HPC ADVISORY COUNCIL

OpenAtom Performance Benchmark and Profiling

Feb 2010

AMD The future is fusion

DELL

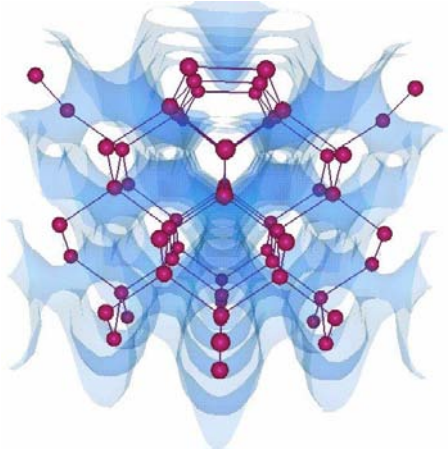
Mellanox TECHNOLOGIES

Note

- **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://charm.cs.uiuc.edu/OpenAtom

OpenAtom

- **OpenAtom a massively parallel quantum chemistry application**
  - Implements the Car-Parrinello ab-initio Molecular Dynamics method to solve problems in
    - Material science
    - Chemistry
    - Solid-state physics
    - Biophysics



- **Using the Charm++ parallel programming framework**
- **OpenAtom is an open source software developed by Parallel Programming Lab at UIUC**

Objectives

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

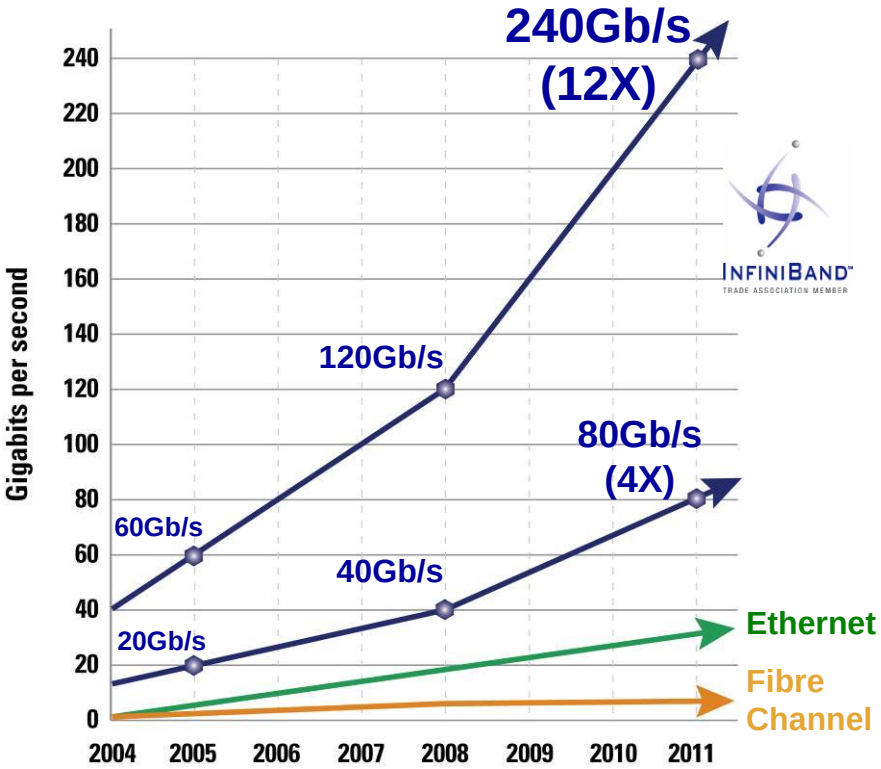
Test Cluster Configuration

- **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 1.4.1 InfiniBand SW stack**
- **MPI: OpenMPI-1.4, MVAPICH2-1.4, charm-6.1.3**
- **Application: OpenAtom 1.0**
- **Benchmark Workload**
  - Water 32M 70Ry (32 water molecules and an energy cutoff of 70 Ry)

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**

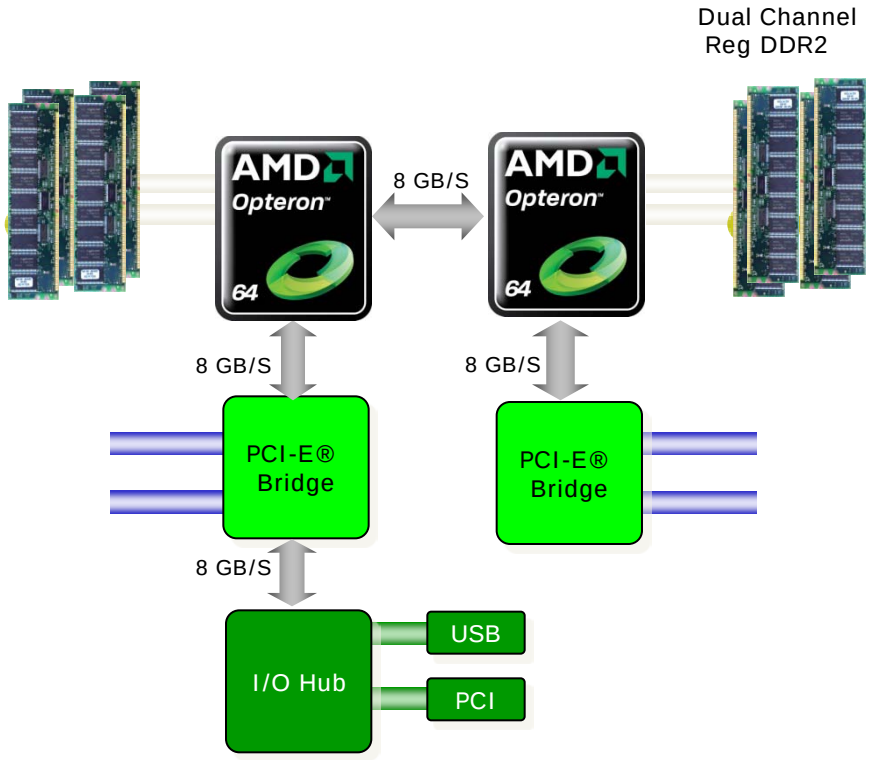
The InfiniBand Performance Gap is Increasing



InfiniBand Delivers the Lowest Latency

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/clk 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

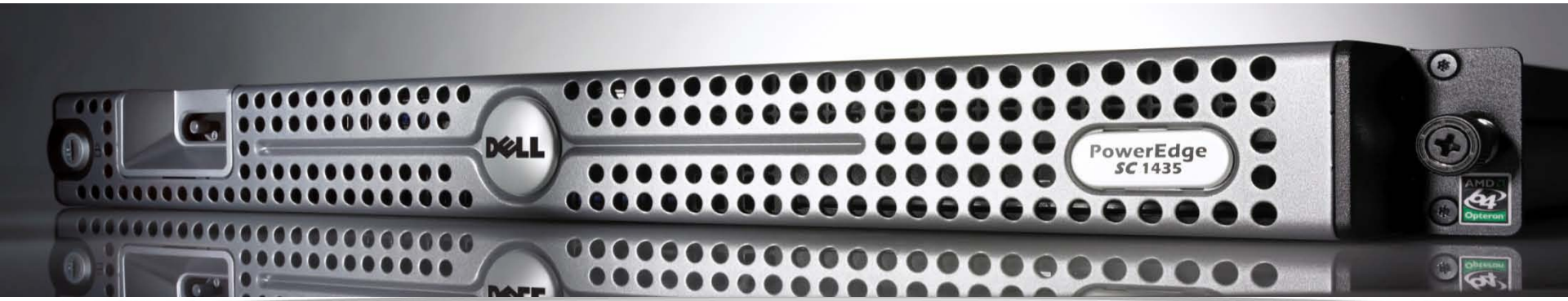


Dell PowerEdge Servers helping Simplify IT

- **System Structure and Sizing Guidelines**
  - 24-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

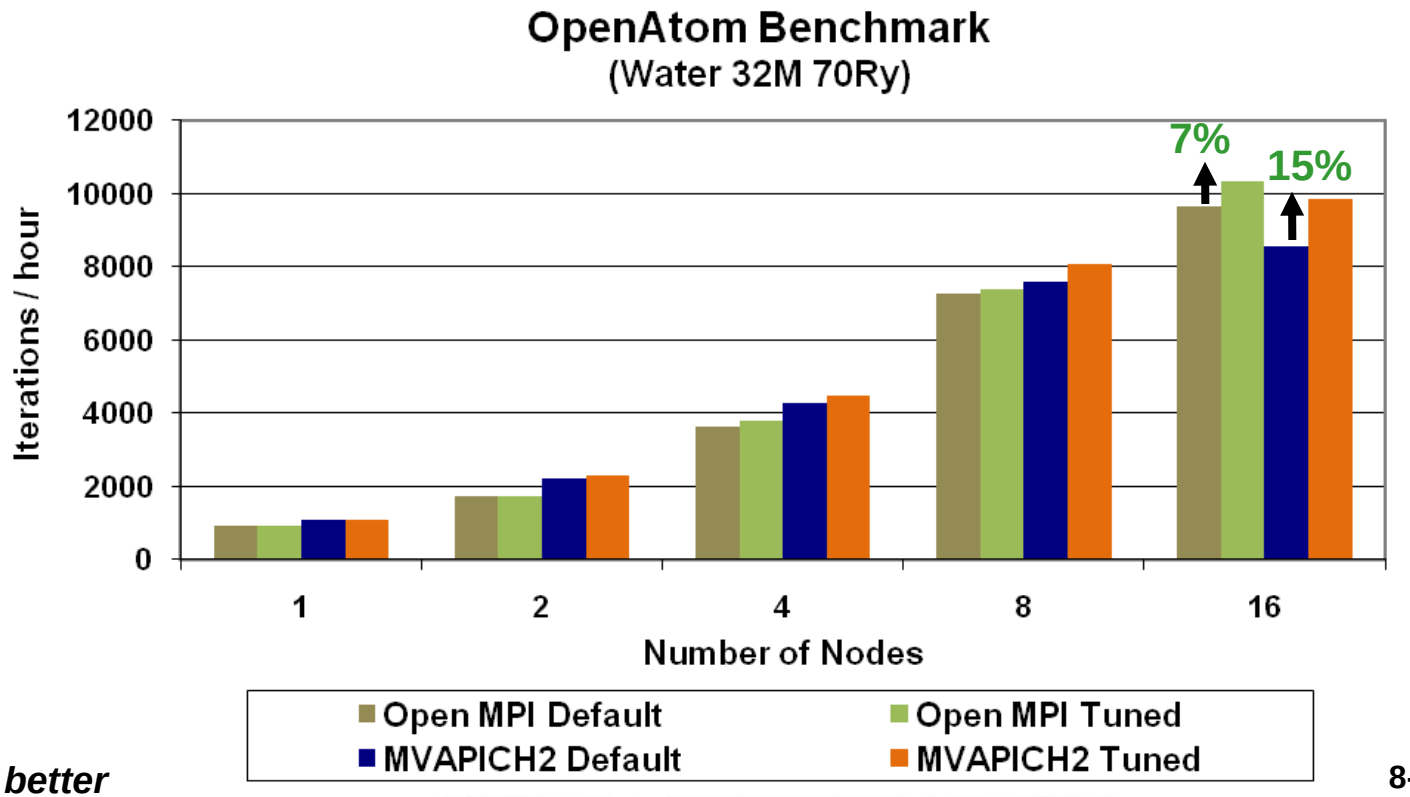
Dell PowerEdge™ Server Advantage

- **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**



OpenAtom Benchmark Results - MPI

- **Customized MPI parameters provide better performance**
  - **Up to 7% higher performance with Open MPI**
    - --mca mpi_affinity_alone 1 --mca btl_openib_eager_limit 65536 --mca btl_openib_eager_rdma_num 64 --mca btl_openib_eager_rdma_threshold 8
  - **Up to 15% higher performance with MVAPICH2**
    - MV2_USE_RDMA_FAST_PATH=0 MV2_USE_RDMA_ONE_SIDED=0

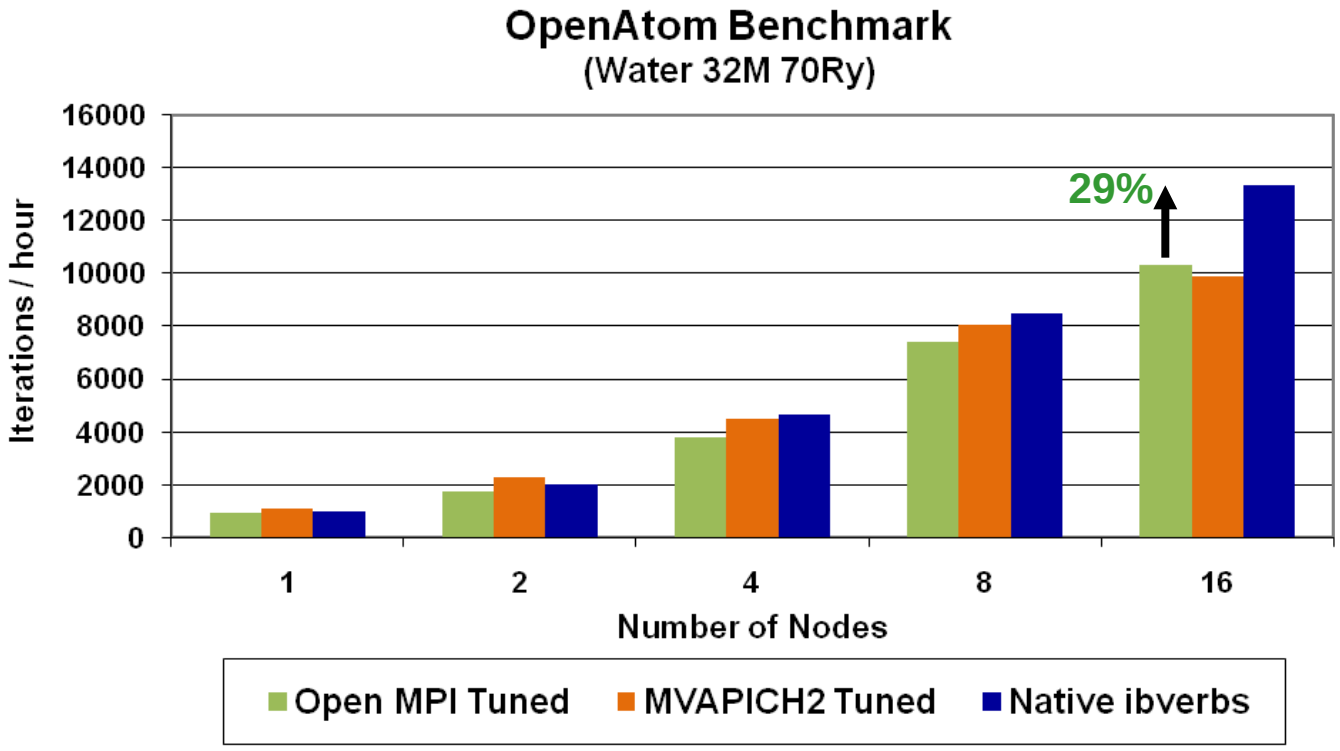


Higher is better

8-cores per node

OpenAtom Benchmark Results - ibverbs

- **Native ibverbs enables higher performance and better scalability**
  - Up to 29% higher performance versus tuned Open MPI
  - Performance advantage increases as cluster size scales

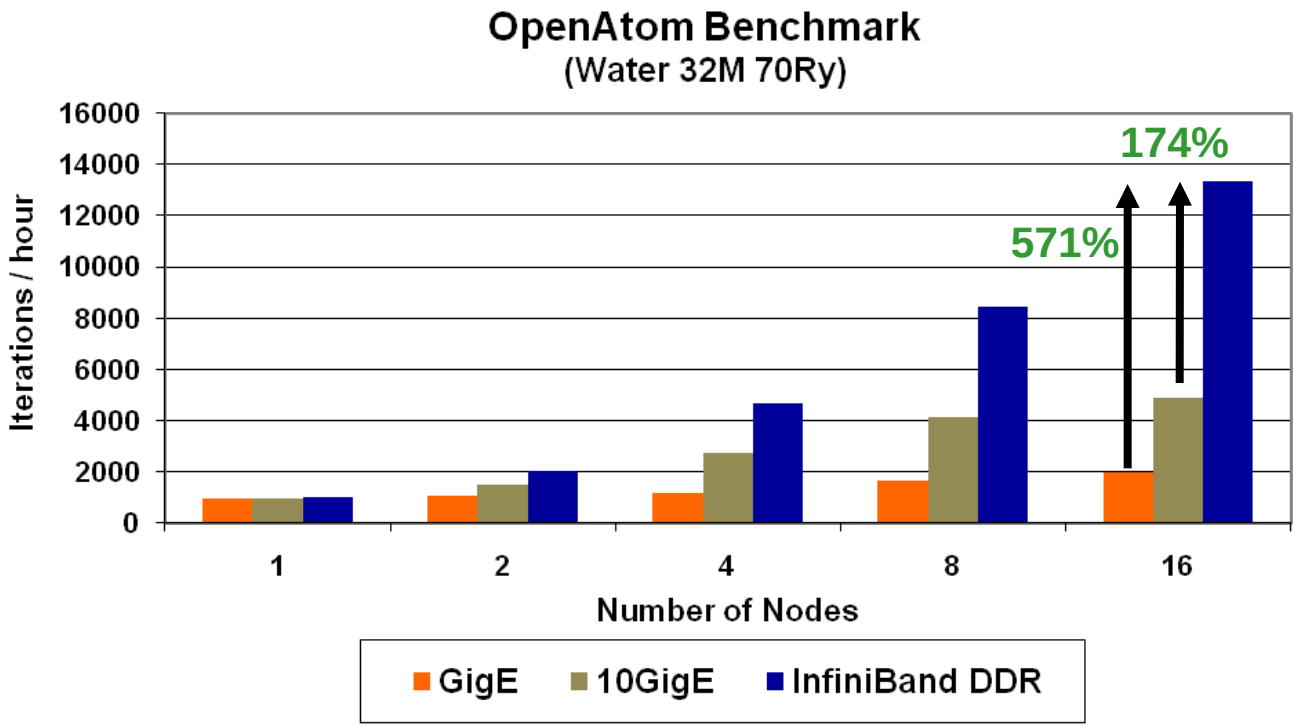


Higher is better

8-cores per node

OpenAtom Benchmark Results - Interconnect

- **InfiniBand enables higher application performance**
  - Up to 174% higher performance than 10GigE and 571% than GigE
- **Application performance over InfiniBand scales as cluster size increases**

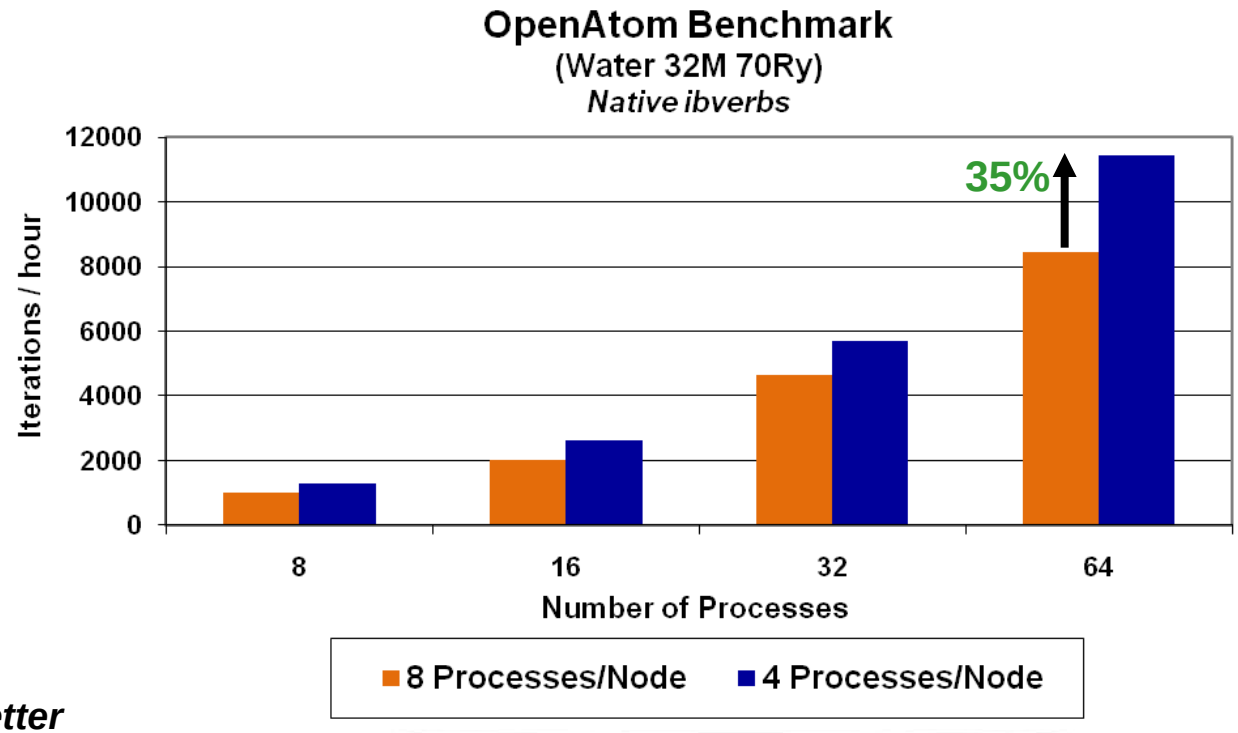


Higher is better

8-cores per node

OpenAtom Benchmark Results - Job Placement

- **Test Scenarios**
  - Scenario 1: Running job over all CPU cores of each node
  - Scenario 2: Running job over half of cores of each CPU
- **Scenario 2 provides better performance**
  - With same number of processes
  - Up to 35% higher performance than scenario 1
- **Given fixed number of nodes, fully utilizing all CPU cores provides highest performance**

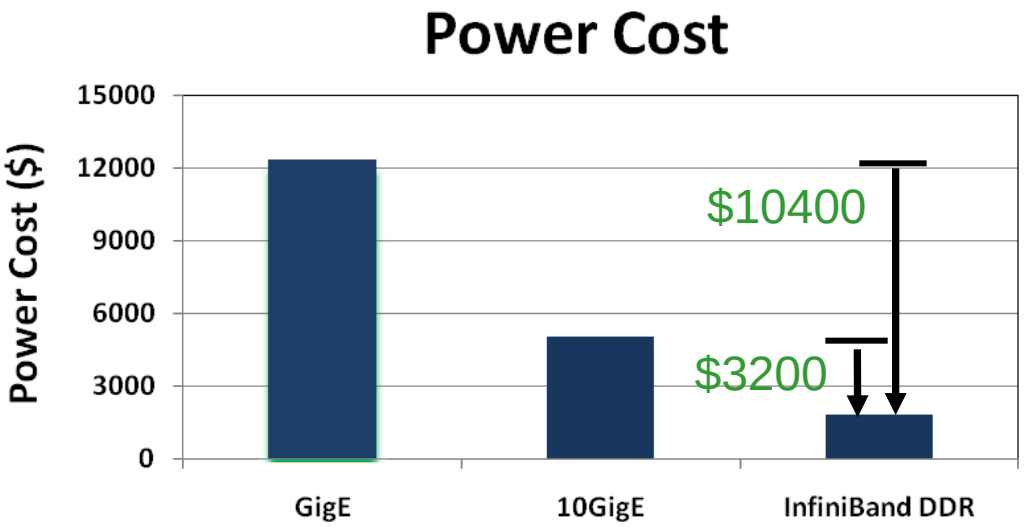


Higher is better

8-cores per node

Power Cost Savings with Different Interconnect

- **Dell economical integration of AMD CPUs and Mellanox InfiniBand**
  - To achieve same number of OpenAtom jobs over GigE
  - InfiniBand saves power up to $3200 versus 10GigE and $10400 versus 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

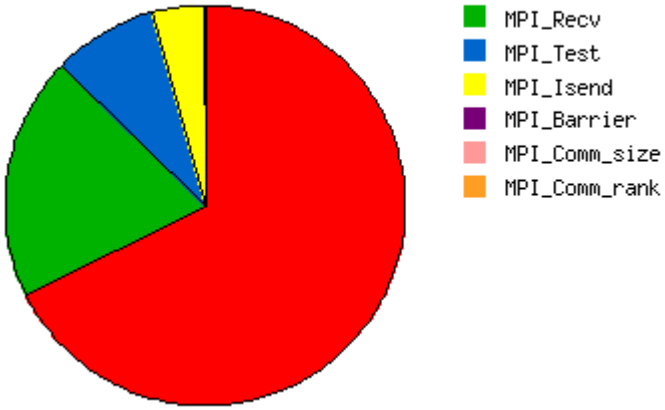
OpenAtom Benchmark Results Summary

- **Tuned MPI parameters provides better performance**
  - Both Open MPI and MVAPICH2 gain extra performance by adjusting parameters
- **Native ibverbs implementation enables higher OpenAtom performance**
  - Up to 29% faster than running with MPI
- **Interconnect comparison shows**
  - InfiniBand delivers superior performance in every cluster size versus GigE and 10GigE
  - Performance advantage extends as cluster size increases
- **Reasonable job placement delivers higher productivity**
- **InfiniBand enables power saving**
  - Up to $10400/year power savings versus GigE and $3200 versus 10GigE
- **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

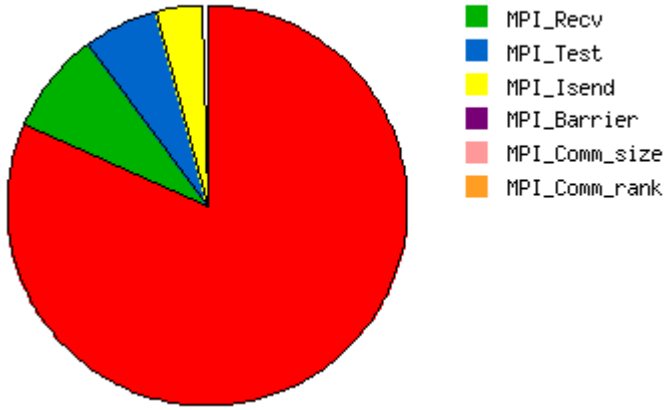
OpenAtom MPI Profiling – MPI Functions

- **Mostly used MPI functions**
  - MPI_Iprobe, MPI_Recv, MPI_Test, and MPI_Isend
  - Point-to-point messages create biggest overhead

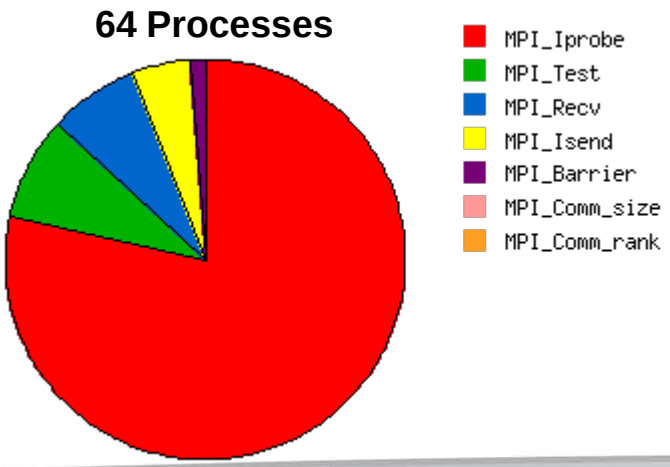
16 Processes



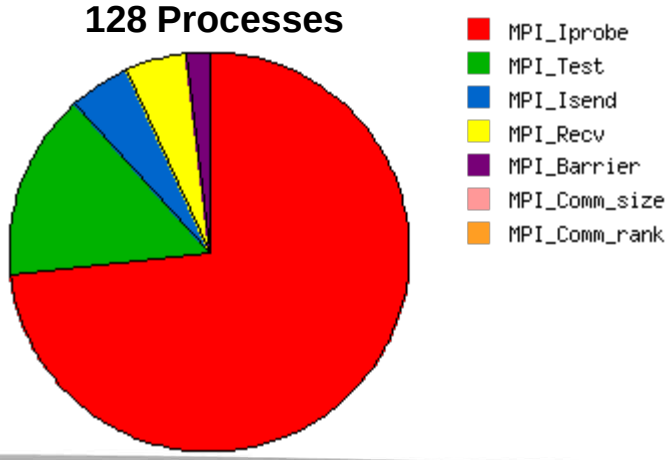
32 Processes



64 Processes

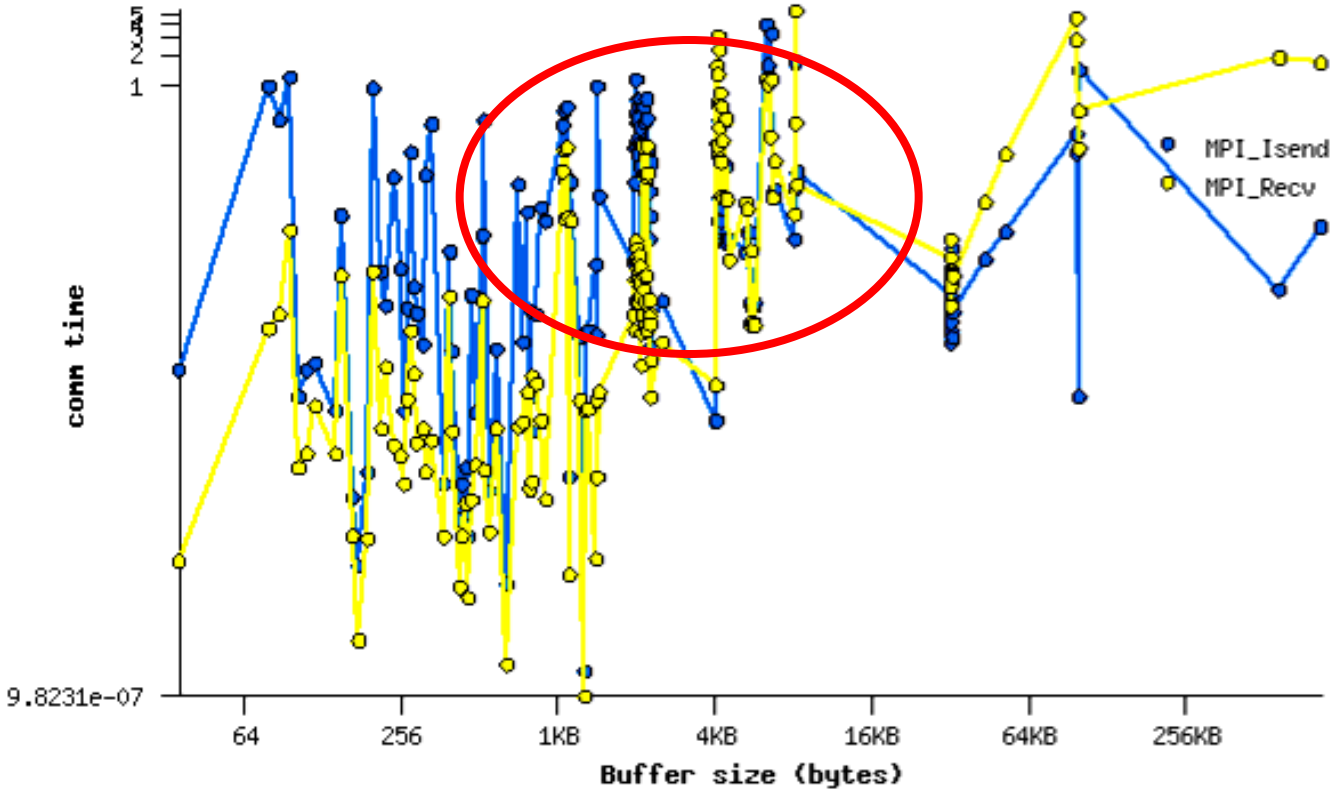


128 Processes



OpenAtom MPI Profiling – MPI Timing

- **Majority messages are between 1KB-16KB**



128 Processes

OpenAtom Profiling Summary

- **OpenAtom was profiled to identify its communication patterns**
  - MPI point-to-point functions create the biggest communication overhead
  - Number of messages increases with cluster size
- **Interconnects effect to OpenAtom performance**
  - Most messages are between 1KB-16KB
  - Both bandwidth and latency is critical to OpenAtom performance
- **Balanced system – CPU, memory, Interconnect that match each other capabilities, is essential for providing application efficiency**

Thank You

HPC Advisory Council