# Future of opto-electronics in CMOS-based systems

**A. F. J. Levi**

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

**http://www.usc.edu/dept/engineering/eleceng/Adv_Network_Tech/**

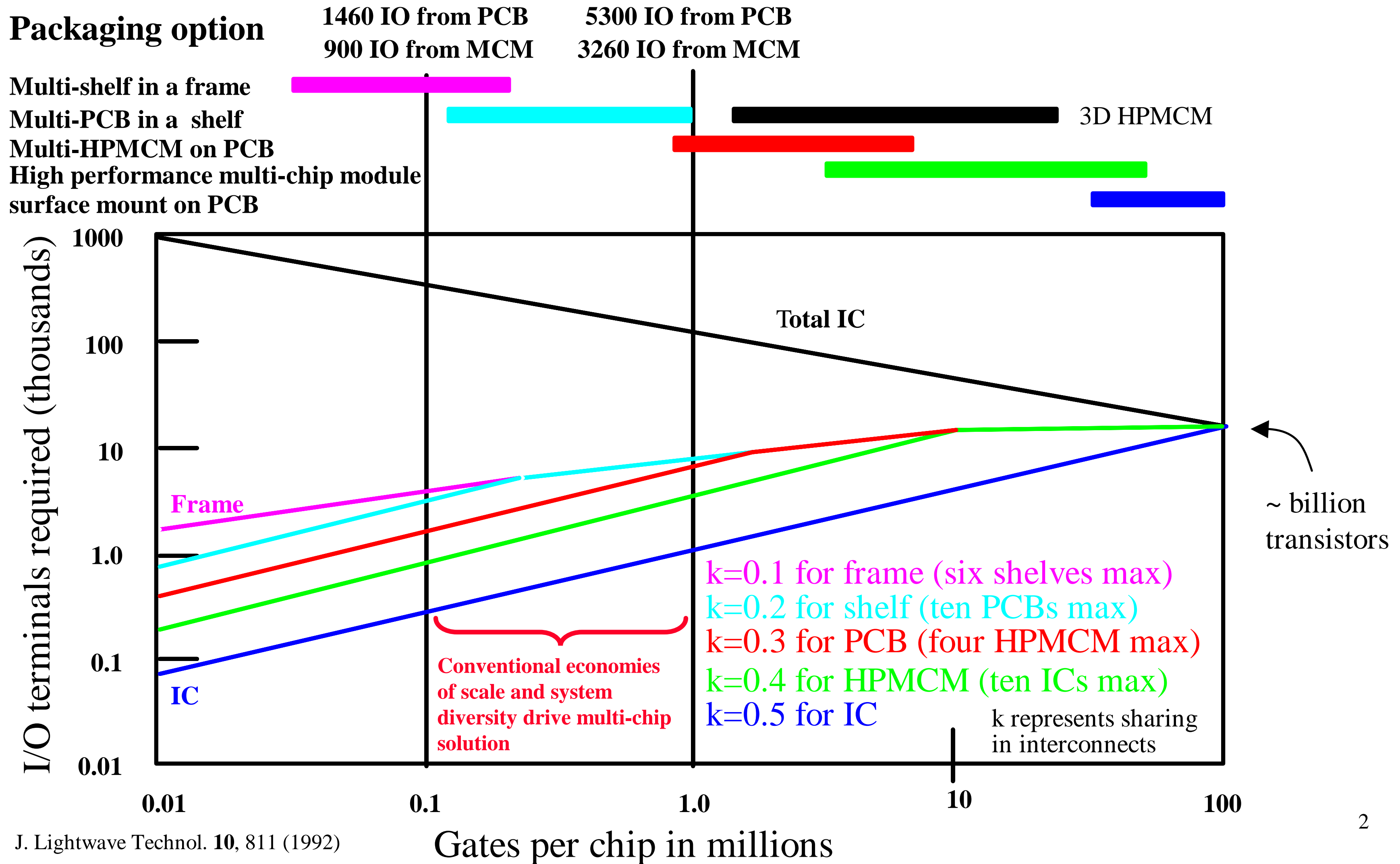
| | |
|---|---|
| **voice** | **(213) 740 -7318** |
| **email** | **alevi@usc.edu** |

**Presented at the Optoelectronics Industry Development Association workshop on "Opto-electronics on CMOS", May 12, 1999, Hilton Santa Fe, Santa Fe, New Mexico.**
**OIDA voice (202)-785-4426, fax (202)-785-4428**

USC
UNIVERSITY
OF SOUTHERN
CALIFORNIA

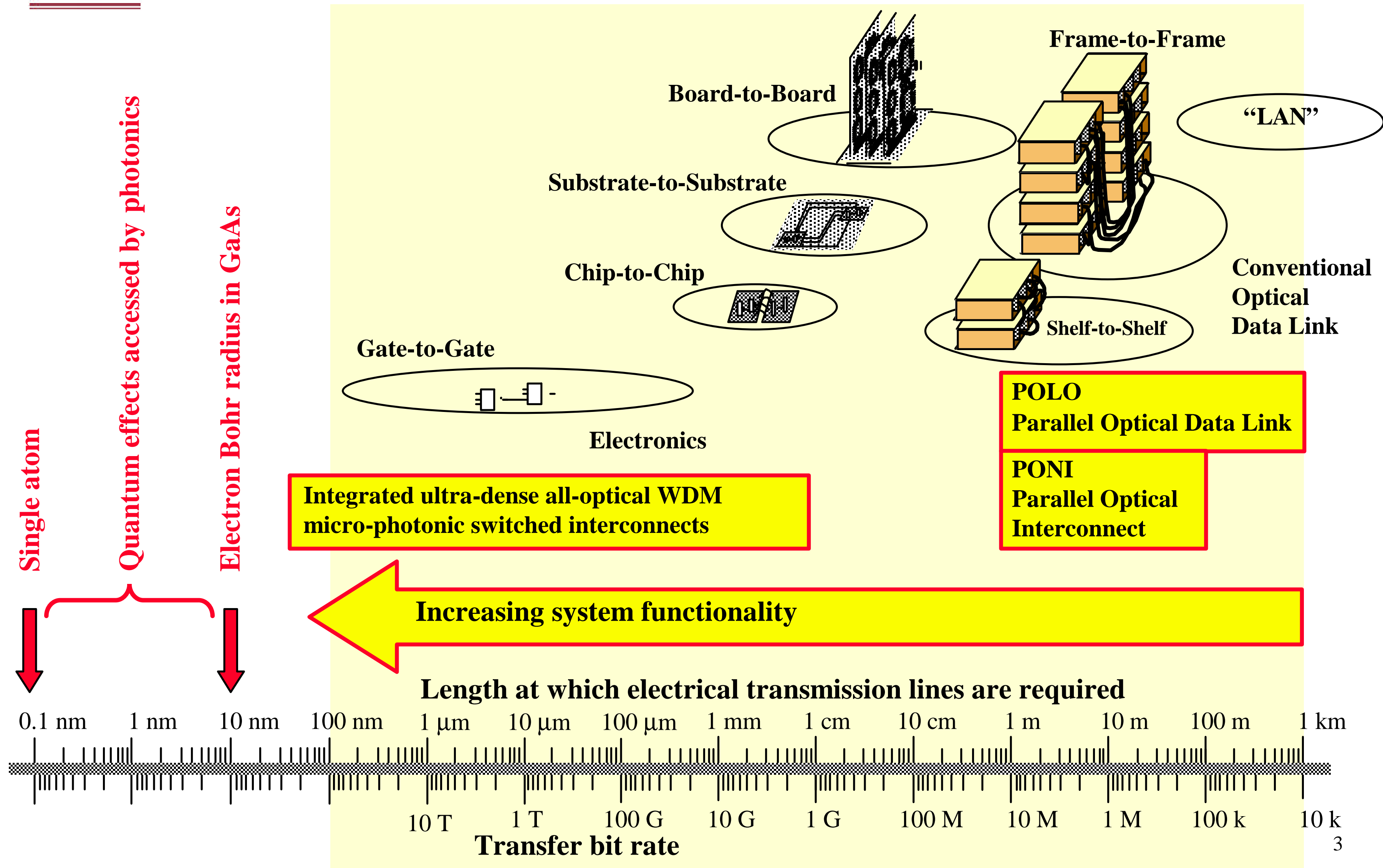
# IO interconnection and packaging in a large system

Large system IO as function of gates per chip follows Rents rule $(IO/k)^{1.8}$ = number of gates



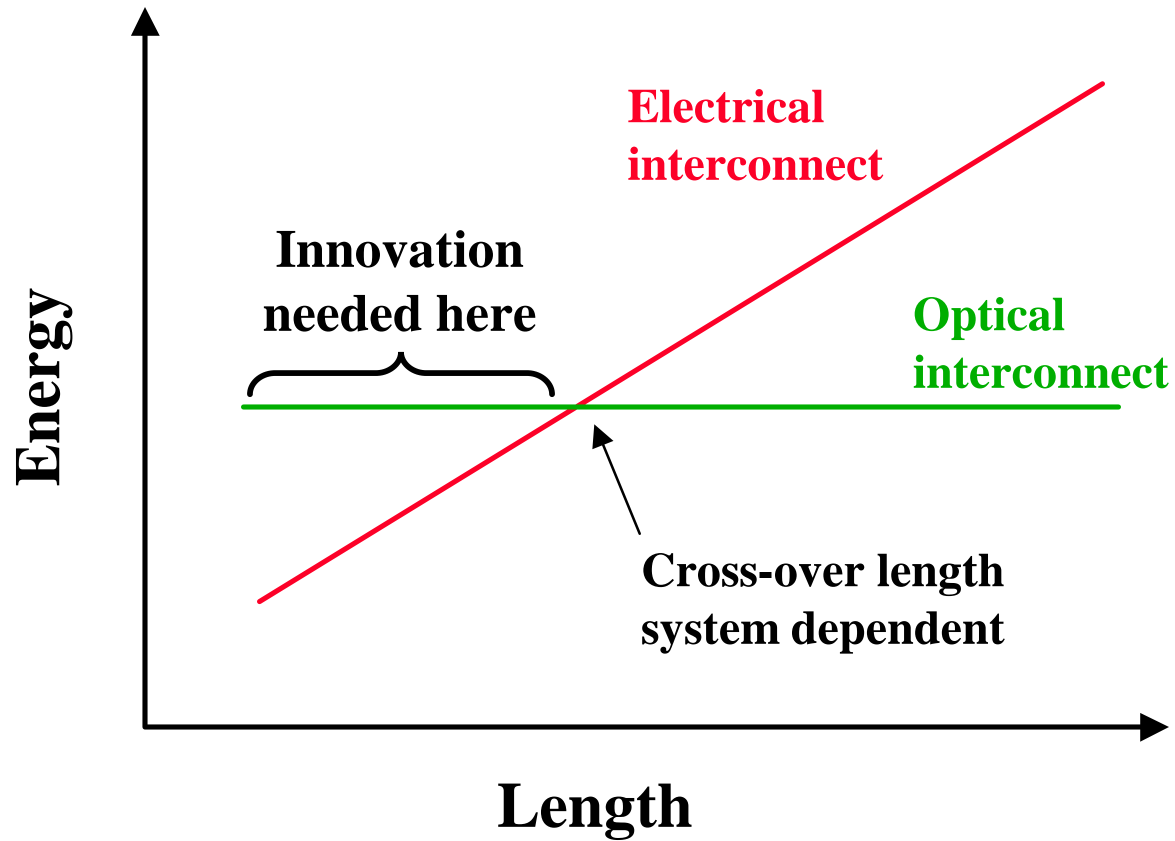

J. Lightwave Technol. **10**, 811 (1992)

# System interconnect hierarchy and advanced optical solutions



USC UNIVERSITY OF SOUTHERN CALIFORNIA

# Communication energy



# Synchronous interconnection price comparison

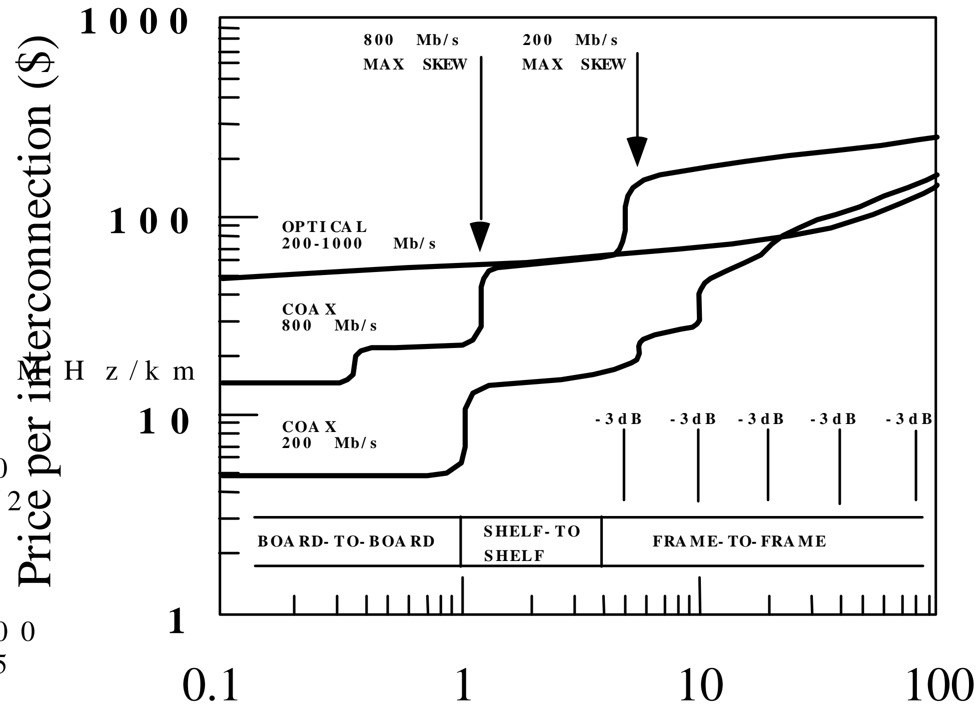
**Assumptions-price per line:**

**optical (200/1000 Mb/s)**
pin detector $7
Driver and receiver $2
Submount $4
Hybrid package $2
Fiber (per meter) $1
Array connector $11
Laser array $25

Fiber skew 3σ = 9 ps/m
Fiber loss = 1 dB/m
Fiber modal dispersion 500 MHz/km

**Electrical (200/800 Mb/s)**
Driver and receiver $2/10
4 Backplane pins $1.2/2
PCB $2
Coax per meter $1
Coax per connection $1/3
Mux/demux $1/100
Clock recovery $5/35

Cable skew 3σ = 300 ps/m
Cable loss = 0.3 dB / 0.6 dB
At every 3 dB point demultiplex
Clock recovery at skew line
Max. skew 1600 ps / 400 ps




Interconnection length (m)

**Innovation needed here**

USC UNIVERSITY OF SOUTHERN CALIFORNIA

# Example: SGI Origin switch-based system architecture

Electrical interconnect:

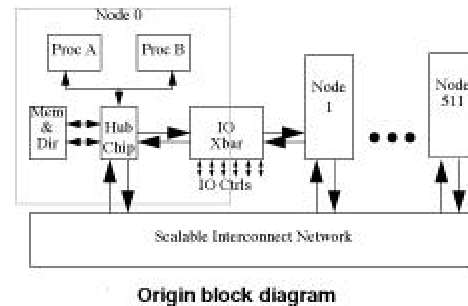
- 44 signal pins per direction
- up to 5 m electrical cable

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

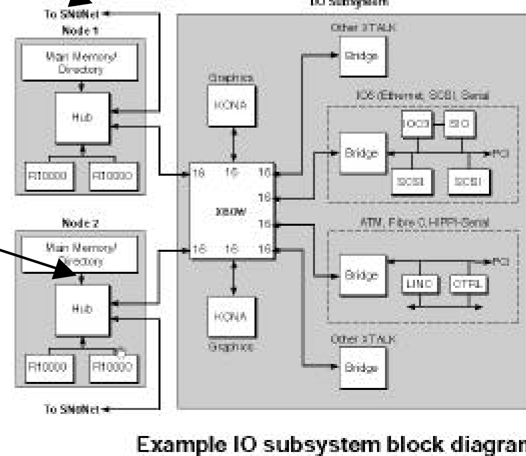
**Memory access**
0.78 GByte/s peak total
0.78 GByte/s sustained total

**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

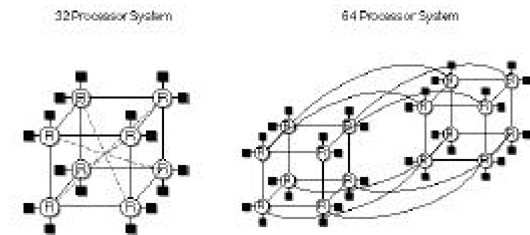
Node-to-node 16 kB page block transfer < 30 µs



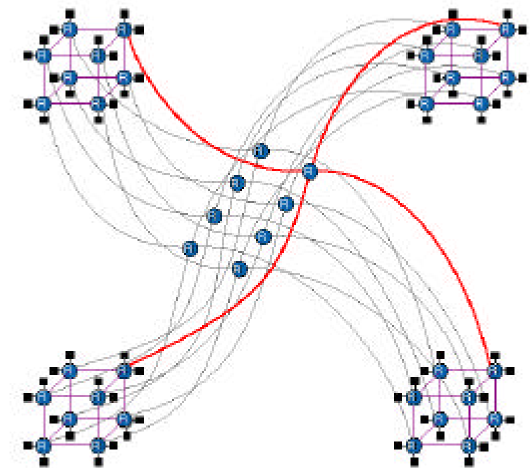

Origin block diagram




Example IO subsystem block diagram

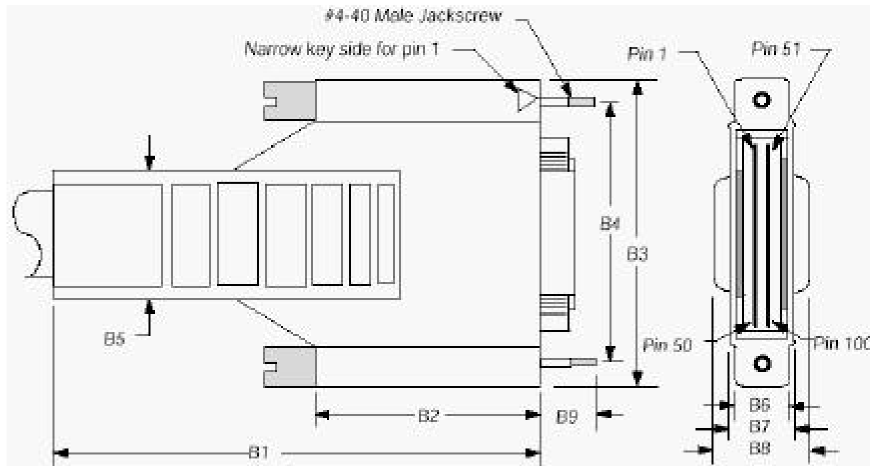



32P and 64P Bristled Hypercubes



128P Heirarchical Fat Bristled Hypercube

# HIPPI-6400 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 GByte/s/inch)
- Limited bandwidth (< 1 Gb/s)
- Limited distance (< 50 m)

HIPPI-6400 electrical connector
- 2.31”x0.75” edge dimension
- 2 row, 100 pin connector (23x2x2=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 GByte/s peak bandwidth per direction, 1.6 GByte/s bisection bandwidth (bisection bandwidth density 0.7 GByte/s/inch or 2.2 Gb/s/cm)

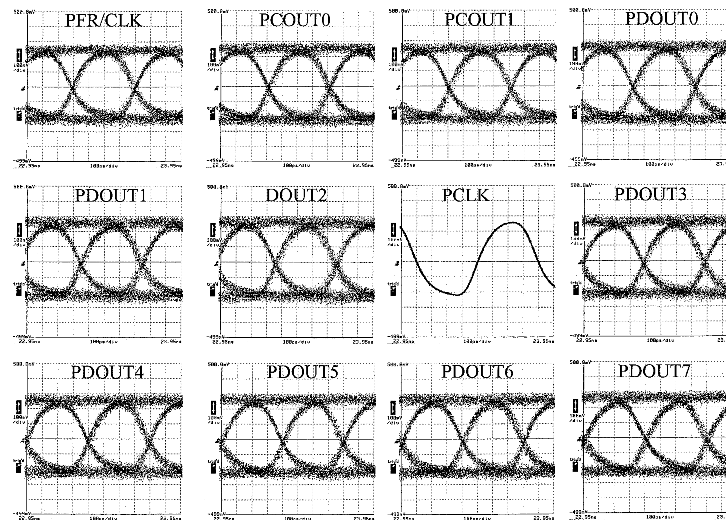
# High-performance opto-electronic interface to CMOS

## Potential power savings using all-optical signal processing

**USC PONI MUX IC (v1) 0.5 mm CMOS**
**10 x 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 x 2.5 Gb/s + CLK**
**1:2 / 2:1 MUX**

**55 Gb/s bi-section data bandwidth per cm**

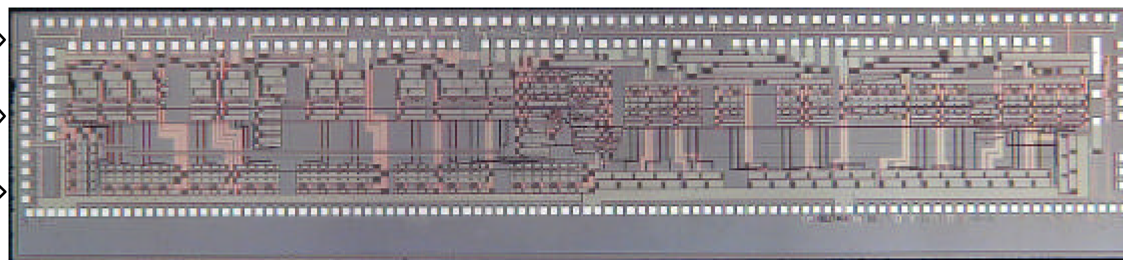
**2.7 ns Tx/Rx mux/demux end-to-end latency**

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

*Power consumption using 0.5 mm 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

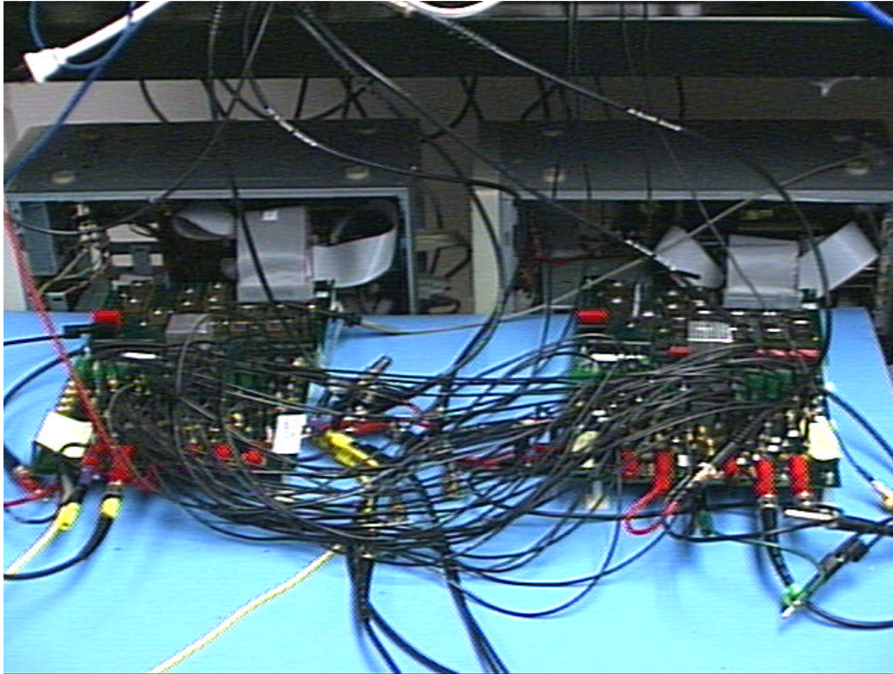


"A 55 Gb/s/cm data bandwidth density interface in 0.5 μm CMOS ...",
B. Madhavan and A. F. J. Levi, Electron. Lett. **34**, 1846-1847 (1998)

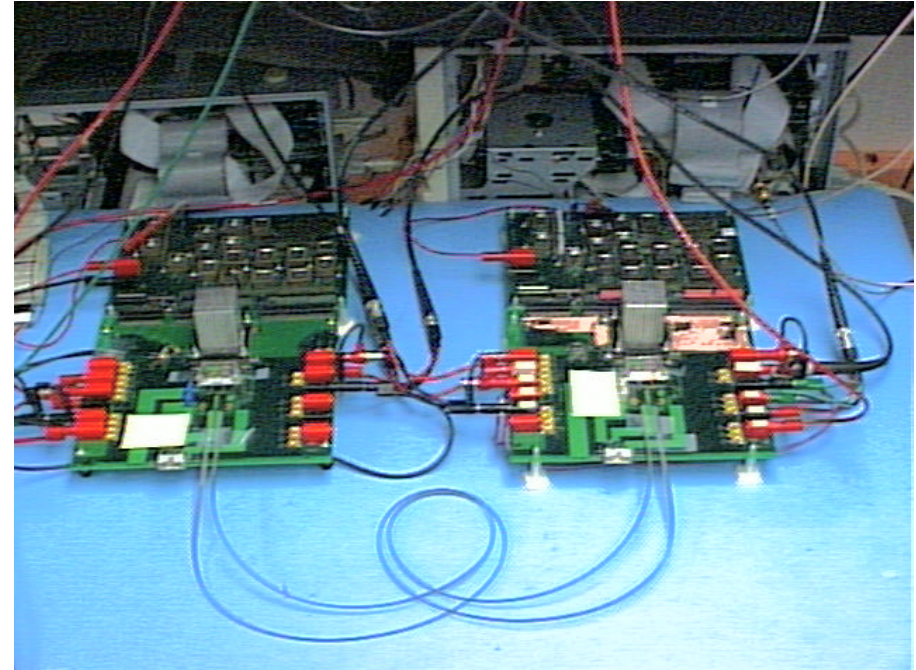
USC UNIVERSITY OF SOUTHERN CALIFORNIA

# Fiber form factor advantage for GByte/s interconnect

## USC electrical and optical system test between two Pentium hosts



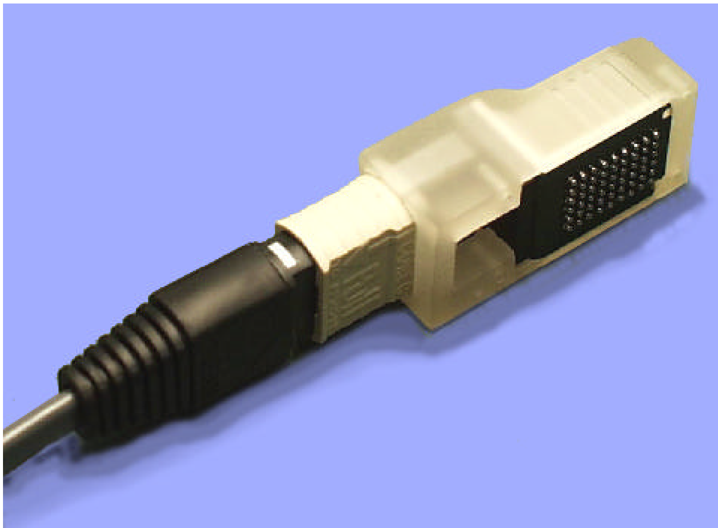
Electrical test fixture for LA chip requires 40 coaxial cables



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

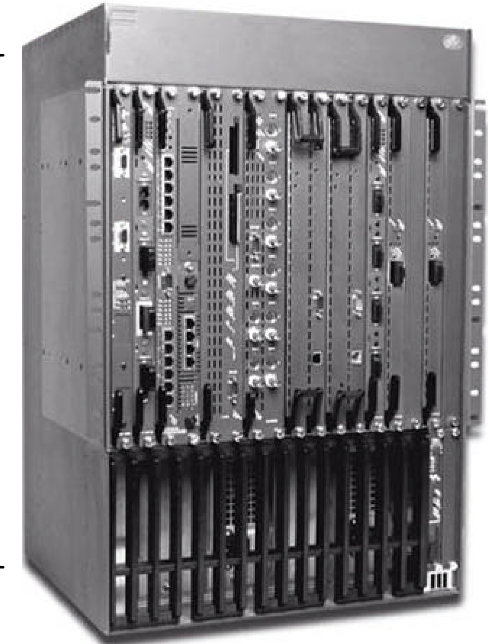
# System impact of > 50 Gb/s per linear cm

- Multi-GByte/s data rate per linear cm



HP PONI Tx-module

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



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

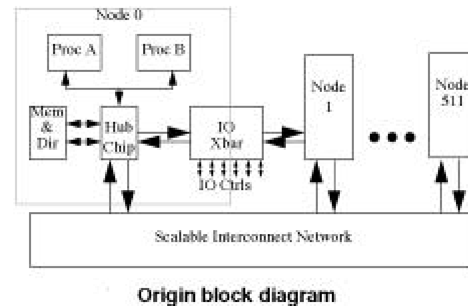
- **USC - HP PONI DARPA program can provide > 50 Gb/s per cm using**
  - **USC interface IC**
  - **12-wide fiber ribbon, 2.5 Gb/s per fiber signaling, and MTP connector**
  - **BGA surface mount to PCB**

USC UNIVERSITY OF SOUTHERN CALIFORNIA

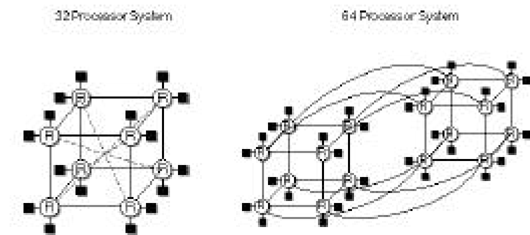
# Example: SGI Origin switch-based system architecture

Electrical interconnect:

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




Origin block diagram

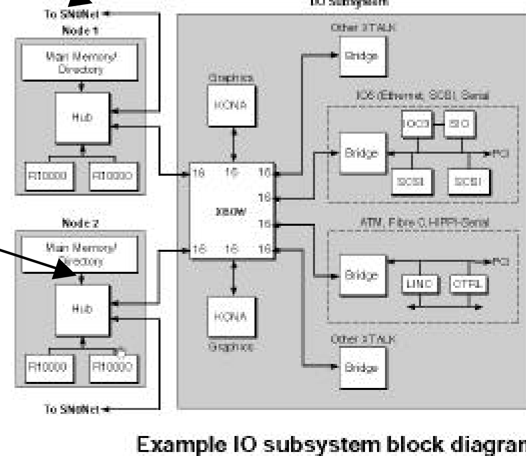



32P and 64P Bristled Hypercubes

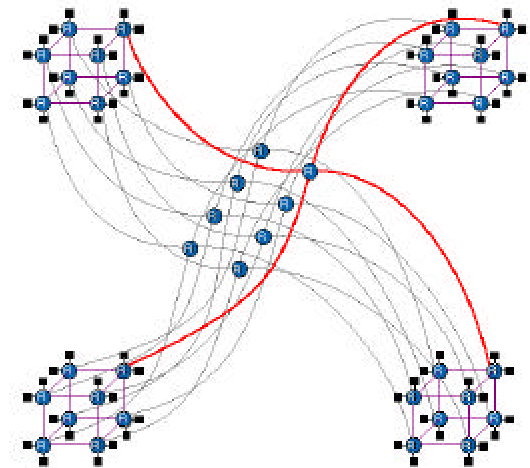
**Node-to-node access**
0.73 GByte/s peak per direction
0.625 GByte/s sustained per direction

**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




Example IO subsystem block diagram

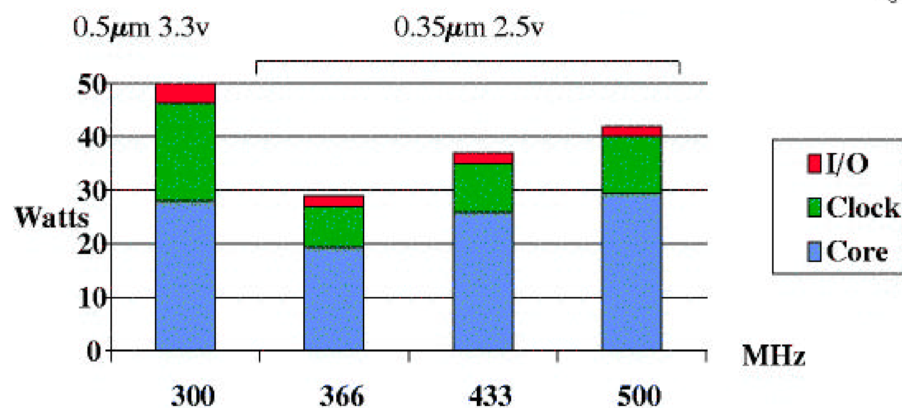


128P Heirarchical Fat Bristled Hypercube

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

# Alpha high-performance, high-power microprocessor design

- **Alpha 21264, 9.6M transistors, 600 MHz, 2.4 BIPS, 64b, 45 W, internal $V_{DD}$ = 2.4 V, 14.4x14.5 = 208 $mm^2$ four-metal 0.35 mm CMOS, 499 PGA package.**
- **11 W (25%) power for clock distribution**
- **8 kB Icache, 8 kB Dcache**
- **96 kB L2 cache**




Source: http://infopad.EECS.Berkeley.EDU/HotChips8/1.2/1.2.10.html

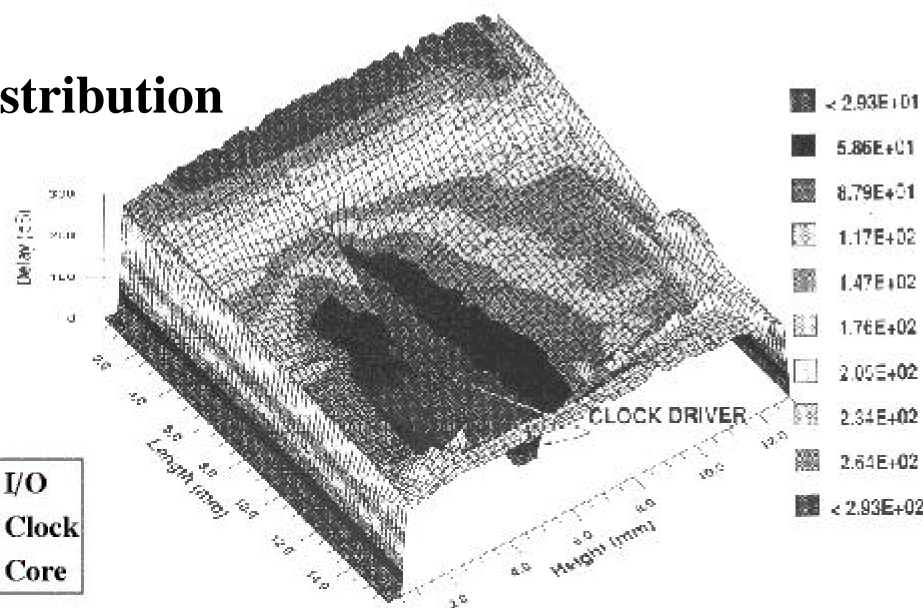
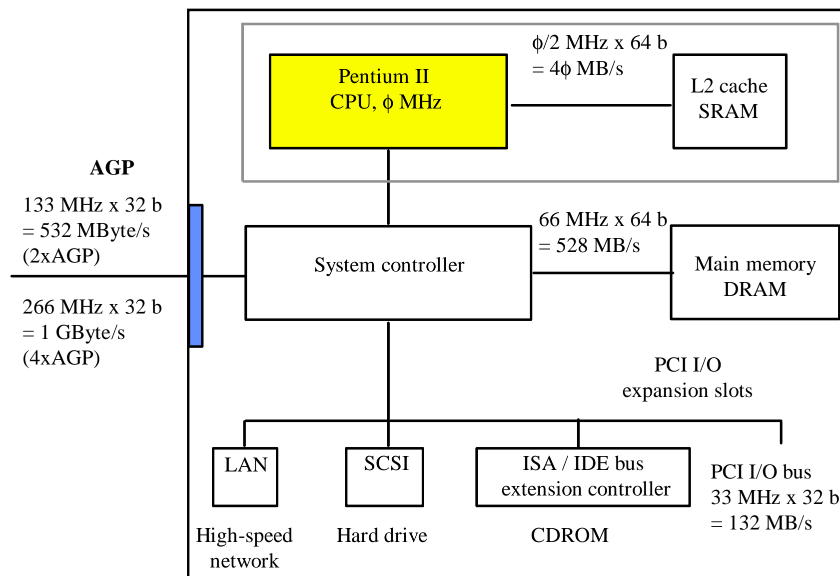



*Figure 7 CPU Clock Skew*

Alpha CPU clock skew in 1.7M transistor, 30 W, 200 MHz, 400 MIPS, 64b. $V_{DD}$ = 3.3 V, 16.8x13.9 $mm^2$ three-metal 0.75 μm CMOS, 431 pin PGA package.
12 W for clock distribution.
Source: Digital Technical Journal **4**, 1 (1992)

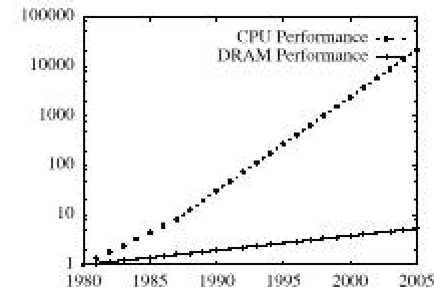
# Microprocessor - DRAM performance gap

- **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 - Jan. Intel Pentium MMX (150 - 233 MHz)
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)
1999 - Intel Merced

| *Processor* | *Clock (MHz)* | *SPECint95* | *SPECfp95* |
|---|---|---|---|
| Pentium II | 266 | 10.8 | 6.9 |
| DEC Alpha | 266 | 7.9 | 11.8 |
| Pentium II | 300 | 11.6 | 7.2 |
| DEC Alpha | 333 | 9.8 | 12.5 |
| Pentium II | 450 | 18 | 13.3 |
| Pentium III | 450 | 18.7 | 13.7 |
| DEC Alpha | 500 | 15.0 | 20.4 |
| Pentium III | 500 | 20.6 | 14.7 |




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

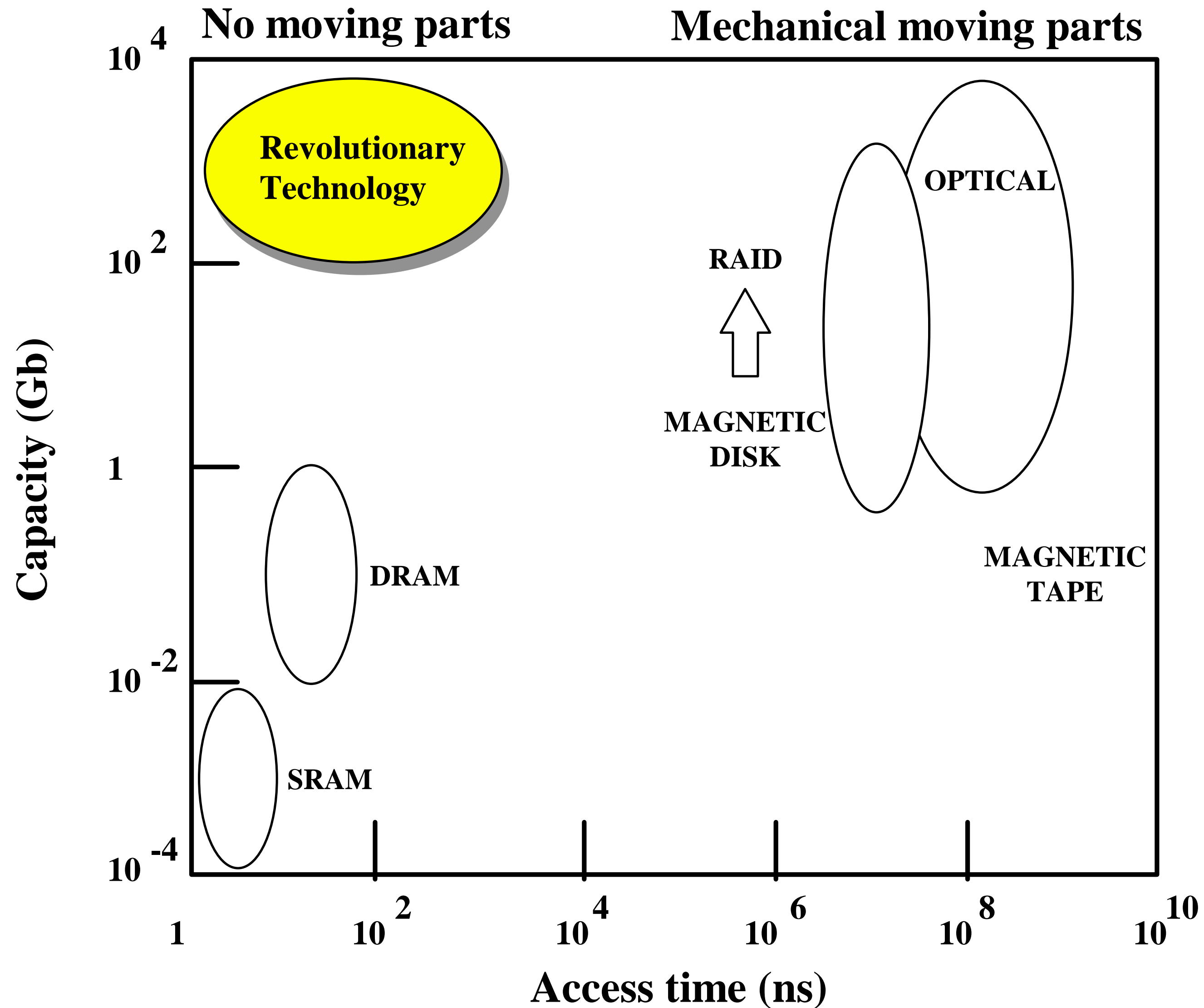
## StrongARM low-power microprocessor design

- 2.5M transistors, 160 MHz, 2.1 MIPS, 32b, 0.5 W, internal $V_{DD}$ = 1.6 V, 7.8x6.4 = 50 $mm^2$ three-metal 0.35 mm CMOS, 144 pin QFP package
- To minimize pin power and support a high-speed internal core, 50% of chip area is devoted to two 16 kB Dcache and Icache
- 90% of the transistors are devoted to Dcache and Icache
- The pad ring occupies 33% of chip area and the processor core fills the remaining 17% of chip area.

Source: IEEE J. Solid-state circuits **31**, 1703 (1996)

| **Power dissipation** | |
|---|---|
| Icache | 27% |
| Ibox | 18% |
| Dcashe | 16% |
| Clock | 10% |
| IMMU | 9% |
| Ebox | 8% |
| DMMU | 8% |
| Write buffer | 2% |
| Bus interface | 2% |
| PLL | <1% |

# Memory access-time bottleneck



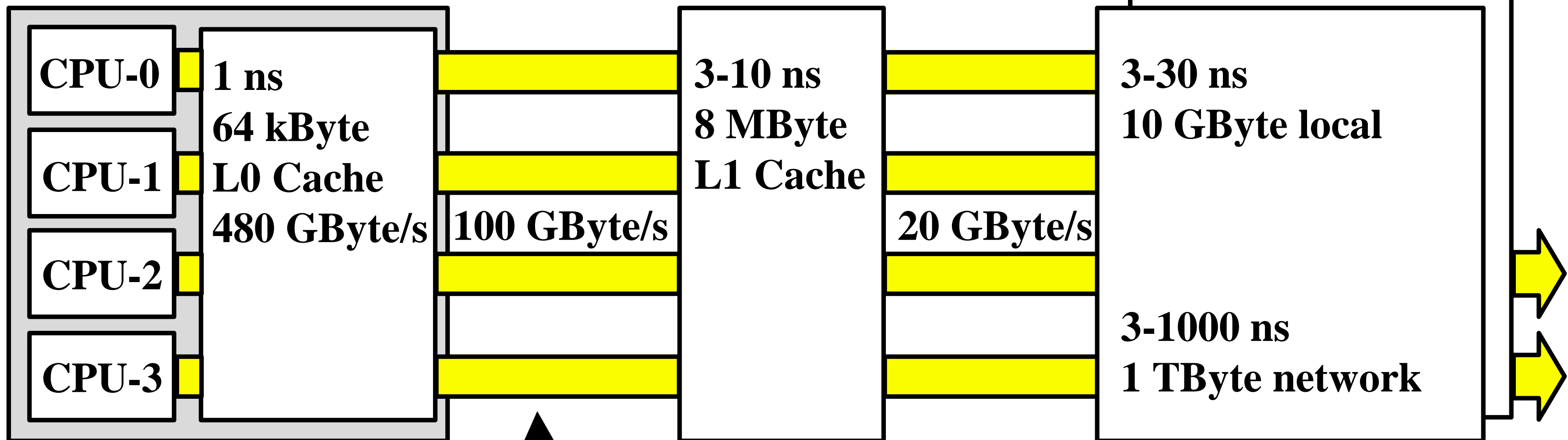
USC UNIVERSITY OF SOUTHERN CALIFORNIA

# Future high-performance microprocessor architecture

**Single chip VLIW or MP with integrated L0 Cache**
**80% scalar hit rate**

**L1 Cache**
**80% scalar hit rate**

**Main Memory**




**64 bit words**
**f = 10 GHz**
**Total 8, 32-bit instructions / f**
**Total 2, 64-bit data / f**
**Power = 100 W**

***Scalar*: 800 Gb/s communication channel**
**< 4 W total micro-photonic power**
**162 W electrical rambus power consumption**
***Vector*: 4.8 Tb/s communication channel**
**< 19 W total micro-photonic power**
**7.8 kW electrical rambus power consumption**

# The Rambus approach

## Physical organization of a Rambus-based system

### Rambus technology

- Vdd = 3.3 V, 50 ns access time
- f = 500 - 600 MHz Byte-wide bus
- 9-chip 72 MB module 3.5 x 3.5 x 1.5 $cm^3$
- Power per 8 MByte chip:

  Evaluate request = 0.55 W
  Read from device = 0.76 W
  Write from device = 0.86 W

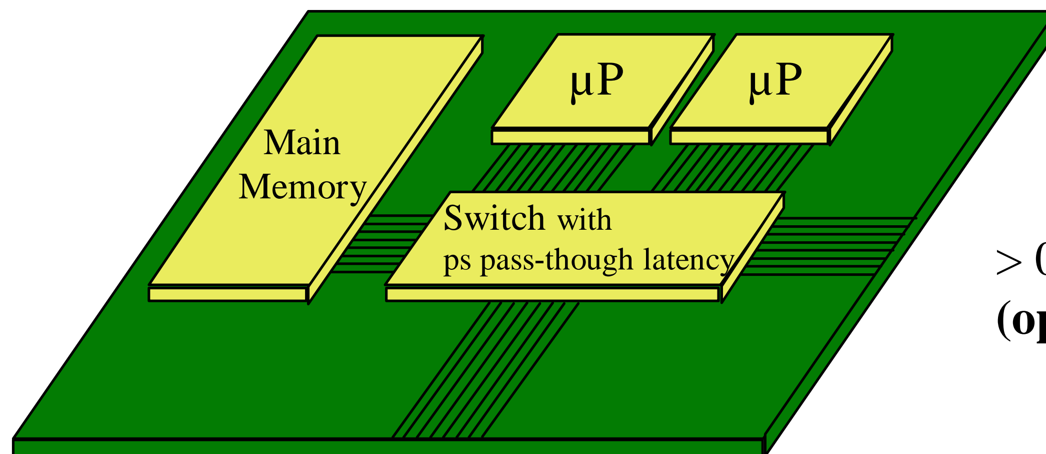
Memory Modules
Rambus DRAMs
Extent of Secondary Channels
Rambus DRAMs
ASIC
Extent of Rambus Primary Channel Along PC Board
Socket for Expansion
ASIC
CPU

⇨ Power consumption dominated by data transfer
10 GByte/s Rambus data transfer: read 15.2 W
write 17.2 W

# Micro-photonic technology in systems

**Reduce high-speed communication power (x40 less than Rambus)**
**Lower electrical noise in system (x10 less dI/dt)**
**> 0.1 m backplane interconnects can maintain bisection bandwidth (multiple Tb/s)**
**WDM *all-optical* functionality can give ps node routing latency (useful for < 1 m), deadlock protection, and collisionless adaptive routing**

**Processor node 10x10 cm$^2$**

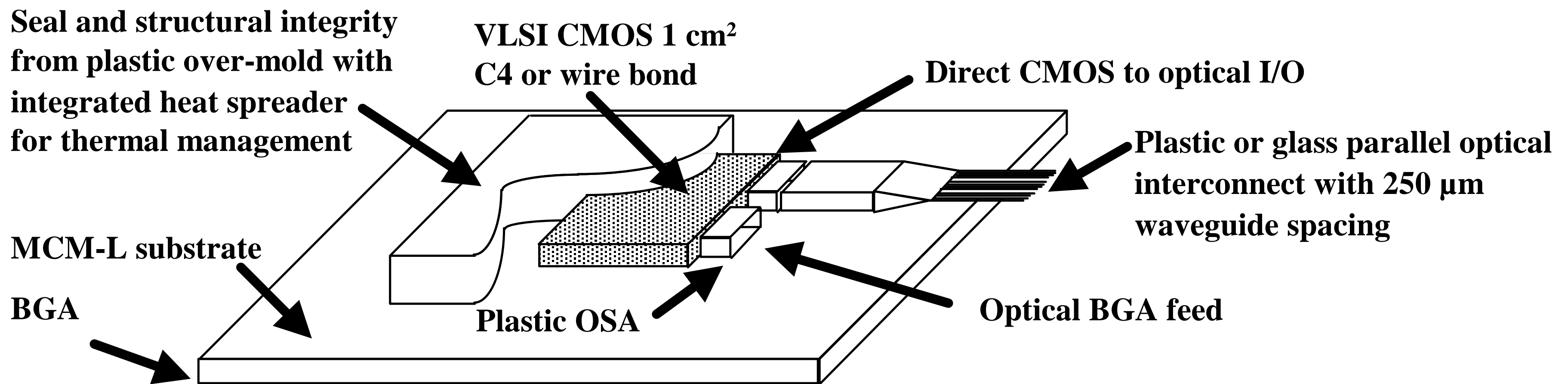


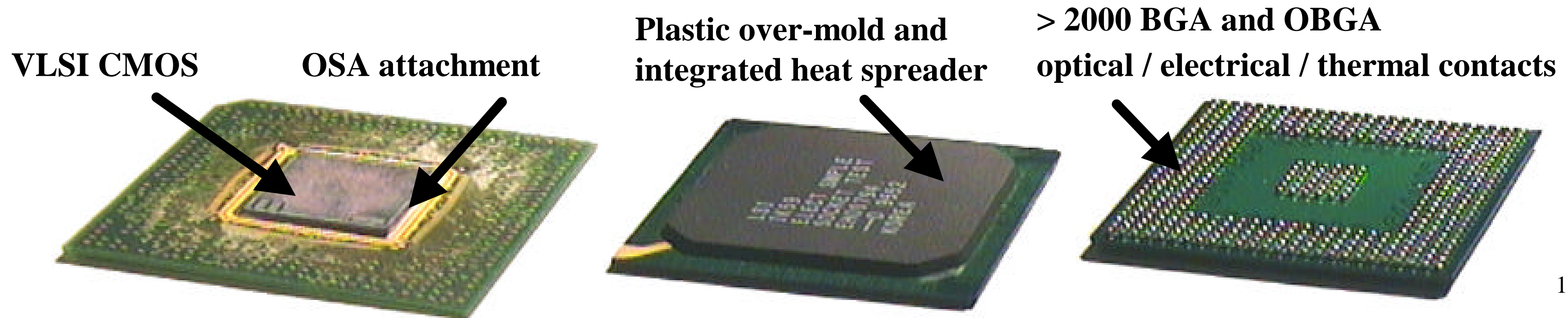

**> 0.1m backplane interconnect (optical 400ƒ Gb/s/cm per layer)**

**> 0.1m backplane interconnect**

# Future MCM-L packaging for direct CMOS to optical I/O

- Combines CMOS VLSI function with dense high-speed optical I/O
- Standard-cell opto-electronic CMOS interface
- *Separate* optical and electrical thermal management using integrated heat spreader
- Signal rate > 2.5 Gb/s per I/O

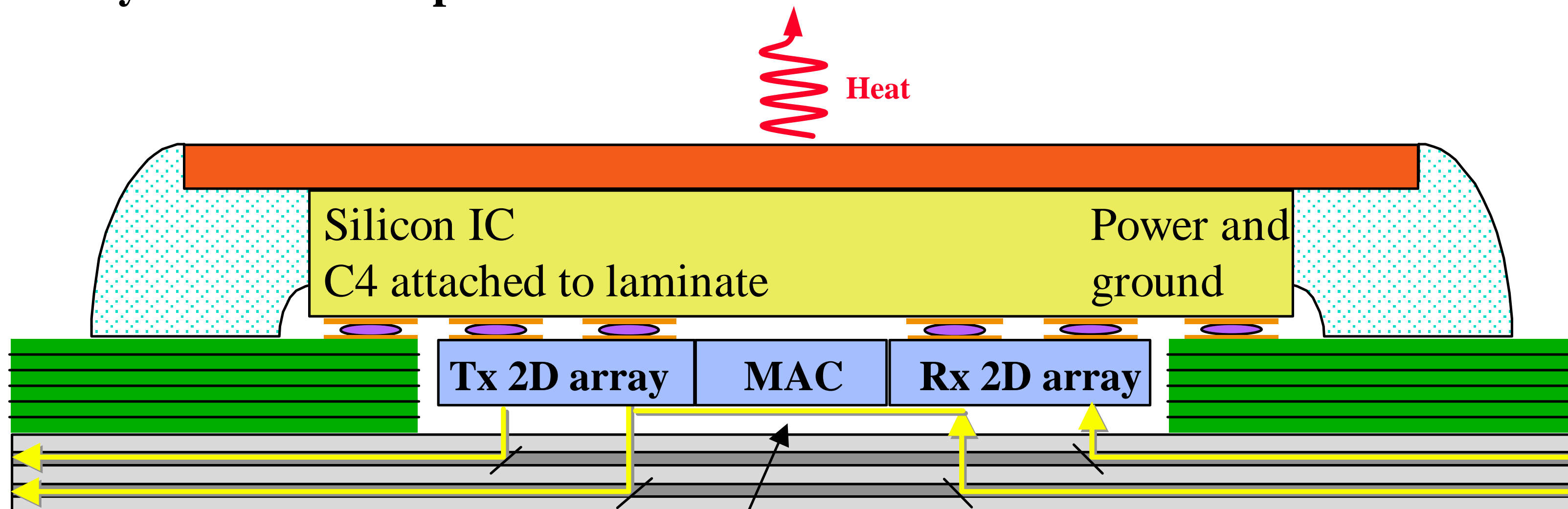


USC UNIVERSITY OF SOUTHERN CALIFORNIA

# Heterogeneous integration of photonic Media Access Control

**Ultra-low node latency using emerging functional micro-photonic technology. WDM all-optical routing using active micro-resonators have potential for ps node latency and deadlock protection.**



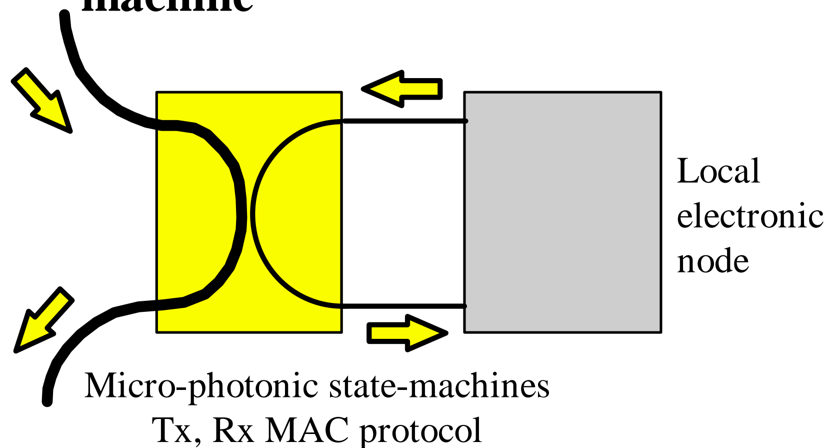

**Silica-glass or polymer multi-layer waveguide low-temperature laminate integrated with electrical laminate**

**All-optical functional micro-photonic circuit has pass-through mode for ultra-low node latency**

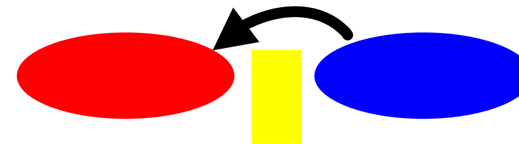
# Increasing functionality of micro-photonic devices

- **Challenge: Rapid switching of optical power from one high-Q resonator to another.**
  - **$U(t) = U_0(t = 0)e^{-(w_0 t / Q)}$**

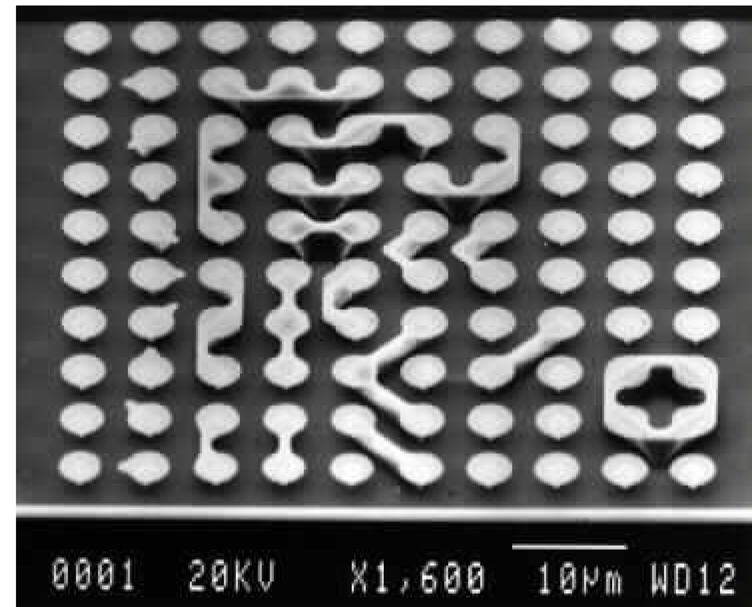
- **Functionality beyond point-to-point interconnect: Example**
  - **System Area Network MAC state machine**




Transfer of optical power between resonators



(i) Spoil Q to rapidly transfer optical power
(ii) Transient dynamics of cavity formation
(iii) Typical values, $\lambda_0 = 1.5$ µm, $\omega_0 = 10^{15}$ rads$^{-1}$, Q ~ $10^3$ - $10^4$ and $\tau_{ph}$ ~ 1 - 10 ps




**Functional micro-photonic circuits**

USC UNIVERSITY OF SOUTHERN CALIFORNIA

# The promise of optics: "Free, infinite-bandwidth, anywhere, anytime !"

- **New system design optimization and functionality**
  - **any distance for a given bandwidth**
  - **less heat 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**

USC UNIVERSITY OF SOUTHERN CALIFORNIA

# Optoelectronics "*inside*"

- **Optoelectronics 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**

- **DARPA**
  - **Focus**
    - **integrated optoelectronic / CMOS *inside* systems**
    - **WDM microphotonic functionality *inside* systems**
  - **Support *one-stop technology shopping***