The Effect of In-Network Computing-Capable Interconnects on the Scalability of CAE Simulations

Ophir Maor, Gerardo Cisneros, David Cho, Yong Qin, Gilad Shainer

(HPC Advisory Council)

1 Abstract

The Co-Design collaboration is a collaborative effort among industry leaders, academia and manufacturers, whose mission is to reach the next level of application performance by exploiting system efficiency and optimizing performance, achieved through creating synergy between the hardware and the software. One of the major outcomes of this collaboration is In-Network Computing technology. This technology enables data algorithms, traditionally managed by the software on general processors, to be managed and executed by the data center interconnect by utilizing dedicated hardware components. This new approach dramatically improves application performance and overall data center return-on-investment (ROI). In this paper we describe and test the performance of computer-aided engineering (CAE) applications, benchmarked over the new architecture, and demonstrate its scaling and efficiency capabilities.

2 Introduction

High-performance computing (HPC) is a critical tool for numerical engineering simulations such as Finite Element Method for Structural Analysis, Computational Fluid Dynamics, and Multibody Simulation. HPC helps large enterprises drive faster time-to-market, realize significant cost reductions over laboratory testing, and achieve tremendous flexibility. HPC's strength and efficiency hinges on its ability to achieve sustained top performance by driving the CPU performance toward its limits. The motivation for high-performance computing in the industry has long been its tremendous cost savings and product improvements.

The recent trends in HPC cluster environments, ranging from multi-core CPUs, GPUs, and advanced high speed, to low-latency interconnect with offloading capabilities, are changing the dynamics of cluster-based simulations. Software applications are being reshaped for higher degrees of parallelism and multi-threading, and hardware is being reconfigured to solve new emerging bottlenecks to maintain high scalability and efficiency. CAE software is a toolbox for the development of customized numerical solvers such as continuum mechanics problems, including computational fluid dynamics (CFD).

Those HPC applications rely on Message Passing Interface (MPI), the de-facto messaging library for high performance clusters 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 cluster interconnect exponentially increase 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.

2.1 In-Network Computing

The latest revolution in HPC is the effort around the Co-Design collaboration, a collaborative effort among industry thought leaders, academia, and manufacturers to reach Exascale performance by taking a holistic system-level approach to fundamental performance improvements. Co-Design exploits system efficiency and optimizes performance by creating synergies between the hardware and the software components, and between the different hardware elements within the data center.

Co-Design recognizes that the CPU has reached the limits of its scalability, and offers an intelligent network as the new "co-processor" to share the responsibility for handling and accelerating application workloads. By placing data-related algorithms on an intelligent network, one can dramatically improve data center and application performance.

Smart interconnect solutions are based on offloading architectures, which can offload all network functions from the CPU to the network, freeing up CPU cycles and increasing system efficiency. Such interconnect solutions have long been proven to enable performance leadership and better scalability. In more recent efforts in Co-Design approach, the trend is to have the interconnect include data algorithms that will be managed and executed within the network, allowing users to run data algorithms on the data as it is being transferred within the system interconnect, rather than waiting for the data to reach the CPU. This technology is referred to as In-Network Computing, which is the leading approach to achieving Exascale system performance and scalability. In-Network Computing transforms the data center interconnect into “distributed CPU” and “distributed memory,” overcoming performance walls and enabling faster and more scalable data analysis.

2.2 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 EDR InfiniBand solutions. With increases in the amount of data that need to be analyzed and higher simulation complexity, the traditional concept 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 the performance wall by migrating these operations to the network, and performing them while the data is being transferred (Figure 1).

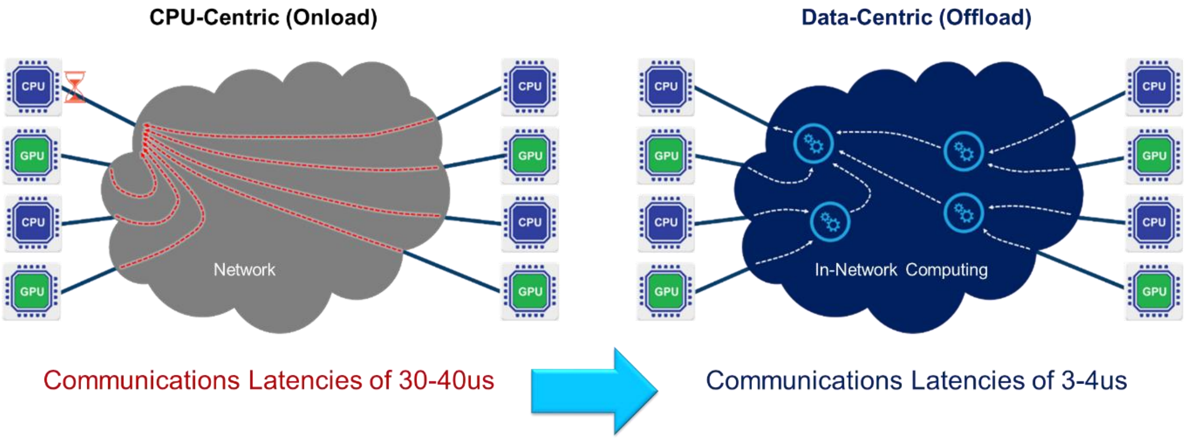

Figure 1: Illustration of SHARP Technology

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.

3 Performance Evaluation with In-Network Computing

3.1 Benchmark System Configuration

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

- 32 servers, each with the characteristics:
  - Dual socket Intel(R) Xeon(R) 16-Core CPU E5-2697V4 at 2.6GHz
  - Mellanox ConnectX®-5 EDR 100Gb/s InfiniBand adapters
  - Intel® Omni-Path 100Gb/s Host Fabric Adapter
  - 256GB DDR4 2400MHz RDIMMs
  - 1TB 7.2K RPM SSD 2.5" hard drive
  - Operating system: Red Hat® Enterprise Linux® 7.4
- Mellanox Switch-IB™ 2 SB7800 36-Port 100Gb/s EDR InfiniBand
- Intel Omni-Path 100Gb/s Switch

3.2 OpenFOAM

OpenFOAM is a free, open source CFD software released and developed primarily by OpenCFD Ltd since 2004. It has a large user base across most areas of engineering and science, from both commercial and academic organizations. OpenFOAM has an extensive range of features to solve anything from complex fluid flows involving chemical reactions, turbulence and heat transfer, to acoustics, solid mechanics and electromagnetics. OpenFOAM is released every six months to include customer sponsored developments and contributions from the community.

3.3 OpenFOAM Performance Comparison - Interconnect

We have tested OpenFoam application with the MotorBike_160 input file. It was run on the same cluster twice, but each time over a different interconnect: once with Intel Omni-Path, and once with EDR InfiniBand. The performance metric is the number of jobs per day (higher is better).

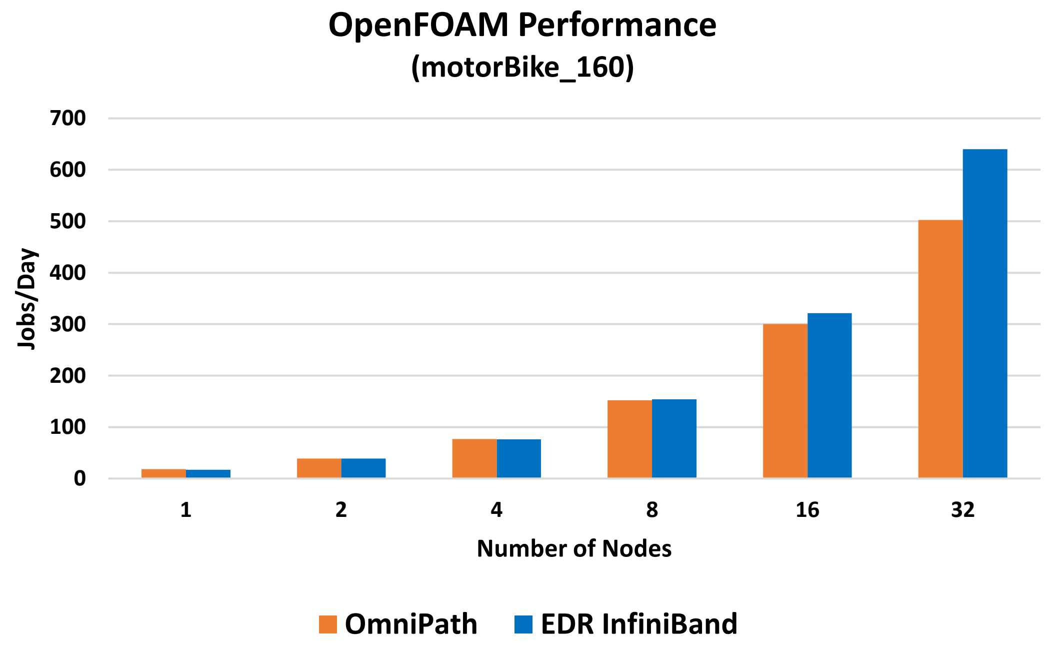

Figure 2: OpenFOAM Performance (motorBike_160)

Omni-Path is based on “onload” network architecture, an architecture that utilizes the CPU to manage and execute the network operations. In contrast, InfiniBand is based on “offload” network architecture that manages and executes the network operations at the network level, therefore freeing up the CPU cycles to be used for other application processes.

Figure 2 shows the performance results for EDR InfiniBand and for Omni-Path. The performance metric is the number of jobs per day (taking the job runtime and calculating the number of jobs that can be executed in 24 hours, and higher is better). The results showcase that In-Network Computing technology provides higher performance and efficiency. InfiniBand delivers nearly 30% higher performance at 32-nodes, and the performance advantage increases with the number of nodes.

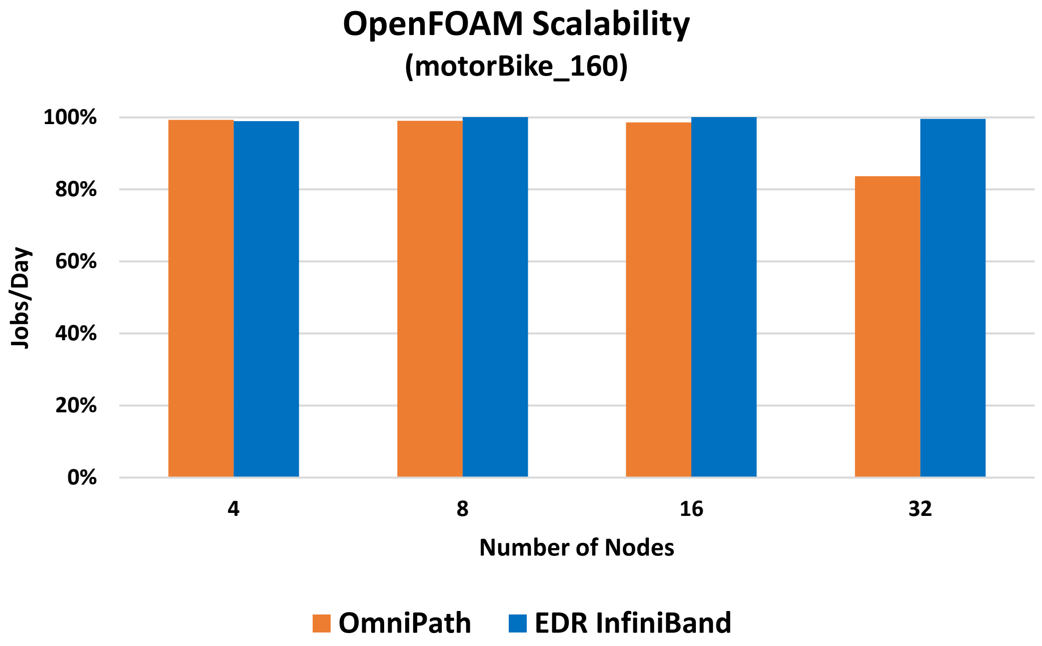

Figure 3: OpenFOAM Scalability per Interconnect Technology(motorBike_160)

Figure 3 explores the scaling performance between the two interconnect technology. We have looked on the scalability achieved when doubling the number of nodes. 100% means that the performance was also doubled when doubling the number of nodes. We can see that at lower number of nodes, the scalability between the two interconnect is nearly the same, but as the number of nodes increases, while InfiniBand maintains linear scalability, OmniPath scalability drops to 80%. With more nodes and more CPU cores being used, the load on the interconnect increases. Interconnect technologies that requires CPUs to be part of the data and control path will suffer from performance degradation, while interconnect technologies that are based on In-Network Computing architecture are set to continue and provide the highest performance, as they are not tied to the CPU load effects.

3.4 OpenFOAM Performance Comparison - MPI

The second set of tests compared two MPI (message passing interface) libraries – HPC-X and Intel MPI. The HPC-X MPI suite is based on OpenMPI with the addition of the available In-Network Computing technology that is part of the latest EDR InfiniBand solutions. The comparison to of was done in order to try and isolate the advantages of the In-Networking Computing elements.

We tested the same benchmark input file for OpenFoam (MotorBike_160). The performance for this test is shown in Figure 4. The unit of performance used for the graphs is jobs per day – how many similar jobs can be executed in a 24-hour period (higher is better).

The results showcase the performance and scalability advantages of In-Network Computing (HPC-X). For the benchmark, HPC-X provides more than 20% higher performance versus IntelMPI and better scalability.

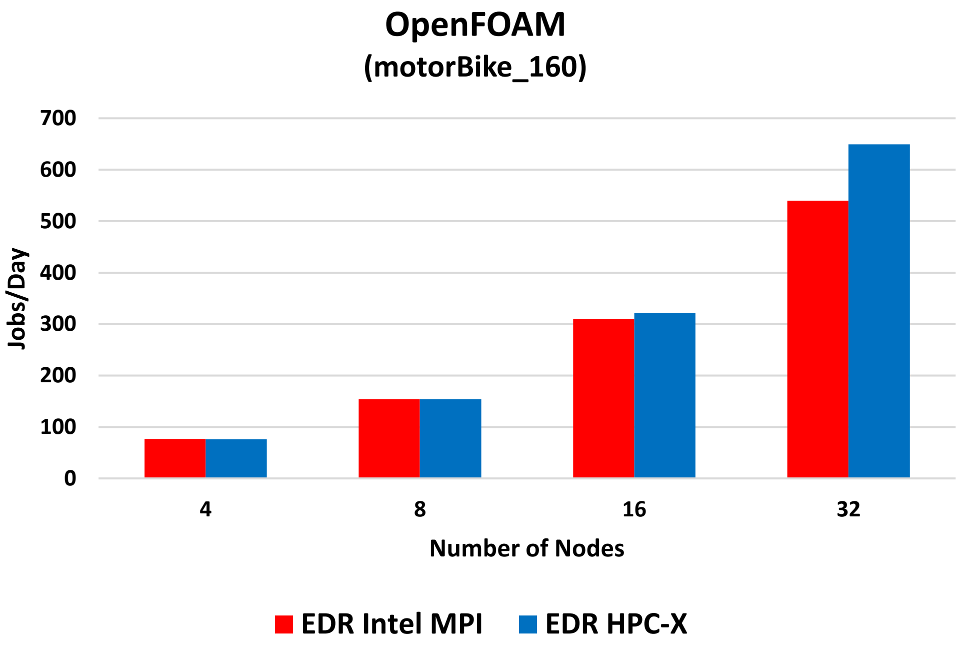

Figure 4: OpenFOAM Intel MPI to HPC-X Comparision

4 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.

The OpenFoam application was benchmarked for this study to demonstrate the advantages of In-Network Computing technology, implemented in the latest EDR InfiniBand interconnect. We have witness nearly 30% performance advantage and linear scalability with InfiniBand.