HPC ADVISORY COUNCIL

OpenMX
Performance Benchmark and Profiling

May 2011

Mellanox TECHNOLOGIES

OPEN

Note

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

- **For more info please refer to**
  - http://www.hp.com/go/hpc
  - www.intel.com
  - www.mellanox.com
  - http://www.openmx-square.org

OpenMX

- **OpenMX (Open source package for Material eXplorer)**

- **OpenMX is designed for nano-scale material simulations based on**
  - Density functional theories (DFT)
  - Norm-conserving pseudopotentials
  - Pseudo-atomic localized basis functions

- **OpenMX is used in a wide variety of systems**
  - Bio-materials, carbon nanotubes, magnetic materials, and nanoscale conductors

- **OpenMX is a freely available (GPL) program from Japan**

Objectives

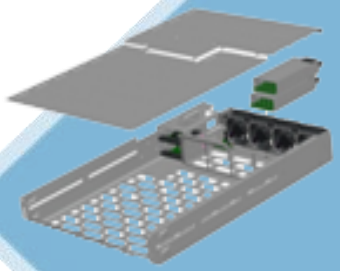
- **The presented research was done to provide best practices**
  - MPI libraries comparisons
  - Interconnect performance benchmarking
  - OpenMX Application profiling
  - Understanding OpenMX communication patterns
- **The presented results will demonstrate**
  - Balanced compute environment determines application performance

Test Cluster Configuration

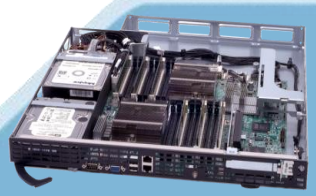
- **HP ProLiant SL2x170z G6 16-node cluster**
  - Six-Core Intel X5670 @ 2.93 GHz CPUs
  - Memory: 24GB per node
  - OS: CentOS5U5, OFED 1.5.3 InfiniBand SW stack
- **Mellanox ConnectX-2 InfiniBand QDR adapters and switches**
- **Fulcrum based 10Gb/s Ethernet switch**
- **MPI**
  - Intel MPI 4, Open MPI 1.5.3 with KNEM 0.9.6, Platform MPI 8.0.1, MVAPICH2-1.6rc1
- **Compilers: Intel Compilers 11.1.064**
- **Application: OpenMX 3.5**
- **Libraries: Intel MKL 2011.3.174**
- **Benchmark workload**
  - DIA512-1.dat

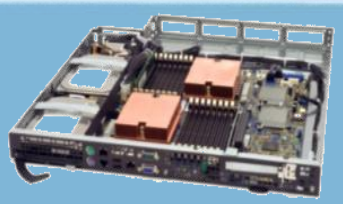
About HP ProLiant SL6000 Scalable System

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



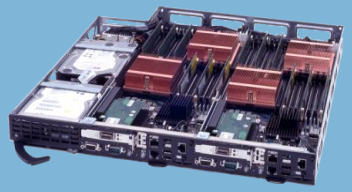
ProLiant z6000 chassis
Shared infrastructure
– fans, chassis, power





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

#1
Power
Efficiency*

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

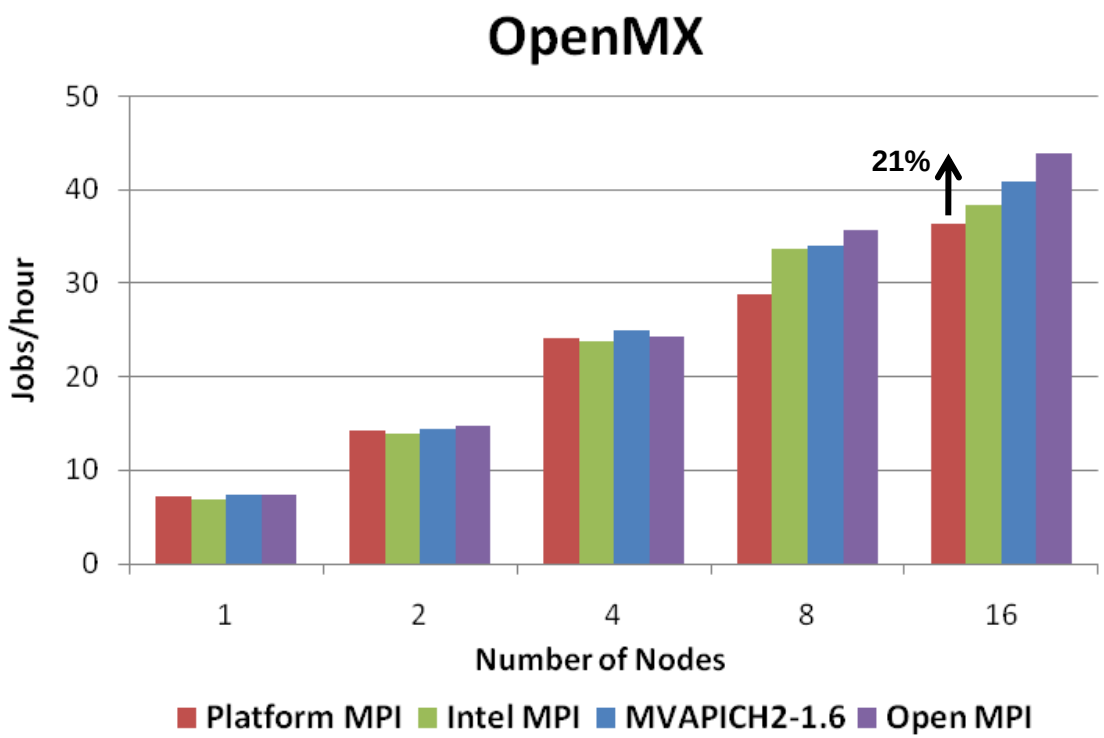
Mix and match configurations

Deploy with confidence

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

OpenMX Benchmark Results – MPI Libraries

- **Input Dataset**
  - DIA512-1.dat
- **OpenMPI delivers best OpenMX performance**
  - Up to 21% better than Platform MPI

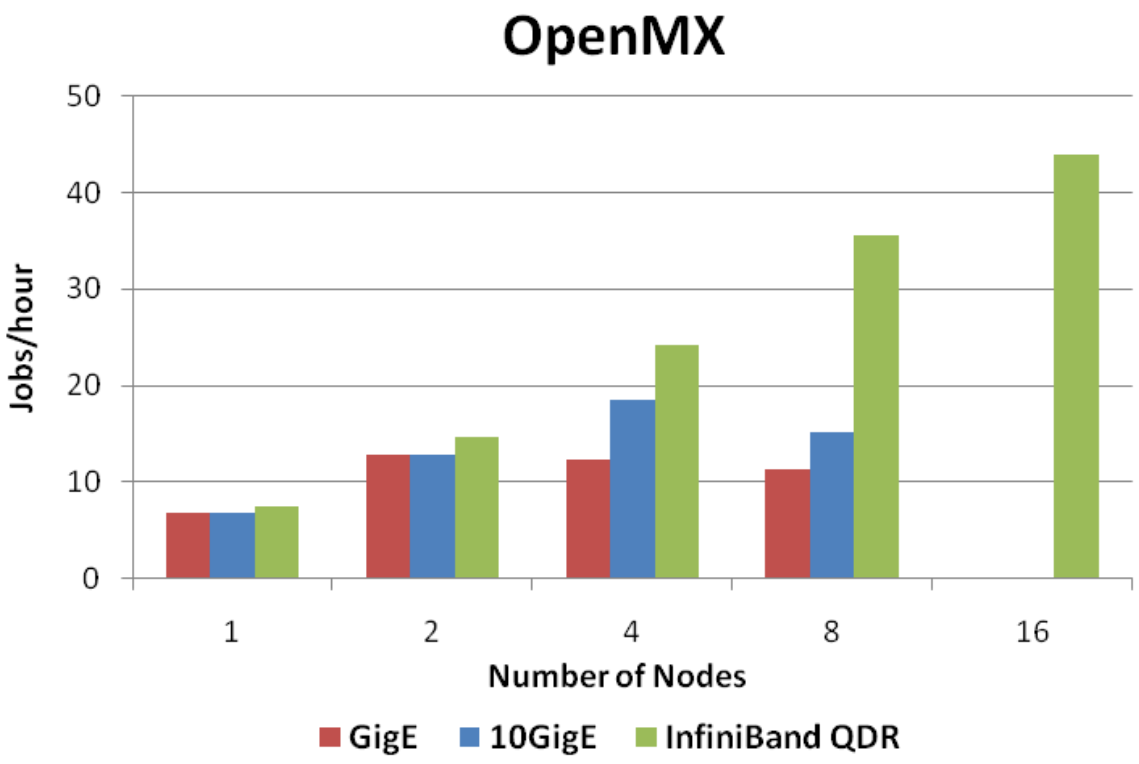


Higher is better

12-cores per node

OpenMX Benchmark Results – Interconnects

- **InfiniBand enables highest performance and scalability for OpenMX**
- **GigE stops scaling after 2 nodes, 10GigE after 4 nodes**



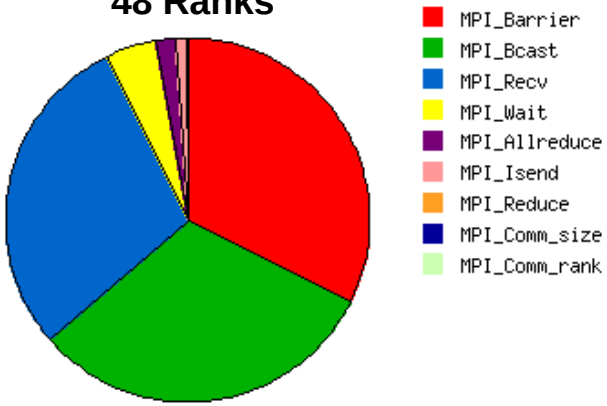
Higher is better

12-cores per node

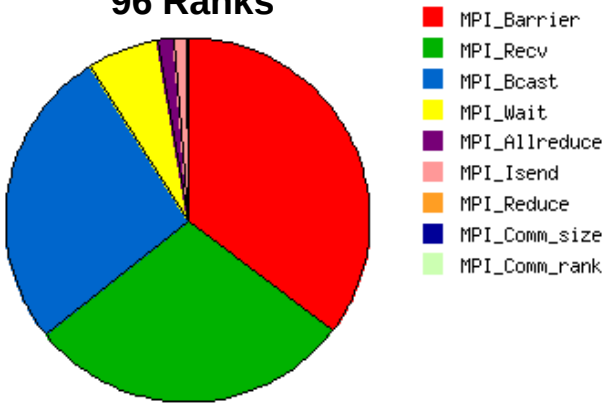
OpenMX MPI Profiling – MPI Functions

- **Both MPI collectives and point-to-point creates big communication time**
  - Collectives: MPI_Barrier and MPI_Bcast
  - Point-to-point: MPI_Isend/Recv

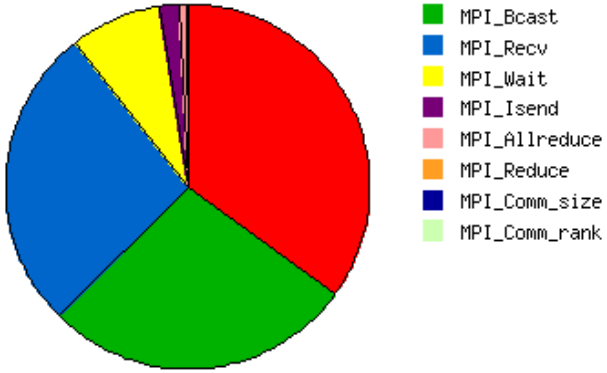
48 Ranks



96 Ranks

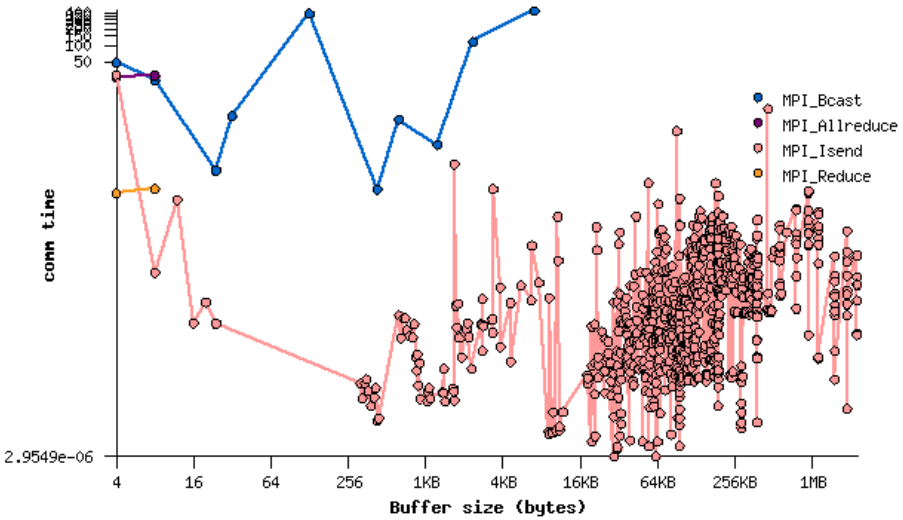


192 Ranks



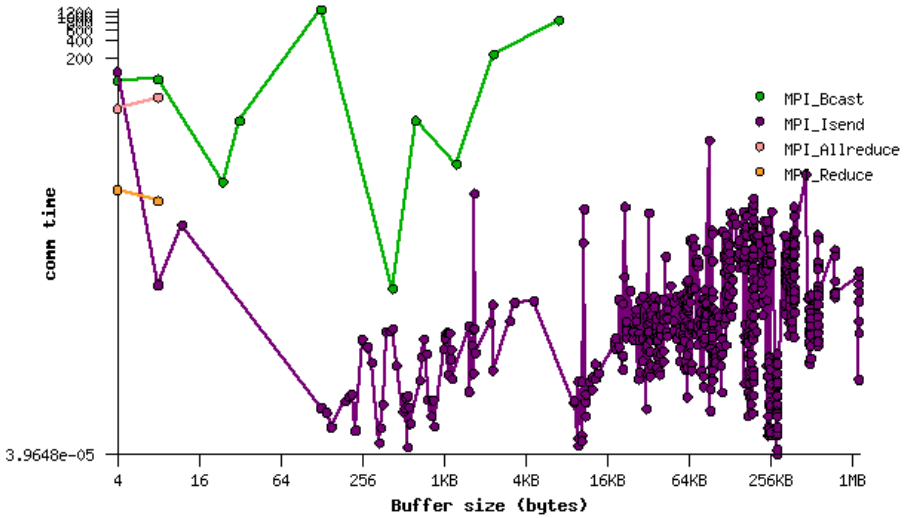
OpenMX MPI Profiling – Message Size

- **Both large and small messages are used**



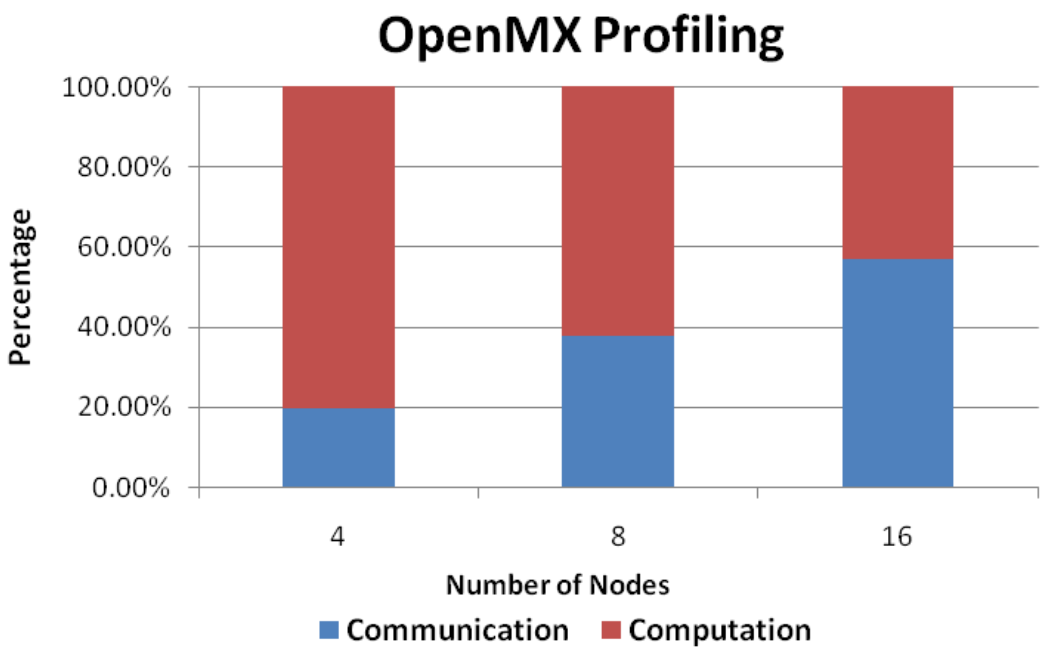
96 Ranks

192 Ranks



OpenMX MPI Profiling – Runtime Distribution

- **Percentage of communication keeps increasing as cluster size scales**



48 Ranks

OpenMX Benchmark Summary

- **OpenMX performance benchmark demonstrates**
  - InifiniBand QDR enables higher application performance and scalability
  - Neither GigE nor 10GigE can scale beyond 4 nodes

- **OpenMX MPI profiling**
  - MPI_Bcast, MPI_Barrier, and MPI_Recv create big communication overhead
  - Both large and small message are used by OpenMX
  - Interconnect latency and bandwidth are critical to OpenMX performance

Thank You

HPC Advisory Council