Maximizing Cluster Utilization for LS-DYNA® Using 100Gb/s InfiniBand

Pak Lui, Gilad Shainer, Scot Schultz
Mellanox Technologies, Inc.

Abstract

From concept to engineering and from design to test and manufacturing, the automotive industry relies on powerful virtual development solutions. Crash simulations are performed in an effort to secure quality, safety and accelerate the development process. As the models become more complex to better simulate the physical behavior in crash simulations, the computers that run as a cluster also need to be higher to meet the needs of the higher standards for simulating these more elaborate models. Among the various components in a compute cluster, the high performance network interconnect is an integral factor which is key in making the simulation run efficiently. The Mellanox Connect-IB™ InfiniBand adapter has introduced a novel high-performance and scalable architecture for high-performance clusters. The architecture was designed from the ground up to provide high performance and maximize scalability for the largest supercomputers in the world today and in the future. This paper demonstrates the new features and technologies driven by the Connect-IB InfiniBand adapters. Besides its raw abilities of delivering sub-microsecond latency and a full bandwidth of over 100Gbps using two of the FDR links, its hardware capabilities also includes CPU offloads, MPI collective operations acceleration and message transport services that make LS-DYNA to perform at scale. This paper also demonstrates running multiple parallel simulations to achieve higher cluster productivity, in an effort to exploit with this new level of performance available from the network.

Introduction

High-performance computing (HPC) is a crucial tool for automotive design and manufacturing. It is used for computer-aided engineering (CAE) from component-level to full vehicle analyses: crash simulations, structure integrity, thermal management, climate control, engine modeling, exhaust, acoustics and much more. HPC helps drive faster time-to-market, significant cost reductions, and tremendous flexibility. The strength in HPC is the ability to achieve best sustained performance by driving the CPU performance towards its limits. The motivation for high-performance computing in the automotive 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 crash test, and the same cluster can serve as the platform for every test simulation going forward.

The recent trends in cluster environments, such as multi-core CPUs, GPUs, and new interconnect speeds and offloading capabilities are changing the dynamics of clustered-based simulations. Software applications are being reshaped for higher parallelism and multi-threading, and hardware is being configured for solving the new emerging bottlenecks, in order to maintain high scalability and efficiency. LS-DYNA software from Livermore Software Technology Corporation is a general purpose structural and fluid analysis simulation software package capable of simulating complex real world problems. It is widely used in the automotive industry for crashworthiness analysis, occupant safety analysis, metal forming and much more. In most cases, LS-DYNA is being used in cluster environments as they provide better flexibility, scalability and efficiency for such simulations.

LS-DYNA relies on Message Passing Interface (or MPI) for cluster or node-to-node communications, the de-facto messaging library for high performance clusters. MPI relies on fast server and storage interconnect in order to 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.

Mellanox introduced the latest Connect-IB™ 56Gb/s FDR InfiniBand adapter, which has a novel high-performance and scalable architecture for high-performance clusters. The architecture was planned from the outset to provide the highest-possible performance and scalability, specifically designed for use by the largest supercomputers in the world.

HPC Clusters

LS-DYNA simulations are typically carried out on high-performance computing (HPC) clusters based on industry-standard hardware connected by a private high-speed network. The main benefits of clusters are affordability, flexibility, availability, high-performance and scalability. A cluster uses the aggregated power of compute server nodes to form a high-performance solution for parallel applications such as LS-DYNA. When more compute power is needed, it can sometimes be achieved simply by adding more server nodes to the cluster.

The manner in which HPC clusters are architected has a huge influence on the overall application performance and productivity – number of CPUs, usage of GPUs, the storage solution and the cluster interconnect. By providing low-latency, high-bandwidth and extremely low CPU overhead, InfiniBand has become the most deployed high-speed interconnect for HPC clusters, replacing proprietary or low-performance solutions. The InfiniBand Architecture (IBA) is an industry-standard fabric designed to provide high-bandwidth, low-latency computing, scalability for tens of thousands of nodes and multiple CPU cores per server platform, and efficient utilization of compute processing resources.

Some of the key features of the Connect-IB architecture that enable its cluster performance superiority are described in the following section.

Connect-IB Architecture

Connect-IB is the first InfiniBand adapter on the market that enables 100Gb/s uni-directional throughput (200 Gb/s bi-directional throughput) by expanding the PCI Express 3.0 bus to 16-lanes and through dual 56Gb/s FDR InfiniBand network ports. In addition, the internal data path of the device can also deliver over 100Gb/s data throughput. Thus, MPI and other parallel programming languages can take advantage of this high throughput, utilizing the multi-rail capabilities built into the software. While Mellanox ConnectX®-3 adapters enabled applications running on the Intel Ivy Bridge systems to realize the full capabilities and bandwidth of the PCI Express 3.0 x8 bus, Connect-IB adapters increase these capabilities. This increase is important for bandwidth sensitive applications, and even more critical with the advent of increased CPU cores such as the new Intel Ivy Bridge systems. In addition, this level of throughput will be required to satisfy the needs of new heterogeneous environments such as GPGPU and Intel Xeon

PHI based endpoints.

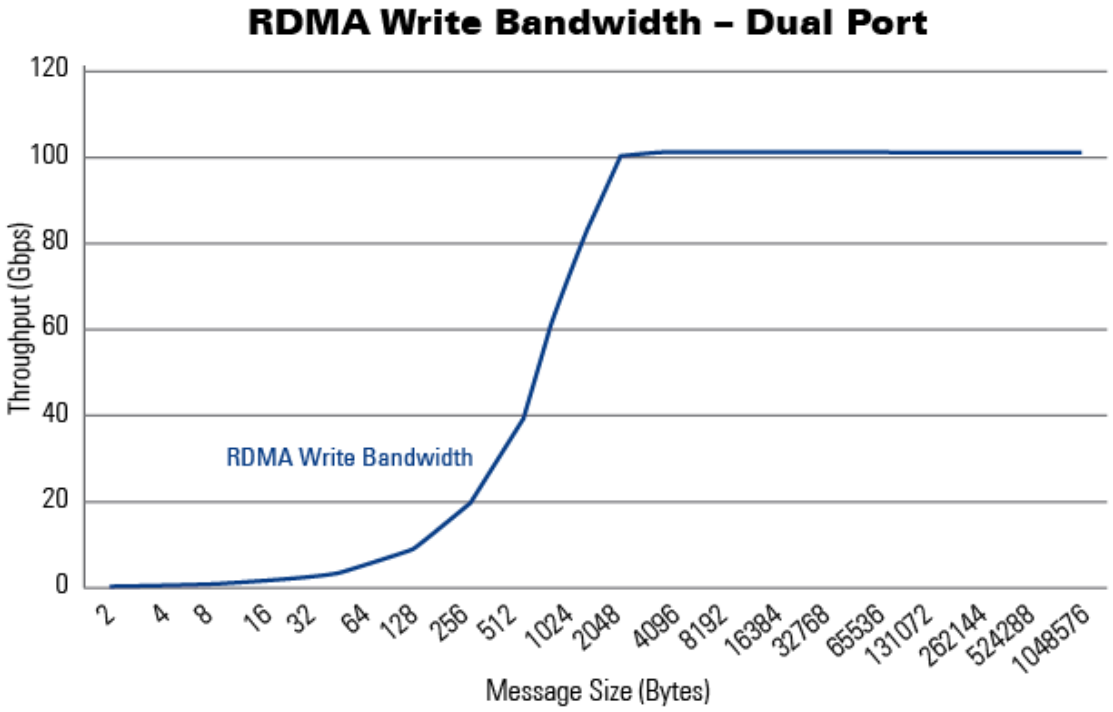

Figure 1: 100Gb/s RDMA Write Performance

Many HPC applications are based on communications patterns that use many small messages between parallel processes within the job. It is critical that the interconnect used to transport these messages provides low latency and high message rate capabilities to assure that there are no bottlenecks to the application. The new Connect-IB architecture provides an increase in the message rate of previous InfiniBand offerings by over 4 times. Connect-IB can deliver over 137 million single-packet (non-coalesced) native InfiniBand messages per second to the network. This increase assures that there are no message rate limitations for applications and that the multiple cores on the server will communicate to other machines as fast as the cores are capable, without any slowdown from the network interface.

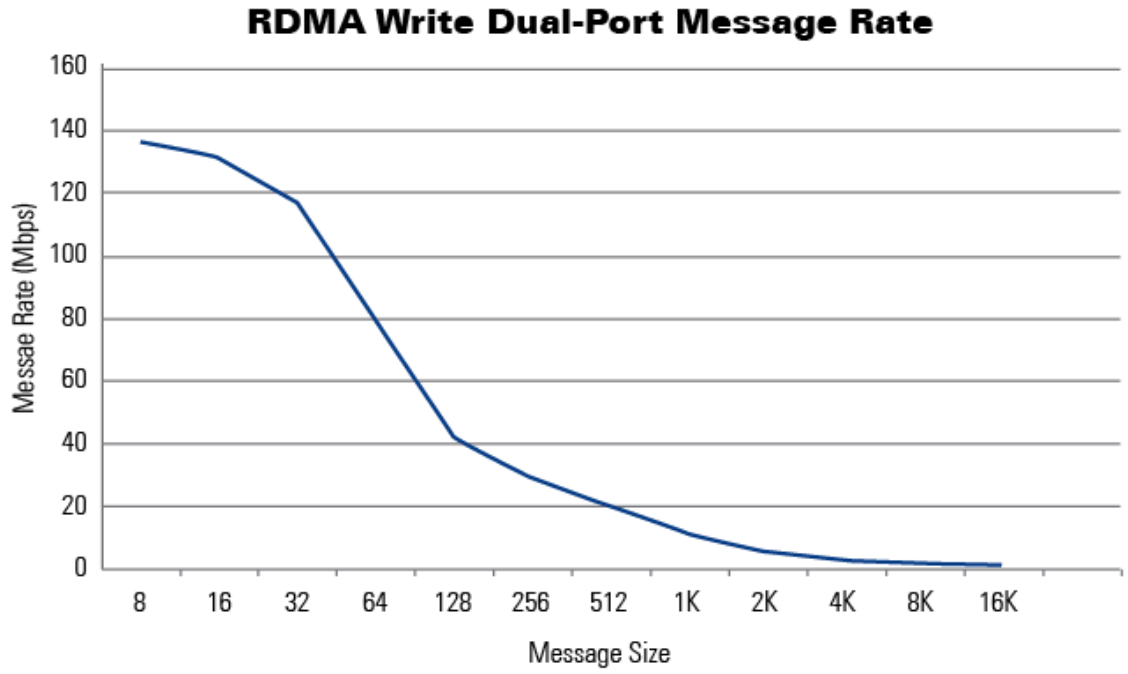

Figure 2: Message Rate Example

Impact of Interconnect on LS-DYNA Cluster Performance

The cluster interconnect is very critical for efficiency and performance of the application in the multi-core era. When more CPU cores are present, the overall cluster productivity increases only by the presence of a high-speed interconnect.

We have compared the elapsed time with LS-DYNA using 1Gb/s Ethernet, 10Gb/s Ethernet, 40Gb/s Ethernet, and 56Gb/s FDR InfiniBand. This study was conducted at the HPC Advisory Council Cluster Center (www.hpcadvisorycouncil.com) on an Intel Cluster Ready certified cluster comprised of Dell™ PowerEdge™ R720xd/R720 32-node cluster with 1 head node, each node with Dual Socket Intel® Xeon® 10-core CPUs E5-2680 v2 at 2.80 GHz, Mellanox Connect-IB 56Gb/s FDR InfiniBand adapter, and with 64GB of 1600MHz DDR3 memory. The nodes were connected into a network using a Mellanox SwitchX® SX6036 36-Port VPI switch which supports 40Gb/s Ethernet and 56Gb/s FDR InfiniBand. The Operating System used was RHEL6.2, the InfiniBand driver version was OFED 2.1-1.0.0, and the File System is shared over NFS from the Dell PowerEdge R720xd head node, which provides 24 250GB 7.2K RPM SATA 2.5" hard drives over RAID 0. The MPI library used was Platform MPI 9.1, the LS-DYNA version was LS-DYNA MPP971_s_R3.2.1, and the benchmark workload was the Three Vehicle Collision test simulation, as shown in Figure 3.

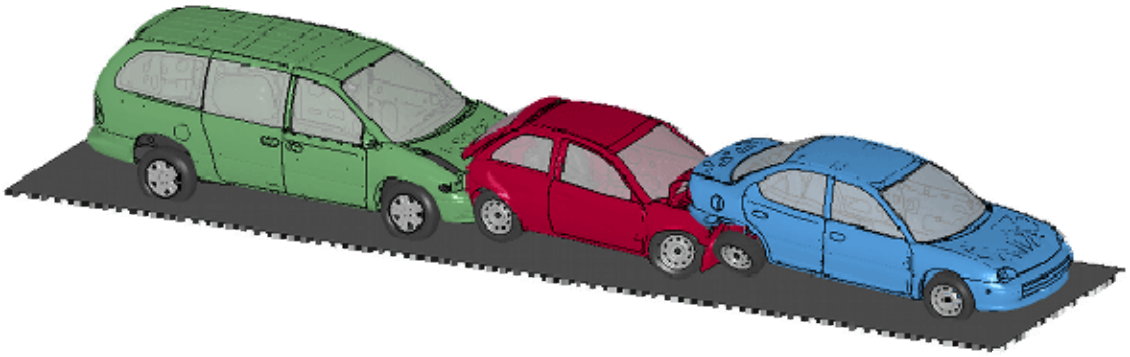

Figure 3: 3 Vehicle Collisions

Figure 4 below shows the elapsed time for the InfiniBand and Ethernet interconnects for a range of core/node counts for the Three Vehicle Collision case.

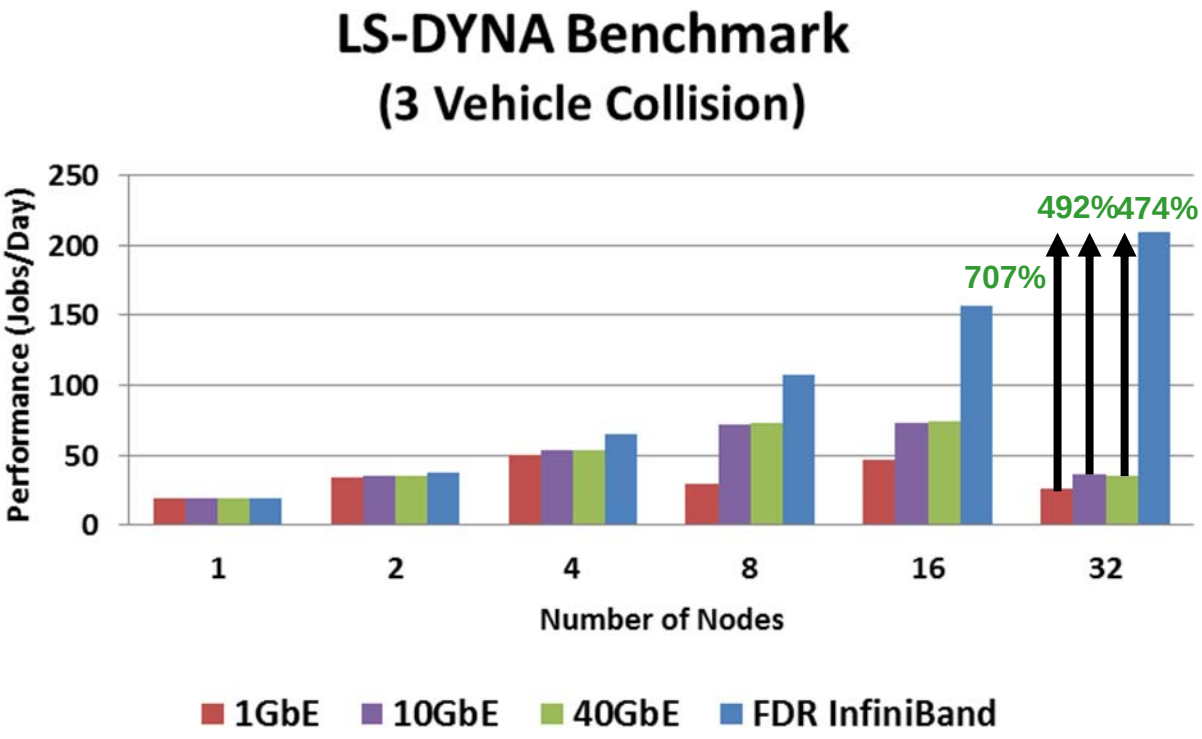

Figure 4: Interconnect Comparison with Three Vehicle Collision

FDR InfiniBand delivered superior scalability in application performance, resulting in faster run time, providing the ability to run more jobs per day. The 56Gb/s FDR InfiniBand-based simulation performance (measured by number of jobs per day) was 707% higher than 1GbE, over 492% higher than 10GbE and over 474% higher than 40GbE on a LS-DYNA simulation

which runs on 32 nodes or 640 MPI processes. While Gigabit Ethernet showed a loss of performance (increase in run time) beyond 4 nodes, FDR InfiniBand demonstrated good scalability throughout the various tested configurations. Because LS-DYNA uses MPI for the interface between the application and the networking layer, it requires scalable and efficient send-receive semantics as well as good scalable collectives operations. While InfiniBand provides an effective method for those operations, the Ethernet TCP stack leads to CPU overheads which translate into higher network latency, reducing the cluster efficiency and scalability.

Accelerating Performance by Hardware Offloads in HPC-X™

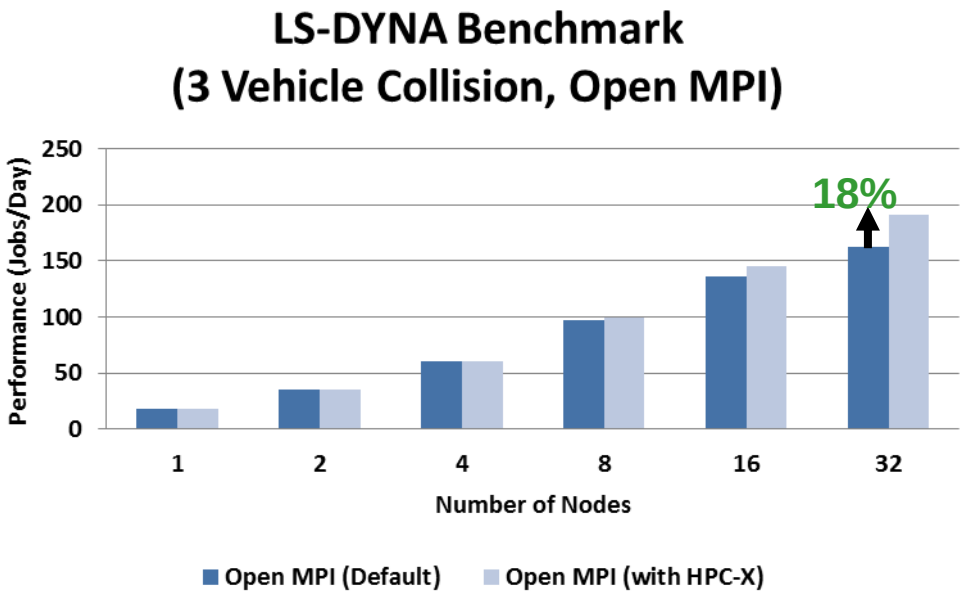

Figure 5: LS-DYNA 3 Vehicle Collision Performance per acceleration software in HPC-X

HPC-X™ is a supported scalable HPC tool kit from Mellanox. It contains a fully supported MPI which is based on OpenMPI as well as other acceleration components that unlock the performance capabilities of the underlying interconnect hardware. It is possible to improve LS-DYNA performance in MPI by deploying acceleration software which offloads the MPI collective communications onto the networking hardware. When placed side-by-side using the same set of systems and additionally by using the software acceleration from the Mellanox Fabric Collective Accelerator (FCA) and Mellanox Messaging Accelerator (MXM) that are available in the HPC-X offering, it clearly demonstrates a significant improvement over the default case which uses the CPU to process the network communication. Connect-IB is unique in that it utilizes improved memory resource management and a more efficient transport service, allowing the HPC cluster to run at its highest scalability.

FDR InfiniBand showed an 18% improvement in performance at 32 nodes over the default case where additional hardware offload capabilities were not used. -Also, the margin for additional performance improvement is expected to be wider as more nodes are involved, as the effects of FCA and MXM allow for even greater scalability of the application.

The effect of the Interconnect in LS-DYNA Performance

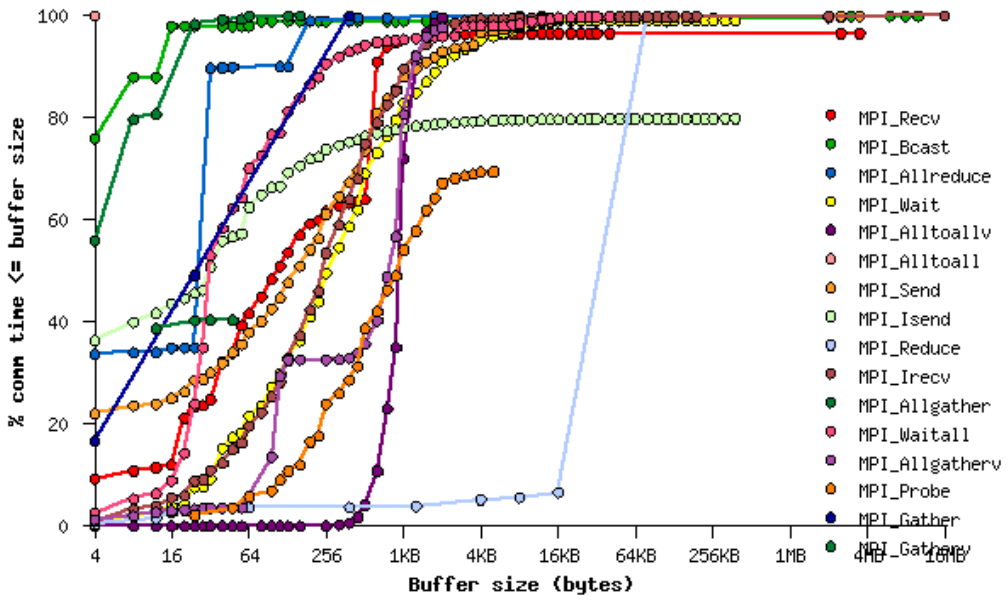

Figure 6: MPI Profiling on Three Vehicle Collision at 640 MPI processes

Figure 6 is the MPI profiling for the 3 Vehicle Collision run at 640 MPI processes. IPM, or Integrated Performance Monitoring, is the MPI profiler used to generate this MPI communication profile. This tool is a portable profiling infrastructure for parallel codes. It provides a low-overhead profile of the performance aspects and resource utilization in a parallel program. IPM is also a support component of the HPC-X scalable toolkit from Mellanox.

By inspecting the output from the MPI profiler, it is observed that the majority of the MPI messaging appears to be concentrated on the small message buffer sizes. For the range of the messages that the Connect-IB HCA message rates appears to perform. We can then understand that the underlying communication patterns that make LS-DYNA scalable on the low latency interconnect like InfiniBand. We see that most of the MPI messages are appeared in the small message sizes under 256 bytes. For the most time consuming MPI calls, MPI_Recv messages are concentrated 4KB. Majority of the MPI_Bcast messages are in buffers that are less than 16B. For MPI_Allreduce, mst messages are less than 256B.

Maximizing Cluster Utilization with 100Gb/s InfiniBand

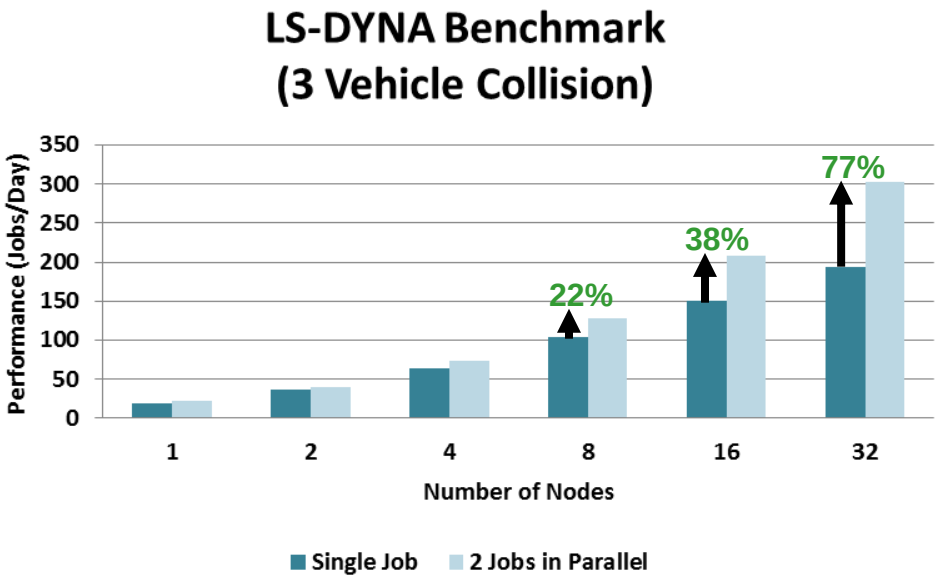

Figure 7: Cluster productivity between utilizing compute nodes for a single job versus 2 jobs in parallel

In the traditional HPC environment, it is a common practice to utilize all of the CPU cores and hardware resources available on the compute nodes to run exclusively for a single workload. By having dedicated compute nodes for a particular parallel job, the underlying hardware would not run into resource contention or interference with other processes, it is regarded as a good approach for achieving highest performance for a parallel job.

While running a single parallel job exclusively is perceived to guarantee that the hardware resources are capable of running at running at its maximum capacity, it is observed that running two jobs in parallel would actually yield higher productivity than a single job.

To experiment with running two jobs in parallel, one will need to ensure that resources are partitioned so that contention and conflict to the hardware resources can be avoided. One way is to explicitly bind the process by specifying which CPU cores that the MPI job is scheduled to run. For instance, on a compute node with two CPU sockets, one would explicitly specify the set of CPU cores to use for each of the parallel jobs. This would prevent the likelihood for the MPI runtime to oversubscribe the CPU cores which will lead to degraded performance. Similarly, by specifying each of the parallel jobs to run on its dedicated port on the Connect-IB InfiniBand adapter, it guarantees the full FDR link to associate for each of the parallel job. As Connect-IB HCA provides 2 FDR links to the network, therefore each link provides a true 56Gbps FDR InfiniBand connection to the network and provides the needed throughput for each of the parallel job.

By introducing the idea of running a second job to run in parallel, it is demonstrated that up to 77% of higher job productivity can be gained compared to running a single parallel job on 32 compute nodes.

Performance Improvements by System Architecture

Because Connect-IB supports the PCIe Gen3 standard, it is perfectly suited to run on either the Intel Xeon E5-2600 v2 Series (Ivy Bridge) or E5-2600 Series (Sandy Bridge) based-platforms. Connect-IB allows the MPI applications to takes advantage of the increases in CPU core processing, memory bandwidth by providing the necessary network throughput. Notably, the Ivy Bridge architecture enables the Connect-IB InfiniBand device to run at its maximum throughput and lowest latency. The results are shown in Figure 8.

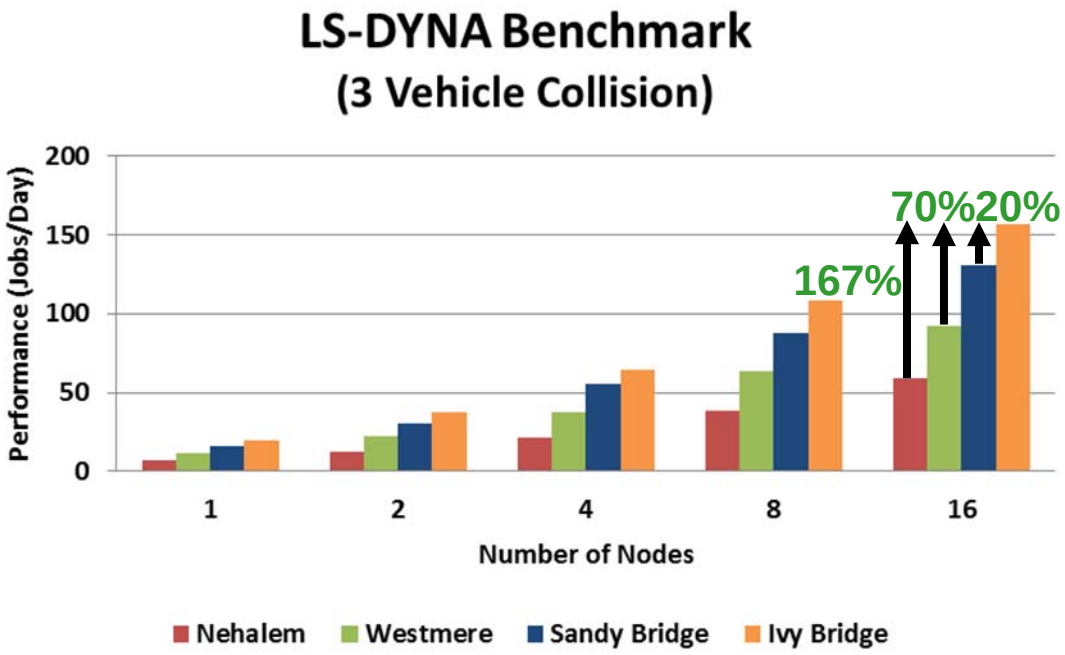

Figure 8: LS-DYNA 3 Vehicle Collision Performance per CPU technology

Compared to previous system generations, the Intel Xeon E5-2680 v2 (Ivy Bridge) cluster outperforms the Intel Xeon E5-2680 (Sandy Bridge) cluster by up to 20%, it also outperforms Intel Xeon X5670 (Westmere) cluster by up to 70%, and provided gains up to 167% higher performance over the older Intel Xeon X5570 (Nehalem) cluster.

To conduct the performance comparison tests, the following system configurations were used:

- Each Nehalem system consisted of the Dell PowerEdge m610 system with a dual-socket Intel Xeon X5570 running at 2.93GHz, 1333MHz DIMMs, and Mellanox ConnectX-2 QDR InfiniBand.
- Each Westmere system used the same Dell PowerEdge m610 system, with a dual-socket Intel Xeon X5670 running at 2.93GHz, 1333MHz DIMMs, and Mellanox ConnectX-2 QDR InfiniBand.
- Each Sandy Bridge system used the aforementioned Dell PowerEdge R720xd, each with a dual-socket Intel Xeon E5-2680 running at 2.7GHz, 1600MHz DIMMs, and Mellanox Connect-IB FDR InfiniBand.

- Each Ivy Bridge system used the aforementioned Dell PowerEdge R720xd, each with a dual-socket Intel Xeon E5-2680 v2 running at 2.8GHz, 1600MHz DIMMs, and Mellanox Connect-IB FDR InfiniBand.

TopCrunch Results

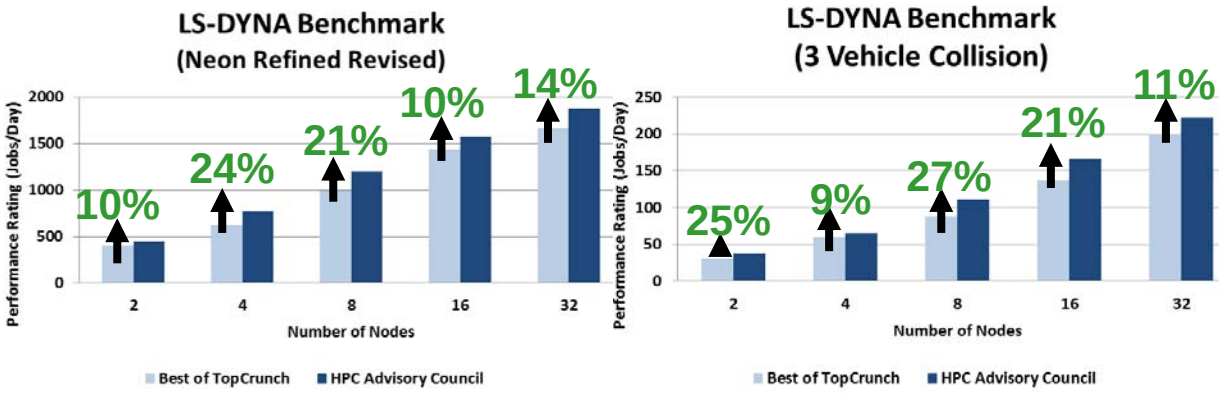

Figure 9: TopCrunch Results on Neon Refined Revised and 3 Vehicle Collision

Figure 9 shows the performance benchmarking effort conducted by the HPC Advisory Council which demonstrates with the latest improvement in the computer architecture which improves the communication throughput, MPI tuning, which performs better than the previously best published results.

The effects of the InfiniBand performance advancement in both acceleration software and hardware are demonstrated in TopCrunch results. With the latest CPU and InfiniBand networking technologies available in the Dell PowerEdge R720xd compute nodes, the HPC Advisory Council was able to establish new records on the LS-DYNA performance results on per node basis.

The results achieved by the HPC Advisory Council demonstrates from 9% to 27% of higher performance than best published results on TopCrunch (February 2014) comparing to all other platforms listed on TopCrunch. The HPC Advisory Council results are world best for cases on 2, 4, 8, 16, and 32 nodes for the Neon Refined Revised and 3 Vehicle Collision cases. The Connect-IB architecture enables LS-DYNA to achieve superior scalability performance at scale.

Conclusions

From concept to engineering and from design to test and manufacturing, engineering organizations rely on powerful virtual development solutions. Finite Element Analysis (FEA) and Computational Fluid Dynamics (CFD) are used in an effort to secure quality and speed up the development process. Cluster solutions maximize the total value of ownership for FEA and CFD environments and extend innovation in virtual product development.

HPC cluster environments impose high demands for connectivity throughput, low-latency, low CPU overhead, network flexibility and high-efficiency in order to maintain a balanced system in order to achieve high application performance and scaling. Low-performance interconnect solutions, or lack of interconnect hardware capabilities will result in application performance and result in an extended time-to-market process.

Livermore Software Technology Corporation (LSTC) LS-DYNA software was investigated. In all InfiniBand-based cases, LS-DYNA demonstrated high parallelism and scalability, which enabled it to take full advantage of multi-core HPC clusters. Moreover, according to the results, a lower-speed interconnect, such as 1, 10 or 40Gb/s Ethernet, is ineffective on mid to large cluster size, and can cause a dramatic reduction in performance beyond 8 server nodes (that is, the application run time actually gets slower).

We have compared the performance levels of various adapter throughputs on different network architectures to measure the effect on LS-DYNA software. The evidence has shown that the inherent advantages offered by the Connect-IB 56Gb/s FDR InfiniBand adapter – namely, the unparalleled message rate and support for the PCI Gen3 standard – offer increased bandwidth, lower latency, and greater scalability than when using 40GbE, 10GbE, or 1 GbE Ethernet interconnects. This has decreased LS-DYNA's run time, enabling LS-DYNA to run significantly more jobs per day than ever before. With the additional improvements in the system architecture and networking technologies such as MPI collective offloads, it enables LS-DYNA to achieve superior performance at scale and establish new records on the TopCrunch benchmarks.