LS-DYNA Productivity and Power-aware Simulations in Cluster Environments

Gilad Shainer[1], Tong Liu[1], Jacob Liberman[2], Jeff Layton[2], Onur Celebioglu[2],

Scot A. Schultz[3], Joshua Mora[3], David Cownie[3], Ron Van Holst[4]

[1]Mellanox Technologies [2]Dell, Inc. [3]Advanced Micro Devices (AMD) [4]Platform Computing

Summary

From concept to engineering and from design to test and manufacturing; engineering relies 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.

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

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 GigE or 10 Gigabit Ethernet are ineffective on mid to large cluster size, and can cause a reduction in performance beyond 16 or 20 server nodes (i.e. the application run time actually gets slower)

We have profiled the communications over the network of LS-DYNA software to determine LS-DYNA sensitivity points, which is essential in order to estimate the influence of the various cluster components, both hardware and software. We evidenced the large number of network latency sensitive small messages through MPI_AllReduce and MPI_Bcast operations that dominate the performance of the application on mid to large cluster size. The results indicated also that large data messages are used and the amount of the data sent via the large message sizes increased with cluster size. From those results we have concluded that the combination of a very high-bandwidth and extremely low-latency interconnect, with low CPU overhead, is required to increase the productivity at mid to large node count.

We have also investigated the increase of productivity from single job to multiple jobs in parallel across the cluster. The increase of productivity is based on two facts. First, good scalability of AMD architecture that allows to run multiple jobs on a given compute node without saturating the memory controller. Second, the low latency and high bandwidth available on the InfiniBand interconnect that allowed us to offload the CPU to CPU data traffic from MPI communications via the interconnect instead of in compute node. The net result of that practice is an increase of the productivity by a factor of 200% with respect to the single job run.

Finally, the increase of productivity on single job runs with high speed interconnects has been analyzed from the point of view of power consumption leading to a 60% reduction or energy savings when using InfiniBand with respect to Ethernet.

Keywords:

Performance, Productivity, Scalability, Power-aware simulations

1 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 accelerated speed 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. Total cost of real vehicle crash-tests in order to determine its safety characteristics is in the range of $250,000 or more. On the other hand, the cost of a high-performance compute cluster can be just a fraction of the price of a single crash test, while providing a system that can be used for every test simulation going forward.

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 the needed flexibility, scalability and efficiency for such simulations

Cluster productivity, sometimes not measured by just how fast an application runs, is the most important factor for cluster hardware and software configuration. Achieving the maximum number of jobs executed per day is of higher importance than the wall clock time of a single job. Maximizing productivity in today's cluster platforms requires using enhanced messaging techniques even on a single server platform. These techniques also help with parallel simulations by using efficient cluster interconnects.

Power-aware simulations are an emerging challenge in clustering simulations. Power consumption has become a significant share of the total cost over the typical three year life span of a cluster. Approaches such as consolidating the cluster components and using power efficient components can help reduce the overall system cost but they needs to be implemented judiciously to maintain productivity growth and future proof the cluster design.

2 Architecture of multi-core HPC clusters for LS-DYNA

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 (i.e. multi-core, multi-processor based HPC servers with high-speed interconnects) has a huge influence on the overall application performance and productivity. To meet the demand of more powerful HPC servers more execution cores (e.g. dual, quad-core) are being integrated into each processor and more processors are being tightly connected. For example, 2, 4 and 8 processors are connected through a packet-based, high-bandwidth, scalable, low-latency point-to-point technology that links processors to each other, processors to coprocessors and processors to I/O and peripheral controllers. AMD's HyperTransport™ technology is a prime example of this technology. There are important challenges in this strategy (e.g. larger scale integration, reduction of voltages and core frequencies) to keep the power consumption as low as possible while increasing the computational capabilities of the HPC servers.

In addition to the processor, the cluster interconnect is a very critical component for delivering efficiency and scalability The networking requirements of each CPU core must be adequately addressed without imposing additional networking overhead so that the application can scale. In a multi-core multi-socket HPC server-based cluster, the driving factors of performance and scalability for LS-DYNA have shifted from the core's clock frequency and cache size to the memory and

interconnect throughput per core. The memory bottleneck can be addressed by using interconnects that support Direct Memory Access (DMA), Remote DMA and zero-copy transactions.

3 Dell HPC Clustering

Over the past decade, commodity clusters have largely replaced proprietary supercomputers for high performance computing applications. According to the Nov 2008 Top500 list of the world's fastest supercomputers, cluster architectures are used in more than 50% of the top 100 systems. This is primarily due to the highly competitive price for performance they can achieve. Dell designs, integrates, and tests HPC clusters built from industry-standard servers, leading open source and commercial software, and high speed interconnects. These clusters combine the performance of proprietary systems with the simplicity, value, and flexibility of standards based hardware.

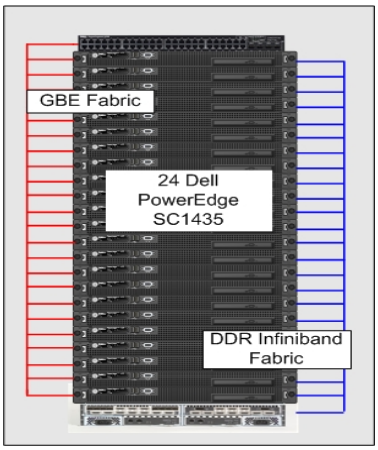

Figure 1 – cluster topology

This study was conducted on a test cluster comprised of 24 Dell PowerEdge SC1435 servers. The SC1435 is an AMD-based 1U 2-socket server that can support up to 32GB of DDR-2 memory in 8 DIMM sockets. It has one 8x PCI-Express expansion slot and two integrated Gigabit Ethernet Network Interface Cards. When used in conjunction with Mellanox InfiniBand HCAs and 3rd generation Quad-Core AMD Opteron™ processors, the SC1435 provides an ideal building block for HPC clusters due to the rack density, energy efficiency, and price/performance it can deliver.

For this study, each cluster server was equipped with a Mellanox ConnectX® 20Gb/s InfiniBand HCA for inter-node communication. The servers were also deployed and configured via a Gigabit Ethernet management fabric. The cluster topology is depicted in Figure 1.

4 Quad-Core AMD Opteron™ processors and tools for HPC servers

The HPC server configuration utilized in the LS-DYNA performance study is based on the latest available AMD processor architecture. AMD's third generation AMD Opteron™ processors with Direct Connect Architecture are designed for advanced scaling linearity in systems with up to 32 cores (i.e. 8 Quad-Core processors). AMD's Direct Connect Architecture helps eliminate the bottlenecks inherent in a front-side bus by directly connecting the processors, the memory controller, and the I/O to the central processor unit to enable improved overall system performance and efficiency in applications such as LS-DYNA.

AMD technology offers an integrated DDR2 DRAM Memory Controller with AMD Memory Optimizer Technology which allows for lower cost High-bandwidth, energy-efficient DDR2 memory. With third-generation Quad-Core AMD Opteron processors, the instruction fetch bandwidth, data cache bandwidth, and memory controller to cache bandwidth have all been doubled over the previous generation technology to help keep the 128-bit floating-point pipeline full.

Deployment of clusters can become both an energy consumption and cost challenge. With AMD's enhanced AMD PowerNow!TM and AMD CoolCoreTM Technologies, today's clusters can deliver performance on demand while minimizing power consumption. Second generation AMD Opteron based platforms can be upgraded to AMD's third generation processors in the same thermal envelope, allowing for increased computational capacity without altering datacenter power and cooling infrastructures.

AMD also offers additional tools such as the AMD Core Math library (ACML), AMD Performance Primitives (APP), and AMD CodeAnalyst for extensive profiling single and multithreaded applications. AMD software tools can be downloaded for free from http://developer.amd.com.

5 InfiniBand high-speed interconnect technology

Choosing the right interconnect technology is essential for maximizing HPC system efficiency. Slow interconnects delay data transfers between servers, causing poor utilization of the compute resources and slow execution of simulations. An interconnect that requires CPU cycles as part of the networking process will decrease the compute resources available to the application and therefore will slow down and limit the number of simulations that can be executed on a given cluster. Furthermore, unnecessary overhead on the CPU increases the system jitter which in return limits the cluster's scalability.

Interconnect flexibility is another requirement for multi-core systems. As various cores can perform different tasks, it is necessary to provide Remote Direct Memory Access (RDMA) along with the traditional semantics of Send/Receive. RDMA and Send/Receive in the same network provides the user with a variety of tools that are crucial for achieving the best application performance and the ability to utilize the same network for multiple tasks. Moreover, in multi-core multi-processor environments, it is essential to have an interconnect that provides the same low-latency for each processor/core (zero scalable latency), regardless of the number of cores and processors that operate simultaneously, in order to guarantee linear application scalability.

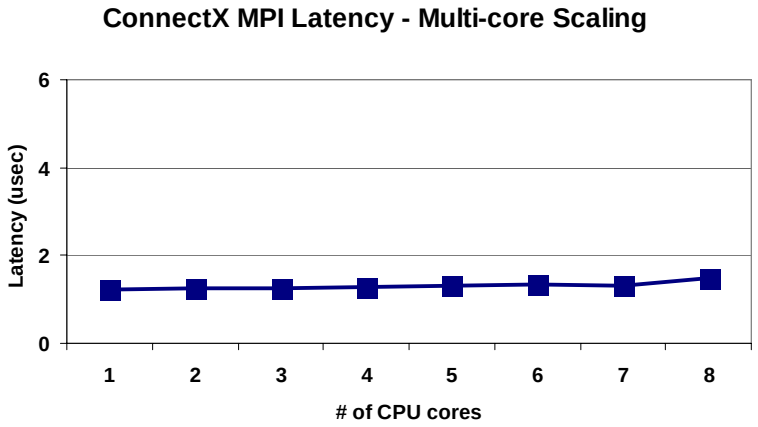

Figure 2 – MPI multi-core latency with Mellanox ConnectX InfiniBand HCA

By providing low-latency, high-bandwidth and extremely low CPU overhead, InfiniBand has become the most deployed high-speed interconnect, 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 ten-thousand nodes and multiple CPU cores per server platform and efficient utilization of compute processing resources. Mellanox ConnectX InfiniBand adapters and InfiniScale® IV-based switches are the leading-edge InfiniBand solutions that have been designed for HPC clustering technology. ConnectX and InfiniScale IV deliver up to 40Gb/s of bandwidth between servers and up to 120Gb/s between switches. This high-performance bandwidth is matched with ultra-low application latency of 1µsec, and switch latencies under 100ns that enable efficient, scale-out compute systems. For multi-core systems, it is essential to provide zero scalable latency, which means to provide the same low-latency regardless of the number of processes running between cluster nodes. Figure 2 shows the MPI multi-core latency benchmark results between two 8-core servers using Mellanox ConnectX InfiniBand adapters. The benchmark measured the latency of eight different cases – from a single process running between the two systems using a single core per system, up to eight parallel processes running between the two systems using all the available cores. According to the results, ConnectX adapters enable zero scalable latency that guarantees the same low latency for each of the CPU cores, regardless on how many cores communicate at the same time. Moreover, InfiniBand was designed to be fully offloaded, meaning all the communications are being handled within the interconnect without involving the CPU. This further enables the ability to scale up with linear performance by reducing the system jitter and enabling efficient synchronizations between the execution cores.

6 LS-DYNA performance scalability and profiling analysis

We selected two standard benchmark cases provided by LSTC - Three Vehicle Collision and Neon-Refined Revised Crash Test simulation. Using the benchmarks cases, we performed the following analysis:

- Compared the performance of different interconnect technologies, namely 20Gb/s InfiniBand, 10 Gigabit Ethernet and Gigabit Ethernet
- Measured LS-DYNA scalability at increasing core counts
- Identified methods for increasing productivity through job placement
- Compared different MPI libraries
- Compared between Opteron CPUs versions
- Measured the power consumed per job

The performance and profiling analysis was carried out as part of the HPC Advisory Council (http://hpcadvisorycouncil.mellanox.com) research activities, using the HPC Advisory Council Cluster Center. The cluster configuration is summarized in table 1 and 2 below.

Table 1. Cluster 1 benchmark configuration

Application	LS-DYNA MPP971
Servers	24 Dell PowerEdge SC1435 servers
Processors	2 Quad-Core AMD Opteron™ 2358 processors (“Barcelona”) at 2.4 GHz per node
Memory	8 x 2 GB, 667 MHz Registered DDR-2 DIMMs per node
OS	Red Hat® Enterprise Linux® 5 Update 1 OS
Message Passing Interface (MPI)	• HP MPI 2.2.7 • Platform MPI 5.6.5
Interconnect	Mellanox MT25408 ConnectX DDR InfiniBand OpenFabrics Enterprise Distribution (OFED) 1.3 software stack

Table 2. Cluster 2 benchmark configuration

Application	LS-DYNA MPP971
Servers	24 Dell PowerEdge SC1435 servers
Processors	2 Quad-Core AMD Opteron™ 2382 processors (“Shanghai”) at 2.6Ghz per node
Memory	8 x 2 GB, 800 MHz Registered DDR-2 DIMMs per node
OS	Red Hat® Enterprise Linux® 5 Update 1 OS
Message Passing Interface (MPI)	• HP MPI 2.2.7 • Platform MPI 5.6.5
Interconnect	Mellanox MT25408 ConnectX DDR InfiniBand OpenFabrics Enterprise Distribution (OFED) 1.3 software stack

7 AMD Opteron CPU Comparison

The two available clusters for our performance differ in the CPU speed and the memory (which is being driven by the CPU capabilities). To asses the performance advantage that AMD Opteron “Shanghai” provides versus the previous Opteron generation (“Barcelona”) we have compared the elapsed time for both 3 Vehicle Collision and Neon-Refined Revised Crash benchmarks using InfiniBand interconnect. The results are shown in figures 3 and 4. Using AMD Opteron “Shanghai” processors enables us to achieve an average 20% reduction in run time, or 20% increase in the number of LS-DYNA jobs one can run per day.

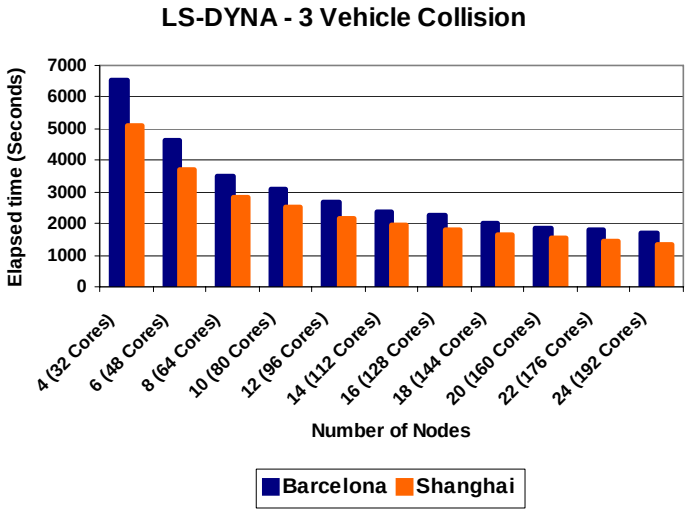

Figure 3 – AMD Opteron CPU comparison with 3 Vehicle Collision

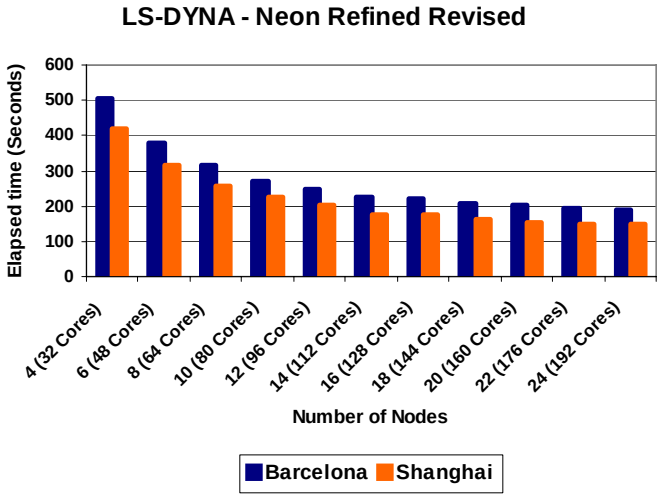

Figure 4 – AMD Opteron CPU comparison with Neon Refined Revised

8 The Importance of the cluster interconnect

The cluster interconnect is very critical for efficiency and performance of the application in multi-core platforms. When more CPU cores are present, the overall cluster productivity increases. The performance results from this section until the end of this paper were performed on cluster 2. We have compared the elapsed time with both LS-DYNA benchmarks using 20Gb/s InfiniBand, 10 Gigabit Ethernet and Gigabit Ethernet. Figure 5 below shows the elapsed time for these interconnects for a range of core/node counts for the Neon Refined Revised case. Figure 6 is the same basic chart but for the 3 vehicle collision case.

In both cases, InfiniBand delivered superior scalability in performance, resulting in faster run time, providing the ability to run more jobs per day. The 20Gb/s InfiniBand based simulations run-time was reduced by up to 35% compared to 10GigE and 50% compared to GigE with 16 nodes and was reduced by up to 60% compared to 10GigE and 61% compared to GigE with 24 nodes. Plus, with InfiniBand, LS-DYNA showed good scalability up to 24 nodes, while Ethernet (both 10G and 1G)

showed no scalability beyond 16 nodes. In addition, GigE and 10GigE results showed a loss of performance (increase in run time) at higher core counts (i.e. the performance actually decreased when the number of cores was increased). For 3 vehicle collision and on 24 node count, the InfiniBand based simulation run time was reduced by up to 40% compared to 10GigE and 55% compared to GigE. LS-DYNA uses MPI for the interface between the application and the networking layer, and as such, requires scalable and efficient send-receive semantics, as well as good scalable collective operations. While InfiniBand provides an effective way for those operations, the Ethernet TCP stack which leads to CPU overheads that translate to higher network latency, reduces the cluster efficiency and scalability.

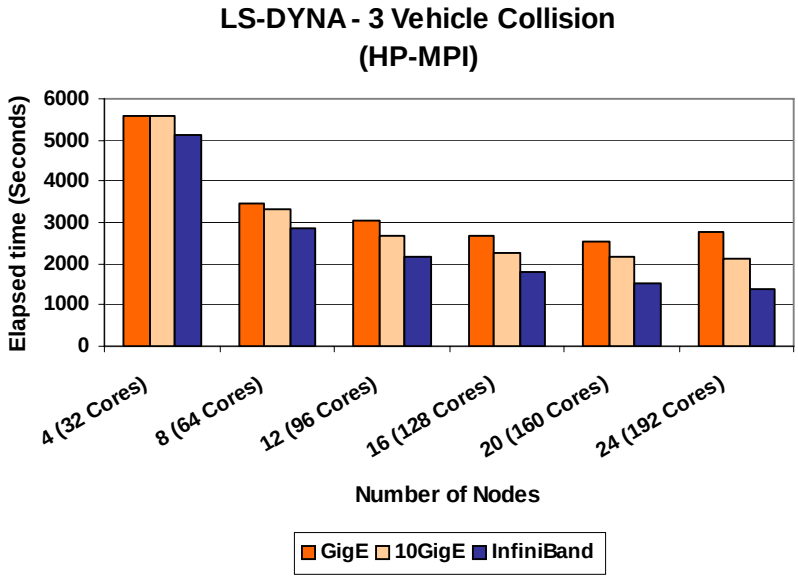

Figure 5 – Interconnect comparison with Neon Refined Revised

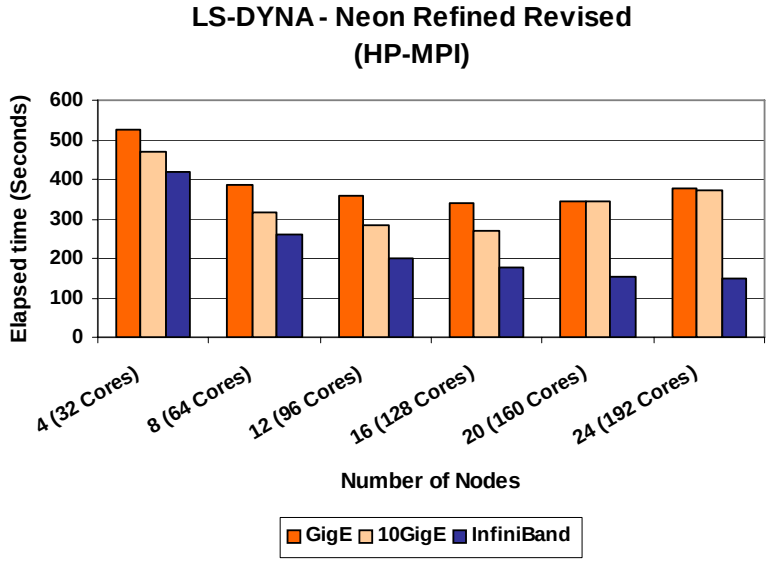

Figure 6 – Interconnect comparison with 3 Vehicle Collision

9 Utilizing job placement for higher productivity

We define productivity as the measure of the number of application jobs that can be achieved in a given time, usually one day. A larger number of jobs per day equals higher productivity. With the

increased complexity of high-performance applications, a single job consuming all the cluster resources might create bottlenecks within the CPU to CPU or CPU to memory communications. We have shown near-linear scaling when using the InfiniBand interconnect between nodes. Therefore, the question becomes how much scalability will we see if a single job is limited to running within each cluster node? These bottlenecks and productivity capabilities can be explored by comparing a single job run on the entire cluster versus several jobs run concurrently (i.e. in parallel) using job placement to a specific CPU socket.

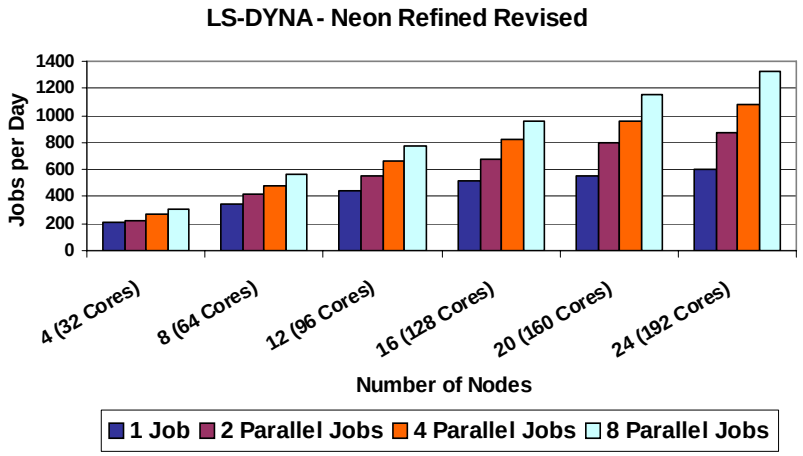

Figure 7 – Neon Refined Revised productivity

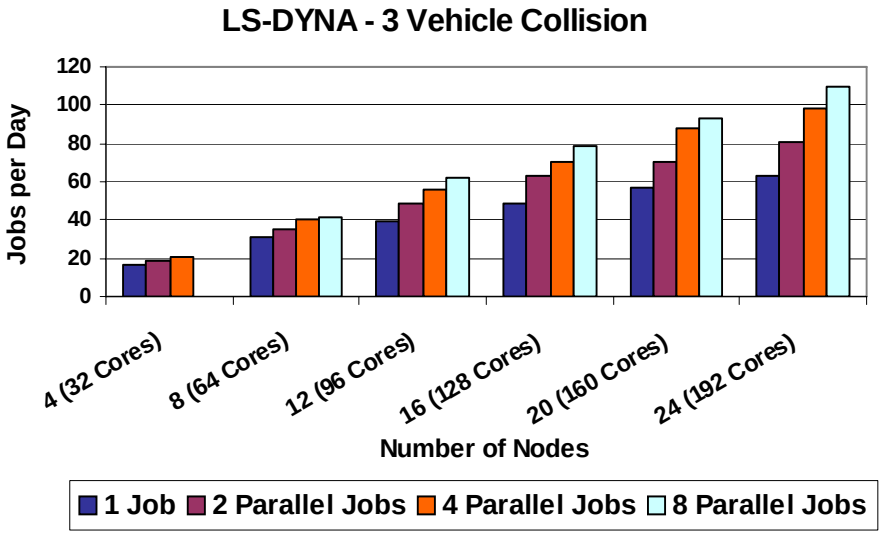

Figure 8 – 3 Vehicle Collision productivity

Figure 7 and 8 show the productivity results (using cluster 2) comparing a single job run to two, four and eight parallel jobs. In the case of the two parallel jobs, each job uses a single socket per server, on the entire cluster, therefore half of the system CPU cores. In the case of four parallel jobs, each uses two cores in a socket, and in the case of eight parallel jobs, each is using a single core of each node across the entire cluster. In the parallel test, the CPU to CPU communication in node has been reduced and the CPU to CPU communication over network has been increased. The net result of this strategy leads to increases of productivity up to a factor of two on eight parallel jobs with respect to the single job on InfiniBand interconnect. The productivity increase over the single InfiniBand port is

enabled by offloading the data traffic among CPUs to the interconnect, and leveraging the interconnect's high throughput capabilities.

10 LS-DYNA profiling

Profiling the application is essential for the understanding of application's performance dependency on the various cluster subsystems. In particular, application's communication profiling can help in choosing the most efficient interconnect and MPI library, and in identifying the critical communication sensitivity points that greatly influence on the application's performance, scalability and productivity.

LS-DYNA 3 Vehicle Collision profiling data is presented in Figure 9 which shows the number of messages that were sent between the cluster nodes in a logarithmic scale. Most of the messages are in the range of 0-4KBytes. The 0-256B area represents mostly the synchronization or control messages and the 256B-4KB the compute messages. The performance of messages in that range will have the greatest impact on LS-DYNA overall performance and is therefore considered the sensitivity area.

We have also noticed that the number of messages increases with the core/node count. This indicates the need for lowest latency interconnect to avoid creating a bottleneck for the fast synchronizations happening in MPI operations.

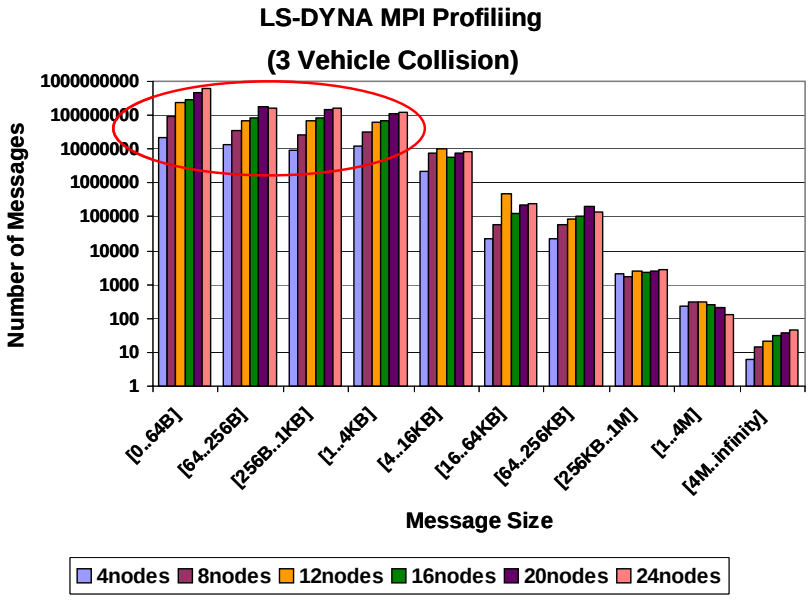

Figure 9 – 3 Vehicle Collision network profiling

MPI operations that arise in the application and that are network latency sensitive are MPI_AllReduce and MPI_Bcast. These two collective operations account for the majority of MPI communication overhead.

11 MPI libraries comparison

We have identified three critical sensitivity points in LS-DYNA – latency for 0-256B messages, throughput for 256B-4KB messages and scalability and latency for MPI_AllReduce and MPI_Bcast.

We have compared the performance of two MPI libraries, HP MPI and Platform MPI. First, we have looked at the MPI collective performance, and noticed that each MPI shows an advantage with a different MPI collective, as shown in Figures 10 and 11.

Figure 12 shows the communication overhead of MPI_AllReduce and MPI_Bcast with respect to the total run time at different node count on 3 Vehicle Collision benchmark. MPI_AllReduce has bigger influence on the application performance. Notice the change of protocol from 32 to 64 cores.

In figure 13, we have compared the 3 Vehicle Collision benchmark performance between HP MPI and Platform MPI. We believe the measured performance differences of the application are related with the MPI_AllReduce performance differences among MPI implementations.

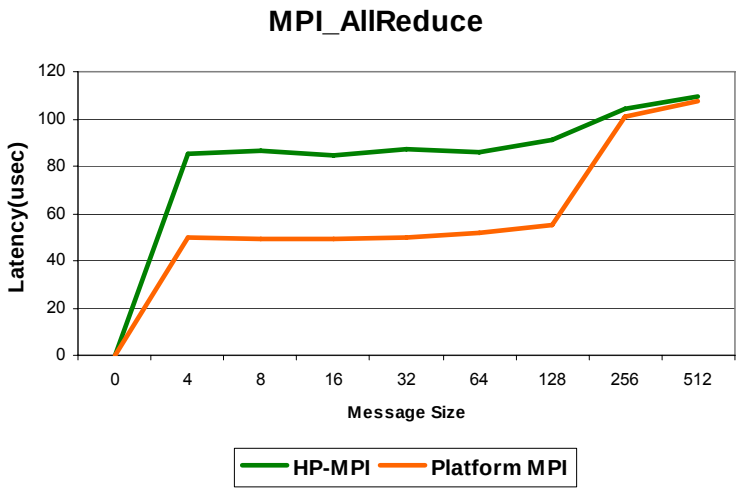

Figure 10 – MPI_AllReduce performance comparison

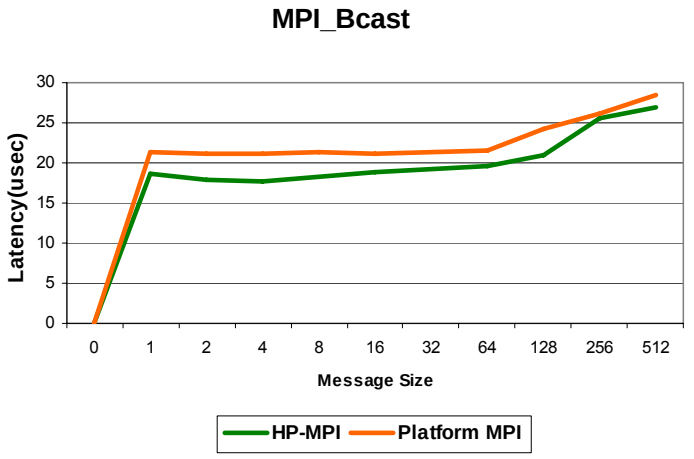

Figure 11 – MPI_Bcast performance comparison

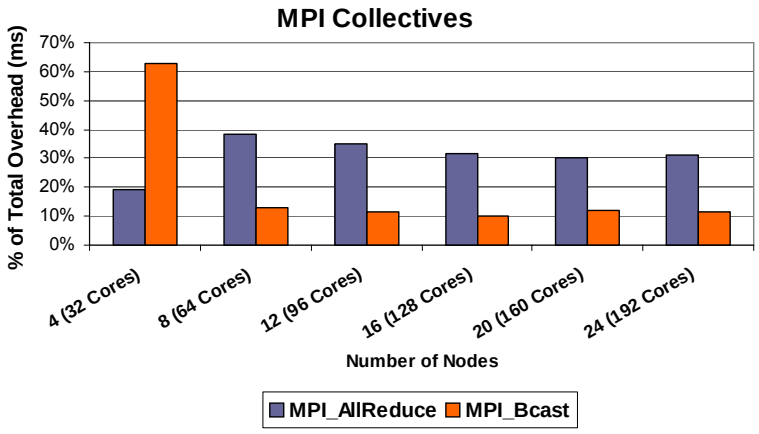

Figure 12 - 3 Vehicle Collision MPI collective profiling

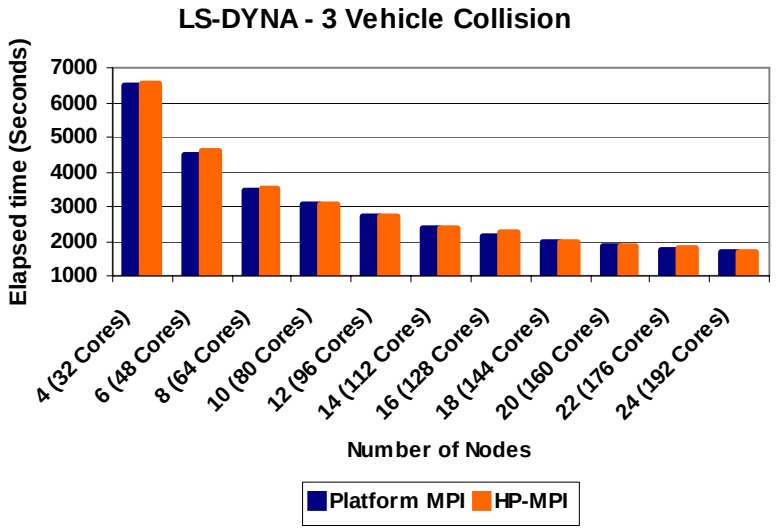

Figure 13 – HP MPI versus Platform MPI 3 Vehicle Collision performance

12 Power aware simulations

One of the major issues with today's compute infrastructures is the power consumption, which can result in power costs a over 3-4 year period that is actually larger than the system cost. Given a certain number of compute nodes, a faster interconnect leads to a faster run of LS-DYNA. Hence the reduction in power consumption of the run is proportional to the runtime. Therefore, the ability to execute more jobs per day not only increases efficiency and faster time for market of products design, but it also includes the benefits of reducing power cost per job, which in turn decreases power and cooling costs associated with a specific product design.

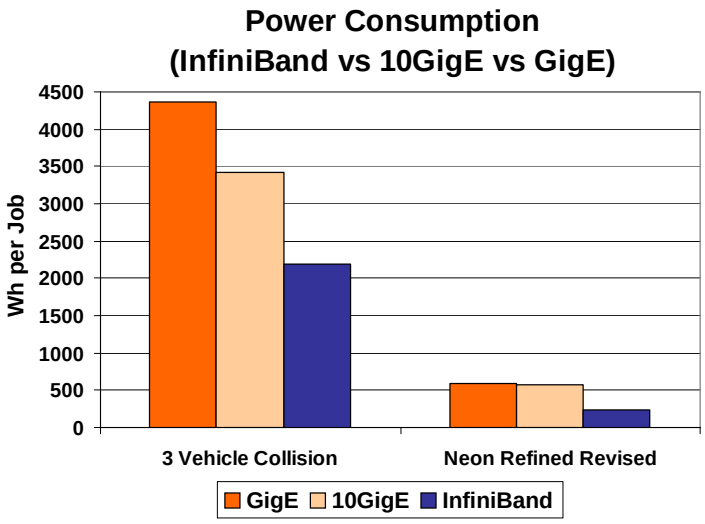

Figure 14 – Power consumption comparison using 24 nodes

Figure 14 shows the power consumption per job comparison between 20Gb/s InfiniBand, 10 Gigabit Ethernet and Gigabit Ethernet (lower is better) for a 16 nodes single job run. Using InfiniBand to connect the clustered servers together (cluster nodes) results in a reduction of the power consumption per job by up to 60%.

13 Future work

Future investigation required on the power consumption analysis of increased productivity by means of multiple parallel jobs. It is expected to keep same levels of power consumption of single job while increasing the productivity.

14 Conclusions

From concept to engineering and from design to test and manufacturing; engineering relies 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.

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

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 GigE or 10 Gigabit Ethernet are ineffective on mid to large cluster size, and can cause a reduction in performance beyond 16 or 20 server nodes (i.e. the application run time actually gets slower)

We have profiled the communications over the network of LS-DYNA software to determine LS-DYNA sensitivity points, which is essential in order to estimate the influence of the various cluster components, both hardware and software. We evidenced the large number of network latency sensitive small messages through MPI_AllReduce and MPI_Bcast operations that dominate the performance of the application on mid to large cluster size. The results indicated also that large data messages are used and the amount of the data sent via the large message sizes increased with cluster size. From those results we have concluded that the combination of a very high-bandwidth and extremely low-latency interconnect, with low CPU overhead, is required to increase the productivity at mid to large node count.

We have also investigated the increase of productivity from single job to multiple jobs in parallel across the cluster. The increase of productivity is based on two facts. First good scalability of AMD architecture that allowed us to run multiple jobs on a given compute node without saturating the memory controller. Second, the low latency and high bandwidth available on the InfiniBand interconnect that allowed us to offload the CPU to CPU data traffic from MPI communications via the interconnect instead of in compute node. The net result of that practice is an increase of the productivity by a factor of 200% with respect to the single job run.

Finally, the increase of productivity on single job runs with high speed interconnects has been analyzed from the point of view of power consumption leading to a 60% reduction or energy savings when using InfiniBand with respect to Ethernet.