



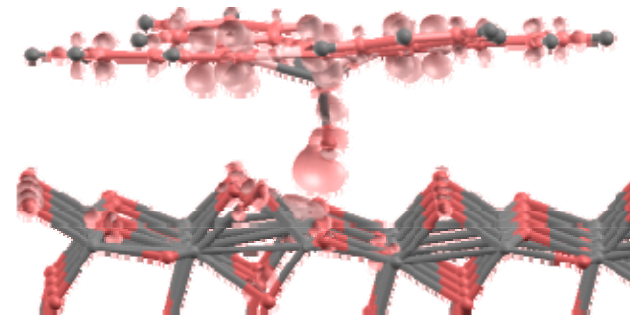
Quantum ESPRESSO Performance Benchmark and Profiling

March 2010



- **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**
 - www.mellanox.com, www.dell.com/hpc, www.amd.com
 - <http://www.quantum-espresso.org>

- Quantum ESPRESSO stands for opEn Source Package for Research in Electronic Structure, Simulation, and Optimization
- It is an integrated suite of computer codes for electronic-structure calculations and materials modeling at the nanoscale
- It is based on
 - Density-functional theory
 - Plane waves
 - Pseudopotentials (both norm-conserving and ultrasoft)
- Open source under the terms of the GNU General Public License



- **The presented research was done to provide best practices**
 - Quantum ESPRESSO performance benchmarking
 - Performance tuning with different communication libraries and compilers
 - Interconnect performance comparisons
 - Understanding Quantum ESPRESSO communication patterns
 - Power-efficient simulations
- **The presented results will demonstrate**
 - Balanced compute system enables
 - Good application scalability
 - Power saving

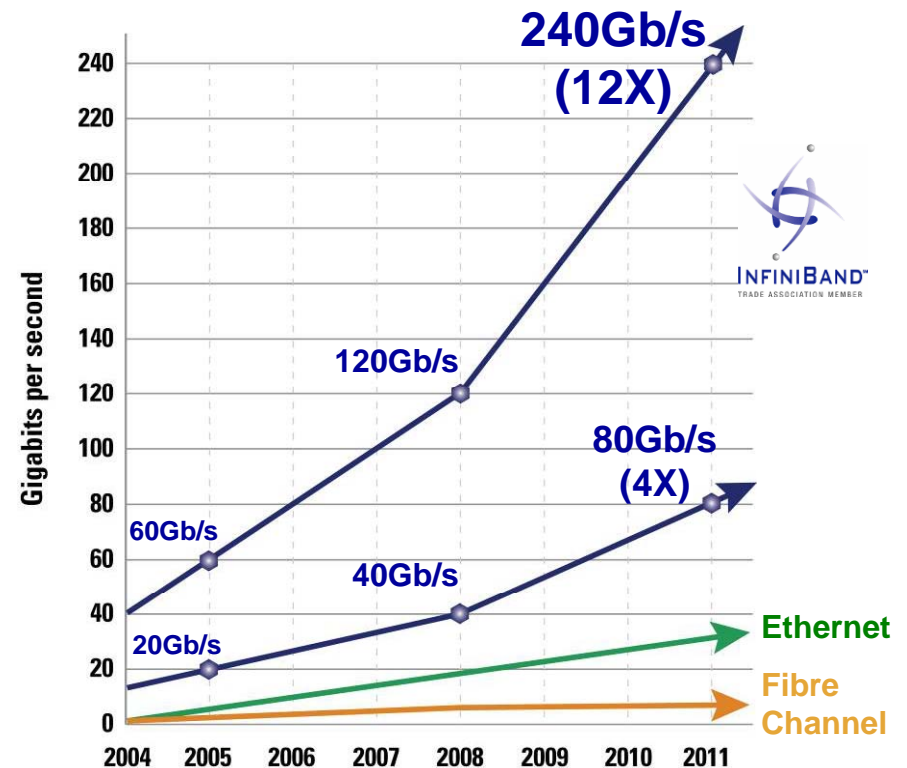
Test Cluster Configuration

- **Dell™ PowerEdge™ SC 1435 24-node cluster**
- **Quad-Core AMD Opteron™ 2382 (“Shanghai”) CPUs**
- **Mellanox® InfiniBand ConnectX® 20Gb/s (DDR) HCAs**
- **Mellanox® InfiniBand DDR Switch**
- **Memory: 16GB memory, DDR2 800MHz per node**
- **OS: RHEL5U3, OFED 1.5 InfiniBand SW stack**
- **MPI: OpenMPI-1.3.3, MVAPICH2-1.4, Platform MPI 5.6.7**
- **Application: Quantum ESPRESSO 4.1.2**
- **Benchmark Workload**
 - Medium size DEISA benchmark AUSURF112
 - Gold surface (112 atoms)

Mellanox InfiniBand Solutions

- **Industry Standard**
 - Hardware, software, cabling, management
 - Design for clustering and storage interconnect
- **Performance**
 - 40Gb/s node-to-node
 - 120Gb/s switch-to-switch
 - 1us application latency
 - Most aggressive roadmap in the industry
- **Reliable with congestion management**
- **Efficient**
 - RDMA and Transport Offload
 - Kernel bypass
 - CPU focuses on application processing
- **Scalable for Petascale computing & beyond**
- **End-to-end quality of service**
- **Virtualization acceleration**
- **I/O consolidation Including storage**

The InfiniBand Performance Gap is Increasing



InfiniBand Delivers the Lowest Latency

Quad-Core AMD Opteron™ Processor

- **Performance**

- Quad-Core

- Enhanced CPU IPC
- 4x 512K L2 cache
- 6MB L3 Cache

- Direct Connect Architecture

- HyperTransport™ Technology
- Up to 24 GB/s peak per processor

- Floating Point

- 128-bit FPU per core
- 4 FLOPS/clock peak per core

- Integrated Memory Controller

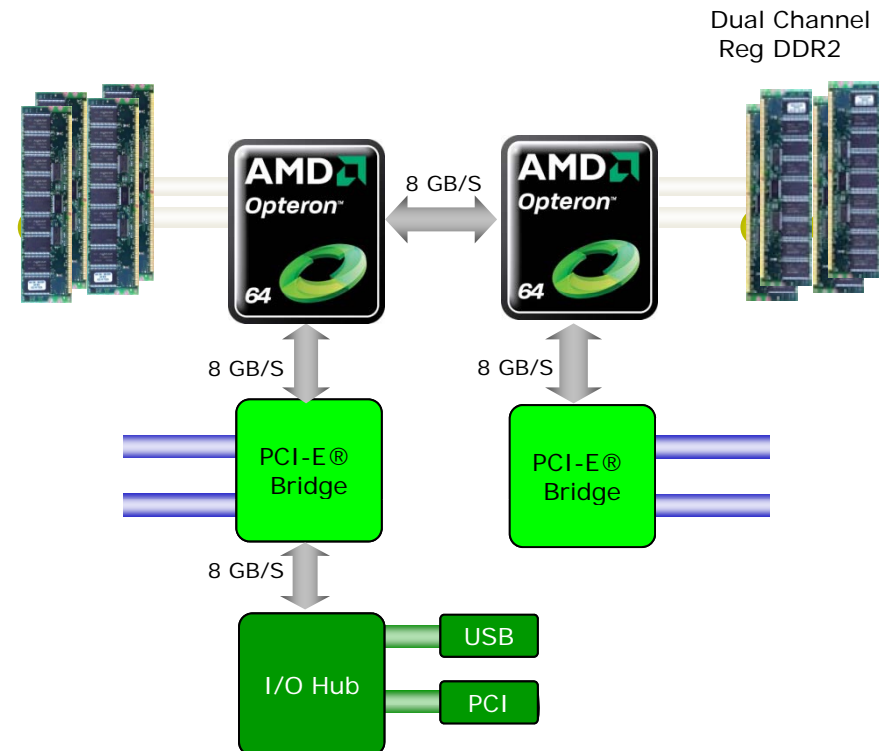
- Up to 12.8 GB/s
- DDR2-800 MHz or DDR2-667 MHz

- **Scalability**

- 48-bit Physical Addressing

- **Compatibility**

- Same power/thermal envelopes as 2nd / 3rd generation AMD Opteron™ processor



Dell PowerEdge Servers helping Simplify IT

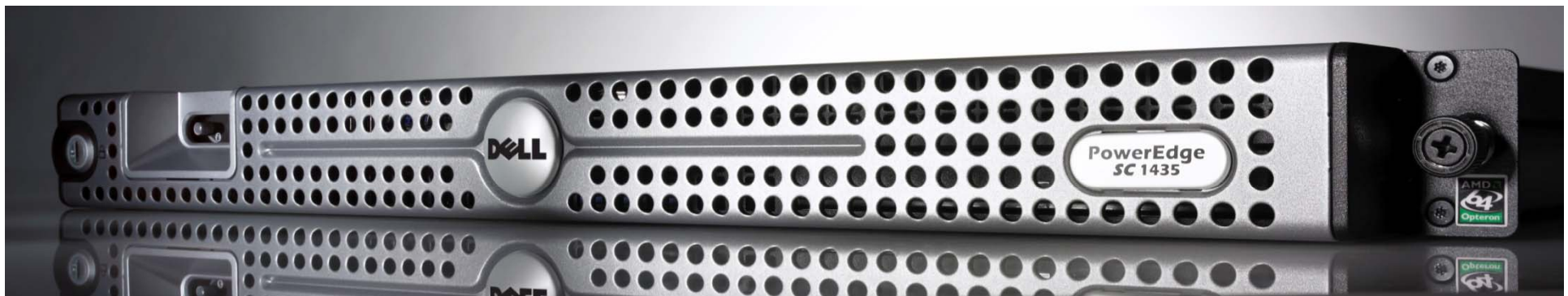
- **System Structure and Sizing Guidelines**
 - 24-node cluster build with Dell PowerEdge™ SC 1435 Servers
 - Servers optimized for High Performance Computing environments
 - Building Block Foundations for best price/performance and performance/watt
- **Dell HPC Solutions**
 - Scalable Architectures 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



Dell PowerEdge™ Server Advantage



- Dell™ PowerEdge™ servers incorporate AMD Opteron™ and Mellanox ConnectX InfiniBand to provide leading edge performance and reliability
- Building Block Foundations for best price/performance and performance/watt
- Investment protection and energy efficient
- Longer term server investment value
- Faster DDR2-800 memory
- Enhanced AMD PowerNow!
- Independent Dynamic Core Technology
- AMD CoolCore™ and Smart Fetch Technology
- Mellanox InfiniBand end-to-end for highest networking performance

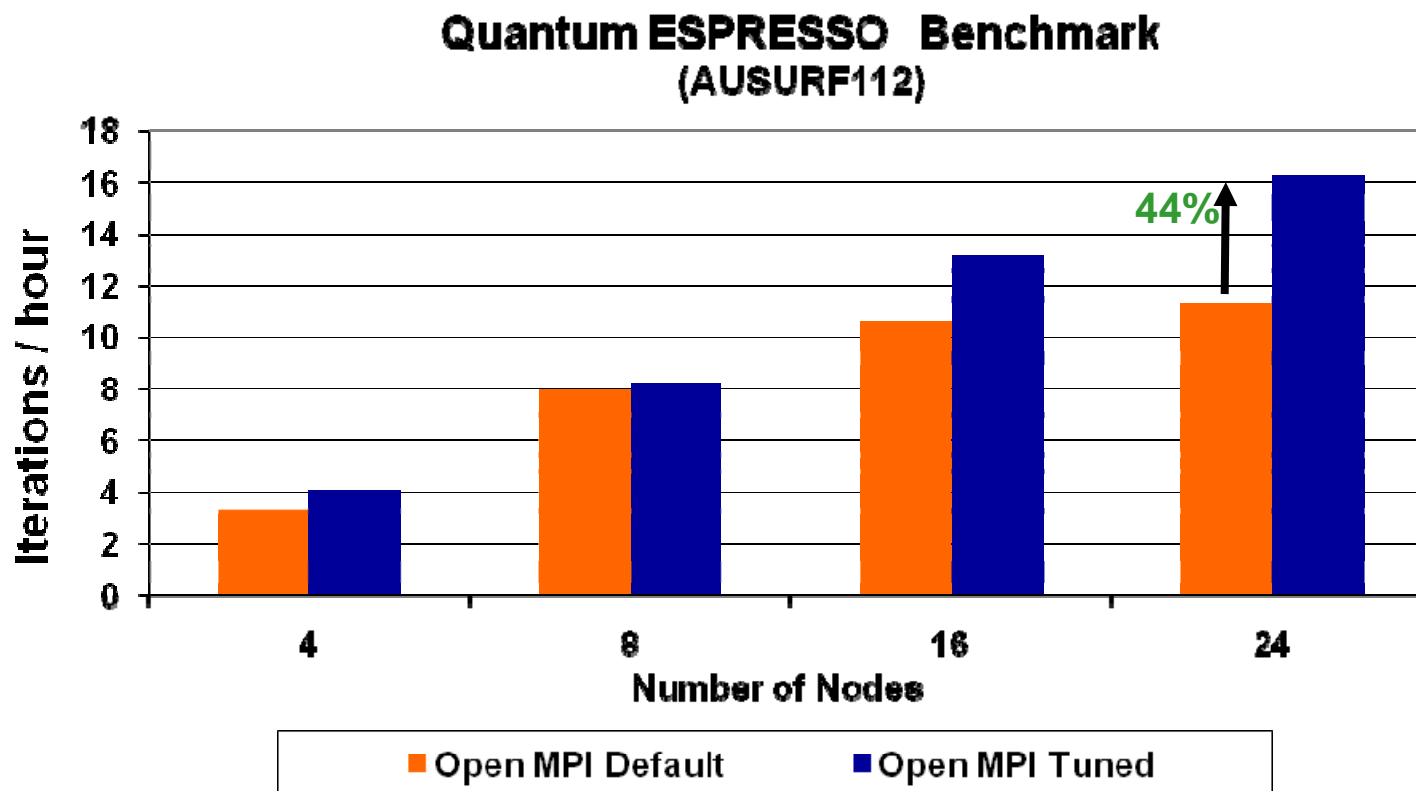


Quantum ESPRESSO Benchmark Results

- **Customized MPI parameters provide better performance**

- **Up to 44% higher performance with Open MPI**

- `--mca mpi_affinity_alone 1 --mca coll_tuned_use_dynamic_rules 1 --mca coll_tuned_alltoallv_algorithm 2 --mca coll_tuned_allreduce_algorithm 0 --mca coll_tuned_barrier_algorithm 6`

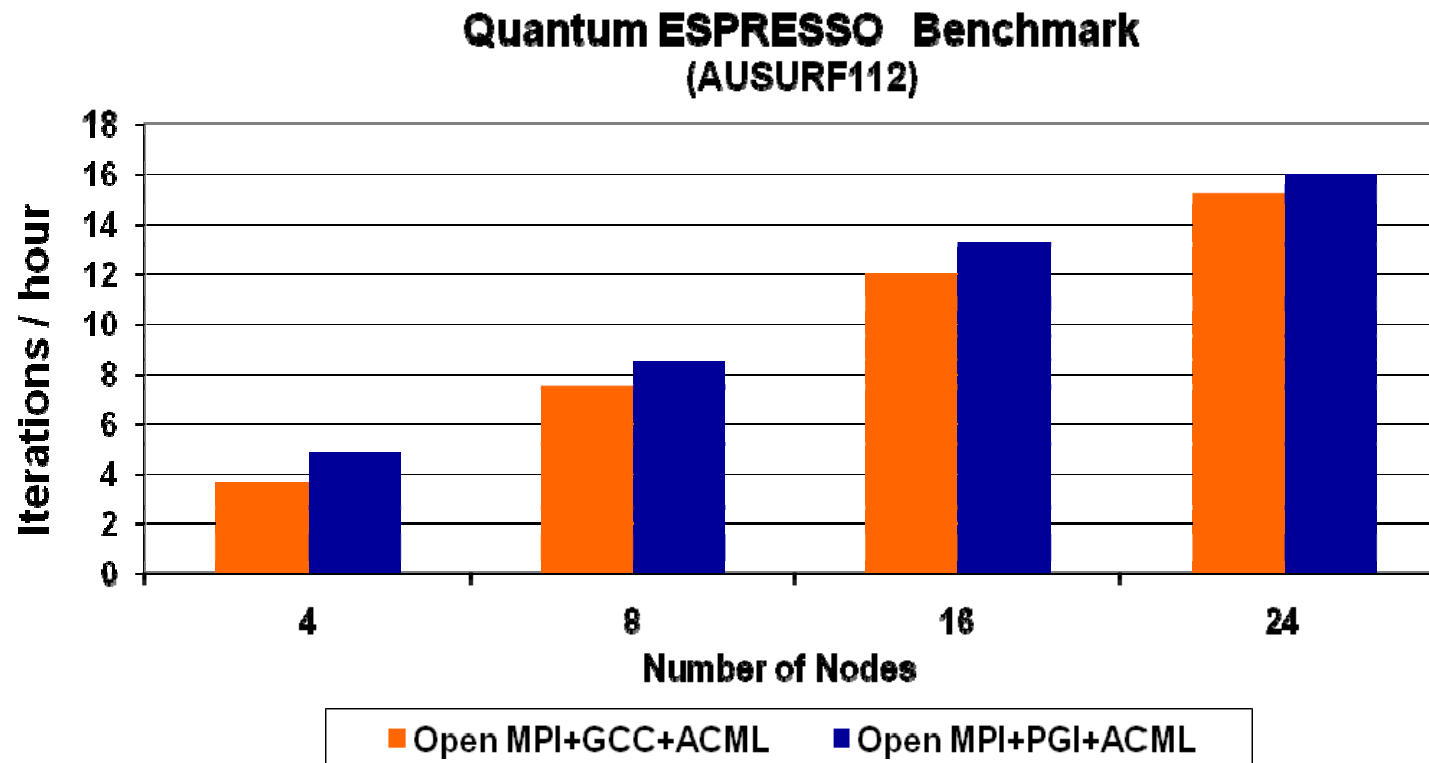


Higher is better

8-cores per node

Quantum ESPRESSO Benchmark Results

- **PGI compiler enables higher performance and better scalability**
 - Up to 4% higher performance versus GCC compiler

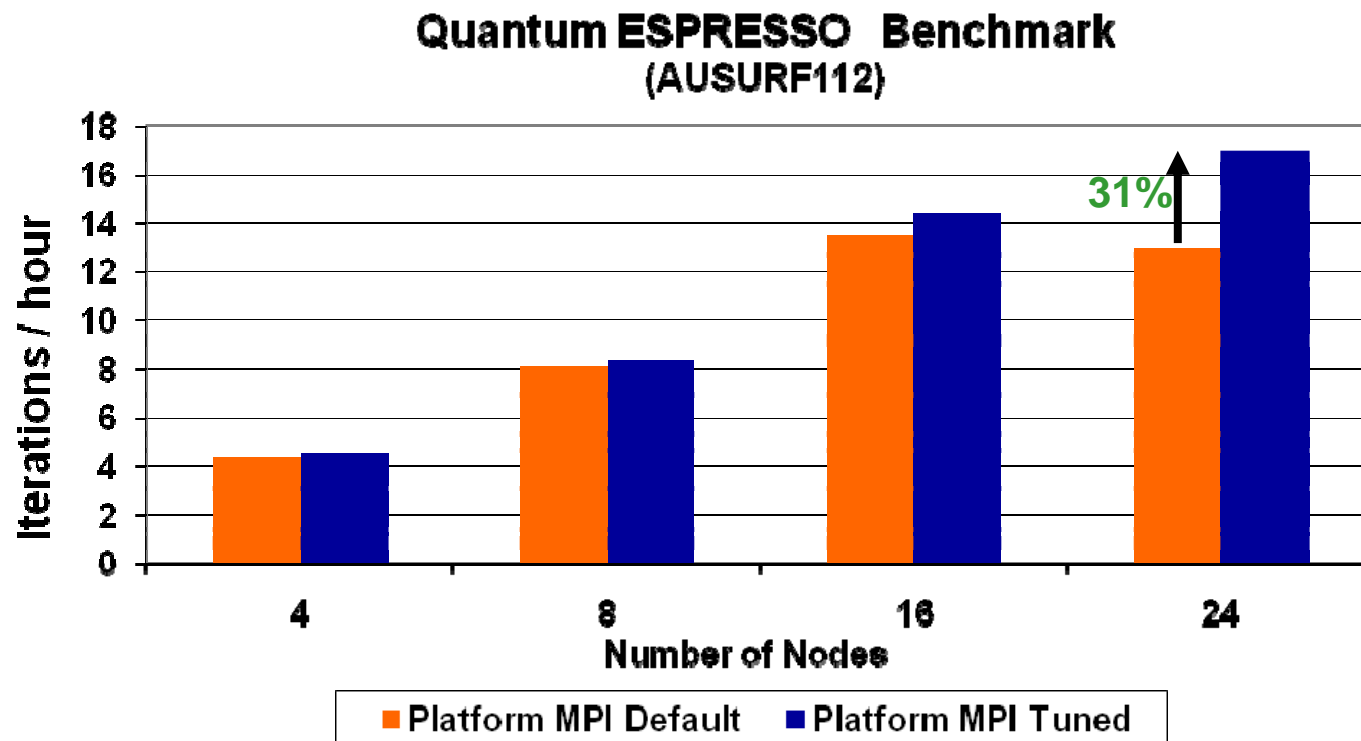


Higher is better

8-cores per node

Quantum ESPRESSO Benchmark Results

- **Optimized MPI parameter provide better performance**
 - Up to 31% higher performance with customized MPI_Alltoallv parameter
 - SCAMPI_ALLTOALLV_ALGORITHM=7

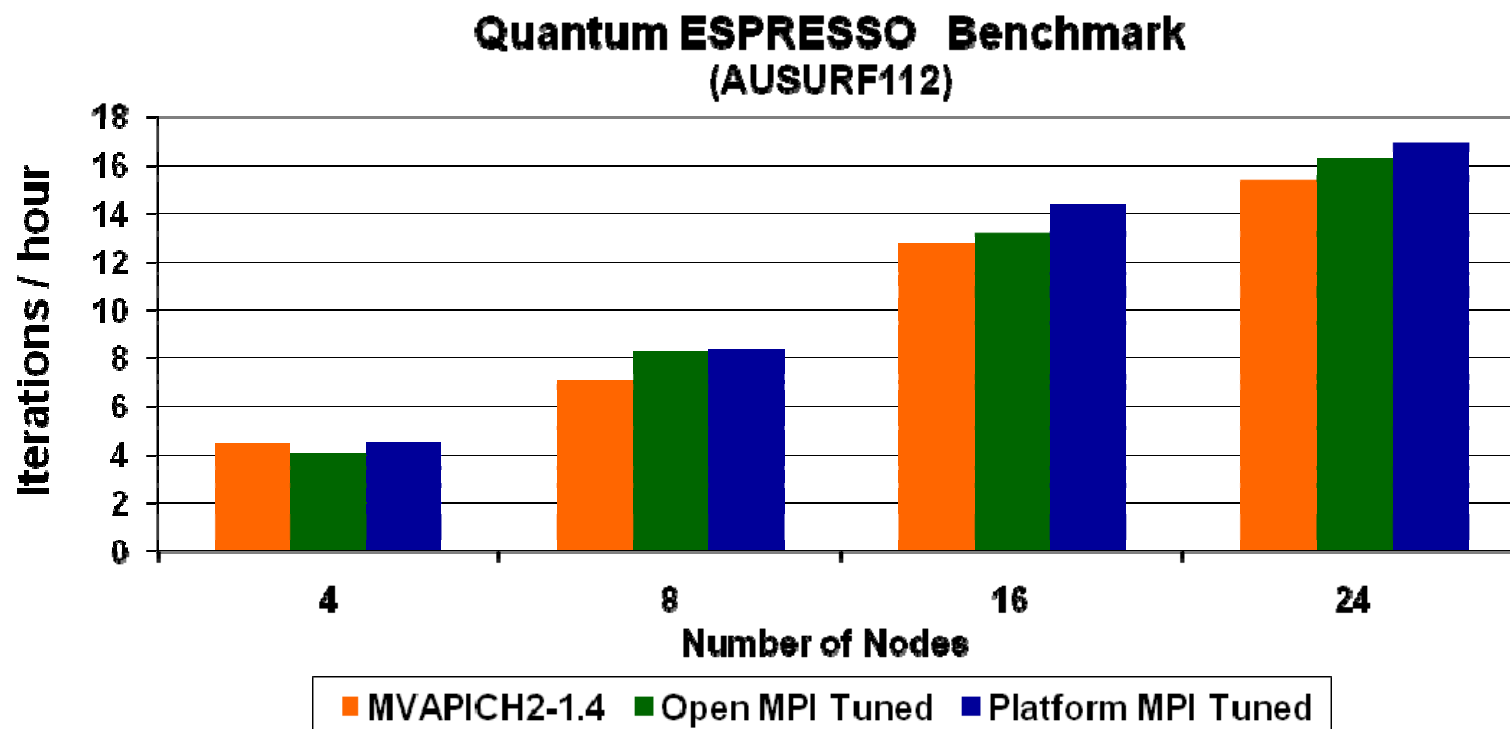


Higher is better

8-cores per node

Quantum ESPRESSO Benchmark Results

- **Platform MPI with ACML enables higher performance**
 - Up to 4% higher performance than Open MPI and 10% than MVAPICH

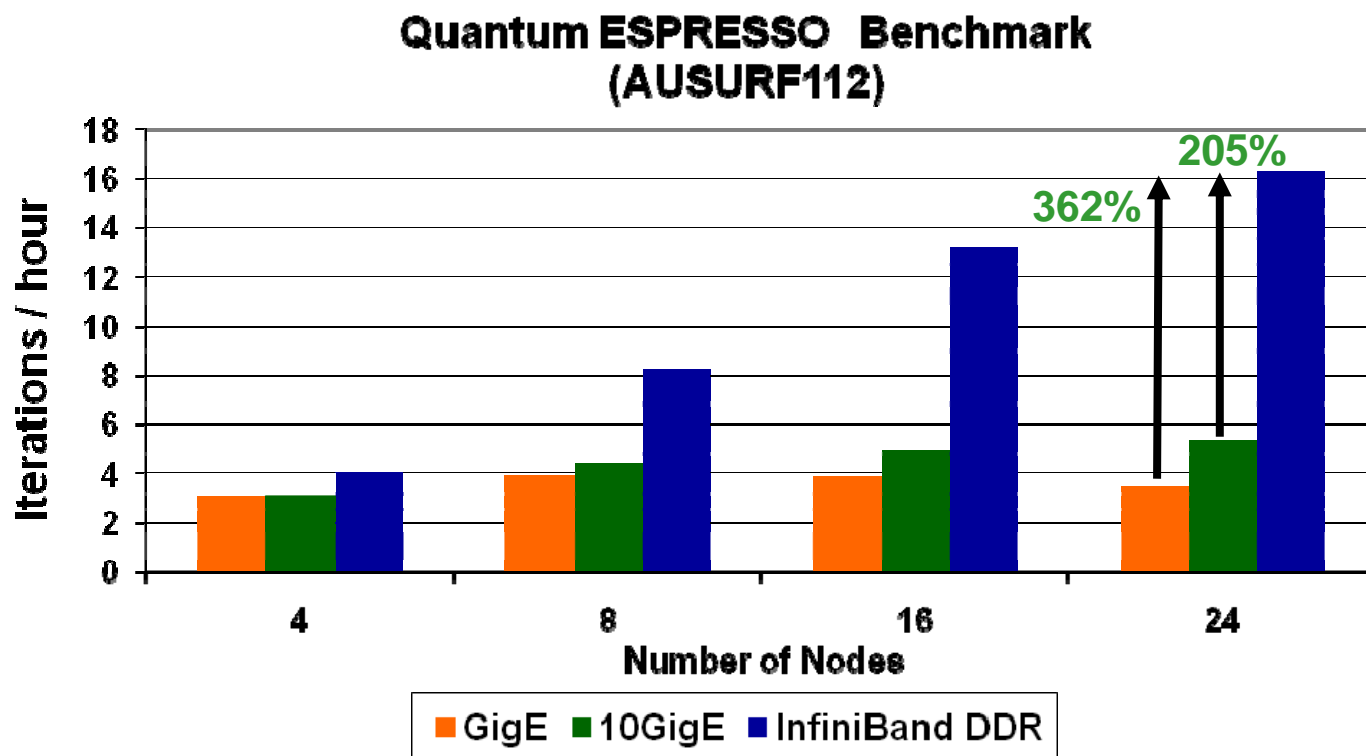


Higher is better

8-cores per node

Quantum ESPRESSO Benchmark Results

- **InfiniBand enables better application performance and scalability**
 - Up to 205% higher performance than 10GigE and 362% than GigE
- **Application performance over InfiniBand scales as cluster size increases**

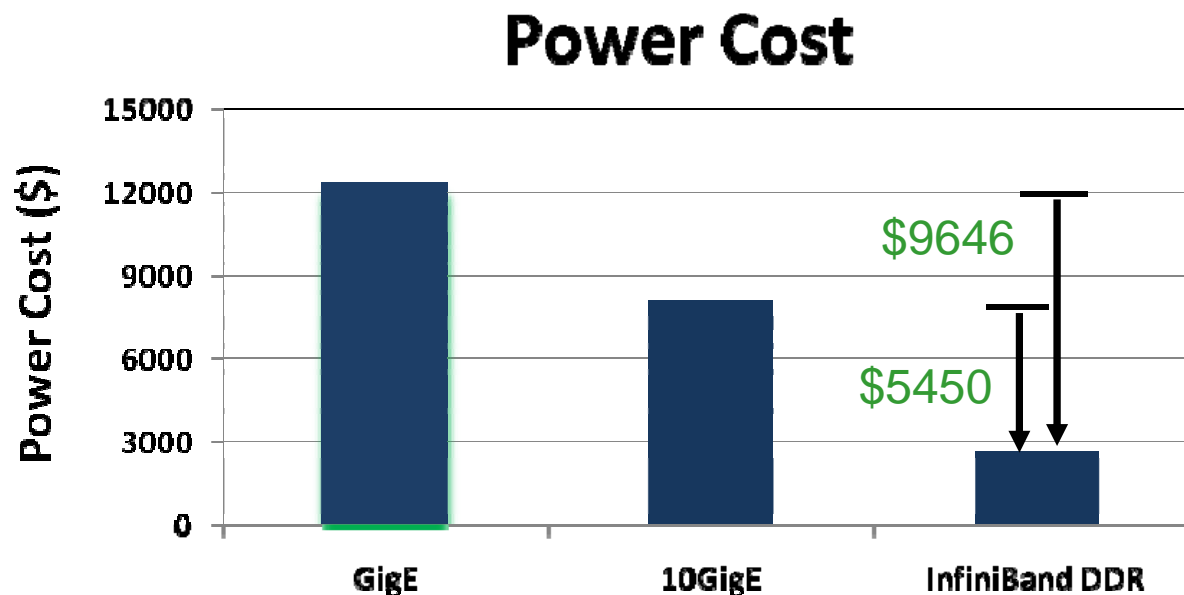


Higher is better

8-cores per node

Power Cost Savings with Different Interconnect

- **Dell economical integration of AMD CPUs and Mellanox InfiniBand**
 - To achieve same number of Quantum ESPRESSO jobs over GigE
 - InfiniBand saves power up to \$5450 versus 10GigE and \$9646 versus GigE
 - Yearly based for 24-node cluster
- **As cluster size increases, more power can be saved**



$\$/KWh = KWh * \0.20

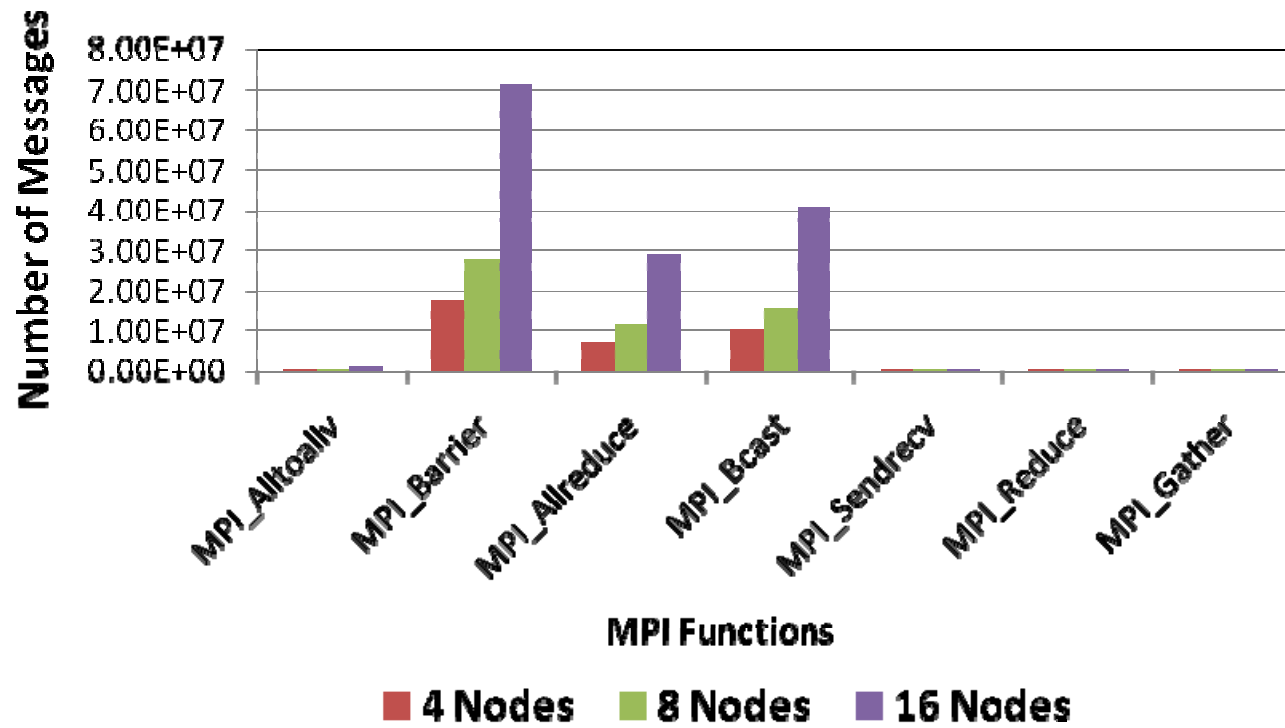
For more information - <http://enterprise.amd.com/Downloads/svrpwrusecompletefinal.pdf>

Quantum ESPRESSO Benchmark Summary

- **Tuned MPI parameters provides better performance**
 - Both Open MPI and Platform MPI gain extra performance by adjusting parameters
- **PGI compiler with ACML enables higher performance than GCC**
- **Interconnect comparison shows**
 - InfiniBand delivers superior performance in every cluster size versus GigE and 10GigE
 - Performance advantage extends as cluster size increases
- **InfiniBand enables power saving**
 - Up to \$9646/year power savings versus GigE and \$5450 versus 10GigE
- **Dell™ PowerEdge™ server blades provides**
 - Linear scalability (maximum scalability) and balanced system
 - By integrating InfiniBand interconnect and AMD processors
 - Maximum return on investment through efficiency and utilization

- **Mostly used MPI functions**
 - MPI_Barrier, MPI_Allreduce, MPI_Bcast, and MPI_Alltoallv
 - Collective functions are the mostly used MPI functions
 - Number of MPI calls increases as cluster size scales

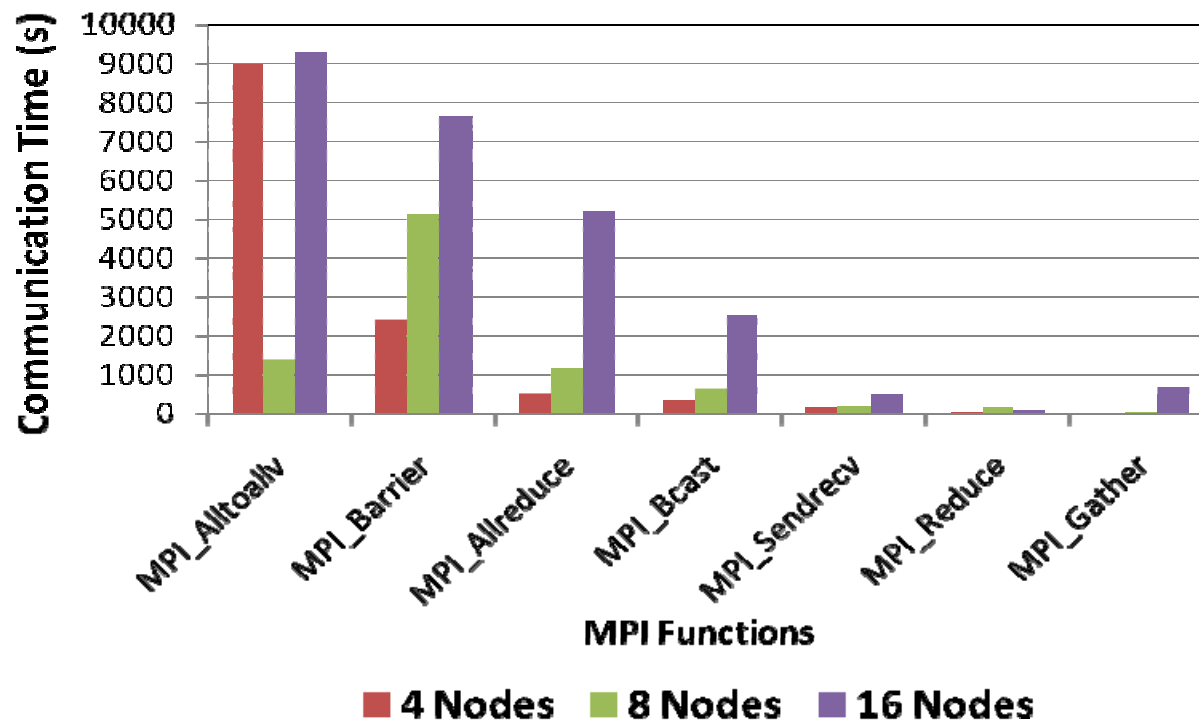
MPI Profiling of Quantum ESPRESSO (Number of MPI messages)



- **Mostly used MPI functions**

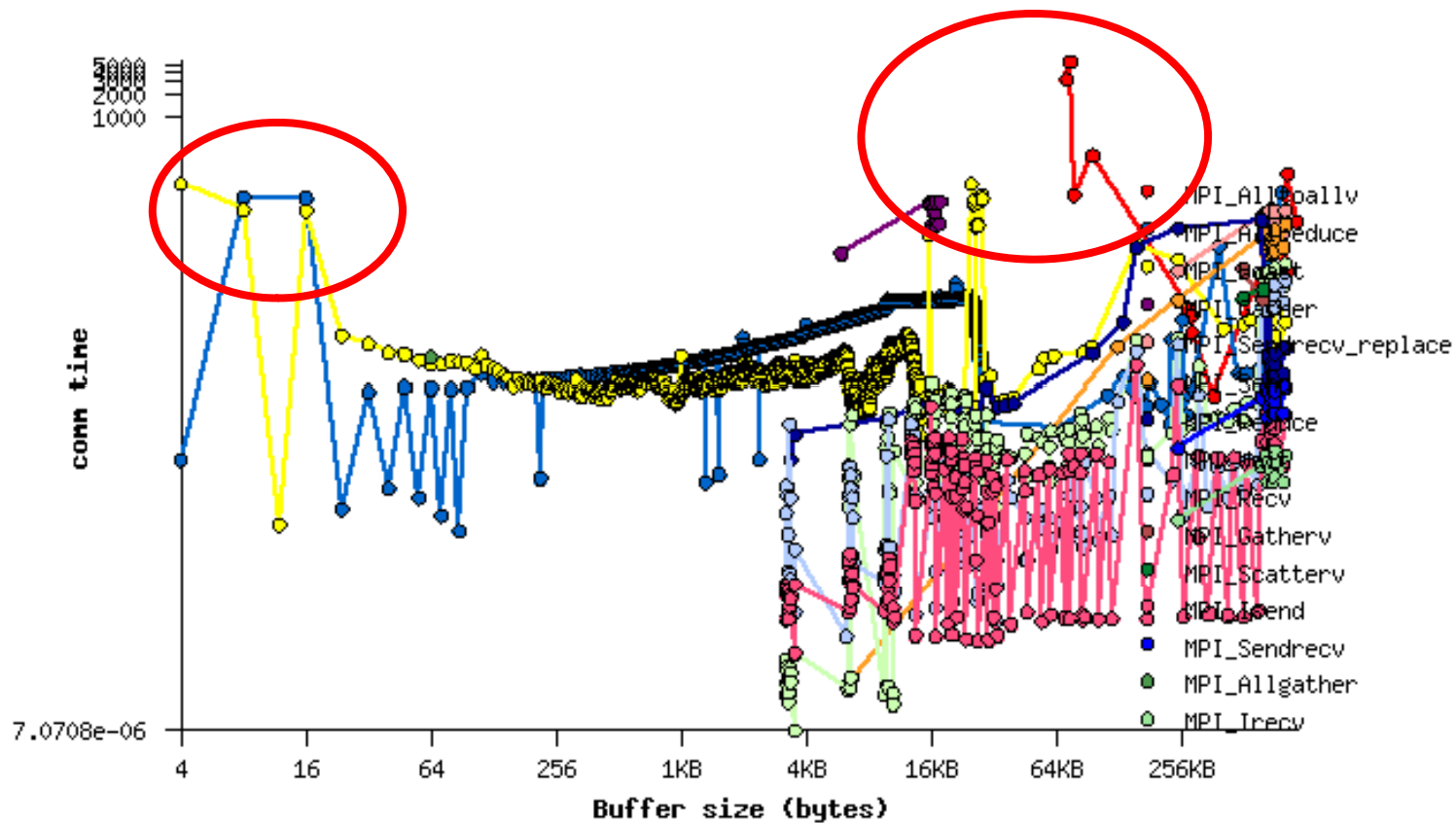
- MPI_Alltoallv, MPI_Barrier, MPI_Allreduce, and MPI_Bcast create the biggest overhead
- Communication overhead increases as cluster size increases

MPI Profiling of Quantum ESPRESSO (MPI Timing)



Quantum ESPRESSO MPI Profiling – Message Size

- **Messages with big communication overhead are**
 - Large messages >64KB
 - Small message <16Bytes



128 Processes

- **Quantum ESPRESSO was profiled to identify its communication patterns**
 - MPI collective functions create the biggest communication overhead
 - Number of messages increases with cluster size
- **Interconnects effect to Quantum ESPRESSO performance**
 - Most messages are either large than 64KB or smaller than 16Bytes
 - Both bandwidth and latency is critical to application performance
- **Balanced system – CPU, memory, Interconnect that match each other capabilities, is essential for providing application efficiency**

Thank You

HPC Advisory Council



All trademarks are property of their respective owners. All information is provided "As-Is" without any kind of warranty. The HPC Advisory Council makes no representation to the accuracy and completeness of the information contained herein. HPC Advisory Council Mellanox undertakes no duty and assumes no obligation to update or correct any information presented herein