

NAMD Performance Benchmark and Profiling

November 2010



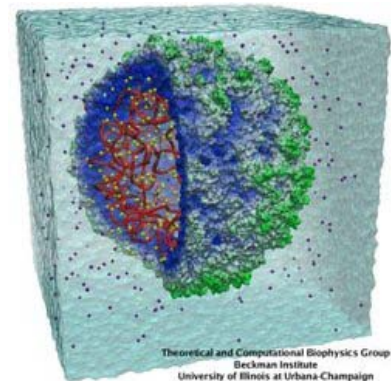
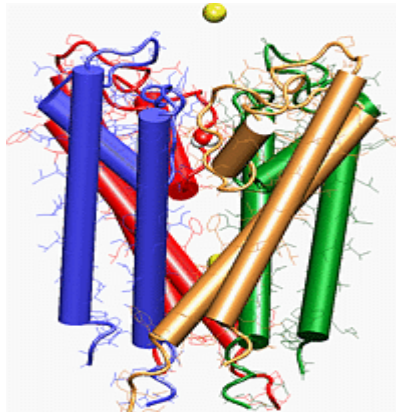
- **The following research was performed under the HPC Advisory Council activities**
 - Participating vendors: HP, Mellanox
 - Compute resource - HPC Advisory Council Cluster Center

- **For more info please refer to**
 - www.mellanox.com
 - <http://www.hp.com/go/hpc>
 - <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
- 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



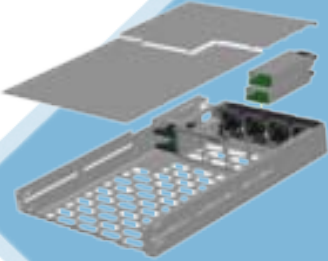
Theoretical and Computational Biophysics Group
Beckman Institute
University of Illinois at Urbana-Champaign

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

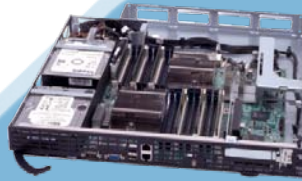
- **HP ProLiant SL2x170z G6 16-node cluster**
 - Six-Core Intel X5670 @ 2.93 GHz CPUs
 - Memory: 24GB per node
 - OS: CentOS5U4, OFED 1.5.1 InfiniBand SW stack
- **Mellanox ConnectX-2 adapters and switches**
- **Fulcrum based 10GigE switch**
- **MPI: Open MPI 1.4.1, MVAPICH-1.2.0**
- **Libraries: Charm++ 6.2.2**
- **Application: NAMD v2.7**
- **Benchmark Workload**
 - **STMV (Satellite tobacco mosaic virus) benchmark (1,066,628 atoms, periodic, PME)**

About HP ProLiant SL6000 Scalable System

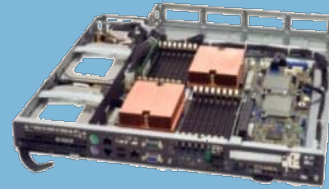
- **Solution-optimized for extreme scale out**



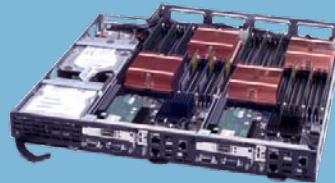
ProLiant z6000 chassis
Shared infrastructure
– fans, chassis, power



ProLiant SL160z G6 ProLiant SL165z G7
Large memory
-memory-cache apps



ProLiant SL170z G6
Large storage
-Web search and database apps



ProLiant SL2x170z G6
Highly dense
- HPC compute and
web front-end apps

Save on cost and energy -- per node, rack and data center

Mix and match configurations

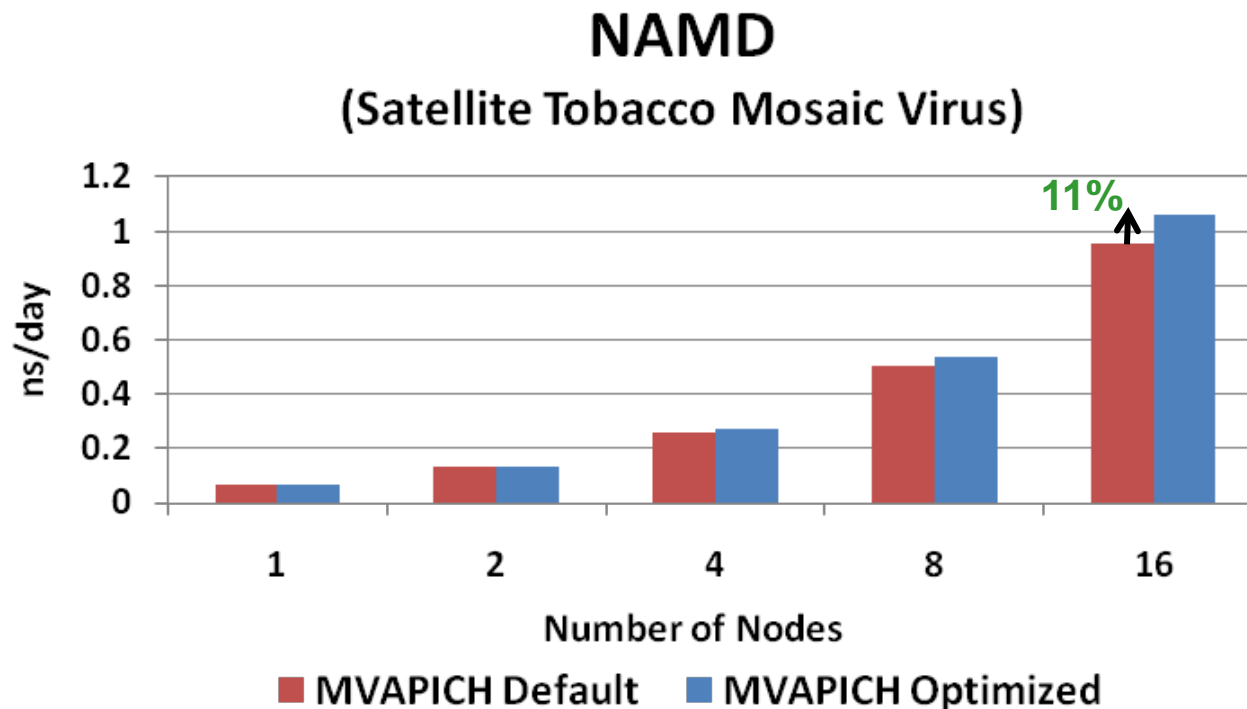
Deploy with confidence



#1
Power
Efficiency*

* SPECpower_ssj2008
www.spec.org
17 June 2010, 13:28

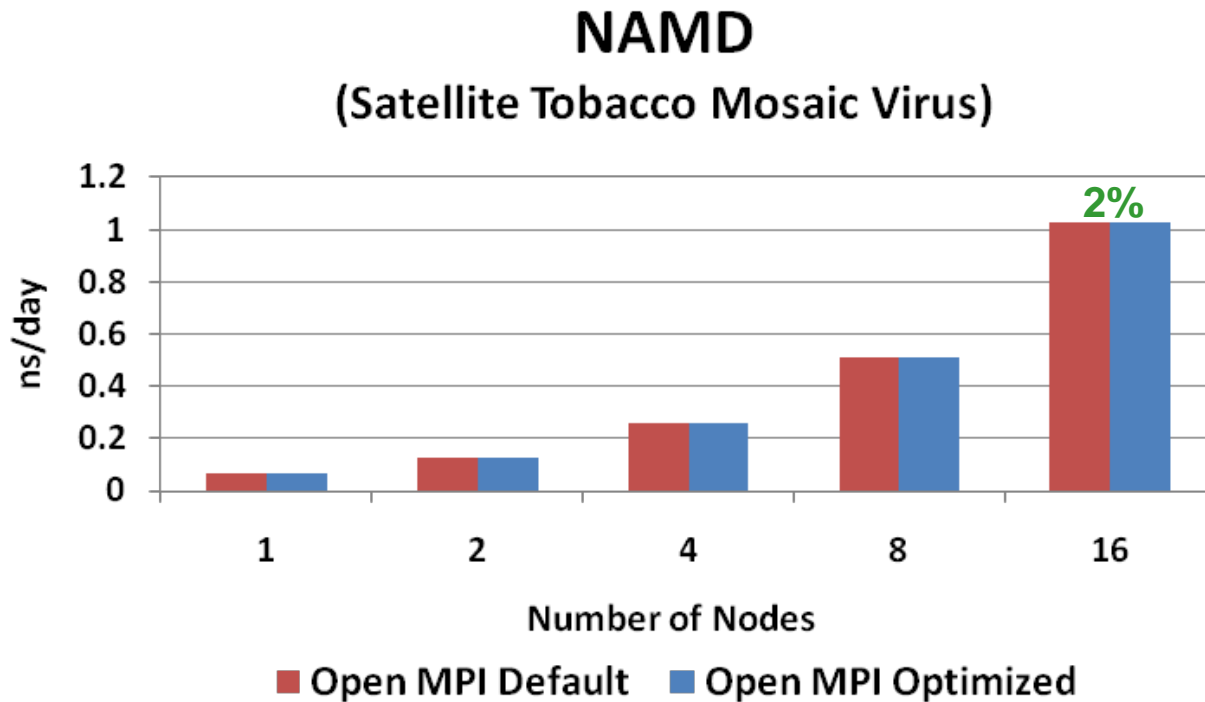
- **Performance improved by up to 11% with optimization**
 - VIADEV_RENDEZVOUS_THRESHOLD=50000
 - Advantage extends as node count increases



Higher is better

12 Cores/Node

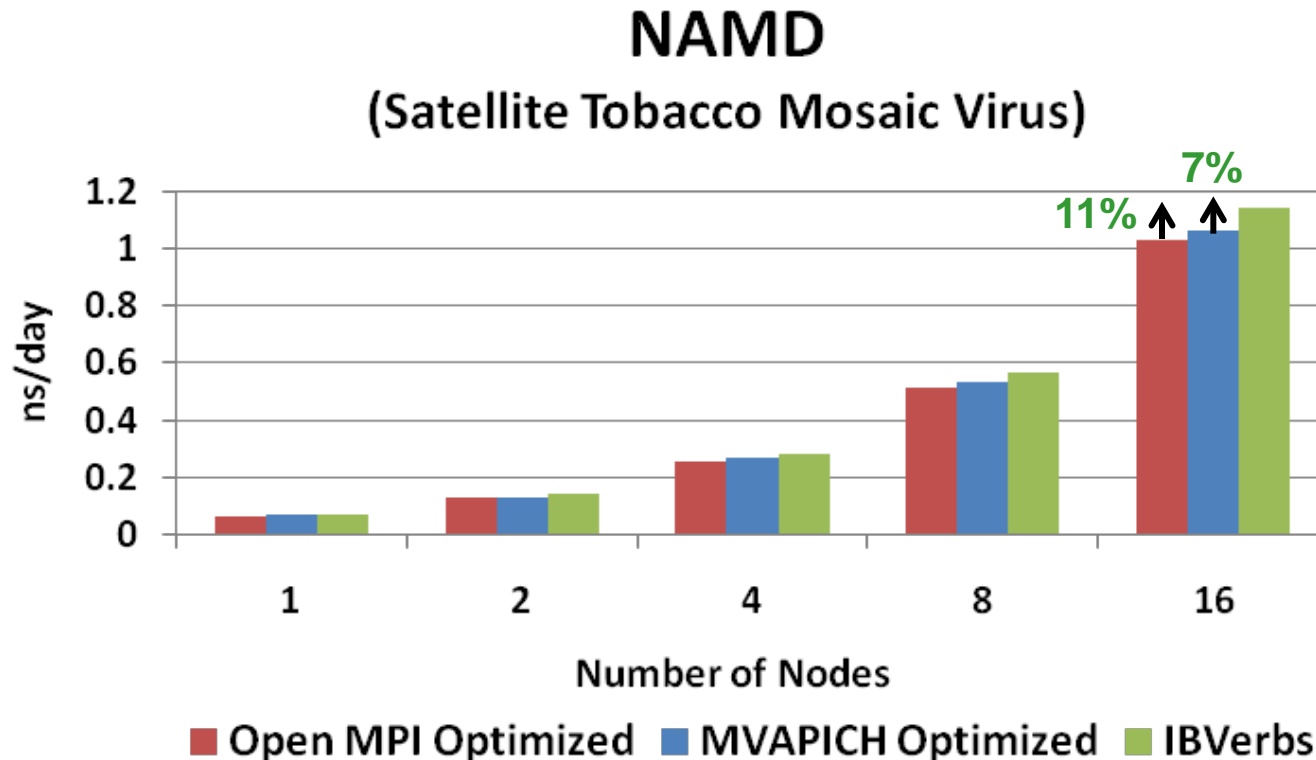
- **Performance doesn't change much with different Rendezvous value**
 - 2% performance gain compared to default mode



Higher is better

12 Cores/Node

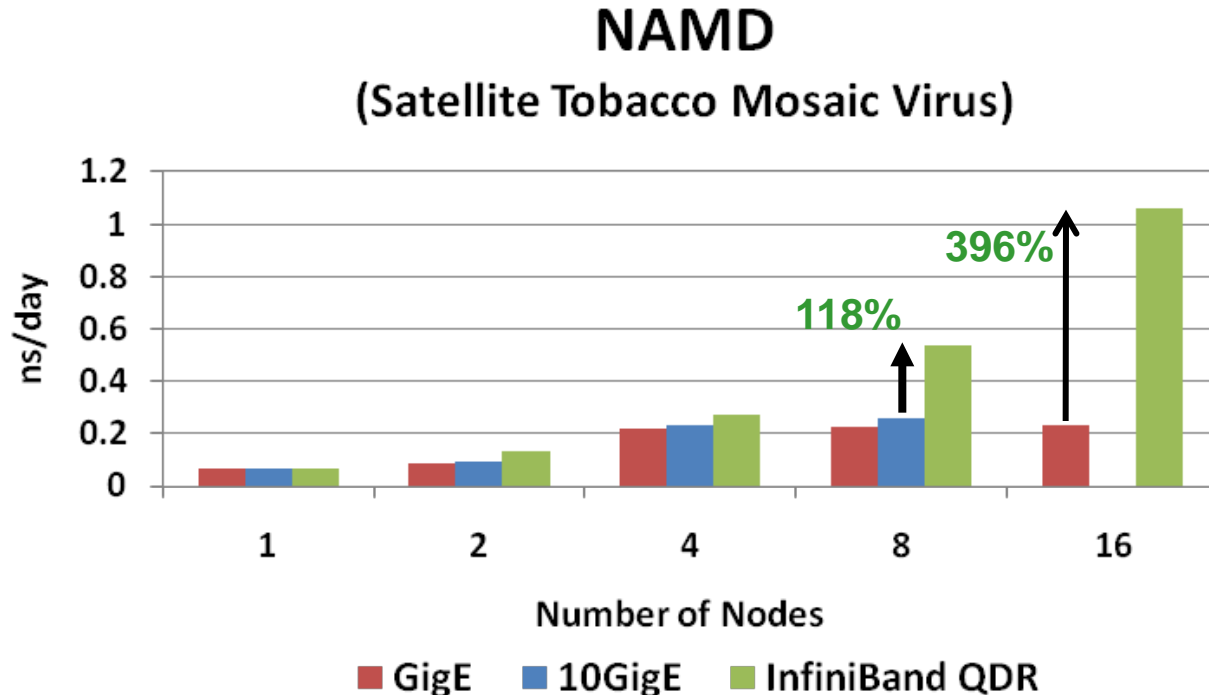
- **Native IBverbs enables highest performance**
 - 11% higher than Open MPI and 7% higher than optimized MVAPICH at 16 nodes
 - Performance advantage increases as cluster scales



Higher is better

12 Cores/Node

- **InfiniBand QDR enables linear scalability**
 - 396% higher performance than GigE at 16 nodes
 - 118% higher performance than 10GigE at 8 nodes
 - Performance advantage extends as cluster size increases
- **InfiniBand reduces electrical energy/job**
 - by 80% or more compared to GigE
 - by 54% or more compared to 10GigE



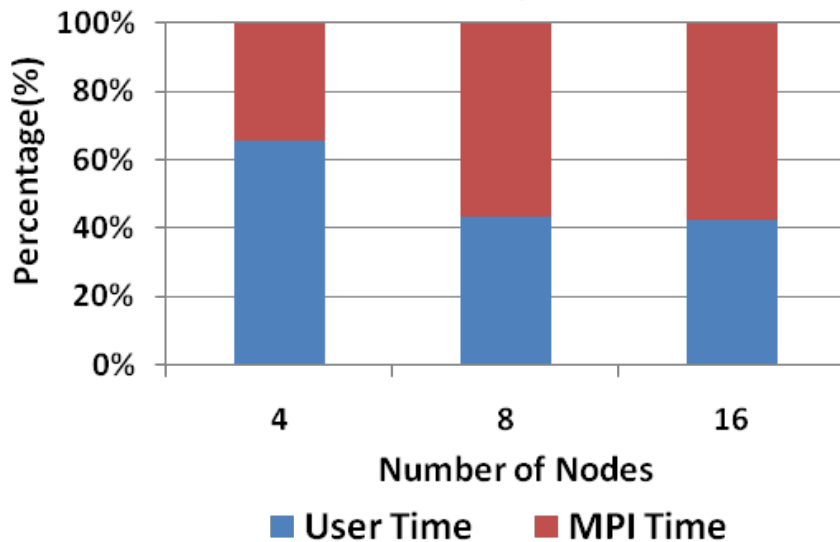
Higher is better

12 Cores/Node

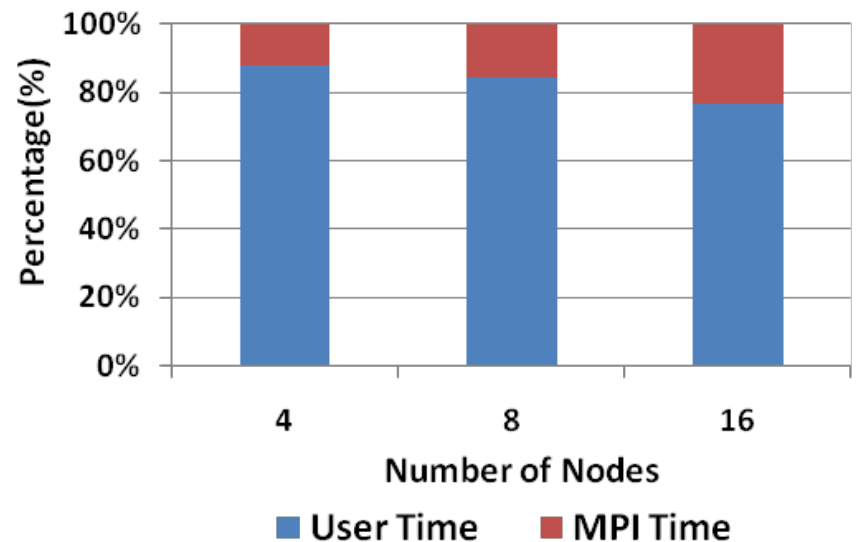
- **Runtime Distribution**

- MPI overhead becomes dominated with GigE as node count increases
- InfiniBand enables much smaller MPI communication overhead comparing to GigE

Runtime Distribution (GigE)



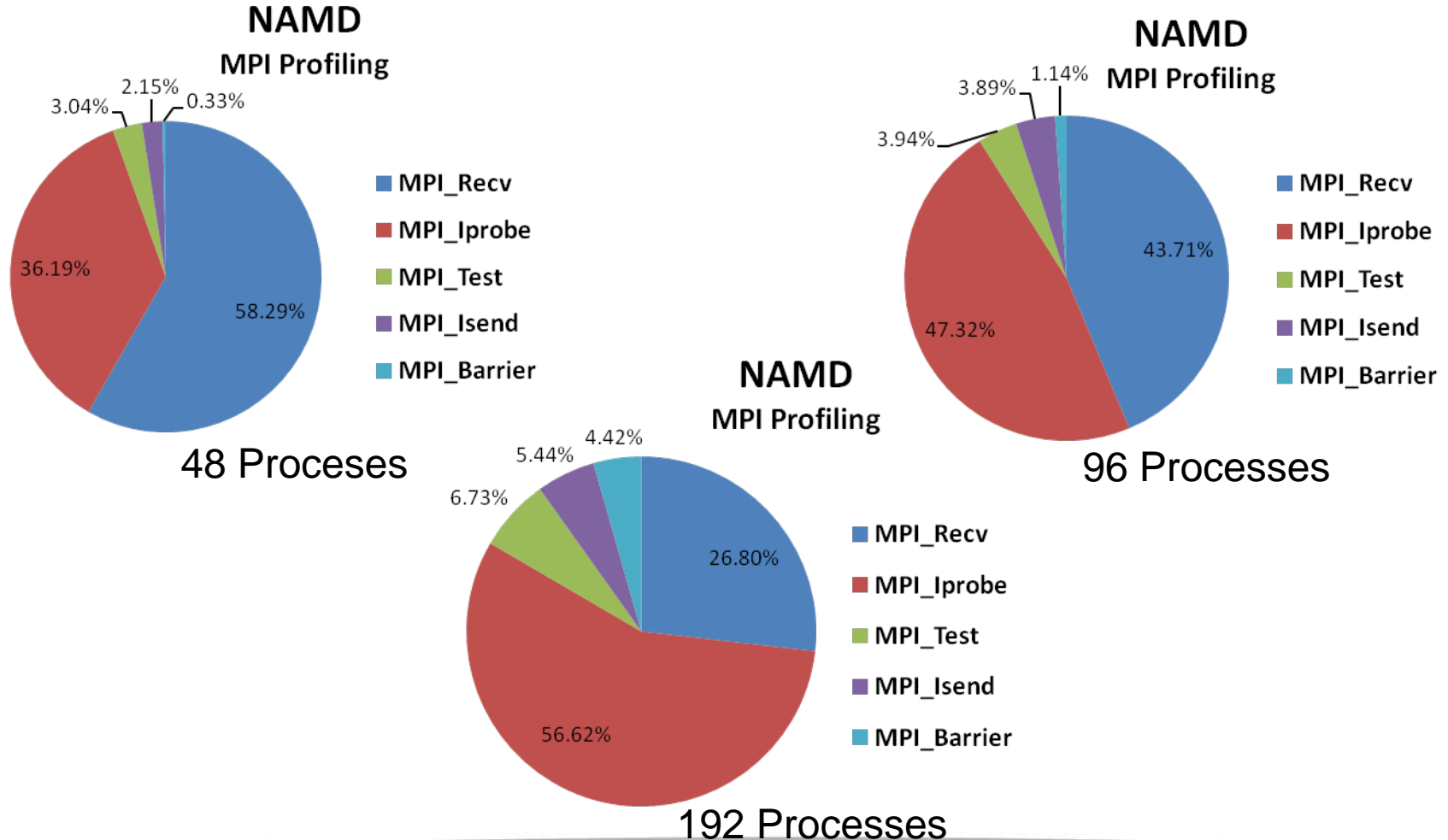
Runtime Distribution (InfiniBand QDR)



12 Cores/Node

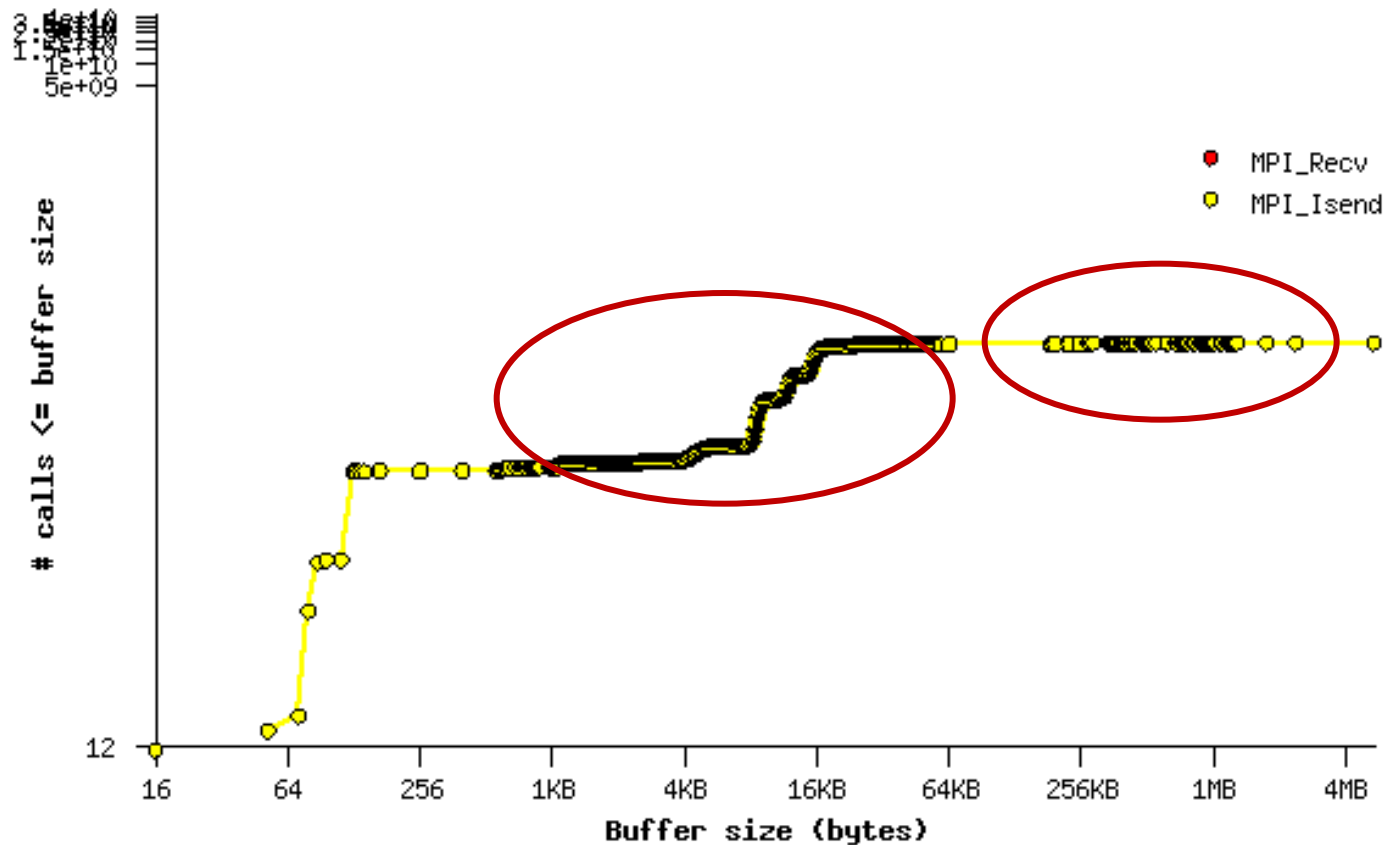
- **Mostly used MPI functions**

- MPI_Iprobe and MPI_Recv create large overhead
- MPI_Iprobe overhead increases faster than others as cluster size scales



NAMD MPI Profiling - MPI Message Size

- **Most message size**
 - 512Bytes to 64KB, and 256KB to 2MB



- **Interconnect comparison shows**
 - InfiniBand delivers superior performance in every cluster size
 - Low latency InfiniBand enables much higher scalability than 10GigE and GigE
- **Communication libraries comparison shows**
 - MVAPICH tuning enables 11% higher performance than default setting
 - Native IBverbs delivers 7% or higher performance than MPI library
 - Performance gains further in both cases as cluster size grows
- **InfiniBand QDR saves power**
 - Reduces power consumption/job by
 - 80% comparing to GigE at 16 nodes (more saving expected at higher node count)
 - 54% comparing to 10GigE at 8 nodes (more saving expected at higher node count)
- **MPI Profiling shows both interconnect bandwidth and latency are important to enable NAMD performance and scalability**

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