HPC ADVISORY COUNCIL

MSC Nastran 2012 Performance Benchmark and Profiling

September 2012

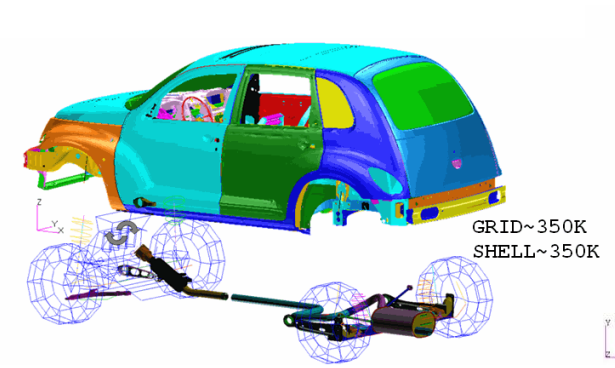
AMD DELL Mellanox TECHNOLOGIES MSC Software Simulating Reality, Delivering Certainty™

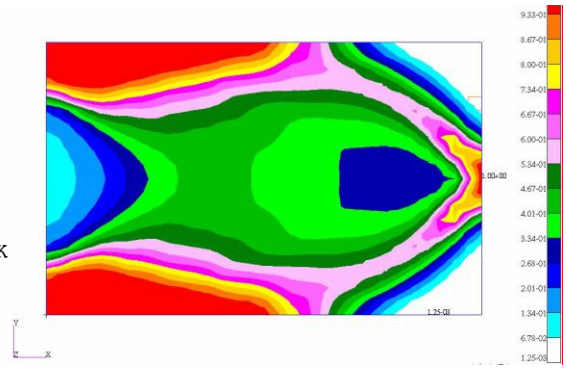
Note

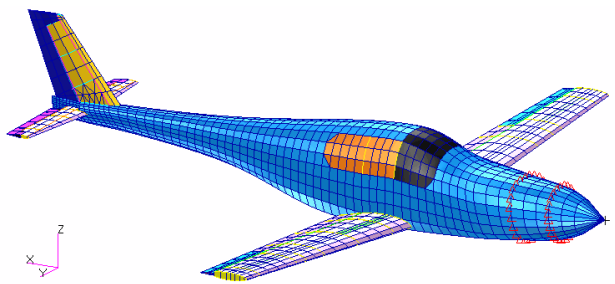
- **The following research was performed under the HPC Advisory Council activities**
  - Participating vendors: AMD, Dell, Mellanox
  - Compute resource - HPC Advisory Council Cluster Center
- **For more info please refer to**
  - http:// www.amd.com
  - http:// www.dell.com/hpc
  - http://www.mellanox.com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

MSC Nastran Application

- **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**







Objectives

- **The following was done to provide best practices**
  - MSC Nastran performance benchmarking
  - Interconnect performance comparisons
  - Understanding MSC Nastran communication patterns
  - Ways to increase MSC Nastran productivity
  - MPI libraries comparisons

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

Test Cluster Configuration

- **Dell™ PowerEdge™ C6145 6-node (384-core) cluster**
  - Memory: 128GB memory per node DDR3 1600MHz, BIOS version 2.6.0
  - 4 CPU sockets per server node
- **AMD™ Opteron™ 6276 (code name "Interlagos") 16-core @ 2.3 GHz CPUs**
- **Mellanox ConnectX®-2 VPI Adapters and IS5030 36-Port InfiniBand switch**
- **MLNX-OFED 1.5.3 InfiniBand SW stack**
- **OS: RHEL 6 Update 2**
- **Storage: 4x 15K 6Gbps 300GB on RAID 0 per node**
- **MPI (vendor provided): HP MPI 2.3, Intel MPI 4.0 Update 1, Open MPI 1.2.2**
- **Application: MSC.Nastran version 2012.2**
- **Benchmark workload: MSC.Nastran 2012.1 Parallel Tests Benchmarks**
  - Power Train (xl0tdf1) (Ndof 529,257, SOL108, Direct Frequency)
  - Car Body (xx0wmd0) (Ndof 3,799,278, SOL103, Large ACMS)

Dell™ PowerEdge™ C6145 6-node cluster

- **HPC Advisory Council Test-bed System**

- **New 6-node 384 core cluster - featuring Dell PowerEdge™ C6145 servers**
  - Replacement system for Dell PowerEdge SC1435 (192 cores) cluster system following 2 years of rigorous benchmarking and product EOL
    - System to be redirected to explore HPC in the Cloud applications

- **Workload profiling and benchmarking**
  - Characterization for HPC and compute intense environments
  - Optimization for scale, sizing and configuration and workload performance
  - Test-bed Benchmarks
    - RFPs
    - Customers/Prospects, etc
  - ISV & Industry standard application characterization
  - Best practices & usage analysis

About Dell PowerEdge™ Platform Advantages

Best of breed technologies and partners

Combination of AMD Opteron™ 6200 series platform and Mellanox ConnectX®-3 InfiniBand on Dell HPC

Solutions provide the ultimate platform for speed and scale

- Dell PowerEdge C6145 system delivers 8 socket performance in dense 2U form factor
- Up to 64 core/32DIMMs per server – 2688 core in 42U enclosure

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

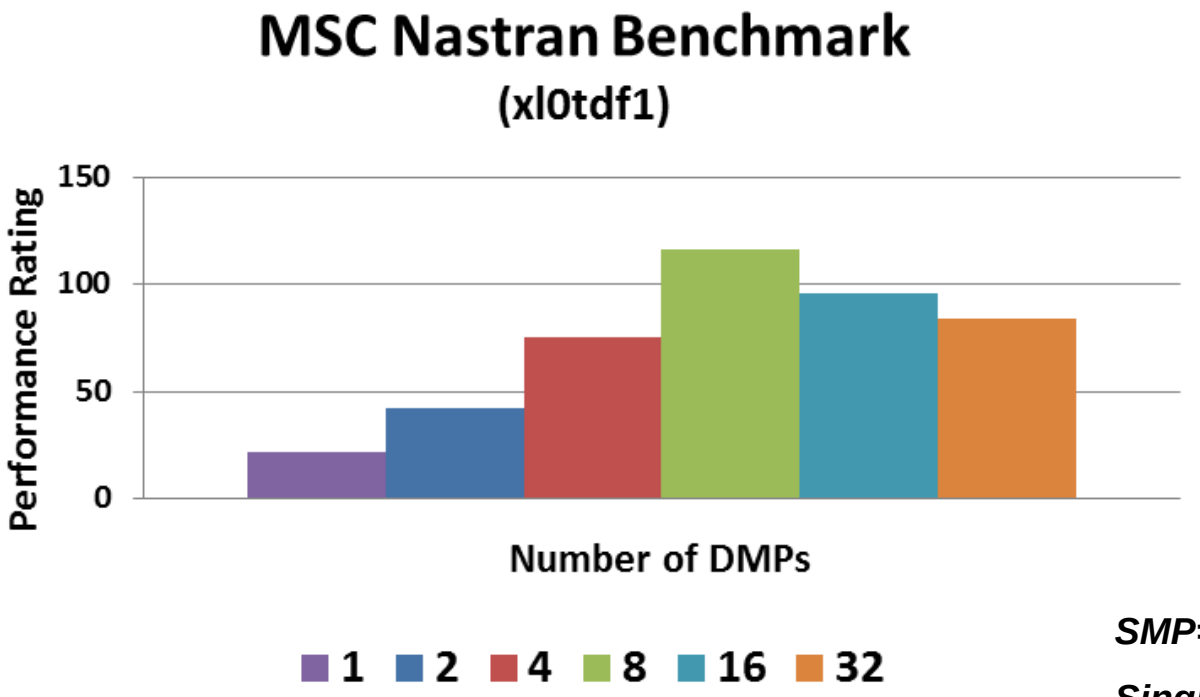
- 2x more memory and processing power in half of the space
- Energy optimized low flow fans, improved power supplies and dual SD modules

Optimized for long-term capital and operating investment protection

- System expansion
- Component upgrades and feature releases

MSC Nastran Performance – # of DMPs

- **For running DMP (Distributed Memory Parallel) job on a single node**
  - Running 8-way parallel by decomposing domains enable better performance
  - Performance drops after 8 DMPs
- **More systems are needed to achieve higher performance with more DMPs**
  - Running on multiple systems allow MSC.Nastran to scale



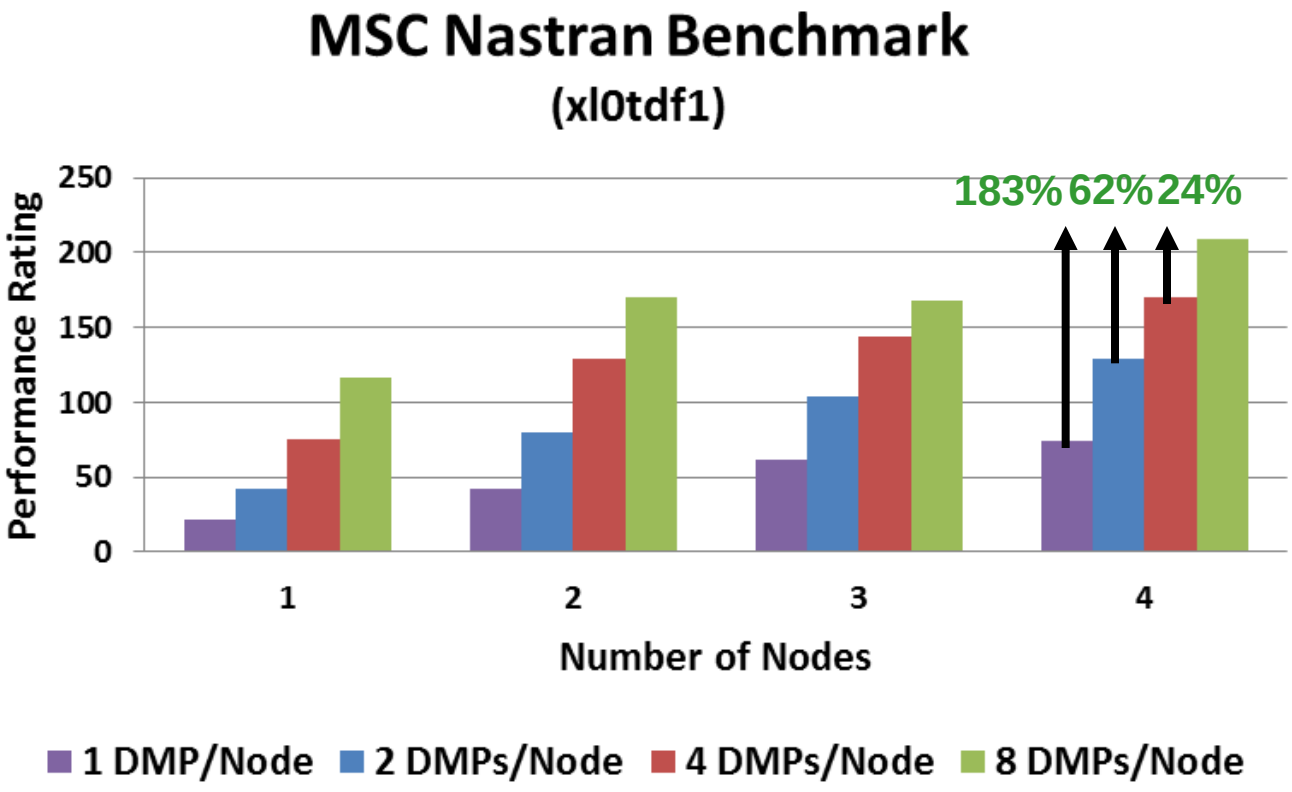
SMP=1

Single Node

Higher is better

MSC Nastran Performance – # of DMPs/Node

- **Achieve higher performance by allocating DMPs into systems**



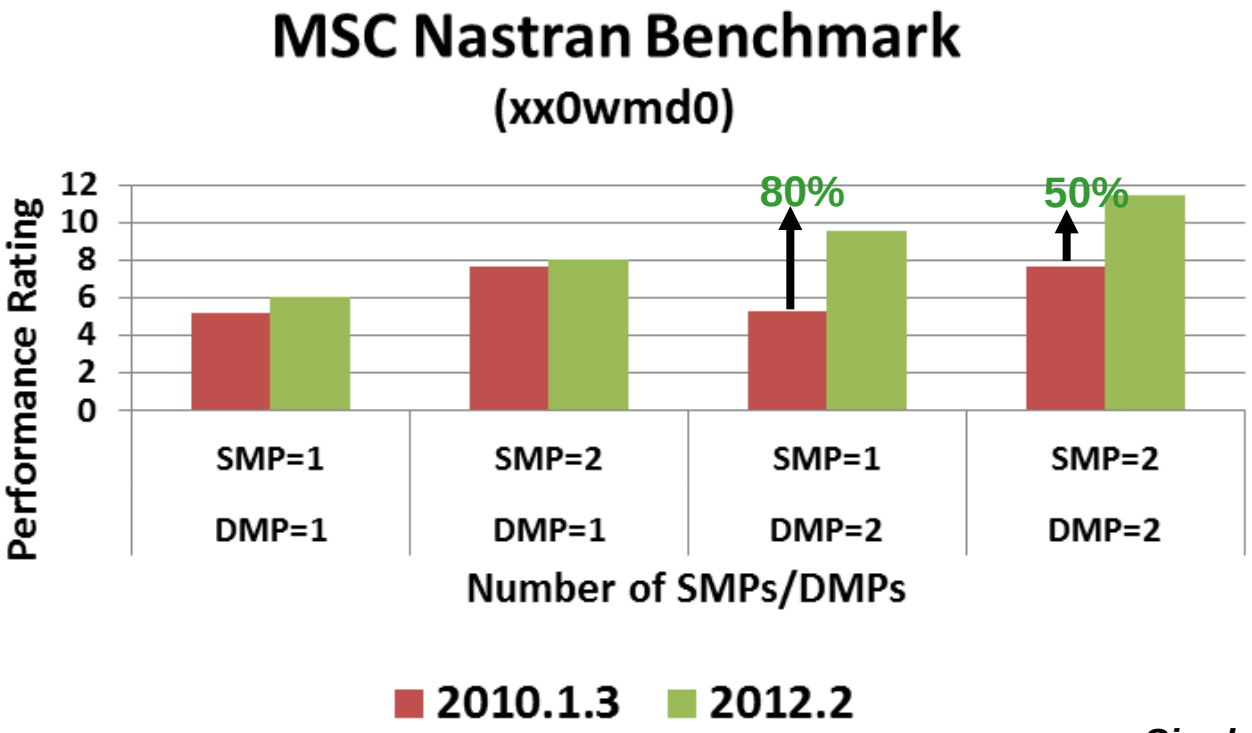
SMP=1

Higher is better

Multiple Nodes

MSC Nastran Performance – Software Version

- **Major changes in MSC.Nastran 2012.2 SOL103 includes:**
  - Improved math kernel support (w/ AVX), SMP scalability, and DMP strategy
- **Huge performance improvement observed in MSC.Nastran 2012.2**
  - Up to 80% performance gain compared to 2010.1.3 version on same system

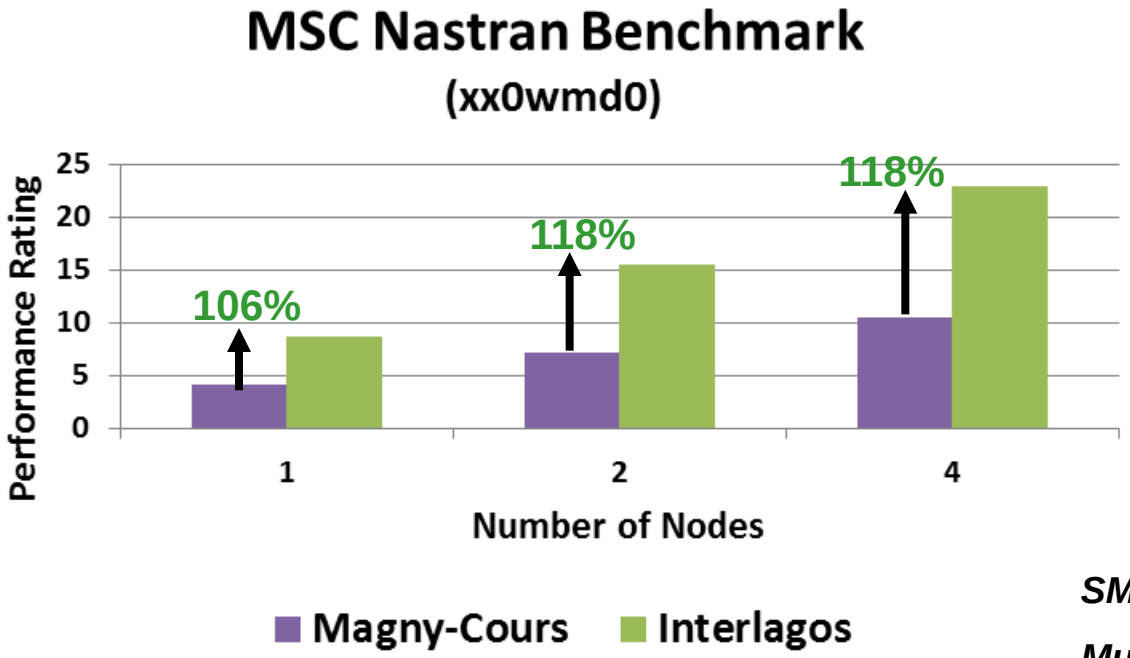


Higher is better

Single Node

MSC Nastran Performance – Generations

- **Latest CPU and software enable faster runtime over previous generation**
  - Up to 118% of performance improvement
- **Configuration differences:**
  - Magny-Cours (prior): AMD Opteron 6174 @ 2.2GHz, MSC Nastran 2010.1.3
  - Interlagos (current): AMD Opteron 6276 @ 2.3GHz, MSC Nastran 2012.2



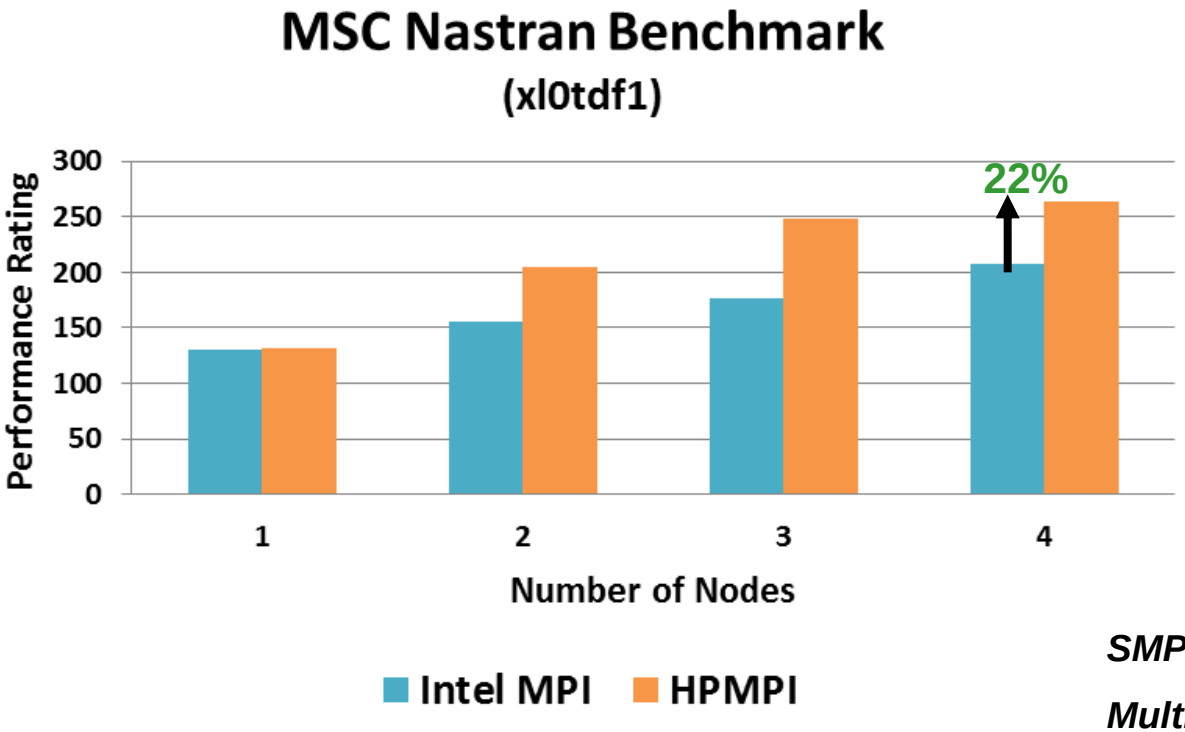
Higher is better

SMP=2, DMP=1/Node

Multiple Nodes

MSC Nastran Performance – MPI

- **HP MPI shows higher performance compared to Intel MPI**
  - Shows a gain of 22% over Intel MPI
- **The vendor-provided Open MPI was not built with InfiniBand support**
  - The openib BTL was not built with the Open MPI shipped with MSC.Nastran
  - User can build Open MPI 1.2.x separately and run with Nastran



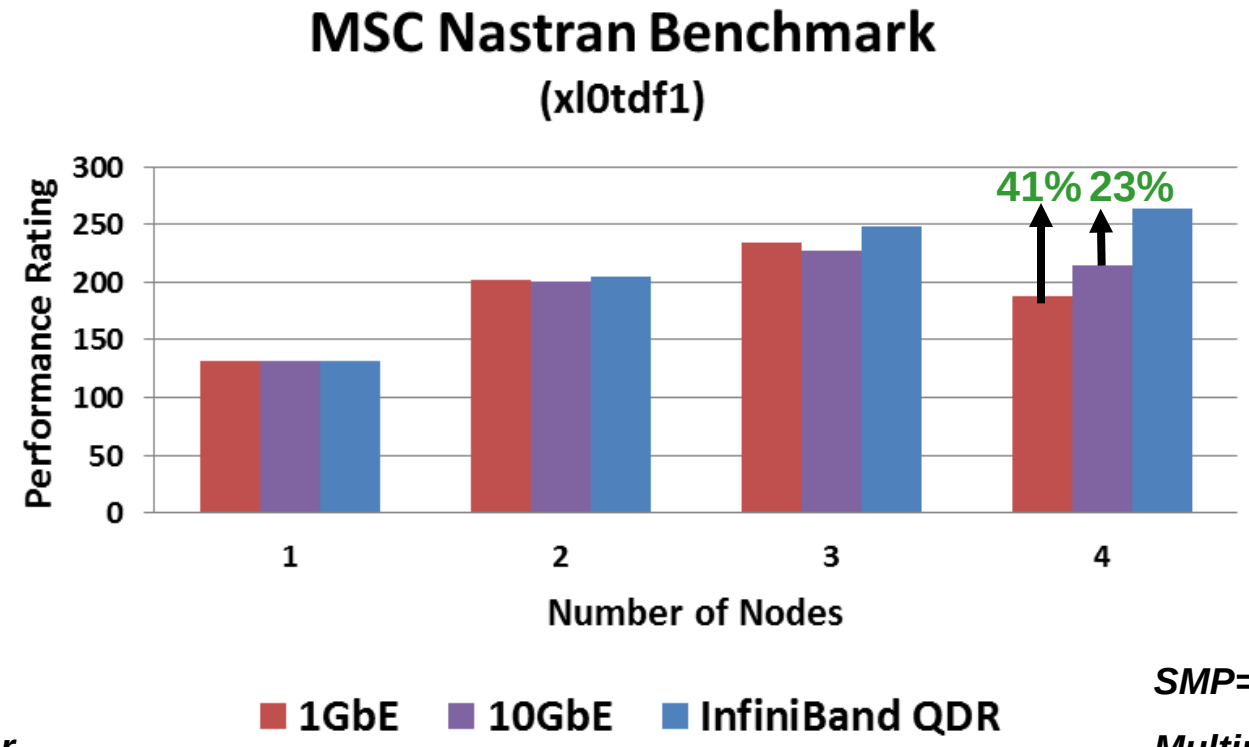
Higher is better

SMP=8, DMP=4/Node

Multiple Nodes

MSC Nastran Performance – Interconnects

- **InfiniBand leads among the network interconnects as the cluster scales**
  - Up to 23% higher performance than 10GbE on xl0tdf1
  - Up to 41% higher performance than 1GbE on xl0tdf1



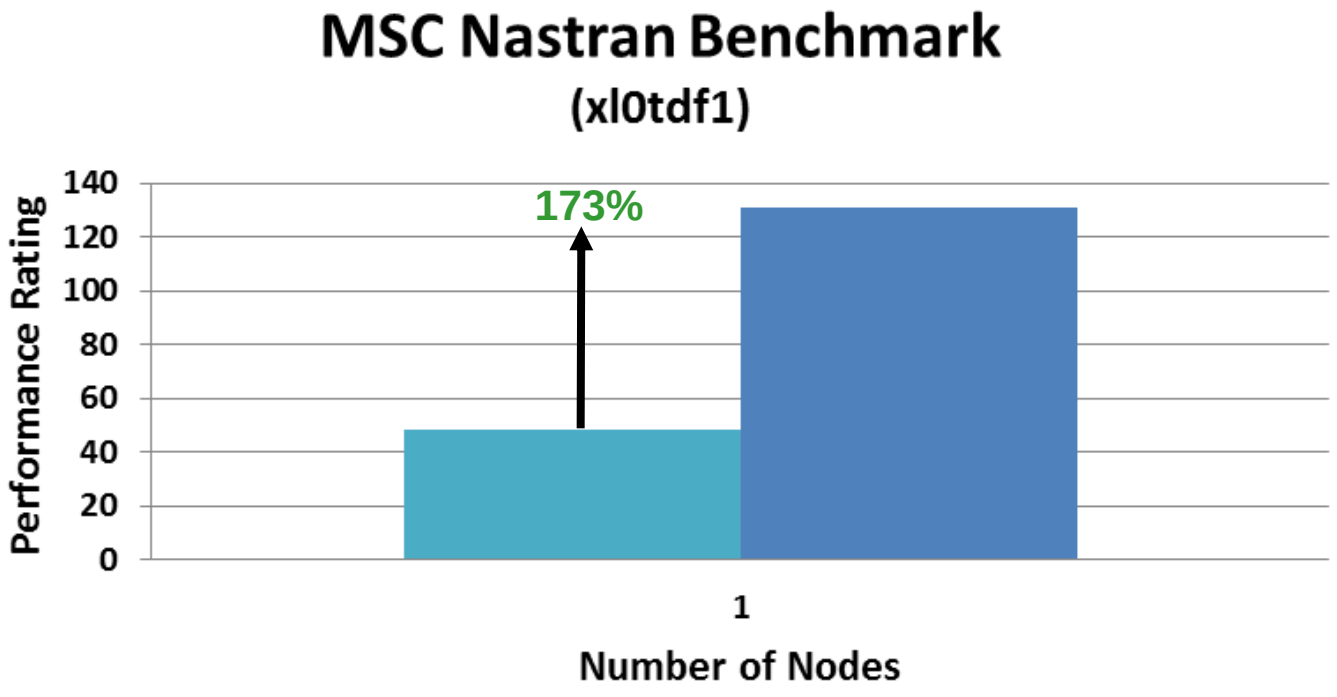
Higher is better

SMP=8, DMP=4/Node

Multiple Nodes

MSC Nastran Performance – Disk IO

- **Using 4 hard drives on RAID 0 versus 1 drives deliver higher performance**
  - Up to 173% higher performance with 4 HDDs on RAID versus 1HDD
  - Reflects that Nastran performance depends heavily on disk IO



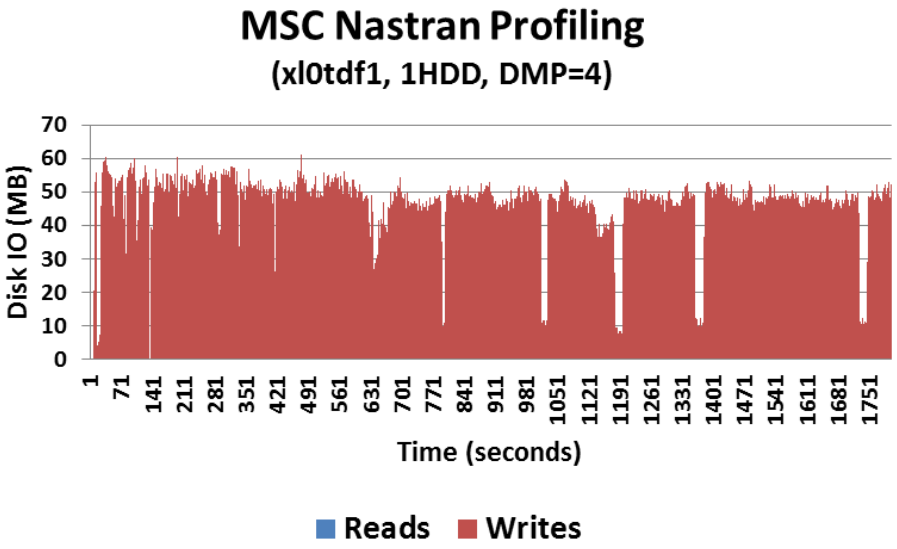
SMP=8, DMP=4

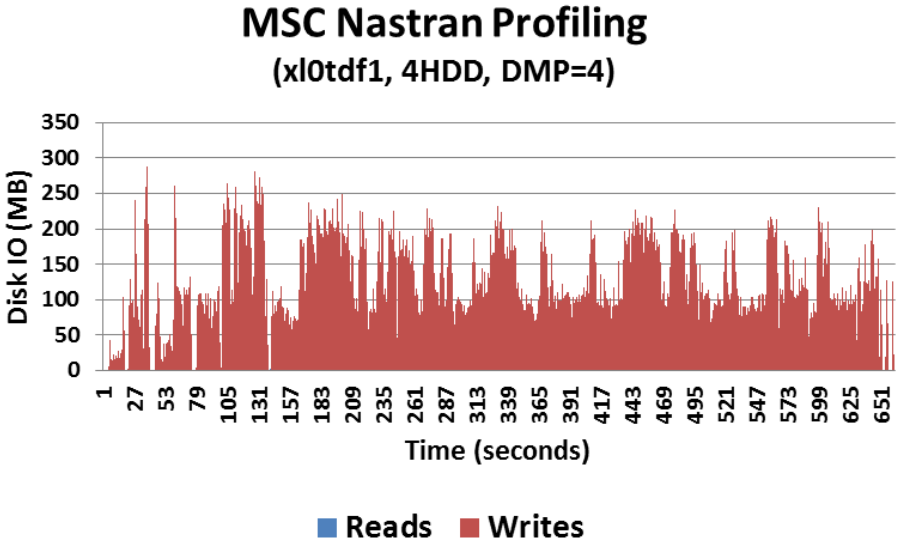
Single Node

Higher is better

MSC Nastran Profiling – Disk IO

- **IO Profiling shows benefits of having more hard drives per system**
  - Majority of the disk IO activities are write access
  - Having more HDDs can improve on IO load and reduces the overall runtime



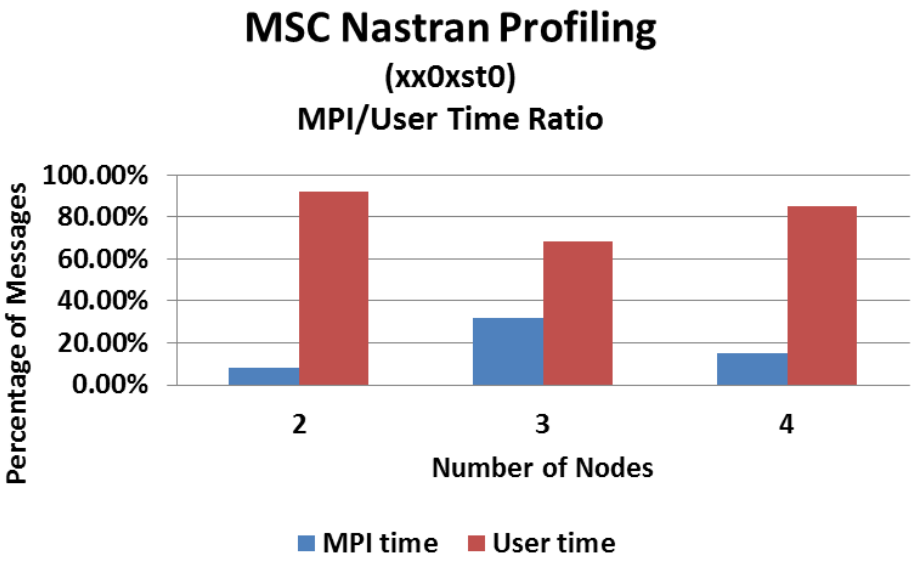


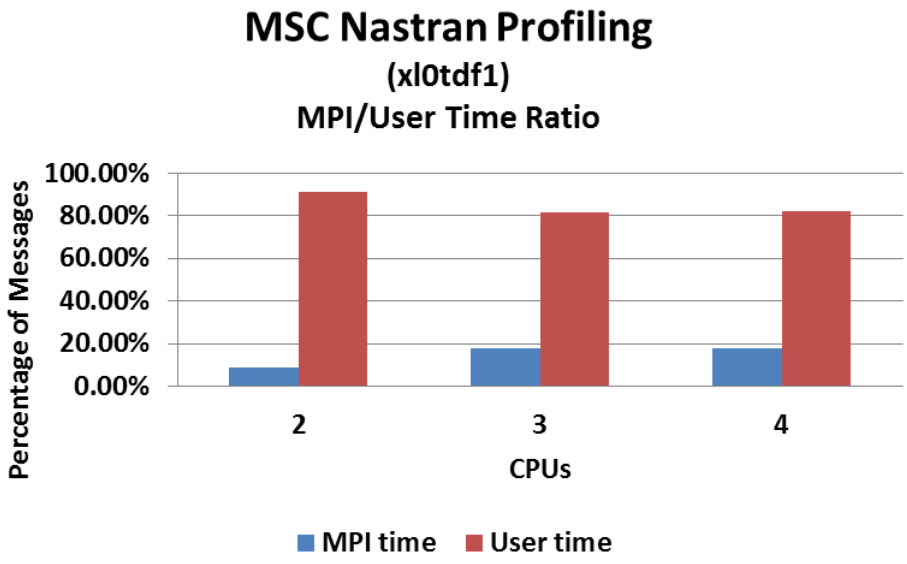
SMP=8, DMP=4

Single Node

MSC Nastran Profiling – MPI/User Time Ratio

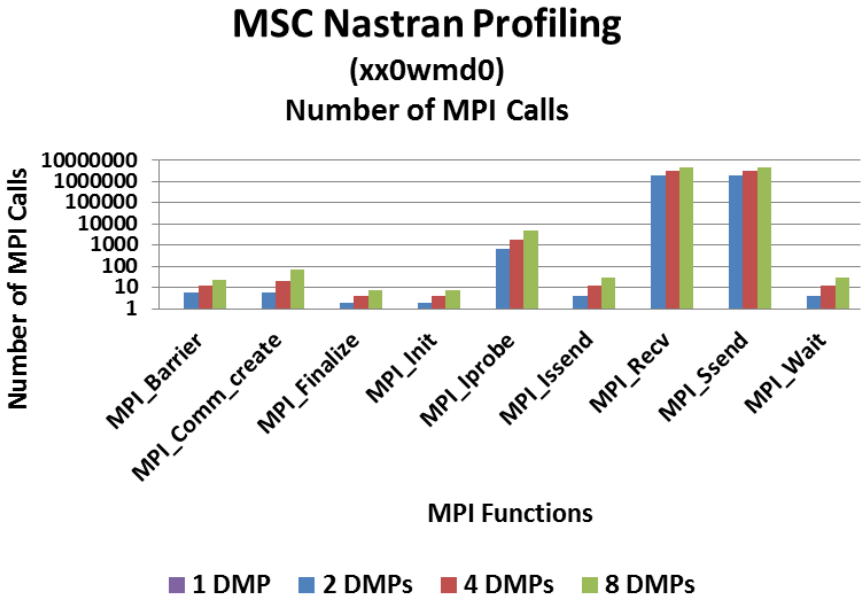
- **Both datasets demonstartes that more time spent on computation**
  - Communication time is not as significant
  - Reflects that MSC.Nastran is a compute-intensive application

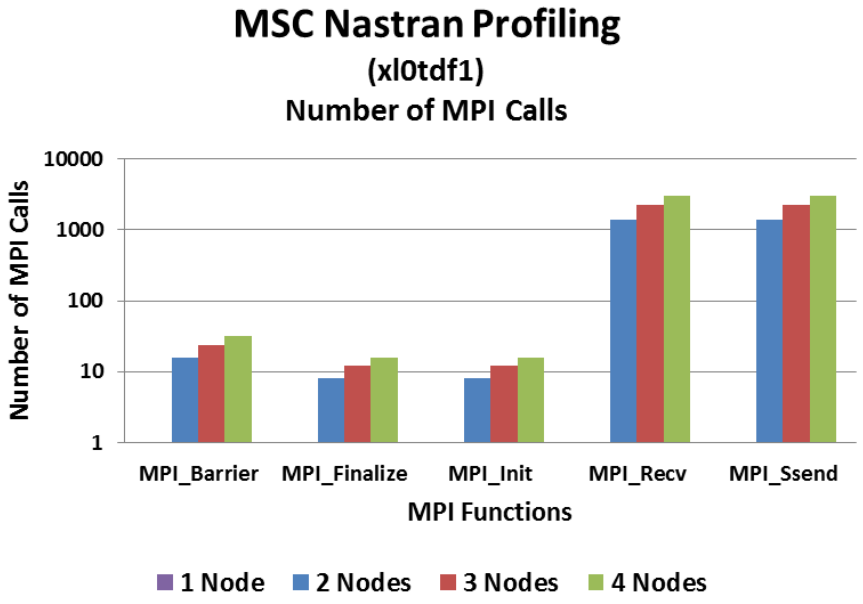




MSC Nastran Profiling – Number of MPI Calls

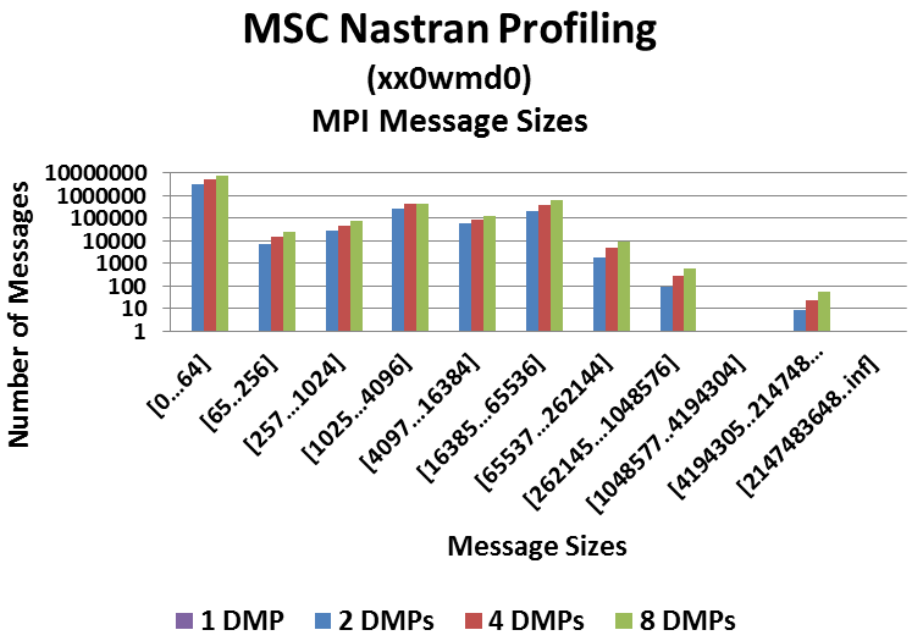
- **MPI_Ssend and MPI_Recv are almost used exclusively**
  - MPI_Ssend is a blocking synchronized send
  - Each of these MPI functions is accounted for nearly half of all MPI functions
  - Only point-to-point communications, and no MPI collectives, are used
- **Diverse views between xx0yst0 and xl0tdf1**
  - Significant MPI data communication for the xx0wmd0 (hence large # of Ssend/Recv)
  - The xx0yst0 is network bound and requires good network bandwidth
  - The xl0tdf1 has some data communication but small compare to xx0wmd0

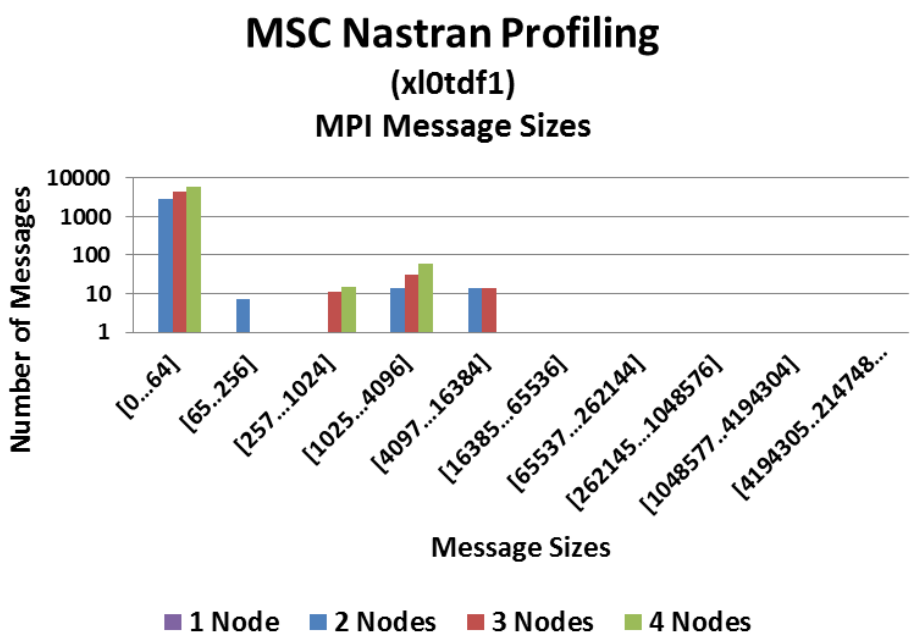




MSC Nastran Profiling – MPI Message Size

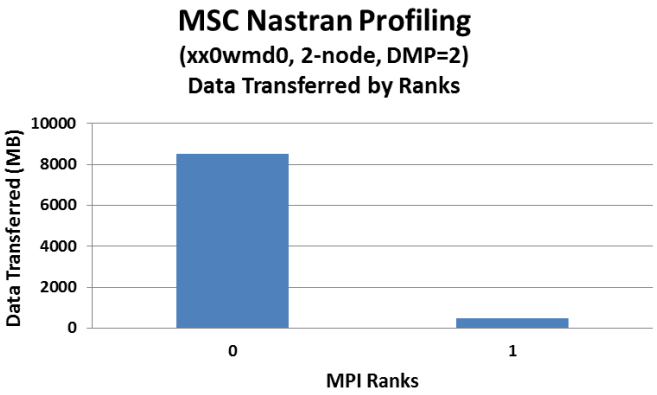
- **Majority of MPI messages are small messages**
  - Large percentage of messages falls in the range between 0 and 64 bytes
  - Small message sizes are typically used for synchronization
- **Depends on the dataset, large messages are also seen**
  - Some messages between 4MB and 2GB range.
  - Large message sizes are typically used for data communication (Send/Recv)
  - Each of the large messages are at around 180MB

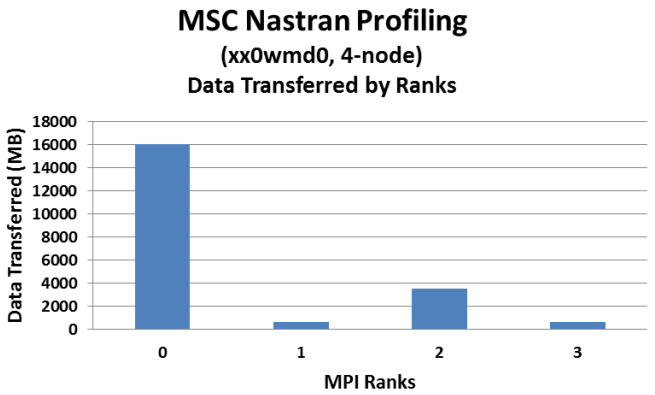


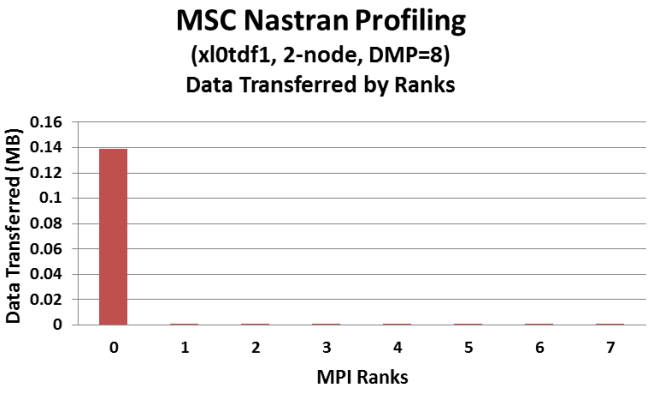


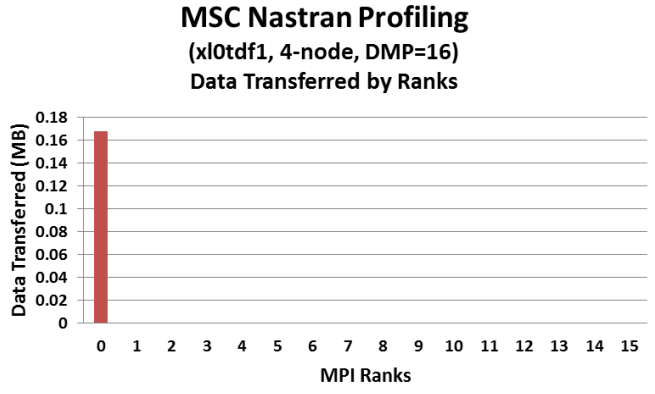
MSC Nastran Profiling – MPI Data Transfer

- **Different communication patterns for different datasets**
  - Show larger amount of data distributed to even number of nodes with xl0tdf1 dataset
  - Shows little data distributions from first MPI process with the xx0wmd0 dataset



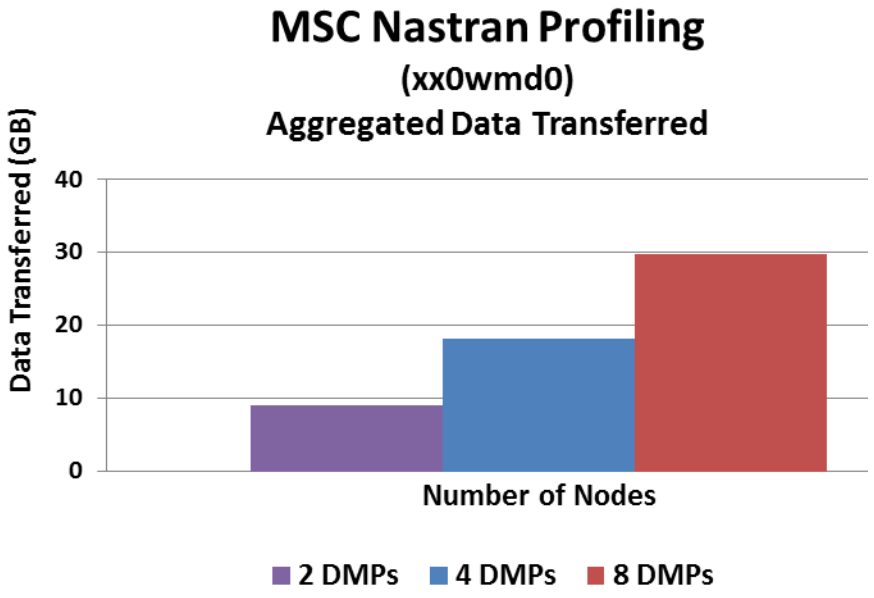


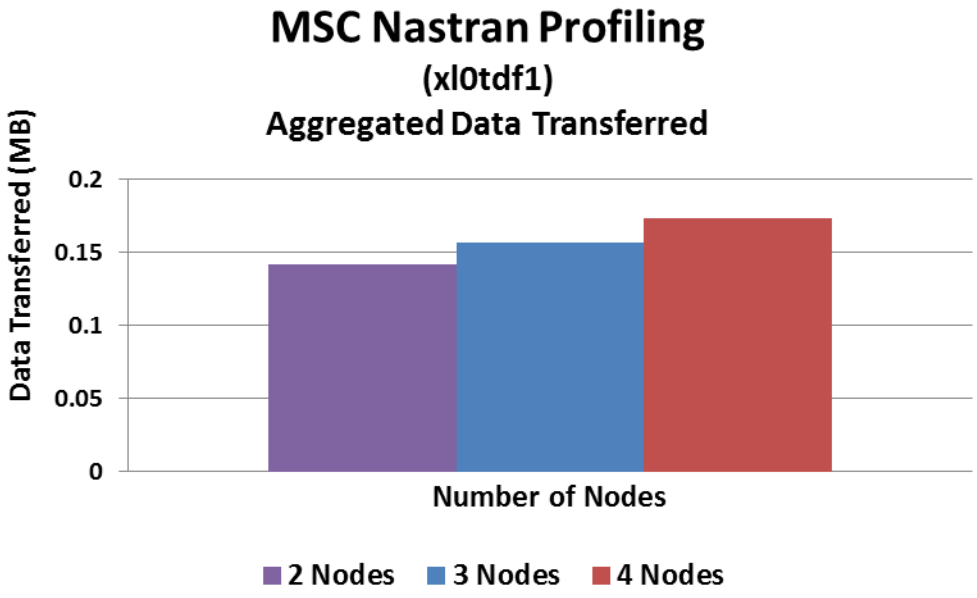




MSC Nastran Profiling – Aggregated Transfer

- **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**
  - The xx0wmd0 requires large network throughput
  - InfiniBand QDR is the best network interconnect that can provide high network bandwidth





MSC Nastran – Summary

- **MSC.Nastran 2012.2 demonstrates huge performance improvement**
  - Up to 80% performance gain compared to 2010.1.3 version on same system
- **Latest AMD CPU and Nastran 2012.2 enables faster runtime over prior generation**
  - Up to 118% of improvement over prior generation of CPU and software
    - Magny-Cours (prior): AMD Opteron 6174 @ 2.2GHz, MSC Nastran 2010.1.3
    - Interlagos (current): AMD Opteron 6276 @ 2.3GHz, MSC Nastran 2012.2
- **MSC.Nastran demonstrates large disk IO and CPU utilization**
  - MSC.Nastran performance jumped 173% by using 4 HDDs on RAID0 versus 1HDD
- **Networking:**
  - InfiniBand QDR allows Nastran to maintain scalability and distribute workload to systems
  - InfiniBand QDR Provides 41% better performance than 1GbE and 23% better than 10GbE
- **MPI:**
  - HP-MPI shows better performance than Intel MPI by 22% at 4 nodes
  - The stock Open MPI from MSC Nastran does not support InfiniBand
  - Only MPI point-to-point communications, and no MPI collectives, are used

Thank You

HPC Advisory Council