HPC ADVISORY COUNCIL


# LS-DYNA
# Performance Benchmark and Profiling

April 2012

DELL

LSTC Livermore Software Technology Corp.

intel

Mellanox TECHNOLOGIES

# Note

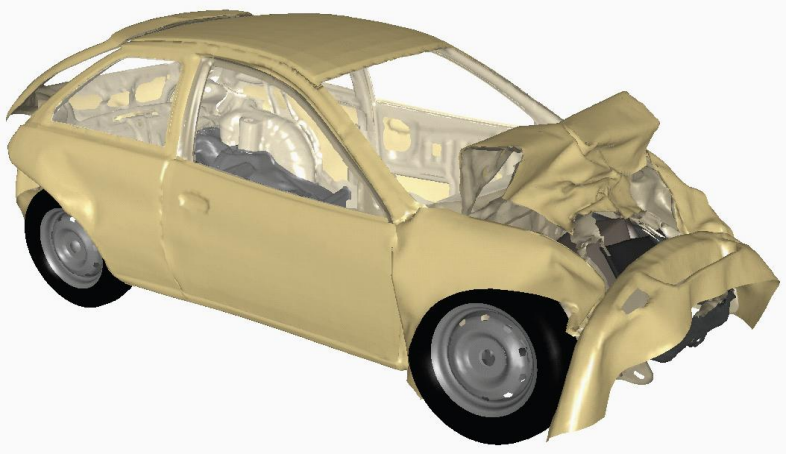
- **The following research was performed under the HPC Advisory Council activities**
  - Participating vendors: Intel, Dell, Mellanox, LSTC
  - Compute resource - HPC Advisory Council Cluster Center

- **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

- **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 16-node (256-core) cluster**
  - Dual-Socket Eight-Core Intel E5-2680 @ 2.70 GHz CPUs (Static max Perf in BIOS)
  - Memory: 64GB memory, DDR3 1600 MHz
  - OS: RHEL 6.2, OFED 1.5.3 InfiniBand SW stack
  - Hard Drives: 24x 250GB 7.2 RPM SATA 2.5” on RAID 0
- **Intel Cluster Ready certified cluster**
- **Mellanox ConnectX-3 VPI InfiniBand adapters**
- **Mellanox SwitchX 6036 VPI InfiniBand switch**
- **MPI: Intel MPI 4 U3, Open MPI 1.5.5 (KNEM 0.9.8), Platform MPI 8.2**
- **Application: LS-DYNA mpp971_s_r6.0.0**
- **Benchmark datasets:**
  - 3cars: 3 Vehicle Collision

# 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

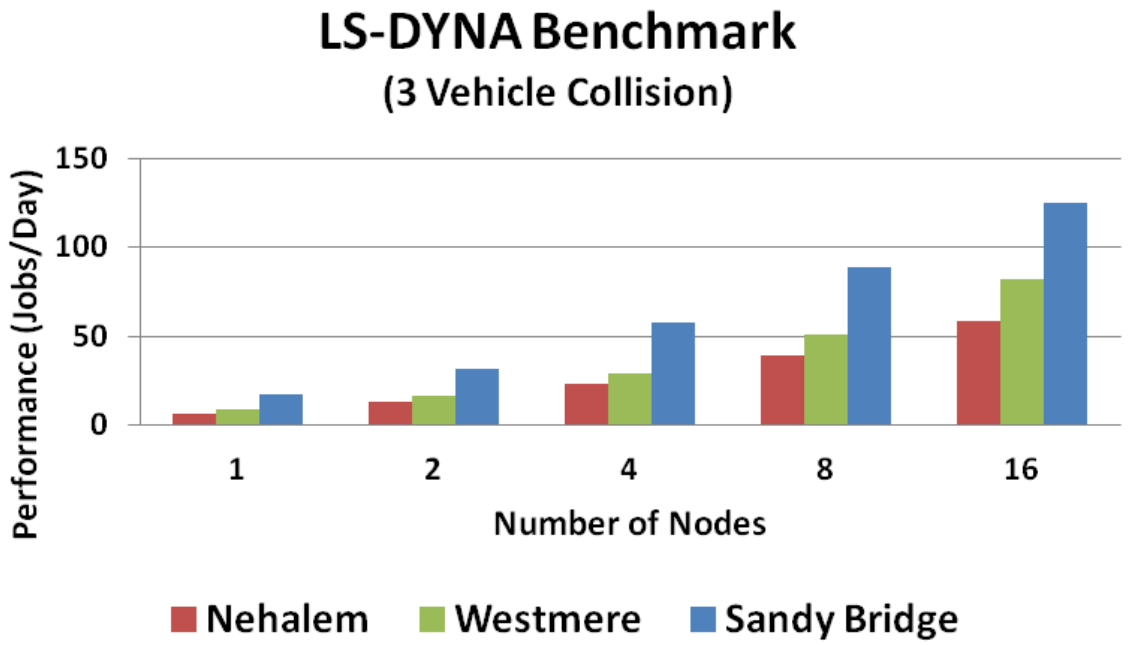
# PowerEdge R720xd

## 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 – Processor Generations

- **Intel E5-2600 Series (Sandy Bridge) outperforms prior generations**
  - Up to 61% higher performance than Intel Xeon X5670 (Westmere)
  - Up to 123% higher performance than Intel Xeon X5570 (Nehalem)
- **System components used:**
  - Sandy Bridge: Dual-socket Intel E5-2680 @ 2.7GHz, 1600MHz DIMMs, FDR IB
  - Westmere: Dual-socket Intel x5670 @ 2.93GHz, 1333MHz DIMMs, QDR IB
  - Nehalem: Dual-socket Intel x5570 @2.93GHz, 1333MHz DIMMs, QDR IB

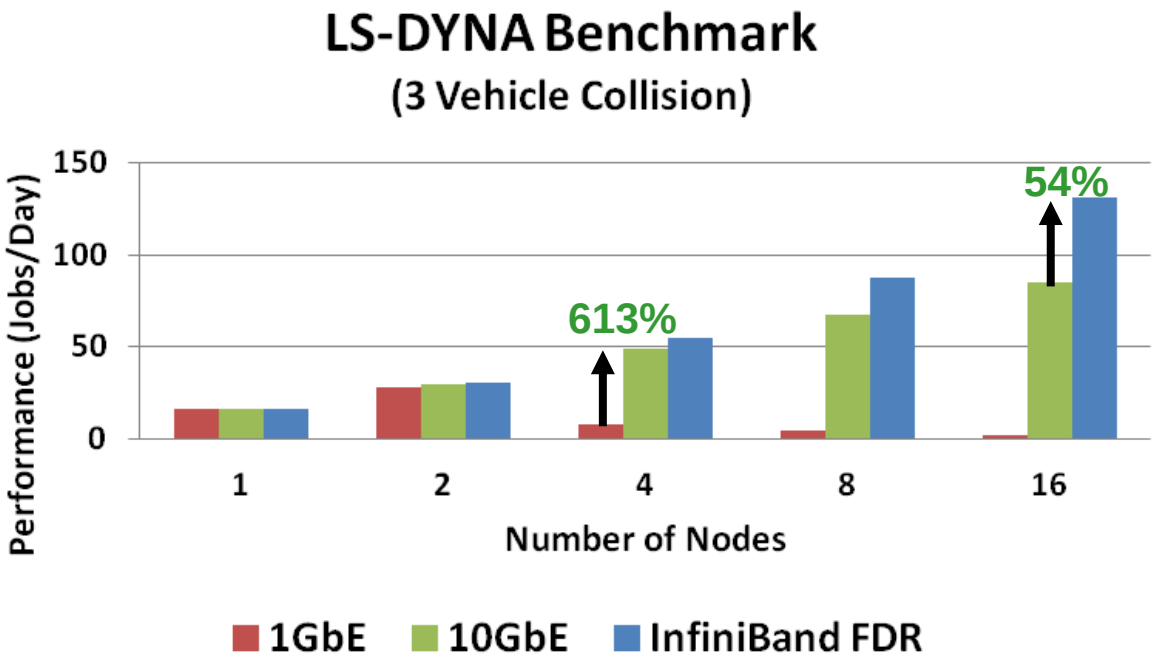



*Higher is better*

*InfiniBand FDR*

# LS-DYNA Performance – Interconnects

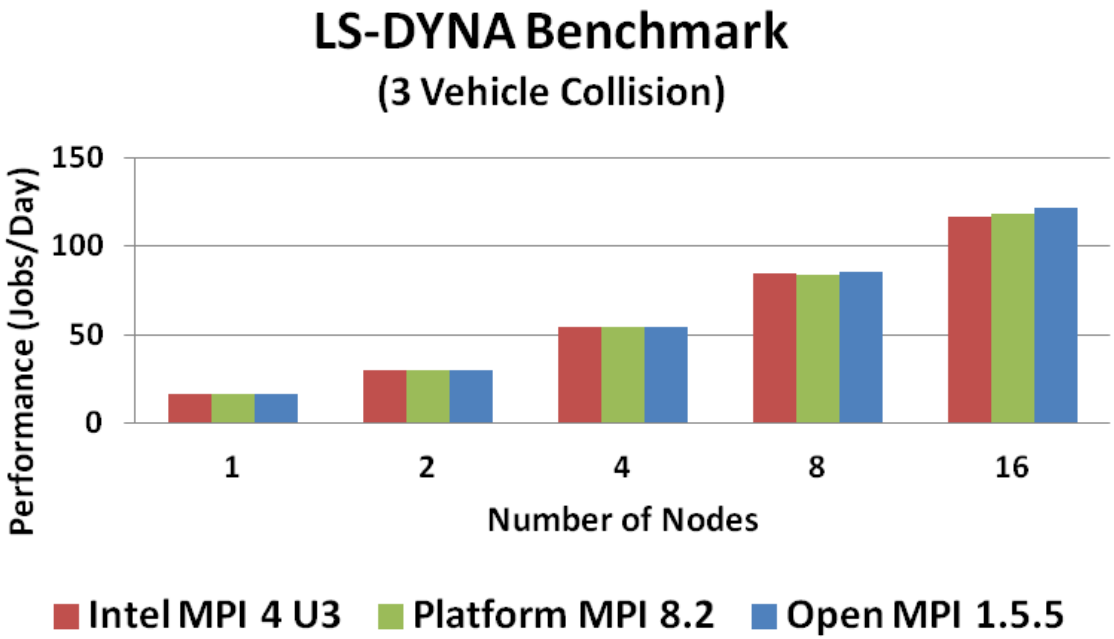
- **InfiniBand FDR enables the highest cluster productivity**
  - Increasing the performance by up to 613% over 1GbE at 4-node
  - Increasing the performance by up to 54% over 10GbE at 16-node
- **Ethernet performance begins to plummet after reaching to a few nodes**
  - 1GbE performance degradation begins after 4-node
  - 10GbE shows scalability issue beyond 8-node




***Higher is better***

# LS-DYNA Performance – MPI Implementations

- **All MPI performs similarly in performance**
  - Reflects each MPI implementation handles efficiently for the MPI pt-to-pt transfers
  - Profiling shows around 42% of time spent on MPI_Recv (point-to-point transfer)

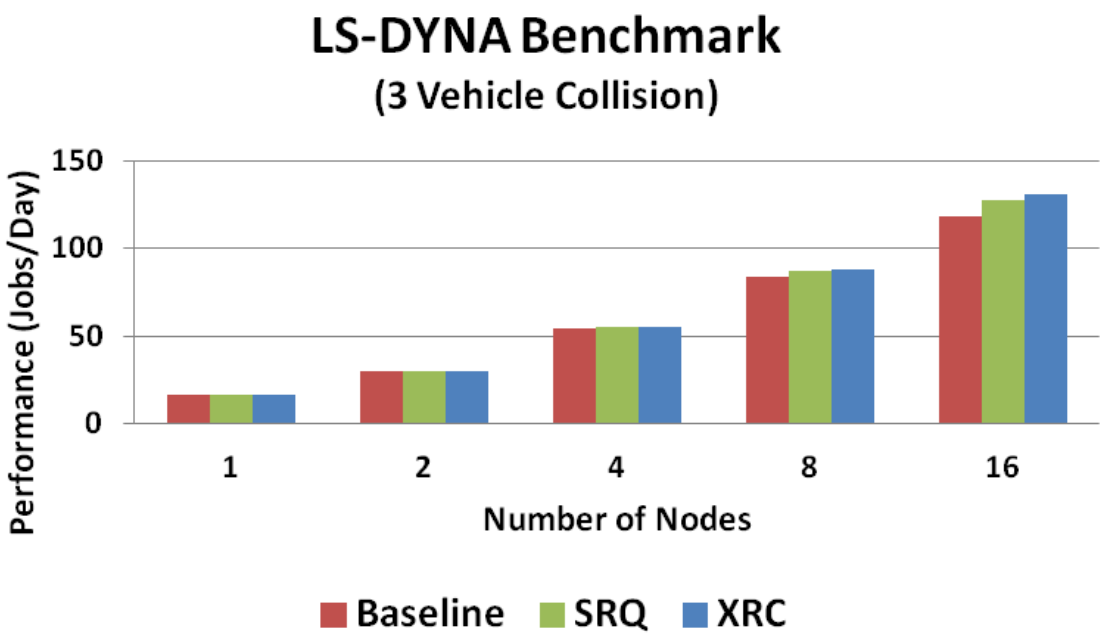



*Higher is better*

*InfiniBand FDR*

# LS-DYNA Performance – MPI Tuning

- **Using XRC and SRQ would provide marginally better in speedup**
  - XRC: Provide a speedup of 6% over baseline
  - SRQ: Provide a speedup of 3% over baseline

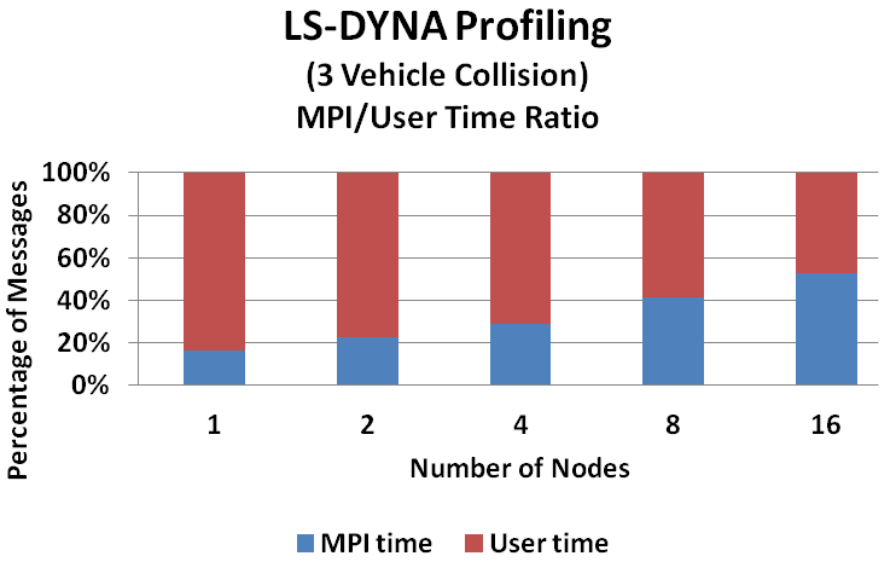


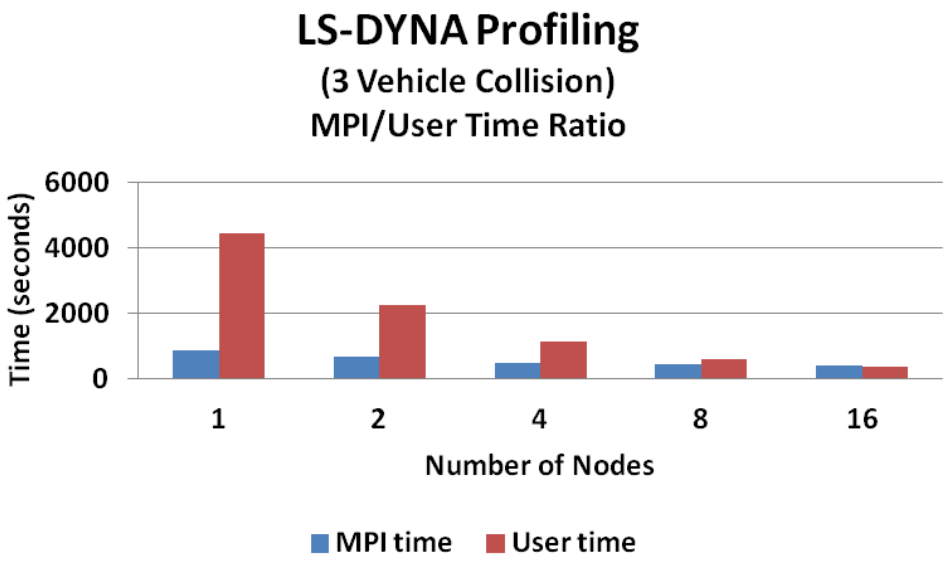

*Higher is better*

*InfiniBand FDR*

# 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

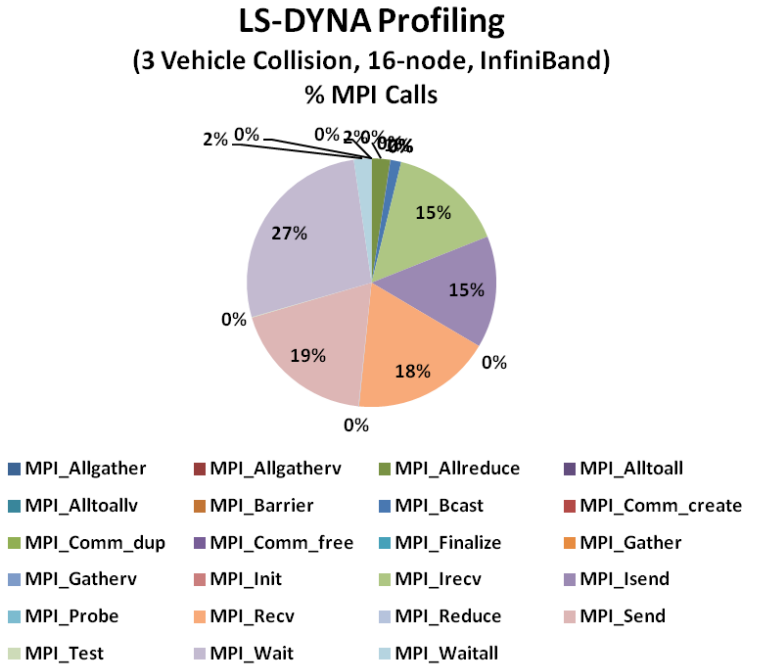


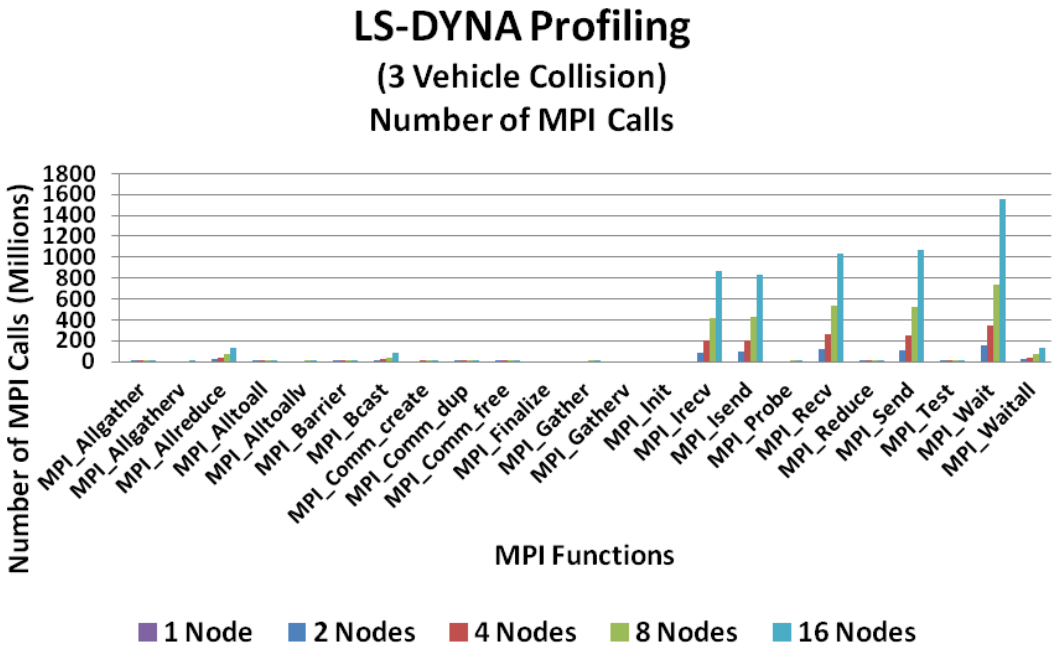




***InfiniBand FDR***

# LS-DYNA Profiling – MPI Calls

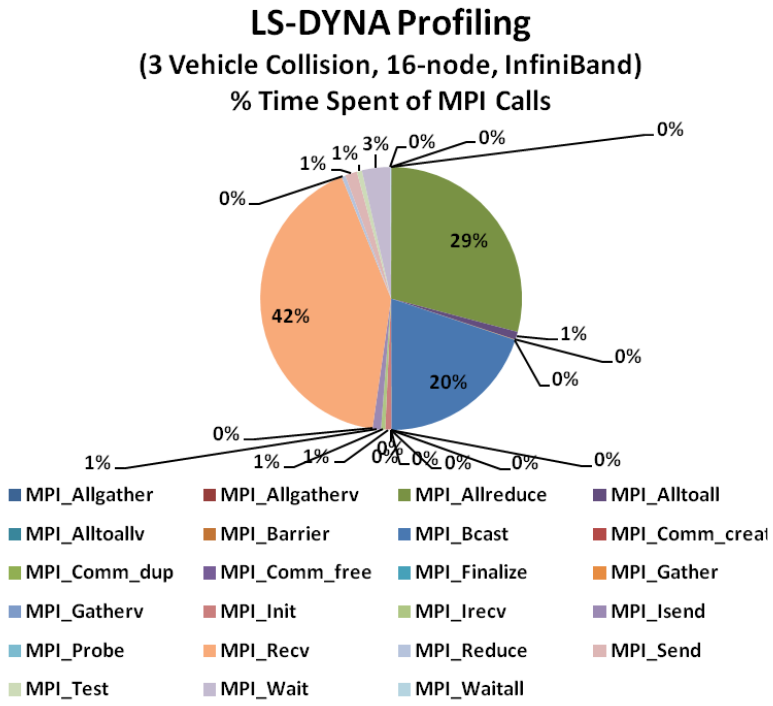
- **MPI_Wait, MPI_Send and MPI_Recv are the most used MPI calls**
  - MPI_Wait(27%), MPI_Send(19%), MPI_Recv(18%), MPI_Isend(15%), MPI_Irecv(15%)
- **LS-DYNA has majority of MPI point-to-point calls for data transfers**
  - Either blocking or non-blocking point-to-point transfers are seen

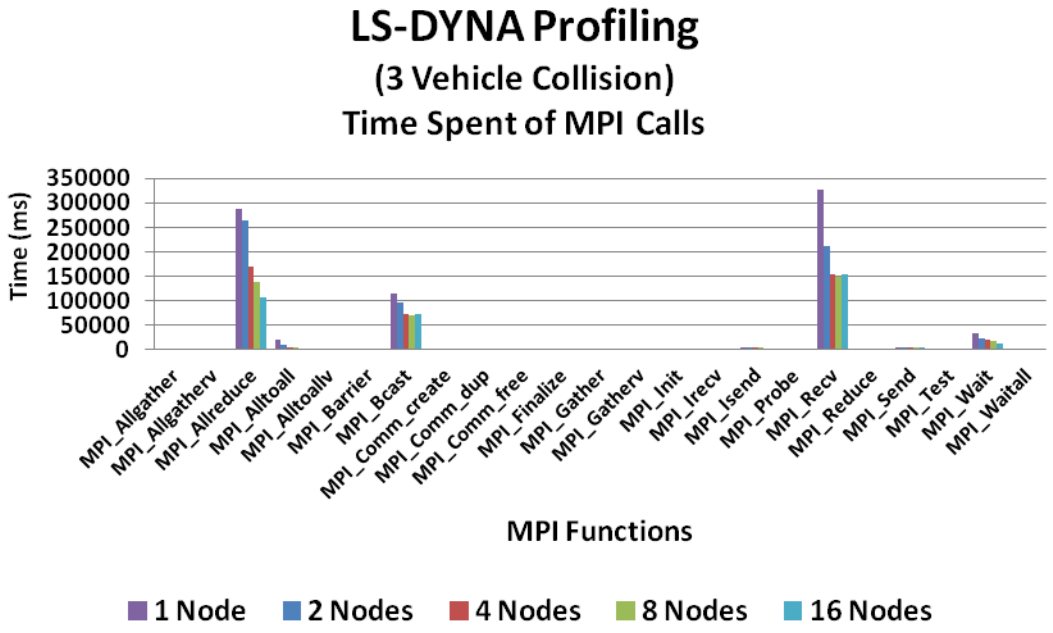


# LS-DYNA Profiling – Time Spent by MPI Calls

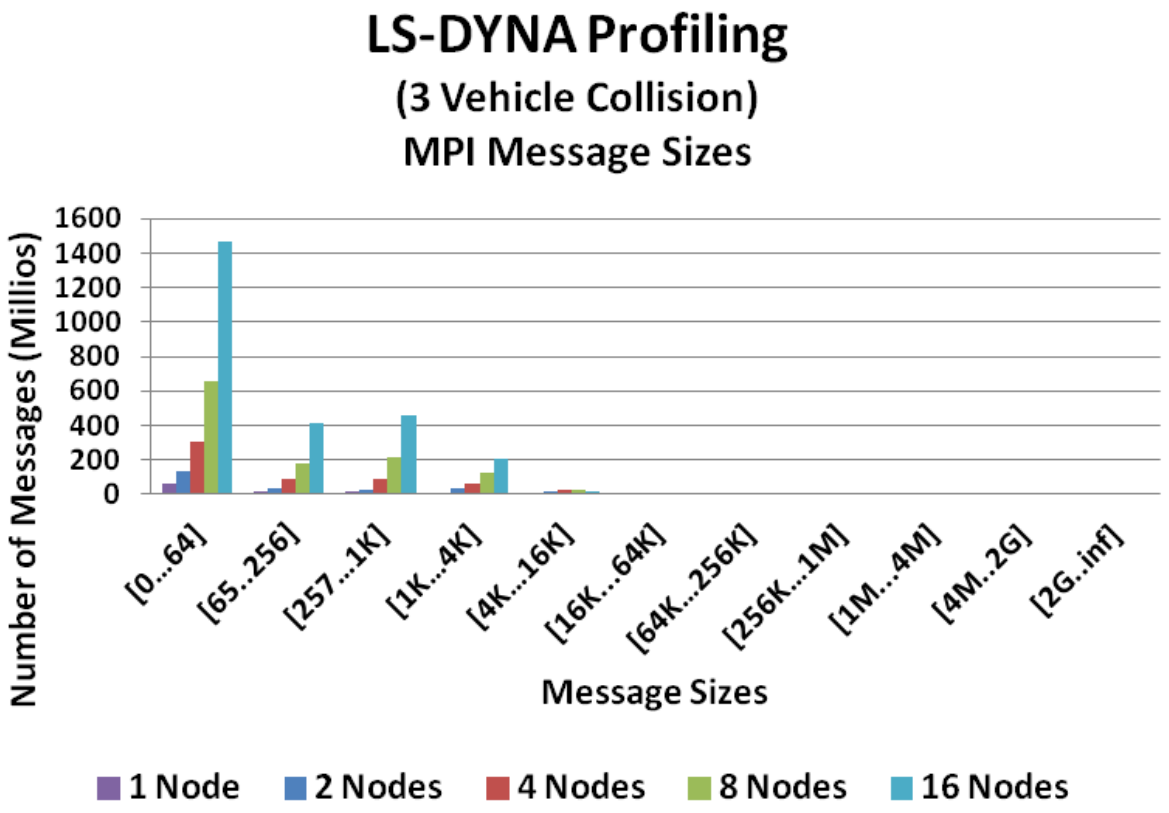
- **Majority of the MPI time is spent on MPI_recv and MPI Collective Ops**
  - MPI_Recv(42%), MPI_Allreduce(29%), MPI_Bcast(20%)
- **MPI communication time lowers gradually as cluster scales**
  - Due to the faster total runtime, as more CPUs are working on completing the job faster
  - Reducing the communication time for each of the MPI calls



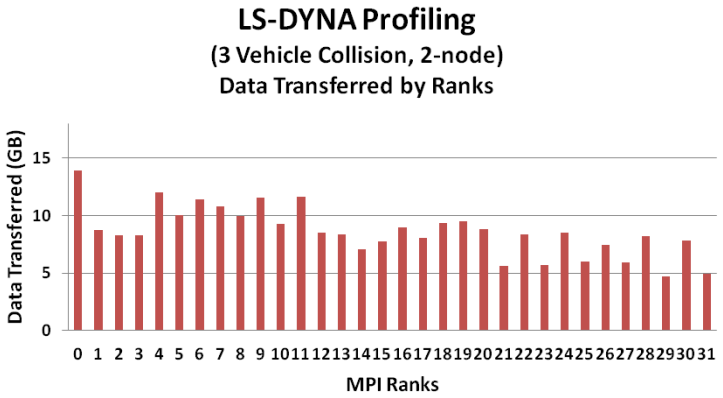
# LS-DYNA Profiling – MPI Message Sizes

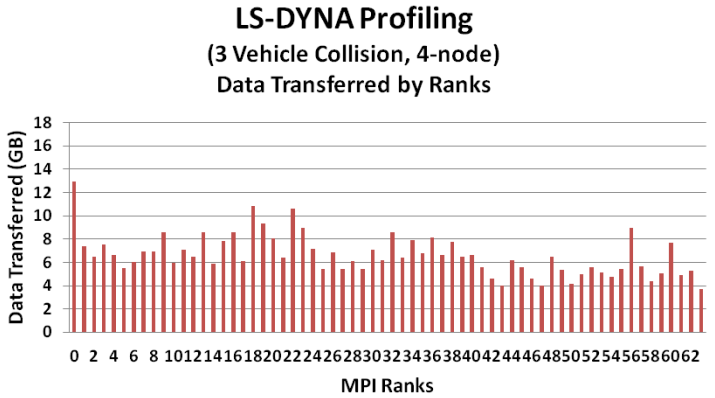
- **Most of the MPI messages are in the medium sizes**
  - Most message sizes are between 0 to 64byte
  - MPI messages are concentrated in the small message sizes under 4KB

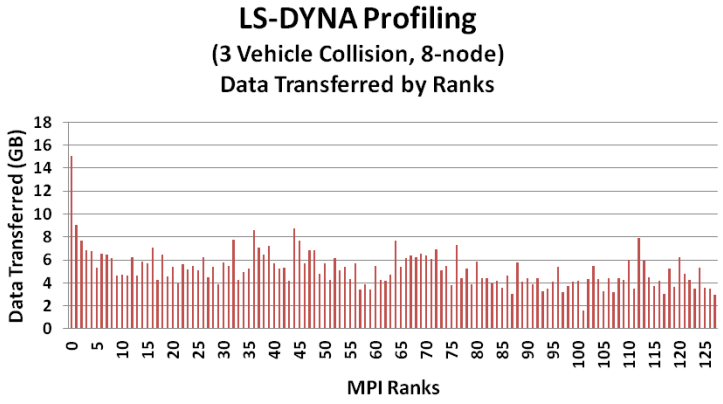


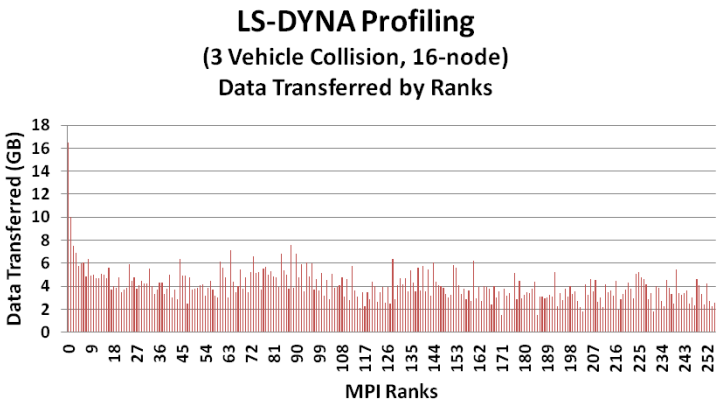
# LS-DYNA Profiling – MPI Data Transfer

- **As the cluster grows, substantial less data transfers between MPI processes**
  - Drops from ~8-10GB per rank at 1-node vs to ~4GB at 16-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



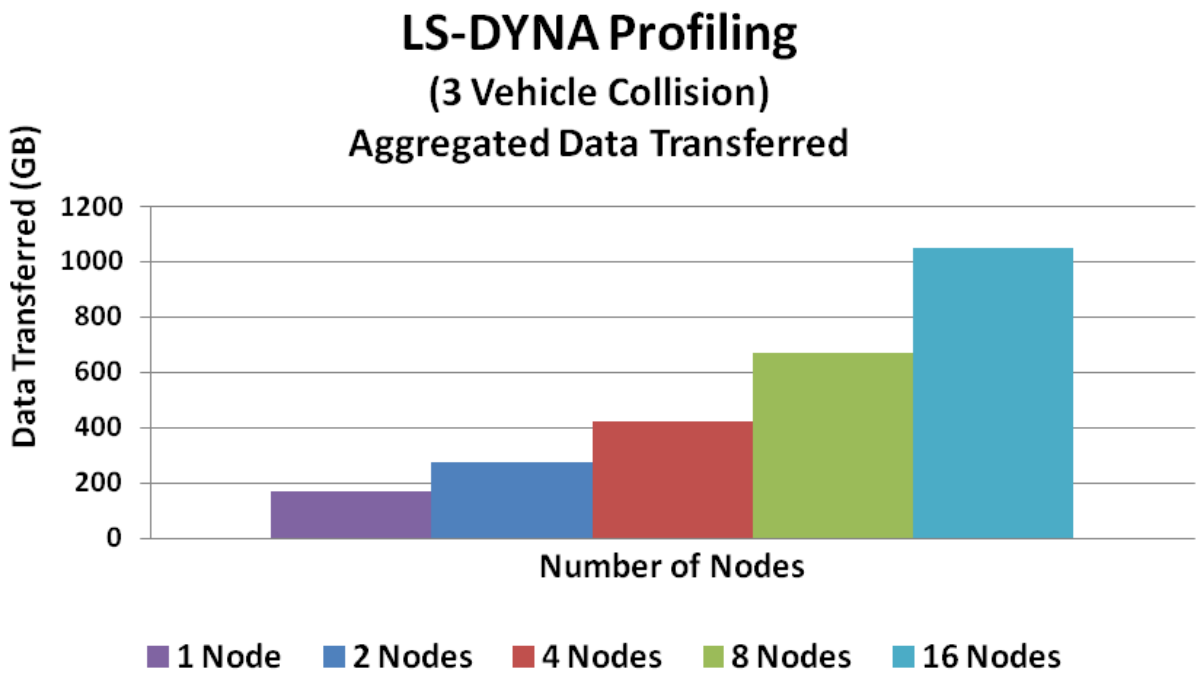



# LS-DYNA Profiling – Aggregated Transfer

- **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 1TB at 16-node for the amount of data being exchanged between the nodes




*InfiniBand FDR*

# LS-DYNA – Summary

- **Performance**
  - Intel Xeon E5-2670 procs (Sandy Bridge) and InfiniBand FDR enable LS-DYNA to scale
    - Provide up to 61% over the X5670 (Westmere)
    - Provide up to 123% over the X5570 (Nehalem)
  - InfiniBand FDR allows LS-DYNA to run at the highest network throughput at 56Gbps
  - Ethernet would not allow scale, ended up wasting valuable system resources
  - All MPI implementations tested (Intel, Platform, Open MPI) show good performance
- **Tuning**
  - MPI tuning (with XRC) provides some benefits for 6% at 16-node
  - As the CPU/MPI time ratio shows significantly more computation is taken place
  - Spreading the computational workload to more nodes can get job done faster
- **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