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

Abstract

From concept to engineering, and from design to test and manufacturing, engineers from wide ranges of industries face ever increasing needs for complex, 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 is developed to tackle these needs for the finite element-based Computational LS-DYNA simulations with superior robustness, speed, and accuracy. Those simulations are designed to carry out on large-scale computational High-Performance Computing (HPC) systems effectively.

The new generation of InfiniBand In-Network Computing technology includes several elements – the Scalable Hierarchical Aggregation and Reduction Protocol (SHARP), a technology that enables to execute data reduction algorithm on the network devices instead of the host-based processor. Other elements include smart MPI Tag Matching and rendezvoused protocol, and more. These technologies are in use at some of the recent large-scale supercomputers around the world, including the top TOP500 platforms.

HPC-AI Advisory Council performed performance investigations including low level benchmarks and applications cases, to evaluate its performance and scaling capabilities with the InfiniBand interconnect.

In Network Computing

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

The past focus for smart interconnects development was to offload the network functions from the CPU to the network. With the new efforts in the co-design approach, the new 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 being transferred within the system interconnect, rather than waiting for the data to reach the CPU. This technology is being referred to as In-Network Computing, which is the leading approach to achieve performance and scalability for Exascale systems. In-Network Computing transforms the data center interconnect to become a “distributed CPU”, and “distributed memory”, enables to overcome performance walls and to enable faster and more scalable data analysis.

SHARP - Scalable Hierarchical Aggregation and Reduction Protocol

The Scalable Hierarchical Aggregation and Reduction Protocol (SHARP) is a technology that enables data reduction and aggregation operations on the interconnect components. SHARP technology has been implemented in the latest generation of InfiniBand solutions. With increases in the amount of data that need to be analysed and higher simulation complexity, the traditional concept of analysing 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 the performance wall by migrating these operations to the network, and performing them while the data is being transferred (Figure 1).

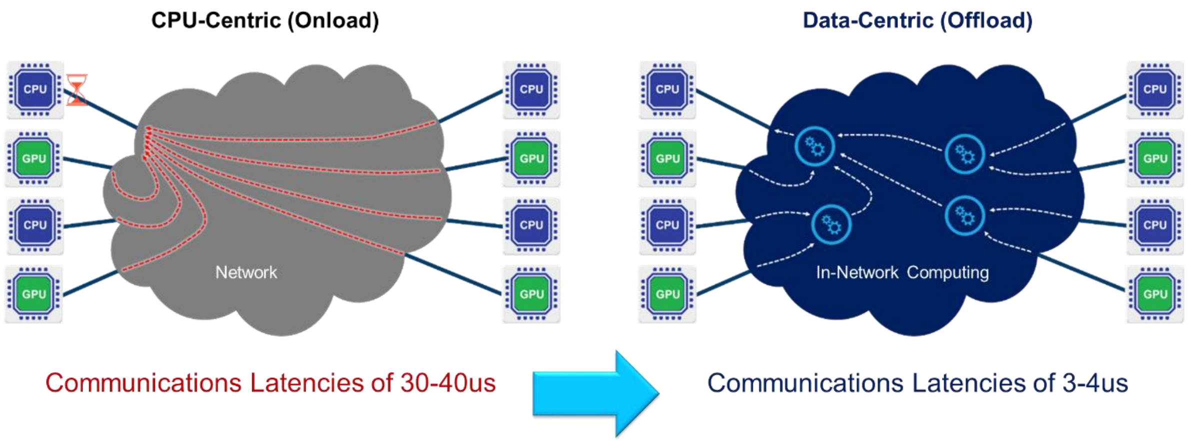

Figure 1: Illustration of SHARP Technology

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 of small amounts of data, 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 ends 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.

HDR InfiniBand

HDR InfiniBand is the latest InfiniBand generation in the market today. 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 neighbour processes and lower total cost of ownership. The HDR InfiniBand technology also includes the second generation of SHARP, to enhance its acceleration capabilities for deep learning applications as well as for HPC workloads.

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 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 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

MPI Micro Benchmarks - MPI AllReduce

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

In the following test, we've tested MPI AllReduce micro-benchmark for EDR InfiniBand SHARP, compares to native EDR InfiniBand, and 100GbE RoCE (RDMA over Ethernet).

Figure 2 demonstrates the AllReduce throughout performance with 32 servers nodes and 1 process per node (PPN).

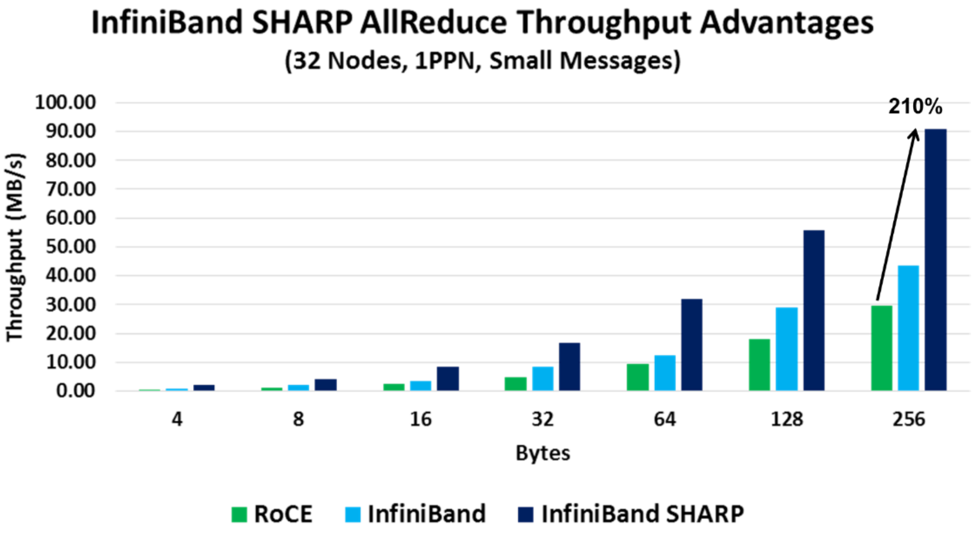

Figure 2: MPI AllReduce Performance

Figure 2 demonstrates the performance advantages of EDR InfiniBand SHARP, enabling 210% higher performance versus 100GbE RoCE and 109% higher performance versus native EDR InfiniBand.

LS-DYNA Application Benchmarks

In this section we have focused on the initial HDR and HDR100 InfiniBand performance compared to the proprietary OmniPath network. The main difference between InfiniBand and OmniPath 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 OmniPath 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 OmniPath, as both options provides 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 the testing to cover that part.

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

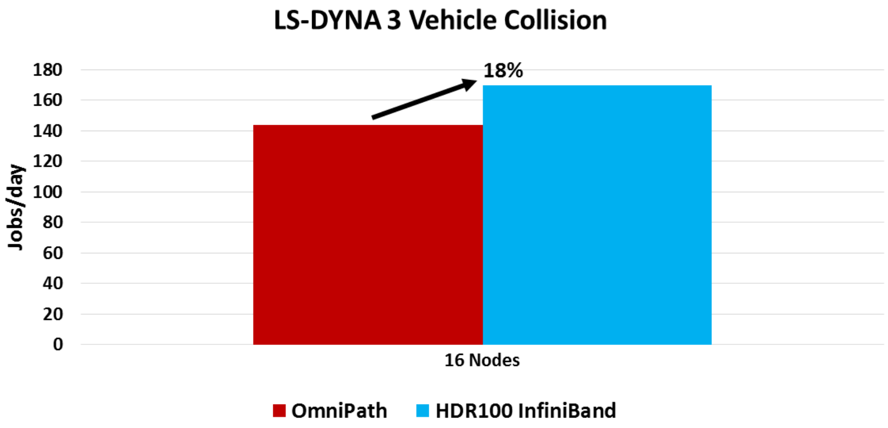

Figure 3: LS-DYNA 3 Vehicle Collision Performance Results of HDR100 InfiniBand and 100Gb/s OmniPath

While both HDR100 and OmniPath 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 Gen 3 servers, HDR InfiniBand adapters are configured as 2 devices, each has 16 lanes of PCIe Gen3. It means that the main adapter is needed to be plugged into one PCIe Gen3 server slot, and its extension card into a second PCIe Gen3 server slot. One can pick the PCIe Gen3 slots to be such that each slot connected 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 present the performance results of HDR InfiniBand with this approach, compared to the proprietary 100Gb/s OmniPath network (as presented in Figure 3).

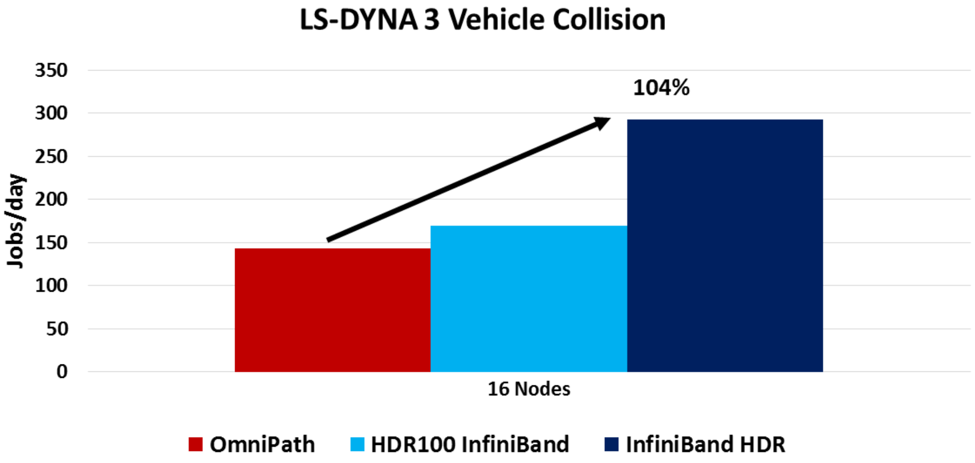

Figure 4: LS-DYNA 3 Vehicle Collision Performance Results of HDR InfiniBand, HDR100 InfiniBand and 100Gb/s OmniPath

HDR InfiniBand expand the performance advantage of InfiniBand, demontrating 104% higher performance compare to OmniPath.

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 scaling. With the increase in number of CPU cores and application threads, in simulation-complexity and in data volume requiring analysis, there is a need to develop a new HPC cluster architecture—a data-focused architecture rather than the traditional CPU-focused architecture. The Co-Design collaboration, the In-Network Computing technologies and the higher network speeds enable higher applications performance and overall data center efficiency.