Optimizing HPC clusters with 10 Gigabit Ethernet iWARP technology

By Tom Stachura

As high-performance computing experts push the limits of cluster efficiency and price/performance, Dell™ servers in an iWARP-enabled Intel® 10 Gigabit Ethernet fabric can help reach new performance and efficiency levels—and a TOP500 ranking.

New supercomputers are continually coming online as business and academic researchers strive to keep up with ever-growing scientific and engineering computational demands. In the persistent quest to do more work with greater efficiency and less cost than ever before, high-performance computing (HPC) has moved from mainframe computers to economical and easily maintained clusters, including clusters made up of cost-effective servers based on the x86 architecture.

Today, some of the world's most powerful supercomputers are composed of x86-based clusters with performance delivered on the network side using high-speed InfiniBand connectivity. But for HPC experts striving toward reduced fabric costs and simplified use, InfiniBand can present several challenges. Now, Internet Wide Area RDMA Protocol (iWARP) provides a way to deliver Remote Direct Memory Access (RDMA) clustering on 10 Gigabit Ethernet (10GbE) network adapters—an approach that enables cluster users to overcome fabric challenges and continue to push the supercomputing envelope.

Assessing HPC connectivity challenges

InfiniBand connectivity presents several challenges for supercomputing. First, InfiniBand is a highly specialized switched-fabric communications link that requires special expertise and tools for setup, configuration, and management, and costs for tools, training, and outside expertise can be high. Second, because Ethernet is the de facto standard fabric for local networking traffic, using InfiniBand requires the HPC team to maintain two networking technologies—InfiniBand connectivity to run the application traffic and Ethernet connectivity to manage the servers in the cluster. Third, InfiniBand remains relatively costly in terms of per-port pricing, so hardware costs can be an impediment to increasing cluster cost-efficiency. Despite these factors, InfiniBand connectivity has been a logical technology for HPC because of its high throughput, low latency, and scalability.

Ethernet, in contrast, is widely used for connecting users and network resources, but has not been traditionally preferred for low-latency supercomputing. Ethernet does have the advantage of being extremely cost-effective for general-purpose LAN traffic. The bandwidth of Ethernet has increased 10-fold with the mainstream availability of 10GbE networking hardware, providing viable connectivity performance for HPC clusters. And Ethernet has broad scalability, providing the ability to dynamically add and remove nodes in cluster environments.

However, achieving low latency and high bandwidth for HPC applications requires enhancements to standard Ethernet. Standard Ethernet communicates using a kernel network protocol stack that adds overhead in terms of compute load, memory bandwidth, and network latency. In HPC environments, this overhead can greatly reduce performance, and is therefore unacceptable.

Avoiding Ethernet overhead barriers with iWARP

A full implementation of currently available iWARP technology helps avoid virtually all processor networking overhead, returning those cycles to the application. The Internet Engineering Task Force (IETF) standardized the iWARP specification in 2007; this standard specifies a set of extensions to the TCP/IP protocol that define a transport mechanism for RDMA. As such, iWARP provides a low-latency means of passing RDMA over Ethernet (see Figure 1). Together, these extensions address the three major sources of networking overhead: application context switches, intermediate buffer copies, and transport (TCP/IP) processing. These sources collectively account for nearly 100 percent of processor overhead related to networking.

The iWARP extensions help reduce processor overhead, memory bandwidth utilization, and latency using several advanced techniques:

- **Kernel bypass (OS bypass):** Context switching can be a costly process in terms of overhead. By allowing the application to communicate directly with the network controller, iWARP bypasses the kernel. This permits the kernel-to-user context switches to be avoided and helps reduce latency and processor load.
- **Intermediate buffer copies avoidance:** Data is placed directly in application buffers rather than being copied multiple times to driver and network stack buffers, which helps reduce latency as well as memory and processor use.
- **Transport (TCP/IP) processing acceleration:** TCP/IP processing is performed in hardware instead of the OS network stack software, enabling reliable connection processing at speed and scale. Data management and network protocol processing can be executed on the Ethernet adapter, which provides hardware acceleration of the process.

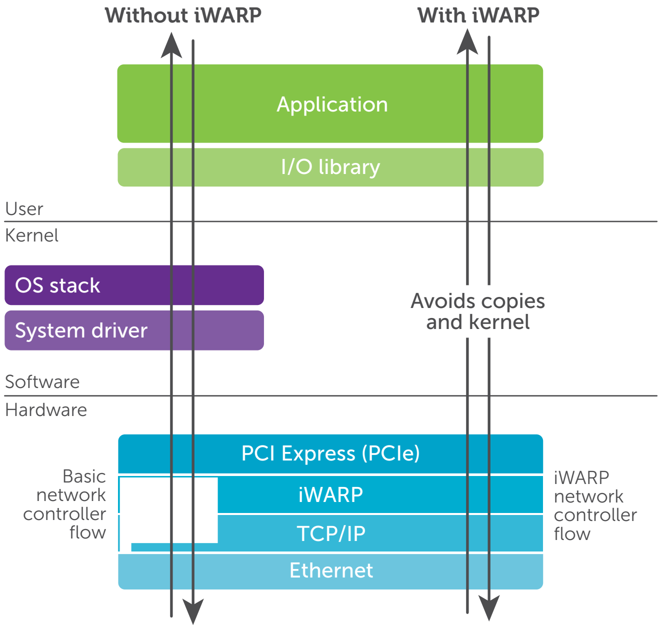

Figure 1. Network controller data flow with and without iWARP

Understanding iWARP

In this Intel technology brief, learn more about how iWARP helps reduce Ethernet overhead and latency to enable organizations to gain the full network throughput benefits of 10 Gigabit Ethernet.

download.intel.com/support/network/sb/understanding_iwarp_final.pdf

TOP500 Supercomputing Sites rankings for iWARP

The TOP500 Supercomputing Sites list ranks the most powerful high-performance computing (HPC) clusters in the world using the High-Performance Linpack (HPL) benchmark for distributed-memory computers—helping provide a reliable basis for tracking and detecting trends in HPC. The list is compiled by HPC experts at the University of Mannheim in Germany, the University of Tennessee at Knoxville, and the National Energy Research Scientific Computing Center (NERSC) at Lawrence Berkeley National Laboratory. Rankings are updated twice a year, coinciding with the annual Supercomputing Conference (SC) in November and the International Supercomputing Conference in June.

Several clusters on the list use Dell servers, including one 4,032-core cluster at a large biomedical research facility that uses Internet Wide Area RDMA Protocol (iWARP) and 10 Gigabit Ethernet (10GbE) technologies. The June 2010 list ranked this cluster at number 208 for performance, achieving higher performance than many of the world's top clusters. Reranking for efficiency shows that this cluster comes in at number 84, boasting a TOP100 efficiency—the highest of all the listed Ethernet-based supercomputers. For more information on the TOP500 list, visit top500.org.

Using iWARP-enabled 10GbE fabrics in HPC clusters

Using these techniques, the iWARP standard enables low-latency network connectivity that can be well suited for HPC clusters. Intel's NetEffect™ iWARP-enabled 10GbE server cluster adapters are specifically designed to deliver this capability in HPC environments.

In fact, a large biomedical research facility has achieved excellent performance and near-linear scalability using iWARP-enabled 10GbE adapters on a cluster of 4,032 cores, as measured using the High-Performance Linpack (HPL) benchmark.[1] The cluster supports large-scale workloads in a range of critical research areas, including bioinformatics, image analysis, and sequencing, and was ranked at number 208 on the June 2010 TOP500 Supercomputing Sites list (see the "TOP500 Supercomputing Sites rankings for iWARP" sidebar in this article).

The cluster consists of 14 server racks with 36 servers per rack, for a total of 504 servers. The compute nodes are two-way Dell PowerEdge™ R610 servers based on the Intel Xeon® processor X5550 architecture at 2.66 GHz, with 24 GB of RAM and one 80 GB Serial ATA (SATA) hard drive in each server. Cluster RDMA network connectivity is provided by Intel's NetEffect iWARP-enabled 10GbE server cluster adapters.

At the rack level, each server has two connections to one of two 48-port Arista 7148SX switches: one 10GbE link (using direct-attach twinaxial cable) for RDMA traffic and one Gigabit Ethernet (GbE) link for all other traffic. Each switch has eight 10GbE uplinks (16 per rack) to a group of additional switches. Software running on the cluster includes the Red Hat® Enterprise Linux® 5.3 OS, OpenFabrics Enterprise Distribution (OFED) 1.4.1, and Intel Message Passing Interface (MPI) 3.2.1.

In October 2009, project engineers running this cluster in the lab with the HPL benchmark

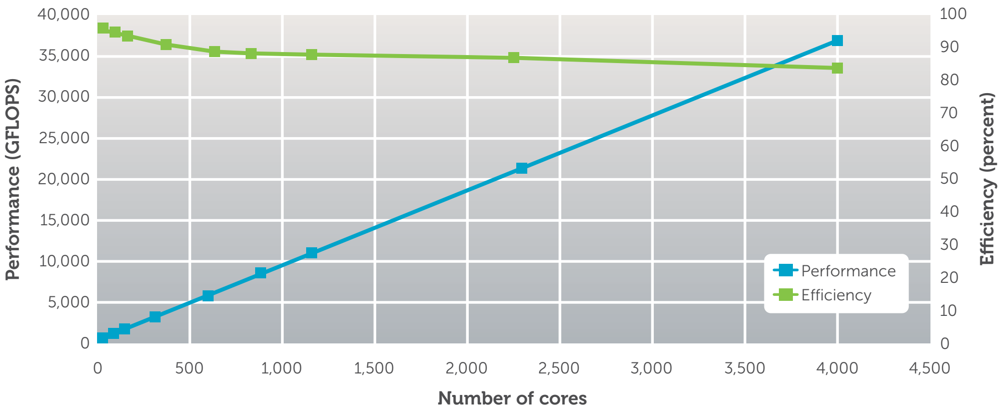

Figure 2. Performance and efficiency test results for an HPC cluster using an iWARP-enabled 10GbE fabric

[1] For more information on HPL, visit netlib.org/benchmark/hpl.

attained performance of up to 35.81 TFLOPS at 84.14 percent efficiency (see Figure 2). An HPL problem size of 1,200,000 was used, and the problem size necessary to achieve half the performance (*n*/2 problem size) was 300,000. The performance data scales in a nearly linear fashion as the number of cores applied to the benchmark workload increases. From an engineering perspective, the linearity of scaling in the results helps ensure the viability of the topology for large-scale computational problems.

Based on these results, the cluster ranks at number 84 for efficiency compared with other TOP500 clusters—a level more efficient than many InfiniBand clusters, and the highest of the listed Ethernet solutions.

Maximizing the advantages of iWARP and Ethernet

A key advantage of iWARP-enabled networking is its compatibility with existing network infrastructure, management tools, and solution stacks. Using mainstream Ethernet connectivity for compute clusters can now provide highly favorable performance, efficiency, and scalability. Taking advantage of iWARP-enabled 10GbE adapters allows RDMA traffic to be passed effectively over an Ethernet fabric. Organizations can now obtain the low latency that HPC clusters need while capitalizing on the ease of use and familiarity of Ethernet.

Using Ethernet connectivity as a unified fabric for cluster interconnects, LANs, and storage can help lower total cost of ownership by significantly reducing the number of switches and cables required. As 10GbE products and technology—including switches with high port density and technologies to further drive down latency—continue to evolve, future work can provide additional value in building supercomputing platforms on Dell PowerEdge servers interconnected with iWARP-enabled Intel 10GbE adapters. PS

Tom Stachura is the product line manager for HPC Ethernet products at Intel. He has over 17 years of industry experience in engineering, architecture, strategic planning, and product marketing.

Learn more

Intel Ethernet server adapters:
intel.com/go/ethernet
intelethernet-dell.com

Dell PowerEdge servers:
dell.com/poweredge