

MM5 Modeling System Performance Research and Profiling

March 2009



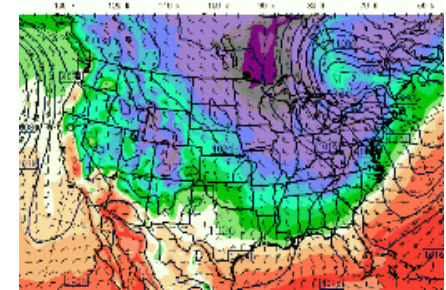
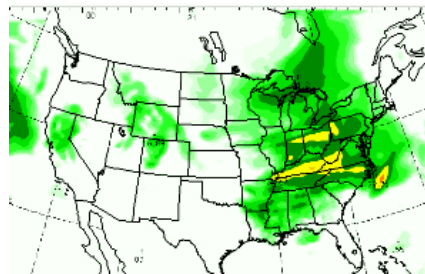
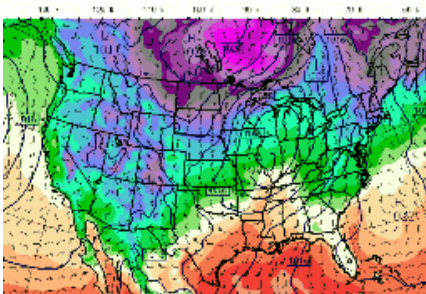
- **The following research was performed under the HPC Advisory Council activities**
 - AMD, Dell, Mellanox
 - HPC Advisory Council Cluster Center
- **For more info please refer to**
 - www.mellanox.com, www.dell.com/hpc, www.amd.com

- **Fifth-Generation NCAR/Penn State Mesoscale Model**

- Designed to simulate or predict mesoscale and regional-scale atmospheric circulation
- Mesoscale Meteorology is the study of weather systems smaller than synoptic scale systems but larger than microscale and storm-scale cumulus systems
 - Horizontal dimensions generally range from around 5 kilometers to several hundred kilometers
 - Examples: sea breezes, squall lines, and mesoscale convective complexes.
- <http://www.mmm.ucar.edu/mm5/>

- **Parallel version of MM5 with MPI enabled**

- Support execution of the model on distributed memory (DM) parallel machines

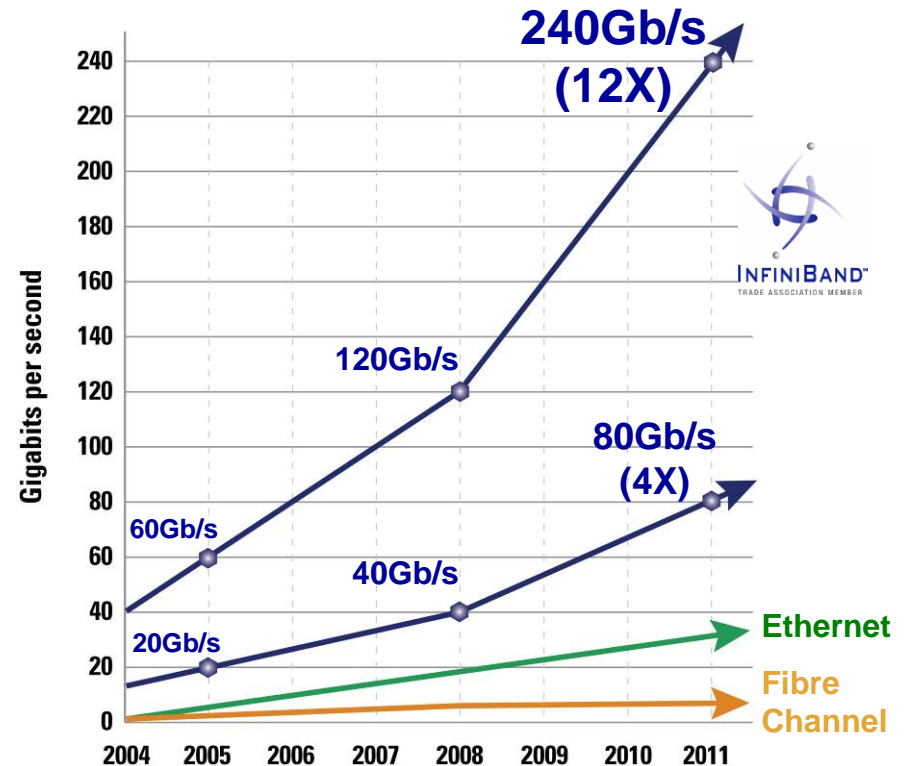


- **The presented research was done to provide best practices**
 - MM5 performance benchmarking
 - Interconnect performance comparisons
 - Ways to increase MM5 productivity
 - Understanding MM5 communication patterns

- **Dell™ PowerEdge™ SC 1435 24-node cluster**
- **Quad-Core AMD Opteron™ 2382 (“Shanghai”) CPUs**
- **Mellanox® InfiniBand ConnectX® DDR HCAs**
- **Mellanox® InfiniBand DDR Switch**
- **Memory: 16GB memory, DDR2 800MHz per node**
- **OS: RHEL5U2, OFED 1.4 InfiniBand SW stack**
- **MPI: Platform MPI 5.6.4, MVAPICH-1.1.0**
- **Application: MM5 Version 3**
- **Benchmark Workload**
 - T3A benchmark test case from NCAR

- **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/clock peak per core

- Integrated Memory Controller

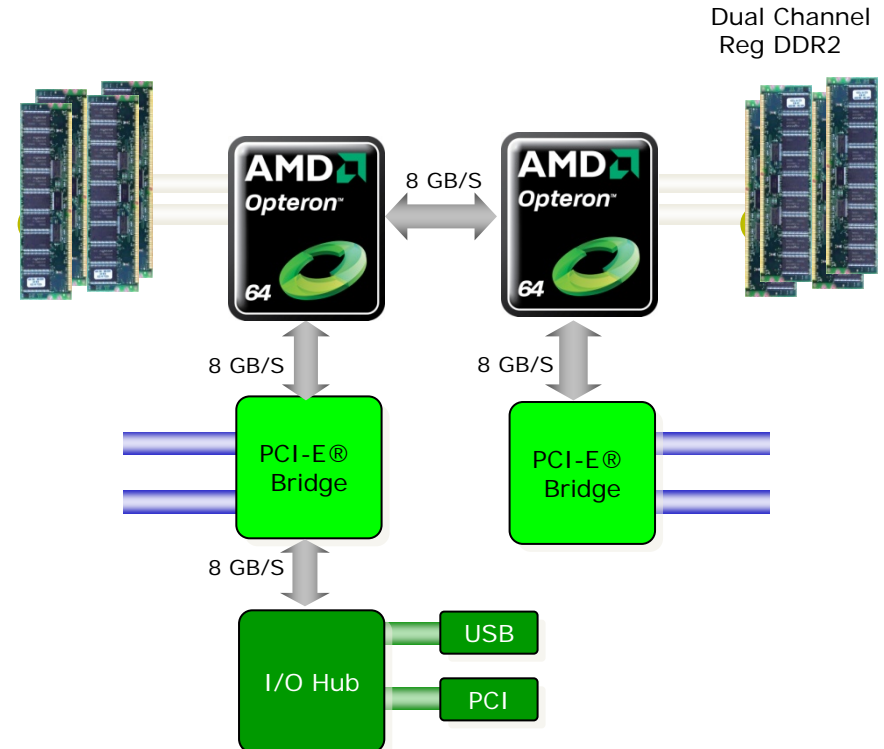
- Up to 12.8 GB/s
- DDR2-800 MHz or DDR2-667 MHz

- **Scalability**

- 48-bit Physical Addressing

- **Compatibility**

- Same power/thermal envelopes as 2nd / 3rd generation AMD Opteron™ processor



- **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

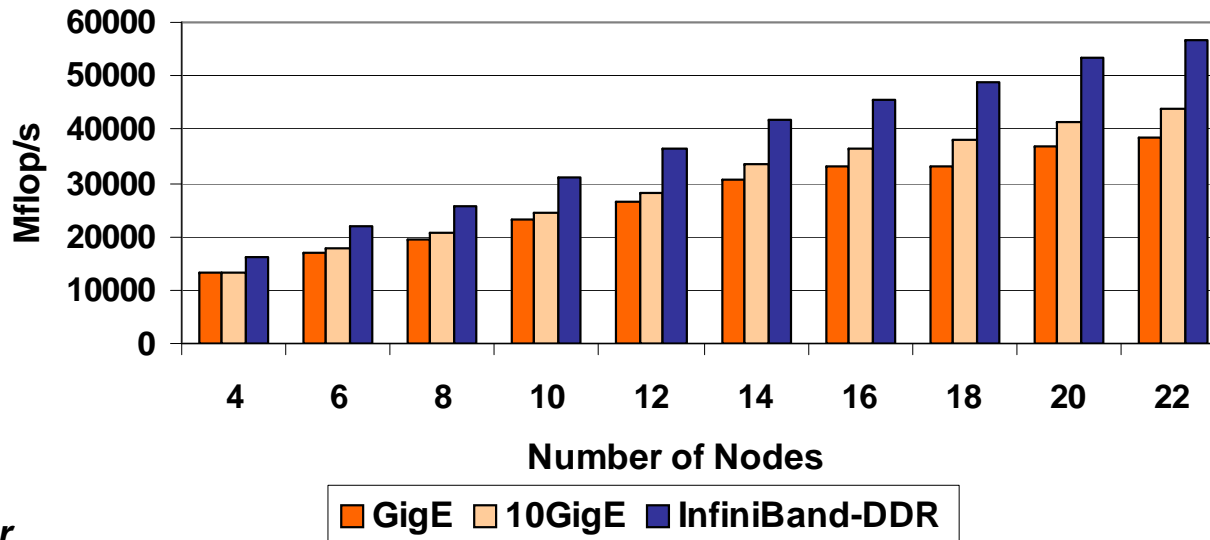
- **Workload Modeling**

- Optimized System Size, Configuration and Workloads
- Test-bed Benchmarks
- ISV Applications Characterization
- Best Practices & Usage Analysis



- **Input Data: T3A**
 - Resolution 36KM, grid size 112x136, 33 vertical levels
 - 81 second time-step, 3 hour forecast
- **InfiniBand DDR delivers higher performance in any cluster size**
 - Up to 46% versus GigE, 30% versus 10GigE

MM5 Benchmark Results - T3A



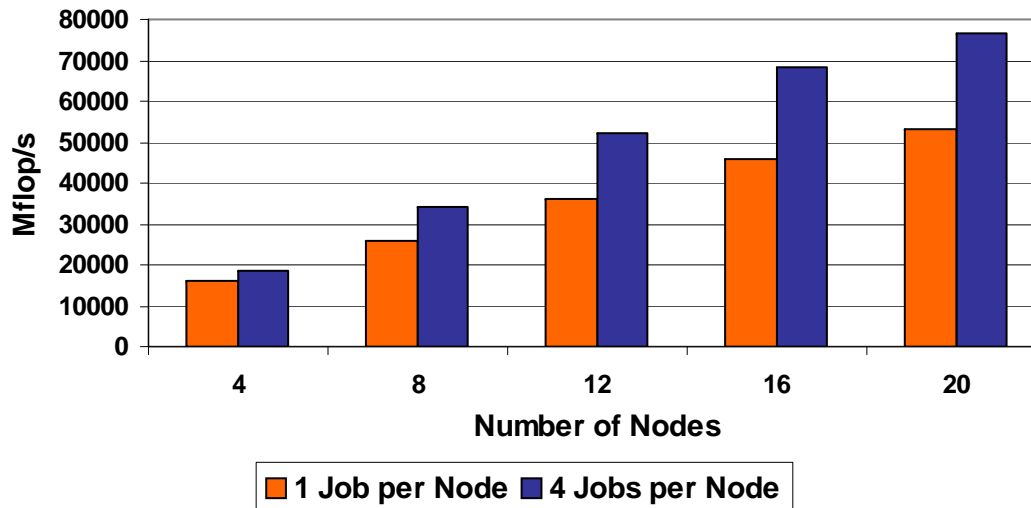
Higher is better

Platform MPI

MM5 Benchmark Results - Productivity

- **InfiniBand DDR increases productivity by allowing multiple jobs to run simultaneously**
 - Providing required productivity for weather simulations
- **Two cases are presented**
 - Single job over the entire systems
 - Four jobs, each on two cores per CPU per server
- **Four jobs per node increases productivity by up to 35%**
 - Utilizing Platform MPI dynamic binding of processes to cores
- **Multiple MPI jobs per node works well using Platform MPI due to its dynamic binding of processes to cores**

**MM5 Benchmark Results - T3A
InfiniBand DDR**

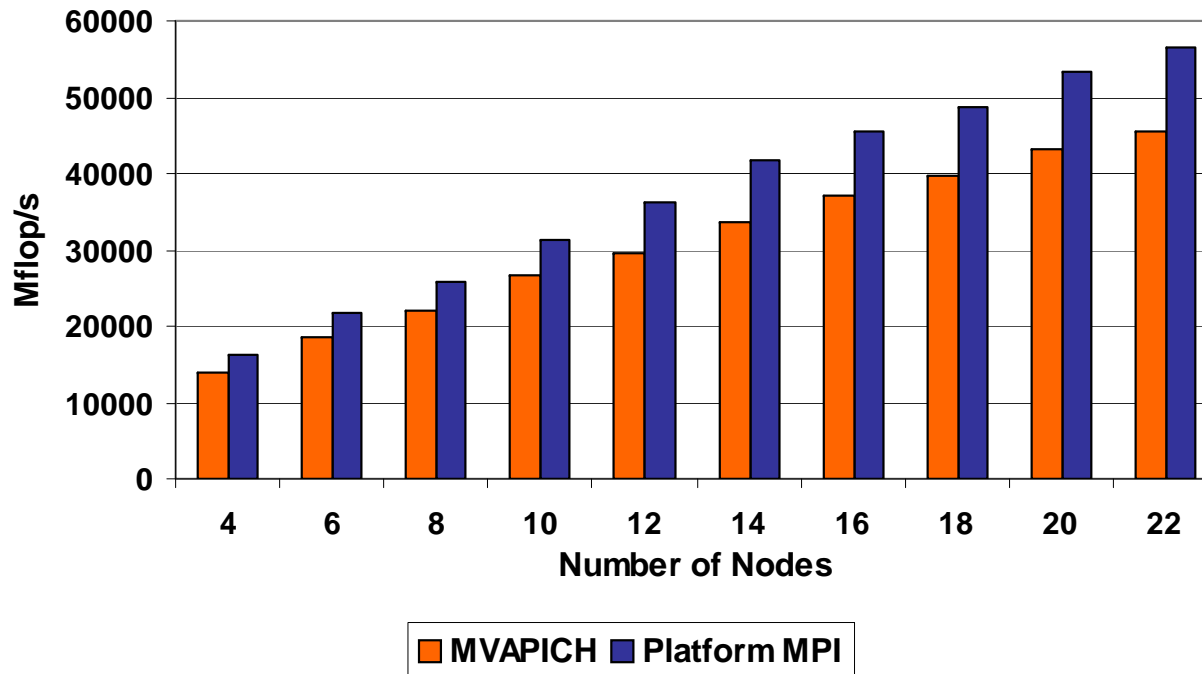


Higher is better

Platform MPI

- **Platform MPI demonstrates higher performance versus MVAPICH**
 - Up to 25% higher performance
 - Platform MPI advantage increases with increased cluster size

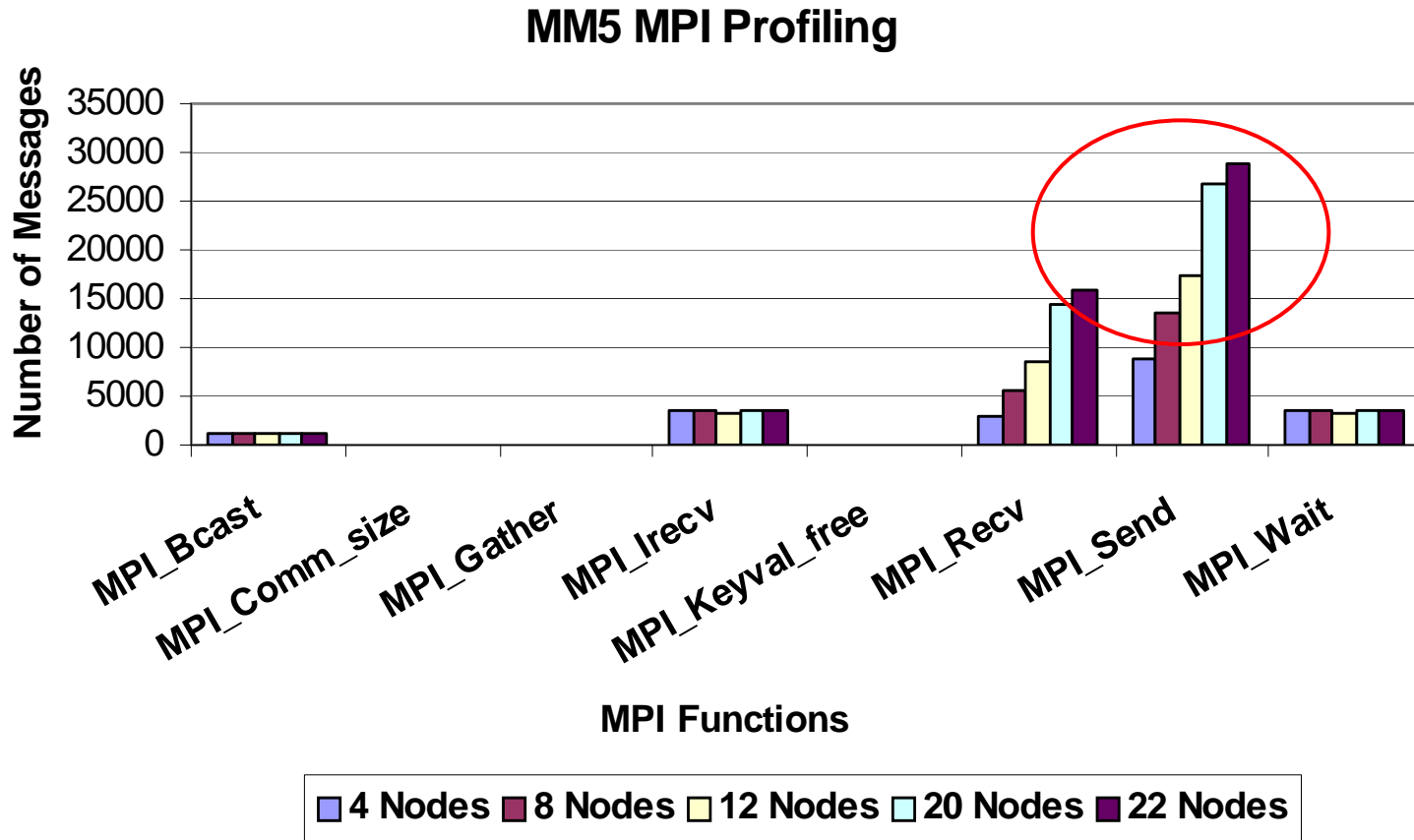
MM5 Benchmark Results - T3A



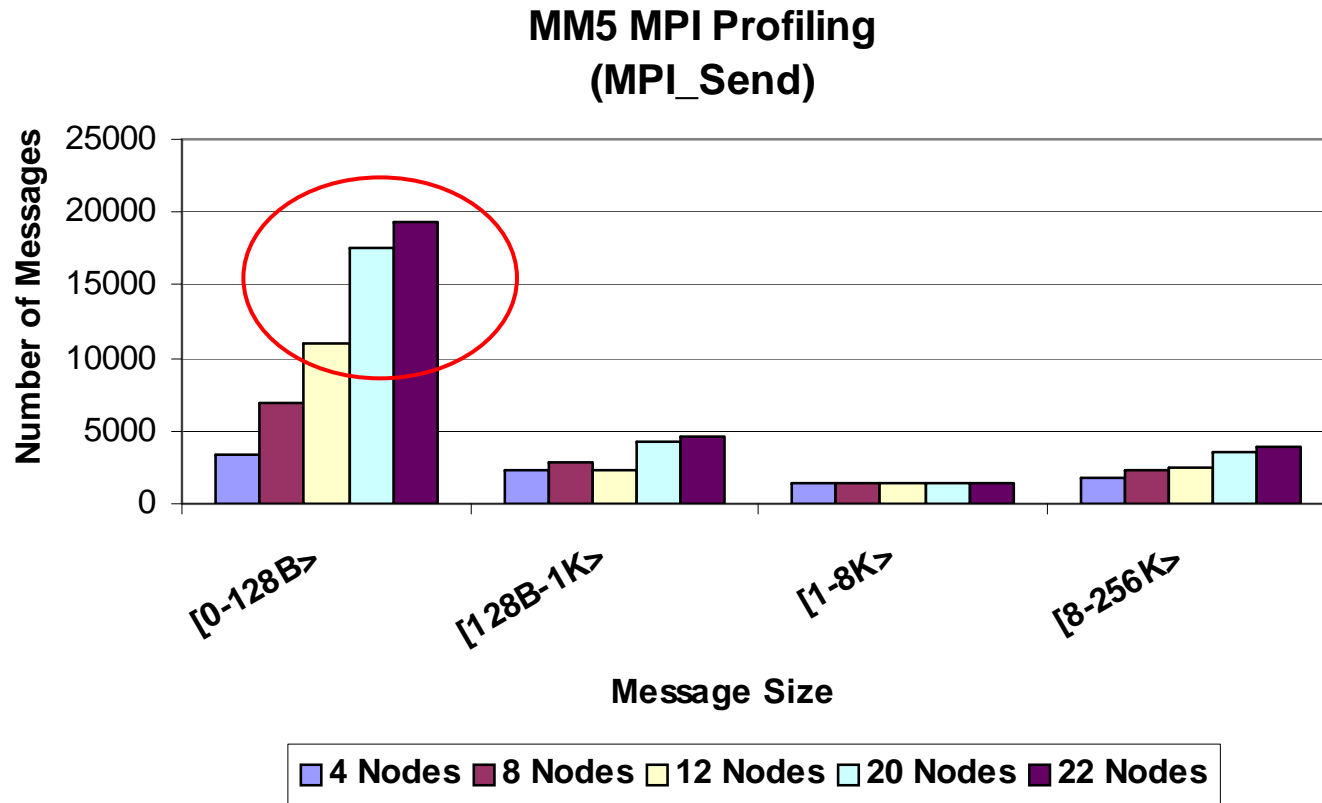
Higher is better

Single job on each node

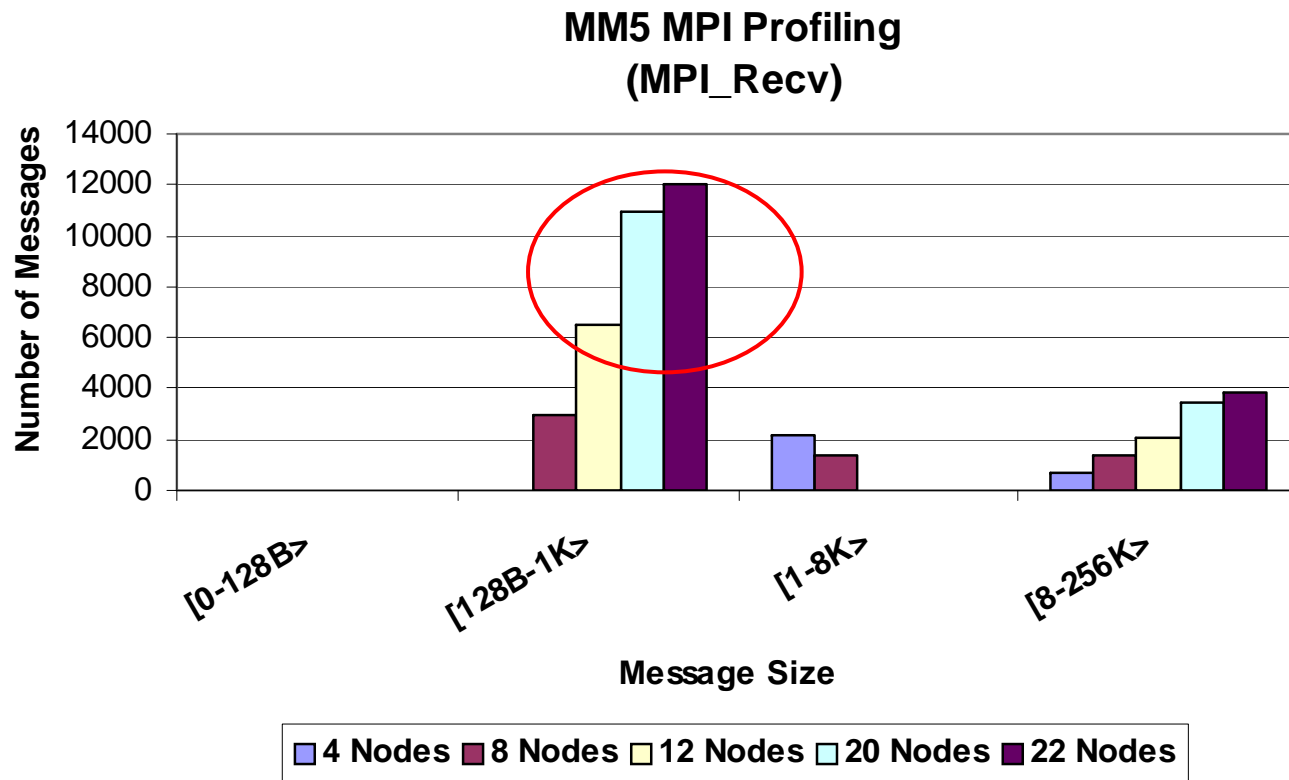
- **MPI_Send** and **MPI_Recv** are the mostly used MPI functions in MM5



- Most outgoing MPI messages are smaller than 128B



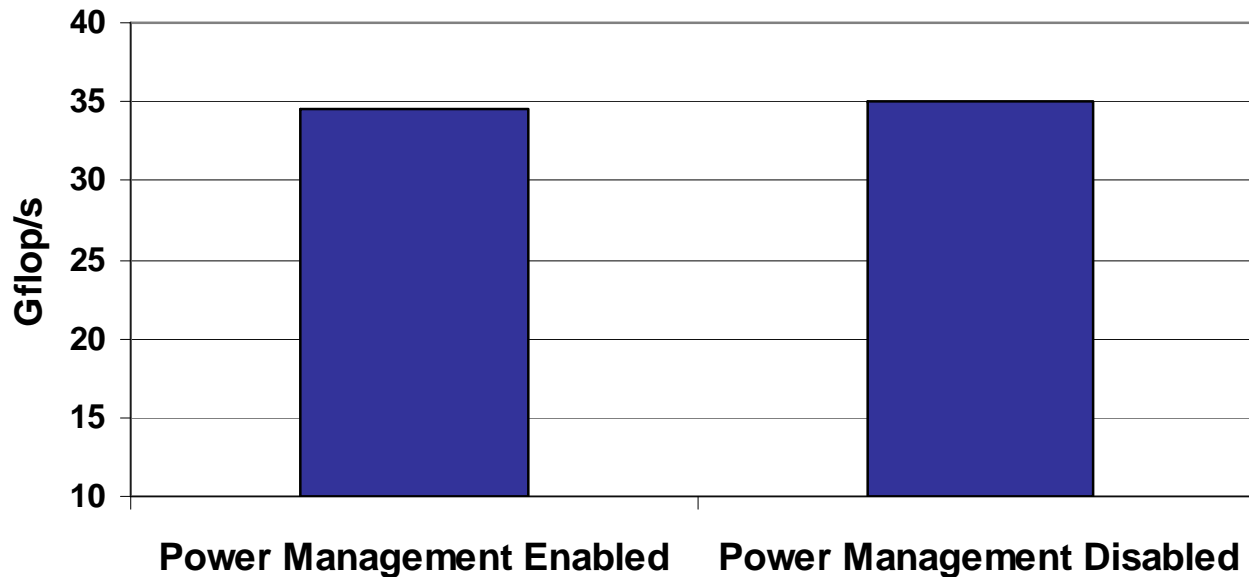
- Most received MPI messages are within 128Bytes to 1K
- Total number of medium size messages increases with cluster size



- **MM5 was profiled to determine networking dependency**
- **Majority of data transferred between compute nodes**
 - Small to medium size messages
 - Data transferred increases with cluster size
- **Most used message sizes**
 - <128B messages – MPI_Send
 - 128B-1KB and 8K-256K – MPI_Recv
- **Total number of messages increases with cluster size**
- **Interconnects effect to MM5 performance**
 - Interconnect latency and throughput for <256KB message range
 - Interconnect latency and throughput become critical with cluster size

- **Test Scenario**
 - 24 servers, 4 processes per node, 2 processes per CPU (socket)
- **Similar performance with power management enabled or disabled**
 - Only 1.4% performance degradation

MM5 Benchmark Results - T3A

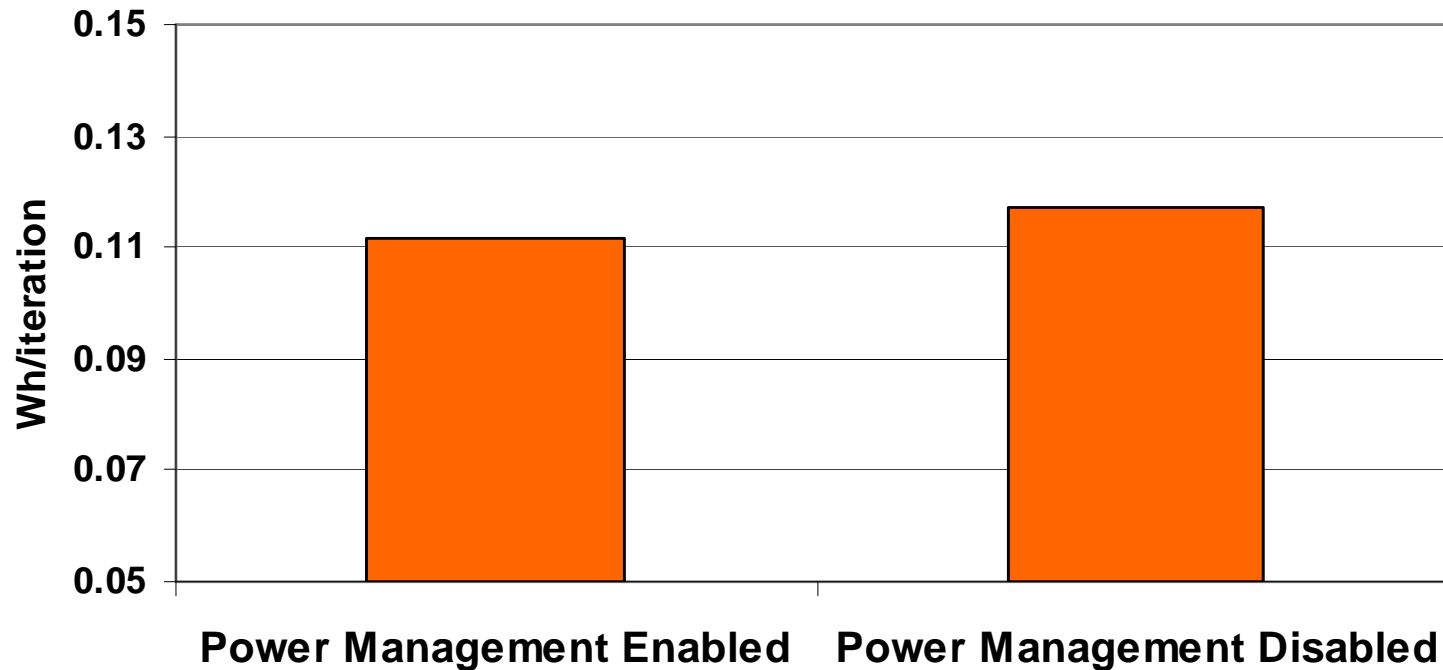


Higher is better

InfiniBand DDR

- Power management reduces 5.1% of total system power consumption

MM5 Benchmark - T3A Power Consumption

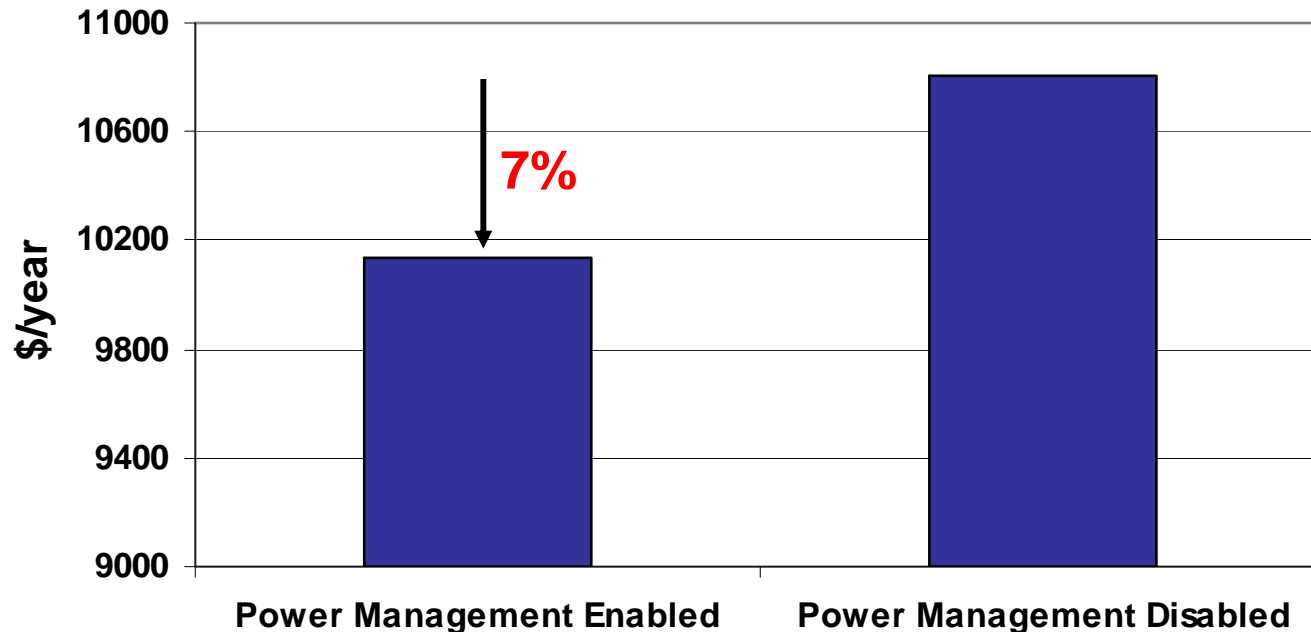


InfiniBand DDR

MM5 Benchmark – Power Cost Savings

- Power management saves 673\$/year for the 24-node cluster
- As cluster size increases, bigger saving are expected

**MM5 Benchmark - T3A
Power Cost Comparison**



24 Node Cluster

$\$/year = \text{Total power consumption/year (KWh)} * \0.20

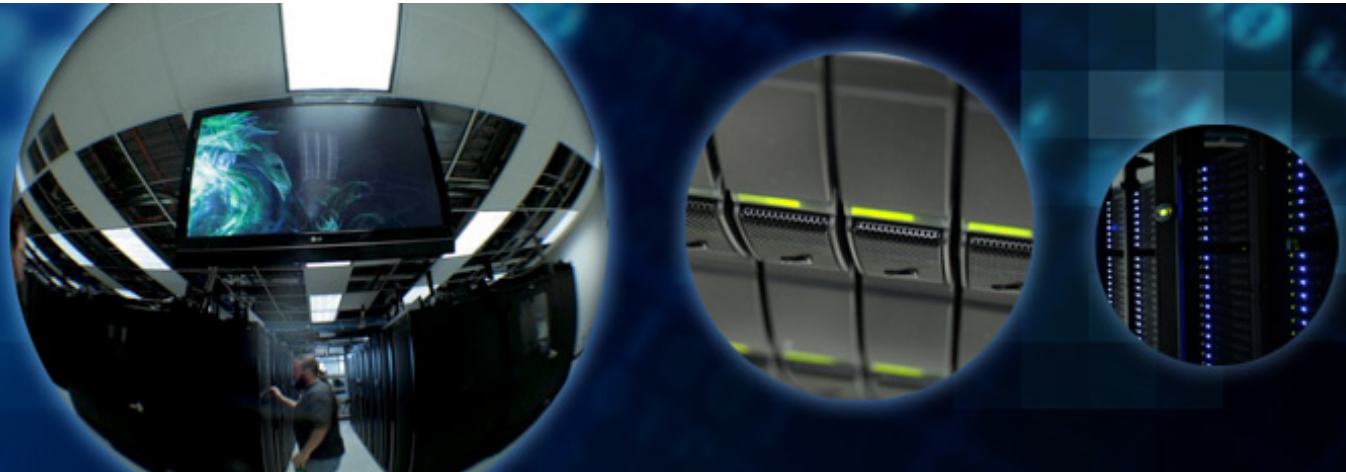
For more information - <http://enterprise.amd.com/Downloads/svrpwrusecompletefinal.pdf>

InfiniBand DDR

- **MM5 is widely used weather simulation software**
- **MM5 performance and productivity relies on**
 - Scalable HPC systems and interconnect solutions
 - Low latency and high throughput interconnect technology
 - NUMA aware application for fast access to memory
 - Reasonably job distribution can dramatically improves productivity
 - Increasing number of jobs per day while maintaining fast run time
- **Interconnect comparison shows**
 - InfiniBand DDR delivers superior performance and productivity
 - Versus GigE and 10GigE, Due to throughput and latency advantage
- **Power management provide 5.1% saving in power consumption**
 - Per 24-node system
 - \$673 power savings per year for 24-node cluster
 - Power saving increases with cluster size

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