Myri-10G

Myrinet Converges with Ethernet

David PeGan

VP, Sales

dave@myri.com

(Substituting for Tom Leinberger)

4 October 2006

Oklahoma Supercomputing Symposium

New Directions for Myricom

- Although Myricom has done well (and we hope some good) with Myrinet in HPC, this market is limited
- We foresee little future for “specialty networks”
  - Technical convergence/standardization and business consolidation has been evident in the computer industry over the past decade
  - Ethernet will prevail over all of the specialty (“Ethernot”) networks
- 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

Myri-10G is …

- ***4th-generation Myricom products,*** a *convergence* that leverages 10-Gigabit Ethernet technology into the HPC world, and HPC techniques into the Ethernet world
  - 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

Myri-10G NICs

10GBase-CX4

10GBase-R

XAUI over ribbon fiber

These programmable NICs are PCI Express x8

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

10-Gigabit Myrinet NICs. Use them with Myricom's Myrinet Express (MX) 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

Myri-10G Switches

- **Switches** are 10-Gigabit Myrinet
  - Cut-through switching with source routing
  - Low latency, economical, and scalable
- Full-bisection Clos networks
  - Based on 16-port and 32-port single-chip crossbar switches
  - These networks are scalable to thousands of nodes using efficient, layer-2 switching
- Connection of a 10-Gigabit Myrinet switch fabric to interoperable 10-Gigabit Ethernet ports requires only simple protocol conversion
  - Now available in special switch line cards

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

Myri-10G Software

- Driver and firmware for 10-Gigabit Ethernet operation is included (bundled) 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
  - Includes TCP/IP, UDP/IP, MPICH-MX, and Sockets-MX
    - MPICH2-MX coming soon. Also available: OpenMPI, HP-MPI, …
  - MX-2G and MX-10G are fully compatible at the application level
    - MX-2G applications don't even need to be recompiled. Just change LD_LIBRARY_PATH to point to the MX-10G libraries.

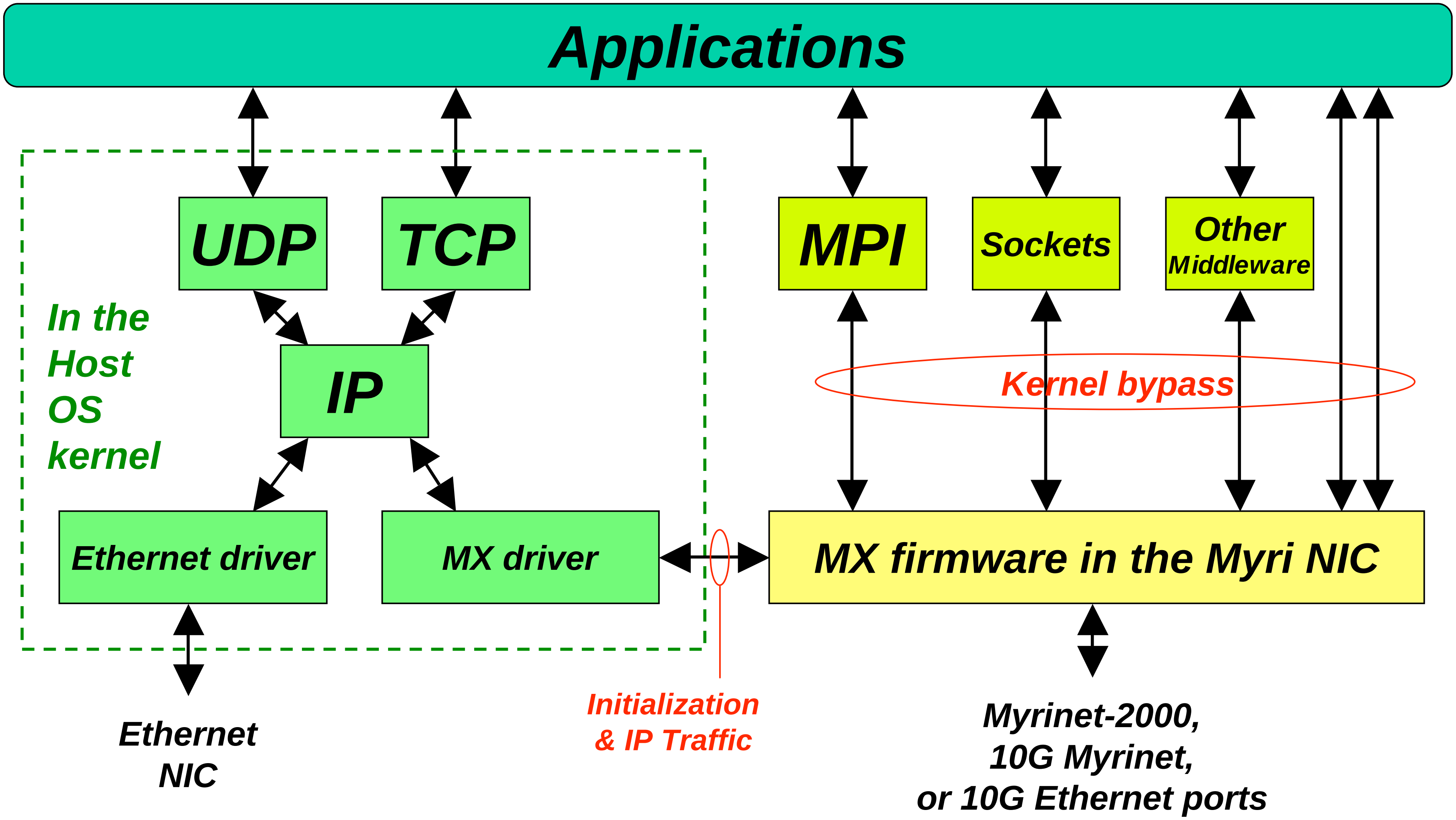
What's New: MX over Ethernet

- ***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 ☞

Open MXoE Invitation

- Myricom is publishing the "on the wires" protocols for MX over Ethernet
- Myricom invites proposals from universities to produce interoperable, open-source implementations of MXoE using other Ethernet NICs, *e.g.,* Intel or Broadcom GbE NICs, or other 10GbE NICs
  - Myricom will support such research with Myri-10G components and a modest research stipend
  - Given the low system-call overhead with Linux 2.6, we expect the performance of MXoE to be excellent even with ordinary GbE NICs
    - Not kernel bypass, but bypasses the host-OS TCP/IP stack
  - Performance boost for Beowulf clusters

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 + Sockets over MX + MPI over MX	TCP/IP, UDP/IP host-OS network stack + MX kernel bypass	MX-10G	IPoE IP over Ethernet MXoE MX over Ethernet	
			IPoM IP over Myrinet MXoM MX over Myrinet	Myrinet

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.

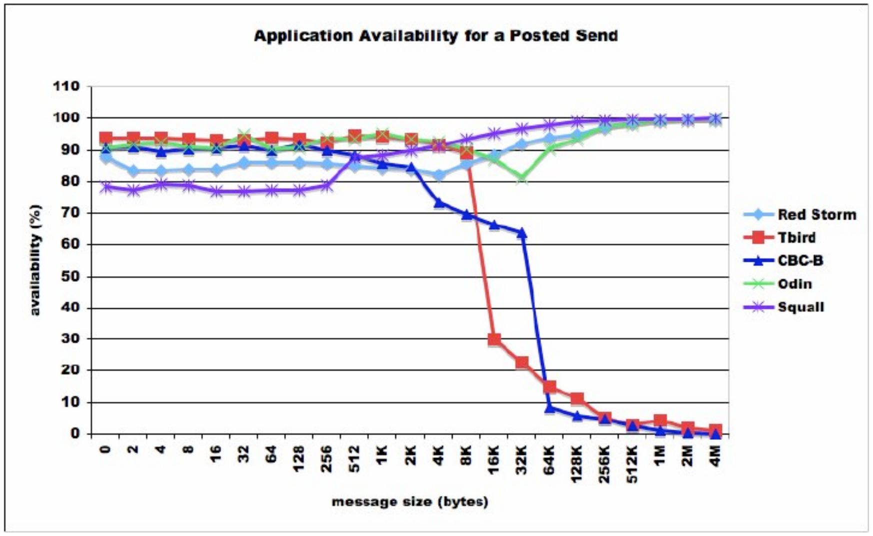
MX MPI Performance Measurements

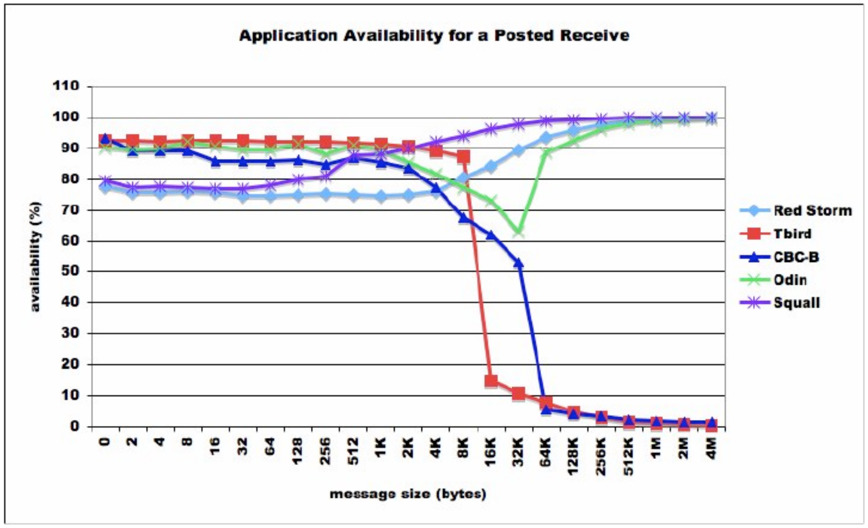
MPI Benchmark	**MX over Myrinet** Myricom 128-port 10G Myrinet Switch	**MX over Ethernet** Fujitsu XG700 12-port or MB8AA3020 20-port 10G Ethernet switch	**OpenIB** with Intel MPI Mellanox InfiniBand
PingPong latency	2.3 µs	2.80 µs (12-port) 2.63 µs (20-port)	4.0 µs
One-way data rate (PingPong)	1204 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.

Communication/Computation Overlap


Application Availability for a Posted Send

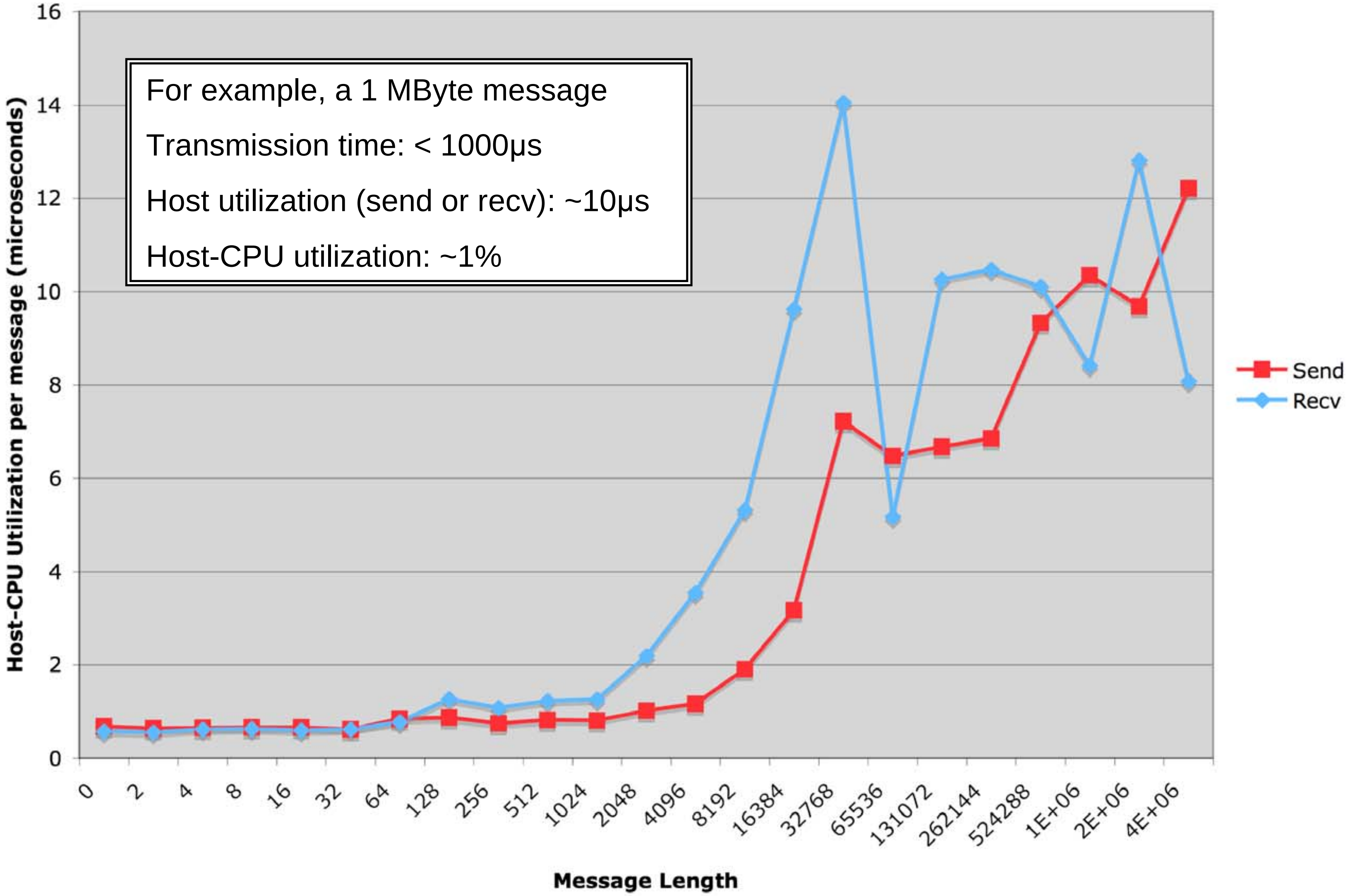

Application Availability for a Posted Receive

The ability of application programs to make progress while communication is taking place concurrently 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 et al (Sandia), *EuroPVM/MPI, September 2006*

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

MXoE Host-CPU Utilization



For example, a 1 MByte message

Transmission time: < 1000µs

Host utilization (send or recv): ~10µs

Host-CPU utilization: ~1%

Ethernet or Myrinet Switching?

- For small clusters, up to the size that can be supported from a single switch, 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 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.

Scalability of Myrinet Switching

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

Extras

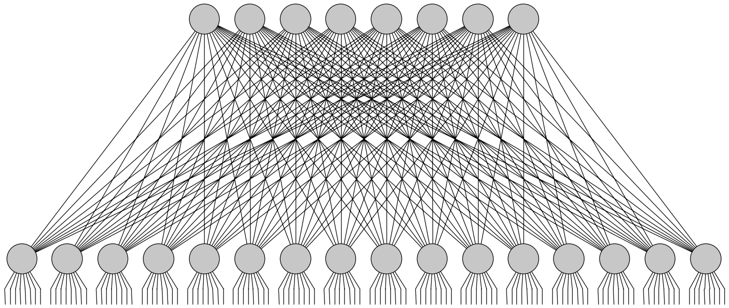
More good things about MX

10-Gigabit Ethernet software and performance

High-performance Interoperability

MX over Myrinet Adaptive Dispersive Routing

- MX takes advantage of the multiple paths through large Myrinet networks to spread packet traffic
  - MX mapping provides multiple routes (usually 8) to each other host
  - Flow-control backpressure sensed on packet injection indicates contention on the route
  - MX changes route when contention is sensed in the network
  - *Note:* Dispersive, multi-path routing may cause packets to be received out-of-order, but MX can reorder packets on receive
    - Matching (message level) is always in-order


Clos Network for 128 hosts

- **Eliminates "hot spots" in switch networks**
  - Adapts the routes to the communication patterns of the application
  - Fault tolerance on the time scale of route selection rather than remapping
  - Extremely valuable for large switch networks
  - Only possible with source-routed networks

Fabric Management System (FMS)

- An all-in-one Myrinet Fabric Management package
  - Switch, host, and link monitoring
  - Fabric control
  - Has largely replaced the earlier mapping software
    - Fabric Management Agent (FMA) process on each host reports to the Fabric Management System (FMS) process
    - FMS maintains a database in order to compare the current state of the fabric with the desired state
- Increased uptime
  - Proactive problem diagnosis and reporting
  - Tools allow monitoring and maintenance during production
- See the documentation on the web
  - http://www.myri.com/scs/fms/

The Lustre File System over MX

- ***New*** software package over MX-2G and MX-10G
  - Runs “native” on Myrinet-2000 and Myri-10G networks
- Leverages MX Advantages
  - Two-sided communication
  - Simple kernel library mirrors user API
  - Lower memory requirements than IB
    - No queue pairs (IB suffers from N^2 memory growth)
    - MX needs only ~128 Bytes per peer
- Excellent Performance
  - Currently with MX-10G: 1165 MB/s read, 1175 MB/s write, and 36K metadata/s
    - With IB: reads and writes at 900 MB/s and 30K metadata/s

Myri-10G in the 10GbE Market

- Highest throughput 10-Gigabit Ethernet NICs on the market
  - First PCI-Express 10GbE NICs
  - Near wire speed TCP/IP or UDP/IP with 9KByte jumbo frames
- Demonstrated interoperability with 10-Gigabit Ethernet switches from Foundry, Extreme, HP, Quadrics, Fujitsu, SMC, Force10, Cisco, (more to come)
- The Myri10GE driver and firmware is currently available for Linux, Windows, Solaris, Mac OS X, and FreeBSD
  - Driver was contributed to and accepted in the Linux kernel; included in the 2.6.18 and later kernels.
- Also an iSCSI target and initiator
  - See http://www.myri.com/scs/iSCSI/
- Bell Micro is Myricom's distributor for 10-Gigabit Ethernet NICs in the US and Europe. OEMs and cluster integrators continue to buy direct from Myricom.

10-Gigabit Ethernet Performance

http://www.myri.com/scs/performance/Myri10GE/

includes current performance measurements with Linux, Windows, and Solaris. The performance is quite good both in throughput and in host-CPU load with 9K-Byte jumbo frames. Our current development projects aim at improving the performance with 1500-Byte frames. The netperf benchmark results below are with Linux.

```
Netperf Test   MTU    BW        TX_CPU %   RX_CPU %
------------   ----   -------   --------   --------
UDP_STREAM_TX  9000   9915.38   38.73      00.00
UDP_STREAM_RX  9000   9908.09   00.00      44.26
TCP_STREAM     9000   9565.49   49.95      49.97
TCP_SENDFILE   9000   9496.91   12.40      50.67
```

10GbE Offloads

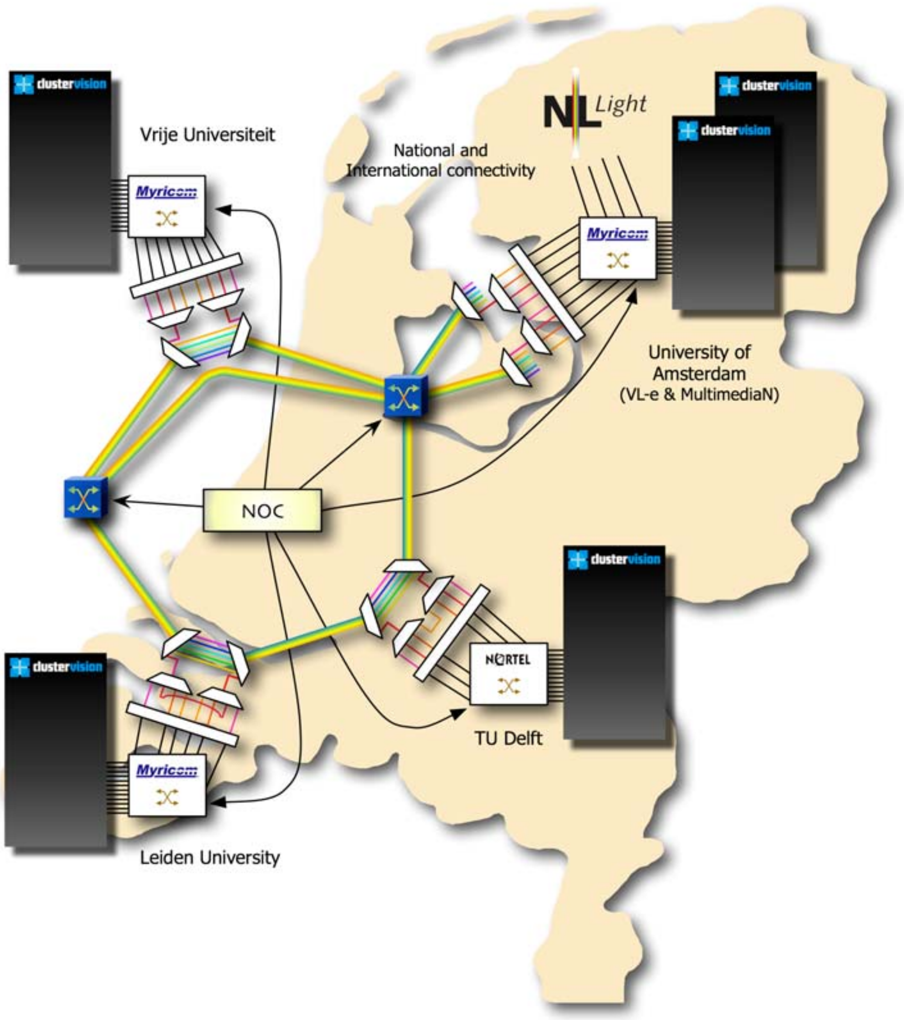
We implement zero-copy on the send side with all OSes, and, depending on the OS, use a variety of stateless offloads in the NIC, including:

- Interrupt Coalescing
- IP and TCP checksum offload, send and receive
- TSO (TCP Segmentation Offload, also known as Large Send Offload)
- RSS (Receive-Side Scaling)
- LRO (Large Receive Offload)
- Multicast filtering

We do not currently implement any “stateful” offloads. These NICs and their software are not TOEs. Note that TCP offload generally requires OS-kernel patches.

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
  - Myrinet protocols within each cluster; IP protocols between clusters


clustervision
Vrije Universiteit
Myricom
National and International connectivity
NLLight
clustervision
clustervision
Myricom
University of Amsterdam (VL-e & MultimediaN)
NOC
clustervision
NORTEL
TU Delft
clustervision
Myricom
Leiden University