The Myricom Myrinet-10G (Myri-10G) interconnect can offer high bandwidth and low latency in high-performance computing environments. This article describes the Myri-10G interconnect architecture and software stack as well as test results comparing the latency, throughput, and performance of Myri-10G, Myrinet-2000, and Gigabit Ethernet on ninth-generation Dell PowerEdge servers.

Introducing Myri-10G
Myri-10G, based on 10 Gigabit Ethernet (10GbE) technology, is the fourth-generation Myrinet interconnect, following the third-generation Myrinet-2000. Myri-10G is supported in two modes: Ethernet mode and Myrinet Express (MX) mode. The MX protocol offers kernel bypass with high-bandwidth, low-latency software support for intra-process cluster communication.

Myri-10G network interface cards (NICs) can be connected to Myri-10G or 10GBase-T switches. The current Myri-10G switch is a 14U enclosure with a 128-port switch that supports both 10GBase-CX4 copper interfaces and 10G-Q quad ribbon-fiber cables for 10GbE Attachment Unit Interface (XAUI).

The x8 PCI Express (PCIe) bandwidth supports 2+2 GB/sec full duplex I/O transfer capacity, which is wide enough to accommodate Myri-10G NICs that support 1.25+1.25 GB/sec bidirectional bandwidth. The NIC offers a total data rate of 10+10 Gbps full duplex. (Myrinet-2000, by comparison, offers a total bandwidth of 2+2 Gbps.)

At the physical layer, the Myri-10G NIC is available in two types: 10GBase-CX4, which connects to a switch through copper cables, and 10GBase-R, which connects to a switch through fiber cables. Both types can operate in Ethernet mode or MX mode; in MX mode, they can deliver low-latency, high-bandwidth kernel bypass Message Passing Interface (MPI) communication that helps reduce host processor utilization.

Myri-10G hardware architecture
The Myri-10G hardware architecture consists of processors, memory, and firmware. This architecture enables the offloading of network protocol processing, minimizing host processor utilization and enabling direct low-latency kernel bypass communication with applications. The Myri-10G NIC has a very-large-scale integration (VLSI) chipset known as the Lanai Z8E, and includes an internal processor at 313 MHz and built-in static RAM (SRAM).

The architecture also includes internal packet buffers. One advantage of on-chip buffers is that they provide most of the required memory bandwidth needed during data communication, thus limiting SRAM memory use on the NIC. The NIC has approximately 2 MB of local SRAM available with a bandwidth of 2,400 MB/sec, which is primarily used for firmware execution and packet header buffering. Packets are transferred through dynamic memory allocation, helping reduce host processor utilization.

When building a high-performance computing (HPC) system for communication-intensive applications, selecting an interconnect that offers high bandwidth and low latency at a good price/performance ratio can be critical. The Myricom Myrinet-10G (Myri-10G) interconnect can do just that, binding multiple nodes and enabling intra-node cluster communication to form the backbone of an HPC architecture.
Figure 1. Chipset architecture on Myri-10G NICs

Myri-10G software protocols

Myri-10G NICs support both Gigabit Ethernet and MX. MX is a low-level MPI implementation designed to provide low latency for small messages, support a virtually unlimited number of pending send and receive requests, and handle configurations with no memory registration.

The type of switch to which the NIC is connected determines whether network packets are transferred using TCP/IP or MX. In a cluster configuration with the default 10GbE drivers installed, the NIC uses TCP/IP; if MX drivers are installed, however, the NIC uses MX.

When MX drivers are installed, if the NIC is connected to a 10GbE switch, then MX uses its TCP/IP mode. If the NIC is connected to a Myri-10G switch—which is designed to handle MX packets—then MX uses Sockets-MX (see Figure 2). Unlike TCP/IP, Sockets-MX bypasses the traditional kernel stack, which can result in the MX protocol having lower latency than TCP/IP, in which the kernel traps typically take up communication time and create overhead.

Testing Myri-10G on ninth-generation Dell PowerEdge servers

In May 2007, Dell engineers from the High-Performance Computing Group performed tests to compare the performance of Myri-10G in MX mode with that of Myrinet-2000 in GM mode and Gigabit Ethernet. The Dell team used the Ohio State University (OSU) MPI-level latency test to evaluate latency, the Intel MPI Benchmarks (IMB) suite to evaluate unidirectional and bidirectional bandwidth, and High-Performance Linpack (HPL) to evaluate performance in clusters using each of the three interconnects. They compiled the tests to run on Myri-10G using the MPICH MPI implementation.

Cluster configurations in the test environment

Figure 3 details the two cluster configurations used in the test environment, which were based on Dell PowerEdge servers using either Intel® Xeon™ processors or AMD Opteron™ processors.

---

1For more information on these benchmarks, visit mvpich.cse.ohio-state.edu/benchmarks, www.intel.com/cd/software/products/asmo-na/eng/307696.htm#mpibenchmarks, and www.netlib.org/benchmark/hpl.
The team ran all tests on both clusters, then chose the best performance for the final results.

The BIOS versions on the cluster nodes support write-combining. Write-combining is supported on Myri-10G NICs by default on the PCIe chipset, and helps optimize Myri-10 performance by allowing data to be combined and temporarily stored in the buffer, then written (in “write-bursts”) whenever the buffer is full. This approach can be particularly helpful in frame buffers, and can increase the performance of write operations, especially when attempting to avoid excess switching overhead between read and write operations.

Test results: Latency, bandwidth, and performance

Figure 4 shows normalized latency for Myri-10G and Myrinet-2000 as a percentage of Gigabit Ethernet latency, as measured using the OSU latency test. Myri-10G had a very low latency of 5.4 percent of the total latency of Gigabit Ethernet, while Myrinet-2000 had approximately 12 percent of the total latency of Gigabit Ethernet.

Figure 5 shows normalized unidirectional bandwidth for Myri-10G and Myrinet-2000 as a multiple of Gigabit Ethernet unidirectional bandwidth, as measured using the IMB PingPong test. The test team measured Myri-10G both with and without the MX_RCACHE=1 variable. When this variable is enabled, MX uses registration cache rather than fetching each communication from memory, helping maximize bandwidth performance with the MX driver stack. Myri-10G with this variable provided 12.11 times the bandwidth of Gigabit Ethernet, or approximately 91 percent of the theoretical unidirectional bandwidth offered by the NIC. Myri-10G without this variable provided 9.97 times the bandwidth of Gigabit Ethernet, or approximately 75 percent of the theoretical unidirectional bandwidth offered by the NIC.

Figure 6 shows normalized bidirectional bandwidth for Myri-10G and Myrinet-2000 as a multiple of Gigabit Ethernet bidirectional bandwidth, measured using the IMB SendRecv test. In these tests, Myri-10G with the MX_RCACHE=1 variable provided 21.4 times the bandwidth of Gigabit Ethernet, while Myri-10G without this variable provided 17.8 times the bandwidth of Gigabit Ethernet.
Figure 6. Myri-10G and Myrinet-2000 bidirectional bandwidth as a multiple of Gigabit Ethernet bandwidth, measured using the Intel MPI Benchmarks SendRecv test

Figure 7. Myri-10G and Gigabit Ethernet performance, measured using the High-Performance Linpack application

Gigabit Ethernet. As with unidirectional bandwidth, Myri-10G with and without this variable provided approximately 91 percent and 75 percent of the theoretical bidirectional bandwidth offered by the NIC, respectively.

Figure 7 compares the performance in gigaflops of Myri-10G and Gigabit Ethernet, as evaluated when running HPL, a communication- and computation-intensive application well suited to serve as a cluster-level benchmark. The difference between the two interconnects is emphasized when scaling up the clusters. Using Myri-10G with 8 cores (two nodes) provided 81 percent of the processors’ theoretical maximum gigaflops, as compared with 75.48 percent when using Gigabit Ethernet. When scaled to 16 cores (four nodes), Myri-10G provided 80.84 percent of the theoretical maximum gigaflops, as compared with 73.73 percent when using Gigabit Ethernet. As the number of processing cores increased, communication across nodes became a bottleneck for Gigabit Ethernet, reducing performance relative to the theoretical maximum performance possible for a given number of cores.

Deploying Myri-10G on ninth-generation Dell PowerEdge servers

Dell has partnered with Myricom to deliver a pre-validated cluster solution using Myri-10G NICs and the 128-port Myri-10G switch in a 14U enclosure. Dell servers support these NICs and the switch in MX mode only. The switch enclosure houses the hot-swappable power supplies, cooling and network monitoring systems, a status display, and a backplane that supports up to 21 hot-swappable line cards. It also includes a monitoring line card that fits into the left line-card slot. The four center slots also contain line cards, with eight 16-port crossbar switches that can make up the spine of a Clos network; Dell-supported 10GBase-CX4 10G-SW16LC-8C line cards can be used in the 16 remaining slots.

81 percent of the processors’ theoretical maximum gigaflops, as compared with 75.48 percent when using Gigabit Ethernet. When scaled to 16 cores (four nodes), Myri-10G provided 80.84 percent of the theoretical maximum gigaflops, as compared with 73.73 percent when using Gigabit Ethernet. As the number of processing cores increased, communication across nodes became a bottleneck for Gigabit Ethernet, reducing performance relative to the theoretical maximum performance possible for a given number of cores.

Dell recommends deploying clusters using Platform Open Cluster Stack (OCS), which is pre-validated as a part of Dell HPC clusters. The Myri-10G software stack and drivers form a separate OCS Roll that administrators can first install on the front end during initial configuration and then deploy on the compute nodes through Preboot Execution Environment (PXE) kickstart images. Device drivers are loaded in the kernel space during compute node installation. Platform OCS enables both Myri-10G and Myrinet-2000 Rolls to coexist; these Rolls are different from and independent of each other.

Boosting HPC cluster performance with Myri-10G

As shown in Dell benchmarking tests, the Myricom Myri-10G interconnect can provide low latency and increased throughput in communication-intensive environments, including those with parallel applications running different MPI stacks such as MPICH and Open MPI. Myri-10G also remains interoperable with 10GbE and can minimize host processor utilization when running using MX. Enterprises can take advantage of these features to increase the performance of their HPC applications.

Sreeram Vedantham is a development engineer adviser in the Dell High-Performance Computing Group. His current interests include high-speed interconnects, HPC cluster stacks, and schedulers.

Shivaraj Nidoni is a senior engineering analyst in the Dell High-Performance Computing Group. His current interests include high-end interconnects, parallel file systems, and HPC clustering packages. Shivaraj has an M.Tech. in Computer Network Engineering from Visvesvaraya Technological University.

Munira Hussain is a systems engineer and adviser in the Dell High-Performance Computing Group specializing in HPC systems and architecture design. She has a bachelor’s degree in Electrical Engineering from the University of Illinois at Urbana-Champaign.