Overview & Technology Roadmap

Charles L. Seitz, Ph.D.
CEO & CTO of Myricom, Inc.
chuck@myri.com

20 February 2003
Presentation for the PC-Cluster Consortium

Outline

- What is a cluster?
- What is Myrinet?
- Technology Roadmap

What is a Cluster?

A working definition:

A computing cluster is a collection of interconnected computers that employ distributed-computing techniques to work together (by teamwork) on a computing problem.

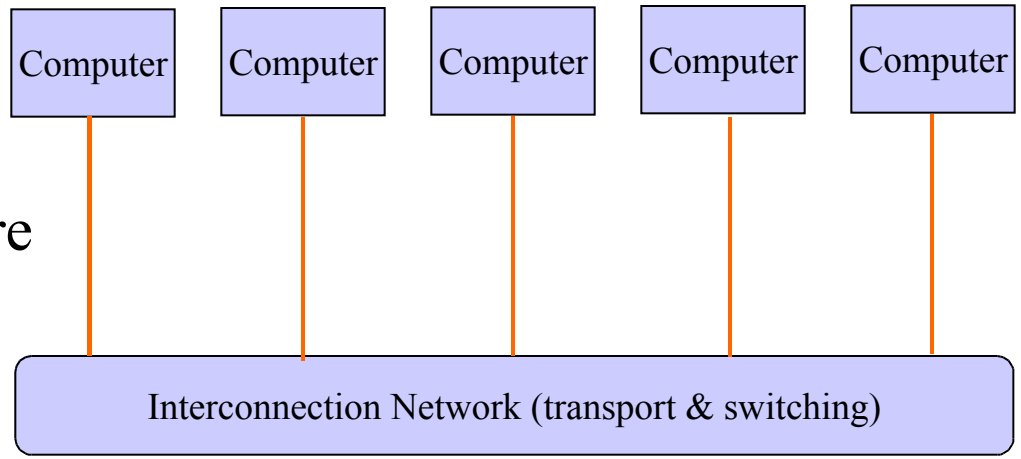
Relevance:

Clusters are of practical interest for high-performance computing because they offer excellent performance/cost, and essentially open-ended performance for most large-scale computations.

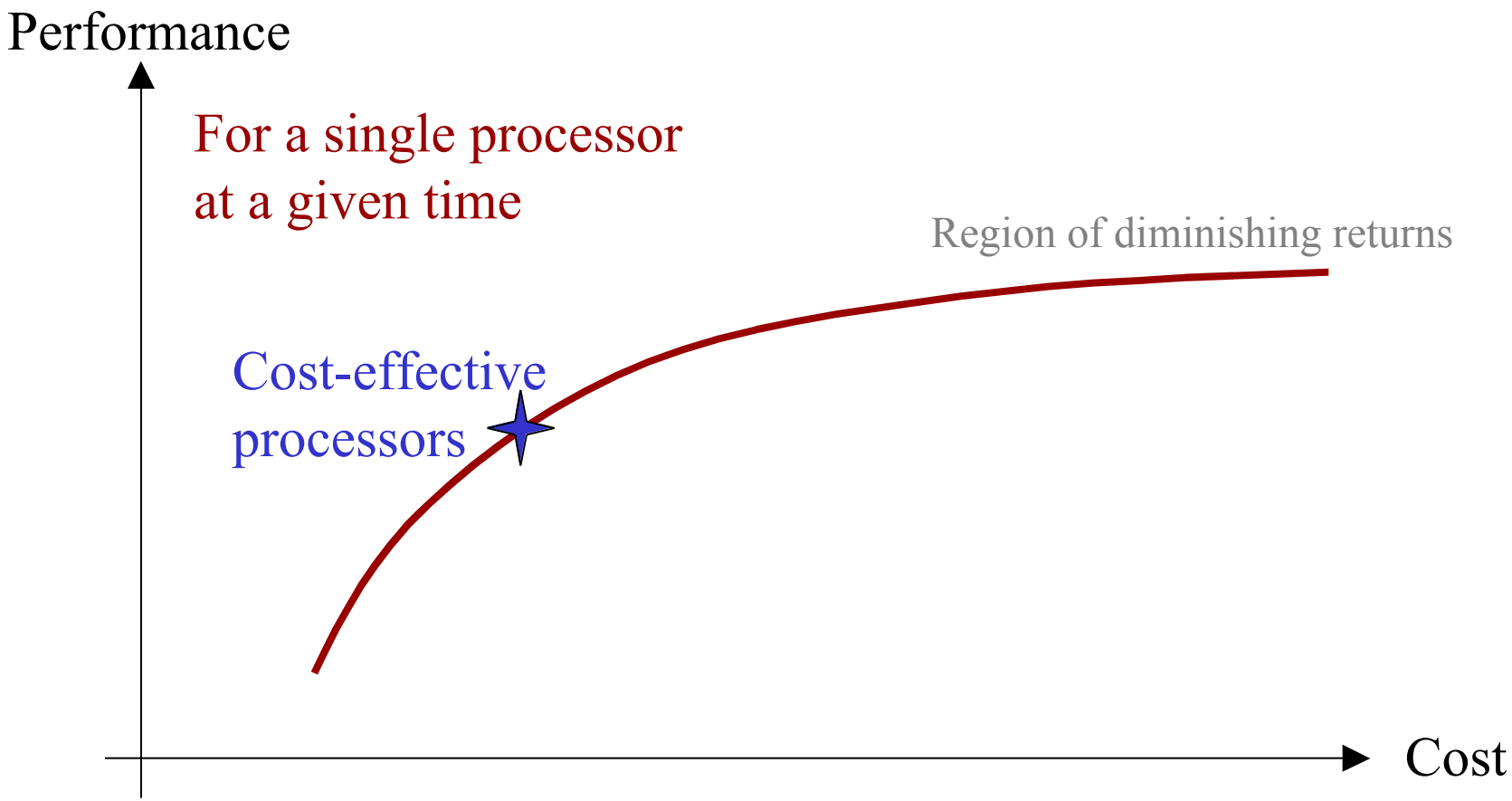
Clusters also provide high availability: a 50-host cluster with two hosts “down” is a perfectly good 48-host cluster.

Cluster Computing Builds Upon …

- Distributed-computing Applications
- Programming Systems
- Operating Systems
- Commodity Computers
- Low-level Network Software
- Interconnection Network



The performance/cost argument for clusters



Cluster computing shifts the burden of extending performance from the processors and memories, where it is limited by physics and technology, to algorithms and programming (distributing an application across a collection of cost-effective hosts).

The success of clusters (data from the TOP500)

- 1 cluster (the Berkeley NOW) in June 1997
- . . .
- 33 clusters in June 2001
- 43 clusters in November 2001
- 80 clusters in June 2002
- 93 clusters in November 2002

However, nearly all of the supercomputers sites in the TOP500 are distributed-memory systems. The authors of the TOP500 reserve the "cluster" classification for systems in which the interconnect is not proprietary. Thus, for example, the IBM SP – clearly a cluster in architecture – is classified as an MPP because it uses an IBM-proprietary interconnection network.

Design choices for a cluster today

- **Hosts.** The best peak performance/cost (the sweet spot) is with 2-4 processors per host (small-scale SMP architecture).
  - Dual Xeon hosts are doing very well right now for HPC clusters.
- **Operating System.** Linux, of course.
  - Why would you want to use a proprietary OS?
- **Interconnection Network.** A useful "rule of thumb" is that the (summed-bidirectional) data rate to and from a host need not be more than a modest fraction, 5-20%, of the host's memory bandwidth.
  - A distributed computation that consumes so much memory bandwidth in communication that it impacts compute performance is over-distributed.
  - 10% of a Xeon's ~4GB/s memory bandwidth, ~400MB/s, is too much for GbE, but just right for Myrinet.
  - Communication rates may be limited by the host I/O bus (*e.g.,* PCI bus).
  - **Message latency and host-CPU overhead.** The other important micro-benchmarks for cluster interconnect.

Typical Technical-Computing Cluster Applications

- Finite-element codes (MSC Software; LSTC LS-Dyna -- automotive crash simulations, metal forming, metal cutting, biomedical, earthquake engineering)
- Computational Fluid Dynamics (Fluent, Zeus-MP, and others)
- Weather-simulation codes (MM5)
- Quantum chemistry & physics (GAMESS)
- Computational astrophysics (Cactus)
- Computational chemistry (CHARMM, AMBER, and others)
- Computational biology (macromolecular modeling, gene sequencing)
- Signal and image processing, including military applications

Typical Commercial/Business/Web Cluster Applications

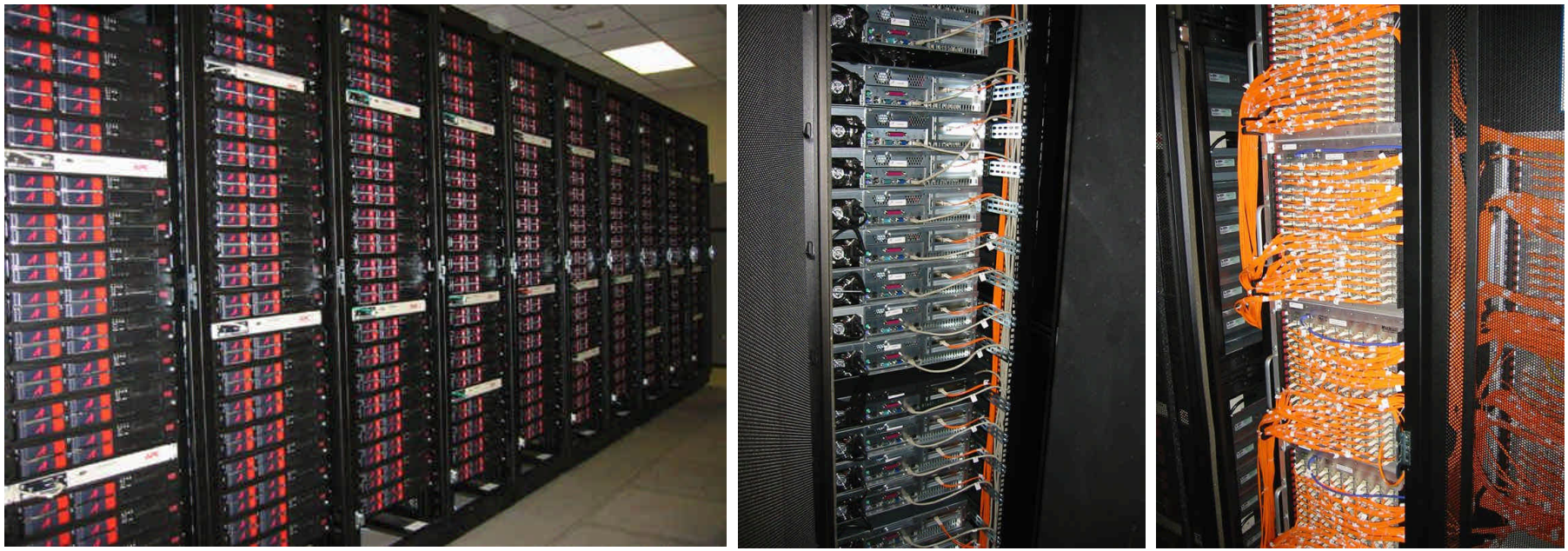
- Web search engines
- Databases (e.g., Oracle & DB2) and data mining
- File systems
- Graphics (fast rendering, display walls)
- Oil and gas exploration
- Automotive collision and engine-performance simulations
- Pharmaceutical research
- Financial transaction processing
- Medical imaging

What is Myrinet?

- A set of **standard products**
  - Interfaces, software (bundled, open-source), switches, and cables.
  - All you need to make a high-performance cluster from a set of host computers.
- A network architecture, protocol, and **technology**
  - A descendant of packet communication and routing in MPPs, but open.
  - ANSI Standard (ANSI/VITA 26-1998).
  - The architectural dual of Ethernet: Processing power is concentrated in the hosts and interfaces, allowing an elegant, streamlined, switching technology.
  - A technology with a very wide range of application.
- The clear **market leader** for high-performance, high-availability interconnect for computing **clusters**
  - ~90% of the interface and switch ports shipped in this niche.
  - First customer shipments: 1994. Installed base: thousands of sites. Current shipping levels: $5{,}000^{+}$ interfaces & $10{,}000^{+}$ switch ports per quarter.

Myrinet Technology "in the Large"

LSU "SuperMike" Cluster

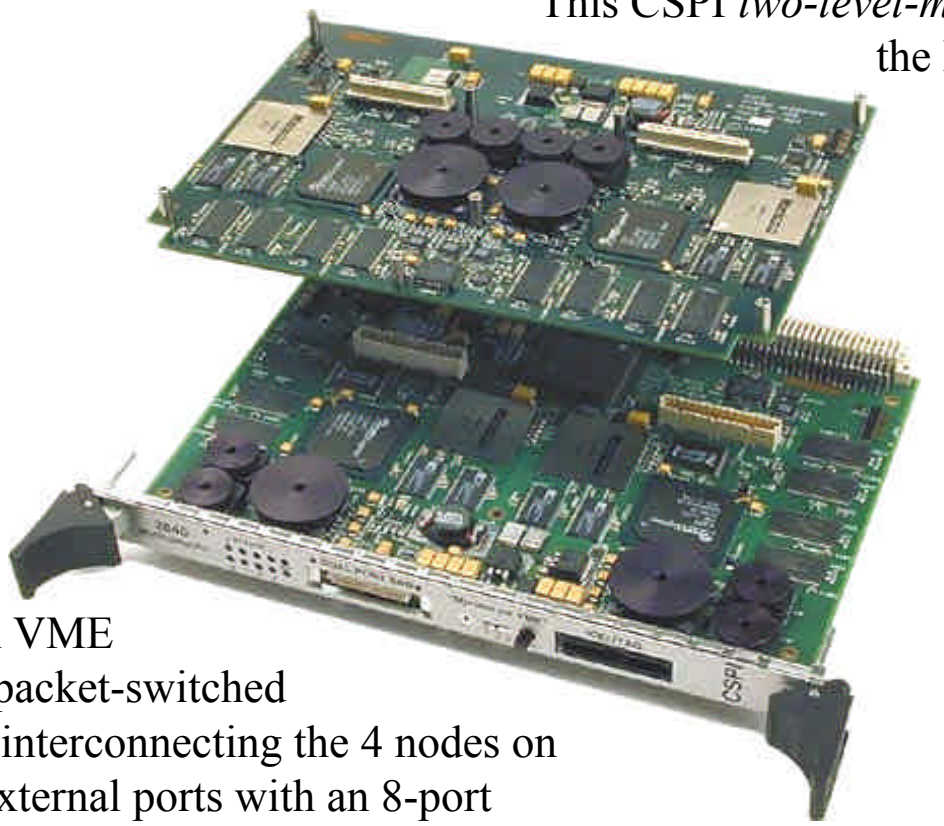


The SuperMike cluster was built for LSU by Atipa Technologies in the Spring 2002. SuperMike consists of 512 dual 1.8GHz Xeon servers connected with Myrinet interfaces, switches, and nearly 10 km of optical-fiber cable. The bisection data rate of the 512-host Myrinet Clos network is 1.024 Terabits/s. SuperMike achieved a LINPACK performance of 2207 Gigaflops for the November TOP500 supercomputer list, making it the 17th fastest system in this list.

Myrinet Technology "in the Small"

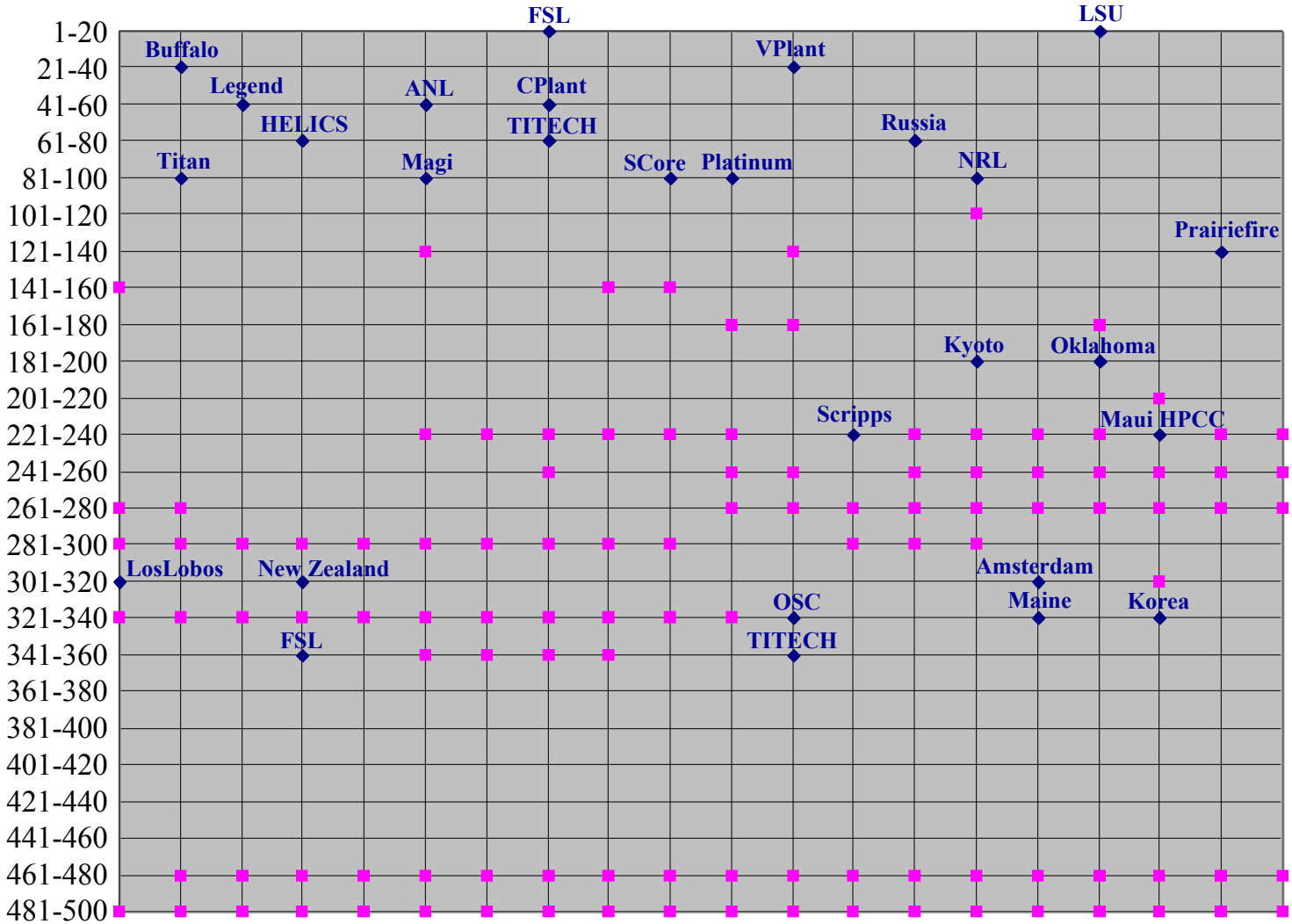
CSPI Quad-PowerPC VME Signal-Processing Board

This CSPI *two-level-multicomputer* product uses the Myricom LANai-9 chip to interface the PowerPCs to the message-passing network.



This single-width VME board includes a packet-switched Myrinet network interconnecting the 4 nodes on the board and 4 external ports with an 8-port Myrinet switch (barely visible in this photo).

Myrinet in the November 2002 TOP500



15 of the top 100

140 of the

TOP500 SUPERCOMPUTER SITES

use Myrinet technology

Myricom

Myricom, Inc. 325 N. Santa Anita Ave. Arcadia CA 91006

626-821-5555 Fax: 626-821-5316 http:/ /www.myri.com

Why Myrinet? The "selling points."

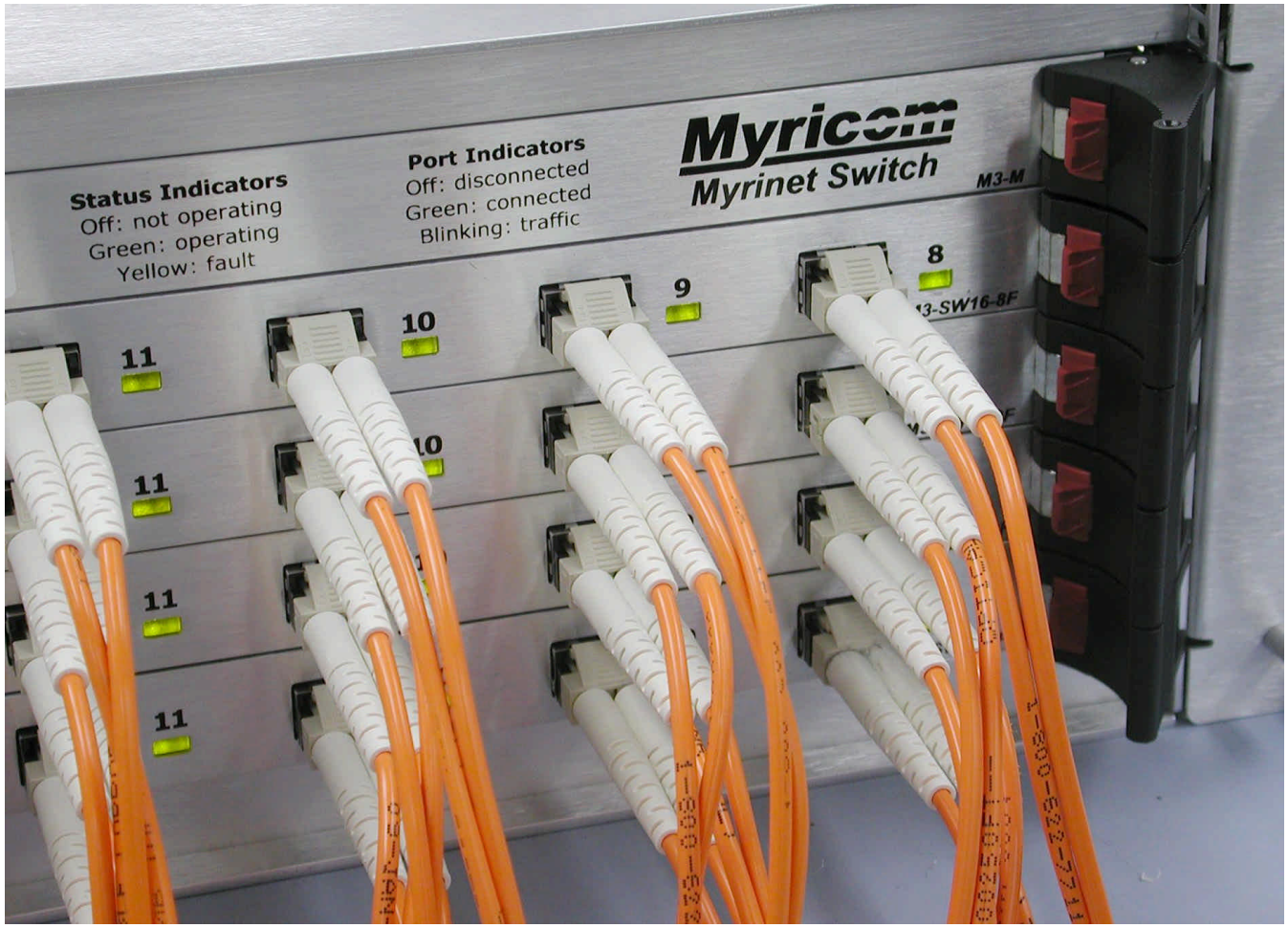
- Low latency *(The lower the latency, the wider the application span)*
  - ~7µs today (UNIX user process to user process, fully protected, with end-to-end data integrity checking)
  - ~5.5µs & ~4.8µs soon.
- High data rate
  - 2+2 Gb/s (250+250 MB/s) links
  - Interfaces with multiple links soon
- Unlimited scalability
  - Full-bisection Clos networks up to 8192 hosts (or more)
- Very low host-CPU utilization
  - logP < 1µs
- Multimode-fiber links
  - Lightweight, small diameter, reliable
- High Availability features
  - Self-mapping, self-healing
  - Link-continuity monitoring
- Data Integrity features
  - Memory and bus parity
  - Link and packet-payload CRCs
- Software drivers for almost all major platforms
  - Download them from the Web
  - Open source
  - Low-level API + TCP|UDP/IP + MPI + VI + PVM + Sockets
- Hybrid Myrinet/GbE networks
  - Coming very soon

Technology Roadmap

Designers'-eye view of Myrinet cluster interconnect:

- Links
  - Data transport and packet protocols.
  - Optical-fiber cables or electrical signals on circuit boards.
- Switches
  - Route data from any host to any other host. The network should be scalable to many hosts.
- Interfaces
  - The physical interface between the host I/O or memory bus and the packet network. The interface must support "zero-copy" data transfers, and offload protocol processing from the host.
- Software
  - Interface firmware, device driver, and "middleware" to support APIs.

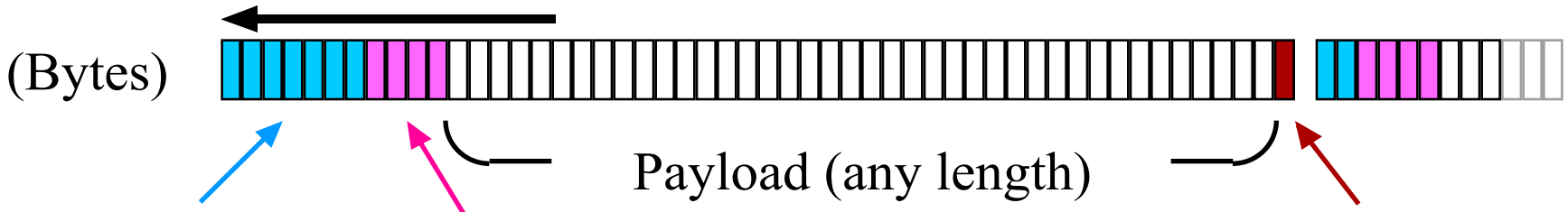
Myrinet-2000 Links, 2+2 Gbit/s, full duplex



The signaling rate on a fiber is 2.5 GBaud, which, after 8b/10b encoding, yields a data rate of 2 Gbits/s

Myrinet = ANSI/VITA 26-1998

Myrinet is defined at the Data-Link level (level 2 of the ISO reference model for computer networks) by its packet format and flow control. *Think of Myrinet as the simplest packet-switched network you can devise.*



Source route used by the switches, which strip the bytes as they are used

Type (allows multiple protocols on one Myrinet)

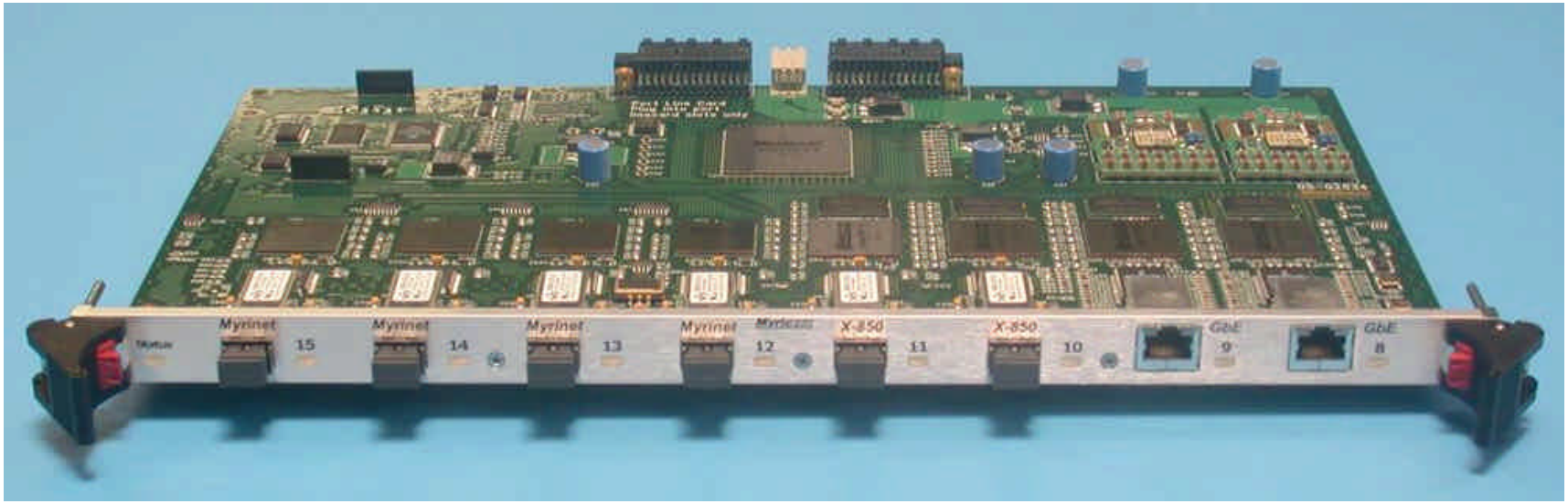
http://www.myri.com/open-specs/

There is flow-control on every link.

Changes in Myrinet Links

- *June 2002* - first chips (Lanai XM) with multi-protocol (X) ports
  - A multi-protocol (X) port can act as a Myrinet port, long-range-Myrinet port, GbE port, or InfiniBand port.
  - Interoperability between Myrinet, long-range Myrinet, GbE, & InfiniBand.
- *2Q03* - PCI-X interfaces with two ports
  - Two links acting as one is provided by multi-path (dispersive) routing in GM 2.
  - 2 x (250+250) MB/s = 1GB/s, a good match to 1 GB/s PCI-X.
- *Mid 2003* - SerDes integrated into Myricom custom-VLSI chips
  - To be used initially in a SAN-Serial chip, then in a 32-port Xbar-switch chip.
  - 2+2 Gb/s data rate, 2.5+2.5 GBaud (8b/10b encoded) links are also used as the base PHY by PCI Express. Myricom plans to support PCI Express -- initially 4x -- as soon as PCI Express hosts become available (expected early 2004).
- *2004* - "4x" Myrinet (multi-protocol) links
  - Most product volume is expected to continue with "1x" links (or multiple 1x links on interfaces) through 2006.

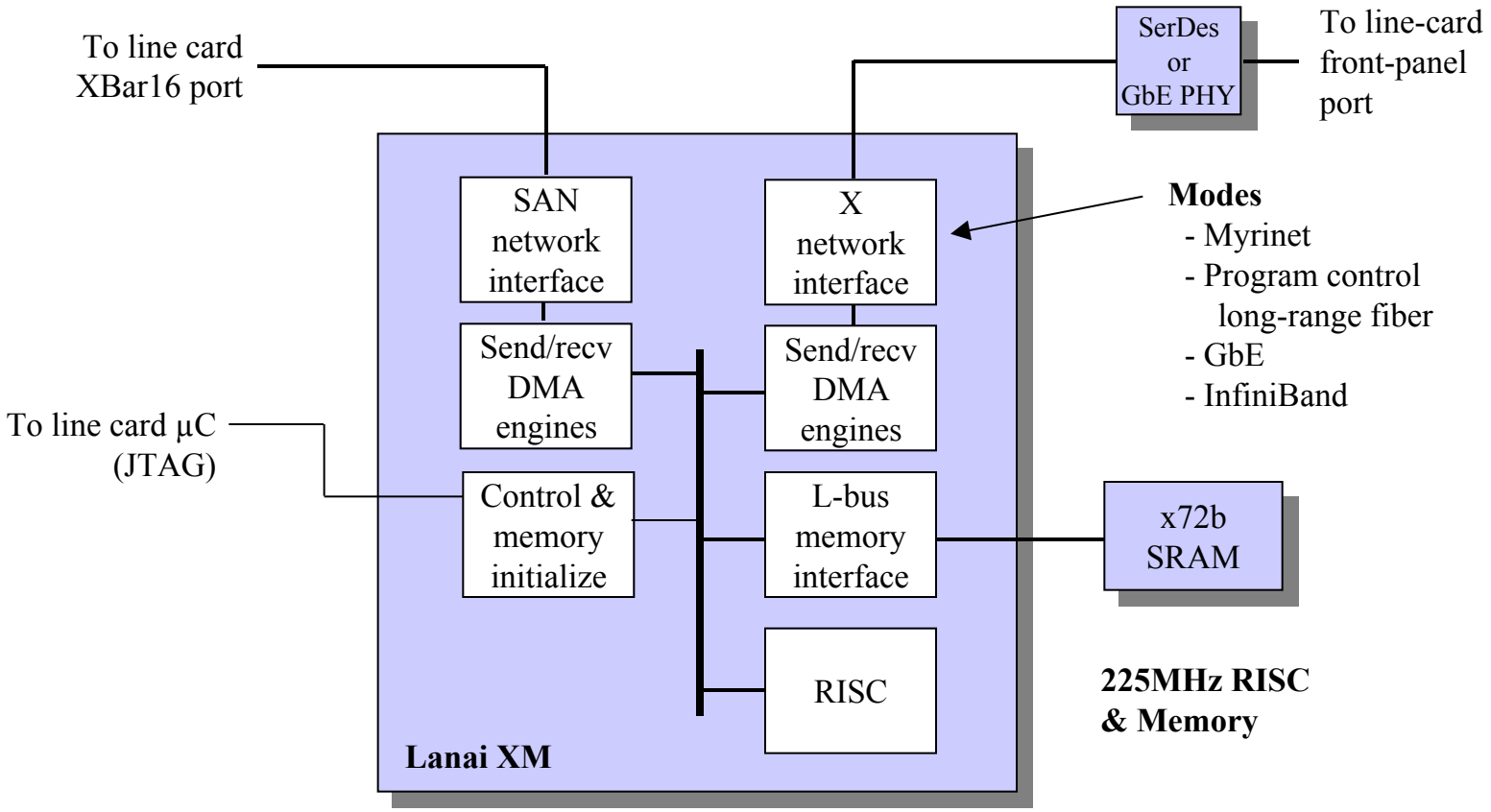
Prototype multi-protocol switch line card



Firmware-development prototype (M3-SW16-2E2X4F) of a switch line card with 2 GbE ports, 2 "X" ports over 850nm fiber, and 4 Myrinet ports.

The Physical-level and protocol conversion between the Myrinet switch on the line card and the GbE or "X" ports on the front panel is performed by a Lanai XM chip on each of these ports. The protocol-conversion firmware – *e.g.,* between native GbE and ethernet over Myrinet – for each Lanai XM is loaded during the line-card initialization process. The Lanai XM is the first in the family of Lanai X (10) chips.

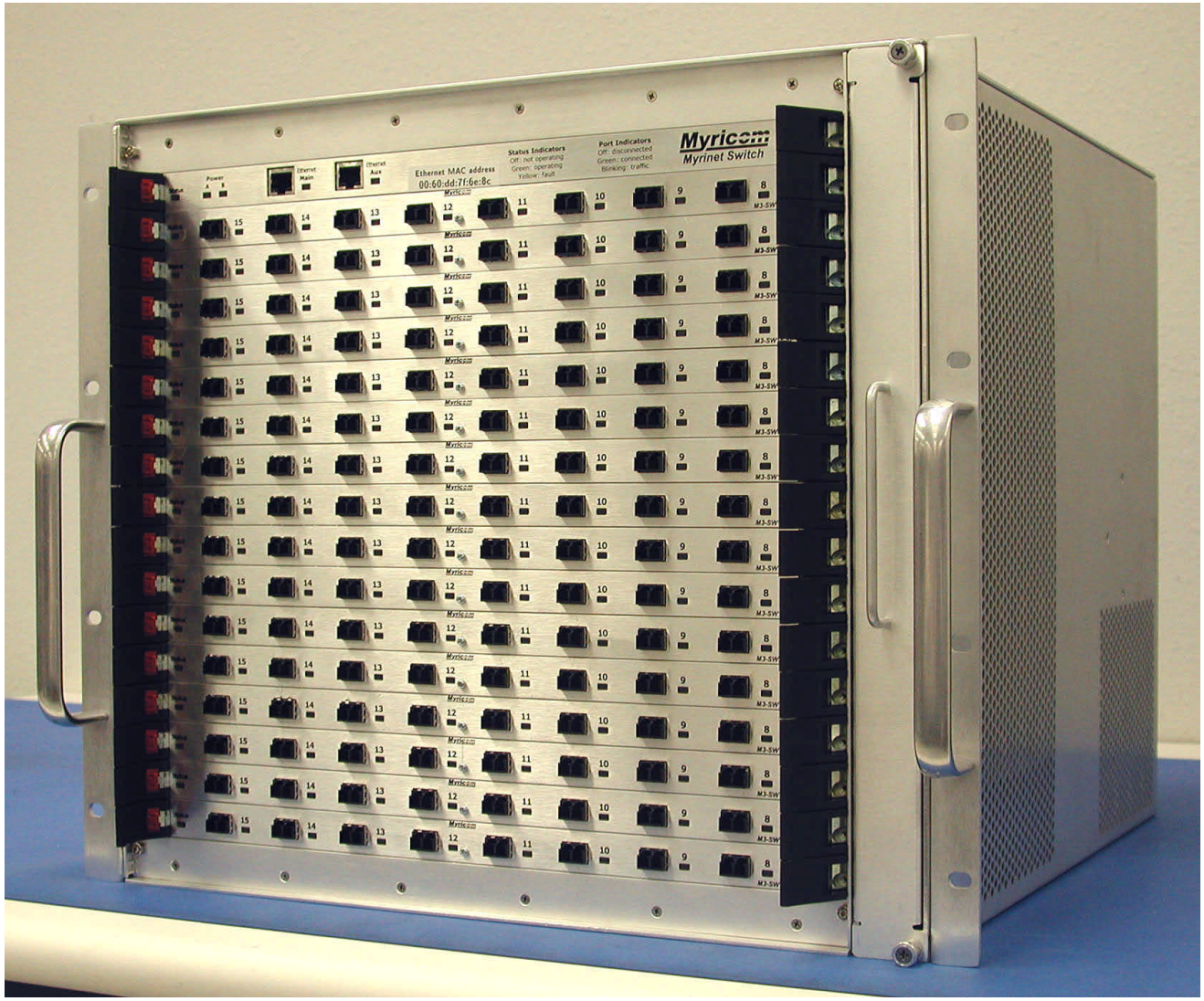
Lanai XM as a protocol converter



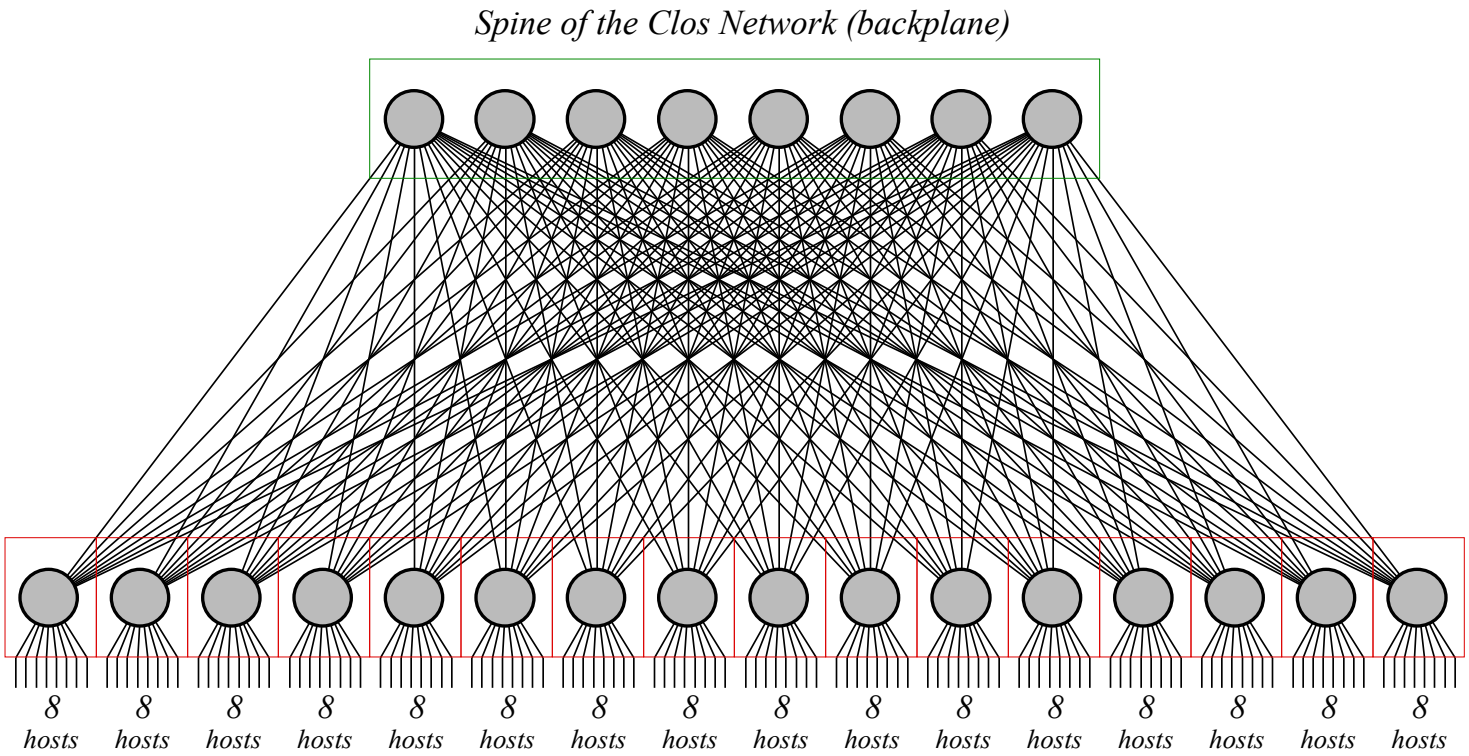
This circuitry is repeated for each "special" line-card port.

16-Port Myrinet Switch

128-Host Clos Network



Myrinet Switches & Switch Networks

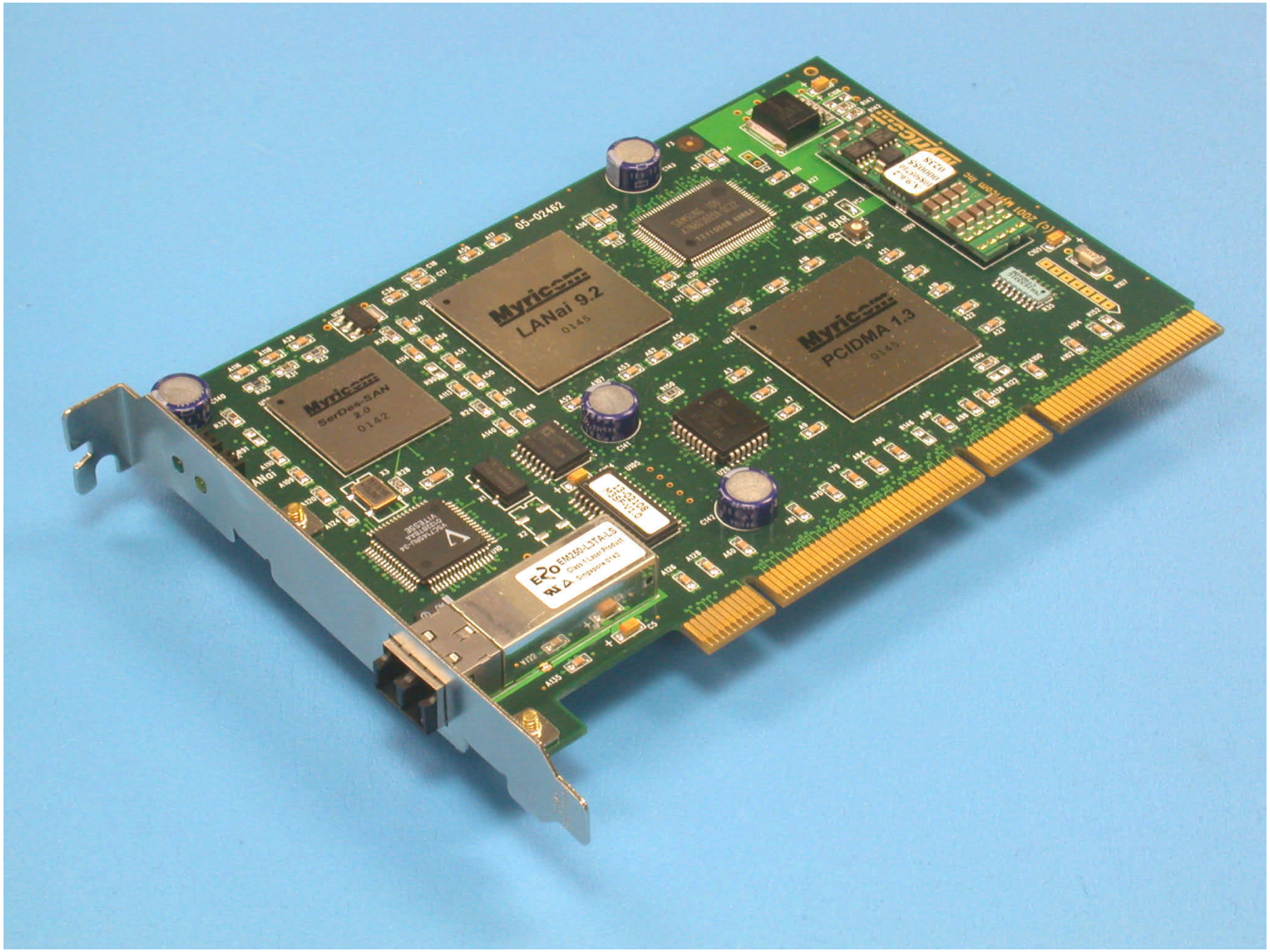


This family of products support hot-plugging of line cards and fans. Microcomputer monitoring (SNMP and a web server accessed through an Ethernet port on the monitoring line card) provides extensive monitoring and diagnostic capabilities, and management features needed for high-availability applications. See www.myri.com/myrinet/m3switch/guide/ for a full exposition.

Switches: Changes Planned

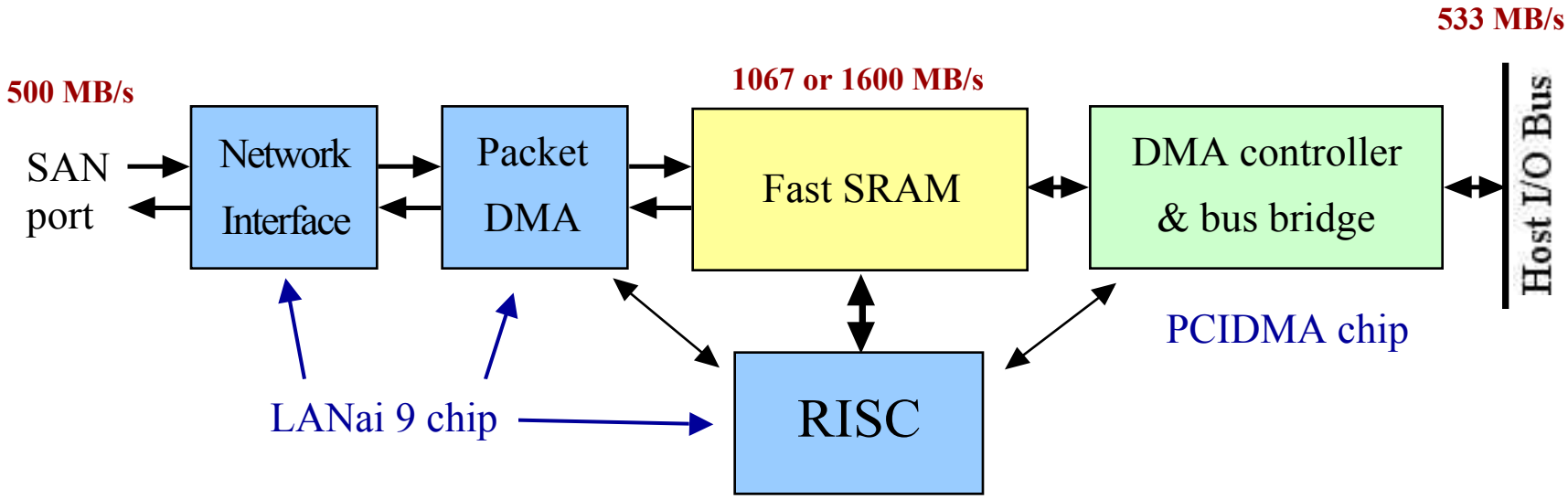
- Switches with a mix of Myrinet, long-range-Myrinet, GbE *(starting 1Q03)*, and *(possibly)* InfiniBand ports.
  - Convenient, low-cost interoperability.
  - High-degree switches with GbE ports may find a market for "Beowulf" clusters that use new-generation hosts with GbE on the motherboard.
- The use of multi-path (dispersive) routing (GM 2) allows better utilization of Myrinet Clos networks (also HA at a finer time scale).
- More capable monitoring line card that can run Linux *(1Q03)*.
- Late-2003 introduction of switches based on an XBar32.
  - "Clos256+256" switch in a 14U (?) rack-mount enclosure.
- Switches with "4x Myrinet" links *(2004)*.
- Switch pricing (list price of ~$400/port) is expected to remain unchanged through the end of 2003.

Interfaces: Current production M3F-PCI64C-2



64-bit, 66MHz, Myrinet/PCI interfaces

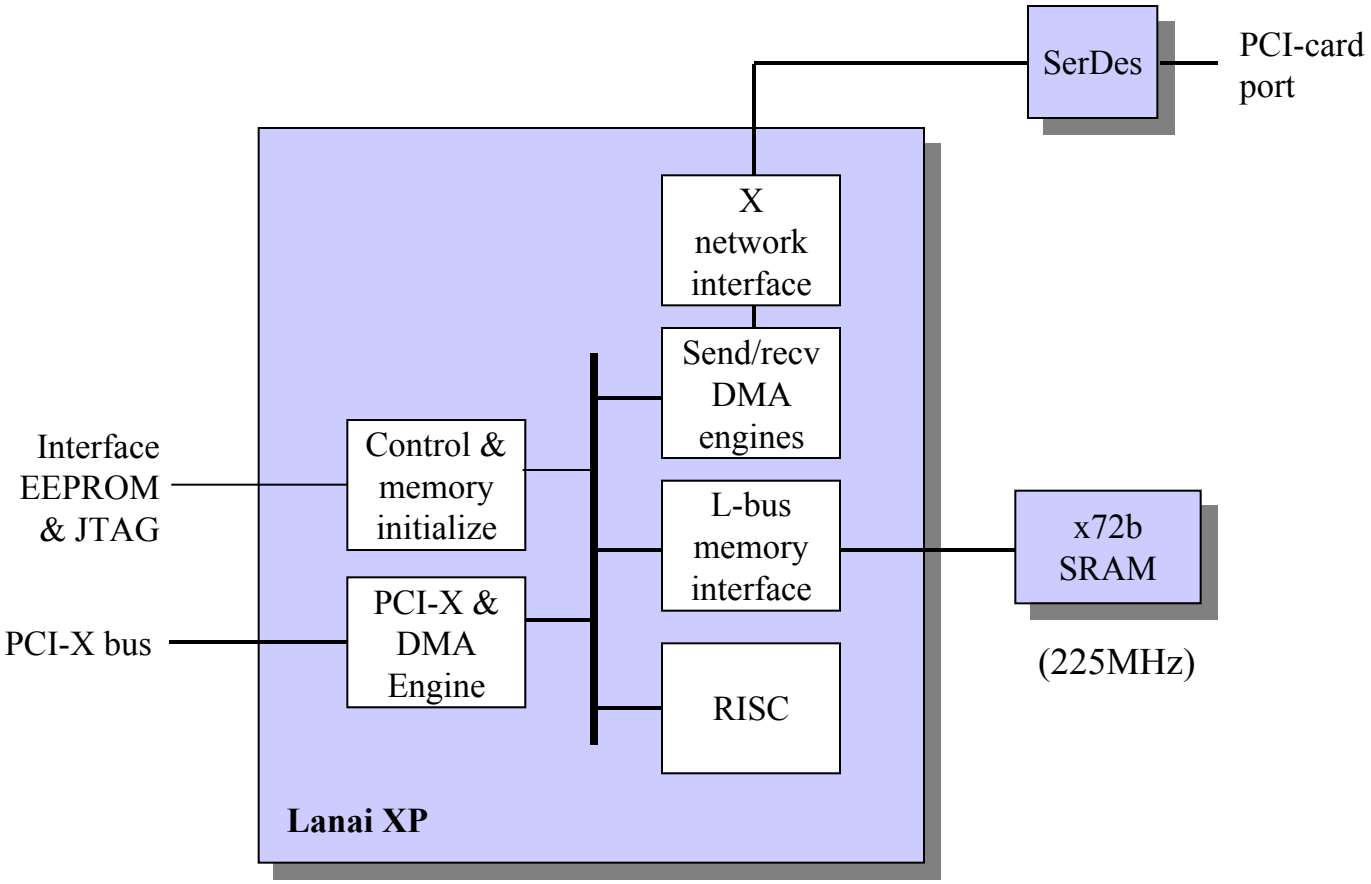
- PCI64B, 133MHz RISC and memory
  - 1067 MB/s memory bandwidth
- PCI64C, 200MHz RISC and memory
  - 1600 MB/s memory bandwidth



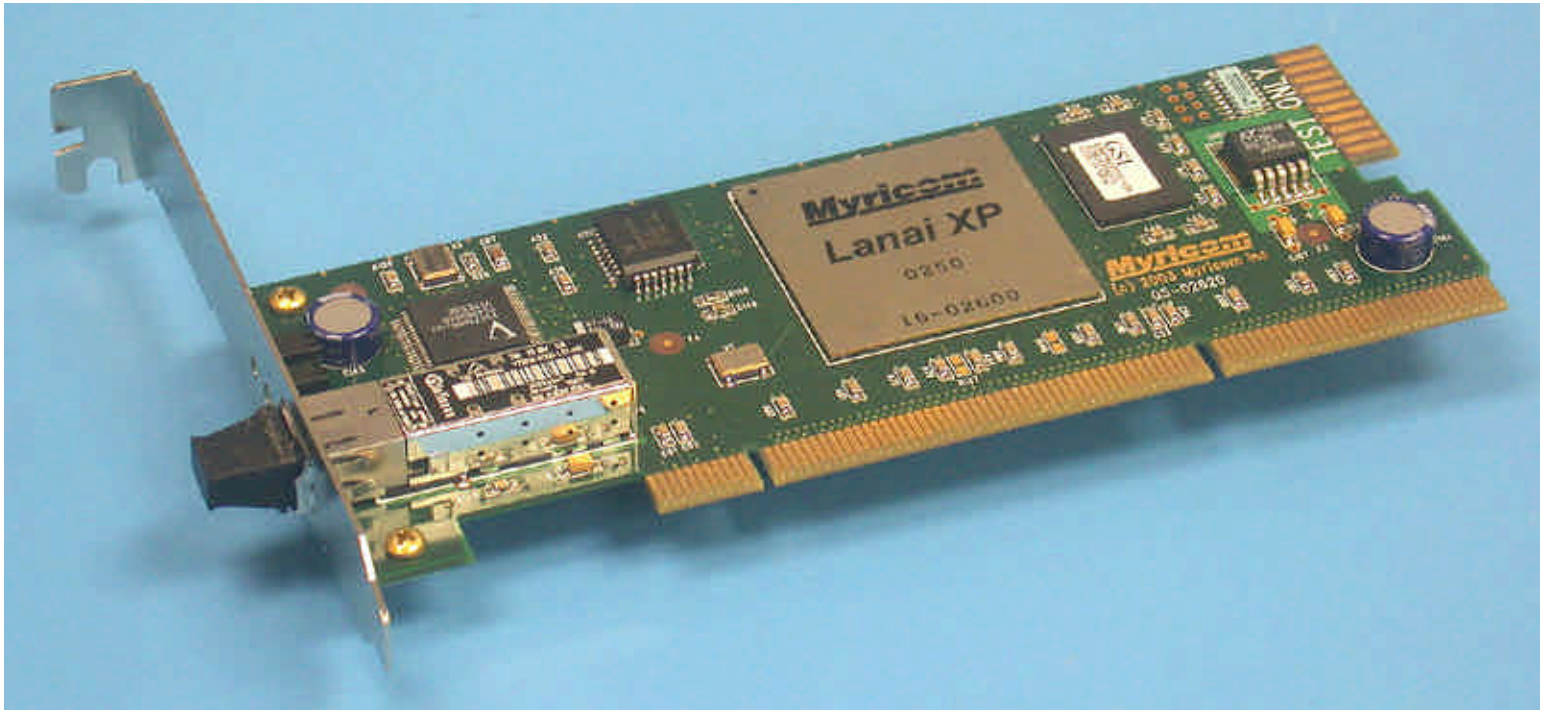
Interfaces: changes in progress and planned

- Myrinet/PCI-X interfaces with several new features
  - Initially *(1Q03):* 225MHz x 8B memory & 225MHz RISC; one port (Lanai XP)
    - ~5.5µs GM latency (*vs* ~7µs today). MPI, VI, etc, latency will decrease correspondingly.
    - Higher level of integration, and the starting point for low-cost Myrinet/PCI-X interfaces.
      - Pricing of low-end Myrinet/PCI-X interfaces is expected to decrease to ~$595 by early 2004.
  - Then *(2Q03):* 333MHz x 8B memory & 333MHz RISC; two ports (Lanai 2XP)
    - ~4.8µs GM latency.
    - The starting point for high-end Myrinet/PCI-X interfaces.
  - Self initializing
    - Simplified initialization; self-test capability.
    - Allows diskless booting through the Myrinet.
    - Blade and other high-density applications.
  - PCI-X series of interfaces will require GM 2.
- Myrinet/PCI-Express interfaces in early 2004
- Lanai with 4 Myrinet ports (or 4x port) in 2004
  - Open-ended performance growth to 600^{+}MHz RISC and memory.

Lanai XP PCI-X Interface



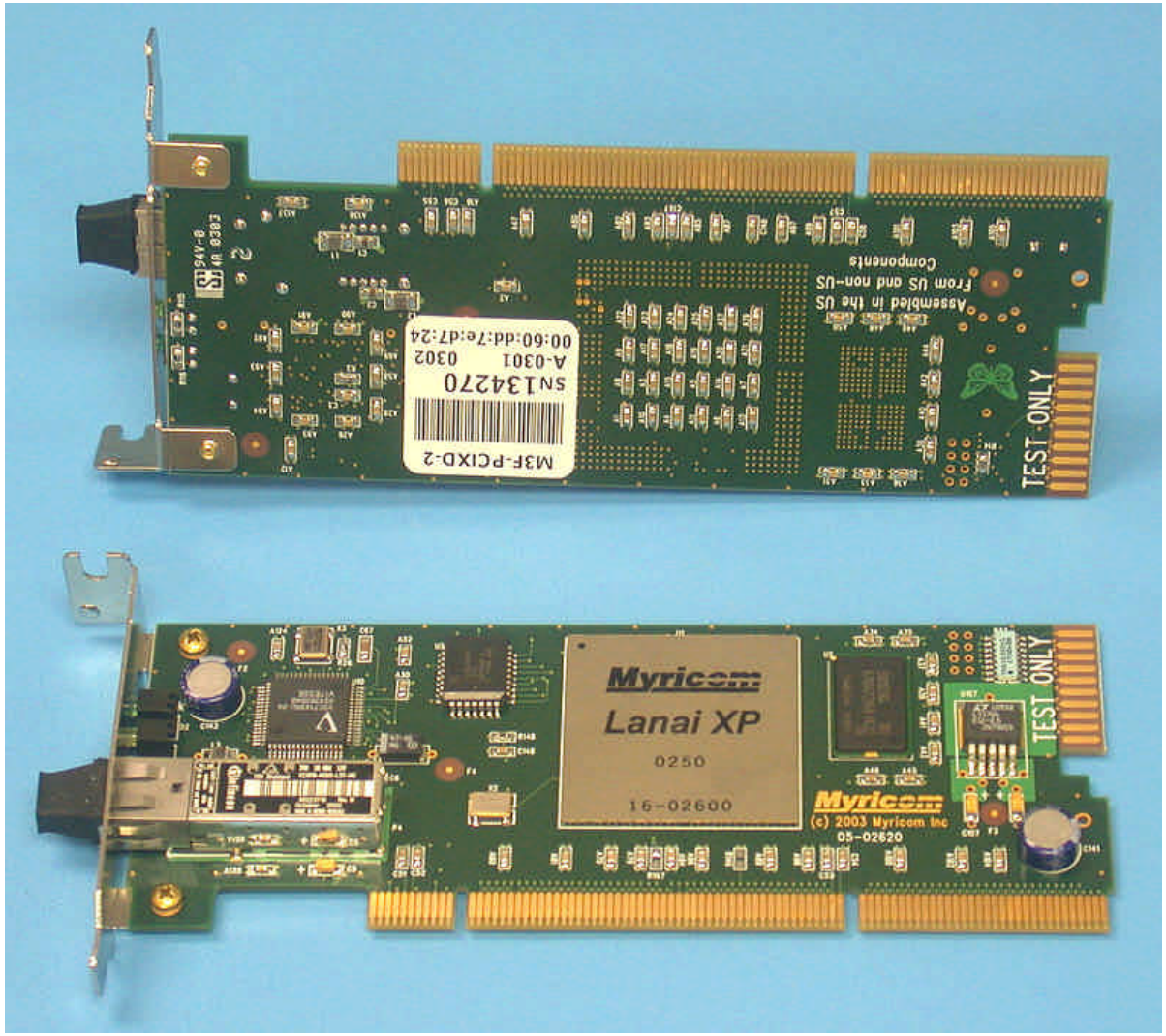
Lanai-XP-based Interface (M3F-PCIXD-2)



"Low Profile" PCI short card. PCI-X & PCI, 3.3V only.
Myricom will also produce this interface in other form factors.

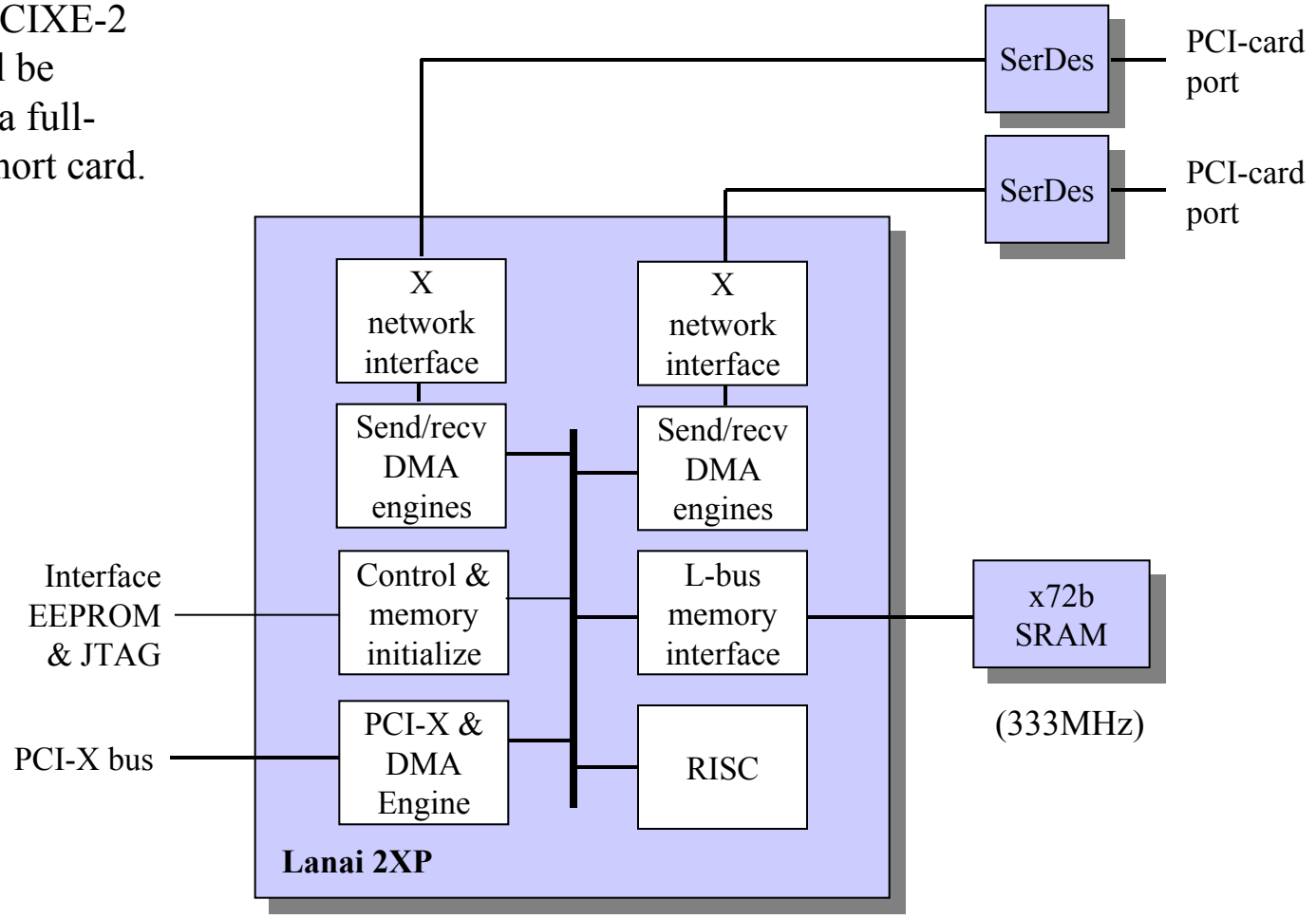
On a dual-2.4GHz Xeon with the Serverworks chip set:
PCI-X DMA read: 859 MB/s PCI-X DMA write: 1054 MB/s

Another view of (2) M3F-PCIXD-2 interfaces

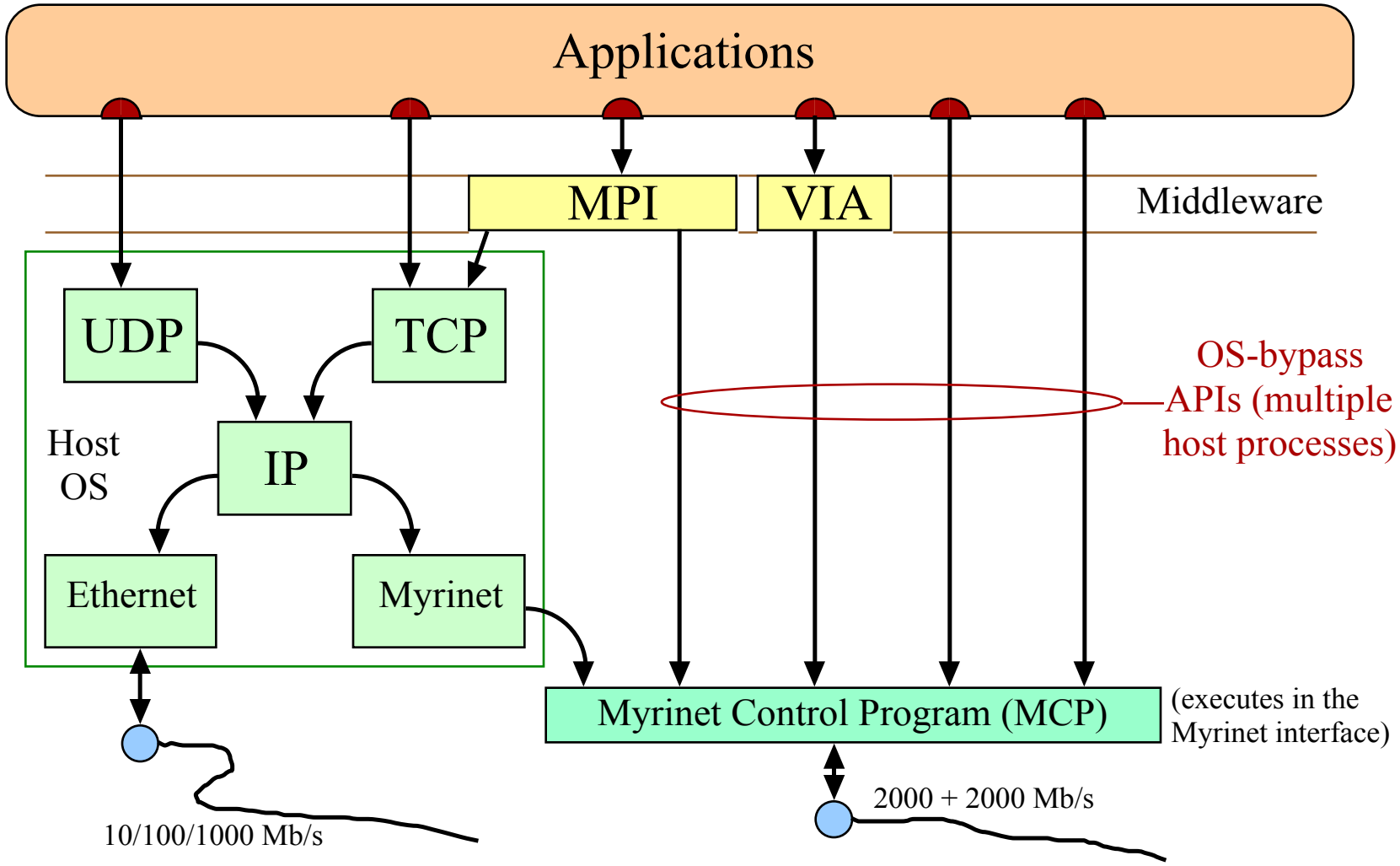


High-End Lanai 2XP PCI-X Interface

The M3F2-PCIXE-2 interface will be packaged as a full-height PCI short card.



Myrinet Software: Basic OS-Bypass Structure

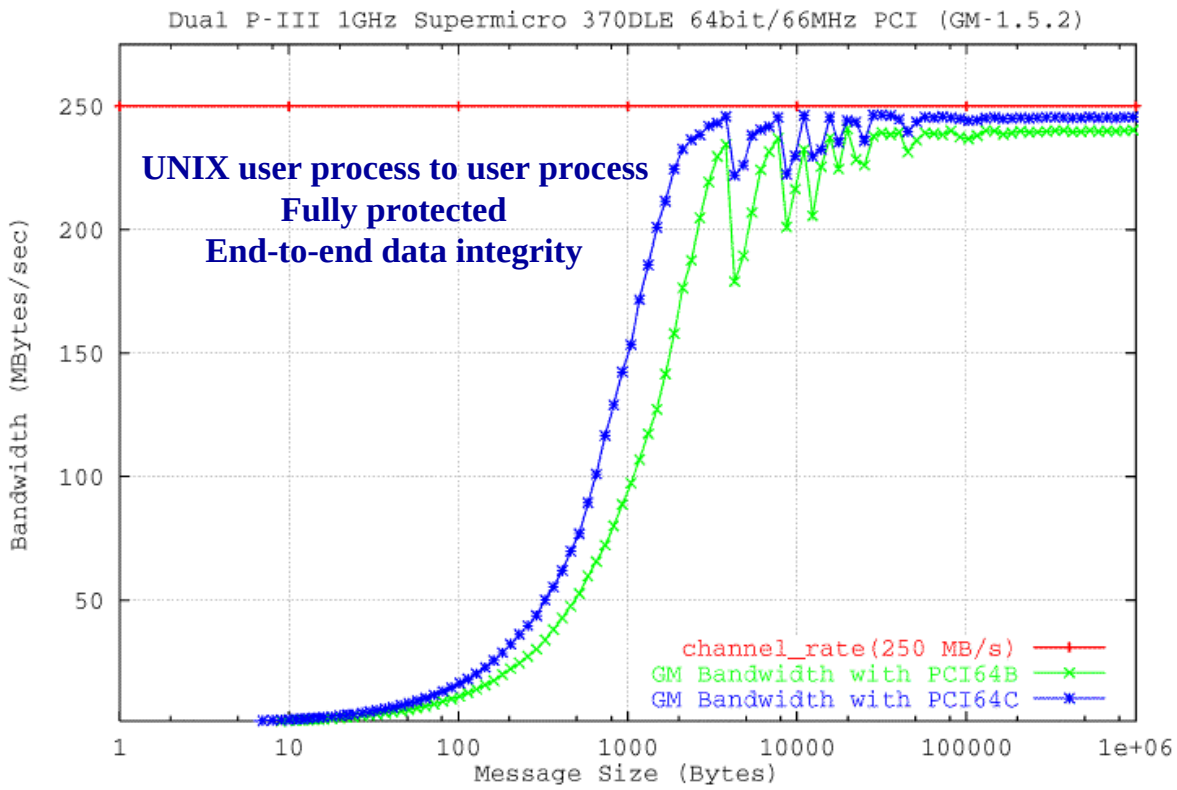


The GM Message-Passing System

No Compromises

- Concurrent, protected, user-level access
- Reliable, ordered message delivery
- Very low CPU overhead
- Robust under network faults
- Mapping
- Segmentation and reassembly of long messages
- High-level flow control
- "Clean" API, with exception handling
- Zero-copy layering of other APIs

GM Data-Rate Performance (Myrinet-2000 Fiber Interfaces)



GM short-message latency (Myrinet-2000 interfaces)
~ 7µs (PCI64C) or ~9µs (PCI64B)
GM CPU overhead < 1µs per message (LogP)

Current GM Software Distributions

OS	Platforms
Linux	IA-32, IA-64, Alpha, PowerPC, IBM Power 3 & 4
Win2000/XP	IA-32, IA-64
Solaris	UltraSPARC
Tru64	Alpha
AIX	IBM Power
Irix *	MIPS
VxWorks *	PowerPC
MacOS X	Apple Macintosh G4
FreeBSD, …	IA-32 & Alpha

Current Choice of Myrinet Software Interfaces

- The GM API
  - Low level, but some applications are programmed at this level
- TCP/IP
  - Actually, "ethernet emulation," included in all GM releases
    - 1.9 Gb/s TCP/IP (netperf benchmarks)
- MPICH-GM
  - An implementation of the Argonne MPICH directly over GM
- VI-GM
  - An implementation of the VI Architecture API directly over GM
- Sockets-GM
  - An implementation of UNIX or Windows sockets (or DCOM) over GM. Completely transparent to application programs. Use the same binaries!
- PVM
- A variety of software packages developed and supported by third parties.
  - SCore is the most widely used.

Myrinet Software: Changes Planned

- GM 2, a restructuring of the GM control program
  - *Alpha release available now.*
  - Slightly extended API, including the **`gm_get`** function.
  - Improvements in mapping and in multi-path traffic dispersion to take better advantage of the Clos networks, and to use multiple-port NICs.
  - Interrupt coalescing in ethernet emulation to reduce host-CPU utilization.
  - Thanks to more flexible buffer management, certain MPI functions can be supported at a lower level.

- Increasing emphasis on Myrinet interoperability with GbE

- Myrinet Express (MX): Specialized firmware and software to support MPI and Ethernet Emulation. 3.5µs (?) MPI latency with PCIXE interfaces.
  - *3Q03 release.* Linux, PCIXD & PCIXE interfaces only.
    - Not a replacement for GM.
  - Designed for MPI to be able to use multiple interfaces.

Technology Roadmap beyond 2004

- 1 GB/s on single fibers
  - 10 GBaud VCSEL transceivers.
  - Requires 1 GB/s ports on switches and interfaces.
    - Entirely new chip set.

- New internal organization for Myrinet crossbar-switch chips
  - Arbitration on a token ring outside of the crossbar.

- Compile interface firmware more fully into silicon
  - Buffer memory entirely within the interface chip.