HPC ADVISORY COUNCIL

CPMD
Performance Benchmark and Profiling

February 2014

CPMD

hp

Mellanox TECHNOLOGIES

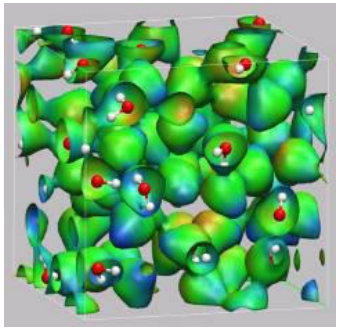
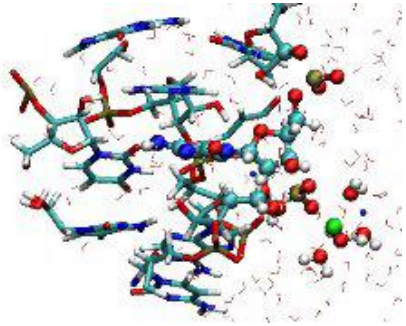
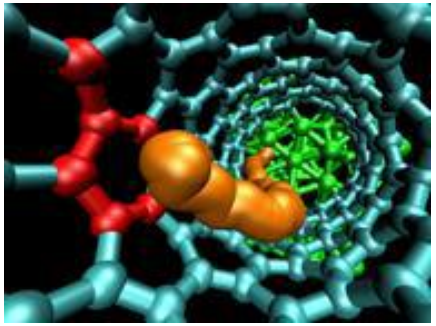
Note

- **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.cpmd.org/

- **CPMD**
  - A parallelized implementation of density functional theory (DFT)
  - Plane wave / pseudopotential implementation of DFT
  - Particularly designed for ab-initio molecular dynamics
  - Brings together methods
    - Classical molecular dynamics
    - Solid state physics
    - Quantum chemistry
  - CPMD supports MPI and Mixed MPI/SMP
  - CPMD is distributed and developed by the CPMD consortium

Objectives

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

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

Test Cluster Configuration

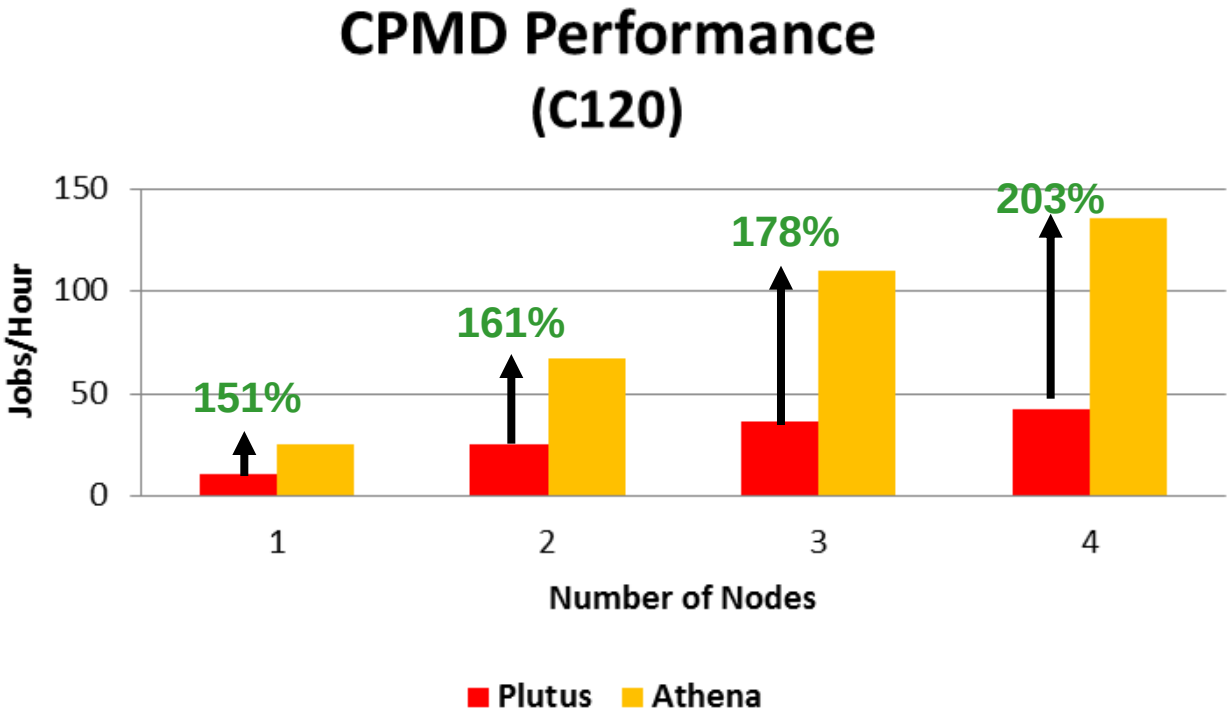
- **HP ProLiant SL230s Gen8 4-node "Athena" cluster**
  - Processors: Dual-Socket 10-core Intel Xeon E5-2680v2 @ 2.8 GHz CPUs
  - Memory: 32GB per node, 1600MHz DDR3 Dual-Ranked DIMMs
  - OS: RHEL 6 Update 2, OFED 2.1-1.0.0 InfiniBand SW stack
- **Mellanox Connect-IB FDR InfiniBand adapters**
- **Mellanox ConnectX-3 VPI adapters**
- **Mellanox SwitchX SX6036 56Gb/s FDR InfiniBand and Ethernet VPI Switch**
- **MPI: Platform MPI 8.3, Open MPI 1.7.4 (with MXM 2.1 and FCA 2.5)**
- **Compiler: Intel Composer XE 2013 (2013.5.192)**
- **Application: CPMD 3.17.1**
- **Benchmark Workload:**
  - C120 – 120 carbon atoms

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

CPMD Performance - Processors

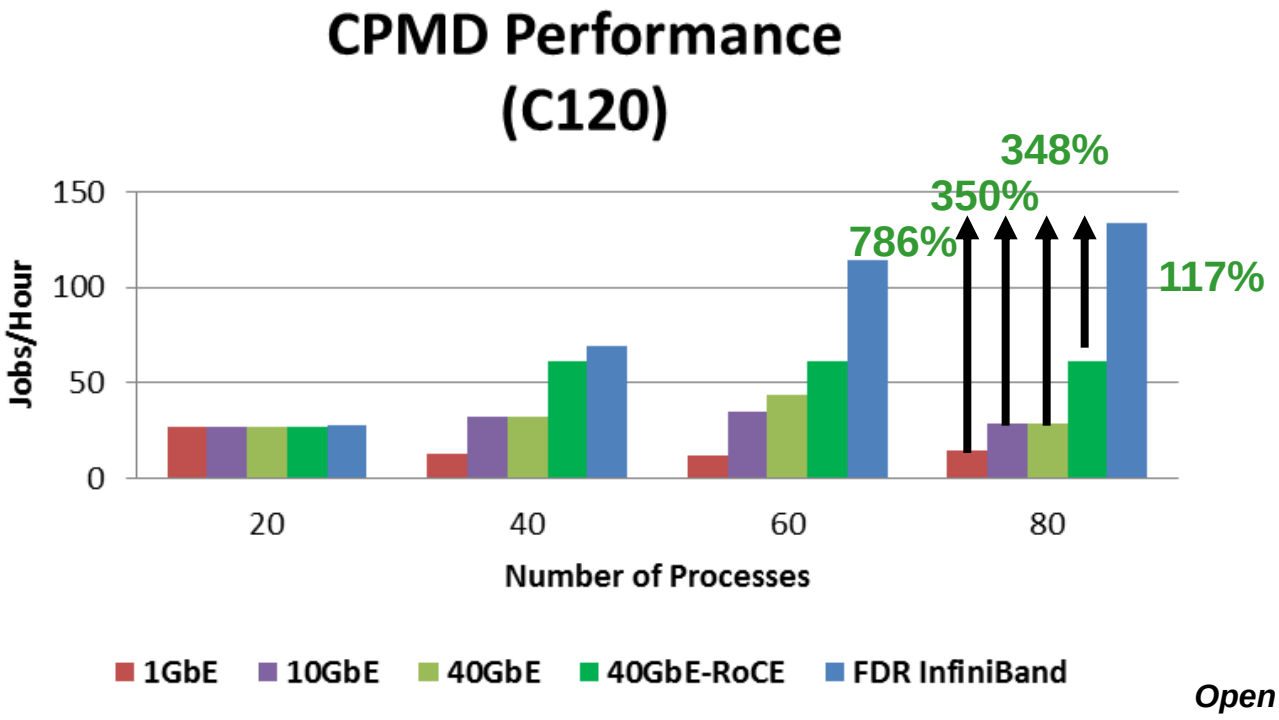
- **Intel E5-2680v2 processors (Ivy Bridge) cluster outperforms prior CPU generation**
  - Performs up to 203% higher than Xeon X5670 (Westmere) cluster at 4 nodes
- **Configurations used:**
  - Athena: 2-socket Intel E5-2680v2 @ 2.8GHz, 1600MHz DIMMs, FDR IB, 20PPN
  - Plutus: 2-socket Intel X5670 @ 2.93GHz, 1333MHz DIMMs, QDR IB, 12PPN
  - Same set of compiler optimization flags used in both cases



Higher is better

CPMD Performance - Interconnect

- **FDR InfiniBand is the most efficient inter-node communication for CPMD**
  - FDR outperforms 1GbE by over 7 times at 80 MPI processes
  - FDR also outperforms 10GbE and 40GbE by over 3.5 times at 80 MPI processes
  - FDR provides 117% higher performance over 40GbE-RoCE
  - The performance benefit of InfiniBand expects to grow at larger CPU core counts

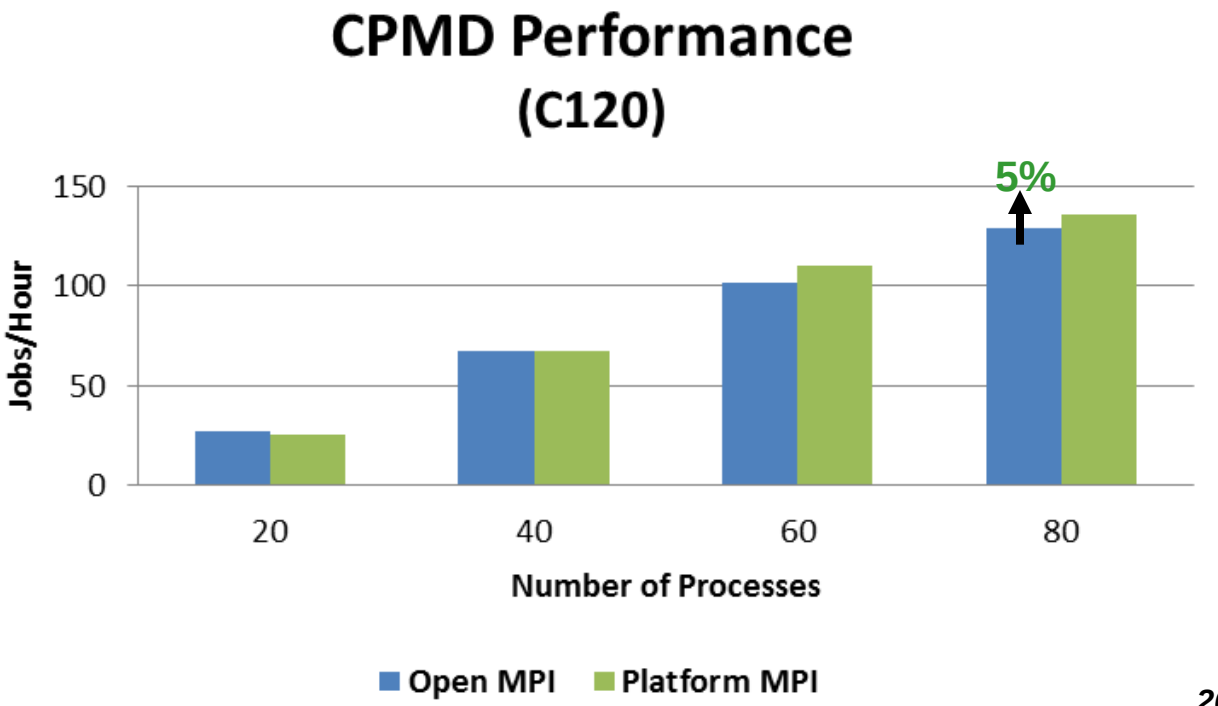


Higher is better

Open MPI

CPMD Performance - Interconnect

- **Tuned Platform MPI performs better than Open MPI**
  - Slight advantage is seen at 80 MPI processes over Open MPI
  - Same compiler flags have been used for both cases

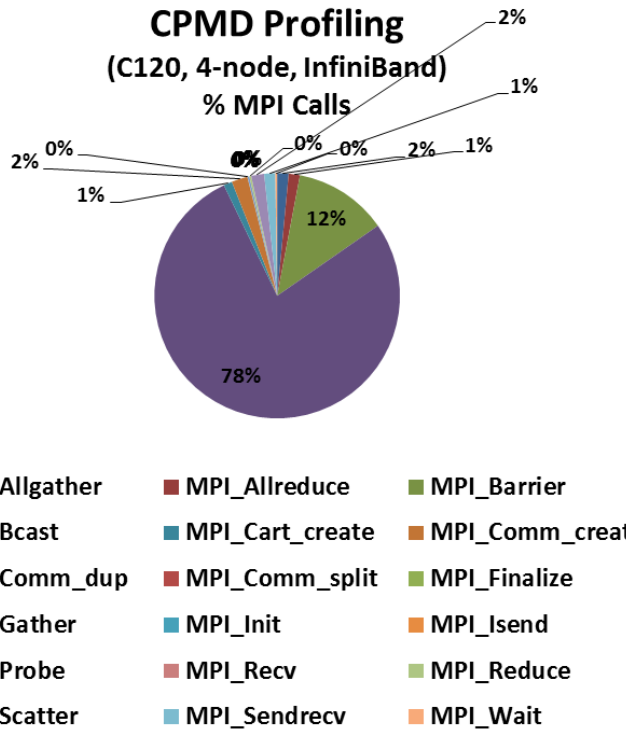


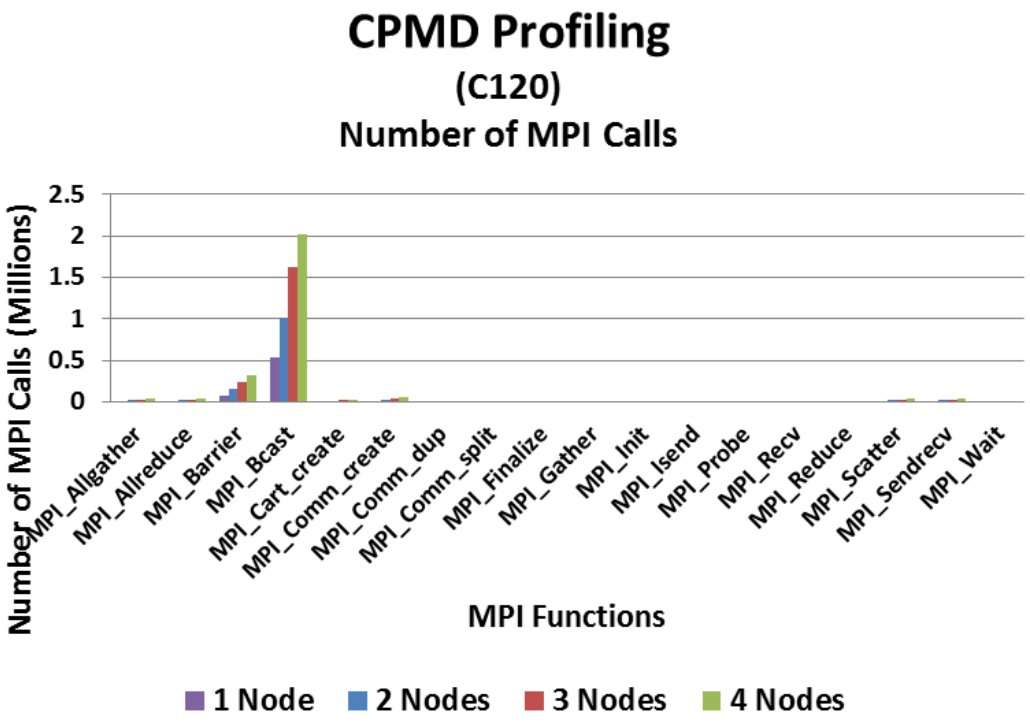
Higher is better

20 Processes/Node

CPMD Profiling – # of MPI Functions

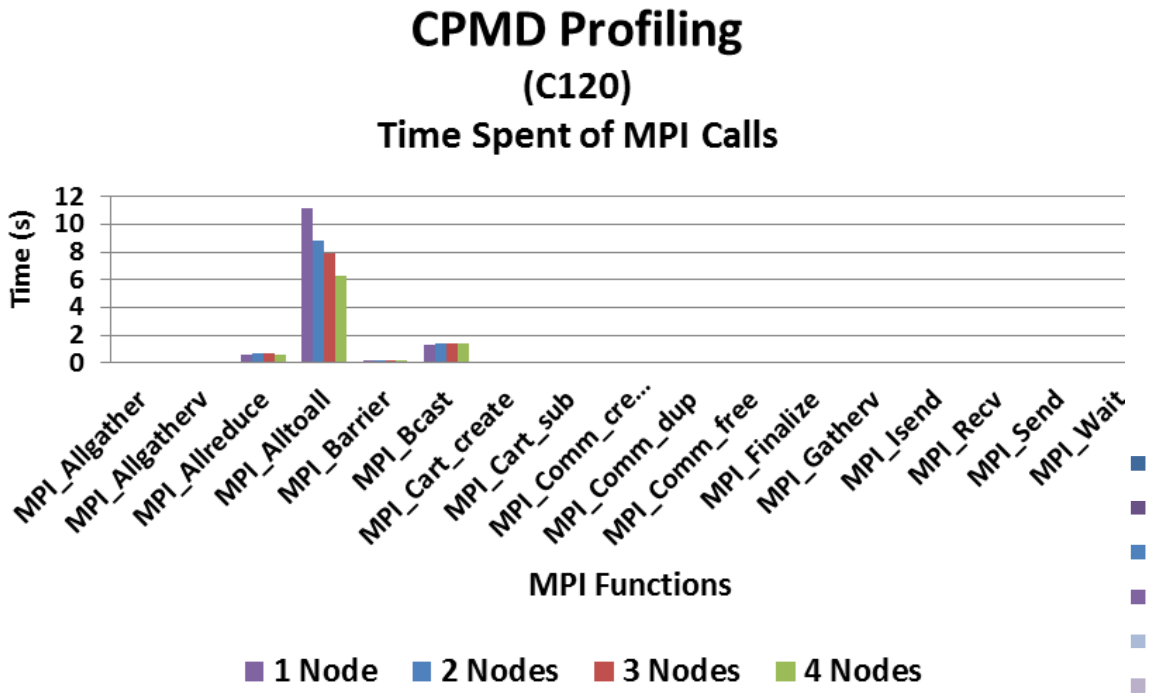
- **Mostly used MPI functions**
  - MPI_Bcast (78%) and MPI_Barrier (12%)

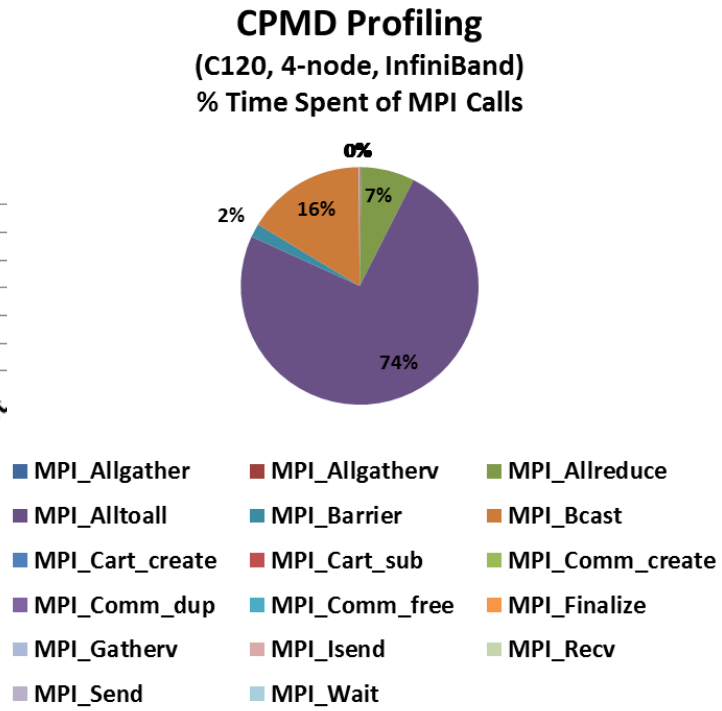




CPMD Profiling – Time Spent on MPI

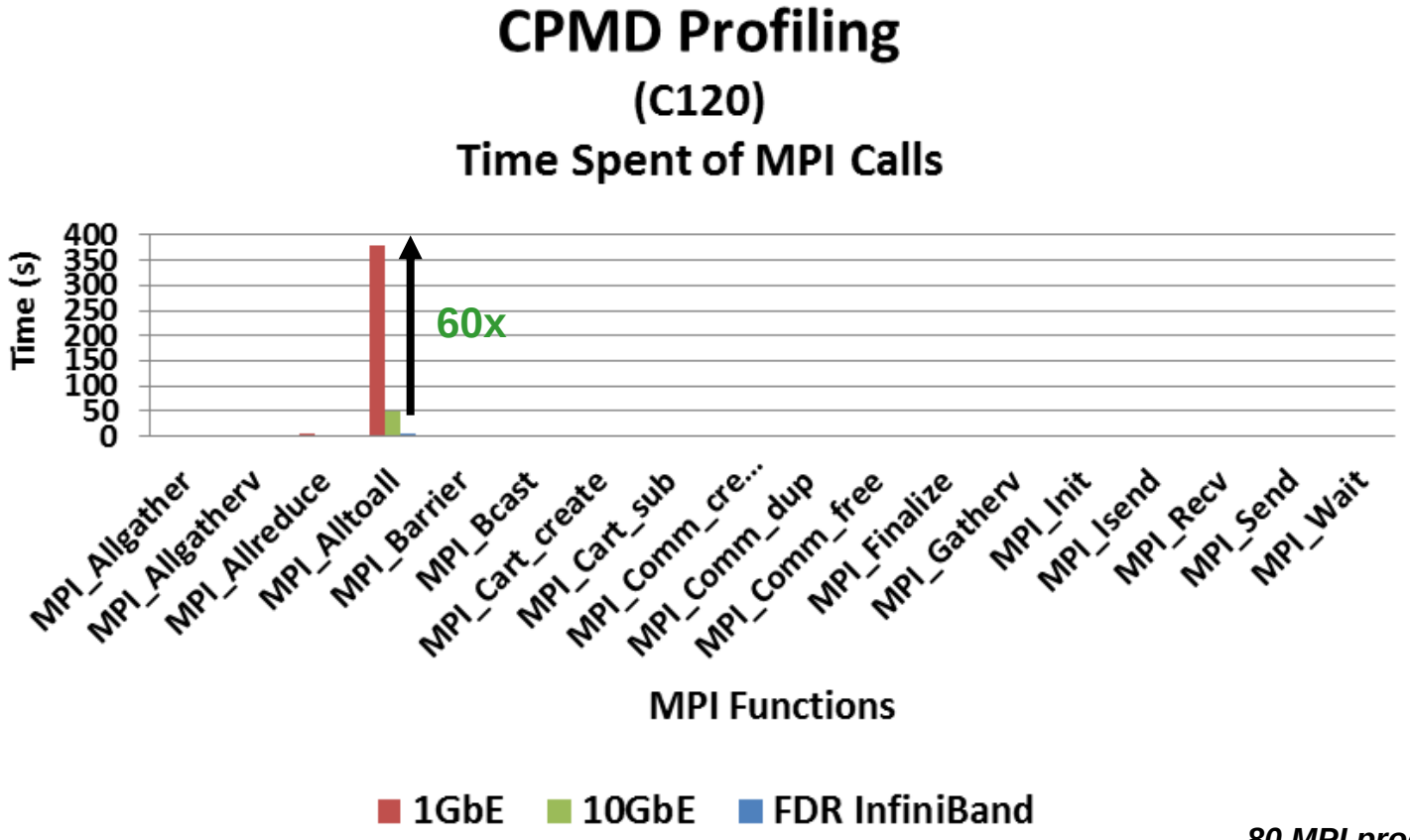
- **The most time consuming MPI functions:**
  - MPI_Alltoall (74%), MPI_Bcast (16%), MPI_Allreduce (7%), MPI_Barrier (2%)
  - Mostly MPI collective operations used





CPMD Profiling – MPI Time Ratio

- **FDR InfiniBand reduces the communication time at scale for MPI Collectives**
  - MPI_Alltoall: 1GbE takes >60x longer, and 10GbE spends ~7x longer than FDR IB
  - MPI_Allreduce: 1GbE takes ~7x longer, while 10GbE run about 82% longer than IB



80 MPI processes

CPMD Summary

- **HP ProLiant Gen8 servers delivers better CPMD Performance than its predecessor**
  - ProLiant Gen8 equipped with Intel E5 2600 V2 series processors and FDR InfiniBand
  - Provides 203% higher performance than the ProLiant G7 (Westmere) servers at 4 nodes
- **FDR InfiniBand is the most efficient inter-node communication for CPMD**
  - FDR IB outperforms 10/40GbE by >3.5x with 4 nodes, and beat 1GbE over 7x with 4 nodes
  - FDR IB also outperforms 40GbE-RoCE by 117% at 4 nodes
- **CPMD Profiling**
  - FDR InfiniBand reduces communication time; leave more time for computation
    - MPI_Alltoall: 1GbE takes >60x longer, and 10GbE spends ~7x longer than FDR IB
    - MPI_Allreduce: 1GbE takes ~7x longer, while 10GbE run about 82% longer than FDR IB
  - Collective operations communications are seen:
    - Time spent: MPI_Alltoall (74%), MPI_Bcast (16%), MPI_Allreduce (7%), MPI_Barrier (2%)
    - Most used: MPI_Bcast (78%) and MPI_Barrier (12%)

Thank You

HPC Advisory Council