HPC ADVISORY COUNCIL

MPQC Performance Benchmark and Profiling

May 2009

AMD The future is fusion

DELL

Lawrence Livermore National Laboratory

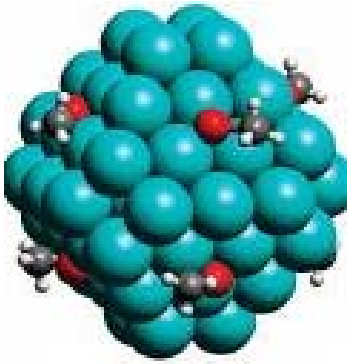
Mellanox TECHNOLOGIES

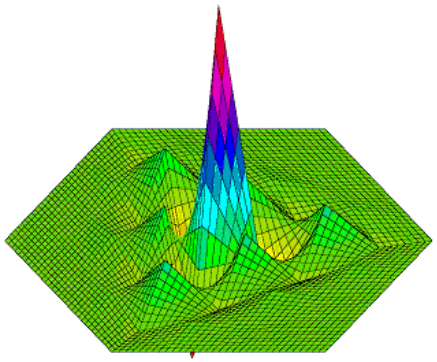
Note

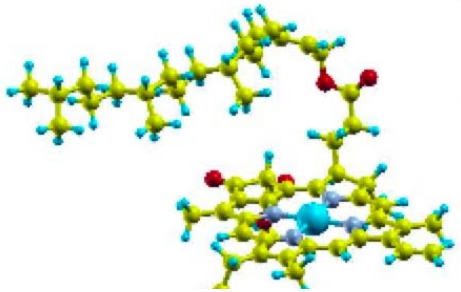
- **The following research was performed under the HPC Advisory Council activities**
  - Participating vendors: AMD, Dell, Mellanox
  - Compute resource - HPC Advisory Council Cluster Center
- **The participating members would like to thank Matt L. Leininger (LLNL) for his support and guidelines**
- **For more info please refer to**
  - www.mellanox.com, www.dell.com/hpc, www.amd.com

MPQC

- **Massively Parallel Quantum Chemistry Program (MPQC)**
  - Computes properties of atoms and molecules using the time independent Schrödinger equation
    - Closed shell and general restricted open-shell Hartree-Fock energies and gradients
    - Second order open-shell perturbation and Z-averaged perturbation theory energies
    - Second order closed shell Moeller-Plesset perturbation theory energies and gradients
- **MPQC runs on a wide range of architecture**
  - From individual workstations to massively parallel computers







Objectives

- **The presented research was done to provide best practices**
  - MPQC performance benchmarking
  - Interconnect performance comparisons
  - Ways to increase MPQC productivity
  - Power-aware consideration

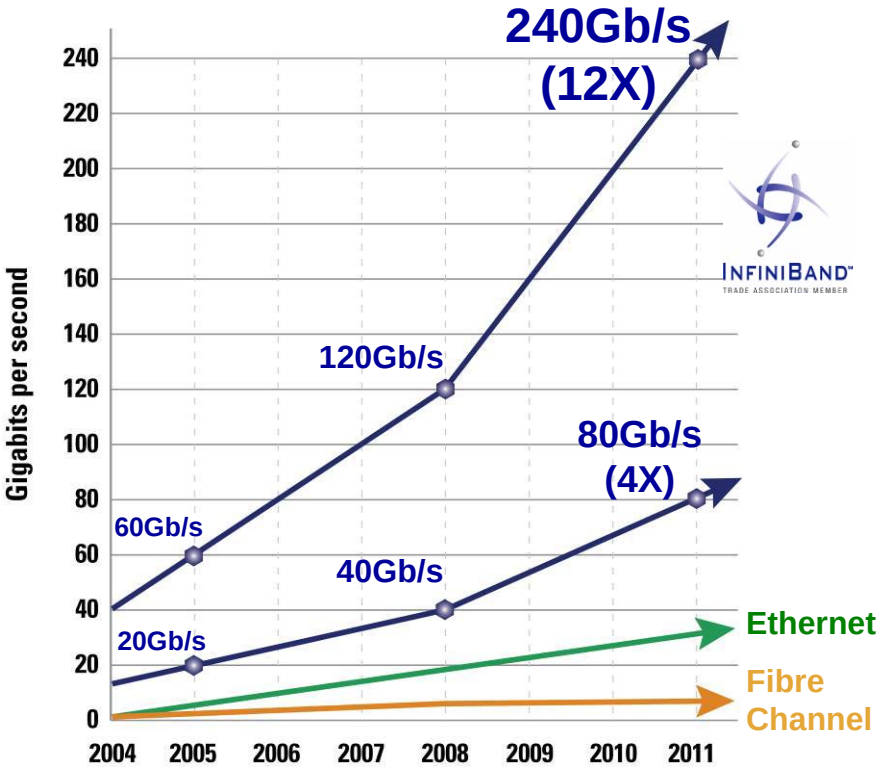
Test Cluster Configuration

- Dell™ PowerEdge™ SC 1435 24-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: RHEL5U2, OFED 1.4 InfiniBand SW stack
- MPI: OpenMPI 1.3.2 (MPI thread enabled, the only MPI supported by MPQC)
- Application: MPQC 2.3.1
- Benchmark Dataset
  - Uracil dimer aug-cc-pVDZ basis set
  - Uracil dimer aug-cc-pVTZ basis set

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**

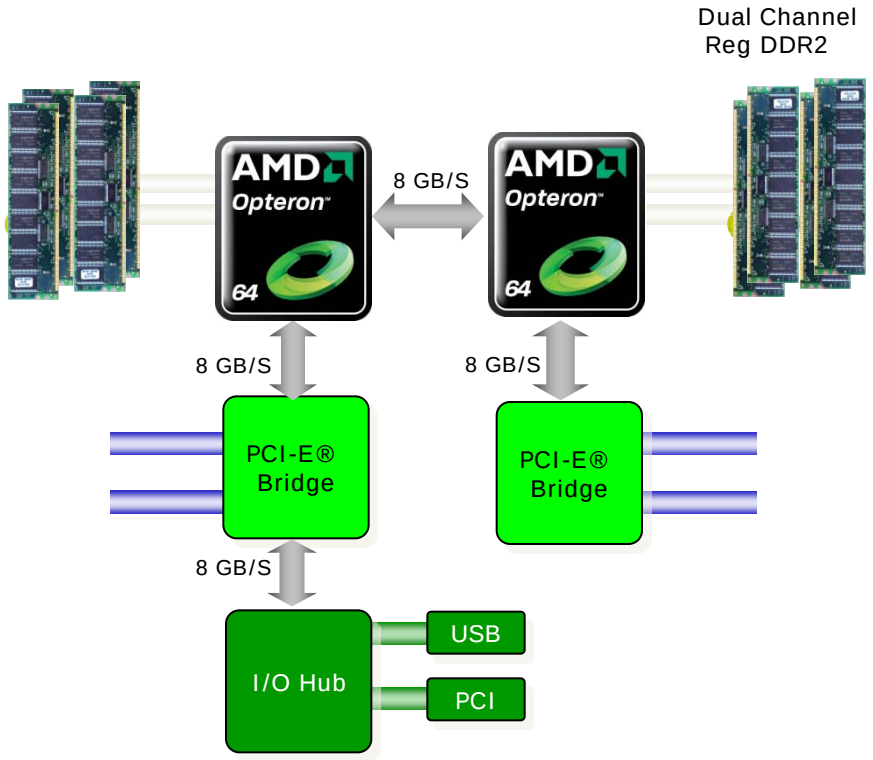
The InfiniBand Performance Gap is Increasing



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

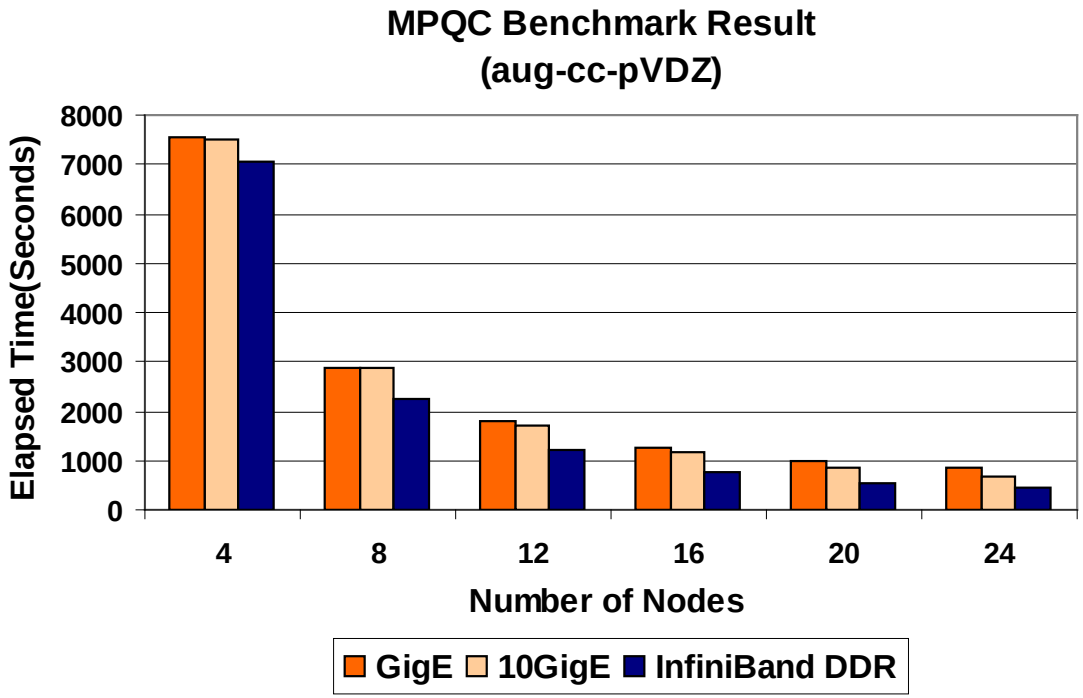


Dell PowerEdge Servers helping Simplify IT

- **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

MPQC Benchmark Results

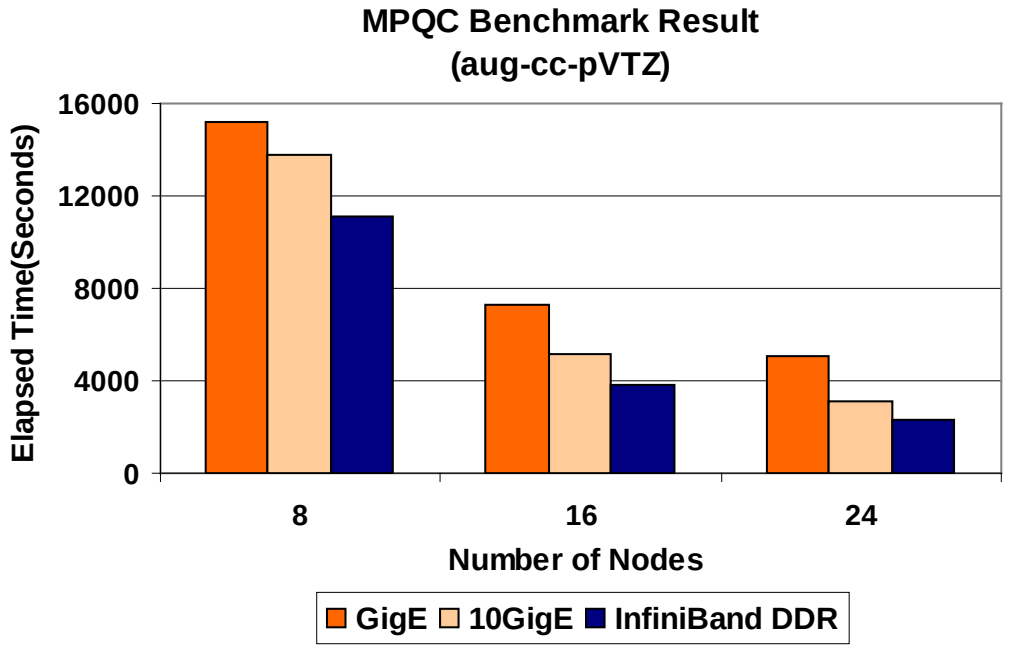
- **Input Dataset**
  - MP2 calculations of the uracil dimer binding energy using the aug-cc-pVDZ basis set
- **InfiniBand delivers higher performance and scalability**
  - Outperforms GigE by up to 87% and 10GigE by up to 47%



Lower is better

MPQC Benchmark Results

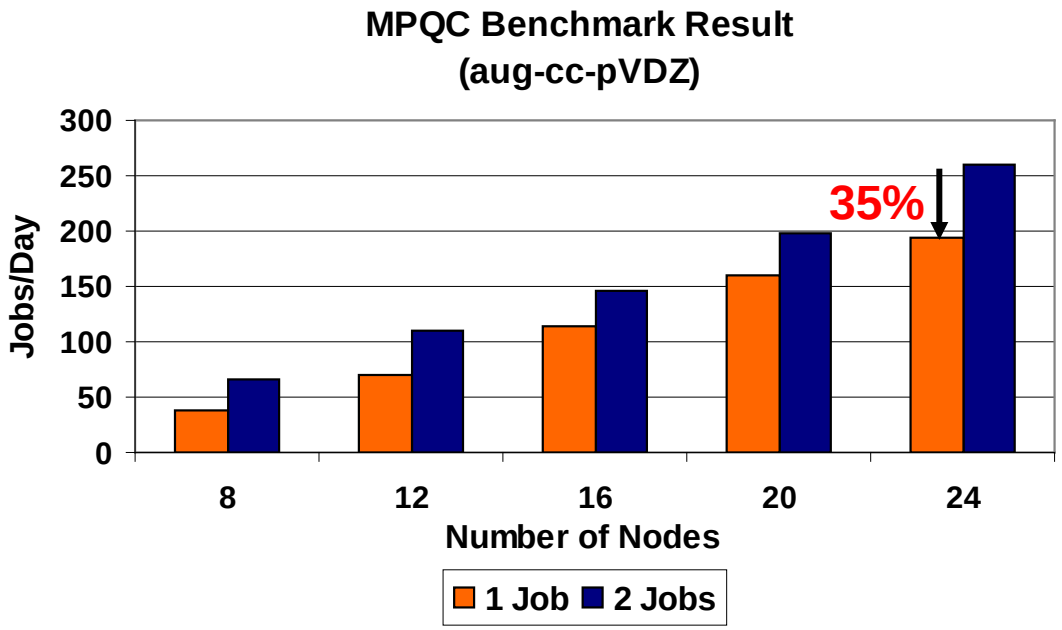
- **Input Dataset**
  - MP2 calculations of the uracil dimer binding energy using the aug-cc-pVTZ basis set
- **InfiniBand enables best scalability**
  - Performance accelerates with cluster size
  - Outperforms GigE by up to 124% and 10GigE by up to 38%



Lower is better

MPQC Performance Results - Productivity

- **InfiniBand increases productivity by allowing multiple jobs to run simultaneously**
  - Providing required productivity for MPQC computation
- **Two cases are presented**
  - Single job over the entire systems
  - Two jobs, each on four cores per server
- **Two jobs per node increases productivity by up to 35%**

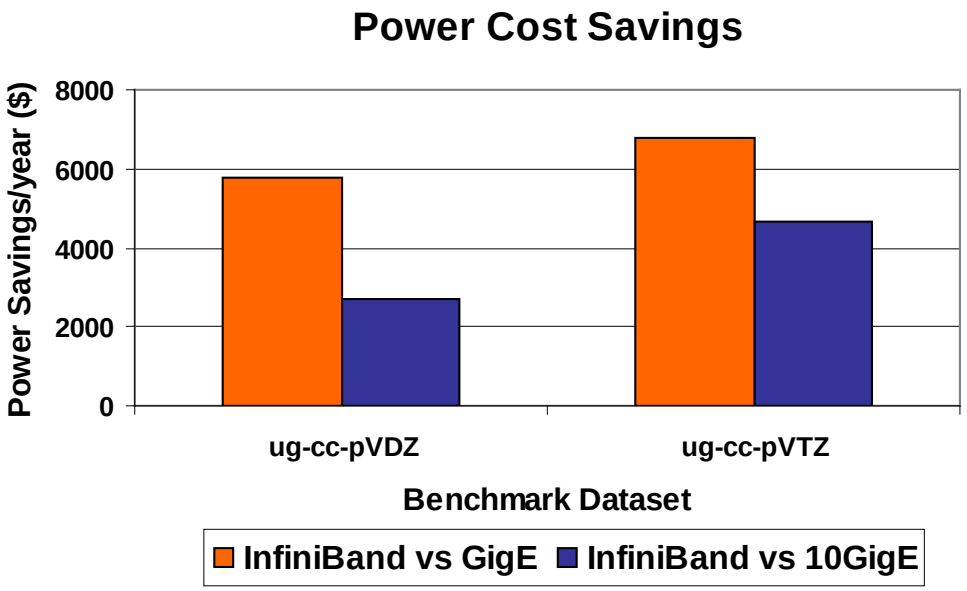


Higher is better

InfiniBand DDR

Power Cost Savings with InfiniBand

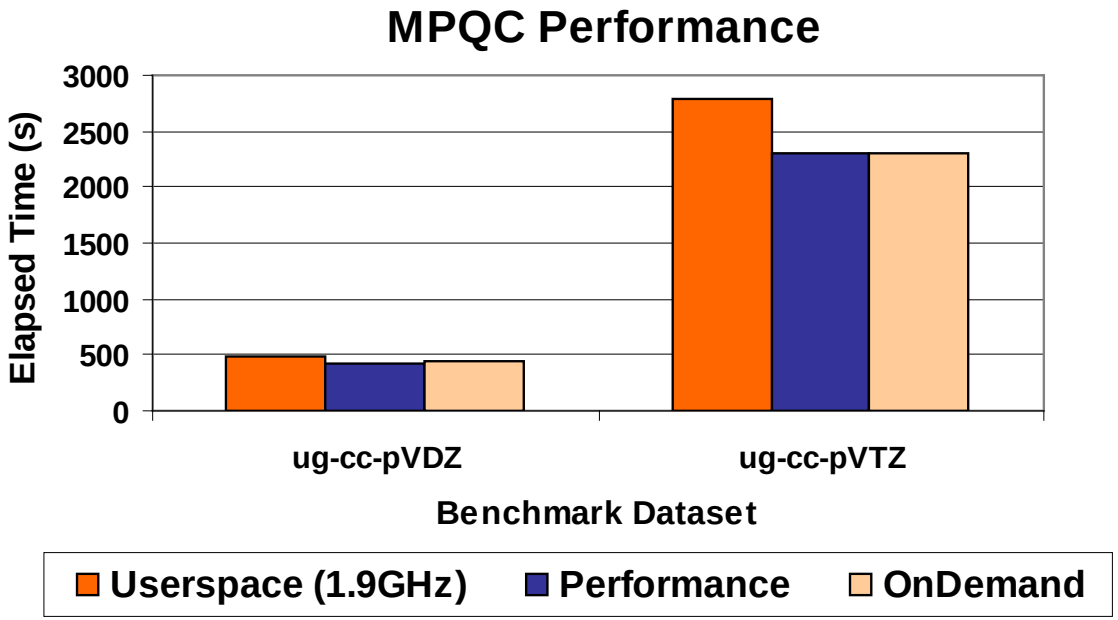
- **InfiniBand saves up to ~$6500 power to finish the same number of jobs run over GigE**
  - Yearly based for 24-node cluster
- **As cluster size increases, more power can be saved**



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

MQPC Performance with CPU Frequency Scaling

- **Enabling CPU Frequency Scaling**
  - Userspace – reducing CPU frequency to 1.9GHz
  - Performance – setting for maximum performance (CPU frequency of 2.6GHz)
  - OnDemand – Maximum performance per application activity
- **Userspace increases job run time since CPU frequency is reduced**
- **Performance and OnDemand enable similar performance**
  - Due to high resource demands from the application

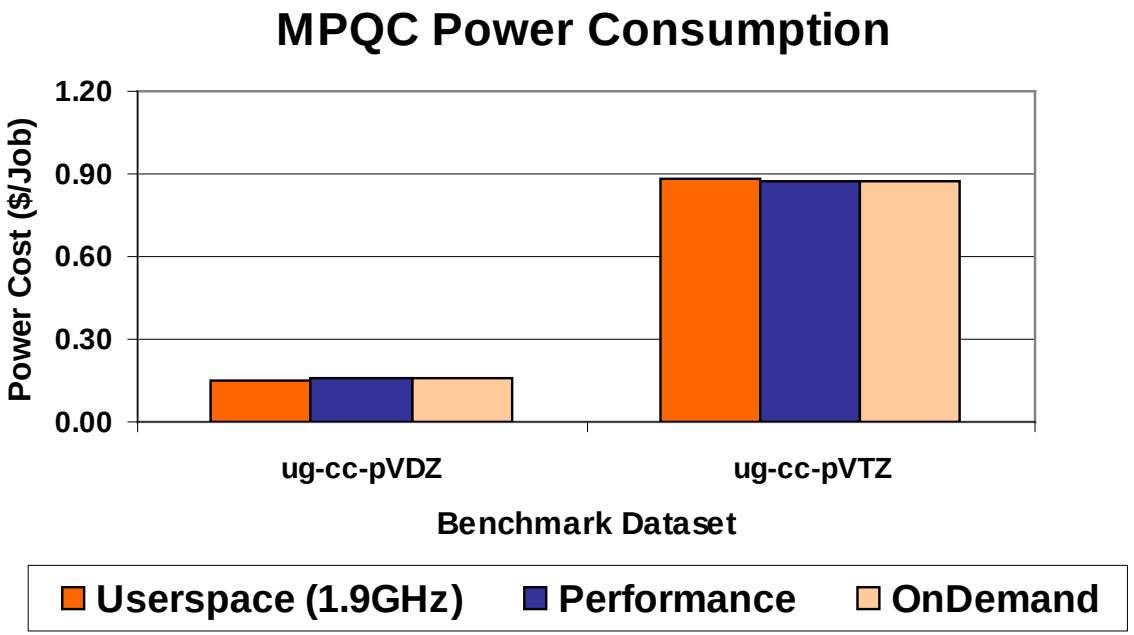


24 Node Cluster

InfiniBand DDR

Power Consumption with CPU Frequency Scaling

- **MPQC has similar power/job consumption with Performance and OnDemand options**
- **Setting CPU frequency to a fixed low value causes higher power consumption with compute-intensive test case**
  - Does provide benefits for low-intensive workload



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

24 Node Cluster

InfiniBand DDR

Power Consumption with CPU Frequency Scaling

- **Dell platform utilizing CPU Frequency with latest AMD CPUs and InfiniBand improves power efficiency and reduce job cost**

- **Using OnDemand (automatic reduction of CPU power consumption per application usage) is proven to be effective**
  - Reduces power consumption in idle, and provide same performance capabilities compared to static setting the CPUs for maximum performance

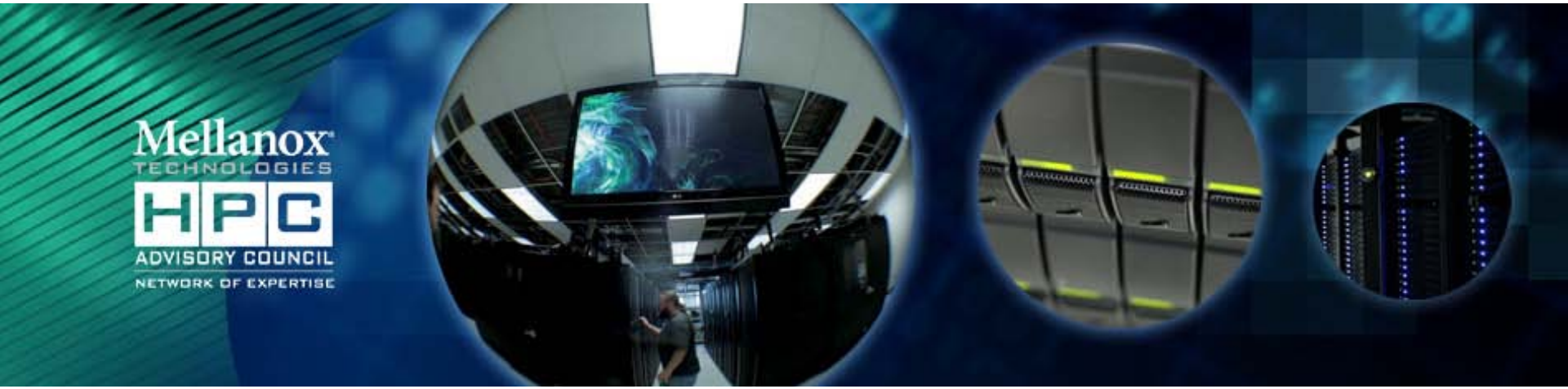
- **Reducing the CPU frequency for power saving was proven to be inefficient and actually increase power consumption per job for compute intensive applications**
  - Manually setting CPU frequency to does not provide value for compute intensive workloads

MPQC Summary and Conclusions

- **MPQC performance relies on**
  - Scalable HPC systems and interconnect solutions
  - Low latency and high throughput interconnect technology
  - NUMA aware application for fast access to local memory
- **Efficient job placement can increase MPQC productivity**
- **Dell platform utilizing InfiniBand and On Demand CPU power schemes enables big power savings**
  - Above $6500 per year for a 24-node cluster configuration
- **CPU Frequency Scaling with latest AMD CPUs improves power efficiency**
  - OnDemand option can dynamically adjust CPU frequency to meet performance requirements while reducing power consumption during system idle time
  - Manually setting CPU frequency to does not provide value for compute intensive workloads

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