Octopus
Performance Benchmark and Profiling
May 2011
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

- [www.intel.com](http://www.intel.com)
- [www.mellanox.com](http://www.mellanox.com)
- [http://www.tddft.org/programs/octopus](http://www.tddft.org/programs/octopus)
Octopus

- **Octopus is designed for**
  - Density-functional theory (DFT)
  - Time-dependent density functional theory (TDDFT)

- **Octopus is aimed at the simulation of the electron-ion dynamics of 1, 2, 3, and 4 dimensional finite systems**

- **Octopus is one of selected 22 applications for the PRACE application benchmark suite**

- **Octopus is a freely available (GPL) software**
Objectives

- The presented research was done to provide best practices
  - MPI libraries comparisons
  - Interconnect performance benchmarking
  - Octopus Application profiling
  - Understanding Octopus 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, Platform MPI 8.0.1, MVAPICH2-1.6rc1
- **Compilers: Intel Compilers 11.1.064**
- **Application: Octopus 3.2.0**
- **Libraries: Intel MKL 2011.3.174**
- **Benchmark workload**
  - Benezne molecule
About HP ProLiant SL6000 Scalable System

• Solution-optimized for extreme scale out

ProLiant z6000 chassis
- Shared infrastructure
  - fans, chassis, power

ProLiant SL160z G6
- Large memory
  - memory-cache apps

ProLiant SL165z G7

ProLiant SL170z G6
- Large storage
  - Web search and database apps

ProLiant SL2x170z G6
- Highly dense
  - HPC compute and web front-end apps

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

#1 Power Efficiency*

* SPECpower_ssj2008
www.spec.org
17 June 2010, 13:28
Octopus Benchmark Results – MPI Libraries

- Input Dataset
  - Benezne molecule
- Intel MPI with default setting is 41% slower than other MPI

Octopus Benchmark

Higher is better

12-cores per node
Octopus Benchmark Results – MPI Optimization

- Intel MPI with tuning runs 46% faster than default mode at 16 nodes
  - `genv I_MPI_RDMA_TRANSLATION_CACHE 1`
  - `genv I_MPI_RDMA_RNDV_BUF_ALIGN 65536`
  - `genv I_MPI_SPIN_COUNT 121`
  - `genv I_MPI_DAPL_DIRECT_COPY_THRESHOLD 65536`
  - `genv I_MPI_ADJUST_ALLREDUCE '2:4-4:5:4-8'`

Higher is better

Octopus Benchmark

12-cores per node
Octopus Benchmark Results – Interconnects

• InfiniBand enables highest performance and scalability for Octopus
  – 151% faster than 10GigE and 700% faster than GigE at 16 nodes

Octopus Benchmark

Higher is better

12-cores per node
MPI collective communication overhead is dominated

- Collectives: MPI_Allreduce, MPI_BARRIER, MPI_Bcast, MPI_Alltoall, and MPI_Allgatherv
- Point-to-point: MPI_Isend/Irecv
Octopus MPI Profiling – Message Size

- Both large and small messages are used
  - Small messages: <64B
  - Medium to large: 1KB-64KB
Octopus Benchmark Summary

- **Octopus performance benchmark demonstrates**
  - InfiniBand QDR enables higher application performance and scalability
    - 151% higher performance than 10GigE and 700% higher than GigE
  - MPI tuning can provide significant performance boost
    - 46% with Intel MPI tuning

- **Octopus MPI profiling**
  - MPI collectives create big communication overhead
  - Both large and small message are used by Octopus
  - Interconnect latency and bandwidth are critical to Octopus performance
Thank You
HPC Advisory Council