HPC ADVISORY COUNCIL


# Altair RADIOSS Performance Benchmark and Profiling

May 2013

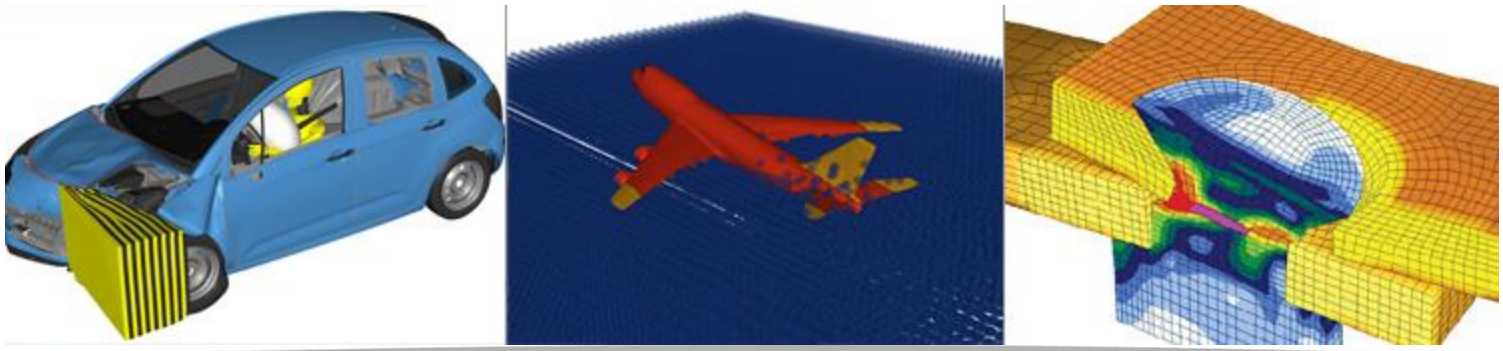
Altair AMD DELL Mellanox TECHNOLOGIES

# Note

- **The following research was performed under the HPC Advisory Council activities**
  - Participating vendors: Altair, 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.altair.com

# 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



# Objectives

- **The following was done to provide best practices**
  - RADIOSS performance benchmarking
  - Interconnect performance comparisons
  - CPU performance
  - Understanding RADIOSS communication patterns
  - Ways to increase RADIOSS productivity

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

# Test Cluster Configuration

- **Dell™ PowerEdge™ R815 11-node (704-core) "Vesta" cluster**
  - AMD™ Opteron™ 6380 (code name "Abu Dhabi") 16-cores @ 2.5 GHz CPUs
- **4 CPU sockets per server node**
- **Mellanox ConnectX-3 VPI adapters for 40Gb/s QDR InfiniBand and 40Gb/s Ethernet**
- **Mellanox SwitchX™ 6036 36-Port InfiniBand switch**
- **Memory: 128GB memory per node DDR3 1333MHz**
- **OS: RHEL 6.2 MLNX-OFED 1.5.3 InfiniBand SW stack**
- **MPI: Intel MPI 4.1**
- **Application: Altair RADIOSS version 12.0**
- **Benchmark workload:**
  - Neon benchmarks: 1 million elements (80ms, SP and DP)

# Dell™ PowerEdge™ R815 11-node cluster

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

- **New 11-node 704 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™ 6300 series platform and Mellanox ConnectX®-3 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 64 core/32DIMMs per server – 1344 core in 42U enclosure

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

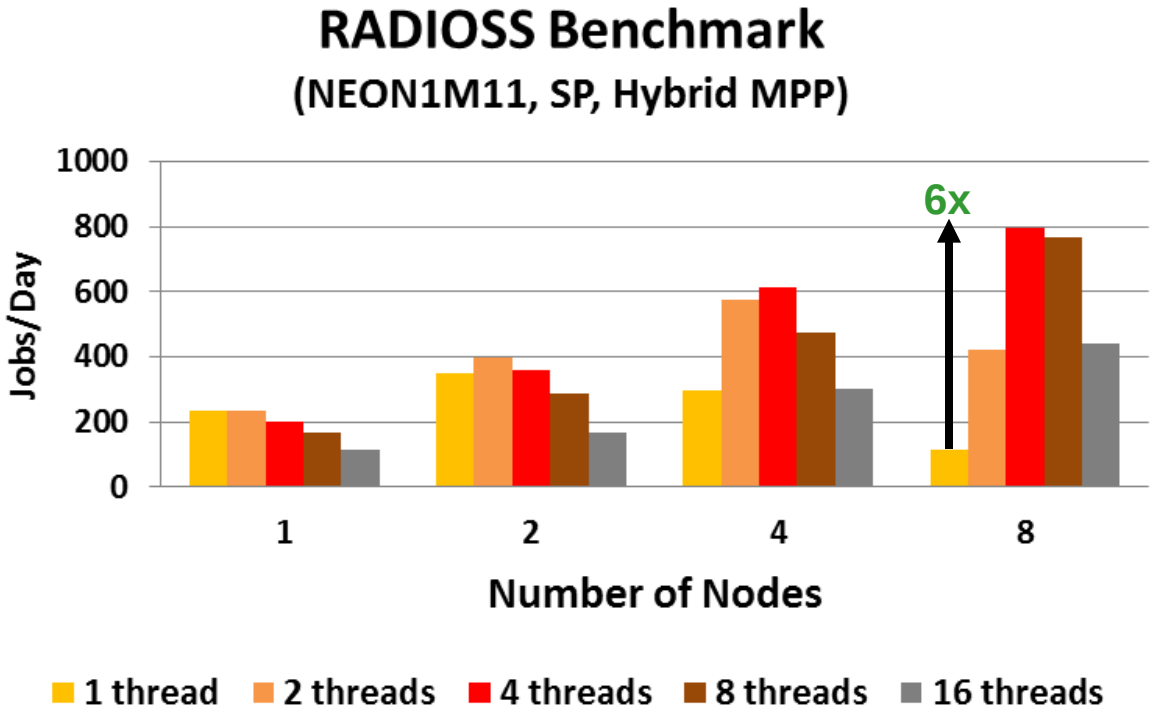
- 2x more memory and processing power in half of the space
- Energy optimized low flow fans, improved power supplies and dual SD modules

## Optimized for long-term capital and operating investment protection

- Platform longevity across 3 CPU generations (AMD™ Opteron™ 6100, 6200 & 6300 series)
- System expansion, component upgrades and feature releases

# RADIOSS Performance – MPP vs Hybrid MPP

- **RADIOSS offers both MPP and hybrid MPP modes for launching jobs**
  - In pure MPP mode, only MPI processes are launched, (e.g. 1 thread)
  - In Hybrid MPP mode, multiple threads spawned for every MPI process launched
  - Hybrid MPP mode reduces the MPI processes, which reduces data exchanges
- **Good improvement seen when using hybrid MPP at high core counts**
  - Up to 6 times higher productivity achieved at 8 nodes with single precision tests

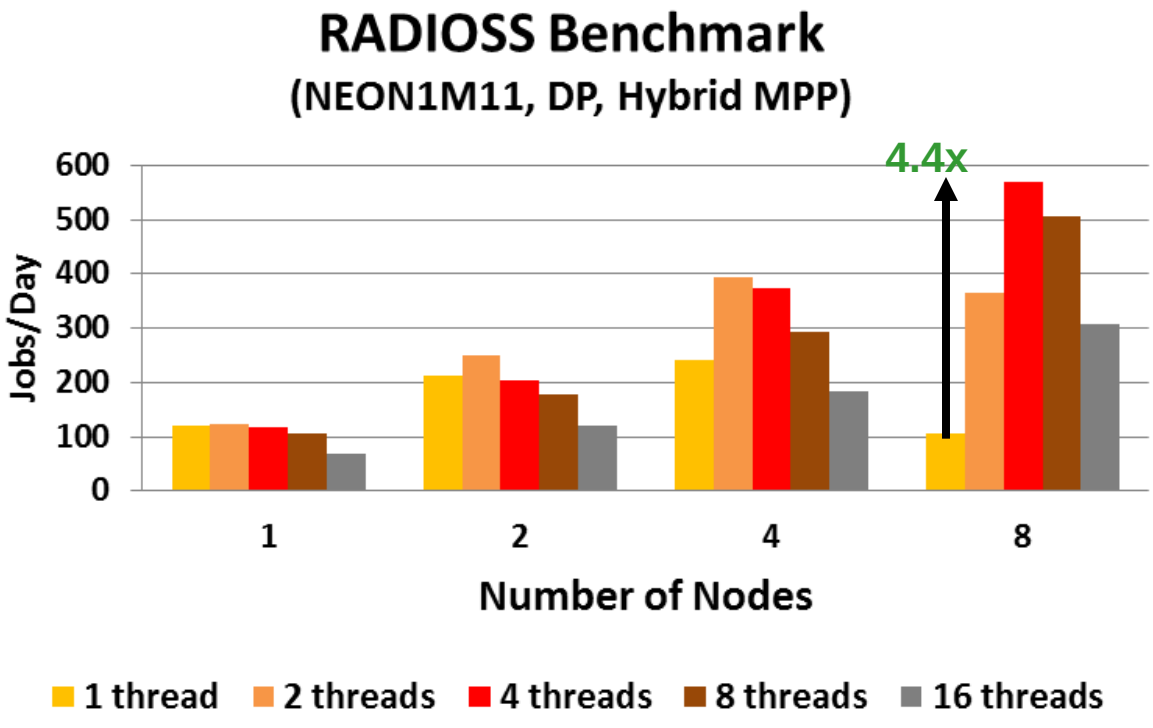



*Higher is better*

*64 Cores/Node*

# RADIOSS Performance – MPP vs Hybrid MPP

- **Similar improvement seen when using HMPP for double precision tests**
  - Up to 4.4 times higher productivity achieved at 8 nodes with double precision tests

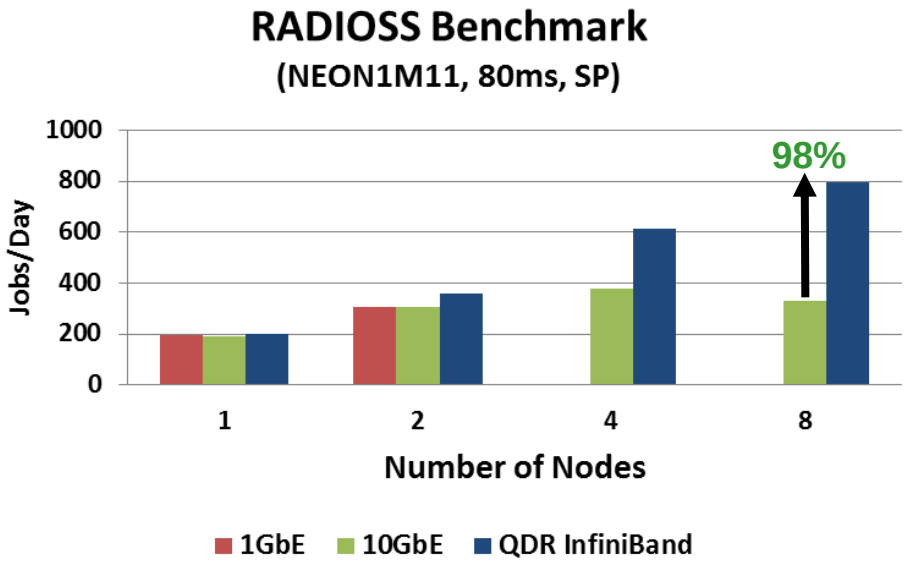


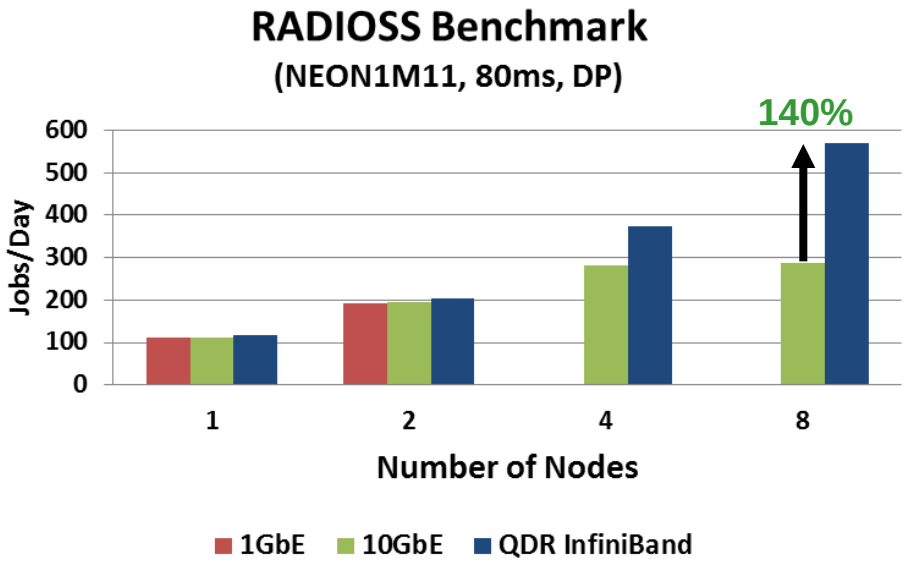

*Higher is better*

*64 Cores/Node*

# RADIOSS Performance – Interconnects

- **InfiniBand shows continuous gain as the cluster scales**
  - Up to 98% to 140% higher productivity compared to 10GbE at 8 nodes
- **Ethernet does not scale**
  - 1GbE performance drops after 2 nodes
  - 10GbE scalability declines from 4 nodes and beyond



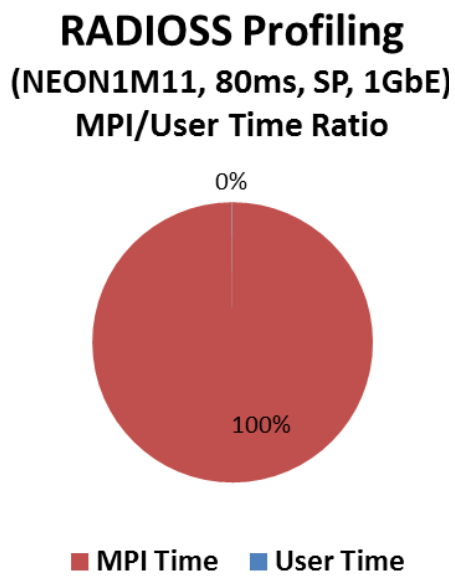



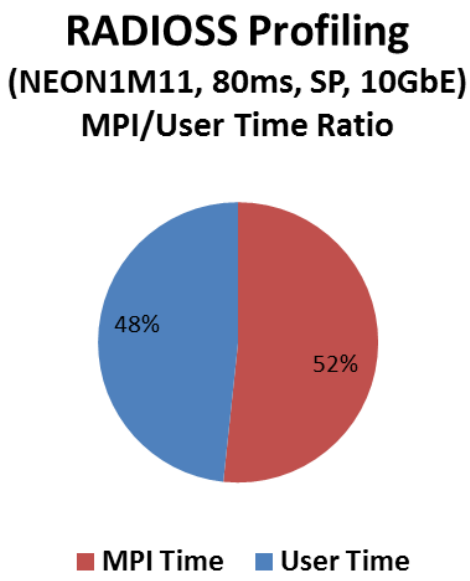

*Higher is better*

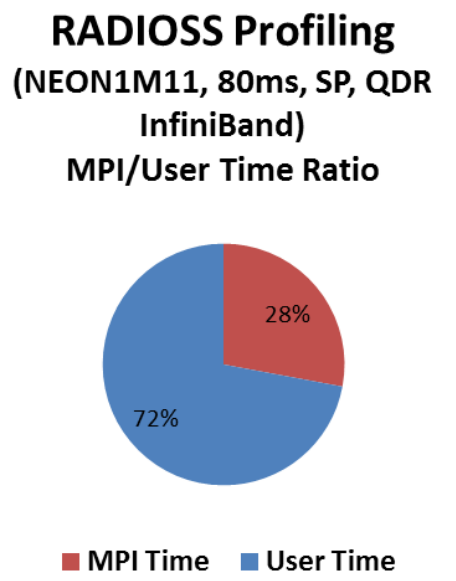
*4 Threads/Core*

# RADIOSS Profiling – MPI/User Time Ratio

- **InfiniBand reduces time spent on network communications**
  - InfiniBand takes around 28% of time for MPI communications
  - Ethernet can takes from 52% to almost 100% for MPI communications
  - Network infrastructure is essential for RADIOSS to run at scale



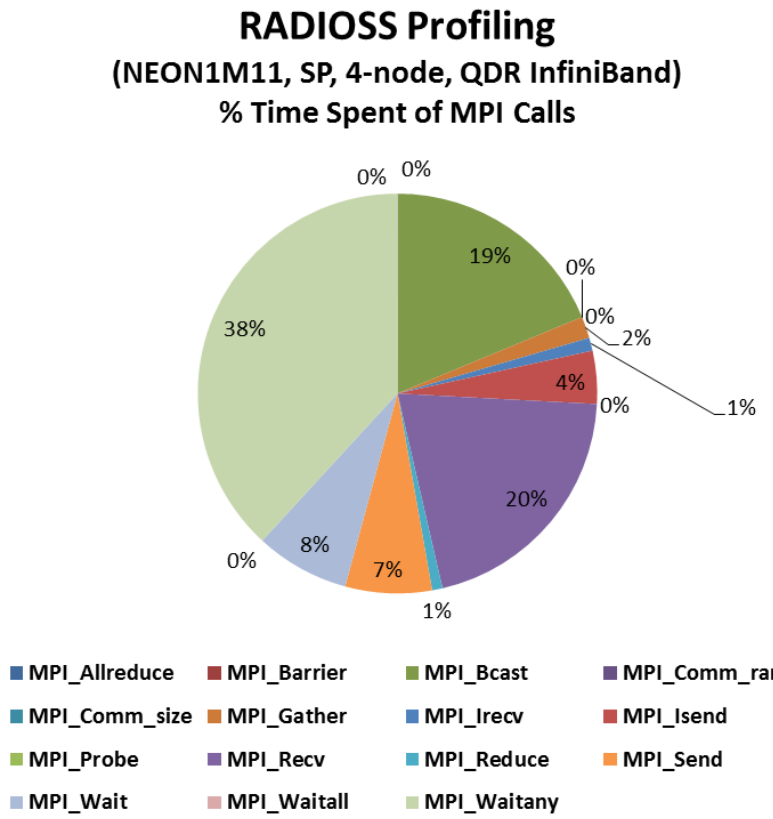



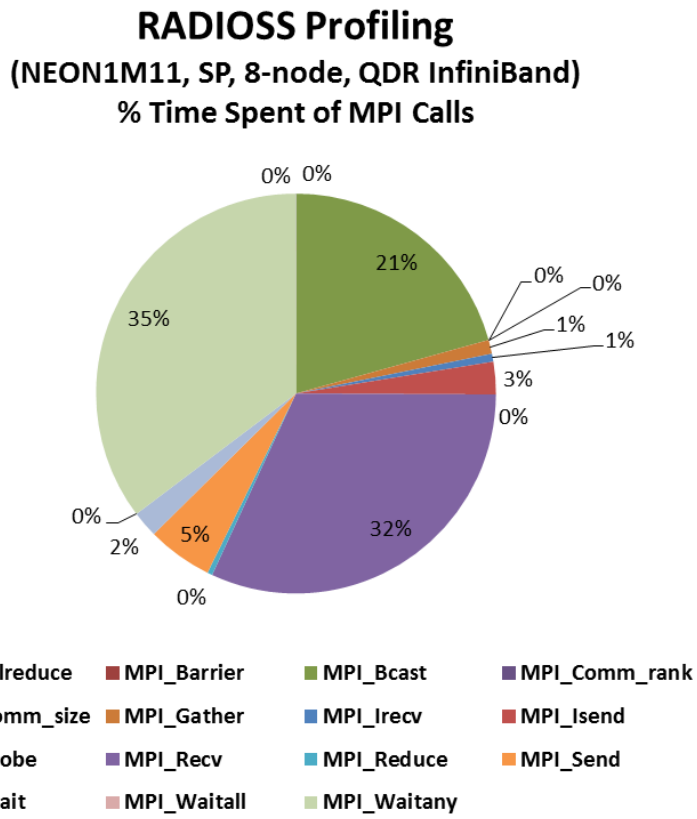

***8 Nodes***

***64 Cores/Node***

# RADIOSS Profiling – Time Spent of MPI Calls

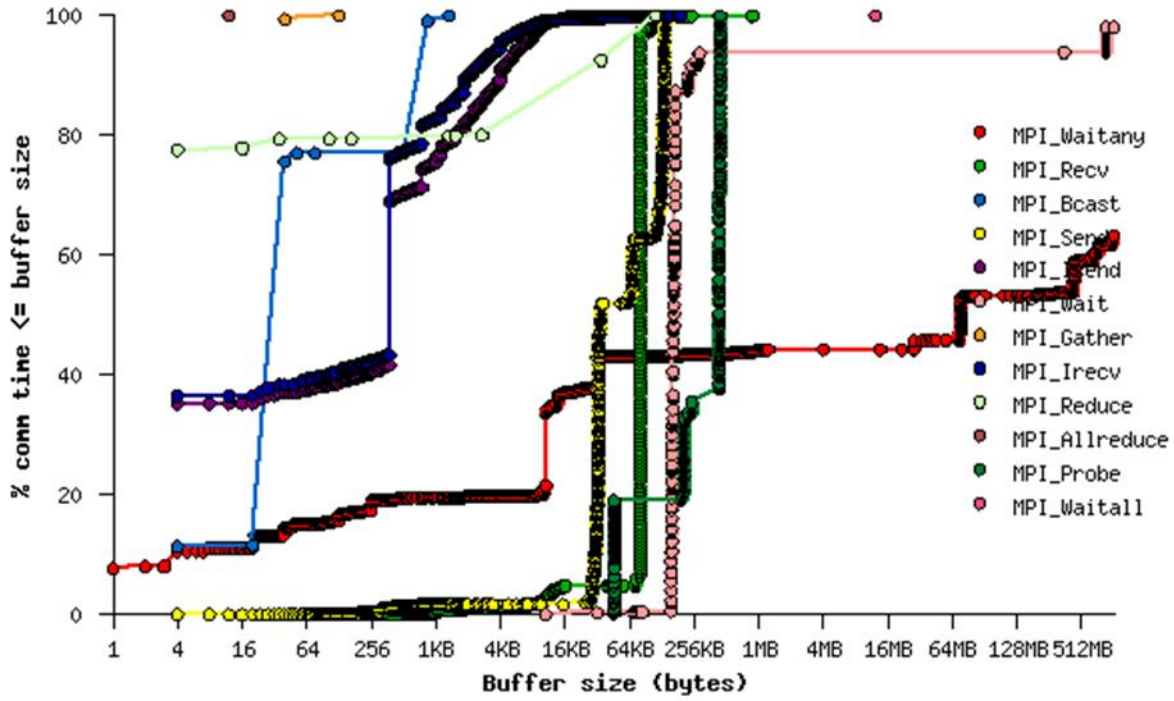
- **MPI_Waitany and MPI_Recv consume the most time for communications**
  - Occupies 35% for MPI_Waitany and 32% for MPI_Recv at 8 nodes
  - MPI_Waitany is for data transfers takes place in non-blocking communications



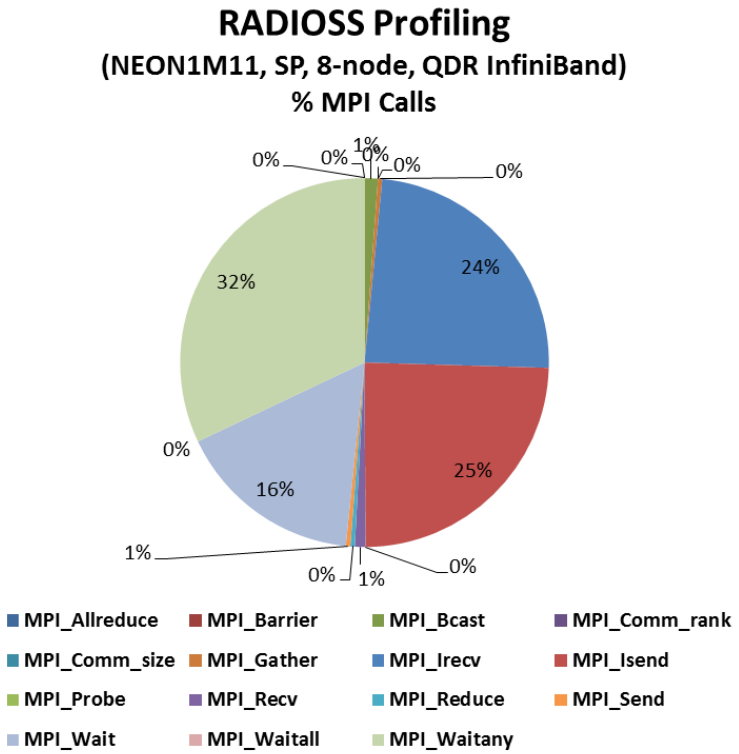
# RADIOSS Profiling – Distribution of Messages

- **MPI message sizes are concentrated in range of midrange message sizes**
  - Majority are in the range of 64KB
  - Large message sizes do exist for non-blocking communications
  - Mid to large messages (>16KB) responsible for data transfers between the MPI ranks
- **Reflects RADIOSS requires good network throughput for data movement**



# RADIOSS Profiling – Number of MPI Calls

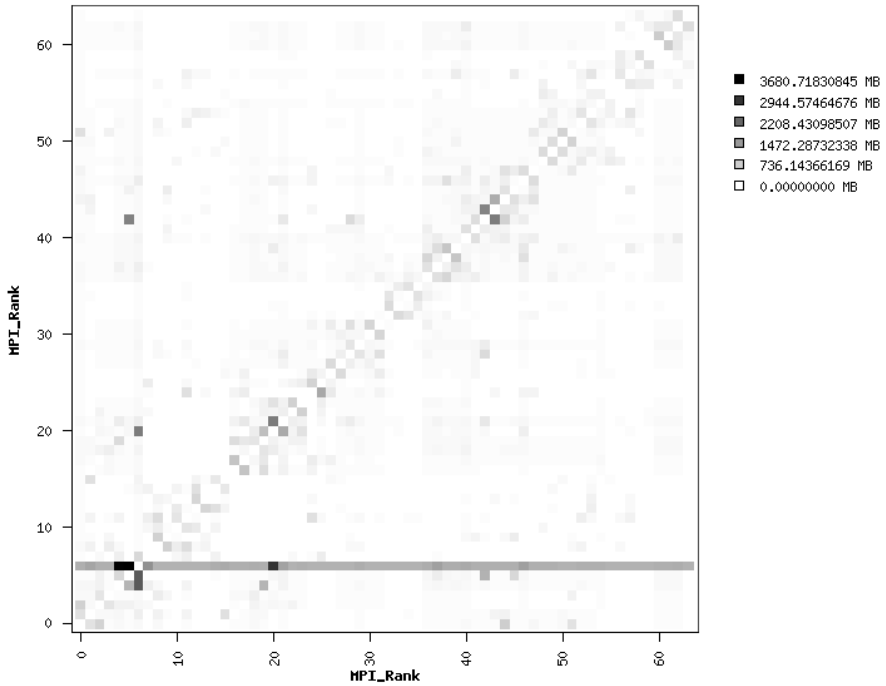
- **Majority of MPI calls are for non-blocking communications**
  - MPI_Waitany (32%), MPI_Isend (25%) and MPI_Irecv (24%)
- **RADIOSS demonstrates the advantages of non-blocking communication**
  - Allows efficient computation by overlapping communications and computation
  - Reduces the time needed for waiting for data inflight to complete, which enhances scalability



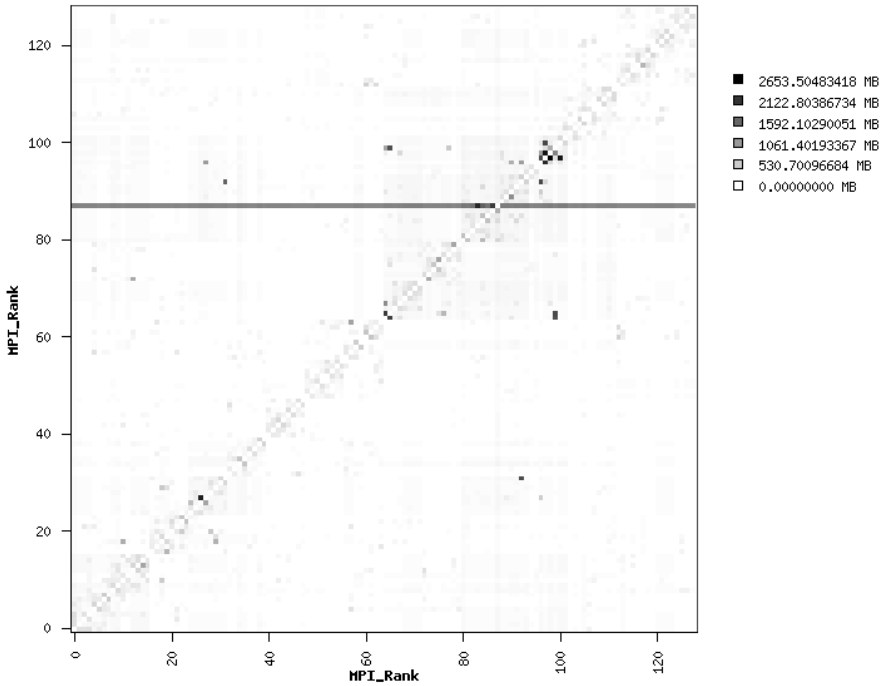
# RADIOSS Profiling – Communication Patterns

- **Majority of data transfers takes place between close neighboring MPI ranks**
  - Communications shown on the diagonal line in the graph
  - Except for 1 MPI process which does data exchanges with other processes
- **Application input data was staged on each node before the job start**
  - Which helps to reduce the amount of data transferred at runtime between compute nodes

**4 Nodes**




**8 Nodes**



# Summary

- **RADIOSS demonstrates ability to perform at large scale**
  - Hybrid MPP and non-blocking communications are features designed for scalability
- **Hybrid MPP mode**
  - Reduced processes which eliminates many data exchanges on network
  - HMPP mode performs 6 times faster than pure MPP mode at 8 nodes
- **Networking**
  - InfiniBand shows continuous gain as the cluster scales
    - Up to 98% to 140% higher productivity compared to 10GbE at 8 nodes
    - 10GbE scalability declines from 4 nodes and beyond
- **MPI/Data Communications on Network**
  - RADIOSS demonstrates good use of non-blocking communications by overlapping time for computation and data exchanges to achieve good scalability

# Thank You

## HPC Advisory Council