A Study of MPP LS-DYNA® Performance on Hardware Configurations

Yih-Yih Lin and Toshihiro Ishibashi
Hewlett Packard Enterprise

Abstract

With vehicle crash simulation models of sizes up to 10M elements, this paper will investigate how LS-DYNA performance is affected by the four hardware components: processor, I/O, memory, and interconnect network. First, two aspects of processor will be studied: performance gain from Intel Turbo Boost and performance gain from the AVX2 over the SSE2 instruction set architecture. Second, performances of using a local storage, a shared NFS file system, and a shared LUSTRE file system will be measured. Third, three aspects of memory will be studied: DIMM's frequency, cache coherence snooping modes, and balance-ness of memory configurations. And fourth, performances of two interconnect network switches will be compared: Mellanox IB-FDR and IB-EDR.

Introduction

MPP LS-DYNA performance is affected by four hardware components: processor, I/O, memory, and interconnect network. In this paper, we will study how factors in these four hardware components affect performance on vehicle crash simulation.

First, we will investigate two aspects of processor: Intel Turbo Boost and Advanced Vector Extensions (AVX). Turbo Boost, enabled by a BIOS setting, is a technology implemented by Intel in Nehalem and later processors. When it is enabled, Turbo Boost allows processor cores to run at frequencies higher than the rated operating frequency via dynamic control of cores' clock rates. SSE2, first introduced by Intel in 2001, is a processor supplementary instruction set that is intended to accelerate operation on structured data, also called vector data. To further accelerate vector operations, Intel started to support the AVX instruction set with the release of Xeon E5-2600v2 processors and the AVX2 instruction set with the release of E5-2600v3 processors.

Second, we will compare performances among using local storage, a shared NFS (Network File System), and a shared LUSTRE file system for I/O in MPP LS-DYNA.

Third, we will examine three aspects of memory performance: (1) performance comparison between two DIMMs with different frequencies; (2) performance comparison between the two snoop modes, Home Snoop and Cluster on Die Snoop; and (3) how the balance-ness of memory affects performance.

And fourth, we will compare performances of two different interconnect-network switches: Mellanox's IB-FDR and IB-EDR.

Unless otherwise noted, the following MPP LS-DYNA version, cluster, interconnect switch, and snoop mode were used:

- Version R8.0.0 SEE2 with Platform MPI
- A cluster consisted of nodes, each of which comprises two 16-core Xeon E5-2698 v3 processors
- IB-FDR
- Home Snoop

The following two vehicle crash models from www.TopCruch.org were used:

- The Car2Car model of 2.4M elements
- The ODB-10M model of 10M elements

Unless otherwise noted, the Car2Car job was a full 120ms simulation, and the ODB-10M job was a 80ms simulation. Furthermore, all cores in a node were fully used, and in figures below we will use the string "Cc Nn" to denote that a job uses all C cores in N nodes.

Turbo Boost and AVX2

MPP LS-DYNA is a CPU intensive application, and hence Turbo Boost will benefit MPP LS-DYNA performance by allowing all cores to operate at higher frequencies than the rated operating frequencies. Figure 1 shows that for the Car2Car model, Turbo Boost speeds up MPP LS-DYNA performance by 6 to 9 percent from 1 to 32 nodes. Turbo Boost's frequency upside and availability depends on the workload and the operating environment, including temperature and power consumption; consequently, performance may vary from job to job, from site to site, and from time to time, as also demonstrated by Figure 1.

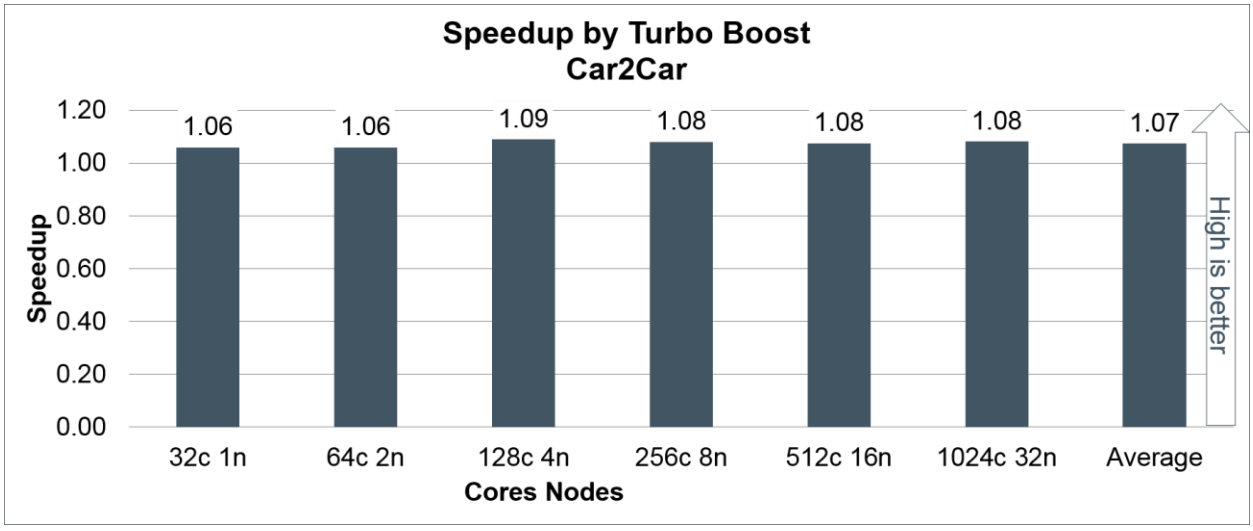

Figure 1

AVX achieves acceleration on vector operations by processing multiple vector data on multiple Arithmetic Logic Units within a single clock cycle. A large portion of data in MPP LS-DYNA is coded as vector data, and hence AVX will benefit MPP LS-DYNA performance. Figure 2 shows that for the Car2Car model, AVX2 speeds up MPP LS-DYNA performance by 9 to 14 percent from 1 to 32 nodes.

Figure 3 shows that for the Car2Car model, enabling both Turbo Boost and AVX2 speeds up MPP LS-DYNA performance by 11 to 20 percent. By comparing the speedups in Figures 1, 2, and 3, we note that the speedup by Turbo Boost and AVX2 are not additive, which is due to the fact that more power will be consumed when AVX2 is enabled.

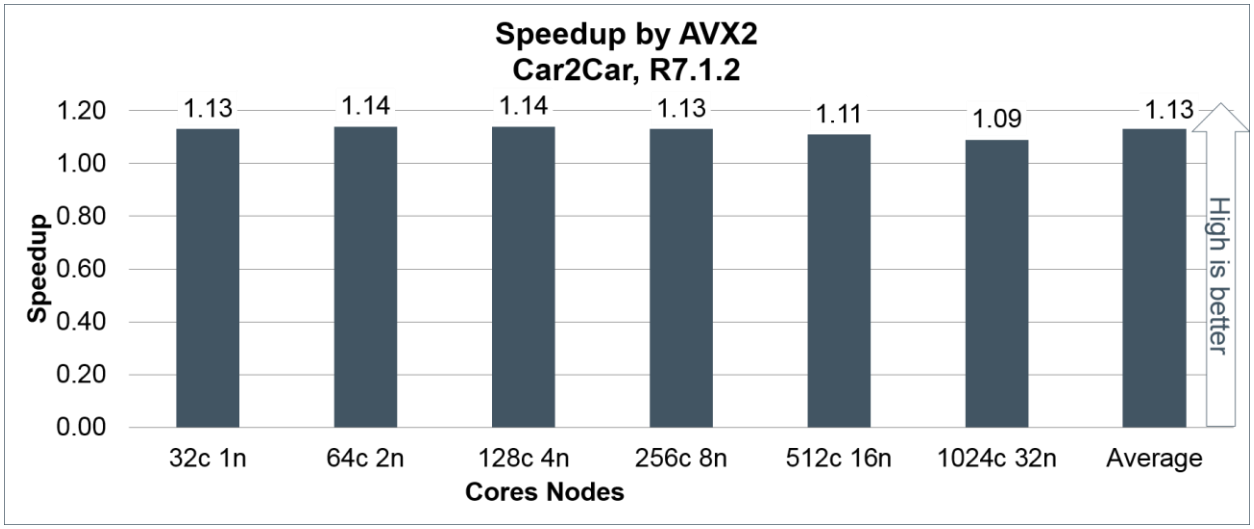

Figure 2

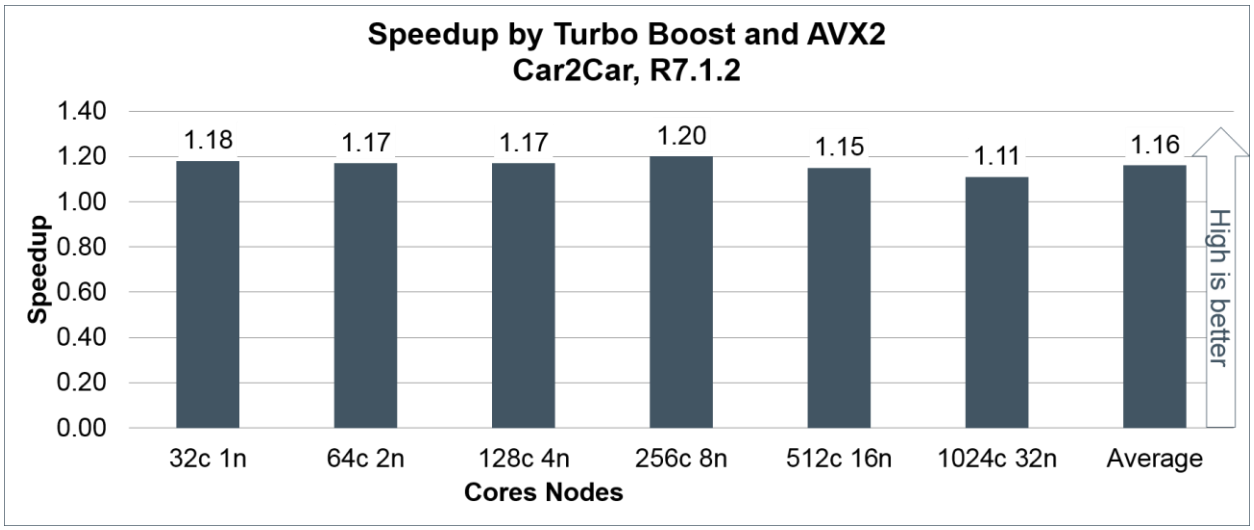

Figure 3

File Systems

MPP LS-DYNA outputs two kinds of files during its computing course: one called global, the other called local. Global files, like the plot files, are output by rank 0; local files, like the decomposition scratch files, are output by each rank. Paths to which the global files and the local files are output are called the global directory and the local directory respectively. It is generally recommended that for better performance that the local directory should be a local storage and that the global directory can be a shared file system. We have tested the validity of this recommendation and found it only partially valid. Figure 4 shows that if you have the slow shared file system NFS, you may be better off to use local storages for both the global directory and the local directory. However, if you have the fast, parallel LUSTRE file system, you can use it for both the global and the local directories without suffering performance loss. Figure 5 also shows that using LUSTRE for throughput of single-node jobs is on par with using local storages.

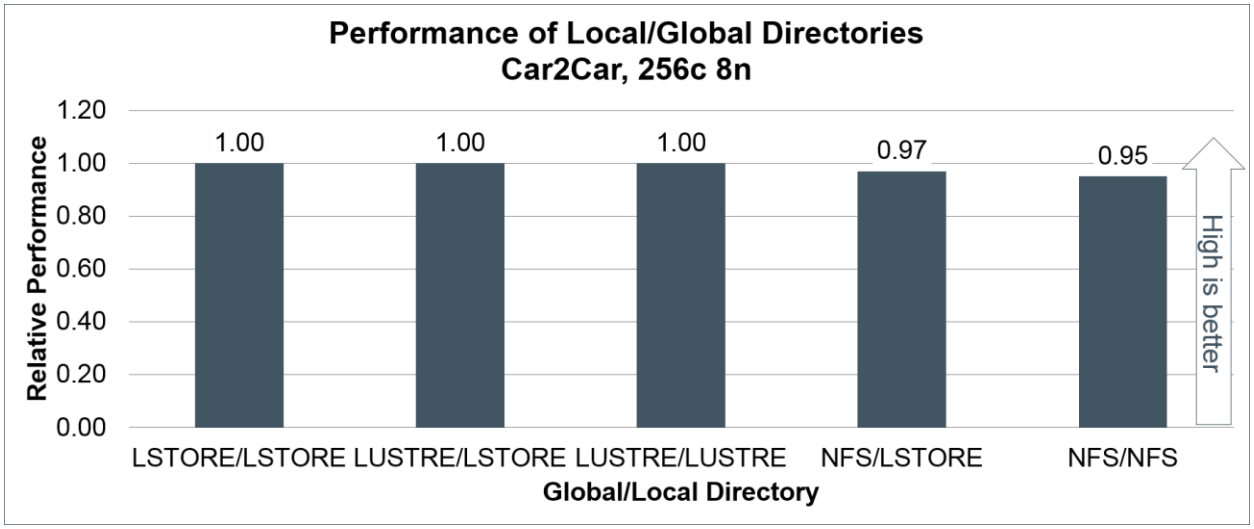

Figure 4

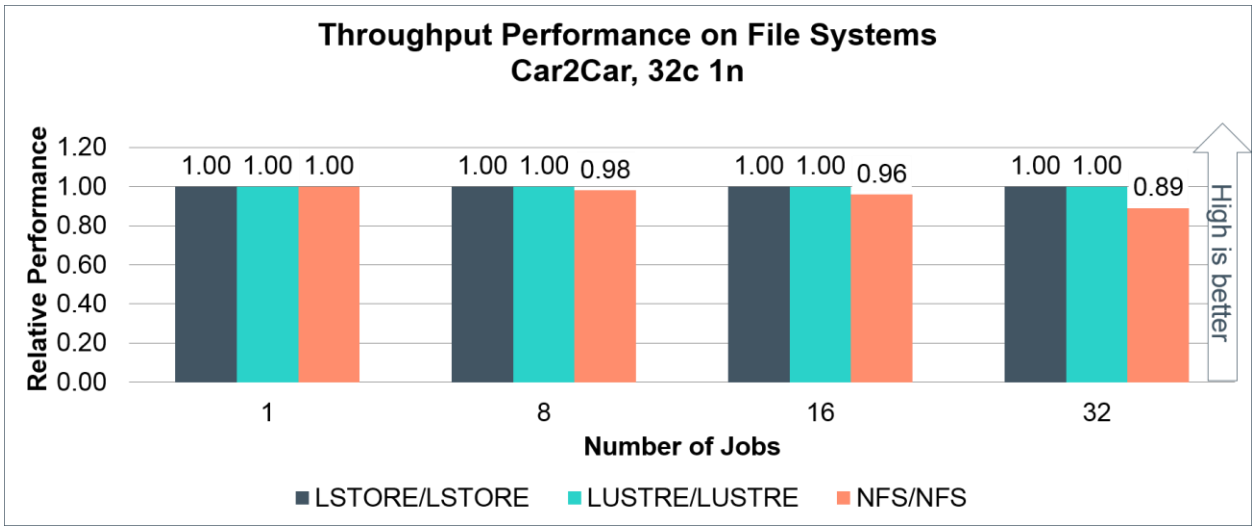

Figure 5

Memory: Frequency, Balance-ness, and Snoop Filter

The latest DRAM edition is DDR4 and comes with two clock speeds: 2133 MHz and 2400 MHz. Figure 6 shows that there is a performance advantage of 2 to 3 percent for MPP LS-DYNA by raising the clock speed by 12.5 percent. The processor used here for this comparison is the recently released Broadwell, whose memory system is very similar to that of Haswell.

Memory DIMMs come with different sizes, including 8 GB, 16 GB, and 32 GB. An E5-2600v3 processor has 4 memory channels, each of which has 3 DIMM slots. A server node, used in this study, comprises of two E5-2600v3 processors (sockets), and so it has a total of 8 channels. If all the 8 channels are installed with DIMMs of the same size, the memory configuration is called balanced; otherwise it is called unbalanced. It is important that customers have choices between a balanced and an unbalanced configuration: The former constrains the customer to a small set of node memory sizes, and the latter gives customers more choices in the memory capacity of a server node. In the following we will compare performances of various unbalanced memory configurations. Not to complicate the discussion, we will only consider configurations in which at most one DIMM is installed in every channel. If the four channels of Socket 1 are installed with a, b, c, and d-GB DIMMs, and those of Socket 2 are installed with e, f, g, and h-GB DIMMs, we will denote its memory configuration as {aG+bG+cG+dG,eG+fG+gG+hG}. If Socket 1 is installed with DIMMs of n-GB and Socket 2 with DIMMs of m-GB in all 4 channels, we will denote its configuration as {nGx4,mGx4}. Furthermore, if each of the two sockets is installed with the same sub-configuration, instead of repeating the sub-configuration we will simply append the string "x2S" after the right curly bracket to denote the configuration. For example, {16Gx4}x2S is another way to denote the configuration {16Gx4,16Gx4}.

Nodes used in this study are multiprocessor system with two sockets, each with a separate cache memory. As the same data from the main memory is often stored in the two cache memories at the same time, the memory system must work to maintain consistency between the two copies of data. Snoop filter is a mechanism to maintain this cache data consistency by "snooping other socket." Intel Xeon E5-2600v3 processor has provided three BIOS settings for snoop modes: Early Snoop (ES), Home Snoop (HS), and Cluster on Die Snoop (COD).

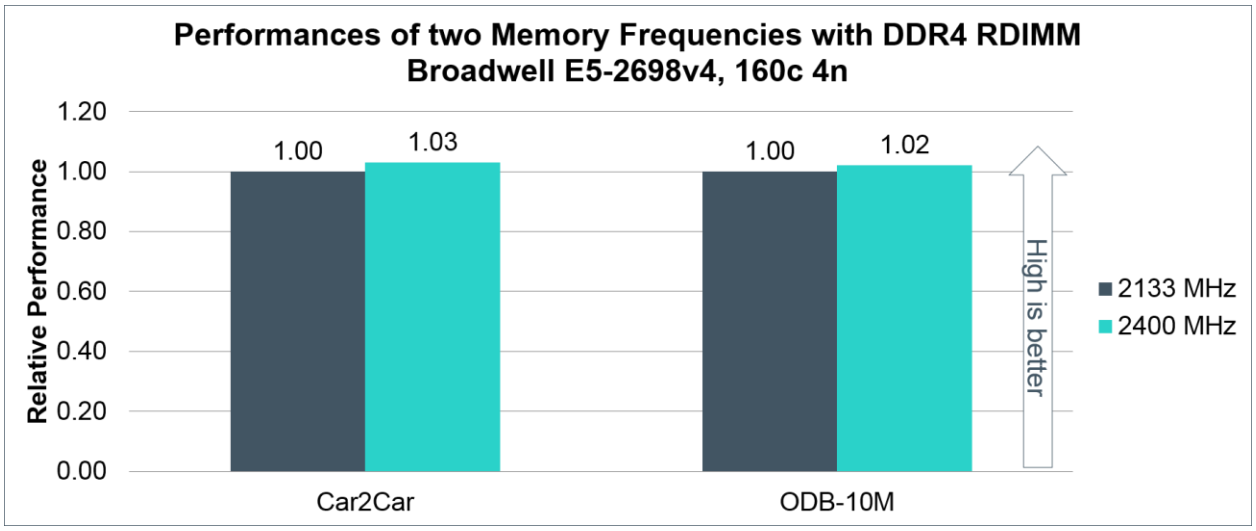

Figure 6

ES incurs severe penalty in bandwidth for accessing main memory that is asked by the remote processor, which is called in computer science as nonuniform memory access (NUMA). Each of the two sockets can be considered as to form a region for memory access, and thus each is often called a NUMA domain. Due to its severe penalty in bandwidth for NUMA access, ES should not be used for an application, such as MPP LS-DYNA, that is memory bandwidth sensitive. Therefore, we will not investigate it here.

The issue of choosing between HS and COD for MPP LS-DYNA is complicated and depends a lot on memory configurations. To gain insight on the issue, we have measured the performance of HS and COD, using a two 12-core E5-2690v3 node, with the following four memory configurations:

- The {16Gx4}x2S configuration, which is balanced
- The {16Gx4,8Gx4} configuration, which is unbalanced between sockets
- The {16G+16G+8G+8G}x2S configuration, which is unbalanced between memory channels in a socket
- The {16G+16G+16G+0G}x2S configuration, which is unbalanced in that one memory channel in a socket is unfilled.

Figure 7 shows that with the balanced memory configuration of {16Gx4}x2S COD has only a small performance advantage of 1 percent over HS for both Car2Car and ODB-10M models. For a general snoop setting for High Performance Computing, we recommend HS. The small advantage of COD shown here should not overwrite our recommendation because HS has been shown to gain advantages over COD in other applications.

Figure 8 shows that with HS, there is no performance difference between the balanced configuration of {16Gx4}x2S and the unbalanced configuration of {16Gx4,8Gx4}. This result shows that HS's performance is insensitive to the memory type in each NUMA domain as long as all channels in a NUMA domain are installed with the same type.

COD logically splits a socket into two NUMA domains and exposed them to the operating system so as to see two NUMA domains per socket. For the unbalanced configuration {16G+16G+8G+8G}x2S, the operating system in the COD mode decides naturally that the four memory channels in a socket comprise two NUMA domains, 16G+16G and 8G+8G, and thus handle memory traffic efficiently to give high memory bandwidth. This explains the result shown in Figure 9, which shows that for the unbalanced configuration {16G+16G+8G+8G}x2S COD has a performance advantage up to 1.23X over HS.

Figure 10 shows that if one of the memory channels in a socket is unfilled, both HS and COD underperform up to 17 percent, relative to a balanced configuration in which all 4 channels of a socket are filled. This result demonstrates that the performance advantage arising from COD's ability to see two NUMA domains in socket is lost if one of channels in a socket is unfilled.

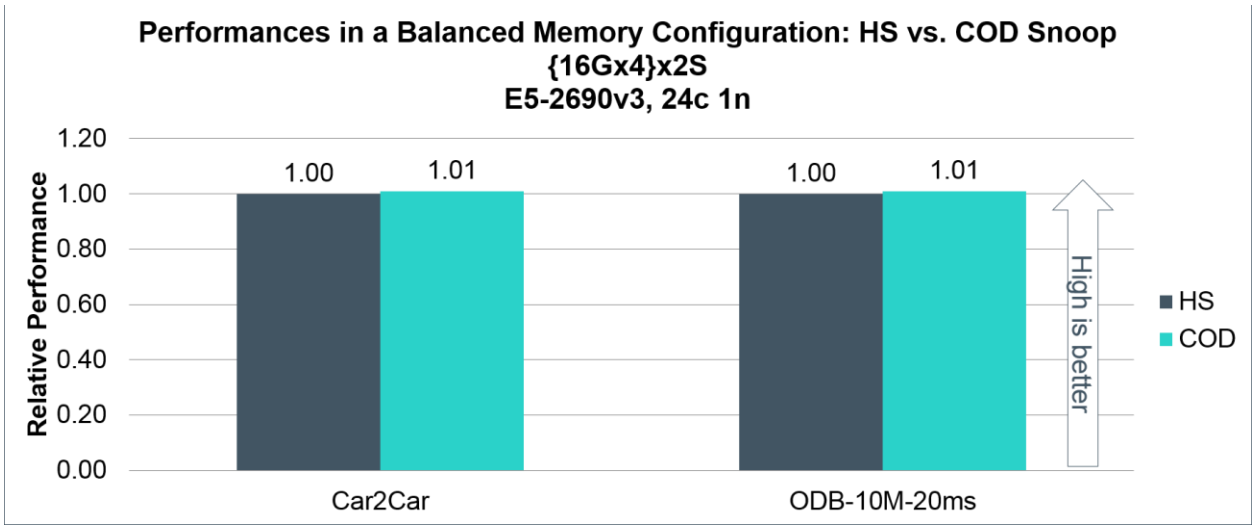

Figure 7

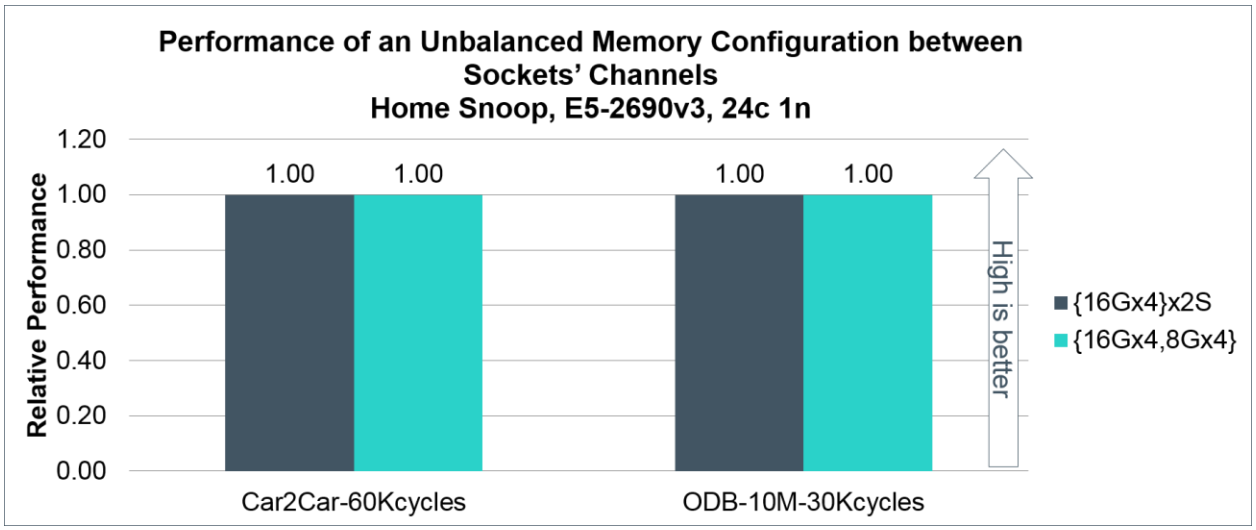

Figure 8

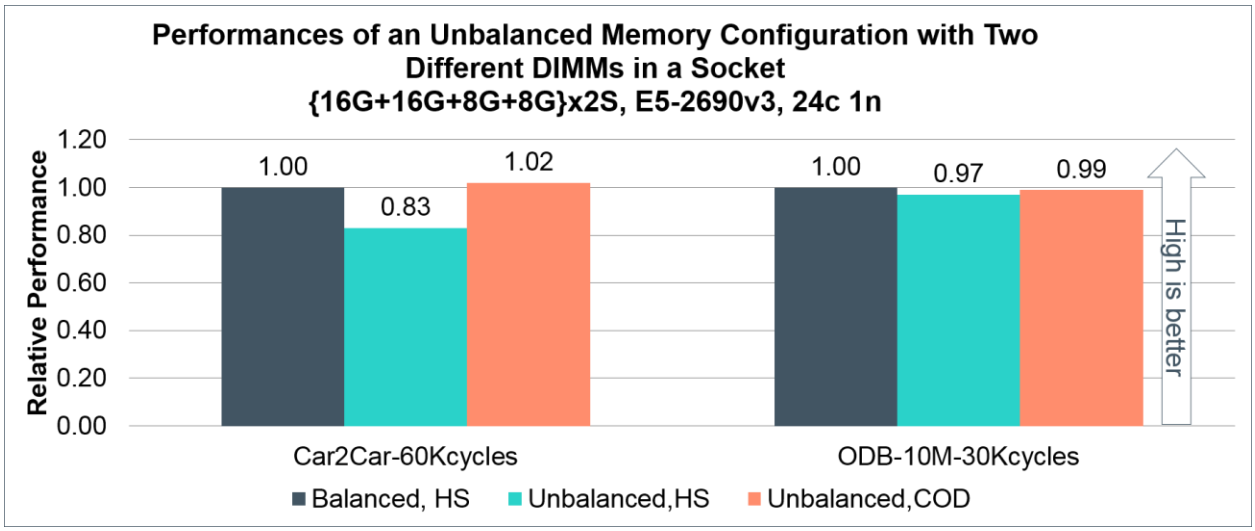

Figure 9

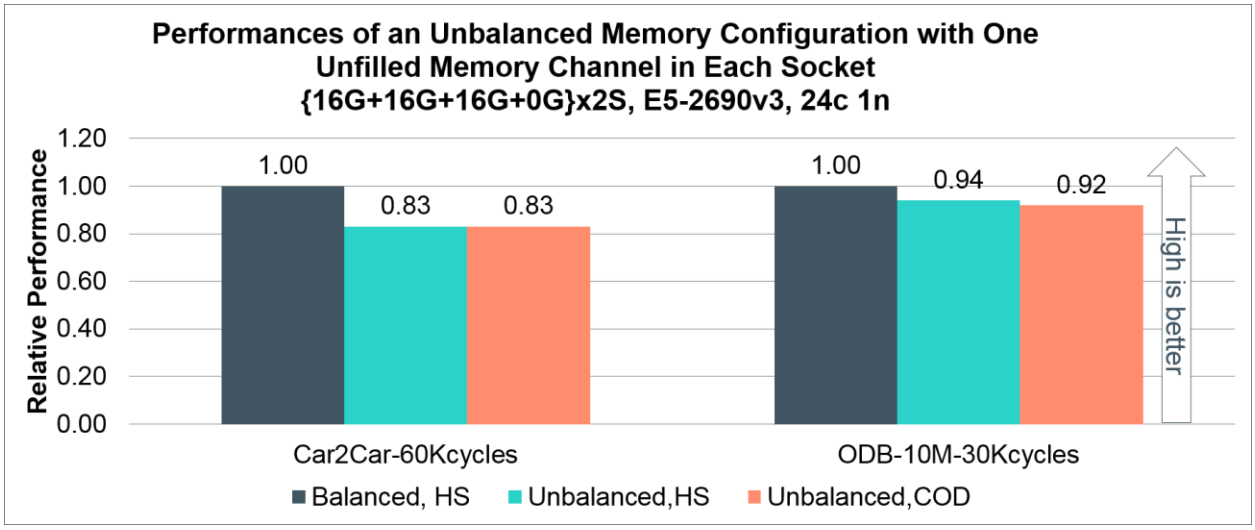

Figure 10

Interconnect Network

The total cost of an MPP LS-DYNA simulation can be divided into two parts: compute cost and communication cost. All the hardware components discussed so far contribute to compute cost. The only hardware component that contributes to communication cost is interconnect network. Interconnect speed is determined by its bandwidth and latency. The current generation of Mellanox's interconnect switch is ConnectX-4 IB-EDR, while its predecessor is ConnectX-3 IB-FDR. From Table 1, which shows the measured bandwidth and latency of these two networks, we can conclude that IB-EDR bandwidth is much faster than IB-FDR—up to 90 percent. However, this much improvement in bandwidth from IB-EDR does not translate too much performance gain, as shown in Figure 11. This result demonstrates that IB-FDR has sufficient network bandwidth to perform these workloads.

	Max Unidirectional BW (GB/sec)	Max Bidirectional BW (GB/sec)	Latency (nsec) with 1 switch hop	Switch hop latency (nsec)	Bidirectional Message rate
					32 ranks
Mellanox Connectx-3 IB-FDR	6.4	12.8	1090	170	78
Mellanox Connectx-4 IB-EDR	12.2	24.3	930	<90	124

Max BW are the highest measurements across all message sizes
Message rate is the maximum, for very short messages

Table 1

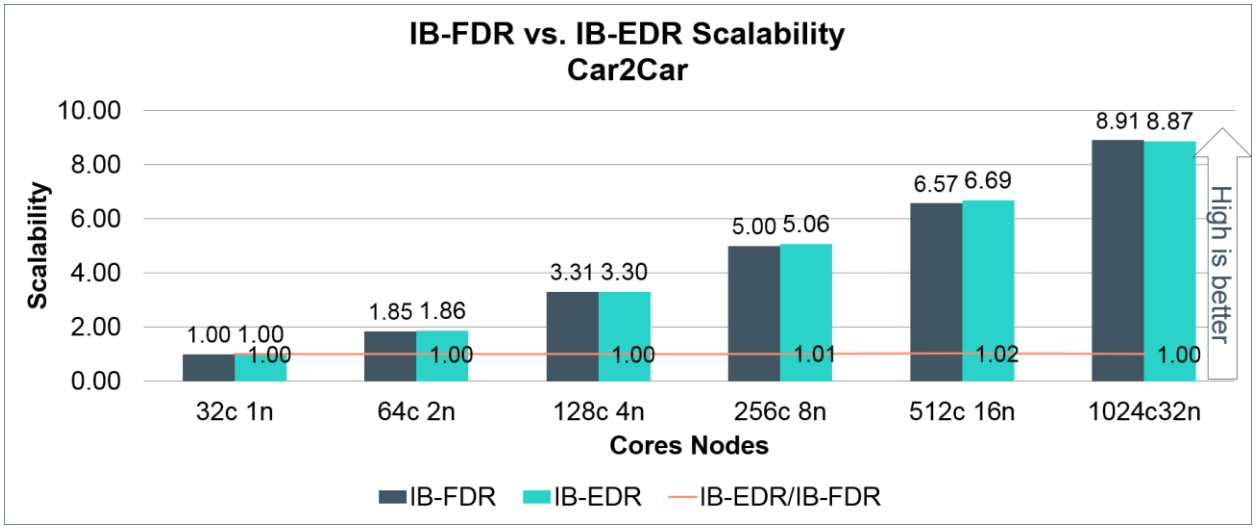

Figure 11

Conclusion

In summary, we investigate many aspects of hardware components that effect MPP LS-DYNA performance and find the following:

- Enabling Turbo Boost and using the AVX2 enabled version can help MPP LS-DYNA performance substantially.
- The performance of MPP LS-DYNA using the highly parallel file system LUSTRE is on par with using a local storage.
- Raising the DIMM frequency by 10 percent can increase MPP LS-DYNA performance by 2 percent.
- For a balanced memory configuration, the performance of MPP LS-DYNA with Home Snoop mode is on par with COD mode. However, for an unbalanced memory configuration using two different DIMMS in a socket, the performance with COD can have a substantial edge over that with Home Snoop.
- IB-FDR has sufficient network bandwidth to perform Car2Car workloads up to 1024 cores.