Performance Optimizations for LS-DYNA with Mellanox HPC-X™ Scalable Software Toolkit

Pak Lui[1], David Cho[1], Gilad Shainer[1], Scot Schultz[1], Brian Klaff[1]

[1]Mellanox Technologies, Inc.

1 Abstract

From concept to engineering, and from design to test and manufacturing, the automotive industry relies on powerful virtual development solutions. CFD and crash simulations are performed in an effort to secure quality and accelerate the development process. Modern-day engineering simulations are becoming more complex and higher in accuracy in order to model closely to real world scenarios. To accomplish such design simulations virtually on a cluster of computer systems, LS-DYNA® decomposes large simulations into smaller problem domains. By distributing the workload and compute requirements across powerful HPC compute nodes that are connected via a high-speed InfiniBand network, the time required to solve such problems is reduced dramatically. To orchestrate such a complex level of communication between compute systems, the solvers of LS-DYNA were implemented with an interface to the Message Passing Interface (MPI) library (the de-facto messaging library for high performance clusters). MPI covers the communications between tasks within an HPC compute cluster. The recently introduced Mellanox HPC-X™ Toolkit is a comprehensive MPI, SHMEM, and UPC software suite for high performance computing environments. HPC-X also incorporates the ability to offload network collectives communication from the MPI processes onto the Mellanox interconnect hardware. 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®-4 HCA, which runs at 100Gb/s EDR InfiniBand and is supported by HPC-X, will be analyzed as well.

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 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, significant cost reductions, and tremendous flexibility. The strength in HPC is the ability to achieve sustained top performance by driving the CPU performance towards 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 interconnect speeds and offloading capabilities, are changing the dynamics of clustered-based simulations. Software applications are being reshaped for higher parallelism and multi-threading, and hardware is being configured to solve new emerging bottlenecks, in order 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 crashworthiness analysis, occupant safety analysis, metal forming, and much more. In most cases, LS-DYNA is used in cluster environments, as they provide better flexibility, scalability, and efficiency for such simulations.

LS-DYNA relies on Message Passing Interface (MPI) the de-facto messaging library for high performance clusters, for cluster or node-to-node communication,. MPI relies on fast server and storage interconnect to provide low latency and high messaging rate. Performance demands from the cluster interconnect increase dramatically as the simulation requires more complexity to properly simulate the physical model behavior.

In 2015, Mellanox introduced the ConnectX®-4 100Gb/s EDR InfiniBand adapter, which has a novel high-performance and scalable architecture for high-performance clusters. The architecture was planned from the outset to provide the highest-possible performance and scalability, specifically designed for use by the largest supercomputers in the world.

3 HPC Clusters

LS-DYNA simulations are typically carried out on high-performance computing (HPC) clusters based on industry-standard 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 compute power is needed, it can sometimes be achieved simply by adding more server nodes to the cluster.

The manner in which HPC clusters are architected – number of CPUs, usage of GPUs, the storage solution, and the 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-4 architecture that enable its cluster performance superiority are described in the following section.

4 ConnectX-4 Architecture

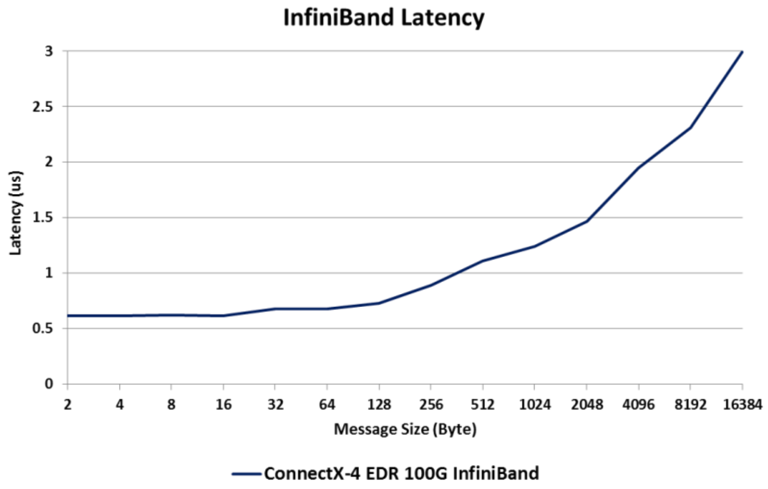

Fig.1: Point-to-point latency of 0.61us achieved on EDR InfiniBand between 2 systems

ConnectX-4 is the first InfiniBand adapter on the market that enables 100Gb/s uni-directional throughput (~195 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 Intel Haswell systems to realize the full capabilities and bandwidth of the PCI Express 3.0 x16 bus. This represents a technology improvement from the previous generation of InfiniBand adapter that runs at the FDR 56Gb/s rate on a single InfiniBand network port.Thus, MPI and other parallel programming languages can take advantage of this high throughput, utilizing the multi-rail capabilities built into the software.

This increase is important for bandwidth-sensitive applications, and even more critical with the advent of increased CPU cores such as the new Intel Haswell system. In addition, this level of throughput will be required to satisfy the needs of new heterogeneous environments such as GPGPU and Intel Xeon Phi-based endpoints.

Many HPC applications are 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. ConnectX-4 can deliver over 150 million single-packet (non-coalesced) native InfiniBand messages per second to the network. This increase assures that there are no message rate limitations for applications and that the multiple cores on the server will communicate to other machines as fast as the cores are capable, without any slowdown from the network interface.

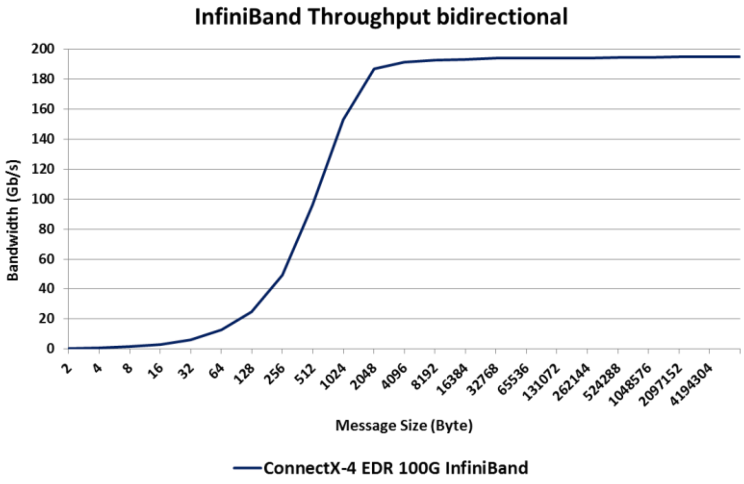

Fig.2: Bidirectional throughput performance on ConnectX-4 EDR 100Gb/s InfiniBand at ~195Gb/s

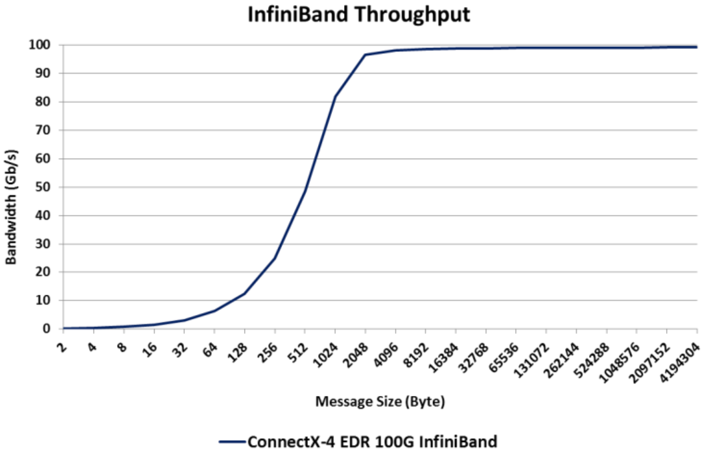

Fig.3: Unidirectional bandwidth of ConnectX-4 EDR at 100Gb/s

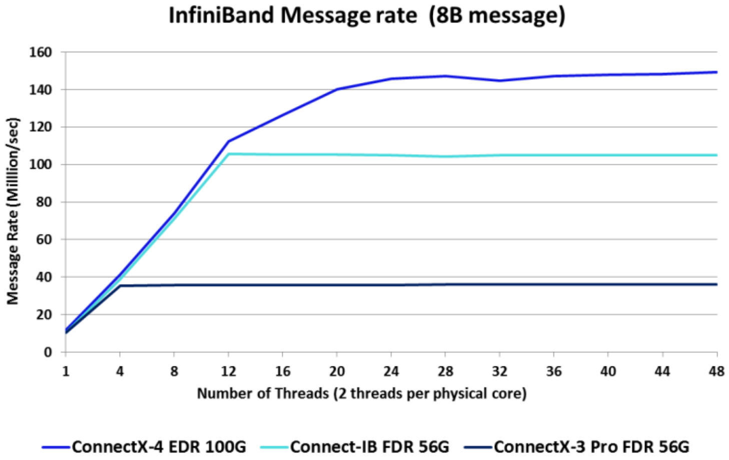

Fig.4: Message rates comparision of recent generations of InfiniBand adapters

5 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 by 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 1Gb/s Ethernet, 10Gb/s Ethernet, 40Gb/s Ethernet, 56Gb/s FDR InfiniBand, and 100Gb/s EDR InfiniBand.

This study was conducted at the HPC Advisory Council Cluster Center[1] comprised of a Dell PowerEdge™ R730 32-node cluster with 1 head node, each node with dual socket Intel Xeon® 14-core E5-2697v3 CPUs at 2.60 GHz, Mellanox ConnectX-4 100Gb/s EDR InfiniBand adapters, as well as Mellanox ConnectX-3 56Gb/s FDR InfiniBand adapters, and 64GB of 2133MHz DDR4 memory. The nodes were connected into two separate communication networks. One of the networks was connected using a Mellanox Switch-IB® SB7700 36-port switch, which supports 100Gb/s EDR InfiniBand, while the other network was connected to a Mellanox SwitchX® SX6036 36-port VPI switch, which supports 40Gb/s Ethernet and 56Gb/s FDR InfiniBand. The Operating System that was used was RHEL6.5, the InfiniBand driver version was MLNX_OFED_LINUX-2.4-1.0.5.1_20150408_1555, and the File System was shared over NFS from the Dell PowerEdge R730 head node, which provides eight 1TB 7.2K RPM SATA 2.5" hard drives over RAID 5. The MPI library that was used was IBM Platform MPI 9.1, 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].

5.1 InfiniBand Advantages over Ethernet

Fig. 5 below shows the elapsed time for the InfiniBand and Ethernet interconnects for a range of core/node counts for the Neon Refined Revised benchmark case.

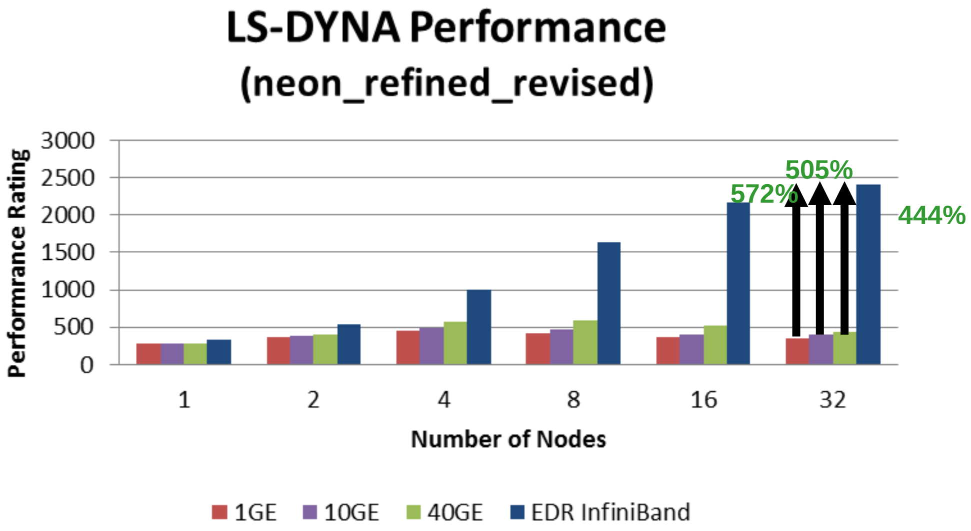

Fig.5: Network interconnect comparision on Neon Refined Revised

When comparing EDR InfiniBand and various Ethernet bandwidths, InfiniBand delivered superior scalability in application performance, resulting in faster run time, providing the ability to run more jobs per day. The 100Gb/s EDR InfiniBand-based simulation performance (measured by number of jobs per day) was 572% higher than 1GbE, over 505% higher than 10GbE, and over 444% higher than 40GbE on an LS-DYNA simulation that runs on 32 nodes

[1] HPC Advisory Council: http://www.hpcadvisorycouncil.com
[2] The LS-DYNA benchmarks are obtainable from the TopCrunch Website: http://www.topcrunch.org

While 1GbE Ethernet showed a loss of performance beyond 4 nodes, EDR InfiniBand demonstrated good scalability throughout the various tested configurations. Because LS-DYNA uses MPI for the interface between the application and the networking layer, it requires scalable and efficient send-receive semantics, as well as good scalable collective operations. While InfiniBand provides an effective method for such operations, the Ethernet TCP stack causes CPU overhead, which translate into higher network latency, reducing the cluster efficiency and scalability.

5.2 EDR InfiniBand Advantage on LS-DYNA Scalability Performance

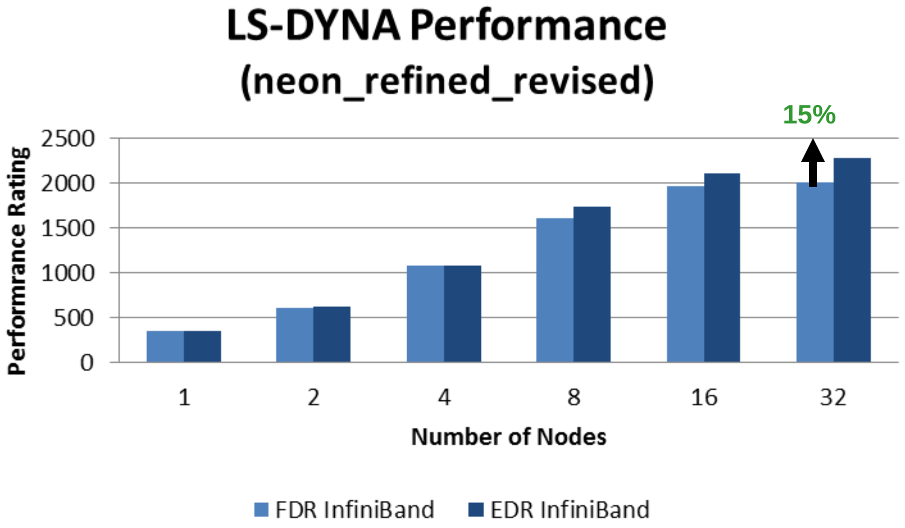

Fig.6: Scalability performance comparision of EDR InfiniBand and FDR InfiniBand

When comparing EDR and FDR InfiniBand, EDR InfiniBand delivered superior scalability in application performance, resulting in faster run time and providing the ability to run more jobs per day. The 100Gb/s EDR InfiniBand-based simulation performance (measured by number of jobs per day) was 15% higher performance than 56Gb/s FDR InfiniBand-based simulation performance on an LS-DYNA simulation that runs on 32 nodes (896 MPI processes). While both are based on InfiniBand technologies, the performance beyond 4 nodes for EDR InfiniBand demonstrated even better scalability. The improvement in EDR InfiniBand is due to the unique capability of ConnectX-4 utilizing the improved memory resource management and more efficient transport service, allowing the HPC cluster to run at its highest scalability.

6 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.

6.1 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 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.

6.2 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, cycles that 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, named CORE-Direct® (Collectives Offload Resource Engine), provides the most advanced solution available for handling collective operations, thereby ensuring maximal scalability, minimal CPU overhead, and providing the ability to overlap communication operations with computation, allowing applications to maximize asynchronous communication.

FCA 3.1 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 3.1, 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".

7 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 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.

HPC-X introduces a new Point-to-Point Management Layer (PML) that is called Yalla, which is a specialized module in the Mellanox's HPC-X Software Toolkit. This unique module reduces overhead by bypassing legacy layers in message transports in Open MPI, priority accessing to MXM directly. Consequently, the microbenchmark shows that for message sizes that are less than 4KB, it yields a

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

latency improvement of up to 5%, message rate improvement of up to 50%, and bandwidth improvement of up to 45%.

7.1 Comparison of HPC-X over Open MPI

The UD transport and memory optimization in HPC-X reduce overhead. MXM provides a speedup of 38% over the untuned Open MPI baseline run at 32 nodes (768 cores), using the neon_refined_revised benchmark case.

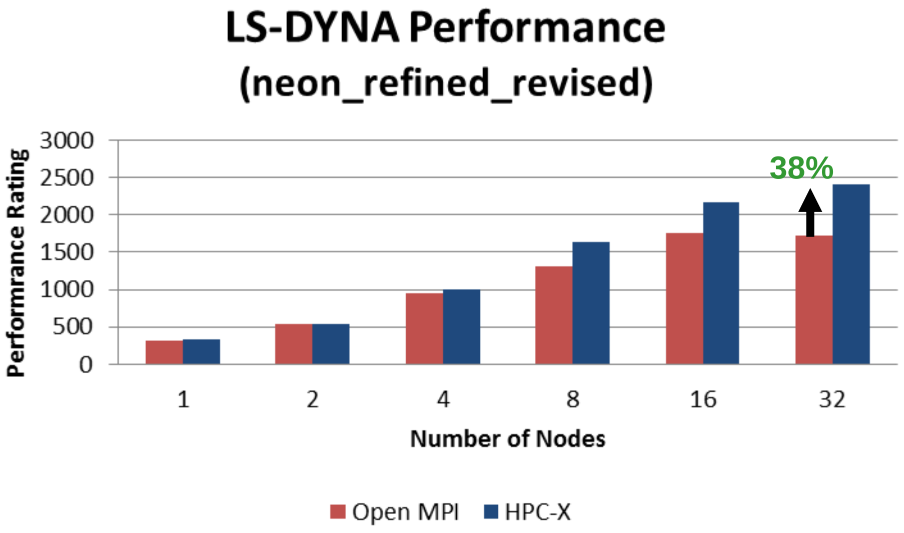

Fig.7: Performance inprovement of HPC-X over Open MPI using the neon_refined_revised case

The following list details the MCA parameters used for running the benchmark with Open MPI:

```
-bind-to core, -mca btl_sm_use_knem 1, -mca btl openib,self,sm -x
MALLOC_MMAP_MAX_=0 -x MALLOC_TRIM_THRESHOLD_=-1
```

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 -x
MXM_RDMA_PORTS=mlx5_0:1
```

The following list details the MCA parameters for enabling hierarchical collective support for HPC-X:

```
-mca coll_hcoll_enable 1 -x coll_hcoll_np=0 -mca coll_fca_enable 0 -x
HCOLL_IB_IF_INCLUDE=mlx5_0:1 -x MXM_RDMA_PORTS=mlx5_0:1 -x
MXM_LOG_LEVEL=FATAL -x HCOLL_ENABLE_MCAST_ALL=1
```

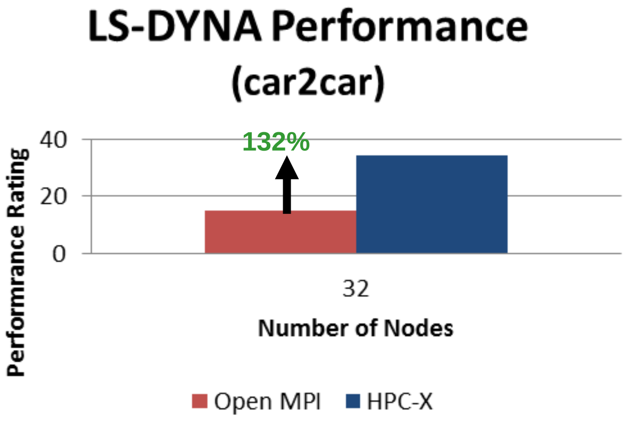

Fig.8: Performance improvement of HPC-X over Open MPI using the car2car case

Similarly, the performance improvement of HPC-X is also dramatically improved on the more compute-intensive benchmark case of car2car. The performance jumped by 132% on a 32-node (896 cores) run by deploying a tuned message library inside HPC-X.

8 Comparisons of HPC-X to Other MPI Libraries

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

8.1 MPI Libraries Comparison: Neon Refined Revised

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

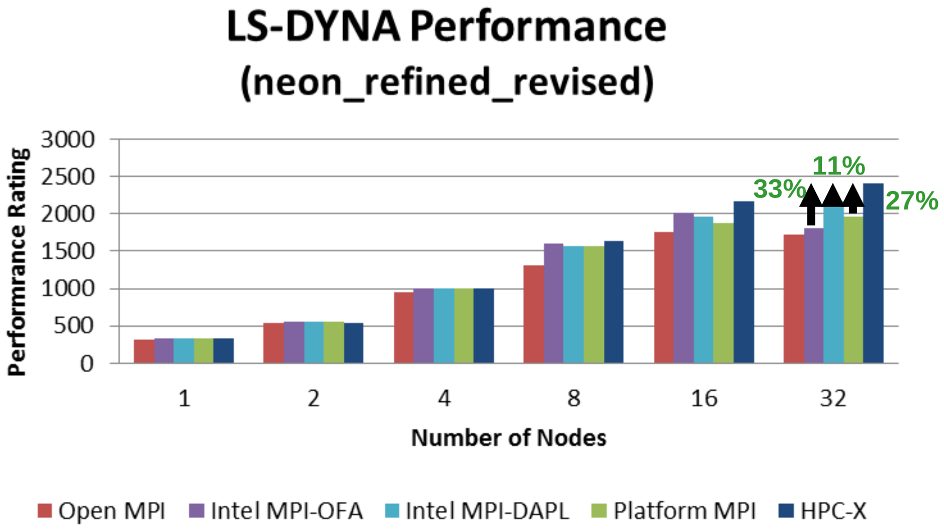

Fig.9: Comparision of HPC-X versus other MPI Libraries with the Neon Refined Revised case

We found that HPC-X outperforms Platform MPI and Open MPI in scalability performance using the neon_refined_revised benchmark. We overlaid the results gathered for Fig. 7 onto Fig. 9 with the 3 additional sets of data points for Platform MPI, and the Intel MPI using the OFA and DAPL providers for InfiniBand support.

We first compared the runtimes achieved on the same workload for Platform MPI and HPC-X, and we found that HPC-X outperforms Platform MPI by 27%. The following list details the tuning parameters for running Platform MPI:
`-IBV -cpu_bind, -xrc`

Similarly, we compared the runtimes achieved on the same workload by using Intel MPI and HPC-X. There are 2 different MPI providers available for Intel MPI. By default, Intel MPI automatically selected the DAPL provider for the communications used for InfiniBand. The following list details the tuning parameters for running Intel MPI on both the OFA and DAPL providers:
`I_MPI_ADJUST_REDUCE 2, I_MPI_ADJUST_BCAST 0, I_MPI_DAPL_SCALABLE_PROGRESS 1, I_MPI_RDMA_TRANSLATION_CACHE 1, I_MPI_FAIR_CONN_SPIN_COUNT 2147483647, I_MPI_FAIR_READ_SPIN_COUNT 2147483647, I_MPI_RDMA_TRANSLATION_CACHE 1, I_MPI_RDMA_RNDV_BUF_ALIGN 65536, I_MPI_SPIN_COUNT 121`

The following list details the tuning parameters specifically for running Intel MPI with the OFA provider:
`-IB, MV2_USE_APM 0, I_MPI_OFA_USE_XRC 1`

The following list details the tuning parameters specifically for running Intel MPI with the DAPL provider:

```
-DAPL,   I_MPI_DAPL_DIRECT_COPY_THRESHOLD   65536,   I_MPI_DAPL_UD   enable,
I_MPI_DAPL_PROVIDER ofa-v2-mlx5_0-1u
```

HPC-X delivers higher scalability performance than Intel MPI, even though the tuned parameters are used to allow the DAPL and OFA providers to perform better. HPC-X outperforms the Intel MPI with the OFA provider by 33%, and the Intel MPI with the DAPL provider by 11% at 32 nodes, as shown in the Fig. 9. The gap for additional performance improvement for HPC-X is expected to widen as the cluster scales, as the effects of FCA and MXM allow for even greater scalability of the application.

8.2 MPI Libraries Comparison: 3 Vehicle Collision

For the following comparison, we switched to a more compute-intensive input data. The 3 Vehicle Collision benchmark shows the performance exhibiting similar behavior. The performance measured for HPC-X is about 20% higher than for Intel MPI using the OFA provider with the parameters mentioned for InfiniBand support in the aforementioned section. Similarly, HPC-X also outperforms Platform MPI by 8% on the 3cars input data at 32 nodes (768 CPU cores).

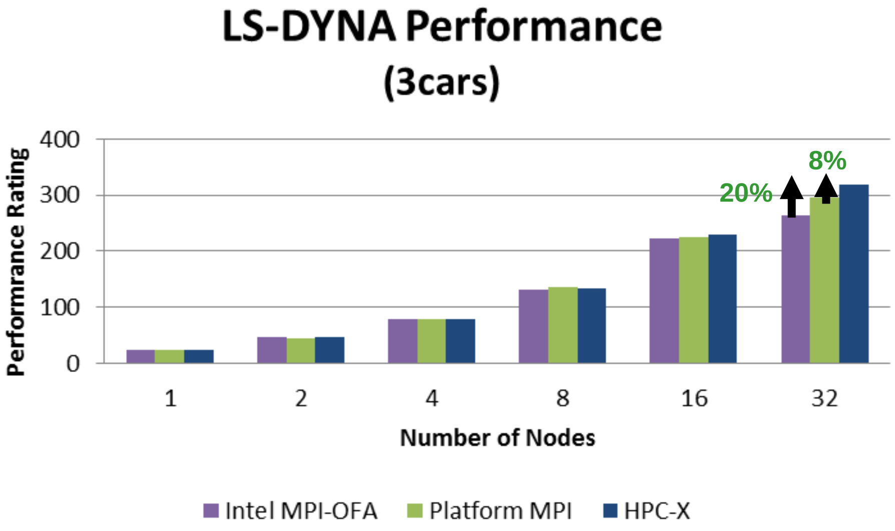

Fig.10: Comparision of HPC-X versus other MPI Libraries with the 3 Vehicle Collision case

9 Performance Improvements by System Architecture

Because ConnectX-4 supports the PCIe Gen3 standard, it is perfectly suited to run on system architecture that supports the PCIe Gen3 x16 slots, which includes the most current Intel Xeon E5-2600 v3 series (Haswell), the Intel Xeon E5-2600 v2 Series (Ivy Bridge), and the E5-2600 Series (Sandy Bridge) based platforms.

The ConnectX-4 HCA allows the MPI applications to take advantage of the increases in CPU core processing and memory bandwidth by providing the necessary network throughput. Notably, the Haswell architecture enables the ConnectX-4 InfiniBand device to run at its maximum throughput and lowest latency. The results are shown in Fig. 11.

Compared to previous system generations, the Intel Xeon E5-2697 v3 (Haswell) cluster outperforms the Intel Xeon E5-2680 v2 (Ivy Bridge) cluster by up to 47%; the Haswell cluster outperforms the Intel Xeon E5-2680 (Sandy Bridge) cluster by up to 75%, it also outperforms Intel Xeon X5670 (Westmere) cluster by up to 148%, and provided performance gains up to 290% over the older Intel Xeon X5570 (Nehalem) cluster.

To conduct the performance comparison tests, the following system configurations were used:

- Each Haswell system consisted of a Dell PowerEdge R730, each with a dual-socket Intel Xeon E5-2697v3 running at 2.6GHz, 2133MHz DIMMs, and Mellanox ConnectX-4 EDR InfiniBand.

- Each Ivy Bridge system consisted of a Dell PowerEdge R720xd, each with a dual-socket Intel Xeon E5-2680 v2 running at 2.8GHz, 1600MHz DIMMs, and Mellanox Connect-IB FDR InfiniBand.
- Each Sandy Bridge system used the aforementioned Dell PowerEdge R720xd, each with a dual-socket Intel Xeon E5-2680 running at 2.7GHz, 1600MHz DIMMs, and Mellanox Connect-IB FDR InfiniBand.
- Each Westmere system consisted of a Dell PowerEdge m610 system, with a dual-socket Intel Xeon X5670 running at 2.93GHz, 1333MHz DIMMs, and Mellanox ConnectX-2 QDR InfiniBand.
- Each Nehalem system used the aforementioned Dell PowerEdge m610 system with a dual-socket Intel Xeon X5570 running at 2.93GHz, 1333MHz DIMMs, and Mellanox ConnectX-2 QDR InfiniBand.

Contributing factors to the improved Haswell system performance over prior system generations are the better scalability support in the EDR InfiniBand in the hardware and the Mellanox HPC-X provided in the MPI library layer.

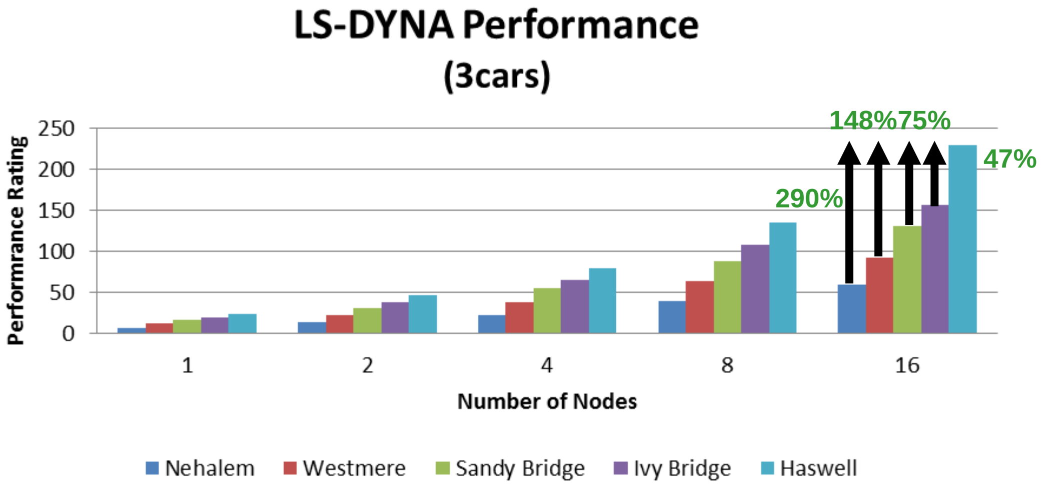

Fig.11: Comparisons of performance on recent system generations of Intel system architecture

10 Conclusions

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

We reviewed the novel architecture used in the HPC-X MPI library and explored some of the features in HPC-X, which can maximize LS-DYNA performance by exploiting the underlying InfiniBand hardware architecture, which also uses HPC-X to outperform other MPI libraries.

We also compared the performance levels of various adapter throughputs on different network architectures to measure the effect on LS-DYNA software. The evidence showed that the inherent advantages offered by the ConnectX-4 100Gb/s EDR InfiniBand adapter – namely, the unparalleled message rate and support for the PCI Gen3 standard – produce increased bandwidth, lower latency, and greater scalability than when using 40GbE, 10GbE, or 1GbE interconnects. The recently debuted ConnectX-4 HCA has decreased LS-DYNA's run time, enabling LS-DYNA to run significantly more jobs per day than ever before. With the additional improvements in the system architecture and networking technologies, such as MPI collective offloads, it enables LS-DYNA to achieve superior performance at scale.