MSC Nastran
Performance Benchmark and Profiling

May 2011
The following research was performed under the HPC Advisory Council activities

- Participating vendors: AMD, Dell, Mellanox, MSC
- Compute resource - HPC Advisory Council Cluster Center

For more info please refer to

- [http://www.amd.com](http://www.amd.com)
- [http://www.dell.com/hpc](http://www.dell.com/hpc)
- [http://www.mellanox.com](http://www.mellanox.com)
- [http://www.mscsoftware.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™ R815 11-node (528-core) cluster**
- **AMD™ Opteron™ 6174 (code name “Magny-Cours”) 12-cores @ 2.2 GHz CPUs**
- **4 CPU sockets per server node**
- **Mellanox ConnectX-2 VPI adapters for 40Gb/s QDR InfiniBand and 10Gb/s Ethernet**
- **Mellanox MTS3600Q 36-Port 40Gb/s QDR InfiniBand switch**
- **Fulcrum based 10Gb/s Ethernet switch**
- **Memory: 128GB memory per node DDR3 1333MHz**
- **OS: RHEL 5.5, MLNX-OFED 1.5.2 InfiniBand SW stack**
- **Storage: 3x 15K 6Gbps 300GB on RAID 5**
- **MPI: HP MPI 2.3, Open MPI 1.2.2 & 1.2.9**
- **Application: MD Nastran / MSC Nastran version 2010.1.3**
- **Benchmark workload: MD Nastran R3 Benchmarks**
Dell™ PowerEdge™ R815 11-node cluster

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

- **New 11-node 528 core cluster - featuring Dell PowerEdge™ R815 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™ 6100 series platform and Mellanox ConnectX InfiniBand on Dell HPC

Solutions provide the ultimate platform for speed and scale
• Dell PowerEdge R815 system delivers 4 socket performance in dense 2U form factor
• Up to 48 core/32DIMMs per server – 1008 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
**Input dataset: xx0wmd0**
- Car Body (Ndof 3,799,278, SOL103, Freq Response w/ (interior) acoustics, and ACMS)
- Memory: 2000MB, SCR Disk: 169GB, Total I/O 3600GB
- Require large scratch disk space

**Time reduced as more nodes is being utilized**
- Up to 76% in time saved by running on a 8-node cluster versus 1-node
- Over 4 times faster when running with 8-node than 1-node

---

**MSC Nastran Benchmark**

(xx0wmd0)

**Elapsed Time (s)**

- 0
- 5000
- 10000
- 15000
- 20000
- 25000

**Number of Nodes**

- 1
- 2
- 4
- 8

**SMP=2**

InfiniBand QDR

*Lower is better*
MSC Nastran Results - Performance

- **Input dataset:** xl0tdf1
  - Power Train (Ndof 529,257, SOL108, Direct Freq)
  - 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 11-node cluster versus 1-node
  - Almost 9 times faster than running on a single node

---

**MSC Nastran Benchmark**

(xl0tdf1)

![Graph showing time reduction with number of nodes]

**Lower is better**

InfiniBand QDR

SMP=1

Elapsed Time (s)

Number of Nodes

Legend:
- 1
- 2
- 4
- 8
- 11

89%
MSC Nastran Performance – MPI

- **HP MPI shows slightly higher performance on larger number of nodes**
  - While Open MPI runs better on smaller number of nodes
- **Modified the shipped Open MPI to allow InfiniBand support**
  - The openib BTL was not built with the Open MPI shipped with MSC Nastran
  - Processor binding is used to enhance performance with the MCA “mpi_paffinity_alone 1”

![MSC Nastran Benchmark (xl0tclf1)](chart)

- Jobs/Day
- Number of Nodes
- Open MPI
- HPMPi

*Higher is better*
**MSC Nastran Performance – Interconnects**

- **InfiniBand leads among the network interconnects as the cluster scales**
  - Up to 14% higher performance than 10GigE on xl0tdf1
  - Up to 42% higher performance than 1GigE on xl0tdf1

- **InfiniBand continue to scales while Ethernet performance drops off**
  - Seen less performance for Ethernet after 8 node

---

**MSC Nastran Benchmark (xl0tdf1)**

![Bar chart showing performance comparison]

- **Higher is better**
- **48 Cores/Node**
• Increasing CPU core frequency enables higher job efficiency
  – An increase of 22-25% of higher job performance between 1800MHz vs 2200MHz
  – An increase of 46-52% of higher job performance between 1400MHz vs 2200MHz

• The increase in performance gain exceeds the increase CPU frequencies
  – CPU bound application can see higher benefit of using CPU with higher frequencies
MSC Nastran Profiling – MPI/User Time Ratio

- **Different communication patterns with different datasets**
  - The xx0wmd0 is heavy on data communication while xl0tdf1 is compute-bound
  - The xx0wmd0 spends time in network communication as the cluster scales
  - The xl0tdf1 spends its time almost strictly on computation

![MSC Nastran Profiling](image)
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 xx0wmd0 and xl0tdf1**
  - Significant MPI data communication for the xx0wmd0 (hence large # of Ssend/Recv)
  - The xx0wmd0 is network bound and requires good network bandwidth
  - The xl0tdf1 has some data communication but small compare to xx0wmd0

![MSC Nastran Profiling](image)
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 (xxwmd0)

**MPI Message Sizes**

MSC Nastran Profiling (x0tldf1)

**MPI Message Sizes**
MSC Nastran Profiling – Time Spent by MPI

- **MPI_Recv** is the biggest time consumer for data communicative dataset
  - Xx0wmd2 shows 88% of time in **MPI_Recv**
- **MPI_Ssend** consumes more in x10tdf1 but being overtaken by **MPI_Recv** later
MSC Nastran Profiling – MPI Data Transfer

- Data transferred to each process mainly from the first MPI rank
- 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 shows large CPU utilization and also on the network**
  - It can achieve higher performance by scaling out
  - Take advantage by clustering with more computation resources with InfiniBand QDR

- **MPI**
  - HP-MPI scales better Open MPI
  - Only MPI point-to-point communications, and no MPI collectives, are used

- **Data distribution**
  - First MPI rank responsible for data distribution
  - Majority of messages are small messages between 0 and 64 bytes

- **Networking:**
  - InfiniBand QDR allows best scaling on MSC Nastran

- **CPU:**
  - Shows gains in job productivity by using CPU with higher frequencies
Thank You
HPC Advisory Council