

CP2K

Performance Benchmark and Profiling

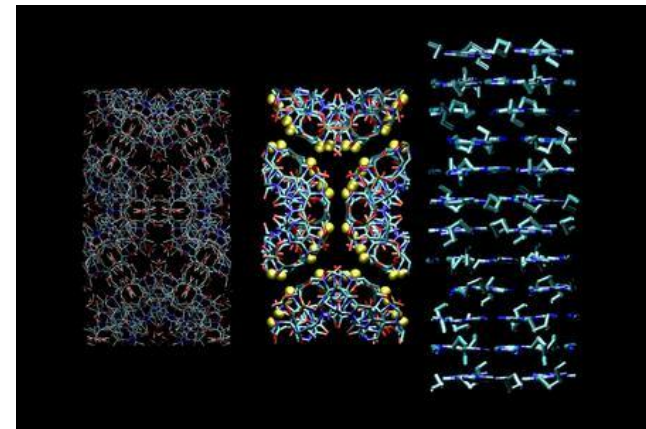
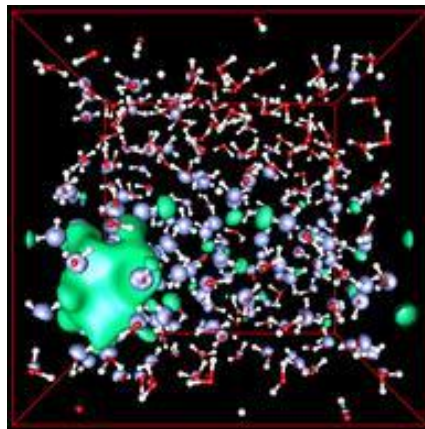
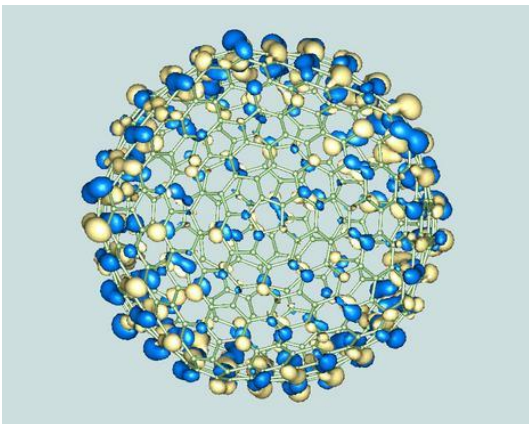
April 2011



- **The following research was performed under the HPC Advisory Council HPC|works working group activities**
 - Participating vendors: HP, Intel, Mellanox
 - Compute resource - HPC Advisory Council Cluster Center

- **For more info please refer to**
 - <http://www.hp.com/go/hpc>
 - www.intel.com
 - 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**

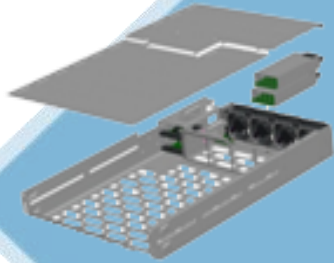


- **The presented research was done to provide best practices**
 - MPI libraries comparisons
 - Interconnect performance benchmarking
 - CP2K Application profiling
 - Understanding CP2K communication patterns
 - CP2K performance optimization
- **The presented results will demonstrate**
 - Balanced compute environment determines application performance
 - Tips to tune MPI to achieve maximum CP2K scalability

- **HP ProLiant SL2x170z G6 16-node cluster**
 - Six-Core Intel X5670 @ 2.93 GHz CPUs
 - Memory: 24GB per node
 - OS: CentOS5U5, OFED 1.5.3 InfiniBand SW stack
- **Mellanox ConnectX-2 InfiniBand QDR adapters and switches**
- **Fulcrum based 10Gb/s Ethernet switch**
- **MPI: Intel MPI 4, Open MPI 1.5.3 with KNEM 0.9.6, Platform MPI 8.0.1**
- **Compilers: Intel Compilers 11.1**
- **Application: CP2K version 2.2.196**
- **Libraries: Intel MKL 10.1, BLACS, ScaLAPACK 1.8.0**
- **Benchmark workload**
 - H2O-128.inp

About HP ProLiant SL6000 Scalable System

- **Solution-optimized for extreme scale out**



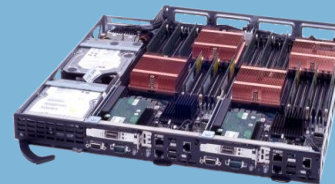
ProLiant z6000 chassis
Shared infrastructure
– fans, chassis, power



ProLiant SL160z G6 ProLiant SL165z G7
Large memory
-memory-cache apps



ProLiant SL170z G6
Large storage
-Web search and database apps




ProLiant SL2x170z G6
Highly dense
- HPC compute and
web front-end apps

Save on cost and energy -- per node, rack and data center

Mix and match configurations

Deploy with confidence



#1
Power
Efficiency*

* SPECpower_ssj2008
www.spec.org
17 June 2010, 13:28

- **Input Dataset**

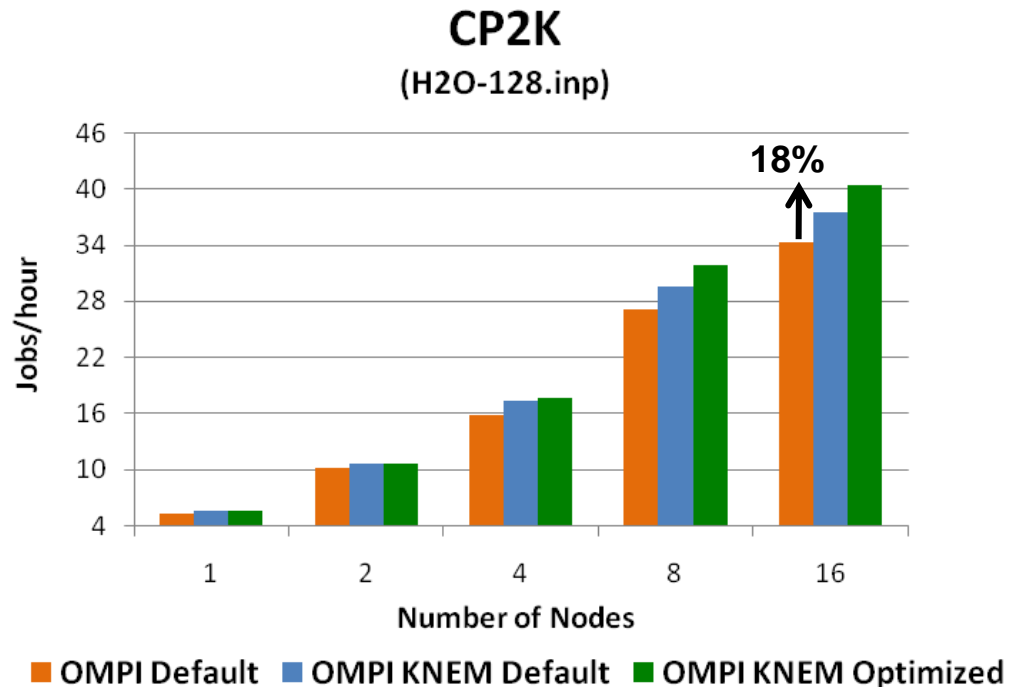
- H2O-128.inp

- **MPI tuning enables nearly 18% performance gain at 16 nodes / 192 cores**

- KNEM accelerates intra node communication

- MPI RDMA optimization reduce inter-node communication latency

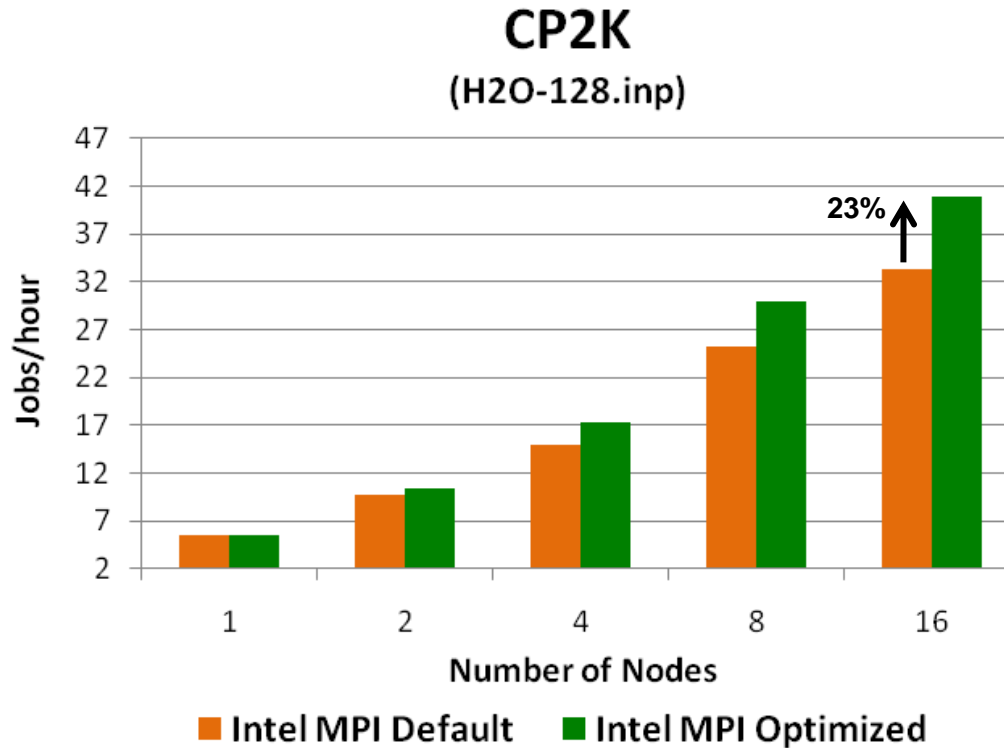
- `--mca btl_openib_eager_limit 65536 --mca btl_openib_max_eager_rdma 8 --mca btl_openib_eager_rdma_num 8`



Higher is better

12-cores per node

- **Applying optimized parameters improves CP2K performance by 23%**
 - Increase alignment for the sending buffer
 - Use right Alltoall and Alltoallv algorithm

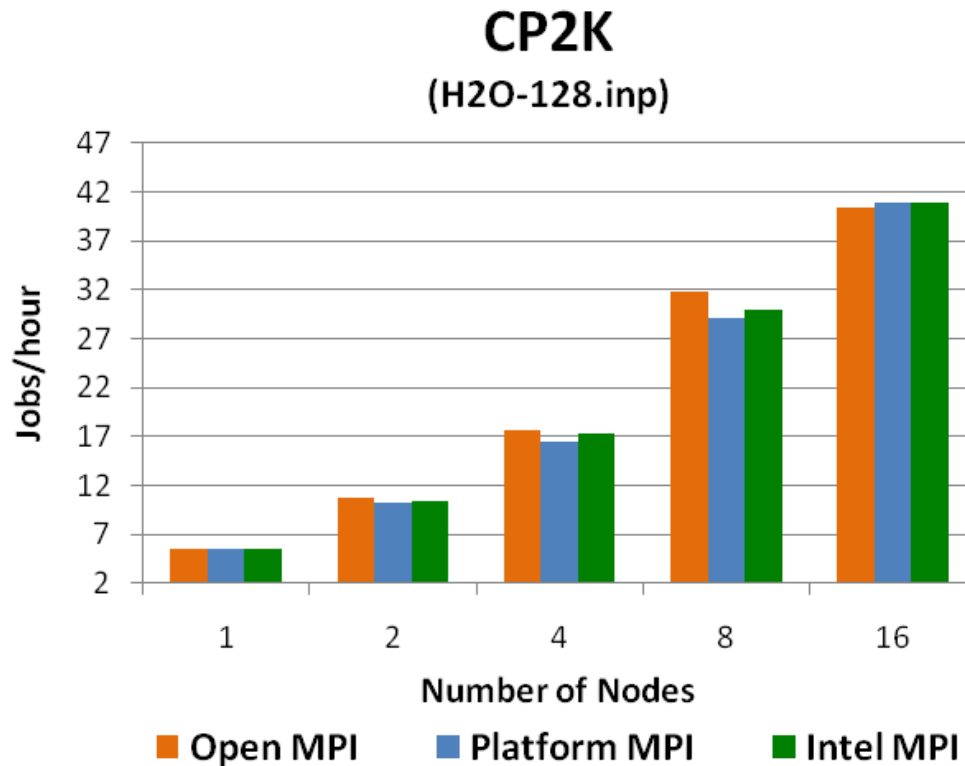


Higher is better

12-cores per node

CP2K Benchmark Results – MPI Libraries

- Open MPI is better at smaller node count
- All tested MPI libraries have similar performance at 16 nodes

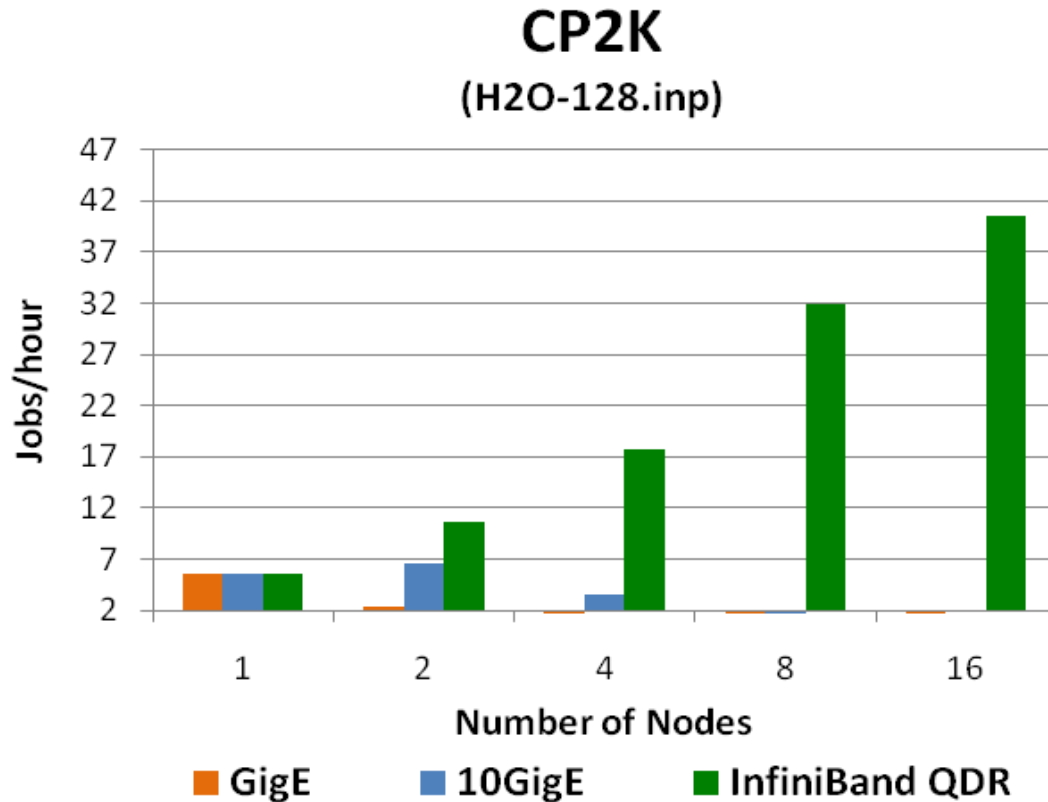


Higher is better

12-cores per node

CP2K Benchmark Results – Interconnects

- **Only InfiniBand enables CP2K application scalability**
 - GigE can't scale even on 2 nodes
 - 10GigE stops scaling after 2 nodes

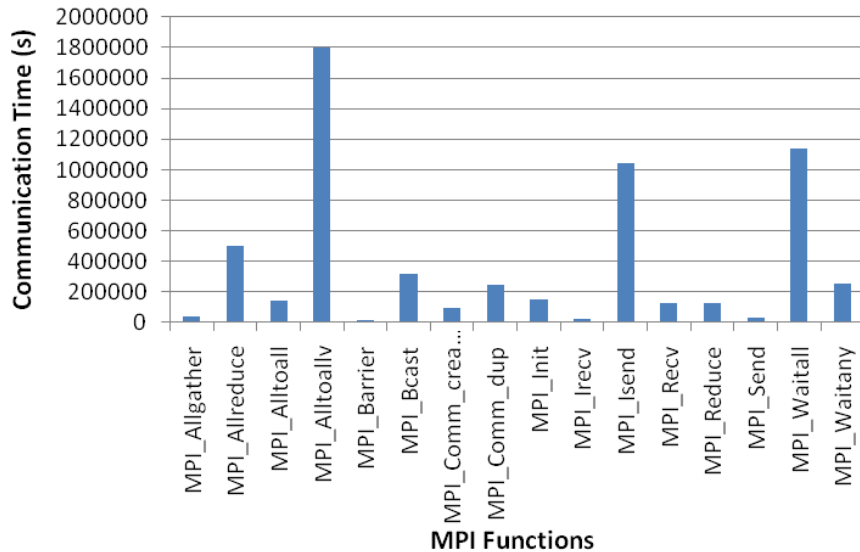


Higher is better

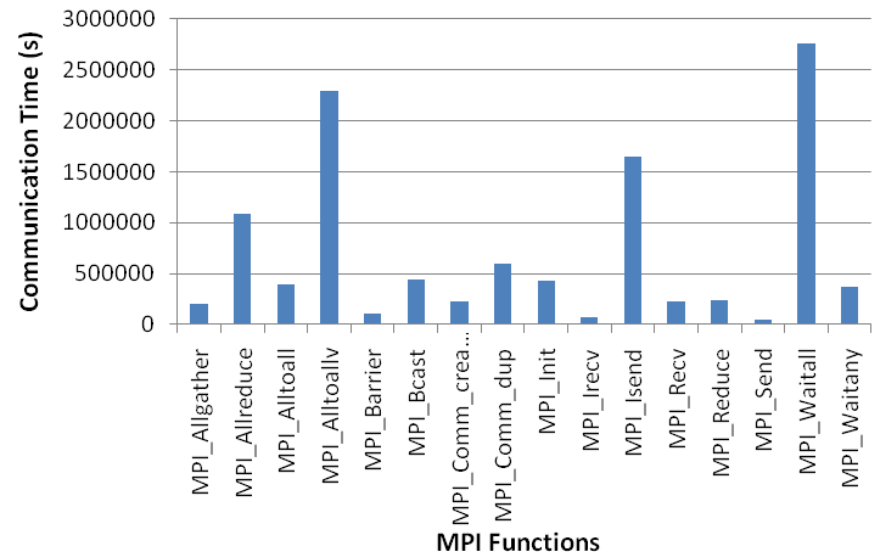
12-cores per node

- **Both MPI collectives and point-to-point creates big communication time**
 - MPI_Alltoallv and MPI_Allreduce
 - MPI_Waitall and MPI_Isend

CP2K Profiling
96 Ranks



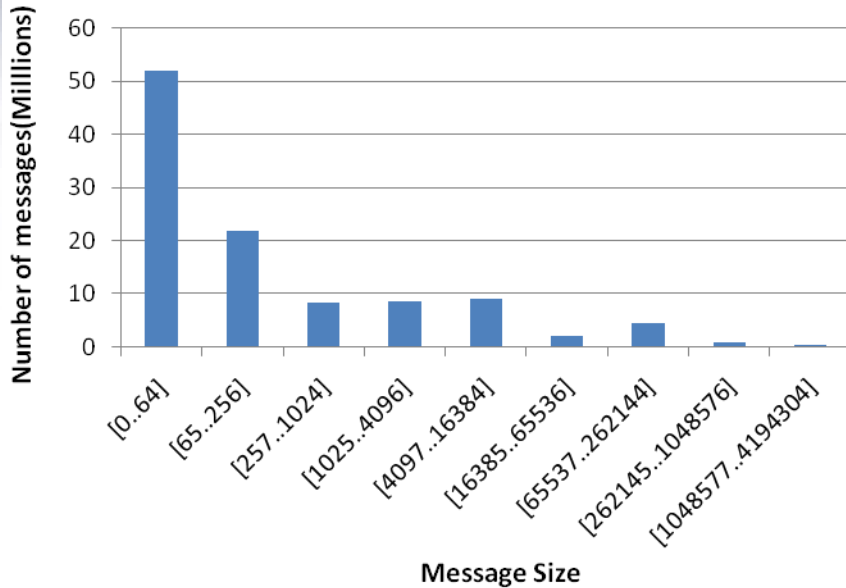
CP2K Profiling
192 Ranks



- Majority messages are smaller than 64KB

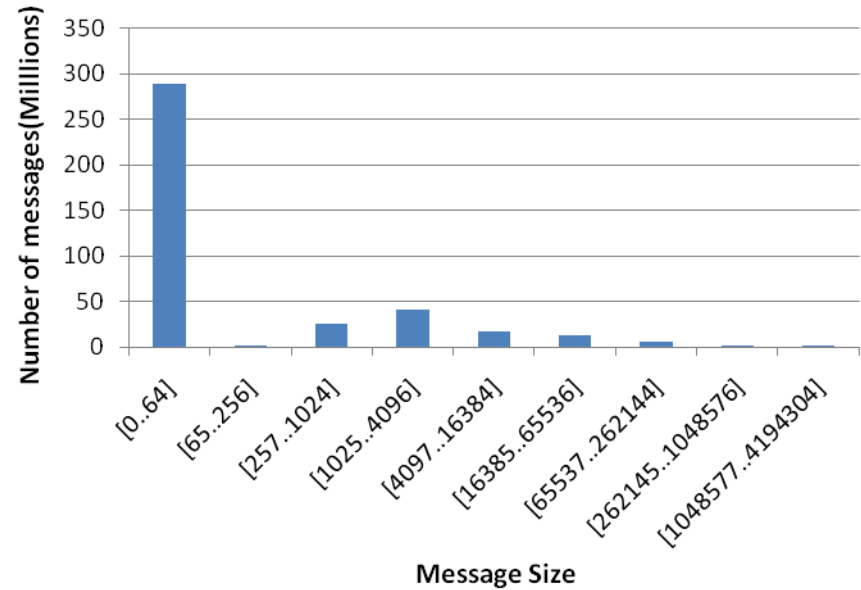
CP2K Profiling

96 Ranks



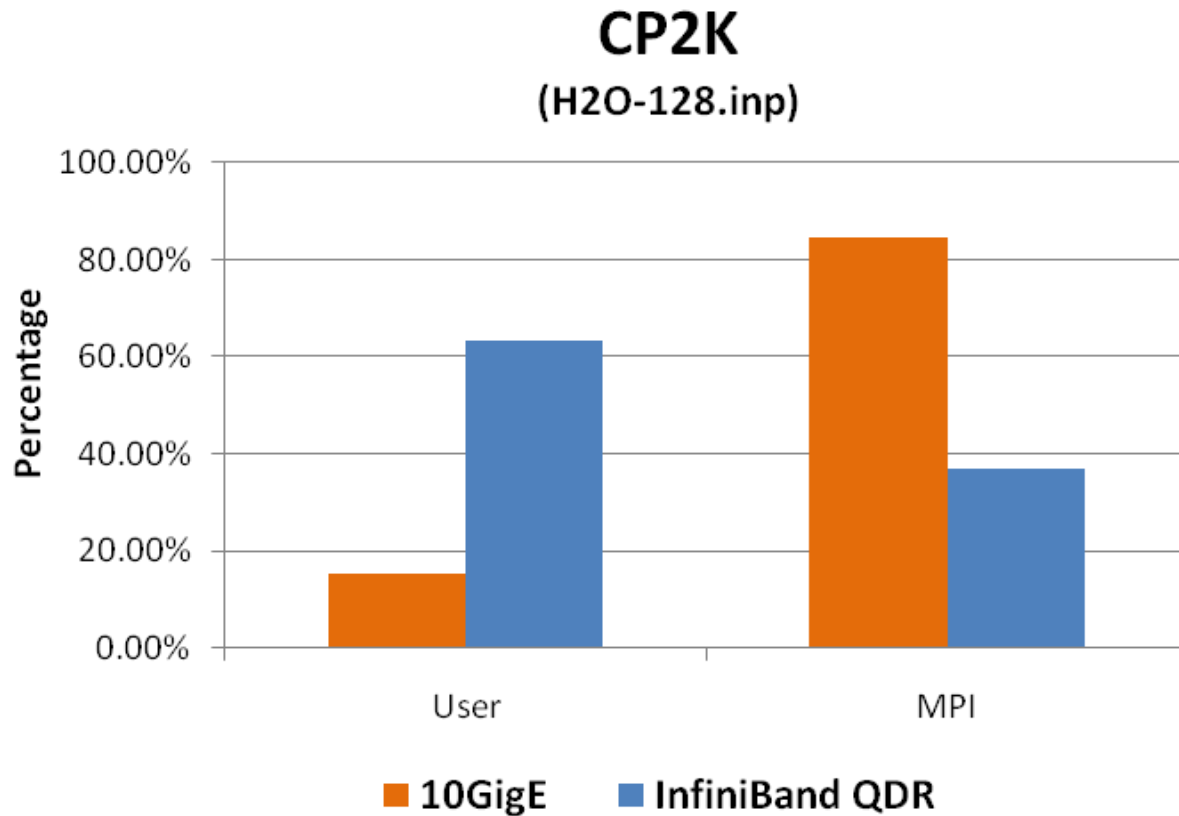
CP2K Profiling

192 Ranks



CP2K MPI Profiling – 10GigE vs IB

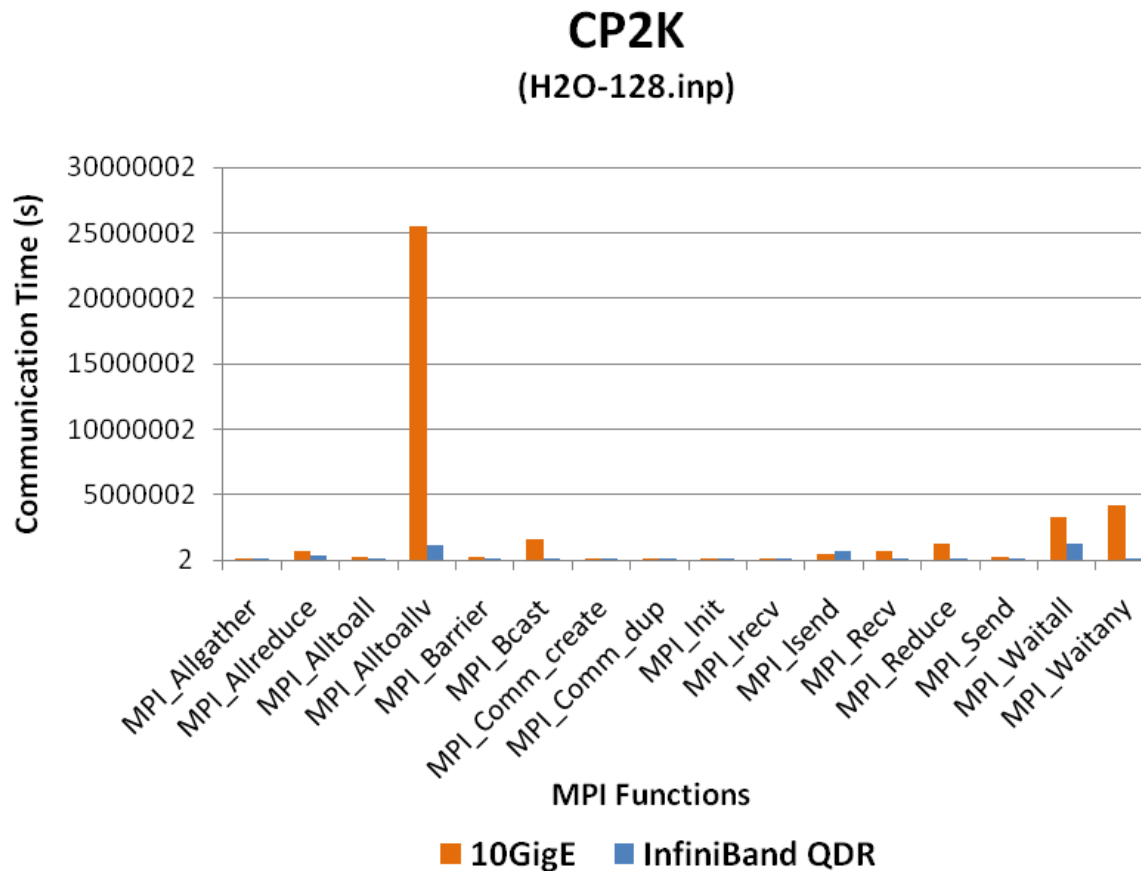
- 10GigE spends most time in communication rather than computation



48 Ranks

CP2K MPI Profiling – 10GigE vs IB

- **10GigE has big communication overhead with MPI collectives**
 - 22 times longer than InfiniBand QDR



48 Ranks

- **CP2K performance benchmark demonstrates**
 - InfiniBand QDR enables application performance and scalability
 - Neither GigE nor 10GigE meet CP2K network requirement
- **CP2K MPI profiling**
 - Large number of small messages are used by CP2K
 - MPI_Alltoallv, MPI_Alltoall, and MPI_Allreduce are major collectives
 - Point-to-point has big communication overhead
 - Interconnect latency is critical to CP2K performance
- **MPI tuning accelerates CP2K performance**
 - More than 20% at 16 nodes / 192 cores

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