WHITE PAPER

June 2013

Performance Optimizations via Connect-IB™ and Dynamically Connected Transport™ Service for Maximum Performance on LS-DYNA®

Abstract

From concept to engineering, and from design to test and manufacturing, the automotive industry relies on powerful virtual development solutions. CFD and crash simulations are performed in an effort to secure quality and accelerate the development process. LS-DYNA® relies on Message Passing Interface (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. The more complex simulation being performed to better simulate the physical model behavior, the higher the performance demands from the cluster interconnect are.

The recently launched Mellanox Connect-IB™ InfiniBand adapter 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 scalability for the largest supercomputers in the world, today and in the future. The device includes a new network transport mechanism called Dynamically Connected Transport™ Service (DCT), which was invented to provide a Reliable Connection Transport mechanism — the service that provides many of InfiniBand's advanced capabilities such as RDMA, large message sends, and low latency kernel bypass — at an unlimited cluster size. The paper will review the novel Connect-IB architecture, the new transport service, and their performance effect on LS-DYNA simulations.

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 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 – 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 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 recently launched the 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. One of its primary features is a new network transport mechanism, Dynamically Connected Transport™ Service (DCT), which can provide an unlimited cluster size with a Reliable Connection Transport mechanism – the service that provides many of InfiniBand's advanced capabilities such as RDMA, large message sends, and low latency kernel bypass.

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.

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 in 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 QDR InfiniBand, and 56Gb/s FDR InfiniBand. This study was conducted at the HPC Advisory Council systems 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® 8-core CPUs E5-2680 at 2.70 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 2.0, 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 8.3, the LS-DYNA version was LS-DYNA MPP971_s_R3.2.1, and the benchmark workload was the Three Vehicle Collision test simulation.

Figure 1 shows the elapsed time for the InfiniBand and Ethernet interconnects for a range of core/node counts for the Three Vehicle Collision case.

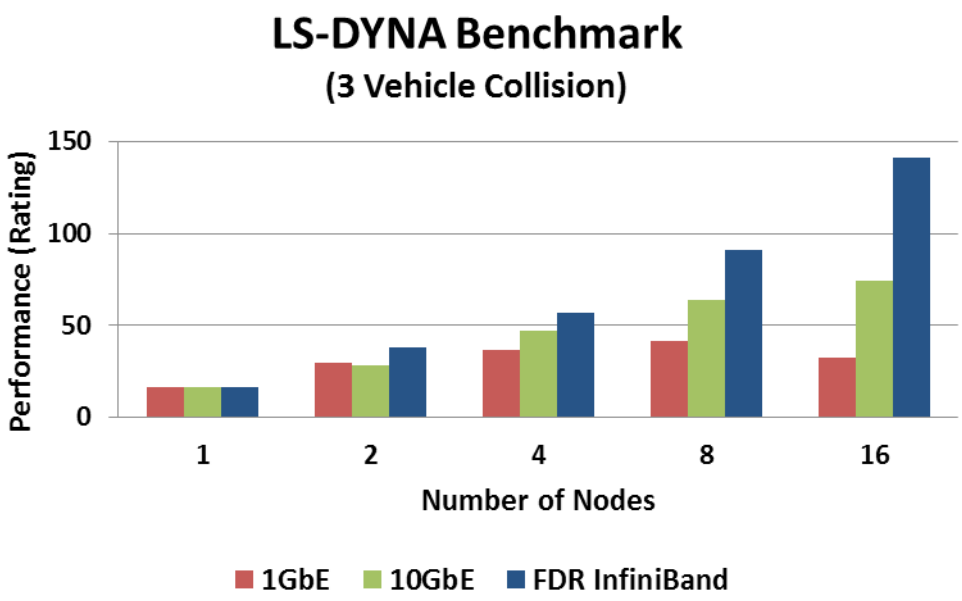

Figure 1. *Interconnect Comparison with Three Vehicle Collision*

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 336% higher than 1GbE and over 90% higher than 10GbE. While 1 Gigabit Ethernet showed a loss of performance (increase in run time) beyond 8 nodes, FDR InfiniBand demonstrated good scalability throughout the various tested configurations. LS-DYNA uses MPI for the interface between the application and the networking layer, requiring 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.

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 Sandy 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.

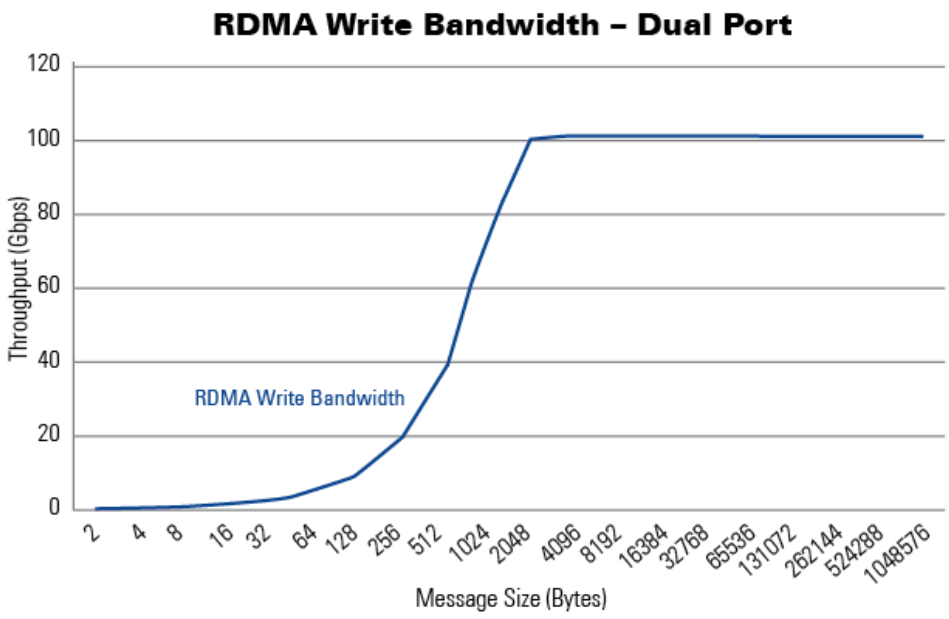

Figure 2. *100Gb/s RDMA Write Performance*

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.

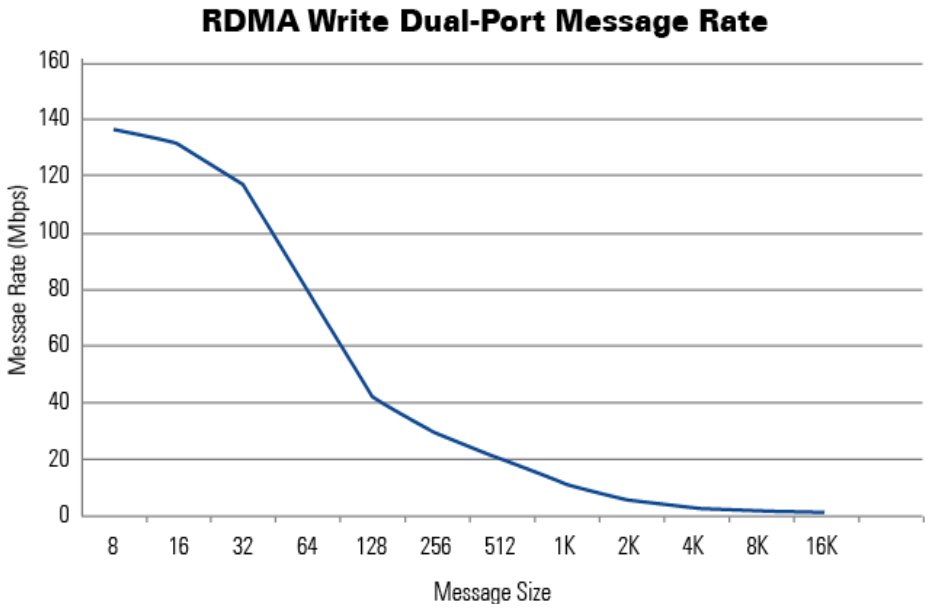

Figure 3. *Message Rate Example*

Dynamically Connected Transport™ (DCT)

One of the major strengths of InfiniBand is the capabilities it enables with Reliable Connection (RC) Transport Services. These provide a number of advantages for parallel computing communications including end-to-end reliability performed by the adapter hardware, full transport offload, large send/receive messages, and the capability of remote memory access through RDMA. Because the RC Transport service requires connections to be established between the two endpoints of the connection, context for these connections must be established and stored on the endpoints. The amount of context grows with the size of the job, thus the amount of connections that need to be established grows. At extreme large scale this can cause a higher amount of memory consumption on the endpoint host's memory. It can also affect performance of the endpoint at large scale when the adapter resources become heavily used and context retrieval from the system memory to adapter happens at a higher frequency.

For example, with RC connections, a single connection is established between every CPU core and every other CPU core within the running application. This means that each endpoint will hold P2*N connections (where P=PPN or processors per node, and N=number of nodes participating in the job). Thus, a 16-core machine running across 256 nodes would require 65,536 connections on each endpoint. Even using Extended Reliable Connected Transport Service (XRC), introduced in 2007, provides changes in the transport layer mechanism to reduce the number of connections to P*N, such that in our example the number of connections per endpoint is reduced down to 4,096.

By using DCT, the number of connections per endpoint can be reduced from 65,536 using RC connections to 4,096 using XRC connections to only 16 with DCT, as the adapter resources and host memory consumption are no longer related to the system size in the matter of number of compute nodes, but rather are only related to the capacity and performance of the processors.

Future testing will include the evaluation of the DCT transport.

Performance Improvements in LS-DYNA Simulations

Because Connect-IB supports the PCIe Gen3 standard, it is perfectly suited to run on the Intel E5-2600 Series (Sandy Bridge) and to take advantage of the increases in CPU core processing, memory bandwidth, and network throughput. Similarly, the Sandy Bridge architecture enables the Connect-IB InfiniBand device to run at its maximum throughput and lowest latency. The results are shown in Figure 4.

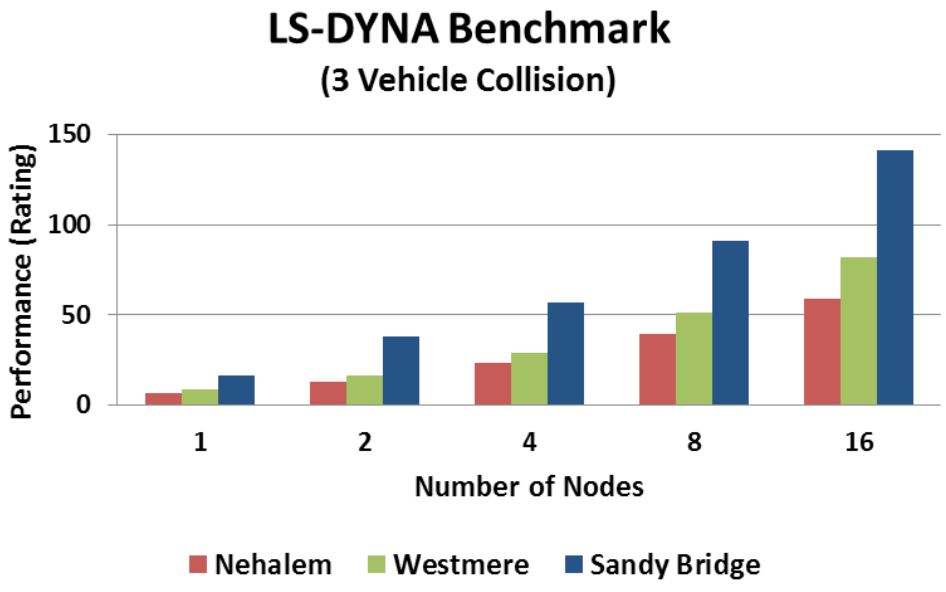

Figure 4. *LS-DYNA 3 Vehicle Collision Performance per CPU technology*

Compared to previous system generations, the Sandy Bridge cluster outperforms the Intel Xeon X5670 (Westmere) cluster by up to 73%, and gains up to 141% 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.

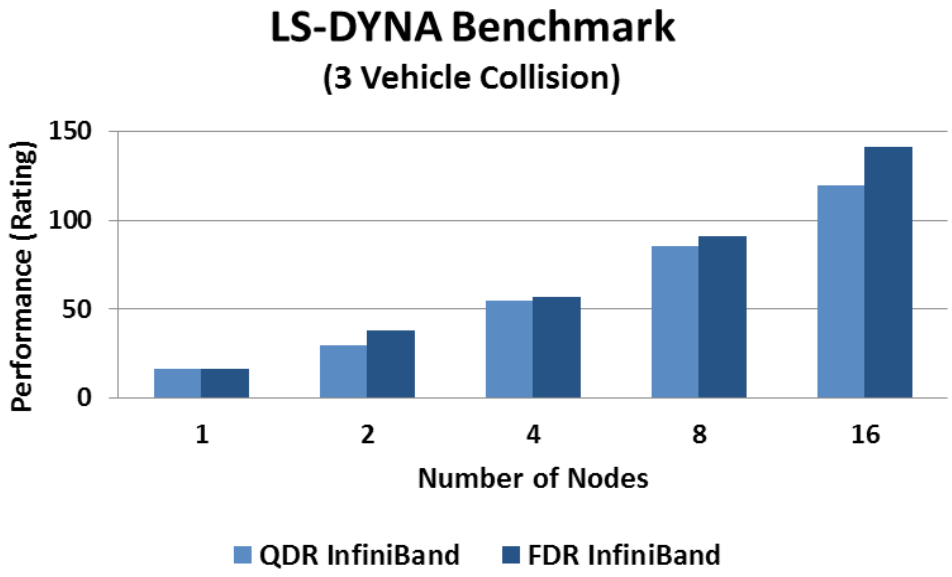

Figure 5: *LS-DYNA 3 Vehicle Collision Performance per InfiniBand technology*

Furthermore, it is possible to distinguish between QDR InfiniBand and FDR InfiniBand using the same set of systems. When placed side-by-side, Connect-IB clearly demonstrates a raw improvement over its predecessor ConnectX-3. Connect-IB is unique in that it utilizes improved memory resource management and more efficient transport service, allowing the HPC cluster to run at its highest scalability.

FDR InfiniBand showed an 18% improvement in performance at 16 nodes over QDR InfiniBand, and the margin for additional performance improvement is expected to be wider as more nodes are involved, as the effects of DCT allow for even greater scalability.

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.

HPC 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 1 or 10 Gb/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, support for the PCI Gen3 standard, and unique Dynamically Connected Transport service – offer increased bandwidth, lower latency, and greater scalability than when using QDR InfiniBand, 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.

Mellanox TECHNOLOGIES

350 Oakmead Parkway, Suite 100, Sunnyvale, CA 94085
Tel: 408-970-3400 • Fax: 408-970-3403
www.mellanox.com