HPC ADVISORY COUNCIL

LS-DYNA
Performance Benchmark and Profiling

August 2012

AMD The future is fusion

DELL

LSTC Livermore Software Technology Corp.

Mellanox TECHNOLOGIES

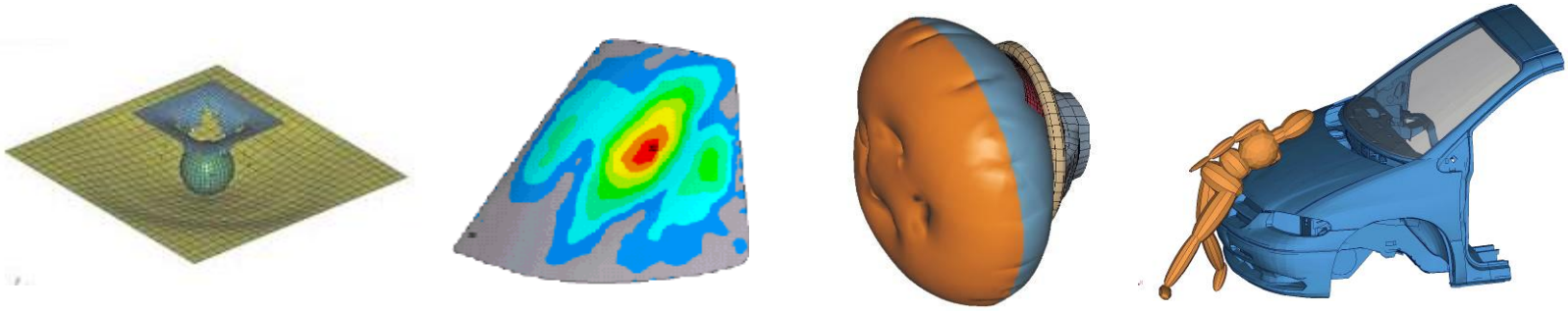
Note

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

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

LS-DYNA

- **LS-DYNA SMP (Shared Memory Processing)**
  - Optimize the power of multiple CPUs within single machine
- **LS-DYNA MPP (Massively Parallel Processing)**
  - The MPP version of LS-DYNA allows to run LS-DYNA solver over High-performance computing cluster
  - Uses message passing (MPI) to obtain parallelism
- **Many companies are switching from SMP to MPP**
  - For cost-effective scaling and performance



Objectives

- **The following was done to provide best practices**
  - LS-DYNA performance benchmarking
  - Understanding LS-DYNA communication patterns
  - Ways to increase LS-DYNA productivity
  - Network interconnects comparisons

- **The presented results will demonstrate**
  - The scalability of the compute environment
  - The capability of LS-DYNA to achieve scalable productivity
  - Considerations for performance optimizations

Test Cluster Configuration

- **Dell™ PowerEdge™ R815 11-node Quad-socket (704-core) cluster**
- **AMD™ Opteron™ 6276 (code name "Interlagos") 16-core @ 2.3 GHz CPUs**
- **Memory: 128GB memory per node DDR3 1333MHz**
- **Mellanox ConnectX-3 VPI adapters**
- **Mellanox SwitchX 6036 36-Port 40Gb/s Ethernet Switch**
- **OS: RHEL 6.2, MLNX-OFED 1.5.3 InfiniBand SW stack**
- **MPI: Platform MPI 8.2**
- **Application: LS-DYNA mpp971_s_R6.1.0 (MPP)**
- **Benchmark workload:**
  - 3 Vehicle Collision Test simulation (3cars)
    - The 3 Vehicle Collision model comprises 0.8 million elements
  - Caravan Model (car2car)
    - The minivan model is based on NCAC, comprises 2.4 million elements

About Dell PowerEdge™ Platform Advantages

Best of breed technologies and partners

Combination of AMD™ Opteron™ 6200 series platform and Mellanox ConnectX InfiniBand on Dell HPC

Solutions provide the ultimate platform for speed and scale

- Dell PowerEdge R815 system delivers 4 socket performance in dense 2U form factor
- Up to 64 core/32DIMMs per server – 1344 core in 42U enclosure

Integrated stacks designed to deliver the best price/performance/watt

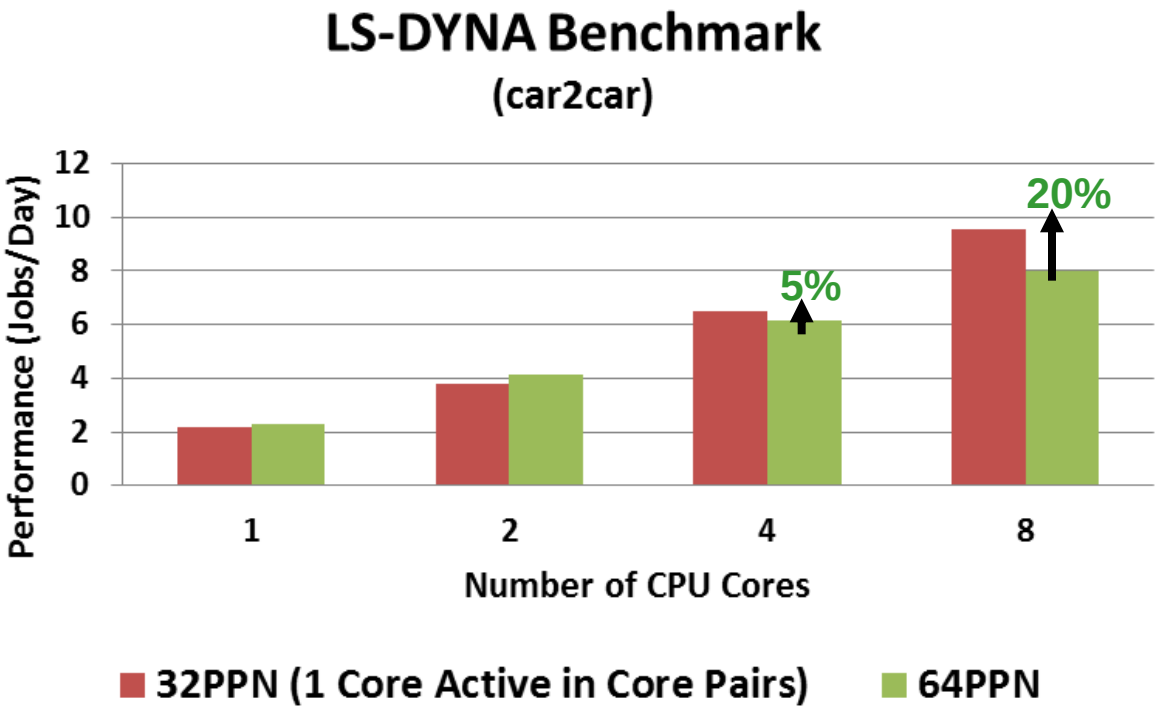
- 2x more memory and processing power in half of the space
- Energy optimized low flow fans, improved power supplies and dual SD modules

Optimized for long-term capital and operating investment protection

- System expansion
- Component upgrades and feature releases

LS-DYNA Performance – Core Pairs (car2car)

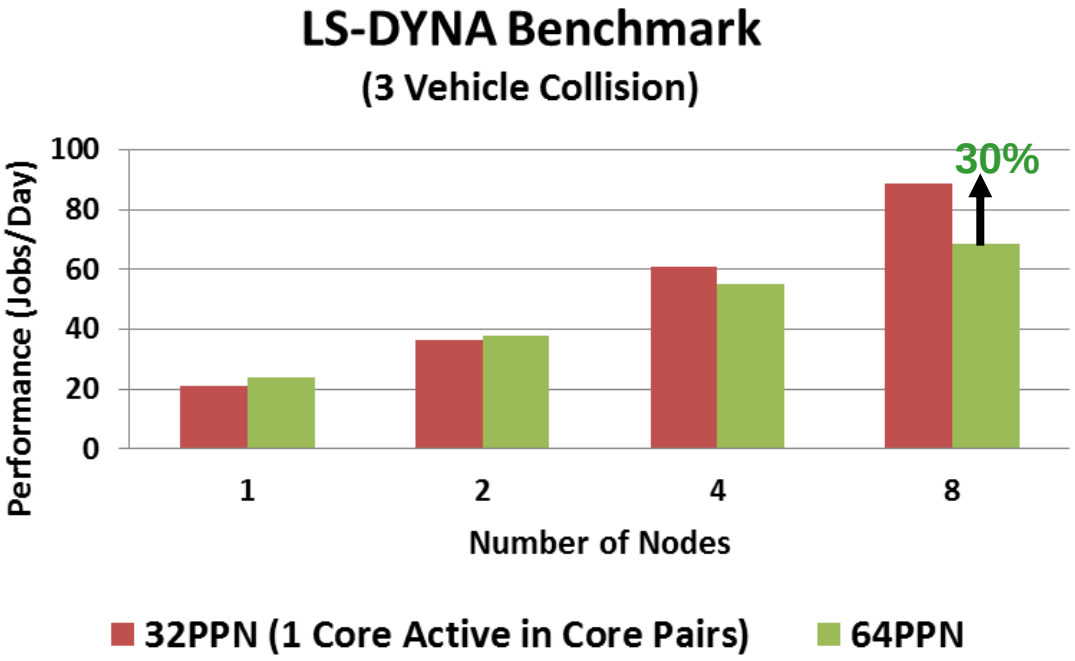
- **Input Dataset: car2car**
  - The minivan model is based on NCAC, comprises of 2.4 million elements
- **Boost from core pairs drives higher performance boost**
  - Improves the overall job performance by up to 20% than when both core pairs are active
  - Performance boost is enabled only when 1 of the 2 cores in a core pair is being used



Higher is better

LS-DYNA Performance – Core Pairs (3cars)

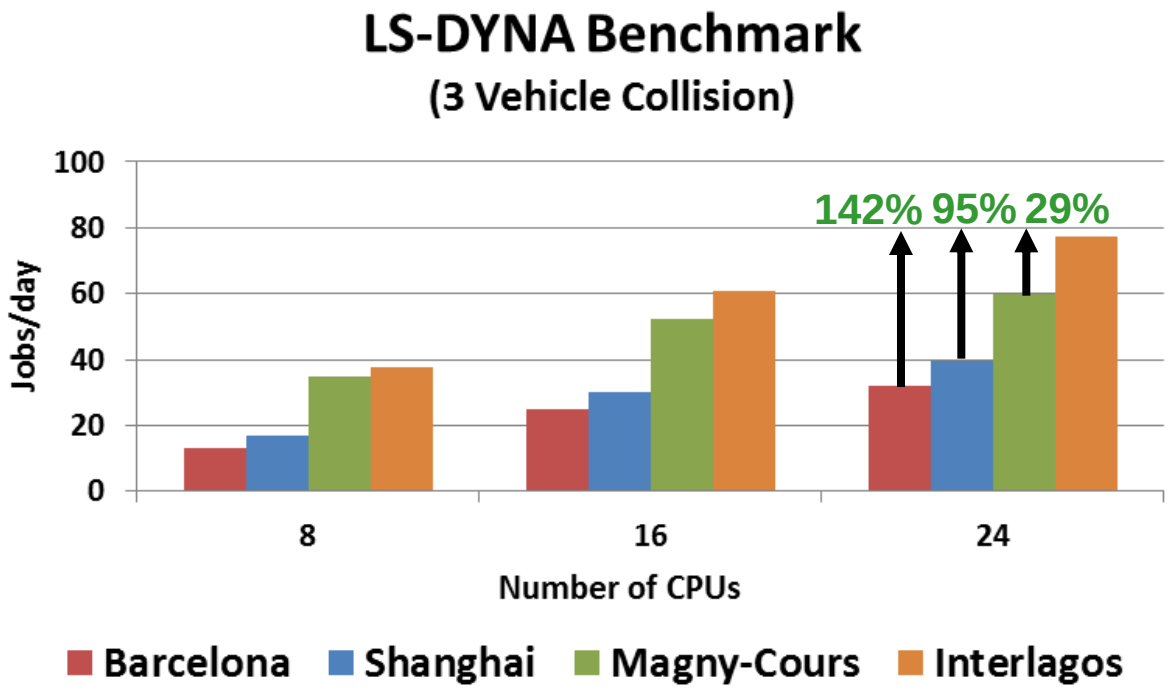
- **Input Dataset: 3cars**
  - A van crashes into the rear of a compact car, in turns crashes into a midsize car
- **Running jobs with 1 core per Core Pairs allows efficient usage**
  - Using 1 active core per Core Pairs would boost performance by 30% on a 8-node case
- **Reducing the CPU core counts while delivering greater performance**
  - Seen as an advantage as the license is based on CPU core count



Higher is better

LS-DYNA Performance – Processor Generations

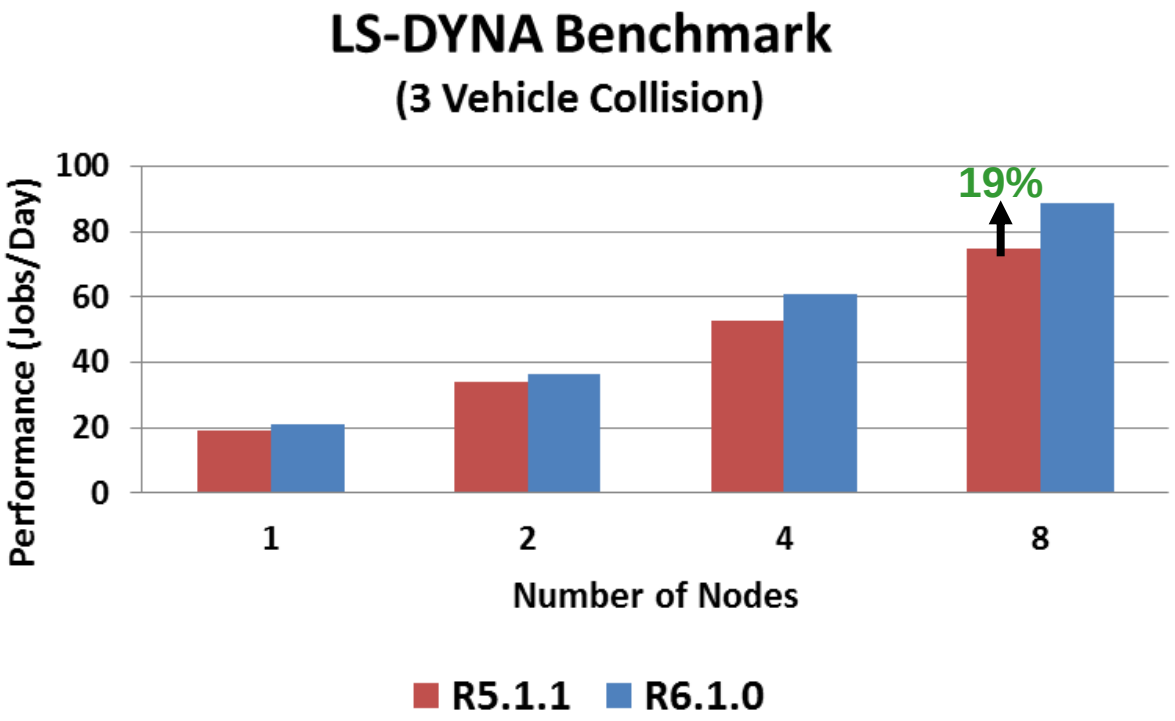
- **AMD "Interlagos" provides higher scalability than previous generations**
  - Improved by: 29% vs "Magny-Cours", 95% vs "Barcelona", 142% vs "Shanghai"
  - Barcelona /Shanghai with InfiniBand DDR and PCIe Gen1
  - Interlagos / Magny-Cours with InfiniBand QDR and PCIe Gen2



Higher is better

LS-DYNA Performance – Executable Version

- **Version 6.1.0 Executable delivers better LS-DYNA performance**
  - Improves overall job performance by up to 19% with 8 nodes compared to Version 5.1.1
- **Executable used:**
  - R6.1.0: ls-dyna_mpp_s_r6_1_0_74904_x64_suse111_open64455_avx_platformmpi
  - R5.1.1: mpp971_s_R5.1.1_71519_Open64avx_linux86-64_hpmpi_120

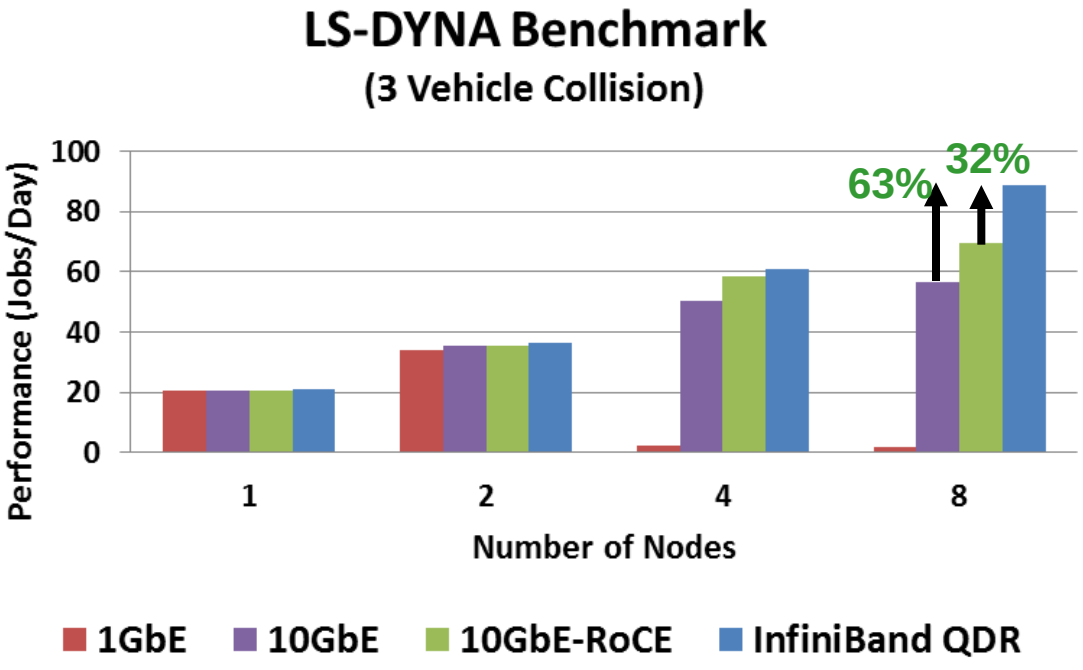


Higher is better

32 Cores/Node

LS-DYNA Performance – Interconnects (3cars)

- **InfiniBand QDR allows LS-DYNA to achieve the highest scalability**
  - Shows 63% better than using 10GbE
  - Shows 32% better than RDMA over Ethernet on 10GbE
  - Productivity of 1GbE interconnect plummeted after 2 nodes
- **RDMA over Ethernet avoids CPU involvement in processing network data**
  - Thus improve CPU productivity by offloading network transfer to the network adapter

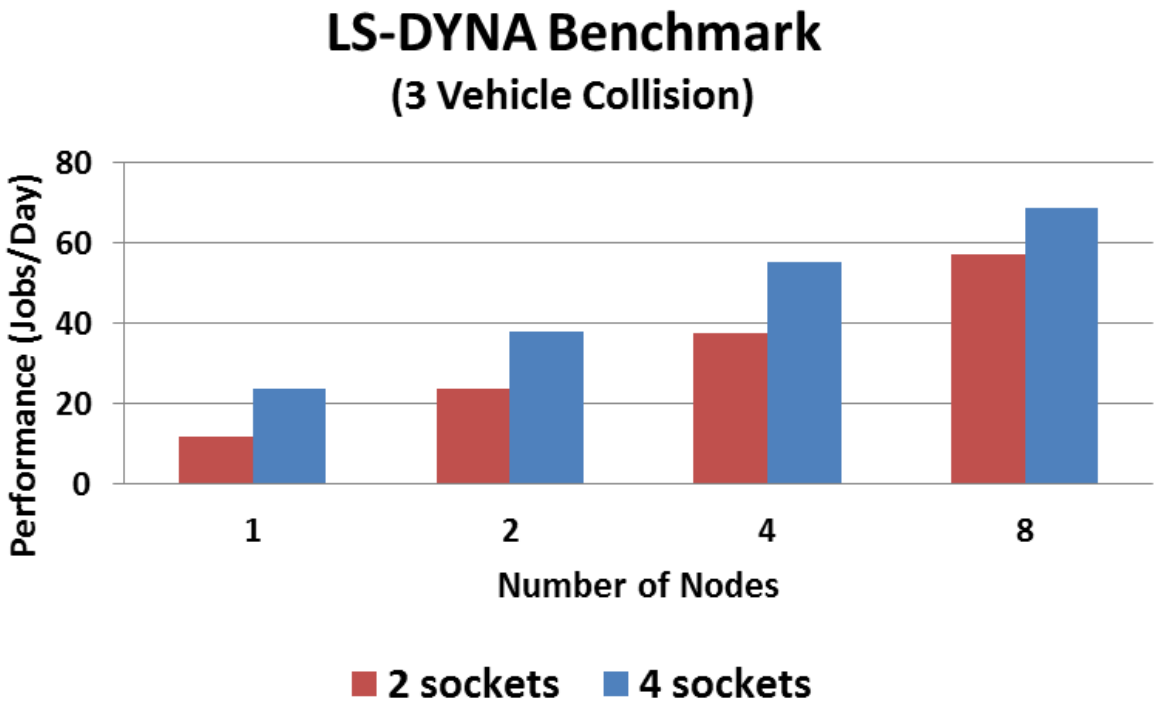


Higher is better

32 Cores/Node

LS-DYNA Performance – Processes Per Node (3cars)

- **Difference between 2-socket (32 PPN) and 4-socket (64 PPN) systems**
  - Using the MPP version of LS-DYNA, the 2-socket system provides better scalability

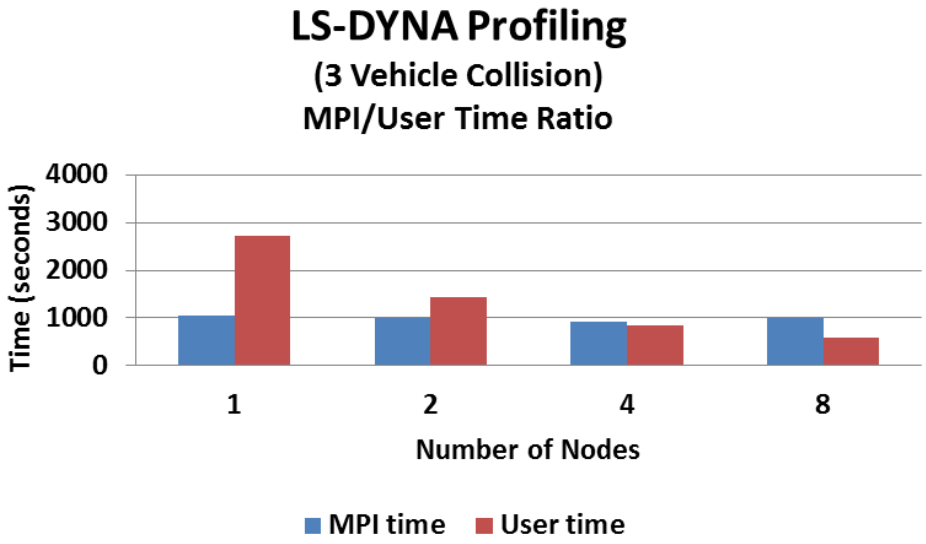


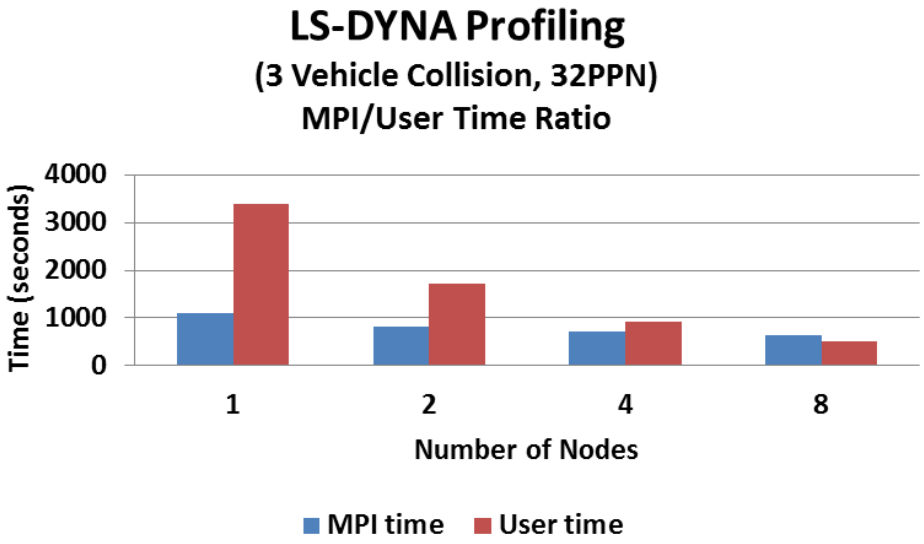
Higher is better

InfiniBand QDR

LS-DYNA Profiling – MPI/User Run Time

- **Both compute and communication time reduce when using 32PPN**
  - Less time spent on both MPI and compute time with less processes per node
- **InfiniBand reduces the overall runtime by spreading the workload**
  - Compute time cuts roughly by half as the node count doubles
  - MPI time also reduces as more nodes taking on the workload

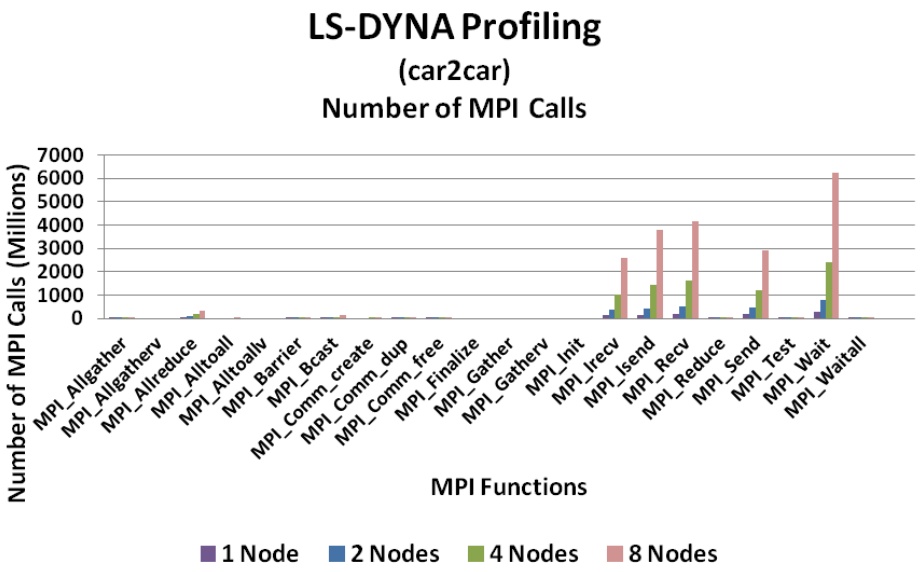


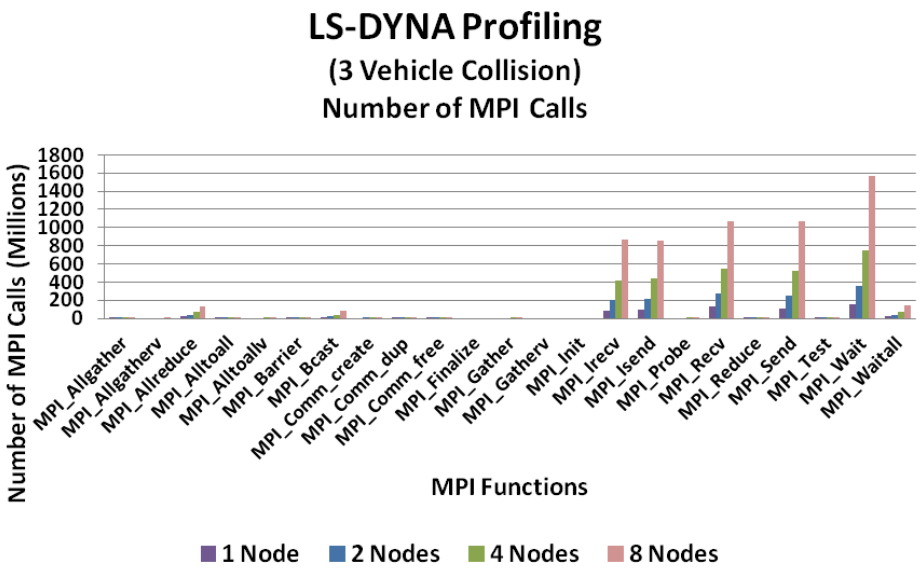


InfiniBand QDR

LS-DYNA Profiling – Number of MPI Calls

- **The most used MPI functions are for data transfers**
  - Car2car: MPI_Wait(31%), MPI_Recv(20%), MPI_Isend(19%) and MPI_Send(14%)
  - 3cars: MPI_Wait(27%), MPI_Send(19%), MPI_Recv(18%) and MPI_Isend(15%)
  - Reflects that LS-DYNA requires good network throughput for network communications
- **The number of calls increases proportionally as the cluster scales**

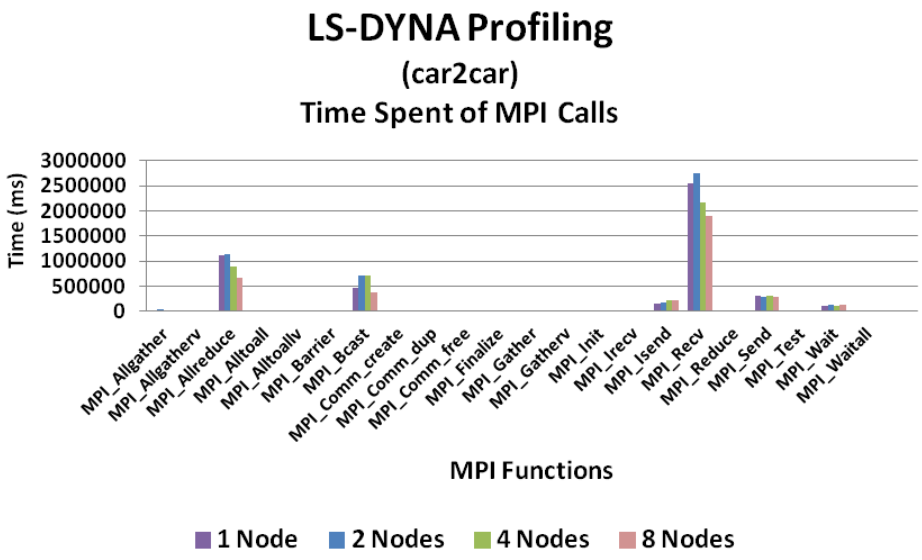


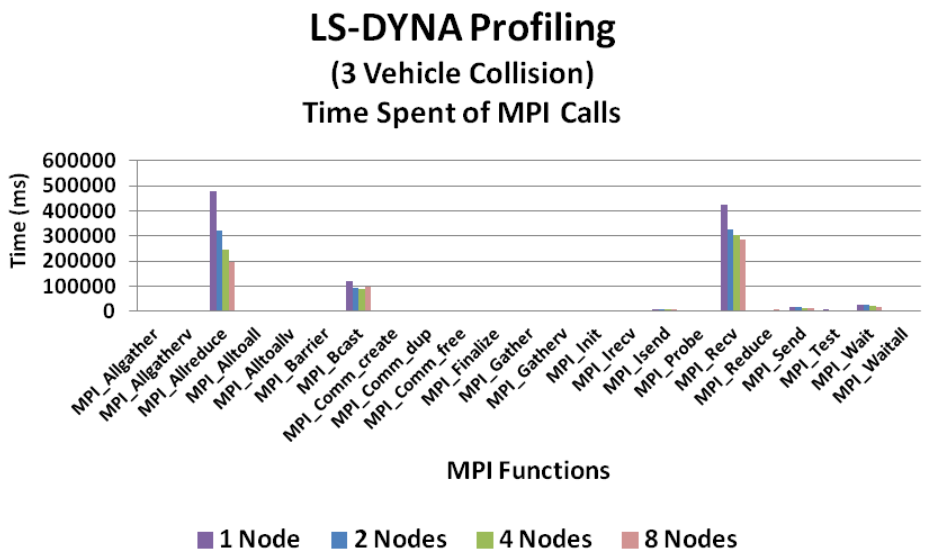


32 Cores/Node

LS-DYNA Profiling – Time Spent of MPI Calls

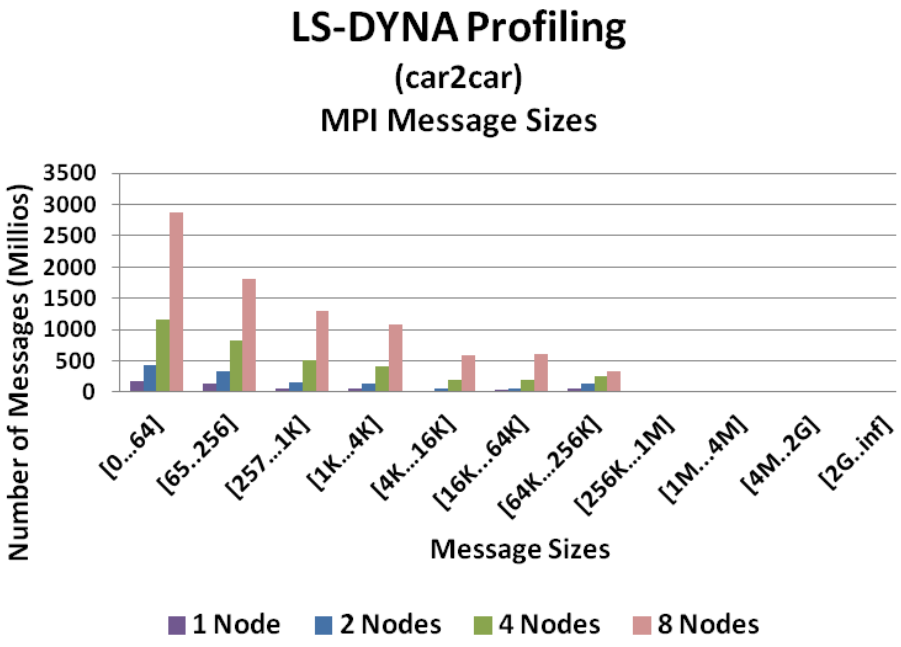
- **The time spent in MPI calls for both datasets are generally the same**
  - The car2car dataset involves significantly more data communications than 3car case
- **The MPI time is taken place in the following MPI functions:**
  - 3cars: MPI_Recv(45%), MPI_Allreduce(31%), MPI_Bcast(15%)
  - Car2car: MPI_Recv(52%), MPI_Allreduce(18%), MPI_Bcast(10%)
- **Communication time per call is reduced as more cluster nodes being added**

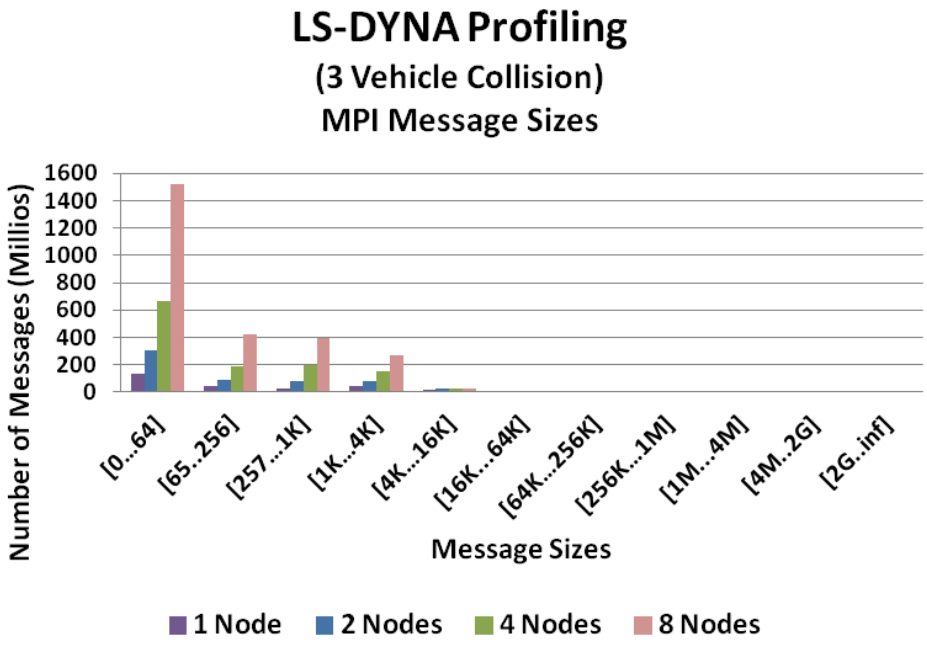




LS-DYNA Profiling – MPI Message Sizes

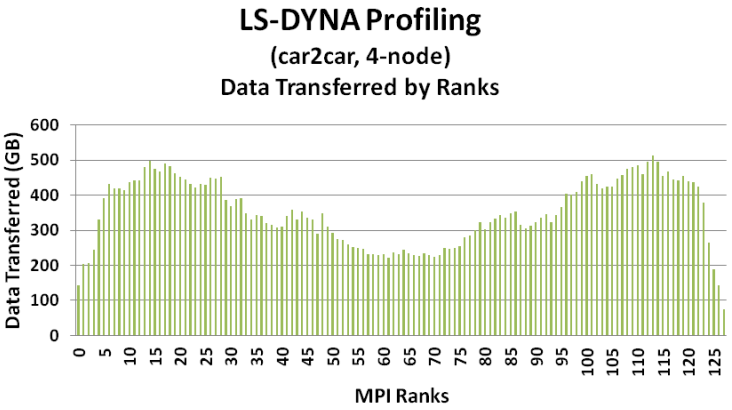
- **Majority of the MPI messages are small message sizes**
  - The largest cencentration is between 0B and 64B
  - Car2car appears to have the message size range from small to medium messages
  - 3cars appears to only concentrate in the small messages

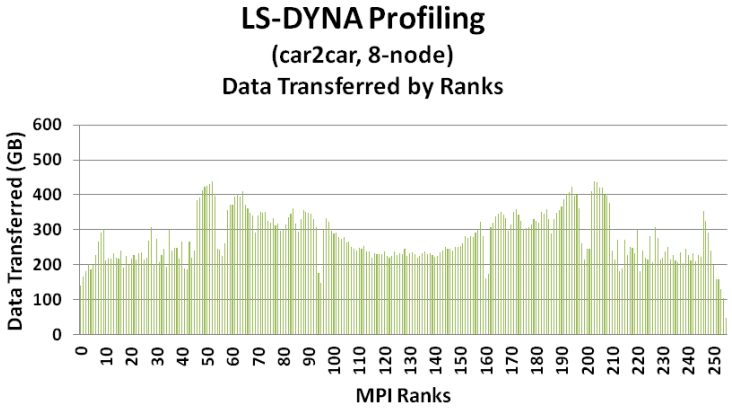


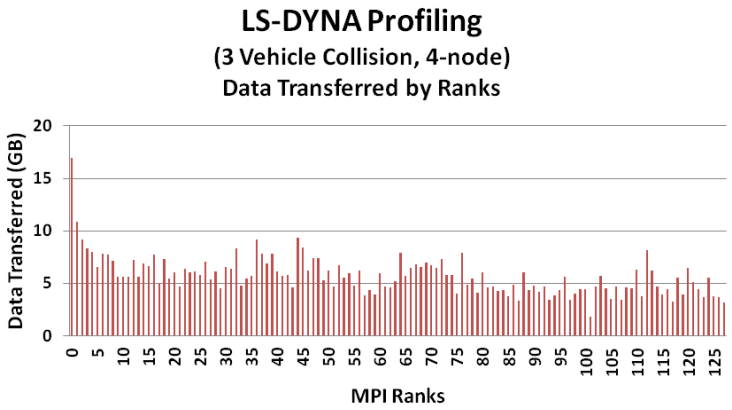


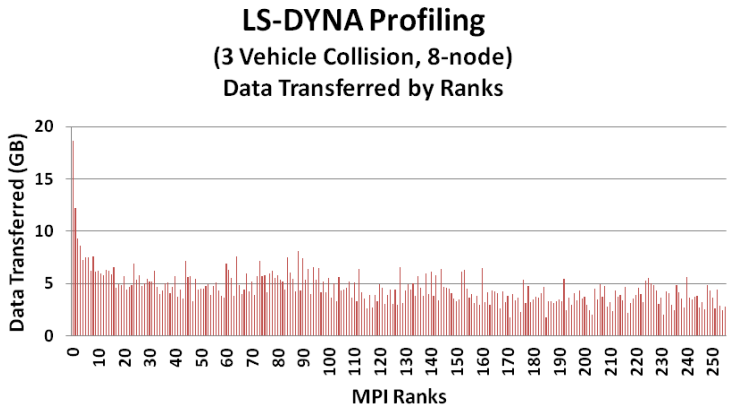
LS-DYNA Profiling – Data Transfer By Process

- **Data transferred to each MPI rank are not evenly distributed**
  - Car2car: The 2 valleys where MPI message communications are more concentrated
- **Amount of data per rank reduces as the cluster scales**
  - While the overall data transfers increase steadily



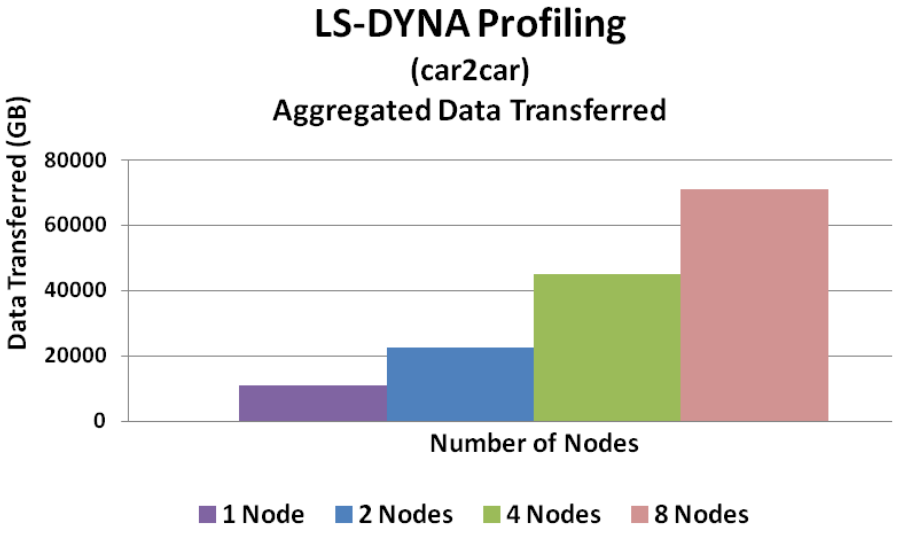


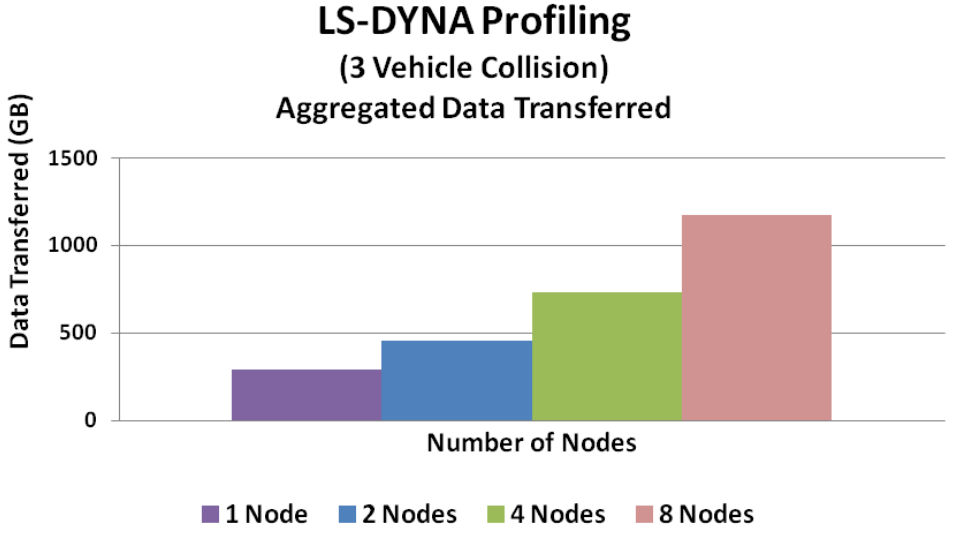




LS-DYNA Profiling – Aggregated Data Transfer

- **Aggregated data transfer refers to:**
  - Total amount of data being transferred in the network between all MPI ranks collectively
- **The total data transfer for car2car is significantly larger than 3cars**
  - Roughly 60 times more data being used for car2car on the 8-node case
  - For both datasets, a sizable amount of data being sent and received across the network





InfiniBand QDR

Summary

- **LS-DYNA shows great needs for CPU computation and network scalability**
  - Best performance can seen when more CPU and nodes are involved
- **CPU:**
  - Running with 1 active core in core pairs allows CPU to run at higher clock frequency
  - Shows around 30% of improvement with 1 active core enabled on 3cars at 8-node
- **Interconnects:**
  - InfiniBand QDR can deliver great network bandwidth needed for scaling to many nodes
  - Using RoCE to offload network processing can offload CPU involvement in processing network data and improve overall job productivity
- **Profiling:**
  - Both data models require for good network bandwidth (~70TB of data on 8 node for car2car)

Thank You

HPC Advisory Council