HPC ADVISORY COUNCIL

NETWORK OF EXPERTISE

AMBER 11
Performance Benchmark and Profiling

July 2011

AMD

DELL

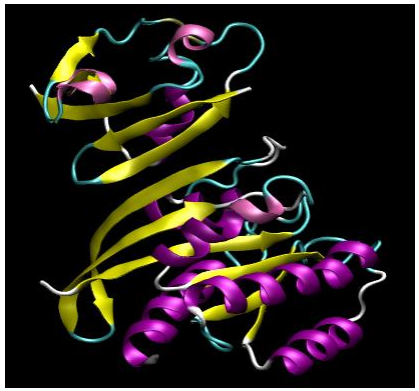
Mellanox TECHNOLOGIES

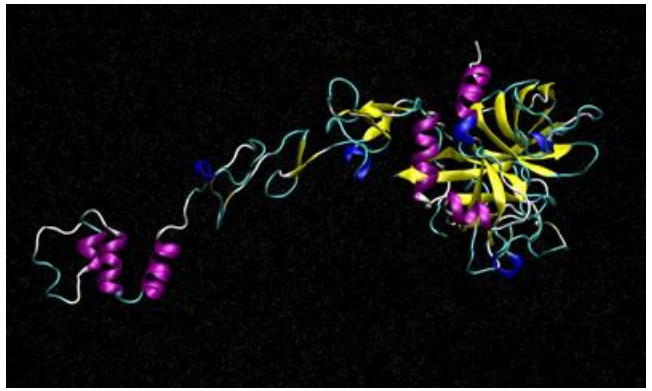
Note

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

AMBER Application

- **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 Fortan90 and C, and its GPU kernels are written with NVIDIA's CUDA API to achieve maximum GPU performance





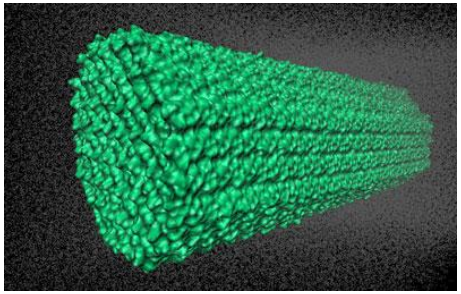
Objectives

- **The following was done to provide best practices**
  - AMBER performance benchmarking
  - Understanding AMBER communication patterns
  - Ways to increase AMBER productivity
  - Compilers and MPI libraries comparisons

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

Test Cluster Configuration

- Dell™ PowerEdge™ R815 11-node (528-core) cluster
- AMD Opteron™ 6174 Series processors (codenamed “Magny-Cour”) 12-cores @ 2.2 GHz
- 4 CPU sockets per server node
- Mellanox ConnectX-2 VPI adapters for 40Gb/s QDR InfiniBand and 10Gb/s Ethernet
- Mellanox MTS3600Q 36-Port 40Gb/s QDR InfiniBand switch
- Memory: 128GB memory per node DDR3 1333MHz
- OS: RHEL 5.5, MLNX-OFED 1.5.2 InfiniBand SW stack
- MPI: Open MPI 1.5.3 with KNEM 0.9.6, Platform MPI 8.1.1
- Compilers: PGI 10.9, GNU Compilers 4.1.2
- Application: AMBER 11 (PMEMD), AmberTools 1.5
- Benchmark workload:
  - Cellulose_production_NVE_256_128_128 (408,609 atoms)



About Dell™ PowerEdge™ R815 11-node cluster

- **HPC Advisory Council Test-bed System**

- **New 11-node 528 core cluster - featuring Dell PowerEdge™ R815 servers**
  - Replacement system for Dell PowerEdge SC1435 (192 cores) cluster system following 2 years of rigorous benchmarking and product EOL
    - System to be redirected to explore HPC in the Cloud applications

- **Workload profiling and benchmarking**
  - Characterization for HPC and compute intense environments
  - Optimization for scale, sizing and configuration and workload performance
  - Test-bed Benchmarks
    - RFPs
    - Customers/Prospects, etc
  - ISV & Industry standard application characterization
  - Best practices & usage analysis

About Dell PowerEdge™ Platform Advantages

Best of breed technologies and partners

Combination of AMD Opteron™ 6000 Series platform and Mellanox ConnectX InfiniBand on Dell HPC. Solutions provide the ultimate platform for speed and scale

- Dell PowerEdge R815 system delivers 4 socket performance in dense 2U form factor
- Up to 48 core/32DIMMs per server – 1008 core in 42U enclosure

Integrated stacks designed to deliver the best price/performance/watt

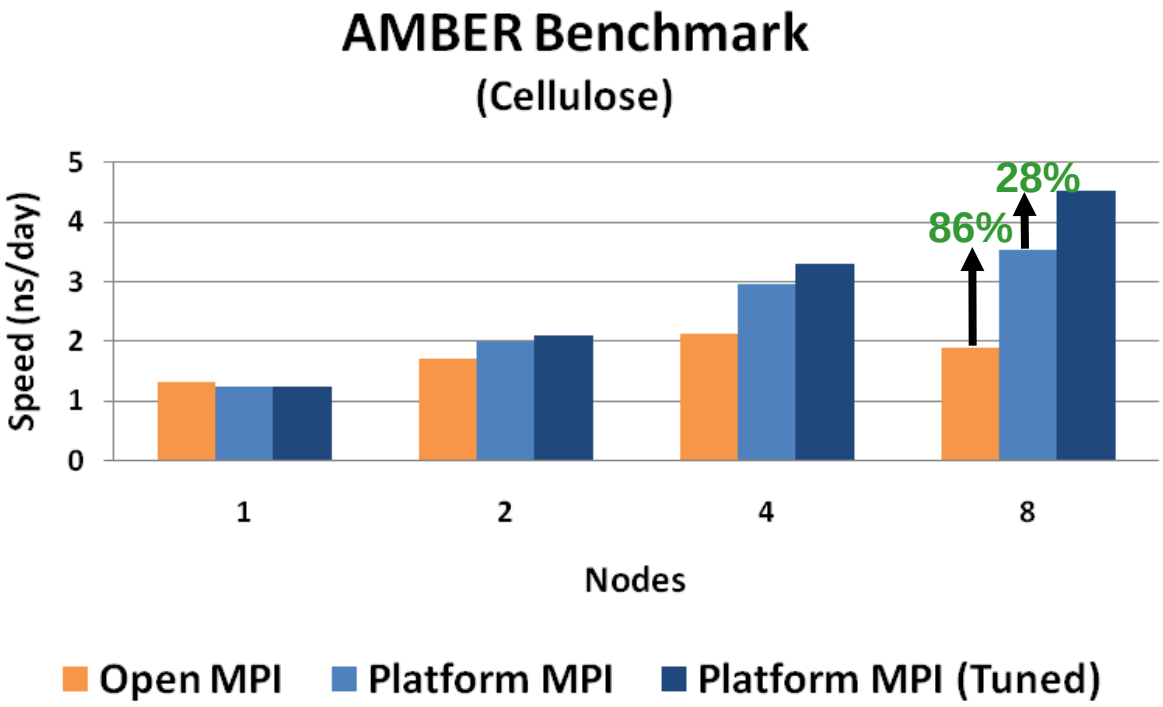
- 2x more memory and processing power in half of the space
- Energy optimized low flow fans, improved power supplies and dual SD modules

Optimized for long-term capital and operating investment protection

- System expansion
- Component upgrades and feature releases

AMBER Performance – MPI

- **Platform MPI enables better scalability than Open MPI**
  - Seen up to 86% better performance with Platform MPI
- **Tuned Platform MPI provides better performance for InfiniBand**
  - Seen an speed improvement of up to 28% at 8-node
  - Tuning parameters used: -cpu_bind -e MPI_RDMA_MSGSIZE=65536,65536,4194304 -e MPI_RDMA_NSRQRECV=2048 -e MPI_RDMA_NFRAGMENT=128

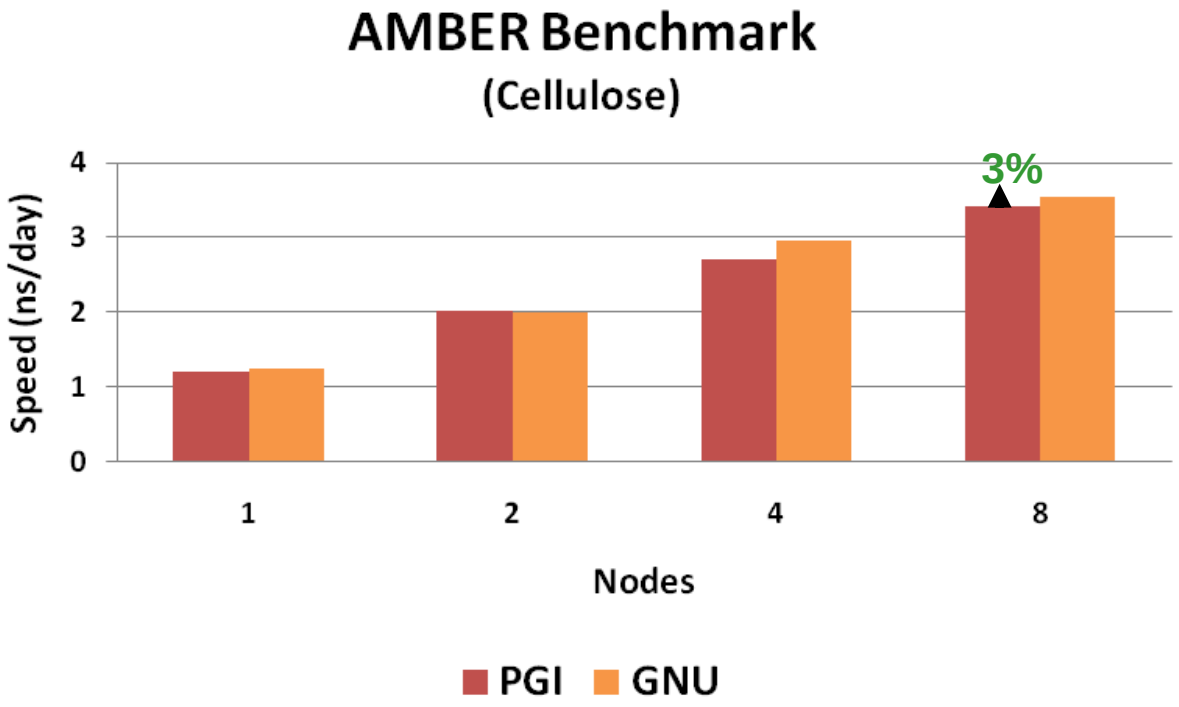


Higher is better

48 Cores/Node

AMBER Performance – Compilers

- **Executable generated by the GNU compilers runs slightly faster**
  - Up to 3% faster is seen than with PGI
- **Using the default optimization and linker flags:**
  - PGI: “-fast -O3 -fastsse”
  - GNU: “-O3”



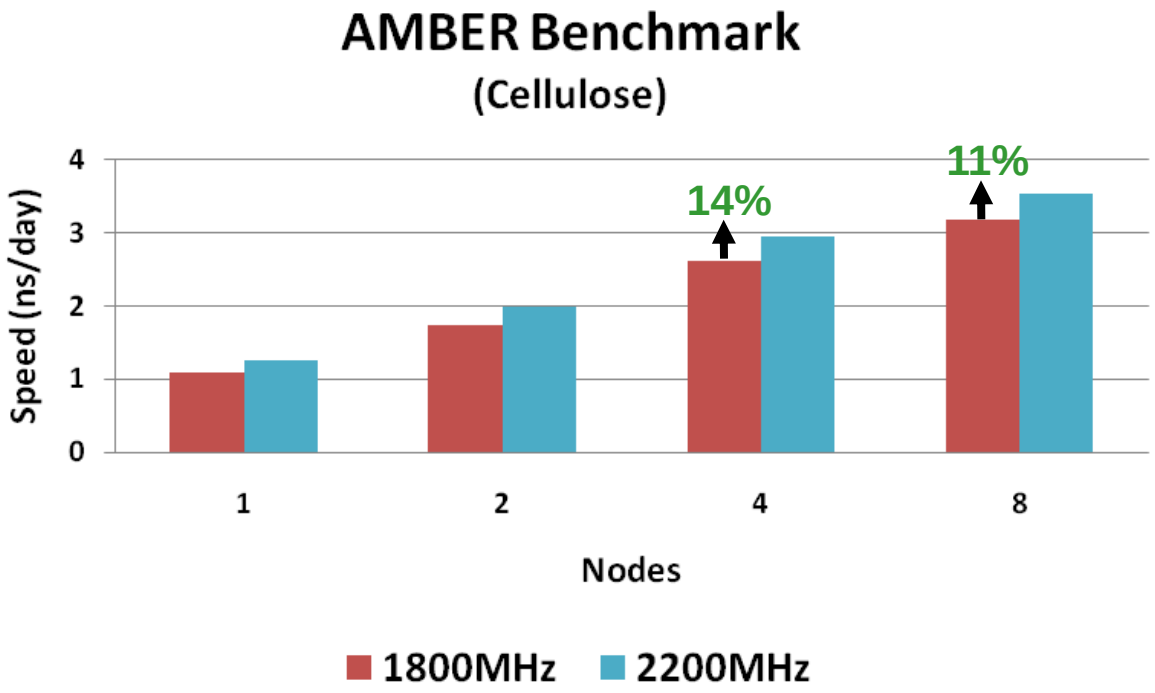
Higher is better

Platform MPI

48 Cores/Node

AMBER Performance – CPU Frequency

- **Higher CPU core frequency enables higher job performance**
  - Up to 11-14% better job performance between 2200MHz vs 1800MHz
  - Shows the performance is somewhat affected by the change in the CPU frequency, and the percentage of Computation versus MPI Communication

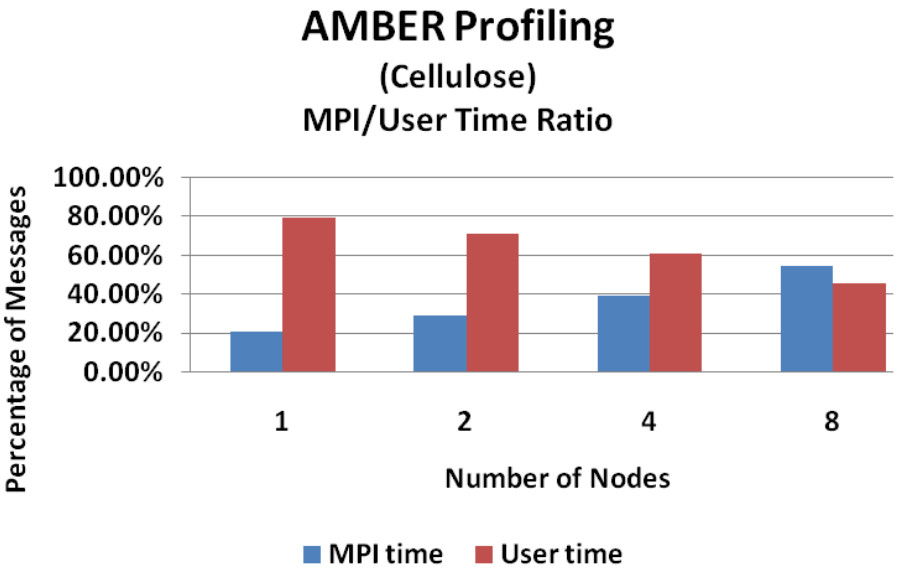


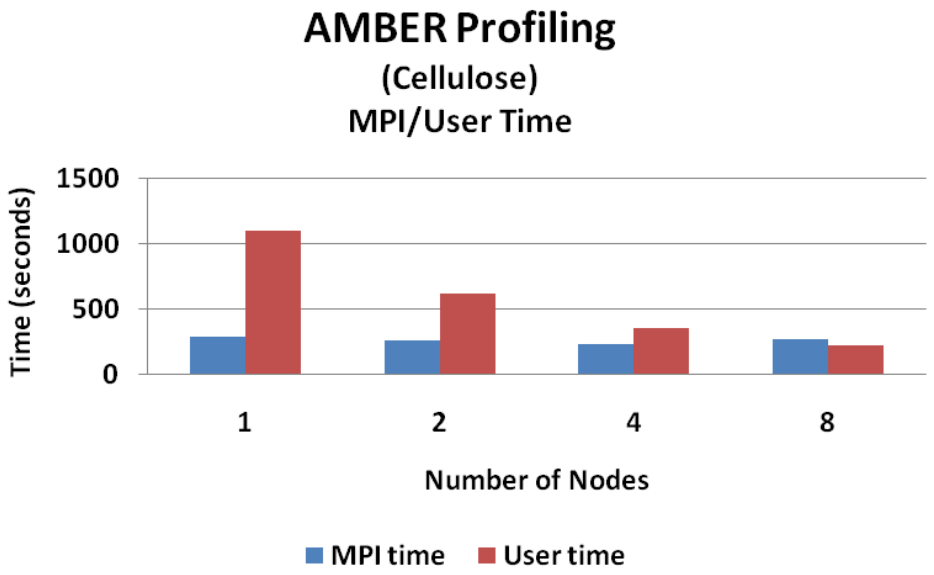
Higher is better

48 Cores/Node

AMBER Profiling – MPI/User Time Ratio

- **Large data communications are seen**
  - Shows heavy data communications between parallel tasks
  - More communications than computation happens between 4 to 8 node
- **MPI time stays constant while CPU time reduces as more nodes in cluster**
  - Demonstrate the importance of the interconnect to handle addition network throughput



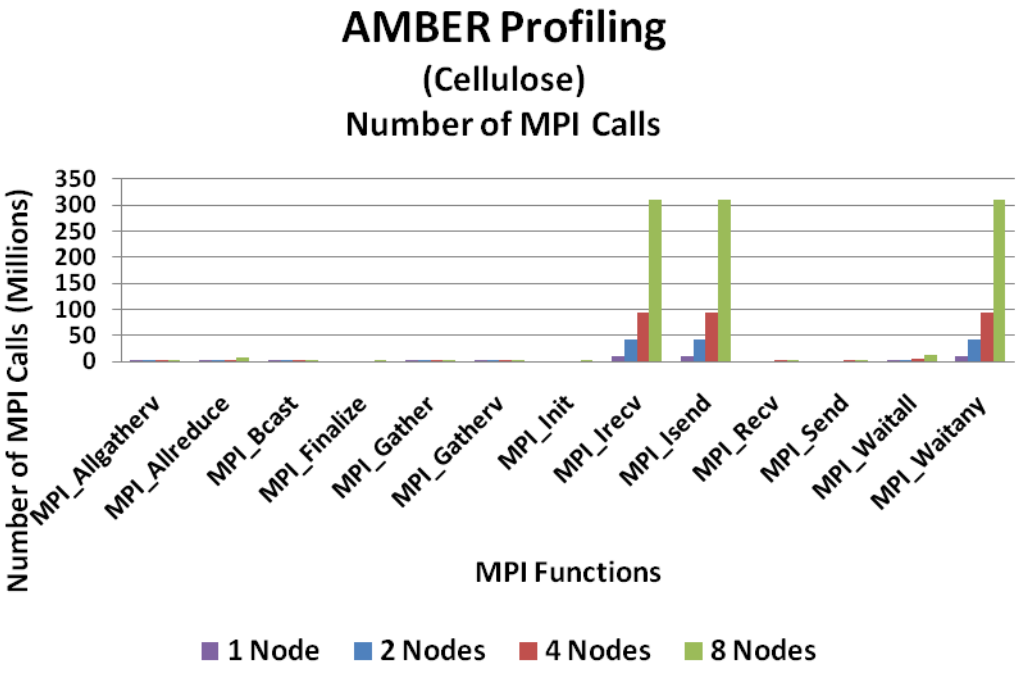


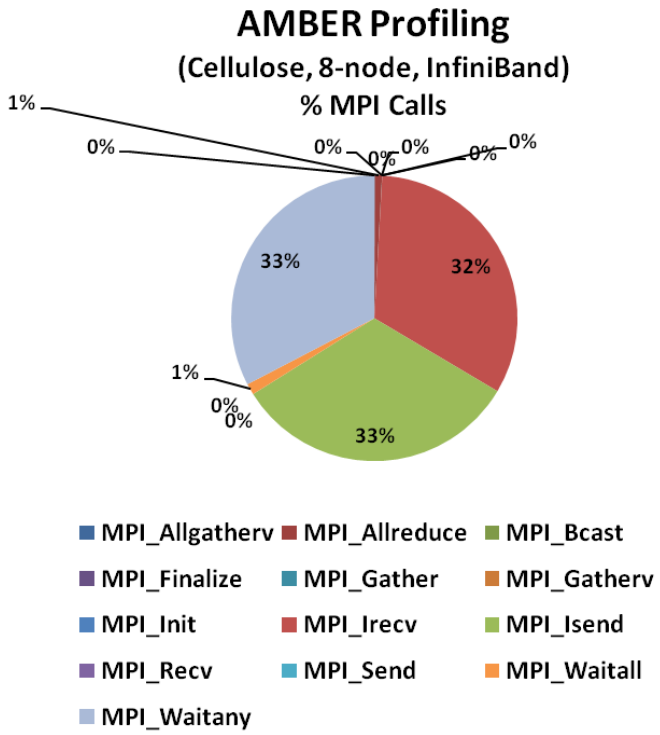
InfiniBand QDR

48 Cores/Node

AMBER Profiling – Number of MPI Calls

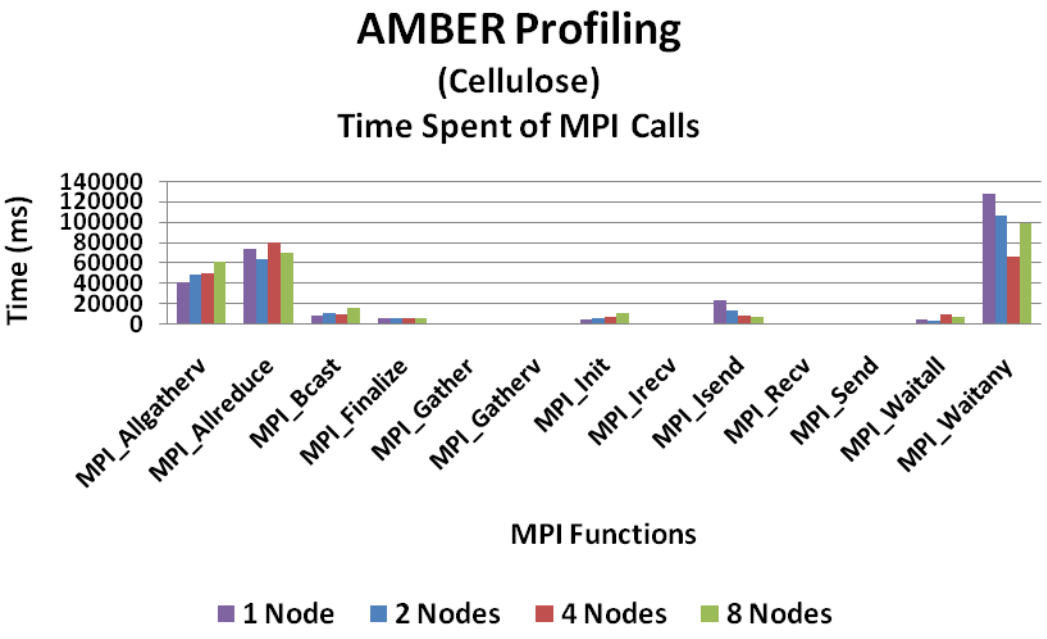
- **The most used MPI functions are MPI_Isend, MPI_Irecv, MPI_Waitany**
  - These non-blocking MPI calls allows computation while communications take place
  - Each of these dominates as a third of MPI calls used on a 8-node job
- **The number of calls accelerates rapidly as the cluster scales**

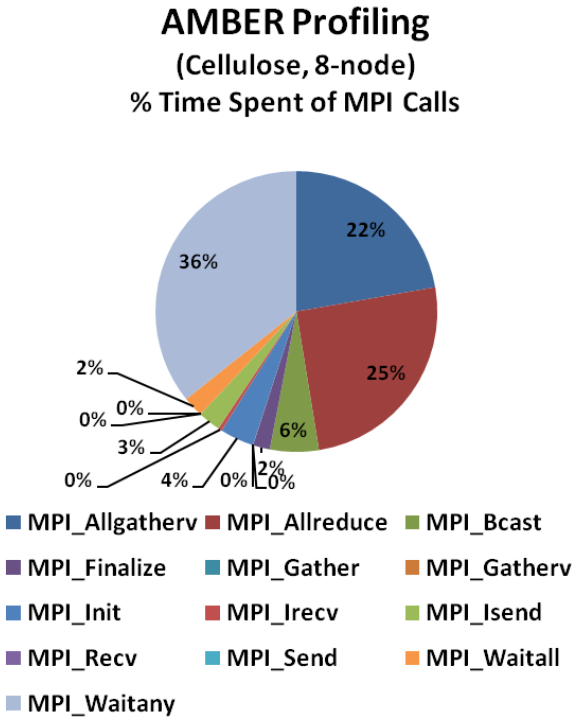




AMBER Profiling – Time Spent of MPI Calls

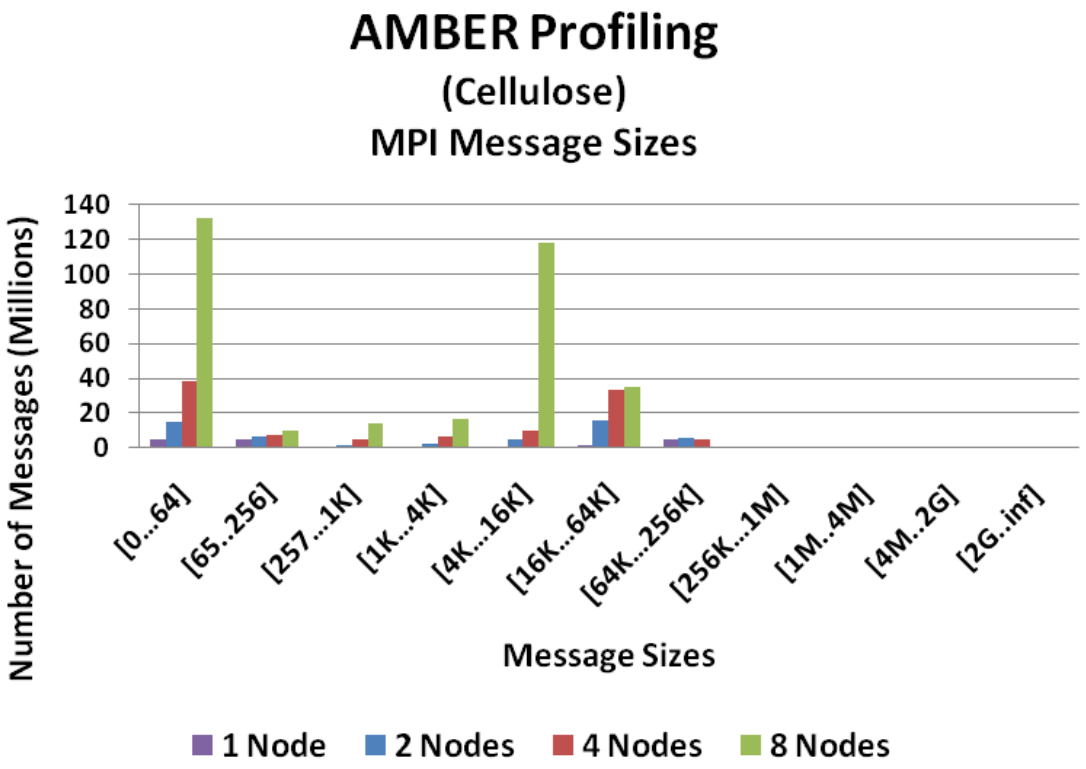
- **The time in communications is taken place in the following MPI functions:**
  - MPI_Waitany (36%)
  - MPI_Allreduce (25%)
  - MPI_Allgather (22%)
  - MPI_Bcast (6%)





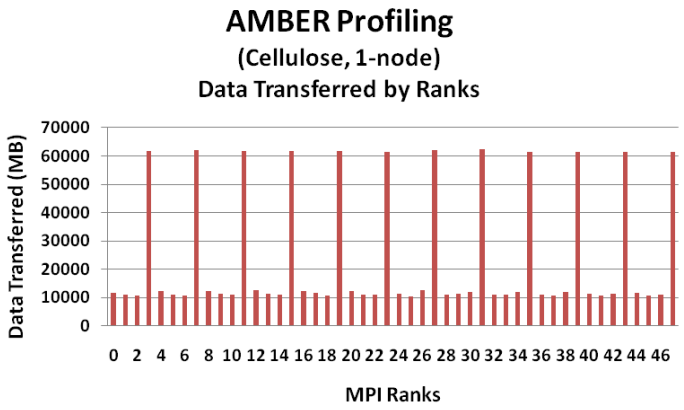
AMBER Profiling – MPI Message Sizes

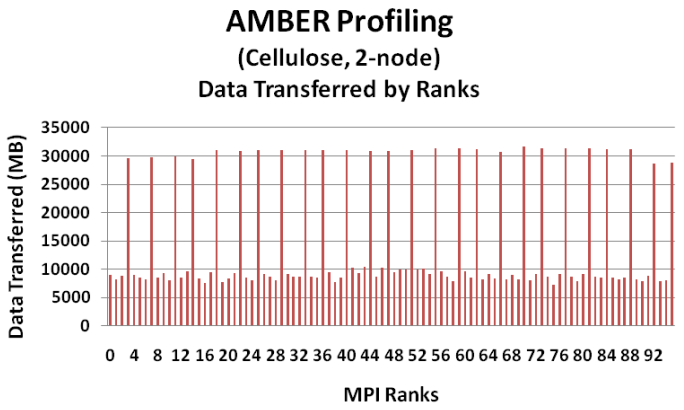
- **Majority of the MPI messages are small and median message sizes**
  - In the ranges of less than 64 bytes and between 4K and 16K

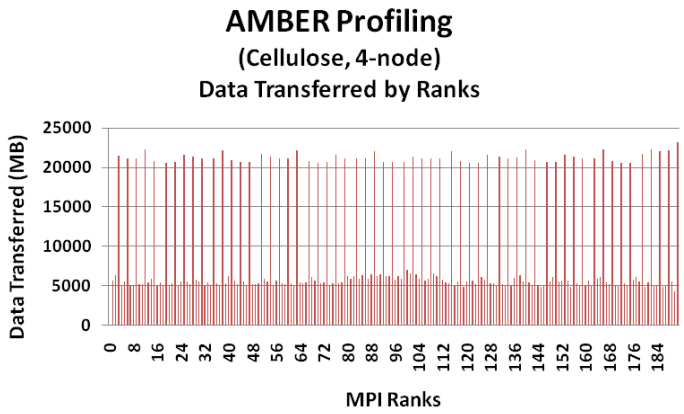


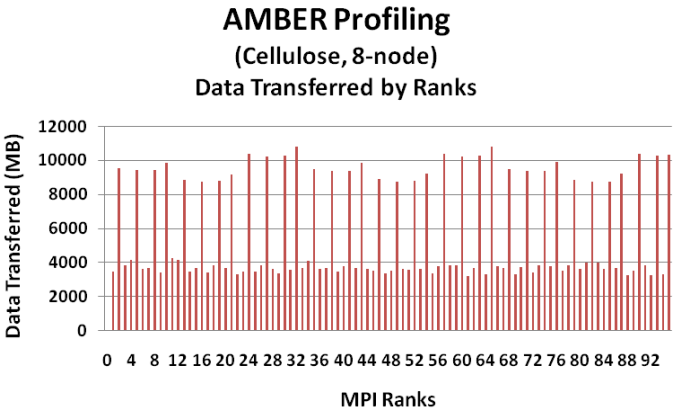
AMBER Profiling – Data Transfer By Process

- **Data transferred to each MPI rank is consistent for any number of processes**
  - Shows large amount of data transfers happened
  - Amount of data transfer to each rank is reduced as more nodes are in the job



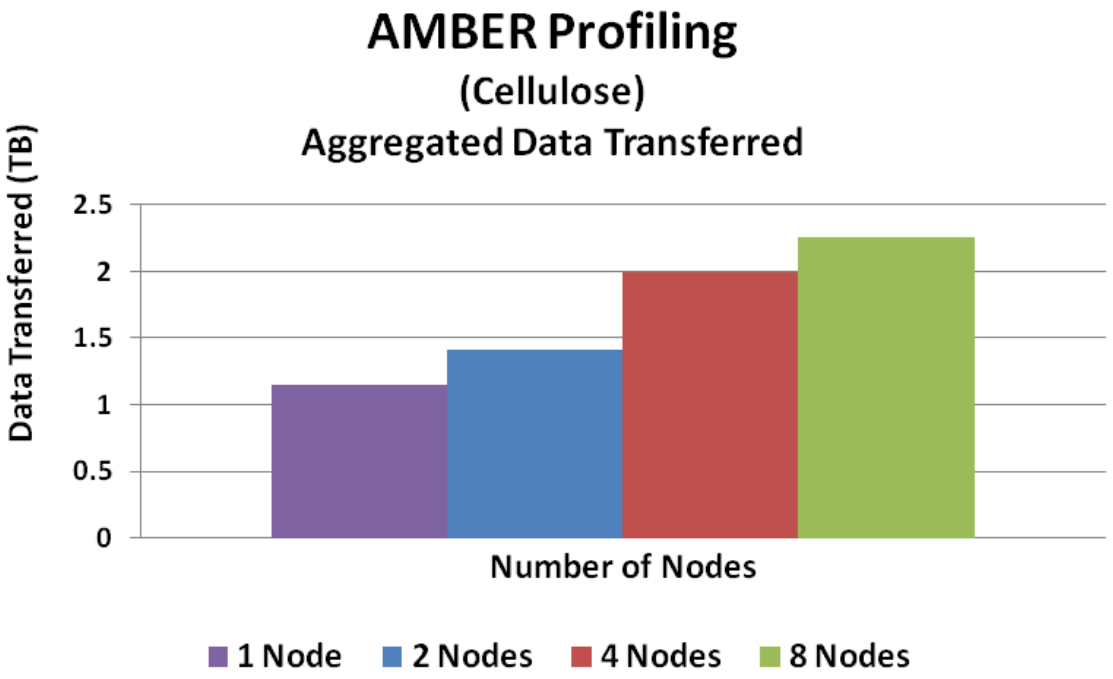






AMBER Profiling – Aggregated Data Transfer

- **Aggregated data transfer refers to:**
  - Total amount of data being transferred in the network between all MPI ranks collectively
- **The total data transfer increases steadily as the cluster scales**
  - Hugh amount of data being sent and received across the network
  - As a compute node being added, more data communications will take place



InfiniBand QDR

Summary

- **AMBER is a compute and data communications intensive application**
  - Which has a high demand for CPU power and good network throughput
  - Using InfiniBand enables good scalability by spreading workload to compute nodes

- **MPI:**
  - Better scalability is seen with Platform MPI than with Open MPI

- **CPU:**
  - Shows higher job productivity when using CPU with higher core frequency

- **MPI Communications:**
  - Large data communications take place between MPI processes

Thank You

HPC Advisory Council