• 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
  – RADIOSS performance overview
  – Understanding RADIOSS communication patterns
  – Ways to increase RADIOSS productivity
  – MPI libraries comparisons

• For more info please refer to
  – http://www.altair.com
  – http://www.dell.com
  – http://www.intel.com
  – http://www.mellanox.com
Objectives

- **The following was done to provide best practices**
  - RADIOSS performance benchmarking
  - Interconnect performance comparisons
  - MPI performance comparison
  - Understanding RADIOSS communication patterns

- **The presented results will demonstrate**
  - The scalability of the compute environment to provide nearly linear application scalability
  - The capability of RADIOSS to achieve scalable productivity
RADIOSS by Altair

- **Altair® RADIOSS®**
  - Structural analysis solver for highly non-linear problems under dynamic loadings
  - Consists of features for:
    - multiphysics simulation and advanced materials such as composites
    - Highly differentiated for Scalability, Quality and Robustness
- **RADIOSS is used across all industry worldwide**
  - Improves crashworthiness, safety, and manufacturability of structural designs
- **RADIOSS has established itself as an industry standard**
  - for automotive crash and impact analysis for over 20 years
Test Cluster Configuration

• **Dell™ PowerEdge™ R720xd 32-node (512-core) “Jupiter” cluster**
  – Dual-Socket Eight-Core Intel E5-2680 @ 2.70 GHz CPUs (Static max Perf in BIOS)
  – Memory: 64GB memory, DDR3 1600 MHz
  – OS: RHEL 6.2, OFED 1.5.3 InfiniBand SW stack
  – Hard Drives: 24x 250GB 7.2 RPM SATA 2.5” on RAID 0

• **Mellanox ConnectX-3 FDR InfiniBand VPI adapters**

• **Mellanox SwitchX SX6036 InfiniBand VPI switch**

• **Intel Cluster Ready certified cluster**

• **MPI: Intel MPI 4.1.0**

• **Application: Altair RADIOSS 12.0**

• **Benchmark datasets:**
  – Neon benchmarks: 1 million elements (8ms, SP)
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

• RADIOSS is Intel Cluster Ready
PowerEdge R720xd
Massive flexibility for data intensive operations

• **Performance and efficiency**
  – Intelligent hardware-driven systems management with extensive power management features
  – Innovative tools including automation for parts replacement and lifecycle manageability
  – Broad choice of networking technologies from GbE to IB
  – Built in redundancy with hot plug and swappable PSU, HDDs and fans

• **Benefits**
  – Designed for performance workloads
    • from big data analytics, distributed storage or distributed computing where local storage is key to classic HPC and large scale hosting environments
    • High performance scale-out compute and low cost dense storage in one package

• **Hardware Capabilities**
  – Flexible compute platform with dense storage capacity
    • 2S/2U server, 6 PCIe slots
  – Large memory footprint (Up to 768GB / 24 DIMMs)
  – High I/O performance and optional storage configurations
    • HDD options: 12 x 3.5” - or - 24 x 2.5 + 2x 2.5 HDDs in rear of server
    • Up to 26 HDDs with 2 hot plug drives in rear of server for boot or scratch
**FDR InfiniBand provides better scalability performance than Ethernet**
- 743% better performance than 1GbE at 8 nodes
- 214% better performance than 10GbE at 16 nodes
- 1GbE does not scale beyond 4 nodes with pure MPI

**RADIOSS Benchmark**
*(NEON1M11, MPP)*

<table>
<thead>
<tr>
<th>Number of Nodes</th>
<th>1GbE</th>
<th>10GbE</th>
<th>FDR InfiniBand</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>4</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>8</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>16</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

*Higher is better*

**Intel MPI**

16 Processes/Node
• Enabling SRQ allows to run at larger MPI process counts
  – Up to 148% higher performance at 16 nodes than without SRQ being used
  – Running with SRQ reduces the memory footprint needed for communications
  – No other optimization flags are used between the 2 cases
• **Running in Hybrid MPP (HMPP) mode can enhance RADIOSS scalability**
  – In normal MPP mode, only MPI processes are launched
  – In Hybrid MPP mode, multiple threads spawned for every MPI process launched
  – Threads shown represents the number of threads spawned by each MPI process

• **Hybrid mode improves scalability at higher core counts**
  – Hybrid mode starts to improve runtime when running beyond 8 nodes (128 cores)

---

**RADIOSS Benchmark**
(NEON1M11, Hybrid)

*Lower is better*
Enabling Hybrid MPP mode unlocks the RADIOSS scalability
- At larger scale, productivity improves as more threads involves
- As more threads involved, amount of communications by processes are reduced
- At 32 nodes (or 512 cores), the best configuration is 2 PPN with 8 threads each

The following environment setting and tuned flags are used:
- Intel MPI flags: -genv I_MPI_PIN_DOMAIN auto -genv OMP_NUM_THREADS $OMP_NUM_THREADS -genv I_MPI_ADJUST_BCAST 1 -genv I_MPI_ADJUST_REDUCE 2 -genv KMP_AFFINITY verbose,compact -genv KMP_STACKSIZE 400m
- User environment: “ulimit -s unlimited”

RADIOSS Benchmark
(NEON1M11, Hybrid)

Higher is better

Intel MPI
FDR InfiniBand
RADIOSS Performance – Interconnect (HMPP)

- **FDR InfiniBand provides better scalability performance than Ethernet**
  - 290% better performance than 1GbE at 16 nodes
  - 93% better performance than 10GbE at 16 nodes

**RADIOSS Benchmark**
(NEON1M11, Hybrid)

- **Higher is better**
- **Intel MPI**
- **8 Threads/MPI proc**
Intel E5-2680 (Sandy Bridge) cluster outperforms prior generations
- Performs up to 50% better than X5670 cluster at 16 nodes

System components used:
- Jupiter: 2-socket Intel E5-2680 @ 2.7GHz, 1600MHz DIMMs, FDR IB, 24 HDDs
- Janus: 2-socket Intel X5670 @ 2.93GHz, 1333MHz DIMMs, QDR IB, 1 HDD

**RADIOSS Benchmark**
(NEON1M11, Hybrid)

<table>
<thead>
<tr>
<th>Number of Nodes</th>
<th>Jobs/Day</th>
<th>Janus</th>
<th>Jupiter</th>
<th>Intel MPI</th>
<th>8 Threads/MPI proc</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td></td>
<td>0</td>
<td>0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>38%</td>
<td>0</td>
<td>0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>38%</td>
<td>0</td>
<td>0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>50%</td>
<td>0</td>
<td>0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>16</td>
<td>47%</td>
<td>0</td>
<td>0</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Higher is better**
“OFA provider” in Intel MPI delivers better scalability performance

- Up to 15% better application performance than DAPL provider at 32 nodes
**RADIOSS Profiling – MPI/User Time Ratio**

- **MPI communication time grows rapidly between 8 to 16 nodes**
  - Reflects that more time spent on computation than communications
  - Dramatic increase indicates the Neon input file becomes too “small” to scale beyond 8 nodes

---

**RADIOSS Profiling**
(NEON1M11)

**MPI/User Time Ratio**

![Graph showing MPI/User Time Ratio](image)

- **Pure MPP**
  - 16 Processes/Node
• **RADIOSS utilizes non-blocking communications in most data transfers**
  - MPI_Wait, MPI_Waitany, MPI_Irecv and MPI_Isend are almost used exclusively
  - MPI_Wait(37%), MPI_Waitany(26%) and MPI_Isend/Irecv (18% each) at 32 nodes
• The most time MPI consuming calls is MPI_Waitany() and MPI_Wait()
  – MPI_Waitany(43%), MPI_Wait(22%), MPI_Bcast(18%), MPI_Recv(12%)
• Time spent on MPI_Wait and Waitany are for MPI_Isend/Irecv
  – Wait time are accounted for time spent on pending non-blocking transfers
RADIOSS Profiling – MPI Message Sizes

- **RADIOSS uses small MPI message sizes**
  - Most message sizes are between 0B to 64B, and 257B to 1KB
Uneven distribution of data transfers between the MPI processes
- Non-blocking data communications between processes are involved
RADIOSS Profiling – Aggregated Transfer

- **Aggregated data transfer refers to:**
  - Total amount of data being transferred in the network between all MPI ranks collectively

- **Substantially larger data transfer takes place in RADIOSS**
  - As node count doubles, amount of data transferred is more than double

---

**RADIOSS Profiling**
(NEON1M11, SP, MPI)
Aggregated Data Transferred

![Graph showing aggregated data transferred with different node counts.](image)
RADIOSS – Summary

- **RADIOSS is designed to perform at large scale HPC environment**
  - Shows excellent scalability over 512 cores (32 nodes) and beyond with Hybrid MPP
  - Hybrid MPP version enhanced RADIOSS scalability
    - At 32 nodes, the best Hybrid MPP configuration is 2 MPI processes per socket with 8 threads each
- **Intel Xeon E5-2600 series and FDR InfiniBand enable RADIOSS to scale**
  - The E5-2680 cluster outperforms X5670 cluster by 50% at 16 nodes
- **Network and MPI comparisons**
  - For MPP version, FDR InfiniBand provides better scalability performance than Ethernet
    - Over 7.4 times better performance than 1GbE at 8 nodes
    - Over 2.1 times better performance than 10GbE at 16 nodes
  - For Hybrid MPP, FDR InfiniBand provides better scalability performance than Ethernet
    - Over 2.9x better performance than 1GbE at 32 nodes
    - Up to 93% better performance than 10GbE at 32 nodes
  - OFA provider in Intel MPI delivers better application performance
    - Up to 15% better scalability performance than DAPL at 32 nodes
  - Enabling SRQ allows to run at larger MPI process counts
    - Up to 147% higher performance at 16 nodes (256 MPI processes) than without SRQ being used
Thank You
HPC Advisory Council