PENNSTATE

WHITE PAPER

Michael Fenn
Lazaro Calderin
Jeffrey Nucciarone
Vikas Argod

Evaluation of iWARP versus InfiniBand® Performance

High Performance Computing over 10 Gigabit Ethernet Fabric

Internet Wide Area RDMA Protocol (iWARP) has the potential to enable cluster-based high-performance computing using Ethernet fabric in place of conventional InfiniBand®, in some cases, resulting in lower cost and complexity of the environment. This paper tests side-by-side performance of iWARP and InfiniBand on several high performance computing applications.

AN ALTERNATIVE NETWORK TOPOLOGY FOR LARGE-SCALE COMPUTATION

Options have conventionally been limited for providing the low network latency and high network bandwidth required for large-data computations on high performance computing clusters. InfiniBand® has been the fabric of choice in many research environments for its ability to support the low-latency, large-scale data transfer required for calculations associated with both natural phenomena such as weather and molecular dynamics and commercial problems such as optimizing oil and gas recovery and industrial-design simulation.

InfiniBand has typically been a suitable choice for these applications, but its use also has a number of limitations, chiefly with regard to system cost and complexity. Since nearly all computational facilities also have Ethernet environments in place, adding an InfiniBand fabric means that administrators must manage at least two network fabrics, which adds both to cost and to the amount of time and effort staff must spend maintaining the network, rather than accomplishing meaningful work. Additionally, running a separate fabric solely for computation-related communication increases the power footprint of such an environment.

While Ethernet fabrics have the advantage of reducing that cost and complexity, InfiniBand has traditionally been the performance leader in HPC. With the mainstream proliferation of 10 Gigabit Ethernet (10GbE), network bandwidth has become less of an issue, but since very low latency is not a key design consideration of the Ethernet protocols, 10GbE server adapters have only recently been able to provide a suitable alternative to InfiniBand. Internet Wide Area RDMA Protocol (iWARP) has lately emerged to take advantage of the lower cost and complexity of Ethernet fabric in low-latency applications.

This paper reports on performance testing performed by the Research Computing and Cyberinfrastructure unit of Information Technology services at Penn State to identify how well iWARP fabrics support workloads on widely used high performance computing applications compared to InfiniBand. It begins with a description of iWARP itself before describing the test environment and procedure. Then, for each application under test, the paper gives a brief description of the application and test scenario, as well as comparative test results under iWARP and InfiniBand.

IWARP ARCHITECTURE

The key design goal of iWARP is to virtually eliminate the processor overhead associated with Ethernet networking, to help bring dramatic improvements in networking performance at low cost. The approach consists of the following mechanisms:

- **Kernel bypass** removes the need for context switching from kernel-space to user-space. Traditionally, when an application issues commands such as reads and writes to a server adapter, those commands are transmitted through user-space layers of the application to kernel-space layers in the OS stack. This requires a compute-intensive context switch between user space and the OS. The iWARP extensions use Remote Direct Memory Access (RDMA) to enable the application to post commands directly to the server adapter. This capability eliminates expensive calls to the OS, and that lower overhead reduces latency.
- **Direct data placement** eliminates intermediate buffer copies by reading and writing directly to application memory. Under conventional Ethernet, data is copied (and re-cached each time) by the processor several times as it passes from the server adapter's receive buffer to the application buffer. Those operations consume time and memory bandwidth that the application could otherwise use. Using RDMA, iWARP enables direct copies from the server adapter's receive buffer to the application buffer. This provides a direct data placement implementation that eliminates the intermediate operations, which significantly reduces latency.
- **Transport acceleration** performs transport processing on the network controller instead of the processor. With traditional Ethernet, the processor dedicates substantial resources to maintaining the network stack. It must maintain connection context, segment and reassemble payloads, and process interrupts. This overhead increases linearly with wire speed, limiting scalability. The iWARP extensions enable the transport processing to be done in the network controller. This enables the processor to perform more application processing, providing a deterministic, low-latency solution that is optimized for applications that demand low latency.

By addressing the key sources of Ethernet overhead, iWARP provides several potential benefits. LAN and RDMA traffic can pass over a single wire, as well as enabling application and management traffic to be converged, reducing requirements in terms of cables, server ports, and switches. Network administrators can use standard IP tools to manage traffic in an iWARP network, with an often-favorable impact on the overall cost and complexity of operations. And because iWARP uses Ethernet and the standard IP stack, it can be supported with standard equipment and be routed across IP subnets using general-purpose network switches, appliances, and cabling.

TEST ENVIRONMENT

The test bed to compare the performance of 10GbE iWARP with that of Quad Data Rate (QDR) InfiniBand was configured as shown in Table 1.

The application software under test included the following, the results for each of which are provided in a separate section in the body of this paper:

- Abaqus 6.10
- LAMMPS 15 Jan 2010
- LS-DYNA 971_R4.2.1
- Quantum ESPRESSO 4.2.1
- VASP 5.2
- WRF 3

IWARP VERSUS INFINIBAND PERFORMANCE: ABAQUS

Abaqus® Unified Finite Element Analysis (FEA) software suite is provided by SIMULIA®, a brand owned by Dassault Systèmes, for realistic simulations of multiphysics engineering problems and lifecycle management solutions for managing simulation data, processes, and intellectual property. The Abaqus FEA suite consists of three core products (in addition to optional add-on products that address the requirements of specific targeted applications):

- **Abaqus/Standard** is a comprehensive, general-purpose finite element analysis tool that includes a variety of time-domain and frequency-domain analysis procedures.

Servers	Dell PowerEdge™ R710 Server Two Intel® Xeon® processors X5560 48 GB RAM
Network Adapters	10Gb iWARP-enabled NetEffect™ Ethernet Server Cluster Adapter from Intel Mellanox Connect-X MT26428 QDR InfiniBand® Host Channel Adapter
System Software	Red Hat Enterprise Linux 5.6 OpenFabrics Enterprise Distribution™ 1.5.2 OpenMPI 1.4.2 (except Abaqus, which uses its own HP-MPI)
Switches	iWARP: Arista 7148SX with Jumbo Frames enabled InfiniBand: Mellanox MTS3600

Table 1. Test system configuration.

- **Abaqus/Explicit**, a complement to Abaqus/Standard, is a finite element analysis program designed to serve advanced, nonlinear continuum and structural analysis requirements. The program addresses highly nonlinear transient dynamic phenomena and certain nonlinear quasi-static simulations.
- **Abaqus/CAE** provides a modeling and visualization environment for Abaqus.

Abaqus FEA is intended for use in understanding the detailed behavior of complex assemblies, refining design concepts, understanding the behavior of new materials, and simulating discrete manufacturing processes. It addresses non-linear problems, large-scale linear dynamics applications, and routine design simulations, and it includes user-programmable features, scripting, and GUI customization.

For more information, see the Abauqus FEA product website.[1]

TEST SCENARIO

The problem set in the testing scenario involved modeling a tapered, hollow wing with internal longitudinal and cross stiffeners. The wing is 5 m long and has an aerofoil profile throughout. Four- and three-noded shell elements with six degrees of freedom at each node are used. The thickness of the shells that form the skin of the wing varies from 2.5 mm on the tip to 5 mm at the root. The root of the wing is free to move in the x-y plane (i.e., the translation about the z axis and rotation about the x and y axes are constrained). A spring and damper constrain the motion of the root of the wing in the x-y plane.

At t=0.05 seconds, load application was begun in the –y direction on the top edge of the wing tip, which was linearly ramped up to 1000 N at t=0.15 seconds and then linearly ramped down to 0 at t=0.20 seconds. This model has 16,027 elements with 33,081 degrees of freedom. The total simulation time is 200 ms.

TEST RESULTS

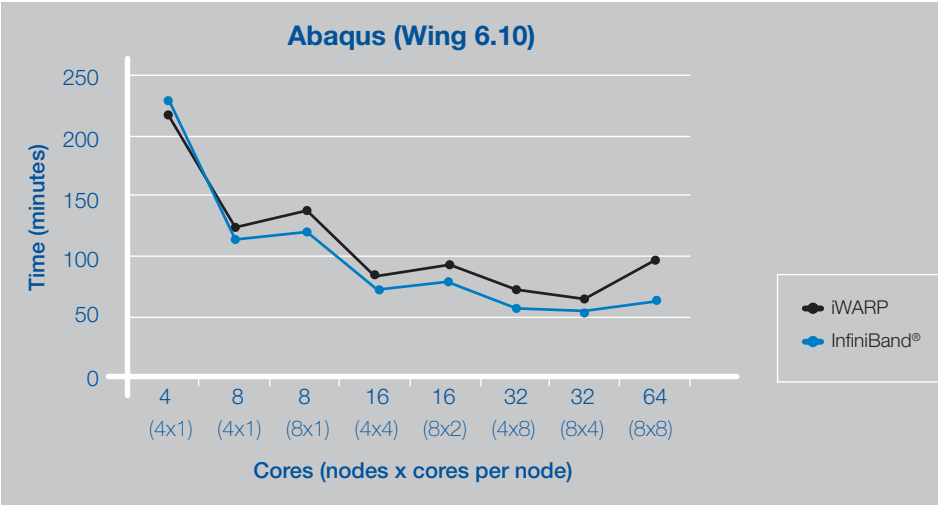

Figure 1. Abaqus iWARP versus InfiniBand® performance-testing results (lower y-axis figures are better).

IWARP VERSUS INFINIBAND PERFORMANCE: LAMMPS

LAMMPS (Large-scale Atomic/Molecular Massively Parallel Simulator) is a classical molecular dynamics code that models an ensemble of particles in a liquid, solid, or gaseous state. It can model atomic, polymeric, biological, metallic, granular, and coarse-grained systems using a variety of force fields and boundary conditions. The application can model systems with only a few particles up to several billion.

In the most general sense, LAMMPS integrates Newton's equations of motion for collections of atoms, molecules, or macroscopic particles that interact via short- or long-range forces with a variety of initial and/or boundary conditions. For computational efficiency, LAMMPS uses neighbor lists to keep track of nearby particles. The lists are optimized for systems with particles that are repulsive at short distances, so that the local density of particles never becomes too large. On parallel machines, LAMMPS uses spatial-decomposition techniques to partition the simulation domain into small 3D sub-domains, one of which is assigned to each processor. Processors communicate and store "ghost" atom information for atoms that border their sub-domain. LAMMPS is most efficient (in a parallel sense) for systems whose particles fill a 3D rectangular box with roughly uniform density.

LAMMPS is designed to be easy to modify or extend with new capabilities, such as new force fields, atom types, boundary conditions, or diagnostics. LAMMPS is a freely-available open-source code, distributed under the terms of the GNU Public License.[2] The current version is written in C++. Earlier versions were written in F77 and F90. LAMMPS was originally developed under a US Department of Energy (DOE) Cooperative Research and Development Agreement between two DOE labs and three companies. It is distributed by Sandia National Labs.[3]

LAMMPS runs efficiently on single-processor desktop or laptop machines, but it is designed for parallel computers. It will run on any parallel machine that compiles C++ and supports the MPI[4] message-passing library. This includes distributed- or shared-memory parallel machines and Beowulf-style clusters.

For more information, see the LAMMPS FAQ page.[5]

TEST SCENARIO

The lithium-ion batteries used in cell phones and laptop computers are based on a liquid electrolyte in which a lithium salt is dissolved, and lithium is the cation that is transferred across the electrolyte during charge and discharge. Replacing the liquid electrolyte with a polymer based "solid" electrolyte, termed "solid polymer electrolyte" offers advantages in weight, size, flexibility, safety, and end-of-life disposal. However, the conductivity of these electrolytes falls short of required standards. The study of cation transport in solid polymer electrolytes is very important for overcoming this challenge.

While experimental techniques provide information on the diffusion coefficient, polymer segmental relaxation, and the content of mobile ions, it is difficult to determine a transport mechanism from these measurements. This testing uses molecular dynamics simulation to study ion transport and backbone mobility of a polyethylene oxide-based single-ion conductor for potential lithium ion battery application. In single-ion conductors, or ionomers, the anion is incorporated in the polymer chain. The conductivity then arises exclusively from the cation, which can eliminate unwanted buildup of anions on the electrodes.

The simulation contains 27 molecules with a total number of atoms close to 6,000. Although this is a modest size, observation of cation dynamics into the diffusive regime requires simulation runs up to 500 ns, depending on the cation identity, the anion identity, and the temperature.

IWARP VERSUS INFINIBAND PERFORMANCE: LS-DYNA

LS-DYNA, developed by Livermore Software Technology Corporation (LSTC), is a general-purpose, transient-dynamic finite-element program designed to simulate complex real-world problems. It is optimized for shared and distributed memory UNIX®, Linux, and Windows®-based platforms.

TEST RESULTS

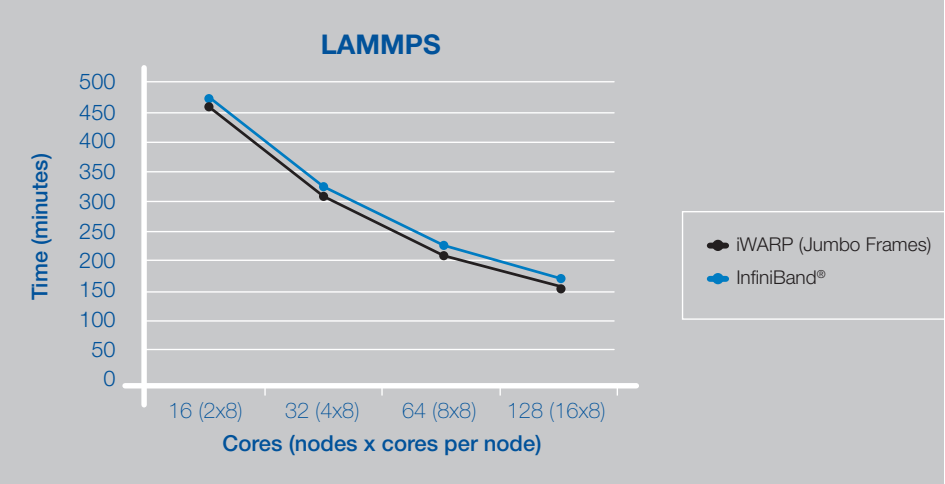

Figure 2. LAMMPS iWARP versus InfiniBand® performance-testing results (lower y-axis figures are better).

Related products include LS-OPT, a standalone Design Optimization and Probabilistic Analysis package with an interface to LS-DYNA; and LS-PrePost, an advanced interactive program used for preparing input data for LS-DYNA and processing the results from LS-DYNA analyses.

For more information, see the LS-DYNA website.[6]

TEST SCENARIO

In the test case simulated here, a van crashes into the rear of a compact car, which in turn crashes into a midsize car. Vehicle models were created by the National Crash Analysis Center (NCAC) and assembled into the input file by Mike Berger, consultant to LSTC. For this study, the input files were downloaded from the Top Crunch Project.[7] This model has 794,780 elements with six contact surfaces and 1,052 materials. The simulation time of collision is 150 ms.

TEST RESULTS

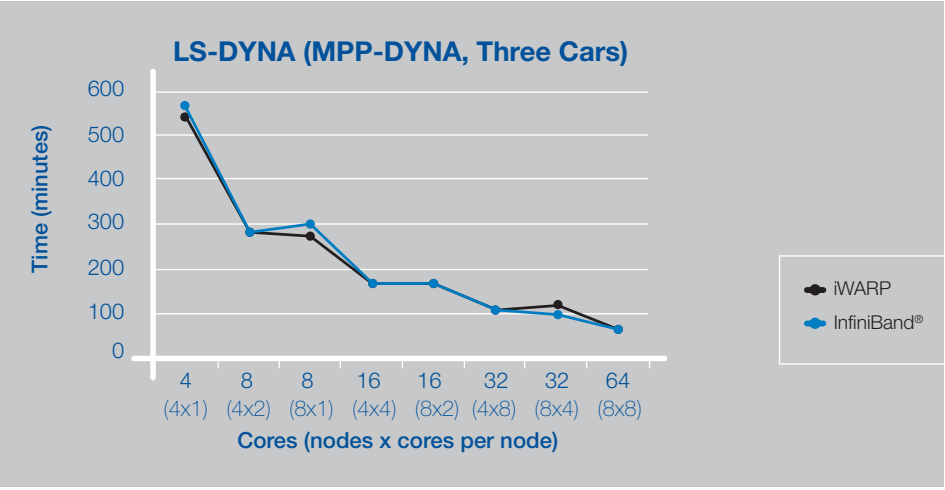

Figure 3. LS-DYNA iWARP versus InfiniBand® performance-testing results (lower y-axis figures are better).

IWARP VERSUS INFINIBAND PERFORMANCE: QUANTUM ESPRESSO

Quantum ESPRESSO (opEn Source Package for Research in Electronic Structure, Simulation, and Optimization) is an integrated suite of computer codes for electronic-structure calculations and materials modeling at the nanoscale. It is based on density-functional theory, plane waves, and pseudopotentials (both norm-conserving and ultrasoft). It is freely available under the terms of the GNU General Public License.

The package builds onto newly restructured electronic-structure codes (PWscf, PHONON, CP90, FPMD, Wannier) that have been developed and tested by some of the original authors of novel electronic-structure algorithms—from Car-Parrinello molecular dynamics to density-functional perturbation theory—and applied in the last twenty years by some of the leading materials modeling groups worldwide. The Quantum ESPRESSO distribution consists of a "historical" core set of packages and a set of plug-ins that performs more advanced tasks.

Quantum ESPRESSO is an initiative of the DEMOCRITOS National Simulation Center,[8] (Trieste) and SISSA[9] (Trieste), in collaboration with the CINECA National Supercomputing Center in Bologna,[10] the Ecole Polytechnique Fédérale de Lausanne,[11] the Université Pierre et Marie Curie,[12] Princeton University,[13] Massachusetts Institute of Technology,[14] and Oxford University.[15]

For more information, see the Quantum ESPRESSO website.[16]

TEST SCENARIO

To broadly test the performance of iWARP and InfiniBand with Quantum ESPRESSO, the following workloads were developed, results for each of which are presented later in this section:

- **Ausurf112 (DEISA benchmark)**: Self-consistent cycle of the surface of Ag made of 112 atoms, using a cut-off of 25 Ry and 4 kpoints (2x2x1).
- **Al256 (1 kpoints)**: Two steps of molecular dynamics of liquid Al made of 256 atoms using a cutoff energy of 50 Ry and the gamma point.
- **Al256 (8 kpoints)**: Identical to the previous description except using 8 kpoints (2x2x2).

TEST RESULTS

Figure 4. Quantum Espresso iWARP versus InfiniBand® performance-testing results (lower y-axis figures are better).

IWARP VERSUS INFINIBAND PERFORMANCE: VASP

VASP (Vienna Ab-initio Simulation Package) is an application maintained by the Institut fur Materialphysik – Computational Material Science at the University of Vienna. The approach implemented in VASP is based on a finite-temperature local-density approximation (with the free energy as variational quantity) and an exact evaluation of the instantaneous electronic ground state at each MD-step using efficient matrix diagonalization schemes and an efficient Pulay mixing. These techniques avoid problems occurring in the original Car-Parrinello method, which is based on the simultaneous integration of electronic and ionic equations of motion. The interaction between ions and electrons is described using ultrasoft Vanderbilt pseudopotentials (US-PP) or the projector augmented wave method (PAW). Both techniques allow a considerable reduction of the necessary number of plane-waves per atom for transition metals and first row elements. Forces and stress can be easily calculated with VASP and used to relax atoms into their instantaneous ground-state.

For more information, see the VASP website.[17]

TEST SCENARIO

To demonstrate the behaviors of iWARP and InfiniBand connectivity with variously sized problem sets, three custom workloads were developed:

- **Small**: Five molecular dynamics steps of a system of 50 atoms of Hg using only the gamma point and 10.29 Ry for the energy cutoff.
- **Medium**: Self-consistent cycle of a supercell of 32 HfO2 using a cutoff energy of 36.75 Ry and 8 kpoints (2x2x2).
- **Large**: Self-consistent cycle of a slab of 360 TiO2 using 55.12 Ry and the gamma point.

TEST RESULTS

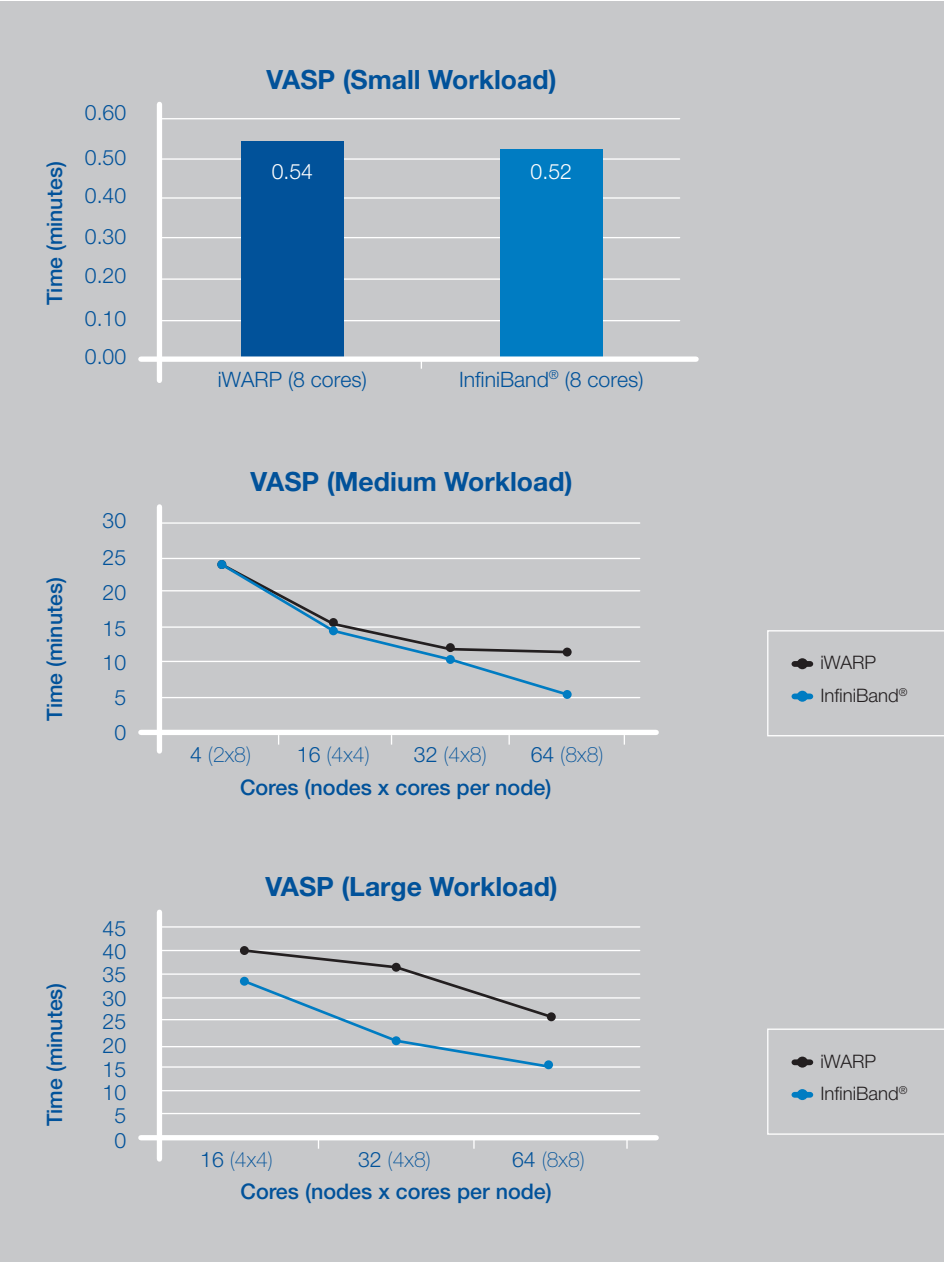

Figure 5. VASP iWARP versus InfiniBand® performance-testing results (lower y-axis figures are better).

During the large-workload VASP testing, a large portion of the system's total CPU time was spent in system calls instead of in the user-space application code. This result indicates that VASP is interacting with the iWARP driver in an unexpected, non-optimal way.

IWARP VERSUS INFINIBAND PERFORMANCE: WRF

The WRF (Weather Research and Forecasting) Modeling System development project is a collaborative partnership, principally among the National Center for Atmospheric Research (NCAR), the National Oceanic and Atmospheric

Administration (NOAA), the National Centers for Environmental Prediction (NCEP), the Forecast Systems Laboratory (FSL), the Air Force Weather Agency (AFWA), the Naval Research Laboratory, Oklahoma University, and the Federal Aviation Administration (FAA). WRF allows researchers the ability to conduct simulations reflecting either real data or idealized configurations. It is an operational forecasting model that is flexible and computationally efficient, while offering advances in physics, numerics, and data assimilation contributed by the research community.

For more information, see the WRF Model website.[18]

TEST SCENARIO

This study uses the Weather Research and Forecasting (WRF) system's Advanced Research WRF (ARW) version 2.2.1 (Skamarock et al. 2005).[19] Meteorological models are often significant contributors to errors in atmospheric transport and dispersion predictions. Wind errors can be especially large in the nocturnal stable boundary layer (SBL). Because turbulence tends to be so weak in the shallow nocturnal SBL, compared to deep convective boundary layers, these cases are much more likely to exhibit poor dispersion characteristics, thus maintaining high concentrations of airborne contaminants for many hours.

The example research production run used in this benchmark study came from research conducted at Penn State. The research continued recent DTRA-sponsored numerical research at Penn State investigating SBL predictability at very fine mesoscale resolutions.

To study the evolution of SBL flows, ARW is configured with four nested domains, each having a one-way interface with the next smaller grid. The finest domain covers ~67 x 67 km, has a horizontal resolution of 444 m, and is centered over the Nittany Valley of central Pennsylvania. This region is dominated by narrow, quasi-parallel ridges oriented southwest-to-northeast, which flank broad valleys, with the Allegheny Mountains located in the northwest part of the domain. The 1.333-km domain covers ~256 x 224 km, encompassing almost the entire Allegheny Mountain region, but it resolves the narrow ridge-and-valley topography of Central Pennsylvania with lower fidelity.

TEST RESULTS

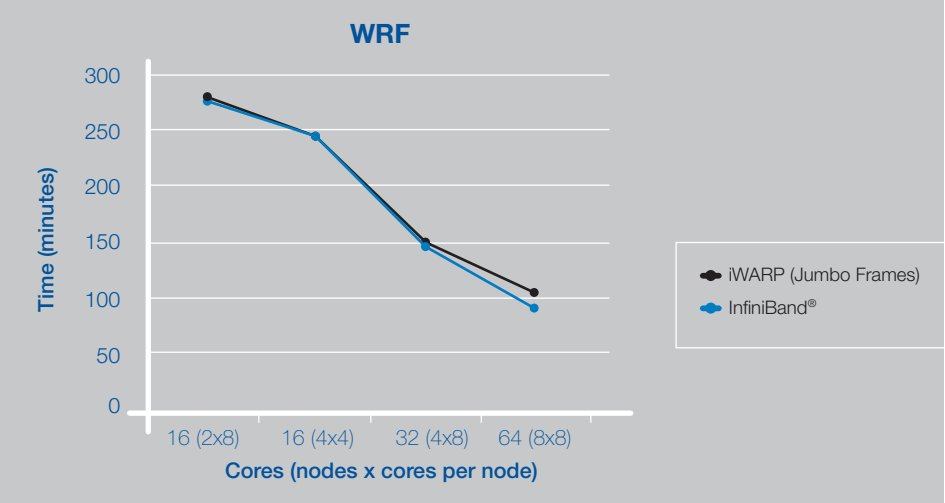

Figure 6. WRF iWARP versus InfiniBand® performance-testing results (lower y-axis figures are better).

CONCLUSION

The testing conducted to date shows that an iWARP-enabled 10GbE Ethernet network is a credible competitor to a dedicated InfiniBand fabric for the purposes of computational applications. Of the scientific codes tested, Abaqus, LAMMPS, LS-DYNA, Quantum ESPRESSO, and WRF showed that iWARP performance tracks very closely with InfiniBand performance.

Only the large VASP workload exhibits a distinct preference for the InfiniBand fabric. However, due to the interaction that was observed between VASP and the iWARP driver, it seems unlikely that this benefit is due to a fundamental hardware limitation. The issue that was observed is most likely due to the driver and can be resolved with software updates. Since iWARP and InfiniBand hardware performed similarly despite the fact that InfiniBand has over three times the useful data rate of 10GbE (32 Gb/s and 10 Gb/s respectively), a larger conclusion can be drawn about the factors constraining the performance of scientific codes. That is, these applications currently do not take advantage of the bandwidth available on today's InfiniBand fabrics, but instead benefit mostly from the low latency offered by any RDMA-enabled transport.

Given the testing conducted for this report, it is clear that the age of having separate, dedicated networks for computation-related communication is coming to a close. For reasons of cost, complexity, and energy-efficiency, networks are converging into a single data center fabric. As the dominant interconnect technology, Ethernet is uniquely positioned to be the fabric of choice in tomorrow's data centers. Because the traditional TCP/IP protocol stack still introduces too much latency to be viable for use in latency-sensitive computational applications, iWARP with its RDMA implementation is a key technology for enabling high performance communication across Ethernet networks.

CALL TO ACTION

To learn more about High Performance Computing at Penn State, see the **Research Computing and Cyberinfrastructure Group website**.[20]

To learn more about 10Gb iWARP-enabled NetEffect™ Ethernet Server Cluster Adapters from Intel, see **the product website**.[21]

The authors wish to thank Chris Bellmare at Arista Networks for providing cabling for the test equipment, as well as Julie Cummings and Gary Interdonato of Intel Corporation for furnishing network hardware, time, and expertise.

[1] http://www.simulia.com/products/abaqus_fea.html.
[2] http://www.gnu.org/copyleft/gpl.html.
[3] http://www.sandia.gov/.
[4] http://www-unix.mcs.anl.gov/mpi.
[5] http://lammps.sandia.gov/FAQ.html.
[6] http://www.ls-dyna.com/.
[7] http://www.topcrunch.org/.
[8] http://www.democritos.it/.
[9] http://www.sissa.it/.
[10] http://www.cineca.it/.
[11] http://www.epfl.ch/.
[12] http://www.impmc.upmc.fr/.
[13] http://www.princeton.edu/.
[14] http://web.mit.edu/.
[15] http://www.materials.ox.ac.uk/.
[16] http://www.quantum-espresso.org/.
[17] http://cms.mpi.univie.ac.at/vasp/vasp/vasp.html.
[18] http://www.wrf-model.org/index.php.
[19] http://wrf-model.org/wrfadmin/docs/arw_v2.pdf.
[20] http://rcc.its.psu.edu.
[21] http://www.intel.com/content/www/us/en/network-adapters/gigabit-network-adapters/neteffect-ethernet-server-cluster.html.

PENNSTATE

1855