Myri-10G
HPC mit 10-Gigabit Ethernet

Dr. Markus Fischer
Senior Software Architect
fischer@myri.com

5 March 2010
ZKI-Arbeitskreis Supercomputing
Hamburg
**Myri-10G**

*4th-generation Myricom products*, a *convergence* at 10-Gigabit/s data rates of Myrinet with Ethernet

- Based on 10-Gigabit Ethernet PHYs (layer 1)
  - Standard 10-Gigabit Ethernet cabling, copper and fiber
- NICs support both Ethernet and Myrinet network protocols at the Data Link level (layer 2)
  - and, like earlier Myricom NICs, include processors and firmware for offload and kernel-bypass operation
- Switches retain the efficiency and scalability of layer-2 Myrinet switching internally
  - but can have a mix of 10-Gigabit Myrinet and 10-Gigabit Ethernet ports externally
The Spectrum of Myri-10G Applications

• **10-Gigabit Ethernet solutions for general networking**
  o **10-Gigabit Ethernet NICs** with wire-speed TCP/IP performance, low cost, fully compliant with Ethernet standards, interoperable with the 10-Gigabit Ethernet products of other companies

• **Optional capabilities enabled by NIC firmware**
  o for low-latency, low-host-CPU-load, kernel-bypass operation
    **over either 10Gb Ethernet or 10Gb Myrinet networks**
    ▪ MX (Myrinet Express) for HPC; Video PumpTM for IPTV video streaming; Sniffer10G for 10-Gigabit Ethernet packet capture; DBL for low-latency UDP communication, …

• **10-Gigabit Myrinet solutions for HPC (Myrinet 10G)**
  o A complete, low-latency, scalable, cluster-interconnect solution – NICs, software, and switches – software-compatible with Myrinet-2000, and highly interoperable with 10-Gigabit Ethernet
Myri-10G 10-Gigabit Ethernet Solutions

Although Myri-10G products continue to be used with Myrinet protocols for traditional HPC clusters, most of Myricom’s sales today are for 10-Gigabit Ethernet applications with demanding latency or throughput requirements.

“HPC” techniques are today not just for MPI applications. They are also for High Performance Communication applications.
Myri-10G Network Interface Cards (NICs)

Fastest, lowest cost, lowest power, 10-Gigabit Ethernet NICs available
Myri-10G TCP/IP Performance

- The excellent **netperf benchmark** results below are between one-port Myri-10G “8B” NICs with the Myri10GE 1.5.0 driver in servers with two Intel quad-core 2.93GHz Xeon X5570s running Linux (2.6.18):

<table>
<thead>
<tr>
<th>MTU</th>
<th>BW</th>
<th>TX_CPU %</th>
<th>RX_CPU %</th>
</tr>
</thead>
<tbody>
<tr>
<td>9000</td>
<td>9910.33</td>
<td>4.52</td>
<td>2.84</td>
</tr>
<tr>
<td>9000</td>
<td>9910.32</td>
<td>2.71</td>
<td>2.82</td>
</tr>
<tr>
<td>1500</td>
<td>9477.10</td>
<td>4.62</td>
<td>5.57</td>
</tr>
<tr>
<td>1500</td>
<td>9452.54</td>
<td>2.56</td>
<td>5.63</td>
</tr>
</tbody>
</table>

Performance/Cost Leadership

The performance/cost leadership of Myri-10G NICs is the result of several factors that derive from Myricom’s 16 years of experience in high-performance networking.

• **Architecture:** Processors and firmware in the NIC
  - Unique in the 10-Gigabit Ethernet NIC market in being internally programmable, and capable of kernel-bypass operation

• **Software support:** Broad and mature
  - Bundled software distribution for conventional TCP/IP and UDP/IP operation is available for Linux, Windows, Solaris, Mac OS X, FreeBSD, and VMware ESX
  - Also, optional firmware-accelerated software for demanding applications

• **Technology:** Custom-silicon implementations
  - High speed, low cost, and low power
Examples of Lanai-Z8ES-based NICs

Network Ports
10 Gb/s
Lanai Z8ES
PCIe x8

Network Ports
20 Gb/s
Lanai Z8ES
PCIe x8
Lanai Z8ES
PCIe x8

Network Ports
20 Gb/s
Lanai Z8ES
PCIe x8
Lanai Z8ES
PCIe x8
“Gen 2” PCIe x8

Simple NIC
Two ports for failover
10G-PCIE-8B-2S, -2I, -2C, -QP -S, -C

Double NIC
Two ports for performance
blades, motherboards
10G-PCIE-8B2-4I

Double NIC
Two ports for performance
Gen2 PCIe Add-in Cards
10G-PCIE2-8B2-2QP, -2S, -2C

Myricom
www.myri.com
© 2010 Myricom, Inc.
Myri-10G High Speed Expansion Cards (HSECs) for the IBM BladeCenter H

- 2 Ports for failover
  - Single PCIe device
  - 10 Gb/s throughput
  - 3 Watts
  - Product Code: 10G-PCIE-8B-2I

- 4 Ports, 2 for performance, 2 for failover
  - Two PCIe devices
  - 20 Gb/s throughput
  - 6 Watts
  - Product Code: 10G-PCIE-8B2-4I

Fastest, lowest cost, lowest power, Ethernet expansion cards available for the IBM BladeCenter H (and also great for Myrinet 10G applications)
This two-port NIC appears to the host operating system as two independent PCI Express devices, one for each network port. Two ports are provided not only for failover, but also for performance. The two devices can carry TCP/IP traffic concurrently at an aggregate data rate of 19.8 Gb/s with a 9KB MTU, or 18.9 Gb/s with a 1500B MTU. Fastest two-port 10-Gigabit Ethernet NIC available.
“Gen2” PCIe v2.0 (5GT/s) NIC with 2 10GBase-CX4 ports

Product code: 10G-PCIE2-8B2-2C

2 CX4 Network Ports

PCI Express 5GT/s x8 Host Port

Different PHY, but the same story. Fastest two-port CX4 10-Gigabit Ethernet NIC on the market. Aggregate data rate of 19.8 Gb/s with a 9KB MTU, or 18.9 Gb/s with a 1500B MTU.
One-port NIC for HPC-cluster applications

Indiana University and TU Dresden won first place in the SC08 Cluster Challenge, best performance within a limited power budget, using these 3.3W NICs in an IBM iDataPlex system with a Myri-10G switch.
Preferred two-port NIC for HPC clusters

Myricom 10G-PCIE2-8B2-2QP
Two PCI Express devices
Nearly 20Gb/s throughput

Two QSFP network ports for performance and failover

“Gen2” PCI Express (5GT/s) x8 Host Port
Low-cost one-port Myri-10G NICs
Software Distributions, & Software Interfaces

Myri10GE for conventional TCP/IP and UDP/IP
MX (Myrinet Express) for HPC
Video Pump for IPTV video streaming
Sniffer10G for Ethernet packet capture
DBL (Datagram Bypass Layer) for low-latency UDP
(others in development)
The Myri10GE Software Distribution

• For conventional TCP/IP and UDP/IP operation
• Included (bundled) with all Myri-10G NICs
• The Myri10GE driver and firmware is currently available for Linux, Windows, Solaris, Mac OS X, FreeBSD, and VMware ESX
  o The driver was contributed to and accepted in the Linux kernel; included in the 2.6.18 and later kernels
    ▪ Also available in some Linux distributions; however, the software available for download is more up-to-date and has better performance.
  o WHQL-certified for Windows XP, Windows Server 2003 and 2008, Vista
    ▪ NDIS 5.1, 5.2, and NDIS 6.0
  o VMware certified for ESX 4
• For the Myri10GE software distribution, the firmware in the NIC is used for highly effective stateless offloads
• Not a TOE; no troublesome stateful offloads; no OS patches
Myri-10G 10GbE Offloads

The Myri10GE software distribution implements zero-copy on the send side with all operating systems, and, depending on the OS, uses a variety of stateless offloads in the driver and NIC firmware, 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

Thanks to the flexibility of the firmware, we were also able a year ago to start taking advantage of DCA on receive, and we also have firmware for ATS on Nehalems.
MX-10G Software Distribution

- MX-10G is an optional, low-level, message-passing system for low-latency, low-host-CPU-utilization, kernel-bypass operation of Myri-10G NICs
  - over either 10Gb Myrinet or 10Gb Ethernet networks
- MX (Myrinet Express) is used principally for HPC clusters
  - However, many of the kernel-bypass techniques for the other optional software distributions originated with MX
- Includes TCP/IP, UDP/IP, MPICH-MX, and Sockets-MX
  - Also available: MPICH2-MX, OpenMPI, HP-MPI, Intel-MPI, …
- Supports NIC bonding (teaming)
  - With \( N \) NICs, nearly \( N \) times the data rate at the same latency
- Cluster file systems operate directly over MX
  - Lustre-MX, PVFS2-MX, NFS-MX
MX Software Interfaces

Applications

UDP  TCP  IP

Ethernet driver  MX driver

In the Host OS kernel

Conventional Ethernet NIC

Initialization & IP Traffic

MX firmware in the Myri NIC

MPI  Sockets  Other Middleware

Myrinet-2000, 10G Myrinet, or 10G Ethernet ports

Kernel bypass
IMB with 4-way MX bonding between Nehalem hosts

Real 40 Gb/s
IMB data rate, not marketing-spin link-signaling rate

From the Dalco demo cluster in the Myricom booth at ISC09.

The mebibyte/s units of the IMB output are converted to MB/s, and then to Gb/s for the upper meter.

4912 MB/s
Video PumpTM Software Distribution

- Optional software+firmware solution for UDP streaming
  - Developed principally for IPTV video applications
- Typical demonstration maintains 9.4 Gb/s, 2350 streams @ 4 Mb/s each, at less than 5% host-CPU load
  - 1380-Byte UDP packets (1338-Byte payload)
- The NIC firmware does most of the work
  - Host CPU performs only non-time-critical operations
  - Fine-grain packet scheduling (constant bit rate, low jitter)
    - Also referred to as “traffic shaping”
  - Zero-copy; checksum offload
- At the same time, Video Pump supports TCP/IP and UDP/IP operation through the host OS protocol stack
  - Video traffic normally has priority
Sniffer10G Software Distribution

• An optional software distribution for 10-Gigabit Ethernet packet capture
  o Available for all Myri-10G NICs with Linux and FreeBSD hosts
  o Writes all received packets into host memory in an easily interpreted data structure (compatible with pcap and libpcap)
  o Small-packet coalescing and an efficient zero-copy path to the host memory allow Sniffer10G to capture streams at high packet rates and at full line speed
    ▪ With “8B” NICs, Sniffer10G achieves 14.88Mpps (theoretical maximum for 10GbE) sustained packet rate into host memory
  o See http://www.myri.com/scs/SNF/doc/ for full documentation

• The Myricom software team is currently adding a “replay” capability (packet capture in reverse).
  o Discussion point: Could you use Sniffer10G for MPI applications?
**DBL (Datagram Bypass Layer)**

- **Low-latency UDP/IP communication**
  - User-level latency less than ~2.5µs (process to process)
  - *Used principally for financial-trading PubSub applications*
- **DBL software with Myri-10G NICs** leverages user-level, kernel-bypass, messaging techniques originally developed for MX HPC applications, and applies them to UDP communication over Ethernet in a multicast and/or unicast environment
  - Interoperable and wire-compatible with other UDP/IP implementations
  - DBL software distributions are currently available for Linux, Windows, and Solaris
  - Also includes a Java interface (instead of UDP Sockets)
    - Many financial-trading applications are written in Java
  - The current beta DBL software for Myri-10G NICs requires an MX license and a mutual NDA
Myri-10G Modular Switches

Myricom’s modular Myri-10G switches are based on a 110ns-latency 32-port Myrinet-protocol crossbar-switch chip. The photos show the current choice of enclosures for these modular switches. Different mixes of Ethernet or Myrinet external ports, as well as ports with different PHYs, are configured by using different types of line cards. The line cards plug into a backplane that connects the switch chips on the line cards in a diameter-3 full-bisection Clos network. For applications requiring more than 512 host ports, the switch networks are scalable by interconnecting the internal Myrinet fabrics.
Myri-10G in the TOP500 List

We’re pleased about the “Myrinet 10G” clusters in the TOP500 list, e.g., the T2K cluster at the University of Tokyo (#45), and the clusters at USC (#71) and Clemson (#79). However, in keeping with Myricom’s diversification, our Myri-10G products are used in many other TOP500 systems without being listed explicitly, for example:

- ORNL Jaguar (#1), LANL Roadrunner (#2), and UT Kraken (#3) all use Myri-10G 10GbE NICs for Ethernet or storage connectivity
- The IBM Blue Gene/P system at Argonne National Laboratory (#8) uses a Myri-10G dual-protocol switch with nearly 1000 ports, 10Gb Ethernet to the Blue Gene/P racks and 10Gb Myrinet to the file-system cluster
  - Other TOP500 Blue Gene/P systems that use Myri-10G switches for storage switching: IDRIS (#32) and EDF (#49), both in France
- The central 10GbE switch for the SciNet cluster (#22) is from Myricom
- and similar off-the-list use in many other TOP500 systems