Ethernet Will Eradicate Ethernot

The Rise and Fall of Specialty Networks for HPC Clusters

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

13 September 2006
CCGSC 2006

TOP500 Data

Interconnect Family

Top500 List:
06/2006

Statistics Type:
Interconnect Family

Generate

Interconnect Family	Count	Share %	Rmax Sum (GF)	Rpeak Sum (GF)	Processor Sum
Myrinet	87	17.40 %	369286	560601	100546
Quadrics	14	2.80 %	131249	173218	39108
Gigabit Ethernet	256	51.20 %	795582	1511012	249726
Infiniband	36	7.20 %	221074	316047	53068
Crossbar	12	2.40 %	114397	149382	16674
Mixed	5	1.00 %	71924	91853	15396
NUMAlink	9	1.80 %	55509	61028	9728
SP Switch	42	8.40 %	281098	406406	66036
Proprietary	26	5.20 %	629527	806058	291296
Fireplane	1	0.20 %	2054	3053	672
Cray Interconnect	9	1.80 %	109965	133803	28988
RapidArray	3	0.60 %	8388	10374	2357
Totals	**500**	**100%**	**2790054.02**	**4222834.82**	**873595**

This table is from www.top500.org

The first TOP500 cluster appeared in 1997 (Berkeley NOW, a Myrinet cluster)

Today: 72.8% Clusters! Gigabit Ethernet is the interconnect for more than half of the TOP500 systems.

Disclaimer:
The TOP500 data distorts the true situation in several ways: Architecture classification, the HPL benchmark, vendor “bragging rights” motivations, …

Personal view of how we arrived here (1)

- Before clusters there were MPPs (multicomputers)
  - Multicomputer = message-passing, distributed-memory, concurrent computer (Gordon Bell terminology)
  - Multicomputer interconnect technology from my DARPA-sponsored Caltech research group was used in several DARPA projects and in commercial multicomputers from Intel, Cray, …
- 1991? - Dave Patterson's observation
  - MPP processors become obsolete before the first machine ships
  - Therefore, clusters will displace MPPs
- 1992 - I became a cluster convert
  - Documented in my first "clusters will displace MPPs" paper
- Parallel historical track for Beowulf clusters
  - 1993-1994, Thomas Sterling and Donald Becker

Personal view of how we arrived here (2)

- 1993-1994 - Myricom founded
  - By 8 DARPA researchers from Caltech and USC/ISI
- Early specialty networks for HPC clusters (Myrinet, Quadrics) were multicomputer networks with robust cables and that could operate from host I/O buses
  - Source-routed, cut-through switching, with arbitrary topology
  - Processor and firmware in the NIC
- 1997 - First Myrinet cluster on TOP500
- Nov-03 TOP500 has 193 Myrinet, 26 Quadrics, & 3 IB clusters, the peak for specialty networks
  - 111 GbE clusters
- Jun-06 TOP500: Specialty networks down to ~140, and GbE up to 256

We already realized in 2002 that Ethernet will prevail

- Although Myricom has done well (and done some good) with Myrinet in the HPC market, this market is limited
  - *Some quantification in the next two slides* ☞
- We foresee little future for "specialty networks"
  - Technical convergence/standardization and business consolidation has been evident in the computer industry over the past decade
  - Thus, new directions for Myricom (and for Quadrics)
- 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 new generation of high-performance networking products, is converging with Ethernet
  - Diversification strategy: Dual-use 10G Ethernet & 10G Myrinet
  - Programmable NICs are a feature crucial to both modes of operation
  - The latest Quadrics products are also Ethernet hybrids

HPC Market Data from IDC

IDC HPC Technical Server Forecast

11-Sep-10

WW HPC System Revenue by Applications 2005 - 2010 Forecast ($M)							
	2005	2006	2007	2008	2009	2010	CAGR
BioSci	1,434	1,608	1,786	1,946	2,087	2,219	9.1%
CAE	1,109	1,282	1,468	1,647	1,814	1,978	12.3%
Chem Eng	222	266	314	359	403	447	15.0%
DCC&D	514	558	596	619	628	628	4.1%
Econ Fin	255	302	351	398	442	484	13.7%
EDA	648	740	834	912	978	1,037	9.8%
GeoSci	489	559	630	693	748	801	10.3%
CAD	156	171	184	193	198	199	5.0%
Defense	811	952	1,105	1,270	1,452	1,651	15.3%
Govt Lab	1,376	1,463	1,529	1,576	1,623	1,672	4.0%
Software	20	20	21	21	20	19	-1.3%
Tech Mgmt	102	104	102	96	86	73	-6.4%
Academic	1,700	1,877	2,049	2,201	2,348	2,499	8.0%
Weather	359	391	423	450	474	494	6.6%
Other	3	12	23	35	49	65	82.0%
Total	9,198	10,305	11,417	12,417	13,350	14,265	9.2%

Source: IDC, 2006

Size of the HPC-cluster interconnect market

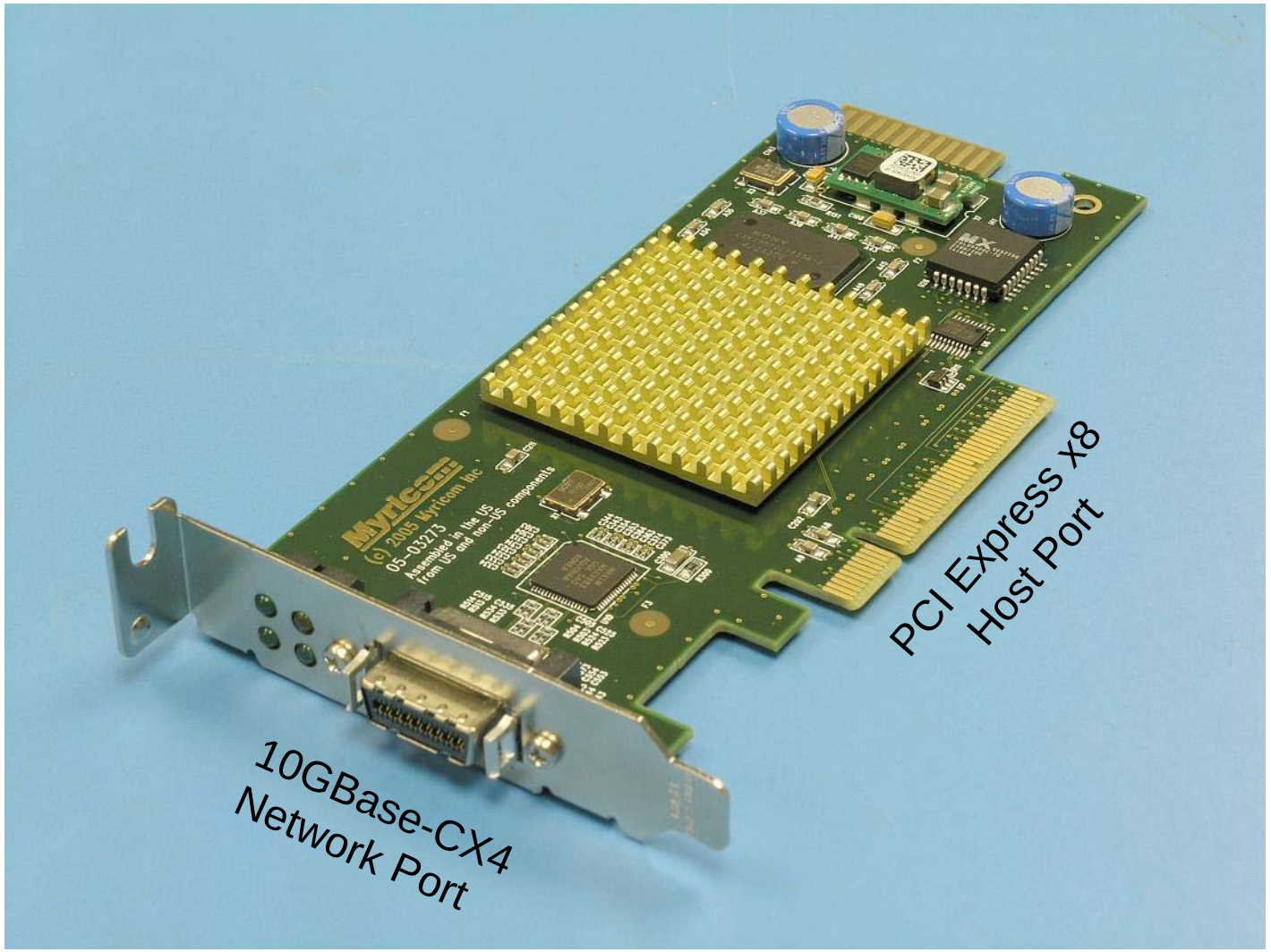
- The IDC data is for servers, software, services, interconnect, …
- IDC estimates that clusters are currently 50% of the market: ~$5B in 2006
- “Tightly-coupled” or large clusters (non-Ethernet) are ~20% of the total: ~$1B/yr
- The interconnect is 10-15% of the cost of a tightly-coupled cluster: $100M-$150M/yr
  - Obviously insufficient revenue to sustain aggressive technical developments just for HPC clusters

- By contrast, the 10-Gigabit Ethernet market is expected to approach ~$1B in 2006 with a CAGR of ~100%.

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***
  - 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

A Myri-10G NIC



Family of Modular Myri-10G Switches

- Up to 128 host ports in the ***Clos*** configuration
- 64 host ports + 64 interswitch ports in the ***Leaf*** configuration
- Mixed PHYs OK ☞
- Enterprise features
  - Redundant hot-swap power supplies and fans
  - Hot-swap line cards
  - Functional and physical monitoring via dual 10/100 Ethernet ports and a TFT display

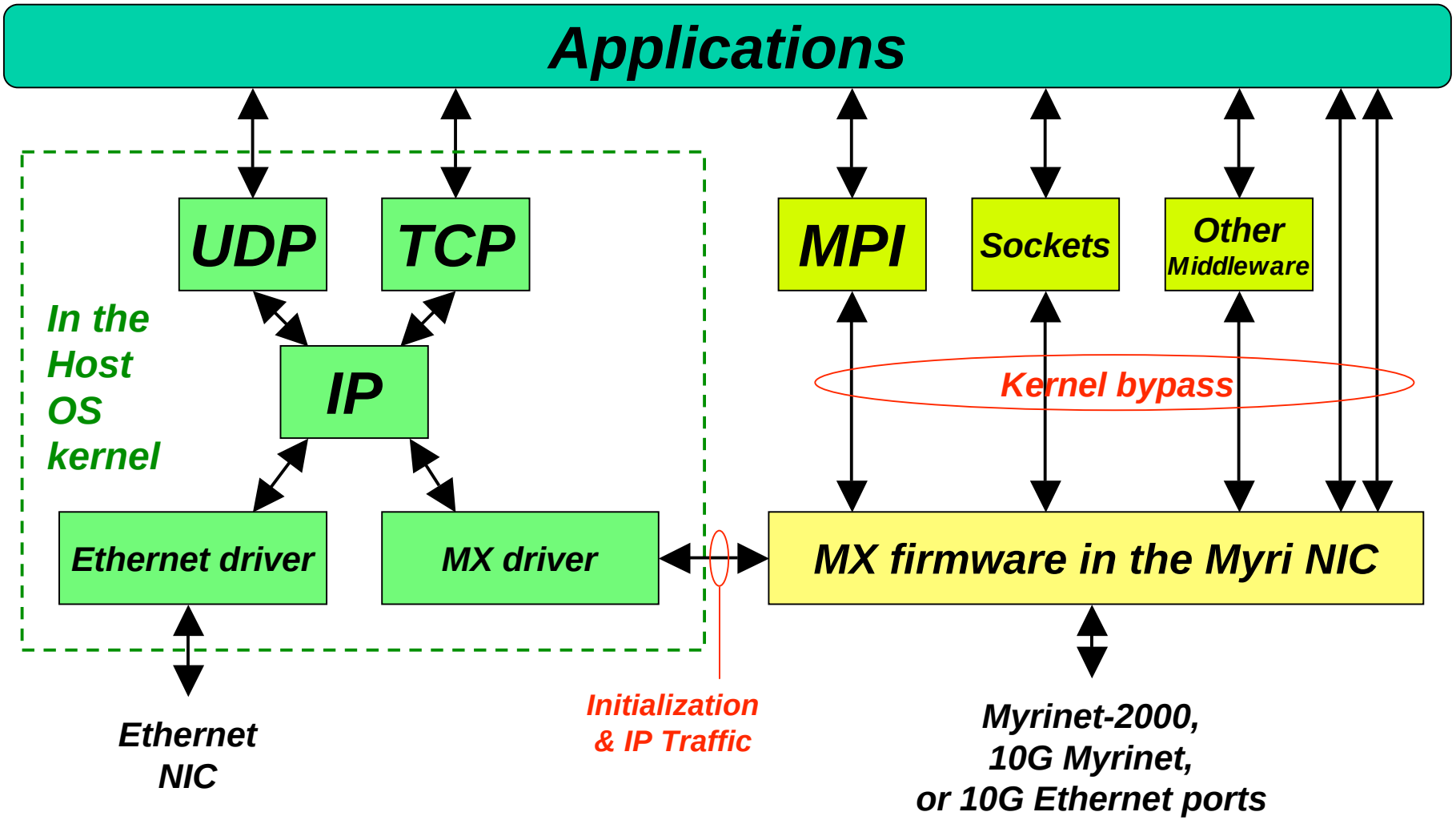
Quick Lesson about 10-Gigabit Ethernet

- Not your grandfather's Ethernet
  - No CSMA/CD. Full-duplex operation only. Designed for switches and fiber. LAN, MAN, or WAN
- External 10GbE PHYs (per IEEE 802.3ae,)
  - 10GBase-{S|L|E}{R|W} for full-duplex fiber, *e.g.,* 10GBase-SR
  - 10GBase-CX4 copper cables to 15m
  - 10GBase-T unlikely to get much traction before advances in fiber components make it irrelevant
- Internal 10GbE Interfaces
  - XAUI (X Attachment Unit Interface) 4x 3.125 GBaud, 8b/10b
    - Blade backplanes, internal to switches
  - XFI (X Fiber Interface) 1x 10.3125 GBaud, 64b/66b
- Otherwise, it's Ethernet
  - The familiar APIs and software

Can we live with 10-Gigabit Ethernet for tightly-coupled clusters?

Low-latency, low-host-CPU-load,
10-Gigabit Ethernet

MX Software Interfaces



Myri-10G Software Matrix

<table>
<tr><th>Host APIs</th><th>protocols</th><th>Driver &
NIC firmware</th><th>Network
protocols</th><th>Network</th></tr>
<tr><td>IP Sockets</td><td>TCP/IP,
UDP/IP
host-OS
network stack</td><td>Myri10GE</td><td>IPoE
IP over Ethernet</td><td rowspan="2">Ethernet</td></tr>
<tr><td rowspan="2">IP Sockets
+
Sockets over
MX
+
MPI over MX</td><td rowspan="2">TCP/IP,
UDP/IP
host-OS
network stack
+
MX
kernel bypass</td><td>MX-10G
for Ethernet</td><td>IPoE
IP over Ethernet
MXoE
MX over Ethernet</td></tr>
<tr><td>MX-10G
for Myrinet</td><td>IPoM
IP over Myrinet
MXoM
MX over Myrinet</td><td>Myrinet</td></tr>
</table>

For MX, we have 2 APIs (Sockets, MPI) **x** 2 protocols (IP, MX) **x** 2 networks (Ethernet, Myrinet) = 8 possible combinations, all supported and all useful.

MX over Ethernet

- ***Myricom recently extended MX-10G to operate over 10-Gigabit Ethernet as well as 10-Gigabit Myrinet***
- MXoE works with Myri-10G NICs (kernel bypass) and standard 10-Gigabit Ethernet switches
- 2.4-2.8µs MPI latency, 1.2 GByte/s one-way data rate
  - Pallas/IMB benchmarks with low-latency, layer-2, 10-Gigabit Ethernet switches
  - Nearly on-par with results with MX over Myrinet (MXoM)
- MXoE uses Ethernet as a layer-2 network with an MX EtherType to identify MX packets (frames)
  - The Myri-10G NICs can carry IP traffic (IPoE) together with MX (MXoE) traffic
  - Myricom is making the MXoE protocols open

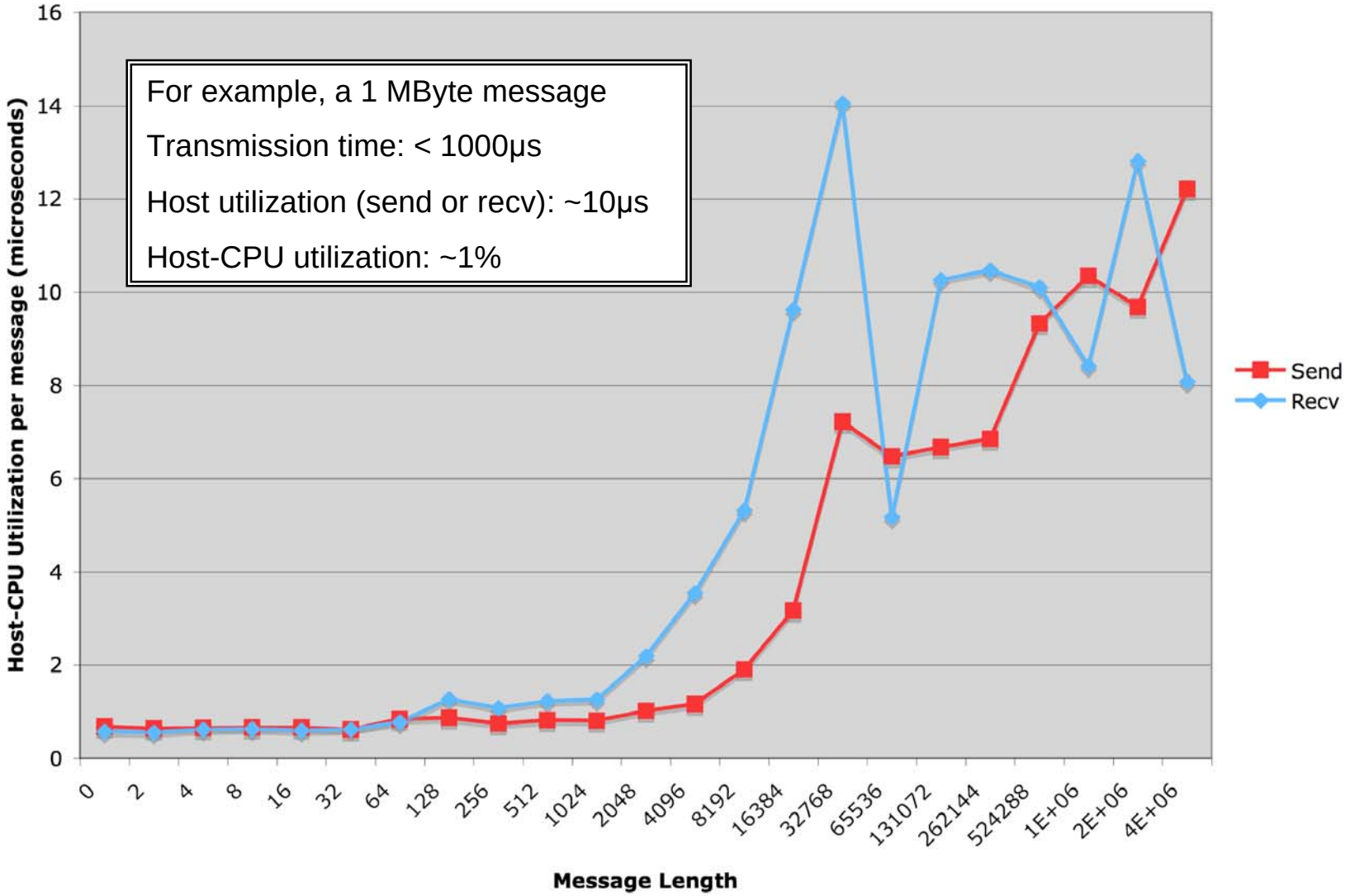
MX MPI Performance Measurements

MPI Benchmark	**MX over Myrinet** Myricom 128-port 10G Myrinet Switch	**MX over Ethernet** Fujitsu XG700 (Fujitsu MB8AA3020) 10G Ethernet switch	**OpenIB** with Intel MPI Mellanox InfiniBand No switch
PingPong latency	2.3 µs	2.8 µs (2.63 µs)	4.0 µs
One-way data rate (PingPong)	1204 MByte/s	1201 MByte/s	964 MByte/s
Two-way data rate (SendRecv)	2397 MByte/s	2387 MByte/s	1902 MByte/s

The MPI benchmarks for MX are the standard Pallas, now Intel, MPI benchmarks. The data rates are converted from the Mebibyte (2^{20} Byte) per second measure reported to the standard MByte/s measure.

The MPI benchmarks for OpenIB (with Intel MPI) are from a published OSU Benchmark Comparison, May 11, 2006. The numbers cited are typical of the best of 45 benchmarks reported. The reported latency does not include the latency of an InfiniBand switch; thus, the actual in-system latency will be higher. The data rates are from streaming tests, which are less demanding than and produce better throughput numbers than PingPong tests.

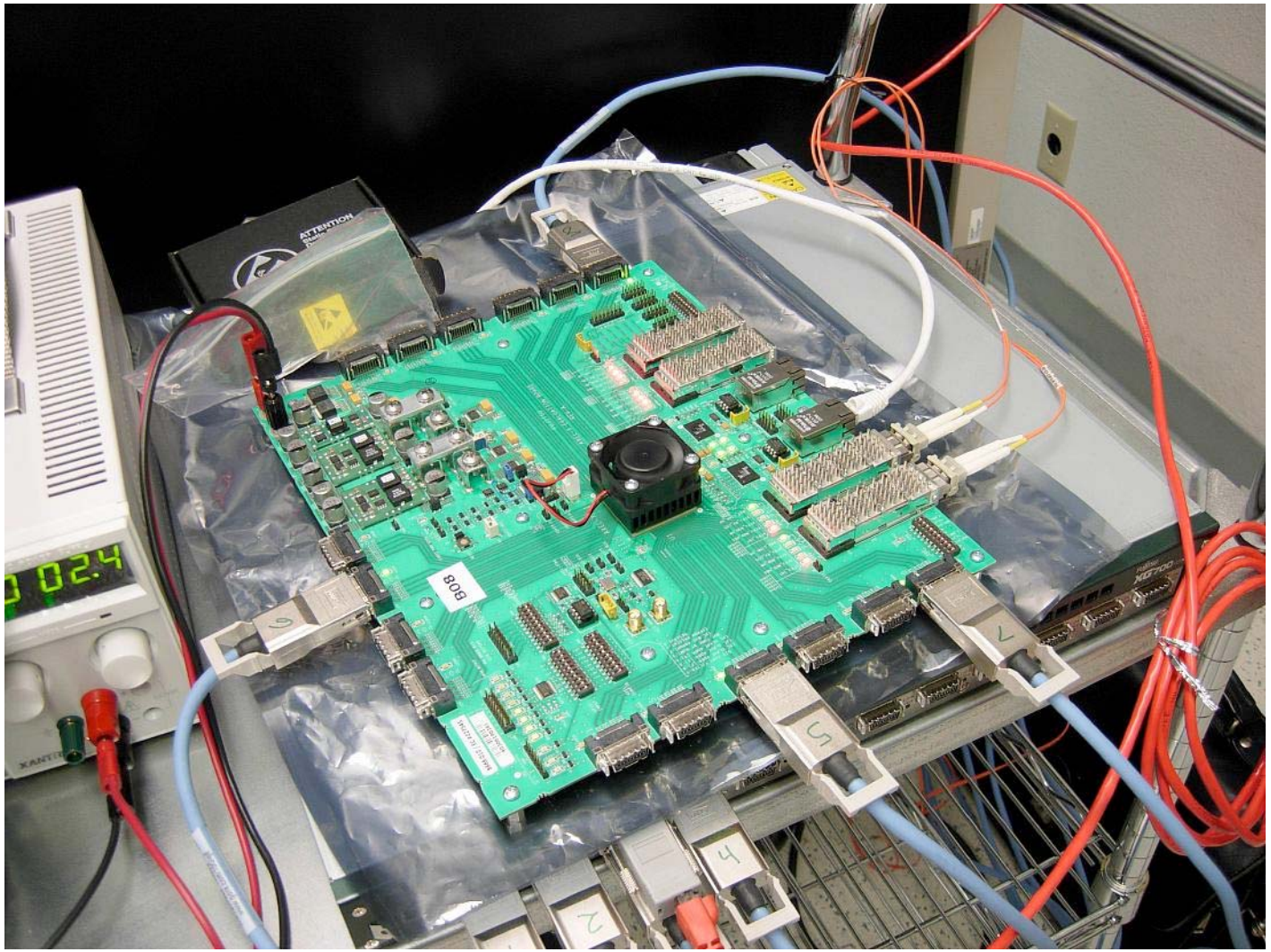
MXoE Host-CPU Utilization



Ethernet or Myrinet Switching?

- **The solved problem:** 10-Gigabit Ethernet is capable with MXoE and Myri-10G NICs of performance formerly associated only with specialty cluster interconnects
  - Not only low latency, but low-host-CPU utilization, thanks to MX's kernel-bypass mode of operation
- **The unsolved problem:** These Ethernet solutions are limited to smaller clusters that can be served with a single 10-Gigabit Ethernet switch
  - There are performance losses in building larger Ethernet networks by connecting multiple Ethernet switches
  - Inasmuch as there are no high-port-count, low-latency, full-bisection, 10-Gigabit Ethernet switches on the market today, MX over Myrinet with 10-Gigabit Myrinet switches will continue to be preferred for large clusters because of the economy and scalability of Myrinet switching

Advances in 10GbE Switches



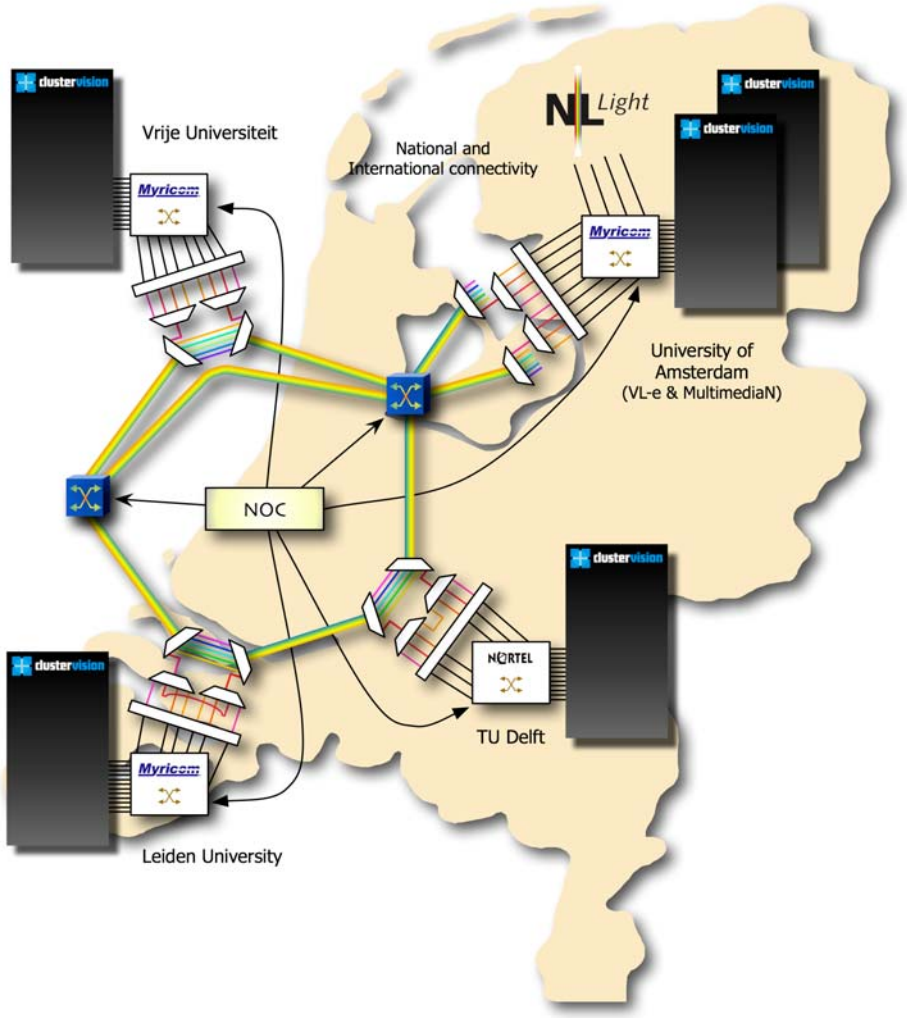
Fujitsu 20-port low-latency 10GbE switch being benchmarked in a Myricom lab

10G Myrinet compared with *10G Ethernet*

- Sharing all the same physical-layer devices/systems …
- Discovery:
  - 10G Myrinet: hosts rely on mapping to learn the location of other hosts
  - 10G Ethernet: switches learn the location of hosts by observing traffic
- Flow Control:
  - 10G Myrinet: Byte-level flow control
  - 10G Ethernet: packet-level, MAC-layer-PAUSE flow control
- Switching:
  - 10G Myrinet: source-based routing, cut-through
  - 10G Ethernet: destination-based routing, nominally store-and-forward
- Scalability:
  - 10G Myrinet: virtually unlimited, utilizing Clos topologies
  - 10G Ethernet: limited by the spanning tree, so multi-switch configurations typically use Layer-3, IP routing between switches

DAS-3: High-Performance Interoperability

- Being installed in the Netherlands
  - Operational by the end of 2006
- Five supercomputing clusters connected by a private DWDM fiber network
- “Seamless” cluster and grid operation thanks to Myri-10G interoperability
  - MX within each cluster; IP between clusters



Questions?

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