

# Low-Latency 10-Gigabit Ethernet

## *HPC Clustering with Myri-10G NICs and Standard 10-Gigabit Ethernet Switches*

Myricom supplies optional message-passing software for Myri-10G network-interface cards (NICs) to deliver low latency and low host-CPU utilization over 10-Gigabit Ethernet. Myricom extended its Myrinet Express (MX) software, already widely used in High-Performance Computing (HPC) clusters interconnected with Myrinet, to work also over 10-Gigabit Ethernet. “MX over Ethernet” operates by kernel bypass with Myricom’s dual-protocol Myri-10G network-interface cards and standard 10-Gigabit Ethernet switches to achieve latencies 5 to 10 times lower than with TCP/IP over Ethernet. As is detailed below, MX over Ethernet performance metrics with low-latency 10-Gigabit Ethernet switches are nearly on par with those achieved by MX over Myrinet.

The MX over Ethernet (MXoE) protocols are open, and Myricom encourages implementations using other Ethernet NICs. The technique is transparent to Ethernet switch makers, less expensive than proprietary HPC solutions, and applicable both to HPC and to enterprises.

**How it Works.** Myricom’s Myri-10G solutions introduced a convergence at 10-Gigabit/s data rates of Myrinet, the most successful specialty network for HPC applications, and mainstream Ethernet. Dual-protocol Myri-10G NICs initially achieved optimal performance running MX software with Myrinet network protocols through Myri-10G switches. MX’s kernel-bypass techniques achieve low latency and low host-CPU utilization by allowing application programs to communicate directly with firmware in the programmable Myri-10G NICs. Now, the availability of MXoE extends MX’s advantages to standard 10-Gigabit Ethernet switching. OEMs and cluster integrators can achieve HPC performance with mainstream Ethernet technology. Myricom is making the MXoE protocols fully open and accessible, just as with earlier Myrinet protocols and source code.

MXoE uses 10-Gigabit Ethernet as a layer-2 network with an MX EtherType to identify MX frames (packets). The EtherType, a part of the Ethernet standards since the earliest days, identifies the protocol of an Ethernet frame. For example, there are EtherTypes for the Internet Protocol (IP), Address Resolution Protocol (ARP), AppleTalk, and many other protocols. All of these protocols can be carried concurrently on the same Ethernet network. Ethernet switches normally ignore the EtherType. Myri-10G NICs carry TCP/IP and other traffic along with MX traffic, but achieve the best performance by circumventing TCP/IP. MX provides its own, highly efficient, reliability layer.

MXoE is plug-and-play with any 10-Gigabit Ethernet switch, although you get the best performance with low-latency switches such as the Fujitsu XG700 or XG2000. The table below of MPI benchmarks<sup>1</sup> starts with MX over Myrinet with Myri-10G NICs and a 10-Gigabit Myrinet switch as a baseline. The performance of MX over Ethernet with the low-latency Fujitsu XG2000 20-port 10-Gigabit Ethernet switch is nearly as good as the MX over Myrinet performance. The last column of the table cites published<sup>2</sup> MPI benchmarks for Mellanox InfiniBand to show that, even with standard 10-Gigabit Ethernet switches, MX with Myri-10G NICs outperforms InfiniBand.

MPI Benchmark	<b>MX over Myrinet</b> Myricom 128-port 10G Myrinet switch	<b>MX over Ethernet</b> Fujitsu 20-port 10G Ethernet switch	<b>OpenIB</b> with Intel MPI Mellanox InfiniBand
PingPong latency	2.3μs	2.63μs	4.0μs
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

In addition to low latency, MX exhibits host-CPU utilization that is dramatically lower than the typical TCP/IP utilization and service demand reported in standard benchmarks such as netperf. The host-CPU utilization for MPI communication for MXoE ranges from less than 1μs of host-CPU time at the sender or receiver to transfer

messages up to 2 KBytes, then increasing gradually to  $\sim 10\mu\text{s}$  of host-CPU time to transfer messages in the range from 64KBytes to many MBytes. At a 1MByte message size, for example, a data transfer that MXoE accomplishes across 10-Gigabit Ethernet in less than 1000 $\mu\text{s}$ , the 10 $\mu\text{s}$  host-CPU utilization corresponds to a host-CPU utilization of  $\sim 1\%$ , an unheard-of low host-CPU load in the TCP/IP world. Even applications that are not sensitive to latency can benefit from MXoE due to the savings in host-CPU load.

These MX/Ethernet results show that for small clusters, up to the size that can be supported from a single switch, ***10-Gigabit Ethernet is capable of performance formerly associated only with specialty cluster interconnects.*** These solutions will be limited to smaller clusters that can be served with a single 10-Gigabit Ethernet switch, because of performance losses in building larger 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.

This MXoE innovation provides strong new evidence that 10-Gigabit Ethernet is a good choice for the interconnect for small HPC clusters. 10-Gigabit Ethernet will be used for larger clusters as 10-Gigabit Ethernet switch technology advances. Clusters have come to dominate the TOP500 supercomputer list in recent years. Over the past three years, commodity Gigabit Ethernet has eclipsed specialty interconnects, including Myricom's earlier Myrinet-2000 interconnect, in the number of systems in the TOP500 list. However, Gigabit Ethernet is not fast enough for leading-edge cluster hosts with their multiple, multi-core processors. In anticipation of these trends, Myricom's latest generation of products, Myri-10G, was designed as a convergence at 10-Gigabit/s data rates of Myrinet, the most successful specialty network for HPC applications, and mainstream Ethernet. As these MX over Ethernet results demonstrate, Myricom's Myri-10G technology combines the best of both worlds.



*Dual-protocol Myri-10G NIC with a PCI-Express x8 host port and a 10GBASE-CX4 network port.*

---

<sup>1</sup> 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.

<sup>2</sup> These MPI benchmarks for SDR Mellanox InfiniBand are corroborated from many sources. For example, in an OSU Benchmark Comparison, May 11, 2006, the numbers cited are with Intel MPI, and are typical of the best of 45 benchmarks reported. The latency reported in these OSU benchmarks 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. The link for "MPI performance of interconnects" from this Paderborn Center for Parallel Computing web page, <http://www.wcs.uni-paderborn.de/pc2/index.php?id=277>, provides similar results with bandwidths in Mebibyte units. The same table provides data for DDR InfiniBand that shows little improvement over SDR InfiniBand, particularly for bidirectional SendRecv throughput, because the throughput is limited by PCI Express and software.