HPC ADVISORY COUNCIL

LAMMPS
Performance Benchmark and Profiling

July 2012

DELL intel Mellanox TECHNOLOGIES Sandia National Laboratories

Note

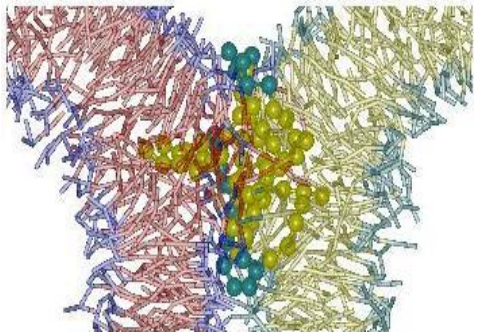
- **The following research was performed under the HPC Advisory Council activities**
  - Participating vendors: Intel, Dell, Mellanox
  - Compute resource - HPC Advisory Council Cluster Center

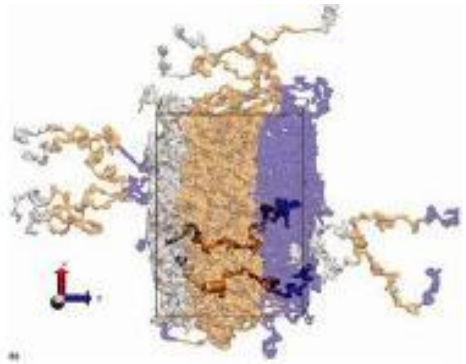
- **The following was done to provide best practices**
  - LAMMPS performance overview
  - Understanding LAMMPS communication patterns
  - Ways to increase LAMMPS productivity
  - MPI libraries comparisons

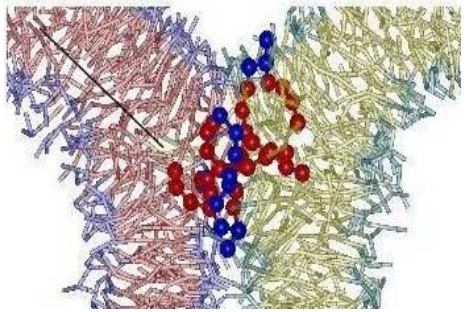
- **For more info please refer to**
  - http://www.dell.com
  - http://www.intel.com
  - http://www.mellanox.com
  - http://lammps.sandia.gov/

LAMMPS

- **Large-scale Atomic/Molecular Massively Parallel Simulator**
  - Classical molecular dynamics code which can model:
    - Atomic
    - Polymeric
    - Biological
    - Metallic
    - Granular, and coarse-grained systems
- **LAMMPS runs efficiently in parallel using message-passing techniques**
  - Developed at Sandia National Laboratories
  - An open-source code, distributed under GNU Public License







Objectives

- **The following was done to provide best practices**
  - LAMMPS performance benchmarking
  - Interconnect performance comparisons
  - Understanding LAMMPS communication patterns
  - Power-efficient simulations

- **The presented results will demonstrate**
  - The scalability of the compute environment to provide nearly linear application scalability
  - The capability of LAMMPS to achieve scalable productivity

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**
- **MPI and libraries: Intel MPI 4 Update 3, Platform MPI 8.2, Intel MKL 10.3 Update 10**
- **Application: LAMMPS-4Jul12**
- **Benchmarks:**
  - Rhodo - Rhodopsin protein in solvated lipid bilayer, CHARMM force field with a 10 Angstrom LJ cutoff

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

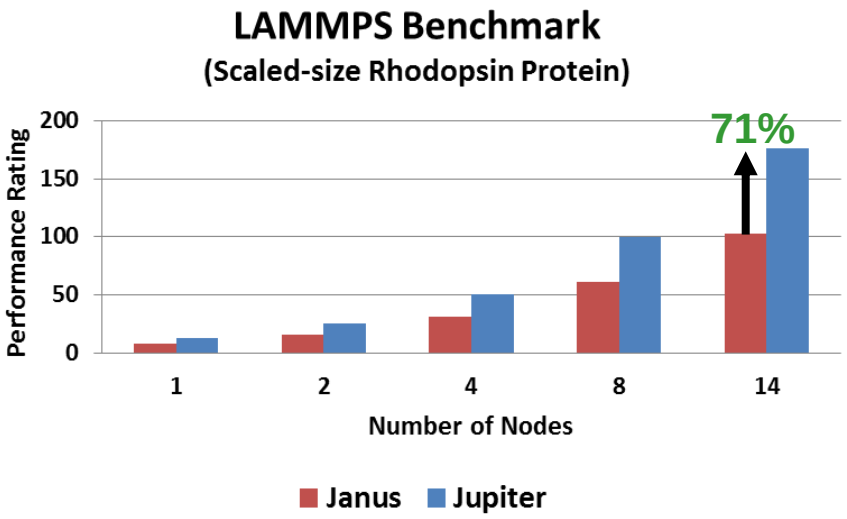
About PowerEdge R720xd

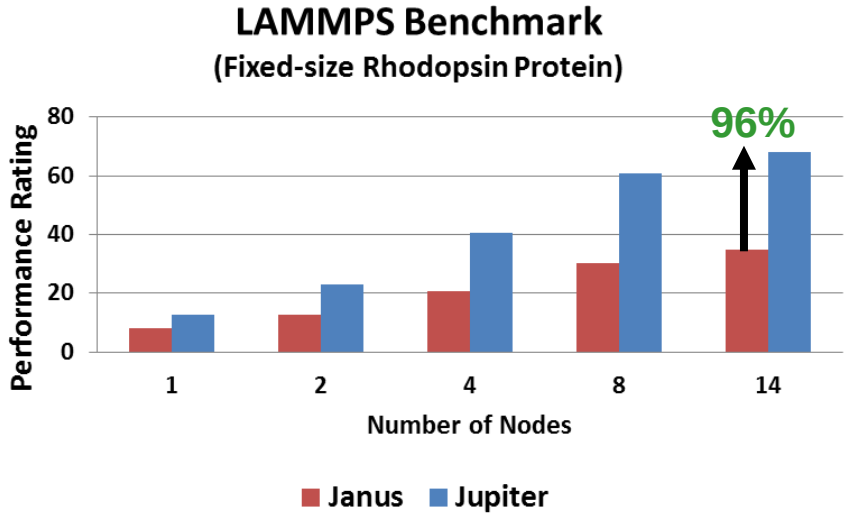
Massive flexibility for data intensive operations

- **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

LAMMPS Performance – Processors

- **Intel E5-2680 (Sandy Bridge) cluster outperforms prior generations**
  - Performs 96% better than X5670 cluster at 14 nodes with fixed-size Rhodo test
  - Performs 71% better than X5670 cluster at 14 nodes with scaled-size Rhodo test
- **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
- **14 nodes are used in the comparison**
  - In order to compare with results previously done on Janus cluster



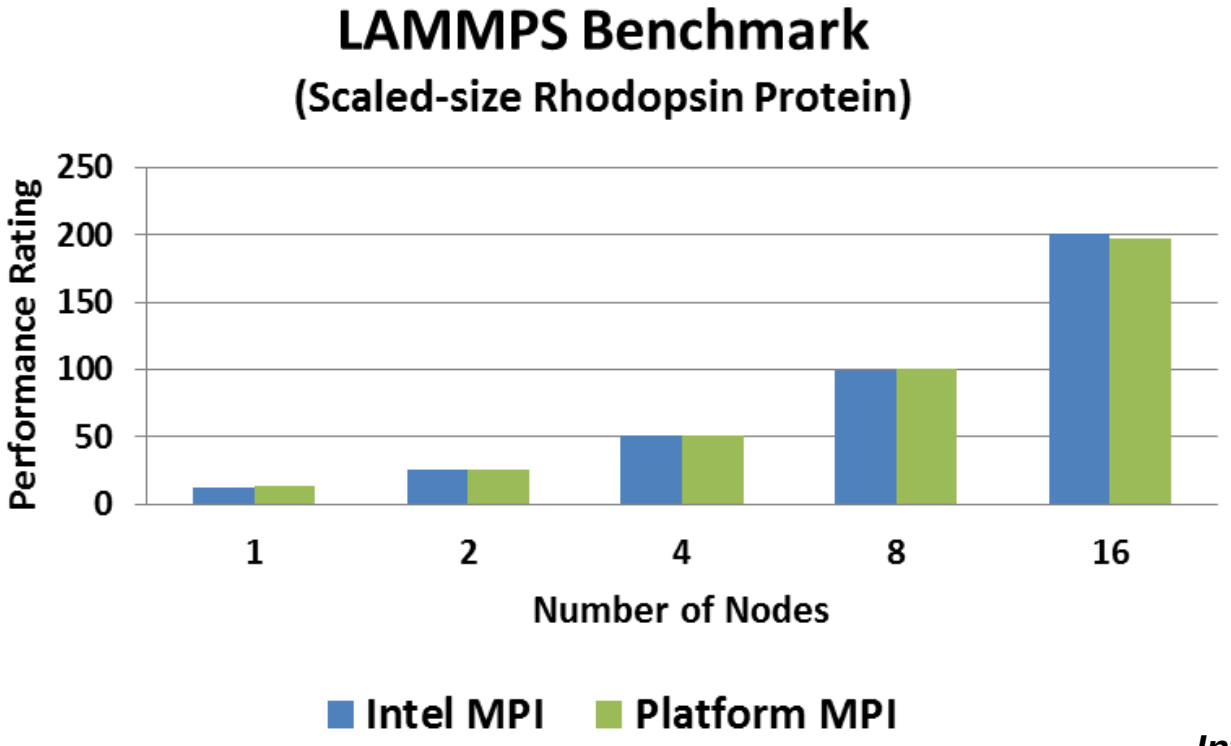


Higher is better

InfiniBand FDR

LAMMPS Performance – MPI

- **Both Intel MPI and Platform MPI perform equally good**
- **CPU binding optimization flag used in all cases shown**
  - No other optimization flags are used

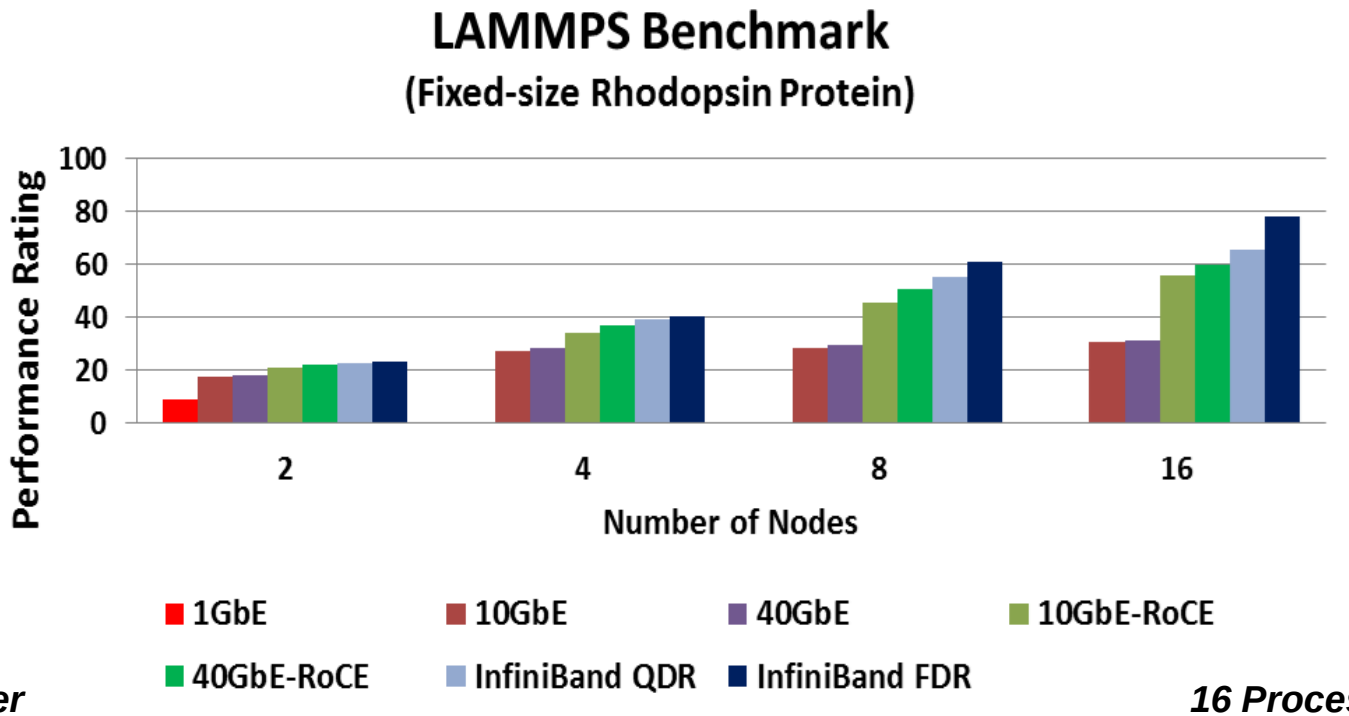


Higher is better

InfiniBand FDR

LAMMPS Performance – Network

- **InfiniBand delivers the highest performance**
  - FDR InfiniBand provides 20% higher performance than QDR InfiniBand
    - Performance gap increases with cluster size
- **Ethernet RoCE provides highest performance for Ethernet**
  - Nearly 2X performance increase versus TCP (valid for 10GbE and 40GbE)
  - 1GbE does not scale at all



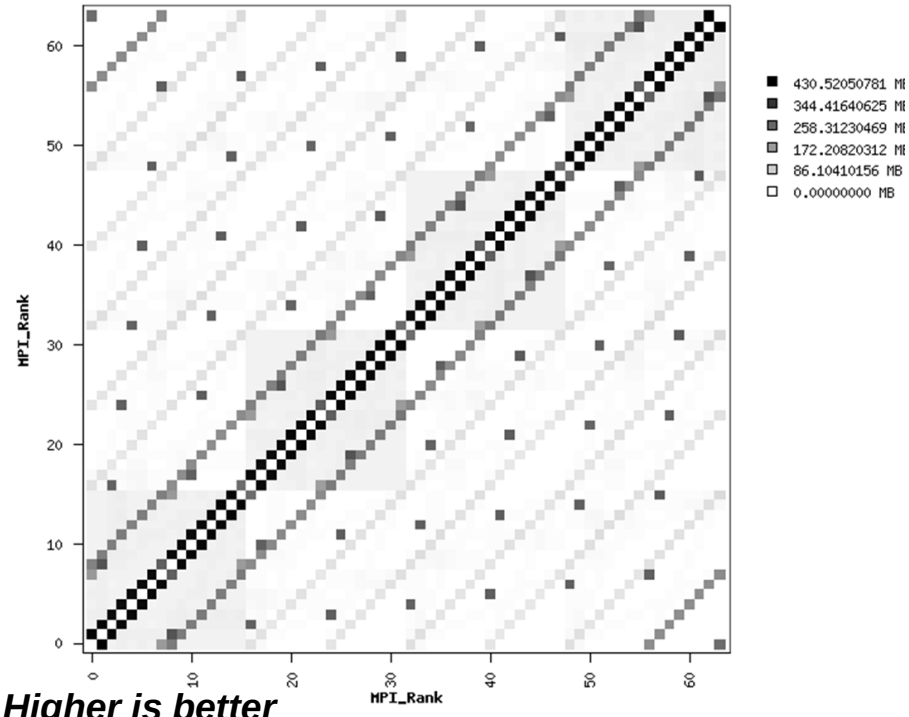
Higher is better

16 Processes/Node

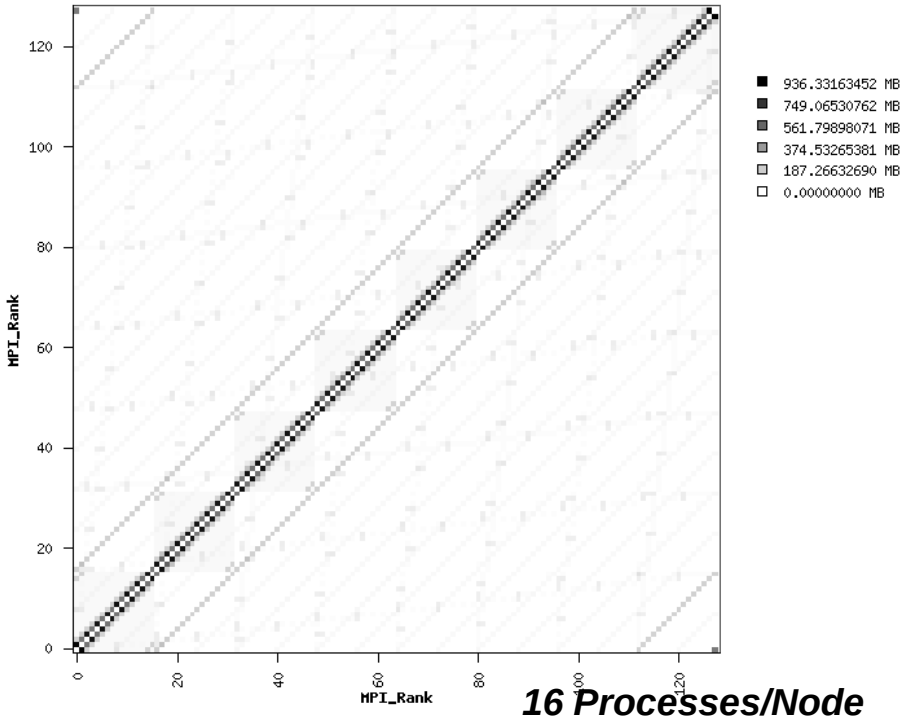
LAMMPS Profiling – Point-to-point Dataflow

- **Communications takes place between close and far processes**
  - Heavier communications between 1 rank above and below
  - Lighter communication happens to ranks further away
- **As more processes added, the scaled-size problem set grows in size**
  - The number of atoms involved grows as more processes are being added

64 MPI processes



128 MPI processes



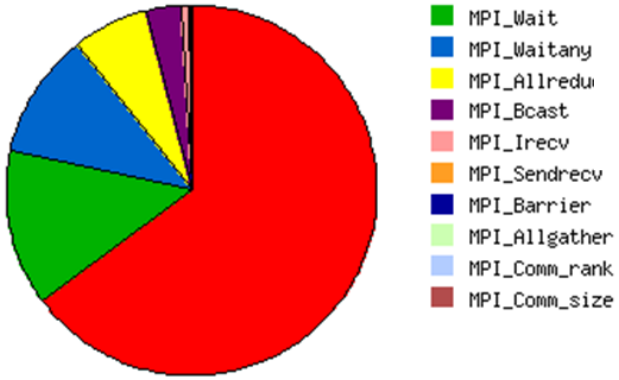
Higher is better

16 Processes/Node

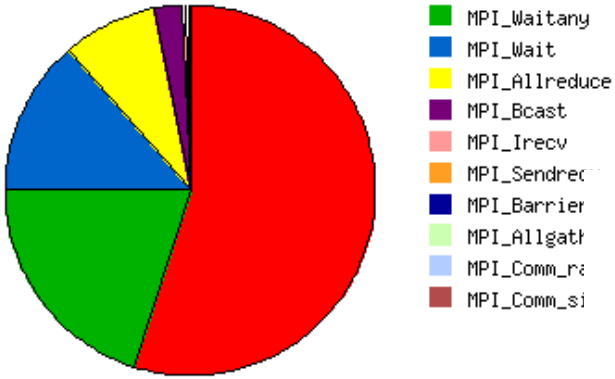
LAMMPS Profiling – MPI Comm. Time

- **Majority of MPI communication time is spent on MPI_Send**
  - MPI_Send, MPI_Waitany, MPI_Wait and MPI_Allreduce
  - Demonstrates that LAMMPS is heavy on data communications

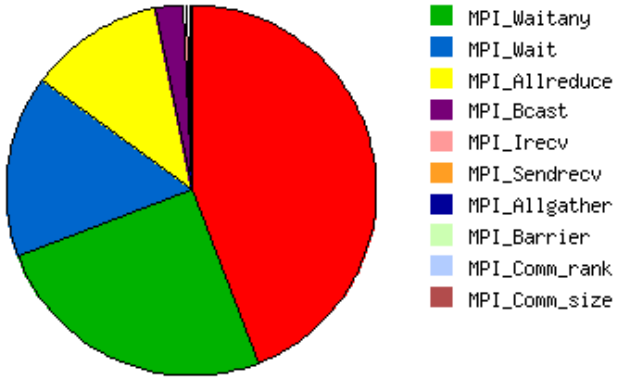
4 Nodes



8 Nodes



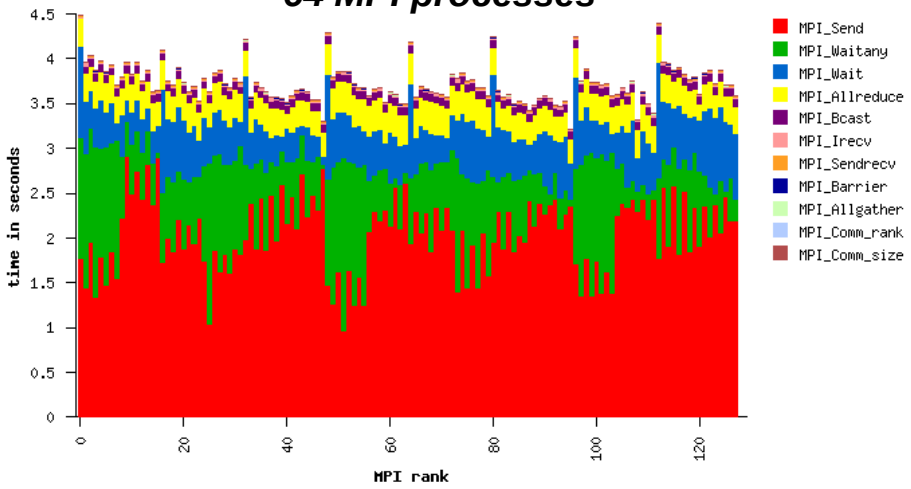
16 Nodes



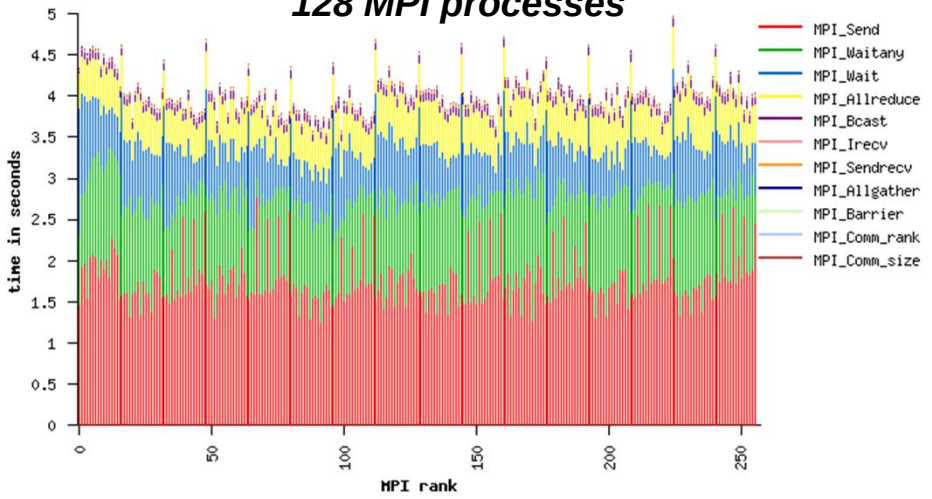
LAMMPS Profiling – MPI Data Transfer

- **Amount of data transfer grows as more processes are added to the cluster**
  - Time spikes are shown for one MPI rank out of each node

64 MPI processes

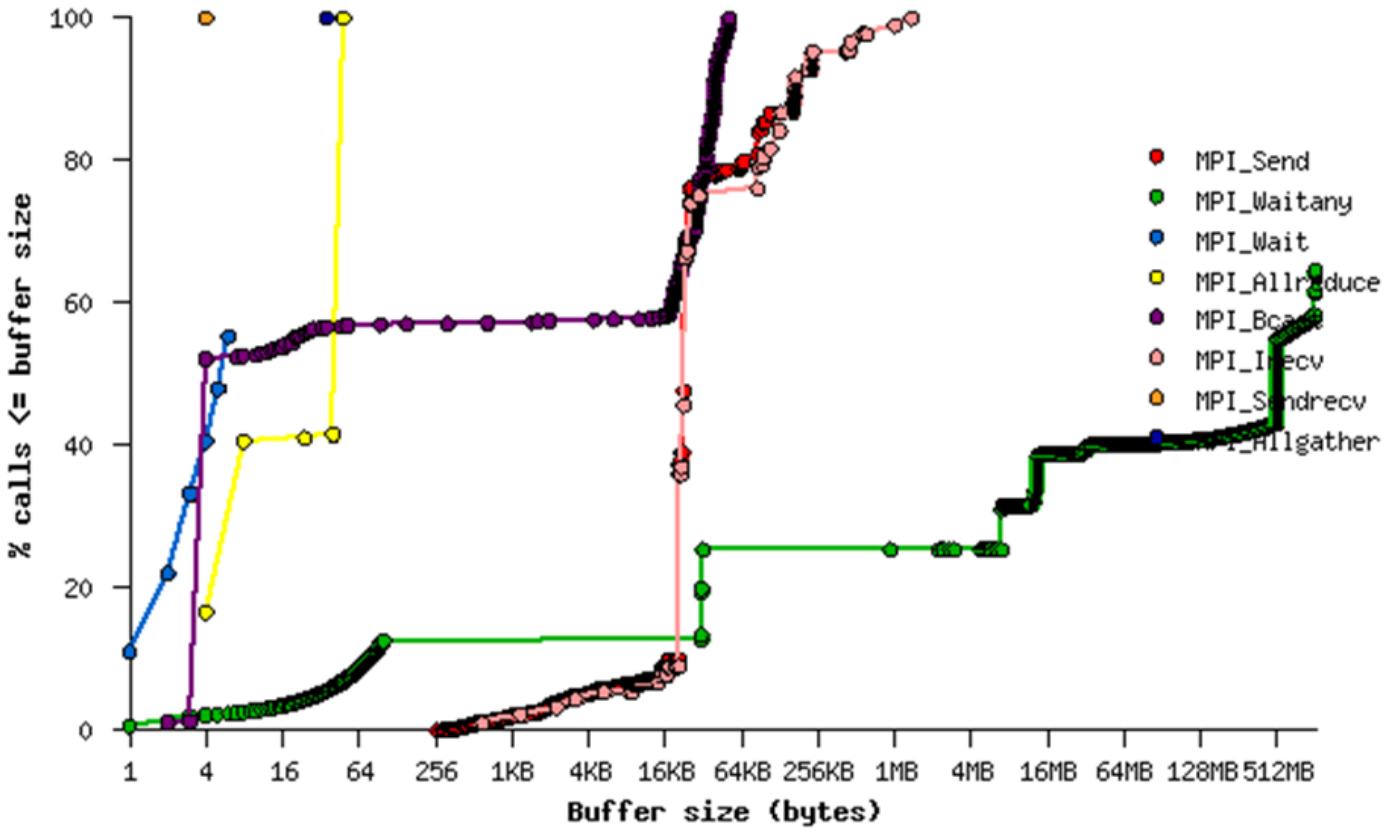


128 MPI processes



LAMMPS Profiling – MPI Message Distributions

- **Message Distribution for the percentage of calls**
  - Large percentage of MPI_Send calls are in the midrange between 16KB to 256KB



and FDR

LAMMPS – Summary

- **Performance**
  - Intel Xeon E5-2600 series and InfiniBand FDR enable LAMMPS to scale with 16 nodes
  - The E5-2680 cluster outperforms X5670 cluster by 96% for fixed-size Rhodo test
  - The E5-2680 cluster outperforms X5670 cluster by 71% on scaled-sized Rhodo test
- **Network**
  - InfiniBand (QDR or FDR) provides higher performance than Ethernet (10GbE and 40GbE)
  - FDR InfiniBand 56Gb/s delivers the highest performance for LAMMPS
    - 20% higher performance than QDR InfiniBand, and performance gap increases with cluster size
  - RoCE provides best network scalability performance for Ethernet
    - 2X performance increase (40GbE RoCE vs 40GbE TCP, 10GbE RoCE vs 10GbE TCP)
  - 1GbE would not scale beyond two nodes
- **Profiling**
  - Good network throughput is required for delivering the network bandwidth needed
  - Large percentage of MPI_Send calls are in the midrange between 16KB to 256KB

Thank You

HPC Advisory Council