THE EFFECT OF IN-NETWORKING COMPUTING CAPABLE INTERCONNECTS ON SCALABILITY OF CAE SIMULATIONS

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

HPC Advisory Council

ABSTRACT

From concept to design, testing and manufacturing, engineers from a wide range of industries face an ever-increasing need for complex, realistic models to analyze the most challenging industrial problems. Finite Element Analysis (FEA) simulations help to both secure quality and speed up the development process. These simulations are designed effectively to run on large-scale computational High-Performance Computing (HPC) systems.

The latest revolution in HPC platforms is the move to a co-design architecture, to reach Exascale performance by taking a holistic system-level approach to fundamental performance improvements. Co-design architecture exploits system efficiency and optimizes performance by creating synergies between the hardware and the software, 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, we can dramatically improve data center and application performance.

Smart interconnect solutions are based on an "offloading architecture" that can offload all network functions from the CPU to the network, freeing CPU cycles and increasing the system's efficiency. With the innovative efforts of the co-design approach, the newer generations of interconnects are including more and more data algorithms that can be managed and executed within the network, thus allowing users to run data algorithms on the data as the data is being transferred within the system interconnect, rather than waiting for the data to reach the CPU. Today, In-Network Computing and In-Network Memory is the leading approach to achieving performance and scalability for Exascale systems.

HPC Advisory Council performed deep investigations on a few popular CFD software to evaluate its performance and scaling capabilities and to explore potential optimizations. The study reviews the recent developments of in-network computing architectures, and how they can influence on the runtime, scalability and performance of CAE simulations.

THE EFFECT OF IN-NETWORKING COMPUTING-CAPABLE INTERCONNECTS ON SCALABILITY OF CAE SIMULATIONS

1. Introduction

High-performance computing (HPC) is a critical tool for computer-aided engineering (CAE): from component-level design to full simulation analysis. 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 CAE industry has long been 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 test, and the same cluster can serve as the platform for every test simulation going forward.

The recent trends in high-performance computing 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. Fluent software Package from ANSYS Inc is a general-purpose structural and fluid analysis simulation software package, capable of simulating complex real-world problems. It is widely used in the CAE industry.

Fluent relies 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 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 behaviours.

2. 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 an offloading architecture, which can offload all network functions from the CPU to the network, freeing up CPU cycles and increasing the system's efficiency. These kind of interconnect solutions have proven in the past to enable performance leadership and better scalability. With more recent efforts in the Co-Design approach, the interconnect will also 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 achieve performance and scalability for Exascale systems. In-Network Computing transforms the data center interconnect into a "distributed CPU" and "distributed memory," overcoming performance walls and enabling faster and more scalable data analysis.

3. Scalable Hierarchical Aggregation and Reduction Protocol (SHARP)

SHARP is a technology that enables data reduction and aggregation operations on the interconnect components. SHARP has been implemented in the latest generation of EDR 100Gb/s InfiniBand solutions. With increases in the amount of data that needs to be analyzed and with higher simulation complexity, the traditional concept of analyzing data only on the compute elements has reached a latency wall. Adding more cores to handle the various data reduction or aggregation operations does not result in any performance improvement. SHARP helps to overcome the performance wall by migrating data reduction and aggregation operations to the network and by performing them while the data is being transferred.

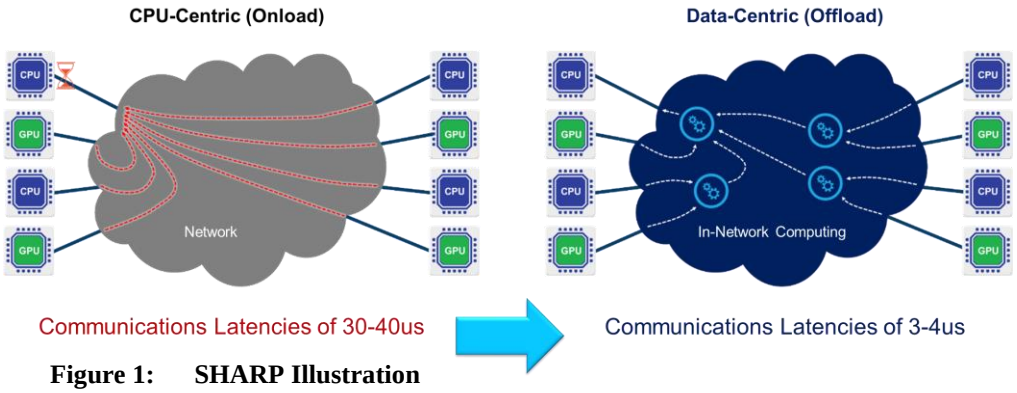

Figure 1: SHARP Illustration

The goal of In-Network Computing architecture is to optimize the completion time of frequently used global communication patterns and to minimize CPU utilization. The first set of targeted patterns is global reductions of small amounts of data, including barrier synchronization and small data reductions. SHARP provides an abstraction layer describing the 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 leaves, and makes its way up the tree, with data reductions occurring at each AN, with the global aggregate ending up at the root of the tree. This result is distributed in a way 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. Performance Evaluation with In-Network Computing

The following performance tests were conducted using the resources of the HPC Advisory Council HPC cluster center. We used a cluster based on Dell PowerEdge R730 servers. Each server consists of:

- Dual socket Intel(R) E5-2697A V4 CPUs at 2.60GHz
- Mellanox ConnectX-5 EDR 100Gb/s InfiniBand adapters
- Memory of 256GB DDR4 2400MHz RDIMMs per node
- 1TB 7.2K RPM SSD 2.5" hard drive per node.

The networking switch is Mellanox Switch-IB 2 SB7800 36-Port 100Gb/s EDR InfiniBand switch.

In the following tests, we used **ANSYS Fluent v19**, and the operating system was Red Hat Enterprise Linux 7.4.

For the test, we selected two of the fluent benchmarks **ice_2m** and **oil_rig_7m**, to perform the analysis on.

The first set of tests compared four MPI libraries:

- OpenMPI 3.1 “Vanilla”
- Intel MPI
- Platform MPI
- HPC-X

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 the other MPI libraries was done in order to try and isolate the advantages of the In-Networking Computing elements.

We tested both the **ice_2m** and **oil_rig_7m** Fluent benchmarks. The performance for these tests appears in the following figures. The unit of performance used for the graphs is solver rating – it defines the performance levels of the job (higher is better).

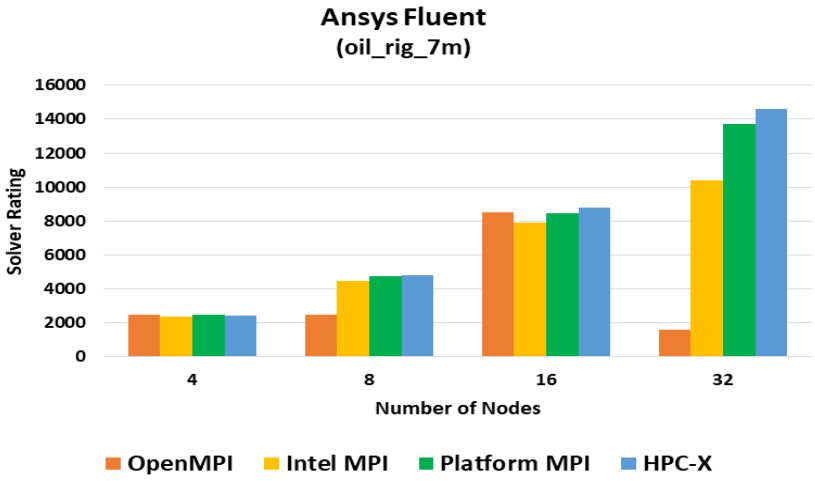

Figure 2: MPI Performance Comparison - oil_rig_7m

The results showcase the performance and scalability advantages of In-Network Computing (HPC-X). For the **oil_rig_7m** benchmark**,** using HPC-X MPI provides higher performance than OpenMPI (which doesn't scale beyond 16 node), and 41% performance improvement over Intel MPI.

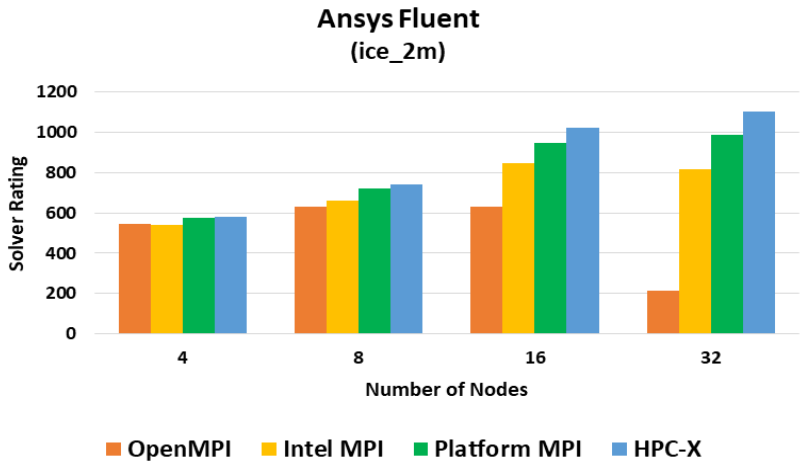

Figure 3: MPI Performance Comparison – ice_2m

For the **ice_2m** benchmark case, HPC-X exhibits higher performance and better scalability than OpenMPI, and 35% better scalability than Intel MPI. Neither Intel MPI nor Open MPI scale beyond 16 nodes.

The second part of the testing drew a comparison between InfiniBand and Omni-Path using the oil_rig_7m input file. Omni-Path is an "onload" network architecture, which relies on the CPU to manage and execute the network operations, while InfiniBand is an "offload" network architecture that manages

and executes the network operations at the network level; therefore InfiniBand frees up expensive CPU cycles to be used for other applications.

The following figure shows the performance results for EDR InfiniBand and Omni-Path. The performance metric is the Solver Rating (higher is better). The results showcase the advantage of the In-Network Computing technology in providing higher performance and data center efficiency. InfiniBand extracts higher productivity from the cluster, resulting in a higher number of jobs that can be run during a given time period.

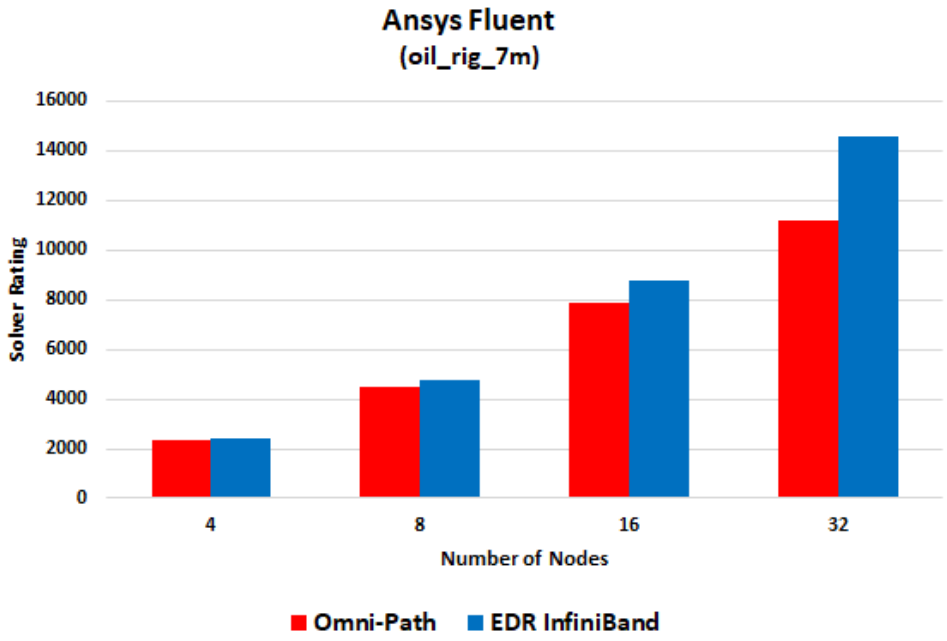

Figure 4: Network Adapter Comparison – oil_rig_7m

Similarly, InfiniBand extracts 30% higher productivity than Omni-Path when using the oil_rig_7m input file, for an overall better return on investment. In contrast, when using ice_2m, InfiniBand performs 29% better.

As the oil_rig_7m and ice_2m benchmarks demonstrate good scaling capability, we believe that the gap between InfiniBand and Omni-Path will continue to increase with cluster size.

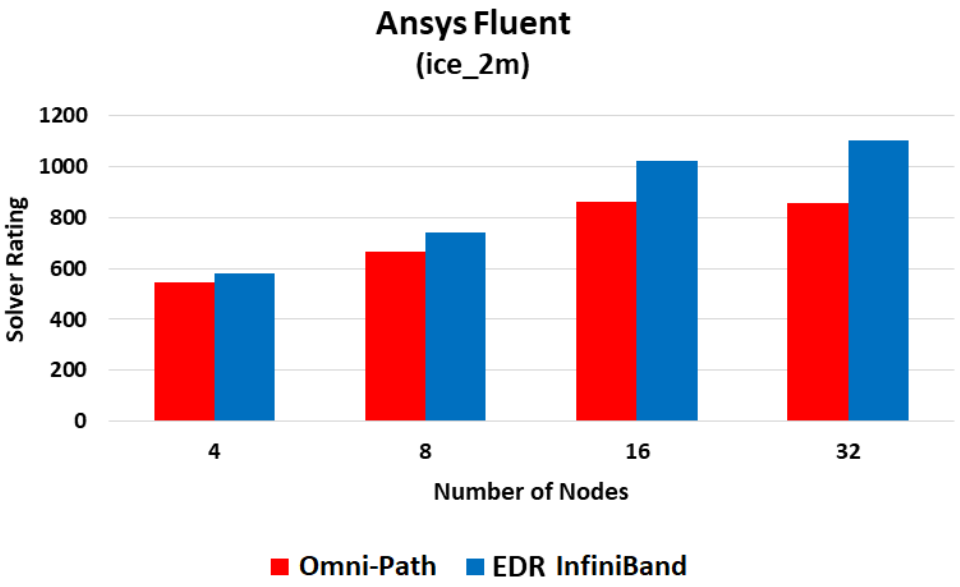

Figure 5: Network Adapter Comparison – ice_2m

5. Conclusions

HPC cluster environments impose high demands for connectivity throughput and low latency with low CPU overhead, network flexibility, and high-efficiency. Fulfilling these demands allows for 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, simulation complexity and in-data volume requiring analysis, there is a need to develop a new HPC cluster architecture - one that will be data-focused instead of the traditional CPU- focused architecture. The Co-Design collaborations enable the development of In-Network Computing technology that breaks the performance and scalability barriers, and moves us to the next generation of HPC systems.

ANSYS Fluent software was benchmarked for this study to understand the advantages of In-Network Computing technology that was implemented in the latest EDR InfiniBand interconnect solution. In all cases, ANSYS Fluent demonstrated higher performance and scalability with EDR InfiniBand.