



AMBER 11

Performance Benchmark and Profiling

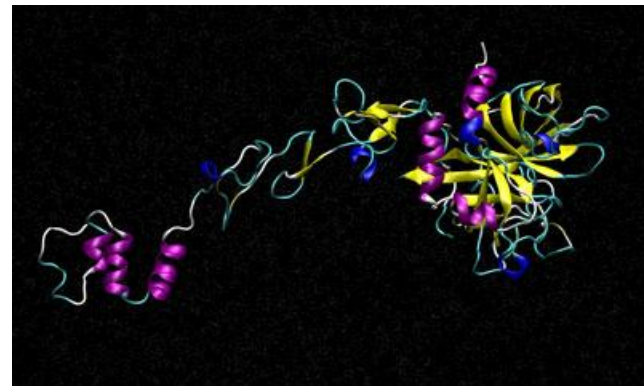
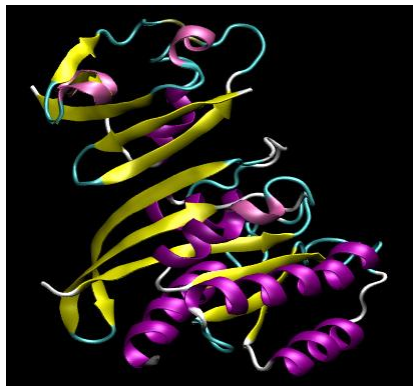
July 2011



- **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**
 - AMBER performance overview
 - Understanding AMBER communication patterns
 - Ways to increase AMBER productivity
 - MPI libraries comparisons
- **For more info please refer to**
 - <http://www.dell.com>
 - <http://www.intel.com>
 - <http://www.mellanox.com>
 - <http://ambermd.org>

- **AMBER**

- Software for analyzing large-scale molecular dynamics (MD) simulation trajectory data
- Reads either CHARMM or AMBER style topology/trajectory files as input, and its analysis routines can scale up to thousands of compute cores or hundreds of GPU nodes with either parallel or UNIX file I/O
- AMBER has dynamic memory management, and each code execution can perform a variety of different structural, energetic, and file manipulation operations on a single MD trajectory at once
- The code is written in a combination of Fortran90 and C, and its GPU kernels are written with NVIDIA's CUDA API to achieve maximum GPU performance



- **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**
- **Mellanox Fabric Collectives Accelerator™ 2.1**
- **Compiler: GNU Compilers 4.1.2 Intel Compilers 11.1, PGI 10.9**
- **MPI: Intel MPI 4, Open MPI 1.5.3 with KNEM 0.9.6, Platform MPI 8.1.1**
- **Application: AMBER 11 (PMEMD), AmberTools 1.5**
- **Benchmark datasets:**
 - Myoglobin – 2492 atoms in Generalized Born (GB) implicit solvent
 - FactorIX – 90,906 atoms in Particle Mesh Ewald (PME) explicit solvent

- **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

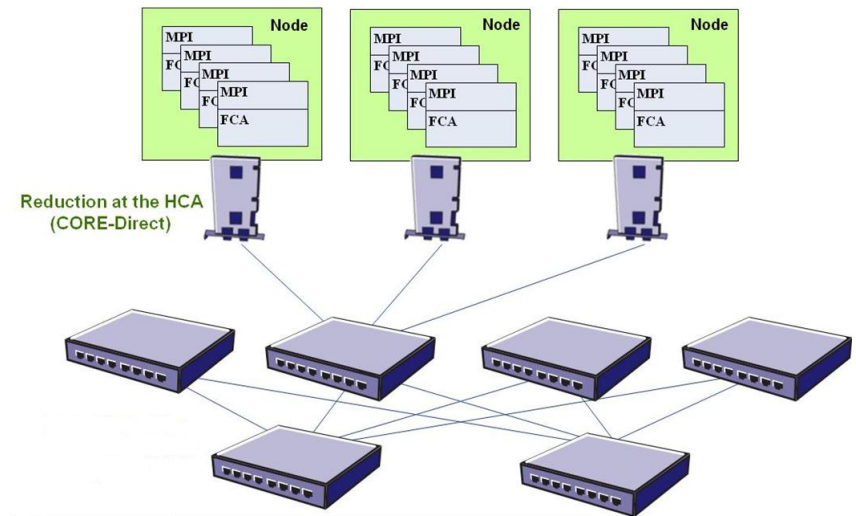


- **System Structure and Sizing Guidelines**
 - 38-node cluster build with Dell PowerEdge™ M610 blade servers
 - Servers optimized for High Performance Computing environments
 - Building Block Foundations for best price/performance and performance/watt
- **Dell HPC Solutions**
 - Scalable Architectures 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

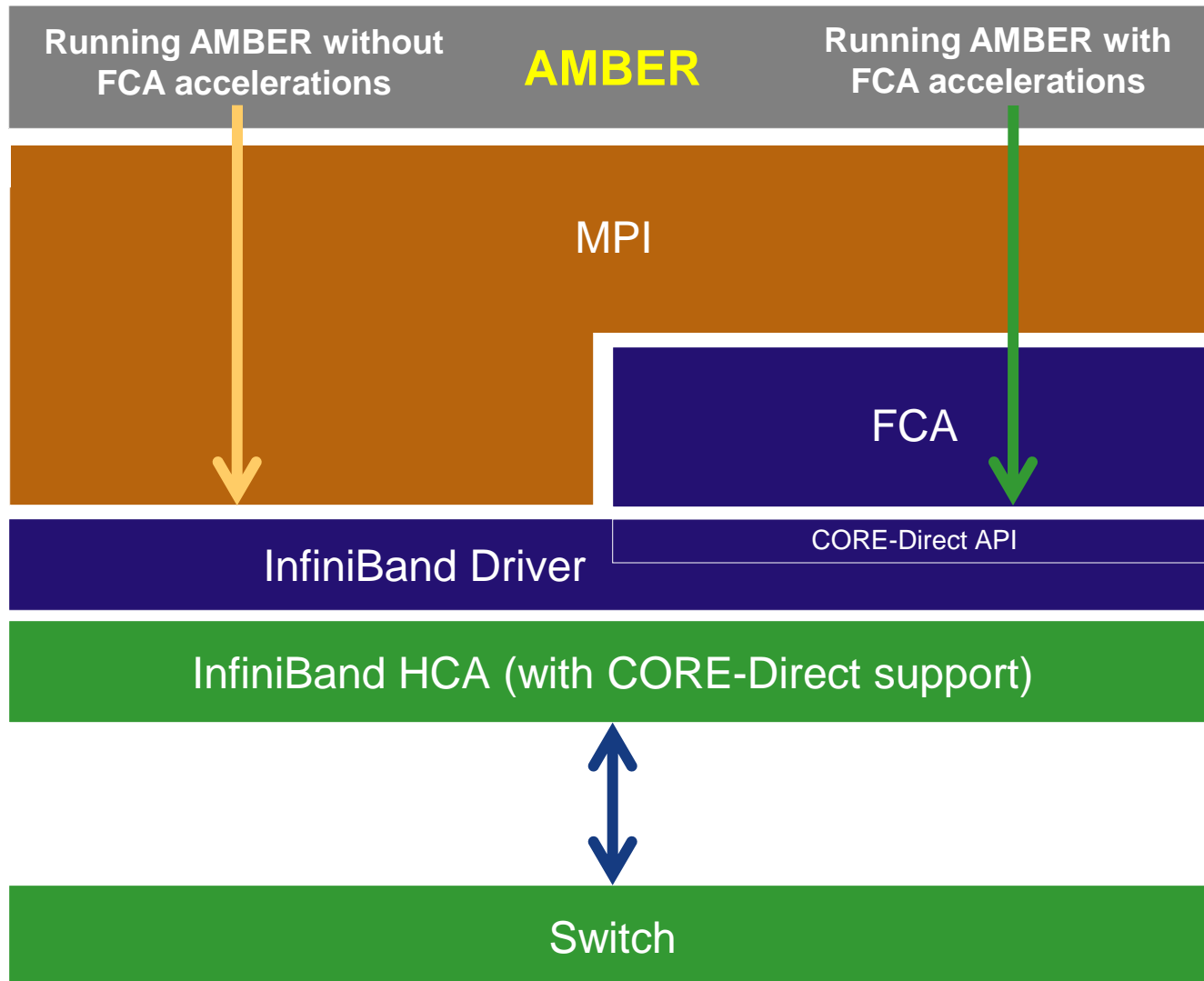


- **Mellanox Fabric Collectives Accelerator (FCA)**
 - Utilized hardware accelerations on the adapter (CORE-Direct)
 - Accelerating MPI collectives operations by offloading them to the network
 - The world first complete solution for MPI collectives offloads

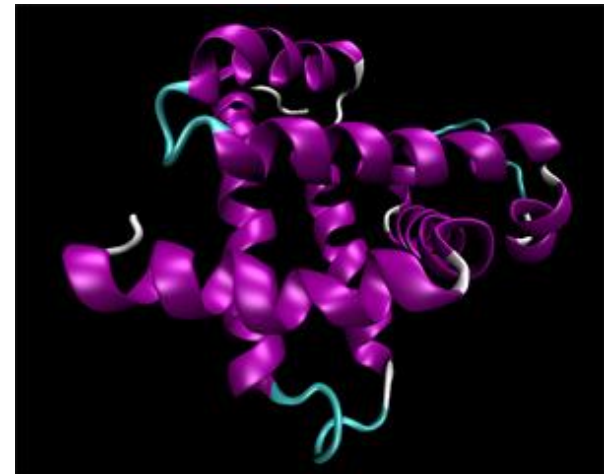
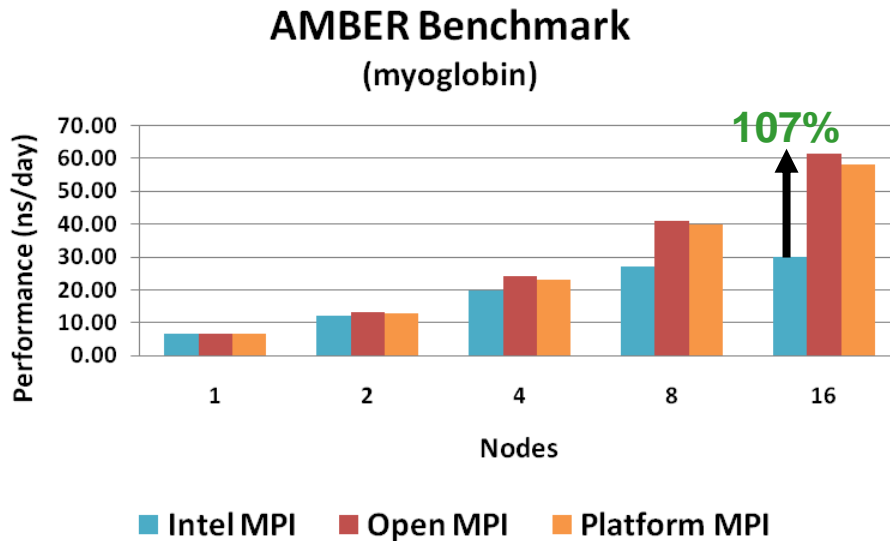
- **FCA 2.1 supports accelerations/offloading for**
 - MPI Barrier
 - MPI Broadcast
 - MPI AllReduce and Reduce
 - MPI AllGather and AllGatherv



Software Layers Overview



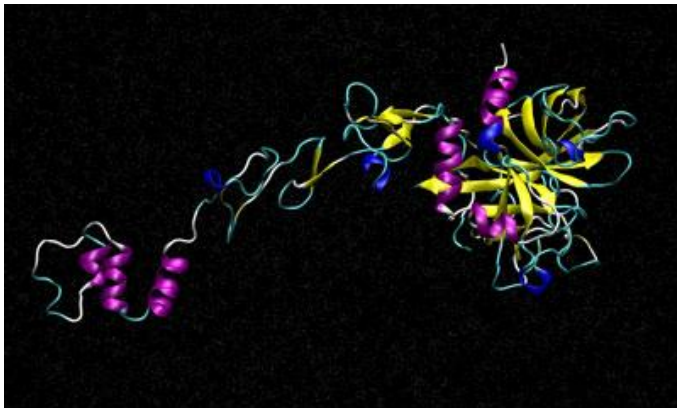
- **Dataset: myoglobin**
 - This is a Myoglobin simulation, 2492 atoms in Generalized Born implicit solvent, setup to have no cutoff for the nonbond terms and 15 angstroms for the calculation of GB radii
- **Platform MPI and Open MPI (with FCA) provides higher scalability**
 - Up to 107% higher performance
 - These MPI implementations enables higher performance as the cluster scales



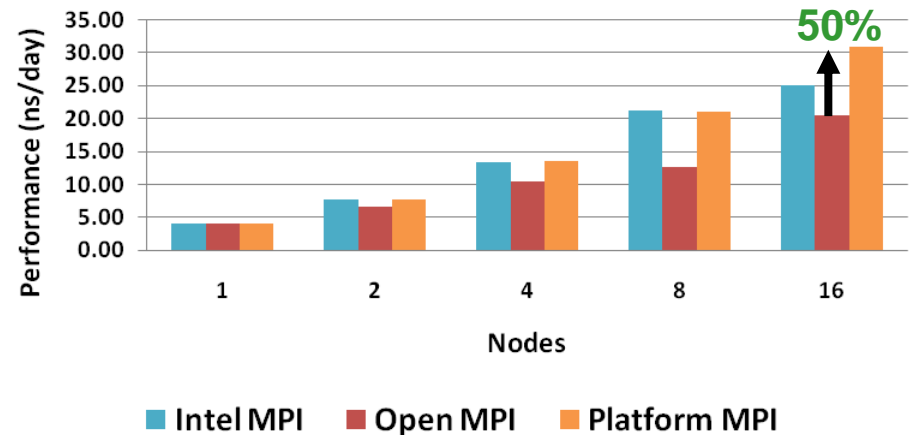
Higher is better

InfiniBand QDR

- **Dataset: FactorIX**
 - FactorIX in TIP3P Water box - 90,906 atoms, shake with a 2fs timestep, 8A cutoff
- **Platform MPI allows FactorIX dataset to achieve the highest performance**
 - Up to 50% performance gains by using Platform versus Open MPI
- **No use of collectives operations (see next slides)**



AMBER Benchmark
(FactorIX)

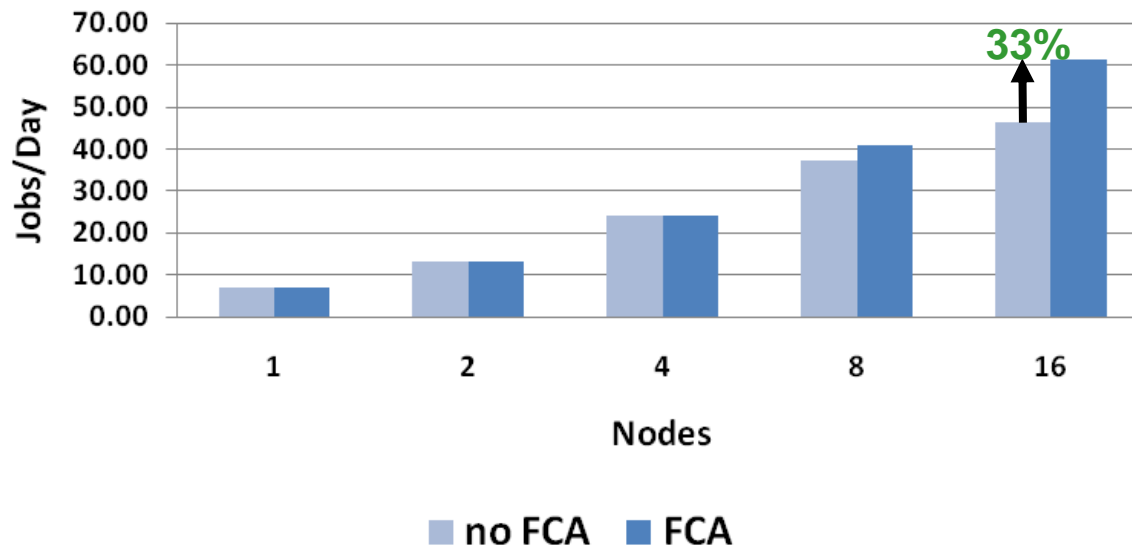


Higher is better

InfiniBand QDR

- **AMBER demonstrates the benefits of having MPI collectives offloads**
 - Mellanox FCA offloads MPI collectives to the InfiniBand HCA hardware
 - By freeing up CPU resources to the InfiniBand hardware, more CPU computation can be done
- **Mellanox FCA enables nearly 33% performance gain at 16 nodes / 192 cores**
 - Expect to continue to show advantage expected at higher node count / core count

AMBER Benchmark (myoglobin)



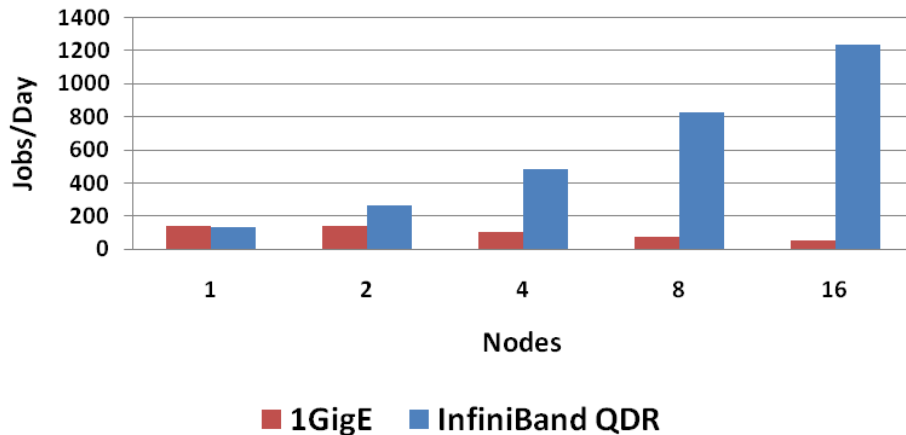
Higher is better

Open MPI

InfiniBand QDR

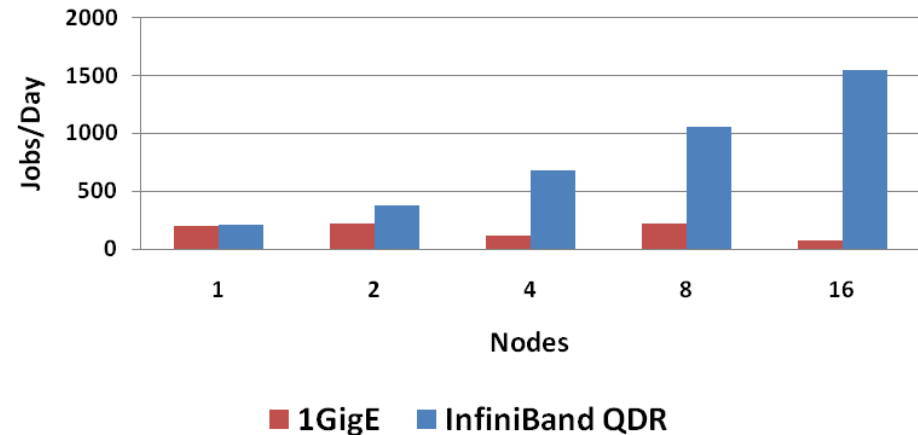
- **InfiniBand enables higher scalability and cluster productivity**
 - Provides the needed network infrastructure to deliver cluster scalability
 - 1GigE would not scale for more than 2 nodes
- **AMBER requires good network throughput for data communications**

**AMBER Benchmark
(myoglobin)**



Higher is better

**AMBER Benchmark
(FactorIX)**

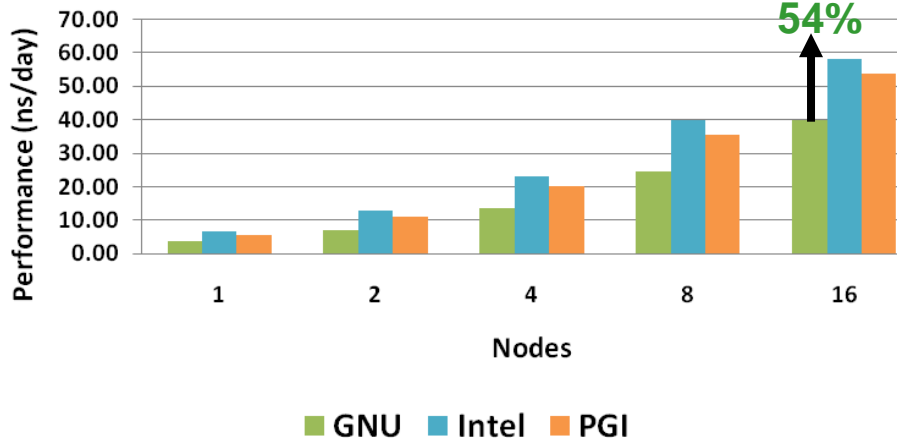


InfiniBand QDR

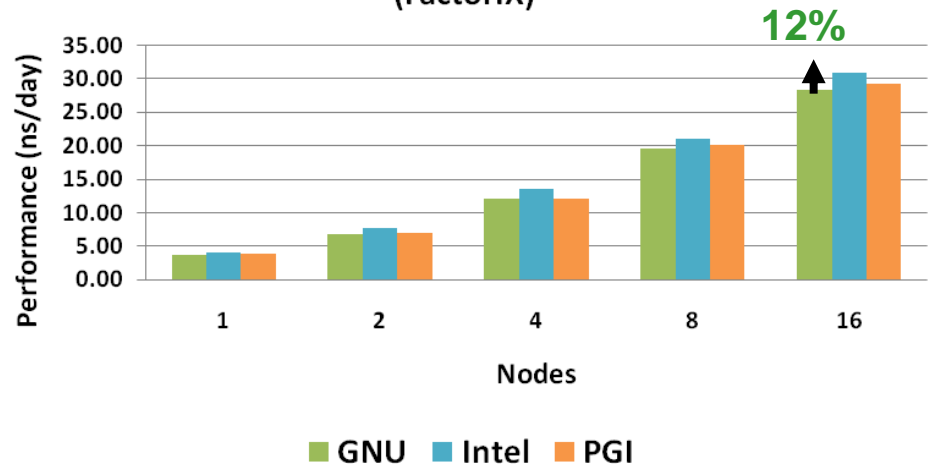
AMBER Performance – Compilers

- **Intel Compilers enables higher CPU utilization on both datasets**
 - Shows up to 54% better utilization with the myoglobin dataset
 - Shows around 12% gain in performance with the FactorIX dataset

**AMBER Benchmark
(myoglobin)**



**AMBER Benchmark
(FactorIX)**

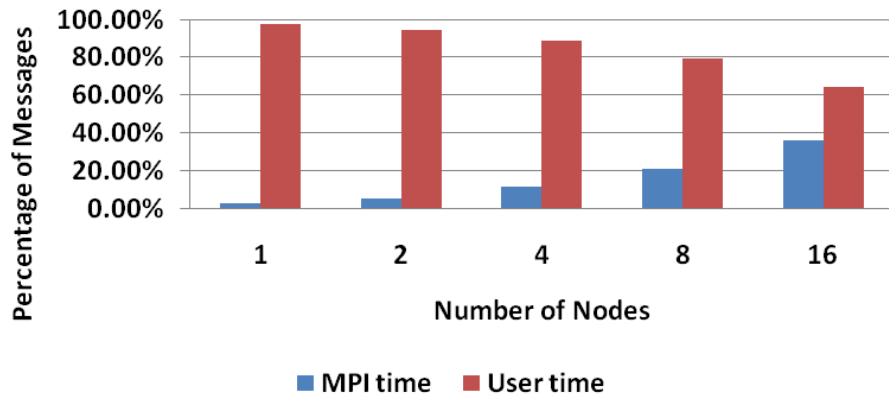


Higher is better

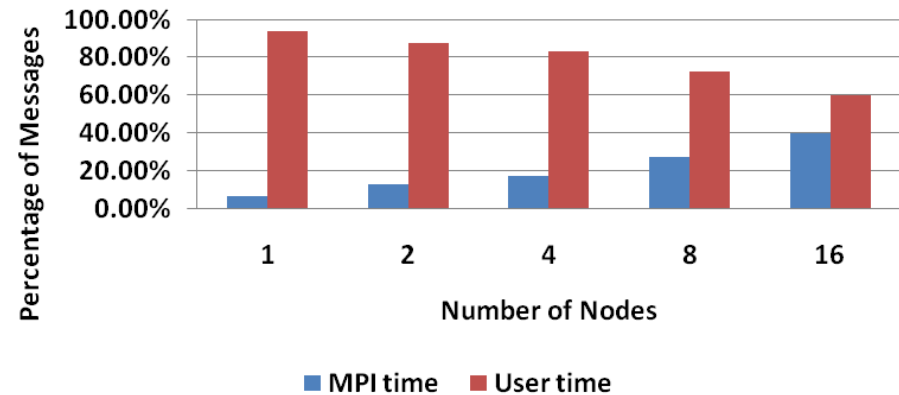
*Platform MPI
InfiniBand QDR*

- **Computation time dominates run time**
 - Reflects that more time spent on computation than communications
- **Communication percentage increases substantially as the cluster scales**

AMBER Profiling
(myoglobin)
MPI/User Time Ratio

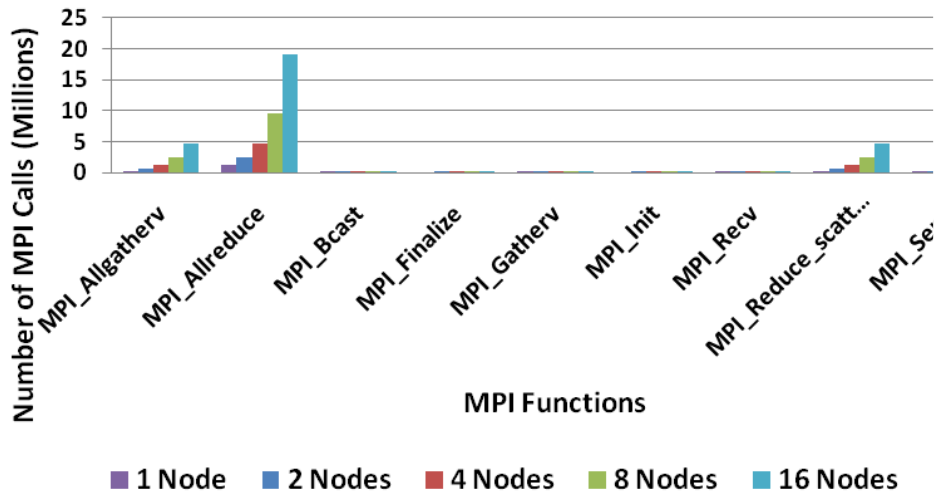


AMBER Profiling
(FactorIX)
MPI/User Time Ratio

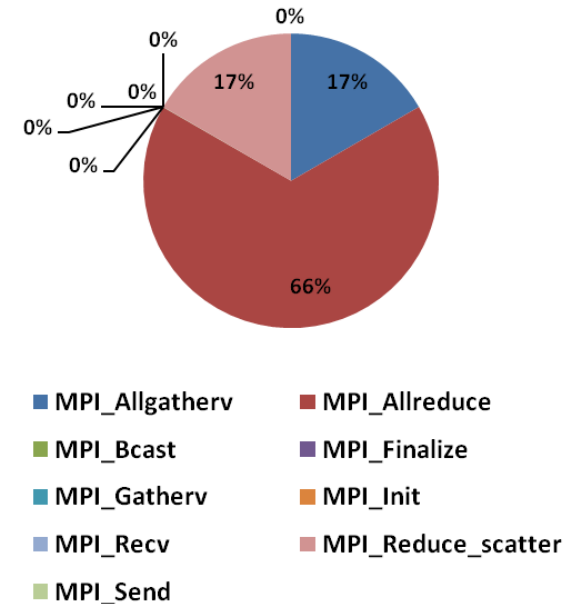


- **Dataset: Myoglobin**
- **The myoglobin uses a large fraction of MPI collective operations**
 - Up to 66% of the time is spent on MPI_Allreduce
- **Potential savings for MPI Collective Offloading with FCA**

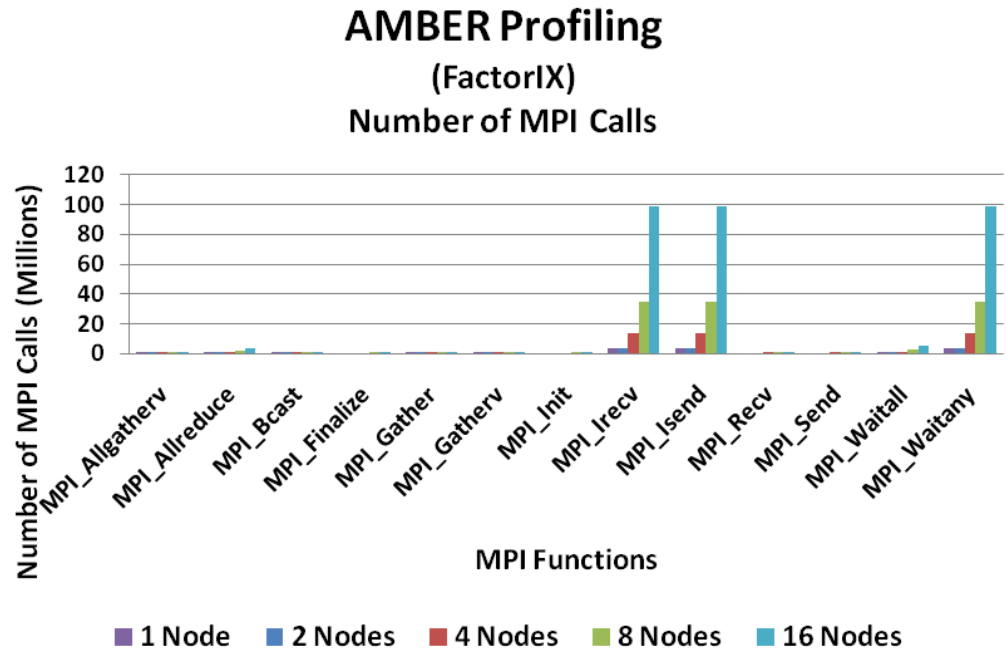
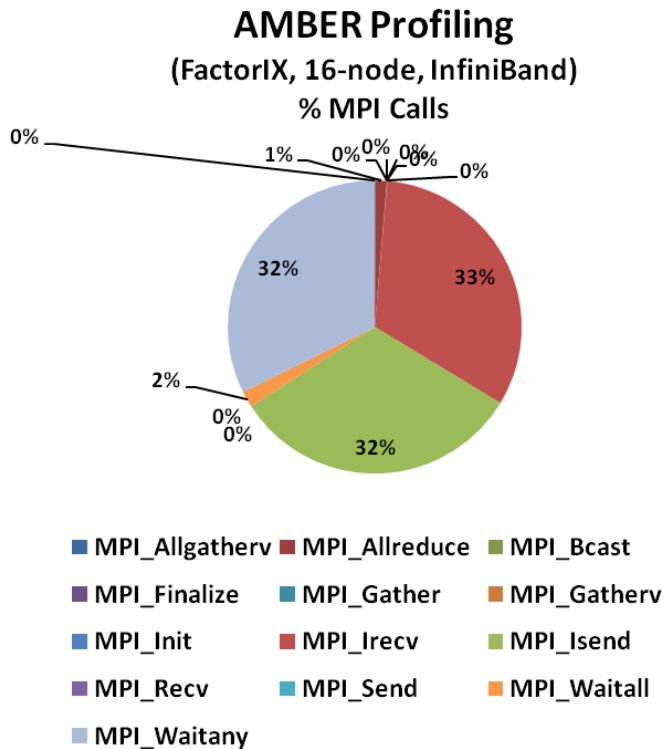
AMBER Profiling
(myoglobin)
Number of MPI Calls



AMBER Profiling
(myoglobin, 16-node, InfiniBand)
% MPI Calls



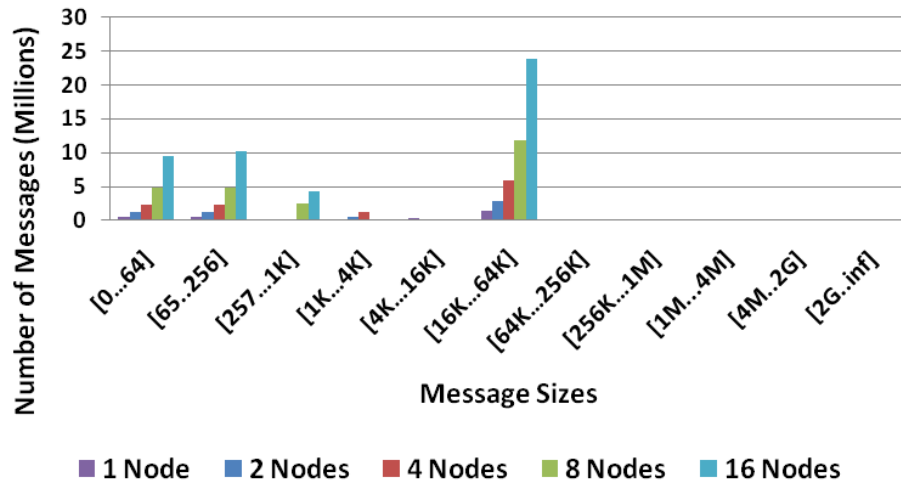
- **Dataset: FactorIX**
- **The FactorIX dataset uses a different set of MPI functions**
 - Majority of the calls are non-blocking MPI calls
 - MPI_Irecv (33%)
 - MPI_Isend (32%)
 - MPI_Waitany (32%)



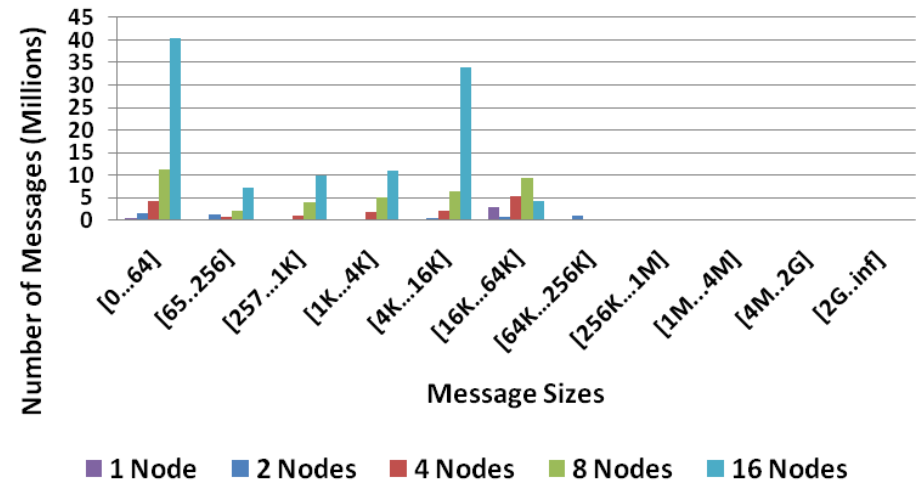
AMBER Profiling – MPI Message Size

- **For Myoglobin, majority of MPI messages are messages in the midrange**
 - In the range of 16K to 64K bytes
- **For FactorIX, Messages are scattered in the small and mid-ranges.**
 - Higher spikes are in the range of 0 to 64 bytes, and 4K to 16K bytes

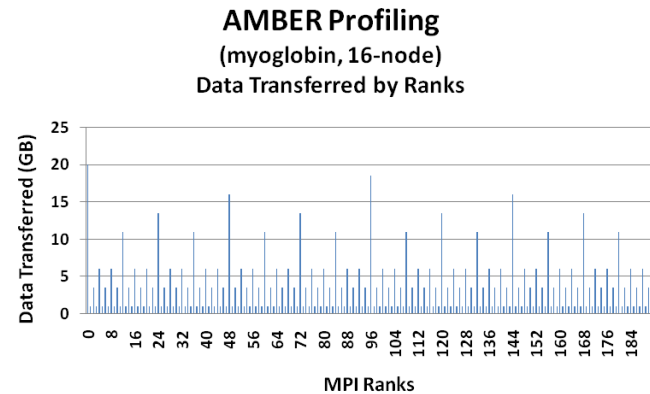
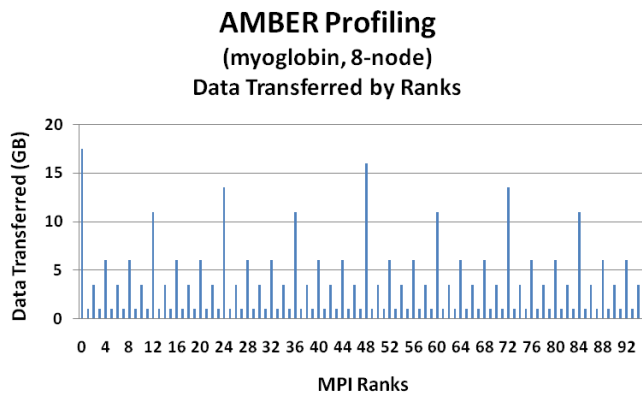
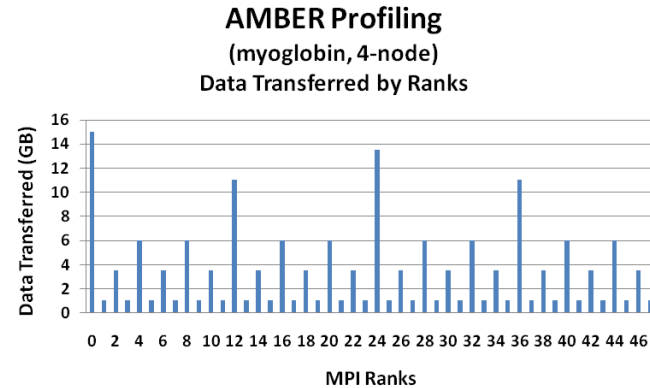
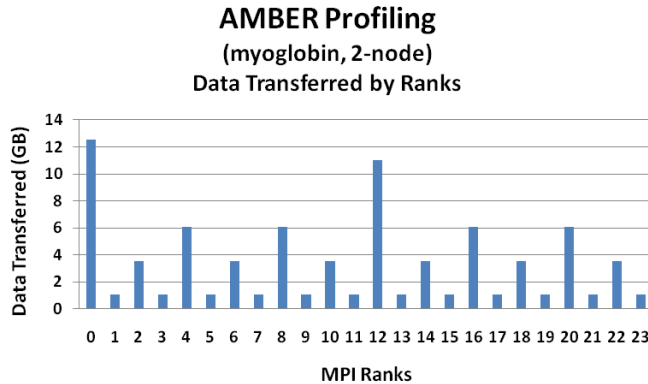
AMBER Profiling
(myoglobin)
MPI Message Sizes



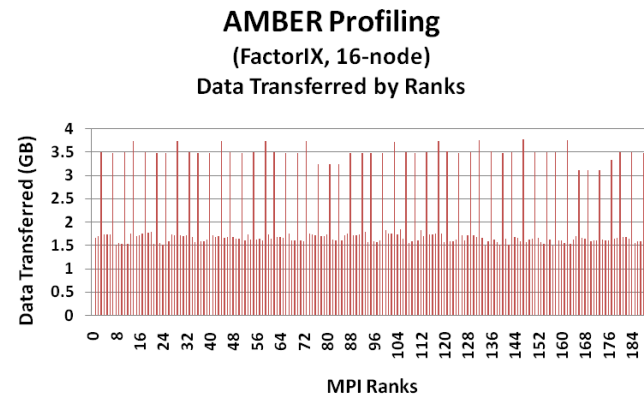
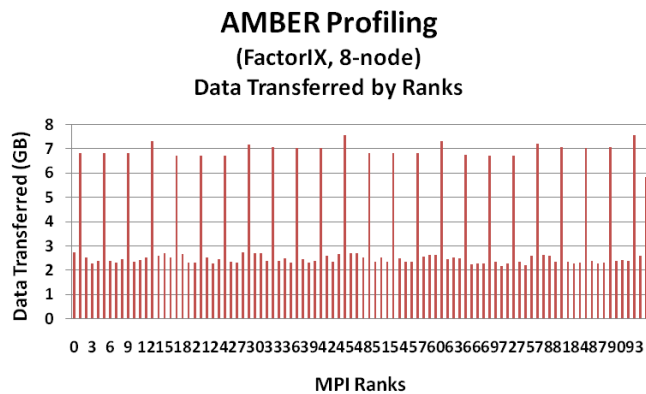
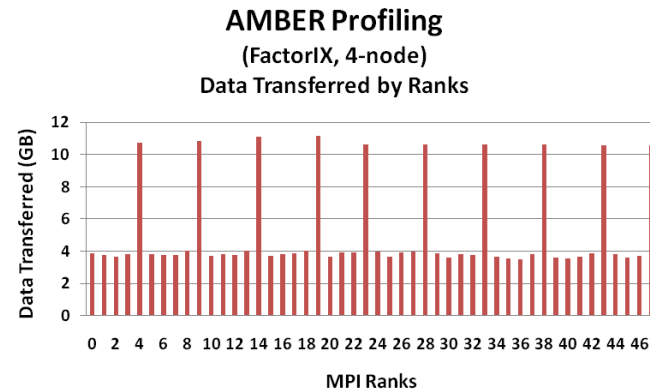
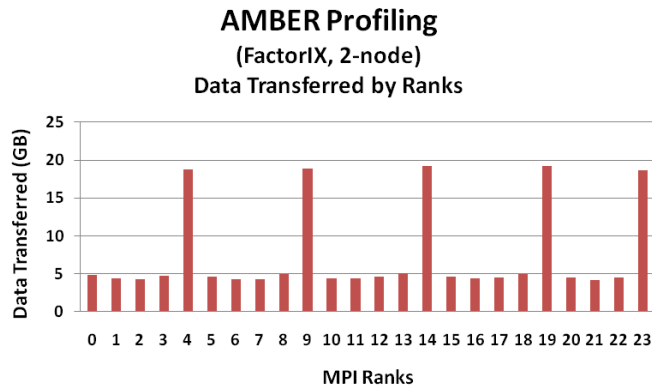
AMBER Profiling
(FactorIX)
MPI Message Sizes



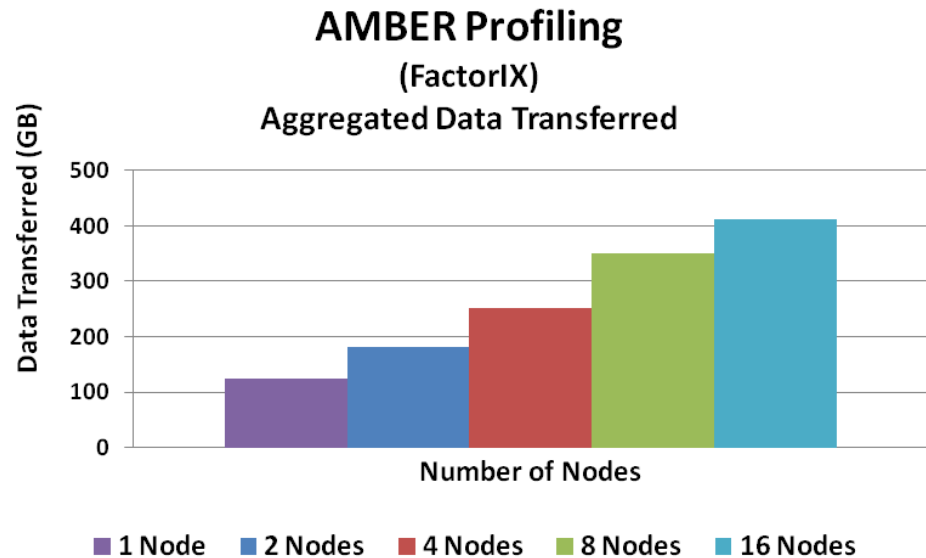
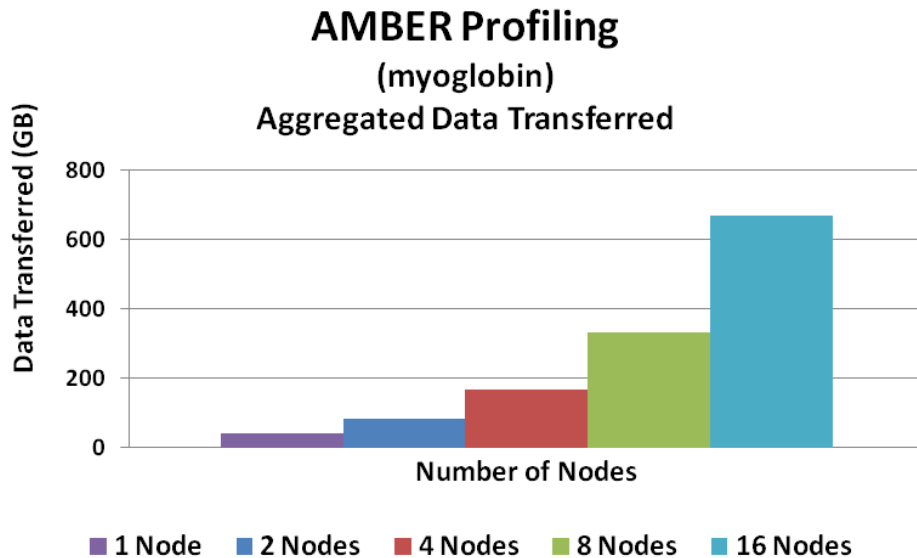
- **Myoglobin shows substantial data transfers between the compute nodes**
 - In the range of 1GB to 12GB per node
 - Shows data bifurcation in data communications



- **FactorIX shows spikes in data communications**
 - Shows spike in data communication for 1 in every 5 ranks
 - The data spikes drops from 19GB (on 1-node) to 3.5 GB (on 16-node)
 - Data transfers for all other MPI ranks lower as more nodes are being used in cluster



- **Aggregated data transfer refers to:**
 - Total amount of data being transferred in the network between all MPI ranks collectively
- **Both datasets shows huge data transfer takes place in the network**
 - The total data transfer grows exponentially for Myoglobin
 - The total data transfer increases proportionally for FactorIX



InfiniBand QDR

- **AMBER is a network intensive application that shows high communications**
- **InfiniBand shows better performance as more compute nodes are used**
- **Intel Compilers enables the highest CPU utilization**
- **Platform MPI shows good cluster scalability**
 - that allows maximizing the computation resource available
- **FCA allows offloading MPI collective operations to the InfiniBand hardware**
 - Which allows the CPU to focus on computation
- **Different MPI communication patterns are seen from the 2 datasets**
 - Myoglobin uses a majority of the MPI collective operations
 - FactorIX communicates using the MPI one-sided communications

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