HPC ADVISORY COUNCIL

OpenFOAM Performance Benchmark and Profiling

Jan 2010

AMD
The future is fusion

DELL

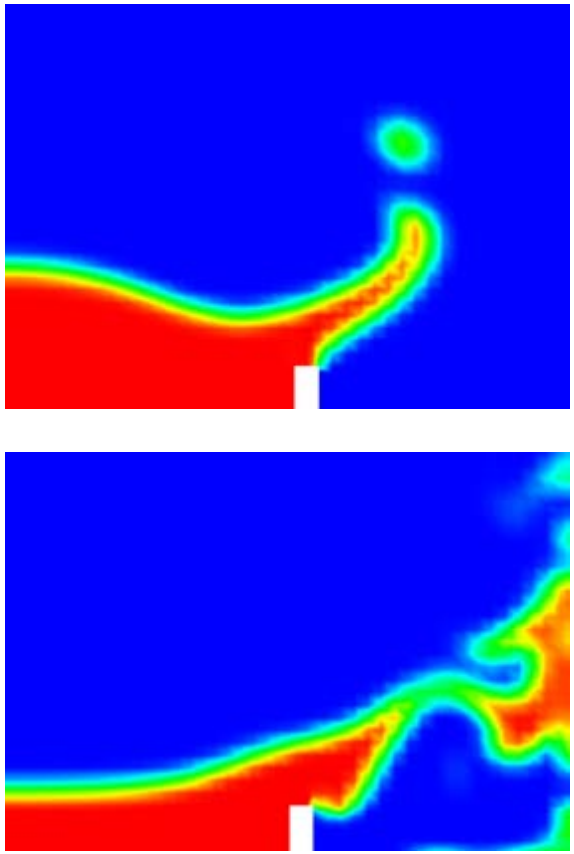
Mellanox
TECHNOLOGIES

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.opencfd.co.uk/openfoam

OpenFOAM

- **OpenFOAM® (Open Field Operation and Manipulation) CFD Toolbox can simulate**
  - Complex fluid flows involving
    - Chemical reactions
    - Turbulence
    - Heat transfer
  - Solid dynamics
  - Electromagnetics
  - The pricing of financial options



- **OpenFOAM is Open source, produced by OpenCFD Ltd**

Objectives

- **The presented research was done to provide best practices**
  - OpenFOAM performance benchmarking
  - Interconnect performance comparisons
  - Understanding OpenFOAM communication patterns
  - Power-efficient simulations
  - Compilation tips
- **The presented results will demonstrate**
  - Balanced compute system enables
    - Good application scalability
    - Power saving

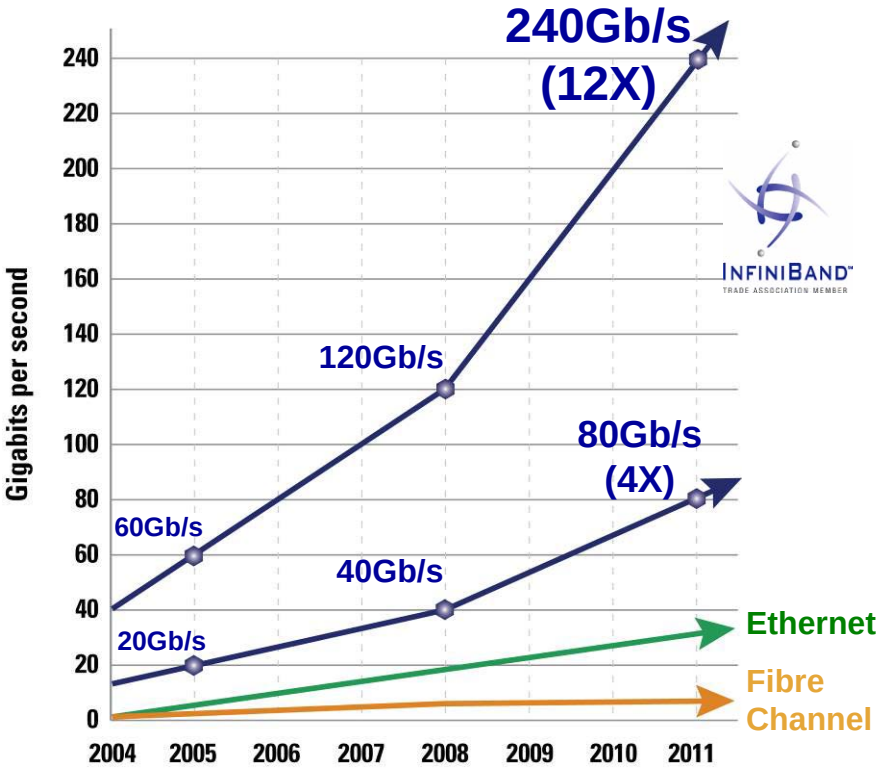
Test Cluster Configuration

- Dell™ PowerEdge™ SC 1435 24-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 1.4.1 InfiniBand SW stack
- MPI: OpenMPI-1.3.3
- Application: OpenFOAM 1.6
- Benchmark Workload
  - Lid-driven cavity flow

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**

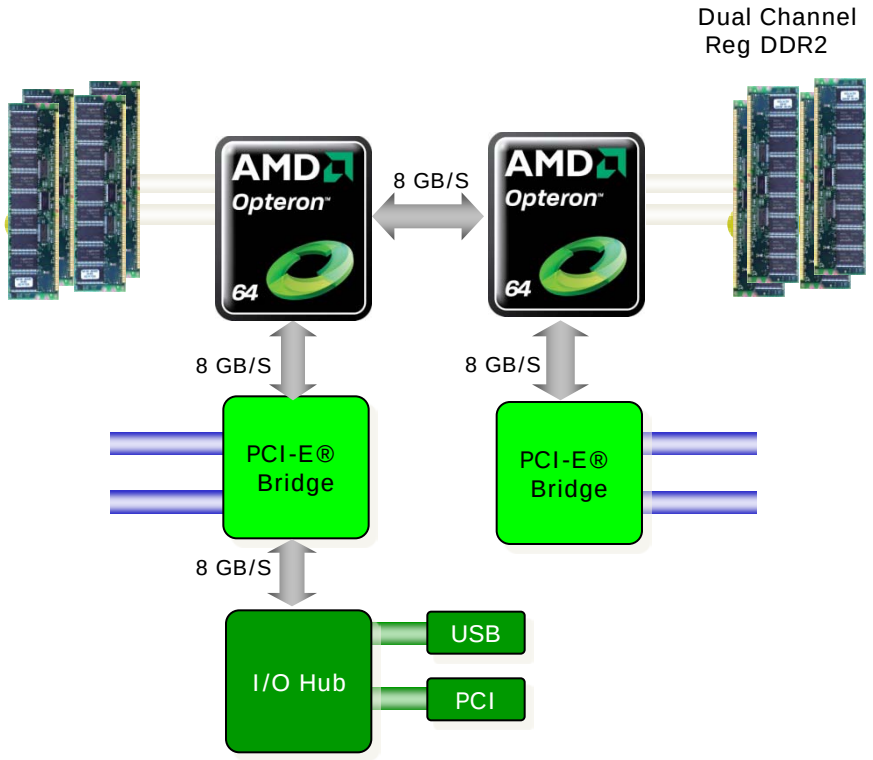
The InfiniBand Performance Gap is Increasing



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

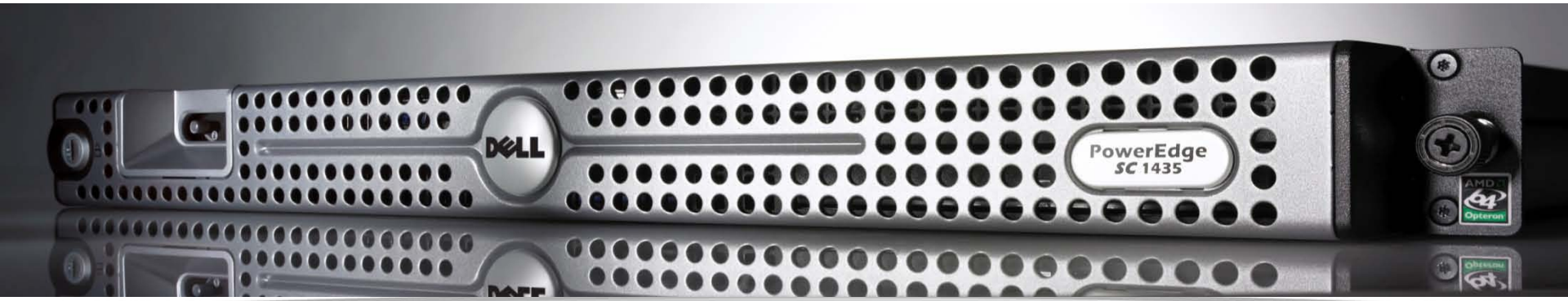


Dell PowerEdge Servers helping Simplify IT

- **System Structure and Sizing Guidelines**
  - 24-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

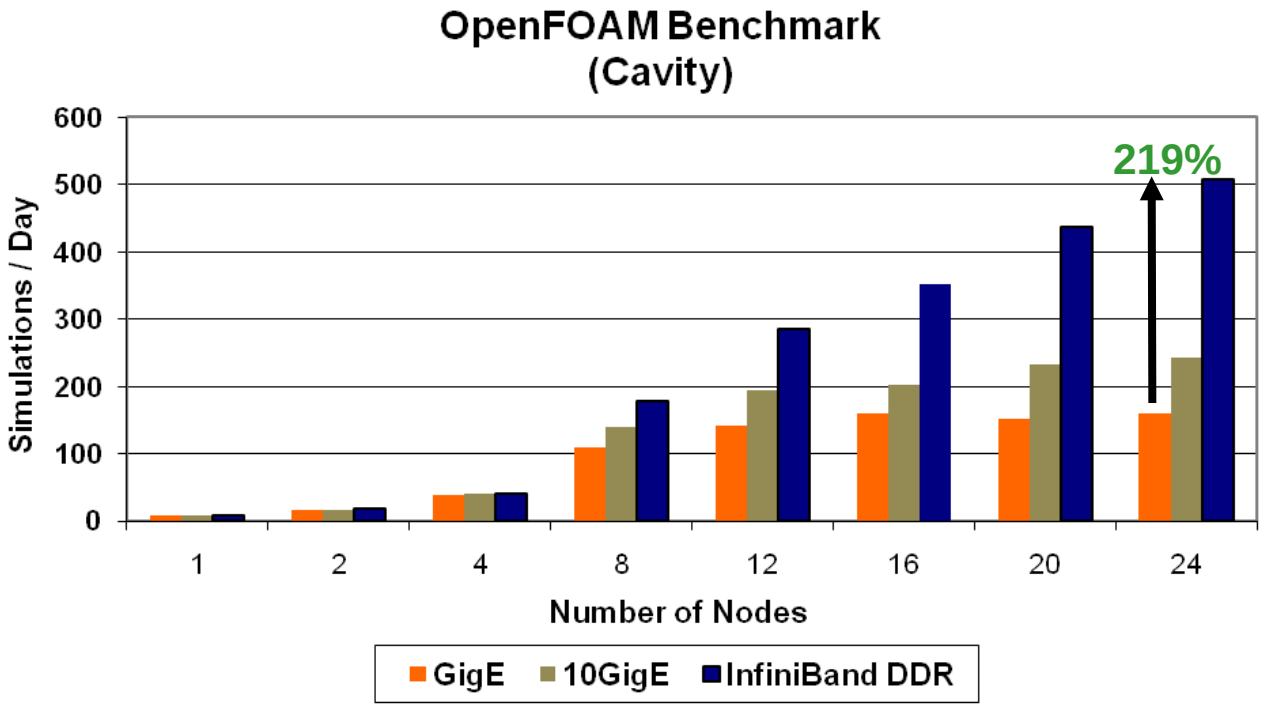
Dell PowerEdge™ Server Advantage

- **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**



OpenFOAM Benchmark Results

- **Input Dataset: Lid-driven cavity flow**
  - Mesh of 1000x1000 cells, icoFoam solver for laminar, 2D, 1000 steps
- **InfiniBand provides higher utilization, performance and scalability**
  - **Up to 219% higher performance versus GigE and 109% higher than 10GigE**

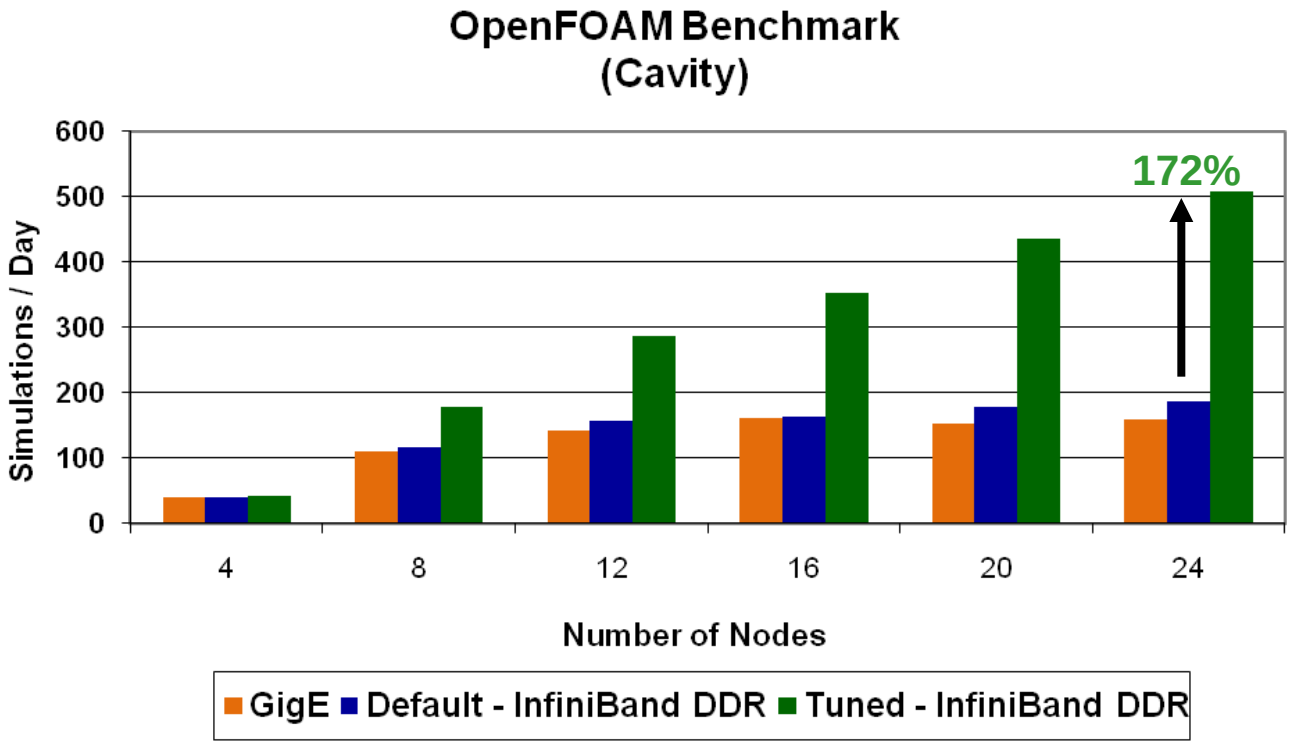


Higher is better

8-cores per node

OpenFOAM Performance Enhancement

- **Default OpenFOAM binary is not optimized over InfiniBand**
  - Precompiled Open MPI doesn't solve the issue
  - The ways to compile OpenFOAM properly is provided in the next slide
- **With proper optimization, InfiniBand based performance improves by 172%**



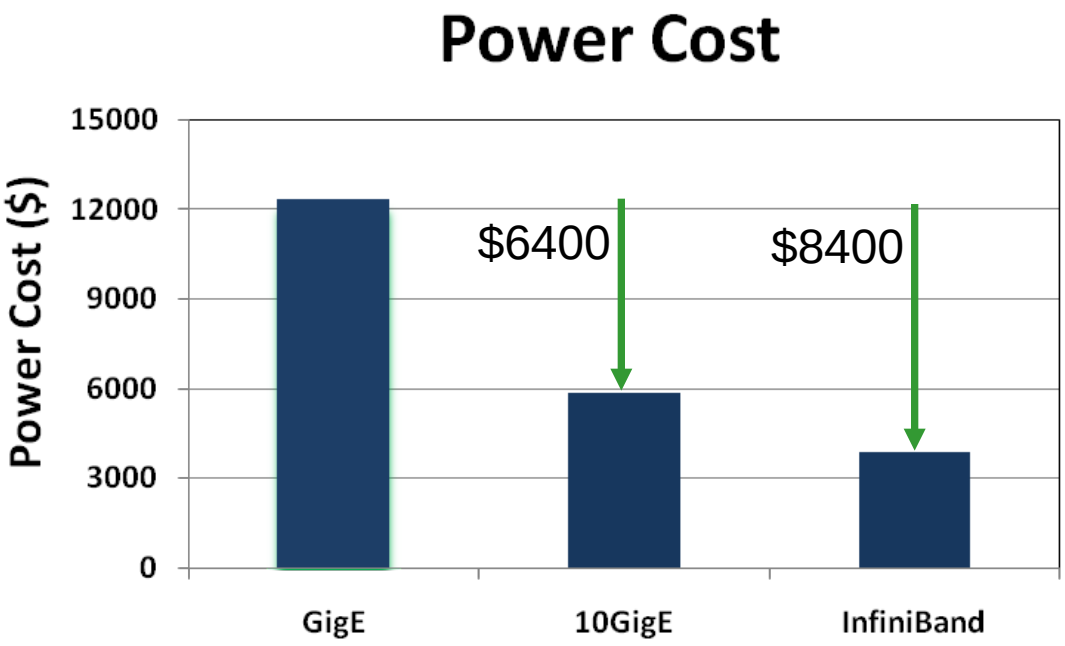
Higher is better

OpenFOAM Compilation Best Pratices

- **Two ways to compile OpenFOAM**
  - **Option 1:**
    - **Modify OpenFOAM-1.6/etc/bashrc to use MPI entry rather OPENMPI**
      - WM_MPLIB:=MPI
    - **Change MPI entry within settings.sh to system OpenMPI**
      - export MPI_HOME=/usr/mpi/gcc/openmpi-1.3.3
    - **Add the following to wmake/rules/linux64Gcc/mplib**
      - PFLAGS = -DOMPI_SKIP_MPICXX
      - PINC = -I$(MPI_ARCH_PATH)/include
      - PLIBS = -L$(MPI_ARCH_PATH)/lib64 –lmpi
  - **Option 2:**
    - **Keep the default OPENMPI entry in bashrc**
    - **Modify default Open MPI compiler option in ThirdParty-1.6/Allwmake**
      - Refer to Open MPI website for full compiling options
    - **Compiling with this option will take much longer (> 4 hours)**

Power Cost Savings with Different Interconnect

- **Dell economical integration of AMD CPUs and Mellanox InfiniBand**
  - Saves power up to $8400 to achieve same number of application jobs over GigE
  - Up to $6400 to achieve same number of application jobs with 10GigE
  - Yearly based for 24-node cluster
- **As cluster size increases, more power can be saved**



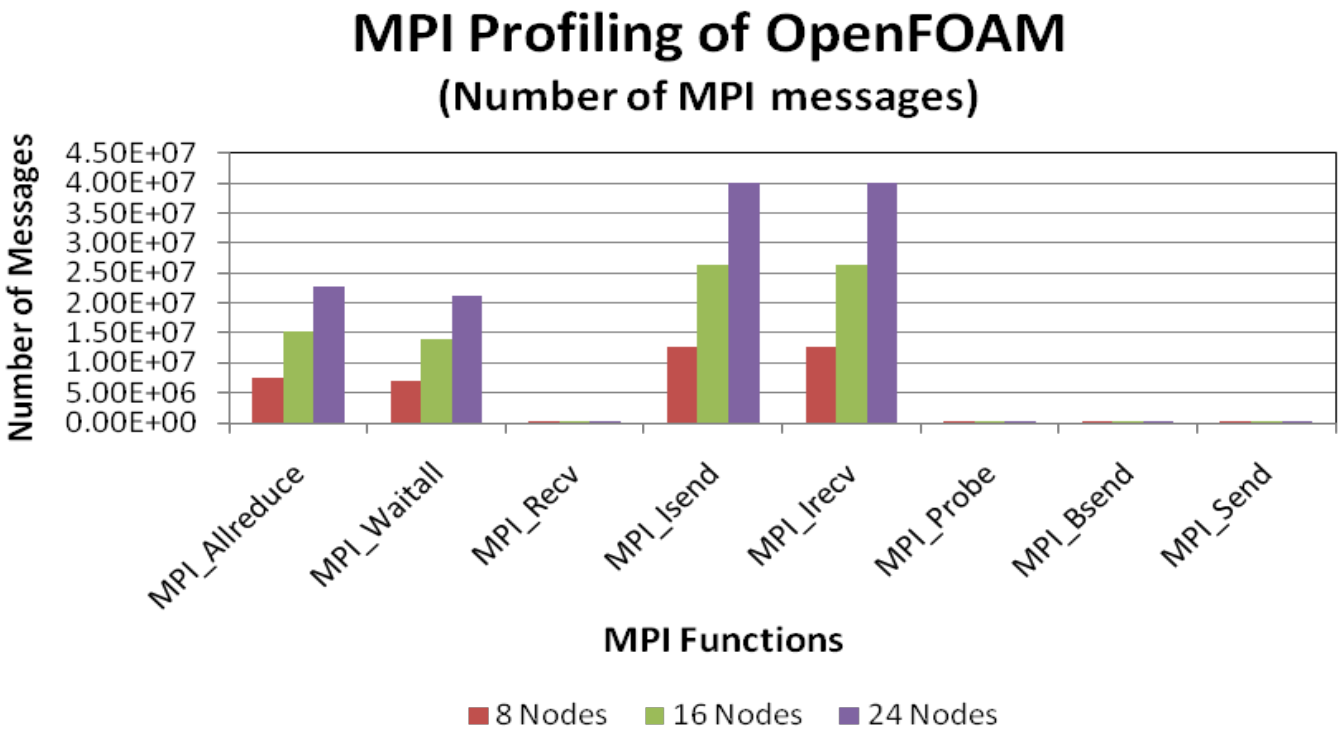
*$/KWh = KWh * $0.20*
For more information - http://enterprise.amd.com/Downloads/svrpwrusecompletefinal.pdf

OpenFOAM Benchmark Results Summary

- **Interconnect comparison shows**
  - InfiniBand delivers superior performance in every cluster size
  - Performance advantage extends as cluster size increases
- **InfiniBand enables power saving**
  - Up to $8400/year power savings versus GigE
  - Up to $6400/year power savings versus 10GigE
- **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

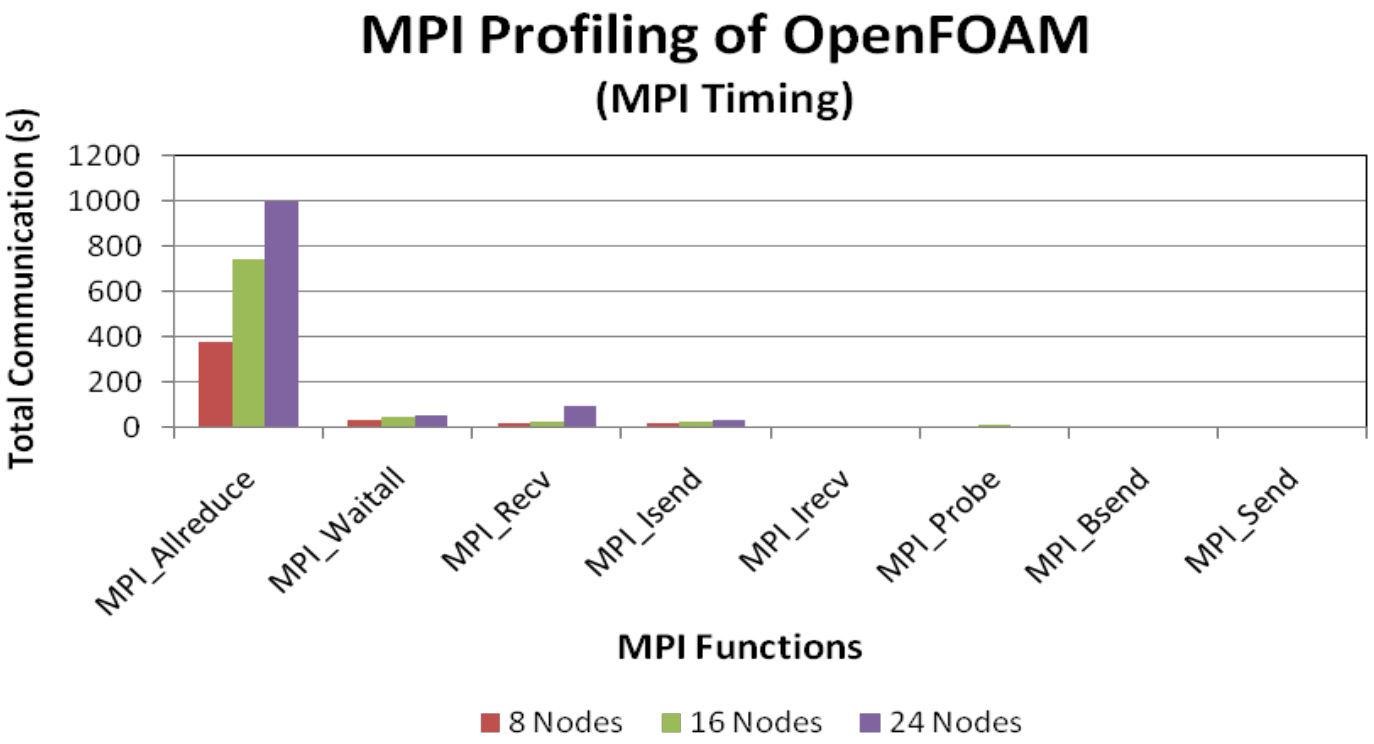
OpenFOAM MPI Profiling – MPI Functions

- **Mostly used MPI functions**
  - MPI_Allreduce, MPI_Waitall, MPI_Isend, and MPI_recv
  - Number of MPI functions increases with cluster size



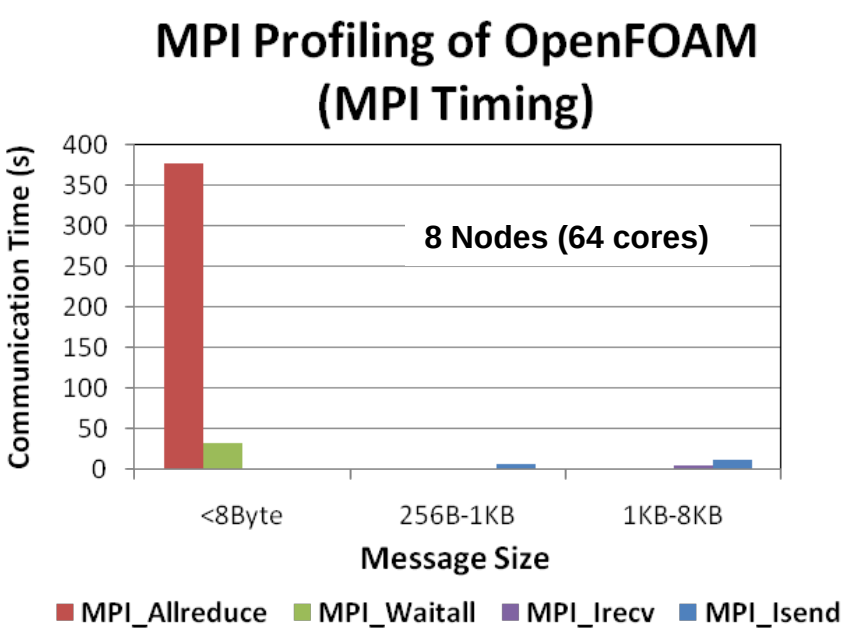
OpenFOAM MPI Profiling – MPI Timing

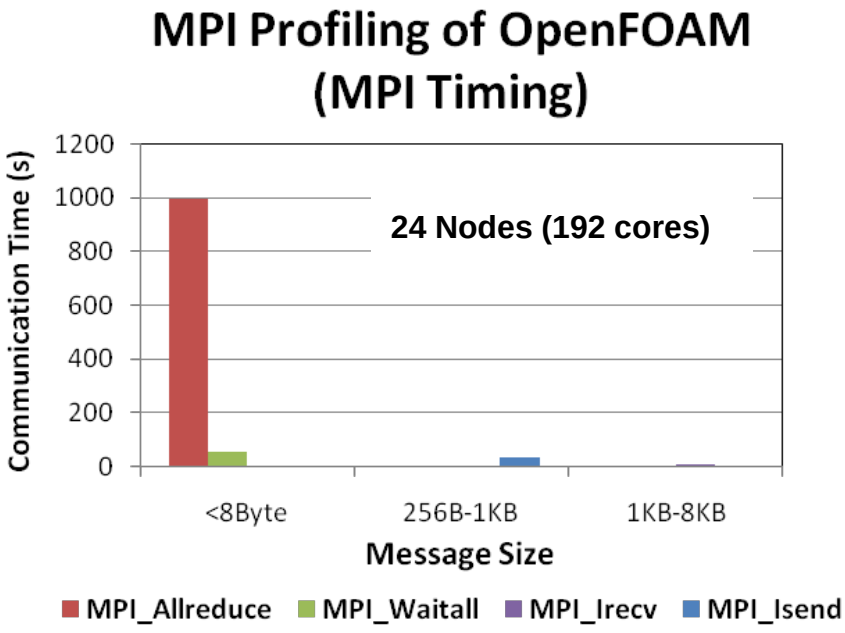
- **MPI_Allreduce, MPI_Recv, and MPI_Waitall show the highest communication overhead**



OpenFOAM MPI Profiling – Message Size

- **Large communication overhead is caused by**
  - Small messages handled by MPI_Allreduce





OpenFOAM Profiling Summary

- **OpenFOAM was profiled to identify its communication patterns**
  - MPI collective functions create the biggest communication overhead
  - Number of messages increases with cluster size
- **Interconnects effect to OpenFOAM performance**
  - Interconnect latency is critical to OpenFOAM performance
- **Balanced system – CPU, memory, Interconnect that match each other capabilities, is essential for providing application efficiency**

Thank You

HPC Advisory Council