

# HPC 应用程序 可扩展性的优化

刘通

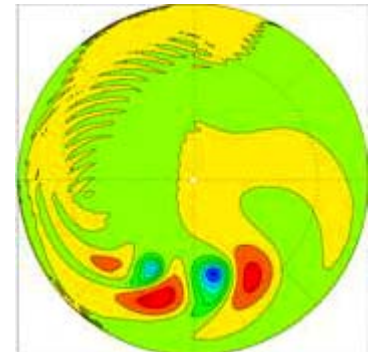
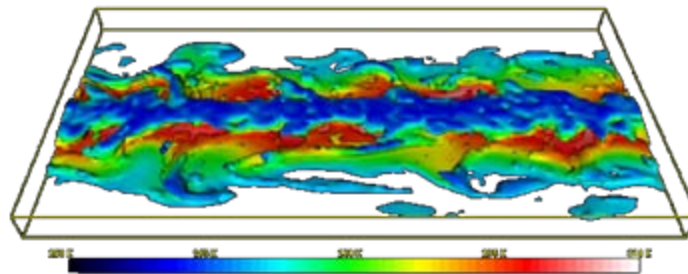
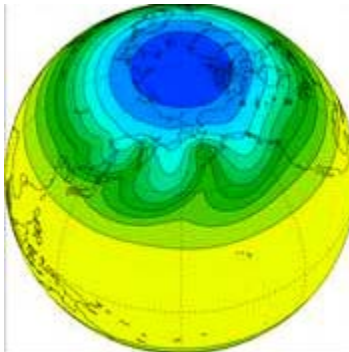
2010年10月

国际高性能计算咨询委员会中国区总监

- **硬件– Scalability depends on balanced System**
  - Intra-node
    - CPU/Memory/Disk
  - Inter-node
    - Networking
- **软件**
  - Libraries
  - Optimized network driver stack
  - File system
  - MPI tuning
- **程序本身**

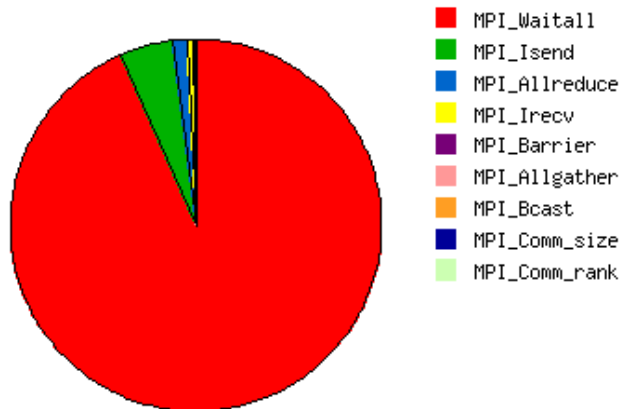
- **High-Order Methods Modeling Environment (HOMME)**

- 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)

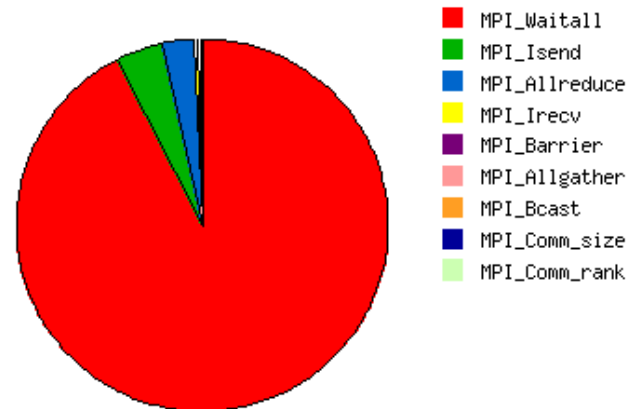


# HOMME MPI 性能分析- Timing

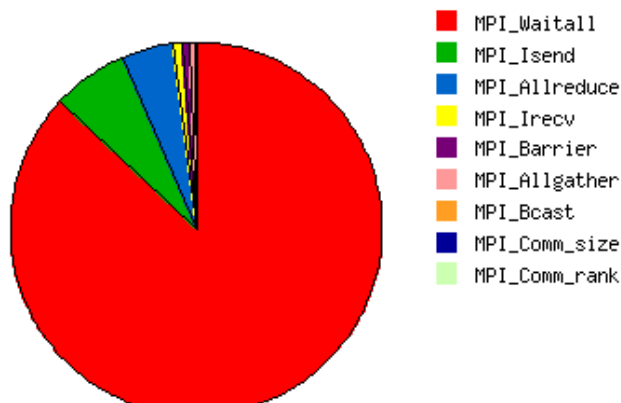
- **MPI\_Waitall** consumes large portion of MPI time
- **Time for MPI\_Allreduce** keeps increasing as cluster size scales



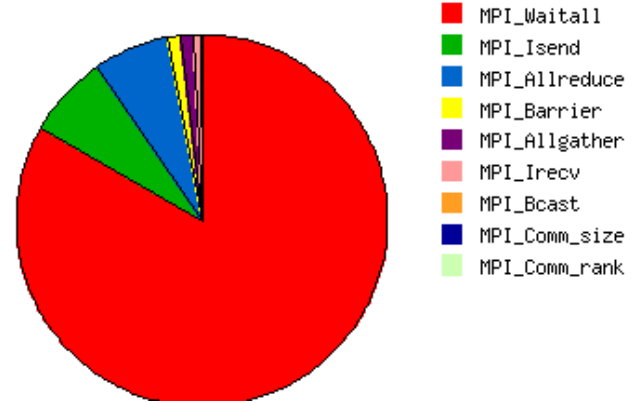
**32 processes**



**64 processes**

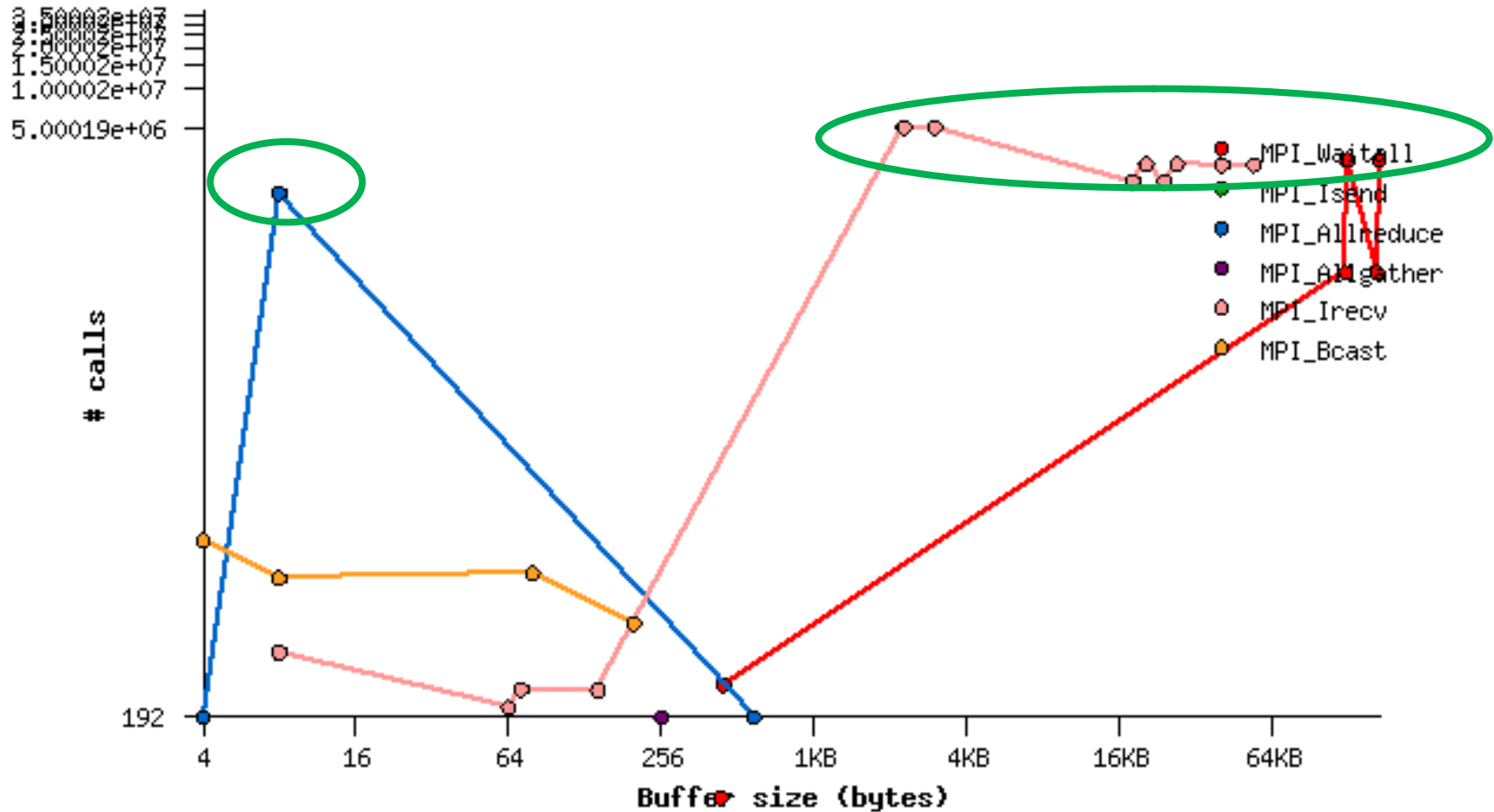


**128 processes**



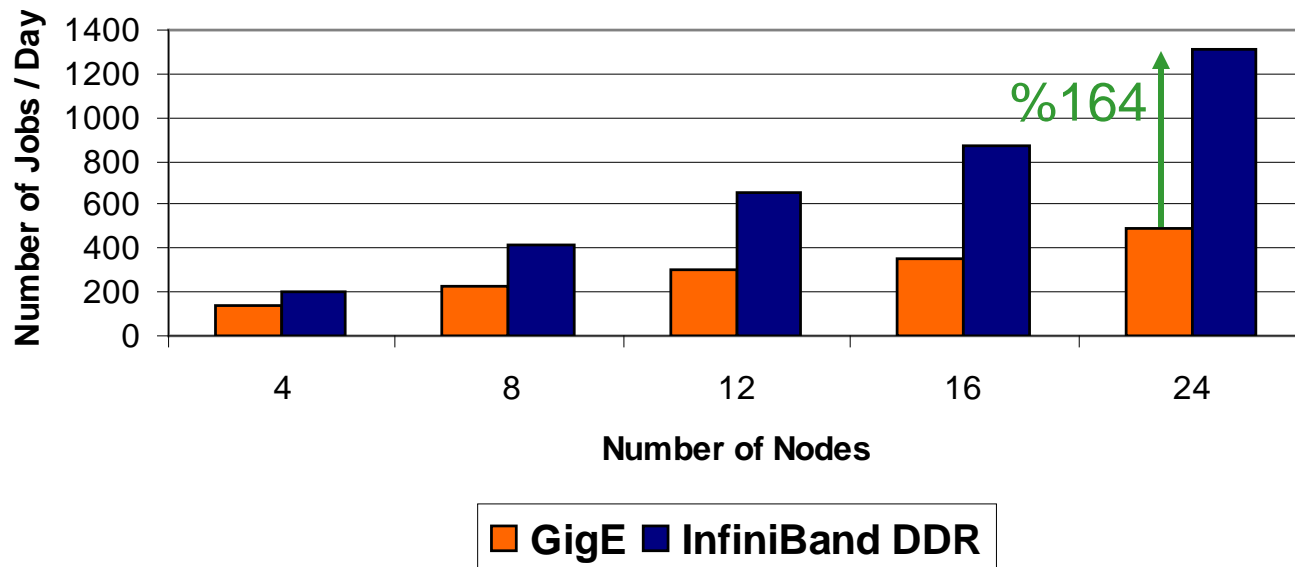
**192 processes**

# HOMME MPI 性能分析- 消息大小



- **InfiniBand provides higher utilization, performance and scalability**
  - Up to 164% faster than GigE with 24 nodes configuration
  - Performance advantage of InfiniBand versus GigE increases as cluster size scales

### HOMME Performance Results



*Higher is better*

8-cores per node

*Open MPI*

- **PSP\_ONDEMAND**

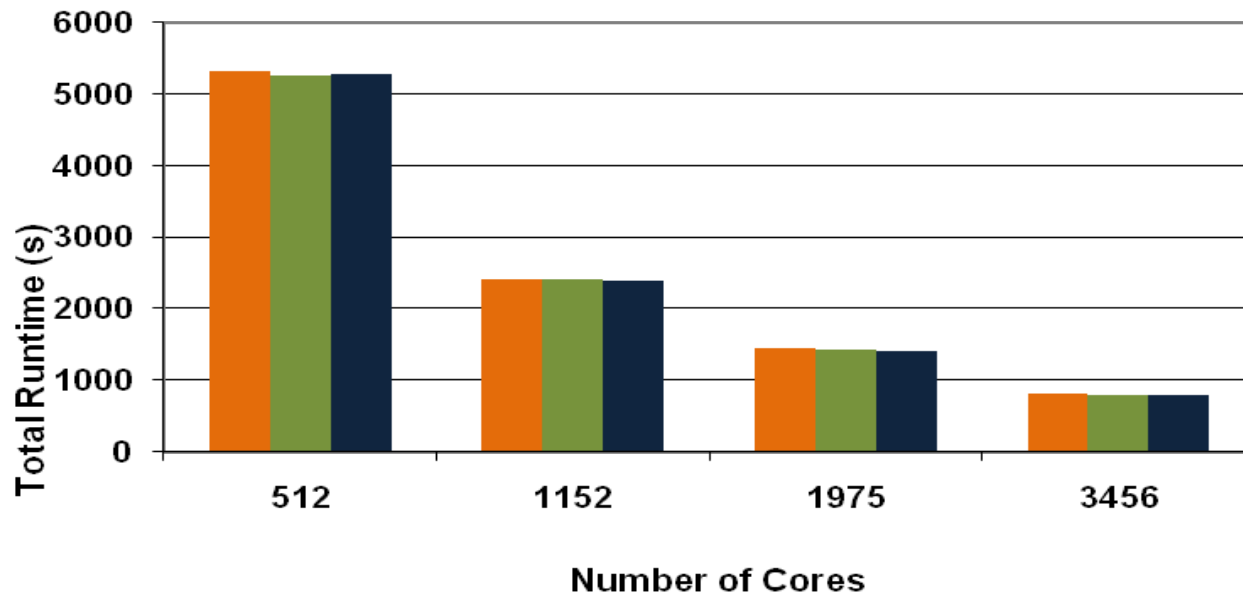
- Each MPI connection needs a certain amount of memory by default (0.5MB)
- PSP\_ONDEMAND disabled
  - MPI connections establish when application starts
  - Reduce overhead to start connection dynamically
- PSP\_ONDEMAND enabled
  - MPI connections establish per need
  - Reduce unnecessary message checking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

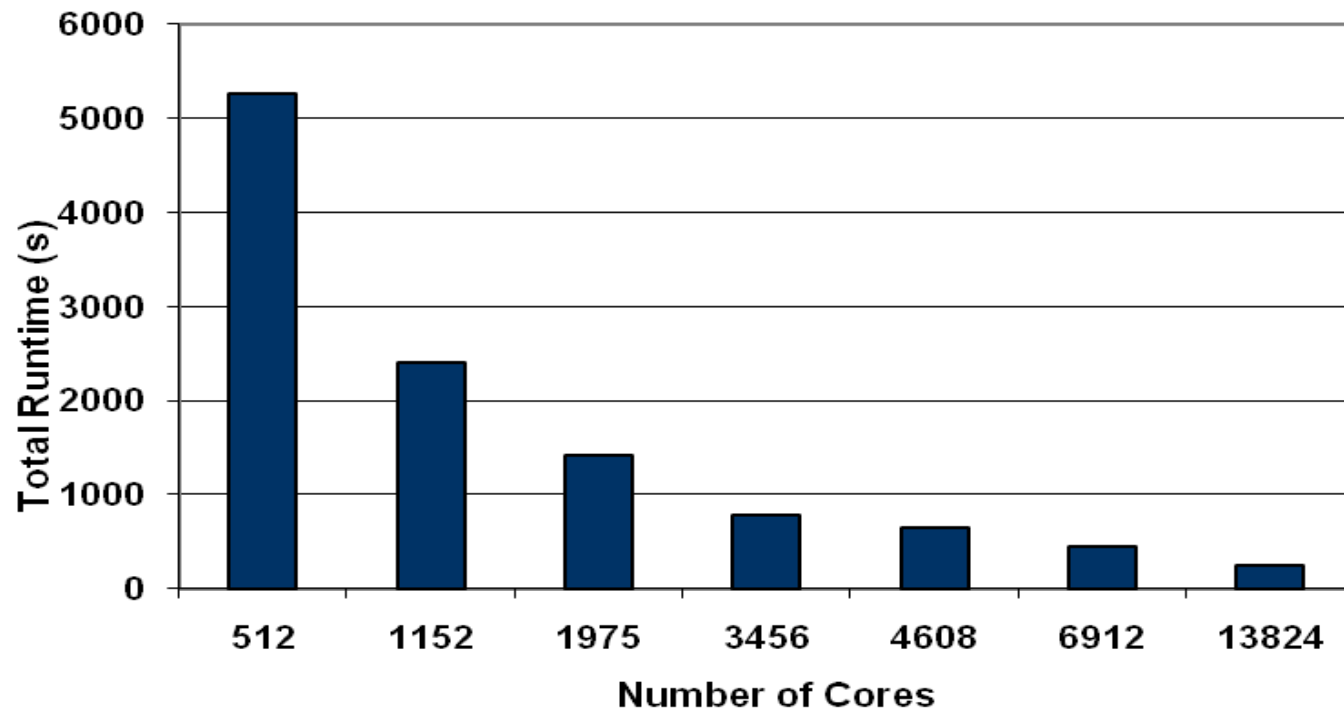
- **每个程序需要不同的MPI参数设置**
  - Dynamic MPI connection establishment should be enabled if
    - Simultaneous connections setup is not a must
    - Some MPI function needs to check incoming messages from any source
  - Dynamic MPI connection establishment should be disabled if
    - MPI\_Alltoall is used in the application
    - Number of calls to check MPI\_ANY\_SOURCE is small
    - Enough memory in the system
- **根据系统规模大小不同, 要采用不同的参数**
  - 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)

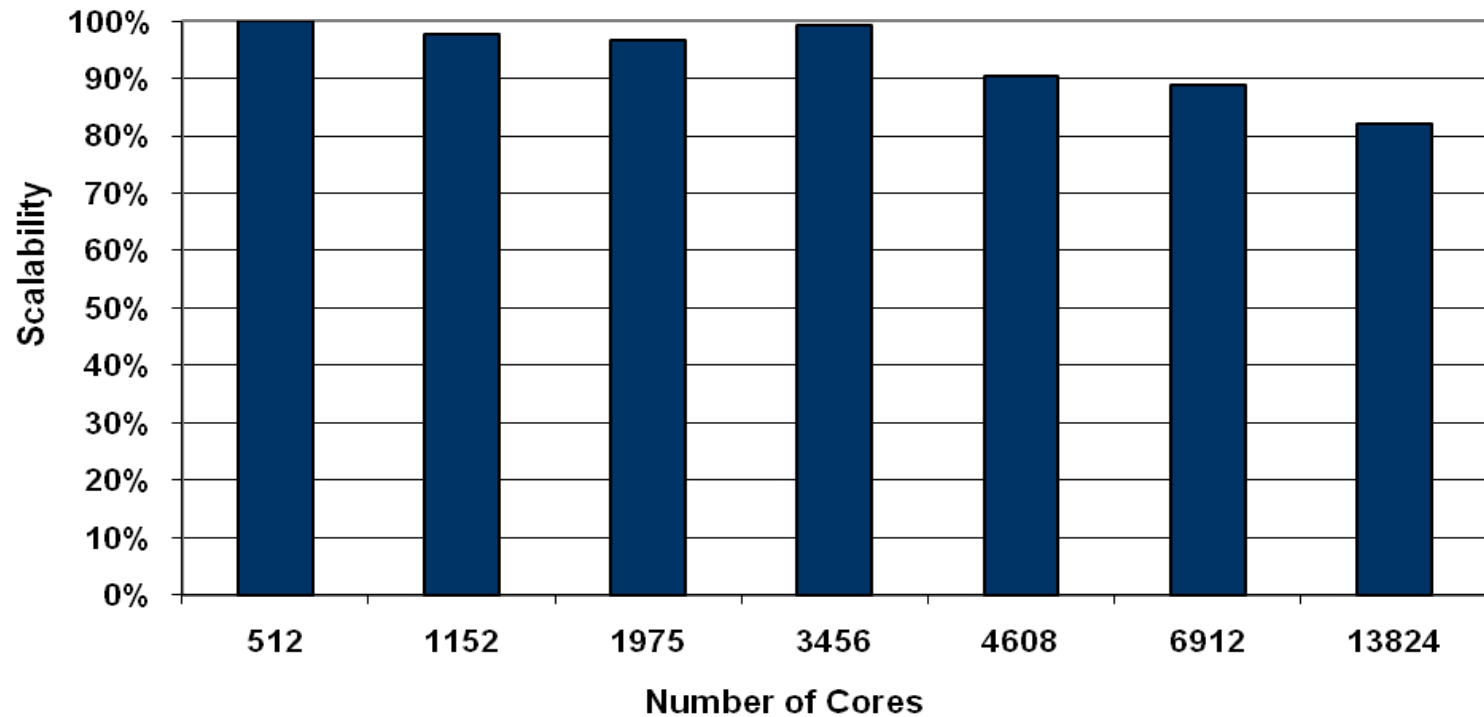


- PSP\_ONDEMAND=1 (no memory allocation)
- No PSP\_ONDEMAND (with memory allocation), PSP\_OPENIB\_SENDQ & RECVQ\_SIZE=8
- No PSP\_ONDEMAND (with memory allocation), PSP\_OPENIB\_SENDQ & RECVQ\_SIZE=16

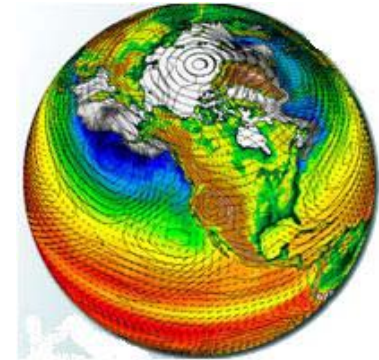
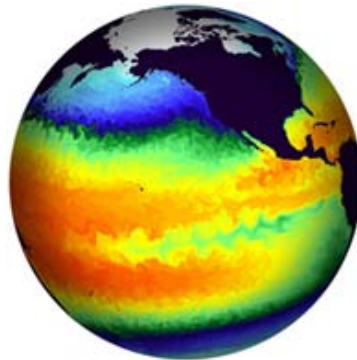
## HOMME Performance Results (Standard.nl, ndays=12)



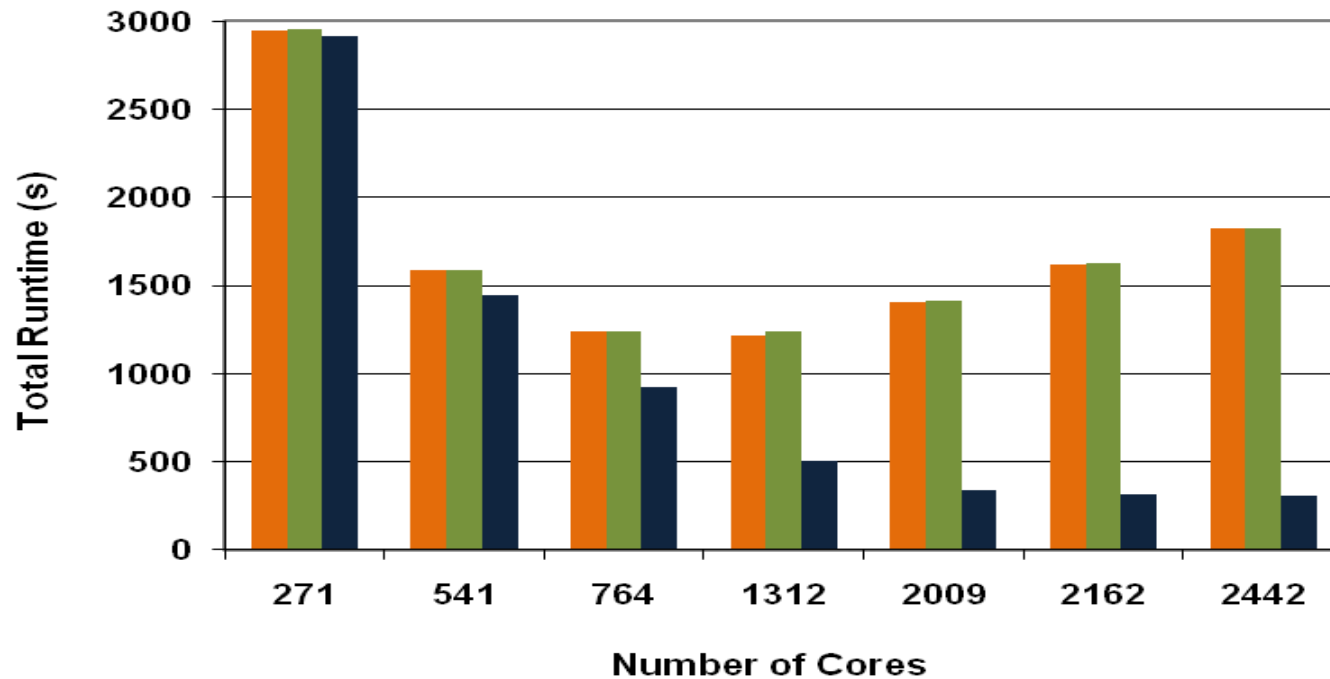
## HOMME Scalability Results (Standard.nl, ndays=12)



- **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**

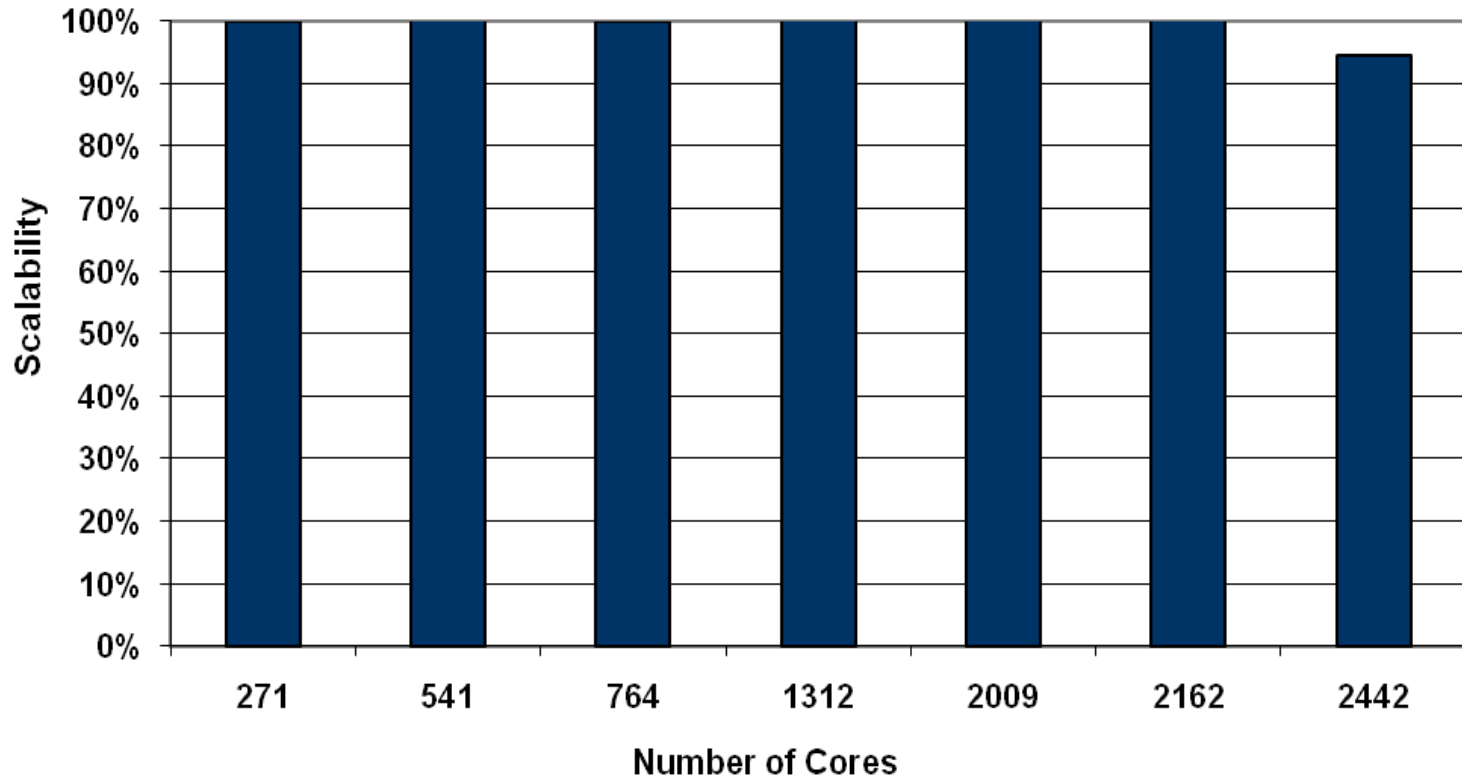


## POPperf Performance Results

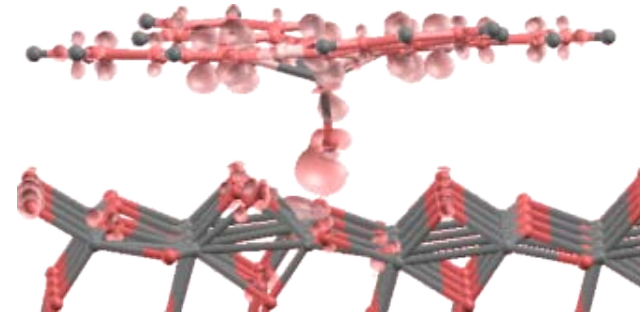


- No PSP\_ONDEMAND (with memory allocation)
- No PSP\_ONDEMAND (with memory allocation), PSP\_OPENIB\_SENDQ & RECVQ\_SIZE=8
- PSP\_ONDEMAND=1 (No memory allocation)

## POPperf Scalability Results

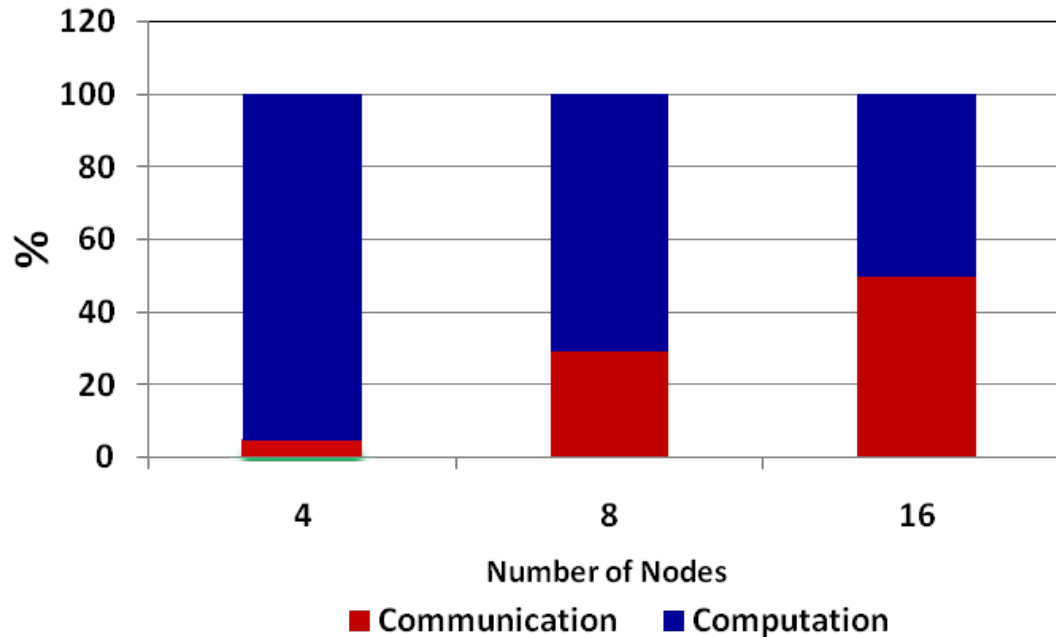


- Quantum ESPRESSO stands for opEn Source Package for Research in Electronic Structure, Simulation, and Optimization
- It is an integrated suite of computer codes for electronic-structure calculations and materials modeling at the nanoscale
- It is based on
  - Density-functional theory
  - Plane waves
  - Pseudopotentials (both norm-conserving and ultrasoft)
- Open source under the terms of the GNU General Public License

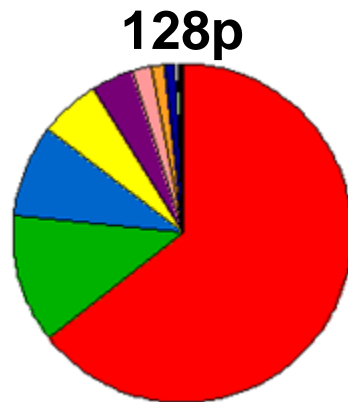
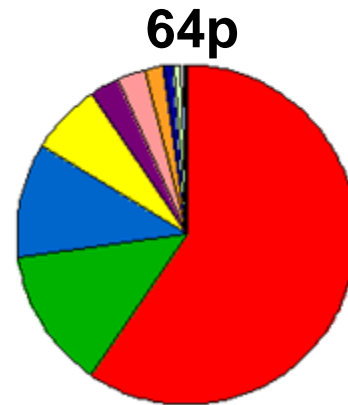
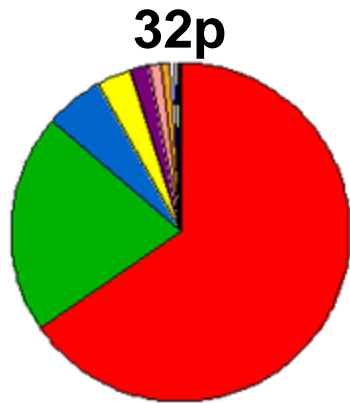


- **Percentage of communication time increases as cluster size scales**
  - 5% at 32 processes, increases up to 50% at 128 processes

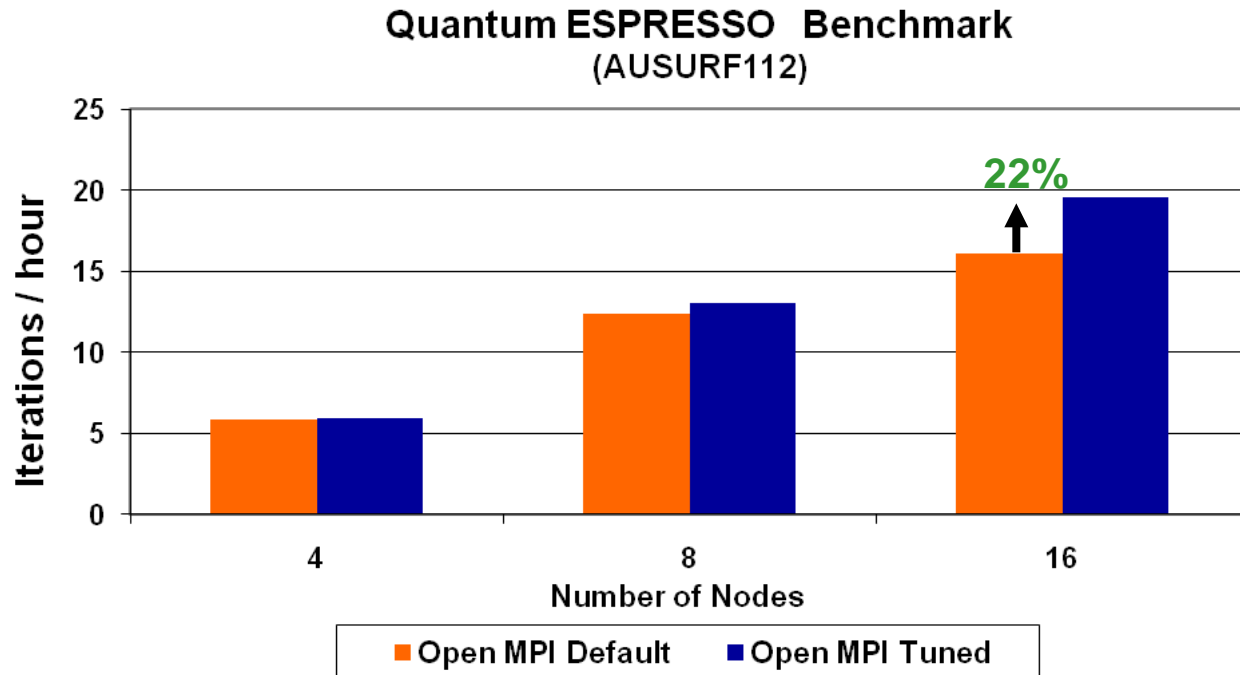
## Runtime Distribution



- Three MPI collectives (MPI\_Barrier, MPI\_allreduce, and MPI\_Alltoallv) consume more than 80% of total MPI time



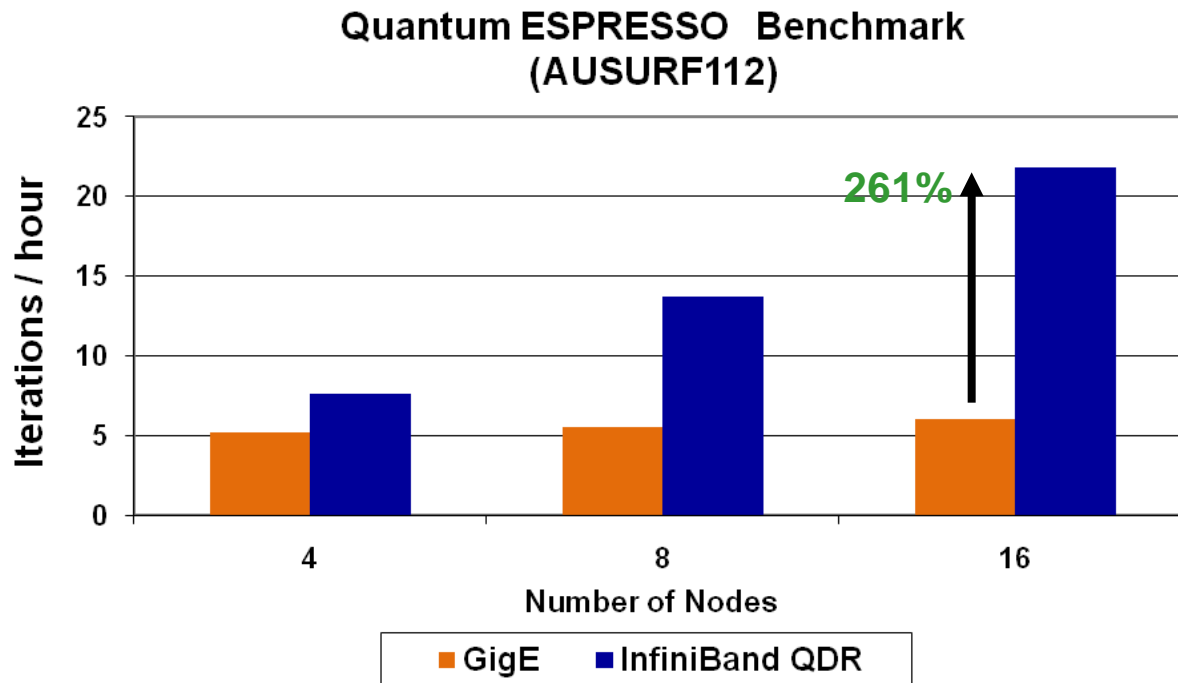
- **Customized MPI parameters provide better performance**
  - **Up to 22% higher performance with Open MPI**
    - `--mca mpi_affinity_alone 1 --mca coll_tuned_use_dynamic_rules 1 --mca coll_tuned_alltoallv_algorithm 2 --mca coll_tuned_allreduce_algorithm 0 --mca coll_tuned_barrier_algorithm 6`



*Higher is better*

8-cores per node

- **InfiniBand enables better application performance and scalability**
  - Up to 261% higher performance than GigE
  - GigE stops scaling after 8 nodes
- **Application performance over InfiniBand scales as cluster size increases**

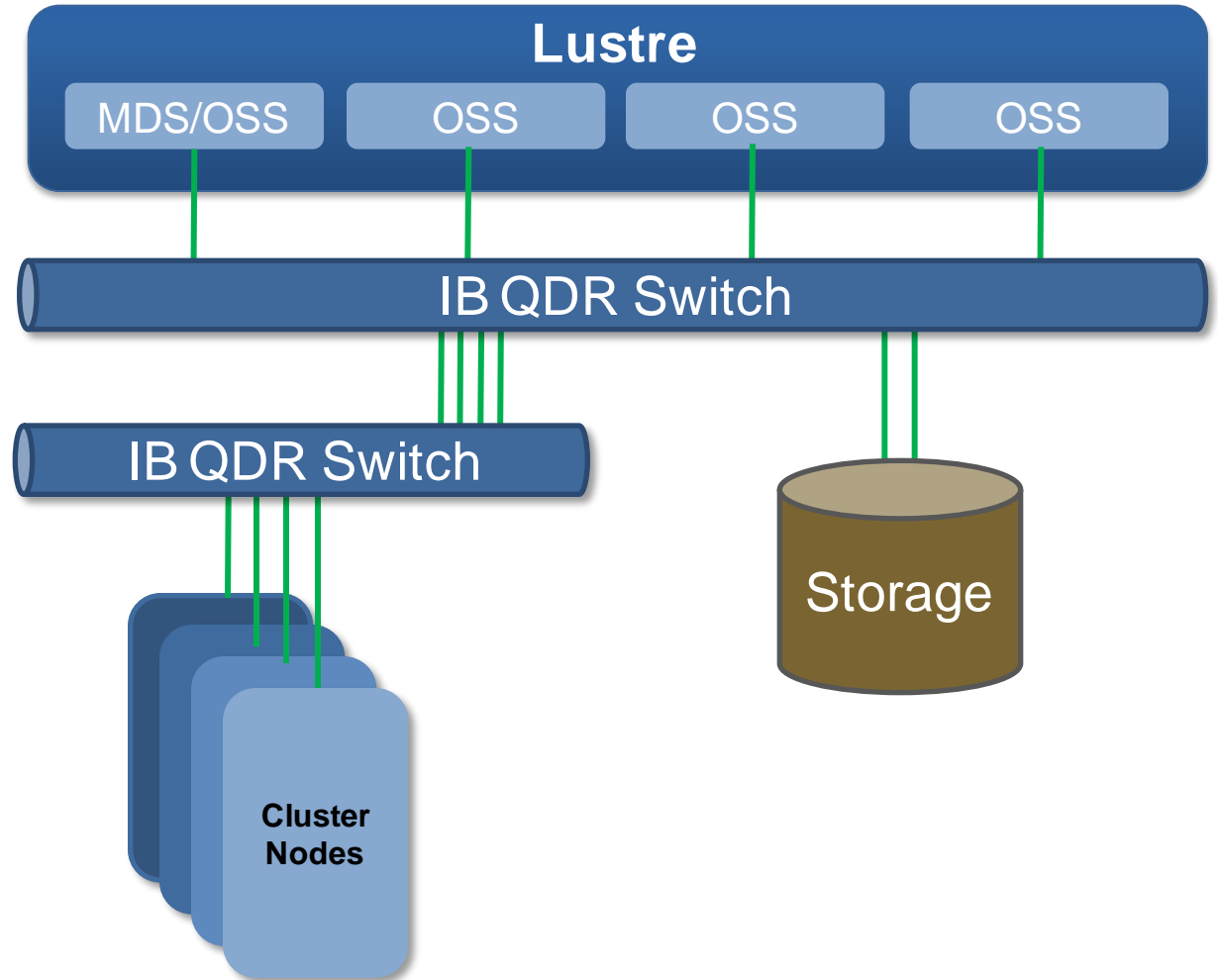


*Higher is better*

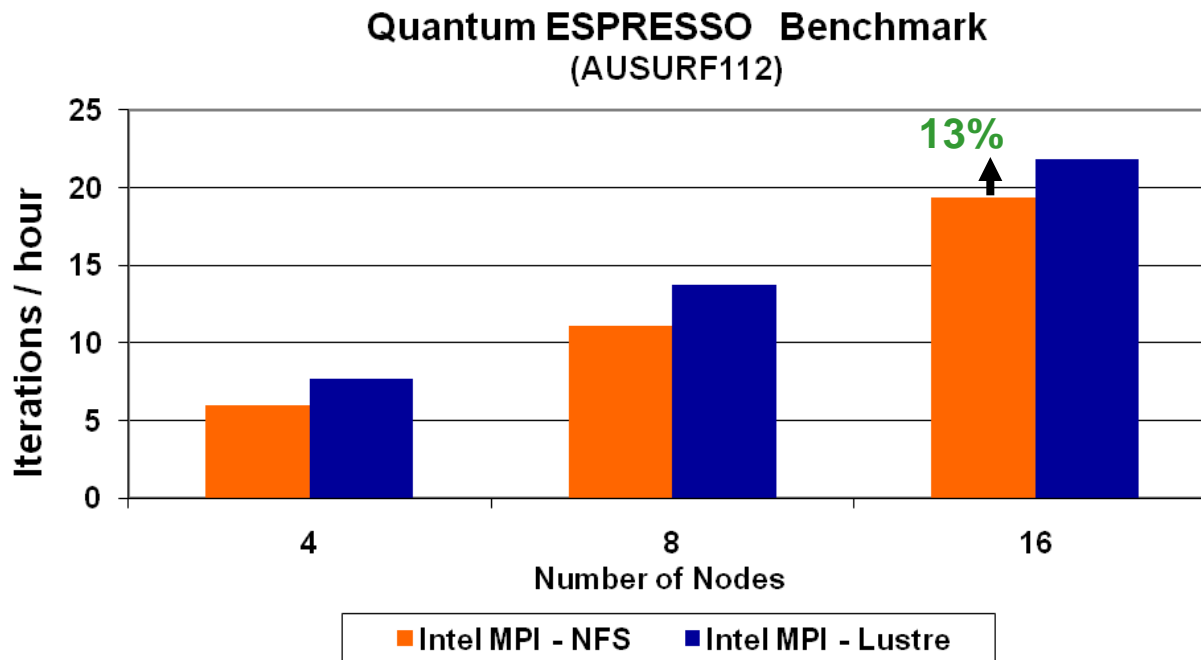
8-cores per node

- **Lustre Configuration**

- 1 MDS
- 4 OSS (Each has 2 OST)
- InfiniBand based Backend storage
- All components are connected through InfiniBand QDR interconnect



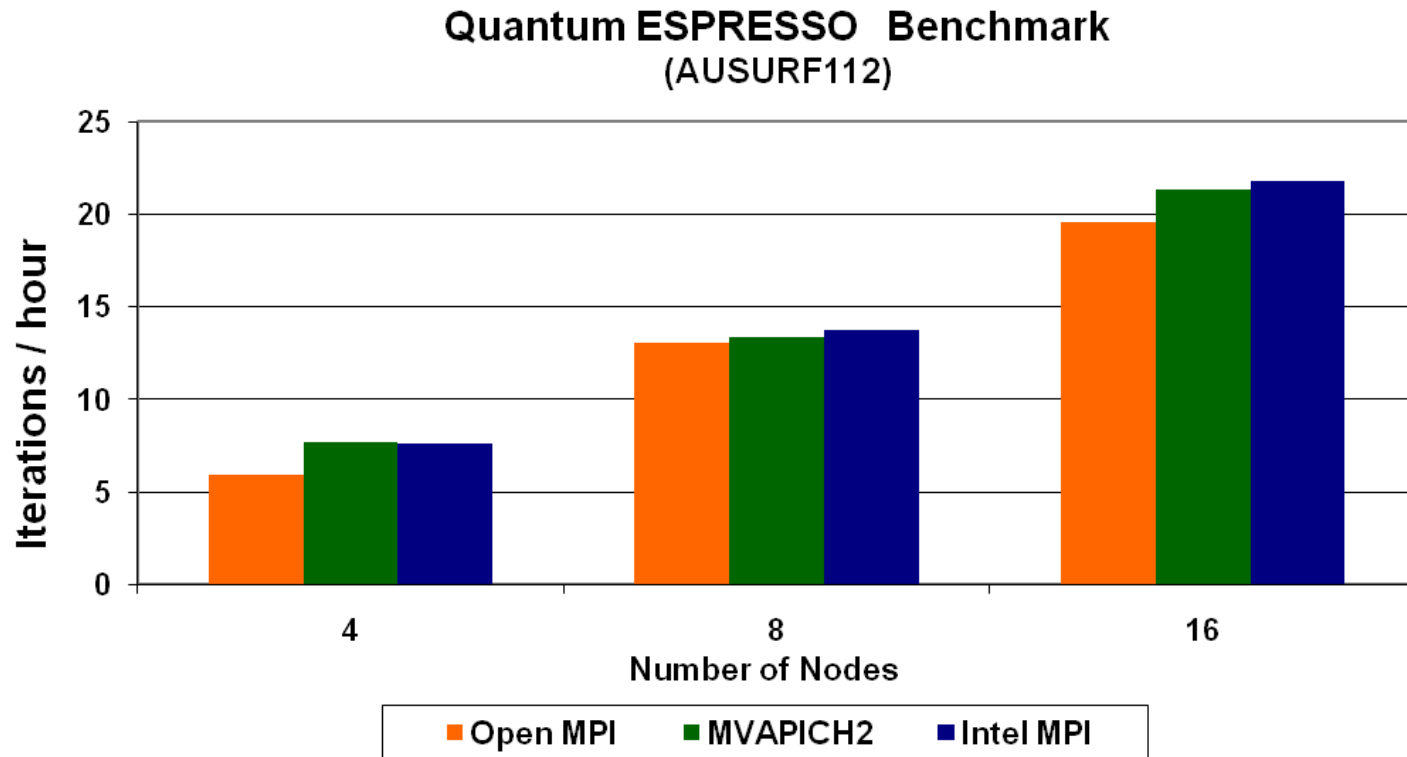
- **Intel MPI has native Lustre support**
  - `mpiexec-genv I_MPI_EXTRA_FILESYSTEM on -genv I_MPI_EXTRA_FILESYSTEM_LIST lustre`
- **Lustre enables higher performance**
  - Up to 13% faster than local hard disk at 16 nodes



*Higher is better*

8-cores per node

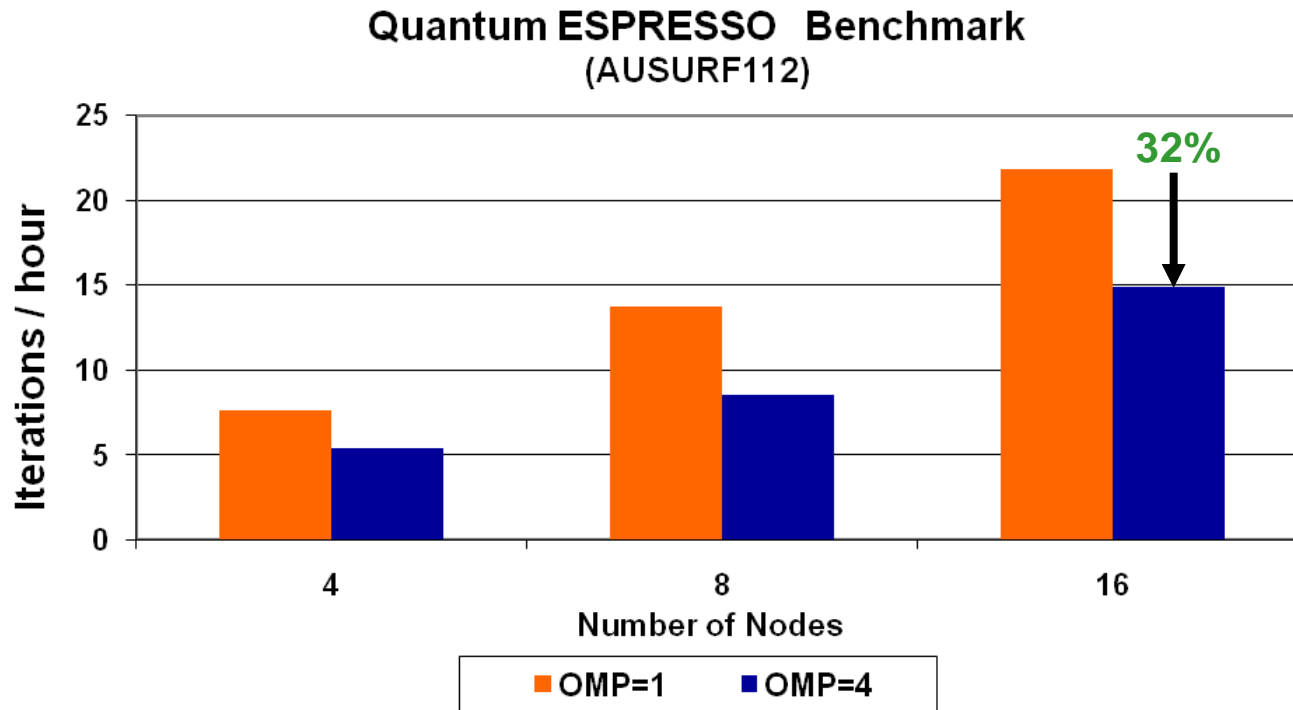
- Intel MPI has comparable performance to MVAPICH2
- 12% faster than Open MPI



*Higher is better*

8-cores per node

- **Multi-thread Intel MPI doesn't provide higher performance**
  - Up to 32% slower than non-threaded application performance



*Higher is better*

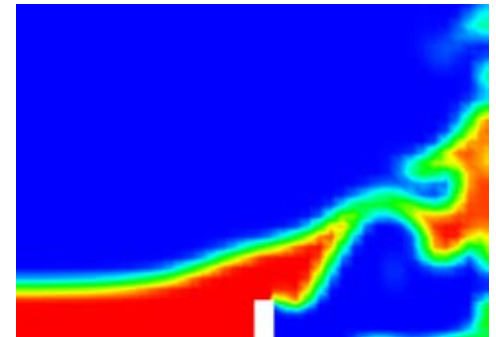
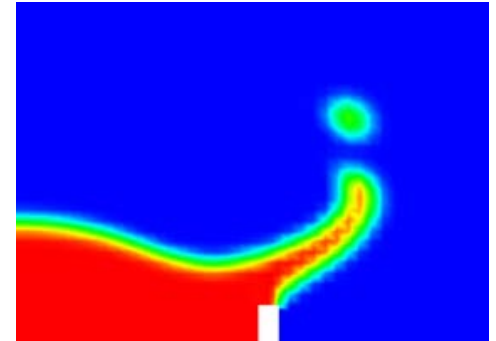
8-cores per node

- **Time in communication increases faster relative to computation**
- **MPI Collective functions dominate total MPI communication time**
  - More than 90% MPI time is spent in MPI collectives
  - Total number of messages increases with cluster size
  - Both small and large messages are used by Quantum ESPRESSO
  - Interconnect latency and bandwidth are critical to application performance
- **Performance Optimization**
  - MPI libraries showed comparable performance overall
  - Lustre with IB delivers increased performance
  - Enabling multi-thread does not yield performance increase
  - InfiniBand continues to deliver superior performance across a broad range of system sizes

- **OpenFOAM® (Open Field Operation and Manipulation) CFD**

**Toolbox can simulate**

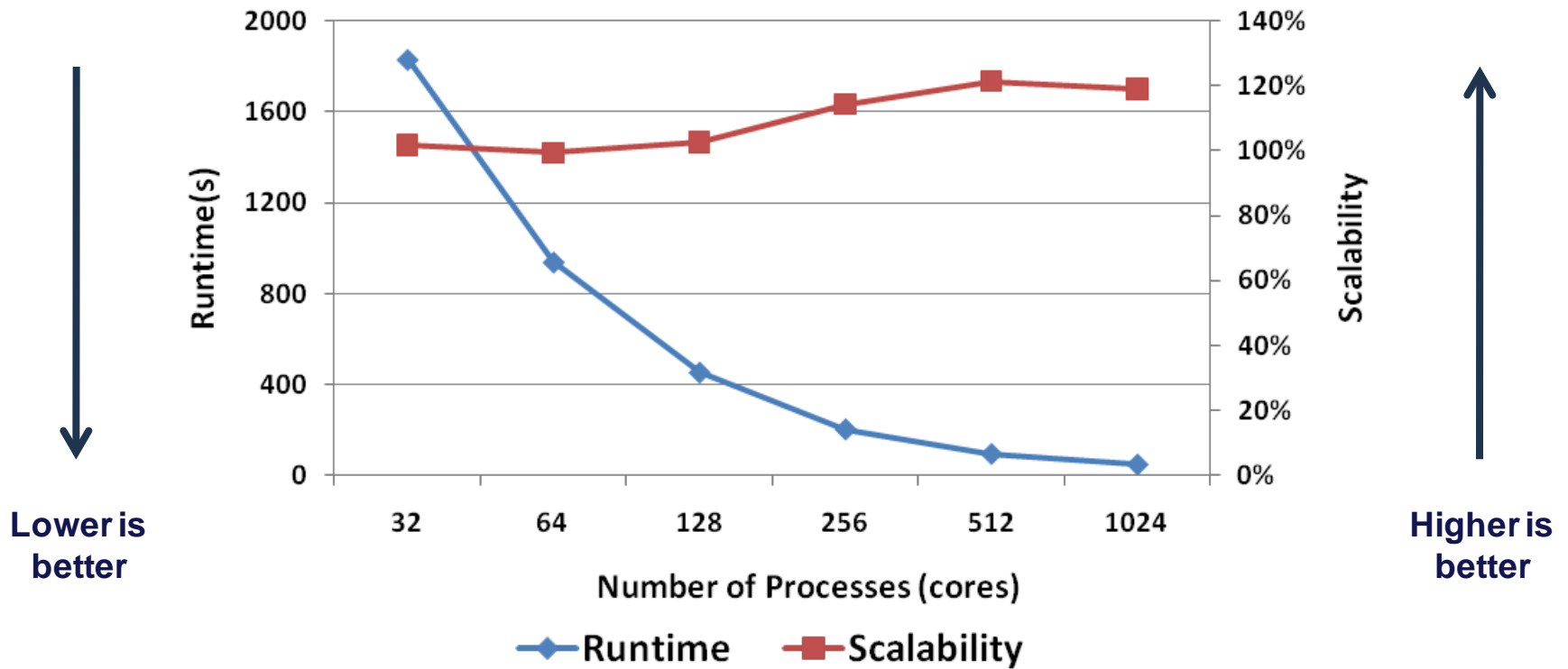
- Complex fluid flows involving
  - Chemical reactions
  - Turbulence
  - Heat transfer
- Solid dynamics
- Electromagnetics
- The pricing of financial options



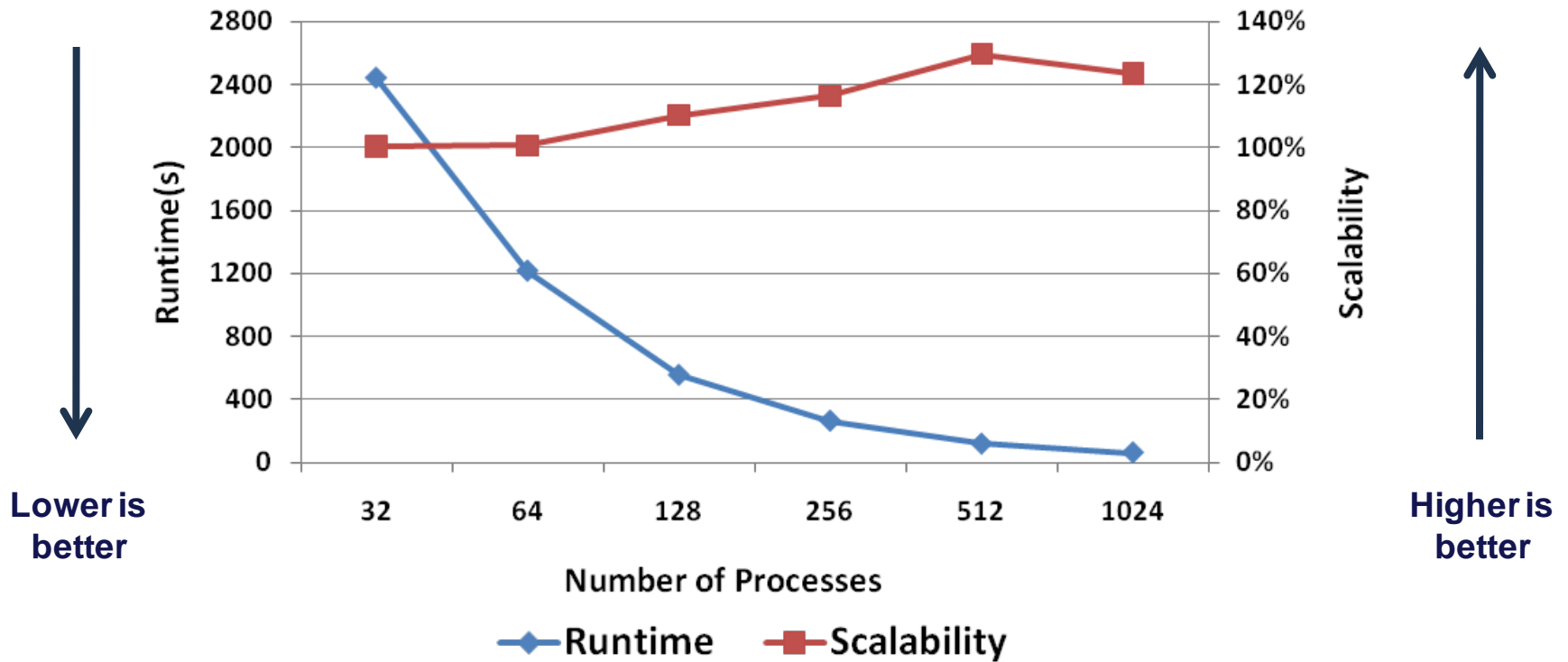
- **OpenFOAM is Open source, produced by OpenCFD Ltd**

- **JuRoPa supercomputer at the Jülich Supercomputing Centre**
  - Dual socket Intel Xeon X5570 quad-core @2.93 GHz
  - 24 GB memory (DDR3, 1066 MHz)
  - Mellanox InfiniBand QDR, non-blocking network configuration
  - SUSE SLES 11
  - ParTec MPI
- **Itasca supercomputer at the Minnesota Supercomputing Institute**
  - HP ProLiant BL280c G6 blade servers
  - Dual socket Intel Xeon X5560 quad-core @2.80 GHz
  - Mellanox InfiniBand QDR, 2:1 blocking network configuration
  - SUSE SLES 11
  - OpenMPI-1.4.2, Platform MPI 2.7.1
- **Application**
  - OpenFOAM 1.7.1
  - Benchmark Dataset: Laminar Cavity Flow (2D, 16Million Cells)

## OpenFOAM 1.7.1 Performance (JuRoPa) (Laminar Cavity Flow Benchmark)

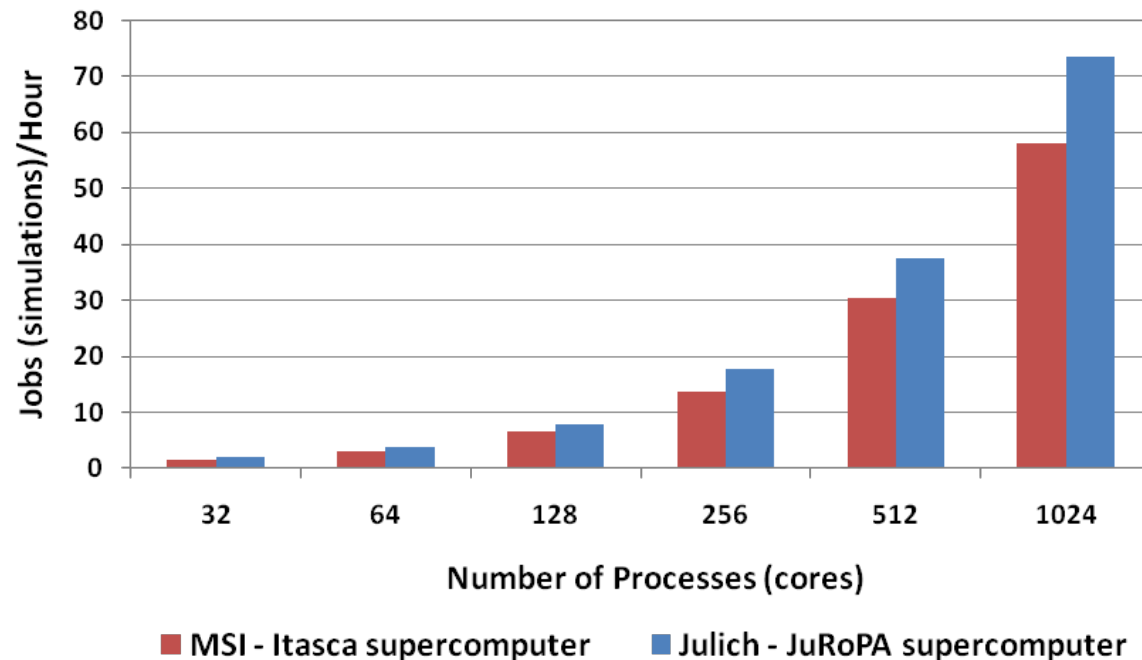


## OpenFOAM 1.7.1 Performance (Itasca) (Laminar Cavity Flow Benchmark)



- **Non-blocking InfiniBand network provides ~27% higher performance versus 2:1 network blocking configuration**

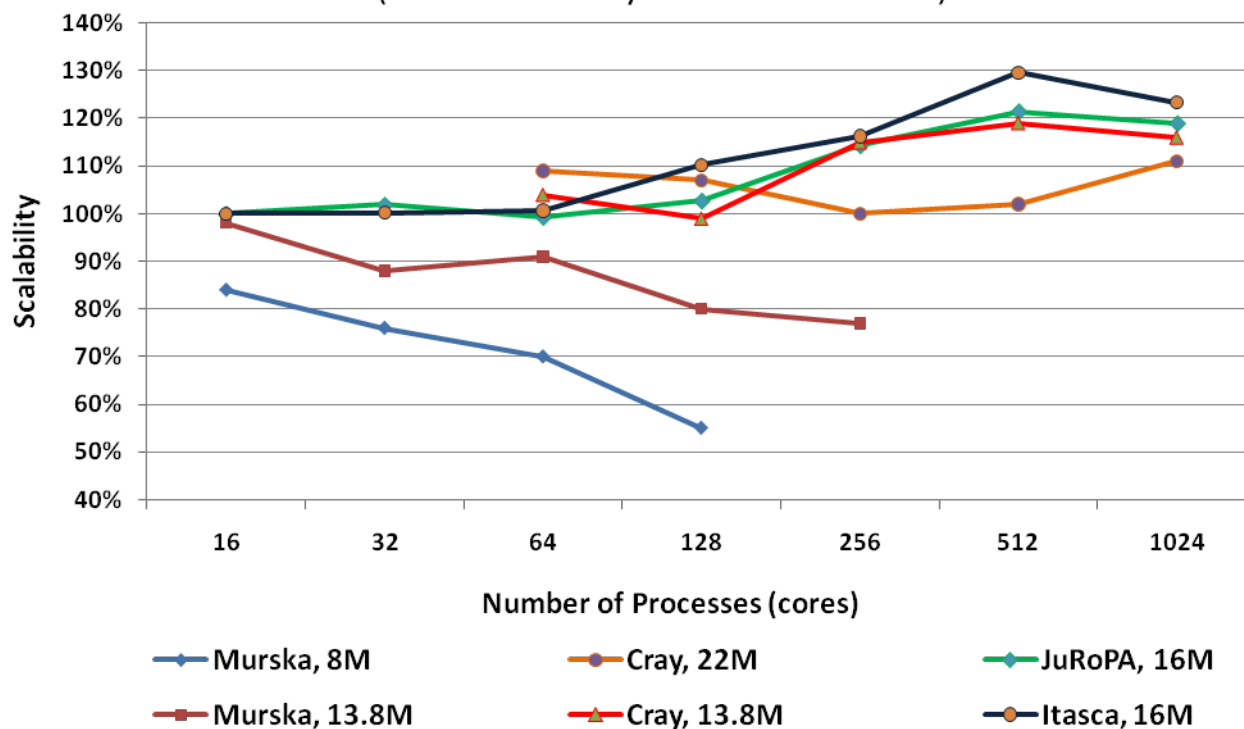
**OpenFOAM 1.7.1 Performance**  
(Laminar Cavity Flow Benchmark)



- **Murska supercomputer at CSC**
  - HP CP4000 BL ProLiant supercluster
  - Dual socket dual-core 2.6GHz AMD Opteron 64-bit CPUs
  - Mellanox InfiniBand DDR, non-blocking network
- **Louhi supercomputer at CSC**
  - Cray XT5 Massively Parallel Processor (MPP) supercomputer
  - Dual socket quad-core AMD 2.3GHz AMD Opteron 64-bit CPUs
  - RHEL 4 Linux operating systems
- **The reference environment are only used for scalability comparisons, and not for performance comparison**

- **InfiniBand QDR delivers highest scalability**
  - Versus Cray XT5, and versus InfiniBand DDR
  - Murska – IB DDR, JuRoPa and Itasca – IB QDR

### OpenFOAM 1.7.1 Performance (Laminar Cavity Flow Benchmark)



- **OpenFoam demonstrates good scaling capabilities**
  - Testing includes systems configuration up to 1K cores
- **For OpenFOAM, non blocking network delivers higher performance compared to 2:1 blocking configuration**
  - 27% higher performance in average
- **InfiniBand QDR demonstrates highest scalability**
  - Compared to Cray XT5 and InfiniBand DDR

- 硬件基础上的集群通信卸载 (Collective Offload)
- 更先进的并行编程方法

# 谢谢

HPC Advisory Council

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

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

