

Scheduling Strategies for HPC as a Service (HPCaaS) for Bio-Science Applications

Sep 2009

Gilad Shainer, Tong Liu (Mellanox); Jeffrey Layton (Dell); Joshua Mora (AMD)

High Performance Interconnects for Distributed Computing (HPI-DC)



IEEE
Cluster 2009

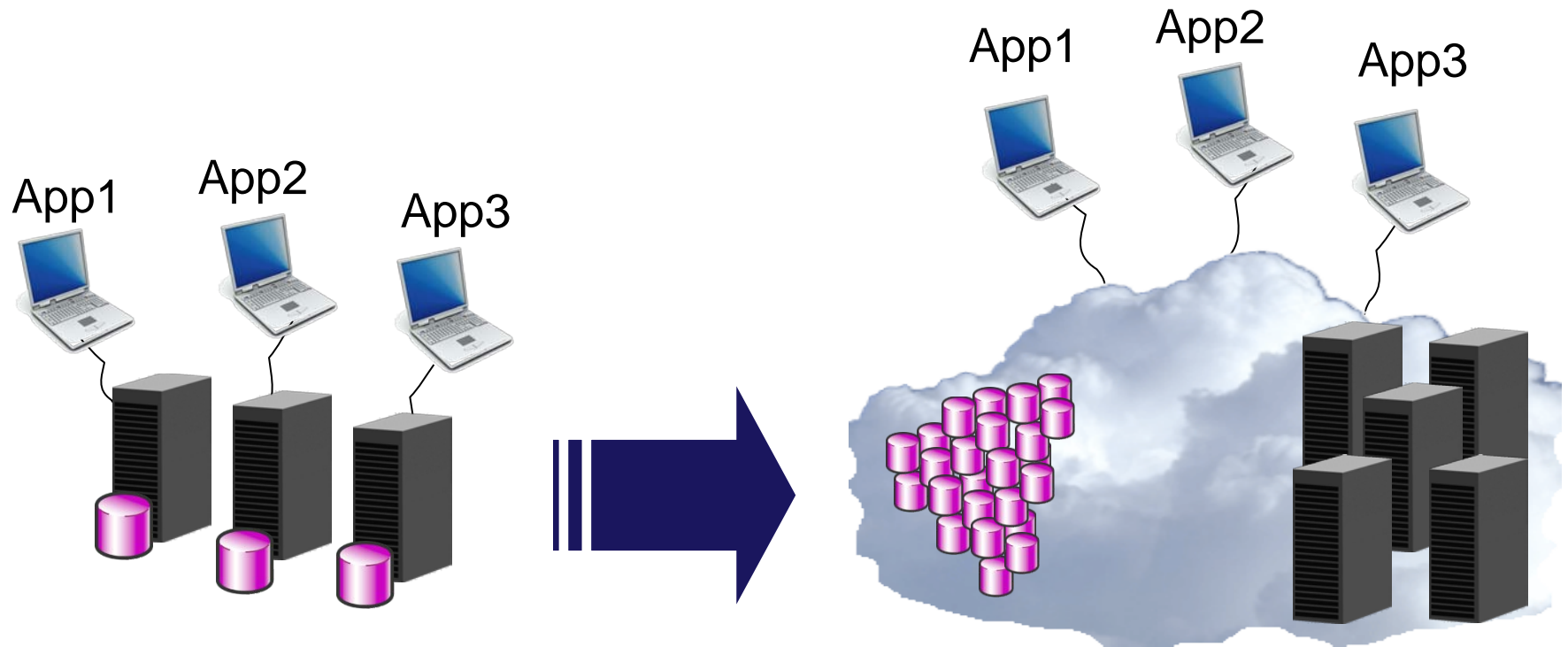
New Orleans,
Louisiana

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



- **WW HPC organization (90 members)**
 - Bridges the gap between HPC usage and its potential
 - Provides best practices and a support/development center
 - Explores future technologies and future developments
 - Explores advanced topics – HPC in a cloud, HPCaaS etc.
 - Leading edge solutions and technology demonstrations
- **For more info: <http://www.hpcadvisorycouncil.com>**



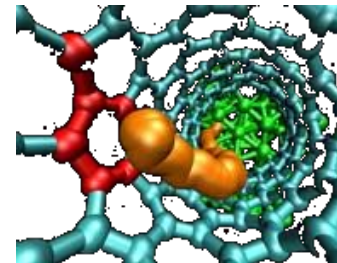
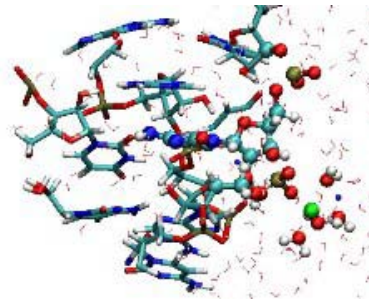
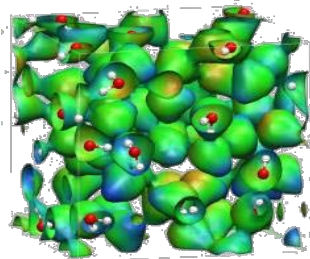


- From equipment warehouse to service provider
- From dedicated HW per application to application services
- Higher productivity, simplicity and efficiency

- **Investigate HPC as a Service for bio-science applications**
 - In particular NAMD and CPMD
- **Performance and productivity impact**
- **Cluster interconnect effects on applications performance**

- **For specific information on NAMD and CPMD, please refer to additional studies at**
 - http://hpcadvisorycouncil.mellanox.com/best_practices.php

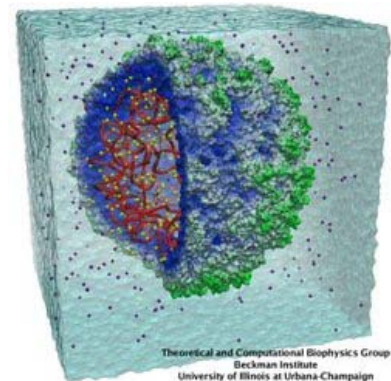
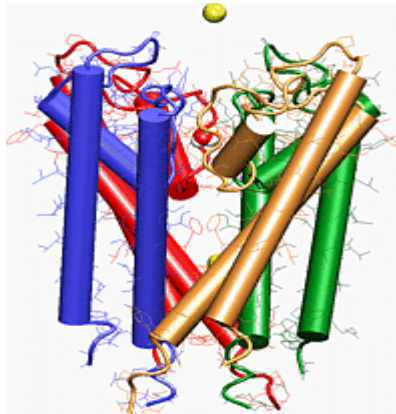
- **A parallelized implementation of density functional theory (DFT)**
- **Particularly designed for ab-initio molecular dynamics**
- **Brings together methods**
 - Classical molecular dynamics
 - Solid state physics
 - Quantum chemistry
- **CPMD supports MPI and Mixed MPI/SMP**
- **CPMD is distributed and developed by the CPMD consortium**



- **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

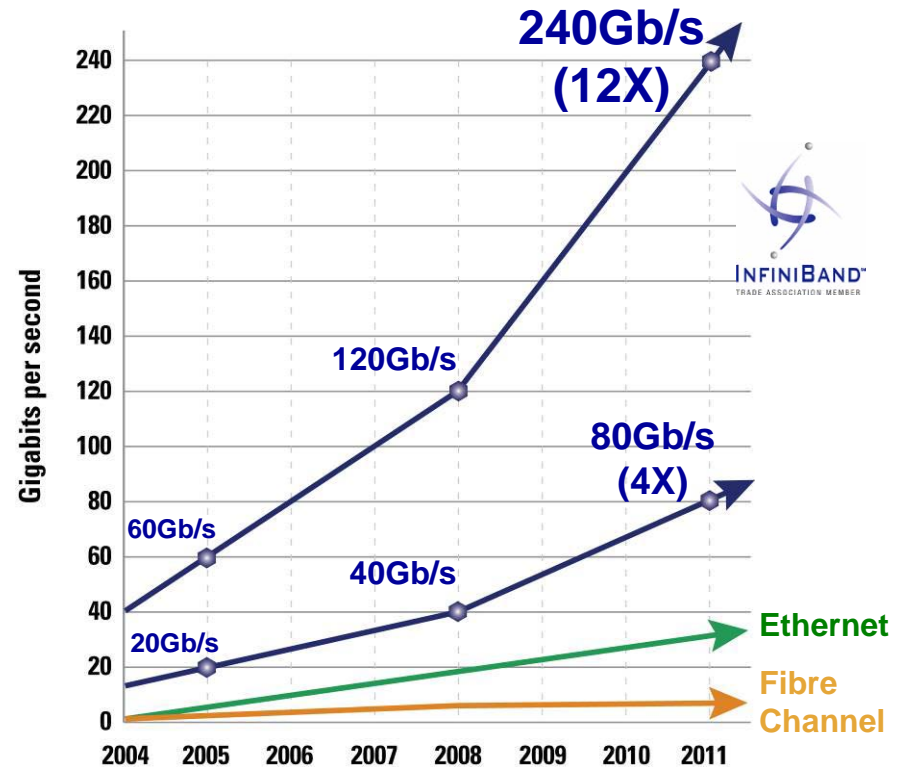


Theoretical and Computational Biophysics Group
Beckman Institute
University of Illinois at Urbana-Champaign

- **Dell™ PowerEdge™ SC 1435 24-node cluster**
- **Quad-Core AMD Opteron™ 2382 (“Shanghai”) CPUs**
- **Mellanox® InfiniBand ConnectX® DDR HCAs and InfiniBand DDR Switch**
- **Memory: 16GB memory, DDR2 800MHz per node**
- **OS: RHEL5U2, OFED 1.3 InfiniBand SW stack**
- **MPI: Open MPI 1.3, Platform MPI 5.6.4**
- **Compiler: GCC 4.2.0**
- **Benchmark Application:**
 - CPMD 3.13
 - Benchmark Dataset: C120
 - NAMD 2.6 with fftw3 libraries and Charm++ 6.0
 - Benchmark Dataset: **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



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

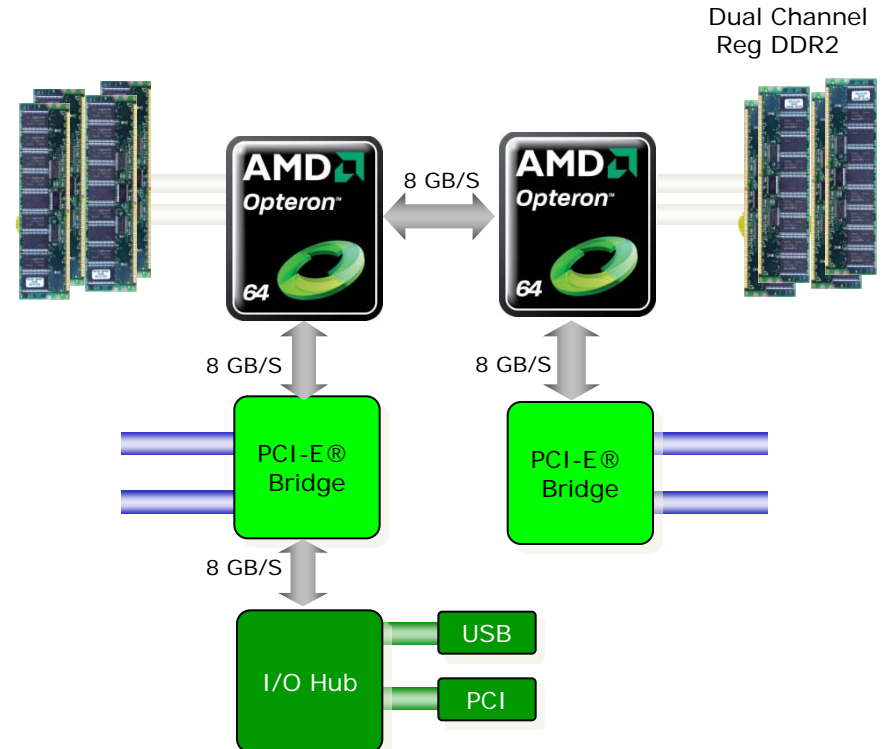
- Up to 12.8 GB/s
- DDR2-800 MHz or DDR2-667 MHz

- **Scalability**

- 48-bit Physical Addressing

- **Compatibility**

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



- **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**

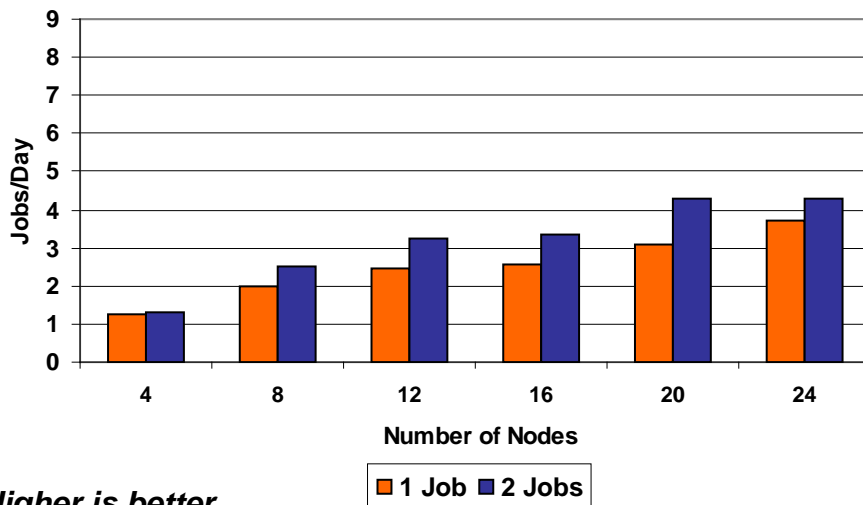
- Optimized System Size, Configuration and Workloads
- Test-bed Benchmarks
- ISV Applications Characterization
- Best Practices & Usage Analysis



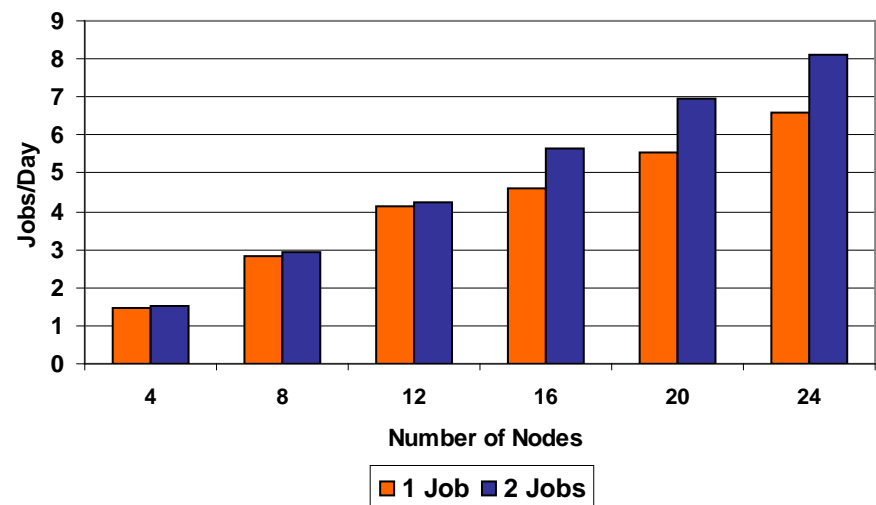
NAMD Benchmark Results – Productivity

- **Case 1: Dedicated hardware resource for NAMD**
- **Input Date: ApoA1**
 - Benchmark comprises 92K atoms of lipid, protein, and water
 - Models a bloodstream lipoprotein particle
 - One of the most used data sets for benchmarking NAMD
- **Increasing number of concurrent jobs increases cluster productivity**
- **InfiniBand enables higher performance and productivity**

NAMD
(ApoA1) - GigE



NAMD
(ApoA1) - InfiniBand DDR

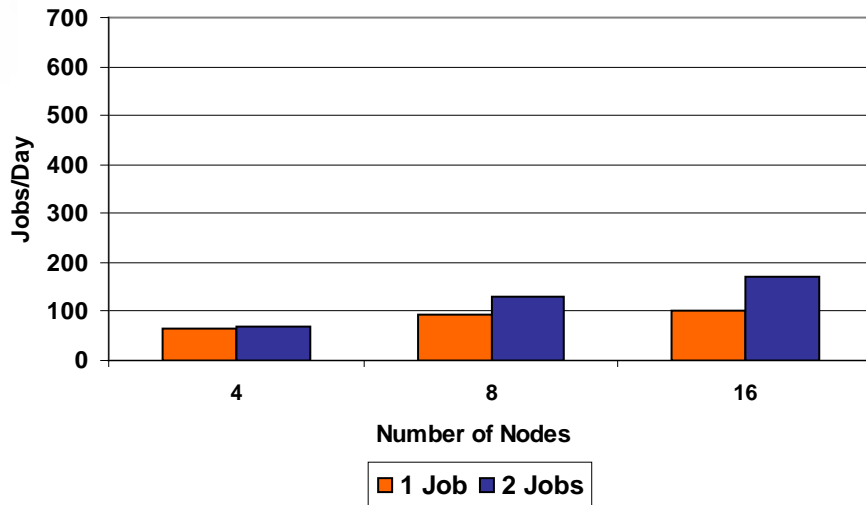


Higher is better

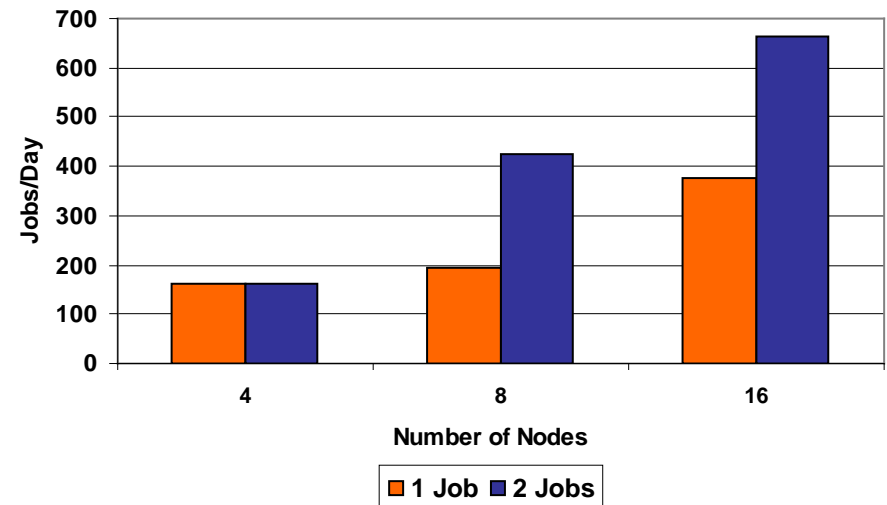
CPMD Benchmark Results – Productivity

- **Case 2: Dedicated hardware resource for CPMD**
- **Benchmark Data: C₁₂₀ - 120 carbon atoms**
- **Running two jobs in parallel increases cluster productivity**
- **InfiniBand enables higher performance and scalability than GigE**

CPMD
C₁₂₀ - GigE

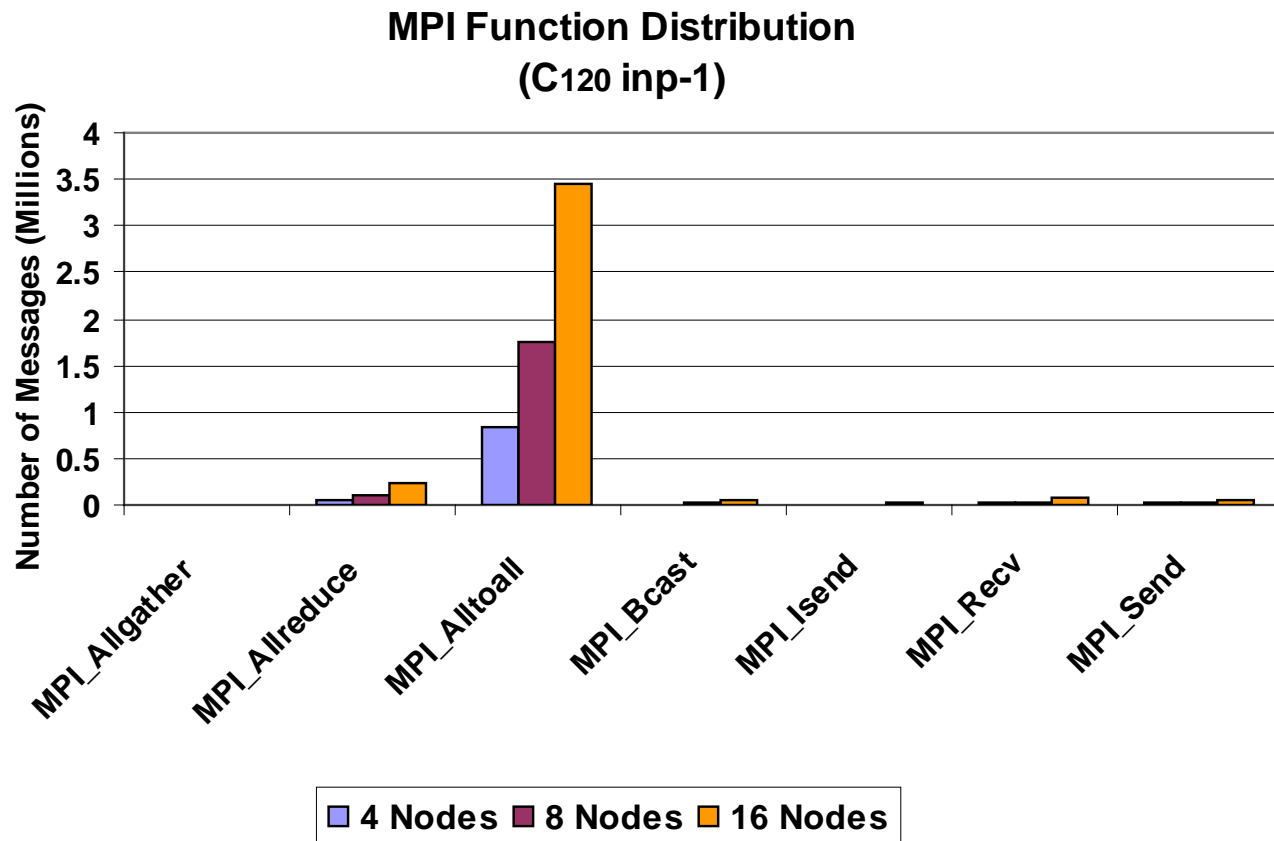


CPMD
C₁₂₀ - InfiniBand DDR



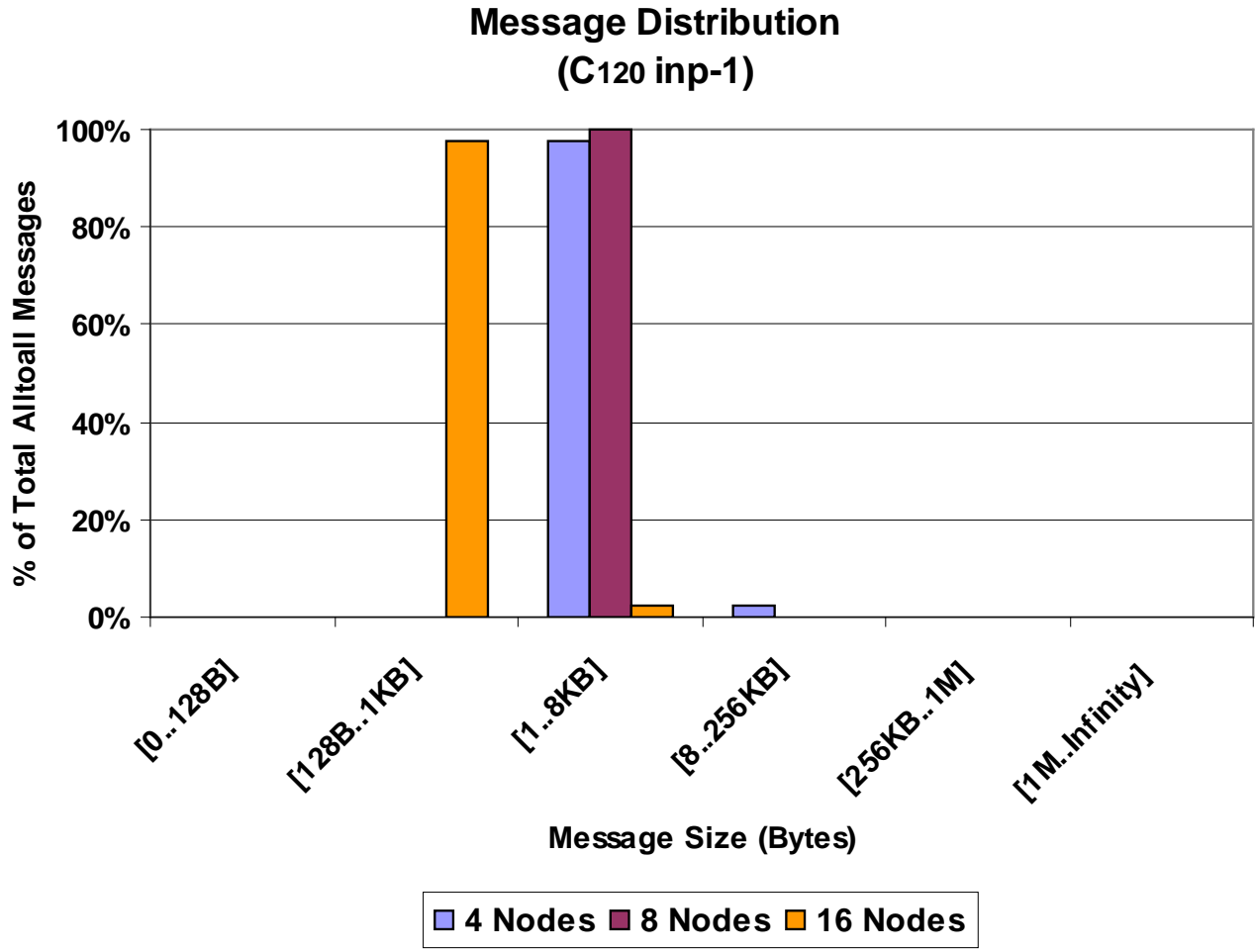
Higher is better

- **MPI_AlltoAll is the key collective function in CPMD**
 - Number of AlltoAll messages increases dramatically with cluster size



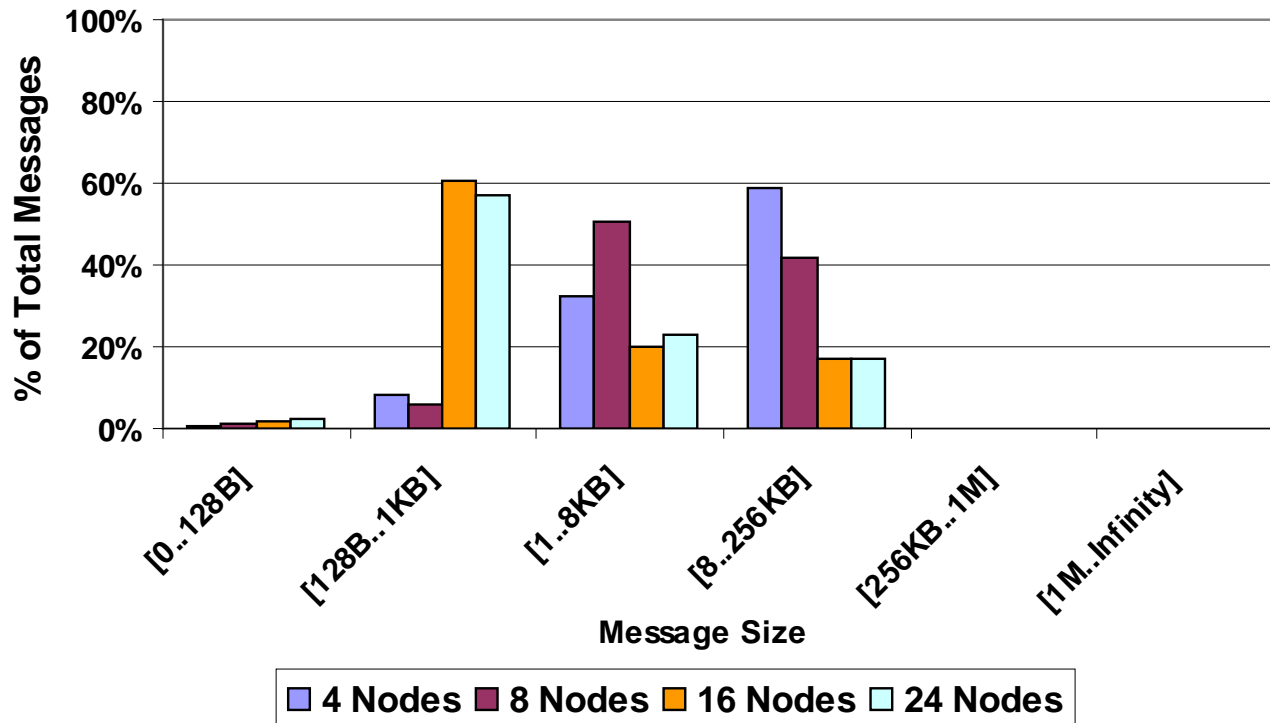
CPMD MPI Message Size Distribution

- Majority messages are medium size



- 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)

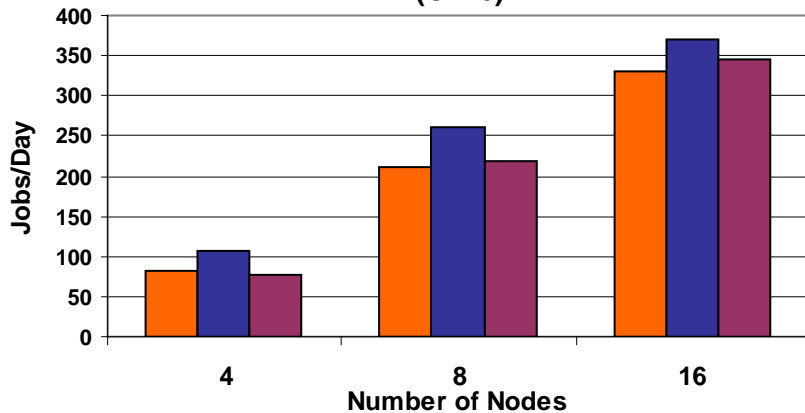


- **Case 3 – HPC as a Service (HPCaaS)**
- **HW platform to serve multiple applications at the same time**
 - CPMD and NAMD
- **Multiple test scenarios will be presented in the following slides**
 - Each describes different allocation methods of the HW system per service
 - Service refers to a single application
 - At least 2 applications will be served at a given time
 - Each scenario will be compared to a dedicated HW per applications approach
 - Evaluation metric: productivity (number of jobs per day)

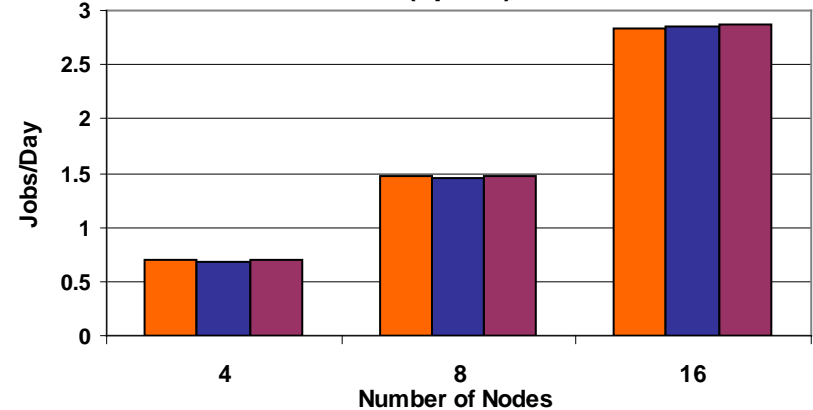
Multiple Applications – CPMD and NAMD

- **Test Scenario:**
 - Single Application approach:
 - Two NAMD jobs in parallel for half day then two CPMD jobs for the other half day
 - Multiple Applications approach
 - One CPMD job and one NAMD job simultaneously on the cluster for a full day
 - Case I: 4 cores for each application (2 cores on each CPU)
 - Case II: One application per CPU socket
- **Running CPMD and NAMD in parallel improves CPMD productivity**
- **Distributing CPMD processes to two sockets has better performance**
 - Versus using an entire CPU (socket) per applications
- **NAMD shows negligible productivity difference under the three scenarios**

CPMD Application Productivity (C120)



NAMD Application Productivity (ApoA1)



■ Single Application ■ Multiple Applications I ■ Multiple Applications II

■ Single Application ■ Multiple Applications I ■ Multiple Applications II

Multiple Applications – CPMD + NAMD

- **Test Scenario:**

- Single Application

- Two NAMD jobs in parallel for 3/4 day then two CPMD jobs for 1/4 day

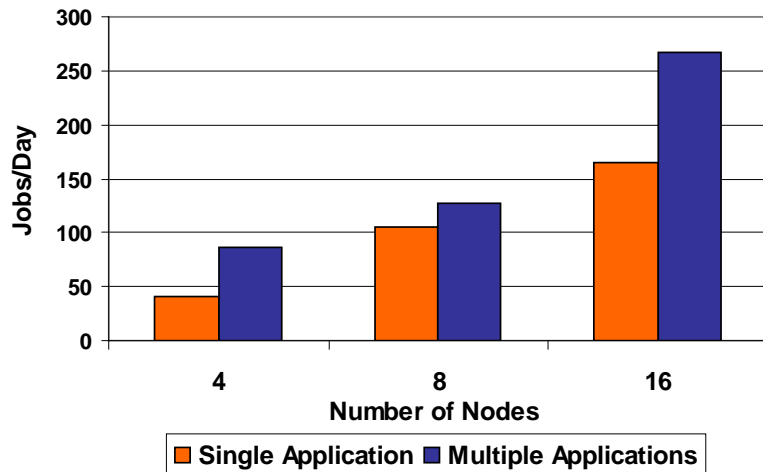
- Multiple Applications

- One CPMD job and one NAMD job simultaneously on the cluster for a full day
- 6 cores for NAMD (3 cores on each CPU), 2 cores for CPMD (1 core on each CPU)

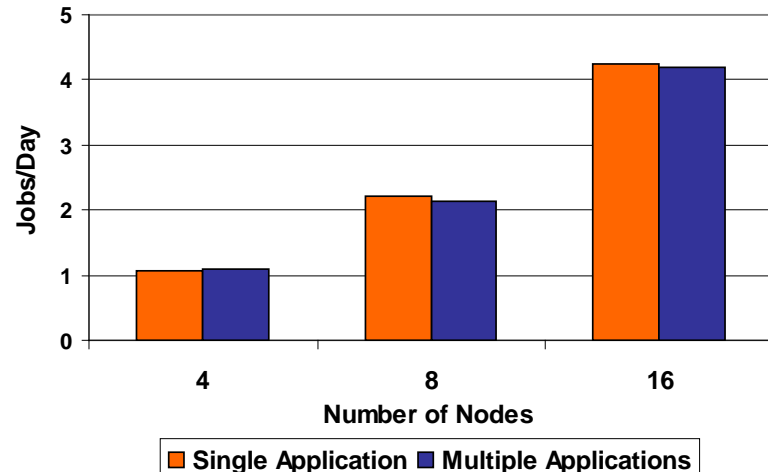
- **Running CPMD and NAMD in parallel improves CPMD productivity by up to 61%**

- **NAMD shows negligible productivity difference under the two scenarios**

CPMD Application Productivity (C120)



NAMD Application Productivity (ApoA1)



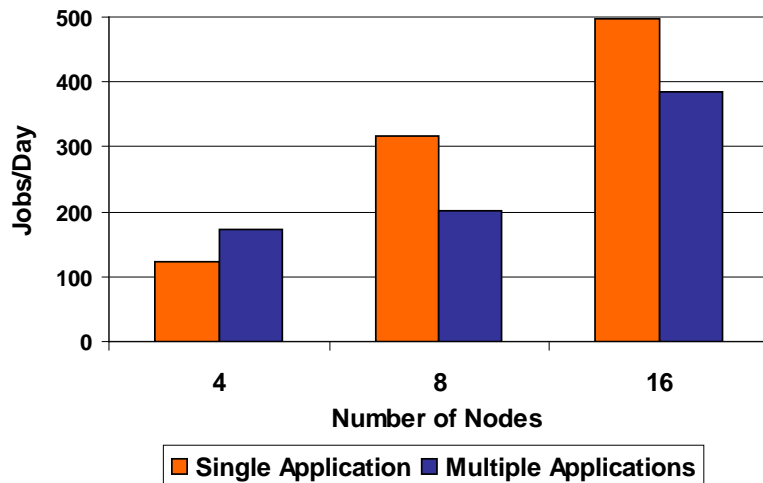
Higher is better

InfiniBand DDR

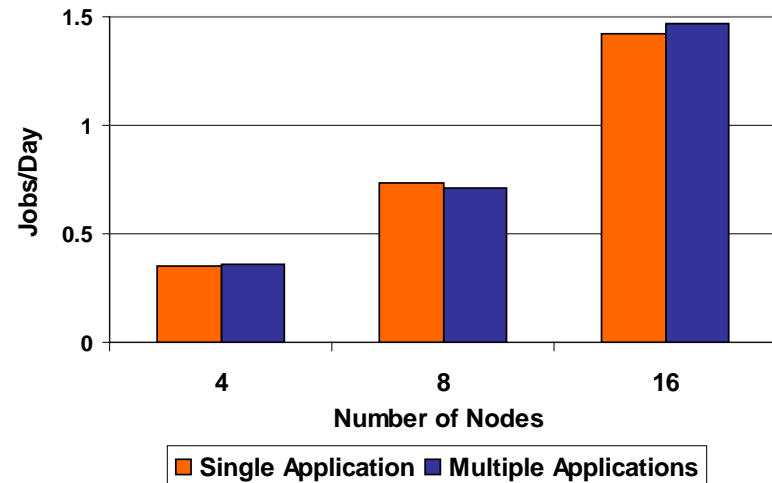
Multiple Applications – CPMD + NAMD

- **Test Scenario:**
 - Single Application
 - Two NAMD jobs in parallel for 1/4 day then two CPMD jobs for 3/4 day
 - Multiple Applications
 - One CPMD job and one NAMD job simultaneously on the cluster for a full day
 - 2 cores for NAMD (1 core on each CPU), 6 cores for CPMD (3 cores on each CPU)
- **Running CPMD with less cores decreases CPMD productivity**
- **NAMD shows negligible productivity difference under the two scenarios**

CPMD Application Productivity (C120)



NAMD Application Productivity (ApoA1)



Higher is better

InfiniBand DDR

- **NAMD**
 - Increase number of jobs running on the each node improves productivity
 - InfiniBand provides nearly doubled performance of GigE
 - GigE does not scale beyond 20 nodes
- **CPMD**
 - Higher productivity is gained with 2 parallel CPMD jobs on the cluster
 - InfiniBand delivers up to 300% higher productivity vs GigE
- **CPMD and NAMD simultaneously – HPC as a Service (HPCaaS)**
 - It is feasible and productive to run CPMD and NAMD simultaneously on a single system
 - When enough core were allocated, CPMD productivity was increased
 - NAMD demonstrates same level of productivity
 - NAMD consumes large portion of the systems resources
 - Having more than a single NAMD and CPMD jobs will not increase productivity

- **HPC as a Service enables greater systems flexibility**
 - Eliminates the need for dedicated HW resources per applications
 - Simplifies usage models
 - Enables dynamic allocation per given task
- **Effectively model needs to take into consideration**
 - Applications sensitivity points and applications bottlenecks
 - Minimum HW resource requirements per applications
 - Matching up applications with different hardware requirements
- **HPC as a Service for Bio-science applications (CMPD and NAMD)**
 - Enables increased or equal productivity versus dedicated HW resource
 - Method: allocation of 4 cores or less for CPMD, 4 cores or more for NAMD
 - Cores per application allocation – using both sockets demonstrate higher productivity
 - Better allocation of cores, memory and interconnect resources to minimize contention
 - NAMD requires more or equal compute resources than CPMD

Thank You

HPC Advisory Council

HPC Advisory Council Workshop

October 28th, 2009

Changsha, Hunan, China