CPMD Performance Benchmark and Profiling

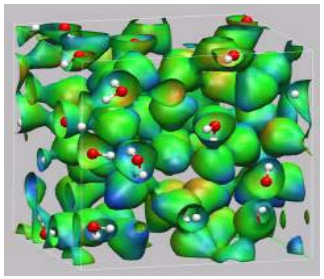
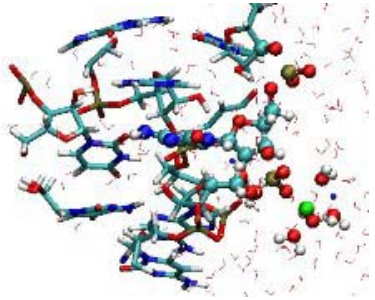
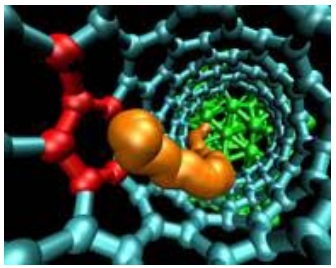
Feb 2010

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**
  - www.mellanox.com, www.dell.com/hpc, www.intel.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 presented research was done to provide best practices**
  - CPMD performance benchmarking
    - MPI Library performance comparisons
    - Interconnect performance comparisons
  - Understanding CPMD communication patterns
  - Power-efficient simulations
- **The presented results will demonstrate**
  - The scalability of the compute environment to provide good application scalability
  - Considerations for power saving through balanced system configuration

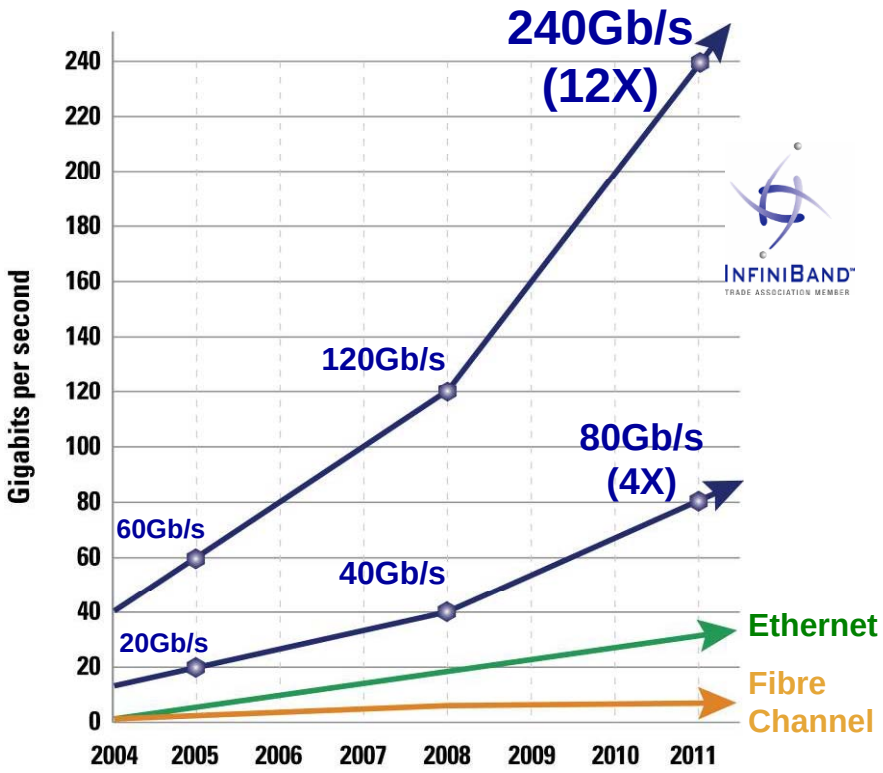
Test Cluster Configuration

- **Dell™ PowerEdge™ M610 16-node cluster**
- **Quad-Core Intel X5570 @ 2.93 GHz CPUs**
- **Intel Cluster Ready certified cluster**
- **Mellanox ConnectX MCQH29-XCC 4X QDR InfiniBand mezzanine card**
- **Mellanox M3601Q 32-Port Quad Data Rate (QDR-40Gb) InfiniBand Switch**
- **Memory: 24GB memory per node**
- **OS: RHEL5U3, OFED 1.5 InfiniBand SW stack**
- **MPI: Open MPI 1.3.3, MVAPICH2-1.4, Intel MPI 4.0**
- **Application: CPMD 3.13.2**
- **Benchmark Workload**
  - **Si512 (Wavefunction cutoff 20 Ry, no. of plane waves ~320K, Real space mesh 108X108X108)**
    - Wavefunction cutoff 20 Ry (inp-1 and inp-2)
    - Wavefunction cutoff 60 Ry (inp-1.60 and inp-2.60)

Mellanox InfiniBand Solutions

- **Industry Standard**
  - Hardware, software, cabling, management
  - Design for clustering and storage interconnect
- **Performance**
  - 40Gb/s node-to-node
  - 120Gb/s switch-to-switch
  - 1us application latency
  - Most aggressive roadmap in the industry
- **Reliable with congestion management**
- **Efficient**
  - RDMA and Transport Offload
  - Kernel bypass
  - CPU focuses on application processing
- **Scalable for Petascale computing & beyond**
- **End-to-end quality of service**
- **Virtualization acceleration**
- **I/O consolidation Including storage**

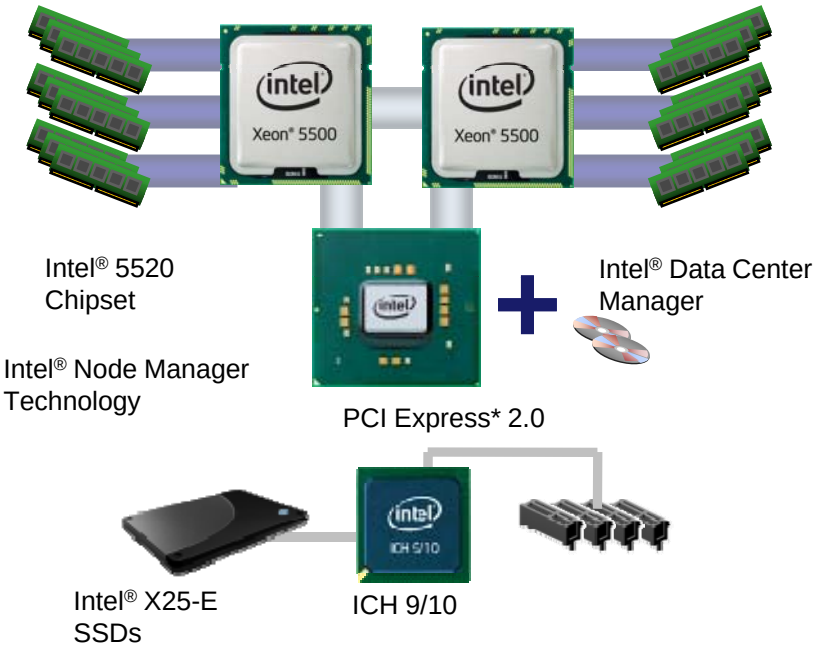
The InfiniBand Performance Gap is Increasing



InfiniBand Delivers the Lowest Latency

Delivering Intelligent Performance

Next Generation Intel® Microarchitecture



Bandwidth Intensive

- Intel® QuickPath Technology
- Integrated Memory Controller

Threaded Applications

- 45nm quad-core Intel® Xeon® Processors
- Intel® Hyper-threading Technology

Performance on Demand

- Intel® Turbo Boost Technology
- Intel® Intelligent Power Technology

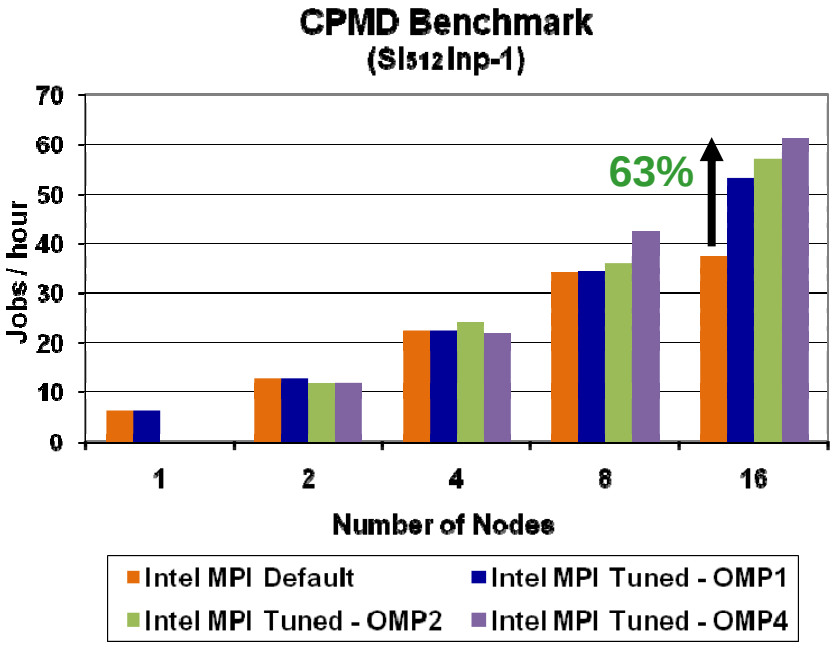
Performance That Adapts to The Software Environment

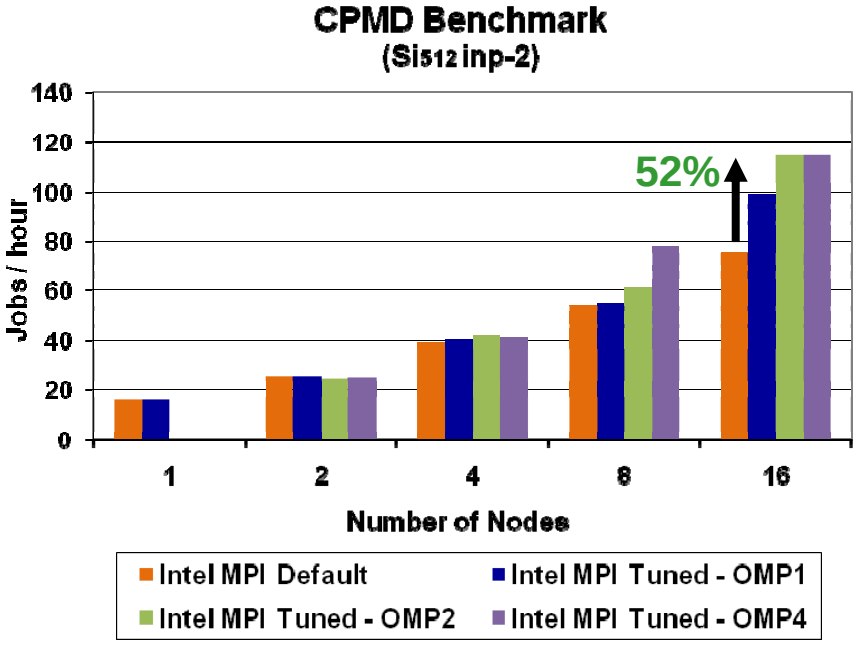
Dell PowerEdge Servers helping Simplify IT

- **System Structure and Sizing Guidelines**
  - 16-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 Benchmark Results – Intel MPI

- **Input Dataset: Si512**
  - inp-1: Wavefunction optimization
  - inp-2: Molecular dynamics simulation
- **Customized MPI settings dramatically improve CPMD performance**
  - I_MPI_ADJUST_ALLTOALL 4, I_MPI_ADJUST_ALLREDUCE 5
- **Intel MPI with OpenMP enabled further enhances application performance**



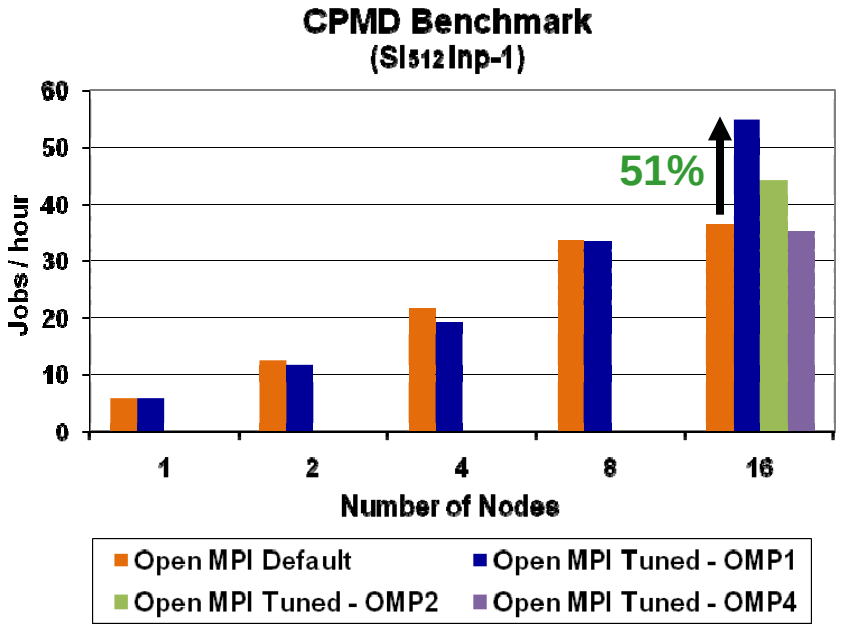


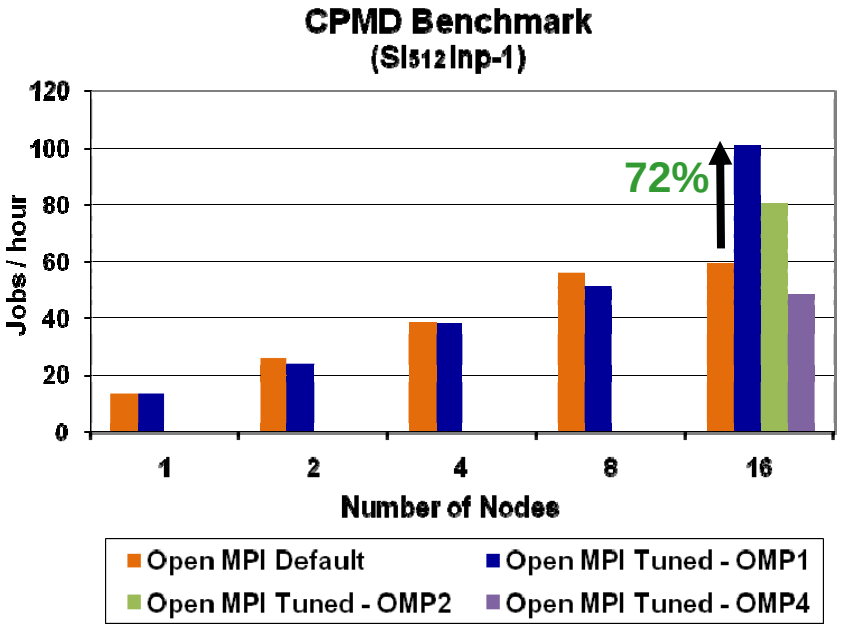
Higher is better

8-cores per node

CPMD Benchmark Results – Open MPI

- **Optimal Alltoall algorithm can boost CPMD performance by up to 72%**
  - mca mpi_paffinity_alone 1 --mca coll_tuned_use_dynamic_rules 1 --mca coll_tuned_alltoall_algorithm 3 --mca coll_tuned_allreduce_algorithm 4
- **Open MPI with OpenMP does not provide performance advantage**



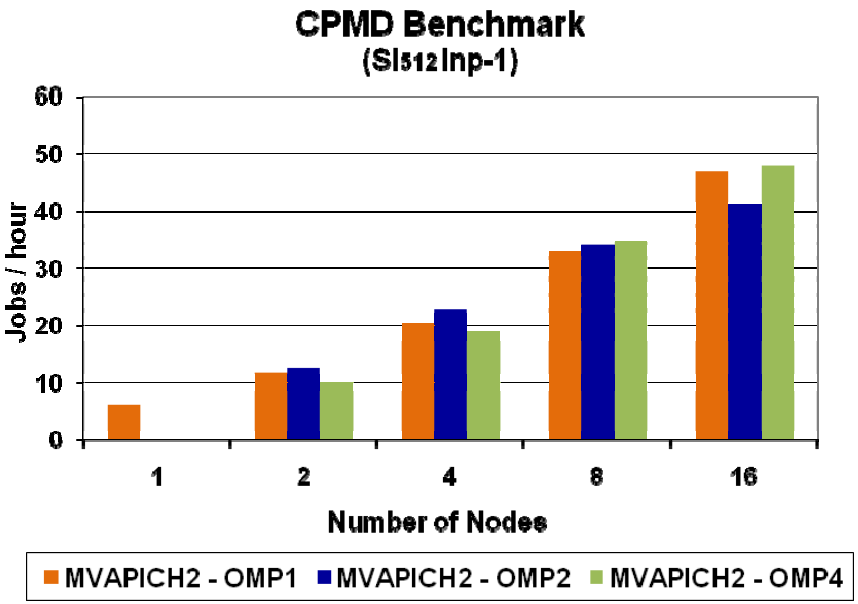


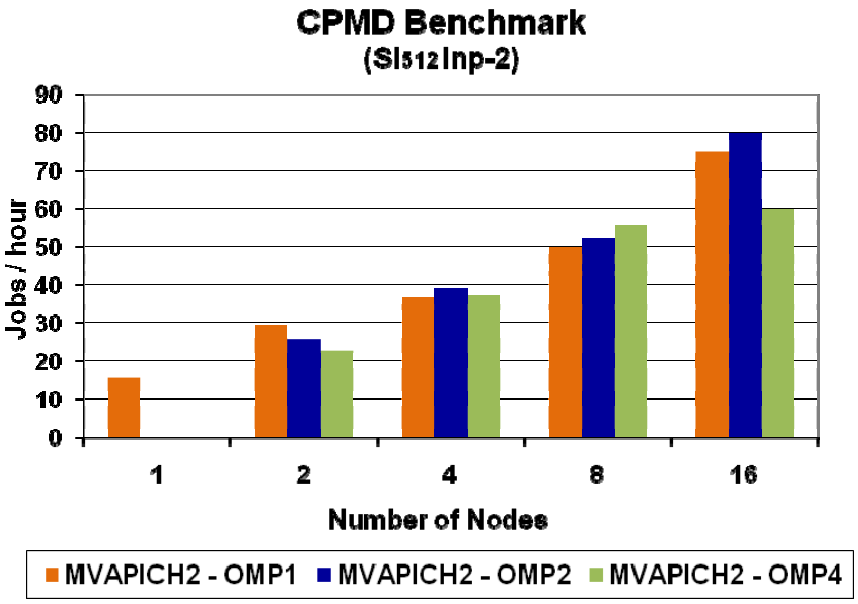
Higher is better

8-cores per node

CPMD Benchmark Results – MVAPICH2

- **Multiple threads do not provide major performance increase at this cluster size**



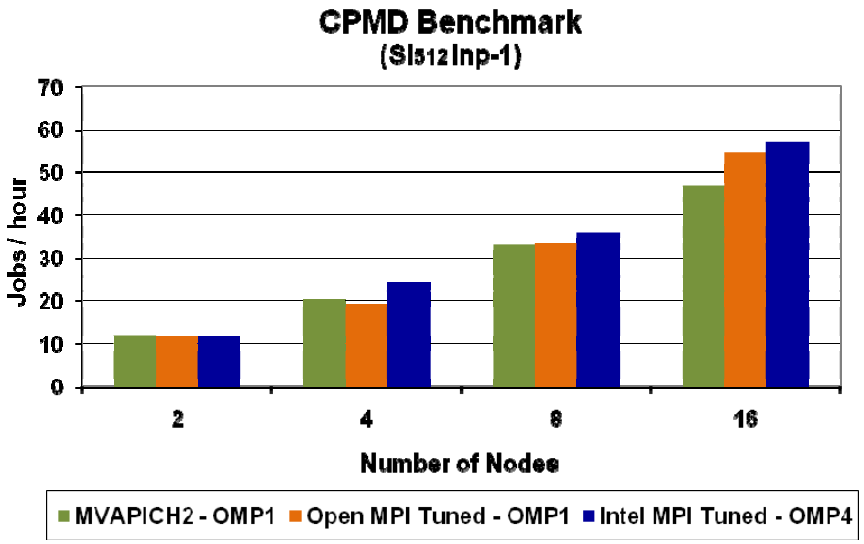


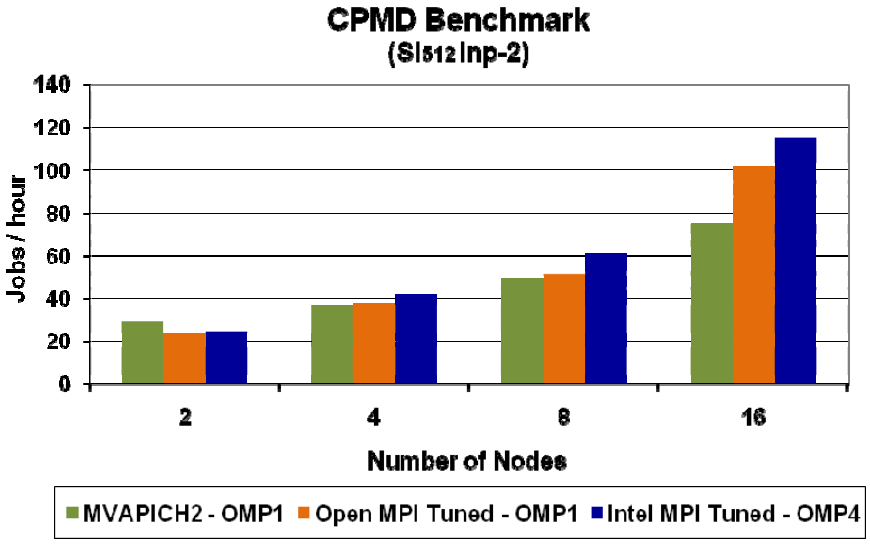
Higher is better

8-cores per node

CPMD Benchmark Results – MPI Comparison

- **Performance scales with all three MPI implementations over InfiniBand QDR**
- **Intel MPI with threads provides higher performance and scalability**



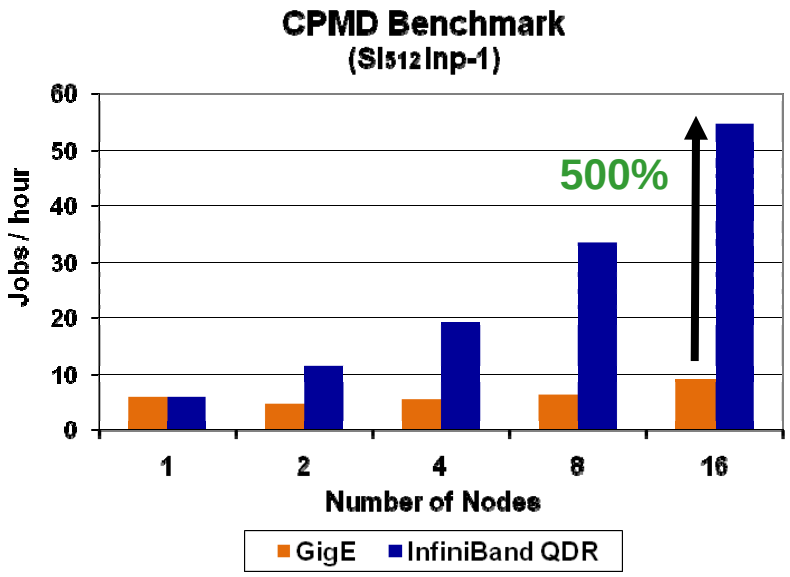


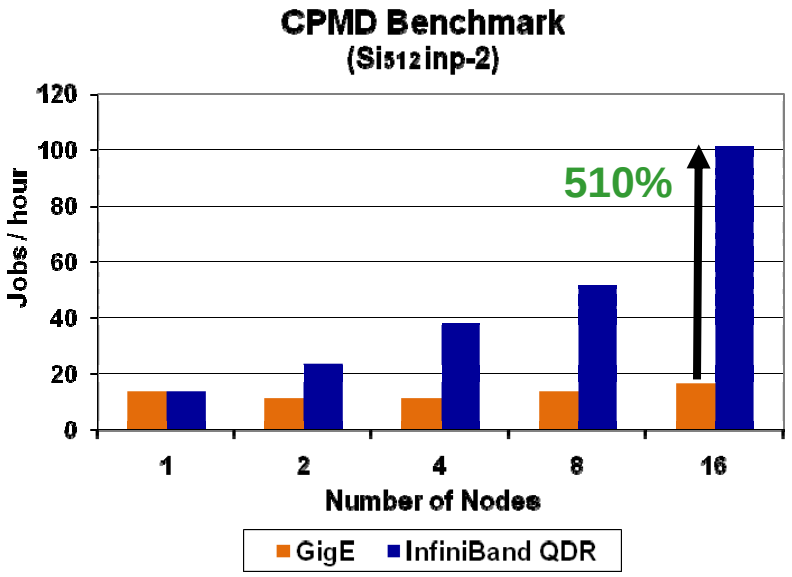
Higher is better

8-cores per node

CPMD Benchmark Results – Interconnect

- **InfiniBand QDR delivers highest application performance and scalability**
  - GigE stops scaling even with two nodes
  - InfiniBand performance scales as cluster size increases



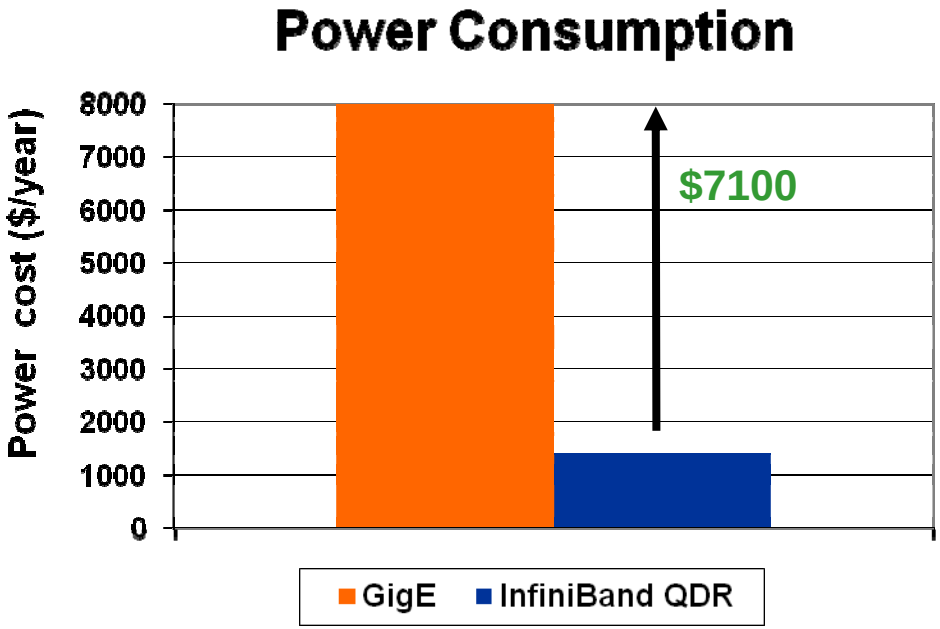


Higher is better

8-cores per node

Power Cost Savings with Different Interconnect

- **InfiniBand saves up to $7100 power to finish the same number of CPMD jobs compared to GigE**
  - Yearly based for 16-node cluster
- **As cluster size increases, more power can be saved**



*$/KWh = KWh * $0.20*
For more information - http://enterprise.amd.com/Downloads/svrpwrusecompletefinal.pdf

CPMD Benchmark Results Summary

- **MPI Performance Optimization**
  - MPI_Alltoall is the main MPI routine impacts CPMD performance
  - Optimal MPI_Alltoall settings can dramatically enhance CPMD performance
  - Intel MPI enables best performance with MPI/OpenMP combination
- **Interconnect comparison shows**
  - Both Interconnect latency and bandwidth are important to CPMD performance
  - InfiniBand delivers superior performance in every cluster size
  - GigE performance can't scale beyond one node
- **InfiniBand enables power saving**
  - Up to $7200/year power savings versus GigE
- **Balanced system – CPU, memory, Interconnect that match each other capabilities, is essential for providing application efficiency**

Productive Systems = Balanced System

- **Balanced system enables highest productivity**
  - Interconnect performance to match CPU capabilities
  - CPU capabilities to drive the interconnect capability
  - Memory bandwidth to match CPU performance
- **Applications scalability relies on balanced configuration**
  - “Bottleneck free”
  - Each system components can reach it’s highest capability
- **Dell M610 system integrates balanced components**
  - Intel “Nehalem” CPUs and Mellanox InfiniBand QDR
    - Latency to memory and Interconnect latency at the same magnitude of order
  - Provide the leading productivity and power/performance system for Desmond simulations

Thank You

HPC Advisory Council