HPC ADVISORY COUNCIL

CAM-SE Performance Benchmark and Profiling

May 2014

hp

Mellanox TECHNOLOGIES

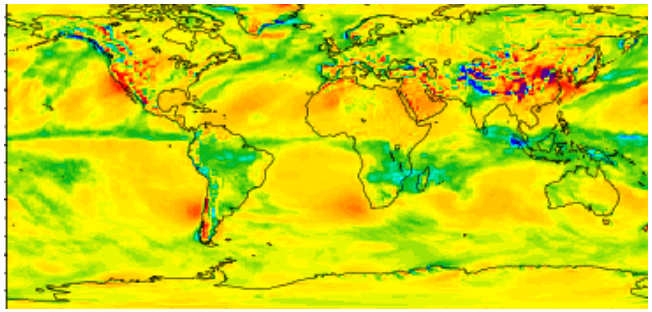
Note

- **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://asc.llnl.gov/CORAL-benchmarks/#camse

- **CAM-SE**
  - Stands for: Community Atmosphere Model – Spectral Element
  - Widely used by climate scientists
    - As the default atmospheric model in Community Earth System Model (CESM)
    - For climate projections in Inter-governmental Panel on Climate Change (IPCC)
  - Comprised of a dynamic core and a physics package
  - Dynamic core
    - Is called HOMME (High-Order Methods Modeling Environment)
    - Solves for wind, energy and mass
  - Models the stratified, compressible, hydrostatic Euler equations on the sphere with the added multi-scale physics representing climate-related processes.
  - Parallelized in MPI and hybrid OpenMP



Objectives

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

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

Test Cluster Configuration

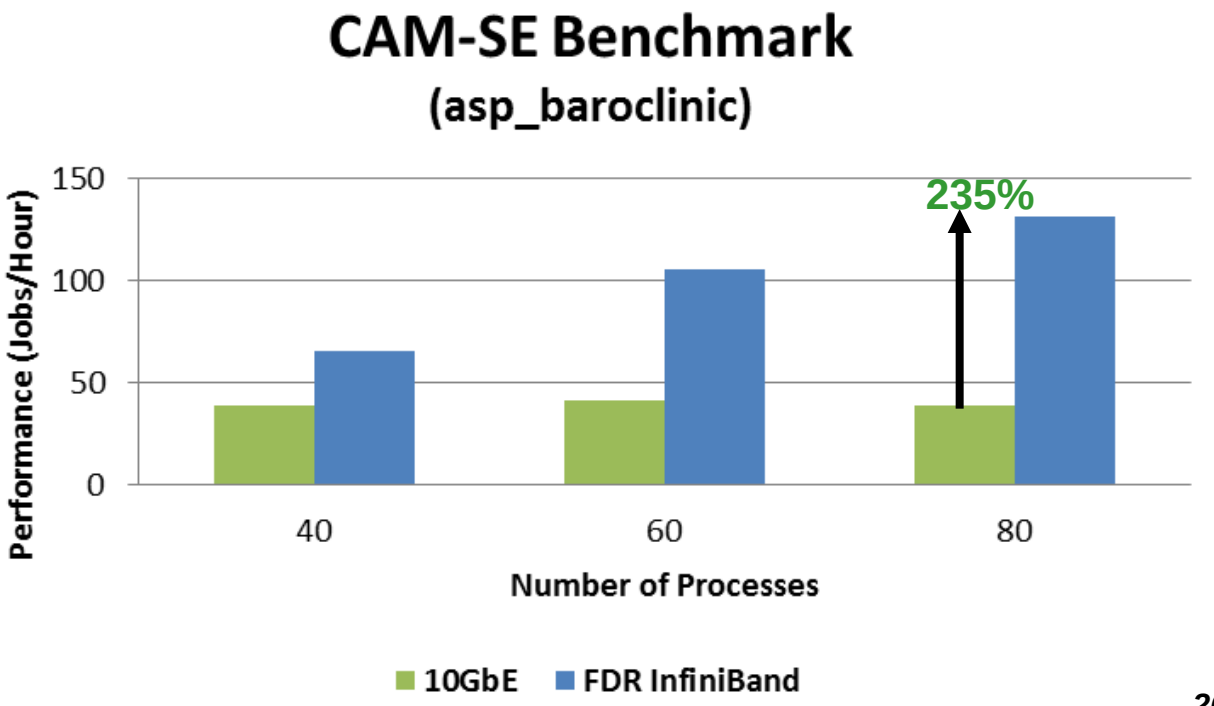
- **HP ProLiant SL230s Gen8 4-node "Athena" cluster**
  - Processors: Dual-Socket 10-core Intel Xeon E5-2680v2 @ 2.8 GHz CPUs
  - Memory: 32GB per node, 1600MHz DDR3 Dual-Ranked DIMMs
  - OS: RHEL 6 Update 2, OFED 2.1-1.0.6 InfiniBand SW stack
- **Mellanox Connect-IB FDR InfiniBand adapters**
- **Mellanox ConnectX-3 VPI Ethernet adapters**
- **Mellanox SwitchX SX6036 56Gb/s FDR InfiniBand and Ethernet VPI Switch**
- **MPI: Open MPI 1.6.5**
- **Compiler: GNU Compilers, NetCDF 4.1.3**
- **Application: HOMME 1.3.6**
- **Benchmark Workload:**
  - Asp_baroclinic, 1 simulated day

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

CAM-SE Performance - Interconnect

- **FDR InfiniBand is the most efficient inter-node communication**
  - Outperforms 10GbE by 235% at 80 MPI processes
  - Performance benefit of InfiniBand expects to grow at larger CPU core counts

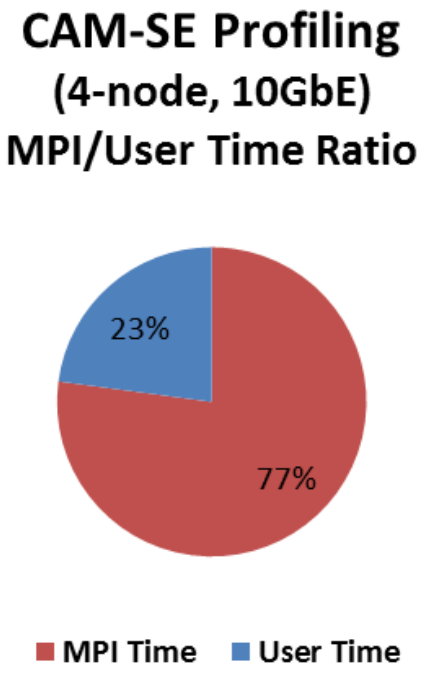


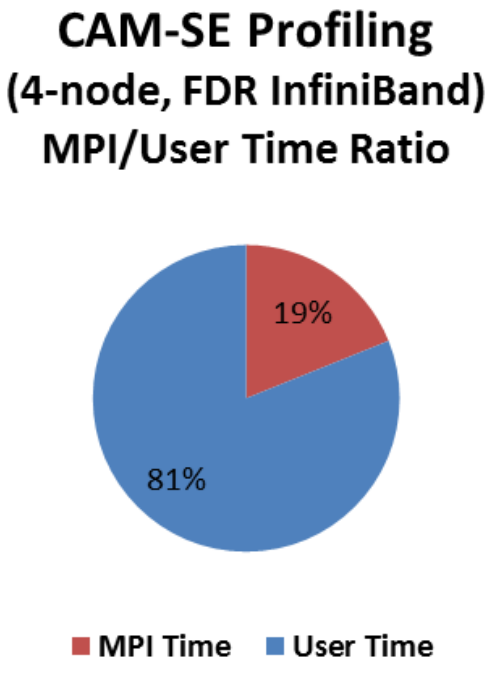
Higher is better

20 Processes/Node

CAM-SE Profiling – MPI Time Ratio

- **FDR InfiniBand reduces the communication time at scale**
  - FDR InfiniBand consumes about 19% of total runtime
  - 10GbE consumes about 77% of total runtime

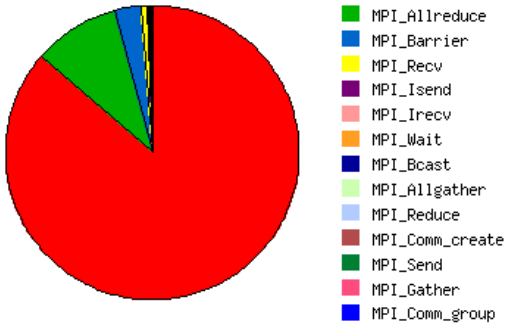




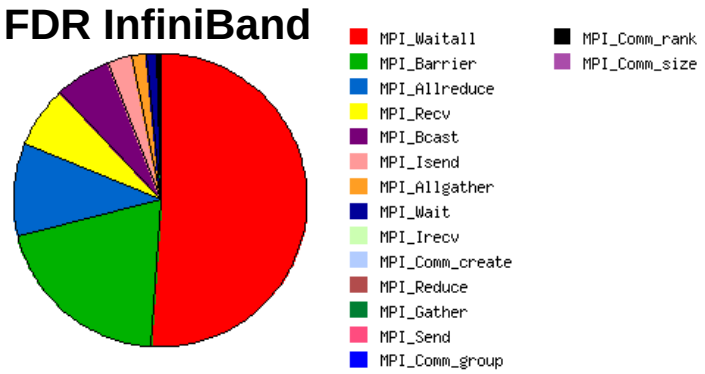
20 Processes/Node

CAM-SE Profiling – MPI Functions

10GbE



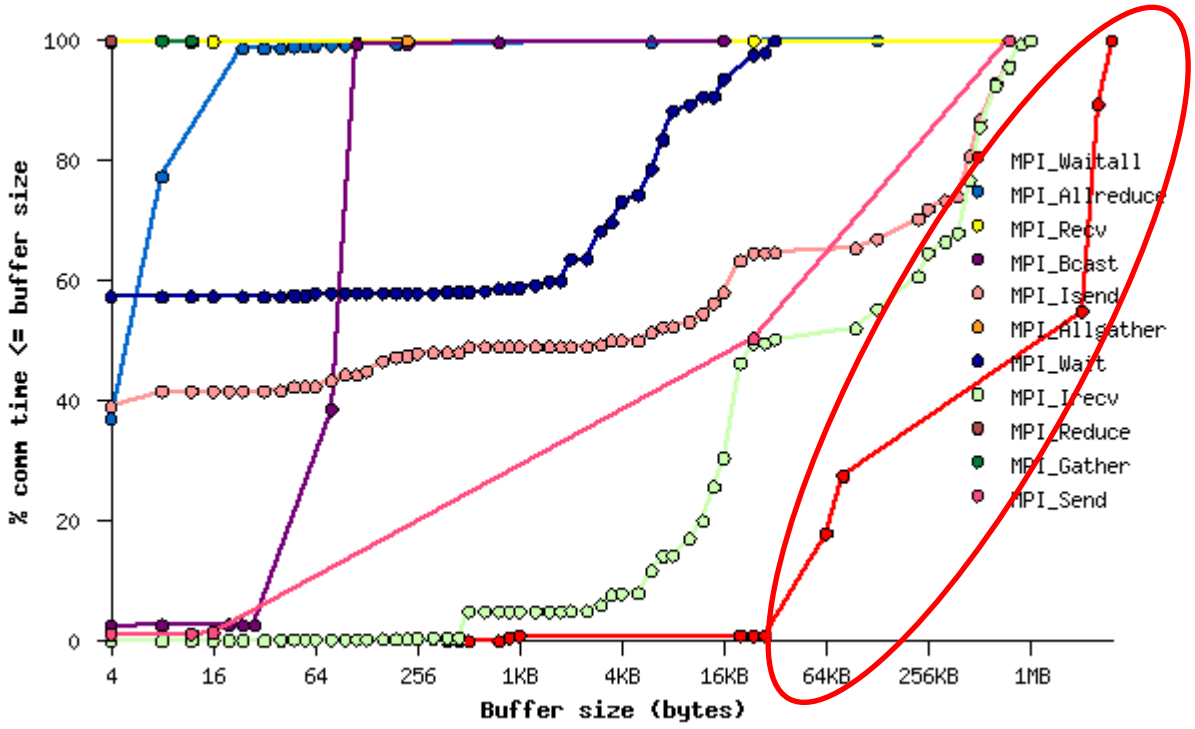
FDR InfiniBand



- **The most time consuming MPI functions:**
  - 10GbE:
    - MPI_Waitall (63% Wall),
    - MPI_Allreduce (8% Wall),
    - MPI_Barrier (3%)
  - FDR InfiniBand:
    - MPI_Waitall (9% Wall),
    - MPI_Barrier (3%),
    - MPI_Allreduce (1%)

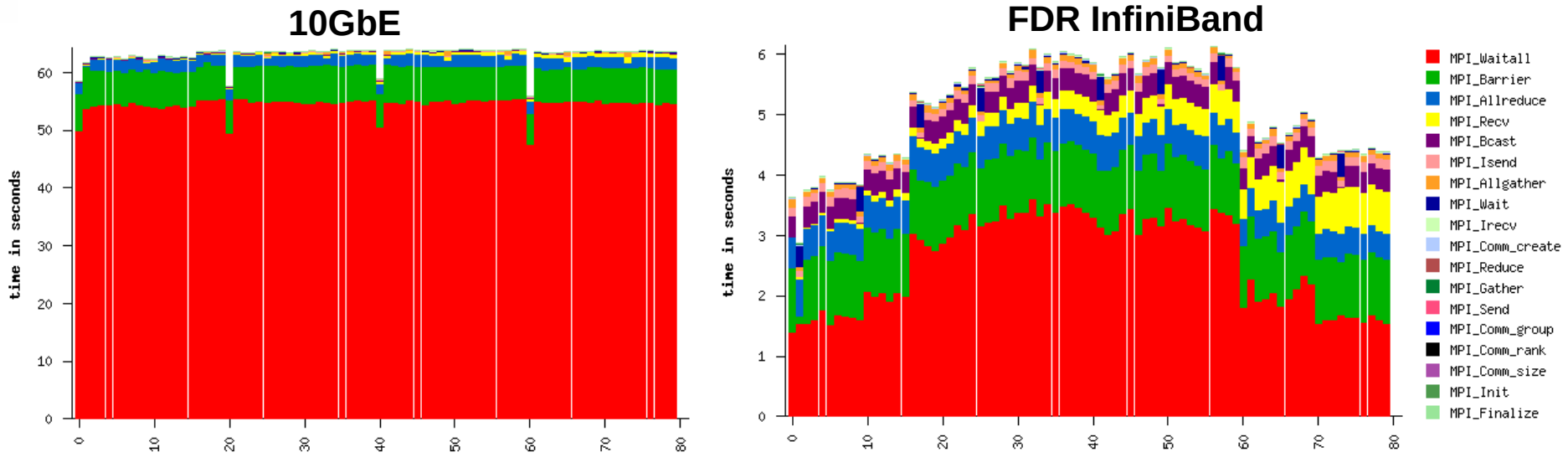
CAM-SE Profiling – MPI Message Sizes

- **The most communication time consumer is MPI_Waitall**
  - Concentrated at 64KB, 81KB, 2MB and 3MB
  - These are for non-blocking send and receive messages
- **The 2nd largest time consumer is the MPI_Barrier**
  - Concentrated at 0B



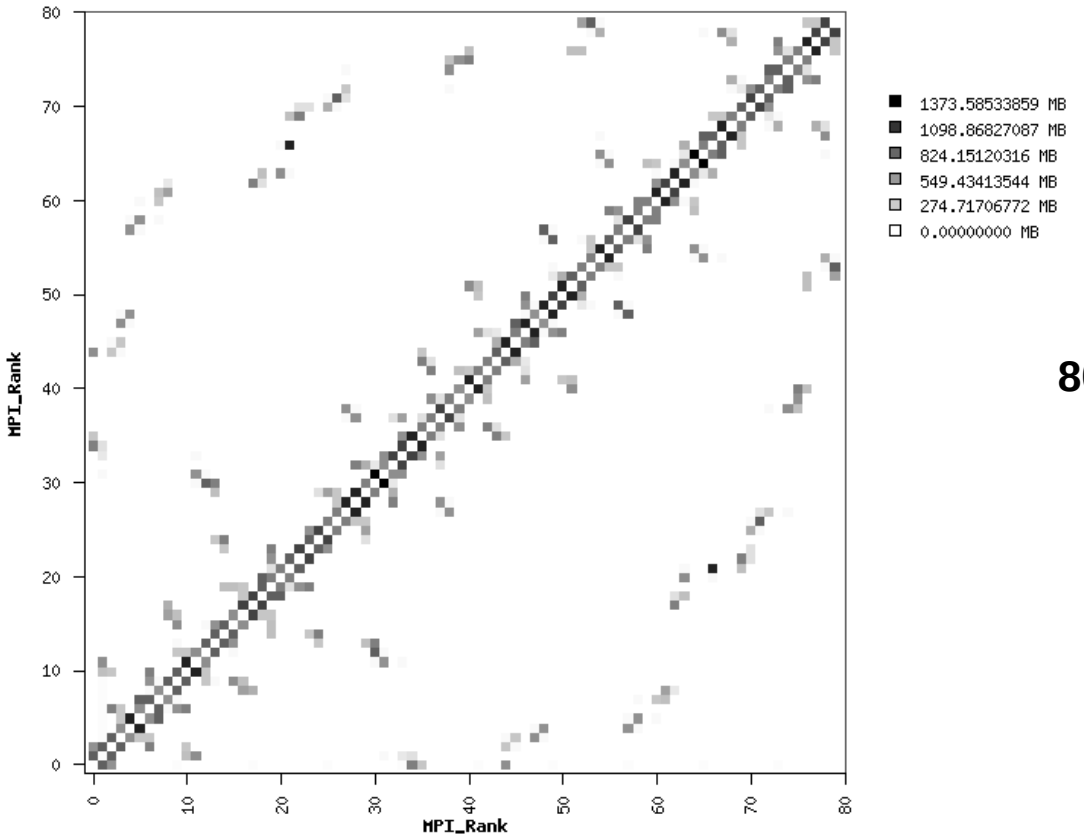
CAM-SE Profiling – MPI Time Ratio

- **FDR InfiniBand reduces MPI communication time**
  - FDR IB reduces most on MPI_Waitall communications compared to 10GbE



CAM-SE Profiling – Comm. Data Flow

- **The point to point data flow shows the communication pattern of CAM-SE**
  - mainly communicates mainly its neighbors and close ranks
  - The pattern stays the same as the cluster scales



80 Processes

FDR InfiniBand

CAM-SE Summary

- **Performance of CAM-SE is directly affected by the network latency**
- **FDR InfiniBand delivers the best network communication for CAM-SE**
  - Outperforms 10GbE up to 234% at 4 nodes (or 80 MPI processes)
- **MPI Profiling**
  - FDR InfiniBand reduces communication time; leave more time for computation
    - FDR InfiniBand consumes 19% of total time
    - Versus 77% against 10GbE
  - MPI_Wait is the most time-consumed operation
  - Blocking communications are seen:
    - Time spent: MPI_Wait (66% Wall Time)

Thank You

HPC Advisory Council