HPC ADVISORY COUNCIL

Weather Research and Forecasting (WRF) Performance Benchmark and Profiling

July 2012

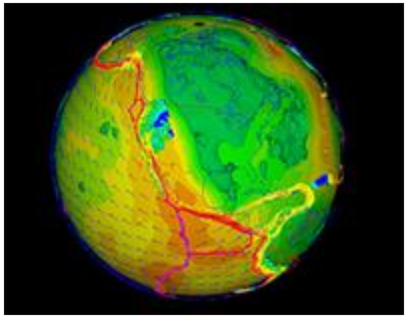
DELL intel Mellanox TECHNOLOGIES WRF THE WEATHER RESEARCH & FORECASTING MODEL

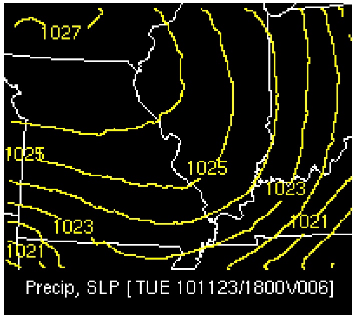
Note

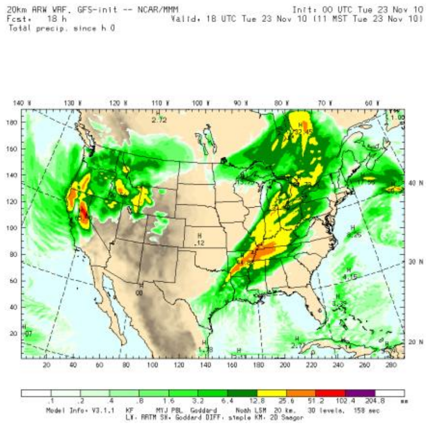
- **The following research was performed under the HPC Advisory Council activities**
  - Participating vendors: Intel, Dell, Mellanox
  - Compute resource - HPC Advisory Council Cluster Center
- **The following was done to provide best practices**
  - WRF performance overview
  - Understanding WRF communication patterns
  - Ways to increase WRF productivity
  - MPI libraries comparisons
- **For more info please refer to**
  - http://www.dell.com
  - http://www.intel.com
  - http://www.mellanox.com
  - http://wrf-model.org

Weather Research and Forecasting (WRF)

- **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)



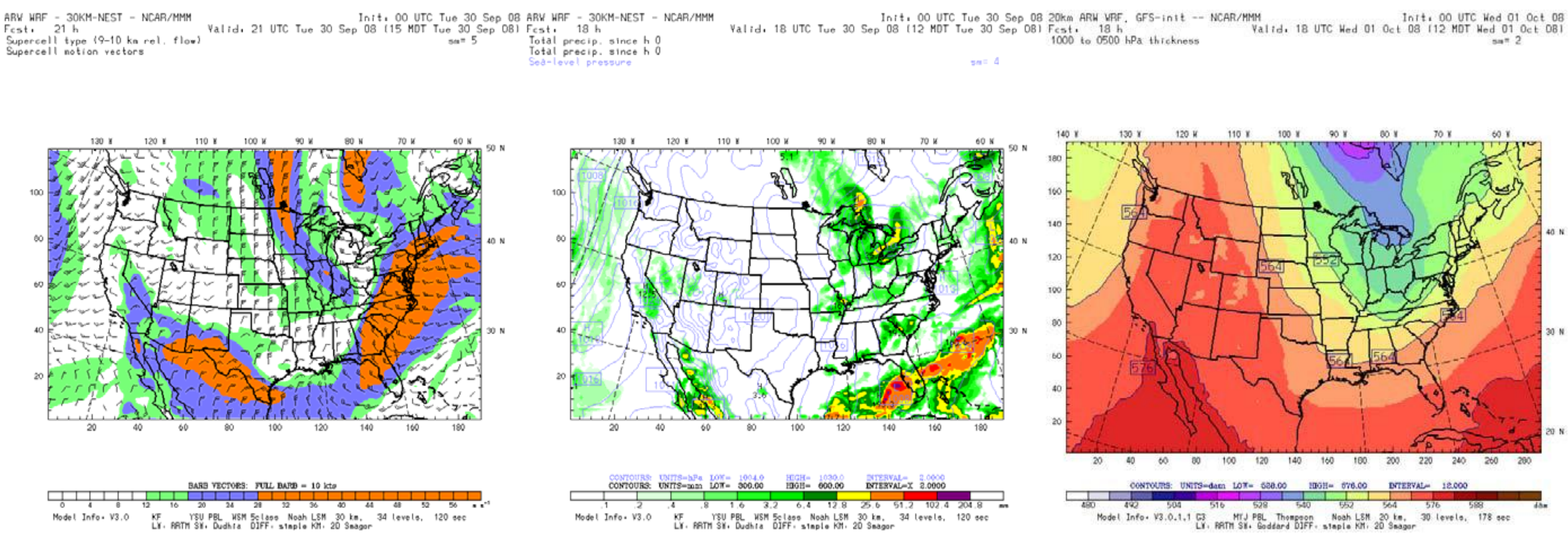




WRF Usage

HPC ADVISORY COUNCIL

- **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



Objectives

- **The presented research was done to provide best practices**
  – WRF performance benchmarking
    - MPI Library performance comparison
    - Interconnect performance comparison
    - CPUs comparison
    - Compilers comparison
- **The presented results will demonstrate**
  – The scalability of the compute environment/application
  – Considerations for higher productivity and efficiency

Test Cluster Configuration

- **Dell™ PowerEdge™ R720xd 16-node (256-core) "Jupiter" cluster**
  - Dual-Socket Eight-Core Intel E5-2680 @ 2.70 GHz CPUs (Static max Perf in BIOS)
  - Memory: 64GB memory, DDR3 1600 MHz
  - OS: RHEL 6.2, OFED 1.5.3 InfiniBand SW stack
  - Hard Drives: 24x 250GB 7.2 RPM SATA 2.5" on RAID 0
- **Intel Cluster Ready certified cluster**
- **Mellanox ConnectX-3 FDR InfiniBand VPI adapters**
- **Mellanox SwitchX SX6036 InfiniBand switch**
- **Compilers: GNU 4.6.3, Intel Composer XE 2011. NetCDF 4.1.3**
- **MPI: Intel MPI 4 U3, Open MPI 1.6 (KNEM 0.9.8), Platform MPI 8.2**
- **Application and benchmarks: WRF 3.4, CONUS-12km - 48-hour, 12km resolution case over the Continental US from October 24, 2001**

About Intel® Cluster Ready

- **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

PowerEdge R720xd

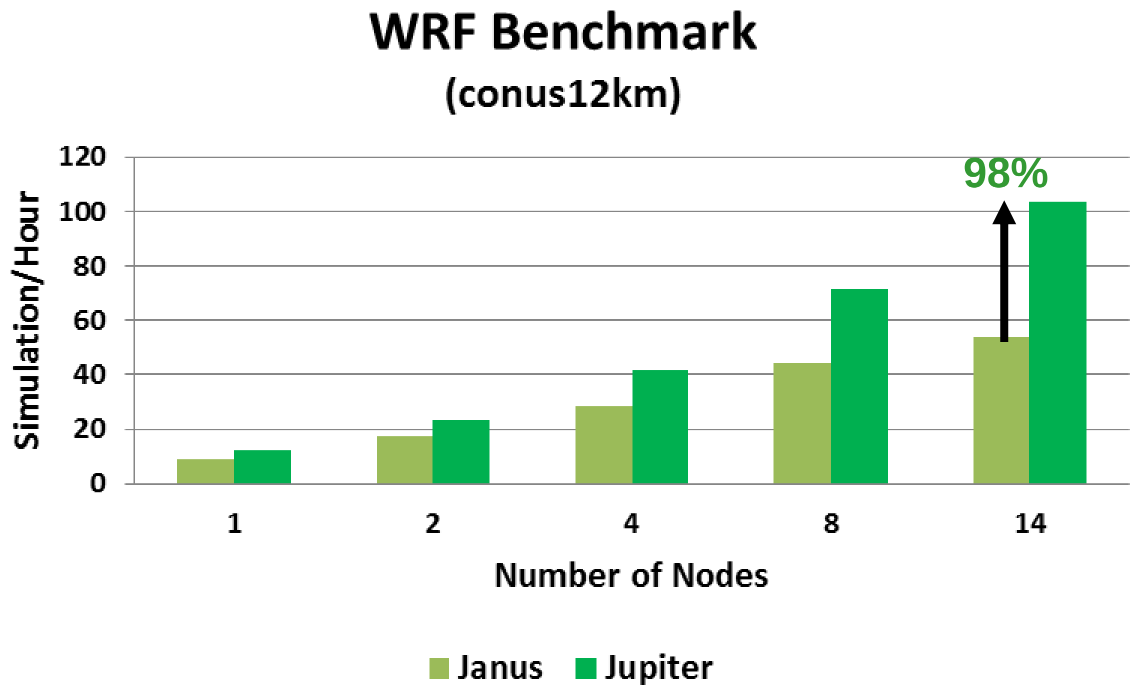
Massive flexibility for data intensive operations

- **Performance and efficiency**
  - Intelligent hardware-driven systems management with extensive power management features
  - Innovative tools including automation for parts replacement and lifecycle manageability
  - Broad choice of networking technologies from GigE to IB
  - Built in redundancy with hot plug and swappable PSU, HDDs and fans
- **Benefits**
  - Designed for performance workloads
    - from big data analytics, distributed storage or distributed computing where local storage is key to classic HPC and large scale hosting environments
    - High performance scale-out compute and low cost dense storage in one package
- **Hardware Capabilities**
  - Flexible compute platform with dense storage capacity
    - 2S/2U server, 6 PCIe slots
  - Large memory footprint (Up to 768GB / 24 DIMMs)
  - High I/O performance and optional storage configurations
    - HDD options: 12 x 3.5" - or - 24 x 2.5 + 2x 2.5 HDDs in rear of server
    - Up to 26 HDDs with 2 hot plug drives in rear of server for boot or scratch

WRF Performance – Processor Generations

HPC ADVISORY COUNCIL

- **Intel E5-2600 Series (Sandy Bridge) outperforms prior generations**
  - Up to 98% higher performance than Intel Xeon X5670 (Westmere) at 14-node
- **System components used:**
  - Jupiter: 2-socket Intel E5-2680 @ 2.7GHz, 1600MHz DIMMs, FDR IB, 24 disks
  - Janus: 2-socket Intel x5670 @ 2.93GHz, 1333MHz DIMMs, QDR IB, 1 disk

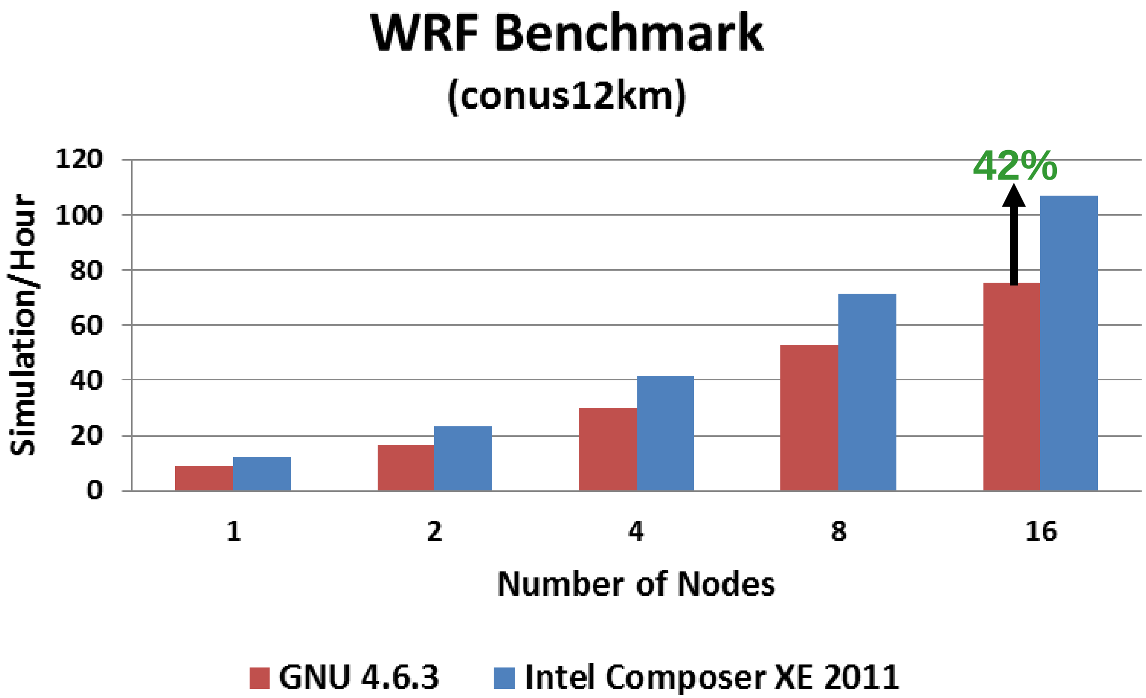


Higher is better

InfiniBand FDR

WRF Performance – Compilers

- **Using Intel Composer XE 2011 compiler shows the best performance**
  - Up to 42% better performance than compiling WRF using GNU 4.6.3 compilers
- **Default compile options are being tested**
  - The standard compiler options in the “configure.wrf” file

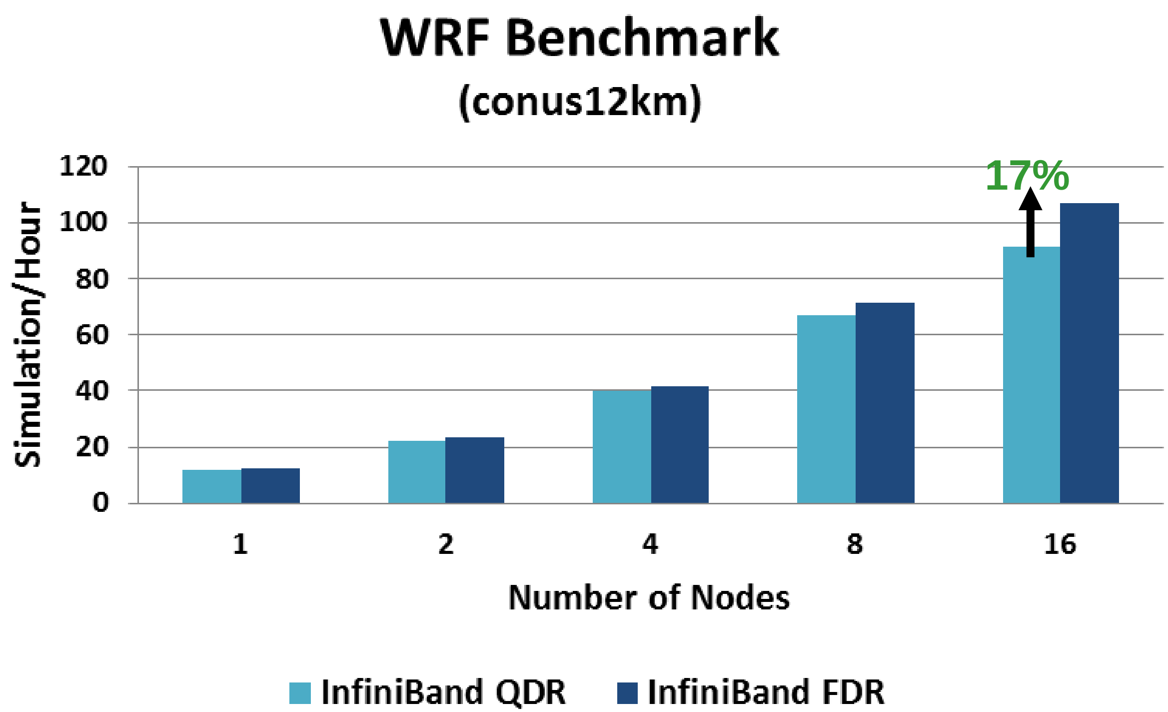


Higher is better

InfiniBand FDR

WRF Performance – Interconnects

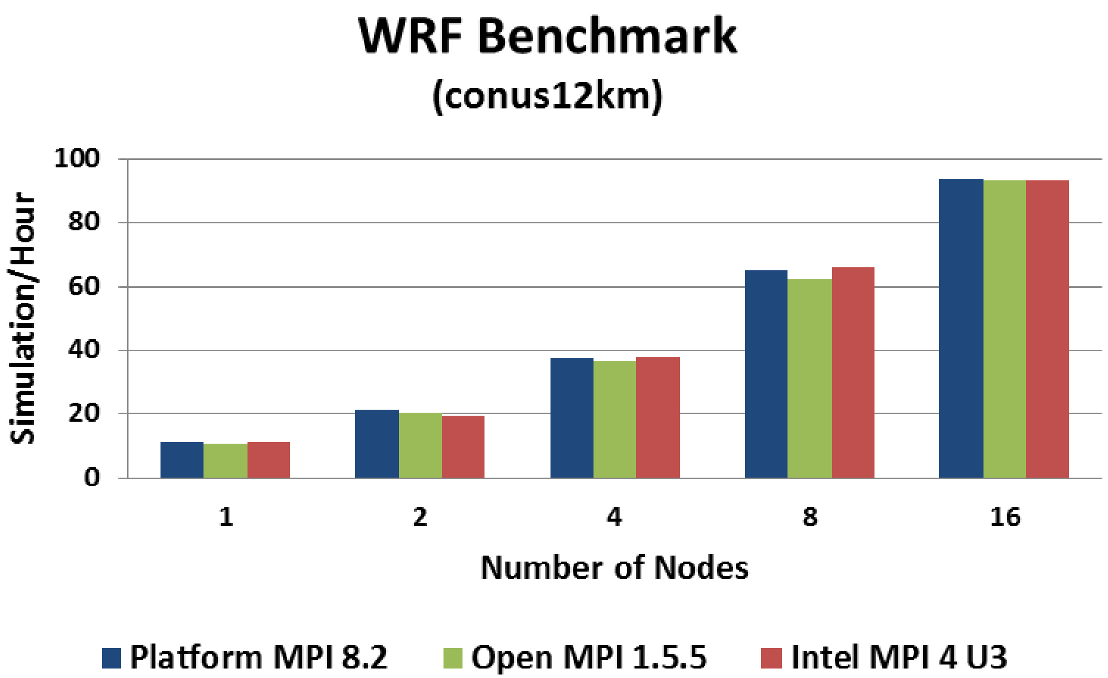
- **InfiniBand FDR enables the highest cluster productivity**
  - Increasing the performance by up to 17% over InfiniBand QDR at 16-node



Higher is better

WRF Performance – MPI Implementations

- **All MPI performs similarly in performance**
  - All MPI implementations tested (Intel, Platform, Open MPI) show good performance
  - Reflects each MPI implementation handles efficiently for the MPI transfers



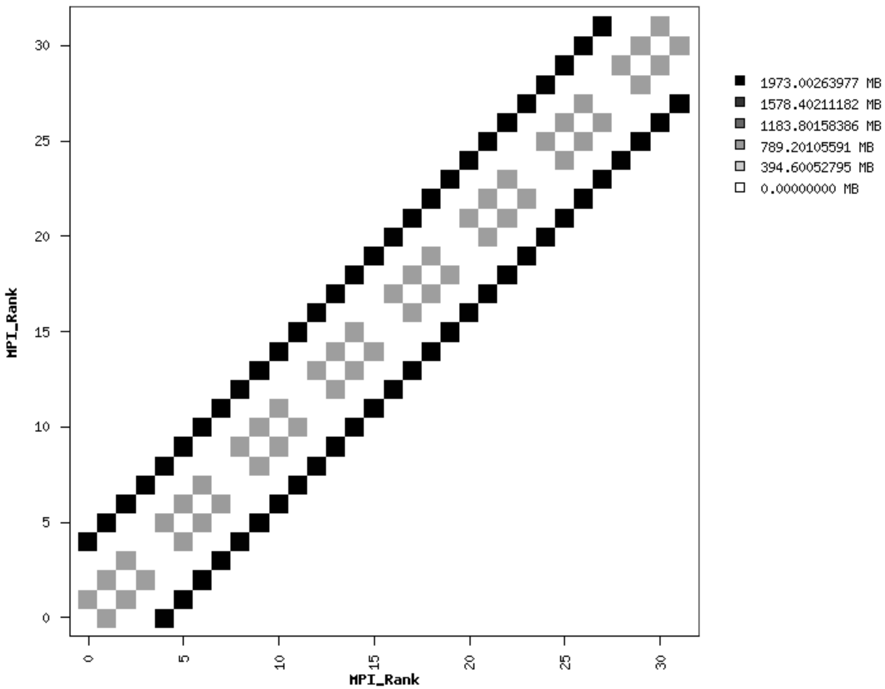
Higher is better

GNU Compilers

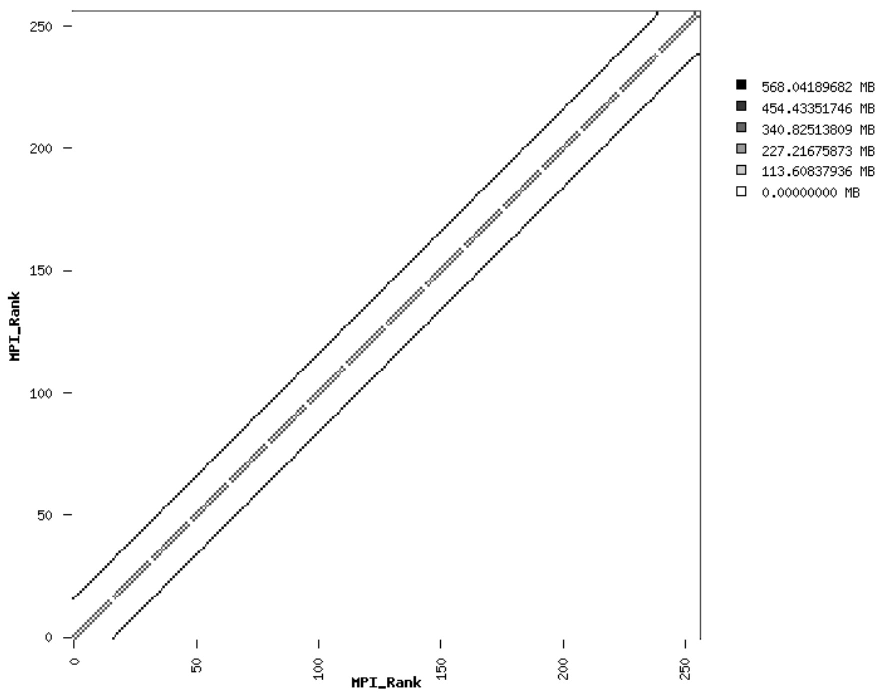
WRF Profiling – Point-to-point dataflow

- **Communication seems to be limited to MPI ranks that is closer to self**
  - Heavy communications between self and 4 ranks above and below
- **Communication pattern does not change as the cluster scales**
  - However, the amount of data being transferred is reduced as the node scales

2 nodes

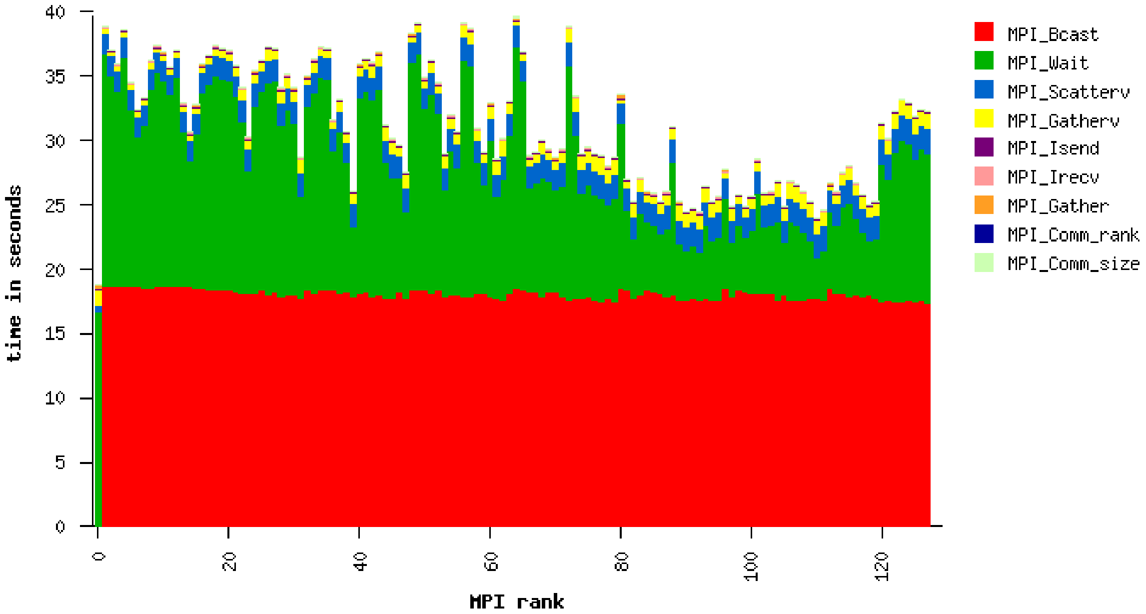


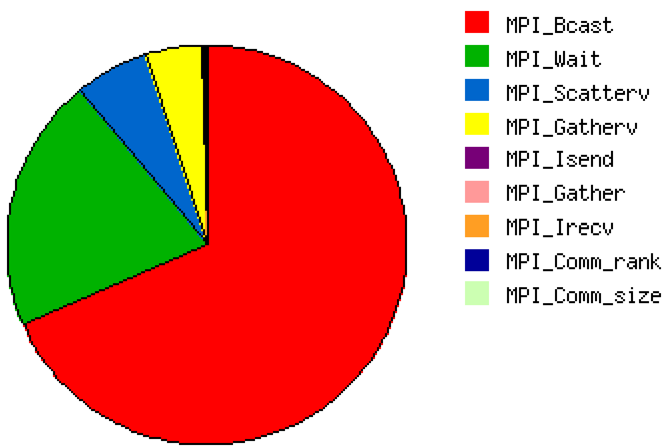
16 nodes



WRF Profiling – Time Spent by MPI Calls

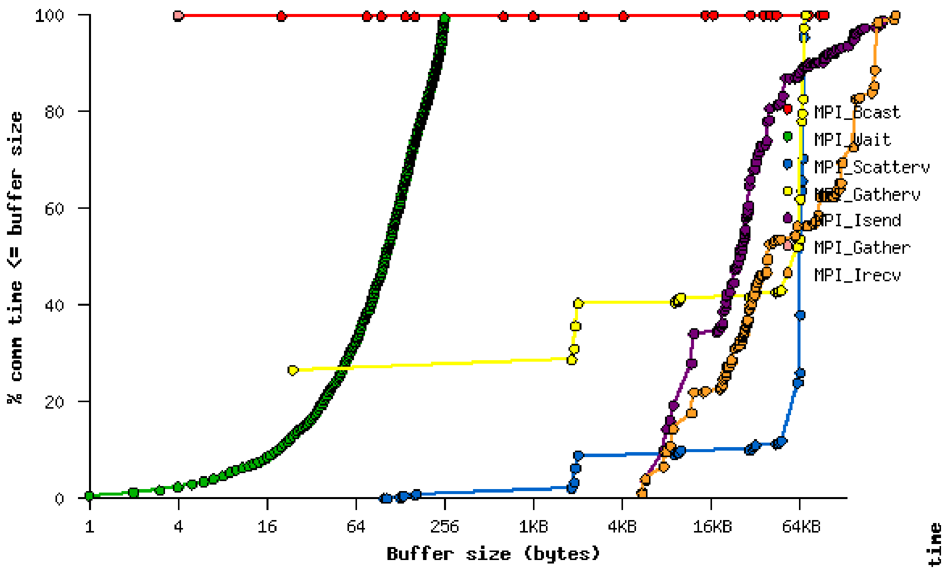
- **Majority of the MPI time is spent on MPI_Bcast**
  – For waiting for pending non-blocking sends and receives to complete
- **Some differences can be seen in MPI time consumption by each MPI rank**
  – MPI_Wait time differences
  – Rank 0 does not participate on MPI_Bcast

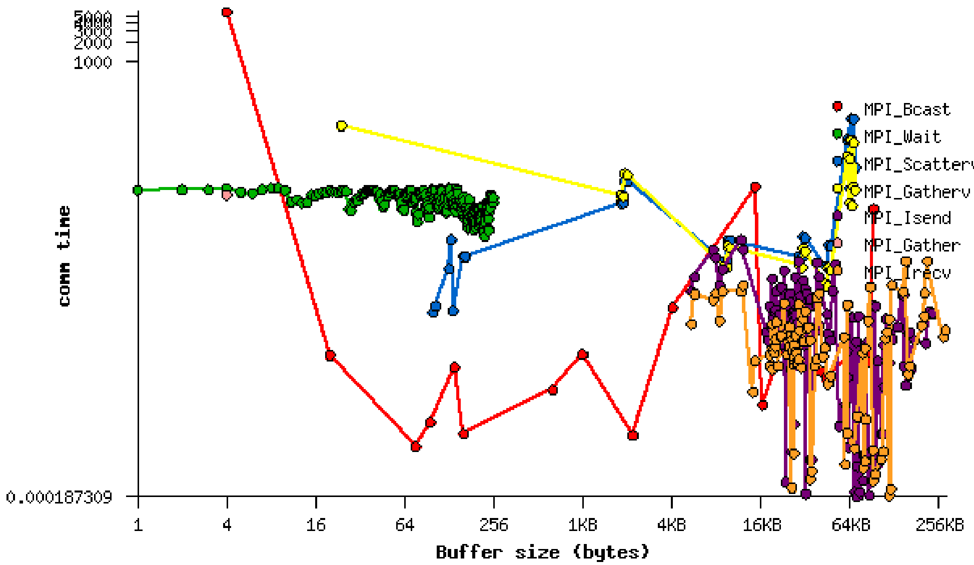




WRF Profiling – MPI Message Sizes

- **Majority of data transfer messages are medium sizes, except for:**
  - MPI_Bcast has a large concentration in small sizes (e.g. 4 byte size)
  - MPI_Wait: Large concentration of 1 to less than 256 bytes





WRF – Summary

- **Performance**
  – Intel Xeon E5-2680 on the “Jupiter” cluster and InfiniBand FDR enable WRF to scale
  – “Jupiter”, the E5-2680 cluster performs up to 98% over “Janus” the X5670 cluster
- **Network**
  – InfiniBand FDR allows WRF to run at the highest network throughput at 56Gbps
  – InfiniBand FDR provides up to 19% of performance gain over QDR rate for WRF
  – All MPI implementations tested (Intel, Platform, Open MPI) show good performance
- **Compilers**
  – Intel Composer compiler provides gains of 42% at 16-node over GNU 4.6 compilers
- **Profiling**
  – Heavy usage in midrange message sizes for MPI communications
  – Majority of MPI time is spent on MPI_Wait for pending non-blocking sends and receives

Thank You

HPC Advisory Council