NETWORK OF EXPERTISE

CPMD Performance Benchmark, Profiling and Tuning

November 2010

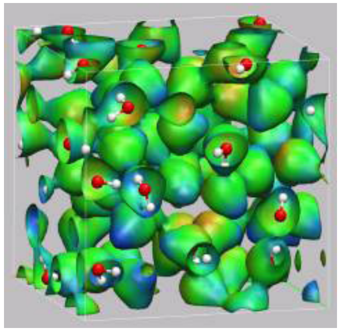
Note

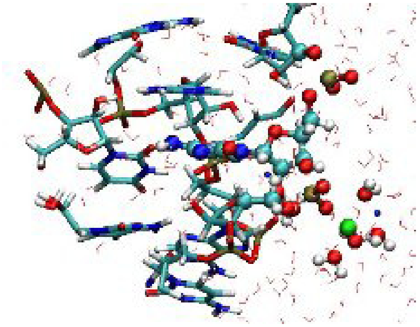
- **The following research was performed under the HPC Advisory Council activities**
  - Participating vendors: Intel, Dell, Mellanox
  - Compute resource - HPC Advisory Council Cluster Center

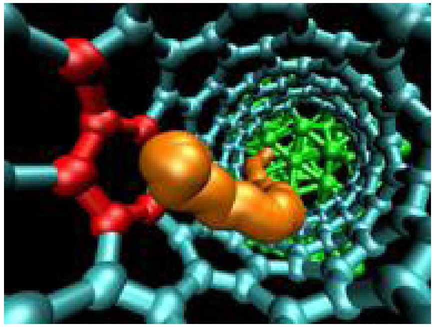
- **For more info please refer to**
  - http://www.dell.com
  - http://www.intel.com
  - http://www.mellanox.com
  - http://www.cpmd.org

Car-Parrinello Molecular Dynamics (CPMD)

- **CPMD**
  - A parallelized implementation of density functional theory (DFT)
  - Particularly designed for ab-initio molecular dynamics
  - Brings together methods
    - Classical molecular dynamics
    - Solid state physics
    - Quantum chemistry
- **CPMD supports MPI and Mixed MPI/SMP**
- **CPMD is distributed and developed by the CPMD consortium**







Objectives

- **The following was done to provide best practices**
  - – CPMD performance benchmarking
  - – Interconnect performance comparisons
  - – Understanding CPMD communication patterns
  - – Power-efficient simulations

- **The presented results will demonstrate**
  - – The scalability of the compute environment to provide nearly linear application scalability
  - – The capability of CPMD to achieve scalable productivity
  - – Considerations for power saving through balanced system configuration

Test Cluster Configuration

- **Dell™ PowerEdge™ M610 14-node cluster**
  - Six-Core Intel X5670 @ 2.93 GHz CPUs
  - Memory: 24GB memory, DDR3 1333 MHz
  - OS: CentOS5U4, OFED 1.5.1 InfiniBand SW stack
- **Intel Cluster Ready certified cluster**
- **Mellanox ConnectX-2 InfiniBand adapters and switches**
- **MPI: Intel MPI 4.0 U1, Open MPI 1.5, Platform MPI 8.0.1**
- **Compilers: GNU Compilers 4.1.2, Intel Compilers 12.0.0**
- **Math Libraries: ATLAS 3.8.3, BLAS 3.0.8, LAPACK, FFTW 2.1.5, MKL 10.3**
- **Application: CPMD 3.13.2_01**
- **Benchmark: Si512 Inp-1**

About Intel® Cluster Ready

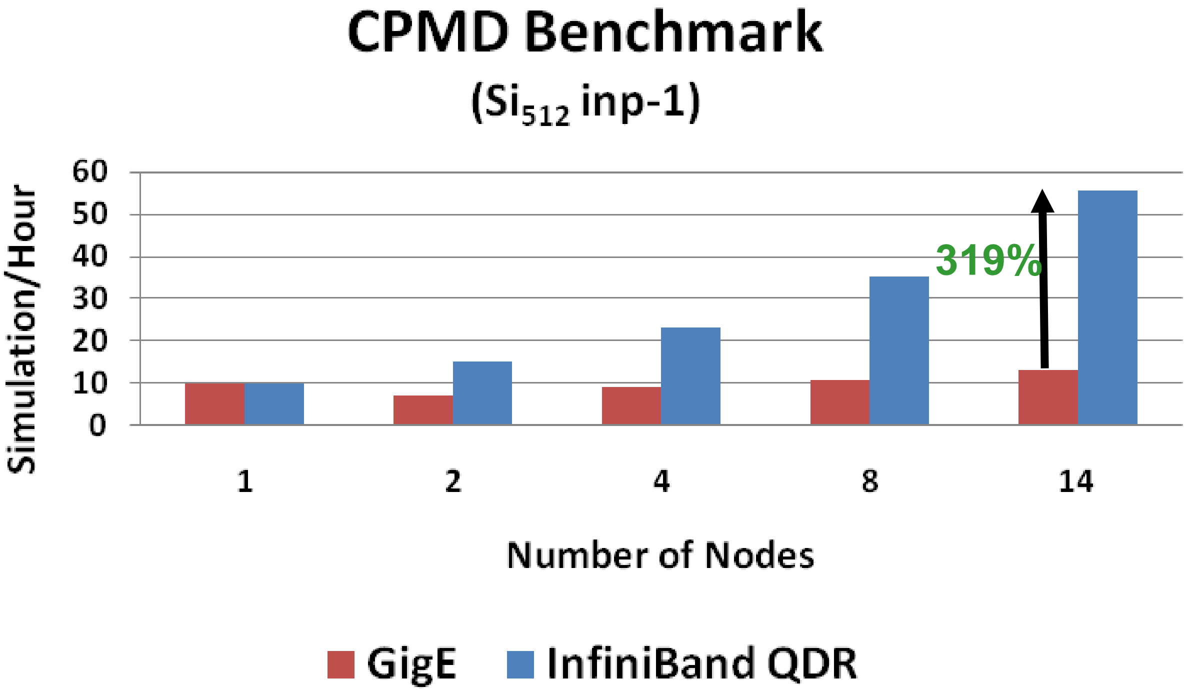
- **Intel® Cluster Ready systems make it practical to use a cluster to increase your simulation and modeling productivity**
  - Simplifies selection, deployment, and operation of a cluster
- **A single architecture platform supported by many OEMs, ISVs, cluster provisioning vendors, and interconnect 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

Dell PowerEdge Servers helping Simplify IT

- **System Structure and Sizing Guidelines**
  - 14-node cluster build with Dell PowerEdge™ M610 blades server
  - 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

CPMD Performance – Interconnect

- **InfiniBand enables higher scalability**
  - Up to 319% higher performance than Ethernet at 14 nodes
- **Ethernet would not scale beyond 1 node**
  - Show virtually no gain by increasing nodes

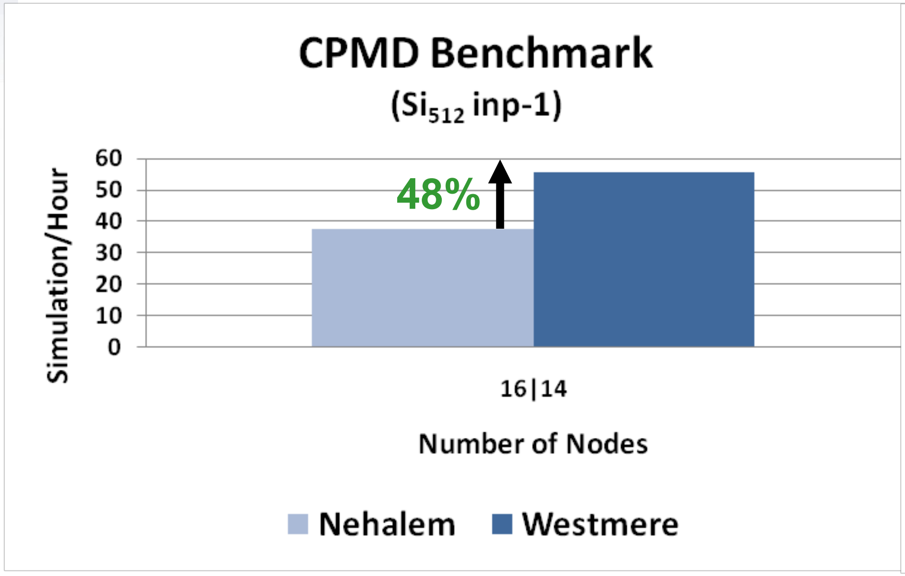


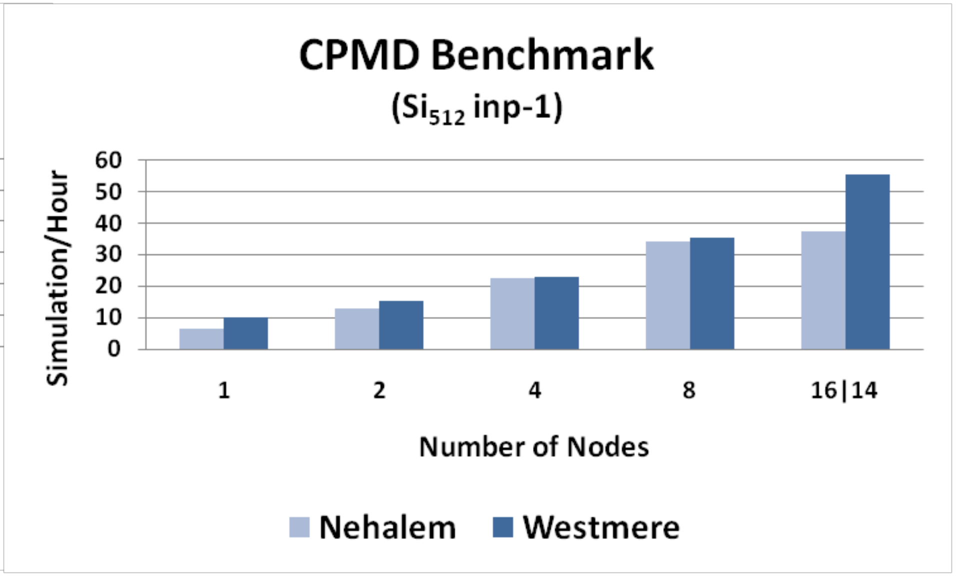
Higher is better

12 Cores/Node

CPMD Performance – Nehalem vs Westmere

- **Westmere processors enabled better performance**
  - Up to 48% gain with 14 Westmere node compared to 16 Nehalem nodes



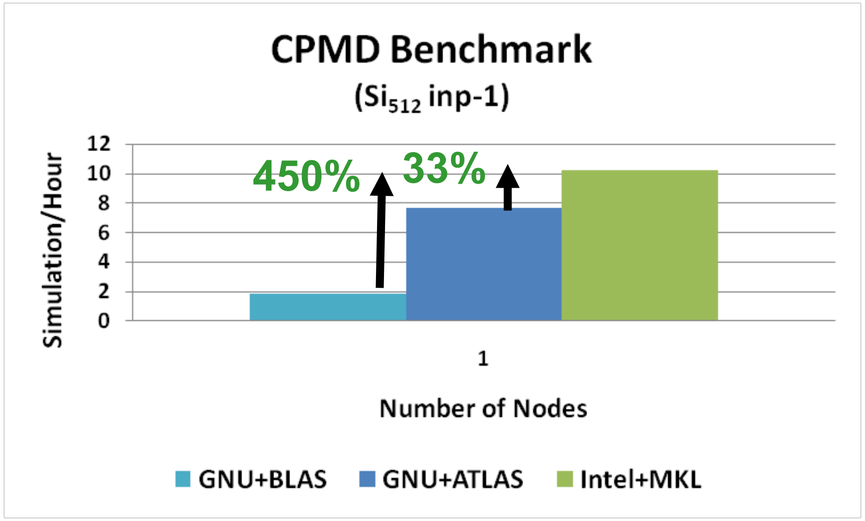


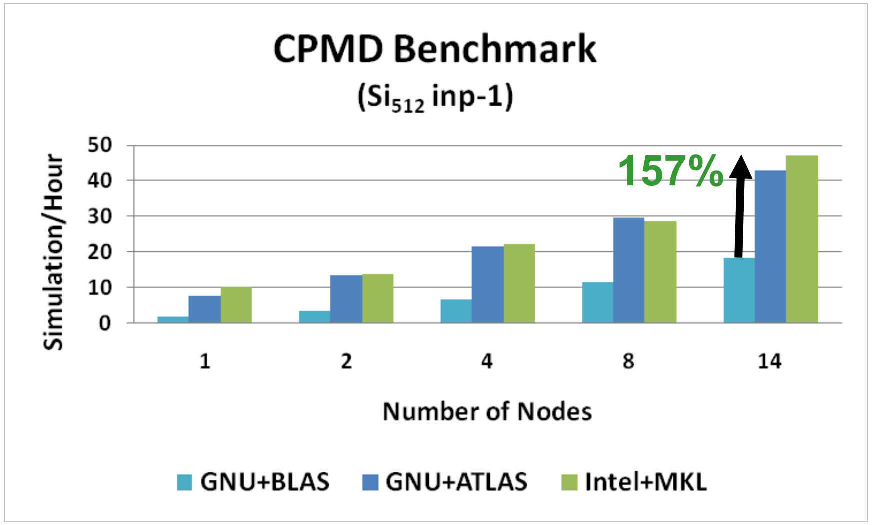
Higher is better

12 Cores/Node

CPMD Performance – Compilers and Libraries

- **Intel compilers and MKL enable the best performance**
  - Up to 450% gain on a single node versus GNU compilers with the BLAS libraries
  - Up to 33% gain on a single node versus GNU compilers with the ATLAS libraries
- **ATLAS can be a good alternative to BLAS**
  - ATLAS (Automatically Tuned Linear Algebra Software)





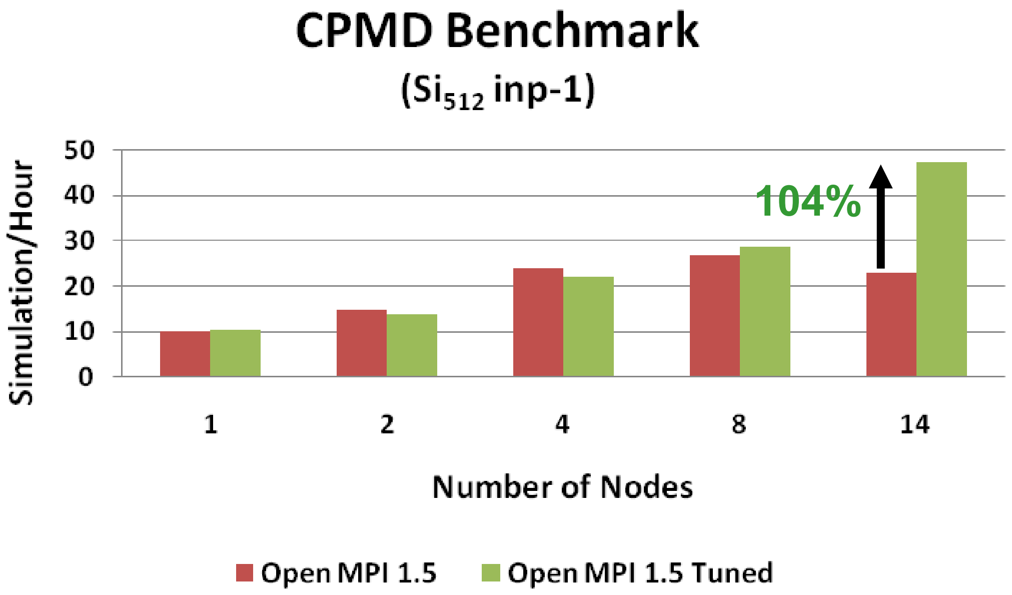
Open MPI 1.5

12 Cores/Node

Higher is better

CPMD Performance – Open MPI Tuning

- **Selecting tuned collectives algorithms can provide boost performance**
  - Up to 104% in performance improvement
  - The difference is more apparent on larger number of nodes or processes
  - Tuning MPI_Alltoall and MPI_Allreduce can make a positive impact
- **Optimize Open MPI using the following MCA parameters**
  - coll_tuned_use_dynamic_rules 1, coll_tuned_alltoall_algorithm 3, coll_tuned_allreduce_algorithm 4, mpi_paffinity_alone 1

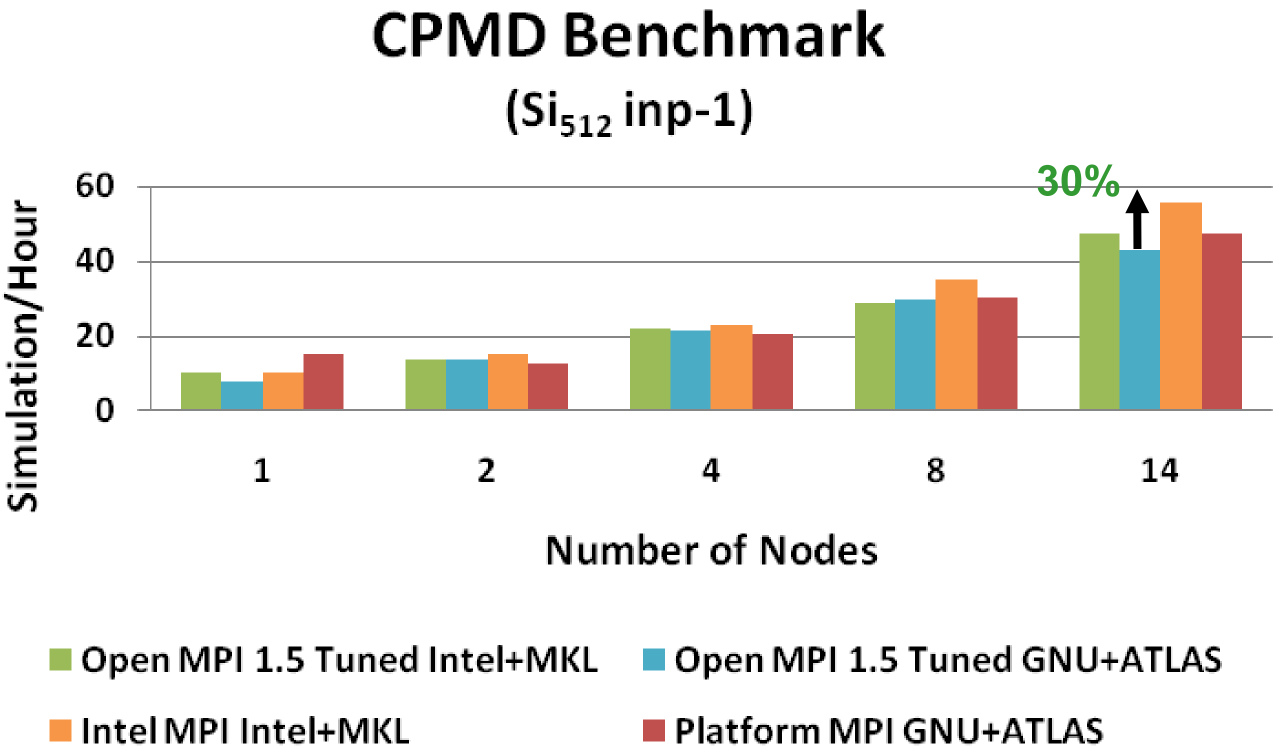


Higher is better

12 Cores/Node

CPMD Performance – MPI

- **Intel MPI shows better scalability over the tuned Open MPI**
  - Shows 30% gain over tuned Open MPI with GNU compilers and ATLAS

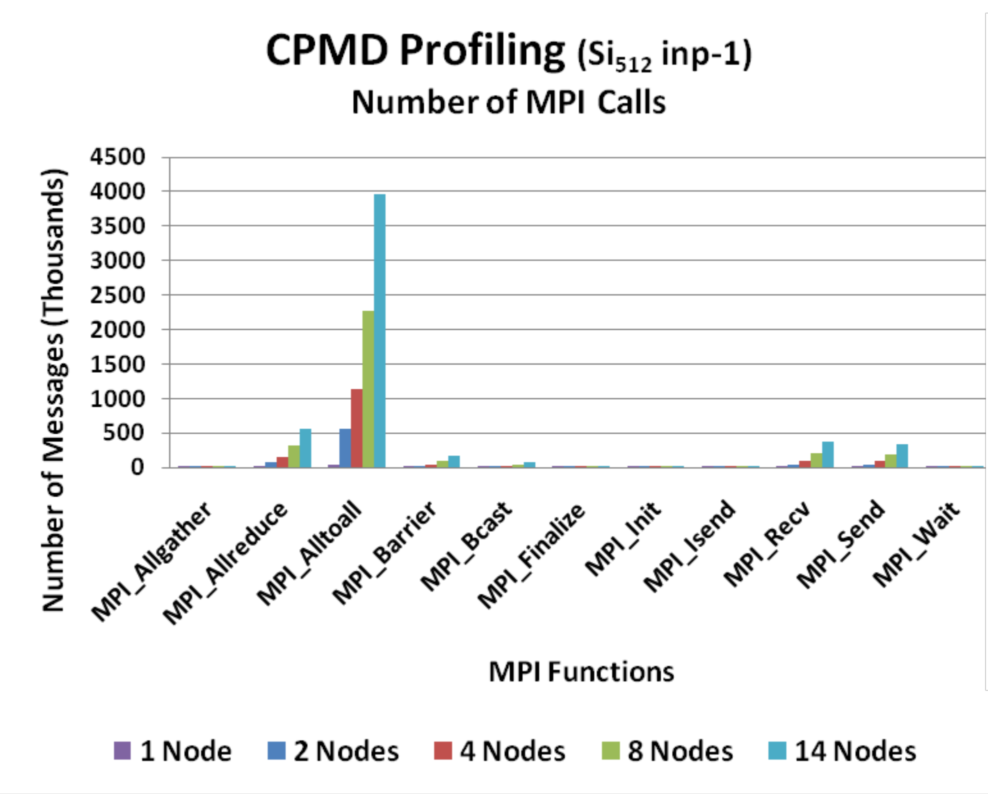


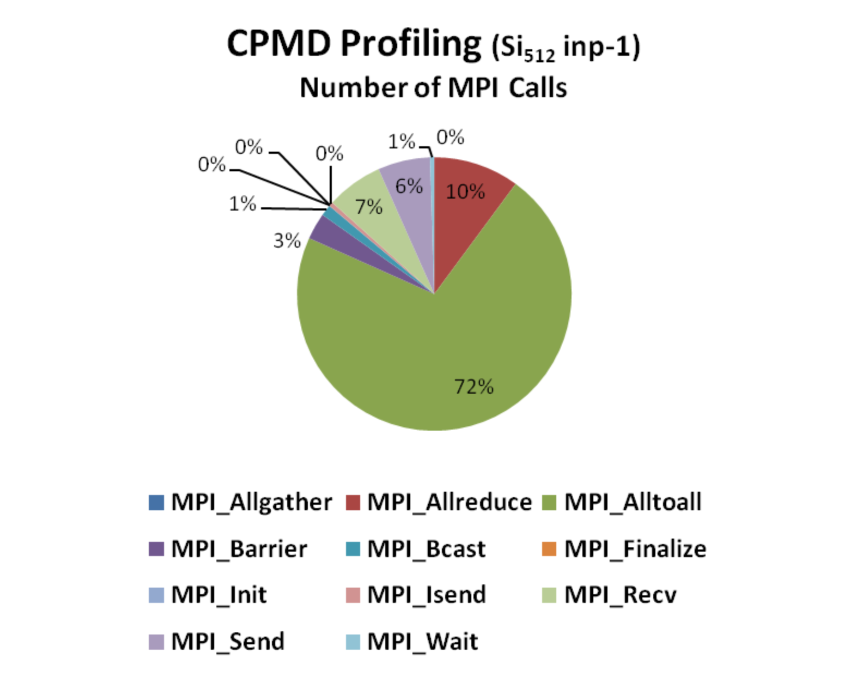
Higher is better

12 Cores/Node

CPMD Profiling – Number of MPI Calls

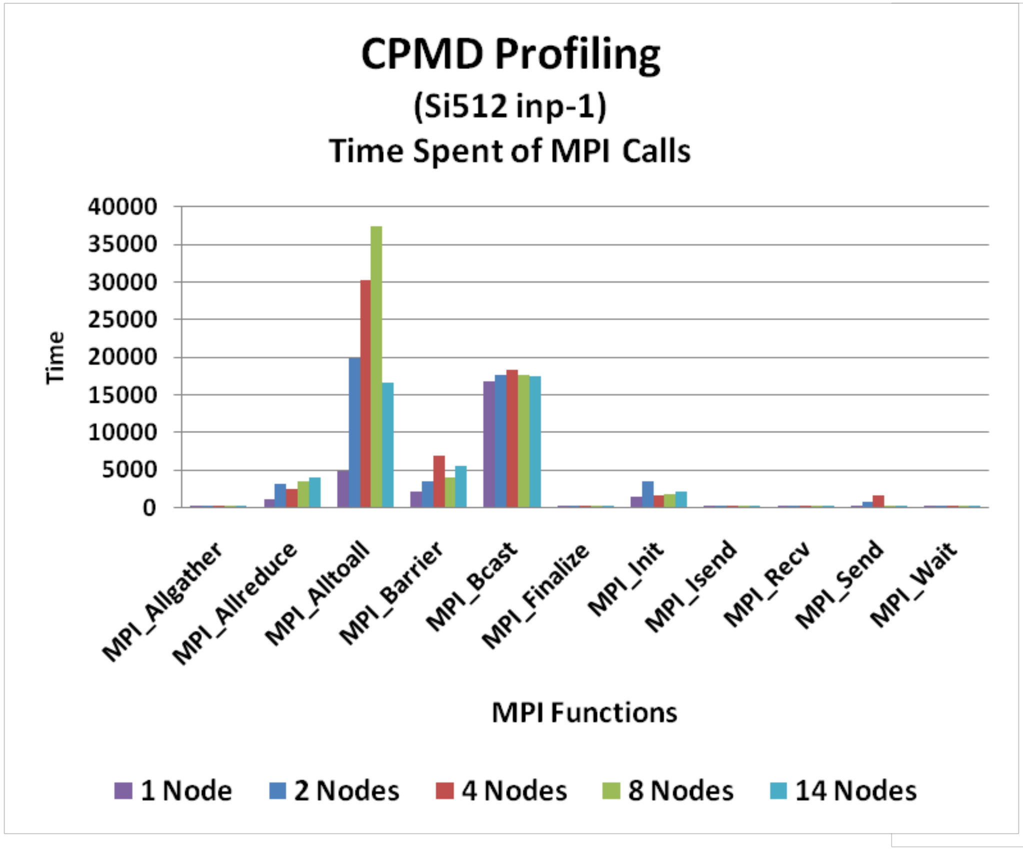
- **The most used MPI functions with this dataset is MPI_Alltoall**
- **MPI_Alltoall accounted for 72% of all MPI calls on a 14-node job**

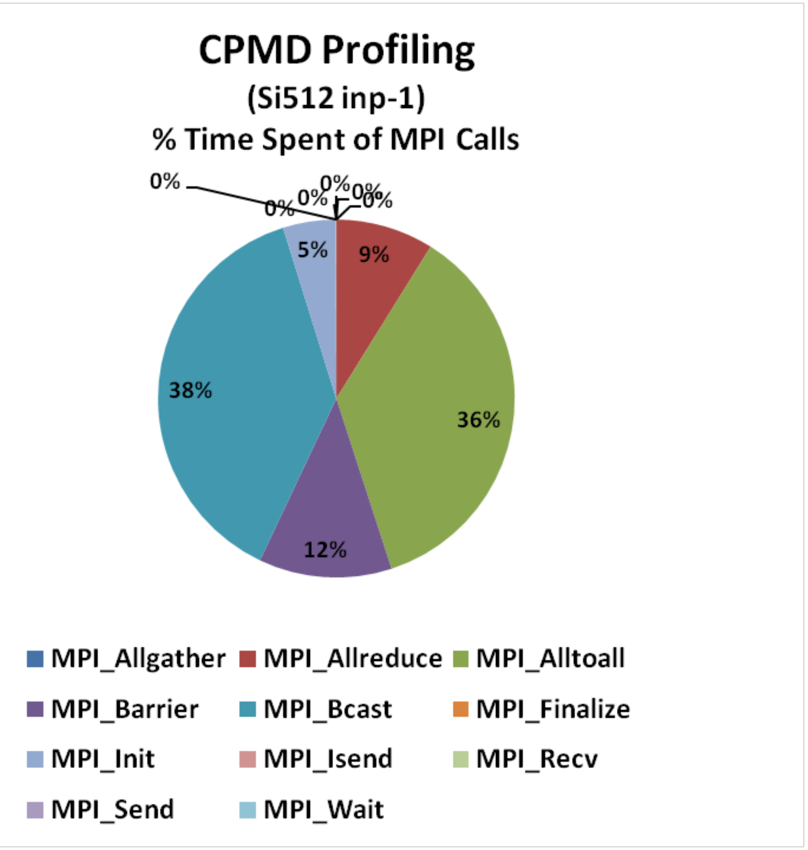




CPMD Profiling – Time Spent of by MPI Calls

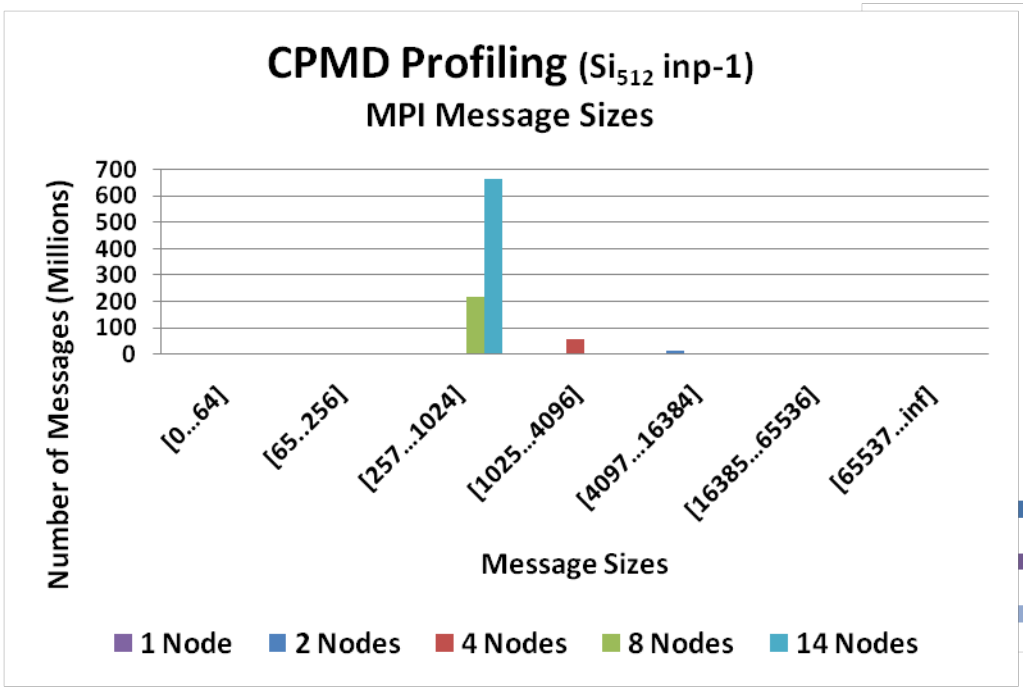
- **Majority of time is spent on MPI_Alltoall and MPI_Bcast**
  - MPI_Alltoall is accounted for 38% of time spent on a 14-node job
  - MPI_Bcast is accounted for 36% of time spent on a 14-node job

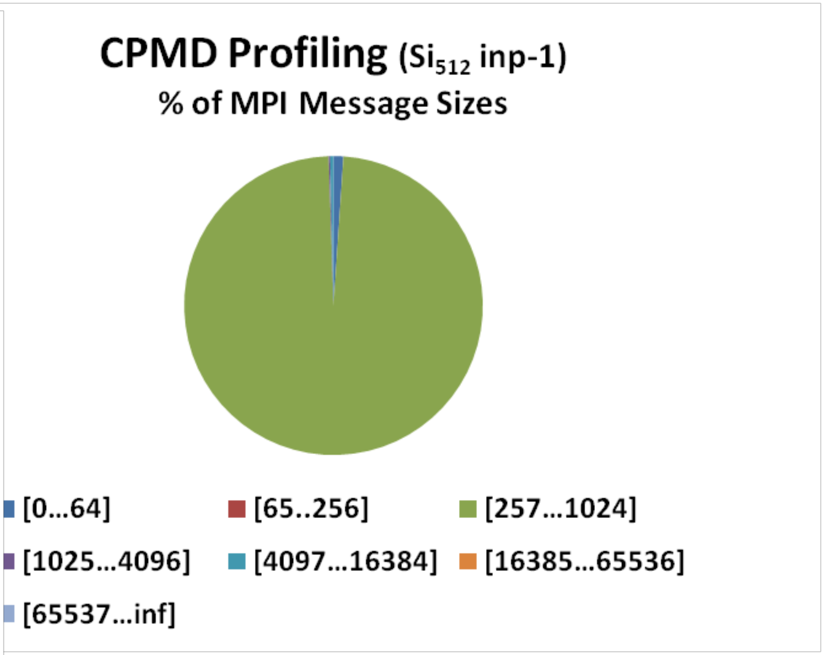




CPMD Profiling – MPI Message Sizes

- **Majority of messages are small messages**
  - Messages between 256B and 1KB are the majority for the 14-node and 8-node runs
  - Accounted for 98% of the MPI message sizes on the 14 nodes





Summary

- **Interconnects effect to CPMD performance**
  - InfiniBand enables higher performance/scalability
  - Up to 319% higher performance than Ethernet at 14 nodes
- **Intel Westmere processor delivers better performance**
  - Up to 48% gain on a 14-node Westmere processors versus 16-node on Nehalem processors
- **Intel compilers and MKL provides better performance**
  - Up to 450% gain on a single node over GNU compilers with BLAS
  - Up to 33% gain on a single node over GNU compilers with ATLAS
- **Intel MPI shows better scalability over the tuned Open MPI**
- **Tuning MPI_Alltoall and MPI_Allreduce provide up to 104% improvement for Open MPI**
- **Majority of MPI messages are between 256B and 1KB**
- **MPI_Alltoall is the most used MPI functions**
- **Majority of MPI time is spent on MPI_Alltoall and MPI_Bcast**

Thank You

HPC Advisory Council

DELL intel Mellanox TECHNOLOGIES

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