WHITE PAPER
Internet Wide Area RDMA Protocol (iWARP)
NetEffect™ 10 gigabit Ethernet
Server Cluster Adapters
Technical and High-Performance Computing

A High-Performance Cluster for Biomedical Research Using 10 Gigabit Ethernet iWARP Fabric

A large research institute has achieved performance of nearly 36 TeraFLOPS at greater than 84 percent efficiency using the HPL benchmark on a cluster of 4,032 cores. Intel iWARP and Arista 10Gigabit Ethernet switching technology enabled these results using 10 gigabit Ethernet (10GbE) by reducing the overhead associated with kernel-to-user context switches, intermediate buffer copies, and TCP/IP processing.

March 2010
Version 2.0

Tom Stachura
Intel LAN Access Division

Brian Yoshinaka
Intel LAN Access Division

EXECUTIVE SUMMARY

Using ubiquitous, standards-based Ethernet technology, iWARP (Internet Wide Area RDMA Protocol) enables low-latency network connectivity suitable for high-performance clusters. A key advantage of iWARP networking is its compatibility with existing network infrastructure, management solutions, and solution stacks.

This paper demonstrates the viability of cluster computing based on iWARP to achieve very high performance using 10GbE. It begins with a description of the architecture of a cluster based on iWARP connectivity before moving to a brief overview of iWARP technology. The paper concludes by reporting performance achieved using that cluster and observations about the value of iWARP to future work in this area.

A large research facility has achieved excellent performance and near-linear scalability using 10 gigabit Ethernet iWARP and NetEffect™ Adapters and Arista Networking on a cluster of 4,032 cores, as measured using the HPC LINPACK benchmark. This result represents a relatively low-cost approach to processing very large technical workloads using commercial off-the-shelf network hardware.

Architecture of an iWARP Cluster for Medical Research

To support large-scale workloads in a range of areas critical to its research, including bioinformatics, image analysis, and sequencing, a research institution has built a large (4,032 cores) cluster using Intel iWARP and Arista switching technology. For the compute nodes, they chose two-way Dell PowerEdge* R610 servers based on Intel® Xeon® processors x5550 running at 2.66 GHz with 24 GB RAM and a single 80 GB SATA hard drive in each server.
For RDMA (remote direct memory access) network connectivity, the design uses

NetEffect™ 10GbE Server Cluster Adapters. The interconnect is based on Arista 7500 spine and 7100 leaf switches.

The cluster will be used to run a variety of workloads, such as image analysis, various bioinformatics software and tools, CFD modeling, computational chemistry software, and many other open source, commercial, and in-house applications. The cluster is designed to meet all of the current scientific computational demands as well as provide a platform that will be able to handle other kinds of workloads over the cluster's lifespan.

The cluster topology, which is shown in Figure 1, consists of 14 server racks with 36 servers per rack, for a total of 504 servers. At the rack level, each server has two connections to one of two 48-port, 1U Arista 7148SX switches: one 10GbE link (using direct-attach Twinax cable) for RDMA traffic and one GbE link for all other traffic. Each Arista 7148SX switch has eight 10GbE uplinks (16 per rack) to two Arista 7508 switches configured as one MLAG (Multi-Chassis Link Aggregation Group) group.

Software running on the cluster includes Red Hat Enterprise Linux* 5.3, OFED (OpenFabrics Enterprise Distribution) 1.4.1, and Intel® MPI (Message Passing Interface) 3.2.1.

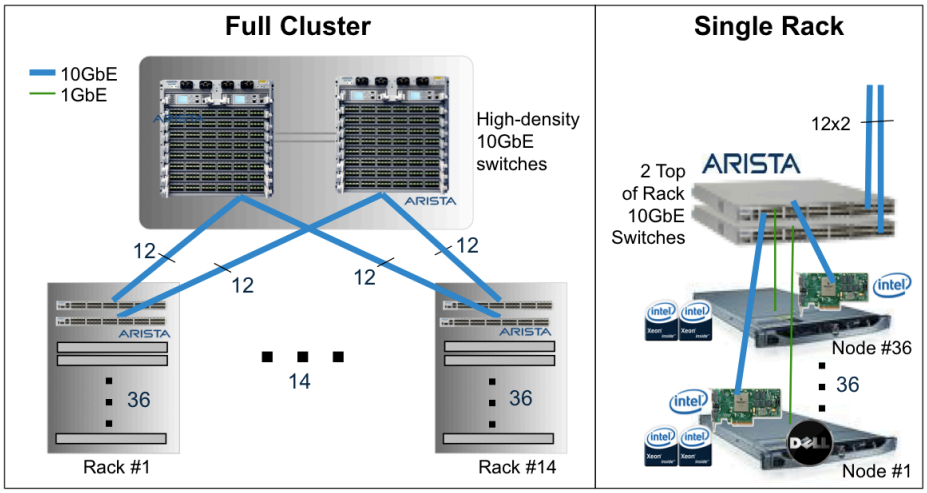

Figure 1. The cluster consists of 504 servers with two quad-core processors each, uplinked to two rack-level switches per rack, which are uplinked to a central network fabric.

Using iWARP to Lower Overhead and Latency in Multi-Gigabit Networks

Ethernet sales volume makes it extremely cost-effective for general-purpose local area network traffic, but its suitability as the underlying topology for high-performance compute clusters created a series of challenges that had to be met. The first of these was for line rate to reach a sufficiently high level, which has been achieved with the mainstream availability of 10GbE networking equipment.

To take full advantage of 10GbE line rate, however, the latency related to Ethernet networking had to be overcome. iWARP specifies a standard set of extensions to TCP/IP that define a transport mechanism for RDMA. As such, iWARP provides a low-latency means of passing RDMA over Ethernet, as depicted in Figure 2:

- Delivering a Kernel-Bypass Solution. Placing data directly in user space avoids kernel-to-user context switches, reducing latency and processor load.
- Eliminating Intermediate Buffer Copies. Data is placed directly in application buffers rather than being copied multiple times to driver and network stack buffers, reducing latency as well as memory and processor usage.
- Accelerated TCP/IP (Transport) Processing. TCP/IP processing is done in hardware instead of the operating system network stack software, enabling reliable connection processing at speed and scale.

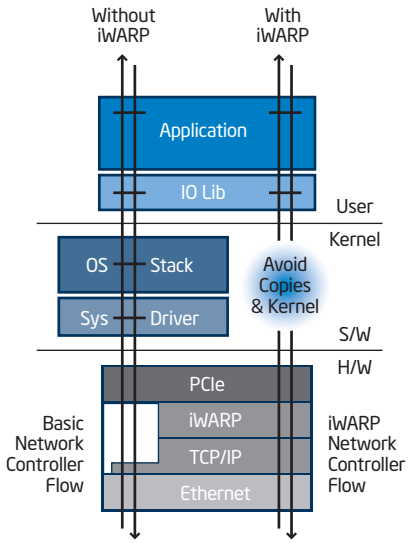

Figure 2. iWARP improves throughput by reducing the overhead associated with kernel-to-user context switches, intermediate buffer copies, and TCP/IP processing.

The iWARP protocol was developed to perform within an Ethernet infrastructure, and thus does not require any modifications to existing Ethernet networks or equipment. At the same time, iWARP's Ethernet compatibility enables IT organizations to take advantage of enhancements to Ethernet, such as Data Center Bridging, low-latency switches, and IP security.

Standard Ethernet switches and routers carry iWARP traffic over existing TCP/IP protocols. Because iWARP is layered over TCP, network equipment doesn't need to process the iWARP layer, nor does it require any special-purpose functionality. This enables the use of industry-accepted management consoles that use existing IP management protocols. The Open Fabrics Alliance (www.openfabrics.org) provides an open source RDMA software stack that is hardware-agnostic and application-agnostic for iWARP. These characteristics allow iWARP to be readily integrated into existing environments while meeting stringent cost and performance requirements.

Performance and Scalability Results

Using this cluster in the lab with the HPL benchmark running on 4,000 cores, project engineers attained performance of 35.81 TeraFLOPS at 84.14 percent efficiency, as shown in Figure 3. The HPL problem size used was 1,200,000, and the problem size necessary to achieve half the performance (N/2 problem size) was 300,000. Importantly, 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. This cluster demonstrates the best efficiency in an Ethernet solution compared to the systems on the June 2009 Top500* list, as well as a placement within the range of the top 30 x86 clusters for efficiency on that list overall. Moreover, because the data does not show an obvious drop-off in efficiency at this cluster size, it suggests that the solution is scalable beyond the size shown here, although that hypothesis would need to be tested to verify its validity. From a budgetary perspective, the results demonstrate that each compute node added to the cluster, up to at least 500 nodes, provides value commensurate with the overall cost of the cluster.

These performance and efficiency results must be considered in the context that this cluster configuration oversubscribes the connections to the Arista 7508 switches by a factor of 2.475 to 1. Making additional connections from the racks to the network fabric using free ports to reduce the oversubscription could potentially result in higher performance. This is a possible area for future inquiry.

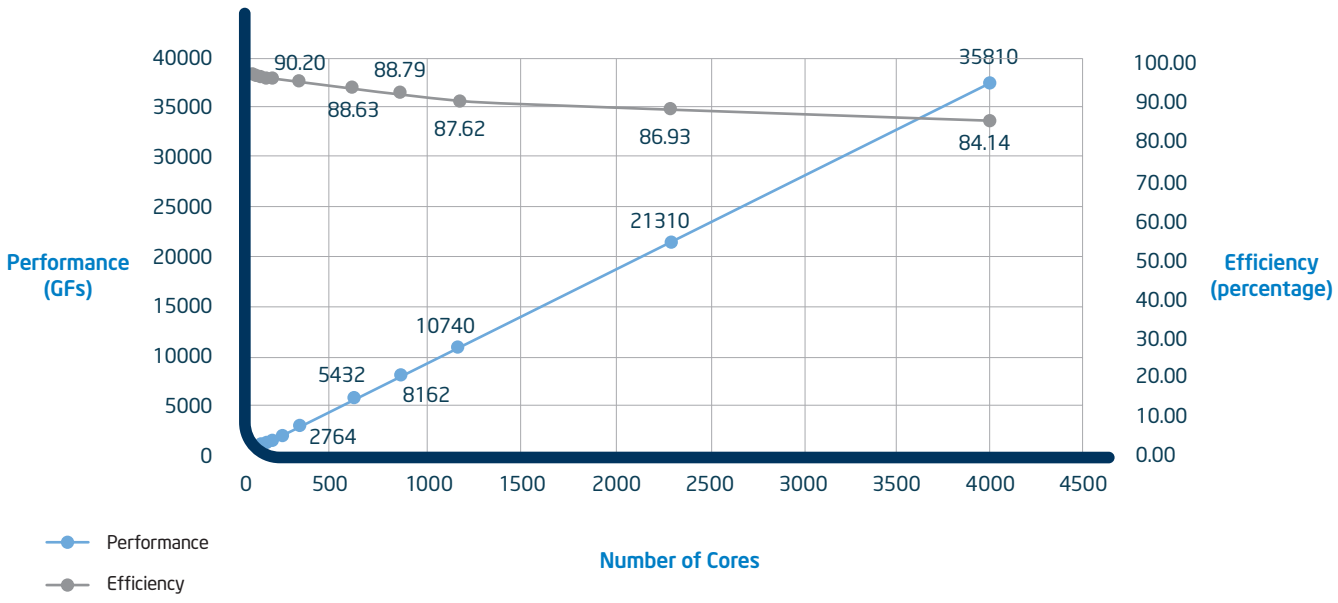

Figure 3. As measured using the HPC LINPACK benchmark, the cluster achieves performance of 35.81 TeraFLOPS at 84.14 percent efficiency using iWARP and 10 gigabit Ethernet.

Conclusion

The lab results with a synthetic benchmark described in this paper suggest substantial value of this cluster topology for future research. These findings show that using mainstream Ethernet technology for compute clusters can now provide very favorable performance, efficiency, and scalability. Taking advantage of iWARP with NetEffect 10GbE Server Cluster Adapters allows RDMA traffic to be passed effectively over Ethernet fabric. Taking advantage of Arista's high density, low latency switching infrastructure allows to build clusters with thousands of nodes. Future work as 10GbE products and technology continue to evolve, including higher-port-density switches and technologies to further drive down latency, promises additional value in using Ethernet to build supercomputing platforms.

ADDITIONAL RESOURCES

For more information on the technologies, products, and implementations described in this paper, see the following resources:

10 Gb iWARP-enabled NetEffect™ Ethernet Server Cluster Adapters:
www.intel.com/Products/Server/Adapters/Server-Cluster/Server-Cluster-overview.htm

Arista switches: www.aristanetworks.com

Dell PowerEdge* R610 servers:
www.dell.com/us/en/business/servers/server-poweredge-r610/pd.aspx?refid=server-poweredge-r610&cs=04&s=bsd

HPL benchmark Web site: www.netlib.org/benchmark/hpl/

For more information about iWARP, see the paper, "Understanding iWARP: Eliminating Overhead and Latency in multi-Gb Ethernet Networks," which is available at: http://download.intel.com/support/network/adapter/pro100/sb/understanding_iwarp.pdf.

More information about the Arista 7508 switch can be found at: http://www.aristanetworks.com/en/products/7508

More information about the Arista 7100 series of switches can be found at: http://www.aristanetworks.com/en/products/7100series

More information about Arista Multi-Chassis Link Aggregation (MLAG) can be found at: http://www.aristanetworks.com/media/system/pdf/AristaMLAG_v2_tn.pdf

Printed in USA 1109/BY/MESH/PDF ♻ Please Recycle 322957-001US