The Effect of MPI Collective Operations and MPI Collective Offloads on LS-DYNA® Performance

Gilad Shainer[1], Tong Liu[1], Pak Lui[1], Dave Field[2]
[1]*Mellanox Technologies* [2]*Hewlett-Packard*

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. The recent trends in cluster environments, such as multi-core CPUs, GPUs, cluster file systems and new interconnect speeds and offloading capabilities are changing the dynamics of clustered-based simulations. Software applications are being reshaped for higher parallelism and hardware configuration for solving the new emerging bottlenecks, in order to maintain high scalability and efficiency. In this paper we cover new hardware based accelerations and offloads for MPI collectives communications and how is affect LS-DYNA performance and productivity.

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, while providing a system that can be used 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-threads, and hardware configuration 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 these environments provide better flexibility, scalability and efficiency for such simulations.

LS-DYNA relies on the Message Passing Interface (MPI) for cluster or node-to-node communications, the de-facto messaging library for high performance clusters. Collectives communications are the point to multipoint messaging operations frequently used by MPI for operations like broadcasts for sending around initial input data, reductions for consolidating data from multiple sources and barriers for global synchronization among the cluster nodes. Any collective communication executes some global communication operation by coupling all processes in a given group. As such, collective communications have a crucial impact on the application's scalability. In addition, the explicit and implicit communication coupling, used in high-performance implementations of collective algorithms, tends to magnify the effects of system-noise on application performance further hampering application scalability.

Mellanox 40Gb/s QDR InfiniBand adapters (ConnectX-2 and later) addresses the collective communication scalability problem by offloading a sequence of data dependent communications to the Host Channel Adapter (HCA). This solution, named CORE-direct, provides the mechanism needed to support computation and communication overlap, allowing the communication to progress asynchronously in hardware while at the same time computations are processed by the CPU. It is also provides a means to reduce the effects of system noise and application skew on application scalability.

In this paper we review the most used collectives operations by LS-DYNA, namely MPI AllReduced and MPI Broadcast, their performance affect on LS-DYNA simulations, and first results of LS-DYNA with the new CORE-direct technology which offload the collectives operations from the software to the network.

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 ten-thousand nodes and multiple CPU cores per server platform and efficient utilization of compute processing resources.

This study was conducted at the HPC Advisory Council systems center (www.hpcadvisorycouncil.com) on an HP ProLiant SL2x170z G6 16-node cluster, each server has dual socket six-core Intel X5670 @ 2.93 GHz CPUs and 24GB memory. The operating system is CentOS5U4 with InfiniBand drivers OFED 1.5.2. Mellanox ConnectX-2 InfiniBand

QDR adapters and InfiniBand QDR switches are used for the cluster interconnect. The MPI version is Open MPI 1.4.3, LS-DYNA LS-DYNA mpp971_s_R5.0 with the 2001 Ford Taurus, detailed model (1,057,113 elements) benchmark.

MPI Collectives Communications

Scientific simulation codes frequently use collective communications. The ordered communication patterns used by high performance implementation of collective algorithms present an application scalability challenge. This impediment is further magnified by application load imbalance and system activity, or system noise, delaying the collective operations. MPI Collective communications provide the capabilities for doing broadcasts for MPI applications for sending around initial input data, reductions for consolidating data from multiple sources and barriers for global synchronization. Any collective communication executes some global communication operation by coupling all processes in a given group. This behavior tends to have the most significant negative impact on the application's scalability. In addition, the explicit and implicit communication coupling, used in high-performance implementations of collective algorithms, tends to magnify the effects of system-noise on application performance further hampering application scalability.

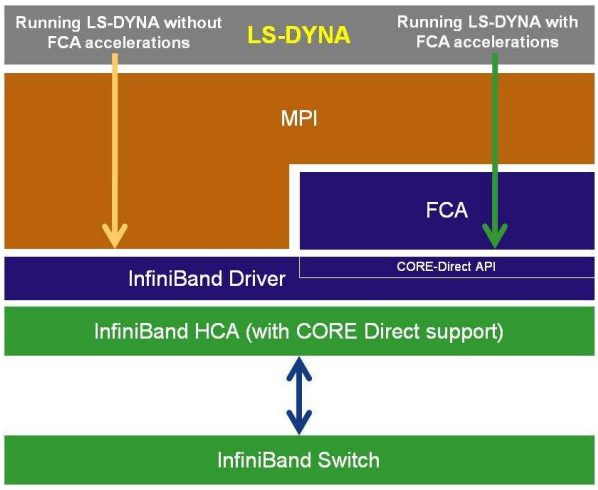

Figure 1 – software layers overview with and without FCA

Mellanox adapters and switches address the collective communication scalability problem by offloading the MPI collective communications to the network. This solution provides the mechanism needed to support computation and communication overlap, allowing the communication to progress asynchronously in hardware while at the same time computations are processed by the CPU. It is also provides a means to reduce the effects of system noise and application skew on application scalability. Mellanox Fabric Collectives Acceleration (FCA) includes the CORE-Direct hardware offload engine within the Mellanox HCAs and a sophisticated software implementation of MPI collectives communication algorithm. FCA is being licensed by Mellanox Technologies and can be integrated with leading MPI libraries. In

this paper we have used Open MPI, and Mellanox FCA. Since FCA utilizes the interconnect hardware engines, it reduces the collectives runtime, increases the CPU availability to the application and allows overlap of communications and computations with asynchronous collective operations.

Figure 1 shows the software architecture of typical InfiniBand based system, with and without FCA. CORE-Direct is hardware accelerations for collectives operations which is exposed through the InfiniBand drivers to third party libraries. FCA is one of the implementations that utilizes CORE-Direct for MPI collectives accelerations, and it is licensed from Mellanox. Figure 2 shows the system architecture. FCA take advantage of both CORE-Direct as well as CPU resources if exist in the switch fabric.

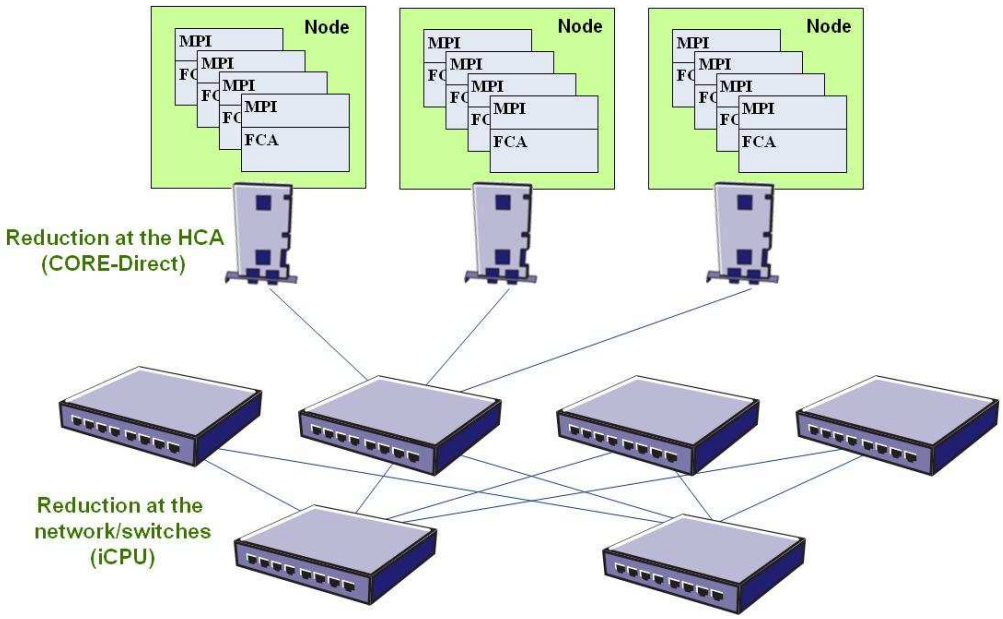

Figure 2 – System architecture with FCA

LS-DYNA MPI Profiling

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

LS-DYNA MPI profiling data is presented in Figure 3which shows the usage of the different MPI communications in several cluster configurations (32-cores, 64-cores 128-cores).

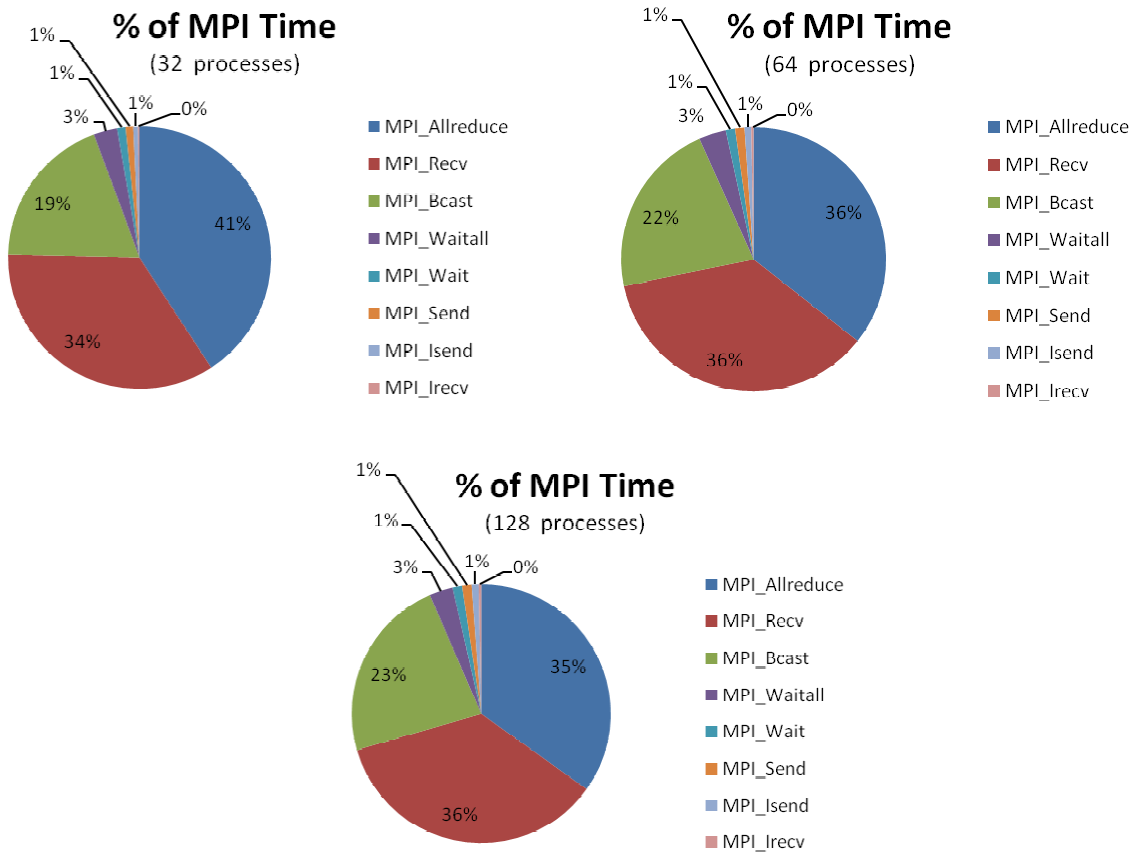

Figure 3 – Distribution of the different MPI communications

From figure 3 it is clear that the two MPI collectives, MPI_Allreduce and MPI_Bcast (Broadcast) consume most of the total MPI time and hence is critical to LS-DYNA performance. MPI libraries and offloading related to those two collectives operation will greatly influence the system performance.

LS-DYNA Performance Results with FCA

MPI collectives offloads via FCA is targeted for large scale system to minimize the effect of system noise and jitter, and to provide future proof solution for future MPI-3 usage which will include asynchronous collectives operations. As a side effect, FCA also accelerates the collectives operations and reduces the MPI consumed time and the CPU overhead, hence increases the applications performance also at low scale. We have tested a 16-node cluster in order to identify the system size in which FCA will start showing a performance and productivity advantage.

Figure 4 shows the performance results of the Ford Taurus benchmark with and without the Fabric Collectives Acceleration and CORE-Direct offloading hardware and software package.

Up to 4 nodes configuration, we have see no much of an advantage with FCA, but at 8 nodes configurations the performance increase was 5% and at 16 nodes, 192 cores, the performance increase was 15%.

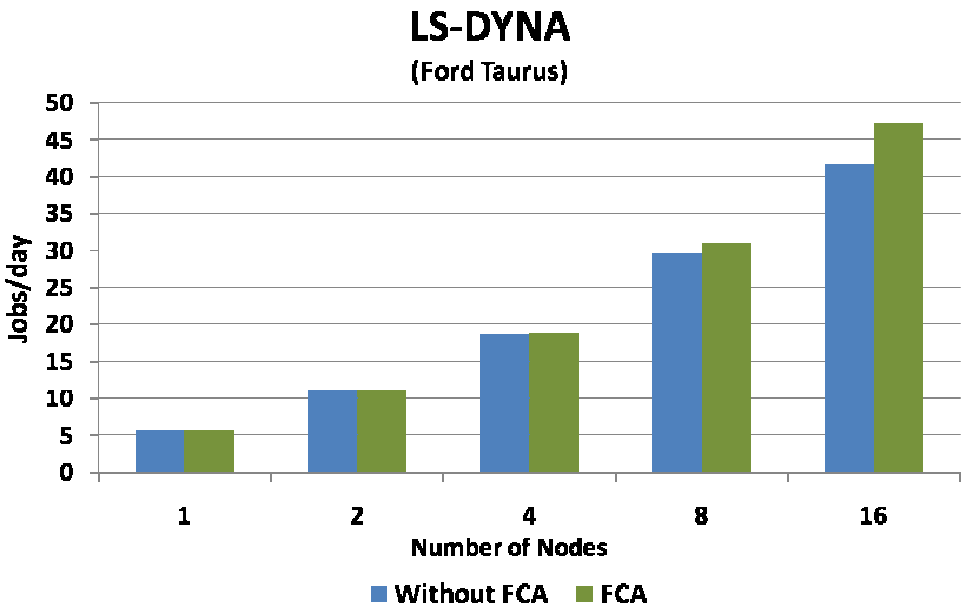

Figure 4 – LS-DYNA performance results with and with FCA

The performance advantage with FCA increases with cluster size, and we expect to see more than 20% at 32 nodes etc. From the results we can conclude that any system size above 128 cores will benefit from FCA and CORE-Direct, where the benefit will increase as more cores will be used.

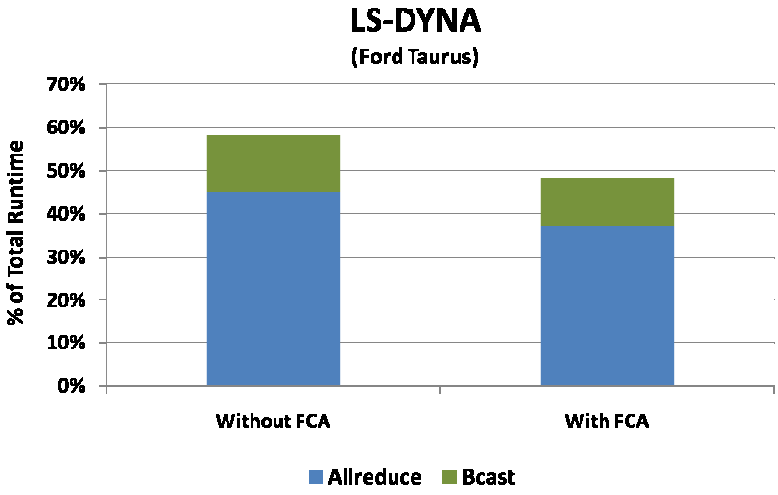

Figure 5 – LS-DYNA MPI time for AllRedcue and Broadcast collectives operations

Figure 5 shows the time spend in the MPI library for the MPI All Reduce and MPI Broadcast operations at 192 cores configuration (16 nodes in our setup) with and without FCA and CORE-Direct. One of the side benefits of FCA is the accelerations of the collectives communications and therefore we saw the increase in the LS-DYNA performance even at low scale. According to the results in figure 5, FCA reduces the MPI All Reduce and MPI Broadcast time by more than 10% each. We do expect higher acceleration and reduction in MPI time at a larger core-count.

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.

The new concept of MPI offloads provided by Mellanox Technologies was developed and designed together with Oak Ridge National Lab and provides the necessary capabilities for efficient large scale computing environments.

In order to examine the benefits of FCA andCORE-Direct at large scale commercial HPC, the Livermore Software Technology Corporation (LSTC) LS-DYNA software was investigated. We showed that FCA and CORE-Direct becomes essential for higher productivity beyond 128, and for example provided 15% performance increase at 192 cores. We expect to see larger impact on performance at larger cluster sizes. As FCA and CORE-Direct already provide the capability for MPI asynchronous collectives (MPI-3) we expect to see higher gain with the next generation MPI solutions that will be based on the MPI-3 specifications.