HPC ADVISORY COUNCIL

Nekbone Performance Benchmark and Profiling

January 2014

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

Nekbone

- **Nekbone**
  - Captures basic structure and user interface of the Nek5000 software
    - Nek5000 is a more extensive software which does high order, incompressible Navier-Stokes solver based on the spectral element method
  - Solves a standard Poisson equation
    - Using the spectral element method with an conjugate gradient (CG) iteration with a simple pre-conditioner on a block or linear geometry
  - Exposes the principal computational kernel to reveal the essential elements of the algorithmic architectural coupling that is pertinent to Nek5000
- **The Nekbone benchmark**
  - Is scalable and can accommodate a wide range of problem sizes
    - By specifying the number of spectral elements and the polynomial order of the elements
  - Consists of a setup phase and a solution phase
    - The solution phase consists of CG iterations that call the main computational kernel
    - which performs a matrix vector multiplication operation in an element-by-element fashion
  - Each iteration consists of:
    - vector, matrix-matrix multiply, nearest neighbor communication, and MPI_Allreduce operations
  - Written in Fortran and C, where C routines are used for the nearest neighbor communication and the rest of the compute kernel routines are in Fortran
  - The current version uses MPI parallelism with no threading

Objectives

HPC ADVISORY COUNCIL

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

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

Test Cluster Configuration

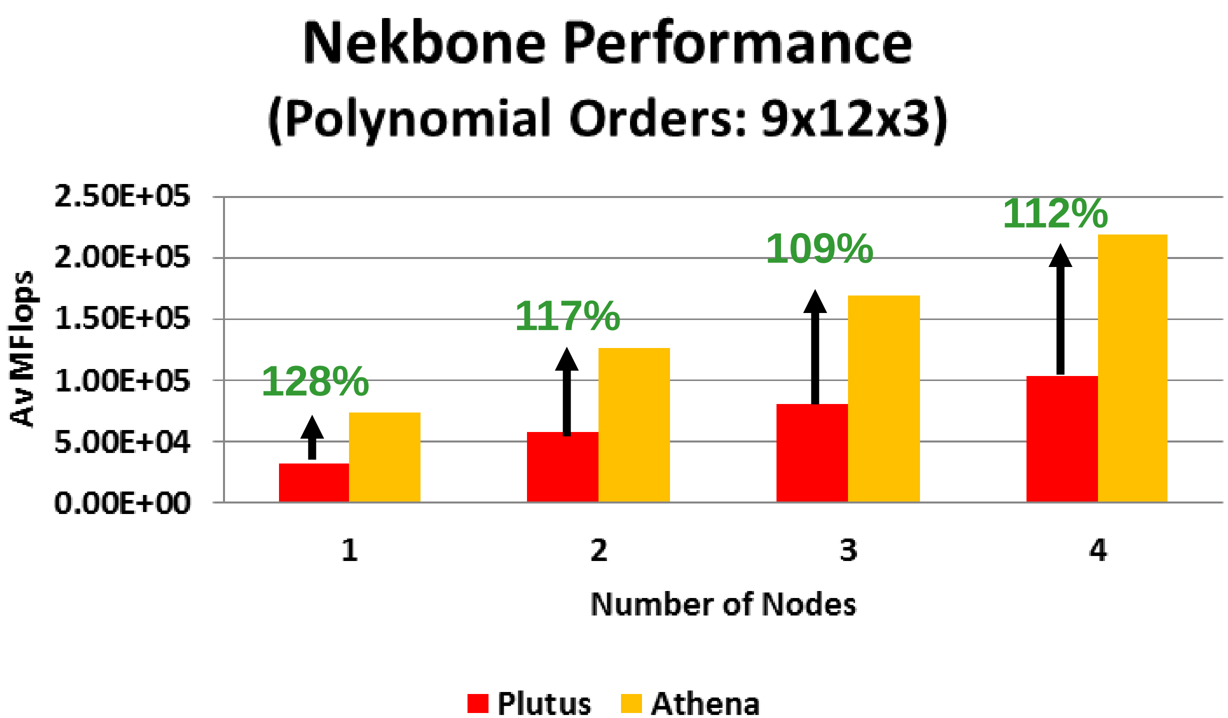
- **HP ProLiant SL230s Gen8 4-node "Athena" cluster**
  - Processors: Dual-Socket 10-core Intel Xeon E5-2680v2 @ 2.8 GHz CPUs
  - Memory: 32GB per node, 1600MHz DDR3 Dual-Ranked DIMMs
  - OS: RHEL 6 Update 2, OFED 2.0-3.0.0 InfiniBand SW stack
- **Mellanox Connect-IB FDR InfiniBand adapters**
- **Mellanox 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 (with MXM 2.5 and FCA 2.1)**
- **Compiler: GNU Compilers**
- **Application: Nekbone 2.1.2**
- **Benchmark Workload:**
  - Polynomial orders (nx0, nxN, and nxD) = (9x12x3), 8 elements per rank

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

Nekbone Performance - Processors

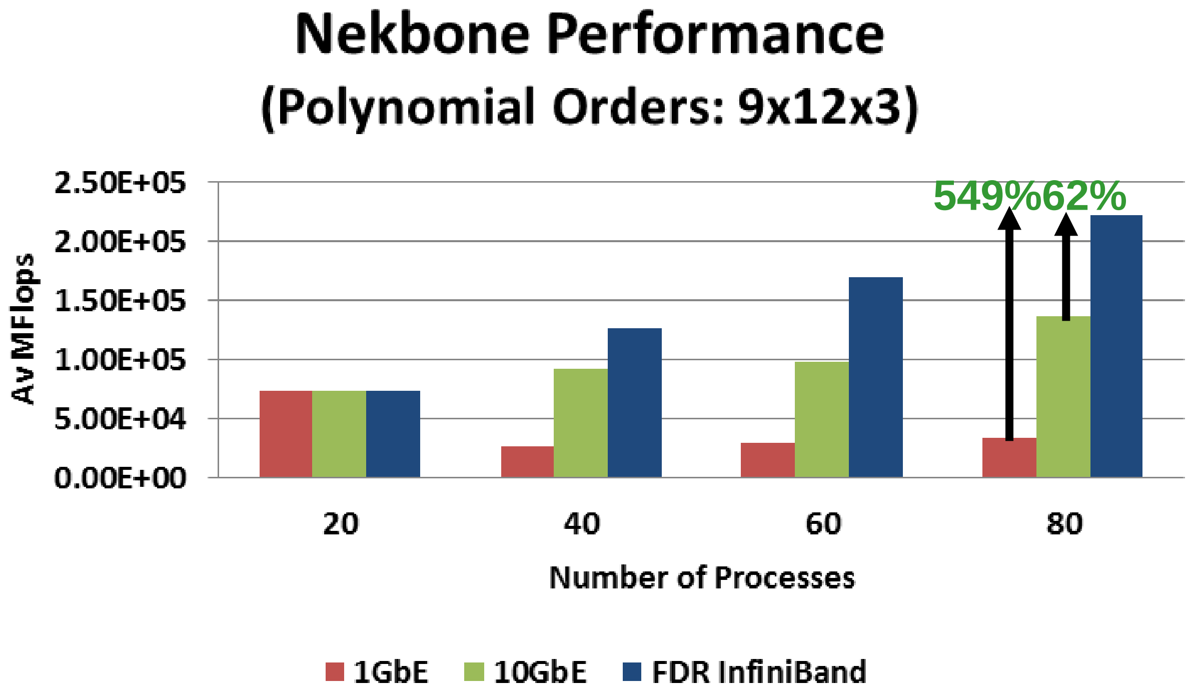
- **Intel E5-2680v2 processors (Ivy Bridge) cluster outperforms prior CPU generation**
  - Performs up to 128% higher than Xeon X5670 (Westmere) cluster
- **Configurations used:**
  - Athena: 2-socket Intel E5-2680v2 @ 2.8GHz, 1600MHz DIMMs, FDR IB, 20PPN
  - Plutus: 2-socket Intel X5670 @ 2.93GHz, 1333MHz DIMMs, QDR IB, 12PPN
  - Compiler optimization flags: “CFLAGS=-O3”



Higher is better

Nekbone Performance - Interconnect

- **FDR InfiniBand is the most efficient inter-node communication for Nekbone**
  - Outperforms 10GbE by 549% at 80 MPI processes
  - Outperforms 1GbE by 62% at 80 MPI processes
  - The performance benefit of InfiniBand expects to grow at larger CPU core counts

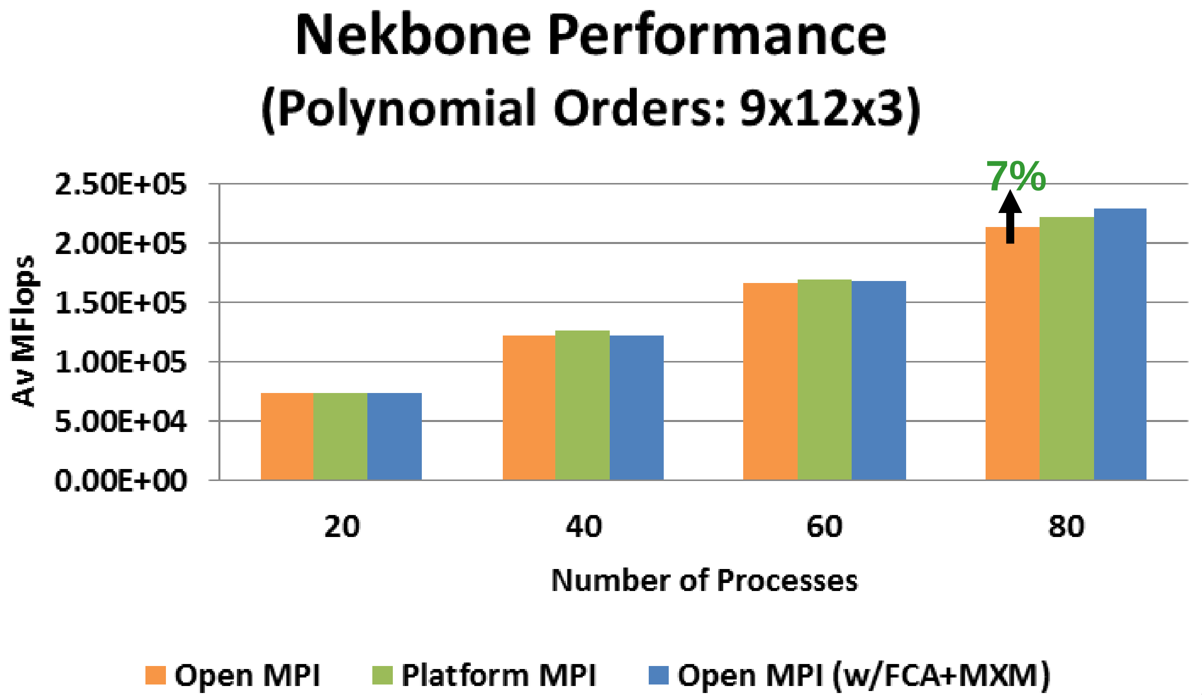


Higher is better

20 Processes/Node

Nekbone Performance - Interconnect

- **Tuned Open MPI performs better than untuned Open MPI and Platform MPI**
  - Up to 7% improved performance seen at 80 MPI processes over untuned Open MPI
  - Same compiler flags have been used for all 3 cases

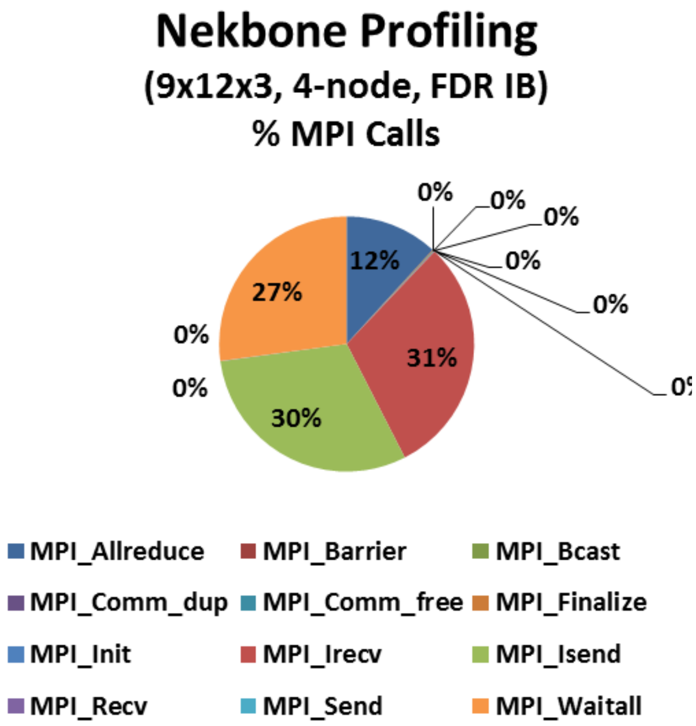


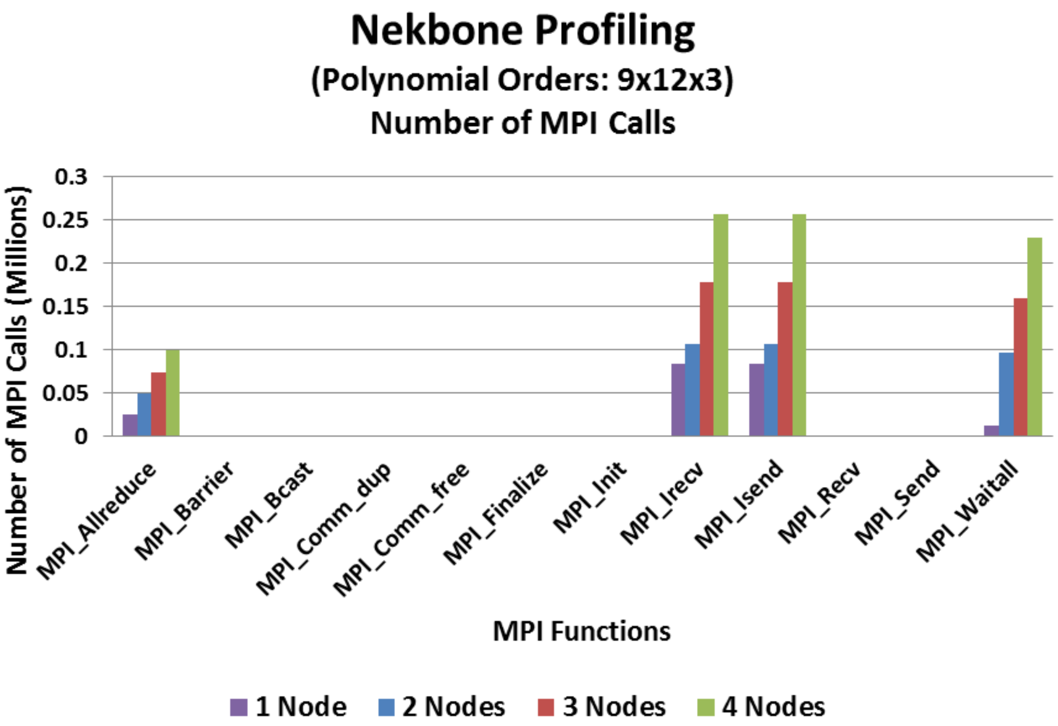
Higher is better

20 Processes/Node

Nekbone Profiling – MPI Functions

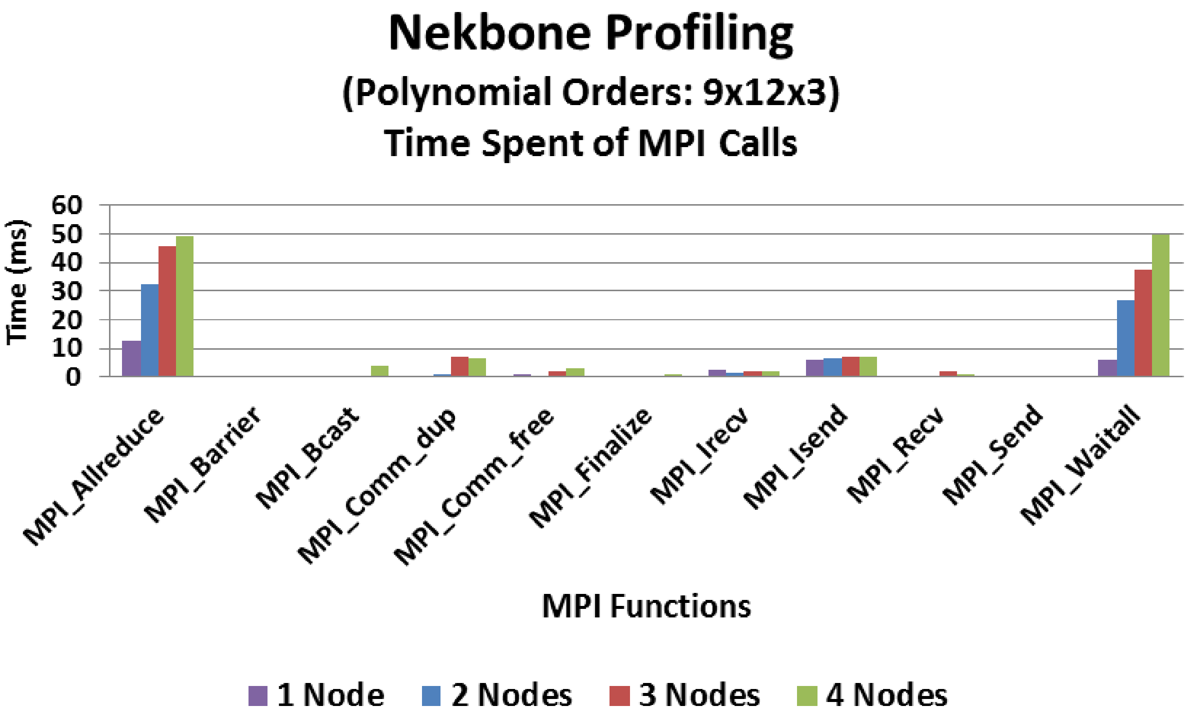
- **Mostly used MPI functions**
  - MPI_Irecv (31%) and MPI_Isend (30%), MPI_Waitall (27%), MPI_Allreduce (12%)

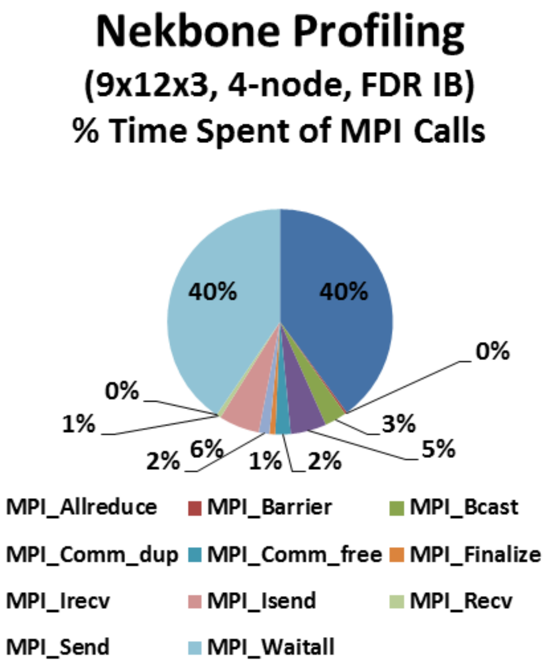




Nekbone Profiling – MPI Functions

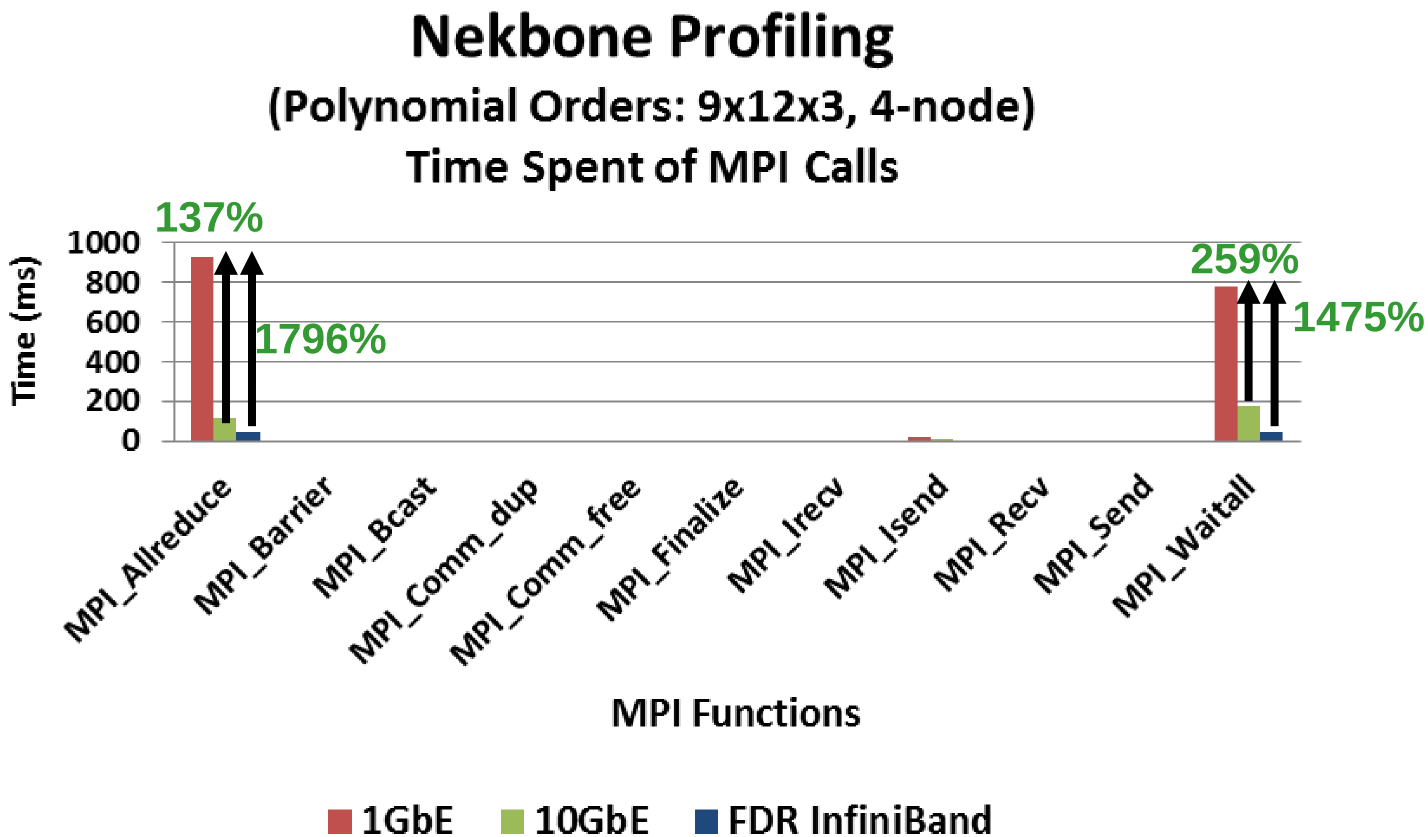
- **The most time consuming MPI functions:**
  - MPI_Allreduce (40%), MPI_Waitall (40%), MPI_Isend (6%), MPI_Bcast (5%)





Nekbone Profiling – MPI Time Ratio

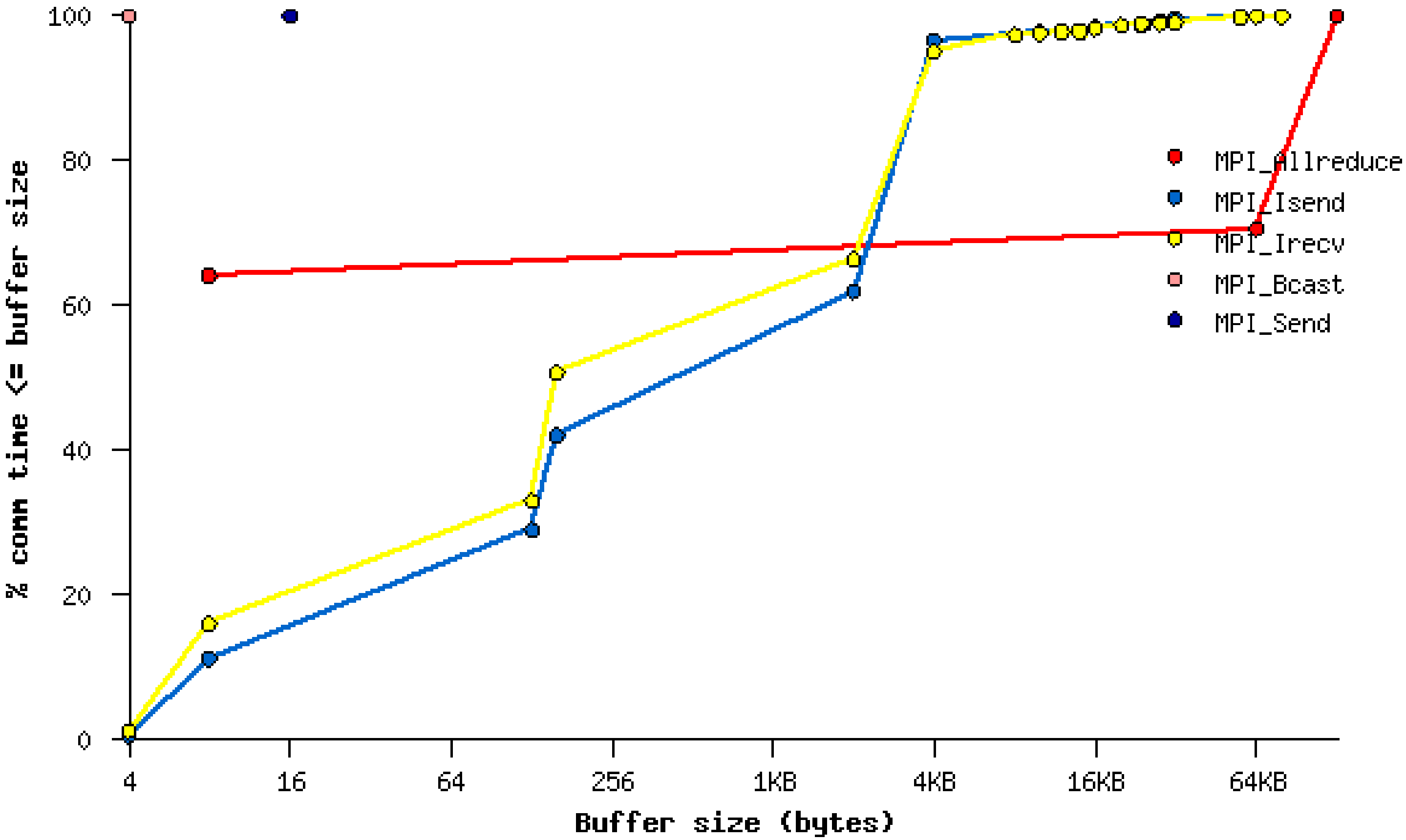
- **FDR InfiniBand reduces the communication time at scale**
  - MPI_Allreduce: 1GbE takes ~18x longer, and 10GbE spends 1.37x longer than FDR IB
  - MPI_Waitall: 1GbE takes ~15x longer, while 10GbE consumes about 2.59x longer than IB



20 Processes/Node

Nekbone Profiling – Message Size

- **Distribution of message sizes for the MPI calls**
  - Concentrate of MPI_Allreduce at around 8B
  - Majority of MPI_Irecv and MPI_Isend occur below 4KB



20 MPI Processes

Nekbone Summary

- **HP ProLiant Gen8 servers delivers better Nekbone Performance than its predecessor**
  - ProLiant Gen8 equipped with Intel E5 2600 V2 series processors and FDR InfiniBand
  - Provides 128% higher performance than the ProLiant G7 (Westmere) servers at 4 nodes
- **FDR InfiniBand is the most efficient inter-node communication for Nekbone**
  - Outperforms 10GbE by 62% with 4 nodes, and beat 1GbE over 5.5x with 4 nodes
- **Nekbone Profiling**
  - FDR InfiniBand reduces communication time; leave more time for computation
    - MPI_Allreduce: 1GbE takes 18x longer, and 10GbE spends 1.37x longer than FDR IB
    - MPI_Waitall: 1GbE takes 15x longer, while 10GbE consumes about 2.59x longer than FDR IB
  - Non-blocking and collective operations communications are seen:
    - Time spent: MPI_Allreduce (40%), MPI_Waitall (40%), MPI_Isend (6%), MPI_Bcast (5%)
    - Most used: MPI_Irecv (31%) and MPI_Isend (30%), MPI_Waitall (27%), MPI_Allreduce (12%)
  - Distribution of MPI messages:
    - MPI_Allreduce at 8B, MPI_Irecv/MPI_Isend <4KB

Thank You

HPC Advisory Council