HPC ADVISORY COUNCIL

FLOW-3D
Performance Benchmark and Profiling

September 2012

DELL

FLOW Science
Improving the world through accurate flow modeling

intel

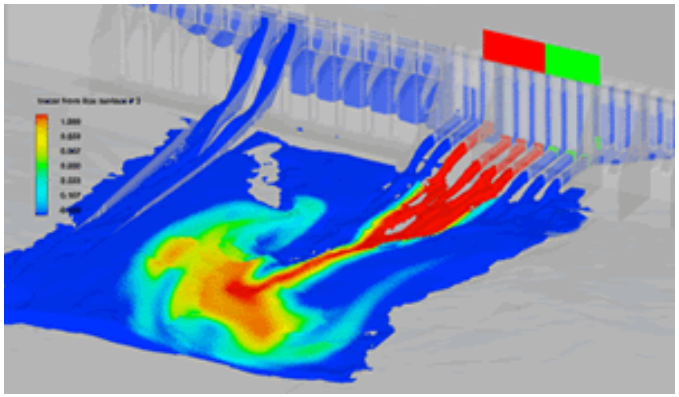
Mellanox
TECHNOLOGIES

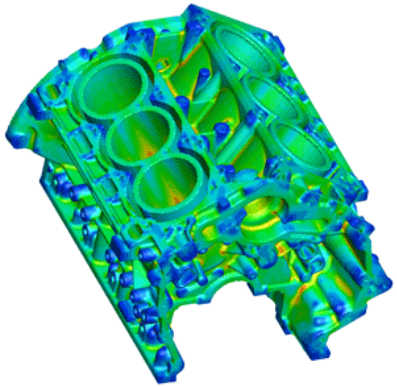
Note

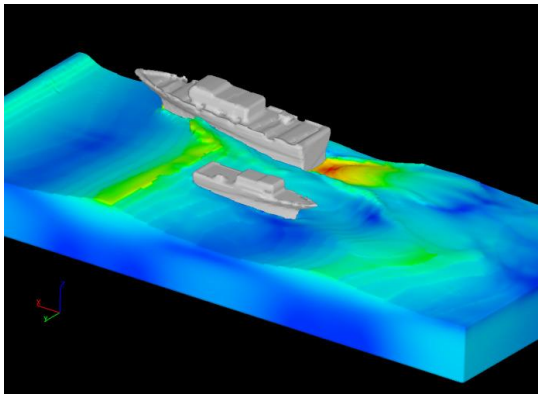
- **The following research was performed under the HPC Advisory Council activities**
  - Participating vendors: FLOW-3D, Dell, Intel, Mellanox
  - Compute resource - HPC Advisory Council Cluster Center
- **For more info please refer to**
  - http://www.dell.com/hpc
  - http://www.flow3d.com
  - http://www.intel.com
  - http://www.mellanox.com

FLOW-3D

- **FLOW-3D is a powerful and highly-accurate CFD software**
  - Provides engineers valuable insight into many physical flow processes
- **FLOW-3D is the ideal computational fluid dynamics software**
  - To use in the design phase as well as in improving production processes
  - Provides special capabilities for accurately predicting free-surface flows
- **FLOW-3D is a standalone, all-inclusive CFD package**
  - Includes an integrated GUI that ties components from problem setup to post-processing







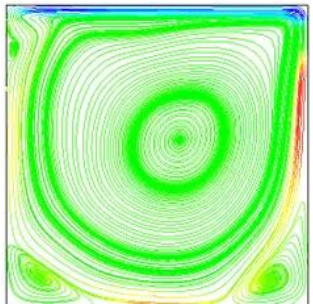
Objectives

- **The following was done to provide best practices**
  - FLOW-3D performance benchmarking
  - Interconnect performance comparisons
  - Understanding FLOW-3D communication patterns
  - Ways to increase FLOW-3D productivity
  - MPI libraries comparisons

- **The presented results will demonstrate**
  - The scalability of the compute environment
  - The capability of FLOW-3D to achieve scalable productivity
  - Considerations for performance optimizations

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 (vendor provided): Intel MPI 3.2.0.011**
- **Application: FLOW-3D MP 4.2**
- **Benchmarks:**
  - Lid Driven Cavity Flow
  - P2 Engine Block



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

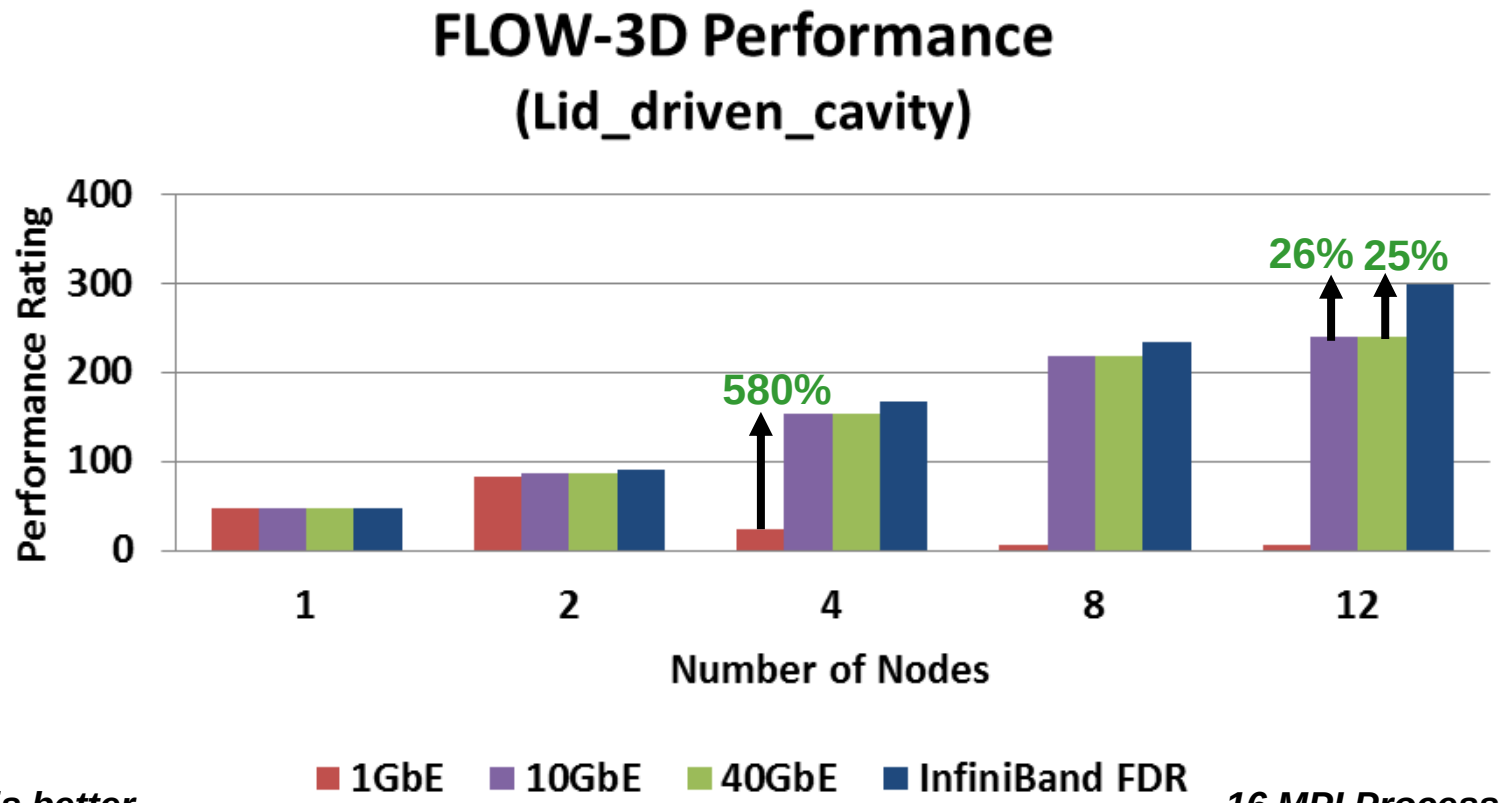
PowerEdge R720xd

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 1GbE 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

FLOW-3D Performance – Network

- **Input dataset: Lid Driven Cavity**
- **InfiniBand FDR provides best scalability performance than Ethernet**
  - Provides up to 580% better performance than 1GbE at 4-node
  - Provides up to 26% better performance than 10GbE at 12-node
  - Provides up to 25% better performance than 40GbE at 12-node

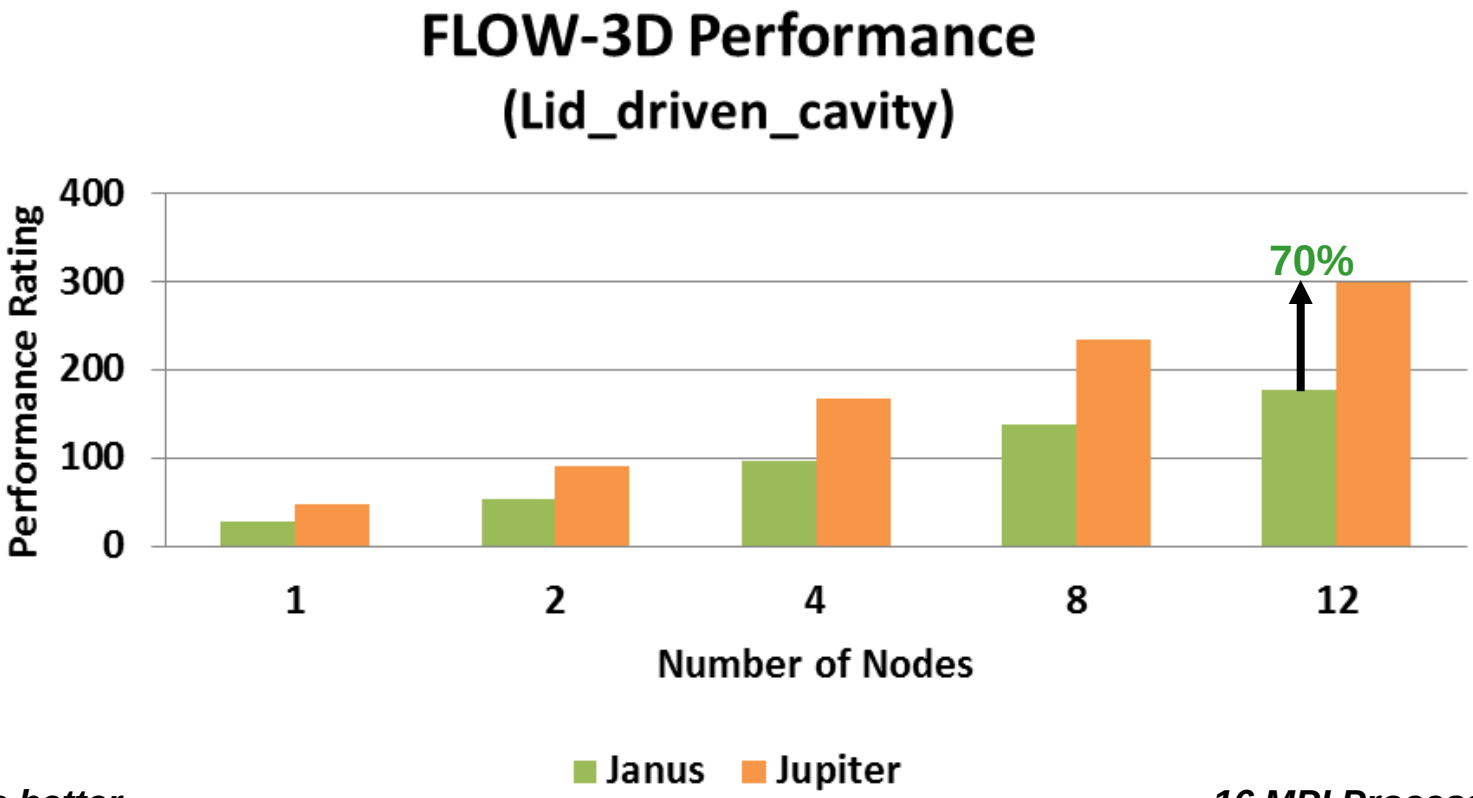


Higher is better

16 MPI Processes/Node

FLOW-3D Performance – Processors

- **Intel E5-2680 (Sandy Bridge) cluster outperforms prior generations**
  - Performs 70% 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

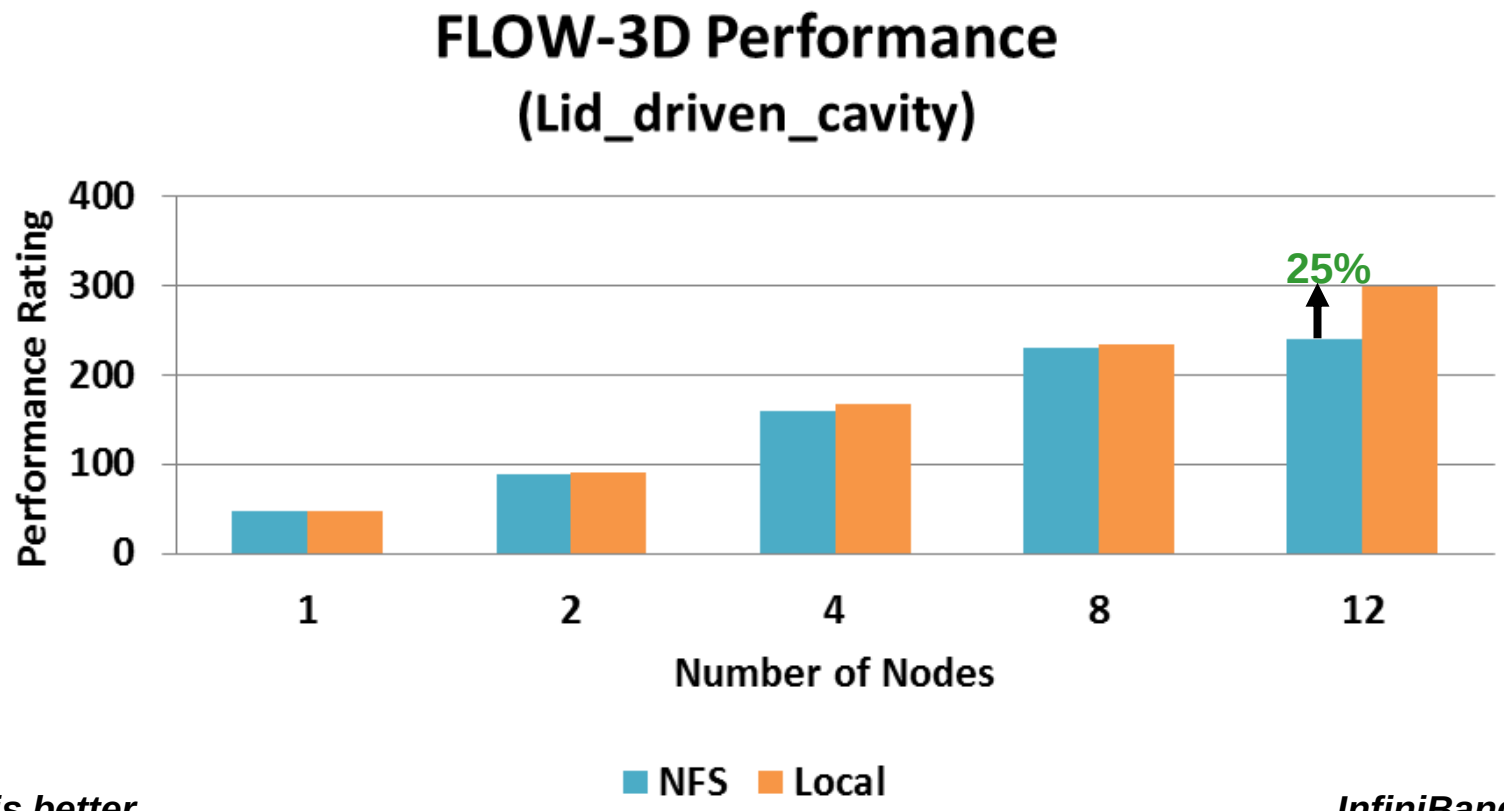


Higher is better

16 MPI Processes/Node

FLOW-3D Performance – Processors

- **Storing data files on local FS or tmpfs would improve performance**
  - Scalability is limited by NFS when running at scale after 8 nodes
  - NFS used in this case is over 1GbE network

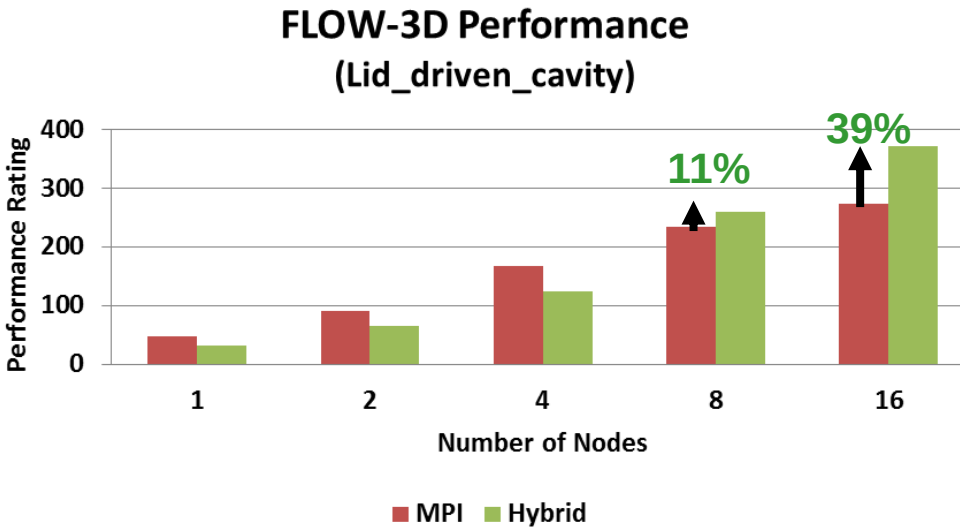


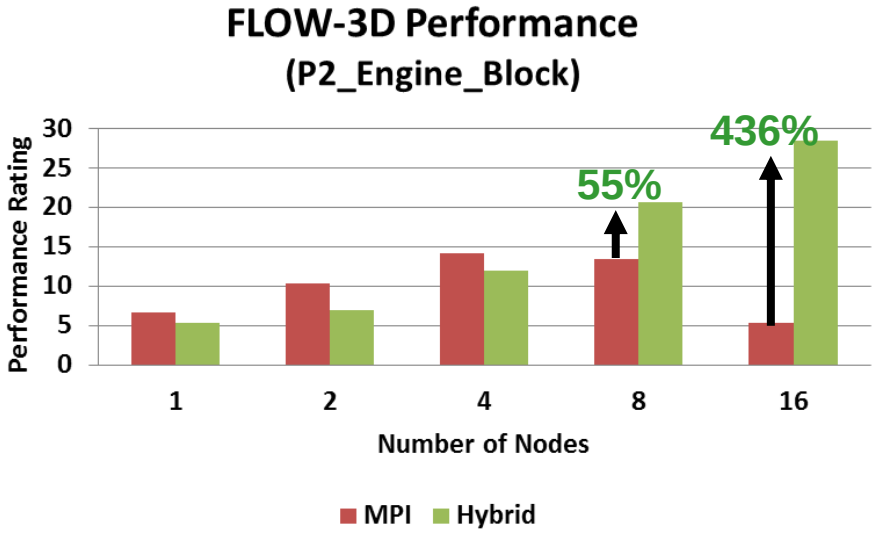
Higher is better

InfiniBand FDR

FLOW-3D Performance – MPI vs Hybrid

- **FLOW-3D/MP supports OpenMP Hybrid mode**
  - Up to 39% of higher performance than MPI at 16 nodes (Lid Driven Cavity)
  - Up to 436% of higher performance than MPI at 16 nodes (P2 Engine Block)
- **Hybrid version enables higher scalability versus pure MPI version**
  - Hybrid version delivers better scalability after 4 nodes
  - MPI processes would spawn OpenMP threads for computation on CPU cores
  - Streamline and reduce communication endpoints to improve scalability



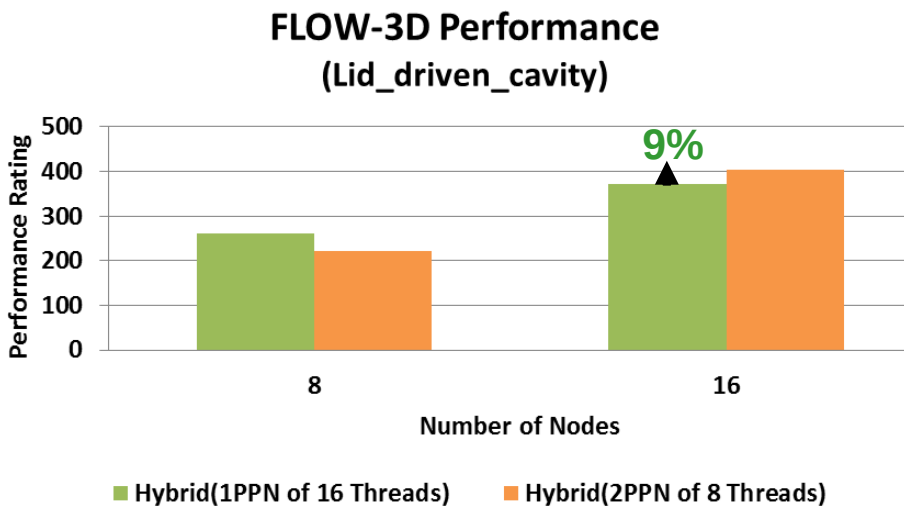


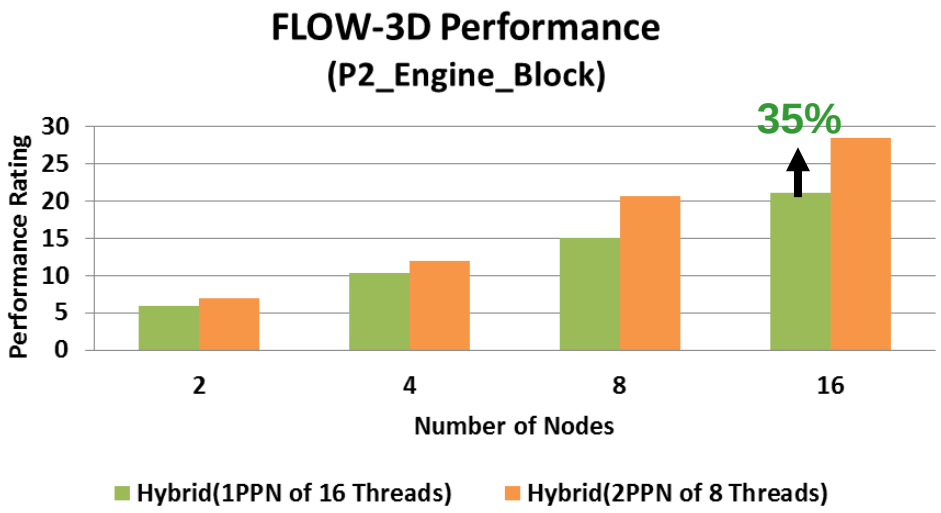
Higher is better

InfiniBand FDR

FLOW-3D Performance – Hybrid (Socket/Node)

- **2PPN of 8 threads can provide better performance than 1PPN of 16 threads**
  - Threads of the MPI process causes threads to spawn within the same socket
  - With the "-genv I_MPI_PIN_DOMAIN **socket**" specified in the **runhyd_par** script
- **Default for FLOW-3D/MP hybrid is to run 1 PPN of 16 threads**
  - With the "-genv I_MPI_PIN_DOMAIN **node**" specified in the **runhyd_par** script
- **The flag is modified to "socket" to allow spawning of threads within a socket**
  - For the case of 2PPN of 8 threads

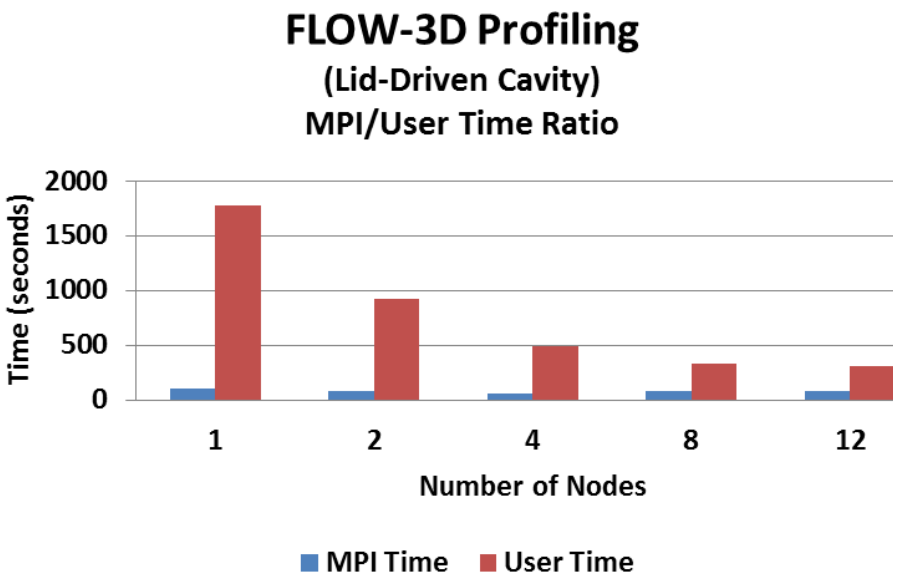


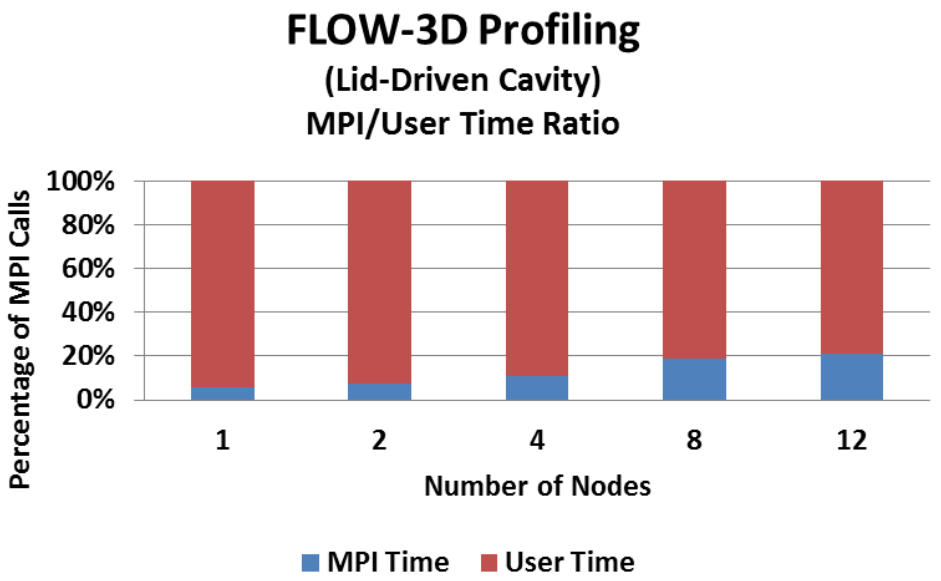


Higher is better ***InfiniBand FDR***

FLOW-3D Profiling – # of MPI Calls

- **The overall runtime reduces as more nodes take part of the MPI job**
  - More compute nodes can reduce the runtime by spreading out the workload
- **Computation time drops while the communication time stays flat**
  - As cluster scales, MPI time stays constantly at the same level



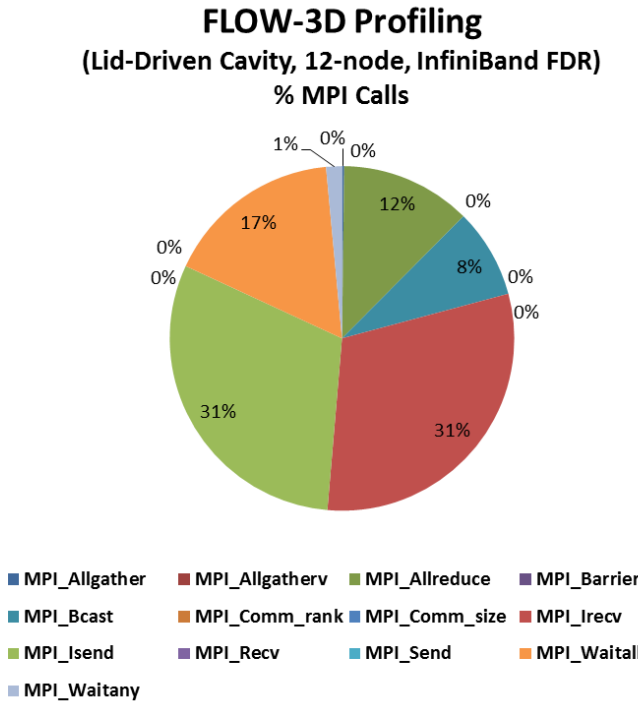


Higher is better

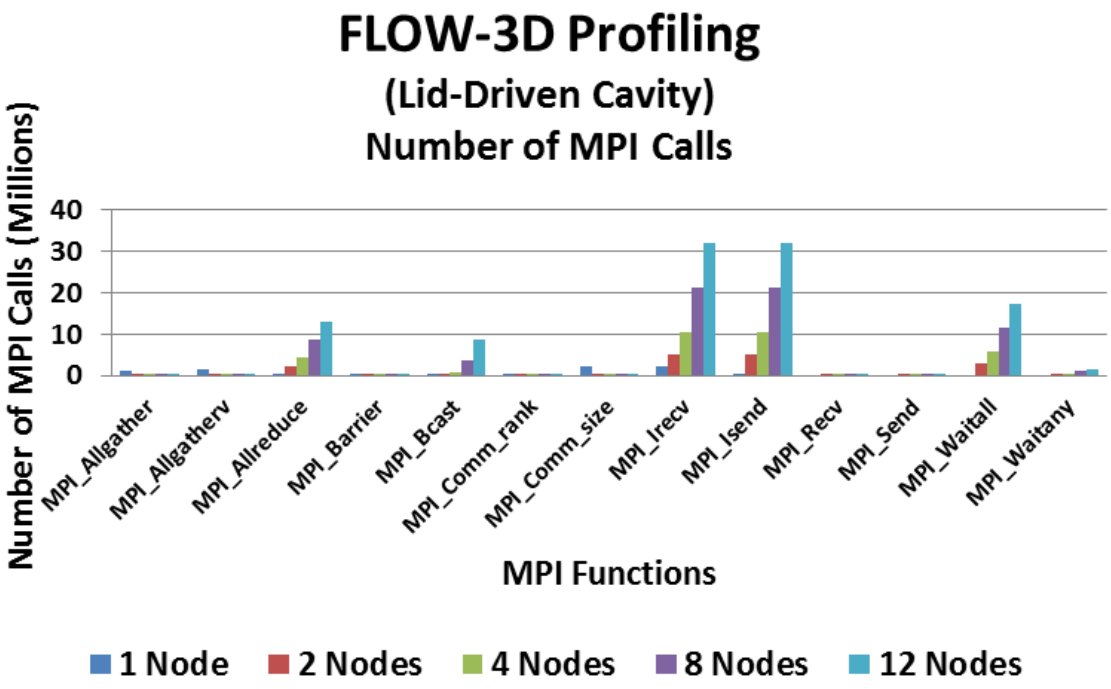
16 MPI Processes/Node

FLOW-3D Profiling – # of MPI Calls

- **The most used MPI calls are MPI_Isend and MPI_Irecv**
  - MPI_Isend(31%), MPI_Irecv(31%), MPI_Waitall(17%), MPI_Allreduce (12%)
  - The number of calls scales proportionally with the number of MPI processes



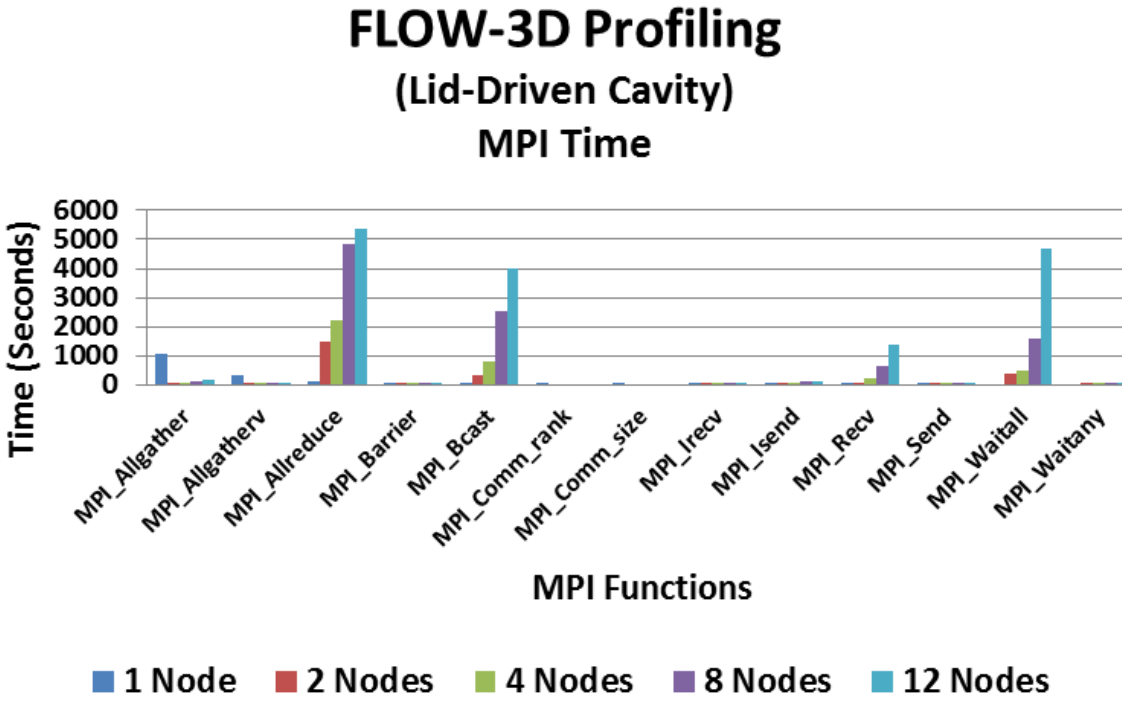
Higher is better

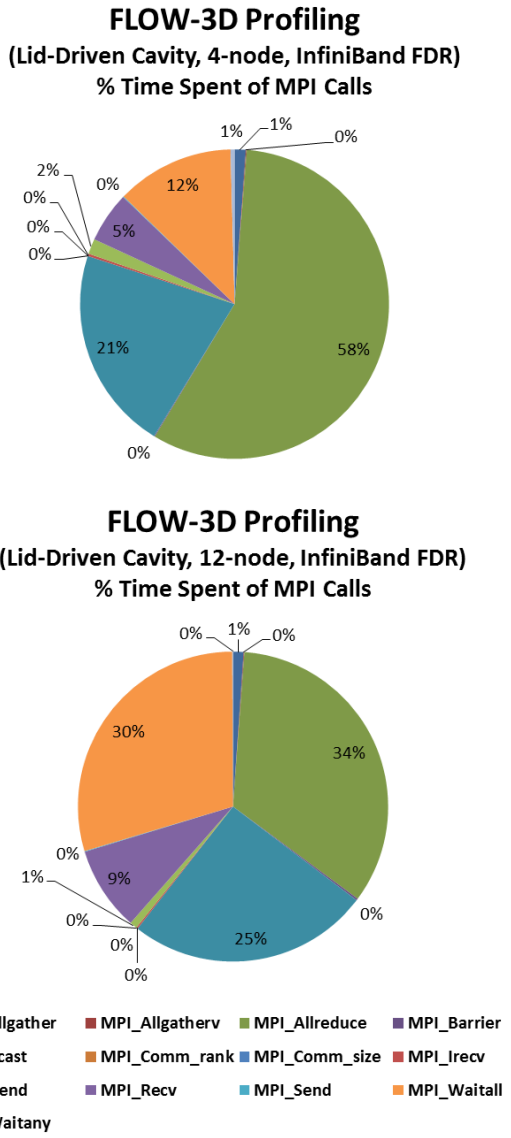


16 MPI Processes/Node

FLOW-3D Profiling – MPI Communication Time

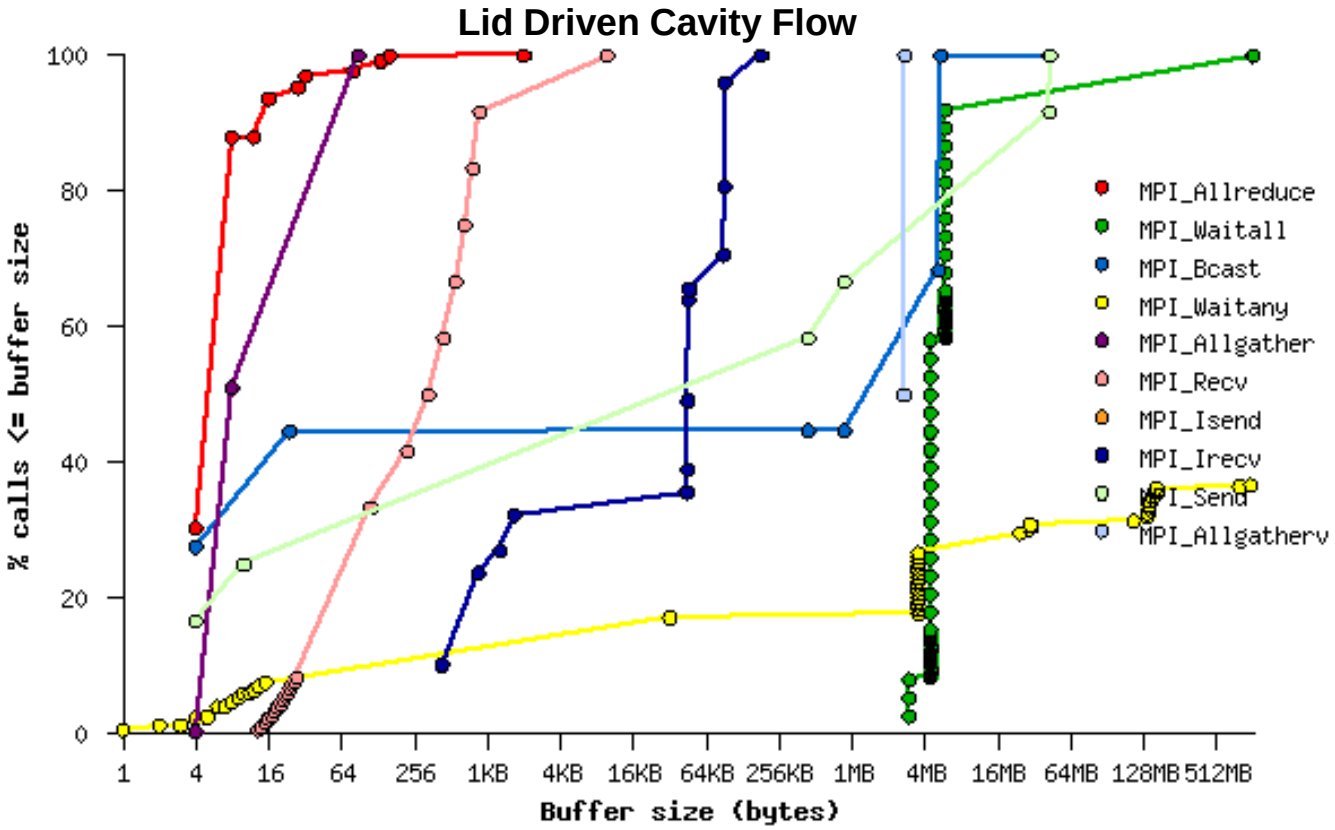
- **The most time consumed MPI functions are:**
  - MPI_Allreduce (34%)
  - MPI_Waitall (30%)
  - MPI_Bcast (25%)
- **MPI_Allreduce time shrinks when cluster size grows**
  - while MPI_Bcast and MPI_Waitall grows





FLOW-3D Profiling – # of MPI Calls

- **There is a wide range of message sizes seen:**
  - MPI_Allreduce: Concentration between 4B to 16B
  - MPI_Waitall: Around 4MB message sizes
  - MPI_Bcast: Around 1MB

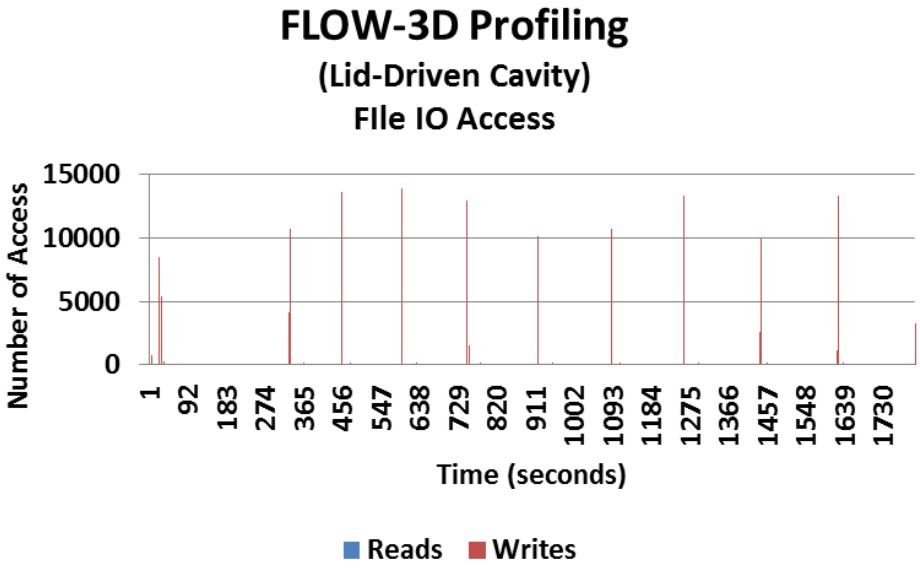


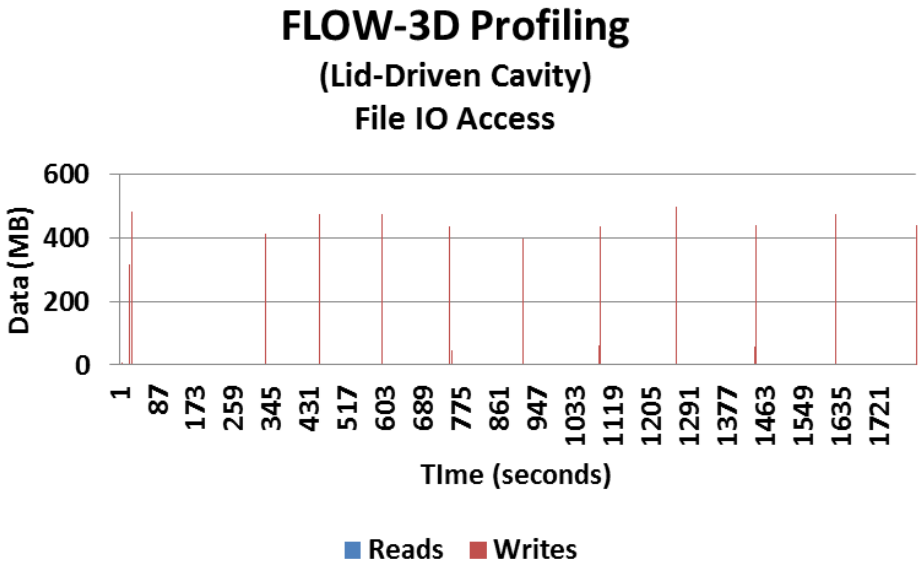
Higher is better

16 MPI Processes/Node

FLOW-3D Profiling – MPI Communication Time

- **File IO access occurs during certain period during the MPI solver**
  - Large spikes for writing the restart and spatial data
  - Files are directed to write to local instead of NFS to avoid IO bottleneck





FLOW-3D – Summary

- **Scalability**
  - FLOW-3D/MP Hybrid version enables higher scalability versus pure MPI version
    - Hybrid version can deliver good scalability to 16 nodes
- **Performance**
  - Intel Xeon E5-2600 series and InfiniBand FDR enable FLOW-3D to scale with 16 nodes
  - The E5-2680 cluster outperforms X5570 “Nehalem” cluster by 70% at 12 nodes
  - The Hybrid mode allows FLOW-3D to scale at 16 nodes, up to 39% better at 16 nodes
- **Network**
  - InfiniBand FDR allows the best scalability performance with 56Gbps rate
    - Outperforms by 580% over 1GbE at 4-node
    - Outperforms by 26% over 10GbE at 12-node
    - Outperforms by 25% over 40GbE at 12-node
- **Profiling**
  - The overall runtime reduces as more nodes take part of the MPI job
  - More compute nodes can reduce the runtime by spreading out the workload
  - MPI Communication time is spent mostly on MPI_Allreduce at 34% of overall MPI time
  - Large concentration on small messages, typical for latency sensitive HPC applications

Thank You

HPC Advisory Council