Innovations in Cluster Computing

Charles L. Seitz
President & CEO of Myricom, Inc.
chuck@myri.com

13 March 2007
RIKEN HPC Symposium

New Directions for Myricom

- Although Myricom has done well (and done some good) with Myrinet in the HPC market, this market is limited
- We foresee little future for “specialty networks”
  - Forces of technical convergence/standardization and business consolidation in the computer industry over the past decade
  - Thus, new directions for Myricom (developments started in 2002)
- Myricom has great technology for 10-Gigabit Ethernet
  - And Myricom products have always installed like Ethernet, carried Ethernet traffic, and been interoperable with Ethernet
- Thus, ***Myri-10G***, our 4th generation of high-performance networking products, has converged with Ethernet
  - Diversification strategy: Dual-use 10G Ethernet & 10G Myrinet
  - Programmable NICs are important for both modes of operation

Myri-10G is …

- ***4th-generation Myricom products,*** a *convergence* at 10-Gigabit/s data rates of Ethernet and Myrinet
  - Based on 10G Ethernet PHYs (layer 1), 10 Gbit/s ***data rates***
  - Network Interface Cards (NICs) support both Ethernet and Myrinet network protocols at the Data Link level (layer 2)
- ***10-Gigabit Ethernet products from Myricom***
  - High performance, low cost, fully compliant with IEEE 802.3ae, interoperable with 10G Ethernet products of other companies
- ***4th-generation Myrinet***
  - A complete, low-latency, cluster-interconnect solution – NICs, software, and switches – software-compatible with Myrinet-2000
  - Switches retain the efficiency and scalability of layer-2 Myrinet switching internally, but may have a mix of 10-Gigabit Myrinet and 10-Gigabit Ethernet ports externally

Myri-10G for HPC Clusters

A technically superior cluster-interconnect solution:

- 2.3µs MPI latency (including one Myrinet switch)
- 1.2 GByte/s MPI PingPong data rate
  - Nearly 2.4 GByte/s MPI SendRecv data rate
- High application availability with MPI (and Sockets)
  - Low host-CPU load; excellent computation-communication overlap
- Small and constant memory footprint in the host
- If higher data rates are required, NIC bonding available
  - 2.3 GByte/s MPI data rate with 2 NICs, 4.5 GByte/s with 4 NICs
- Native support for popular cluster file systems (Lustre, PVFS2)
- Carries TCP/IP and UDP/IP traffic at >9.6 Gbits/s
  - Highly interoperable with 10-Gigabit Ethernet
- ***Main announcement today:*** New series of modular switches

Myri-10G NICs

10GBase-CX4

10GBase-R

XAUI over ribbon fiber

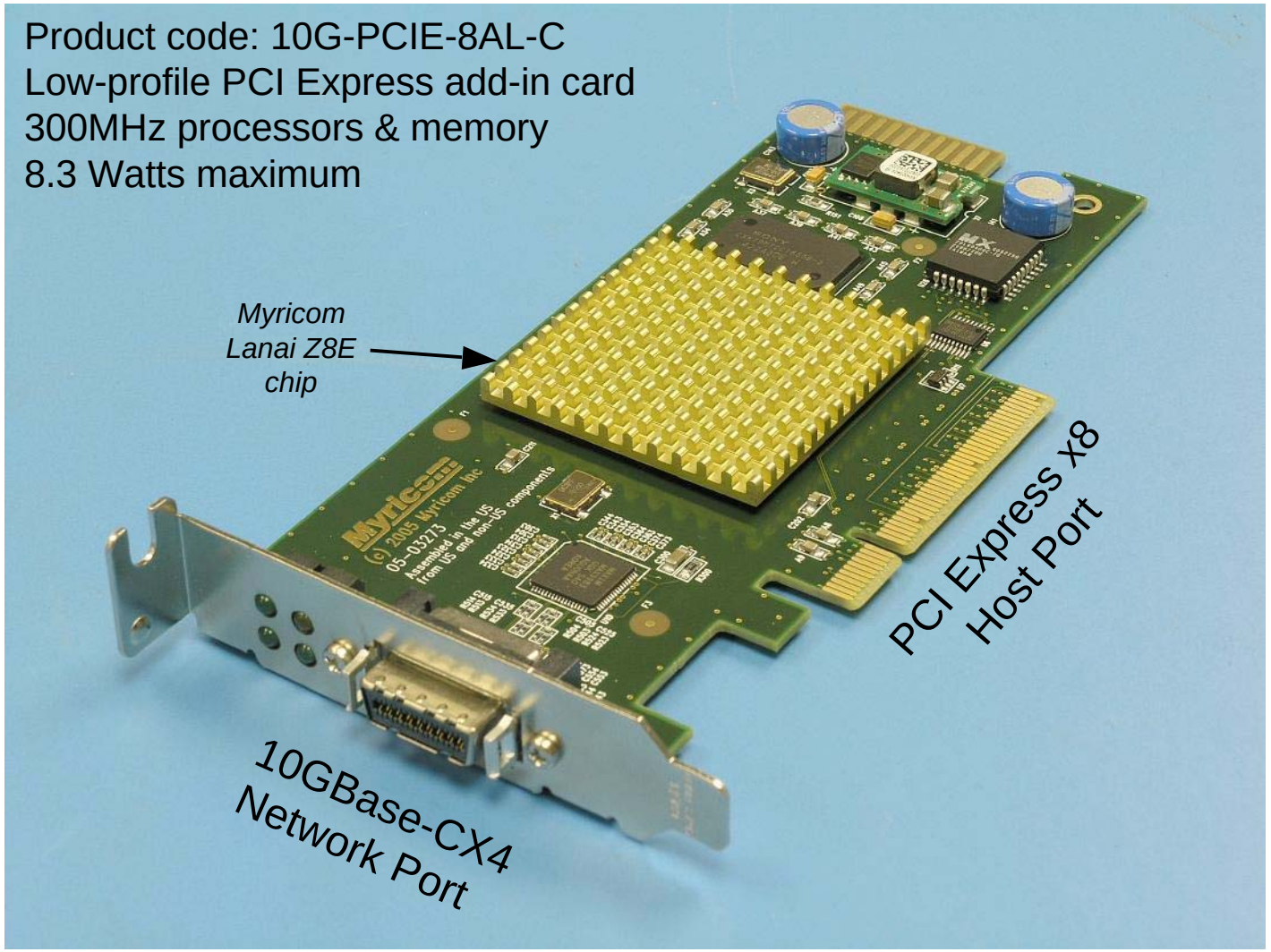
These NICs are PCI Express x8, and are based on the Myricom Lanai-Z8E chip

Protocol-offload 10-Gigabit Ethernet NICs. Use them with Myricom's bundled driver and a 10G Ethernet switch. You'll see >9.6 Gbit/s TCP/IP or UDP/IP data rates (Linux 2.6, netperf benchmark, jumbo frames).

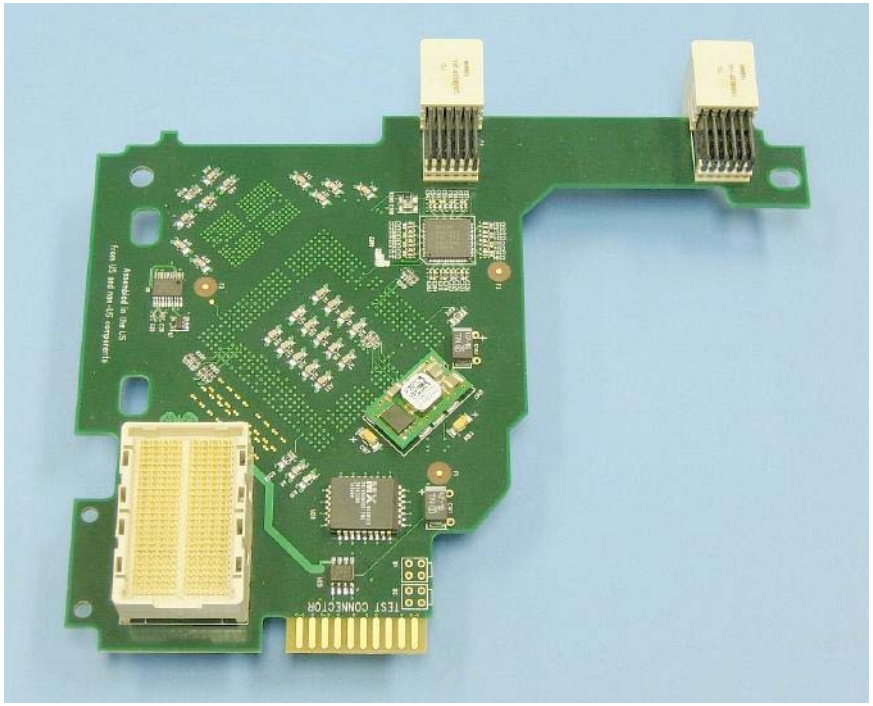
10-Gigabit Myrinet NICs. Use them with Myricom's MX (Myrinet Express) software and a 10G Myrinet switch. You'll see performance metrics of:

- 2.3µs MPI latency
- 1.2 GBytes/s data rate
- Very low host-CPU utilization

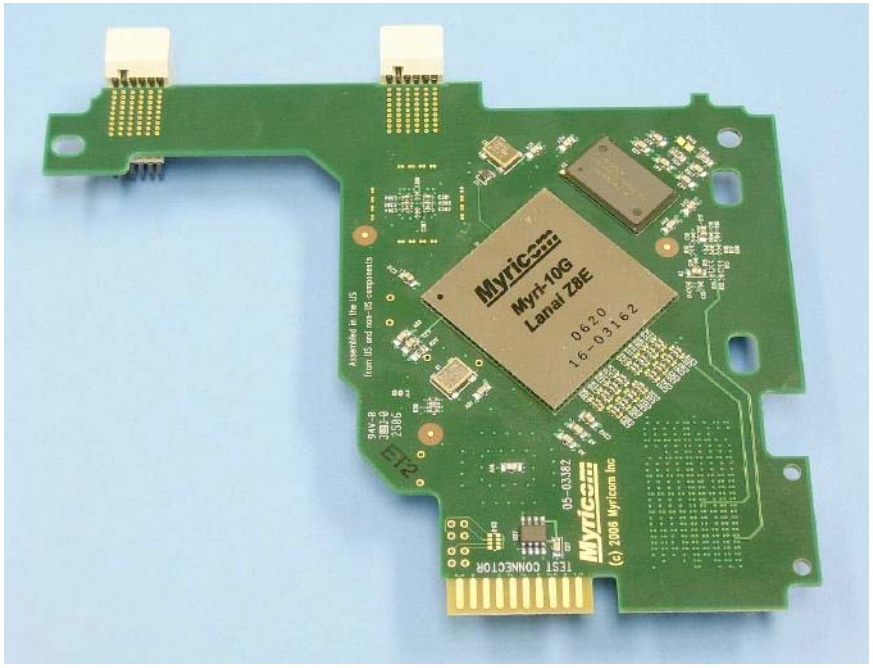
Closer View of a Myri-10G NIC



Myricom also makes special NICs using the Lanai Z8E chip, for example …



Bottom View



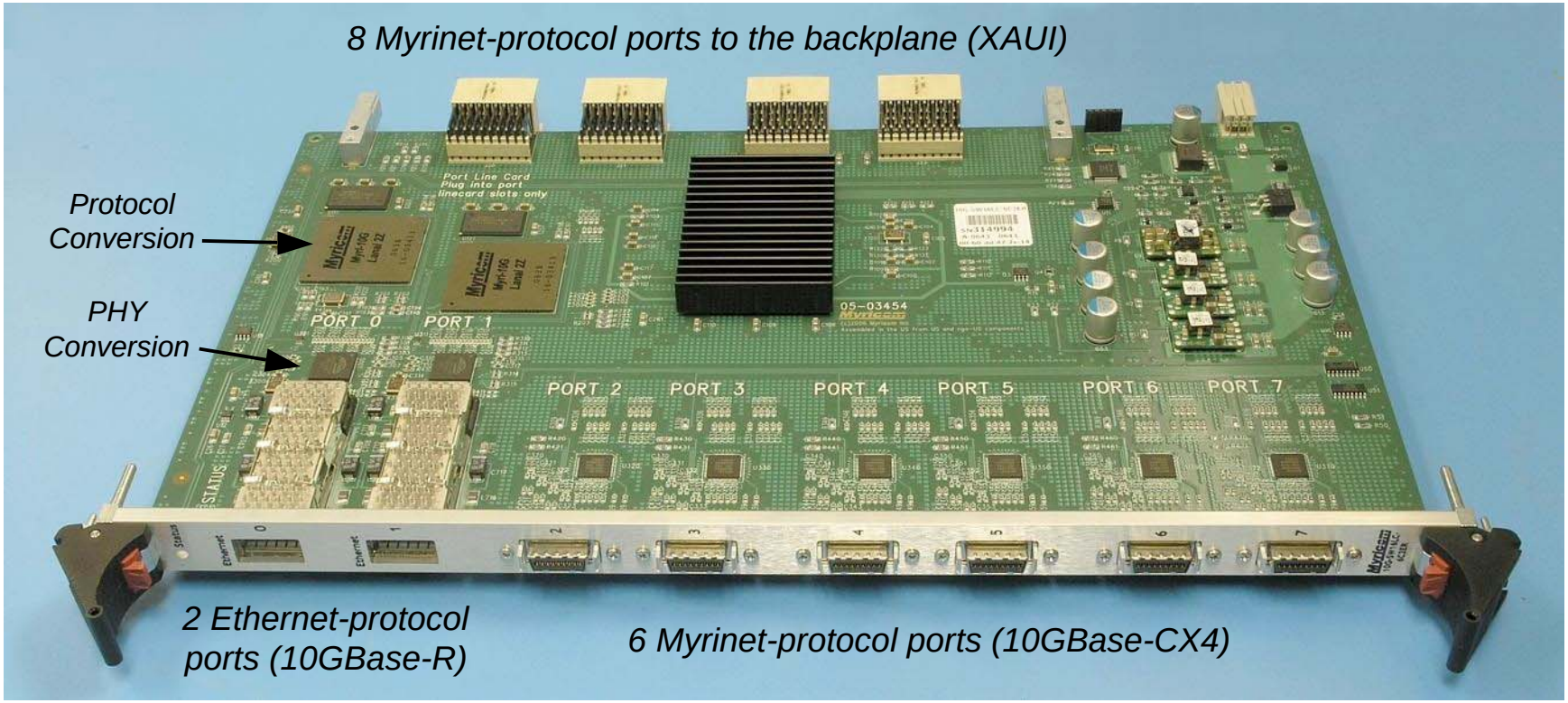
Top View

IBM BladeCenter H “HSEC”
(High Speed Expansion Card)

Myri-10G Switches

- 10-Gigabit Myrinet
  - Similar except for data rates to Myrinet-2000 switches
  - Cut-through switching with source routing
  - Low latency, economical, and scalable to thousands of hosts using efficient, layer-2 switching
- Full-bisection Clos networks
  - Modular packaging
- Multi-protocol switching
  - Connection of a 10-Gigabit Myrinet switch fabric to interoperable 10-Gigabit Ethernet ports is performed by protocol conversion on line cards (next slide ☞)

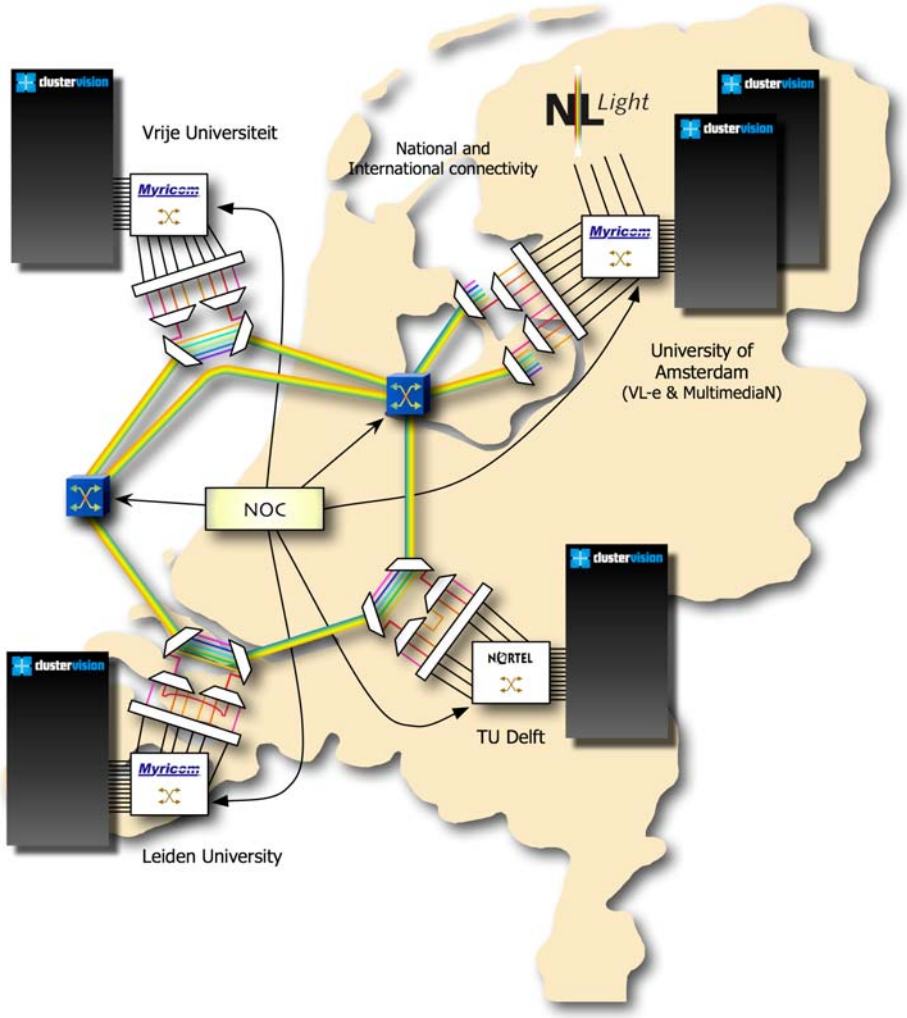
Protocol Conversion on a Line Card



Photograph of a 10G-SW16LC-6C2ER Myri-10G Switch Line Card

DAS-3: High-Performance Interoperability

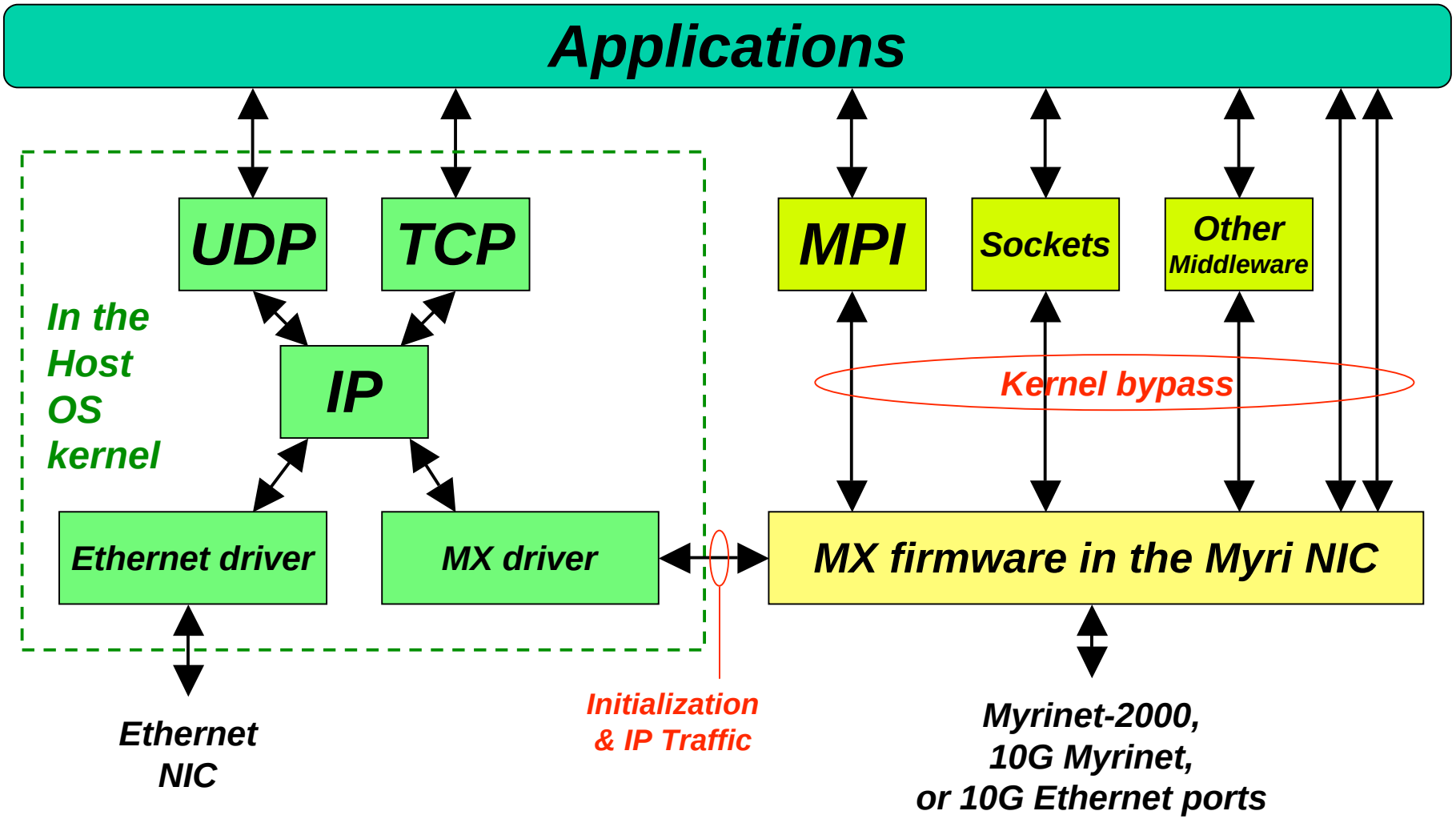
- Grid of clusters in the Netherlands
- Five supercomputing clusters connected by a private DWDM fiber network
- “Seamless” cluster and grid operation thanks to Myri-10G
  - Myrinet protocols within each cluster; IP protocols between clusters



Myri-10G Software

- Driver and firmware for 10-Gigabit Ethernet operation (Myri10GE software distribution) is included with the NIC
- The broader software support for 10-Gigabit Myrinet and Low-Latency 10-Gigabit Ethernet is MX (Myrinet Express)
  - MX-10G is the message-passing system for low-latency, low-host-CPU-utilization, kernel-bypass operation of Myri-10G NICs over either 10-Gigabit Myrinet or 10-Gigabit Ethernet
  - MX-2G for Myrinet-2000 PCI-X NICs was released in June 2005
    - Myricom software support always spans two generations of NICs
    - MX-2G and MX-10G are fully compatible at the application level
  - Includes TCP/IP, UDP/IP, MPICH-MX, and Sockets-MX
    - MPICH2-MX coming soon. Also available: OpenMPI, HP-MPI, …
  - Also includes cluster file systems directly over MX
    - Lustre-MX, PVFS2-MX, others in progress

MX Software Interfaces

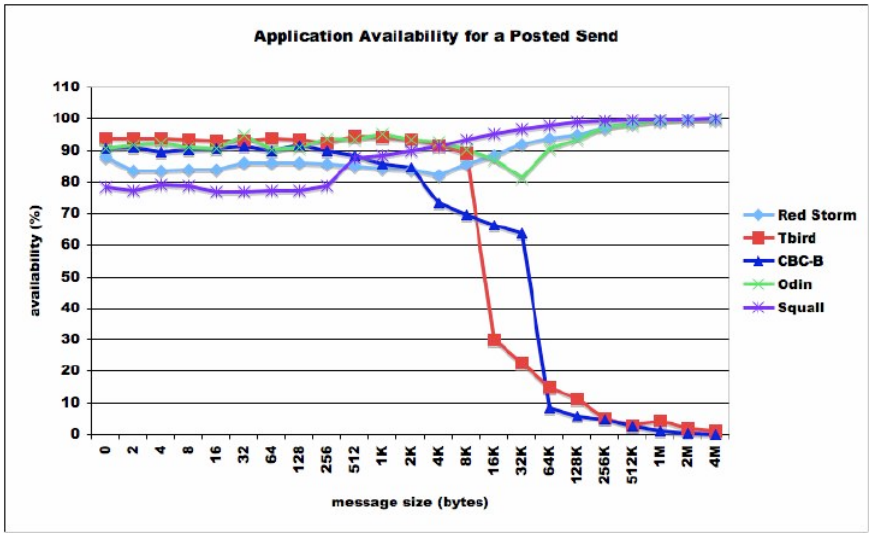


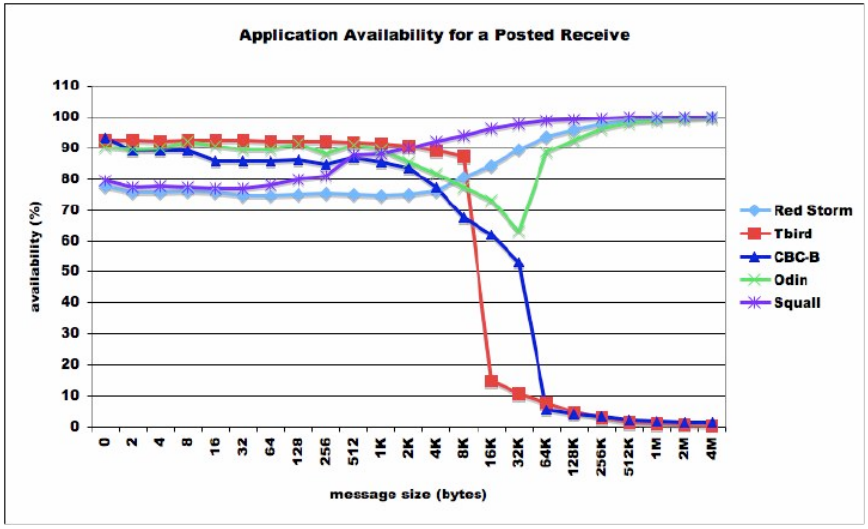
Myri-10G Software Matrix

Host APIs	protocols	Driver & NIC firmware	Network protocols	Network
IP Sockets	TCP/IP, UDP/IP host-OS network stack	Myri10GE	IPoE IP over Ethernet	Ethernet
IP Sockets	TCP/IP, UDP/IP host-OS network stack	MX-10G	IPoE IP over Ethernet MXoE MX over Ethernet	
Sockets over MX + MPI over MX	MX kernel bypass		IPoM IP over Myrinet MXoM MX over Myrinet	Myrinet

Note that the MX-10G software distribution is a superset of the Myri10GE software

Communication/Computation Overlap





The ability of applications to execute concurrently with communication is crucial to the performance of many production computing tasks.

The graphs on the left are from "Measuring MPI Send and Receive Overhead and Application Availability in High Performance Network Interfaces," Doerfler and al (Sandia), *EuroPVM/MPI, September 2006*

The graphs show good results for Myri-10G, Quadrics, and Cray SeaStar, and poor results for InfiniBand.

Red Storm: Cray SeaStar/Portals
Tbird: Cisco (Mellanox) Infiniband/MVAPICH
CBC-B: Qlogic Infinipath
Odin: Myricom Myri-10G/MPICH-MX
Squall: Quadrics QsNetII

Myrinet-protocol Switching

Proven Scalability of Myrinet Switching

MareNostrum Cluster in Barcelona. The central Myrinet-2000 switch has 2560 host ports. Photo courtesy of IBM.

The 10G XBar32 Chip

The technology behind the new Myri-10G switch products

- Myrinet-protocol crossbar-switch chip with 32 XAUI ports
- 0.1µs cut-through latency
- 15 Watts
- Fabricated in 0.13µm CMOS on 300mm wafers
- Flip-chip packaging, 912 pins
- Scan path for monitoring packet counts, misaddressed packets, CRC errors, etc.
- Switch-ID feature for mapping acceleration

Starting with these chips, the rest is topology and packaging

New Series of Modular Myri-10G Switches

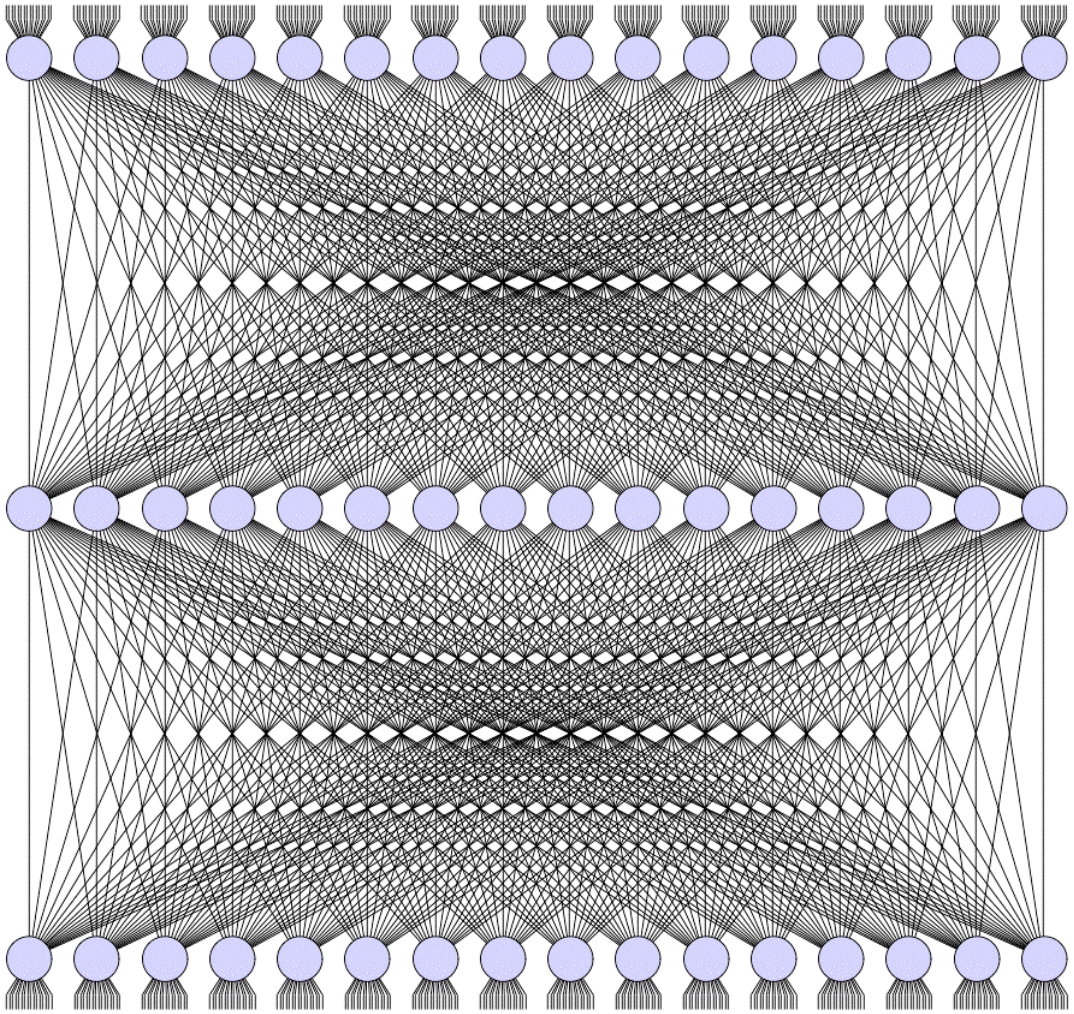
- Based on the low-latency 10G XBar32 chip
  - Standalone 32-port switches
  - Diameter-3 512-port Clos network in a single enclosure
  - Economical diameter-5 switch-network scaling to 8192 nodes
- Horizontal line cards that can be used across the entire family of enclosures from 2U (32 ports) to 21U (512 or 640 ports)
  - “Flagship” is 512-port Clos “Network in a Box,” 21U enclosure
  - Also, 256-port Clos in 12U, 128-port in 7U, down to …
  - Standalone, modular, 32-port switch in 2U enclosure for entry-level or edge-switch applications uses the same line cards
  - Choice of line cards with up to 16 external ports.
    - Choice of Myri-10G PHYs on external ports
    - Line cards with Ethernet-protocol external ports
    - Myrinet protocols inside, Myrinet or Ethernet protocols outside

Topology of the 512-Port Switch Network

Diameter = 3 Clos network of 32-port crossbar-switch chips

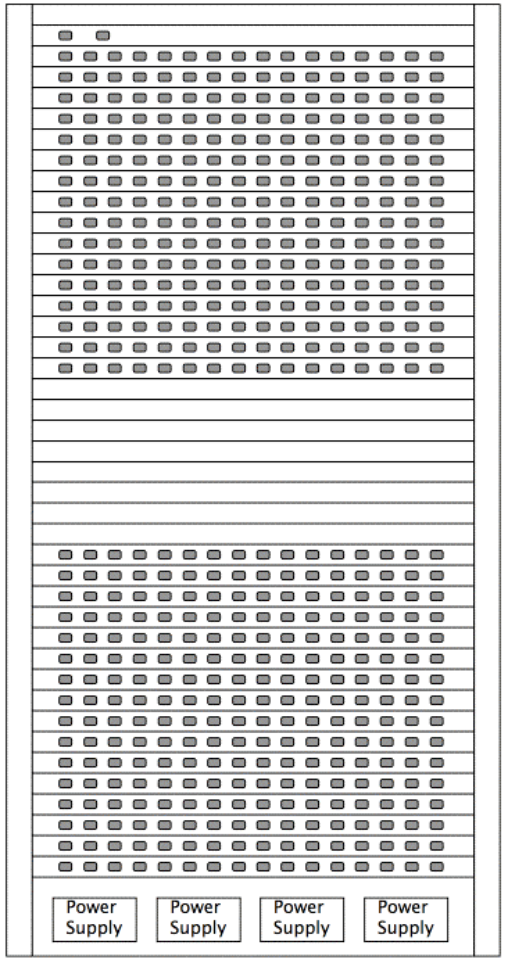
A total of 48 10G_XBar32 chips

PHY and/or protocol conversion at external ports



Packaging of the 512-Port Switch Network

- 21U, 19-inch, rack-mount enclosure, 27 inches deep
  - Similar in style and finish to Myricom's current 14U enclosures
- Front (inlet air) has only hot-swap fans for the line-card section, inlet grill for the power-supply section, and labels
- Rear (exhaust air) has 41 horizontal line cards and up to 4 hot-swap power supplies
  - 40 port line cards and the monitoring line card
  - IEC inlets are on the power supplies, allowing operation from dual AC-power sources.
- The 8 central (10G-2SW32LC) line cards include the 16 XBar32 switches for the spine
- Monitoring via dual-redundant 10/100/1000Base-T Ethernet ports
- 1680 Watts in maximal configuration



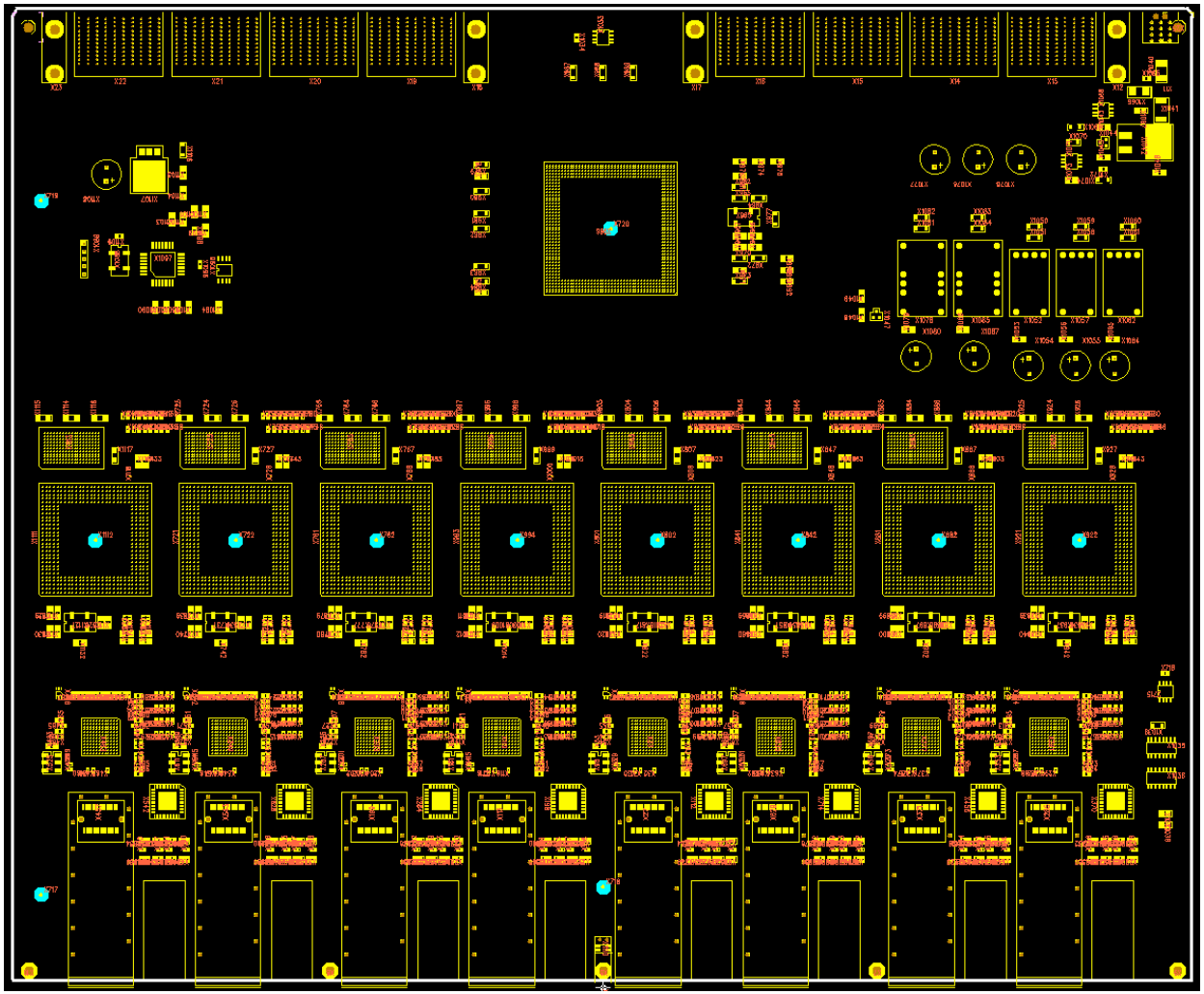
Rear View

21U Switch Enclosure *(without backplanes)*

Front View

Rear View

10G-SW32LC-8M8ER Circuit Board

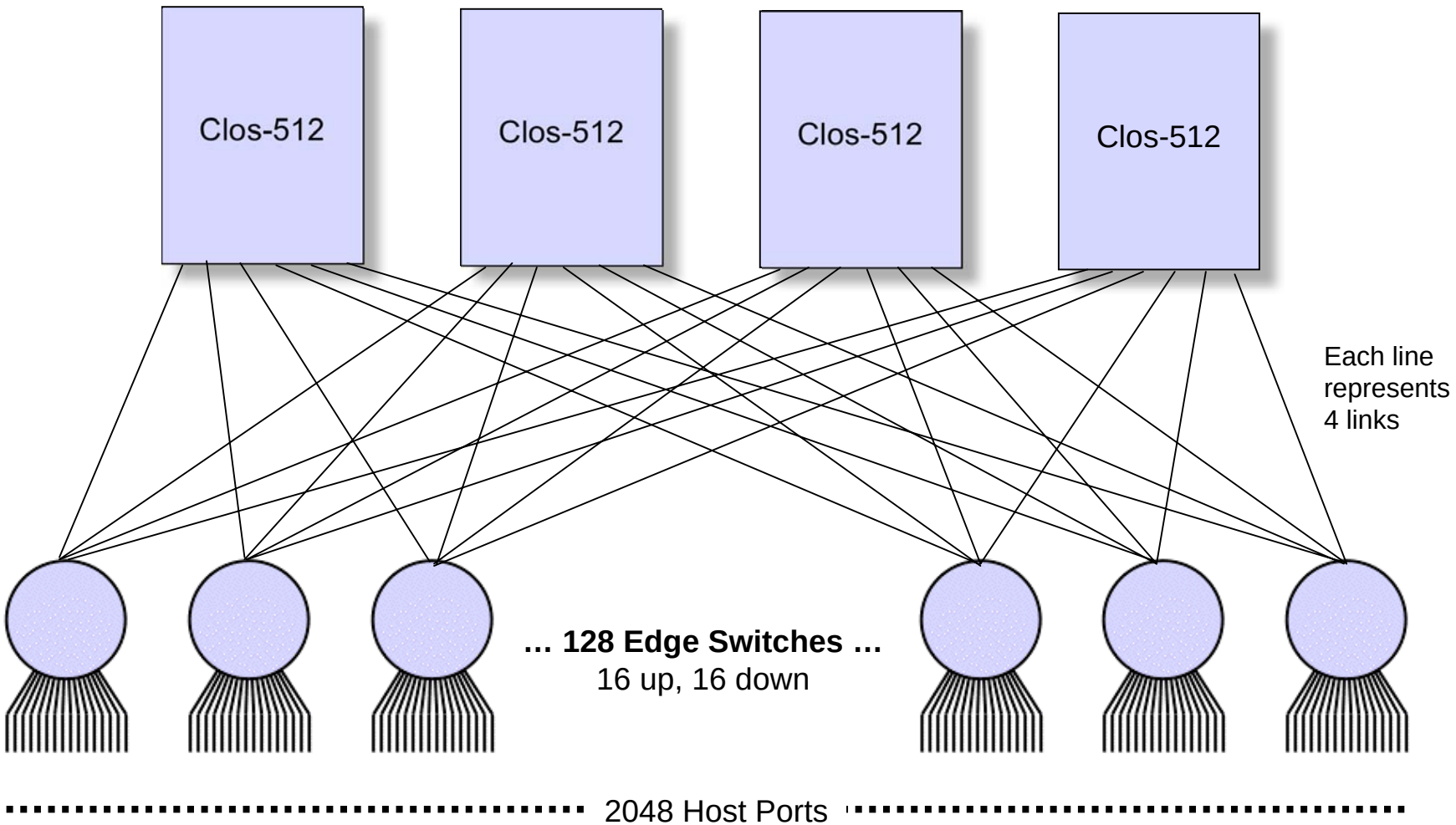


(Part of a packaging feasibility study to optimize the size of the line cards)

Scaling to 8192 hosts

- For 512 < *N* ≤ 8192 with 32-port crossbar-switch chips, the Clos network has a diameter = 5
- One approach is to use multiple 512-port Clos switches (diameter = 3) as the spine of the network
- Use 32-port switches, “16 up, 16 down,” as first-level (edge) switches for each 16 hosts
  - For blade systems, an edge switch may be in the blade chassis
- In a network with *k* spine switches, up to 512 ports each, *~1/k* of the “up” ports of the first-level switches are cabled to each of the spine switches
  - Note that this network is *k*-fold redundant
- Switch-switch links use Myrinet protocols, but connections to hosts may use either Myrinet or Ethernet protocols

Example: “Federated” 2048 Host Ports



An application of large-scale multi-protocol switching