HPC ADVISORY COUNCIL

HOOMD-blue
Performance Benchmark and Profiling

February 2014

DELL

HOOMD-blue

Mellanox TECHNOLOGIES

NVIDIA

Note

- **The following research was performed under the HPC Advisory Council activities**
  - Participating vendors: Dell, Mellanox, NVIDIA
  - Compute resource –The Wilkes cluster at the University of Cambridge, HPC Advisory Council Cluster Center
- **The following was done to provide best practices**
  - HOOMD-blue performance overview
  - Understanding HOOMD-blue communication patterns
  - MPI libraries comparisons
- **For more info please refer to**
  - http://www.dell.com
  - http://www.mellanox.com
  - http://www.nvidia.com
  - http://codeblue.umich.edu/hoomd-blue
  - J. A. Anderson, C. D. Lorenz, and A. Travesset. General purpose molecular dynamics simulations fully implemented on graphics processing units Journal of Computational Physics 227(10): 5342-5359, May 2008. 10.1016/j.jcp.2008.01.047

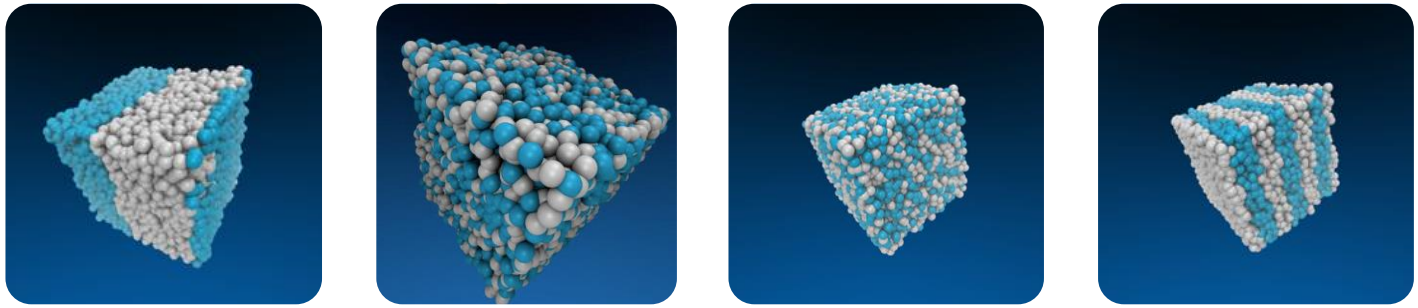
Objectives

- **The following was done to provide best practices**
  - HOOMD-blue performance benchmarking
  - Interconnect performance comparisons
  - MPI performance comparison
  - Understanding HOOMD-blue communication patterns

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

HOOMD-blue

- **Highly Optimized Object-oriented Many-particle Dynamics - Blue Edition**
- **Performs general purpose particle dynamics simulations**
- **Takes advantage of NVIDIA GPUs**
- **Free, open source**
- **Simulations are configured and run using simple python scripts**
- **The development effort is led by Glotzer group at University of Michigan**
  - Many groups from different universities have contributed code to HOOMD-blue

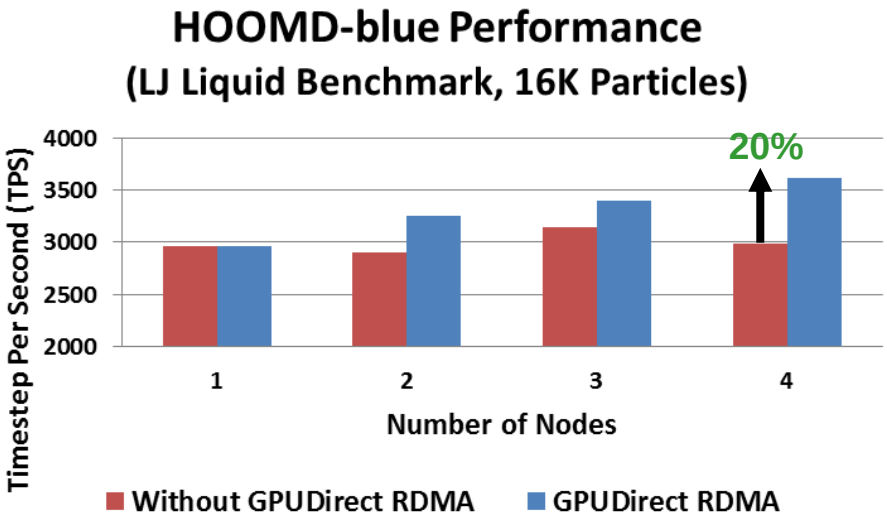


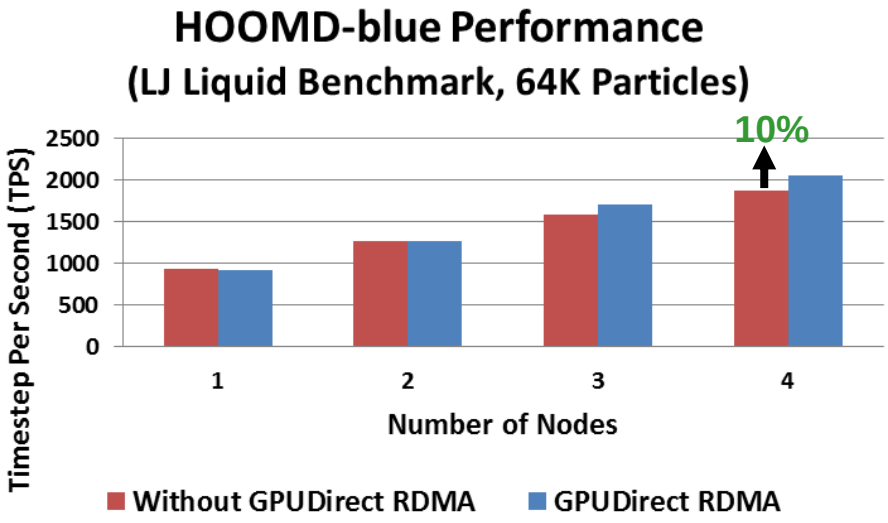
Test Cluster Configuration 1

- **Dell™ PowerEdge™ R720xd/R720 cluster**
  - Dual-Socket Octa-core Intel E5-2680 V2 @ 2.80 GHz CPUs (Static max Perf in BIOS)
  - Memory: 64GB DDR3 1600 MHz Dual Rank Memory Module
  - Hard Drives: 24x 250GB 7.2 RPM SATA 2.5” on RAID 0
  - OS: RHEL 6.2, MLNX_OFED 2.1-1.0.0 InfiniBand SW stack
- **Mellanox Connect-IB FDR InfiniBand**
- **Mellanox SwitchX SX6036 InfiniBand VPI switch**
- **NVIDIA® Tesla K40 GPUs (1 GPU per node)**
- **NVIDIA® CUDA® 5.5 Development Tools and Display Driver 331.20**
- **GPUDirect RDMA (nvidia_peer_memory-1.0-0.tar.gz)**
- **MPI: Open MPI 1.7.4rc1**
- **Application: HOOMD-blue (git master 28Jan14)**
- **Benchmark datasets: Lennard-Jones Liquid Benchmarks (16K, 64K Particles)**

HOOMD-blue Performance – GPUDirect RDMA

- **GPUDirect RDMA enables higher performance on a small GPU cluster**
  - Demonstrated up to 20% of higher performance at 4 nodes for 16K particles
  - Showed up to 10% of performance gain at 4 nodes for 64K particles
- **Adjusting OMPI MCA param can maximize GPUDirect RDMA usage**
  - Based on MPI profiling, limits for GDR for 64K particles was tuned to 65KB
- **MCA Parameter to enable and tune GPUDirect RDMA for Open MPI:**
  - -mca btl_openib_want_cuda_gdr 1 -mca btl_openib_cuda_rdma_limit XXXX





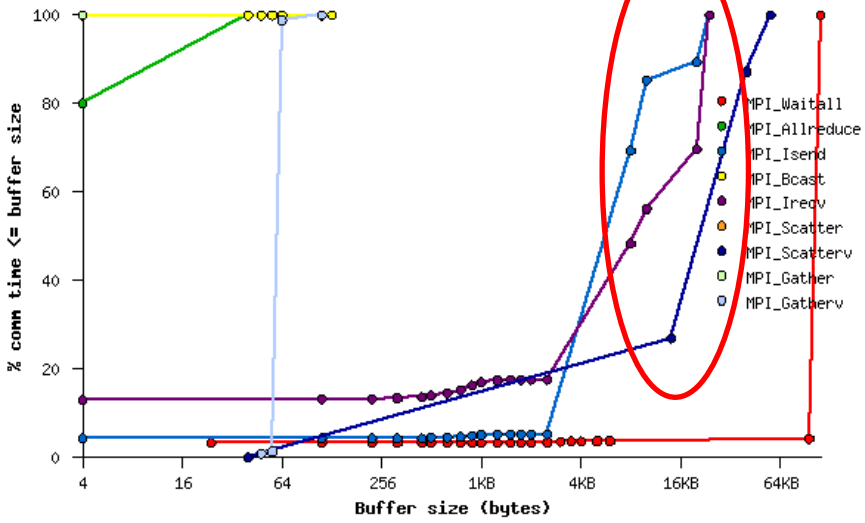
Higher is better

Open MPI

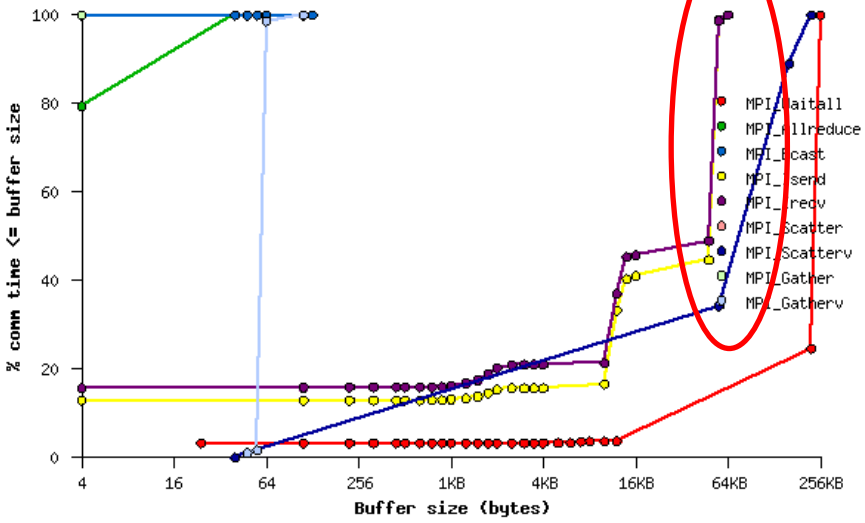
HOOMD-blue Profiling – MPI Message Sizes

- **HOOMD-blue utilizes non-blocking and collectives for most data transfers**
  - 16K particles: MPI_Isend/MPI_Irecv are concentrated between 4B to 24576B
  - 64K particles: MPI_Isend/MPI_Irecv are concentrated between 4B to 65536B
- **MCA parameter used to enable and tune for GPUDirect RDMA**
  - 16K particles: Default would allow all send/recv to use GPUDirect RDMA
  - 64K particles: Maximize GDR by tuning MCA param to include up to 65KB
    - -mca btl_openib_cuda_rdma_limit 65537 (Change for 64K particles case)

4 Nodes – 16K Particles



4 Nodes – 64K Particles



1 MPI Process/Node

Test Cluster Configuration 2

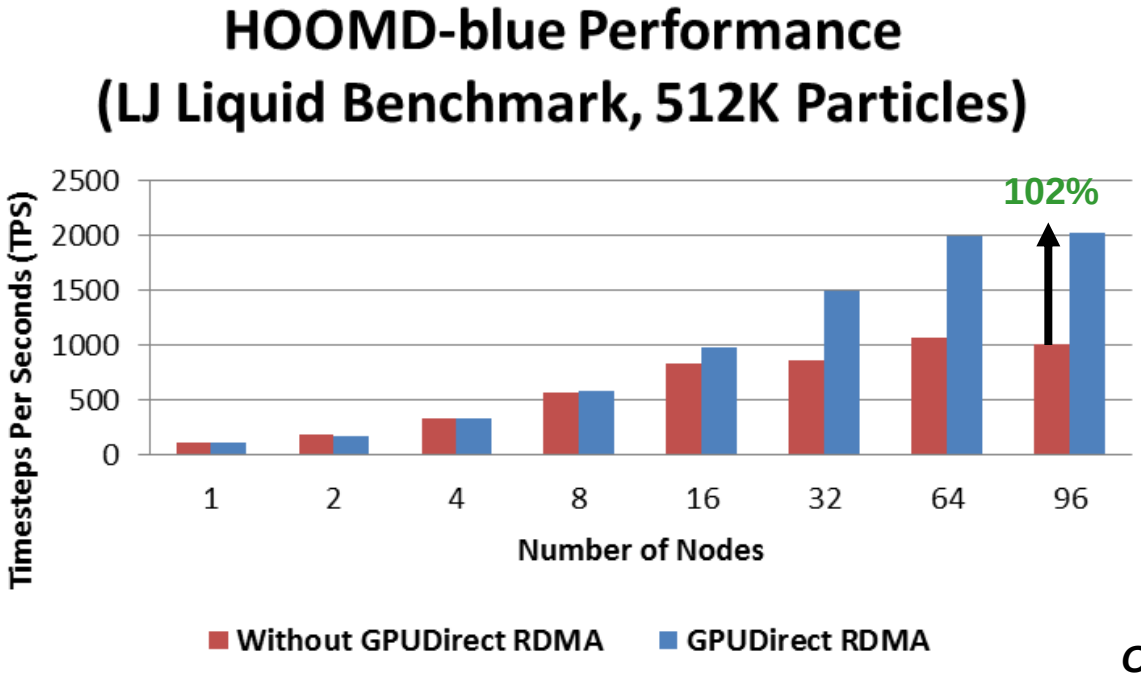
- **Dell™ PowerEdge™ T620 128-node (1536-core) Wilkes cluster at Univ of Cambridge**
  - Dual-Socket Hexa-Core Intel E5-2630 v2 @ 2.60 GHz CPUs
  - Memory: 64GB memory, DDR3 1600 MHz
  - OS: Scientific Linux release 6.4 (Carbon), MLNX_OFED 2.1-1.0.0 InfiniBand SW stack
  - Hard Drives: 2x 500GB 7.2 RPM 64MB Cache SATA 3.0Gb/s 3.5”
- **Mellanox Connect-IB FDR InfiniBand adapters**
- **Mellanox SwitchX SX6036 InfiniBand VPI switch**
- **NVIDIA® Tesla K20 GPUs (2 GPUs per node)**
- **NVIDIA® CUDA® 5.5 Development Tools and Display Driver 331.20**
- **GPUDirect RDMA (nvidia_peer_memory-1.0-0.tar.gz)**
- **MPI: Open MPI 1.7.4rc1, MVAPICH2-GDR 2.0b**
- **Application: HOOMD-blue (git master 28Jan14)**
- **Benchmark datasets: Lennard-Jones Liquid Benchmarks (256K and 512K Particles)**

The Wilkes Cluster at University of Cambridge

- **The University of Cambridge in partnership with Dell, NVIDIA and Mellanox**
  - The UK's fastest academic cluster, deployed November 2013
- **Produces a LINPACK performance of 240TF**
  - on the Top500 position of 166 in the November 2013 list
- **Ranked most energy efficient air cooled supercomputer in the world**
- **Ranked second in the worldwide Green500 ranking**
  - Extremely high performance per watt of 3631 MFLOP/W
- **Architected to utilize the NVIDIA RDMA communication acceleration**
  - Significantly increase the system's parallel efficiency

HOOMD-blue Performance – GPUDirect RDMA

- **GPUDirect RDMA unlocks performance between GPU and IB**
  - Demonstrated up to 102% of higher performance at 96 nodes
- **GPUDirect RDMA provides a direct P2P data path between GPU and IB**
  - This new technology significantly lowers GPU-GPU communication latency
  - Completely offload CPU from all GPU communications across the network
- **MCA param to enable GPUDirect RDMA between 1 GPU and IB per node**
  - --mca btl_openib_want_cuda_gdr 1 (Default value for btl_openib_cuda_rdma_limit)

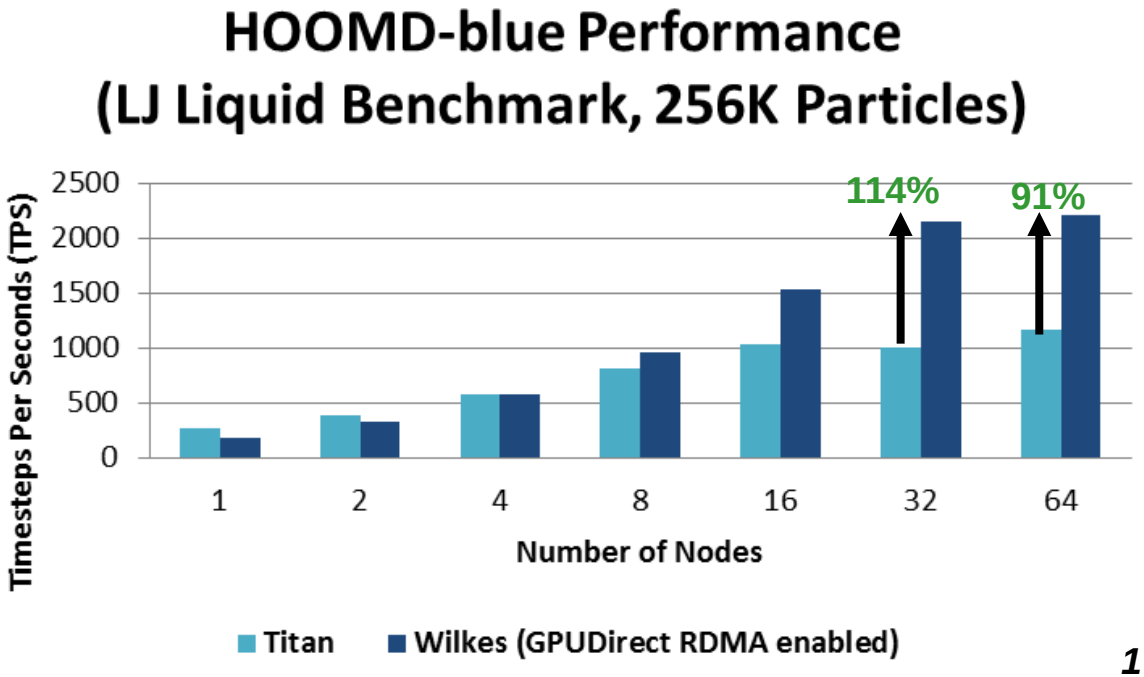


Higher is better

Open MPI

HOOMD-blue Performance – Scalability

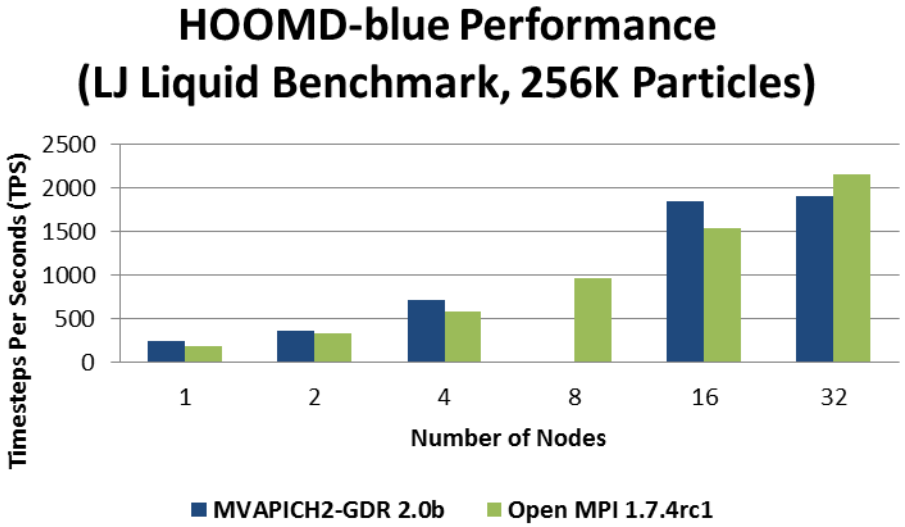
- **FDR InfiniBand empowers Wilkes to surpass Titan on scalability**
  - Titan showed higher per-node performance but Wilkes outperformed in scalability
  - Titan: K20x GPUs which computes at higher clock rate than the K20 GPU
  - Wilkes: K20 GPUs at PCIe Gen2, and FDR InfiniBand at Gen3 rate
- **Wilkes exceeds Titan in scalability performance with FDR InfiniBand**
  - Outperformed Titan by up to 114% at 32 nodes

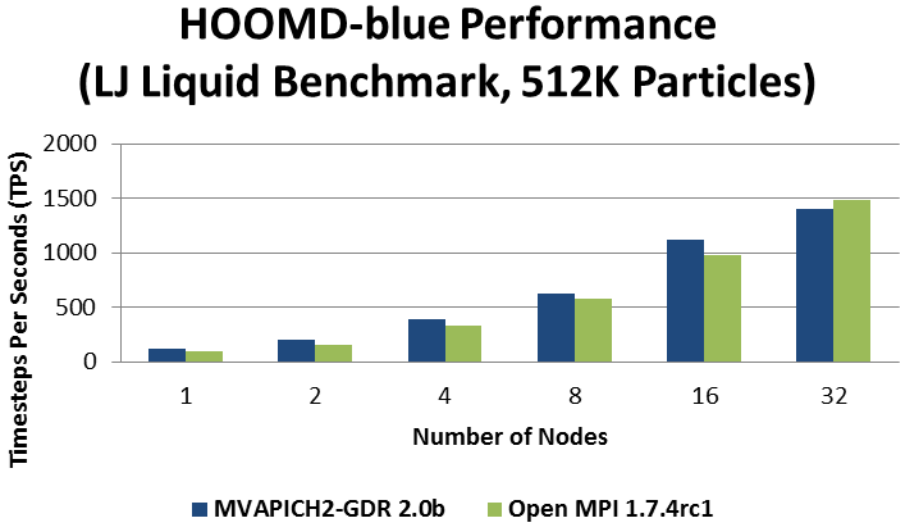


1 Process/Node

HOOMD-blue Performance – MPI

- **Open MPI performs better than MVAPICH2-GDR**
  - At lower scale, MVAPICH2-GDR performs slight better
  - At higher scale, Open MPI shows better scalability
  - Locality of IB interface used was explicitly specified with flags when tests were run
- **Both MPI implementations are in their beta releases**
  - Scalability performance are expected to improve on their release versions
  - An Issue prevented MVAPICH2-GDR from running for 8 and 64 nodes

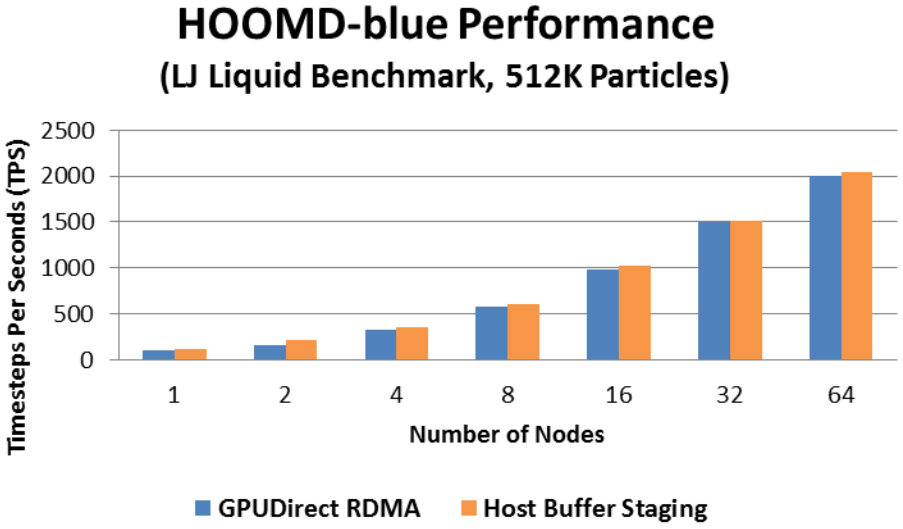


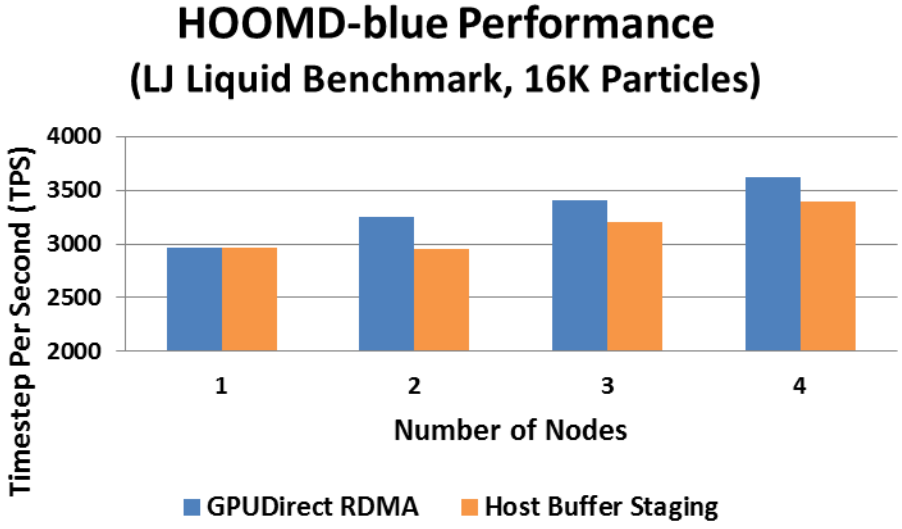


Higher is better ***1 Process/Node***

HOOMD-blue Performance-Host-buffer Staging

- **HOOMD-blue can run w/ non-CUDA aware MPI using Host Buffer Staging**
  - HOOMD-blue is built using “ENABLE_MPI=ON" and "ENABLE_MPI_CUDA=OFF” flags
  - Non-CUDA aware (or host) MPI has lower latency than CUDA aware MPI
  - With GDR: CUDA-aware MPI is copied Individually. Slightly higher latency with MPI
  - With HBS: Only single large buffers are copied as needed. Lower latency using MPI
- **GDR performs on par with HBS on large scale, better in some cases**
  - On large scale, HBS performance appears to perform slightly faster than GDR
  - On small scale, GDR can be faster than HBS when small number of particles per GPU





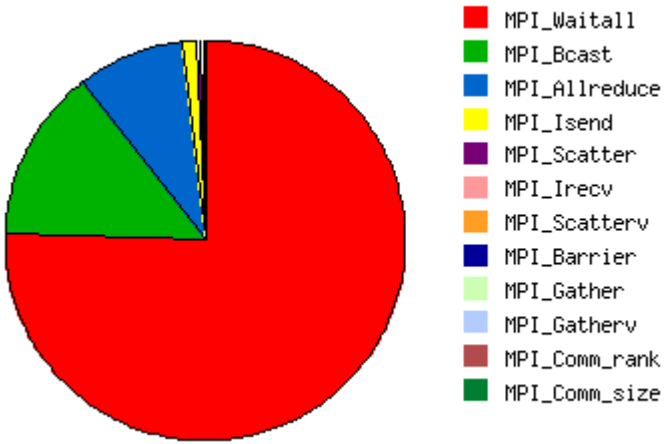
Higher is better

1 Process/Node

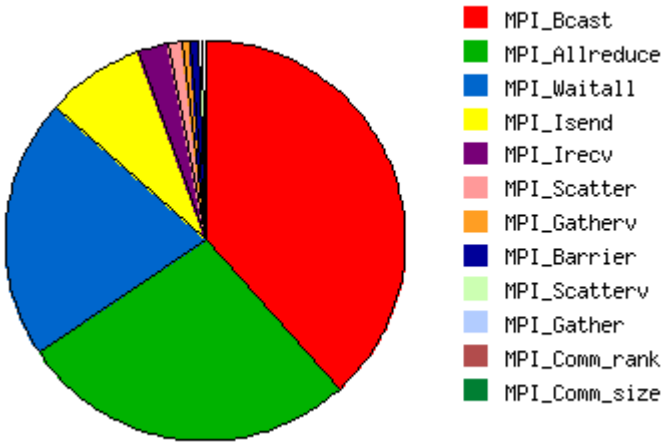
HOOMD-blue Profiling – % Time Spent on MPI

- **HOOMD-blue utilizes both non-blocking and collective ops for comm**
  - Changes in network communications take place as cluster scales
  - 4 nodes: MPI_Waitall(75%), the rest are MPI_Bcast and MPI_Allreduce
  - 96 nodes: MPI_Bcast (35%), the rest are MPI_Allreduce, MPI_Waitall

4 Nodes – 512K Particles



96 Nodes – 512K Particles

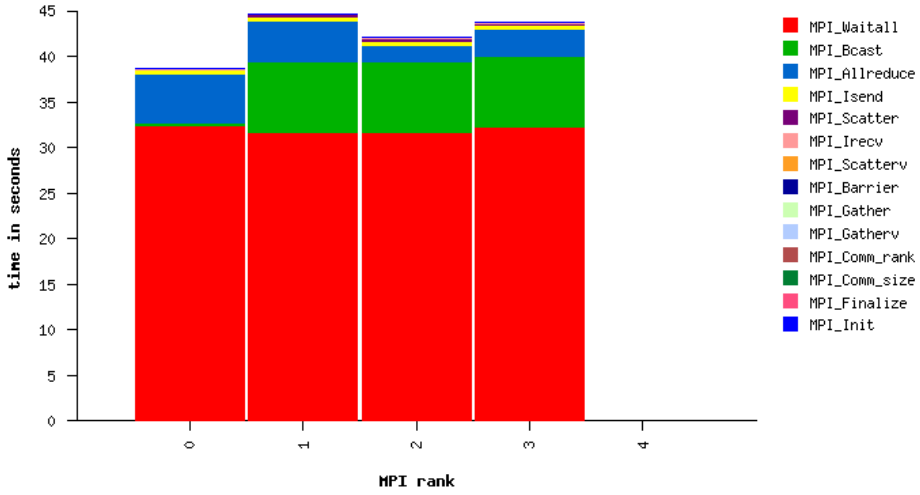


Open MPI

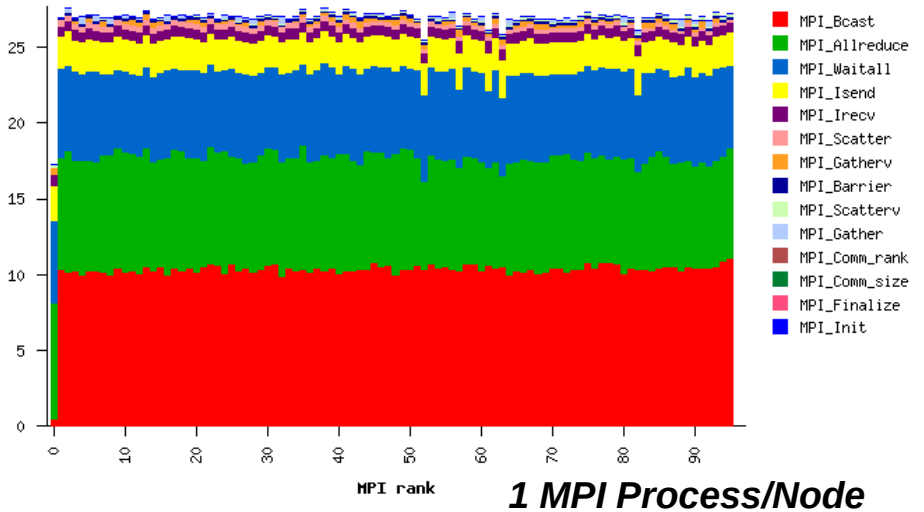
HOOMD-blue Profiling – MPI Communication

- **Each rank engages in similar network communication**
  - Except for rank 0, which spends less time in MPI_Bcast

4 Nodes – 512K Particles



96 Nodes – 512K Particles

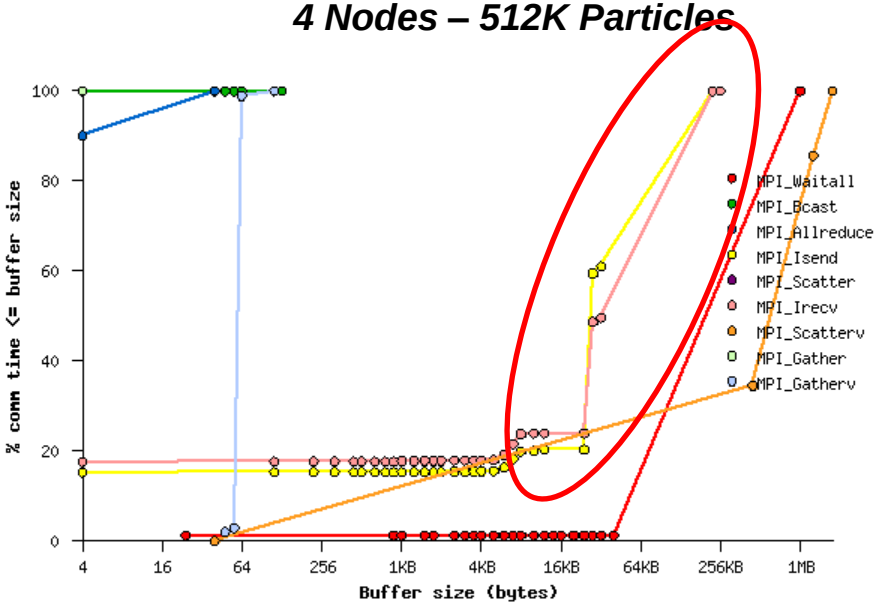


1 MPI Process/Node

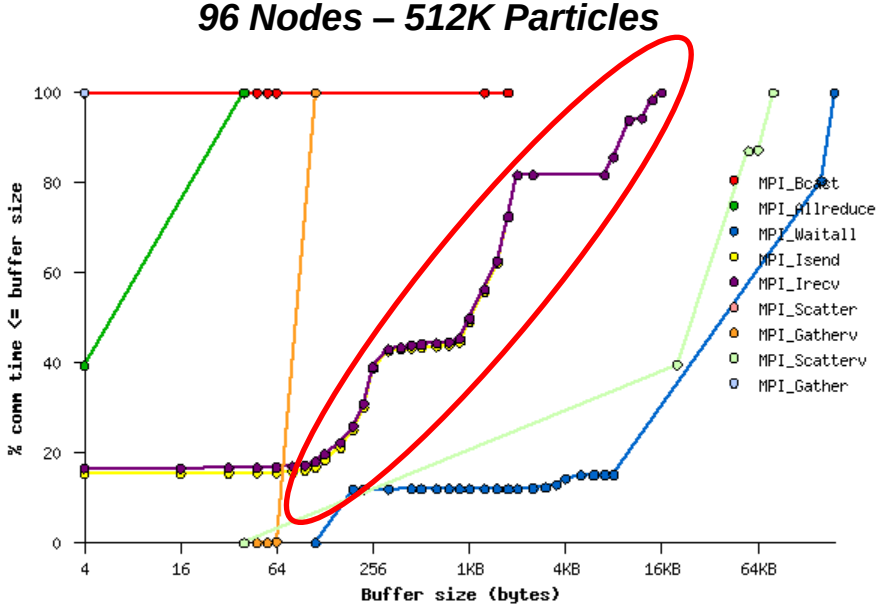
HOOMD-blue Profiling – MPI Message Sizes

- **HOOMD-blue utilizes non-blocking and collectives for most data transfers**
  - 4 Nodes: MPI_Isend/MPI_Irecv are concentrated between 28KB to 229KB
  - 96 Nodes: MPI_Isend/MPI_Irecv are concentrated between 64B to 16KB
- **GPUDirect RDMA is enabled for messages between 0B to 30KB**
  - MPI_Isend/_Irecv messages are able to take advantage of GPUDirect RDMA
  - Messages fitted within the (tunable default of) 30KB window can be benefited

4 Nodes – 512K Particles



96 Nodes – 512K Particles

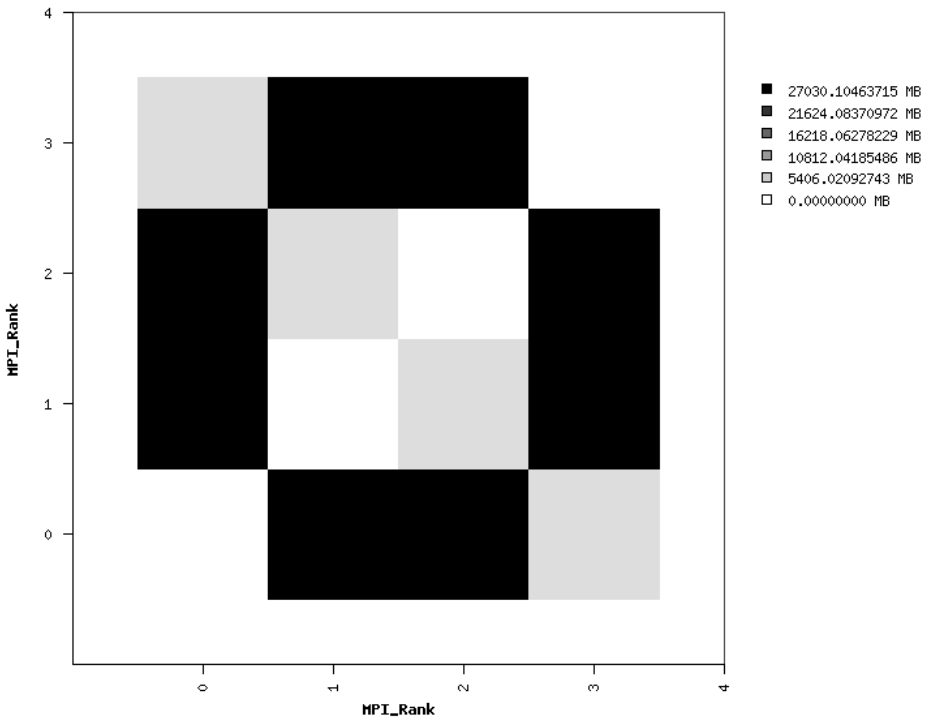


1 MPI Process/Node

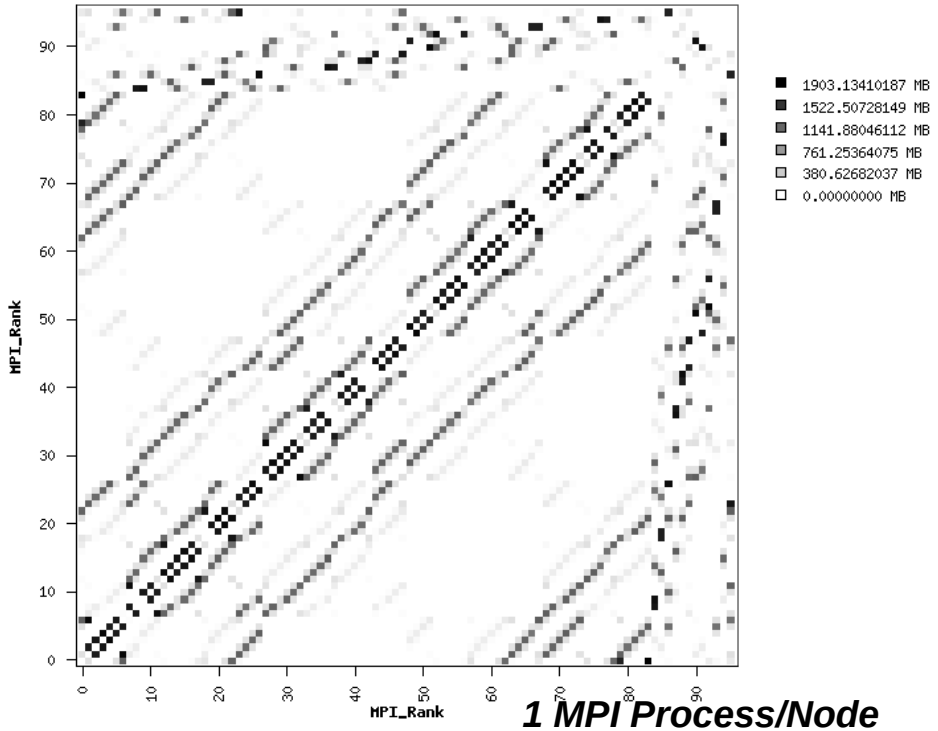
HOOMD-blue Profiling – Point to Point Transfer

- **Distribution of data transfers between the MPI processes**
  - Non-blocking point-to-point data communications between processes are involved
  - Less data is being transferred as more ranks being part of the run

4 Nodes – 512K Particles



96 Nodes – 512K Particles



1 MPI Process/Node

HOOMD-blue – Summary

- **HOOMD-blue demonstrates good use of GPU and InfiniBand at scale**
  - FDR InfiniBand is the interconnect allows HOOMD-blue to scale
  - Ethernet solutions would not scale beyond 1 node
- **GPUDirect RDMA**
  - This new technology provides a direct P2P data path between GPU and IB
  - This provides a significant decrease in GPU-GPU communication latency
- **GPUDirect RDMA unlocks performance between GPU and IB**
  - Demonstrated up to 20% of higher performance at 4 nodes for 16K case
  - Demonstrated up to 102% of higher performance at 96 nodes for 512K case
- **InfiniBand empowers Wilkes to surpass Titan on scalability performance**
  - Titan has higher per-node performance but Wilkes outperforms in scalability
  - Outperforms Titan by 114% at 32 nodes
- **GPUDirect RDMA performs on par with Host Buffer Staging**
  - On large scale, HBS performance appears to perform slightly faster than GDR
  - On small scale, GDR can be faster than HBS when small num. of particles per GPU

Thank You

HPC Advisory Council