HPC ADVISORY COUNCIL

OCTOPUS
Performance Benchmark and Profiling

June 2015

Mellanox TECHNOLOGIES

octopus

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.tddft.org/programs/octopus

Octopus

- **Octopus is designed for**
  - Density-functional theory (DFT)
  - Time-dependent density functional theory (TDDFT)
- **Octopus is aimed at the simulation of the electron-ion dynamics of 1, 2, 3, and 4 dimensional finite systems**
- **Octopus is one of selected 22 applications for the PRACE application benchmark suite**
- **Octopus is a freely available (GPL) software**

Objectives

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

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

Test Cluster Configuration

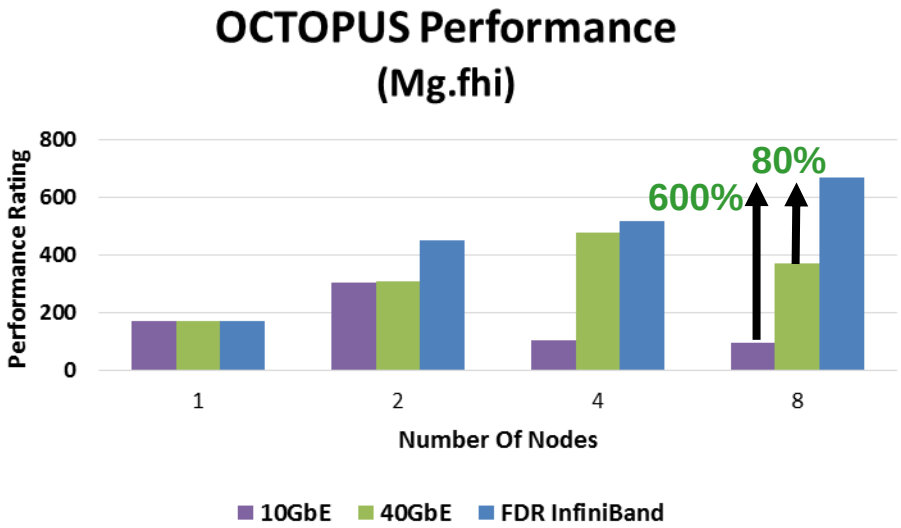
- **HP Apollo 6000 “Heimdall” cluster**
  - HP ProLiant XL230a Gen9 10-node “Heimdall” cluster
  - Processors: Dual-Socket 14-core Intel Xeon E5-2697v3 @ 2.6 GHz CPUs (Turbo Mode off; Home Snoop as default)
  - Memory: 64GB per node, 2133MHz DDR4 DIMMs
  - OS: RHEL 6 Update 6, OFED 2.4-1.0.0 InfiniBand SW stack
- **Mellanox Connect-IB FDR InfiniBand adapters**
- **Mellanox ConnectX-3 Pro Ethernet adapters**
- **Mellanox SwitchX SX6036 56Gb/s FDR InfiniBand and Ethernet VPI Switch**
- **MPI: Intel MPI 5.0.2**
- **Compiler and Libraries: Intel Composers and MKL 2015.1.133, FFTW 3.3.4, GSL 1.16, libxc 2.0.3, pfft 1.0.8 alpha**
- **Application: OCTOPUS 4.1.2**
- **Benchmark Workload: Magnesium (Ground State, maximum iteration set to 1)**

HP ProLiant XL230a Gen9 Server

Item	HP ProLiant XL230a Gen9 Server
Processor	Two Intel® Xeon® E5-2600 v3 Series, 6/8/10/12/14/16 Cores
Chipset	Intel Xeon E5-2600 v3 series
Memory	512 GB (16 x 32 GB) 16 DIMM slots, DDR3 up to DDR4; R-DIMM/LR-DIMM; 2,133 MHz
Max Memory	512 GB
Internal Storage	1 HP Dynamic Smart Array B140i SATA controller HP H240 Host Bus Adapter
Networking	Network module supporting various FlexibleLOMs: 1GbE, 10GbE, and/or InfiniBand
Expansion Slots	1 Internal PCIe: 1 PCIe x 16 Gen3, half-height
Ports	Front: (1) Management, (2) 1GbE, (1) Serial, (1) S.U.V port, (2) PCIe, and Internal Micro SD card & Active Health
Power Supplies	HP 2,400 or 2,650 W Platinum hot-plug power supplies delivered by HP Apollo 6000 Power Shelf
Integrated Management	HP iLO (Firmware: HP iLO 4) Option: HP 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	10 servers in 5U chassis

OCTOPUS Performance - Interconnects

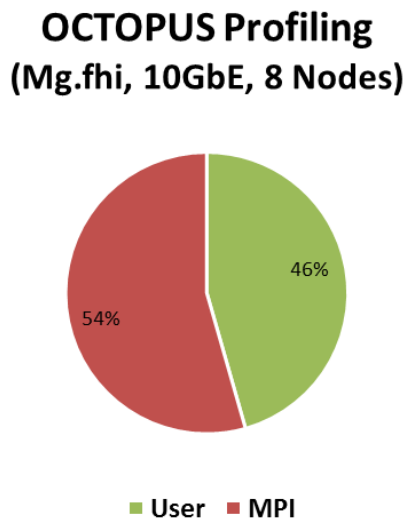
- **FDR InfiniBand is the most efficient network interconnect for OCTOPUS**
  - FDR IB outperforms 10GbE by 600% at 8 nodes (224 MPI processes)
  - FDR IB outperforms 40GbE by 80% at 8 nodes (224 MPI processes)

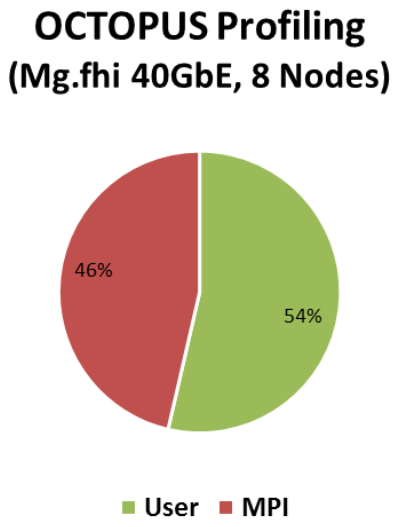


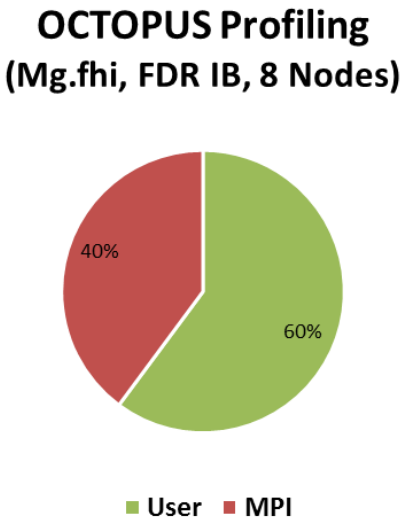
Higher is better

OCTOPUS Profiling – Interconnects

- **FDR InfiniBand reduces the communication time at scale**
  - FDR InfiniBand consumes about 40% of total runtime at 8 nodes (224 processes)
  - 10GbE consumes 54% of total time in communications, while 40GbE consumes 46%



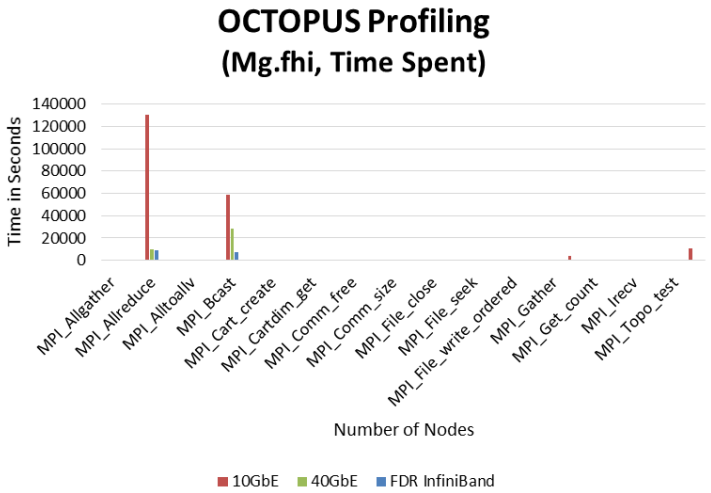




28 Processes Per Node

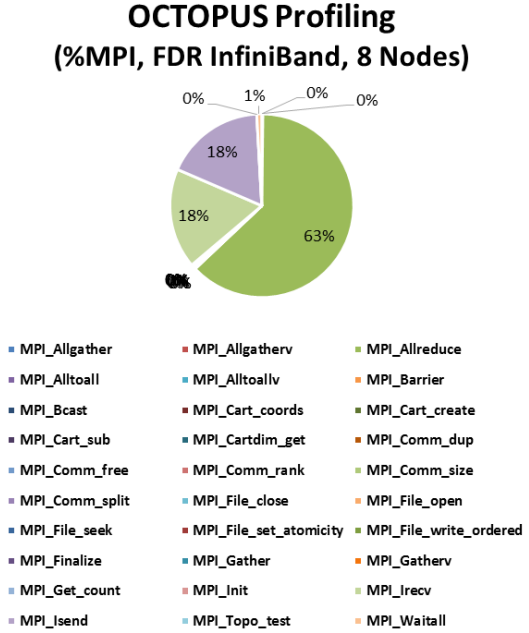
OCTOPUS Profiling – Time Spent in MPI

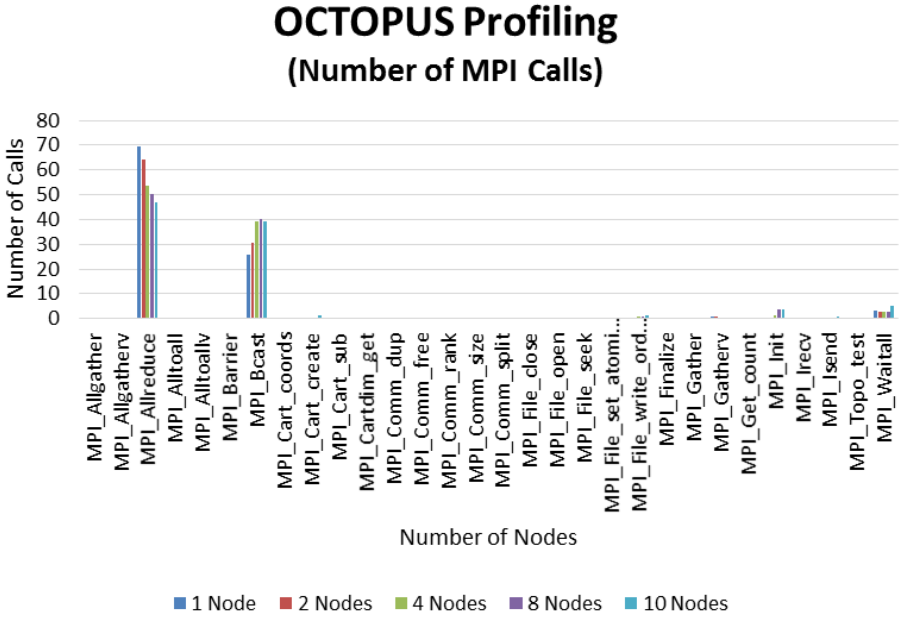
- **The most time consuming MPI functions for OCTOPUS:**
  - MPI_Reduce (50%), MPI_Bcast (40%) among all MPI calls
- **Time spent on network bandwidth differentiates among interconnects**
  - 10GbE/40GbE spent more time in MPI_Reduce/MPI_Bcast
  - Demonstrated that InfiniBand performs better than Ethernet networks for MPI collectives ops



OCTOPUS Profiling – Number of MPI Calls

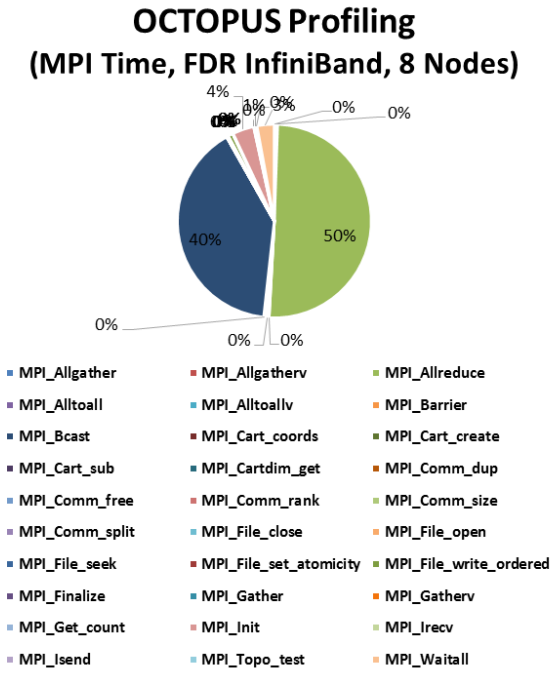
- **MPI calls for collective communication is dominated in OCTOPUS**
  - MPI_Allreduce (63% of calls)
  - MPI_Irecv, MPI_Isend (18%) on a 8 node (224 processes)

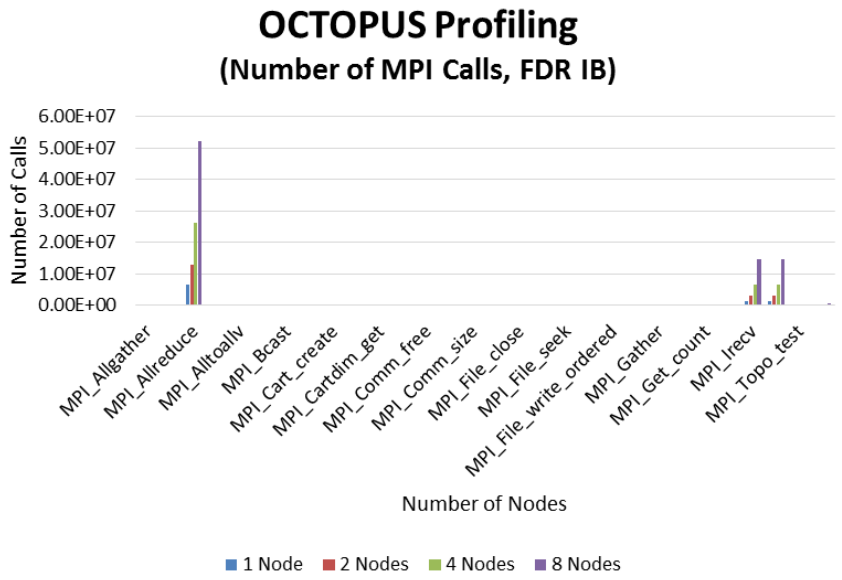




OCTOPUS Profiling – Time Spent in MPI

- **OCTOPUS: More time spent on MPI collective operations:**
  - MPI_Allreduce(50%), MPI_Bcast(40%)





OCTOPUS Summary

- **InfiniBand FDR is the most efficient cluster interconnect for OCTOPUS**
  - FDR IB outperforms 10GbE by 600% at 8 nodes (224 MPI processes)
  - FDR IB outperforms 40GbE by 80% at 8 nodes (224 MPI processes)
- **FDR InfiniBand reduces the communication time at scale**
  - FDR InfiniBand consumes about 40% of total runtime at 8 nodes (224 processes)
  - 10GbE consumes 54% of total time in communications, while 40GbE consumes 46%
- **Octopus MPI profiling**
  - MPI collectives create big communication overhead
  - Both large and small message are used by Octopus
  - Interconnect latency and bandwidth are critical to Octopus performance

Thank You

HPC Advisory Council