GPU TECHNOLOGY CONFERENCE

MELLANOX EDR UPDATE & GPUDIRECT

MELLANOX SR. SE 정연구

Leading Supplier of End-to-End Interconnect Solutions

Analyze

Mellanox TECHNOLOGIES

Enabling the Use of Data

Store

Comprehensive End-to-End InfiniBand and Ethernet Portfolio

ICs | Adapter Cards | Switches/Gateways | Software and Services | Metro / WAN | Cables/Modules

At the Speeds of 10, 25, 40, 50, 56 and 100 Gigabit per Second

Entering the Era of 100Gb/s

Mellanox TECHNOLOGIES

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

Switch

Spectrum™

32 100GbE Ports, 64 25/50GbE Ports

(10 / 25 / 40 / 50 / 100GbE)

Throughput of 6.4Tb/s

Interconnect

LinkX™

Copper (Passive, Active)

Optical Cables (VCSEL)

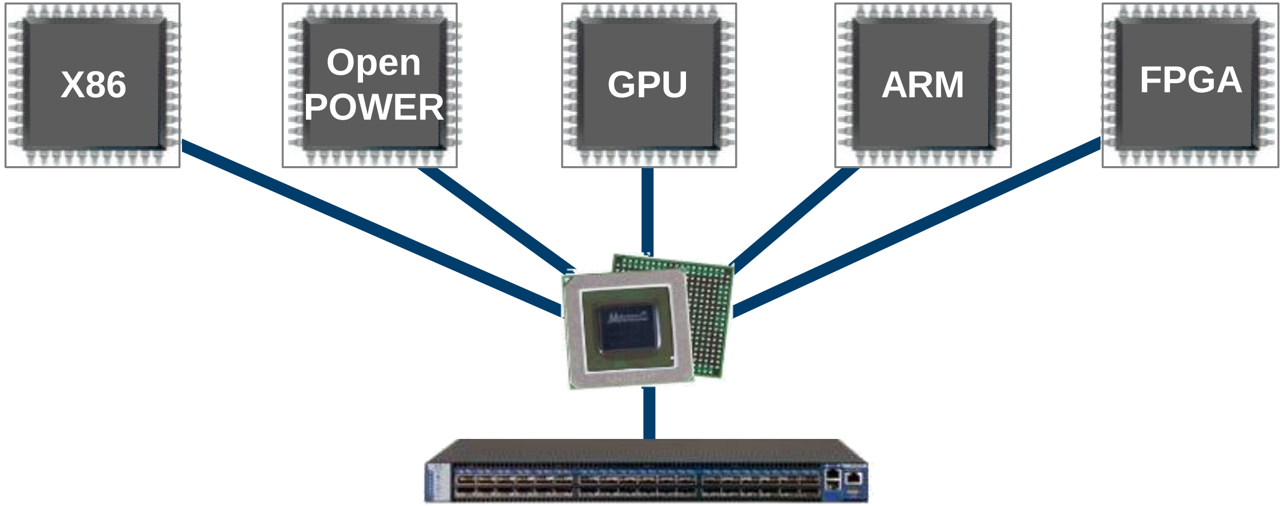
Silicon Photonics

End-to-End Interconnect Solutions for All Platforms

Highest Performance and Scalability for

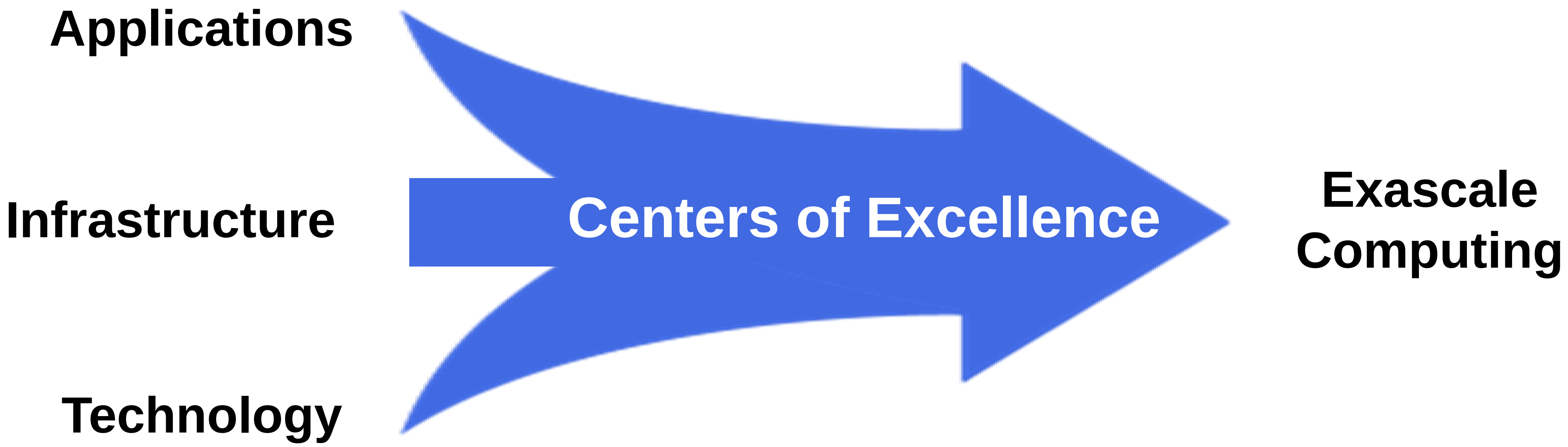
X86, Power, GPU, ARM and FPGA-based Compute and Storage Platforms

10, 20, 25, 40, 50, 56 and 100Gb/s Speeds



Smart Interconnect to Unleash The Power of All Compute Architectures

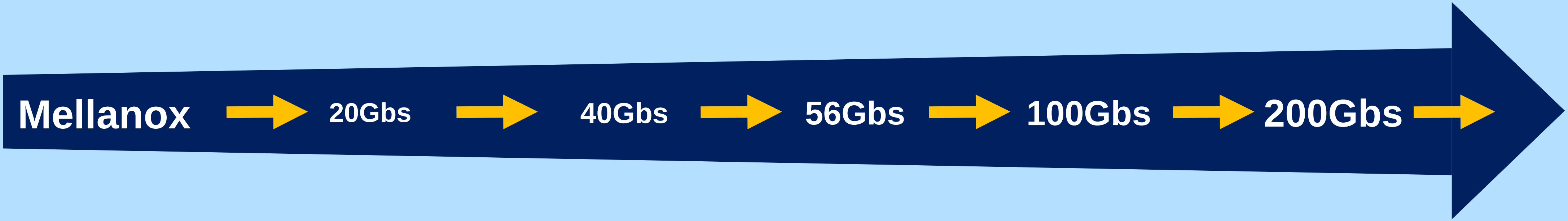
Mellanox Centers of Excellence



- Early access to new technologies (EDR, Multi-Host, HPC-X etc.)
- Co-Design effort to optimize and accelerate applications performance and scalability
- Participate in the Mellanox advisory board

Together We Can Develop the Solutions of Tomorrow

Technology Roadmap – One-Generation Lead over the Competition



Terascale

Petascale

Exascale

3rd

TOP500 2003
Virginia Tech (Apple)

1st

"Roadrunner"
Mellanox Connected

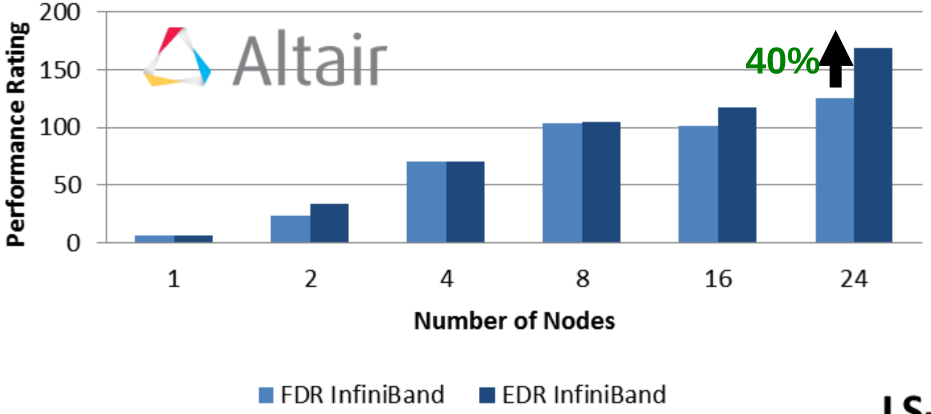
OAK RIDGE National Laboratory
"Summit" System

Lawrence Livermore National Laboratory
"Sierra" System

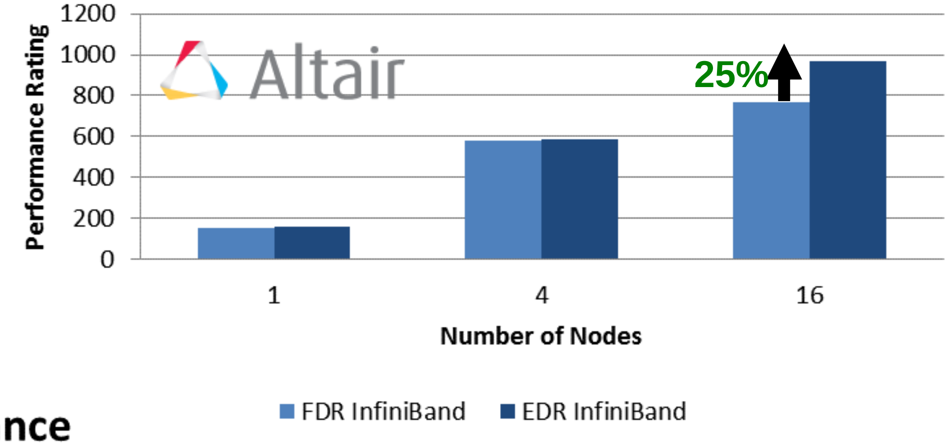
2000 **2005** **2010** **2015** **2020**

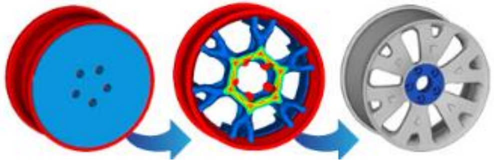
EDR InfiniBand Performance – Commercial Applications

OptiStruct Performance (Engine_Assy.fem)

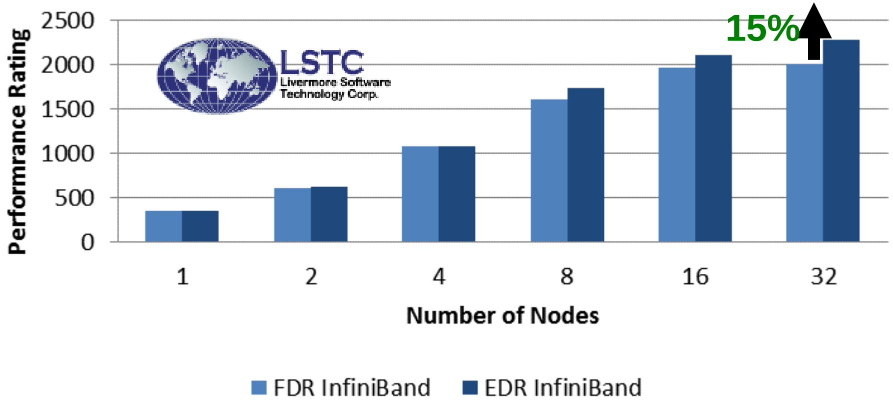


RADIOSS 13.0 Performance (NEON1M11, MPP)



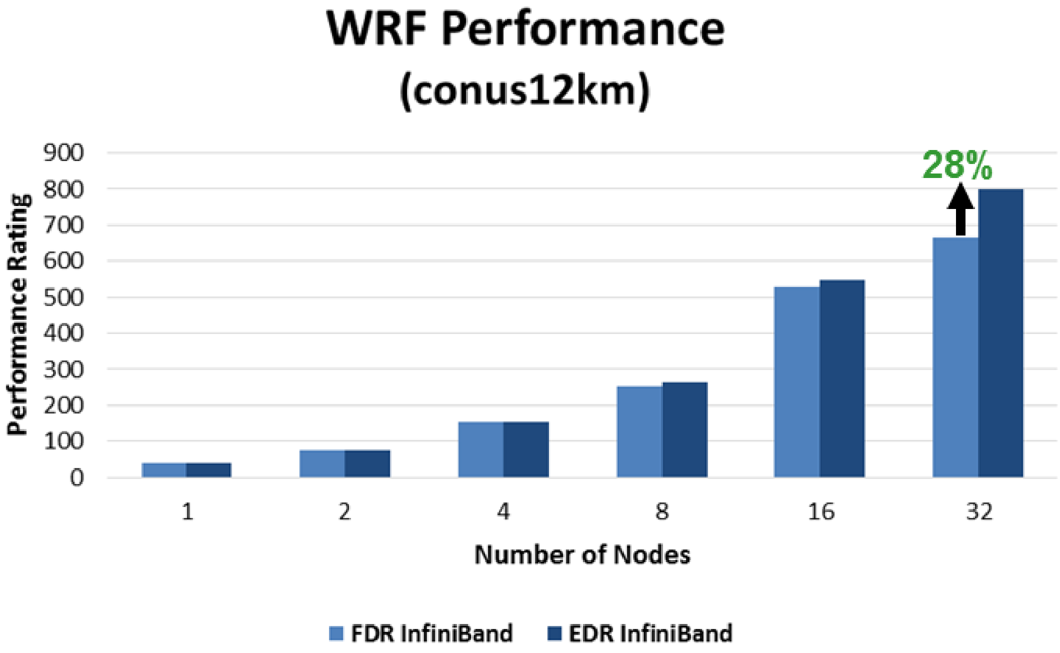


LS-DYNA Performance (neon_refined_revised)

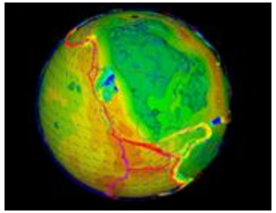


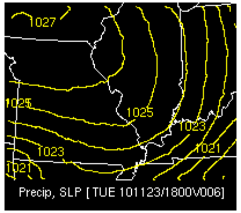
EDR InfiniBand Performance – Weather Simulation

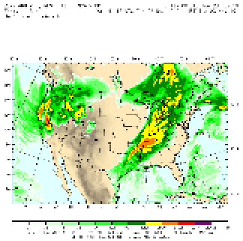
- Weather Research and Forecasting Model
- Optimization effort with the HPCAC
- EDR InfiniBand delivers 28% higher performance
  - 32-node cluster
  - Performance advantage increase with system size



WRF THE WEATHER RESEARCH & FORECASTING MODEL







HPC ADVISORY COUNCIL

InfiniBand Adapters Performance Comparison

Mellanox Adapters Single Port Performance	ConnectX-4 EDR 100G	Connect-IB FDR 56G	ConnectX-3 Pro FDR 56G
Uni-Directional Throughput	100 Gb/s	54.24 Gb/s	51.1 Gb/s
Bi-Directional Throughput	195 Gb/s	107.64 Gb/s	98.4 Gb/s
Latency	0.61 us	0.63 us	0.64 us
Message Rate	149.5 Million/sec	105 Million/sec	35.9 Million/sec

Mellanox Interconnect Advantages

- Proven, scalable and high performance end-to-end connectivity
- Flexible, support all compute architectures: x86, Power, ARM, GPU, FPGA etc.
- Standards-based (InfiniBand, Ethernet), supported by large eco-system
- Offloading architecture – RDMA, application acceleration engines etc.
- Flexible topologies: Fat Tree, mesh, 3D Torus, Dragonfly+, etc.
- Converged I/O– compute, storage, management on single fabric
- Backward and future compatible
- EDR InfiniBand delivers highest applications performance

Speed-Up Your Present, Protect Your Future

Paving The Road to Exascale Computing Together

Mellanox PeerDirect™ with NVIDIA GPUDirect™ RDMA

Native support for peer-to-peer communications
between Mellanox HCA adapters and NVIDIA GPU devices

Industry Adoption of GPUDirect RDMA

- GPUDirect RDMA was released in May 2014 and is available for download from Mellanox
  - Adoption and development continues to grow in various areas of technical disciplines
    - Leveraging RDMA and NVIDIA GPUs in today’s energy-efficient datacenters

Big Data

Bioscience

Defense

Database

Green Computing

Government

Healthcare

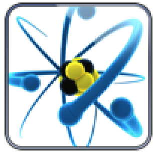
HPC

Risk / Analysis

Space Exploration

Transportation

Oil & Gas


Physics

Research/Education

Financial

What is PeerDirect™

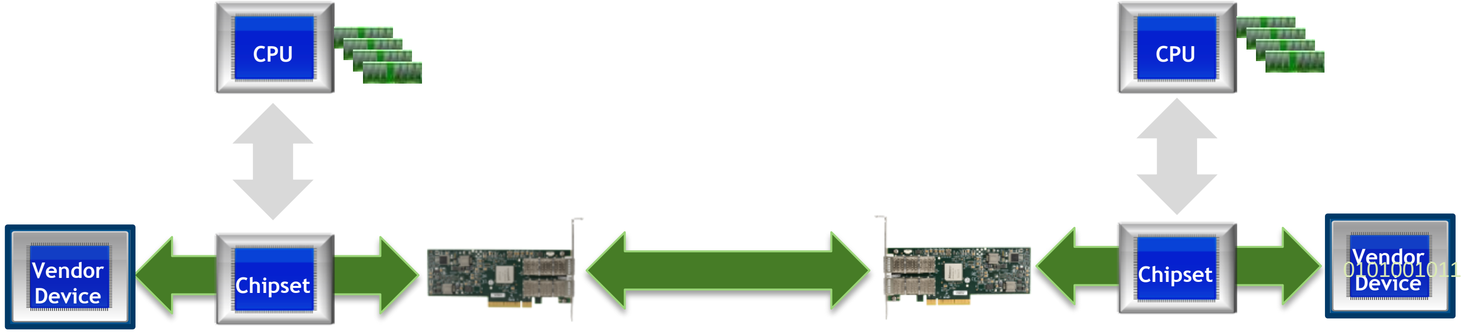
PeerDirect is natively supported by Mellanox OFED 2.1 or later distribution

Supports peer-to-peer communications between Mellanox adapters and third-party devices

No unnecessary system memory copies & CPU overhead

- No longer needs a host buffer for each device
- No longer needs to share a host buffer either

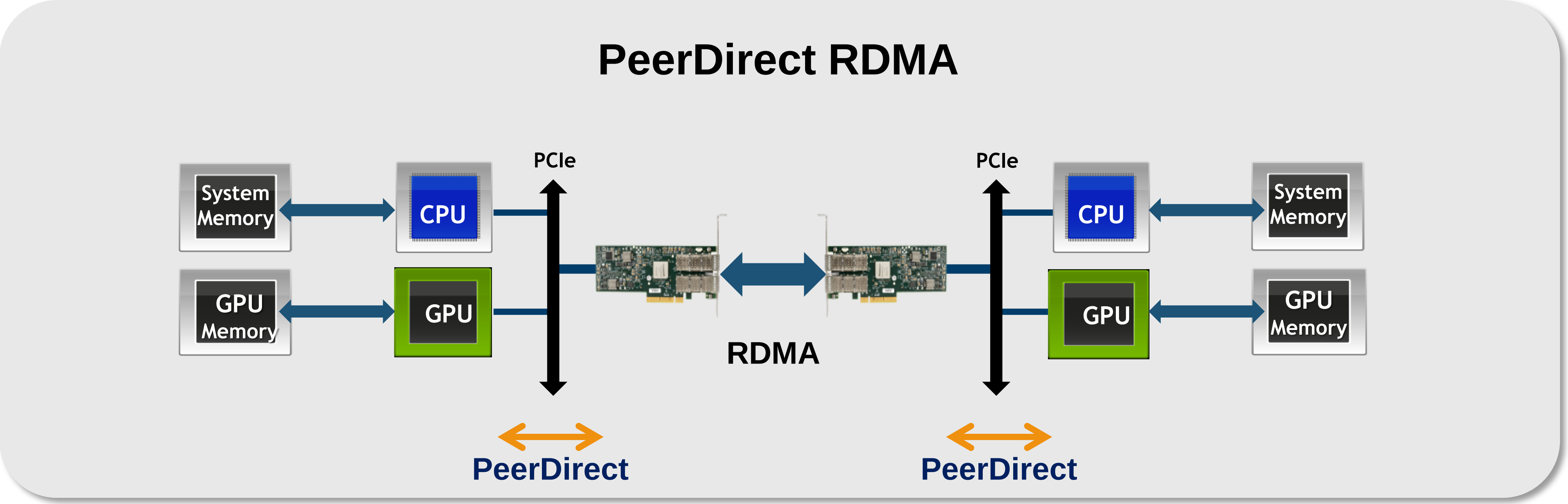
- Supports NVIDIA® GPUDirect RDMA with a separate plug-in
- Support for RoCE protocol over Mellanox VPI



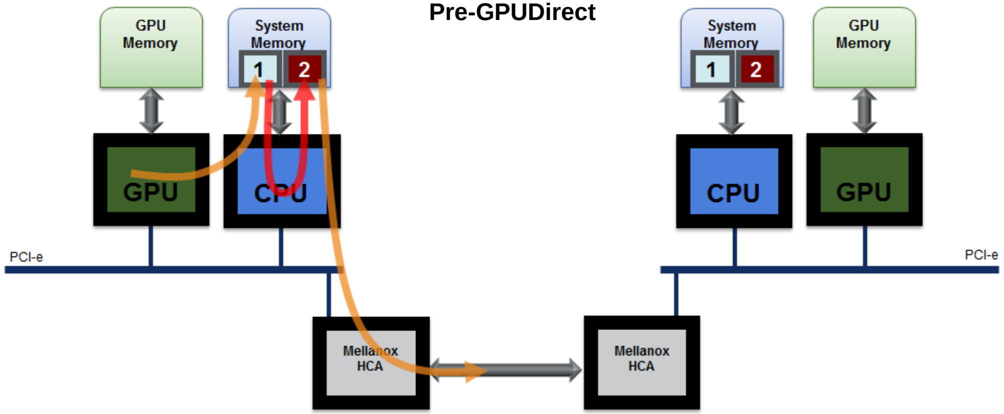
Supported with all Mellanox ConnectX-3 and Connect-IB Adapters

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

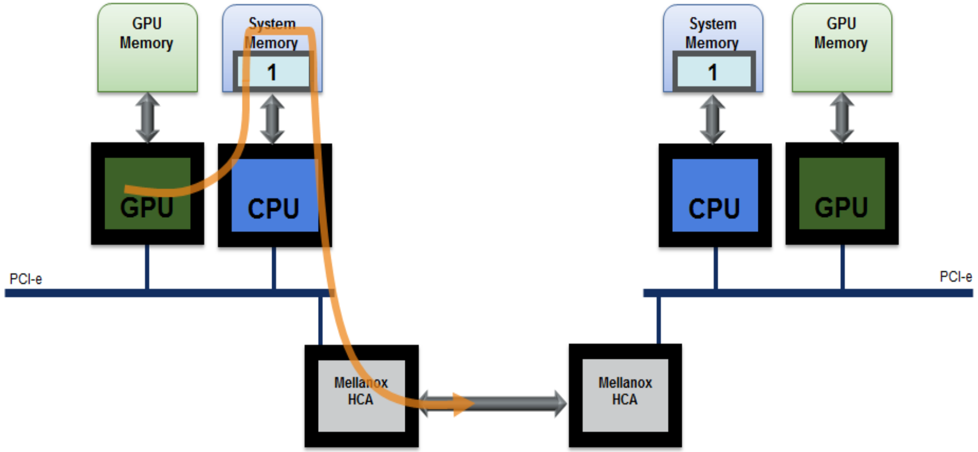


Evolution of GPUDirect RDMA



Before GPUDirect

- Network and third-party device drivers, did not share buffers, and needed to make a redundant copy in host memory.

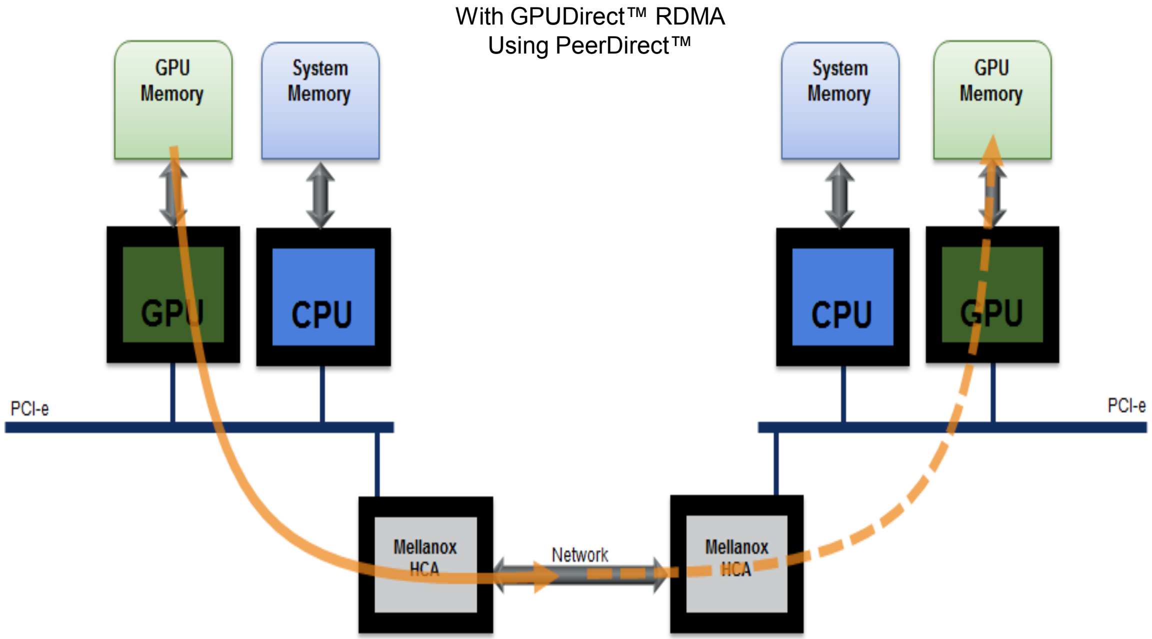


GPUDirect Shared Host Memory Pages Model

With GPUDirect Shared Host Memory Pages

- Network and GPU could share “pinned” (page-locked) buffers, eliminating the need to make a redundant copy in host memory.

GPUDirect™ RDMA (GPUDirect 3.0)



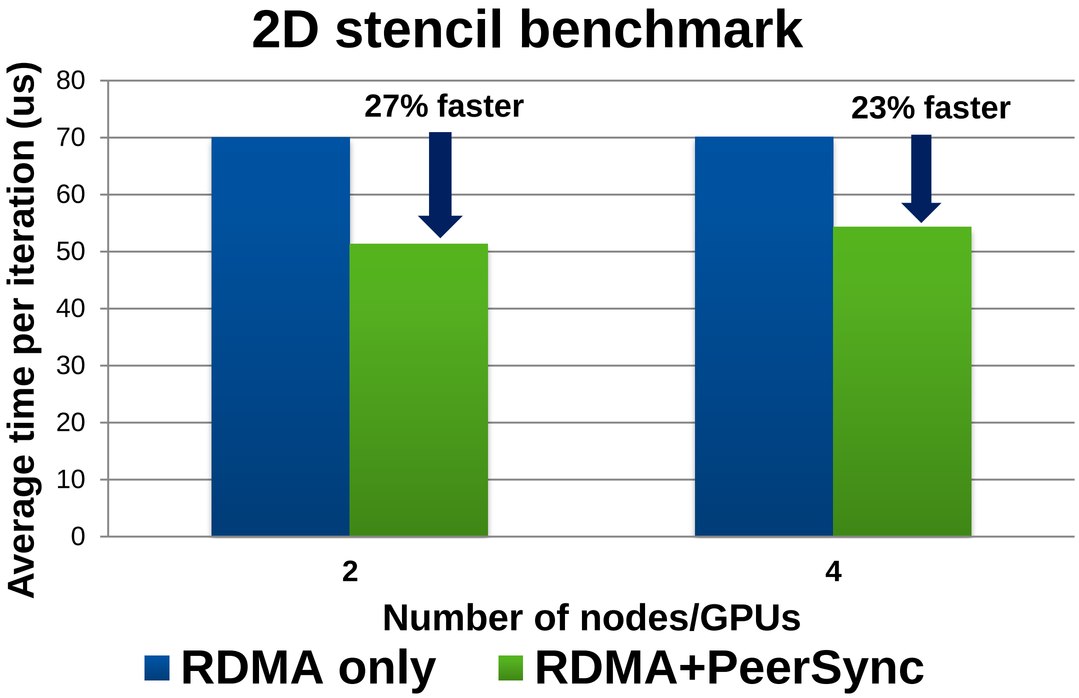
- Eliminates CPU bandwidth and latency bottlenecks
- Uses remote direct memory access (RDMA) transfers between GPUs
- Resulting in significantly improved MPI efficiency between GPUs in remote nodes
- Based on PCIe PeerDirect technology

GPUDirect Sync (GPUDirect 4.0)

- GPUDirect RDMA (3.0) – direct data path between the GPU and Mellanox interconnect
  - Control path still uses the CPU
    - CPU prepares and queues communication tasks on GPU
    - GPU triggers communication on HCA
    - Mellanox HCA directly accesses GPU memory

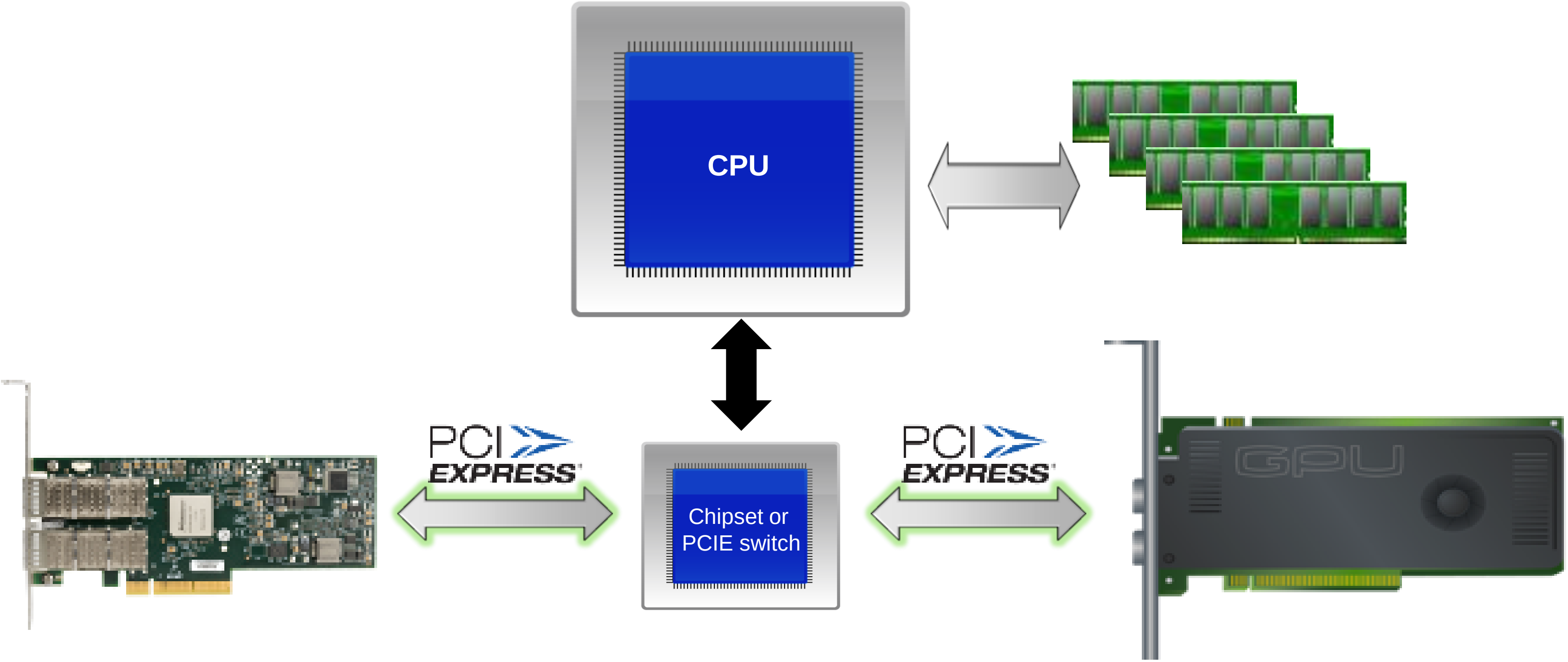
- GPUDirect Sync (GPUDirect 4.0)
  - Both data path and control path go directly between the GPU and the Mellanox interconnect

Maximum Performance For GPU Clusters



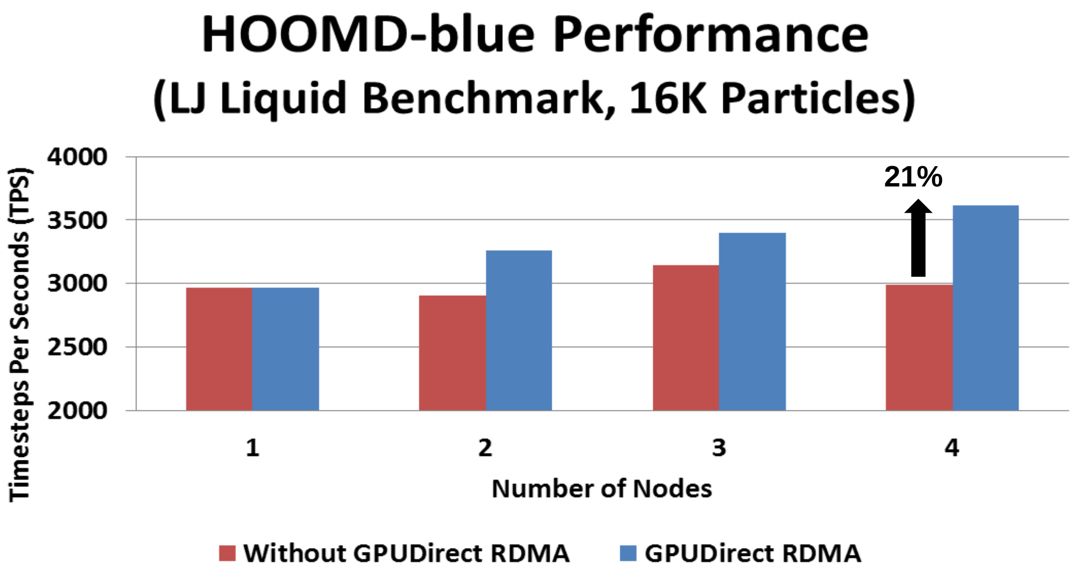
Hardware considerations for GPUDirect RDMA

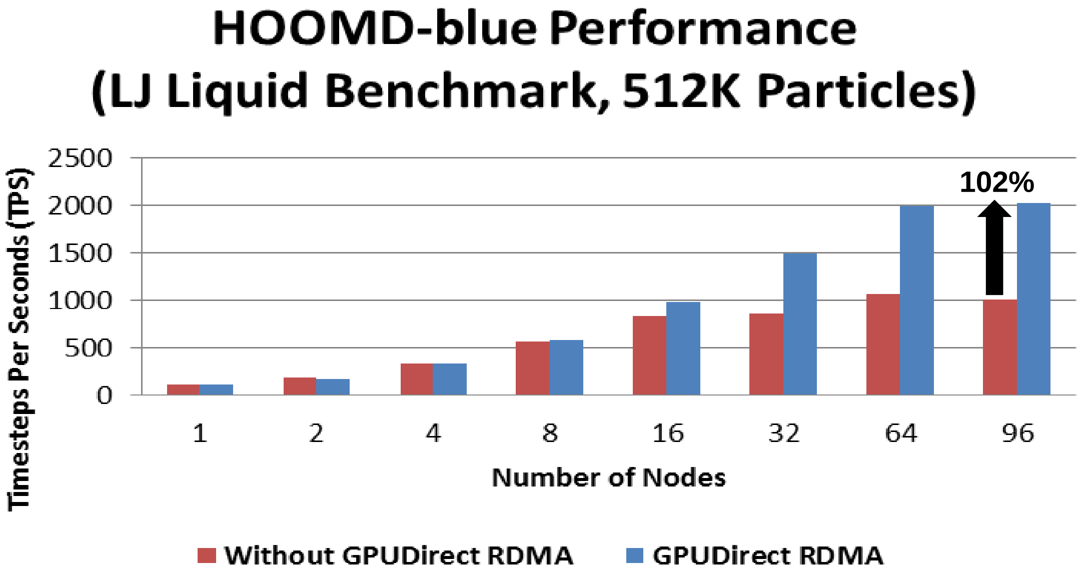
Note : A requirement on current platforms for GPUDirect RDMA to work properly is that the NVIDIA GPU and the Mellanox InfiniBand Adapter share the same root complex… Only a limitation of current hardware today, not GPUDirect RDMA



Mellanox PeerDirect™ with NVIDIA GPUDirect RDMA

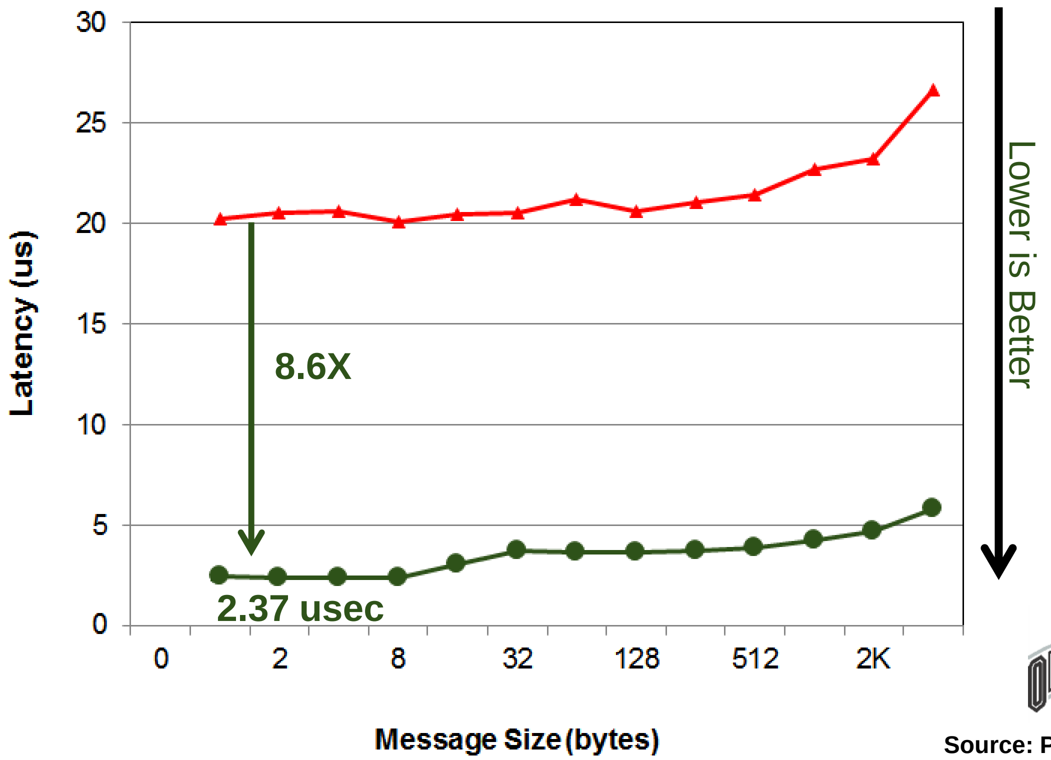
- HOOMD-blue is a general-purpose Molecular Dynamics simulation code accelerated on GPUs
- GPUDirect RDMA allows direct peer to peer GPU communications over InfiniBand
  - Unlocks performance between GPU and InfiniBand
  - This provides a significant decrease in GPU-GPU communication latency
  - Provides complete CPU offload from all GPU communications across the network
- Demonstrated up to 102% performance improvement with large number of particles



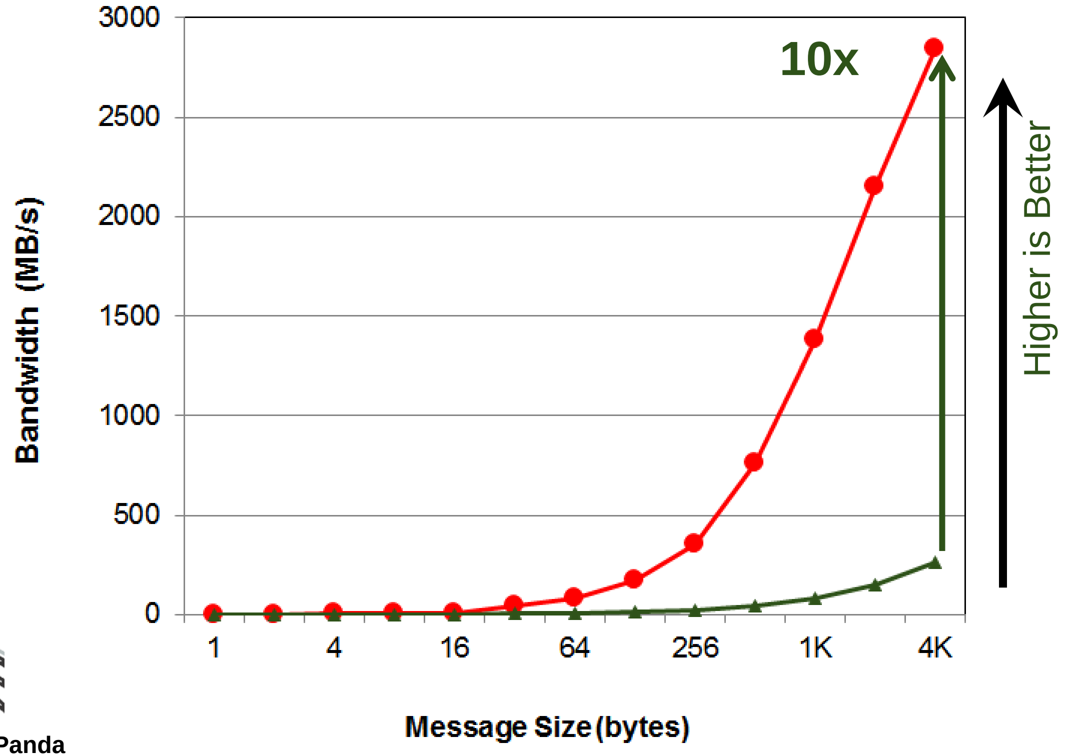


Performance of MVAPICH2 with GPUDirect RDMA

GPU-GPU Internode MPI Latency



GPU-GPU Internode MPI Bandwidth



Source: Prof. DK Panda

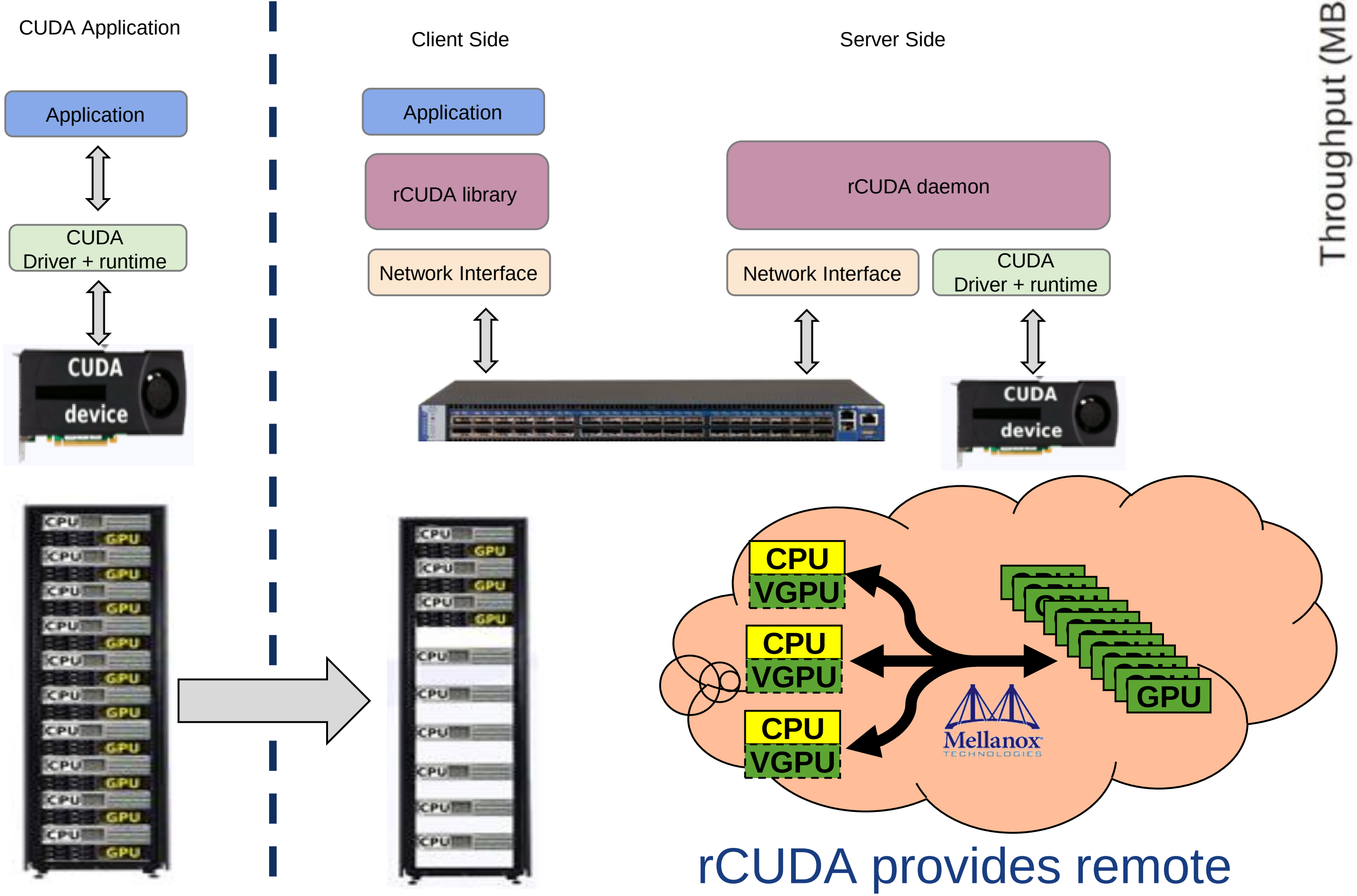
88% Lower Latency

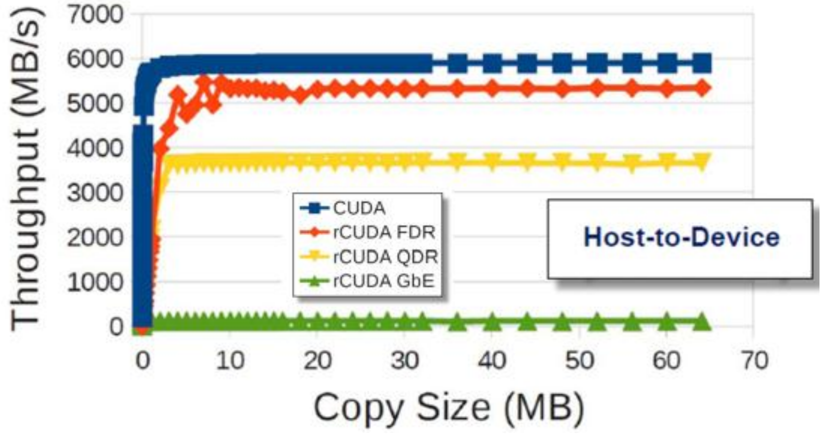
10X Increase in Throughput

Remote GPU Access through rCUDA

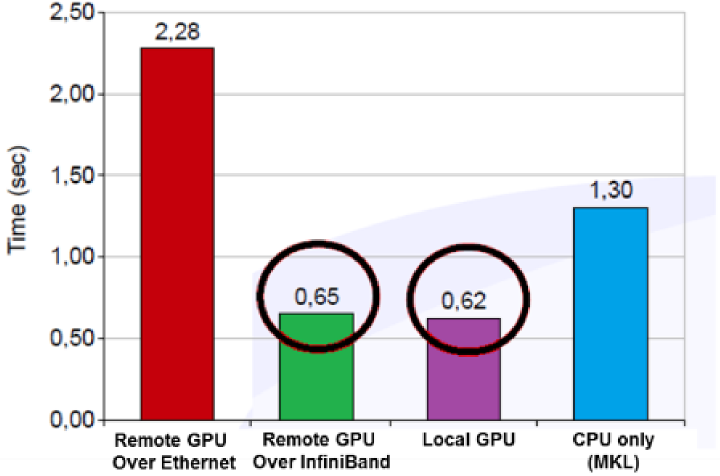
GPU servers

GPU as a Service



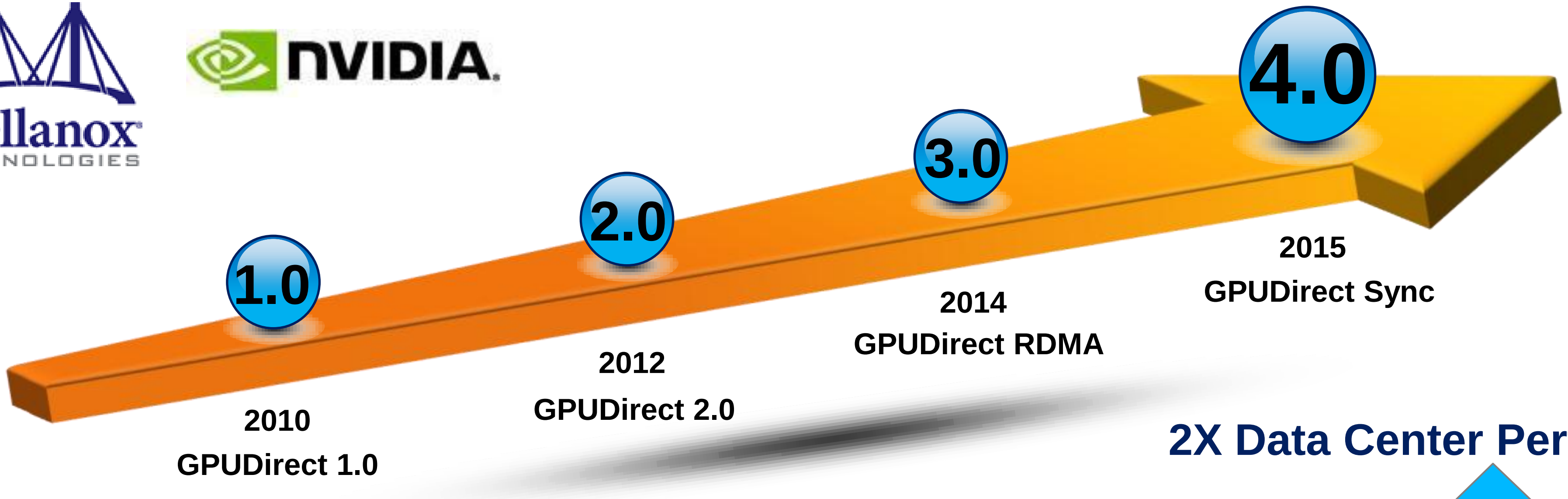


Time for matrix-matrix product (4096x4096)



rCUDA provides remote access from every node to any GPU in the system

Mellanox GPUDirect 기술을 통한 GPU기반 어플리케이션 가속화 솔루션

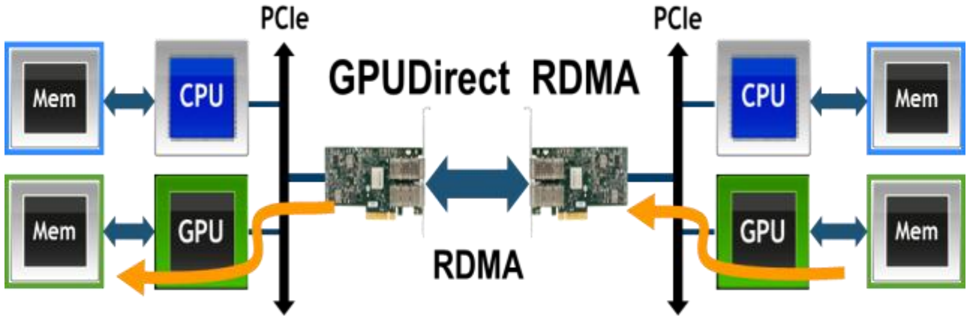


2010
GPUDirect 1.0

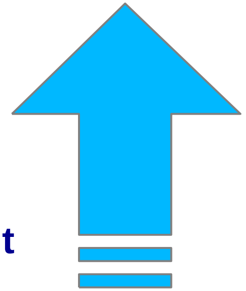
2012
GPUDirect 2.0

2014
GPUDirect RDMA

2015
GPUDirect Sync



2X Data Center Performance



5X Higher Throughput

5X Lower Latency

UNIVERSITY OF CAMBRIDGE

GPU TECHNOLOGY CONFERENCE

THANK YOU

JOIN THE CONVERSATION

#GTC15