40Gb Ethernet: A Competitive Alternative to InfiniBand

LAMMPS, WRF and Quantum ESRESSO Modeling with 40Gb iWARP Technology

Ron Kunkel
IBM Systems and Technology Group

Tom Reu
Chelsio Communications

Executive Overview

The use of InfiniBand as interconnect technology for High-Performance Computing (HPC) applications has been increasing over the past few years, replacing aging Gigabit Ethernet technology as the most commonly used fabric in the Top 500 list. One of the main reasons for preferring InfiniBand over Ethernet is InfiniBand's native support for Remote Direct Memory Access (RDMA), a technology that forms the basis for high performance Message Passing Interface (MPI) implementations.

Today, a mature competitive RDMA solution over Ethernet – the iWARP protocol – is available at 40Gbps and enables MPI applications to run unmodified over the familiar and preferred Ethernet technology. Offering the same API to applications and included within the same middleware distributions, iWARP (Internet Wide Area RDMA Protocol) can be dropped in seamlessly in place of the InfiniBand fabric. With the availability of 40Gb Ethernet, the performance gap between Ethernet and InfiniBand options has been virtually closed. This paper supports this conclusion with three real application benchmarks running on IBM's Rackswitch G8316, a 40Gb Ethernet aggregation switch, in conjunction with Chelsio Communications' 40Gb Ethernet Unified Wire network adapter. This paper shows how iWARP offers comparable application level performance at 40 Gbps with the latest InfiniBand FDR speeds.

What is iWARP?

iWARP, the standard for RDMA over Ethernet, is a low latency solution for supporting high-performance computing over TCP/IP. Standardized by the Internet Engineering Task Force (IETF) and supported by the industry's leading Ethernet vendors, iWARP works with existing Ethernet switches and routers to deliver low latency fabric technology for high-performance data centers.

In addition to providing all of the total cost of ownership benefits of Ethernet, iWARP delivers several distinct advantages for use with Ethernet in HPC environments:

- It is a multivendor solution that works with legacy switches
- It is an established IETF standard
- It is built on top of IP, making it routable and scalable from just a few nodes to thousands of collocated or geographically dispersed endpoints
- It is built on top of TCP, making it highly reliable and resilient to adverse network conditions
- It uses the familiar TCP/IP/Ethernet stack and therefore leverages all the existing traffic monitoring and debugging tools
- It allows RDMA and MPI applications to be ported from InfiniBand interconnect to IP/Ethernet interconnect in a seamless fashion

What is LAMMPS?

LAMMPS ("Large-scale Atomic/Molecular Massively Parallel Simulator") is a molecular dynamics program from Sandia National Laboratories. LAMMPS makes use of MPI for parallel communication. LAMMPS was originally developed under a Cooperative Research and Development Agreement (CRADA) between two laboratories from United States Department of Energy and three other laboratories from private sector firms. It is currently maintained and distributed by researchers at the Sandia National Laboratories and is free, open-source software, distributed under the terms of the GNU General Public License

What is WRF?

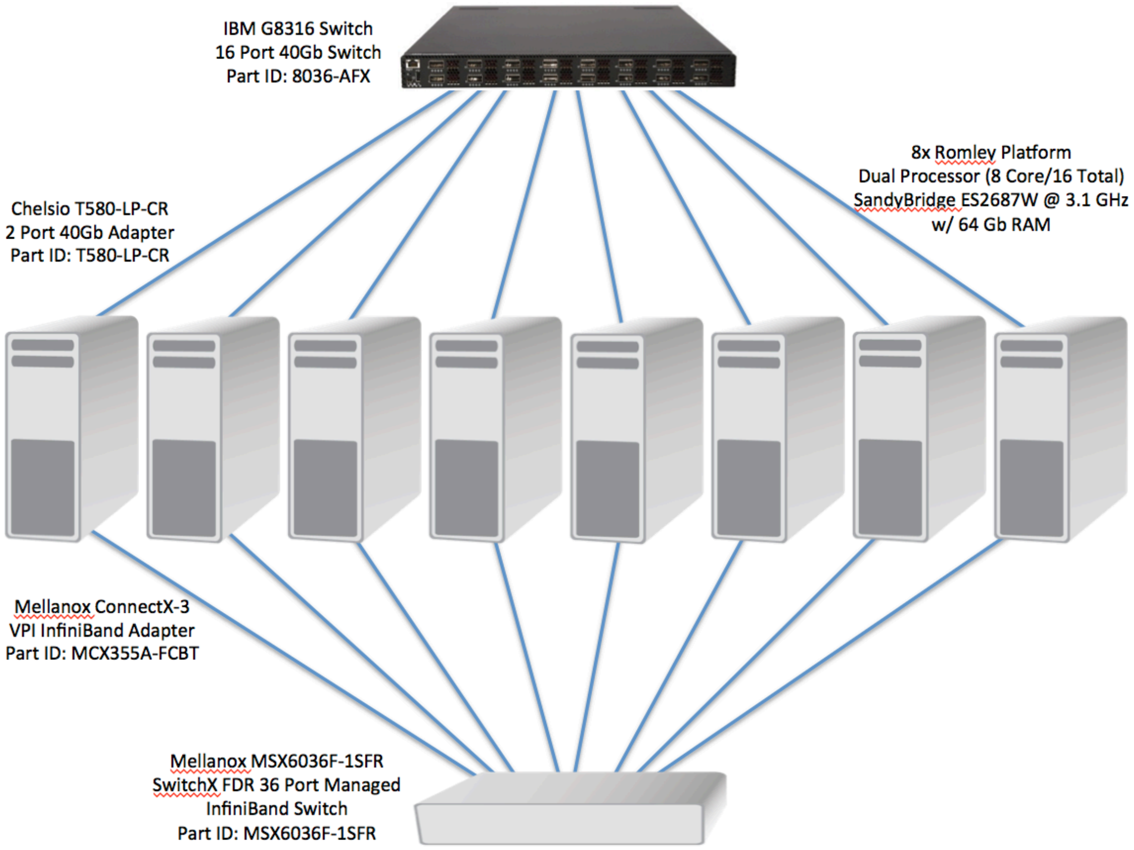
The Weather Research and Forecasting model (WRF) is a freely available program used for weather forecasting and research. It was created through a partnership of the National Oceanic and Atmospheric Administration (NOAA), the National Center for Atmospheric Research (NCAR), and more than 150 other organizations and universities in the US and other countries.

What is Quantum ESPRESSO?

Quantum ESPRESSO (opEn-Source Package for Research in Electronic Structure, Simulation, and Optimization) is an integrated suite of open-source computer codes under the GNU General Public License. This suite of software tools is typically used for electronic-structure calculations and nanoscale materials modeling.

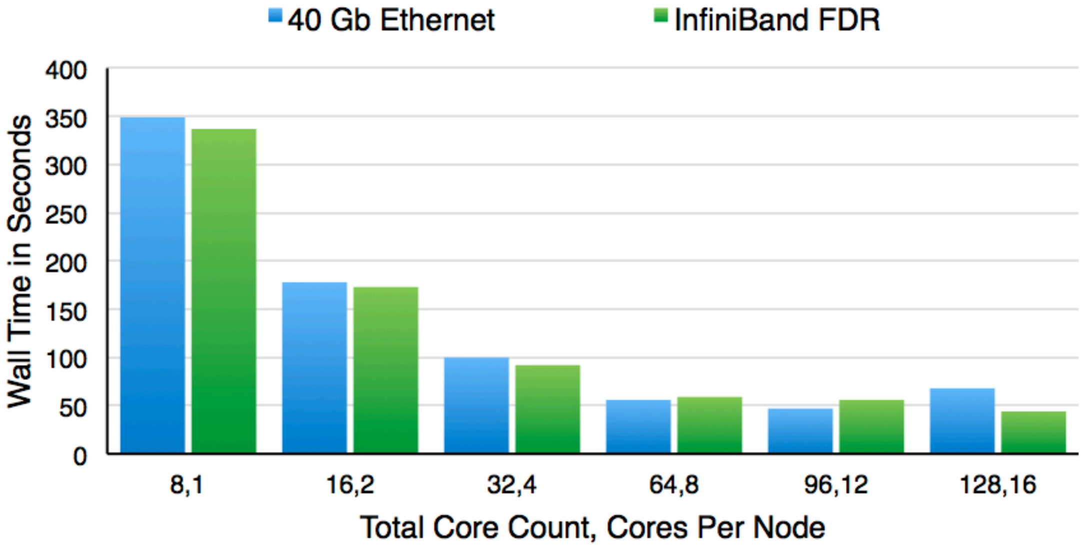
Test Setup

The following figure shows the testbed configuration employed for this evaluation effort.



The testbed used in these tests consists of eight servers dual connected to a 40Gb Ethernet network and to the latest InfiniBand FDR fabric. Identical tests were run using the two fabrics for an objective comparison. The cluster is implemented using the oneSIS open-sourced software tool, from Sandia National Laboratories, where there is a head node that hosts the root file systems over Network File System (NFS) for the other nodes. The NFS traffic as well as PXE boot goes through the same Chelsio Communications' T580-LP-CR network adapter as iWARP traffic, thereby providing a total converged fabric cluster where a single Ethernet link carries all communications for the cluster.

LAMMPS Test Results

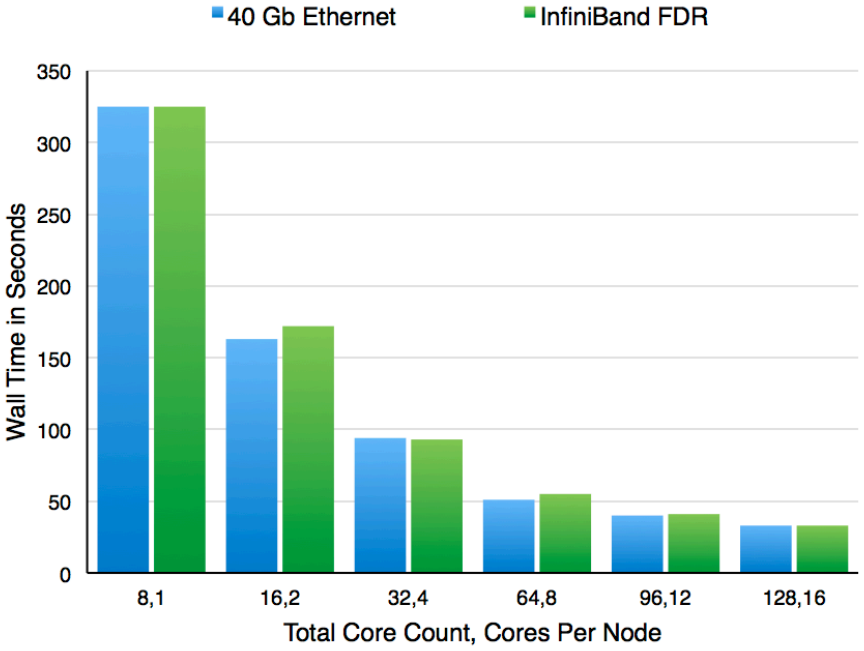


LAMMPS Command Line Used

```
mpirun -np [x] -npernode [y] --hostfile $HOME/hostfile --bynode --mca btl openib,sm,self  --mca btl_openib_if_include [cxgb4_0|mlx4_0] --mca btl_openib_connect_rdmacm_port 64000 /root/lammps-16Aug13/src/lmp_openmpi <in.melt
```

The test results clearly illustrate the fact that with real applications, 40Gb Ethernet iWARP and InfiniBand FDR performance is nearly identical.

WRF Test Results

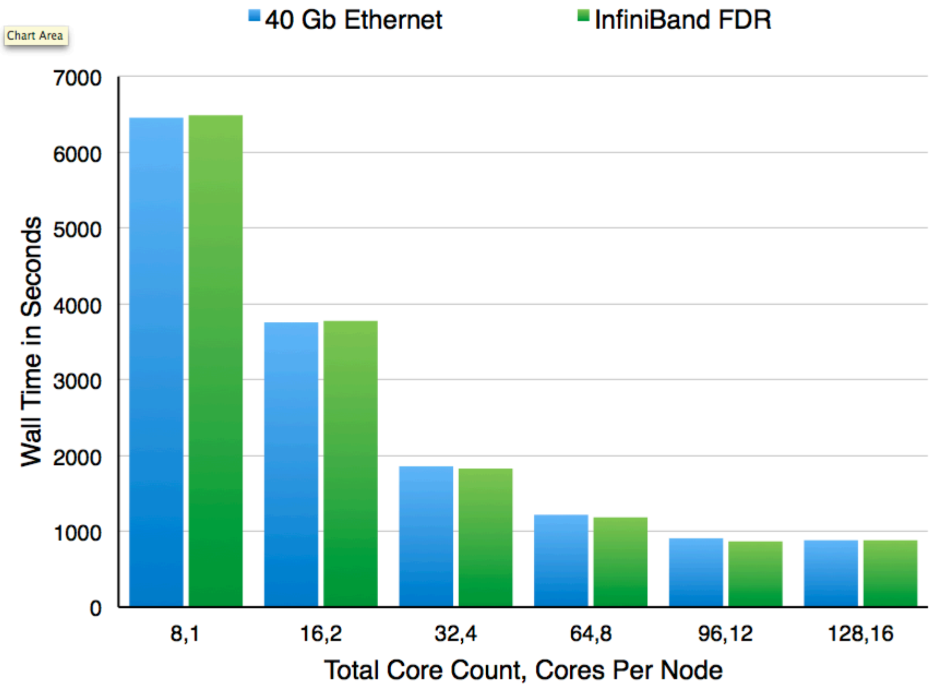


WRF Command Line Used

```
mpirun -np x -npernode x --hostfile $HOME/hostfile --bynode --mca btl openib,sm,self  --mca btl_openib_if_include [cxgb4_0|mlx4_0] --mca btl_openib_connect_rdmacm_port 6 00 numactl -c 0  ./wrf.exe
```

This test is a subset of the 12KM CONUS benchmark. Results for WRF testing again show parity between 40Gb Ethernet iWARP and InfiniBand FDR testing, with a slight edge to 40Gb Ethernet iWARP.

Quantum ESPRESSO Test Results



Quantum ESPRESSO Command Line Used
mpirun -np x -npernode x --hostfile $HOME/hostfile –bynode –bind-to-core -bycore --mca btl openib,sm,self --mca btl_openib_if_include [cxgb4_0|mlx4_0] /root/espresso/bin/pw.x –npool 2 –ntg 1 –ndiag 25

The results above are from running AUSURF112 which is a medium size input dataset for Plane-Wave Self-Consistent Field (PWscf), a part of the DEISA benchmark test. Again, 40Gb Ethernet iWARP shows a slight edge over InfiniBand FDR, corroborating the trend seen in the previous benchmark tests.

Conclusions

iWARP RDMA technology, as employed by IBM's Rackswitch G8316 40Gb Ethernet switch and Chelsio Communications' Unified Wire line of 40Gb Ethernet adapters, is an attractive plug-and-play alternative to InfiniBand FDR that provides equivalent application performance levels, and closes the gap that so far has separated the raw capabilities of these two fabrics. This eliminates any perceived drawback to Ethernet, allowing unqualified access to its advantages of ubiquity, familiarity, ease of use and flexibility. In fact, using the IBM Rackswitch G8316 GbE switch and the Chelsio Communications Unified Wire adapter, users can create and maintain a true converged fabric cluster. In this configuration all storage and networking cluster traffic runs over a single 40Gb Ethernet network, rather than having to build and maintain multiple networks, which results in significant acquisition and operational savings.

About the IBM System Networking Rackswitch G8316

Designed with top performance in mind, the IBM Rackswitch G8316 provides line-rate, high-bandwidth switching, filtering, and traffic queuing without delaying data. Large data center grade buffers keep traffic moving. Hot-swappable, redundant power and fans along with numerous high availability features enable the RackSwitch G8316 to be available for business-sensitive traffic. The RackSwitch G8316 offers up to 16×40 Gb Ethernet ports, which can also be used as a high-density 10Gb Ethernet switch, with 1.28 Tbps—in a 1U footprint. The G8316 provides a cost-efficient way to aggregate multiple racks of servers compared to other expensive core switches, while allowing massive scalability for your data center network.

About Chelsio Communications

Chelsio Communications is a leading technology company focused on solving high performance networking and storage challenges for virtualized enterprise data centers, cloud service installations, and cluster computing environments. Now shipping its fourth generation protocol acceleration technology, Chelsio Communications is delivering hardware and software solutions including Unified Wire Ethernet network adapter cards, unified storage software, high performance storage gateways, unified management software, bypass cards, and other solutions focused on specialized applications. Chelsio’s T580-LP-CR is a dual port 40 Gigabit Ethernet Unified Wire adapter with PCI Express 3 host bus interface, optimized for cloud computing, HPC, virtualization, storage, and other data center applications.

Chelsio
Communications
Accelerate

For More Information

IBM System Networking RackSwitch	http://www.ibm.com/systems/networking/switches/rack.html
Chelsio T5 Unified Wire Adapters	http://www.chelsio.com/nic/t5-unified-wire-adapters/

Legal Information

IBM Systems and Technology Group
Dept. U2SA
3039 Cornwallis Road
Research Triangle Park, NC 27709

Produced in the USA
September 2013

For a copy of applicable product warranties, write to: Warranty Information, P.O. Box 12195, RTP, NC 27709, Attn: Dept. JDJA/B203. IBM makes no representation or warranty regarding third-party products or services including those designated as ServerProven® or ClusterProven®. Telephone support may be subject to additional charges. For onsite labor, IBM will attempt to diagnose and resolve the problem remotely before sending a technician.

IBM, the IBM logo, ibm.com, RackSwitch, RackSwitch Solution Partner, are trademarks of IBM Corporation in the United States and/or other countries. If these and other IBM trademarked terms are marked on their first occurrence in this information with a trademark symbol (® or ™), these symbols indicate U.S. registered or common law trademarks owned by IBM at the time this information was published. Such trademarks may also be registered or common law trademarks in other countries. For a list of additional IBM trademarks, please see www.**ibm.com**/legal/copytrade.shtml.

InfiniBand is a trademark of InfiniBand Trade Association.

Other company, product, and service names may be trademarks or service marks of others.

IBM reserves the right to change specifications or other product information without notice. References in this publication to IBM products or services do not imply that IBM intends to make them available in all countries in which IBM operates. IBM PROVIDES THIS PUBLICATION “AS IS” WITHOUT WARRANTY OF ANY KIND, EITHER EXPRESS OR IMPLIED, INCLUDING THE IMPLIED WARRANTIES OF MERCHANTABILITY AND FITNESS FOR A PARTICULAR PURPOSE. Some jurisdictions do not allow disclaimer of express or implied warranties in certain transactions; therefore, this statement may not apply to you.

This publication may contain links to third party sites that are not under the control of or maintained by IBM. Access to any such third party site is at the user's own risk and IBM is not responsible for the accuracy or reliability of any information, data, opinions, advice or statements made on these sites. IBM provides these links merely as a convenience and the inclusion of such links does not imply an endorsement.

Information in this presentation concerning non-IBM products was obtained from the suppliers of these products, published announcement material or other publicly available sources. IBM has not tested these products and cannot confirm the accuracy of performance, compatibility or any other claims related to non-IBM products. Questions on the capabilities of non-IBM products should be addressed to the suppliers of those products.

MB, GB and TB = 1,000,000, 1,000,000,000 and 1,000,000,000,000 bytes, respectively, when referring to storage capacity. Accessible capacity is less; up to 3GB is used in service partition. Actual storage capacity will vary based upon many factors and may be less than stated.

Performance is in Internal Throughput Rate (ITR) ratio based on measurements and projections using standard IBM benchmarks in a controlled environment. The actual throughput that any user will experience will depend on considerations such as the amount of multiprogramming in the user’s job stream, the I/O configuration, the storage configuration and the workload processed. Therefore, no assurance can be given that an individual user will achieve throughput improvements equivalent to the performance ratios stated here.

Maximum internal hard disk and memory capacities may require the replacement of any standard hard drives and/or memory and the population of all hard disk bays and memory slots with the largest currently supported drives available. When referring to variable speed CD-ROMs, CD-Rs, CD-RWs and DVDs, actual playback speed will vary and is often less than the maximum possible.