

NAMD

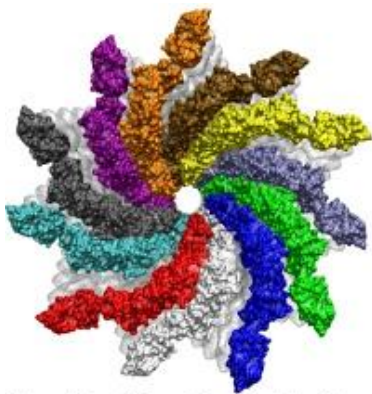
Performance Benchmark and Profiling

February 2012

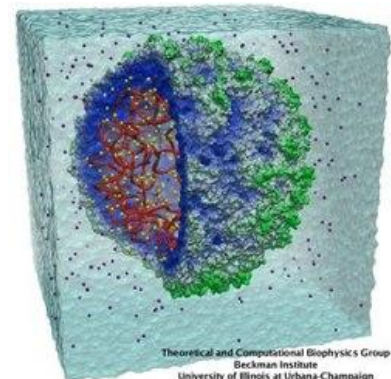
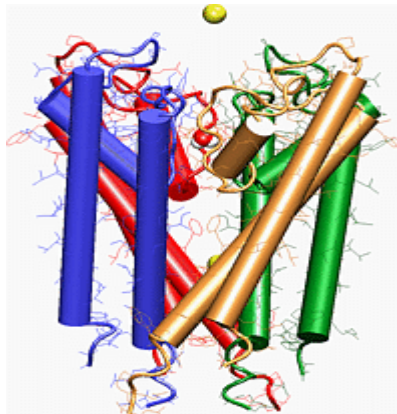


- **The following research was performed under the HPC Advisory Council activities**
 - Participating vendors: AMD, Dell, Mellanox
 - Compute resource - HPC Advisory Council Cluster Center
- **For more info please refer to**
 - [http:// www.amd.com](http://www.amd.com)
 - [http:// www.dell.com/hpc](http://www.dell.com/hpc)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

- **The following was done to provide best practices**
 - NAMD performance benchmarking
 - Interconnect performance comparisons
 - Ways to increase NAMD productivity
 - MPI libraries comparisons
- **The presented results will demonstrate**
 - The scalability of the compute environment to provide nearly linear application scalability
 - The capability of NAMD to achieve scalable productivity
 - Considerations for performance optimizations

- **Dell™ PowerEdge™ R815 11-node (704-core) cluster**
- **AMD™ Opteron™ 6276 (code name “Interlagos”) 16-core @ 2.3 GHz CPUs**
- **4 CPU sockets per server node**
- **Mellanox ConnectX®-3 FDR InfiniBand Adapters**
- **Mellanox SwitchX™ 6036 36-Port InfiniBand switch**
- **Memory: 128GB memory per node DDR3 1333MHz**
- **OS: RHEL 6.2, MLNX-OFED 1.5.3 InfiniBand SW stack**
- **MPI: Open MPI 1.5.5rc2, Platform MPI 8.2**
- **Compilers: GNU Compilers 4.6**
- **Application: NAMD 2.8 (External libraries used: charm-6.2.3, fftw-2.1.3, TCL 8.3)**
- **Benchmark workload:**
 - ApoA1 bloodstream lipoprotein particle model (92,224 atoms, periodic, PME, 12A cutoff)
 - ATPase benchmark (327,506 atoms, periodic, PME)

- **HPC Advisory Council Test-bed System**
- **New 11-node 704 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™ 6200 series platform and Mellanox ConnectX®-3 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 64 core/32DIMMs per server – 1344 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

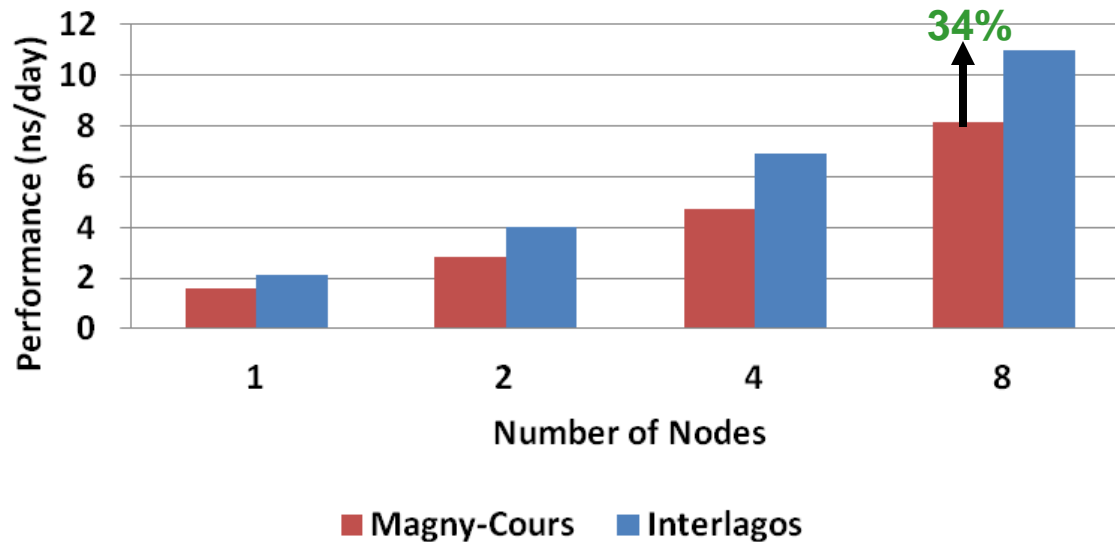
Optimized for long-term capital and operating investment protection

- System expansion
- Component upgrades and feature releases



- **Interlagos CPUs provides better performance than Magny-Cours CPUs**
 - Up to 34% gain in performance with Open MPI versus Magny-Cours CPUs
- **Processors used:**
 - Magny-Cours: AMD Opteron™ 6174 (2200MHz)
 - Interlagos: AMD Opteron™ 6276 (2300MHz)

NAMD Performance
(Apoa1, Open MPI)



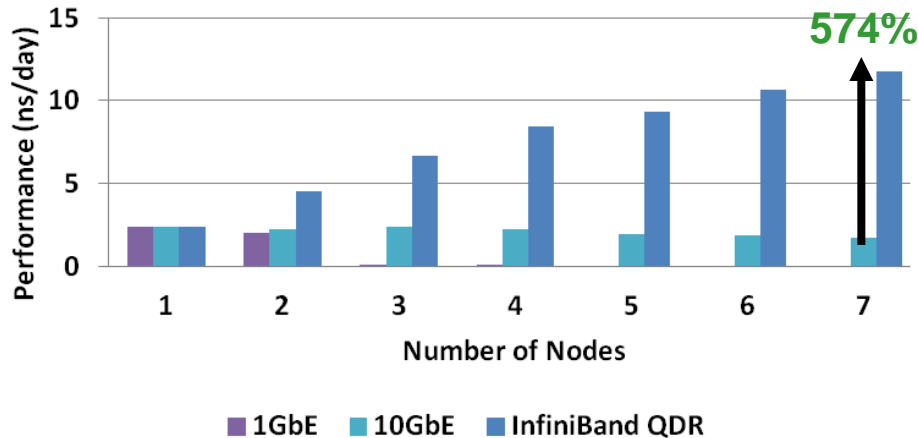
Higher is better

64 Cores/Node

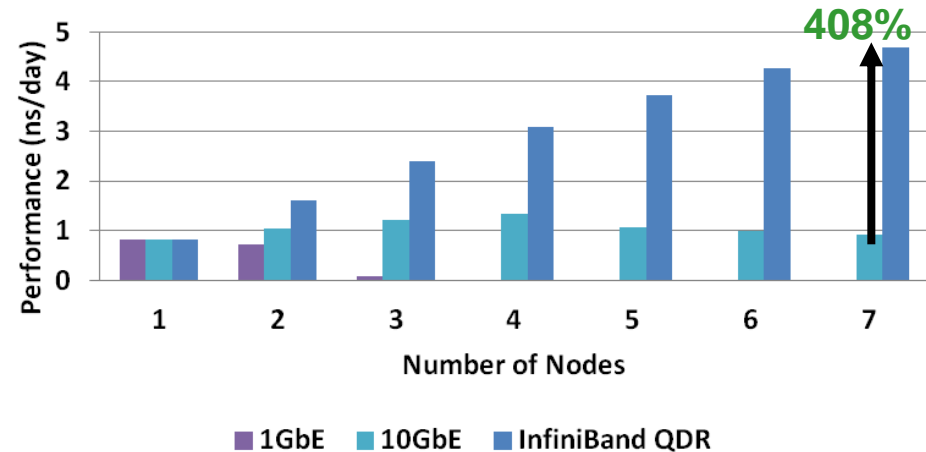
NAMD Performance – Interconnects

- **InfiniBand enables better scalability for NAMD**
 - Showing unlimited continuous gain to 7-node
- **Ethernet does not allow good scalability**
 - The performance of 1GbE plummet after 2 nodes (128 processes)
 - Both 1GbE and 10GbE do not show gain in productivity
 - The effect of MPI communications overwhelms the Ethernet network

**NAMD Performance
(Apoa1)**



**NAMD Performance
(f1atpase)**

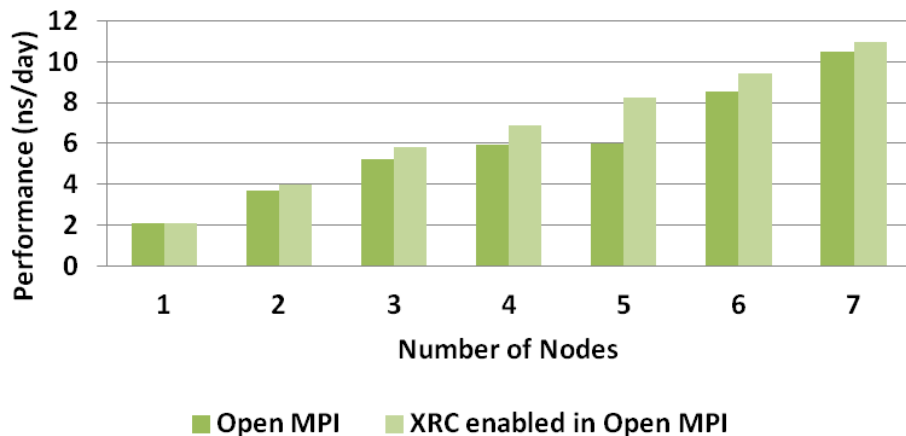


Higher is better

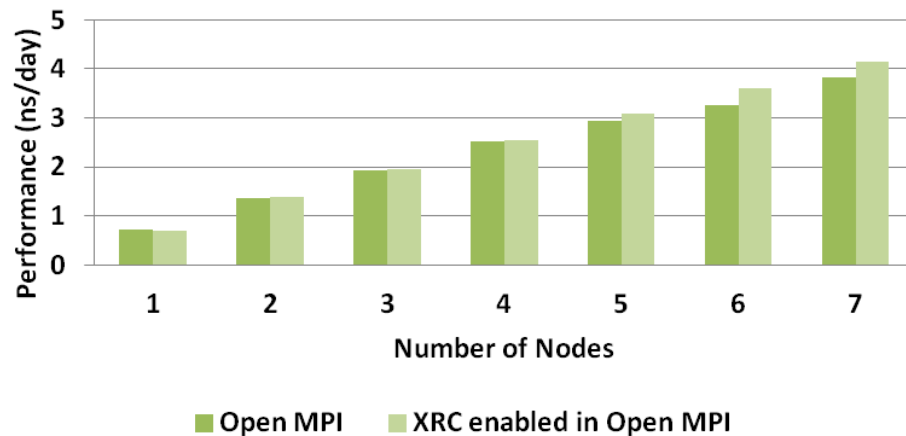
64 Cores/Node

- **Using XRC in Open MPI allows better performance and scalability**
 - Stands for eXtended Reliable Connection
 - Reduces memory footprint and is essential for scaling
 - Flags used: `-mca btl_openib_receive_queues X,128,256,192,128:X,2048,256,128,32:X,12288,256,128,32:X,65536,256,128,32`
- **Open MPI optimization flags used in both cases:**
 - `-bind-to-core -mca btl openib,sm,self`

**NAMD Performance
(Apoa1)**



**NAMD Performance
(f1atpase)**



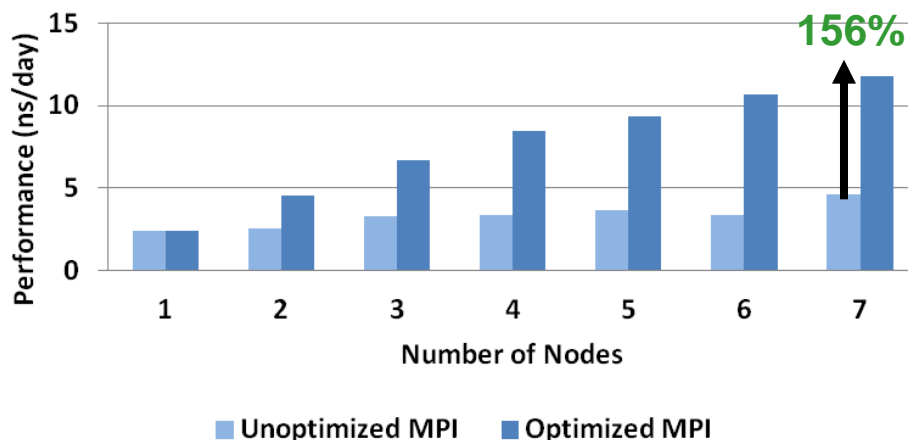
Higher is better

64 Cores/Node

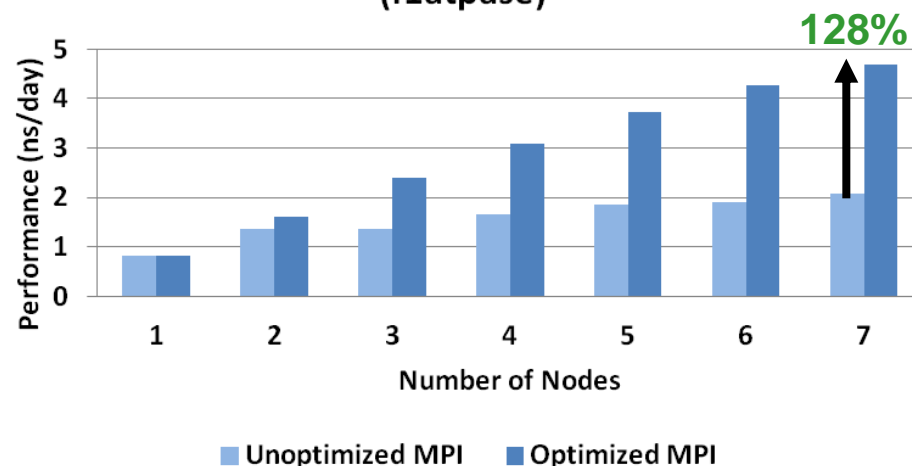
NAMD Performance – Platform MPI Tuning

- **Using SRQ (Shared Receive Queue) allow Platform MPI to scale**
 - Reduces memory footprint and is essential for scaling
- **Explicit mapping of CPU cores to ensure MPI ranks are placed sequentially**
 - To ensure CPU core enumeration, check with “lstopo” from hwloc or “numactl --hardware”
- **Extra flags in optimized case for SRQ, RDMA params and core bindings:**
 - `-srq -e MPI_RDMA_MSGSIZE=32768,32768,4194304 -e MPI_RDMA_NSQRQRECV=2048 -e MPI_RDMA_NFRAGMENT=128 -cpu_bind=v,map_cpu:0,4,8,12,16,20,24,28,32,36,40,44,48,52,56,60,2,6,10,14,18,22,26,30,34,38,42,46,50,54,58,62,3,7,11,15,19,23,27,31,35,39,43,47,51,55,59,63,1,5,9,13,17,21,25,29,33,37,41,45,49,53,57,61`

NAMD Performance (Apoa1)



NAMD Performance (f1atpase)



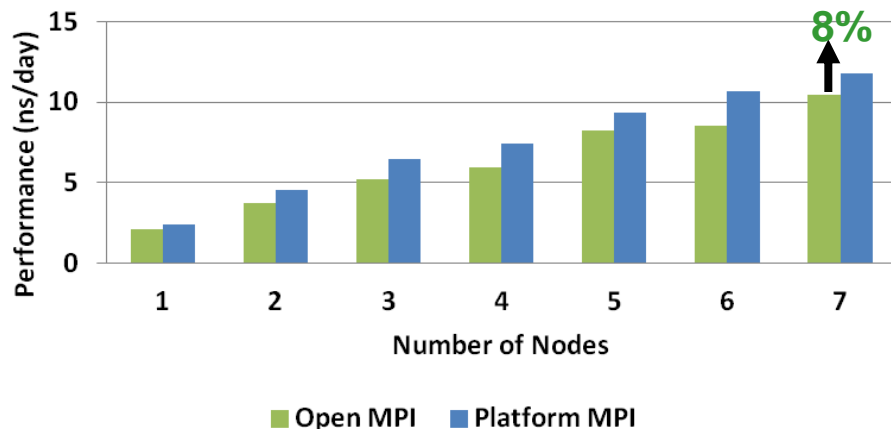
Higher is better

64 Cores/Node

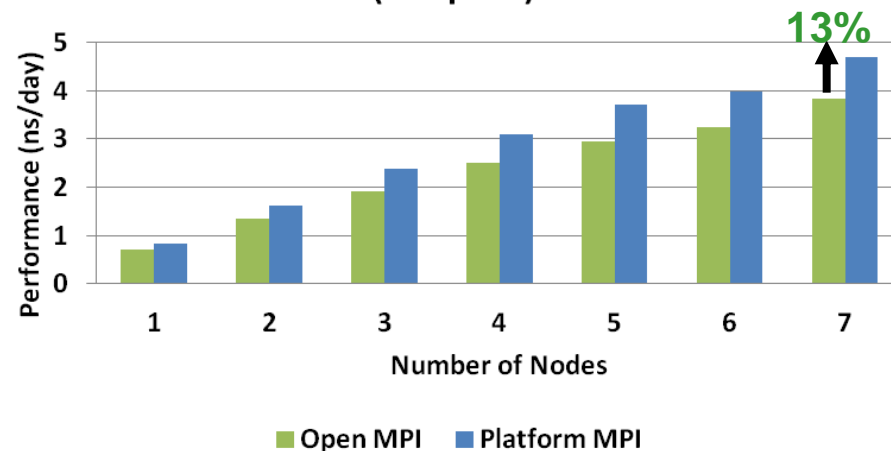
NAMD Performance – MPI Implementations

- **The Platform MPI performs better than Open MPI**
 - Up to 8% better than Open MPI for apoa1 at 7-node
 - Up to 13% better than Open MPI for f1atpase at 7-node
- **The tuned flags are used for both Platform MPI and Open MPI**

**NAMD Performance
(Apoa1)**



**NAMD Performance
(f1atpase)**



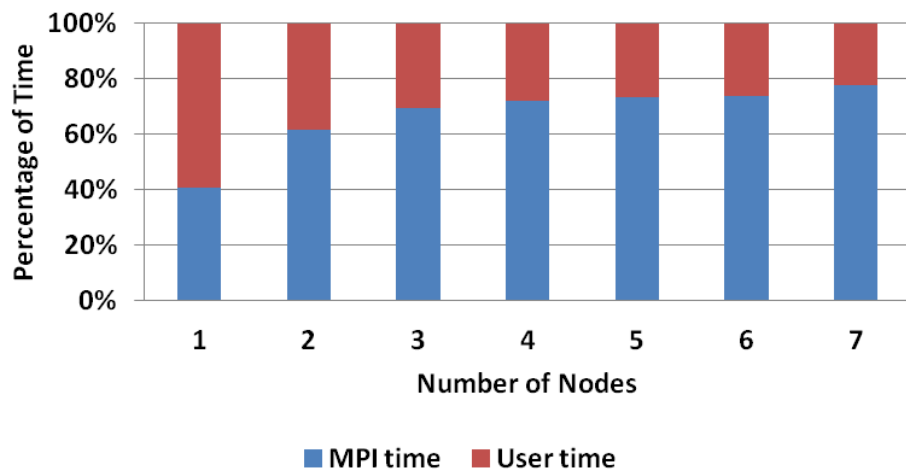
Higher is better

64 Cores/Node

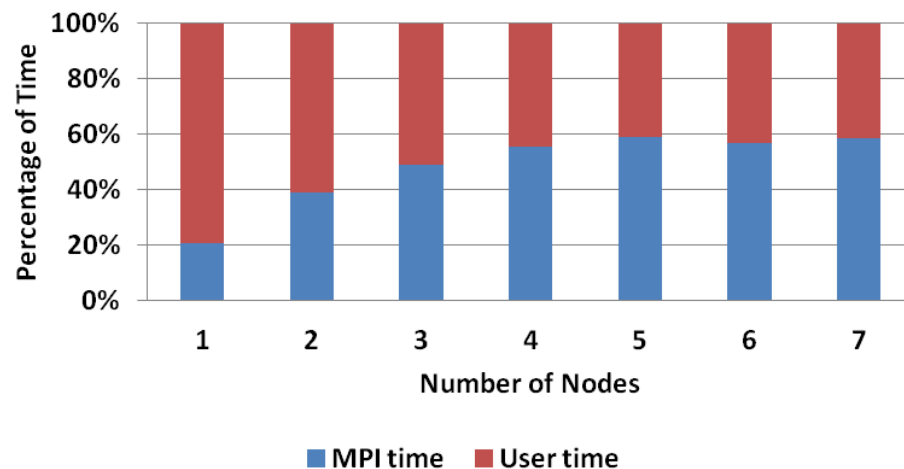
NAMD Profiling – MPI/User Time Ratio

- **NAMD becomes highly communicative starting from 2-3 nodes**
 - Due to the high core counts per node
 - The stmv contains more CPU and MPI communication times than apoa1
- **MPI communication time dominates the overall time**
 - Shows low latency interconnect such as InfiniBand is required for good scalability

NAMD Profiling
(Apoa1)
MPI/User Time Ratio



NAMD Profiling
(stmv)
MPI/User Time Ratio



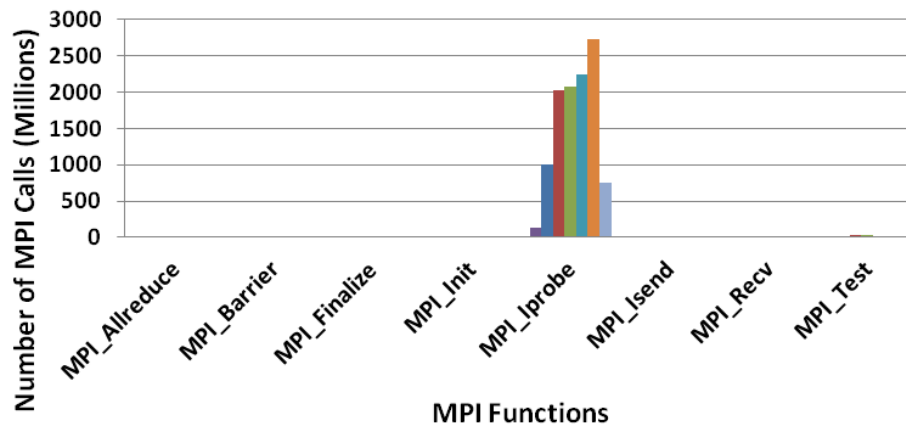
Higher is better

64 Cores/Node

NAMD Profiling – Number of MPI Calls

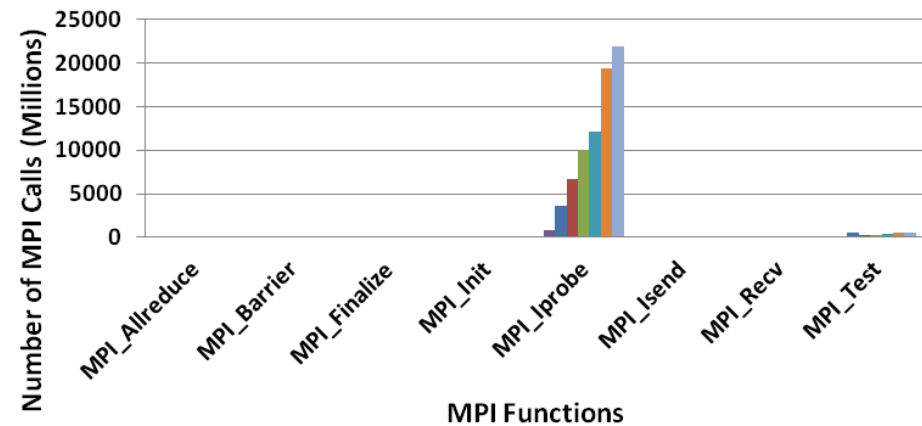
- **The most used MPI function is MPI_Iprobe**
 - MPI_Iprobe is used for testing non-blocking messages
 - Accounted for 97% of all MPI calls

**NAMD Profiling
(Apoa1)**
Number of MPI Calls



■ 1 Node ■ 2 Nodes ■ 3 Nodes ■ 4 Nodes
■ 5 Nodes ■ 6 Nodes ■ 7 Nodes

**NAMD Profiling
(stmv)**
Number of MPI Calls

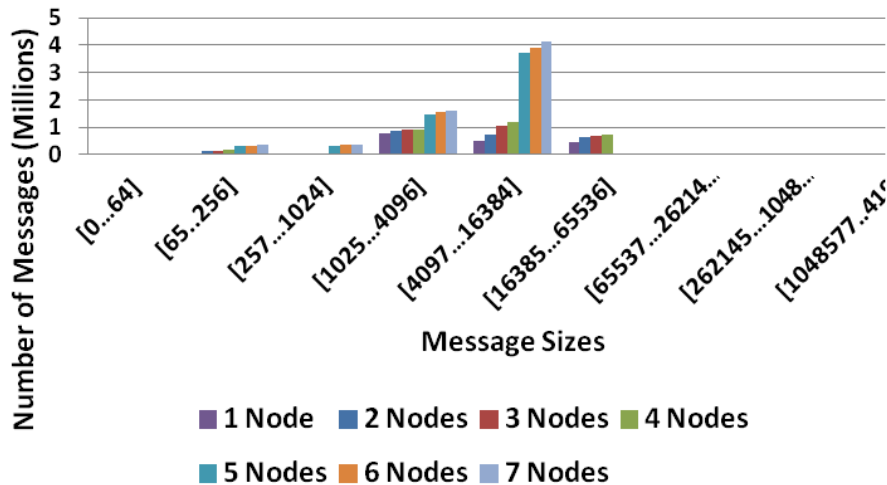


■ 1 Node ■ 2 Nodes ■ 3 Nodes ■ 4 Nodes
■ 5 Nodes ■ 6 Nodes ■ 7 Nodes

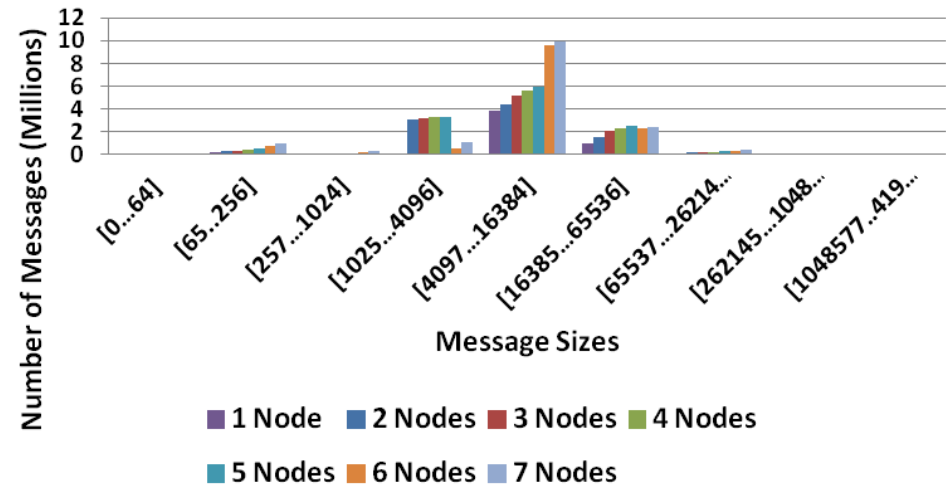
NAMD Profiling – MPI Message Sizes

- Majority of the MPI message sizes are
 - in the range from 4KB to 16KB
- The increase in messages accelerates starting around 5-6 nodes

NAMD Profiling
(Apoa1)
MPI Message Sizes



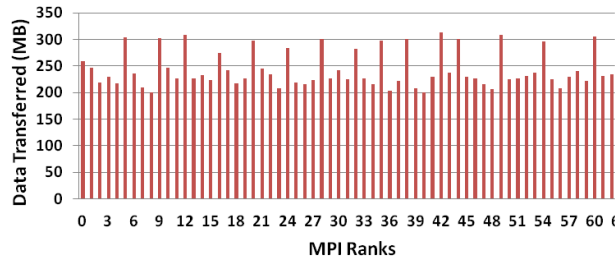
NAMD Profiling
(stmv)
MPI Message Sizes



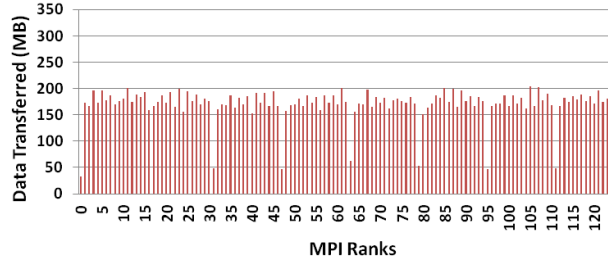
NAMD Profiling – Data Transfer Per Process

- **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**
 - 300-600MB per rank in 1-node job versus 150-200MB per rank in a 4-node job

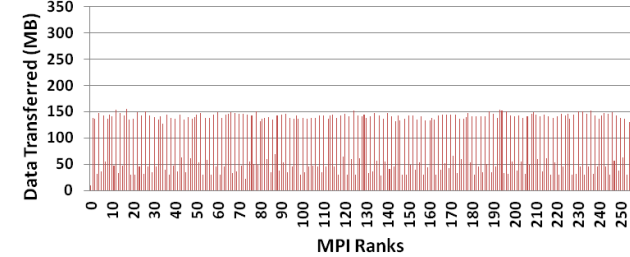
**NAMD Profiling
(Apoa1, 1-node)
Data Transferred by Ranks**



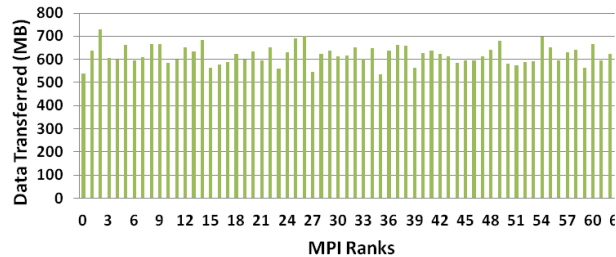
**NAMD Profiling
(Apoa1, 2-node)
Data Transferred by Ranks**



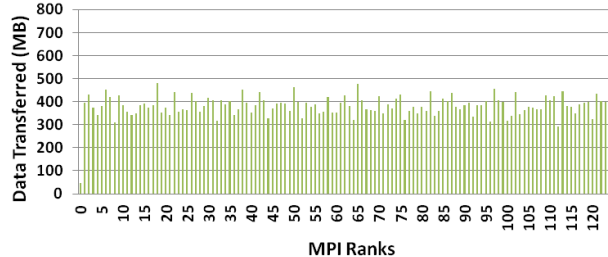
**NAMD Profiling
(Apoa1, 4-node)
Data Transferred by Ranks**



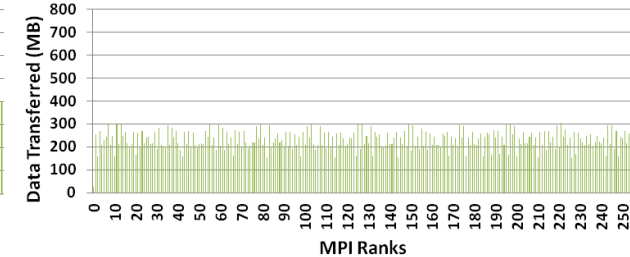
**NAMD Profiling
(f1atpase, 1-node)
Data Transferred by Ranks**



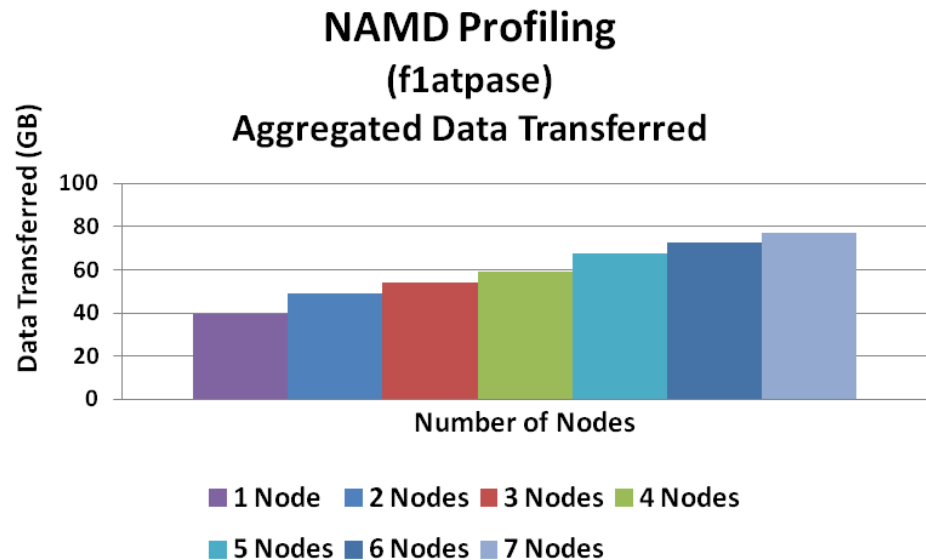
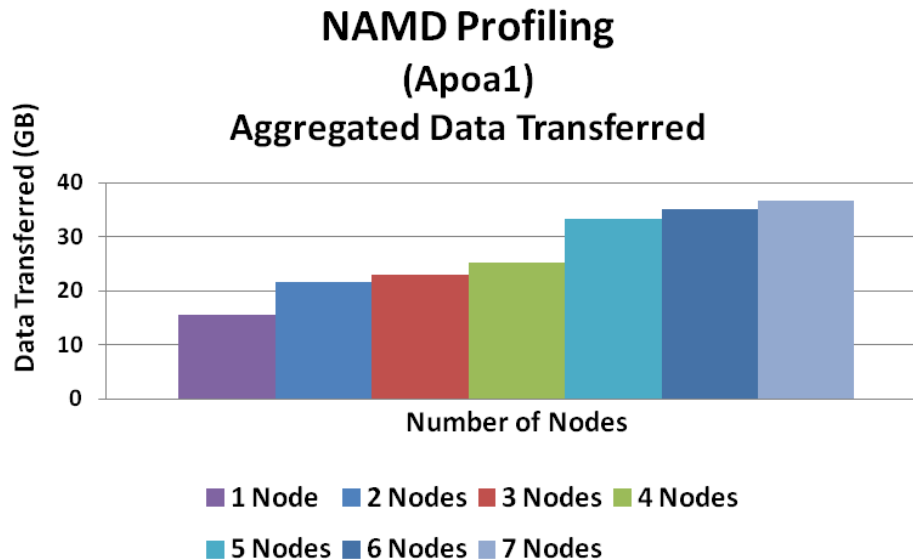
**NAMD Profiling
(f1atpase, 2-node)
Data Transferred by Ranks**



**NAMD Profiling
(f1atpase, 4-node)
Data Transferred by Ranks**



- **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 importance of scalable network interconnect**
 - InfiniBand can deliver bandwidth needed to push data in 40GB+ across the network



- **Interlagos provides higher performance than Magny-Cours CPUs**
 - Up to 34% performance gain with Open MPI
 - AMD Opteron™ 6276 (code name “Interlagos”) 16-core @ 2.3 GHz CPUs
 - AMD Opteron™ 6174 (code name “Magny-Cours”) 12-core @ 2.2GHz CPUs
- **Mellanox ConnectX®-3 proves significantly higher scalability for NAMD**
 - 4x to 5x higher performance versus 10GbE
- **Open MPI and Platform MPI benefit from tuned parameters**
 - Having XRC and SRQ enabled the MPIs to scale at large core counts
- **The tuned Platform MPI performs better than the tuned Open MPI**
 - By 8-13% on 2 different datasets

Thank You

HPC Advisory Council



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