HYCOM Performance Benchmark and Profiling

Jan 2010

Acknowledgment:
- The DoD High Performance Computing Modernization Program
• The following research was performed under the HPC Advisory Council activities
  – Participating vendors: HP, Mellanox
  – Compute resource - HPC Advisory Council Cluster Center

• We would like to acknowledge
  – The DoD High Performance Computing Modernization Program for providing access to the FY 2009 benchmark suite

• For more info please refer to
HYCOM (HYbridCoordinateOceanModel)

- A primitive equation ocean general circulation model
  - Evolved from the Miami Isopycnic-Coordinate Ocean Model

- HYCOM provides the capability of selecting several different vertical mixing schemes for
  - The surface mixed layer
  - The comparatively weak interior diapycnal mixing

- HYCOM is fully parallelized

- Open source and joined developed by:
  - University of Miami, the Los Alamos National Laboratory, and the Naval Research Laboratory physics
Objectives

- **The presented research was done to provide best practices**
  - HYCOM performance benchmarking
    - Interconnect performance comparisons
    - File system performance comparisons
    - MPI libraries performance comparisons
  - Understanding HYCOM communication patterns
- **The presented results will demonstrate**
  - The scalability of the compute environment
  - Considerations for power saving through balanced system configuration
Test Clusters Configuration

- **HP ProLiant SL170z G6 16-node cluster**
  - Six-Core Intel X5670 @ 2.93 GHz CPUs
  - Memory: 24GB per node
  - OS: CentOS5U4, OFED 1.5.2 InfiniBand SW stack

- **Intel Cluster Ready certified cluster**

- **Mellanox ConnectX2 InfiniBand adapters and switches**

- **MPI**: OpenMPI-1.4.2, MVAPICH2-1.5.1, Intel MPI 4.0, Platform MPI 8.0

- **Application**: HYCOM 2.2.10

- **Benchmark Workload**
  - HYCOM standard benchmark dataset
    - 26-layer 1/4 degree fully global HYCOM benchmark
HP ProLiant SL6000 Scalable System

- Solution-optimized for extreme scale out

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

- Mix and match configurations

- Deploy with confidence

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

ProLiant z6000 chassis
Shared infrastructure
- fans, chassis, power

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

#1 Power Efficiency*
• Lustre over InfiniBand enables better application performance and scalability
  – Up to 10% faster than NFS
  – Advantage increases as cluster scales
**HYCOM Benchmark Results – Interconnects**

- **InfiniBand enables better application performance and scalability**
  - Up to 226% higher performance than GigE
  - 15% higher performance than 10GigE at 96 cores
    - 5% higher than 10GigE at 47 cores
    - Performance gap increases as core count grows
  - Application performance over InfiniBand scales as cluster size increases

![HYCOM Performance Graph](image_url)
HYCOM Benchmark Results – MPI Libraries

- All MPIs show similar performance and scalability over InfiniBand
  - MVAPICH2 is slightly better than others
HYCOM Profiling – MPI Overhead

- **MPI_Allreduce, MPI_Waitall are major functions**
  - MPI_Waitall generates largest overhead
  - Allreduce overhead grows faster than other functions as cluster size increases

---

![Pie charts showing MPI overhead for 47, 96, and 124 processes.](chart.png)
HYCOM Profiling – MPI Message Size

- Majority MPI_Waitall messages are large size
  - >64KB
- MPI_Allreduce messages are small size

47 Processes

124 Processes
HYCOM Profiling Summary

• HYCOM was profiled to identify its communication patterns

• MPI_Waitall and MPI_Allreduce generate most overhead
  – Majority MPI_waitall messages are large size
  – MPI_allreduce messages are small size

• Interconnect bandwidth is important for HYCOM performance
  – As cluster scales, percentage of small messages increases
    • Hence interconnect latency becomes crucial too
Thank You
HPC Advisory Council