HPC ADVISORY COUNCIL

LS-DYNA Performance Benchmark and Profiling

March 2010

intel DELL LSTC Livermore Software Technology Corp. Mellanox TECHNOLOGIES

Note

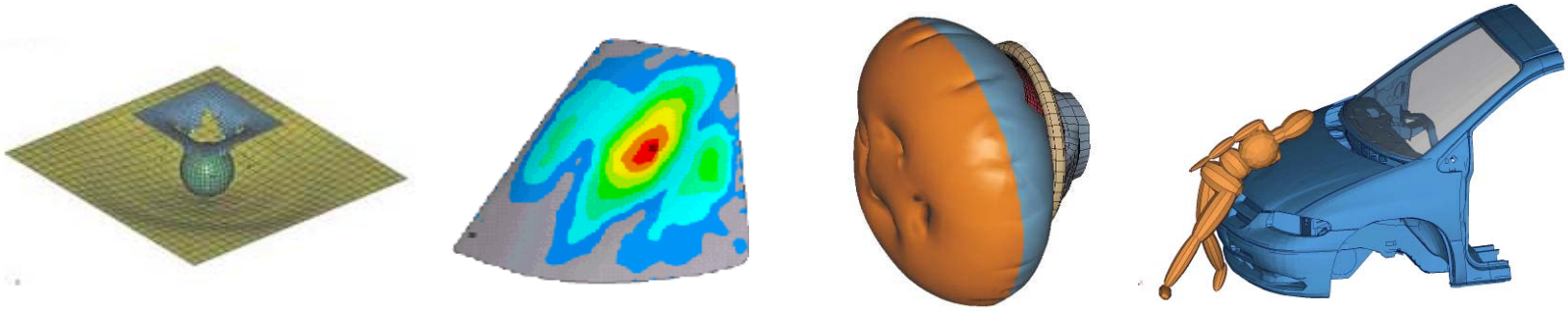
- **The following research was performed under the HPC Advisory Council activities**
  - Participating vendors: Intel, Dell, Mellanox, LSTC
  - Compute resource - HPC Advisory Council Cluster Center
- **For more info please refer to**
  - www.mellanox.com, www.dell.com/hpc, www.intel.com, www.lstc.com

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

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
    - MPI Library performance comparisons
    - Interconnect performance comparisons
  - Understanding LS-DYNA communication patterns
  - Power-efficient simulations
- **The presented results will demonstrate**
  - The scalability of the compute environment to provide good application scalability
  - Considerations for power saving through balanced system configuration

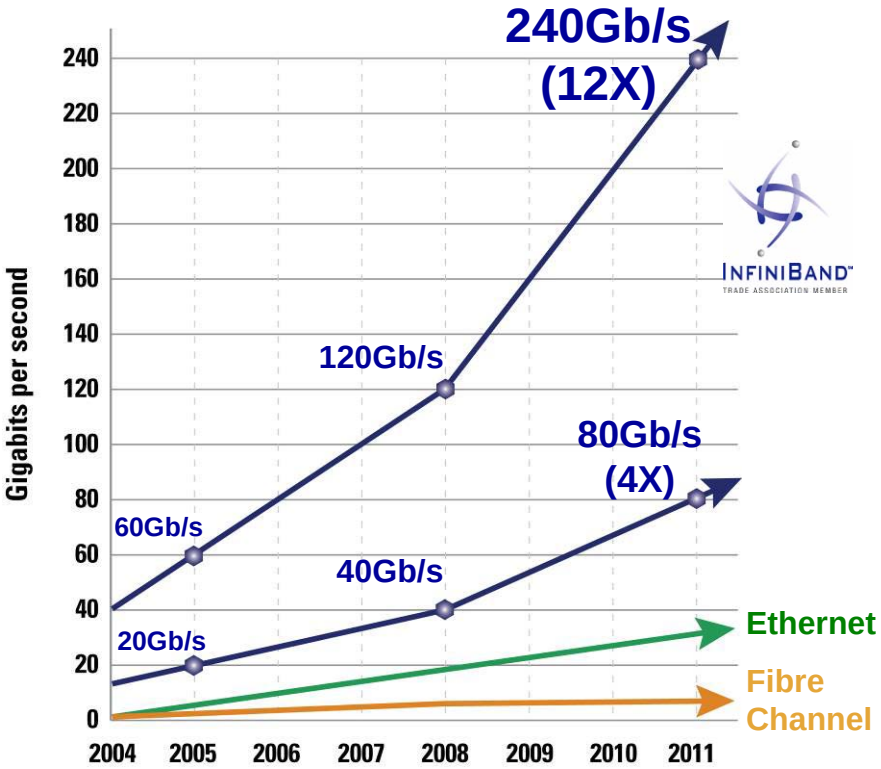
Test Cluster Configuration

- **Dell™ PowerEdge™ M610 16-node cluster**
- **Quad-Core Intel X5570 @ 2.93 GHz CPUs**
- **Intel Cluster Ready certified cluster**
- **Mellanox ConnectX2 QDR InfiniBand mezzanine card**
- **Mellanox M3601Q 32-Port Quad Data Rate (QDR-40Gb) InfiniBand Switch**
- **Memory: 24GB memory per node**
- **OS: RHEL5U3, OFED 1.5 InfiniBand SW stack**
- **File system: Lustre 1.8.2**
- **MPI: Open MPI 1.3.3, HP-MPI 2.7.1, Platform MPI 5.6.7, Intel MPI 4.0**
- **Application: LS-DYNA MPP971_s_R4.2.1**
- **Benchmark Workload**
  - Three Vehicle Collision 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**

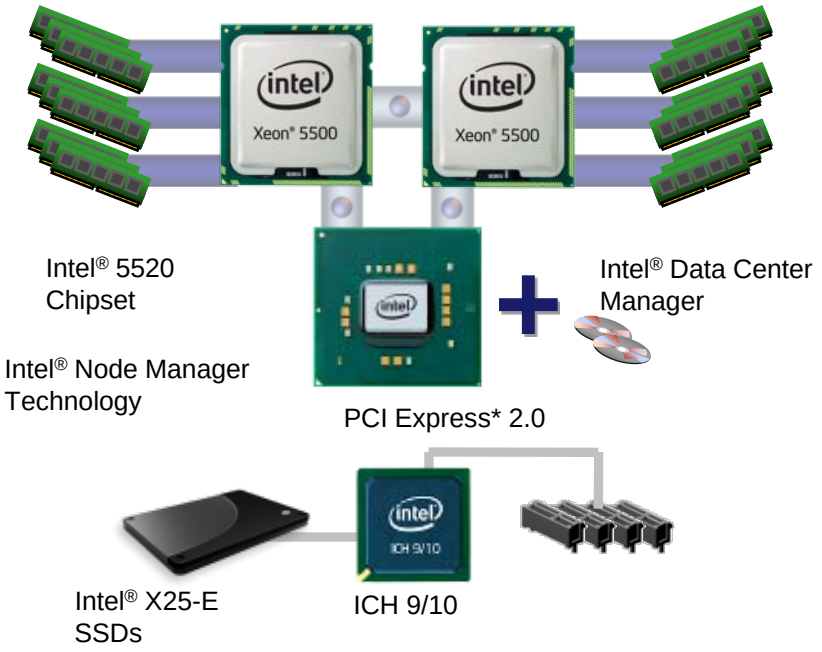
The InfiniBand Performance Gap is Increasing



InfiniBand Delivers the Lowest Latency

Delivering Intelligent Performance

Next Generation Intel® Microarchitecture



Bandwidth Intensive

- Intel® QuickPath Technology
- Integrated Memory Controller

Threaded Applications

- 45nm quad-core Intel® Xeon® Processors
- Intel® Hyper-threading Technology

Performance on Demand

- Intel® Turbo Boost Technology
- Intel® Intelligent Power Technology

Performance That Adapts to The Software Environment

Intel® Cluster Ready

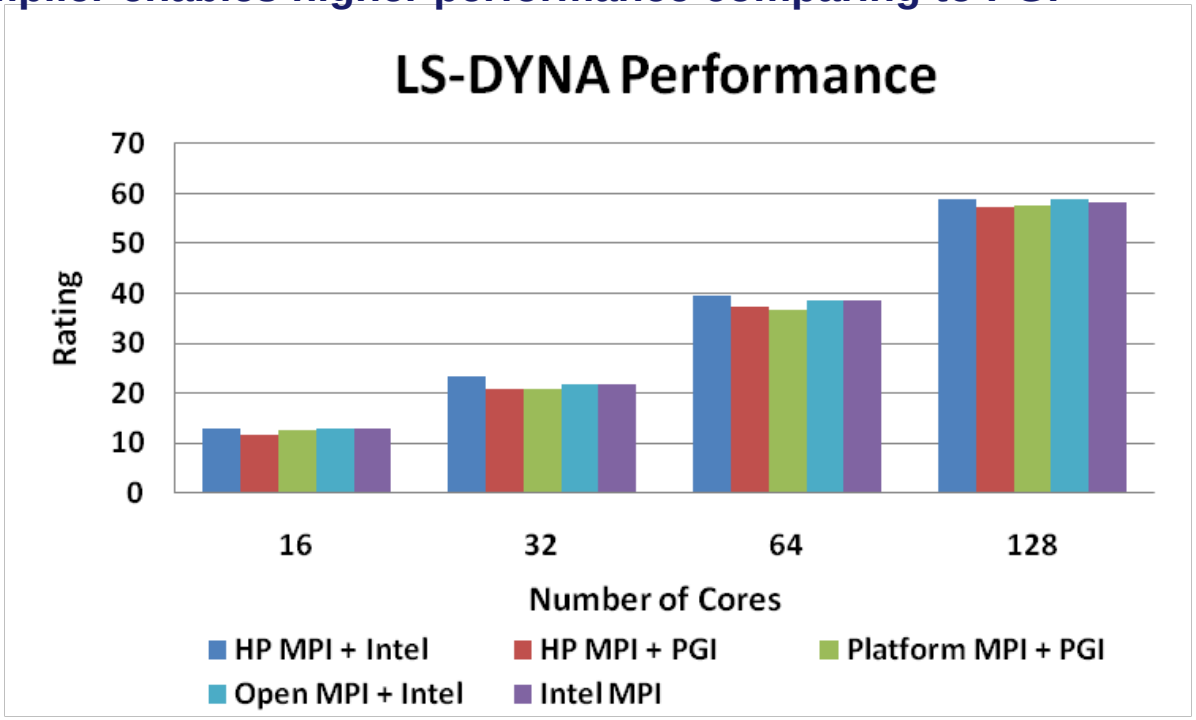
- **Intel® Cluster Ready is a consistent reference Linux platform architecture for Intel-based systems**
  - Makes it easier to design, develop, and build applications for clusters
- **A single architecture platform supported and used by a wide range of OEMs, ISVs, cluster provisioning vendors, and interconnect providers**
- **Includes**
  - Platform specification, that defines the Intel Cluster Ready platforms
  - Program branding, that makes it easier to identify compliant solutions and applications
  - Hardware certifications, confirming solutions that are delivered ready to run
  - Application registration, validating applications that execute on top of Intel Cluster Ready architecture
  - Intel® Cluster Checker tool, to validate hardware and software configuration and functionality

Dell PowerEdge Servers helping Simplify IT

- **System Structure and Sizing Guidelines**
  - 16-node cluster build with Dell PowerEdge™ M610 blades server
  - 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 Benchmark Results - Libraries

- **Input Dataset**
  - Three Vehicle Collision Test
- **All MPIs provide same level of performance**
  - HP-MPI slightly better than others
- **Intel compiler enables higher performance comparing to PGI**

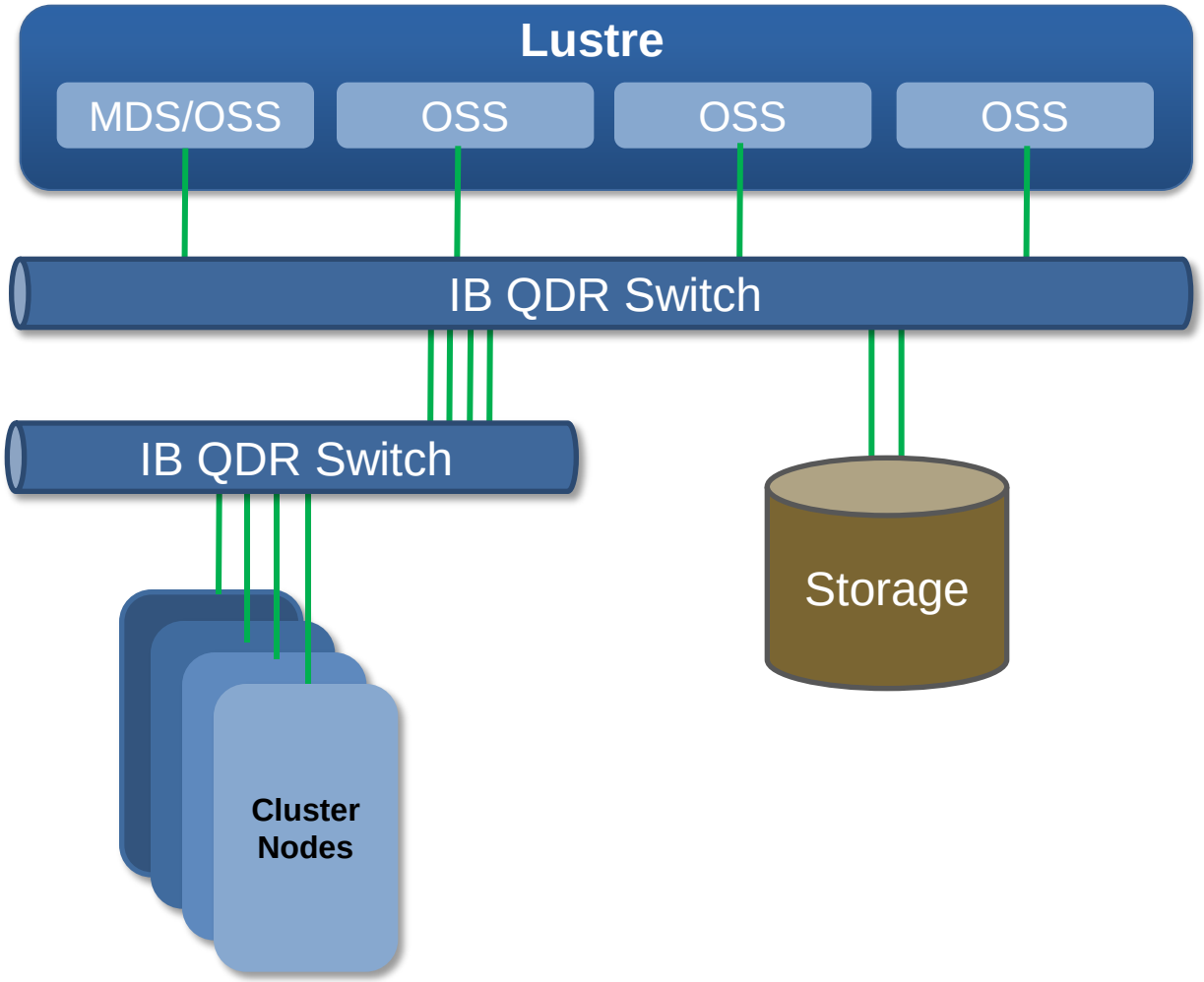


Higher is better

8-cores per node

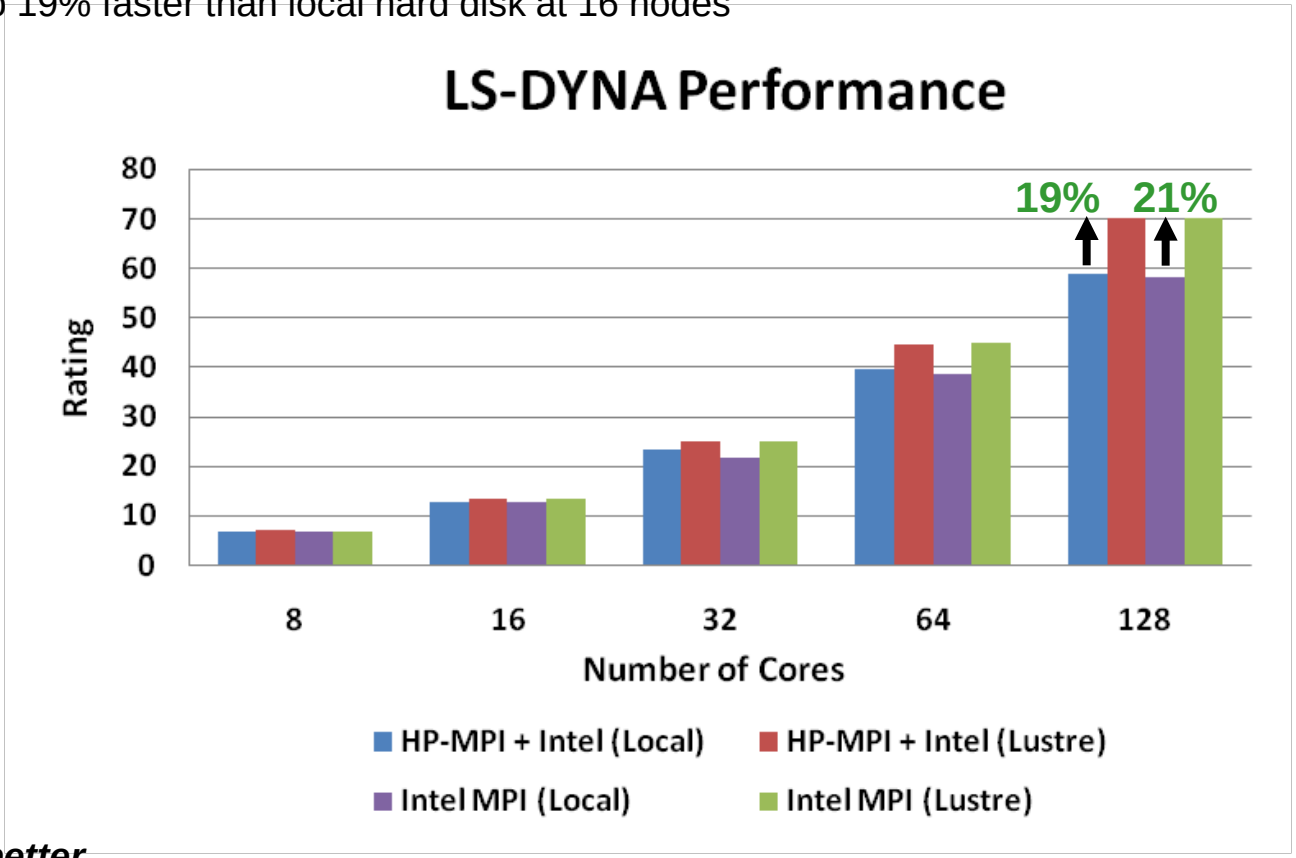
Lustre File System Configuration

- **Lustre Configuration**
  - 1 MDS
  - 4 OSS (Each has 2 OST)
  - InfiniBand based Backend storage
  - All components are connected through InfiniBand QDR interconnect



LS-DYNA Benchmark Results - File System

- **Intel MPI has native Lustre support**
  - mpiexec -genv I_MPI_ADJUST_BCAST 5 -genv I_MPI_EXTRA_FILESYSTEM on -genv I_MPI_EXTRA_FILESYSTEM_LIST lustre
- **Lustre enables higher performance**
  - Up to 19% faster than local hard disk at 16 nodes

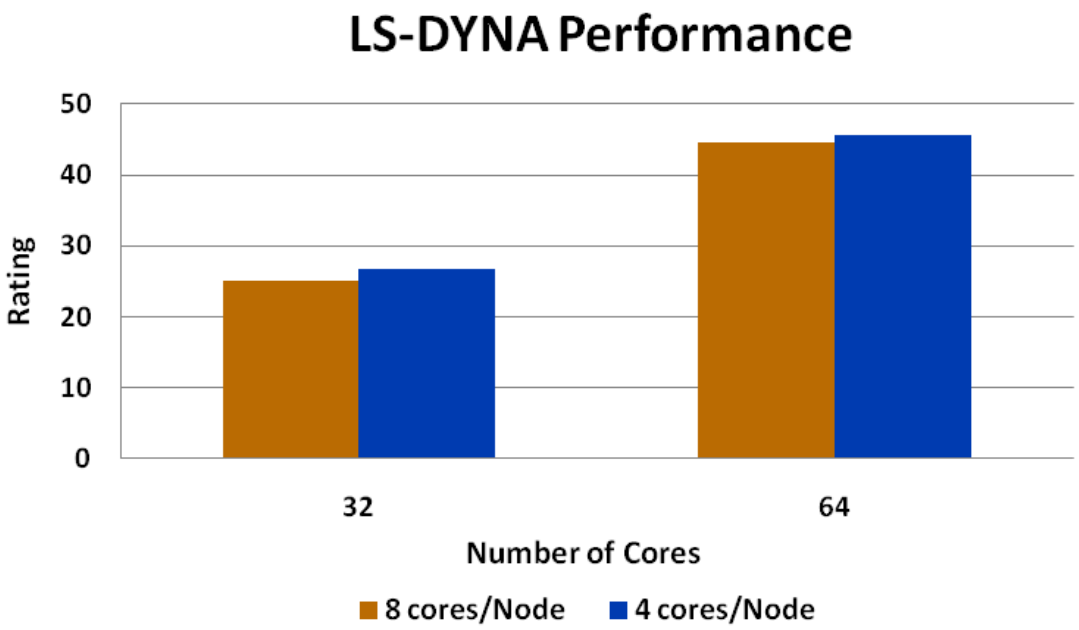


Higher is better

8-cores per node

LS-DYNA Benchmark Results – ½ Cores

- **Running LS-DYNA over half number of cores doesn't generate performance benefit**
  - Performance difference is negligible

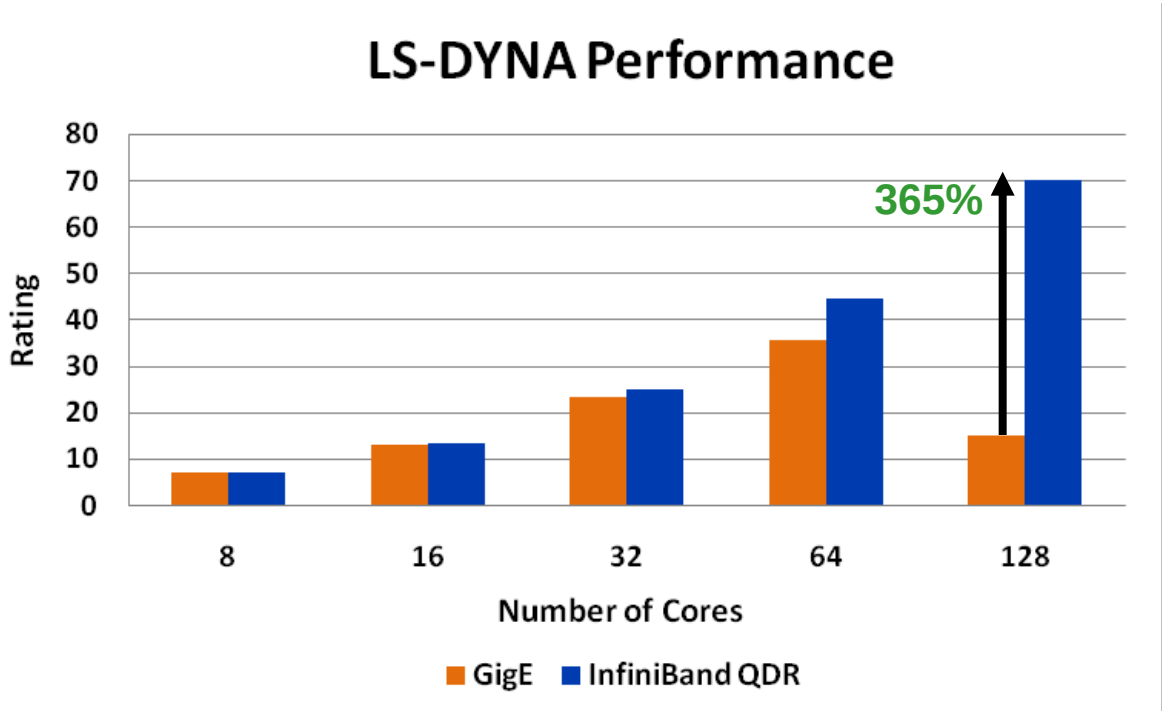


Higher is better

InfiniBand QDR based

LS-DYNA Benchmark Results - Interconnects

- **InfiniBand enables better application performance and scalability**
  - Up to 365% higher performance than GigE
  - GigE stops scaling after 8 nodes
- **Application performance over InfiniBand scales as cluster size increases**

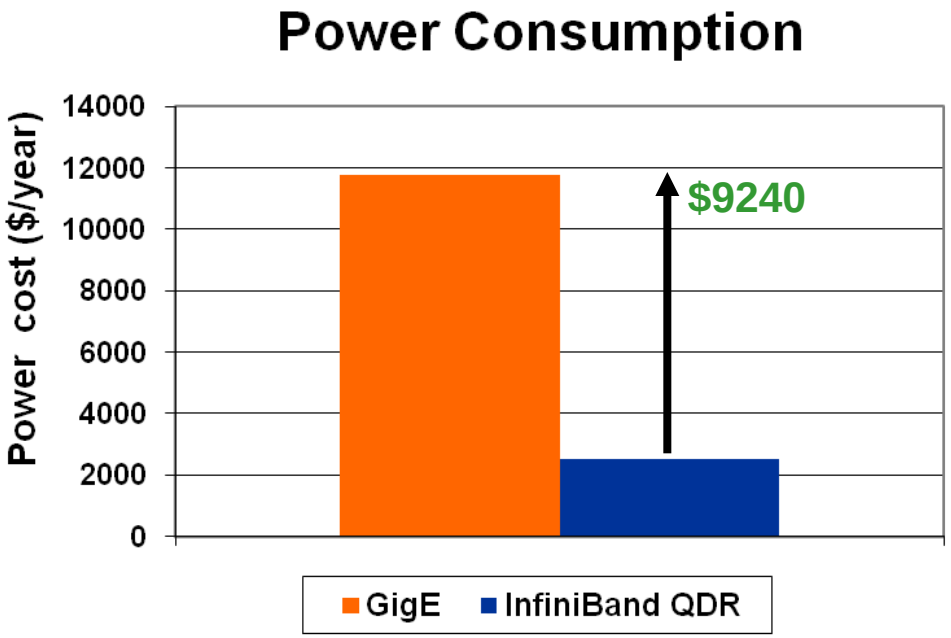


Higher is better

8-cores per node

Power Cost Savings with Different Interconnect

- **InfiniBand saves up to $9240 power compared to GigE**
  - To finish the same number of LS-DYNA jobs
  - Yearly based for 16-node cluster
- **As cluster size increases, more power can be saved**



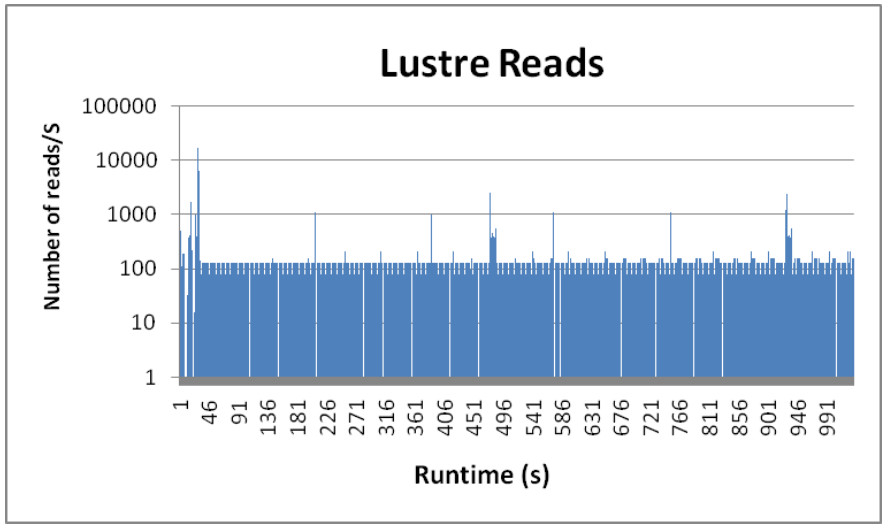
*$/KWh = KWh * $0.20*
For more information - http://enterprise.amd.com/Downloads/svrpwrusecompletefinal.pdf

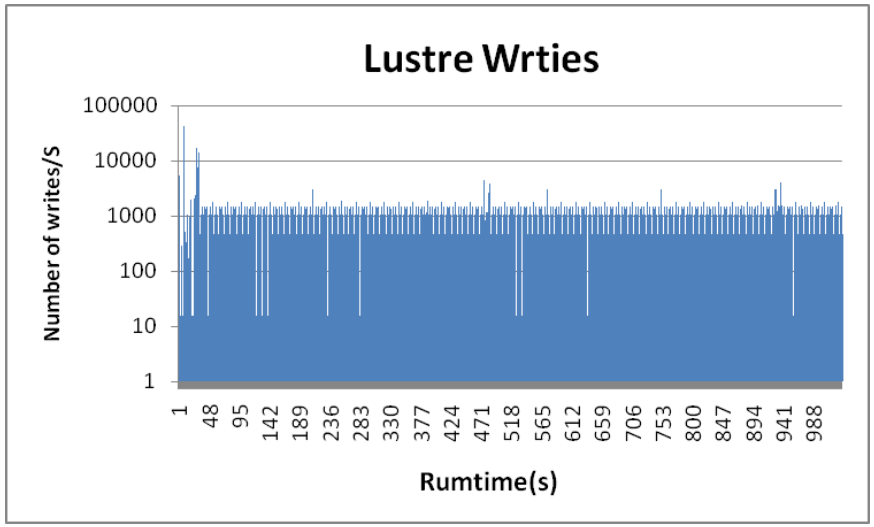
LS-DYNA Benchmark Results Summary

- **Balanced system – CPU, memory, Interconnect that match each other capabilities - is essential for providing application efficiency**
- **Performance Optimization**
  - MPI libraries showed comparable performance overall
    - Intel compilers provide moderate performance boost
  - Lustre with IB delivers increased performance
  - Reducing core utilization/access (from 4 to 2 core, for example) does not yield performance increase
- **Interconnect Characterization**
  - InfiniBand continues to deliver superior performance across a broad range of system sizes
  - GigE scalability is limited beyond 8 nodes
- **Power Analysis**
  - System architecture can yield nearly $10K annually in power savings

LS-DYNA File I/O Performance - Lustre

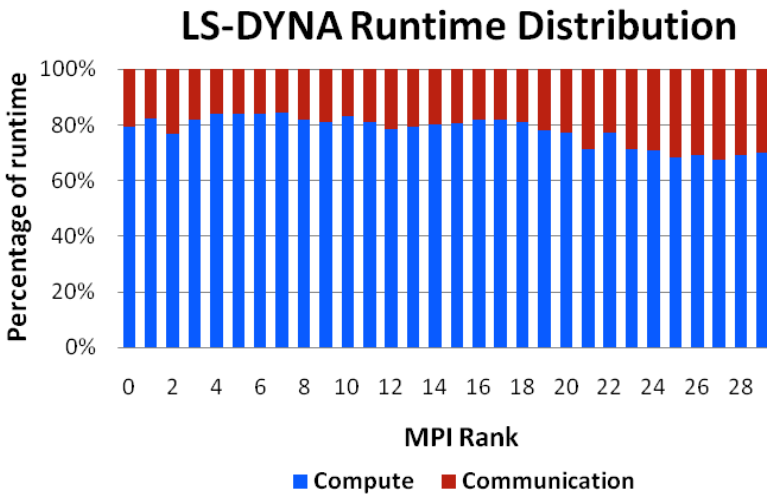
- **Large number of file I/O operations are performed**
  - Larger than 100 reads and1K writes per second
  - On 16 nodes cluster

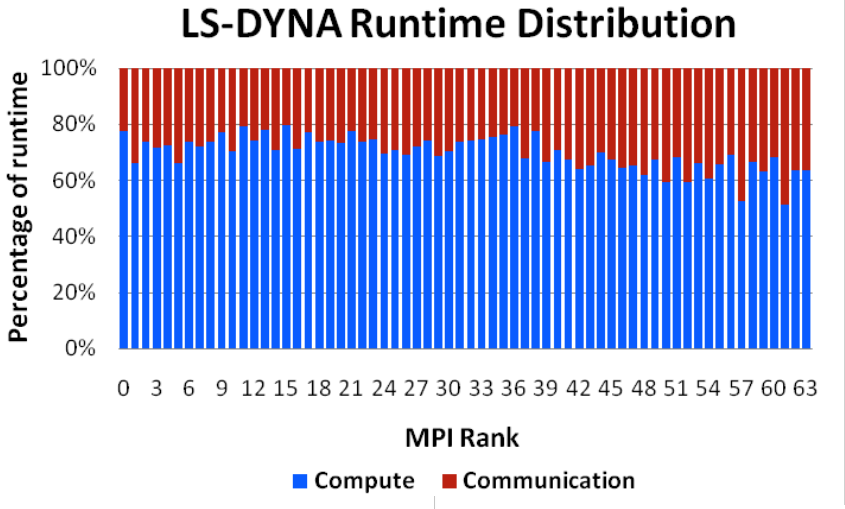


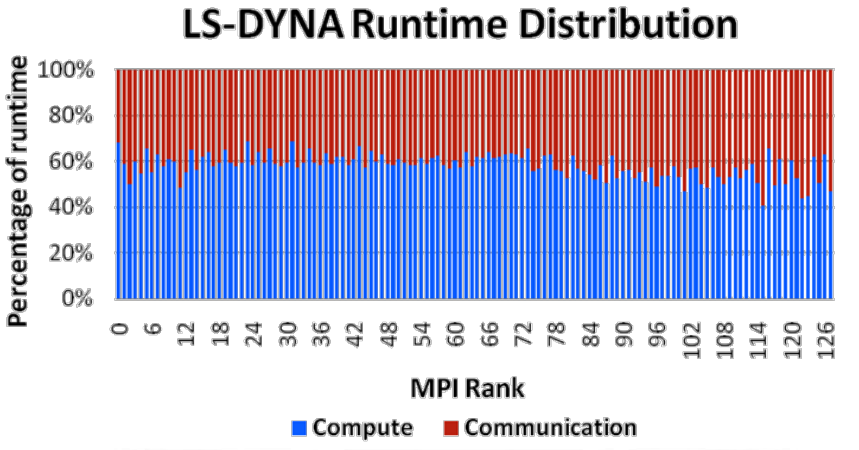


LS-DYNA MPI Profiling - % of Total Runtime

- **Percentage of communication time increases as cluster size scales**
  - 35% at 32 processes, increases up to 55% at 128 processes

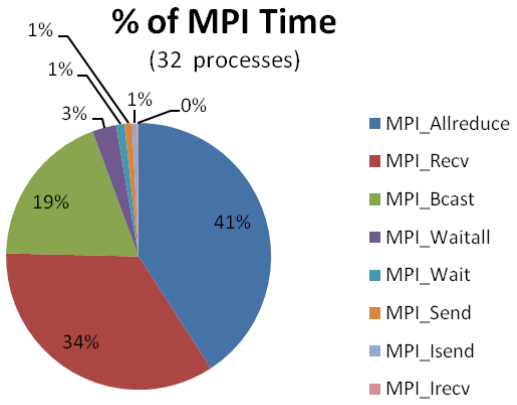


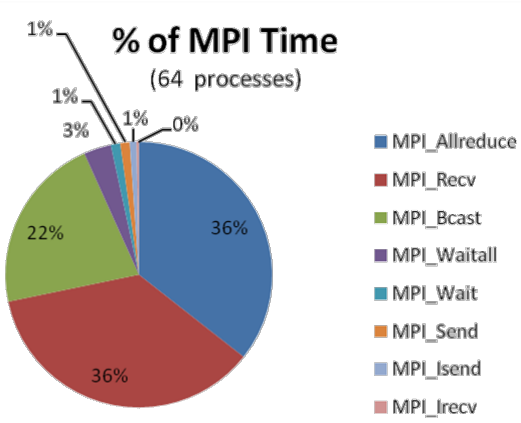


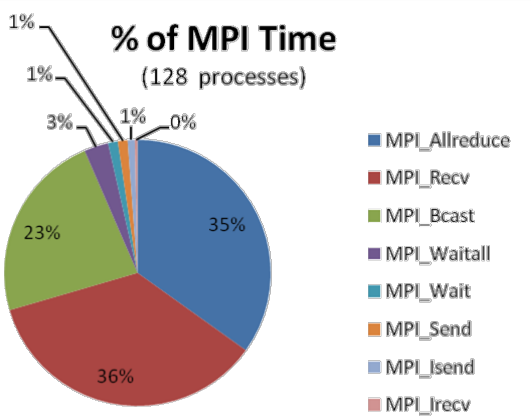


LS-DYNA MPI Profiling - % of MPI Time

- **Two MPI collectives (MPI_Allreduce and MPI_Bcast) and MPI_Recv consume more than 90% of total MPI time**

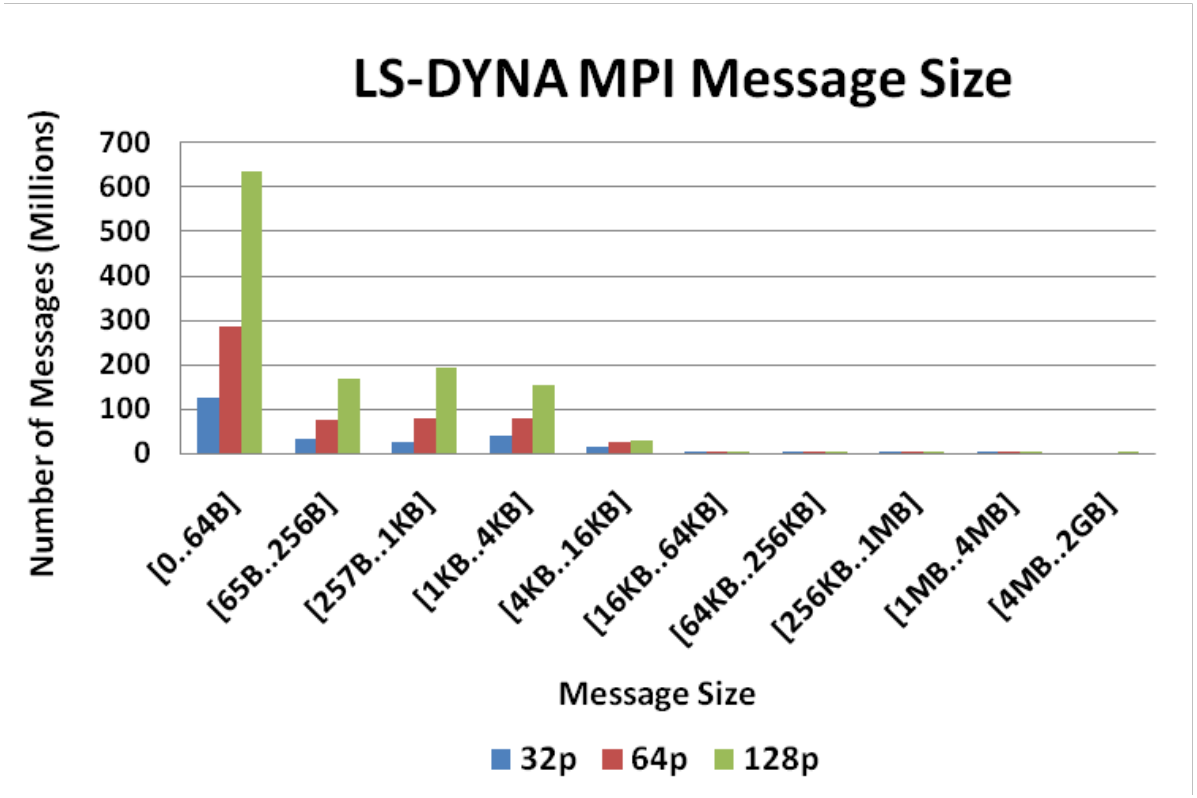






LS-DYNA MPI Profiling - Message Size

- **Majority message are small messages**
  - < 64B
- **Number of messages increases linearly as cluster size scales**



LS-DYNA Profiling Summary

- **LS-DYNA was profiled to identify its communication patterns**
- **LS-DYNA generates large number of file I/O operations**
- **Percentage of time in communication grows faster relative to computation**
- **MPI Collective functions dominate total MPI communication time**
  - More than 50% MPI time is spent in MPI_Allreduce and MPI_Bcast
  - MPI_Recv consumes 36% of total MPI time
- **Majority MPI messages are smaller than 4KB**
  - Total number of messages increases with cluster size
- **Interconnects effect to LS-DYNA performance**
  - Interconnect latency is critical to LS-DYNA performance
- **Balanced system – CPU, memory, Interconnect that match each other capabilities, is essential for providing application efficiency**

Productive Systems = Balanced System

- **Balanced system enables highest productivity**
  - Interconnect performance to match CPU capabilities
  - CPU capabilities to drive the interconnect capability
  - Memory bandwidth to match CPU performance
- **Applications scalability relies on balanced configuration**
  - “Bottleneck free”
  - Each system components can reach it’s highest capability
- **Dell M610 system integrates balanced components**
  - Intel “Nehalem” CPUs and Mellanox InfiniBand QDR
    - Latency to memory and Interconnect latency at the same magnitude of order

Thank You

HPC Advisory Council