LAMMPS and WRF on iWARP vs. InfiniBand FDR

The use of InfiniBand as interconnect technology for HPC applications has been increasing over the past few years, replacing the aging Gigabit Ethernet as the most commonly used fabric in the Top500 list. One of the main reasons for preferring IB over Ethernet is IB'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 inboxed within the same middleware distributions, iWARP (Internet Wide Area RDMA Protocol) can be dropped in seamlessly in place of the esoteric fabric. With the availability of 40Gbs Ethernet, the performance gap between Ethernet and InfiniBand has been virtually closed. This paper supports this conclusion with two real application benchmarks running on Chelsio's 40Gb Unified Wire, showing how iWARP offers comparable application level performance at 40Gbps and the latest FDR IB 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 (TCO) 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 (IB) interconnect to IP/Ethernet interconnect in a **seamless** fashion

What is LAMMPS?

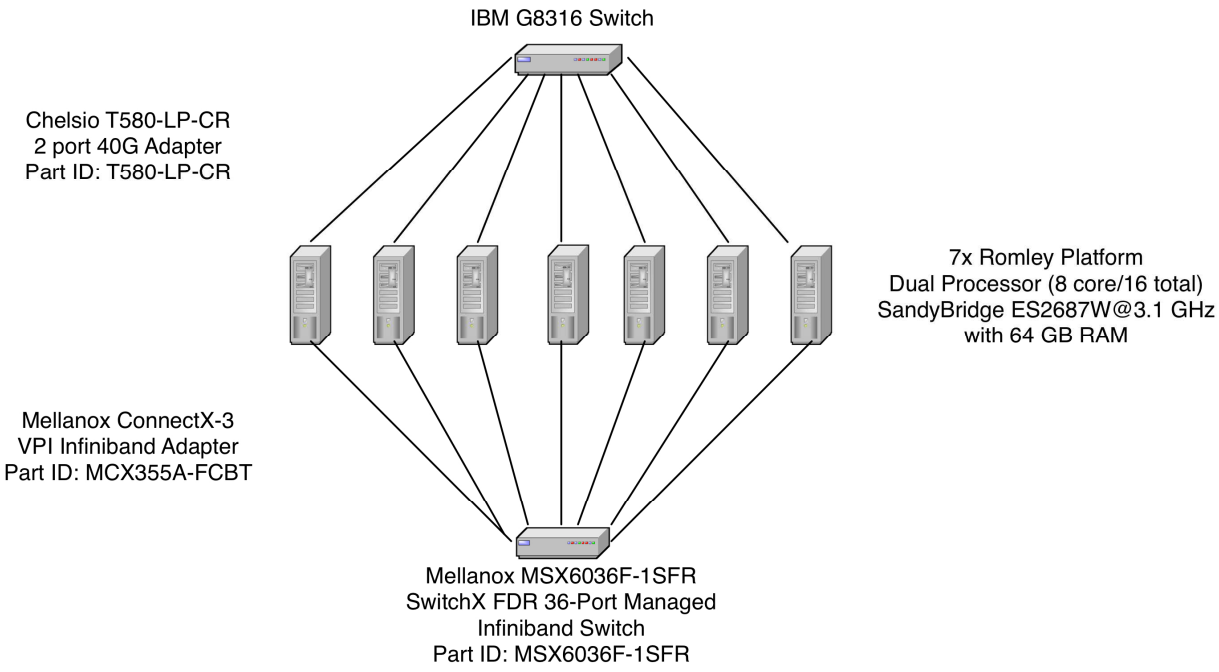
LAMMPS ("Large-scale Atomic/Molecular Massively Parallel Simulator") is a molecular dynamics program from Sandia National Laboratories. LAMMPS makes use of MPI for parallel communication. LAMMPS was originally developed under a Cooperative Research and Development Agreement (CRADA) between two laboratories 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.

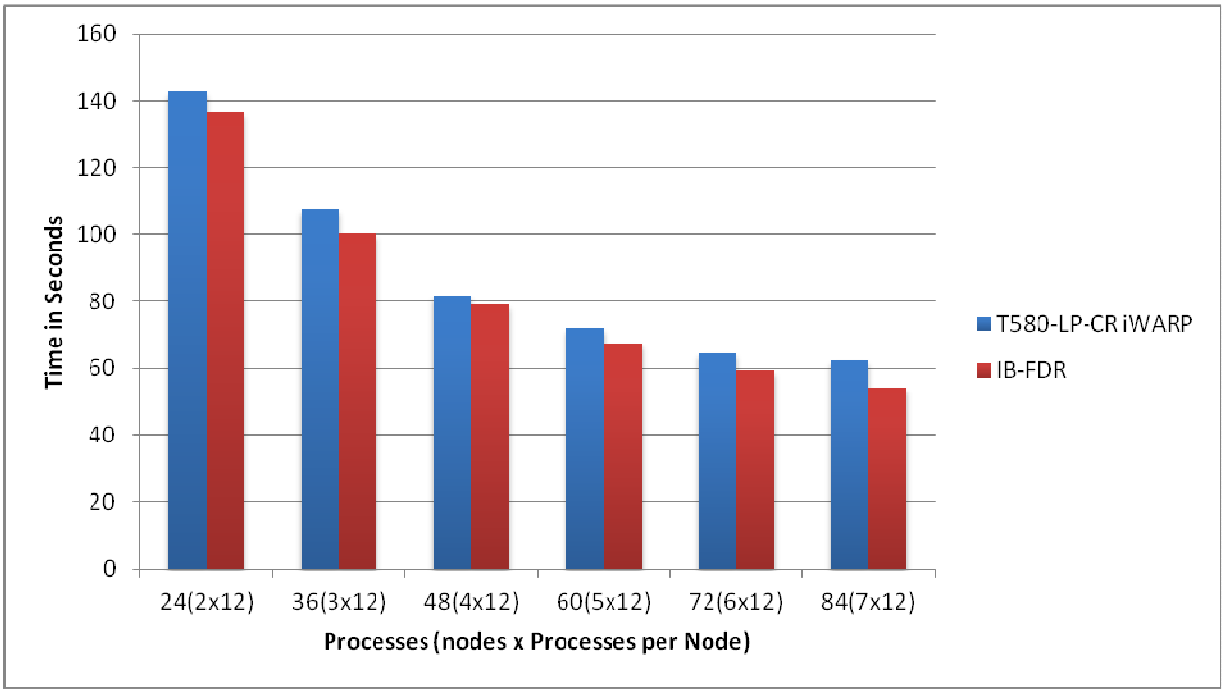
Test Setup

The following figure shows the testbed configuration.



The testbed used in these tests consists of seven servers dual connected to a 40Gbs Ethernet network and to the latest FDR IB fabric. Identical tests were run using the two fabrics for an objective comparison. The cluster is implemented using one SIS where there is a head node that hosts the root file systems over NFS for the other nodes. The NFS traffic as well as PXE boot goes through the same Chelsio T580-LP-CR as iWARP traffic, thereby providing a total converged fabric cluster where a single Ethernet link carries all communications for the cluster!

LAMMPS Test Results

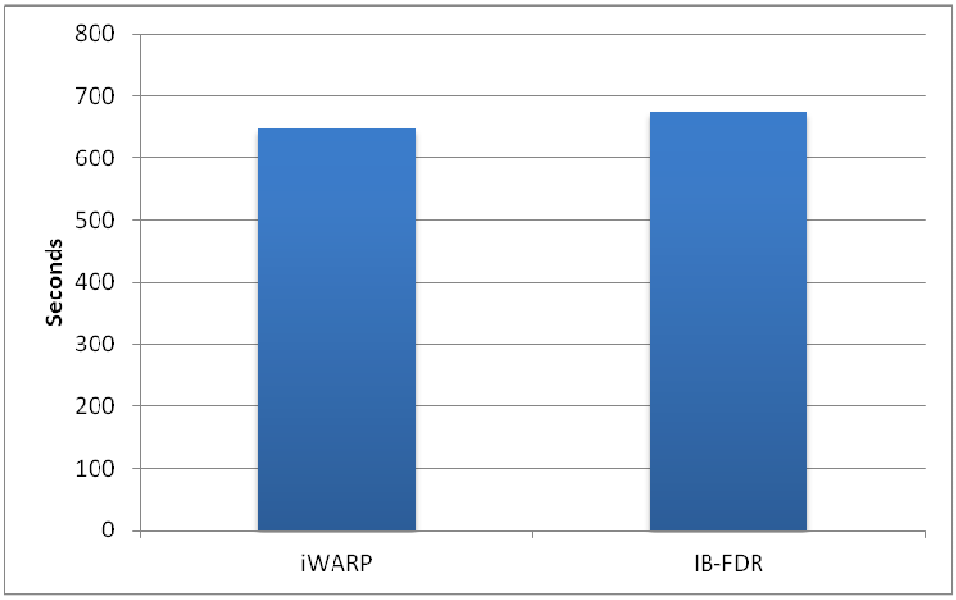


LAMMPS Command Line Used

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, 40Gbps iWARP and FDR IB performance is nearly identical.

WRF 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 64000 numactl -c 0 ./wrf.exe

This test is a subset of the 12KM Conus benchmark. The results for WRF again show parity between the results running over iWARP and those over InfiniBand, with a slight edge to iWARP in this case.

Conclusions

iWARP RDMA, as provided by Chelsio's new line of 40Gb adapters, is an attractive plug-and-play alternative to FDR InfiniBand that provides equivalent application performance levels, and closes the gap that so far has separated the raw capabilities of the 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 a Chelsio adapter along with the Unified Wire Software package available as part of the Chelsio solution, users can create and maintain a true Converged Fabric cluster where all storage and networking cluster traffic runs over a single 40Gb network, rather than having to build and maintain multiple networks, resulting in significant acquisition and operational savings.

About Chelsio

Chelsio 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 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.