increased available bandwidth: Reduced message latency and improved scaling

Bandwidth impact on no-load message latency

2 processors/node

- 2048 Processors
- 512 Processors
- 128 Processors
- 32 Processors

10^3

Average message latency (ns)

10^2

N = 1024 → 214 ns

N = 256 → 123 ns

N = 64 → 78 ns

N = 16 → 55 ns

Bandwidth (GB/s): 0, 5, 10, 15, 20, 25, 30, 35

3.2 GB/s = 25.6 Gb/s

32 GB/s = 256 Gb/s

➢ **The 4-SAN ports can be used to design a 2-D torus with $N = k^2$ processors ($n = 2$, $N = [16, 64, 256, 1024]$)**

➢ **Message latency is**

$$t_{message\_latency} = \frac{k}{2}(t_r + t_s + t_w) + \frac{L}{BW}$$

➢ **For 32 processor network**

  – **32 GB/s, 4-port switch achieve × 1.5 better no-load average message latency compared with to a 20 GB/s, 6-port switch**

  • **(× 1.36 better no-load average message latency for 2048 processors)**

➢ **Bisection-bandwidth and message latency for a k-array n-cube network**

  – **A network with *n*-dimensions and k-nodes per dimension**

$$BW_{bisection} = 2 \cdot BW_{port} \cdot k^{n-1}$$

$$t_{message\_latency} = D \cdot (t_r + t_s + t_w) + \frac{L}{BW}$$

$$k = \sqrt[n]{N}$$

$$D = n \cdot \frac{k}{4}$$

$$n = \frac{Ports}{2}$$

Where N → Total number of nodes

k → Number of nodes in each dimension

n → number of dimensions

D → Average distance between any pair of nodes
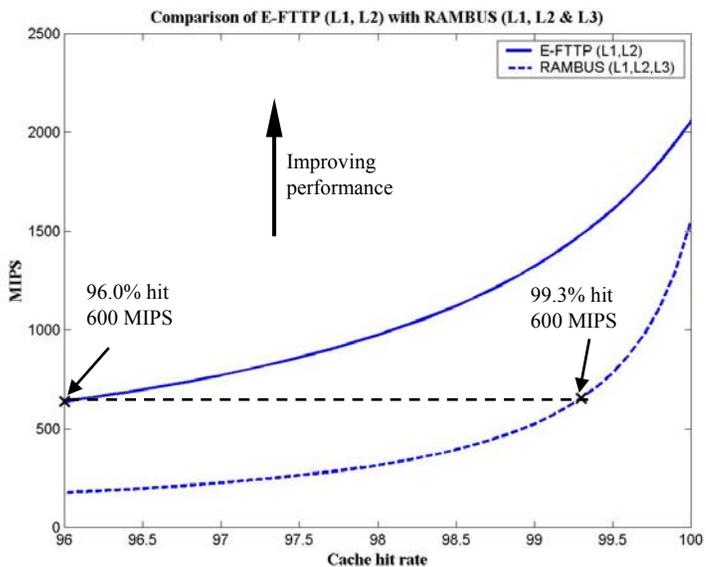
t_r → Time to make routing decision (10 cycles, < 20 ns)

t_s → The delay through switch (6 cycles, < 20 ns)

t_w → The interconnection delay (1.0 m hop length)

BW → Bandwidth of each port = B × W, Where B is the bandwidth of each line, and W is port width

L → Packet length (1 kB)
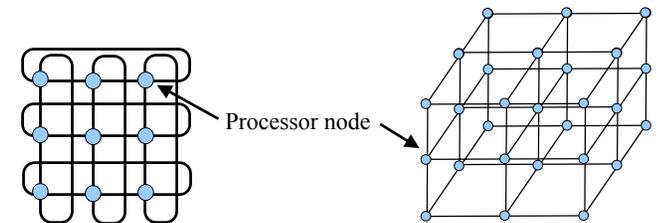
Processor node

3-array, 2-cube (2-D torus)

3-array, 3-cube (3-D torus) wrap-around not shown

System impact of reduced cache miss

Impact of memory access latency on system performance: With L1, L2 & L3 Caches

Comparison of E-FTTP (L1, L2) with RAMBUS (L1, L2 & L3)
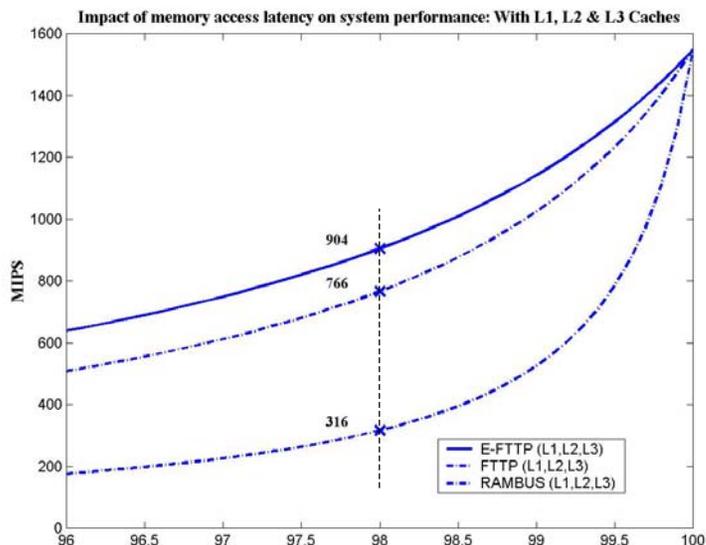
Simulation assumptions
- **L1 hit rate - 90% (based on third party test results)**
  - http://www.aceshardware.com/Spades/read.php?article_id=20000190
- **L2 access latency - 9 cycles (based on P4)**
  - http://www.aceshardware.com/Spades/read.php?article_id=20000190
- **L3 access latency - 20 cycles (based on Merced)**
  - http://www.geek.com/procspec/features/itanium/index.htm
  - **Assume 96% of the memory access is satisfied by L1 and L2.**
- **5.0 GHz processor speed**
- **1.3 cycles per instruction**
  - **Using Intel assumptions**
    - http://developer.intel.com/design/pentium4/manuals/248966.htm
    - **Each instruction is sub-divided into micro-ops during execution**

Impact of memory access bandwidth on cache hit rate not taken into account
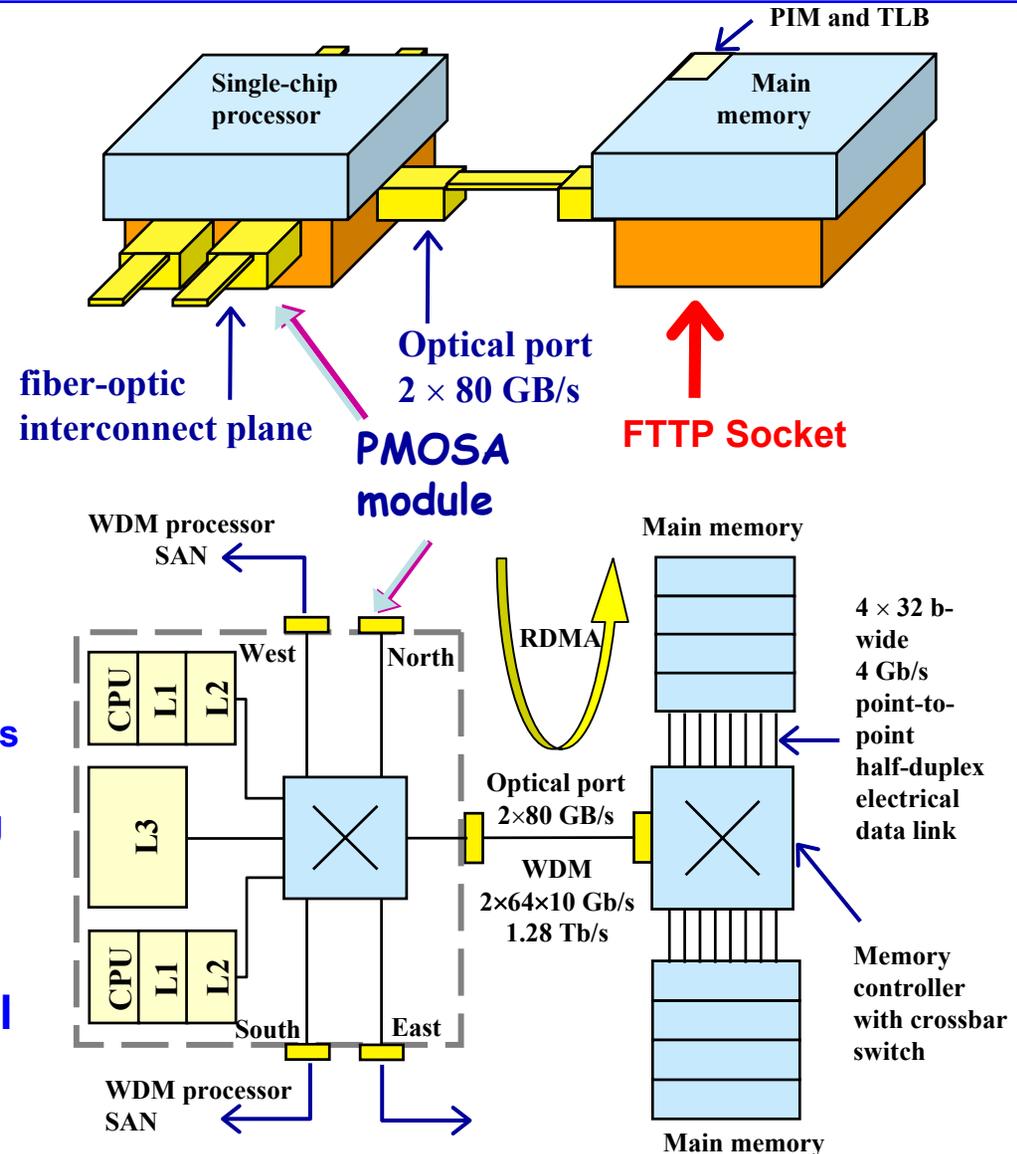- **Improved BW improves hit-rate because of reduced pre-fetch distance**

Performance of FTTP with only L2 cache and 96% cache hit rate is equal to RAMBUS with L2 and L3 with 99.3% cache hit rate
- **Adding a L3 cache to hide memory access latency does not out perform FTTP**
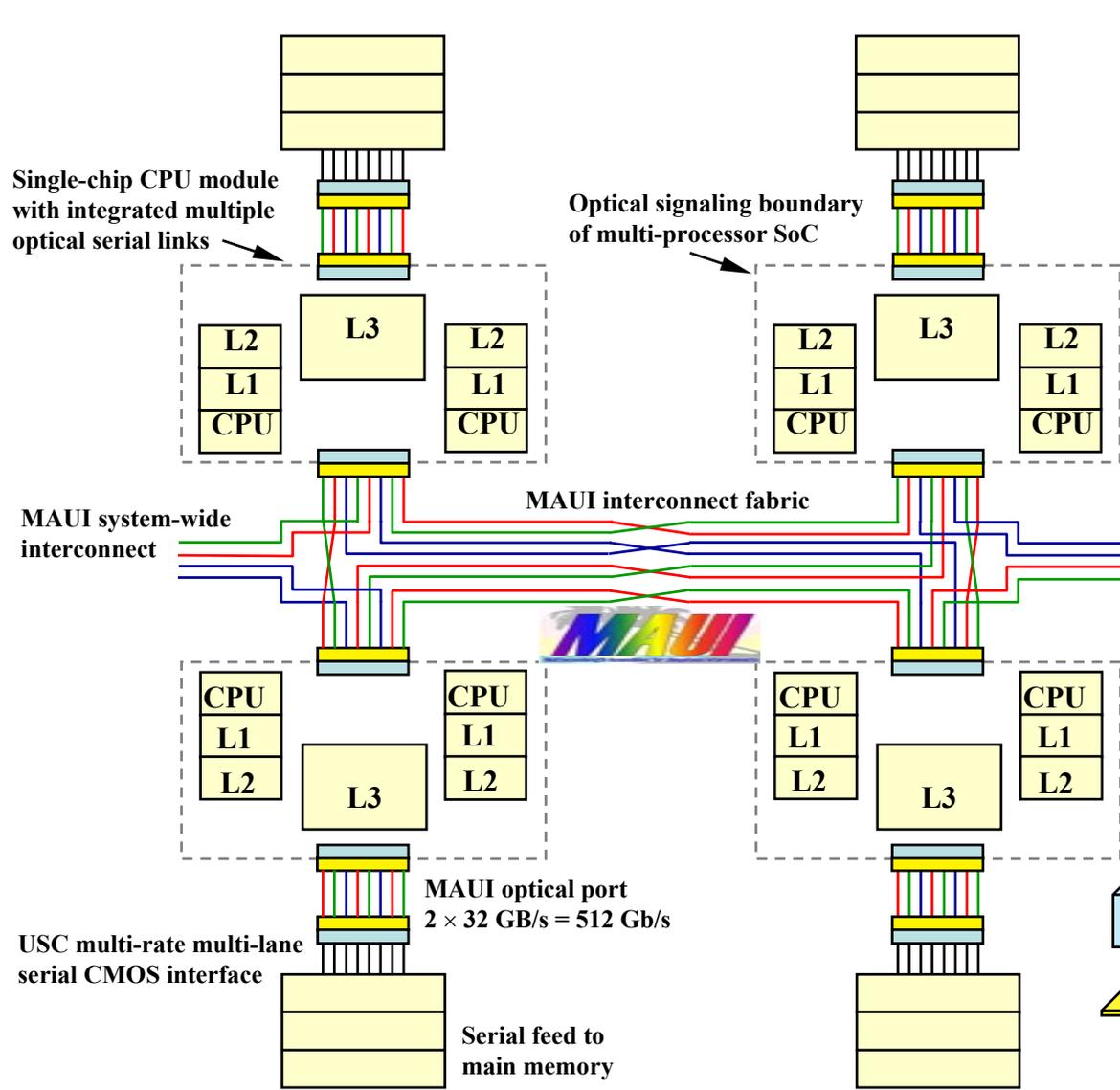
Fiber-to-the processor: Exposing raw CPU performance

System level issues

- ➢ **Single-chip multi-CPU module with integrated switch and optical system area network (SAN)**
  - – **SoC internal bandwidth 10GHz×128×2×2=5.12Tb/s**
- ➢ **Main memory module with high-performance optical IO port**
- ➢ **All off-chip high-speed signals are optical**
  - – **1.28 Tb/s×5 ports = 6.4 Tb/s SoC IO bisection bandwidth**
    - • **RDMA ready**
    - • **1RU electrical backplane supports only two (2) SoC processors**
    - • **Number of SoC processors using FTTP backplane determined by power dissipation**
- ➢ **All off-chip slow-speed signals are electrical (including electrical power)**

PIM and TLB

Single-chip processor

Main memory

Optical port 2 × 80 GB/s

fiber-optic interconnect plane

PMOSA module

FTTP Socket

WDM processor SAN

Main memory

RDMA

West   **North**

CPU L1 L2

L3

Optical port 2×80 GB/s

WDM 2×64×10 Gb/s 1.28 Tb/s

CPU L1 L2

South   **East**

WDM processor SAN

4 × 32 b-wide 4 Gb/s point-to-point half-duplex electrical data link

Memory controller with crossbar switch

Main memory

FTTP exposes raw CPU performance with multiple serial optical chip-to-chip interconnects

Single-chip CPU module with integrated multiple optical serial links

Optical signaling boundary of multi-processor SoC

L2	L3	L2
L1		L1
CPU		CPU

MAUI interconnect fabric

MAUI system-wide interconnect

CPU	L3	CPU
L1		L1
L2		L2

MAUI optical port
2×32 GB/s = 512 Gb/s

USC multi-rate multi-lane serial CMOS interface

Serial feed to main memory

Single-chip processor

PIM and TLB

Main memory

Optical port
2×32 GB/s

Socket

fiber-optic interconnect plane

Socket

- ➤ **Single-chip CPU module (SoC) with FTTP optical interface**
- ➤ **Main memory module with high-performance optical port**
  - – **Serial main memory fed by optical/CMOS interface**
- ➤ **All off-chip high-speed signals are optical**
- ➤ **All off-chip slow-speed signals are electrical (including electrical power)**
- ➤ **Key FTTP enablers:**
  - **Agilent MAUI optical sub-assembly**
  - **USC multi-rate multi-lane serial CMOS interface**
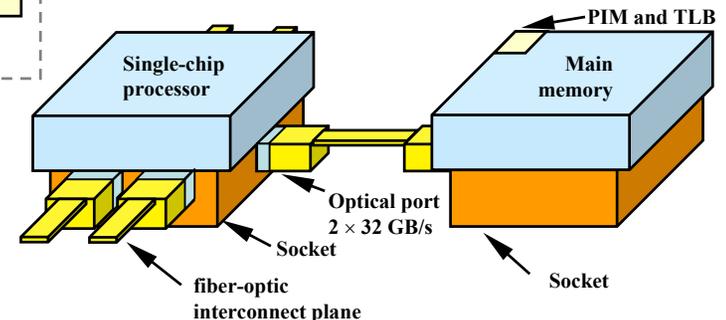
Flip-chip optical socket LGA concept
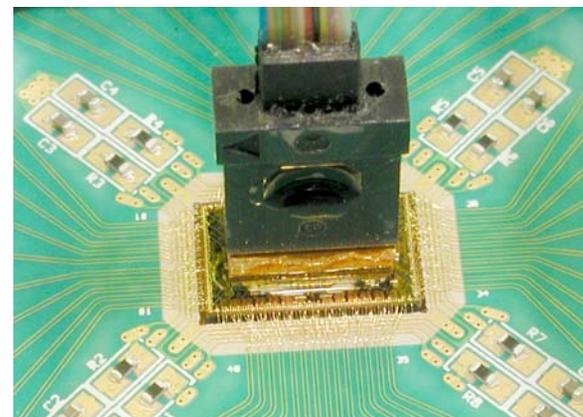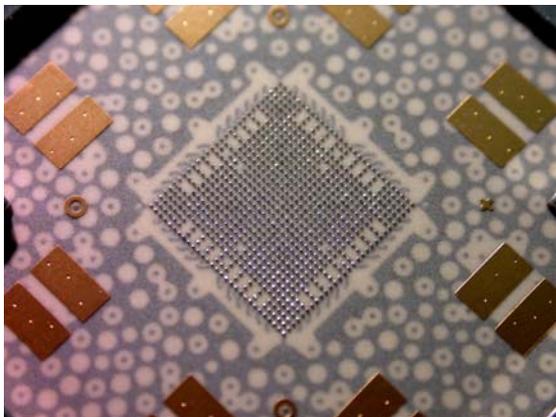
Agilent / MAUI – DARPA program

- Today at USC: 1.27mm pitch FC-LGA, 40 x 40 mm^2, 960-pin, Rogers 2800 dielectric, estimated price $30 in 10k volume
- 212.5 mm center-to-center IC pad-pitch
  - Option 1: 6.5 x 6.5 mm^2 IC = 216 diff IO
  - Option 2: 5.0 x 5.0 mm^2 IC = 108 diff IO
- Package performance
  - -3dB > 20 GHz, NEXT < -30 dB
  - Can be improved to -3dB ~40 GHz, NEXT < -30 dB
- Easily modified to implement "optical socket" for fiber to the processor

- Package level optical interconnect for inter-chip *optical buses*
- 8mm x 5mm chip scale optical port is a prototype today
- Today: 0.48 Tb/s, <2W unidirectional fiber-optic port
- Future: >1 Tb/s, <1W unidirectional fiber-optic port
- Includes alignment pins for MT-ferrule with 12-fiber ribbon
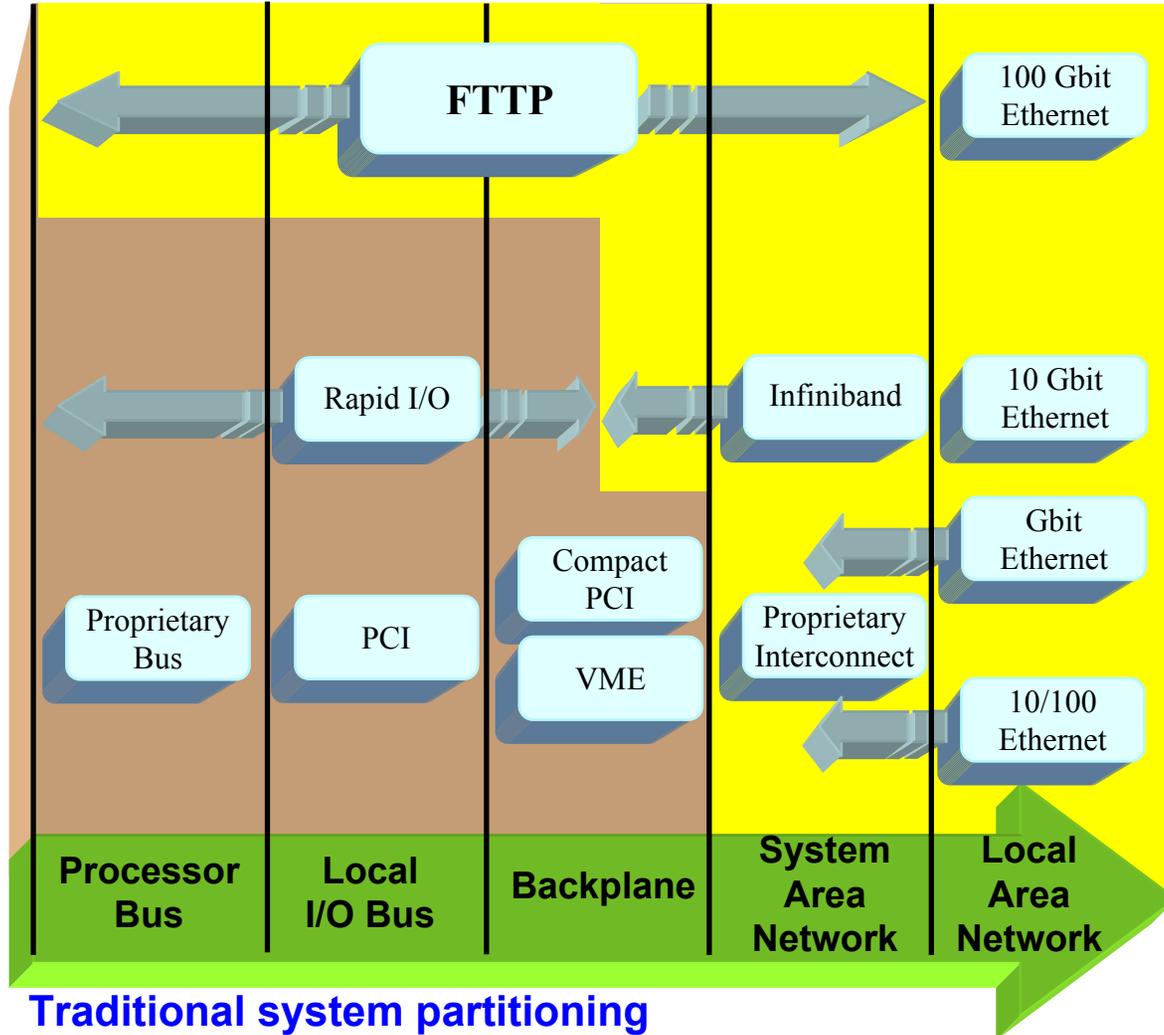
A system architecture roadmap: The FTTP opportunity

Minimum 1 Tb/s/port × 5 ports/chip

Technology insertion

Increasing system integration

Time

Processor Bus	Local I/O Bus	Backplane	System Area Network	Local Area Network

FTTP — 100 Gbit Ethernet

Rapid I/O — Infiniband — 10 Gbit Ethernet

Compact PCI — Gbit Ethernet

Proprietary Bus — PCI — VME — Proprietary Interconnect — 10/100 Ethernet

2010

2000

Traditional system partitioning and increasing interconnect length scale
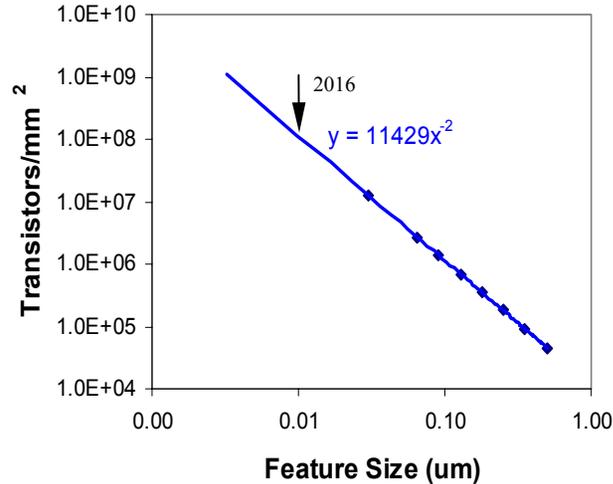
The cost of myths

- ➢ '*Optics will not speed up memory access*'
  - – **said Howard Davidson, OIDA, October 21, 2004, Burlingame, CA.**
  - – **Actually only true for for SMP and its current programming model in which latency is dominated by global directory coherency**
    - • **NUMA, which has local coherency, does not suffer from this problem – but you have to change your software**
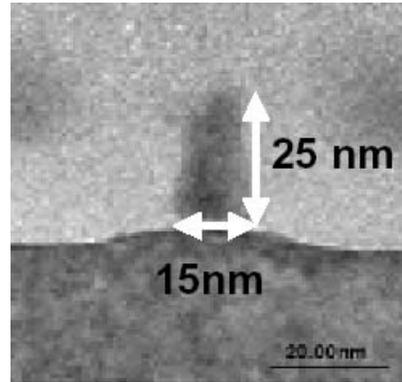- ➢ **Embracing myths as truths avoids the need to innovate**
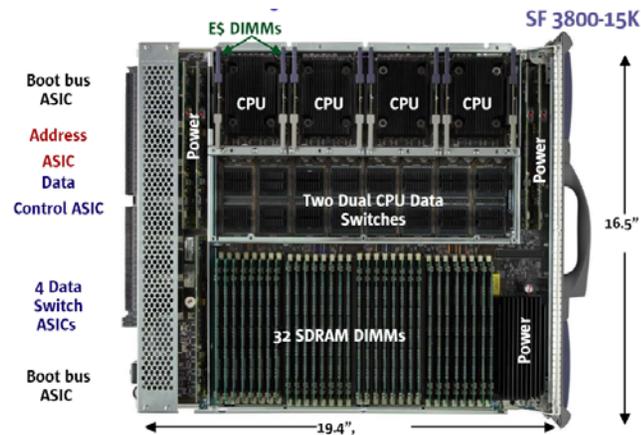
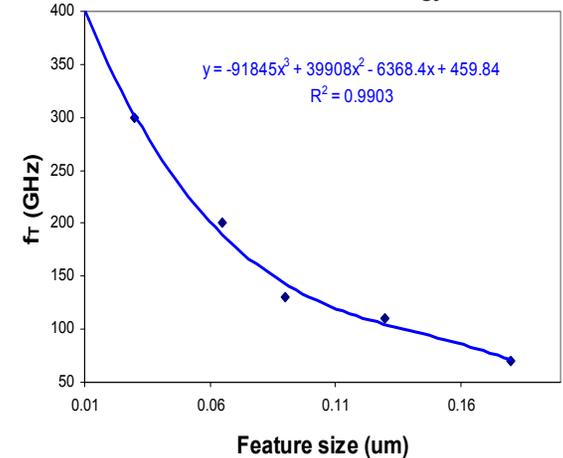# Impact of decreasing CMOS device feature size on interconnect: 80 Gb/s serial IO

Transistor density versus minimum CMOS feature size

2016

$y = 11429x^{-2}$

Intel 11/2001

25 nm

15nm

20.00 nm

f_T versus CMOS Technology

$y = -91845x^3 + 39908x^2 - 6368.4x + 459.84$
$R^2 = 0.9903$

f_T (GHz)

Feature size (um)

Pad Characteristics

FC pad pitch (um)
High Peformance IO pad count

FC Pad Pitch (mm)

High Performance ASIC IO count

Year since 1958

➢ **Transistor scaling to 10 nm CMOS by 2016**

- **100 M transistors/mm² (2 Intel Pentium-IV processors)**
  - **Scaling fails due to IO, on-chip wiring, and V_{dd} ~ 0.8 V to give 10-60 W power dissipation**
- **80 Gb/s IO based on PAM-4, f_T > 400 GHz and 400 mW**
- **High-speed IO pad-pitch improvement limited by crosstalk and package material properties**
- **75 μm pad diameter and 150 μm pitch**
- **36 bond-pads/mm²**
- **9 differential pair IO/mm²**
- **18 power and ground pads/mm²**

IC IO density 75 μm dia

150 μm

150 μm
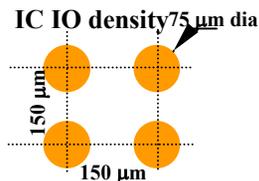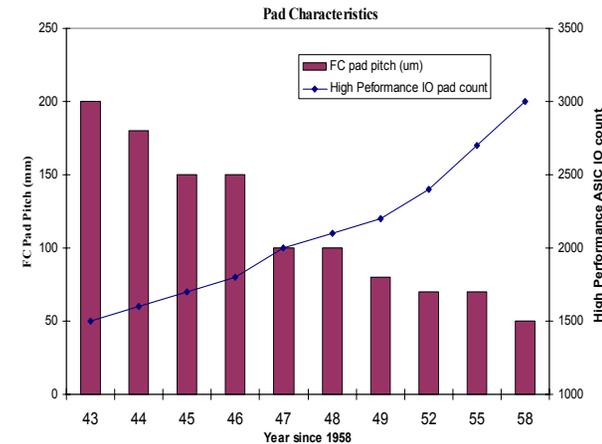
NRZ
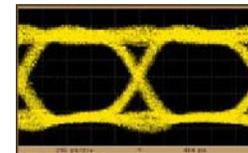
PAM-4
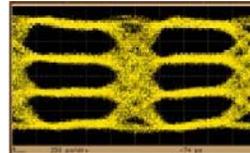
Challenges for electronics and photonics driven by CMOS scaling

Electronics

Computation

| Proc. | Mem | Comm |

Communication

| Pkg | trace | Connector |

- **10 nm CMOS, f_T > 400 GHz, < 10^{-18} J switching energy**
  - **10 – 12 metal layers**
  - **100 transistors/μm^2 for random logic**
  - **500 transistors/μm^2 for SRAM cells**
    - **0.0122 μm^2 /SRAM single-port cell**
  - **100M transistors/mm^2**
    - **2 Pentium-IV/mm^2**
  - **80 Gb/s IO (PAM-4 and f_T > 400 GHz)**
- **Integration implies high power density ~ 10-60 W/mm^2**
  - **Assumes 110 ^oC junction temperature**
  - **Si thermal conductivity κ = 1.5 W/cm ^oC**
  - **Forces 10 mm^2 area (~ 1-6 W/mm^2) for 100M transistor circuit in 10 nm CMOS (*or* liquid cooling …)**
    - **Distributed architecture on chip**
    - **Benefit from large f_T to reduce power and use high-speed serial IO to reduce packaging cost**
    - **Remaining area for power regulation, RF-style and analog elements, self-test, calibration**

- **Controlled-impedance launch to package trace with S11 < -10 dB restricts flip-chip IO pitch on IC/Pkg to 150 μm pitch**
  - **9 Differential IO/mm^2, suggests high-speed serial that also reduces backplane design effort**
- **Low-loss (< -3 dB), low-crosstalk (< -30 dB), dense IO electrical packages requires**
 - **tan δ < 0.002**
  - **ε_r < 2.5**
  - **Via technology**
    - **High-aspect ratio, blind-via, tight pad overlap of via, relatively tight registration**
- **Low-loss tangent PCB dielectric (tan δ < 0.002)**
- **High density, *perfect* electrical backplane connector is required that is mechanically reliable, manufacturable, low-cost, low-NEXT, and impedance-matched at data rate**

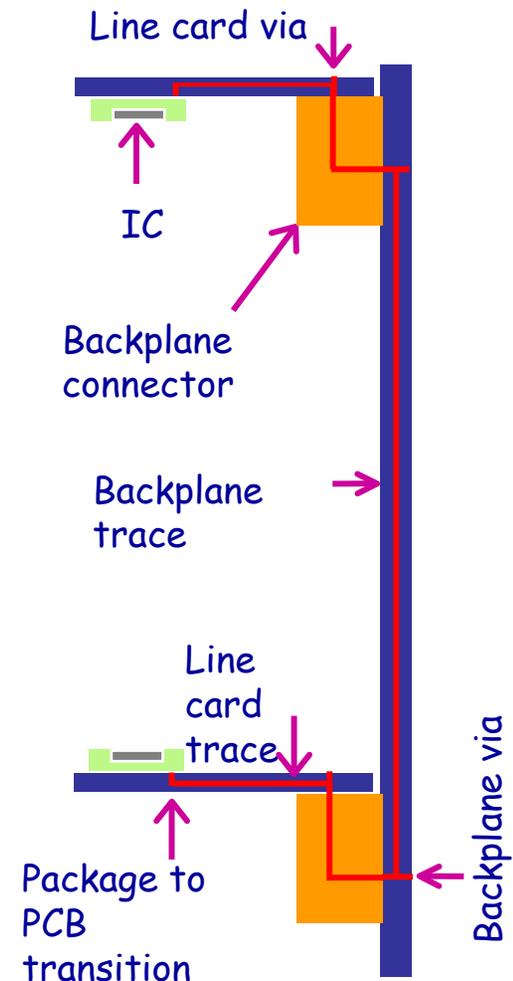# Challenges for electronics and photonics driven by Moore's Law CMOS scaling

Photonics

| Computation | | | Communication | | |

| Proc. | Mem | Comm | Pkg | trace | Connector |
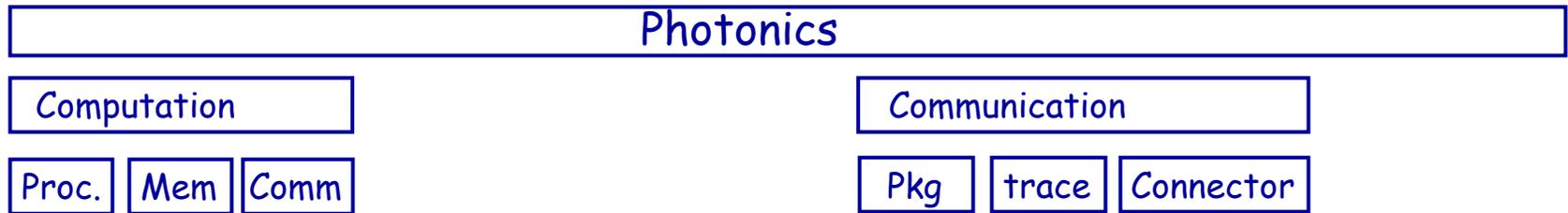
- **Optical logic and memory not practical at present time**

- **Optical devices cannot match electronic feature size (100 transistors/μm^2 in 10 nm CMOS) and efficiency or approach computational equivalence for digital processing**

- **Electronic interface to optical devices potentially limited by:**
  - **Bias voltage and current**
  - **Drive voltage and current**
  - **Intimacy of integration requiring fan-in/fan-out of controlled impedance lines**
  - **Harsh thermal, mechanical, electromagnetic environment**
  - ***Slow speed* photonic devices!**
    - **≤ 20 Gb/s digital modulation of laser diodes**

- **Fiber optics superior to electrical interconnect on length scales ≥ 1 m, using metrics of signal loss, power dissipation and bandwidth**
  - **Lower-power, higher-impedance lines can be used to interface electronics to optical devices.**
  - **"*Optical PCB-trace*" required for intra-chassis interconnect**

- **Optical connector has superior form-factor (3× – 10×) compared to electrical connector**
  - **Low-cost line-card to backplane version of parallel-optics connector needed to enable optical interconnect in chassis**

- **Conclude photonics useful for communication in systems but presently limited by *slow speed* photonic devices and incompatibility with PAM-4**
  - **≤ 20 Gb/s digital modulation of laser diodes**
  - **Message latency**
    - **0.5 ns conversion latency**
    - **20 Gb/s optical vs. 80 Gb/s electrical**
      - **64 B message per signal line 25.6 ns optical, 6.4 ns electrical**

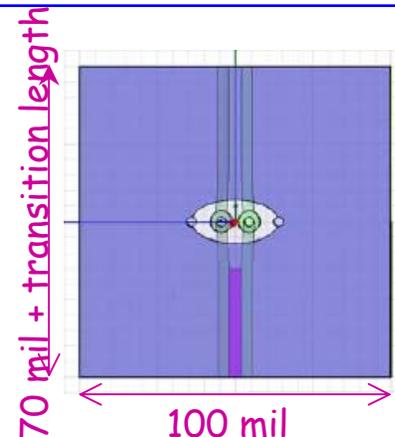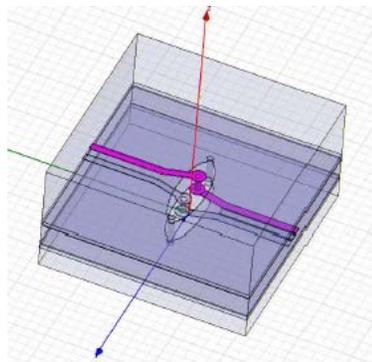# IO bandwidth example for 10 nm / 50 nm CMOS IC

➢ **10 GHz Pentium-X with two-cores, 2 IPC, 64-bit wide internal bus**

  – **5.12 Tb/s bi-directional total internal data bandwidth of two-core IC**

  – **Estimate 64 bits × 10 Gb/s = 0.64 Tb/s bi-directional external-CPU bandwidth**

    • **1.28 Tb/s bisection bandwidth with dedicated unidirectional IO buses for wide-slow interconnect or multiple thin-fast serial links**

➢ **10 nm CMOS (640 Gb/s/mm^2)**

  – **Thin-fast, 150 μm pad-pitch = 18 pads/mm^2 for IO**

    • **8-bit wide datapath using 80 Gb/s = 640 Gb/s/mm^2 (unidirectional)**

      – **Requires 16 signal pins, 50% power-pad/ground-pad rule total 32 pins**

  – **Wide-slow**

    • **128-bit wide datapath using 5 Gb/s = 640 Gb/s (unidirectional)**

      – **Requires 256 signal pins, 50% power-pad/ground-pad rule total 512 pins**

➢ **50 nm CMOS (320 Gb/s/mm^2)**

  – **Thin-fast, 150 μm pad-pitch = 18 pads/mm^2 for IO**

    • **8-bit wide datapath using 40 Gb/s = 320 Gb/s (unidirectional)**

  – **Wide-slow**

    • **128-bit wide datapath using 2.5 Gb/s = 320 Gb/s (unidirectional)**

Line card via

IC

Backplane connector

Backplane trace

Line card trace

Package to PCB transition

Backplane via

40GHz Differential PCB via simulation test fixture

70 mil + transition length

100 mil

➤ **Parameterized Ansoft HFSSv9.1 test structure for 100-ohm differential microstrip-stripline transition**

  – **RO4503 (ε_r=3.48, tan δ = 0.004), trace is copper (5.8E7 S/m), surface roughness not considered, Radiation boundaries on all sides**

  – **7-mil wide trace, 8-mil space, 1.2mil thick planes**

    • **Microstrip: 1.2-mil thick trace, 4-mil dielectric**
    • **Stripline : 0.7-mil thick trace, 16.7-mil dielectric**

  – **100-mil microstrip, 100-mil stripline, 15.7-mil tall via, NO via stub**

➤ **Number of geometrical parameters associated with transition varied to determine best fit (least SDD11, max SDD21)**

  – **Ground plane opening (major and minor axes of ellipse), which affects spacing of guard vias**

  – **Relative spacing of trace vias, transition length to vias**

# Six via model, 33"model - 3 sections of microstrip(0.5")-stripline(10")-microstrip(0.5")
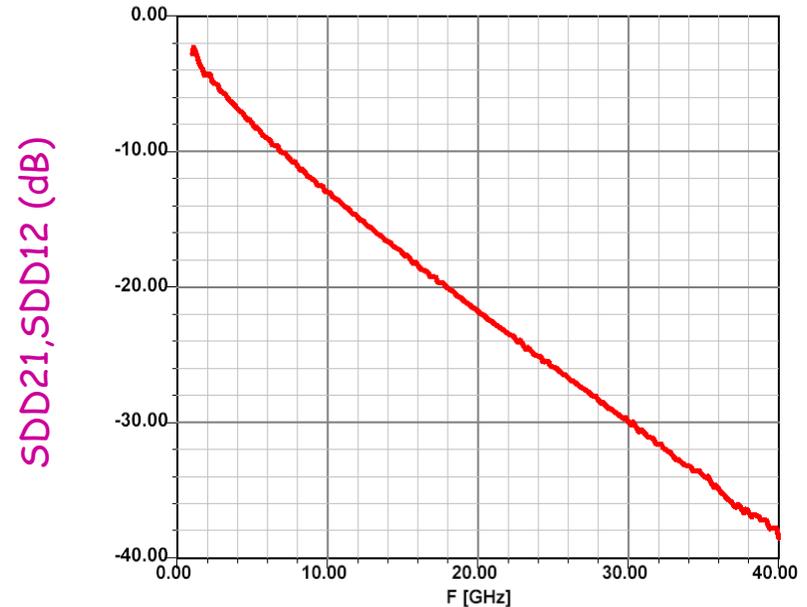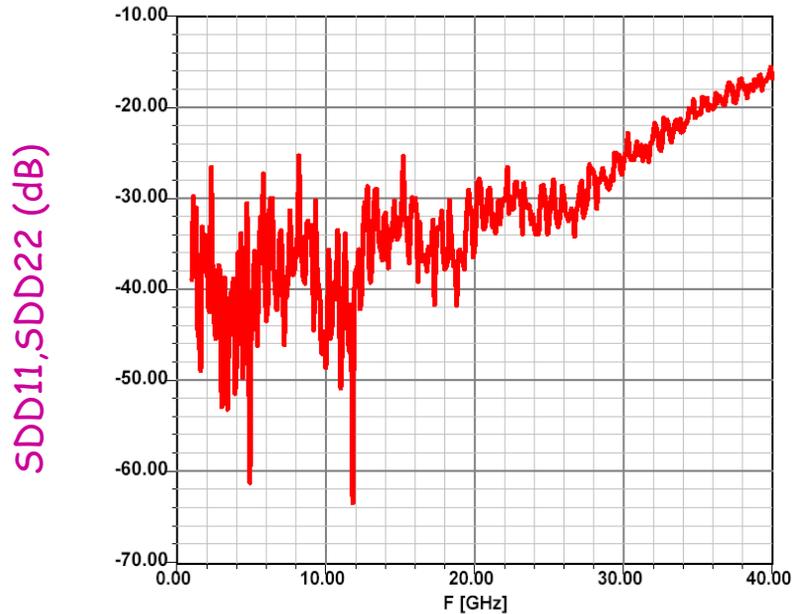
- ➢ **3 sections of microstrip (0.5")-stripline (10")-microstrip (0.5") transition, nominal 100-ohm differential structures**
  - – **Axis ratios of ground plane ellipse opening=2 for major radius=14mil, via offset = 9mil from line of symmetry of coupled line structure and transition length=20mil**
- ➢ **Near linear roll off - no significant notches or ripples in SDD21/SDD12**
- ➢ **TxLine gives trace loss alone is 35.91 dB at 40 GHz = (30"x1.1 dB/"+ 3"x0.97152 dB/")**

IC interconnect paradigm bifurcation: Optical interconnect insertion in intra-chassis communication

➤ **Packaging bifurcation**

– **Thin-fast electrical IO fewer by a factor 16 on low-loss package with vastly reduced tradeoff between interconnect loss, NEXT and routing density**

– **Wide-slow electrical 16× IO pads compared to Thin-fast and tradeoff between interconnect loss, NEXT, and routing density in package *and* backplane**

– **Wide-slow FTTP optical technology**

  • **0 m – 500 m distributed systems**

  • **Optical backplane**

  • **Optical isolation**

IC   Package   Electrical transmission line
PCB
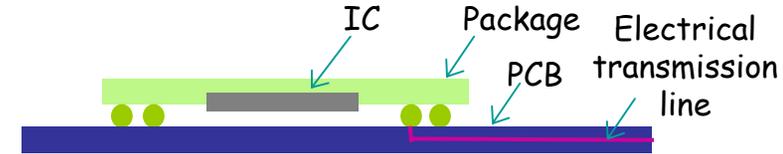
Thin-fast electrical (Intel) 40 Gb/s – 80 Gb/s per IO

Wide-slow electrical (IBM) 5 Gb/s per IO with 16 × IO pads compared to Thin-fast

VCSEL/PIN   IC   Package   Optical Waveguide
PCB
Optional lens
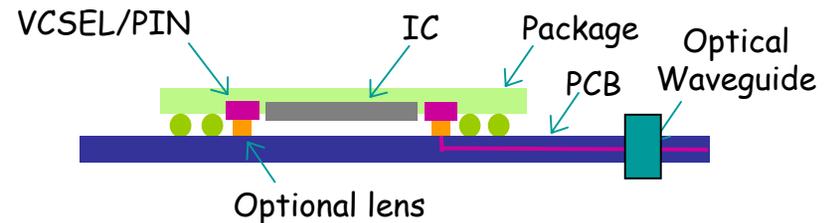
Wide-slow FTTP optical technology 10 Gb/s – 20 Gb/s per IO with 8× - 4× IO pads compared to Thin-fast
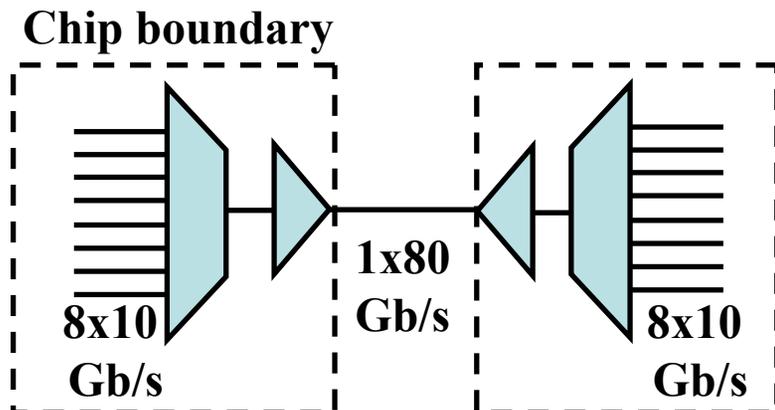
Time

Thin-fast optical technology compatible with 80 Gb/s PAM4 is yet to be determined

Incompatible technology paths:
Thin-fast electrical IO versus wide-slow optical IO

### Thin-fast electrical IO

Serial electrical IO at chip boundary

- **Electrical – need 20 dB+ equalization at 28 GHz for 80 Gb/s serial PAM-4**
- **Power: 400 mW estimate per 80 Gb/s serial link in 10 nm CMOS**
- **Challenge: PCB connector is the key enabler! Material loss must also be lowered to enable continued use of low-cost electrical links power-efficiently**

Wide-slow optical IO
Achievable practical data rate limited by laser modulation frequency / power / size
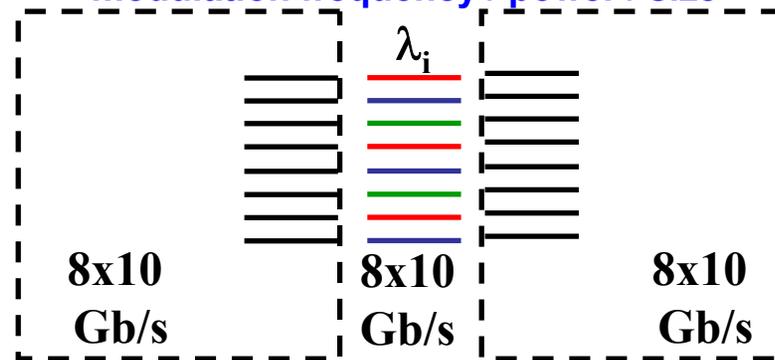
Parallel IO at chip boundary

- **Optics – need 8×10 Gb/s or 4×20 Gb/s parallel fiber-optics or WDM**
- **Power: > 320 mW (8×40 mW) per 8x10 Gb/s parallel link**
- **Challenge: Per-lane CDR must be avoided and traces from IC to Tx/Rx electronics of optical module must be ~ 1 mm to be competitive in power with electrical; need high yields, thermal regulation and low-cost test**
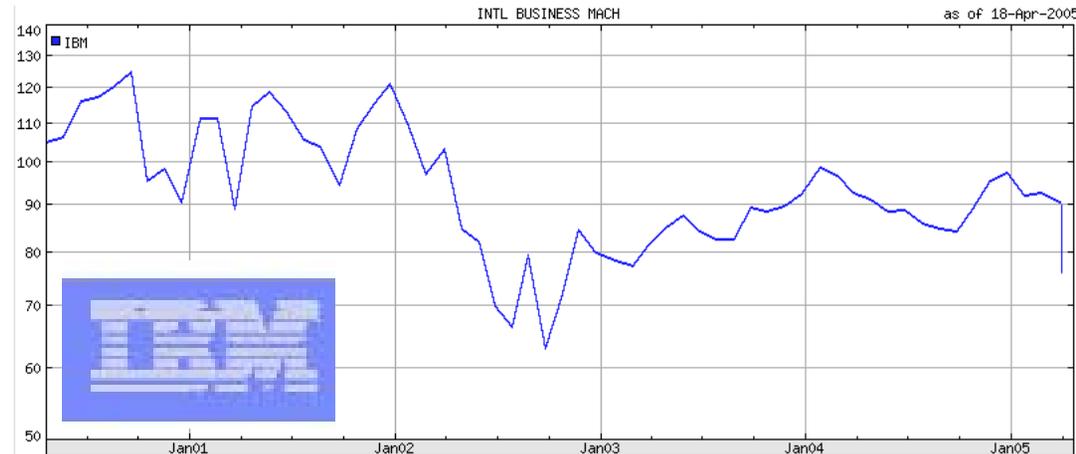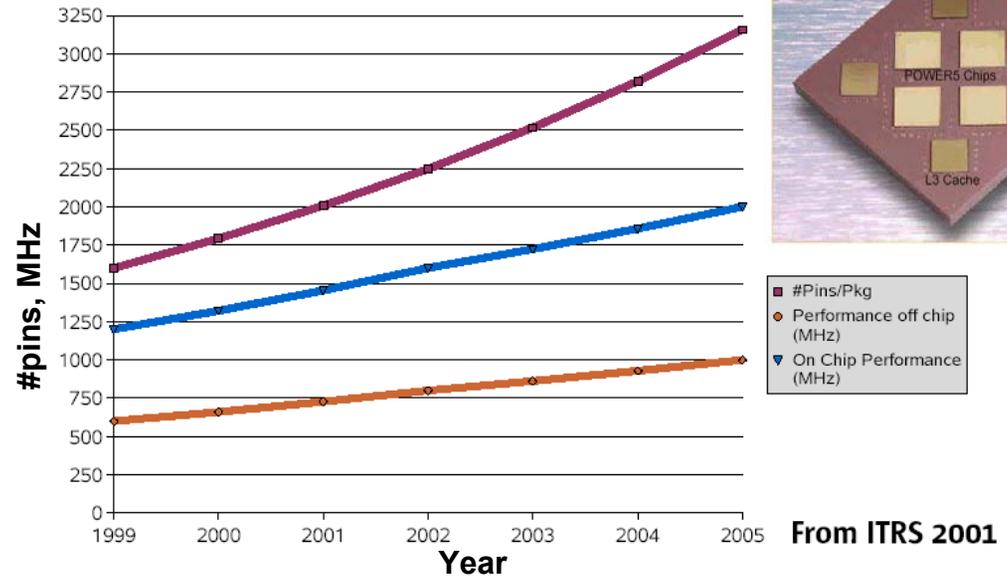
Slow-wide packaging solution

> **The IBM way**

- – **Large number of IO limited by size of pads and die**

- – **Increase packaging complexity, cost, system integration**

- – **Keep electrical interconnect by using relatively slow signaling rate**

> **… and IBM microelectronics failed to make money in past XQs**

Packaging roadmap

From ITRS 2001

Roadmaps

➢ **Following the directions of roadmaps only makes sense if you can make money on the journey**

    – **Big companies have a vested interest in following the *yellow brick road* especially if they can exclude direct competition from using the same road**

➢ ***However*, if the road turns into a dirt track**

    – **Off-road technology can win and dinosaurs following the dirt track will die**
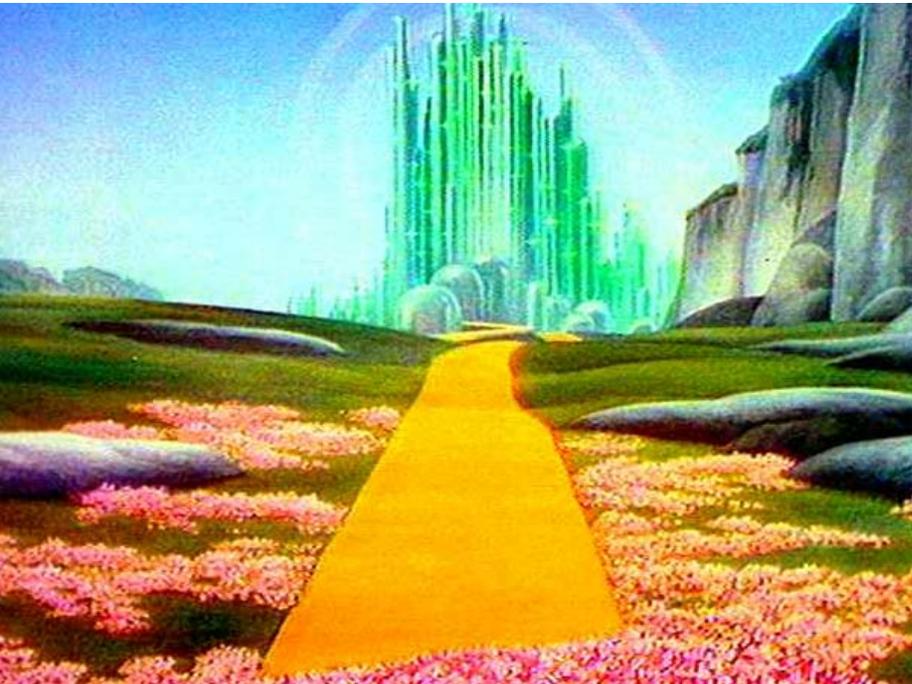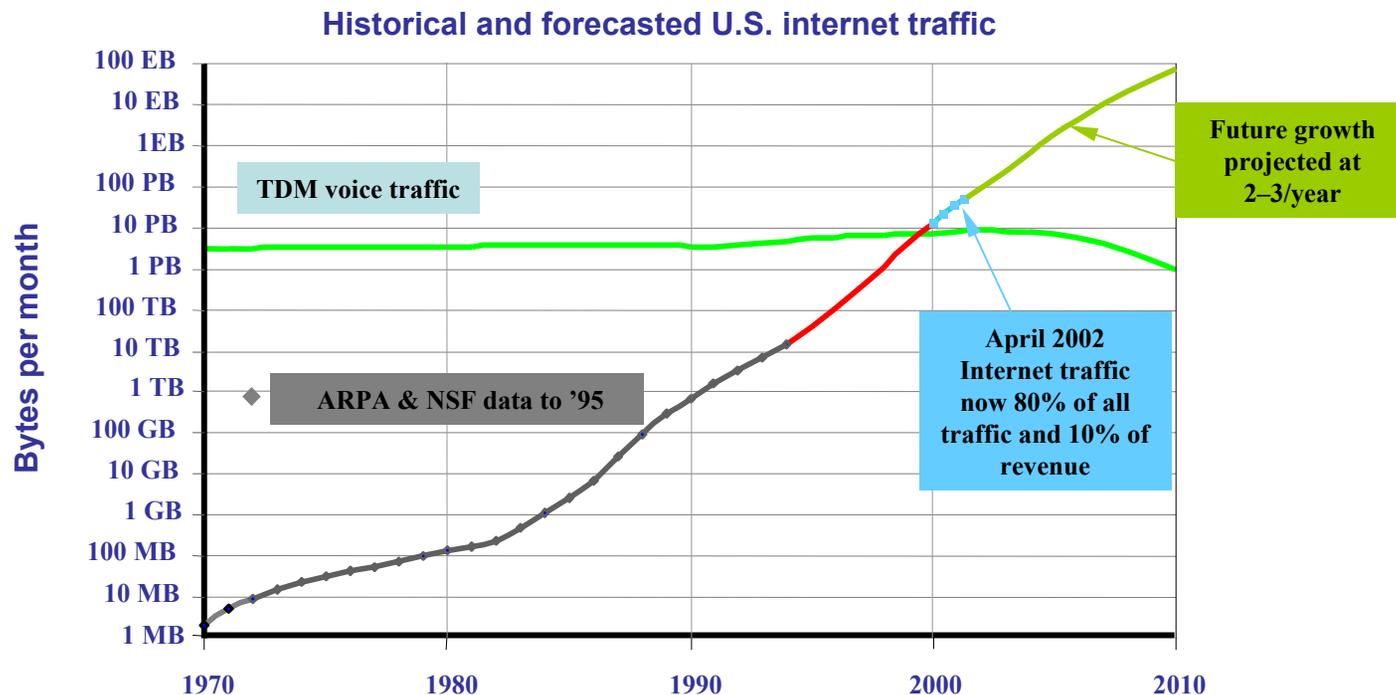

The yellow brick road to the emerald city

When the road turns to dirt, the dinosaurs die

Driving force: Opening the 'fat' photonic pipe for global application-on-demand

Historical and forecasted U.S. internet traffic

- TDM voice traffic
- ARPA & NSF data to '95
- April 2002 Internet traffic now 80% of all traffic and 10% of revenue
- Future growth projected at 2–3/year

Source: http://www.caspiannetworks.com/library/presentations/traffic/GEthernet.ppt
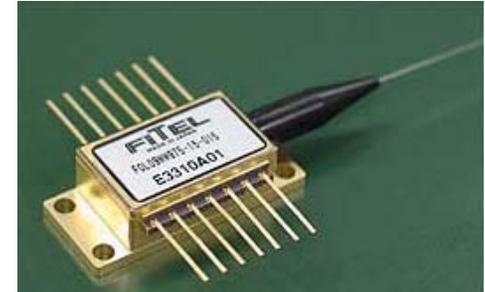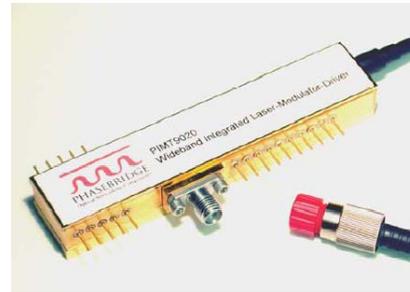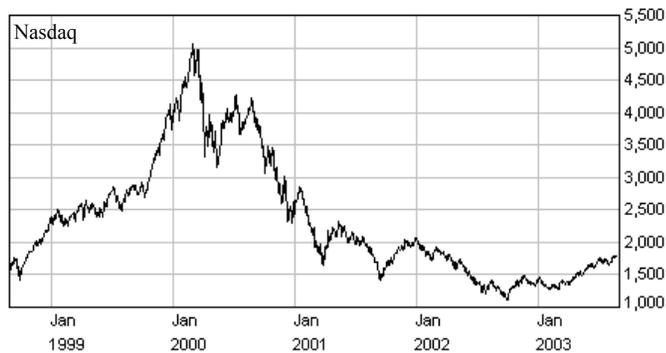
Who is going to provide the components, modules, and system integration?

Where are the new devices going to come from?

Volume manufacture and component integration:
The new path forward for fiber-optic system development

Volume production

Fiber-optic components and modules

➢ **Since the Telco meltdown technology base has moved from US to pacific rim (China) to remove labor cost from products.**

➢ **Even with zero labor cost, components are _still_ too expensive!**

➢ **_Need_**

➢ **_New_ high-volume markets (metro-FTTH, FTTP, automotive, …)**

➢ **_New_ cutting-edge technologies must be characterized by:**

➢ **Ultra-low cost (small, light-weight, low-power, few sub-component parts, approach cost-of-materials)**

➢ **High added value (e.g. integration of multiple functions)**

➢ **High level of volume manufacturability (10M/month, true 6σ)**

➢ **A new platform based on**

➢ **Ultra-precise metal coining with nm tolerance**

➢ **Advanced photonic devices**

➢ **High levels of integration with CMOS electronics**

Volume production with nano-scale precision Example: The fiber-optic connector!

➢ **Fiber connector average selling price is too high (e.g. $4 per installed plug in 2006, 500M units)**

  ➢**Tolerance scale set by wavelength of light λ_0 = 1550 nm and mode diameter in fiber**

  ➢**SMF-28e lateral displacement induced loss (dB) = 4.343 $(d/r)^2$, d = lateral off-set, r = mode field radius**

  ➢**± 300 nm typical finish tolerance on 2.5 mm diameter ferrule (l / $\triangle l$ = 8,333)**

  ➢**Volume production (>10M/month, >250/min) best if true 6σ or < 2 PPB failure rate, c.f. Motorola 'six sigma process' ≡ 4.5σ or 3.4 PPM failure rate**

  ➢**Assuming normal distribution, true 6σ requires better than σ = 50 nm tolerance**

➢ **New volume production nano-technology!**

  ➢**Production cost must approach cost-of-materials**

  ➢**Ultra low-cost, high-volume, precision fiber-optic manufacturing enables revolutionary wide-scale adoption of optics in systems**

Normal distribution: single-sided probability of error
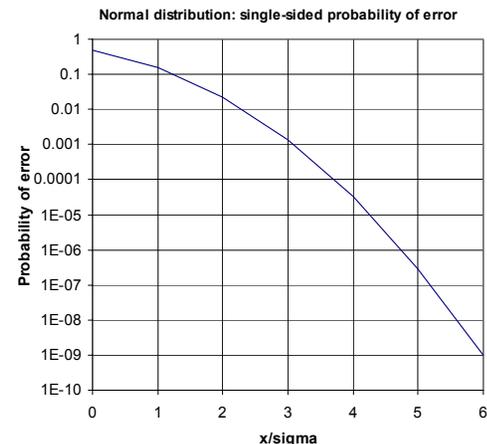
Stamping process is path to cost-of-materials manufacturing

Precision stamping of SMF MT-RJ:
- ➢ **Closed die process**
- ➢ **Small clearances between punch and punch holder**
- ➢ **The linear gauge reader is attached to punch and hydraulic pressure monitored for future active tooling**
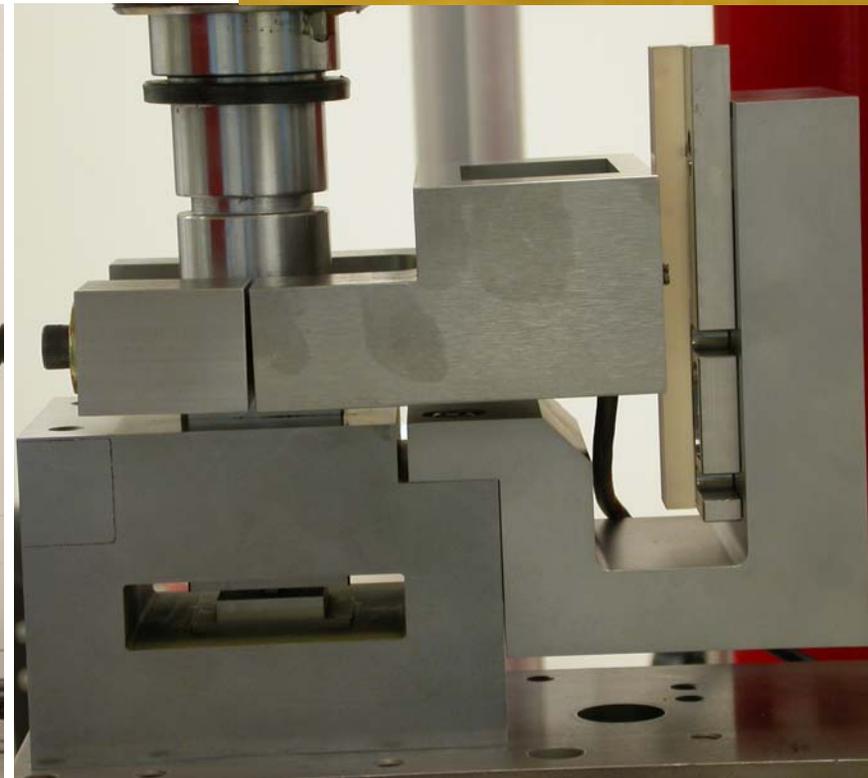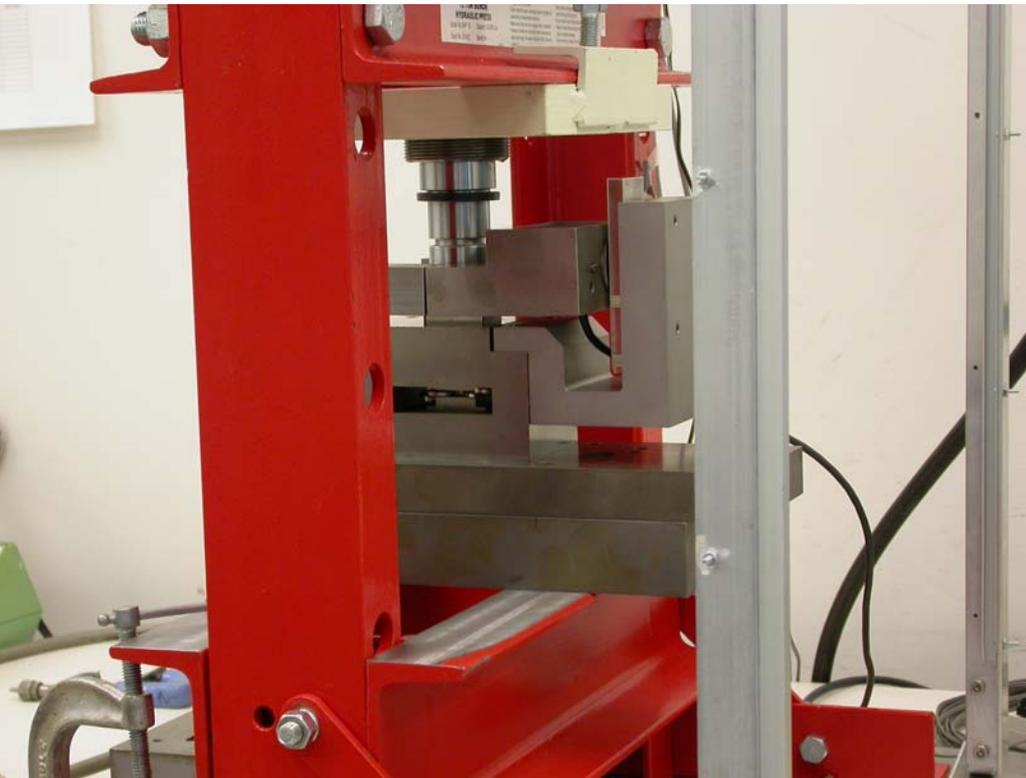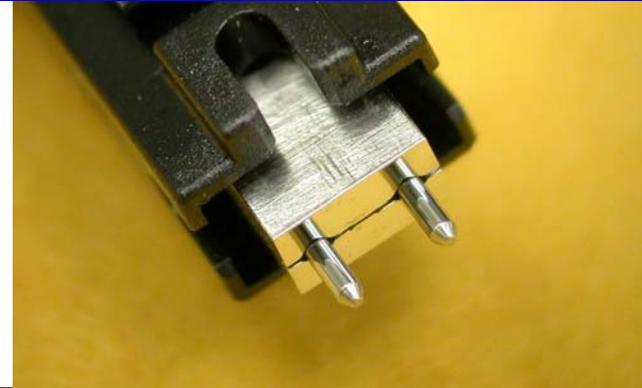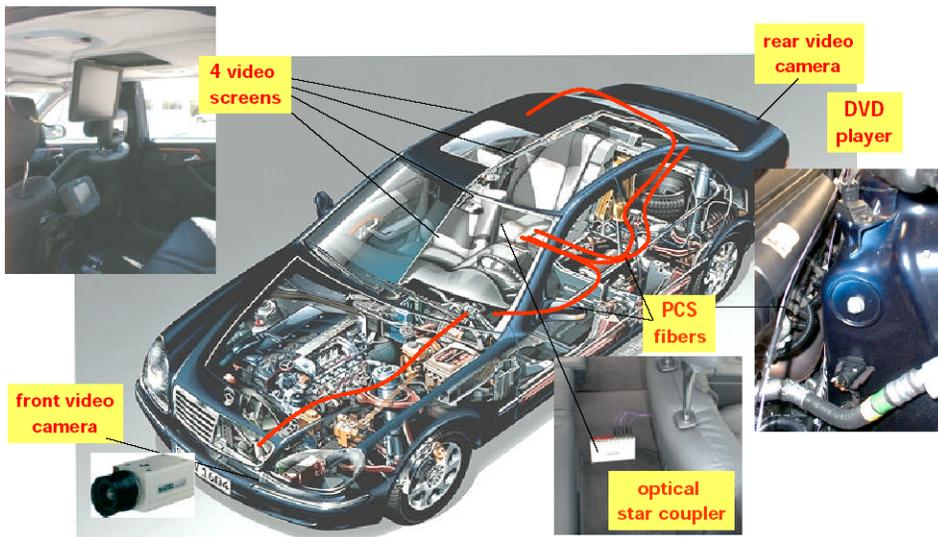
New volume markets for optical interconnects: The automobile

- Mercedes-Benz S-class model year 2005 has a fiber-optic data bus backbone operating at Gb/s rates and for the first time using VCSELs (E-class and other models already use LED based systems at ~5 Mb/s)

- Data carried includes several video channels, the entertainment channels, and all sensor data / telemetry

- Fiber beats copper!
  - 30M fiber links in 2005, over 120M fiber links in 2010

DAIMLERCHRYSLER — Research and Technology

Car Demonstrator: PCS fiber star network @ 622 MBit/s

4 video screens | rear video camera | DVD player | PCS fibers | front video camera | optical star coupler

DAIMLERCHRYSLER — Research and Technology

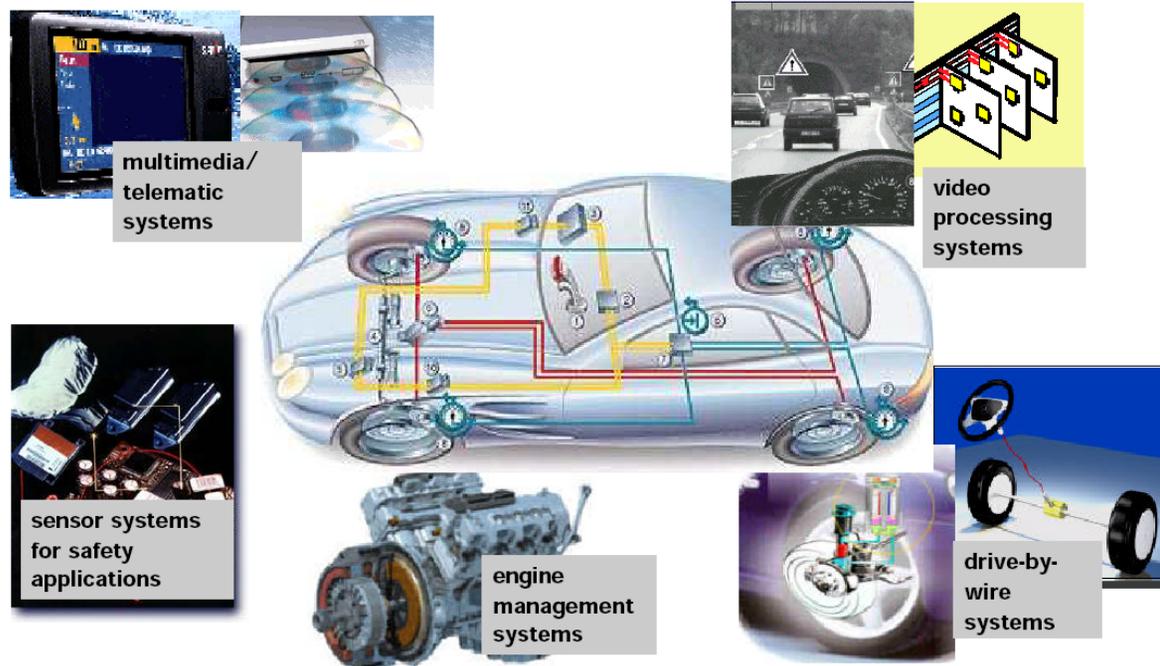Car Demonstrator: PCS fiber star network @ 622 MBit/s

Future needs for optical interconnects in multi-processor automobile systems

➢ **12-fiber ribbon and multi-Gb/s/fiber**

➢ **Ultra-high reliability for real-time processing of drive-by-fiber data in multi-processor embedded system environment**

    ➢ **MOST protocol, physical layer standards**

    ➢ **Aircraft as secondary market!**

DAIMLERCHRYSLER   Research and Technology

Future needs for robust high speed data networks

multimedia/ telematic systems

video processing systems

sensor systems for safety applications

engine management systems

drive-by-wire systems

Summary

➢ **Development of *perfect* electrical connector would be significant technical barrier to optics penetrating ≤ 0.5 m interconnect length in systems**

  – **Electronic interconnect distance is collapsing to ≤ 0.5 m**

    • **1RU electrical bisection bandwidth limited to ≤ 18.7 Tb/s**

➢ **Challenge for optics is to be competitive with electronic solutions**

  – **Opportunity to implement new architectures such as FTTP (8 Tb/s/SoC) that require optical interconnect inside the box**

    • **New optical devices**

      – **Optical – electrical socket for FTTP**
      – **Optical – electrical PCB, optical backplane connectors**
      – **New PAM-4 compatible optical components that directly interface to 80 Gb/s data bandwidth PAM-4 electrical signaling *or* > 40 Gb/s VCSELs, I_{th} < 0.5 mA at 100 ^{o}C, I_d < 2 mA, η > 0.5**
      – **Cost-of-materials manufacturing**

    • **Complete optical solution for system designer**

      – **Standards for socket, PCB, connectors, testing**
      – **One-stop shopping**
      – **Multi-sourcing of components**
      – **Design tools that are transparent to system designer**

➢ **Adapt and innovate or die !**