HPC ADVISORY COUNCIL

NWChem Performance Benchmark and Profiling

July 2009

AMD The future is fusion

DELL

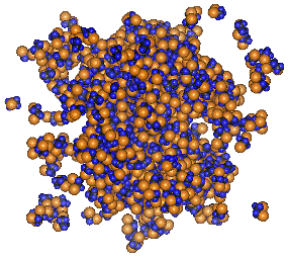
Mellanox TECHNOLOGIES

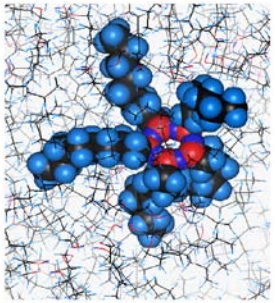
Note

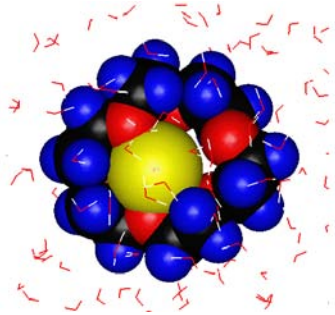
- **The following research was performed under the HPC Advisory Council activities**
  - Participating vendors: AMD, Dell, Mellanox
  - Compute resource - HPC Advisory Council Cluster Center
- **For more info please refer to**
  - www.mellanox.com, www.dell.com/hpc, www.amd.com,

NWChem

- **NWChem is a computational chemistry package**
  - NWChem has been developed by the Molecular Sciences Software group of the Environmental Molecular Sciences Laboratory (EMSL) at the Pacific Northwest National Laboratory (PNNL)
- **NWChem provides many methods to compute the properties of molecular and periodic systems**
  - Using standard quantum mechanical descriptions of the electronic wavefunction or density
- **NWChem has the capability to perform classical molecular dynamics and free energy simulations**
  - These approaches may be combined to perform mixed quantum-mechanics and molecular-mechanics simulations







Objectives

- **The presented research was done to provide best practices**
  - NWChem performance benchmarking
  - Performance comparison with different MPI libraries
  - Interconnect performance comparisons
  - Understanding NWChem communication patterns
  - Power-efficient simulations

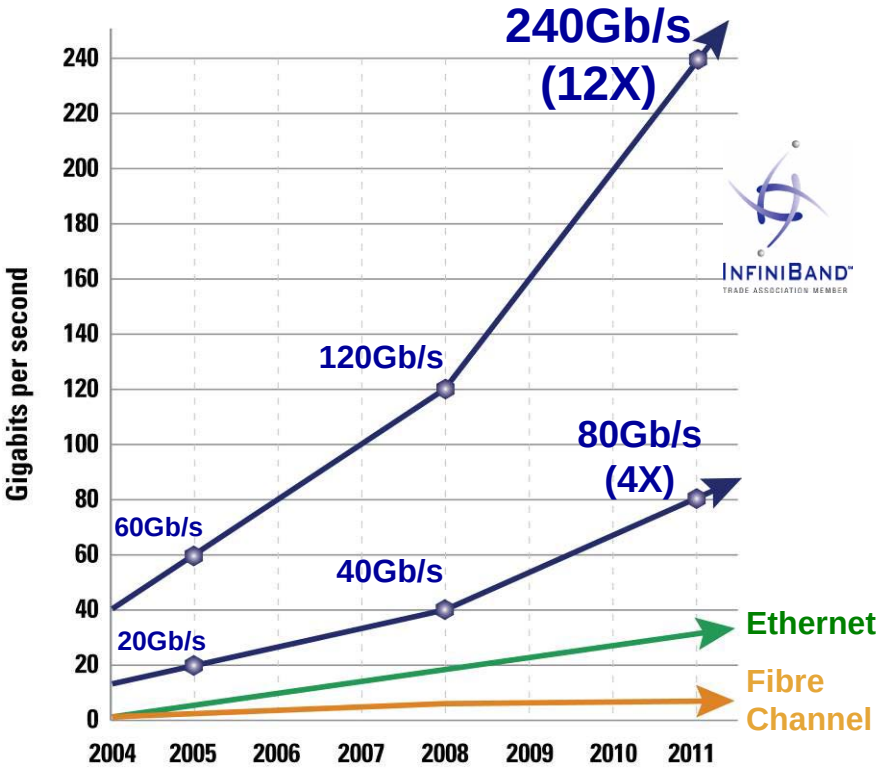
Test Cluster Configuration

- 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: HP-MPI 2.3, Open MPI 1.3.2, and Mvapich-1.1
- Application: NWChem 5.1.1
- Math Library: AMD Core Math Library (ACML)
- Benchmark Workload
  - MP2 gradient calculation of the (H2O)7 molecule - H_2O_7
- Flags for ifort compiler
  - -i8 -align -w -g -vec-report1 -i8 -O3 -prefetch -unroll -tpp7 -ip

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**

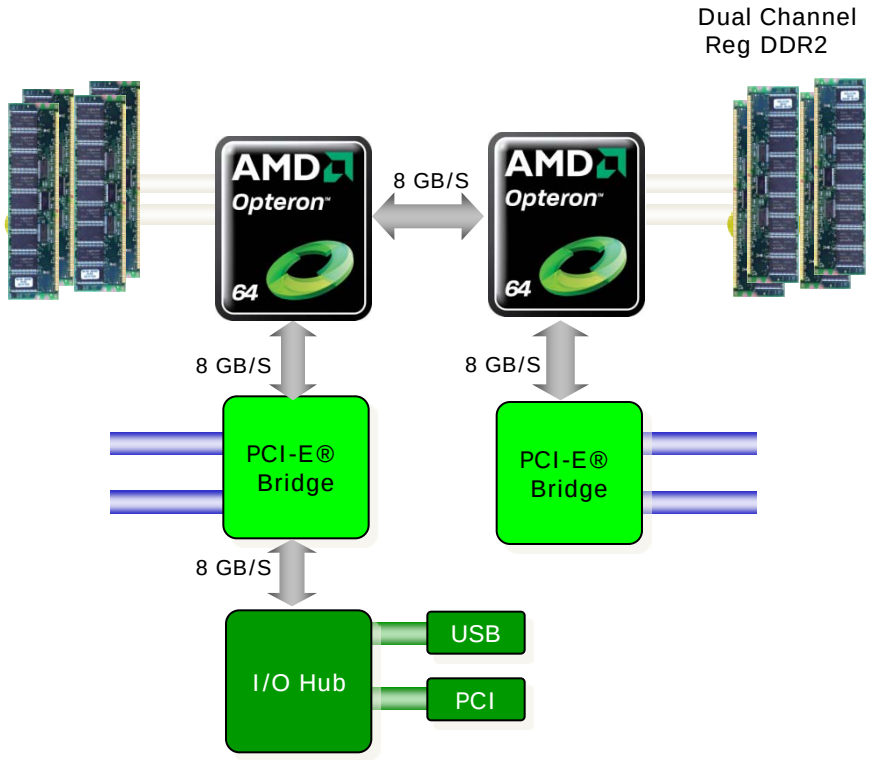
The InfiniBand Performance Gap is Increasing



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

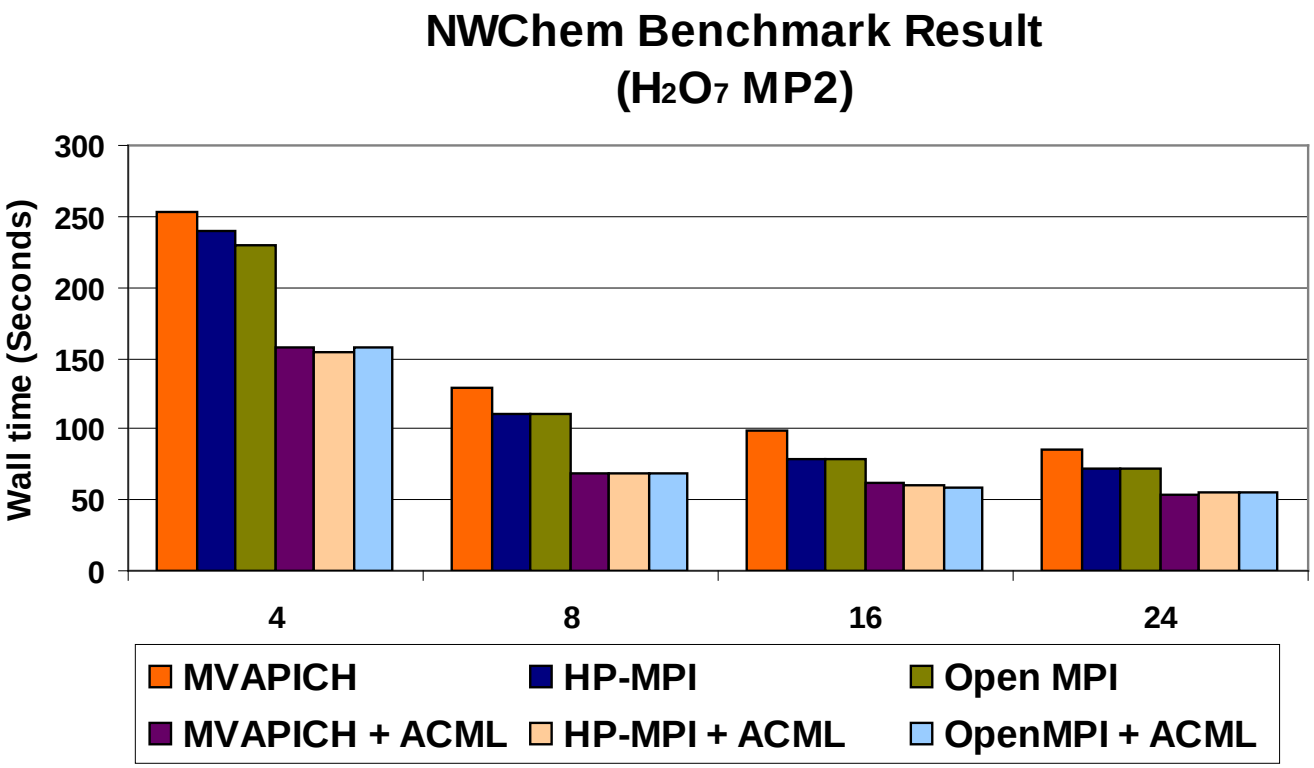


Dell PowerEdge Servers helping Simplify IT

- **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

NWChem Benchmark Results

- **Input Dataset - H_2O_7**
- **ACML provides higher performance and scalability versus the default BLAS library**

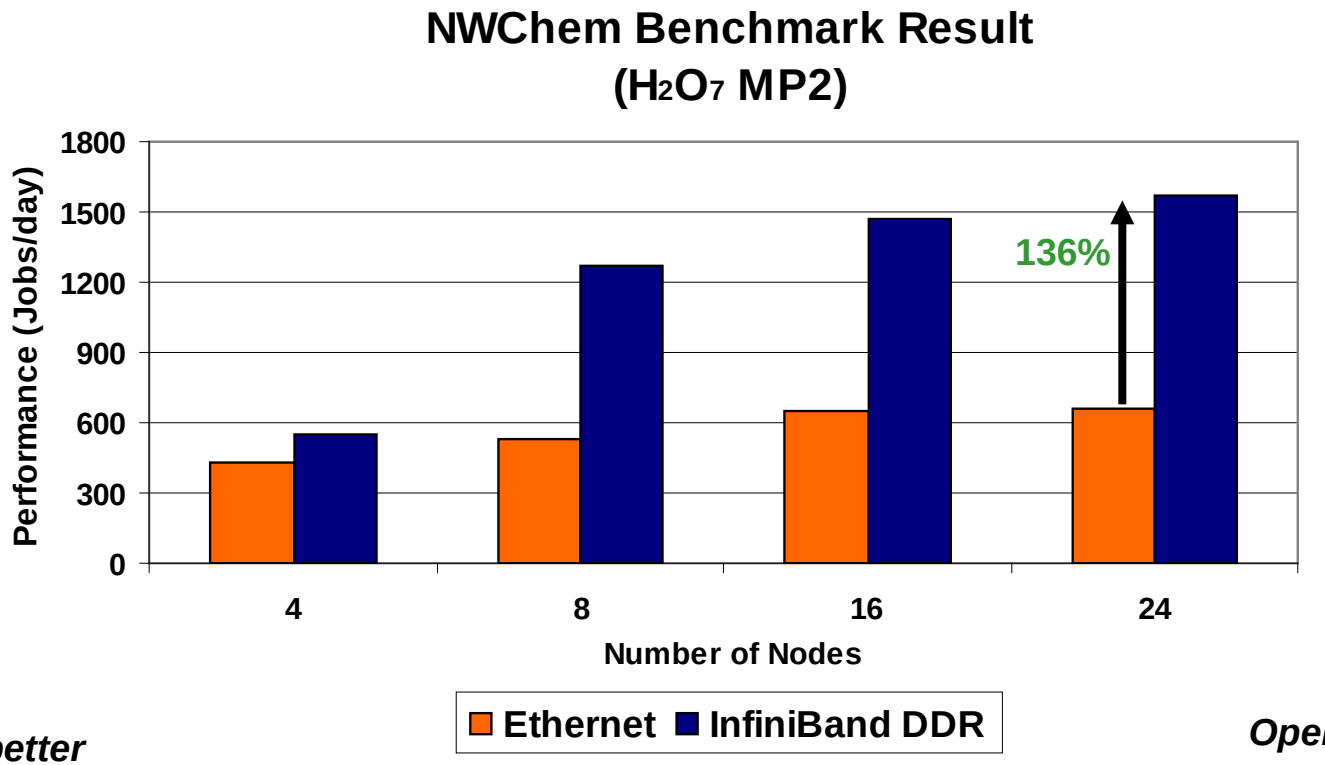


Lower is better

InfiniBand DDR

NWChem Benchmark Results

- **Input Dataset - H2O7**
- **InfiniBand enables better performance and scalability**
  - Up to 136% higher productivity versus Gigabit Ethernet

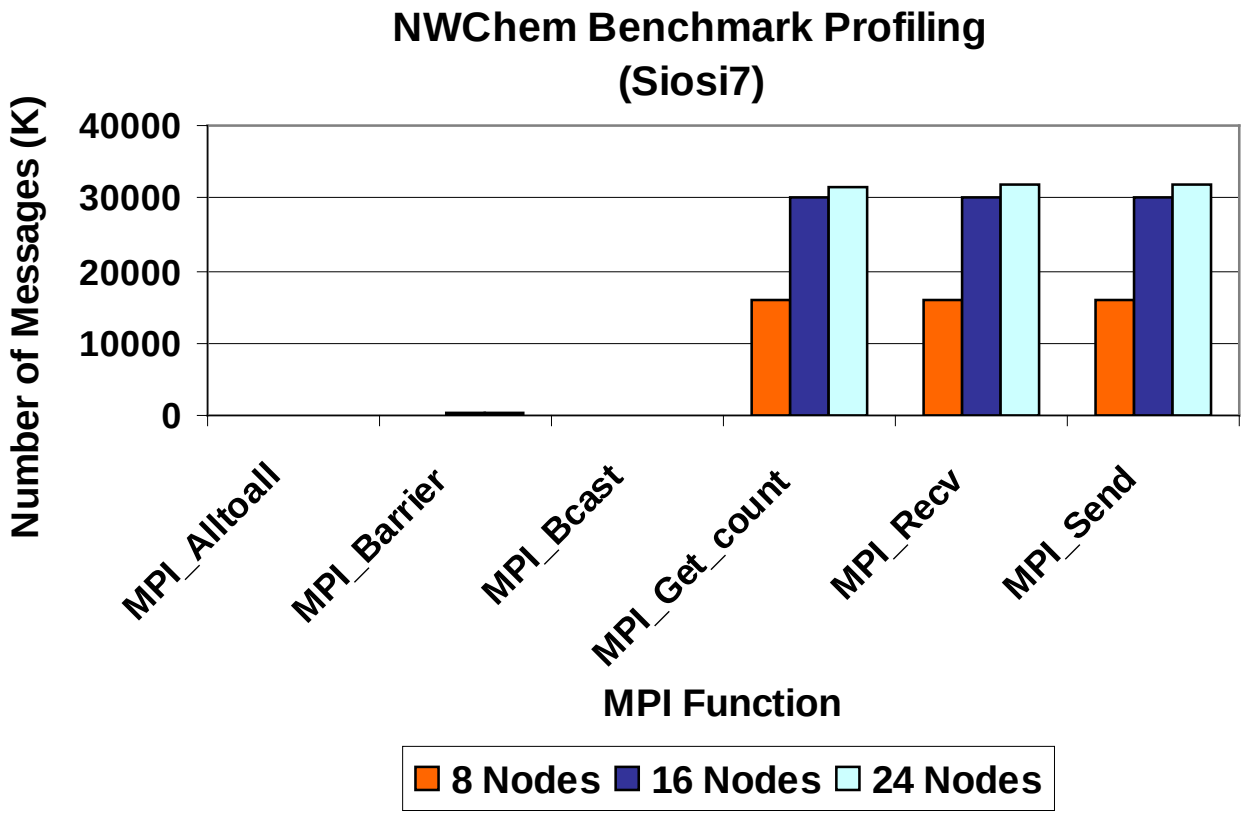


Higher is better

Open MPI + ACML

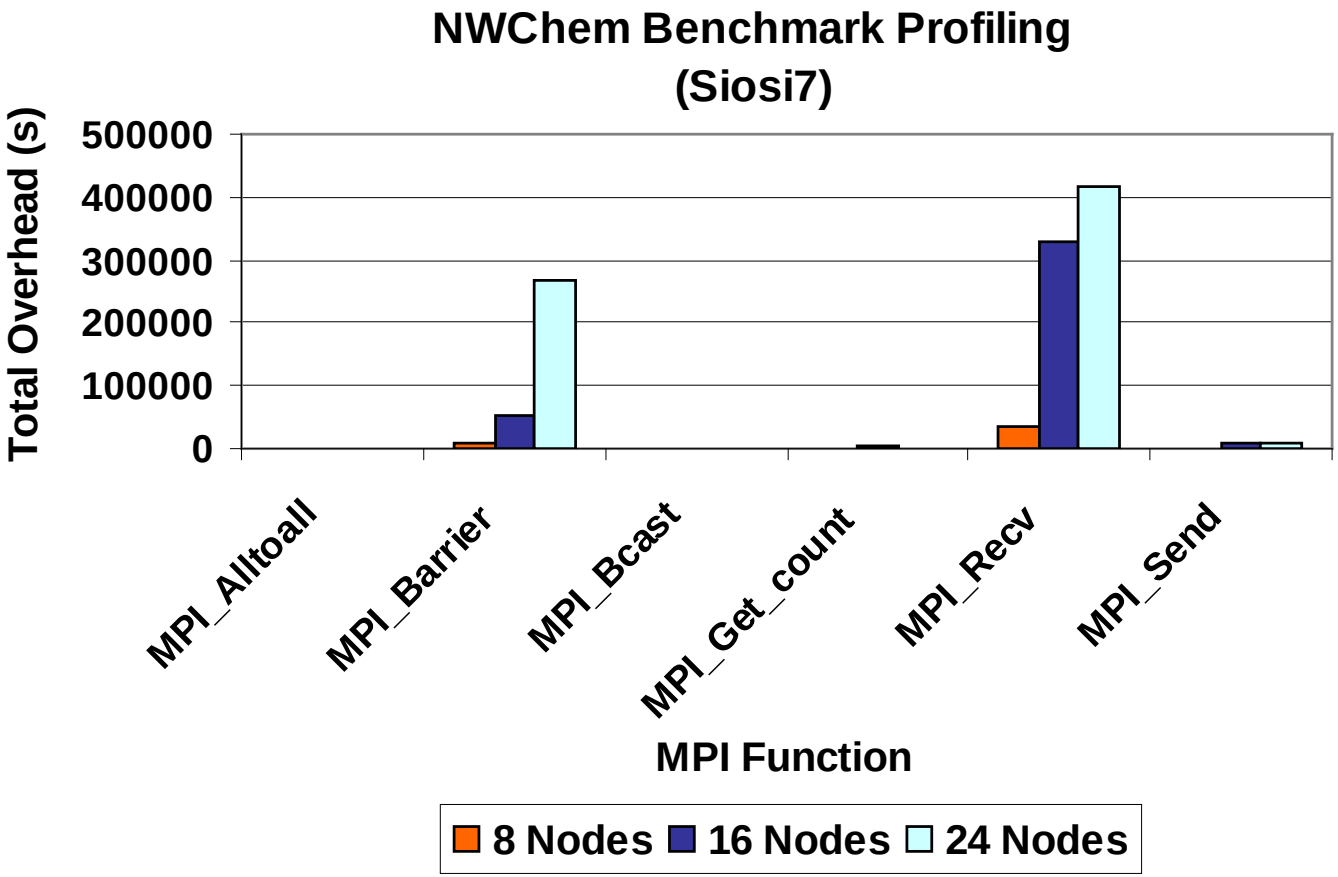
NWChem Profiling – MPI Functions

- **Mostly used MPI functions**
  - MPI_Get_Count, MPI_Recv, and MPI_Send



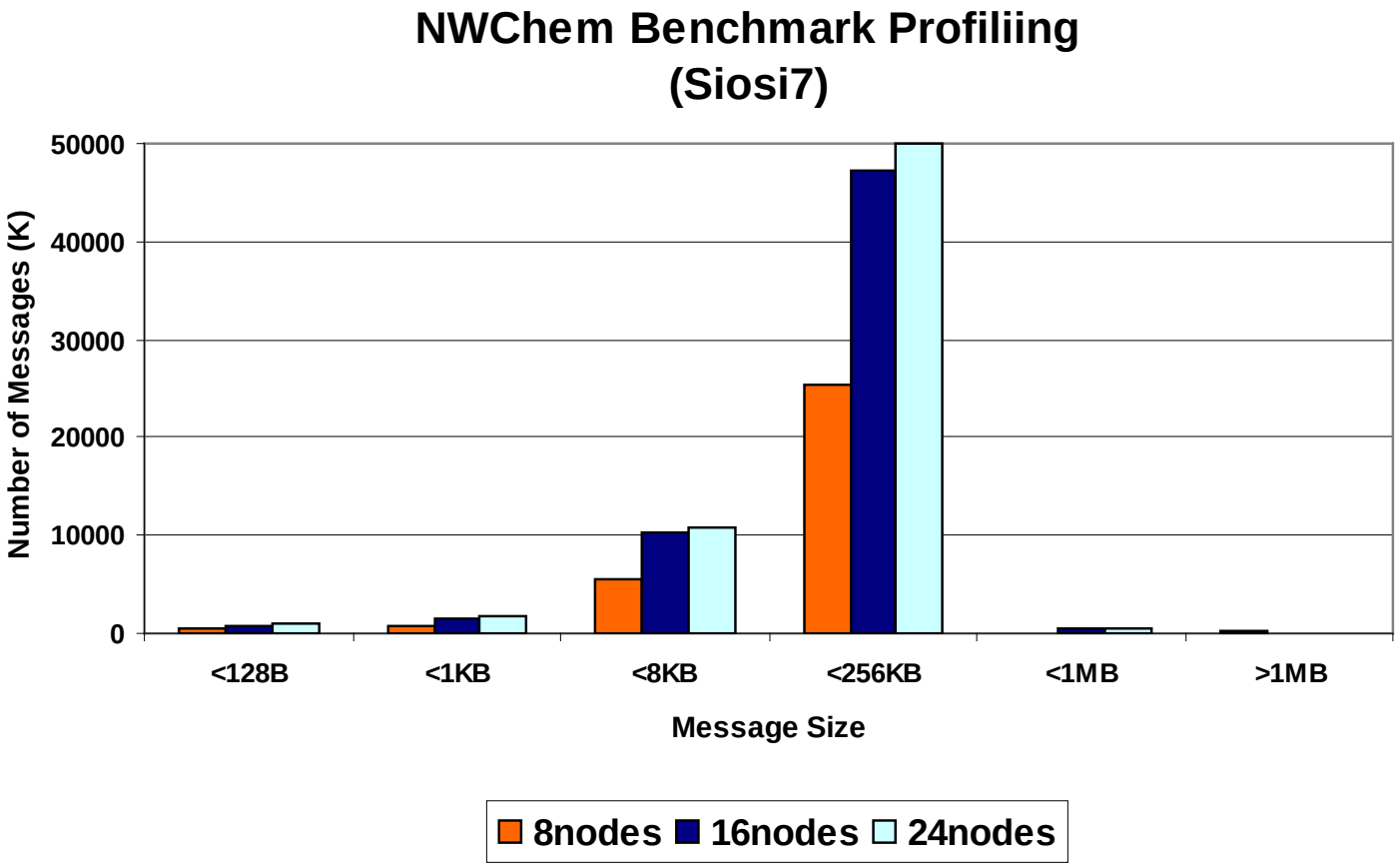
NWChem Profiling – Timing per MPI Function

- **MPI_Recv and MPI_Barrier show high communication overhead**



NWChem Profiling – Message Transferred

- Most data related MPI messages are within 8KB-256KB in size
- Number of messages increases with cluster size



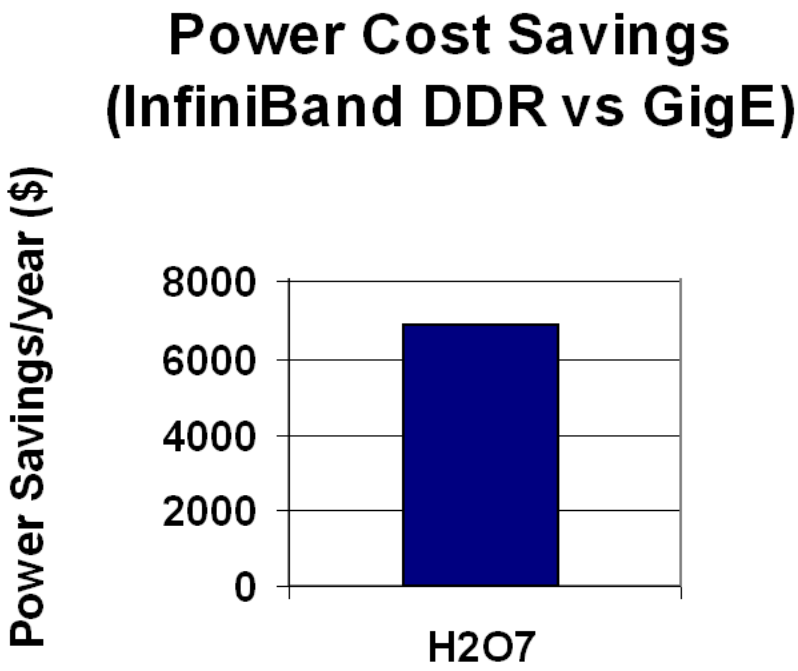
NWChem Profiling Summary

- **NWChem is profiled to identify its communication pattern**
- **Frequent used message sizes**
  - 8KB-256KB messages for data related communications
  - Number of messages increases with system size
  - Message size kept with system size

- **Interconnects effect to NWChem performance**
  - Interconnect throughput significantly influences NWChem performance
  - The need for higher throughput increases with system size

Power Cost Savings

- **Dell economical integration of AMD CPUs and Mellanox InfiniBand saves up to $7000 in power**
  - Versus using Gigabit Ethernet as the connectivity solutions
  - 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

Conclusions

- **ACML enables higher NWChem performance**
  - Faster than GCC compiler with default BLAS library
- **HP MPI and Open MPI shows better performance than MVAPICH**
- **NWChem relies on interconnect with highest throughput**
  - Most transferred messages are 8KB-256KB messages
  - Number of messages scales up as number of processes increases
- **InfiniBand enables highest NWChem performance and scalability**
  - Nearly 136% higher productivity versus GigE
  - performance gain increases with system size
    - Higher performance expected with more nodes
- **Balanced system enables high productivity**
  - Optimal job placement can maximize NWchem simulations

Thank You

HPC Advisory Council