HPC ADVISORY COUNCIL

Quantum ESPRESSO Performance Benchmark and Profiling

December 2014

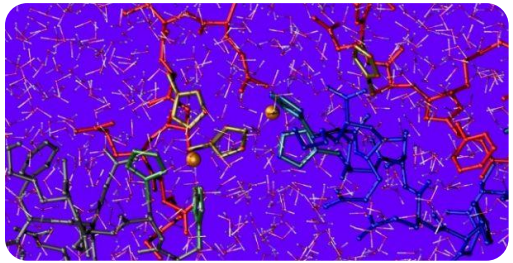
DELL intel Mellanox TECHNOLOGIES QUANTUMESPRESSO

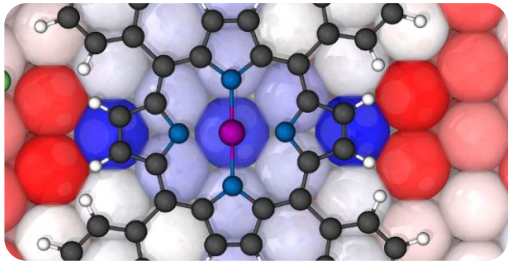
Note

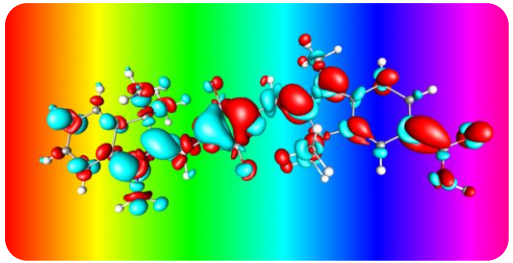
- **The following research was performed under the HPC Advisory Council activities**
  - Participating vendors: Intel, Dell, Mellanox
  - Compute resource - HPC Advisory Council Cluster Center
- **The following was done to provide best practices**
  - Quantum ESPRESSO performance overview
  - Understanding Quantum ESPRESSO communication patterns
  - Ways to increase Quantum ESPRESSO productivity
  - MPI libraries comparisons
- **For more info please refer to**
  - http://www.dell.com
  - http://www.intel.com
  - http://www.mellanox.com
  - http://www.quantum-espresso.org

Quantum ESPRESSO

- **Quantum ESPRESSO**
  - Stands for opEn Source Package for Research in Electronic Structure, Simulation, and Optimization
  - Is an integrated suite of computer codes for
    - electronic structure calculations
    - materials modeling at the nanoscale
  - 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 comparison
    - Interconnect performance comparison
    - CPUs comparison
    - Compilers comparison
- **The presented results will demonstrate**
  - The scalability of the compute environment/application
  - Considerations for higher productivity and efficiency

Test Cluster Configuration

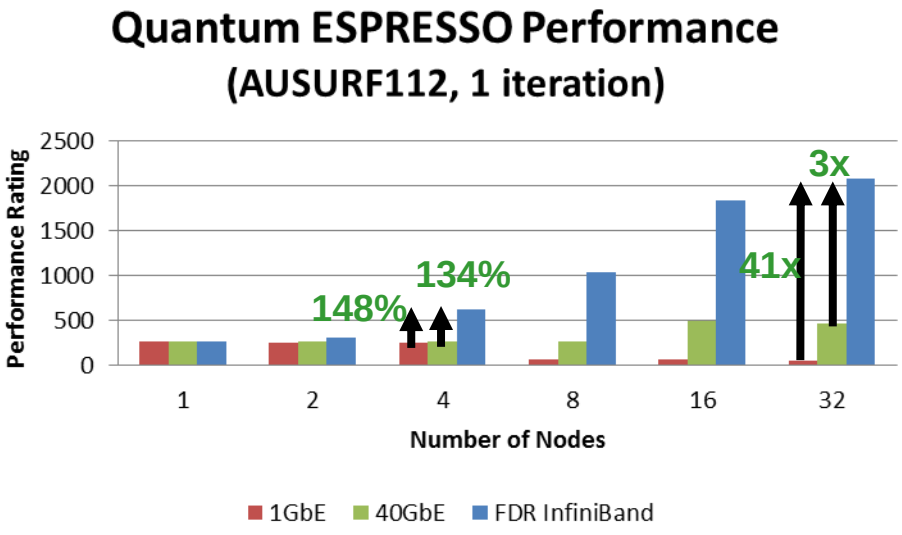
- **Dell PowerEdge R730 32-node (896-core) "Thor" cluster (Default cluster unless otherwise stated)**
  - Dual-Socket 14-Core Intel E5-2697v3 @ 2.60 GHz CPUs; Memory: 64GB memory, DDR4 2133 MHz
  - OS: RHEL 6.5, OFED 2.3-2.0.2 InfiniBand SW stack; Hard Drives: 2x 1TB 7.2 RPM SATA 2.5" on RAID 1
  - Memory Snoop Mode: Cluster-on-Die
- **Dell PowerEdge R720xd 32-node (640-core) "Jupiter" cluster**
  - Dual-Socket 10-Core Intel E5-2680v2 @ 2.80 GHz CPUs; Memory: 64GB memory, DDR3 1600 MHz
  - OS: RHEL 6.2, OFED 2.3-2.0.2 InfiniBand SW stack; Hard Drives: 24x 250GB 7.2 RPM SATA 2.5" on RAID 0
- **Interconnect**
  - Mellanox Connect-IB FDR InfiniBand adapters; Mellanox ConnectX-3 QDR InfiniBand and 40GbE VPI adapters
  - Mellanox SwitchX SX6036 VPI InfiniBand and Ethernet switches
- **MPI: Mellanox HPC-X v1.2.0-318, Intel MPI 5.0.2.044**
- **Compilers: Intel Composer XE 2015.1.133**
- **Application: Quantum ESPRESSO 5.1.1**
- **Benchmarks:**
  - Unified European Application Benchmark Suite (UEABS) DEISA pw benchmark Test Case A
    - AUSURF112 - Gold surface (112 atoms) DEISA pw benchmark

PowerEdge R730 and R730xd
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 GigE 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

Quantum ESPRESSO Performance - Interconnect

- **FDR InfiniBand is the most efficient network interconnects for Quantum ESPRESSO**
  - FDR IB outperforms by 134% vs 40GbE, and 148% vs 1GbE at 4 nodes (112 MPI cores)
  - The performance gap widen as higher core count
  - The “electron_maxstep = 1” is set in the ausurf.in to run for 1 iteration



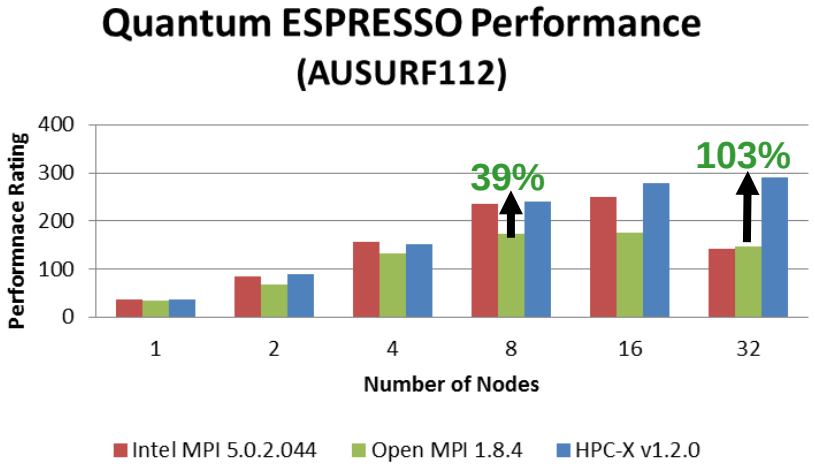
Higher is better

MPI Mode

28 Cores Per Node

Quantum ESPRESSO Performance – MPI Libraries

- **Both Intel MPI and HPC-X outperform Open MPI at low MPI process count**
  - HPC-X and Intel MPI outperform Open MPI by up to 36% at 8 nodes (224 cores)
  - DAPL is used for Intel MPI, MXM is used for the HPC-X
    - Flags for MXM: -mca pml yalla -x MXM_TLS=self,shm,ud -x MALLOC_MMAP_MAX_=0 -x MALLOC_TRIM_THRESHOLD_=-1
- **HPC-X delivers the best performance over Intel MPI and Open MPI at higher scale**
  - The performance levels off at high MPI core counts; workload is too small to scale to many MPI processes
  - Use Hybrid (OpenMP-MPI) instead of MPI in Quantum ESPRESSO for better performance at higher core counts



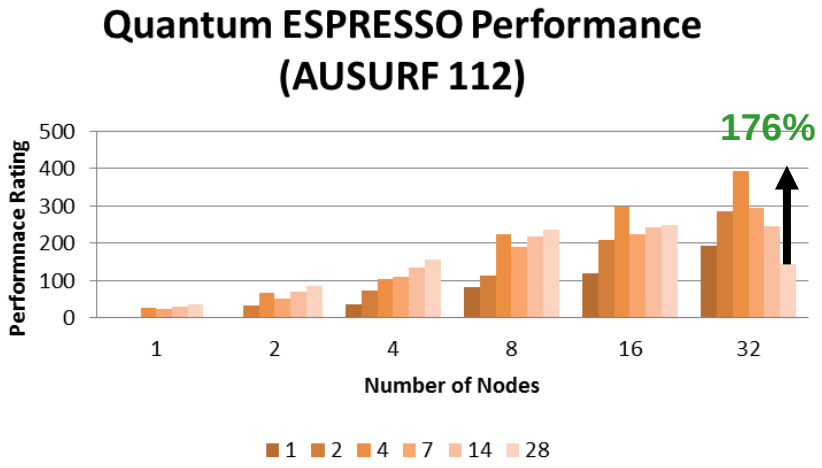
Higher is better

MPI Mode

28 Cores Per Node

Quantum ESPRESSO Performance – Hybrid vs MPI

- **Hybrid mode scales while MPI mode performs better at low core counts**
  - MPI mode runs better at low core counts (~224 cores/8 nodes)
  - Hybrid (OpenMP-MPI) allows Quantum ESPRESSO to scale (896 cores/32 nodes)
  - Best case is to use 4 PPN with 7 threads at large scale
  - The MPI mode runs with 28 MPI processes per node (PPN)
  - The rest of the cases were run in hybrid mode. E.g. in the 4 PPN case, each MPI rank spawns 7 thread



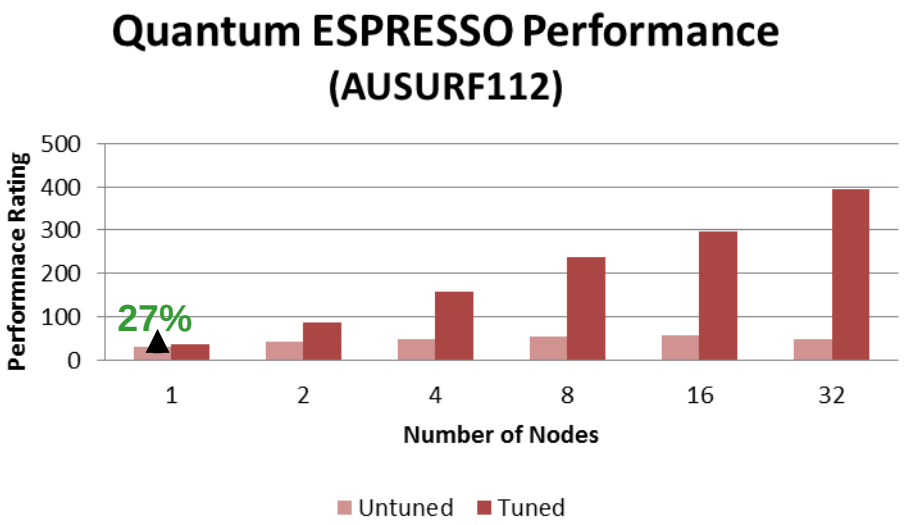
Higher is better

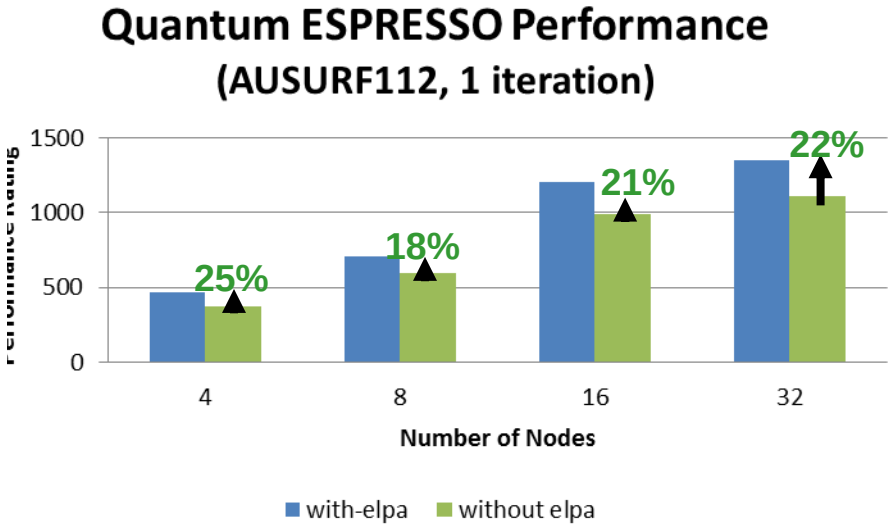
Hybrid & MPI Modes

28 Cores Per Node

Quantum ESPRESSO Performance – Compiler Tuning

- **Using compiler tuning improves performance over default options**
  - Tuning includes: Enabling ELPA, ScalaPACK, and OpenMP
  - Compiler Flags: -O3 -xCORE-AVX2 -fno-alias -ansi-alias -g -mkl -openmp
  - "Eigenvalue Solvers for Petaflop-Architectures" (ELPA) library provides ~25% improvement



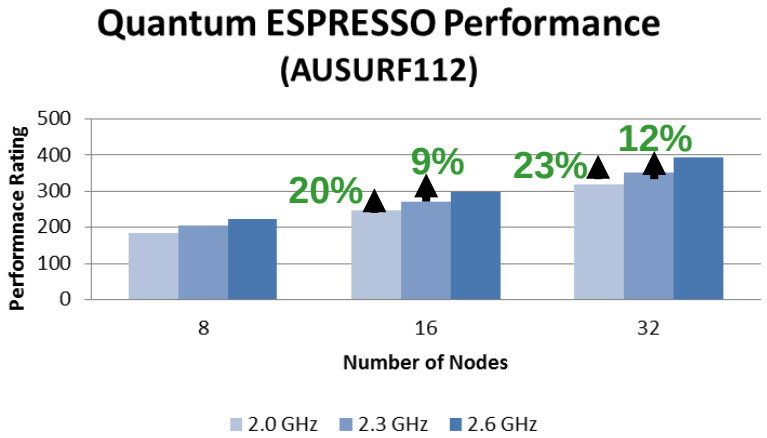


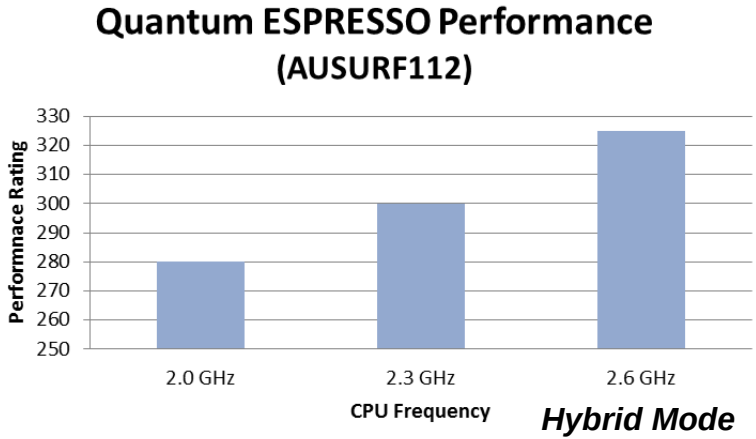
Higher is better

28 Cores Per Node

Quantum ESPRESSO Performance – Core Frequency

- **Running at high CPU clock rate allows good improvement**
  - Up to 23% higher performance from 2 GHz to 2.6 GHz
  - Up to 12% higher performance from 2.3 GHz to 2.6 GHz
  - Turbo mode is turned off throughout these tests





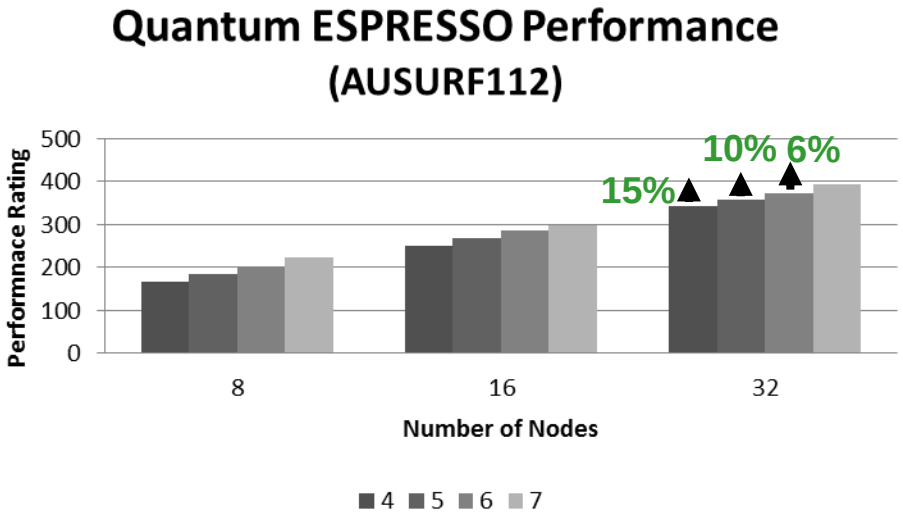
Hybrid Mode

28 Cores Per Node

Higher is better

Quantum ESPRESSO Performance – Threads/Node

- **Running more available threads cores adds more marginal performance**
  - Hybrid (OpenMP-MPI) cases shown: 4 MPI processes, each MPI process spawns # of threads
  - The number of threads per MPI process is indicated in the legend on the chart
  - If 28 CPU cores available per node, it adds ~15% of additional performance than 16 cores per node
  - Used to estimate the effects of using processors with less available cores



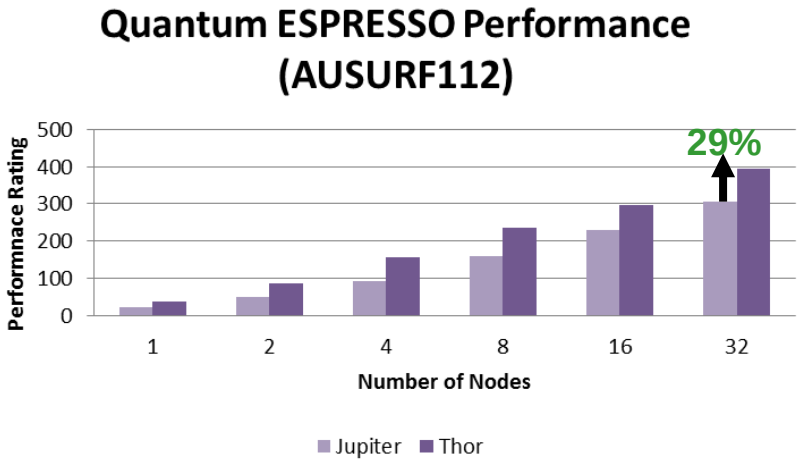
Hybrid OpenMP-MPI

Hybrid Mode

4 MPI Processes

Quantum ESPRESSO Performance – CPU Generation

- **Intel E5-2697v3 (Haswell) cluster outperforms prior CPU generation**
  - Performs ~29% higher than E5-2680v2 (Ivy Bridge) Jupiter cluster (on a per node basis)
- **System components used:**
  - Jupiter: R720: 2-socket 10c E5-2680v2 @ 2.8GHz, 1600MHz DIMMs, FDR IB
  - Thor: R730: 2-socket 14c E5-2697v3 @ 2.6GHz, 2133MHz DIMMs, FDR IB



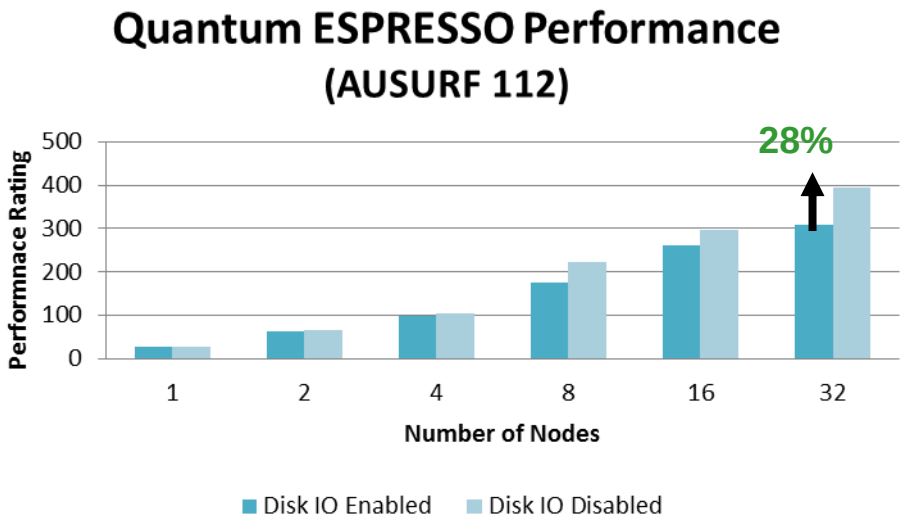
Higher is better

Hybrid Mode

Intel MPI

Quantum ESPRESSO Performance – Disk IO

- **Disabling IO in input data shows the effect of disk IO**
  - ~28% of the performance improvement seen with IO disabled at 32 nodes / 896 cores
  - Disabling I/O activities by using the disk_io = 'none‘
  - The performance differences increase at scale



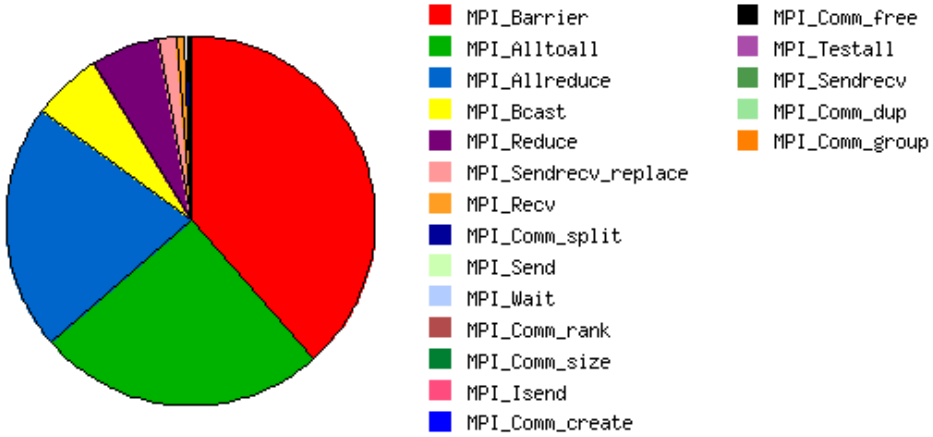
Higher is better

Hybrid Mode

28 Cores Per Node

Quantum ESPRESSO Profiling – MPI calls

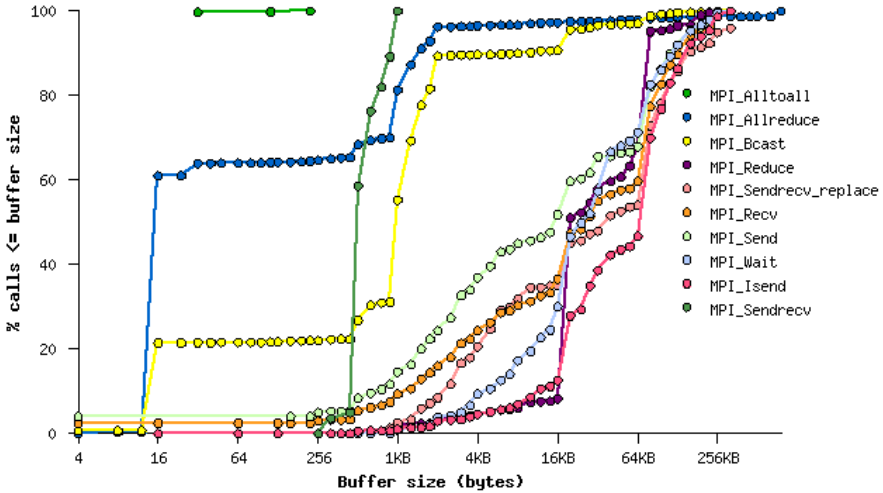
- **Quantum ESPRESSO shows high usage for MPI collective operations:**
  - Majority of the communications involve collective operations, such as:
  - MPI_Barrier (38%), MPI_Alltoall (24%), MPI_Allreduce (20%)



MPI Mode

Quantum ESPRESSO Profiling – Time Spent in MPI

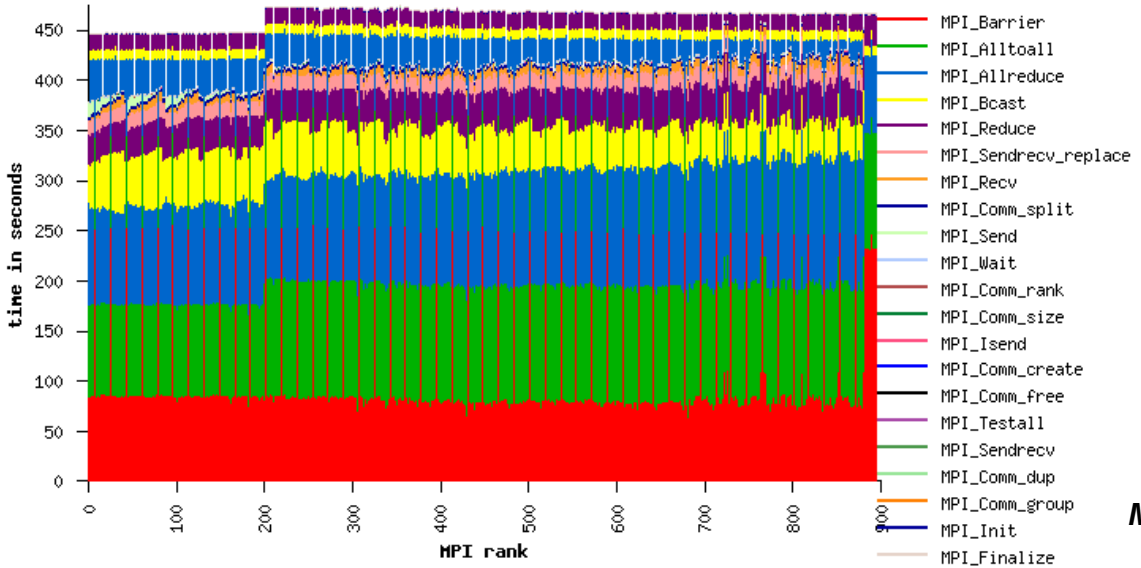
- **Quantum ESPRESSO shows high usage for collective operations**
  - MPI_Barrier at 0B (~38% of MPI time)
  - MPI_Alltoall at 32B (~32% of MPI time)
  - MPI_Allreduce at 786KB (15% of MPI time)



MPI Mode

Quantum ESPRESSO Profiling – Time Spent in MPI

- **Quantum ESPRESSO: More time spent on MPI collective operations:**
  - Some imbalance in work load which causes communication imbalances



MPI Mode

Quantum ESPRESSO Summary

- **Performance of Quantum ESPRESSO can be improved significantly through multiple level tuning**
  - Compiler tuning:
    - Enabling compiler optimization, ELPA and SCALAPACK libraries can improve scalability performance
  - Runtime:
    - Running Quantum ESPRESSO with hybrid mode unlocks scalability performance for beyond ~8 nodes/224 cores
  - MPI:
    - HPC-X provides the highest scalability at 32 nodes (896 cores) and deliver up to 39% over Open MPI at 8 nodes (224 cores)
  - Network Interconnect:
    - InfiniBand FDR is the most efficient cluster interconnect for Quantum ESPRESSO
  - CPU:
    - The additional CPU cores and higher CPU clock frequency can yield additional performance
  - The latest generation of servers outperform previous generation of servers
    - Provided up to 29% higher performance on a per node basis

Thank You

HPC Advisory Council