HPC ADVISORY COUNCIL

FLOW-3D/MP v5.0 Performance Benchmark and Profiling

September 2013

DELL

FLOW Science
Improving the world through accurate flow modeling

intel

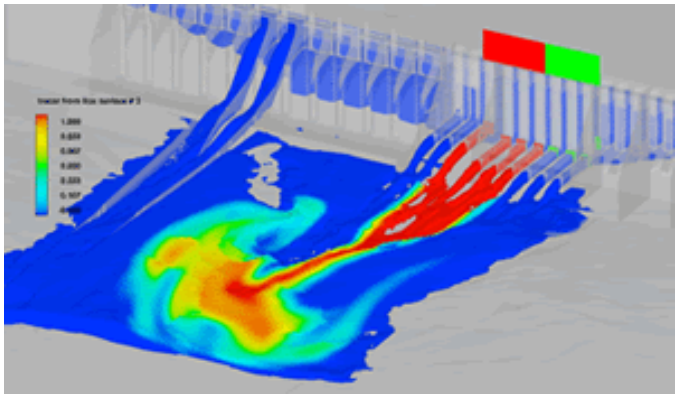
Mellanox TECHNOLOGIES

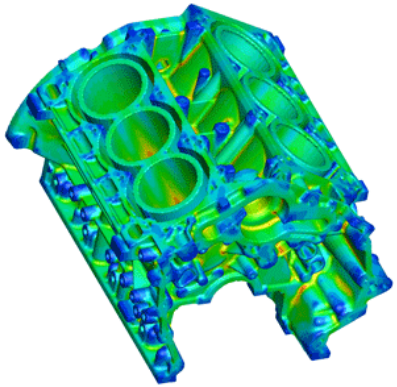
Note

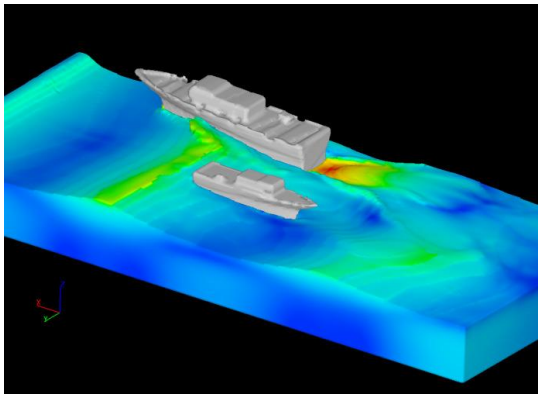
- **The following research was performed under the HPC Advisory Council activities**
  - Participating vendors: FLOW-Science, Dell, Intel, Mellanox
  - Compute resource - HPC Advisory Council Cluster Center
- **For more info please refer to**
  - http://www.dell.com/hpc
  - http://www.flow3d.com
  -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

FLOW-3D/MP

- **FLOW-3D/MP is a powerful and highly-accurate CFD software**
  - Provides engineers valuable insight into many physical flow processes
- **FLOW-3D/MP 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/MP is a standalone, all-inclusive CFD package**
  - Includes an integrated GUI that ties components from problem setup to post-processing







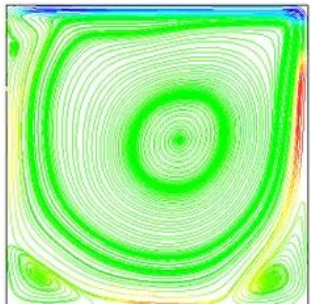
Objectives

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

- **The presented results will demonstrate**
  - The scalability of the compute environment
  - The capability of FLOW-3D/MP 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
  - Dual-Socket Ten-Core Intel E5-2680 V2 @ 2.80 GHz CPUs
  - Memory: 64GB memory, DDR3 1600 MHz
  - OS: RHEL 6.2, MLNX_OFED 2.0-3.0.0 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 (v4.2) and 4.0.0.025 (v5.0)**
- **Application: FLOW-3D/MP MP 4.2 and 5.0**
- **Benchmarks:**
  - Lid Driven Cavity 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

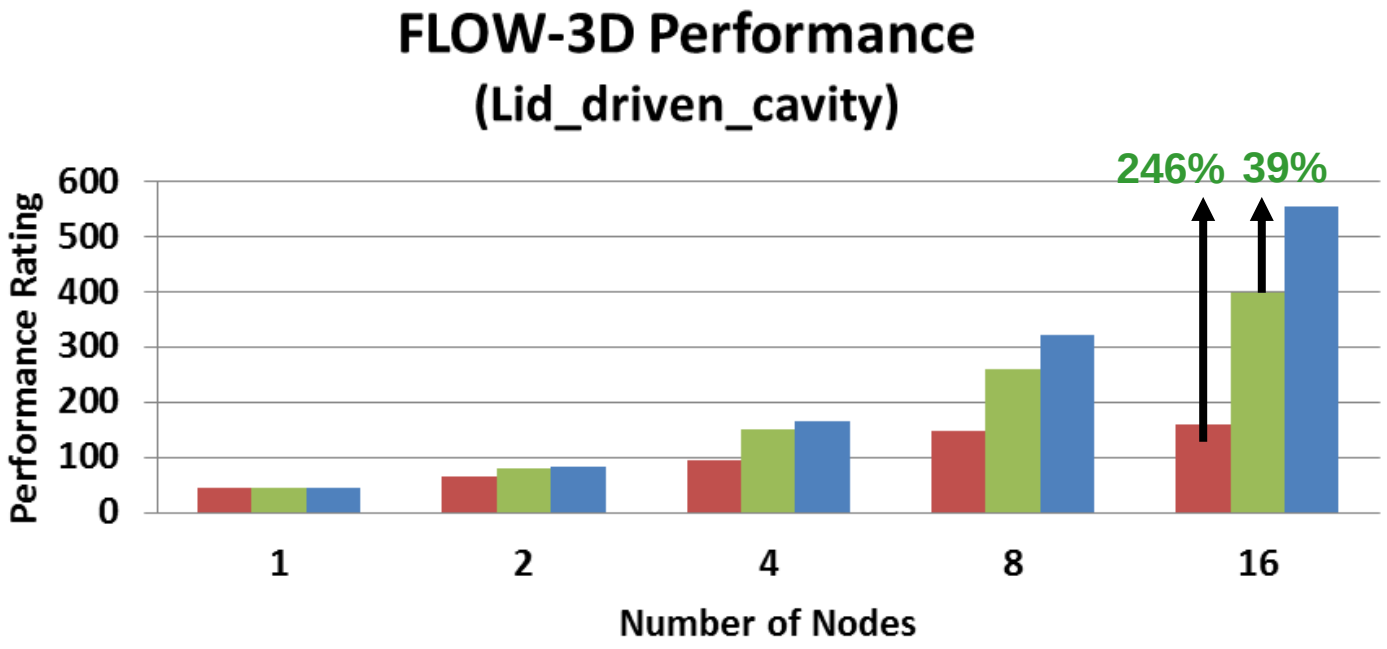
PowerEdge R720xd

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 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/MP Performance – Network

- **InfiniBand FDR provides better scalability performance than Ethernet**
  - Scalability gap widens as more nodes involved in simulation
  - FDR InfiniBand provides up to 246% better performance than 1GbE
  - FDR InfiniBand delivers up to 39% better performance than 10GbE
  - Hybrid mode is shown

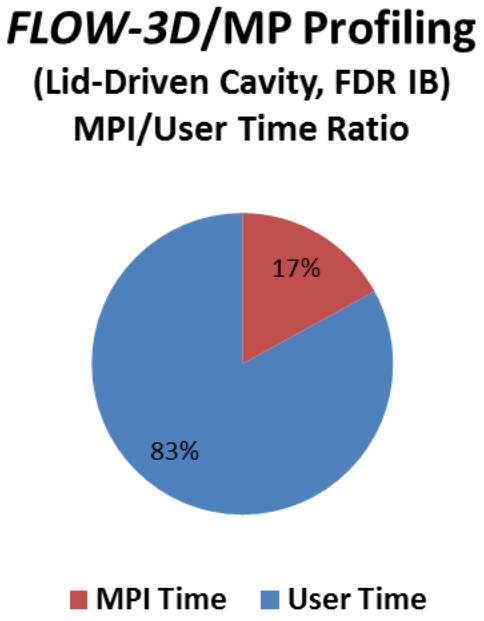


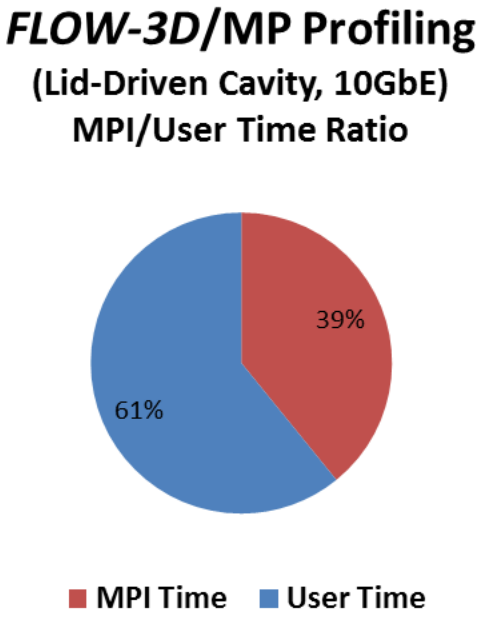
**Performance Rating = Jobs/Day*

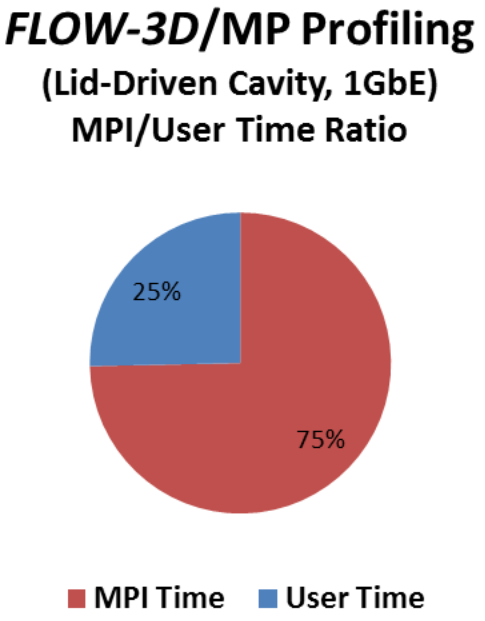
FLOW-3D/MP v5.0

FLOW-3D/MP Profiling – MPI/User Time Ratio

- **InfiniBand FDR reduces the communication time at scale**
  - InfiniBand FDR consumes about 17% of total runtime on 16-node Hybrid job
  - 10GbE consumes 39% of total time, while 1GbE consumes about 75%
- **IB RDMA technology allows communication to bypass CPU involvement**
  - Reduces CPU overhead in handling communication
  - Which leaves more time for application processing

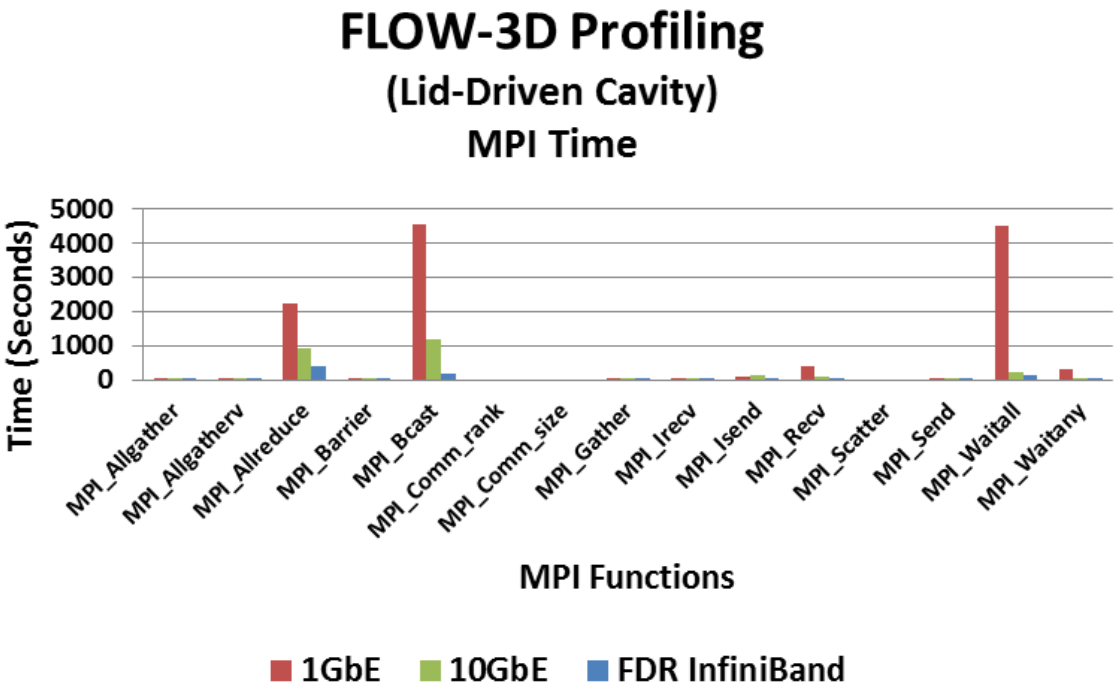


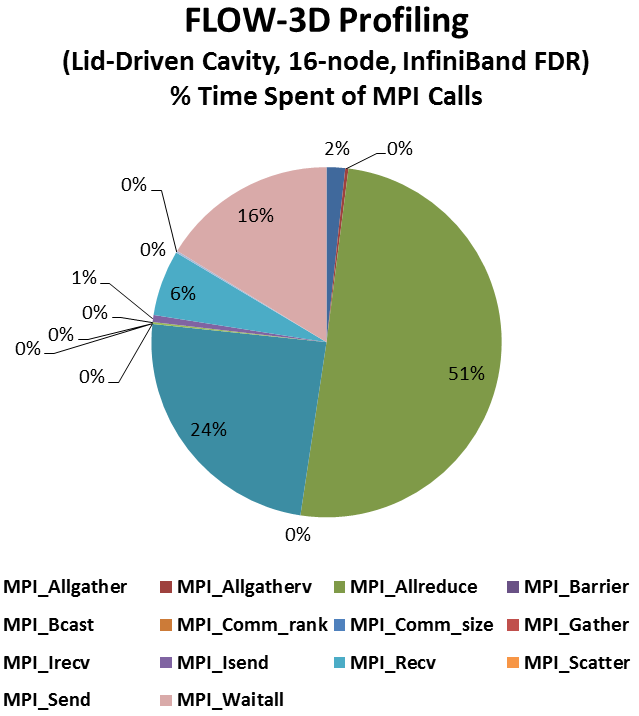




FLOW-3D/MP Profiling – MPI Communication Time

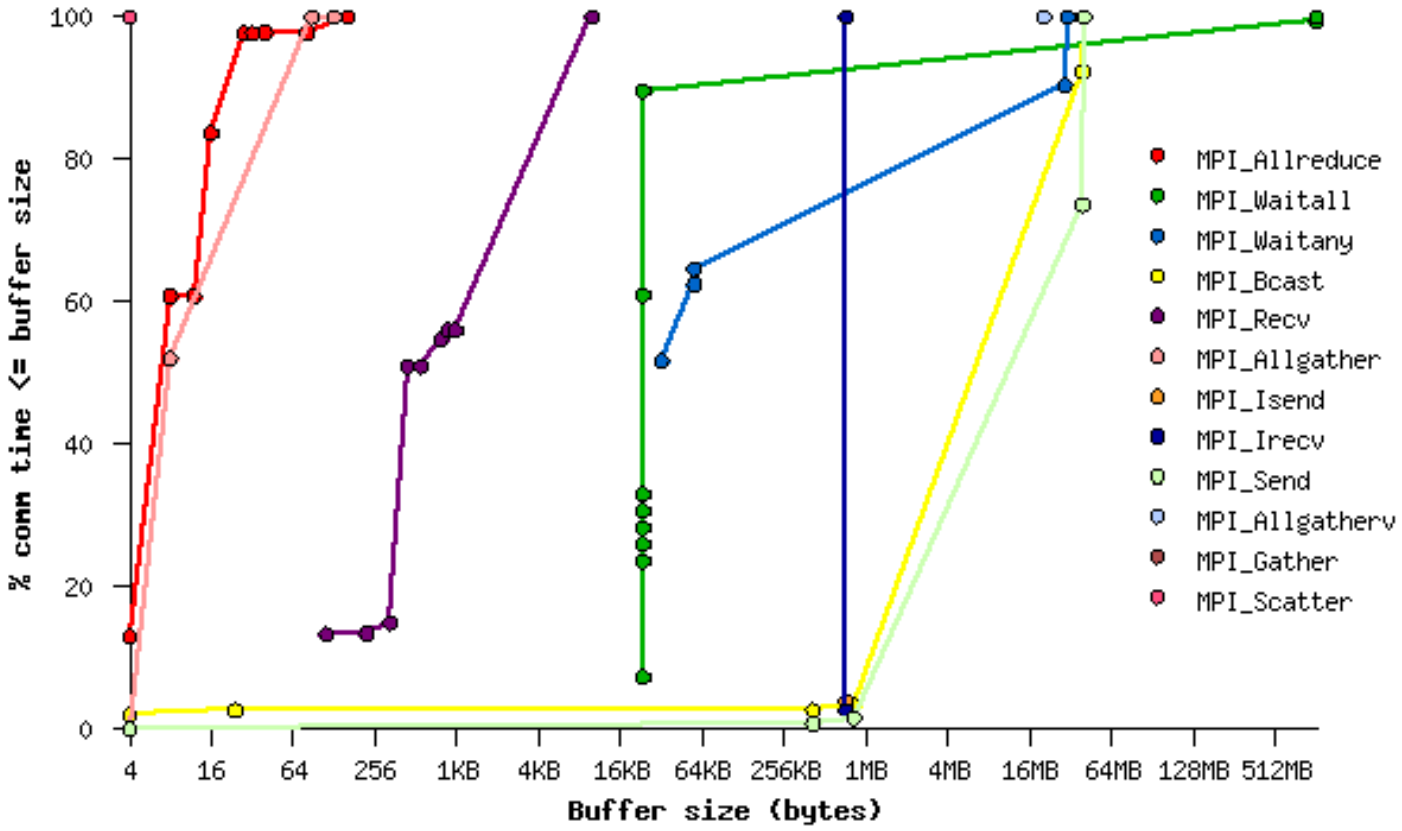
- **The most time consumed MPI functions are:**
  - FDR: MPI_Allreduce(51%), MPI_Bcast(24%), MPI_Waitall(16%)
- **InfiniBand reduces more time in Collective Operations than Ethernet**
  - Collective communications are most used in FLOW-3D/MP v5.0
  - Those communications account for the highest communication time





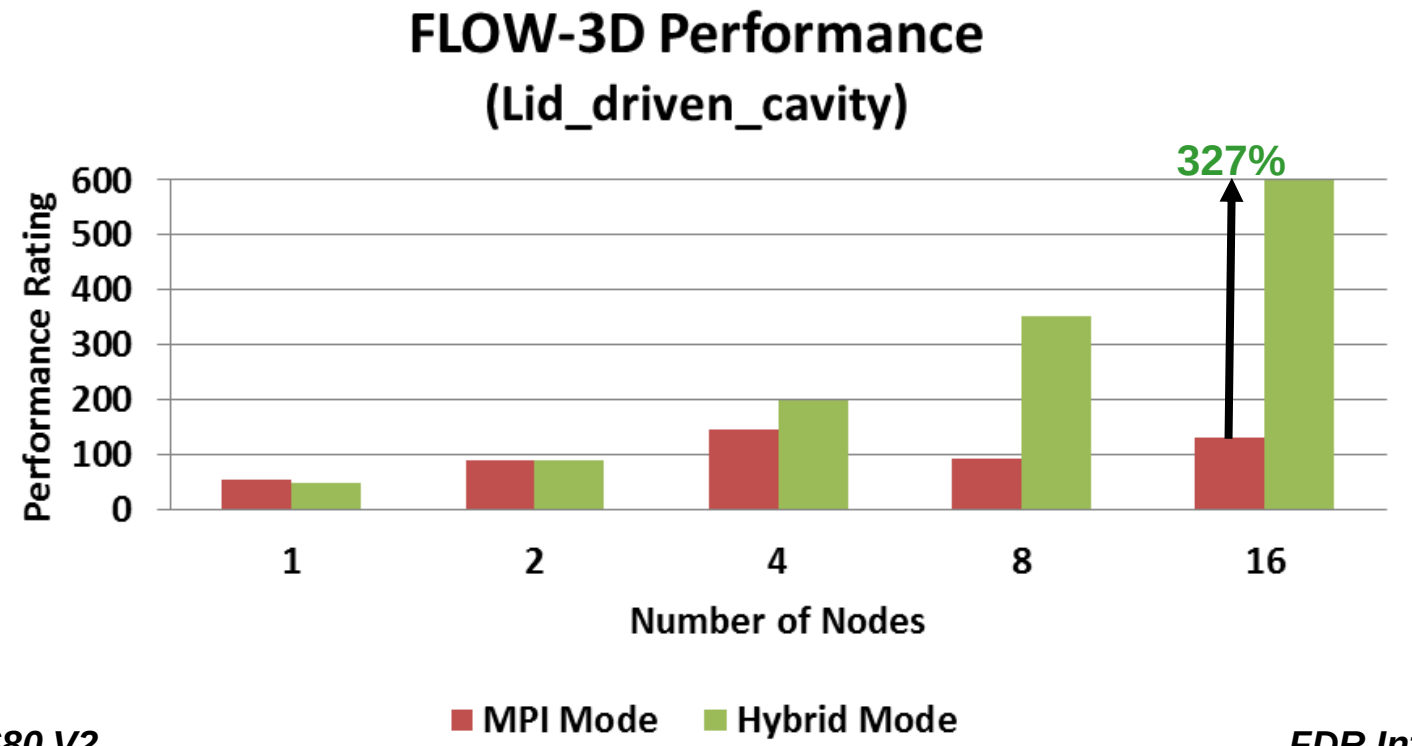
FLOW-3D/MP Profiling – # of MPI Calls

- **There is a wide range of message sizes seen:**
  - MPI_Allreduce: Concentration between 4B to 16B
  - MPI_Waitall: Around 16KB
  - MPI_Bcast: Around 1-64MB



FLOW-3D/MP Performance – MPI vs Hybrid

- **FLOW-3D/MP supports OpenMP Hybrid mode**
  - Up to 327% of higher performance than MPI mode at 16 nodes
- **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

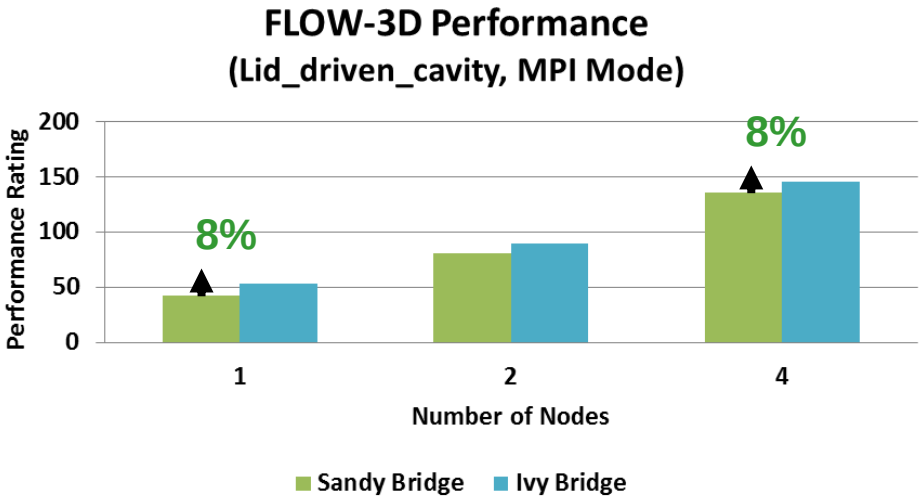


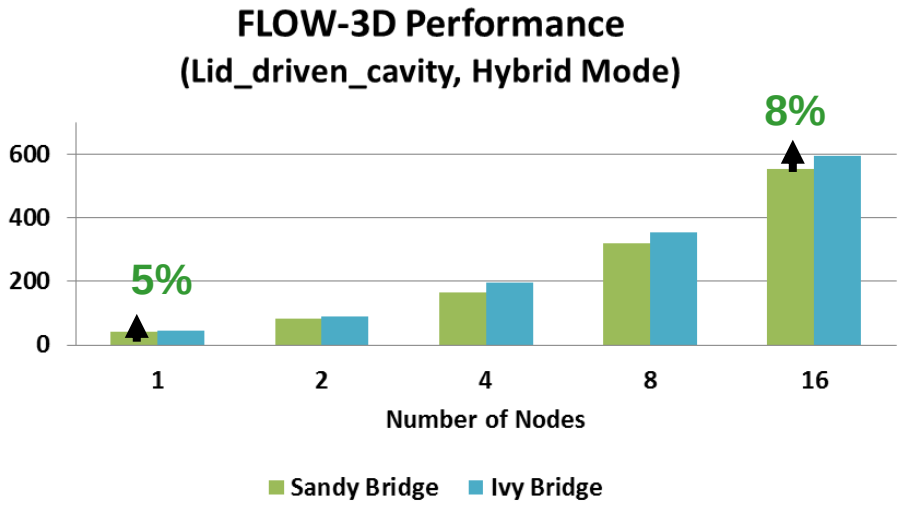
Intel E5-2680 V2

FDR InfiniBand

FLOW-3D/MP Performance – Processors

- **Intel E5-2680 (Sandy Bridge) cluster outperforms prior generations**
  - Performs 8% better than X5670 cluster at 16 nodes
- **System components used:**
  - Sandy Bridge: 2-socket 8-core Intel E5-2680 @ 2.7GHz
  - Ivy Bridge: 2-socket 10-core Intel E5-2680 V2 @ 2.8GHz



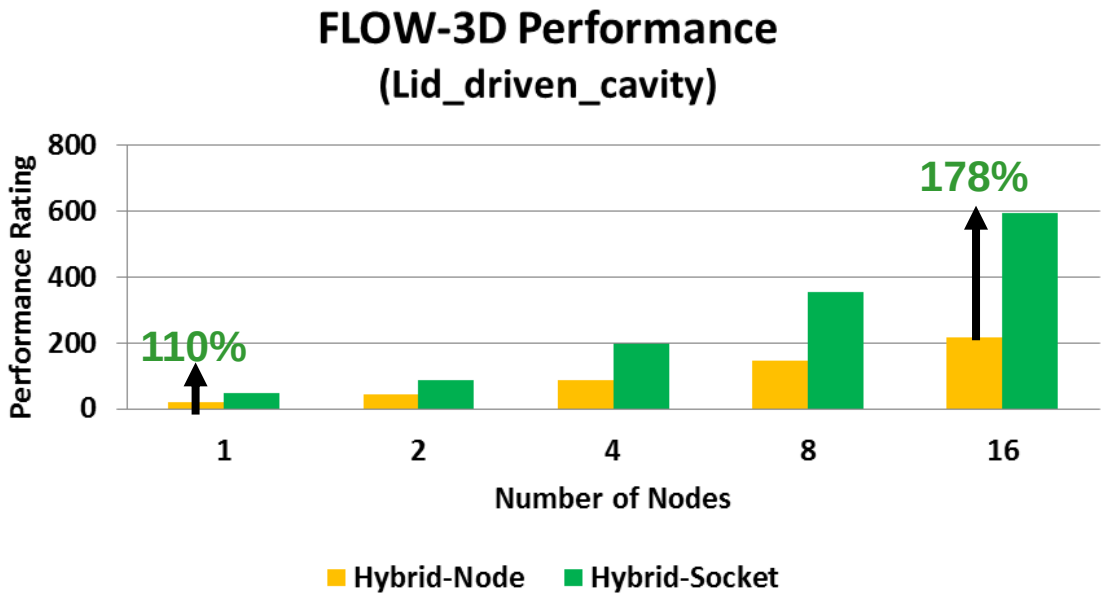


FLOW-3D/MP v5.0

FDR InfiniBand

FLOW-3D/MP Performance – Hybrid Mode

- **Default for FLOW-3D/MP Hybrid mode is to run 1 PPN of 16 threads**
  - Which uses **"-genv I_MPI_PIN_DOMAIN node"** as specified in **runhyd_par** script
- **For best performance on 2P Sandy Bridge or Ivy Bridge Platform:**
  - Use "-**genv I_MPI_PIN_DOMAIN socket**" in **runhyd_par** script
  - 2PPN of 8/10 threads can yield better performance than 1PPN of 16/20 threads
  - The flag allows to each MPI process to spawn threads within its own socket
  - Instead of both MPI processes sharing the same socket

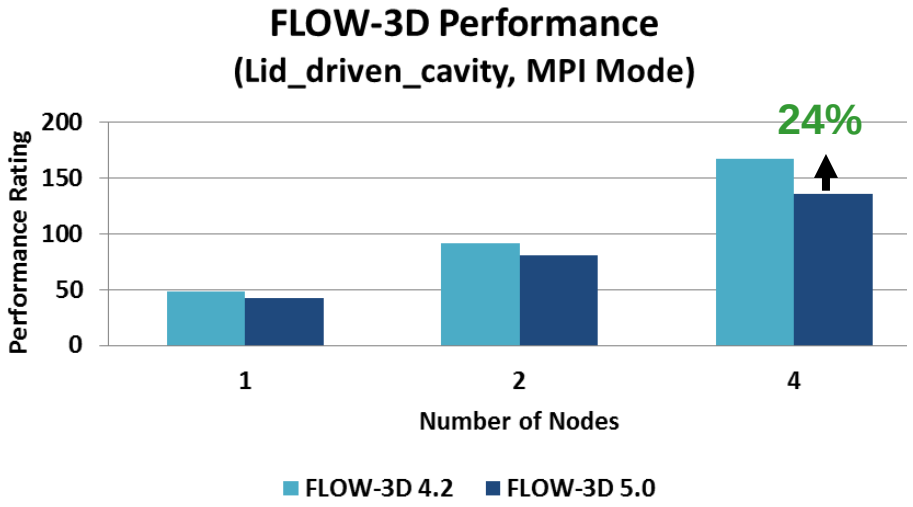


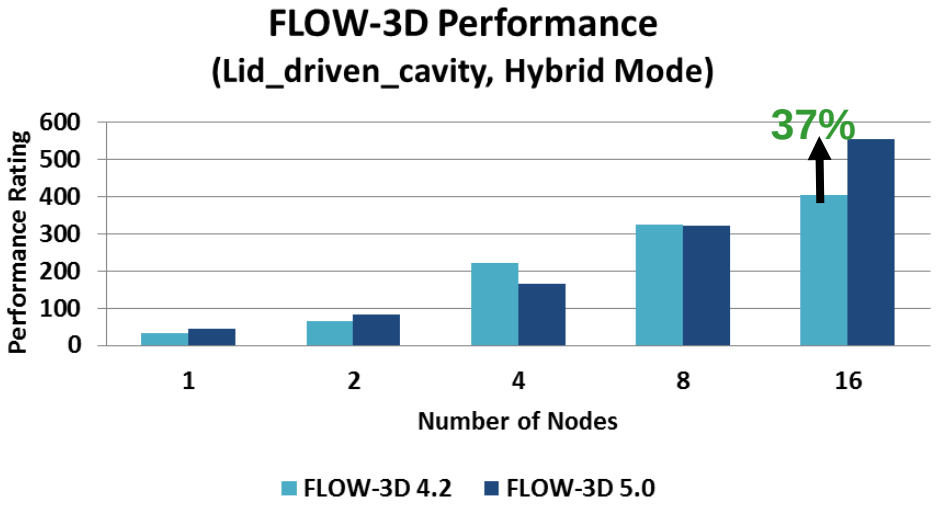
Intel E5-2680 V2

FDR InfiniBand

FLOW-3D/MP Performance – SW Versions

- **FLOW-3D/MP v5.0 outperforms v4.2 in scalability performance**
  - Provides up to 37% faster in Hybrid mode at 16-node
- **MPI mode shows longer runtime than previous version**
  - Which seems to cause by a change in communication algorithm
  - More MPI collective operations are being used compared to prior version





Intel E5-2680

16 MPI Processes/Node

FLOW-3D/MP – Summary

- **Scalability**
  - FLOW-3D/MP v5.0 Hybrid mode enables higher scalability versus pure MPI version
    - Hybrid version can deliver good scalability to 16 nodes; over 37% faster than FLOW-3D/MP v4.2
- **Performance**
  - Xeon E5-2600 v2 series and FDR InfiniBand enable FLOW-3D/MP to scale to 320 cores
  - Allocate MPI process to “proper” socket in Hybrid mode allows performance to jump 178%
  - Hybrid mode allows FLOW-3D/MP to scale at 16 nodes, over 327% faster than MPI mode
- **Network**
  - InfiniBand FDR allows the best scalability performance with 56Gbps rate
    - Outperforms by 246% over 1GbE at 16-node (on 320 CPU cores)
    - Outperforms by 39% over 10GbE at 16-node (on 320 CPU cores)
  - RDMA technology in InfiniBand allows bypassing CPU for network transfer
    - This offload reduces CPU overhead in handling communication; thus CPU can focus on application
- **Profiling**
  - MPI Communication time is spent mostly on MPI_Allreduce at 51% of overall MPI time
  - InfiniBand can process Collective Operations in network faster than Ethernet
  - Large concentration on small messages, typical for latency sensitive HPC applications

Thank You

HPC Advisory Council