Performance Analysis of LS-DYNA® in Huawei HPC Environment

Pak Lui, Zhanxian Chen, Xiangxu Fu, Yaoguo Hu, Jingsong Huang
Huawei Technologies

Abstract

LS-DYNA is a general-purpose finite element analysis application from LSTC. LS-DYNA is capable of simulating and solving complex real-world structural mechanics problems in an HPC cluster environment. In this paper, we are analyzing different areas that can impact on the performance of LS-DYNA by comparing different hardware components in Huawei HPC cluster environment. By evaluating the components, such as CPUs, network interconnects, system and software tuning on the latest Huawei HPC cluster solutions, we can demonstrate the sensitivity of the components on LS-DYNA performance which may help achieve higher productivity on LS-DYNA workloads.

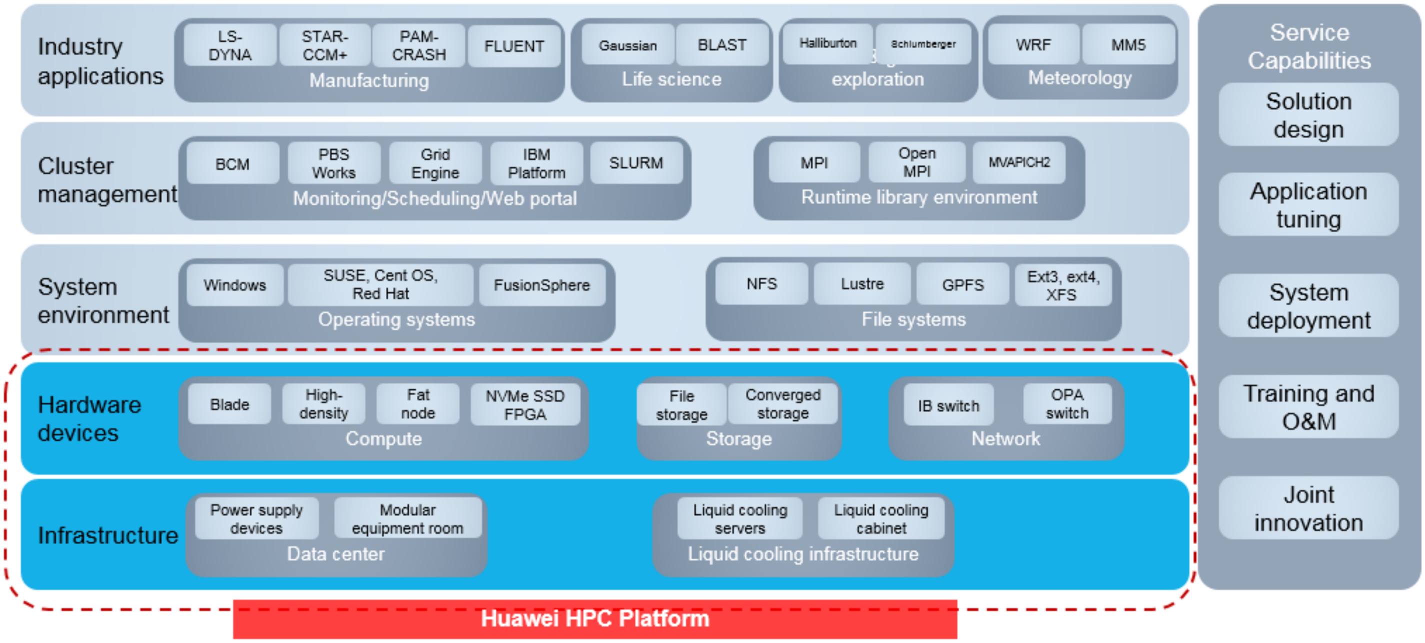
Introduction

The objective of this paper is to analyze various components that can impact LS-DYNA performance in a High performance computing (HPC) cluster environment. The LS-DYNA/MPP version is capable of simulating and solving complex real-world structural mechanics problems in an HPC cluster environment. In the subsequent sections, we would describe the performance benefits of using certain hardware system components that would benefit LS-DYNA performance which is used in the Huawei HPC Cluster Solutions.

Huawei HPC Cluster Solutions

Huawei is a leading global information and communications technology (ICT) solutions provider. The Huawei HPC infrastructure is designed to solve complex simulation problems and speed up the time-to-solution by utilizing the cluster solution where powerful, highly-efficient systems are deployed to solve scientific, engineering, and data analysis problems.

Huawei provides a complete end-to-end HPC solutions, which range from data center infrastructure, cooling solutions, hardware resources, system environment, cluster management, service platform, and HPC industrial applications. Huawei has a complete portfolio of products which include highly integrated blade servers, high-density servers, and the “KunLun” line of large SMP supercomputing platform. In addition, Huawei supports various storage devices, high-speed, low-latency InfiniBand and Ethernet networking switches to build high-performance computing clusters. In addition to infrastructure, Huawei provides modular data centers, container data center solutions, and liquid cooling solutions. At the software level, Huawei cooperates with many high-performance cluster software vendors and application software vendors to perform integration tests and optimization for mature HPC commercial products and components and provide high-performance solutions that are most suitable for user services.



Hardware and Software Configuration

Huawei HPC cluster solution utilizes Huawei high-performance servers, large-capacity storage, and innovative cluster and device management software, providing the power of HPC to easily solve these complex problems. The tests shown in the subsequent sections were conducted on two generations of Intel-based HPC clusters. One that is based on the Intel "Broadwell" CPUs on a cluster of Huawei FusionServer X6000 servers that each server includes 4 of the XH321 compute nodes; the other cluster is based on the "Skylake" CPUs on a cluster of Huawei FusionServer E9000 blade enclosure that each contains 16 of the CH121 V5 compute nodes.

Figure 1: (Left) FusionServer X6000 with four XH321 blade compute nodes. (Right) FusionServer E9000 with sixteen CH121 V5 half-width blade compute nodes, and integrated InfiniBand switch module. Both chassis support air and liquid cooling

The hardware configuration of the cluster is the same. Briefly speaking, we used a cluster of 32 nodes, each node has dual Intel Xeon E5-2690 v4 CPUs; all nodes are connected with EDR Infiniband.
The following shows the detailed information about the hardware and software used in every compute node.

Hardware configuration for Intel E5-2600v4 “Broadwell” series cluster:

Compute Node	Huawei FusionServer X6000 server, XH321 V3 compute nodes
CPU	Dual Socket Intel Xeon E5-2680 v4, 14 cores @ 2.4GHz, 35MB cache Dual Socket Intel Xeon E5-2690 v4, 14 cores @ 2.6GHz , 35MB cache Dual Socket Intel Xeon E5-2697A v4, 16 cores @ 2.6GHz, 40MB cache
Memory	256GB, DDR4 2400 MHz
OS	RHEL 7.2
InfiniBand	Mellanox ConnectX-4 EDR InfiniBand
Storage	Huawei OceanStor 9000 Scale-out NAS storage system

Hardware configuration for Intel Xeon 6100 “Skylake” series cluster

Compute Node	Huawei FusionServer E9000 server, CH121 V5 compute nodes
CPU	Dual Socket Intel Xeon Gold 6138, 20 cores @ 2.0GHz, 27.5MB cache Dual Socket Intel Xeon Gold 6140, 18 cores @ 2.3GHz, 24.75MB cache Dual Socket Intel Xeon Gold 6148, 20 cores @ 2.4GHz, 27.5MB cache
Memory	192GB, DDR4 2666 MHz
OS	RHEL 7.3
InfiniBand	Mellanox ConnectX-4 EDR InfiniBand

Software Configuration

LS-DYNA	LS-DYNA/MPP ls971 R9.1.0, single precision
MPI	Intel MPI 2018 Mellanox HPC-X v1.9.7 Platform MPI 9.1.4.3

Benchmark Datasets

The benchmark dataset that will be described in the subsequent sections are obtained from the TopCrunch (topcrunch.org) web site. The TopCrunch project provides a web site that list the aggregate performance of HPC systems used with engineering simulation software.

- neon_refined_revised: it is a dataset is a frontal crash with initial speed at 31.5 miles/hour of a 1996 Plymouth Neon. The model consists of about 500 thousand elements with a simulation time for 30ms. The model is created by National Crash Analysis Center (NCAC) at George Washington University.
- 3cars: The 3car dataset involves a van crashes into the rear of a compact car, which in turn, crashes into a midsize car. Vehicle models created by NCAC. The simulation time is 150ms.
- Caravan2m-ver10: The caravan model is created by National Crash Analysis Center (NCAC) at George Washington University (GWU). The model is consists of 2.4 million elements with simulation time of 120ms. The run writes around 200MB.
- Odb10m-ver16: This LS-DYNA model comprises of 10 million elements, configured with a simulation time of 120 milliseconds. This model has been developed by the NCAC of the GWU under a contract with the FHWA and NHTSA of the US DOT.

The performance analysis were conducted with LS-DYNA on the two clusters. Various factors affecting LS-DYNA performance were evaluated.

Performance Metric

After a completion of an LS-DYNA run, the elapsed time will be reported in the output. The metric used in this study is called Performance Rating, which is essentially the number of simulation jobs can be run per day. The higher of the performance rating value represents better performance, which simulations can be generated faster. We can see that as the LS-DYNA job scales to run more nodes, we would typically expect the LS-DYNA performance to improve as more cores are used in the simulation run; it also means that the elapsed time would be reduced significantly as more compute nodes are used to process the simulation. By using performance rating, we can easily see how well the LS-DYNA simulation scales, because it would be hard to see in graph the differences in runtime at larger node counts.

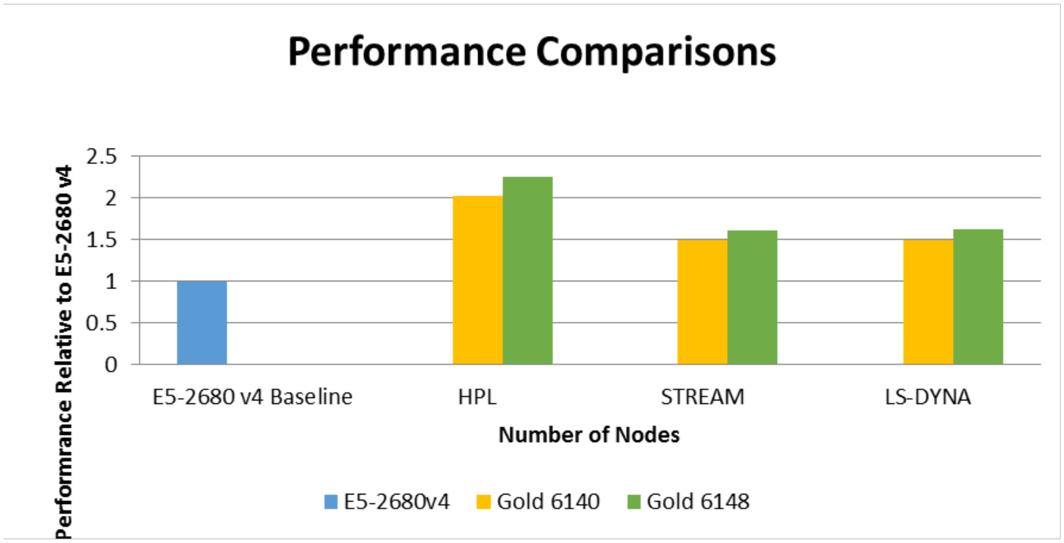
The BIOS Turbo mode was enabled for the test. Here the IVB processors performed 12-16% better compared to the SB processors (as shown in Figures 1 and 2). It will come as little surprise those familiar with LS-DYNA that this observed performance increase tracks linearly with the 16% memory bandwidth increase of IVB platform over the SB platform, proving that application performance is tied to much more than just the sheer number of CPU cores or clock frequency.

Performance on a Single Node

In this section, we are comparing the performance differences between different SKUs of the Skylake CPU, and using Broadwell CPU as a baseline to compare with. The types of CPUs in the comparisons are as follows:

- Broadwell: Intel Xeon E5-2680 v4, 14 cores @ 2.4GHz, 35MB cache
- Skylake: Intel Xeon Gold 6140, 18 cores @ 2.3GHz, 24.75MB cache
- Skylake: Intel Xeon Gold 6148, 20 cores @ 2.4GHz, 27.5MB cache

When we look at the differences between the two CPU generations, we first strictly look the differences in terms of the number of CPU cores and normal clock frequency, we can see that the Gold 6140 has about 28% compared to E5-2680v4. Likewise, the Gold 6148 has roughly 42% more cores compared to E5-2680v4. The difference in CPU clock frequency is roughly the same. The normal CPU clock on Gold 6140 is roughly 4% less compared to E5-2680v4, and the same for Gold 6150.



Between the two Skylake CPUs, we observe that Gold 6148 has 11% more CPU cores and 4% higher clock frequency than the Gold 6140. If it is directly translated into performance, we would expect about 15% of better performance at best.

We ran a couple of the HPC benchmarks to show differences between the two CPU generations, we noticed that High Performance LINPACK (HPL) demonstrates over 103% of the performance improvement on the Gold 6140 compared to the E5-2680v4, and Gold 6148 demonstrates 125% of the improvement, thanks to the new additional of the AVX-512 instruction set which allows up to 32 FLOPs per clock cycle on certain instructions used by the matrix multiplication inside HPL. Compared to only 16 FLOPs per cycle available in the Broadwell generation of the E5-2680v4 CPU.

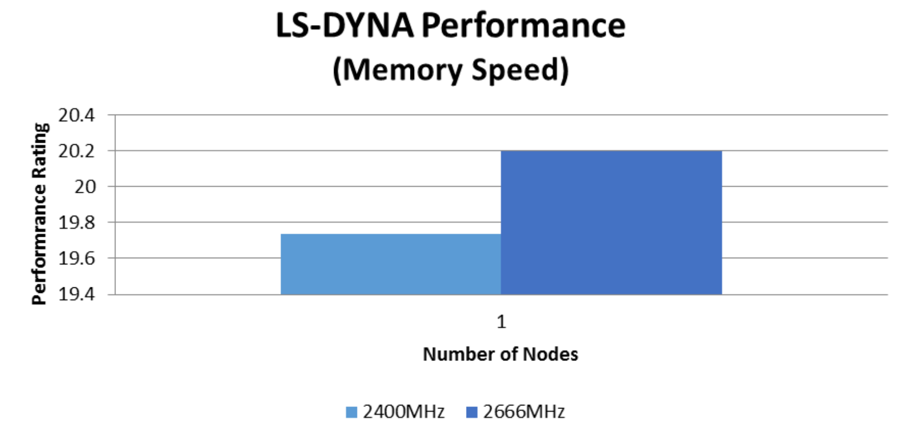
The other workload we run is STREAM benchmark which measure the system memory bandwidth. The Skylake architecture introduces 6 memory channels in the architecture, compared to only 4 memory channels in the Broadwell architecture. Skylake demonstrated about 50-61% of the improved performance compared to the Broadwell architecture.

Coincidently, the single-node LS-DYNA performance we run yields the same performance gain achieved by STREAM. Based on this finding, it is natural to assume that LS-DYNA is very sensitive to memory bandwidth. The increase of the memory channels open up additional memory bandwidth operations for the CPU to utilize.

Memory Speed

We look further into the memory subsystem by comparing 2 different memory speeds supported on the Huawei Skylake platform. The system memory we compared here are the 2400MHz and 2666MHz DIMMs. The calculated difference between the 2 types of DIMMs is about 11%, as 2666MHz DIMMs is 11% faster than the 2400MHz DIMMs.

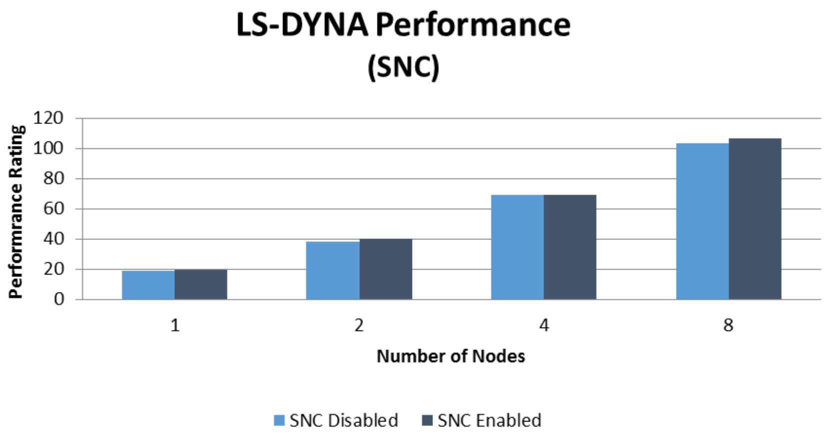
When we use LS-DYNA to measure the performance difference, LS-DYNA reports only about 2% of the improvement on a single node. Only part of the speed difference is translated into LS-DYNA performance gain.



Sub-NUMA Clustering

Sub-NUMA Clustering (SNC) is a new Intel technology that is similar to a cluster-on-die (COD) in Xeon 2600v3/v4 (or Haswell/Broadwell) generation. On the system that has SNC enabled, CPU cores and memory of socket would be split into 2 separate NUMA domains. Compared to COD, the idea of SNC is to improve memory throughput between remote NUMA regions. In this section, we want to see if there is an effect of Sub NUMA Cluster (SNC) modes in the BIOS on the LS-DYNA performance.

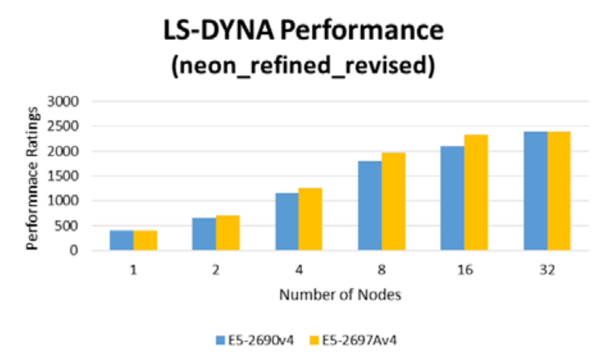
In general, SNC would show some benefits for applications that requires good NUMA locality. By utilizing SNC, which allows for better memory access to remote NUMA domains, we are able to see a performance gain of 3% on a single node basis.

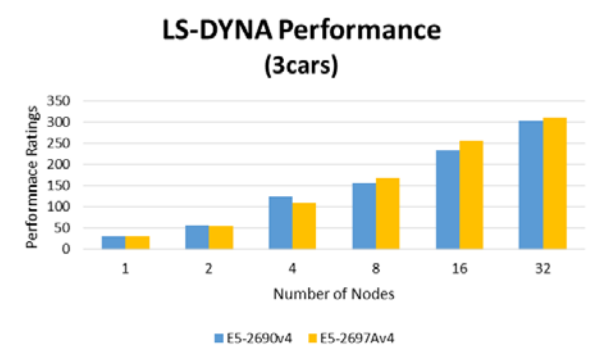


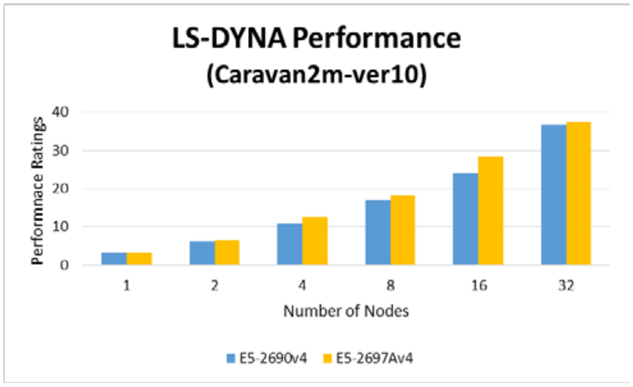
CPU Performance Differences at Scale

When we measure the performance at scale, the CPU performance difference appeared appear to have the less effect. Here we are comparing runs that has the E5-2690v4 CPU and the E5-2697Av4 CPUs. Each of the E5-2690v4 CPU has 14 cores running at 2.6GHz, another is E5-2697Av4 which is a 16-core version that runs at a 2.6GHz of normal clock.

With the small dataset, neon_refined_revised, similar performance is observed on single node and 32 nodes, and the cluster of E5-2697Av4 performs better than the cluster consisted of the E5-2690v4 CPUs.







We assume that dataset with larger number of elements would have higher CPU utilization. With the larger dataset, caravan2m-ver10, it appears that LS-DYNA dataset that can demonstrate better speedup at scale, as the larger dataset requires more CPU utilization, the performance benefit from the better CPU become up transparent on larger dataset. The E5-2697A v4 has more cache, and with the additional 4 cores available per node, and a faster turbo speed, the performance advantage due to the additional core count is calculated to be 14.2%. The measured performance gain of a cluster of E5-2697Av4 compared to a cluster of E5-2690v4 is as high as 14% at 16 nodes (448 cores).

As job scales to more CPU cores which also participates in MPI communications, the percentage of time that the application process spent on MPI communications would be greater in the overall runtime, which means that CPU/computation percentage would be smaller. At 32 nodes (896 cores), the performance differences between the cluster of different processes would be minimal, because MPI communications become more dominates than computation, the effect of computational performance becomes less.

CPU turbo speed

The CPU core speed can be configured to run at turbo speed to achieve better performance. In the specification of the Gold 6148 processor, it shows that base clock runs at the frequency of 2.4GHz, or 2.4 billion cycles per second. It is the rate at which the processor's transistors would open and close for operations. Each processor has the defined base clock frequency which the CPU would operates at the Thermal Design Power (TDP) which is at 150Watts for this CPU.

The max turbo frequency defines the maximum core frequency at which the processor is capable of operating using the Turbo boost technology. For the Gold 6148 CPUs, the max turbo defined by the specification is at 3.7GHz. Typically the rate of the turbo mode depends on a few factors, such as the workload, the number of concurrent active CPU cores, power consumption and processor temperature. For MPI workload where all of the CPU cores are in used to process the workload concurrently, the actual turbo clock speed observed would be less. In the case of the LS-DYNA run during an actual simulation, the CPU cores measured are running between 2.7-2.8GHz.

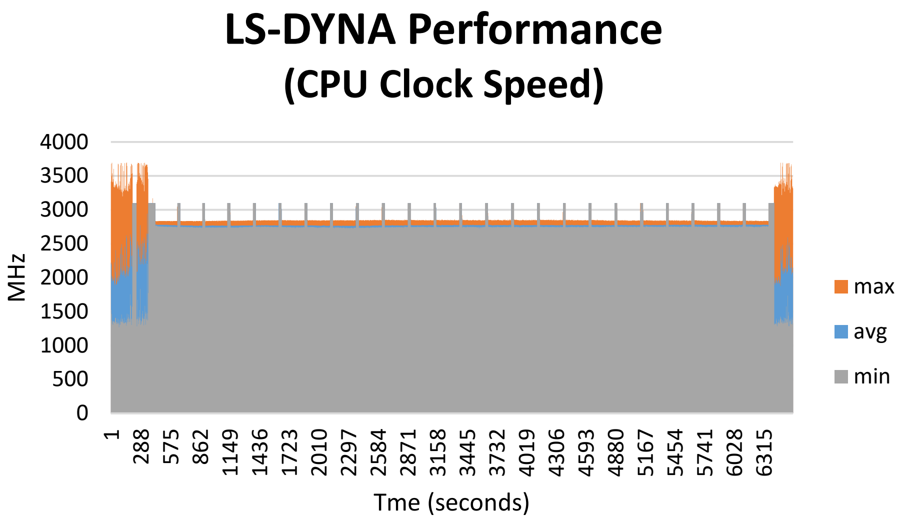


Figure 2: Turbo speed of the Xeon 6148 CPU cores on a compute node

CPU extensions support

LS-DYNA provides a range of different types of executables to support systems with various architectural capabilities. One of the executables provided by LS-DYNA is the executable that supports the AVX2 CPU instruction set for the Broadwell and the AVX-512 support on the Skylake systems. With Intel AVX2 technology, it allows the processor to execute 16 floating point arithmetic per cycle; and with Intel AVX512 instructions, it allows the processor to execute 32 FLOPs per cycle.

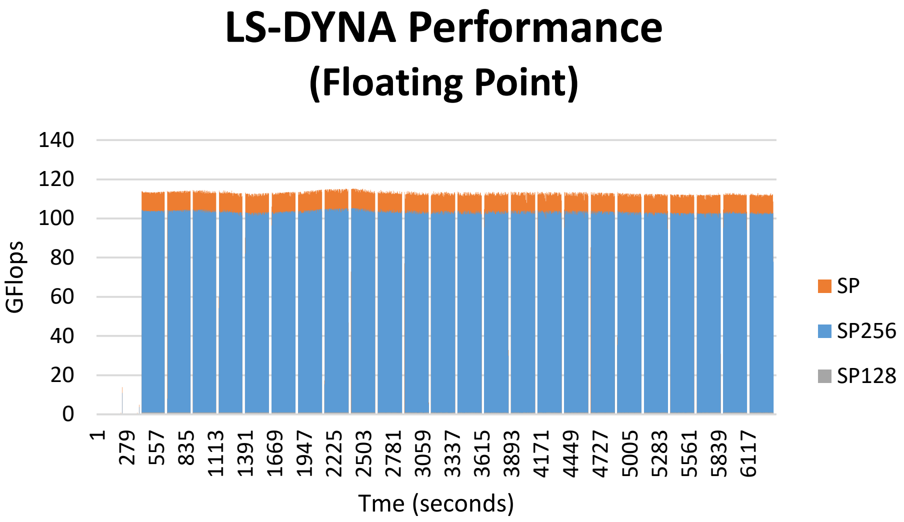
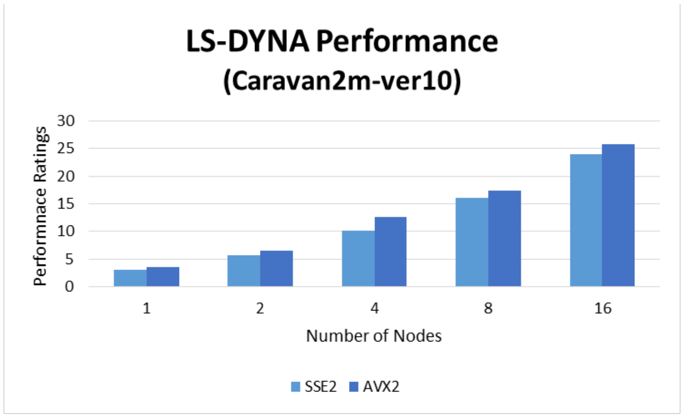


Figure 3: Example of floating point bandwidth observed using single precision using LS-DYNA AVX2 executable

SSE2 versus AVX2

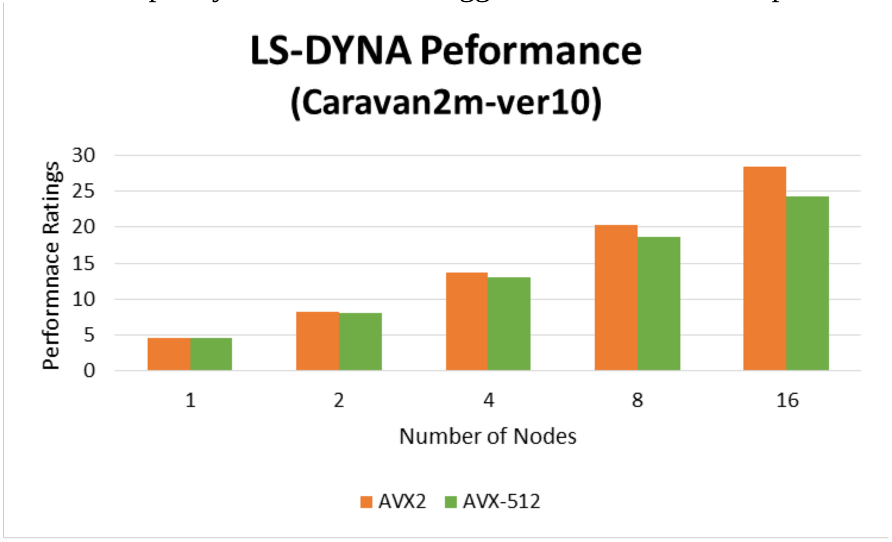
In the comparison of SSE2 and AVX2 on same cluster with Broadwell CPUs, we are able to show higher performance than SSE2 executable on "Broadwell" CPU. The performance gain of 7-23% by using AVX2 over SSE2 executable. It is important to note AVX2 instructions runs at a reduced clock frequency as a typical normal clock. With AVX2, it provides speedups of floating-point multiplication and addition operations. The benefit of AVX2 appears to be larger on dataset with larger number of elements.

For all of runs, LS-DYNA with AVX2 binary executable performs better than the SSE2 versions by 7 to 23% on the Broadwell platform. On a single node basis, the AVX2 is 14% better than the SSE2 code. At 448 cores (16 nodes), the performance delta is about 7%. This difference decreases at higher core counts, to 7% at 448 cores or 16 nodes.



AVX2 and AVX-512

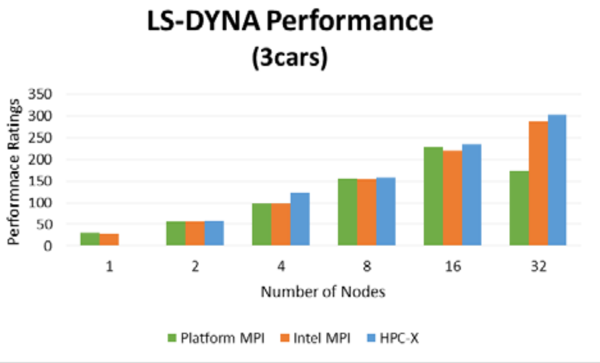
On the Skylake platform, we compared between LS-DYNA binaries that built with the AVX2 and AVX512 support using the Caravan2m dataset. It is observed that AVX2 outperforms both AVX-512 and SSE2 executables on Skylake CPUs. AVX2 has a performance gain of 17% over AVX-512 executables, that is surprising, despite the improved vectorization in AVX-512. It is observed by inspecting the 'turbostat' output during the run that AVX-512 instructions runs at a reduced clock frequency as AVX2 and normal clocks. When running with AVX2 executable, the clock are reported in the range of 2.3GHz-2.5GHz. When running AVX512 executable, the clock on the CPU cores are running consistently in the range of 2.2-2.3GHz. The CPU core runs slower on the AVX512 instructions, despite the improved vectorization available to make be able to processes twice the number of instructions per cycle. The results suggest that AVX2 would perform better than AVX-512.

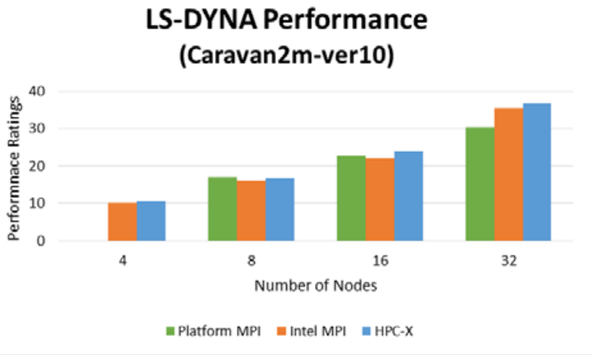


MPI Libraries

MPI library is responsible for passing messages between application processes. The difference in the algorithm used in communications would have a real impact on scalability performance. We compare 3 of the popular MPI implementations to study their effect for scalability.

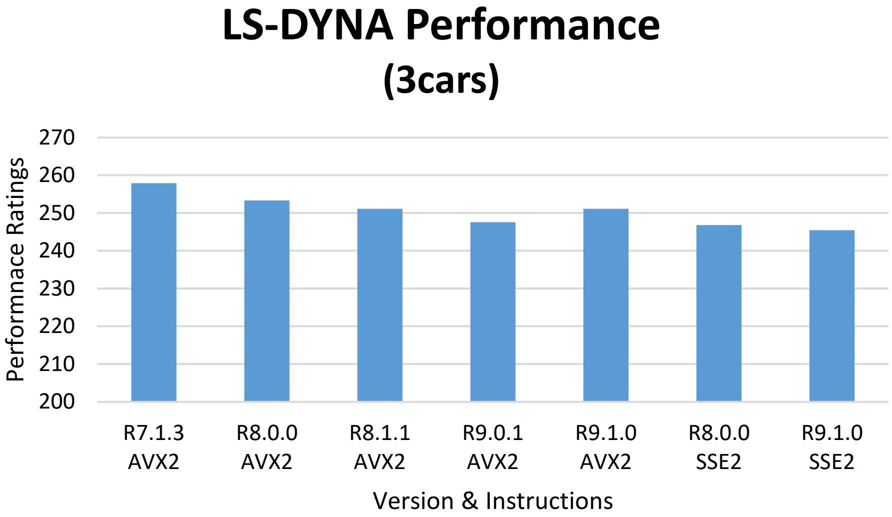
At lower CPU core counts, all 3 MPI implementations tested perform similarly. As more nodes and CPU cores are involved, we observed that the performance of the MPI libraries start to differentiate. The MPI implementations become more noticeable after 8 nodes. For this study, Mellanox HPC-X MPI Toolkit, and Intel MPI library perform noticeably better than Platform MPI at larger core counts. Platform MPI performs better at small node count, while HPC-X shows better performance at scale. HPC-X demonstrates 18% advantage at 32 nodes for Caravan2m-ver10. Platform MPI was runs with these parameters to provide better scalability: -IBV -cpu_bind, -xrc





LS-DYNA versions

We have compared with the recent versions of LS-DYNA, we found that performance wise, the earlier version of LS-DYNA, such as R7.1.3 performance slightly higher than the more recent versions. As noted earlier in this study, the AVX2 binaries would performance better than the SSE2 counterparts. The difference appeared to be small, about 4% lower performance with R9.0.1 than R7.1.3 when both are using AVX2 executable.



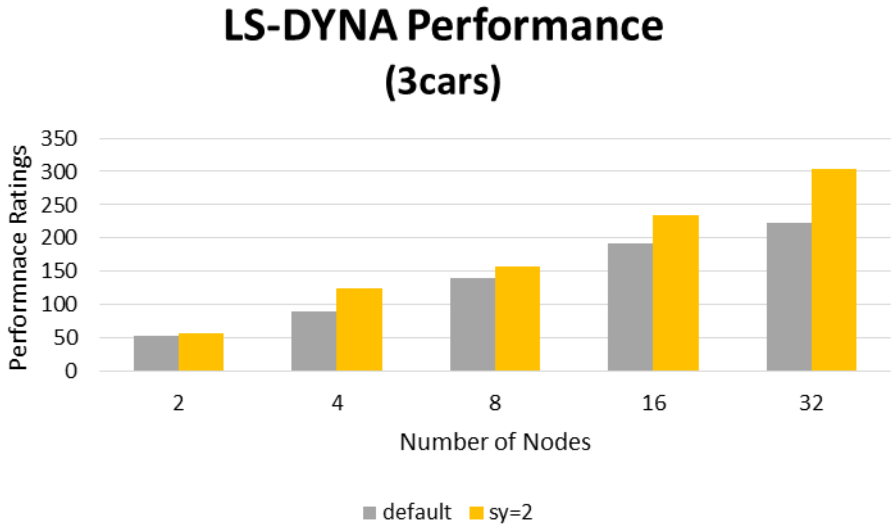
Domain Decomposition

As the performance and capability of a computational system improves, models with finer meshes are becoming more commonplace. To solve these larger models efficiently on a larger cluster due to the increase of computational time, the Massively Parallel Process (MPP) version of LS-DYNA has been improved to solve that.

The domain decomposition is an important aspect to achieve better performance when running LS-DYNA at scale. It works by breaking down a large problem into smaller piece, so each MPI process is responsible for its domain of the calculation of a bigger problem. The method that determines how the work is divided among the processes is called domain decomposition, and it would also affect how the processes would communicate with other processes which would general network traffic and CPU load among the MPI processes. The domain decomposition method would also determine if the MPI processes would perform the same number of computations and communication exchanges to finish at the simultaneously.
The keyword "decomp" in the pfile describes how the domain decomposition method is defined. By defining a decomposition method below in the pfile of the 3cars benchmark.

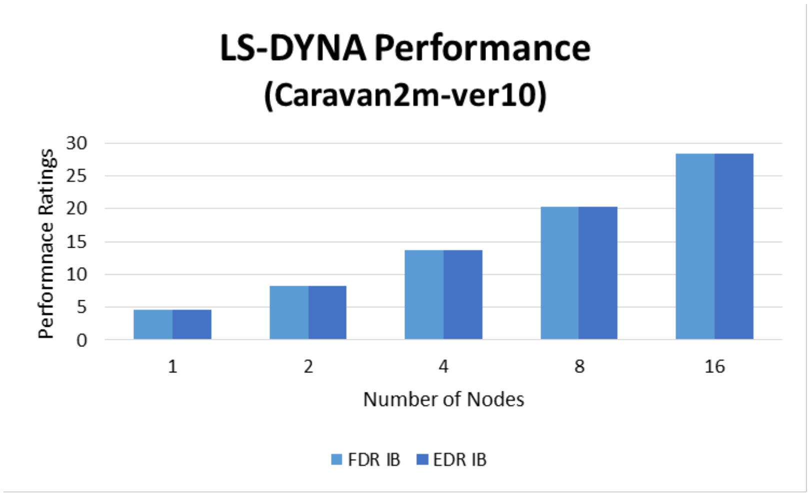
```
decomp { sy 2 }
```

We can observed an efficiency improvement by 36% at 32 nodes (896 cores). Scalability is improved as the workload is better distributed among the MPI processes.



Cluster Network Interconnect

The communication network used in MPI communications is important factor to LS-DYNA scalability. Here we are comparing between EDR InfiniBand versus FDR InfiniBand. We conduct the tests on the same InfiniBand infrastructure, to create an environment where the difference are due to the bandwidth on the InfiniBand connections. To test for the FDR speed, we lower the speed on the 100Gbps EDR InfiniBand switch to run at 56Gbps FDR rate 56Gbps to conduct this comparison.



The results in the graph shows that with the reduced speed in FDR run, we notice the same performance between LS-DYNA runs on EDR and FDR speeds. When we measure the type of MPI communications in an MPI profile, we do not noticed usage of large messages which typically would be beneficial to use EDR InfiniBand. It is safe to conclude that LS-DYNA is to network bandwidth sensitive.

It is important to note that the scalability performance impact of InfiniBand comes not only from the speed, but also from the architecture designs of different generations of the InfiniBand adapters and switches, and the implementation in software support and drivers. If we use an older generation of FDR InfiniBand hardware for this comparison, the scalability performance of FDR InfiniBand is very likely to be affected.

MPI Tuning

In order to get better performance on MPI performance, some of the MPI tuning parameters were added which helps getting best scalability performance. The neon_refined_revised is having the highest percentage of time spent in MPI communications, compared to the overall runtime. We use the neon_refined_revised case is the most network sensitive case.

By using the UD transport supported in the HPC-X MXM and memory optimization helps reducing network transfer overhead. Other tuning parameters include turning off support for HCOLL to reduce some overheads, and increasing the range of messages that can use ZCOPY within Mellanox MXM point-to-point communication library.

```
-mca coll_hcoll_enable 0 -x MXM_SHM_RNDV_THRESH=32768 -x
MXM_ZCOPY_THRESH=inf -x MXM_UD_HARD_ZCOPY_THRESH=inf -x
MXM_UD_MSS=8mb
```

Conclusions

In this article, we have identified few areas that are relevant to LS-DYNA performance.

- Skylake generation performs better than Haswell generation due to the additional memory channels, which has a direct impact on LS-DYNA performance
- The performance gain of switch from 2400MHz to 2666MHz DIMMs is 2% on a single node
- SNC provides 3% of benefits on a single node
- AVX2 executable performs better, compared to executables with SSE2 and AVX-512 instructions
- R7.1.3 executable performs better than newer LS-DYNA releases, by about 4%
- Mellanox HPC-X and Intel MPI both perform better than Platform MPI at scale
- EDR and FDR InfiniBand performs generally the same up for all node counts being tested
- Domain decomposition method can have significant impact on LS-DYNA scalability