THE EFFECT OF HPC CLUSTER ARCHITECTURE ON THE SCALABILITY OF CAE SIMULATIONS

Pak Lui, David Cho, Gilad Shainer

HPC Advisory Council

SUMMARY

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 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 breakthrough in HPC parallel computing that allows such complex analyses to be performed that generate the high-quality results, while reducing simulation time from days to just hours. Those simulations involve complex calculations and data exchanges among computational systems in HPC systems. The more complex simulations are being performed, the higher demands from the cluster performance are.

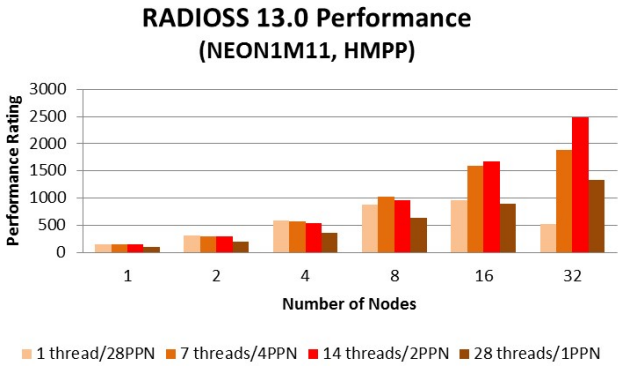
The recent trends in HPC cluster environments, such as the use of multi-core CPUs, and low-latency, high-speed Ethernet and InfiniBand cluster interconnect with offloading capabilities are changing the dynamics of clustered-based simulations. Applications are being reshaped in order to maintain high scalability and efficiency.

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 HPC clustering architectures, CPUs and interconnect solutions and how they can influence on the runtime, scalability and performance of CAE simulations. This study will present optimization techniques and networking communication profiling to further understand various software's dependencies on communication networks and the underlying hardware. Deep analyses of communication profiles will be explored and guidelines for optimized productivity will be introduced. The research will review the effects by comparing various hardware using different simulation models.

1. Consideration of Modern CPU Cores

The proliferation of CPU cores has become a commonplace for the compute node used in the modern HPC cluster. As more compute nodes are being deployed, more of the CPU cores can be utilized for computation. As large number of CPU cores being available, performance does not always correlate linearly to the amount of cores being used. It often requires careful placement of user processes to the CPU cores to achieve good performance. One paradigm to achieve good scalability performance is by deploying the application processes via hybrid of MPI processes which spawn threads to run on CPU cores.

As demonstrated by the prior study[1] on Altair RADIOSS simulation of a car crash model, each compute node provides 28 CPU cores. As more CPU cores involved, the scalability would be limited if all cores are involved in MPI communication. With the use of hybrid execution model available in RADIOSS, it would allow a subset of MPI processes to launch, and the associated MPI process would spawn additional worker threads onto the CPU cores to allow maximum performance and scalability performance. At 32 nodes/896 cores, it is observed that the best configuration to run is by spawning 2 MPI processes to each node, each MPI process would be allocated to a socket and spawn additional threads to fill the CPU cores. On the other hand, spawning a single MPI process to each node is not advised due to breach of data locality in system memory across different CPU sockets.



[1] HPC Advisory Council RADIOSS 13.0 Performance Benchmark and Profiling. Available at: http://hpcadvisorycouncil.com/pdf/RADIOSS_Analysis_and_Profiling_Intel_2697.pdf

Figure 1: Analysis for best combination of process per node (PPN) and threads in Altair RADIOSS

2. The Impact by Network Technologies

CAE software utilizes Message Passing Interface (or MPI) for HPC cluster communications. MPI is the de-facto messaging library for HPC clusters; the fast messaging among the systems depend on modern network that can provide low latency and high messaging rate. Performance demands from the cluster interconnect increase dramatically as the simulation requires more complexity to properly simulate the physical model behavior.

The cluster interconnect is critical for efficiency and scalability performance in the multi-core era. The overall cluster productivity can increase only by the presence of a high-speed interconnect that is capable of handing the increase of messages due to the high CPU core counts available. We have compared the performance of a few simulations using Ethernet and InfiniBand.

Ethernet is typically used in the data center to provide a cost-effective solutions for transferring data between the compute nodes and storage systems. However, due to the communication that takes place in HPC simulations also involved group and multicast communications, the communication patterns would overwhelm Ethernet network even at a small scale. The details in the comparisons are shown from prior studies[23] revealed that Ethernet would either limit performance or would cause bottleneck when running in the HPC environment.

InfiniBand is a network interconnect with a scalable architecture for delivery performance on large high-performance clusters environment, specifically designed for use by the large HPC cluster and supercomputers. With low-latency, high-bandwidth and extremely low CPU overhead, InfiniBand has become the most deployed high-speed interconnect for HPC clusters. The InfiniBand architecture is an industry-standard that is designed to scale to tens of thousands of nodes. The Zero-copy and RDMA technologies in InfiniBand allow

---

[2] HPC Advisory Council. STAR-CCM+ 10.02.012 Performance Benchmark and Profiling: http://www.hpcadvisorycouncil.com/pdf/STAR-CCM_Analysis_Intel_E5_2697v3.pdf

[3] HPC Advisory Council. ANSYS Fluent 16.1 Performance Benchmark and Profiling: http://hpcadvisorycouncil.com/pdf/Fluent_Analysis_and_Profiling_Intel_E5_2697v3.pdf

communication processing to be offloaded to the network hardware, allowing CPU to be focused on processing application computation.

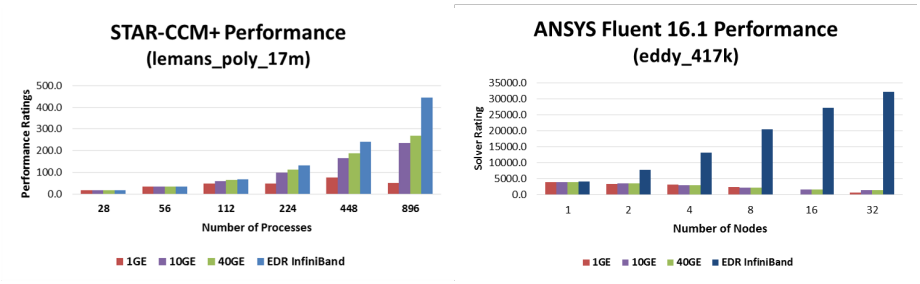


Figure 2: Interconnect Comparisons between Ethernet and InfiniBand on CD-Adapco STAR-CCM+ and ANSYS Fluent

InfiniBand demonstrated good scalability throughout the various tested configurations. The results show InfiniBand delivered better scalability in application performance. The faster run time translates into higher performance rating, or higher productivity. InfiniBand-based simulation performance consistently outperforms other interconnect, while the scalability performance improves, while Ethernet showed some loss of performance.

3. To Understand Scalability Bottleneck through MPI Profiling

By inspecting the MPI profile of some of CAE applications, we can observe the underlying MPI communications that can contribute to bottleneck in performance. By understanding the underlying communication patterns, we can observe in the type of MPI calls that are involved. We will review the communication patterns of a few CAE applications to demonstrate the MPI communication could have an effect on the scalability performance of an application.

REFERENCES

HPC Advisory Council, "RADIOSS 13.0 Performance Benchmark And Profiling".http://hpcadvisorycouncil.com/pdf/RADIOSS_Analysis_and_Profiling_Intel_2697.pdf
HPC Advisory Council, "STAR-CCM+ 10.02.012 Performance Benchmark and Profiling".
http://www.hpcadvisorycouncil.com/pdf/STAR-CCM_Analysis_Intel_E5_2697v3.pdf
HPC Advisory Council, "ANSYS Fluent 16.1 Performance Benchmark and Profiling ".
http://hpcadvisorycouncil.com/pdf/Fluent_Analysis_and_Profiling_Intel_E5_2697v3.pdf