HPC ADVISORY COUNCIL

LS-DYNA
Performance Benchmark and Profiling

February 2014

DELL

LSTC Livermore Software Technology Corp.

intel

Mellanox TECHNOLOGIES

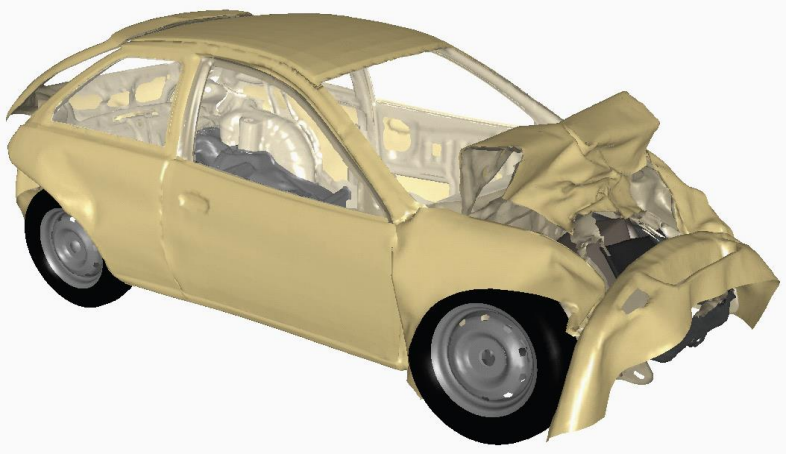
Note

- **The following research was performed under the HPC Advisory Council activities**
  - Participating vendors: Intel, Dell, Mellanox, LSTC
  - 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**
  - 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
    - Compilers comparison
- **The presented results will demonstrate**
  - The scalability of the compute environment/application
  - Considerations for higher productivity and efficiency

Test Cluster Configuration

- **Dell™ PowerEdge™ R720xd/R720 32-node (640-core) cluster**
  - Dual-Socket Octa-Core Intel E5-2680 V2 @ 2.80 GHz CPUs (Turbo Mode enabled)
  - Memory: 64GB DDR3 1600 MHz Dual Rank Memory Module (Static max Perf in BIOS)
  - Hard Drives: 24x 250GB 7.2 RPM SATA 2.5” on RAID 0
  - OS: RHEL 6.2, MLNX_OFED 2.1-1.0.0 InfiniBand SW stack
- **Intel Cluster Ready certified cluster**
- **Mellanox Connect-IB FDR InfiniBand and ConnectX-3 Ethernet adapters**
- **Mellanox SwitchX 6036 VPI InfiniBand and Ethernet switches**
- **MPI: Intel MPI 4.1, Platform MPI 9.1, Open MPI 1.6.5 w/ FCA 2.5 & MXM 2.1**
- **Application: LS-DYNA**
  - mpp971_s_R3.2.1_Intel_linux86-64 (for TopCrunch)
  - ls-dyna_mpp_s_r7_0_0_79069_x64_ifort120_sse2
- **Benchmark datasets: 3 Vehicle Collision, Neon refined revised**

About Intel® Cluster Ready

- **Intel® Cluster Ready systems make it practical to use a cluster to increase your simulation and modeling productivity**
  - Simplifies selection, deployment, and operation of a cluster
- **A single architecture platform supported by many OEMs, ISVs, cluster provisioning vendors, and interconnect providers**
  - Focus on your work productivity, spend less management time on the cluster
- **Select Intel Cluster Ready**
  - Where the cluster is delivered ready to run
  - Hardware and software are integrated and configured together
  - Applications are registered, validating execution on the Intel Cluster Ready architecture
  - Includes Intel® Cluster Checker tool, to verify functionality and periodically check cluster health

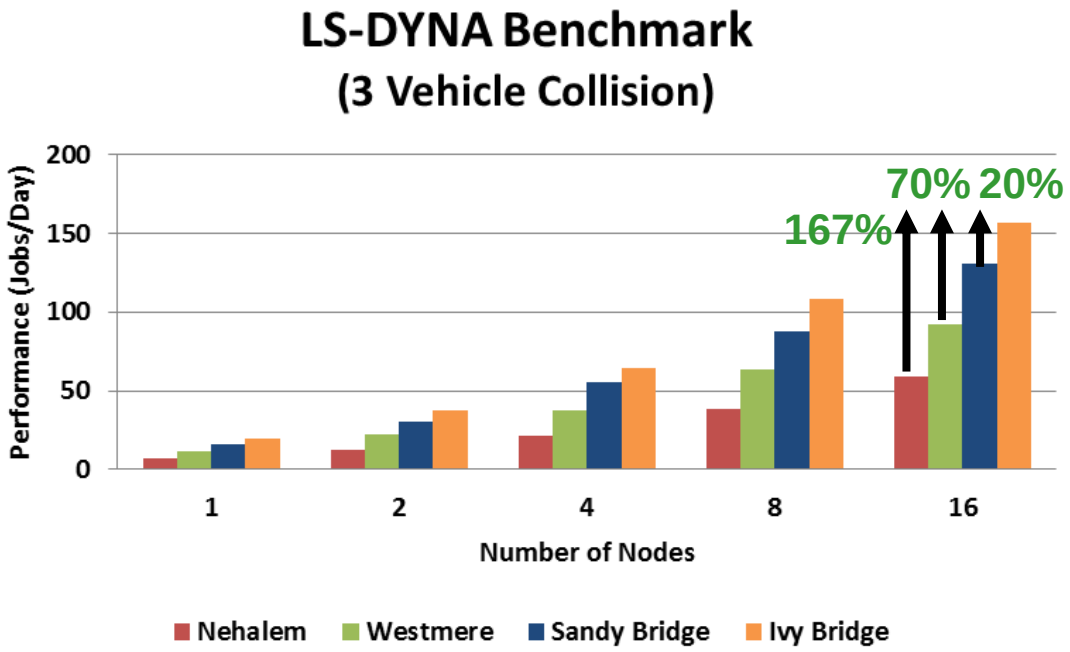
PowerEdge R720xd

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 – System Generations

- **Intel E5-2600v2 Series (Ivy Bridge) outperforms prior generations**
  - Up to 20% higher performance than Intel Xeon E5-2680 (Sandy Bridge) cluster
  - Up to 70% higher performance than Intel Xeon X5670 (Westmere) cluster
  - Up to 167% higher performance than Intel Xeon X5570 (Nehalem) cluster
- **System components used:**
  - Ivy Bridge: 2-socket 10-core E5-2680v2@2.8GHz, 1600MHz DIMMs, Connect-IB FDR
  - Sandy Bridge: 2-socket 8-core E5-2680@2.7GHz, 1600MHz DIMMs, ConnectX-3 FDR
  - Westmere: 2-socket 6-core x5670@2.93GHz, 1333MHz DIMMs, ConnectX-2 QDR
  - Nehalem: 2-socket 4-core x5570@2.93GHz, 1333MHz DIMMs, ConnectX-2 QDR

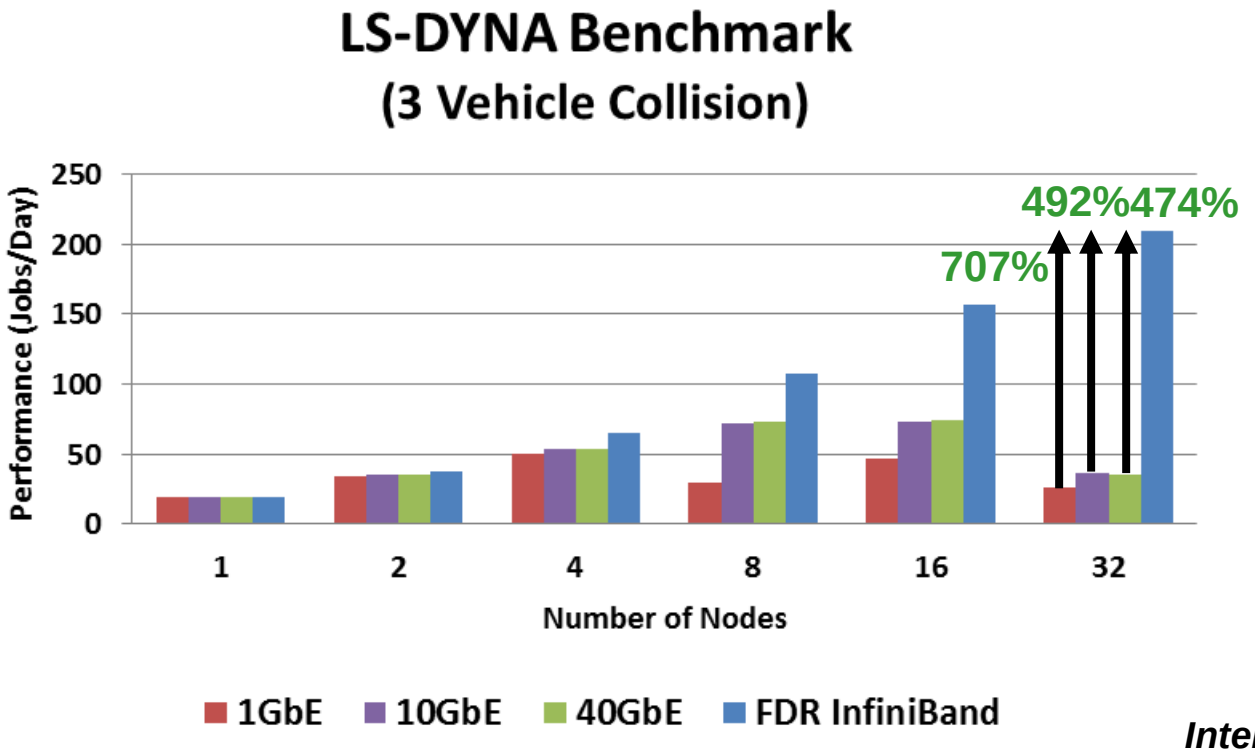


Higher is better

FDR InfiniBand

LS-DYNA Performance – Interconnects

- **FDR InfiniBand delivers superior scalability in application performance**
  - Provides higher performance by over 7 times % for 1GbE
  - Almost 5 times faster than 10GbE and 40GbE
  - 1GbE stop scaling beyond 4 nodes, and 10GbE stops scaling beyond 8 nodes
  - Only FDR InfiniBand demonstrates continuous performance gain at scale

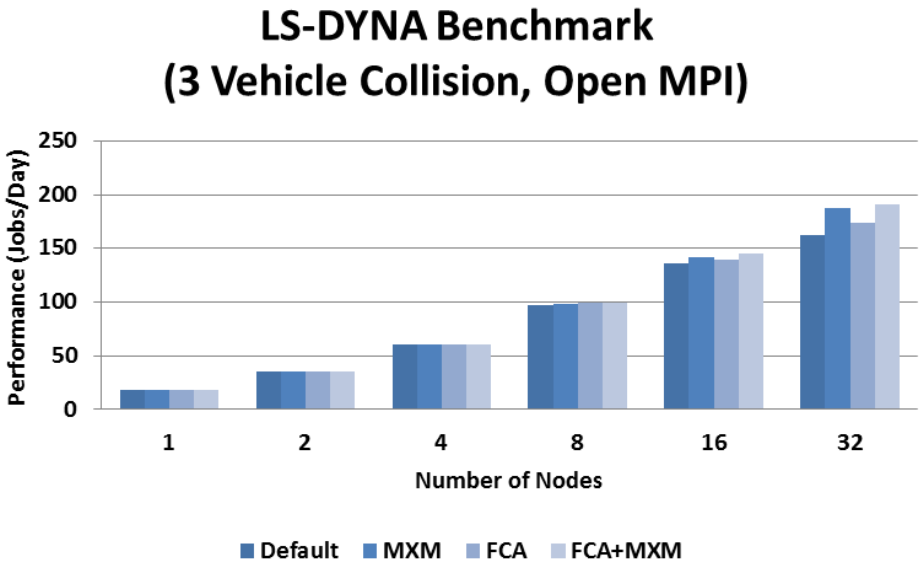


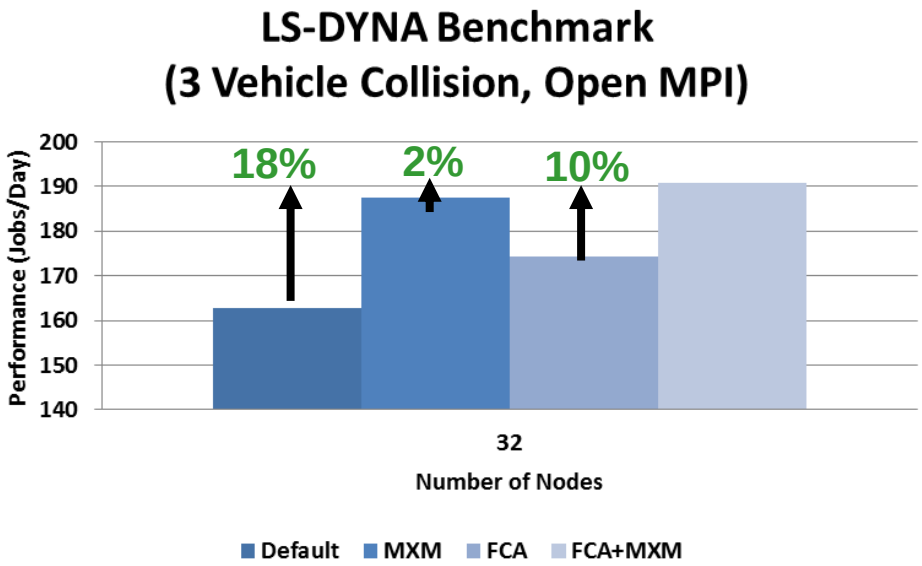
Higher is better

Intel E5-2680v2

LS-DYNA Performance – Open MPI Tuning

- **FCA and MXM enhance LS-DYNA performance at scale for Open MPI**
  - FCA allows MPI collective operation offloads to hardware while MXM provides memory enhancements to parallel communication libraries
  - FCA and MXM provide a speedup of 18% over untuned baseline run at 32 nodes
- **MCA parameters for enabling FCA and MXM:**
  - For enabling MXM:
    ```
    -mca mtl mxm -mca pml cm -mca mtl_mxm_np 0 -x MXM_TLS=ud,shm,self -x
    MXM_SHM_RNDV_THRESH=32768 -x MXM_RDMA_PORTS=mlx5_0:1
    ```
  - For enabling FCA:
    ```
    -mca coll_fca_enable 1 -mca coll_fca_np 0 -x fca_ib_dev_name=mlx5_0
    ```



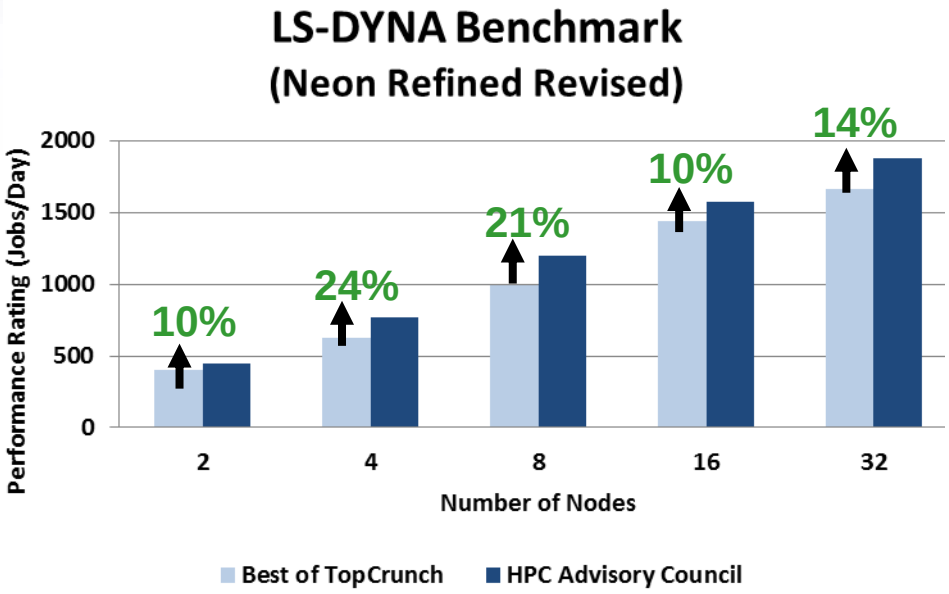


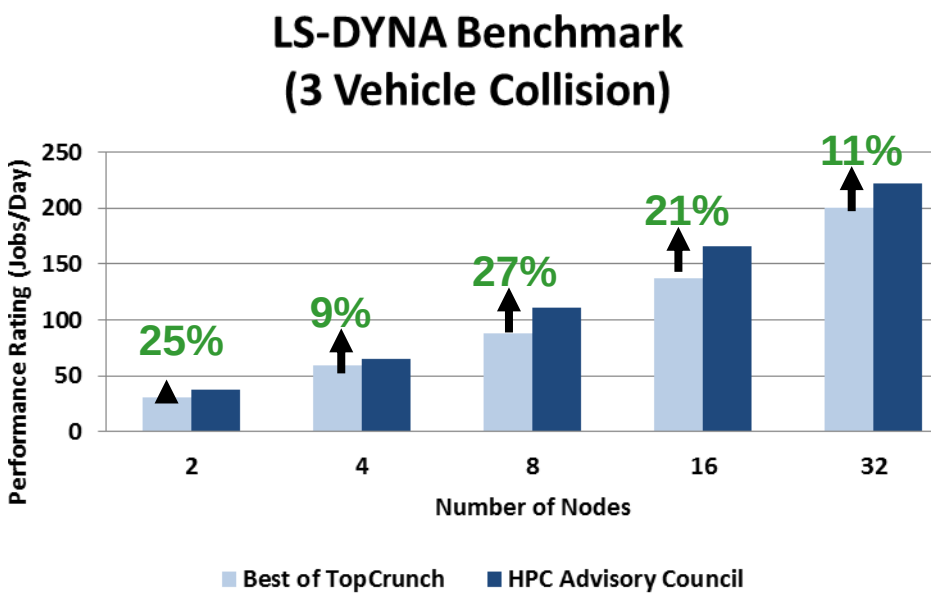
Higher is better

FDR InfiniBand

LS-DYNA Performance – TopCrunch

- **HPC Advisory Council performs better than the previous best published results**
  - TopCrunch (www.topcrunch.com) publishes LS-DYNA performance results
  - HPCAC achieved better performance on per node basis
  - 9% to 27% of higher performance than best published results on TopCrunch (Feb2014)
- **Comparing to all platforms on TopCrunch**
  - HPC Advisory Council results are world best for systems for 2 to 32 nodes
  - Achieving higher performance than larger node count systems



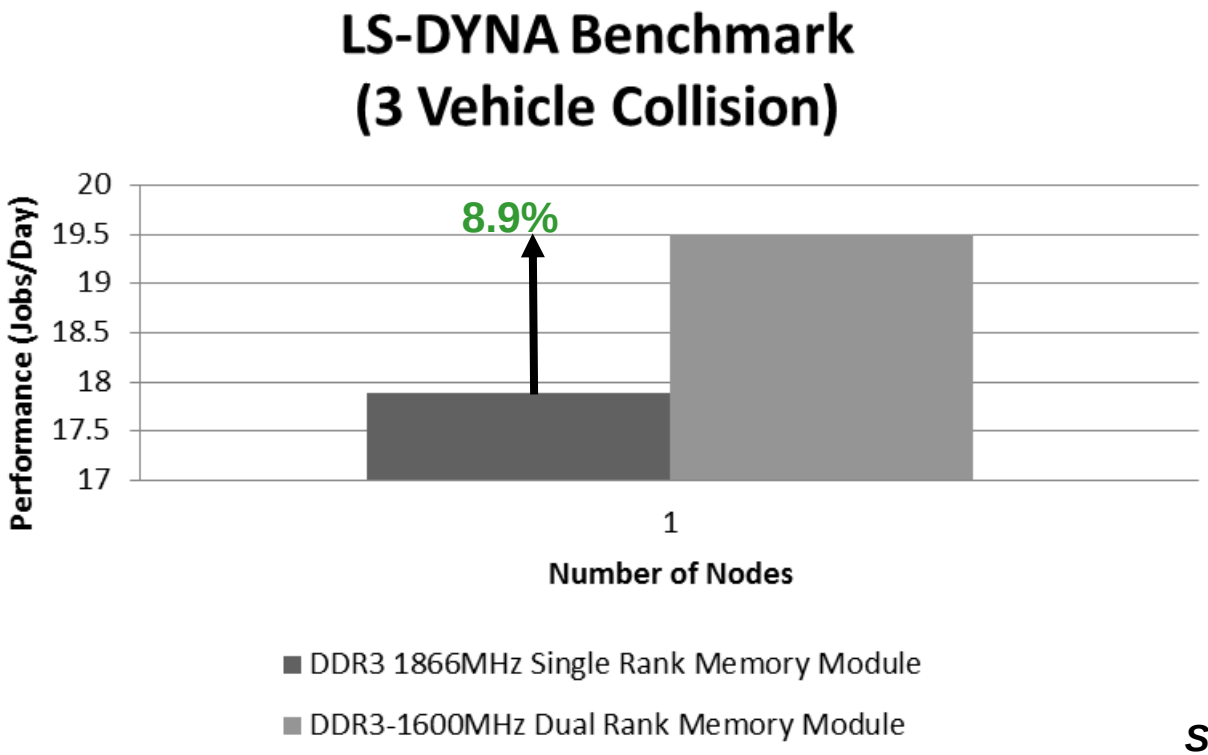


Higher is better

FDR InfiniBand

LS-DYNA Performance – Memory Module

- **Dual rank memory module provides better speedup for LS-DYNA**
  - Using Dual Rank 1600MHz DIMM is 8.9% faster than Single Rank 1866MHz DIMM
  - Ranking of the memory has more importance to performance than speed
- **System components used:**
  - DDR3-1866MHz PC14900 CL13 Single Rank Memory Module
  - DDR3-1600MHz PC12800 CL11 Dual Rank Memory Module

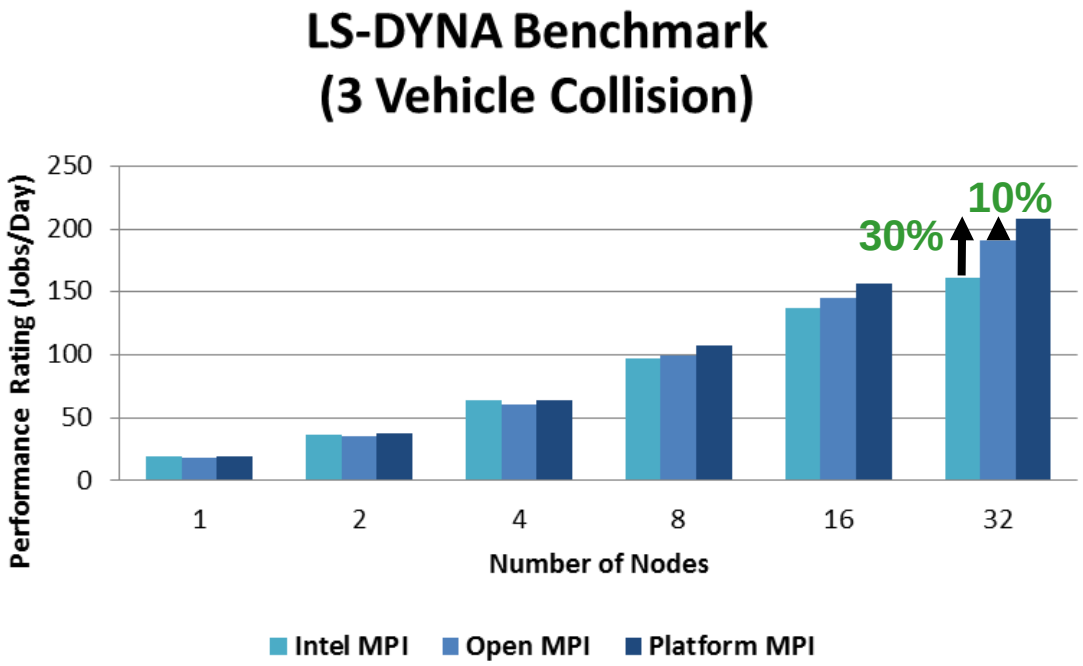


Higher is better

Single Node

LS-DYNA Performance – MPI

- **Platform MPI performs better than Open MPI and Intel MPI**
  - Up to 10% better than Open MPI and 30% better than Intel MPI
- **Tuning parameter used:**
  - Open MPI: -bind-to-core, FCA, MXM, KNEM
  - Platform MPI: -cpu_bind, XRC
  - Intel MPI: I_MPI_FABRICS shm:ofa, I_MPI_PIN on, I_MPI_ADJUST_BCAST 1 I_MPI_DAPL_SCALABLE_PROGRESS 1

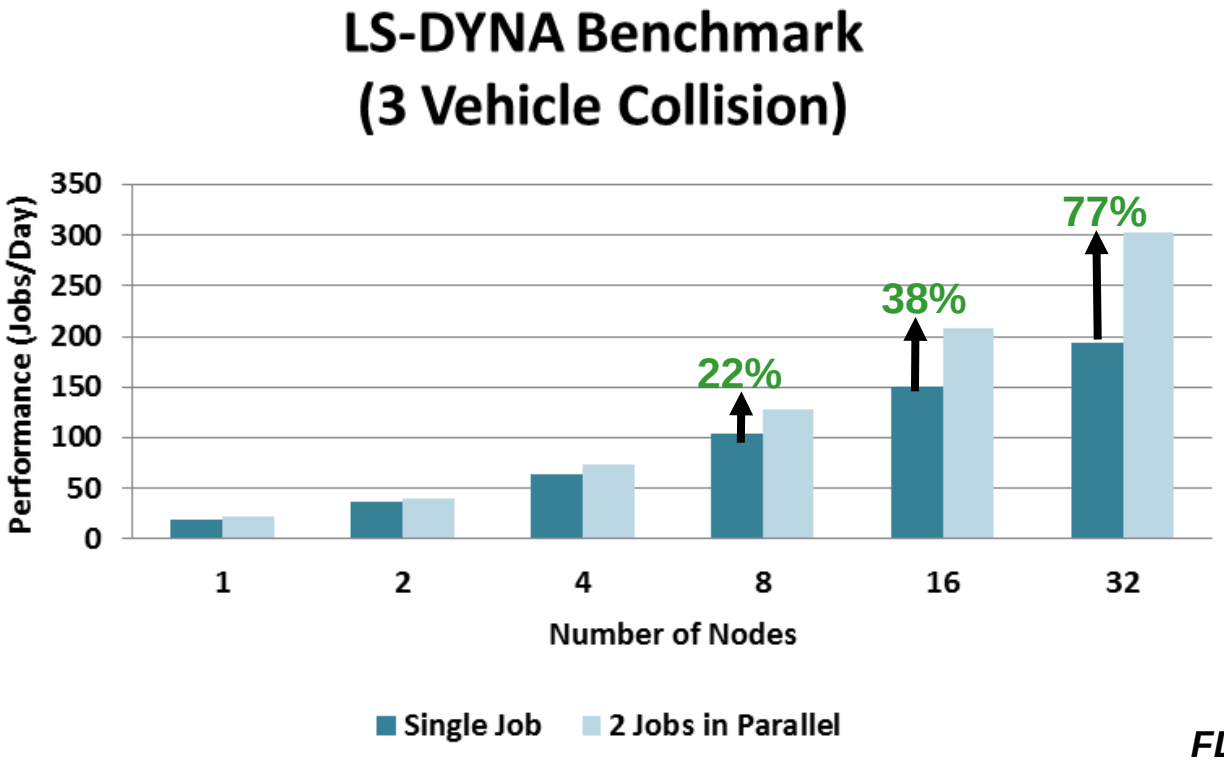


Higher is better

FDR InfiniBand

LS-DYNA Performance – System Utilization

- **Maximizing system productivity by running 2 jobs in parallel**
  - Up to 77% of increased system utilization at 32 nodes
  - Run each job separately in parallel by splitting system resource in half
- **System components used:**
  - Single Job: Use all cores and both IB ports to run a single job
  - 2 Jobs in parallel: Cores of 1 CPU and 1 IB port for a job, the rest for another

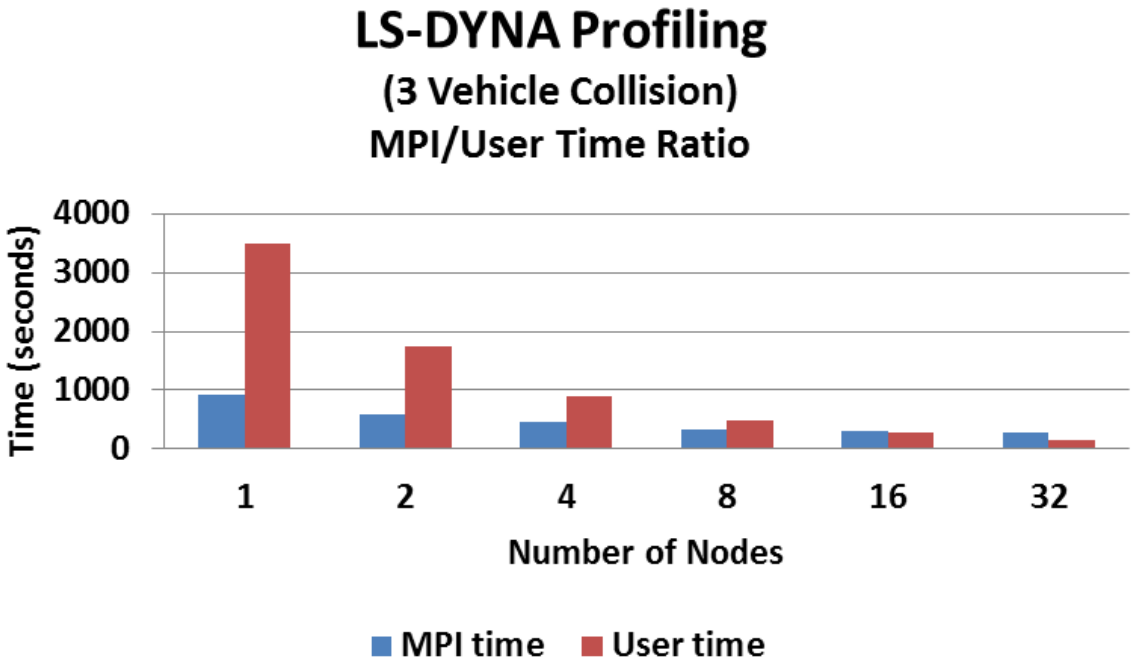


Higher is better

FDR InfiniBand

LS-DYNA Profiling – MPI/User Time Ratio

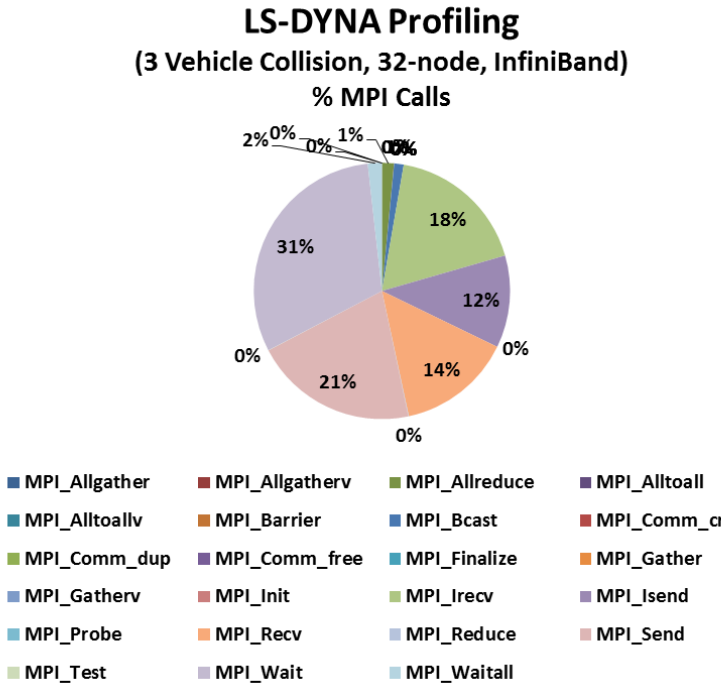
- **Computation time is dominant compared to MPI communication time**
  - MPI communication ratio increases as the cluster scales
- **Both computation time and communication declines as the cluster scales**
  - The InfiniBand infrastructure allows spreading the work without adding overheads
  - Computation time drops faster compares to communication time
  - Compute bound: Tuning for computation performance could yield better results

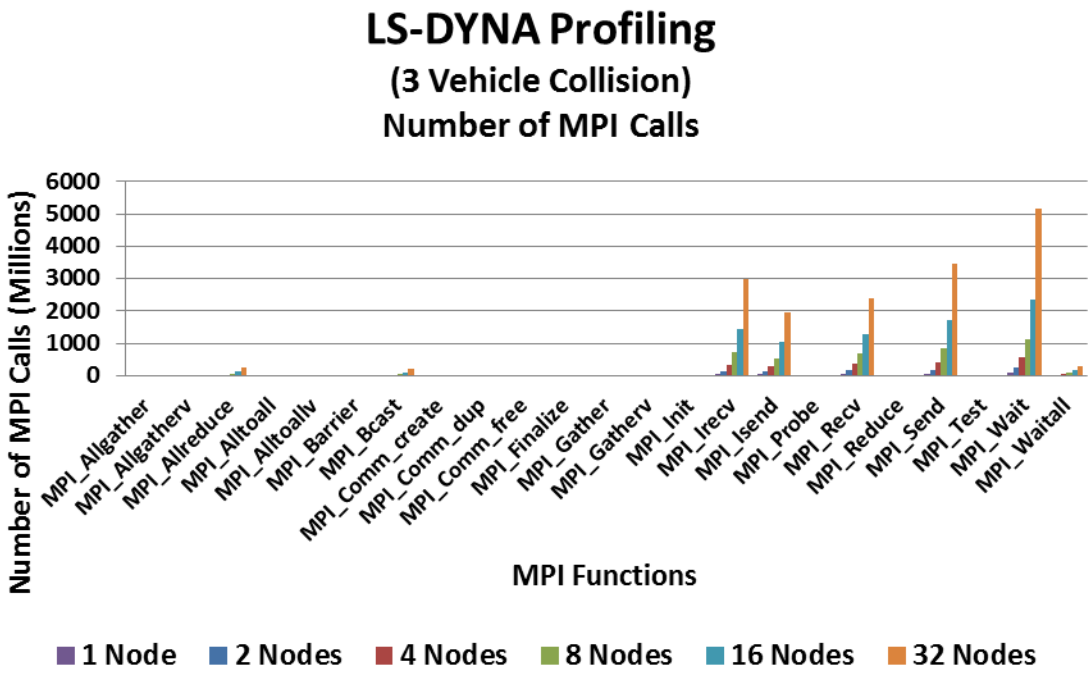


FDR InfiniBand

LS-DYNA Profiling – MPI Calls

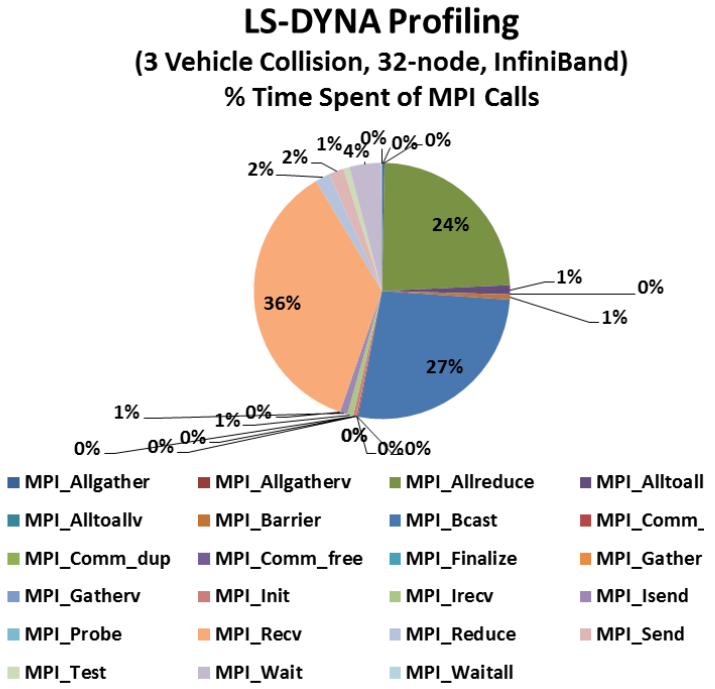
- **MPI_Wait, MPI_Send and MPI_Recv are the most used MPI calls**
  - MPI_Wait(31%), MPI_Send(21%), MPI_Irecv(18%), MPI_Recv(14%), MPI_Isend(12%)
- **LS-DYNA has majority of MPI point-to-point calls for data transfers**
  - Either blocking or non-blocking point-to-point transfers are seen
  - LS-DYNA has an extensive use of MPI APIs, over 23 MPI APIs are used

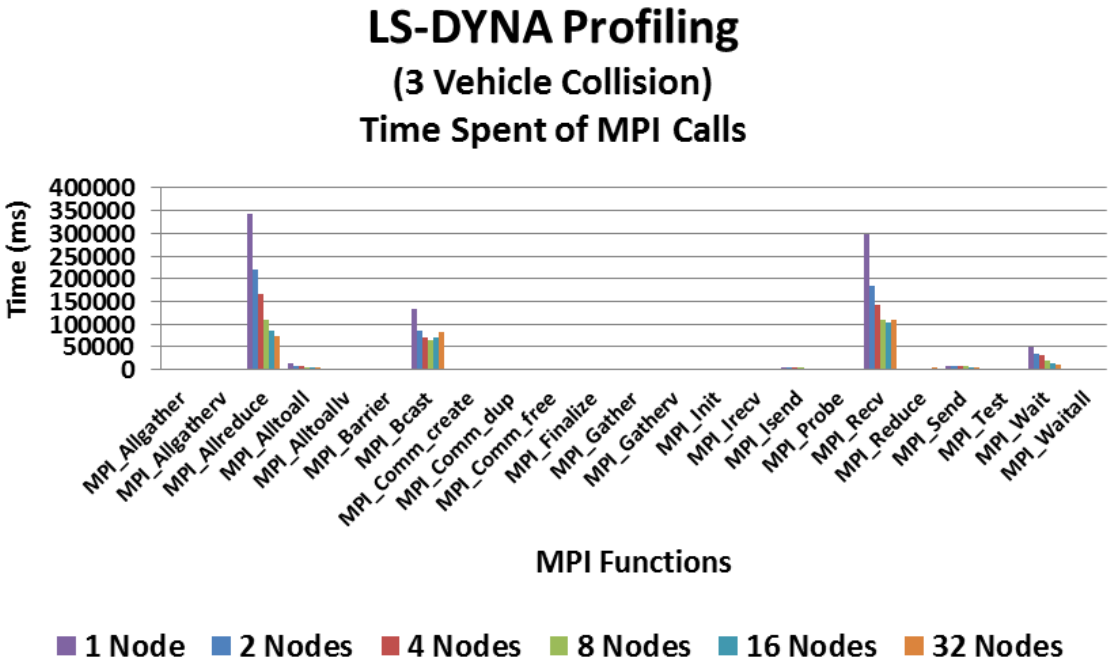




LS-DYNA Profiling – Time Spent by MPI Calls

- **Majority of the MPI time is spent on MPI_recv and MPI Collective Ops**
  - MPI_Recv(36%), MPI_Allreduce(27%), MPI_Bcast(24%)
- **MPI communication time lowers gradually as cluster scales**
  - Due to the faster total runtime, as more CPUs are working on completing the job faster
  - Reducing the communication time for each of the MPI calls

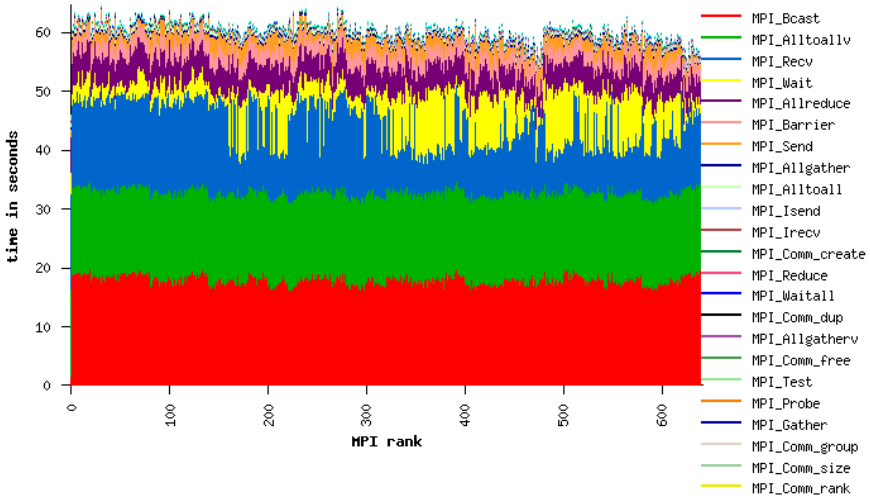




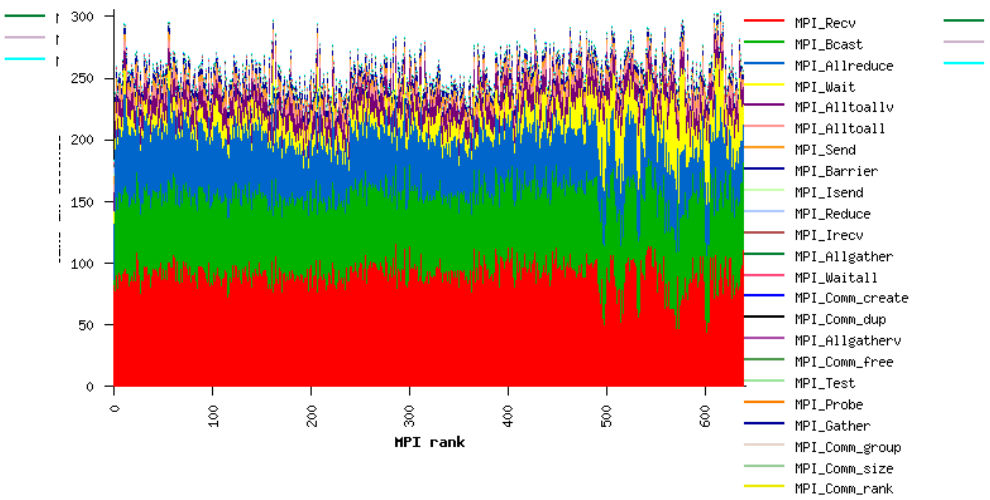
LS-DYNA Profiling – MPI Message Sizes

- **Similar communication characteristics seen on both input dataset**
  - Both exhibit similar communication patterns

Neon_refined_revised – 32 nodes

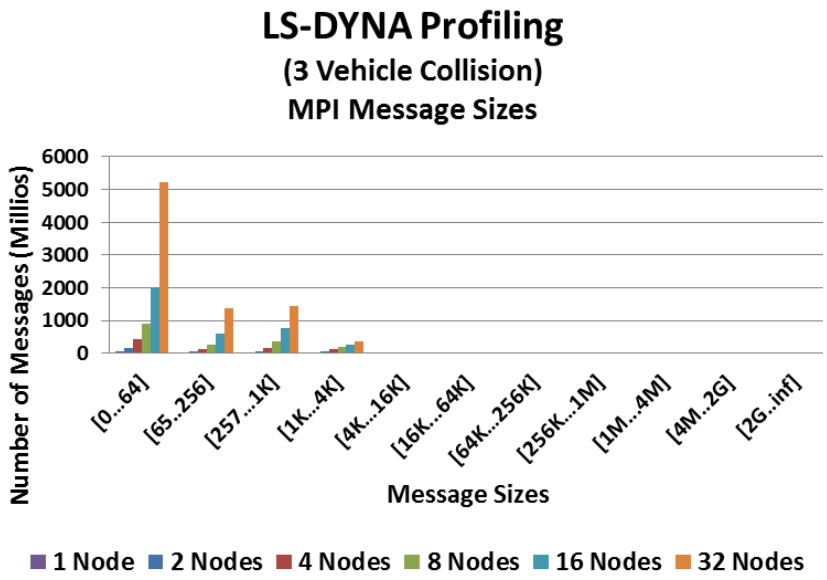


3 Vehicle Collision – 32 nodes

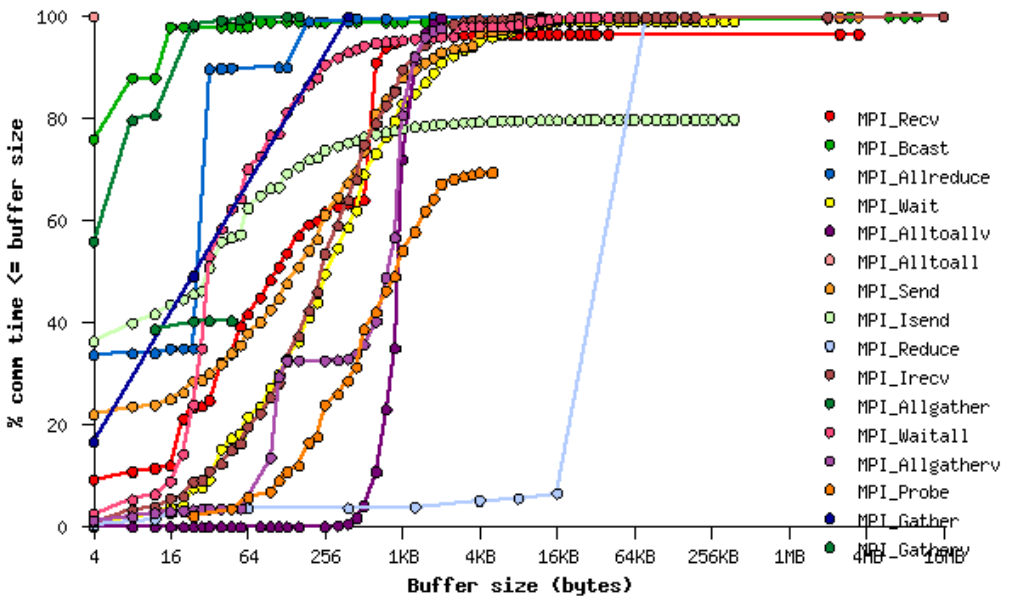


LS-DYNA Profiling – MPI Message Sizes

- **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

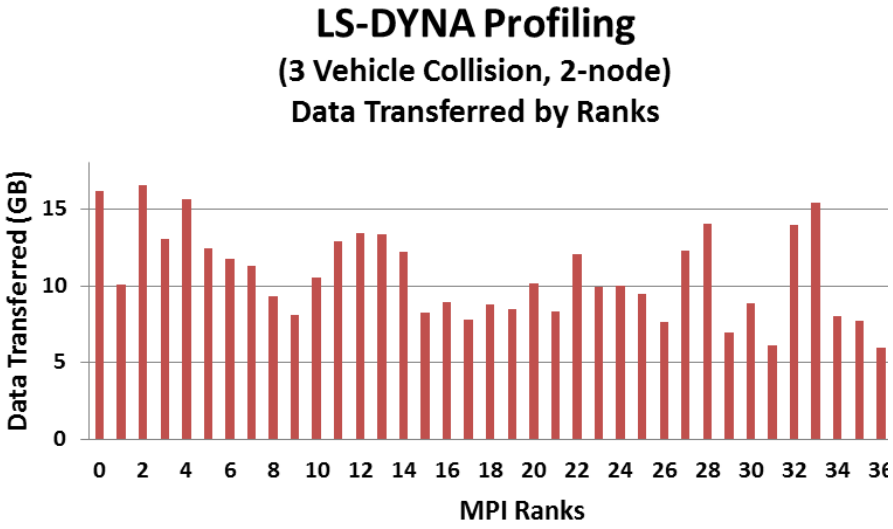


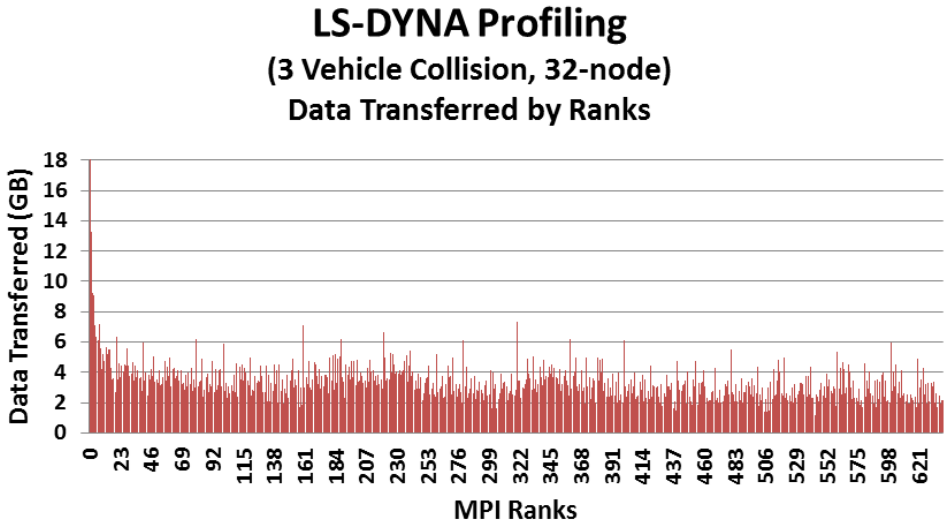
3 Vehicle Collision – 32 nodes



LS-DYNA Profiling – MPI Data Transfer

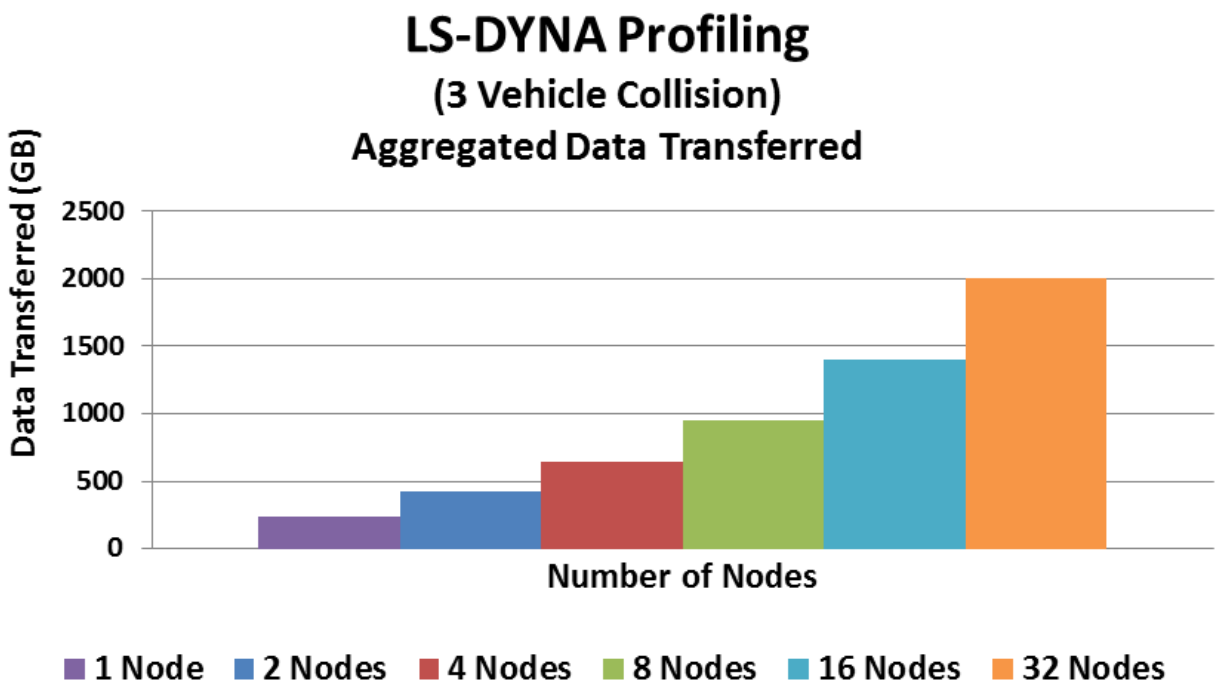
- **As the cluster grows, substantial less data transfers between MPI processes**
  - Drops from ~10GB per rank at 2-node vs to ~4GB at 32-node
  - Rank 0 contains higher transfers than the rest of the MPI ranks
  - Rank 0 responsible for file IO and uses MPI to communicates with the rest of the ranks





LS-DYNA Profiling – Aggregated Transfer

- **Aggregated data transfer refers to:**
  - Total amount of data being transferred in the network between all MPI ranks collectively
- **Large data transfer takes place in LS-DYNA**
  - Seen around 2TB at 32-node for the amount of data being exchanged between the nodes

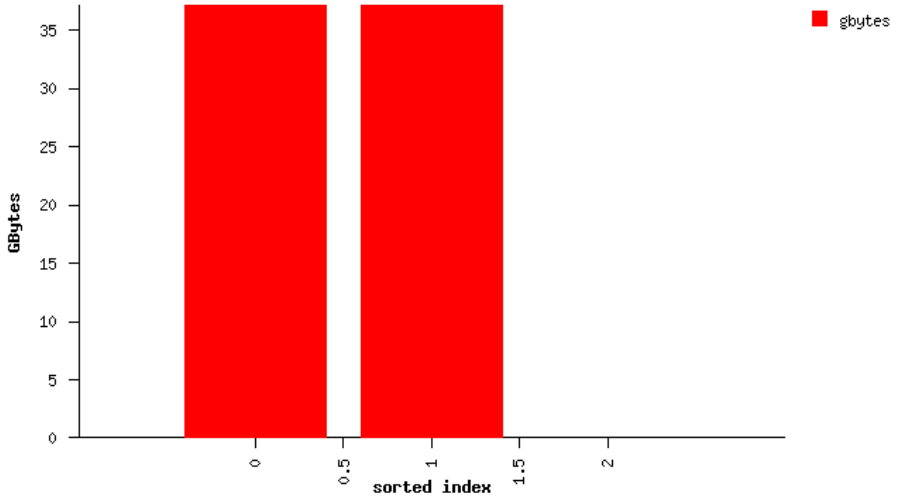


FDR InfiniBand

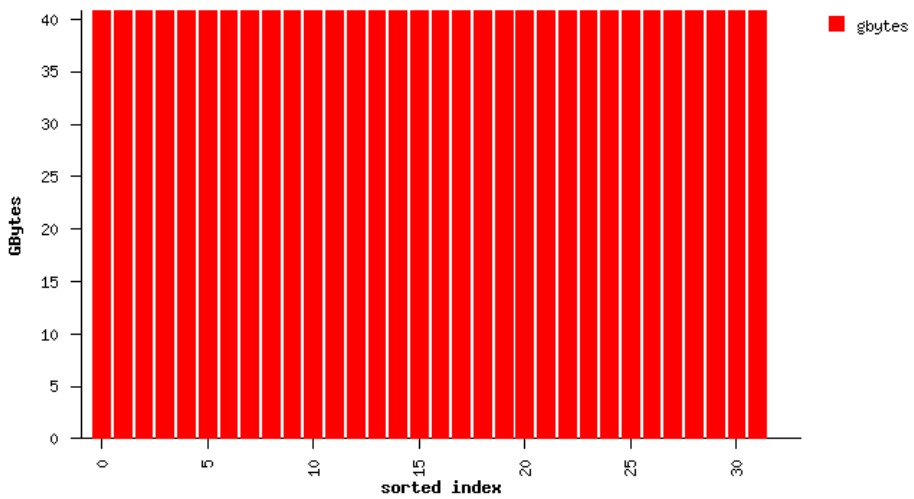
LS-DYNA Profiling – Memory Usage By Node

- **Uniform amount of memory consumed for running LS-DYNA**
  - About 38GB of data is used for per node
  - Each node runs with 20 ranks, thus about 2GB per rank is needed
  - The same trend continues for 2 to 32 nodes

3 Vehicle Collision – 2 nodes



3 Vehicle Collision – 32 nodes



LS-DYNA – Summary

- **Performance**
  - Intel Xeon E5-2680v2 (Ivy Bridge) and FDR InfiniBand enable LS-DYNA to scale
    - Up to 20% over Sandy Bridge, Up to 70% over Westmere, 167% over Nehalem
  - FDR InfiniBand delivers superior scalability in application performance
    - Provides higher performance by over 7 times for 1GbE and over 5 times for 10/40GbE at 32 nodes
  - Dual rank memory module provides better speedup
    - Using Dual Rank 1600MHz DIMM is ~9% faster than single rank 1866MHz DIMM
  - HPC Advisory Council performs better than the best published results on TopCrunch
    - 9% to 27% of higher performance than best published results on TopCrunch (Feb 2014)
- **Tuning**
  - FCA and MXM enhances LS-DYNA performance at scale for Open MPI
    - Provide a speedup of 18% over untuned baseline run at 32 nodes
  - As the CPU/MPI time ratio shows significantly more computation is taken place
- **Profiling**
  - Majority of MPI calls are for (blocking and non-blocking) point-to-point communications
  - Majority of the MPI time is spent on MPI_recv and MPI Collective operations

Thank You

HPC Advisory Council