

LS-DYNA Performance Benchmark and Profiling on Windows

July 2009



- **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
www.microsoft.com/hpc

- **LS-DYNA**

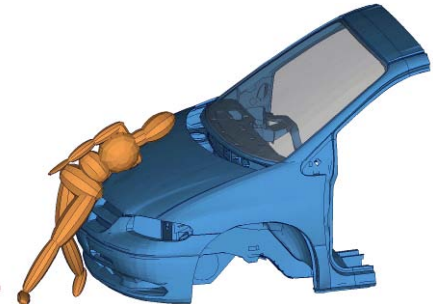
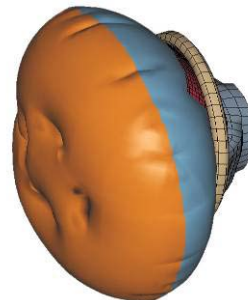
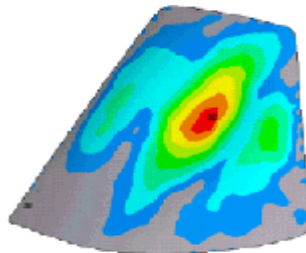
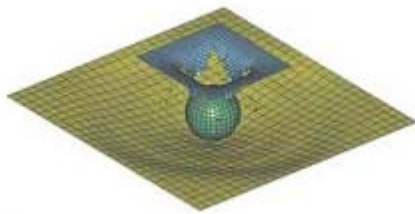
- A general purpose structural and fluid analysis simulation software package capable of simulating complex real world problems
- Developed by the Livermore Software Technology Corporation (LSTC)

- **LS-DYNA used by**

- Automobile
- Aerospace
- Construction
- Military
- Manufacturing
- Bioengineering



- **LS-DYNA 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

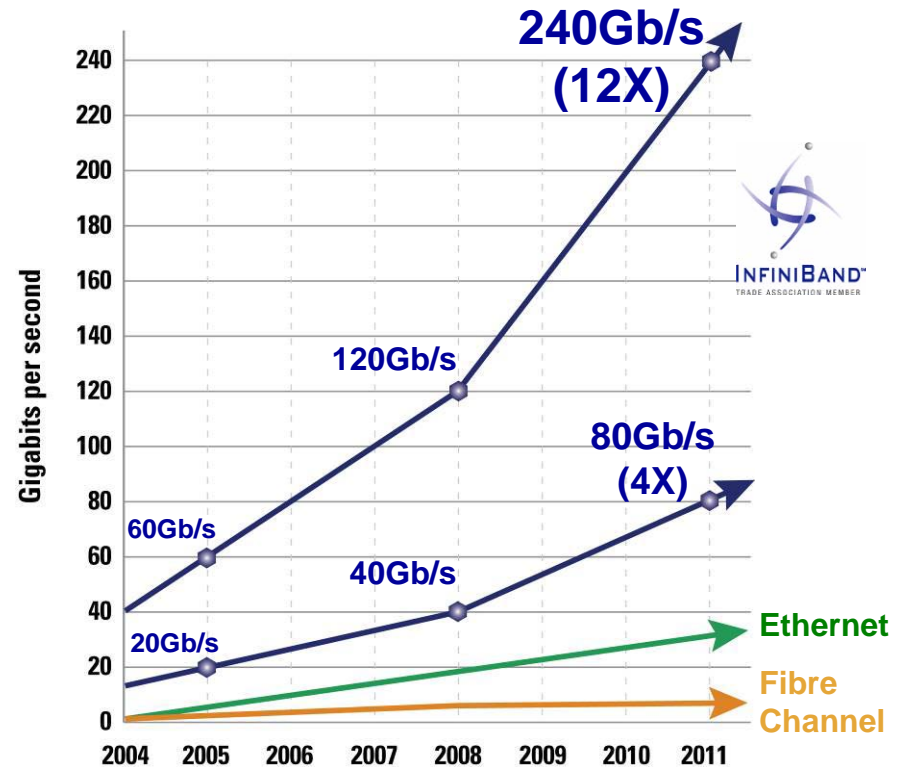


- **The presented research was done to provide best practices**
 - LS-DYNA performance benchmarking
 - LS-DYNA scaling with Windows and Linux
 - Power consumption comparison between Windows and Linux

- **Dell™ PowerEdge™ M605 10-node cluster**
- **Quad-Core AMD Opteron™ 2389 (“Shanghai”) CPUs**
- **Mellanox® InfiniBand ConnectX® 20Gb/s (DDR) Mezz card**
- **Mellanox® InfiniBand DDR Switch Module**
- **Memory: 8GB memory, DDR2 800MHz per node**
- **Windows Server 2008 HPC edition, Mellanox WinOF v2.0, MS MPI**
- **Linux RHEL5U3, OFED1.4, HP-MPI**
- **Application: LS-DYNA MPP971_S_R4.2.1**
- **Benchmark Workload**
 - Three Vehicle Collision Test simulation

- **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
- 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/clock peak per core

- Integrated Memory Controller

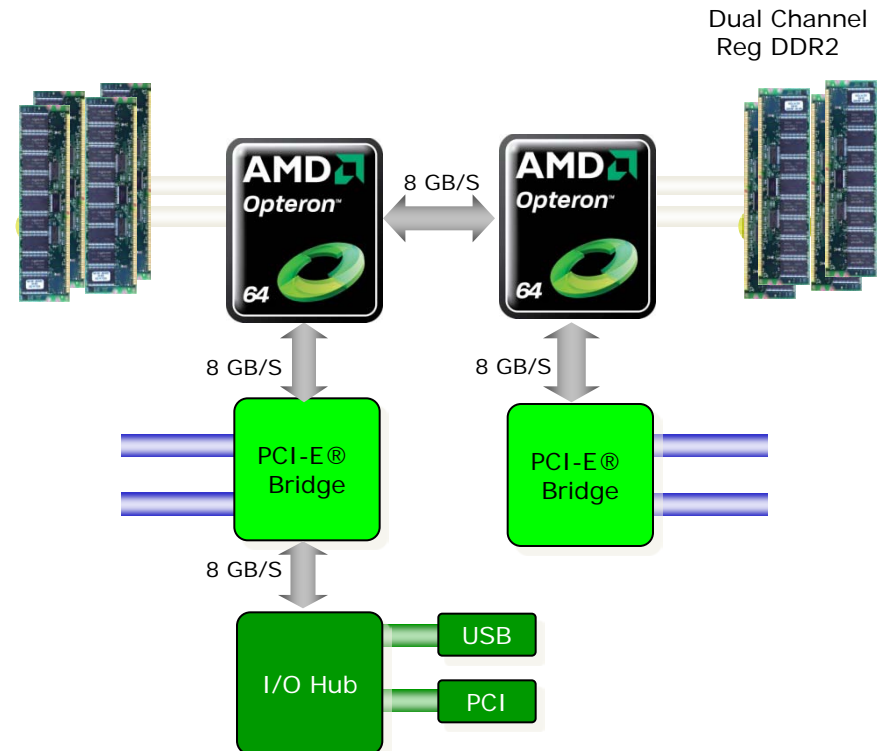
- Up to 12.8 GB/s
- DDR2-800 MHz or DDR2-667 MHz

- **Scalability**

- 48-bit Physical Addressing

- **Compatibility**

- Same power/thermal envelopes as 2nd / 3rd generation AMD Opteron™ processor



- **System Structure and Sizing Guidelines**

- 8-node cluster build with Dell PowerEdge™ M605 blades
- 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

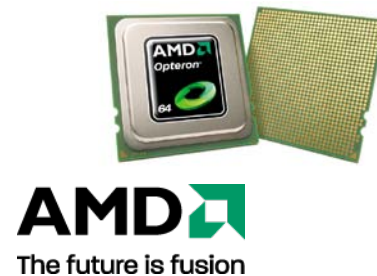
- **Workload Modeling**

- Optimized System Size, Configuration and Workloads
- Test-bed Benchmarks
- ISV Applications Characterization
- Best Practices & Usage Analysis



Dell PowerEdge™ Server Advantage

- Dell™ PowerEdge™ servers incorporate AMD Opteron™ and Mellanox ConnectX InfiniBand to provide leading edge performance and reliability
- Building Block Foundations for best price/performance and performance/watt
- Investment protection and energy efficient
- Longer term server investment value
- Faster DDR2-800 memory
- Enhanced AMD PowerNow!
- Independent Dynamic Core Technology
- AMD CoolCore™ and Smart Fetch Technology
- Mellanox InfiniBand end-to-end for highest networking performance



Current Issues

- ❖ HPC and IT data centers merging: isolated cluster management
- ❖ Developers can't easily program for parallelism
- ❖ Users don't have broad access to the increase in processing cores and data



How can Microsoft help?

- ❖ Well positioned to mainstream integration of application parallelism
- ❖ Have already begun to enable parallelism broadly to the developer community
- ❖ Can expand the value of HPC by integrating productivity and management tools



Microsoft Investments in HPC

- ❖ Comprehensive software portfolio: Client, Server, Management, Development, and Collaboration
- ❖ Dedicated teams focused on Cluster Computing
- ❖ Unified Parallel development through the Parallel Computing Initiative
- ❖ Partnerships with the Technical Computing Institutes

NetworkDirect

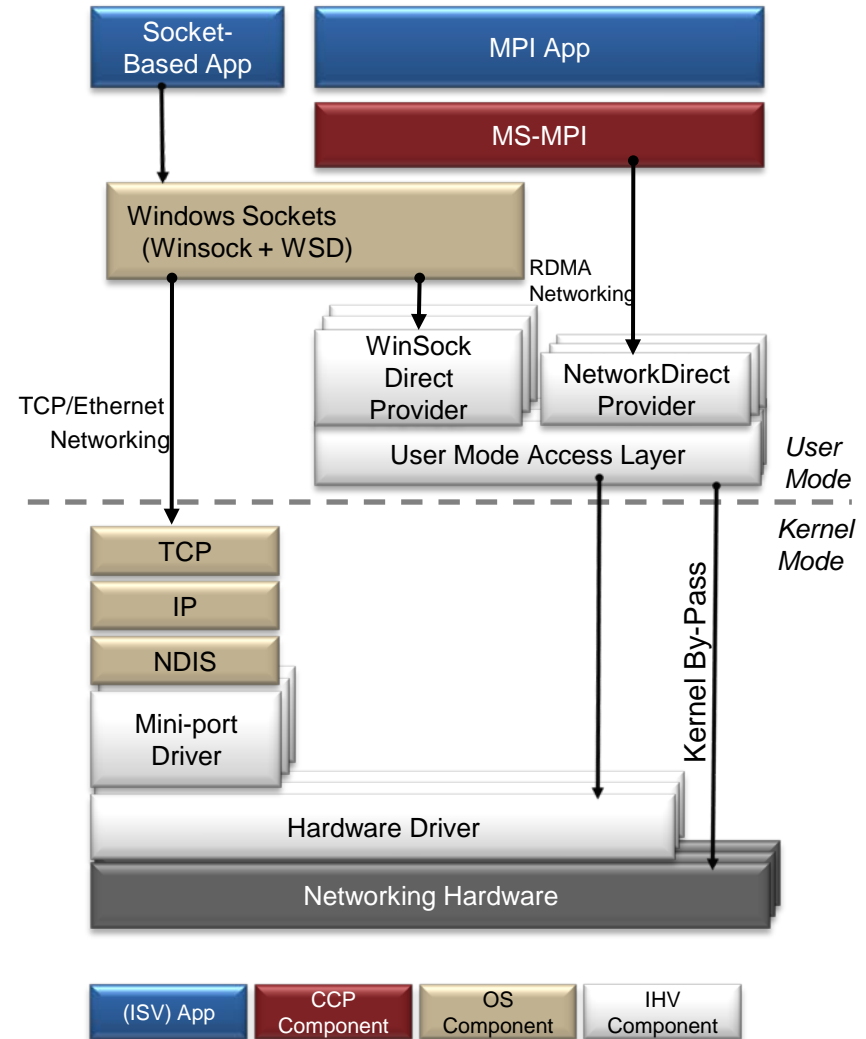
A new RDMA networking interface built for speed and stability

- **Priorities**

- Comparable with hardware-optimized MPI stacks
 - Focus on **MPI-Only Solution for version 2**
- Verbs-based design for close fit with native, high-perf networking interfaces
- Coordinated w/ Win Networking team's long-term plans

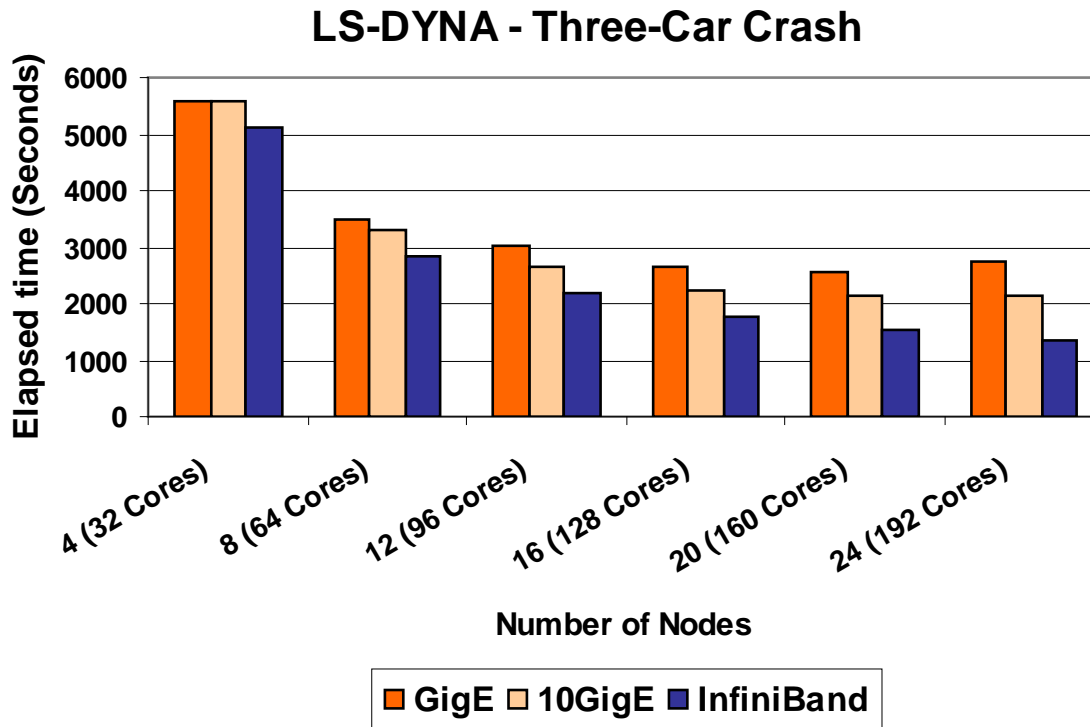
- **Implementation**

- MS-MPIv2 capable of 4 networking paths:
 - Shared Memory between processors on a motherboard
 - TCP/IP Stack (“normal” Ethernet)
 - Winsock Direct (and SDP) for sockets-based RDMA
 - New RDMA networking interface
- HPC team partners with networking IHVs to develop/distribute drivers for this new interface



LS-DYNA Performance Results - Linux

- InfiniBand 20Gb/s vs 10GigE vs GigE, 24-node system
- InfiniBand 20Gb/s (DDR) outperforms 10GigE and GigE in all test cases
 - Reducing run time by up to 25% versus 10GigE and 50% vs GigE
- Performance loss shown beyond 16 nodes with 10GigE and GigE
- InfiniBand 20Gb/s maintain scalability with cluster size

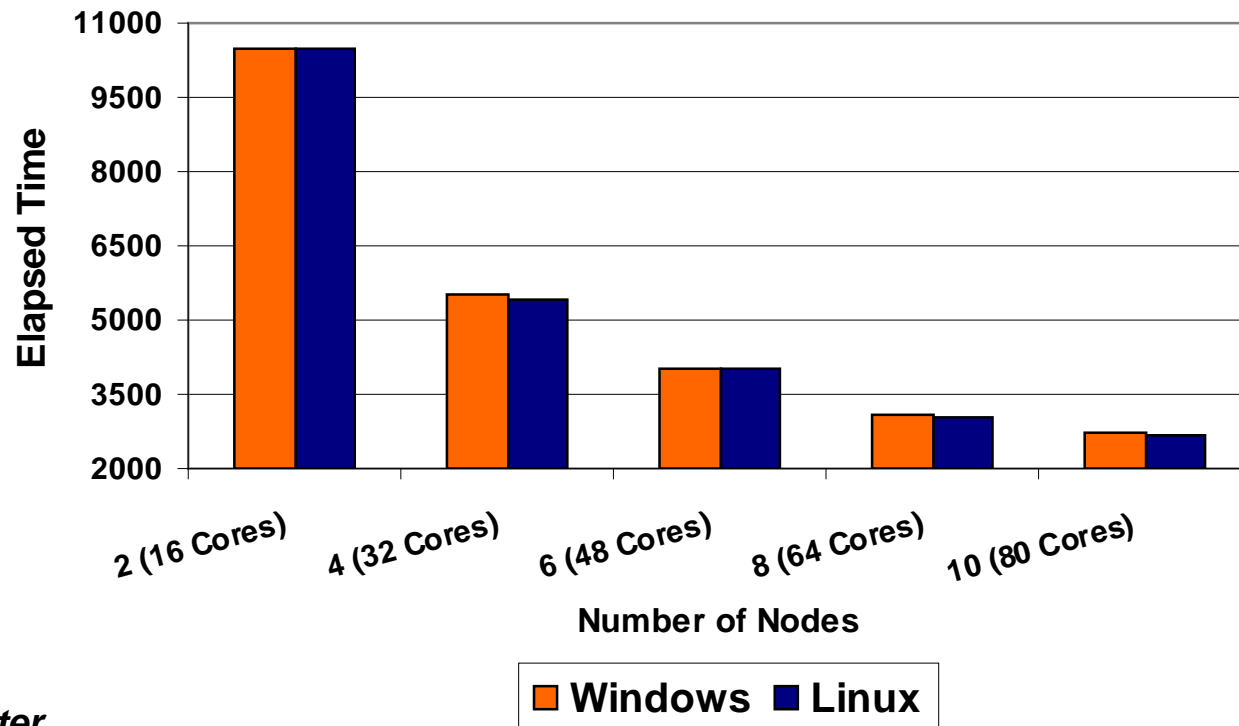


Lower is better

LS-DYNA Performance – Linux vs Windows

- The testing were limited to 10-nodes system at the given time
- Windows delivers comparable performance to Linux
- InfiniBand enables high scalability for both systems

LS-DYNA Benchmark Result
(Three-Car Crash)

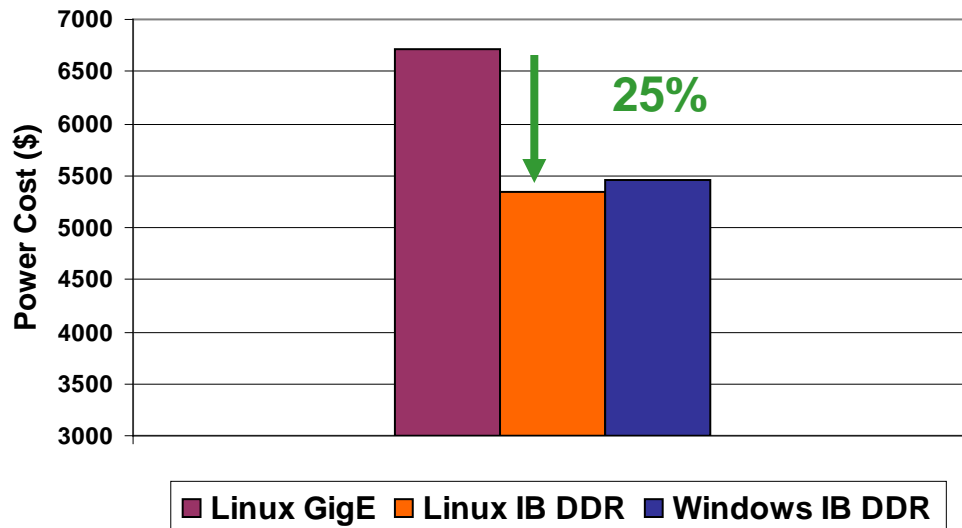


InfiniBand DDR

- **Dell economical integration of AMD CPUs and Mellanox InfiniBand saves up to 25% in power**
 - 10-node system comparison
 - In the 24-node system configuration, power saving was up to 50% as shown in previous publications
 - Versus using Gigabit Ethernet as the connectivity solutions
 - As cluster size increases, more power can be saved
- **Windows and Linux consumes similar power with InfiniBand**



**Power Consumption
(Three-Car Crash)**



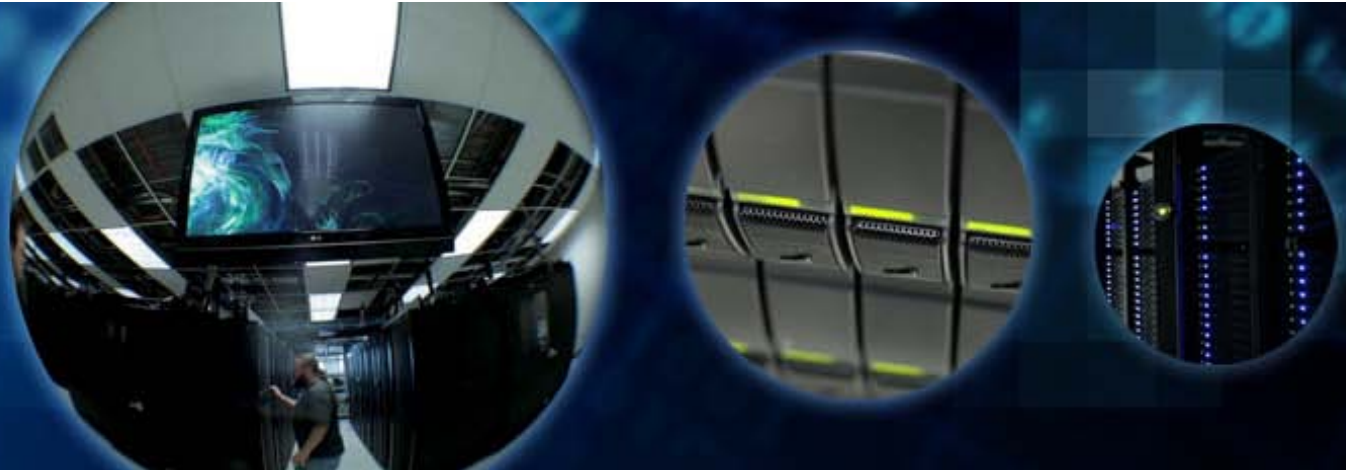
$\$/KWh = KWh * \0.20

For more information - <http://enterprise.amd.com/Downloads/svrpwrusecompletefinal.pdf>

- **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
- **LS-DYNA Performance shows**
 - Windows and Linux provide comparable performance figures
 - InfiniBand enables high scalability for both windows and Linux
- **System power consumption**
 - InfiniBand enables big power saving compared to GigE
 - Windows and Linux has same level of power consumption

Thank You

HPC Advisory Council



All trademarks are property of their respective owners. All information is provided "As-Is" without any kind of warranty. The HPC Advisory Council makes no representation to the accuracy and completeness of the information contained herein. HPC Advisory Council Mellanox undertakes no duty and assumes no obligation to update or correct any information presented herein