



AcuSolve

Performance Benchmark and Profiling

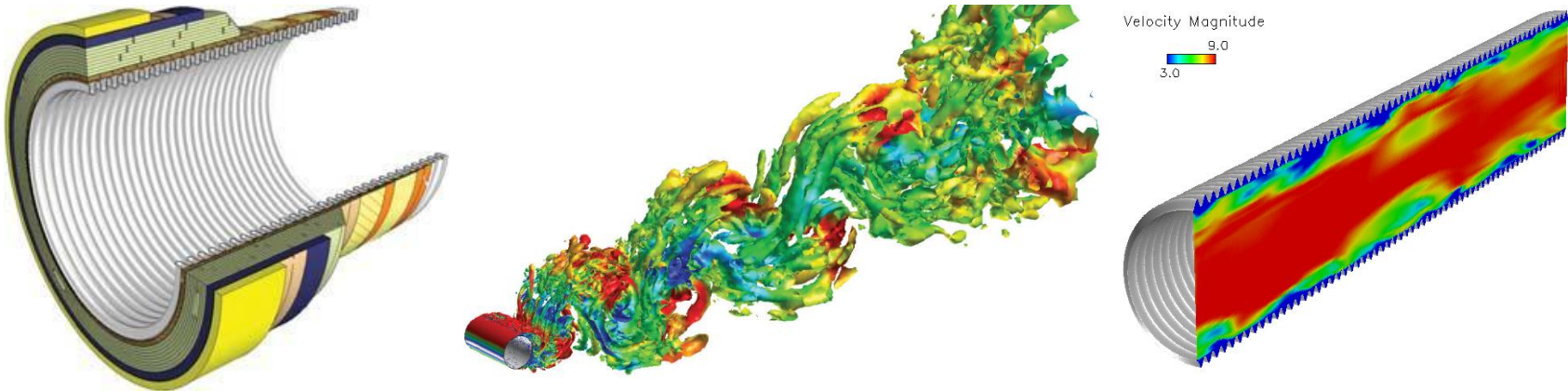
October 2011



- **The following research was performed under the HPC Advisory Council activities**
 - Participating vendors: Intel, Dell, Mellanox, Altair
 - Compute resource - HPC Advisory Council Cluster Center
- **The following was done to provide best practices**
 - AcuSolve performance overview
 - Understanding AcuSolve communication patterns
 - Ways to increase AcuSolve productivity
 - MPI libraries comparisons
- **For more info please refer to**
 - <http://www.dell.com>
 - <http://www.intel.com>
 - <http://www.mellanox.com>
 - <http://www.altairhyperworks.com/Product,54,AcuSolve.aspx>

- **AcuSolve**

- AcuSolve™ is a leading general-purpose finite element-based Computational Fluid Dynamics (CFD) flow solver with superior robustness, speed, and accuracy
- AcuSolve can be used by designers and research engineers with all levels of expertise, either as a standalone product or seamlessly integrated into a powerful design and analysis application
- With AcuSolve, users can quickly obtain quality solutions without iterating on solution procedures or worrying about mesh quality or topology



- **Dell™ PowerEdge™ M610 38-node (456-core) cluster**
 - Six-Core Intel X5670 @ 2.93 GHz CPUs
 - Memory: 24GB memory, DDR3 1333 MHz
 - OS: RHEL 5.5, OFED 1.5.2 InfiniBand SW stack
- **Intel Cluster Ready certified cluster**
- **Mellanox ConnectX-2 InfiniBand adapters and non-blocking switches**
- **MPI: Intel MPI 3.0, MVAPICH2 1.0, Platform MPI 7.1**
- **InfiniBand-based Lustre Storage: Lustre 1.8.5**
- **Application: AcuSolve 1.8a**
- **Benchmark datasets:**
 - Pipe_fine (700 axial nodes, 3.04 million mesh points total, 17.8 million tetrahedral elements)
 - The test computes the steady state flow conditions for the turbulent flow ($Re = 30000$) of water in a pipe with heat transfer. The pipe is 1 meter in length and 150 cm in diameter. Water enters the inlet at room temperature conditions.

- **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

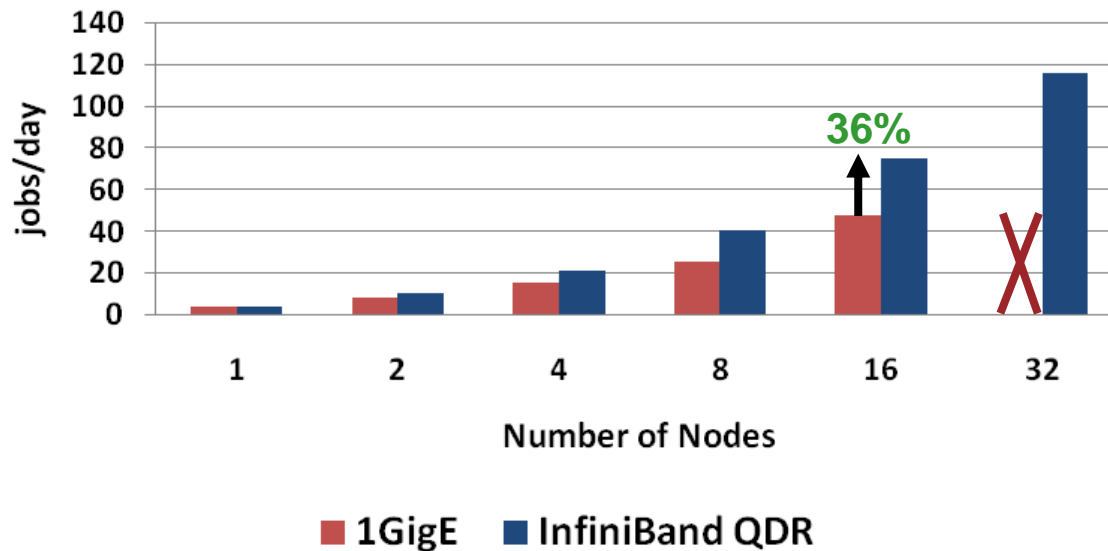


- **System Structure and Sizing Guidelines**
 - 38-node cluster build with Dell PowerEdge™ M610 blade servers
 - Servers optimized for High Performance Computing environments
 - Building Block Foundations for best price/performance and performance/watt
- **Dell HPC Solutions**
 - Scalable Architectures for High Performance and Productivity
 - Dell's comprehensive HPC services help manage the lifecycle requirements.
 - Integrated, Tested and Validated Architectures
- **Workload Modeling**
 - Optimized System Size, Configuration and Workloads
 - Test-bed Benchmarks
 - ISV Applications Characterization
 - Best Practices & Usage Analysis



- **InfiniBand QDR enables higher cluster productivity**
 - Provides more than 36% of job productivity than 1GigE network on benchmark problem
 - Savings in job productivity increases as cluster size increases
- **1GigE performance has a limited effect on performance for this benchmark**
 - Infers that the application is not as sensitive to network latency
- **Test stops at 16-node for 1GigE due to switch port limitation**

AcuSolve Benchmark
(pipe_fine)

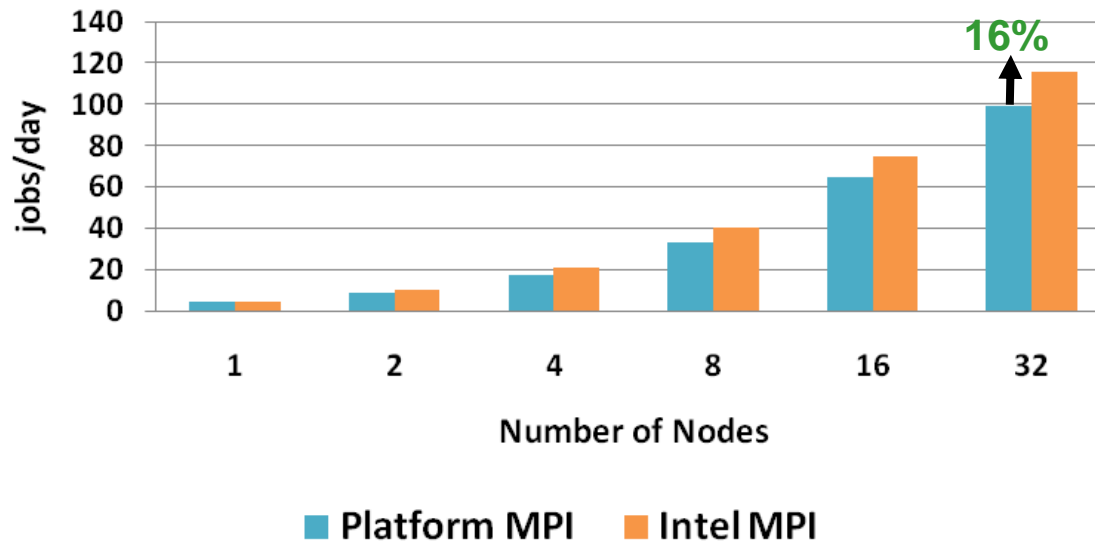


Higher is better

InfiniBand QDR

- **Intel MPI performs better than Platform MPI**
 - Seen around 16% higher performance at 32-node
 - Reflects that each Intel MPI handles efficiently for the MPI data transfers
- **MVAPICH2 executable is only built with ch3:sock support for TCP network**
 - Thus it does not reflect the true ibverbs performance as other MPI implementations
 - Comparison is not shown since the executable does not contain the InfiniBand support

AcuSolve Benchmark (pipe_fine)

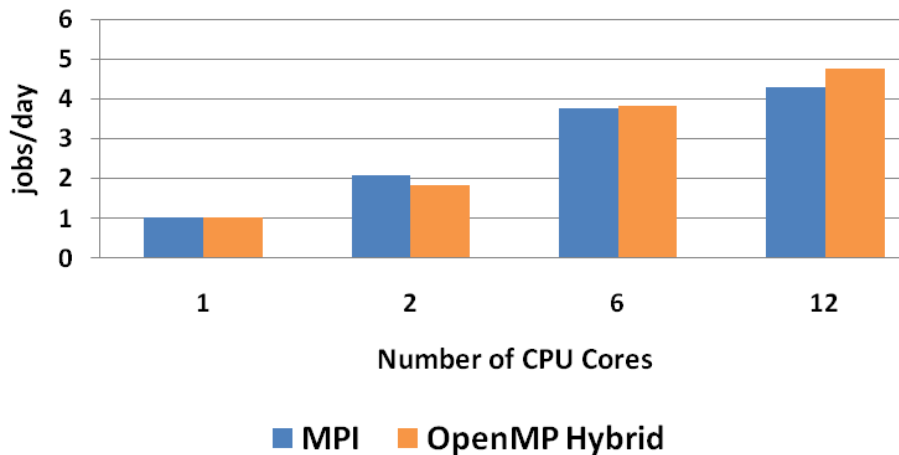


Higher is better

InfiniBand QDR

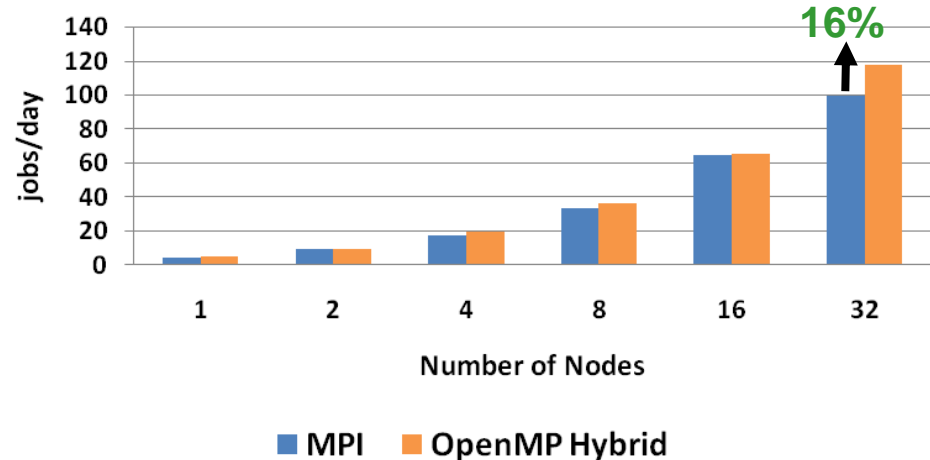
- **On a single node, OpenMP Hybrid performs better than pure MPI**
 - OpenMP provides faster results starting with 6 CPU cores (or 6 OpenMP threads)
 - OpenMP hybrid threads is a lighter weight alternative compared to MPI processes
- **Hybrid process enables scalability by minimizing process and communications**
 - MPI communications are done by an MPI-OpenMP hybrid process on each node
 - The hybrid process is responsible for communications and spawning off worker threads
 - The OpenMP worker threads subsequently responsible for computation
- **Graphs below compare Platform MPI to Platform MPI/OpenMP hybrid**

AcuSolve Benchmark
(pipe_fine)



Higher is better

AcuSolve Benchmark
(pipe_fine)

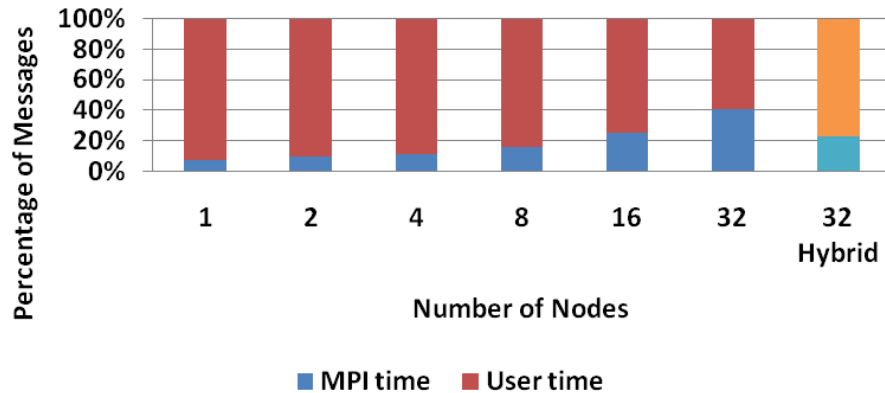


InfiniBand QDR

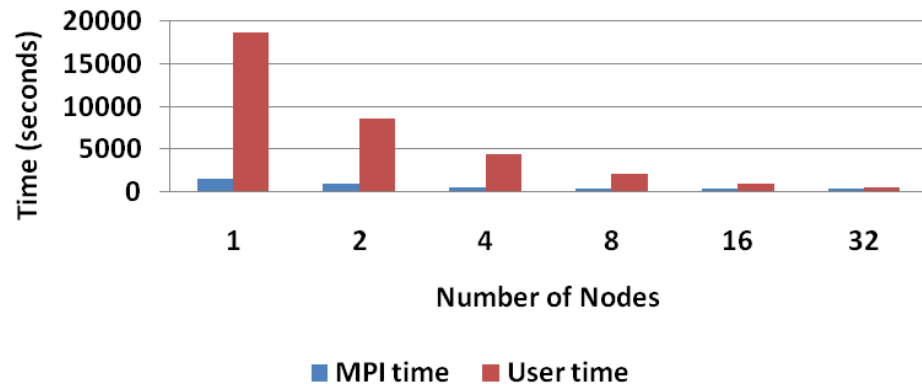
AcuSolve Profiling – MPI/User Time Ratio

- **Time spent in computation is more dominant than the MPI communication**
 - MPI time only accounts for around 40% at 32-node
 - Actual computation run time reduces as the cluster scales
- **OpenMP hybrid mode reduces overheads and yields more time for computation**
 - Computation time: From 60% in pure MPI mode versus 77% in OpenMP hybrid mode

AcuSolve Profiling
(pipe_fine)
MPI/User Time Ratio



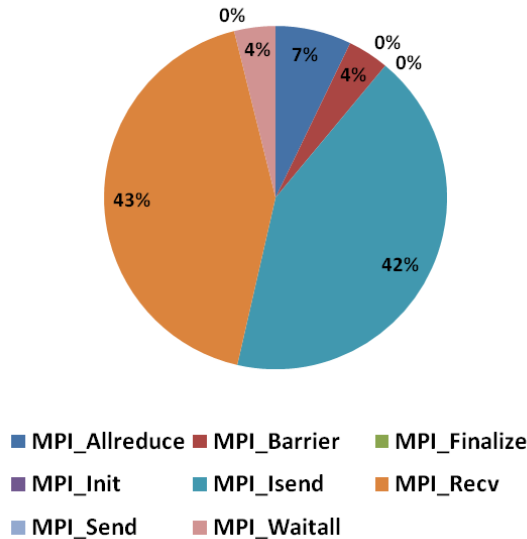
AcuSolve Profiling
(pipe_fine)
MPI/User Time Ratio



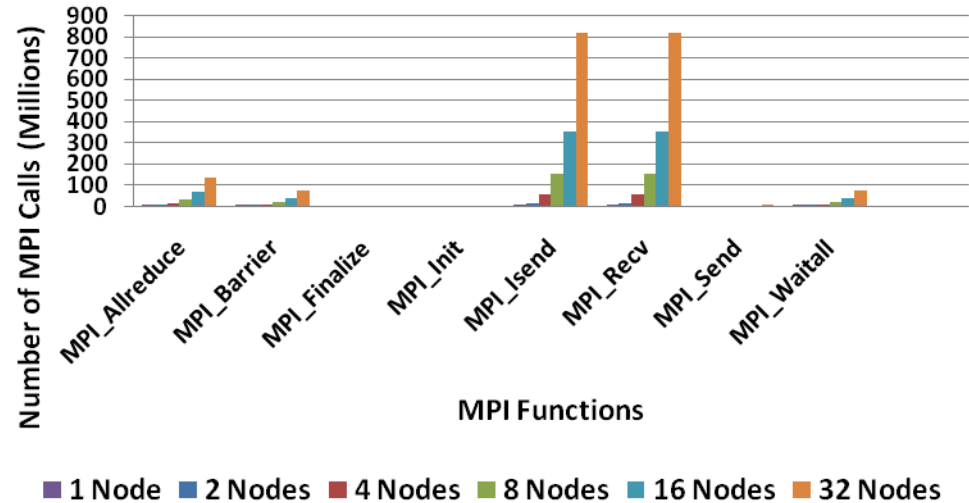
InfiniBand QDR

- **MPI_Recv and MPI_Isend are the most used MPI calls**
 - Each is accounted for around 42-43% of the MPI function calls on a 32-node job
- **AcuSolve has a large percent of MPI calls for non-blocking data transfers**
 - The non-blocking APIs allows transferring data while overlaps computation
 - Along with minimizing communications by using OpenMP hybrid
 - These 2 measures allow slow network to maintain decent productivity

AcuSolve Profiling
(pipe_fine, 32-node, InfiniBand, MPI)
% MPI Calls



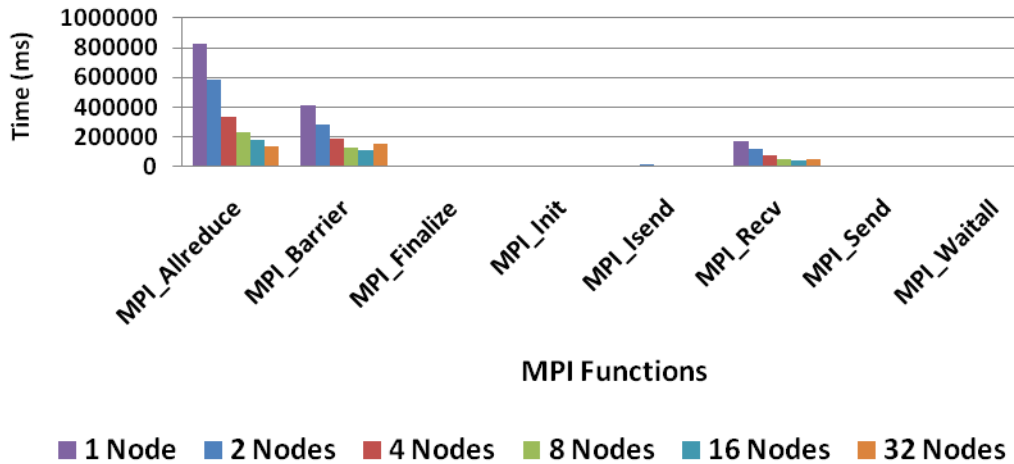
AcuSolve Profiling
(pipe_fine)
Number of MPI Calls



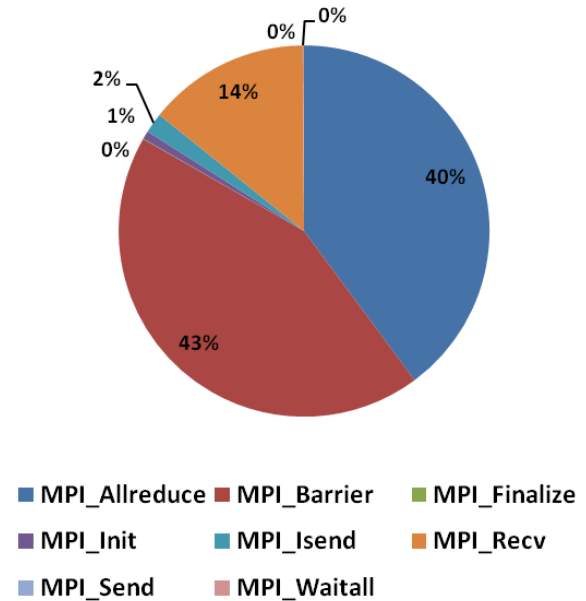
AcuSolve Profiling – Time Spent by MPI Calls

- **Majority of the MPI time is spent on MPI_Allgather and MPI_Allgather**
 - MPI_Allgather(43%), MPI_Allgather(40%), MPI_Commfree(14%) on 32-node
- **MPI communication time drops as cluster scales**
 - Due to the faster total runtime, as more CPUs are working on completing the job faster
 - Reducing the communication time for each of the MPI calls

AcuSolve Profiling
(pipe_fine)
Time Spent of MPI Calls



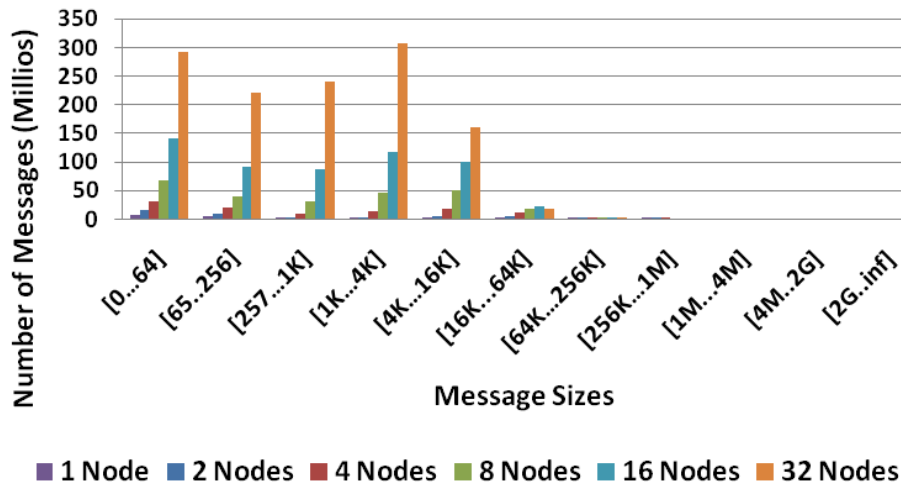
AcuSolve Profiling
(pipe_fine, 32-node, InfiniBand, MPI)
% Time Spent of MPI Calls



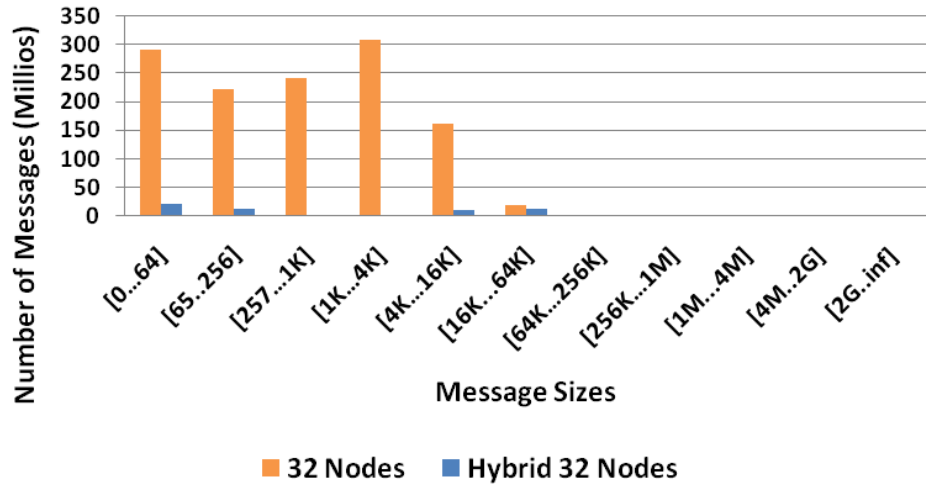
AcuSolve Profiling – MPI Message Sizes

- **Most of the MPI messages are in the range of small to medium sizes**
 - Most message sizes are less than 4KB
- **The volume of MPI messages in MPI are significantly higher than hybrid**
 - While the concentration of the messages stay within the same range

AcuSolve Profiling
(pipe_fine)
MPI Message Sizes

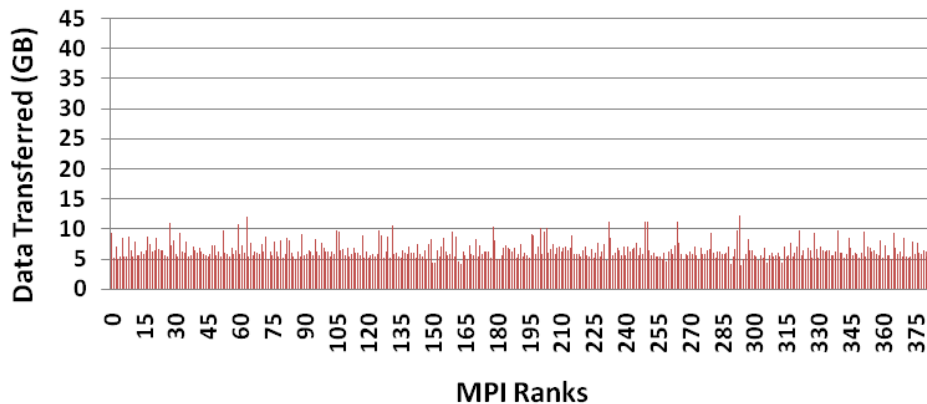


AcuSolve Profiling
(pipe_fine)
MPI Message Sizes

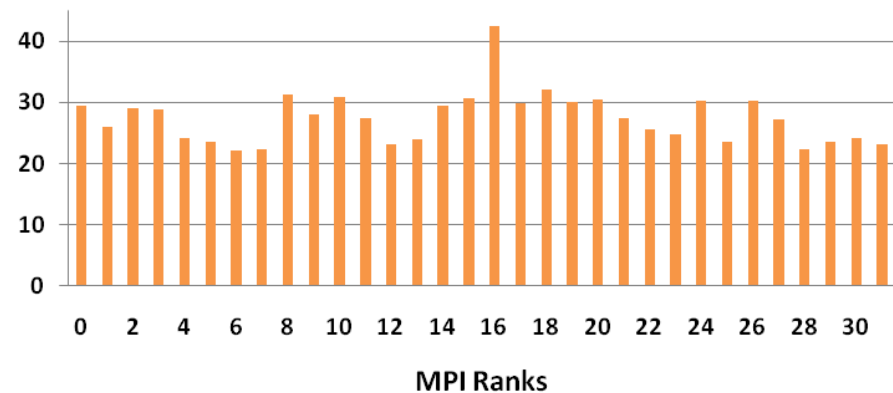


- **The amount of communications becomes more concentrated with hybrid mode**
 - With 1 hybrid process launched for each node that is responsible for communications
 - Leaving the worker OpenMP threads for doing parallel computational routines
- **At a result, the hybrid mode becomes a more efficient mode at scale**
 - Even though larger data transfers takes place between MPI processes on each node

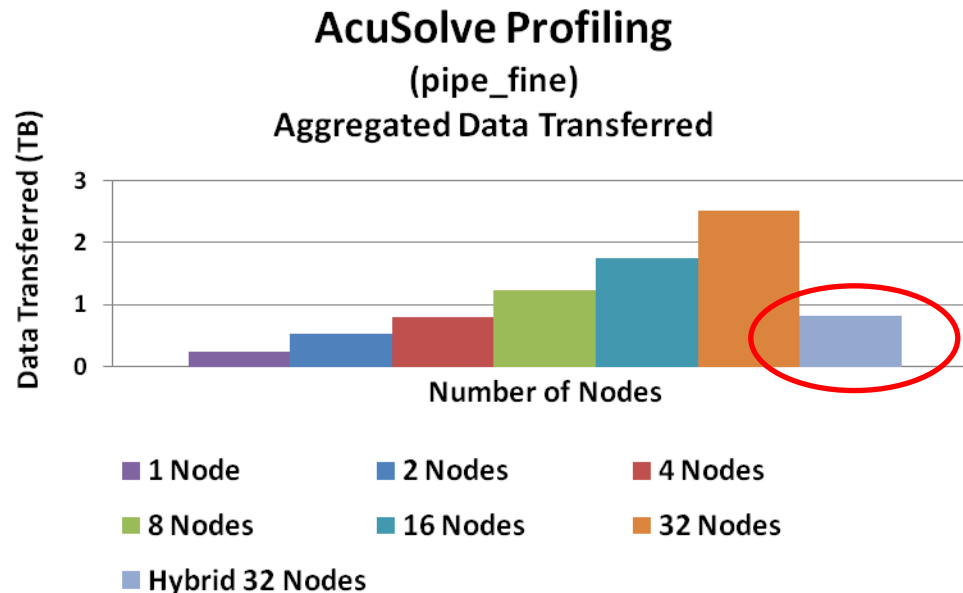
AcuSolve Profiling
(pipe_fine, 32-node, MPI)
Data Transferred by Ranks



AcuSolve Profiling
(pipe_fine, 32-node, Hybrid)
Data Transferred by Ranks



- **Aggregated data transfer refers to:**
 - Total amount of data being transferred in the network between all MPI ranks collectively
- **Large sum of data transfer takes place in AcuSolve**
 - Seen around 2.5TB of data being exchanged between the nodes at 32-node in MPI
- **The OpenMP hybrid mode reduces the overall traffic between the MPI processes**
 - OpenMP has less than 870GB of data transferred, compared to 2.5TB for pure MPI case



- **Performance**

- Acusolve is designed for superior performance and scalability
- InfiniBand allows AcuSolve to run at the most efficient rate
- Intel MPI produces higher parallel job efficiency than Platform MPI
- The MVAPICH2 executable does not support communications over InfiniBand verbs

- **MPI**

- By deploying non-blocking MPI calls, it overlaps computation with in-flight communications
- Thus allowing it to achieve higher job performance while reducing communication needed

- **OpenMP hybrid mode**

- By using the hybrid model, less data is needed be exchanged between nodes in a cluster
- Thus allowing job to be done faster as more resources available for the computation

- **Profiling**

- MPI_Isend and MPI_Recv are the most used MPI functions
- OpenMP mode reduces the amount of network data transfer that needs to take place

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