HPC ADVISORY COUNCIL

CCSM Performance Benchmark and Profiling

April 2010

AMD
The future is fusion

DELL

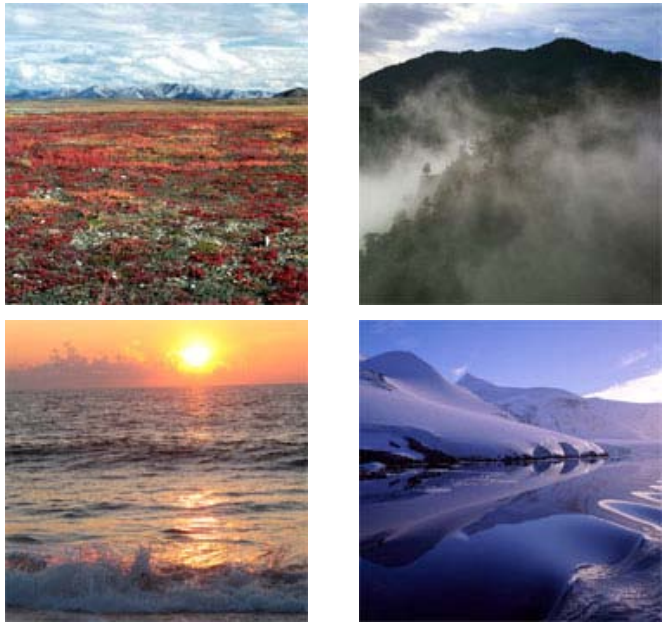
Mellanox
TECHNOLOGIES

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**
  - www.mellanox.com, www.dell.com/hpc, www.amd.com
  - http://www.ccsm.ucar.edu/models/ccsm4.0

CCSM 4.0 - Community Climate System Model

- **CCSM is a coupled climate model for simulating the earth's climate**
- **Composed of four separate models**
  - Atmosphere (CAM4)
  - Ocean (POP2)
  - Land surface (CLM4)
  - Sea-ice (CICE4)
- **CCSM was developed in cooperation with NFS, DOE, NASA, and NCAR**



Objectives

- **The presented research was done to provide best practices**
  - CCSM 4.0 performance benchmarking
    - Performance tuning with different communication libraries and compilers
    - Interconnect performance comparisons
  - Understanding CCSM communication patterns
  - Power-efficient simulations
- **The presented results will demonstrate**
  - Balanced compute system enables
    - Good application scalability
    - Power saving

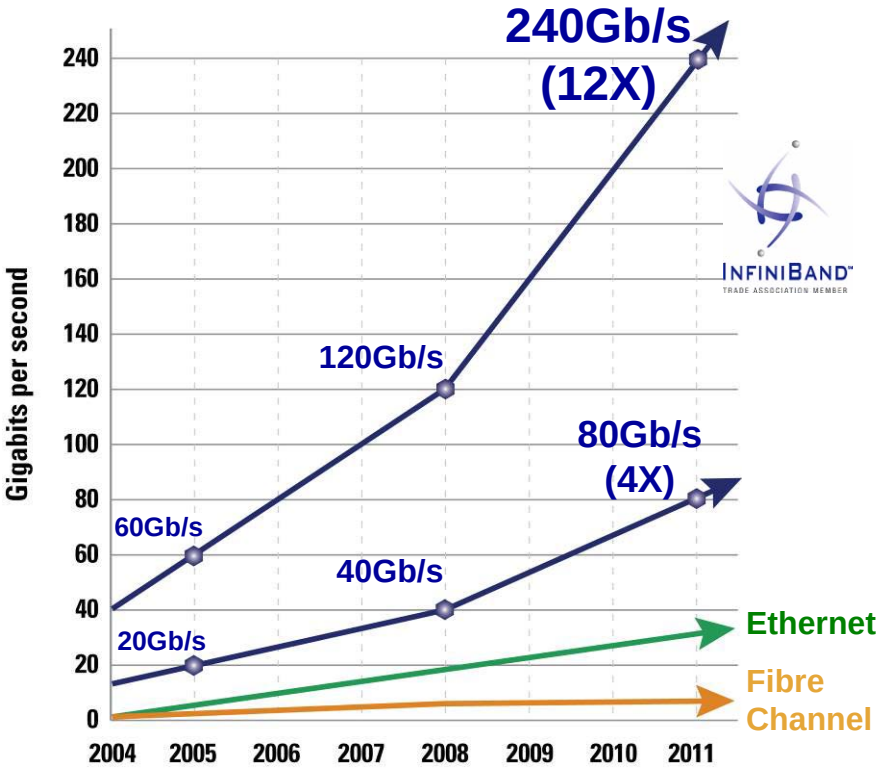
Test Cluster Configuration

- **Dell™ PowerEdge™ SC 1435 16-node cluster**
- **Quad-Core AMD Opteron™ 2382 (“Shanghai”) CPUs**
- **Mellanox® InfiniBand ConnectX® 20Gb/s (DDR) HCAs**
- **Mellanox® InfiniBand DDR Switch**
- **Memory: 16GB memory, DDR2 800MHz per node**
- **OS: RHEL5U3, OFED 1.5 InfiniBand SW stack**
- **Compiler and Math library: Intel compiler 11.1**
- **MPI: OpenMPI-1.3.3, MVAPICH2-1.4, Intel MPI 4.0**
- **Application: CCSM 4.0**
- **Benchmark Workload**
  - Compset: B1850CN
    - Resolution: 0.9x1.25_gx1v6

Mellanox InfiniBand Solutions

- **Industry Standard**
  - Hardware, software, cabling, management
  - Design for clustering and storage interconnect
- **Performance**
  - 40Gb/s node-to-node
  - 120Gb/s switch-to-switch
  - 1us application latency
  - Most aggressive roadmap in the industry
- **Reliable with congestion management**
- **Efficient**
  - RDMA and Transport Offload
  - Kernel bypass
  - CPU focuses on application processing
- **Scalable for Petascale computing & beyond**
- **End-to-end quality of service**
- **Virtualization acceleration**
- **I/O consolidation Including storage**

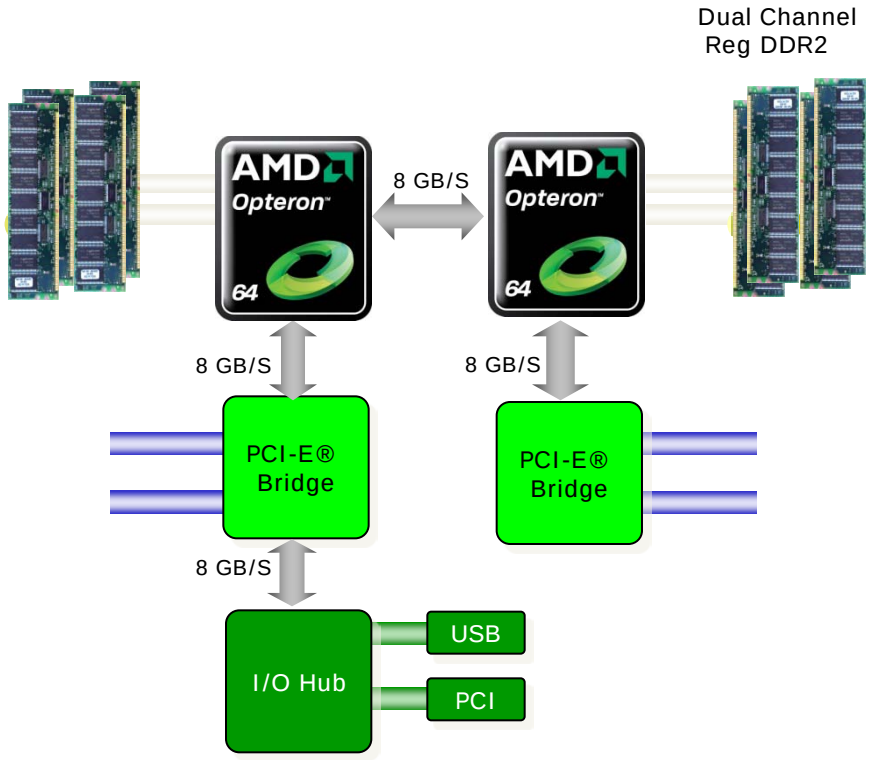
The InfiniBand Performance Gap is Increasing



InfiniBand Delivers the Lowest Latency

Quad-Core AMD Opteron™ Processor

- **Performance**
  - Quad-Core
    - Enhanced CPU IPC
    - 4x 512K L2 cache
    - 6MB L3 Cache
  - Direct Connect Architecture
    - HyperTransport™ Technology
    - Up to 24 GB/s peak per processor
  - Floating Point
    - 128-bit FPU per core
    - 4 FLOPS/clk peak per core
  - Integrated Memory Controller
    - Up to 12.8 GB/s
    - DDR2-800 MHz or DDR2-667 MHz
- **Scalability**
  - 48-bit Physical Addressing
- **Compatibility**
  - Same power/thermal envelopes as 2nd / 3rd generation AMD Opteron™ processor

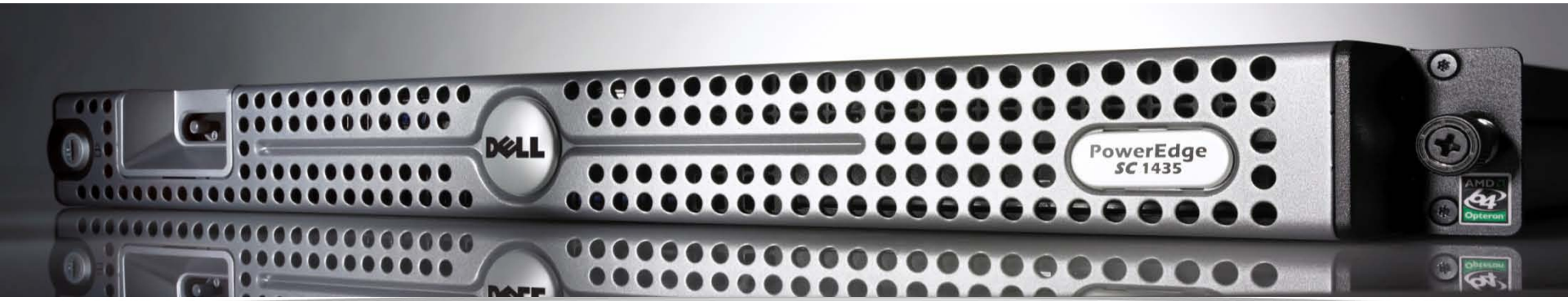


Dell PowerEdge Servers helping Simplify IT

- **System Structure and Sizing Guidelines**
  - 16-node cluster build with Dell PowerEdge™ SC 1435 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

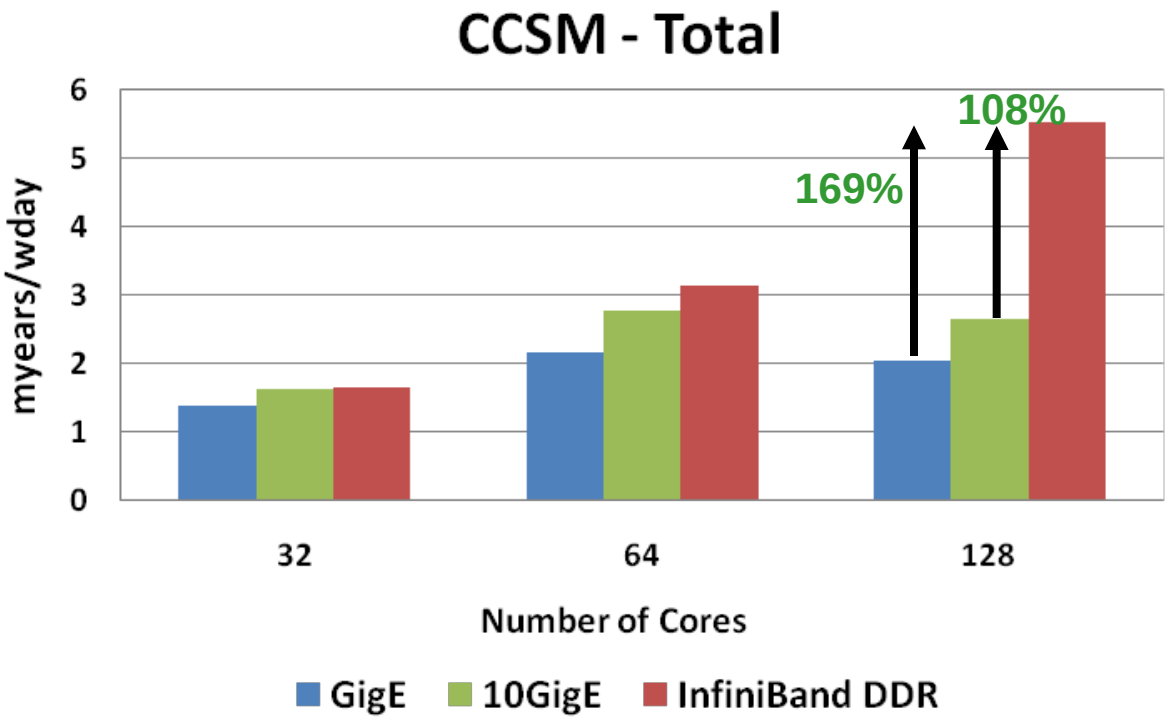
Dell PowerEdge™ Server Advantage

- **Dell™ PowerEdge™ servers incorporate AMD Opteron™ and Mellanox ConnectX InfiniBand to provide leading edge performance and reliability**
- **Building Block Foundations for best price/performance and performance/watt**
- **Investment protection and energy efficient**
- **Longer term server investment value**
- **Faster DDR2-800 memory**
- **Enhanced AMD PowerNow!**
- **Independent Dynamic Core Technology**
- **AMD CoolCore™ and Smart Fetch Technology**
- **Mellanox InfiniBand end-to-end for highest networking performance**



CCSM 4.0 Benchmark Results

- **InfiniBand enables higher performance and scalability**
  - Up to 169% higher performance than GigE and 108% higher than 10GigE
  - Both GigE and 10GigE stop scaling after 8 nodes

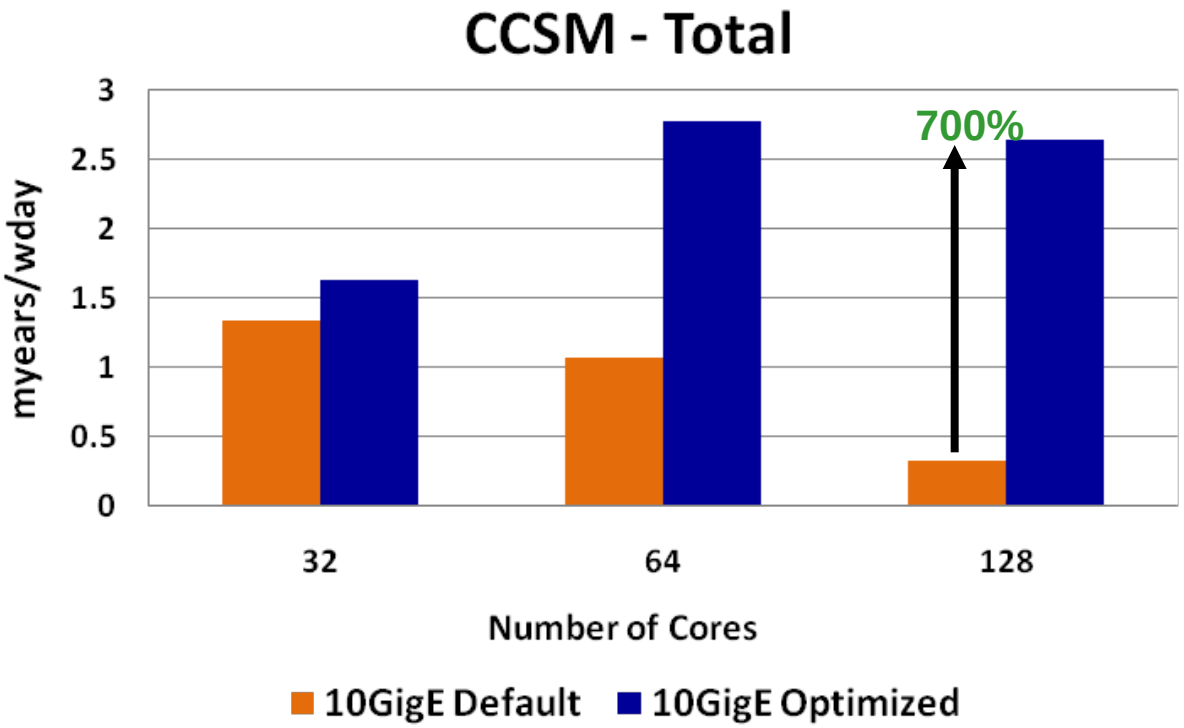


Higher is better

8-cores per node

CCSM 4.0 Benchmark Results – 10GigE Tuning

- **Default 10GigE NIC setting is not optimal for CCSM**
  - LRO (Large Receive Offload) might be enabled by default for 10GigE
  - LRO dramatically increases large message latency (>1MB)
  - Disable LRO and increase MTU size enable 700% higher performance

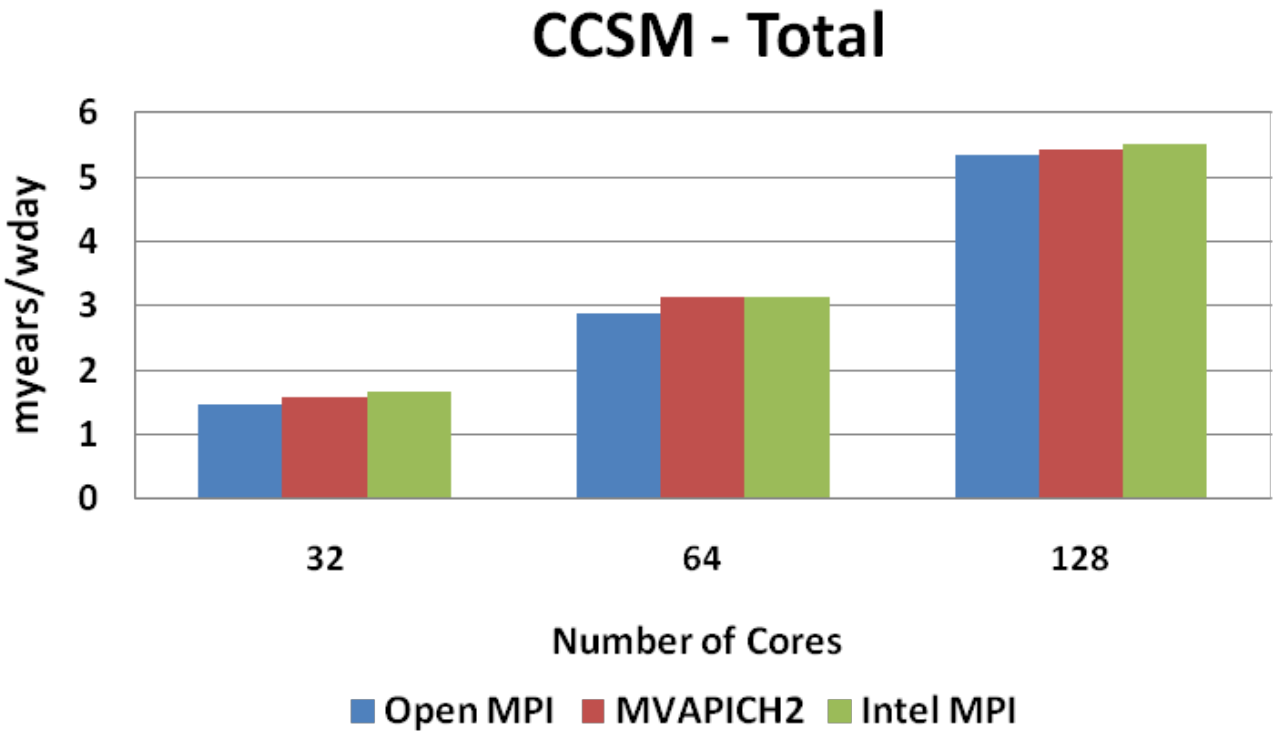


Higher is better

8-cores per node

CCSM 4.0 Benchmark Results

- **Total throughput for the complete CCSM model**
  - Three different MPI provide similar performance

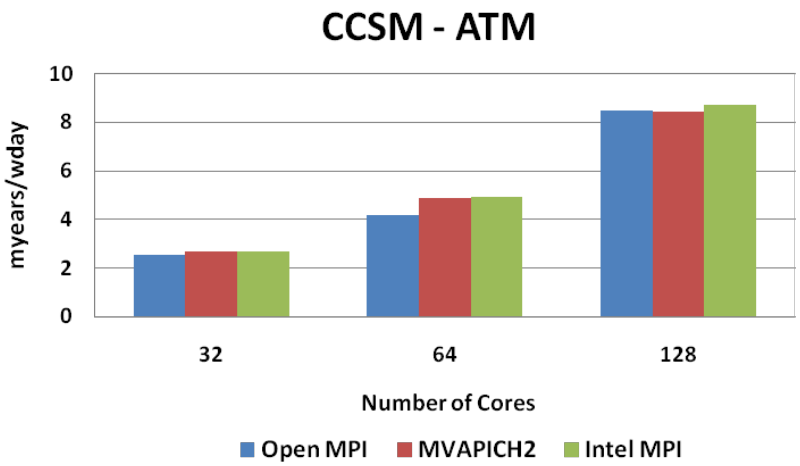


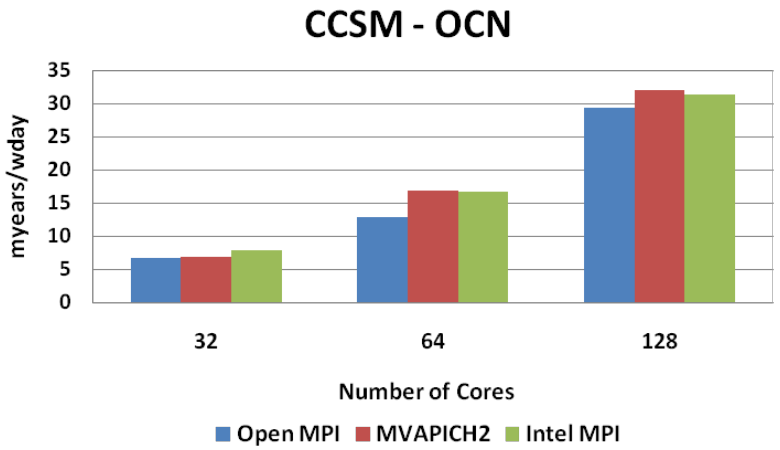
8-cores per node

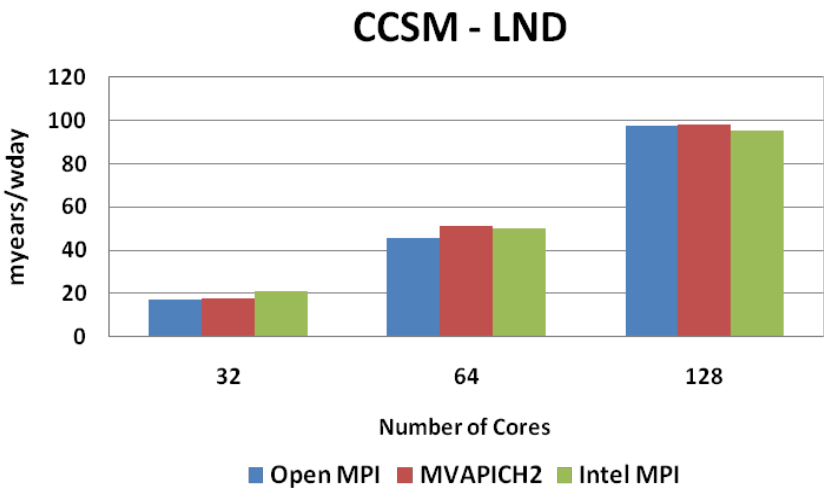
Higher is better

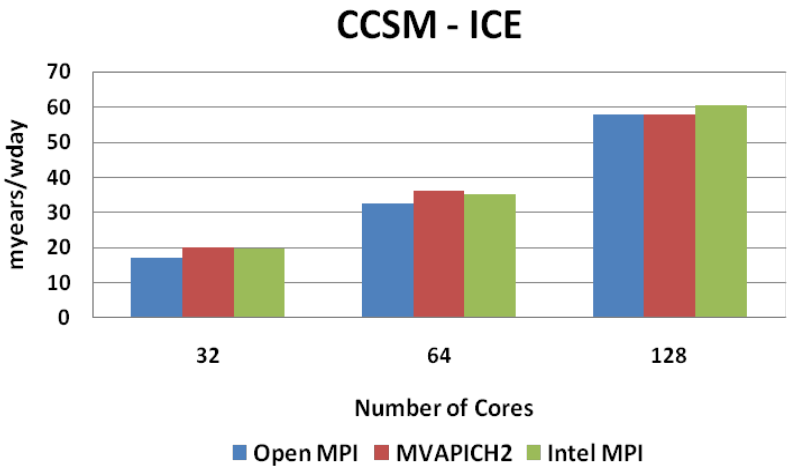
CCSM 4.0 Benchmark Results

- **Performance for each component**







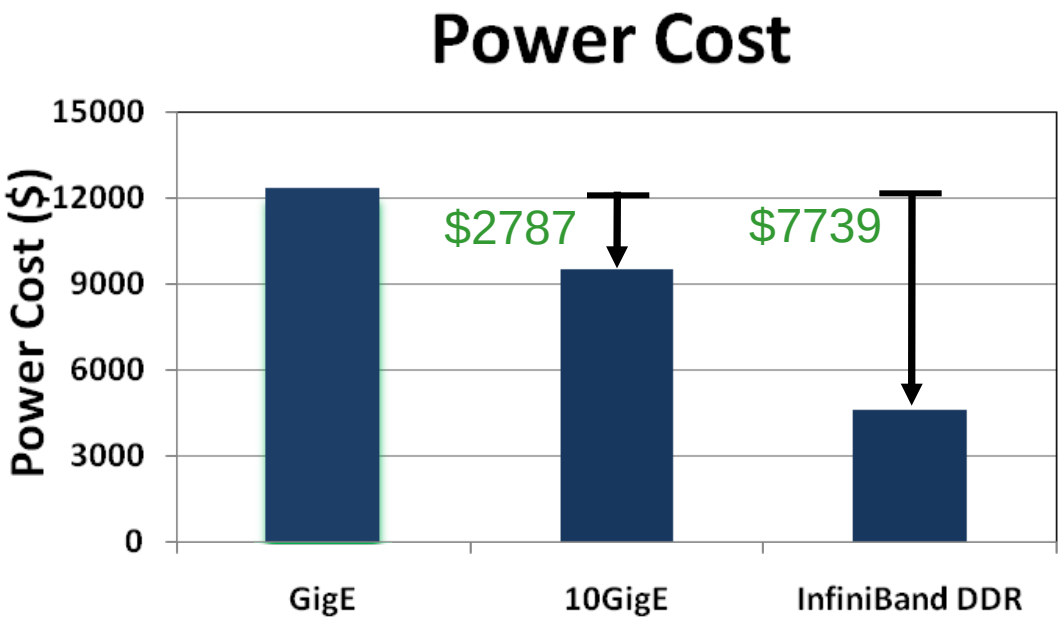


Higher is better

8-cores per node

Power Cost Savings with Different Interconnect

- **Dell economical integration of AMD CPUs and Mellanox InfiniBand**
  - To achieve same number of CCSM simulations over GigE
  - InfiniBand saves power up to $7739 versus GigE and $2787 versus 10GigE
  - Yearly based for 16-node cluster
- **As cluster size increases, more power can be saved**



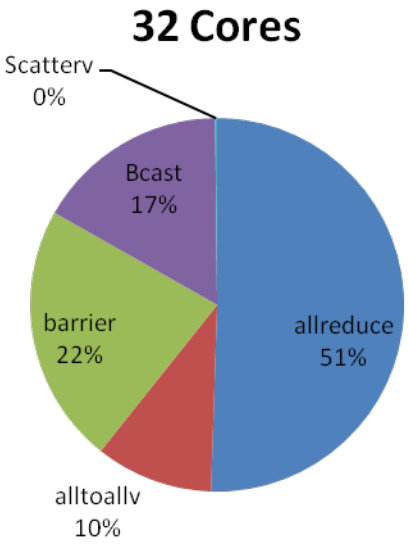
*$/KWh = KWh * $0.20*
For more information - http://enterprise.amd.com/Downloads/svrpwrusecompletefinal.pdf

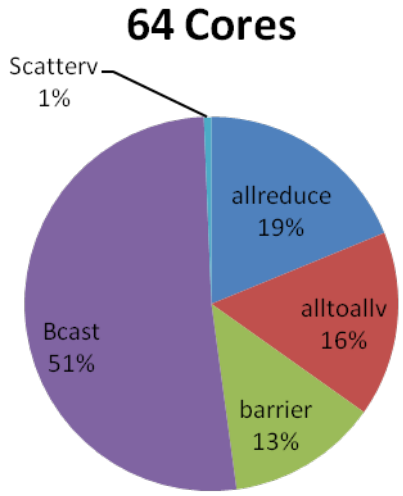
CCSM 4.0 Benchmark Summary

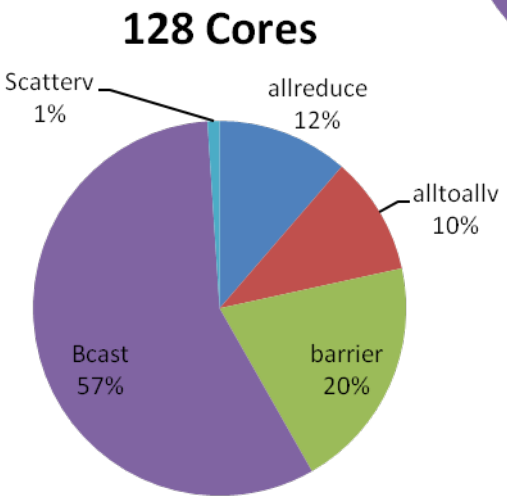
- **Interconnect comparison shows**
  - InfiniBand delivers superior performance in every cluster size versus GigE and 10GigE
  - Performance advantage extends as cluster size increases
- **Three different MPIs provides similar performance**
- **InfiniBand enables power saving**
  - Up to $7739/year power savings versus GigE and $2787 versus 10GigE on16 node cluster
- **Dell™ PowerEdge™ server blades provides**
  - Linear scalability (maximum scalability) and balanced system
    - By integrating InfiniBand interconnect and AMD processors
  - Maximum return on investment through efficiency and utilization

CCSM 4.0 MPI Profiling – Communication Overhead

- **Bcast, Allreduce, Barrier, and Alltoallv are four major collectives**
- **Communication overhead of Bcast increases faster than functions**

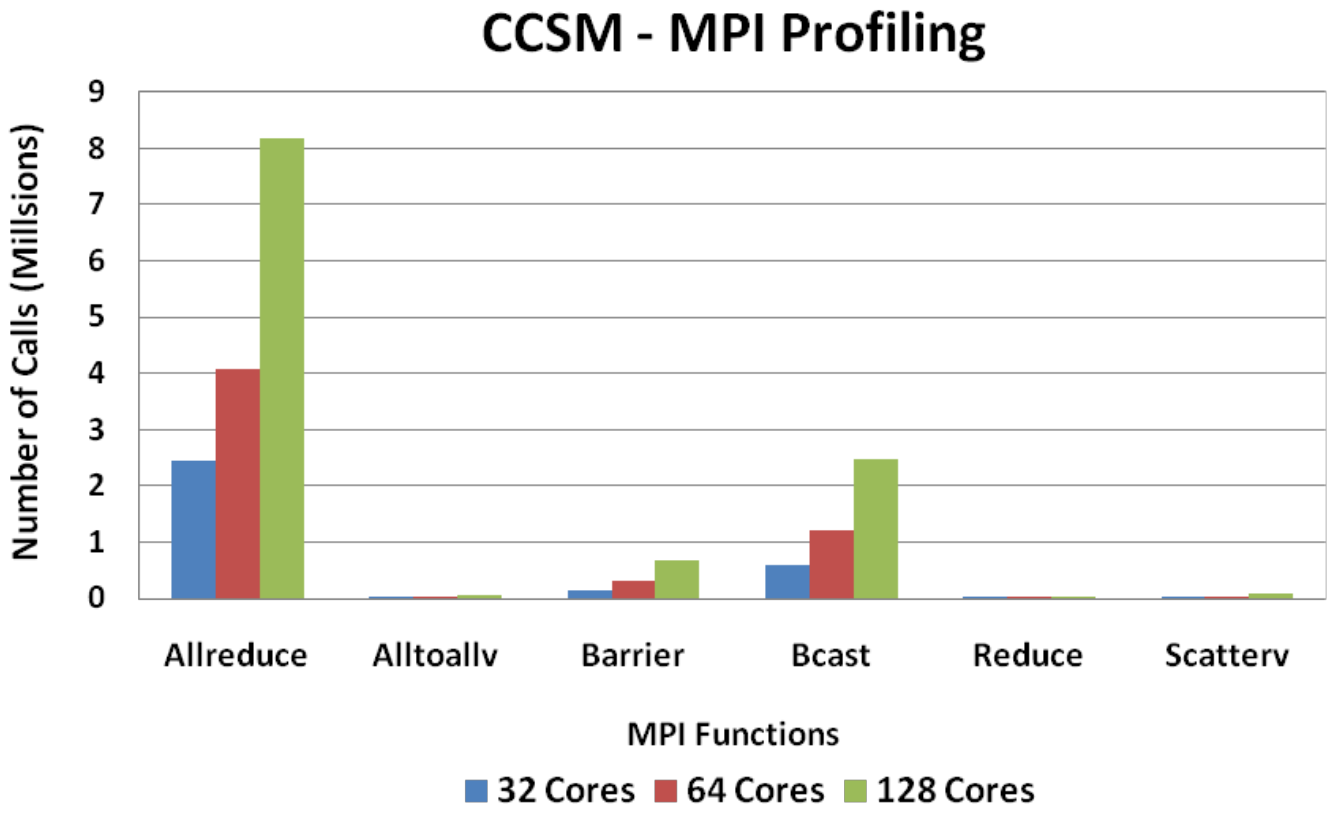






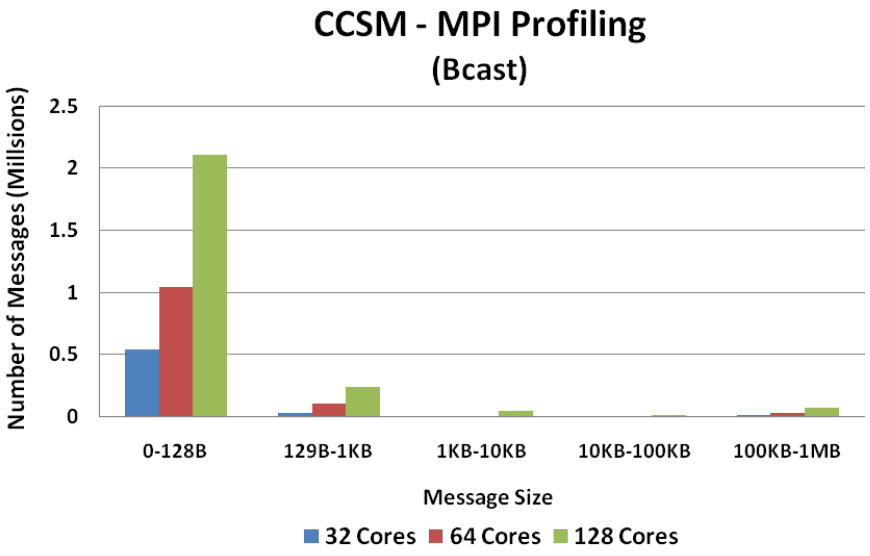
CCSM 4.0 MPI Profiling – MPI Functions

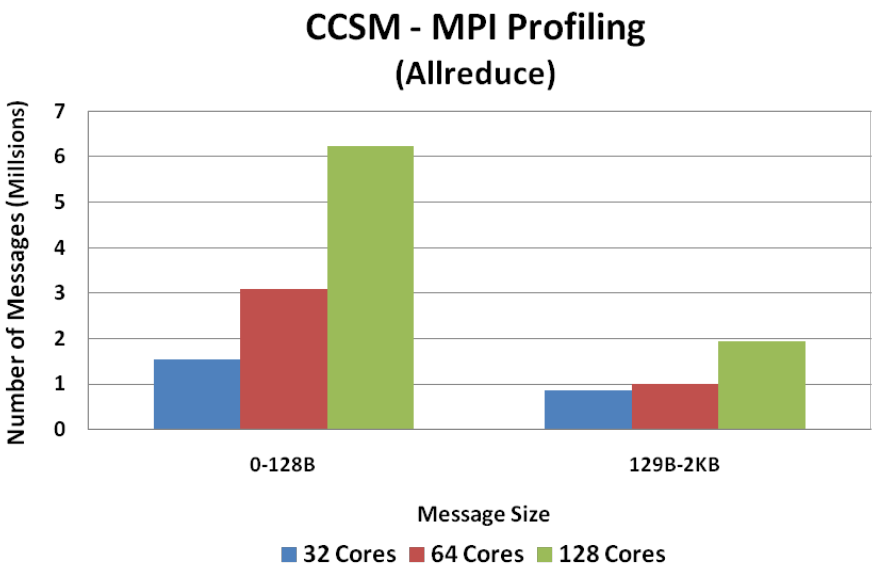
- **Mostly used MPI functions**
  - MPI_Allreduce and MPI_Bcast are the mostly used MPI functions
  - Total number of messages increases linearly as cluster size scales



CCSM 4.0 MPI Profiling – Message Size

- **Most messages are small size message**
  - For two main MPI functions: MPI_Bcast and MPI_Allreduce





CCSM 4.0 Profiling Summary

- **CCSM 4.0 was profiled to identify its communication patterns**
  - MPI collectives create the big communication overhead
  - Large number of small messages are used
  - Number of messages increases with cluster size
- **Interconnects effect to CCSM 4.0 performance**
  - Latency is critical to application performance
- **Balanced system – CPU, memory, Interconnect that match each other capabilities, is essential for providing application efficiency**

Thank You

HPC Advisory Council