Quantum ESPRESSO Performance Benchmark and Profiling

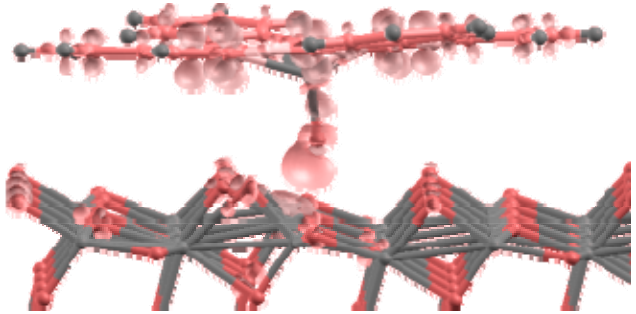
May 2010

Note

- **The following research was performed under the HPC Advisory Council activities**
  - Participating vendors: Intel, Dell, Mellanox
  - Compute resource - HPC Advisory Council Cluster Center
- **For more info please refer to**
  - www.mellanox.com, www.dell.com/hpc, www.intel.com, http://www.quantum-espresso.org

Quantum ESPRESSO

- 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



Objectives

- **The presented research was done to provide best practices**
  - Quantum ESPRESSO performance benchmarking
    - MPI Library performance comparisons
    - Interconnect performance comparisons
  - Understanding Quantum ESPRESSO communication patterns
  - Power-efficient simulations
- **The presented results will demonstrate**
  - The scalability of the compute environment to provide good application scalability
  - Considerations for power saving through balanced system configuration

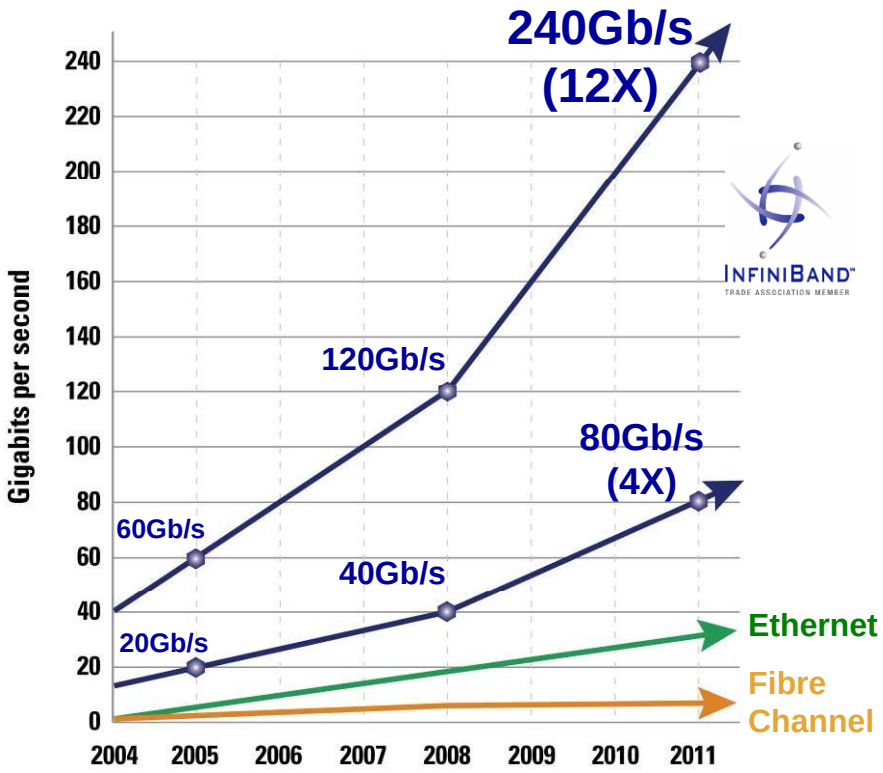
Test Cluster Configuration

- **Dell™ PowerEdge™ M610 16-node cluster**
- **Quad-Core Intel X5570 @ 2.93 GHz CPUs**
- **Intel Cluster Ready certified cluster**
- **Mellanox ConnectX2 QDR InfiniBand mezzanine card**
- **Mellanox M3601Q 32-Port Quad Data Rate (QDR-40Gb) InfiniBand Switch**
- **Memory: 24GB memory per node**
- **OS: RHEL5U3, OFED 1.5 InfiniBand SW stack**
- **File system: Lustre 1.8.2**
- **MPI: Open MPI 1.3.3, MVAPICH2-1.4, Intel MPI 4.0**
- **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**

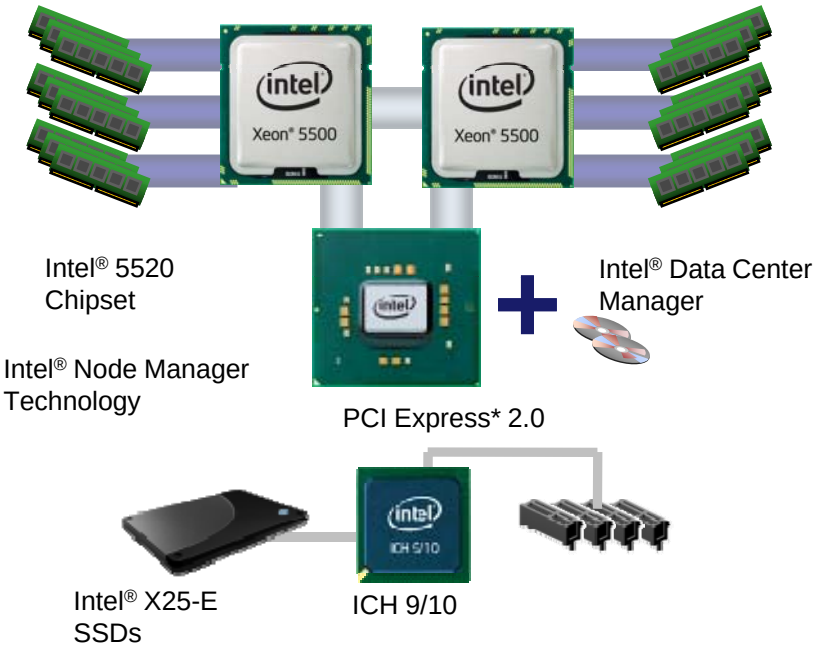
The InfiniBand Performance Gap is Increasing



InfiniBand Delivers the Lowest Latency

Delivering Intelligent Performance

Next Generation Intel® Microarchitecture



Bandwidth Intensive

- Intel® QuickPath Technology
- Integrated Memory Controller

Threaded Applications

- 45nm quad-core Intel® Xeon® Processors
- Intel® Hyper-threading Technology

Performance on Demand

- Intel® Turbo Boost Technology
- Intel® Intelligent Power Technology

Performance That Adapts to The Software Environment

Intel® Cluster Ready

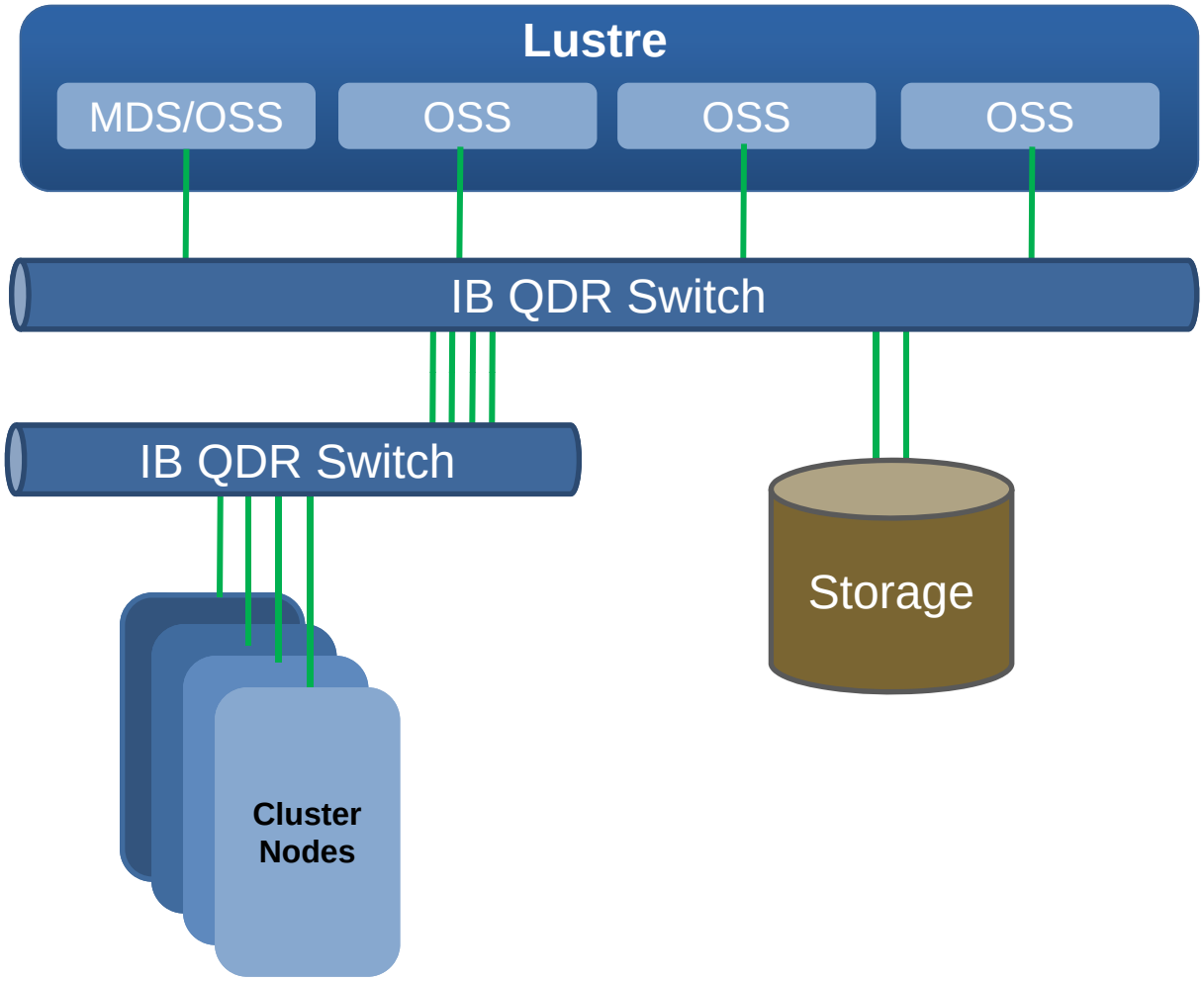
- **Intel® Cluster Ready is a consistent reference Linux platform architecture for Intel-based systems**
  - Makes it easier to design, develop, and build applications for clusters
- **A single architecture platform supported and used by a wide range of OEMs, ISVs, cluster provisioning vendors, and interconnect providers**
- **Includes**
  - Platform specification, that defines the Intel Cluster Ready platforms
  - Program branding, that makes it easier to identify compliant solutions and applications
  - Hardware certifications, confirming solutions that are delivered ready to run
  - Application registration, validating applications that execute on top of Intel Cluster Ready architecture
  - Intel® Cluster Checker tool, to validate hardware and software configuration and functionality

Dell PowerEdge Servers helping Simplify IT

- **System Structure and Sizing Guidelines**
  - 16-node cluster build with Dell PowerEdge™ M610 blades server
  - 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

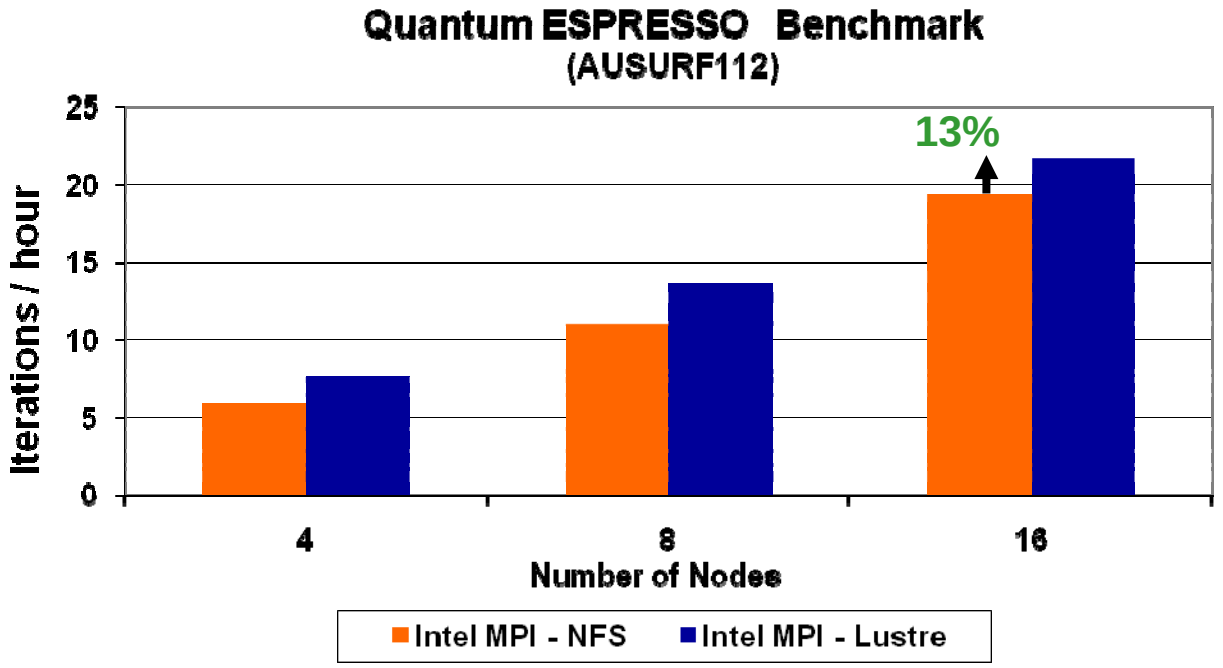
Lustre File System Configuration

- **Lustre Configuration**
  - 1 MDS
  - 4 OSS (Each has 2 OST)
  - InfiniBand based Backend storage
  - All components are connected through InfiniBand QDR interconnect



Quantum ESPRESSO Benchmark Results - File System

- **Intel MPI has native Lustre support**
  - mpiexec -genv I_MPI_EXTRA_FILESYSTEM on -genv I_MPI_EXTRA_FILESYSTEM_LIST lustre
- **Lustre enables higher performance**
  - Up to 13% faster than local hard disk at 16 nodes

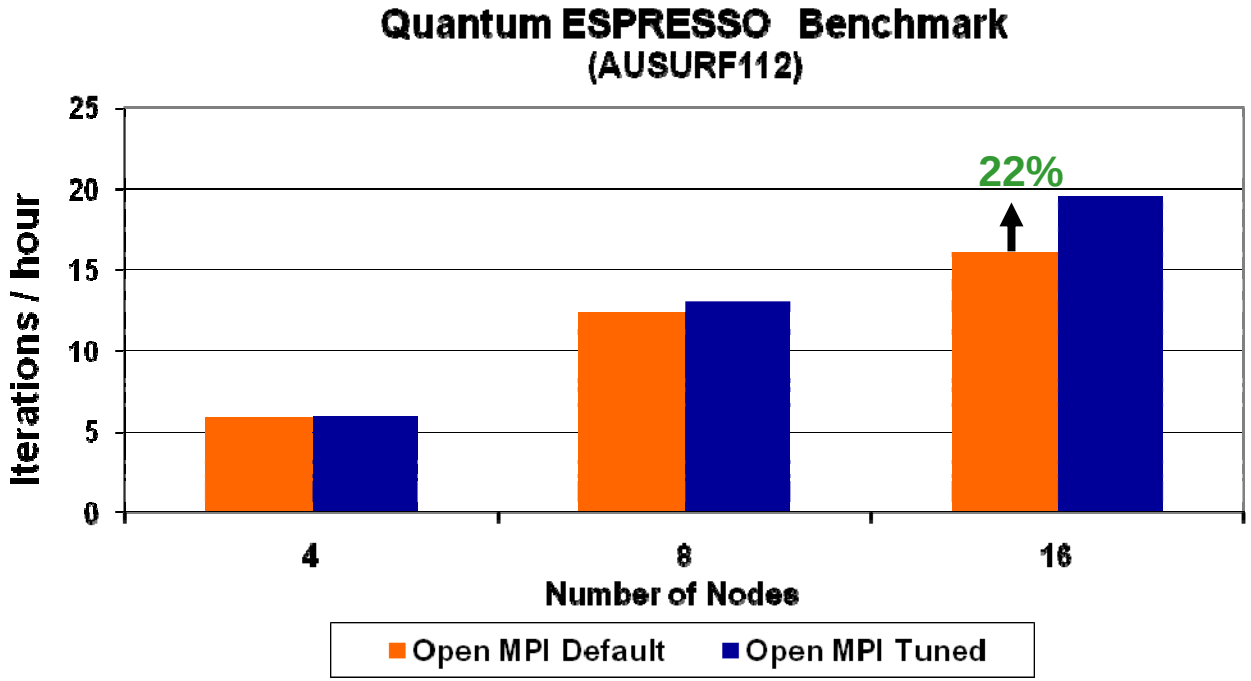


Higher is better

8-cores per node

Quantum ESPRESSO Benchmark Results

- **Customized MPI parameters provide better performance**
  - **Up to 22% 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

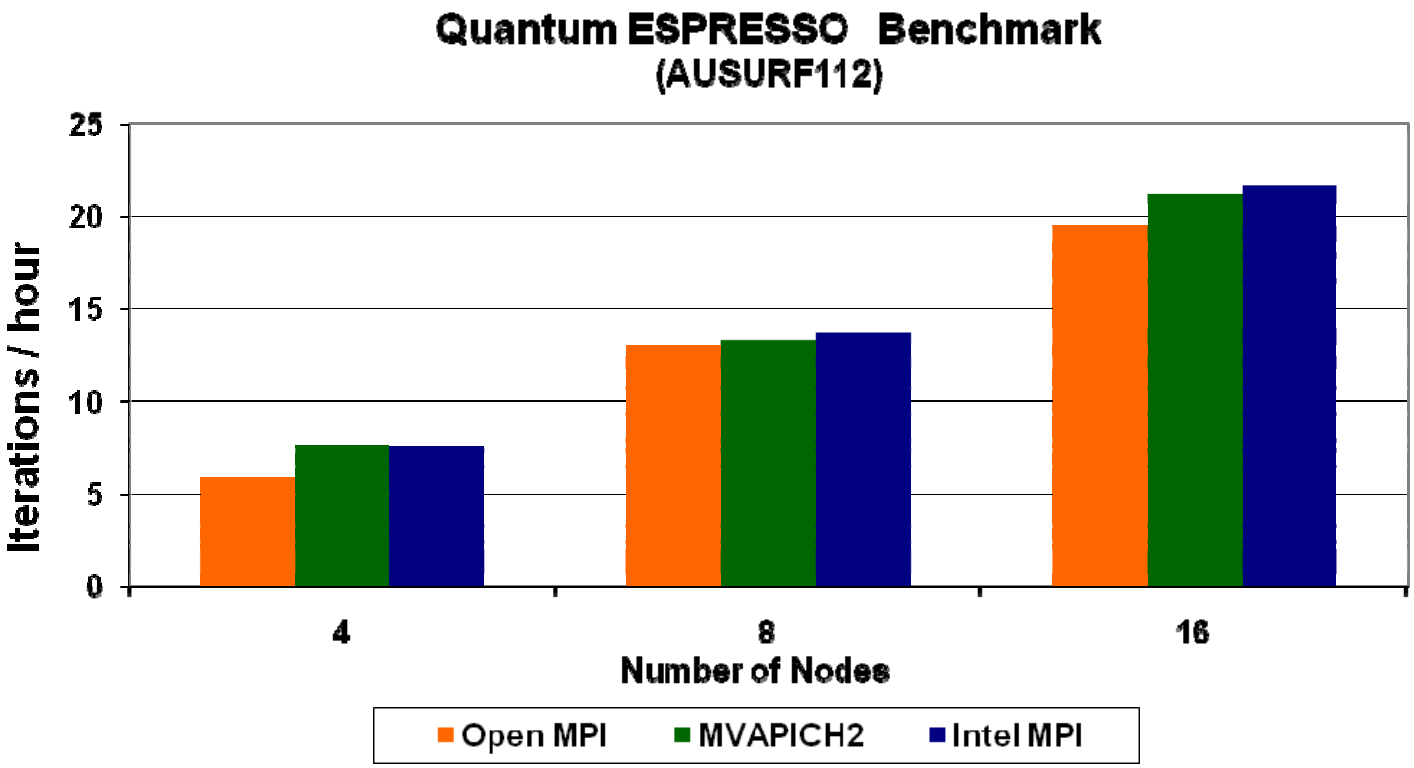


Higher is better

8-cores per node

Quantum ESPRESSO Benchmark Results

- **Intel MPI enables higher performance**
  - Up to 2% higher performance than MVAPICH2 and 12% than Open MPI

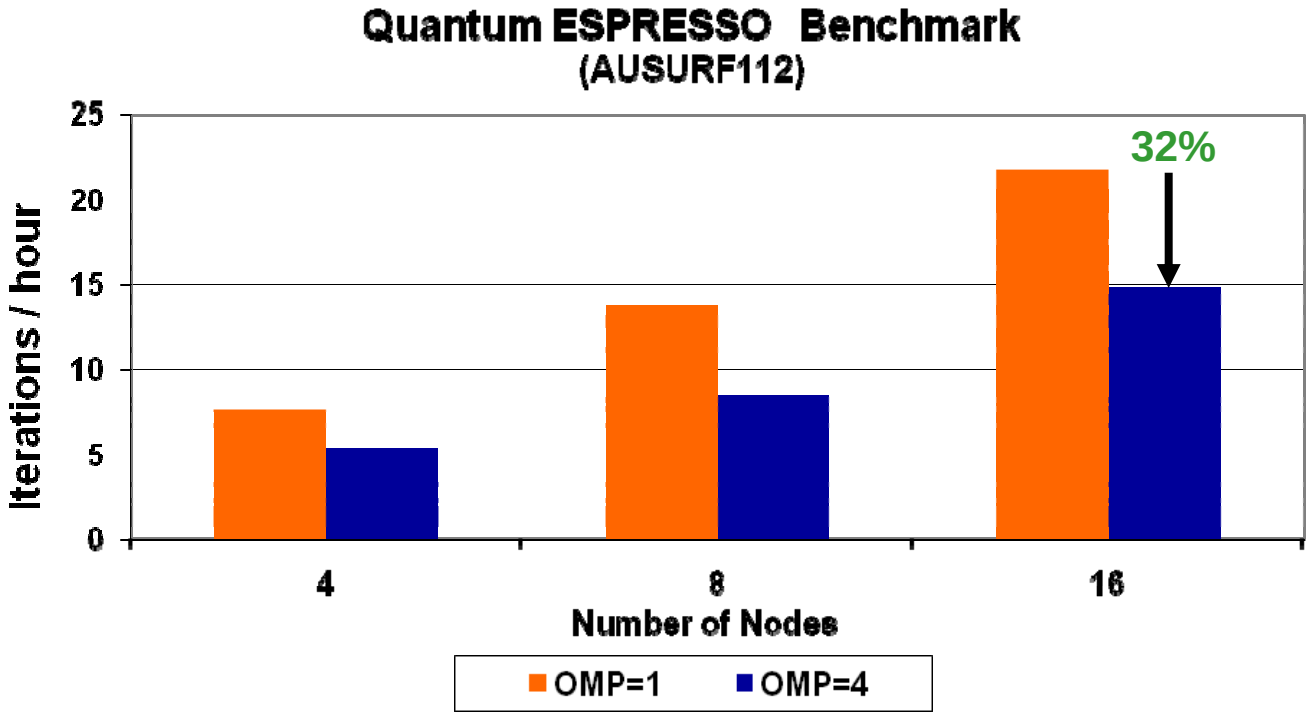


Higher is better

8-cores per node

Quantum ESPRESSO Benchmark Results

- **Multi-thread Intel MPI doesn't provide higher performance**
  - Up to 32% slower than non-threaded application performance

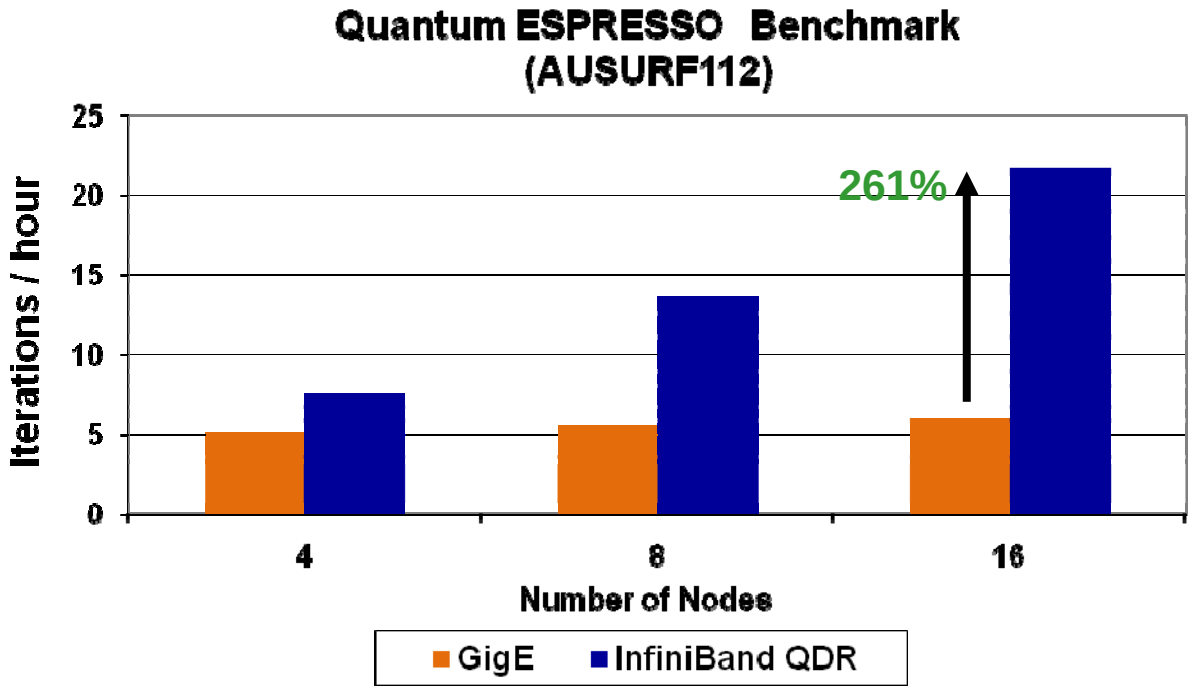


Higher is better

8-cores per node

Quantum ESPRESSO Benchmark Results

- **InfiniBand enables better application performance and scalability**
  - Up to 261% higher performance than GigE
  - GigE stops scaling after 8 nodes
- **Application performance over InfiniBand scales as cluster size increases**

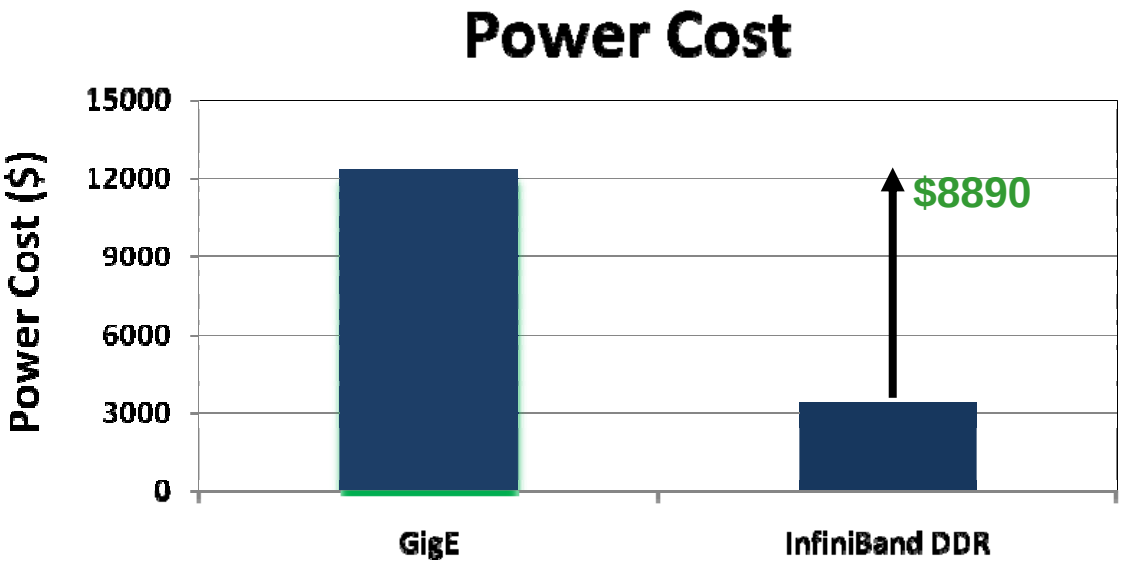


Higher is better

8-cores per node

Power Cost Savings with Different Interconnect

- **InfiniBand saves up to $8890 power compared to GigE**
  - To finish the same number of Quantum ESPRESSO jobs
  - Yearly based for 16-node cluster
- **As cluster size increases, more power can be saved**



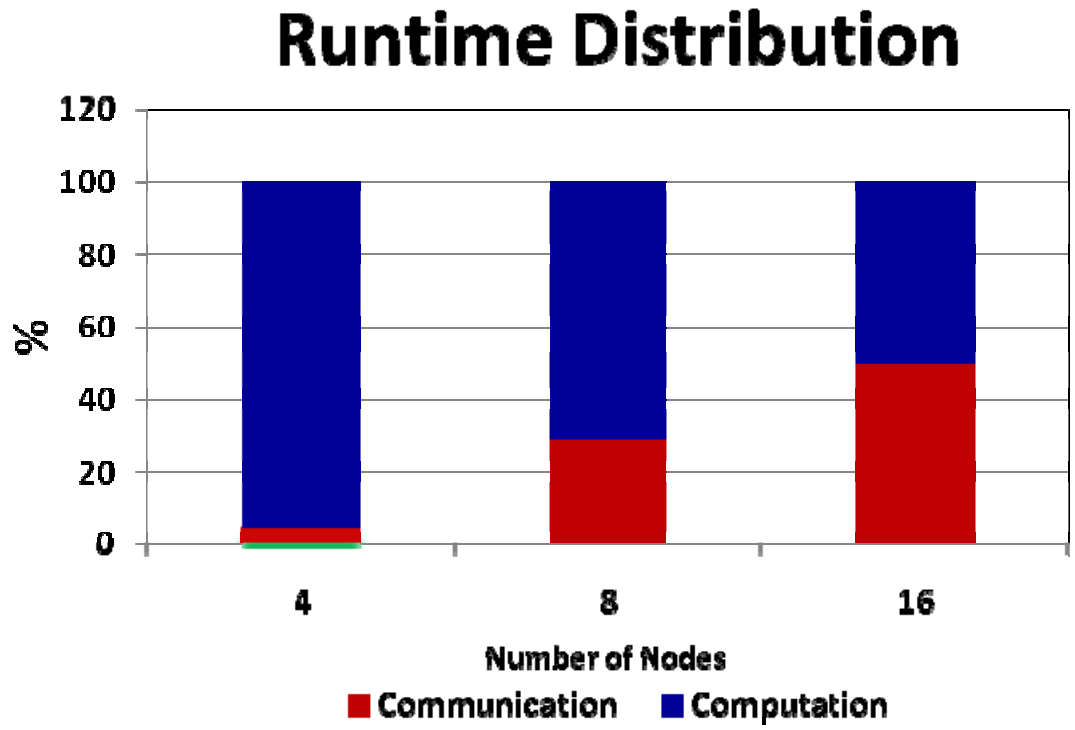
*$/KWh = KWh * $0.20*
For more information - http://enterprise.amd.com/Downloads/svrpwrusecompletefinal.pdf

Quantum ESPRESSO Benchmark Results Summary

- **Balanced system – CPU, memory, Interconnect that match each other capabilities - is essential for providing application efficiency**
- **Performance Optimization**
  - MPI libraries showed comparable performance overall
    - Intel MPI enables slightly higher performance
  - Lustre with IB delivers increased performance
  - Enabling multi-thread does not yield performance increase
- **Interconnect Characterization**
  - InfiniBand continues to deliver superior performance across a broad range of system sizes
  - GigE scalability is limited beyond 8 nodes
- **Power Analysis**
  - System architecture can yield nearly $9K annually in power savings

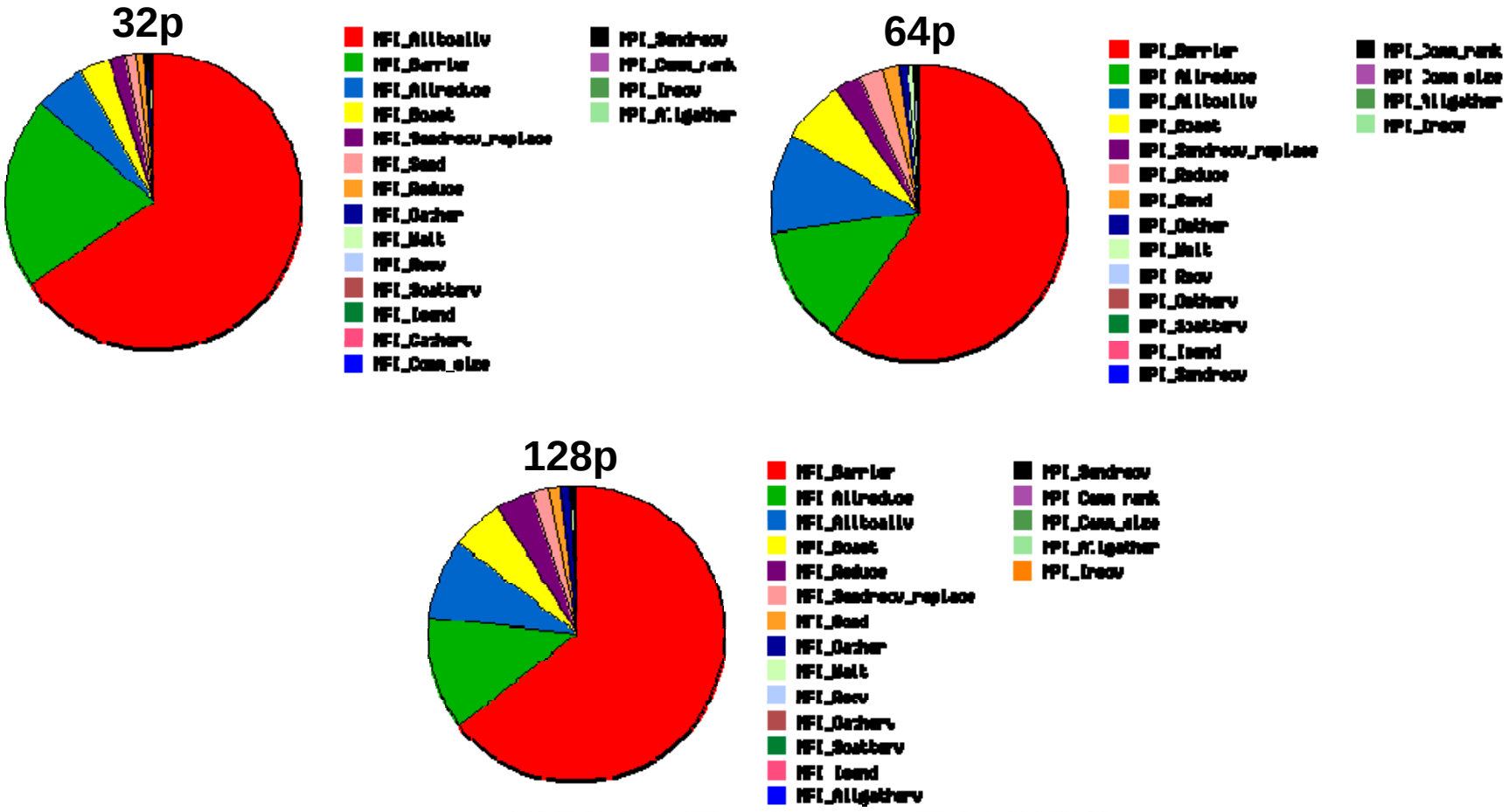
Quantum ESPRESSO Profiling – Runtime Distribution

- **Percentage of communication time increases as cluster size scales**
  - 5% at 32 processes, increases up to 50% at 128 processes



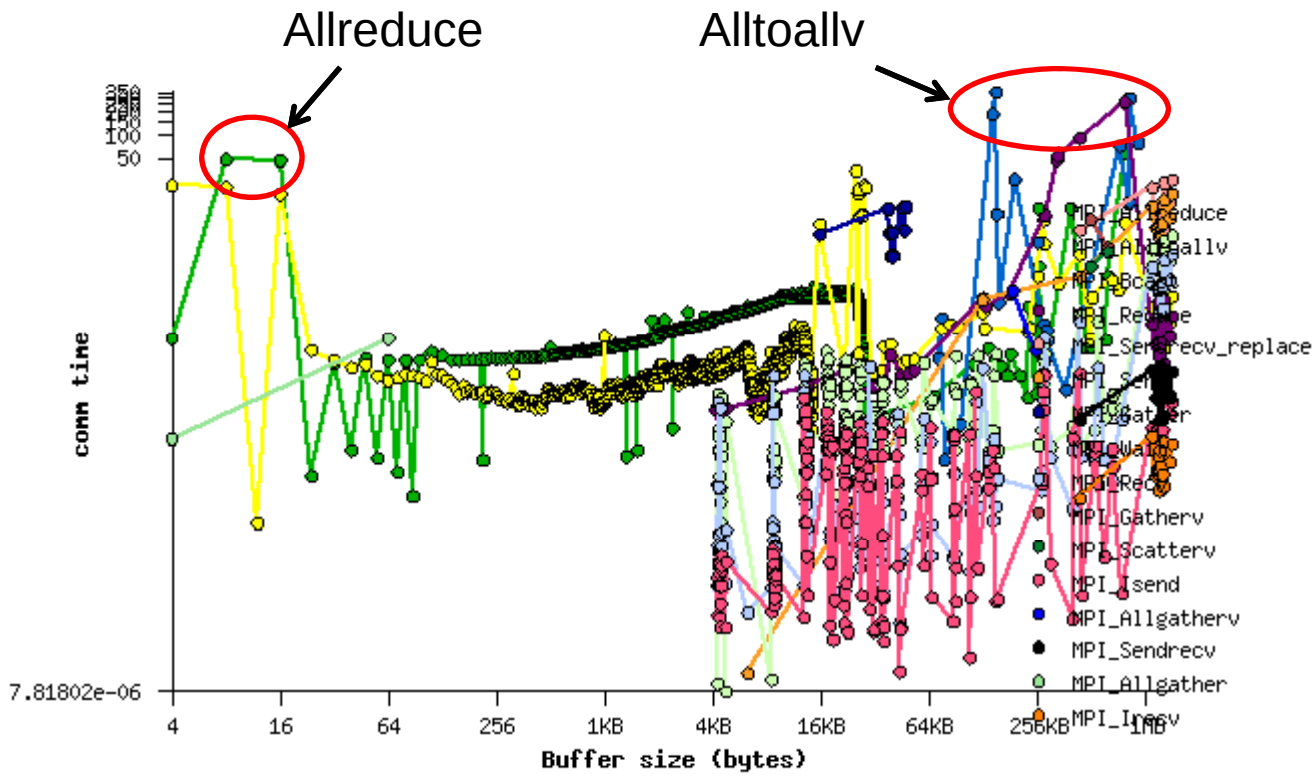
Quantum ESPRESSO Profiling - % of MPI Time

- **Three MPI collectives (MPI_Barrier, MPI_allreduce, and MPI_Alltoallv) consume more than 80% of total MPI time**



Quantum ESPRESSO Profiling - Message Size

- **Both small and large messages are creating communication overhead**
  - Most messages called by Barrier and Allreduce are small messages (<16B)
  - Alltoallv and Reduce messages are large size (>128KB)



128 processes

Quantum ESPRESSO Profiling Summary

- **Quantum ESPRESSO was profiled to identify its communication patterns**
- **Time in communication increases faster relative to computation**
- **MPI Collective functions dominate total MPI communication time**
  - More than 90% MPI time is spent in MPI collectives
  - Total number of messages increases with cluster size
- **Interconnects effect to Quantum ESPRESSO performance**
  - Both small and large messages are used by Quantum ESPRESSO
  - Interconnect latency and bandwidth are critical to application performance
- **Balanced system – CPU, memory, Interconnect that match each other capabilities, is essential for providing application efficiency**

Productive Systems = Balanced System

- **Balanced system enables highest productivity**
  - Interconnect performance to match CPU capabilities
  - CPU capabilities to drive the interconnect capability
  - Memory bandwidth to match CPU performance
- **Applications scalability relies on balanced configuration**
  - “Bottleneck free”
  - Each system components can reach it’s highest capability
- **Dell M610 system integrates balanced components**
  - Intel “Nehalem” CPUs and Mellanox InfiniBand QDR
    - Latency to memory and Interconnect latency at the same magnitude of order
  - Provide the leading productivity and power/performance system for Desmond simulations

Thank You

HPC Advisory Council