

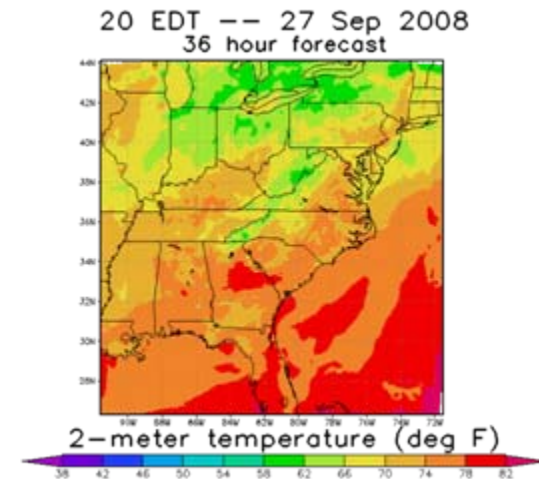
# Weather Research and Forecasting (WRF) Model Performance Research and Profiling

October 2008



- **The following research was performed under the HPC Advisory Council activities**
  - AMD, Dell, Mellanox
  - HPC Advisory Council Cluster Center
- **The participating members would like to thank John Michalakes for his support and guidelines**
- **For more info please refer to**
  - [www.mellanox.com](http://www.mellanox.com), [www.dell.com/hpc](http://www.dell.com/hpc), [www.amd.com](http://www.amd.com)

- **The Weather Research and Forecasting (WRF) Model**
  - Numerical weather prediction system
  - Designed for operational forecasting and atmospheric research
- **WRF developed by**
  - National Center for Atmospheric Research (NCAR),
  - The National Centers for Environmental Prediction (NCEP)
  - Forecast Systems Laboratory (FSL)
  - Air Force Weather Agency (AFWA)
  - Naval Research Laboratory
  - Oklahoma University
  - Federal Aviation Administration (FAA)

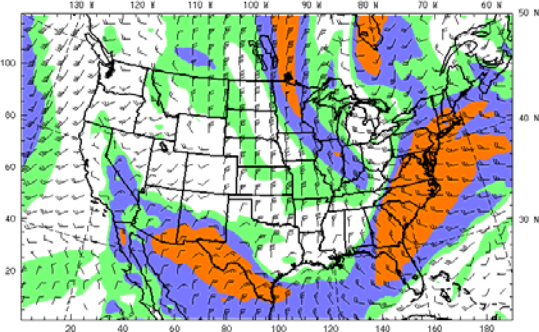


- **An application for weather forecasting and research**
- **The latest numerical program model to be adopted by**
  - The National Weather Service
  - The U.S. military and private meteorological services
  - Meteorological services worldwide
- **The Hurricane Weather Research and Forecasting (HWRF) model is a specialized version of WRF**
  - Became operational in 2007
- **WRF is a mesoscale model**
  - Uses a grid spacing between 4 and 12.5 kilometers
  - Vertical grid spacing between 25 and 37 divisions
- **WRF Real-time Forecasting**
  - <http://www.wrf-model.org/plots/wrfrealtime.php>

# The WRF Usage

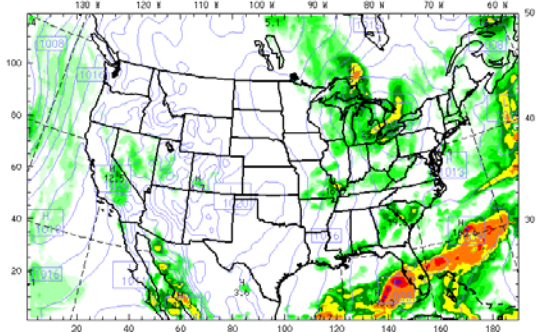
- **The WRF model includes**
  - Real-data and idealized simulations
  - Various lateral boundary condition options
  - Full physics options
  - Non-hydrostatic and hydrostatic
  - One-way, two-way nesting and moving nest
  - Applications ranging from meters to thousands of kilometers

ARW WRF - 30KM-NEST - NCAR/MMM  
Fcst. 21 h Valid: 21 UTC Tue 30 Sep 08 (15 MDT Tue 30 Sep 08)  
Supercell type (9-10 km rel. flow) sa= 5  
Supercell motion vectors



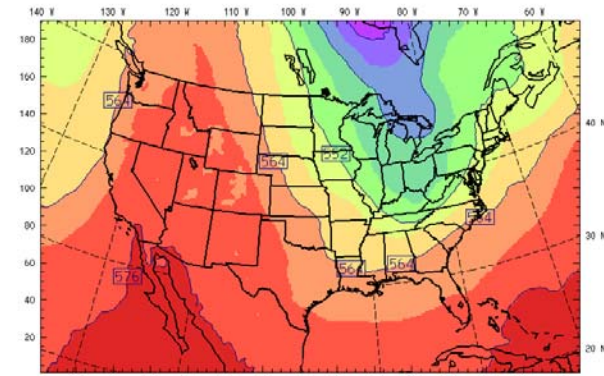
BASE VECTORS: FULL DATA = 10 kts  
Model Info: V3.0 KF YSU PBL WSM ScLasso Noah LEM 30 km, 34 levels, 120 sec  
LX: RRTH S8, Dudkita DIFF, stapsie KM: 20 Saagar

ARW WRF - 30KM-NEST - NCAR/MMM  
Fcst. 18 h Valid: 18 UTC Tue 30 Sep 08 (12 MDT Tue 30 Sep 08)  
Total precip. since h 0  
Total precip. since h 0  
Sea-level pressure sw= 4



CONTOURS: UNITS=hPa L0F= 1014.0 H0F= 1033.0 INTERVAL=X 2.0000  
CONTOURS: UNITS=mm LOT= 300.00 H0H= 600.00 INTERVAL=X 2.0000  
Model Info: V3.0 KF YSU PBL WSM ScLasso Noah LEM 30 km, 34 levels, 120 sec  
LX: RRTH S8, Dudkita DIFF, stapsie KM: 20 Saagar

20km ARW WRF, GFS-init -- NCAR/MMM  
Fcst. 18 h Valid: 18 UTC Wed 01 Oct 08 (12 MDT Wed 01 Oct 08)  
1000 to 0500 hPa thickness sa= 2



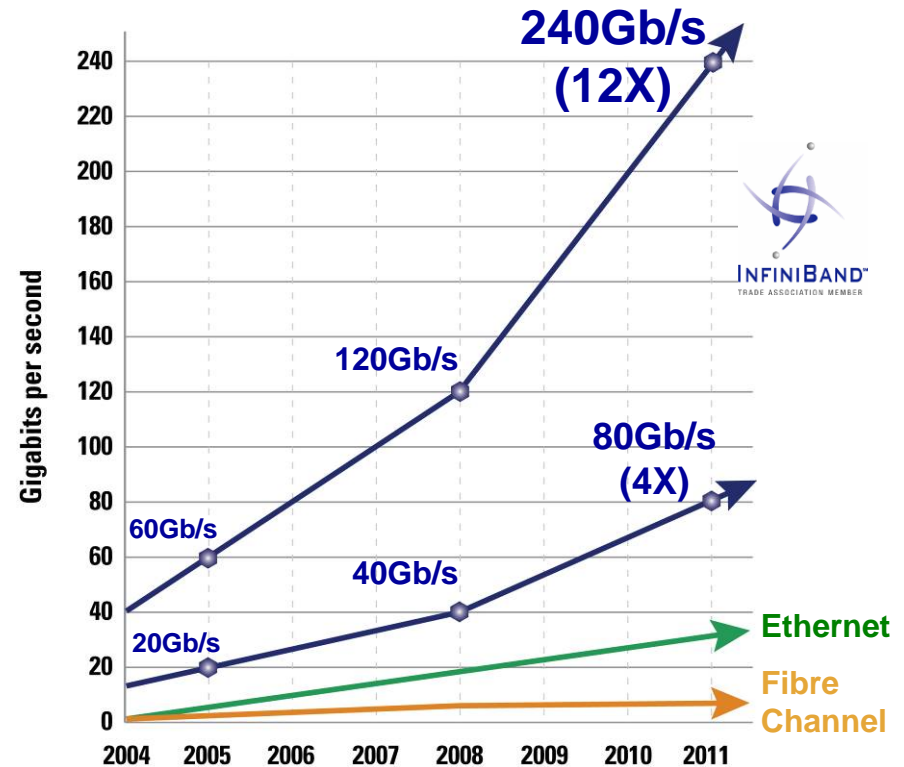
CONTOURS: UNITS=gpm LOT= 658.00 H0H= 678.00 INTERVAL= 18.0000  
Model Info: V3.0.1.1 C3 MFJ PBL Thompson Noah LEM 30 km, 30 levels, 178 sec  
LX: RRTH S8, Goddard DIFF, stapsie KM: 20 Saagar

- **The presented research was done to provide**
  - WRF Model performance benchmarking
  - Interconnect comparisons and how they impact WRF performance
  - Ways to increase WRF productivity
  - WRF networking profiling and determination of sensitivity points
  - MPI libraries comparisons

- **Dell PowerEdge SC 1435 24-node cluster**
- **Quad-Core AMD Opteron™ 2358 SE CPUs**
- **Mellanox® InfiniBand ConnectX® DDR HCAs**
- **Memory: 16GB memory, DDR2 677MHz per node**
- **OS: RH 5.1, OFED 1.3 InfiniBand SW stack**
- **MPI: Open MPI 1.3, MVAPICH 1.1, HP MPI 2.2.7**
- **Application: WRF V3, 12km CONUS benchmark case**
- **Compiler: Gfortran v4.2**
  - **Flags: FCOPTIM= -O3 -ffast-math -ftree-vectorize -ftree-loop-linear -funroll-loops**

- **Industry Standard**
  - Hardware, software, cabling, management
  - Design for clustering and storage interconnect
- **Price and 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
- 2MB L3 Cache

- Direct Connect Architecture

- HyperTransport™ Technology 1.0
- Up to 8 GB/s

- Floating Point

- 128-bit FPU per core
- 4 FLOPS/clock peak per core

- Memory

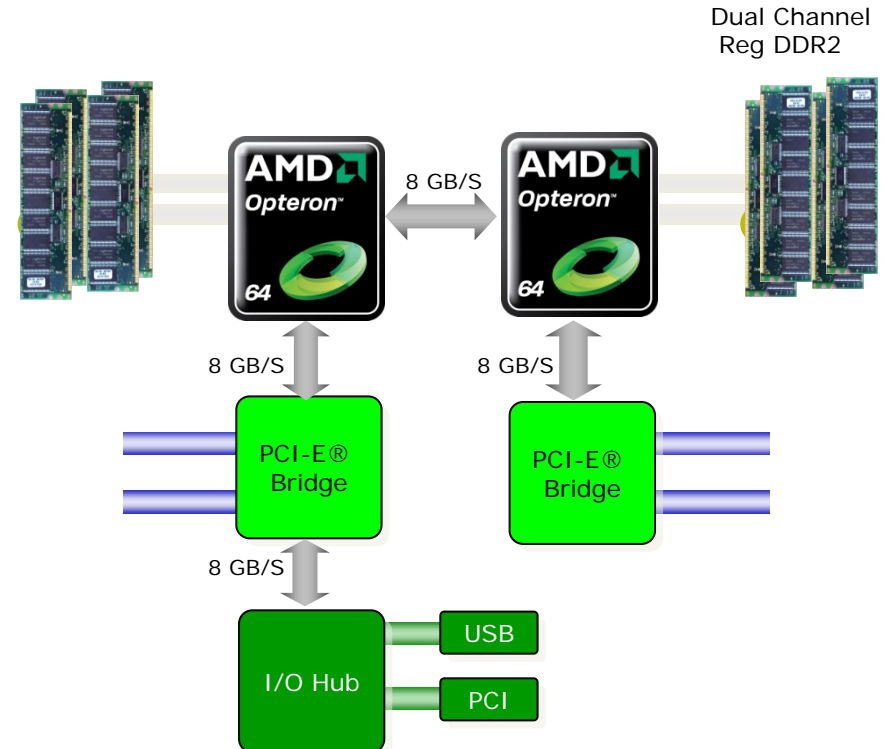
- 1GB Page Support
- DDR-2 667 MHz

- **Scalability**

- 48-bit Physical Addressing

- **Compatibility**

- Same power/thermal envelopes as Second-Generation AMD Opteron™ processor



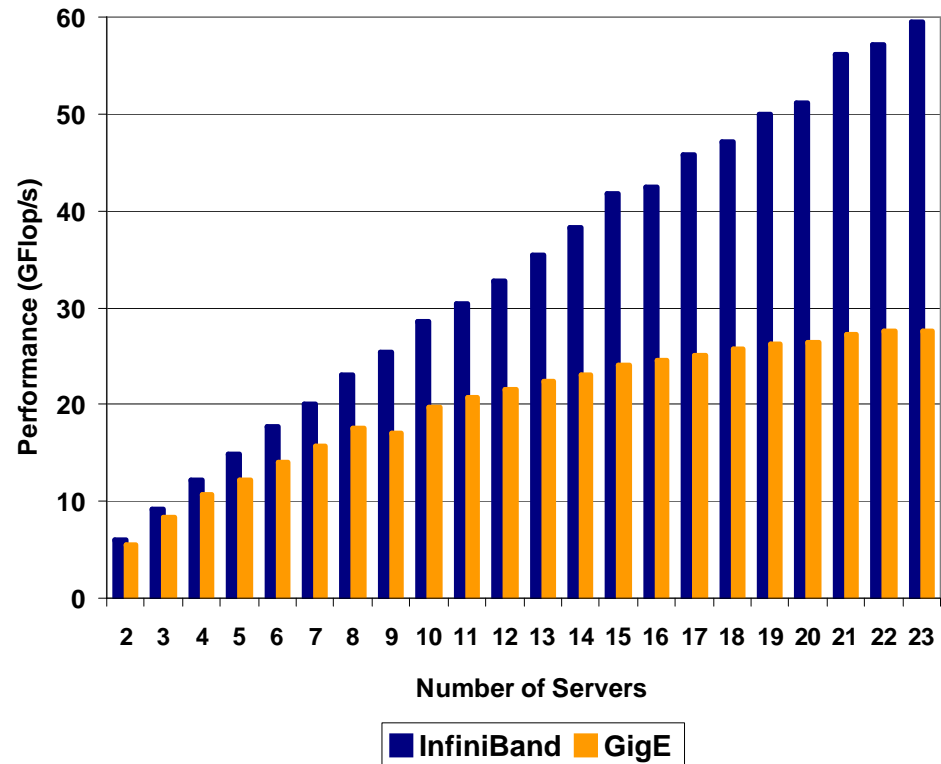
- **Dell HPC Solutions**
  - High Performance, Scalable Architectures
  - Professional Deployment and Support Services
  - Maximum Efficiency, Cost, Power, Performance & Productivity
- **System Sizing Guidelines**
  - Building Block Foundations
  - Integrated and Validated Architectures
- **Workload Modeling**
  - Optimized System Size, Configuration and Workloads
  - Test-bed Benchmarks
  - ISV Applications Characterization
  - Best Practices & Usage Analysis



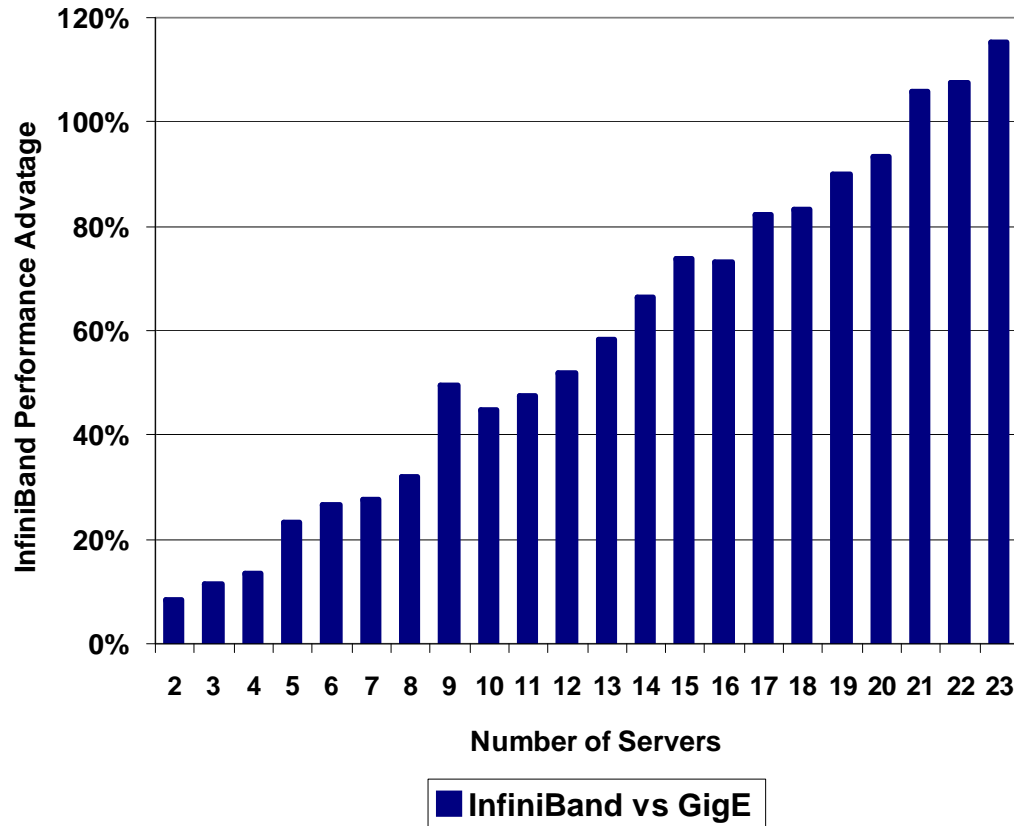
# WRF Performance Results - Interconnect

- Comparison between clustering interconnects - InfiniBand and GigE
- InfiniBand high speed interconnect enables almost linear scaling
  - Maximized system performance and enable faster simulations
- Gigabit Ethernet limits WRF performance and slow down simulations

WRF Benchmark Results - Conus 12Km



### WRF Benchmark Results - InfiniBand vs GigE

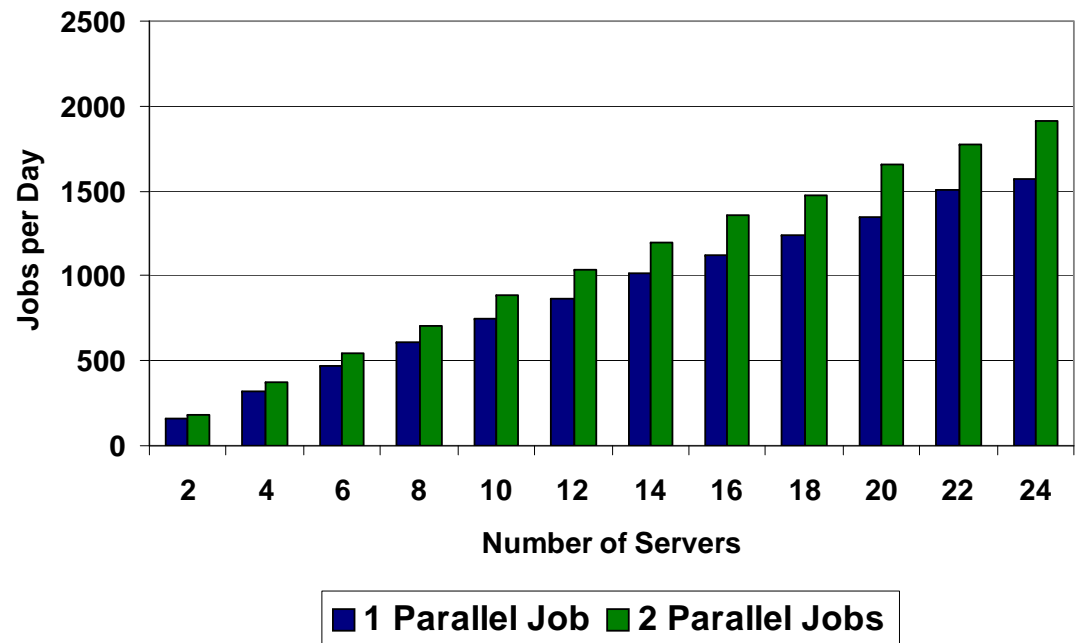


**InfiniBand increases WRF performance by up to 115%**

For cluster size of 24 nodes, higher numbers expected with larger cluster size

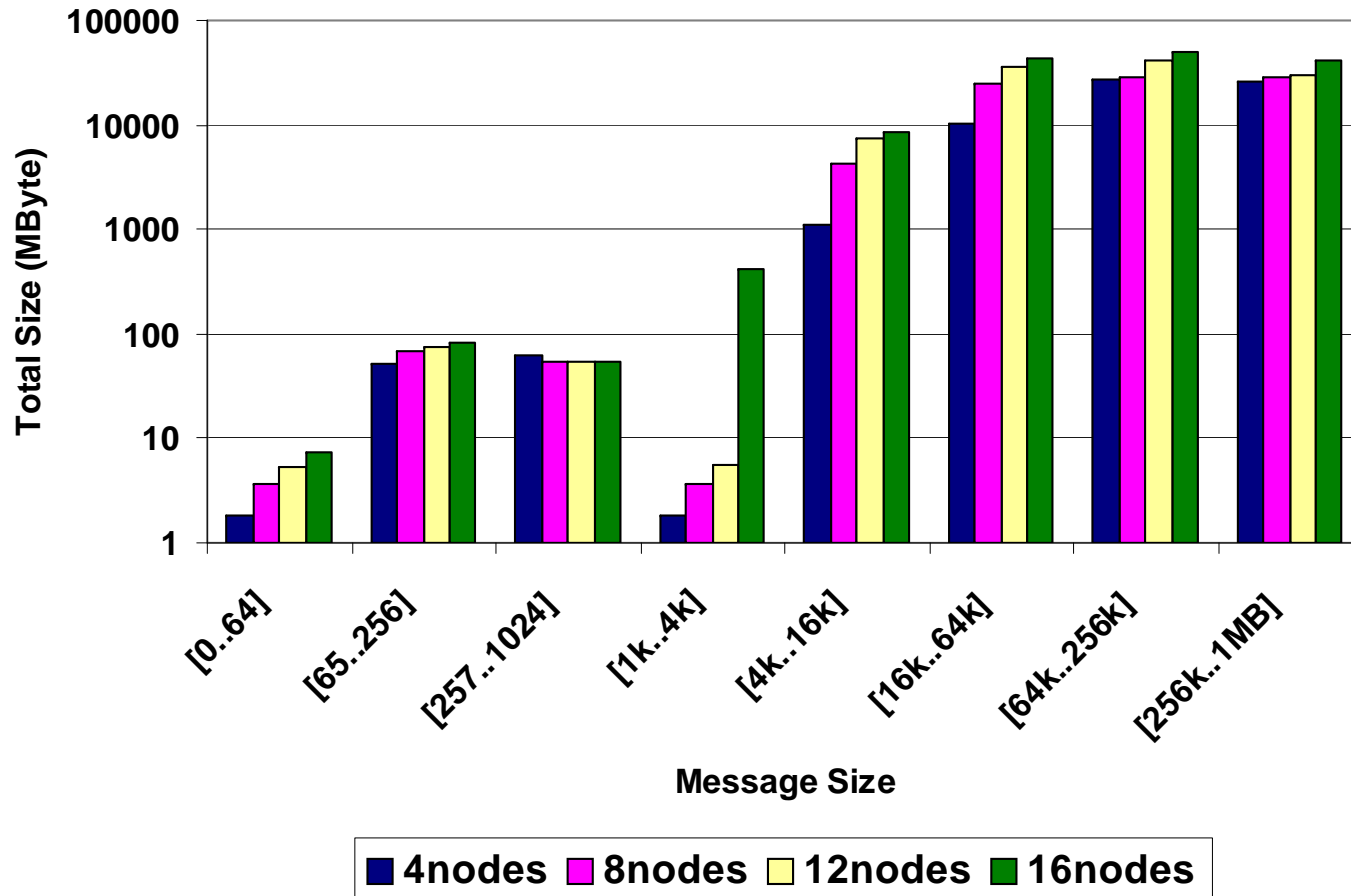
- Utilizing CPU affinity for higher productivity
- Two cases
  - Single job over the entire systems
  - Two jobs, each utilized single CPU in every server (CPU affinity)
- CPU affinity enables up to 20% more jobs per day

Increasing WRF Productivity via CPU Affinity

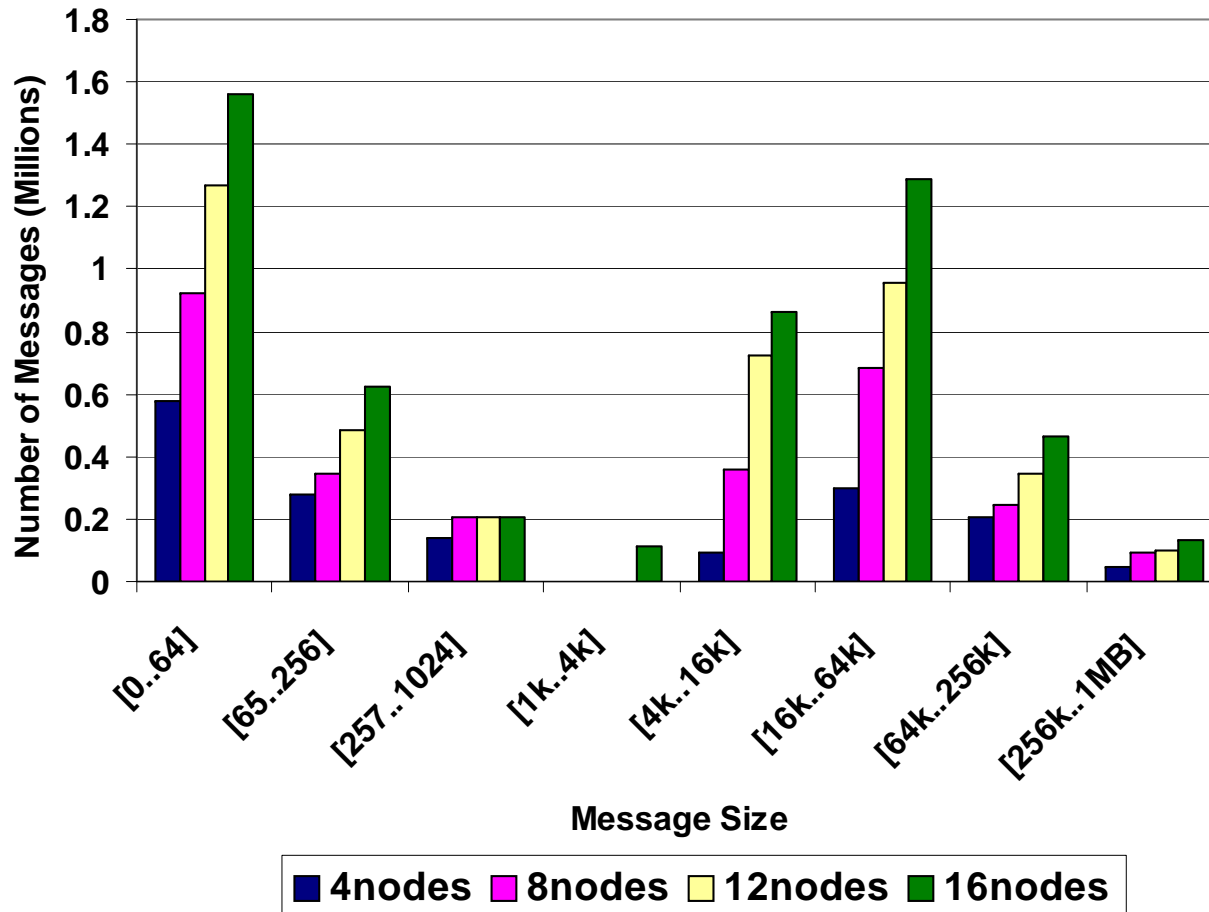


- **WRF model was profiled to determine dependency networking capabilities**
- **Majority of data transferred between compute nodes**
  - Done with 16KB-1MB message size
  - Data transferred increases with cluster size in those message sizes
- **Most used message sizes**
  - <64B messages – mainly synchronizations
  - 16KB-64KB – mainly compute related
- **Message size distribution**
  - With cluster size, there is increase in both small and larger messages
  - From the total number of messages
    - The percentage of large messages increases on behalf of small messages
- **WRF shows dependency on both clustering latency and throughput**
  - Latency – synchronizations
  - Throughput (interconnect bandwidth) – compute

**WRF MPI Profiling**  
**Total Data Send per Message Size per Cluster Size**

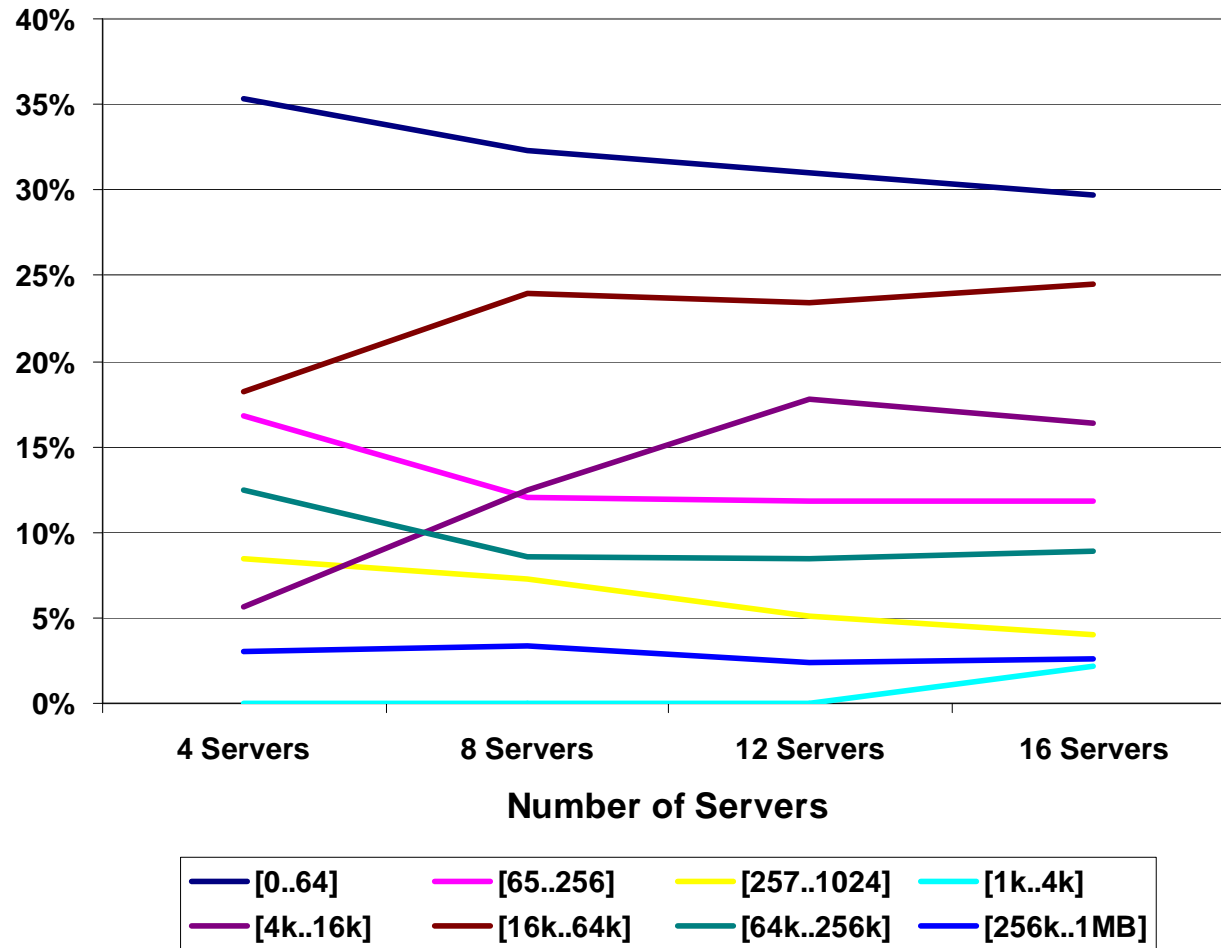


### WRF MPI Profiling Total Number of Messages per Cluster Size



# WRF Profiling – Message Distribution

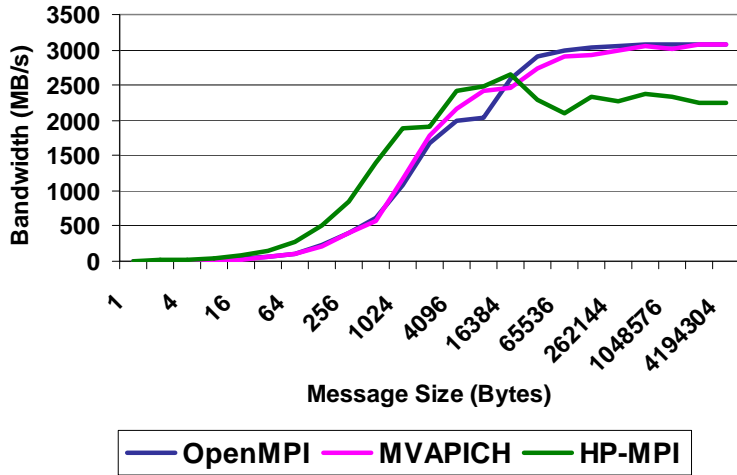
### WRF MPI Profiling Message Distribution



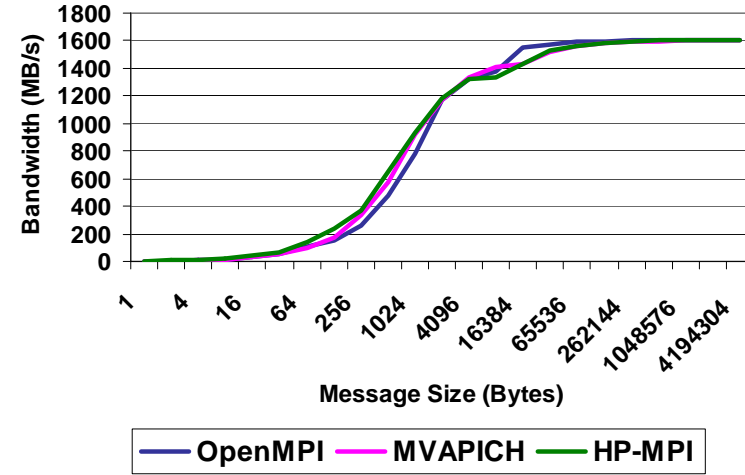
- **WRF model shows dependency on latency**
  - Mainly for <64B messages
- **WRF model shows dependency on throughput**
  - Mainly for 16KB-64KB messages
- **MPI comparison**
  - MPI libraries tested – Open MPI, MVAPICH, HP-MPI
  - All show same latency up to 128B
    - Beyond that MVAPICH and Open MPI show better latency
  - MVAPICH and Open MPI show higher bi-directional throughput
- **WRF mode results**
  - MVAPICH and Open MPI show similar performance results
  - HP-MPI shows average of 10% lower performance results
    - Due to lower bandwidth and higher latency

# MPI Low Level Performance Comparison

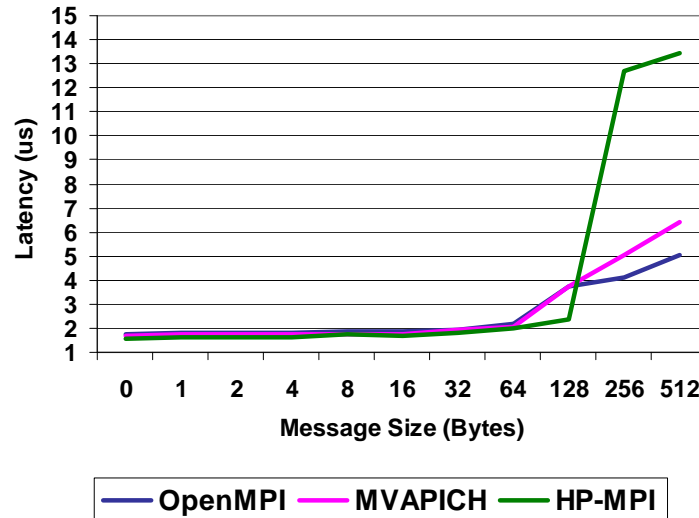
### MPI Bi-Dir Bandwidth



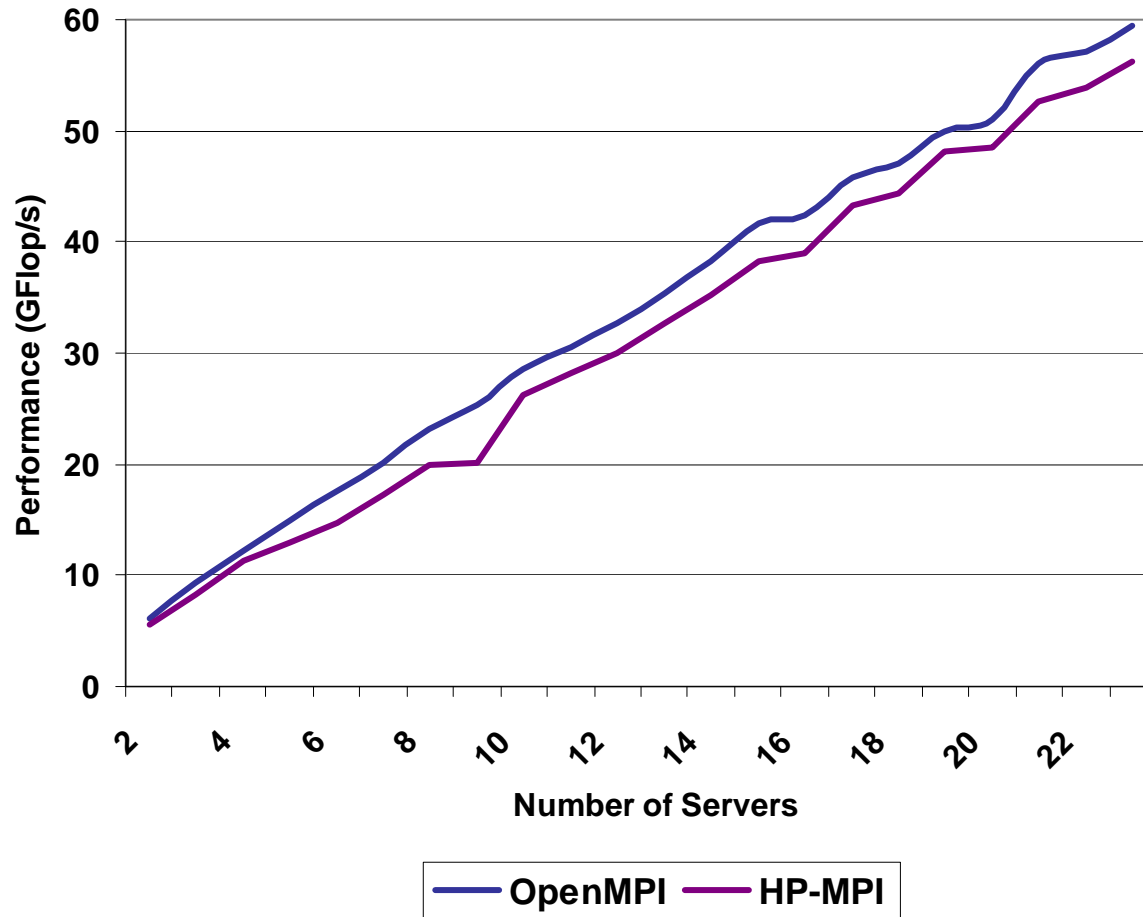
### MPI Uni-Dir Bandwidth



### MPI Latency



## WRF Benchmark Results - Conus 12Km



- **WRF is the next generation model for weather forecasting**
  - Critical tool for severe storms prediction and alerts
  - Operational since 2006, one of the most used models nowadays
- **Efficient WRF Model usage requires HPC systems**
  - Real-time, accurate and large scale weather analysis
- **WRF Model profiling analysis proves the needs for**
  - High throughput and low latency interconnect solution
  - NUMA aware application for fast access to memory
  - Expert integration and the right choice of MPI library
- **Future work**
  - Power-aware simulations, large memory pages effect
  - Optimized MPI collective operations and collectives offload

# Thank You

HPC Advisory Council  
HPC@mellanox.com



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