HPC
ADVISORY COUNCIL

CP2K
Performance Benchmark and Profiling

April 2011

DELL intel Mellanox TECHNOLOGIES

Note

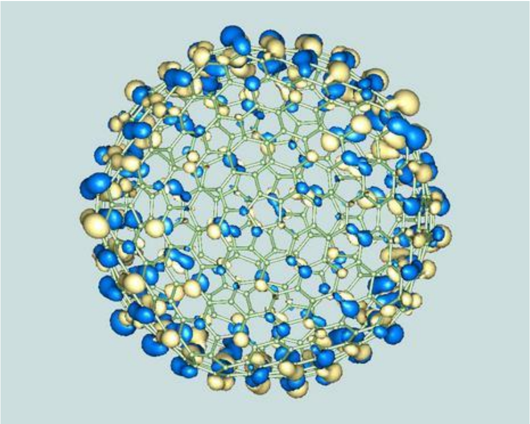
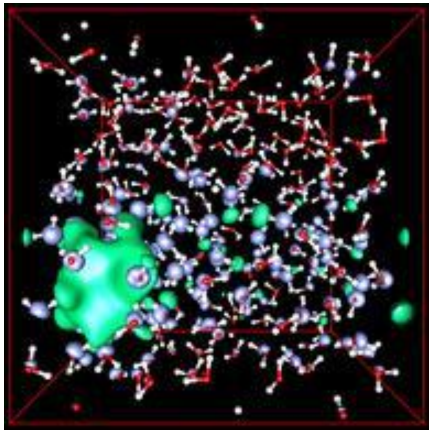
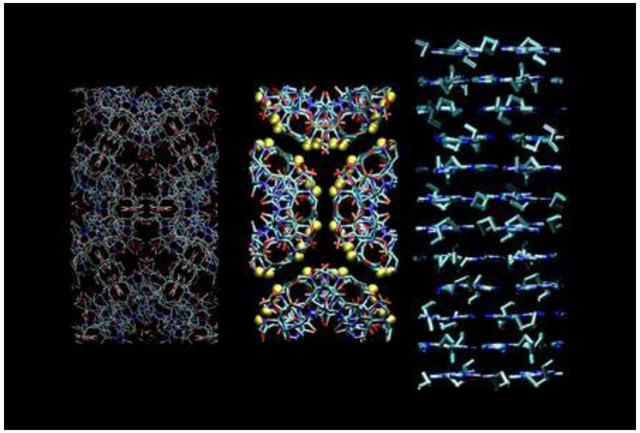
- **The following research was performed under the HPC Advisory Council activities**
  - Participating vendors: Intel, Dell, Mellanox
  - Compute resource - HPC Advisory Council Cluster Center

- **The following was done to provide best practices**
  - CP2K performance overview
  - Understanding CP2K communication patterns
  - Ways to increase CP2K productivity
  - MPI libraries comparisons

- **For more info please refer to**
  - http://www.dell.com
  - http://www.intel.com
  - http://www.mellanox.com
  - http://cp2k.berlios.de

CP2K Application

- **CP2K is an atomistic and molecular simulations software for solid state, liquid, molecular and biological systems**
- **CP2k provides a general framework for different methods, such as:**
  - Density functional theory (DFT) using a mixed Gaussian and plane waves approach (GPW)
  - Classical pair and many-body potentials
- **CP2K is a freely available (GPL) program, written in Fortran 95**

Test Cluster Configuration

- **Dell™ PowerEdge™ M610 38-node (456-core) cluster**
  – Six-Core Intel X5670 @ 2.93 GHz CPUs
  – Six-Core Intel X5675 @ 3.06 GHz CPUs
  – Memory: 24GB memory, DDR3 1333 MHz
  – OS: RHEL 5.5, OFED 1.5.2 InfiniBand SW stack
- **Intel Cluster Ready certified cluster**
- **Mellanox ConnectX-2 InfiniBand adapters and non-blocking switches**
- **Compiler: Intel Compiler 11.1**
- **MPI: Intel MPI 4, Open MPI 1.5.3, Platform MPI 8.0.1**
- **Libraries: Intel MKL 10.1, FFTW3, BLACS, ScaLAPACK 1.8.0, LAPACK 3.3**
- **Application: CP2K version 2.2.188 (Development Version)**
- **Benchmark dataset: H2O-128.inp**

About Intel® Cluster Ready

- **Intel® Cluster Ready systems make it practical to use a cluster to increase your simulation and modeling productivity**
  – Simplifies selection, deployment, and operation of a cluster

- **A single architecture platform supported by many OEMs, ISVs, cluster provisioning vendors, and interconnect providers**
  – Focus on your work productivity, spend less management time on the cluster

- **Select Intel Cluster Ready**
  – Where the cluster is delivered ready to run
  – Hardware and software are integrated and configured together
  – Applications are registered, validating execution on the Intel Cluster Ready architecture
  – Includes Intel® Cluster Checker tool, to verify functionality and periodically check cluster health

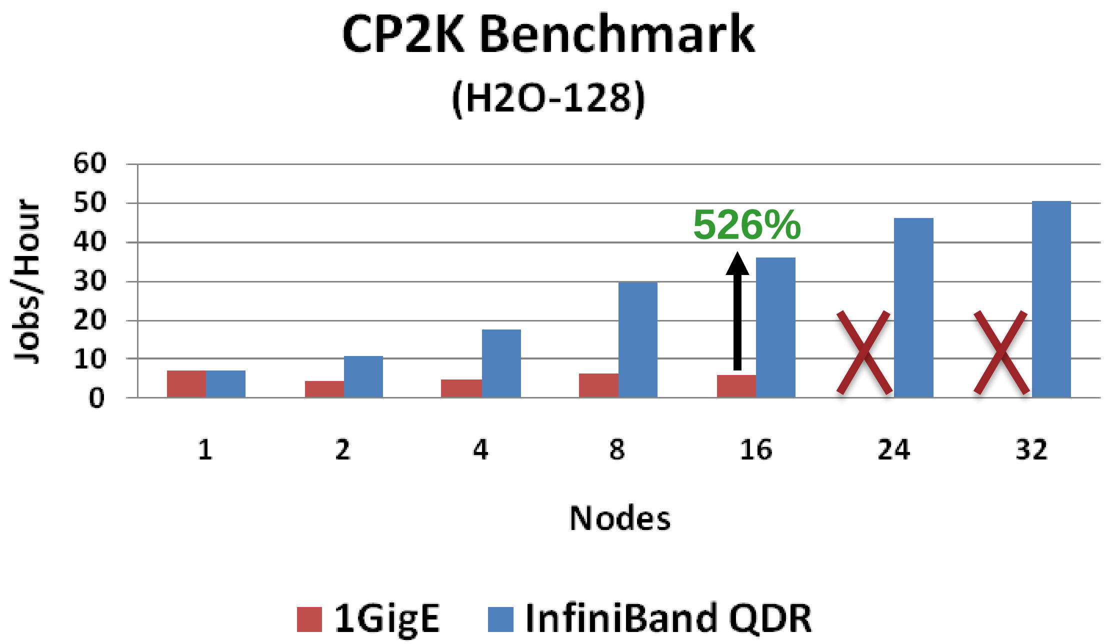
About Dell PowerEdge Servers

- **System Structure and Sizing Guidelines**
  - 38-node cluster build with Dell PowerEdge™ M610 blade 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

CP2K Performance – Interconnect

- **InfiniBand enables higher throughput and cluster productivity**
  - Provides up to 526% gain in job productivity over GigE on a 16-node cluster
    - GigE testing is limited to 16-node due to switch port availability
  - GigE shows virtually no gain in job productivity through all node counts tested

- **Comparison to other interconnect options was not available with the system configurations. Profiling in later slides should provide input data for interconnect performance estimations**

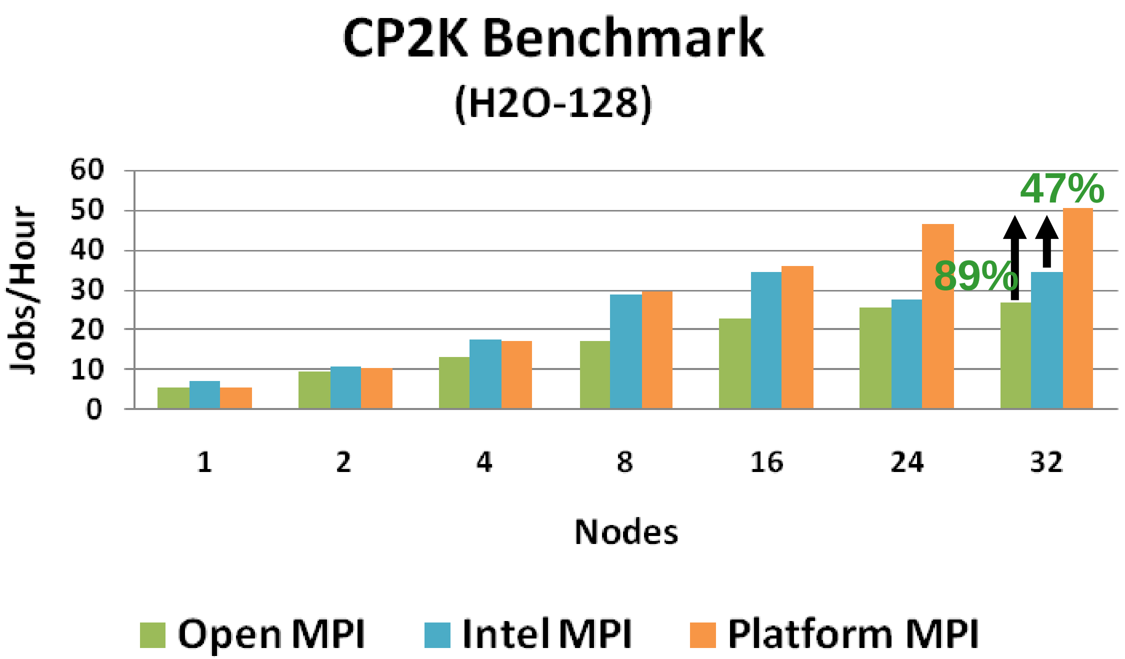


Higher is better

InfiniBand QDR

CP2K Performance – MPI

- **Platform MPI shows superior performance in high scalability**
  - Runs 89% more jobs compared to Open MPI at 32-node
  - Runs 47% more jobs compared to Intel MPI at 32-node
- **Dataset contains wide range of MPI calls**
  - Reflects that Platform MPI implements optimal performance for a range of MPI calls

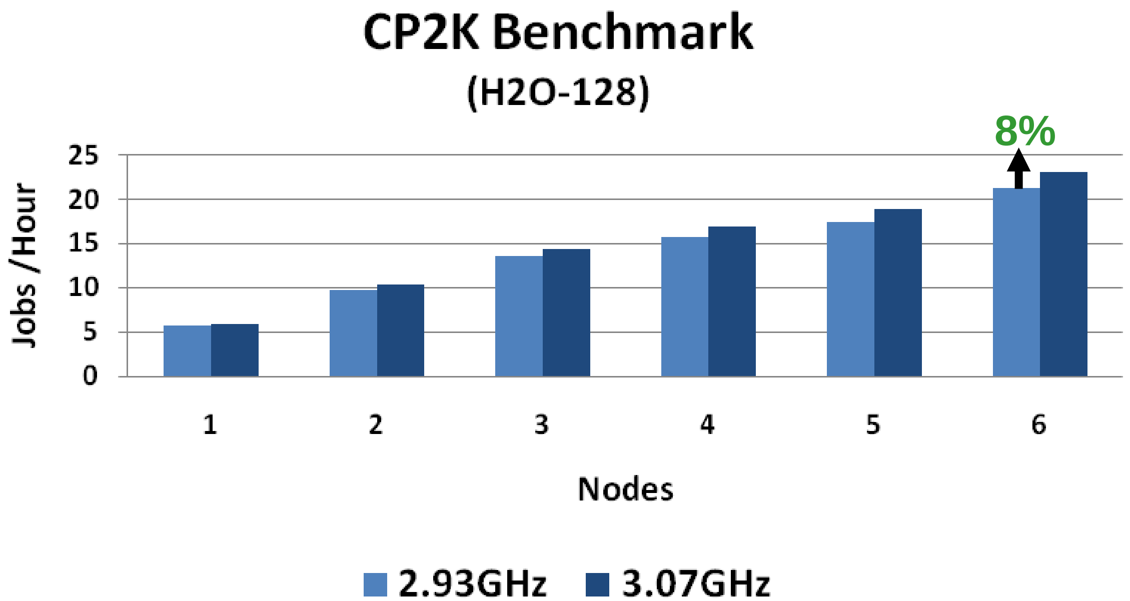


Higher is better

InfiniBand QDR

CP2K Performance – CPU Frequencies

- **Higher CPU frequency provides higher performance**
  - Seen a 6-8% in work improvement by using CPUs with 3.07GHz vs 2.93GHz

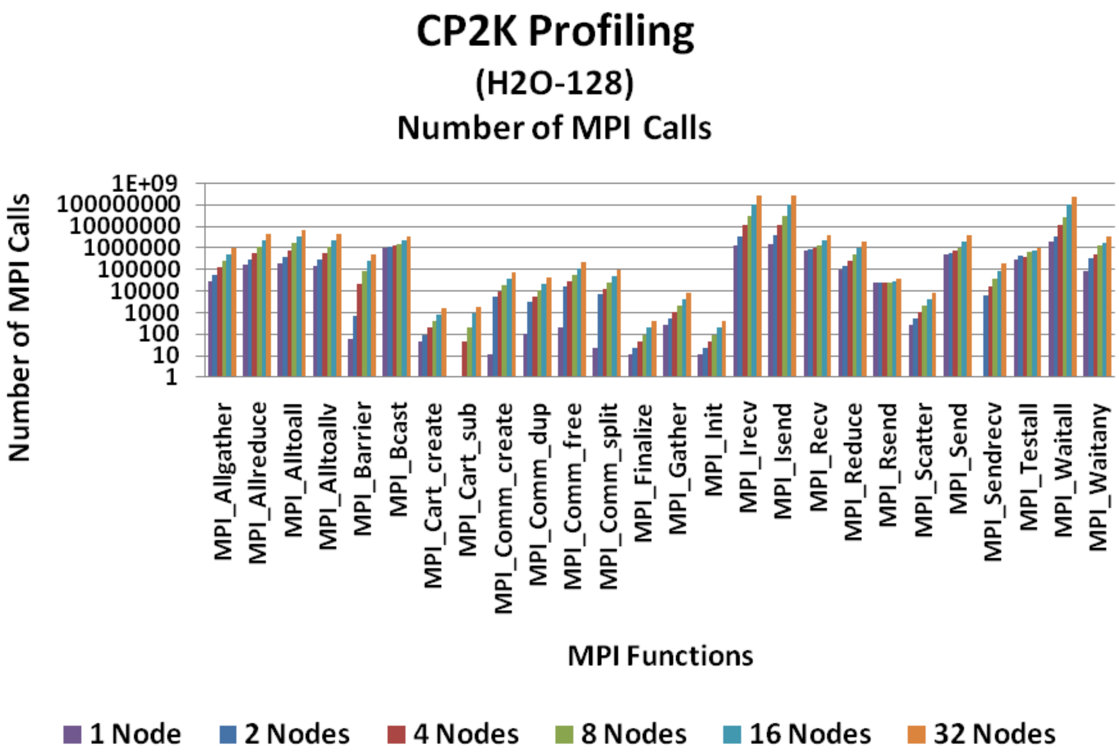


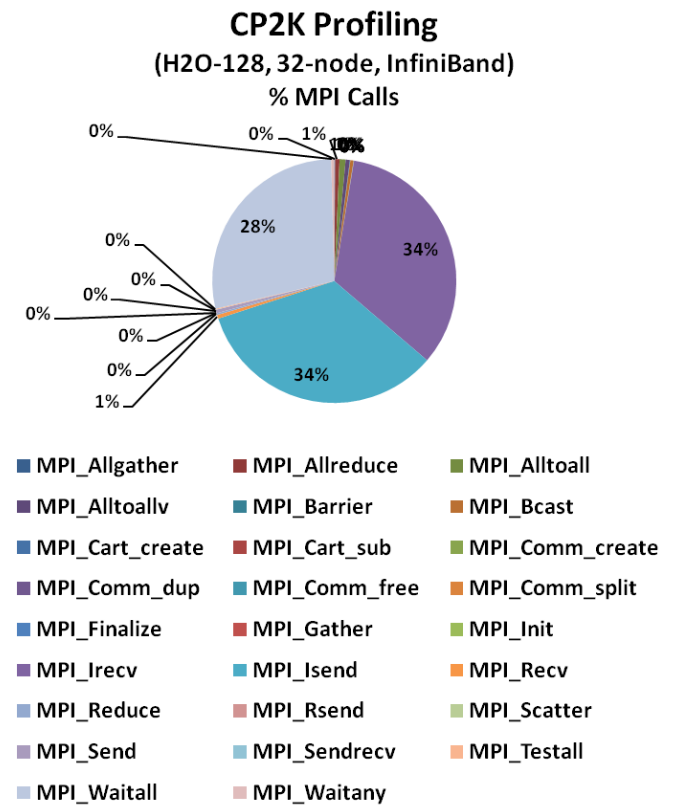
Higher is better

InfiniBand QDR

CP2K Profiling – Number of MPI Calls

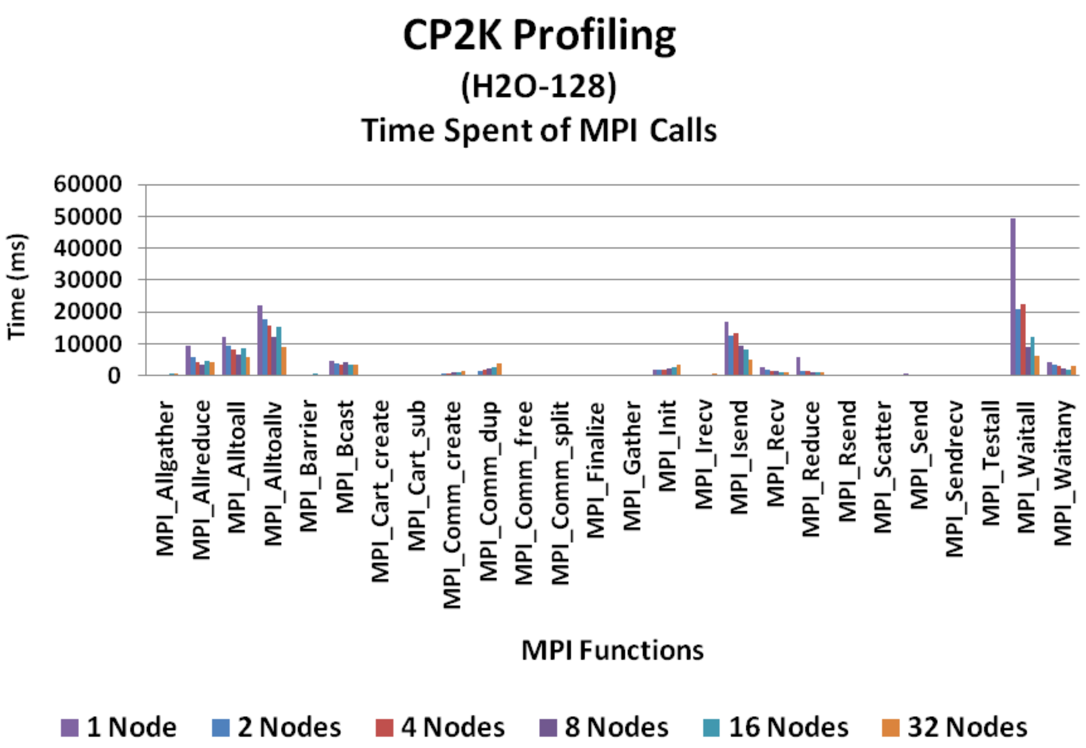
- **CP2K utilizes a wide range of MPI APIs**
  - 26 MPI APIs used in total
  - Shows a heavy use in MPI collectives and non-blocking point-to-point MPI APIs
- **MPI_Isend and MPI_Irecv are almost used exclusively at scale**
  - Each of these MPI functions is accounted for 34% of all MPI functions at 32-node
  - MPI_Waitall represents 28% of all MPI calls at 32-node

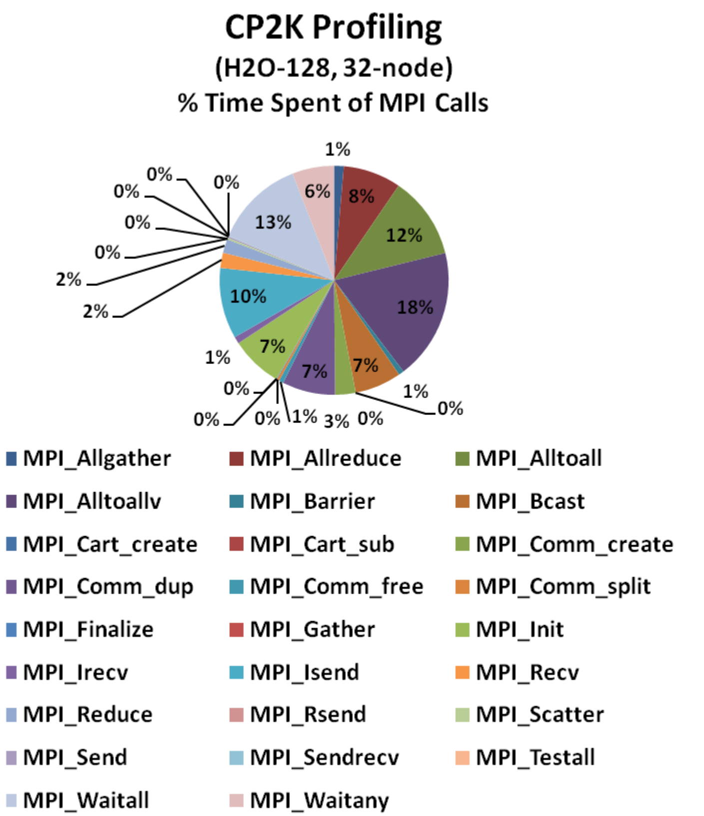




CP2K Profiling – Time Spent by MPI

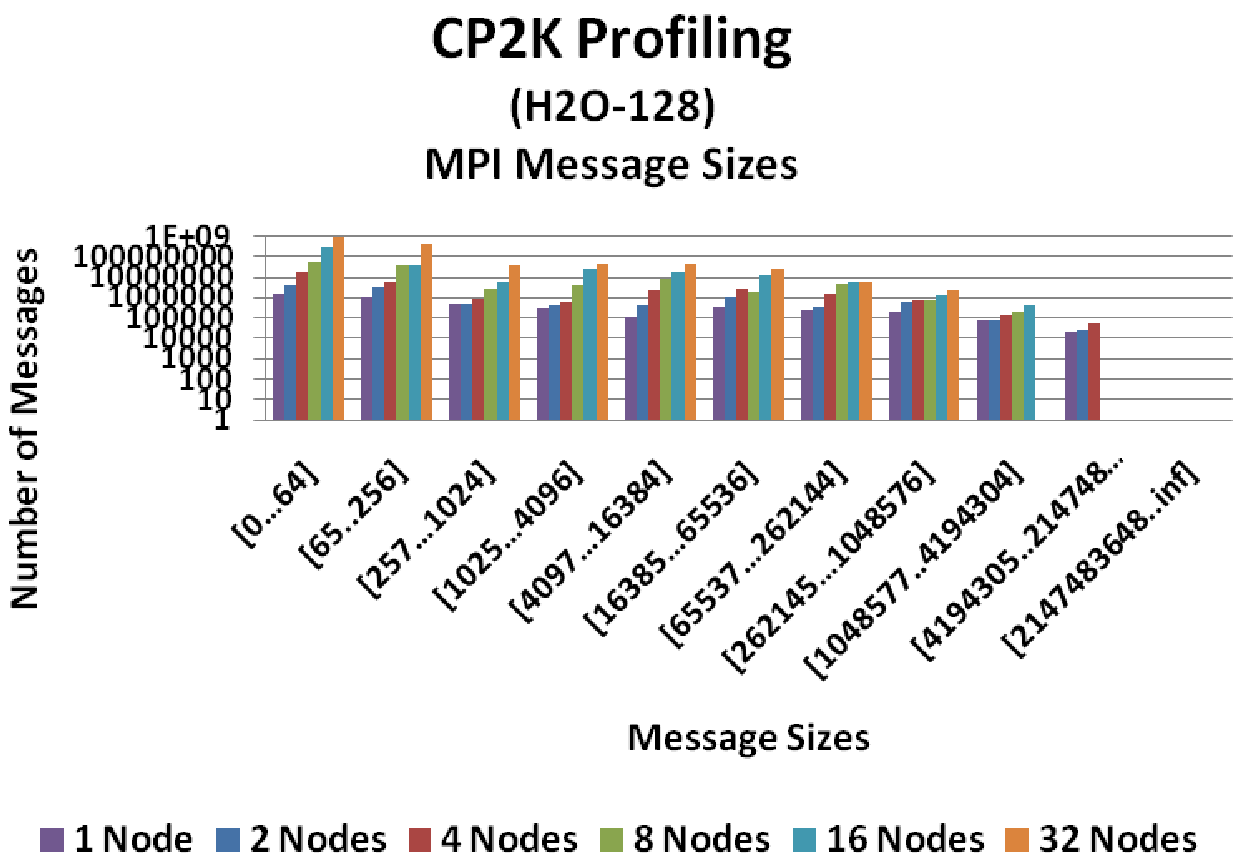
- **MPI collectives the biggest time consumer at 32 node**
  - MPI_Alltoallv(18%), MPI_Waitall(13%), MPI_Alltoall(12%), MPI_Allreduce(8%)
  - MPI non-blocking send: MPI_Isend(8%)
- **Time spent by MPI_Waitall reduces as node count increases**





CP2K Profiling – MPI Message Size

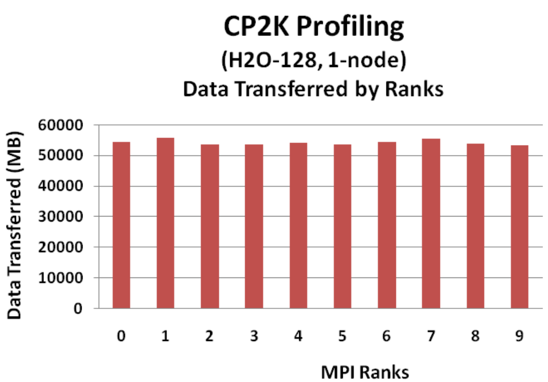
- **Majority of MPI messages are small messages**
  - In the range of 0 to 256 bytes

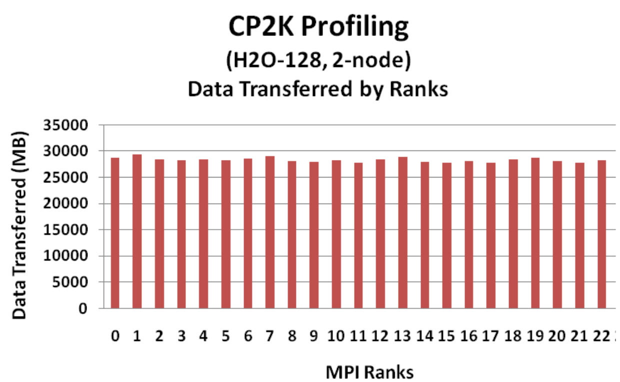


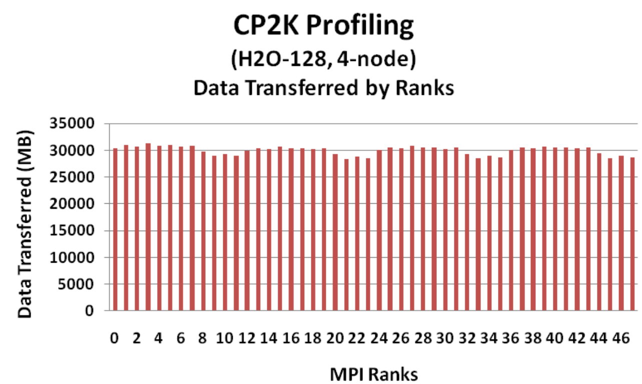
CP2K Profiling – MPI Data Transfer

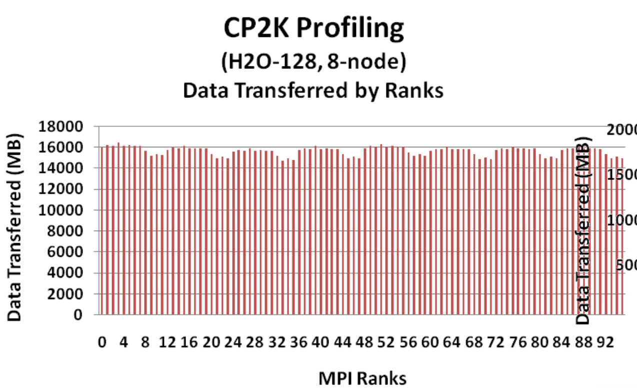
HPC ADVISORY COUNCIL

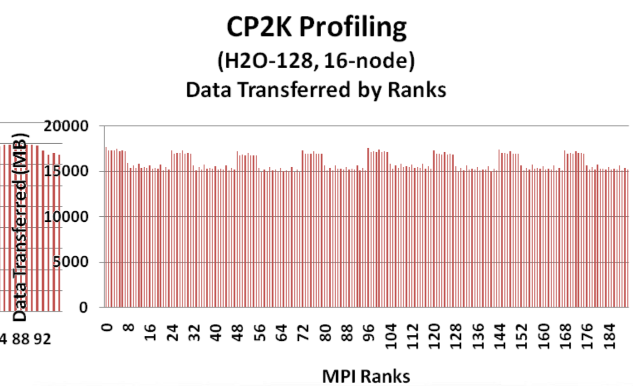
- **Data transferred to each process gradually drops as processes increase**
  - 54GB of data transferred for a process in 1-node, versus
  - 16GB of data transferred for a process in 8-node, versus
  - 9GB of data transferred for a process in 32-node
- **Communication pattern shows all ranks communicate about evenly**

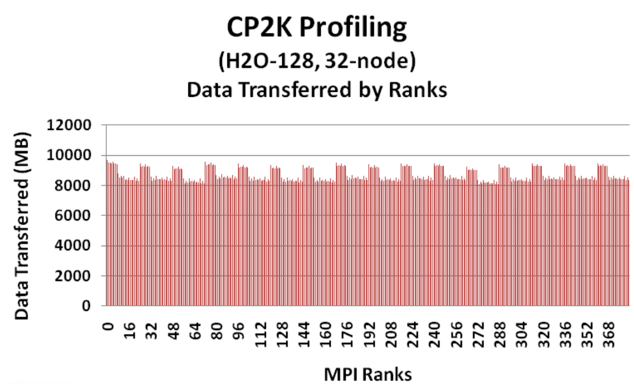






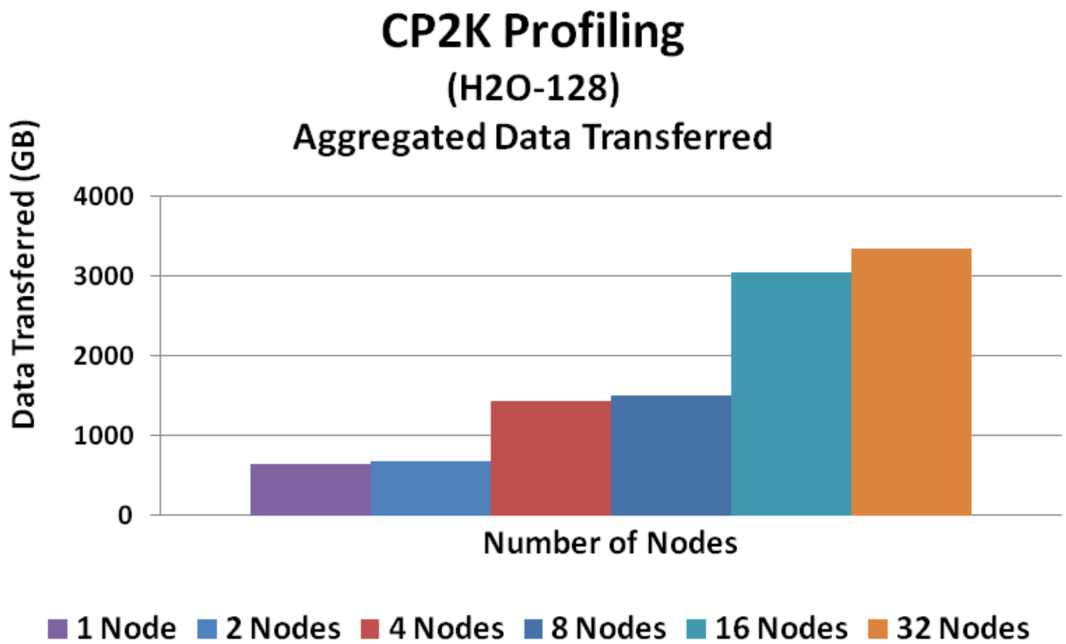






CP2K Profiling – Aggregated Transfer

- **Aggregated data transfer refers to**
  - Total amount of data being transferred in the network between all MPI ranks collectively
- **The total data transfer increases in “steps” as the cluster scales**
  - Almost constant data transferred between 1-2, 4-8, and 16-32 nodes
- **Demonstrates the advantage and importance of high throughput interconnects**
  - InfiniBand QDR was used for the profiling testing



InfiniBand QDR

CP2K – Summary

HPC ADVISORY COUNCIL

- **CP2K uses an exhaustive list of MPI APIs for data communications**
  - Non-blocking, and blocking point-to-point MPI APIs
  - Range of collective MPI APIs
- **InfiniBand QDR demonstrates higher performance at scale**
- **Using nodes with higher CPU frequency enables higher job productivity**
- **Data distribution**
  - MPI data transfer grows in steps between 2-4 nodes and 8-16 nodes.
  - Majority of messages are small messages between 0 and 64 bytes
  - Message sizes are being "shifted" to small range as node count increases

Thank You

HPC Advisory Council