

HPC 最佳实例

Tools, Compilers, and profiling

刘通

2010年10月

国际高性能计算咨询委员会中国区总监

- **免费编译器**

- GNU Compiler
- Open64

- **付费编译器**

- PGI
- Intel

- **數學函数庫**

- ACML
- Intel MKL

- **MPI 库**

- Open MPI
- MVAPICH
- Platform MPI
- Intel MPI
- ParaStation MPI

- **BIOS 设置**
- **CPU/Memory**
- **网络**
 - Driver/Firmware
 - 底层性能测试
 - MPI 性能
 - 点对点以及集群通信性能
 - Performance counters (性能计数器)
- **File system (文件存储系统)**
 - Disk performance (硬盘的性能)
 - Shared file system performance (共享文件系统的性能)

- **Platform MPI (previously HP-MPI)**

- *-i file*

User time: 58.22%

MPI time : 41.78% [Overhead:41.74% Blocking:0.04%]

Routine Summary by Rank:

Rank	Routine	Statistic	Calls	Overhead(ms)	Blocking(ms)
------	---------	-----------	-------	--------------	--------------

0	MPI_Allgather		3078	672.696352	0.000000
---	---------------	--	------	------------	----------

- **Open MPI**

- IPM (LD_PRELOAD=~/.ipm/lib/libipm.so)

#	[time]	[calls]	<%mpi>	<%wall>
# MPI_Allreduce	741.508	1.51281e+07	86.45	22.34
# MPI_Waitall	44.2919	1.40326e+07	5.16	1.33

- **Intel MPI**

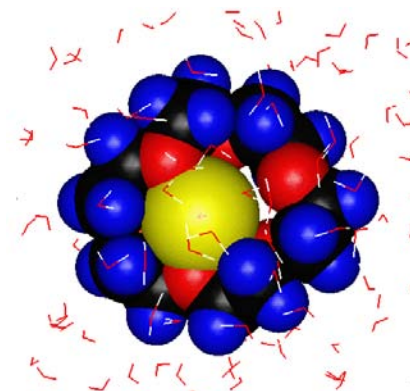
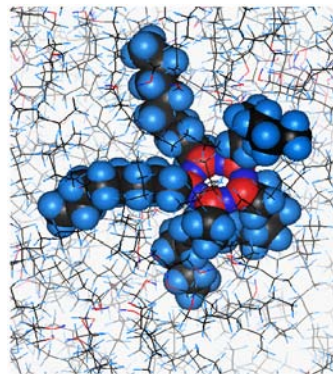
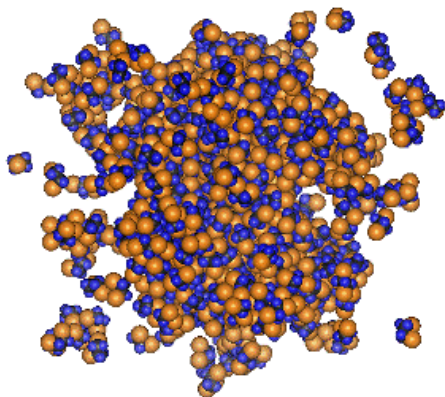
- `-genv I_MPI_STATS=10`

- **Platform MPI (previously Scali MPI)**

- `SCAMPI_TRACE="-f arg;timing"`

- `/opt/scali/bin/scanalyze -m trace tracefile.txt`

- NWChem 是量子化学计算软件
 - 美国太平洋西北国家实验室的环境分子科学研究室开发
- NWChem 为分子学计算提供了多种方法
- NWChem 能进行传统的分子动力学以及自由能模拟



- **编译**

- 环境设置

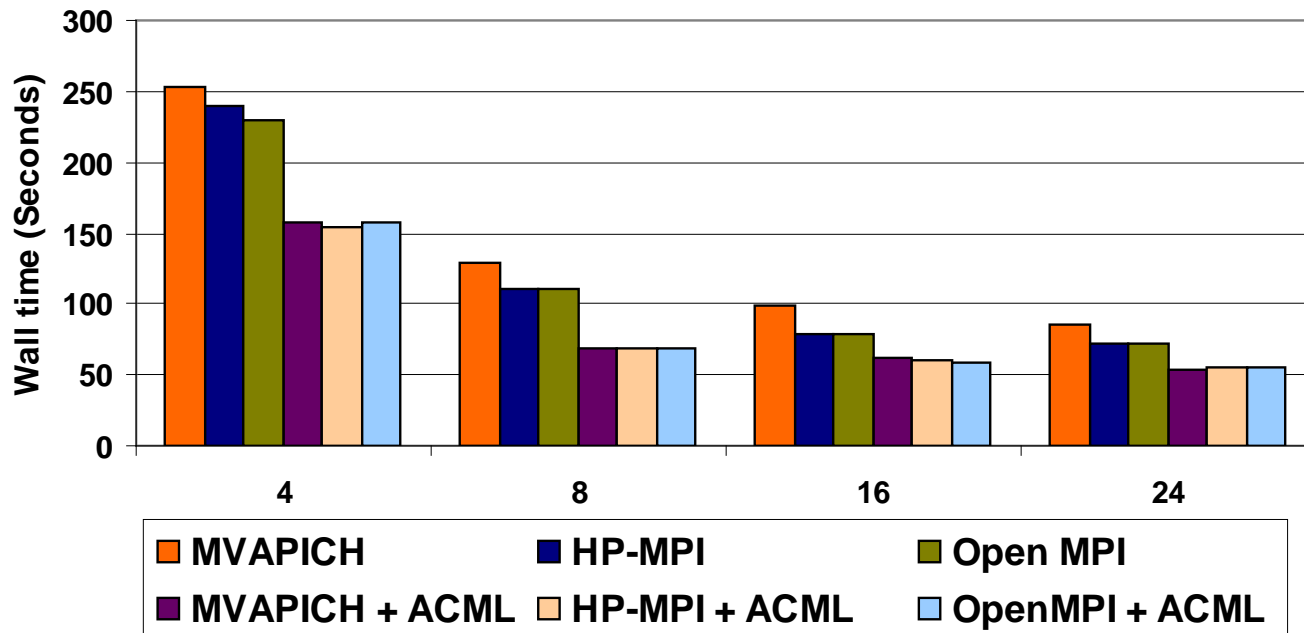
- export ARMCI_NETWORK=OPENIB
- export MPI_LOC=/usr/mpi/gcc/openmpi-1.4.1
- export LIBMPI="-lmpi"
- export BLAS_LIB="-L/acml4.3.0/ifort64/lib -lacml -lacml_mv "
- export BLASOPT="-L/acml4.3.0/ifort64/lib -lacml -lacml_mv "

- **运行**

- 工作路径应该是本地硬盘，最好是并行文件存储系统

- Input Dataset - H2O7
- ACML提供了更好的性能及可扩展性

NWChem Benchmark Result (H₂O₇ MP2)

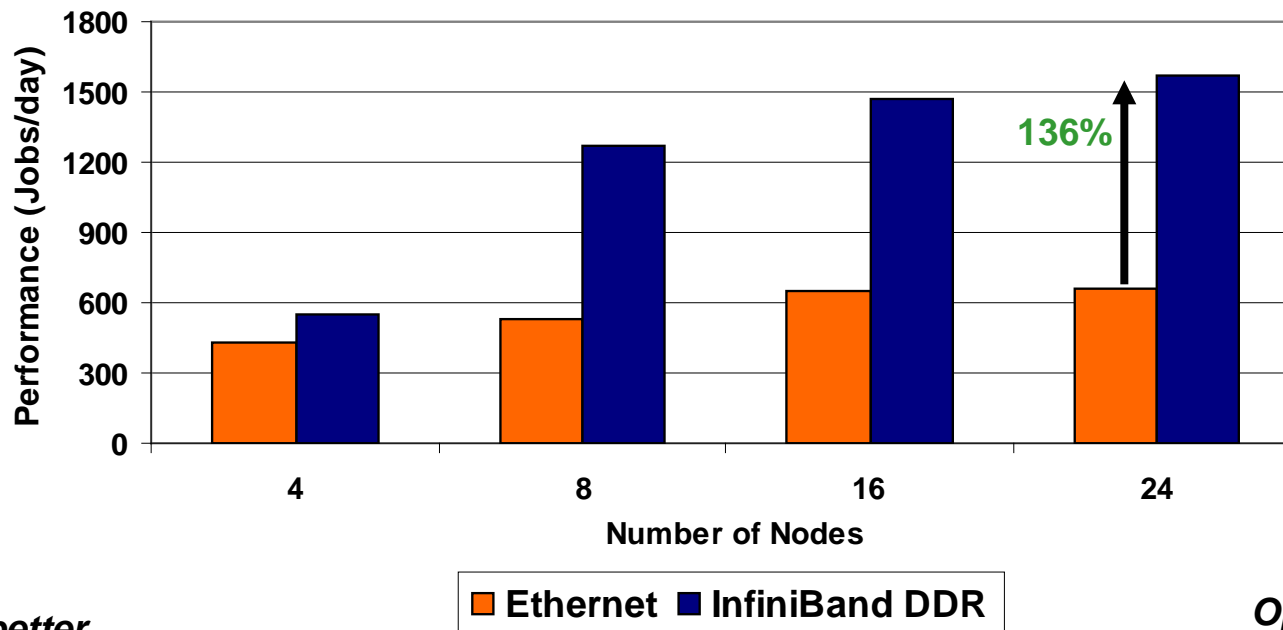


Lower is better

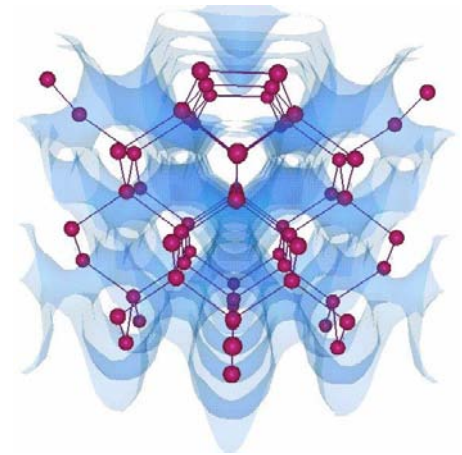
InfiniBand DDR

- Input Dataset - H2O7
- InfiniBand提供了更好的性能及可扩展性
 - 速度是以太网2.36倍

NWChem Benchmark Result
(H₂O₇ MP2)



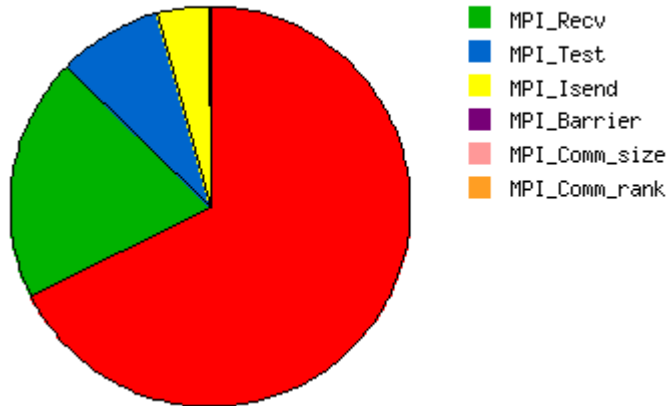
- **OpenAtom是并行量子化学软件**
 - 采用the Car-Parrinello ab-initio Molecular Dynamics 方法解决：
 - Material science
 - Chemistry
 - Solid-state physics
 - Biophysics
- **基于 Charm++ 并行编程开发框架**
- **OpenAtom 是由美国UIUC大徐开发的开源软件**



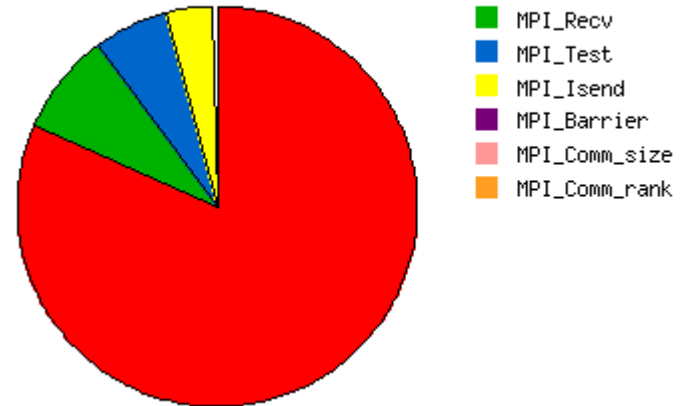
- 主要应用的MPI程式

- MPI_Iprobe, MPI_Recv, MPI_Test, and MPI_Isend
- 点对点通信传输产生最大负载

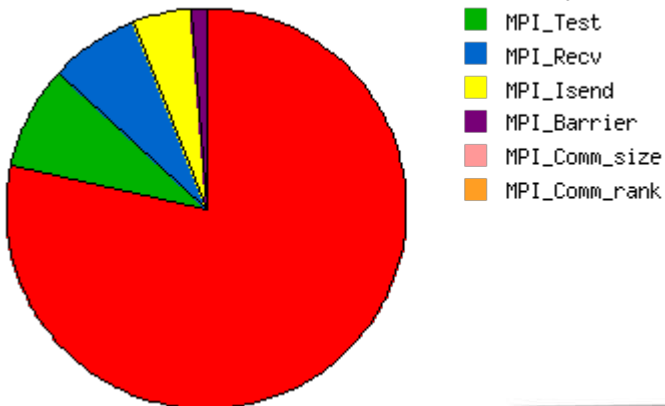
16 Processes



32 Processes



64 Processes

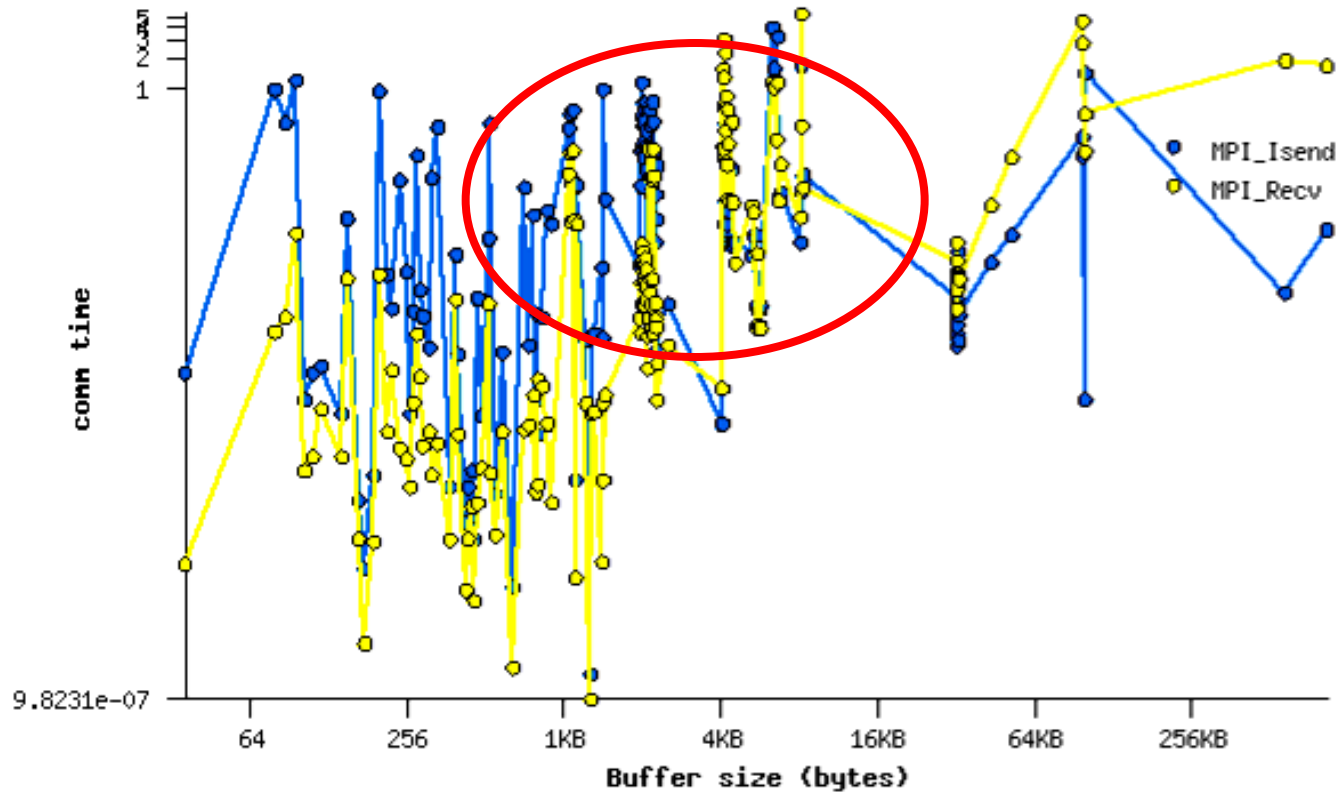


128 Processes



OpenAtom性能分析– MPI 消息大小

- 大部分消息都在1KB-16KB



128 Processes

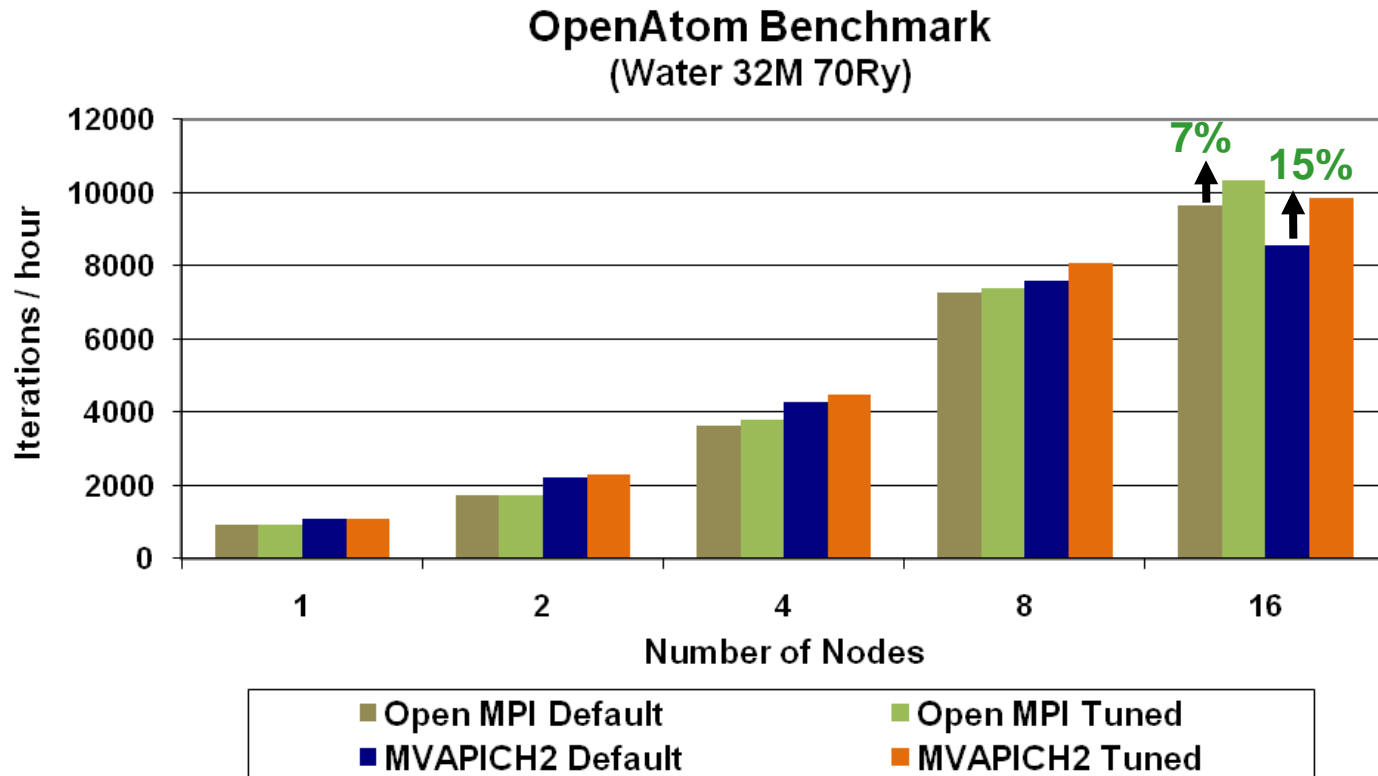
- 优化的MPI 运行参数可以加速程序运行速度

- Up to 7% higher performance with Open MPI

- `--mca mpi_affinity_alone 1 --mca btl_openib_eager_rdma_threshold 8`

- Up to 15% higher performance with MVAPICH2

- `MV2_USE_RDMA_FAST_PATH=0 MV2_USE_RDMA_ONE_SIDED=0`

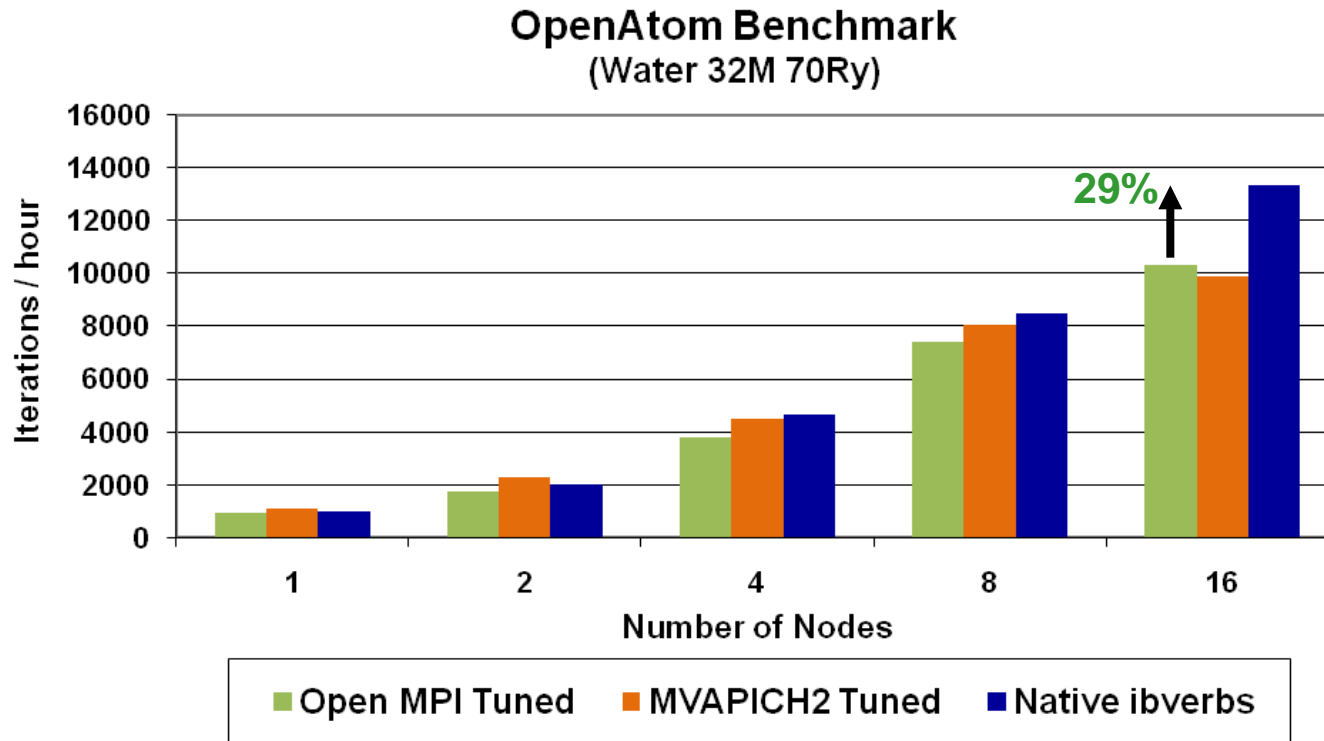


Higher is better

8-cores per node

OpenAtom性能测试结果 - ibverbs

- 直接使用ibverbs 能更好提供运行速度及可扩展性
 - Up 29% higher performance versus tuned Open MPI
 - Performance advantage increases as cluster size scales

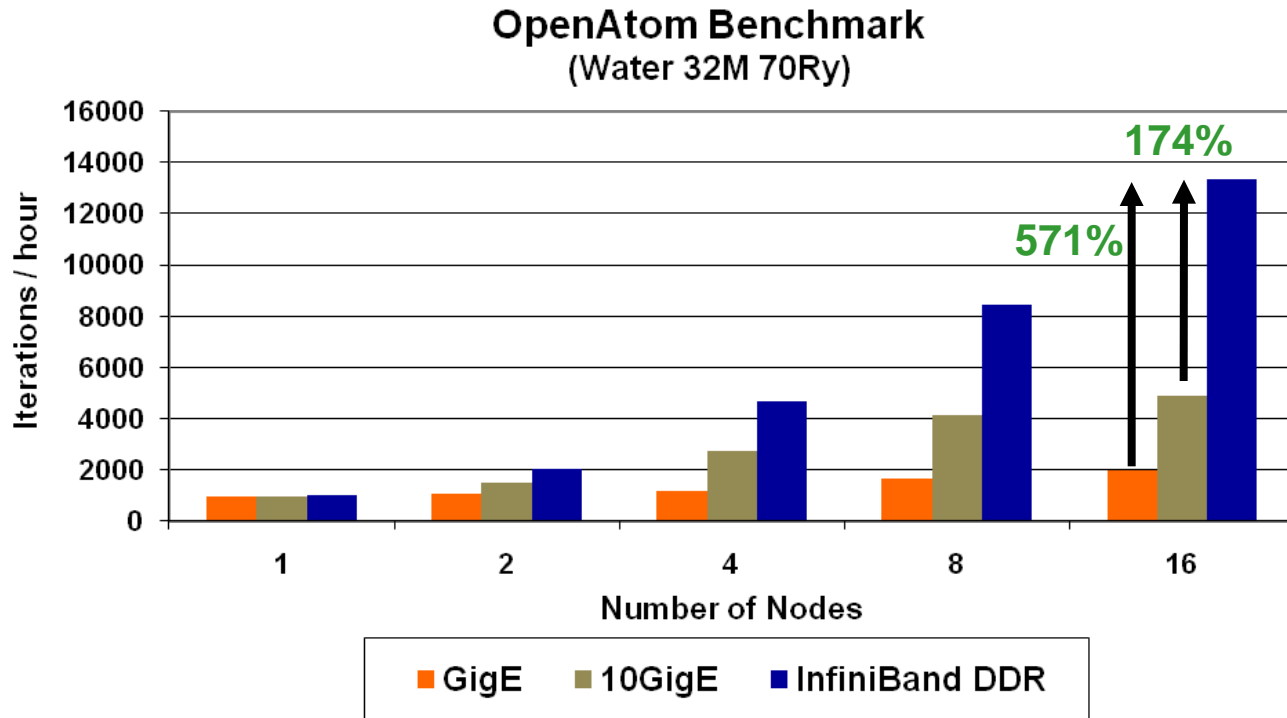


Higher is better

8-cores per node

OpenAtom性能测试结果 - Interconnect

- **InfiniBand 提供了更快的系统性能**
 - 比万兆以太网快1.74倍，比千兆以太网快5.71倍
- **随着系统规模的增加，InfiniBand能保持程序的可扩展性**



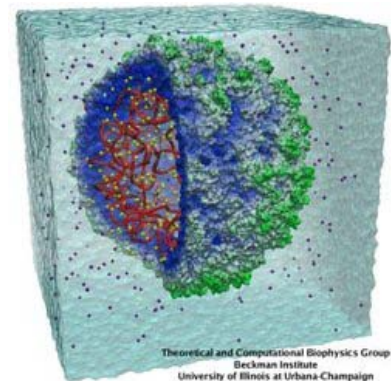
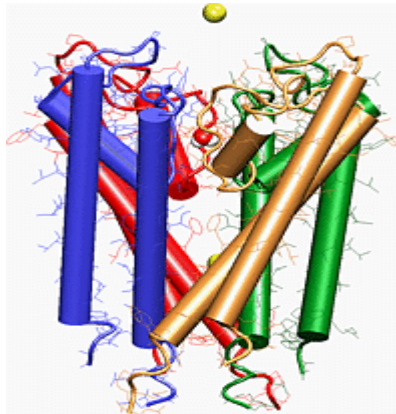
Higher is better

8-cores per node

- 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



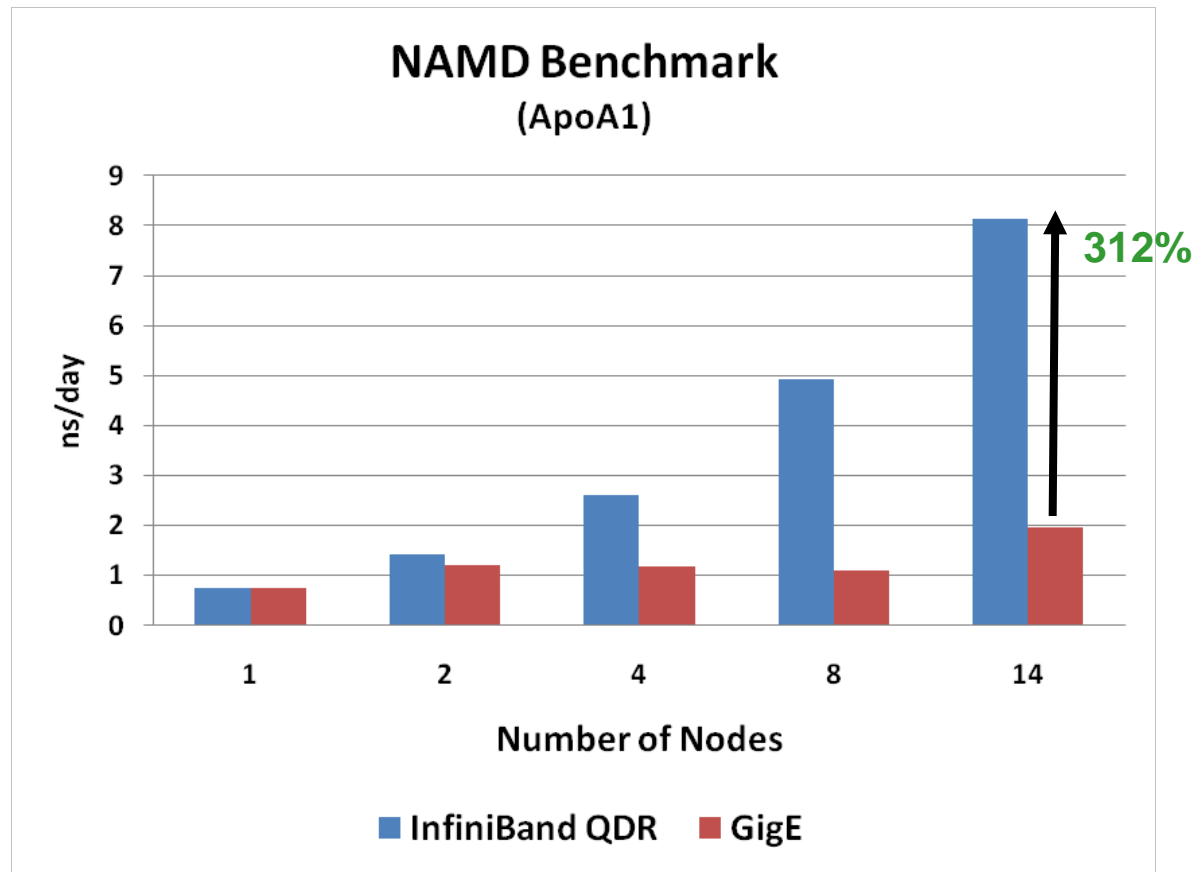
Theoretical and Computational Biophysics Group
Beckman Institute
University of Illinois at Urbana-Champaign



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

- **InfiniBand enables higher scalability**

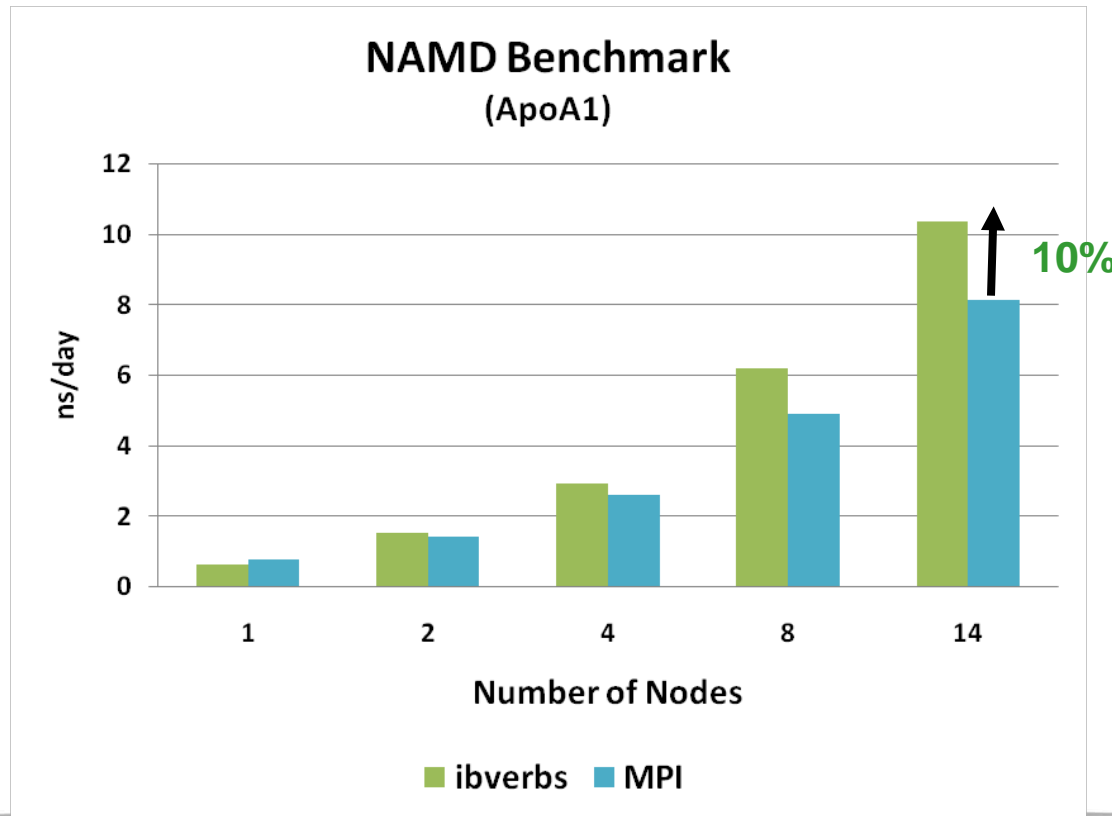
- Up to 312% higher performance than Ethernet at 14 nodes
- Four InfiniBand connected servers deliver higher performance vs 14 Ethernet connected servers



Higher is better

12 Cores/Node

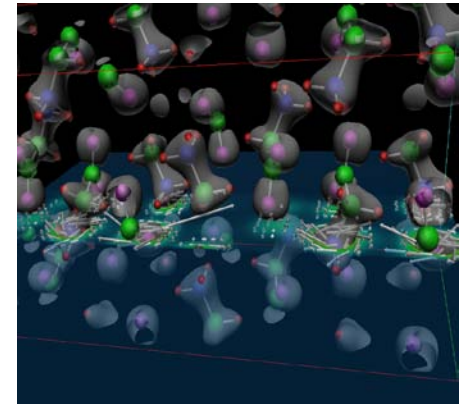
- NAMD can use MPI or the native InfiniBand interface (verbs) as the interface to the interconnect (InfiniBand)
- IB verbs version provides better productivity versus MPI
 - 10% improvement over 14 nodes
 - IB verbs provides a lower level interface to the interconnect versus MPI – lower overhead



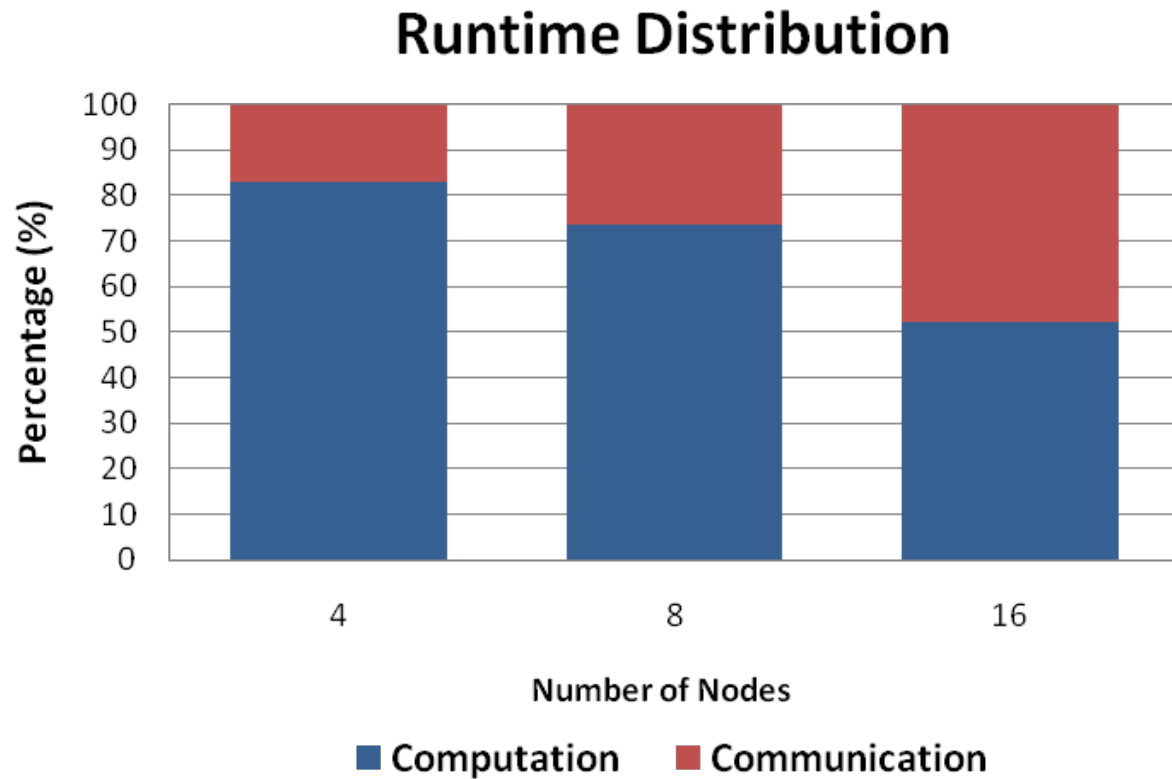
Higher is better

12 Cores/Node

- PARATEC stands for PARAllel Total Energy Code
- Performs ab-initio quantum-mechanical total energy calculations using pseudopotentials and a plane wave basis set
- Designed to run on massively parallel computing platforms and clusters
- Developed through a joint collaboration between
 - LBNL
 - Université Pierre et Marie CURIE
 - University of Montreal
 - University of Cambridge

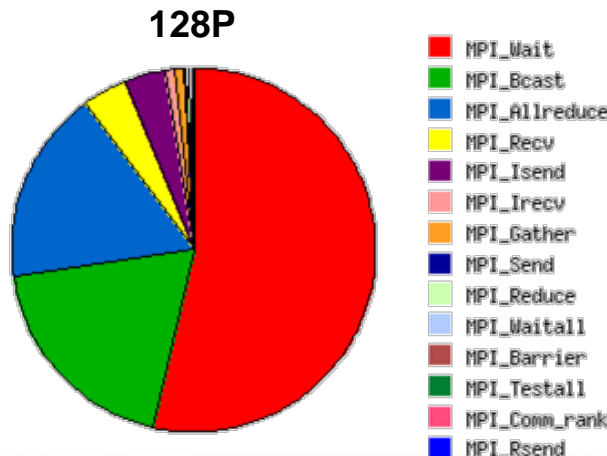
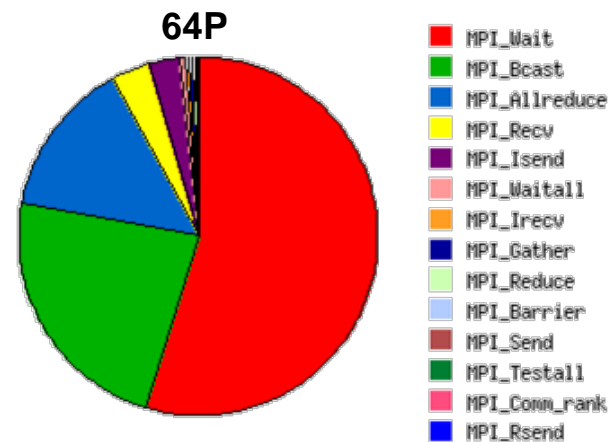
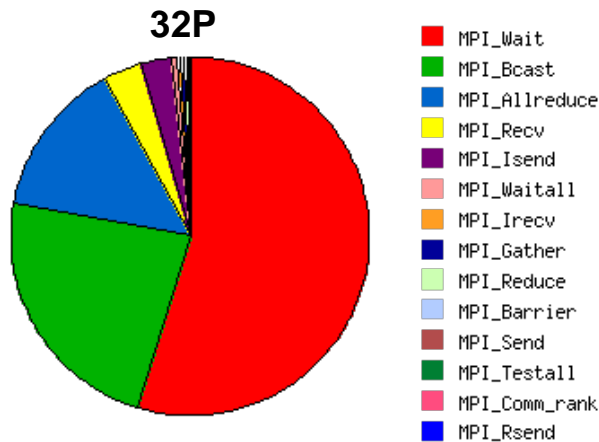


- 随着系统大小的增加，通信时间所占的比例持续增加

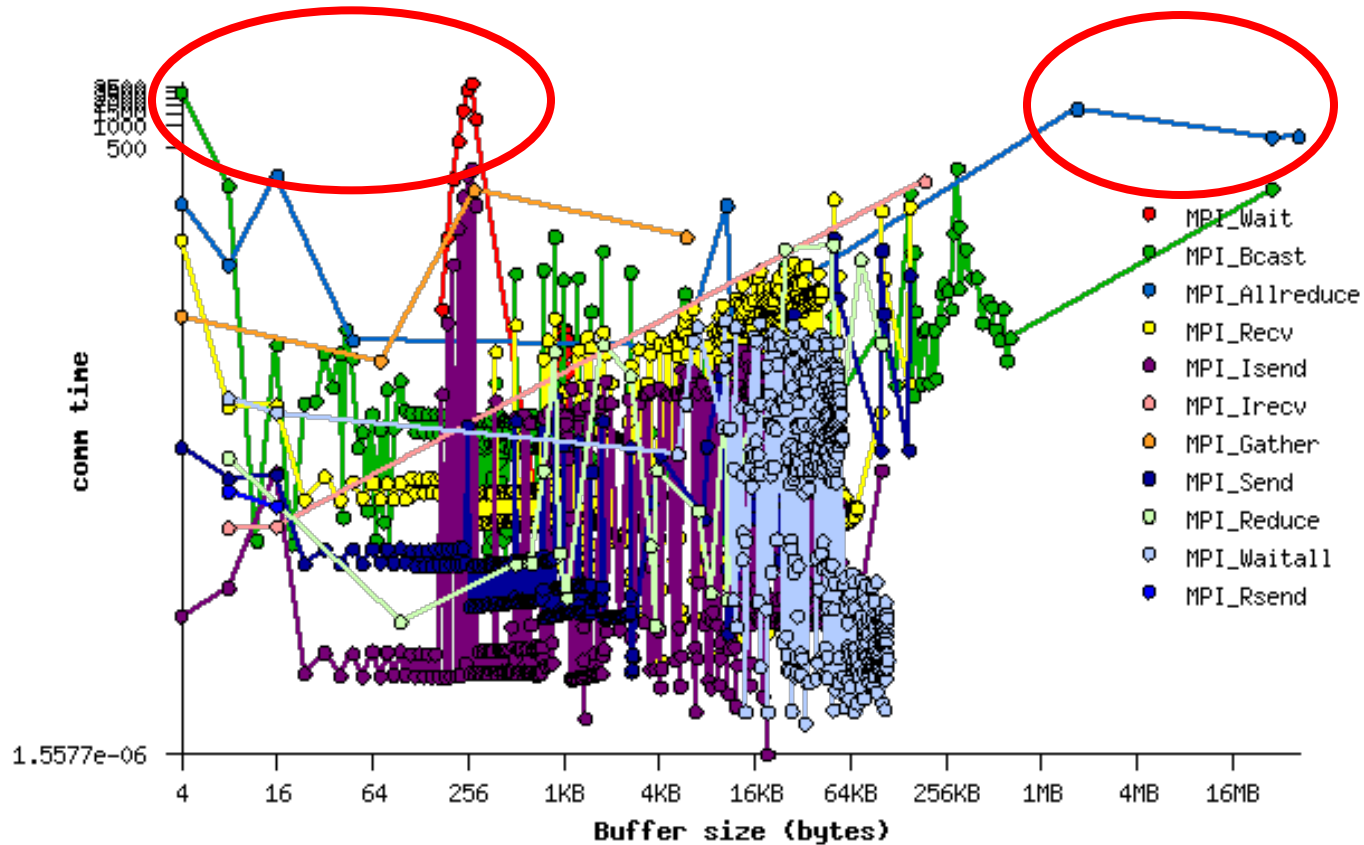


- **Mostly used MPI functions**

- MPI_Wait, MPI_Allreduce, and MPI_Bcast are the mostly used MPI functions
- MPI_Allreduce 负载比例持续增加



- **Messages with big communication overhead are**
 - Large messages >1MB
 - Small message <256Bytes

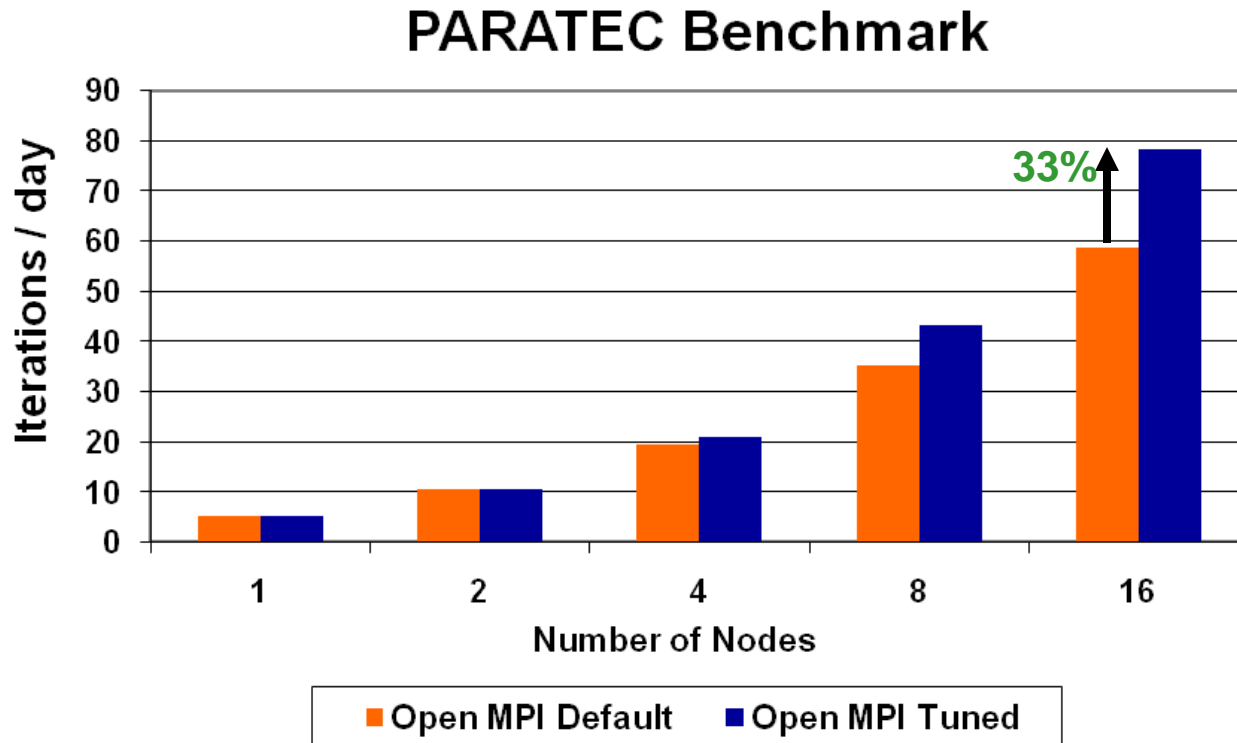


128 Processes

- 优化的MPI 运行参数提供更高程序性能

- Up to 33% higher performance with customized MPI_Gather, barrier, and XRC parameter

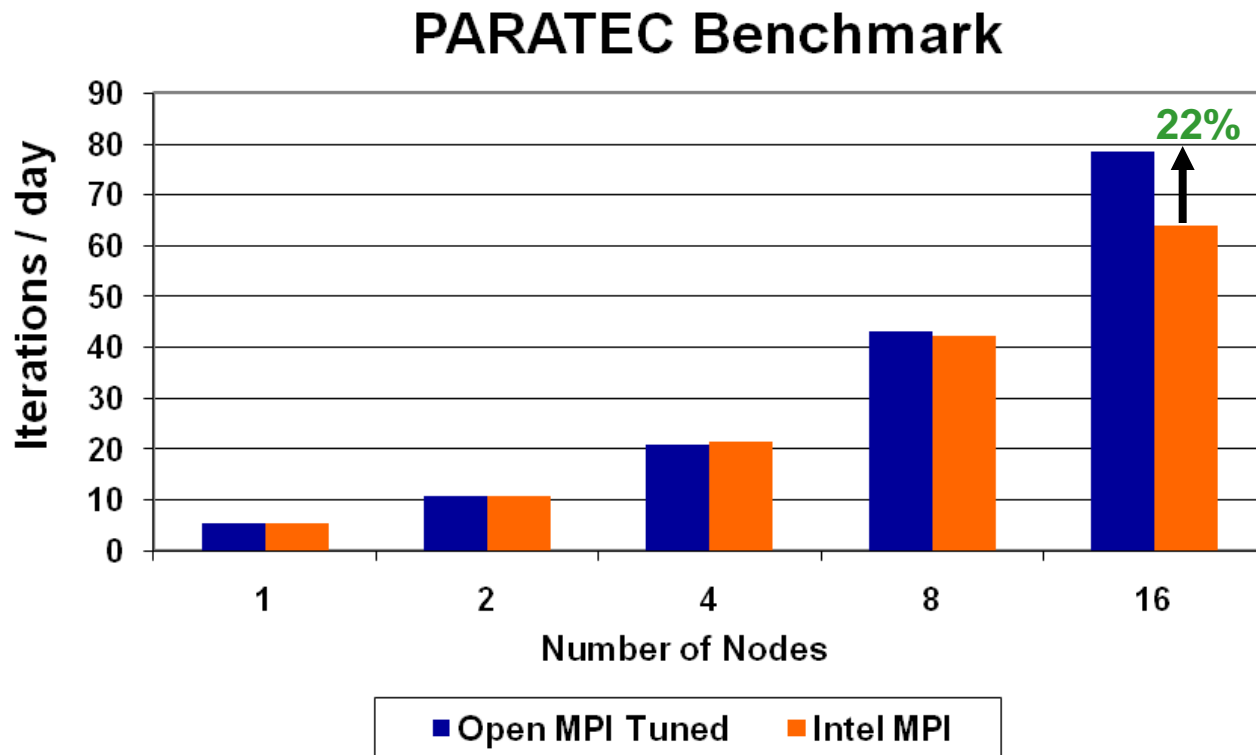
- `--mca btl_openib_receive_queues X,9216,256,128,32:X,65536,256,128,32 --mca coll_tuned_use_dynamic_rules 1 --mca coll_tuned_gather_algorithm 1 --mca coll_tuned_barrier_algorithm 3`



Higher is better

8-cores per node

- **Open MPI with optimization enables higher performance**
 - Up to 22% higher performance than Intel MPI

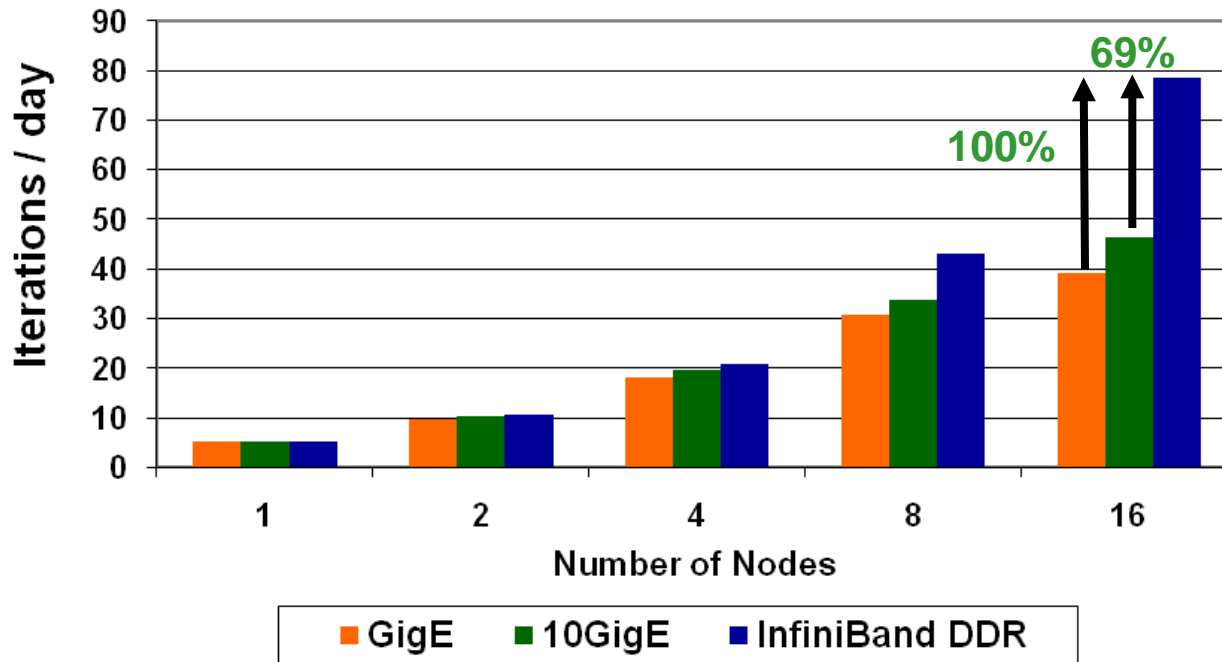


Higher is better

8-cores per node

- **InfiniBand enables better application performance and scalability**
 - Up to 69% higher performance than 10GigE and 100% than GigE
 - 16-node cluster
- **Application performance over InfiniBand scales as cluster size increases**

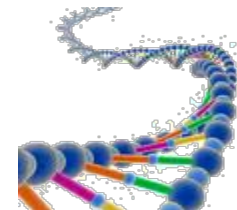
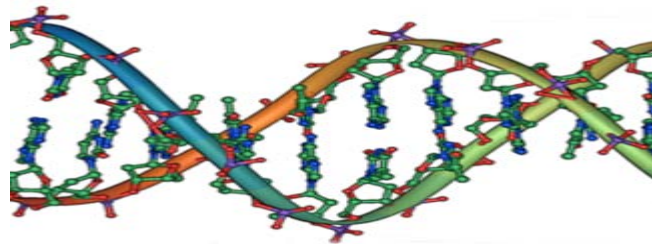
PARATEC Benchmark



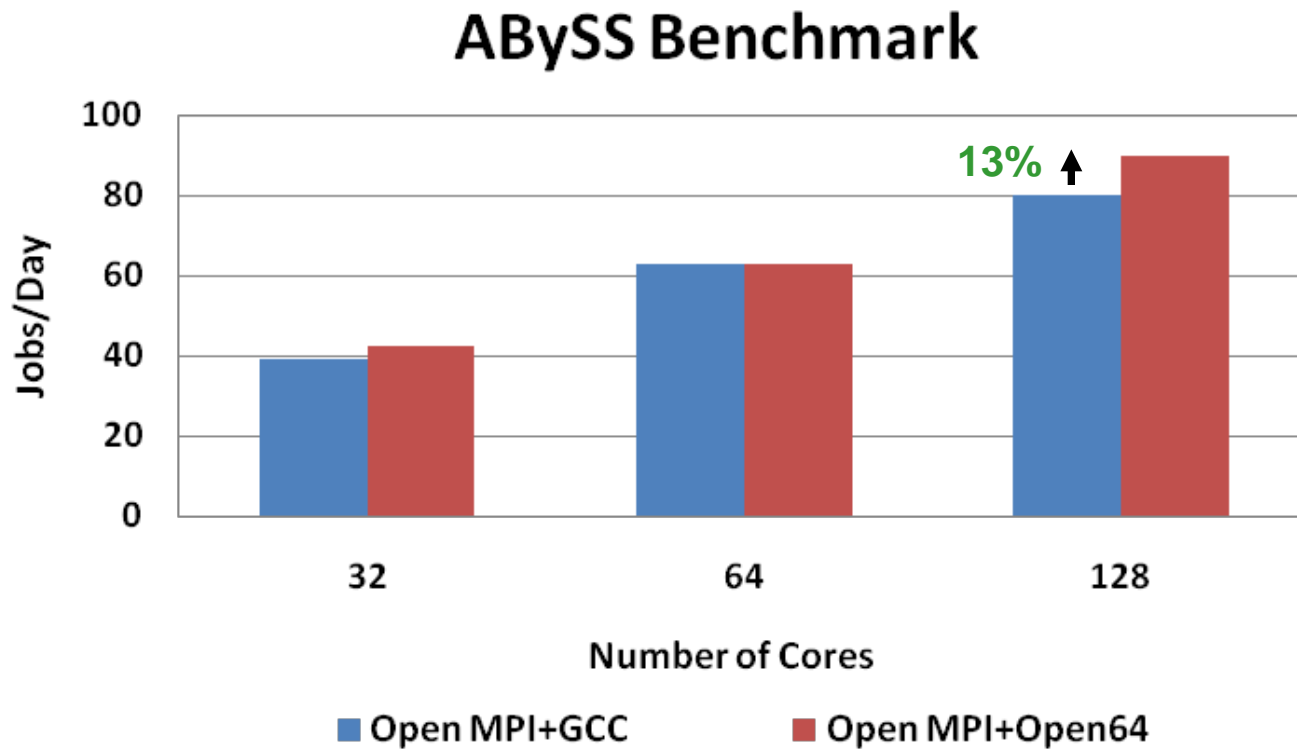
Higher is better

8-cores per node

- **ABYSS is a de novo, parallel, paired-end sequence assembler designed for short reads**
 - Capable of assembling larger genomes
 - Implemented using MPI
- **ABYSS was developed at Canada's Michael Smith Genome Sciences Centre**



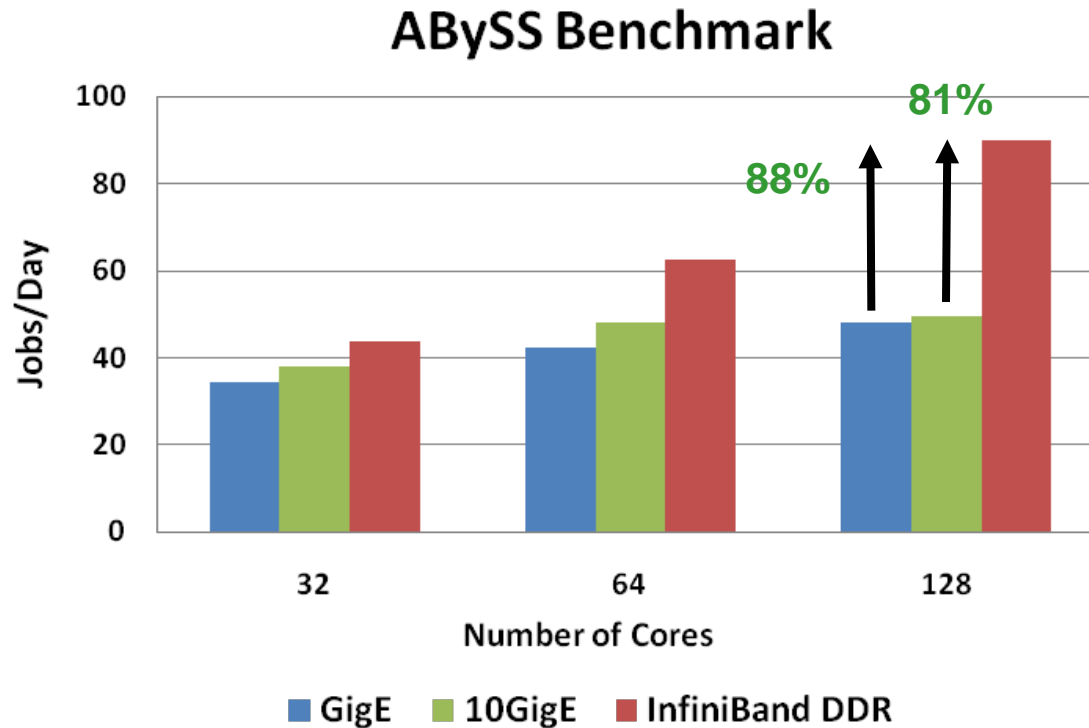
- **Two different compilers were used to compile ABySS**
 - Open64 provides better performance than GCC at 128 cores



Higher is better

8-cores per node

- **InfiniBand enables higher performance and scalability**
 - Up to 88% higher performance than GigE and 81% higher than 10GigE
 - Both GigE and 10GigE don't scale well beyond 8 nodes

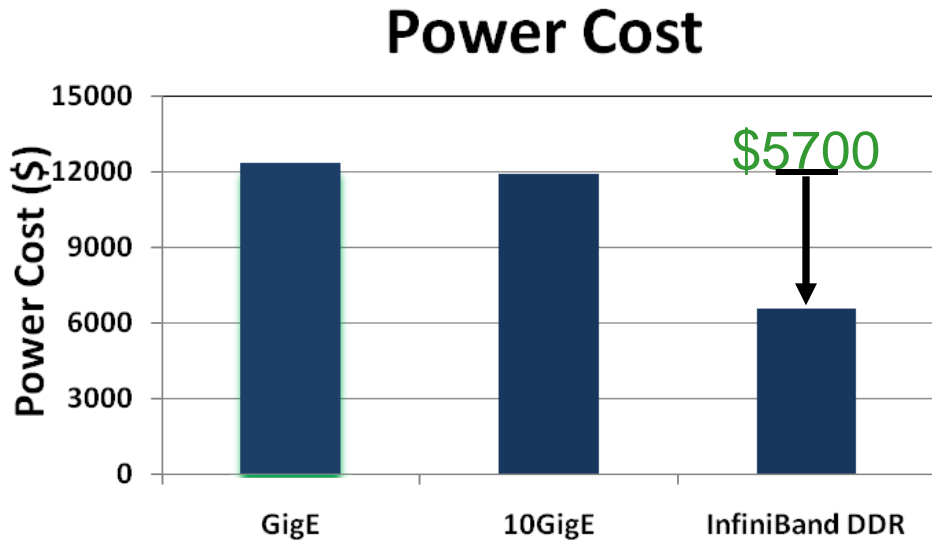


Higher is better

8-cores per node

Power Cost Savings with Different Interconnect

- **To achieve same number of ABySS jobs over GigE**
 - InfiniBand saves power up to \$5700 versus GigE and \$5300 versus 10GigE
 - Yearly based for 16-node cluster
- **As cluster size increases, more power can be saved**



$\$/KWh = KWh * \0.20

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

- **Interconnect comparison shows**
 - InfiniBand delivers superior performance in every cluster size versus GigE and 10GigE
 - Performance advantage extends as cluster size increases
- **Open64 compiler delivers higher performance**
- **InfiniBand enables power saving**
 - Up to \$5700/year power savings versus GigE and \$5300 versus 10GigE on 16 node cluster
 - Maximum return on investment through efficiency and utilization

谢谢
国际高性能计算咨询委员会

www.hpcadvisorycouncil.com

info@hpcadvisorycouncil.com

