The Effect of InfiniBand and In-Network Computing on LS-DYNA® Simulations

Ophir Maor, Gilad Shainer, David Cho, Gerardo Cisneros-Stoianowski, Yong Qin
HPC-AI Advisory Council

Abstract

High-performance computing (HPC) technologies are used in the automotive design and manufacturing industry. One of the applications is computer-aided engineering (CAE), from component-level design to full analyses for a variety of use cases, including: crash simulations, structure integrity, thermal management, climate control, modeling, acoustics, and more. HPC helps drive faster time-to-market, significantly reducing the cost of laboratory testing and enabling 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 has long been associated with 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, for example, and the same cluster can serve as the platform for every test simulation going forward.

Recent trends in cluster environments, such as multi-core CPUs, GPUs, and advanced high-speed and low latency interconnect with In-Network Computing capabilities, are changing the dynamics of cluster-based simulations. Software applications are being reshaped for higher degrees of parallelism and multithreading, and hardware is being reconfigured to solve new emerging bottlenecks to maintain high scalability and efficiency. Applications like LS-DYNA and others are widely used and provide better flexibility, scalability, and efficiency for such simulations, allowing for larger problem sizes and speeding up time-to-results.

CAE applications rely 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 a 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 this paper, we will focus on the value of HDR InfiniBand interconnect technology for LS-DYNA applications, by comparing different InfiniBand network transport options and MPI libraries.

InfiniBand In-Network Computing Technology

The latest revolution in HPC is the effort around "Co-design," a collaboration to reach Exascale performance by taking a holistic system-level approach to fundamental performance improvements, otherwise referred to as In-Network Computing. The CPU-centric approach has reached the limits of its scalability in several aspects. In-Network Computing, acting as a "distributed co-processor," can handle and accelerate the performance of various data algorithms, such as reductions and more.

The past focus of smart interconnect development was to offload the network functions from the CPU to the network. With the new efforts of the co-design approach, the new generation of smart interconnects can also offload data algorithms being managed within the network, allowing users to run these algorithms as the data is transferred within the system interconnect, rather than waiting for the data to reach the CPU. This technology is referred to as In-Network Computing and is the leading approach to achieving performance and scalability for Exascale systems. In-Network Computing transforms the data center interconnect into a "distributed CPU" with "distributed memory," enabling the overcoming of performance walls and faster and more scalable data analysis.

Scalable Hierarchical Aggregation and Reduction Protocol (SHARP)

Mellanox's Scalable Hierarchical Aggregation and Reduction Protocol (SHARP)™ is a technology that enables data reduction and aggregation operations on the interconnect components. Mellanox SHARP technology has been implemented in the latest generations of InfiniBand solutions, running EDR 100Gb/s and HDR 200Gb/s speeds. With increases in the amount of data requiring analysis and in simulation complexity, the traditional approach to analyzing data based solely on the compute elements has reached a performance wall. Adding more cores to handle the various data reduction and aggregation operations does not result in any performance improvement. Mellanox SHARP technology helps overcome this performance wall by migrating these operations to the network and performing them while the data is being transferred (Figure 1).

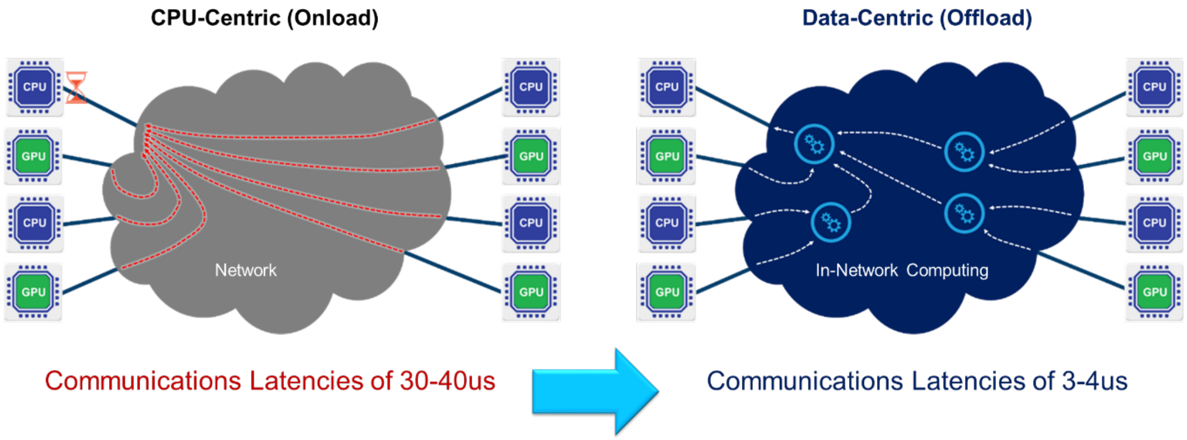

Figure 1: Illustration of Mellanox SHARP Technology

The goal of In-Network Computing architecture is to optimize the completion time of frequently used global communication patterns, and to minimize their impact on CPU utilization. The set of patterns being targeted are global reductions of data, including barrier synchronization and data reductions. The Mellanox SHARP protocol provides an abstraction that describes data reduction. The protocol defines aggregation nodes (ANs) in an aggregation tree, which are basic components of in-network reduction operation offloading. In this abstraction, data enters the aggregation tree from its leaf nodes, and makes its way up the tree with data reductions occurring at each AN, and the global aggregate ends up at the root of the tree.

The method of distributing the aggregation result can be independent of the aggregation pattern. Much of the communication processing of these operations is moved to the network, providing host-independent progress and minimizing application exposure to the negative effects of system noise. The implementation manipulates data as it traverses the network, minimizing data motion. The design benefits from the high degree of network-level parallelism, with the high-radix InfiniBand switches enabling the use of shallow reduction trees.

Other In-Network Computing elements include hardware-based MPI tag matching, MPI rendezvous offloads, and more.

HDR InfiniBand

HDR InfiniBand is the latest InfiniBand generation in the market today. The HDR InfiniBand specification includes two network speeds – 200Gb/s (HDR) and 100Gb/s (HDR100). Beyond their faster data speeds, the HDR InfiniBand products include a switch radix of 40 ports of 200Gb/s, or 80 ports of 100Gb/s. The higher switch radix provides lower latency between neighboring processes, and a lower total cost of ownership. The HDR InfiniBand technology also includes the second generation of Mellanox SHARP, which enhances InfiniBand's acceleration capabilities for deep learning applications as well as for HPC workloads.

InfiniBand Transports (RC, UC and DC)

InfiniBand supports multiple transport services that can be used by applications, including Reliable Connected (RC), Unreliable Connected (UC), Unreliable Datagram (UD), Reliable Datagram (RD), Dynamically Connected (DC), and Raw Datagram (RAW). The main options used today are RC, UD and DC. RC and UC were defined with the creation of the InfiniBand specification, and DC was added a few years later.

Reliable Connected allows two Queue Pair (QP) connections to exchange data in a reliable way. RC guarantees that messages are delivered from a requester to a responder no more than a single time, in order and without corruption. RC also supports RDMA and atomic operations, but it has scalability limitations.

Unreliable Datagram allows one QP to send and receive messages to/from any other UD QP either as unicast (one to one) or multicast (one to many), though in an unreliable way; that is, there is no guarantee that the messages will be received by the other side. Moreover, corrupted packets are silently dropped, which means that the application is responsible for the data reliability (for example MPI). In addition, UD does not support RDMA or atomic operations.

Dynamic Connected transport combines the advantages of RC and UD transport services, achieving the highest scalability and hardware-based data reliability. DC supports RDMA and atomic operations, and it is reliable and scalable as one QP can support multiple destinations. DC addresses several shortcomings of the older RC and UD transport protocols. DC aims to support all of the features provided by RC, such as RDMA, atomics, and hardware reliability, while allowing processes to communicate with any remote process with just one DC QP, similar to UD that scale better. DC provides the best of the RC and UD worlds – small memory footprint, scalability and reliability.

Transport Performance Evaluation

The following testing measures the performance of RC, UD and DC for two LS-DYNA benchmarks, and explores whether DC, which combines the benefits of RC and UD transport services, delivers performance on par with RC and UD.

The testbed setup is as follows:

- OS: CentOS 7.7
- Driver: MLNX_OFED 4.7
- CPU: Intel E5-2697 v4 @2.6GHz, dual socket 16 cores per socket (dual socket)
- Network: InfiniBand HDR100
- LS-DYNA Version: ls-dyna_mpp_s_R11_1_0_x64_centos65_ifort160_avx2_intelmpi-2018
- Input: 3cars_rev02
- IO: RAMFS
- MPI: HPC-X 2.6.0

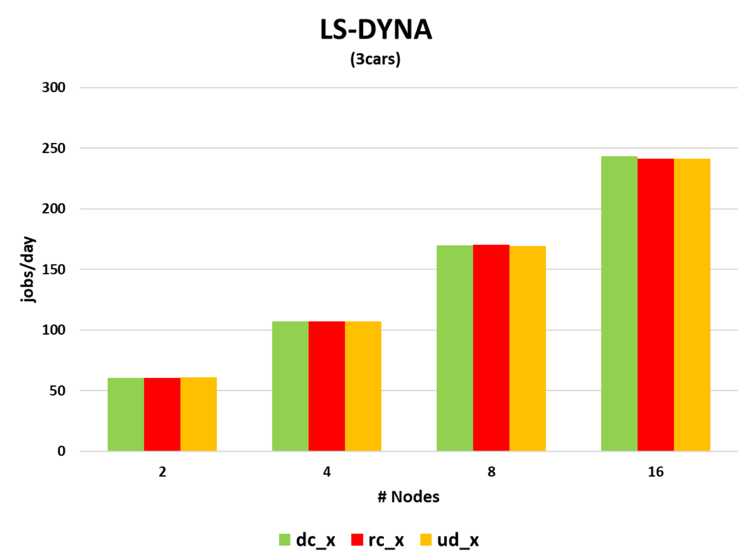

Figure 2: 3cars - InfiniBand Transport Comparison

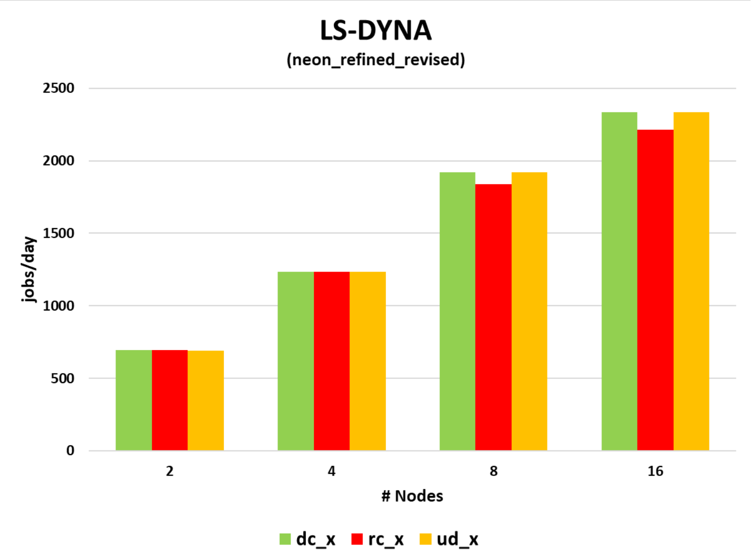

Figure 3: Neon_refined_revised - InfiniBand Transport Comparison

In both examples, DC transport reduces the number of needed connection channels, enabling the highest performance and scalability in comparison to the other two transport options.

Unified Communication X (UCX)

Unified Communication X (UCX) is an open source framework that provides an efficient and relatively easy way to construct widely used HPC protocols, including: MPI tag matching, RDMA operations, rendezvous protocols, stream, fragmentation, remote atomic operations, and others. The framework's design, data structures, and components are designed to provide highly optimized access to the network hardware to achieve the highest performance. UCX supports a high-level API for a broad range of HPC programming models.

UCX is natively supported in HPC-X MPI and was recently added as the preferred P2P library for Intel MPI, starting from Intel MPI 2019 update 5 (declared at SC19). When used with InfiniBand, UCX selects the InfiniBand transports based on performance and system size and characteristics.

UCX is managed by the UCF consortium, alongside other projects such as SparkUCX, Unified Collective Communications (UCC) and more.

MPI Performance Evaluation

The following testing compares the performance of HPC-X MPI and Intel MPI, both utilizing UCX for the underlying communication layer. The testbed setup is as follows:

- OS: CentOS 7.7
- Driver: MLNX_OFED 4.7
- CPU: Intel Gold 6138 @2GHz, dual socket 20 cores per socket (dual socket)
- Network: InfiniBand HDR
- LS-DYNA Version: ls-dyna_mpp_s_R11_1_0_x64_centos65_ifort160_avx2_intelmpi-2018
- Input: 3cars_rev02
- IO: RAMFS
- MPI: HPC-X 2.6.0, Intel MPI 2019 u7

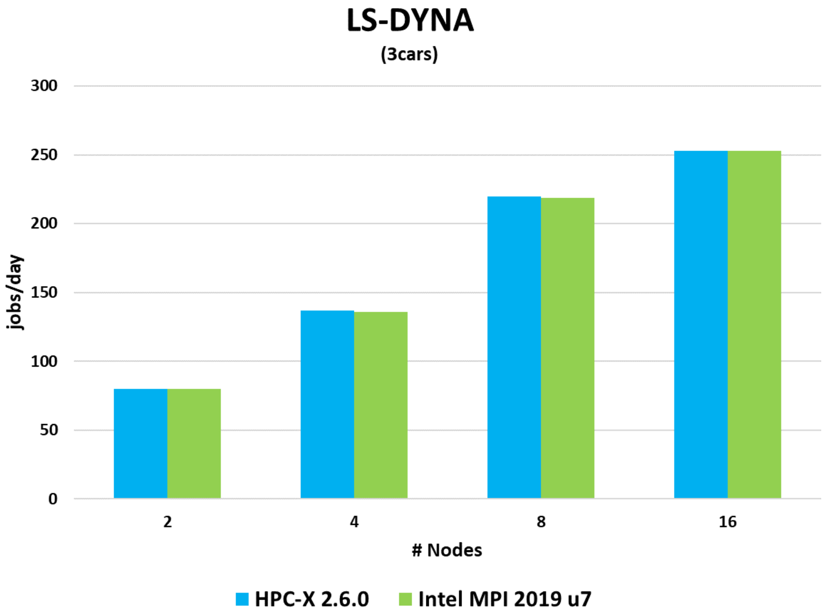

Figure 4: 3cars - InfiniBand MPI Comparison

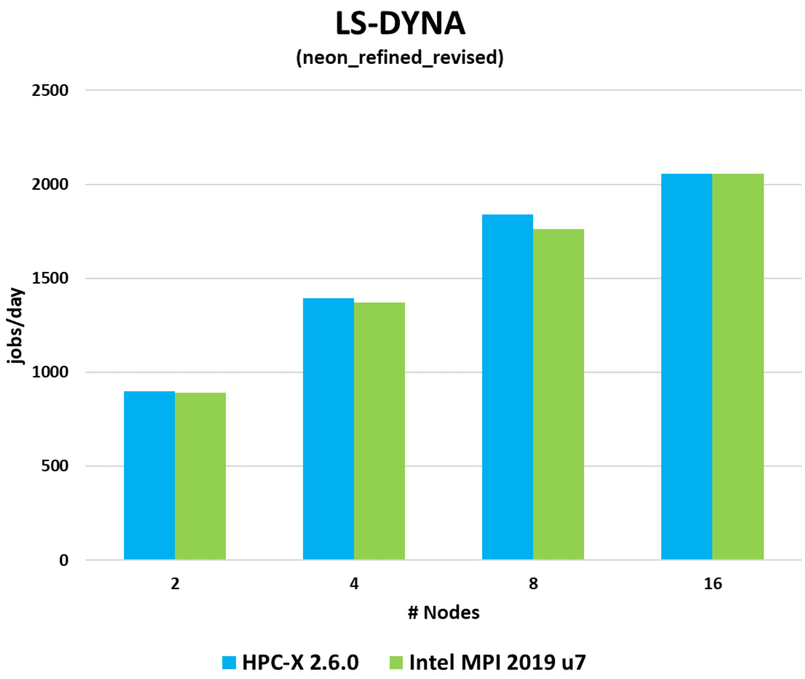

Figure 5: Neon_refined_revised - InfiniBand MPI Comparison

As both MPI libraries utilize UCX as the underlying communication framework, both MPI libraries provide similar performance for the tested benchmarks – 3cars and Neon Refined Revised.

Summary

HPC cluster environments impose high demands for connectivity throughput and low latency with low CPU overhead, network flexibility, and high efficiency. Fulfilling these demands requires the maintenance of a balanced system that can achieve high application performance and high scaling. With the increase in the number of CPU cores and application threads, simulation-complexity and data-volume requiring analysis, there is a need to develop a new HPC cluster architecture—a data-centered architecture rather than the traditional CPU-centered architecture. Co-Design collaboration, In-Network Computing technologies and higher network speeds enable higher application performance and overall data-center efficiency.

In addition, the comparison of HPC-X MPI with traditional transport options showed that DC transport performs better or similar to RC and UD transports.

Lastly, we demonstrated that the adoption of UCX as the P2P transport for the recent Intel MPI 2019 u7 yields, as expected, the same performance results as those of HPC-X 2.6, which also uses UCX for its P2P transport.