Maximizing Cluster Scalability for LS-DYNA®

Pak Lui[1], David Cho[1], Gerald Lotto[1], Gilad Shainer[1]

[1]Mellanox Technologies, Inc.
Sunnyvale, CA, USA

1 Abstract

High performance network interconnect is an integral component that enables all compute resources to work together. It is the key to scaling the LS-DYNA simulation in a cluster environment for both network compute and network storage to accelerate CAE simulations. The latest Mellanox InfiniBand adapters have introduced a novel high-performance and scalable architecture for high-performance clusters. This architecture was enhanced to provide higher performance and scalability for the largest supercomputers in the world, today and in the future. In this paper, we demonstrate the new features and technologies that are driven by the latest InfiniBand adapter. Hardware capabilities featuring CPU offloads, MPI tag matching, MPI collective operations acceleration and message transport services make LS-DYNA perform at scale. In this study, we will review the novel architecture used in the HPC-X™ MPI library and explore some of the features in HPC-X that can maximize LS-DYNA performance by exploiting the underlying InfiniBand hardware architecture. The newly debuted Mellanox ConnectX®-5 HCA, which supports up to 100Gb/s EDR InfiniBand will be analyzed as well. For comparison purposes, we will also contrast the performance and scalability advantages of EDR InfiniBand, which is based on an CPU offload architecture, over Intel Omni-Path interconnect, which is based on an onload architecture, on the LS-DYNA simulations.

2 Introduction

High-performance computing (HPC) is a crucial tool for automotive design and manufacturing. It is used for computer-aided engineering (CAE), from component-level design to full vehicle analyses: crash simulations, structure integrity, thermal management, climate control, engine modeling, exhaust, acoustics, and much more. HPC helps drive faster time-to-market, realizing significant cost reductions over laboratory testing and tremendous flexibility. HPC's strength and efficiency depend on the ability to achieve sustained top performance by driving the CPU performance toward its limits. The motivation for high-performance computing in the automotive industry has long been its tremendous cost savings and product improvements; the cost of a high-performance compute cluster can be just a fraction of the price of a single crash test, and the same cluster can serve as the platform for every test simulation going forward.

The recent trends in cluster environments, such as multi-core CPUs, GPUs, and advanced high speed, low latency interconnect with offloading capabilities, are changing the dynamics of cluster-based simulations. Software applications are being reshaped for higher degrees of parallelism and multi-threading, and hardware is being reconfigured to solve new emerging bottlenecks to maintain high scalability and efficiency. LS-DYNA® software from Livermore Software Technology Corporation is a general purpose structural and fluid analysis simulation software package capable of simulating complex real-world problems. It is widely used in the automotive industry for the analysis of crashworthiness, occupant safety, metal stress and much more. In most cases, LS-DYNA is used in cluster environments, as they provide better flexibility, scalability, and efficiency for such simulations, allowing for larger problem sizes and speeding up time to results.

LS-DYNA relies on Message Passing Interface (MPI), the de-facto messaging library for high performance clusters that is used for node-to-node inter-process communication (IPC). MPI relies on a fast, unified server and storage interconnect to provide low latency and high messaging rate. Performance demands from the cluster interconnect increase exponentially with scale due in part to all-to-all communication patterns. This demand is even more dramatic as simulations involve greater complexity to properly simulate physical model behaviors.

In 2016, Mellanox introduced the ConnectX®-5 100Gb/s EDR InfiniBand adapter, which implements a novel architecture for high-performance clusters that was planned from the outset to provide the highest-possible performance and scalability, specifically designed for use by the largest supercomputers in the world. We will describe the effects that EDR InfiniBand have on LS-DYNA simulation performance.

3 HPC Clusters

LS-DYNA simulations are typically carried out on high-performance computing (HPC) clusters based on industry-standard computational hardware connected by a private high-speed network. The main benefits of clusters are affordability, flexibility, availability, high-performance, and scalability. A cluster uses the aggregated power of compute server nodes to form a high-performance solution for parallel applications such as LS-DYNA. When more computational power is needed, it can often be achieved by simply adding more server nodes to the cluster.

The architecture of an HPC cluster – the number of CPUs, usage of GPUs, storage solution, and cluster interconnect – has a huge influence on the overall application performance and productivity. By providing low-latency, high-bandwidth, and extremely low CPU overhead, InfiniBand has become the most deployed high-speed interconnect for HPC clusters, replacing proprietary or low-performance solutions. The InfiniBand Architecture (IBA) is an industry-standard fabric designed to provide high-bandwidth, low-latency computing, scalability for tens of thousands of nodes and multiple CPU cores per server platform, and efficient utilization of compute processing resources.

Some of the key features of the ConnectX-5 architecture that enable its cluster performance superiority are described in the following section.

4 ConnectX-5 Architecture

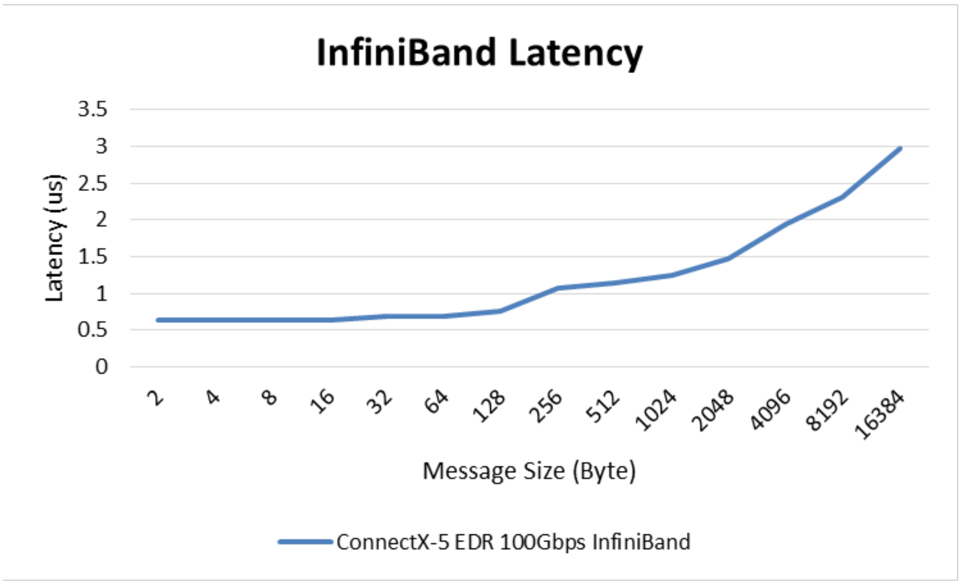

Fig.1: Point-to-point latency of 0.61us achieved on EDR InfiniBand between 2 systems

ConnectX-5 is an intelligent adapter card that introduces new acceleration engines for maximizing High Performance, Web 2.0, Cloud, Data Analytics and Storage platforms. ConnectX-5 is an InfiniBand adapter that enables 100Gb/s uni-directional throughput (~200 Gb/s bi-directional throughput) by expanding the PCI Express 3.0 bus to 16-lanes through a single 100Gb/s EDR InfiniBand network port. The latest adapter-enabled applications running on the latest compute systems can realize the full capabilities and bandwidth of the PCI Express 3.0 x16 bus. Thus, MPI and other parallel programming languages can take advantage of this high throughput, fully utilizing the multi-rail capabilities built into the software.

ConnectX-5 enables higher HPC performance with new Message Passing Interface (MPI) offloads, such as MPI Tag Matching and MPI AlltoAll operations, advanced dynamic routing, and new capabilities to perform various in-network data algorithms.

ConnectX-5 supports Virtual Protocol Interconnect®, which provides up to two ports of either 100Gb/s InfiniBand or Ethernet connectivity, sub-600 nanosecond latency, and very high message rate, plus PCIe switch and NVMe over Fabric offloads, providing the highest performance and most flexible solution for the most demanding applications and markets.

HPC applications are often based on communication patterns that use many small messages between parallel processes within the job. It is critical that the interconnect used to transport these messages provide low latency and high message rate capabilities to assure that there are no bottlenecks to the application performance. ConnectX-5 can deliver over 190 million single-packet (non-coalesced) native InfiniBand messages per second to the network. This performance assures that there are no message rate limitations for applications, allowing multiple cores on each server to communicate with other machines as fast as these cores are capable, without any slowdown from the network interface.

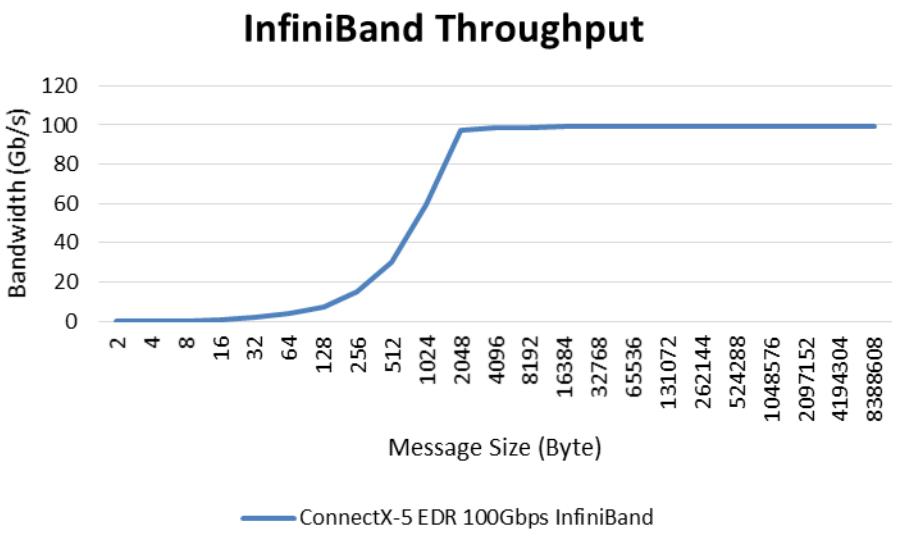

Fig.2: Unidirectional bandwidth of ConnectX-5 EDR at 100Gb/s

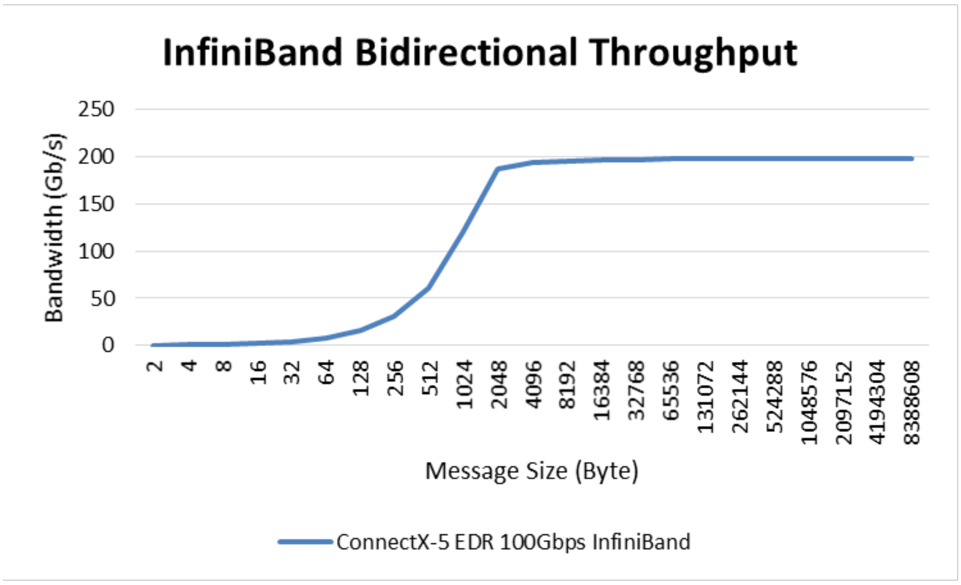

Fig.3: Bidirectional throughput performance on ConnectX-5 EDR 100Gb/s InfiniBand at ~200Gb/s

5 HPC Cluster Configuration

In this section, we will describe the HPC cluster configuration used for this study. The study was conducted at the HPC Advisory Council Cluster Center[1] on the Thor cluster.

The hardware configuration was comprised of a Dell PowerEdge™ R730 32-node cluster with 1 head node, each node with dual socket Intel Xeon® 16-core E5-2697Av4 CPUs at 2.60 GHz, Mellanox ConnectX-5 100Gb/s EDR InfiniBand adapters together with the Intel Omni-Path Host fabric Interface running at 100Gb/s, and 256GB of 2400MHz DDR4 memory and 1TB of 2.5“ 10K RPM SAS drive. The nodes were connected into two separate communication networks. One of the networks was connected using a Mellanox Switch-IB® 2 SB7800 36-port switch, which supports 100Gb/s EDR InfiniBand, while the other network was connected to an Intel Omni-Path 100 Series Edge Switch.

The software configuration was as follows. The operating system that was used was RHEL7.2. The InfiniBand driver version was MLNX_OFED 3.4-1.0.0.0; the Intel Omni-Path Host Fabric Interface driver version was 10.1.1.0.9. The input data was staged on each node to eliminate any performance influence by the network file system. MPI libraries used were IBM Platform MPI 9.1 and Mellanox HPC-X MPI Software Toolkit v1.6 (based on Open MPI v1.10) for InfiniBand, and Open MPI v1.10.2 for the PSM2 support in the Intel Omni-Path Host Fabric Interface. The LS-DYNA version was LS-DYNA MPP971_s_R8.0.0, and the benchmark workloads were the Neon Refined Revised (neon_refined_revised), the Three Vehicle Collision (3cars), and the NCAC Minivan Model (car2car) test simulations[2].

6 Impact of Network Interconnect on LS-DYNA Cluster Performance

The cluster interconnect is very critical for efficiency and performance of the application in the multi-core era. When more CPU cores are present, the overall cluster productivity increases only in the presence of a high-speed interconnect; since more data communications are generated by the increased number of available CPU cores, there must be fast network interconnect to handle the cluster-wide communications efficiently. We have compared the elapsed time with LS-DYNA using Omni-Path and Mellanox EDR InfiniBand, with both network interfaces running up to 100Gb/s.

[1] HPC Advisory Council: http://www.hpcadvisorycouncil.com

[2] The LS-DYNA benchmarks are obtainable from the TopCrunch Website: http://www.topcrunch.org

In this study, we evaluated performance differences between two different cluster networking technologies: EDR InfiniBand and Omni-Path. We begin by discussing their underlying technologies and then demonstrate their effect on performance and scalability.

Before evaluating the scalability of these two network technologies for the HPC network environment, we evaluated the underlying technologies for these implementations. Although both of these HPC networking technologies can run at a maximum data rate of 100Gb/s, EDR InfiniBand technology features CPU offloading, whereas Omni-Path features CPU onloading technology.

Onloading interconnect technology is easier to build, but this creates an issue with CPU utilization; because the CPU must also manage and execute the onloaded network operations, it has less availability for applications, which is its primary purpose. Offloading, on the other hand, seeks to overcome performance bottlenecks in the CPU by performing the network functions, as well as complex communications operations such as collective operations or data aggregation operations, on the data while it moves within the cluster network. When data is highly distributed in an application, a performance bottleneck is often created if it is necessary to wait for data to reach the CPU for analysis.

6.1 Neon Refined Revised

When comparing EDR InfiniBand with Omni-Path, EDR InfiniBand delivered superior scalability in application performance, resulting in faster runtime and providing the ability to run more jobs per day. The 100Gb/s EDR InfiniBand-based simulation (measured by number of jobs per day) delivered 25% higher performance than the Omni-Path-based simulation performance on an LS-DYNA simulation that runs on 32 nodes (1024 MPI processes). While both network interfaces nominally operate at 100Gb/s, there is clear difference in performance due to the onload versus the offload architecture. For performance beyond 4 nodes, EDR InfiniBand demonstrated even better scalability. The improvement in EDR InfiniBand is attributed to the unique capability of ConnectX-5 to utilize improved memory resource management, its more efficient transport service, as well as the CPU offload capability, which allows all CPU cores to focus on the actual computation of the simulation, such that the HPC cluster runs at its highest scalability.

Fig. 4 shows the elapsed time for the InfiniBand and Omni-Path interconnects for a range of core/node counts for the Neon Refined Revised benchmark case.

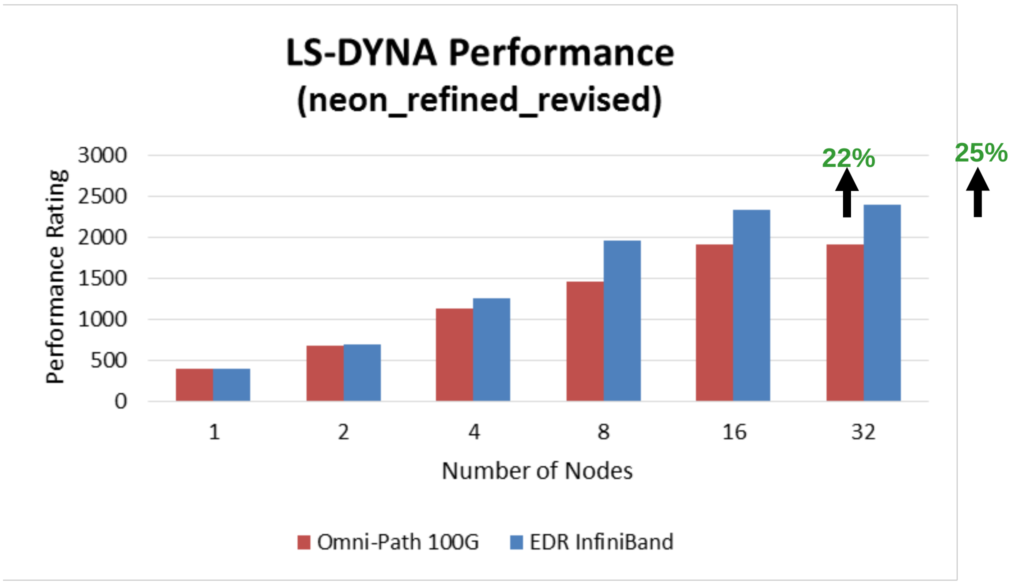

Fig.4: Network interconnect comparision on Neon Refined Revised

We next ran the Neon Refined Revised simulation under an MPI profiler. The MPI profiler allowed us to show the type of underlying MPI network communications that take place during the simulation run. It showed that the majority of the MPI communications that occurred during the simulation were collective operations, which are good candidates to be offloaded by the MPI offloading engine supported by the InfiniBand network interconnect.

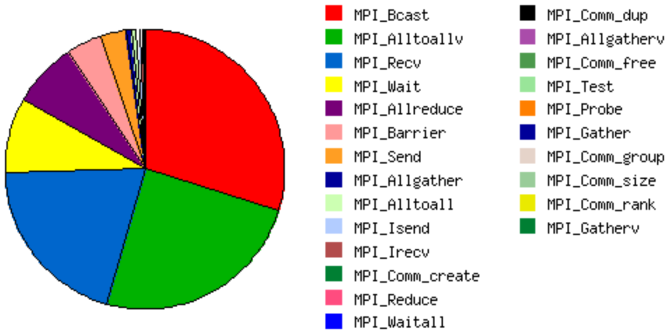

Fig.5: Network interconnect comparision on Neon Refined Revised

6.2 Three Vehicle Collision (3cars)

We also conducted a comparison between the 100Gb/s Omni-Path interconnect and EDR InfiniBand on the same set of compute nodes for an LS-DYNA simulation of the Three Vehicle Collision. As the cluster scales, the performance difference becomes even more apparent. The difference in performance occurs with as few as 8 nodes or 256 cores. EDR outperforms Omni-Path at scale by 34% for 512 cores (16 nodes, with 32 cores running per node). This performance difference becomes wider as more nodes and cores are used. At 32 nodes (1024 cores), EDR performance improves to 109% better than Intel Omni-Path, as shown in Figure 6. It is important to note that Omni-Path performance actually decreased as we increased the cluster size from 16 nodes to 32 nodes. This is clear evidence that the overhead of onload network processing actually detracts from the potential application performance in this case.

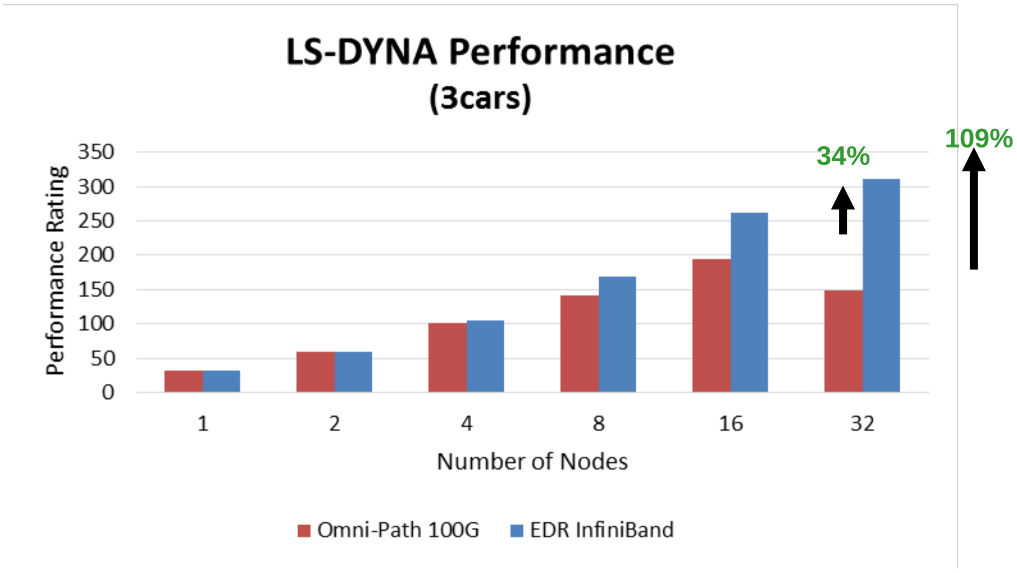

Fig.6: Scalability performance comparison between Omni-Path 100G and EDR InfiniBand

6.3 NCAC Minivan Model (car2car)

Our last comparison between the 100Gb/s Omni-Path interconnect and EDR InfiniBand was performed on the same set of compute nodes for an LS-DYNA simulation of the NCAC Minivan Model. Once again, as the cluster scales, the performance difference becomes more apparent. EDR outperforms Omni-Path at scale by 29% for 1024 cores (32 nodes, with 32 cores running per node). The difference in performance can be seen with as few as 8 nodes or 256 cores, as shown in Figure 7.

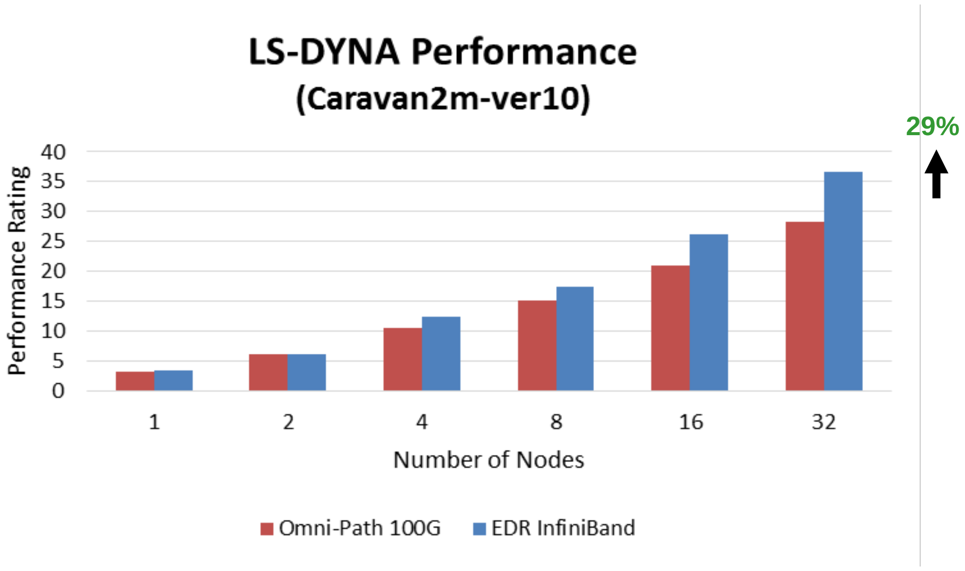

Fig.7: Scalability performance comparison between Omni-Path 100G and EDR InfiniBand

7 HPC-X Software Toolkit Overview

Mellanox HPC-X is a comprehensive software package that includes MPI, SHMEM, and UPC communications libraries. HPC-X also includes various acceleration packages to improve both the performance and scalability of applications running on top of these libraries, including MXM (Mellanox Messaging), which accelerates the underlying send/receive (or put/get) messages, and FCA (Fabric Collectives Accelerations), which accelerates the underlying collective operations used by the MPI/PGAS languages. The subsequent sections will provide an overview of MXM and FCA.

The full-featured, tested and packaged version of HPC software in HPC-X enables MPI, SHMEM, and PGAS programming languages to scale to extremely large clusters by improving on memory and latency-related efficiencies and assuring that the communication libraries are fully optimized with the Mellanox interconnect solutions. Mellanox HPC-X allows OEMs and system integrators to meet the needs of their end-users by deploying the latest available software to take advantage of the features and capabilities available in the most recent hardware and firmware changes.

7.1 Mellanox Messaging Accelerator (MXM)

The Mellanox Messaging Accelerator (MXM)[3] library provides enhancements to parallel communication libraries by fully utilizing the underlying networking infrastructure provided by the Mellanox HCA/switch hardware. This includes a variety of enhancements that take advantage of Mellanox networking hardware, including multiple transport support for RC, DC, and UD, proper management of HCA resources and memory structures, efficient memory registration, and intra-node

[3] MXM Overview: http://www.mellanox.com/products/mxm/

shared memory communication though KNEM. These enhancements significantly increase the scalability and performance of message communications in the network, alleviating bottlenecks within the parallel communication libraries.

7.2 Unified Communication X (UCX)

UCX is a communication library implementing high-performance messaging for MPI/PGAS frameworks. It is a collaboration between industry, laboratories, and academia to create an open-source production-grade communication framework for data-centric and high-performance applications

Traditionally, there have been three popular mainstream communication frameworks to support various interconnect technologies and programming languages: MXM, developed by Mellanox Technologies; PAMI, developed by IBM; and UCCS, developed by ORNL, the University of Houston, and the University of Tennessee. UCX unifies the strengths and capabilities of each of these communication libraries and optimizes them into one unified communication framework that delivers essential building blocks for the development of a high-performance communication ecosystem. "As we drive towards next generation, larger scale systems, the UCX project enables the research needed for emergent exascale programming models that are agnostic to the underlying interconnect and acceleration technology," said Dr. Arthur Bernard Maccabe, division director, Computer Science and Mathematics Division, Oak Ridge National Laboratory.

7.3 Fabric Collective Accelerator (FCA)

Collective communications execute global communication operations to couple all processes/nodes in the system and therefore must be executed as quickly and as efficiently as possible. The scalability of most scientific and engineering applications is bound by the scalability and performance of the collective routines employed. Most current implementations of collective operations will suffer from the effects of system noise at extreme-scale. (System noise increases the latency of collective operations by amplifying the effect of small, randomly-occurring OS interrupts during collective progression.) Furthermore, collective operations will consume a significant fraction of CPU cycles, which could be better spent doing meaningful computation.

Mellanox has addressed the two issues of lost CPU cycles and performance lost to the effects of system noise by offloading the communications to the host channel adapters (HCAs) and switches. This technology, called CORE-Direct® (Collectives Offload Resource Engine), provides the most advanced solution available for handling collective operations, thereby ensuring maximum scalability and minimal CPU overhead, and providing the ability to overlap communication operations with computation, allowing applications to maximize asynchronous communication.

The new FCA 3 also contains support to build runtime configurable hierarchical collectives. It currently supports socket and UMA-level discovery, with network topology slated for future versions. As with FCA 2.x, it also provides the ability to accelerate collectives with hardware multicast. In FCA 3, it also exposes the performance and scalability of Mellanox's advanced point-to-point library, MXM 2.x, in the form of the "mlnx_p2p" BCOL. This allows users to take full advantage of new features with minimal effort. FCA 3.1 and above is a standalone library that can be integrated into any MPI or PGAS runtime. Support for FCA 3.1 is currently integrated into the latest HPC-X Scalable Toolkit. The 3.1 release currently supports blocking and non-blocking variants of "MPI_Allgather", "MPI_Allreduce", "MPI_Barrier", and "MPI_Bcast".

8 Accelerating Performance by Hardware Offloads in HPC-X

HPC-X is a supported scalable HPC toolkit from Mellanox. It contains a fully supported MPI that is based on Open MPI, as well as other acceleration components that unlock the performance capabilities of the underlying interconnect hardware.

It is possible to improve LS-DYNA performance in MPI by deploying acceleration software that offloads the MPI collective communications onto the networking hardware. When placed side-by-side using the same set of systems, and additionally by using the software acceleration from the new Mellanox Fabric Collective Accelerator[4] (FCA) and Mellanox Messaging Accelerator (MXM) that are

[4] Fabric Collective Accelerator (FCA) Overview: http://www.mellanox.com/products/fca/

available in the HPC-X offering, LS-DYNA clearly demonstrates a significant improvement over the default case, which uses the CPU to process network communication.

While Open MPI and HPC-X are based on the same Open MPI distribution, HPC-X offers a few extra modules that provide additional scalability enhancement for large clusters than the baseline in the Open MPI library. The installation best practice for LS-DYNA details the steps to run the application benchmarks.[5]

HPC-X introduces a new Point-to-Point Management Layer (PML) called Yalla, which is a specialized module in the Mellanox HPC-X Software Toolkit. This unique module reduces overhead by bypassing legacy layers in message transports in Open MPI, priority accessing MXM directly. Consequently, the microbenchmark shows that for message sizes that are less than 4KB, HPC-X yields a latency improvement of up to 5%, message rate improvement of up to 50%, and bandwidth improvement of up to 45%.

8.1 Runtime Options for HPC-X

To understand the performance improvement that can be achieved by the HPC-X Software Toolkit on the MPI communication layer, we investigated and performed detailed performance studies with other popular MPI libraries, running the same LS-DYNA simulations on the same set of hardware. The other MPI libraries were also run with their tuned parameters in order to provide fair comparisons.

The UD transport and memory optimization in HPC-X reduces overhead. MXM provides a speedup of 25% over Platform MPI run at 32 nodes (1024 cores), using the neon_refined_revised benchmark case, as shown in Fig. 8. The neon_refined_revised benchmark is used to demonstrate the difference between the two implementations because the percentage of time spent on network communication is high compared to the overall runtime.

The following list details the MCA parameters and options used for benchmarking with HPC-X:

```
-x MALLOC_MMAP_MAX_=0 -x MALLOC_TRIM_THRESHOLD_=-1
-mca coll_hcoll_enable 0
-x MXM_SHM_RNDV_THRESH=32768 -x KMP_BLOCKTIME=0
-mca rmaps_base_mapping_policy slot
```

KNEM is implicitly enabled by HPC-X, and is configured by default to run with the following command line option implicitly:

```
-mca btl_sm_use_knem 1 -x MXM_SHM_KCOPY_MODE=knem
```

In addition to the aforementioned flags for Open MPI that are used in HPC-X, the following list details the MCA parameters for enabling MXM in HPC-X:

```
-mca pml yalla -x MXM_TLS=ud,shm,self -x MXM_SHM_RNDV_THRESH=32768
```

8.2 MPI Libraries Comparison: Neon Refined Revised

By comparing the HPC-X performance to other MPI libraries, we can see how HPC-X outperforms other MPI libraries in scalability by exploiting hardware offload capabilities, which other MPI libraries do not.

[5] LS-DYNA Installation Best Practices: http://www.hpcadvisorycouncil.com/pdf/LS-DYNA_Installation_Best_Practices-HPC-X.pdf

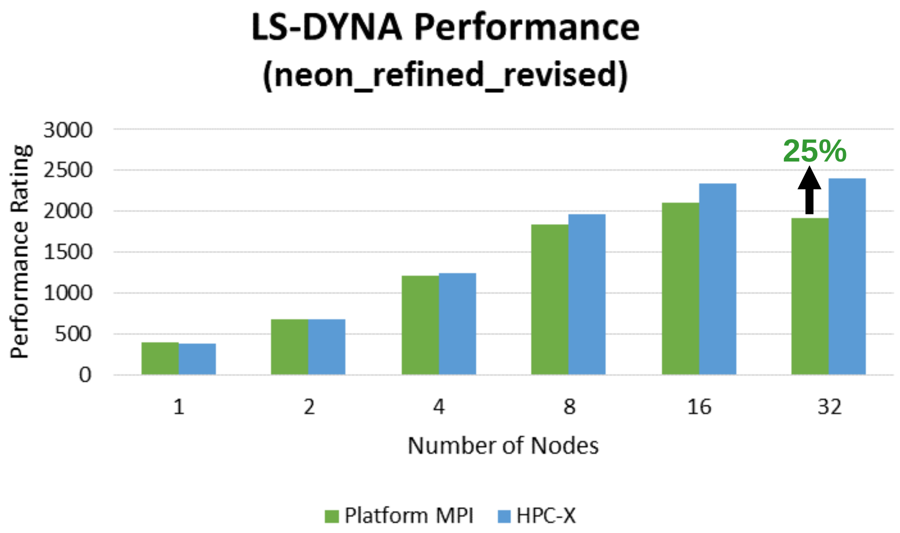

Fig.8: Comparision of HPC-X versus Platform MPI with the Neon Refined Revised case

We found that HPC-X outperforms Platform MPI in scalability performance using the neon_refined_revised benchmark. We compared the results gathered for HPC-X with an additional set of data points for Platform MPI that has been shown to perform well on InfiniBand hardware.

We compared the runtimes achieved on the same workload for Platform MPI and HPC-X, finding that HPC-X outperforms Platform MPI by 25%. The following list details the tuning parameters for running Platform MPI:

```
-IBV -cpu_bind, -xrc
```

HPC-X delivers higher scalability performance than Platform MPI, even with parameters tuned specifically for Platform MPI. The performance improvement gap should increase as the cluster scales further, as the effects of FCA and MXM allow for even greater scalability of the application.

9 Conclusions

HPC cluster environments impose high demands for connectivity throughput and low latency with low CPU overhead, network flexibility, and high-efficiency in order to maintain a balanced system that can achieve high application performance and scaling. Low-performance interconnect solutions, or those that lack interconnect hardware offload capabilities, will result in reduced application performance and extended time-to-market. Livermore Software Technology Corporation (LSTC) LS-DYNA software was benchmarked for this study. In all cases, LS-DYNA demonstrated higher parallelism and scalability with InfiniBand, which enabled it to take full advantage of multi-core HPC clusters.

We reviewed the novel architecture used by the HPC-X MPI library to leverage the unique advantages offered by InfiniBand, and we explored some of the features in HPC-X that maximize LS-DYNA performance by exploiting the underlying InfiniBand hardware architecture, which also uses HPC-X to outperform Platform MPI.

The performance levels of various adapter throughputs on different network architectures were measured to document their effect on LS-DYNA software. The evidence showed the inherent advantages offered by the ConnectX-5 100Gb/s EDR InfiniBand adapter, including: the unparalleled higher message rate, superior bandwidth, lower latency, and greater scalability over the Omni-Path 100G network interconnect. The overall result is that the Mellanox ConnectX-5 HCA has decreased LS-DYNA's runtime, enabling LS-DYNA to run significantly more jobs per day at scale. With the additional improvements in the system architecture and networking technologies, such as MPI collective offloads, EDR InfiniBand enables LS-DYNA to achieve superior performance at scale.