HPC ADVISORY COUNCIL

HPCG
Performance Benchmark and Profiling

July 2014

HP

Mellanox TECHNOLOGIES

Note

- **The following research was performed under the HPC Advisory Council activities**
  – Participating vendors: 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://software.sandia.gov/hpcg

Objectives

HPC ADVISORY COUNCIL

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

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

HPCG

- **HPCG Benchmark project**
  - An effort to create a more relevant metric for ranking HPC systems
  - Potential replacement for the High Performance LINPACK (HPL) benchmark
  - Currently HPL is used by the TOP500 benchmark
- **HPCG**
  - **H**igh **P**erformance **C**onjugate **G**radient
  - Stand-alone code that measures the performance of basic operations
    - Sparse matrix-vector multiplication
    - Sparse triangular solve
    - Vector updates
    - Global dot products
    - Local symmetric Gauss-Seidel smoother
  - Driven by multigrid preconditioned CG algorithm that exercises the key kernels on a nested set of coarse grids
  - Reference implementation is written in C++ with MPI and OpenMP support

Test Cluster Configuration

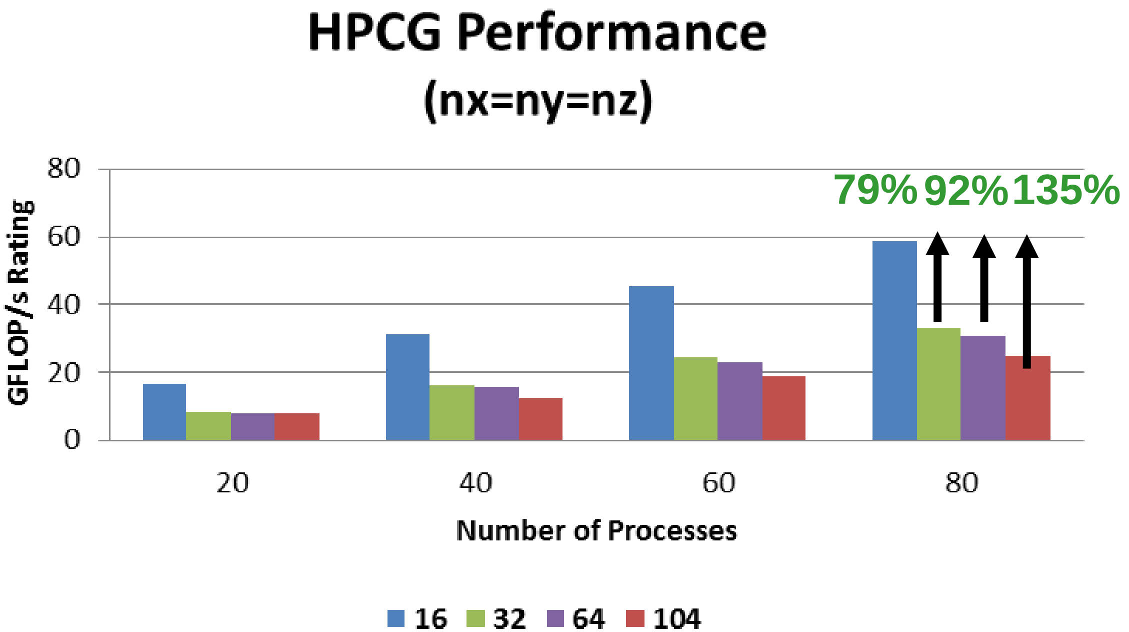
- **HP ProLiant SL230s Gen8 4-node “Athena” cluster**
  - Processors: Dual-Socket 10-core Intel Xeon E5-2680v2 @ 2.8 GHz CPUs
  - Memory: 32GB per node, 1600MHz DDR3 Dual-Ranked DIMMs
  - OS: RHEL 6 Update 2, OFED 2.2-1.0.1 InfiniBand SW stack
- **Mellanox Connect-IB FDR InfiniBand adapters**
- **Mellanox ConnectX-3 VPI Ethernet adapters**
- **Mellanox SwitchX SX6036 56Gb/s FDR InfiniBand and Ethernet VPI Switch**
- **MPI: Mellanox HPC-X v1.0.0, Platform MPI 9.1.2**
- **Compiler: Composer XE 2013 SP1**
- **Application: HPCG 2.4**
- **Benchmark Workload:**
  - Local domain dimensions 16x16x16, Runtime for 60 seconds unless otherwise stated

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

HPCG Performance – Domain Dimensions

- **Adjusting local domain dimensions can affect global problem size**
  - User specifies local domain in hpcg.dat which predicts problem size
- **Higher performance is observed when small problem is specified**
  - Advantageous to tune the local dimension to a lower number
  - Values under 16 will be defaulted to 16 (for a 16x16x16 mesh)
  - Up to 135% higher performance against using the default (104x104x104)

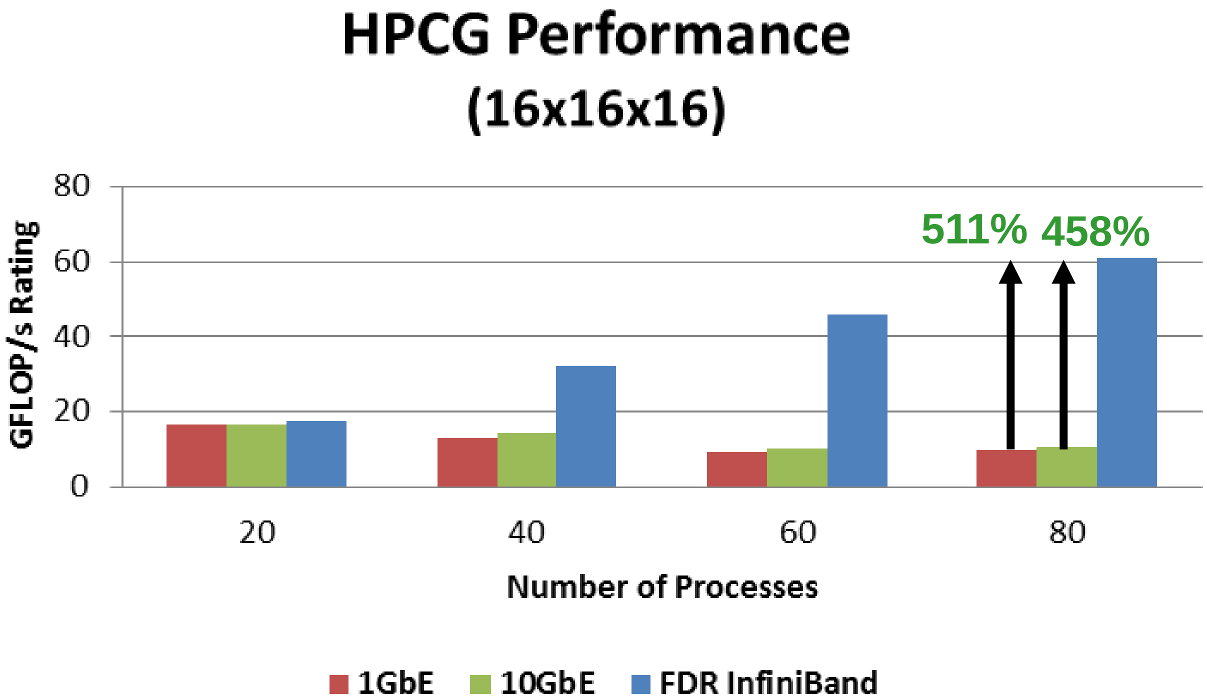


Higher is better

FDR InfiniBand

HPCG Performance – Network

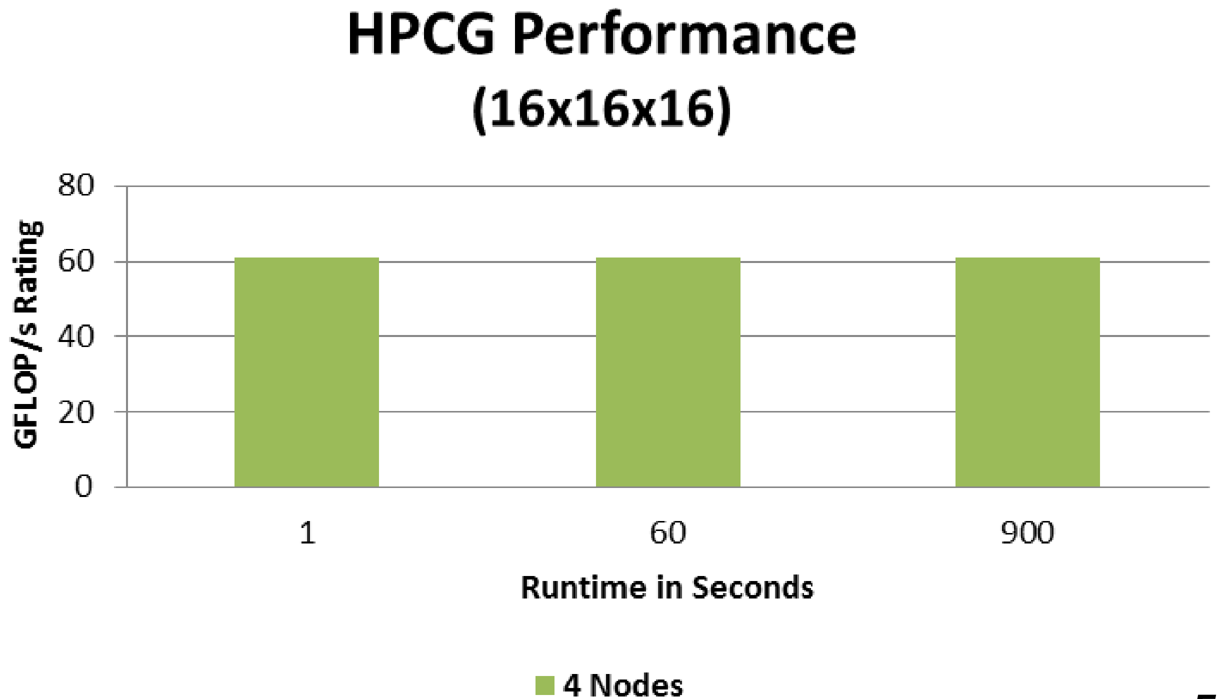
- **FDR InfiniBand delivers higher performance against Ethernet**
  - Over 5 times against 1GbE, and 4.5 times over 10GbE
  - Scalability advantage can be seen beyond a single node for HPCG



Higher is better

HPCG Performance – Runtime

- **No advantage is observed by running at a longer duration**
  - Although official results requires the execution time to be >=3600 seconds
  - Duration of the run does not appear to a factor in the performance at all

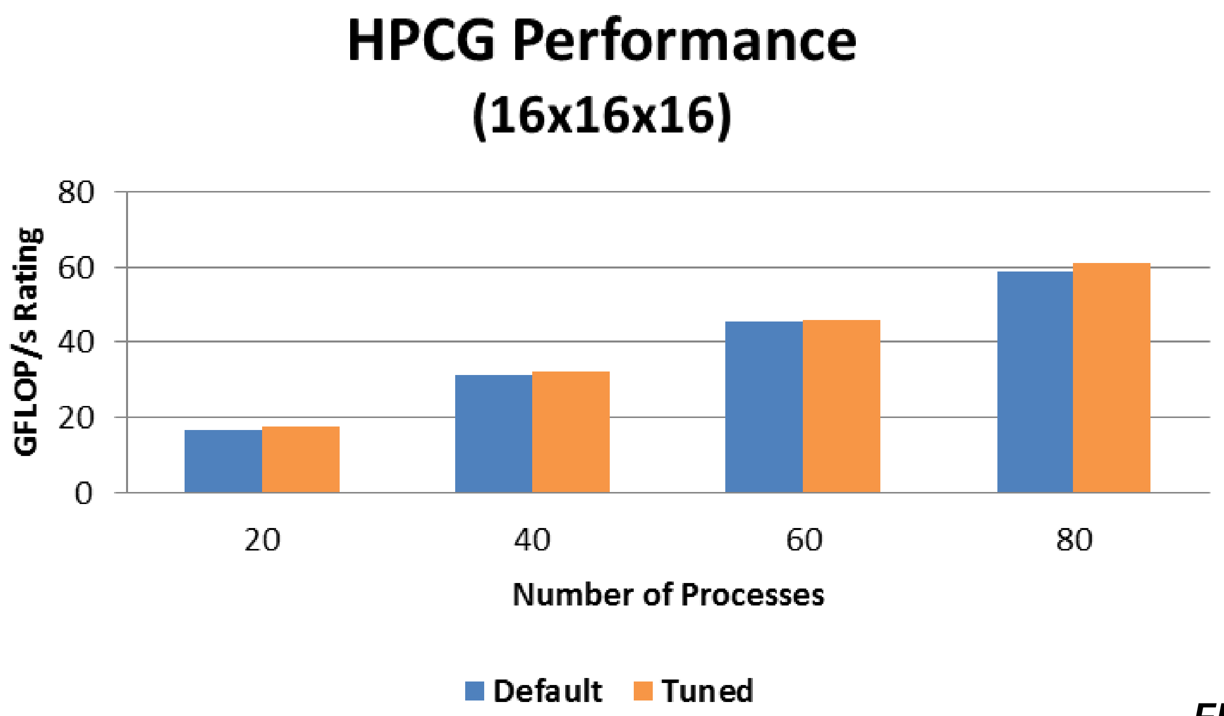


Higher is better

FDR InfiniBand

HPCG Performance – Compiler Options

- **Little advantage is observed by tuning the CXXFLAGS option**
  - Small increase (~2%) of increased performance is seen
  - Default: -O3
  - Tuned: -O3 -unroll-aggressive -no-prec-div -ipo -xHost -axavx

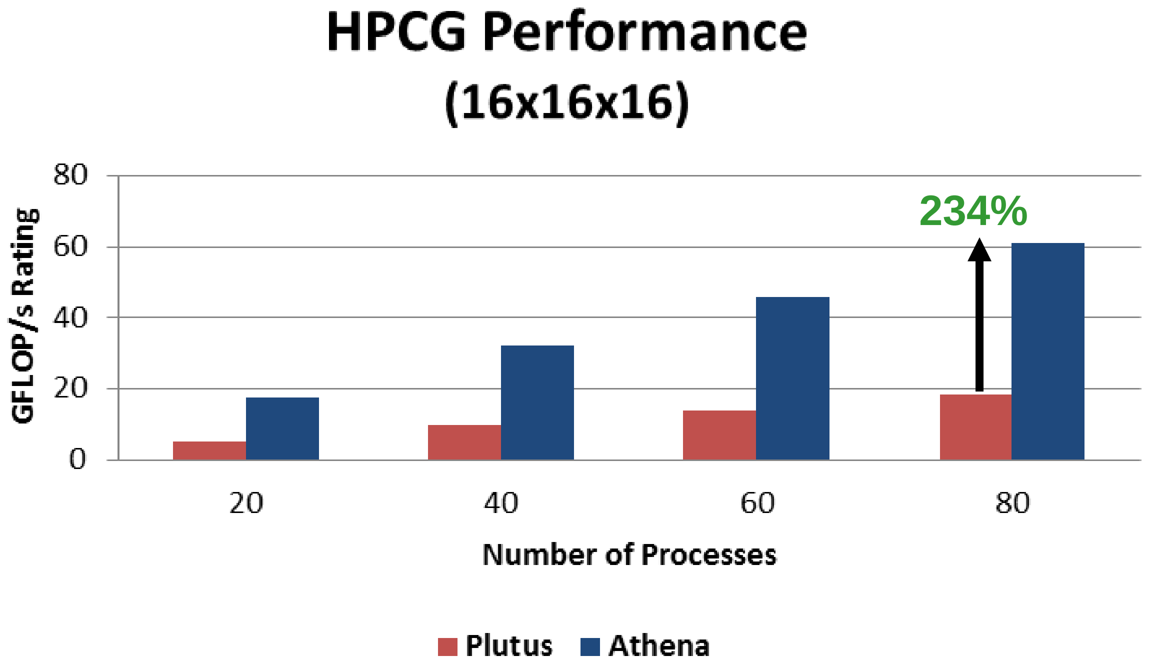


Higher is better

FDR InfiniBand

HPCG Performance – System Generations

- **Athena cluster outperforms prior generation cluster**
  - Up to 234% higher performance than the Plutus cluster
  - Executable for Athena is compiled with AVX while Plutus is with SSE4.2
- **System components used:**
  - Athena: Dual 10-core E5-2680v2@2.8GHz, 1600MHz DIMMs, FDR IB
  - Plutus: Dual 6-core x5670@2.93GHz, 1333MHz DIMMs, QDR IB

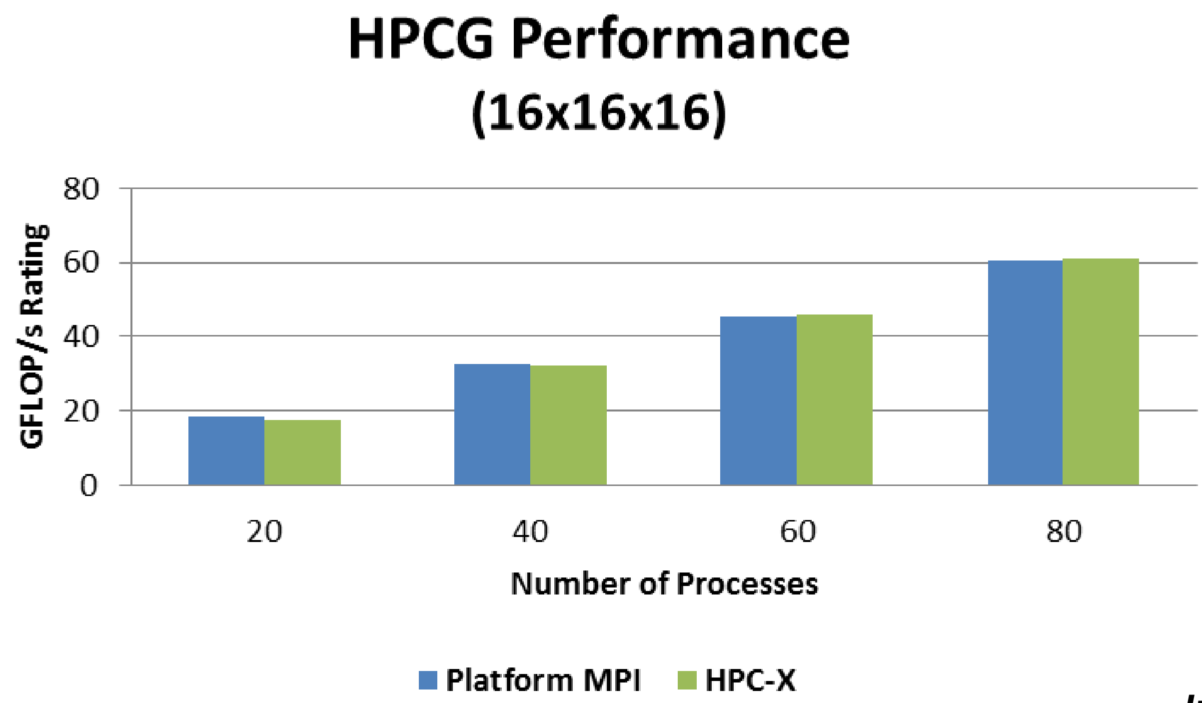


Higher is better

Tuned Compiler

HPCG Performance – MPI

- **Both MPI implementations show comparable performance**
  - Reflect that both MPIs handle MPI calls used in HPCG efficiently
  - Limited variety of calls and different message sizes were made in profiling

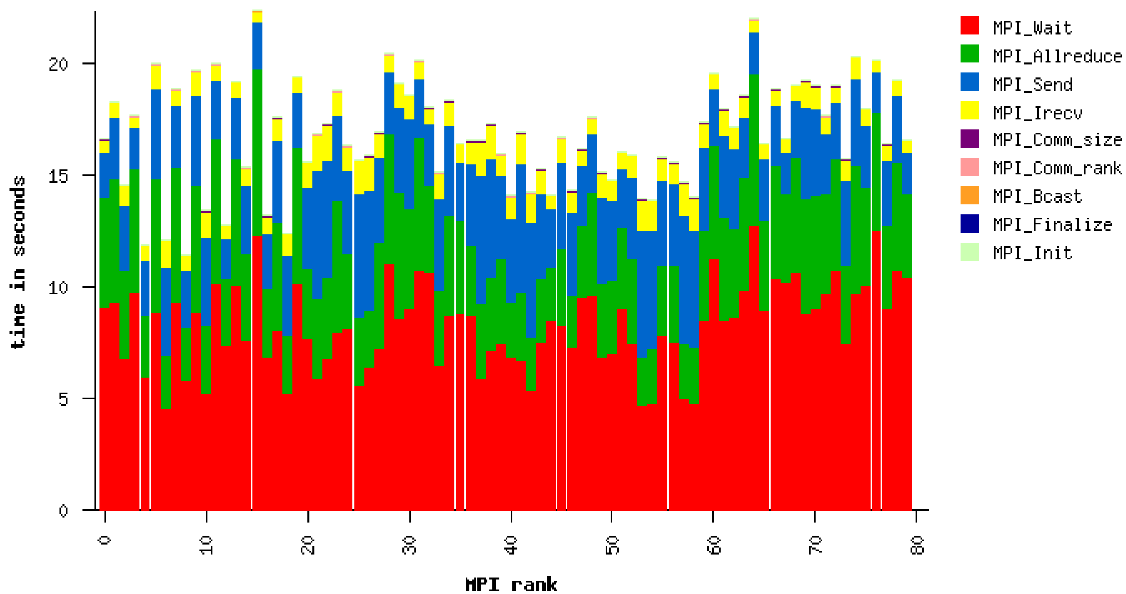


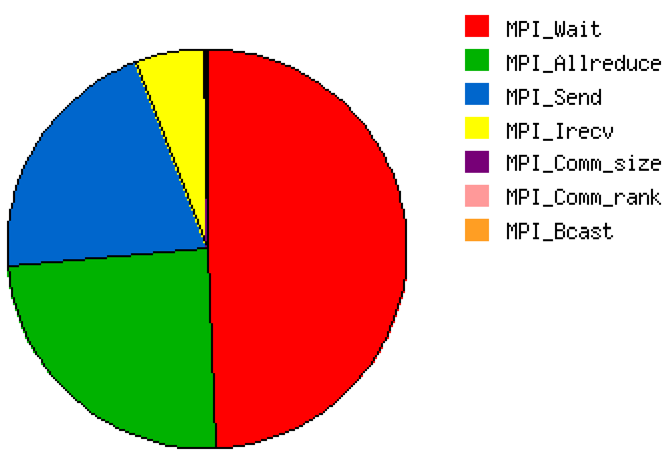
Higher is better

Intel E5-2680v2

- **Majority of the MPI time is spent on MPI_send and MPI_Allreduce**
  - MPI_Wait(~49%), MPI_Allreduce(~24%), MPI_Send(~20%)
  - Some load imbalances are seen
  - About 28% of time spent in MPI communications at 4 nodes (80 processes)

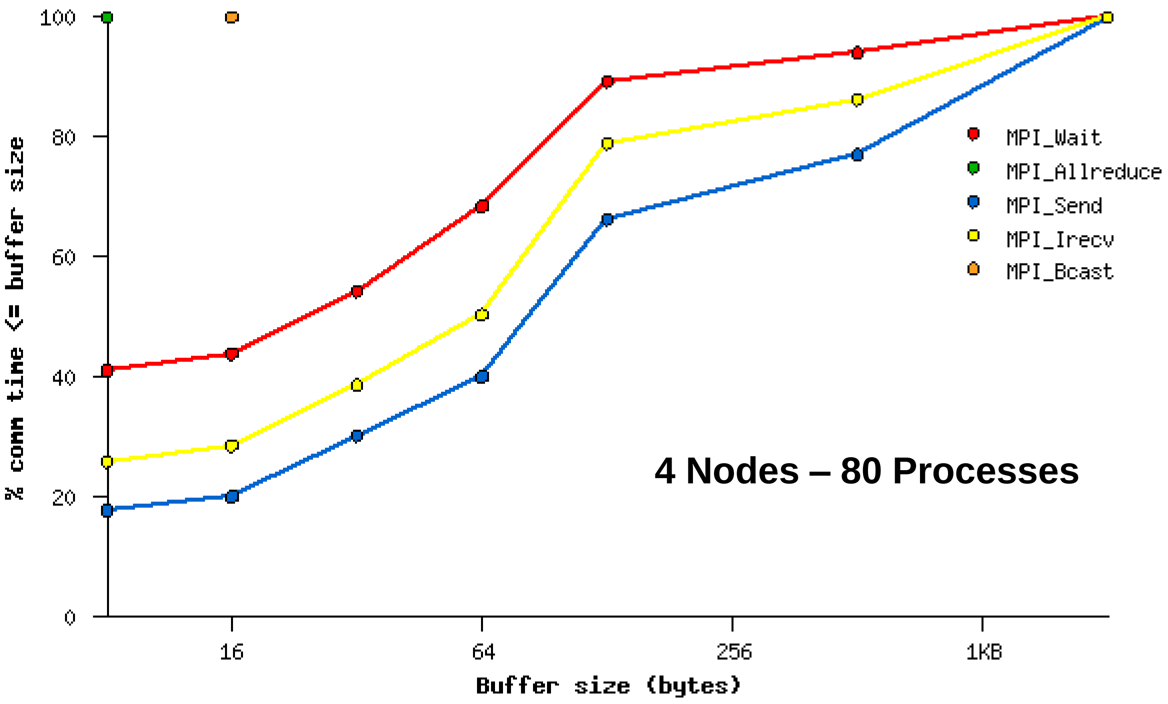
4 Nodes – 80 Processes





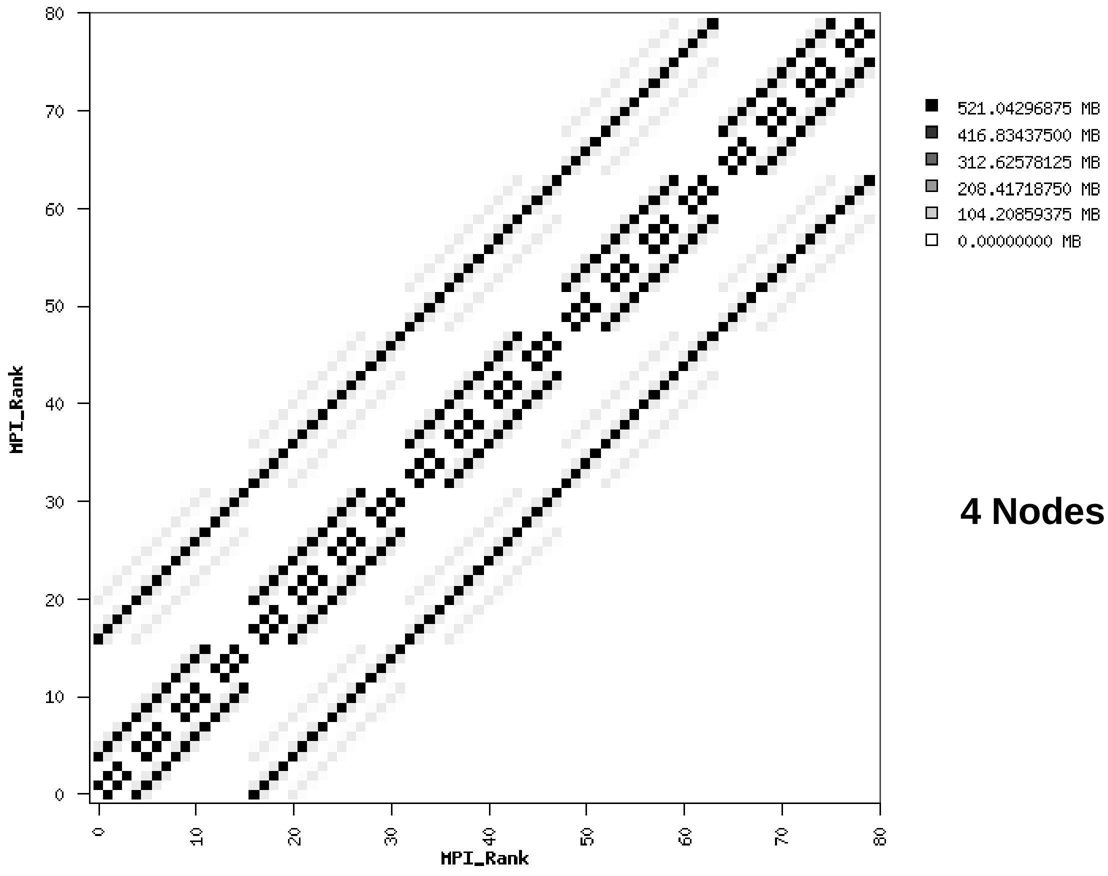
HPCG Profiling – MPI Calls

- **Little variety of MPI calls with limited message sizes were made**
  - Calls are concentrated at these 7 sizes:
  - 0B, 8B, 16B, 32B, 64B, 128B, 512B, 2KB
- **All messages are seen at these quantized sizes**



HPCG Profiling – MPI Data Flow

- **Data transfers between MPI processes the mixed**
  - Up to 521MB between ranks are seen

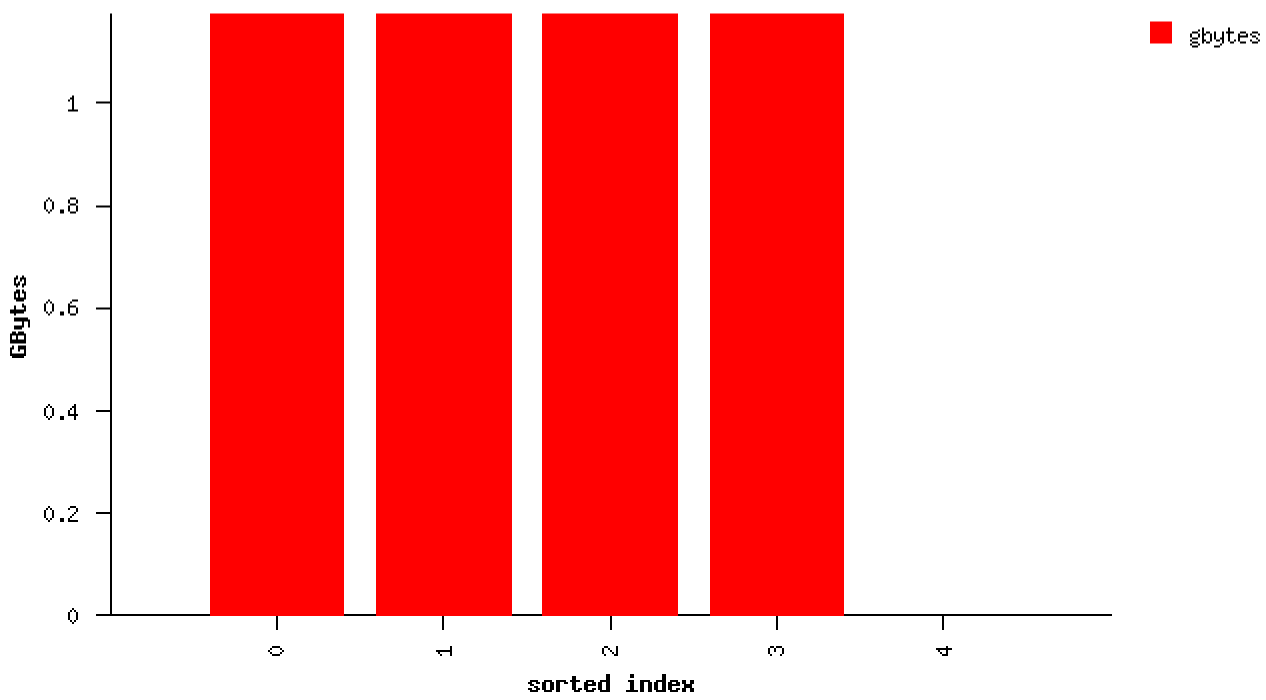


4 Nodes – 80 Processes

HPCG Profiling – Memory Usage By Node

- **The memory usage shown the memory consumption by the compute node**
  - Using the 16x16x16 of input data size, about 1GB of memory is being used by each node

4 Nodes – 80 Processes



HPCG – Summary

- **Performance**
  - Higher performance can be seen by tuning the input value
    - The 16x16x16 mesh yields ~135% higher performance than the default mesh
  - FDR InfiniBand delivers superior scalability in application performance
    - Outperformed 1GbE and 10GbE by over 5 times and 4.5 times, respectively
  - Athena (based on Intel Xeon E5-2680v2) and FDR IB enable HPCG to scale
    - Up to 234% over the Plutus cluster based on Intel Xeon X5670 (Westmere)
  - Tuning compiler with AVX instructions set shows little gain over the default
  - No difference between different MPI implementation
    - Reflect that the 2 MPI implementations handle the MPI calls used in HPCG efficiently
  - No difference in performance by adjusting the runtime duration
- **Profiling**
  - Limited variety of MPI calls and different message sizes were seen
    - MPI calls are MPI_Allreduce, and MPI_Send at certain quantized sizes

Thank You

HPC Advisory Council