HPC ADVISORY COUNCIL


# GROMACS
# Performance Benchmark and Profiling

August 2012

DELL

GROMACS FAST. FLEXIBLE. FREE.

intel

Mellanox TECHNOLOGIES

# Note

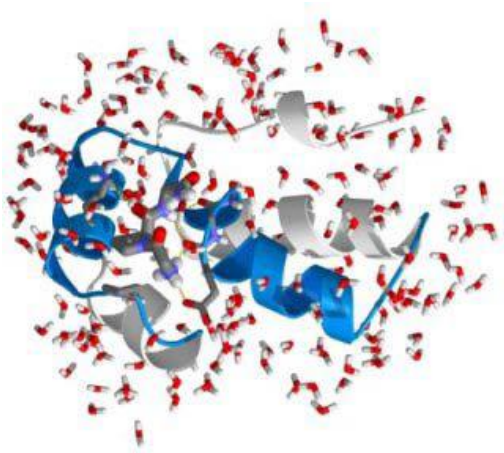
- **The following research was performed under the HPC Advisory Council activities**
  - Participating vendors: Intel, Dell, Mellanox
  - Compute resource - HPC Advisory Council Cluster Center

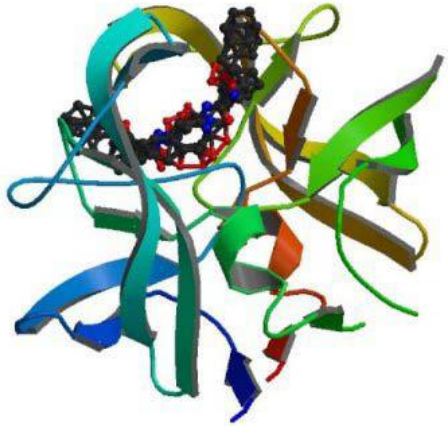
- **The following was done to provide best practices**
  - GROMACS performance overview
  - Understanding GROMACS communication patterns
  - Ways to increase GROMACS productivity
  - MPI libraries comparisons

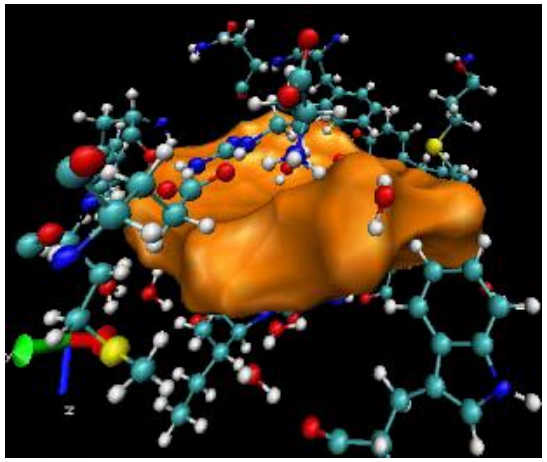
- **For more info please refer to**
  - http://www.dell.com
  - http://www.intel.com
  - http://www.mellanox.com
  - http://www.gromacs.org

# GROMACS

- **GROMACS (GROningen MAchine for Chemical Simulation)**
  - A molecular dynamics simulation package
  - Primarily designed for biochemical molecules like proteins, lipids and nucleic acids
    - A lot of algorithmic optimizations have been introduced in the code
    - Extremely fast at calculating the nonbonded interactions
  - Ongoing development to extend GROMACS with interfaces both to Quantum Chemistry and Bioinformatics/databases
  - An open source software released under the GPL







# 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**
- **SwitchX SX6036 InfiniBand switch**
- **MPI: Intel MPI 4 Update 3, Open MPI 1.5.5 with KNEM 0.9.8, Platform MPI 8.2**
- **Compiler and Libraries: Intel Composer XE 2011 for Linux, Intel MKL 10.3 Update 5**
- **Application: GROMACS 4.5.5**
- **Benchmark datasets:**
  - DPPC in Water (d.dppc, 121856 atoms, 5000 steps)

# 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

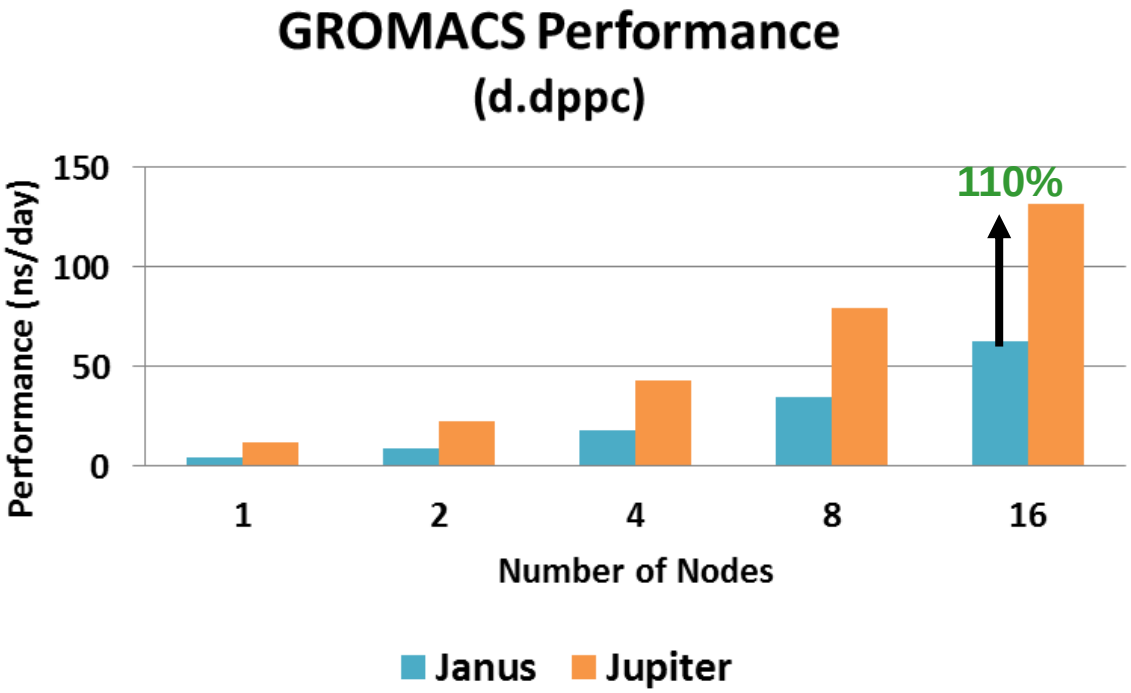
# PowerEdge R720xd

## 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

# GROMACS Performance – Processors

- **Intel E5-2600 Series (Sandy Bridge) outperforms prior generations**
  - Up to 110% higher performance than Intel Xeon X5670 (Westmere) at 16-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

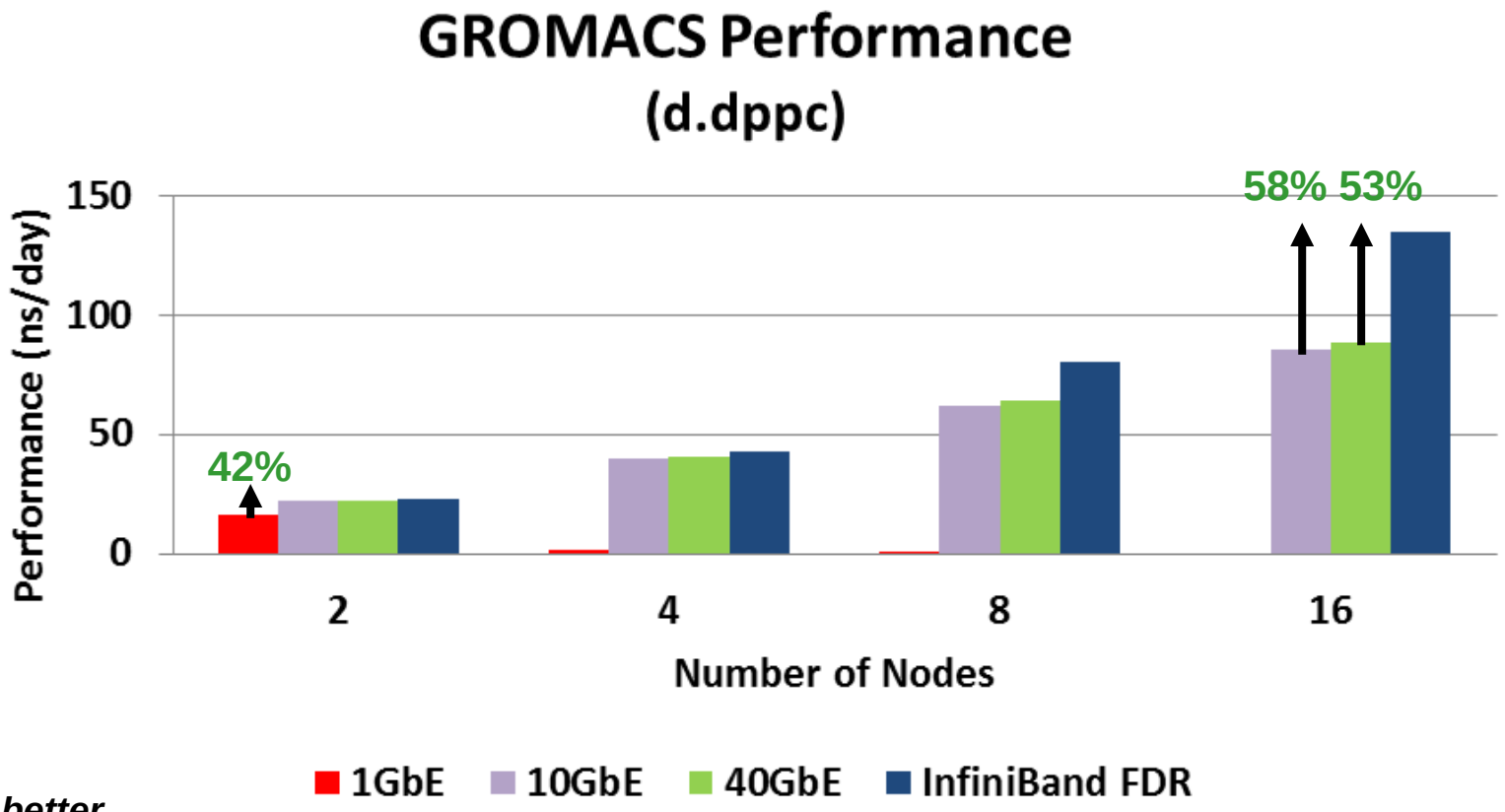



*Higher is better*

*InfiniBand FDR*

# GROMACS Performance – Interconnects

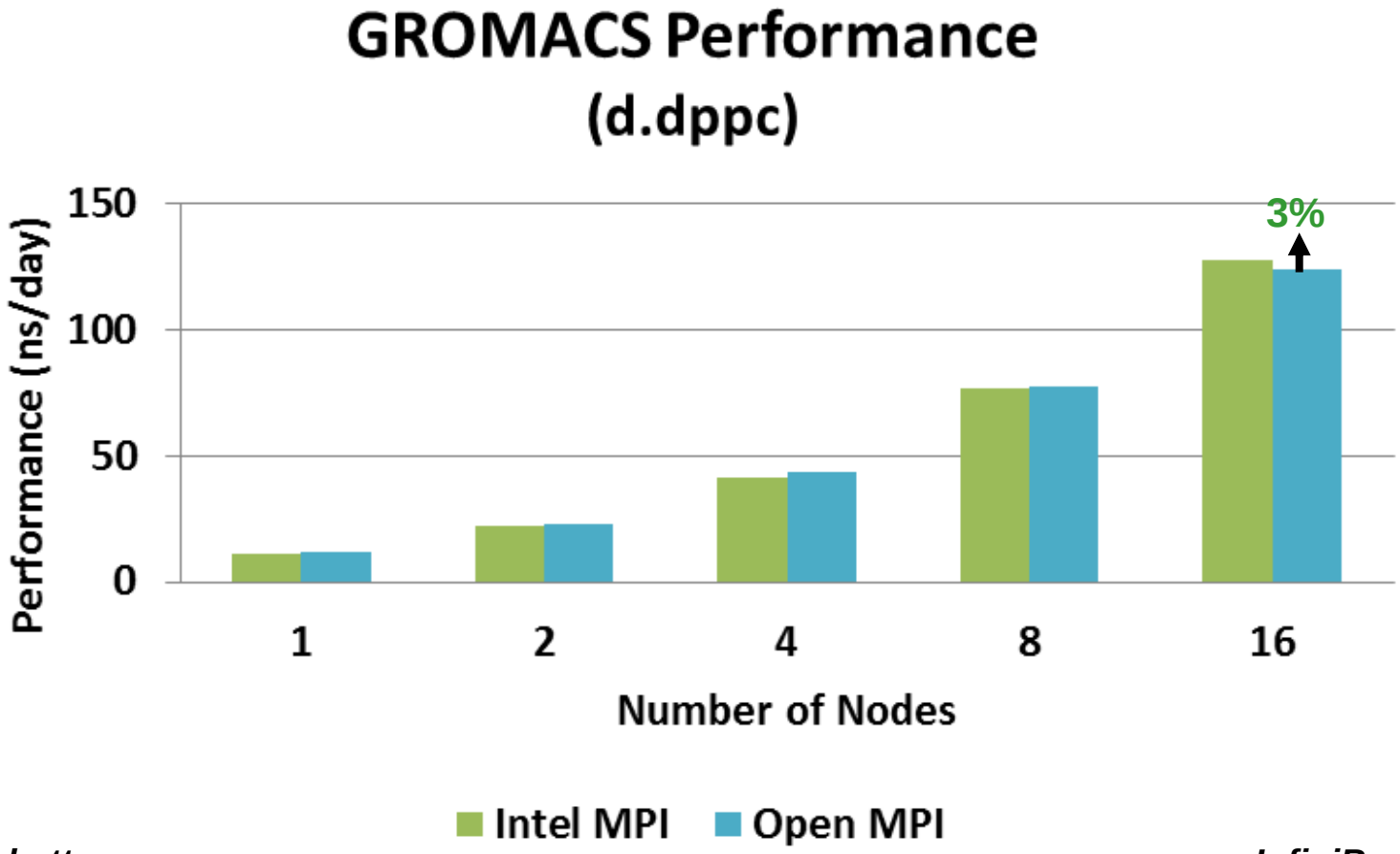
- **InfiniBand FDR enables higher cluster productivity**
  - Outperforms 1GbE by up to 42% at 2 nodes; 1GbE scalability stalls after 2 nodes
  - Outperforms 10GbE by 58% at 16 nodes
  - Outperforms 40GbE by 53% at 16 nodes




*Higher is better*

# GROMACS Performance – MPI Implementations

- **Intel MPI performs slightly better at scale than Open MPI**
  - Up to 3% better than Open MPI
  - The default DAPL provider is used for Intel MPI
  - Processor binding is enabled for Open MPI

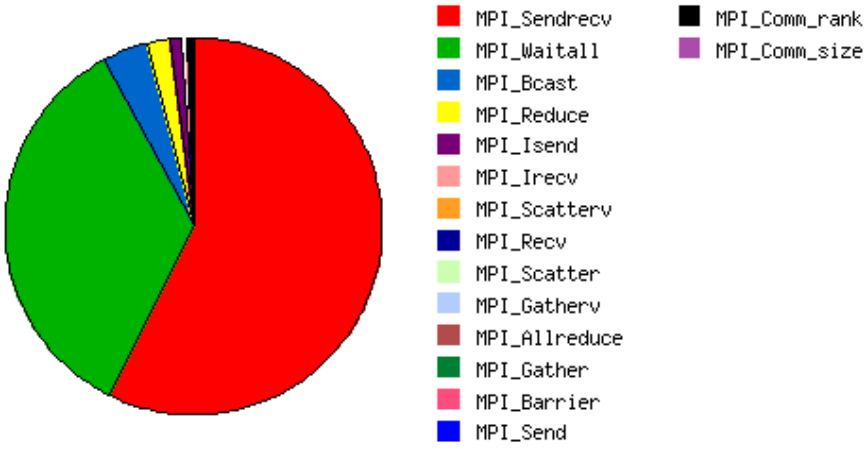



*Higher is better*

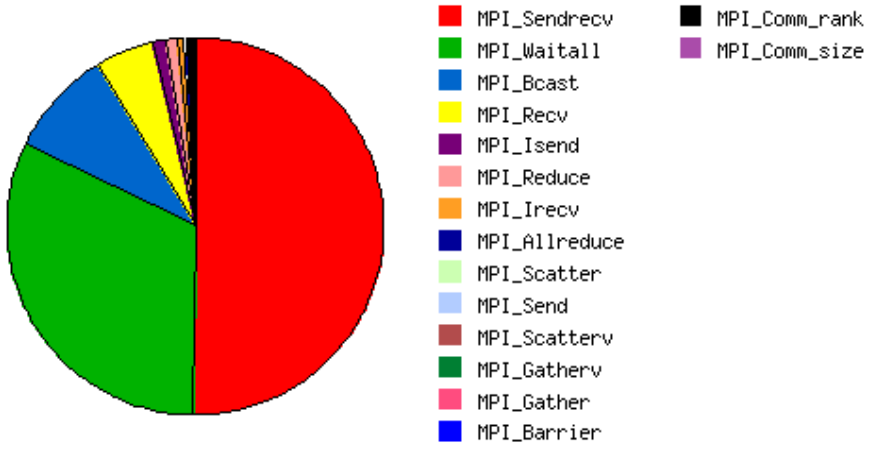
*InfiniBand FDR*

# GROMACS Profiling – MPI Functions

- **Mostly used MPI functions are MPI_Sendrecv and MPI_Waitall**
  - MPI_Sendrecv (50%) and MPI_Waitall (32%) dominate the % of MPI Time at 256 cores

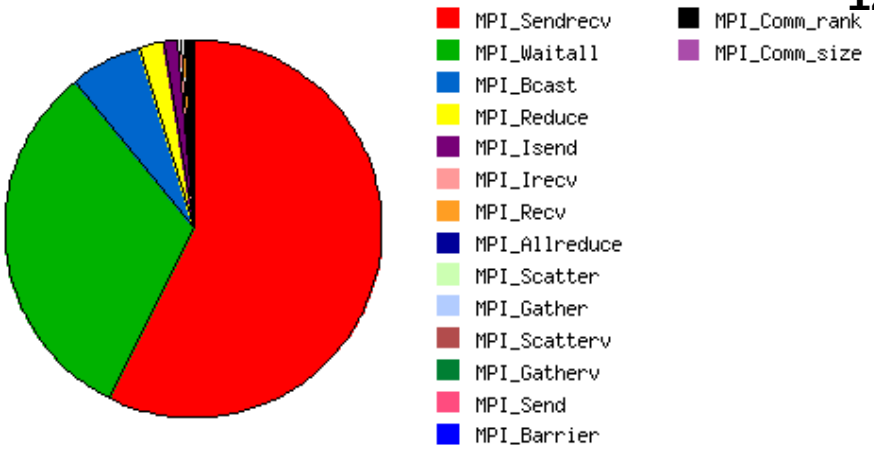



64 Cores




128 Cores

256 Cores

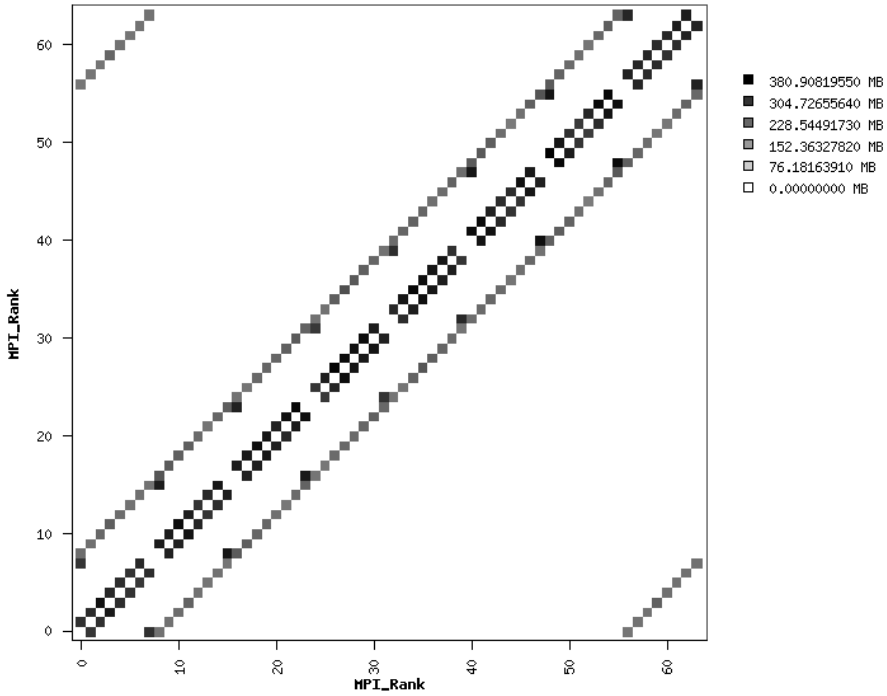



***InfiniBand FDR***

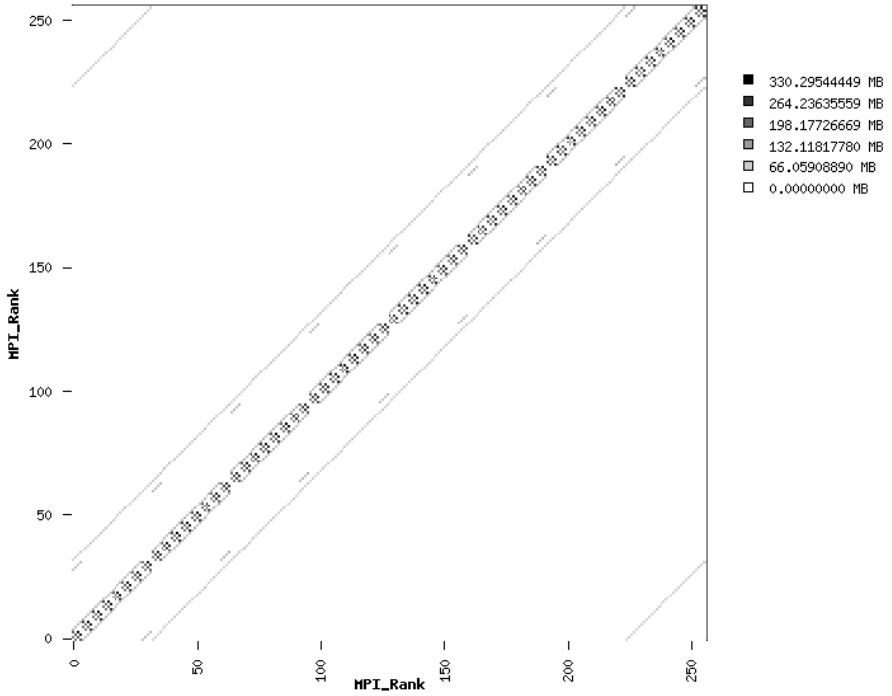
# GROMACS Profiling – Distribution of MPI Calls

- **Majority of communications takes place with the close neighboring ranks**
  - There are also some communications take place between far ranks
  - Data transfer drops from 380MB per rank (4 nodes) to 330MB per rank (16 nodes)

**4 Nodes**



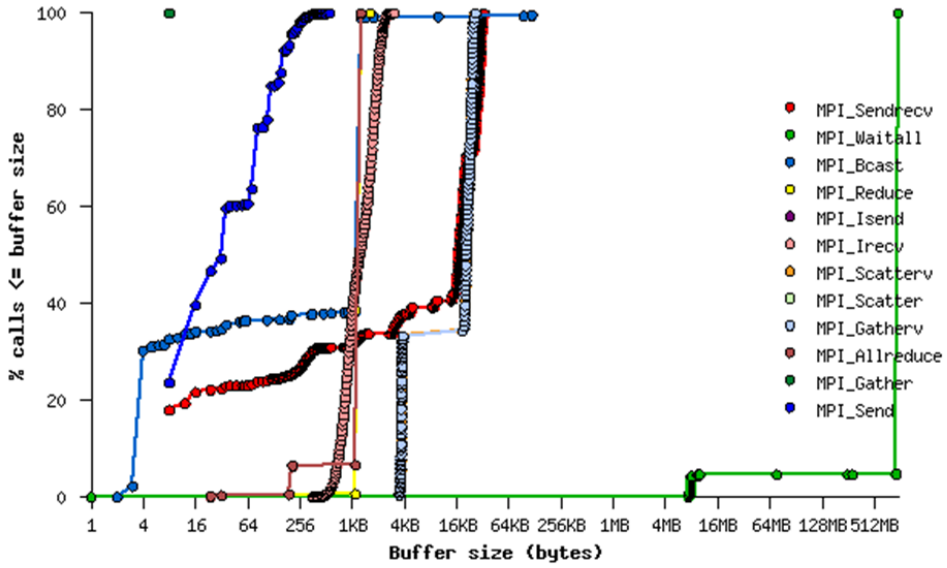

**16 Nodes**

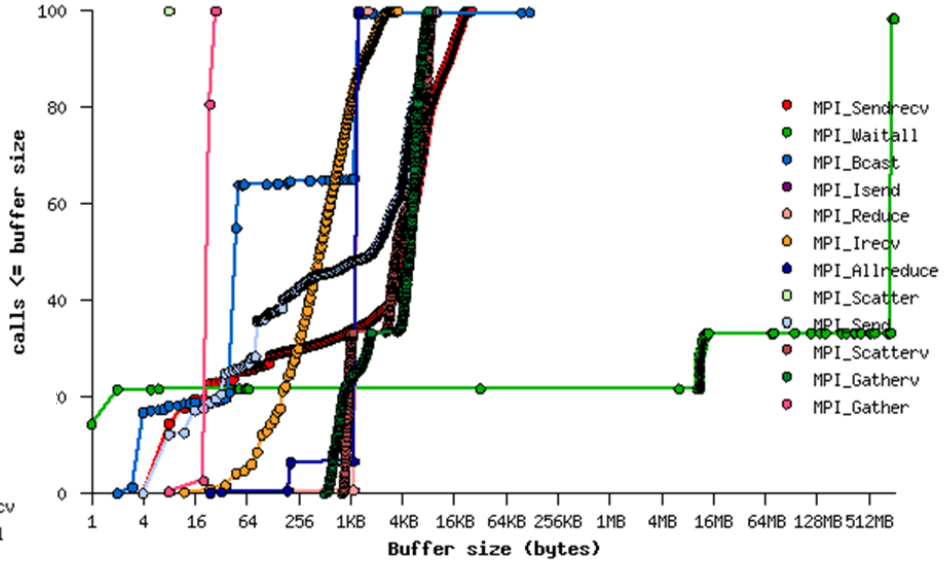


# GROMACS Profiling –Buffer Size Distributions

- **As cluster size grows, message size becomes smaller**
  - MPI_Sendrecv: 16KB to 64KB (16 Nodes). 4KB to 16KB (4 Nodes)
  - MPI_Gatherv/Scatterv: 1KB to 4KB (16 Nodes).  4KB to 16KB (4 nodes)

**4 Nodes**




**16 Nodes**

# GROMACS – Summary

- **Performance**
  - Intel Xeon E5-2680 and InfiniBand FDR enable GROMACS to scale
  - “Jupiter”, the E5-2680 cluster performs up to 110% over “Janus” the X5670 cluster
  - InfiniBand allows GROMACS to run at the most efficient rate
  - InfiniBand FDR outperforms 10GbE by 58% and 40GbE by 53% at 16 nodes
  - Performance of 1GbE stalls when run for more than 2 nodes
- **MPI**
  - Intel MPI performs slightly better than Open MPI (by 3%)
- **Profiling**
  - MPI_Sendrecv and MPI_Waitall are the most used MPI functions
  - MPI Communication takes place between close neighboring ranks
  - Majority of the MPI messages are concentrated in the range between 4KB to 16KB

# Thank You

## HPC Advisory Council