

CP2K

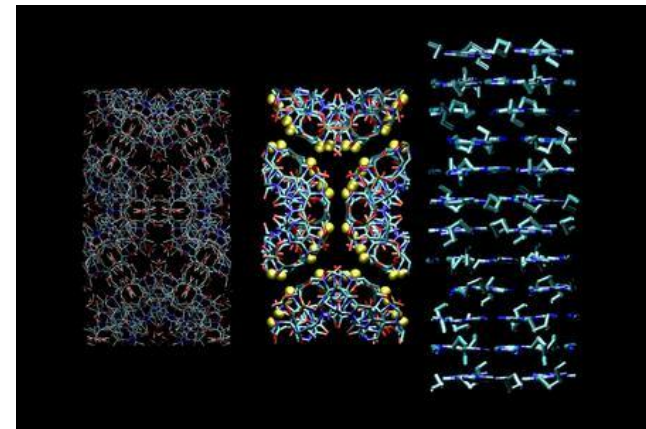
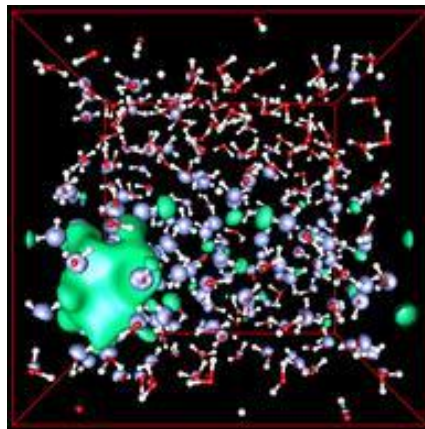
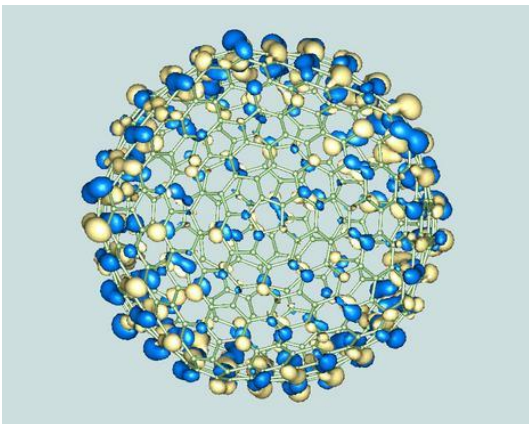
Performance Benchmark and Profiling

April 2011



- **The following research was performed under the HPC Advisory Council activities**
 - Participating vendors: AMD, Dell, Mellanox
 - Compute resource - HPC Advisory Council Cluster Center
- **For more info please refer to**
 - [http:// www.amd.com](http://www.amd.com)
 - [http:// www.dell.com/hpc](http://www.dell.com/hpc)
 - <http://www.mellanox.com>
 - <http://cp2k.berlios.de>

- **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**



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

- **The presented results will demonstrate**
 - The scalability of the compute environment
 - The capability of CP2K to achieve scalable productivity
 - Considerations for performance optimizations

- **Dell™ PowerEdge™ R815 11-node (528-core) cluster**
- **AMD™ Opteron™ 6174 (code name “Magny-Cours”) 12-cores @ 2.2 GHz CPUs**
- **4 CPU sockets per server node**
- **Mellanox ConnectX-2 VPI adapters for 40Gb/s QDR InfiniBand and 10Gb/s Ethernet**
- **Mellanox MTS3600Q 36-Port 40Gb/s QDR InfiniBand switch**
- **Fulcrum based 10Gb/s Ethernet switch**
- **Memory: 128GB memory per node DDR3 1333MHz**
- **OS: RHEL 5.5, MLNX-OFED 1.5.2 InfiniBand SW stack**
- **MPI: Intel MPI 4, Open MPI 1.5.3 with KNEM 0.9.6, Platform MPI 8.0.1**
- **Compilers: Intel Compilers 11.1, PGI 11.4**
- **Application: CP2K version 2.2.196 (External libraries used: Intel MKL 10.1, FFTW3, BLACS, ScaLAPACK 1.8.0, LAPACK 3.3)**
- **Benchmark workload: H2O-256.inp**

- **HPC Advisory Council Test-bed System**
- **New 11-node 528 core cluster - featuring Dell PowerEdge™ R815 servers**
 - Replacement system for Dell PowerEdge SC1435 (192 cores) cluster system following 2 years of rigorous benchmarking and product EOL
 - System to be redirected to explore HPC in the Cloud applications
- **Workload profiling and benchmarking**
 - Characterization for HPC and compute intense environments
 - Optimization for scale, sizing and configuration and workload performance
 - Test-bed Benchmarks
 - RFPs
 - Customers/Prospects, etc
 - ISV & Industry standard application characterization
 - Best practices & usage analysis



About Dell PowerEdge™ Platform Advantages

Best of breed technologies and partners

Combination of AMD™ Opteron™ 6100 series platform and Mellanox ConnectX InfiniBand on Dell HPC

Solutions provide the ultimate platform for speed and scale

- Dell PowerEdge R815 system delivers 4 socket performance in dense 2U form factor
- Up to 48 core/32DIMMs per server – 1008 core in 42U enclosure

Integrated stacks designed to deliver the best price/performance/watt

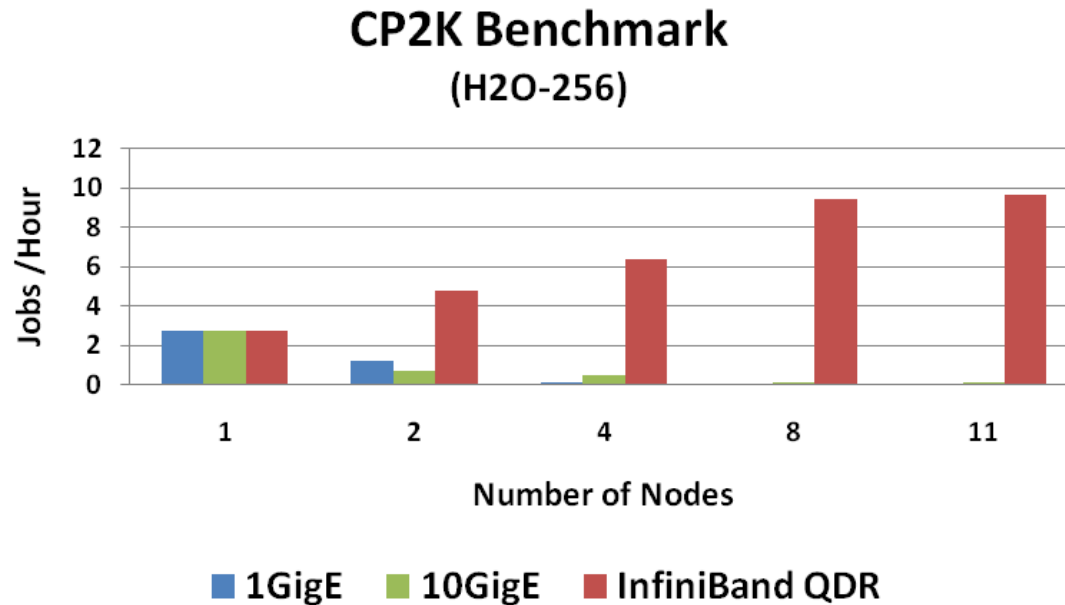
- 2x more memory and processing power in half of the space
- Energy optimized low flow fans, improved power supplies and dual SD modules

Optimized for long-term capital and operating investment protection

- System expansion
- Component upgrades and feature releases



- **InfiniBand shows continuous gain as the cluster scales**
 - The only interconnect that enables higher scalability for CP2K
- **Ethernet performance does not scale beyond 1 node**
 - Both 10GigE and 1GigE performance plummets with 2 or more nodes

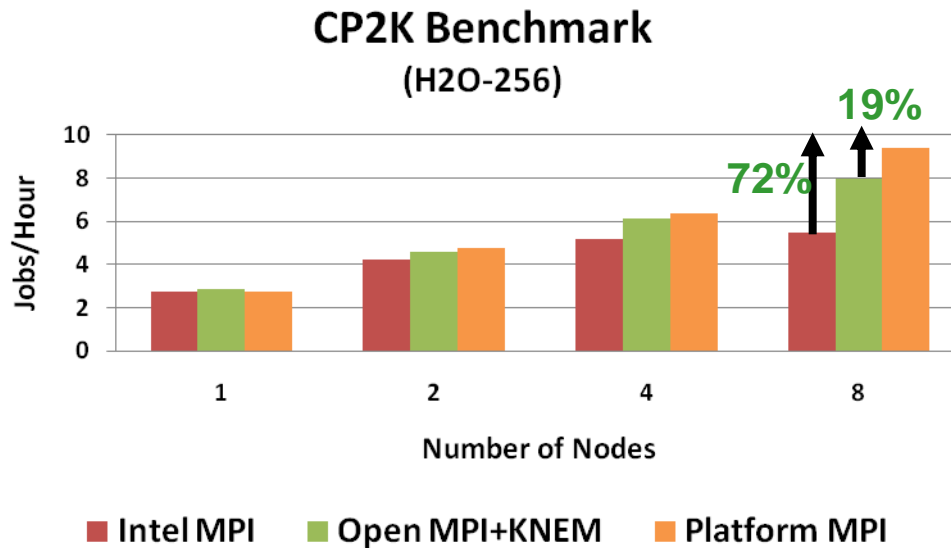


Higher is better

48 Cores/Node

CP2K Performance – MPI Implementations

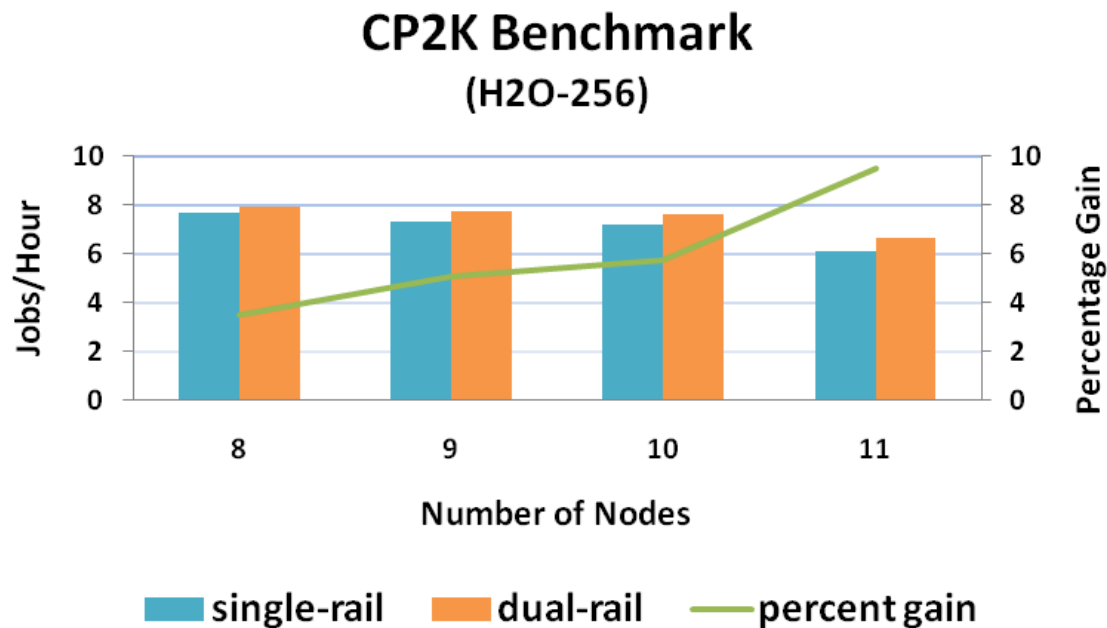
- **Platform MPI enables the highest scalability**
 - Up to 72% faster than Intel MPI on 8 nodes
 - Up to 19% faster than Open MPI with KNEM support on 8 nodes
- **Open MPI runs with KNEM support**
 - KNEM enables intra-node MPI communication for large messages



Higher is better

48 Cores/Node

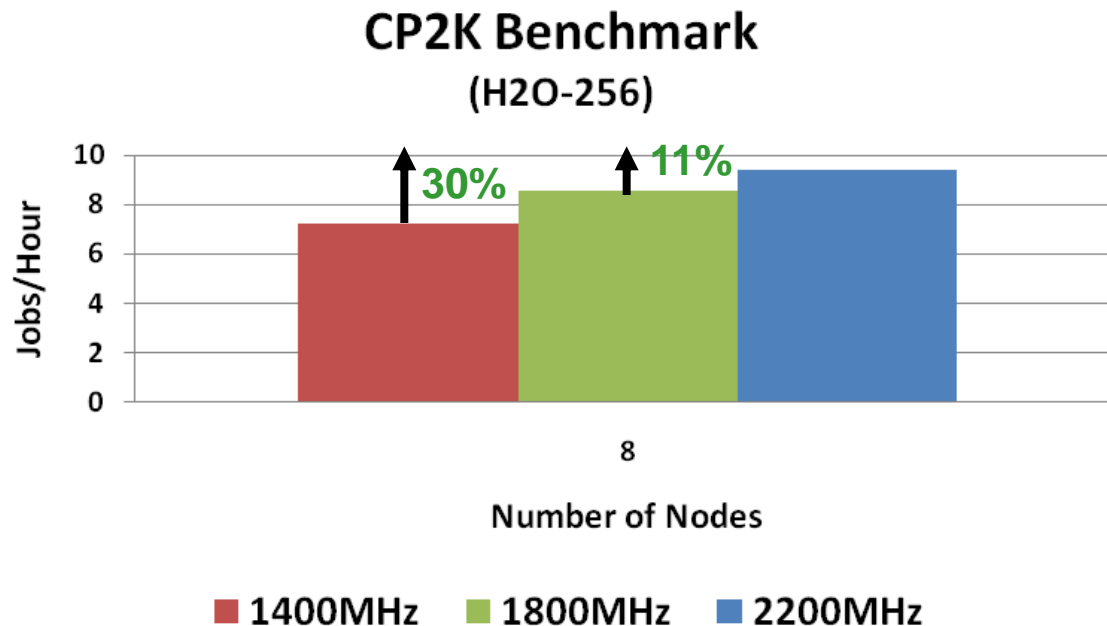
- **Dual-rail (Dual InfiniBand cards) enables better performance than single-rail**
 - Up to 10% better at 11-node
- **The benefit of dual-rail starts to show with 8 nodes**
 - Expect to see more gain as the cluster size increases



Higher is better

*Open MPI
48 Cores/Node*

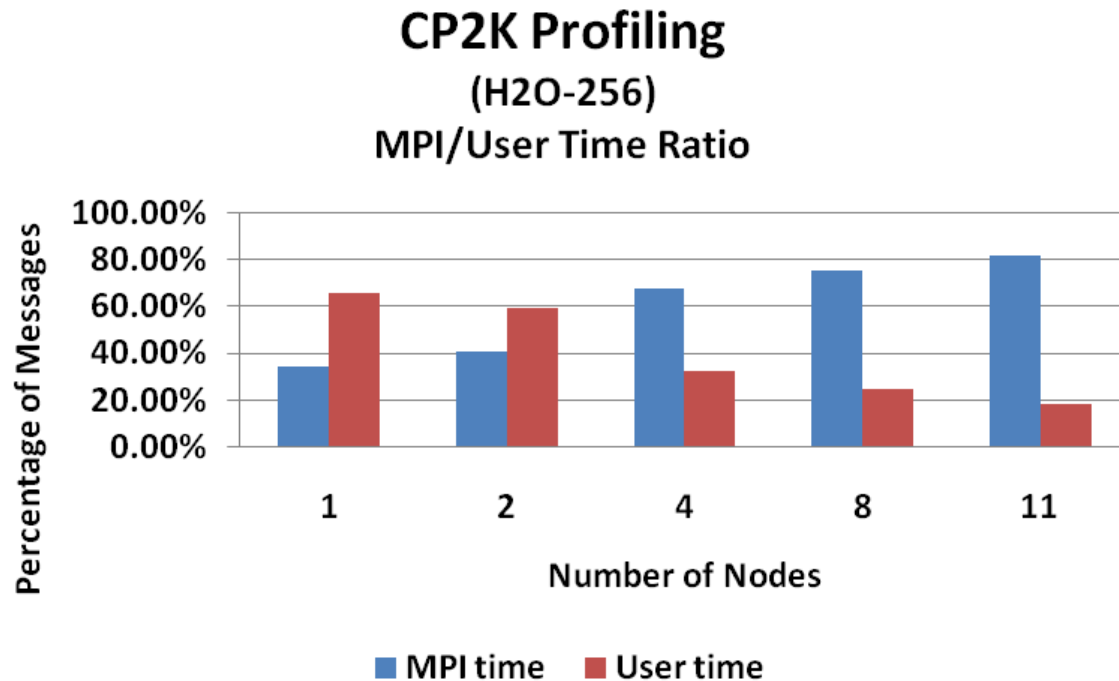
- **Increasing CPU core frequency enables higher job efficiency**
 - Up to 30% better job performance between 2200MHz vs 1400MHz
 - Up to 11% better job performance between 2200MHz vs 1800MHz



Higher is better

48 Cores/Node

- **CP2K becomes communicative at a very fast pace**
 - Due to the high core counts per node
- **MPI communication time dominates the overall time**
 - Shows low latency interconnect such as InfiniBand is required for good scalability

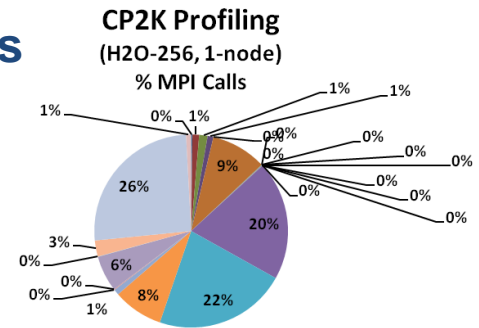


Higher is better

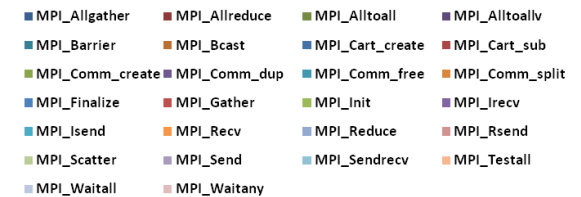
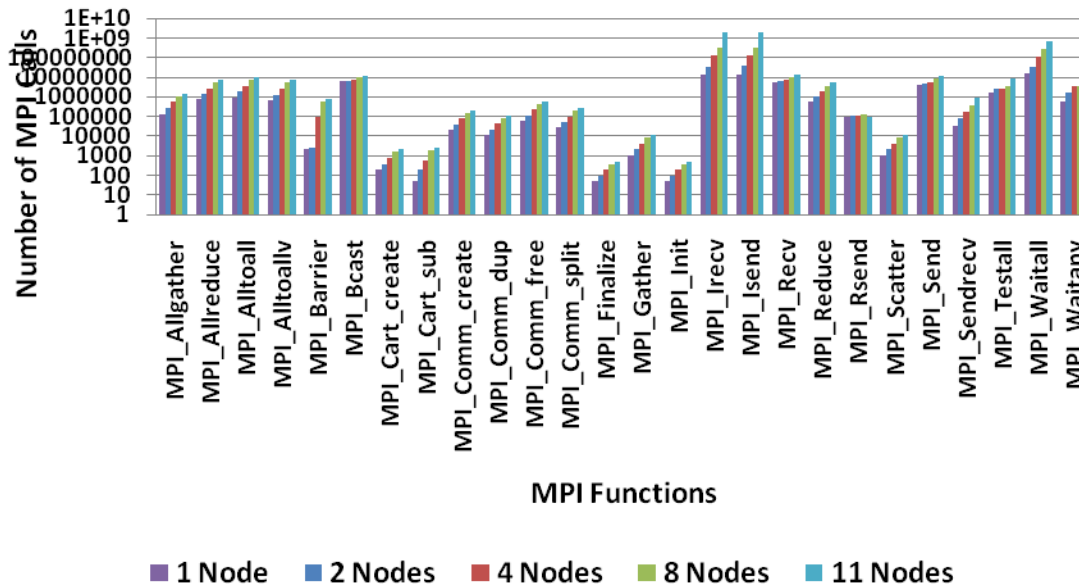
48 Cores/Node

CP2K Profiling – Number of MPI Calls

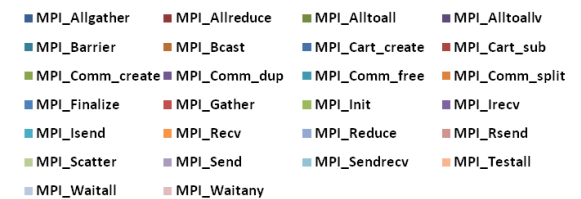
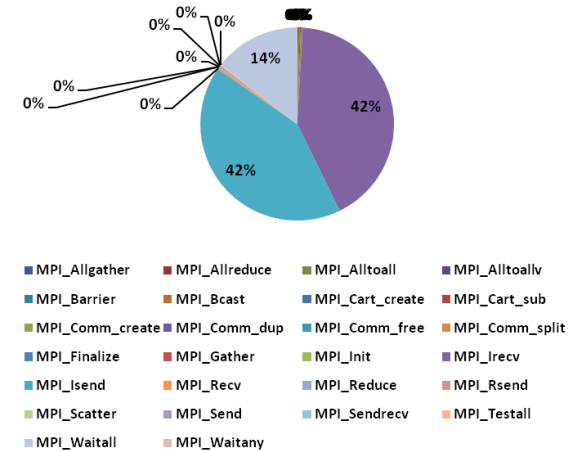
- **CP2K with this dataset uses an extensive list of MPI calls**
 - 26 different MPI APIs are used
- **The most used MPI function is MPI_Isend and Irecv**
 - Each accounted for 42% of all MPI calls on a 11-node job
- **MPI_Waitall represents a smaller ratio as size increases**
 - From 26% down to 14% from 1 node to 11 node



CP2K Profiling (H2O-256)
Number of MPI Calls



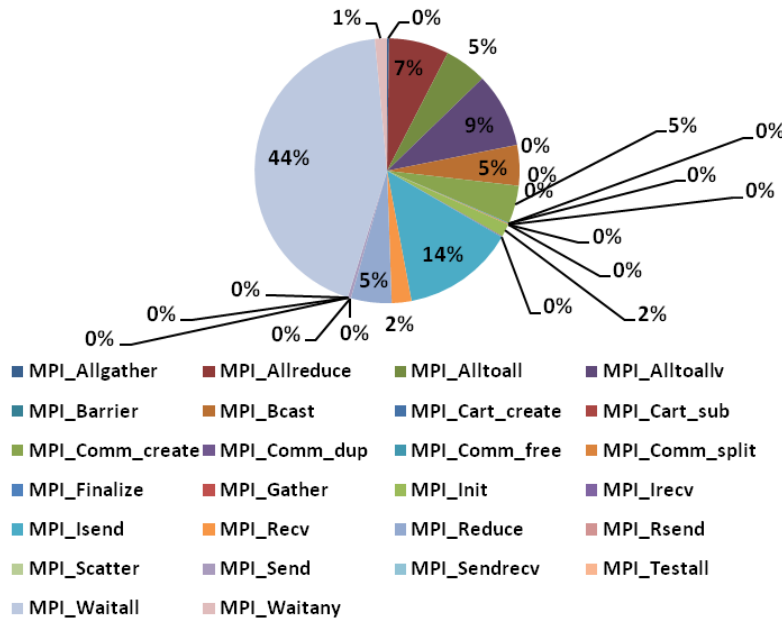
CP2K Profiling (H2O-256, 11-node, InfiniBand) % MPI Calls



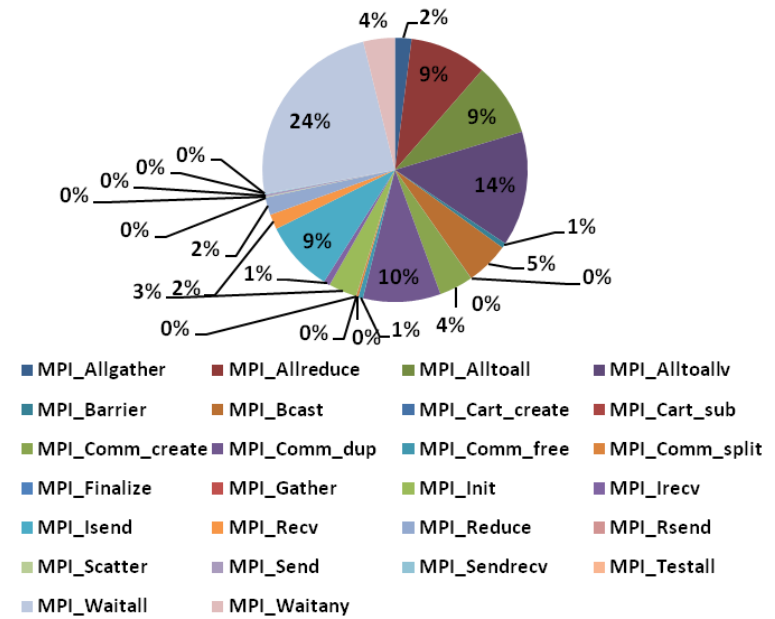
CP2K Profiling – Time Spent of MPI Calls

- **The largest time consumer is calls MPI_Waitall**
 - MPI_Waitall uses for waiting communications to complete
 - Occupies 44% of all MPI time for 1 node
 - Occupies 24% of all MPI time for 11 node
- **Next on the list are MPI_Isend and MPI_Alltoallv**
 - MPI_Isend takes up 14% and MPI_Alltoallv takes 11% on a 11-node run

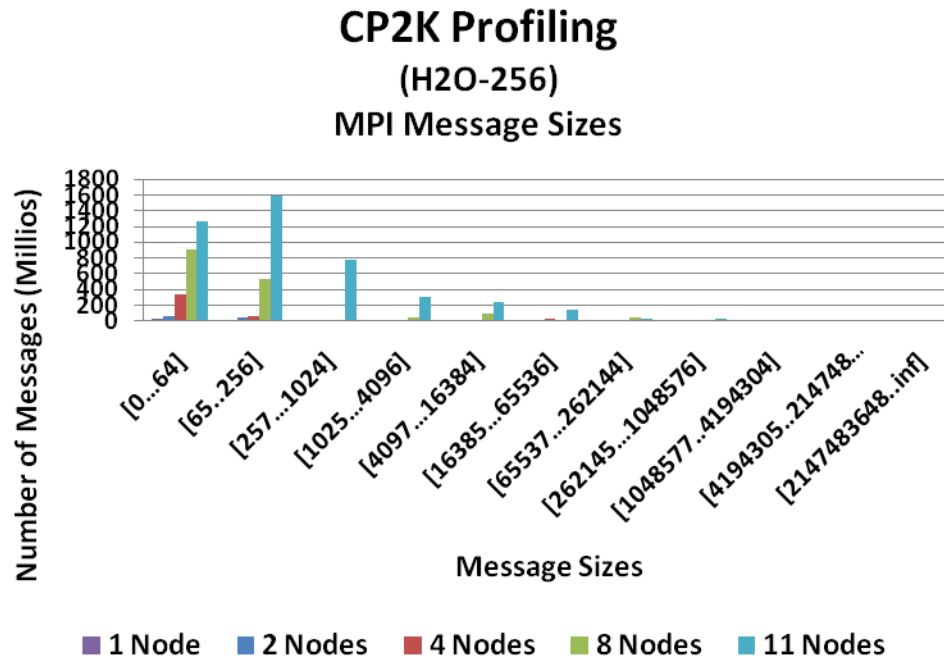
CP2K Profiling
(H2O-256, 1-node)
% Time Spent of MPI Calls



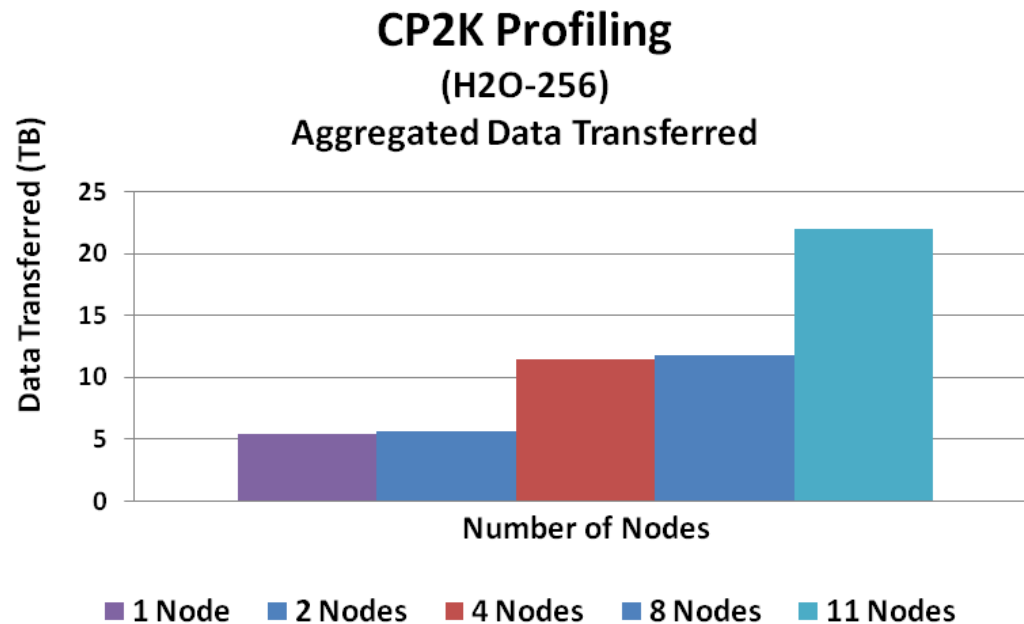
CP2K Profiling
(H2O-256, 11-node)
% Time Spent of MPI Calls



- **Majority of the MPI message sizes are small message sizes**
 - In the range of 0B and 64B for 8 nodes or less
- **Messages increase accelerates with the node count increases**
 - Especially in the range of 65B and 256B for 11 nodes
- **Benefit of Multi-rail begins to emerge starting with 8-node**
 - As the number of smaller messages start to increase dramatically



- **Aggregated data transfer refers to:**
 - Total amount of data being transferred in the network between all MPI ranks collectively
- **The total data transfer increases as the cluster scales**
 - As the cluster size reaches a new level, amount of data being driven would be doubled
- **Demonstrates the advantage and importance of scalable network interconnect**
 - InfiniBand QDR can deliver bandwidth needed to push 21TB of data across the network

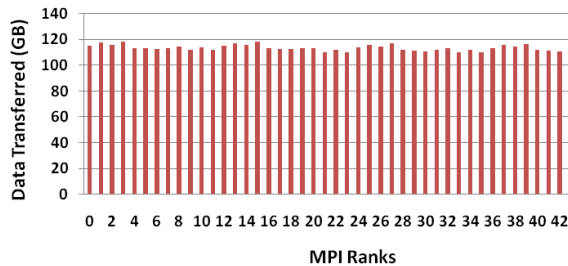


InfiniBand QDR

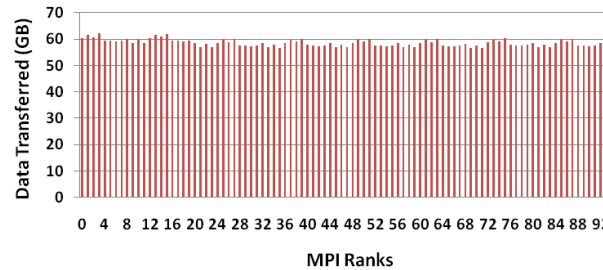
CP2K Profiling – Data Transfer Per Process

- **Data transferred to each MPI rank is driven down in “levels”**
 - From 120GB (1 node) to 60GB (2 ,4 nodes) to 30GB (8 nodes)
 - Aggregated Data Transferred doubled for 11 node, hence the increase in data per rank
- **As the cluster scales, data is spread across to more processes**

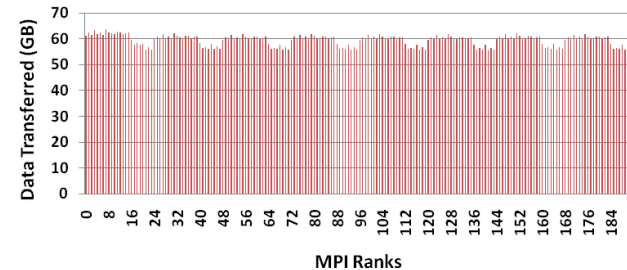
CP2K Profiling
(H2O-256, 1-node)
Data Transferred by Ranks



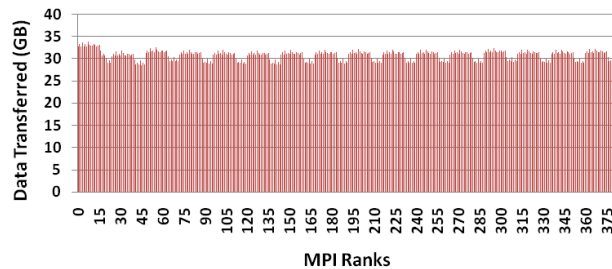
CP2K Profiling
(H2O-256, 2-node)
Data Transferred by Ranks



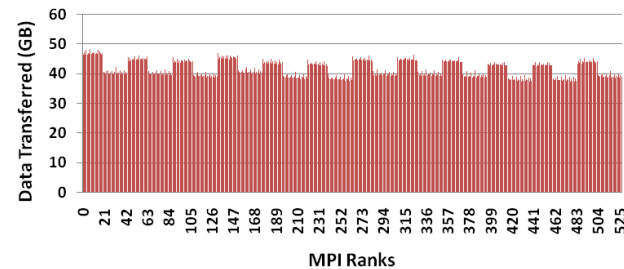
CP2K Profiling
(H2O-256, 4-node)
Data Transferred by Ranks



CP2K Profiling
(H2O-256, 8-node)
Data Transferred by Ranks



CP2K Profiling
(H2O-256, 11-node)
Data Transferred by Ranks



- **CP2K and this dataset shows a high demand for:**
 - Both CPU and network bandwidth throughput
- **Networking:**
 - InfiniBand shows as the preferred interconnect solution for any cluster size
 - Shows benefit for using dual-rail InfiniBand from 8-nodes and up
 - 10GigE and 1GigE do not scale for this application and dataset
- **CPU:**
 - The CPU frequency has a direct impact on job productivity
- **MPI:**
 - Platform MPI allows good scalability for CP2K among the other MPI tested
 - **Data being transferred on the network**
 - Tends to increase in “levels” that can create heavy burden to network as cluster scales

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