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

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

1 Abstract

From concept to engineering, and from design to test and manufacturing, engineers from a wide range of industries face the ever-increasing need for complex and realistic models to analyze the most challenging industrial problems; Finite Element Analysis is performed to secure quality and speed up the development process. Powerful virtual development software aims to tackle the need for finite element-based Computational LS-DYNA simulations with superior robustness, speed, and accuracy. These simulations are designed to run effectively on large-scale computational High-Performance Computing (HPC) systems.

The new generation of InfiniBand In-Network Computing technology includes several elements, such as Scalable Hierarchical Aggregation and Reduction Protocol (SHARP)™, a technology that enables the execution of data reduction algorithms on network devices instead of on host-based processors. Other elements include smart MPI Tag Matching, rendezvous protocols, and more. These technologies are in use at several of the recently deployed large-scale supercomputers around the world, including the top TOP500 platforms.

The HPC-AI Advisory Council has conducted performance investigations, including low-level benchmarks and application use-cases, to evaluate its performance and scaling capabilities with the InfiniBand interconnect.

2 In Network Computing

The latest revolution in HPC involves the effort around a co-design approach – a collaborative effort to reach Exascale performance by taking a holistic system-level approach to fundamental performance improvements. This effort is enabled by "In-Network Computing“ - a new technology displacing the CPU-centric approach, which has reached the limits of its scalability in several ways. In-Network Computing, acting as “distributed co-processor,” can handle and accelerate the performance of various data algorithms, including reductions, and more.

In the past, smart interconnect development focused on offloading the network functions from the CPU to the network. With the new co-design efforts and In-Network Computing, the latest generation of smart interconnects will also offload data algorithms that will be managed within the network, allowing users to run these algorithms as the data traverses within the system interconnect, rather than waiting for the data to reach the CPU. In-Network Computing technology is the leading approach to achieve performance and scalability for Exascale systems. In-Network Computing transforms the data center interconnect into a “distributed CPU” and “distributed memory,“ which can overcome performance walls and enable faster and more scalable data analysis.

3 SHARP - Scalable Hierarchical Aggregation and Reduction Protocol

Scalable Hierarchical Aggregation and Reduction Protocol (SHARP)™ technology enables data reduction and aggregation operations on the interconnect components. SHARP technology has been implemented in the latest generation of InfiniBand solutions. With both the increasing amount of data that needs to be analyzed and higher simulation complexity, the traditional idea of analyzing data 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.

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).

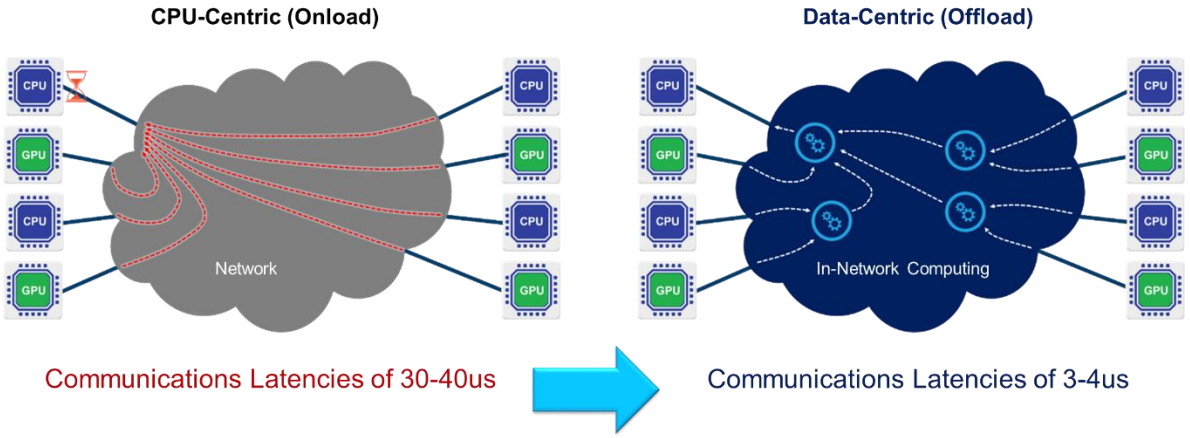

Figure 1: Illustration of SHARP Technology

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 first set of patterns being targeted are global reductions, including barrier synchronization and small data reductions. 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 ending up at the root of the tree.

This result is distributed in a method that may 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 interconnect-based, hardware-based MPI tag matching, MPI rendezvous offloads, and more.

4 HDR InfiniBand

HDR InfiniBand is the latest InfiniBand generation in today‘s market. HDR InfiniBand includes two network speeds – 200Gb/s (HDR) and 100Gb/s (HDR100). Beyond the faster data speeds, the HDR InfiniBand products include higher switch radix with 40 ports of 200Gb/s or 80 ports of 100Gb/s. The higher switch radix provides lower latency between neighbor processes and lower total cost of ownership. HDR InfiniBand technology also includes the second generation of SHARP, to enhance its acceleration capabilities for both deep learning applications and HPC workloads.

5 Performance Evaluation with In-Network Computing

The following performance tests were conducted using the resources of the HPC Advisory Council - HPC Cluster Center:

- 16 servers, each with the following characteristics:
  - Dual Socket Intel(R) Xeon(R) Gold 6138 CPU @ 2.00GHz

  - Mellanox HDR and HDR100 ConnectX-6 InfiniBand adapter
  - Mellanox EDR ConnectX-5 InfiniBand/Ethernet adapter
  - Intel® Omni-Path Host Fabric Adapter
  - 192GB DDR4 2677MHz RDIMMs per node
  - Operating system: Red Hat® Enterprise Linux® 7.5
- Mellanox InfiniBand HDR switch
- Mellanox InfiniBand EDR switch
- Intel Omni-Path Switch
- Mellanox Spectrum Ethernet switch 100Gb/s

In this example we used the following drivers and software:

- OS: CentOS 7.6, kernel 3.10.0-957.1.3.el7.x86_64
- Mellanox OFED: 4.5-1
- Intel IFS 10.9.0.0.2.1.0
- HPC-X 2.4 / IMPI 2018
- LS-DYNA 11 Single Precision
- I/O – local HDD

6 MPI Micro Benchmarks - MPI AllReduce

MPI AllReduce is a collective micro-benchmark that performs multiple iterations on all ranks and reduces a function to one result. A simple example is SUM, MAX, MIN or any other function-based operations, that take an argument from all ranks and reduces it to a single argument. In this example, we used OSU AllReduce implementation.

In the following test, we tested MPI AllReduce micro-benchmark for EDR InfiniBand SHARP, compared to native EDR InfiniBand and 100GbE RoCE (RDMA over Ethernet). Figure 2 demonstrates the AllReduce throughout performance with 32 server nodes and 1 process per node (PPN).

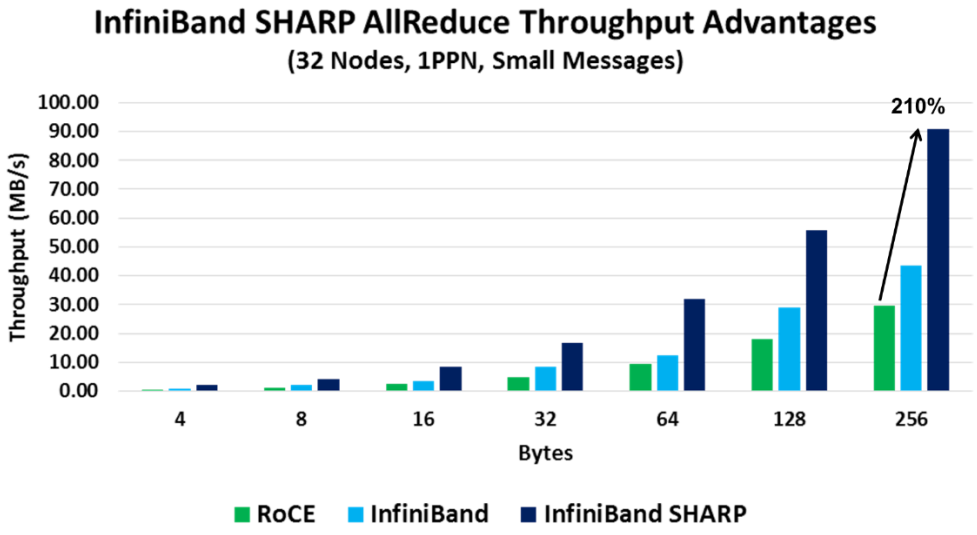

Figure 2: MPI AllReduce Performance

Figure 2 demonstrates the performance advantages of EDR InfiniBand SHARP, which enables 210% higher performance compared to 100GbE RoCE, and 109% higher performance compared to native EDR InfiniBand.

7 LS-DYNA Application Benchmarks

In this section, we have focused on the initial HDR and HDR100 InfiniBand performance compared to the proprietary Omni-Path network. The main difference between InfiniBand and Omni-Path is via its core network architecture. InfiniBand architecture is based on an offload approach, in which the network manages and executes the network function by itself, while Omni-Path architecture is based on an onload approach – leaving the network functions to be executed and managed by the CPU (via software).

We have compared the results of HDR100 InfiniBand to Omni-Path, as both options provide network throughput of 100Gb/s. We have also compared the results of HDR InfiniBand, to review the effect of larger network throughput on the applications performance.
Our testing platform was limited to 16 servers, and therefore we could not test at scale. Future work will expand to cover at-scale testing.

We have benchmarked a LS-DYNA 3 Vehicle Collision benchmark. It involves simulating a 3-car collision, where a compact car is hit from the rear by a van and collides into a midsize car. Figure 3 presents the performance results of HDR100 InfiniBand compared to the proprietary 100Gb/s Omni-Path network.

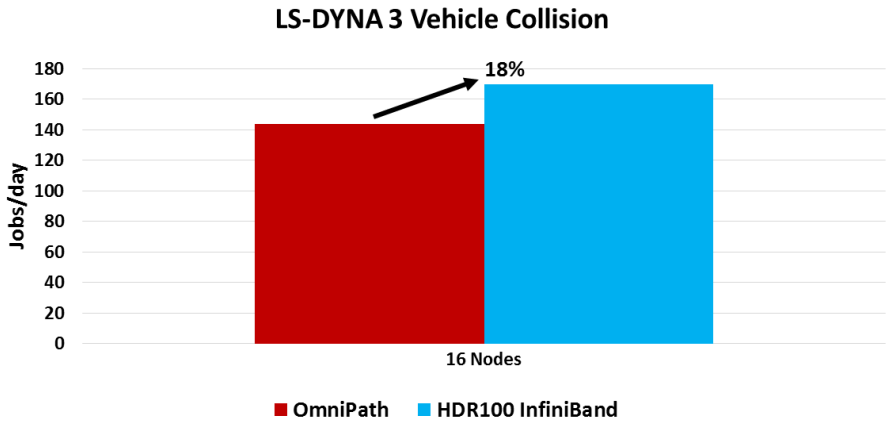

Figure 3: LS-DYNA 3 Vehicle Collision Performance Results of HDR100 InfiniBand and 100Gb/s Omni-Path

While both HDR100 and Omni-Path provide the same data throughput, HDR100 enables 18% higher performance for LS-DYNA, due to its offloading architecture and the In-Network Computing acceleration engines.

For PCIe Gen3 servers, HDR InfiniBand adapters are configured as 2 devices, with each including 16 lanes of PCIe Gen3. This means that the main adapter needs to be plugged into one PCIe Gen3 server slot, and its extension card into a second PCIe Gen3 server slot. One can select from the available PCIe Gen3 slots such that each slot connects to a different CPU socket, and therefore enable direct connectivity from each CPU to the network. This is an ideal situation for running separate LS-DYNA jobs, each on a different CPU. Figure 4 presents the performance results of HDR InfiniBand based on this approach, compared to the proprietary 100Gb/s Omni-Path network (as presented in Figure 3).

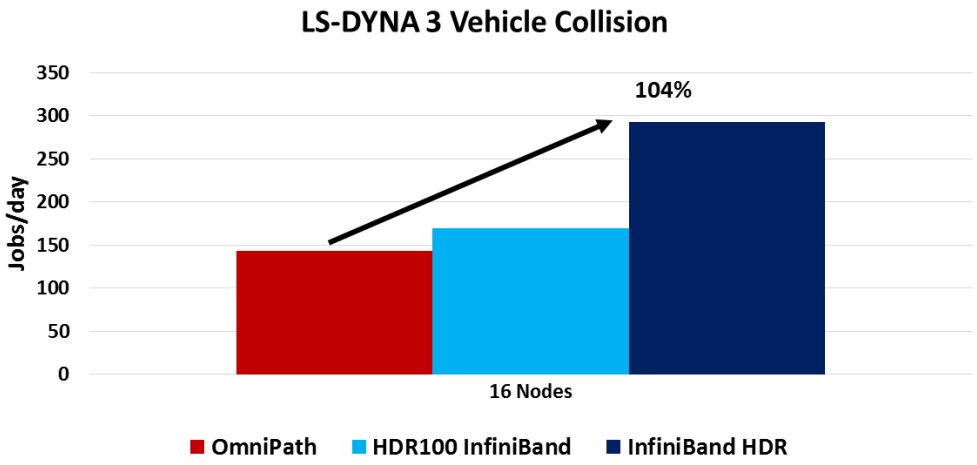

Figure 4: LS-DYNA 3 Vehicle Collision Performance Results of HDR InfiniBand, HDR100 InfiniBand and 100Gb/s Omni-Path

HDR InfiniBand expands the performance advantage of InfiniBand, demonstrating 104% higher performance compared to Omni-Path.

8 Conclusions

HPC cluster environments impose high demands on connectivity throughput and low latency, with low CPU overhead, network flexibility, and high efficiency. Fulfilling these demands enables the maintenance of a balanced system that can achieve high application performance and high scalability. 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 based on a data-focused architecture rather than the traditional CPU-focused architecture. The Co-Design collaboration, In-Network Computing technologies and higher network speeds enable higher application performance and overall data center efficiency.