HPC ADVISORY COUNCIL

BQCD Performance Benchmark and Profiling

July 2010

intel

DELL

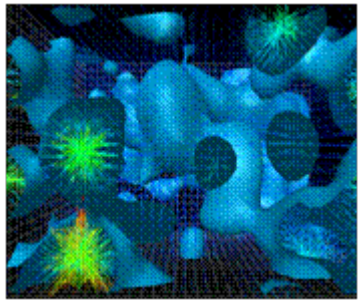
Mellanox TECHNOLOGIES

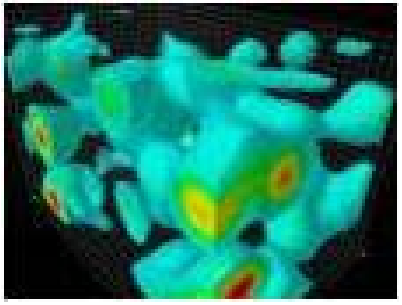
Note

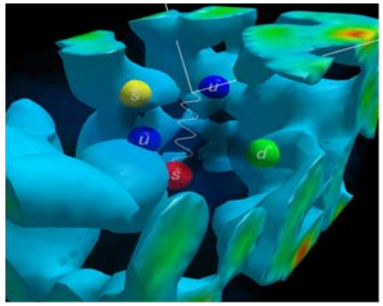
- **The following research was performed under the HPC Advisory Council activities**
  - Participating vendors: Intel, Dell, Mellanox
  - Compute resource - HPC Advisory Council Cluster Center
- **For more info please refer to**
  - www.mellanox.com, www.dell.com/hpc, www.intel.com, http://www.deisa.eu/science/benchmarking/codes/bqcd

BQCD

- **BQCD (Berlin Quantum ChromoDynamics program)**
  - A hybrid Monte-Carlo code
  - The kernel of the program is a standard conjugate gradient solver
  - Simulates Quantum Chromodynamics with dynamical standard Wilson fermions
- **Developed by Dr. Hinnerk Stueben from Konrad-Zuse-Zentrum fuer. Infomationstechnik Berlin**
  - Open source software







Objectives

- **The presented research was done to provide best practices**
  - BQCD performance benchmarking
    - MPI Library performance comparisons
    - Interconnect performance comparisons
  - Understanding BQCD 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 Clusters Configuration

- **Dell™ PowerEdge™ M610 14-node cluster**
  - Six-Core Intel X5670 @ 2.93 GHz CPUs
  - OS: CentOS5U4, OFED 1.5.1 InfiniBand SW stack
- **Intel Cluster Ready certified clusters**
- **Mellanox ConnectX2 QDR InfiniBand mezzanine card**
- **Mellanox M3601Q 32-Port Quad Data Rate (QDR-40Gb) InfiniBand Switch**
- **Memory: 24GB memory per node**
- **MPI: OpenMPI-1.4.1, MVAPICH2-1.4, Platform MPI 7.1, Intel MPI 4.0**
- **Application: BQCD**
- **Benchmark Workload**
  - Lattice size: 48x24x24x48

About Intel® Cluster Ready

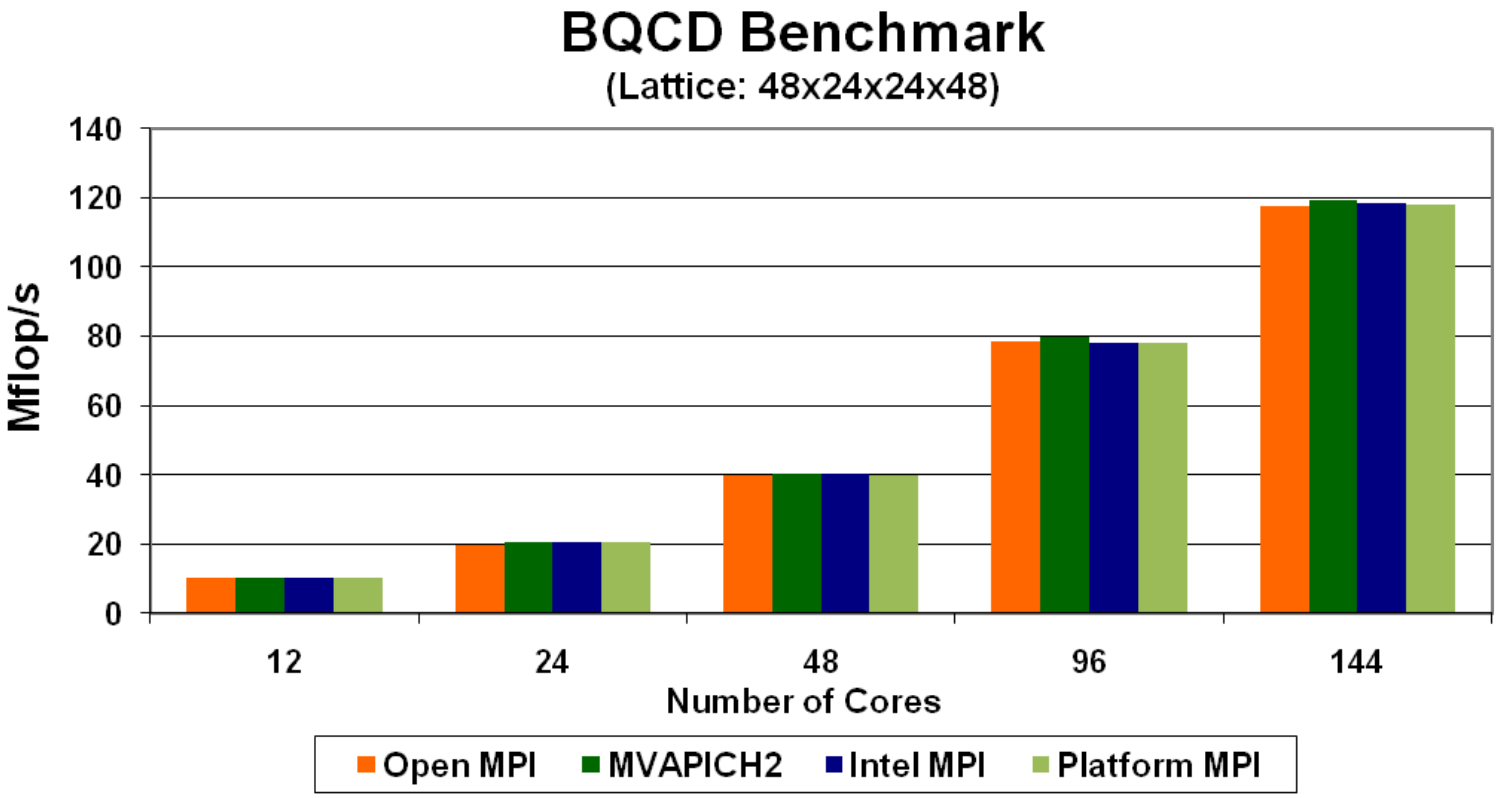
- **Intel® Cluster Ready systems make it practical to use a cluster to increase your simulation and modeling productivity**
  - Simplifies selection, deployment, and operation of a cluster
- **A single architecture platform supported by many OEMs, ISVs, cluster provisioning vendors, and interconnect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

BQCD Benchmark Results

- **All tested MPIs provide similar performance**

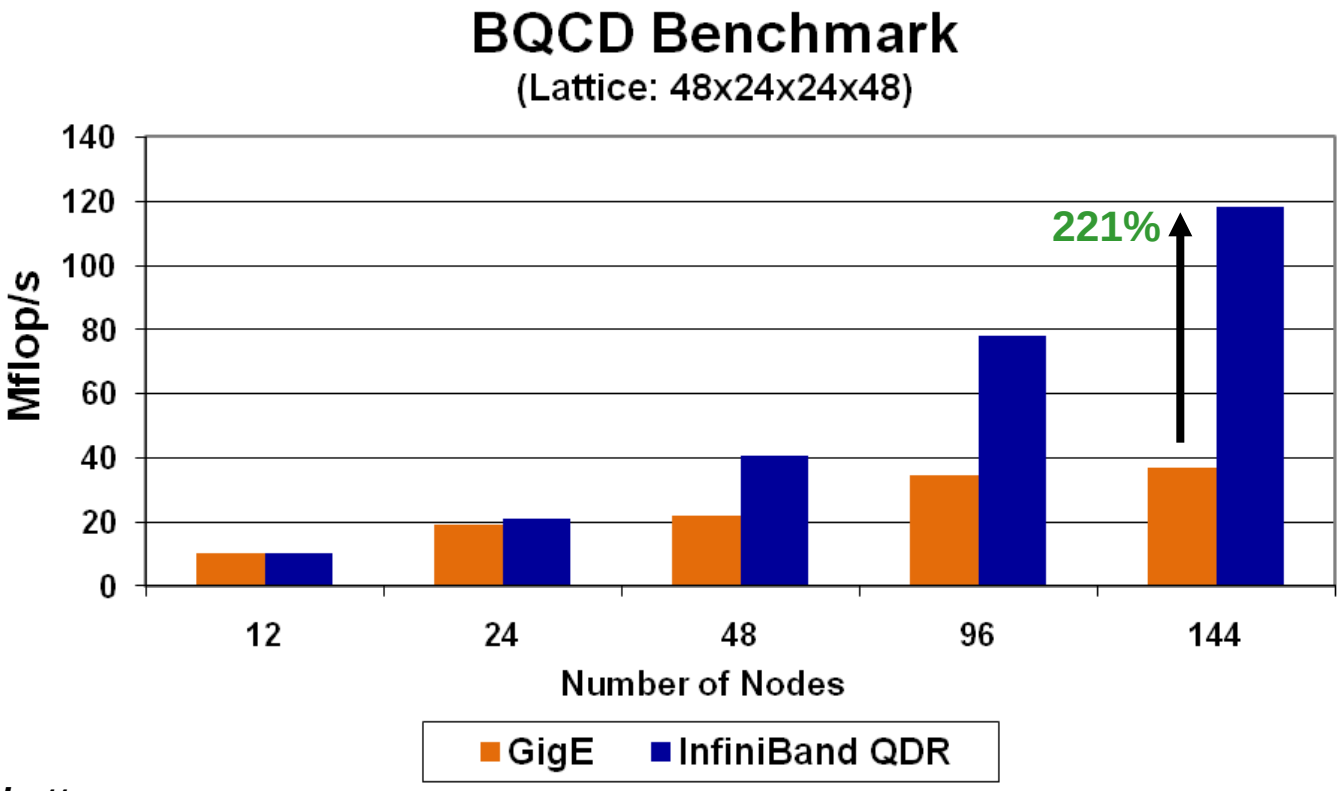


Higher is better

12-cores per node

BQCD Benchmark Results

- **InfiniBand enables better application performance and scalability**
  - Up to 221% higher performance than GigE
  - GigE stops scaling after 8 nodes
- **Application performance over InfiniBand scales as cluster size increases**

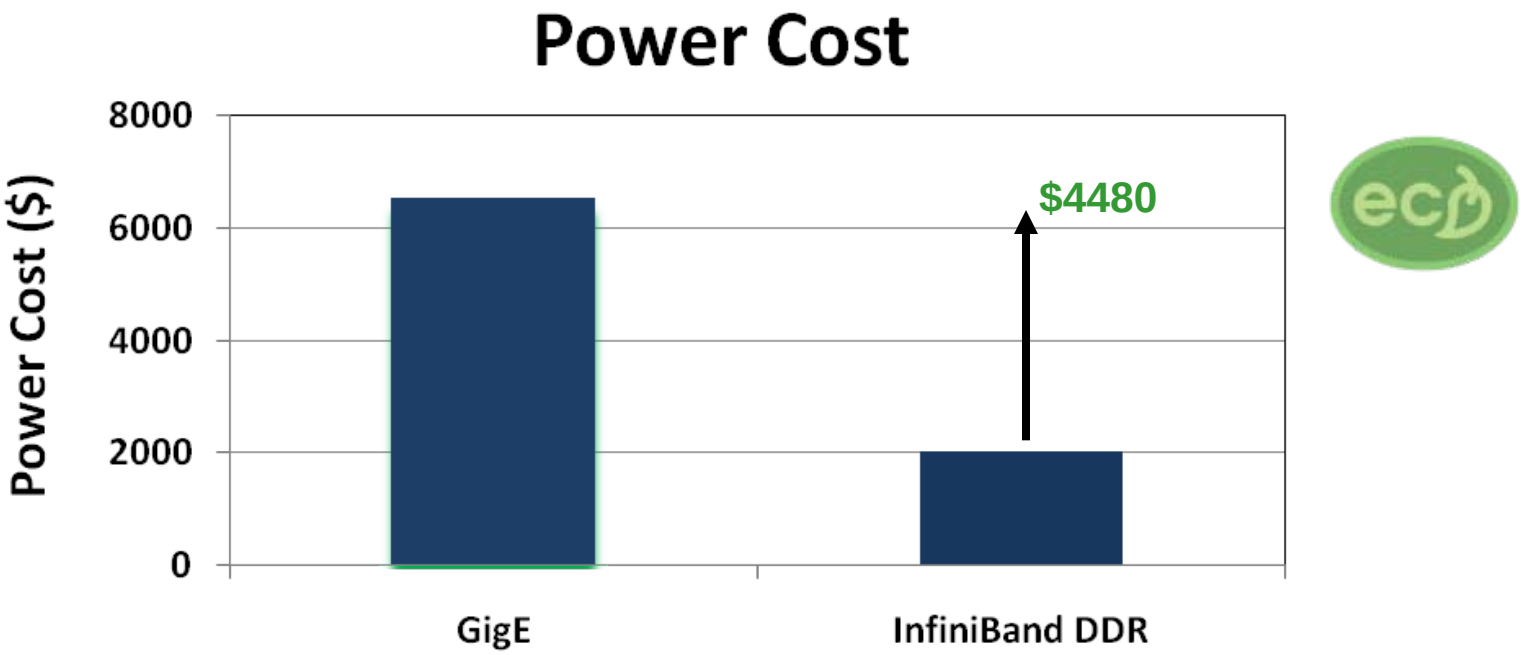


Higher is better

12-cores per node

Power Cost Savings with Different Interconnect

- **InfiniBand saves up to $4480 power compared to GigE**
  - To finish the same number of BQCD jobs
  - Yearly based for 14-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

BQCD Benchmark Results Summary

- **Balanced system – CPU, memory, Interconnect that match each other capabilities - is essential for providing application efficiency**

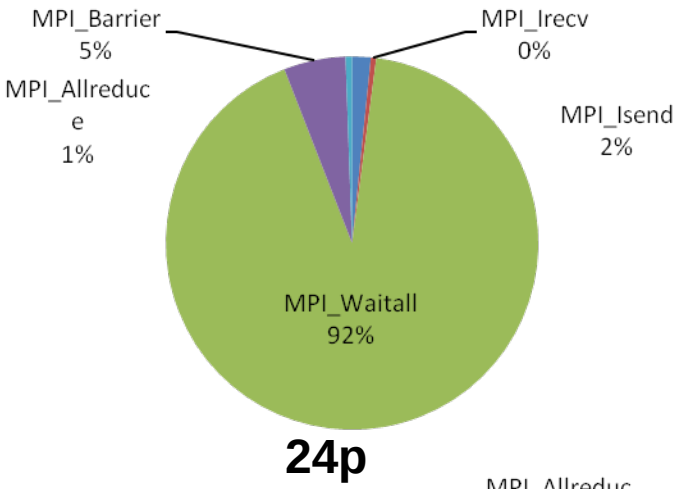
- **Performance Optimization**
  - All tested MPIs enable similar performance

- **Interconnect Characterization**
  - InfiniBand delivers superior performance across all system sizes
  - GigE scalability is limited beyond 8 nodes

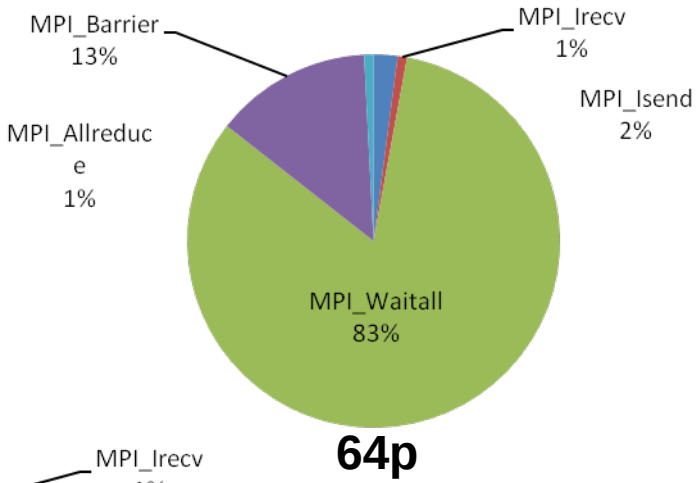
- **Power Analysis**
  - System architecture can yield nearly $5K annually in power savings

BQCD Profiling - % of MPI Time

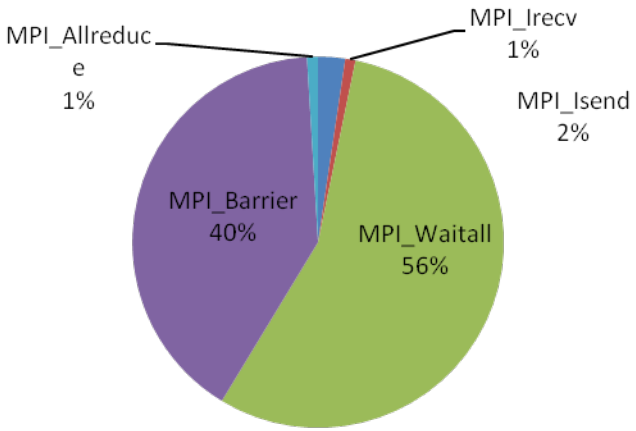
- **MPI P2P and MPI_Barrier consume more than 96% of total MPI time**



24p



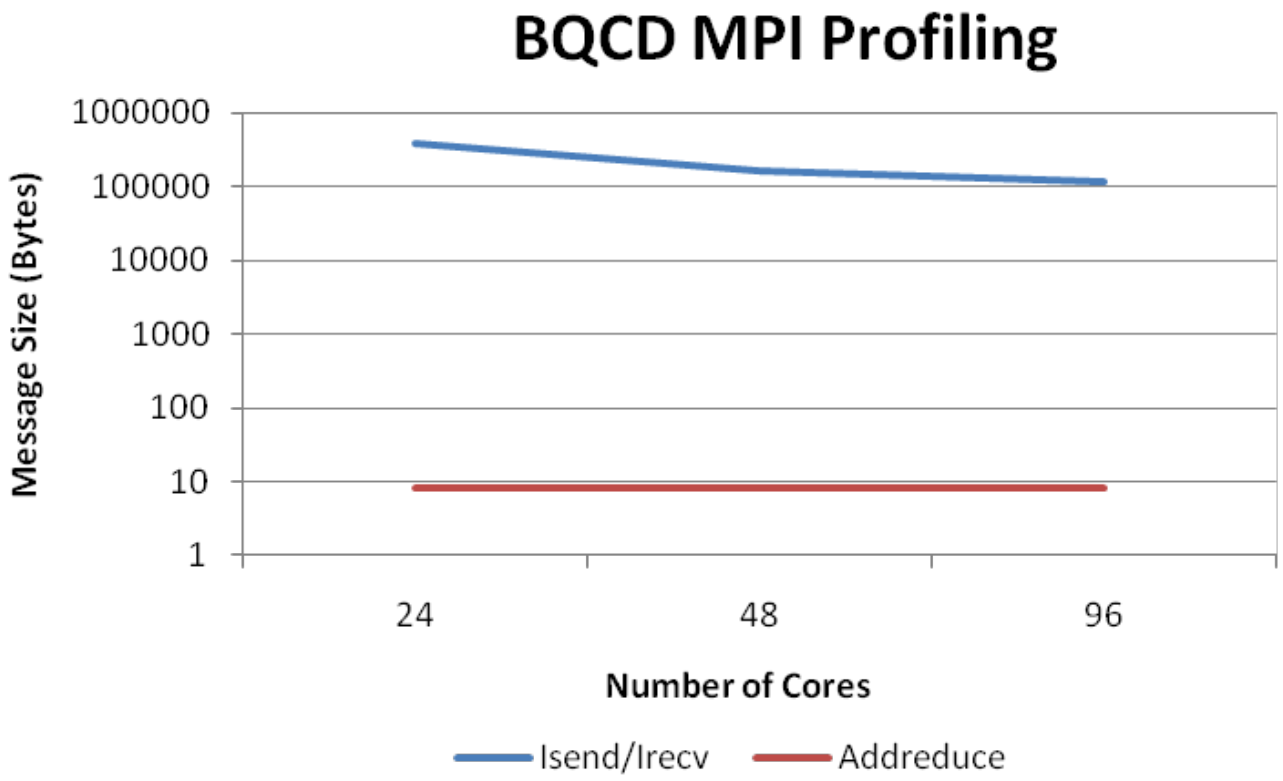
64p



96p

BQCD Profiling – Message Size

- **MPI P2P messages are large message size**
  - Message size becomes smaller as node number increases
- **MPI collective messages are small size**



BQCD Profiling Summary

- **BQCD was profiled to identify its communication patterns**
- **MPI P2P and MPI_Barrier functions dominate total MPI communication time**
  - MPI collective overhead grows faster relative to MPI P2P function
  - Total number of messages increases with cluster size
- **Interconnects effect to BQCD performance**
  - Both small and large messages are used by BQCD
  - Interconnect latency and bandwidth are critical to application performance
- **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
  - 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 “Westmere” 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

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