# GROMACS
# Performance Benchmark and Profiling

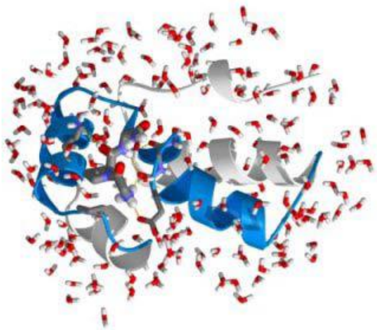
July 2015

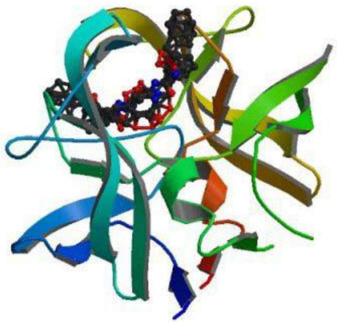
# Note

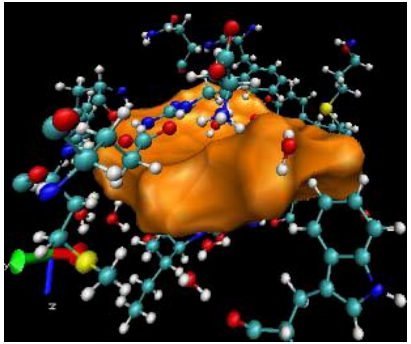
- **The following research was performed under the HPC Advisory Council activities**
  - Participating vendors: Intel, Dell, Mellanox
  - Compute resource - HPC Advisory Council Cluster Center
- **The following was done to provide best practices**
  - GROMACS performance overview
  - Understanding GROMACS 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







# Objectives

- **The presented research was done to provide best practices**
  - GROMACS performance benchmarking
    - CPU performance comparison
    - MPI library performance comparison
    - Interconnect performance comparison
    - System generations comparison
- **The presented results will demonstrate**
  - The scalability of the compute environment/application
  - Considerations for higher productivity and efficiency

# Test Cluster Configuration

- **Dell PowerEdge R730 32-node (896-core) “Thor” cluster**
  - Dual-Socket 14-Core Intel E5-2697v3 @ 2.60 GHz CPUs (Power Management in BIOS sets to Maximum Performance)
  - Memory: 64GB memory, DDR4 2133 MHz, Memory Snoop Mode in BIOS sets to Home Snoop, Turbo Enabled
  - OS: RHEL 6.5, MLNX_OFED_LINUX-3.0-1.0.1 InfiniBand SW stack
  - Hard Drives: 2x 1TB 7.2 RPM SATA 2.5” on RAID 1
- **Mellanox ConnectX-4 EDR 100Gbps EDR InfiniBand Adapters**
- **Mellanox Switch-IB SB7700 36-port 100Gb/s EDR InfiniBand Switch**
- **Mellanox ConnectX-3 FDR InfiniBand, 10/40GbE Ethernet VPI Adapters**
- **Mellanox SwitchX-2 SX6036 36-port 56Gb/s FDR InfiniBand / VPI Ethernet Switch**
- **MPI: Mellanox HPC-X v1.2.0-326**
- **Compiler and Libraries: Intel Composer XE 2015.3.187 and MKL**
- **Application: GROMACS 4.6.7**
- **Benchmark datasets: DPPC in Water (d.dppc, 121856 atoms, 150000 steps, SP) unless stated otherwise**

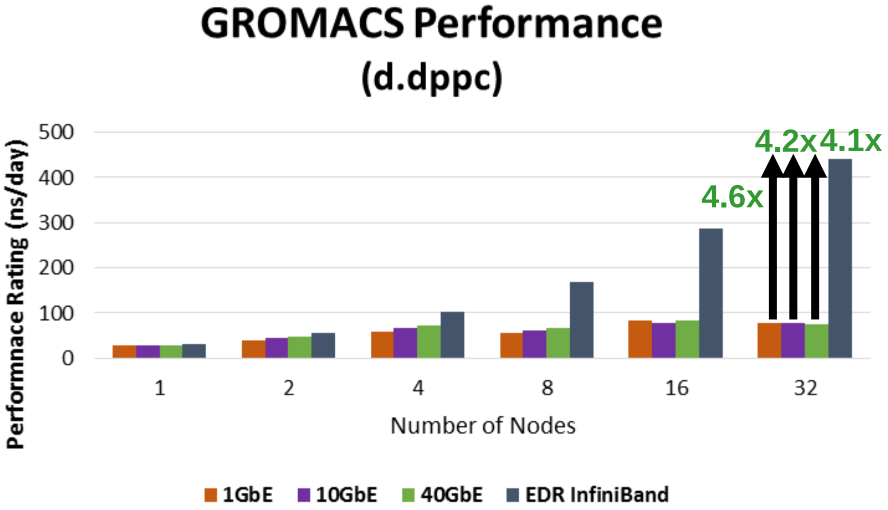
# PowerEdge R730

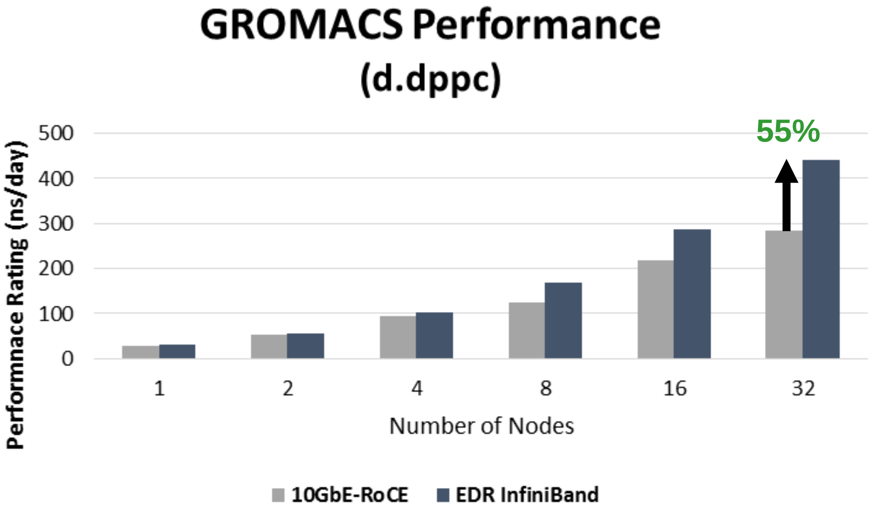
## Massive flexibility for data intensive operations

- **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 – Network Interconnects

- **InfiniBand is the only interconnect that delivers superior scalability performance**
  - EDR InfiniBand provides higher performance and more scalable than 1GbE, 10GbE, or 40GbE
  - Performance for Ethernet stays flat (or stops scaling) beyond 2 nodes
  - EDR InfiniBand outperforms 10GbE-RoCE on scalability performance by 55% at 32 nodes / 896c
  - EDR InfiniBand demonstrates continuous performance gain at scale



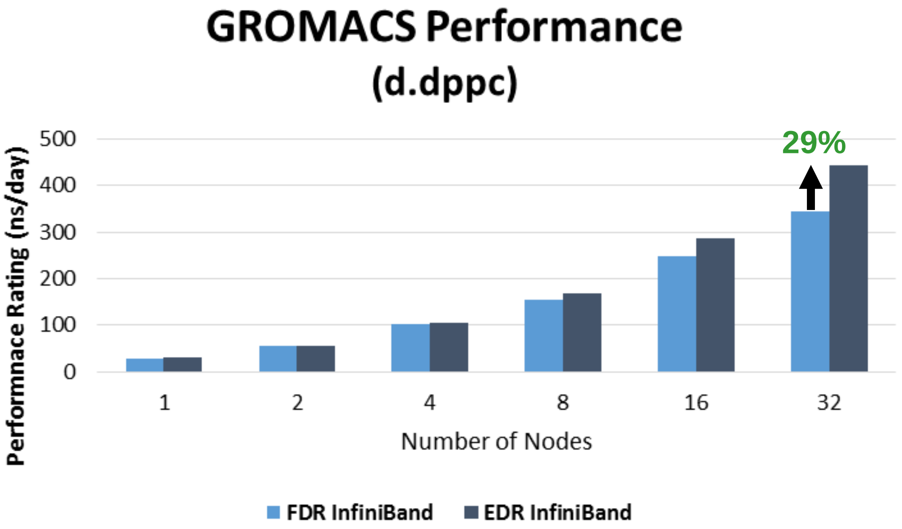


*Higher is better*

*28 MPI Processes / Node*

# GROMACS Performance – EDR vs FDR InfiniBand

- **EDR InfiniBand delivers superior scalability in application performance**
  - As the number of nodes scales, performance gap of EDR IB becomes widen
- **Performance advantage of EDR InfiniBand increases for larger core counts**
  - EDR InfiniBand provides 29% versus FDR InfiniBand at 32 nodes (896 cores)

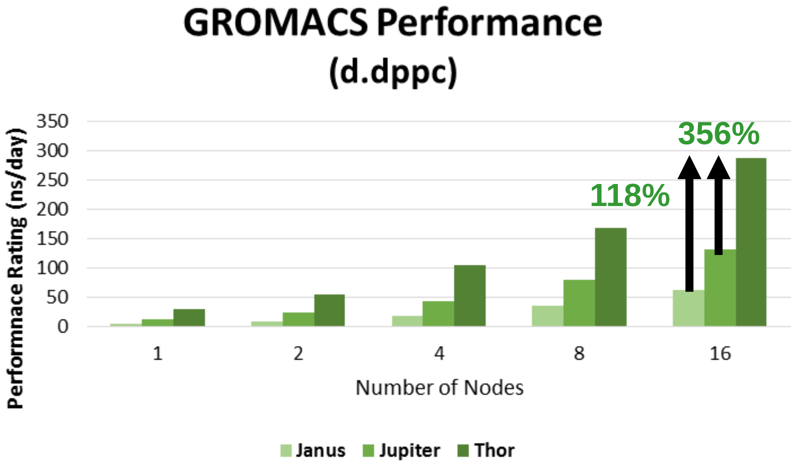


*Higher is better*

*28 MPI Processes / Node*

# GROMACS Performance – System Generations

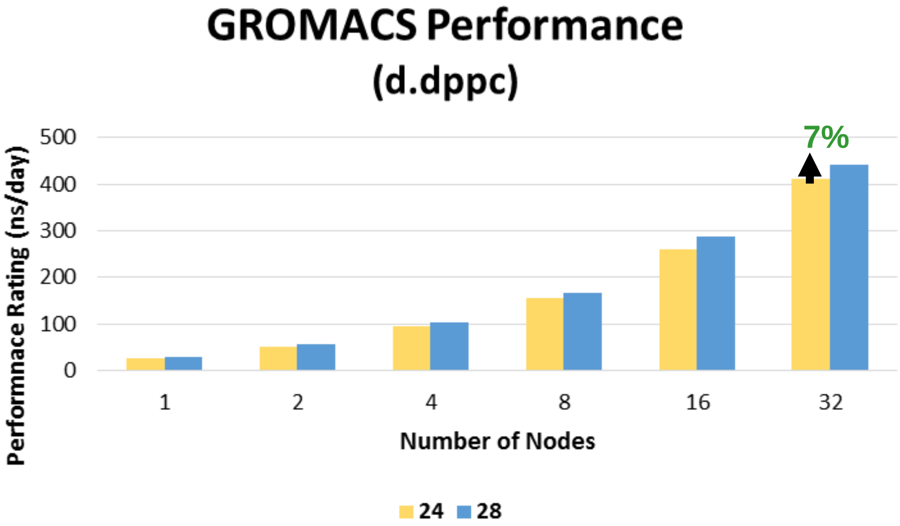
- **Thor cluster (based on Intel E5-2697v3 - Haswell) outperforms prior generations**
  - 1.1 to 3.5x higher performance than clusters based on previous generations of Intel architecture
- **System components used:**
  - Janus: 2-socket 6-core Xeon X5670 @ 2.93GHz, 1333MHz DIMMs, ConnectX-2 QDR IB
  - Jupiter: 2-socket 8-core Xeon E5-2680 @ 2.7GHz, 1600MHz DIMMs, ConnectX-3 FDR IB
  - Thor: 2-socket 14-core Xeon E5-2680V3 @2.6GHz, 2133MHz DIMMs, ConnectX-4 EDR IB



*Higher is better*

# GROMACS Performance – Cores Per Node

- **Running more CPU cores provides higher performance**
  - ~7-10% higher productivity with 28PPN compared to 24PPN
  - Higher demand on memory bandwidth and network might limit performance as more cores are used

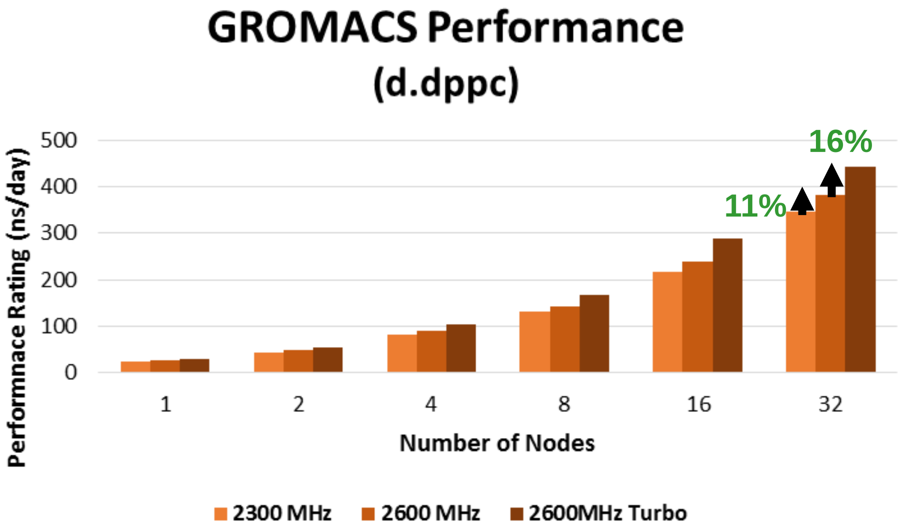


*Higher is better*

*CPU @ 2.6GHz*

# GROMACS Performance – Turbo Mode & CPU Clock

- **Advantages are seen with running higher clock rate**
  - Either by enabling Turbo mode or higher CPU clock frequency
- **Boosting CPU clock rate yields higher performance at lower cost**
  - Increasing to 2600MHz (from 2300MHz) run 11% faster

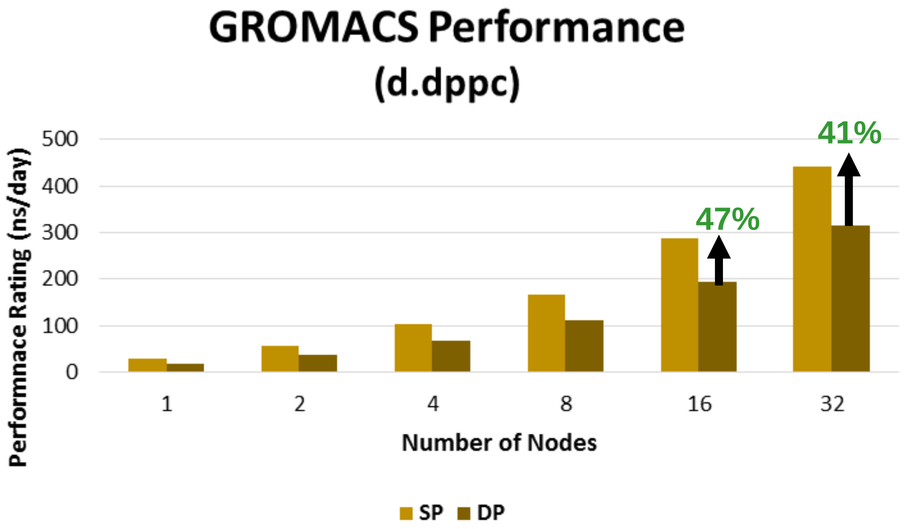


*Higher is better*

*CPU @ 2.6GHz*

# GROMACS Performance – Floating Point Precision

- **GROMACS allows running either SP and DP for floating point precision**
- **Running at SP is shown to be faster than running at DP**
  - Seen around 41%-47% faster running at SP (Single Precision) versus DP (Double Precision)
  - All other slides are running using Single Precision

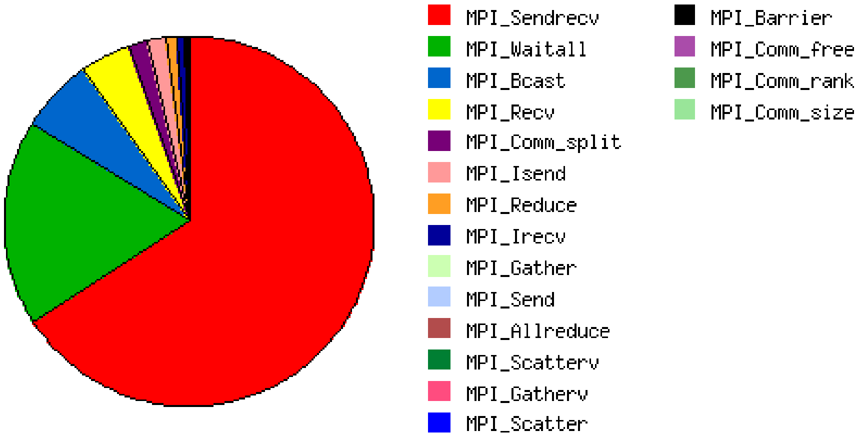


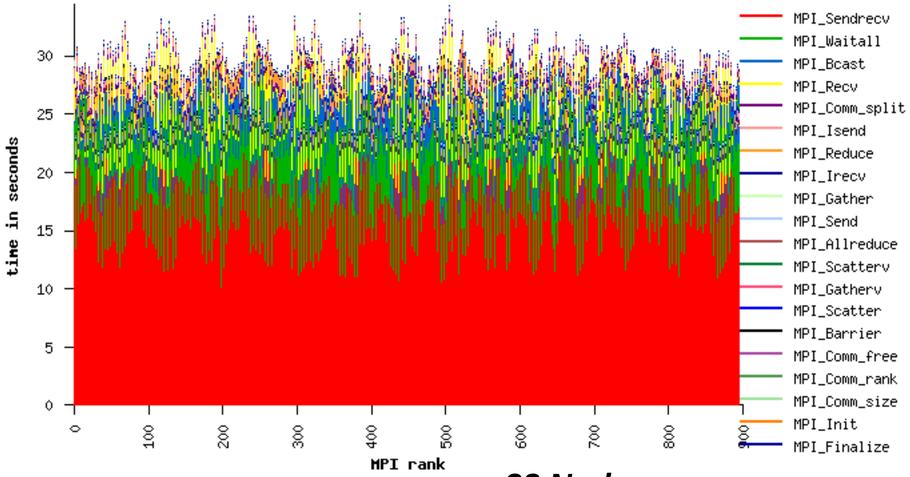
*Higher is better*

*CPU @ 2.6GHz*

# GROMACS Profiling – Time Spent by MPI Calls

- **The most time consuming MPI call os MPI_Sendrecv**
  - MPI_Sendrecv: 66% (or 27% of runtime) at 32 nodes (896 cores)
  - MPI_Waitall: 18% (or 7% of runtime), MPI_Bcast: 6% (or 2% of runtime)
  - Point to point and non-blocking sends and receives consume most time in GROMACS

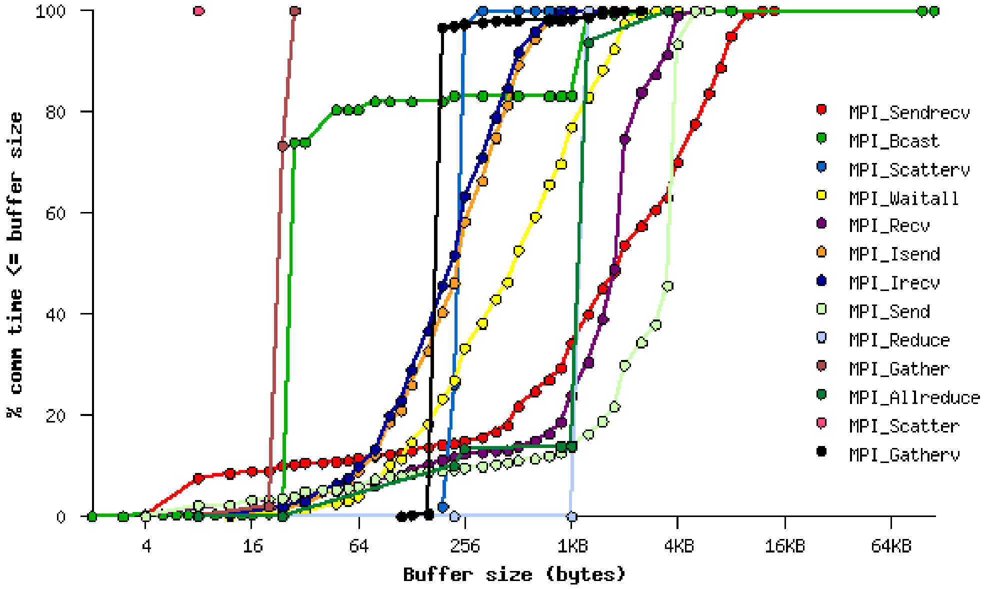




***32 Nodes***

# GROMACS Profiling – MPI Message Sizes

- **Majority of data transfer messages are medium sizes, except for:**
  - MPI_Sendrecv has a large concentration (from 8B to 8KB)
  - MPI_Bcast shows some concentration

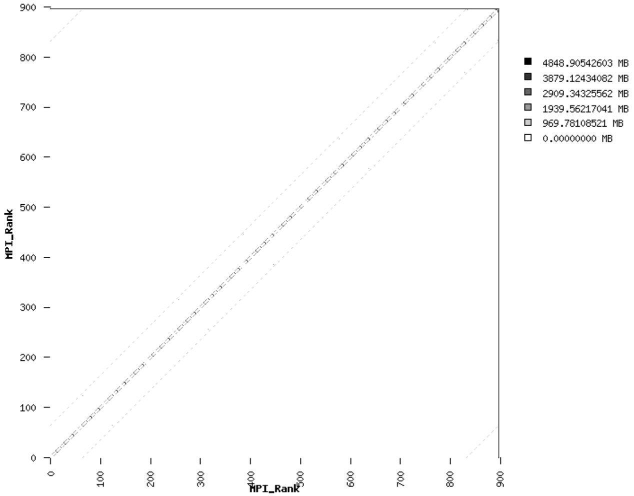


***32 Nodes***

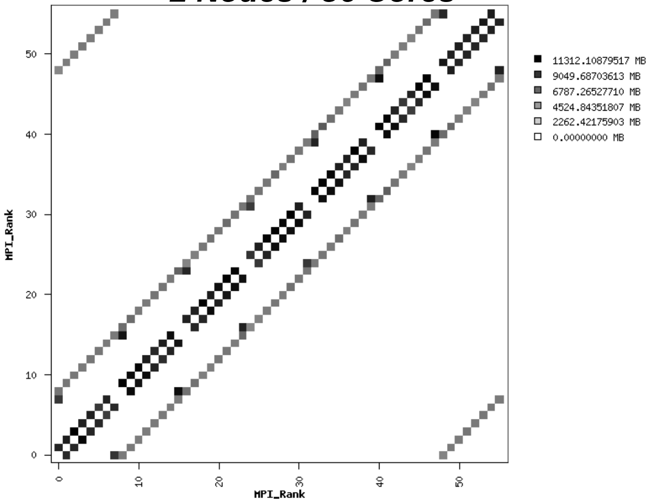
# GROMACS Profiling – MPI Data Transfer

- **As the cluster grows, similar communication behavior is seen**
  - Majority of communications are between neighboring ranks
  - Non-blocking (point to point) data, and point-to-point transfers are shown in the graph
  - Collective data communications are small compared to point-to-point communications

***32 Nodes / 896 Cores***



***2 Nodes / 56 Cores***



# GROMACS Summary

- **Latest system generation improve GROMACS performance at scale**
  - Compute: Intel Haswell cluster outperforms system architecture of previous generations
    - Haswell cluster outperforms Sandy Bridge cluster by 110%, and outperforms Westmere cluster by 350% at 32 node
  - Compute: Running more CPU cores provides higher performance
    - ~7-10% higher productivity with 28PPN compared to 24PPN
  - Network: EDR InfiniBand delivers superior scalability in application performance
    - EDR InfiniBand provides higher performance and more scalable than 1GbE, 10GbE, or 40GbE
    - Performance for Ethernet (1GbE/10GbE/40GbE) stays flat (or stops scaling) beyond 2 nodes
    - EDR InfiniBand outperforms 10GbE-RoCE on scalability performance by 55% at 32 nodes / 896c
  - Running at Single Precision is approximately twice as fast as running at Double Precision
    - Seen around 41%-47% faster running at SP (Single Precision) versus DP (Double Precision)
  - MPI Profile shows majority of data transfer are point-to-point and non-blocking communications
    - MPI_Sendrecv and MPI_Waitall are the most used MPI 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 undertakes no duty and assumes no obligation to update or correct any information presented herein