

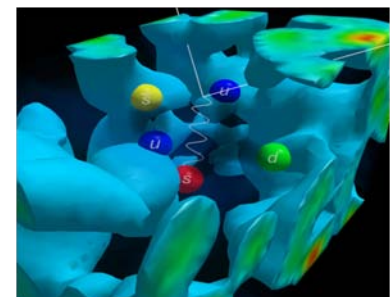
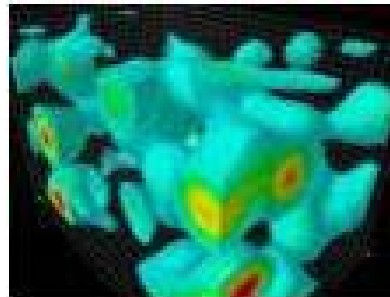
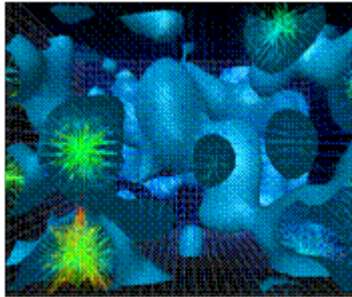
BQCD Performance Benchmark and Profiling

June 2010



- **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**
 - www.mellanox.com, www.dell.com/hpc, www.amd.com
 - <http://www.deisa.eu/science/benchmarking/codes/bqcd>

- **BQCD - Berlin Quantum ChromoDynamics program**
 - Hybrid Monte-Carlo code that simulates Quantum Chromodynamics with dynamical standard Wilson fermions
 - The computations take place on a four-dimensional regular grid with periodic boundary conditions
 - The kernel of the program is a standard conjugate gradient solver with even/odd pre-conditioning
 - The parallelization is done by a regular grid decomposition in the highest 3 dimensions
- **Developed by Dr. Hinnerk Stueben from Konrad-Zuse-Zentrum fuer. Infomationstechnik Berlin**
 - Open source software



- **The presented research was done to provide best practices**
 - BQCD performance benchmarking
 - Performance tuning with different communication libraries and compilers
 - Interconnect performance comparisons
 - Understanding BQCD communication patterns
 - Power-efficient simulations
- **The presented results will demonstrate**
 - Balanced compute system enables
 - Good application scalability
 - Power saving

- **Dell™ PowerEdge™ SC 1435 16-node cluster**
- **Quad-Core AMD Opteron™ 2382 (“Shanghai”) CPUs**
- **Mellanox® InfiniBand ConnectX® 20Gb/s (DDR) HCAs**
- **Mellanox® InfiniBand DDR Switch**
- **Memory: 16GB memory, DDR2 800MHz per node**
- **OS: RHEL5U3, OFED 1.5 InfiniBand SW stack**
- **MPI: OpenMPI-1.4.2, MVAPICH2-1.4.1, Platform MPI 7.1**
- **Application: BQCD**
- **Benchmark Workload**
 - Lattice size: 48 6 12 48

Mellanox Connectivity: Taking HPC to New Heights

World Highest Efficiency

- The world's only full transport-offload
- CORE-Direct - MPI and SHMEM offloads
- GPU-Direct - direct connectivity GPU-IB

World Fastest InfiniBand

- 40Gb/s node to node, 120G IB switch to switch
- Highest dense switch solutions - 51.8TB in a single switch
- World's lowest switch latency – 100ns 100% load

HPC Topologies for Scale

- Fat-tree, mesh, 3D-Torus, Hybrid
- Advanced adaptive routing capabilities
- Highest reliability, lowest bit error rate, real-time adjustments



Quad-Core AMD Opteron™ Processor

- **Performance**

- Quad-Core

- Enhanced CPU IPC
- 4x 512K L2 cache
- 6MB L3 Cache

- Direct Connect Architecture

- HyperTransport™ Technology
- Up to 24 GB/s peak per processor

- Floating Point

- 128-bit FPU per core
- 4 FLOPS/clock peak per core

- Integrated Memory Controller

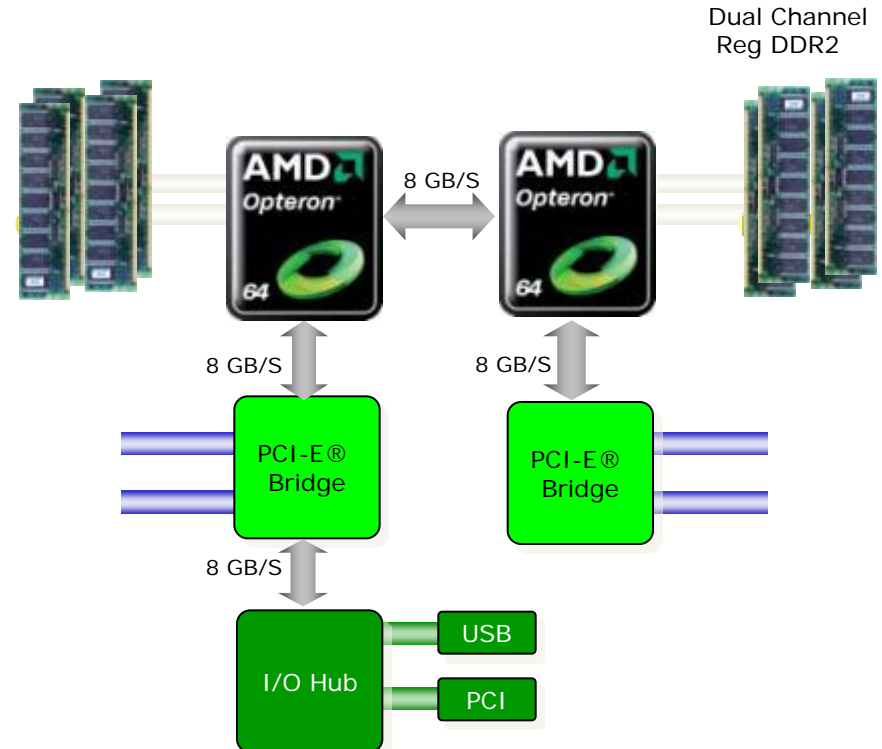
- Up to 12.8 GB/s
- DDR2-800 MHz or DDR2-667 MHz

- **Scalability**

- 48-bit Physical Addressing

- **Compatibility**

- Same power/thermal envelopes as 2nd / 3rd generation AMD Opteron™ processor



- **System Structure and Sizing Guidelines**

- 16-node cluster build with Dell PowerEdge™ SC 1435 Servers
- 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

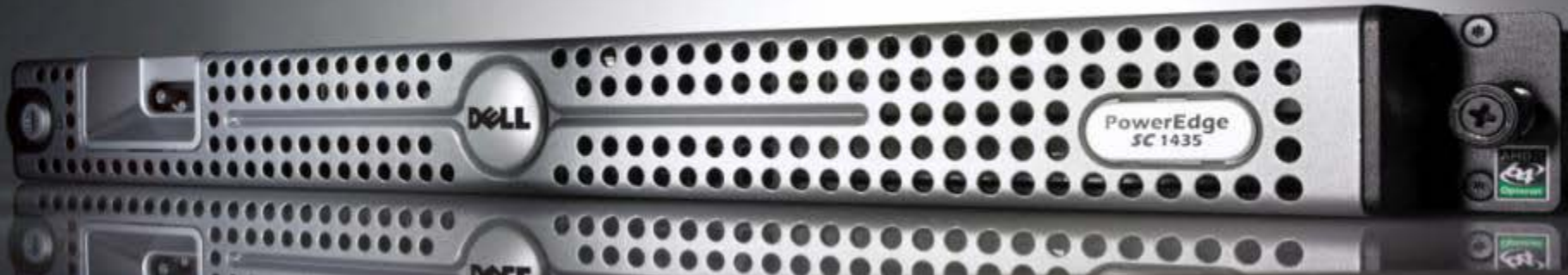
- **Workload Modeling**

- Optimized System Size, Configuration and Workloads
- Test-bed Benchmarks
- ISV Applications Characterization
- Best Practices & Usage Analysis

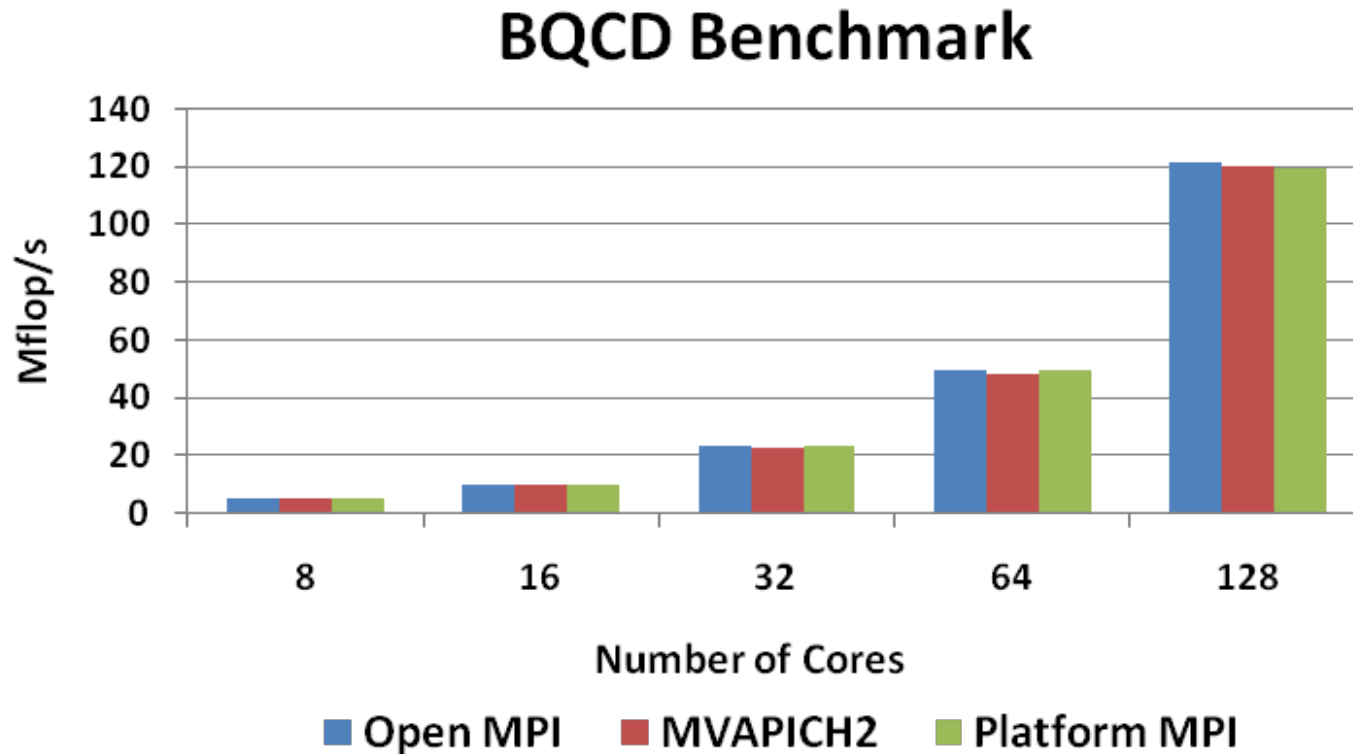


Dell PowerEdge™ Server Advantage

- Dell™ PowerEdge™ servers incorporate AMD Opteron™ and Mellanox ConnectX InfiniBand to provide leading edge performance and reliability
- Building Block Foundations for best price/performance and performance/watt
- Investment protection and energy efficient
- Longer term server investment value
- Faster DDR2-800 memory
- Enhanced AMD PowerNow!
- Independent Dynamic Core Technology
- AMD CoolCore™ and Smart Fetch Technology
- Mellanox InfiniBand end-to-end for highest networking performance



- All three tested MPIs provide similar performance

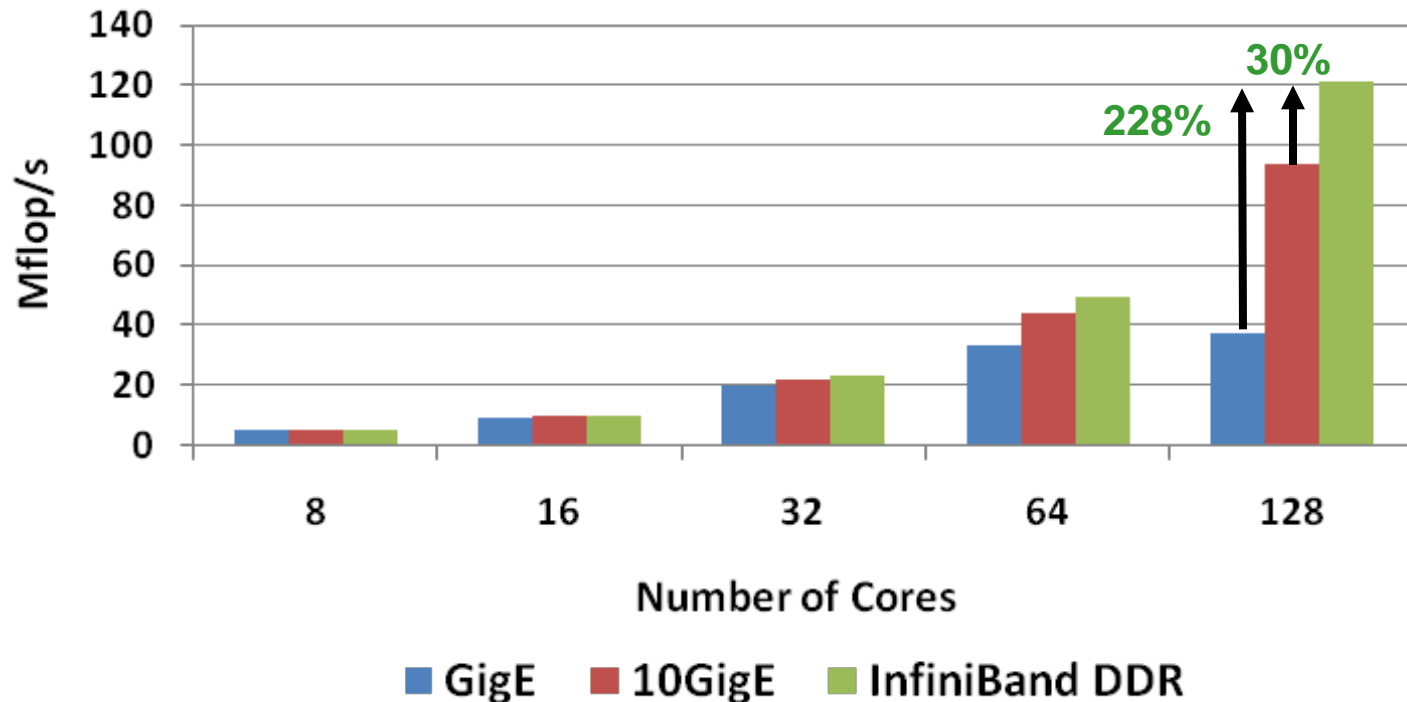


Higher is better

8-cores per node

- **InfiniBand enables higher performance and scalability**
 - Up to 228% higher performance than GigE and 30% higher than 10GigE
 - Performance difference increases as cluster size scales

BQCD Benchmark

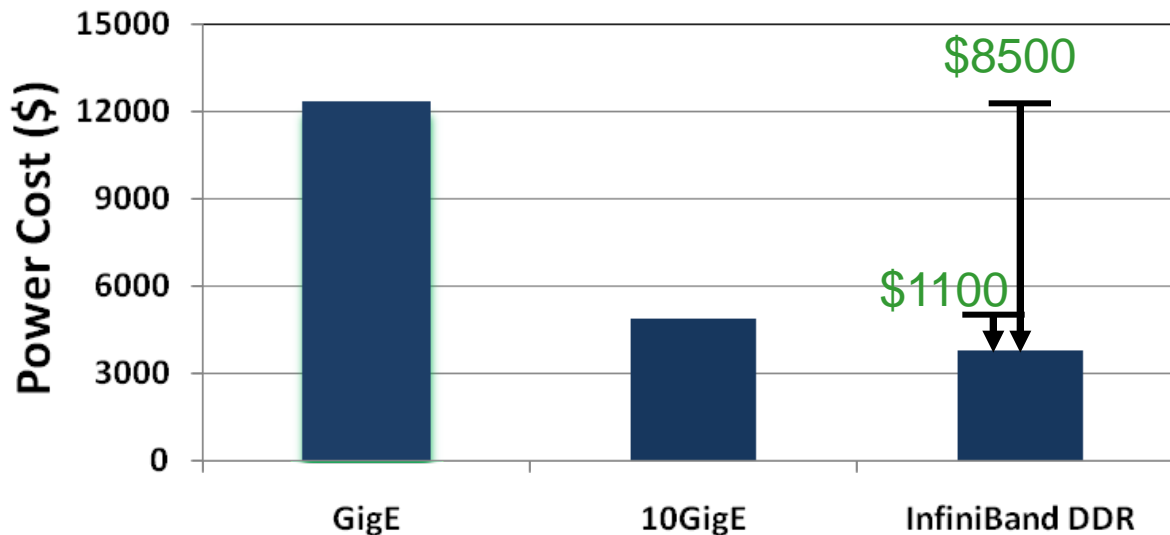


Higher is better

8-cores per node

- **Dell economical integration of AMD CPUs and Mellanox InfiniBand**
 - To achieve same number of BQCD jobs over GigE
 - InfiniBand saves power up to \$8500 versus GigE and \$1100 versus 10GigE
 - Yearly based for 16-node cluster
- **As cluster size increases, more power can be saved**

Power Cost



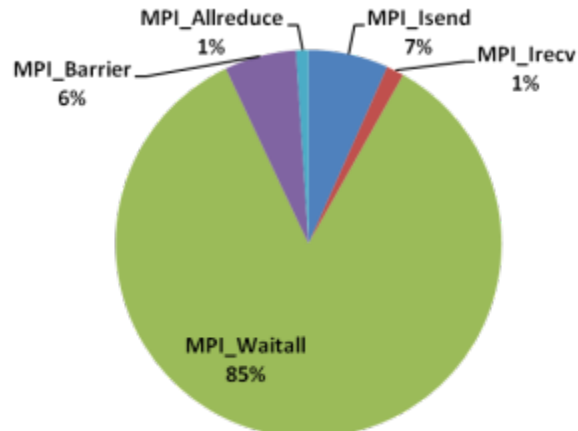
$\$/KWh = KWh * \0.20

For more information - <http://enterprise.amd.com/Downloads/svrpwrusecompletefinal.pdf>

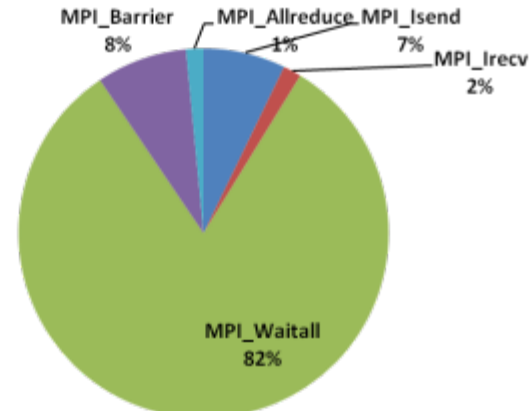
- **Interconnect comparison shows**
 - InfiniBand delivers superior performance in every cluster size versus GigE and 10GigE
 - Performance advantage extends as cluster size increases
- **Open MPI, MVAPICH, and Platform MPI provide similar performance**
- **InfiniBand enables power saving**
 - Up to \$8500/year power savings versus GigE and \$1100 versus 10GigE on 16 node cluster
- **Dell™ PowerEdge™ server blades provides**
 - Linear scalability (maximum scalability) and balanced system
 - By integrating InfiniBand interconnect and AMD processors
 - Maximum return on investment through efficiency and utilization

- **Mostly used MPI functions**

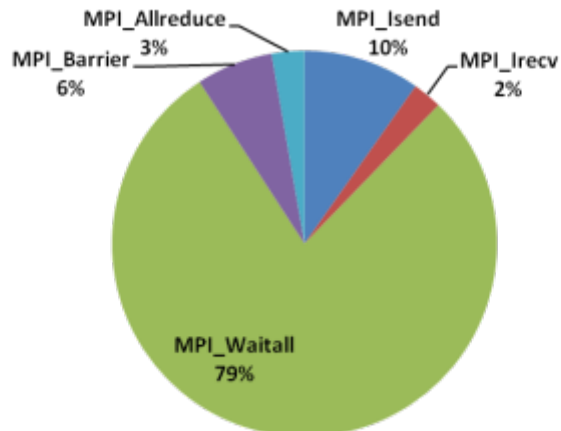
- MPI_Waitall, MPI_Barrier, and MPI_Allreduce create large overhead
- MPI_Barrier and MPI_Allreduce overhead increases faster than as cluster size scales



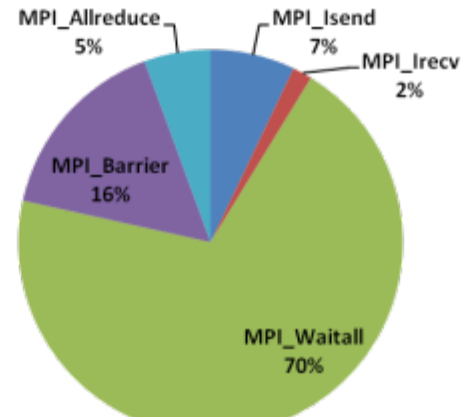
16P



32P

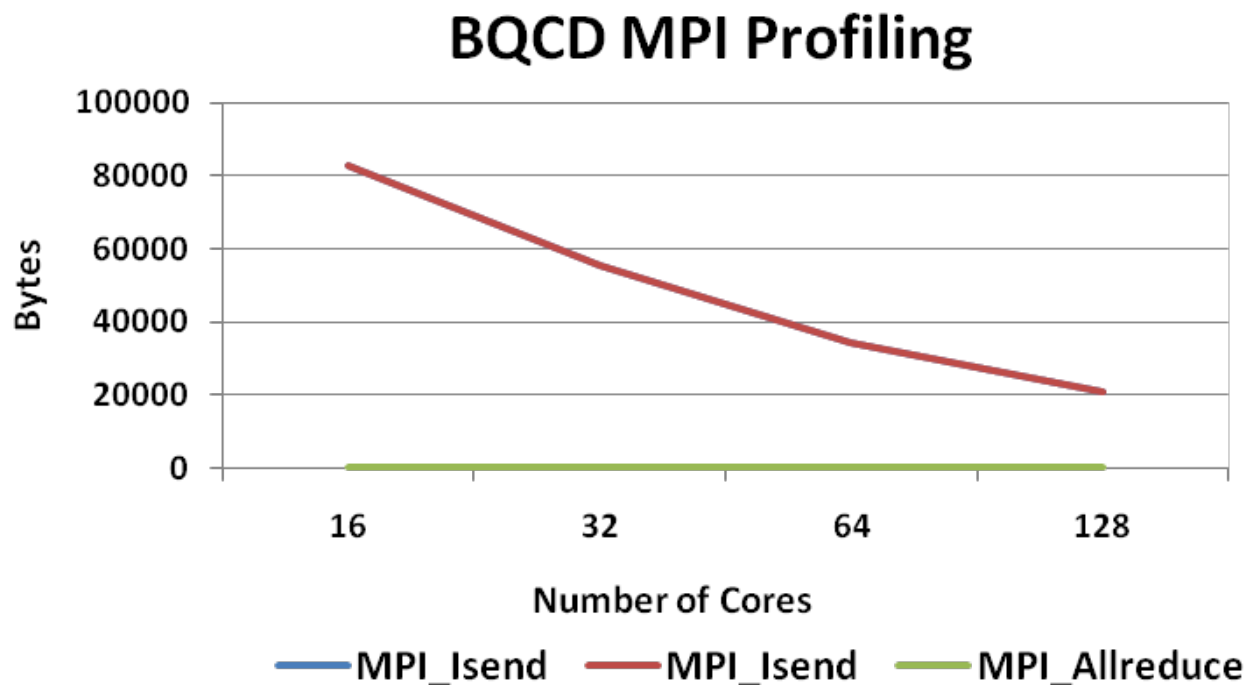


64P



128P

- **Both small and large messages are used**
 - Most point-t-point messages are large size
 - Collective (MPI_Allreduce) messages are small size



- **BQCD was profiled to identify its communication patterns**
 - MPI point-t-point creates the big communication overhead
 - Collective overhead increases with cluster size
- **Interconnects effect to BQCD performance**
 - Both latency and bandwidth are critical to application performance
- **Balanced system – CPU, memory, Interconnect that match each other capabilities, is essential for providing application efficiency**

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