Mellanox TECHNOLOGIES
HPC ADVISORY COUNCIL

NETWORK OF EXPERTISE

LS-DYNA Performance Benchmarks and Profiling

January 2009

AMD Smarter Choice

DELL

Mellanox TECHNOLOGIES

Note

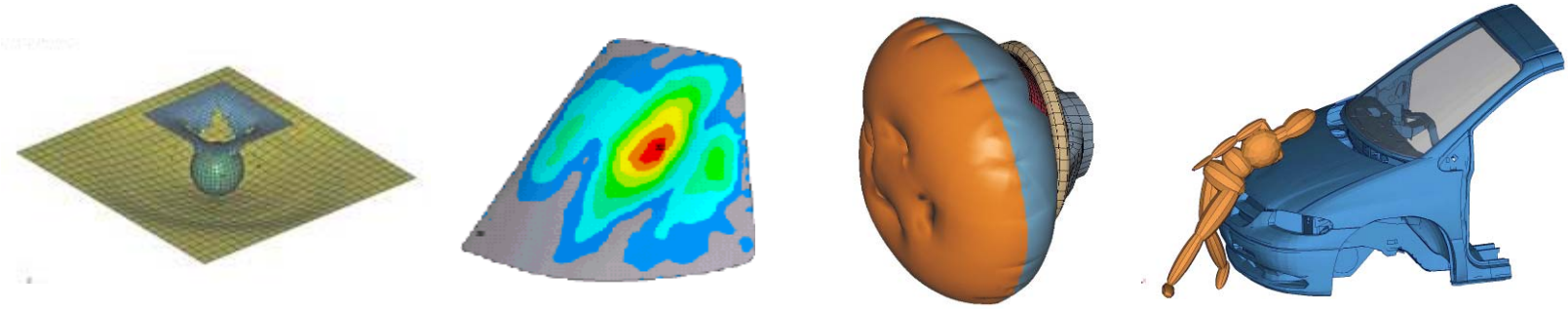
- **The following research was performed under the HPC Advisory Council activities**
  - AMD, Dell, Mellanox
  - HPC Advisory Council Cluster Center
- **The participating members would like to thank LSTC for their support and guidelines**
- **The participating members would like to thank Sharan Kalwani, HPC Automotive specialist, for his support and guidelines**
- **For more info please refer to**
  - www.mellanox.com, www.dell.com/hpc, www.amd.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

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 presented research was done to provide best practices**
  - LS-DYNA performance benchmarking
  - Interconnect performance comparisons
  - Ways to increase LS-DYNA productivity
  - Understanding LS-DYNA communication pattern
  - MPI libraries comparisons
  - Power-aware consideration

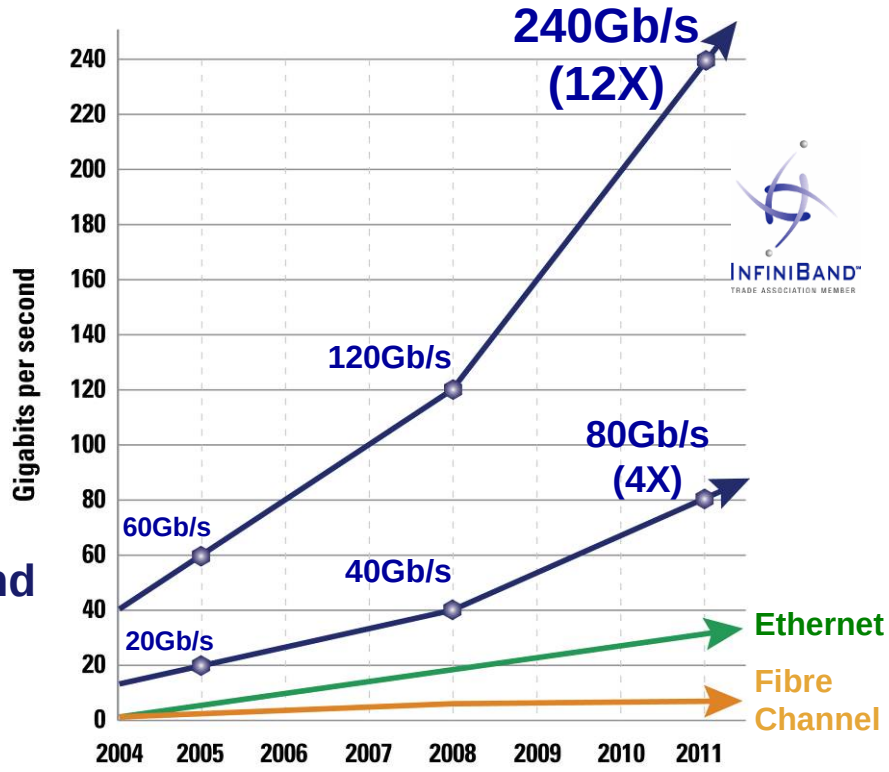
Test Cluster Configuration

- **Dell™ PowerEdge™ SC 1435 24-node cluster**
- **Quad-Core AMD Opteron™ Model 2358 processors (“Barcelona”)**
- **Mellanox® InfiniBand ConnectX® DDR HCAs**
- **Mellanox® InfiniBand DDR Switch**
- **Memory: 16GB memory, DDR2 667MHz per node**
- **OS: RHEL5U2, OFED 1.3 InfiniBand SW stack**
- **MPI: HP MPI 2.2.7, Platform MPI 5.6.5**
- **Application: LS-DYNA MPP971**
- **Benchmark Workload**
  - Three Vehicle Collision Test simulation
  - Neon-Refined Revised Crash Test simulation

Mellanox InfiniBand Solutions

- **Industry Standard**
  - Hardware, software, cabling, management
  - Design for clustering and storage interconnect
- **Performance**
  - 40Gb/s node-to-node
  - 120Gb/s switch-to-switch
  - 1us application latency
  - Most aggressive roadmap in the industry
- **Reliable with congestion management**
- **Efficient**
  - RDMA and Transport Offload
  - Kernel bypass
  - CPU focuses on application processing
- **Scalable for Petascale computing & beyond**
- **End-to-end quality of service**
- **Virtualization acceleration**
- **I/O consolidation Including storage**

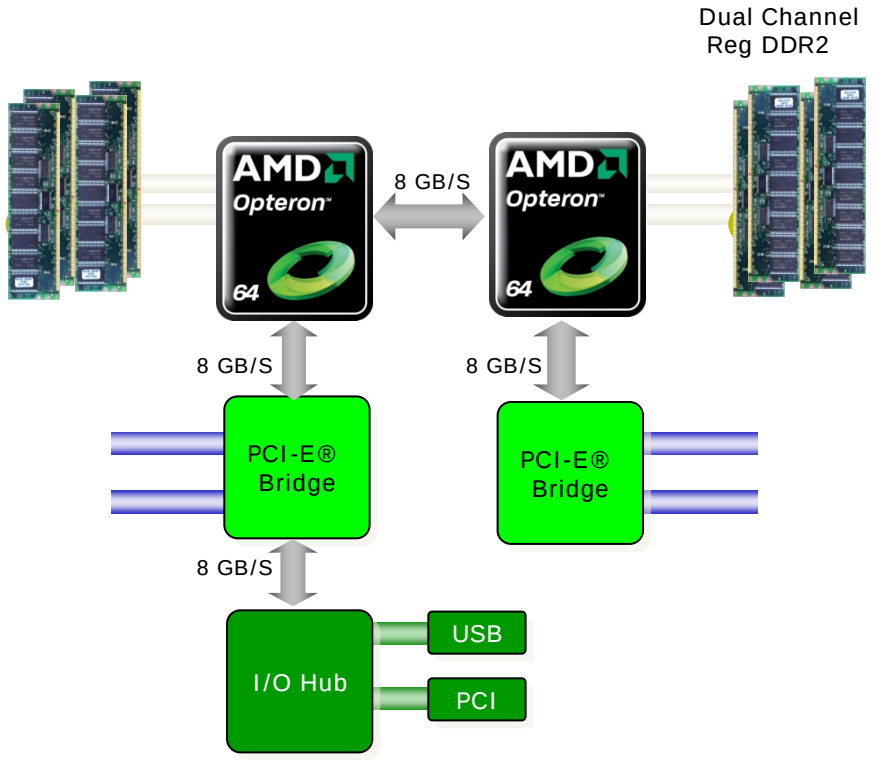
The InfiniBand Performance Gap is Increasing



InfiniBand Delivers the Lowest Latency

Quad-Core AMD Opteron™ Processor

- **Performance**
  - Quad-Core
    - Enhanced CPU IPC
    - 4x 512K L2 cache
    - 2MB L3 Cache
  - Direct Connect Architecture
    - HyperTransport™ technology
    - Up to 24 GB/s
  - Floating Point
    - 128-bit FPU per core
    - 4 FLOPS/clk peak per core
  - Memory
    - 1GB Page Support
    - DDR-2 667 MHz
- **Scalability**
  - 48-bit Physical Addressing
- **Compatibility**
  - Same power/thermal envelopes as Second-Generation AMD Opteron™ processor

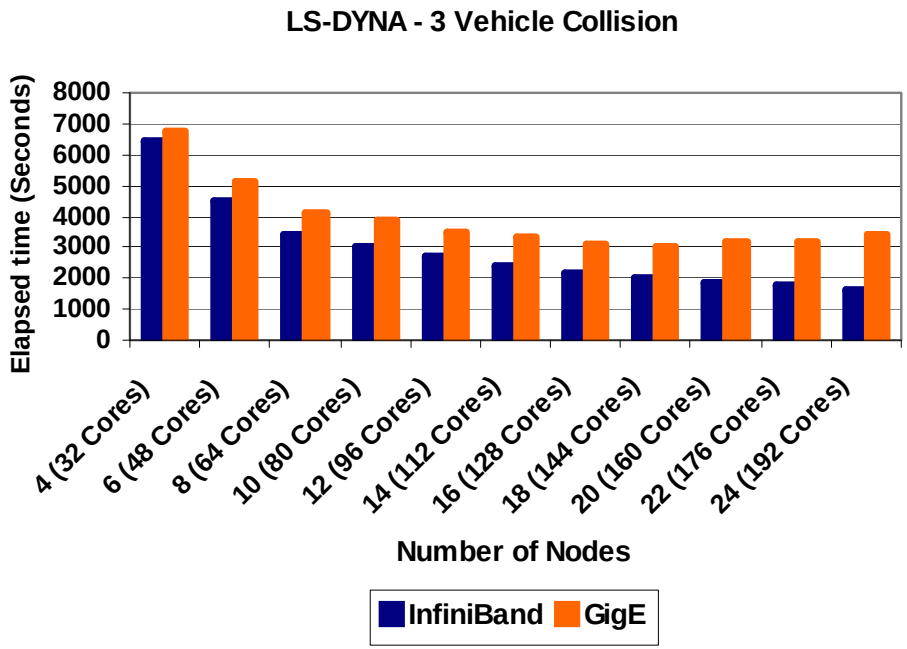


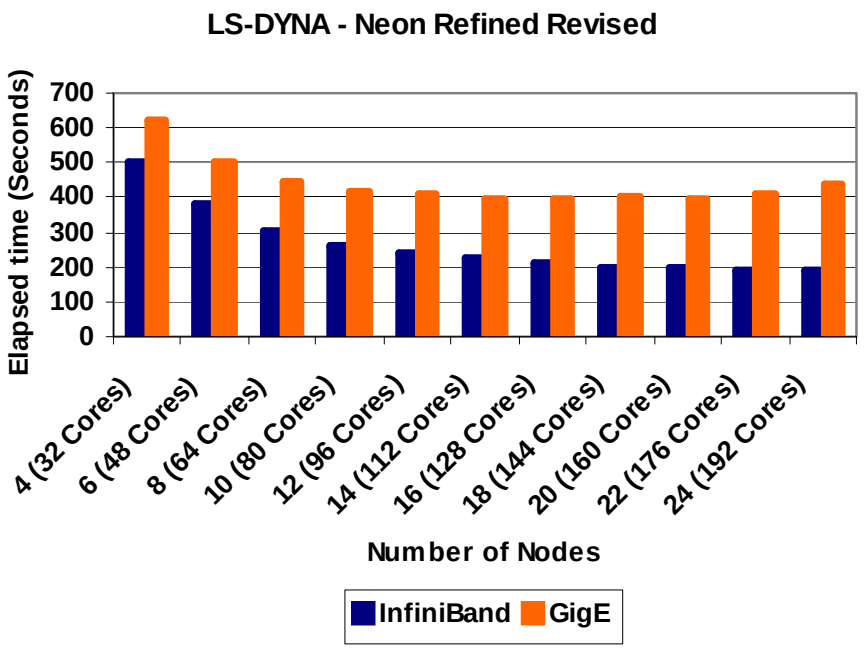
Dell PowerEdge Servers helping Simplify IT

- **System Structure and Sizing Guidelines**
  - 24-node cluster build with Dell PowerEdge™ SC 1435 Servers
  - Servers optimized for High Performance Computing environments
  - Building Block Foundations for best price/performance and performance/watt
- **Dell HPC Solutions**
  - Scalable Architectures for High Performance and Productivity
  - Dell's comprehensive HPC services help manage the lifecycle requirements.
  - Integrated, Tested and Validated Architectures
- **Workload Modeling**
  - Optimized System Size, Configuration and Workloads
  - Test-bed Benchmarks
  - ISV Applications Characterization
  - Best Practices & Usage Analysis

LS-DYNA Performance Results - Interconnect

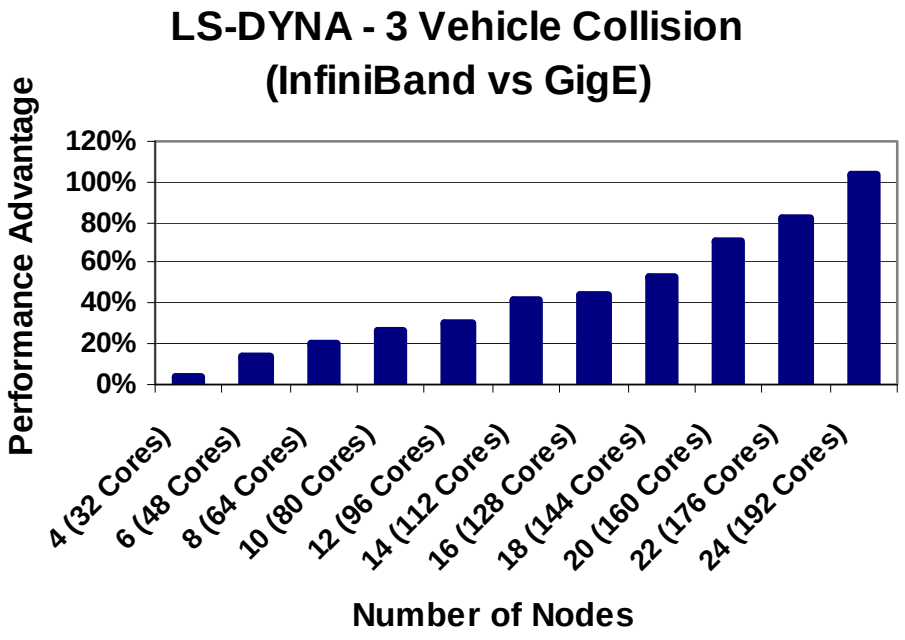
- **InfiniBand high speed interconnect enables highest scalability**
  - Performance gain with cluster size
- **Performance over GigE is not scaling**
  - Slowdown occurs as number of processors increases beyond 16 nodes

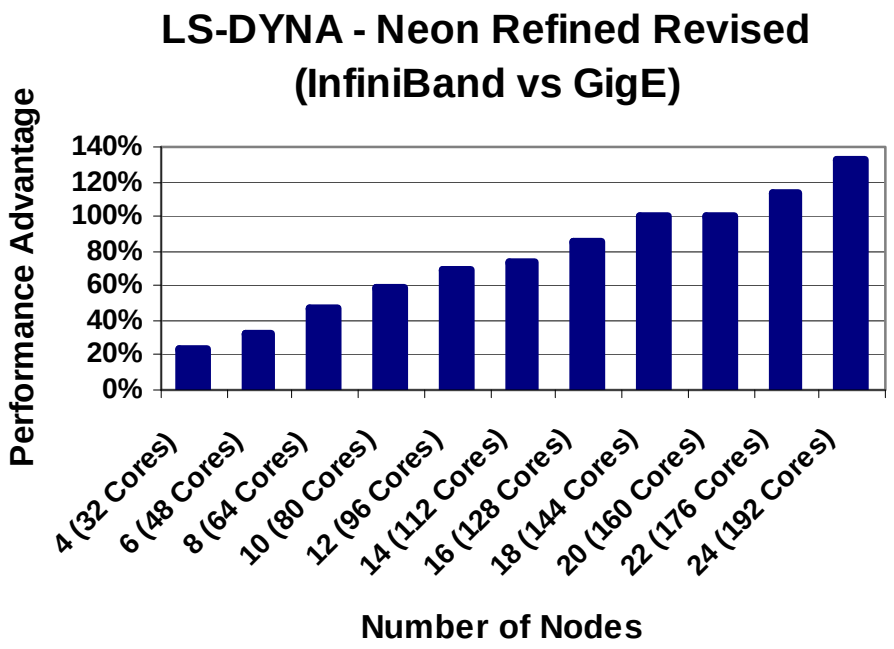




Lower is better

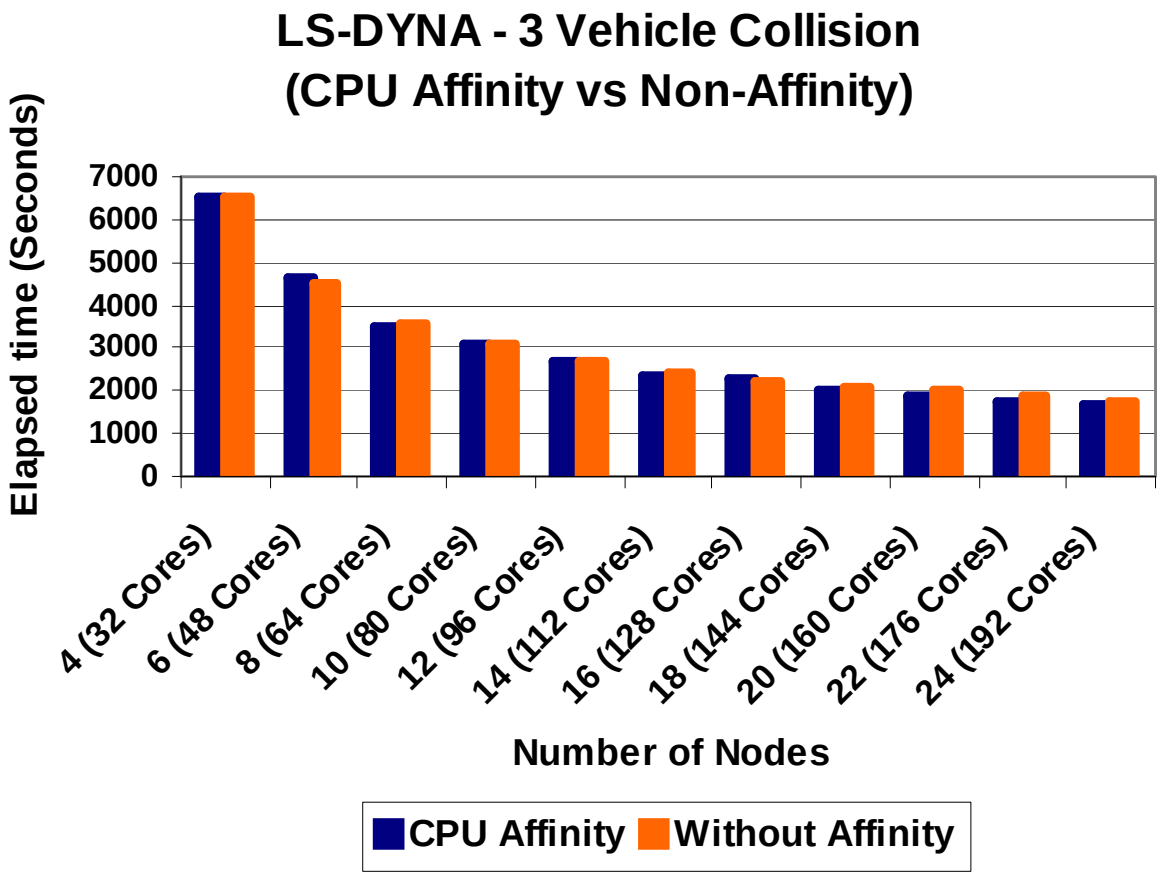
LS-DYNA Performance Results - Interconnect





- **InfiniBand outperforms GigE by up to 132%**
  - As node number increases, bigger advantage is expected

LS-DYNA Performance Results – CPU Affinity

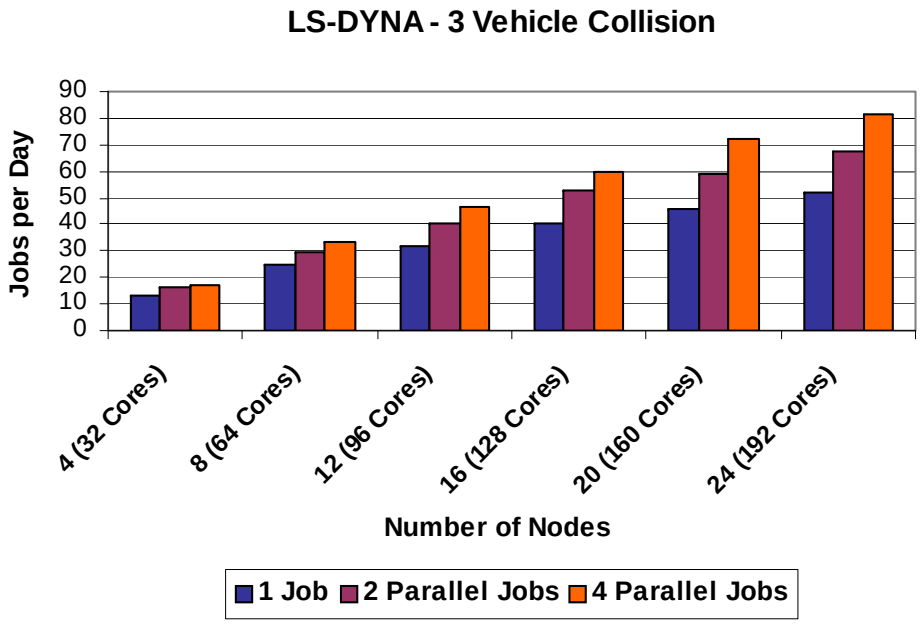


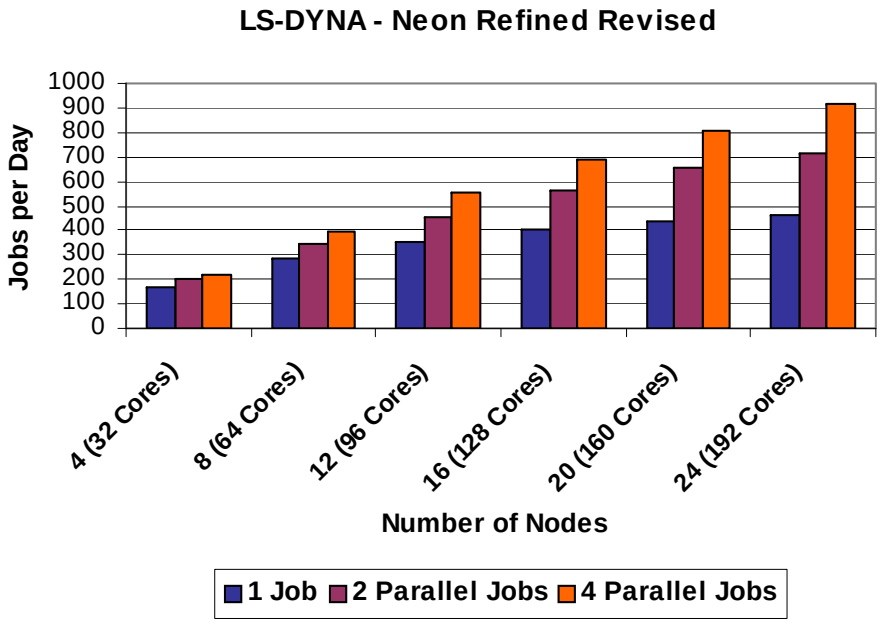
- **CPU affinity accelerates performance up to 10%**
- **Saves up to 177 seconds per simulation**

Lower is better

LS-DYNA Performance Results - Productivity

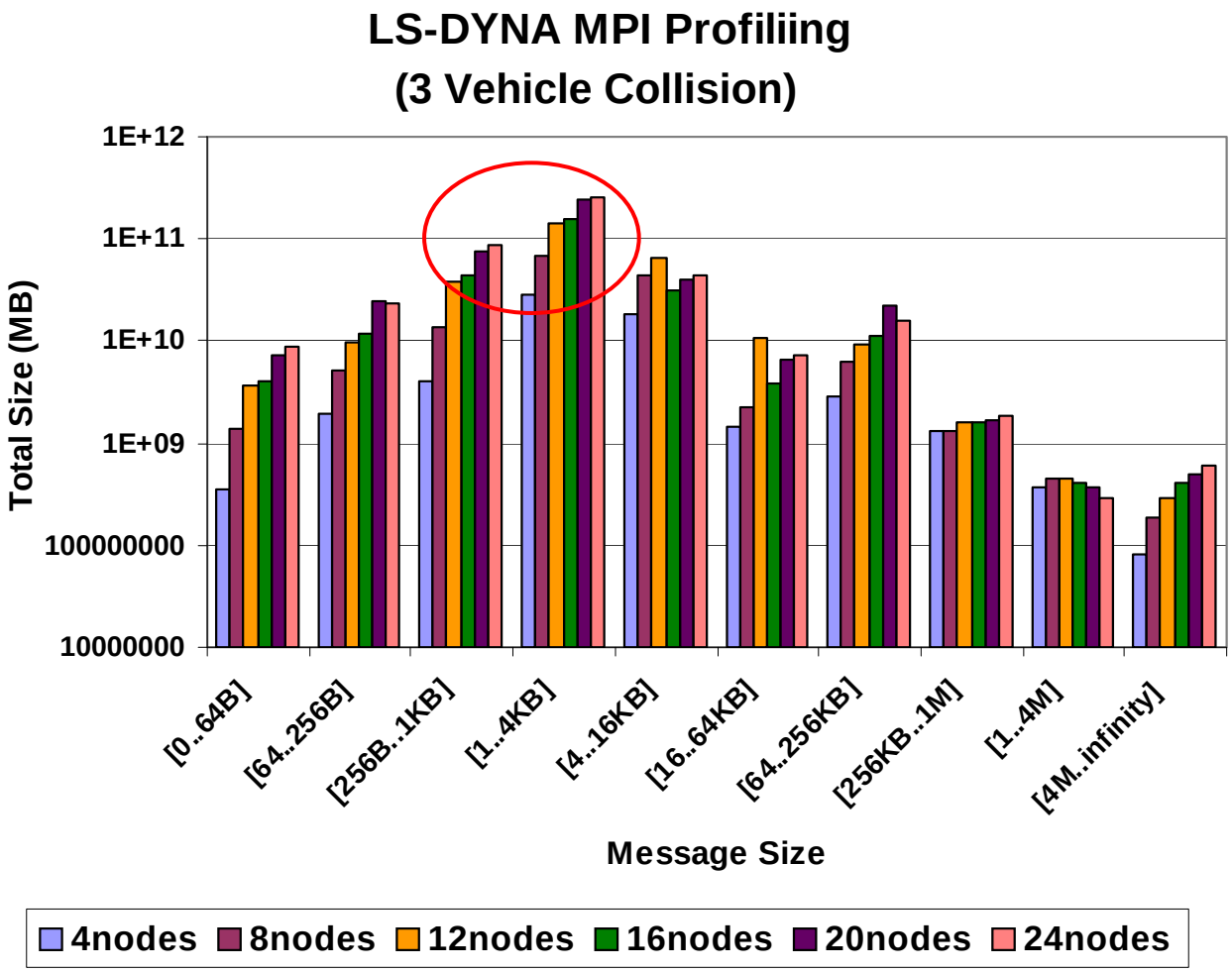
- **InfiniBand increases productivity by allowing multiple jobs to run simultaneously**
  - Providing required productivity for virtual vehicle design
- **Three cases are presented**
  - Single job over the entire systems (with CPU affinity)
  - Two jobs, each on a single CPU per server (job placement , CPU affinity)
  - Four jobs, each on two CPU cores per CPU per server (job placement , CPU affinity)
- **Four jobs per day increases productivity by 97% for Neon Refined Revised, 57% for 3 Car collision case**
- **Increased number of parallel processes (jobs) increases the load on the interconnect**
  - High speed and low latency interconnect solution is required for gaining high productivity





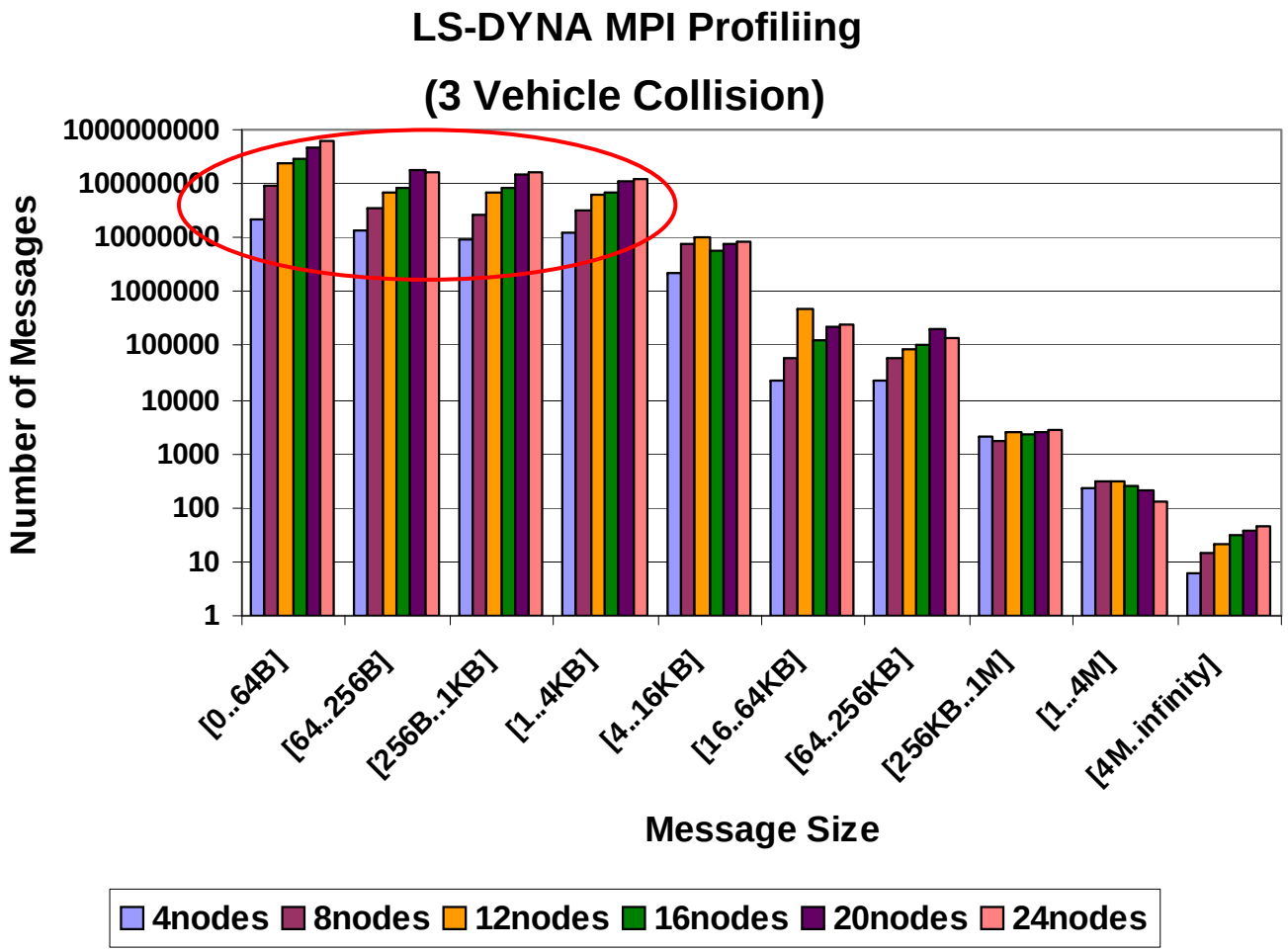
Higher is better

LS-DYNA Profiling – Data Transferred



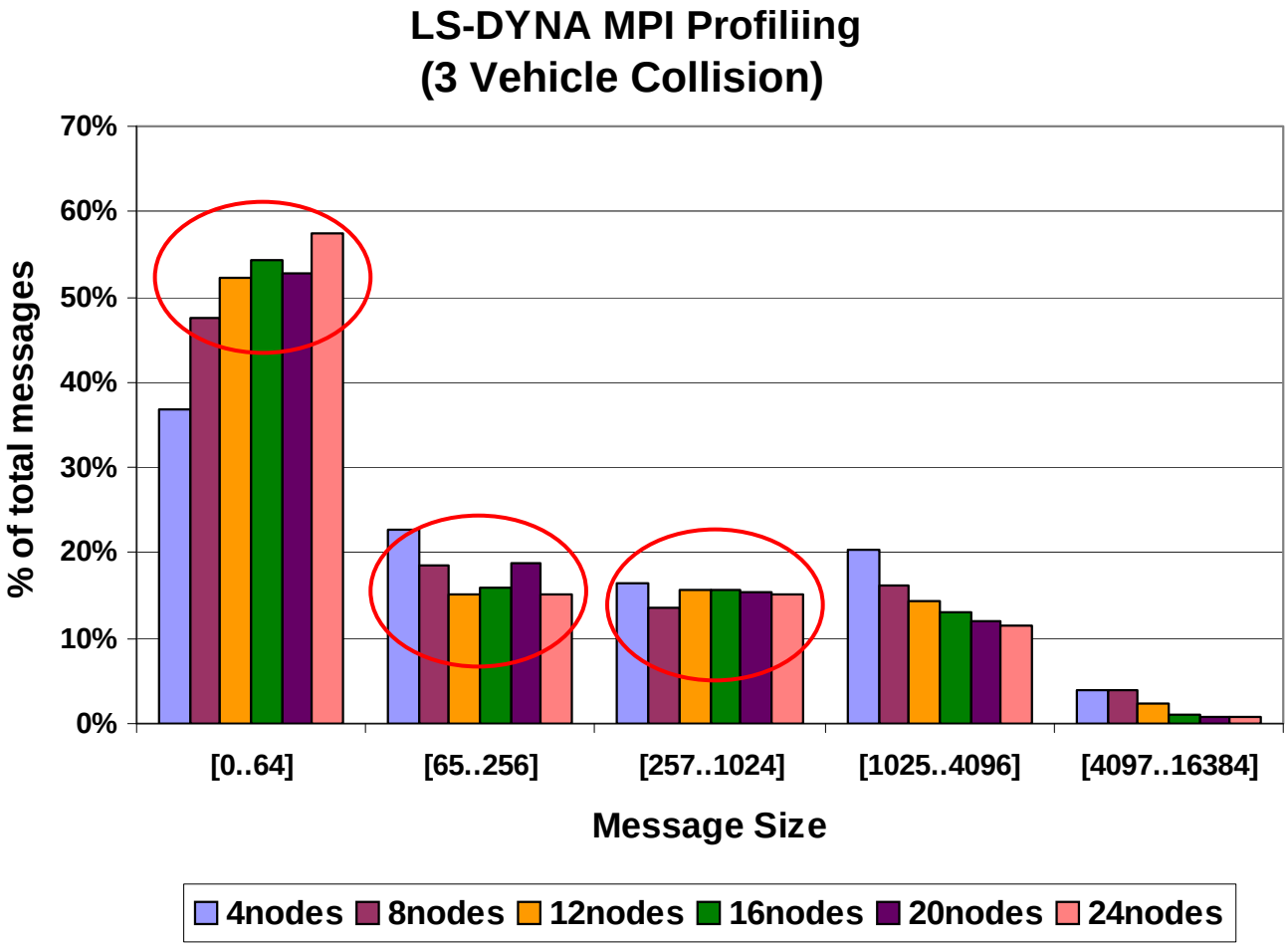
- **Majority of data transfer is done via 256B-4KB message size**

LS-DYNA Profiling – Message Distribution



- **Majority of the messages are in the range of 2B-4KB**
  - 2B-256B for synchronization, 256B-4KB for data communications

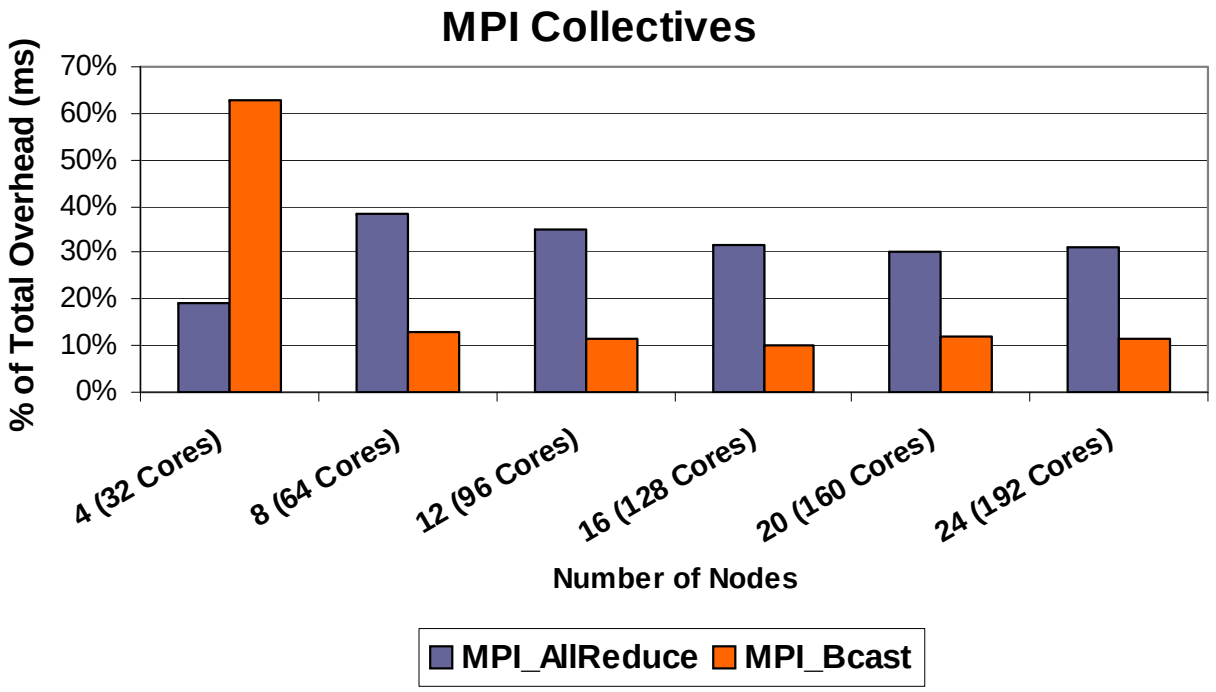
LS-DYNA Profiling – Message Distribution



- **As number of nodes scales, percentage of small messages increases**
- **percentage of 256-1KB messages is relatively consistent with cluster size**
  - Actual number increases with cluster size,

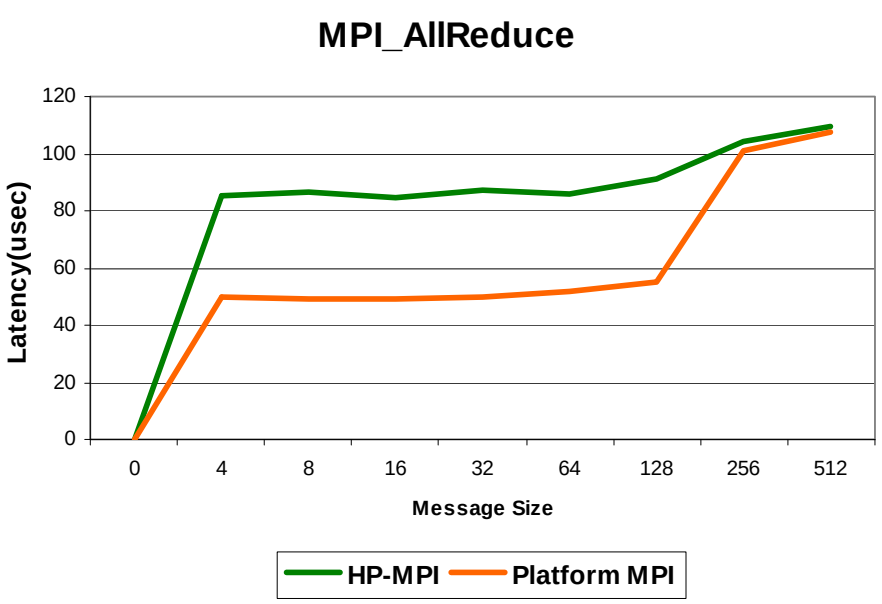
LS-DYNA Profiling – MPI Collectives

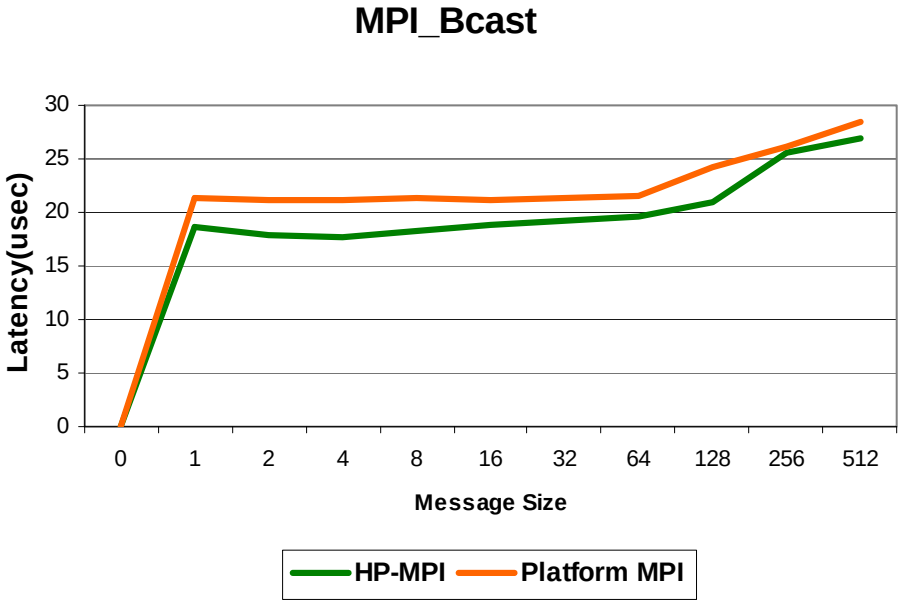
- **Two key MPI collective functions in LS-DYNA**
  - MPI_AllReduce
  - MPI_Bcast
- **Account for the majority of MPI communication overhead**



MPI Collective Benchmarking

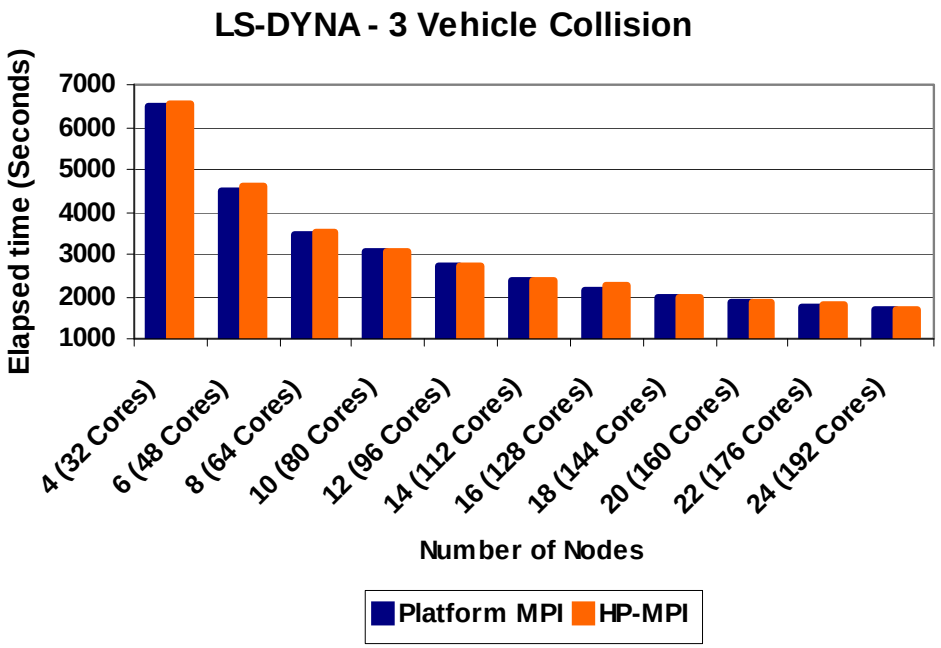
- **MPI collective performance comparison**
  - Two frequently called collection operations in LS-DYNA were benchmarked
    - MPI_Allreduce
    - MPI_Bcast
  - Platform MPI shows better latency for AllReduce operation

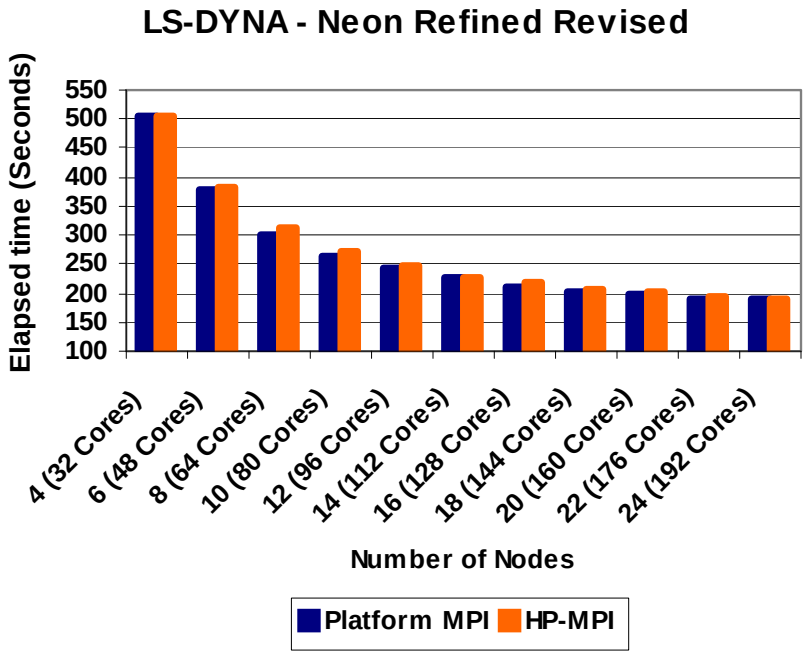




LS-DYNA with Different MPI Libraries

- **LS-DYNA performance Comparison**
  - Each MPI library shows different benefits for latency and collectives
  - As such, HP-MPI and Platform MPI shows comparable performance





Lower is better

LS-DYNA Profiling Summary - Interconnect

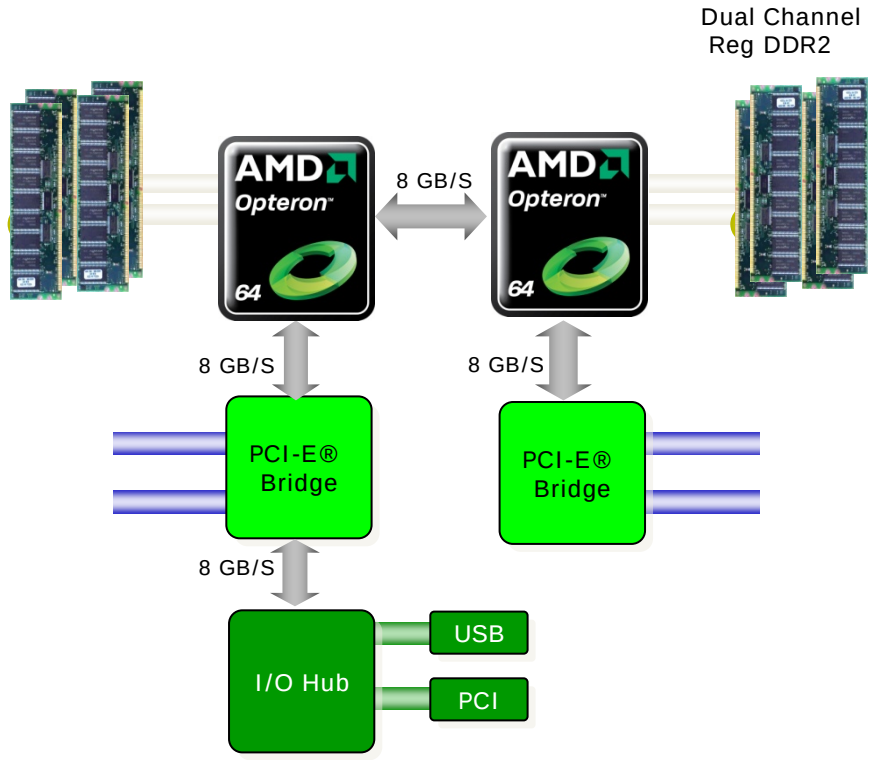
- **LS-DYNA was profiled to determine networking dependency**
- **Majority of data transferred between compute nodes**
  - Done with 256B-4KB message size, data transferred increases with cluster size
- **Most used message sizes**
  - <64B messages – mainly synchronizations
  - 64B-4KB – mainly compute related
- **Message size distribution**
  - Percentage of smaller messages (<64B) increases with cluster size
    - Mainly due to the needed synchronization
  - Percentage of mid-size messages (64B-4KB) is kept the same with cluster size
    - Compute transactions increases with cluster size
  - Percentage of very large messages decreases with cluster size
    - Mainly used for problem data distribution at the simulation initialization phase
- **LS-DYNA interconnect sensitivity points**
  - Interconnect latency and throughput for 64B-4KB message range
  - Collectives operations performance, mainly MPI_Allreduce

Test Cluster Configuration – System Upgrade

- **The following results were achieved after system upgrade (changes are in green)**
  - Dell PowerEdge SC 1435 24-node cluster
  - Quad-Core AMD Opteron™ Model **2382** processors (“Shanghai”) (vs “Barcelona” in previous configuration)
  - Mellanox® InfiniBand ConnectX® DDR HCAs
  - Mellanox® InfiniBand DDR Switch
  - Memory: 16GB memory, DDR2 **800MHz** per node (vs 667MHz in previous configuration)
  - OS: RHEL5U2, OFED 1.3 InfiniBand SW stack
  - MPI: HP MPI 2.2.7, Platform MPI 5.6.5
  - Application: LS-DYNA MPP971
  - Benchmark Workload
    - Three-Car Crash Test simulation
    - Neon-Refined Revised Crash Test simulation

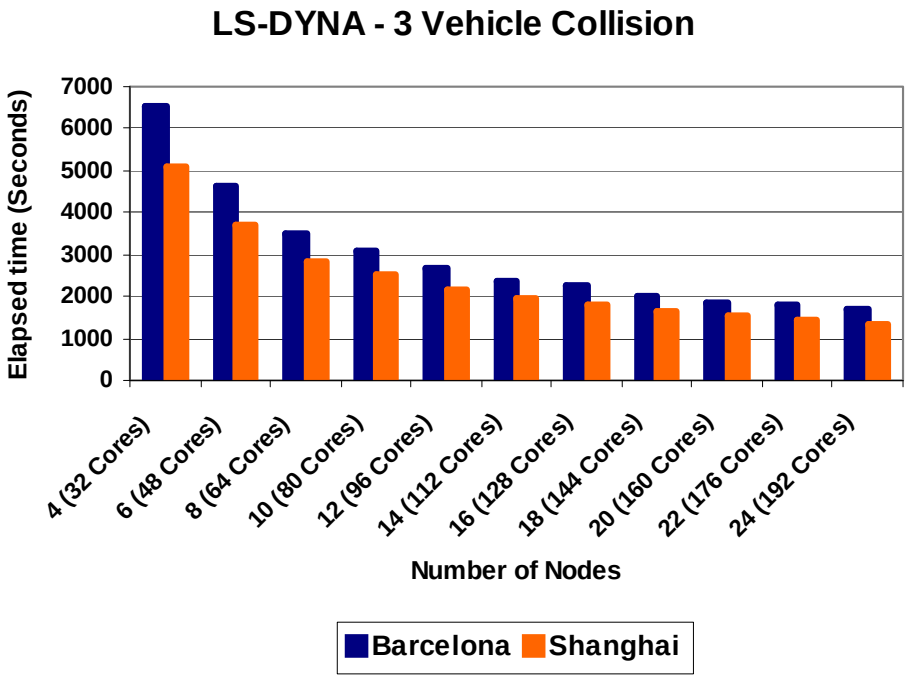
Quad-Core AMD Opteron™ Processor

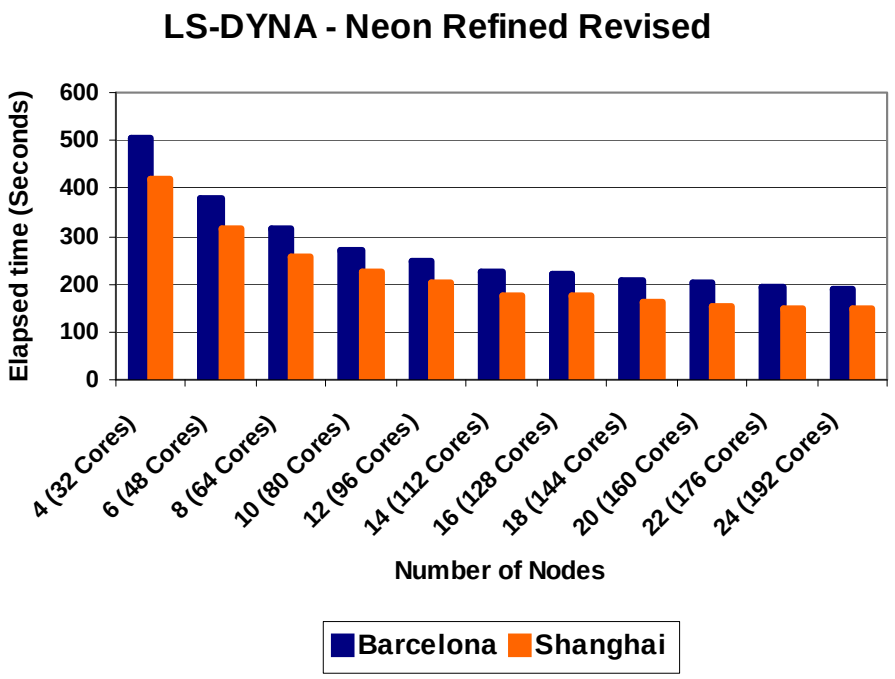
- **Performance**
  - Quad-Core
    - Enhanced CPU IPC
    - 4x 512K L2 cache
    - 6MB L3 Cache
  - Direct Connect Architecture
    - HyperTransport™ technology
    - Up to 24 GB/s peak per processor
  - Floating Point
    - 128-bit FPU per core
    - 4 FLOPS/clk peak per core
  - Integrated Memory Controller
    - Up to 12.8 GB/s
    - DDR2-800 MHz or DDR2-667 MHz
- **Scalability**
  - 48-bit Physical Addressing
- **Compatibility**
  - Same power/thermal envelopes as 2nd / 3rd generation AMD Opteron™ processor



Performance Improvement

- **Upgraded AMD CPU and DDR-2 Memory**
- **LS-DYNA run time decreased by more than 20%**
  - Leveraging InfiniBand 20Gb/s for higher scalability

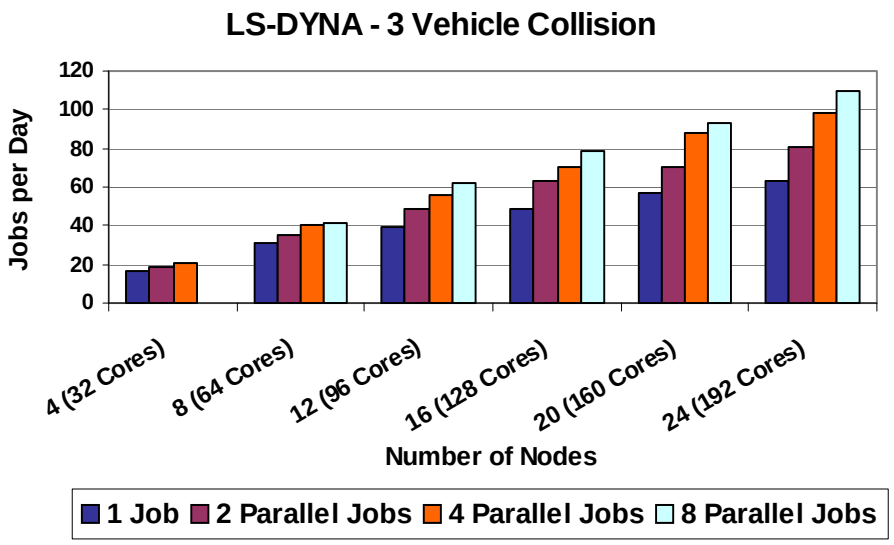


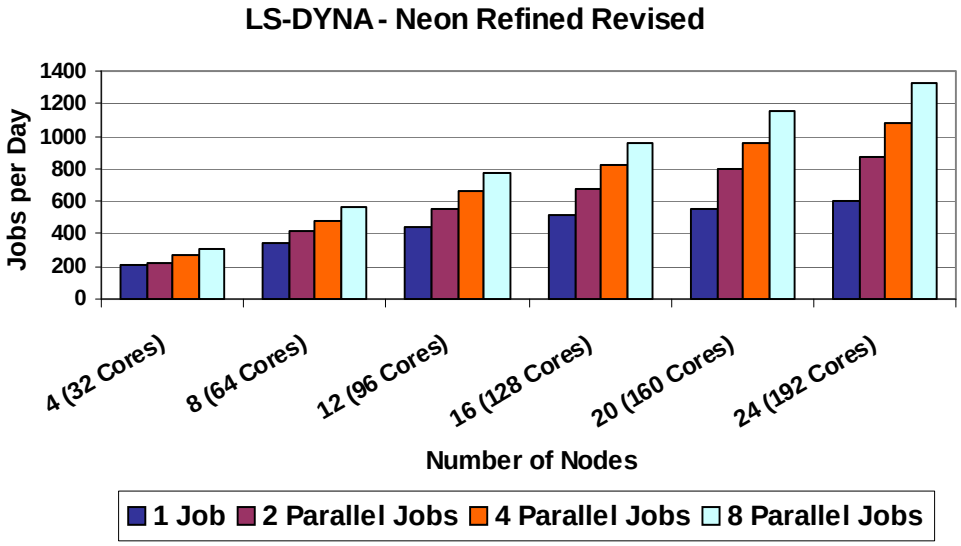


Lower is better

Maximize LS-DYNA Productivity

- **Scalable latency of InfiniBand and latest Shanghai processor deliver scalable LS-DYNA performance**

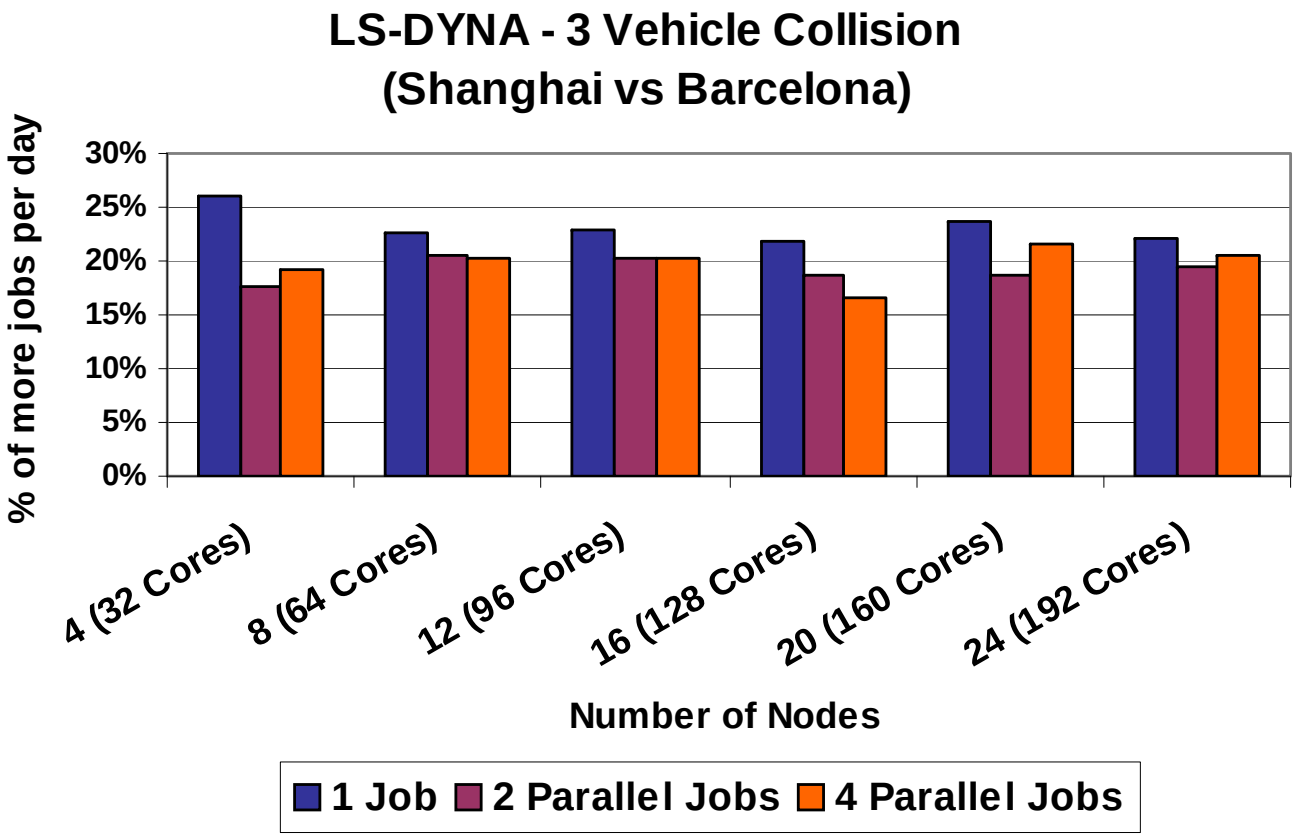




Higher is better

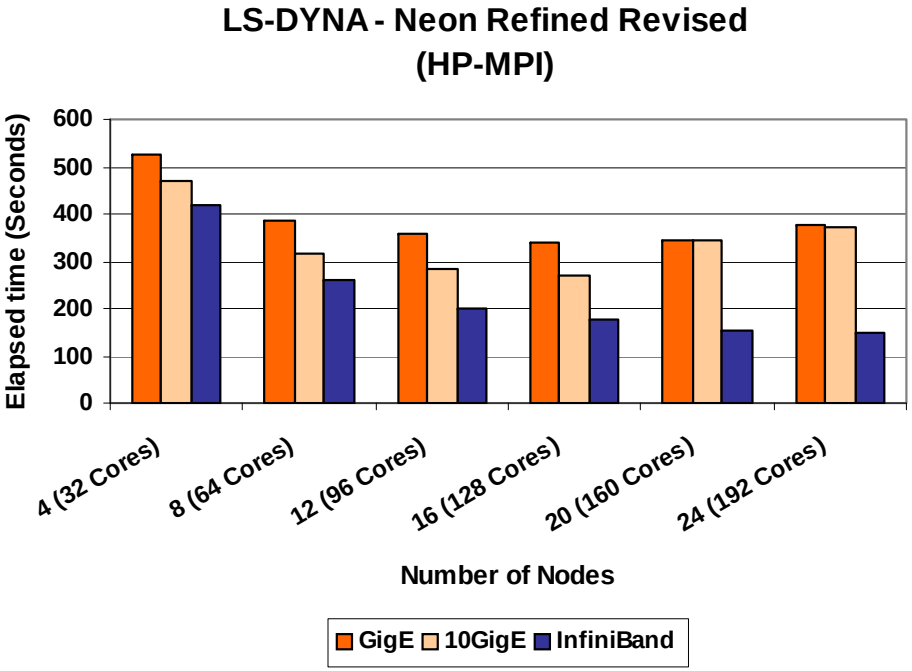
LS-DYNA with Shanghai Processors

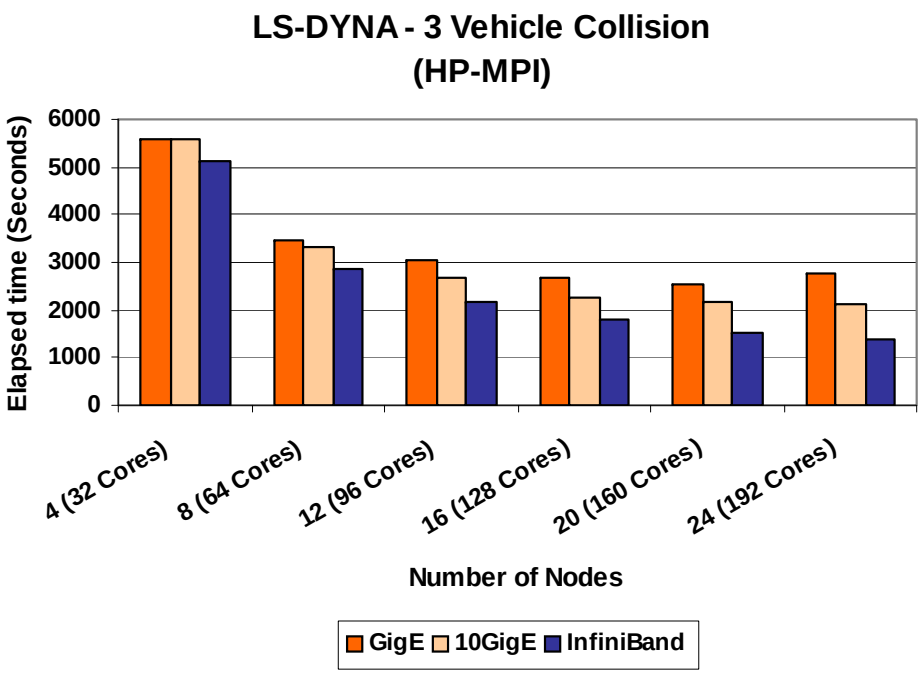
- **"Shanghai" processors provides higher performance compared to "Barcelona'**



LS-DYNA Performance Results - Interconnect

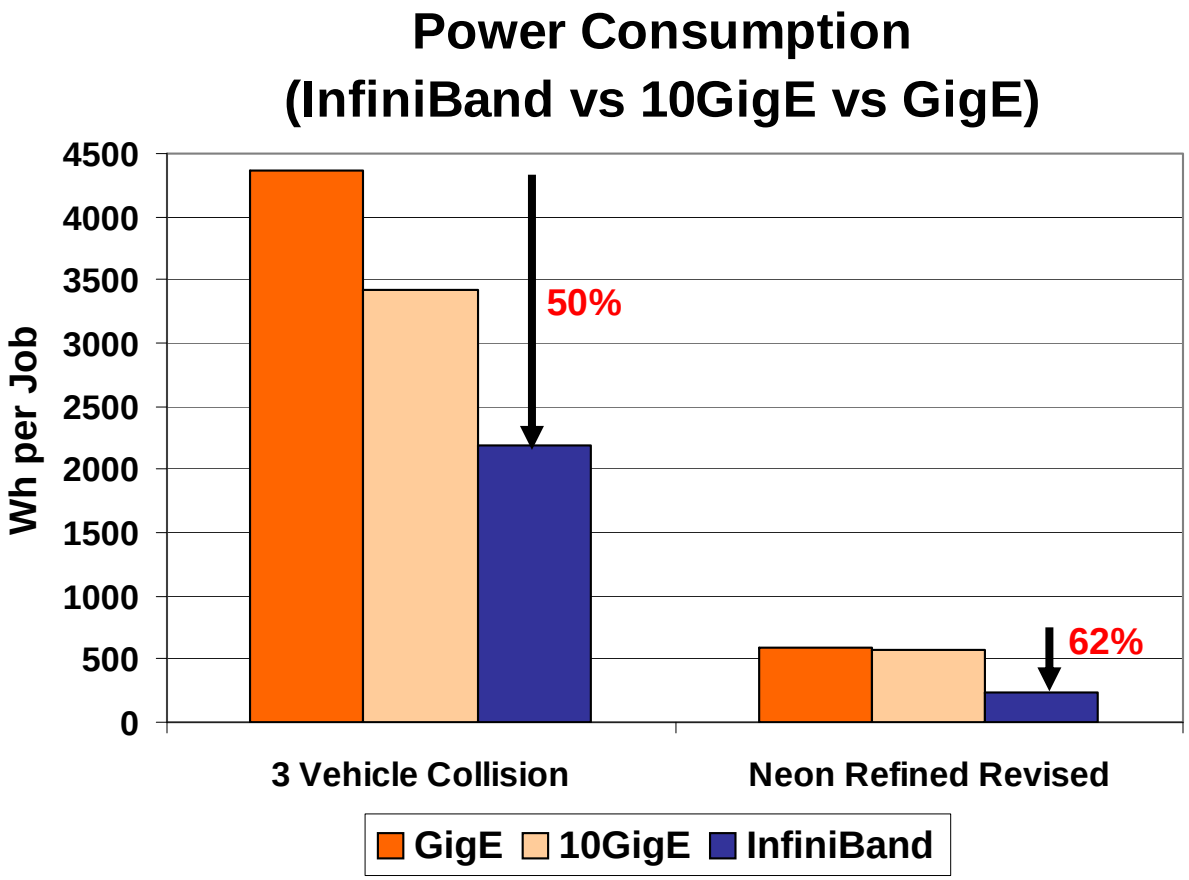
- **InfiniBand 20Gb/s vs 10GigE vs GigE**
- **InfiniBand 20Gb/s (DDR) outperforms 10GigE and GigE in all test cases**
  - Reducing run time by up to 60% versus 10GigE and 61% vs GigE
- **Performance loss shown beyond 16 nodes with 10GigE and GigE**
- **InfiniBand 20Gb/s maintain scalability with cluster size**





Lower is better

Power Consumption Comparison



- **InfiniBand also enables power efficient simulations**
  - Reducing power/job by up to 62%!

24-node comparison

Conclusions

- **LS-DYNA is widely used to simulate many real-world problems**
  - Automotive crash-testing and finite-element simulations
  - Developed by Livermore Software Technology Corporation (LSTC)
- **LS-DYNA performance and productivity relies on**
  - Scalable HPC systems and interconnect solutions
  - Low latency and high throughput interconnect technology
  - NUMA aware application for fast access to local memory
  - Reasonable job distribution can dramatically improve productivity
    - Increasing number of jobs per day while maintaining fast run time
- **Interconnect comparison shows**
  - InfiniBand delivers superior performance and productivity in every cluster size
  - Scalability requires low latency and “zero” scalable latency
  - Lowest power consumption was achieved with InfiniBand
    - Saving in system power, cooling and real-estate

Thank You

HPC Advisory Council

HPC@mellanox.com