HPC ADVISORY COUNCIL

NETWORK OF EXPERTISE

MiniFE
Performance Benchmark and Profiling

December 2013

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:**
  - https://asc.llnl.gov/CORAL-benchmarks/#minife

MiniFE

- **MiniFE**
  - Is a Finite Element mini-application
  - Implements kernels that represents implicit finite-element applications
  - Assembles a sparse linear-system from the steady-state conduction equation
    - on a brick-shaped problem domain of linear 8-node hex elements
  - Solves linear-system using un-preconditioned conjugate-gradient algorithm
- **MiniFE kernels responsible for:**
  - Computation of element-operators
    - Diffusion matrix, source vector
  - Assembly
    - Scattering element-operators into sparse matrix and vector
  - Sparse matrix-vector product (during CG solve)
  - Vector operations
    - level-1 blas: axpy, dot, norm
  - Compare computed solution vs analytic solution for steady-state temperature

Objectives

- **The presented research was done to provide best practices**
  - MiniFE performance benchmarking
  - Interconnect performance comparisons
  - MPI performance comparison
  - Understanding MiniFE communication patterns
- **The presented results will demonstrate**
  - The scalability of the compute environment to provide nearly linear application scalability

Test Cluster Configuration

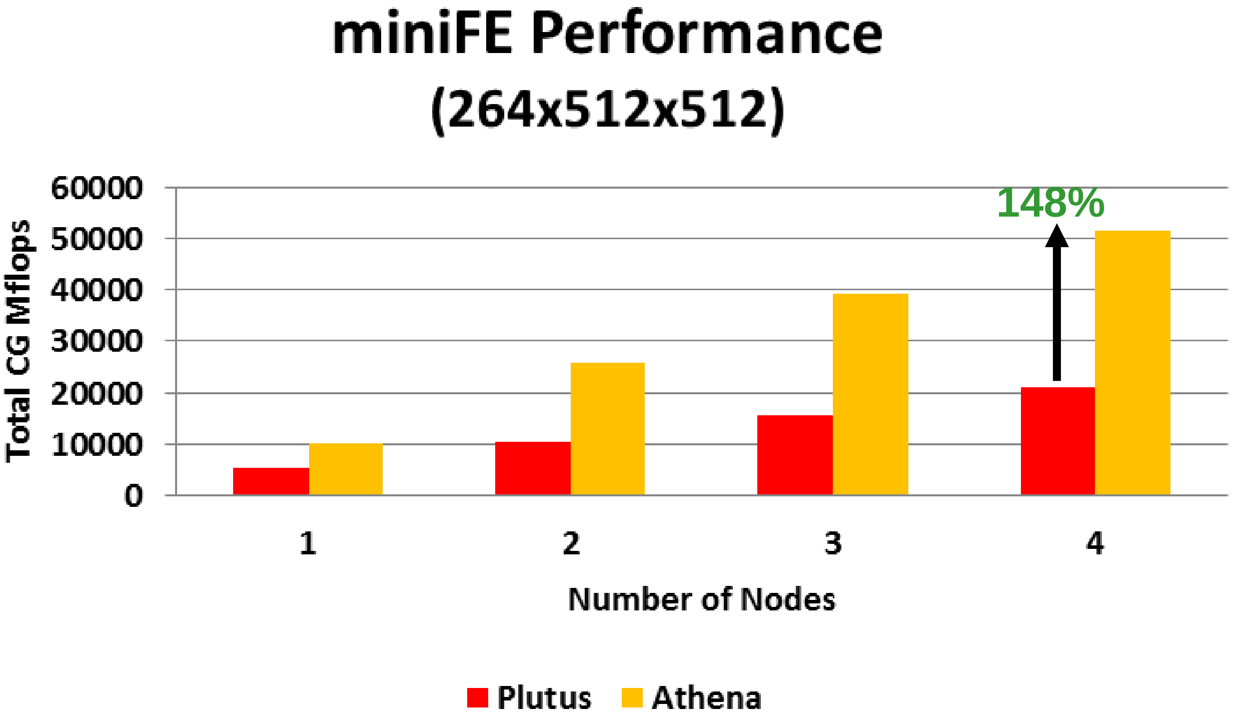
- **HP ProLiant SL230s Gen8 4-node “Athena” cluster**
  - Processors: Dual-Socket 10-core Intel Xeon E5-2680v2 @ 2.8 GHz CPUs
  - Memory: 32GB per node, 1600MHz DDR3 Dual-Ranked DIMMs
  - OS: RHEL 6 Update 2, OFED 2.0-3.0.0 InfiniBand SW stack
- **Mellanox Connect-IB FDR InfiniBand adapters**
- **Mellanox ConnectX-3 VPI adapters**
- **Mellanox SwitchX SX6036 56Gb/s FDR InfiniBand and Ethernet VPI Switch**
- **MPI: Platform MPI 8.3, Open MPI 1.6.5**
- **Compiler: GNU Compilers**
- **Application: miniFE 2.0 rc3**
- **Benchmark Workload:**
- **Input dataset:**
  - 264x512x512 problem size

About HP ProLiant SL230s Gen8

Item	HP ProLiant SL230s Gen8 Server
Processor	Two Intel® Xeon® E5-2600 v2 Series, 4/6/8/10/12 Cores,
Chipset	Intel® Xeon E5-2600 v2 product family
Memory	(256 GB), 16 DIMM slots, DDR3 up to 1600MHz, ECC
Max Memory	256 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

MiniFE Performance - Processors

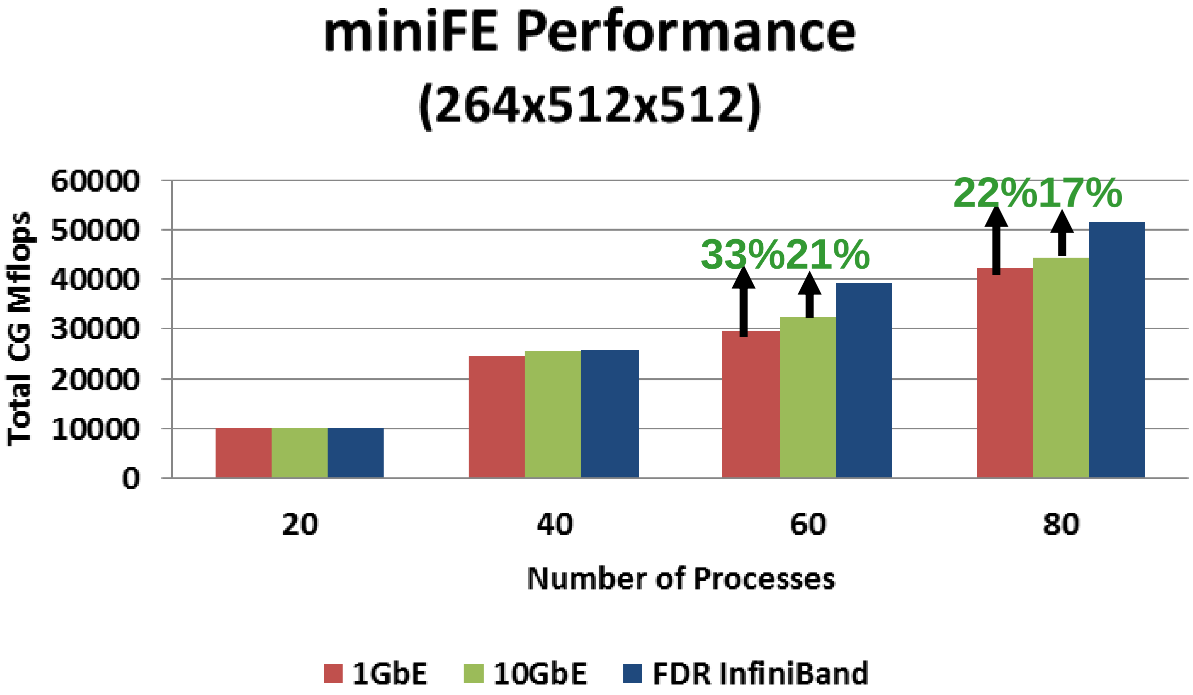
- **Intel E5-2680v2 processors (Ivy Bridge) cluster outperforms prior CPU generation**
  - Performs 148% higher than Xeon X5670 (Westmere) cluster at 4 nodes
- **Configurations compared:**
  - Athena: 2-socket Intel Xeon E5-2680v2 @ 2.8GHz, 1600MHz DIMMs, FDR IB, 20PPN
  - Plutus: 2-socket Intel Xeon X5670 @ 2.93GHz, 1333MHz DIMMs, QDR IB, 12PPN
  - Compiler optimization flags: “CFLAGS=-O3”



Higher is better

MiniFE Performance - Interconnect

- **FDR InfiniBand is the most efficient inter-node communication for MiniFE**
  - Outperforms 10GbE by 21% at 60 MPI processes
  - Outperforms 1GbE by 33% at 60 MPI processes
  - The performance benefit of InfiniBand expects to grow at larger CPU core counts

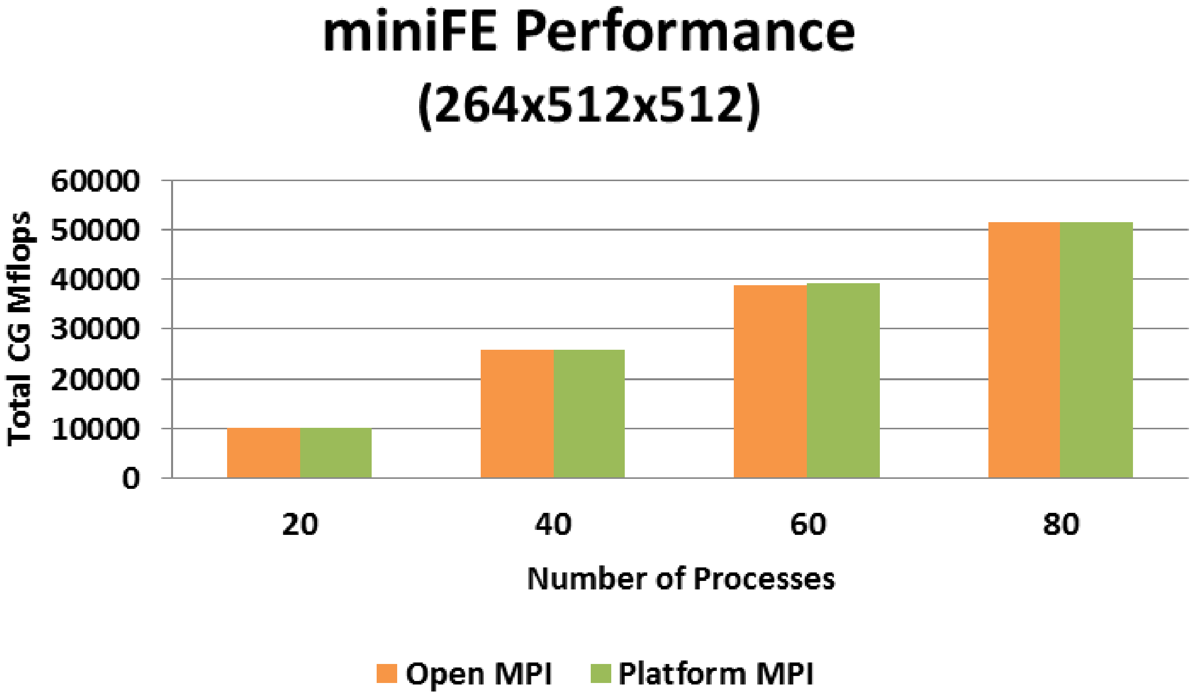


Higher is better

20 Processes/Node

MiniFE Performance - Interconnect

- **Both MPI in comparison shows similar performance**
  - No tuning flags used other than processor binding used in both cases
  - Same compiler flags have been used for both cases

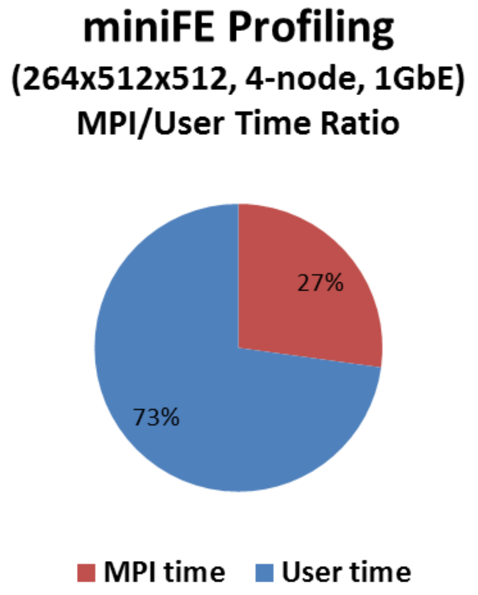


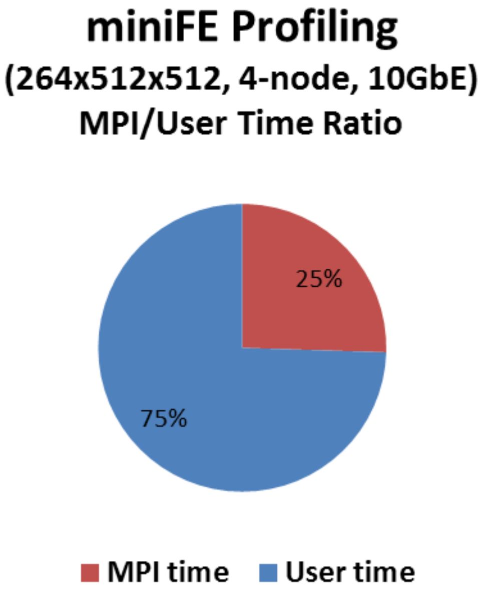
Higher is better

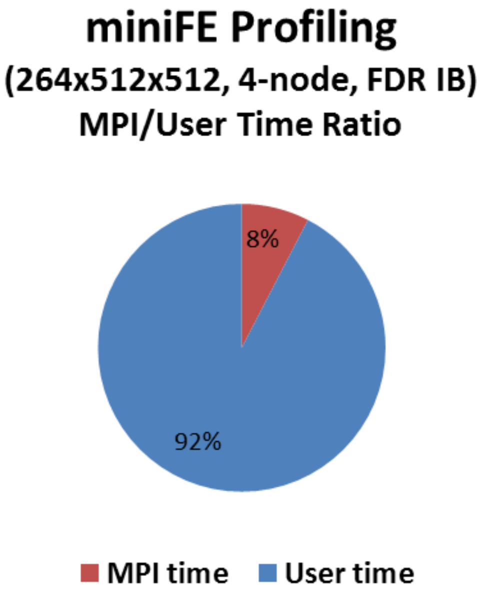
20 Processes/Node

MiniFE Profiling – MPI Time Ratio

- **FDR InfiniBand reduces the communication time at scale**
  - FDR InfiniBand consumes about 8% of total runtime
  - Compared to: 1GbE consumes 27% of total time, while 10GbE consumes about 25%



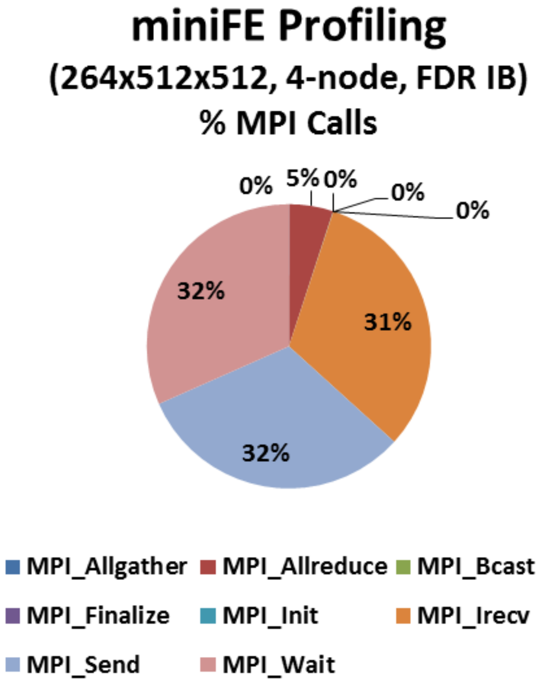


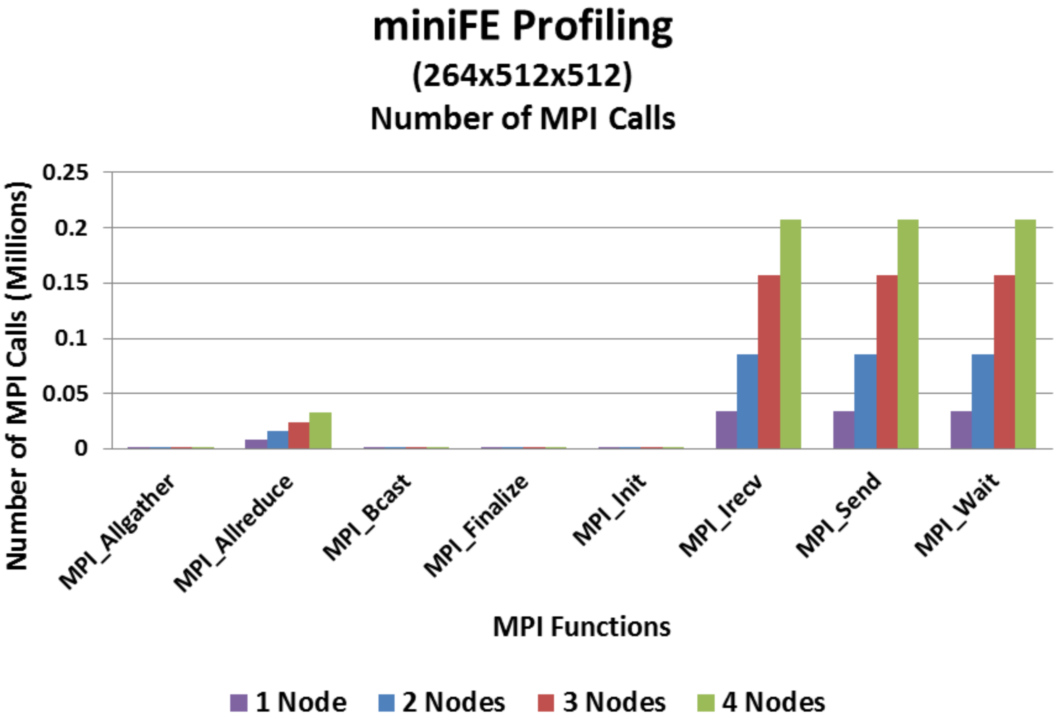


20 Processes/Node

MiniFE Profiling – MPI Functions

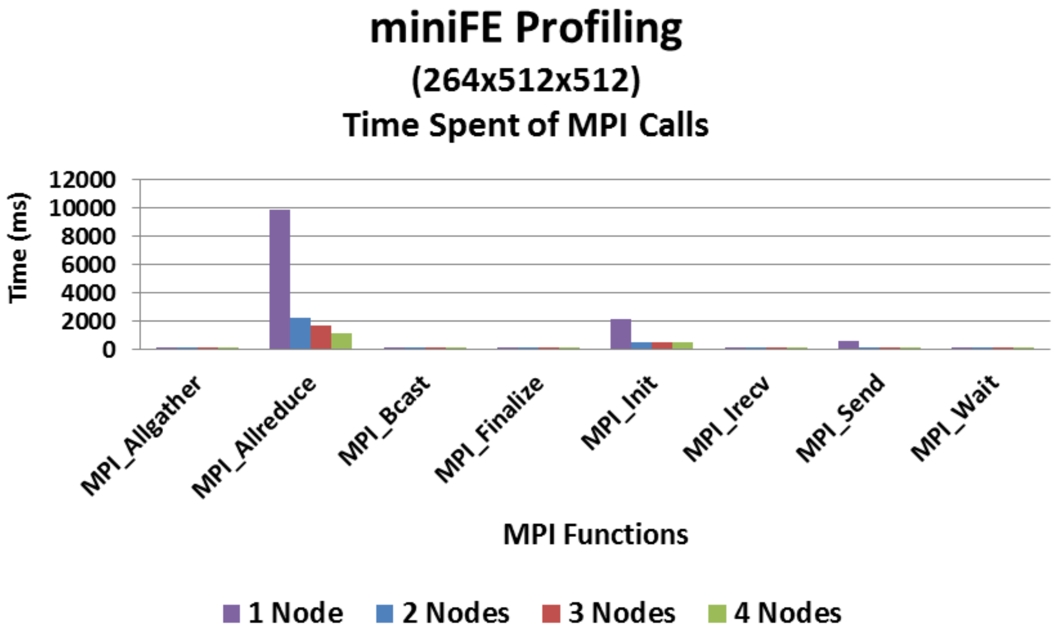
- **Mostl used MPI functions**
  - MPI_Wait (32%) and MPI_Send (32%), MPI_Irecv (31%)

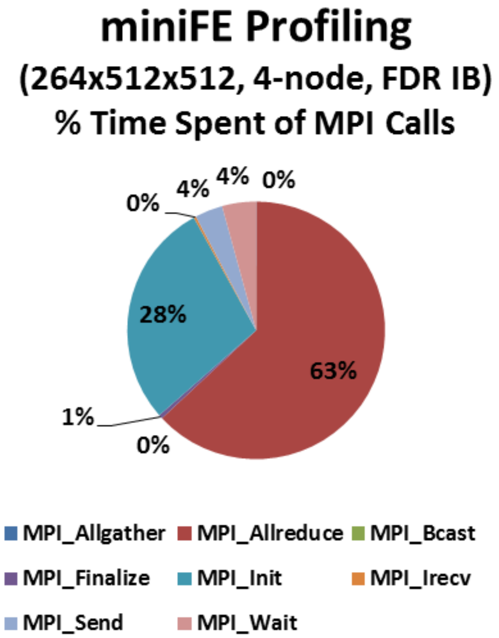




MiniFE Profiling – MPI Functions

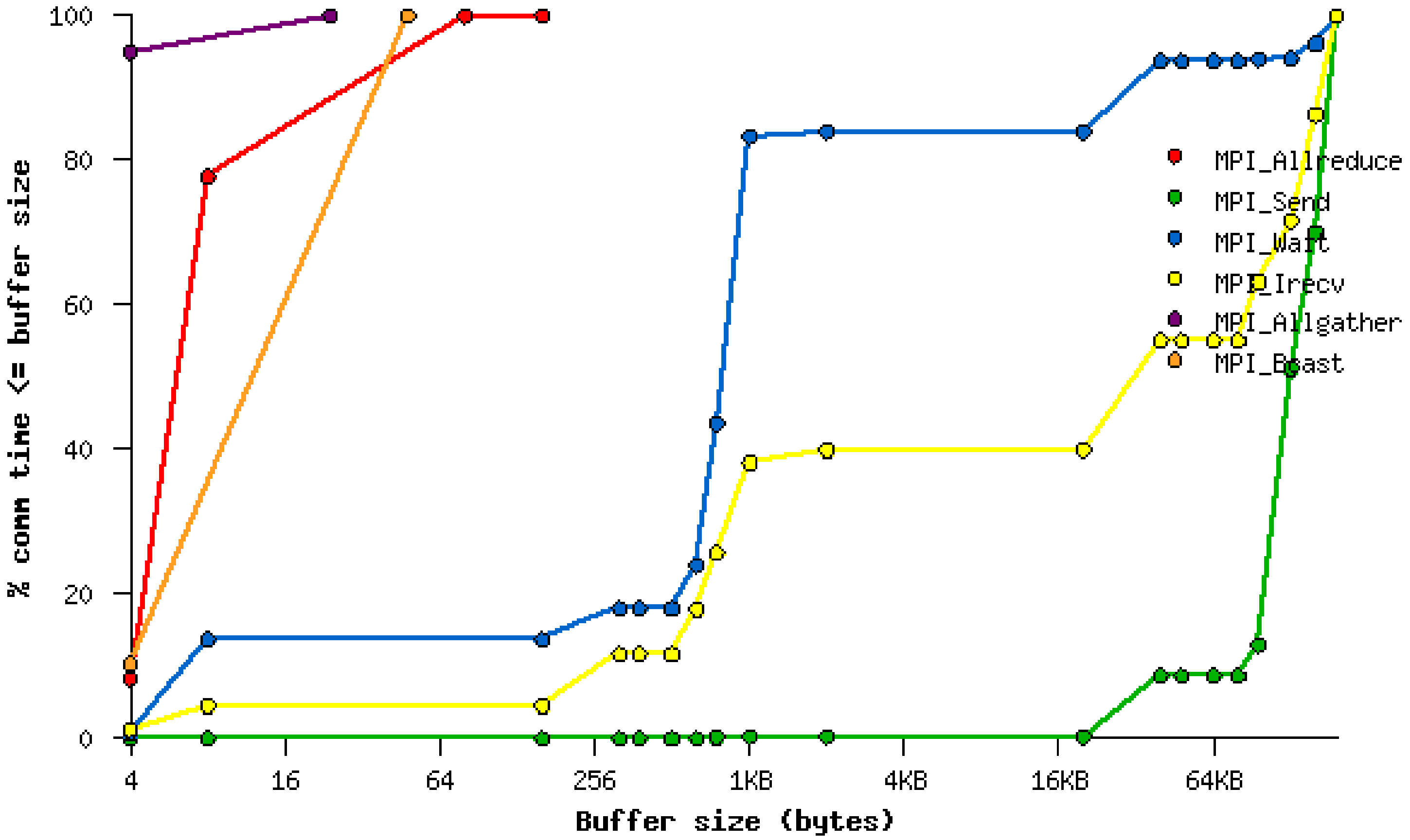
- **The most time consuming MPI functions:**
  - MPI_Allreduce (63%), MPI_Init(28%)





MiniFE Profiling – Message Size

- **Distribution of message sizes for the MPI calls**
  - MPI_Wait and MPI_Irecv between 256B to 1KB
  - MPI_Allreduce: small messages less than 64B



20 MPI Processes

MiniFE Summary

- **HP ProLiant Gen8 servers delivers better MiniFE Performance than its predecessor**
  - ProLiant Gen8 equipped with Intel Xeon E5-2600 V2 series processors and FDR InfiniBand
  - Provides 148% higher performance than the 4-node ProLiant G7 servers with X5670 procs
- **FDR InfiniBand is the most efficient inter-node communication for MiniFE**
  - Outperforms 10GbE up to 17-21% at 3 nodes
  - Outperforms 1GbE up to 22-33% at 3 nodes
- **MiniFE Profiling**
  - MPI Collective Operations (Allreduce) is the most time consumed communications
  - FDR InfiniBand reduces communication time; leave more time for computation
    - FDR InfiniBand consumes 8% of total time, versus 25% 10GbE, versus 27% 1GbE
  - Non-blocking communications are seen:
    - Time spent: MPI_Allreduce (63%)
    - Most used: MPI_Wait (32%) and MPI_Send (32%), MPI_Irecv (31%)

Thank You

HPC Advisory Council

All trademarks are property of their respective owners. All information is provided “As-Is” without any kind of warranty. The HPC Advisory Council makes no representation to the accuracy and completeness of the information contained herein. HPC Advisory Council undertakes no duty and assumes no obligation to update or correct any information presented herein