HPC ADVISORY COUNCIL

NETWORK OF EXPERTISE

CP2K
Performance Benchmark and Profiling

August 2012

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

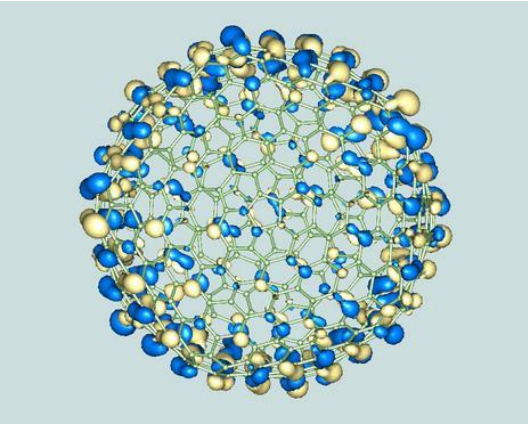
Objectives

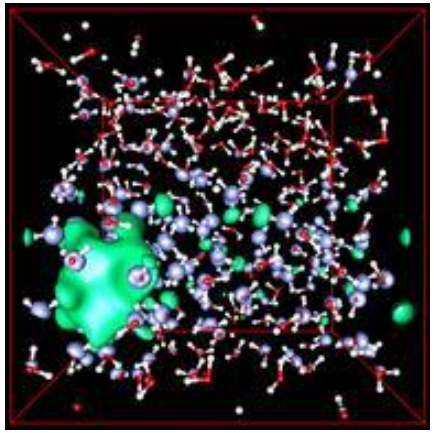
- **The following was done to provide best practices**
  - CP2K performance benchmarking
  - Interconnect performance comparisons
  - MPI performance comparison
  - Understanding CP2K communication patterns

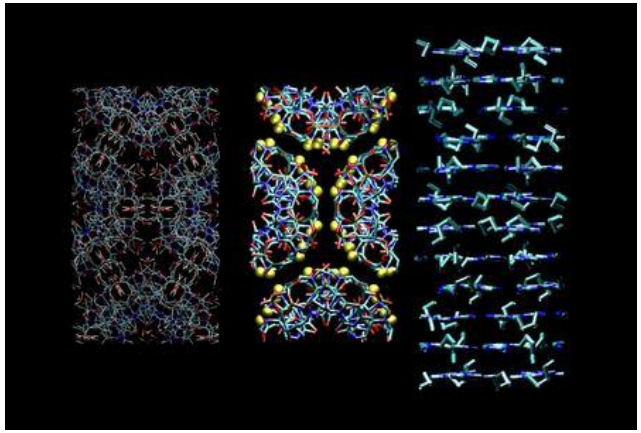
- **The presented results will demonstrate**
  - The scalability of the compute environment to provide nearly linear application scalability
  - The capability of CP2K to achieve scalable productivity

CP2K Application

- **CP2K is used to perform atomistic and molecular simulations:**
  - 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™ R720xd 16-node (256-core) "Jupiter" cluster**
  - Dual-Socket Eight-Core Intel E5-2680 @ 2.70 GHz CPUs (Static max Perf in BIOS)
  - Memory: 64GB memory, DDR3 1600 MHz
  - OS: RHEL 6.2, OFED 1.5.3 InfiniBand SW stack
  - Hard Drives: 24x 250GB 7.2 RPM SATA 2.5" on RAID 0
- **Intel Cluster Ready certified cluster**
- **Mellanox ConnectX-3 FDR InfiniBand VPI adapters**
- **Mellanox SwitchX SX6036 InfiniBand switch**
- **MPI: Intel MPI 4 Update 3, Open MPI 1.5.5**
- **Compilers and libraries: Intel Composer XE 2011 SP1, Intel MKL 10.3**
- **Application: CP2K version 2.3 (Development Version)**
- **Benchmarks 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

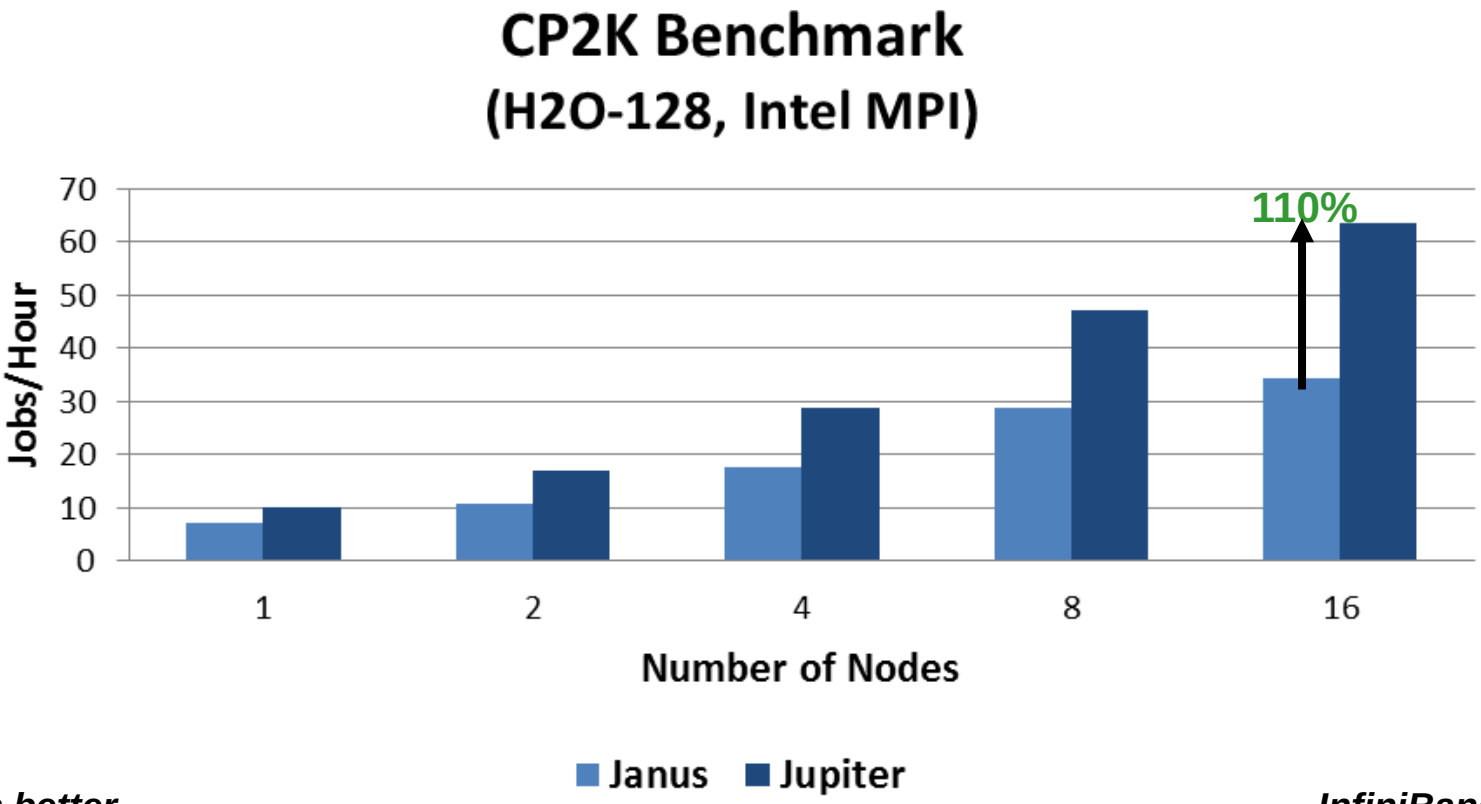
PowerEdge R720xd

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

CP2K Performance – Processors

- **Intel E5-2680 (Sandy Bridge) cluster outperforms prior generations**
  - Performs 110% better than X5670 cluster at 16 nodes
- **System components used:**
  - Jupiter: 2-socket Intel E5-2680 @ 2.7GHz, 1600MHz DIMMs, FDR IB, 24 disks
  - Janus: 2-socket Intel X5670 @ 2.93GHz, 1333MHz DIMMs, QDR IB, 1 disk

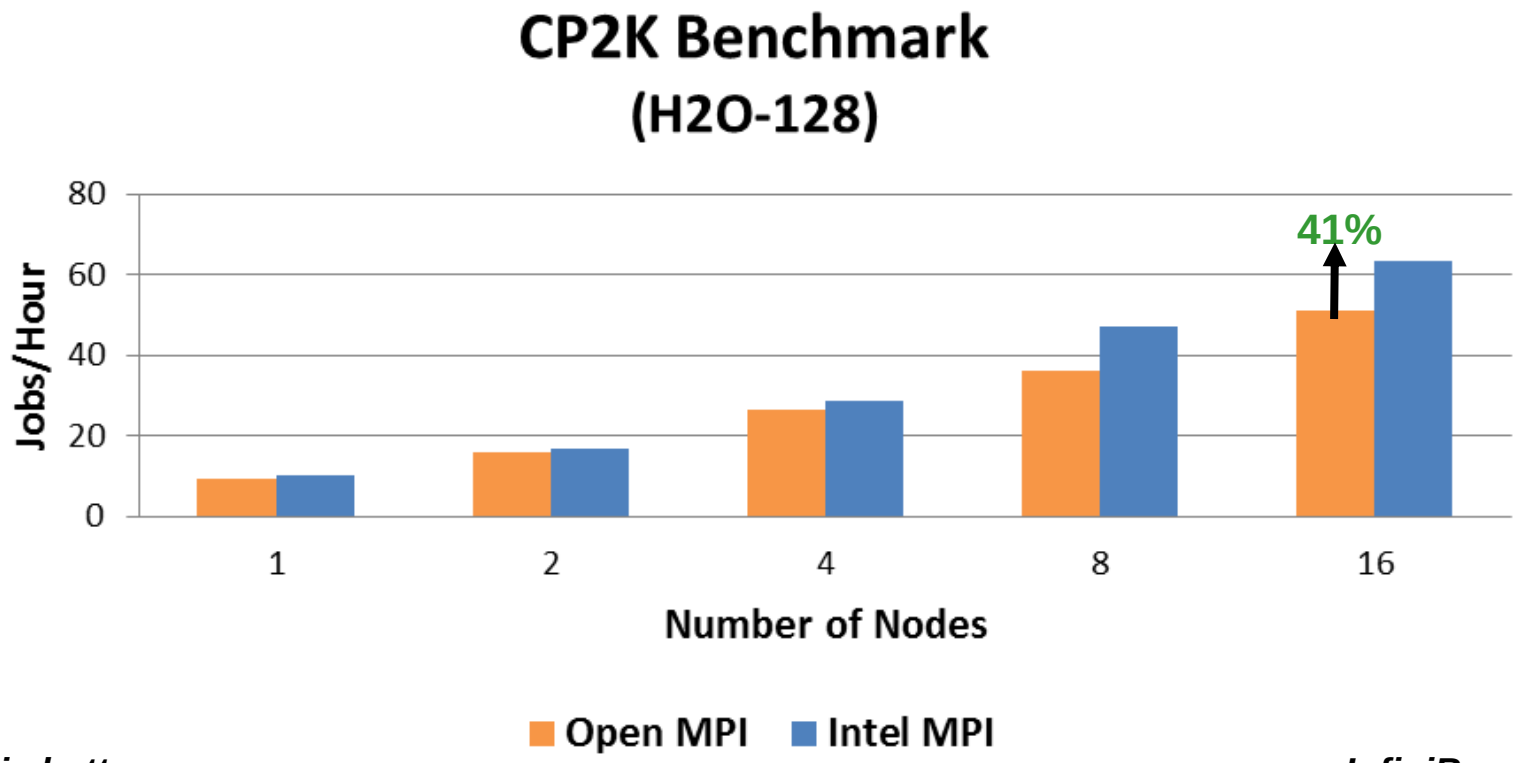


Higher is better

InfiniBand FDR

CP2K Performance – MPI

- **Intel MPI outperforms Open MPI at larger scale**
  - Up to 41% higher performance than Open MPI at 16-node
- **CPU binding optimization flag used in all cases shown**
  - No other optimization flags are used

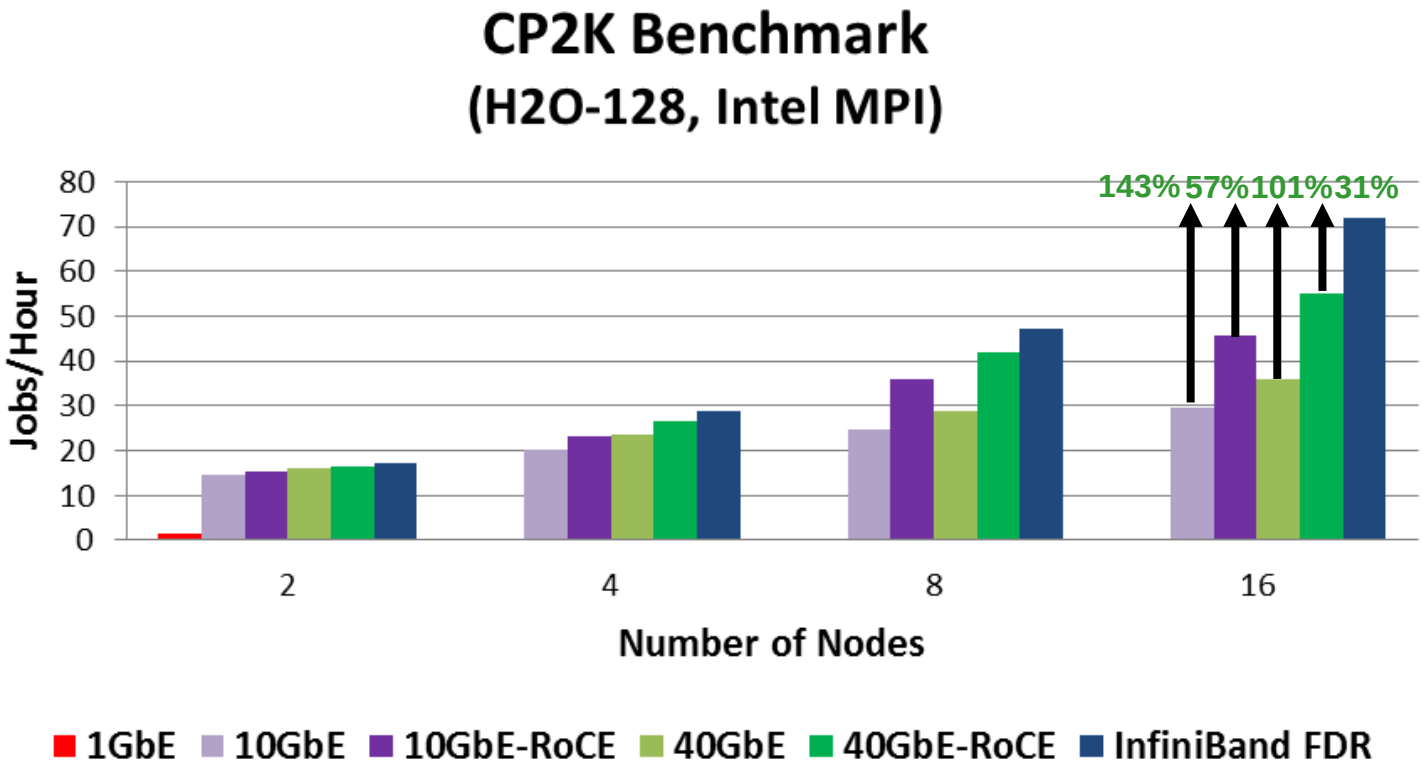


Higher is better

InfiniBand FDR

CP2K Performance – Ethernet Interconnect

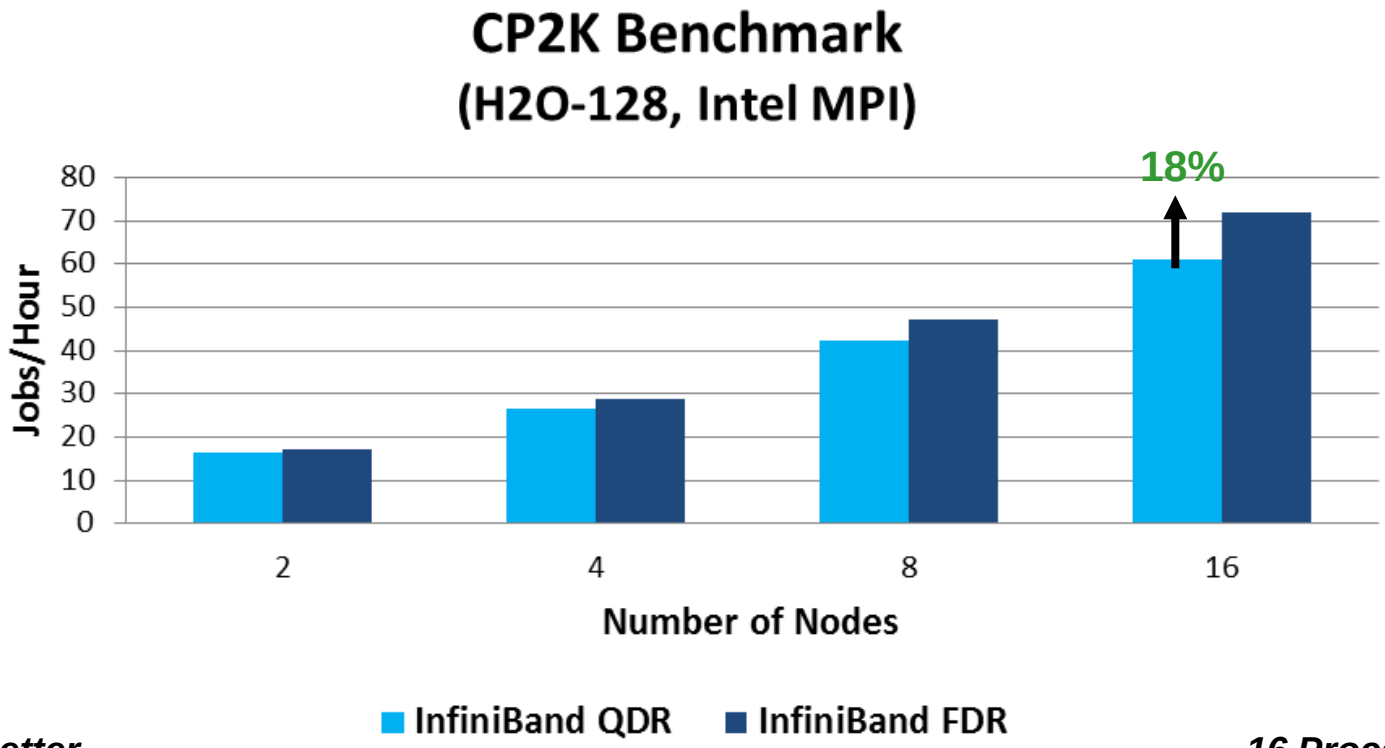
- **InfiniBand FDR provides better scalability performance than Ethernet**
  - 143% better performance than 10GbE at 16 nodes
  - 101% better performance than 40GbE at 16 nodes
  - 57% better performance than 10GbE-RoCE at 16 nodes
  - 31% better performance than 40GbE-RoCE at 16 nodes
  - 1GbE does not scale at all



Higher is better *16 Processes/Node*

CP2K Performance – InfiniBand Interconnect

- **InfiniBand FDR delivers better application performance**
  - Up to 18% better performance than InfiniBand QDR
  - Using Mellanox ConnectX-3 PCIe Gen3 in FDR mode and QDR mode

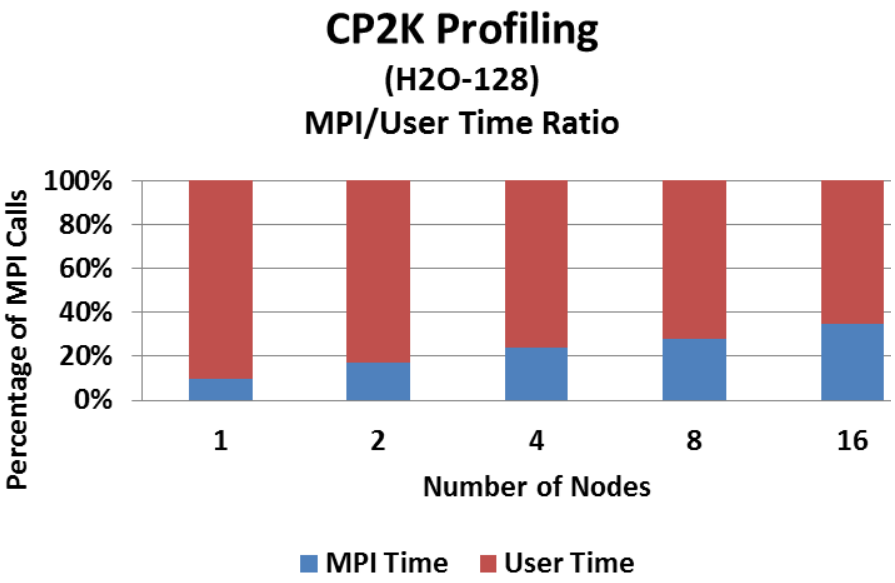


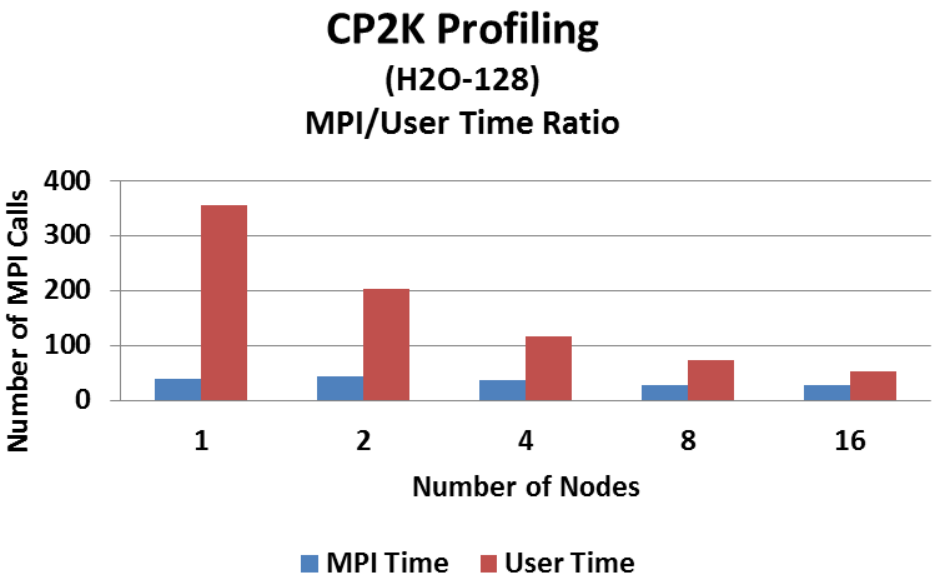
Higher is better

16 Processes/Node

CP2K Profiling – MPI/User Time Ratio

- **MPI communication time stays flat while compute time halves**
  - Reflects that more time spent on computation than communications
  - Spreading workload to InfiniBand-connected nodes without introducing extra overhead

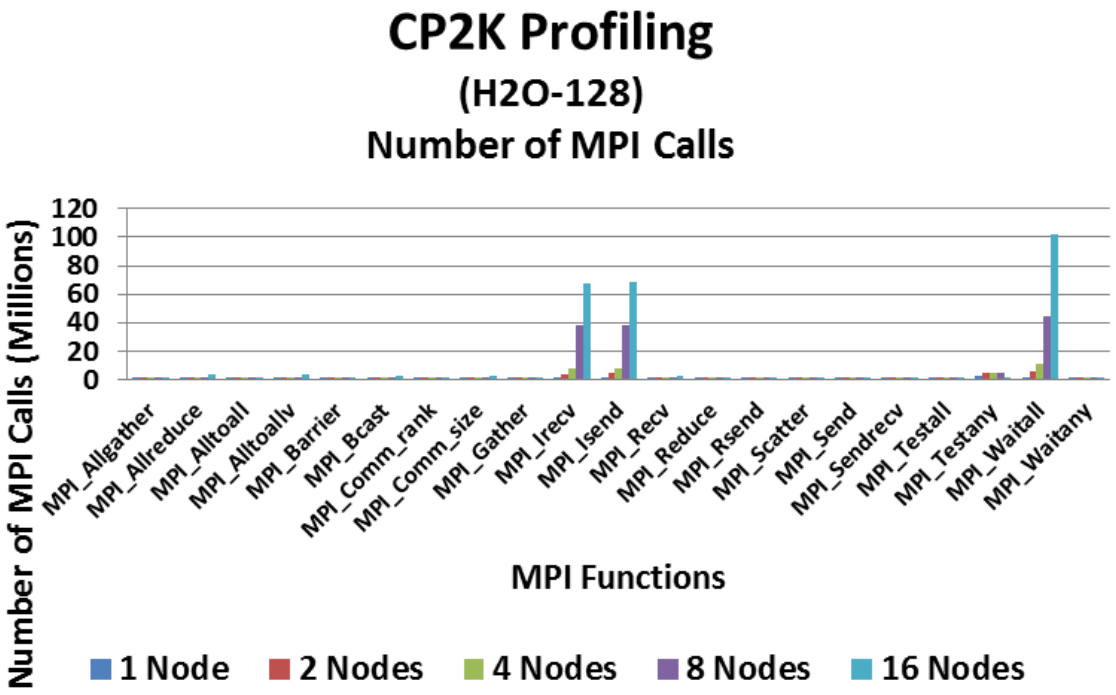


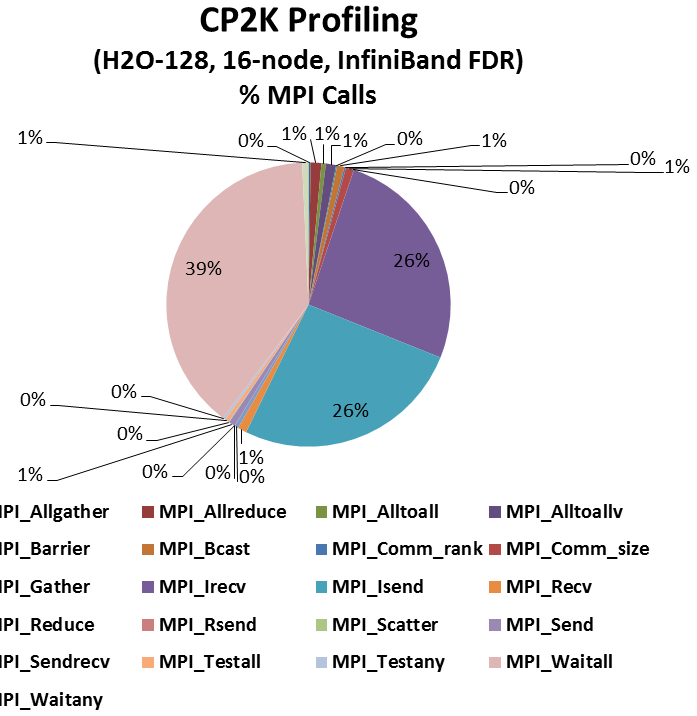


InfiniBand FDR

CP2K Profiling – Number of MPI Calls

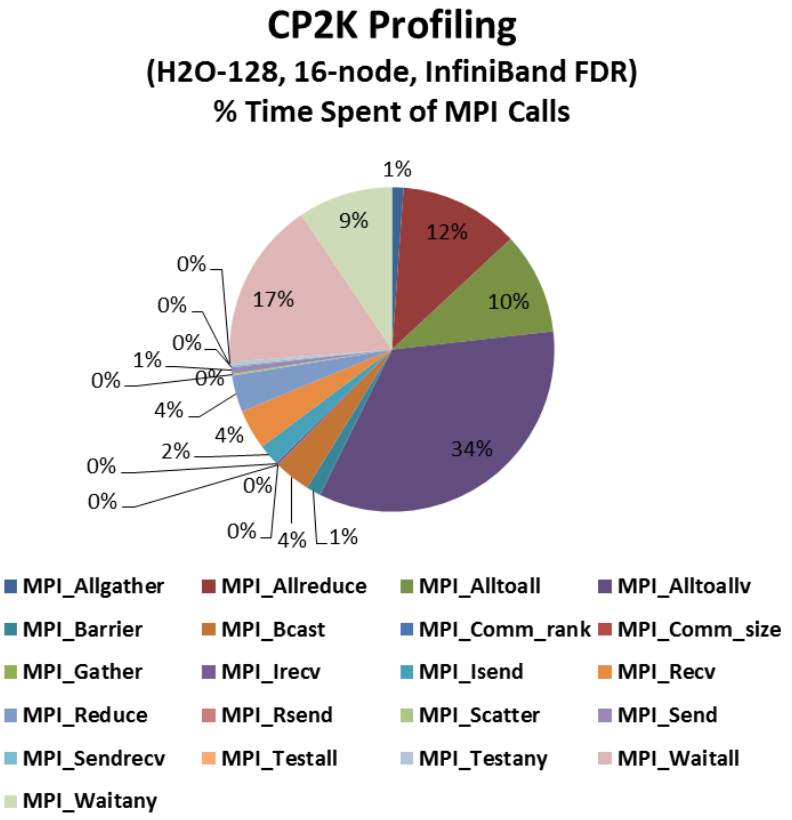
- **CP2K utilizes a wide range of MPI APIs**
  - 21 MPI APIs used in total
- **MPI_Waitall, MPI_Irecv and MPI_Isend are almost used exclusively**
  - MPI_Alltoallv (39%), MPI_Irecv and MPI_Isend (26% each) at 16 nodes

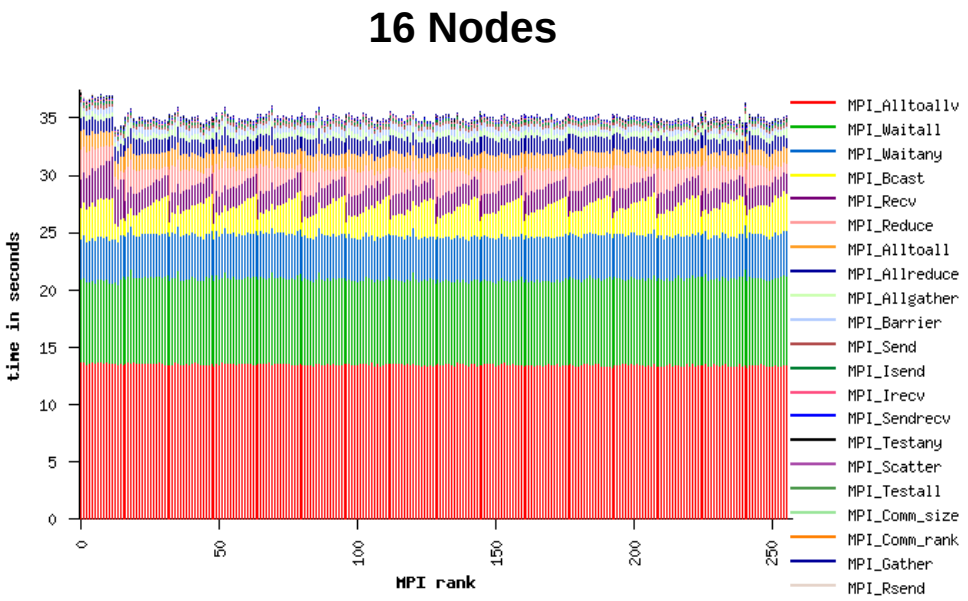




CP2K Profiling – % Time Spent of MPI Calls

- **The most time MPI calls is MPI_Alltoallv**
  - MPI_Alltoallv(34%), MPI_Waitall(17%), MPI_Reduce(12%), MPI_Alltoall(10%)
- **Time consumed by calls are generally balanced**
  - E.g. MPI_Alltoall, MPI_Waitall and MPI_waitany



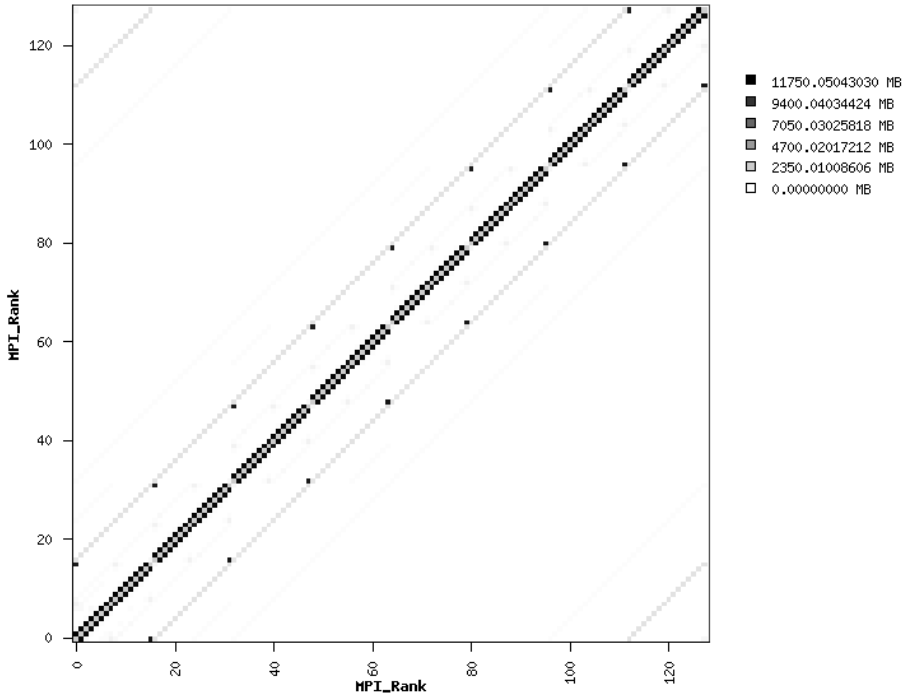


16 Processes/Node

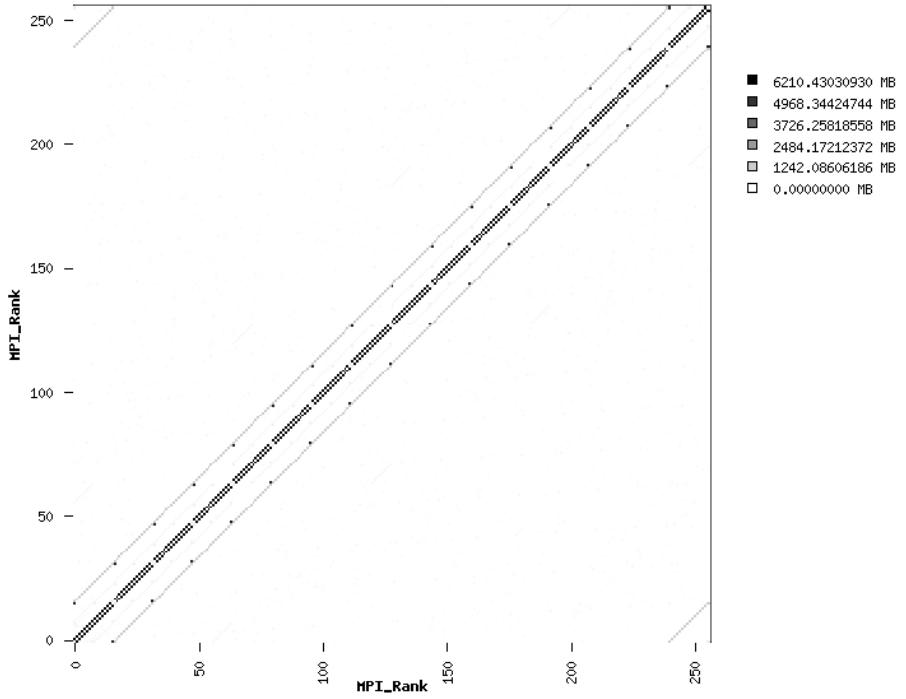
CP2K Profiling – MPI Data Transfer

- **As the cluster grows, less data transfers between MPI processes**
  - Decrease from 11GB max (8 nodes) at to 6GB max per rank (16 nodes)
  - Majority of communications are between neighboring ranks

8 Nodes

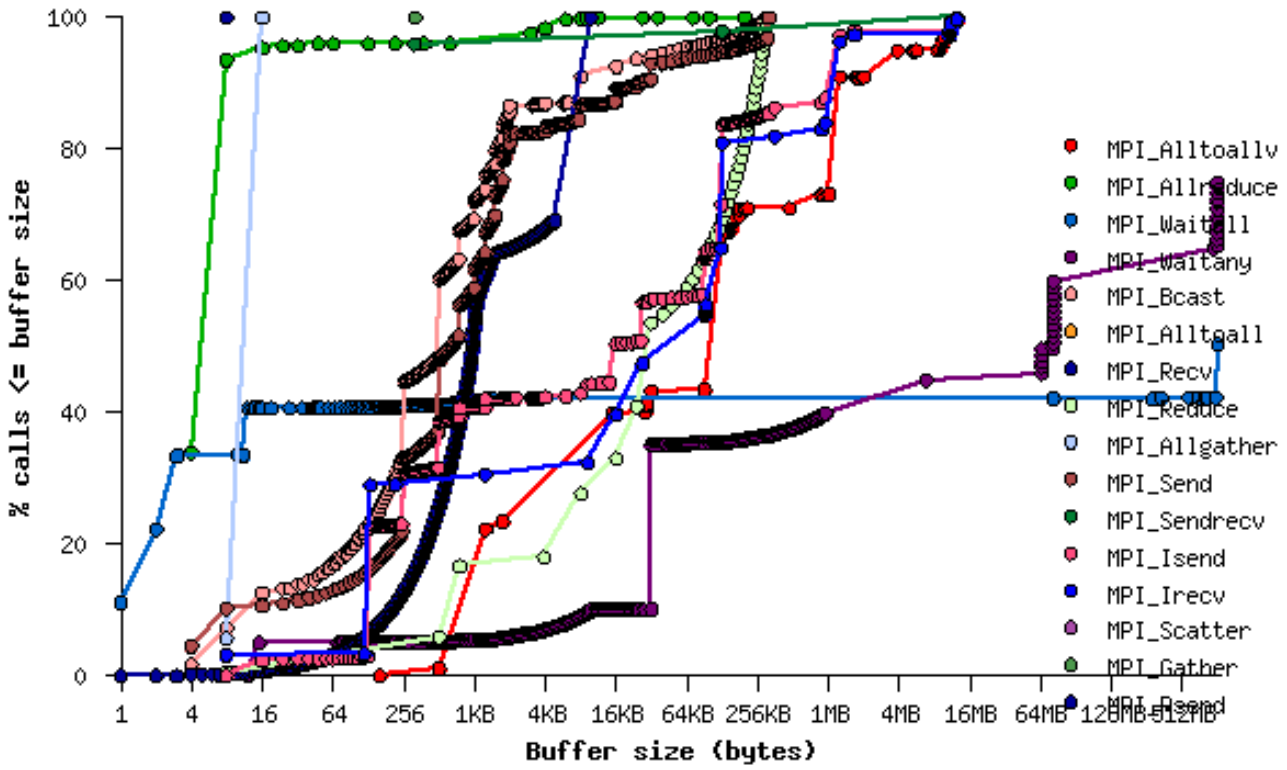


16 Nodes



CP2K Profiling – MPI Message Size

- **Majority of MPI messages are small to midrange messages**
  - In the range of 0B to 256B, and 16KB to 256KB



CP2K – Summary

- **Performance**
  - Intel Xeon E5-2600 series and InfiniBand FDR enable CP2K to scale with 16 nodes
  - The E5-2680 cluster outperforms X5670 cluster by 110% at 16 nodes
  - InfiniBand FDR provides better scalability performance than Ethernet
    - 143% better performance than 10GbE at 16 nodes
    - 101% better performance than 40GbE at 16 nodes
    - 57% better performance than 10GbE-RoCE at 16 nodes
    - 31% better performance than 40GbE-RoCE at 16 nodes
    - 1GbE does not scale beyond 2 nodes
  - InfiniBand FDR provides up to 18% of performance gain over InfiniBand QDR at 16-node
  - Intel MPI scales better than Open MPI at large node counts (16 nodes) by 41%

- **Profiling**
  - The most time MPI calls is MPI_Alltoallv
  - Majority of MPI messages are small to midrange messages

Thank You

HPC Advisory Council