HPC ADVISORY COUNCIL

MILC Performance Benchmark and Profiling

April 2013

hp

Mellanox TECHNOLOGIES

Note

- **The following research was performed under the HPC Advisory Council activities**
  - Special thanks for: HP, Mellanox

- **For more information on the supporting vendors solutions please refer to:**
  - www.mellanox.com, http://www.hp.com/go/hpc

- **For more information on the application:**
  - http://www.physics.utah.edu/~detar/milc/

MILC

- **MILC (MIMD Lattice Computation) QCD Code**
  - Developed by the MIMD Lattice Computation (MILC) collaboration
  - Performs large scale numerical simulations to study quantum chromodynamics (QCD)
  - QCD is the theory of the strong interactions of subatomic physics
  - Simulates 4-dimensional SU(3) lattice gauge theory on MIMD parallel machines
  - Contains a set of codes written in C.
  - Publicly available for research purposes
  - Free, open-source code, distributed under GNU General Public License

- **The MILC Collaboration**
  - Produced application codes to study several different QCD research areas
  - Is engaged in a broad research program in Quantum Chromodynamics (QCD)
  - Its research addresses fundamental questions in high energy and nuclear physics
  - Related to major experimental programs in the fields, including:
  - Studies of the mass spectrum of strongly interacting particles,
  - The weak interactions of these particles,
  - The behavior of strongly interacting matter under extreme conditions

Objectives

- **The presented research was done to provide best practices**
  - MILC performance benchmarking
  - Interconnect performance comparisons
  - MPI performance comparison
  - Understanding MILC communication patterns

- **The presented results will demonstrate**
  - The scalability of the compute environment to provide nearly linear application scalability

Test Cluster Configuration

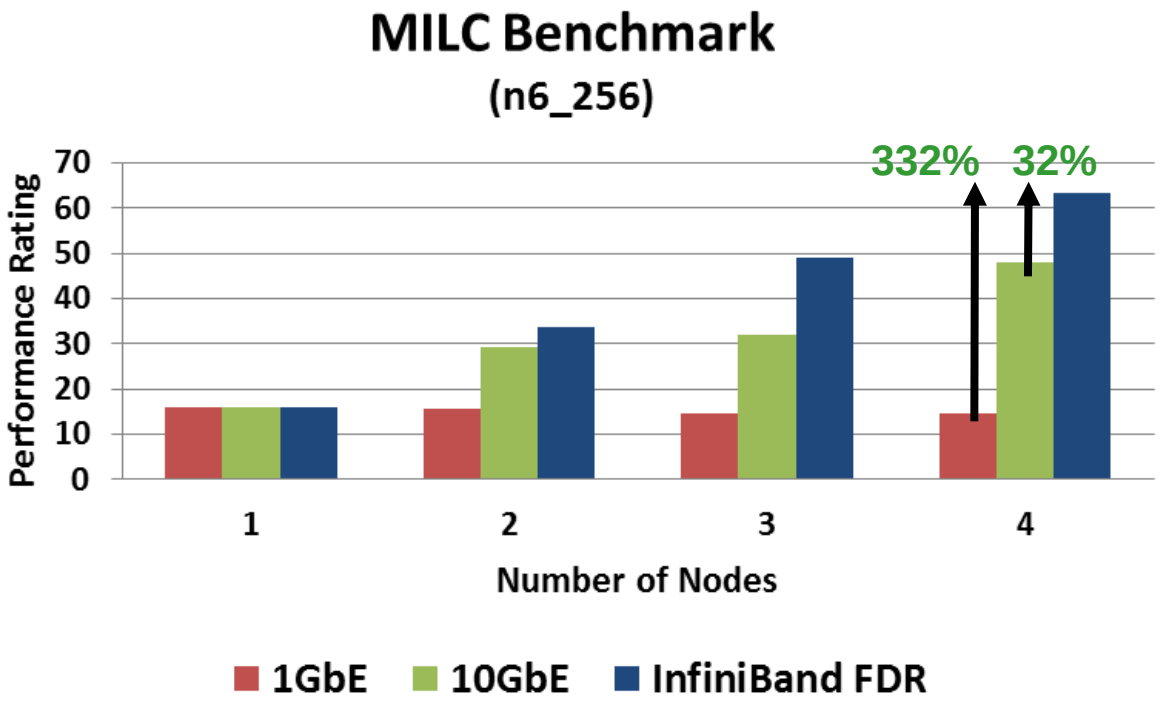
- **HP ProLiant SL230s Gen8 4-node "Athena" cluster**
  - Processors: Dual Eight-Core Intel Xeon E5-2680 @ 2.7 GHz
  - Memory: 32GB per node, 1600MHz DDR3 DIMMs
  - OS: RHEL 6 Update 2, OFED 1.5.3 InfiniBand SW stack
- **Mellanox ConnectX-3 VPI InfiniBand adapters**
- **Mellanox SwitchX SX6036 56Gb/s FDR InfiniBand and 40G/s Ethernet VPI Switch**
- **MPI: Open MPI 1.6.4, Platform MPI 8.2.1**
- **Compiler: Intel Compilers Version 13 (Intel Composer XE 2013)**
- **Application: milc_qcd-7.7.8**
- **Benchmark Workload:**
- **Input dataset:**
  - n6_256.in – Input dataset from NERSC: http://www.nersc.gov/assets/RD/milc7v1.0.tar
  - Global lattice size of (32x32x32x36)

About HP ProLiant SL230s Gen8

Item	SL230 Gen8
Processor	Two Intel® Xeon® E5-2600 Series, 4/6/8 Cores,
Chipset	Intel® Sandy Bridge EP Socket-R
Memory	(512 GB), 16 sockets, DDR3 up to 1600MHz, ECC
Max Memory	512 GB
Internal Storage	Two LFF non-hot plug SAS, SATA bays or Four SFF non-hot plug SAS, SATA, SSD bays Two Hot Plug SFF Drives (Option)
Max Internal Storage	8TB
Networking	Dual port 1GbE NIC/ Single 10G Nic
I/O Slots	One PCIe Gen3 x16 LP slot 1Gb and 10Gb Ethernet, IB, and FlexF abric options
Ports	Front: (1) Management, (2) 1GbE, (1) Serial, (1) S.U.V port, (2) PCIe, and Internal Micro SD card & Active Health
Power Supplies	750, 1200W (92% or 94%), high power chassis
Integrated Management	iLO4 hardware-based power capping via SL Advanced Power Manager
Additional Features	Shared Power & Cooling and up to 8 nodes per 4U chassis, single GPU support, Fusion I/O support
Form Factor	16P/8GPUs/4U chassis

MILC Performance - Interconnect

- **InfiniBand FDR is the most efficient inter-node communication for MILC**
  - Outperforms 10GbE by 32% at 4 nodes
  - Outperforms 1GbE by 332% at 4 nodes
- **1GbE do not show performance gain beyond 1 node**

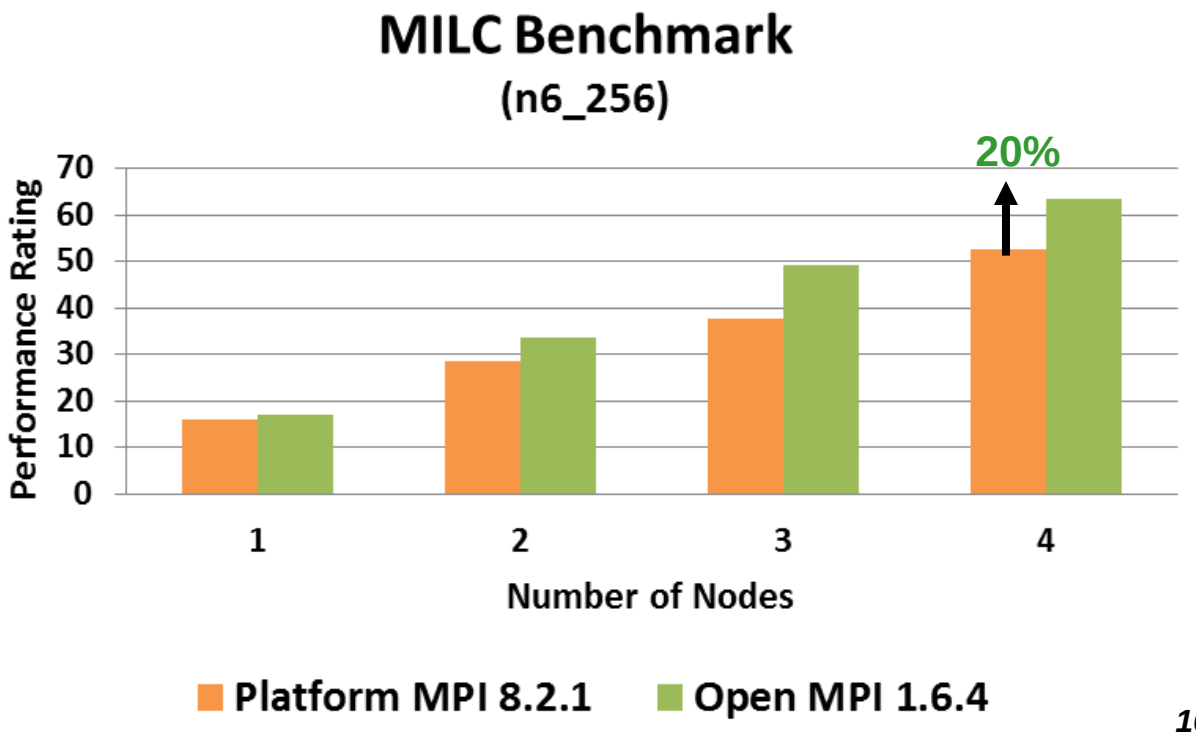


Higher is better

16 Processes/Node

MILC Performance - Interconnect

- **Open MPI performs better than Platform MPI**
  - Outperforms by around 20% at 64 processes
  - No tuning flags used other than processor binding used in both cases
  - Same compiler flags have been used for both cases

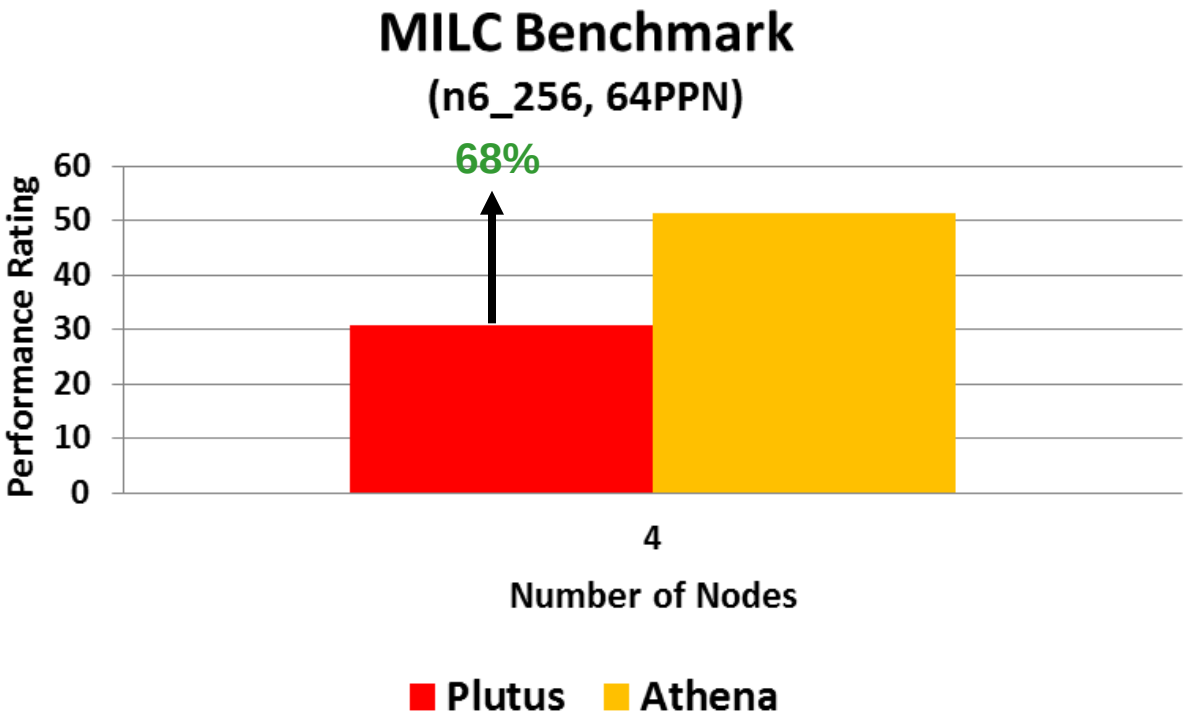


Higher is better

16 Processes/Node

MILC Performance - Processors

- **Intel E5-2680 processors (Sandy Bridge) cluster outperforms prior CPU generation**
  - Performs 68% higher than X5670 cluster at 4 nodes
- **Configurations used:**
  - Athena: 2-socket Intel E5-2680 @ 2.7GHz, 1600MHz DIMMs, FDR InfiniBand, 16PPN
  - Plutus: 2-socket Intel X5670 @ 2.93GHz, 1333MHz DIMMs, QDR InfiniBand, 12PPN
  - Compiler optimization flags: “OCFLAGS=-O3 -ip”. Athena has additional “-xAVX” flag

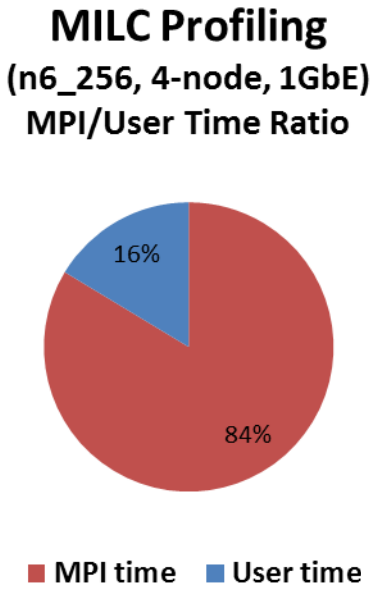


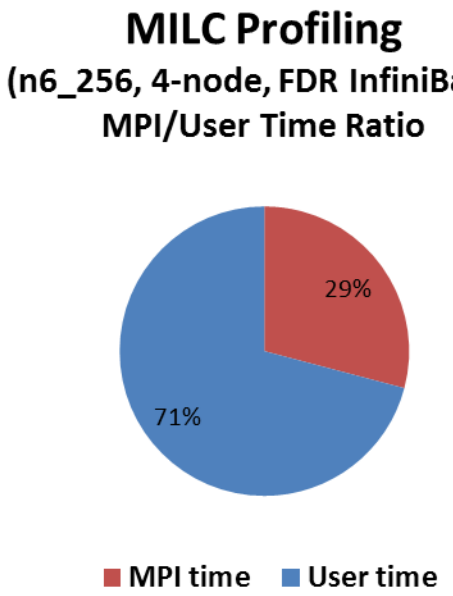
Higher is better

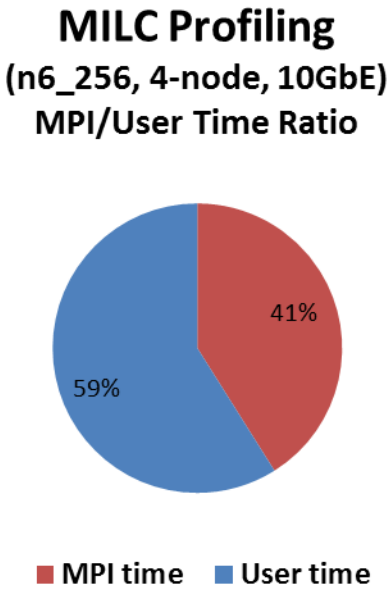
Platform MPI

MILC Profiling – MPI Time Ratio

- **InfiniBand FDR reduces the communication time at scale**
  - InfiniBand FDR consumes about 29% of total runtime
  - 1GbE consumes 84% of total time, while 10GbE consumes about 41%



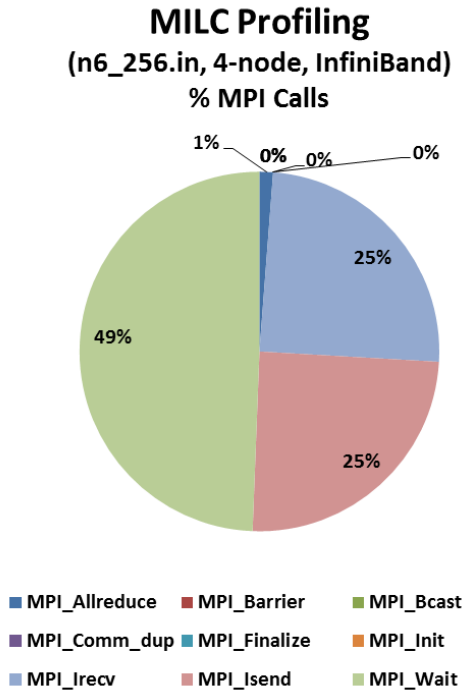


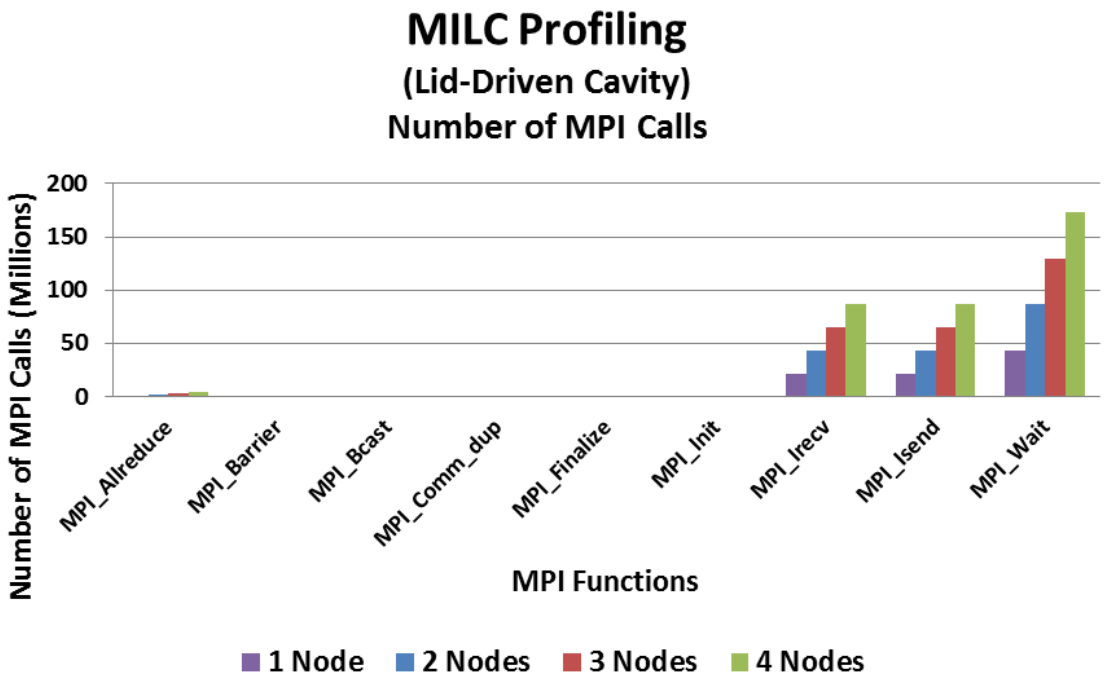


16 Processes/Node

MILC Profiling – MPI Functions

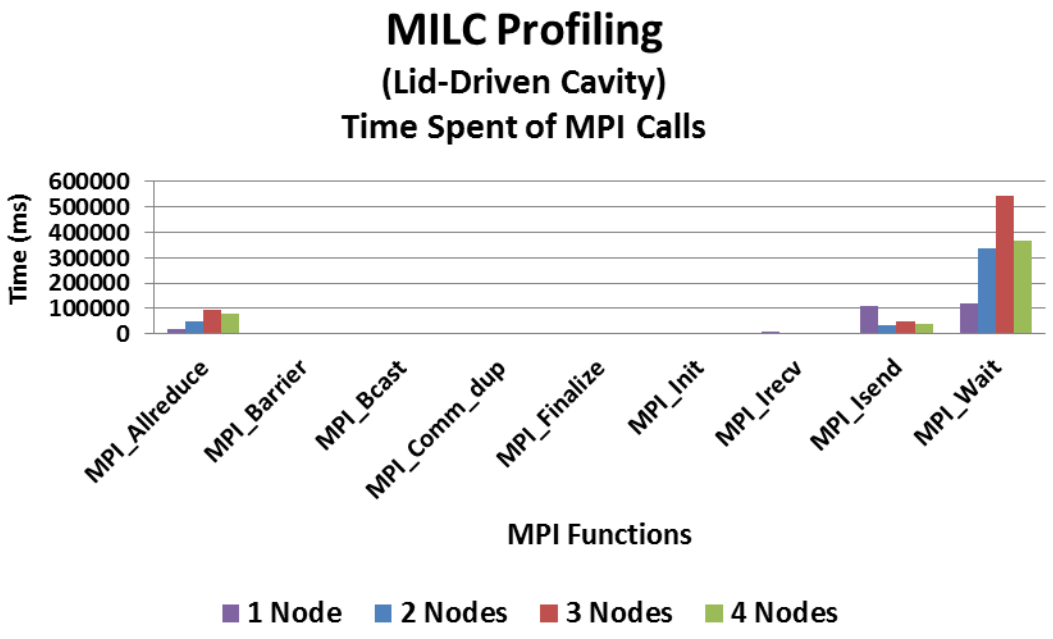
- **Mostl used MPI functions**
  - MPI_Wait (49%) and MPI_Isend (25%), MPI_Irecv (25%)

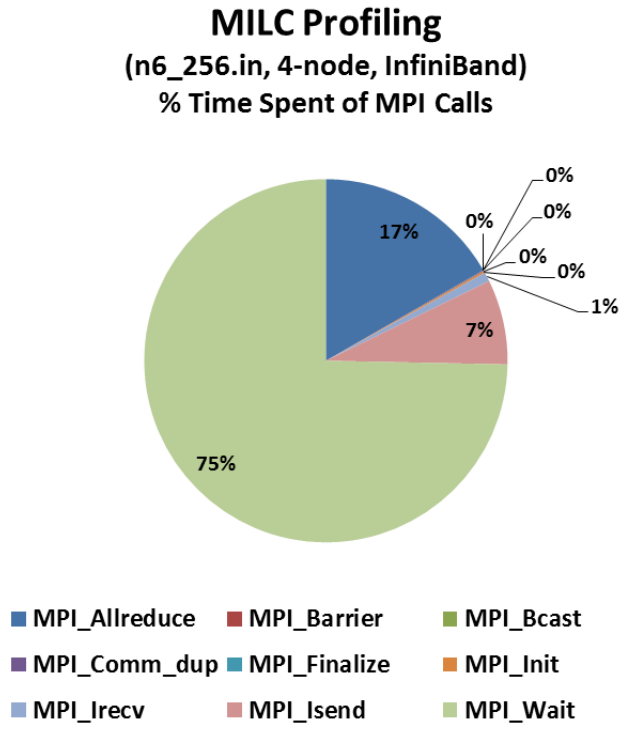




MILC Profiling – MPI Functions

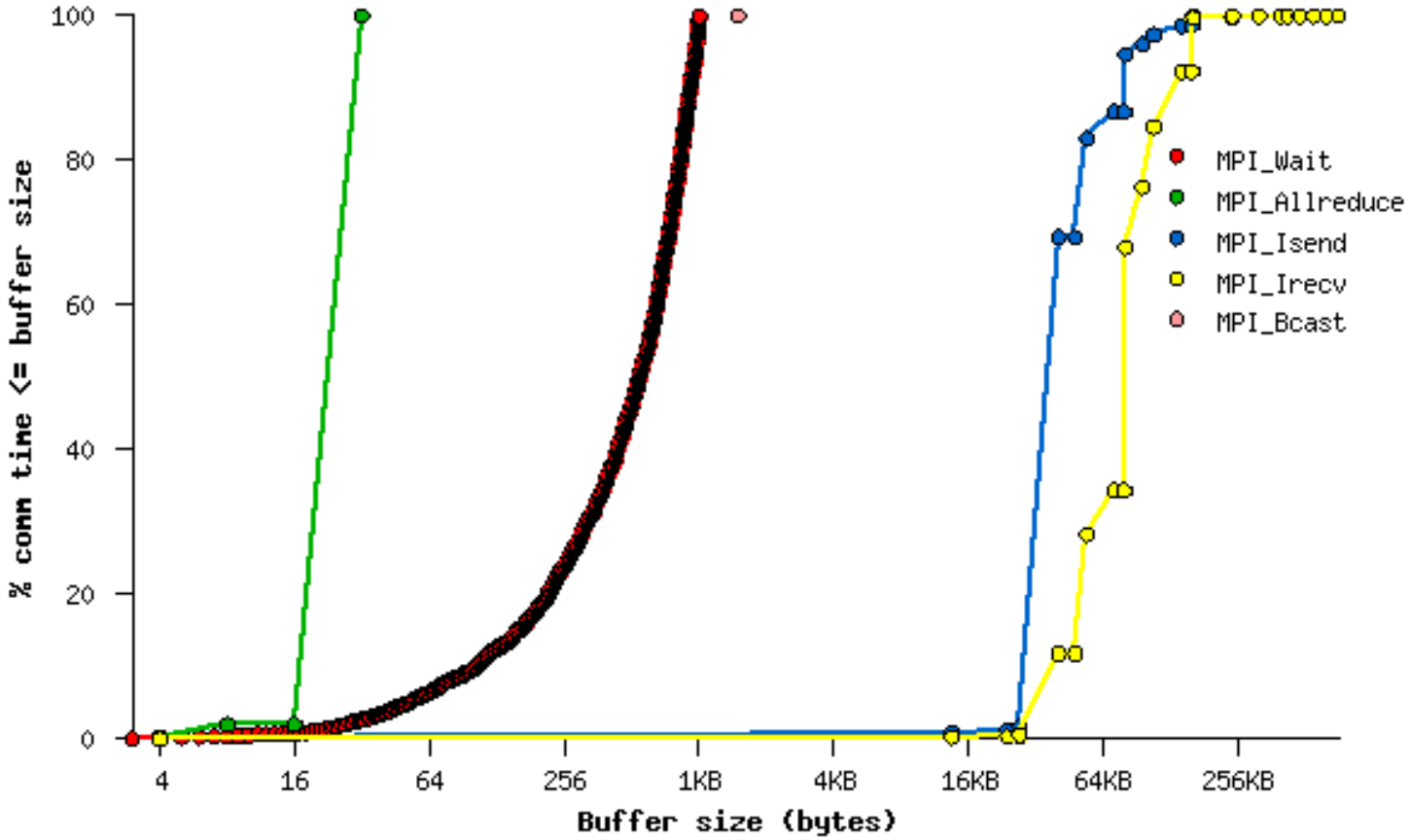
- **The most time consuming MPI functions:**
  - MPI_Wait (75%), MPI_Allreduce (17%)





MILC Profiling – Message Size

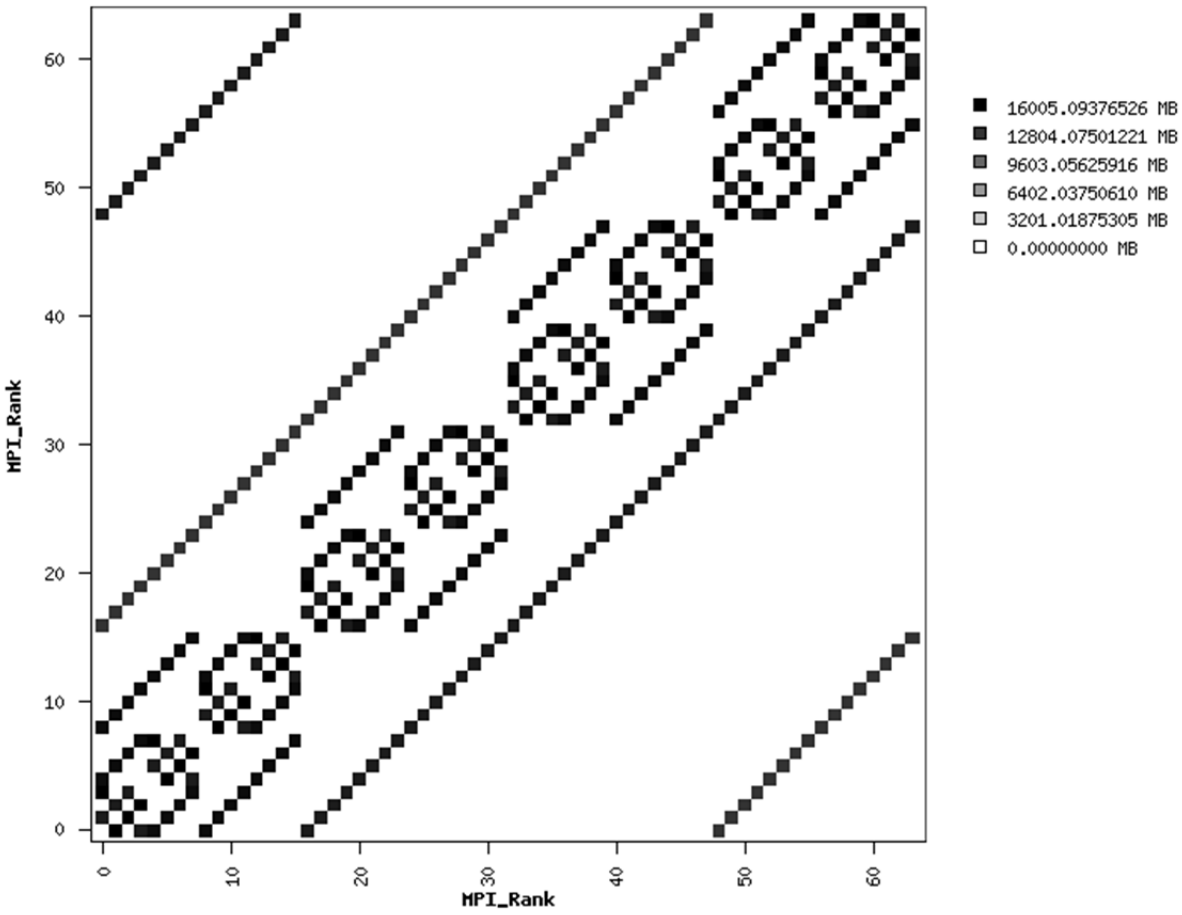
- **Distribution of message sizes for the MPI calls**
  - MPI_Wait between 256B to 1KB
  - MPI_Allreduce: small messages from 16B



64 MPI Processes

MILC Profiling – Point-to-point Flow

- **Heavy MPI communications seen between processes**
  - Mainly concentrated between close neighboring ranks
  - For point-to-point communications, such as the non-blocking communications



64 MPI Processes

MILC Summary

- **HP ProLiant Gen8 servers delivers better MILC Performance than its predecessor**
  - ProLiant Gen8 equipped with Intel E5 series processes and InfiniBand FDR
  - Provides 68% higher performance than the ProLiant G7 servers when compare at 4 nodes
- **InfiniBand FDR is the most efficient inter-node communication for MILC**
  - Outperforms 10GbE by 32% at 4 nodes
  - Outperforms 1GbE by 332% (or by over 3x) at 4 nodes
- **MILC Profiling**
  - Heavy MPI communications are seen between MPI processes
  - InfiniBand FDR reduces communication time; leave more time for computation
    - InfiniBand FDR consumes 29% of total time, versus 41% 10GbE, versus 84% 1GbE
  - Non-blocking communications are seen:
    - Time spent: MPI_Wait (75%), MPI_Allreduce (17%)
    - Most used: MPI_Wait (49%) and MPI_Isend (25%), MPI_Irecv (25%)

Thank You

HPC Advisory Council