



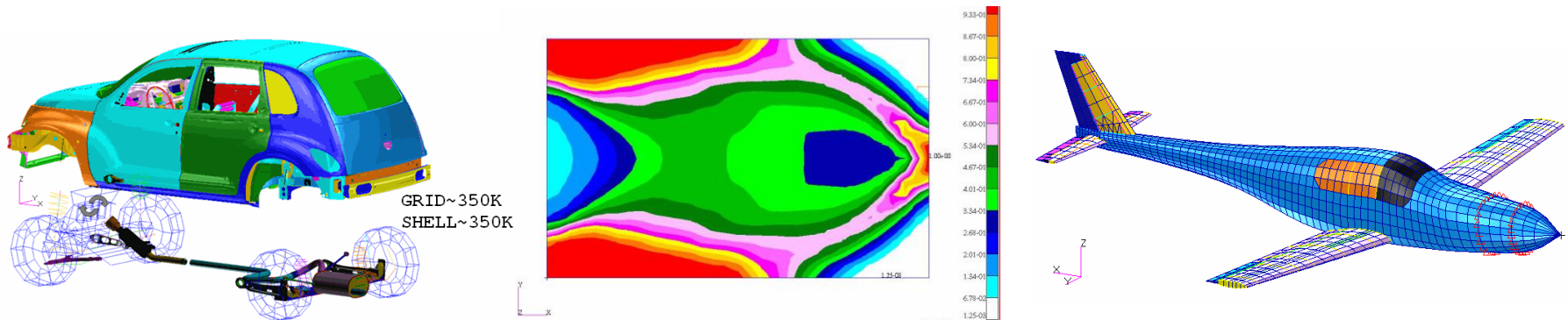
MSC Nastran Performance Benchmark and Profiling

March 2011



- **The following research was performed under the HPC Advisory Council activities**
 - Participating vendors: Intel, Dell, Mellanox
 - Compute resource - HPC Advisory Council Cluster Center
- **The following was done to provide best practices**
 - MSC Nastran performance overview
 - Understanding MSC Nastran communication patterns
 - Ways to increase MSC Nastran productivity
 - MPI libraries comparisons
- **For more info please refer to**
 - <http://www.dell.com>
 - <http://www.intel.com>
 - <http://www.mellanox.com>
 - <http://www.mscsoftware.com>

- **MSC Nastran is a widely used Finite Element Analysis (FEA) solver**
- **Used for simulating stress, dynamics, or vibration of real-world, complex systems**
- **Nearly every spacecraft, aircraft, and vehicle designed in the last 40 years has been analyzed using MSC Nastran**



- **Dell™ PowerEdge™ M610 38-node (456-core) cluster**
 - Six-Core Intel X5670 @ 2.93 GHz CPUs
 - Memory: 24GB memory, DDR3 1333 MHz
 - OS: RHEL 5.5, OFED 1.5.2 InfiniBand SW stack
- **Intel Cluster Ready certified cluster**
- **Mellanox ConnectX-2 InfiniBand adapters and non-blocking switches**
- **MPI: HP MPI 2.3, Intel MPI 3.1, Open MPI 1.2.2**
- **Application: MSC Nastran (2010.1.3)**
- **Benchmark datasets:**
 - xl0tdf1 – Car Body (Ndof 529,257, SOL111, Direct Frequency Response)
 - xx0cmd2 – Car Body (Ndof 1,315,340, SOL103, Normal Modes with ACMS)

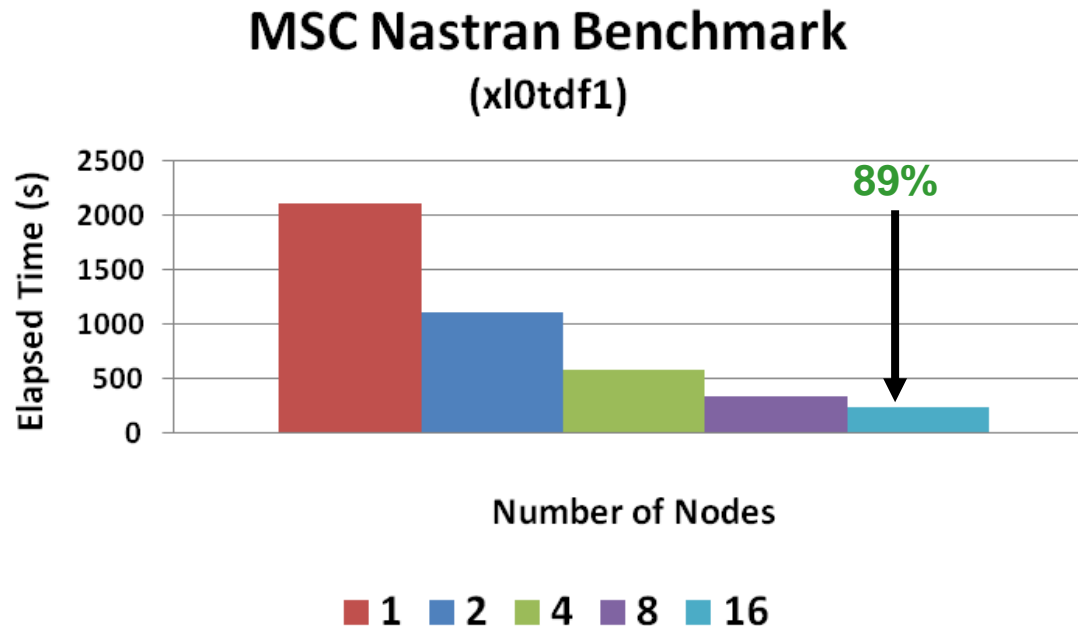
- **Intel® Cluster Ready systems make it practical to use a cluster to increase your simulation and modeling productivity**
 - Simplifies selection, deployment, and operation of a cluster
- **A single architecture platform supported by many OEMs, ISVs, cluster provisioning vendors, and interconnect providers**
 - 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



- **System Structure and Sizing Guidelines**
 - 38-node cluster build with Dell PowerEdge™ M610 blade 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



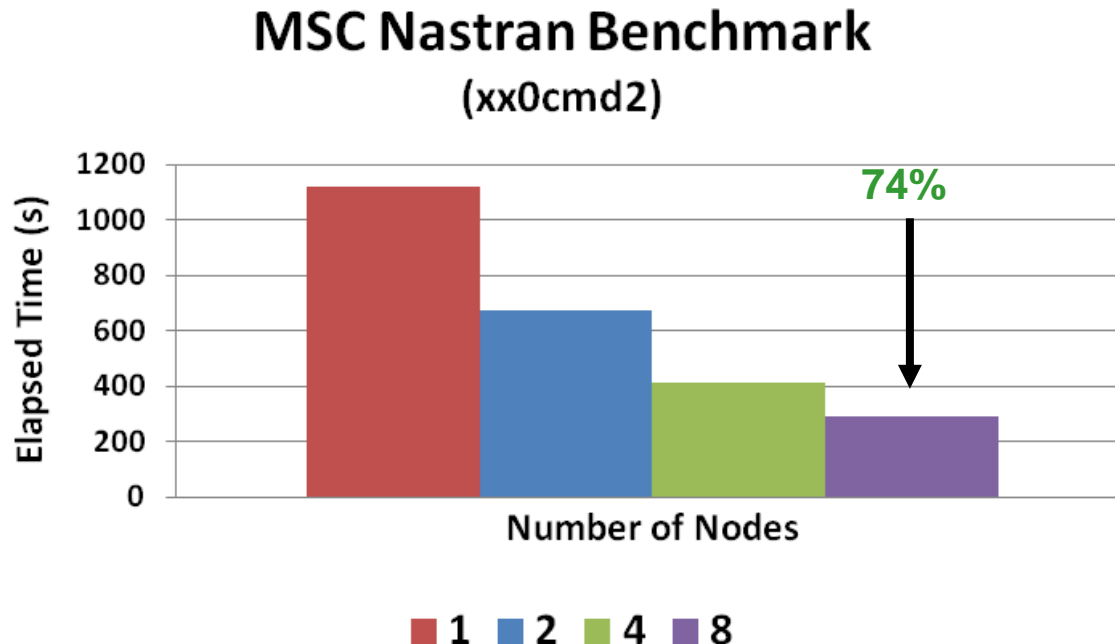
- **Input dataset: xl0tdf1**
 - Car Body (N dof 529,257, SOL111, Direct Frequency Response)
 - Memory: 520MB, SCR Disk: 5GB, Total I/O 190GB
- **Time reduced as more nodes are being utilized for computation**
 - Up to 89% in time saved by running on a 16-node cluster versus 1-node



Lower is better

InfiniBand

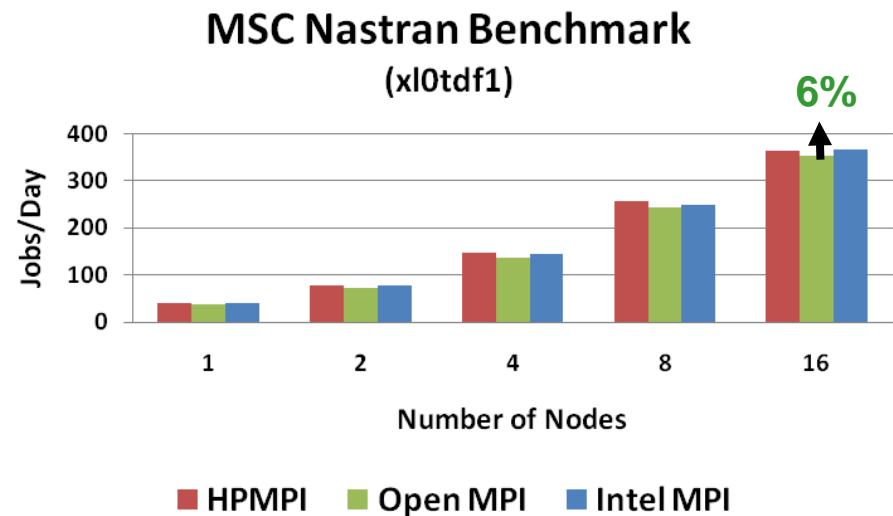
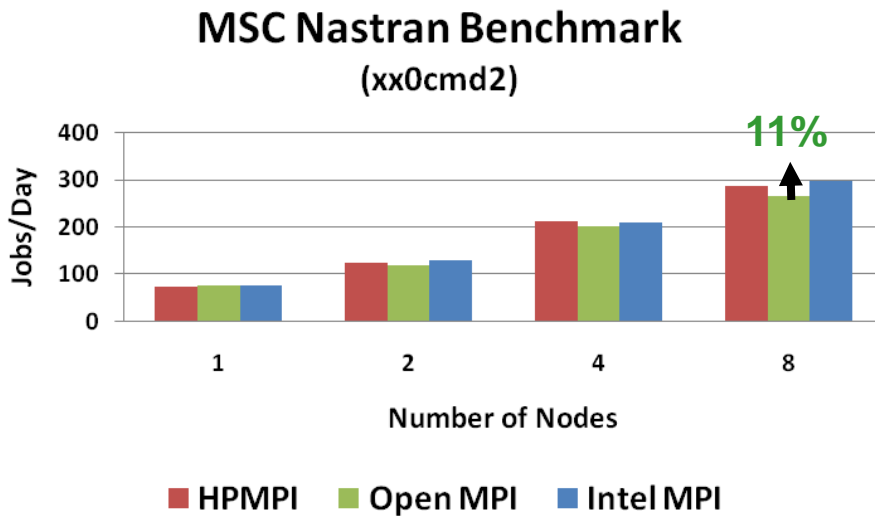
- **Input dataset: xx0cmd2**
 - Car Body (Ndof 1,315,340, SOL103, Normal Modes with ACMS)
 - Memory: 1800MB, SCR Disk: 13GB, Total I/O 202GB
- **Time reduced as more nodes are being utilized for computation**
 - Up to 74% in time saved by running on a 8-node cluster versus 1-node



Lower is better

InfiniBand

- **Intel MPI shows slightly higher performance**
 - Intel MPI runs 11% more jobs compared to Open MPI with the xx0cmd2 dataset
 - Intel MPI runs 6% more jobs compared to Open MPI with the xl0tdf1 dataset
- **Intel MPI reduces time spent in data communications**
 - Reduces time in MPI_Ssend and MPI_Recv as shown in profiling

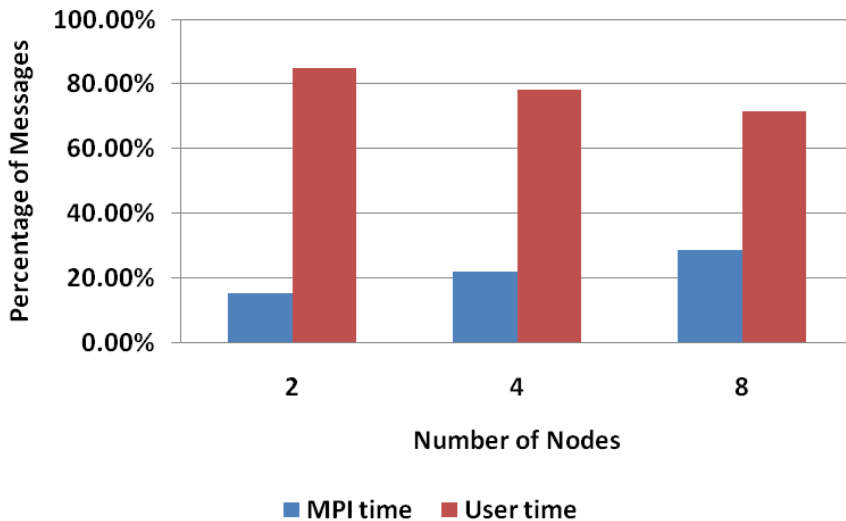


Higher is better

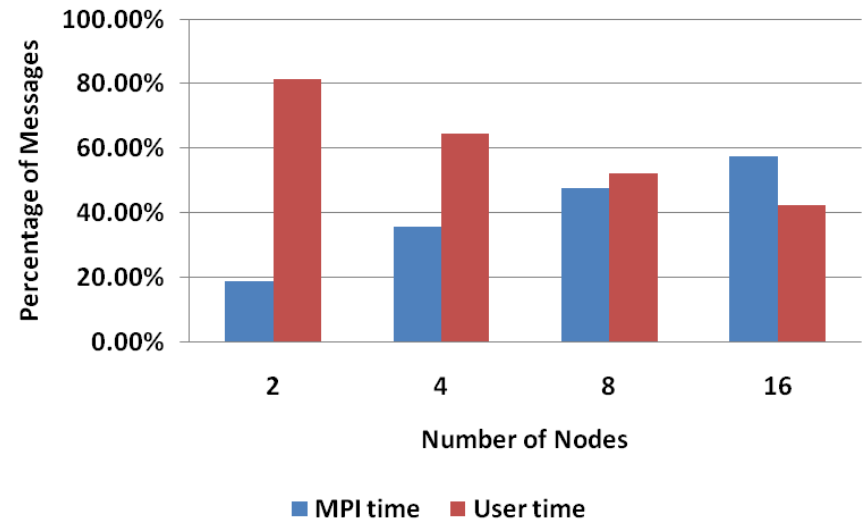
MSC Nastran Profiling – MPI/User Time Ratio

- **Communication percentage increases as the cluster scales**
- **Different communication pattern with different datasets**
 - The xl0imf1 dataset spends more time in MPI than the xx0cmd2 dataset
 - MPI time becomes more dominant than user time computation starting at 8-node

MSC Nastran Profiling
(xx0cmd2)
MPI/User Time Ratio



MSC Nastran Profiling
(xl0imf1)
MPI/User Time Ratio



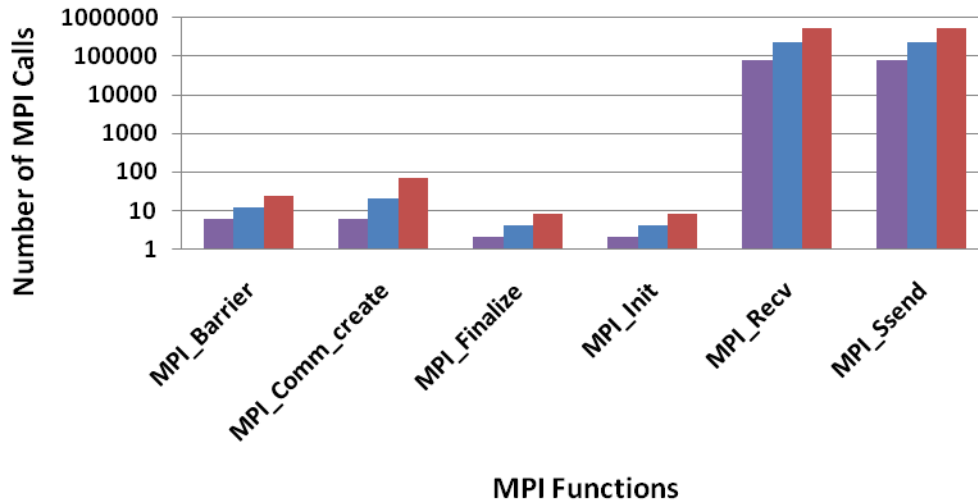
MSC Nastran Profiling – Number of MPI Calls

- **MPI_Ssend and MPI_Recv are almost used exclusively**
 - MPI_Ssend is a blocking synchronized send
 - Each of these MPI functions is accounted for half of all MPI functions
 - Only point-to-point communications, and no MPI collectives, are used
- **MPI calls increase proportionally with the node count**

MSC Nastran Profiling

(xx0cmd2)

Number of MPI Calls

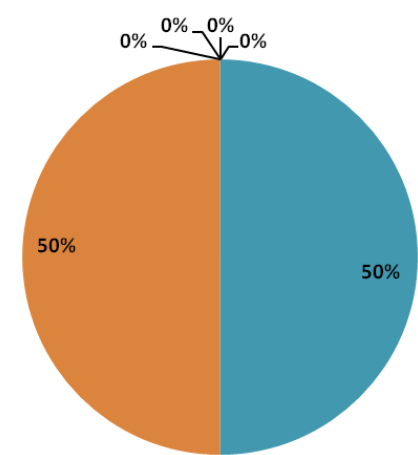


■ 2 Nodes ■ 4 Nodes ■ 8 Nodes

MSC Nastran Profiling

(xx0cmd2, 8-node, InfiniBand)

% MPI Calls

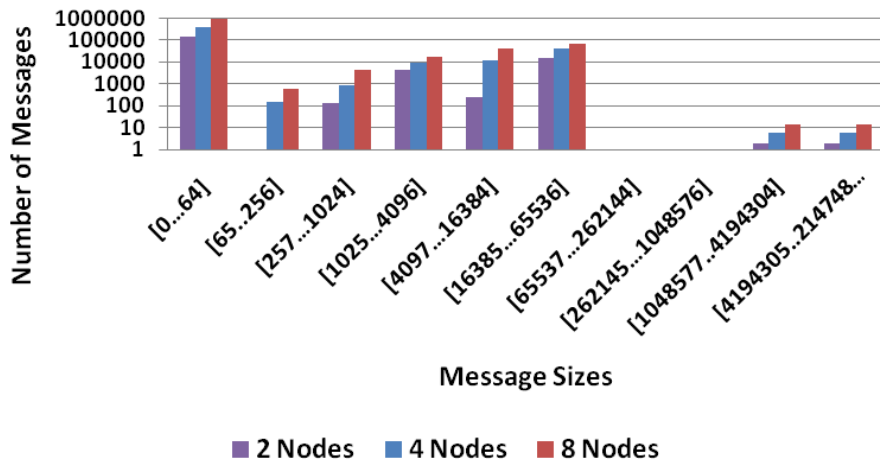


■ MPI_Barrier ■ MPI_Comm_create ■ MPI_Finalize
■ MPI_Init ■ MPI_Recv ■ MPI_Ssend

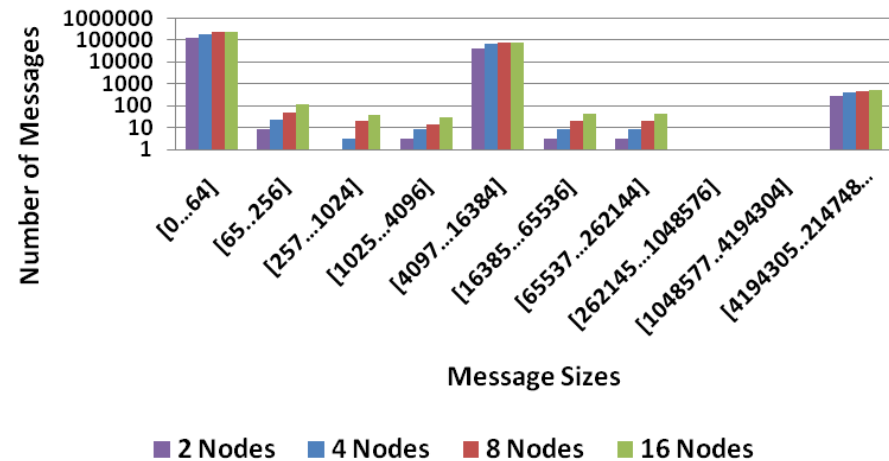
MSC Nastran Profiling – MPI Message Size

- **Majority of MPI messages are small messages**
 - In the range of 0 to 64 bytes
- **Depends on dataset, large messages also seen**
 - Between 4MB to 2GB range
 - Each of the large messages are at around 180MB

MSC Nastran Profiling
(xx0cmd2)
MPI Message Sizes



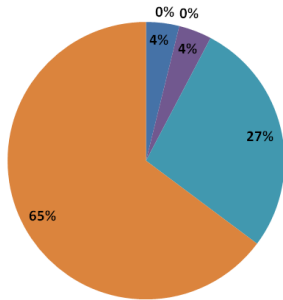
MSC Nastran Profiling
(xl0imf1)
MPI Message Sizes



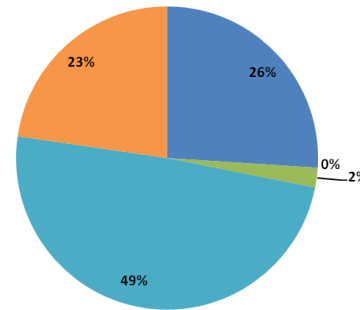
MSC Nastran Profiling – Time Spent by MPI

- **MPI_Recv is the biggest time consumer**
 - Time increases with cluster size
- **MPI_Ssend consumes more in xx0cmd2 but being overtaken by MPI_Recv later**

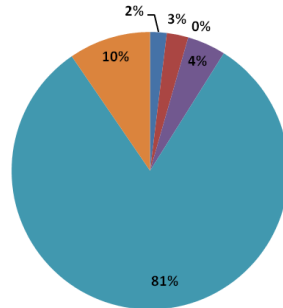
MSC Nastran Profiling
(xx0cmd2, 2-node)
% Time Spent of MPI Calls



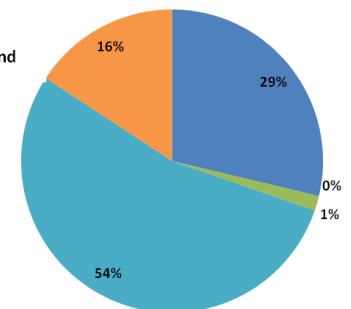
MSC Nastran Profiling
(xl0imf1, 2-node)
% Time Spent of MPI Calls



MSC Nastran Profiling
(xx0cmd2, 8-node)
% Time Spent of MPI Calls



MSC Nastran Profiling
(xl0imf1, 16-node)
% Time Spent of MPI Calls



■ MPI_Barrier ■ MPI_Comm_create ■ MPI_Finalize
■ MPI_Init ■ MPI_Recv ■ MPI_Ssend

■ MPI_Barrier ■ MPI_Finalize ■ MPI_Init ■ MPI_Recv ■ MPI_Ssend

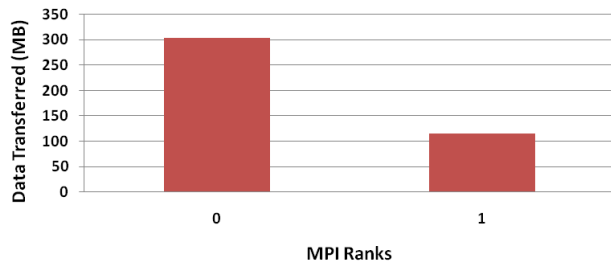
■ MPI_Barrier ■ MPI_Comm_create ■ MPI_Finalize
■ MPI_Init ■ MPI_Recv ■ MPI_Ssend

■ MPI_Barrier ■ MPI_Finalize ■ MPI_Init ■ MPI_Recv ■ MPI_Ssend

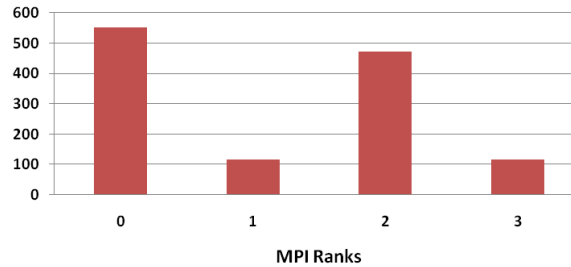
MSC Nastran Profiling – MPI Data Transfer

- Data transferred to each process mainly from the first MPI rank
- Different communication patterns for different datasets
 - Show divide-and-conquer (bifurcation) distribution for xx0cmd2 dataset
 - Shows data distributions from first MPI process for the xl0imf1 dataset

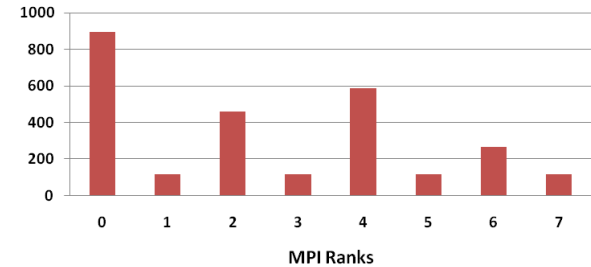
MSC Nastran Profiling
(xx0cmd2, 2-node)
Data Transferred by Ranks



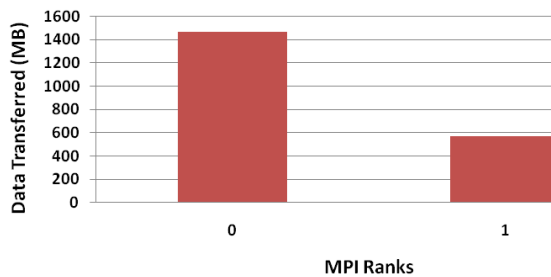
MSC Nastran Profiling
(xx0cmd2, 4-node)
Data Transferred by Ranks



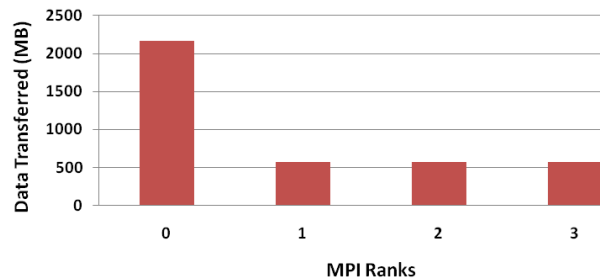
MSC Nastran Profiling
(xx0cmd2, 8-node)
Data Transferred by Ranks



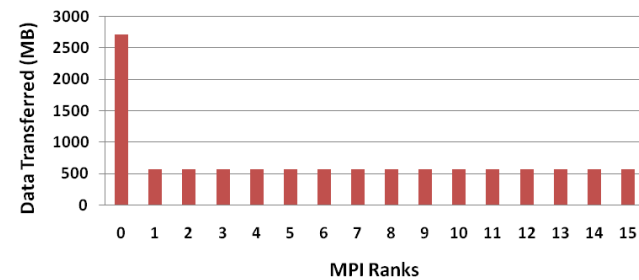
MSC Nastran Profiling
(xl0imf1, 2-node)
Data Transferred by Ranks



MSC Nastran Profiling
(xl0imf1, 4-node)
Data Transferred by Ranks

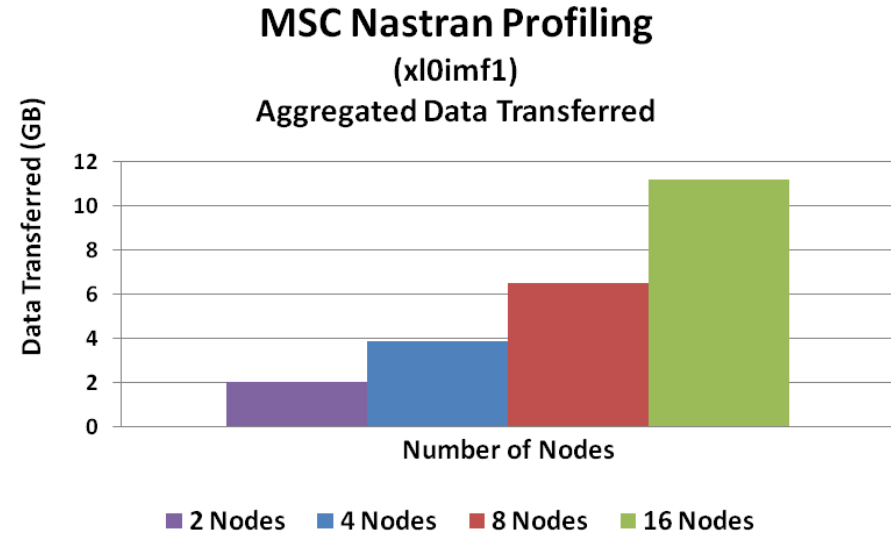
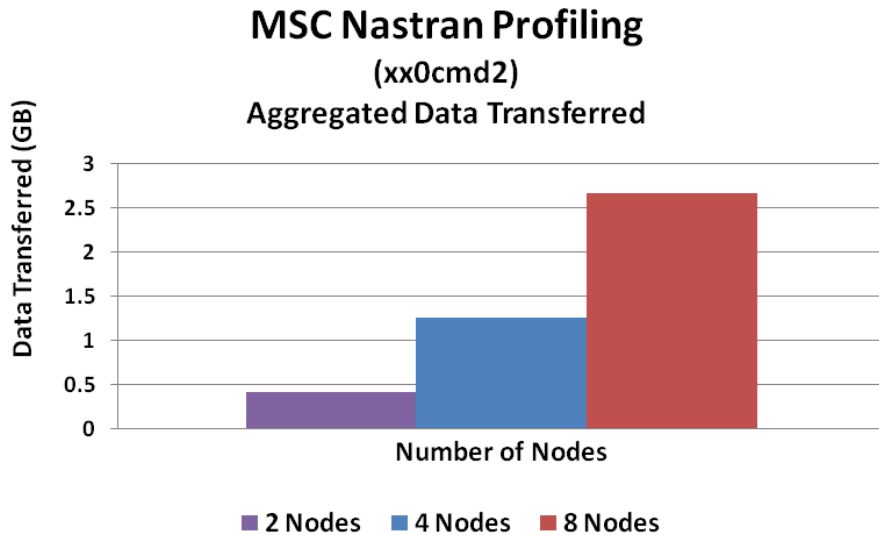


MSC Nastran Profiling
(xl0imf1, 16-node)
Data Transferred by Ranks



MSC Nastran Profiling – Aggregated Transfer

- **Aggregated data transfer refers to:**
 - Total amount of data being transferred in the network between all MPI ranks collectively
- **The total data transfer increases as the cluster scales**
- **Demonstrates the advantage and importance of high throughput interconnects**
 - InfiniBand QDR was used for the profiling testing



InfiniBand QDR

- **MSC Nastran can achieve higher performance by scaling out**
 - Take advantage by clustering with more computation resources with InfiniBand QDR
- **MPI**
 - Intel MPI performs better than Open MPI for both datasets
 - Only MPI point-to-point communications, and no MPI collectives, are used
 - MPI_Recv is the biggest MPI time consumer
- **Data distribution**
 - First MPI rank responsible for data distribution
 - Majority of messages are small messages between 0 and 64 bytes
 - A few sizable data transfer around 180MB can occurred

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