HPC
ADVISORY COUNCIL

OpenFOAM
Performance Benchmark and Profiling

April 2013

DELL
intel
Mellanox TECHNOLOGIES
OpenFOAM

Note

- **The following research was performed under the HPC Advisory Council activities**
  - Participating vendors: Intel, Dell, Mellanox
  - Compute resource - HPC Advisory Council Cluster Center

- **The following was done to provide best practices**
  - OpenFOAM performance overview
  - Understanding OpenFOAM communication patterns
  - Ways to increase OpenFOAM productivity
  - MPI libraries comparisons

- **For more info please refer to**
  - http://www.dell.com
  - http://www.intel.com
  - http://www.mellanox.com
  - http://www.openfoam.org

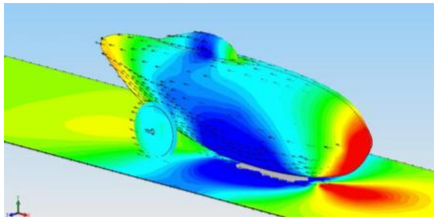
Objectives

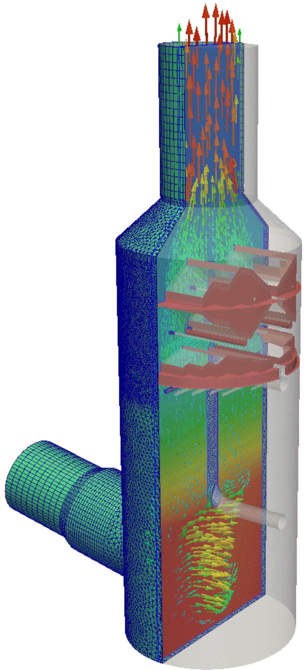
- **The following was done to provide best practices**
  - OpenFOAM performance benchmarking
  - Interconnect performance comparisons
  - MPI performance comparison
  - Understanding OpenFOAM communication patterns

- **The presented results will demonstrate**
  - The scalability of the compute environment to provide nearly linear application scalability
  - The capability of OpenFOAM to achieve scalable productivity

OpenFOAM Applications

- **OpenFOAM® (Open Field Operation and Manipulation) CFD Toolbox in an open source CFD applications that can simulate**
  - Complex fluid flows involving
    - Chemical reactions
    - Turbulence
    - Heat transfer
  - Solid dynamics
  - Electromagnetics
  - The pricing of financial options





- **OpenFOAM support can be obtained from OpenCFD Ltd**

Test Cluster Configuration

- **Dell™ PowerEdge™ R720xd 16-node (256-core) “Jupiter” cluster**
  - Dual-Socket Eight-Core Intel E5-2680 @ 2.70 GHz CPUs (Static max Perf in BIOS)
  - Memory: 64GB memory, DDR3 1600 MHz
  - OS: RHEL 6.2, OFED 1.5.3 InfiniBand SW stack
  - Hard Drives: 24x 250GB 7.2 RPM SATA 2.5” on RAID 0
- **Intel Cluster Ready certified cluster**
- **Mellanox ConnectX-3 FDR InfiniBand VPI adapters**
- **Mellanox SwitchX SX6036 InfiniBand switch**
- **MPI: Intel MPI 4 Update 3, Open MPI 1.6.2**
- **Application: OpenFOAM 2.1.0**
- **Benchmark datasets:**
  - Lid Driven Cavity Flow - 1 Million elements, 2D, icoFoam solver for laminar, isothermal, incompressible flow

About Intel® Cluster Ready

- **Intel® Cluster Ready systems make it practical to use a cluster to increase your simulation and modeling productivity**
  - Simplifies selection, deployment, and operation of a cluster
- **A single architecture platform supported by many OEMs, ISVs, cluster provisioning vendors, and interconnect providers**
  - Focus on your work productivity, spend less management time on the cluster
- **Select Intel Cluster Ready**
  - Where the cluster is delivered ready to run
  - Hardware and software are integrated and configured together
  - Applications are registered, validating execution on the Intel Cluster Ready architecture
  - Includes Intel® Cluster Checker tool, to verify functionality and periodically check cluster health

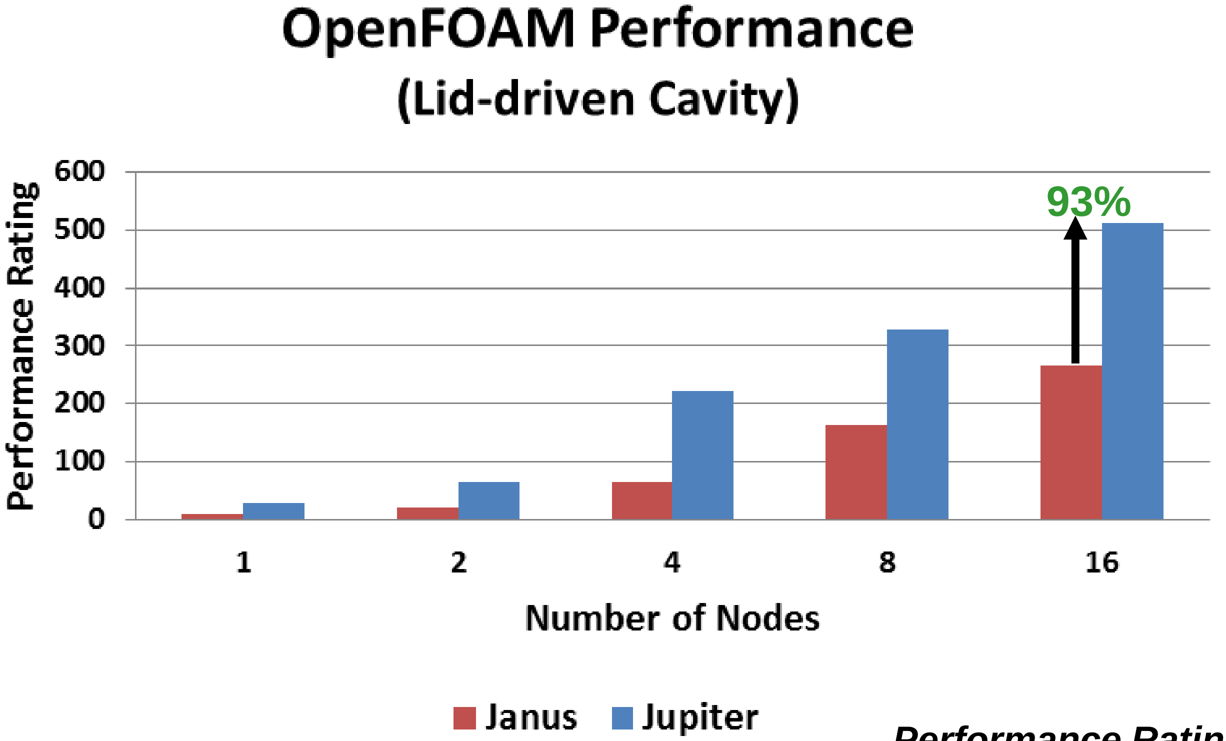
PowerEdge R720xd

Massive flexibility for data intensive operations

- **Performance and efficiency**
  - Intelligent hardware-driven systems management with extensive power management features
  - Innovative tools including automation for parts replacement and lifecycle manageability
  - Broad choice of networking technologies from GigE to IB
  - Built in redundancy with hot plug and swappable PSU, HDDs and fans
- **Benefits**
  - Designed for performance workloads
    - from big data analytics, distributed storage or distributed computing where local storage is key to classic HPC and large scale hosting environments
    - High performance scale-out compute and low cost dense storage in one package
- **Hardware Capabilities**
  - Flexible compute platform with dense storage capacity
    - 2S/2U server, 6 PCIe slots
  - Large memory footprint (Up to 768GB / 24 DIMMs)
  - High I/O performance and optional storage configurations
    - HDD options: 12 x 3.5" - or - 24 x 2.5 + 2x 2.5 HDDs in rear of server
    - Up to 26 HDDs with 2 hot plug drives in rear of server for boot or scratch

OpenFOAM Performance – Processors

- **Intel E5-2680 (Sandy Bridge) cluster outperforms prior generations**
  - Performs 93% better than X5670 cluster at 16 nodes
- **System components used:**
  - Jupiter: 2-socket Intel E5-2680 @ 2.7GHz, 1600MHz DIMMs, FDR IB, 24 disks
  - Janus: 2-socket Intel X5670 @ 2.93GHz, 1333MHz DIMMs, QDR IB, 1 disk

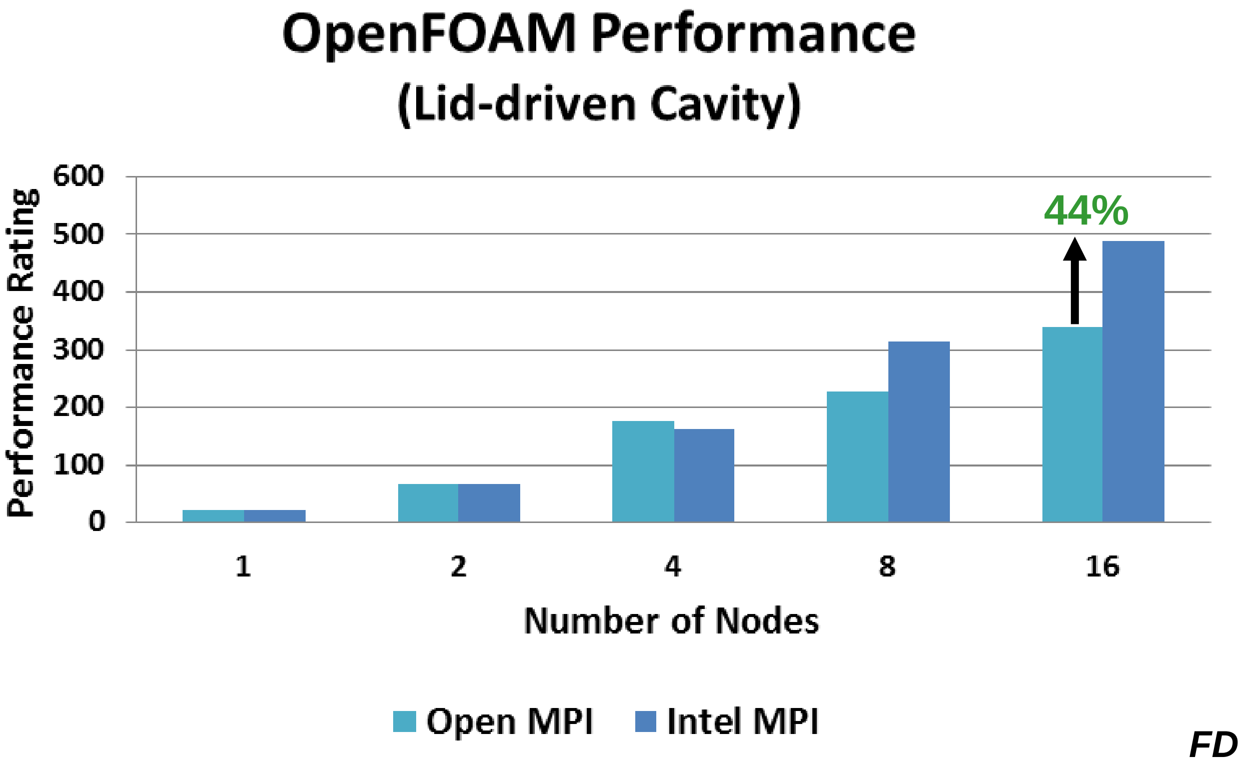


Higher is better

Performance Rating = Jobs/Day

OpenFOAM Performance – MPI

- **Intel MPI outperforms Open MPI at larger scale**
  - Up to 44% higher performance than Open MPI at 16-node
- **CPU binding optimization flag used in all cases shown**
  - No other optimization flags are used

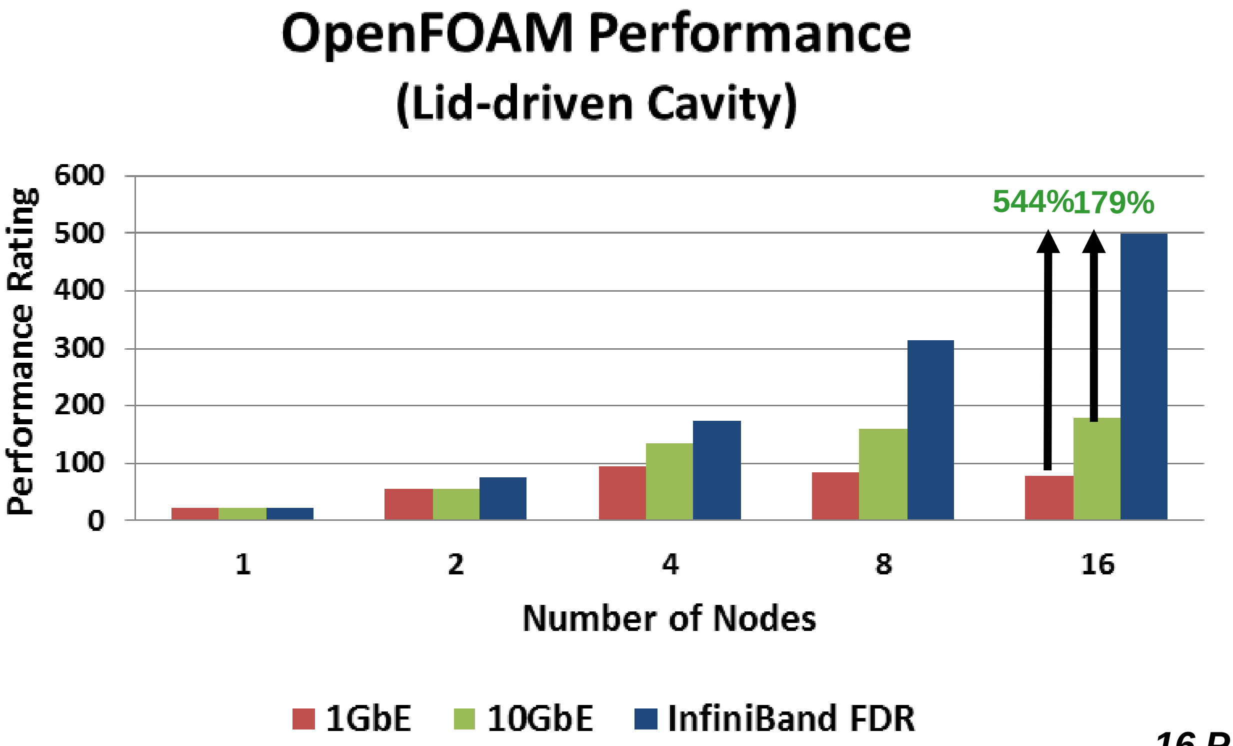


Higher is better

FDR InfiniBand

OpenFOAM Performance – Interconnects

- **FDR InfiniBand provides better scalability performance than Ethernet**
  - 544% better performance than 10GbE at 16 nodes / 256 processes
  - 179% better performance than 1GbE at 16 nodes / 256 processes
  - 1GbE does not scale at all

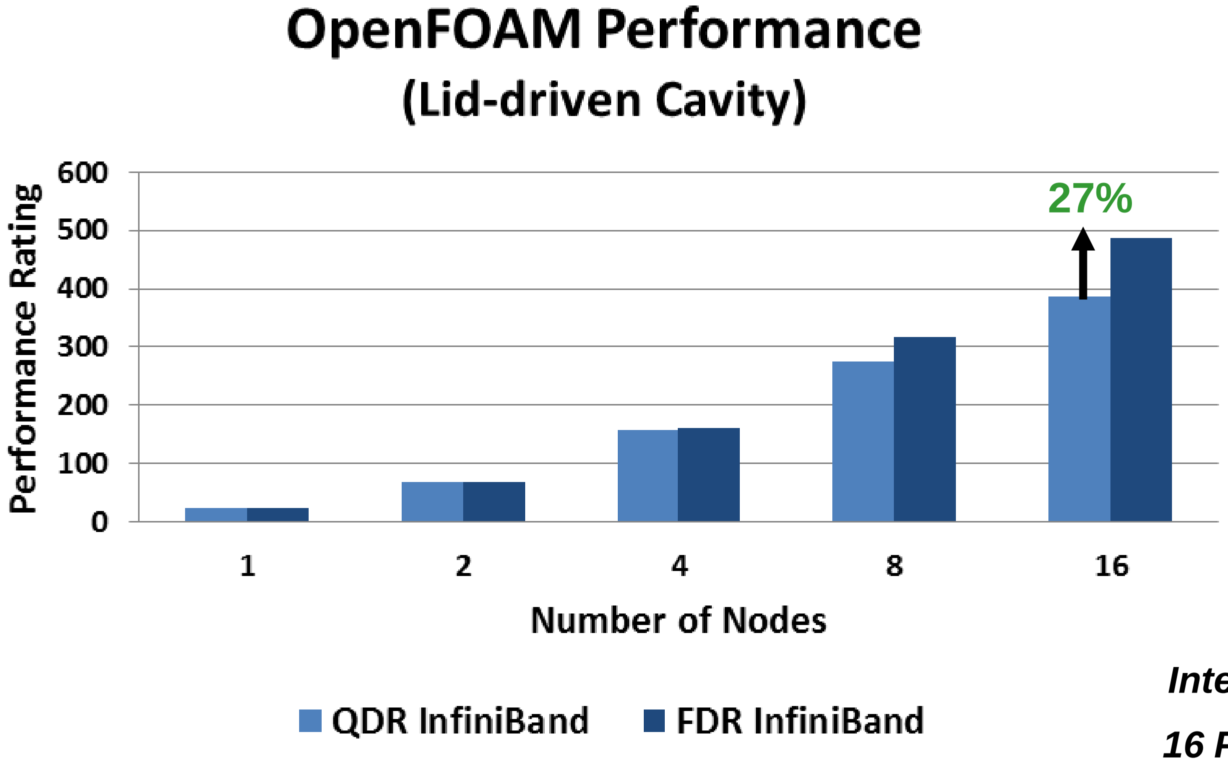


Higher is better

16 Processes/Node

OpenFOAM Performance – Interconnects

- **FDR InfiniBand delivers better application performance**
  - Up to 27% better performance than InfiniBand QDR
  - Using Mellanox ConnectX-3 PCIe Gen3 in FDR mode and QDR mode



Higher is better

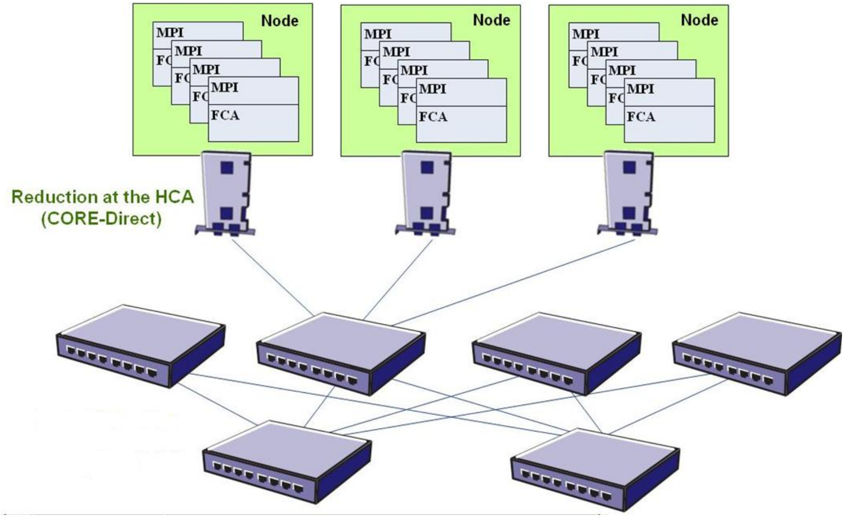
Intel MPI

16 Processes/Node

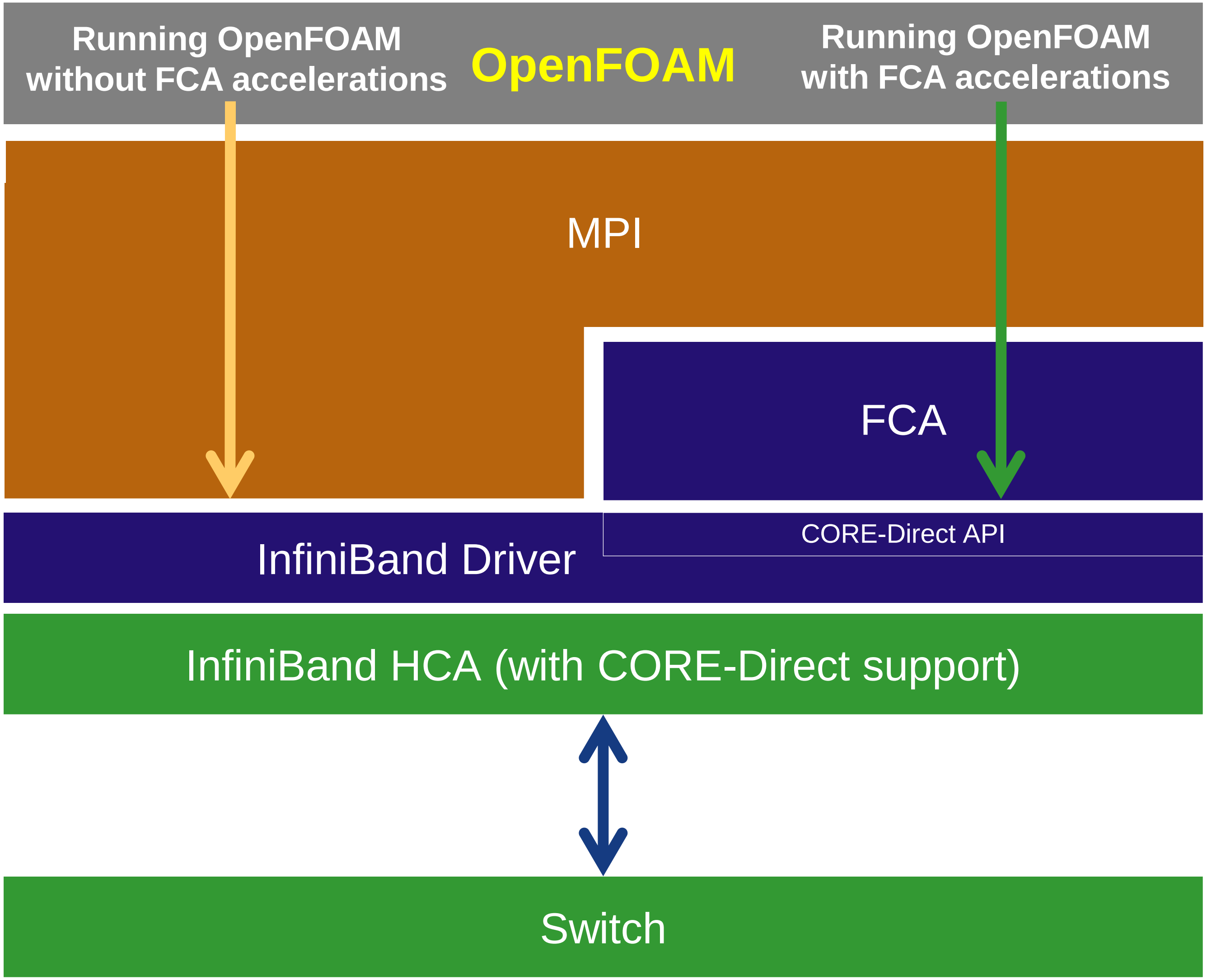
About Mellanox FCA

- **Mellanox Fabric Collectives Accelerator (FCA)**
  - Utilized hardware accelerations on the adapter (CORE-Direct)
  - Accelerating MPI collectives operations by offloading them to the network
  - The world first complete solution for MPI collectives offloads

- **FCA 2.2 supports accelerations/offloading for**
  - MPI_Barrier
  - MPI_Broadcast
  - MPI_Allreduce and MPI_Reduce
  - MPI_Allgather and MPI_Allgatherv

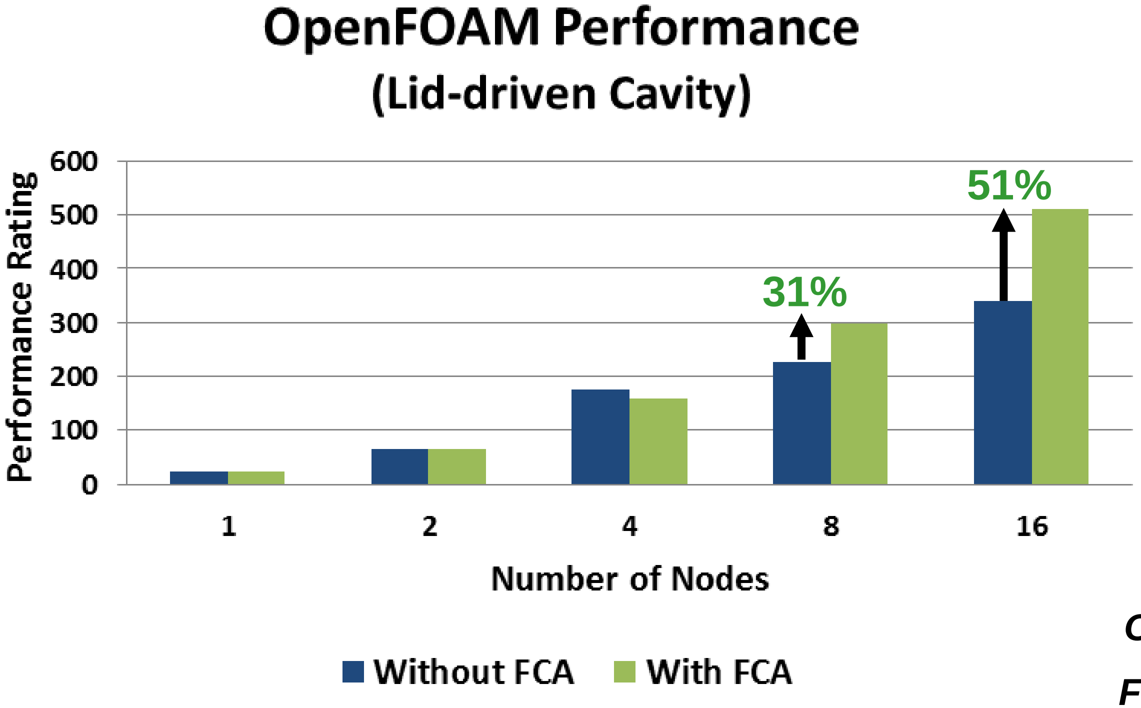


Software Layers Overview


Running OpenFOAM without FCA accelerations
OpenFOAM
Running OpenFOAM with FCA accelerations
MPI
FCA
InfiniBand Driver
CORE-Direct API
InfiniBand HCA (with CORE-Direct support)
Switch

OpenFOAM Performance – FCA

- **FCA enables nearly 51% performance gain at 16 nodes / 256 cores**
  - Bigger advantage expected at higher node count / core count
  - Normally FCA is enabled for >64 cores; FCA is enabled for all processes shown below
- **Flags used:**
  - To enable FCA at runtime: --mca coll_fca_enable 1 --mca coll_fca_np 0
  - Both cases at runtime: --bind-to-core -mca btl openib,sm,self



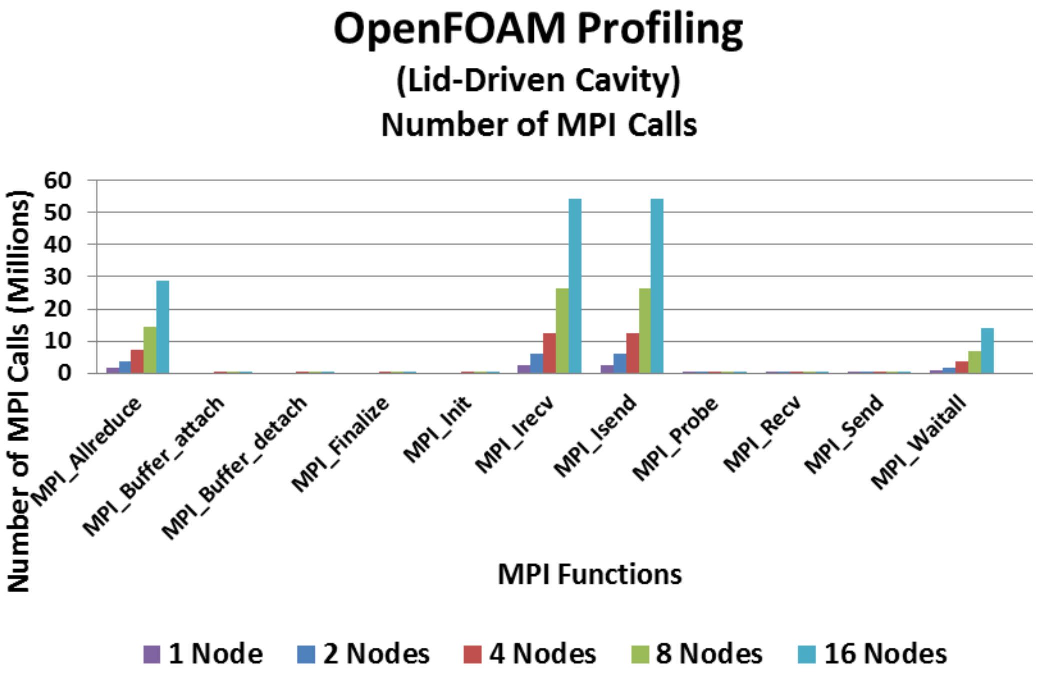
Higher is better

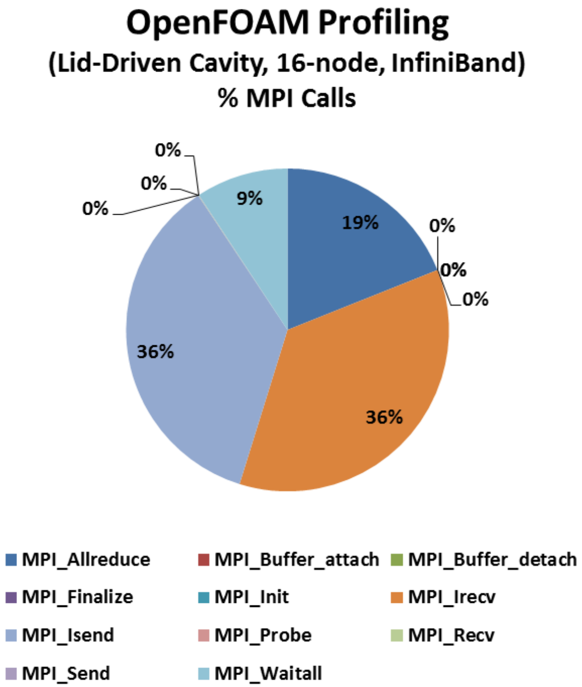
Open MPI

FDR InfiniBand

OpenFOAM Profiling – Number of MPI Calls

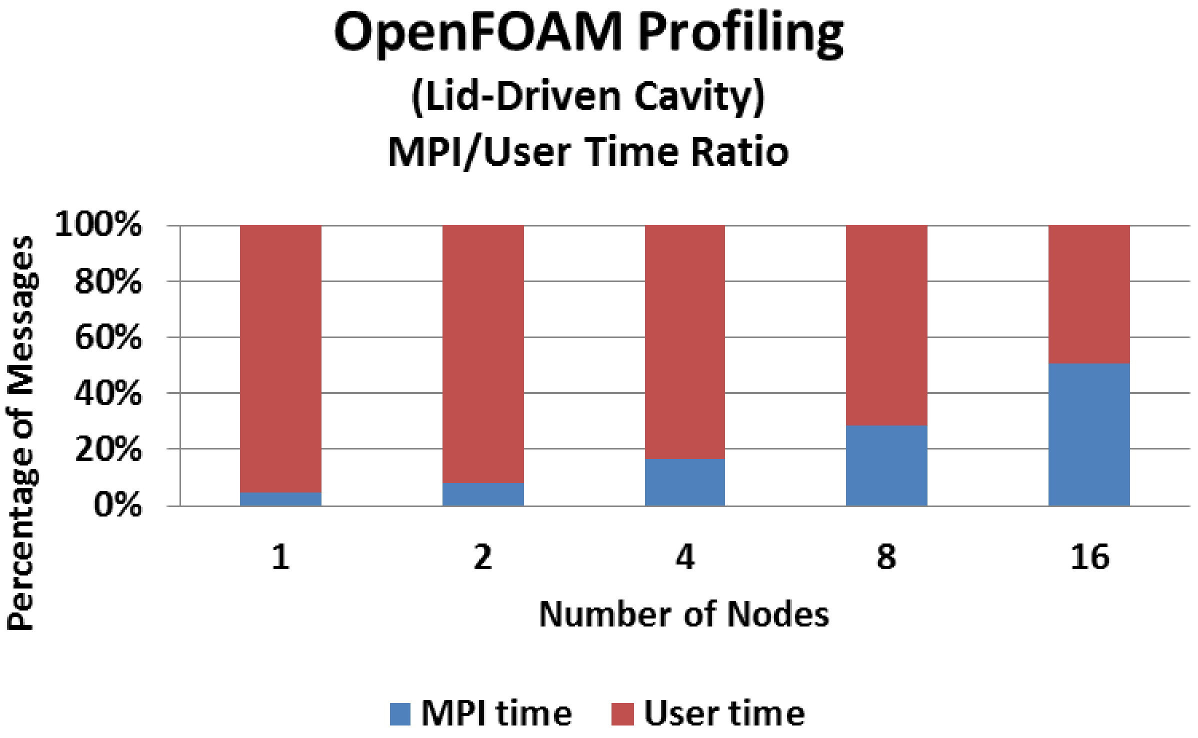
- **OpenFOAM utilizes a wide range of MPI APIs**
  - – 11 MPI APIs used in total
  - – 4 MPI APIs account for almost all of MPI calls
- **MPI_Waitall, MPI_Irecv and MPI_Isend are almost used exclusively**
  - – MPI_Irecv, MPI_Isend (26% each), MPI_Alltoallv (19%) at 16 nodes





OpenFOAM MPI Profiling Results – Time Ratio

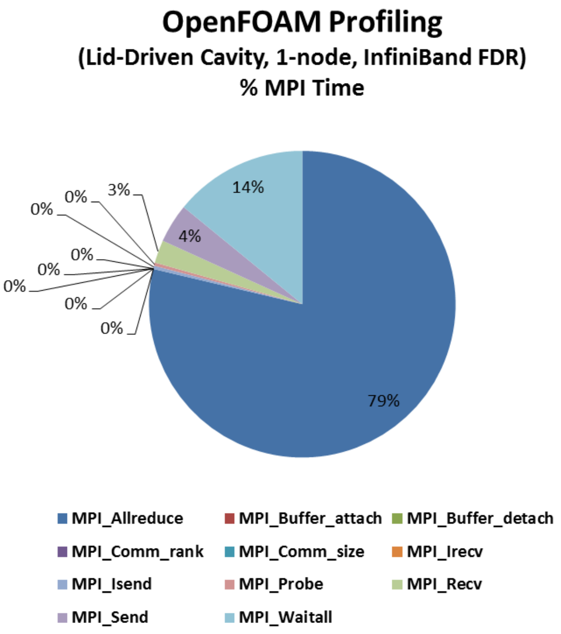
- **MPI communication time accounts for 50%**
  - With 16 nodes / 256 cores
  - The Lid-driven cavity flow is a highly communicative workload

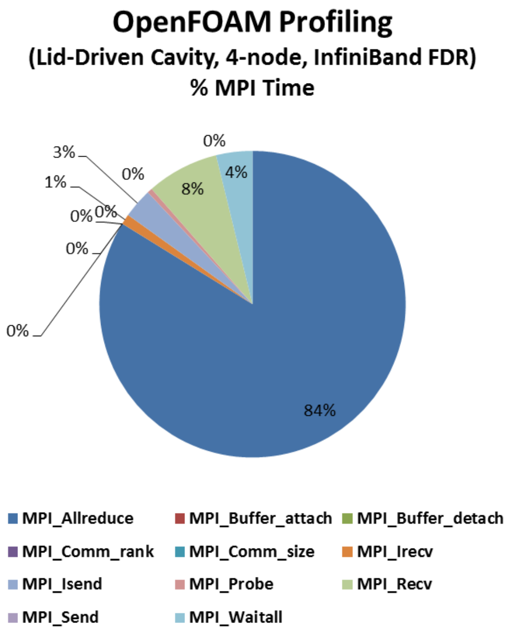


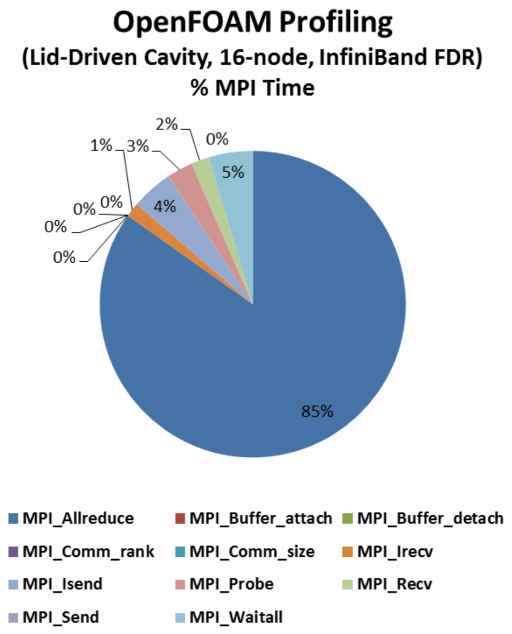
FDR InfiniBand

OpenFOAM Profiling – % MPI Time

- **MPI profiling clearly shows large time usage in MPI collective operations**
  - MPI_Allreduce accounts for 79% to 85% of all MPI time
- **Tuning MPI libraries for MPI collective offloading related to collective operations**
  - Will greatly influence the system performance



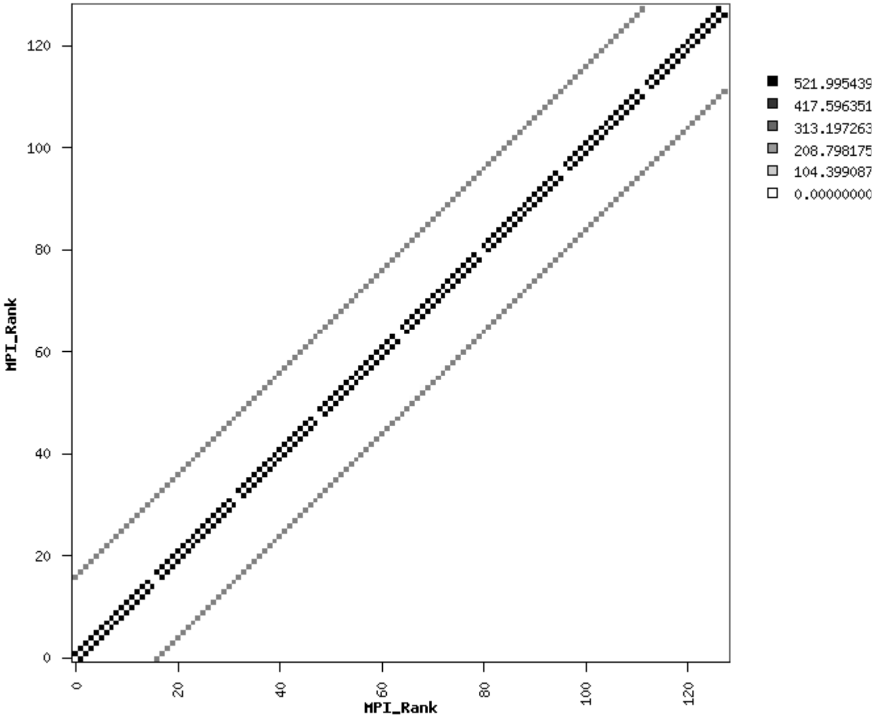




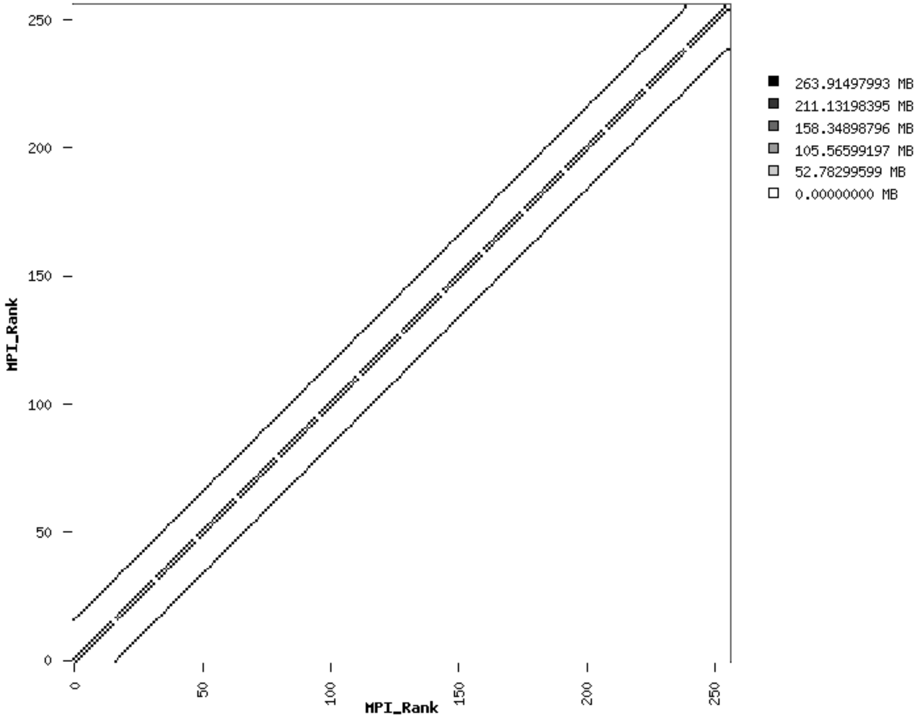
OpenFOAM Profiling – MPI Data Transfer

- **As the cluster grows, less data is transferred between MPI processes**
  - Decrease from 523MB max (8 nodes) at to 263MB max per rank (16 nodes)
  - Majority of communications are between neighboring ranks
  - Non-blocking (point to point) data transfers are shown in the graph
  - Collective data communications are small compared to non-blocking communications

8 Nodes



16 Nodes



OpenFOAM – Summary

- **OpenFOAM performance**
  - Intel Xeon E5-2600 series and FDR InfiniBand enable OpenFOAM to scale with 16 nodes
  - The E5-2680 cluster outperforms X5670 cluster by 93% at 16 nodes
  - Intel MPI scales better than Open MPI at large node counts (16 nodes) by 44%
- **FDR InfiniBand delivers the best application performance for OpenFOAM**
  - Up to 27% higher performance than InfiniBand QDR at 16 nodes
  - Up to 179% higher performance than 10GbE at 16 nodes
  - Up to 544% higher performance than 1GbE at 16 nodes
- **OpenFOAM MPI profiling**
  - Time used by MPI accounts for 50% of total runtime at 16 nodes / 256 processes
  - MPI_Allreduce accounts for 79% to 85% of all MPI time
  - Shows MPI_Allreduce is the main MPI collective routines that impacts OpenFOAM performance
- **FCA package has proven to accelerate application**
  - Nearly 51% faster runtime at 16 nodes / 256 cores for OpenFOAM with Open MPI
  - Higher performance boost expected at larger scale

Thank You

HPC Advisory Council

All trademarks are property of their respective owners. All information is provided "As-Is" without any kind of warranty. The HPC Advisory Council makes no representation to the accuracy and completeness of the information contained herein. HPC Advisory Council Mellanox undertakes no duty and assumes no obligation to update or correct any information presented herein