



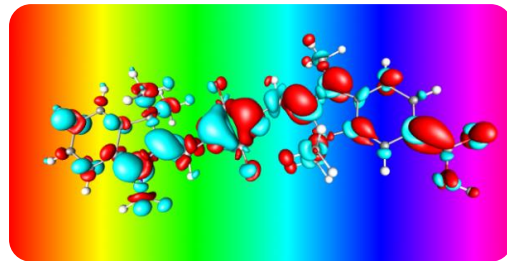
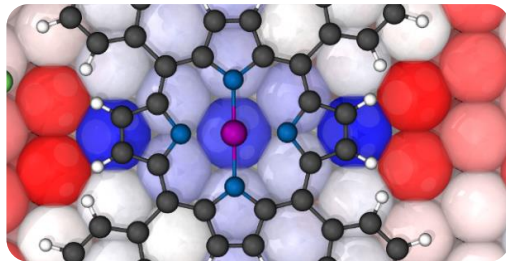
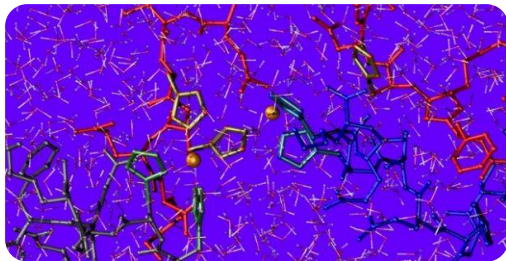
Quantum ESPRESSO

Performance Benchmark and Profiling

February 2017

- **The following research was performed under the HPC Advisory Council activities**
 - 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
- **For more info please refer to**
 - <http://www.quantum-espresso.org/>

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

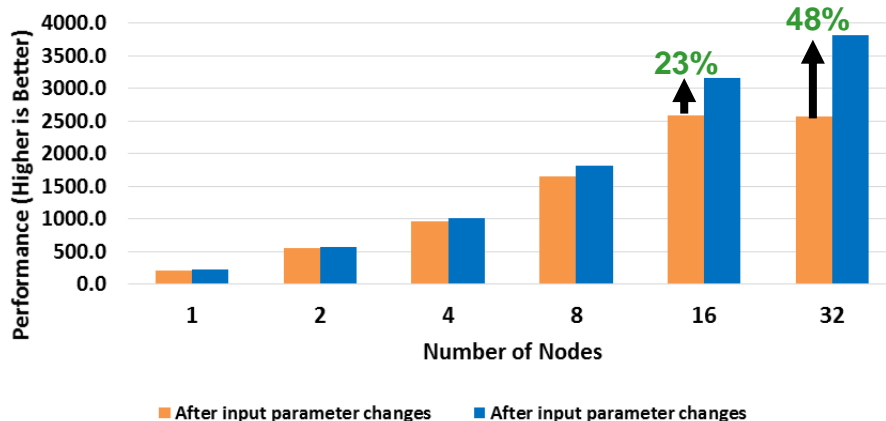
Test Cluster Configuration

- **Dell PowerEdge R730 32-node (1024-core) “Thor” cluster**
 - Dual-Socket 16-Core Intel E5-2697Av4 @ 2.60 GHz CPUs (BIOS: Maximum Performance, Home Snoop)
 - Memory: 256GB memory, DDR4 2400 MHz, Memory Snoop Mode in BIOS sets to Home Snoop
 - OS: RHEL 7.2, MLNX_OFED_LINUX-3.4-2.0.0.0 InfiniBand SW stack
- **Mellanox ConnectX-4 EDR 100Gb/s InfiniBand Adapters**
- **Mellanox Switch-IB SB7800 36-port EDR 100Gb/s InfiniBand Switch**
- **Intel® Omni-Path Host Fabric Interface (HFI) 100Gb/s Adapter**
- **Intel® Omni-Path Edge Switch 100 Series**
- **Dell InfiniBand-Based Lustre Storage based on Dell PowerVault MD3460 and Dell PowerVault MD3420**
- **MPI: Mellanox HPC-X MPI Toolkit v1.7**
- **MPI Profiler: Alinea MAP 6.1.2**
- **Application: Quantum ESPRESSO 5.4 with ELPA support**
- **Benchmarks:**
 - AUSURF112 - Gold surface (112 atoms) DEISA pw benchmark. electron_maxstep is set to 1 for single step

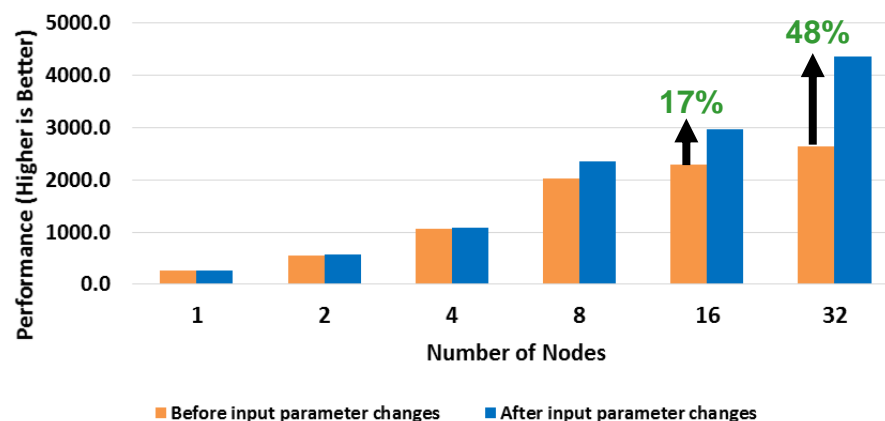
Quantum ESPRESSO Performance – Input Parameters

- **Tuning the input parameters has a positive impact to scalability and parallelization**
 - Up to 48% performance increase at 32 nodes / 1024 cores, for both 28PPN and 32PPN
- **Input parameter changes:**
 - -npools 2 -ntg 4 -ndiag 1024 (for 1024 cores)
 - Instructions for Quantum ESPRESSO set up can be found at the [HPCAC HPC|Works](#) section

Quantum ESPRESSO Performance
(ausurf112, 28PPN)

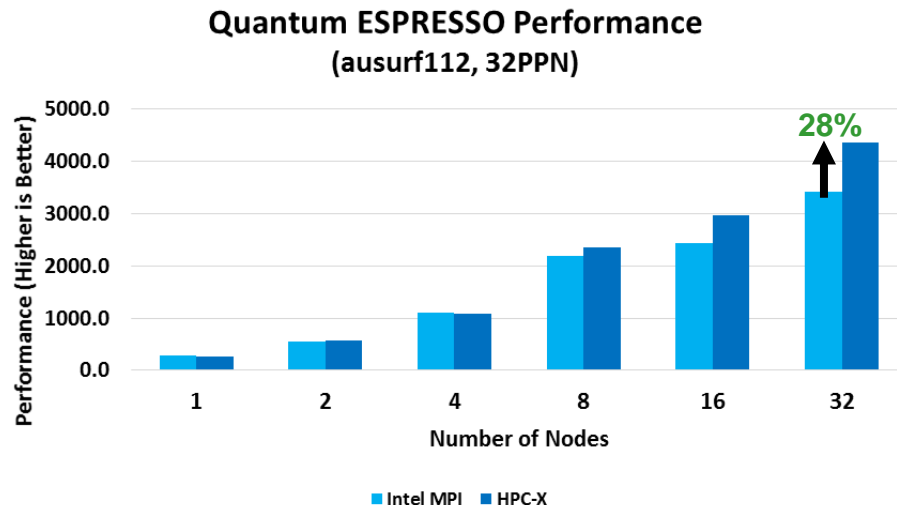


Quantum ESPRESSO Performance
(ausurf112, 32PPN)



Higher is better

- **HPC-X delivers higher performance at higher scale**

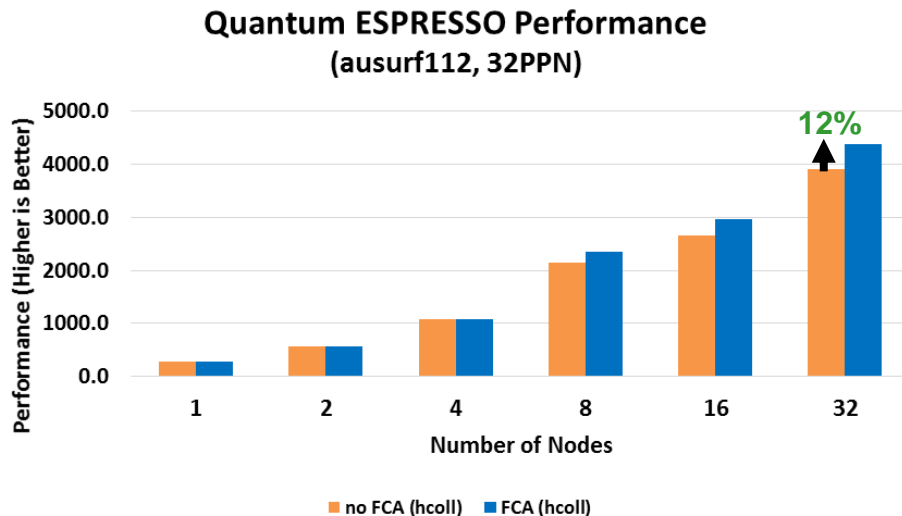


Higher is better

Optimized parameters used

Quantum ESPRESSO Performance - Collective Offloads

- **Optimized collective operations deliver higher scalability**
 - Performance gain by 12% at 32 nodes when FCA (hcoll) is used
 - Due to time increase for collective operations at larger core counts
 - As seen in MPI profiles: MPI_Allreduce, MPI_Barrier, etc

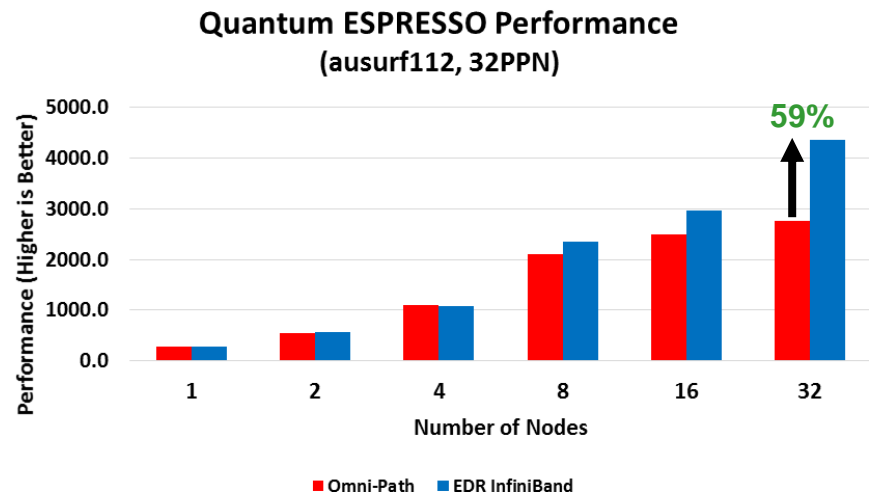
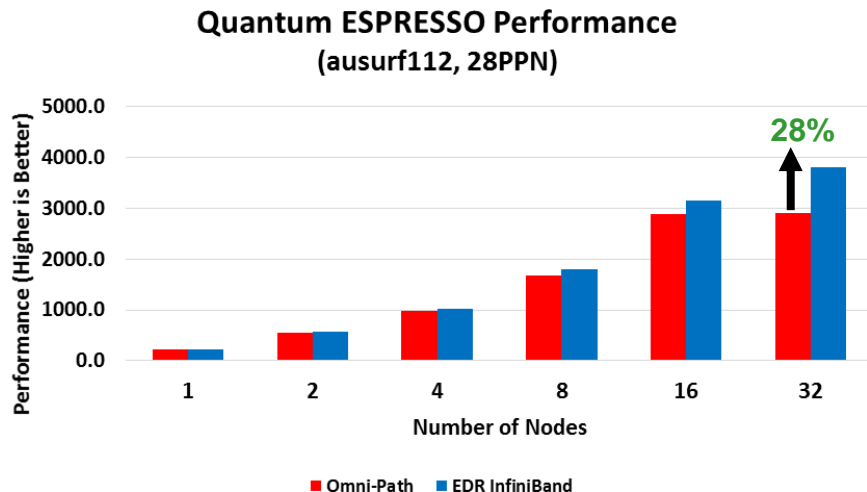


Higher is better

Optimized parameters used

Quantum ESPRESSO Performance –Interconnects

- EDR InfiniBand enables higher performance at various server PPN

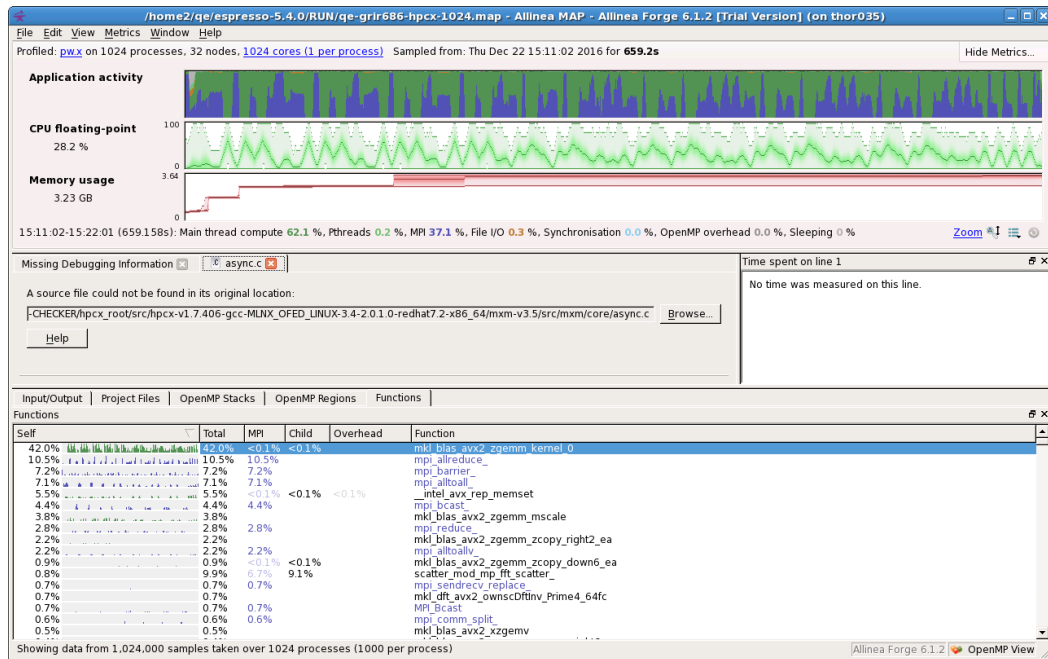


Higher is better

Optimized parameters used


Quantum ESPRESSO Profiling – Alinea MAP

- **Analysis from Alinea MAP indicates heavy time utilization in BLAS routine at 32 nodes**
 - mkl_blas_avx2_zgemm_kernel (42%)
 - MPI communications:
 - MPI_Allreduce (10.5%)
 - MPI_Barrier (7.2%)
 - MPI_Alltoall (7.1%)
 - MPI_Bcast (4.4%)



- Analysis from Allinea Performance Report

Summary: pw.x is **Compute-bound** in this configuration

Compute 62.2% 

Time spent running application code. High values are usually good.
This is **average**; check the CPU performance section for advice.

MPI 37.2% 

Time spent in MPI calls. High values are usually bad.
This is **average**; check the MPI breakdown for advice on reducing it.

I/O 0.6% 

Time spent in filesystem I/O. High values are usually bad.
This is **very low**; however single-process I/O may cause MPI wait times.

This application run was **Compute-bound**. A breakdown of this time and advice for investigating further is in the **CPU** section below.

- Analysis from Alinea Performance Report

CPU

A breakdown of the 62.2% CPU time:

Single-core code	100.0%	<div><div></div></div>
OpenMP regions	0.0%	<div><div></div></div>
Scalar numeric ops	3.2%	<div><div></div></div>
Vector numeric ops	46.3%	<div><div></div></div>
Memory accesses	45.0%	<div><div></div></div>

The CPU performance appears well-optimized for **numerical computation**. The biggest gains may now come from running at larger scales.

Significant time is spent on **memory accesses**. Use a profiler to identify time-consuming loops and check their cache performance.

MPI

A breakdown of the 37.2% MPI time:

Time in collective calls	95.0%	<div><div></div></div>
Time in point-to-point calls	5.0%	<div><div></div></div>
Effective process collective rate	207 MB/s	<div><div></div></div>
Effective process point-to-point rate	533 MB/s	<div><div></div></div>

Most of the time is spent in **collective calls** with an average transfer rate. Using larger messages and overlapping communication and computation may increase the effective transfer rate.


- **Analysis from Allinea Performance Report**


I/O

A breakdown of the 0.6% I/O time:

Time in reads 2.9% |

Time in writes 97.1% 

Effective process read rate 86.1 MB/s 


Effective process write rate 225 MB/s 


Most of the time is spent in write operations with an average effective transfer rate. It may be possible to achieve faster effective transfer rates using asynchronous file operations.

Memory

Per-process memory usage may also affect scaling:

Mean process memory usage 3.01 GiB 

Peak process memory usage 3.39 GiB 

Peak node memory usage 40.0% 

The peak node memory usage is low. Running with fewer MPI processes and more data on each process may be more efficient.

- **Tuning Quantum ESPRESSO for better performance**
 - Adjusting input parameters provides 48% higher performance; improve scalability and parallelization
 - Instructions for Quantum ESPRESSO set up can be found at the HPCAC HPC|Works section
 - Optimized collective operations in FCA (hcoll) provides 12% increase in cluster scalability
 - IB demonstrates better utilization of network, partly due to better collectives offload in EDR
- **Network: EDR InfiniBand enables higher performance**
 - 28%-59% performance advantage at 32 nodes / 1024 cores
- **MPI Library: HPC-X delivers higher performance**
 - Up to 28% higher performance at 32 nodes (1024 cores)
- **MPI Profiling from Allinea MAP**
 - Analysis from Allinea MAP indicates heavy time utilization in BLAS routine at 32 nodes
 - mkl_blas_avx2_zgemm_kernel (42%)
 - MPI communications: MPI_Allreduce (10.5%), MPI_Barrier (7.2%), MPI_Alltoall (7.1%). MPI_Bcast (4.4%)

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 undertakes no duty and assumes no obligation to update or correct any information presented herein