NAMD
Performance Benchmark and Profiling

March 2013

Note

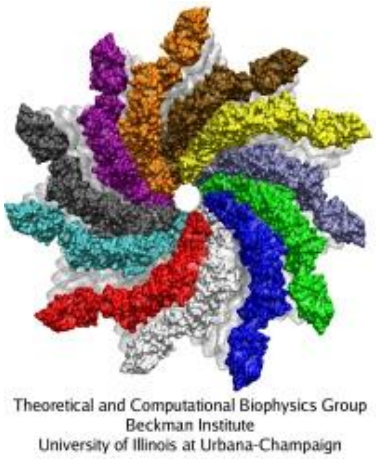
- **The following research was performed under the HPC Advisory Council activities**
  - Participating vendors: Intel, Dell, Mellanox
  - Compute resource - HPC Advisory Council Cluster Center

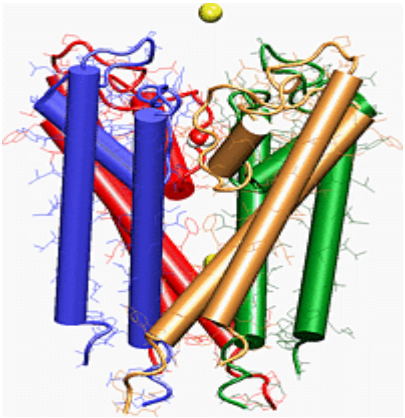
- **The following was done to provide best practices**
  - NAMD performance overview
  - Understanding NAMD communication patterns
  - Ways to increase NAMD productivity
  - MPI libraries comparisons

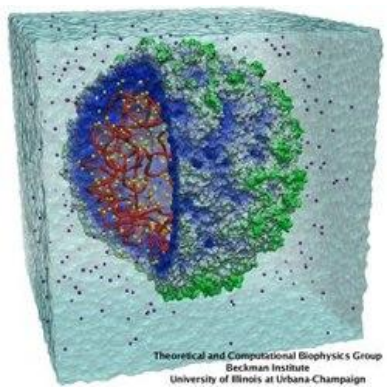
- **For more info please refer to**
  - http://www.dell.com
  - http://www.intel.com
  - http://www.mellanox.com
  - http://www.ks.uiuc.edu/Research/namd

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







Objectives

- **The presented research was done to provide best practices**
  - NAMD performance benchmarking
    - MPI Library performance comparison
    - Interconnect performance comparison
    - CPUs comparison
    - Compilers comparison
- **The presented results will demonstrate**
  - The scalability of the compute environment/application
  - Considerations for higher productivity and efficiency

Test Cluster Configuration

- **Dell™ PowerEdge™ R720xd and R720 32-node "Jupiter" cluster**
  - 16 nodes: Dual-Socket Eight-Core Intel E5-2680 @ 2.70 GHz CPUs (Static max Perf in BIOS)
  - 16 nodes: Dual-Socket Four-Core Intel E5-2603 @ 1.80 GHz CPUs (Static max Perf in BIOS)
  - Memory: 64GB memory, DDR3 1600 MHz
  - OS: RHEL 6.2, OFED 1.5.3 InfiniBand SW stack
  - Hard Drives: 24x 250GB 7.2 RPM SATA 2.5" on RAID 0
- **Intel Cluster Ready certified cluster**
- **Mellanox ConnectX-3 FDR InfiniBand VPI adapters**
- **Mellanox SwitchX SX6036 InfiniBand switch**
- **Compilers: GNU 4.6.3, Intel Composer XE 2011**
- **MPI: Intel MPI 4 U3, Open MPI 1.5.5 (KNEM 0.9.8)**
- **Application: NAMD 2.9**
- **External libraries: charm-6.4.0, fftw-2.1.3, TCL 8.3**
  - Benchmark: ApoA1 bloodstream lipoprotein particle model (92,224 atoms, periodic, PME, 12A cutoff)

About Intel® Cluster Ready

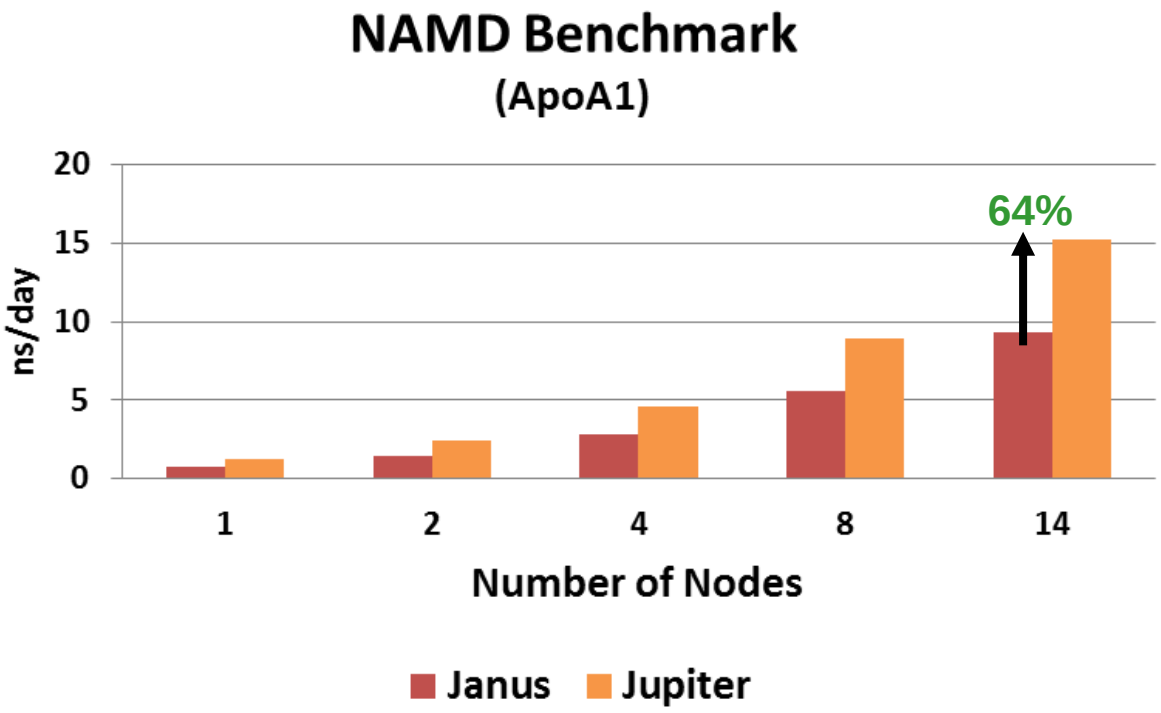
- **Intel® Cluster Ready systems make it practical to use a cluster to increase your simulation and modeling productivity**
  - Simplifies selection, deployment, and operation of a cluster
- **A single architecture platform supported by many OEMs, ISVs, cluster provisioning vendors, and interconnect providers**
  - Focus on your work productivity, spend less management time on the cluster
- **Select Intel Cluster Ready**
  - Where the cluster is delivered ready to run
  - Hardware and software are integrated and configured together
  - Applications are registered, validating execution on the Intel Cluster Ready architecture
  - Includes Intel® Cluster Checker tool, to verify functionality and periodically check cluster health

About Dell PowerEdge R720xd

- **Performance and efficiency**
  - Intelligent hardware-driven systems management with extensive power management features
  - Innovative tools including automation for parts replacement and lifecycle manageability
  - Broad choice of networking technologies from GbE to IB
  - Built in redundancy with hot plug and swappable PSU, HDDs and fans
- **Benefits**
  - Designed for performance workloads
    - from big data analytics, distributed storage or distributed computing where local storage is key to classic HPC and large scale hosting environments
    - High performance scale-out compute and low cost dense storage in one package
- **Hardware Capabilities**
  - Flexible compute platform with dense storage capacity
    - 2S/2U server, 6 PCIe slots
  - Large memory footprint (Up to 768GB / 24 DIMMs)
  - High I/O performance and optional storage configurations
    - HDD options: 12 x 3.5” - or - 24 x 2.5 + 2x 2.5 HDDs in rear of server
    - Up to 26 HDDs with 2 hot plug drives in rear of server for boot or scratch

NAMD Performance – Processor Generations

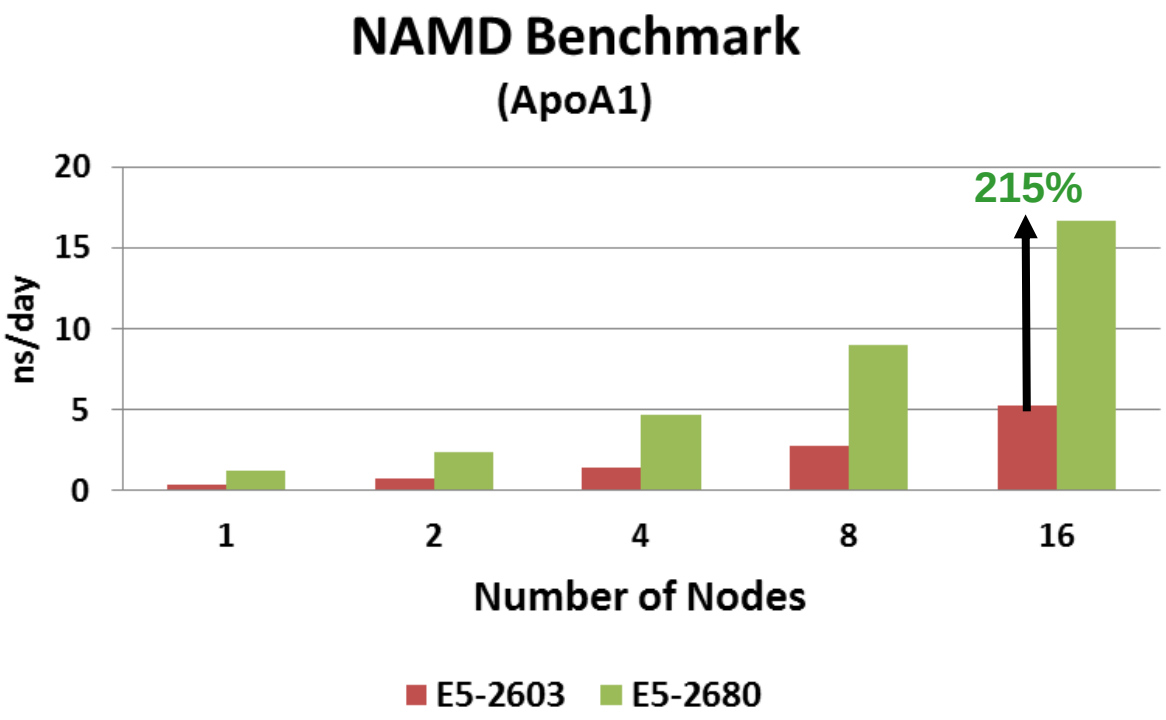
- **Intel Xeon E5-2600 Series (Sandy Bridge) outperforms prior generations**
  - Up to 64% higher performance than Intel Xeon X5670 (Westmere) at 14-node
- **System components used:**
  - Jupiter: 2-socket Intel E5-2680 @ 2.7GHz, 1600MHz DIMMs, FDR IB, 24 disks
  - Janus: 2-socket Intel x5670 @ 2.93GHz, 1333MHz DIMMs, QDR IB, 1 disk
- **14 nodes are used in the comparison**
  - In order to compare with results previously done on Janus cluster



Higher is better

NAMD Performance – Processor Frequencies

- **Higher CPU frequency drives higher productivity NAMD**
  - E5-2680 delivers up to 215% higher performance than E5-2603 at 16-node
- **System components used:**
  - 2-socket Intel E5-2680 @ 2.7GHz, 16PPN, 1600MHz DIMMs, FDR InfiniBand
  - 2-socket Intel E5-2603 @ 1.8GHz, 8PPN, 1600MHz DIMMs, FDR InfiniBand

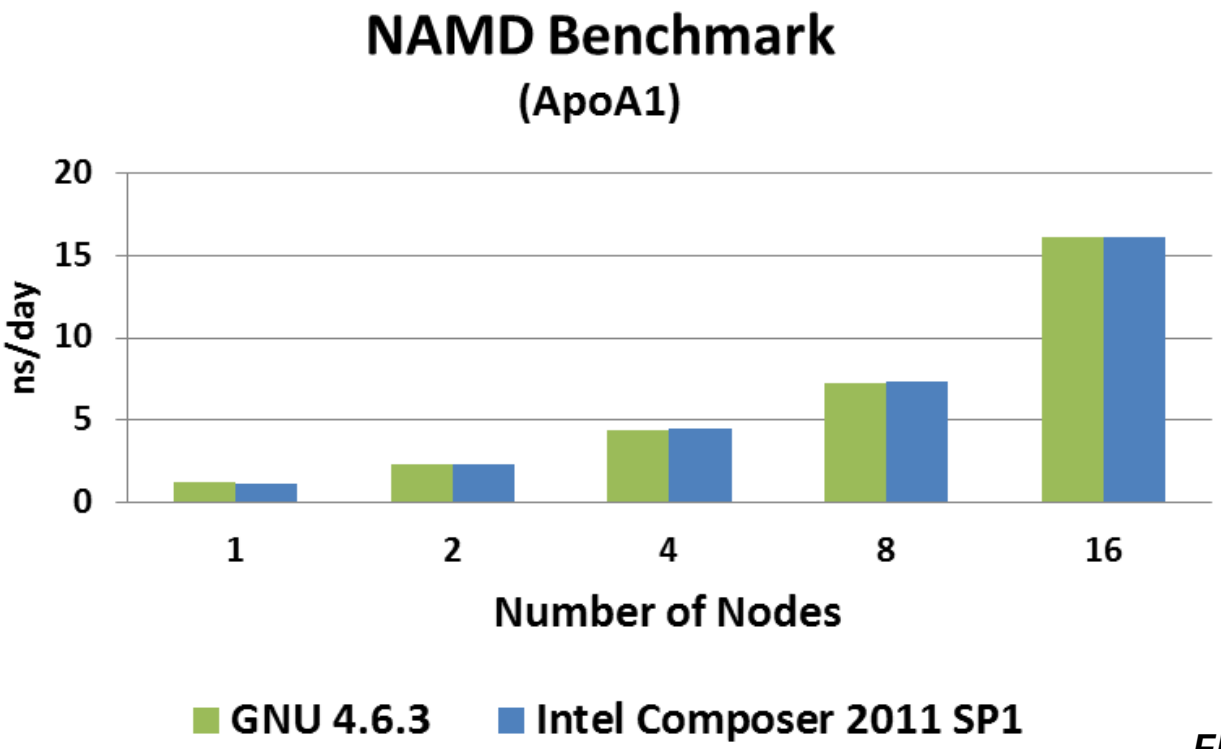


Higher is better

Open MPI

NAMD Performance – Compilers

- **Both Intel Composer and GNU compilers show comparable results**
  - Both compilers are able to generate good results
- **Default compile options are being tested**
  - Standard compiler options are used when building NAMD and its dependencies

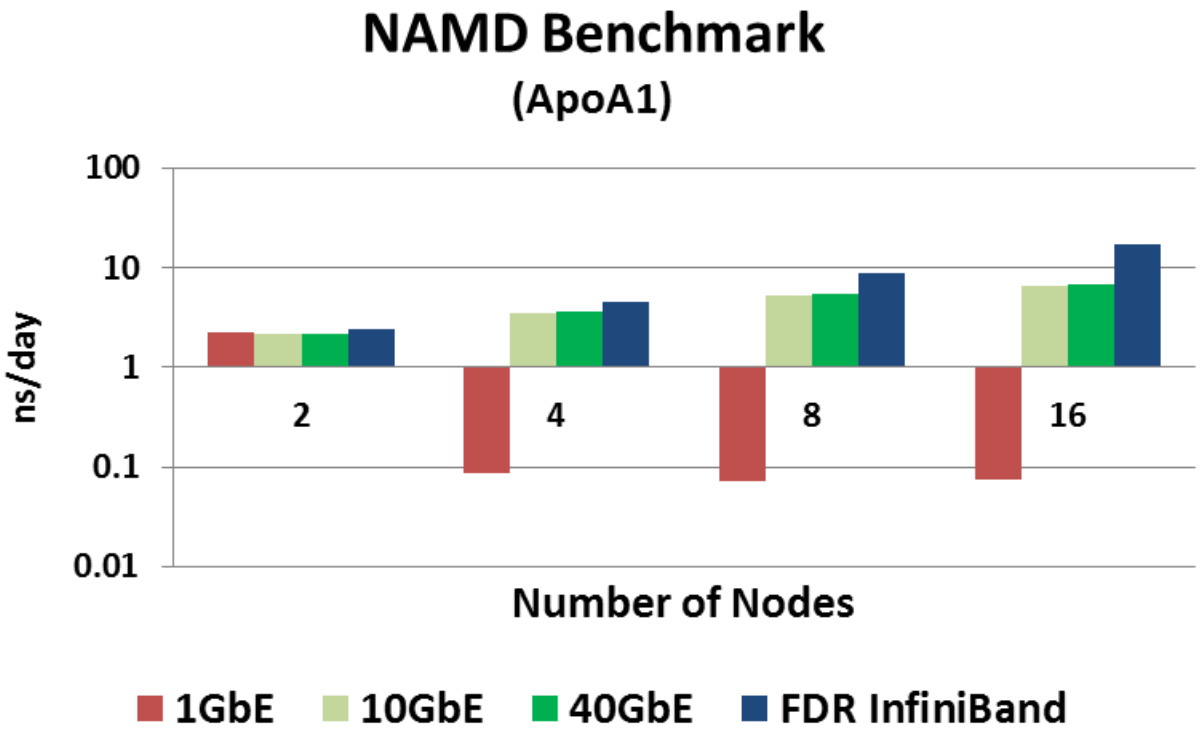


Higher is better

FDR InfiniBand

NAMD Performance – Interconnects

- **FDR InfiniBand enables the highest cluster scalability**
  - Outperformed Ethernet by +150% on 16 nodes
- **Network bottlenecks hinder particularly to the scalability of Ethernet**
  - 1GbE scalability performance is limited at 2 nodes due to high latency
  - 10/40GbE performance is limited significantly at scale; ended up wasting compute time

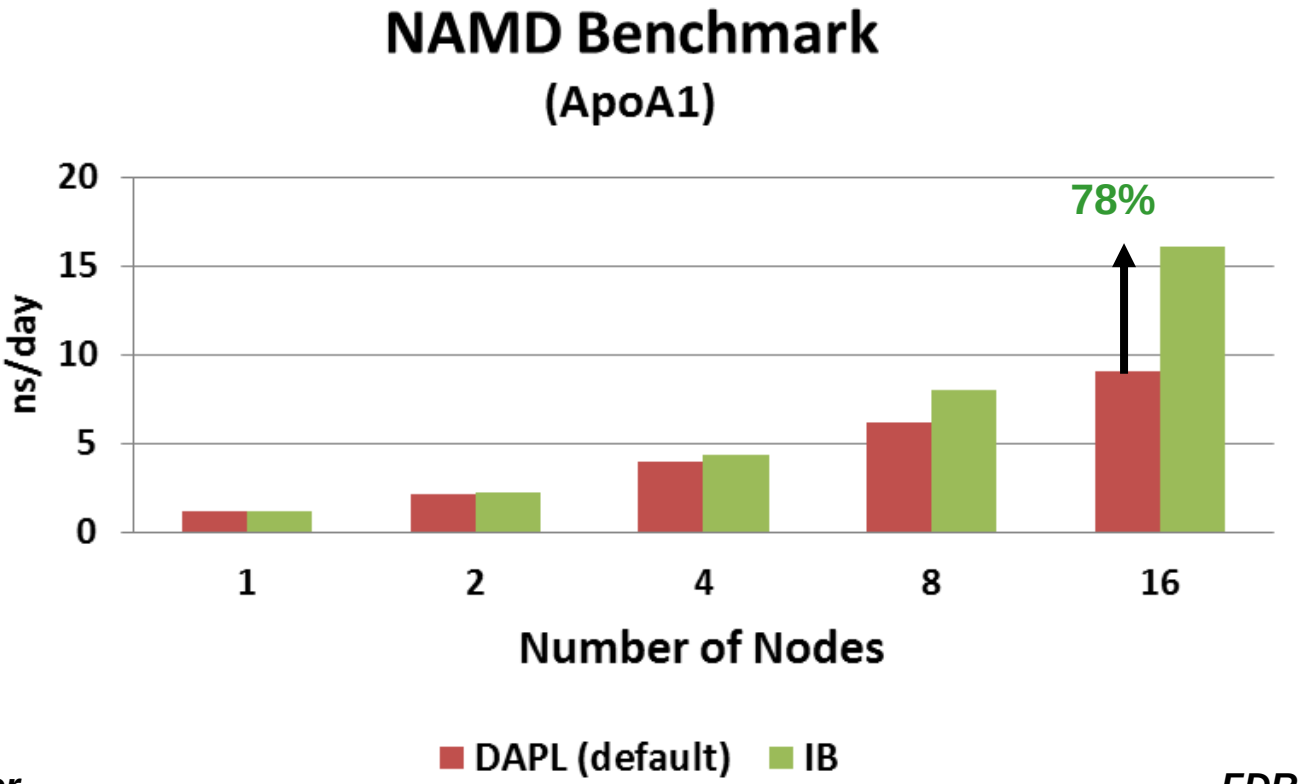


Higher is better

Platform MPI

NAMD Performance – Intel MPI

- **The -IB (OFA) option allows Intel MPI to run at the most efficient rate**
  - The default option for Intel MPI is use the DAPL when not specified in command line
  - The "-IB" options runs faster than the "-DAPL" option by 78% at 16 nodes
  - Both OFA and DAPL are the options available for running on the InfiniBand networks
  - The OFA option is enabled by using the "-IB" option

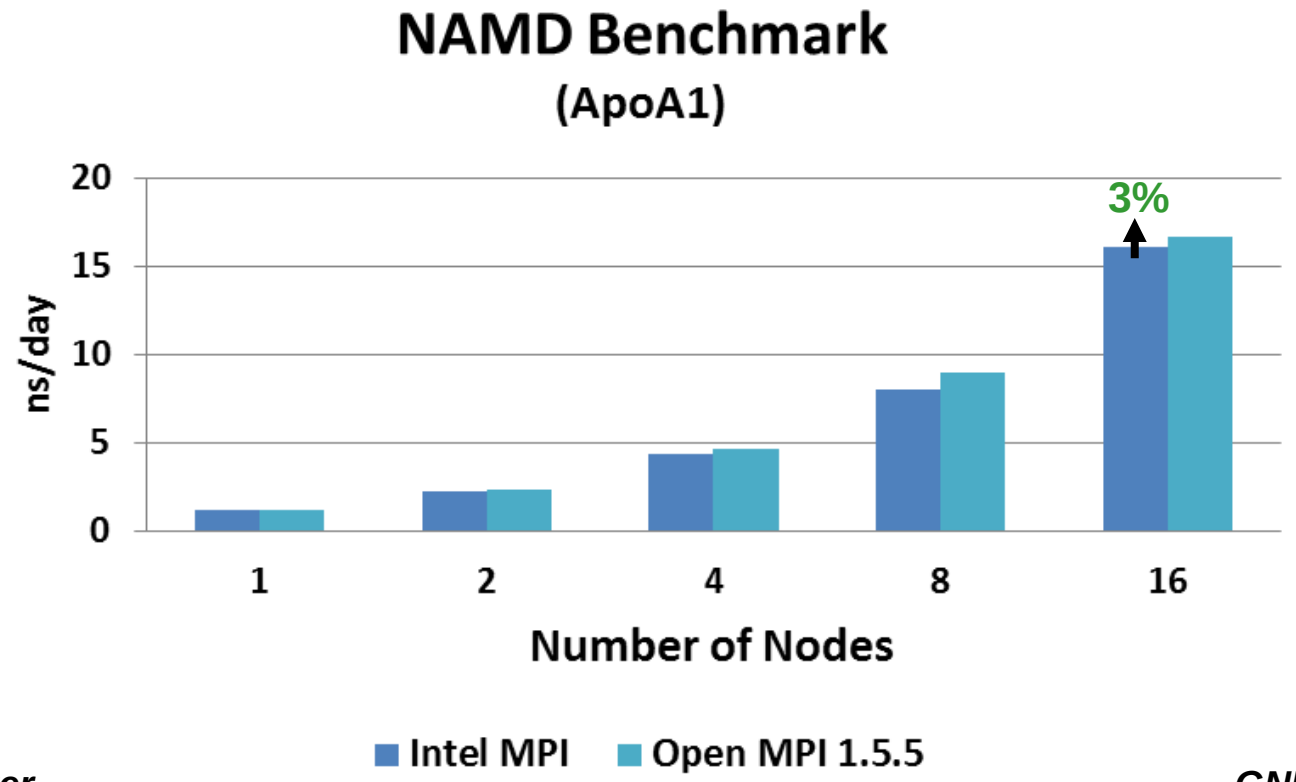


Higher is better

FDR InfiniBand

NAMD Performance – MPI Implementations

- **Both Open MPI and Intel MPI perform comparably**
  - Open MPI runs slightly faster than Intel MPI by 3% at 16 nodes
  - The IBV interface is used for Intel MPI
  - Reflects Intel MPI handles data transfer more efficiently at scale



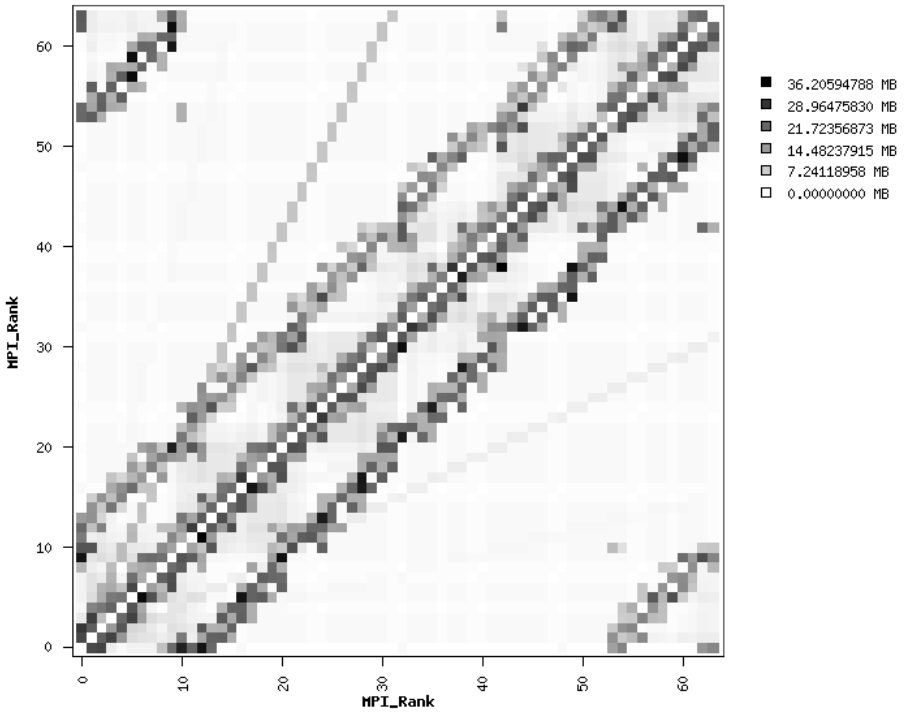
Higher is better

GNU Compilers

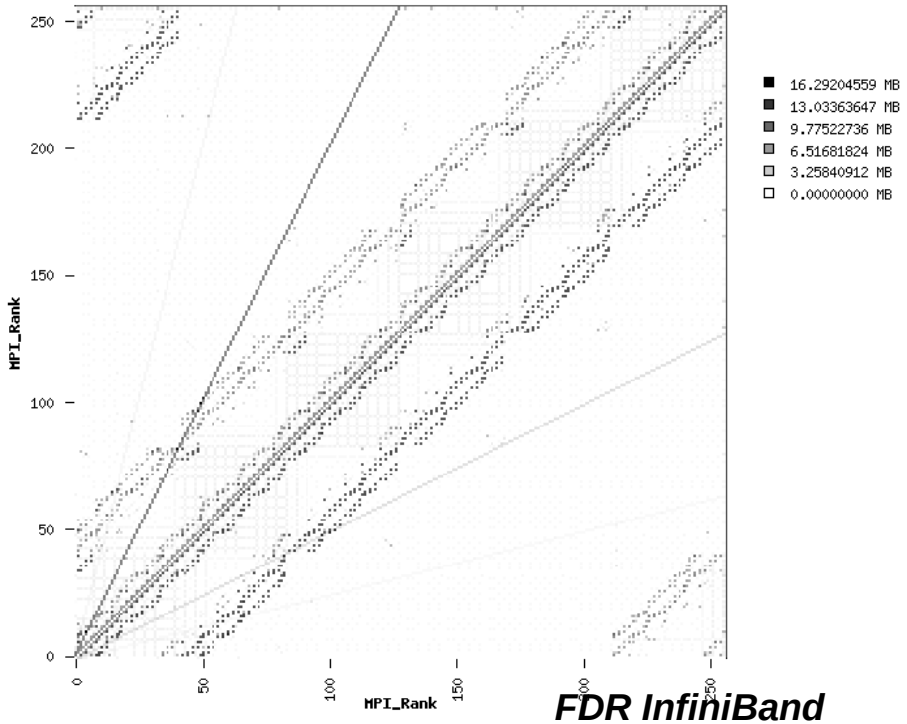
NAMD Profiling – Point-to-point Dataflow

- **Majority of the communication flows occurs to ranks closer to self**
  - Heavier communications between neighboring ranks
- **Similar communication pattern seen as the cluster scales**
  - The amount of data being transferred is reduced as the node scales

4 Nodes



16 Nodes

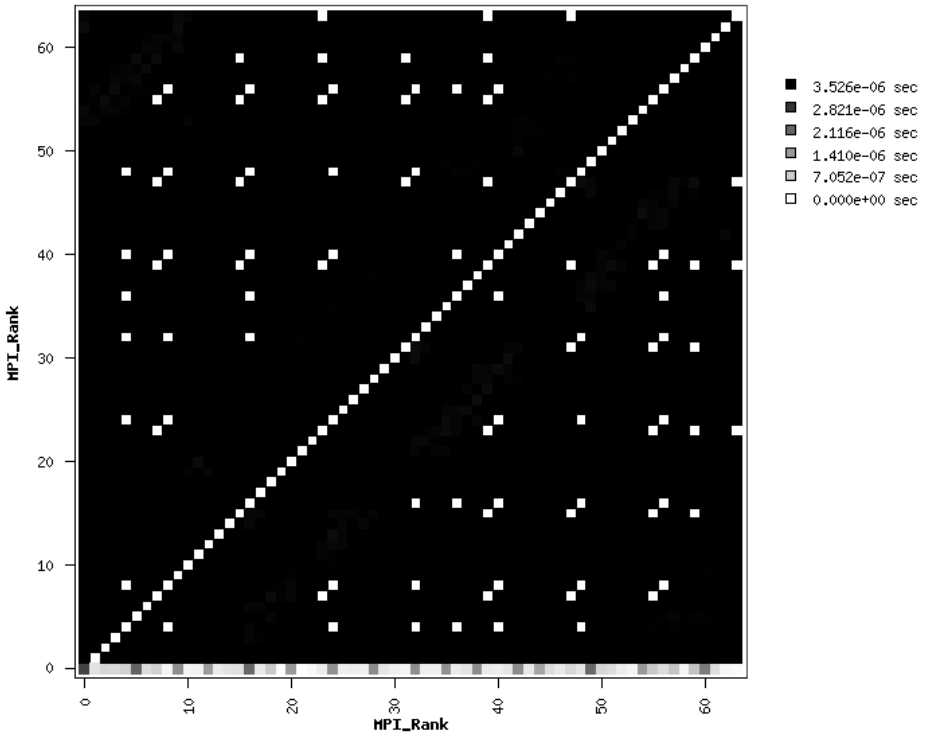


FDR InfiniBand

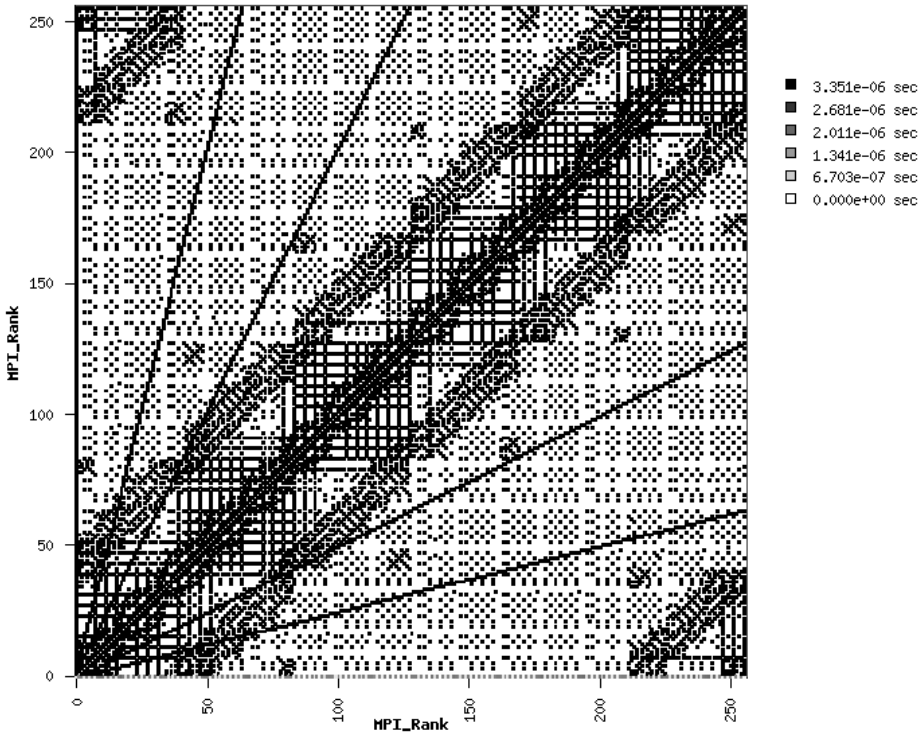
NAMD Profiling – Time Spent Per MPI Rank

- **Communication time remains relatively the same per rank as cluster scales**
  - As cluster scales, the communications pattern and time spent becomes more defined

4 Nodes



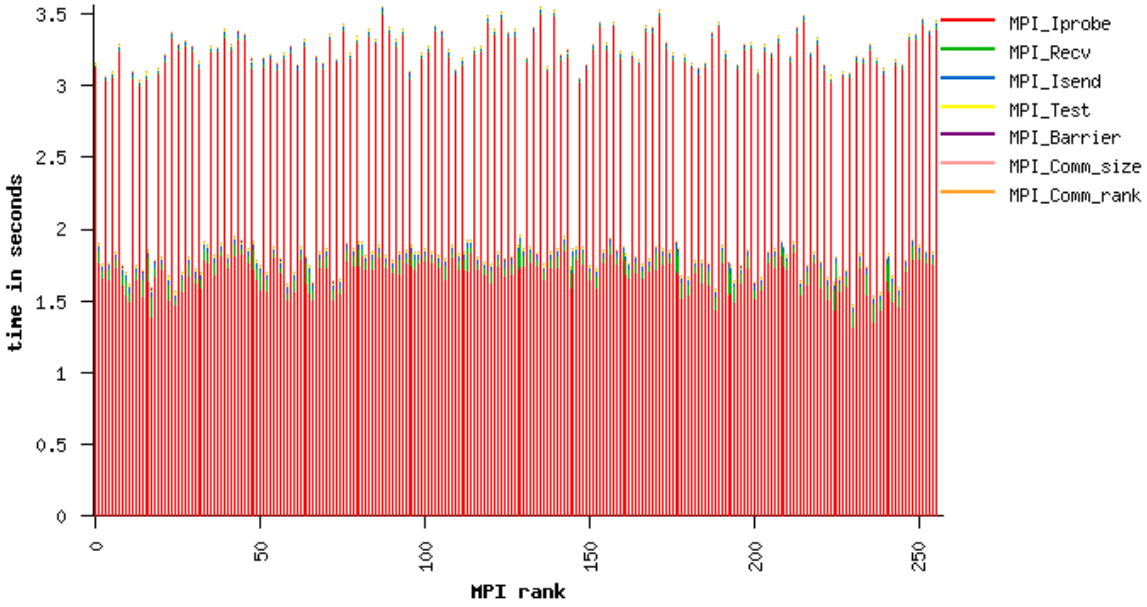
16 Nodes



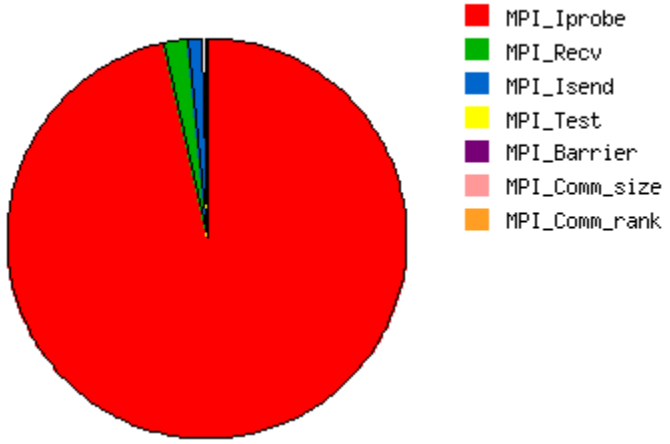
FDR InfiniBand

NAMD Profiling – Time Spent by MPI Calls

- **Majority of the MPI time is spent on MPI_Iprobe**
  - For checking of incoming messages
  - 97% of time is spend on MPI_Iprobe
- **Some variance in time can be seen in MPI time consumption**
  - MPI_Iprobe time differences contributes to this variance

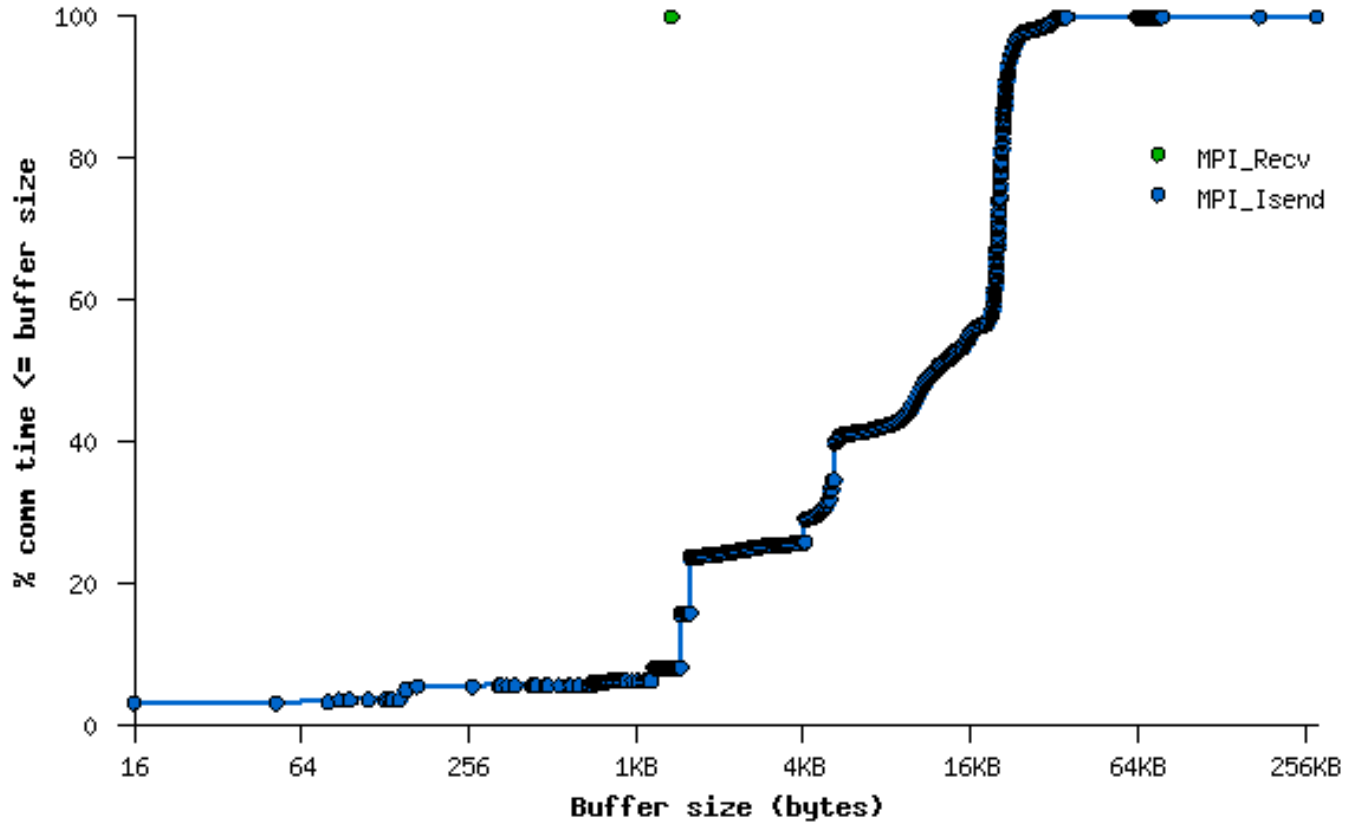


16 Nodes



NAMD Profiling – MPI Message Sizes

- **Majority of data transfer messages are medium sizes:**
  - MPI_Isend has a large concentration of 4KB to 16KB messages



NAMD – Summary

- **Performance:**
  - Intel Xeon E5-2680 on the “Jupiter” cluster and FDR InfiniBand enable NAMD to scale
  - “Jupiter”, the E5-2680 cluster performs up to 64% over “Janus” the X5670 cluster
- **Network:**
  - FDR InfiniBand allows NAMD to run at the highest network throughput at 56Gbps
  - FDR Outperforms Ethernet (10,40) by +150% at 16 nodes for NAMD
- **MPI:**
  - Both Open MPI and Intel MPI runs comparably at 16 nodes
  - The IBV interface runs 78% faster at 16 nodes than default DAPL interface in Intel MPI
- **Compilers:**
  - Both GNU compilers and Intel Composer compilers provides comparable performance
- **Profiling:**
  - Heavy MPI communications in midrange message sizes between 4KB to 16KB
  - 97% of MPI time is spent on MPI_Iprobe for checking for incoming messages

Thank You

HPC Advisory Council