HPC ADVISORY COUNCIL

NAMD
Performance Benchmark and Profiling

February 2011

AMD The future is fusion

DELL

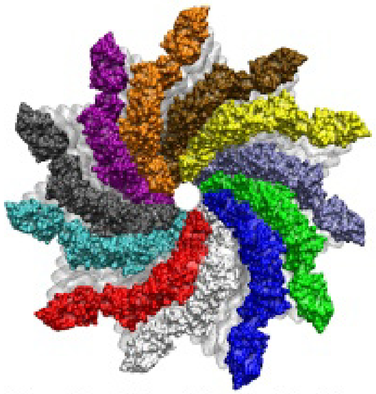
Mellanox TECHNOLOGIES

Note

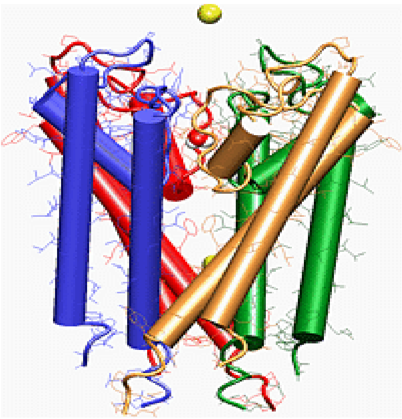
- **The following research was performed under the HPC Advisory Council activities**
  - Participating vendors: AMD, Dell, Mellanox
  - Compute resource - HPC Advisory Council Cluster Center
- **For more info please refer to**
  - http:// www.amd.com
  - http:// www.dell.com/hpc
  - http://www.mellanox.com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

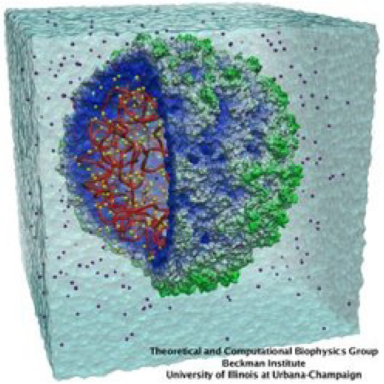
NAMD

- A parallel molecular dynamics code that received the 2002 Gordon Bell Award
- Designed for high-performance simulation of large biomolecular systems
  - **Scales to hundreds of processors and 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

Objectives

- **The following was done to provide best practices**
  - NAMD performance benchmarking
  - Interconnect performance comparisons
  - Understanding NAMD communication patterns
  - Ways to increase NAMD productivity
  - Compilers and MPI libraries comparisons

- **The presented results will demonstrate**
  - The scalability of the compute environment
  - The capability of NAMD to achieve scalable productivity
  - Considerations for performance optimizations

Test Cluster Configuration

- Dell™ PowerEdge™ R815 11-node (528-core) cluster
- AMD™ Opteron™ 6174 (code name “Magny-Cours”) 12-cores @ 2.2 GHz CPUs
- 4 CPU sockets per server node
- Mellanox ConnectX-2 VPI adapters for 40Gb/s QDR InfiniBand and 10Gb/s Ethernet
- Mellanox MTS3600Q 36-Port 40Gb/s QDR InfiniBand switch
- Fulcrum based 10Gb/s Ethernet switch
- Memory: 128GB memory per node DDR3 1333MHz
- OS: RHEL 5.5, MLNX-OFED 1.5.2 InfiniBand SW stack
- MPI: MVAPICH2-1.6RC2, Open MPI 1.4.3, Platform MPI 8.0.1
- Compilers: GNU Compilers 4.1.2
- Application: NAMD 2.7 (External libraries used: charm-6.2.2, fftw-2.1.3, TCL 8.3)
- Benchmark workload: ApoA1 bloodstream lipoprotein particle model (92,224 atoms, 12A cutoff)

Dell™ PowerEdge™ R815 11-node cluster

- **HPC Advisory Council Test-bed System**

- **New 11-node 528 core cluster - featuring Dell PowerEdge™ R815 servers**
  - Replacement system for Dell PowerEdge SC1435 (192 cores) cluster system following 2 years of rigorous benchmarking and product EOL
    - System to be redirected to explore HPC in the Cloud applications

- **Workload profiling and benchmarking**
  - Characterization for HPC and compute intense environments
  - Optimization for scale, sizing and configuration and workload performance
  - Test-bed Benchmarks
    - RFPs
    - Customers/Prospects, etc
  - ISV & Industry standard application characterization
  - Best practices & usage analysis

About Dell PowerEdge™ Platform Advantages

Best of breed technologies and partners

Combination of AMD™ Opteron™ 6100 series platform and Mellanox ConnectX InfiniBand on Dell HPC

Solutions provide the ultimate platform for speed and scale

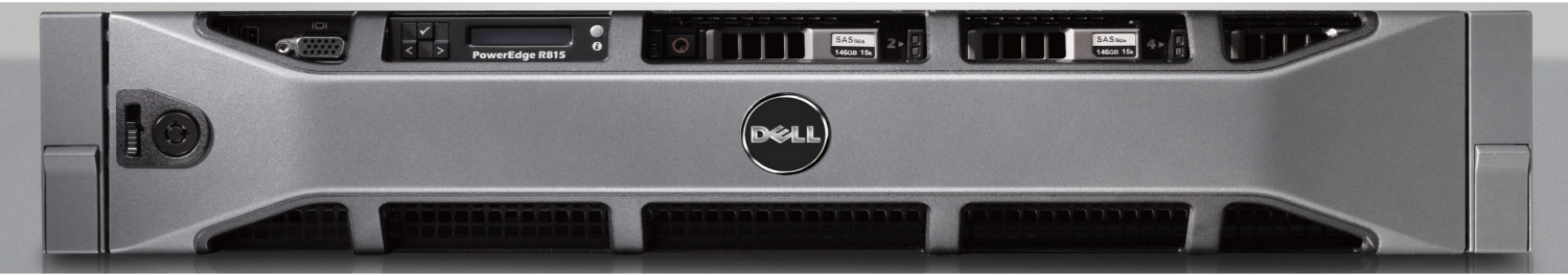
- Dell PowerEdge R815 system delivers 4 socket performance in dense 2U form factor
- Up to 48 core/32DIMMs per server – 1008 core in 42U enclosure

Integrated stacks designed to deliver the best price/performance/watt

- 2x more memory and processing power in half of the space
- Energy optimized low flow fans, improved power supplies and dual SD modules

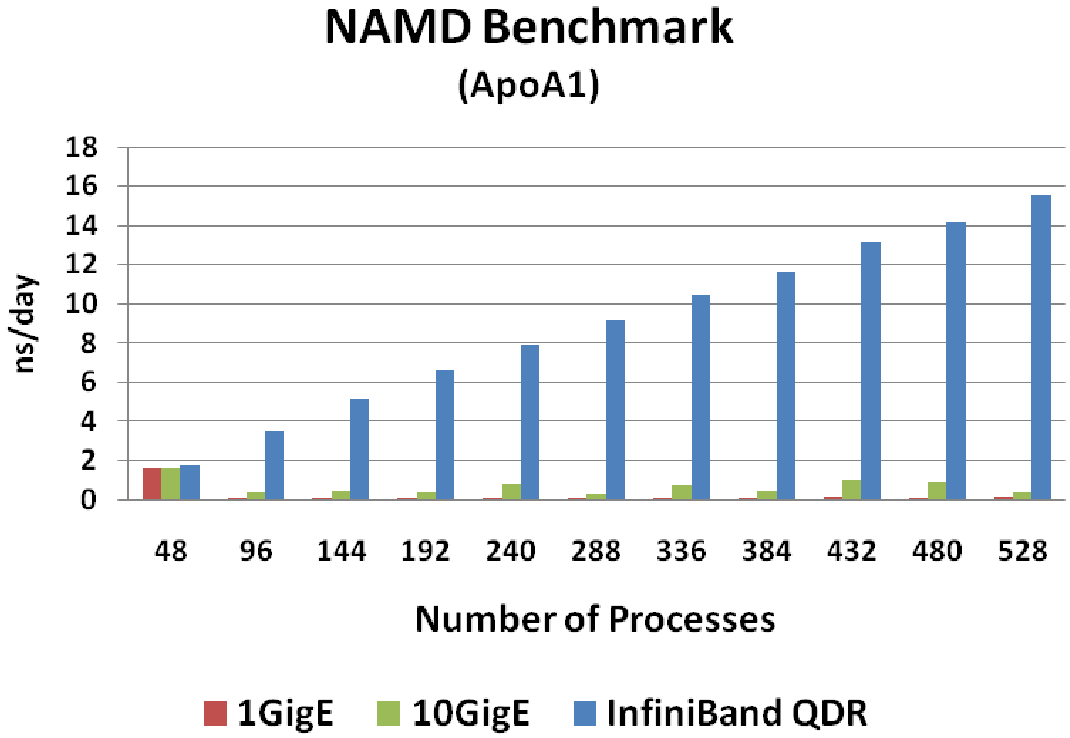
Optimized for long-term capital and operating investment protection

- System expansion
- Component upgrades and feature releases



NAMD Performance – Interconnects

- **InfiniBand shows continuous gain as the cluster scales**
- **Ethernet performance does not scale beyond 48 cores**

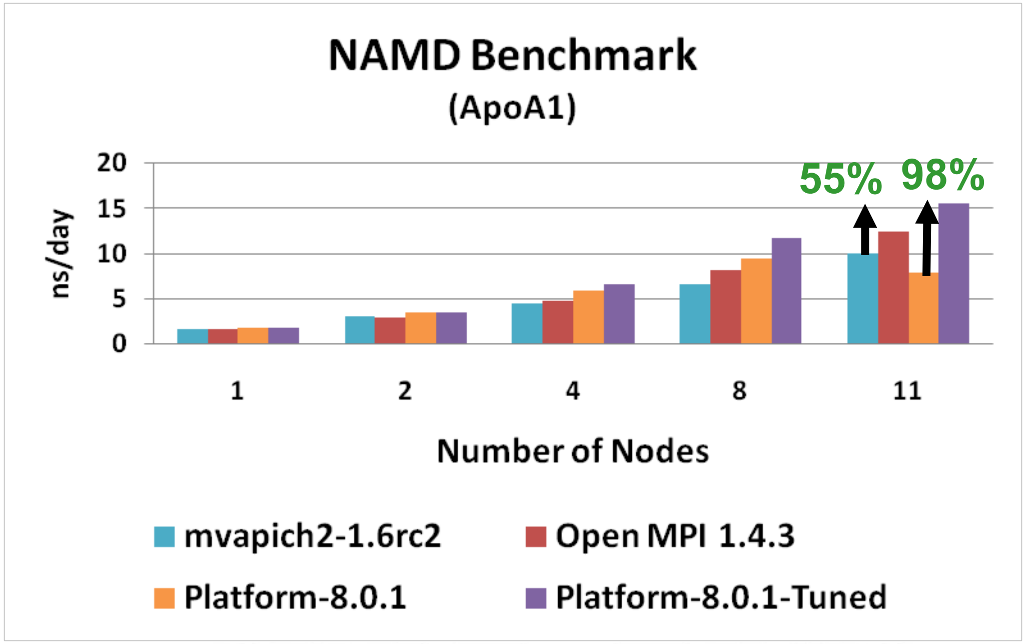


Higher is better

48 Cores/Node

NAMD Performance – MPI Implementations

- **Tuned Platform MPI performs the best**
  - Up to 55% faster than MVAPICH2 at 528 processes
  - Up to 98% improvement over the un-tuned version
    - Un-tuned version hit by performance limitation after 512-core
  - Tuned RDMA message sizes, Shared Receive Queue and related env-vars:
    - -srq -IBV -aff=automatic -e MPI_RDMA_MSGSIZE=16384,16384,4194304 -e MPI_RDMA_NSRQRECV=2048 -e MPI_RDMA_NFRAGMENT=128

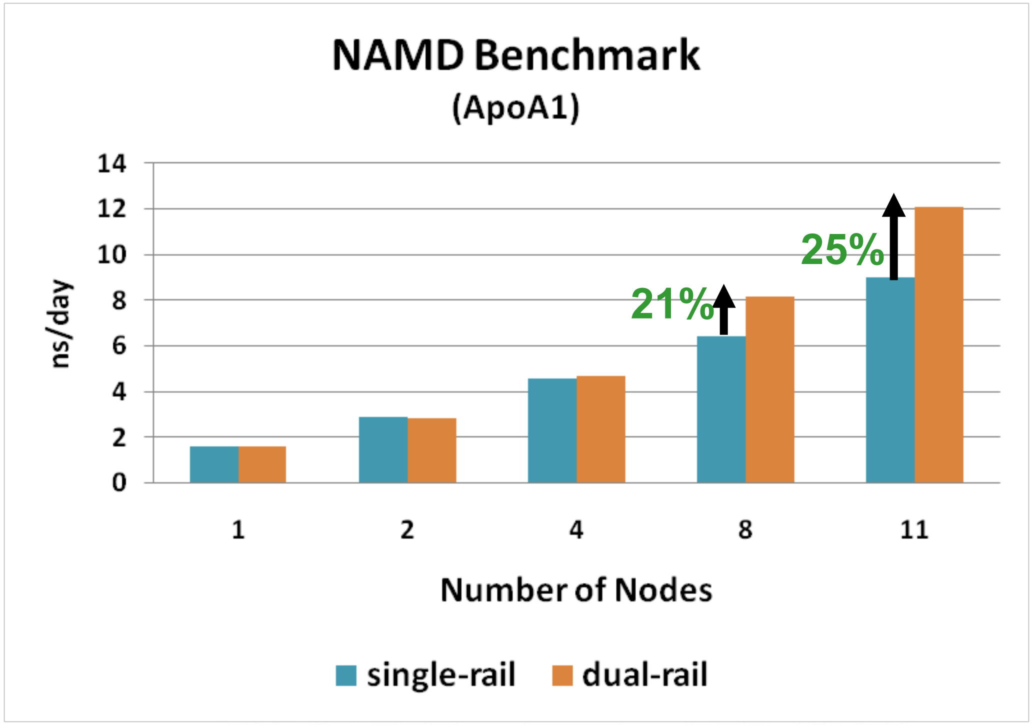


Higher is better

48 Cores/Node

NAMD Performance – InfiniBand Multi-rail

- **Dual-rail (Dual InfiniBand cards) enables better performance than single-rail**
  - Up to 25% better at 11-node
- **The benefit of dual-rail starts to emerge at 8-node**
  - As message profiling shows the volume of messages begins to increase
- **Dual-rail enables round-robin of small messages on the 2 InfiniBand ports**



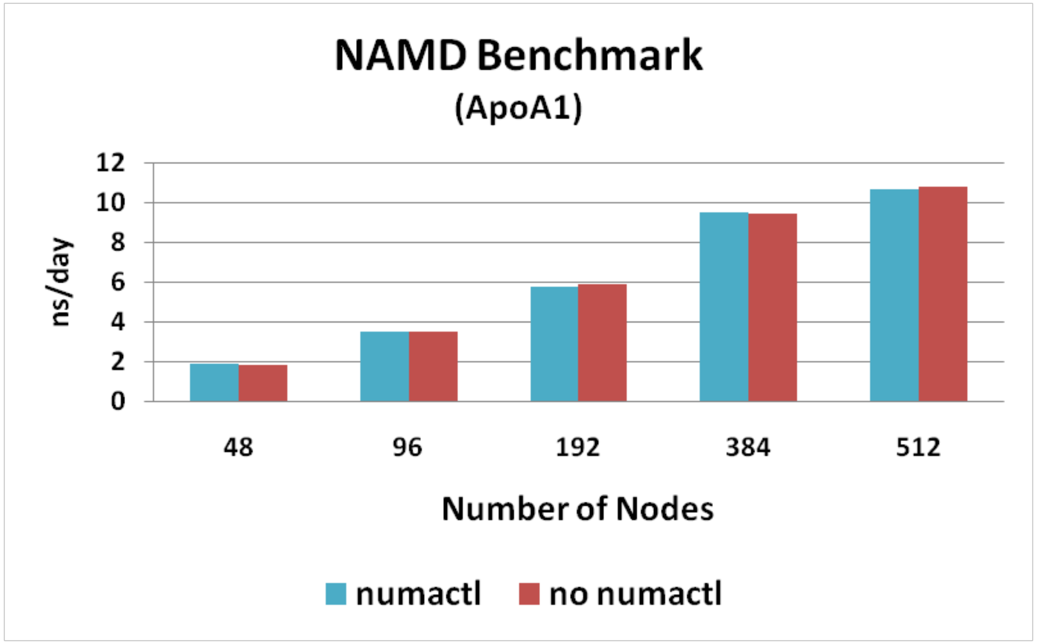
Higher is better

Open MPI

48 Cores/Node

NAMD Performance – Numactl

- **NUMA**
  - Stands for **Non-Uniform Memory Architecture**
  - Memory access depends on memory location relative to a processor
- **Numactl allows assigning processes to CPU node with local memory**
  - Results show no difference in job performance when assigning memory assignment

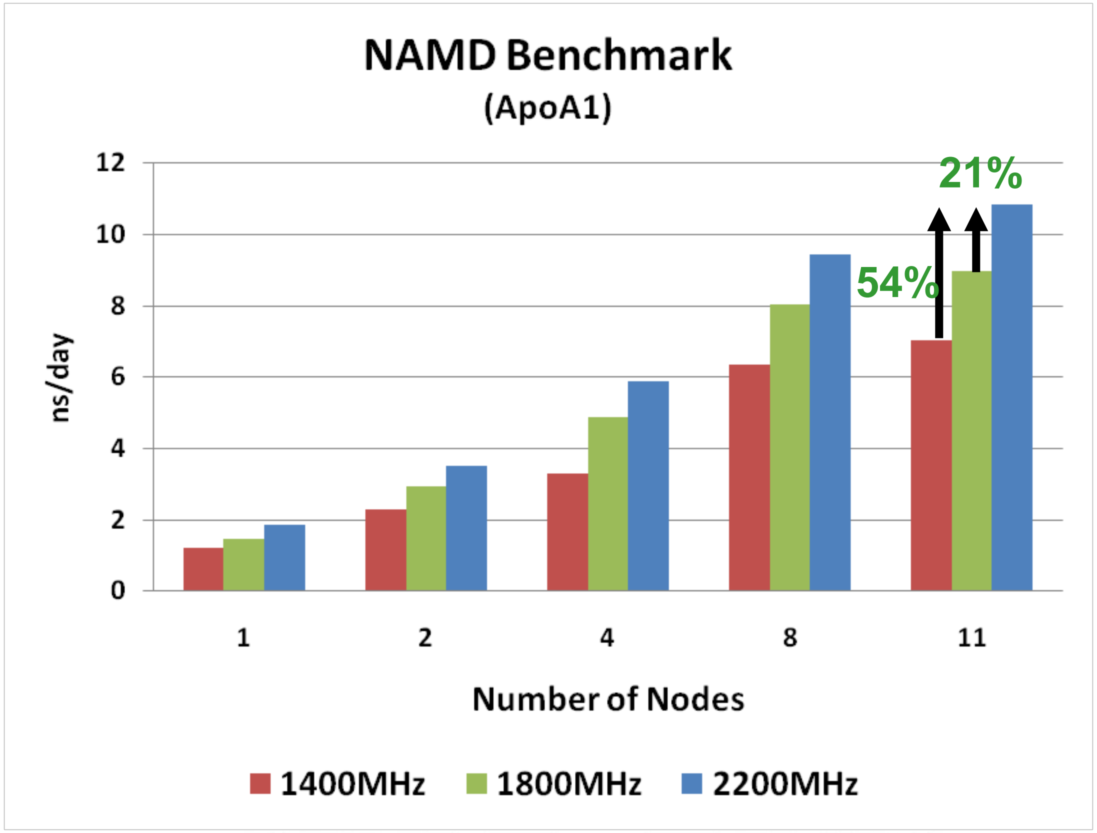


Higher is better

48 Cores/Node

NAMD Performance – CPU Frequency

- **Increasing CPU core frequency has a direct impact on job efficiency**
  - Up to 54% better job performance between 2200MHz vs 1400MHz
  - Up to 21% better job performance between 2200MHz vs 1800MHz
  - Performance improvement similar to the speed improvement

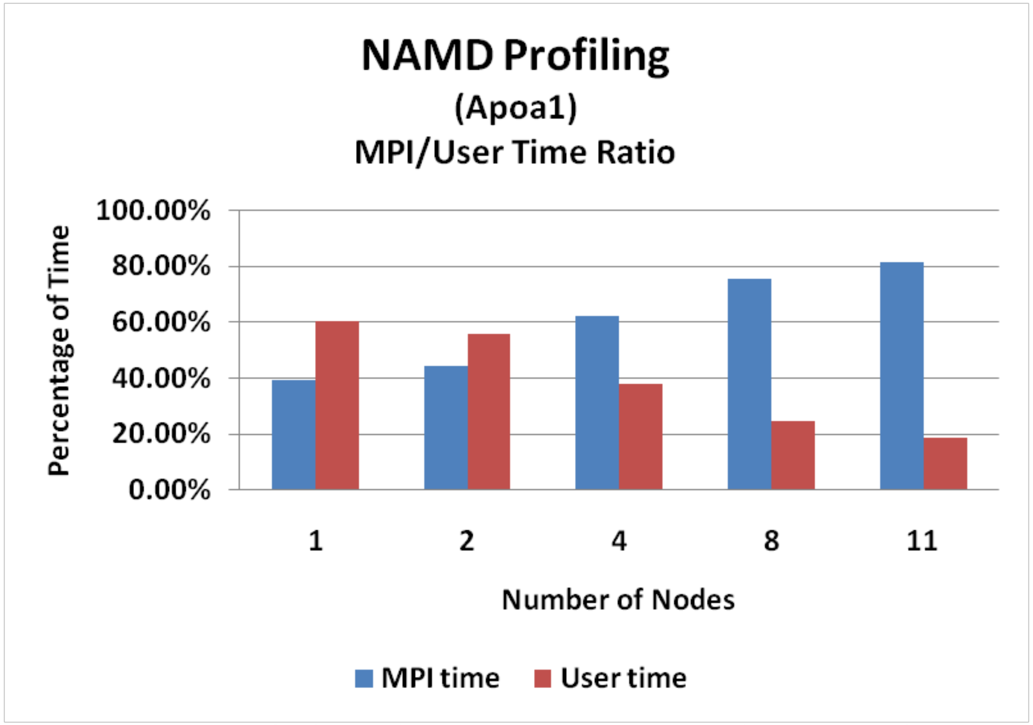


Higher is better

48 Cores/Node

NAMD Profiling – MPI/User Time Ratio

- **NAMD becomes highly communicative starting with 2-node**
  - Due to the high core counts per node
- **MPI communication time dominates the overall time**
  - Shows low latency interconnect such as InfiniBand is required for good scalability

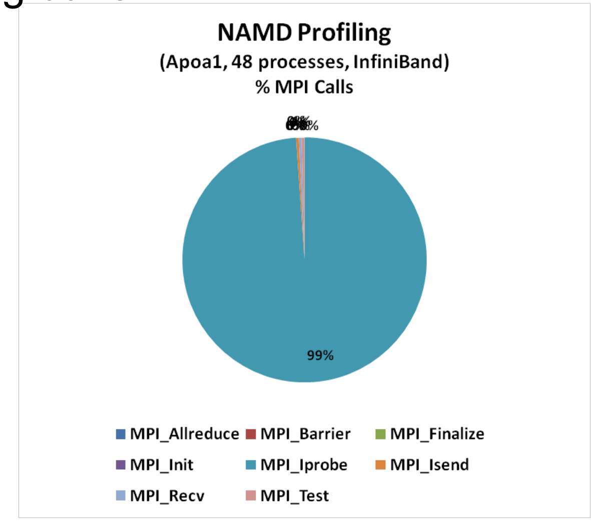


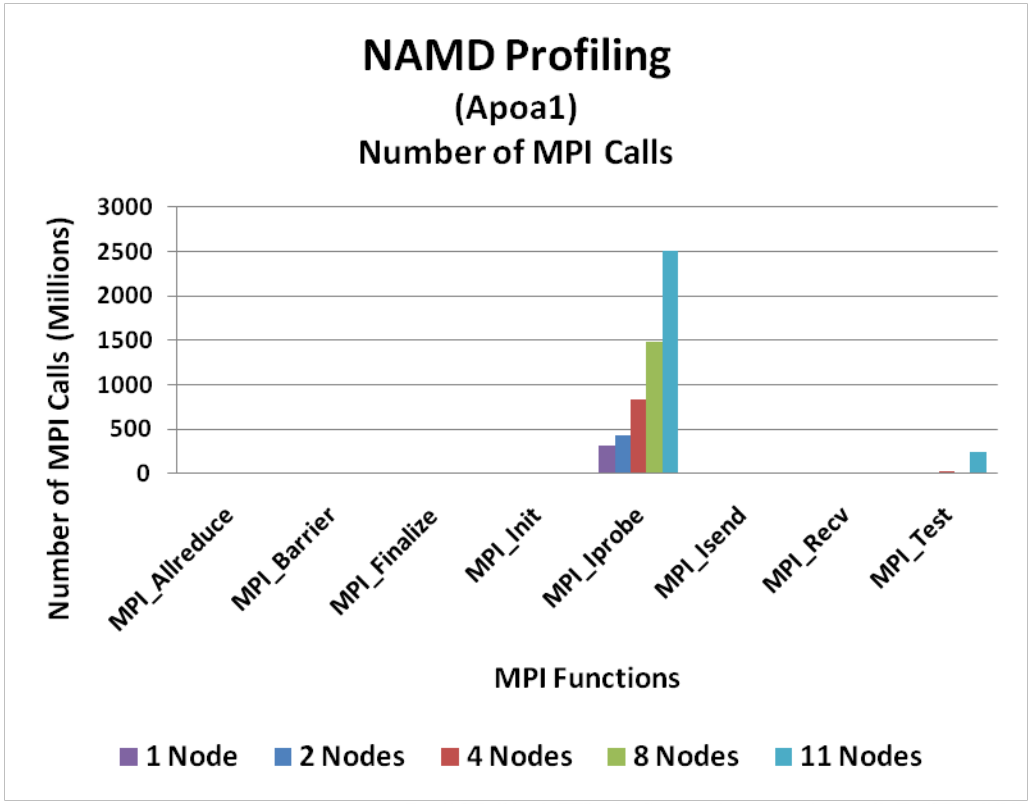
Higher is better

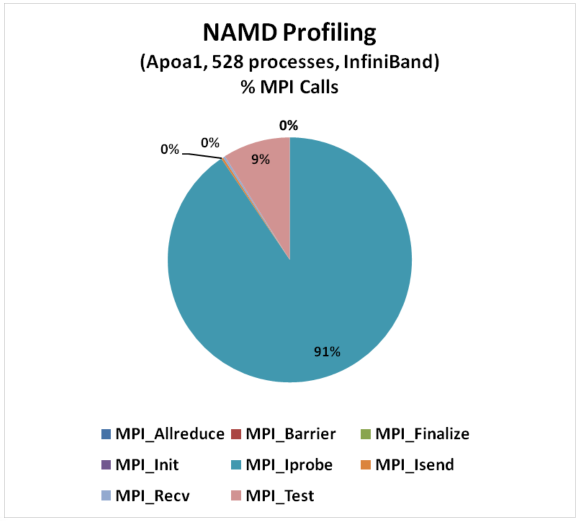
48 Cores/Node

NAMD Profiling – Number of MPI Calls

- **The most used MPI function is MPI_Iprobe**
  - Used for getting receiving message sizes and allows allocating buffer
  - Accounted for 99% of all MPI functions on a 1-node job
  - Accounted for 91% of all MPI functions on a 11-node job

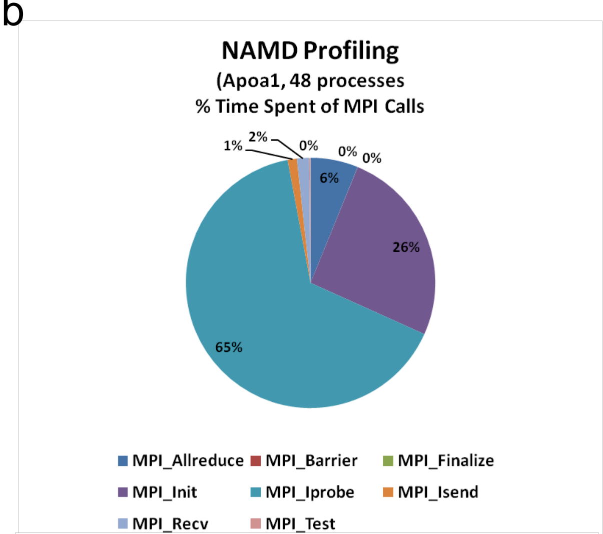


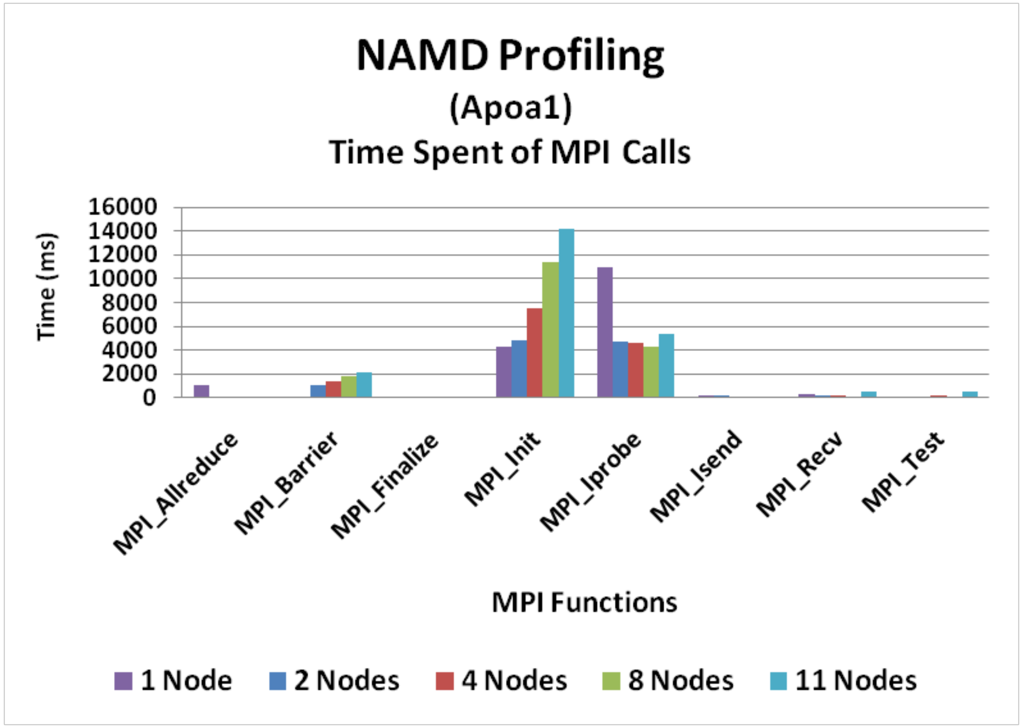


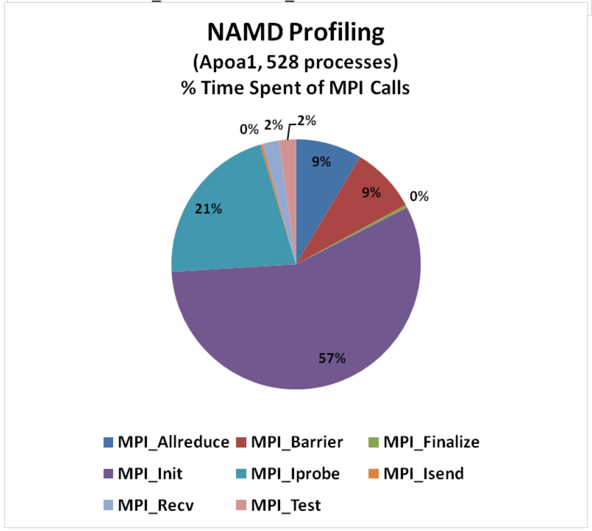


NAMD Profiling – Time Spent of MPI Calls

- **The most used MPI functions are MPI_Init and MPI_Iprobe**
  - Each accounted for 38% of all MPI functions on a 14-node job
- **MPI_Init occupies the largest percentage time**
  - Relatives to other MPI calls
  - Reflects other MPI data transfers are accomplished efficiently

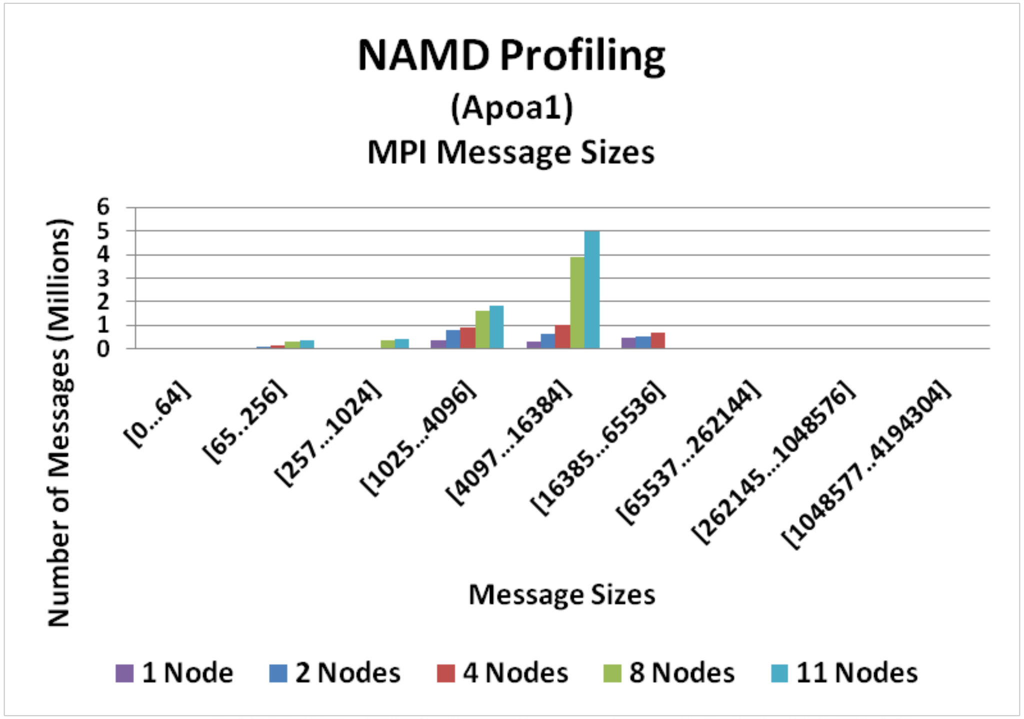






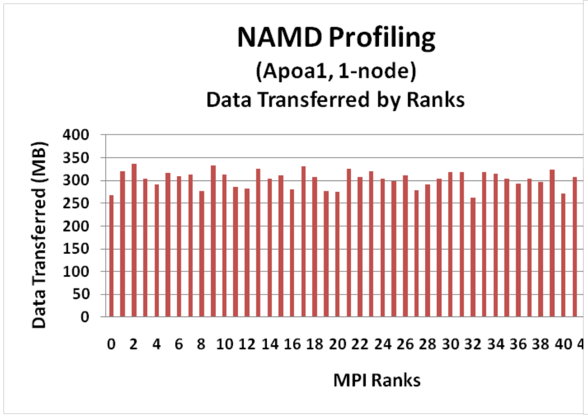
NAMD Profiling – MPI Message Sizes

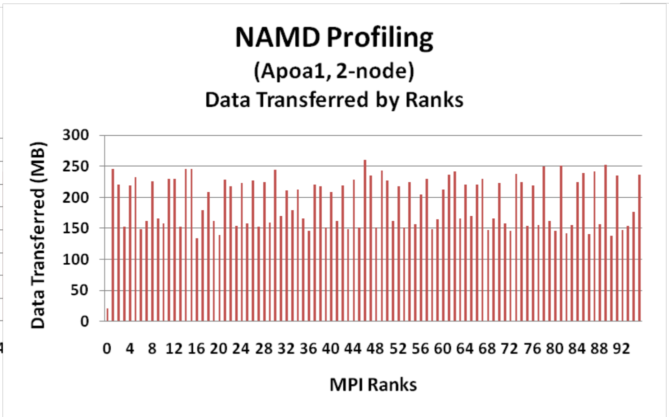
- **Majority of the MPI message sizes are**
  - in the range from 4KB to 16KB
- **Messages increase accelerates with the node count increases**
- **Benefit of Multi-rail begins to emerge starting with 8-node**

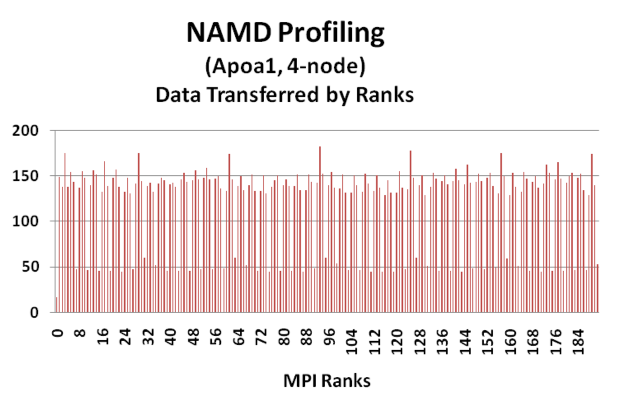


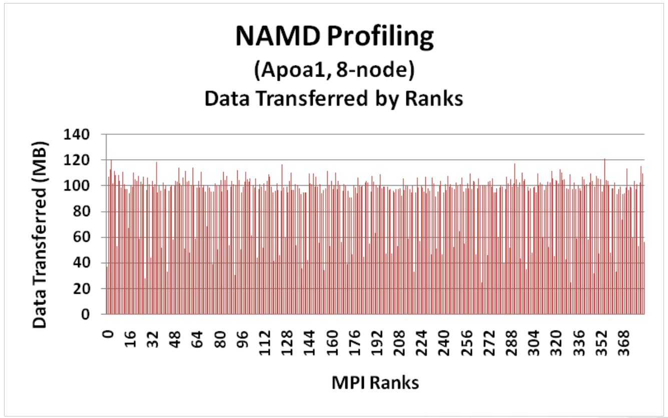
NAMD Profiling – Data Transfer Per Process

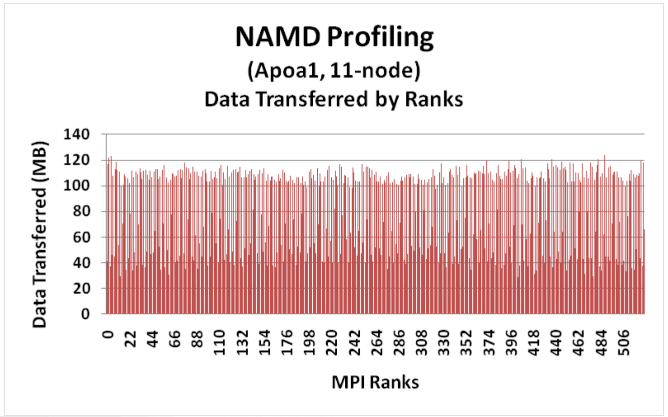
- **Data transferred to each MPI rank is showing some variance**
  - But overall data transfer is roughly the same on a per-node basis
- **As the cluster scales, less data is driven to each rank and each node**
  - 300MB per rank in a 48-process job versus 40-100MB per rank in a 528-process job





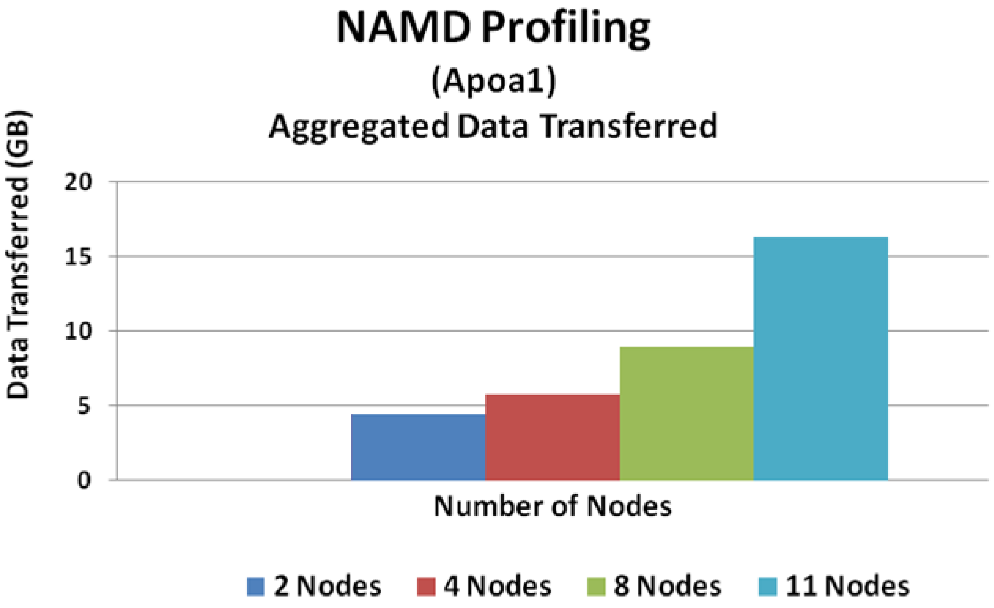






NAMD Profiling – Aggregated Data Transfer

- **Aggregated data transfer refers to:**
  - Total amount of data being transferred in the network between all MPI ranks collectively
- **The total data transfer increases as the cluster scales**
- **Demonstrates the advantage and importance of scalable network interconnect**
  - InfiniBand QDR can deliver bandwidth needed to push 16GB of data across the network



InfiniBand QDR

Summary

- **NAMD is an application for high-performance simulation of large biomolecular systems, is distributed free of charge with source code**
  - One of the leading bioscience application, found in many RFPs as well
- **Networking:**
  - InfiniBand shows as the preferred interconnect solution for any cluster size
    - Due to latency/throughput requirements,
  - Clear benefit for using dual-rail InfiniBand from 8-nodes and up
- **CPU:**
  - The CPU frequency has a direct impact on job productivity
- **MPIs:**
  - Open MPI (open source) and Platform MPI (commercial) are good candidates
  - Depends on the cluster size

Thank You

HPC Advisory Council

AMD The future is fusion

DELL

Mellanox TECHNOLOGIES

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