LAMMPS, LS-DYNA, HPL, 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. The main reason for preferring IB over 10Gbps 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 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. While current iWARP solutions are 10Gbps Ethernet-based, higher speed 40Gbps and 100Gbps implementations are slated for imminent availability. Nevertheless, as this paper shows with real application benchmarks, iWARP today offers competitive application level performance at 10Gbps against 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 10GbE 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 LS-DYNA?

LS-DYNA is a general-purpose transient-dynamic finite-element program designed to simulate complex real-world problems developed by Livermore Software Technology Corporation (LSTC). It is optimized for shared and distributed memory for Linux and Windows platforms. LS-DYNA is scalable code for solving highly nonlinear transient problems enabling the solution of coupled multi-physics and multi-stage problems.

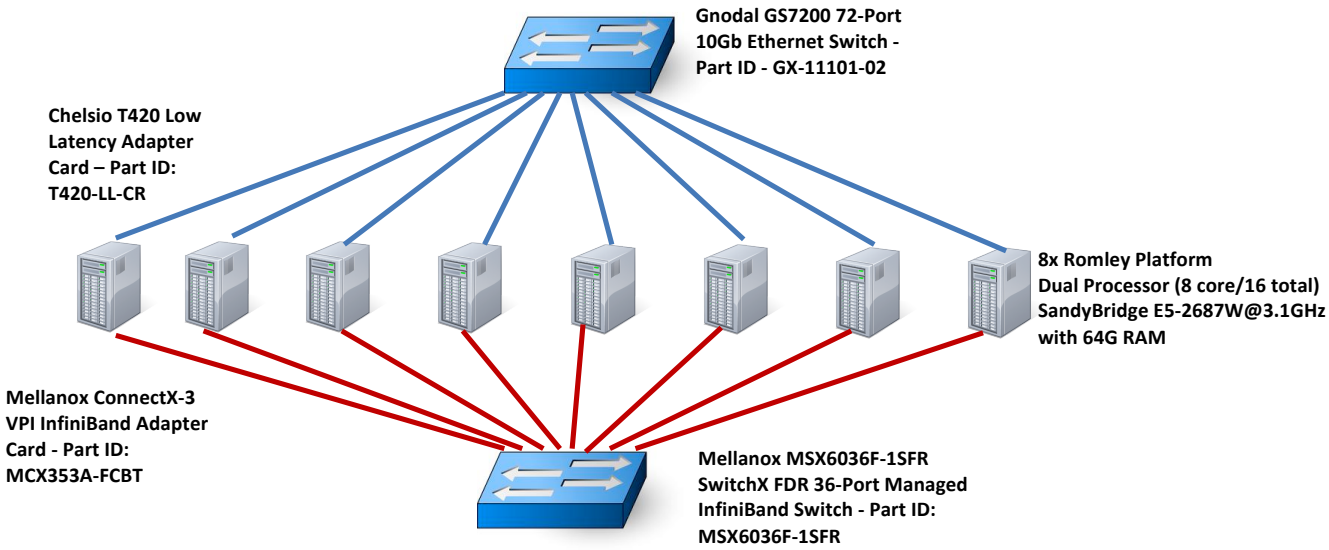
What is HPL?

HPL is a portable as well as freely available implementation of the High Performance Computing Linpack Benchmark. The application solves a random dense linear system in double precision arithmetic on distributed-memory systems. Linpack has long been considered *THE* benchmark to run for the Top-500 HPC system list.

What is WRF?

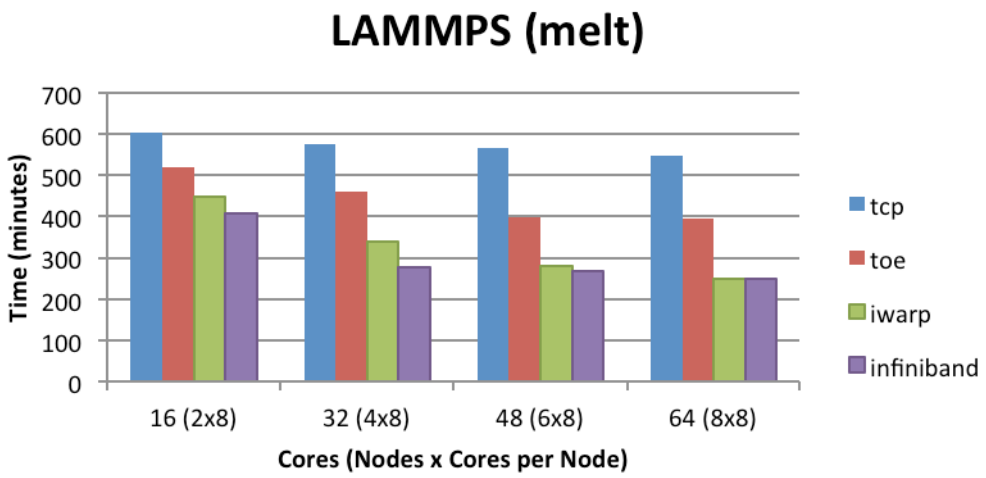
The Weather Research and Forecasting model (WRF) is a freely available program used for weather forecasting and research. It was created through a partnership of the National Oceanic and Atmospheric Administration (NOAA), the National Center for Atmospheric Research (NCAR), and more than 150 other organizations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

Test Setup



The testbed used in the benchmarks consists of a number of servers dual connected to a 10Gbps Ethernet network and to the latest FDR IB fabric. Identical tests were run using the two fabrics for an objective comparison. On the Ethernet side, the tests were conducted using simple NIC attachment (stateless offload), TOE (TCP/IP full offload) and iWARP.

LAMMPS Test Results

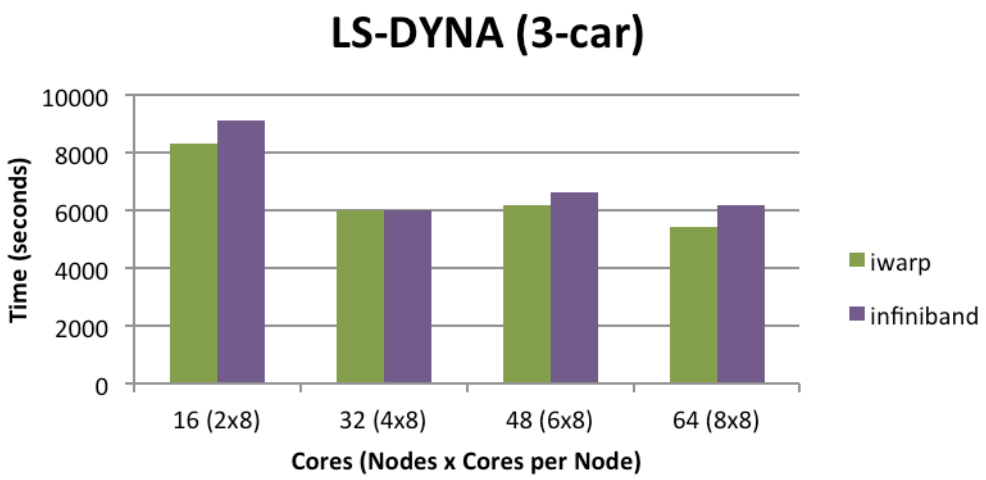


LAMMPS Command Line Used

mpirun -np <nodes> -hostfile /root/hostfile -npernode <processes-per-node> -mca btl_openib_receive_queues P,65536,256,192,128 -mca orte_base_help_aggregate 0 /root/lammps-23Oct12/src/lmp_openmpi </root/lammps-23Oct12/examples/<test-name>/in.<test-name>

The test results clearly illustrate the fact that with real applications, despite the large difference in theoretical bandwidth, 10Gbps iWARP and FDR IB performance is nearly identical, with a diminishing difference as the testbench is scaled up, demonstrating iWARP superior scalability. The numbers also show that TCP offload improves on NIC performance, but mainly demonstrate the value of RDMA in lifting Ethernet up to competitive levels.

LS-DYNA Test Results

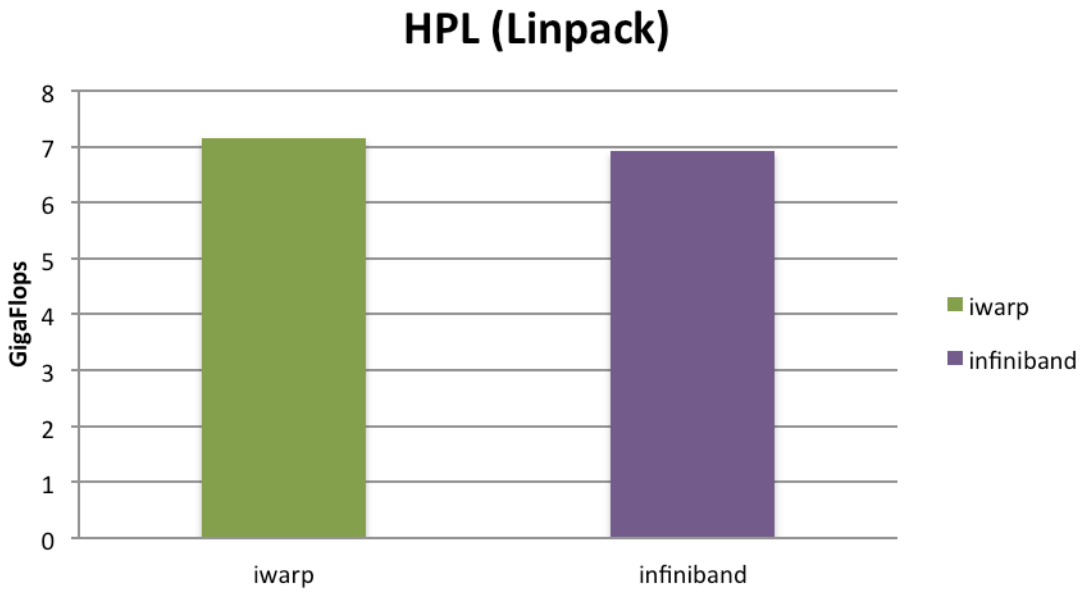


LS-DYNA Command Line Used

mpirun -np <nodes> -hostfile /root/hostfile -npernode <process-per-node> -mca btl_openib_receive_queues P,65536,256,192,128 -mca orte_base_help_aggregate 0 -x LD_LIBRARY_PATH /root/ls-dyna_mpp_d_r6_1_0_74904_x64_redhat54_ifort101_sse2_openmpi151

The test results for LS-DYNA with 10Gbps iWARP besting FDR IB on all counts may be surprising to some, but highlight again the importance of application level benchmarking compared to pure micro-benchmarks, which can often be misleading.

HPL Results



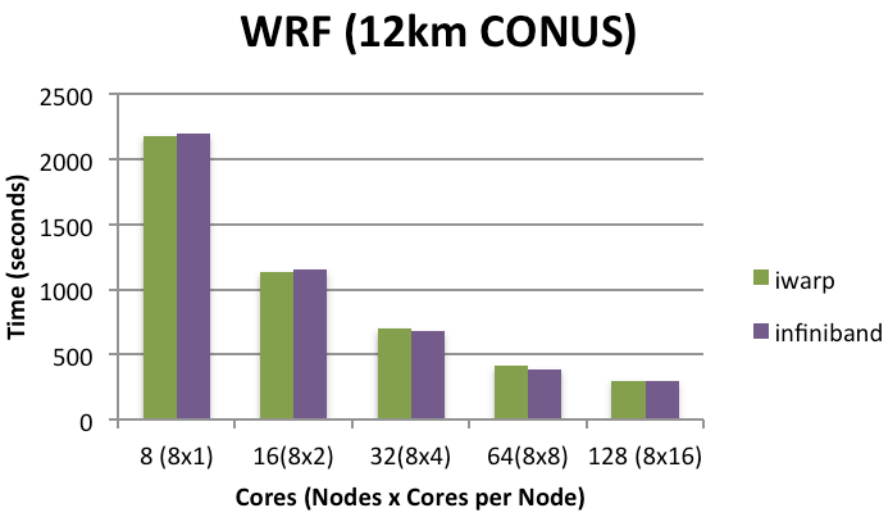
HPL Command Line Used

mpirun -np 180 -hostfile /root/hostfile -npernode 30 -mca btl_openib_warn_no_device_params_found 0 ./xhpl

The HPL.dat file used for this test is available from Chelsio if requested.

The test results for HPL again show 10GBps edging out FDR IB despite the significant difference in raw bandwidth between the two. These numbers further demonstrate the difference between actual application performance and simple microbenchmarks, which can be artificially optimized without real effect on useful applications.

WRF Results



WRF Command Line Used

mpirun -np 128 --hostfile /root/hostfile -mca orte_base_help_aggregate 0 -mca orte_no_session_dir 0 -npernode 16 -mca btl_openib_if_include cxgb4_0 numactl -c 0 ./wrf.exe

The test results for WRF again show parity between the results running over iWARP and those over InfiniBand. These results are a further example where microbenchmarks fail to match actual application performance, confirming the observations made in all previous tests.

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.