HPC ADVISORY COUNCIL

LS-DYNA
Performance Benchmark and Profiling

April 2015

DELL

intel

Mellanox TECHNOLOGIES

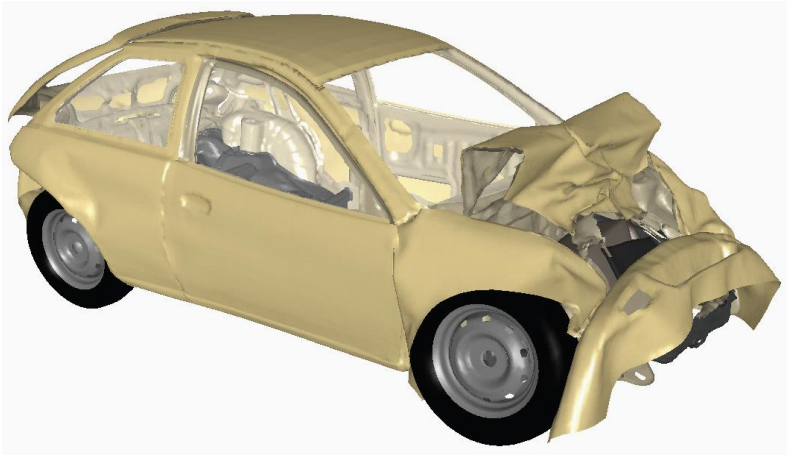
LSTC Livermore Software Technology Corp.

Note

- **The following research was performed under the HPC Advisory Council activities**
  - Participating vendors: Intel, Dell, 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.dell.com
  - http://www.intel.com
  - http://www.mellanox.com
  - http://www.lstc.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
    - CPUs comparison
    - Optimization tuning
- **The presented results will demonstrate**
  - The scalability of the compute environment/application
  - Considerations for higher productivity and efficiency

Test Cluster Configuration

- **Dell PowerEdge R730 32-node (896-core) "Thor" cluster**
  - Dual-Socket 14-Core Intel E5-2697v3 @ 2.60 GHz CPUs (Power Management in BIOS sets to Maximum Performance)
  - Memory: 64GB memory, DDR4 2133 MHz, Memory Snoop Mode in BIOS sets to Home Snoop
  - OS: RHEL 6.5, MLNX_OFED_LINUX-2.4-1.0.5.1_20150408_1555 InfiniBand SW stack
  - Hard Drives: 2x 1TB 7.2 RPM SATA 2.5" on RAID 1
- **Mellanox ConnectX-4 EDR 100Gb/s InfiniBand Adapters**
- **Mellanox Switch-IB SB7700 36-port EDR 100Gb/s InfiniBand Switch**
- **Mellanox ConnectX-3 FDR VPI InfiniBand and 40Gb/s Ethernet Adapters**
- **Mellanox SwitchX-2 SX6036 36-port 56Gb/s FDR InfiniBand / VPI Ethernet Switch**
- **MPI: Open MPI 1.8.4, Mellanox HPC-X v1.2.0-326, Intel MPI 5.0.2.044, IBM Platform MPI 9.1**
- **Application:**
  - LS-DYNA 8.0.0 (builds 95359, 95610), Single Precision
- **Benchmarks: 3 Vehicle Collision, Neon refined revised**

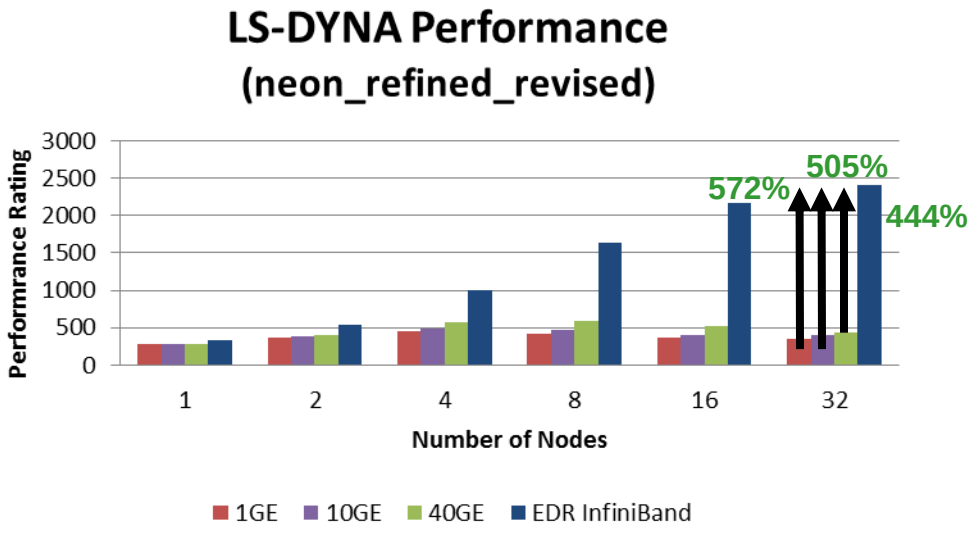
PowerEdge R730

Massive flexibility for data intensive operations

- **Performance and efficiency**
  - Intelligent hardware-driven systems management with extensive power management features
  - Innovative tools including automation for parts replacement and lifecycle manageability
  - Broad choice of networking technologies from GigE to IB
  - Built in redundancy with hot plug and swappable PSU, HDDs and fans
- **Benefits**
  - Designed for performance workloads
    - from big data analytics, distributed storage or distributed computing where local storage is key to classic HPC and large scale hosting environments
    - High performance scale-out compute and low cost dense storage in one package
- **Hardware Capabilities**
  - Flexible compute platform with dense storage capacity
    - 2S/2U server, 6 PCIe slots
  - Large memory footprint (Up to 768GB / 24 DIMMs)
  - High I/O performance and optional storage configurations
    - HDD options: 12 x 3.5” - or - 24 x 2.5 + 2x 2.5 HDDs in rear of server
    - Up to 26 HDDs with 2 hot plug drives in rear of server for boot or scratch

LS-DYNA Performance – Network Interconnects

- **EDR InfiniBand delivers superior scalability in application performance**
  - Provides higher performance by over 4-5 times than 1GbE, 10GbE and 40GbE
  - 1GbE stop scaling beyond 4 nodes, and 10GbE stops scaling beyond 8 nodes
  - InfiniBand demonstrates continuous performance gain at scale

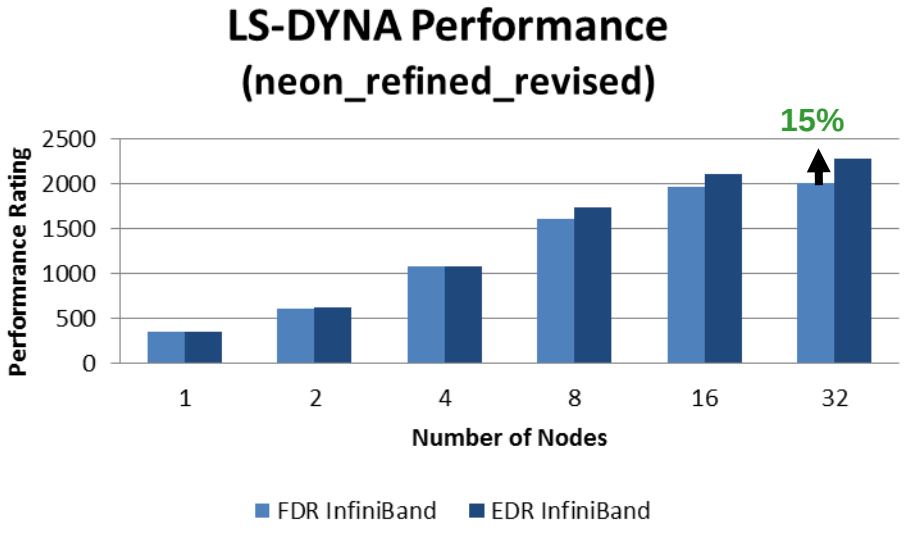


Higher is better

28 MPI Processes / Node

LS-DYNA Performance – EDR vs FDR InfiniBand

- **EDR InfiniBand delivers superior scalability in application performance**
  - As the cluster scales, performance gap of EDR IB becomes widen
- **Performance advantage of EDR InfiniBand increases for larger core counts**
  - EDR IB provides 15% versus FDR IB at 32 nodes (896 cores)

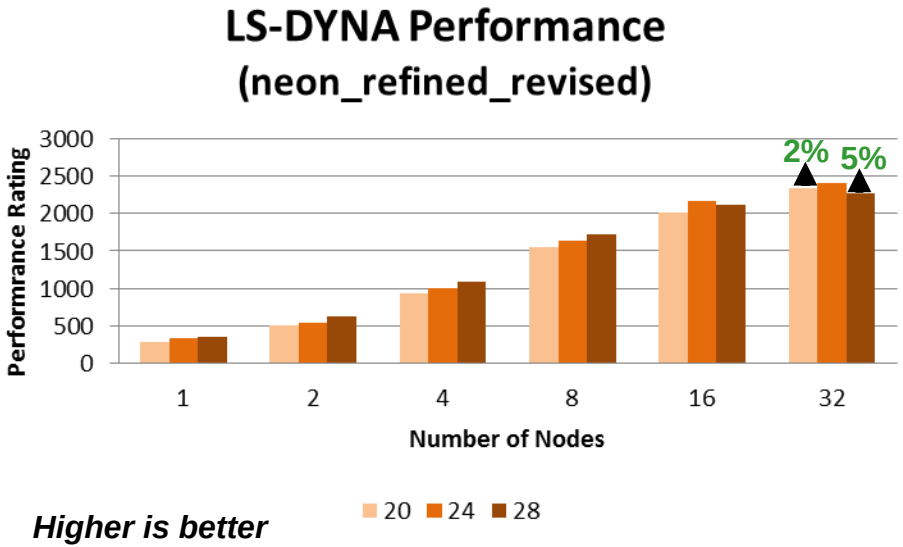


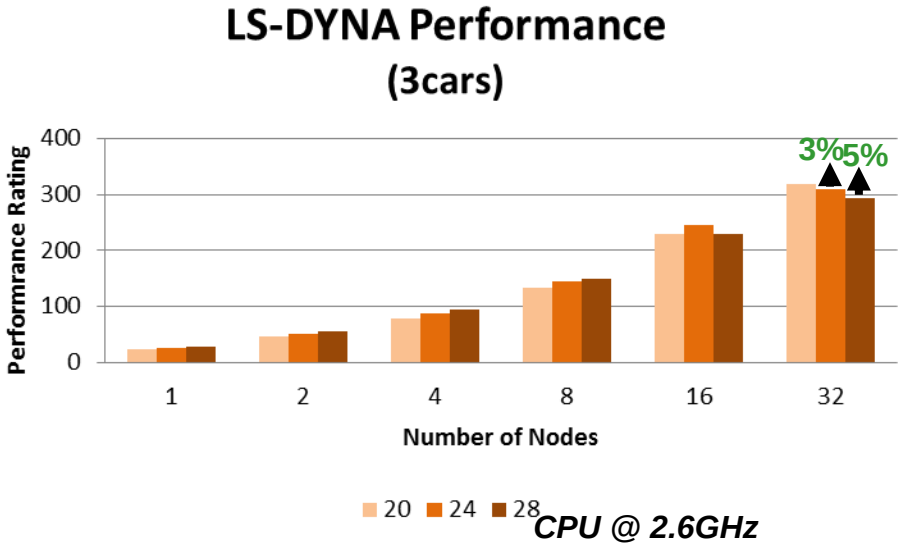
Higher is better

28 MPI Processes / Node

LS-DYNA Performance – Cores Per Node

- **Better performance is seen at scale with less CPU cores per node**
  - At low node counts, higher performance can be achieved with more cores per node
  - At high node counts, slightly better performance by using less cores per node
  - Memory bandwidth might be limited by more CPU cores being used



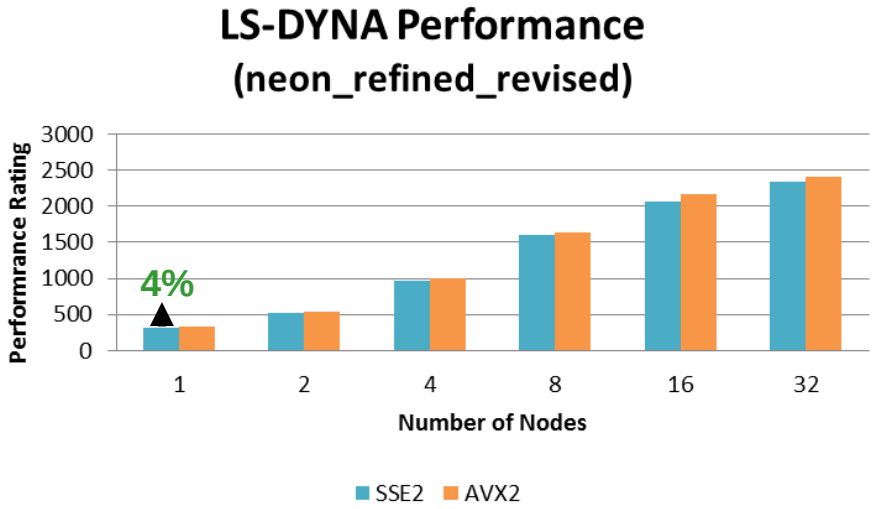


Higher is better

CPU @ 2.6GHz

LS-DYNA Performance – AVX2/SSE2 CPU Instructions

- **LS-DYNA provides executables with supports for different CPU instructions**
  - AVX2 is supported on “Haswell” while SSE2 is supported on previous generations
  - Due to runtime issue, AVX2 executable build 95610 is used, instead of the public build 95359
  - Slight improvement of ~2-4% by using executable with AVX2 instructions
  - The AVX2 instructions runs at a lower clock speed (2.2GHz) than normal CPU clock (2.6GHz)

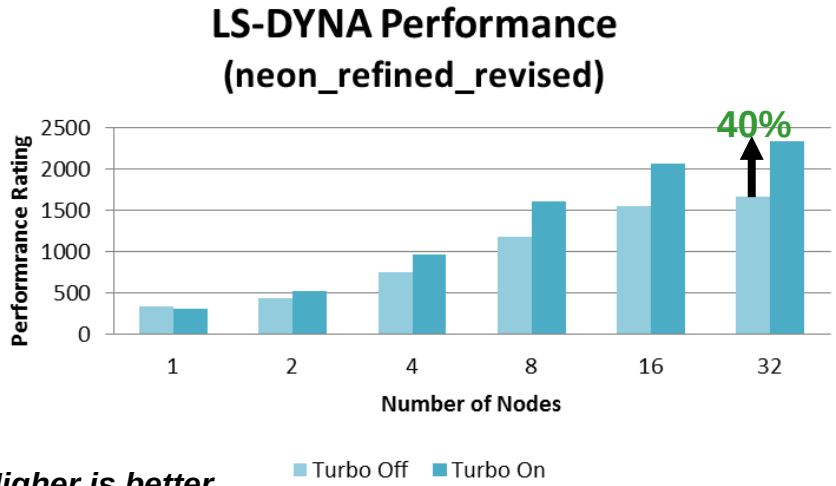


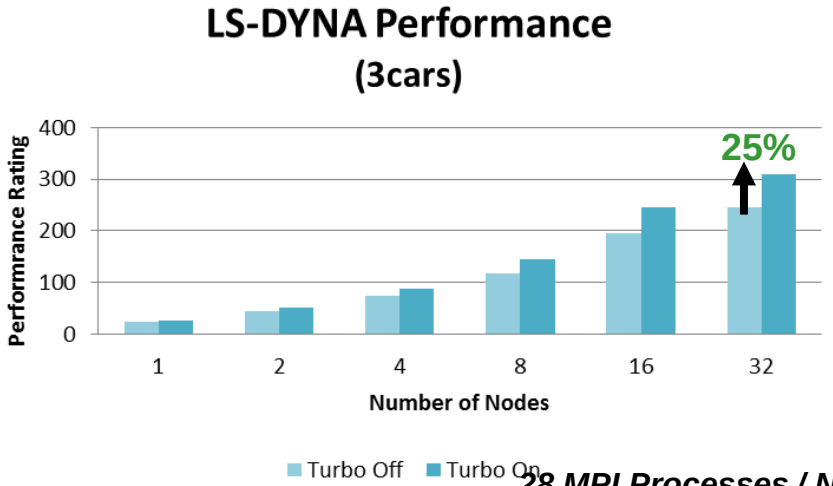
Higher is better

24 MPI Processes / Node

LS-DYNA Performance – Turbo Mode

- **Turbo Boost enables processors to run above its base frequency**
  - Capability to allow CPU cores to run dynamically above the CPU clock
  - When thermal headroom allows the CPU to operate
  - The 2.6GHz clock speed could boost to Max Turbo Frequency of 3.3GHz
  - Running with Turbo Boost translates to a ~25% of performance boost



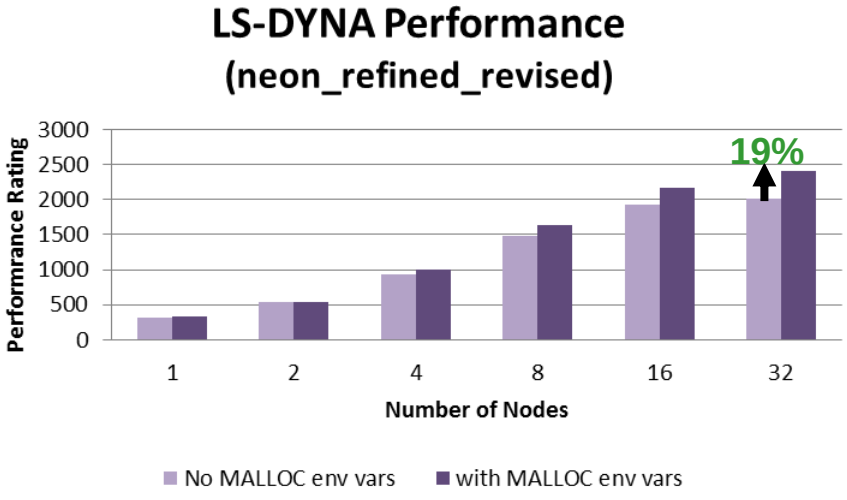


Higher is better

28 MPI Processes / Node

LS-DYNA Performance – Memory Optimization

- **Setting the environment variables for memory allocator improve on performance**
  - Modifying the memory allocator allows faster memory registration for communications
- **Environment variables used:**
  - export MALLOC_MMAP_MAX_=0
  - export MALLOC_TRIM_THRESHOLD_=-1



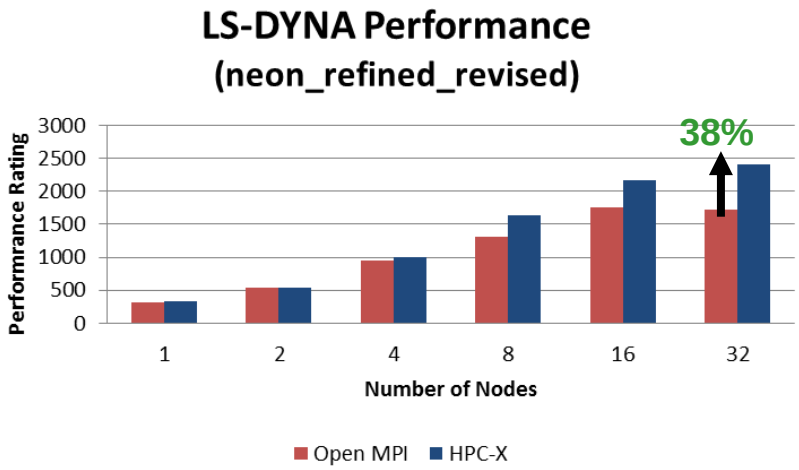
Higher is better

28 MPI Processes / Node

LS-DYNA Performance – MPI Optimization

- **FCA and MXM enhance LS-DYNA performance at scale for HPC-X**
  - Open MPI and HPC-X are based on the Open MPI distribution
  - The “yalla” PML, UD transport and memory optimization in HPC-X reduce overhead
  - MXM provides a speedup of 38% over un-tuned baseline run at 32 nodes (768 cores)
- **MCA parameters for MXM:**
  - For enabling MXM:

```
-mca btl_sm_use_knem 1 -mca pml yalla -x MXM_TLS=ud,shm,self -x MXM_SHM_RNDV_THRESH=32768 -x MXM_RDMA_PORTS=mlx5_0:1
```

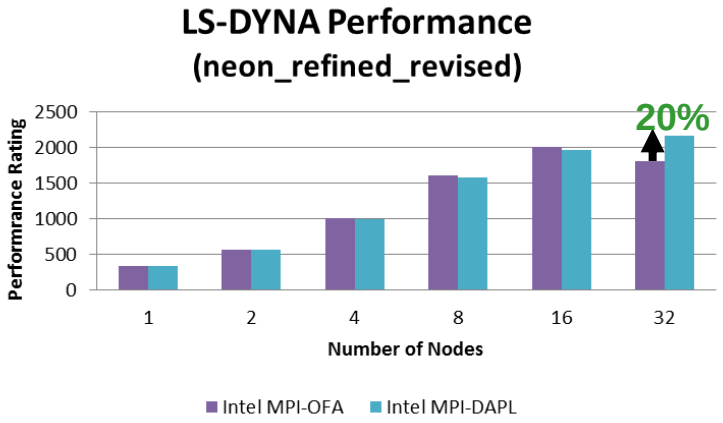


Higher is better

24 MPI Processes / Node

LS-DYNA Performance – Intel MPI Optimization

- **The DAPL provider performs better than OFA provider for Intel MPI**
  - DAPL would provide better scalability performance for Intel MPI on LS-DYNA
- **MCA parameters for MXM:**
  - Common for 2 tests: `I_MPI_DAPL_SCALABLE_PROGRESS 1, I_MPI_RDMA_TRANSLATION_CACHE 1, I_MPI_FAIR_CONN_SPIN_COUNT 2147483647, I_MPI_FAIR_READ_SPIN_COUNT 2147483647, I_MPI_ADJUST_REDUCE 2, I_MPI_ADJUST_BCAST 0, I_MPI_RDMA_TRANSLATION_CACHE 1, I_MPI_RDMA_RNDV_BUF_ALIGN 65536, I_MPI_SPIN_COUNT 121`
  - For OFA: `-IB, MV2_USE_APM 0, I_MPI_OFA_USE_XRC 1`
  - For DAPL: `-DAPL, I_MPI_DAPL_DIRECT_COPY_THRESHOLD 65536, I_MPI_DAPL_UD enable, I_MPI_DAPL_PROVIDER ofa-v2-mlx5_0-1u`

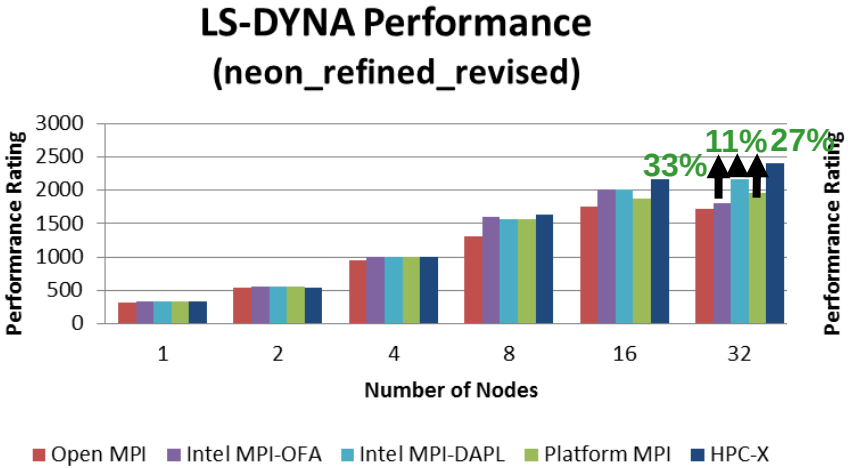


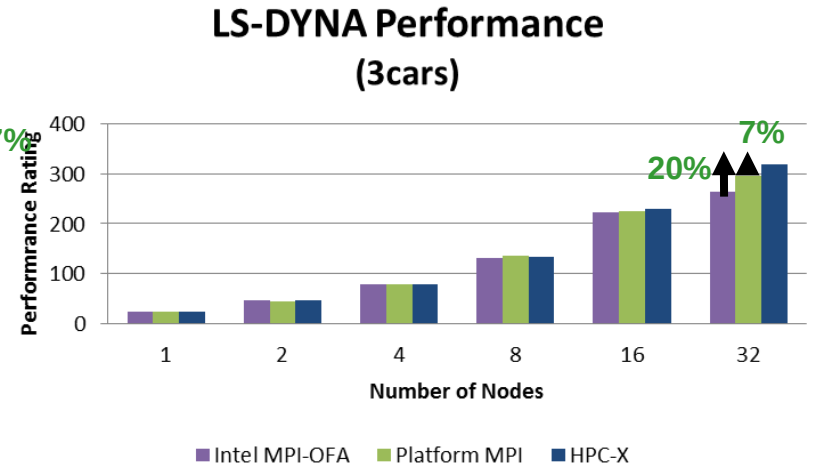
Higher is better

24 MPI Processes / Node

LS-DYNA Performance – MPI Libraries

- **HPC-X outperforms Platform MPI, and Open MPI in scalability performance**
  - HPC-X delivers higher performance than Intel MPI (OFA) by 33%, (DAPL) by 11%, Platform MPI by 27% on neon_refined_revised
  - Performance is 20% higher than Intel OFA, and % 8% better than Platform MPI in 3cars
- **Tuning parameter used:**
  - For Open MPI: -bind-to-core and KNEM. For Platform MPI: -cpu_bind, -xrc. For Intel MPI: see previous slide



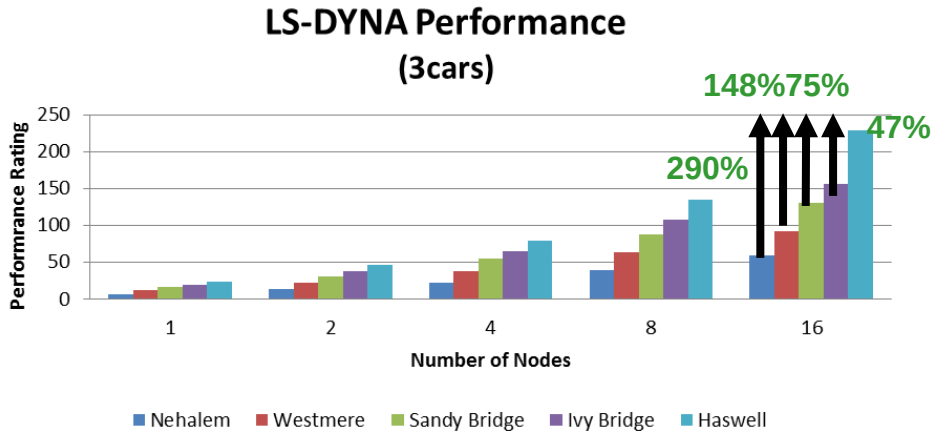


Higher is better

24 MPI Processes / Node

LS-DYNA Performance – System Generations

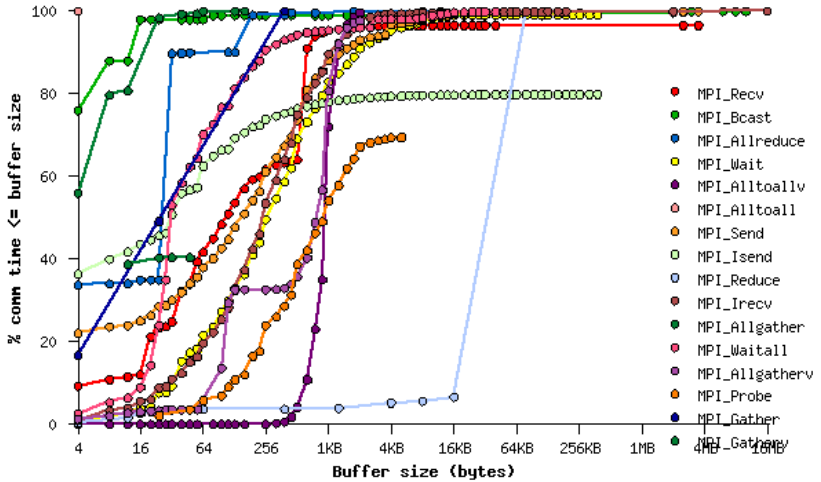
- **Current Haswell system configuration outperforms prior system generations**
  - Current systems outperformed Ivy Bridge by 47%, Sandy Bridge by 75%, Westmere by 148%, Nehalem by 290%
  - Scalability support from EDR InfiniBand and HPC-X provide huge boost in performance at scale for LS-DYNA
- **System components used:**
  - Haswell: 2-socket 14-core E5-2697v3@2.6GHz, 2133MHz DIMMs, ConnectX-4 EDR InfiniBand
  - Ivy Bridge: 2-socket 10-core E5-2680v2@2.8GHz, 1600MHz DIMMs, Connect-IB FDR InfiniBand
  - Sandy Bridge: 2-socket 8-core E5-2680@2.7GHz, 1600MHz DIMMs, ConnectX-3 FDR InfiniBand
  - Westmere: 2-socket 6-core x5670@2.93GHz, 1333MHz DIMMs, ConnectX-2 QDR InfiniBand
  - Nehalem: 2-socket 4-core x5570@2.93GHz, 1333MHz DIMMs, ConnectX-2 QDR InfiniBand



Higher is better

LS-DYNA Profiling – Time Spent in MPI

- **Most of the MPI messages are in the medium sizes**
  - Most message sizes are between 0 to 64B
- **For the most time consuming MPI calls**
  - MPI_Recv: Most messages are under 4KB
  - MPI_Bcast: Majority are less than 16B, but larger messages exist
  - MPI_Allreduce: Most messages are less than 256B

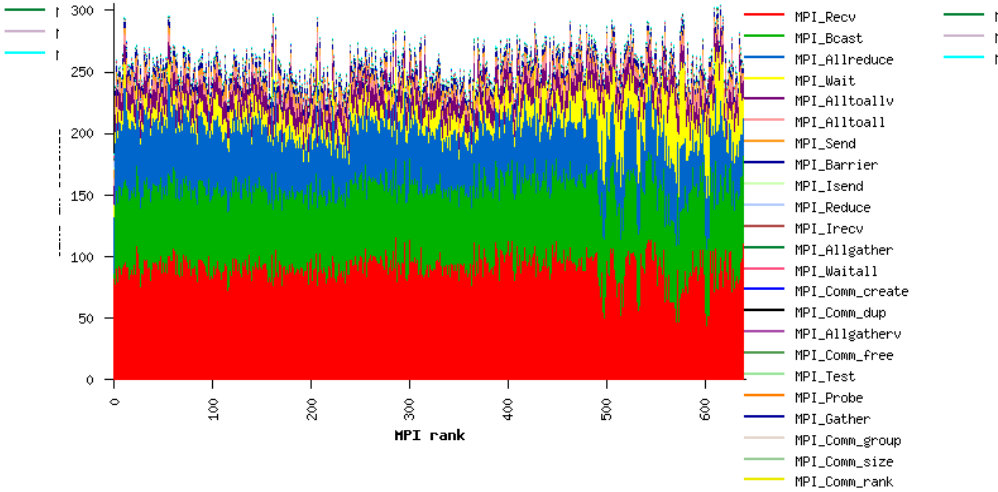


LS-DYNA Profiling – Time Spent in MPI

- **Majority of the MPI time is spent on MPI_recv and MPI Collective Ops**
  - MPI_Recv(36%), MPI_Allreduce(27%), MPI_Bcast(24%)
- **Similar communication characteristics seen on both input dataset**
  - Both exhibit similar communication patterns

Neon_refined_revised – 32 nodes

3 Vehicle Collision – 32 nodes



LS-DYNA Summary

- **Performance**
  - Compute: Intel Haswell cluster outperforms system architecture of previous generations
    - Outperforms Ivy Bridge by 47%, Sandy Bridge by 75%, Westmere by 148%, and Nehalem by 290%
    - Using executable with AVX2 instructions provides slight advantage
    - Slight improvement of ~2-4% by using executable with AVX2 instructions
  - Turbo Mode: Running with Turbo Boost provides ~25% of performance boost in some cases
    - Turbo Boost enables processors to run above its base frequency
  - Network: EDR InfiniBand and HPC-X MPI library deliver superior scalability in application performance
    - EDR IB provides higher performance by over 4-5 times vs 1GbE, 10GbE and 40GbE, 15% vs FDR IB at 32 nodes
- **MPI Tuning**
  - HPC-X enhances LS-DYNA performance at scale for LS-DYNA
    - MXM UD provides a speedup of 38% over un-tuned baseline run at 32 nodes
  - HPC-X outperforms Platform MPI, and Open MPI in scalability performance
    - Up to 27% better than Platform MPI on neon_refined_revised, and 8% better than Platform MPI in 3cars

Thank You

HPC Advisory Council