Cluster Interconnect Installation & Troubleshooting

Susan Blackford
Technical Support
Myricom, Inc.
help@myri.com

Myricom, Inc. 325 N. Santa Anita Ave. Arcadia CA 91006
626-821-5555 Fax: 626-821-5316 http:/ /www.myri.comv

Outline

- Introduction - DST 5B
- Hardware Overview
- Hardware Installation and Troubleshooting
- Software Overview
- GM Software Installation, Troubleshooting and Validation
- Miscellaneous Troubleshooting Hints
- Diagnostic Tools
- Myrinet Technology Roadmap
- Conclusions

Myricom, Inc.

- **Myricom** created Myrinet
  - Myricom handles the development, manufacturing, sales, and technical support of our Myrinet products
  - A growing, profitable, technology company
    - Privately held California Corporation, founded in 1994
    - Completed our 39th consecutive profitable quarter in December 2004.
    - Not dependent on infusions of venture capital
  - Technical staff of 22 people, about half of whom have Ph.D. degrees.

Myricom Technical Support

- The best way to contact Myricom Technical Support is via email, **help@myri.com**.
- Normal business hours are 8am-5pm PST, 626-821-5555. You can telephone if desired, but email is the best way to contact.
- We have tech support personnel on the East and West Coast of the United States, who also address **help@myri.com** requests in the evenings and on weekends.
- All customer correspondence via **help@myri.com** is read by Myricom Management. Thus, if there is a critical situation or escalation, they are fully aware of the status at all times.
- We have a standard 3-year warranty with IBM on all Myrinet hardware. We can cross-ship replacements for RMAs.

Standard Procedure for an RMA

- Send RMA request to Myricom Technical Support, **help@myri.com**.
- Provide the following information:
  - Myricom Product Code (e.g., M3F-PCIXD-2) and Serial number (5-digit or 6-digit number), located on a white sticker on the backside of the component.

  - Error description
  - Return Address (including contact name and phone number)
- Feel free to request a cross-ship replacement if needed.

What is Myrinet?

- ***Market view:*** The clear **market leader** for low-latency interconnect for computing **clusters.**
  - Myrinet is used in many thousands of clusters in more than 50 countries.

- ***Customer view:*** A set of **standard products**
  - Network-interface cards (NICs), software, switches, and cables.
  - All you need to make a high-performance cluster from a set of host computers.

- ***Designer view:*** A network architecture, protocol, and **technology**
  - A descendant of packet communication and routing in MPPs, but open.
  - ANSI Standard (ANSI/VITA 26-1998).
  - The mirror image of Ethernet: Processing power is concentrated in the hosts and NICs, allowing an elegant, streamlined, switching technology.

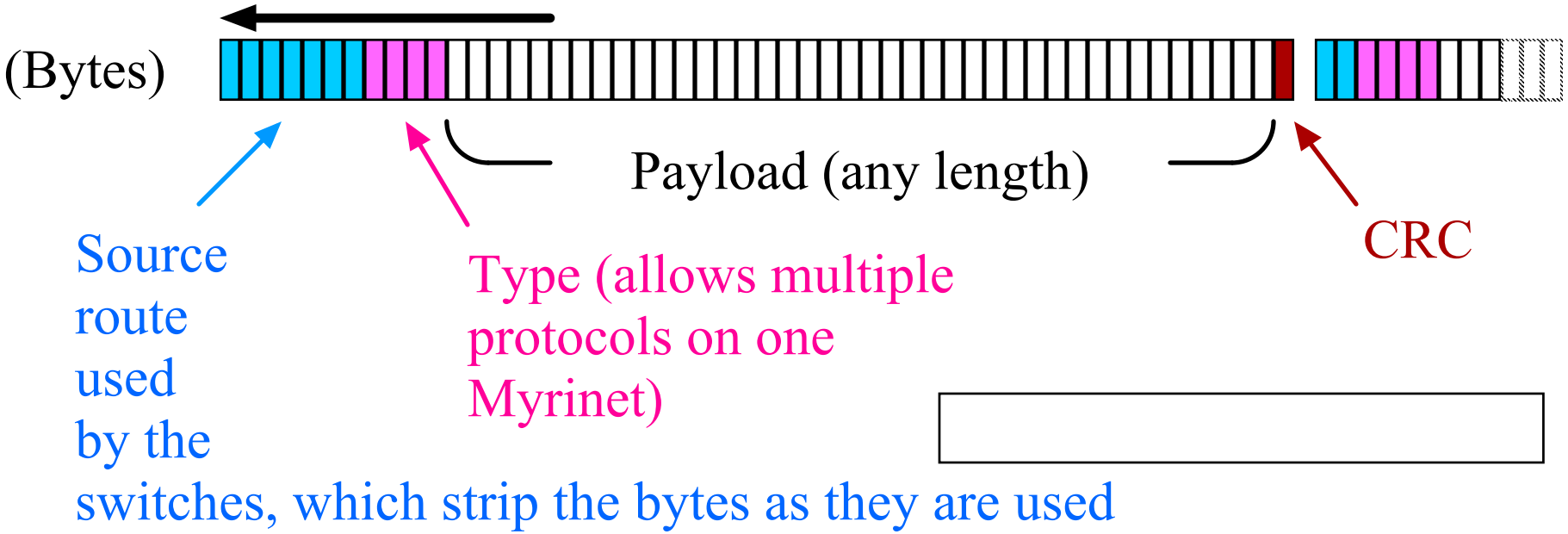
Hardware Overview

Designer View

- A network architecture, protocol, and **technology**
  - A descendant of packet communication and routing in MPPs (Massively Parallel Processors), but open.
  - ANSI Standard (ANSI/VITA 26-1998).
  - **The mirror image of Ethernet:** With Ethernet, the NICs are simple and the switches are complex. With Myrinet, processing power is concentrated in the hosts and NICs, allowing an elegant, streamlined, switching technology.
- A technology with a very wide range of application.
  - *From chips and circuit boards to clusters that fill rooms or buildings.*

Myrinet = ANSI/VITA 26-1998

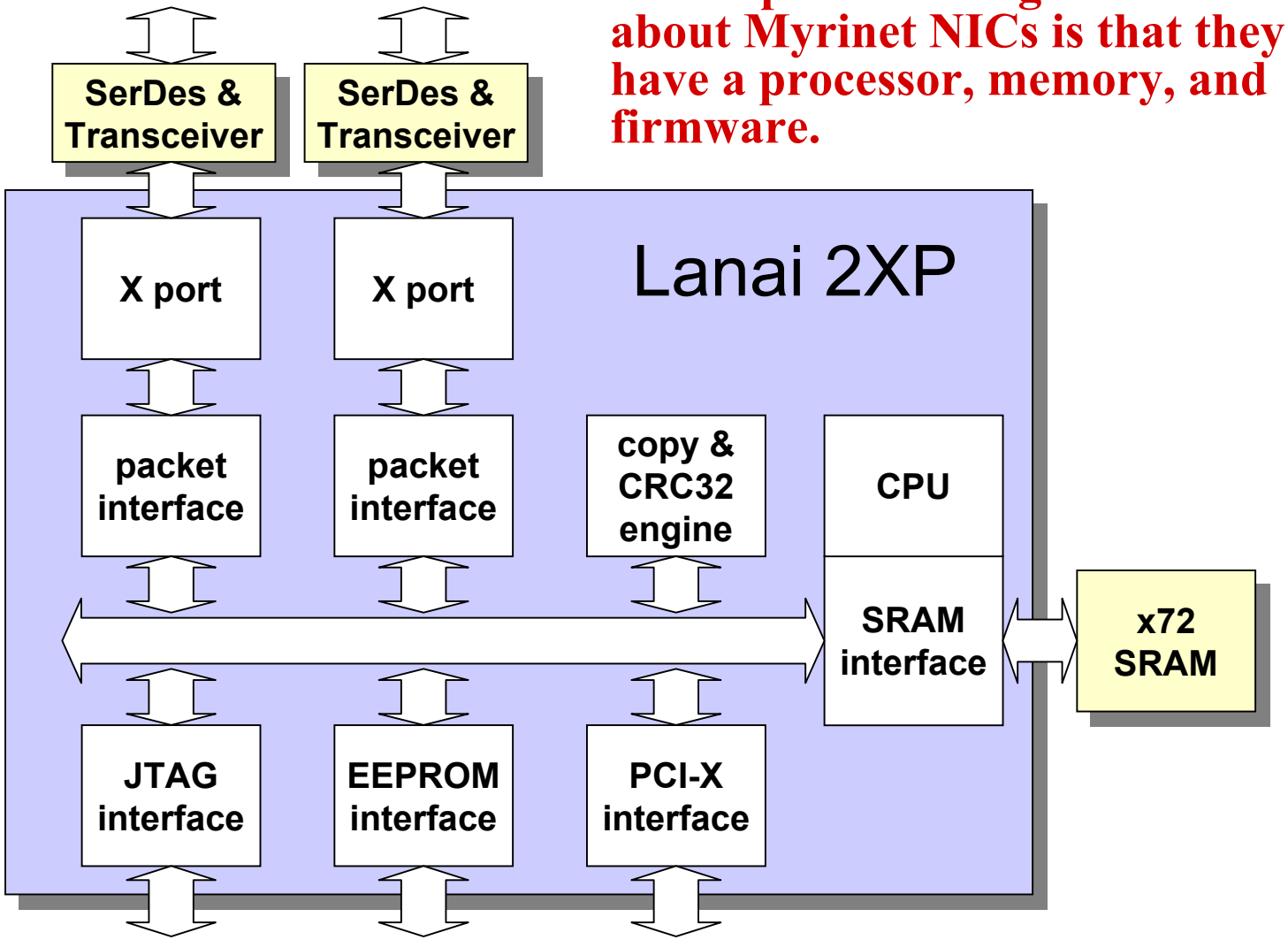
Myrinet is defined at the Data-Link level (level 2 of the ISO reference model for computer networks) by its packet format and flow control.



There is flow-control on every link.

Myrinet NICs = Protocol Offload Engines

The important thing to remember about Myrinet NICs is that they have a processor, memory, and firmware.



One more thing about Myrinet NICs…

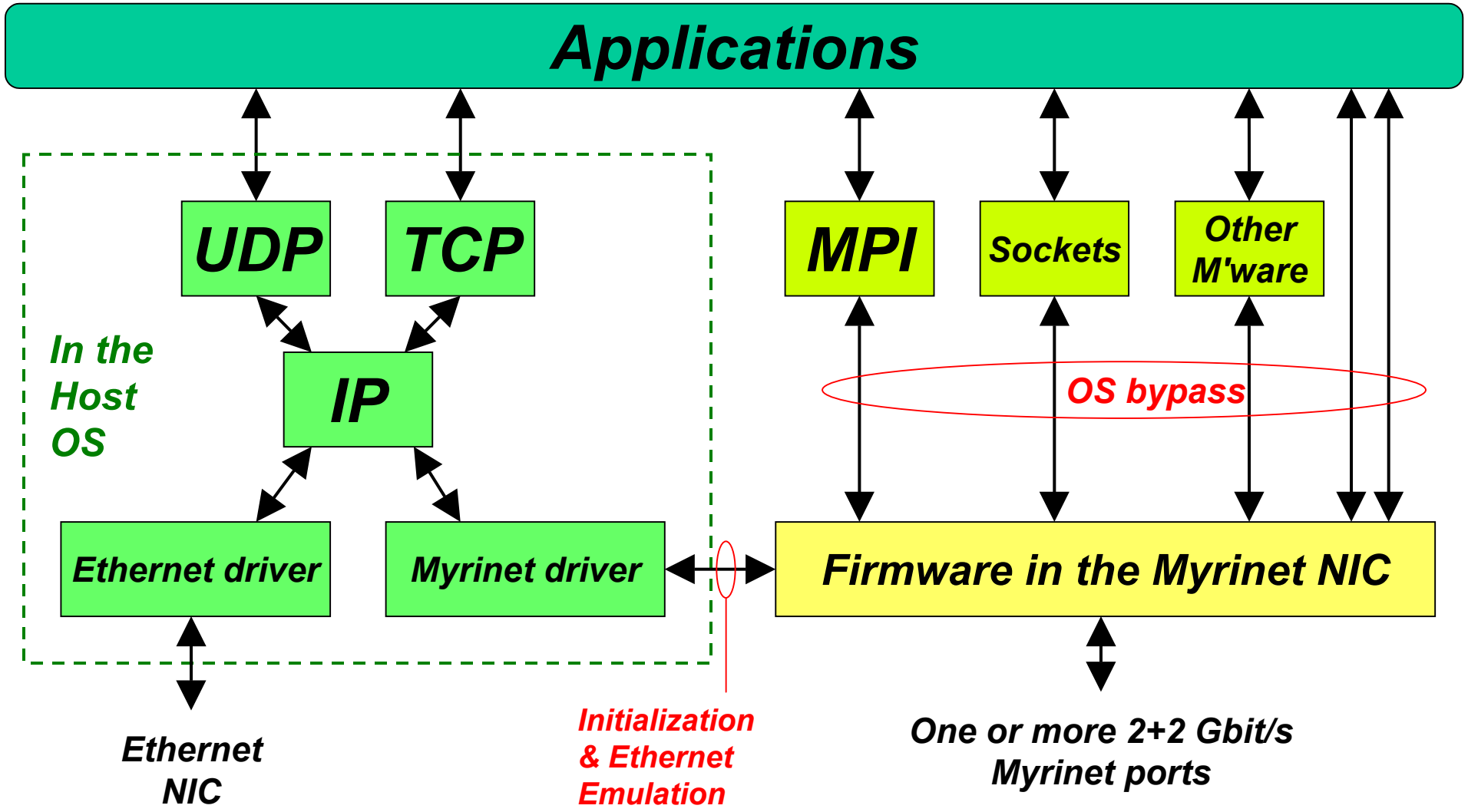
A Myrinet NIC looks from the standpoint of installation exactly like an Ethernet NIC

An Ethernet MAC address

Myrinet NIC product label

The Myrinet device driver advertises itself to the host operating system as an Ethernet driver

Myrinet Software Interfaces



Myricom, Inc. 325 N. Santa Anita Ave. Arcadia CA 91006
626-821-5555 Fax: 626-821-5316 http:/ /www.myri.comv

Why does Myrinet work so well for clusters? (1)

- The "Processor & Firmware in the NIC" Architecture
  - *Versus Ethernet (or IB): Simple or RDMA NICs …*
    - *… depend upon the host to handle low-level network operations*
  - Myrinet NICs are firmware-driven offload engines
  - The NIC processing supports OS-Bypass operation
    - Message operations without system calls, resulting in low latency and low-host-CPU overhead
  - The NIC processing provides type matching
    - Immediate, first-level demultiplexing of incoming messages, resulting in efficient use of IO bandwidth, host CPU, and host memory (no "RDMA window" memory use)
  - The NIC processing handles network protocols
    - Mapping, dispersive routing, reliability layer (reliable ordered delivery with acknowledgments), and exception handling.

Why does Myrinet work so well for clusters? (2)

- Source Routing
  - *Versus Ethernet: Destination Routing, in which …*
    - *… the switch must decide how to route a packet to a destination*
      - and the switch is limited to information local to this switch
    - *… switching is typically store-and-forward*
  - Myrinet Switches are based on simple, high-degree, crossbar switches
    - Inexpensive, low latency, and highly scalable
    - The route is predetermined at the source: the switch just steers packets
      - The network end points have global information, and can be much "smarter" about routing (e.g., dispersive routing) than any single switch
    - Cut-through routing -- low latency even through many switch "hops"

- Note: Myrinet and Quadrics both employ these two techniques: a processor and firmware in the NIC, and source routing
  - The preferred and proven architecture for scalable HPC clusters

Myrinet Summary

- **Low latency**
  - 2.6µs (MPI user level) for the fastest NICs, or
  - 3.5µs for the lowest-cost NICs
- **High data rate**
  - 2+2 Gb/s (250+250 MB/s) data rate links; user level is 95-99% of peak.
  - For higher data rates, use one or more dual-port NICs.
- **Very low host-CPU utilization**
  - Protocol processing is offloaded
  - logP < 0.3µs
- **Unlimited scalability**
  - Switch cost per host scales very well in the range 16 " N " 8192
- **Multimode-fiber links** to 200m
  - 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 APIs + TCP|UDP/IP + MPI + VI + PVM + Sockets + DAPL
- **Hybrid Myrinet/GbE networks**

Customer View: What Hardware Do I Need to Install?

- Myrinet-2000 network consists of :
  - Myrinet-2000 PCI-X Network Interface Cards (NICs)
  - Myrinet-2000 switch(es)
  - Myrinet-2000 fiber cables
- Full product specifications are available:
  - http://www.myri.com/myrinet/product_list.html

What Hardware Do I Need to Install?

- A Myrinet network for the IBM BladeCenter consists of :
  - A PCI-X Host Card Adapter (HCA) in each blade. Most Myricom documentation refers to an “HCA” as an “interface” or “NIC”.
  - Optical Passthru Module (OPM) – converts midplane electrical signaling to fiber.
  - External Myrinet switch(es) – LC-connectorized fiber ports.
  - Myrinet fiber cables – special ribbon-fiber cables are required between the OPM quad-fiber ports and the Myrinet switch(es).

Myrinet-2000 PCI-X Network Interface Cards (NICs)

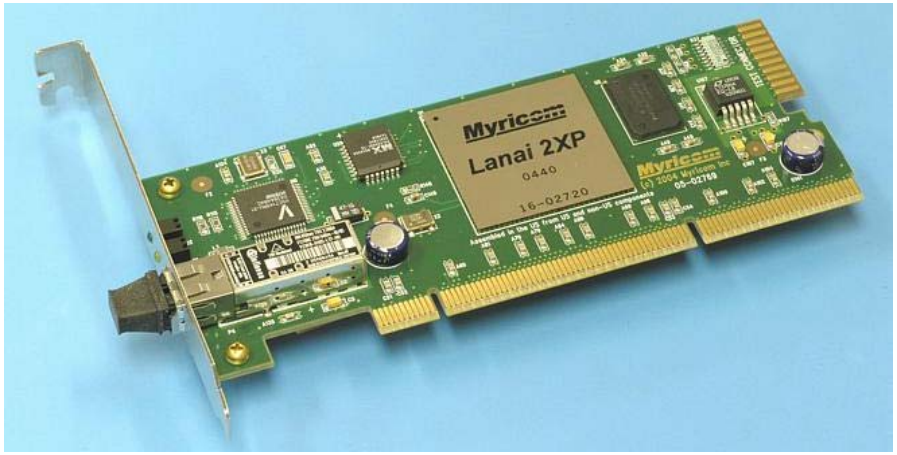
Myrinet-2000 PCI-X Network Interface Cards (NICs)

- Myrinet-2000 PCI-X NICs (64-bit 133MHz)

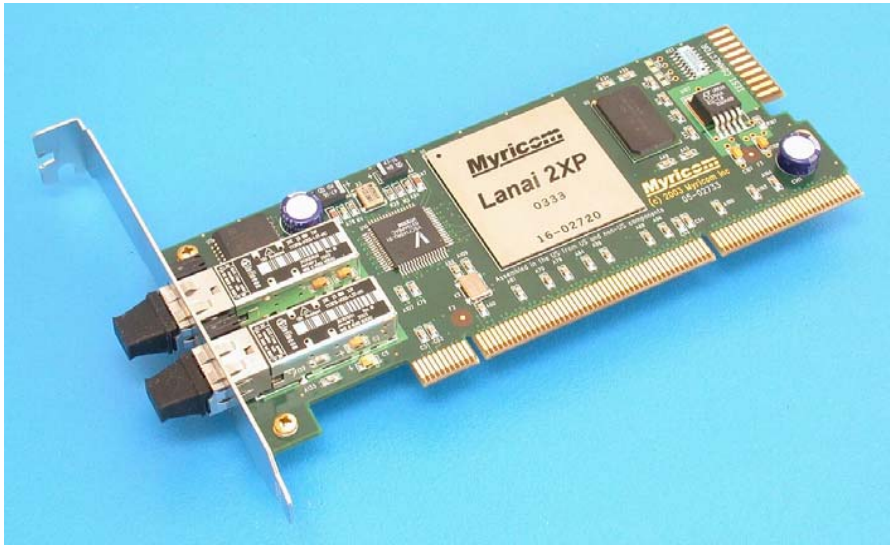
  - M3F-PCIXD-{2,4} 225MHz RISC and Memory
  - M3F2-PCIXE-4 333MHz RISC and Memory
  - M3F-PCIXF-{2,4} 333MHz RISC and Memory

- Low-profile PCI short card
- Available with the standard PCI faceplate or low-profile faceplate
- Available as special-form-factor versions of these NICs for use in blade systems.
- Also function correctly in 100MHz PCI-X slots or 3.3V PCI slots.
- Same installation regardless of 2M or 4M SRAM.
- Open-source software support is included.

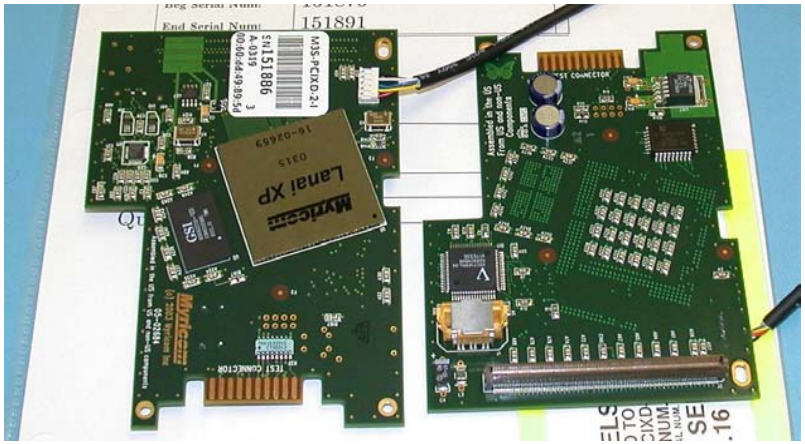
Current-production Myrinet/PCI-X NICs



One-port NICs:
"D card" (225MHz) & "F card" (333MHz)



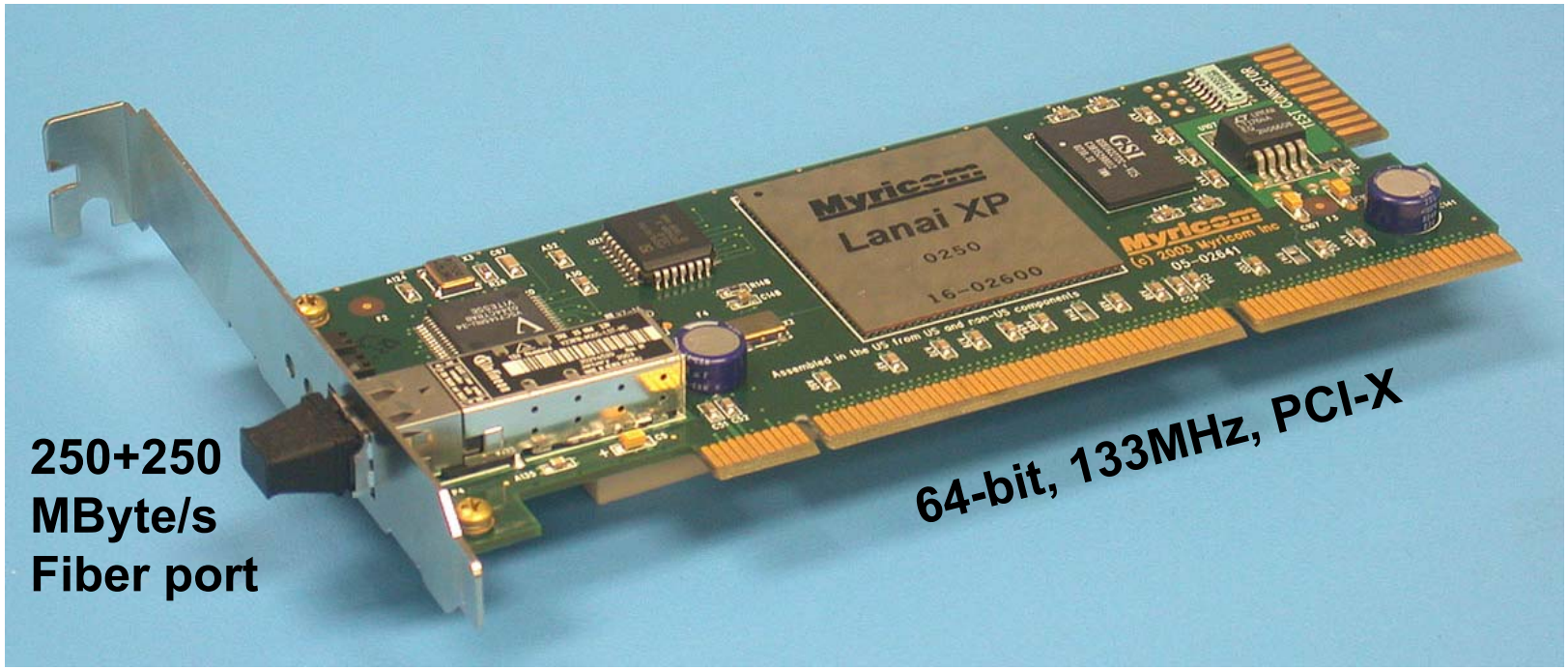
Two-port NIC: "E card" (333MHz)



IBM BladeCenter version of the D card

All PCIX-series NICs are compatible at the network, software, and API levels. PCI-X performance on a dual-2.4GHz Xeon with the Serverworks chip set: 932 MB/s read, 1044 MB/s write. These NICs are self-initializing both for convenience and to allow diskless booting.

M3F-PCIXD NIC



M3F-PCIXD-2 NIC ("D card")

"Low Profile" PCI short card. PCI-X & PCI, 3.3V only.
Self-initializing from EEPROM (convenience, diskless booting).

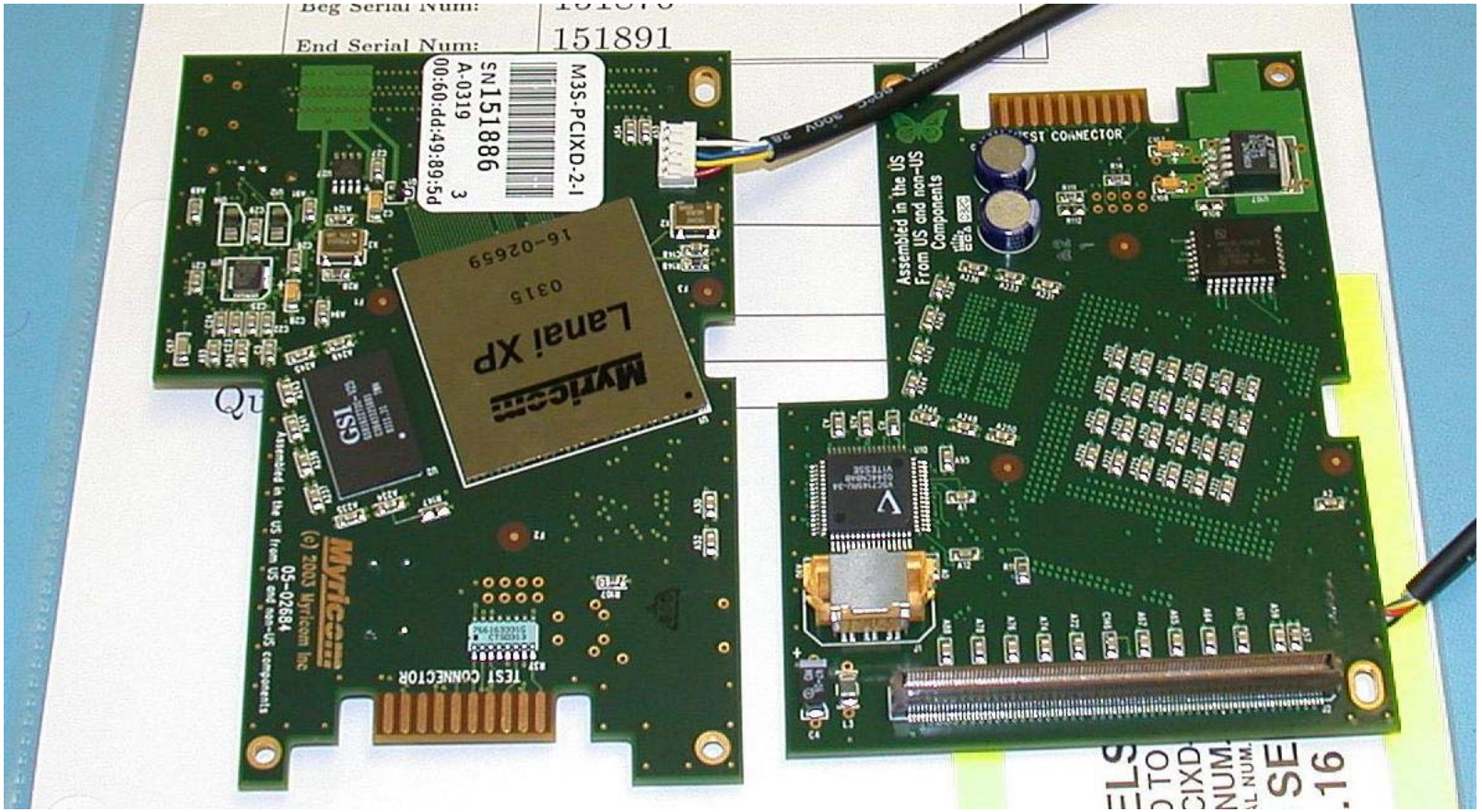
PCI-X performance on a dual-2.4GHz Xeon with the Serverworks chip set:
PCI-X DMA read: 932 MB/s PCI-X DMA write: 1044 MB/s

Myricom

Myricom, Inc. 325 N. Santa Anita Ave. Arcadia CA 91006
626-821-5555 Fax: 626-821-5316 http:/ /www.myri.comv

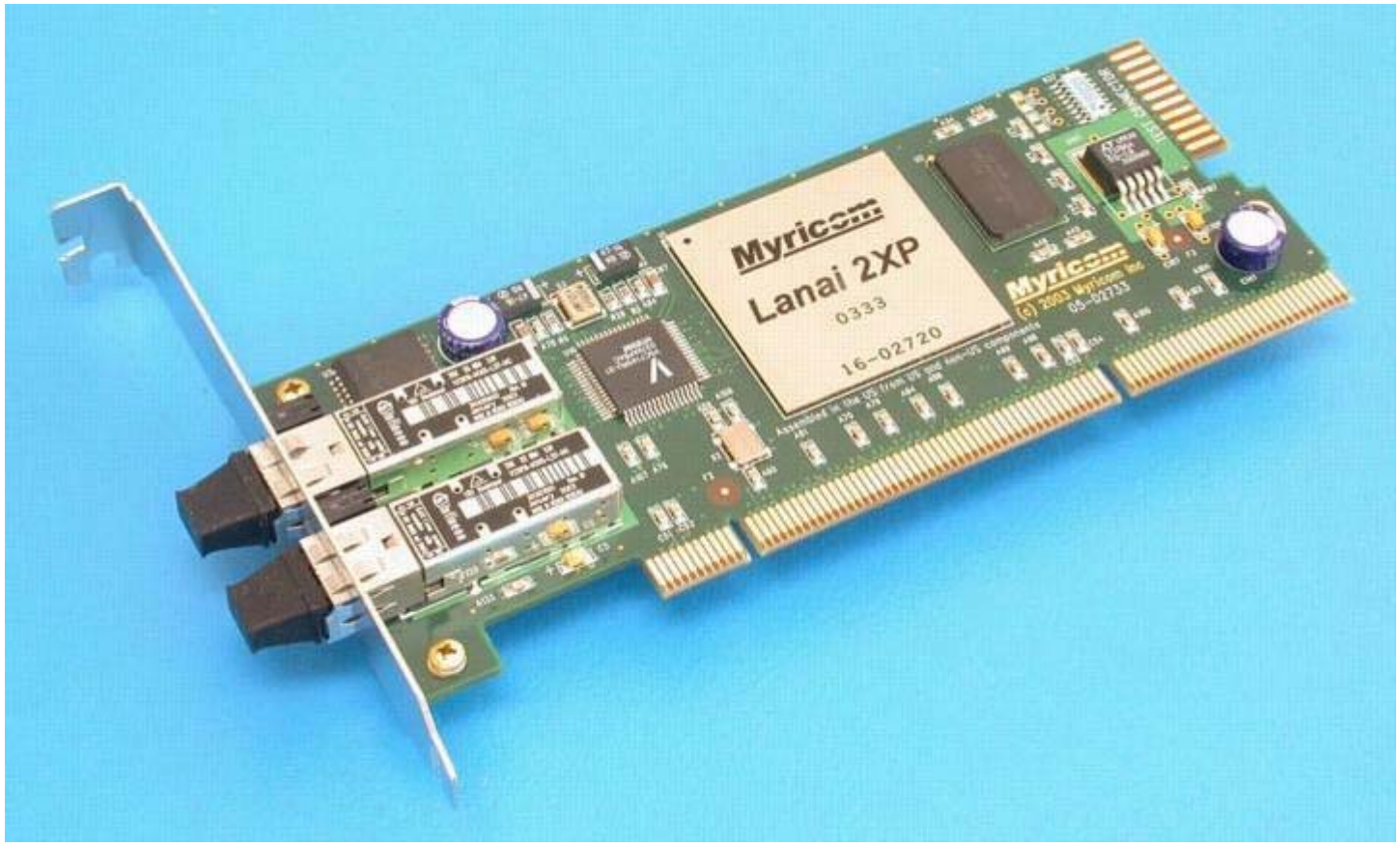
IBM BladeCenter version of the “D card”

(Photograph of 2 HCAs, to show both sides)



“Not a product announcement”

M3F2-PCIXE-2 NIC



M3F2-PCIXE-2 NIC ("E" card)

M3F-PCIXF NIC

M3F-PCIXF NIC ("F" card)

Myrinet-2000 Switches

Myrinet-2000 Switches

- Small Myrinet-2000 Switches
  - M3-E16 3-slot 1-2 port line cards plus monitoring card
- Intermediate-size Switch Networks
  - M3-E32 5-slot 1-4 port line cards plus monitoring card
  - M3-E64 9-slot 1-8 port line cards plus monitoring card
  - M3-E128 17-slot 1-16 port line cards plus monitoring card
- Switch Networks for Large Clusters (>= 128 hosts)
  - M3-CLOS-ENCL 21-slot 1-20 line cards plus monitoring card
  - M3-SPINE-ENCL 21-slot 1-20 line cards plus monitoring card

Myrinet-2000 Switches for Large Clusters

512 Ports

in the Clos256+256 configuration pictured, 256 host ports plus 256 inter-switch ports, **or**

256 Ports

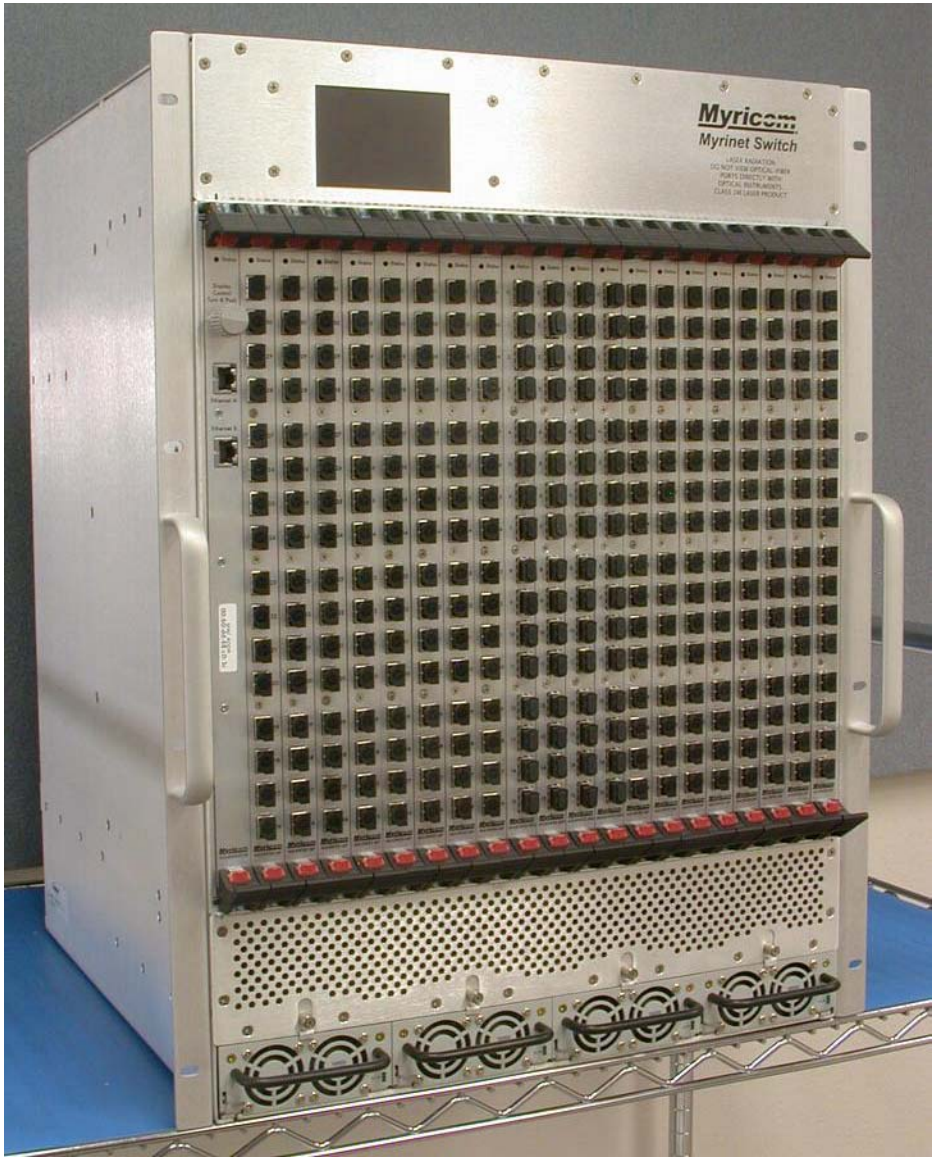
in the low-cost standalone Clos256 configuration, **or**

Up to 1280 Ports

in the Spine configuration used to connect Clos256+256 switches.

Full-bisection Clos networks.

All inter-switch links on quad-fiber ribbon cables.



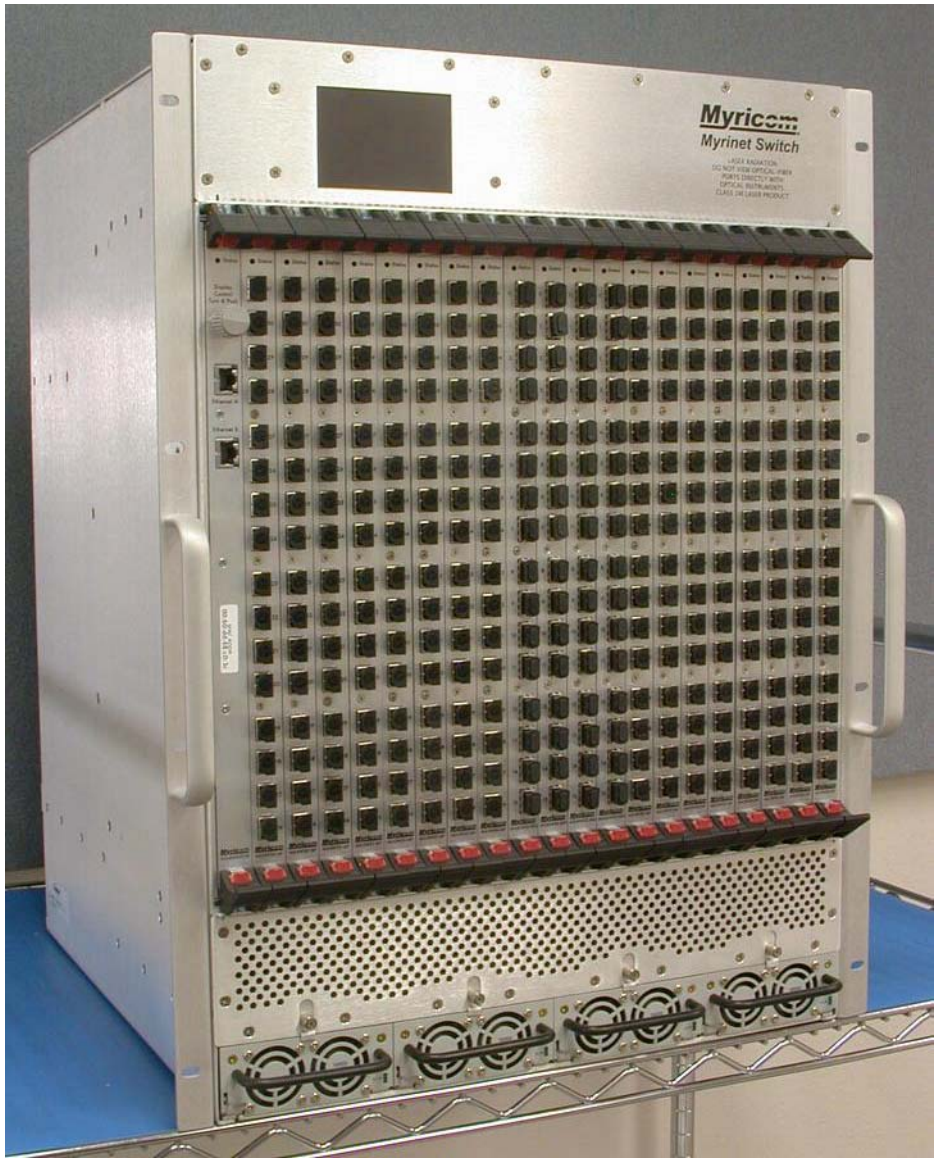
Myricom

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

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

Enterprise Features

- 14U rackmount with front to back airflow.
  - Separate air paths for power supplies and line cards
- Line cards are hot-swappable
- Four hot-swappable power supplies with N+1 redundancy
- Monitoring via 10/100 ethernet and a color TFT display:
  - Link continuity, packet traffic, packet counts, packet errors
  - Internal voltages, currents, temperatures, blower rotation, power-supply function
- Four blowers in rear, two series pairs, with N+2 redundancy

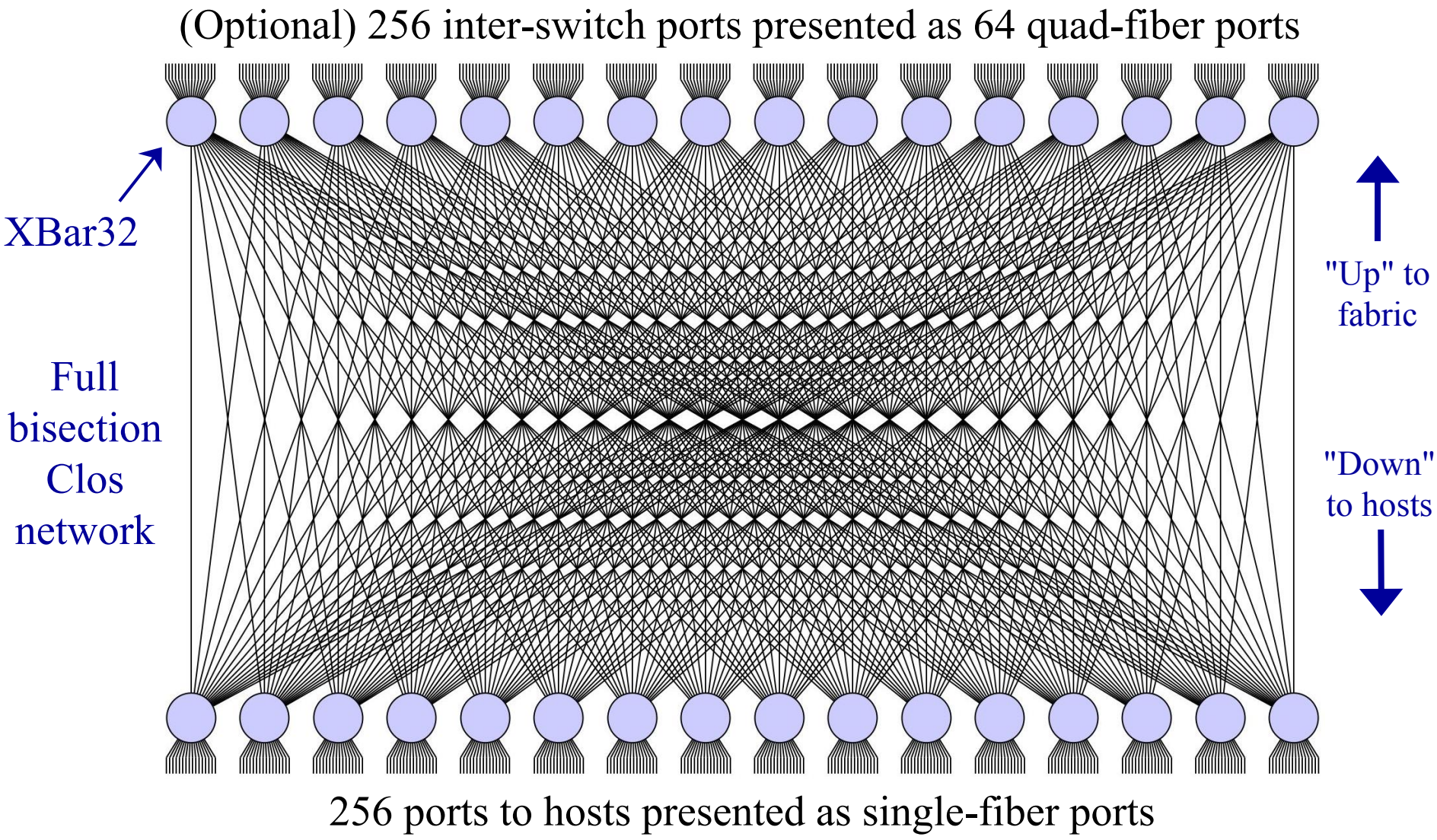


The XBar32 Chip

These new switch products, and others to follow, are based on a new, 32-port, crossbar-switch chip now in production.

- ~7W, 480 BGA, integrated (2.5GBaud Myrinet-serial) PHY
- New capability of disabling a quadrant of the crossbar for hardware assurance of deadlock-free routing in Clos networks
- New capability of accessing the switch ID and absolute port # to accelerate and simplify mapping (internally a 33x33 crossbar)
- **Also ideal for integration in blade systems**

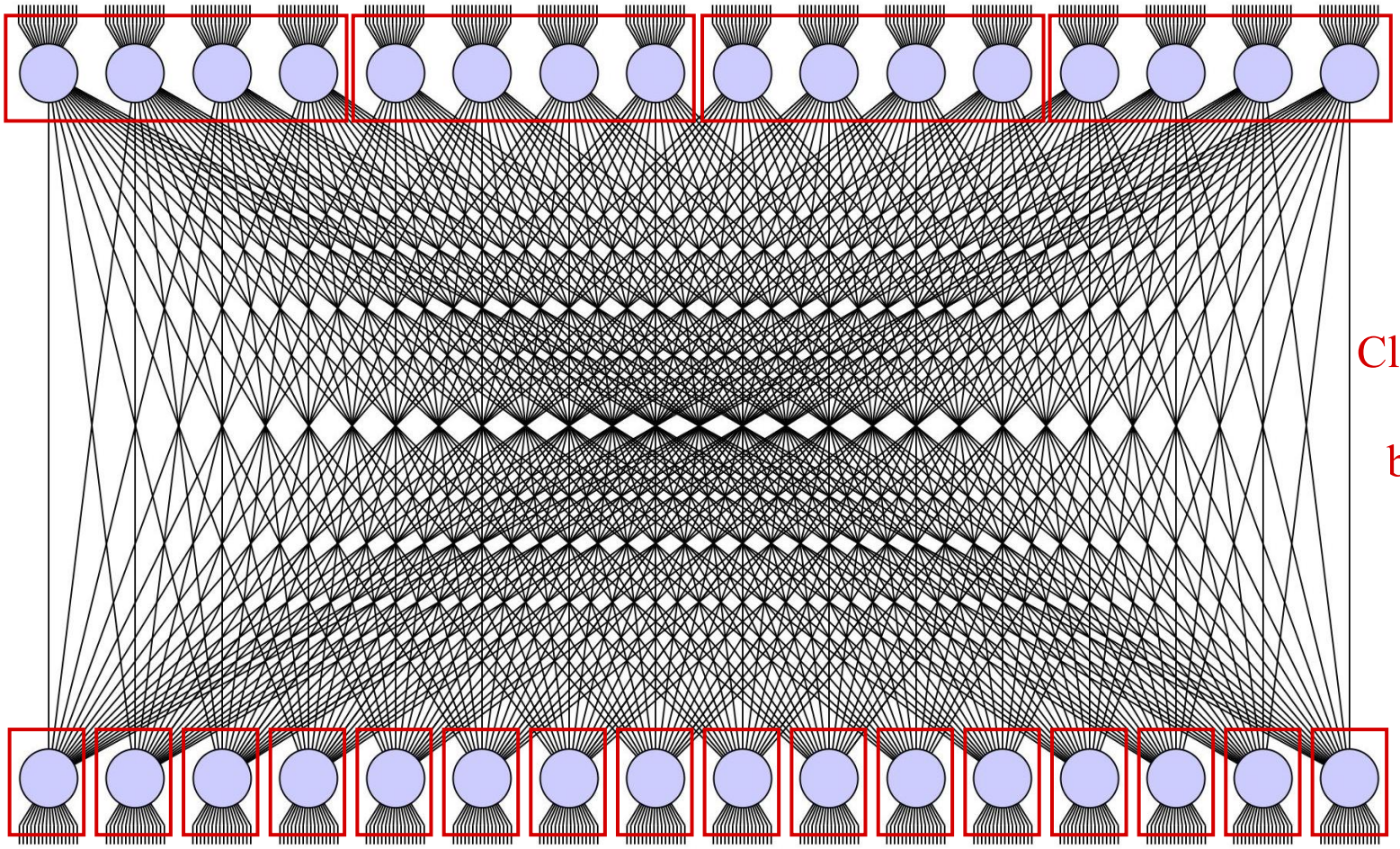
Internal Clos Topology of the Clos256+256



Myricom, Inc. 325 N. Santa Anita Ave. Arcadia CA 91006
626-821-5555 Fax: 626-821-5316 http:/ /www.myri.comv

Packaging

(Optional) 256 inter-switch ports presented as 64 quad-fiber ports



4 quad-SW32 line cards, each with 16 quad-fiber ports

Clos network on the backplane

16 SW32 line cards, each with 16 single-fiber ports

256 ports to hosts presented as single-fiber ports

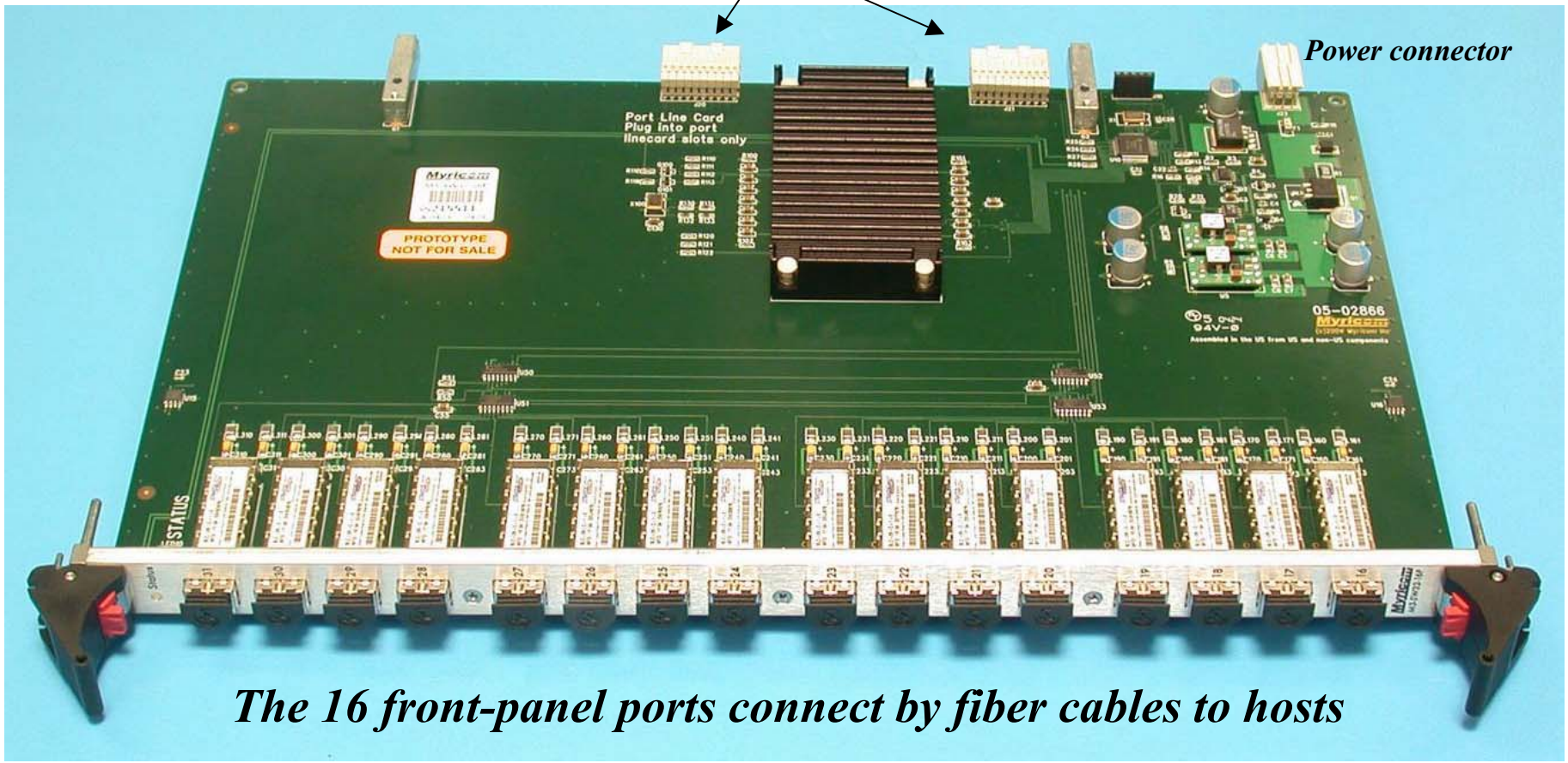
Myricom

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

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

M3-SW32-16F Line Card

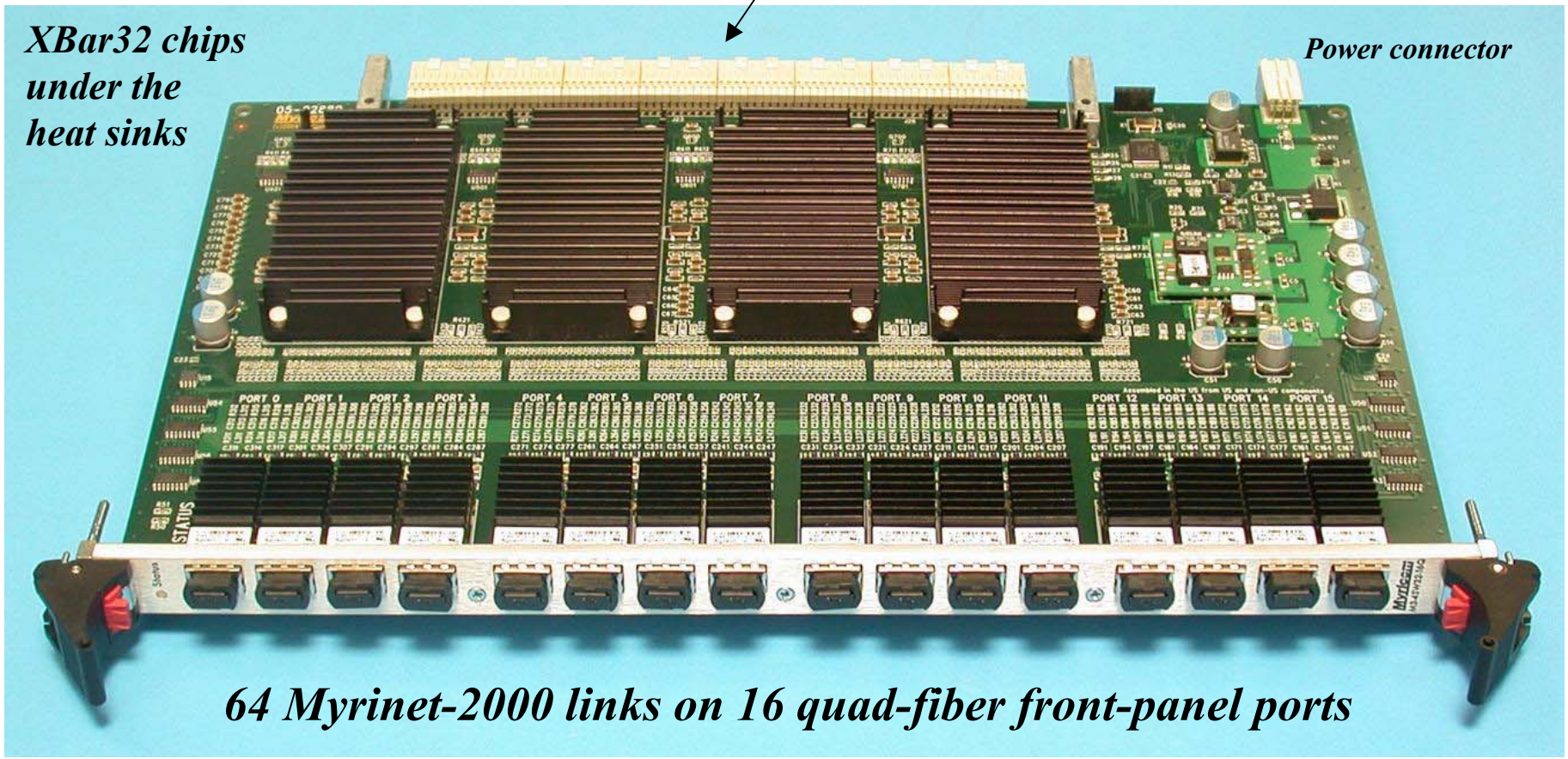
The two high-speed-signal connectors carry 16 links to the backplane



The 16 front-panel ports connect by fiber cables to hosts

M3-4SW32-16Q Line Card

64 Myrinet-2000 links on the high-speed-signal connectors to the backplane



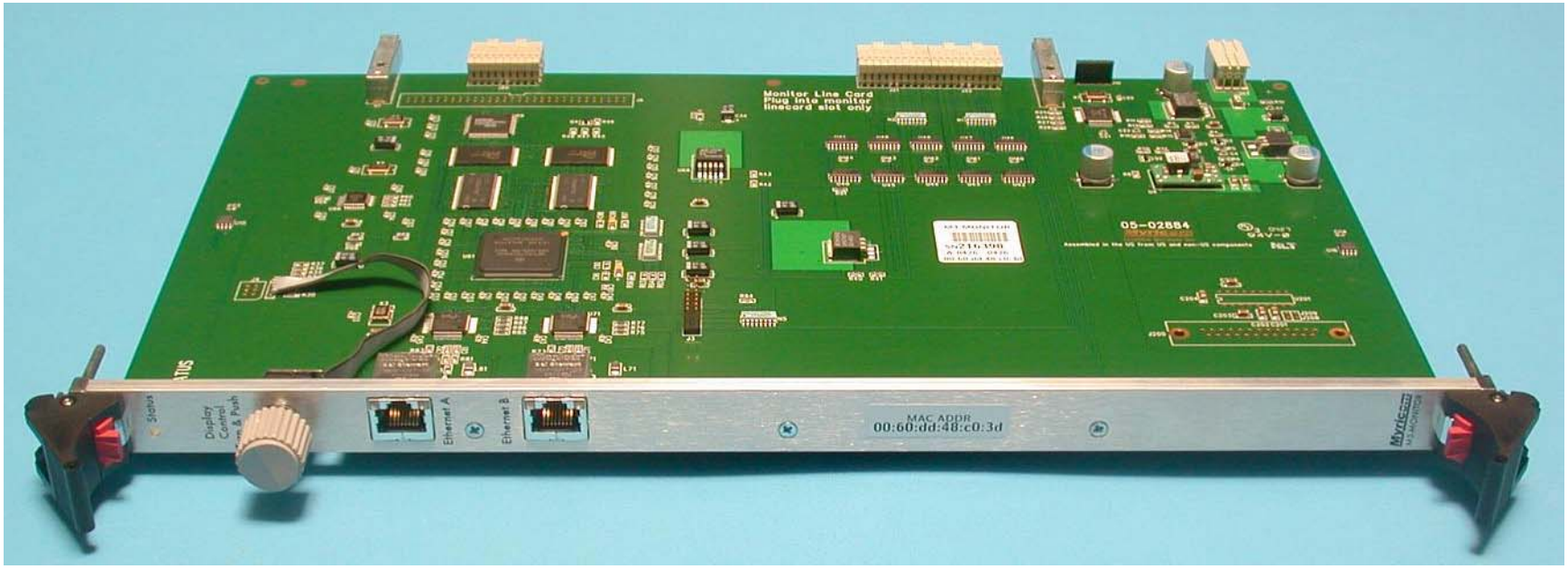
Very high switching density, *e.g.,* much higher switching and port density than components for GbE switches, and on links that operate at twice the data rate.

Myricom

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

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

M3-MONITOR Monitoring Line Card



This monitoring line card, a PPC-based computer that runs Linux, provides status information through dual-redundant 10/100 ethernet ports, and also drives the color TFT display on the enclosure. The turn/push knob controls the display.

Inside View -- M3-CLOS-ENCL

An M3-CLOS-ENCL enclosure (engineering prototype, without final labeling) showing the backplane, and with the monitoring card installed.

The monitoring card, power supplies, and fan units are FRUs, but are included with every enclosure. A rack kit also ships with every enclosure.

Inside View -- M3-SPINE-ENCL

An M3-SPINE-ENCL enclosure (engineering prototype, without final labeling) showing the backplane, and with the monitoring card installed.

The monitoring card, power supplies, and fan units are FRUs, but are included with every enclosure. A rack kit also ships with every enclosure.

M3-RACK-KIT

The rack kit shown to the right is included with the M3-CLOS-ENCL and M3-SPINE-ENCL enclosures, and is designed to mount between the front and rear vertical rack-mount rails, and telescopes in order to accommodate different front-to-rear spacing.

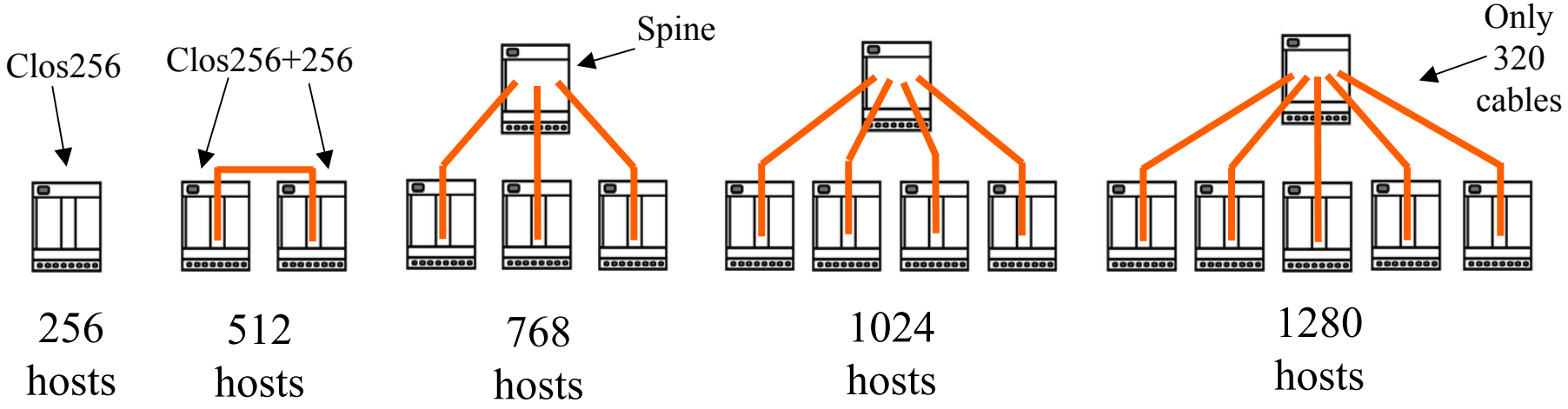
Different mounting-hole patterns on the rack-mount ears can be supplied on special order.

Use of the rack kit is highly recommended both to assist during installation and to support the enclosure in the rack against the shock and vibration that may be encountered when shipping racks with switches installed.

Product Codes

- Enclosures
  - M3-CLOS-ENCL
  - M3-SPINE-ENCL
- FRUs included with each enclosure
  - M3-MONITOR -- monitoring card (1)
  - M3-POWER -- power supply (4)
  - M3-FAN -- blower unit (2)
  - M3-RACK-KIT -- rack kit (1)
- Line cards
  - M3-SW32-16F -- SW32 line card with 16 fiber front-panel host connections
  - M3-4SW32-16Q -- Quad SW32 line card with 16 quad fiber front panel connections
  - M3-2SW32 -- Dual SW32 line card (used as center line cards for Clos256)
  - M3-THRU-16Q -- Line card for 16 quad fiber thru connections
  - M3-AIRDAM -- Blank line card with air dam

Designed to Scale



- 1536 host ports requires 6 Clos256+256 and 2 partially populated Spine units
- …
- 2560 host ports requires 10 Clos256+256 and 2 fully populated Spine units

All inter-switch cabling on quad-link ribbon fiber

A Myrinet Switch Network for 2560 Hosts

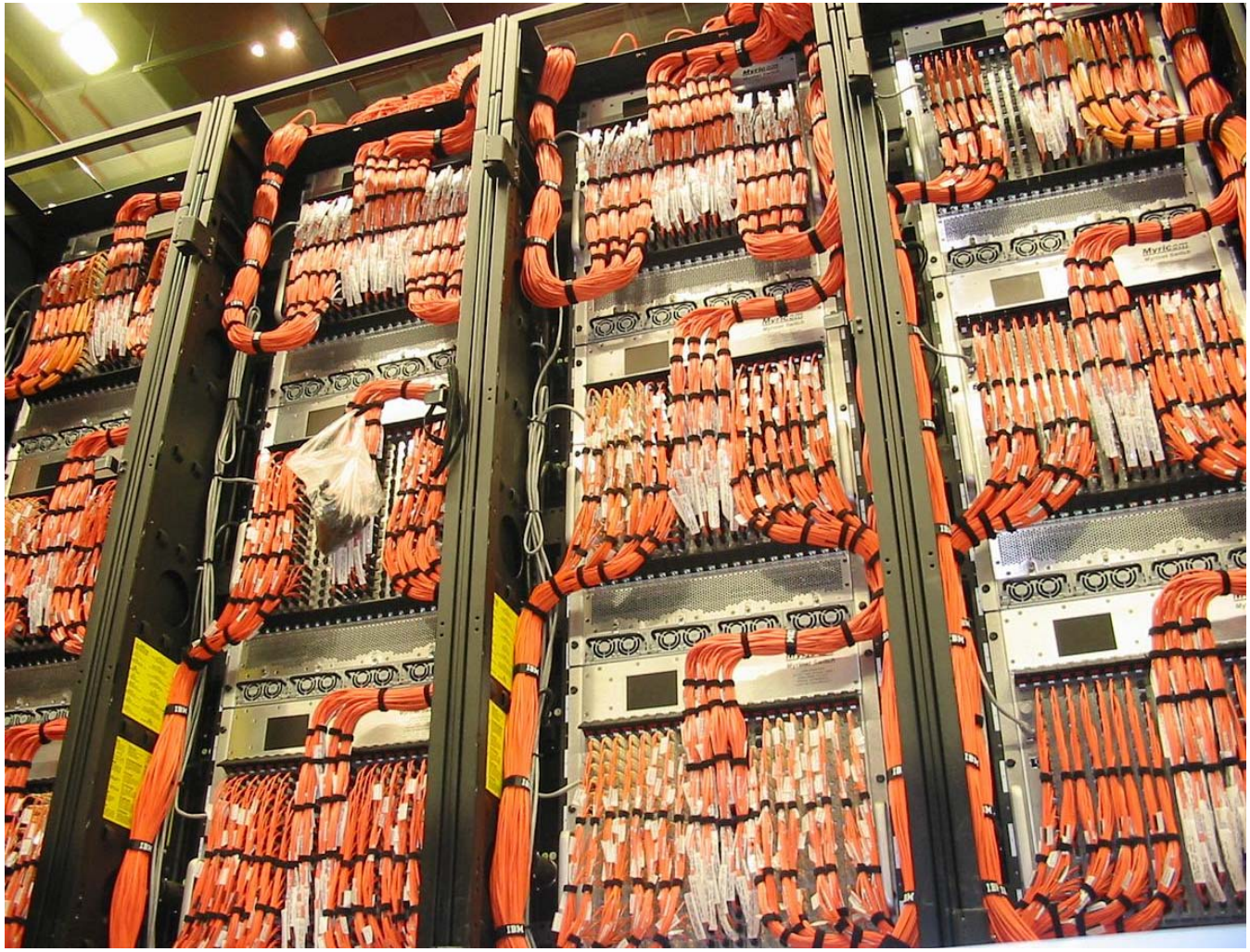
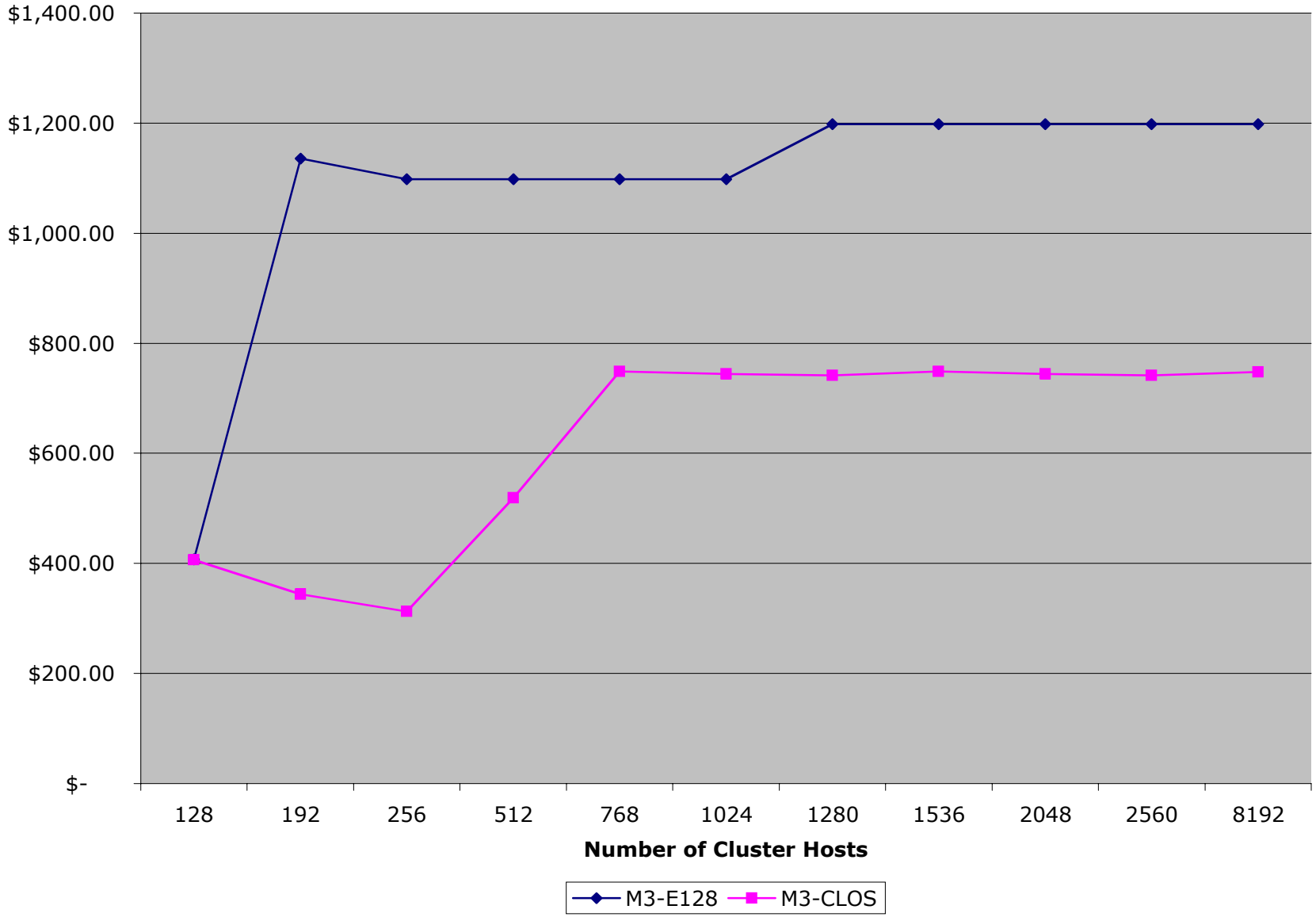


Photo courtesy of IBM

The switch network for the MareNostrum cluster at the Barcelona Supercomputing Center. MareNostrum was ranked #4 in the Nov-04 TOP500, and is the fastest cluster in the world.

An Advance in Myrinet-Switch Costs



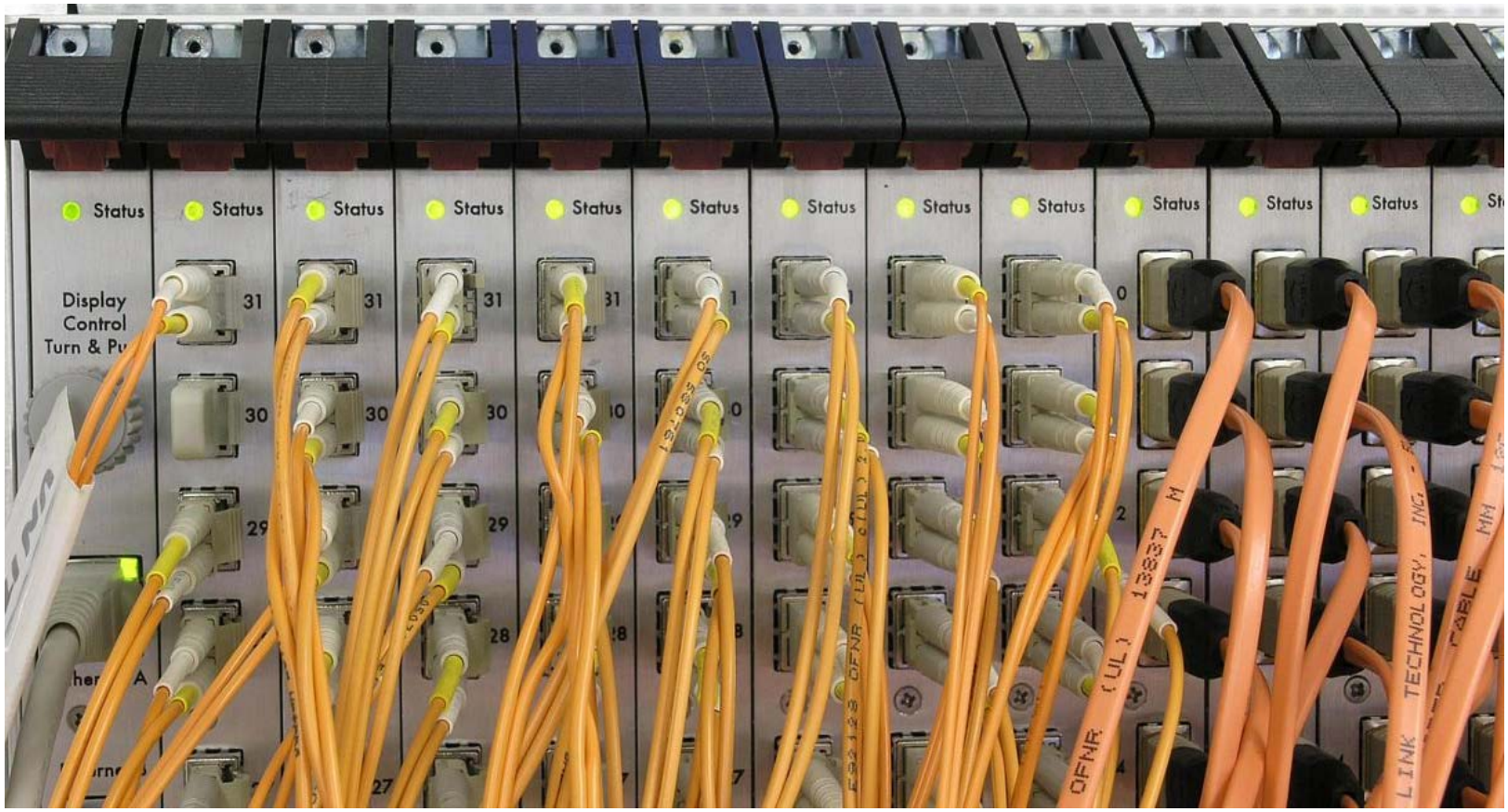
Myricom

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

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

Myrinet-2000 Fiber Cables

Myrinet-2000 fiber links, 2+2 Gbit/s, full duplex



Advantages of fiber: small-diameter, lightweight, flexible cables; reliability; EMC; 200m length; small connectors; low-cost industry-standard 50/125 multimode fiber. The cables on the right side of this photo are quad-link ribbon fiber for ultra-high density for inter-switch links. The optical signaling is 2.5GBaud, 8b/10b encoded.

Myricom

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

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

Myrinet-2000 Fiber cables

- Hosts: (50/125 multimode duplex cables with LC connectors)
- Interswitch: (Quad-link ribbon-fiber cables with MTP/MPO connectors)
- Nine cable lengths from which to choose (1M - 200M)
  - M3F-CB-{1M,3M,5M,10M,25M,50M,100M,150M,200M}
  - M3Q-CB-{1M,3M,5M,10M,25M,50M,100M,150M,200M}
- Cables are hot-pluggable.
- **Cautions:**
  - Myricom does not guarantee correct operation with other than 50/125 fiber cables.
  - Care should be taken when plugging/unplugging cables, otherwise damage can result.
  - Avoid crimping cables (tight angles) as damage can result. The minimum bend radius for fiber cables is a "finger width" (or 1/4" radius).
  - Provide support restraints for cabling of large cluster configurations.

Hardware Installation

Hardware Installation

- Upon receipt of the Myrinet hardware, read and follow the instructions in the following three guides:
  - "Guide to Myrinet-2000 Switches for Large Clusters”
    - http://www.myri.com/myrinet/14U_switches/guide/
  - "Guide to Myrinet/PCI-X Network Interface Cards”
    - http://www.myri.com/scs/doc/guide_to_pcix_nics.pdf
  - “Myrinet Installation and Troubleshooting Guide”
    - http://www.myri.com/scs/doc/troubleshooting_guide.pdf
- After reading these three guides, you should be able to install and connect all of the hardware.
- Difficulties? Refer to the Myrinet FAQ and the Switch Tutorial
  - http://www.myri.com/scs/FAQ/
  - http://www.myri.com/scs/14U_switches/

Installation of the Myrinet PCI-X NICs

- **Step 1**: Power off the host to which you will be installing the NIC(s). Remove the exterior covering of the machine, and locate the PCI-X slot.
  - If the mother board has more than one PCI-X slot, identify which PCI-X slots run at maximum speed (133MHz), and which PCI-X slots share a PCI bus.
  - Note that some vendors ship their mother boards with the speed of the PCI-X slots reduced, and you must manually adjust the jumper next to the PCI-X slots to run at full speed.
  - To minimize the signal degradation between the mother board and the Myrinet NIC (and for best performance), we recommend that the Myrinet NIC is inserted in the PCI-X slot closest to the PCI chipset.
  - We also recommend that the Myrinet NIC not share the PCI bus with another PCI device.

Installation of the Myrinet PCI-X NICs contd.

- **Step 2**: Insert the Myrinet NIC into a PCI-X slot.
  - Inasmuch as some PCI connectors are more close-fitting than others, you may need to push quite hard on the front panel and the edge of the NIC to seat the NIC securely.
  - If at all possible, avoid the use of a riser card.
  - If you must use a riser card and the riser card has multiple slots, insert the Myrinet NIC into the riser card slot closest to the PCI-X slot on the mother board.
  - Although PCI riser cards are commonly used, they will generally violate PCI specifications for the length of signal traces. A riser card may introduce impedance discontinuities and signal degradation between the mother board, riser card, and NIC. If you observe PCI-communication errors when using a riser card, see if the problem persists when you plug the Myrinet NIC directly into the PCI slot. A higher quality riser card may also solve the problem.

Installation of the Myrinet PCI-X NICs contd.

- **Step 3**: Secure the NIC in place with a locking screw (if applicable), replace the exterior cover of the host, and attach a cable (fiber) between the NIC and the switch port.
- **Step 4**: Power on the host and check that the Myrinet NIC is correctly detected by the host operating system.
  - If your host is running Linux, you can issue the command **/sbin/lspci**, which will return all of the devices attached to the PCI bus.
  - You should see a line of text in the output similar to:

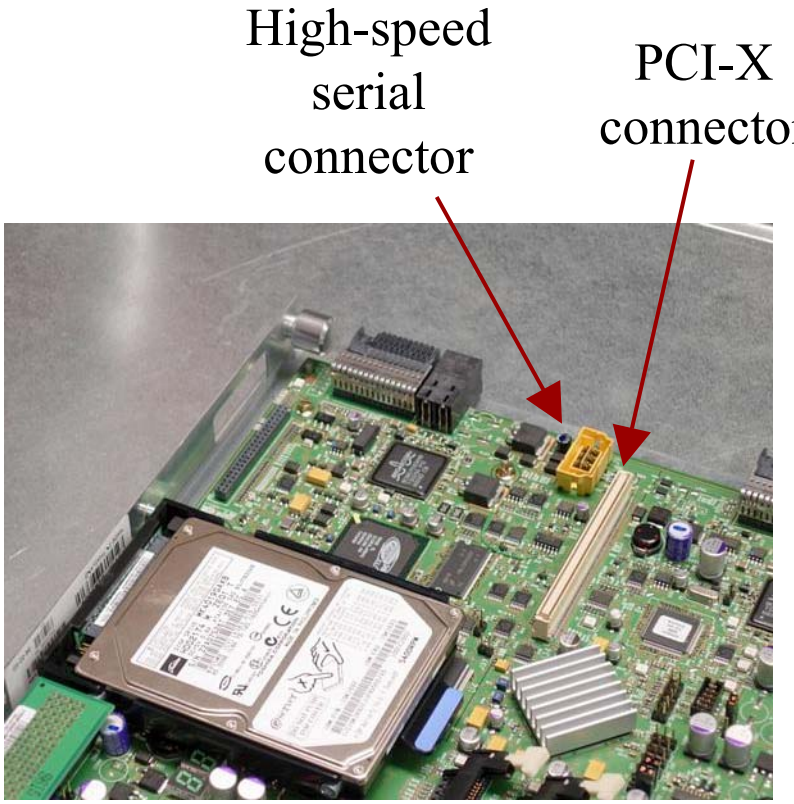
```
Network controller: MYRICOM Inc. Myrinet 2000 Scalable
  Cluster Interconnect (rev 04)
```

  - The (rev 04) denotes a PCIXD NIC, (rev 05) is a PCIXE NIC, and (rev 06) is a PCIXF NIC.

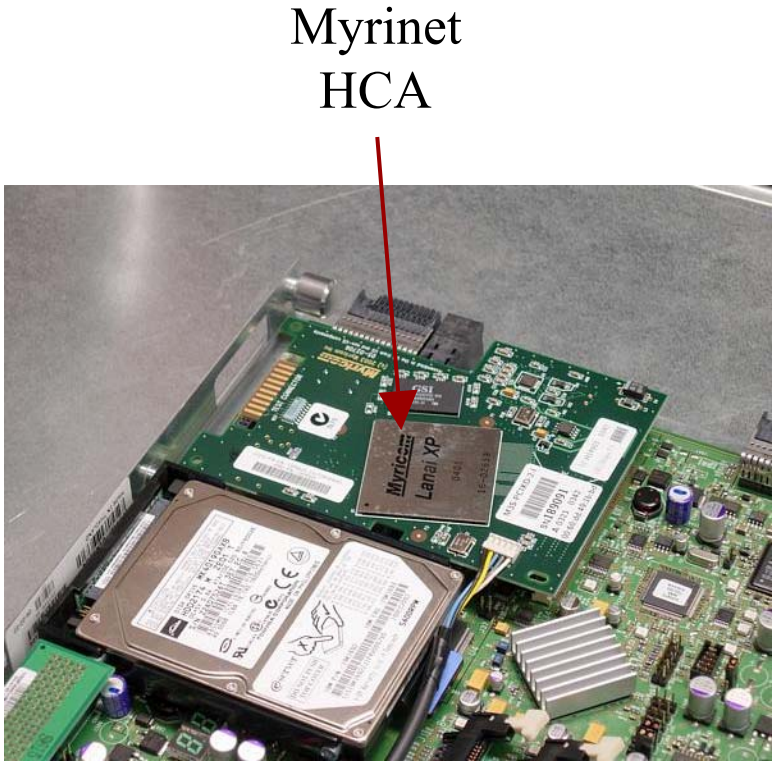
Difficulties?

- If a Myrinet PCI-X NIC is not detected using **/sbin/lspci**, then
  - Have you tried reseating the NIC?
  - Are you using a riser card?
  - Have you tried inserting the NIC into a different slot on the riser card?
  - Have you tried inserting the NIC directly into the PCI slot?
  - Have you tried using a different PCI slot?
  - A different riser card?
  - A newer BIOS for this motherboard?
  - A different Myrinet NIC?
- Examine the **/sbin/lspci -vvv** output for clues.

BladeCenter HCA Installation



Where the HCA is installed



After installation

BladeCenter HCA Installation

OPM Module

BladeCenter Hints

- **Q**: What's the purpose of the dangling cable on the M3S-PCIXD-2-I daughter card for the bladecenter?
- **A**: We don't really know what the cable is for. It is required in the IBM spec for the daughter cards for the blades. When we asked IBM engineers if we really needed to use the cable, the response was yes. In fact, we don't make the cables ourselves, IBM supplies them to us, and we then install them on the boards. We don't know how or when it is to be used, but we do test them.
- The circuit on the board is nothing more than a level converter between signals in the cable and signals on the second mid-plane link. There is no interaction whatsoever with the Myrinet/PCI-X circuitry.

BladeCenter Hints

- **Q**: Are there any special instructions for M3S-PCIXD-2-I installation?
- **A**: After installing the cards into the bladecenter:
- First make sure that the OPM module is in the right slot, which is Bay 4.
- Then hookup the cables from the OPM to the Myrinet switch.
- Note that the newer versions of the bladecenter firmware seem to require you access the blade center maintenance module and turn on the OPM lasers. (Otherwise, there will be no link between the OPM and the Myrinet switch.) The option is under: I/O Module Tasks -> Management -> Advance Management (under the OPM entry in Bay 4) -> External Ports. Set it to enabled.

Installation of the Myrinet Switch and Cables

Mounting the Switch into the Rack

- Enclosures are shipped completely assembled and tested, including power supplies, fans, and the monitoring line card, but without other line cards, which ship separately. The weight of the unit in this shipping configuration is approx. 55 pounds (25 kg).
- The M3-RACK-KIT is included with but ships separately from the enclosure.
- The weight with line cards in all slots can be as high as approx. 101 pounds (46 kg).
- Due to the weight of the unit, it is recommended that the unit be installed in the rack while it is in the shipping configuration (25 kg), before installing the line cards.
- Use of the M3-RACK-KIT is highly recommended, both to assist in installation and to support the enclosure in the rack against shock and vibration.

Thermal and Cooling Specifications for the Switch

- Air flow is front-to-back, with separate air paths for the power-supply and card-cage cooling.
  - Note that the direction of the air flow cannot be changed. Thus, we recommend that switches are mounted with the front face of the switch at the front of the rack.
  - Each of the four power supplies has a pair of high-speed (10,000 rpm) 1.5-inch (38mm) fans. The card cage cooling is accomplished by two M3-FAN assemblies, each of which has a series pair of 4.5-inch (114mm) fans.
  - Fan rotation speed for the card-cage cooling is monitored, and can be read along with the exhaust air temperature through the M3-MONITOR monitoring line card.
  - The small microcontroller card for each M3-FAN unit also controls the fan speed according to the exhaust air temperature.
  - The M3-FAN units operate at maximum speed at exhaust-air temperatures of 40C or higher.

Thermal and Cooling Specifications for the Switch

- Internal temperatures are measured both upstream and downstream in the air flow of each line card.
  - The temperatures can be monitored through the M3-MONITOR monitoring card.
  - The line card will shut down autonomously on an over-temperature condition (55C circuit-board temperature).

Switch Installation

- **Step 1**: Plug in the power cord of the switch and the color TFT display (driven by the monitoring line card) will illuminate and have a color-bar display. After the operating system finishes to boot (about 10 seconds) the color-bar display will change to a virtual image of the switch. Connected switch ports will be illuminated in the image. The turn/push knob on the monitoring line card controls the TFT display. Do not yet connect ethernet to the monitoring line card, as the configuration of the monitoring line card will be performed after all of the fiber cables have been connected.
- **Step 2**: Verify that all of the switch line cards are properly seated. If a switch line card is properly seated, you will the Status LED, located on the top of each front panel, illuminated in green.

Switch Installation

- **Step 3**: A cable should then be connected between the fiber port on each NIC and a port on an M3-SW16-16F switch line card on the M3-CLOS-ENCL enclosure. Configurations of more than 256 hosts employ quad-link ribbon-fiber cables for inter-switch connections.
- **Step 4**: Configure the monitoring line card (located in the left-most slot) in the switch.
  - Read the MAC address from the faceplate of the monitoring line card and register this MAC address with a static IP address in the DHCP server configuration file (/etc/dhcpd.conf) on the DHCP server on the local network. The DHCP server will then serve this static IP address to the monitoring line card when it boots and asks for it.
    - **Note**: A future release of the switch firmware will allow the customer assign a static IP address to the monitoring line card.
  - Connect at least one of the 10/100-Base-T ethernet ports on the monitoring line card to the LAN.

Switch Installation

  - As soon as the ethernet port is connected, the upper green LED on the RJ45 connector will illuminate, and the monitoring line card will immediately start to broadcast DHCP requests. When the monitoring line card has received its IP address, it is reachable. You can ping the card, open a web browser to it, or walk the SNMP MIB.

- Once the monitoring line card has been properly configured, we recommend that you first upgrade the switch firmware on the monitoring line card.

  http://www.myri.com/scs/14U_switches/index-overview-web.html#umcf

- A Tutorial for the Web Interface and the TFT Display for the Switch is available.

  http://www.myri.com/scs/14U_switches/

Hardware Installation Summary

- **Important points to note**:
  - Host cables should be disconnected by carefully depressing the connector tab, otherwise damage can result. To disconnect interswitch cables the beige sleeve/latch must be pulled back on the black connector. No force whatsoever should be applied to the cable itself. We have had a number of reports of physical damage to cables from customers who do not properly disconnect the cable.
  - Avoid crimping cables (tight angles) as damage can result.
  - Provide proper ventilation for the switch(es), otherwise overheating shutdown could occur.
  - Fiber-cable ends and ports on line cards must be kept free of dust particles. Accumulation of dust can cause faults from the port-to-fiber connection. We recommend that you use a moisture-free dust gas to clean the fiber ports.
  - We recommend the use of dust plugs in unused ports, and require blank panels in underpopulated switches.
  - Provide support restraints for cabling of large cluster configurations.

Hardware Installation Summary

- **Important points to note (contd):**
  - Green “link” LEDs should illuminate on NICs and ports on the line cards when the hardware is connected through a cable to an operating component and GM is loaded on the host(s). For more information on LEDs, refer to the FAQ:
    - http://www.myri.com/cgi-bin/fom?file=60
  - Green “link” LEDs should also illuminate on the connected interswitch ports on the switch(es). Refer to the Tutorial for the Web Interface and TFT Display for further details.
    - http://www.myri.com/scs/14U_switches/
  - No riser cards are required for PCI-X NICs on 2U and larger hosts.
  - Riser cards can be a significant source of problems in a hardware configuration.

Troubleshooting Hardware

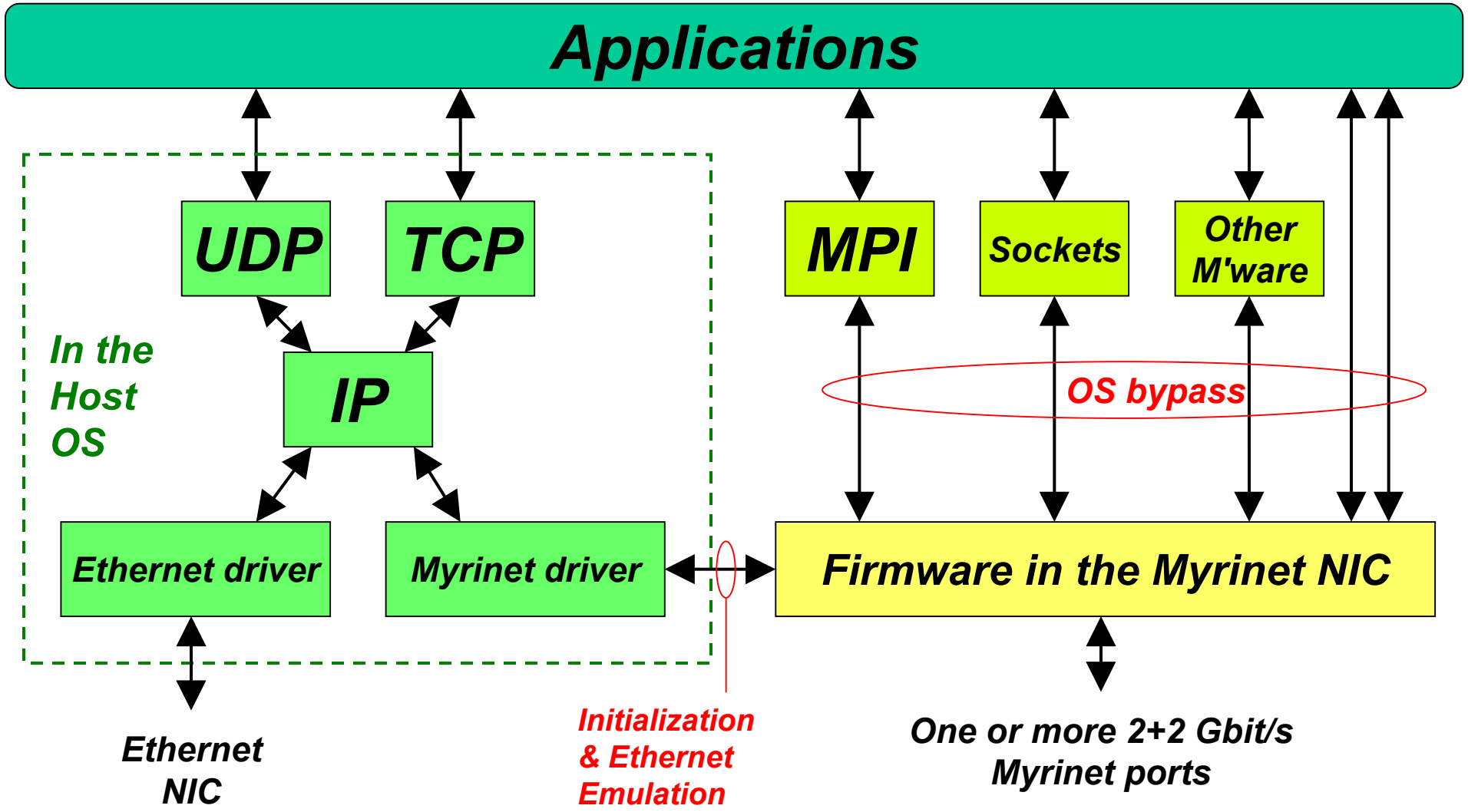
- A number of troubleshooting scenarios are listed on the Tutorial webpage:
  - http://www.myri.com/scs/14U_switches/
- Scenario 1: One of the connected switch ports is not illuminated.
- Scenario 2: None of the connected switch ports on a switch line card are illuminated.
- Scenario 3: How do we tell if the switch firmware is up-to-date?
- Scenario 4: The image on the TFT Display is white/blank. How do I fix this?

Software Overview

Software Overview

- Myricom-supported open-source software is available
  - http://www.myri.com/scs/
- For GM-2 and MX-2G, we support Linux 2.4 and 2.6 on IA32, IA64, AMD64, EM64T, PowerPC, PowerPC64, Power4, and Power5.
- Software Documentation
  - http://www.myri.com/scs/#documentation
- Myrinet Mailing List (myrinet@osc.edu) for discussions and announcements

Myrinet Software Interfaces



Myricom, Inc. 325 N. Santa Anita Ave. Arcadia CA 91006
626-821-5555 Fax: 626-821-5316 http:/ /www.myri.comv

Choice of Myrinet Software Interfaces

Applications need not be tailored for Myrinet.
Myricom provides the APIs the applications require.

- **Low-level APIs**
  - GM 1 (legacy), GM 2 (current standard), MX (new), 3rd party (*e.g.,* SCore)
- **TCP/IP & UDP/IP** – *Commercial Applications*
  - Ethernet emulation, included in all GM and MX releases
    - 1.98 Gb/s (D or F cards) or 3.95 Gb/s (E cards) TCP/IP netperf on Linux (2.6.11smp kernel)
- **MPI** – *HPTC Applications*
  - An implementation of Argonne MPICH directly over GM or MX.
  - Third-party MPI implementations over Myrinet are also available.
- **Sockets** – *High-Performance TCP/IP Applications*
  - An implementation of UNIX or Windows sockets (or DCOM) over GM or MX. Completely transparent to application programs. Use the same binaries!
- **uDAPL and kDAPL** – *Database Cluster Applications*
  - The new standard for distributed databases (Oracle, DB2)

Myrinet Express (MX)

- MX was designed based on earlier experience in writing MPICH-GM and VI-GM middleware
  - MX does generalized matching in the NIC firmware.
    - One of the keys to middleware performance
  - Perfect overlap of computation and communication under MPI.
  - Exceptional MPI ping-pong latency

NIC	D card (one port)	E card (two ports)	F card (one port)
MX and MPICH-MX unidirectional throughput	**230 MB/s**	**475 MB/s**	**240 MB/s**
MX and MPICH-MX bidirectional throughput	**450 MB/s**	**840 MB/s**	**475 MB/s**
MX and MPICH-MX latency	**3.5µs**	**2.7µs**	**2.6µs**

MX-2G-Beta between dual-Opteron hosts, including the latency through one switch

Adaptive Route Dispersion

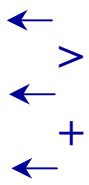
- MX takes advantage of the multiple paths through large networks (*e.g.,* slide 27) to spread packet traffic
  - MX mapping provides multiple routes to each other host.
  - MX measures the time to inject each packet in order to sense contention. Brief flow-control backpressure indicates contention on the route.
  - MX changes route only when contention is sensed in the network
    - Automatically adapts to current traffic patterns.
  - Note: MX can receive packets out of order
    - But matching (message level) is always in-order.
- Eliminates "hot spots" in switch networks
  - Adapts the routes to the communication patterns of the application.
  - Extremely valuable for large switch networks.
- **Only possible with source-routed networks**

MX is the successor to GM

- MX is the successor to GM.
  - http://www.myri.com/scs/MX/doc/mx.pdf
- MX will go into full release with the next generation of Myrinet NICs, “Myrinet-10G” products, which we expect to announce in 2Q05. However, MX was developed on and is available for the PCIX series NICs.
- When MX goes into full release, two versions will be available:
  - MX-2G for PCIX series NICs
  - MX-10G for the Myrinet-10G NICs.
- After the Beta-release stage, Myricom plans to charge nominal license and support fees for MX software support.
- GM will not be supported on any new hardware after PCI-X.

Summary of the NICs & software

NIC	*May 2003* M3F-PCIXD "D card"	*February 2005* M3F-PCIXF "F card"	*Sept 2003* M3F2-PCIXE "E card"
NIC chip	Lanai XP	Lanai 2XP	Lanai 2XP
Myrinet ports	1	1	2
Lanai & memory clock	225MHz	333MHz	333MHz
Local memory data rate	1800 MB/s	2664 MB/s	2664 MB/s
Peak bidirectional Myrinet data rate	500 MB/s	500 MB/s	1000 MB/s
Peak PCI data rate	1067 MB/s	1067 MB/s	1067 MB/s
Myricom software support	GM 2 & MX-2G	GM 2 & MX-2G	GM 2 & MX
MX or GM bidirectional throughput	489 MB/s	489 MB/s	760 MB/s
GM latency	6.3µs	5.0µs	5.7µs
MPI/GM-2 latency	6.8µs	5.5µs	6.2µs
MX and MPI/MX latency	3.5µs	2.6µs	2.7µs



Performance of GM-2 and MX-2G

- Performance graphs for GM-2 and MX-2G are available.
  - http://www.myri.com/myrinet/performance/

IP over Myrinet

Ethernet Emulation

What is IP-over-Myrinet (Ethernet Emulation)?

- In addition to its OS-bypass features, GM/MX also presents itself to the host operating system as an ethernet interface.
- This "ethernet emulation" feature of GM/MX allows Myrinet to carry any packet traffic and protocols that can be carried on ethernet, including TCP/IP and UDP/IP.
- When using “ethernet emulation” over GM/MX, traffic goes from the application through the OS kernel to the GM/MX driver, following the same path as it would for a “real” ethernet interface; traffic does not go directly from the application to the interface, as it does when using GM/MX in its OS-bypass mode.

IP-over-Myrinet (Ethernet Emulation)

- The TCP/IP and UDP/IP performance over GM/MX depends primarily on the host-CPU performance and the host-OS's IP protocol stack.
- This performance varies widely for different hosts and operating systems.
- Unlike GM/MX's OS-bypass mode, which exhibits very low host-CPU overhead, TCP/IP and UDP/IP protocol processing at high data-transfer rates may use a significant fraction of the host-CPU cycles.
- Performance measurements for "ethernet emulation" using PCI-X NICs can be found on the following webpage:
  - http://www.myri.com/myrinet/performance/

Myrinet TCP/IP Benchmark (Benefits of Offload)

Netperf with MX-2G Beta 0.8.8 on E (M3F2-PCIXE-2) cards:

```
>netperf224 -H shout-my -l60 -c -C -- -S262144 -s262144
TCP STREAM TEST to shout-my
Recv   Send    Send                  Utilization       Service Demand
Socket Socket  Message  Elapsed              Send     Recv     Send    Recv
Size   Size    Size     Time     Throughput  local    remote   local   remote
bytes  bytes   bytes    secs.    10^6bits/s  % S      % S      us/KB   us/KB

217088 217088 217088   60.01    3945.89   34.60    39.96    1.437   1.659
```

This Myrinet ethernet-emulation performance closely approaches the dual-port E card's wire speed of 4 Gbits/s at less than 40% CPU utilization. This benchmark was run under the Linux 2.6.11smp kernel between single 3.06GHz P4s with hyperthreading enabled. The MX driver was configured to use a 9K MTU for ethernet emulation. With single-port D cards, the throughput is 1.977 Gbits/s at less than 21% CPU.

Mapper

Why do I need a Mapper?

- Myrinet is a source-routed network. I.e., each host must know the route to all other hosts through the switching fabric.
- The Mapper automatically discovers all of the hosts connected to the Myrinet network, computes a set of deadlock-free minimum-length routes between the hosts, and distributes appropriate routes to each host on the connected network.

Myricom, Inc. 325 N. Santa Anita Ave. Arcadia CA 91006
626-821-5555 Fax: 626-821-5316 http:/ /www.myri.comv

How is the Mapper implemented?

- For GM-1, the mapper ran as a single user process on one node, the active node, with a small piece of mapper code in each MCP.
- For GM-2, there's a mapper running as a user process for each Myrinet NIC.
- For GigE switches, there's a mapper running in a software thread inside the MCP on each XM.
- For MX (Myrinet Express), there's a mapper running as a user process for each Myrinet NIC, and it is started automatically by the MX driver.

Mapper and GM-2 and MX-2G

- Why was the mapper changed from GM-1 to GM-2?
  - Scalability, better routing, and fault tolerance.
- Do GM-1 and GM-2 interoperate?
  - No. If you have a mixture of nodes running GM-1 or GM-2, the GM-1 mapper will only see those nodes running GM-1, and vice versa.
- Do GM-2 and MX-2G interoperate?
  - No.
- Full details on how the Mapper in GM-2 works is explained on the following webpage:
  - http://www.myri.com/scs/mapper_gm2.html
- Fabric Management System (FMS) will eventually replace the Mapper
  - http://www.myri.com/scs/fms/

Software Installation

Software Installation

- The first Myrinet software package you must install is the low-level firmware (GM-2 or MX-2G).
- For MX-2G Beta, contact help@myri.com.
- For the GM-2 Linux Download Page refer to:
  - http://www.myri.com/scs/linux-gm2.html
- These webpages contain the links for software download AND abbreviated installation instructions, as well as any "Current Cautions or Common Problems" that could be encountered.

Software Considerations

- GM-2 is supported on PCIX-based and PCI64-based Myrinet NICs.
- For best performance with GM-2, we recommend:
  - GM-2.0.x (GM-2.0.19) for PCIXD or PCIXF clusters
  - GM-2.1.x (GM-2.1.9) for PCIXE clusters

- MX-2G is supported on PCIX-based Myrinet NICs (PCIXD, PCIXF, and PCIXE).
- MX-2G and GM-2 do not interoperate.
- All hosts on the same Myrinet network must run the same version of MX/GM.
- If you're using GM and you have multiple types of NICs on your network, you'll need to use the higher GM version.

Scalability limit of MX-2G

- MX-2G Beta 0.8.8
  - For one-port NICs (PCIXD-2, PCIXF-2), up to 4096 hosts.
  - For two-port NICs (PCIXE-2), up to 2048 hosts.
- If the cluster has more than 4096 hosts (or 2048 hosts respectively), you would need to use the 4MB Myrinet NICs.

Scalability limit of GM-2

- For the up-to-date scalability limits for GM-2 releases, refer to the SCALABILITY file in the GM-2 distribution tar file.

- GM-2.0.19
  - 32-bit pointers 4KB pages (E.g., x86, PowerPC32)
    - PCIXD-2 supports 2866 nodes
    - PCIXD-4 supports 7962 nodes
  - 64-bit pointers 4KB pages (E.g., AMD64, EM64T, PowerPC64)
    - PCIXD-2 supports 1558 nodes
    - PCIXD-4 supports 6654 nodes
  - 32-bit pointers 8KB pages (E.g., 32-bit UltraSPARC)
    - PCIXD-2 supports 2802 nodes
    - PCIXD-4 supports 7898 nodes

Scalability limit of GM-2

- GM-2.0.19
  - 64-bit pointers 8KB pages (E.g., 64-bit UltraSPARC, Alpha, some Itaniums)
    - PCIXD-2 supports 1750 nodes
    - PCIXD-4 supports 6846 nodes
  - 32-bit pointers 16KB pages
  - 64-bit pointers 16KB pages (E.g., most Itaniums)
    - PCIXD-2 supports 1206 nodes
    - PCIXD-4 supports 6302 nodes
- GM-2.1.9
  - 32-bit pointers 4KB pages (E.g., x86, PowerPC32)
    - PCIXE-2 supports 1279 nodes
    - PCIXE-4 supports 6375 nodes

Scalability limit of GM-2

- GM-2.1.9
  - 64-bit pointers 4KB pages (E.g., AMD64, EM64T, PowerPC64)
    - PCIXE-2 supports 743 nodes
    - PCIXE-4 supports 5065 nodes
  - 32-bit pointers 8KB pages (E.g., 32-bit UltraSPARC)
    - PCIXE-2 supports 1215 nodes
    - PCIXE-4 supports 6311 nodes
  - 64-bit pointers 8KB pages (E.g., 64-bit UltraSPARC, Alpha, some Itaniums)
    - PCIXE-2 supports 677 nodes
    - PCIXE-4 supports 5257 nodes
  - 32-bit pointers 16KB pages
  - 64-bit pointers 16KB pages (E.g., most Itaniums)
    - PCIXE-2 supports 391 nodes
    - PCIXE-4 supports 4713 nodes

Software Installation (cont'd)

- MX/GM installation is performed in 3 easy steps:
  - 1. Configuring, compiling, and loading the MX/GM driver.
  - 2. Enabling IP over Myrinet (Ethernet emulation) (OPTIONAL)
  - 3. Testing/Validation
- Difficulties? Refer to:
  - Linux download page (http://www.myri.com/scs/linux-gm2.html)
  - Myrinet FAQ (http://www.myri.com/scs/FAQ/)
  - README in the MX software distribution tar file.
  - README-linux in the GM software distribution tar file.
  - Send email to help@myri.com.

Abbreviated MX-2G Installation Instructions

1. Configuring, compiling, and loading the MX-2G driver

- Request MX-2G Beta from help@myri.com
- Download MX-2G Beta

gunzip -c mx2g_beta_0.8.8.tar.gz | tar xvf -

cd mx2g_beta_0.8.8

./configure --with-linux=<linux-source-dir>

- where <linux-source-dir> specifies the directory for the Linux kernel source.

make

make install DESTDIR=<install_path>

su root

<install_path>/sbin/mx_local_install

<install_path>sbin/mx_start_stop start

Abbreviated GM Installation Instructions

1. Configuring, compiling, and loading the GM driver

- Download GM
- http://www.myri.com/ftp/pub/GM/gm-2.0.19_Linux.tar.gz

gunzip -c gm-2.0.19_Linux.tar.gz | tar xvf -

cd gm-2.0.19_Linux

./configure --with-linux=<linux-source-dir>

- where <linux-source-dir> specifies the directory for the Linux kernel source.

make

cd binary

./GM_INSTALL <install_path>

su root

<install_path>/sbin/gm_install_drivers

/etc/init.d/gm start

Additional Instructions for AMD64 or EM64T

- Refer to the FAQ entry "How do I build GM-2 on AMD64 or EM64T?" (http://www.myri.com/cgi-bin/fom?file=252) for additional installation instructions required for AMD64 and EM64T processors.
- If you're building GM-2 on SuSE SLES9 on PowerPC64 or AMD64, you may need to explicitly point configure at the kernel source and object trees:

```
configure --with-linux=/lib/modules/`uname -r`/source --with-linux-
build=/lib/modules/`uname -r`/build
```

- Otherwise, if you're using SuSE 9.0 or later, you might also want to refer to the Myrinet FAQ entry "I'm using SuSE Linux but gm_install_drivers complains of a running/source kernel mismatch. What's wrong?"(http://www.myri.com/cgi-bin/fom?file=272).

Additional Instructions for PowerPC64

- Refer to the FAQ entry "How do I build GM-2 on PowerPC64?"(http://www.myri.com/cgi-bin/fom?file=260) for additional installation instructions required for PowerPC64 processors.
- If you're building GM-2 on SuSE SLES9 on PowerPC64 or AMD64, you may need to explicitly point configure at the kernel source and object trees:

```
configure --with-linux=/lib/modules/`uname -r`/source --with-linux-
build=/lib/modules/`uname -r`/build
```

- Otherwise, if you're using SuSE 9.0 or later, you might also want to refer to the Myrinet FAQ entry "I'm using SuSE Linux but gm_install_drivers complains of a running/source kernel mismatch. What's wrong?"(http://www.myri.com/cgi-bin/fom?file=272).

Additional Installation Hints for Large Clusters

- When setting up large ethernet clusters (> 500 nodes), it is helpful to do arp table tuning.
- For example, on a 1000+ node cluster with IP/NFS, the following tuning parameters were used:
- net.ipv4.conf.all.arp_filter = 1
- net.ipv4.conf.all.rp_filter = 1
- net.ipv4.neigh.default.gc_thresh1 = 256
- net.ipv4.neigh.default.gc_thresh2 = 4096
- net.ipv4.neigh.default.gc_thresh3 = 8192
- net.ipv4.neigh.default.gc_stale_time = 120

Troubleshooting: Software Installation

The most common GM installation failures are summarized in the FAQ:

- http://www.myri.com/cgi-bin/fom?file=46

- Running kernel / source kernel mismatch
- Riser cards
- APIC
- AGP (nVidia, ATI) conflicts

Undoubtedly, if you encounter an issue on a specific mother board or version of Linux, someone else has too, and it will be documented on the Myricom web site. If not, contact us at **help@myri.com.**

Enabling IP over Myrinet (Ethernet Emulation) (OPTIONAL)

2. Enabling IP over Myrinet

The Linux command to enable IP over GM/MX (ethernet emulation) is as follows:

/sbin/ifconfig myri0 <ip_address> up

where you must replace *myri0* with the appropriate name (myri1, myri2,etc.) if you have more than one Myrinet NIC per host.

Software Testing and Validation

3. Testing/Validation

Once the MX/GM software is properly installed on all hosts/nodes in your cluster, you are ready to "validate" your Myrinet installation. These steps are described in detail in the "Troubleshooting" section of the FAQ.

- Check the LEDs on each switch port and NIC port
- Run mx_info or gm_board_info on each host
- Run mx_dmabench or gm_debug to test the PCI bandwidth
- Run mx_pingpong or gm_allsize to test the links in the network
- Run the MX Test Suite or gm_stress to test the network

If you can follow the steps outlined in this entry, you will have a solid Myrinet installation.

Check the LEDs on each switch port and NIC port

- After the hardware installation and GM/MX software installation have been completed, there will be a green LED and flashing yellow/amber LED illuminated on each NIC, and a green LED illuminated on each connected switch port.

- If a green LED is not illuminated on each connected NIC and switch port, check that the Myrinet cable is firmly attached both at the switch end and at the NIC end.

- Switch Tutorial Troubleshooting
  - http://www.myri.com/scs/14U_switches/

- Diagnostics: How do I isolate a failed component?
  - Damaged cable?
  - Failed port on a NIC?
  - Failed port on the switch?

Run mx_info or gm_board_info on each host

If all of the LEDs are illuminated, you can now test that the MX/GM mapper has correctly detected all of the hosts in your Myrinet network by typing the following command on one host:

```
cd <install_path>/bin/
  ./mx_info
```

or

```
cd <install_path>/bin/
./gm_board_info
```

If mx_info lists routing information for all of the nodes, then you are ready to communicate over the Myrinet fabric.

If gm_board_info reports **Network is fully configured**, then you are ready to communicate over the Myrinet fabric.

Run mx_info or gm_board_info (cont'd)

- If you see ??? in the output of gm_board_info, refer to the FAQ entry entitled "When using GM-2, what do the ??? signify in the output of gm_board_info?" (http://www.myri.com/cgi-bin/fom?file=318)
- If the list of nodes (routing table) is not complete (that is, it does not contain all of the nodes in the cluster) or you see **Network is not fully Configured**, you will need to determine why some nodes are not listed.
- Refer to the FAQ entry entitled "How do I tell if the GM-2 Mapper has correctly detected all of the hosts in the Myrinet network?" (http://www.myri.com/cgi-bin/fom?file=273) for details.
  - Try gm_board_info again until you see the **Fully configured** message.
  - Or you could use **gm_board_info --wait** which will not return until all of hosts have been mapped.

Run gm_board_info (cont'd)

– If you continue to see **Network is NOT fully configured**, then it sounds like one of the nodes doesn't have a complete map. Please run:

```
gm_board_info --no-yp | fgrep "Map version is"
```

on all nodes and make sure they all report the same map version. If not, there is a likely a flaky or down link preventing complete network configuration. This condition will also cause

```
gm_board_info --no-yp | fgrep fully
```

(on any node) to report the network status as **NOT fully configured**, or by

```
gm_board_info --no-yp --wait
```

waiting forever.

Run gm_board_info (cont'd)

- If not all of the hosts are listed, then it is possible that a cable is not connected, or GM is not properly loaded on all hosts in the Myrinet network. A green LED should be illuminated on the switch for each connection that is active.
- Do you see ??? in the output of **gm_board_info**? If yes, refer to the FAQ entry "When using GM-2, what do the ??? signify in the output of **gm_board_info**?" (http://www.myri.com/cgi-bin/fom?file=318) for further information.
- Do you see any error messages in the kernel log **/var/log/messages**?
- Have you tried running the GM-2 Mapper in verbose mode to obtain additional information? Refer to the FAQ entry "How do I obtain verbose output from a running GM-2 Mapper?" (http://www.myri.com/cgi-bin/fom?file=267) for further information.

Run gm_board_info (cont'd)

- If you see **Network is fully configured**, but not all of the hosts have the same set of nodes, then check if there is a gm_mapper daemon running on all hosts. Run

```
ps -auxw | grep gm_mapper
```

on all hosts to verify this. If any hosts do not have an active mapper running, restart the gm_mapper on that host and check the system log for mapper errors.

If all of the hosts are running mappers, then run the mapper in verbose mode to try to determine what the problem is. Do this by running:

```
gm_mapper --level=2 -v
```

Run mx_pingpong or gm_allsize to test links in the network

- The test program **mx_pingpong** tests the PingPong latency and bandwidth between two hosts, as described in {MX_HOME}/tools/README.

- The test program **gm_allsize** can be used to measure latency and bandwidth, as described in {GM_HOME}/tests/README.

Run the MX Test Suite or gm_stress to test the network

- Run the MX Test Suite in **<install_path>/bin/tests**

- The test program **gm_stress** is designed to stress the network.
  - Full details of how to run **gm_stress** can be found on the FAQ:
    - http://www.myri.com/cgi-bin/fom?file=53
  - Can be run on a subset of nodes or the whole cluster

- What happens when **gm_stress** fails?
  - http://www.myri.com/cgi-bin/fom?file=54
    - Connectivity problem
    - User error
    - Congestion

How do I run gm_stress?

- Points to note:
  - By default, **gm_stress** runs all-to-all communication on the entire cluster.
  - Use the **-f** run-time option to test on a subset of hosts.
  - **gm_stress** must be invoked on all hosts involved in the test.
  - Rsh/ssh must be used to launch **gm_stress** on all hosts involved in the test
  - If you use run-time options to **gm_stress**, they must be the same on all hosts involved in the test.
  - When **gm_stress** terminates on the local host, the test has completed on all hosts involved.
- Example usage:
  - pshd <install_path>/bin/gm_stress -f hostfile

Miscellaneous Troubleshooting Hints

Troubleshooting: IOMMU Limitation on pSeries mother boards

- With Linux, the IBM pSeries hardware limits the amount of IOMMU space that can be mapped/pinned to 256 MB (or 128MB) per each PCI slot. This limitation is enforced by the hardware, and cannot be overridden by our Myrinet GM/MX driver.
- This small IOMMU affects the customer in 2 ways:
  - It limits the number of processes (GM software ports) that can be spawned per processor/NIC at 6 processes per processor.
  - If the application sends large messages, this IOMMU limit will reduce the performance of the application.
- For further details, refer to the Myrinet FAQ entry:
  - http://www.myri.com/cgi-bin/fom?file=384

Troubleshooting: SRAM parity errors

- For the PCI-X NICs, this SRAM parity error is a transient error to be expected occasionally. The errors are caused by high-energy particles that can change the state of memory bits. These and other errors also occur in processors, cache memory, main memory, and IO buses.
- The detection of parity errors in the NIC SRAM is performed to protect the computation from errors.
- In normal environments, the expected rate of parity errors due to particle-induced upset events within the NIC SRAM is approximately one per 100,000 hours. If the cluster is located at a high elevation or near a source of high-energy particles, this would increase the rate at which they are detected.

Troubleshooting: SRAM parity errors

- If an SRAM parity error occurs, you will see a GM error message similar to the following in the kernel log (/var/log/messages) on the affected host.

GM: LANAI[0]: PANIC: mcp/gm_parity.c:114:parity__int():firmware
GM LANAI[0]: LANai detected a PARITY error in board SRAM at rtc time = 0x3b6bda4d299 probably at memory location 0x1ccc9
GM: LANAI[0]: WARNING: libgm/gm_abort.c:44:gm_abort():firmware
GM: LANAI[0]: Program aborted.
GM: NIC firmware error: GM aborted
GM: WARNING: libgm/gm_exit.c:59:gm_exit():kernel
GM: gm_exit() called in the kernel: GM aborted<4>GM: NOTICE: drivers/gm_lanai_command.c:79:gm_lanai_command_report_error():kernel
GM: LANai[0] interface not responding to command 6

Troubleshooting: SRAM parity errors

- If you ever see a GM error message in the kernel log which says that the “interface is not responding”:

  GM: LANai[0] interface not responding to command ...

  Then you need to examine all messages in the kernel log from the time that GM was loaded until you see this message. This message means that the firmware has crashed, and you will need to determine why.

  If it is an SRAM parity error, then you need to reboot the host. Reloading the GM driver is not sufficient.

Troubleshooting: SRAM parity errors

- Please keep us informed of how many SRAM parity errors you see reported. You will need to reboot the machine after it encounters one of these errors.
- If you ever observe two errors in the same NIC, we would like that NIC to be RMA'ed to Myricom for examination.
- Myricom is working on software/firmware workarounds to allow the firmware to deal with SRAM parity errors.
- The MX software has the capability of transparently reinitializing the NIC firmware and data structures. When an SRAM parity error occurs, in many cases it is recoverable. In some case, however, it will not be possible to recover and a reboot of the host will still be required for security reasons.

Troubleshooting: GM NOTICE message

- If you see a GM NOTICE message similar to the following in the kernel log:

  GM: NOTICE: drivers/gm_port_state.c:348:gm_port_state_close():kernel

  GM: There are 1 active subports for port 2 at close

  these message indicate that communication was in progress when one of the processes involved in the message died/disappeared unexpectedly. The messages are not the cause of the problem; they are a side effect.
  - Do you see any GM error messages in the kernel log preceding this message?
  - Is it a specific application that causes these messages to appear, or could it be a variety of applications?
  - If it's a specific application, are you sure that one of the MPI processes isn't segfaulting on one of the nodes? Have you verified this with gdb?

Troubleshooting: Performance

- Check for **badcrcs** in the **mx_counters** or **gm_counters** output, as well as the hardware counters on the switch.

  If you start to see a large number of **badcrcs** (hundreds, thousands), then you may have a flaky hardware component (cable, port on switch, or port on NIC) that needs to be isolated.

  Assuming you have the monitoring line card configured, you could use the **Filter** feature to flag **badcrcs** in the switch

  http://www.myri.com/scs/14U_switches/index-overview-web.html#filter

  and then check the **Log** for any **badcrcs** reported.

  http://www.myri.com/scs/14U_switches/index-overview-web.html#log

Troubleshooting: Performance

- Run **mx_dmabench** or **gm_debug -L** on each node in the cluster to ensure that all nodes report consistent performance.
  - If performance is slower on a specific node, reseat the Myrinet NIC in that host and check the speed settings for the PCI slot.
- Run a sample benchmark (e.g., HPL) (1 node run) on each of the nodes in the cluster to ensure that all nodes report consistent performance. If not, there could be an issue with a particular CPU on one of the hosts.
- Run a sample benchmark (e.g., HPL) on subsets of nodes. Make sure that performance is consistent across all of the subsets of nodes. If you see a particular subset that is slower, then you need to perform a divide-and-conquer approach to isolate the slower component.

Myrinet Monitoring

Why Monitoring?

- Performance
  - How good is the topology?
    - Not necessary in Clos networks (boringly uniform).
  - Finding hot spots in network
  - Observing communication patterns of applications.
- Troubleshooting
  - Cable problems
  - Component failure (rare)
    - Bad port
    - Bad connector
    - Machine room A/C failure.
      - Overtemp shutdowns

Choices for Monitoring the Myrinet Network

- **HTTP** access to the switch
  - Tutorial for the Web Interface
    - http://www.myri.com/scs/14U_switches/
- **TFT Color Display** to the switch
  - Tutorial for the TFT Color Display
    - http://www.myri.com/scs/14U_switches/
- **SNMP** interface to the switch
  - http://www.myri.com/cgi-bin/fom?file=383
- **FMS (Fabric Management System)** **coming soon**
  - http://www.myri.com/scs/fms/

Fabric Management System

- The Fabric Management System (FMS) is a collection of tools and processes used to manage a Myrinet network.
  - **fm_init_db** -- explores a network and creates the database files.
  - **fm_watch_switches** -- periodically monitors a list of switches and reports items of interest, such as badcrc counts.
  - **fm_linktest** -- tests every switch-to-switch link in the network, reporting which ones are missing or marginal.
- FMS will be supported on Myrinet-2000 M3-E* and M3-CLOS-ENCL/M3-SPINE-ENCL switches.
- Full details available on the FMS webpage
  - http://www.myri.com/scs/fms/

fm_init_db

- Command line syntax:

```
fm_init_db -s <switch_list> -m <map_file>
          [ -H <fms_home> ] [ -N <db_name> ]
```

  where <switch_list> is a file which one switch name or IP address per line, and <map_file> is a map file generated by gm_mapper.
- No mappers can be running when this command is issued.
- This tool creates the following database files:
  - enclosures.csv -- a list of all enclosures in the system
  - linecards.csv -- a list of all linecards, and where they are
  - hosts.csv -- a list of all hosts in the fabric
  - nics.csv -- a list of all Myrinet NICs in the fabric
  - links.csv -- a list of all of the fiber connections in the system

fm_watch_switches

- **fm_watch_switches** monitors each crossbar port for:
  - badcrcs -- the number of badcrcs seen
  - rx_timeouts -- the number of receive timeouts
  - tx_timeouts -- the number of transmit timeouts
- **fm_watch_switches** can also be used to report which ports, either transceiver or crossbar ports, have been manually disabled.
- Command line syntax:

```
fm_watch_switches
     [ -t <threshold> ] reporting threshold, default 5
     [ -I <interval> ]  polling interval in seconds, default 30
     [ -g ]             watch goodcrc counts
     [ -d ]             report disabled links
     [ -H <fms_home> ] [ -N <db_name> ]
```

fm_linktest

- **fm_linktest** tests every switch-to-switch link in the system, reporting which ones are missing or marginal.
- Command line syntax:

```
fm_linktest
     [ -B <board_no> ] which NIC to use
     [ -i <host_ifc> ] which interface out of NIC to use, default 0
     [ -l <kbytes> ]   length of test packets in kb, default 4
     [ -H <fms_home> ] [ -N <db_name> ] home and DB name
```

- **Example Usage**: **fm_linktest** should be run periodically as a health check of the fabric.

Myrinet Technology Roadmap

Myrinet Technology Roadmap

- **Myrinet-10G**
  - 4th-generation Myrinet, products to be announced in 2Q05
  - 10+10 Gb/s data rate links
  - Interoperable with 10-Gigabit Ethernet
  - Fully compatible with Myrinet-2000
    - Same architecture, same software, same APIs, same applications
- **Myrinet Express (MX) software**
  - Message-passing system for initial Myrinet-10G products
  - MX-2G-Beta is already available for Myrinet-2000 D, E, & F cards
    - Myricom software support has always spanned two generations of NICs

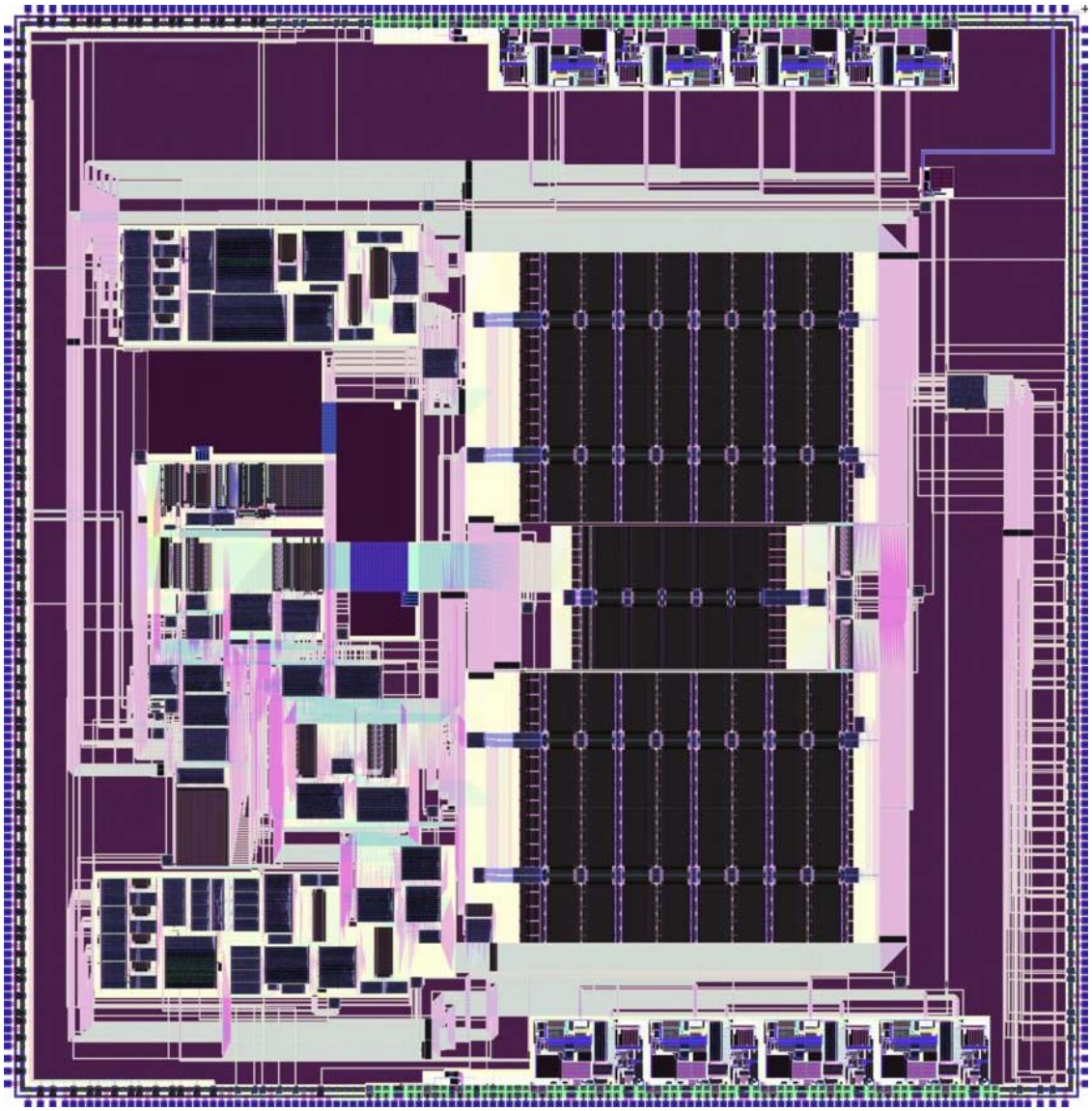
We expect Myrinet-2000 products to continue to sell well through 2006, and to co-exist with the higher priced Myrinet-10G products. Myrinet-2000 is well positioned as a superior alternative to Gigabit Ethernet for clusters, while Myrinet-10G offers performance and cost advantages over 10GbE for clusters.

Myrinet-10G Links, NICs, and Switches

- Myrinet-10G links
  - **Any PHY (cables, connectors, signaling) that can be used for 10-Gigabit Ethernet (10GbE) can be used for Myrinet-10G**
    - The ports of Myrinet-10G chips are XAUI, per IEEE 802.3ae
- Myrinet-10G NICs
  - Even more highly integrated than Myrinet-2000 PCI-X NICs
  - PCI Express, HyperTransport, and other fast host buses
  - **Any Myrinet-10G NIC is also a 10GbE NIC**
    - The standard firmware supports both Myrinet-10G and 10GbE modes
- Myrinet-10G switches
  - Similar except for data rates to Myrinet-2000 switches, based initially on a Myrinet-10G 16-port crossbar-switch chip

The Initial 10G-Ethernet/10G-Myrinet NIC

- PCI Express
- Based on the Lanai ZE chip (on right)
  - Internal packet buffers as well as external SRAM
- 333MHz processors and external memory
- Self-initializes to 10GbE mode, but operates also in Myrinet-10G mode
- 9.5 Gbits/s TCP/IP rate (GbE mode)
- 2µs MPI latency with MX-10G, and 1.2 GBytes/s one-way data rate (Myrinet mode)



Conclusions

- Hardware Product Specifications
  - http://www.myri.com/myrinet/product_list.html
- Hardware and Software documentation is available
  - http://www.myri.com/scs/#documentation
- Troubleshooting?
  - Linux Download Page (http://www.myri.com/scs/linux-gm2.html)
  - Switch Tutorial (http://www.myri.com/scs/14U_switches)
  - Myrinet FAQ (http://www.myri.com/scs/FAQ/)
  - README in the MX software distribution tar file.
  - README-linux in the GM software distribution tar file.
  - Send email to Technical Support at **help@myri.com**.
- Questions?