HPC ADVISORY COUNCIL

Dacapo Performance Benchmark and Profiling

September 2009

AMD The future is fusion

DELL

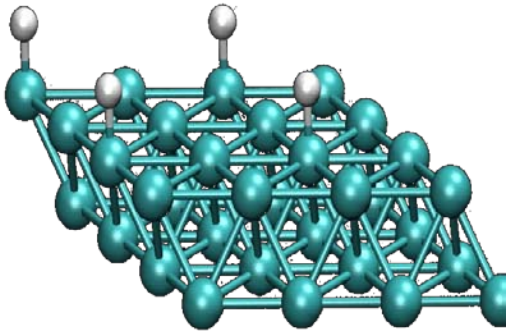
Mellanox TECHNOLOGIES

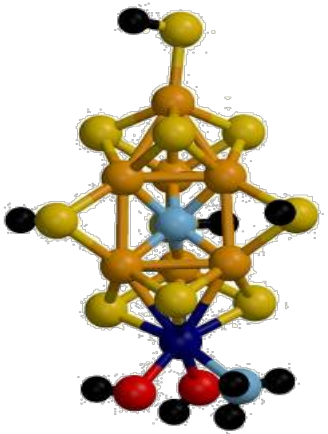
Note

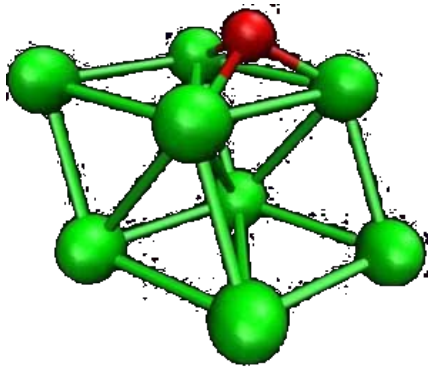
- **The following research was performed under the HPC Advisory Council activities**
  - Participating vendors: AMD, Dell, Mellanox
  - Compute resource - HPC Advisory Council Cluster Center
- **The participating members would like to thank University of Wisconsin-Madison for their guidelines**
- **For more info please refer to**
  - www.mellanox.com, www.dell.com/hpc, www.amd.com

Dacapo

- **Dacapo is a total energy program based on density functional theory**
  - Uses a plane wave basis for the valence electronic states
  - Describes core-electron interactions with Vanderbilt ultrasoft pseudo-potentials
  - Perform molecular dynamics / structural relaxation simultaneous
  - An open-source code, maintained by Technical University of Denmark







Objectives

- **The presented research was done to provide best practices**
  - Dacapo performance benchmarking
  - Interconnect performance comparisons
  - Understanding Dacapo communication patterns
  - Power-efficient simulations
- **The presented results will demonstrate**
  - The scalability of the compute environment to provide nearly linear application scalability
  - Considerations for power saving through balanced system configuration

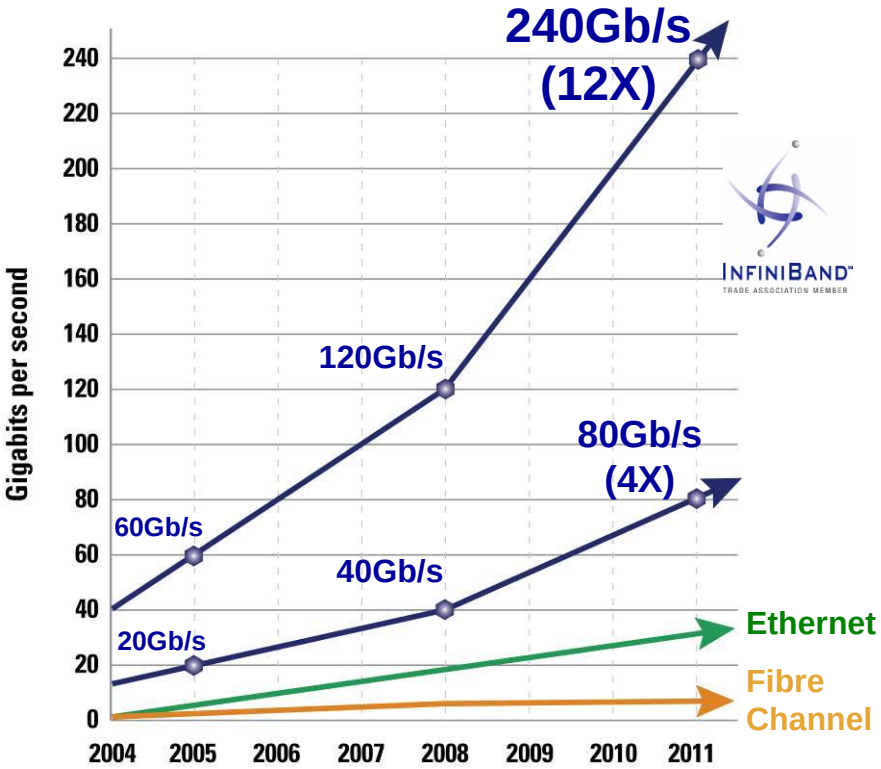
Test Cluster Configuration

- **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: RHEL5U3, OFED 1.4.1 InfiniBand SW stack, PGI compiler 8.0-6**
- **MPI: OpenMPI-1.3.3**
- **Application: Dacapo 2.7.16**
- **Benchmark Workload**
  - **In-Con_Fe110_lbrge.nc (Q**uantum mechanical calculation with 17 total 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**

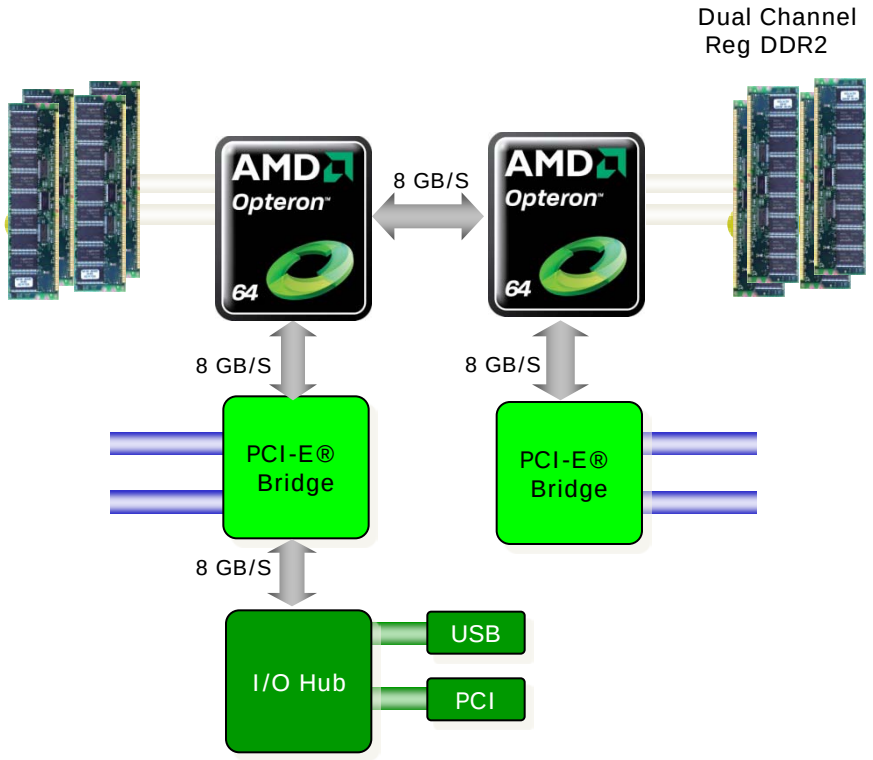
The InfiniBand Performance Gap is Increasing



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

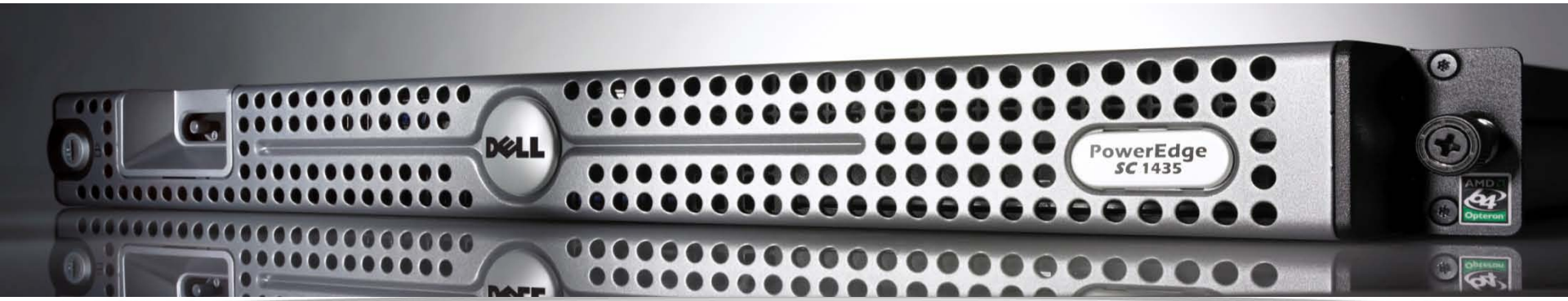


Dell PowerEdge Servers helping Simplify IT

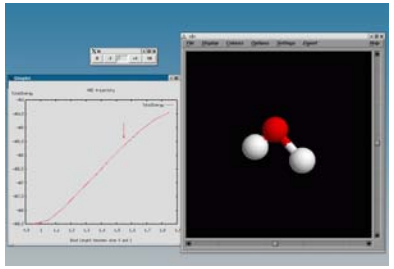
- **System Structure and Sizing Guidelines**
  - 24-node cluster build with Dell PowerEdge™ SC 1435 Servers
  - Servers optimized for High Performance Computing environments
  - Building Block Foundations for best price/performance and performance/watt
- **Dell HPC Solutions**
  - Scalable Architectures 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**



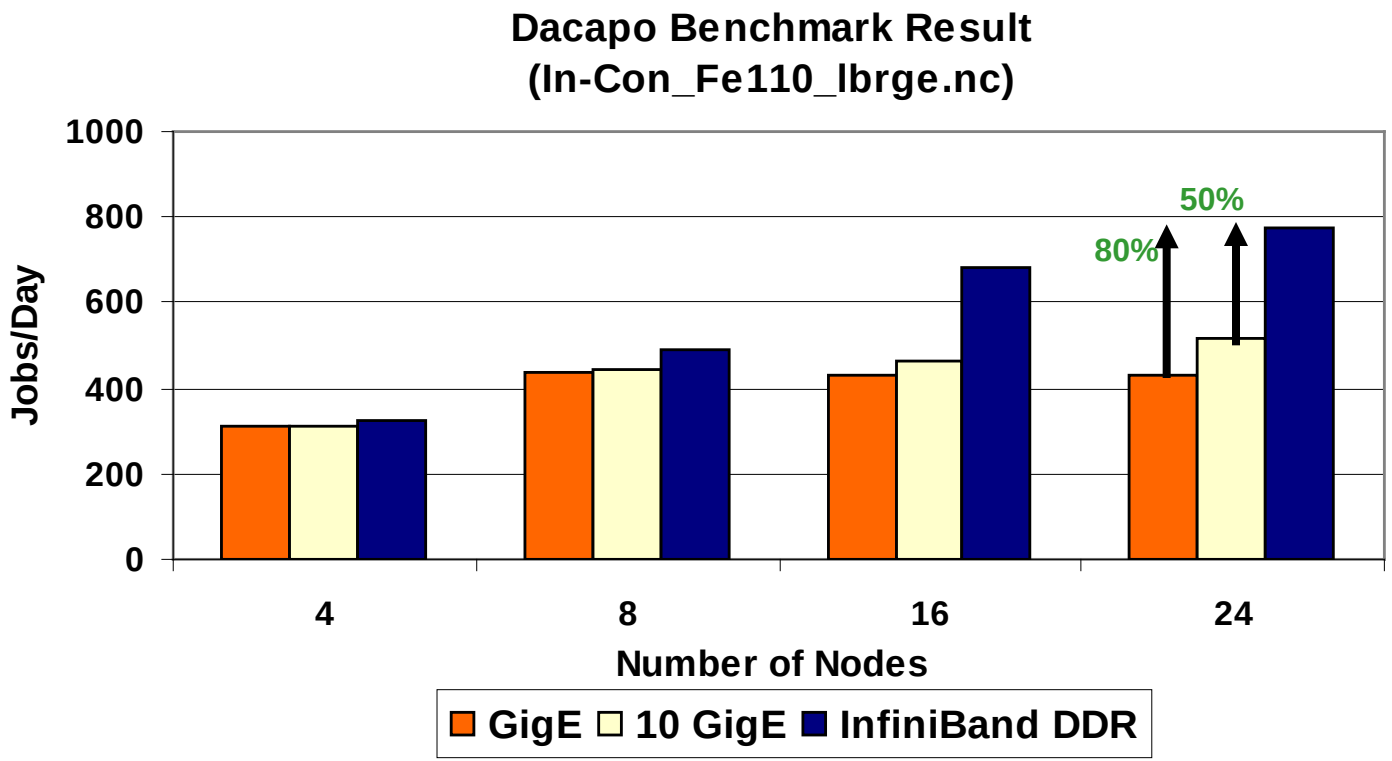
Dacapo Application Summary



- **Dacapo is a simulation software for core-electron interactions**
  - Simulation include atoms movement, energies forces calculations
- **Profiling results shows the scaling capabilities of Dacapo**
  - Good scaling was demonstrated to 24 server nodes
  - No limitations found to hold scaling beyond that size
  - AMD CPUs and memory bandwidth provide the needed capabilities for the continuous increase in performance
- **Beyond 8 server node, Dacapo requires the InfiniBand capabilities**
  - Beyond 10Gb/s bandwidth, lowest latency for MPI collectives operations
  - Networking optimizations for collectives operation, in particular AllReduce and Broadcast, expected to greatly increase performance and efficiency
    - Hardware capabilities to handle MPI collectives
- **Dacapo performance sustainability and scalability rely on reliable integration**
  - Balanced integration of the CPU-memory-interconnect-MPI libraries can affect the overall productivity and reduced power consumption per simulation job

Dacapo Benchmark Results - Interconnect

- **InfiniBand provides higher utilization, performance and scalability**
  - Up to 50% faster than 10 GigE and 80% than GigE with 24 nodes configuration
  - GigE stops scaling after 8 nodes



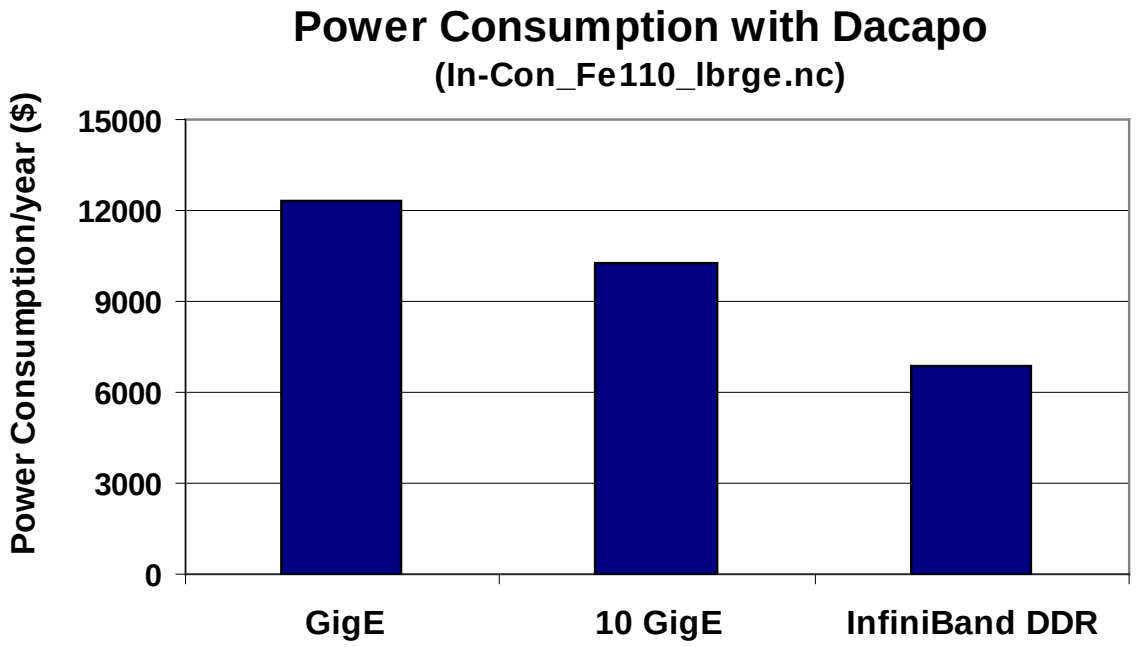
Higher is better

8-cores per node

Open MPI

Power Cost Savings with Different Interconnect

- **Dell economical integration of AMD CPUs and Mellanox InfiniBand saves up to $5500 in power**
  - To achieve same number of application jobs enabled with Gigabit Ethernet
  - Yearly based for 24-node cluster
- **As cluster size increases, more power can be saved**



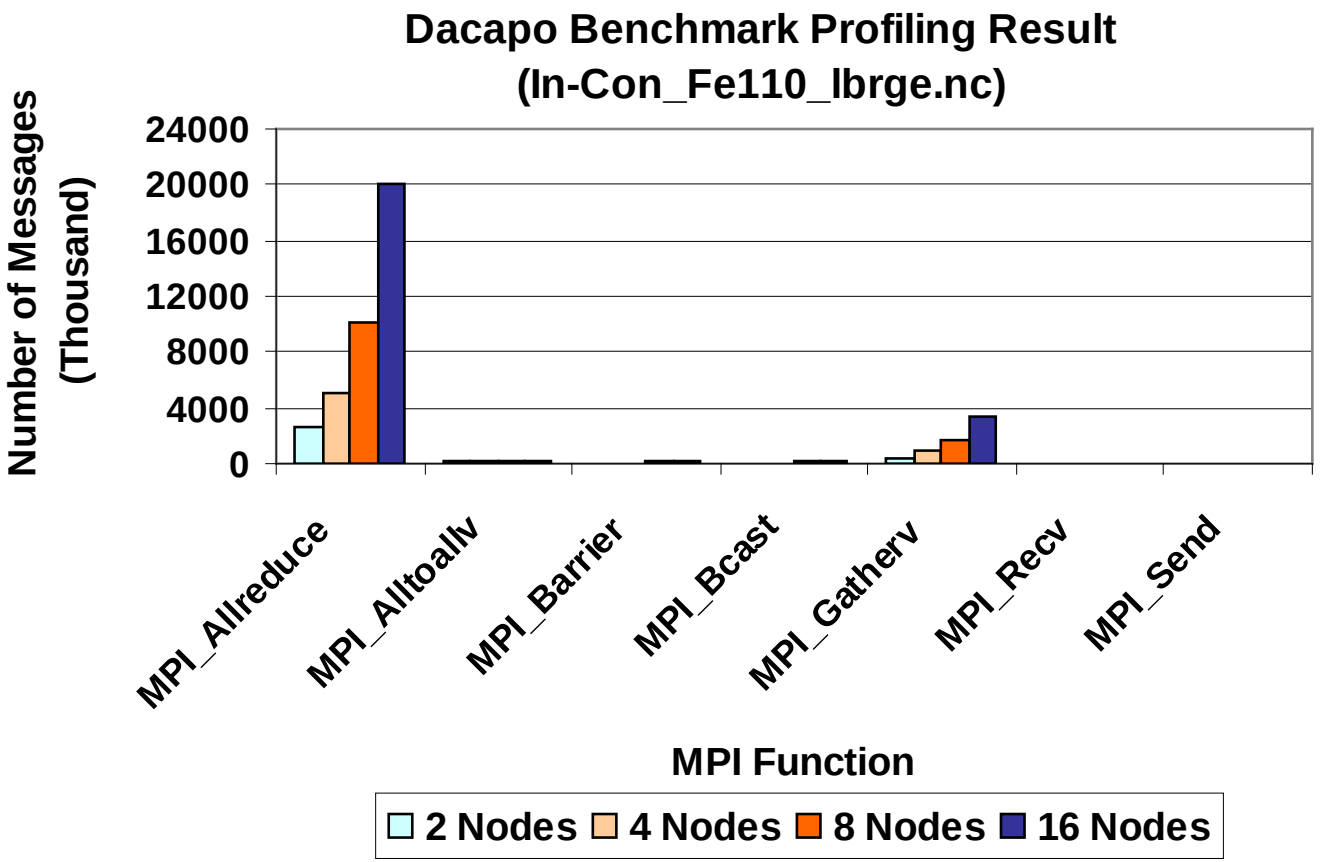
*$/KWh = KWh * $0.20*
For more information - http://enterprise.amd.com/Downloads/svrpwrusecompletefinal.pdf

Dacapo Benchmark Results Summary

- **Interconnect comparison shows**
  - InfiniBand delivers superior performance in every cluster size
  - Performance advantage extends as cluster size increases
- **Power saving**
  - InfiniBand enables up to $5500 and $3400 power savings per year versus GigE and 10 GigE
- **Dell™ PowerEdge™ server provides**
  - Linear scalability (maximum scalability) and balanced system
    - By integrating InfiniBand interconnect and AMD processors
  - Maximum return on investment through efficiency and utilization

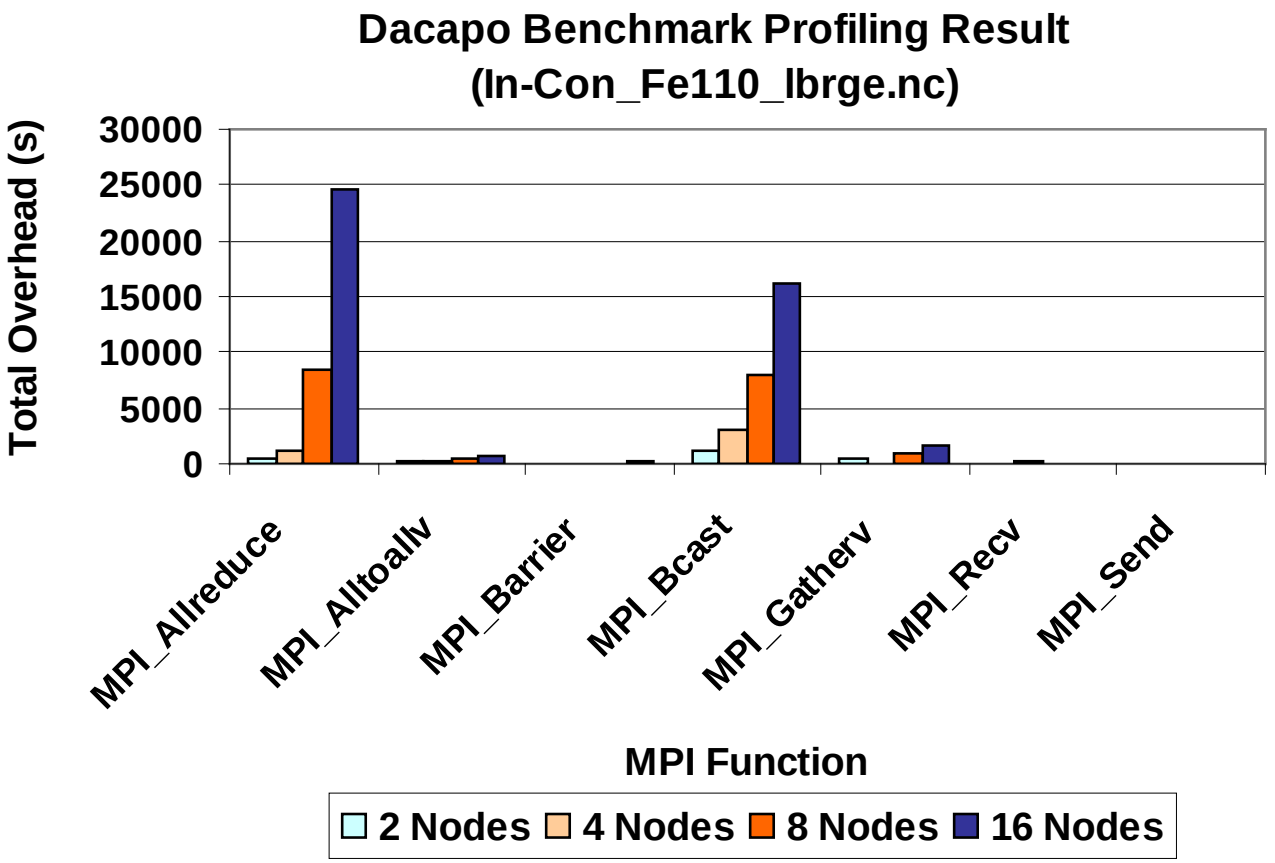
Dacapo Profiling – MPI Functions

- **Mostly used MPI functions**
  - MPI_Allreduce and MPI_AllGatherv



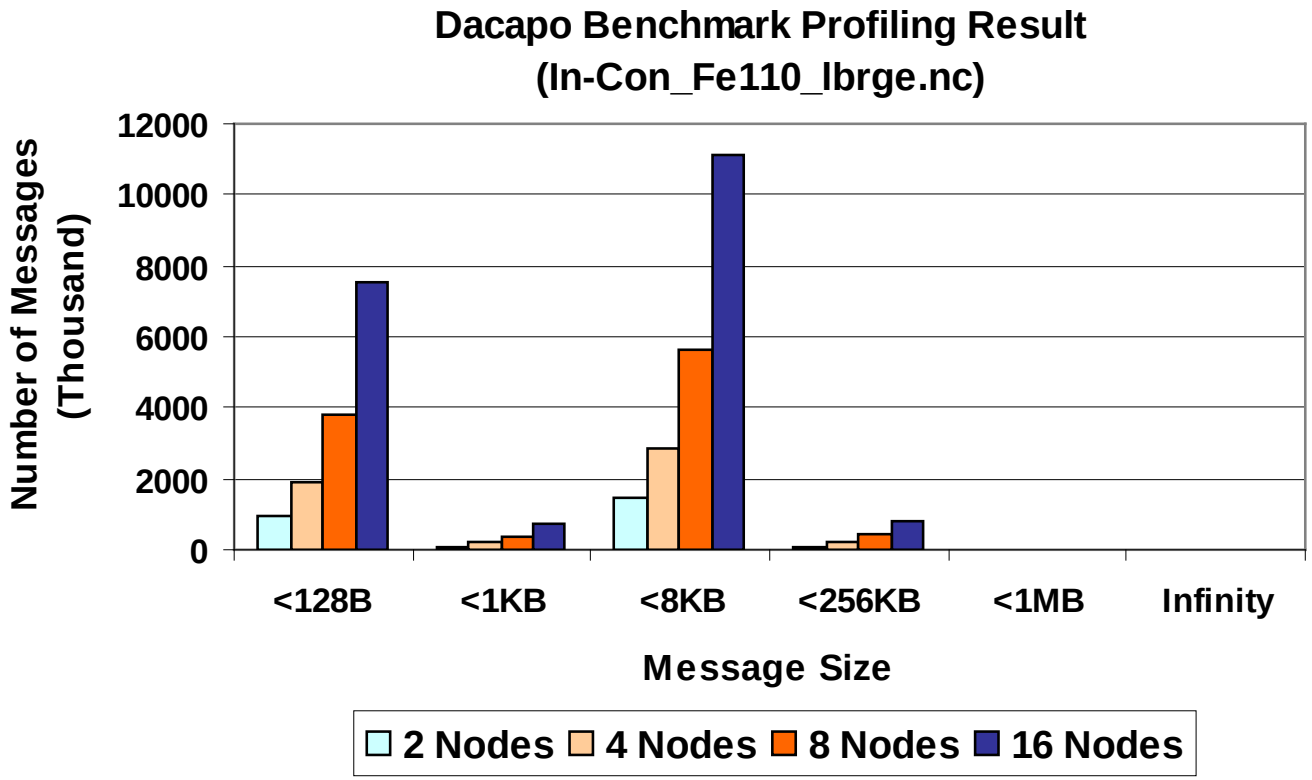
Dacapo Profiling – Timing per MPI Function

- MPI_Allreduce and MPI_Bcast show the highest communication overhead
- Time spend per MPI operation



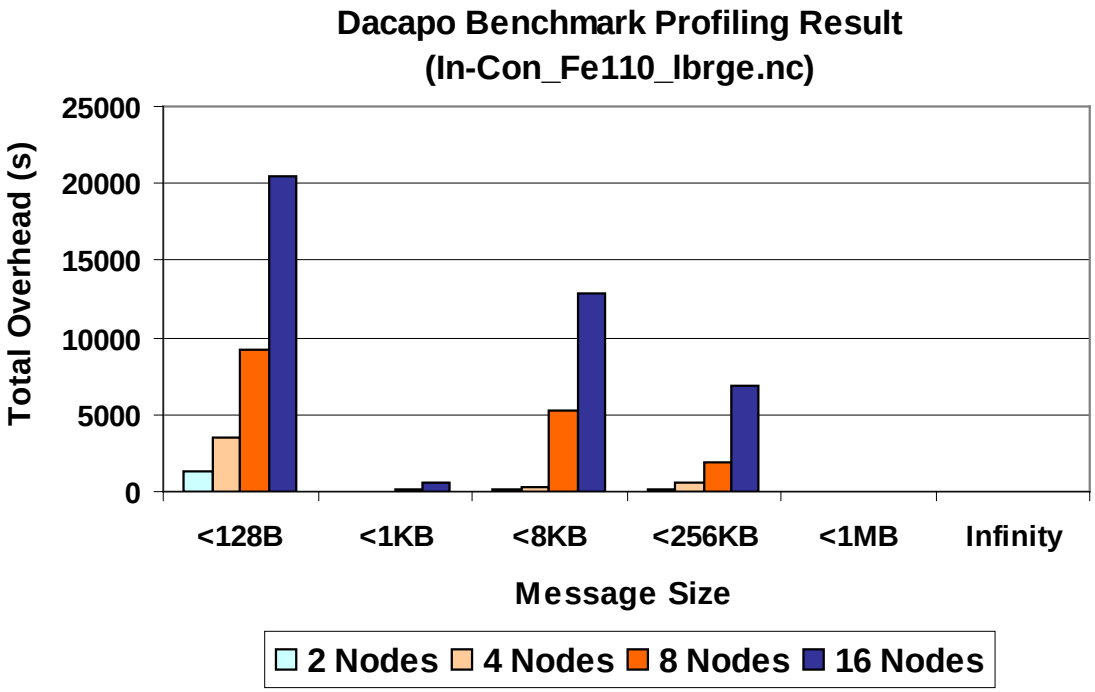
Dacapo Profiling – Message Size

- **Both small and large messages are transferred between ranks**
- **Number of messages increases with cluster size**
- **High bandwidth Interconnect (>10Gb/s starting at 8 nodes) is required**



Dacapo Profiling – Timing per Message Size

- **Small messages overhead is related to MPI collective operations**
  - Efficient or hardware based solution for such operations will greatly effect overall performance and scalability
- **As cluster size increase, Dacapo requires lowest CPU overhead for large data transfer**



Dacapo Profiling Summary

- **Dacapo were profiled to identify its communication patterns**
- **Frequent used message sizes**
  - 1KB-8KB messages for data related communications
  - <128B for synchronizations
  - Number of messages increases with cluster size
- **Interconnects effect to Dacapo performance**
  - Interconnect latency (MPI_Allreduce and MPI_Bcast) highly influence Dacapo performance
- **Balanced system – CPU, memory, Interconnect that match each other capabilities, is essential for providing application efficiency**

Thank You

HPC Advisory Council