NWChem Performance Benchmark and Profiling

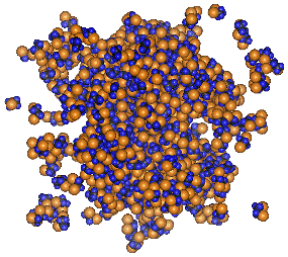
June 2009

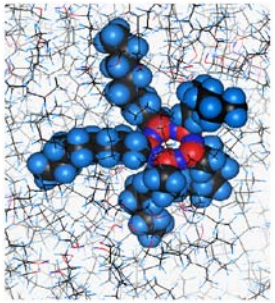
Note

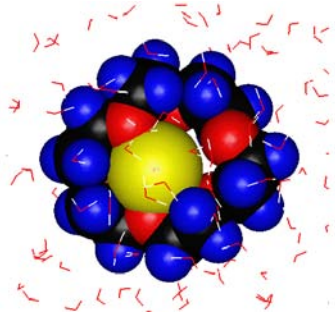
- **The following research was performed under the HPC Advisory Council activities**
  - Participating vendors: Intel, Dell, Mellanox
  - Compute resource - HPC Advisory Council Cluster Center
- **For more info please refer to**
  - www.mellanox.com, www.dell.com/hpc, www.intel.com,

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

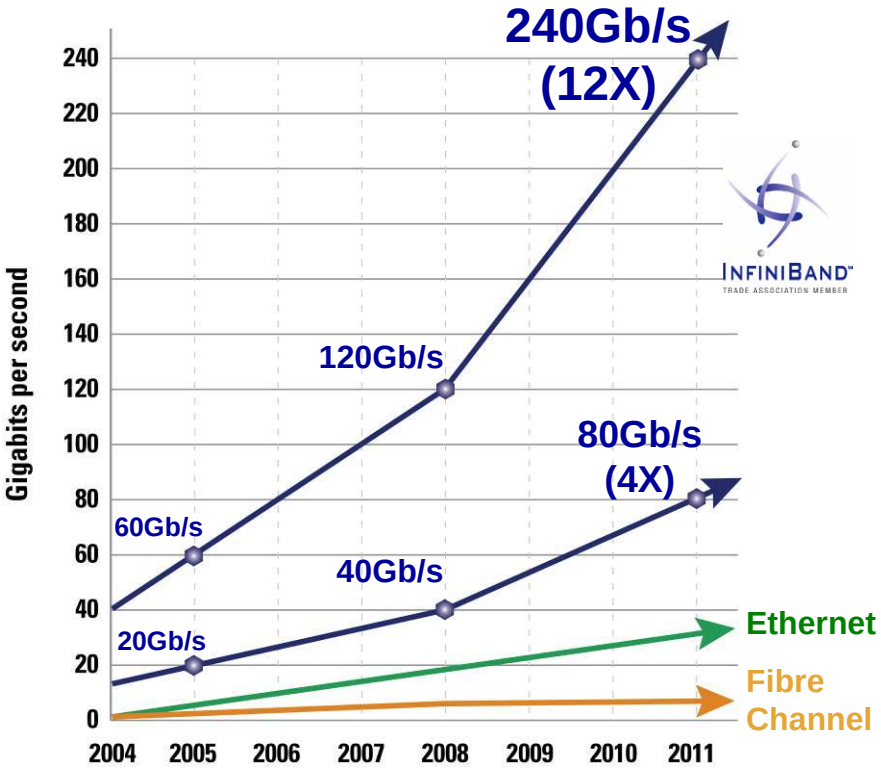
Test Cluster Configuration

- Dell™ PowerEdge™ M610 16-node cluster
- Quad-Core Intel X5570 @ 2.93 GHz CPUs
- Intel Cluster Ready certified cluster
- Mellanox ConnectX MCQH29-XCC 4X QDR InfiniBand mezzanine card
- Mellanox M3601Q 32-Port Quad Data Rate (QDR-40Gb) InfiniBand Switch
- Memory: 24GB memory per node
- OS: RHEL5U3, OFED 1.4 InfiniBand SW stack
- MPI: HP-MPI 2.3, Open MPI 1.3.2
- Application: NWChem 5.1.1
- Benchmarks: LDA calculations of three zeolite fragments (347, 1687, 3554)
  - Siosi6 ($Si_{28}O_{67}H_{30}$) and Siosi7 ($Si_{75}O_{148}H_{66}$)

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**

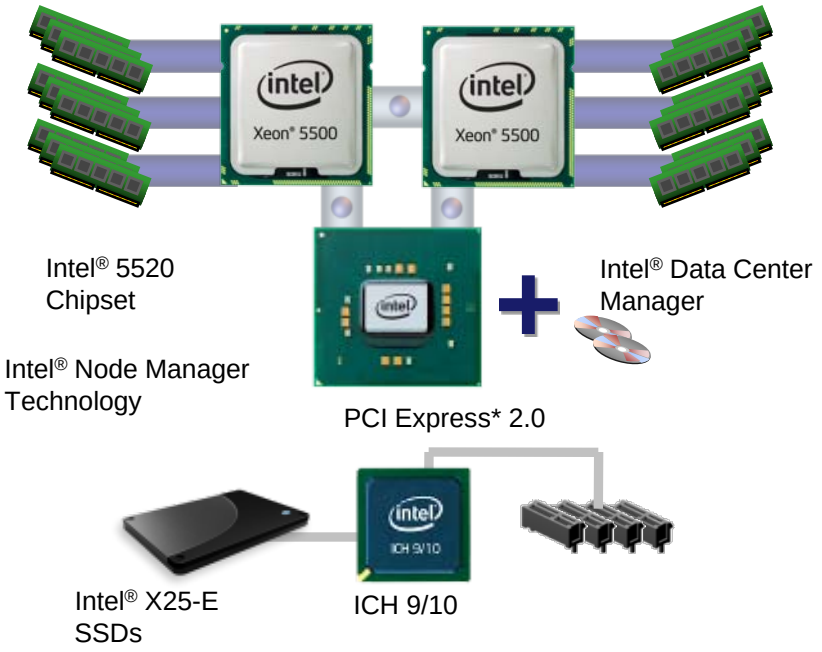
The InfiniBand Performance Gap is Increasing



InfiniBand Delivers the Lowest Latency

Delivering Intelligent Performance

Next Generation Intel® Microarchitecture



Bandwidth Intensive

- Intel® QuickPath Technology
- Integrated Memory Controller

Threaded Applications

- 45nm quad-core Intel® Xeon® Processors
- Intel® Hyper-threading Technology

Performance on Demand

- Intel® Turbo Boost Technology
- Intel® Intelligent Power Technology

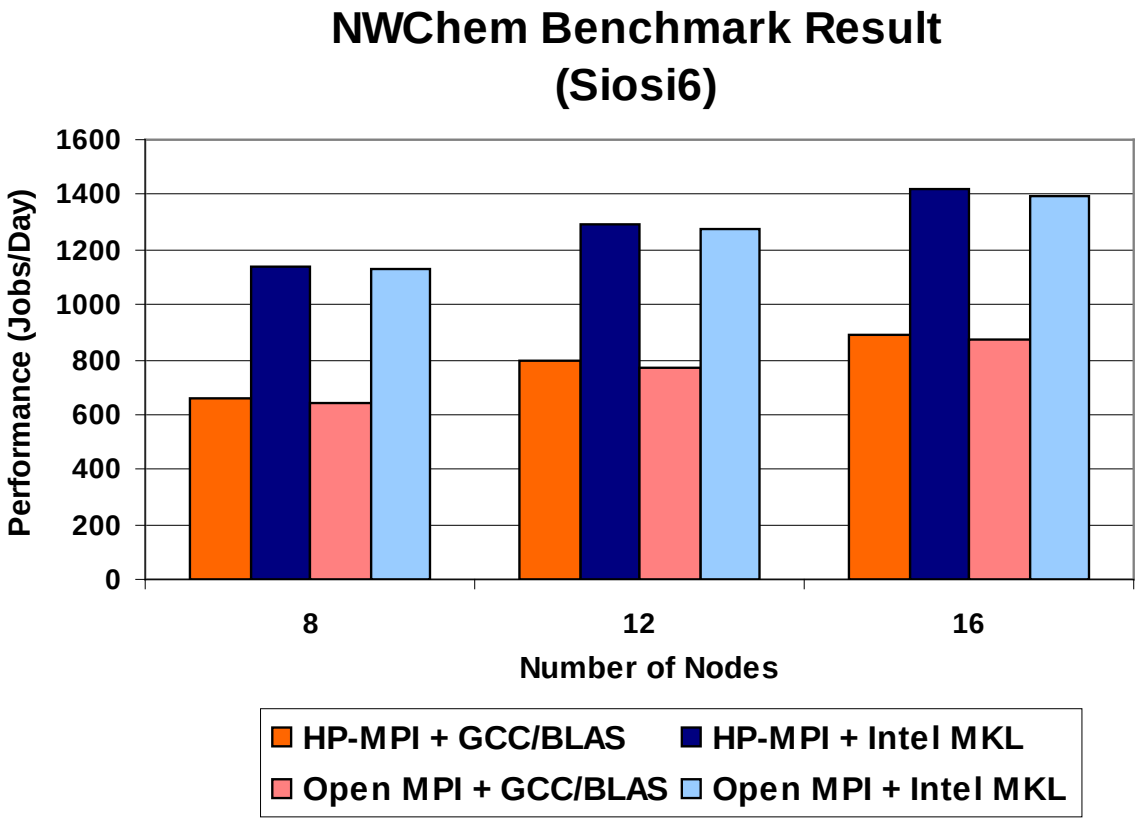
Performance That Adapts to The Software Environment

Dell PowerEdge Servers helping Simplify IT

- **System Structure and Sizing Guidelines**
  - 16-node cluster build with Dell PowerEdge™ M610 blades server
  - 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 - Siosi6**
- **Open MPI and HP-MPI provides similar performance and scalability**
  - Intel MKL library enables better performance versus GCC with BLAS

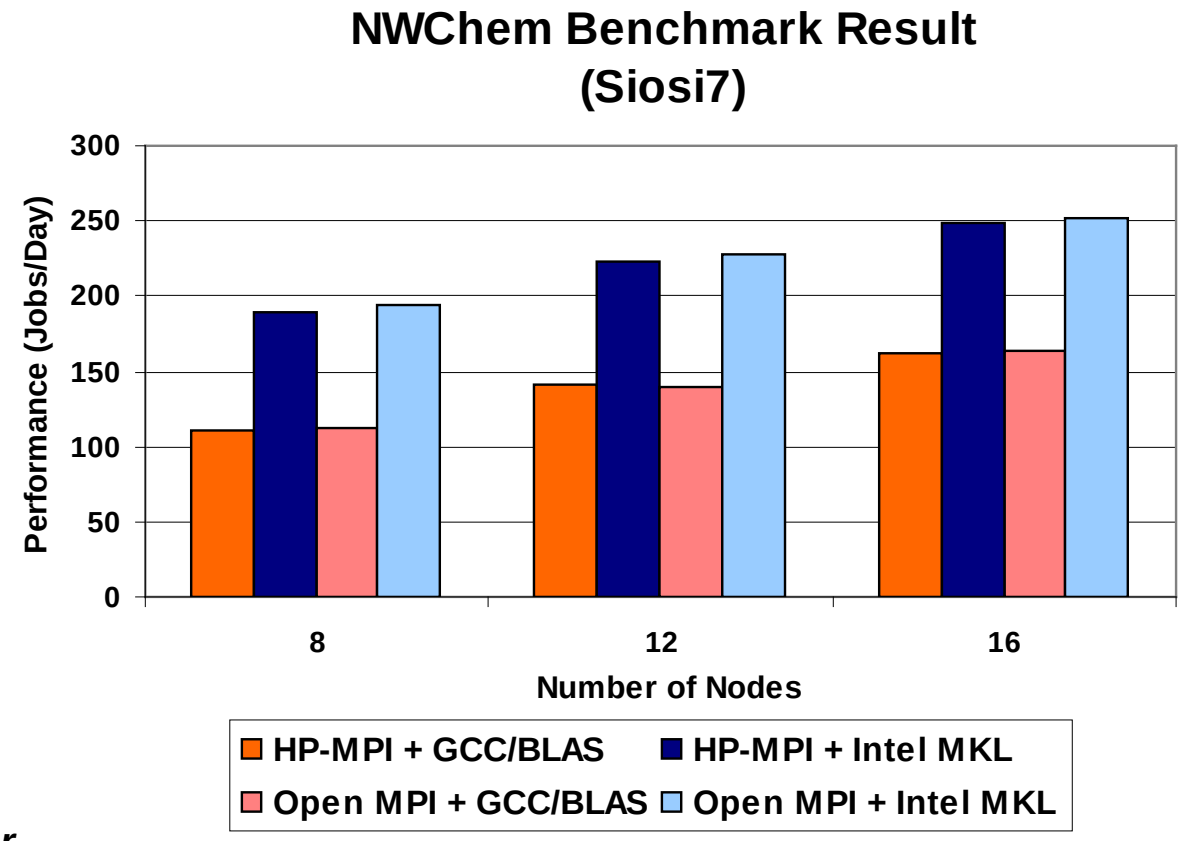


Higher is better

InfiniBand QDR

NWChem Benchmark Results

- **Input Dataset - Siosi7**
- **Open MPI and HP-MPI provides similar performance and scalability**
  - Intel MKL library enables better performance versus GCC with BLAS

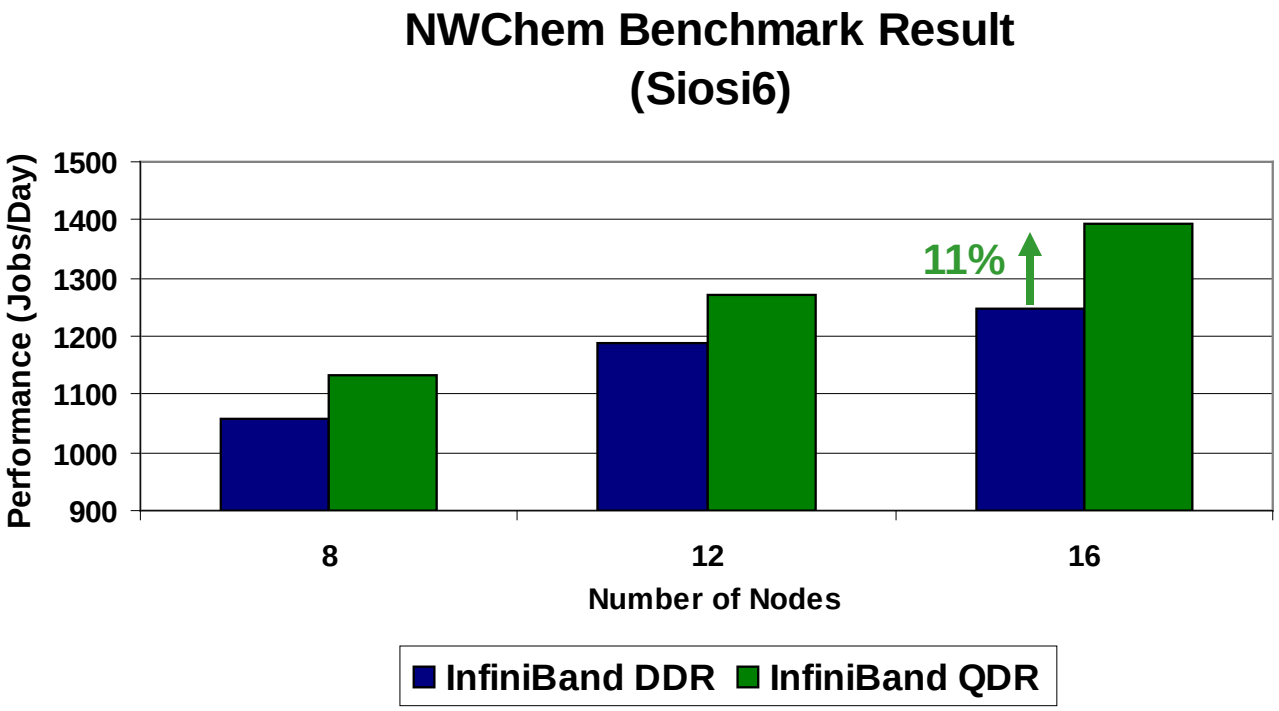


Higher is better

InfiniBand QDR

NWChem Benchmark Results

- **Input Dataset - Siosi6**
- **InfiniBand QDR enables higher performance and scalability**
  - Up to 11% high productivity than InfiniBand DDR
  - InfiniBand QDR demonstrates better scalability versus InfiniBand DDR

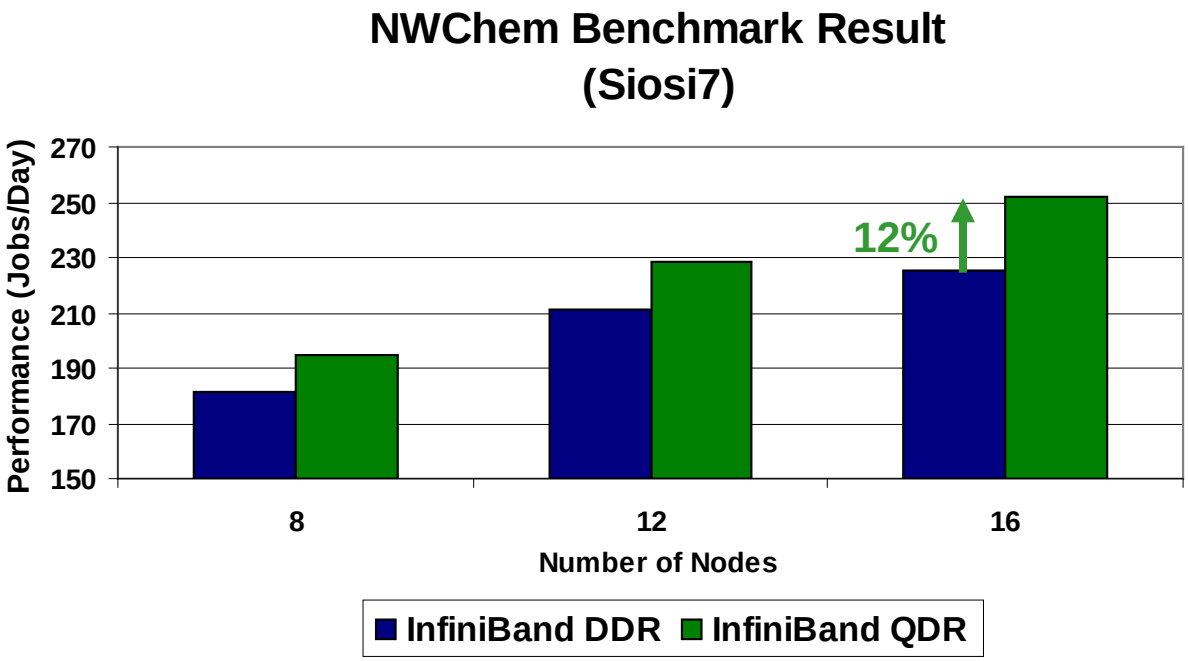


Higher is better

Open MPI + Intel MKL

NWChem Benchmark Results

- **Input Dataset - Siosi7**
- **InfiniBand QDR enables higher performance and scalability**
  - Up to 12% higher productivity versus InfiniBand DDR
  - InfiniBand QDR productivity advantage increases with system size

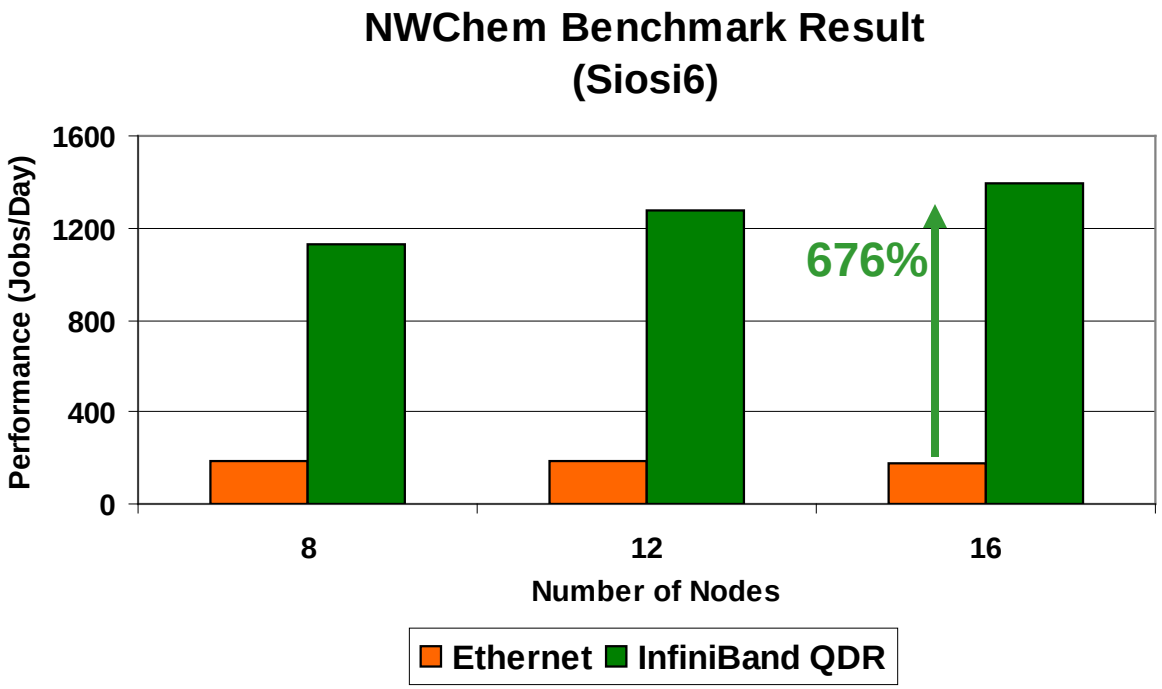


Higher is better

Open MPI + Intel MKL

NWChem Benchmark Results

- **Input Dataset - Siosi6**
- **InfiniBand provides higher performance and scalability versus GigE**
  - Up to 676% higher productivity versus Gigabit Ethernet
  - GigE does not scale beyond 8 nodes

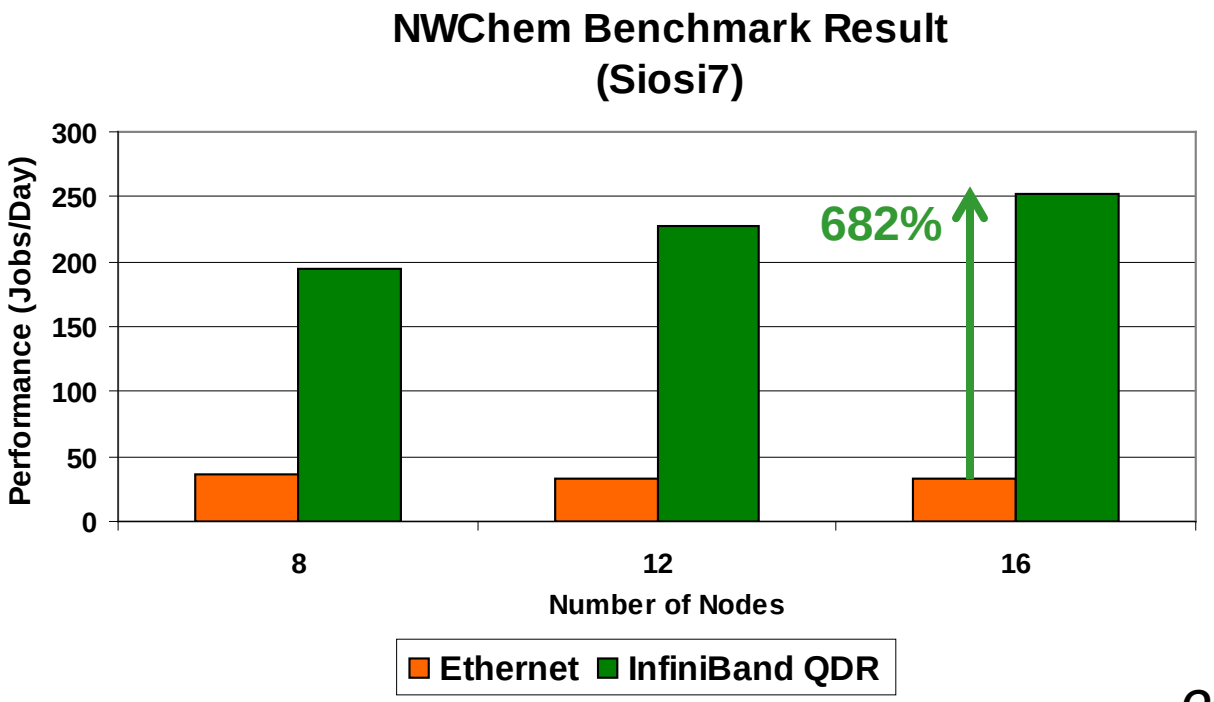


Higher is better

Open MPI + Intel MKL

NWChem Benchmark Results

- **Input Dataset - Siosi7**
- **InfiniBand provides higher performance and scalability versus GigE**
  - Up to 682% higher productivity versus Gigabit Ethernet
  - GigE shows no scalability, and even performance decrease

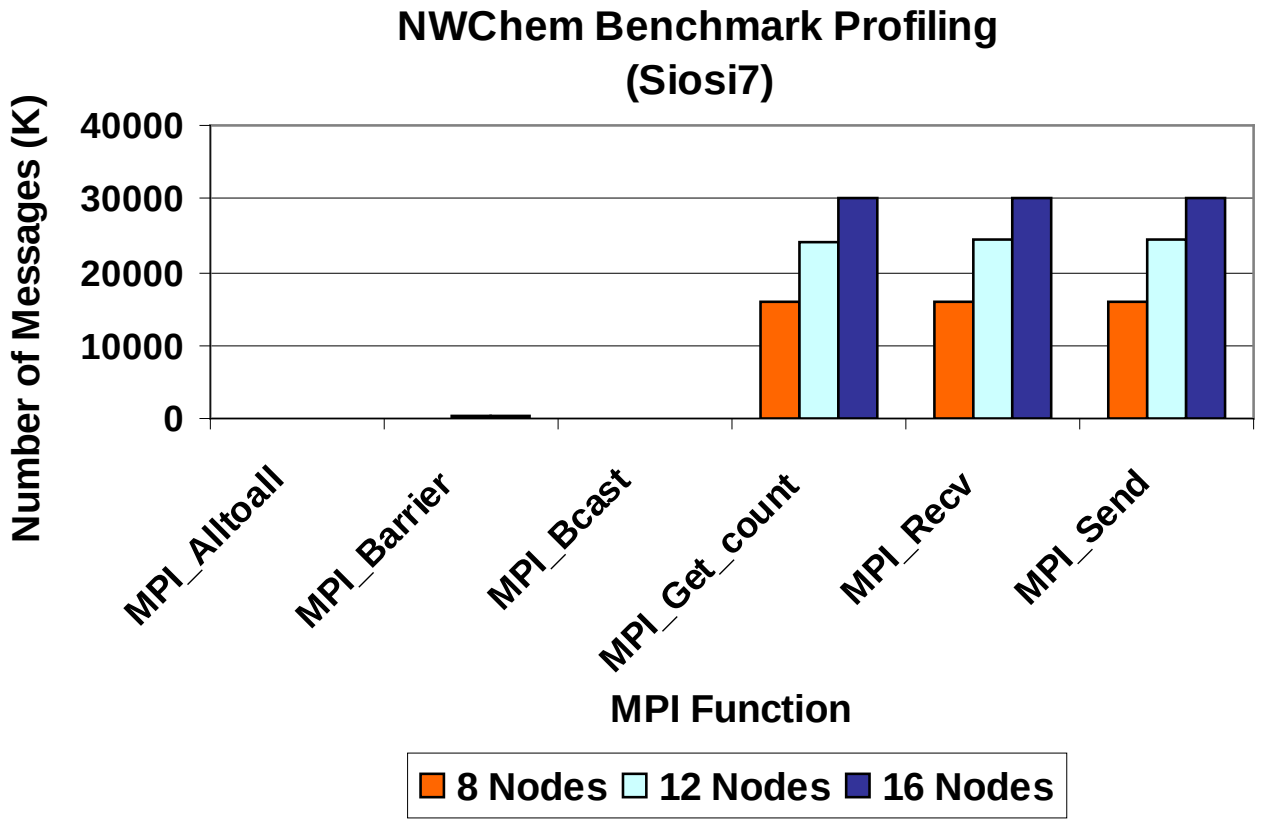


Higher is better

Open MPI + Intel MKL

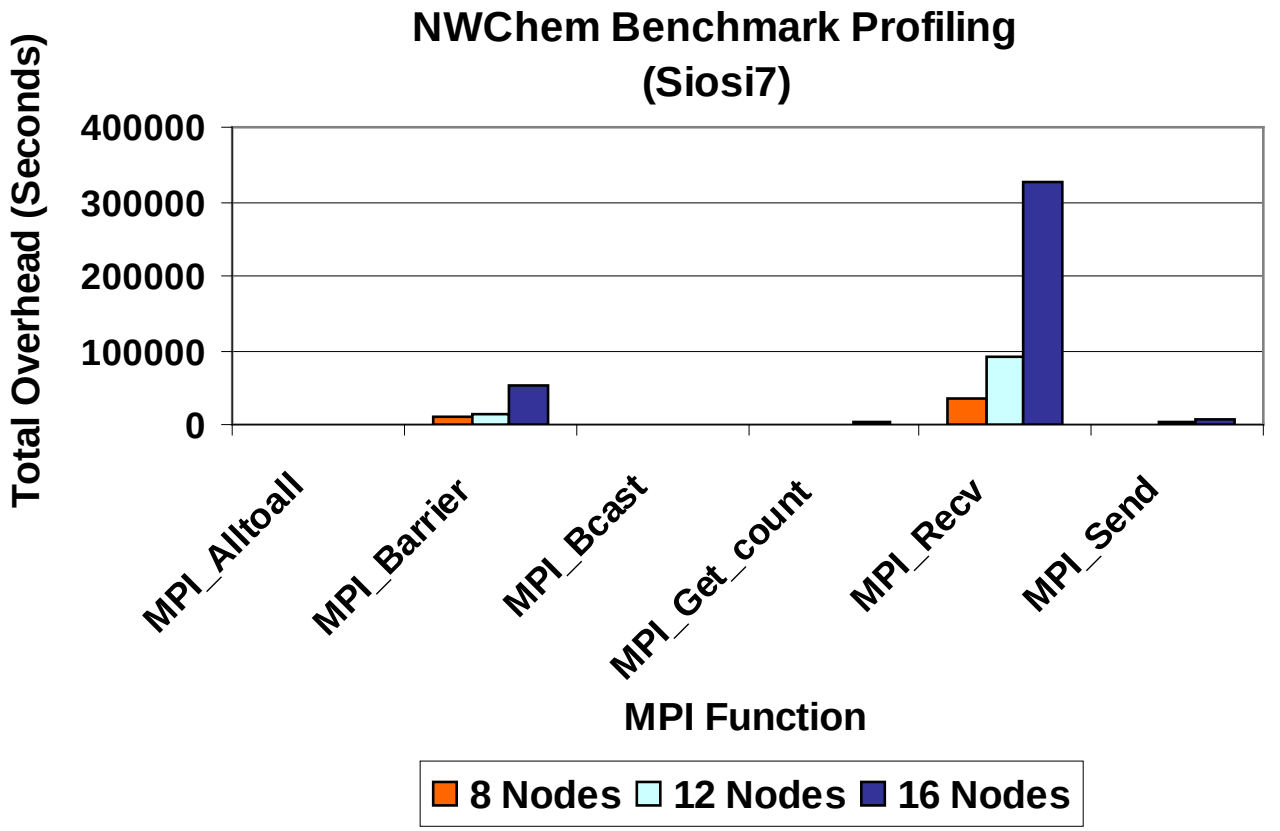
NWChem Profiling – MPI Functions

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



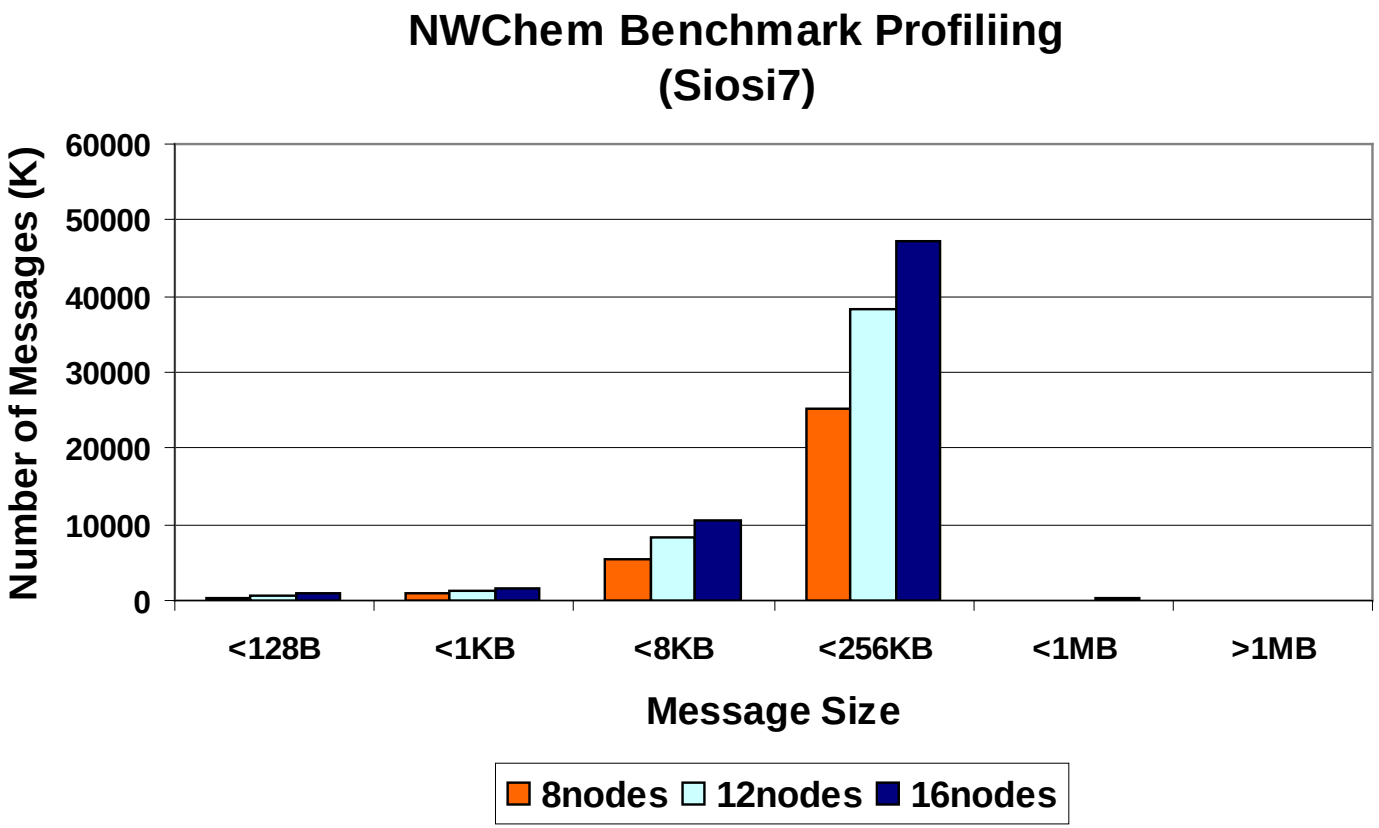
NWChem Profiling – Timing per MPI Function

- **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
- **Shows the need for highest throughput to ensure highest system utilization**



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

Test Results

- **Intel MKL library enables higher NWChem performance**
  - Up to 61% faster than GCC compiler with default BLAS library
- **HP MPI shows slight performance advantage over Open MPI**
- **InfiniBand enables highest NWChem performance and scalability**
  - Nearly 700% higher productivity versus GigE
- **InfiniBand QDR enables highest return in investment**
  - Up to 12% higher productivity versus InfiniBand DDR
    - Measured with16-node system
  - performance gain increases with system size
    - Higher performance expected with more nodes, 30% performance gain estimated for 32 nodes
- **NWChem relies on interconnect with highest throughput**
  - Most transferred messages are 1KB-256KB messages
  - Number of messages scales up as number of processes increases

Estimated System Cost

InfiniBand 40Gb/s Connected

Ethernet Connected

InfiniBand 40Gb/s Connected		Ethernet Connected
$95K (blades servers)	Cost (estimation)	$80K (blade servers)
1400 jobs/day Cost per job: $70	Productivity	200 jobs/day Cost per job: $400
16 servers provide performance equal to 16 servers capability	Utilization	16 servers provide performance equal to 2.2 servers capability
$95K - in actual compute capability $0 - wasted	Return on Investment	$11K - in actual compute capability $69K - wasted

Productive Systems = Balanced System

- **Balanced system enables highest productivity**
  - Interconnect performance to match CPU capabilities
  - CPU capabilities to drive the interconnect capability
  - Memory bandwidth to match CPU performance
- **Applications scalability relies on balanced configuration**
  - “Bottleneck free”
  - Each system components can reach it’s highest capability
- **Dell M610 system integrates balanced components**
  - Intel “Nehalem” CPUs and Mellanox InfiniBand QDR
    - Latency to memory and Interconnect latency at the same magnitude of order
  - Provide the leading productivity and power/performance system for NWchem simulations

NWChem Citation

- **E. J. Bylaska, W. A. de Jong, N. Govind, K. Kowalski, T. P. Straatsma, M. Valiev, D. Wang, E. Apra, T. L. Windus, J. Hammond, P. Nichols, S. Hirata, M. T. Hackler, Y. Zhao, P.-D. Fan, R. J. Harrison, M. Dupuis, D. M. A. Smith, J. Nieplocha, V. Tipparaju, M. Krishnan, Q. Wu, T. Van Voorhis, A. A. Auer, M. Nooijen, E. Brown, G. Cisneros, G. I. Fann, H. Fruchtl, J. Garza, K. Hirao, R. Kendall, J. A. Nichols, K. Tsemekhman, K. Wolinski, J. Anchell, D. Bernholdt, P. Borowski, T. Clark, D. Clerc, H. Dachsel, M. Deegan, K. Dyall, D. Elwood, E. Glendening, M. Gutowski, A. Hess, J. Jaffe, B. Johnson, J. Ju, R. Kobayashi, R. Kutteh, Z. Lin, R. Littlefield, X. Long, B. Meng, T. Nakajima, S. Niu, L. Pollack, M. Rosing, G. Sandrone, M. Stave, H. Taylor, G. Thomas, J. van Lenthe, A. Wong, and Z. Zhang, "NWChem, A Computational Chemistry Package for Parallel Computers, Version 5.1" (2007), Pacific Northwest National Laboratory, Richland, Washington 99352-0999, USA.**
- **"High Performance Computational Chemistry: An Overview of NWChem a Distributed Parallel Application," Kendall, R.A.; Apra, E.; Bernholdt, D.E.; Bylaska, E.J.; Dupuis, M.; Fann, G.I.; Harrison, R.J.; Ju, J.; Nichols, J.A.; Nieplocha, J.; Straatsma, T.P.; Windus, T.L.; Wong, A.T.; Computer Phys. Comm. 2000, 128, 260-283.**

Thank You

HPC Advisory Council