The Effect of InfiniBand In-Network Computing on CAE Simulations

Ophir Maor, Yong Qin, Gerardo Cisneros-Stoianowski, 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 analyse the most challenging industrial problems; Finite Element Analysis is performed in an effort 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 Fluid Dynamics (CFD) 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.

1. 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.

2. SHARP - 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).

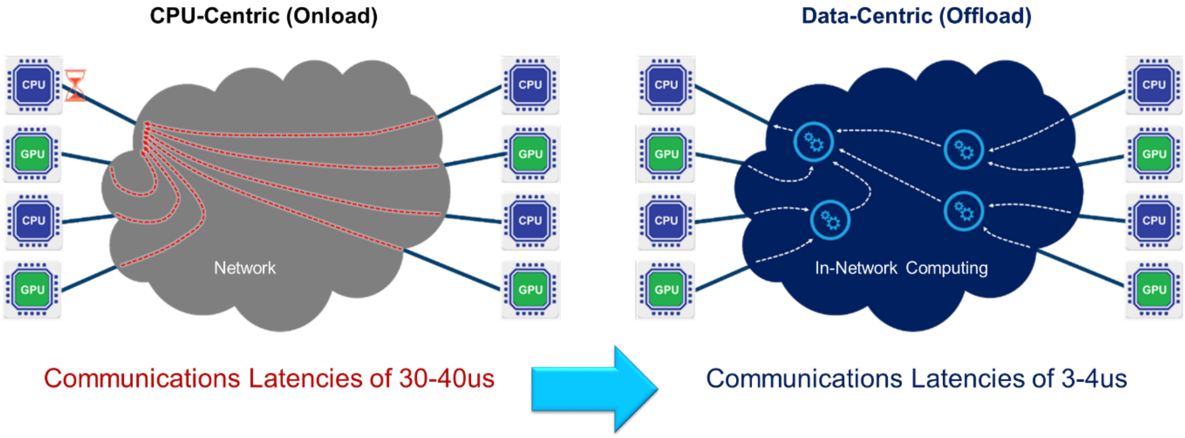

Figure 1: Illustration of SHARP Technology

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.

The latest HDR InfiniBand technology introduce SHARPv2, which allows streaming aggregation collectives. Streaming aggregation is performed on the network switches allows large message collective operations to be reduced in a line speed on the switches, instead of software implementation on one of the nodes.

3. Performance Evaluation with In-Network Computing

Benchmark System Configuration:

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 InfiniBand adapter
    - Intel® Omni-Path Host Fabric Adapter
    - 192GB DDR4 2677MHz RDIMMs per node
    - Operating system: Red Hat® Enterprise Linux® 7.5
- Mellanox InfiniBand switch
- Intel Omni-Path Switch
- Mellanox Spectrum Ethernet switch

4. Micro Benchmarks (MPI allreduce)

In this test, we've tested MPI allreduce micro-benchmark for InfiniBand SHARP, compares to native InfiniBand, and RoCE (RDMA over Ethernet). Figure 2 demonstrates the allreduce throughout performance with 32 servers nodes and 1 process per node (PPN).

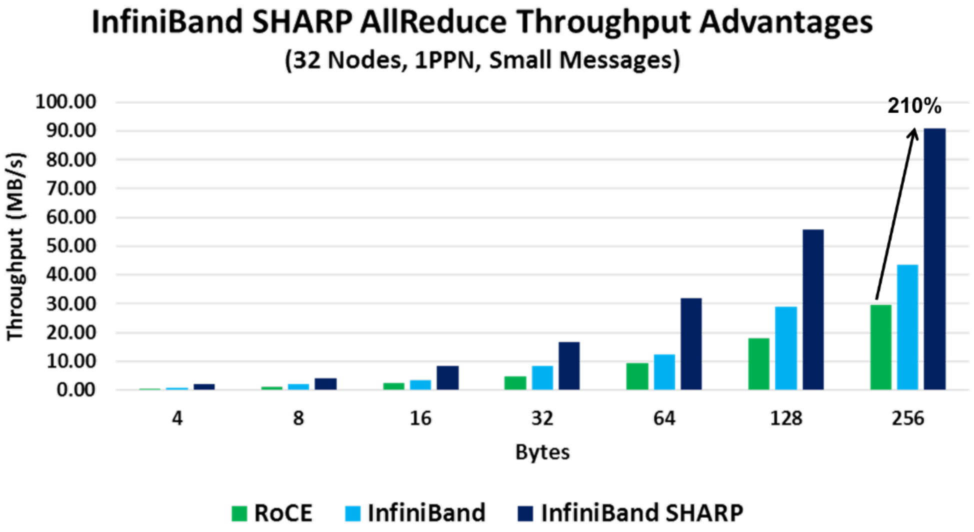

Figure 2: MPI allreduce Performance

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

5. Application Benchmarks

In this section we have compared the results of InfiniBand SHARP versus Omnipath. The main difference between InfiniBand SHARP and Omnipath is via its core network architecture. InfiniBand architecture is based on an offload approach, which deals with the network functions at the network level (hardware based implementation). Omnipath architecture is based on an onload approach – leaving the network functions to be executed and managed by the application. The following graphs compare the performance levels between the two interconnect technologies. Figure 3 and 4 showcase the performance results of OpenFOAM, and Figure 5 showcase the performance results of ANSYS Fluent.

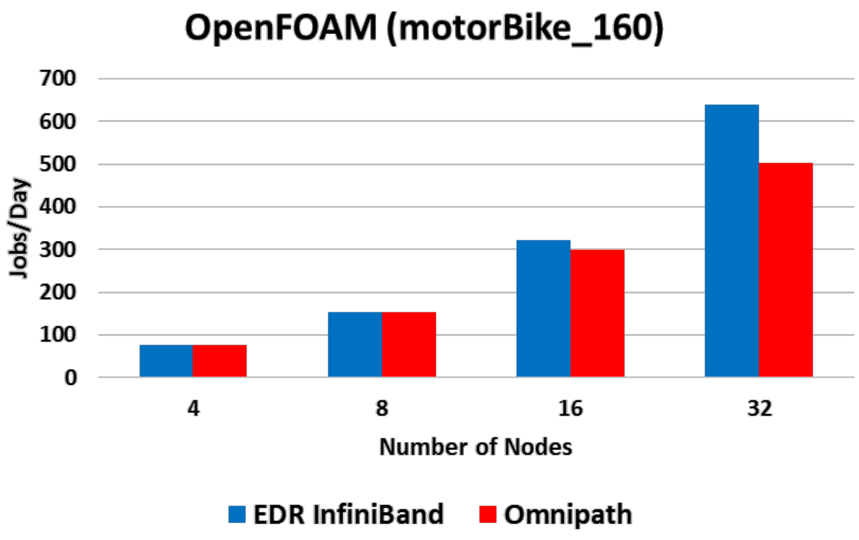

Figure 3: OpenFOAM performance using motorbike_160 benchmark

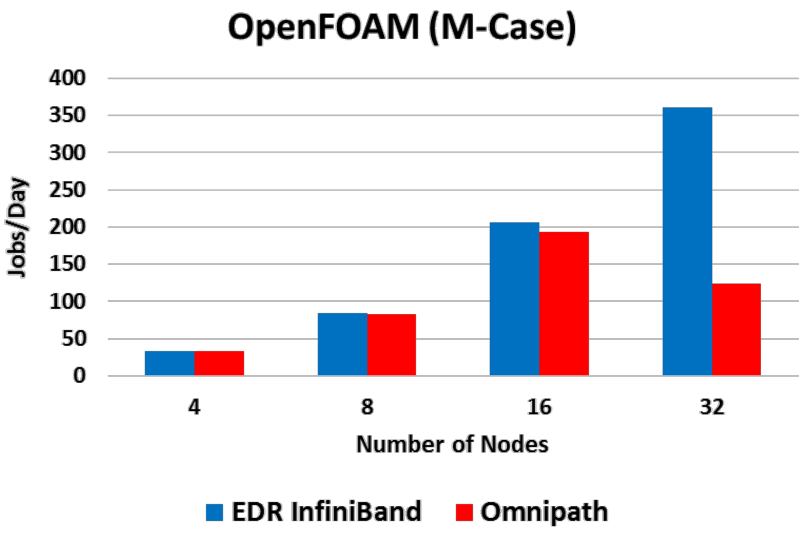

Figure 4: OpenFOAM performance results using M-Case benchmark

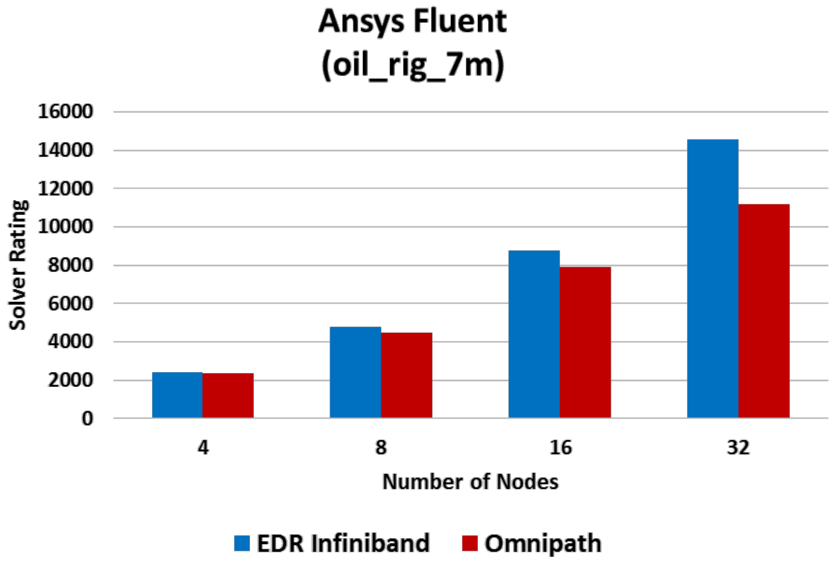

Figure 5: ANSYS Fluent performance results using oil_rig_7m benchmark

The OpenFOAM application performance results demonstrates the performance advantage of InfiniBand SHARP architecture, enabling higher performance from 27% to 190%. The ANSYS Fluent application performance results showcase that InfiniBand SHARP provides 30% higher performance versus Omnipath.

6. 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 enables the development of In-Network Computing technology that breaks the performance and scalability barriers, and moves us toward the next generation of HPC systems.