LS-DYNA® Performance on Intel® Scalable Solutions

Nick Meng, Michael Strassmaier, James Erwin, Intel
nick.meng@intel.com, michael.j.strassmaier@intel.com,
james.erwin@intel.com
Jason Wang, LSTC
jason@lstc.com

Abstract

Along with the Intel® Purley platform launch, a series of cost-effective products such as the Intel® Xeon® processor Scalable Family (formerly code-named Skylake-SP), Intel® Omni-Path Architecture fabric, Intel® SSDs, Intel® MPI 2018 library, and Intel® Math Kernel Libraries (MKL) have been released in 2017. In this paper we study and evaluate the impact of these Intel products on LS-DYNA application. Numerous factors affect application performance and must be investigated and understood to ensure top performance and value to our customers. Intel has characterized LS-DYNA Explicit and Implicit scalability performance through extensive benchmarking and has determined the optimal factors to be considered for the Intel® Omni-Path Architecture fabric, Intel® MPI, Intel MKL and the Skylake-SP processor.

Introduction

The typical problem size of LS-DYNA explicit model in automobile industry has recently increased to more than ten millions elements. More and more we have observed automakers increasing the number of cores used for LS-DYNA simulation workloads to keep pace with these larger models. Scaling of LS-DYNA at the lowest cost and highest performance for customers is becoming more important as these systems grow in core counts and number of nodes. The process of improving application performance on single core and then scaling across many nodes is a major focus area for Intel®. The new round of Intel® products, such as the Intel® Xeon® processor Scalable Family (code named Skylake-SP), Intel® Advanced Vector Extensions 512 (Intel® AVX-512), Intel® Omni-Path Architecture fabric, Intel® Optane™ technology (bringing 3D XPoint™ memory to storage products), enhanced Fortran Compiler, Intel® MPI library, and Intel® MKL library with AVX-512 support contribute to the most effective integrated scalable solution to LS-DYNA customers both in cost and performance. In this paper, we highlight the advantages of Intel solutions, demonstrate benchmark results and share performance tips.

Intel® Xeon® Processor Scalable Family

Intel® uses a tick-tock model associated with its generation of processors. Major microarchitecture changes take place on a "tock," while minor microarchitecture changes and a die shrink occur on a "tick". The Intel® Xeon® processor Scalable Family (formerly code named Skylake-SP) is a "tock" based on 14nm process technology and on the Purley platform introduces many innovative features compared to the previous-generation Intel® Xeon® processors products [1]. These features include higher number of processor cores with mesh-based microarchitecture, increased memory bandwidth, non-inclusive cache, Intel® Advanced Vector Extensions 512 (Intel® AVX-512)[1], Intel® Memory Protection Extensions (Intel® MPX) [1], Intel® Ultra Path Interconnect (Intel® UPI) [1], and sub-NUMA clusters.

On previous generations of Intel® multi-core Xeon® processors, known as Haswell and Broadwell (Grantley platform), the processors, the cores, last-level cache (LLC), memory controller, IO controller and inter-socket Intel® QuickPath Interconnect (Intel® QPI)[1] ports are connected together using a ring-based architecture. As the number of cores on the CPU increased with each generation, the memory access latency has also increased and available bandwidth per core diminished. The Intel® Xeon® processor Scalable Family introduces a mesh-based microarchitecture to mitigate the increased memory latencies and bandwidth constraints associated with previous ring-based microarchitecture. The mesh-based microarchitecture (Figure 1) encompasses an array of vertical and horizontal communication paths allowing traversal from one core to another through the shortest path (hop on vertical path to correct row, and hop across horizontal path to correct column).

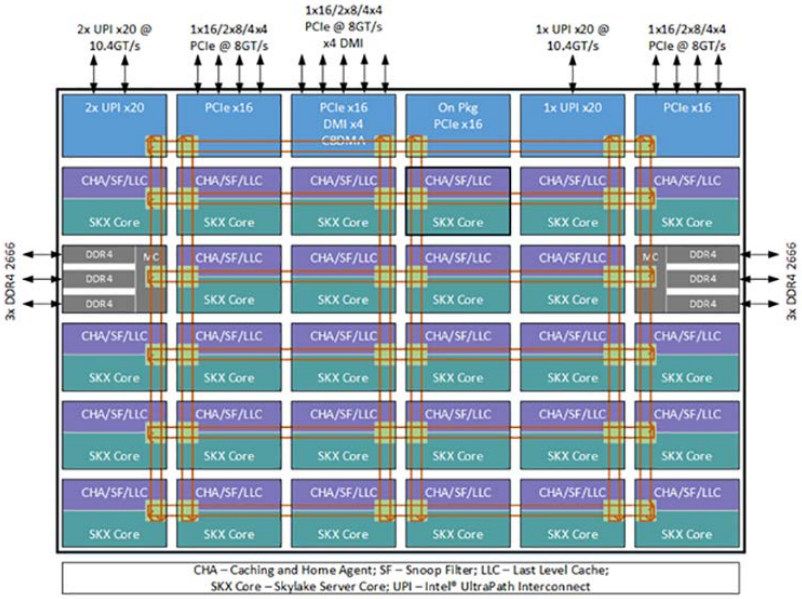

Figure 1*. Intel Xeon Skylake-SP mesh-based microarchitecture*

In addition, the previous generation of Intel® Xeon® processors utilized Intel® QPI®, which has been replaced on the Intel® Xeon® processor Scalable family with Intel® UPI. The Intel® UPI (Figure 2) is a coherent interconnect for scalable systems containing multiple processors in a single shared address space. Intel® Xeon® processors that support Intel UPI, provide either two or three Intel® UPI links for connecting to other Intel® Xeon® processors and do so using a high-speed, low-latency path to the other CPU sockets. Intel® UPI uses a directory-based home snoop coherency protocol, which provides an operational speed of up to 10.4 Giga Transfers per second.

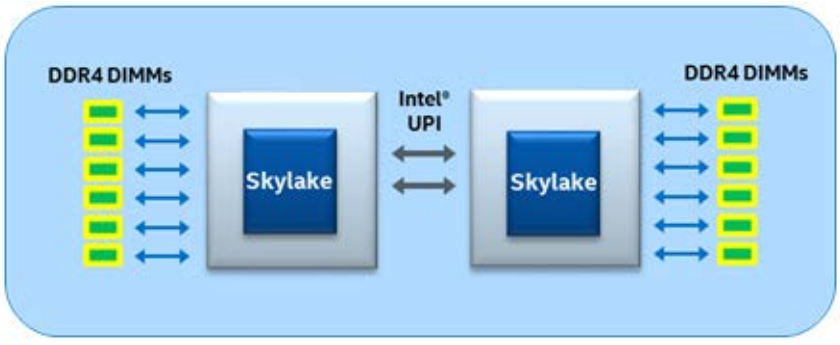

Figure 2*. Two socket configuration with Intel UPI*

The Intel® Xeon® processor Scalable Family also includes six memory channels supporting higher clock speed of DDR4 memory (up to 2667 MHz) compared to four memory channels in Broadwell, providing over 50% improvement in memory bandwidth, useful for HPC applications such as LS-DYNA.

The Intel® Xeon® processor Scalable Family (known as Skylake-SP) introduces new Intel® AVX-512 instruction groups (AVX512CD, AVX512F, AVX512BW, AVX512DQ, and AVX512VL) [1]. Intel® AVX-512 is an extension to the CPU vector extensions, which includes 32 registers each 512-bit wide and eight dedicated mask registers which help in efficient AVX-512 instruction generation for branchy codes. Intel® AVX-512 also doubles the width of the register compared to its AVX2 predecessor in Broadwell, thus computationally doubling the single and double precision floating point operations per clock cycle on Skylake-SP. A comparison of register widths for SSE, AVX/AVX2 and AVX-512 is shown in Figure 3.

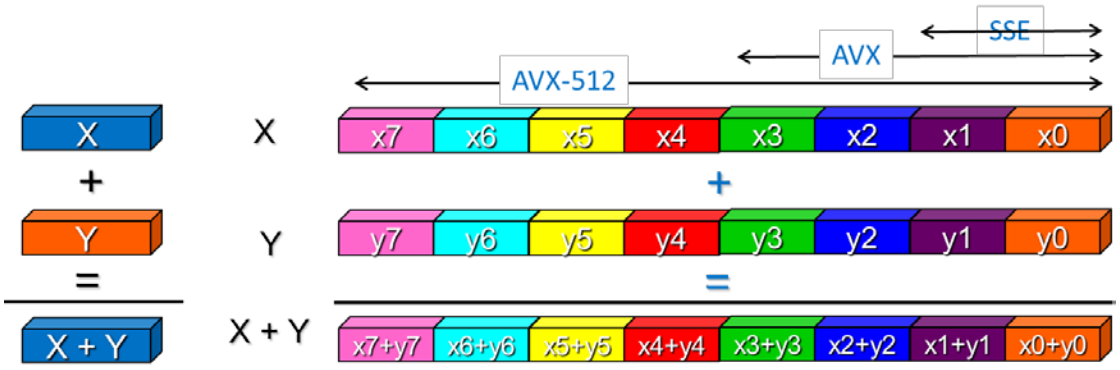

Figure 3. *Intel Advanced Vector Extensions technology*

The important features discussed above, such as Skylake-SP's mesh-based microarchitecture, faster Intel® UPI links, and increased number of memory channels (to six), and AVX-512 provide significant generational performance improvements to large HPC applications such as LS-DYNA on Intel processors as we discuss below.

Single core performance improvement of LS-DYNA on Skylake-SP processor

Figure 4 demonstrates the generational performance LS-DYNA users will see when running on Intel® Xeon® Gold 6148 processor (code named Skylake) in a 2-socket system (2 x 20 cores @ 2.4Ghz) vs. Intel® Xeon® E5-2697v4 processor (code named Broadwell) in a 2-socket system (2 x 18 cores @ 2.3Ghz). Running LS-DYNA with car2car standard benchmark using 40 MPI ranks on Skylake vs. 36 MPI ranks on Broadwell shows 1.47x speedup. Note that accounting for core count (40 vs. 36) and frequency (2.4GHz vs. 2.3GHz) changes one would expect only a 1.16x speedup. But we see a significantly more 1.47x speedup due to improved micro-architecture features in Skylake, increased memory bandwidth and use of Intel® new AVX-512 instructions in LS-DYNA.

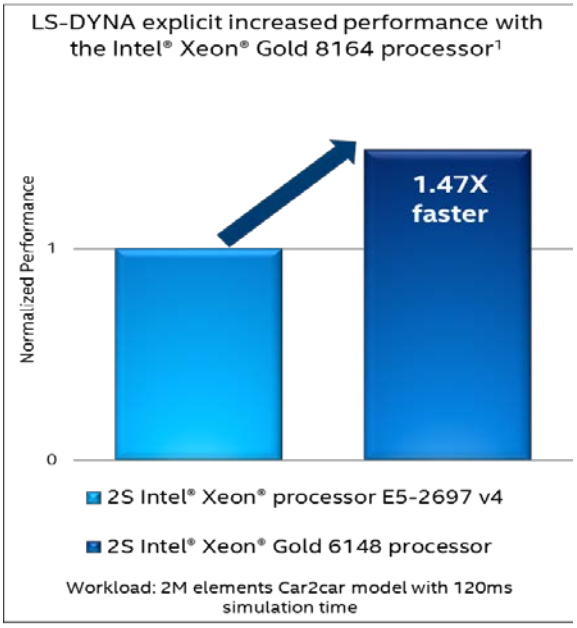

Figure 4. *LS-DYNA MPP explicit performance comparison*

(System configuration: 2S x 20C Intel® Xeon® Gold 6148 processor and 192GB DDR4@2667Mhz and single 800GB SSD. 2S x 18C Intel® Xeon® Processor E5-2697 v4 and 128GB@2400Mhz DDR4, and single 400GB SSD)

Intel® and LSTC® continue to optimize LS-DYNA R9.0 and R9.1 leveraging SSE, AVX2 and AVX-512 instructions for single core / single node performance. A key focus area has been to make source changes to leverage Intel compiler to vectorize performance critical hot loops to generate appropriate vector instructions (for example using –xCORE-AVX512 compiler option for Skylake) for best performance. As seen in Figure 5, LS-DYNA R9.0_SP EXPLICIT improved up to 1.25x using AVX-512 on Skylake, and we have continued to improve LS-DYNA R9.1_SP EXPLICIT as well.

LS-DYNA MPP R9.0_SP EXPLICIT (original code) 40 cores run on Gold 6148 system	**SSE2 binary**	**CORE AVX-512 binary**
Performance ratio of 3cars/150ms model	1.00	1.15
Performance ratio of Car2car/10ms model	1.00	1.25
LS-DYNA MPP R9.1_SP EXPLICIT (optimized code) 40 cores run on Gold 6148 system	**SSE2 binary**	**CORE AVX-512 binary**
Performance ratio of 3cars/150ms model	1.07	1.25
Performance ratio of Car2car/10ms model	1.04	1.32

Figure 5*. LS-DYNA R9.0 and R9.1 explicit performance comparison between SSE2 and AVX-512*

(System configuration: 2S x 20C Intel® Xeon® Gold 6148 processor and 192GB DDR4@2667Mhz and single 800GB SSD, Intel® Turbo Boost enabled, Intel® Hyper-Threading Technology disabled. 40 MPI ranks run with Intel 2018 MPI update 1, Linux release 3.10.0-693.el7.x86_64)

A key optimization strategy for HPC applications is to leverage use Intel MKL/BLAS library where possible. Intel® MKL is optimized for SSE, AVX/AVX2 and AVX-512, and can dynamically use the appropriate vector instruction set at run time depending on the x86 processor it runs on. We have linked LS-DYNA R9.1 and later versions with Intel MKL and as shown in Figure 6, LS_DYNA HYBRID R9.1_DP IMPLICIT takes advantage of AVX-512 vector performance up to 1.17x on Skylake.

LS-DYNA HYBRID R9.1_DP IMPLICIT 16 MPI x 8 OMP run on 8 Gold 6148 nodes/w OPA	CORE AVX2 binary	CORE AVX-512 binary
Performance ratio of 2.5M DOFs implicit model	1.00	1.17

***Figure 6**. LS-DYNA R9.1 implicit performance comparison between AVX2 and AVX-512*

(System configuration: 8 Skylake nodes cluster with OPA network, 2S x 20C Intel® Xeon® Gold 6148 processor and 192GB DDR4@2666Mhz and single 800GB SSD, Intel® Turbo Boost enabled, Intel® Hyper-Threading Technology disabled. 2 MPI ranks x 8 OMP threads per node, Intel® MPI 2018 build 20170713. I_MPI_FABRICS=shm:tmi. Intel® Omni-Path Architecture (Intel® OPA): Intel Fabric Suite 10.5.1.0.2. Intel Corporation Series 100 Host Fabric Interface (HFI), Series 100 Edge Switch – 48 port)

While the above improvements are focused on standard LS-DYNA benchmarks, Intel continues to profile performance hotspots using customer proprietary workloads, and optimize LS-DYNA for the broader user base.

Cluster improvements on LS-DYNA using Intel® Omni-Path Architecture

The Intel® Omni-Path Architecture (Intel® OPA) provides the highest performing network fabric for HPC applications. It is a low-cost solution for entry-level to large supercomputer applications and is the most scalable network fabric with the ability to span over many thousand nodes. Our internal testing on LS-DYNA highlights the improved performance with HPC scaling on OPA versus InfiniBand* EDR* (Figure 7). Starting at 16 nodes, Intel® OPA begins to scale better than EDR, and up to as much as 28% better with the car2car benchmark at 32 nodes.

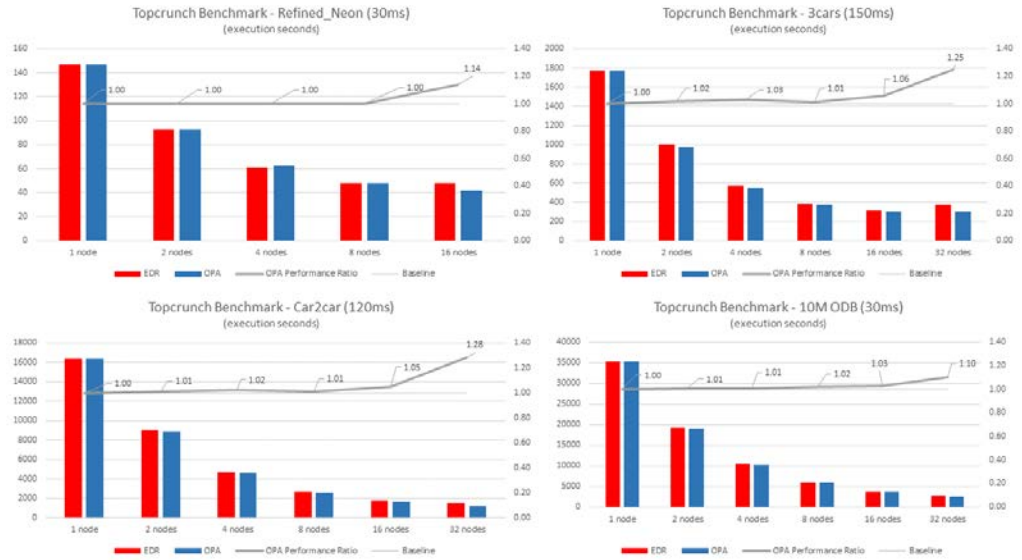

***Figure 7**. LS-DYNA scaling on OPA versus InfiniBand EDR*

(System configuration: 32 Skylake cluster with OPA/EDR network, dual sockets Skylake system with 2 x 20C Intel® Xeon® Gold 6148 processors and 192GB DDR4@2666Mhz and single 800GB SSD, Intel® Turbo Boost enabled, Intel® Hyper-Threading Technology disabled. 40 MPI ranks per node, Intel® MPI 2018 build 20170713. I_MPI_FABRICS=shm:tmi. Intel® Omni-Path Architecture (Intel® OPA): Intel Fabric Suite 10.5.1.0.2. Intel Corporation Series 100 Host Fabric Interface (HFI), Series 100 Edge Switch – 48 port.)

LS-DYNA scales easily because Intel® OPA has been designed from the ground-up to provide users with extreme network performance when running their HPC applications. One of the critical software components of Intel® OPA is known as Performance Scaled Messaging (PSM) and was first introduced with the Intel® True Scale fabric. It has since been further refined and optimized for Intel® OPA with release of the second generation, known as PSM2. PSM2 is semantically matched to MPI, unlike the Verbs interface which is an additional software layer between the network hardware, InfiniBand RDMA Message Transport Service, and finally the user API such as Message Passing Interface 1. In the case of Intel® OPA, the only software layer between the Host Fabric Interface (HFI) driver and MPI is PSM2. PSM2 allows users writing their applications according to the MPI standard to utilize all of the enhanced features of Intel® OPA without having to understand the lower level detail of the architecture [4,5], including an LS-DYNA specific tuning [6] which has become a default setting for Intel® OPA driver software.

Three critical components to the network performance are MPI latency, message rate, and collectives performance. An 8 byte message can be sent from one compute node to another, through an Intel® OPA 48-port switch, from MPI to MPI rank in as little as 0.94 microseconds (940 nanoseconds) [3]. This low latency includes features such as Packet Integrity Protection (PIP), which is a built-in error detection and correction mechanism. PIP has zero latency penalty for detecting errors in the fabric. This low latency and efficient error detection and correction also enables high message rates. Up to 158 million 8-byte messages per second can be sent between two dual sockets Intel® Xeon® Platinum 8180 processor nodes in one direction, and up to 249 million messages per second simultaneously in both directions [3]. Message rates this high mean that individual MPI ranks in the LS-DYNA application can efficiently communicate without the network being a limiter to performance. Both low latency and high message rates contribute to excellent collective performance, as recently demonstrated on a cluster of up to four thousand nodes [4]. An MPI collective is when a message is collectively sent throughout the entire application using highly tuned and efficient algorithms, such as those existing in the latest Intel MPI libraries. The user does not need to understand the details of the collective communication pattern - when coupled with Intel® OPA, the Intel® MPI library ensures the lowest latency is achieved for a wide range of collectives.

Low latency and high message rates contribute to excellent collective performance, such as what we have demonstrated on a cluster of up to four thousand nodes [4]. An MPI collective is when a message is collectively sent throughout the entire application using highly tuned and efficient algorithms, such as those existing in the latest Intel® MPI libraries. The user does not need to understand the details of the collective communication pattern. When coupled with Intel® OPA, the Intel® MPI library ensures the lowest latency is achieved for a wide range of collectives. Figure 8 demonstrates the ability of Intel® OPA to scale better than expectations up to 512 dual socket Intel® Xeon® Gold 6142 processor nodes. The data shows scaling for MPI ALLREDUCE and Broadcast (Bcast), which are two important collectives in LS-DYNA. The 512 node runs involve the communication across 16,384 MPI ranks in under 24 microseconds for an 8-byte ALLREDUCE, and under 8 microseconds for an 8-byte broadcast. The data shows scaling for MPI ALLREDUCE and broadcast, which are two important collectives in LS-DYNA. The 512 node runs involve the communication across 16,384 MPI ranks in under 24 microseconds for an 8-byte ALLREDUCE, and under 8 microseconds for an 8-byte broadcast.

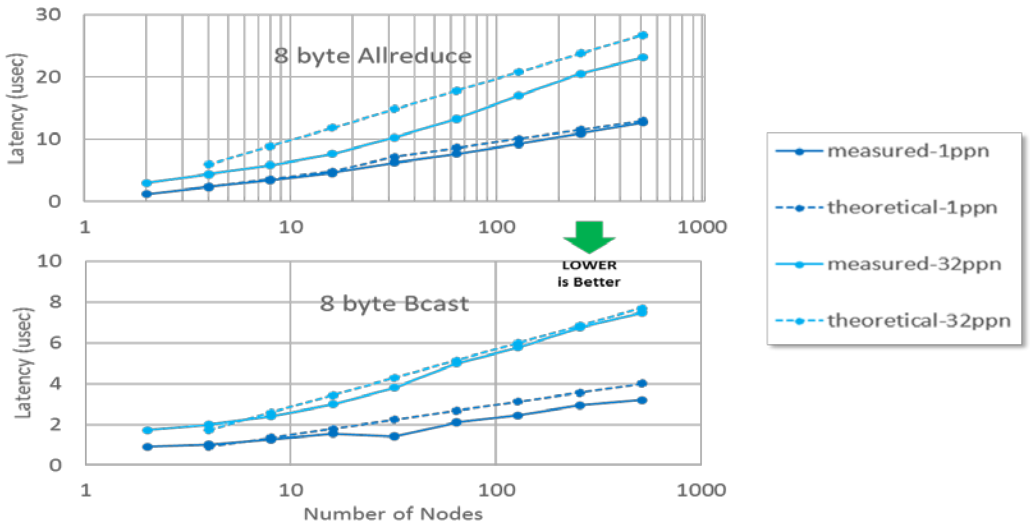

Figure 8. Scaling collectives up to 512 Intel® Xeon® Gold 6142 processor nodes (ppn is number of MPI ranks per node)

(System configuration: Intel® Xeon® Gold 6142 processor, Intel® Turbo Boost and Intel® Hyper-Threading Technology disabled. 1 and 32 MPI ranks per node. Intel MPI Benchmarks 2017, built with Intel® Parallel Studio XE 2018 . Intel® MPI 2018 build 20170713. I_MPI_FABRICS=shm:tmi. Intel® Omni-Path Architecture (Intel® OPA): Intel Fabric Suite 10.5.1.0.2. Intel Corporation Series 100 Host Fabric Interface (HFI), Series 100 Edge Switch – 48 port. Scientific Linux release 7.3 , 3.10.0-514.32.3.el7.x86_64. Theoretical scaling is computed as: Log2(Nn) for Allreduce and Barrier, Log4(Nn) for Bcast, and O(Nr) for Alltoall, where Nn is number of nodes and Nr is number of ranks. Switch hop correction added to 1ppn scaling, where hopcorrection= (nhops-1)*Nsteps*0.12usec/hop , nhops=1 for nodes<=16, nhops=3 for nodes > 16.)

Optimization tips for LS-DYNA

We have discussed the advantage of Intel® Xeon® Skylake-SP microarchitecture, Omni-Path architecture, and Intel® Compiler and Math Kernel Library. Now we will focus on the benefit LS-DYNA will observe through optimal settings across the Intel Integrated Scalable solution.

First, Intel® has optimized the LS-DYNA application with AVX2 & AVX512 instruction using the Intel® Fortran Compiler and Intel Math Kernel Library and LS-DYNA users will gain benefit from this work. Intel® also works with customers and LSTC closely to optimize specific routines for further performance improvements when users build models with specific features which Intel optimization work hasn't covered. As an example, we found a complex loop with COSH function in a hot material routine difficult to vectorize and now we have this function vectorized with the SVML function provided by Intel® FORTRAN compiler. There is 56% performance improvement in the material routine and 2.3% performance improvement in whole job.

Secondly, optimal domain decomposition is an important factor in LS-DYNA scalability and customers have benefitted from this feature. In addition, adjusting the collective algorithm in Intel MPI library and adjusting parameter in Omni-Path driver may help customers gain better scalability on the Intel® Scalable Solution. As an example, when I_MPI_ADJUST_BCAST is set to 1 and I_MPI_ADJUST_ ALLREDUCE is set to 5, many users have observed optimal scalability with these settings while using the Intel® MPI library. In the Omni-Path driver, we observed positive results by adjusting the eager_buffer_size and rcvhdrcnt parameter for LS-DYNA application. With this change, LS-DYNA users who do simulation with more than 16 nodes will see more than 43 % improvement on the Omni-Path network.

Finally, performance per core is still important for many automobile customers. Intel® provides multiple Xeon® Gold processor choices to customers that match workload requirements, such as the Gold 6142, 6418, 6154 and 6137 processors.

Conclusions

Intel® continues to optimize LS-DYNA R9.2, R10, and later version using the Intel® FORTRAN compiler and Intel Math Kernel Library to leverage AVX2 and AVX512 instructions, and these improvements are available in LS-DYNA R9.1.

Use of Intel® MPI and Intel OPA cluster provide great cluster scalability options for LS-DYNA user. Intel MPI and OPA driver stack include many tunable parameters to improve cluster scaling. Intel MPI library provides an optimal configuration file (e.g lsdyna.conf) as option for LS-DYNA customers. Optimal OPA parameters have been added into newer version of Intel® Omni-Path driver.

Intel® provides a range of core counts and price points in the Intel® Xeon® processor Scalable Family (Platinum, Gold and Silver) for LS-DYNA users to consider based on their workload requirements.

References

1. https://software.intel.com/en-us/articles/intel-xeon-processor-scalable-family-technical-overview
2. Grun, Paul, "Introduction to InfiniBand™ for End Users," InfiniBand Trade Association, 2010. http://www.mellanox.com/pdf/whitepapers/Intro_to_IB_for_End_Users.pdf
3. Intel® Xeon® Platinum 8180 processor, 2.50 GHz 28 cores, 64 GB 2666 MHz DDR4 memory per node, 3.10.0-514.el7.x86_64 kernel. Dual socket servers with one switch hop using 2 meter copper cables. Intel® Turbo Boost Technology enabled, Intel® Hyper-Threading Technology enabled. Intel MPI Benchmarks 4.1. Intel® OPA: 1. Open MPI 1.10.4-hfi as packaged with IFS 10.3.1.0.22, 2. Intel MPI version 2017.3.196 with ofi fabric (libfabric-1.2.0). RHEL 7.3. hfi1.conf: krcvqs=4 eager_buffer_size=8388608 max_mtu=10240.
4. R. B. Ganapathi, A. Gopalakrishnan, R. W. McGuire, "Mpi process and network device affinitization for optimal hpc application performance," in 2017 IEEE 25th Annual Symposium on High Performance Interconnects (HOTI), Aug 2017, pp. 80-86.
5. R. B. Ganapathi, A. Gopalakrishnan, R. W. McGuire, "HPC Process and Optimal Network Device Affinitization," paper pending.
6. Ravi, V., Erwin, J., Sivakumar, P., "Host Software Stack Optimizations to Maximize Aggregate Fabric Throughput," IEEE Hot Interconnects, 2017.