S5169 - Maximizing Scalability Performance in HOOMD-blue by Exploiting GPUDirect® RDMA on Green500 Supercomputer

Pak Lui

GPU Technology Conference 2015

GPU TECHNOLOGY CONFERENCE

Mellanox TECHNOLOGIES

Connect. Accelerate. Outperform.™

Technology Roadmap – One-Generation Lead over the Competition

Mellanox TECHNOLOGIES

Mellanox → **20Gbs** → **40Gbs** → **56Gbs** → **100Gbs** → **200Gbs** →

Terascale

Petascale

Exascale

3rd

TOP500 2003
Virginia Tech (Apple)

1st

“Roadrunner”
Mellanox Connected

OAK RIDGE National Laboratory
“Summit” System

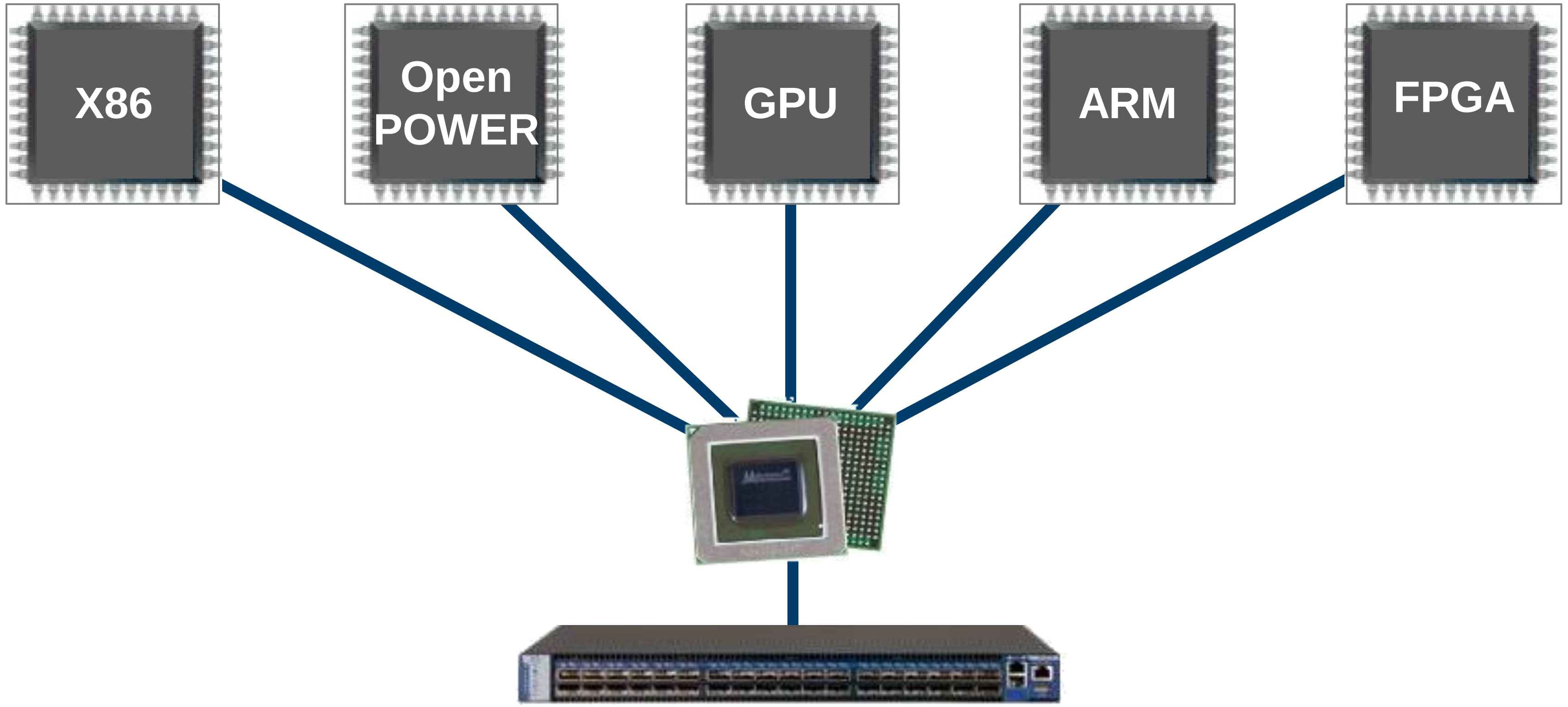
Lawrence Livermore National Laboratory
“Sierra” System

2000 **2005** **2010** **2015** **2020**

End-to-End Interconnect Solutions for All Platforms

Highest Performance and Scalability for

X86, Power, GPU, ARM and FPGA-based Compute and Storage Platforms



Smart Interconnect to Unleash The Power of All Compute Architectures

The Future is Here

Entering the Era of 100Gb/s

Adapters

ConnectX®-4

100Gb/s Adapter, 0.7us latency

150 million messages per second

(10 / 25 / 40 / 50 / 56 / 100Gb/s)

Switch

SwitchIB™

36 EDR (100Gb/s) Ports, <90ns Latency

Throughput of 7.2Tb/s

Interconnect

LinkX™

Copper (Passive, Active)

Optical Cables (VCSEL)

Silicon Photonics

Enter the World of Scalable Performance – 100Gb/s Adapter

ConnectX-4: Highest Performance Adapter in the Market

InfiniBand: SDR / DDR / QDR / FDR / EDR

Ethernet: 10 / 25 / 40 / 50 / 56 / 100GbE

100Gb/s, <0.7us latency

150 million messages per second

OpenPOWER CAPI technology

CORE-Direct technology

GPUDirect RDMA

Dynamically Connected Transport (DCT)

Ethernet offloads (HDS, RSS, TSS, LRO, LSOv2)

Shattering The World of Interconnect Performance!

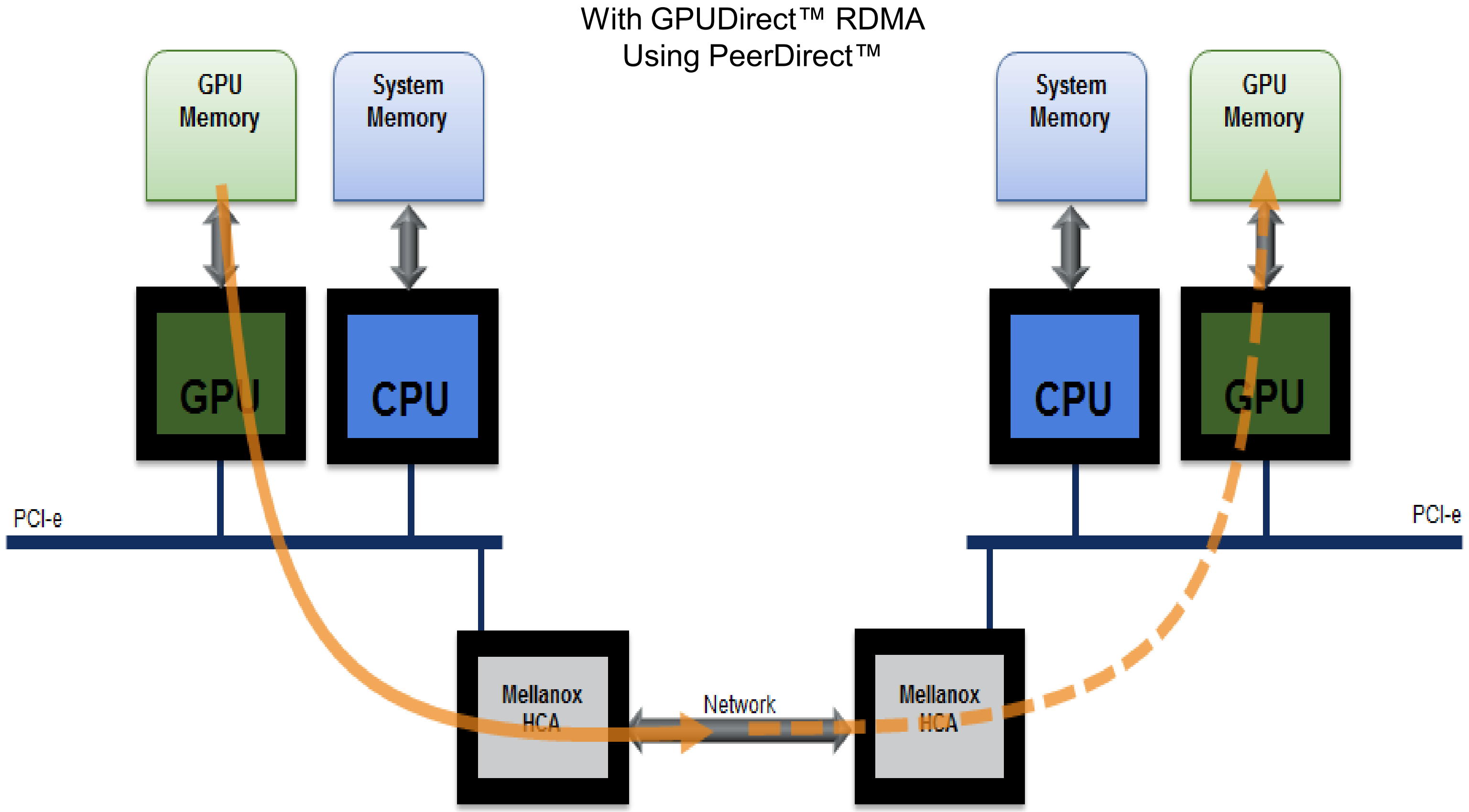
ConnectX-4 EDR 100G InfiniBand	
InfiniBand Throughput	**100 Gb/s**
InfiniBand Bi-Directional Throughput	**195 Gb/s**
InfiniBand Latency	**0.61 us**
InfiniBand Message Rate	**149.5 Million/sec**
HPC-X MPI Bi-Directional Throughput	**193.1 Gb/s**

***First results, optimizations in progress**

GPUDirect RDMA

Introduction

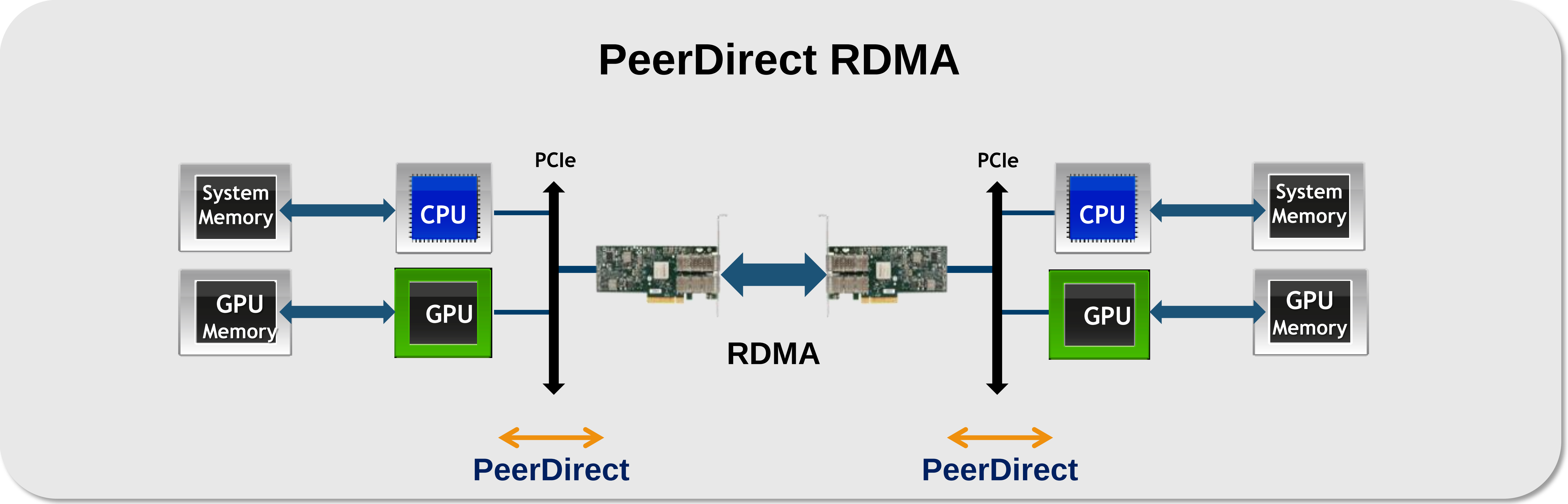
GPUDirect™ RDMA (GPUDirect 3.0)



- Eliminates CPU bandwidth and latency bottlenecks
- Uses remote direct memory access (RDMA) transfers between GPUs
- Resulting in significantly improved MPI SendRecv efficiency between GPUs in remote nodes
- Based on PeerDirect technology

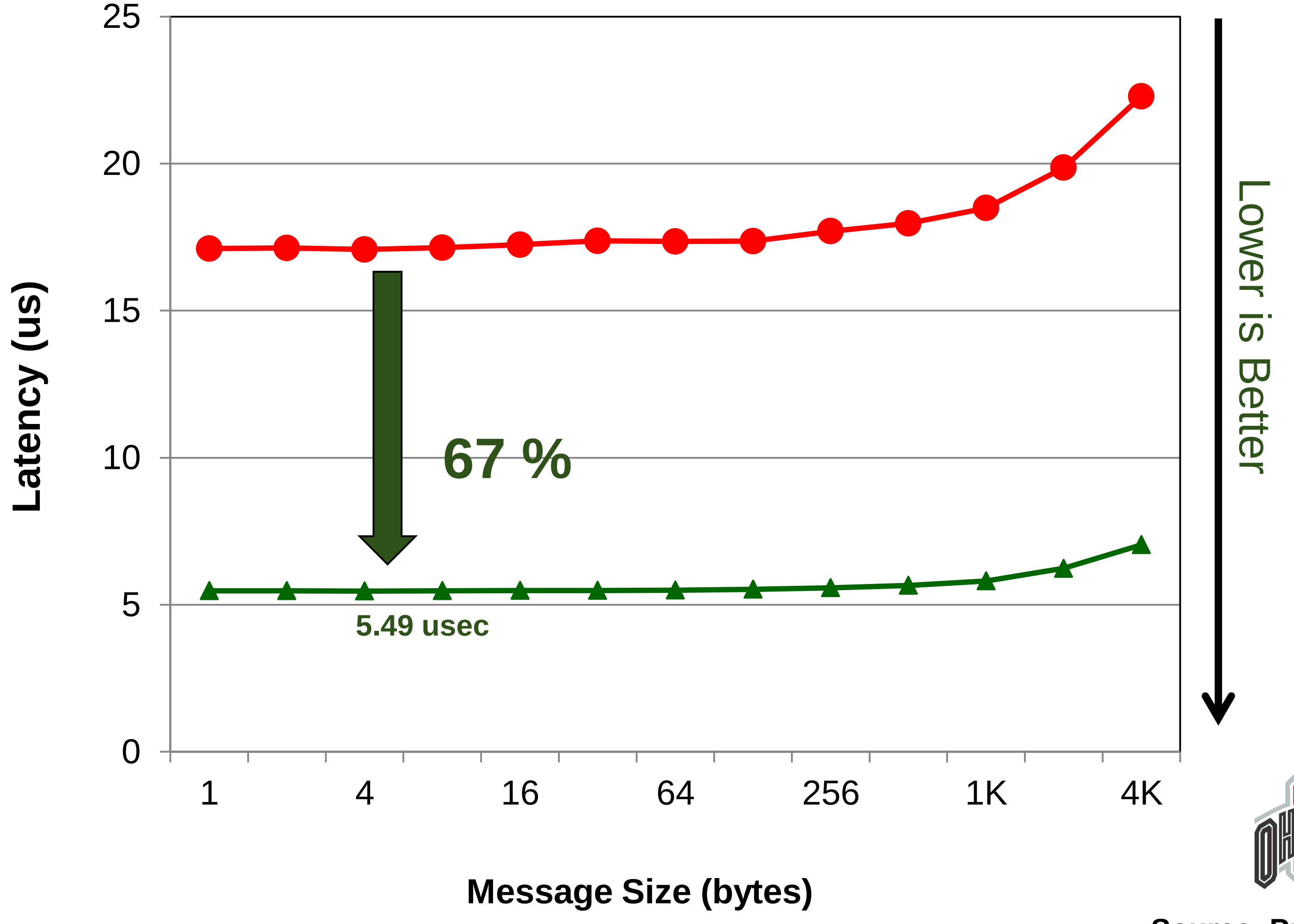
PeerDirect Technology

- Based on Peer-to-Peer capability of PCIe
- Support for any PCIe peer which can provide access to its memory
  - NVIDIA GPU, XEON PHI, AMD, custom FPGA

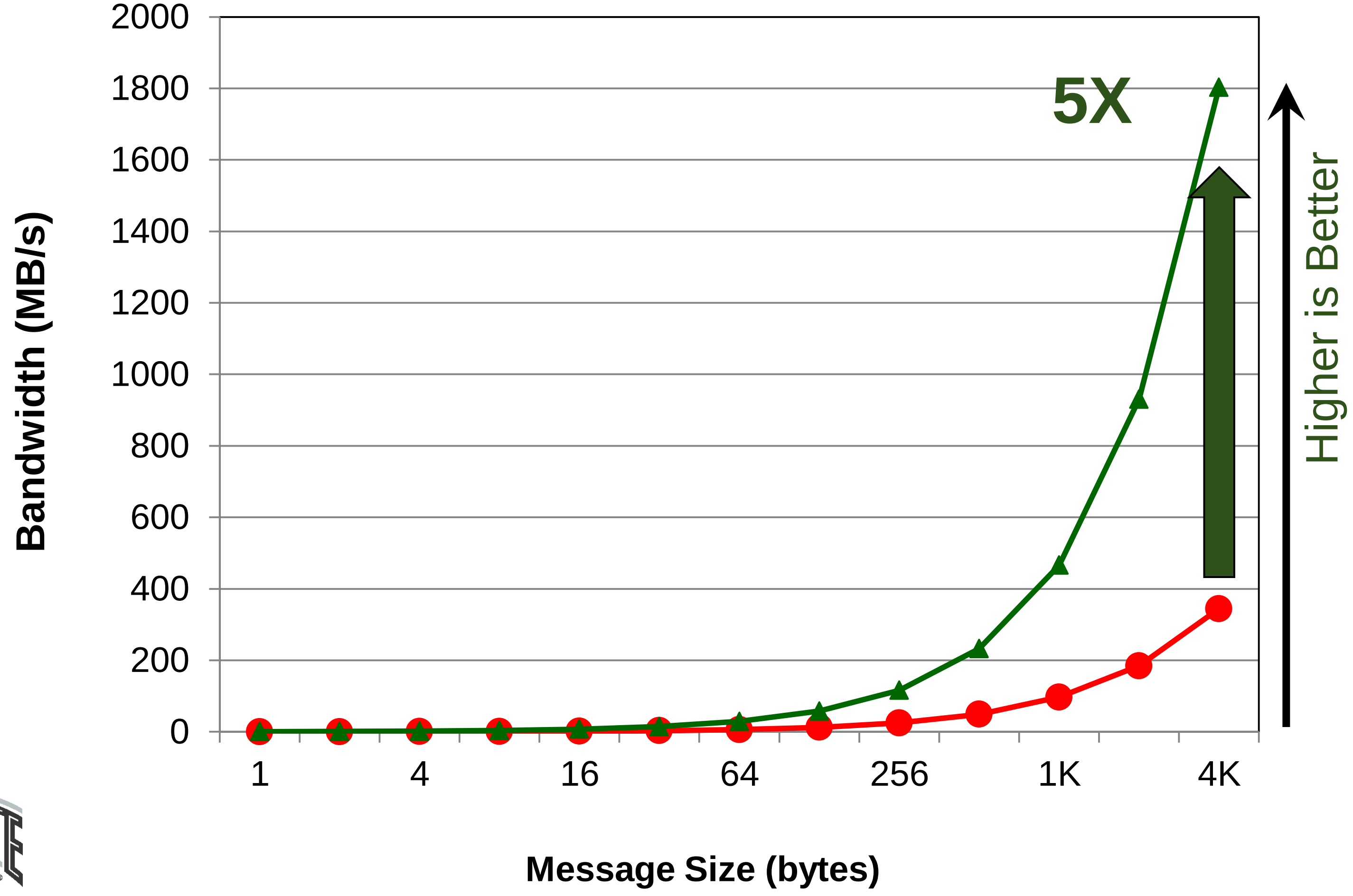


Performance of MVAPICH2 with GPUDirect RDMA

GPU-GPU Internode MPI Latency



GPU-GPU Internode MPI Bandwidth



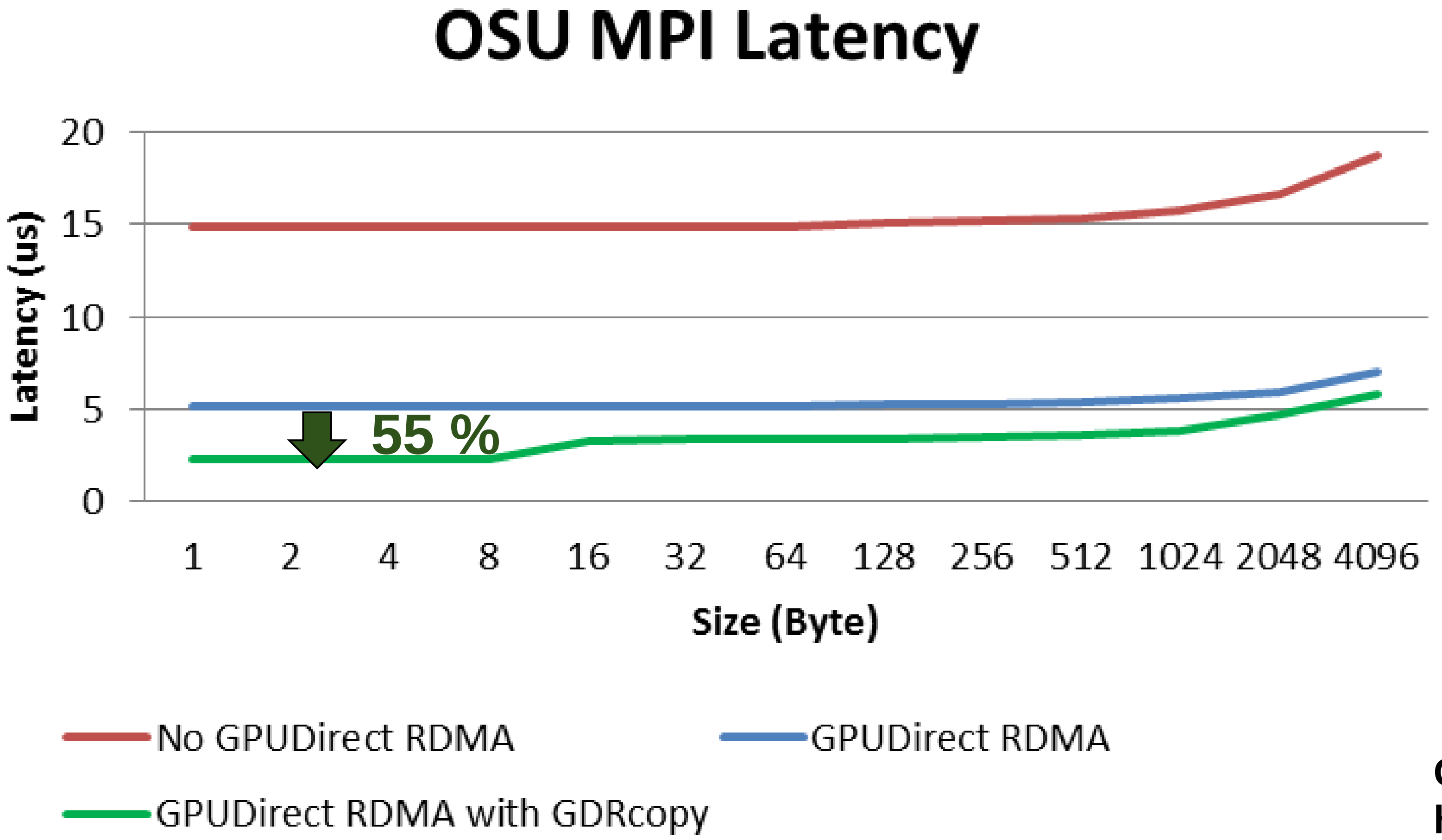
Source: Prof. DK Panda

67% Lower Latency

5X Increase in Throughput

Performance of MVAPICH2 with GPUDirect RDMA + gdrcopy

- gdrcopy: A low-latency GPU memory copy library based on GPUDirect RDMA technology
  - Offers the infrastructure to create user-space mappings of GPU memory
  - Demonstrated further latency reduction by 55%
- S5461 - Latest Advances in MVAPICH2 MPI Library for NVIDIA GPU Clusters with InfiniBand

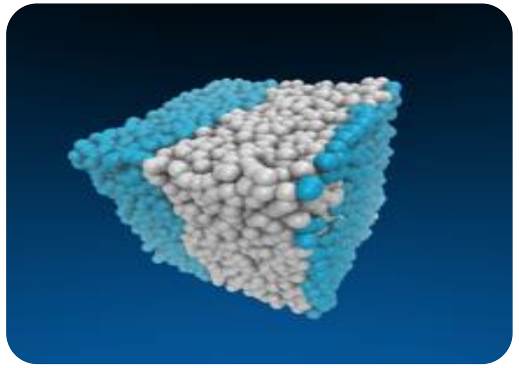


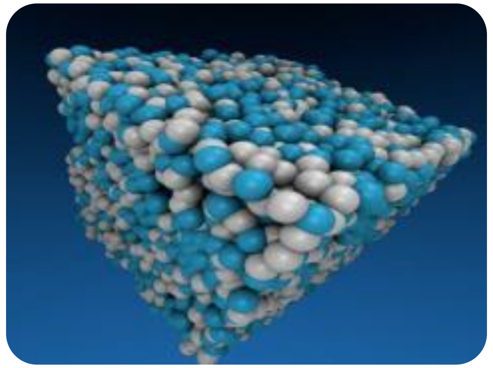
Lower is Better

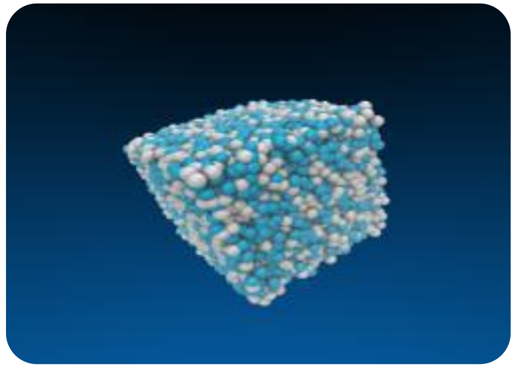
GDRcopy: https://github.com/NVIDIA/gdrcopy

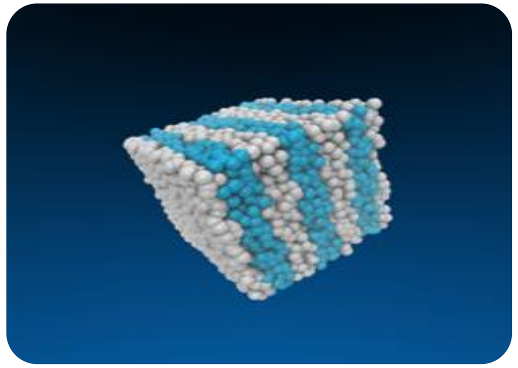
HOOMD-blue

- 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

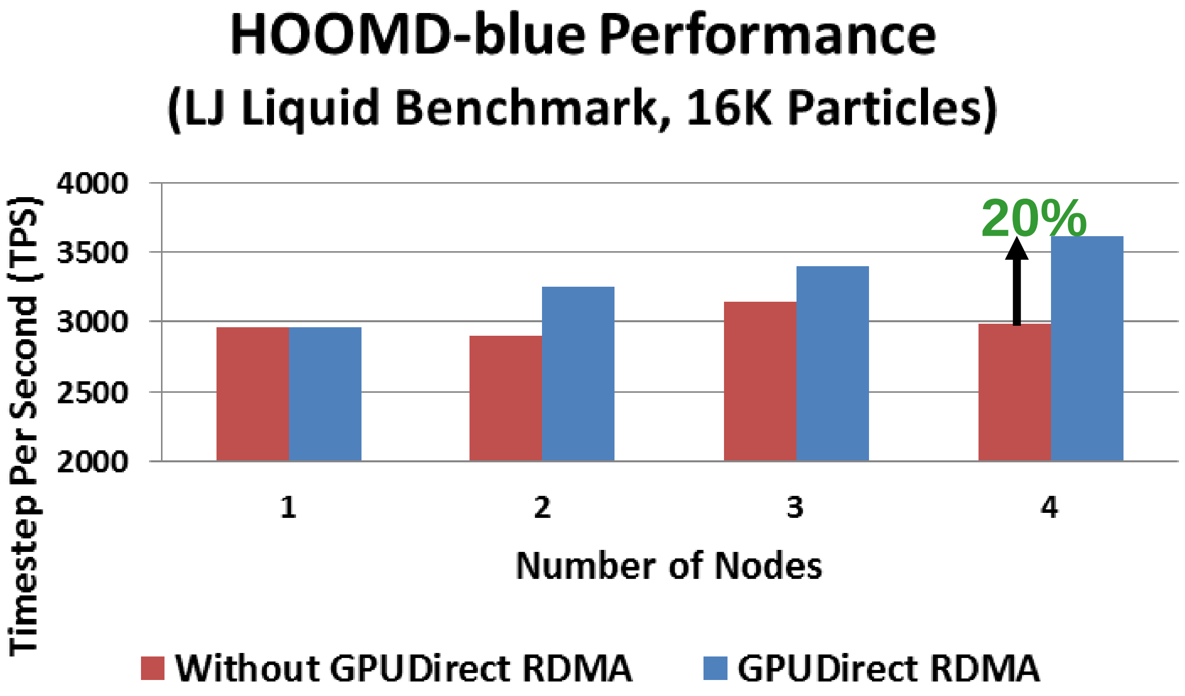
Jupiter Cluster
HPC Advisory Council

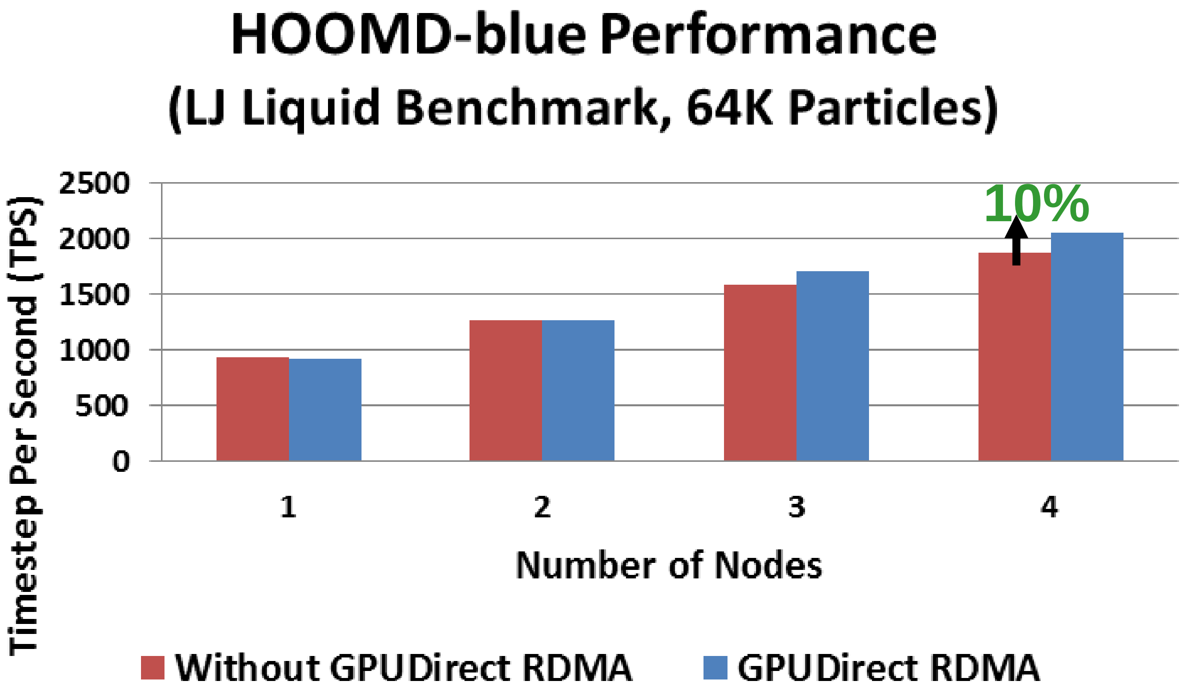
Test Cluster Configuration 1

- Dell™ PowerEdge™ R720xd/R720 “Jupiter” 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





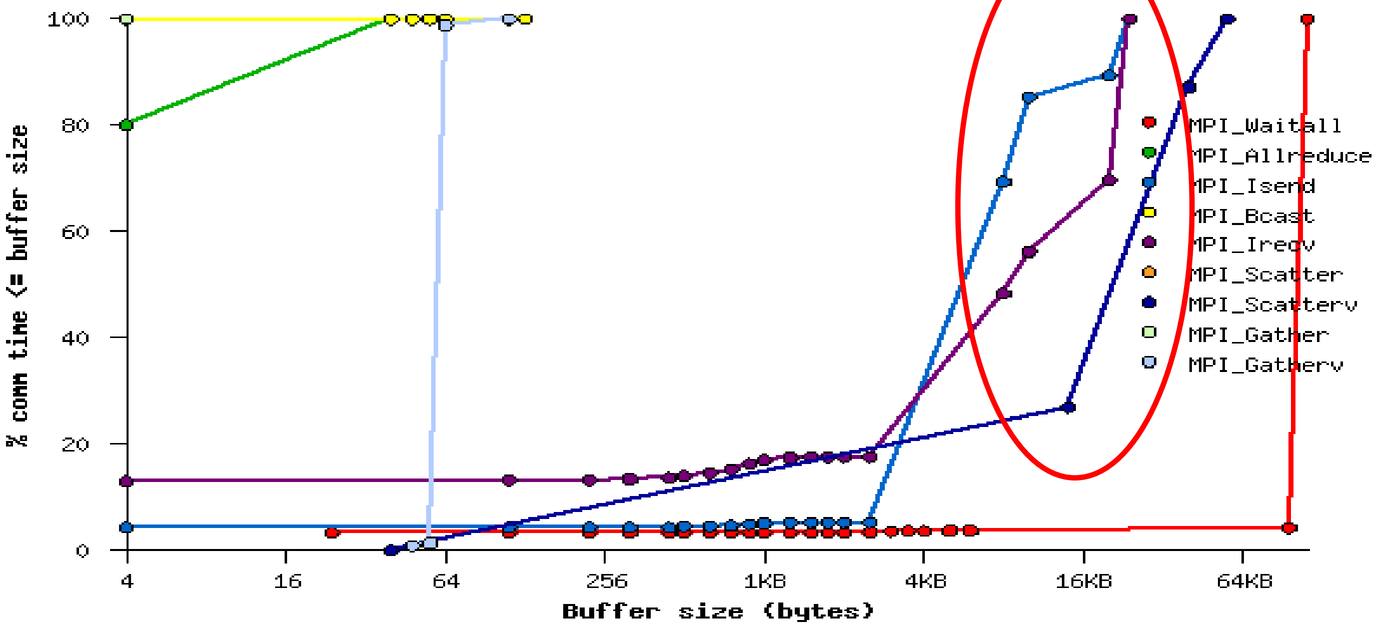
Higher is better

Open MPI

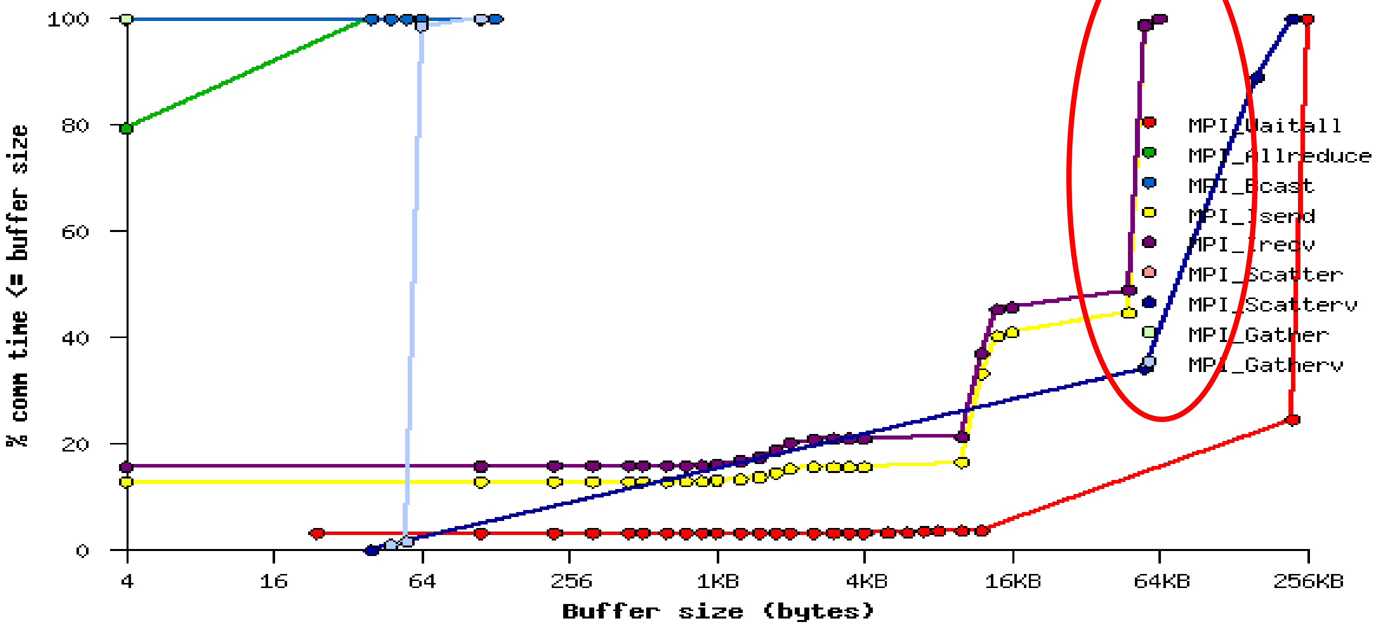
HOOMD-blue Profiling – MPI Message Sizes

- 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

Wilkes Cluster
University of Cambridge

Test Cluster Configuration 2

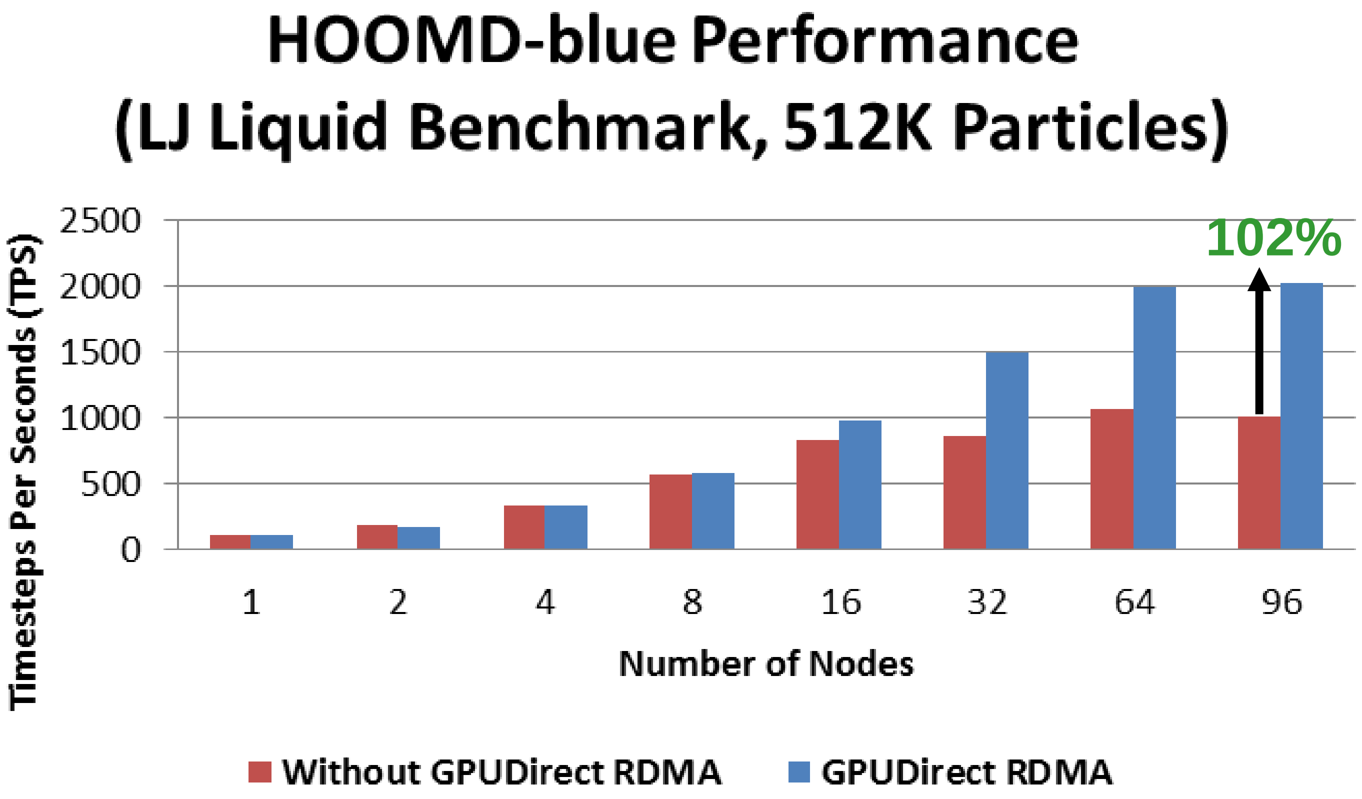
- Dell™ PowerEdge™ T620 128-node (1536-core) Wilkes cluster at Univ of Cambridge
  - Dual-Socket Hexa-Core Intel E5-2630 v2 @ 2.60 GHz CPUs
  - Memory: 64GB memory, DDR3 1600 MHz
  - OS: Scientific Linux release 6.4 (Carbon), MLNX_OFED 2.1-1.0.0 InfiniBand SW stack
  - Hard Drives: 2x 500GB 7.2 RPM 64MB 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 allows direct peer to peer GPU communications over InfiniBand
  - Unlocks performance between GPU and InfiniBand
  - This provides a significant decrease in GPU-GPU communication latency
  - Provides complete CPU offload 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)

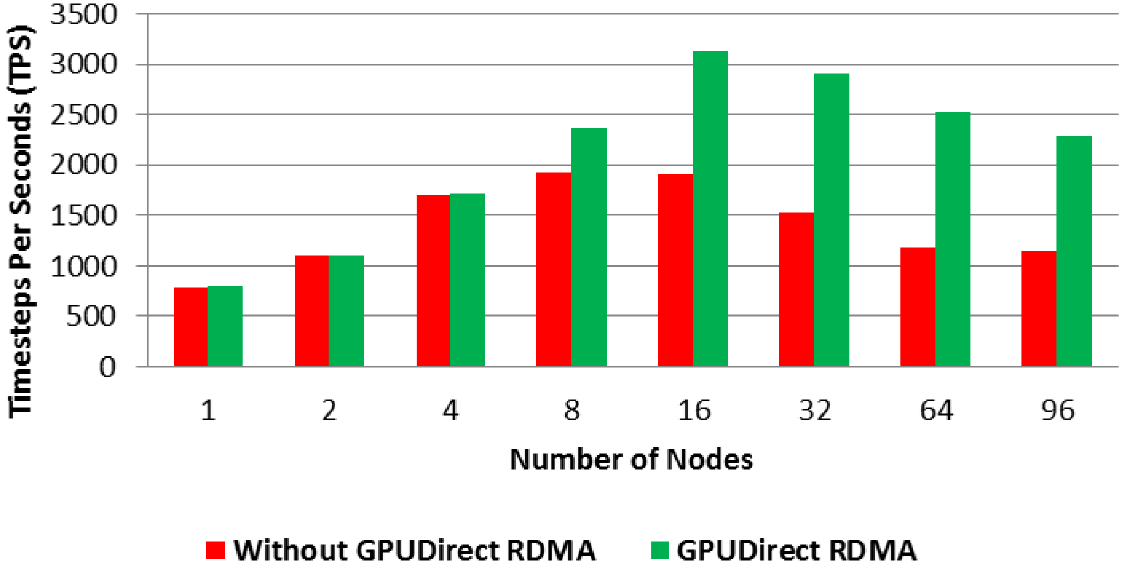


Higher is better

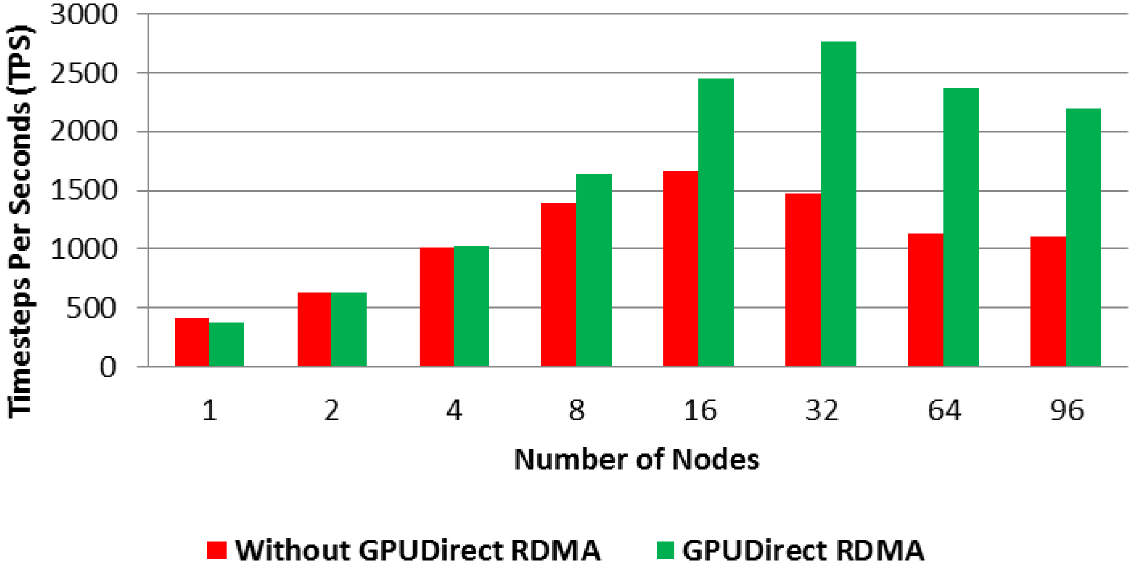
Open MPI

HOOMD-blue Performance – Benefits of GPUDirect RDMA

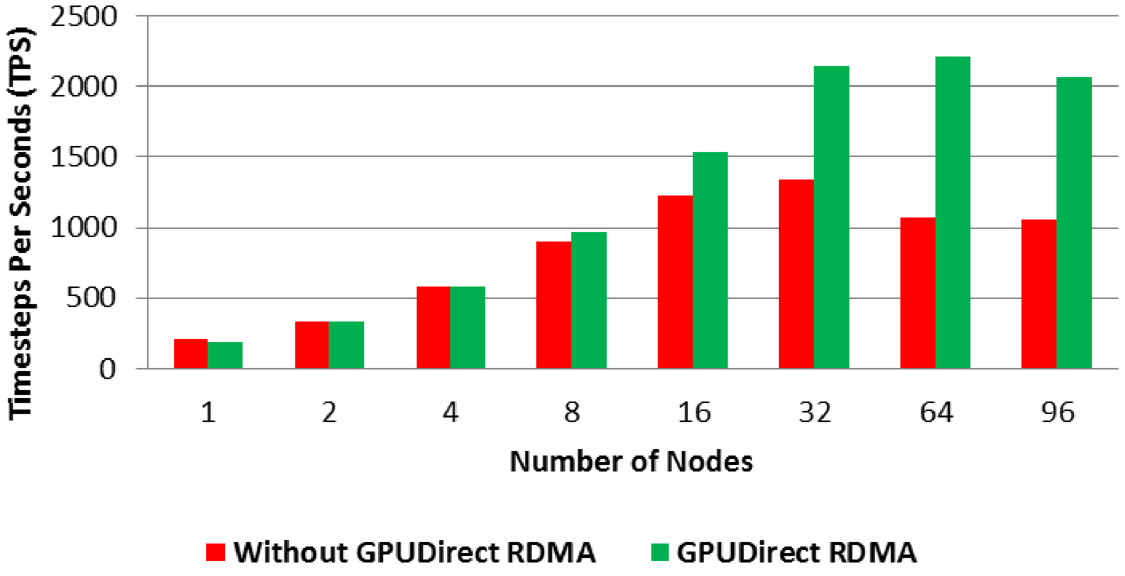
LJ Liquid Benchmark, 64K Particles



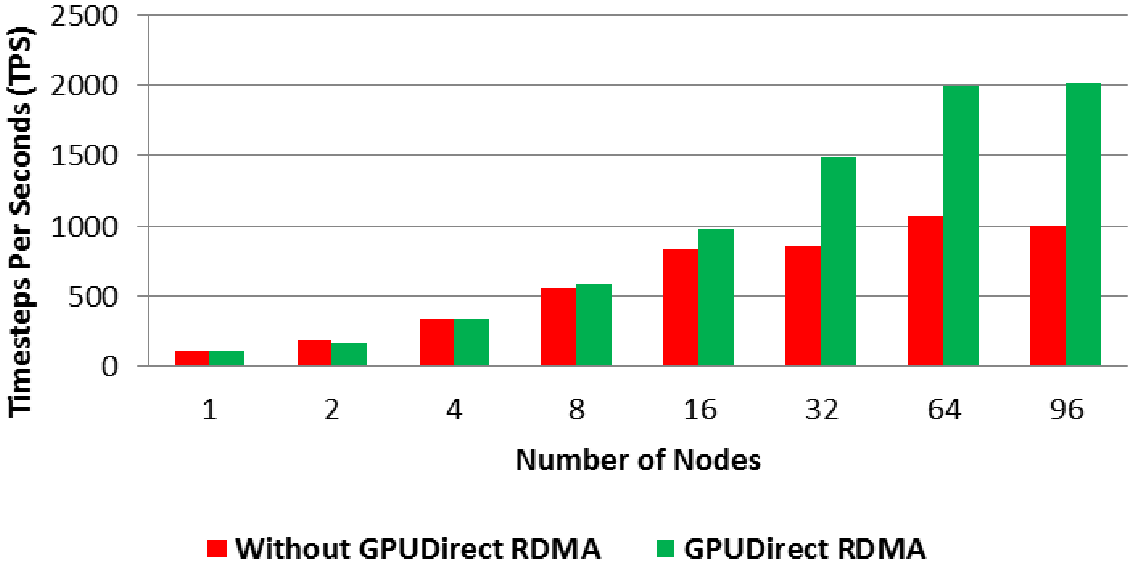
LJ Liquid Benchmark, 128K Particles



LJ Liquid Benchmark, 256K Particles



LJ Liquid Benchmark, 512K Particles

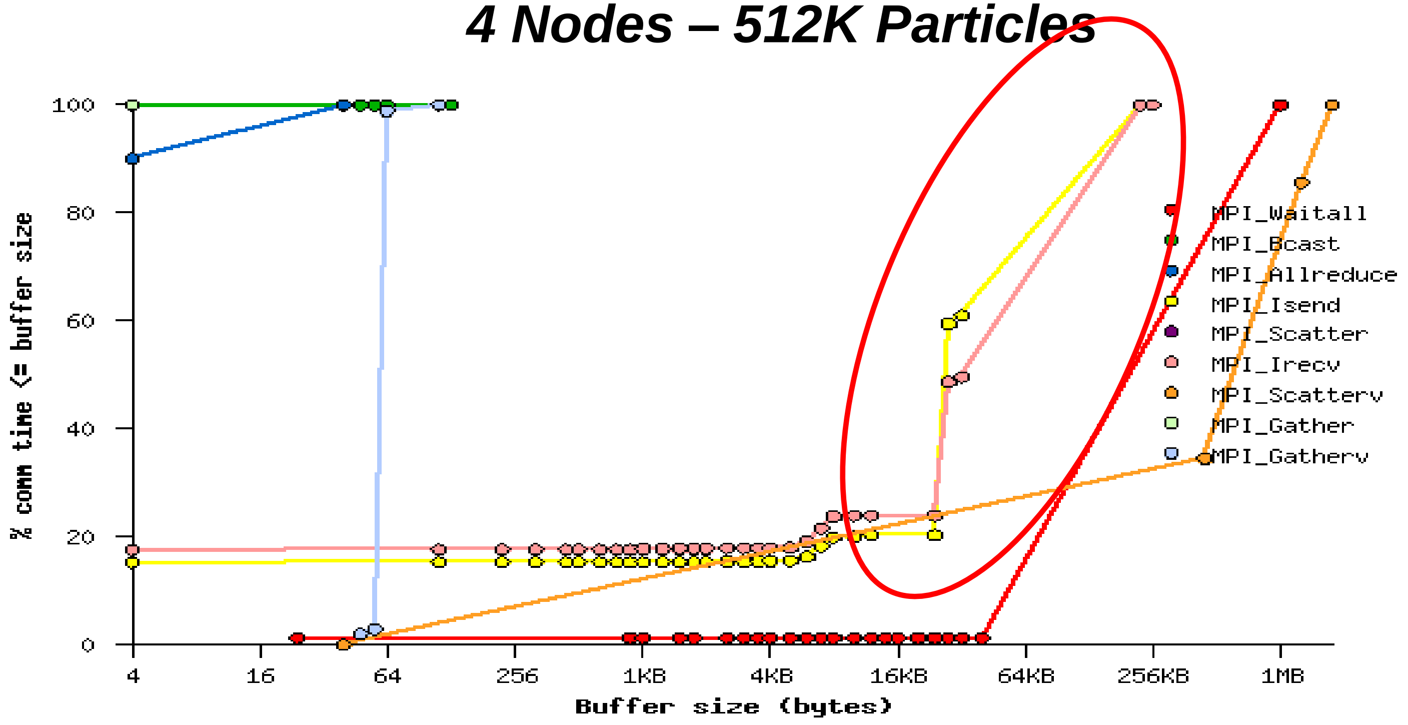


Higher is

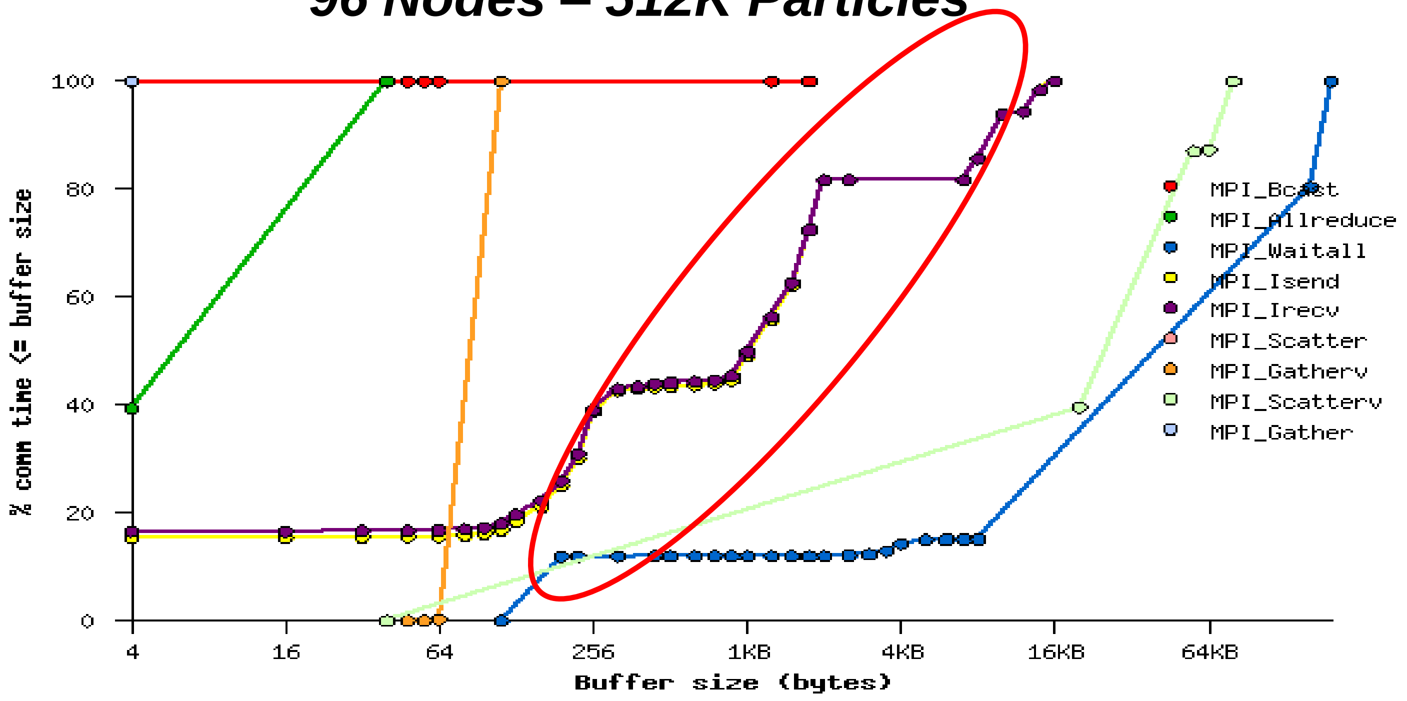
HOOMD-blue Profiling – MPI Message Sizes

- 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

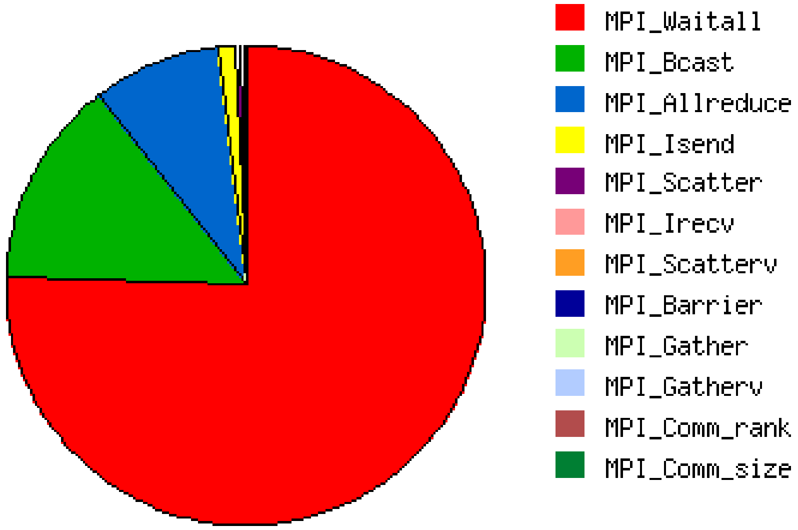


1 MPI Process/Node

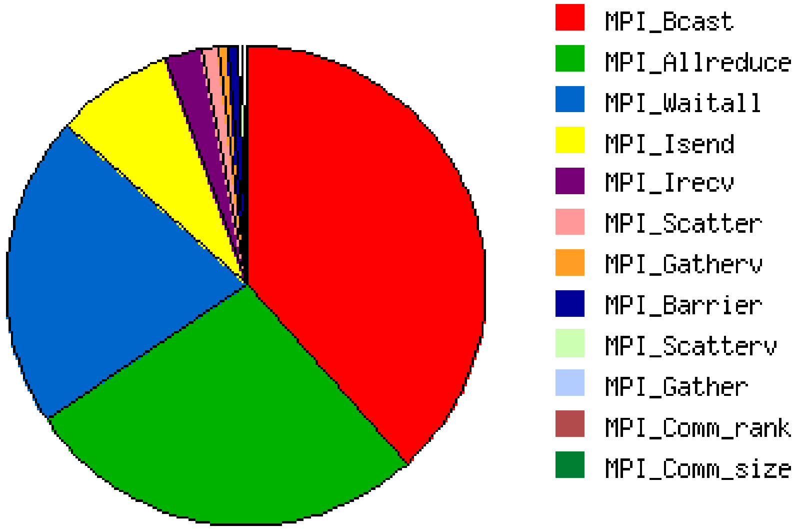
HOOMD-blue Profiling – % Time Spent on MPI

- 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

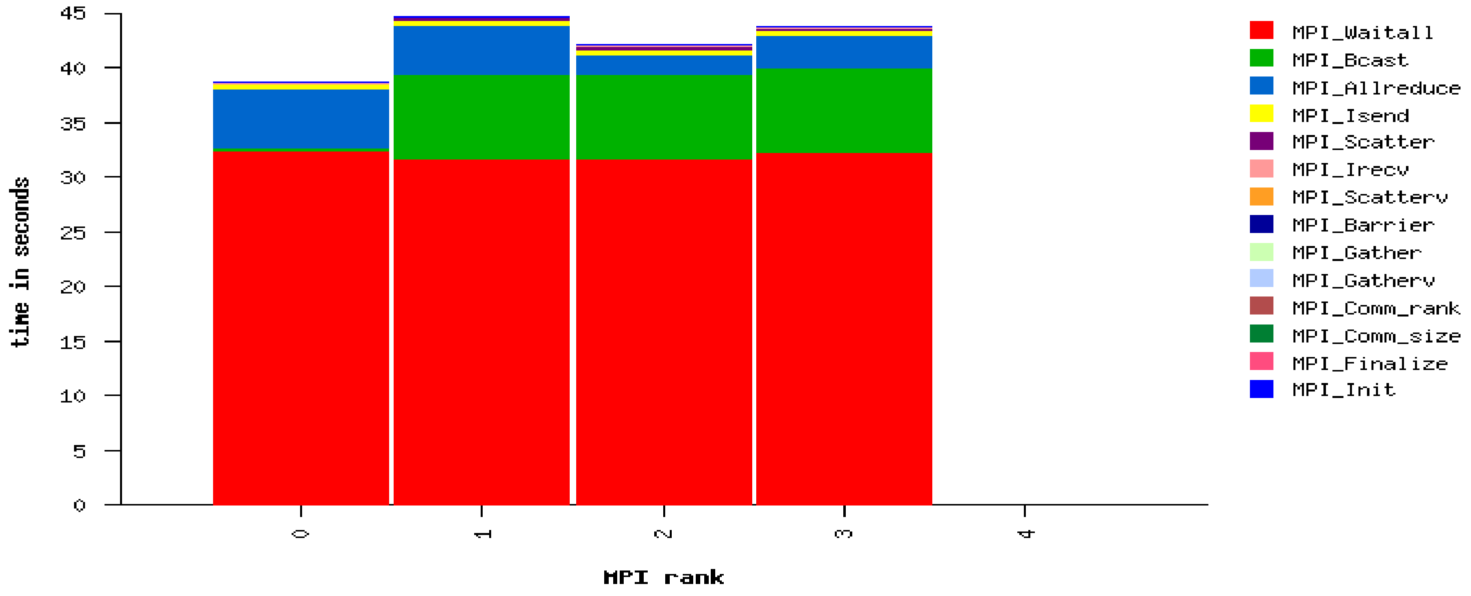


Open MPI

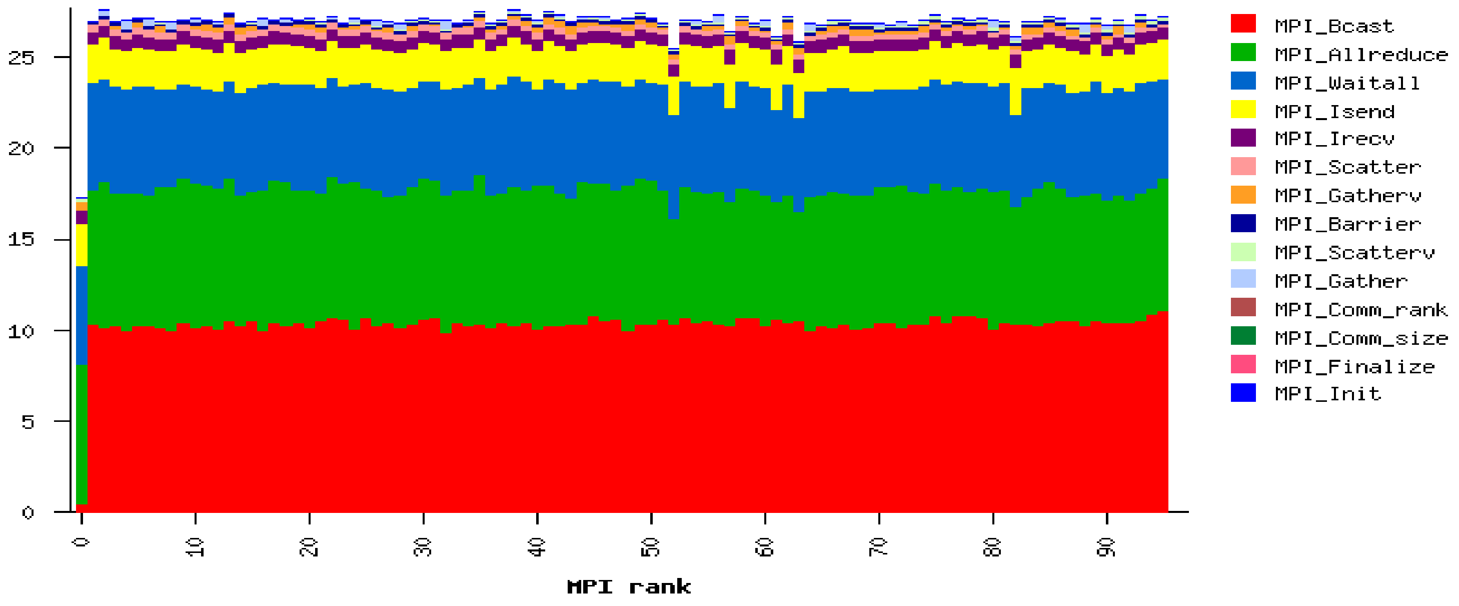
HOOMD-blue Profiling – MPI Communication

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

4 Nodes – 512K Particles



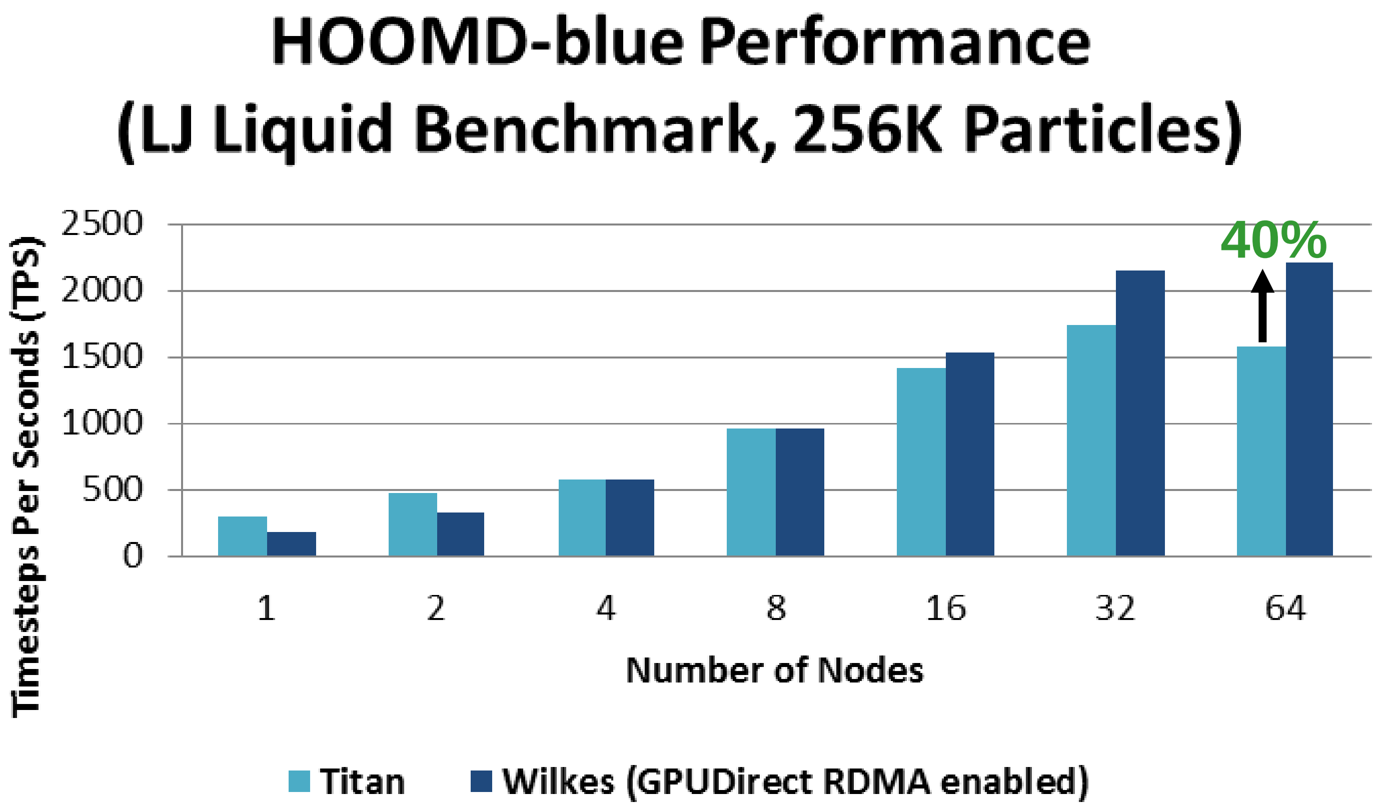
96 Nodes – 512K Particles



1 MPI Process/Node

HOOMD-blue Performance – Scalability

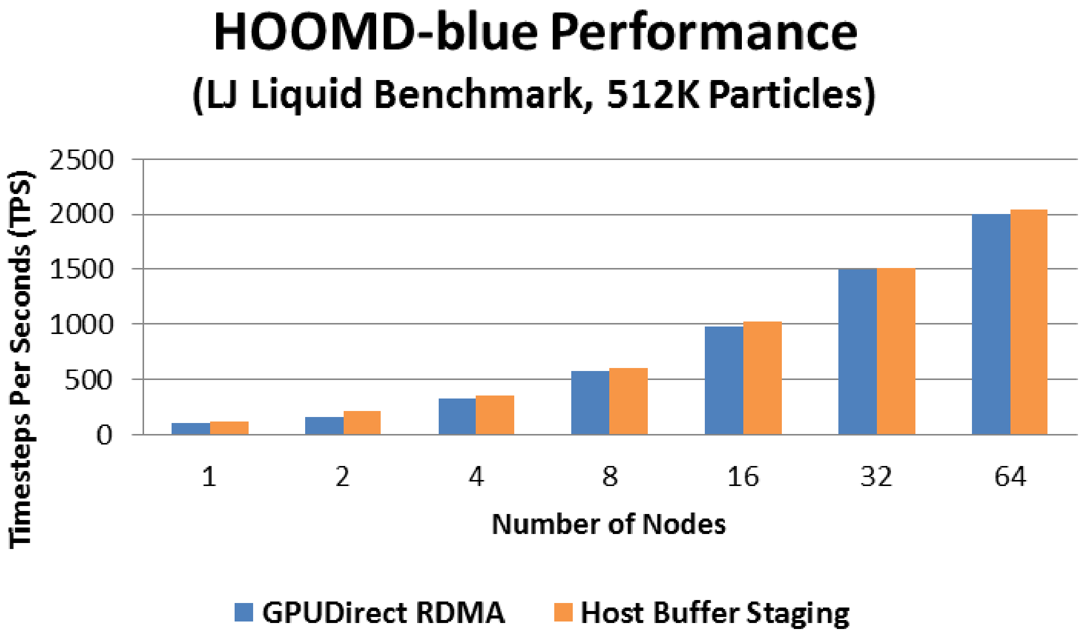
- FDR InfiniBand empowers Wilkes to surpass Titan on scalability
  - Titan showed higher per-node performance but Wilkes outperformed in scalability
  - Titan: K20x GPUs which computes at higher clock rate than the K20 GPU
  - Wilkes: K20 GPUs (using 1 GPU per node) at PCIe Gen2, and FDR InfiniBand at Gen3 rate
- Wilkes exceeds Titan in scalability performance with FDR InfiniBand
  - Outperformed Titan by up to 40% at 64 nodes

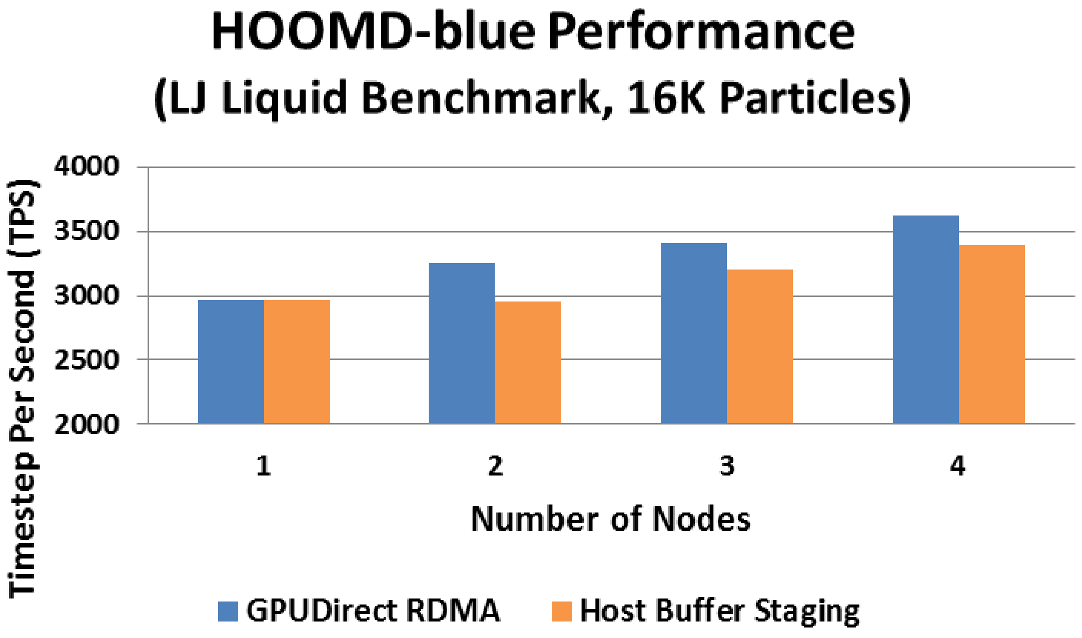


1 Process/Node

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



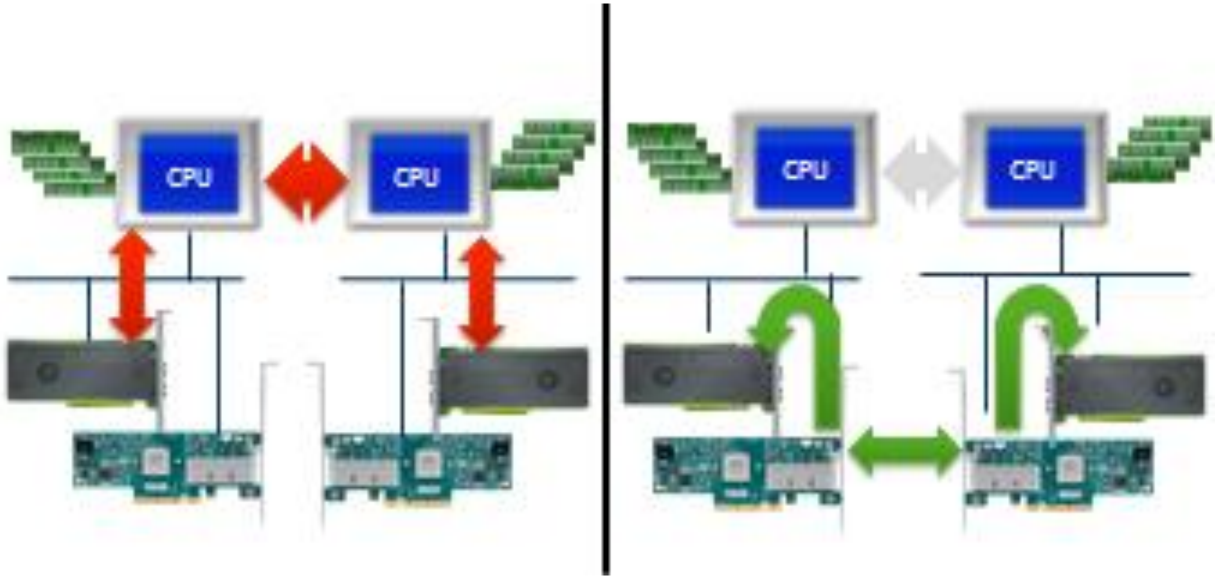


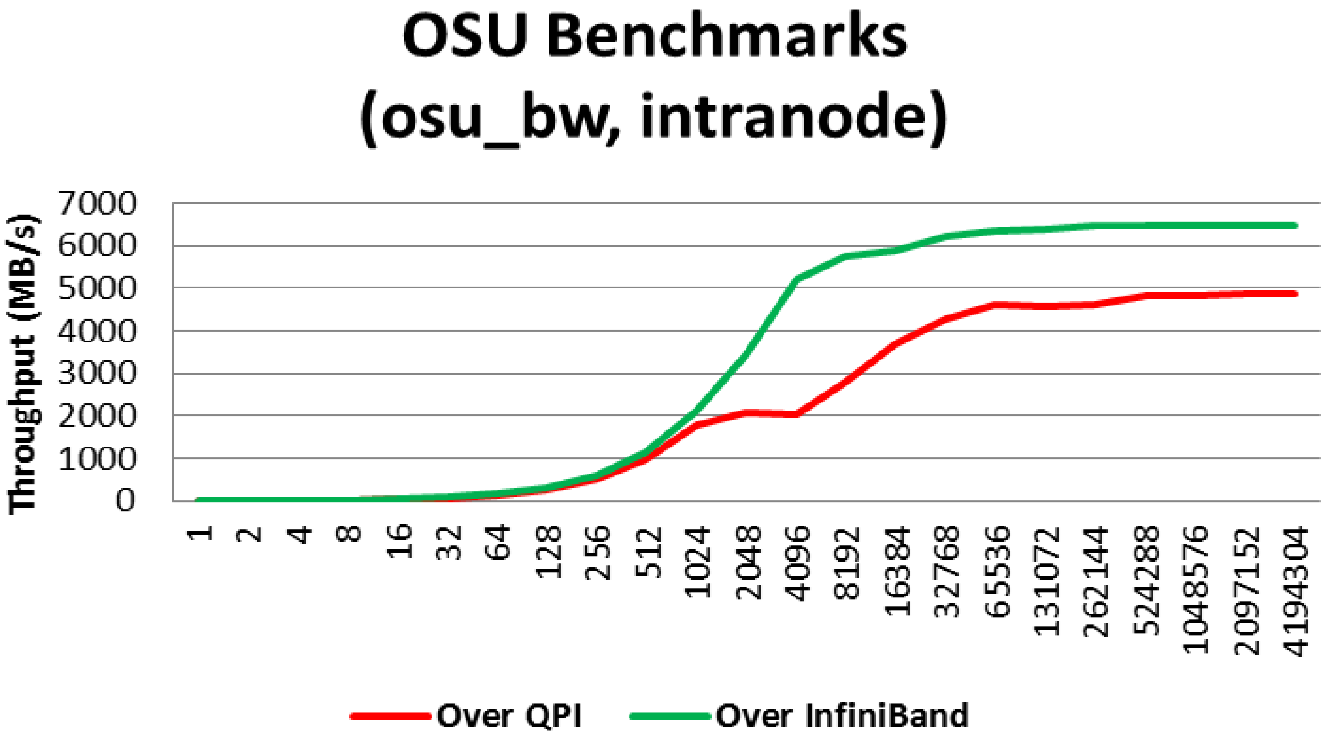
Higher is better

1 Process/Node

HOOMD-blue Performance – Intra-node GPU communications

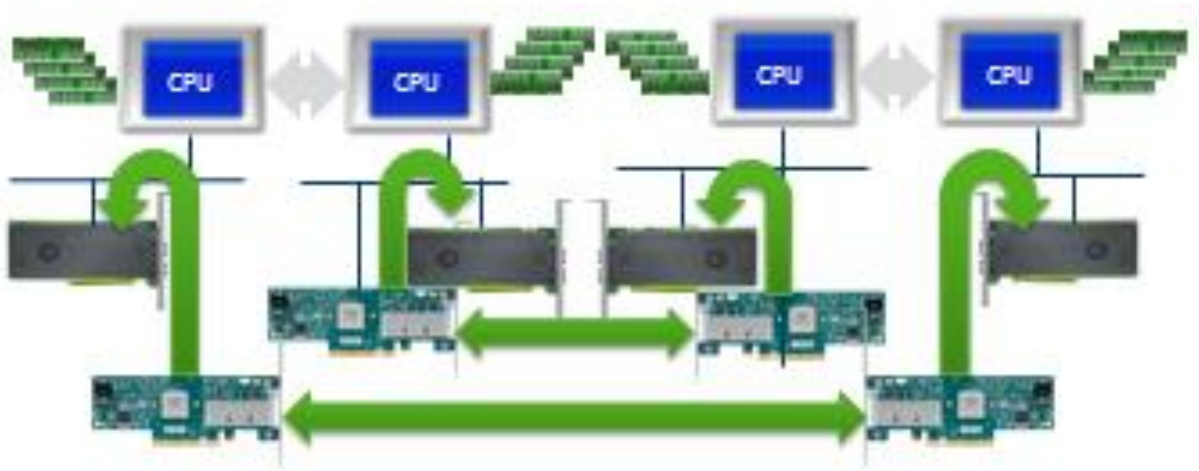
- When data communications between GPU and host memory do not have affinity:
  - Data communications must take place over QPI
  - Performance of such communications would be bottlenecked dramatically
- When MPI communication of GPU data takes place by crossing over the QPI bridge
  - latency and bandwidth that results from the GPU communication would be blocked dramatically compared to a case without crossing over QPI
- Performance difference is between 4.8GB/s over QPI versus 6.5GB/s over FDR InfiniBand

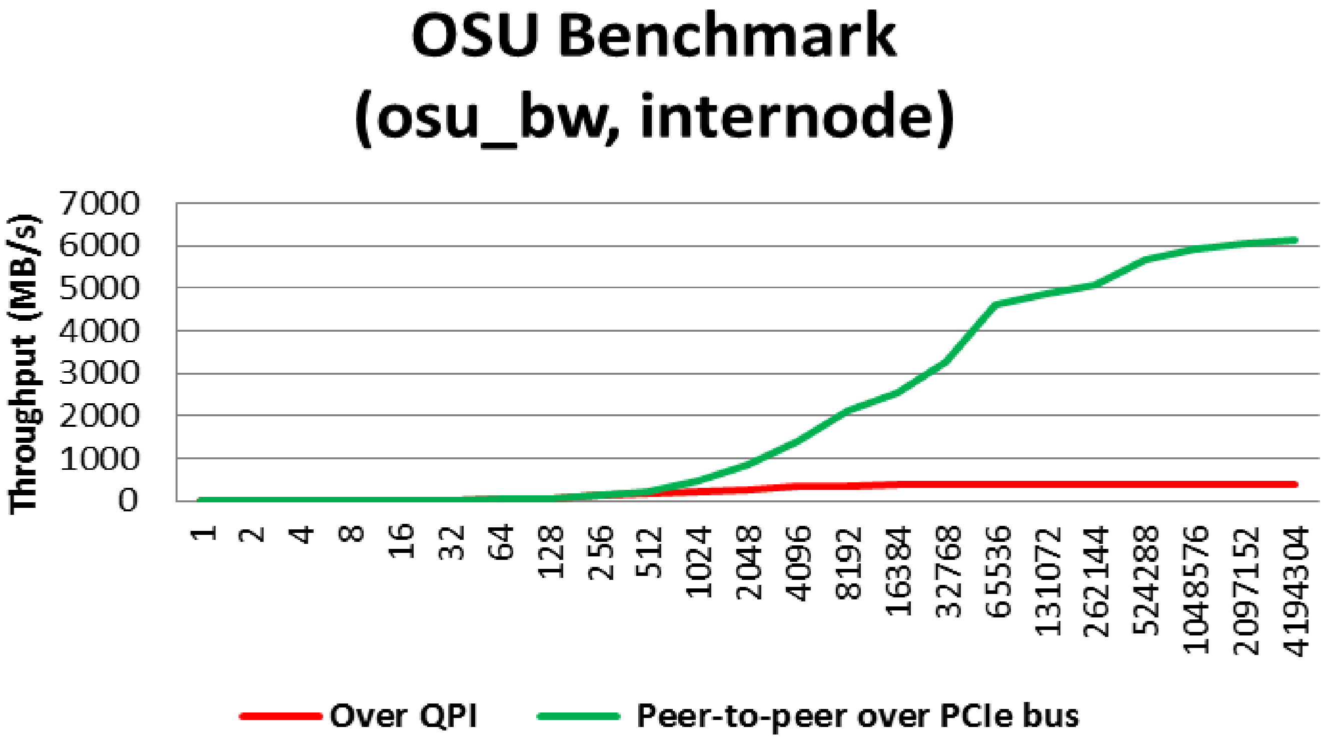




HOOMD-blue Performance – Inter-node GPU communications

- The effect of this QPI penalty would be dramatically worsen for inter-node communications
  - When crossing over the QPI bridge to reach the InfiniBand device
  - Rather than accessing through peer-to-peer method available in GPUDirect RDMA over the PCIe bus
- The OSU bandwidth test confirmed a network bandwidth limitation due to QPI
  - Limitation by QPI to a throughput around 300MB/s instead of more than 6GB/s

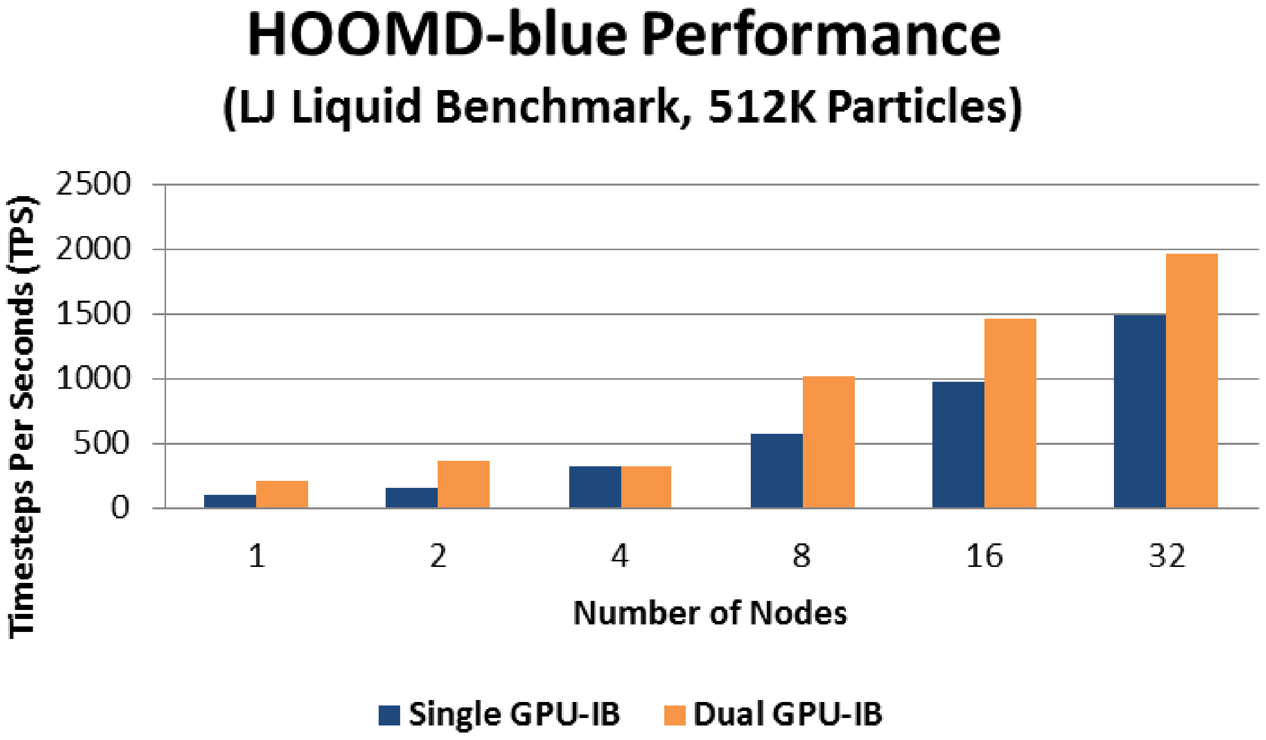




HOOMD-blue Performance – Dual GPU-InfiniBand

-
  ```
  mpirun -np $NP -bind-to socket -display-map -report-bindings --map-by ppr:1:socket \
  --mca mtl ^mxm -mca coll_fca_enable 0 --mca btl openib,self --mca btl_openib_device_selection_verbose 1 \
  --mca btl_openib_warn_nonexistent_if 0 --mca btl_openib_if_include mlx5_0:1,mlx5_1:1 \
  --mca btl_smcuda_use_cuda_ipc 0 --mca btl_smcuda_use_cuda_ipc_same_gpu 1 --mca btl_openib_want_cuda_gdr 1 \
  hoomd lj_liquid_bmark_256000.hoomd
  ```

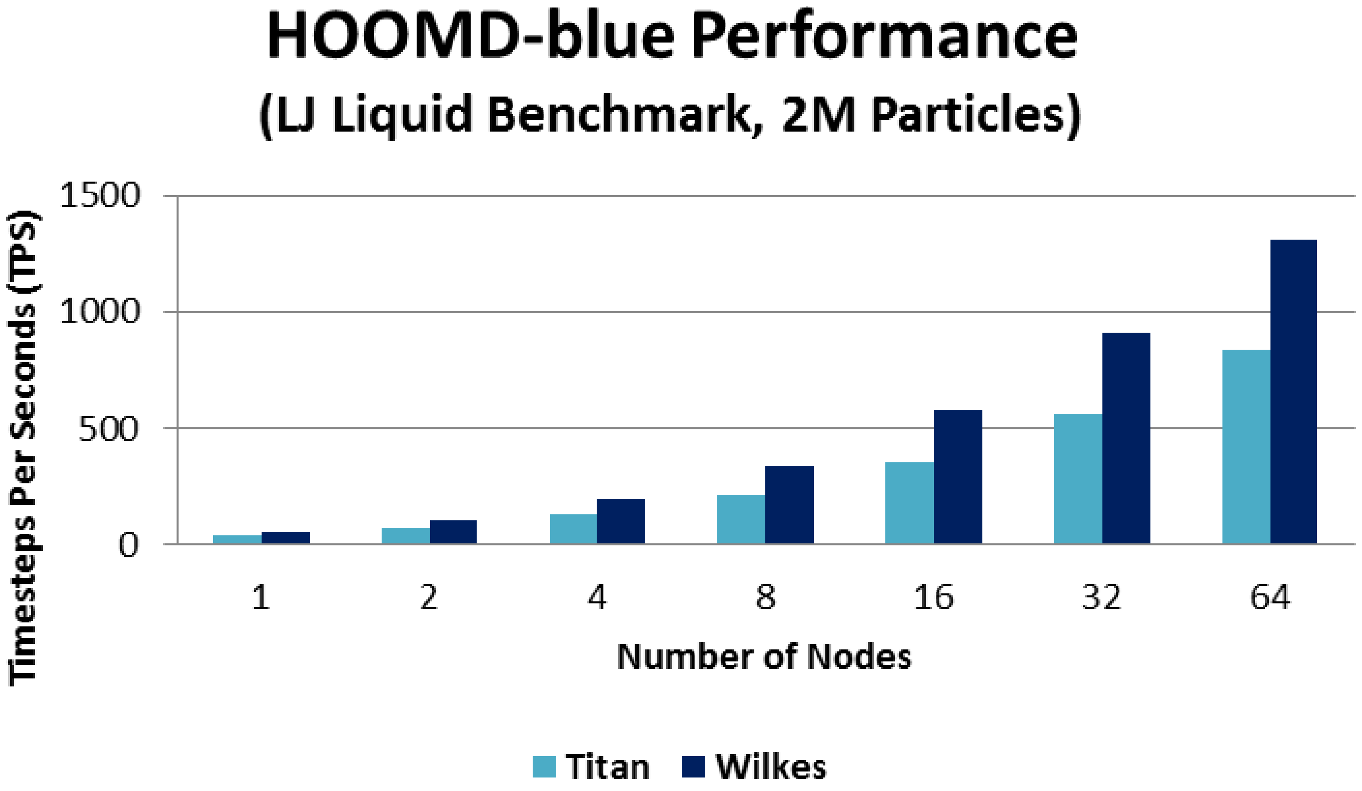
-
  ```
  mpirun -np $NP -ppn 2 -genvall -genv MV2_ENABLE_AFFINITY 1 -genv MV2_CPU_BINDING_LEVEL SOCKET \
  -genv MV2_CPU_BINDING_POLICY SCATTER -genv MV2_RAIL_SHARING_POLICY FIXED_MAPPING \
  -genv MV2_PROCESS_TO_RAIL_MAPPING mlx5_0:mlx5_1
  -genv MV2_USE_CUDA 1 -genv MV2_CUDA_IPC 0 -genv MV2_USE_GPUDIRECT 1 hoomd lj_liquid_bmark_256000.hoomd
  ```



1 Process/Node

HOOMD-blue Performance – Scalability

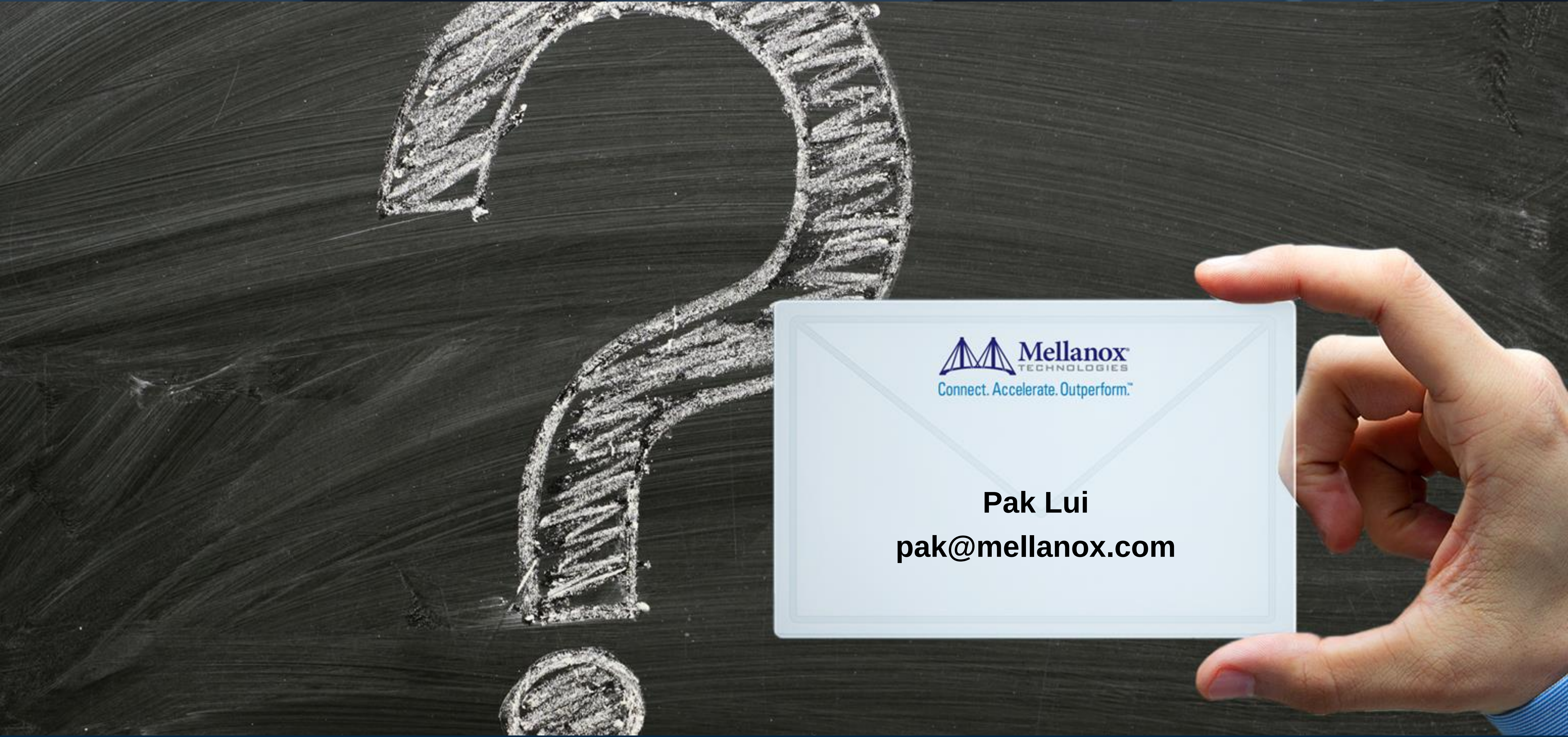
- Dual GPU + dual FDR InfiniBand empowers Wilkes to surpass Titan on scalability
  - Titan: K20x GPUs which computes at higher clock rate than the K20 GPU
  - Wilkes: K20 GPUs (using 2 GPUs per node) at PCIe Gen2, and FDR InfiniBand at Gen3 rate
- Wilkes exceeds Titan in scalability performance with dual FDR InfiniBand
  - Outperformed Titan by up to 67% in some cases



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 technology provides a direct P2P data path between GPU and InfiniBand
  - 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
- QPI can introduce a bottleneck for communications between (intra/internode) GPU devices
  - Bottleneck can be avoided by going over the InfiniBand for communications
- GPUDirect RDMA performs better than Host Buffer Staging in some cases
  - On large scale, HBS performance appears to perform slightly faster than GDR
  - On small scale, GDR can be faster than HBS for small number of particles per GPU

Question Time



Pak Lui

pak@mellanox.com

Thank You

Mellanox TECHNOLOGIES

Connect. Accelerate. Outperform.™