MrBayes
Performance Benchmark and Profiling

May 2011

Note

- **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**
  - MrBayes performance overview
  - Understanding MrBayes communication patterns
  - Ways to increase MrBayes productivity
  - MPI libraries comparisons

- **For more info please refer to**
  - http://www.dell.com
  - http://www.intel.com
  - http://www.mellanox.com
  - http://mrbayes.csit.fsu.edu

MrBayes Application

- **MrBayes**
  - An application for the Bayesian estimation of phylogeny
- **Phylogeny is the history of the evolution of species and group**
  - Especially in reference to lines of descent and relationships among groups of organisms
- **Bayesian inference of phylogeny**
  - Based upon a quantity called the posterior probability distribution of trees
  - It is the probability of a tree conditioned on the observations
- **The conditioning is accomplished using Bayes's theorem**
- **MrBayes uses a simulation technique**
  - Called Markov chain Monte Carlo (or MCMC)
  - It is used to approximate the posterior probabilities of trees

Test Cluster Configuration

- **Dell™ PowerEdge™ M610 38-node (456-core) cluster**
  - Six-Core Intel X5670 @ 2.93 GHz CPUs
  - Six-Core Intel X5675 @ 3.06 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**
- **Compiler: Intel Compiler 11.1, Intel MKL 10.1**
- **MPI: Intel MPI 4, Open MPI 1.5.3, Platform MPI 8.0.1**
- **Application: MrBayes v3.1.2 (for MPI) and v3.1.2h (for Hybrid)**
- **Benchmark dataset:**
  - DNA for primates (12 taxa, 898 characters in each aligned sequence)

About Intel® Cluster Ready

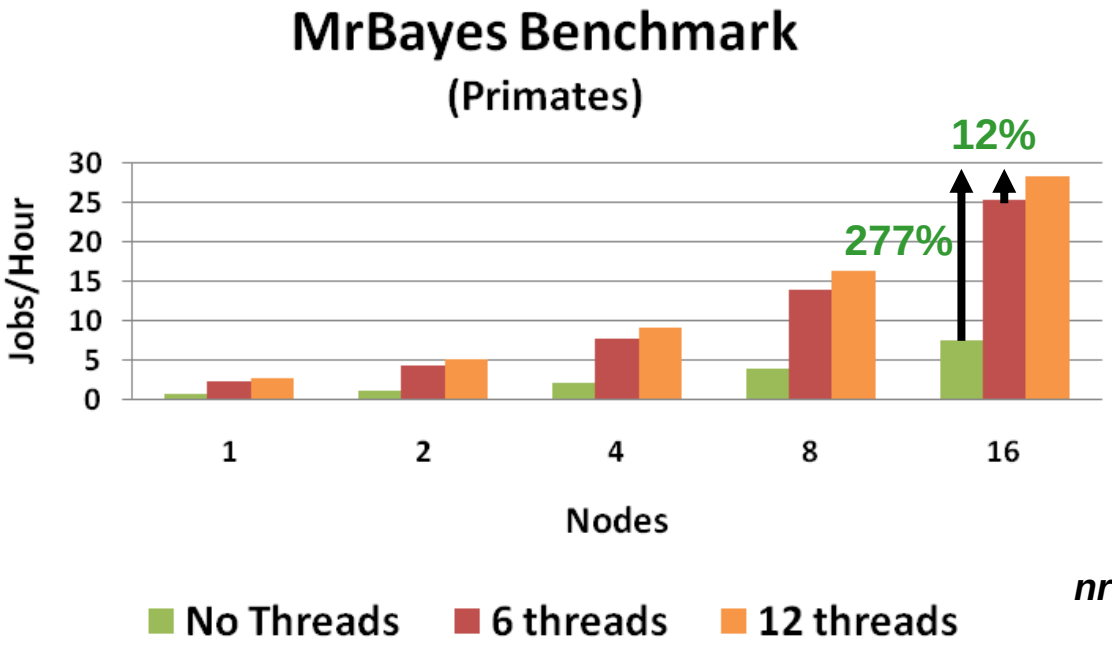
- **Intel® Cluster Ready systems make it practical to use a cluster to increase your simulation and modeling productivity**
  - Simplifies selection, deployment, and operation of a cluster
- **A single architecture platform supported by many OEMs, ISVs, cluster provisioning vendors, and interconnect 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

About Dell PowerEdge Servers

- **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

MrBayes Performance – MPI, OpenMP Hybrid

- **MPI version requires adjusting the chains and runs in the data file**
  - Which affects overall job size
  - The job size is determined by the runtime parameters (nruns=2, nchains=8)
- **Hybrid version runs with MPI and OpenMP threads**
  - Hybrid version is available separately in the hybrid version
- **Increasing the number of OpenMP threads do not affect overall job size**



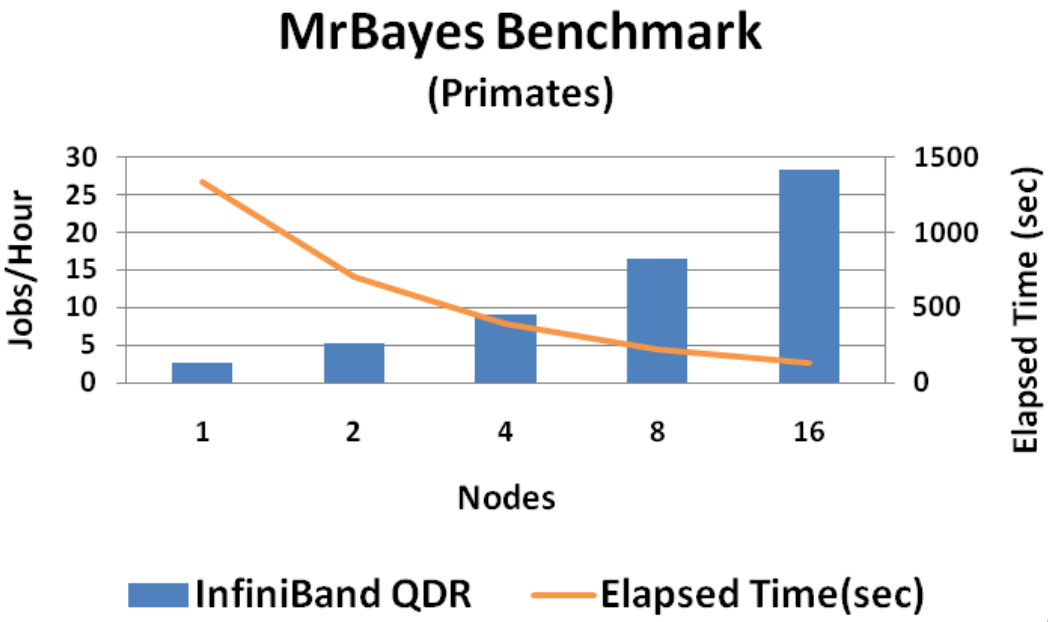
nruns=2,nchains=8

InfiniBand QDR

Higher is better

MrBayes Performance – Interconnect

- **InfiniBand enables higher scalability and cluster productivity**
  - Provides the needed network infrastructure to deliver cluster scalability
- **Inter-node parallelism is achieved by adding compute nodes**
  - Job runtime is reduced at the same pace as more compute nodes are added
  - The number of runs and chains remain constant for all test runs

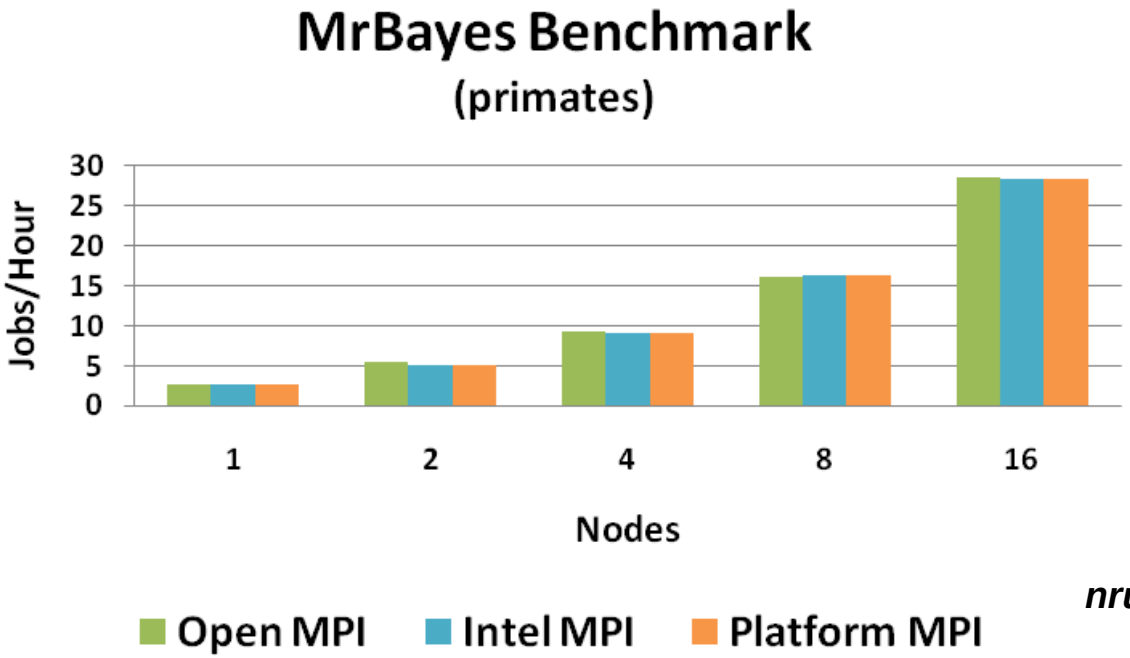


12 OpenMP Threads

nruns=2,nchains=8

MrBayes Performance – MPI

- **All MPI implementations show essentially the same level of performance**
  - The comparison is using 1 MPI process per node with 12 OpenMP threads
- **Processor binding needs to be disabled with Open MPI**
  - Do not use (--bind-to-core) as it can hurt performance when using hybrid (MPI + OpenMP)



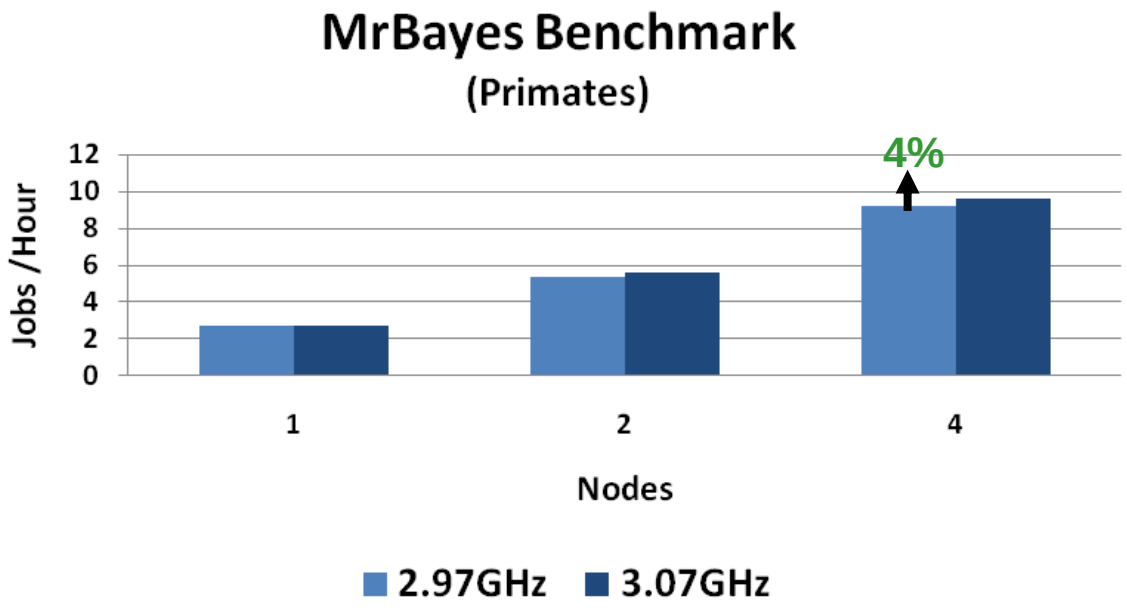
nruns=2,nchains=8

InfiniBand QDR

Higher is better

MrBayes Performance – CPU Frequencies

- **Higher CPU frequency provides some advantage to job productivity**
  - Seen a 4% increase in job productivity by using CPUs with 3.07GHz vs 2.93GHz



12 OpenMP Threads

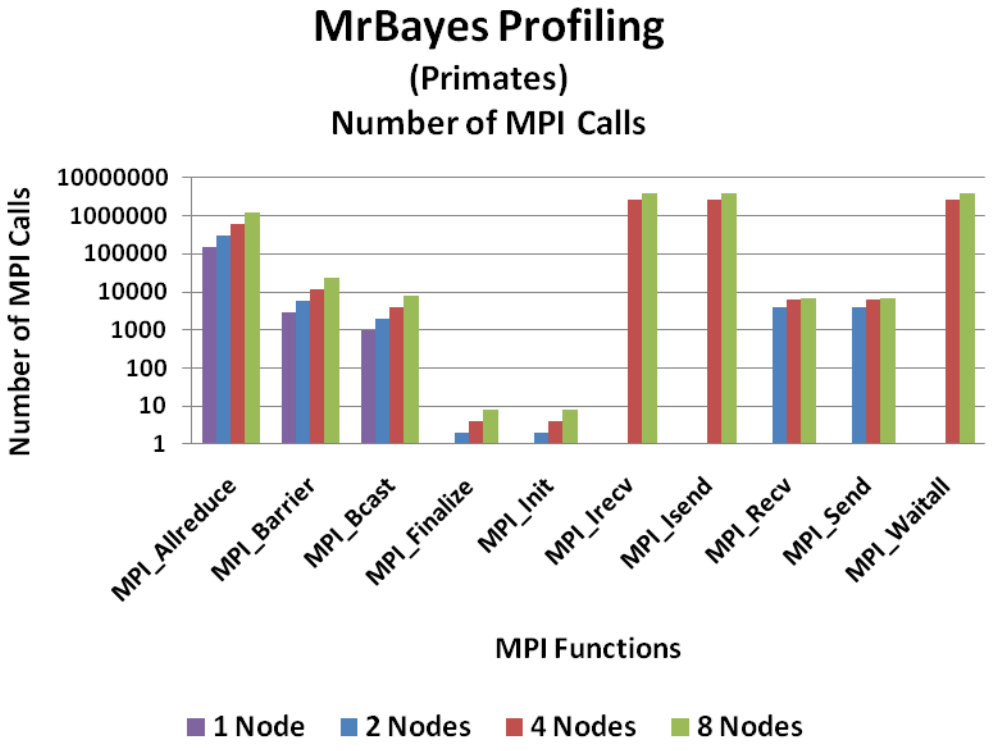
nruns=2,nchains=8

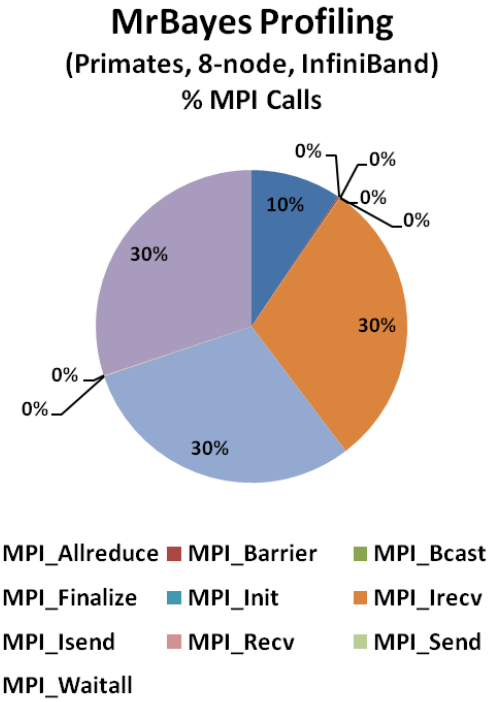
Higher is better

InfiniBand QDR

MrBayes Profiling – Number of MPI Calls

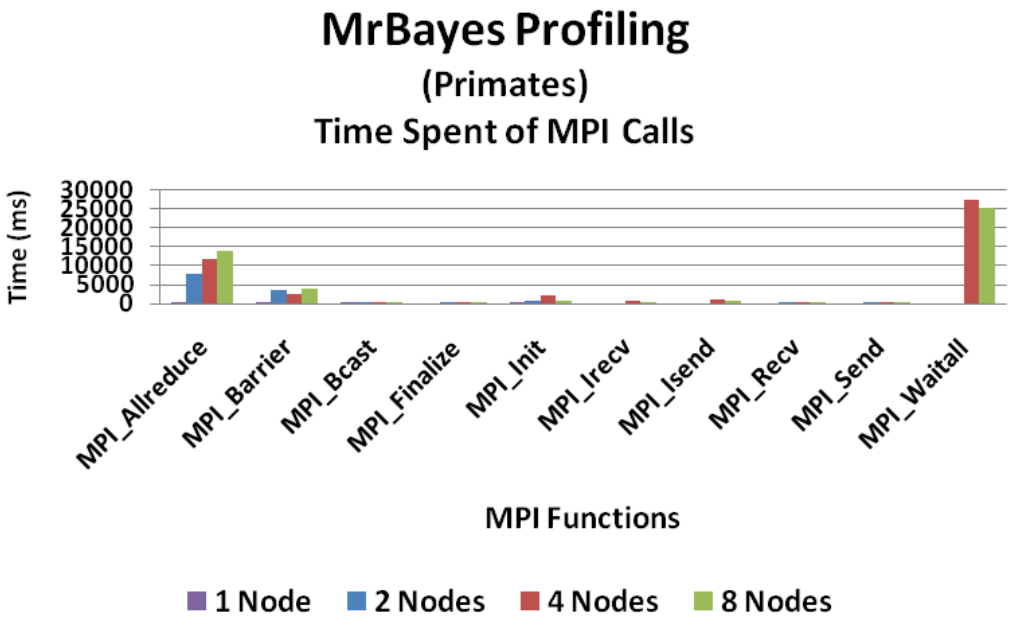
- **MrBayes shows a large number of MPI calls for data communications**
  - Uses in MPI collectives (MPI_Allreduce, MPI_Bcast)
  - Uses point-to-point communications (MPI_Send, MPI_Recv) starting with 2 nodes
  - Uses non-blocking communications (Isend, Irecv, Waitall) starting with 4 nodes

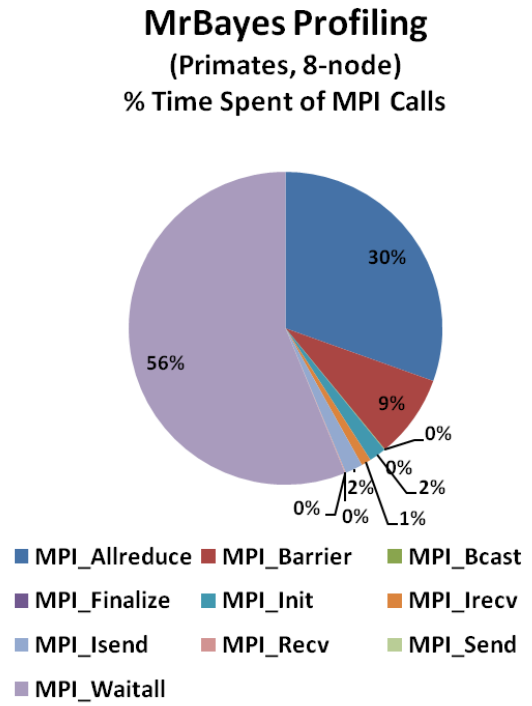




MrBayes Profiling – Time Spent by MPI

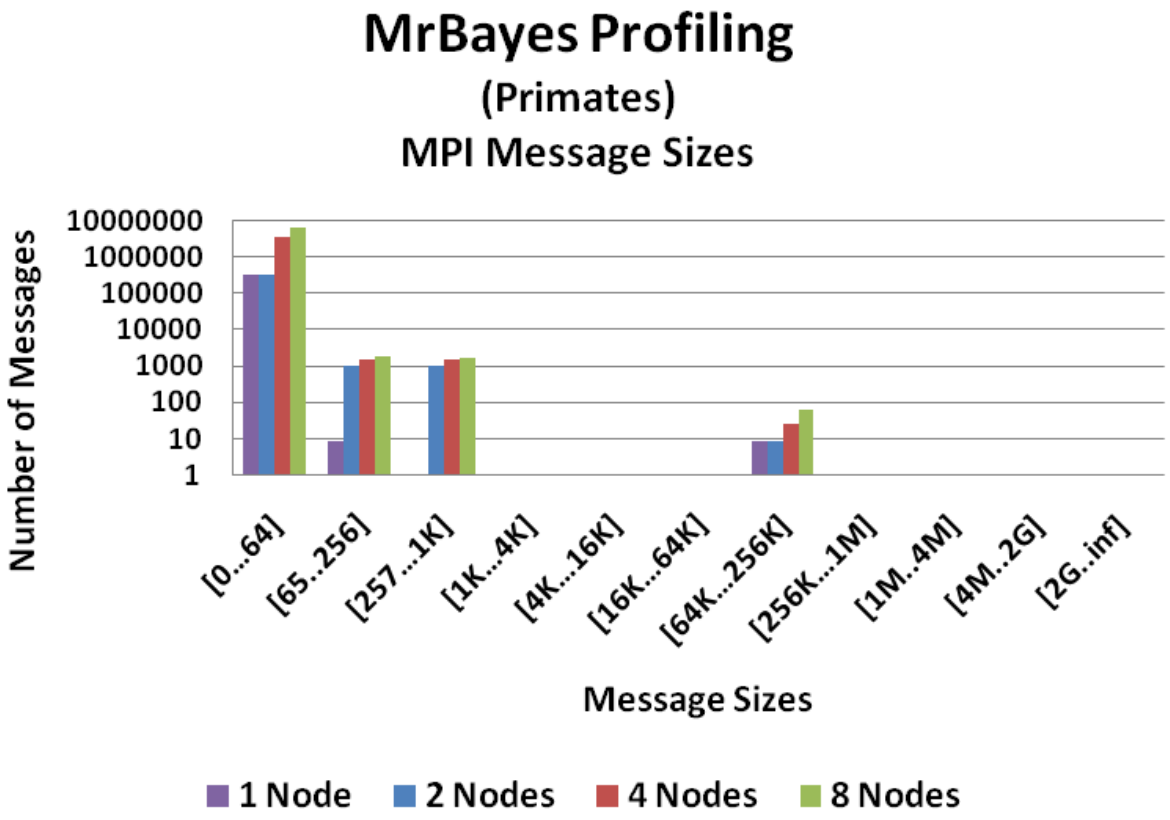
- **MPI_Waitall is the biggest time consumer at 8 node**
  - MPI_Waitall occupies 56% of all MPI time at node
  - MPI_Allreduce takes 30% of all MPI time at 8 node





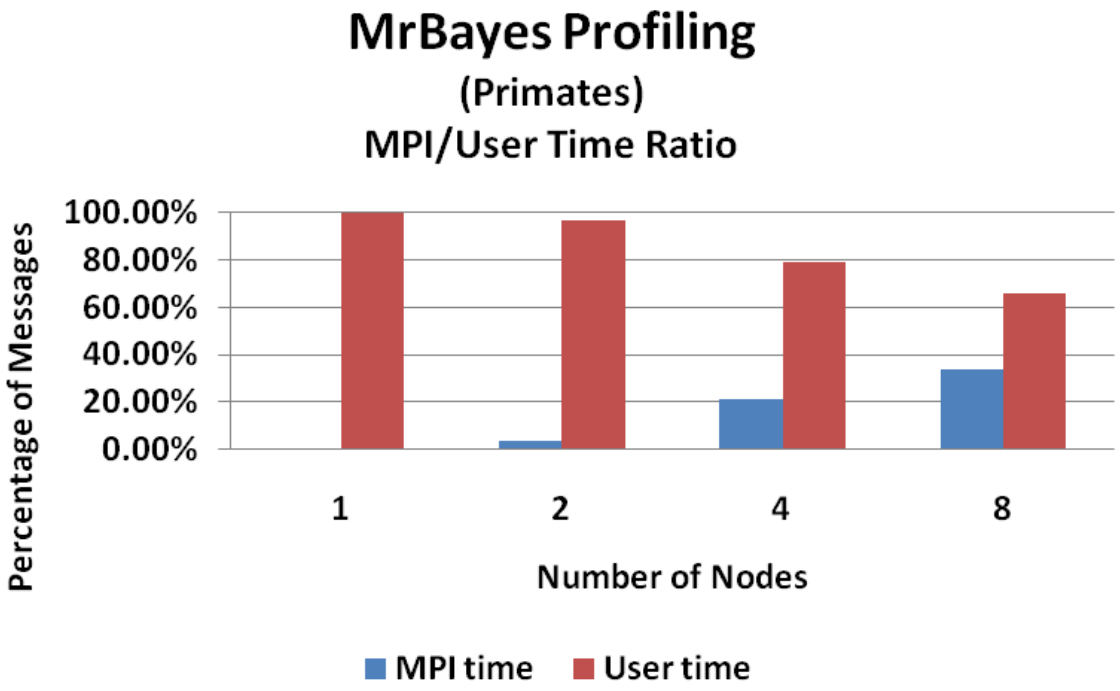
MrBayes Profiling – MPI Message Size

- **Majority of MPI messages are small messages**
  - In the range of 0 to 64 bytes
- **Messages increase for the changes in the MPI communication patterns**
  - Point-to-point communications starting at 2-node
  - Non-blocking communications starting at 4-node



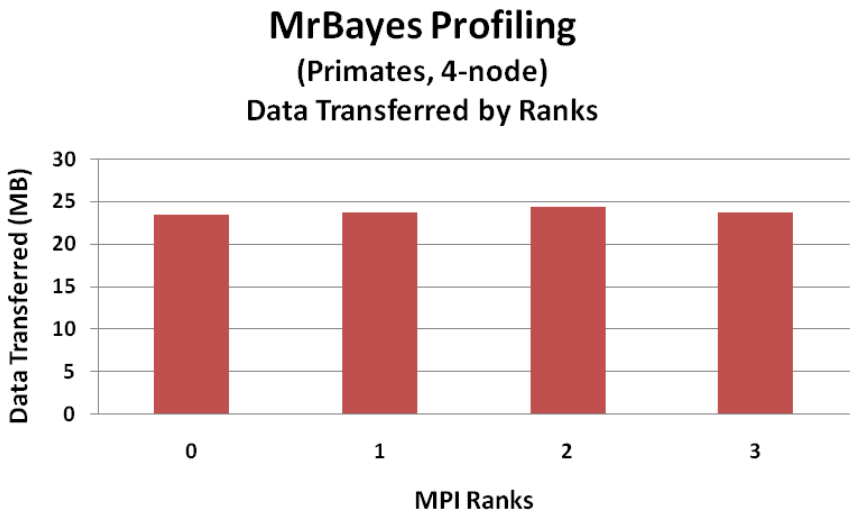
MrBayes Profiling – MPI/User Time Ratio

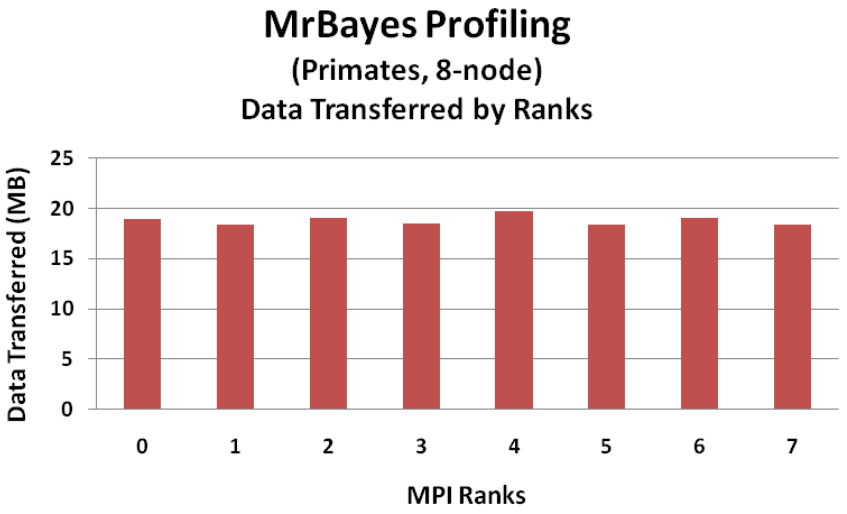
- **Computation time dominates run time**
  - Reflects that more time spent on computation than communications
- **Communication percentage increases substantially as the cluster scales**



MrBayes Profiling – MPI Data Transfer

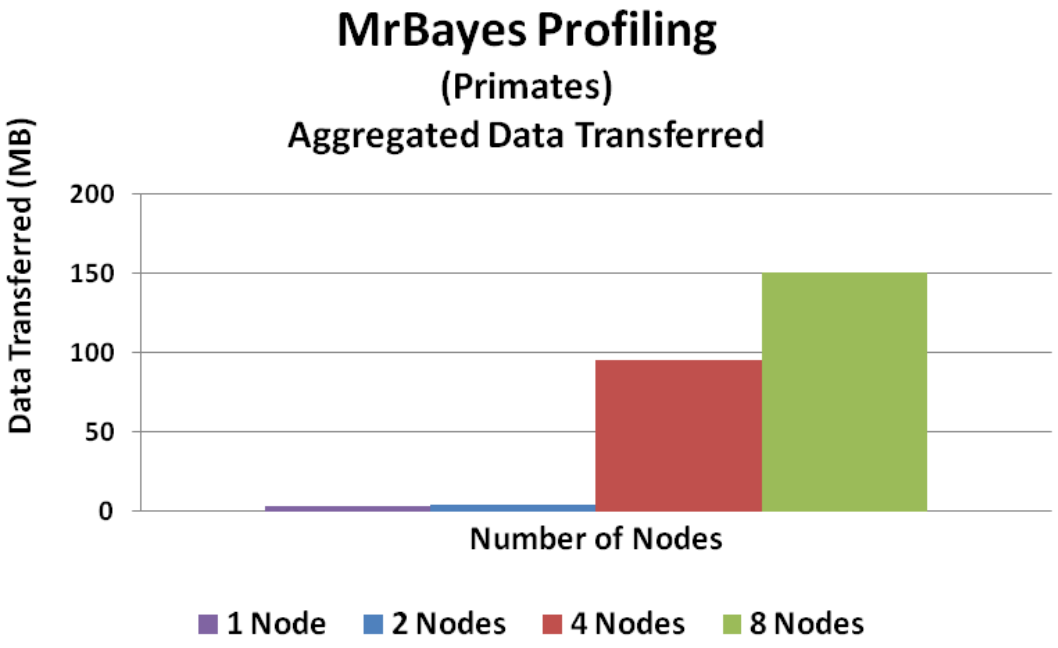
- **Shows fairly limited data transfers between the compute nodes**
  - In the range of 20MB per node
  - Reflects that MrBayes does not require high network bandwidth for communications
  - Shows that small message sizes dominate in MPI communications
- **Communication pattern shows all ranks communicate about evenly**





MrBayes Profiling – Aggregated 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 for 4 nodes**
  - Accounted for the non-blocking communications happening starting at 4 nodes



InfiniBand QDR

MrBayes – Summary

- **MrBayes is a CPU intensive application that would benefit most with Hybrid**
  - Hybrid utilizes both MPI and OpenMP
  - Using OpenMP threads to boost the performance by utilizing all CPU cores
  - Using MPI allow scaling to more compute nodes to run to maximizing the computation resources available
- **Using nodes with higher CPU frequency enables higher job productivity**
  - Shows a 4% increase in performance by using 3.06GHz versus 2.93GHz CPUs
- **All MPI implementations tested show good scalability**
  - Using processor binding in Open MPI can hurt job productivity
- **InfiniBand shows better performance as more compute nodes are used**
- **Profiling shows limited data transfer in data communications**
  - Reflects that MrBayes does not require high bandwidth for network communications
  - Typically means that the application is latency-bound

Thank You

HPC Advisory Council