HPC ADVISORY COUNCIL

LS-DYNA
Performance Benchmark and Profiling

October 2017

Note

- **The following research was performed under the HPC Advisory Council activities**
  - Participating vendors: LSTC, Huawei, Mellanox
  - Compute resource - HPC Advisory Council Cluster Center

- **The following was done to provide best practices**
  - LS-DYNA performance overview
  - Understanding LS-DYNA communication patterns
  - Ways to increase LS-DYNA productivity
  - MPI libraries comparisons

- **For more info please refer to**
  - http://www.lstc.com
  - http://www.huawei.com
  - http://www.mellanox.com

LS-DYNA

- **LS-DYNA**
  - A general purpose structural and fluid analysis simulation software package capable of simulating complex real world problems
  - Developed by the Livermore Software Technology Corporation (LSTC)
- **LS-DYNA used by**
  - Automobile
  - Aerospace
  - Construction
  - Military
  - Manufacturing
  - Bioengineering

Objectives

- **The presented research was done to provide best practices**
  - LS-DYNA performance benchmarking
    - MPI Library performance comparison
    - Interconnect performance comparison
    - Compilers comparison
    - Optimization tuning
- **The presented results will demonstrate**
  - The scalability of the compute environment/application
  - Considerations for higher productivity and efficiency

Test Cluster Configuration

- **Huawei FusionServer X6000 “Broadwell” Cluster with Huawei FusionServer XH321 V3 32 Server Nodes**
  - Dual-Socket 14-Core Intel Xeon E5-2690 v4 @ 2.60 GHz CPUs; Dual-Socket 16-core Intel Xeon E5-2697A v4 @ 2.60 GHz CPUs
  - Memory: 256GB memory, DDR4 2400 MHz, Memory Snoop Mode in BIOS sets to Home Snoop
  - OS: RHEL 7.2, MLNX_OFED_LINUX-4.1-1.0.2.0 InfiniBand SW stack
- **Mellanox ConnectX-4 EDR 100Gb/s InfiniBand Adapters**
- **Mellanox Switch-IB SB7800 36-port EDR 100Gb/s InfiniBand Switch**
- **Huawei OceanStor 9000 Scale-out NAS storage system**
- **Compilers: Intel Parallel Studio XE 2018**
- **MPI: Mellanox HPC-X MPI Toolkit v1.9.7, Platform MPI 9.1.4.3, Intel MPI 2018**
- **Application: MPP LS-DYNA R9.1.0, single precision**
- **MPI Profiler: IPM (from Mellanox HPC-X)**
- **Benchmarks: TopCrunch benchmarks**
  - Neon Refined Revised (neon_refined_revised), Three Vehicle Collision (3cars), NCAC Minivan Model (Caravan2m-ver10, car2car)

X6000 Server Node — XH321 V3

XH321 V3

Entire Server

Node

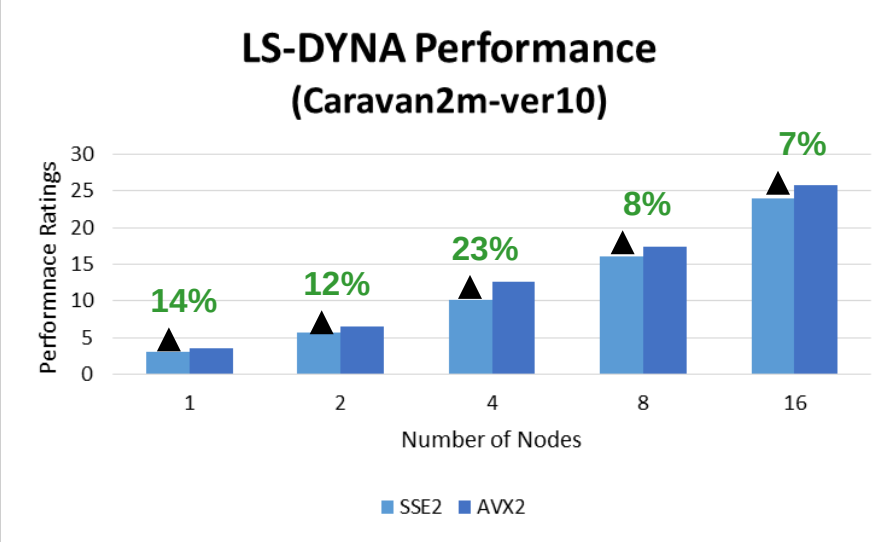
Form Factor	Half-width server node
Processors	1 or 2 Intel® Xeon® E5-2600 v3/v4
Memory	16 DDR4 DIMMs, providing up to 1 TB memory when configured with LRDIMMs
Internal Storage	6 SAS/SATA HDDs or NVMe SSDs
RAID	RAID 0, 1, 10, 5, 50, 6, or 60; supercapacitor
LOM Network Ports	**2 GE or 2 GE + 2 x 10GE**
PCIe Expansion	**Up to 2 PCIe x16 slots** Note: The mainboard has two PCIe slots: PCIe slot 1 (left) shared by a RAID controller card and PCIe slot 2(right) shared by an IB card .
LOM Storage	1 SATADOM and 1 M.2 NVMe SSD
Operating Temperature	120 W to 145 W processors: 5ºC to 35ºC Processors below 120 W: 5ºC to 40ºC
Dimensions	40.5 mm x 177.9 mm x 545.5 mm (1.59-in. x 7.00-in. x 21.48-in.)

Highlights

- **High performance:** Supports 1 or 2 Intel® Xeon® E5-2600 v3/v4 series processors with up to forty-four cores and 55 MB L3 cache capacity
- **High reliability:** Supports multiple RAID levels, supercapacitor for power failure protection, and TPM disk encryption
- **Diverse LOM I/O ports:** mainboards support up to two 1GbE and two 10GbE LOM ports
- **Features**: Hot-swappable fans, 2.5” NVMe SSDs, Mixed NVMe SSD and SAS/SATA SSD, 1m deep cabinet, 40 ℃ operating temperature, Liquid Cooling

LS-DYNA Performance – CPU Instructions

- **AVX2 shows higher performance than SSE2 executable on "Broadwell" CPU**
  - Performance gain of 7-23% by using AVX2 over SSE2 executable
  - AVX2 instructions runs at a reduced clock frequency as normal clock
  - AVX2 provides speedups of floating-point multiplication and addition operations
  - Benefit of AVX2 appears to be larger on bigger dataset (such as car2car)

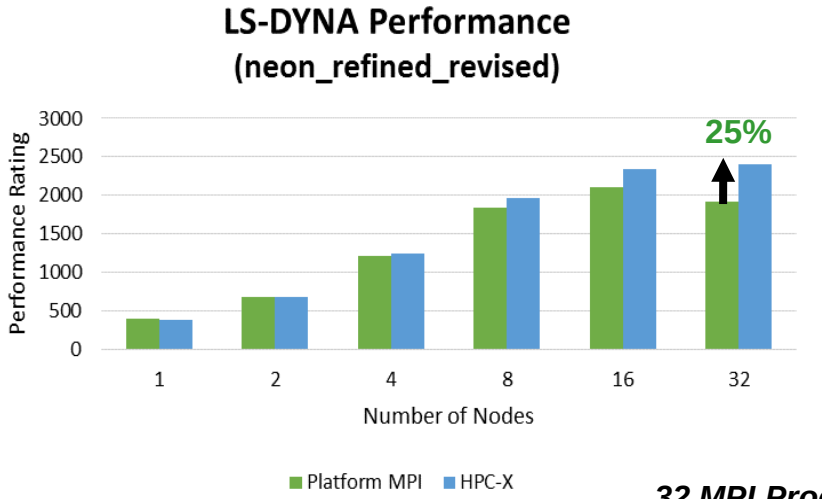


Higher is better

Platform MPI

LS-DYNA Performance – MPI Libraries (small model)

- **HPC-X outperforms Platform MPI in scalability performance**
  - Performance is 25% higher than Platform MPI in neon_refined_revised
  - HPC-X is based on Open MPI, with additional libraries for support offloads
  - HPC-X uses MXM UD transport for messaging and hcoll for collective offload

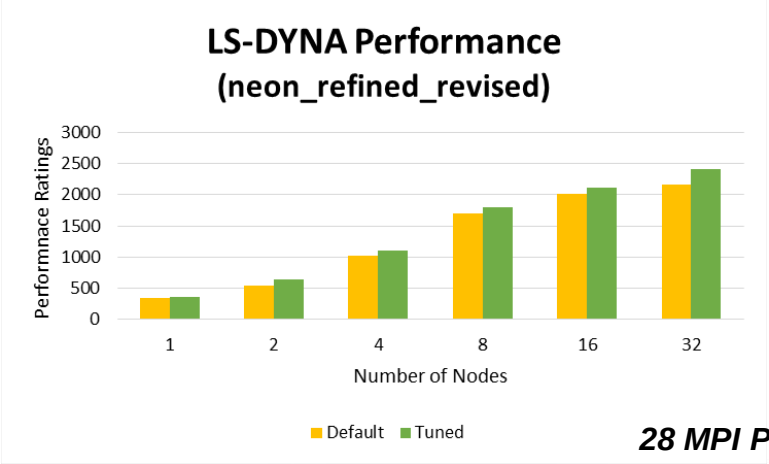


Higher is better

32 MPI Processes / Node, AVX2 executable

LS-DYNA Performance – MPI Tuning

- **MPI tuning helps getting best scalability performance for neon_refined_revised**
  - The neon_refined_revised case is the most network sensitive case
  - UD transport in MXM and memory optimization helps reducing overhead
  - Tuning flags used:
    - `-x MALLOC_MMAP_MAX_=0 -x MALLOC_TRIM_THRESHOLD_=-1 -mca coll_hcoll_enable 0 -x MXM_SHM_RNDV_THRESH=32768 –mca btl_sm_use_knem 1 –x MXM_SHM_KCOPY_MODE=knem -x MXM_ZCOPY_THRESH=inf -x MXM_UD_HARD_ZCOPY_THRESH=inf -x MXM_UD_MSS=8mb`

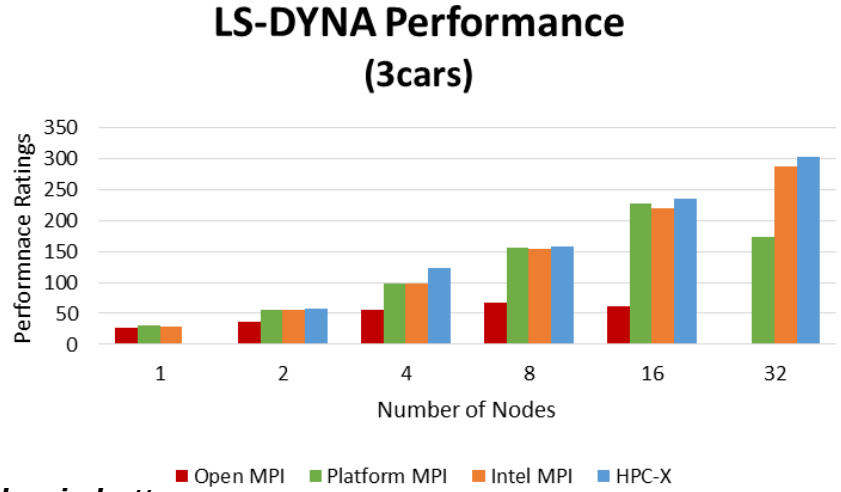


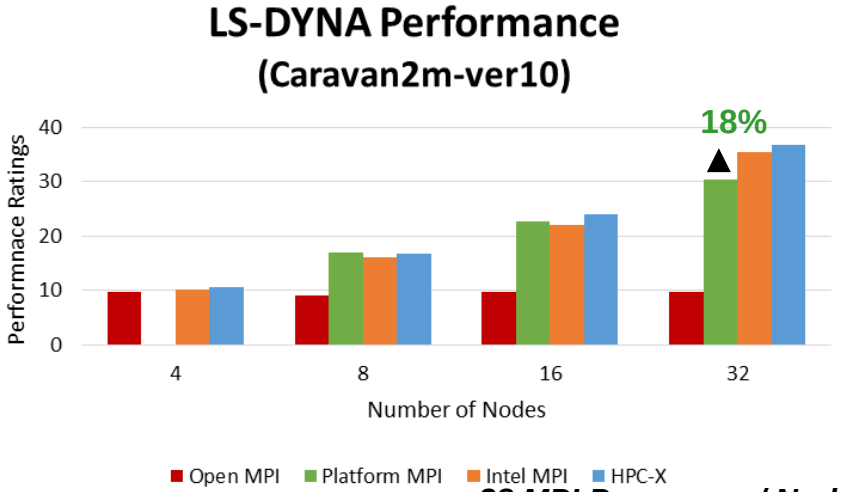
Higher is better

28 MPI Processes / Node, SSE2 executable

LS-DYNA Performance – MPI Libraries (larger models)

- **Platform MPI and HPC-X show good performance at scale**
  - 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
  - Open MPI runs are built and run without Mellanox acceleration modules (such as MXM, HCOLL)
  - Platform MPI runs with these parameters: -IBV -cpu_bind, -xrc



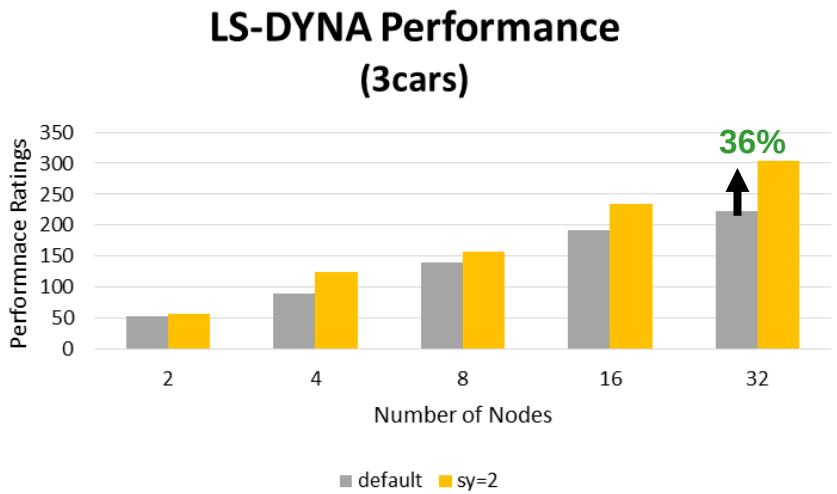


Higher is better

28 MPI Processes / Node

LS-DYNA Performance – Domain Decomposition

- **Domain Decomposition method can impact on scalability performance**
  - LS-DYNA allows specifying decomposition method in pfile
  - It allows MPI processes to get better distribution of workload so all processes can be fully utilized
  - Decomposition method can improve on scalability performance
    - Using “decomp { sy 2 }” in the pfile allows 36% of improved performance at 32 nodes (896 cores)

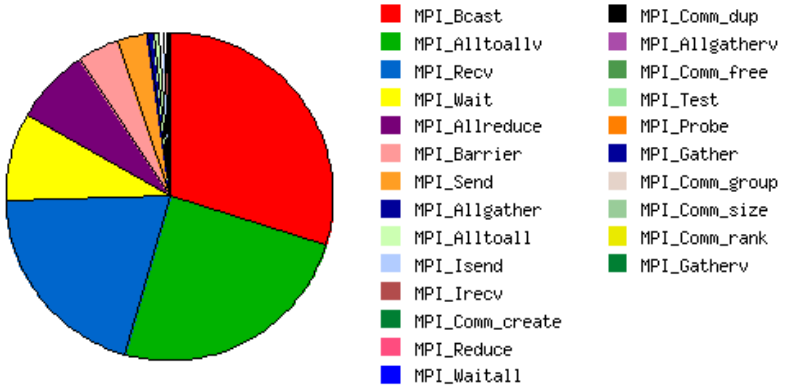


Higher is better

28 MPI Processes / Node

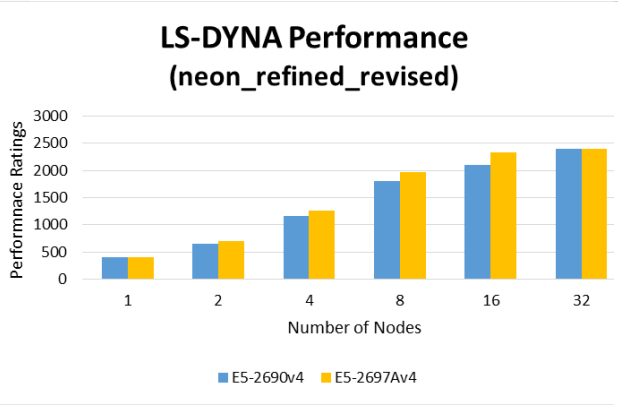
LS-DYNA Profiling – neon_refined_revised

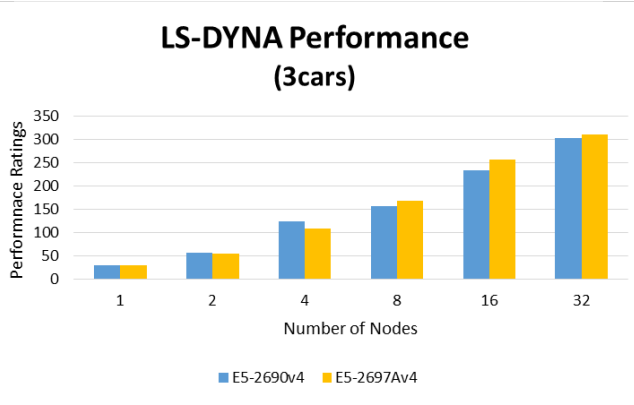
- **MPI profiler shows the type of underlying MPI network communications**
  - Showed that the majority of the MPI communications that occurred were collective operations
  - To be offloaded by the MPI offloading engine supported by the InfiniBand network interconnect
- **Majority of the MPI time is spent on MPI Collective Ops and MPI_Recv at 32 nodes**
  - MPI_Bcast (29%), MPI_Alltoallv (25%), MPI_Recv (20%)

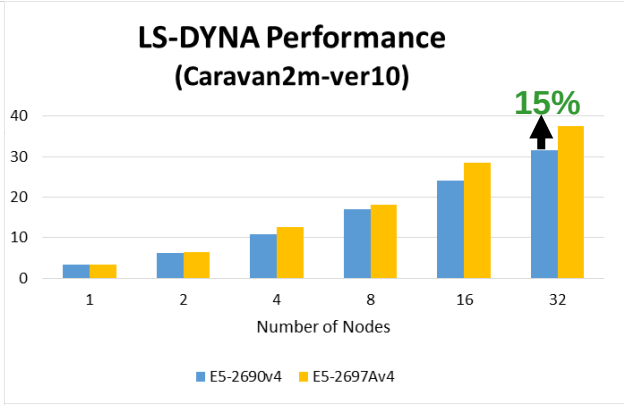


LS-DYNA Performance – CPU SKUs

- **LS-DYNA can benefit from the better CPU on larger dataset**
  - Percentage of time spent on compute is significant on larger dataset
  - E5-2697A v4 has more cache, 4 cores/node, & faster turbo; performance can be ~15%
  - As job scales, benefits would be limited since more time spent on MPI communications
  - Workload saturation would limited perf difference at scale for smaller dataset



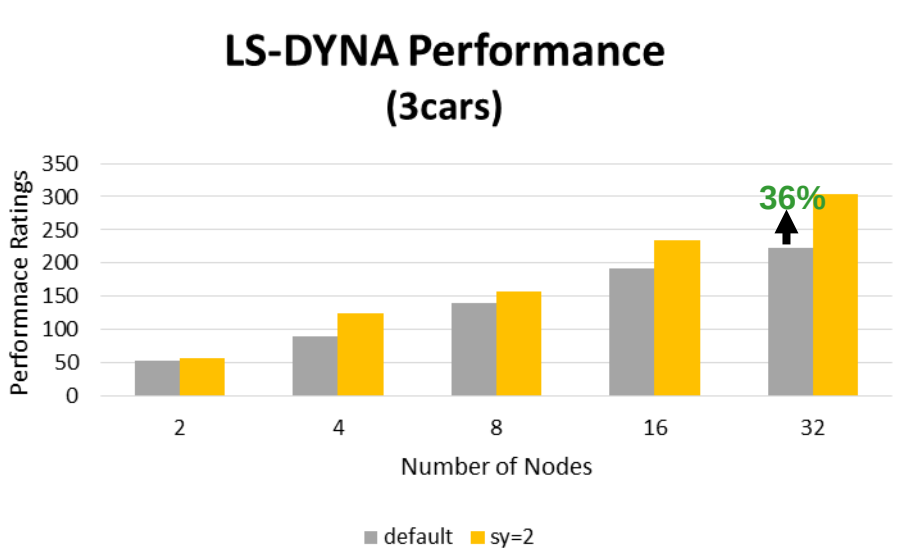




Higher is better

LS-DYNA Performance – Decomposition Method

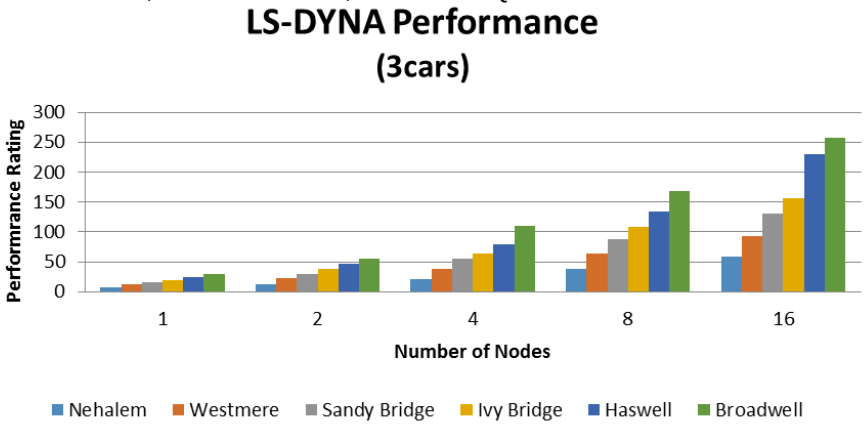
- **Domain decomposition method has impact on LS-DYNA scalability**
  - It affects the way LS-DYNA partition workload for the MPI processes, and also MPI communication
  - The effect in performance of domain decomposition method is widen as LS-DYNA scales
  - Performance gained by 36% at 32 nodes (896 cores) when compared to default with sy=2



Higher is better

LS-DYNA Performance – System Generations

- **Broadwell system configuration outperforms prior system generations**
  - Current Broadwell system outperformed Haswell by 21%, Ivy Bridge by 52%, Sandy Bridge by 132%, Westmere by 222%, Nehalem by 425%
  - Scalability support from EDR InfiniBand and HPC-X provide performance boost for LS-DYNA at 16 nodes
- **System components used:**
  - Broadwell: 2-socket 16-core Xeon E5-2697A v4 2.6GHz, 2400MHz DIMMs, ConnectX-4 EDR InfiniBand
  - Haswell: 2-socket 14-core Xeon E5-2697v3 2.6GHz, 2133MHz DIMMs, ConnectX-4 EDR InfiniBand
  - Ivy Bridge: 2-socket 10-core Xeon E5-2680v2 2.8GHz, 1600MHz DIMMs, Connect-IB FDR InfiniBand
  - Sandy Bridge: 2-socket 8-core Xeon E5-2680 2.7GHz, 1600MHz DIMMs, ConnectX-3 FDR InfiniBand
  - Westmere: 2-socket 6-core Xeon x5670 2.93GHz, 1333MHz DIMMs, ConnectX-2 QDR InfiniBand
  - Nehalem: 2-socket 4-core Xeon x5570 2.93GHz, 1333MHz DIMMs, ConnectX-2 QDR InfiniBand

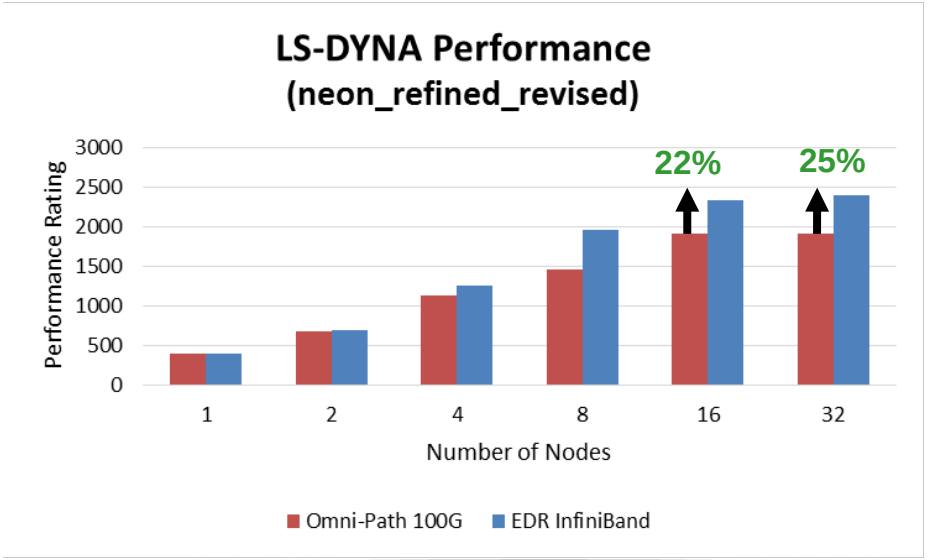


Higher is better

Best results per generation shown

LS-DYNA Performance – Interconnect

- **While both network nominally operate at 100Gb/s, EDR demonstrates performance lead**
  - Due to EDR IB offload architecture vs Omni-Path onload architecture
  - For performance beyond 4 nodes, EDR InfiniBand demonstrated even better scalability
- **EDR InfiniBand delivered superior scalability in application performance**
  - Resulting 25% in faster runtime on an LS-DYNA simulation that runs on 32 nodes (1024 MPI processes)

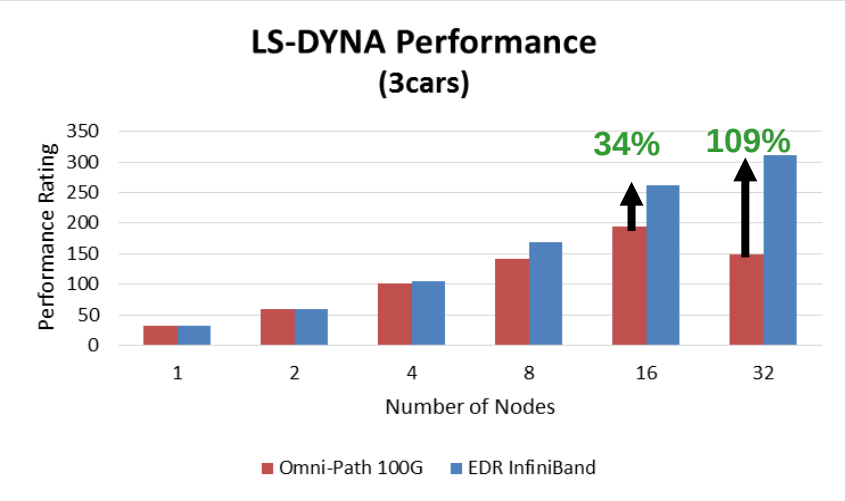


Higher is better

E5-2697A v4

LS-DYNA Performance – Interconnect

- **As the cluster scales, the performance difference becomes even more apparent**
  - Difference in performance occurs with as few as 8 nodes or 256 cores
  - EDR IB outperforms Omni-Path at scale by 34% for 512 cores (16 nodes, 512 ranks)
- **Performance difference becomes wider as more nodes and cores are used**
  - EDR performance improves to 109% better than Intel Omni-Path (32 nodes 1024 ranks)

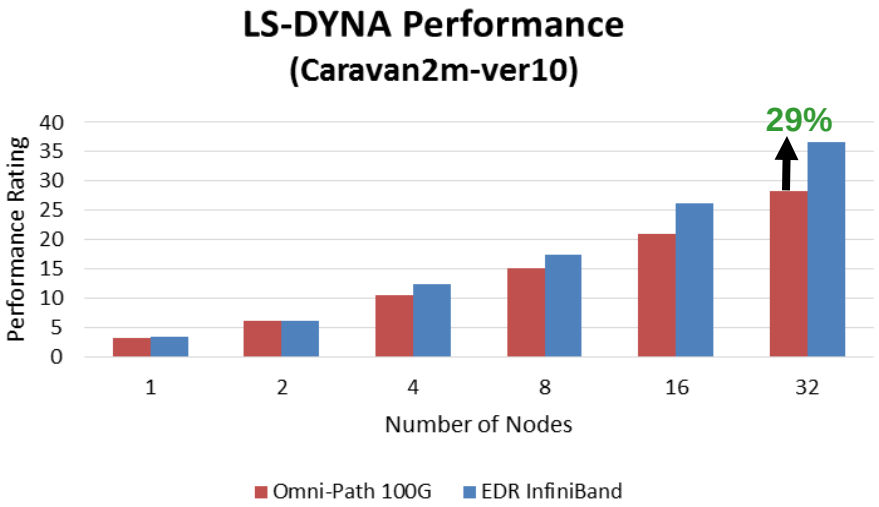


Higher is better

E5-2697A v4

LS-DYNA Performance – Interconnect

- **As the cluster scales, performance difference becomes more apparent**
  - EDR outperforms Omni-Path at scale by 29% for 1024 cores (32 nodes, 1024 ranks).
  - The difference in performance can be seen with as few as 8 nodes or 256 cores
- **Omni-Path performance decreased as cluster size scales to 32 nodes**
  - Overhead of onload network processing actually detracts from potential application performance



Higher is better

E5-2697A v4

LS-DYNA Summary

- **LS-DYNA is multi-purpose explicit and implicit finite element program**
  - Utilizes both compute and network communications
- **Fast network and MPI library are important for scalability performance**
  - HPC-X and tuning can improve performance by up to 18% at 32 nodes for large problem
- **CPU instruction support and number of cores can impact on performance**
  - Performance gain of up to 15% by using a 16-core vs a 14-core CPU SKU
  - Using AVX2 executable provides 7-23% of performance improvement
- **MPI Profiling**
  - Most MPI time is spent on MPI collective operations and non-blocking communications
  - Majority of the MPI time is spent on MPI Collective Ops and MPI_Recv at 32 nodes
    - MPI_Bcast (29%), MPI_Alltoallv (25%), MPI_Recv (20%)

Thank You

HPC Advisory Council