Performance, Tuning and Scalability of MSC Nastran Explicit Nonlinear (SOL 700) over High Performance Interconnects

2013 User Conference

Presented By: Pak Lui, HPC Advisory Council

May 7, 2013
Agenda

- Introduction to HPC Advisory Council
- Benchmark Configuration
- Performance Benchmarking
- MPI Profiling
- Summary
Introduction
HPC Advisory Council
The HPC Advisory Council

- World-wide HPC organization (360+ members)
- Bridges the gap between HPC usage and its full potential
- Provides best practices and a support/development center
- Explores future technologies and future developments
- Working Groups – HPC|Cloud, HPC|Scale, HPC|GPU, HPC|Storage
- Leading edge solutions and technology demonstrations
HPC Council Board

- **HPC Advisory Council Chairman**
  Gilad Shainer - gilad@hpcadvisorycouncil.com

- **HPC Advisory Council Media Relations and Events Director**
  Brian Sparks - brian@hpcadvisorycouncil.com

- **HPC Advisory Council China Events Manager**
  Blade Meng - blade@hpcadvisorycouncil.com

- **Director of the HPC Advisory Council, Asia**
  Tong Liu - tong@hpcadvisorycouncil.com

- **HPC Advisory Council HPC|Works Special Interest Group Chair**
  and Cluster Center Manager
  Pak Lui - pak@hpcadvisorycouncil.com

- **HPC Advisory Council Director of Educational Outreach**
  Scot Schultz – scot@hpcadvisorycouncil.com

- **HPC Advisory Council Programming Advisor**
  Tarick Bedeir - Tarick@hpcadvisorycouncil.com

- **HPC Advisory Council Workshop Program Director**
  Eric Lantz – eric@hpcadvisorycouncil.com

- **HPC Advisory Council Research Steering Committee Director**
  Cydney Stevens - cydney@hpcadvisorycouncil.com

- **HPC Advisory Council HPC|Scale Special Interest Group Chair**
  Richard Graham – richard@hpcadvisorycouncil.com

- **HPC Advisory Council HPC|Cloud Sub-group Chair**
  William Lu – william@hpcadvisorycouncil.com

- **HPC Advisory Council HPC|GPU Special Interest Group Chair**
  Sadaf Alam – sadaf@hpcadvisorycouncil.com

- **HPC Advisory Council India Outreach**
  Goldi Misra – goldi@hpcadvisorycouncil.com

- **Director of the HPC Advisory Council Switzerland Center of Excellence and HPC|Storage Special Interest Group Chair**
  Hussein Harake – hussein@hpcadvisorycouncil.com
Special Interest Subgroups Missions

- **HPC|Scale**
  - To explore usage of commodity HPC as a replacement for multi-million dollar mainframes and proprietary based supercomputers with networks and clusters of microcomputers acting in unison to deliver high-end computing services.

- **HPC|Cloud**
  - To explore usage of HPC components as part of the creation of external/public/internal/private cloud computing environments.

- **HPC|Works**
  - To provide best practices for building balanced and scalable HPC systems, performance tuning and application guidelines.

- **HPC|Storage**
  - To demonstrate how to build high-performance storage solutions and their affect on application performance and productivity. One of the main interests of the HPC|Storage subgroup is to explore Lustre based solutions, and to expose more users to the potential of Lustre over high-speed networks.

- **HPC|GPU**
  - To explore usage models of GPU components as part of next generation compute environments and potential optimizations for GPU based computing.

- **HPC|FSI**
  - To explore the usage of high-performance computing solutions for low latency trading, more productive simulations (such as Monte Carlo) and overall more efficient financial services.
HPC Advisory Council

- HPC Advisory Council (HPCAC)
  - Application best practices, case studies (Over 150)
  - Benchmarking center with remote access for users
  - World-wide workshops
  - Value add for your customers to stay up to date and in tune to HPC market

- 2013 Workshops
  - USA (Stanford University) – January 2013
  - Switzerland – March 2013
  - Germany (ISC’13) – June 2013
  - Spain – Sep 2013
  - China (HPC China) – Oct 2013

- For more information
  - [www.hpcadvisorycouncil.com](http://www.hpcadvisorycouncil.com), info@hpcadvisorycouncil.com
ISC’13 – Student Cluster Challenge

- University-based teams to compete and demonstrate the incredible capabilities of state-of-the-art HPC systems and applications on the ISC’13 show-floor
- The Student Cluster Challenge is designed to introduce the next generation of students to the high performance computing world and community
ISC’13 – Student Cluster Challenge

[Images of student groups and clusters]
Benchmark Configuration
MSC Nastran Explicit Nonlinear (SOL 700)

- **MSC Nastran Explicit Nonlinear (SOL 700)**
  - Explicit finite element analysis (FEA) solution
  - For simulating short-duration events like impact and crash
  - To analyze the complex nonlinear behavior that structures undergo during these events.
  - Performs explicit Transient Dynamic Solution for Crash, Impact and Fluid-Structure Interaction Studies for improved product safety and reduced warranty costs
  - Shown as “SOL 700” in short for this presentation

- **Explicit Nonlinear (SOL700) enables users to study structural integrity of designs**
  - To ensure that final products stand a better chance of meeting customer safety, reliability, and regulatory requirements
  - Uses a unique coupling feature that enables integrated analysis of structural components with fluids and highly deformed materials in one continuous simulation
Objectives

• The following was done to provide best practices
  – MSC Nastran Explicit Nonlinear (SOL 700) performance benchmarking
  – Interconnect performance comparisons
  – Understanding MSC Nastran Explicit Nonlinear communication patterns
  – Ways to increase MSC Nastran Explicit Nonlinear productivity
  – Comparisons between different software versions

• The presented results will demonstrate
  – The scalability of the compute environment
  – The capability of MSC Nastran Explicit Nonlinear to achieve scalable productivity
  – Considerations for performance optimizations
**Test Cluster Configuration**

- **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 1.5.3 InfiniBand SW stack

- **Intel Cluster Ready certified cluster**

- **Mellanox ConnectX-2 InfiniBand adapters**

- **Mellanox M3601Q 32-port 40Gb/s InfiniBand switch**

- **MPI: HP MPI 2.3**

- **Application:**
  - MSC Nastran (Versions 2012.1.2 and 2013) Explicit Nonlinear (SOL700)
About Dell PowerEdge Servers

- **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
About Intel® Cluster Ready

• 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
Performance Benchmarking
SOL 700 Performance – Background

- All of the test models shown in the subsequent slides are:
  - Based on actual customer data
  - Test cases are referred to by the model numbers
  - Consisted of mainly in the fluid domain
  - Spent 10% or less of the total elapsed is structural elements
  - Models #2, #12, #20 and #21 are 100% fluid domain
  - Refer to the Release Notes for results of a list of models
SOL 700 Performance – Interconnects

• **InfiniBand performs better than Ethernet as the cluster scales**
  – Up to 64% higher performance than 1GbE on model #6

• **InfiniBand continue to scales while Ethernet performance drops off**
  – Performance improvement starts beyond 4 processes (or a single node)
SOL 700 Performance – Process Allocation

- Spreading MPI processes to more nodes shows better performance
  - Better to spread the MPI processes to more nodes

![Graph showing SOL 700 Performance](image)

**Lower is better**
SOL 700 Performance – Software Versions

- Improvement in Nastran Explicit Nonlinear (SOL 700) seen between version 2012.2 and 2013
  - Up to 64% improvement seen with test case #20

![Graph showing Nastran Explicit Nonlinear (SOL 700) performance between version 2012.2 and 2013. Lower is better.]
SOL 700 Performance – Software Versions

- For complete list of jobs tested
  - Actual time can be found in release notes

**Nastran Explicit Nonlinear (SOL 700)**
(Serial Tests)

Lower is better
SOL 700 Scalability – Improvement in v2013

- Scalability improvement in SOL 700 over previous version
  - Up to 130% improvement seen with model #14 at 64 MPI processes

- Scalability differences depend on models being tested
  - Some of models may not show good scaling because the size is too small
  - The actual runtime can be found in release notes for all of the models being tested
Scalability differences depend on models being tested
  - The actual runtime can be found in release notes for all of the models being tested

**Nastran Explicit Nonlinear (SOL 700)**
(MPP Tests with 64 MPI Processes)

- **Lower is better**
- **QDR InfiniBand**

2013 Users Conference
A NEW ERA BEGINS.
SOL 700 Results – Fluid Solver Scalability

- The fluid part consists of two algorithms that are DMP parallelized:
  - Fluid solver
  - FSI or Fluid Structure Interface (contact between fluid and structure)

- Good scalability seen on workloads that involves mainly of fluid solver
  - Model #20 has Fluid solver and no FSI, showing 33x speedup from serial run

![Nastran Explicit Nonlinear (SOL 700) (Model #20)](chart.png)

Lower is better
For workloads that involve fluid solver but with some % in FSI
- Both of the Fluid and FSI solvers are DMP parallelized
- Fluid part seems to scale better than FSI
SOL 700 Results – FSI Scalability

• For workloads that involves fluid solver, FSI and others
  – More time spent on FSI, and rest performs in serial

• Performance improvement for 64+ MPI tasks slows down
  – As the serial parts of the code are starting to become time dominant
  – Serial parts include initialization, output processing, and others

![Nastran Explicit Nonlinear (SOL 700) (Model #14)](image)

- Lower is better
- QDR InfiniBand

MSC Software Confidential
• **Relationship between FSI and Euler elements**
  – Ratio = (FSI / Euler Elements)
  – Higher ratio shows lower performance speedup
  – Speedup trend continues at large core counts when time spent on FSI

• **Conclusion**
  – Higher speedup expected when smaller ratio of time is spent on FSI
MPI Profiling
SOL 700 Profiling – MPI/User Time Ratio

• Computation time dominates Nastran Explicit Nonlinear (SOL 700)
  – The model #18 spends time in computation than in network communication

• New solver algorithms in version 2013 leads to better job efficiency
  – Reflects the computation and communication time reduction
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
- Some messages up to 256KB

Improvement in Version 2013 causes less data being transferred
- Efficiency in the code that reduces communication
SOL 700 Profiling – Time Spent by MPI

- MPI_Wait is the biggest time consumer for Version 2012.2
  - Shows 54% of time in MPI_Wait at 32 CPUs
  - MPI_Wait consumes time used for non-blocking communications

- MPI_Allreduce becomes the time consumer for Version 2013
SOL 700 Profiling – MPI Data Transfer

- Slightly less communication shown in version 2013 than in 2012.2
  - Reduction in data transferred from the maximum of ~20GB to ~18GB
**SOL 700 Profiling – Aggregated Transfer**

- **Aggregated data transfer is data transferred between ranks collectively**
  - The total data transfer increases as the cluster scales
- **Demonstrates importance of high throughput interconnects**
  - QDR InfiniBand is the best network interconnect that can provide high network bandwidth
Summary
Nastran Explicit Nonlinear (SOL 700) – Summary

- **Dramatic improvement** in Version 2013 compared over previous version
  - Up to 64% performance improvement in serial run with version 2013 versus 2012.2

- **Scalability improves in version 2013 over version 2012.2**
  - Up to 133% improvement seen with model #14 on a MPP run with 64 MPI tasks

- **Speedup improvement for workloads that involves fluid solver**
  - Delivers 33x speedup on 64 tasks job versus serial run for fluid models

- **InfiniBand performs better than Ethernet as job scales to more tasks**
  - Up to 64% higher performance than 1GbE on model #6
  - InfiniBand reduces network latency which improved MSC Nastran Explicit Nonlinear (SOL 700) MPP performance

- **Performance improvement for 64+ MPI tasks slows down**
  - As the serial parts of the code are starting to become time dominant
Thank you

• Special thanks to Walter Schrauwen of MSC Software
50
YEARS
OF INNOVATION
A NEW ERA BEGINS.