HPC ADVISORY COUNCIL

ABySS Performance Benchmark and Profiling

May 2010

AMD
The future is fusion

DELL

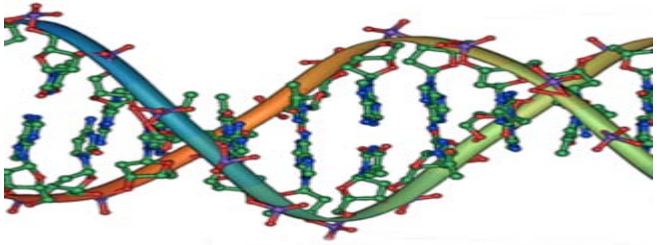
Mellanox
TECHNOLOGIES

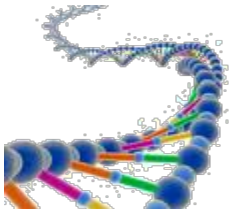
Note

- **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://www.bcgsc.ca/platform/bioinfo/software/abyss

ABySS

- **ABySS is a de novo, parallel, paired-end sequence assembler designed for short reads**
  - Capable of assembling larger genomes
  - Implemented using MPI
- **ABySS was developed at Canada's Michael Smith Genome Sciences Centre**





Objectives

- **The presented research was done to provide best practices**
  - ABySS performance benchmarking
    - Performance tuning with different communication libraries and compilers
    - Interconnect performance comparisons
  - Understanding ABySS 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 1.5 InfiniBand SW stack**
- **Compiler: GCC 4.1.2, Open64-4.2.3.1**
- **Filesystem: Lustre-1.10.0.36**
- **MPI: OpenMPI-1.3.3, MVAPICH2-1.4**
- **Application: ABySS 1.1.2**
- **Benchmark Workload**
  - Illumina Transcriptome Sequencing of Drosophila melanogaster embryo

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

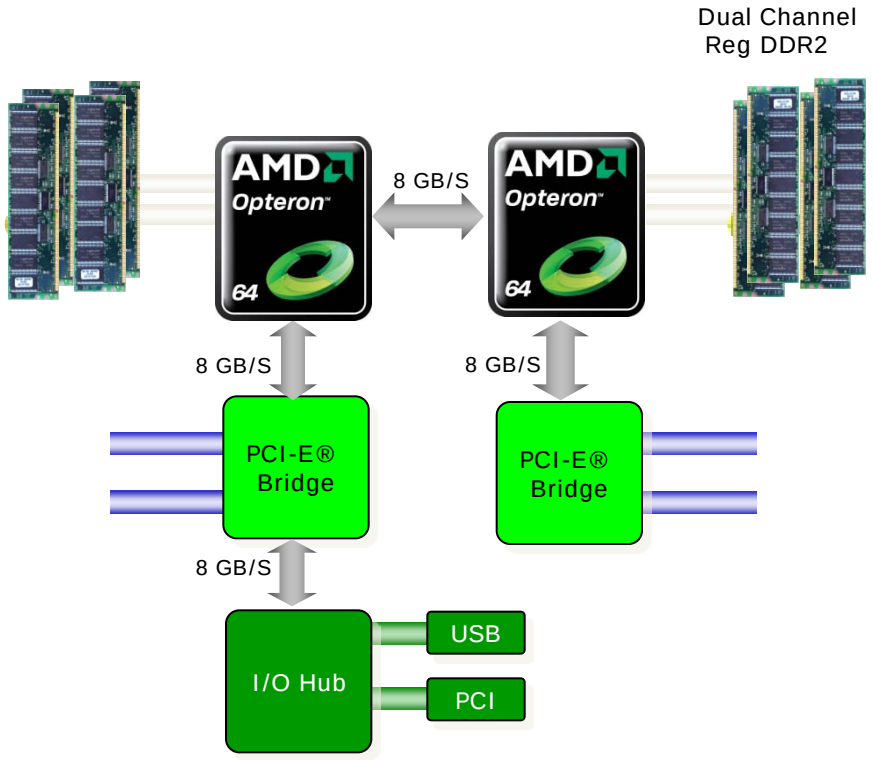
- 40Gb/s node to node, 120G IB switch to switch
- Highest dense switch solutions - 51.8TB in a single switch
- World's lowest switch latency – 100ns 100% load

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/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

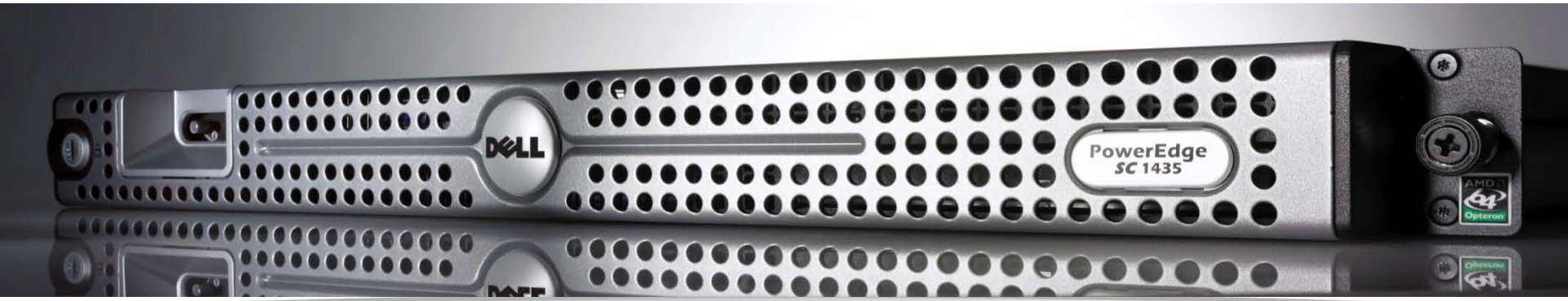


Dell PowerEdge Servers helping Simplify IT

- **System Structure and Sizing Guidelines**
  - 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

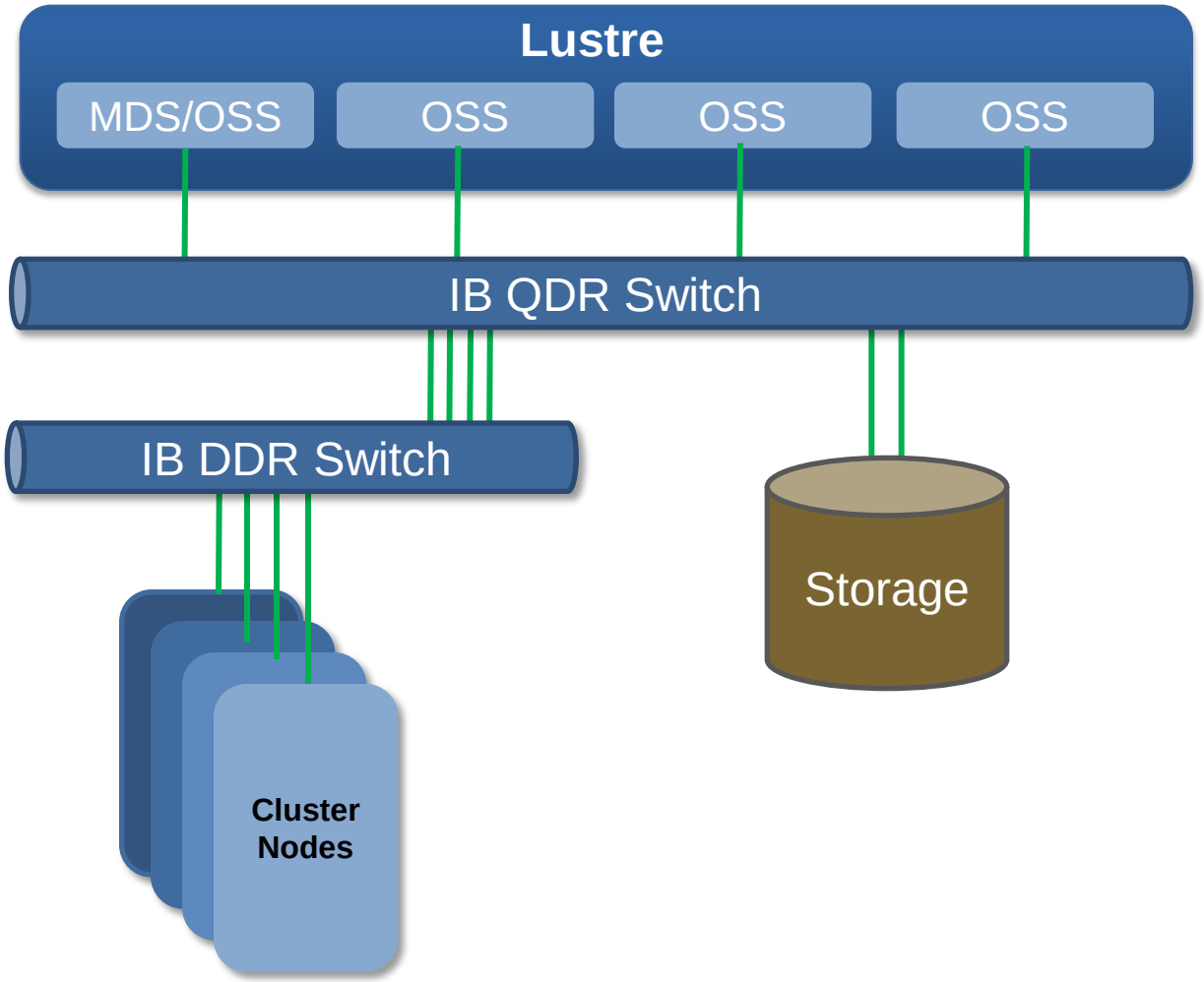
Dell PowerEdge™ Server Advantage

- **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**



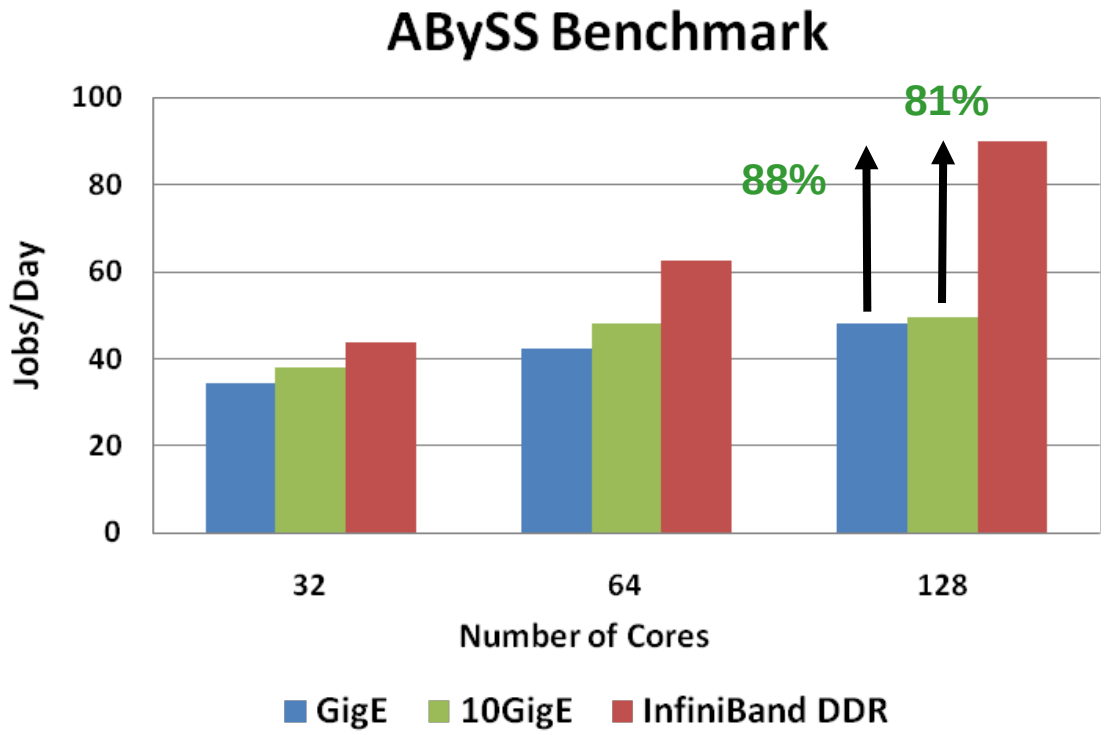
Lustre File System Configuration

- **Lustre Configuration**
  - 1 MDS
  - 4 OSS (Each has 2 OST)
  - InfiniBand based Backend storage
  - All components are connected through InfiniBand interconnect



ABySS Benchmark Results

- **InfiniBand enables higher performance and scalability**
  - Up to 88% higher performance than GigE and 81% higher than 10GigE
  - Both GigE and 10GigE don't scale well beyond 8 nodes

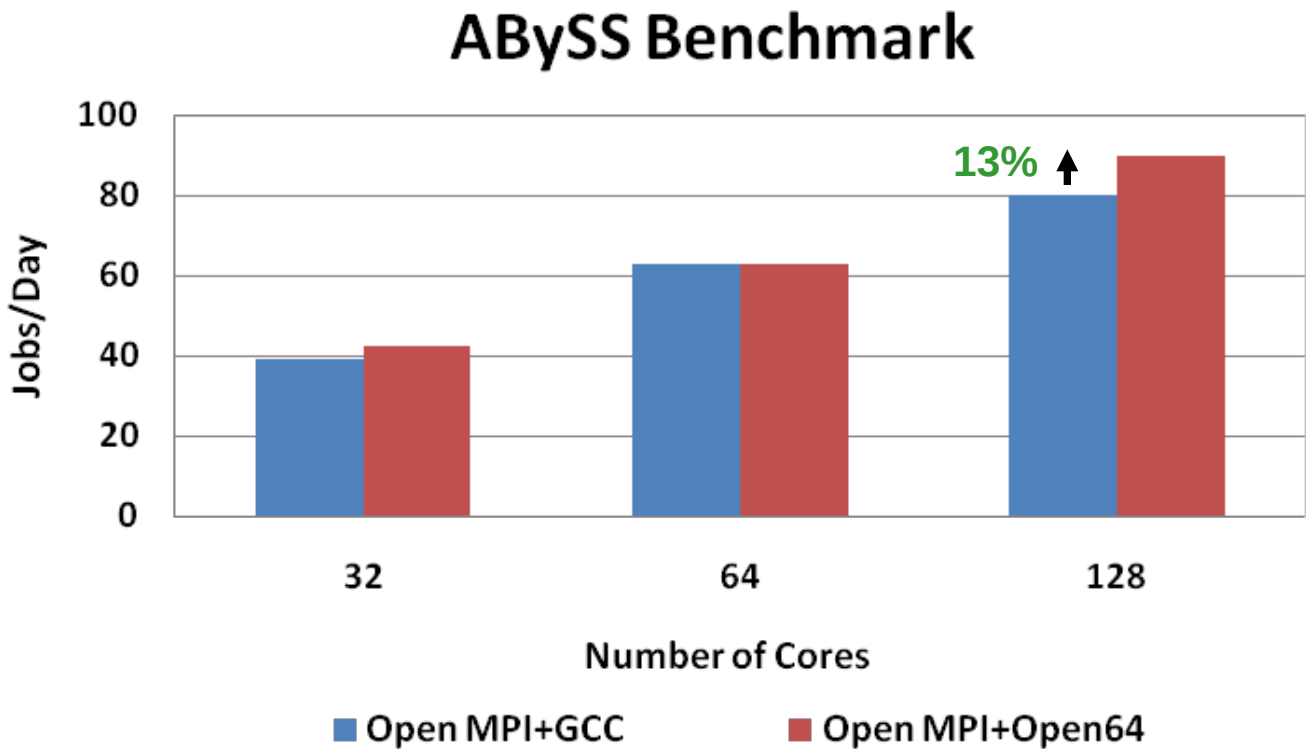


Higher is better

8-cores per node

ABySS Benchmark Results - Compilers

- **Two different compilers were used to compile ABySS**
  - Open64 provides better performance than GCC at 128 cores

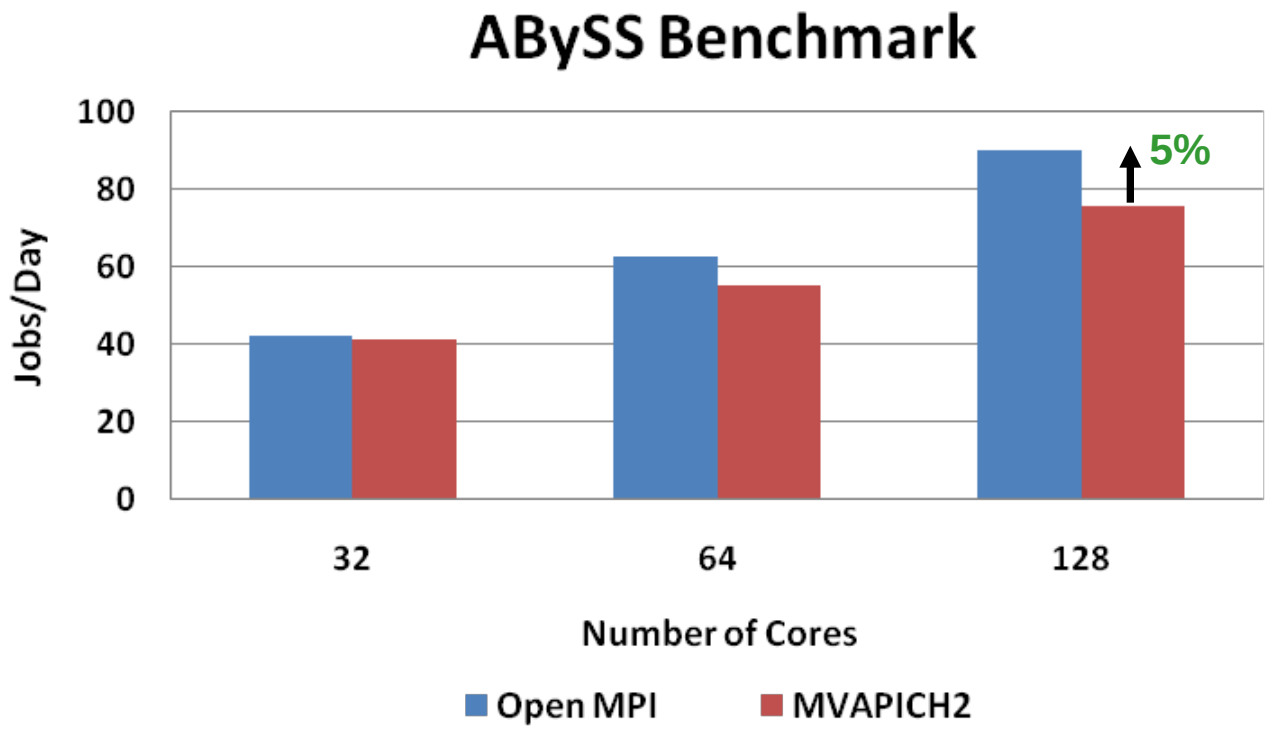


Higher is better

8-cores per node

ABySS Benchmark Results - MPI Libraries

- **Open MPI enables better performance at higher core count**
  - Up to 5% at 128 cores

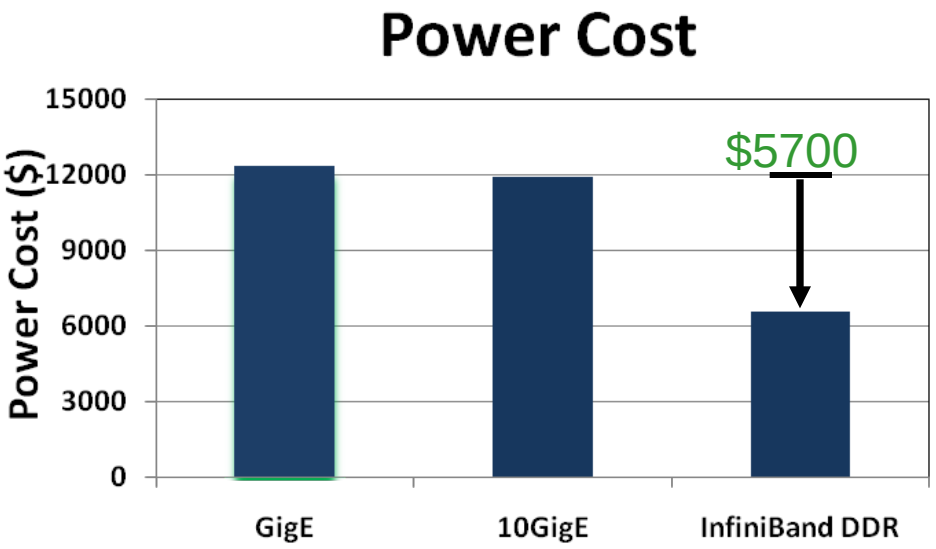


Higher is better

8-cores per node

Power Cost Savings with Different Interconnect

- **Dell economical integration of AMD CPUs and Mellanox InfiniBand**
  - To achieve same number of ABySS jobs over GigE
  - InfiniBand saves power up to $5700 versus GigE and $5300 versus 10GigE
  - Yearly based for 16-node cluster
- **As cluster size increases, more power can be saved**



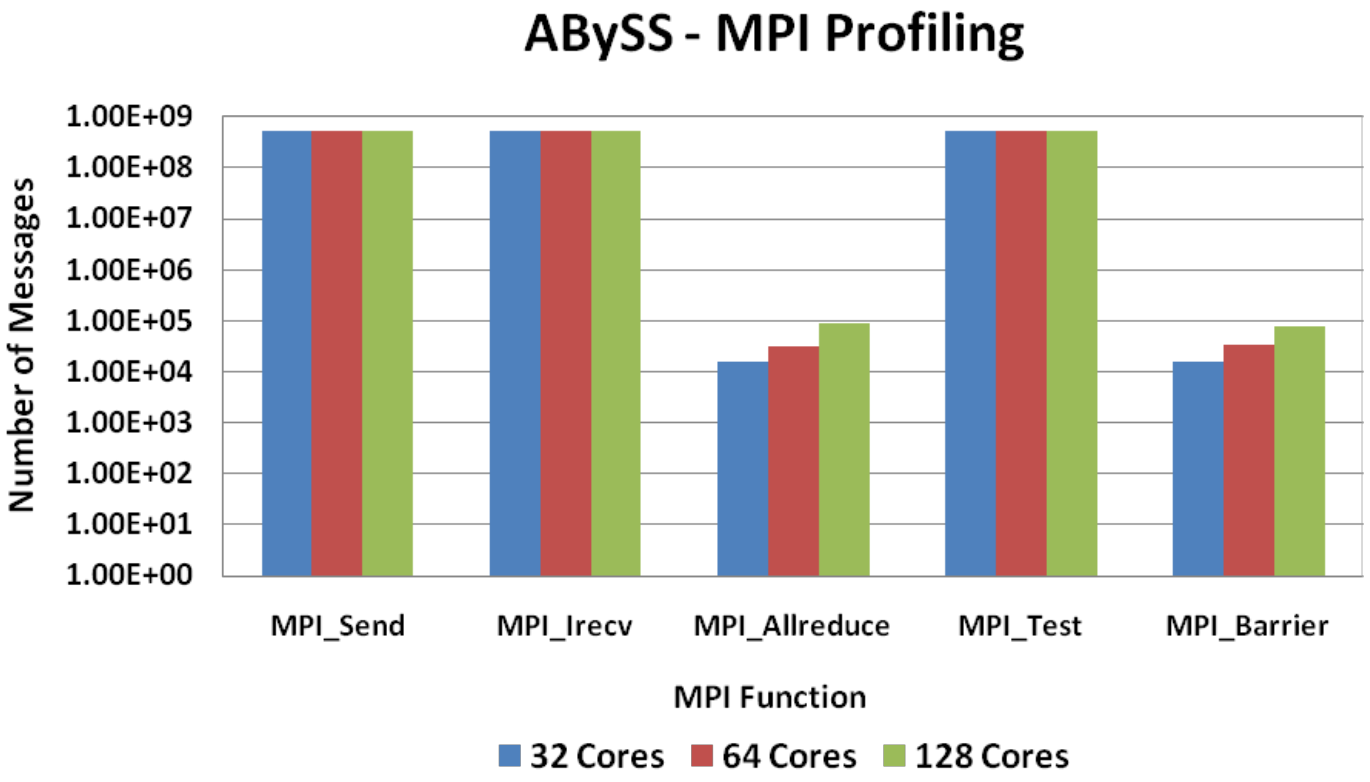
*$/KWh = KWh * $0.20*
For more information - http://enterprise.amd.com/Downloads/svrpwrusecompletefinal.pdf

ABySS Benchmark Summary

- **Interconnect comparison shows**
  - InfiniBand delivers superior performance in every cluster size versus GigE and 10GigE
  - Performance advantage extends as cluster size increases
- **Open64 compiler delivers higher performance**
- **Open MPI provides higher performance**
- **InfiniBand enables power saving**
  - Up to $5700/year power savings versus GigE and $5300 versus 10GigE on16 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

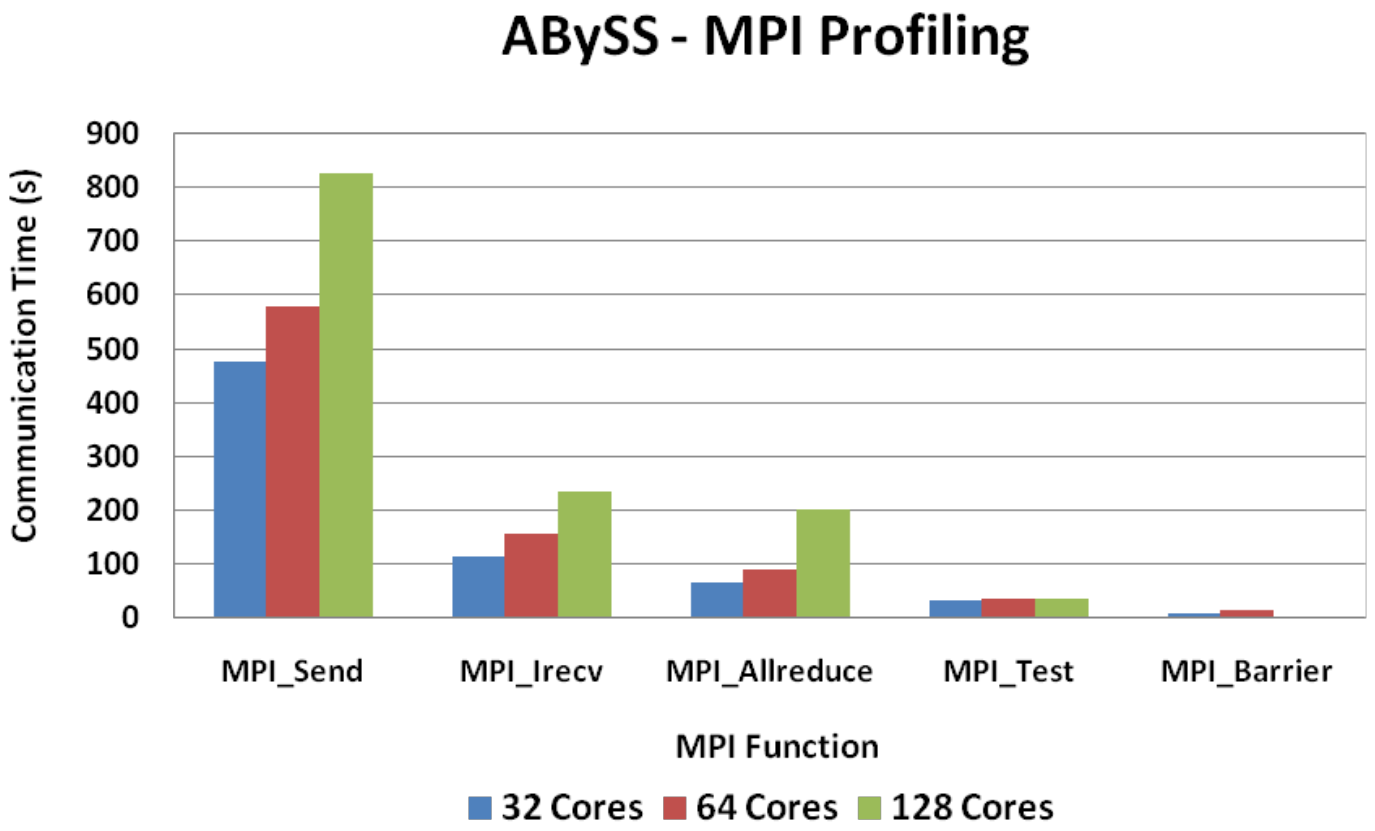
ABySS MPI Profiling – MPI Functions

- **Mostly used MPI functions**
  - MPI_Send/Irecv and MPI_Test are the mostly used MPI functions
  - Number of collectives increases as cluster size scales



ABySS MPI Profiling – Communication Overhead

- Send, Irecv, and Allreduce are three major functions creating big overhead
- Communication overhead increases as cluster size scales



ABySS Profiling Summary

- **ABySS was profiled to identify its communication patterns**
  - MPI point-t-point create the big communication overhead
  - Number of collectives increases with cluster size
- **Interconnects effect to ABySS 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