HPC
ADVISORY COUNCIL

PARATEC Performance Benchmark and Profiling

April 2010

AMD
The future is fusion

DELL

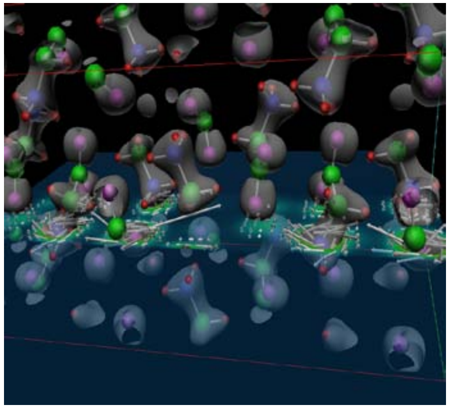
Mellanox
TECHNOLOGIES

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.nersc.gov/projects/paratec/

PARATEC

- PARATEC stands for PARAllel Total Energy Code
- Performs ab-initio quantum-mechanical total energy calculations using pseudopotentials and a plane wave basis set
- Designed to run on massively parallel computing platforms and clusters
- Developed through a joint collaboration between
  - LBNL
  - Université Pierre et Marie CURIE
  - University of Montreal
  - University of Cambridge



Objectives

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

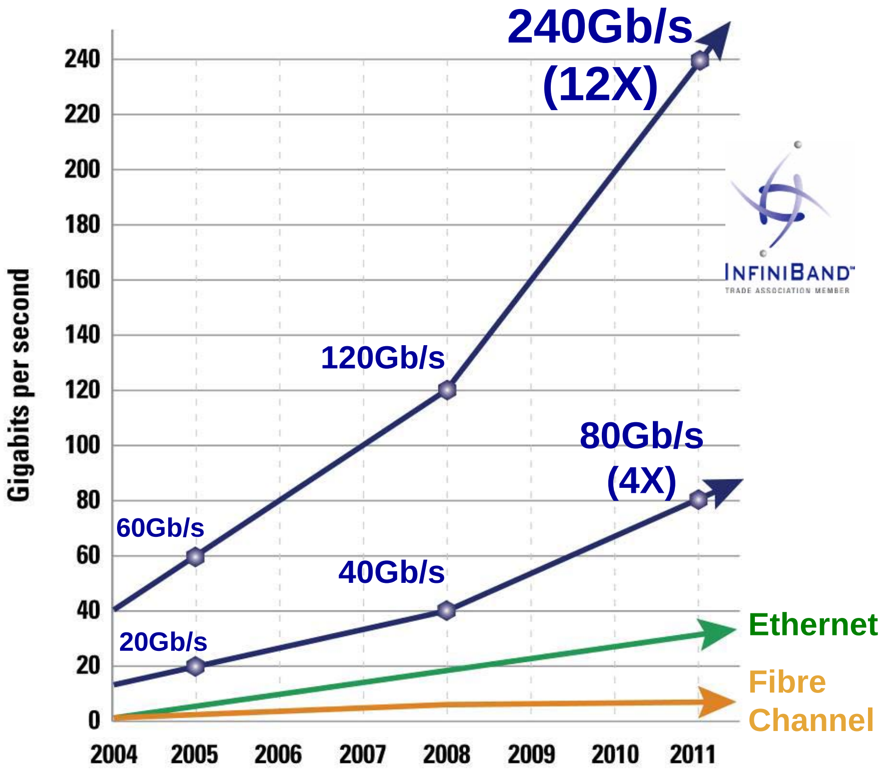
Test Cluster Configuration

- **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 and Math library: Intel compiler 11.1, Intel MKL 11.1**
- **MPI: OpenMPI-1.3.3, Intel MPI 4.0**
- **Application: PARATEC**
- **Benchmark Workload**
  - Large size
    - Silicon in diamond (343 atoms)

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**

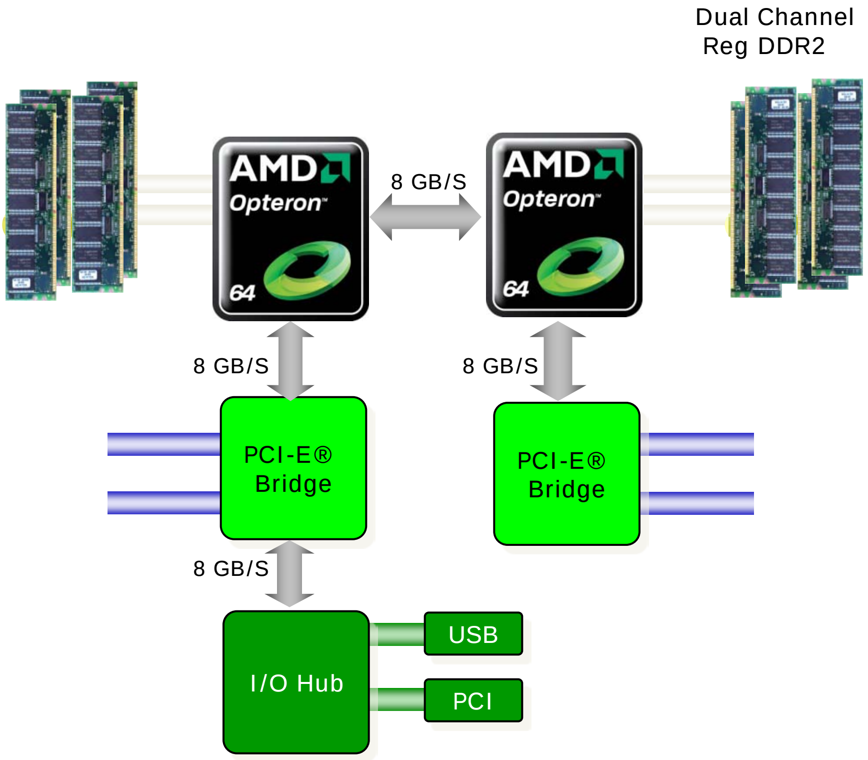
The InfiniBand Performance Gap is Increasing



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

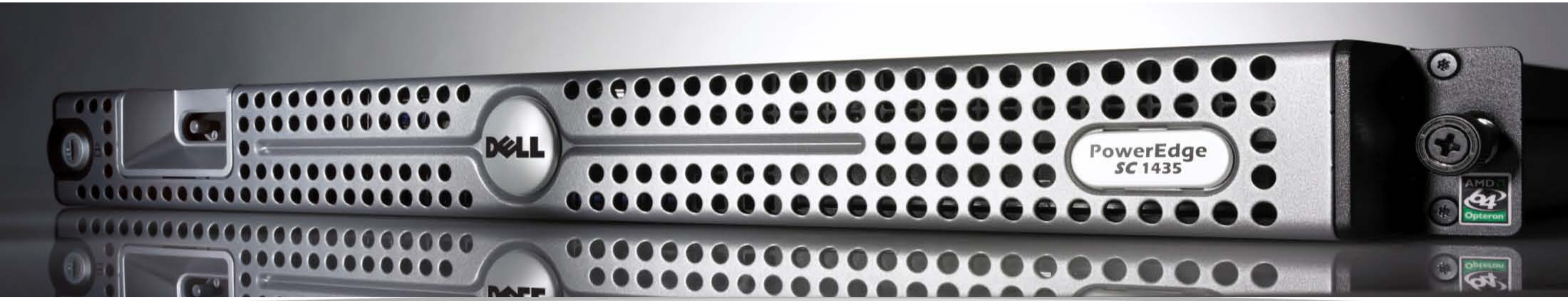


Dell PowerEdge Servers helping Simplify IT

- **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

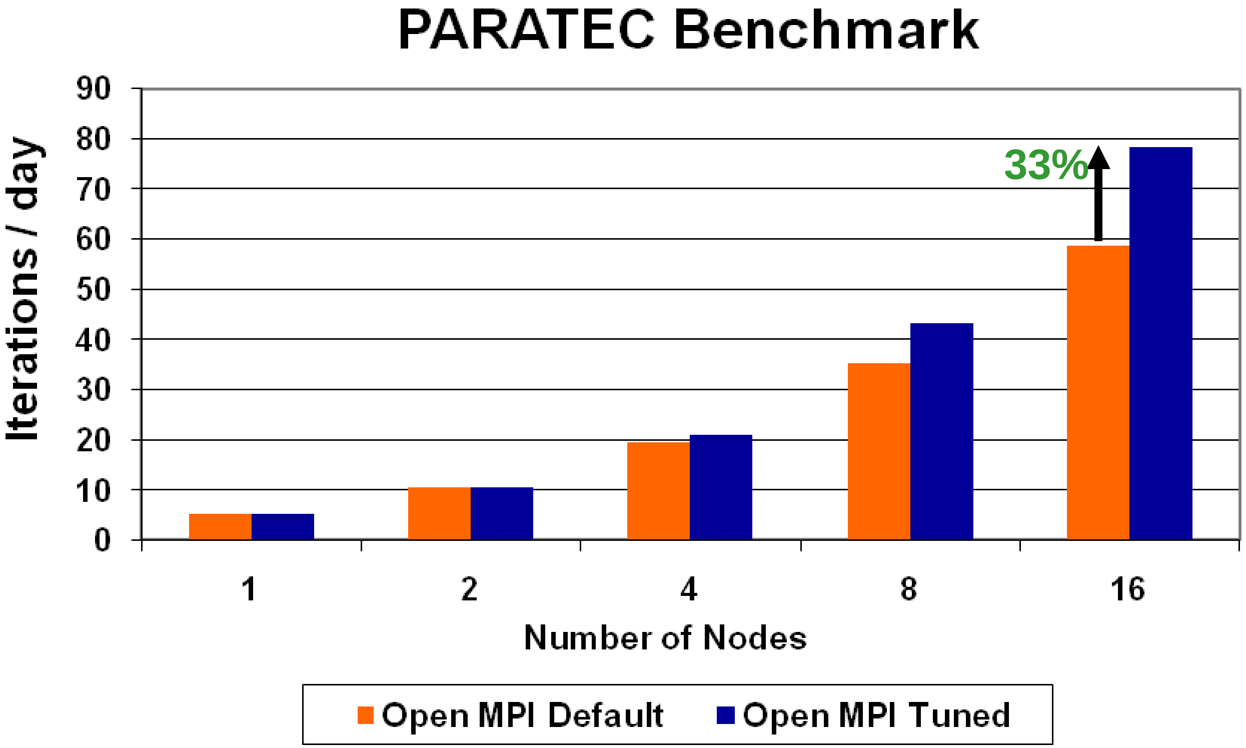
Dell PowerEdge™ Server Advantage

- 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



PARATEC Benchmark Results

- **Optimized MPI parameter provide better performance**
  - **Up to 33% higher performance with customized MPI_Gather, barrier, and XRC parameter**
    - --mca btl_openib_receive_queues X,9216,256,128,32:X,65536,256,128,32 --mca coll_tuned_use_dynamic_rules 1 --mca coll_tuned_gather_algorithm 1 --mca coll_tuned_barrier_algorithm 3

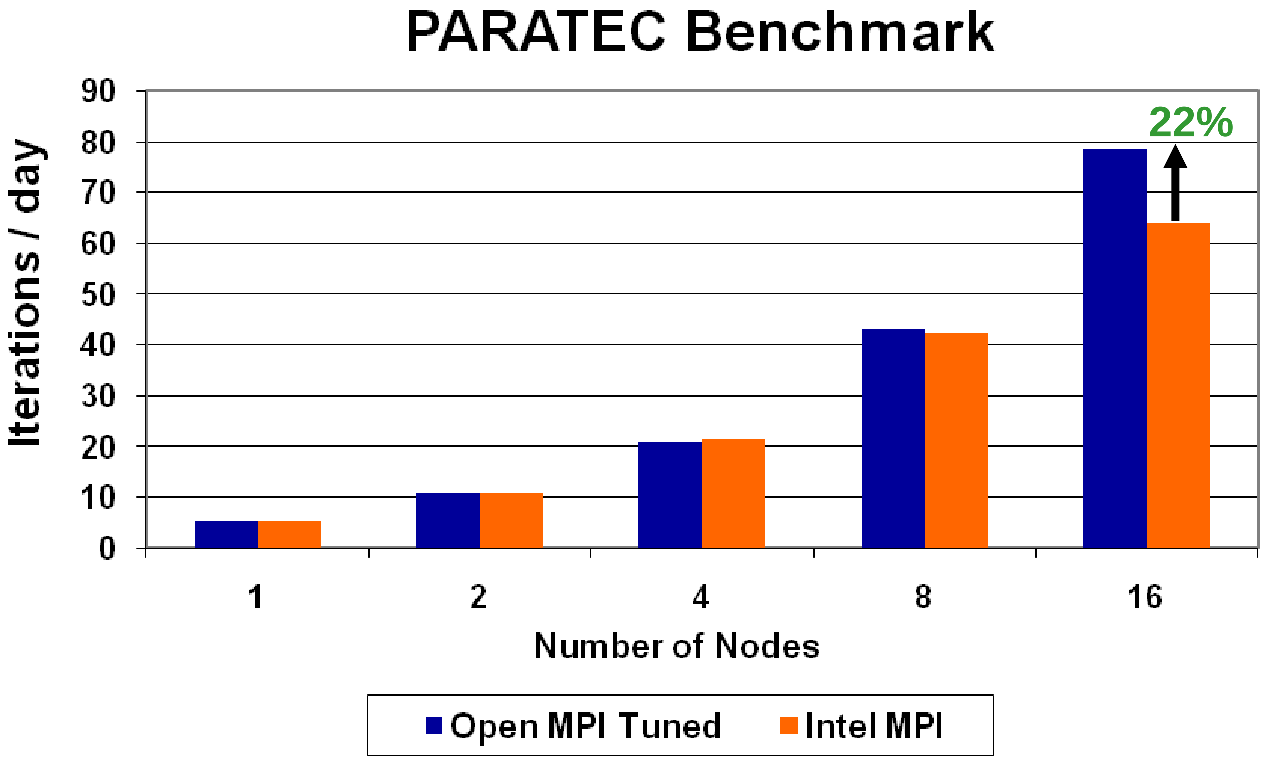


Higher is better

8-cores per node

PARATEC Benchmark Results

- **Open MPI with optimization enables higher performance**
  - Up to 22% higher performance than Intel MPI

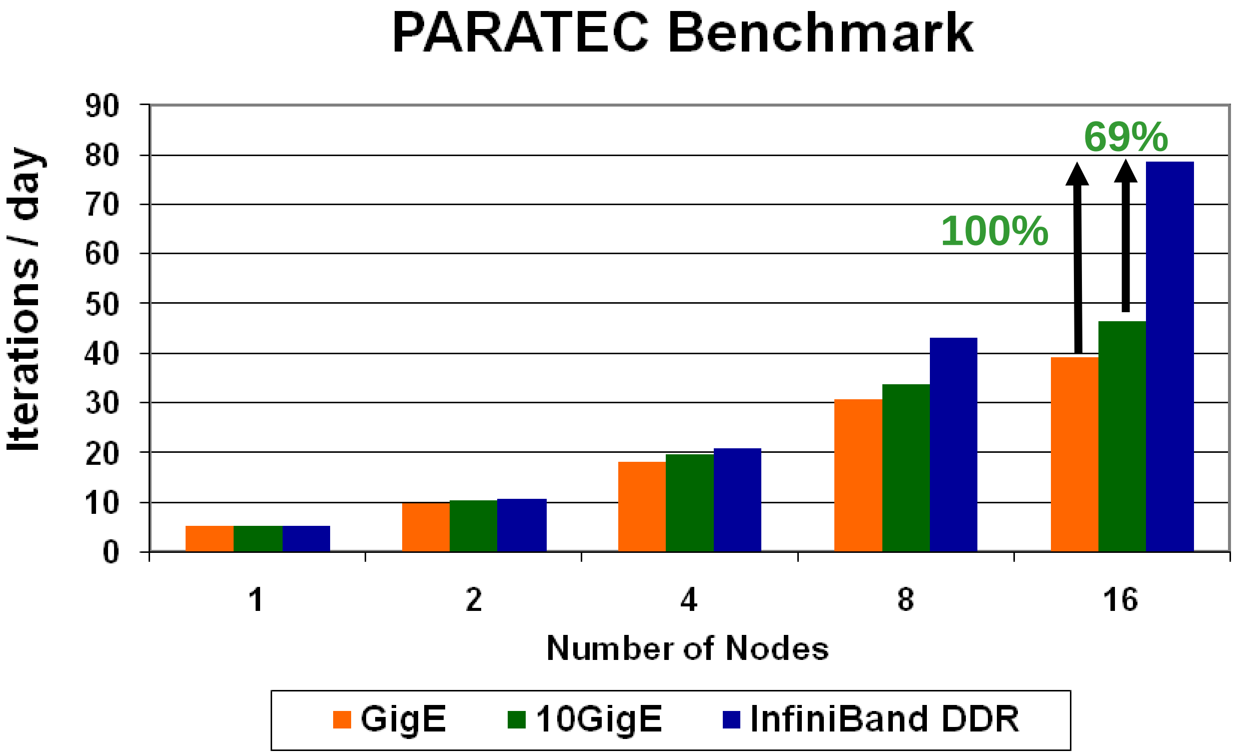


Higher is better

8-cores per node

PARATEC Benchmark Results

- **InfiniBand enables better application performance and scalability**
  - Up to 69% higher performance than 10GigE and 100% than GigE
  - 16-node cluster
- **Application performance over InfiniBand scales as cluster size increases**

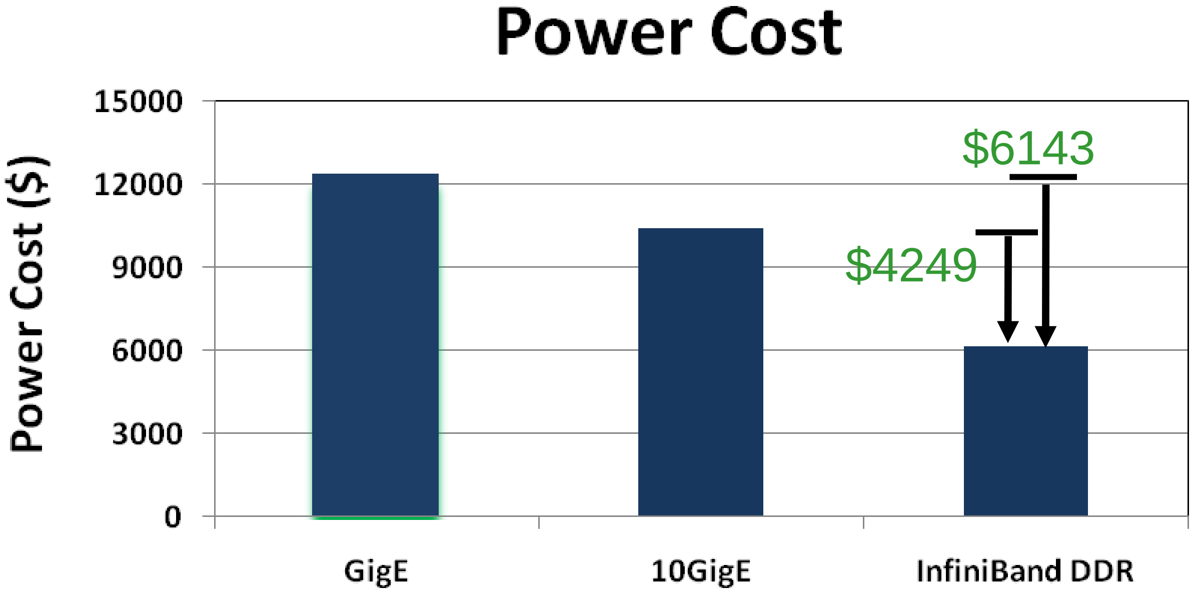


Higher is better

8-cores per node

Power Cost Savings with Different Interconnect

- **Dell economical integration of AMD CPUs and Mellanox InfiniBand**
  - To achieve same number of PARATEC jobs over GigE
  - InfiniBand saves power up to $4249 versus 10GigE and $6143 versus GigE
  - Yearly based for 16-node cluster
- **As cluster size increases, more power can be saved**



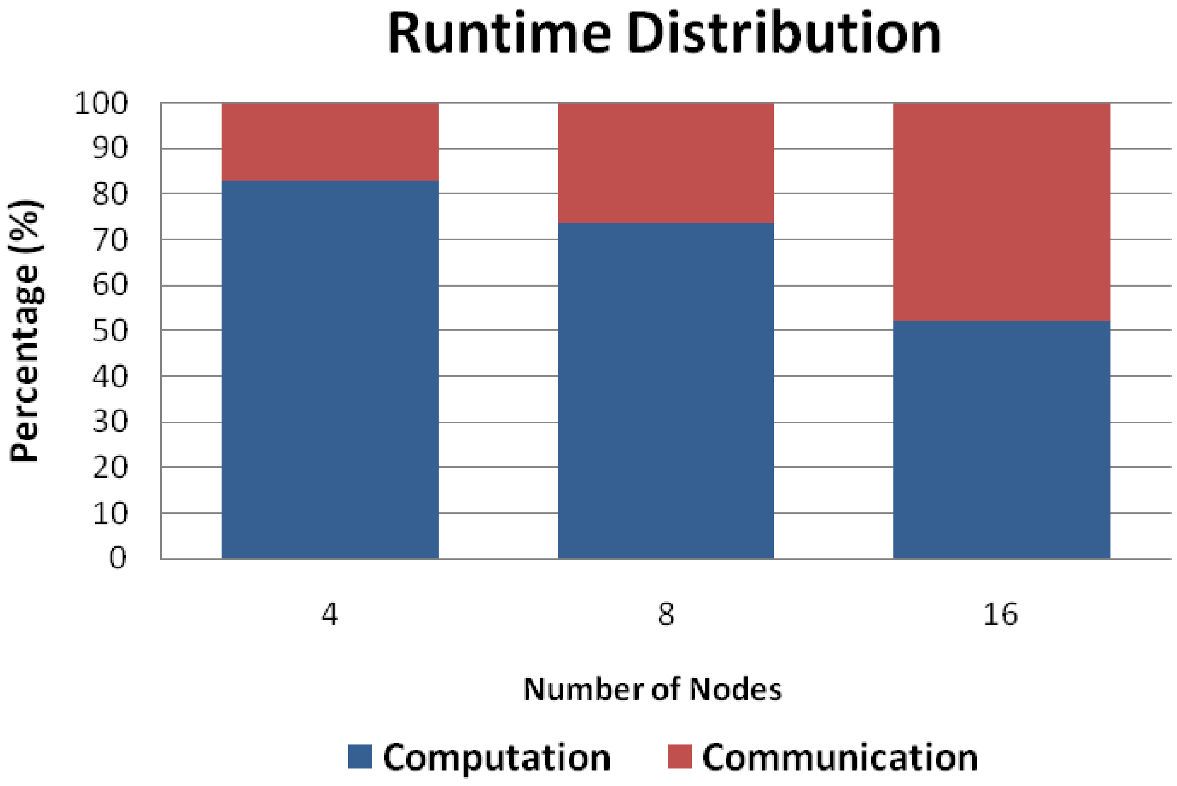
*$/KWh = KWh * $0.20*
For more information - http://enterprise.amd.com/Downloads/svrpwrusecompletefinal.pdf

PARATEC Benchmark Summary

- **Tuned MPI parameters provides better performance**
  - – Customized MPI collectives and XRC algorithm can improve application performance by 33%
- **Interconnect comparison shows**
  - – InfiniBand delivers superior performance in every cluster size versus GigE and 10GigE
  - – Performance advantage extends as cluster size increases
- **InfiniBand enables power saving**
  - – Up to $6143/year power savings versus GigE and $4249 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

PARATEC MPI Profiling – Runtime Distribution

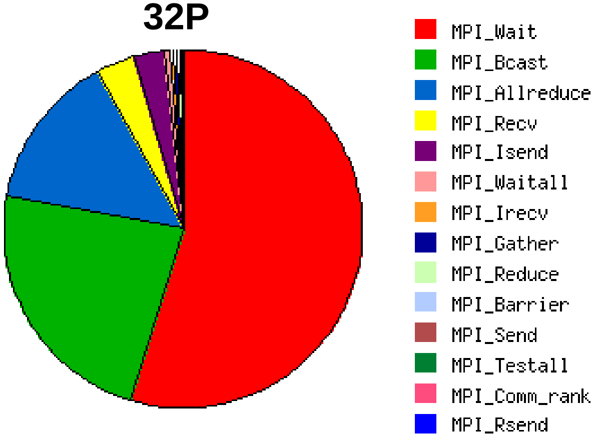
- **Mostly used MPI functions**
  - Percentage of communication increases as cluster size scales



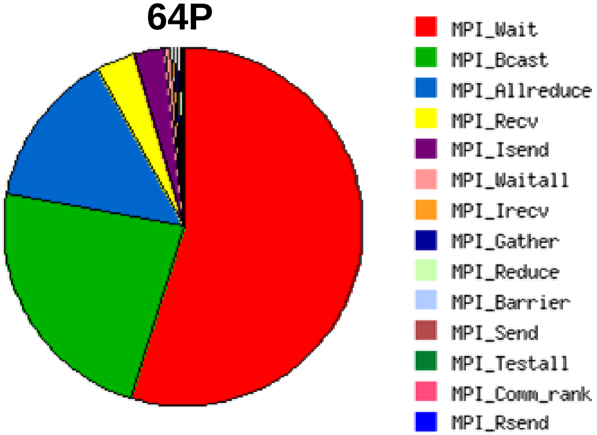
PARATEC MPI Profiling – MPI Functions

- **Mostly used MPI functions**
  - MPI_Wait, MPI_Allreduce, and MPI_Bcast are the mostly used MPI functions
  - MPI_Allreduce overhead becomes large when running processes is larger than 64

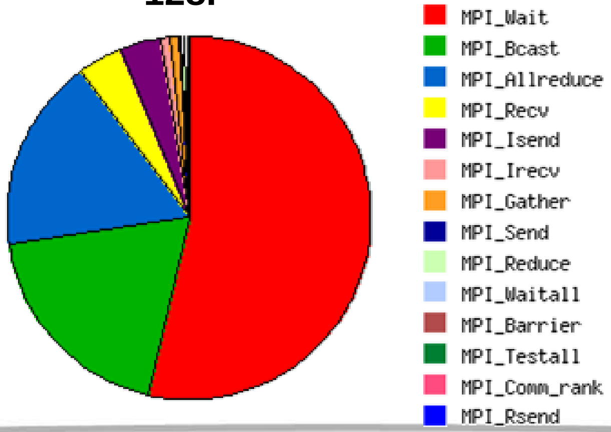
32P



64P

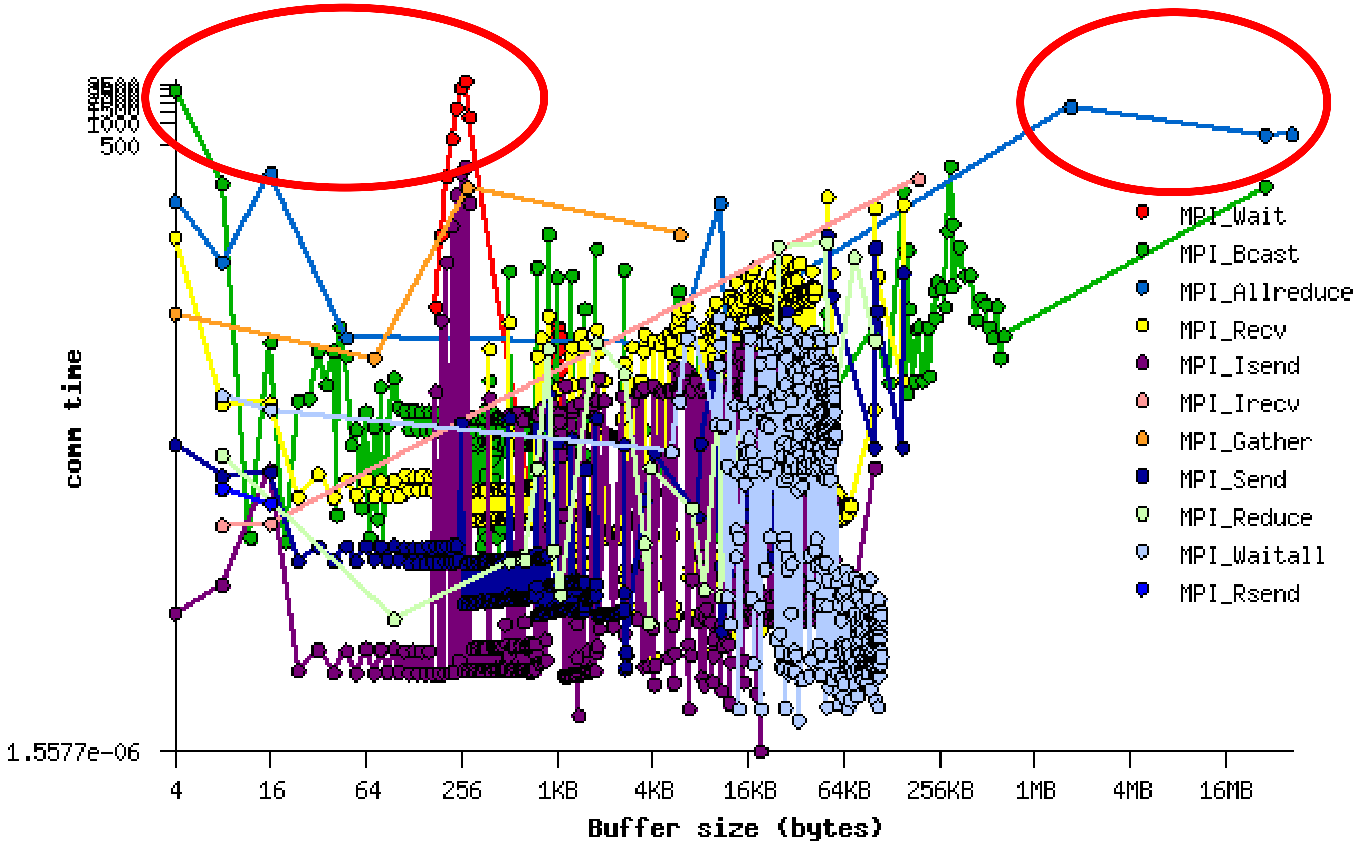


128P



PARATEC MPI Profiling – Message Size

- **Messages with big communication overhead are**
  - Large messages >1MB
  - Small message <256Bytes



128 Processes

PARATEC Profiling Summary

- **PARATEC was profiled to identify its communication patterns**
  - MPI collective and point-to-point create the big communication overhead
  - Both small and large messages are used
  - Number of messages increases with cluster size
- **Interconnects effect to PARATEC performance**
  - 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

AMD The future is fusion

DELL

Mellanox TECHNOLOGIES

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