Interconnects

Session 2490

Richard Kaufmann
High Performance Technical Computing Division

Many slides thanks to M. Krause, M. Levine, M. Maclaren, D. Roweth

Disclaimer

- This is a public talk.
- Some of the slides describe the interconnects in our standard product offerings.
- Other slides speculate about the technologies that may become popular in the industry as a whole. They don't represent plans or commitments from HP.

What I'll Talk About…

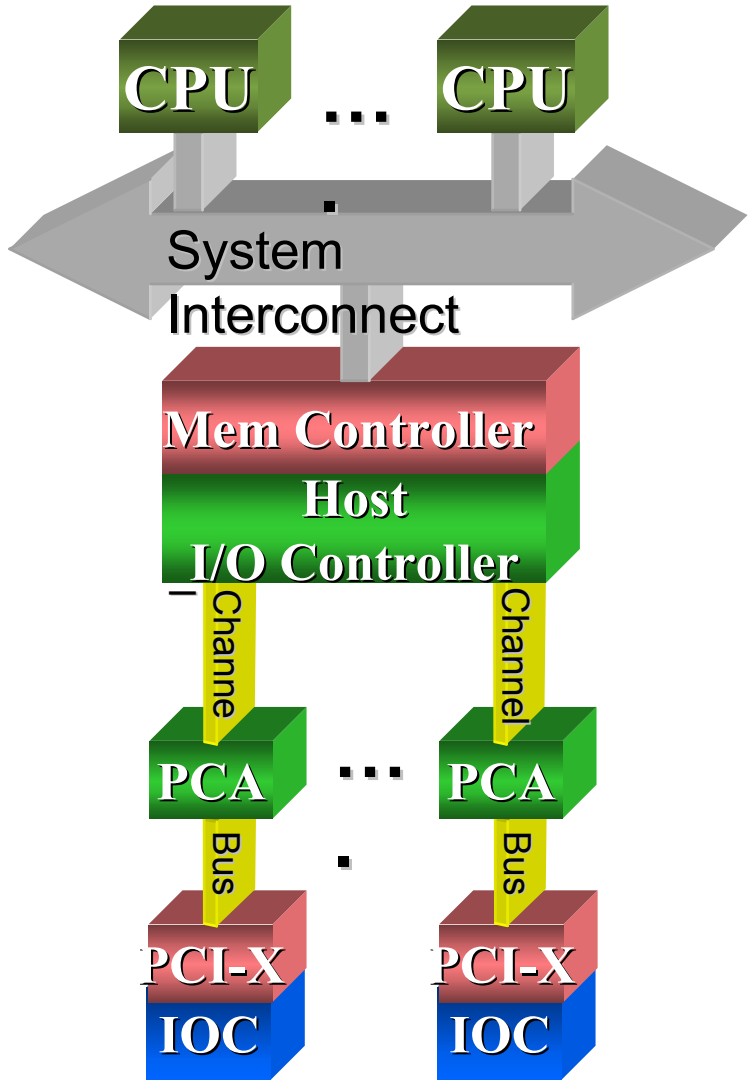
- System Interfaces
- System to System Interconnects
  - LAN
  - Commercial Clusters
  - Supercomputing

Network NICs drive Server Bandwidth

- System I/O interfaces have to keep up with the demands of the various interconnects
- We're currently at 250 – 1,000 MB/s per adapter
  - PCI 64/33 & 66 (ubiquitous adapter availability)
  - PCI-X 133MHz (many adapters available)
- Major change in the next few years
  - Bus → Point-to-Point or Switched Fabric
  - Bandwidth scaling to multi GB/s

PCI / PCI-X

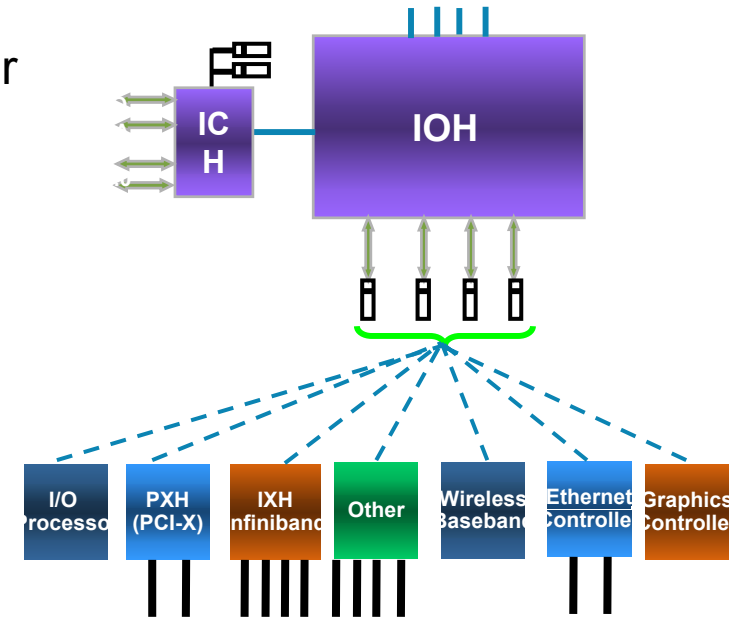
- Technology Focus:
  - Nearly everything uses PCI in some form
    - Servers, storage, desktop, mobile, etc.
  - Variable link widths and signaling
    - 33-133 MHz / wire half-duplex
    - Link widths: 32x, 64x, 128x, 256x
      128 & 256 are logical for new DDR / QDR
  - Primary deployment within chassis
    - Chip-to-chip and Adapter boards
  - Topologies: point-to-point, shared bus
    - Synchronous communication paradigm
  - Software compatibility
    - Ubiquitous deployment, most interop problems resolved
    - No standard I/O or IPC off-load capability
  - PCI-X products available this year
  - PCI-X 2.0 (2-4 GB/s) Very Soon
    - Product launches in 2004-5-…



PCI Express Overview

- **Technology Focus**
  - PCI Replacement for Desktop, Mobile, Server
  - Variable Link Widths and Signaling
    - Initially 2.5 Gbps per wire full-duplex
    - Link widths: 1x, 2x, 4x, 8x, 12x, 16x, 32x
  - Primary deployment within chassis
  - Short distance copper cabling TBD
  - Topologies: point-to-point, switched
    - Synchronous communication
  - Bridge chips to PCI / InfiniBand
  - Software Compatibility
    - New BIOS
    - Backward Compatible to PCI / Graphics Application Paradigm

      No standard I/O or IPC Protocol Off-load



- Products 2004-2007
  - First seen as an AGP replacement

LAN

- Ethernet: 10BaseT → 100BaseT → GigE → 10GigE → 40GigE
- Ubiquitous & completely standard
- Designed to perform in a hostile environment
- 50 – 300 usecs latency
  - Effort to reduce this to O(10usecs)
- As a given speed becomes predominant, cost goes into the noise.

Commercial Clusters

- Various Proprietary Interconnects, plus…
  - InfiniBand: Available now, but at reduced expectations
    - Competent clustering interconnect
    - Good performance for MPI (7usecs)
    - NOT going to take over the world
- Optimized for a machine room
- 5 – 20 usecs latency
- Various levels of aspiration to integrate WAN & Storage fabrics
- Low to moderate cost (2% to 10% of system cost?)

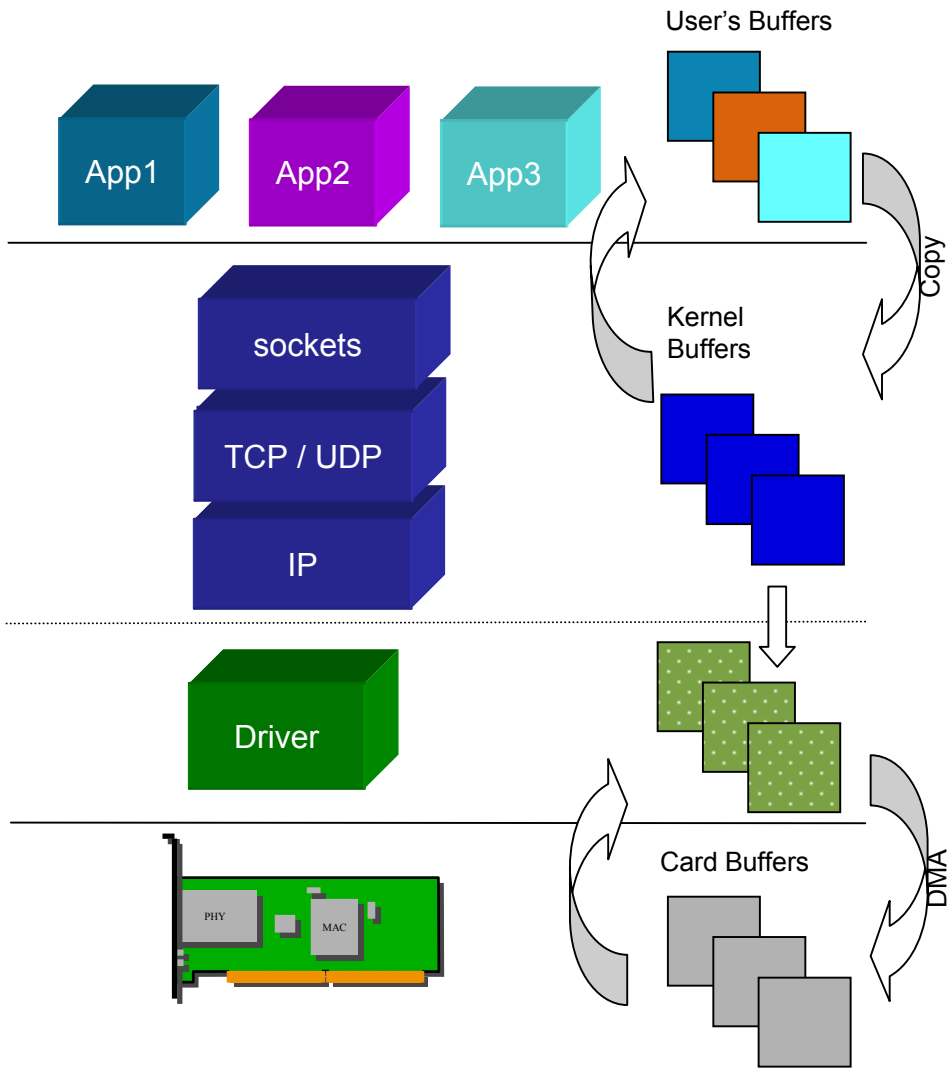
Pressure…

- The commercial interconnects have to differentiate themselves from Ethernet
  - Ethernet is crawling up quickly (if only to handle the increased bandwidth!)
    - RDMA: O/S bypass protocol
    - iSCSI: Integration of storage fabric into LAN
    - Attention paid to latency in NICs, switches
- And the supercomputing interconnects have to defend themselves against the commercial interconnects:
  - O/S bypass protocols used to be the domain of supercomputing interconnects
    - Latency numbers become less differentiated

      2X instead of 10X-30X

Techniques to Conserve Memory B/W

- Direct Memory Placement
  - Zero processor copy solutions
- Protocol offload
  - TCP, iSCSI, SSL, IPSec
- OS Bypass
- RDMA



Direct Memory Placement

- Definition
  - DMA to / from Upper Layer Protocol memory
- Advantages
  - Eliminates copy to/from intermediate I/O buffer
    - Decreases memory pressure for system
  - Proven technique in high-speed OS implementations
    - Dramatic performance benefit for key applications, e.g. ftp, NFS, SPECWeb
- Requirements
  - Application memory must be made accessible to I/O subsystem
    - Set up / tear down costs (can be mitigated by intelligent implementation)
  - Clear buffer ownership delineation to avoid data corruption

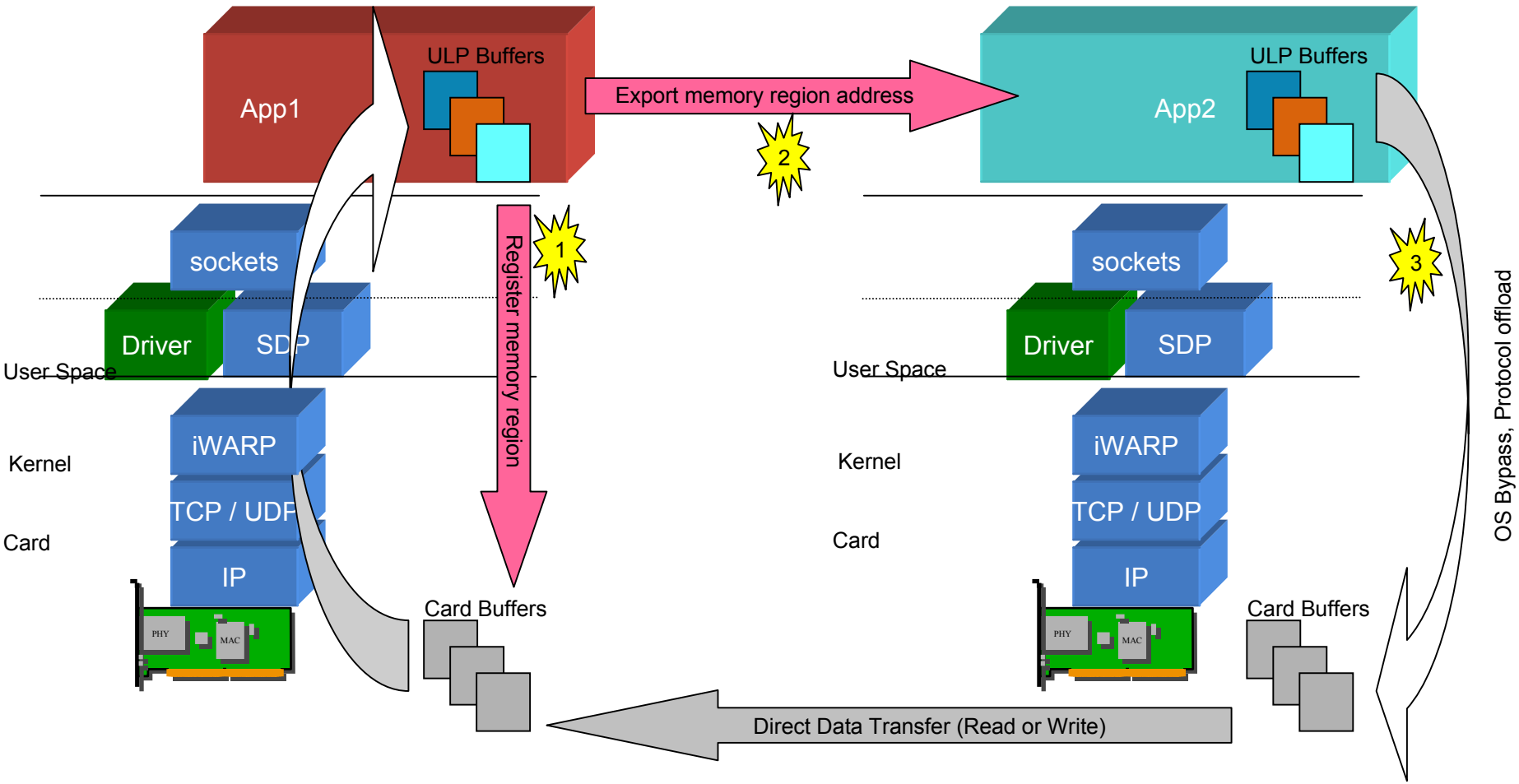
Protocol Offload

- Definition
  - Network protocol off-load services implemented in intelligent NIC / chipset
    - Complete acceleration of main data path, e.g. TCP/IP/Ethernet, UDP, IPSec, etc.
    - Optional acceleration of ULP – iSCSI, iWARP(IETF RDMA), SSL/TLS, NFS, etc.
- Advantages
  - Reduced memory B/W consumption
    - Eliminate CPU I-Cache / D-Cache misses due to protocol processing
  - Supports interrupt coalescence / avoidance
    - Further reduction in system overhead
  - Reduced I/O work requests
    - Fewer I/O transactions per large send / receive operation
    - Enables flexible QoS policies to support latency sensitive applications
  - Easier to maintain system balance
    - Host Processor / Memory / I/O B/W spent on application not protocol

OS Bypass

- Eliminates application invocation of OS system call
  - No OS involvement to send / receive data or perform completion / event processing
- Advantages
  - Eliminates user-kernel context switch overhead from main path
  - Improves application responsiveness
  - Decreases end-to-end message latency

Remote Direct Memory Access (RDMA)



Supercomputing Interconnects

- MPI is the killer app.
  - Various pretenders to the throne.
- All proprietary below the API
  - Actually, model public domain implementation (MPICH) has helped
- System packing density important: distance limited
  - Becomes less true with new optical technologies
    - Better Cost/Bit/Second → More Cost Effective
    - Easier to cable!
- Widely varying bandwidth needs.
  - 100:1 Spread: 0.01 B/FLOP → 1.0 B/FLOP!
- Latency in the 2 usec – 10 usec range
- May well hit 25% of system cost.
  - But for a class of the super-big apps, this is *the* single most important performance factor

Supercomputer Choke Point

- LANs and Commercial Clusters drive the Server marketplace
  - Top Ethernet speed (at any given time) starts as a switch-to-switch technology, then later migrates to Servers.
- The supercomputing interconnects quite often overwhelm the server interface
  - Bandwidth chokes of 2X not unusual
  - Observation: Even this requires great attention from system designers.
  - PCI-X 2.0 (and later PCI Express)

    Will enable next generation of supercomputer networks

Examples of Supercomputing Interconnects

- Vendor-Neutral
  - Dolphin Interconnect Solutions Wulfkit3
    - 3D Torus; kit consists of adapter and 3 cables
  - Myricom's Myrinet
    - Multi-stage switch
  - Quadrics' Qsnet
    - Multi-stage switch, value-add supercomputing features
  - Infiniband
    - Various vendors
    - Moved into HPTC over the last year
- Vendor-Proprietary
  - IBM, NEC, et al

Example: hp AlphaServer SC

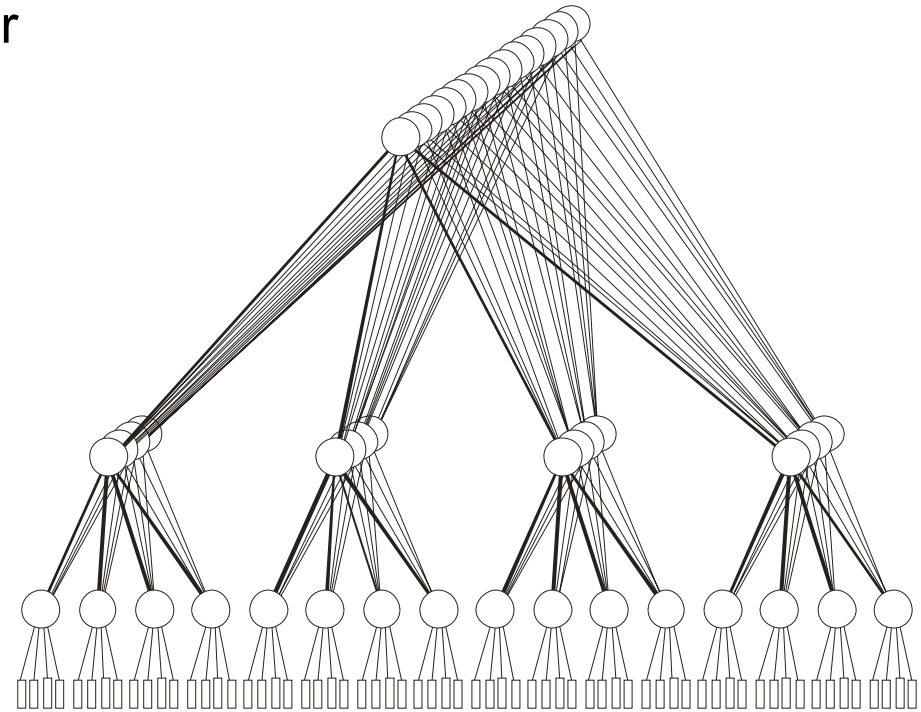
- A distributed-memory supercomputer
  - Made up of quad-processor Alpha servers
    - 750 * 4 = 3,000 CPUs at PSC
  - Tru64 Unix
  - Quadrics Elan-3 Qsnet Network
    - PCI 64/66
    - True, virtual RDMA
      Full get&put
    - No need to lock down pages
    - Full OS Bypass Path
    - H/W Broadcast & barrier
- Three of the top 10 supercomputers in the world

Example: PNNL Intel® Itanium® 2 System

- Made up of hp server rx2600 “long’s peak” nodes
  - Two Processors, each 1.5GHz “Madison”
  - 4 FLOPS/tick → 12GF/server
  - O(1000) Servers
  - Linux
- Connected by two “rails” of Elan-3

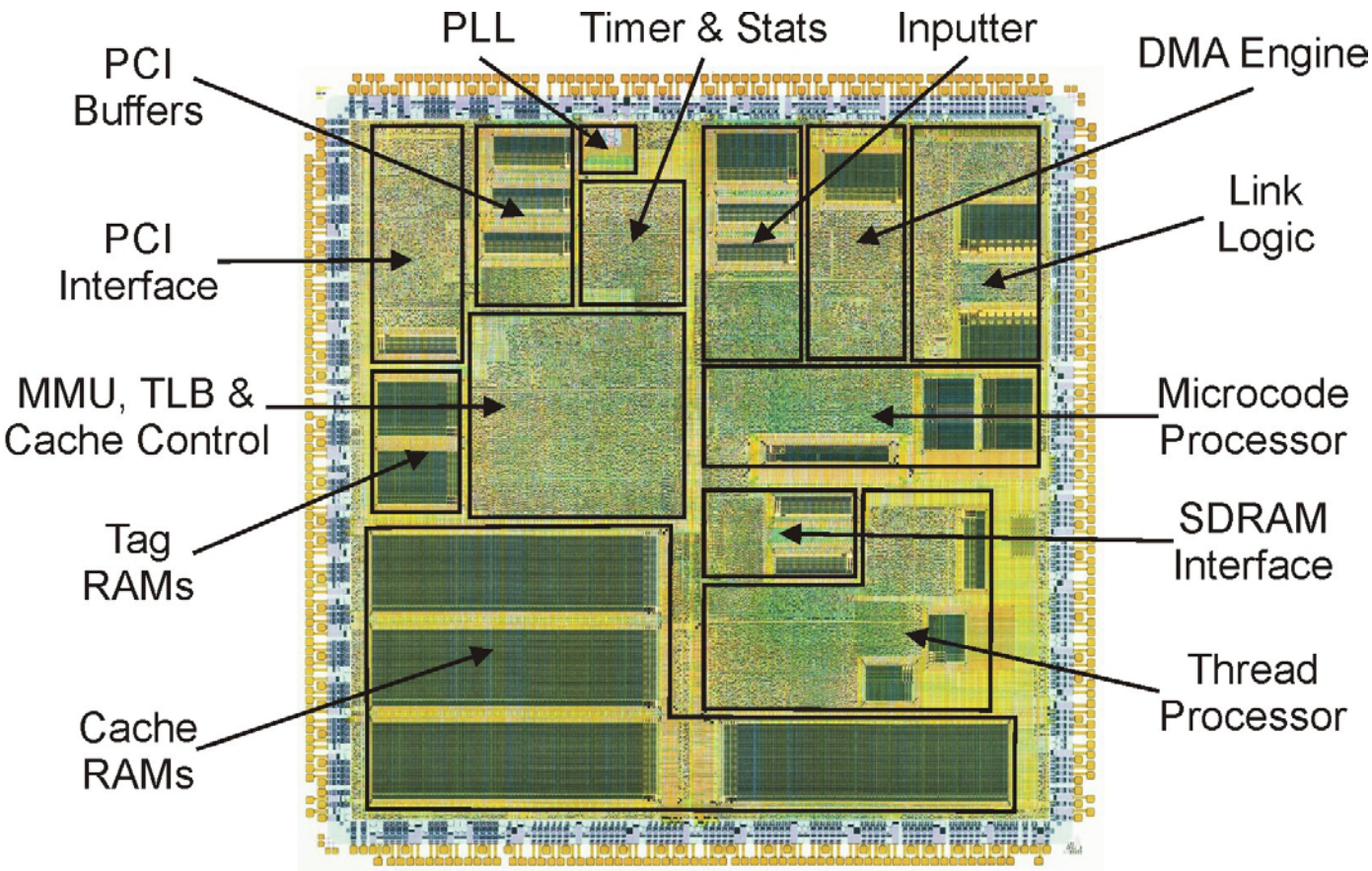
Multi-Stage Switches

- Full bi-sectional bandwidth
- Hardware broadcast & barrier
- Logarithmic cost
- Multiple routes

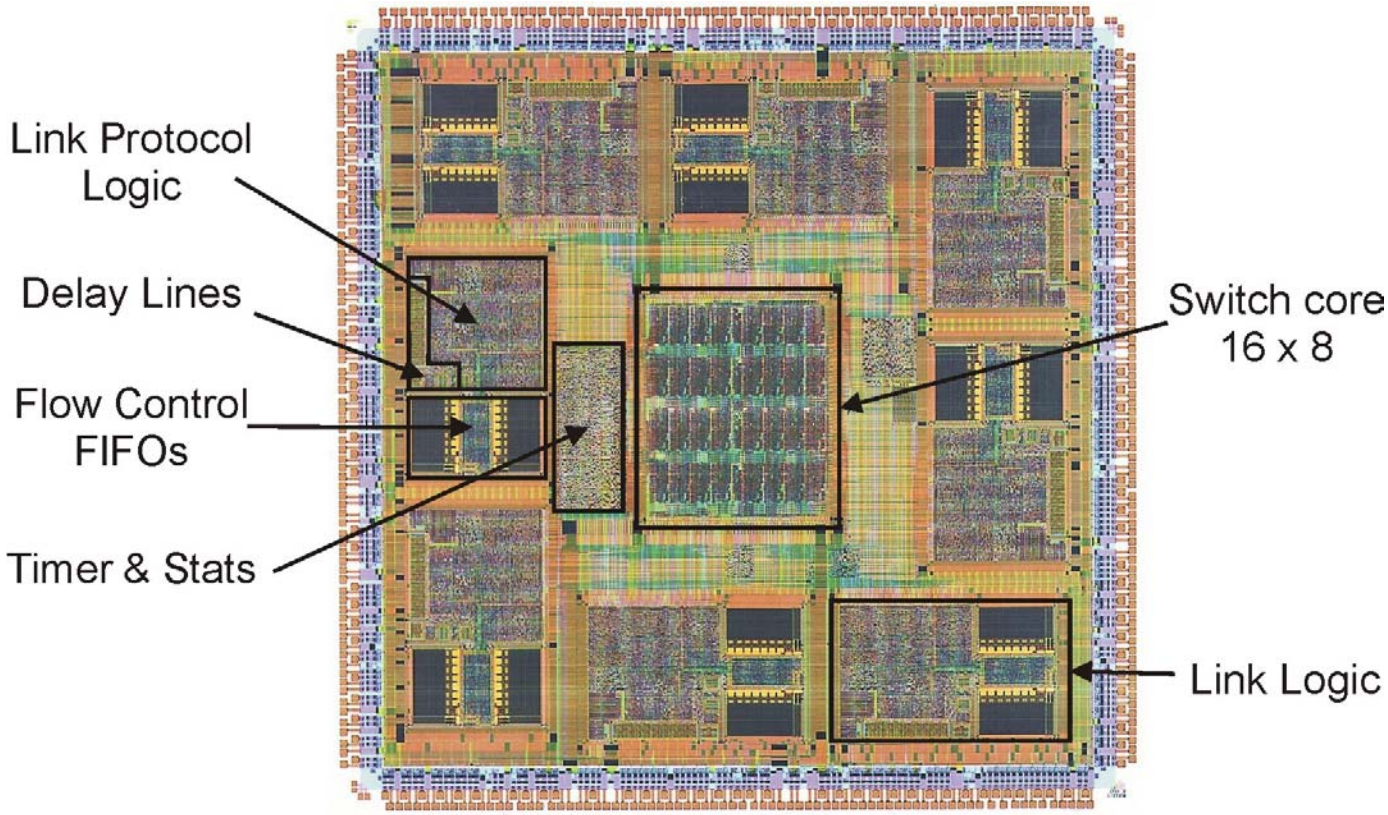


Adapter per node

Elan3 Adapter



Elite3 Switch

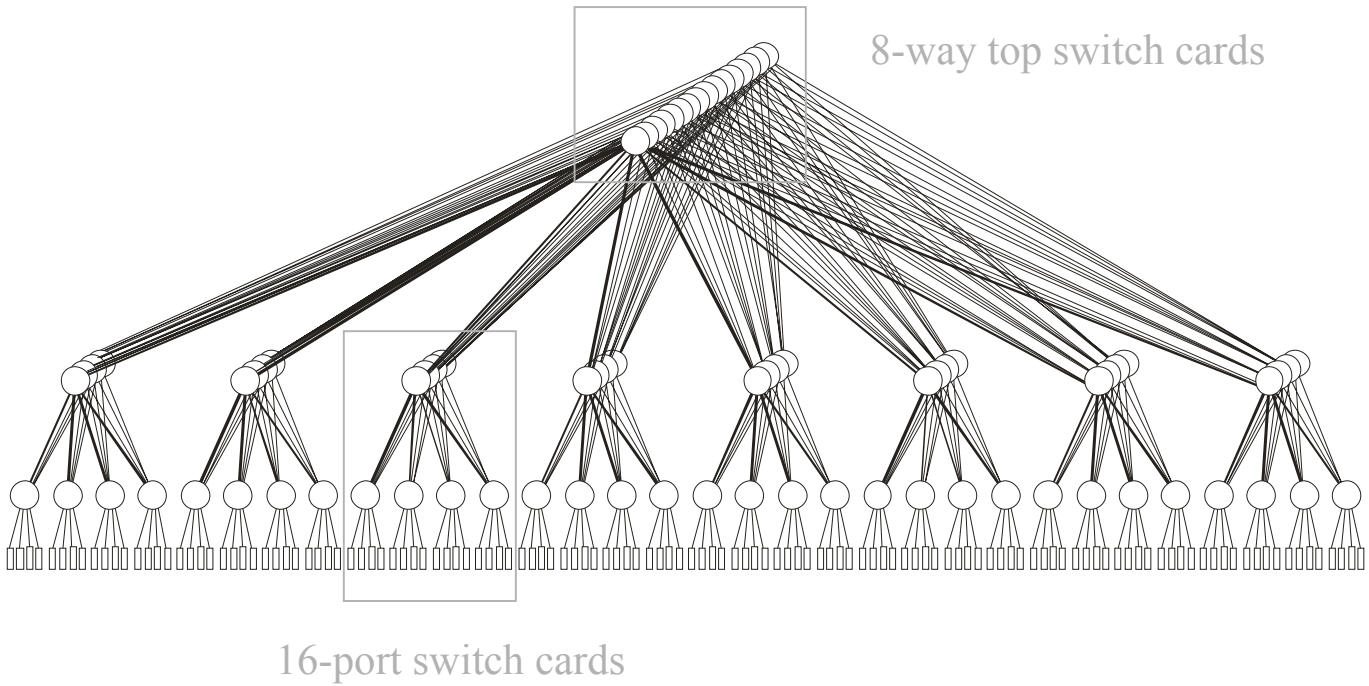


QsNet Packaging

- Network Options
  - 16-128 nodes in units of 16
  - 64-1024 nodes in units of 64
- Switch Chassis
  - 42.5 Gbytes/sec bi-sectional bandwidth
  - 35 - 175 ns unblocked latency
  - Temperature, fan and PSU monitoring
  - Live insertion
  - Embedded controller

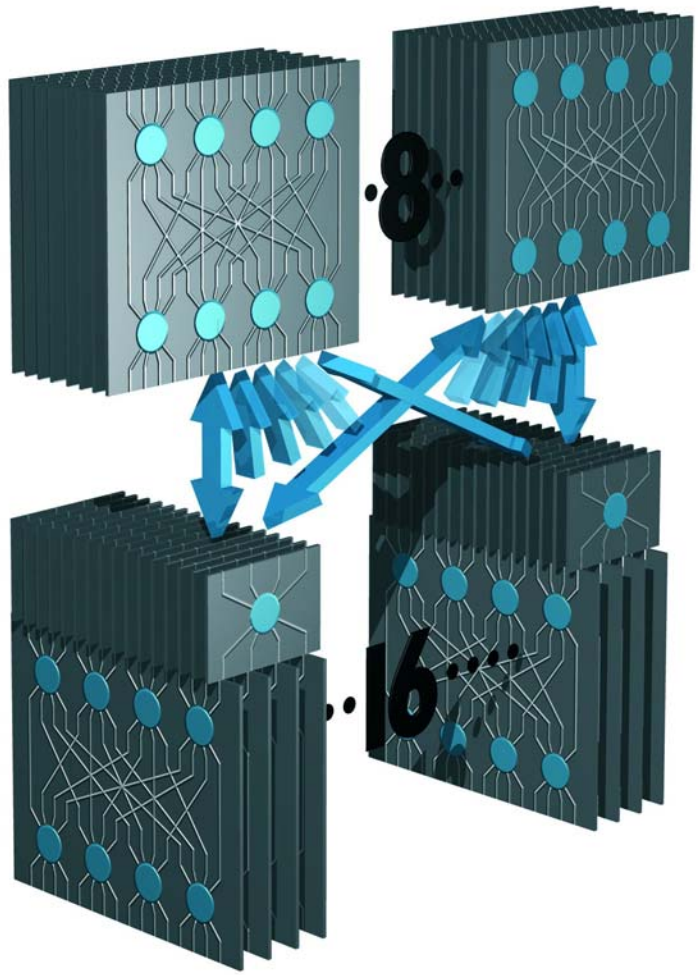
128-Way Switch Chassis

- 1-8 ×16 way switch cards
- 16 × 8 top switch cards

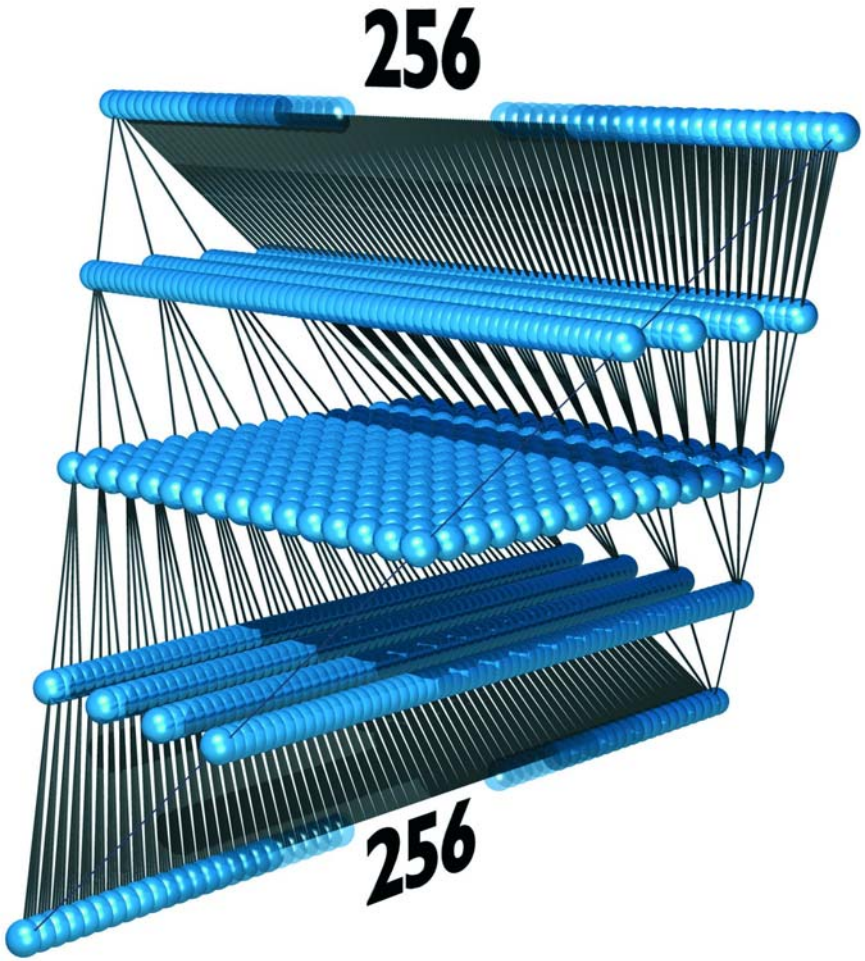


Federated Networks

- Top switches
  - Multiple ×4, ×8, ×16 top switches connecting the node switches together
- Node switches
  - 64 links from nodes
  - 64 links to the top switches
- Common components



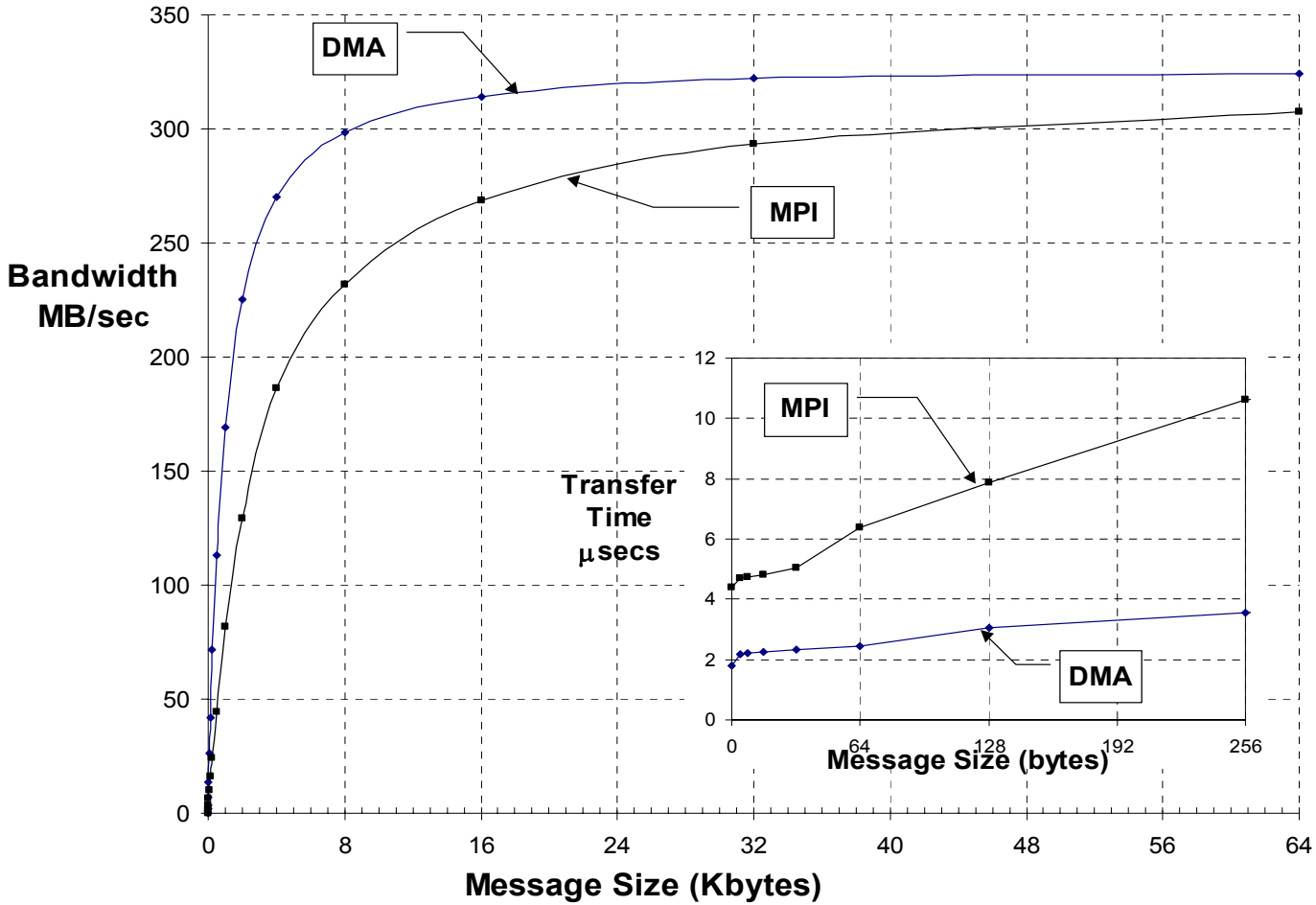
1024-way Network (PSC)



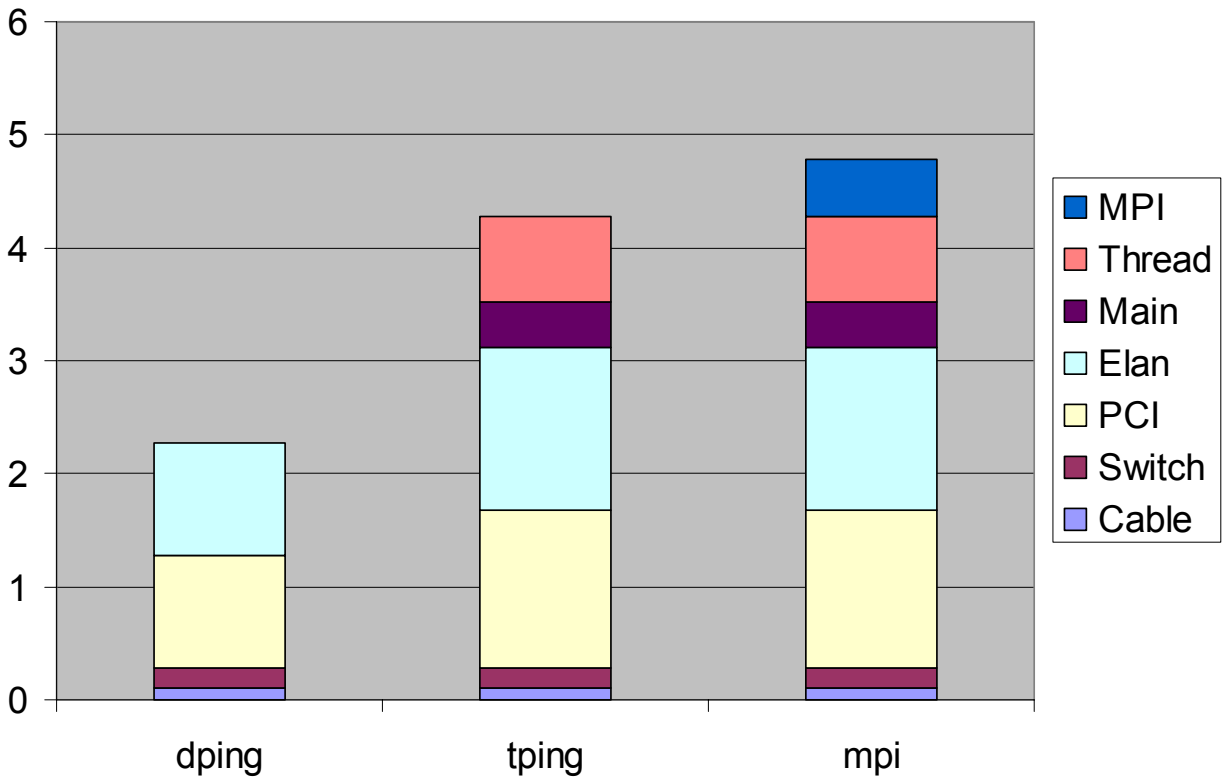
Elan-3 Performance Overview

- Latency
  - MPI send / receive — 4 - 5 μsec
  - Shmem put / get — 2.5 μsec
  - Pipelined put / get — ≤ 1 μsec
- Bandwidth
  - MPI & Shmem — 325 MB/sec
- Depends on PCI implementation
  - AlphaServer ES40 — 2.25 μsec — 210 MB/sec
  - AlphaServer ES45 — 2.0 μsec — 305 MB/sec

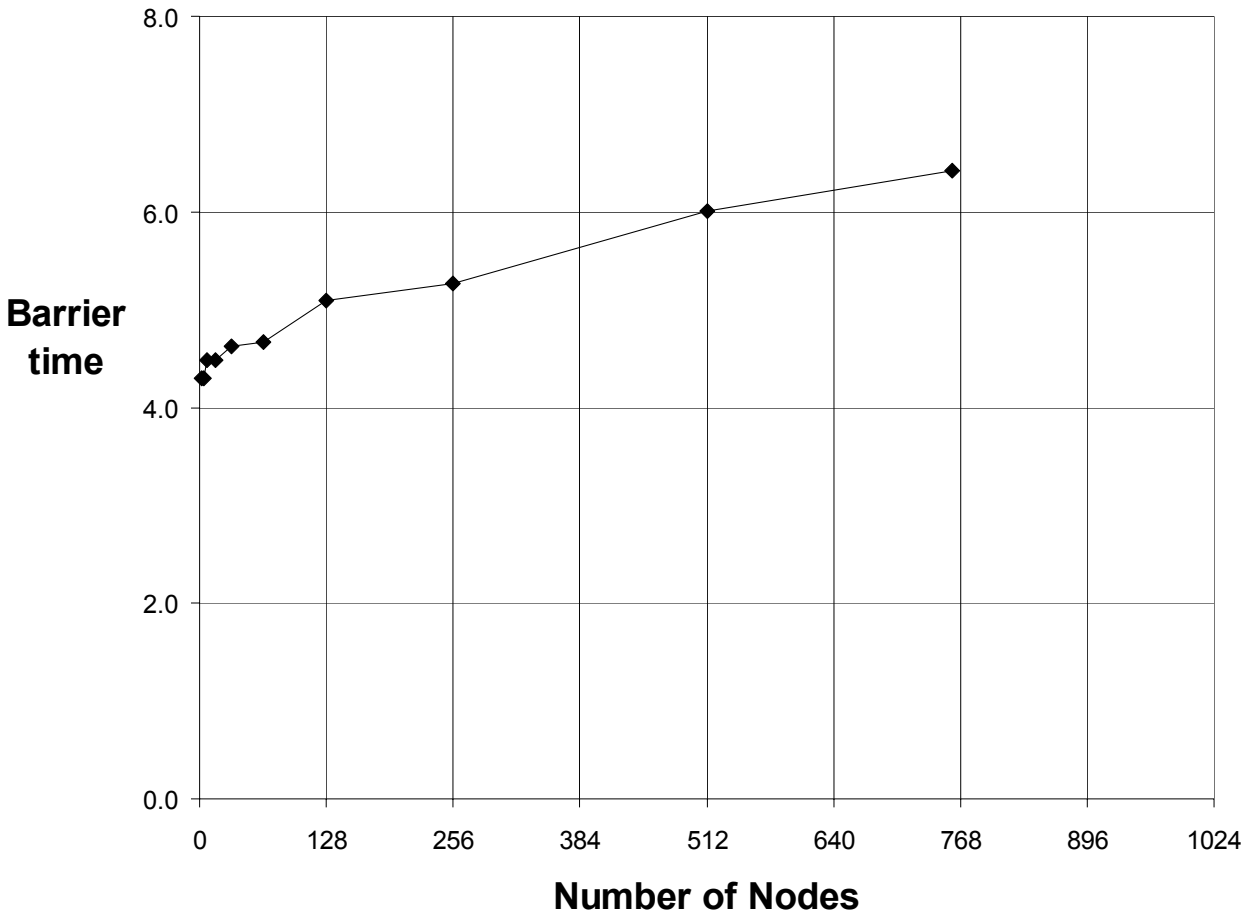
Latency and Bandwidth



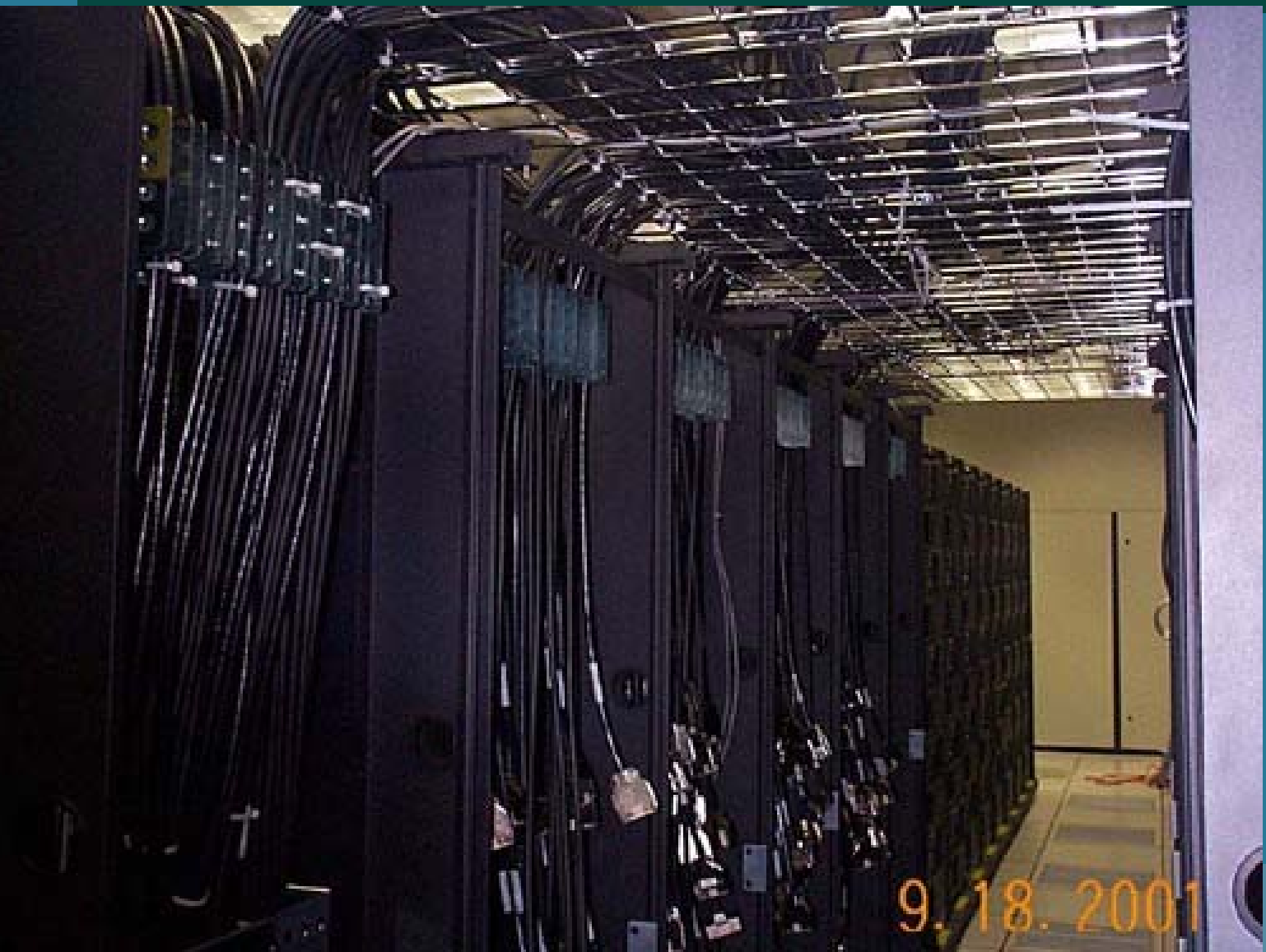
Elan-3 Latency Breakdown



Barrier Scaling



Switch Wiring



PITTSBURGH SUPERCOMPUTING CENTER

Future Supercomputing Interconects

- PCI-X – based Quadrics
  - Respin of NIC, Switch ASICs. New cabling (hybrid copper/optical).
  - You'll see it at Supercomputing 2003
- Personal Predictions
  - PCI-X 2.0 (2-4 GB/s) enables next bandwidth bump
  - PCI Express starts in 2004, becomes predominant by 2006/7
    - Both widely taken advantage of by supercomputing clusters
  - Five Years Out
    - Advanced all-optical switching networks

      High relative cost at the beginning. Initially only for apps that are message passing limited, and have lots of $$$ behind them!
    - Back to grousing about server interfaces!