

NAMD Performance Benchmarks and Profiling

January 2009

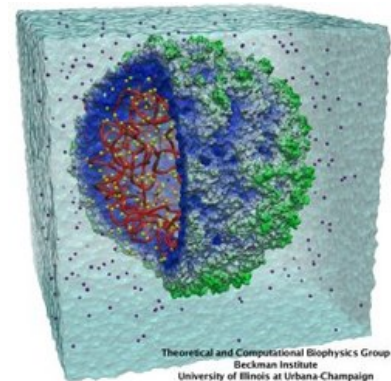
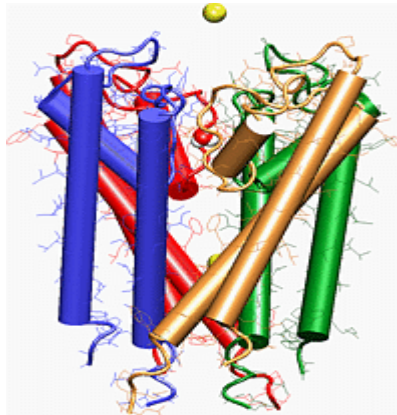


- **The following research was performed under the HPC Advisory Council activities**
 - AMD, Dell, Mellanox
 - HPC Advisory Council Cluster Center
- **For more info please refer to**
 - www.mellanox.com, www.dell.com/hpc, www.amd.com

- A parallel, object-oriented molecular dynamics software
- Designed for high-performance simulation of large biomolecular systems
 - **Millions of atoms**
- Developed by the joint collaboration of the Theoretical and Computational Biophysics Group (TCB) and the Parallel Programming Laboratory (PPL) at the University of Illinois at Urbana-Champaign
- NAMD is distributed free of charge with source code



Theoretical and Computational Biophysics Group
Beckman Institute
University of Illinois at Urbana-Champaign



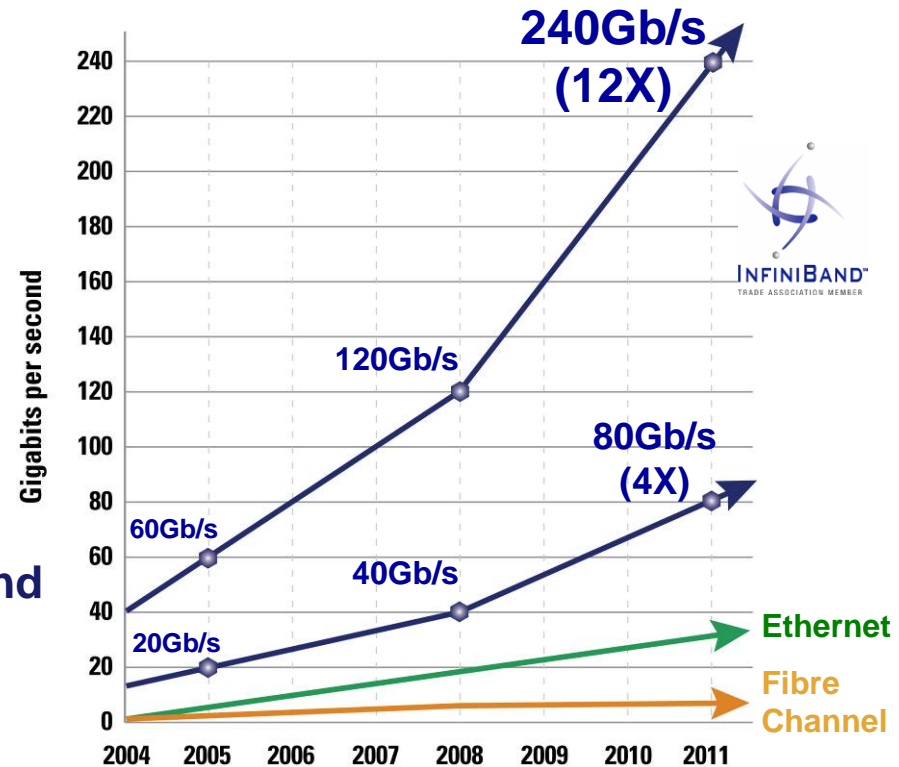
Theoretical and Computational Biophysics Group
Beckman Institute
University of Illinois at Urbana-Champaign

- **The presented research was done to provide**
 - NAMD performance benchmarking
 - Cluster Interconnect effect on NAMD performance
 - NAMD performance comparison with different MPI libraries
 - Understanding NAMD communication pattern
 - Productivity optimization

- **Dell™ PowerEdge™ SC 1435 24-node cluster**
- **Quad-Core AMD Opteron™ 2382 (code name Shanghai) CPUs**
- **Mellanox® InfiniBand ConnectX® DDR HCAs**
- **Mellanox® InfiniBand DDR Switch**
- **Memory: 16GB memory, DDR2 800MHz per node**
- **OS: RHEL5U2, OFED 1.4 InfiniBand SW stack**
- **MPI: Open MPI 1.3, Platform MPI 5.6.4**
- **Application: NAMD 2.6 with fftw3 libraries and Charm++ 6.0**
- **Benchmark Workload**
 - ApoA1 (92,224 atoms, 12A cutoff)

- **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/clock peak per core

- Integrated Memory Controller

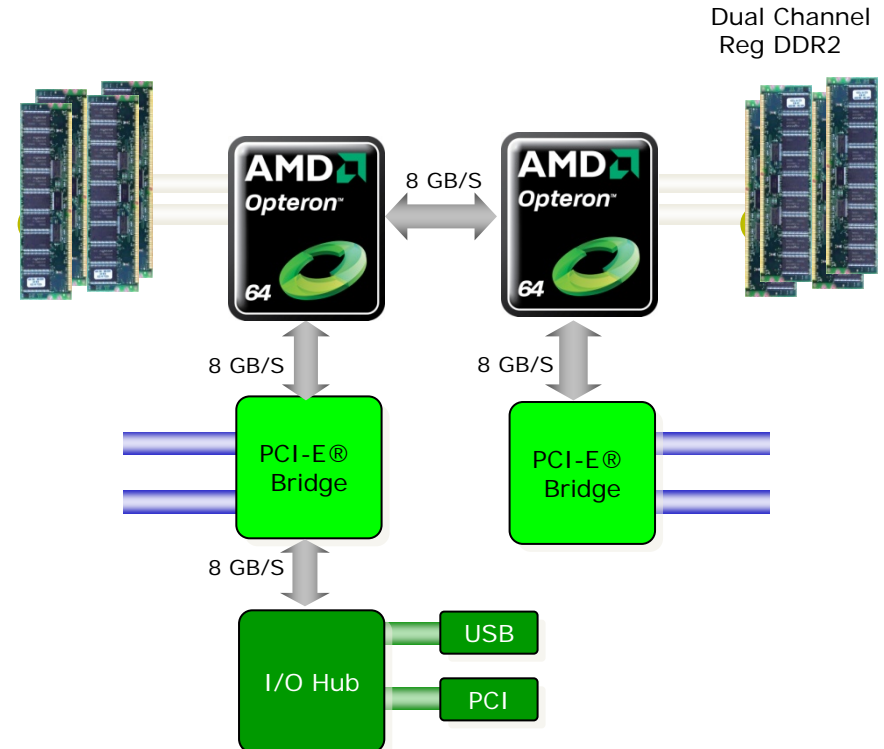
- Up to 12.8 GB/s
- DDR2-800 MHz or DDR2-667 MHz

- **Scalability**

- 48-bit Physical Addressing

- **Compatibility**

- Same power/thermal envelopes as 2nd / 3rd generation AMD Opteron™ processor



- **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

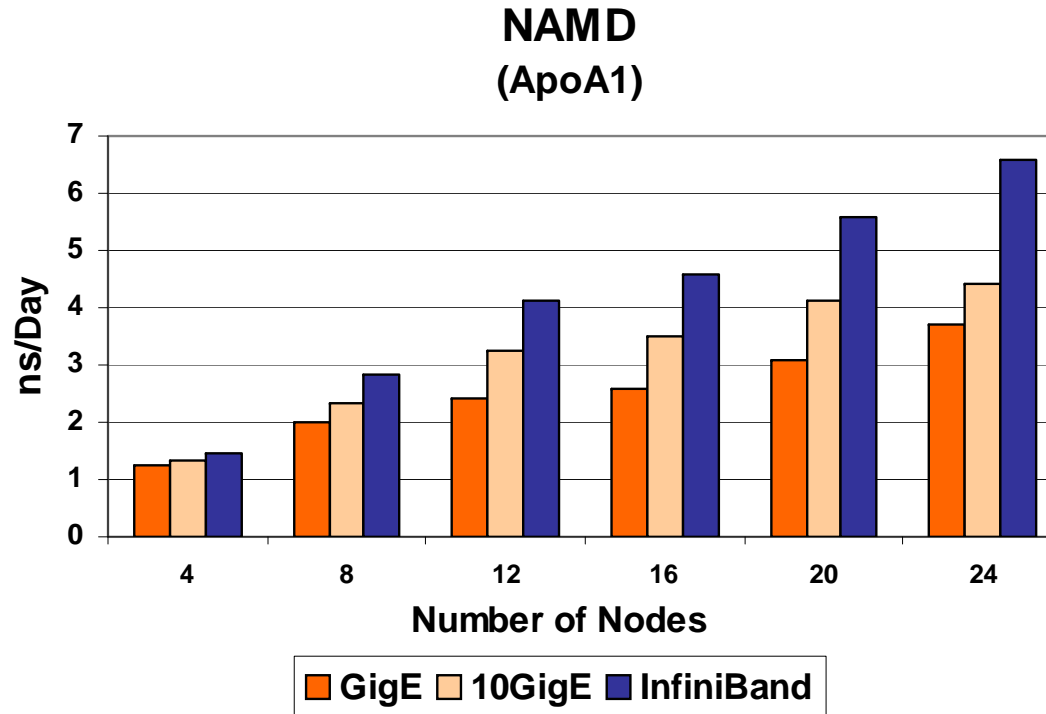
- **Workload Modeling**

- Optimized System Size, Configuration and Workloads
- Test-bed Benchmarks
- ISV Applications Characterization
- Best Practices & Usage Analysis



NAMD Results – ApoA1 Case

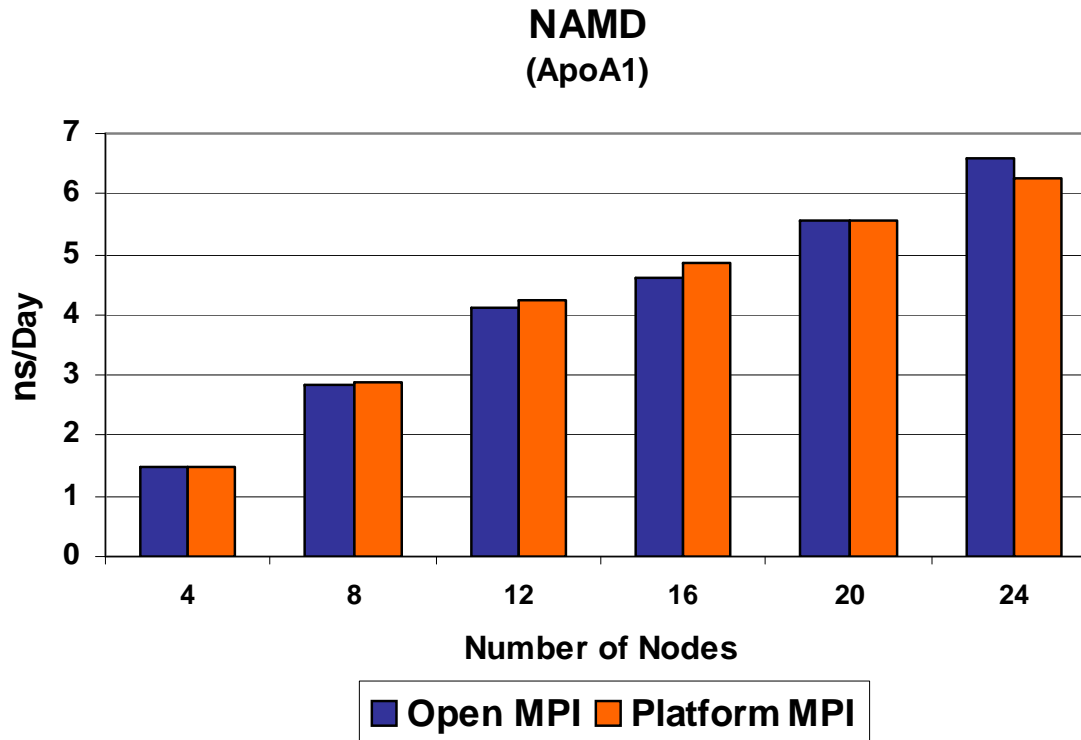
- **ApoA1 case - benchmark comprises 92K atoms of lipid, protein, and water**
 - Models a bloodstream lipoprotein particle
 - One of the most used data sets for benchmarking NAMD
- **InfiniBand 20Gb/s outperforms GigE and 10GigE in every cluster size**
 - InfiniBand provides higher performance up to 79% vs GigE and 49% vs 10GigE



Open MPI

NAMD Performance Comparison - MPI

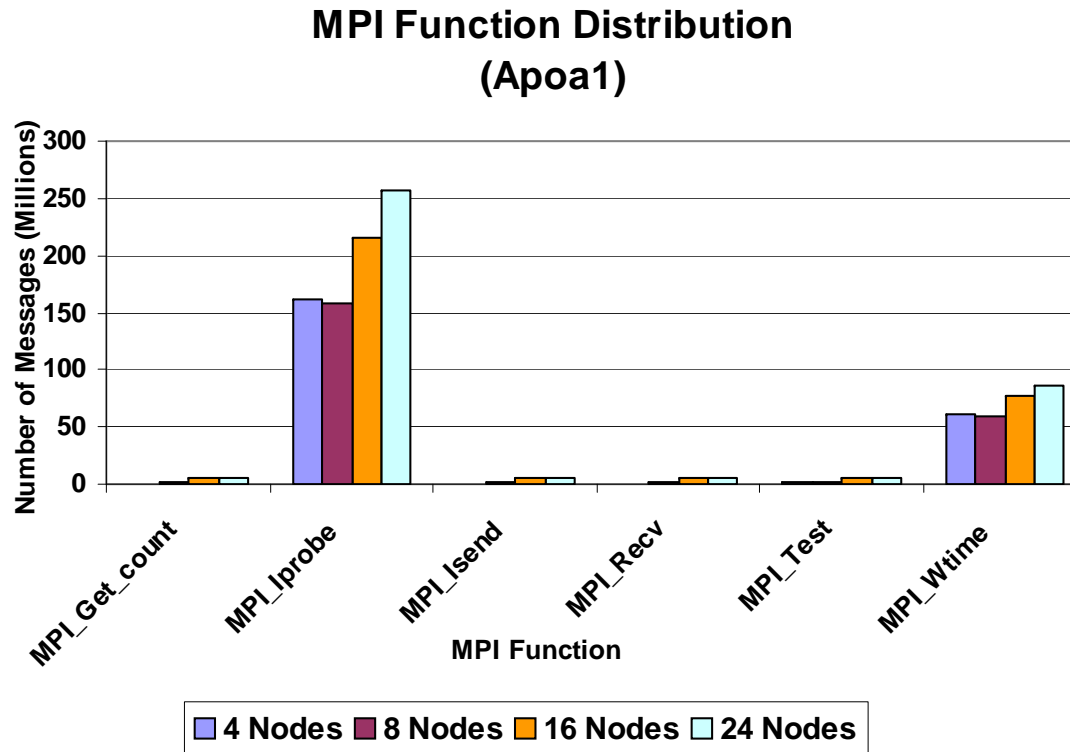
- **Platform MPI and Open MPI provides same level of performance**
 - Platform MPI has better performance for cluster size lower than 20 nodes
 - Open MPI becomes better with 24 nodes
 - Higher configurations than 24 nodes were not tested



Higher is better

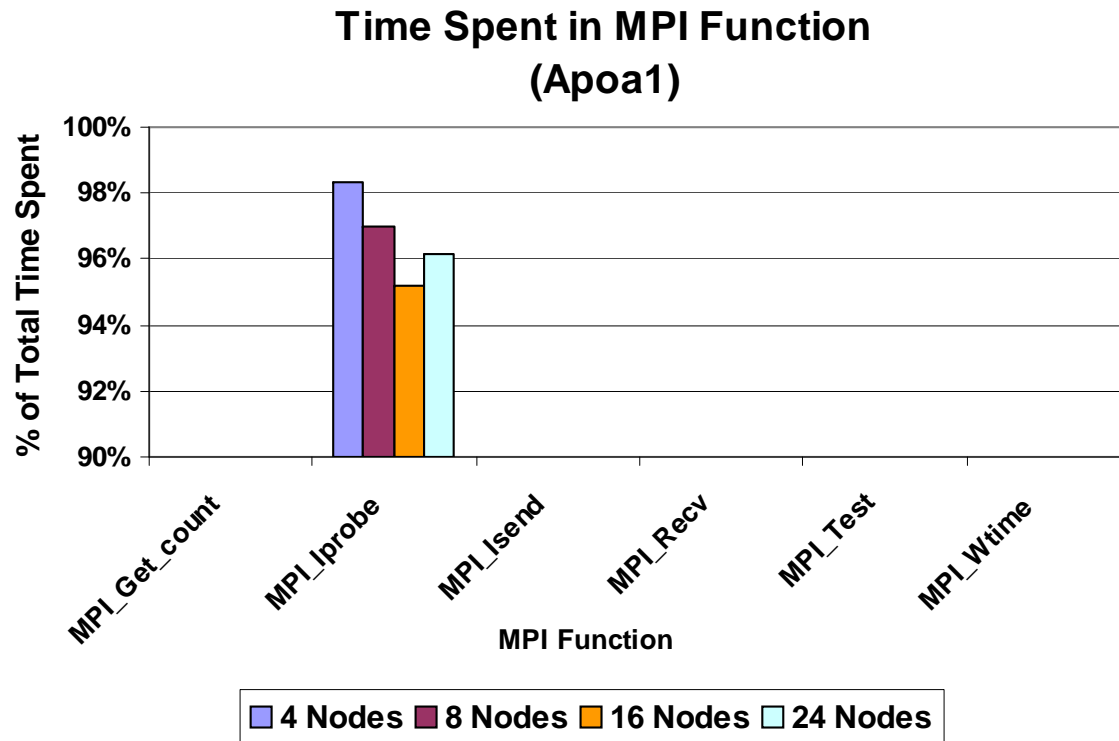
These results are based on InfiniBand

- **MPI_Iprobe** is the most used MPI function in NAMD
 - Number of MPI_Iprobe messages increases dramatically with cluster size



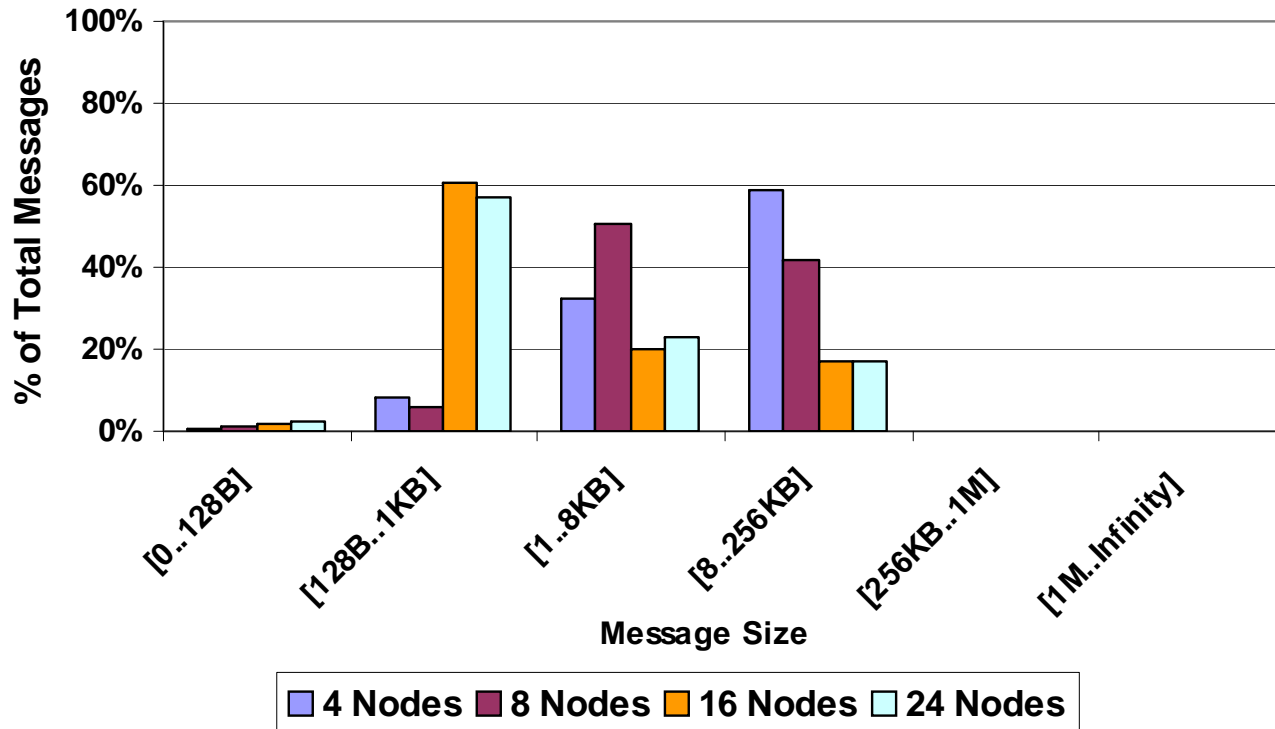
Time Spent in MPI Functions

- **Majority of communication time is spent on MPI_Iprobe**
 - Percentage are relative consistent as number of nodes increases



- As number of nodes scales, percentage of small messages increases
- Percentage of 1KB-256KB messages is relatively consistent for cluster sizes greater than 8 nodes
- Majority of the messages is in the range of 128B-1KB for cluster size greater than 8 nodes

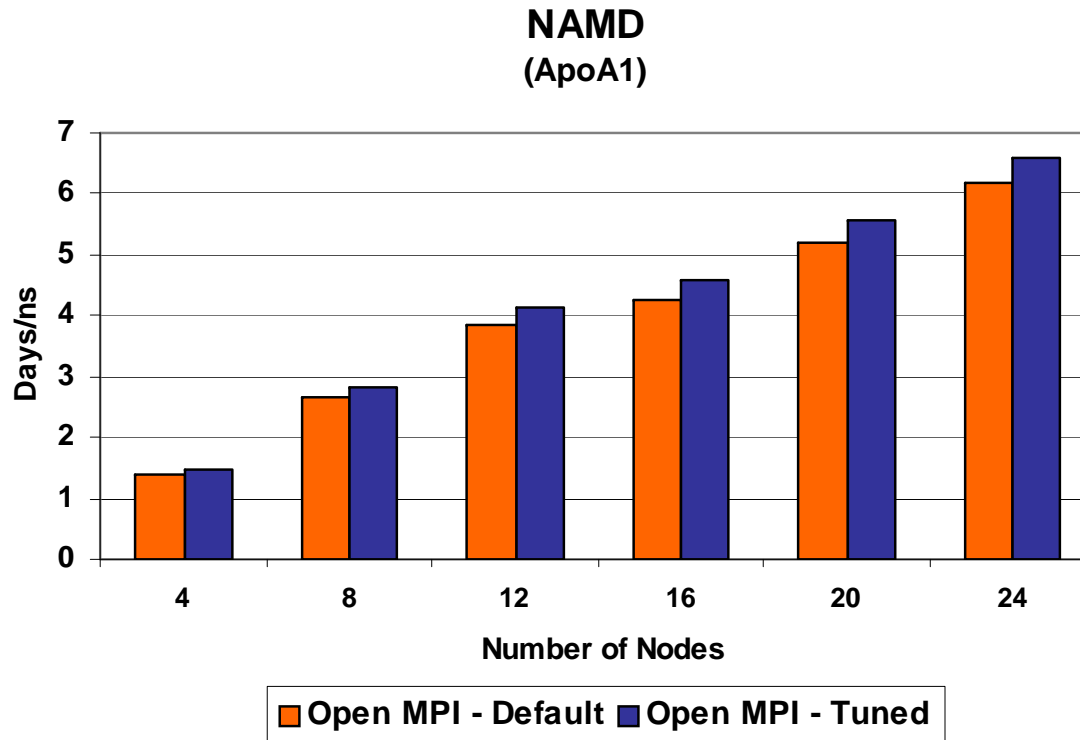
MPI Message Distribution (Apoa1)



- **NAMD was profiled to understand its communication pattern**
- **Message Size Distribution**
 - Most used message are ranging from 1KB to 8KB with node number less than 8
 - Percentage of mid size messages (128B to 1K) increases with cluster size
- **MPI function in NAMD**
 - MPI_Iprobe is the key MPI function
 - MPI_Iprobe counts up to 98% of total communication time
 - As cluster size scales, number of MPI_Iprobe messages increases dramatically
- **Performance Optimization**
 - Tuning eager message passing parameters to minimize MPI_Iprobe overhead

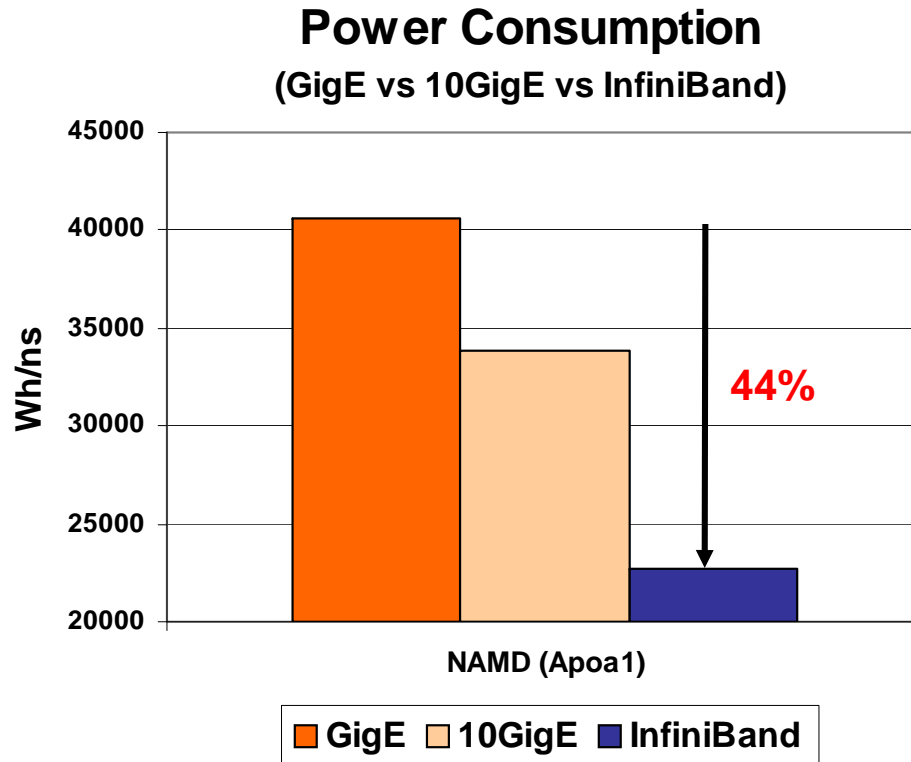
Tuning MPI For Performance Improvements

- **MPI Performance tuning**
 - Enabling CPU affinity
 - `mca mpi_paffinity_alone 1`
 - Increasing eager limit over infiniband to 32K
 - `mca btl_openib_eager_limit 32767`
- **Performance increase of up to 10%**



Higher is better

- InfiniBand enables power efficient simulations
- Reducing system power consumption per job by up to 44%



- **NAMD relies on interconnect with low latency and high throughput**
 - Most messages transferred between processes are 128Bytes - 8KB messages
 - Number of messages scales up quickly as number of processes increases
- **InfiniBand enables NAMD performance scalability**
 - InfiniBand performance is up to 79% vs GigE and 49% vs 10GigE
- **NAMD attains similar performance with Platform MPI versus Open MPI**
 - MPI_Iprobe performance affects NAMD performance
 - MPI tuning enables higher performance
- **Power Efficiency**
 - Less power consumed by finishing same amount of NAMD jobs
 - Saving in system power thus cooling

Thank You

HPC Advisory Council
HPC@mellanox.com



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