
# GROMACS
# Performance Benchmark and Profiling

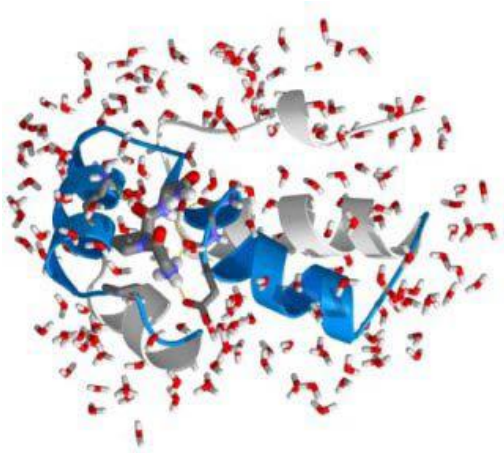
September 2012

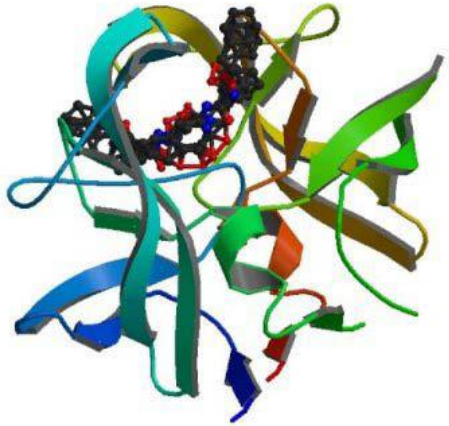
# Note

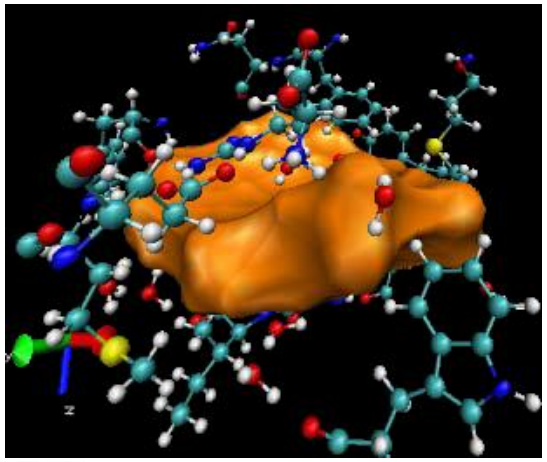
- **The following research was performed under the HPC Advisory Council activities**
  - Participating vendors: AMD, Dell, Mellanox
  - Compute resource –
    - HPC Advisory Council Cluster Center
- **For more info please refer to**
  - http:// www.amd.com
  - http:// www.dell.com/hpc
  - http://www.mellanox.com
  - 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 following was done to provide best practices**
  - GROMACS performance benchmarking
  - Understanding GROMACS communication patterns
  - Ways to increase GROMACS productivity
  - Compilers and network interconnects comparisons

- **The presented results will demonstrate**
  - The scalability of the compute environment
  - The capability of GROMACS to achieve scalable productivity
  - Considerations for performance optimizations

# Test Cluster Configuration

- **Dell™ PowerEdge™ C6145 6-node (384-core) cluster**
  - Memory: 128GB memory per node DDR3 1600MHz, BIOS version 2.6.0
  - 4 CPU sockets per server node
- **AMD™ Opteron™ 6276 (code name "Interlagos") 16-core @ 2.3 GHz CPUs**
- **Mellanox ConnectX®-3 VPI Adapters and IS5030 36-Port InfiniBand switch**
- **MLNX-OFED 1.5.3 InfiniBand SW stack**
- **OS: RHEL 6 Update 2, SLES 11 SP2**
- **MPI: Intel MPI 4 Update 3, Open MPI 1.5.5, Platform MPI 8.2.1**
- **Compilers: GNU 4.7**
- **Application: GROMACS 4.5.5**
- **Benchmark workload:**
  - DPPC in Water (d.dppc) (5000 steps, 10.0 ps.)

# Dell™ PowerEdge™ C6145 6-node cluster

- **HPC Advisory Council Test-bed System**

- **New 6-node 384 core cluster - featuring Dell PowerEdge™ C6145 servers**
  - Replacement system for Dell PowerEdge SC1435 (192 cores) cluster system following 2 years of rigorous benchmarking and product EOL
    - System to be redirected to explore HPC in the Cloud applications

- **Workload profiling and benchmarking**
  - Characterization for HPC and compute intense environments
  - Optimization for scale, sizing and configuration and workload performance
  - Test-bed Benchmarks
    - RFPs
    - Customers/Prospects, etc
  - ISV & Industry standard application characterization
  - Best practices & usage analysis

# About Dell PowerEdge™ Platform Advantages

## Best of breed technologies and partners

Combination of AMD Opteron™ 6200 series platform and Mellanox ConnectX®-3 InfiniBand on Dell HPC

Solutions provide the ultimate platform for speed and scale

- Dell PowerEdge C6145 system delivers 8 socket performance in dense 2U form factor
- Up to 64 core/32DIMMs per server – 2688 core in 42U enclosure

## Integrated stacks designed to deliver the best price/performance/watt

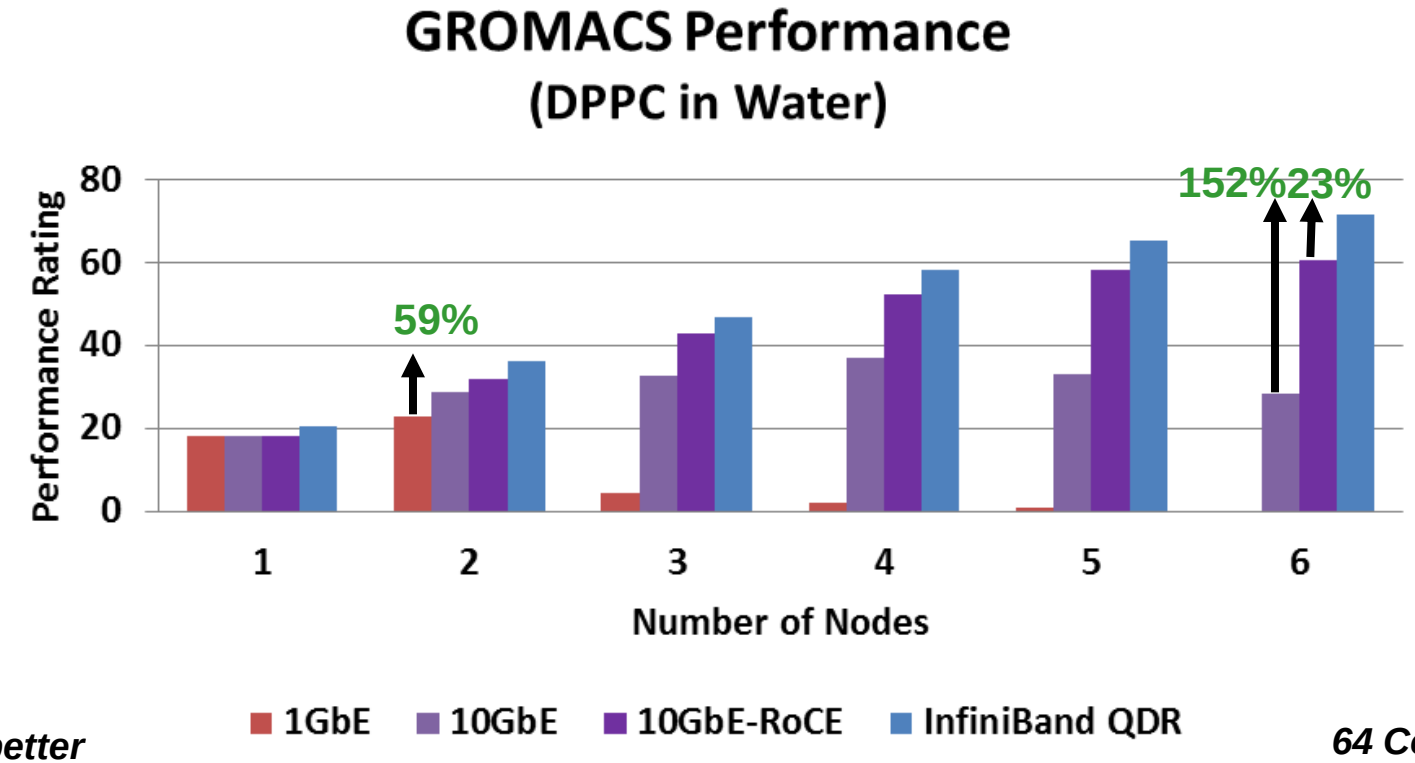
- 2x more memory and processing power in half of the space
- Energy optimized low flow fans, improved power supplies and dual SD modules

## Optimized for long-term capital and operating investment protection

- System expansion
- Component upgrades and feature releases

# GROMACS Performance – Interconnect

- **InfiniBand QDR delivers the best performance for GROMACS**
  - Seen up to 152% better performance than 10GbE on 6 nodes
  - Seen up to 59% better performance than 1GbE on 2 nodes
- **Scalability limitation seen with Ethernet networks**
  - 10GigE performance starts to drop after 3-node
  - 1GigE performance drop takes place after 2-node

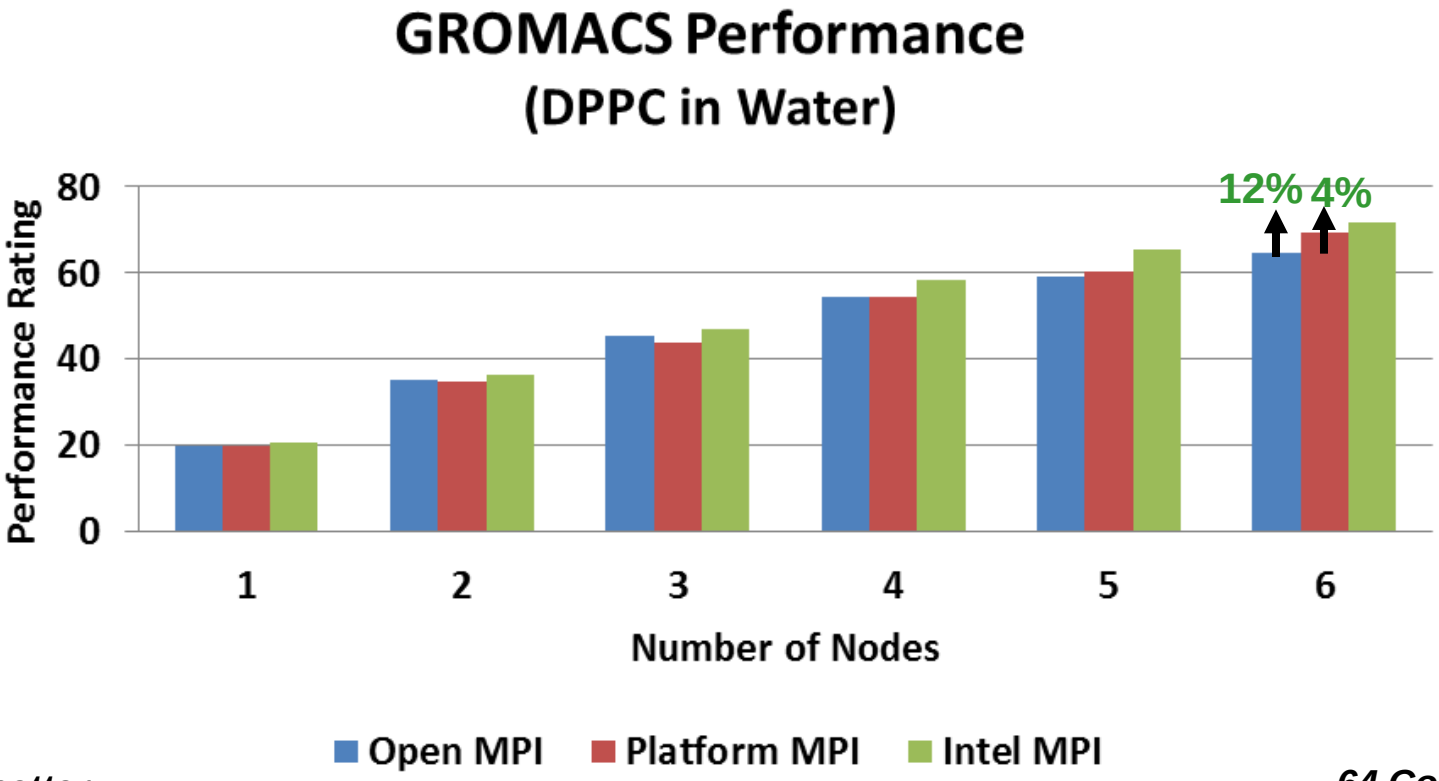



*Higher is better*

*64 Cores/Node*

# GROMACS Performance – MPI

- **Intel MPI delivers better scalability for GROMACS**
  - 12% higher performance than Open MPI at 6 nodes
  - 4% higher performance than Platform MPI at 6 nodes

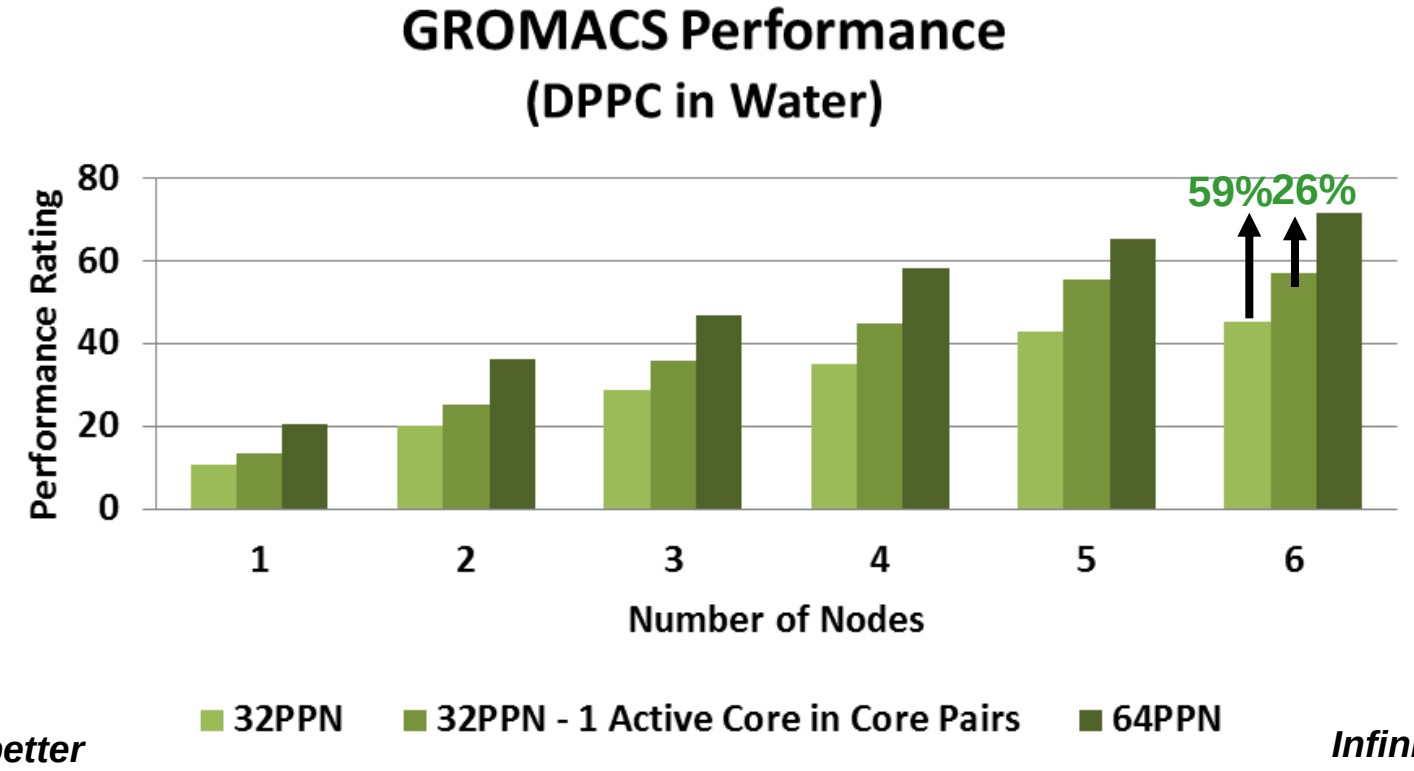



*Higher is better*

*64 Cores/Node*

# GROMACS Performance – Processes Per Node

- **Allocating more processes per node can yield higher system utilization**
  - 59% gain in performance with 4P servers versus 2P servers when comparing at 6 nodes
- **Using 64 PPN delivers higher performance than 32PPN using 1 active core**
  - 26% gain in performance with 64 PPN versus 32 PPN (with 1 active core) for 6 nodes
  - GROMACS can fully utilized all CPU cores available in a system

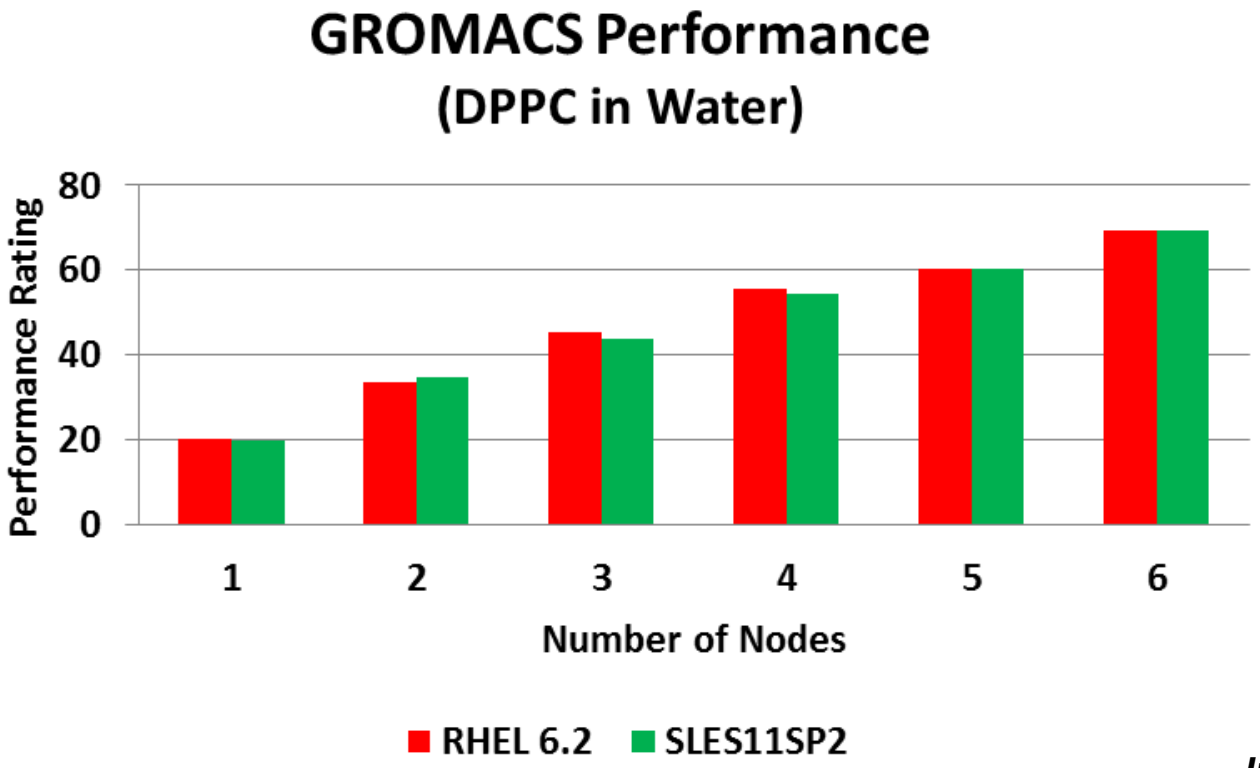



*Higher is better*

*InfiniBand QDR*

# GROMACS Performance – Operating Systems

- **Both SLES11SP2 and RHEL 6.2 perform at the same level of performance**
  - SLES performs slightly better on a single node while RHEL performs better at scale

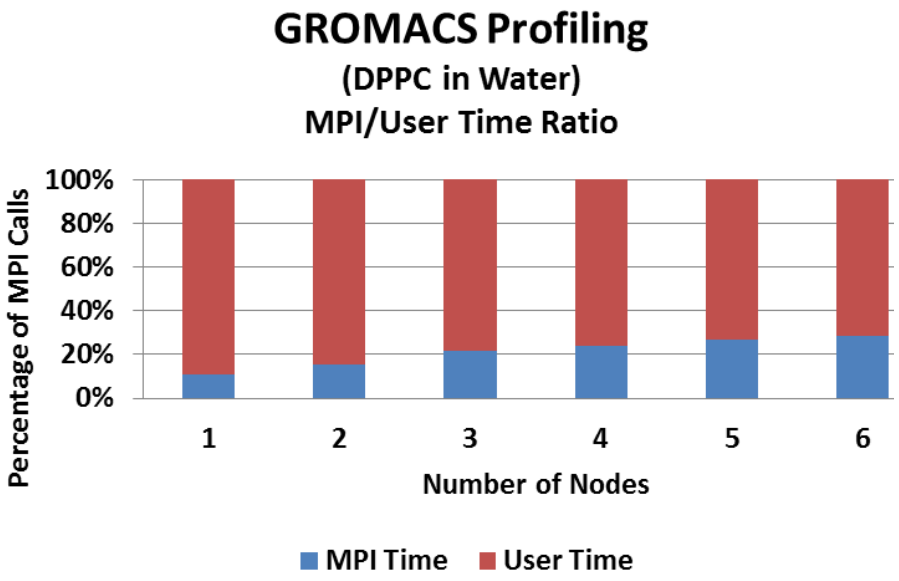


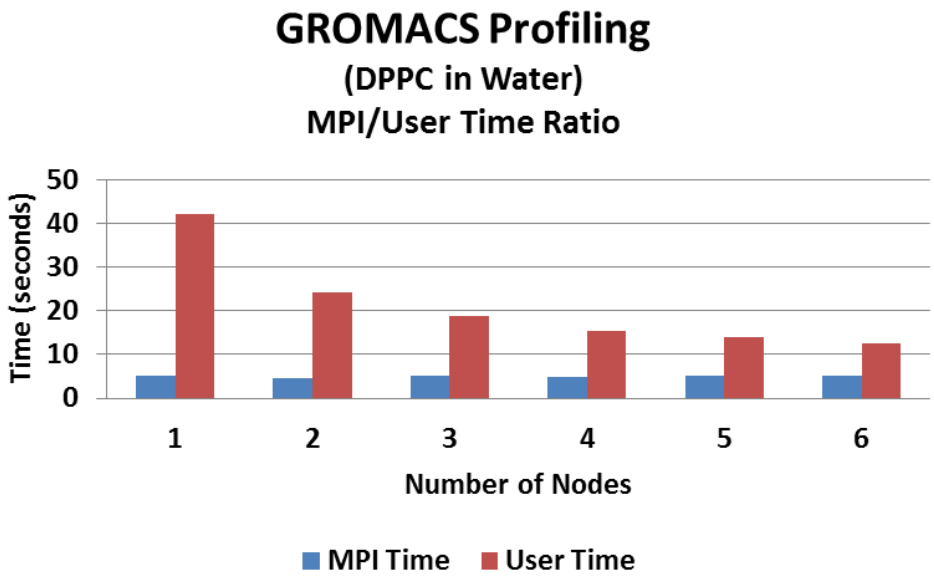

*Higher is better*

*InfiniBand QDR*

# GROMACS Profiling – MPI/User Time Ratio

- **InfiniBand QDR reduces the amount of time for MPI communications**
  - MPI Communication time stays flat as the compute time reduces

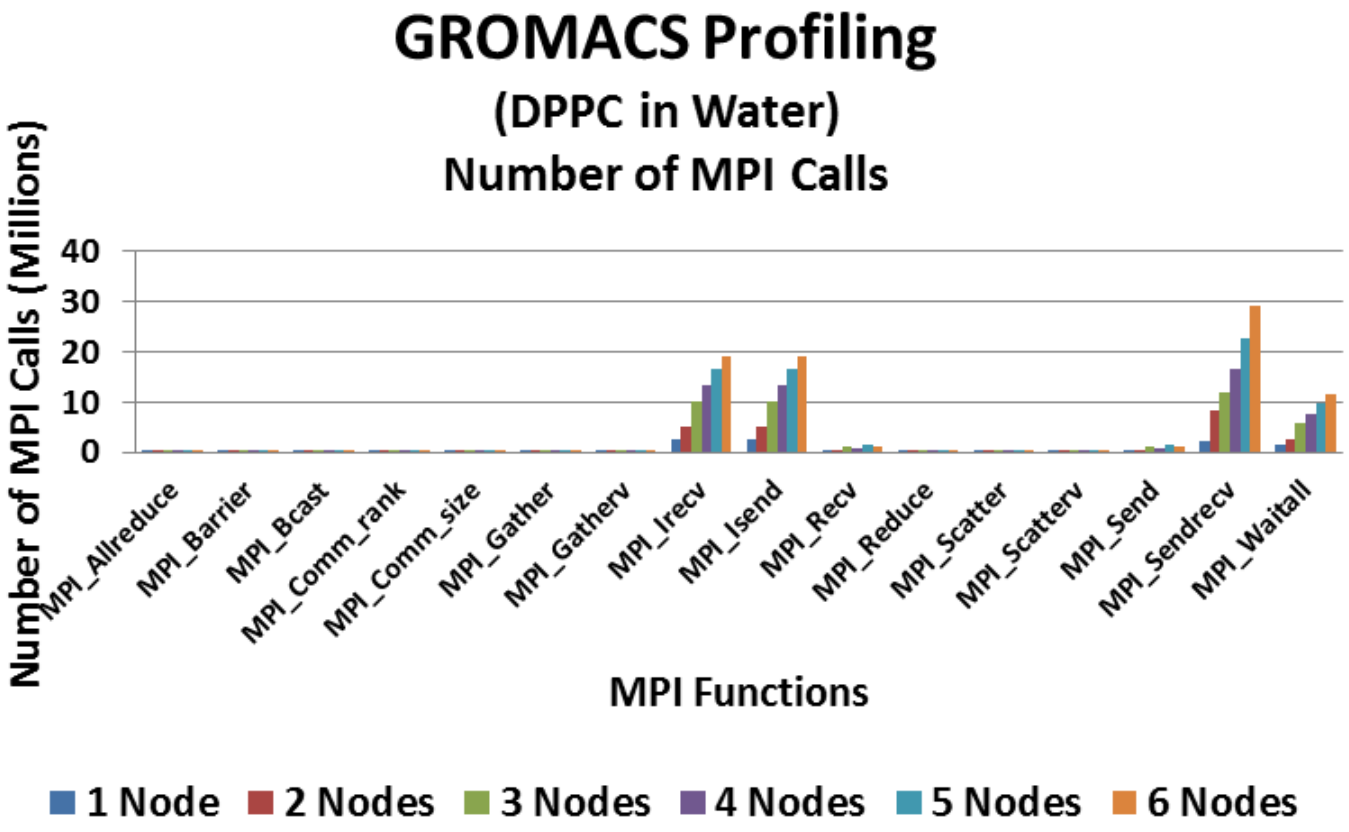



***64 Cores/Node***

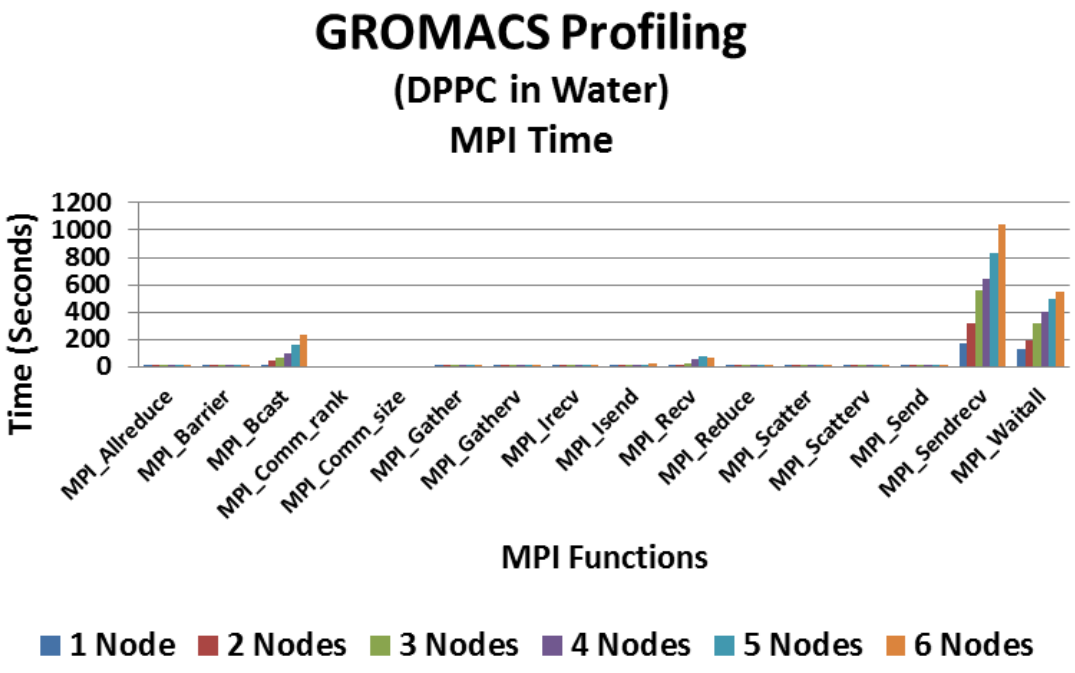
# GROMACS Profiling – Number of MPI Calls

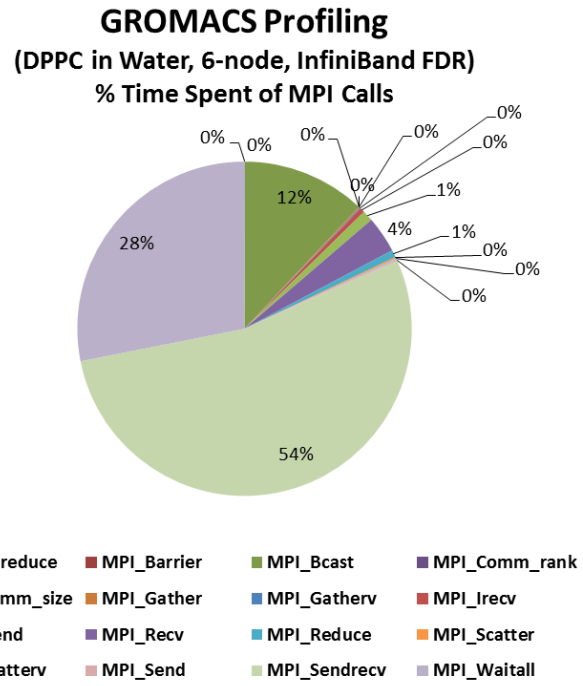
- **The most used MPI functions are for data transfers**
  - MPI_Sendrecv (35%), MPI_Isend (23%), MPI_Irecv (23%), MPI_Waitall (14%)
  - Reflects that GROMACS requires good network throughput
- **The number of calls increases proportionally as the cluster scales**



# GROMACS Profiling – Time Spent of MPI Calls

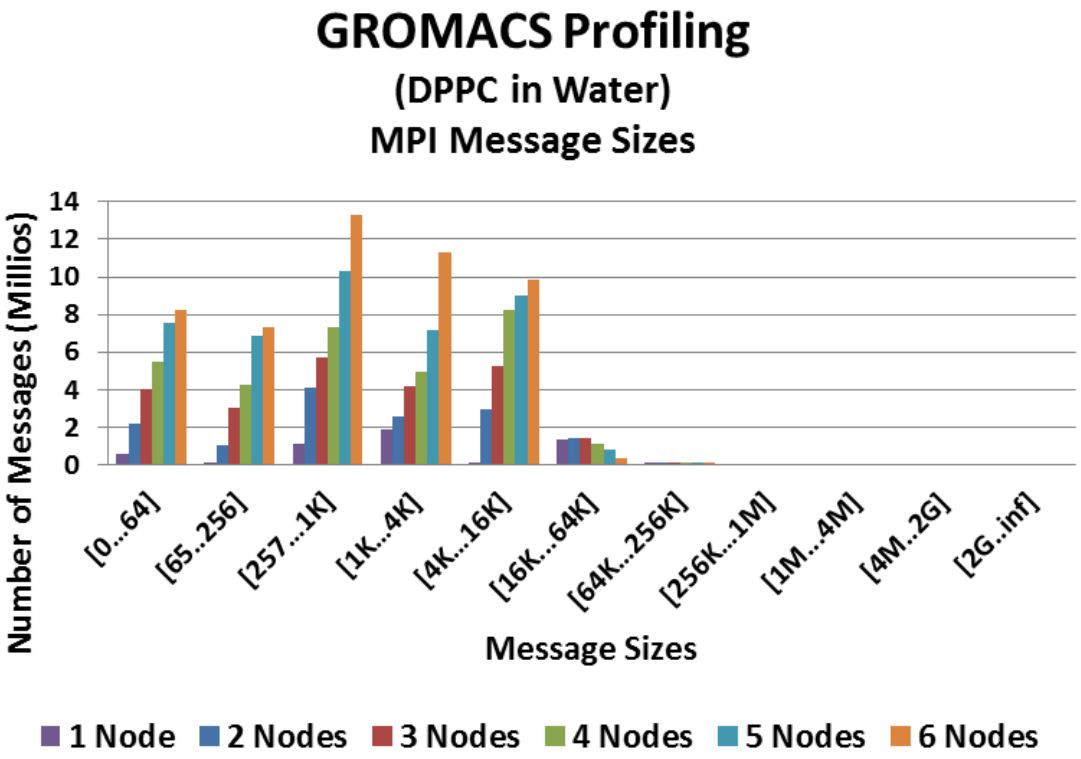
- **The time in communications is taken place in the following MPI functions:**
  - MPI_Sendrecv (54%) MPI_Waitall (28%), MPI_Bcast (12%)



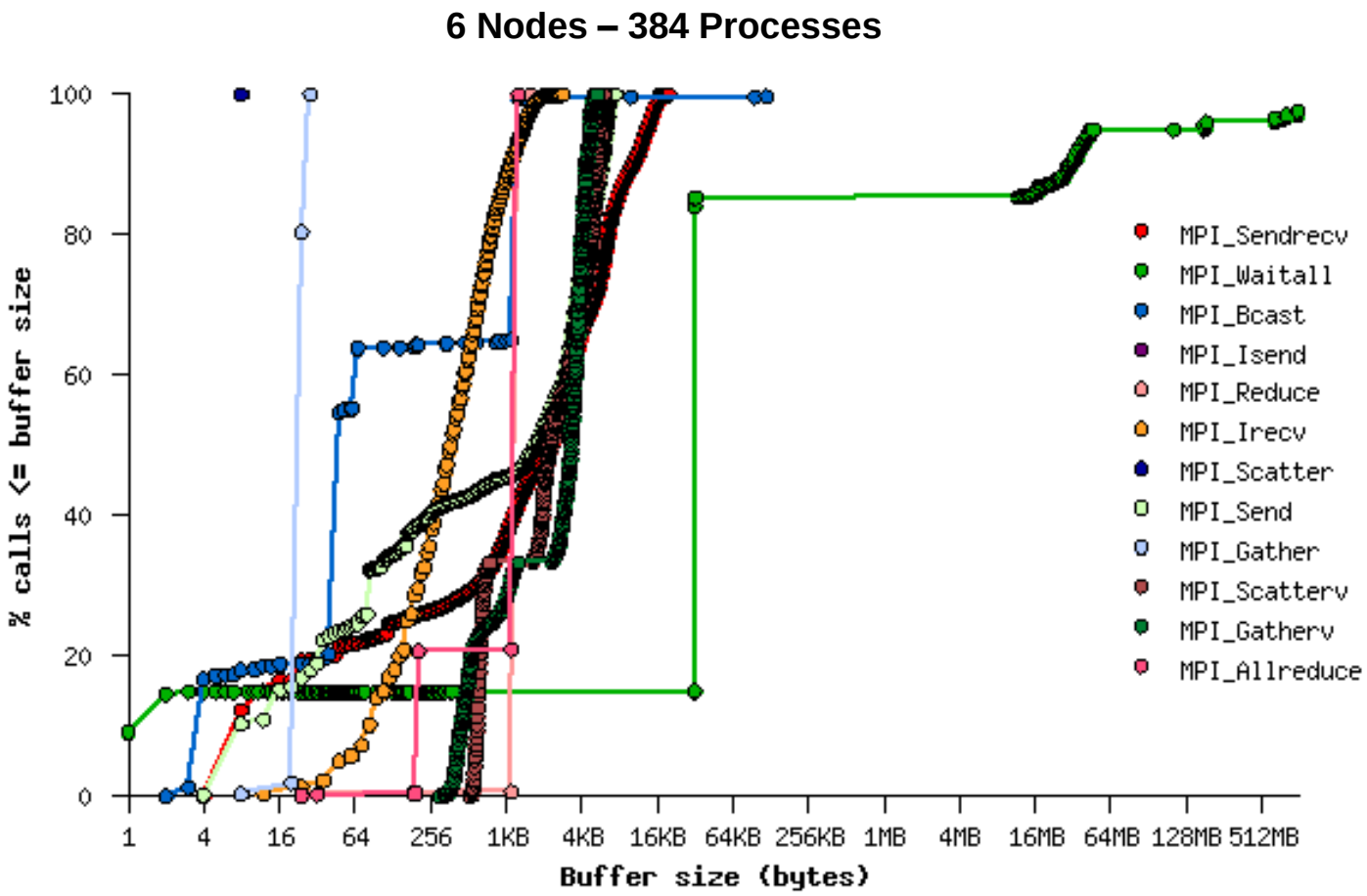
# GROMACS Profiling – MPI Message Sizes

- **Majority of the MPI messages are small to median message sizes**
  - In the ranges of between 257B and 1KB
  - All of the MPI messages are in the sizes less than 256KB
- **Low network latency requires for good small MPI message performance**



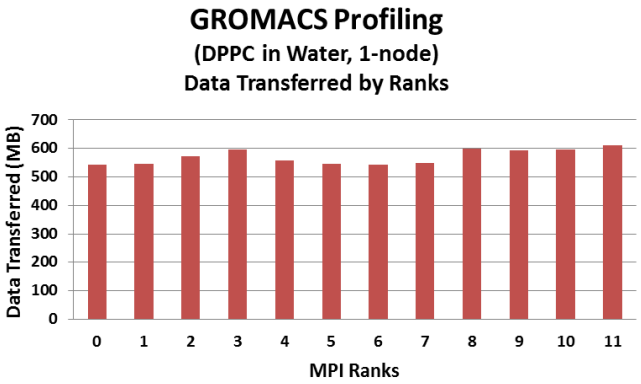
# GROMACS Profiling – MPI Message Sizes

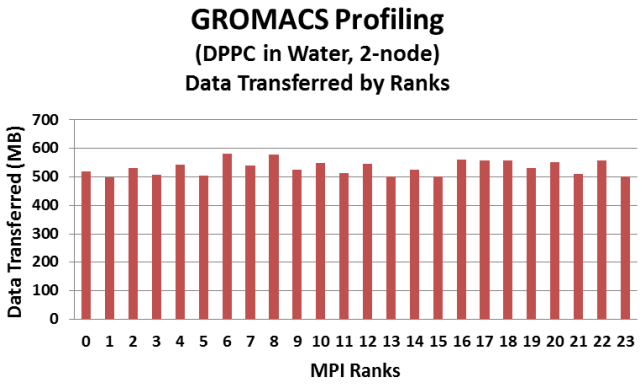
- **Large concentration of MPI calls are small to median message sizes**
  - MPI_Irecv: In the ranges of between 257B and 1KB

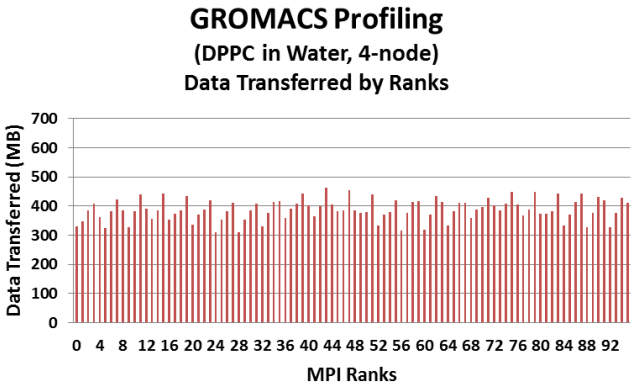


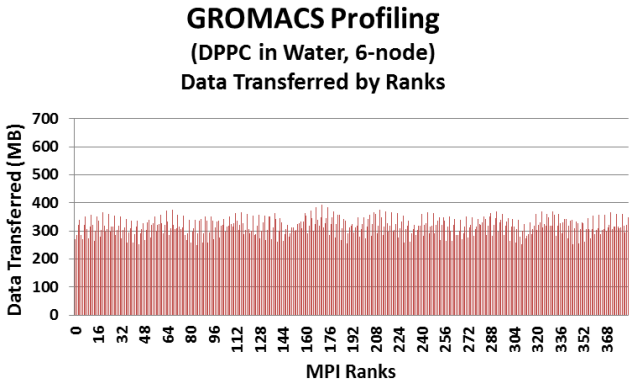
# GROMACS Profiling – Data Transfer By Process

- **Data transferred to each MPI rank is generate constant for all MPI processes**
  - Amount of data transfer to each rank is reduced as more nodes are in the job
  - From around 600MB per rank on 1-node down to around 300MB per rank for 6-node



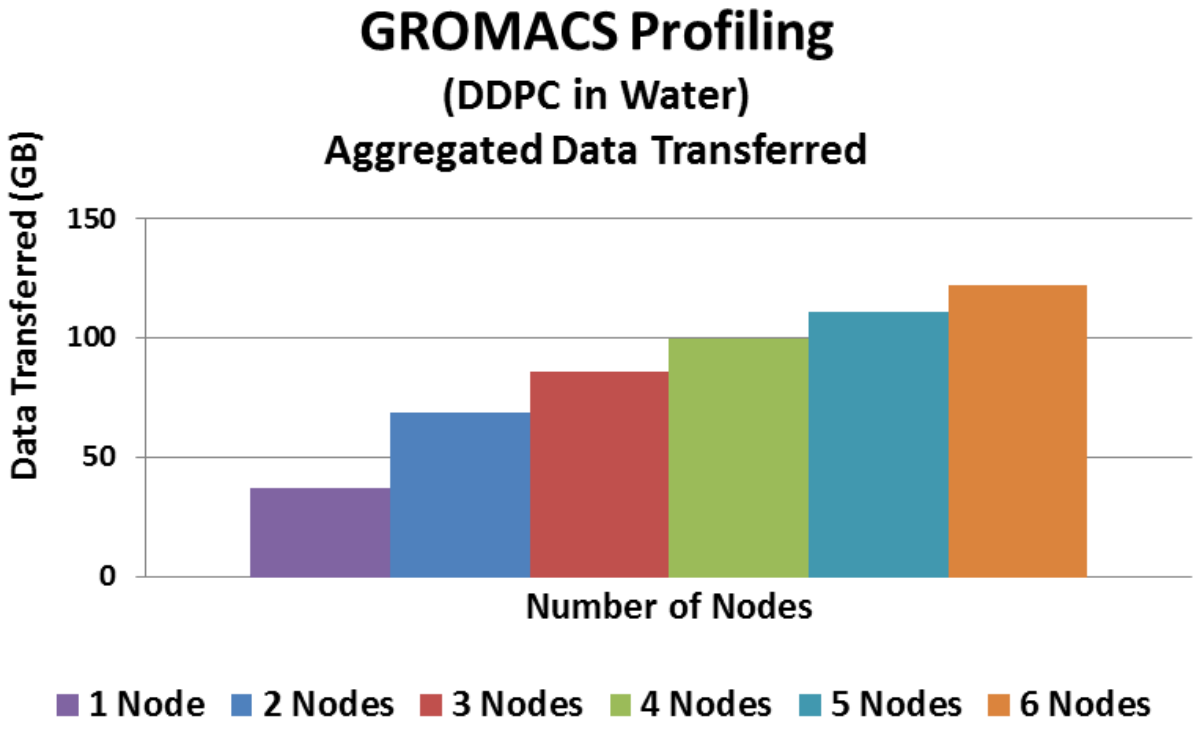



# GROMACS Profiling – Aggregated Data Transfer

- **Aggregated data transfer refers to:**
  - Total amount of data being transferred in the network between all MPI ranks collectively
- **The total data transfer increases steadily as the cluster scales**
  - For this dataset, a good amount of data being sent and received across the network
  - As a compute node being added, more data communications will take place

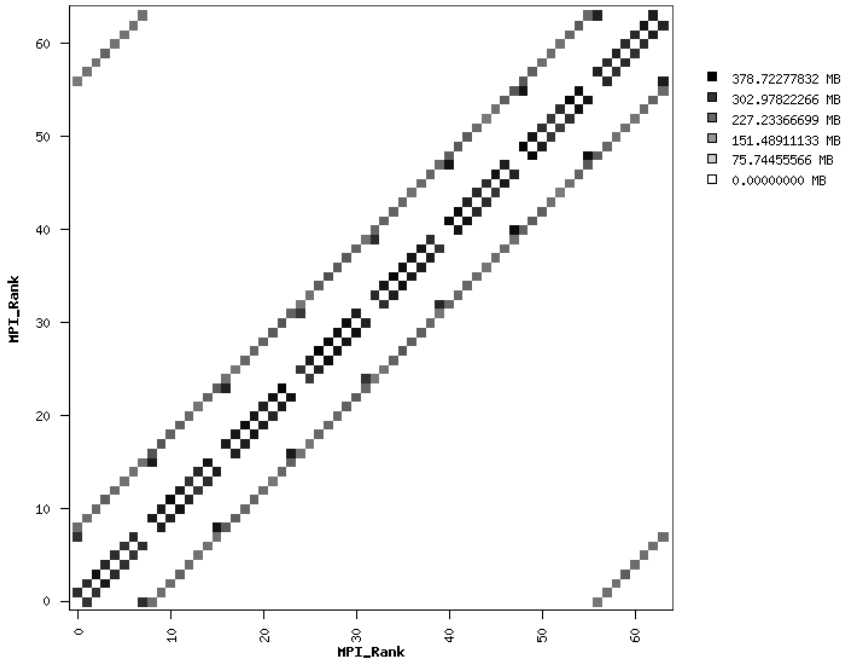



***InfiniBand QDR***

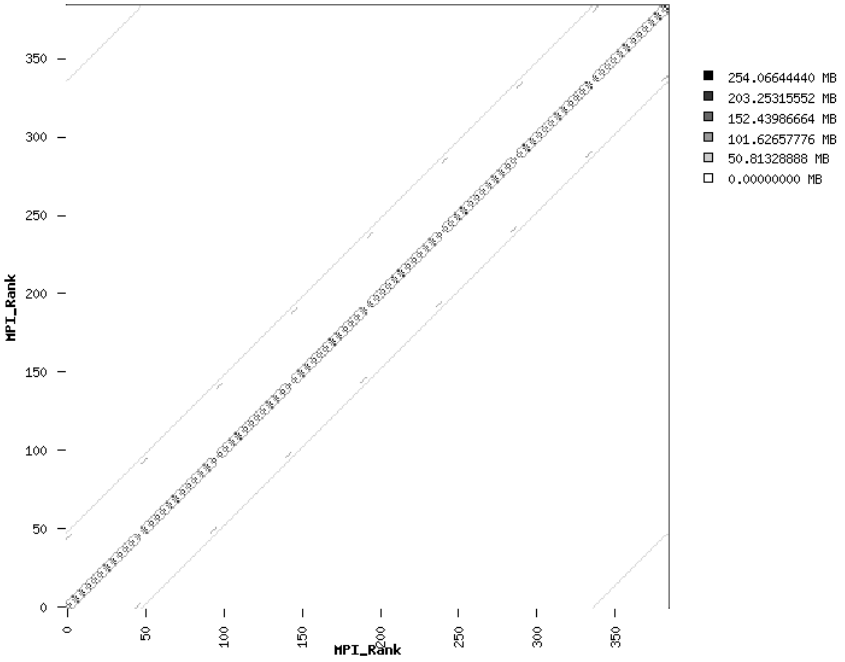
# GROMACS Profiling – Point To Point Data Flow

- **The point to point data flow shows the communication pattern of GROMACS**
  - GROMACS mainly communicates mainly its neighbors and close ranks
  - The pattern stays the same as the cluster scales

**1 Nodes – 64 Processes**




**6 Nodes – 384 Processes**




***InfiniBand QDR***

# Summary

- **GROMACS is a memory and network latency sensitive application**
- **CPU:**
  - Using 4P systems delivers 59% higher performance than 2P systems (at 6 nodes)
  - Using 64 PPN delivers 26% higher performance than 32PPN using 1 active core
- **Interconnects:**
  - InfiniBand QDR can deliver good scalability for GROMACS
    - Provides up to 142% better performance than 10GbE on 6 nodes
    - Provides up to 52% better performance than 1GbE on 2 nodes
  - 10GigE and 1GigE would not scale and become inefficient to run beyond 2-3 nodes
- **MPI:**
  - Intel MPI achieves higher scalability than Open MPI and Platform MPI for GROMACS
- **OS:**
  - Both SLES 11 SP 2 and RHEL 6 Update 2 provides similar level of performance

# Thank You

## HPC Advisory Council