

# CESM

## (Community Earth System Model)

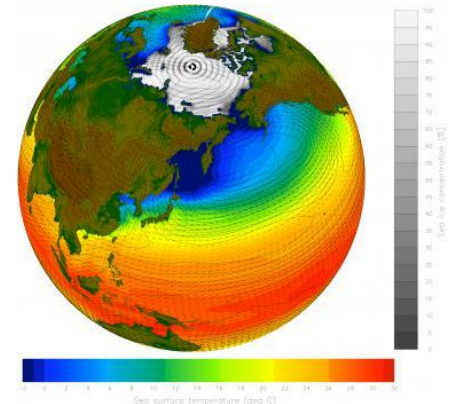
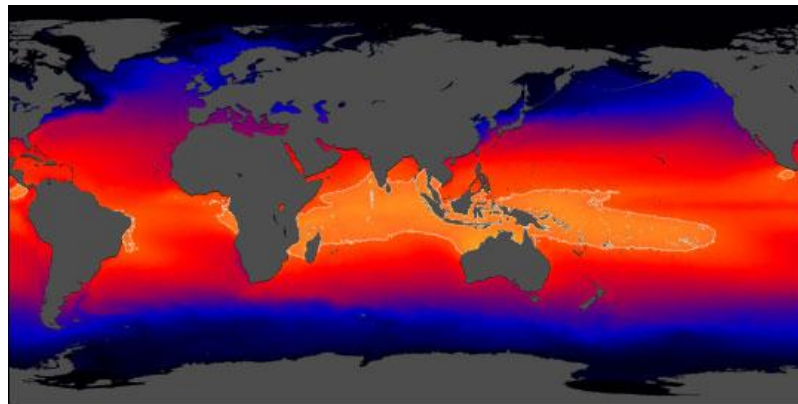
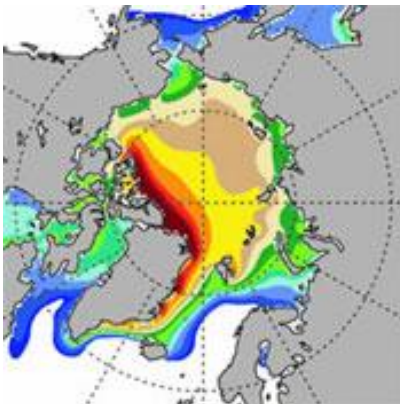
### Performance Benchmark and Profiling

August 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**
  - CESM performance overview
  - Understanding CESM communication patterns
  - Ways to increase CESM productivity
  - MPI libraries comparisons
- **For more info please refer to**
  - <http://www.dell.com>
  - <http://www.intel.com>
  - <http://www.mellanox.com>
  - <http://www.cesm.ucar.edu/models/cesm1.0/>

- **Community Earth System Model (CESM)**
  - A coupled climate model for simulating the earth's climate system
- **Composed of four separate models simultaneously simulating:**
  - Earth's atmosphere
  - Ocean
  - Land surface
  - Sea-ice
- **CESM allows researchers to conduct fundamental research into the earth's past, present and future climate states**
- **CESM1.0.3 supersedes CCSM4.0**



- **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: Intel Compilers 11.1**
- **MPI: Open MPI 1.5.3 with KNEM 0.9.6, Platform MPI 8.1.1**
- **Application: CESM 1.0.3 (Libraries: NetCDF 4.1.3)**
- **Benchmark datasets:**
  - B\_1850\_CN (B1850CN), all active components, pre-industrial (1850), with CN (Carbon Nitrogen) in CLM

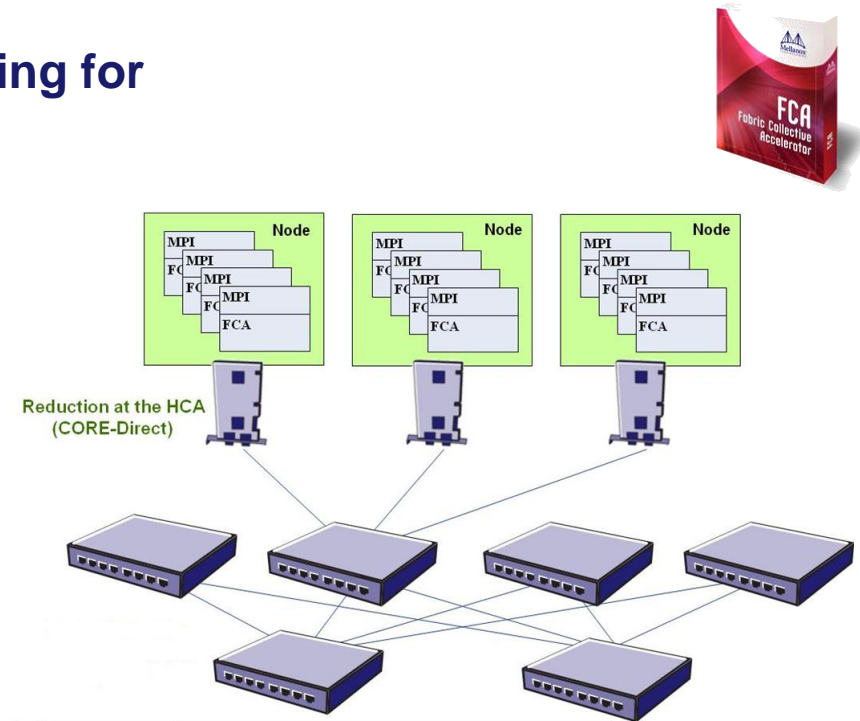
- **Intel® Cluster Ready systems make it practical to use a cluster to increase your simulation and modeling productivity**
  - Simplifies selection, deployment, and operation of a cluster
- **A single architecture platform supported by many OEMs, ISVs, cluster provisioning vendors, and interconnect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



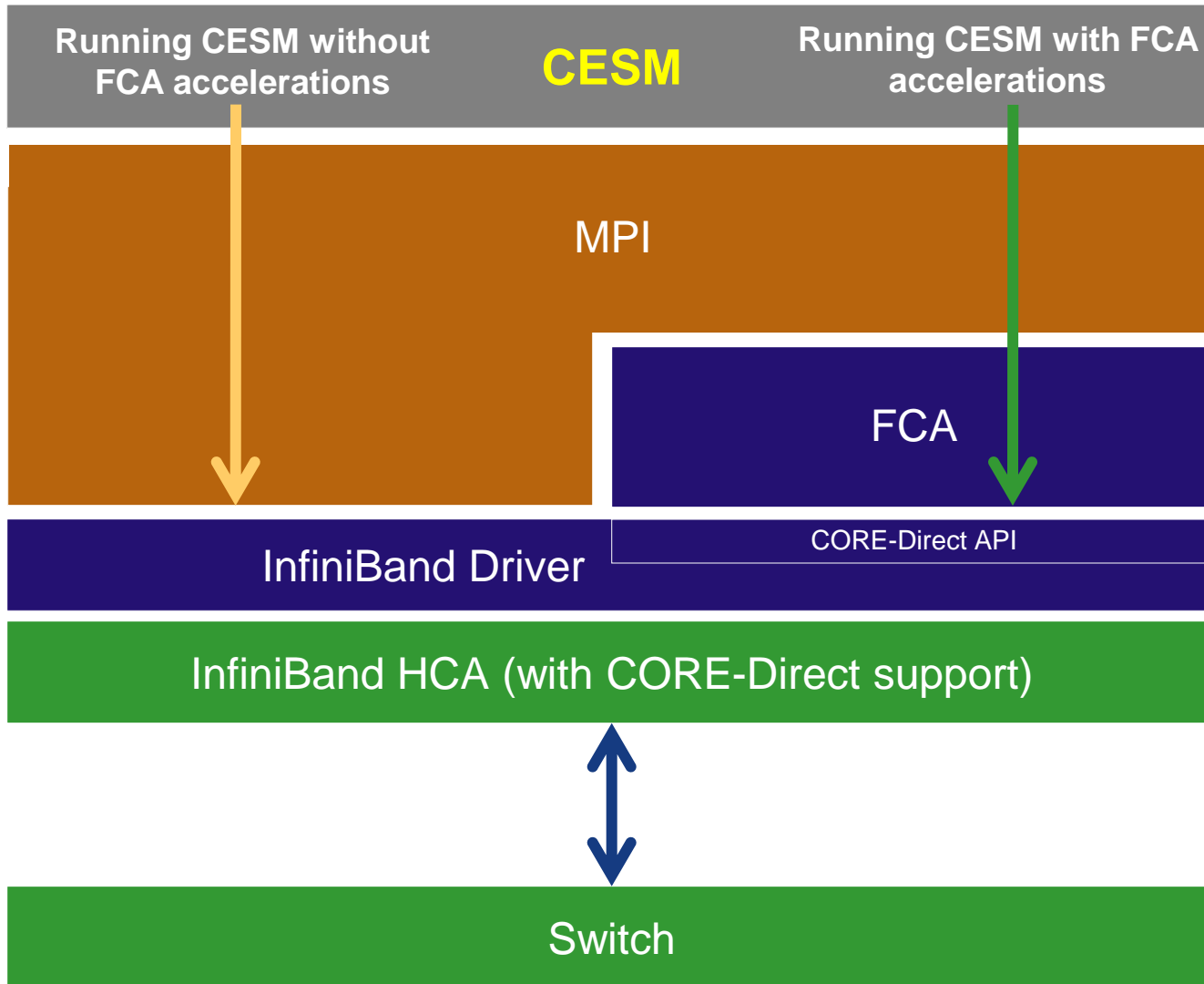
- **System Structure and Sizing Guidelines**
  - 38-node cluster build with Dell PowerEdge™ M610 blade servers
  - Servers optimized for High Performance Computing environments
  - Building Block Foundations for best price/performance and performance/watt
- **Dell HPC Solutions**
  - Scalable Architectures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



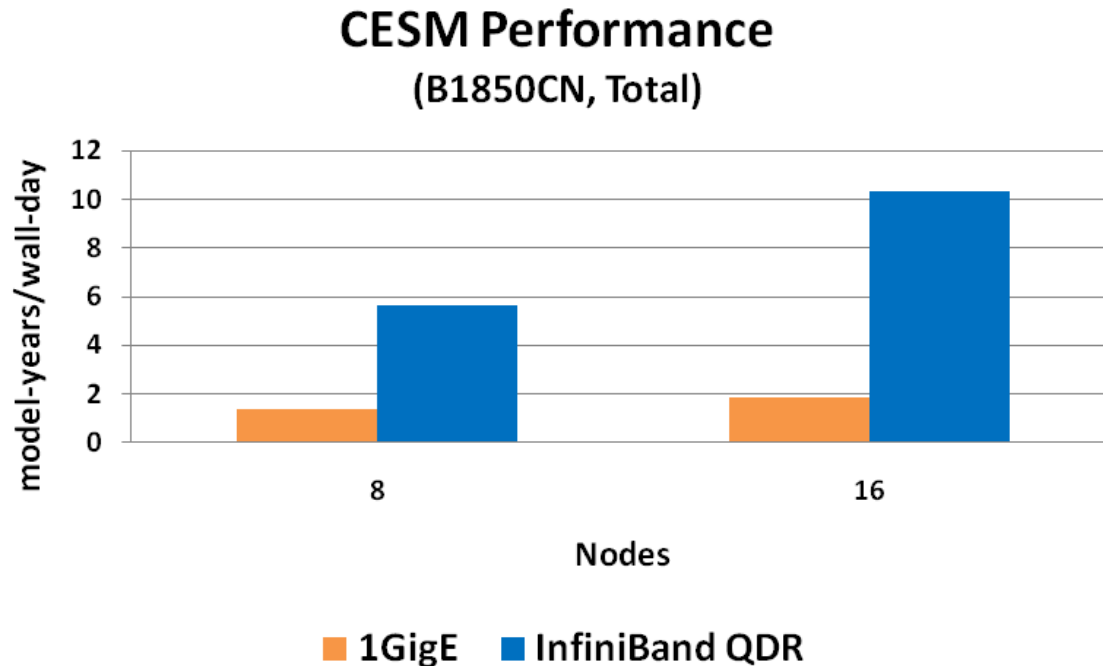
- **Mellanox Fabric Collectives Accelerator (FCA)**
  - Utilized hardware accelerations on the adapter (CORE-Direct)
  - Accelerating MPI collectives operations by offloading them to the network
  - The world first complete solution for MPI collectives offloads
  
- **FCA 2.1 supports accelerations/offloading for**
  - MPI\_Barrier
  - MPI\_Broadcast
  - MPI\_Allreduce and MPI\_Reduce
  - MPI\_Allgather and MPI\_Allgatherv



# Software Layers Overview



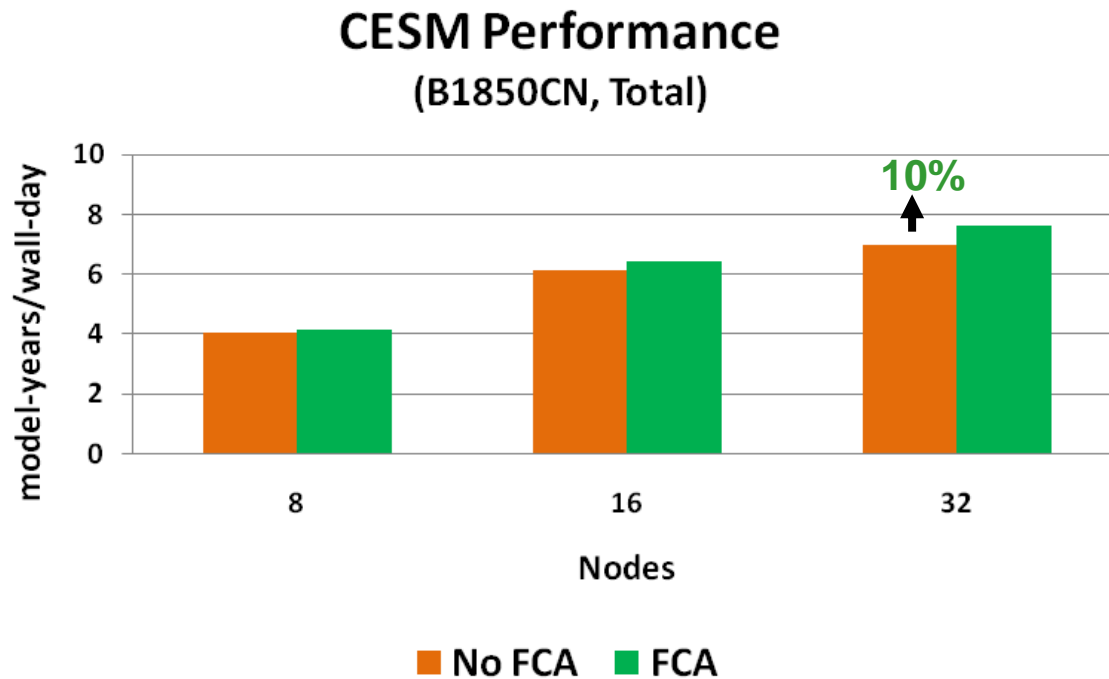
- **Dataset: B1850CN**
  - Requires around 135GB of memory from all the processes to run
- **InfiniBand enables higher scalability and cluster productivity**
  - Provides the needed network infrastructure to deliver cluster scalability
  - 1GigE seen minimal gain by doubling the number of nodes (from 8 to 16)
- **CESM requires good network throughput for data communications**



*Higher is better*

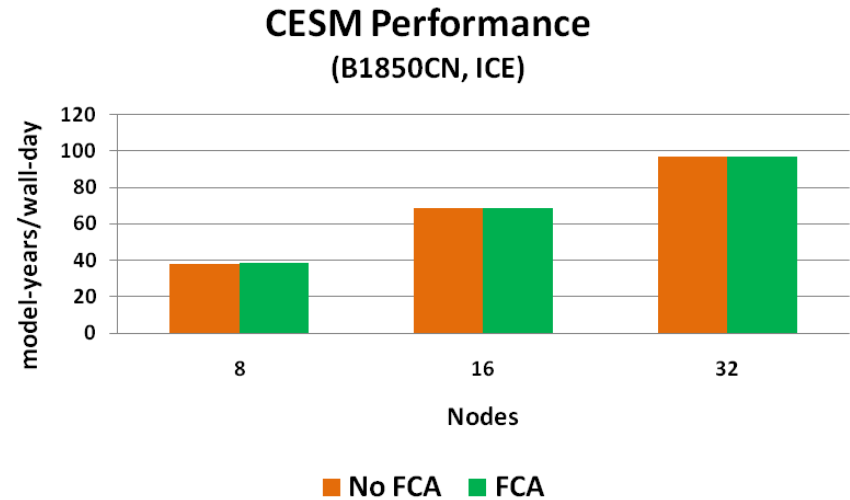
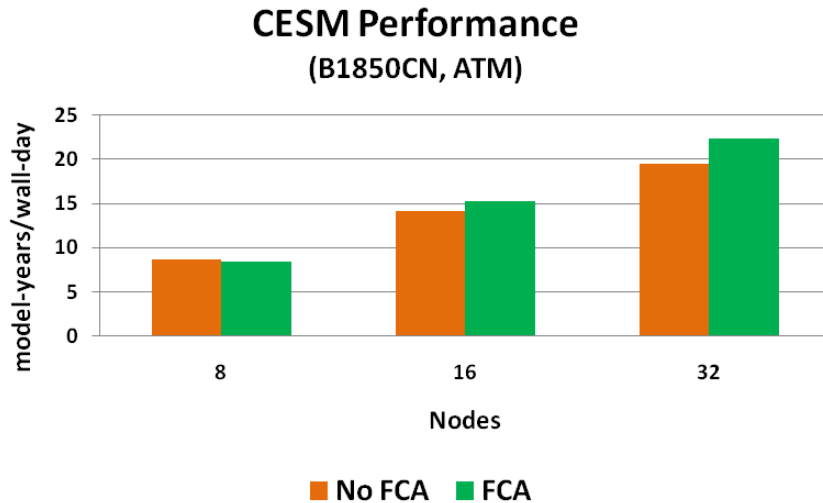
*InfiniBand QDR*

- **CESM demonstrates the benefits of having MPI collectives offloads**
  - By freeing up CPU resources to the InfiniBand hardware, more CPU computation can be done
- **FCA enables 10% performance gain at 32 nodes / 384 cores**
  - Expect to continue to show advantage expected at higher node count / core count
  - Accounts for the time savings due to MPI\_Barrier (51%) and MPI\_Bcast (25%)



*Higher is better*

- The influence of using MPI collectives offloads per CESM component

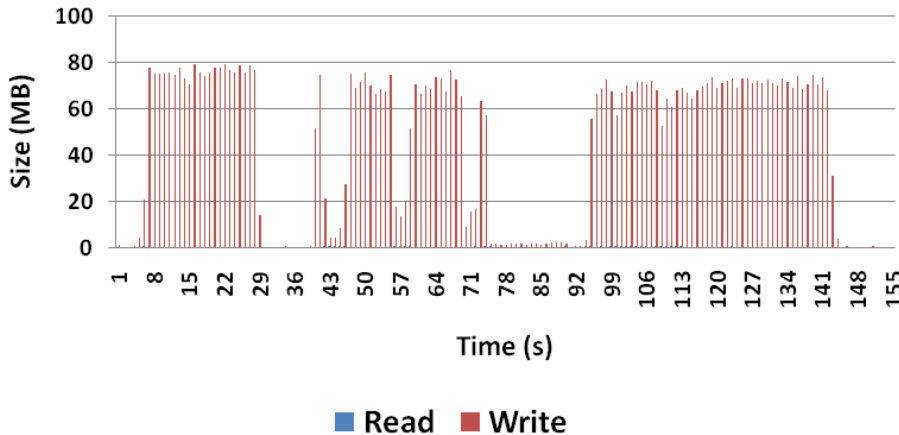


*Higher is better*

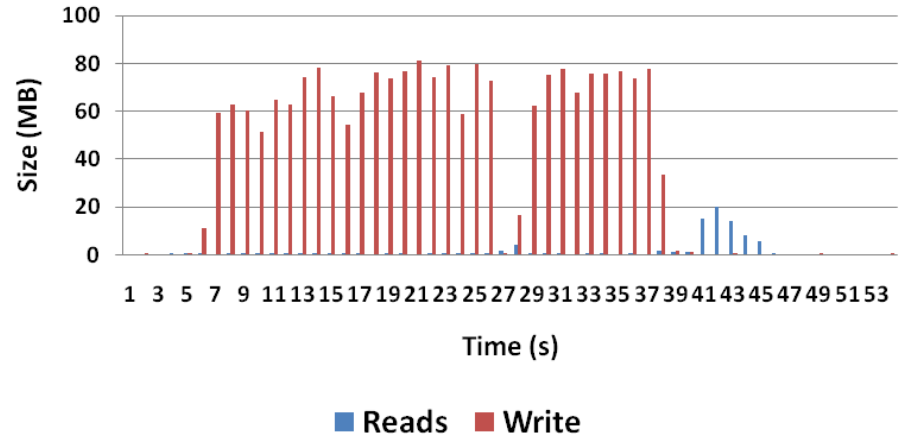
*Open MPI  
InfiniBand QDR*

- **Heavy writes are seen in periods during the test run**
  - Writes are more concentrated at the beginning, middle and near the end of the tests
  - Shows significantly more writes than reads occurred
  - The write speed seems to be limited around 80MB
- **NFS file system is shared from the head node**

**CESM Profiling**  
(B1850CN, 16-node, Total)

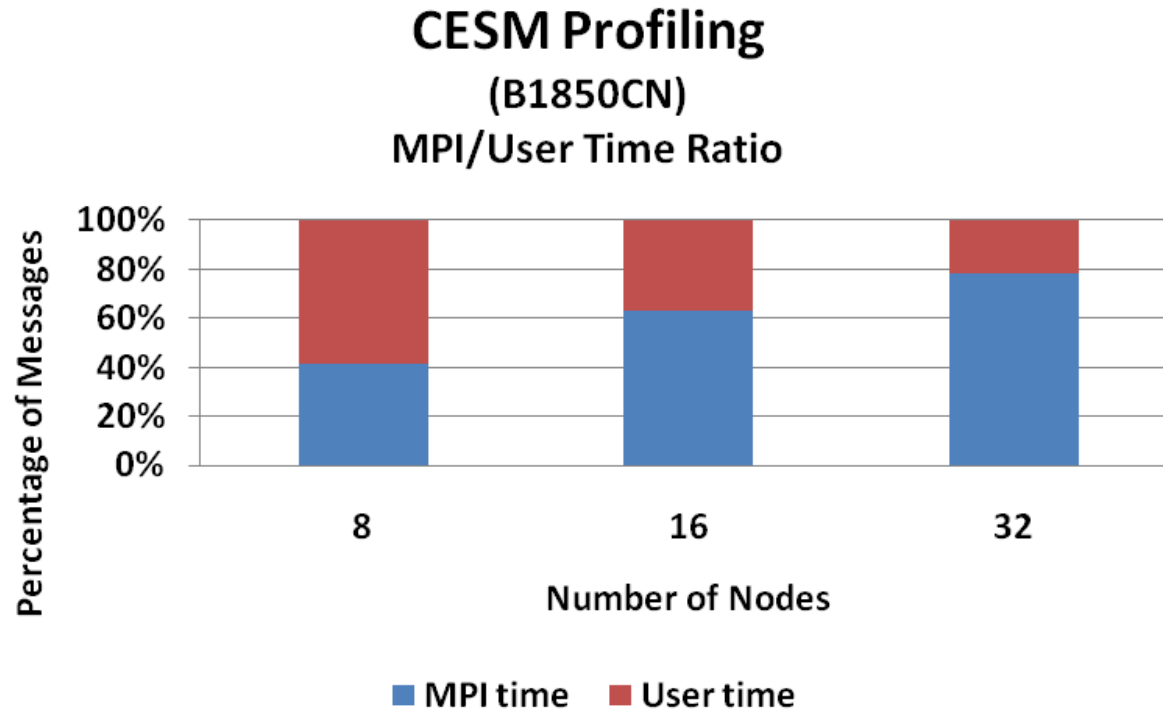


**CESM Profiling**  
(B1850CN, 32-node, Total)



*InfiniBand QDR*

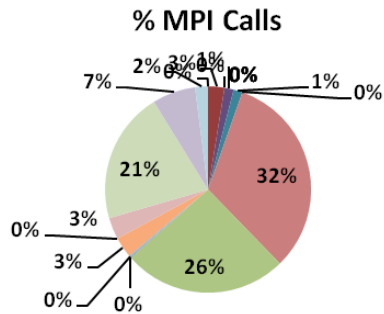
- **Communication percentage increases as the cluster scales**
  - Reflects that more time spent on computation than communications



- **The dataset uses a different set of MPI functions**

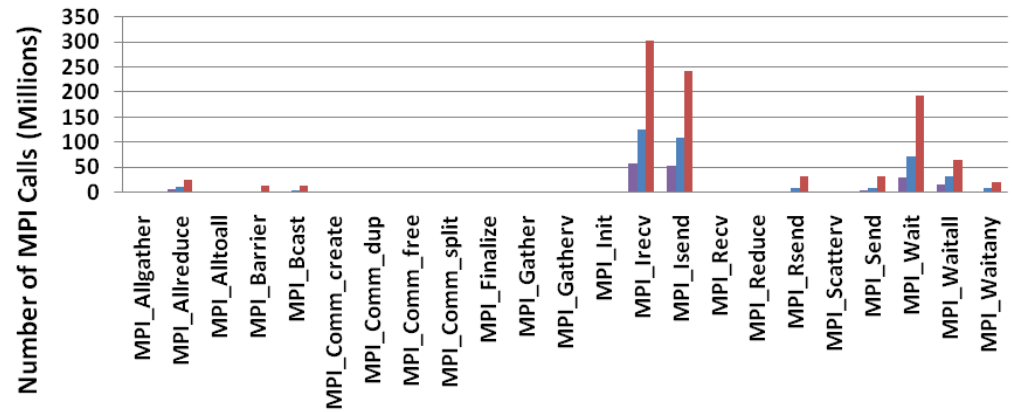
- Majority of the calls are non-blocking MPI calls
- MPI\_Irecv (32%)
- MPI\_Isend (26%)
- MPI\_Wait (21%)
- MPI\_Waitall (7%)

**CESM Profiling**  
(B1850CN, 32-node, InfiniBand)  
% MPI Calls



- MPI\_Allgather
- MPI\_Allreduce
- MPI\_Alltoall
- MPI\_Barrier
- MPI\_Bcast
- MPI\_Comm\_create
- MPI\_Comm\_dup
- MPI\_Comm\_free
- MPI\_Comm\_split
- MPI\_Finalize
- MPI\_Gather
- MPI\_Gatherv
- MPI\_Init
- MPI\_Irecv
- MPI\_Isend
- MPI\_Recv
- MPI\_Reduce
- MPI\_Rsend
- MPI\_Scatterv
- MPI\_Send
- MPI\_Wait
- MPI\_Waitall
- MPI\_Waitany

**CESM Profiling**  
(B1850CN)  
Number of MPI Calls



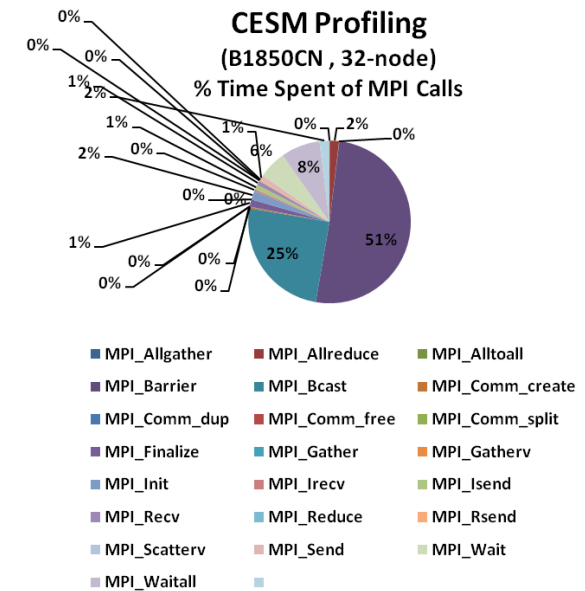
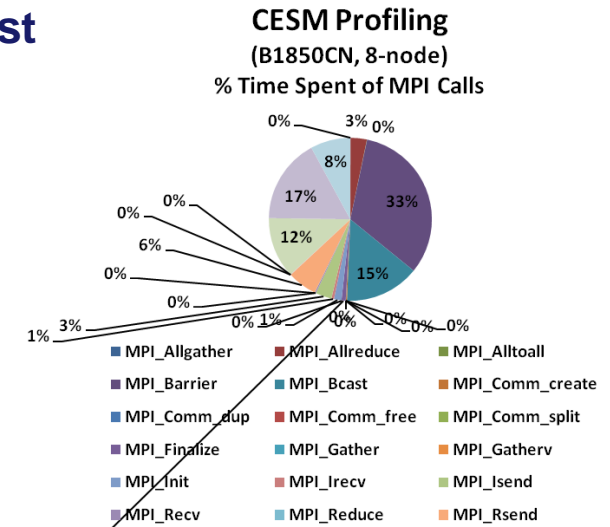
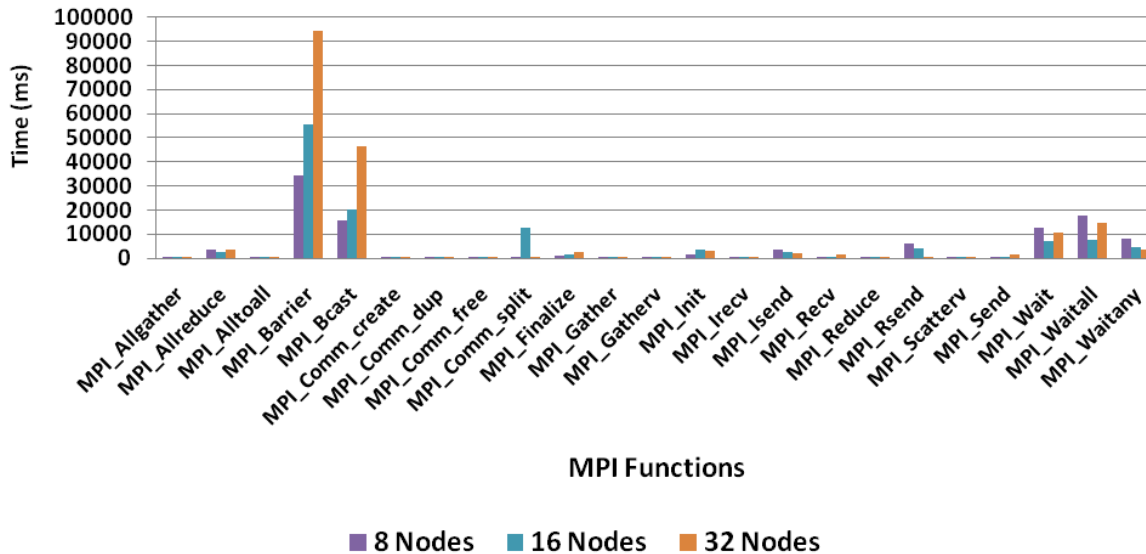
- 8 Nodes
- 16 Nodes
- 32 Nodes

MPI Functions

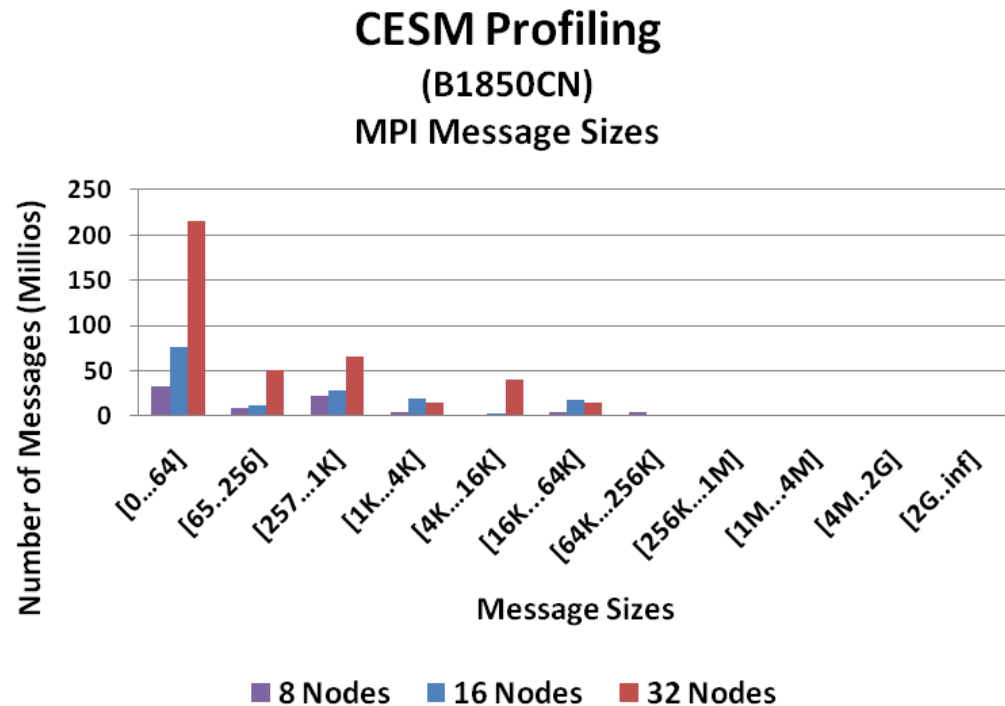
# CESM Profiling – Time Spent by MPI Calls

- Majority of the MPI time are spent on Barrier and Bcast
  - MPI\_Barrier (51% on 32-node to 33% on 8-node)
  - MPI\_Bcast (25% on 32-node to 15% on 8-node)

**CESM Profiling (B1850CN)**  
Time Spent of MPI Calls

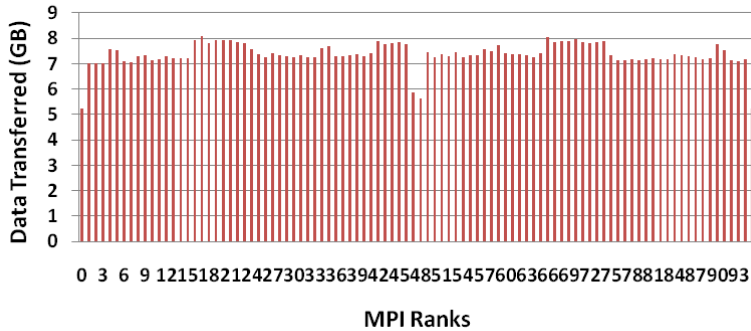


- **Majority of MPI messages are spread across small to medium message sizes**
  - Highest concentration of message sizes are in the range of 0B to 64Bbytes
  - Similar message sizes are seen for 8-, 16- and 32-node runs

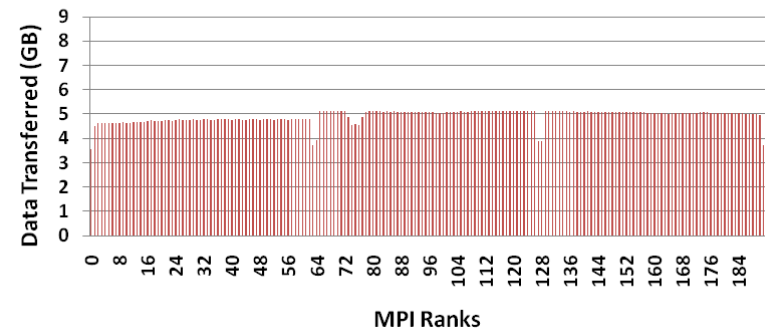


- **CESM shows substantial data transfers between the MPI processes**
  - In the range of 8GB per process in an 8-node test, down to 3GB in a 32-node test
  - Data communications are generally even between MPI processes

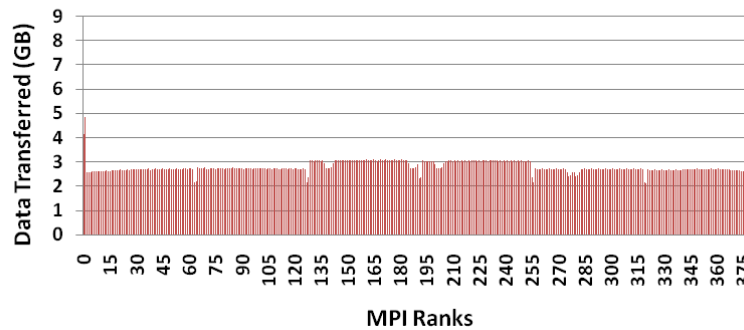
**CESM Profiling**  
(B1850CN, 8-node)  
Data Transferred by Ranks



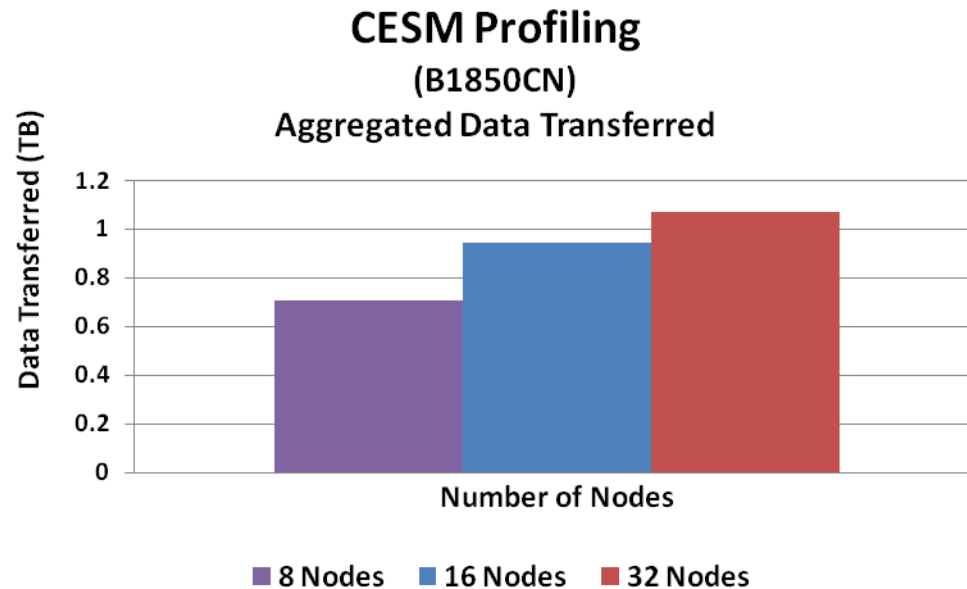
**CESM Profiling**  
(B1850CN, 16-node)  
Data Transferred by Ranks



**CESM Profiling**  
(B1850CN, 32-node)  
Data Transferred by Ranks



- **Aggregated data transfer refers to:**
  - Total amount of data being transferred in the network between all MPI ranks collectively
- **Shows sizable data transfer takes place in the network**
  - The total data transfer increases gradually
  - High throughput network interconnect can be beneficial



*InfiniBand QDR*

- **CESM is a compute intensive application that also shows**
  - CPU and memory
  - Disk IO usage
  - High throughput for data communications
- **InfiniBand shows better performance as more compute nodes are used**
  - More network throughput is needed for cluster scalability
- **MPI Collectives offloads**
  - Allows offloading MPI collective operations to the InfiniBand hardware
  - Which frees up communications on the CPUs
  - It allowing CPUs to focus on computation

# 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