

Weather Research and Forecasting (WRF) Performance Benchmark and Profiling

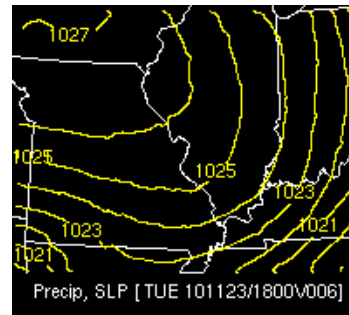
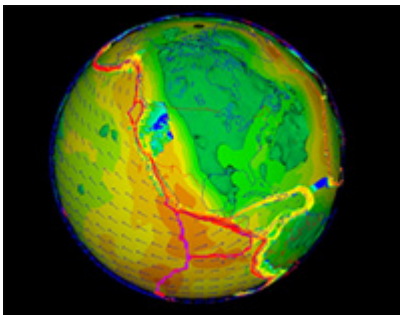
December 2010



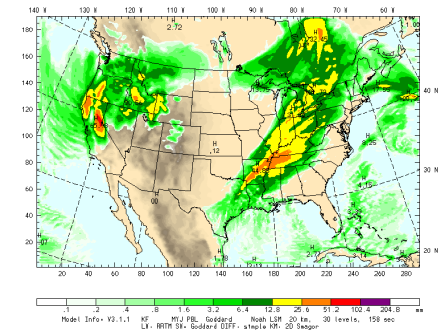
- **The following research was performed under the HPC Advisory Council activities**
 - Participating vendors: Intel, Dell, Mellanox
 - Compute resource - HPC Advisory Council Cluster Center

- **For more info please refer to**
 - <http://www.dell.com>
 - <http://www.intel.com>
 - <http://www.mellanox.com>
 - <http://www.wrf-model.org>

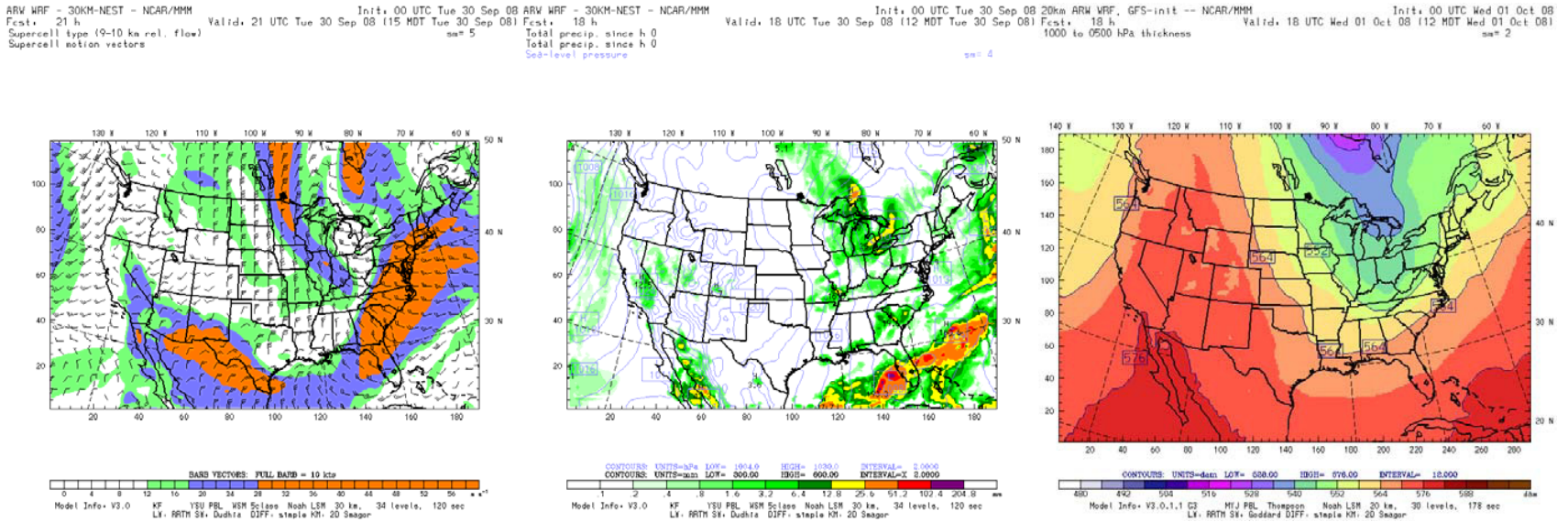
- **The Weather Research and Forecasting (WRF) Model**
 - Numerical weather prediction system
 - Designed for operational forecasting and atmospheric research
- **WRF developed by**
 - National Center for Atmospheric Research (NCAR),
 - The National Centers for Environmental Prediction (NCEP)
 - Forecast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



20km ARW WRF, GFS-init -- NCAR/BBM Init: 00 UTC Tue 23 Nov 10
Exec: 18 h Valid: 18 UTC Tue 23 Nov 10 (11 MST Tue 23 Nov 10)
Total precip. since h 0



- **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



- **The following was done to provide best practices**
 - WRF performance benchmarking
 - Interconnect performance comparisons
 - Understanding WRF communication patterns
 - Ways to increase WRF productivity
 - MPI libraries comparisons

- **The presented results will demonstrate**
 - The scalability of the compute environment to provide nearly linear application scalability
 - The capability of WRF to achieve scalable productivity
 - Considerations for power saving through balanced system configuration

- **Dell™ PowerEdge™ M610 14-node cluster**
 - Six-Core Intel X5670 @ 2.93 GHz CPUs
 - Memory: 24GB memory, DDR3 1333 MHz
 - OS: CentOS5U4, OFED 1.5.1 InfiniBand SW stack
- **Intel Cluster Ready certified cluster**
- **Mellanox ConnectX-2 InfiniBand adapters and switches**
- **MPI: Intel MPI 4.0 U1, Open MPI 1.5, Platform MPI 8.0.1**
- **Compilers: GNU Compilers 4.4.0, Intel Compilers 12.0.0**
- **Miscellaneous package: NetCDF 4.1.1**
- **Application: WRF 3.2.1**
- **Benchmark:**
 - CONUS-12km - 48-hour, 12km resolution case over the Continental US from October 24, 2001

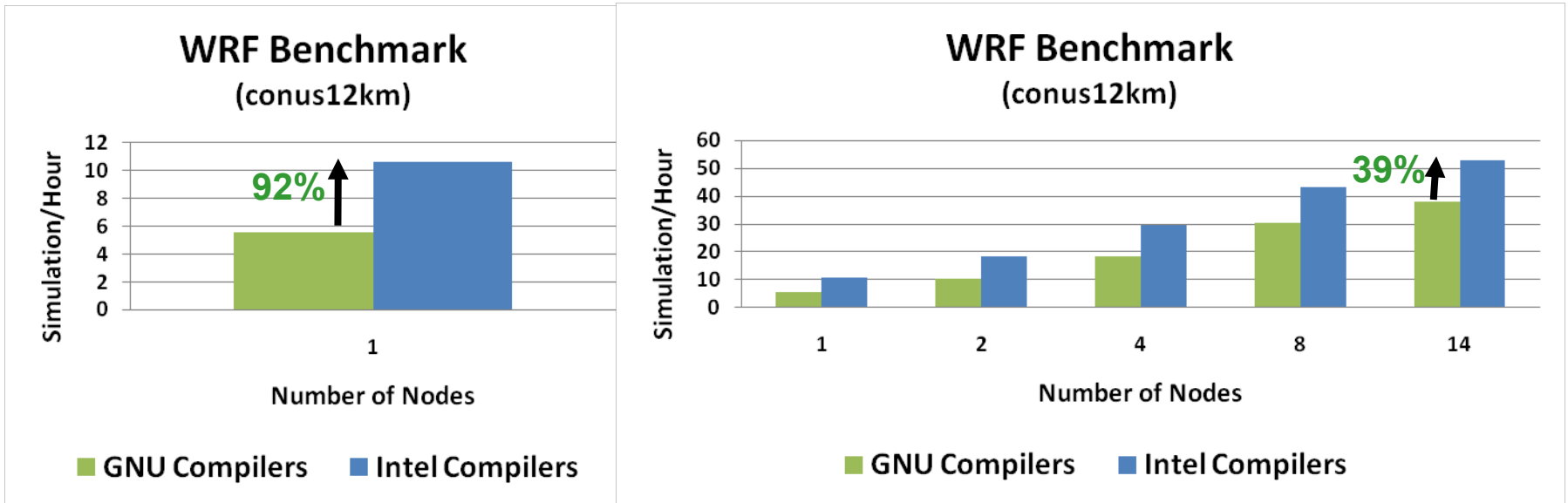
- **Intel® Cluster Ready systems make it practical to use a cluster to increase your simulation and modeling productivity**
 - Simplifies selection, deployment, and operation of a cluster
- **A single architecture platform supported by many OEMs, ISVs, cluster provisioning vendors, and interconnect providers**
 - 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**
 - 14-node cluster build with Dell PowerEdge™ M610 blades server
 - 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



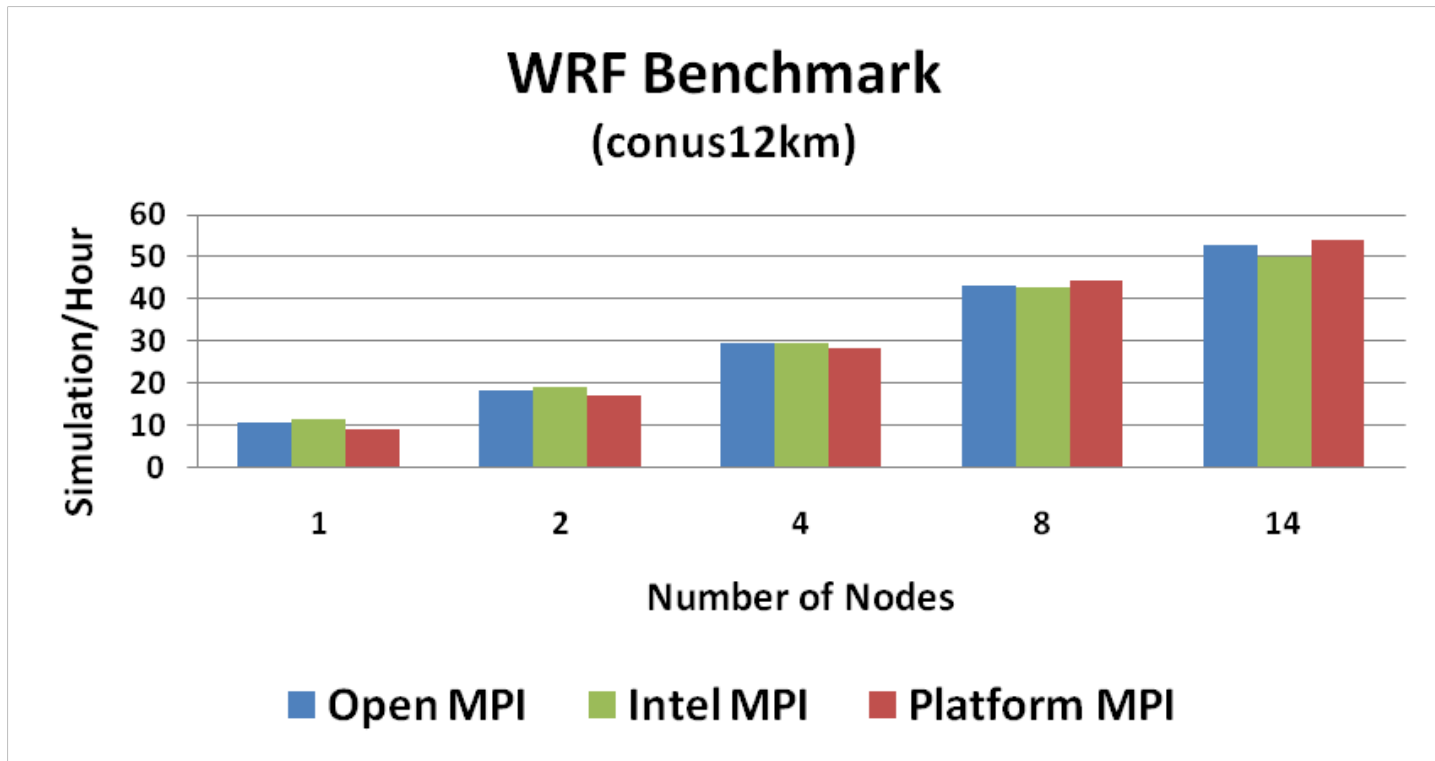
- **Intel compilers enable the better utilization on the cores**
 - Up to 92% gain on a single node versus GNU 4.4 compilers
 - Up to 39% gain on 14-node versus GNU 4.4 compilers



Open MPI 1.5
12 Cores/Node

Higher is better

- All MPI implementations performs generally the same
 - Platform MPI shows slightly better performance as the cluster scales

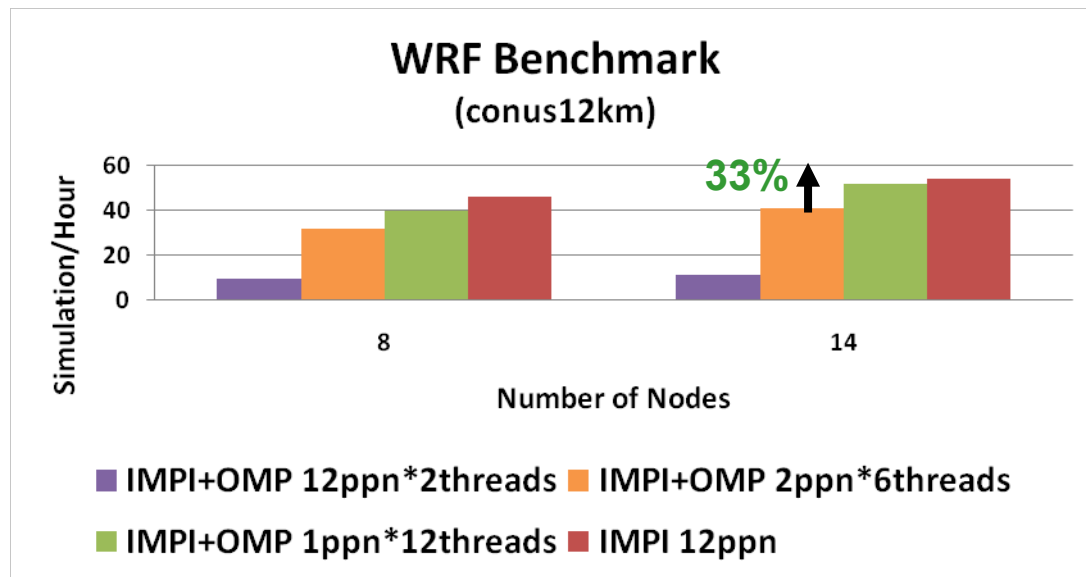


Higher is better

12 Cores/Node

WRF Performance – MPI vs OpenMP Hybrids

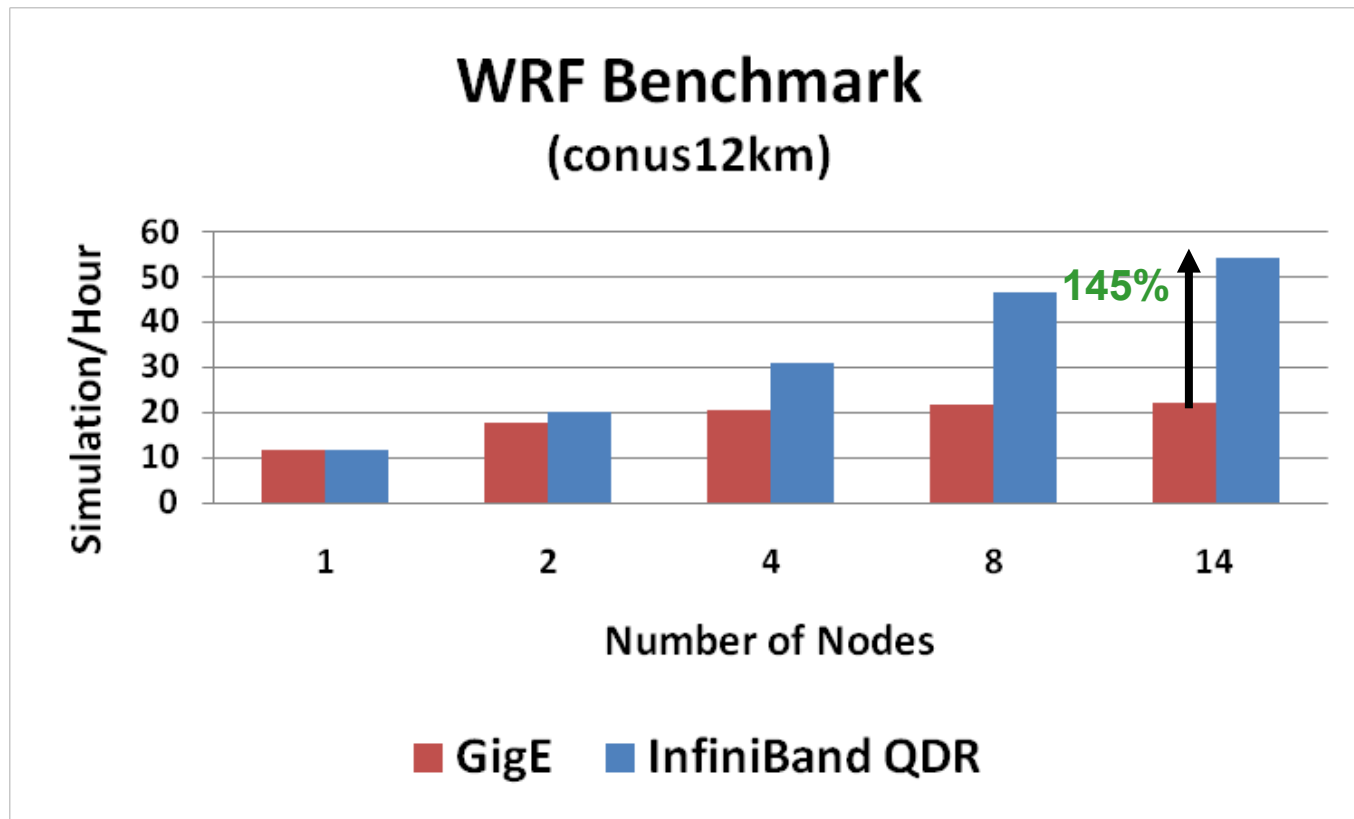
- **MPI beats OpenMP Hybrids in delivering better performance**
- **12 MPI processes per node provides better performance**
 - Up to 33% on 2 MPI processes per node with 6 OpenMP threads, and
 - Up to 6% on 1 MPI process with 12 OpenMP threads
- **Oversubscribing with OpenMP threads does not yield good performance**
 - When using OMP_NUM_THREADS=2 on 12 MPI processes per node



Higher is better

12 Cores/Node

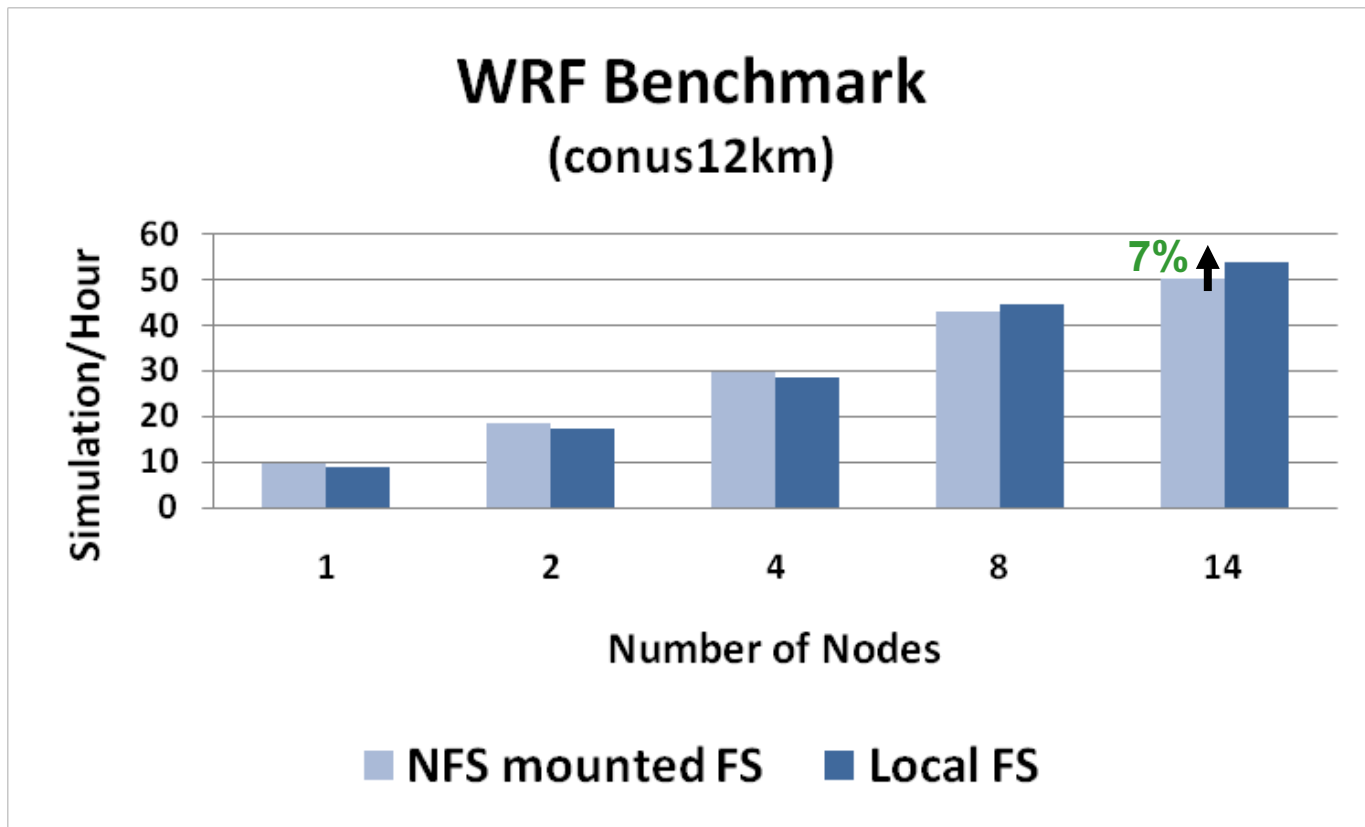
- **InfiniBand enables higher scalability**
 - Up to 145% higher performance than Ethernet at 14-node
- **Ethernet would not scale beyond 4 nodes**
 - Show virtually no gain in work done by increasing nodes



Higher is better

12 Cores/Node

- **Running dataset on Local FS over NFS mounted directory**
 - yields slightly better performance as the node number increases

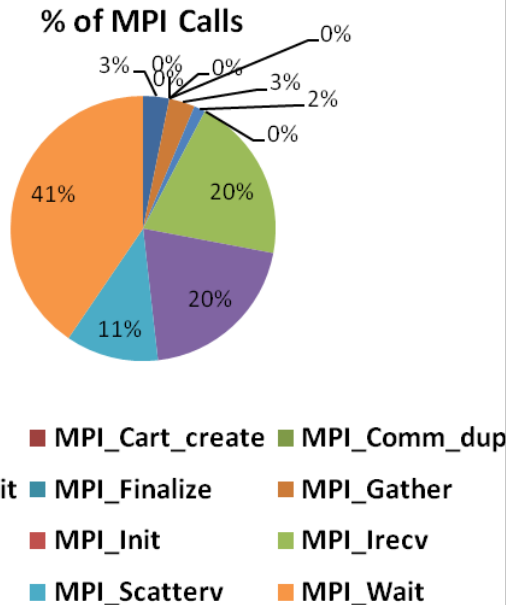


Higher is better

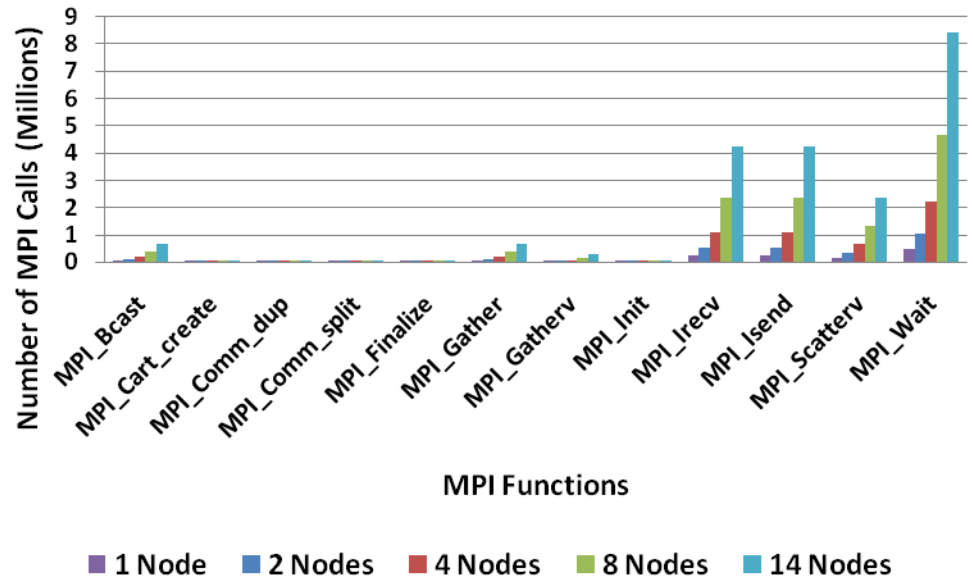
12 Cores/Node

- The most used MPI functions with this dataset are
 - MPI_Wait, MPI_Isend, MPI_Irecv and MPI_Scatterv
- **MPI_Wait accounted for 41% of all MPI calls on a 14-node job**
 - MPI_Wait is called for every non-blocking send and receive to complete

WRF Profiling
(conus12km)
% of MPI Calls



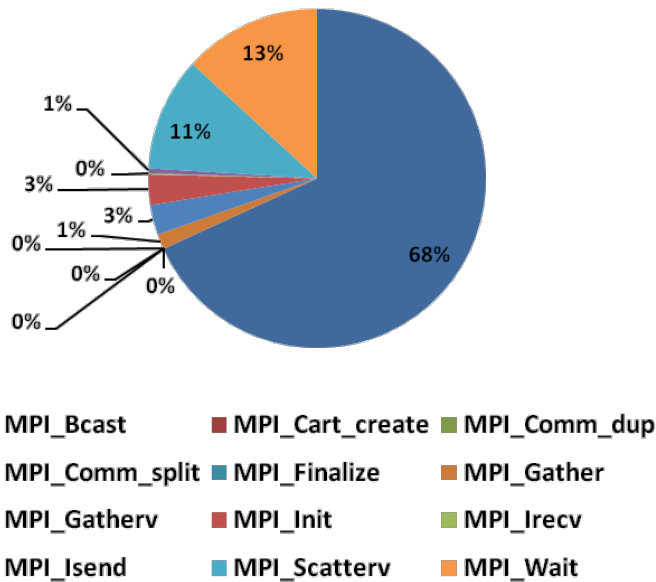
WRF Profiling
(conus12km)
Number of MPI Calls



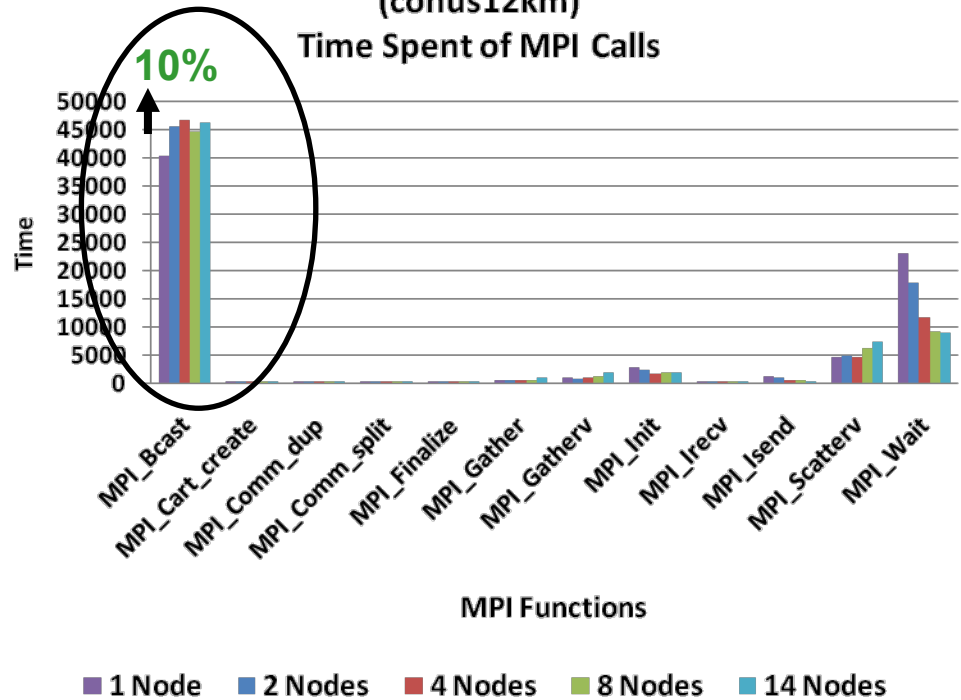
WRF Profiling – Time Spent of by MPI Calls

- **Majority of time is spent on MPI_Bcast and MPI_Wait**
 - MPI_Bcast is accounted for 68% of time spent on a 14-node job
- **The 10% time difference in MPI_Bcast between 1 and 2+ nodes**
 - Reflects the difference for intra node broadcast to inter node broadcast
 - InfiniBand adds only 10% on broadcast time compared to intra node latency

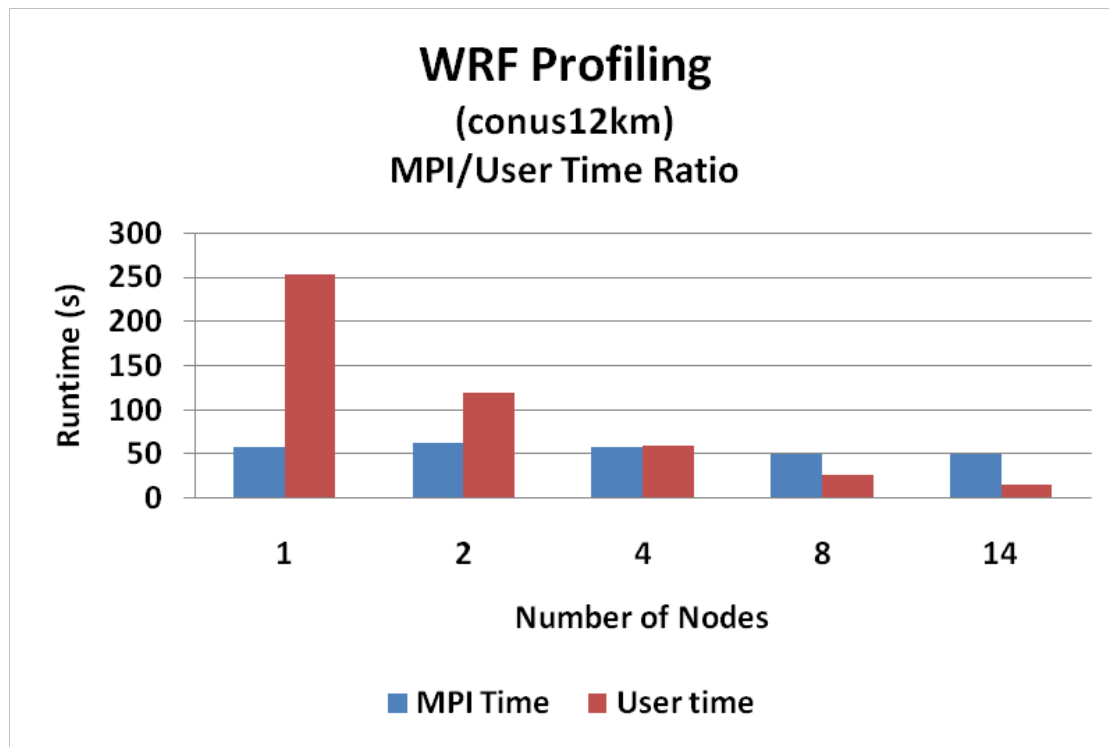
WRF Profiling
(conus12km)
% Time Spent of MPI Calls



WRF Profiling
(conus12km)
Time Spent of MPI Calls



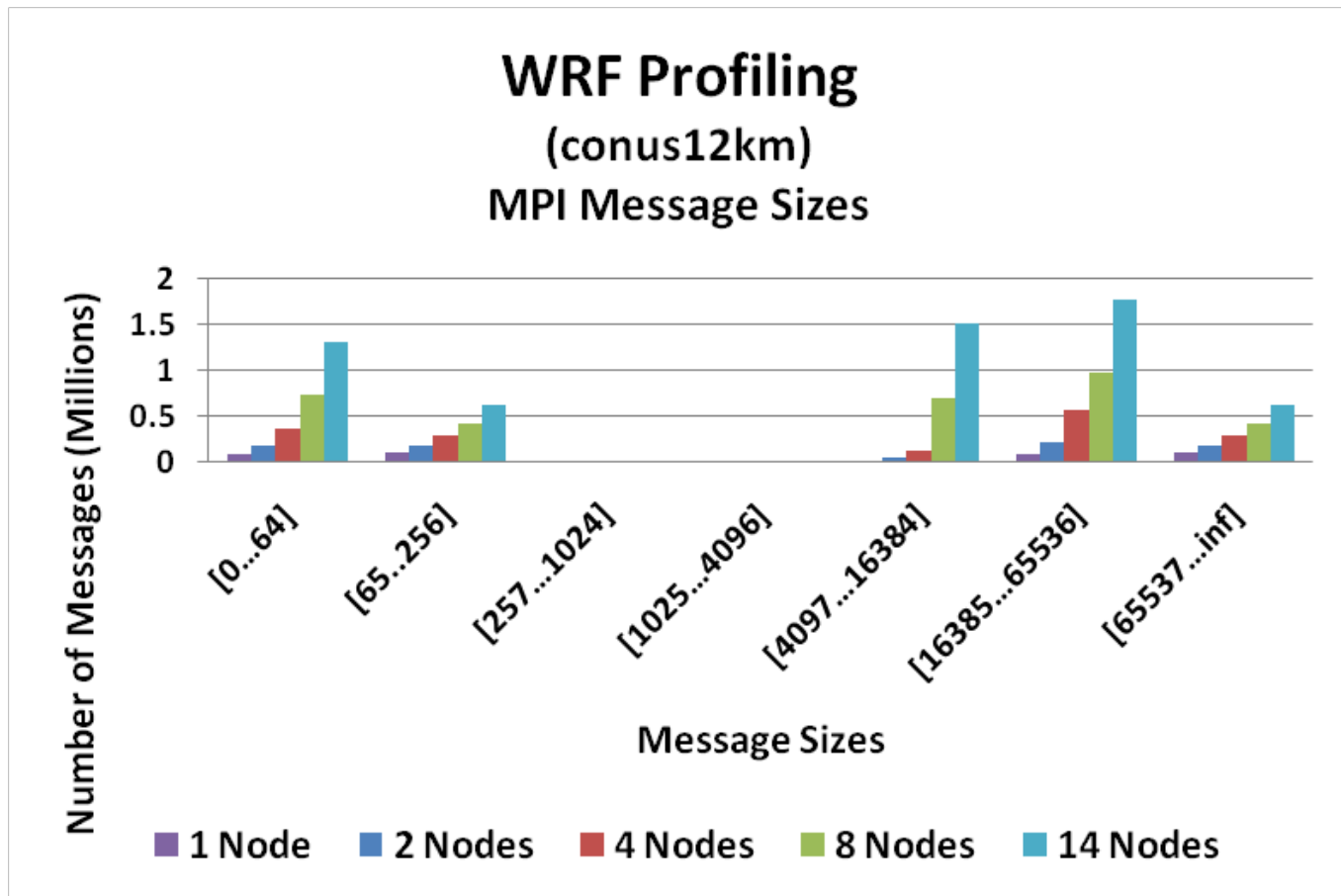
- **The MPI time decreases as the cluster scales up**
 - Reflects the calculation is able to be spread out across the cluster
- **WRF demonstrates the ability to scale as the node count increases**
 - Time spent in user time calculation is able to be reduced



InfiniBand QDR

12 Cores/Node

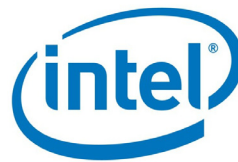
- Majority of the MPI message sizes are in the range from 4K to 64K
- Messages increase proportionally with the cluster size



- **MPI provides better performance over OpenMP Hybrids**
- **Oversubscribing with OpenMP threads does not yield good performance**
 - When using OMP_NUM_THREADS=2 on 12 MPI processes per node
- **InfiniBand enables higher performance/scalability**
 - Up to 145% higher performance than Ethernet at 14-node
- **Intel compilers enable the better processor utilization**
 - Up to 92% gain on a single node versus GNU 4.4 compilers
- **All MPI implementations performs generally the same**
 - Platform MPI shows slightly better performance as the cluster scales
- **The most used MPI functions with this dataset are**
 - MPI_Wait, MPI_Isend, MPI_Irecv and MPI_Scatterv
- **Majority of the MPI message sizes are in the range from 4K to 64K**
- **Majority of time is spent on MPI_Bcast and MPI_Wait**
- **Due to its low latency, InfiniBand adds only 10% on inter node communications latency compared to intra node communication**

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