HYCOM Performance Benchmark and Profiling

Jan 2011

Acknowledgment:
- The DoD High Performance Computing Modernization Program

Note

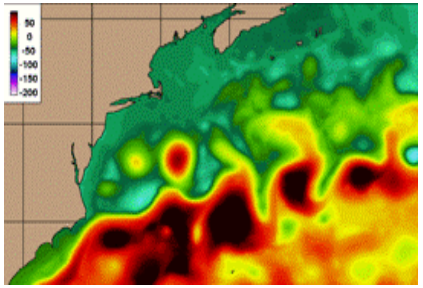
- **The following research was performed under the HPC Advisory Council activities**
  - Participating vendors: AMD, Dell, Mellanox
  - Compute resource - HPC Advisory Council Cluster Center

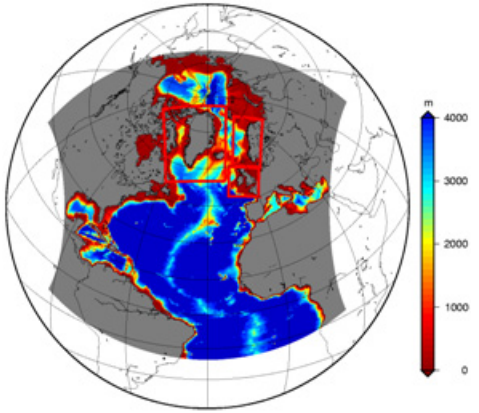
- **We would like to acknowledge**
  - The DoD High Performance Computing Modernization Program for providing access to the FY 2009 benchmark suite

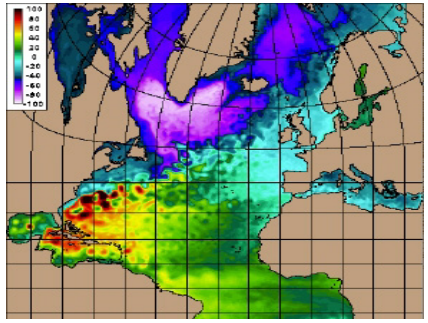
- **For more info please refer to**
  - www.mellanox.com, www.dell.com/hpc, www.amd.com, http://www.hycom.org

HYCOM (HYbrid Coordinate Ocean Model)

- **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
    - MPI libraries performance comparisons
  - Understanding HYCOM communication patterns

- **The presented results will demonstrate**
  - Balanced system to provide good application scalability

Test Cluster Configuration

- Dell™ PowerEdge™ R815 11-node cluster
- AMD™ Opteron™ 6174 (code name “Magny-Cours”) 12-cores @ 2.2 GHz CPUs
- 4 CPUs per server node
- Mellanox ConnectX-2 40Gb/s QDR InfiniBand adapter
- Mellanox M3600 36-Port 40Gb/s QDR InfiniBand switch
- Memory: 128GB memory per node DDR3 1333
- OS: CentOS 5.5, MELNX-OFED 1.5.1 InfiniBand SW stack
- MPI: OpenMPI-1.5.1, Platform MPI 8.0, MVAPICH2-1.5.1
- Application: HYCOM 2.2.10
- Benchmark Workload
  - HYCOM large benchmark dataset
    - 26-layer 1/12 degree fully global HYCOM benchmark

Dell™ PowerEdge™ R815 6-node cluster

- **HPC Advisory Council Test-bed System**

- **New 11-node 528 core cluster - featuring Dell PowerEdge™ R815 servers**
  - Replacement system for Dell PowerEdge SC1435 (192 cores) cluster system following 2 years of rigorous benchmarking and product EOL
    - System to be redirected to explore HPC in the Cloud applications

- **Workload profiling and benchmarking**
  - Characterization for HPC and compute intense environments
  - Optimization for scale, sizing and configuration and workload performance
  - Test-bed Benchmarks
    - RFPs
    - Customers/Prospects, etc
  - ISV & Industry standard application characterization
  - Best practices & usage analysis

About Dell PowerEdge™ Platform Advantages

Best of breed technologies and partners

Combination of AMD™ Opteron™ 6100 series platform and Mellanox ConnectX InfiniBand on Dell HPC

Solutions provide the ultimate platform for speed and scale

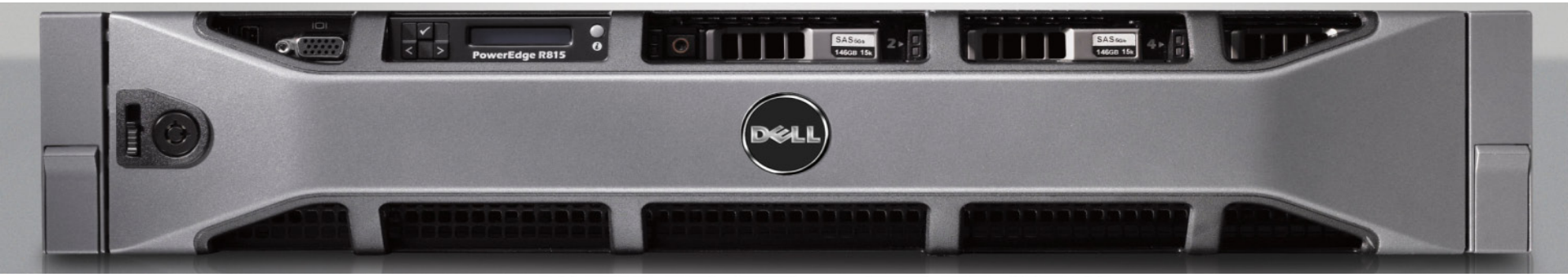
- Dell PowerEdge R815 system delivers 4 socket performance in dense 2U form factor
- Up to 48 core/32DIMMs per server – 1008 core in 42U enclosure

Integrated stacks designed to deliver the best price/performance/watt

- 2x more memory and processing power in half of the space
- Energy optimized low flow fans, improved power supplies and dual SD modules

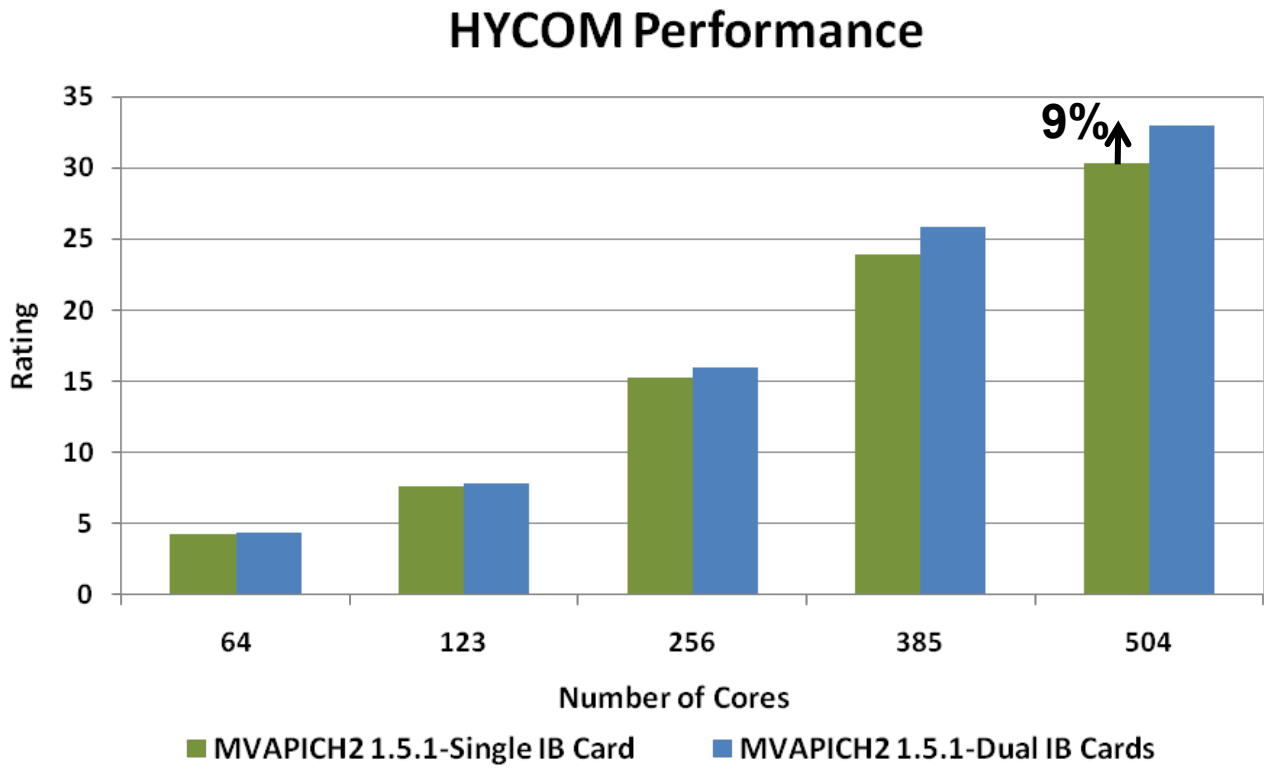
Optimized for long-term capital and operating investment protection

- System expansion
- Component upgrades and feature releases



HYCOM Benchmark Results – Multi Rail

- **Dual InfiniBand Cards enable up to 9% better performance**
- **Higher advantage expected as cluster size scales**

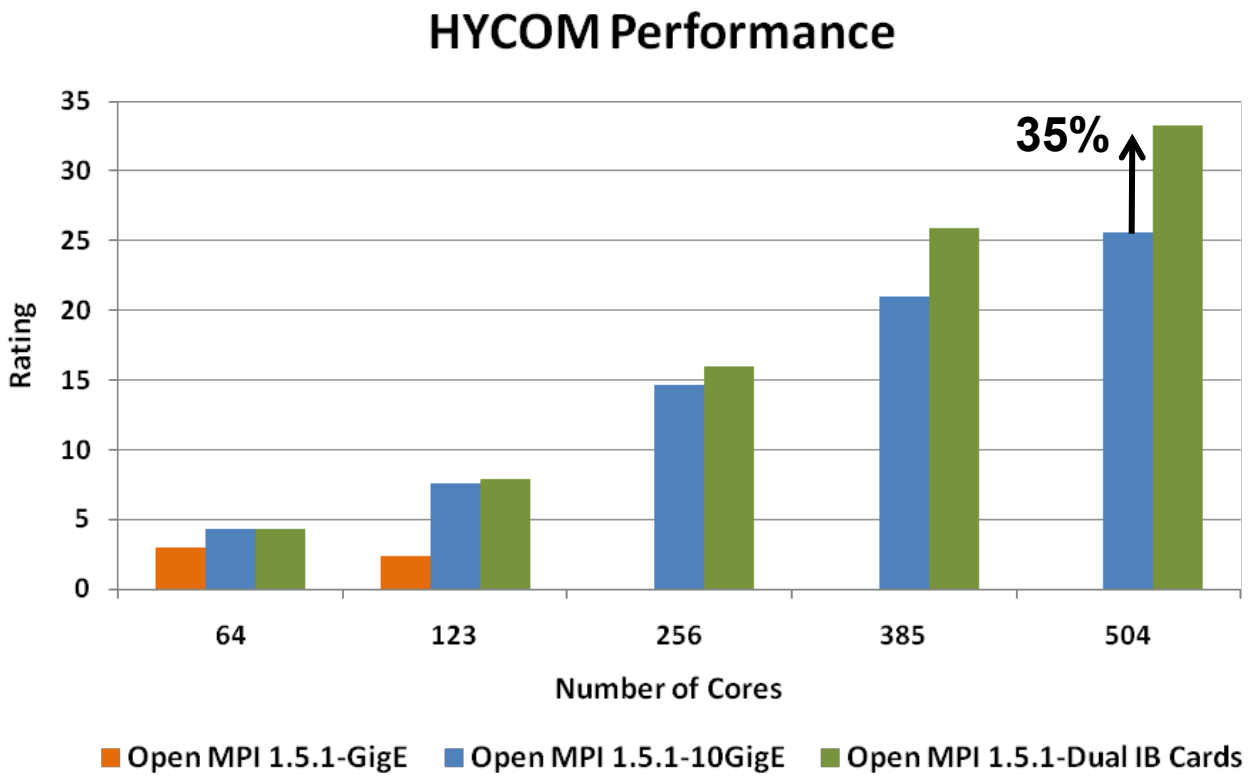


Higher is better

48-cores per node

HYCOM Benchmark Results - Interconnect

- **InfiniBand enables better application performance and scalability**
  - GigE stops scaling after 2 nodes (Extremely slow with 3+ nodes)
  - Up to 35% higher than 10GigE at 504 cores
- **Application performance over InfiniBand scales as cluster size increases**

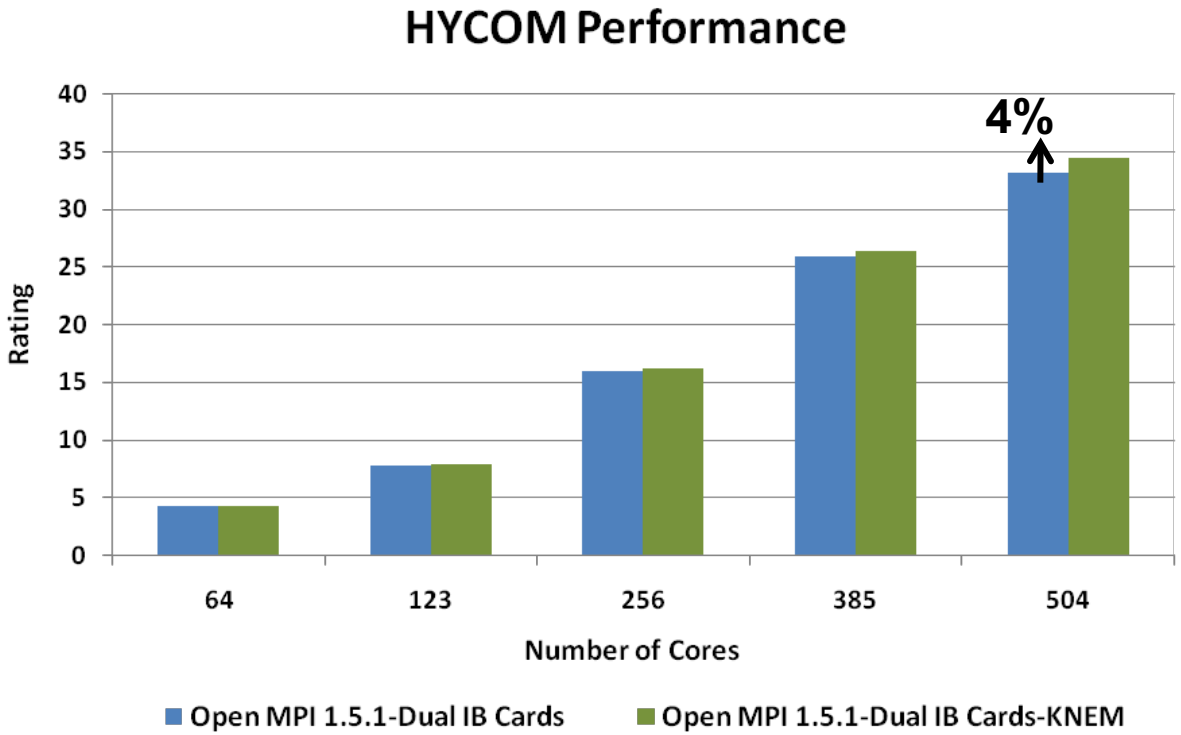


Higher is better

48-cores per node

HYCOM Benchmark Results - KNEM

- **KNEM is a Linux kernel module enabling high-performance intra-node MPI communication for large messages**
- **KNEM improves application performance**
  - By up to 4% at 504 cores
- **Higher performance will gain as cluster size increases**

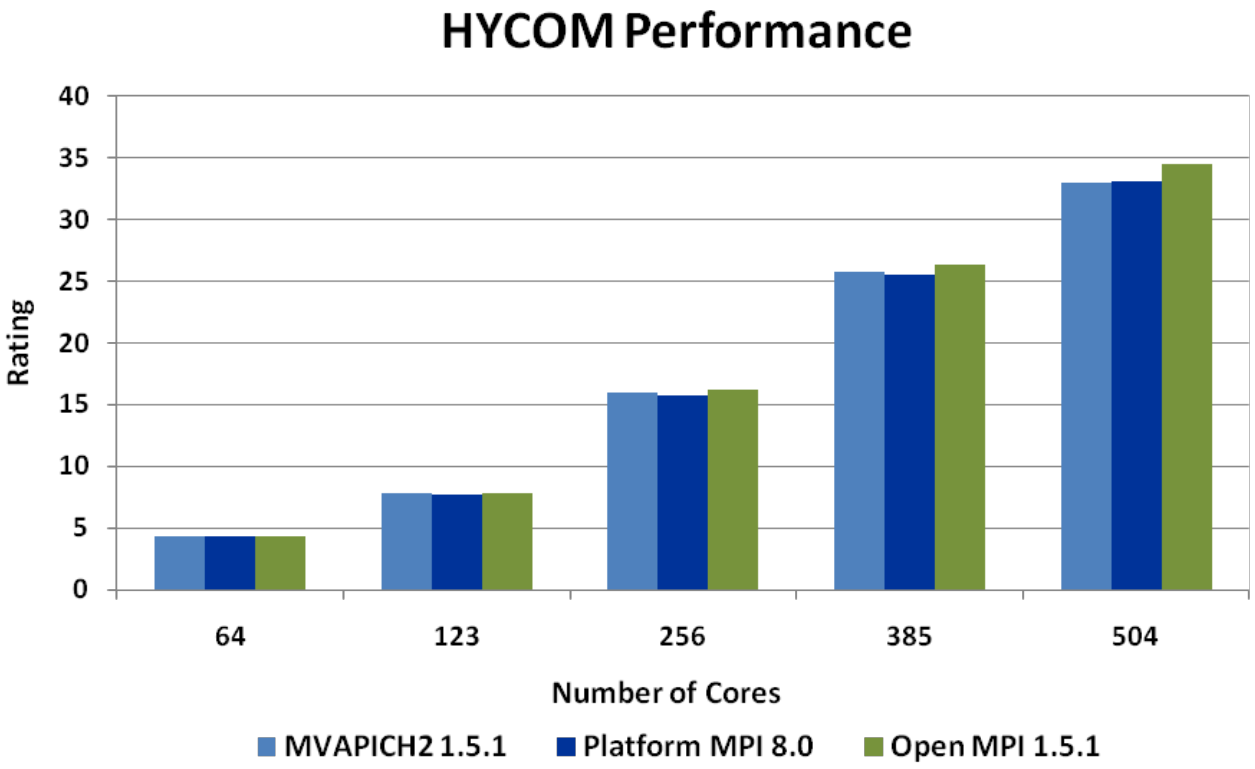


Higher is better

48-cores per node

HYCOM Benchmark Results – MPI Libraries

- **Open MPI with KNEM enables better performance at 504 cores**



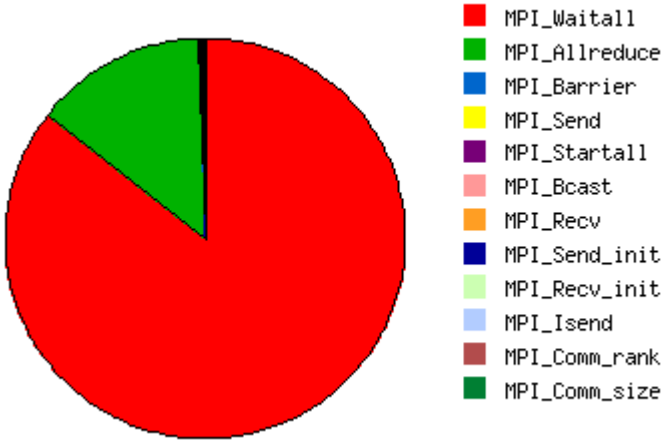
Dual IB cards

48-cores per node

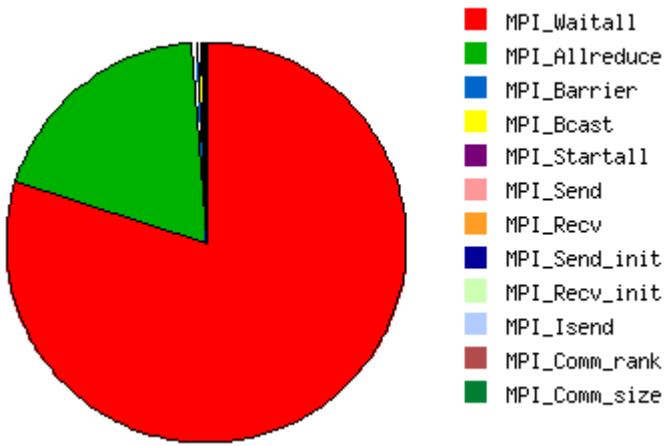
Higher is better

HYCOM MPI Profiling – MPI Functions

- MPI_Waitall and MPI_Allreduce generate more communication overhead
- Time share of MPI_Allreduce increases as cluster size scales

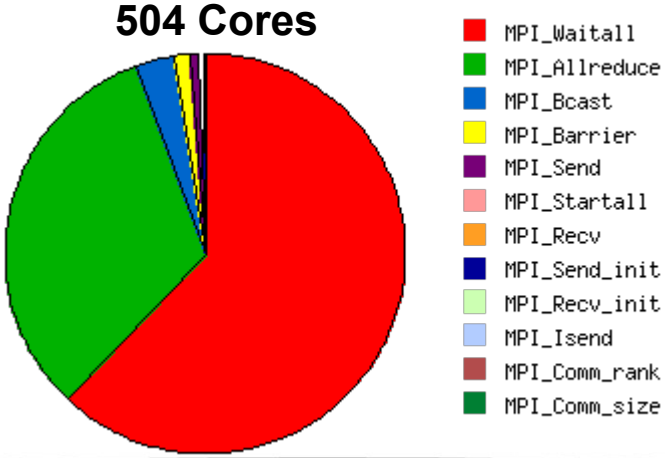


123 Cores



256 Cores

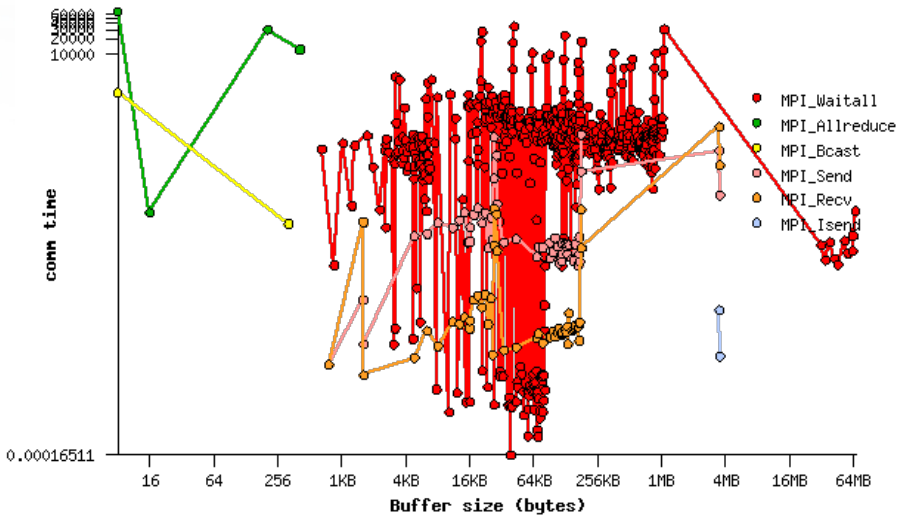
504 Cores



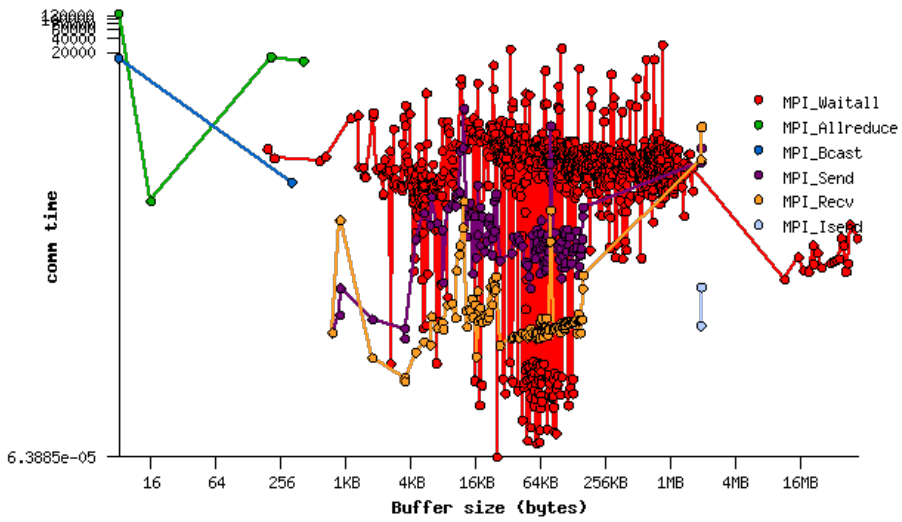
HYCOM MPI Profiling – Message Size

- **Both large and small messages are used by HYCOM**
  - MPI_Waitall uses middle to large size messages
  - MPI_Allreduce uses small size messages

256 Cores



504 Cores



HYCOM Benchmark Results Summary

- **HYCOM Communication Pattern**
  - MPI collective function generates most overhead
    - MPI_Allreduce and MPI_Waitall
  - Both small and large messages are used by application
  - At small scale, bandwidth is critical to HYCOM performance, as cluster size grows, latency becomes more important

- **Performance Optimizations**
  - MPI libraries showed comparable performance overall
  - KNEM enables better performance
  - Dual InfiniBand cards improves higher performance

- **Interconnect Characterization**
  - InfiniBand delivers superior scalability as cluster size increases
  - InfiniBand enables much higher performance than 10GigE and GigE

Thank You

HPC Advisory Council