

# HPC Applications Scalability

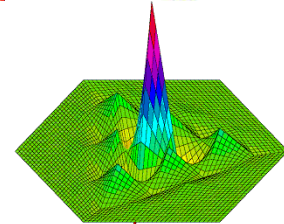
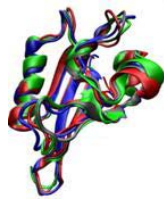
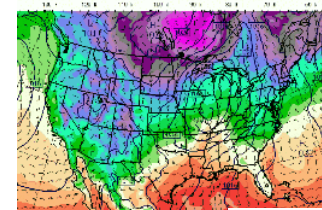
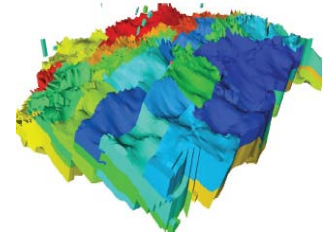
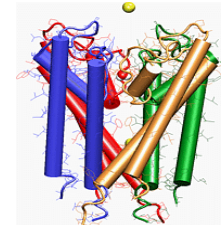
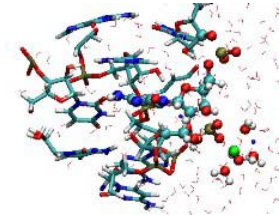
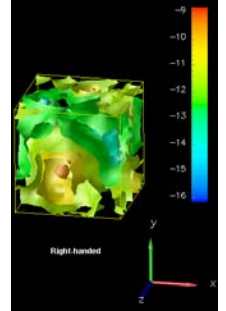
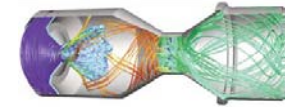
[Gilad@hpcadvisorycouncil.com](mailto:Gilad@hpcadvisorycouncil.com)

HPC Advisory Council  
Switzerland Workshop 2010

March 15-17, 2010  
Lugano Convention Centre, Switzerland

# Applications Best Practices

- **LAMMPS - Large-scale Atomic/Molecular Massively Parallel Simulator**
- **D. E. Shaw Research Desmond**
- **NWChem**
- **MPQC - Massively Parallel Quantum Chemistry Program**
- **ANSYS FLUENT and CFX**
- **CD-adapco STAR-CCM+**
- **CD-adapco STAR-CD**
- **MM5 - The Fifth-Generation Mesoscale Model**
- **NAMD**
- **LSTC LS-DYNA**
- **Schlumberger ECLIPSE**
- **CPMD - Car-Parrinello Molecular Dynamics**
- **WRF - Weather Research and Forecast Model**
- **Dacapo - total energy program based on density functional theory**
- **Lattice QCD**
- **Open Foam**
- **GAMESS**
- **HOMME**
- **OpenAtom**
- **PopPerf**

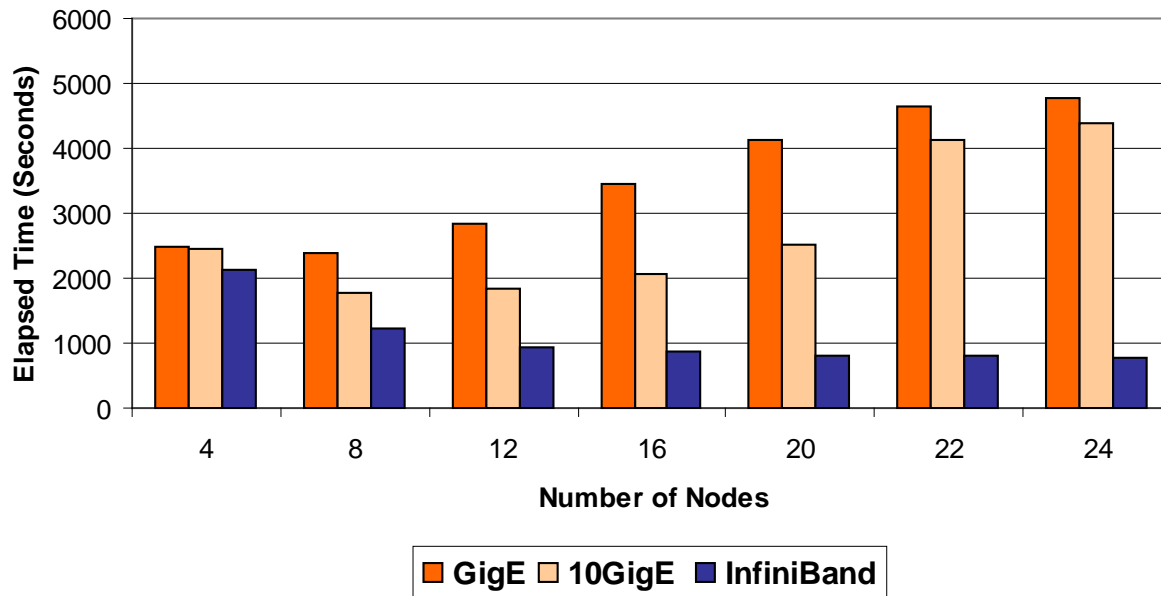


- **Performance evaluation**
- **Profiling – network, I/O**
- **Recipes (installation, debug, command lines etc)**
- **MPI libraries**
- **Power management**
- **Optimizations at small scale**
- **Optimizations at scale**

# HPC Applications Small Scale

- **InfiniBand enables highest scalability**
  - Performance accelerates with cluster size
- **Performance over GigE and 10GigE is not scaling**
  - Slowdown occurs beyond 8 nodes

**Schlumberger ECLIPSE  
(FOURMILL)**



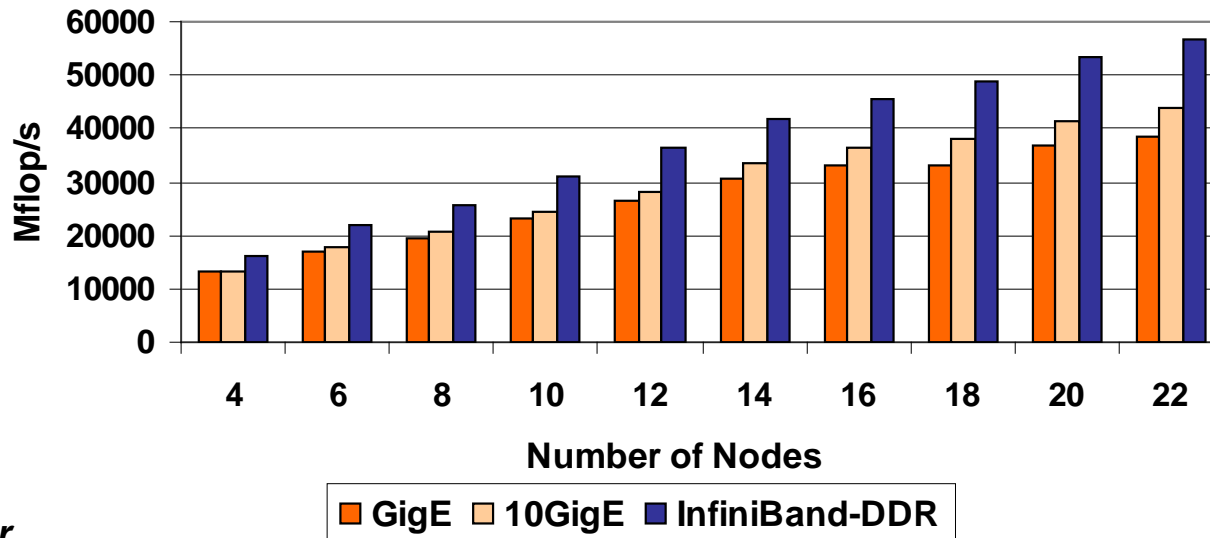
*Lower is better*

*Single job per cluster size*

# MM5 Performance Results - Interconnect

- **Input Data: T3A**
  - Resolution 36KM, grid size 112x136, 33 vertical levels
  - 81 second time-step, 3 hour forecast
- **InfiniBand DDR delivers higher performance in any cluster size**
  - Up to 46% versus GigE, 30% versus 10GigE

**MM5 Benchmark Results - T3A**



*Higher is better*

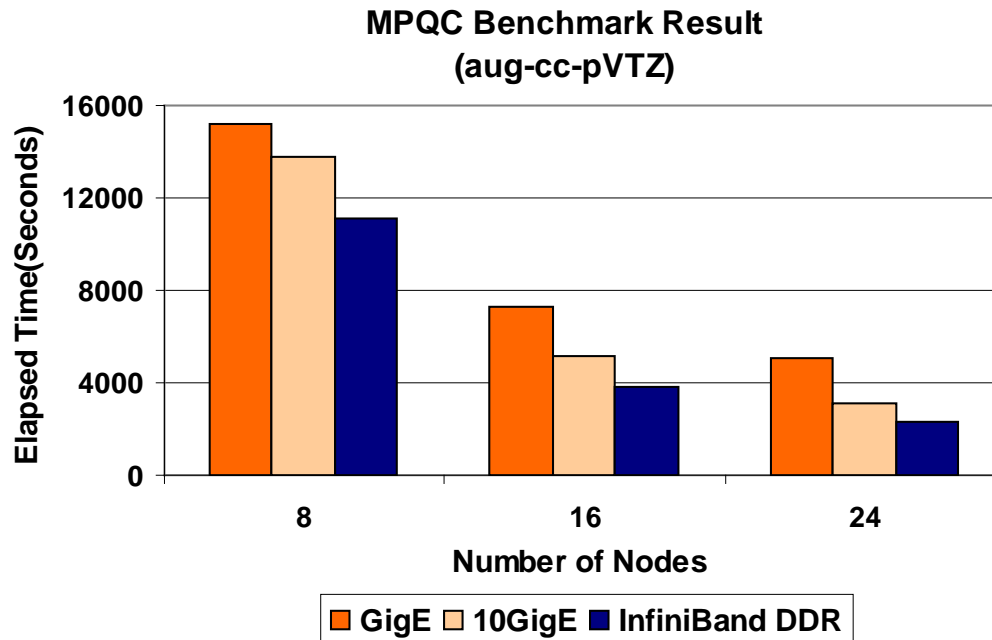
*Platform MPI*

- **Input Dataset**

- MP2 calculations of the uracil dimer binding energy using the aug-cc-pVTZ basis set

- **InfiniBand enables higher scalability**

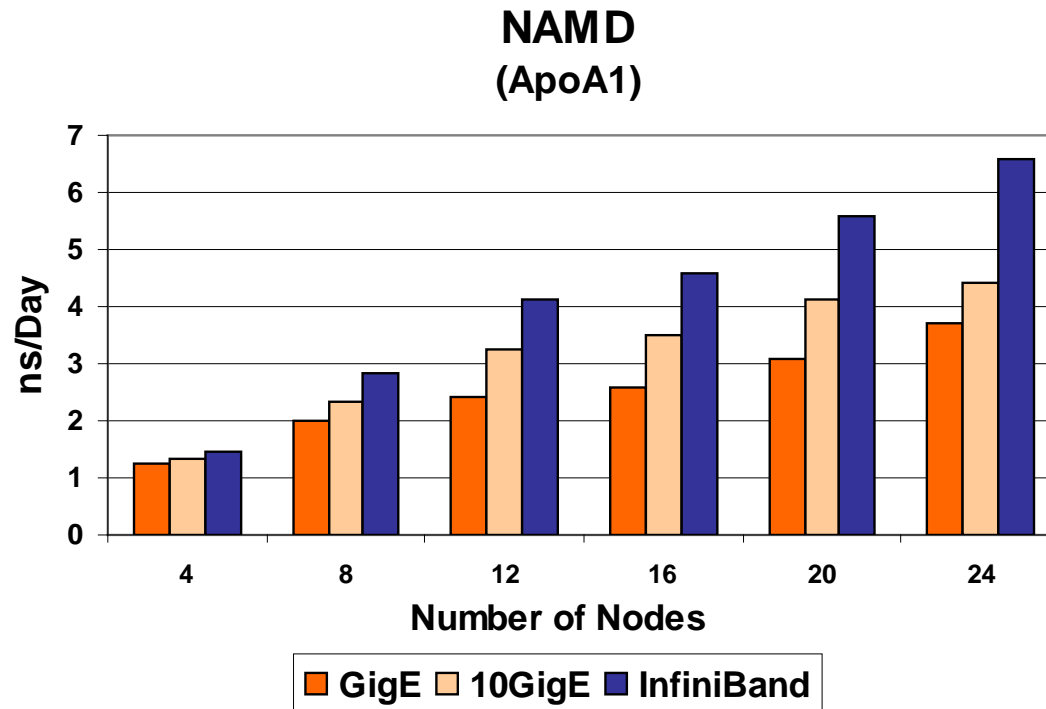
- Performance accelerates with cluster size
- Outperforms GigE by up to 124% and 10GigE by up to 38%



*Lower is better*

# NAMD Performance Results – Interconnect

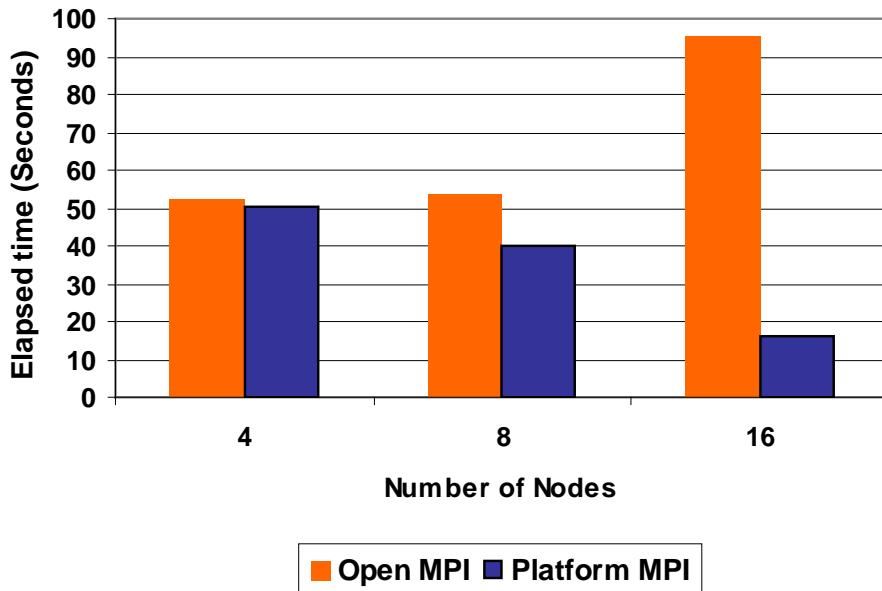
- **ApoA1 case - benchmark comprises 92K atoms of lipid, protein, and water**
  - Models a bloodstream lipoprotein particle
  - One of the most used data sets for benchmarking NAMD
- **InfiniBand 20Gb/s outperforms GigE and 10GigE in every cluster size**
  - InfiniBand provides higher performance up to 79% vs GigE and 49% vs 10GigE



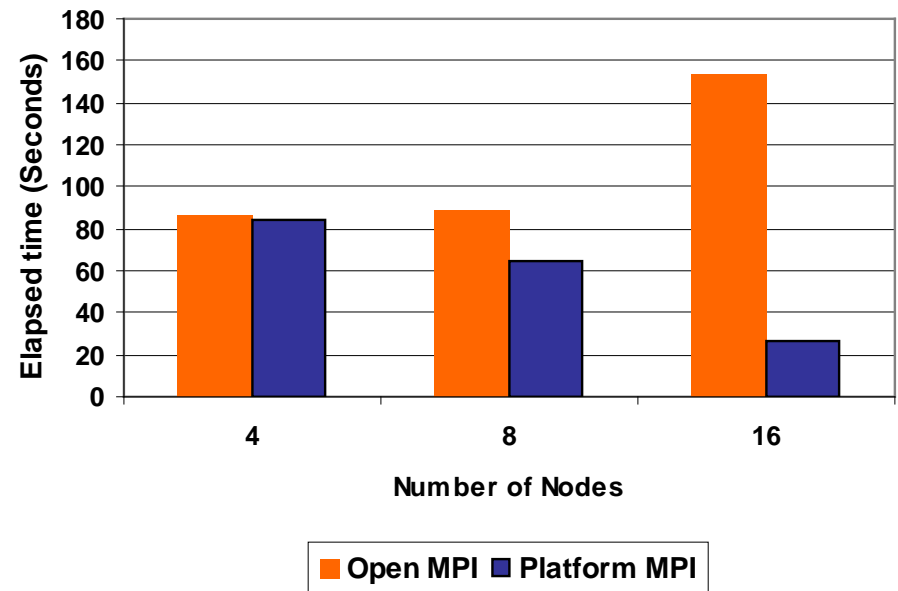
Open MPI

- Platform MPI shows better scalability over Open MPI

CPMD  
(Wat32 inp-1)



CPMD  
(Wat32 inp-2)

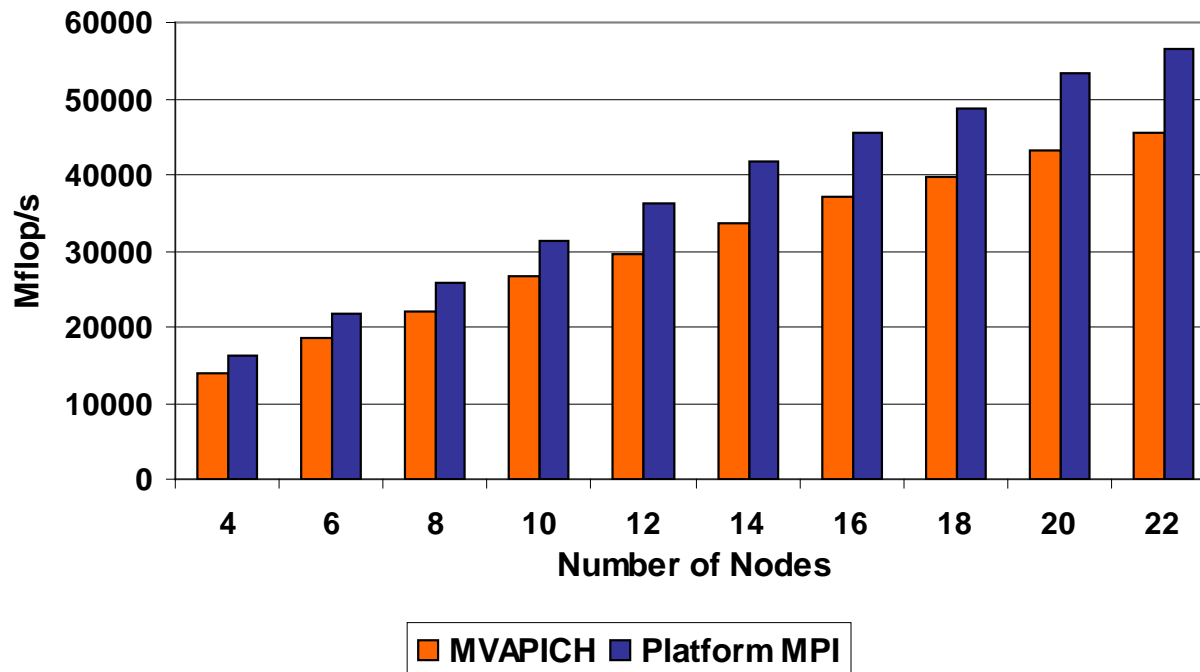


*Lower is better*

*These results are based on InfiniBand*

- **Platform MPI demonstrates higher performance versus MVAPICH**
  - Up to 25% higher performance
  - Platform MPI advantage increases with increased cluster size

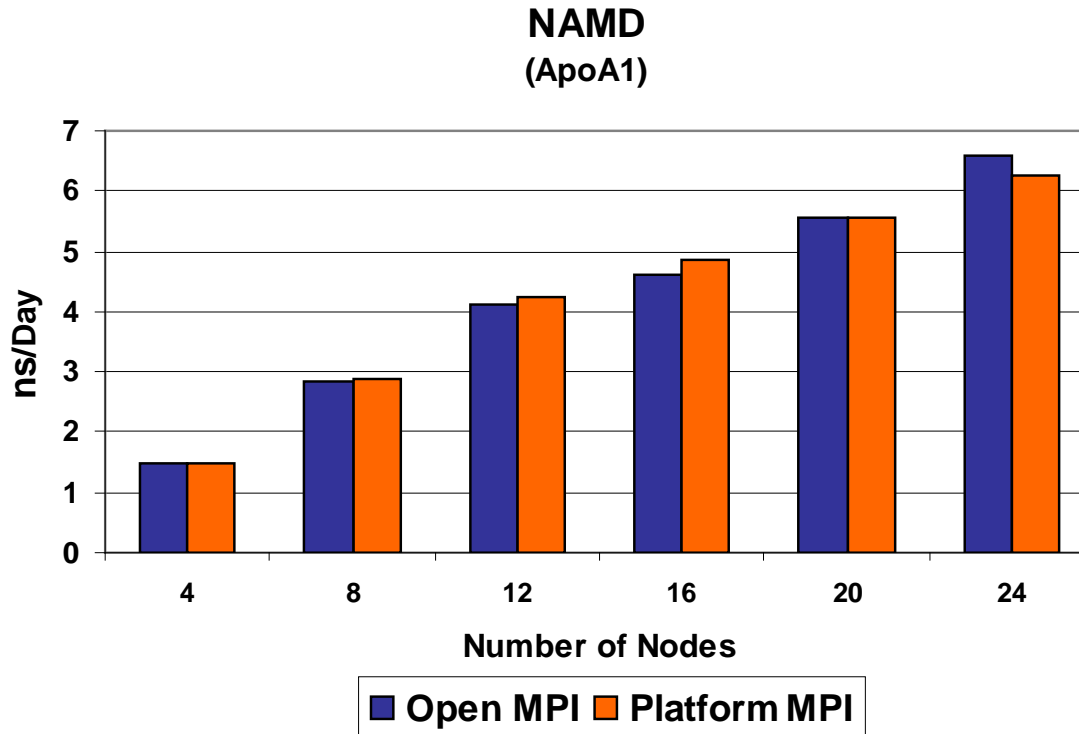
**MM5 Benchmark Results - T3A**



*Higher is better*

*Single job on each node*

- **Platform MPI and Open MPI provides same level of performance**
  - Platform MPI has better performance for cluster size lower than 20 nodes
  - Open MPI becomes better with 24 nodes
    - Higher configurations than 24 nodes were not tested



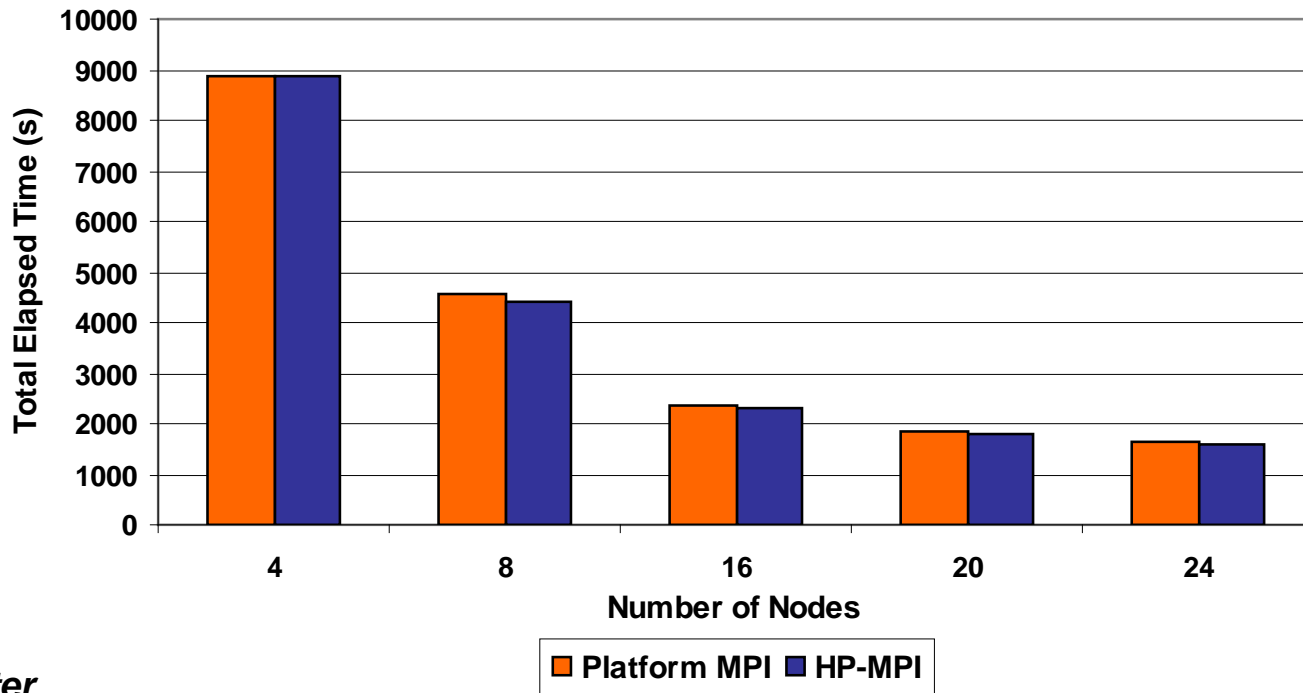
*Higher is better*

*These results are based on InfiniBand*



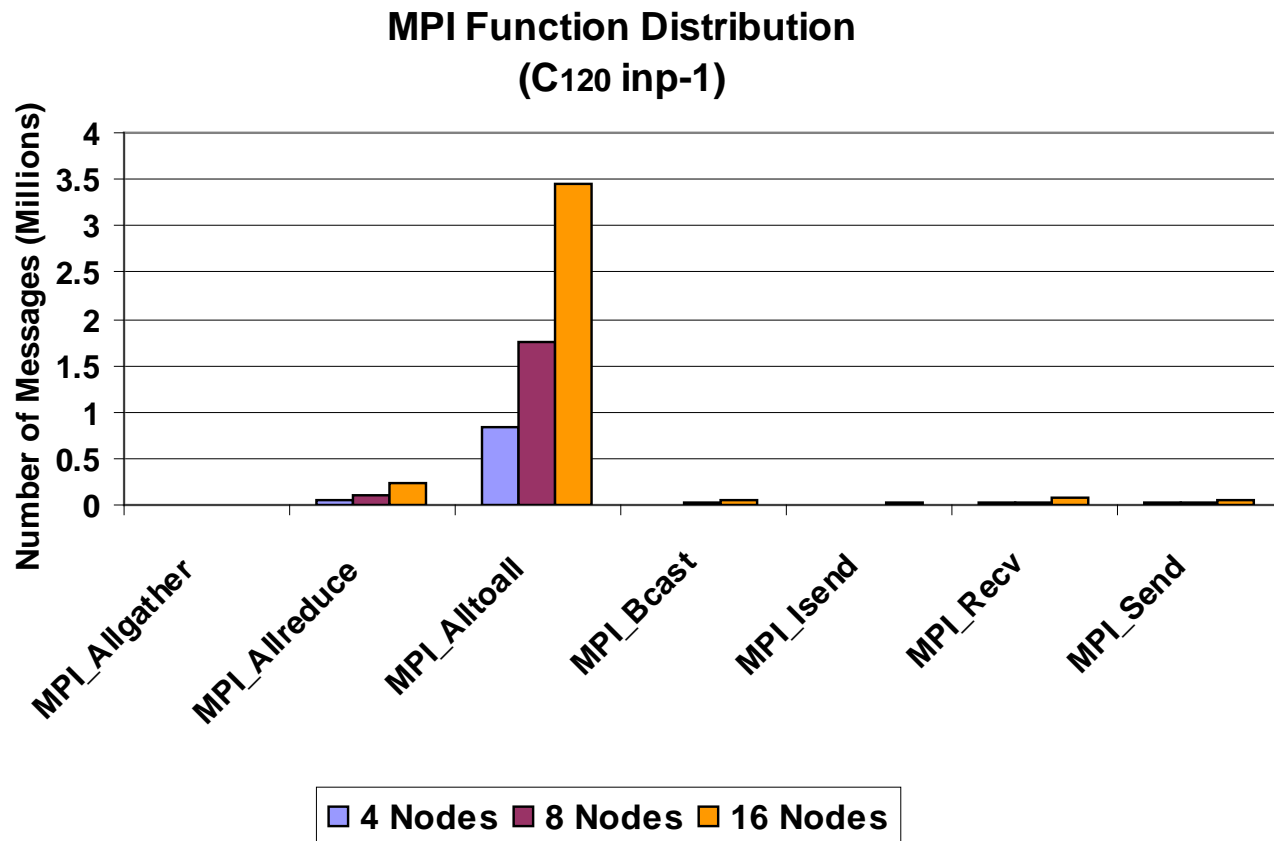
- **Test case**
  - Single job over the entire systems
  - Input Dataset (A-Class)
- **HP-MPI has slightly better performance with CPU affinity enabled**

STAR-CD Benchmark Results  
(A-Class)

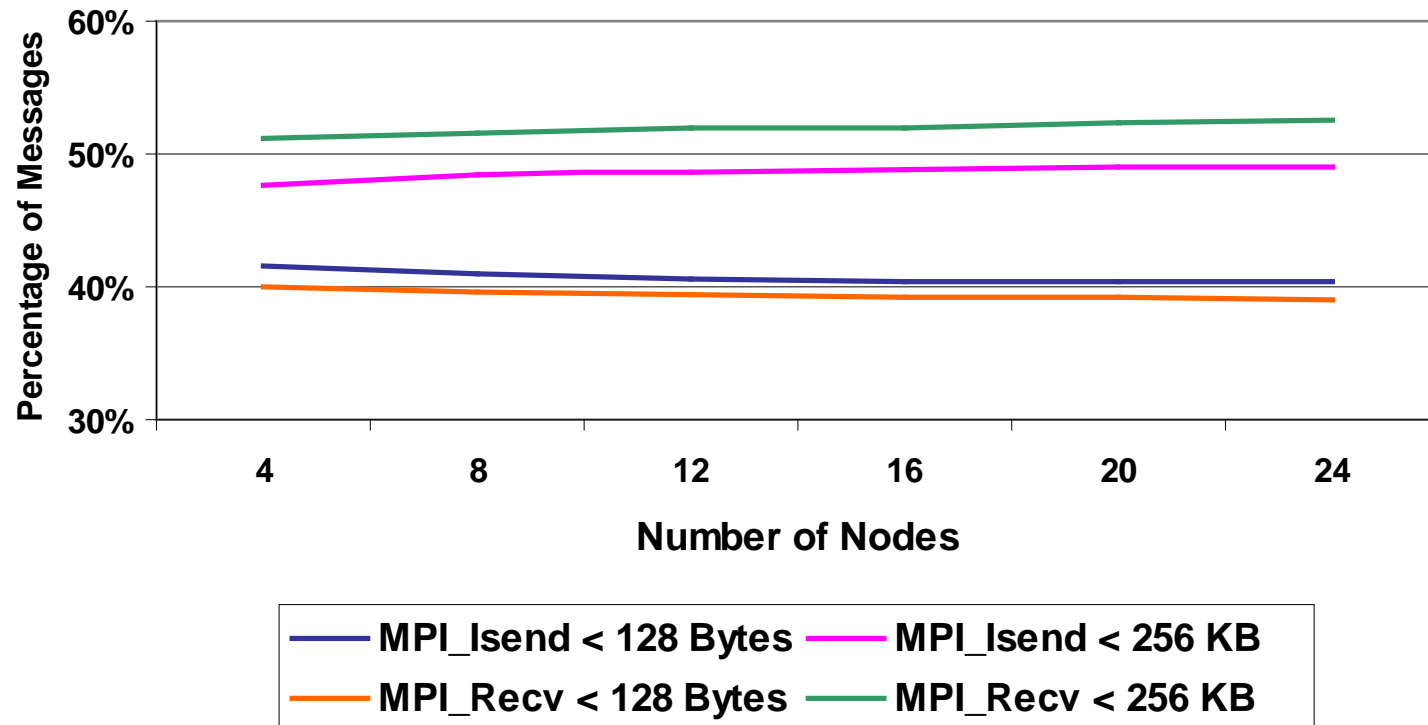


*Lower is better*

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

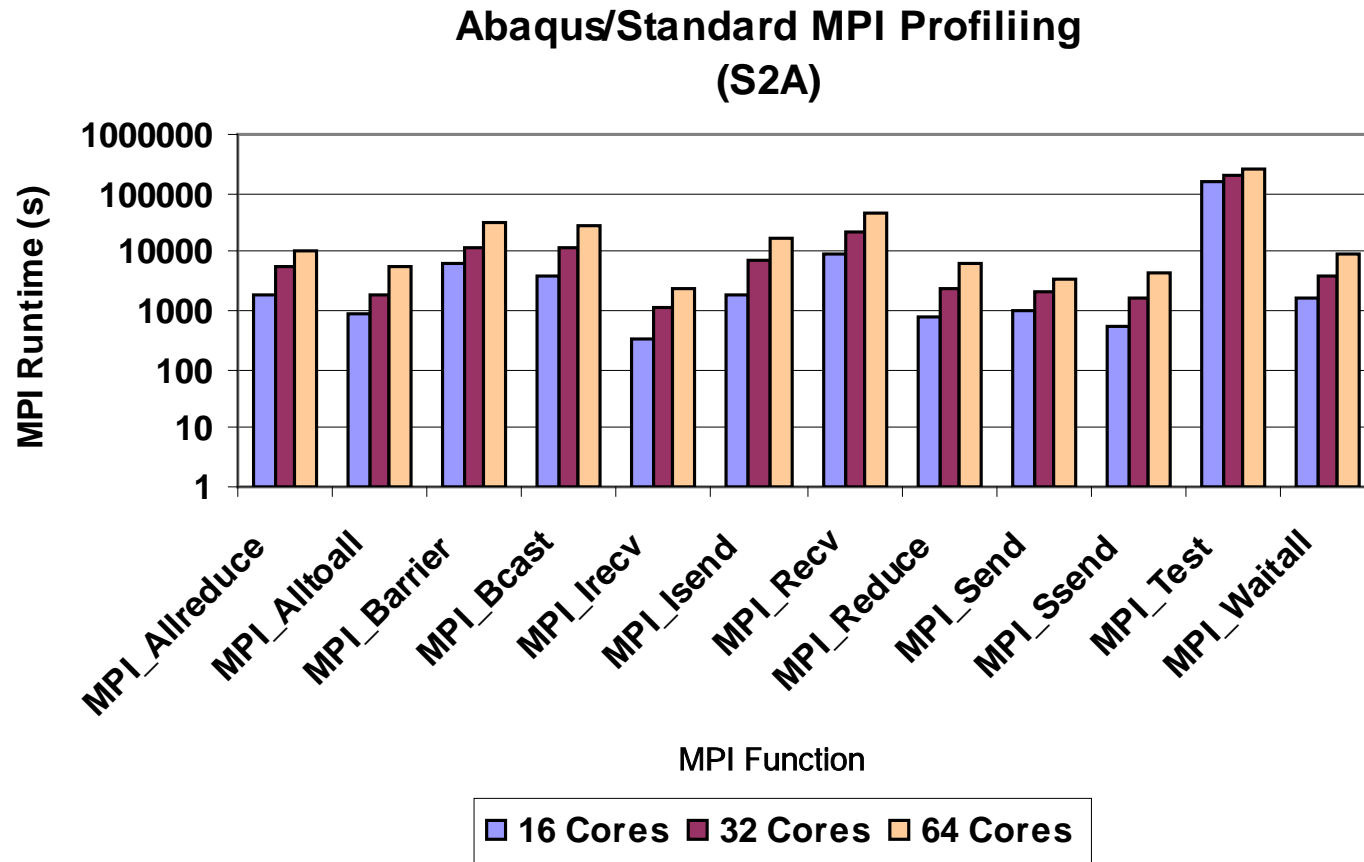


## Eclipse MPI Profiling



- Majority of MPI messages are large size
- Demonstrating the need for highest throughput

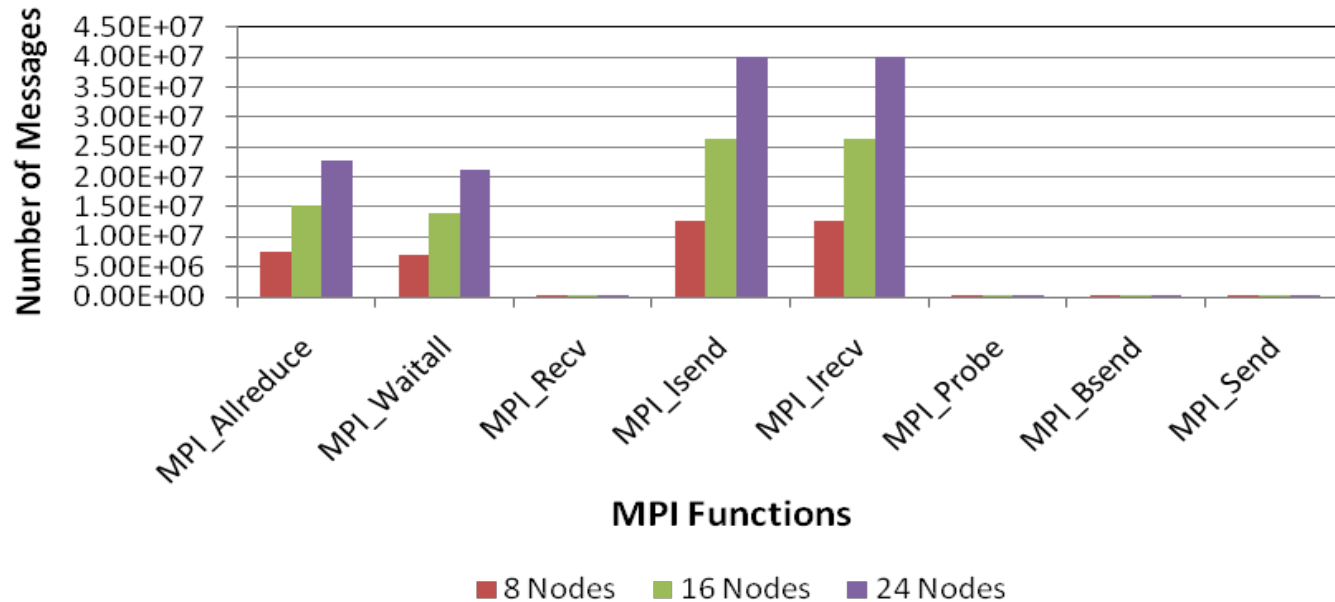
- MPI\_Test, MPI\_Recv, and MPI\_Barrier/Bcast show the highest communication overhead



- **Mostly used MPI functions**

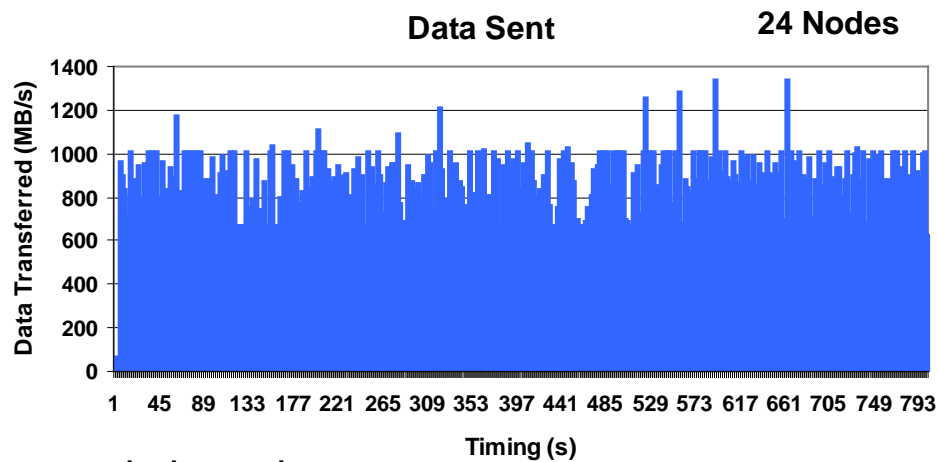
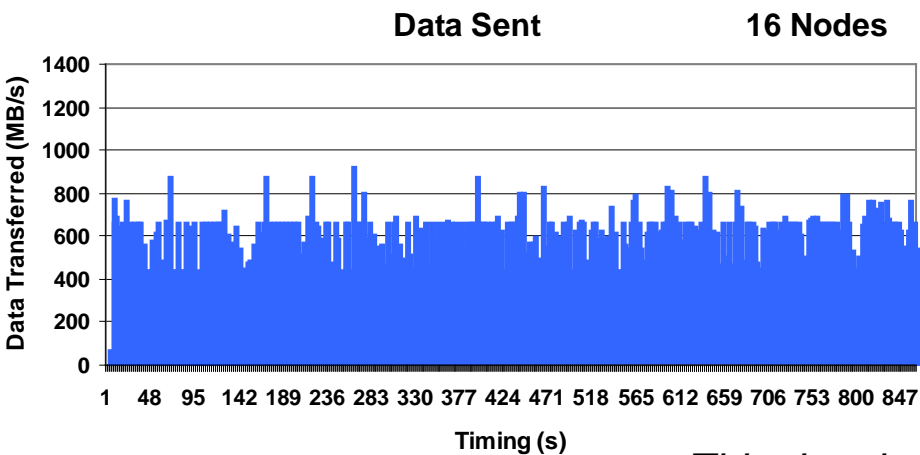
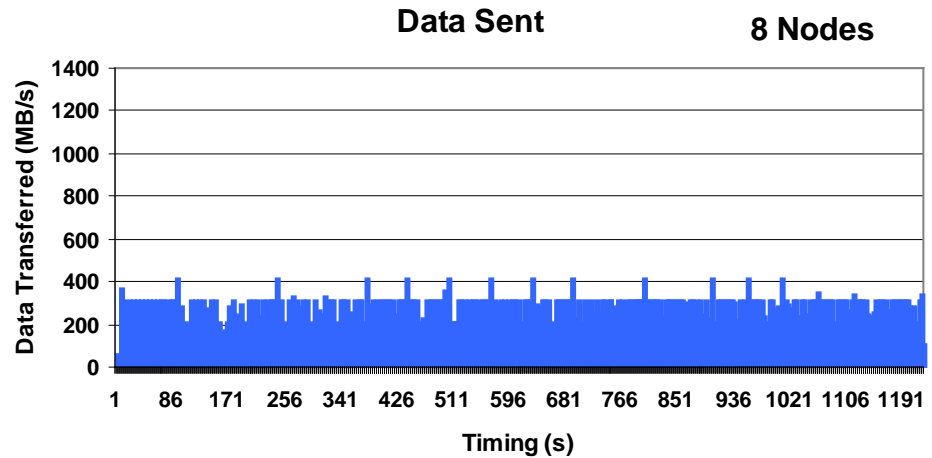
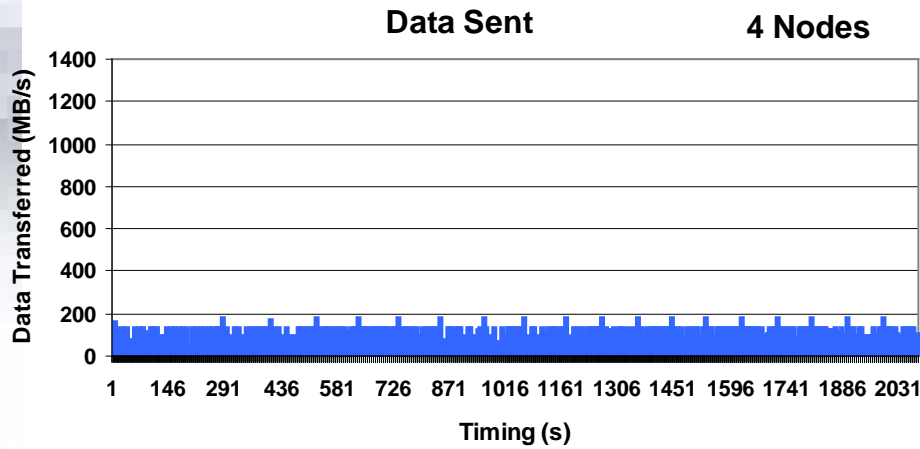
- MPI\_Allreduce, MPI\_Waitall, MPI\_Isend, and MPI\_Recv
- Number of MPI functions increases with cluster size

**MPI Profiling of OpenFOAM**  
(Number of MPI messages)



# ECLIPSE - Interconnect Usage

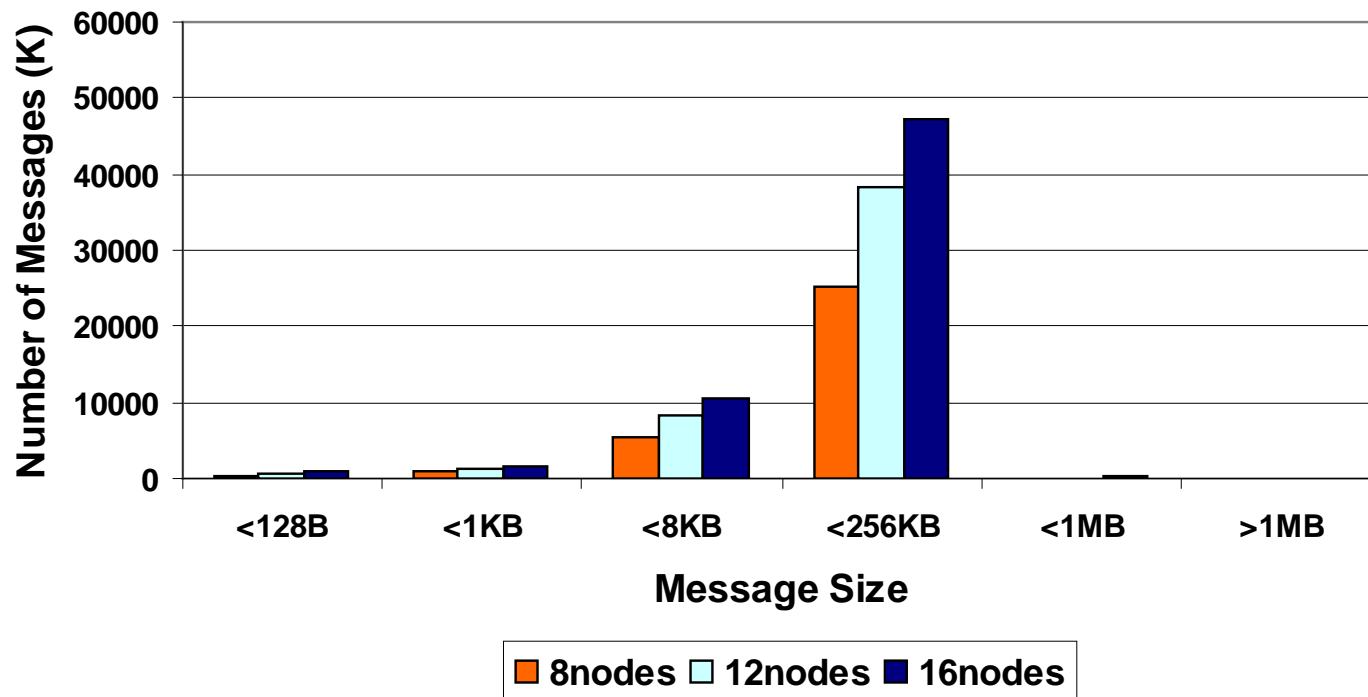
- Total server throughput increases rapidly with cluster size



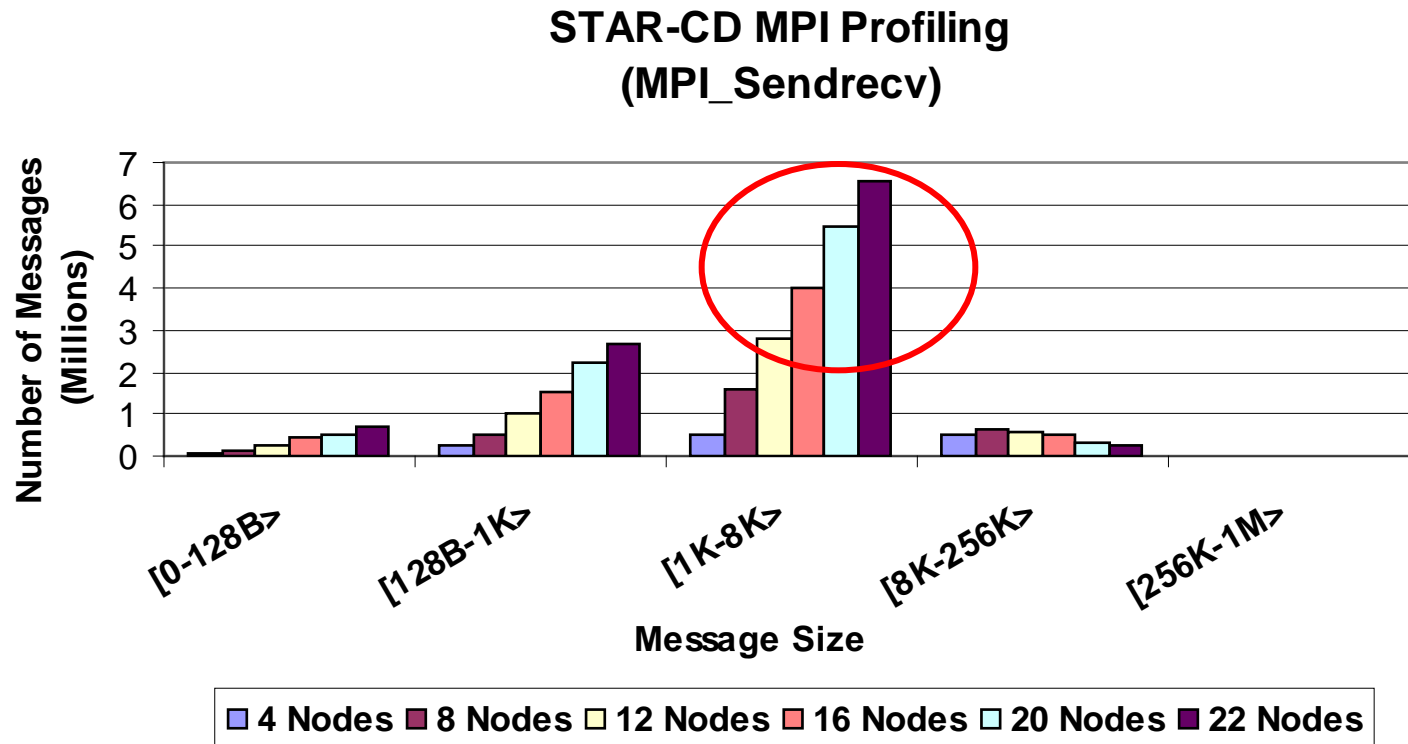
*This data is per node based*

- **Most data related MPI messages are within 8KB-256KB in size**
  - Number of messages increases with cluster size
- **Shows the need for highest throughput to ensure highest system utilization**

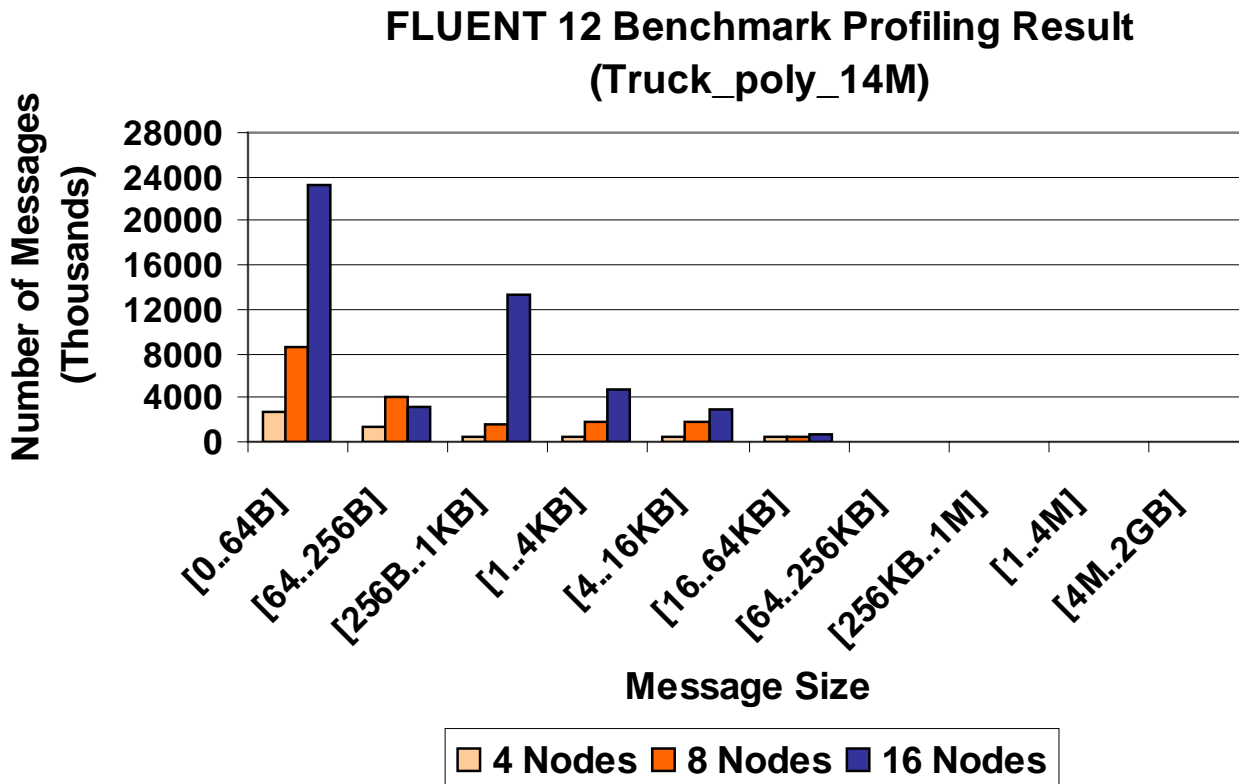
**NWChem Benchmark Profiling  
(Siosi7)**



- Most point-to-point MPI messages are within 1KB to 8KB in size
- Number of messages increases with cluster size

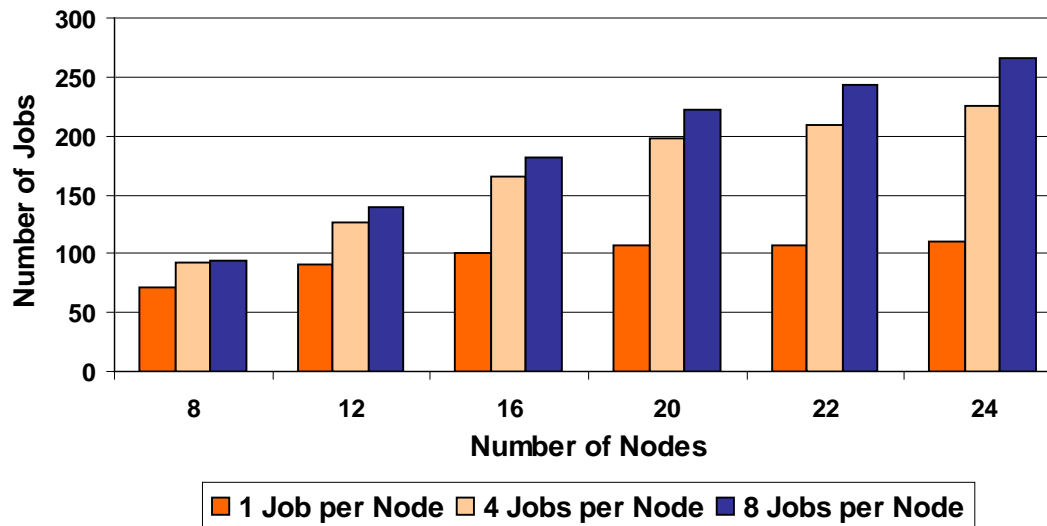


- Most data related MPI messages are within 256B-1KB in size
- Typical MPI synchronization messages are lower than 64B in size
- Number of messages increases with cluster size



- **InfiniBand increases productivity by allowing multiple jobs to run simultaneously**
  - Providing required productivity for reservoir simulations
- **Three cases are presented**
  - Single job over the entire systems
  - Four jobs, each on two cores per CPU per server
  - Eight jobs, each on one CPU core per server
- **Eight jobs per node increases productivity by up to 142%**

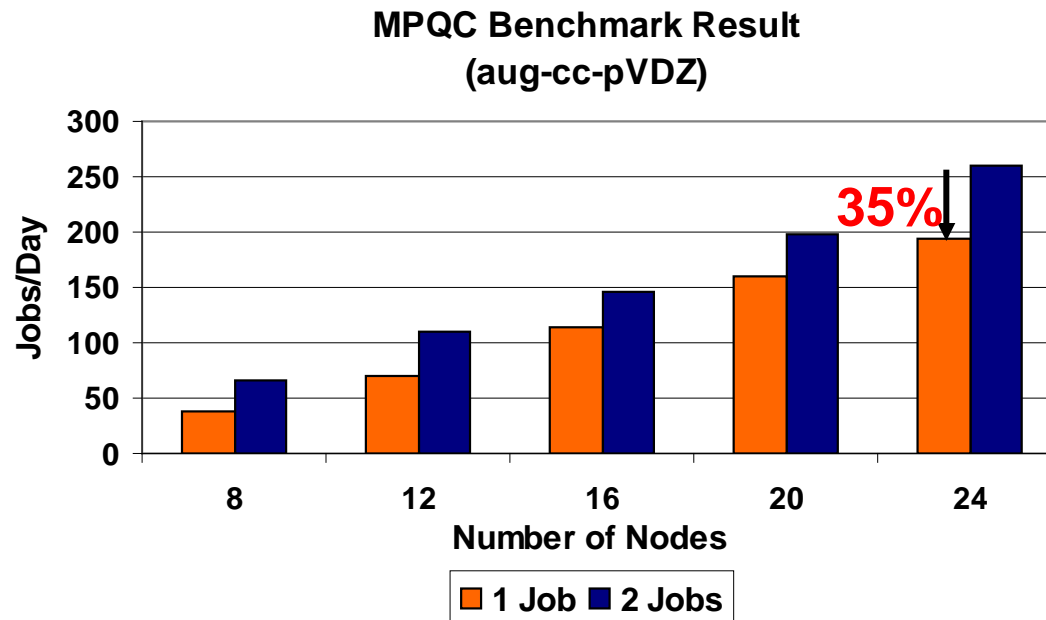
Schlumberger ECLIPSE  
(FOURMILL)



*Higher is better*

*InfiniBand*

- **InfiniBand increases productivity by allowing multiple jobs to run simultaneously**
  - Providing required productivity for MPQC computation
- **Two cases are presented**
  - Single job over the entire systems
  - Two jobs, each on four cores per server
- **Two jobs per node increases productivity by up to 35%**

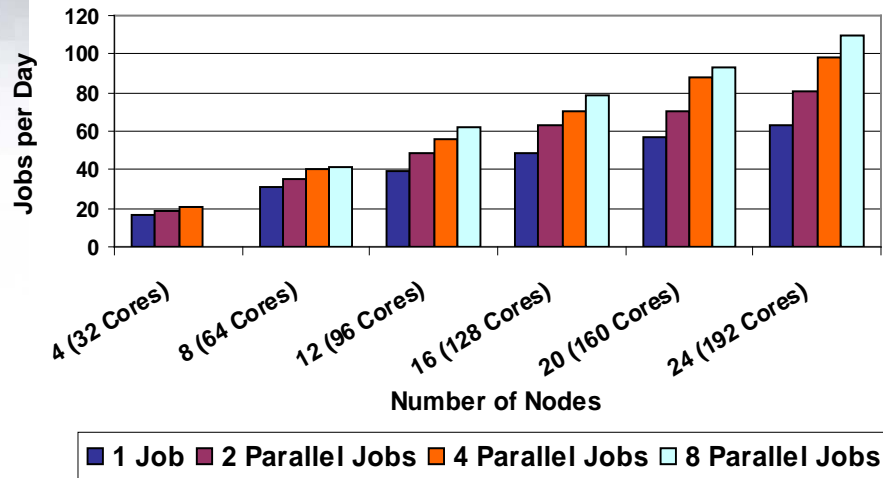


*Higher is better*

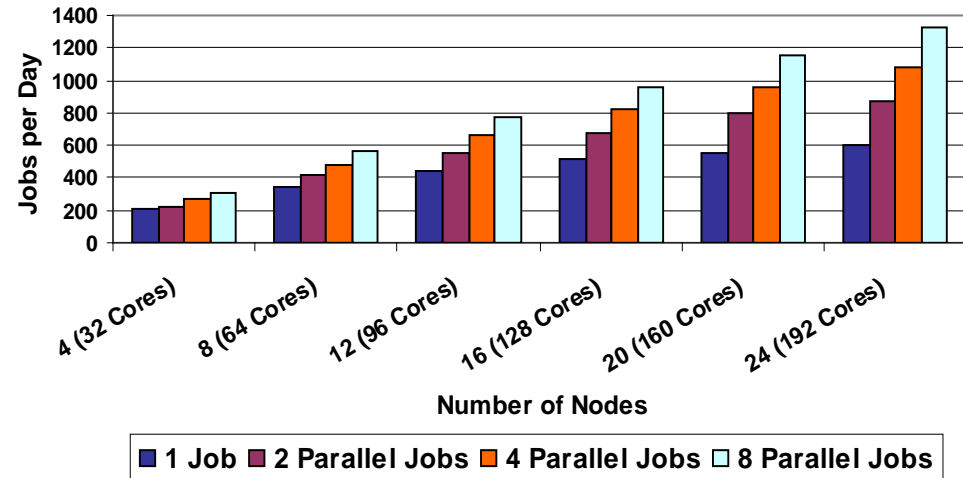
*InfiniBand*

# LS-DYNA Performance - Productivity

### LS-DYNA - 3 Vehicle Collision



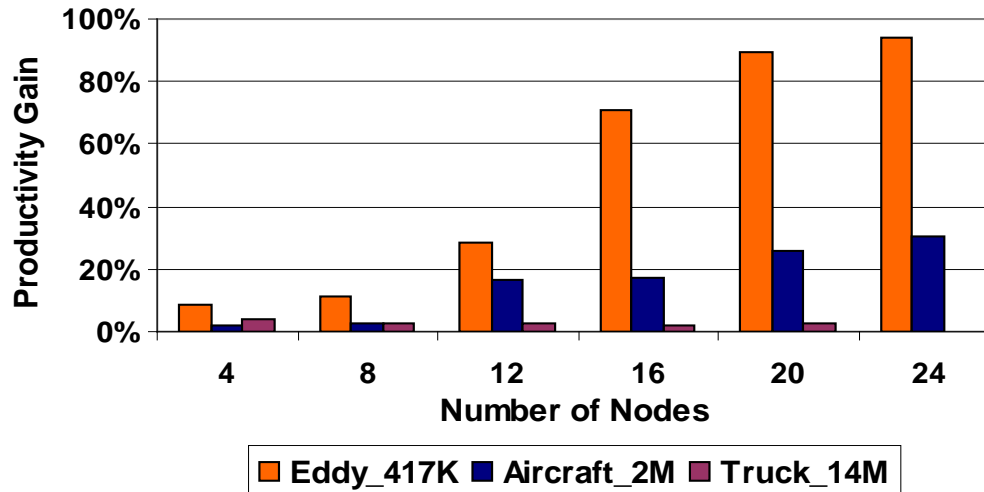
### LS-DYNA - Neon Refined Revised



Higher is better

- **Test cases**
  - Single job over the entire systems
  - 2 jobs, each runs on four cores per server
- **Running multiple jobs simultaneously improves FLUENT productivity**
  - Up to 90% more jobs per day for Eddy\_417K
  - Up to 30% more jobs per day for Aircraft\_2M
  - Up to 3% more jobs per day for Truck\_14M
- **As bigger the # of elements, higher node count is required for increased productivity**
  - The CPU is the bottleneck for larger number of servers

**FLUENT 12.0 Productivity Result**  
(2 jobs in parallel vs 1 job)

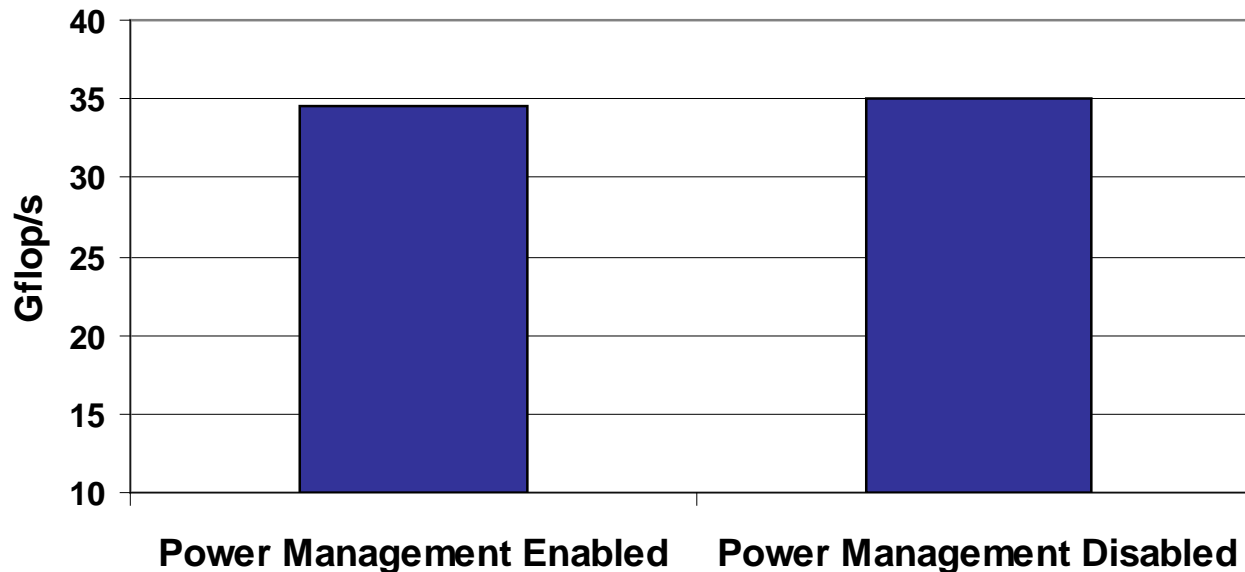


*Higher is better*

*InfiniBand DDR*

- **Test Scenario**
  - 24 servers, 4 processes per node, 2 processes per CPU (socket)
- **Similar performance with power management enabled or disabled**
  - Only 1.4% performance degradation

## MM5 Benchmark Results - T3A



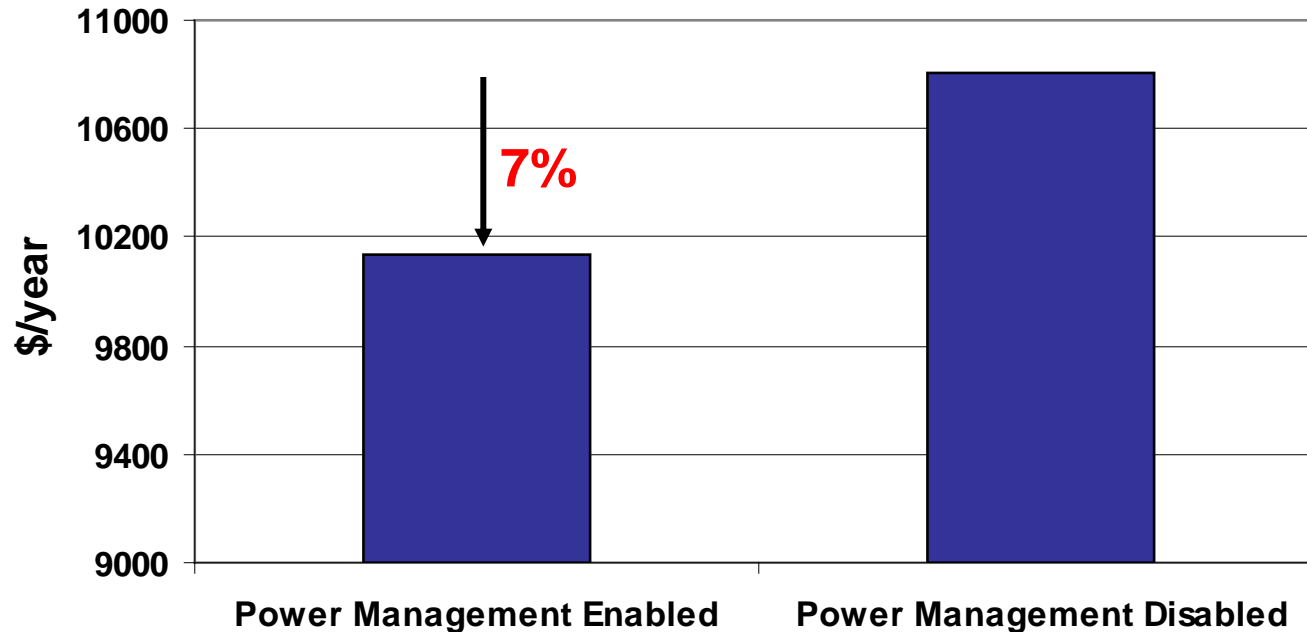
*Higher is better*

*InfiniBand DDR*

# MM5 Benchmark – Power Cost Savings

- Power management saves 673\$/year for the 24-node cluster
- As cluster size increases, bigger saving are expected

**MM5 Benchmark - T3A  
Power Cost Comparison**



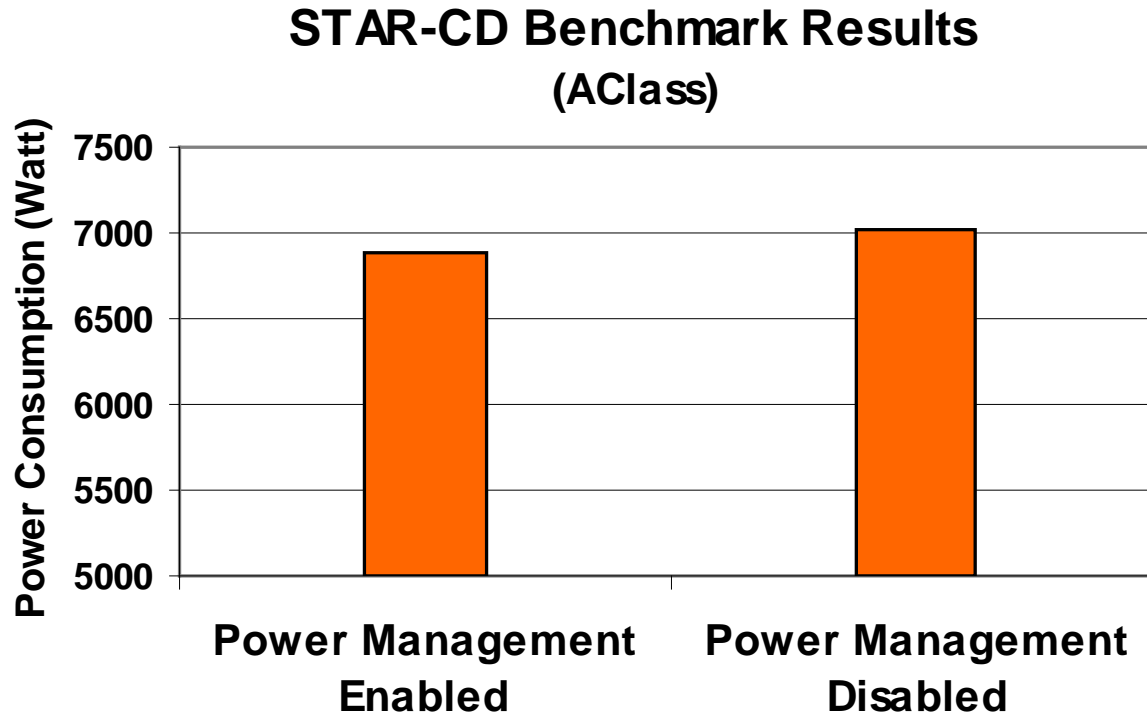
**24 Node Cluster**

$\$/year = \text{Total power consumption/year (KWh)} * \$0.20$

For more information - <http://enterprise.amd.com/Downloads/svrpwrusecompletefinal.pdf>

**InfiniBand DDR**

- Power management reduces 2% of total system power consumption 



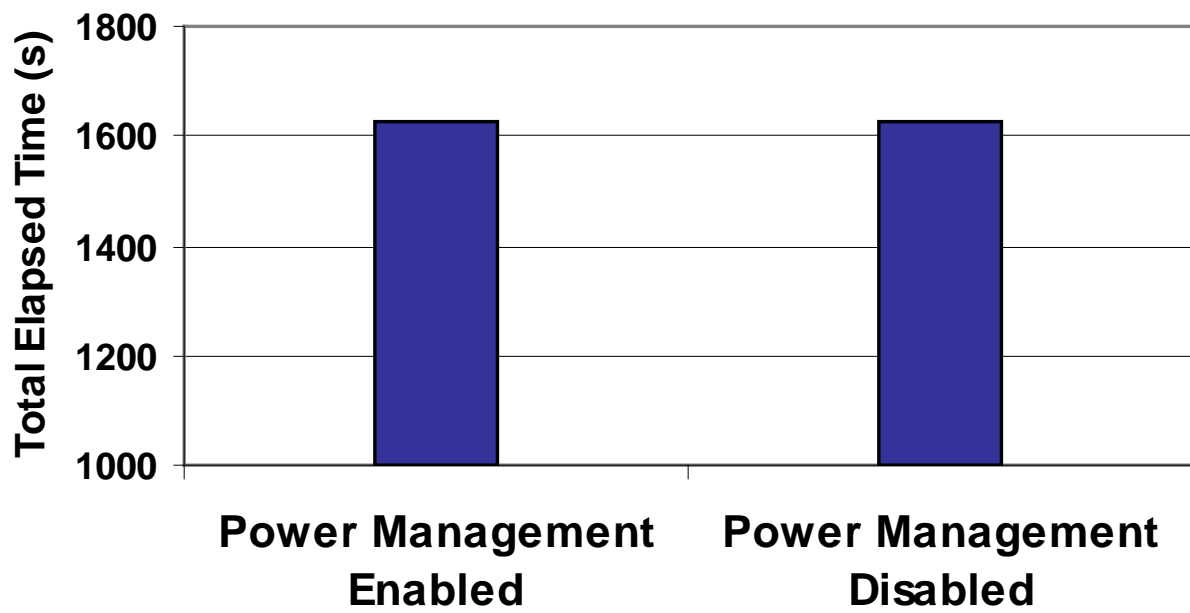
*Lower is better*

*InfiniBand DDR*



- **Test Scenario**
  - 24 servers, 4-Cores/Proc
- **Nearly identical performance with power management enabled or disabled**

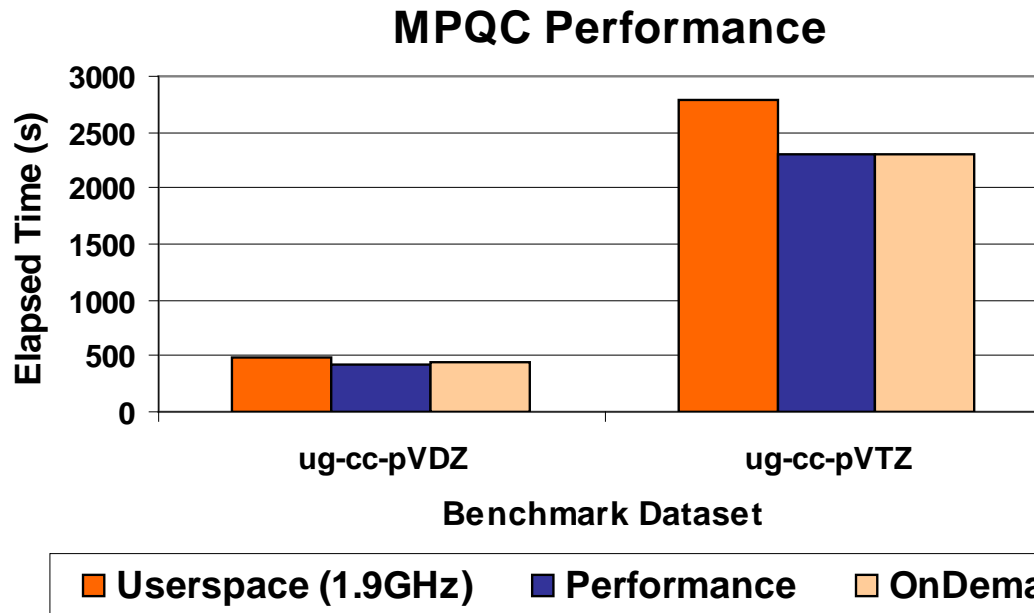
### STAR-CD Benchmark Results (A Class)



*Lower is better*

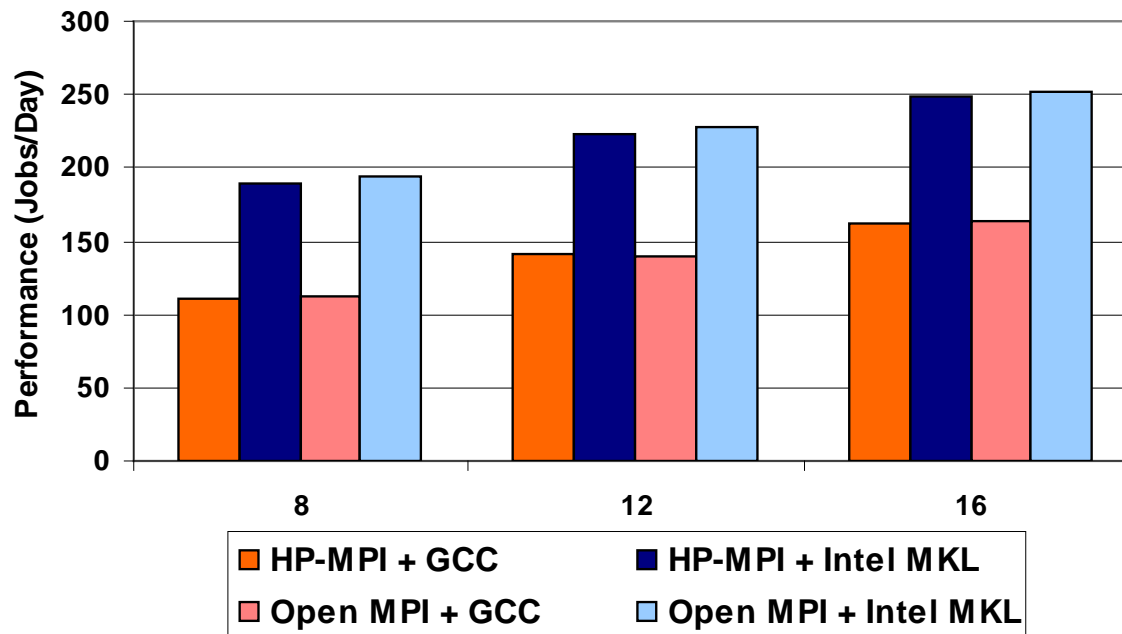
*InfiniBand DDR*

- **Enabling CPU Frequency Scaling**
  - Userspace – reducing CPU frequency to 1.9GHz
  - Performance – setting for maximum performance (CPU frequency of 2.6GHz)
  - OnDemand – Maximum performance per application activity
- **Userspace increases job run time since CPU frequency is reduced**
- **Performance and OnDemand enable similar performance**
  - Due to high resource demands from the application



- **Input Dataset - Siosi7**
- **Open MPI and HP-MPI provides similar performance and scalability**
  - Intel MKL library enables better performance versus GCC

**NWChem Benchmark Result  
(Siosi7)**



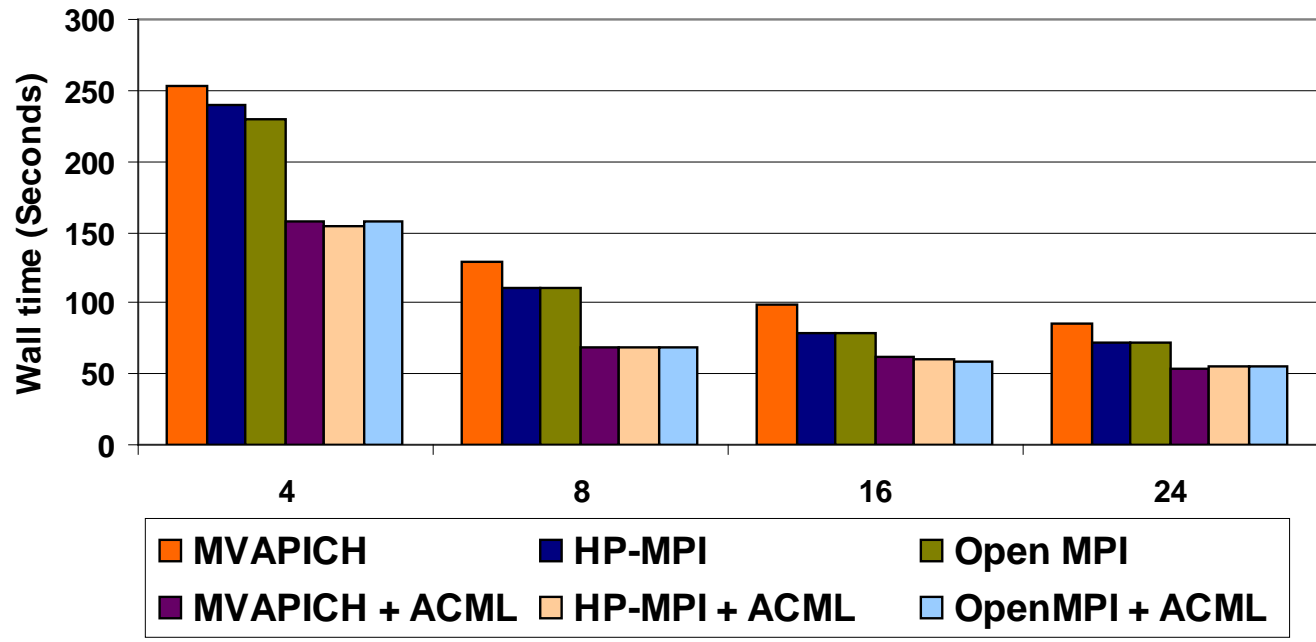
*Higher is better*

*InfiniBand QDR*

# NWChem Benchmark Results

- Input Dataset - H2O7
- AMD ACML provides higher performance and scalability versus the default BLAS library

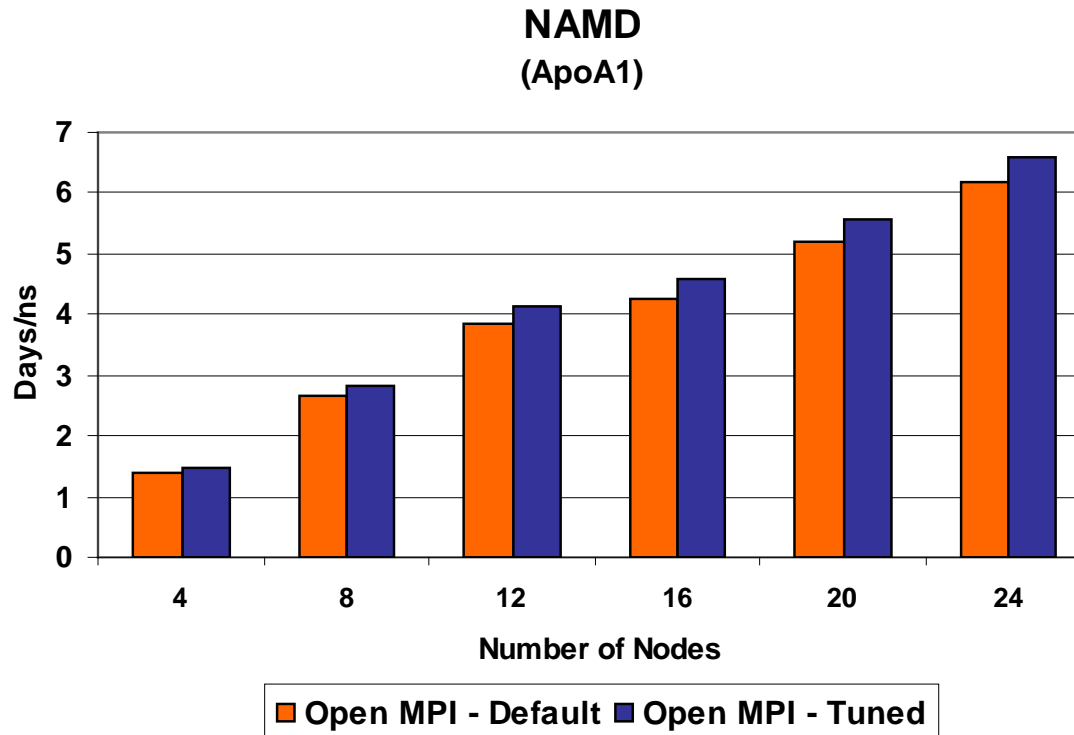
**NWChem Benchmark Result  
(H<sub>2</sub>O<sub>7</sub> MP2)**



*Lower is better*

*InfiniBand DDR*

- **MPI Performance tuning**
  - Enabling CPU affinity
    - `mca mpi_paffinity_alone 1`
  - Increasing eager limit over infiniband to 32K
    - `mca btl_openib_eager_limit 32767`
- **Performance increase of up to 10%**



*Higher is better*

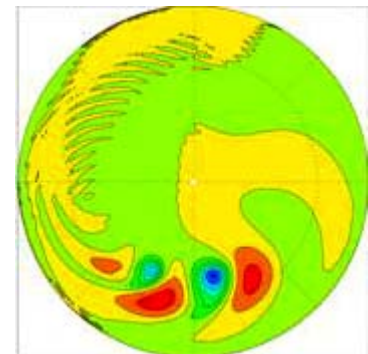
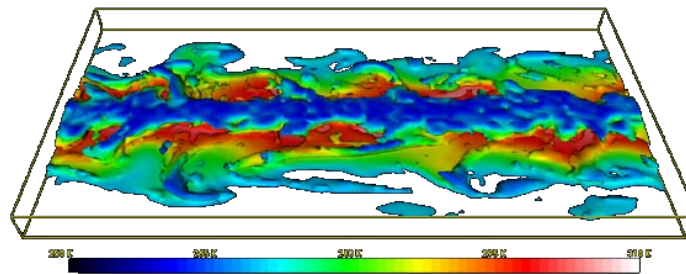
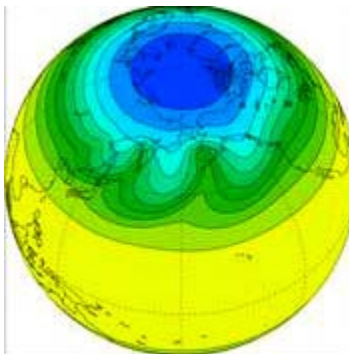
# HPC Applications Large Scale

- **The following research was performed under the HPC Advisory Council activities**
  - Participating members: AMD, Dell, Jülich, Mellanox NCAR, ParTec, Sun
  - Compute resource: HPC Advisory Council Cluster Center, Jülich Supercomputing Centre
- **For more info please refer to**
  - [www.hpcadvisorycouncil.com](http://www.hpcadvisorycouncil.com)

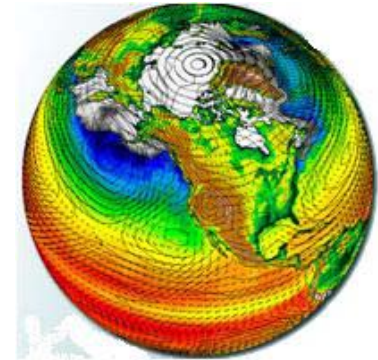
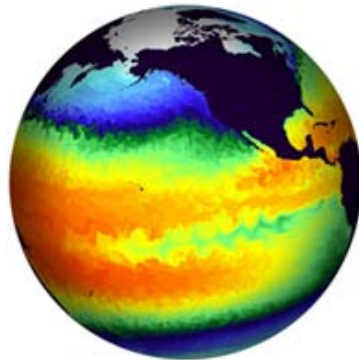


- **High-Order Methods Modeling Environment (HOMME)**

- Framework for creating a high-performance scalable global atmospheric model
- Configurable for shallow water or the dry/moist primitive equations
- Serves as a prototype for the Community Atmospheric Model (CAM) component of the Community Climate System Model (CCSM)
- HOMME supports execution on parallel computers using either MPI, OpenMP or a combination of MPI/OpenMP
- Developed by the Scientific Computing Section at the National Center for Atmospheric Research (NCAR)



- **POP (Parallel Ocean Program) is an ocean circulation model**
  - Simulations of the global ocean
  - Ocean-ice coupled simulations
  - Developed at Los Alamos National Lab
- **POPperf is a modified version of POP 2.0 (Parallel Ocean Program)**
- **POPperf improves POP scalability on large processor counts**
  - Re-writing of the conjugate gradient solver to use a 1D data structure
  - The addition of a space-filling curve partitioning technique
  - Low memory binary parallel I/O functionality
- **Developed by NCAR and freely available to the community**



- **The presented research was done to provide best practices**
  - HOMME and POPperf scalability and optimizations
  - Understanding communication patterns
- **HPC Advisory Council system**
  - Dell™ PowerEdge™ SC 1435 24-node cluster
  - Quad-Core AMD Opteron™ 2382 (“Shanghai”) CPUs
  - Mellanox® InfiniBand ConnectX HCAs and switches
- **Jülich Supercomputing Centre - JuRoPA**
  - Sun Blade servers with Intel Nehalem processors
  - Mellanox 40Gb/s InfiniBand HCAs and switches
  - ParTec ParaStation Cluster Operation Software and MPI

- **PSP\_ONDEMAND**

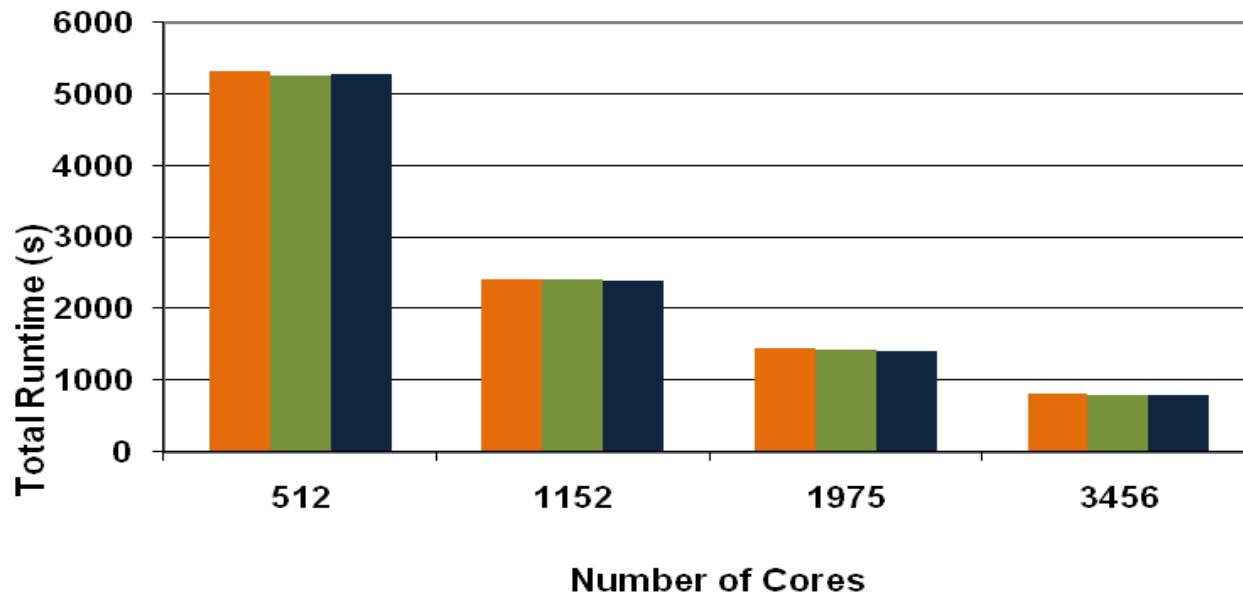
- Each MPI connection needs a certain amount of memory by default (0.5MB)
- PSP\_ONDEMAND disabled
  - MPI connections establish when application starts
  - Reduce overhead to start connection dynamically
- PSP\_ONDEMAND enabled
  - MPI connections establish per need
  - Reduce unnecessary message checking from large number of connections
  - Reduce memory footprint

- **PSP\_OPENIB\_SENDQ\_SIZE and PSP\_OPENIB\_RECVQ\_SIZE**

- Define default send/receive queue size (Default is 16)
- Changing them can reduce memory footprint

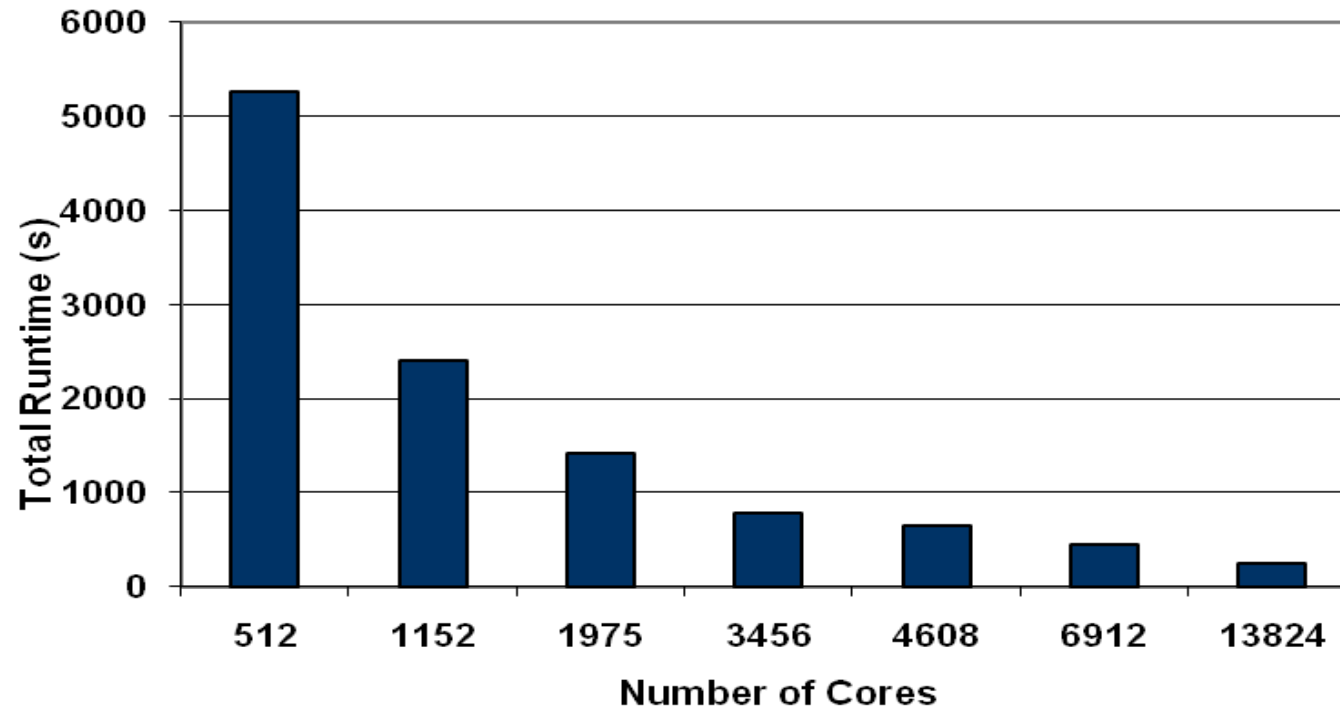
- **Each application needs customized MPI tuning**
  - Dynamic MPI connection establishment should be enabled if
    - Simultaneous connections setup is not a must
    - Some MPI function needs to check incoming messages from any source
  - Dynamic MPI connection establishment should be disabled if
    - MPI\_Alltoall is used in the application
    - Number of calls to check MPI\_ANY\_SOURCE is small
    - Enough memory in the system
- **At different scale, different parameters may be used**
  - PSP\_ONDEMAND shouldn't be enabled at very small scale cluster
- **Different MPI libraries has different characteristics**
  - Parameters change for one MPI may not fit to other MPIs

## HOMME Performance Results (Standard.nl, ndays=12)

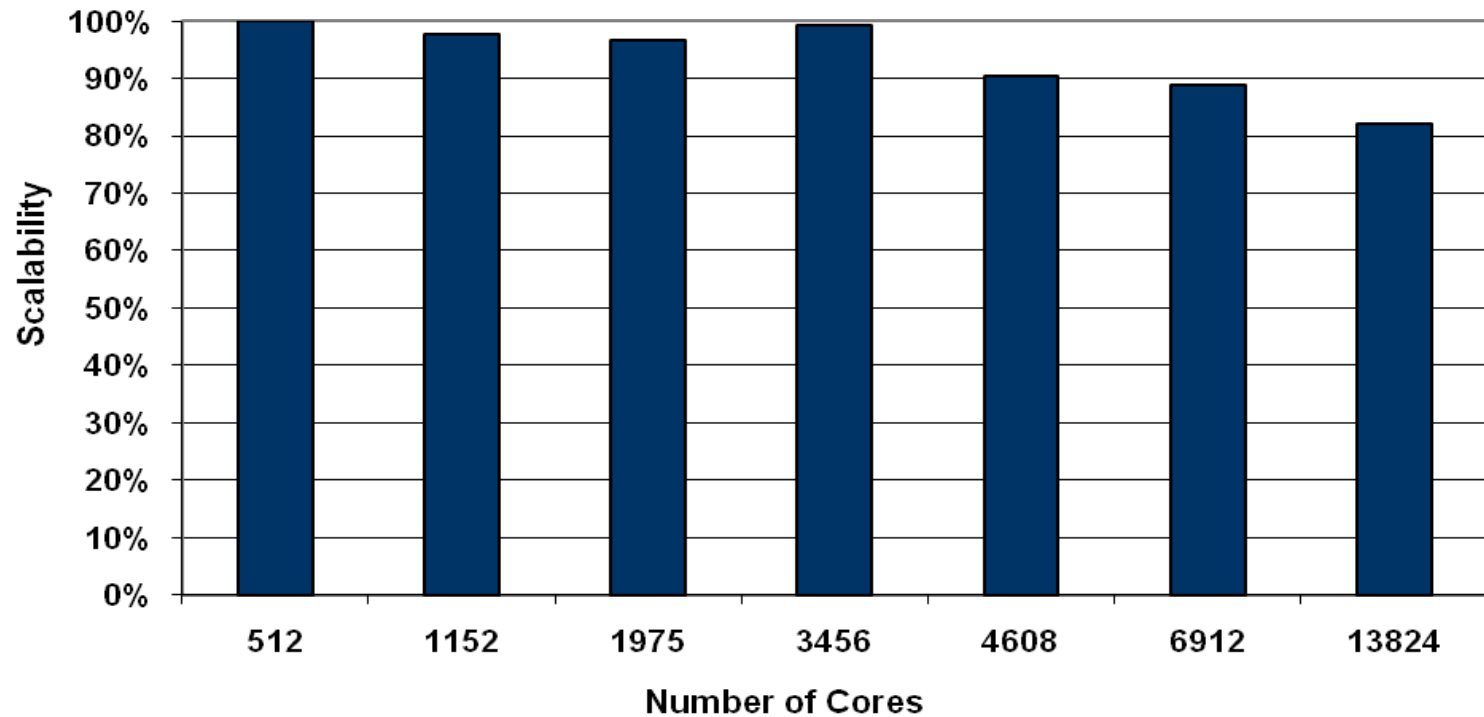


- PSP\_ONDEMAND=1 (no memory allocation)
- No PSP\_ONDEMAND (with memory allocation), PSP\_OPENIB\_SENDQ & RECVQ\_SIZE=8
- No PSP\_ONDEMAND (with memory allocation), PSP\_OPENIB\_SENDQ & RECVQ\_SIZE=16

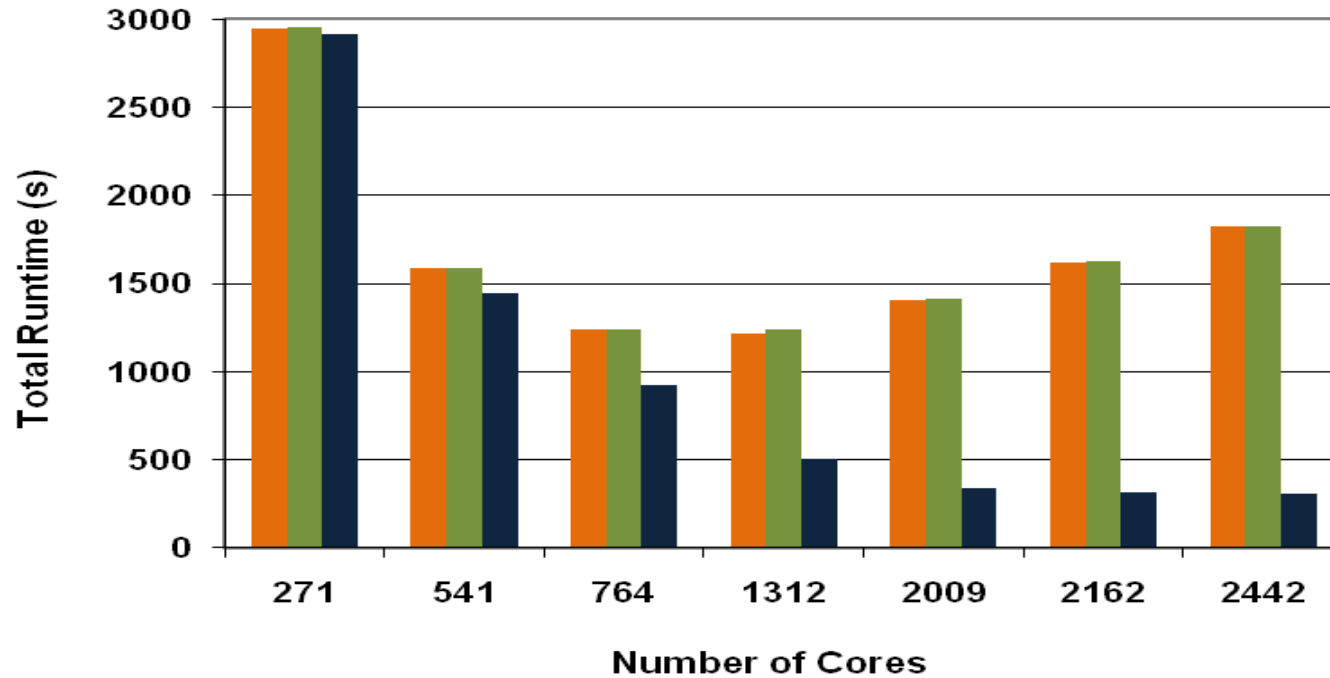
## HOMME Performance Results (Standard.nl, ndays=12)



## HOMME Scalability Results (Standard.nl, ndays=12)

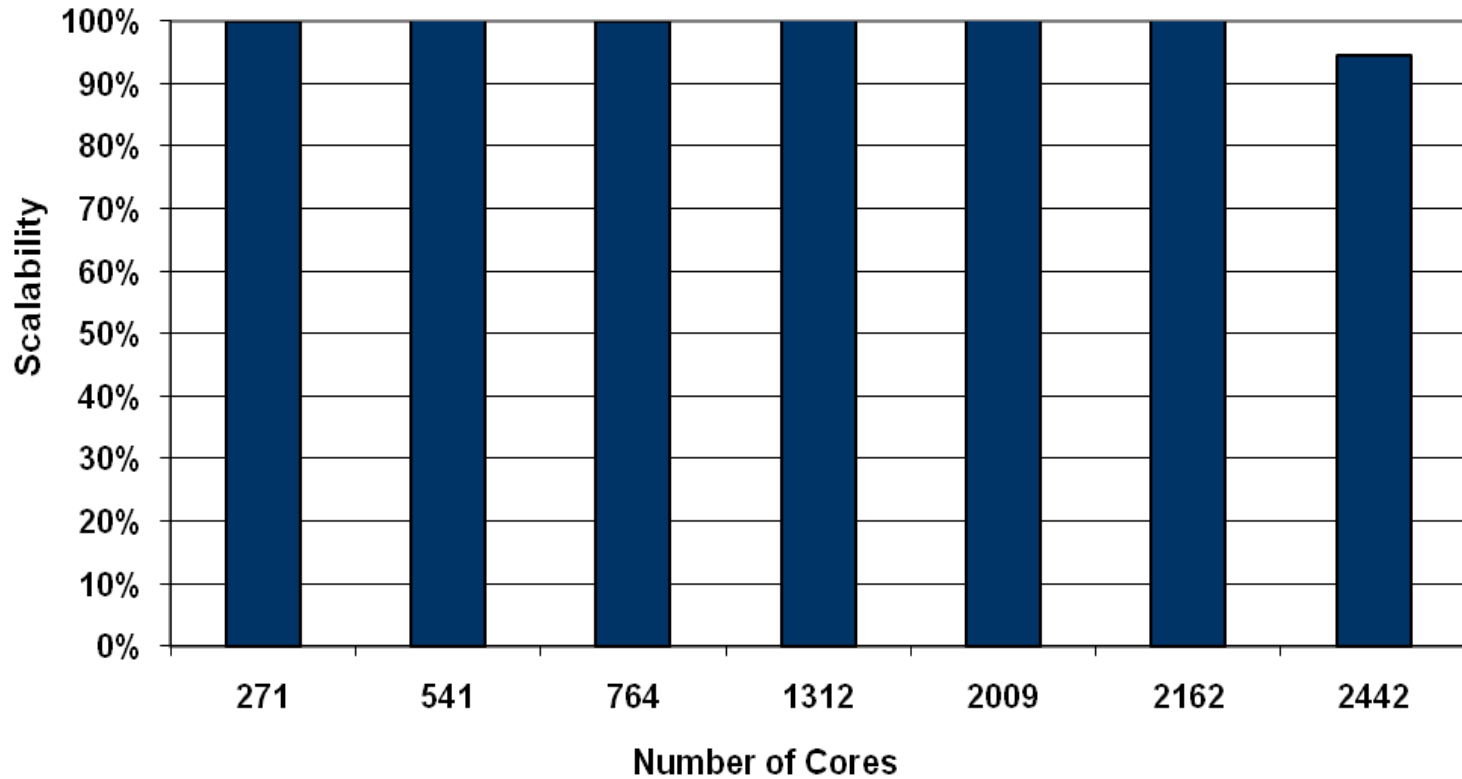


## POPperf Performance Results



- No PSP\_ONDEMAND (with memory allocation)
- No PSP\_ONDEMAND (with memory allocation), PSP\_OPENIB\_SENDQ & RECVQ\_SIZE=8
- PSP\_ONDEMAND=1 (No memory allocation)

## POPperf Scalability Results



- **HOMME and POPperf performance depends on low latency network**
- **InfiniBand enables both HOMME and POPperf to scale**
  - Tested over 13824 cores
  - Same scalability is expected at even higher core count
- **Optimized MPI settings can dramatically improve application performance**
  - Understanding application communication pattern
  - MPI parameter tuning



- **Thanks to all the participated HPC Advisory members**
  - Ben Mayer and John Dennis from NCAR who provided the code and benchmark instructions
  - Bastian Tweddell, Norbert Eicker, and Thomas Lippert who made the JuRoPA system available for benchmarking
  - Axel Koehler from Sun, Jens Hauke and Hugo R. Falter from ParTec who helped review the report



# Thank You

HPC Advisory Council

[www.hpcadvisorycouncil.com](http://www.hpcadvisorycouncil.com)

[info@hpcadvisorycouncil.com](mailto:info@hpcadvisorycouncil.com)

