GPU Direct Accelerate HPC & AI System

Ashrut Ambastha

Sr. Stuff of Solution Architect

Mar, 2019

InfiniBand Accelerates Leading HPC and AI Systems

World's Top 3 Supercomputers

OAK RIDGE National Laboratory

U.S. DEPARTMENT OF ENERGY

1

Summit CORAL System
World's Fastest HPC / AI System
9.2K InfiniBand Nodes

TOP 500 The List.

Lawrence Livermore National Laboratory

U.S. DEPARTMENT OF ENERGY

2

Sierra CORAL System
#2 USA Supercomputer
8.6K InfiniBand Nodes

TOP 500 The List.

国家超级计算无锡中心
National Supercomputing Center in Wuxi

3

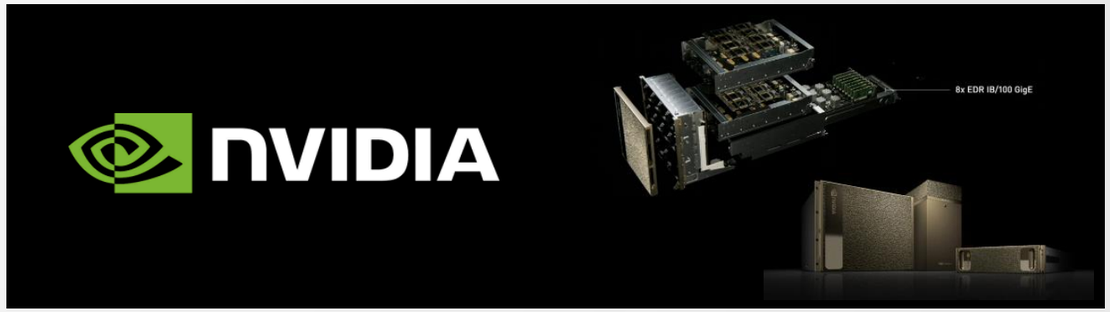
Wuxi Supercomputing Center
Fastest Supercomputer in China
41K InfiniBand Nodes

TOP 500 The List.

Mellanox Accelerates Leading AI Platforms
(Examples)

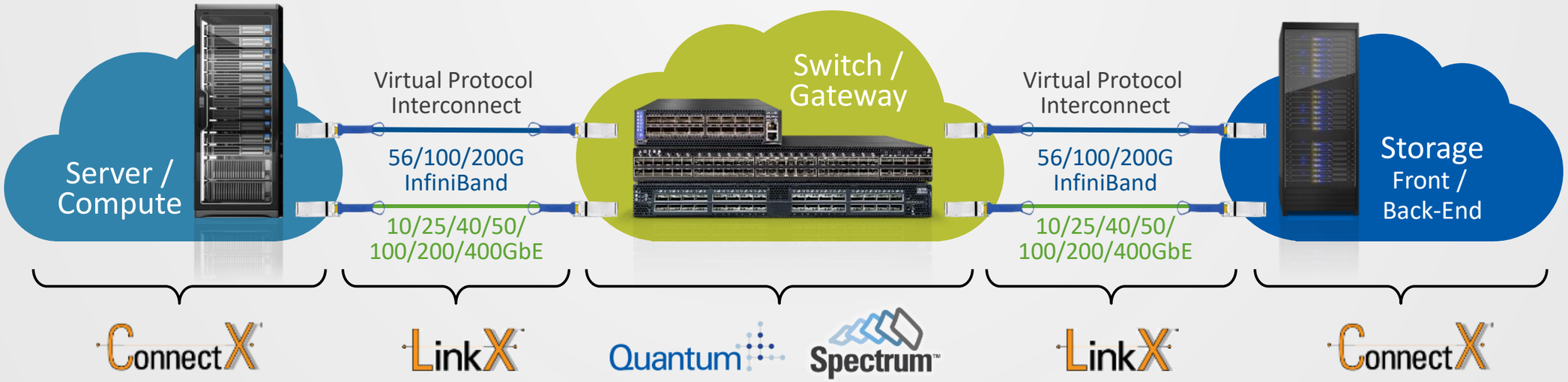
Fastest AI Supercomputer in the World
~27000 GPUs
Dual-Rail Mellanox 100G EDR InfiniBand

NVIDIA® DGX-1 and DGX-2 Platforms
2 Petaflops of AI Performance
Mellanox 100G InfiniBand and Ethernet

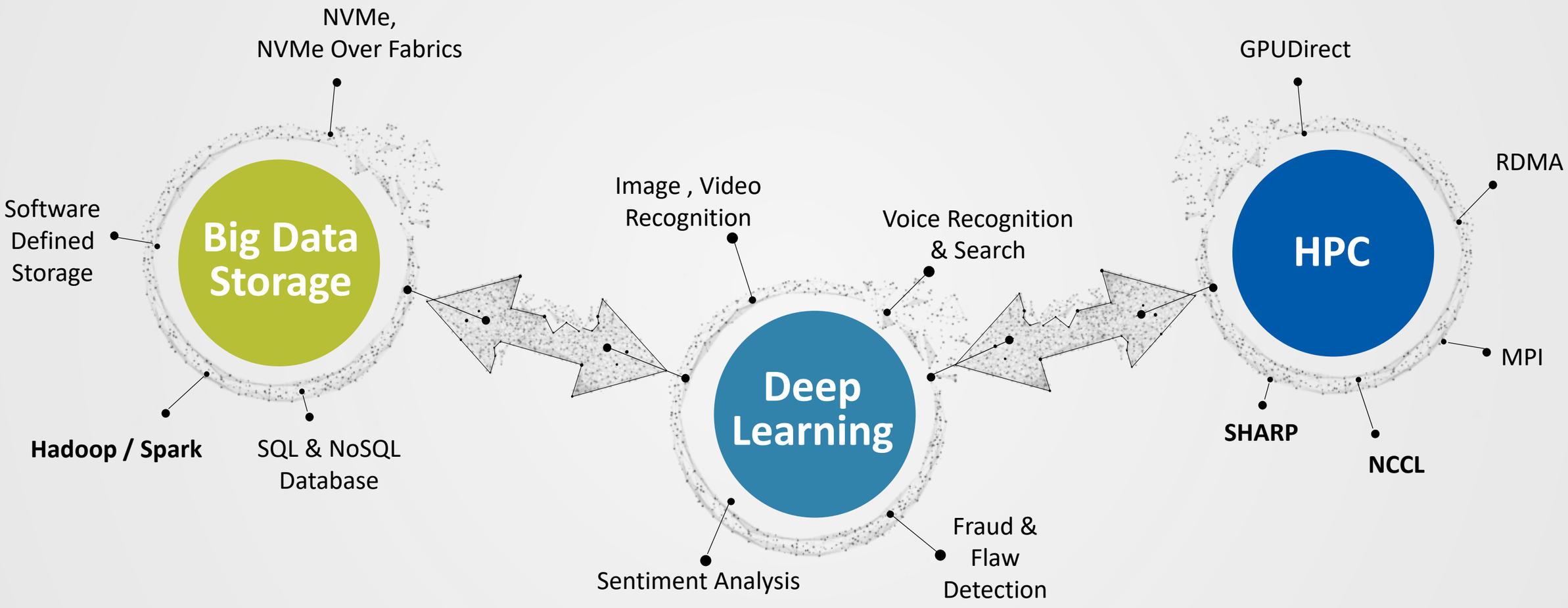


Ranked 50 on the TOP500 list
Facebook AI infrastructure
Mellanox 100G EDR InfiniBand

Leading Supplier of End-to-End Interconnect Solutions

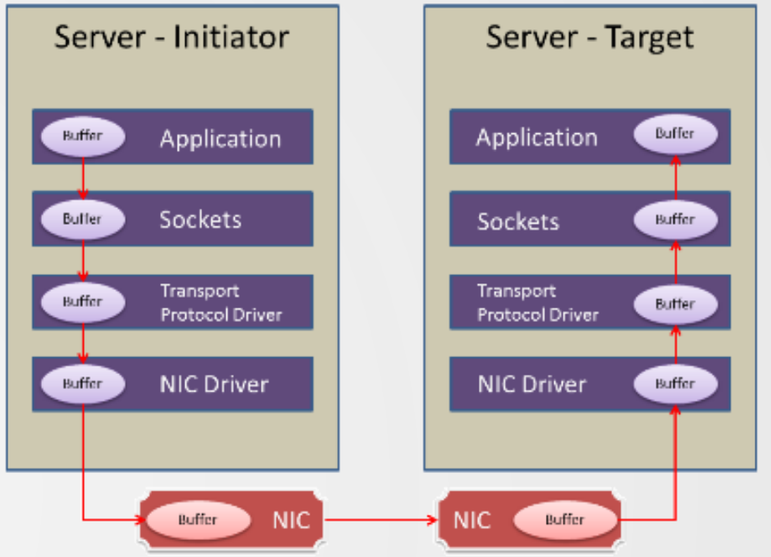


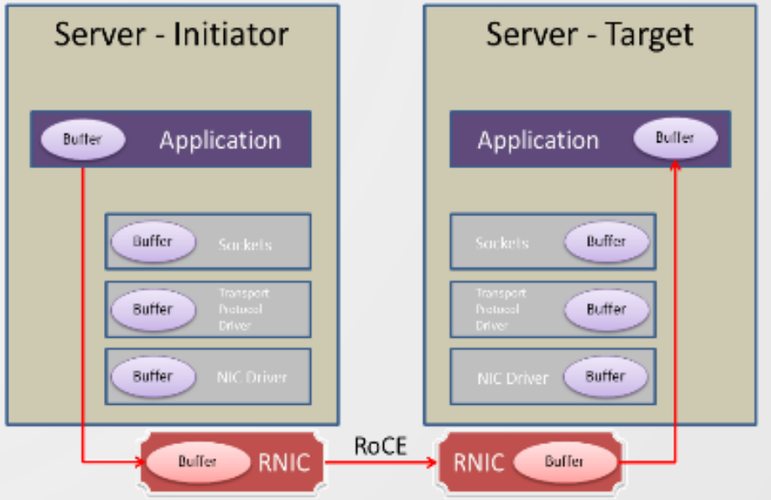
Same Interconnect Technology Enables a Variety of Applications



Unbeatable Performance with RDMA

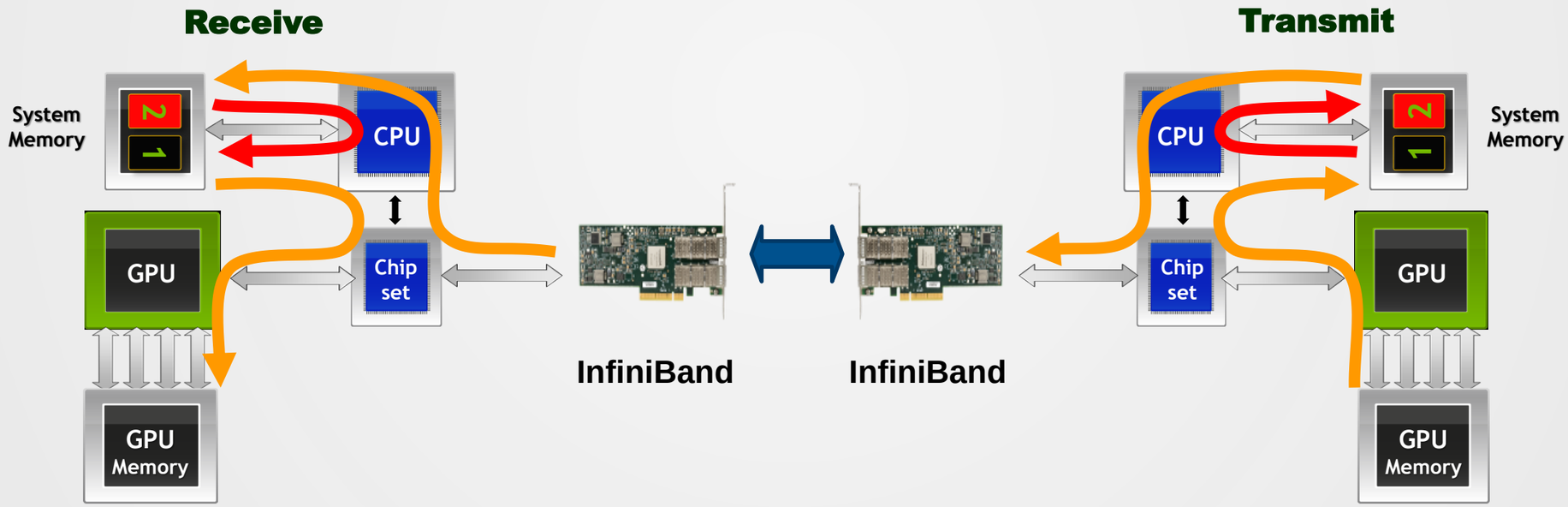
- Remote Direct Memory Access (RDMA)
- Advance transport protocol (same layer as TCP and UDP)
- Main features
  - Remote memory read/write semantics in addition to send/receive
  - Kernel bypass / direct user space access
  - Full hardware offload for network stack
  - Secure, channel based IO
- Application Advantage
  - Lowest latency
  - Highest bandwidth
  - Lowest CPU consumption
- Verbs: RDMA SW Interface (Equivalent to Sockets)





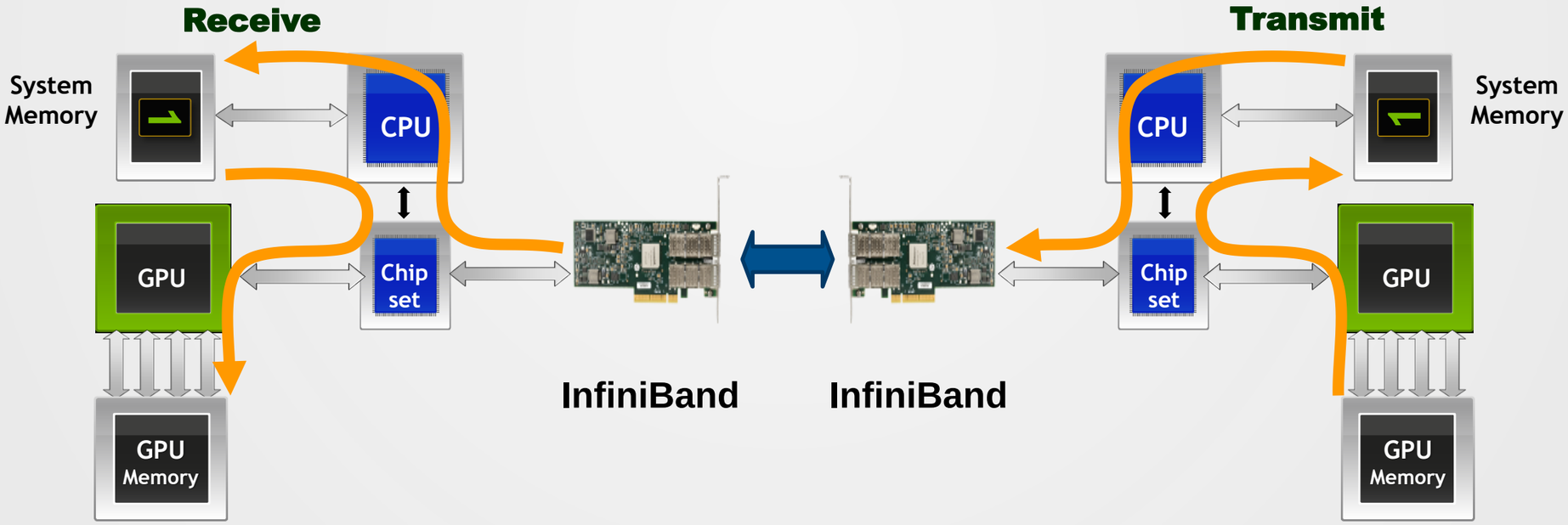
GPU-InfiniBand Bottleneck (pre-GPUDirect)

- Pre-GPUDirect, GPU communications required CPU involvement in the data path
  - Memory copies between the different “pinned buffers”
  - Slow down the GPU communications and creates communication bottleneck



Accelerating GPU Based Supercomputing

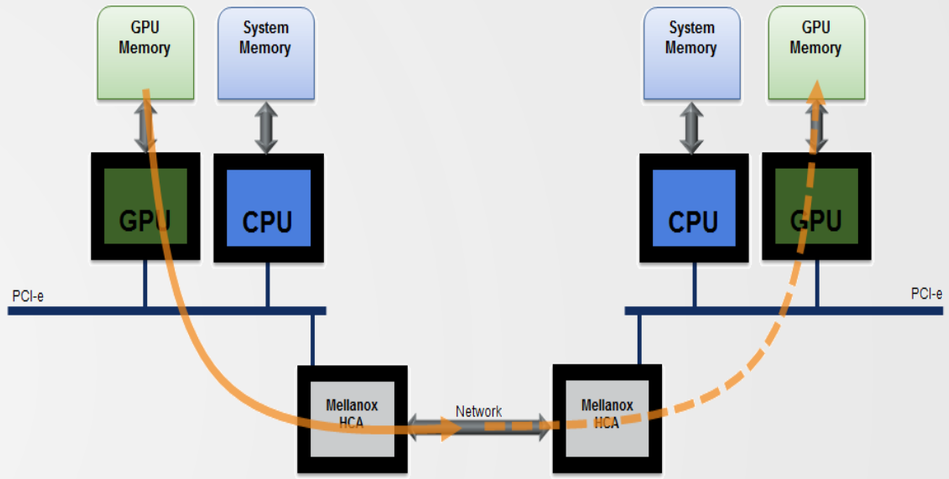
- Fast GPU to GPU communications
- Native RDMA for efficient data transfer
- Reduces latency by 30% for GPUs communication

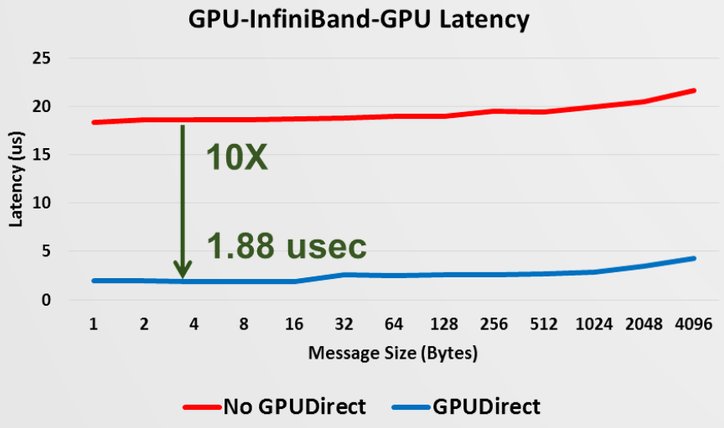


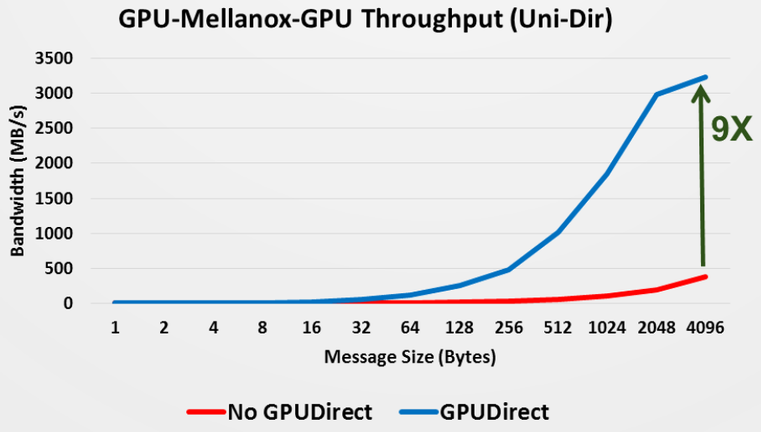
10X Higher Performance with GPUDirect™ RDMA

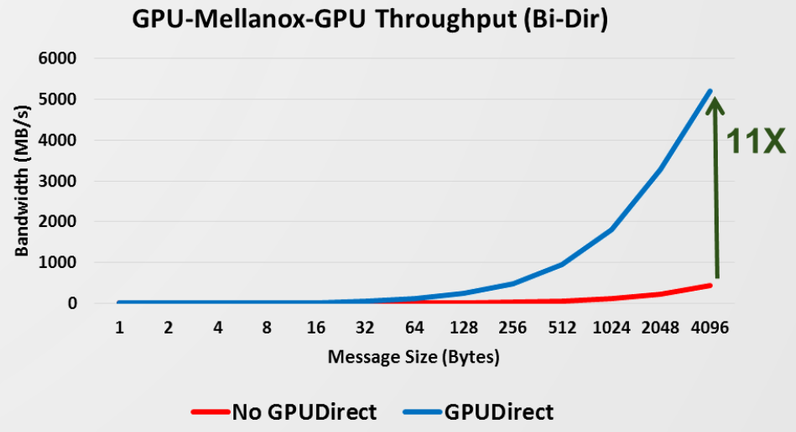
- Accelerates HPC and Deep Learning performance
- Lowest communication latency for GPUs

GPUDirect™ RDMA





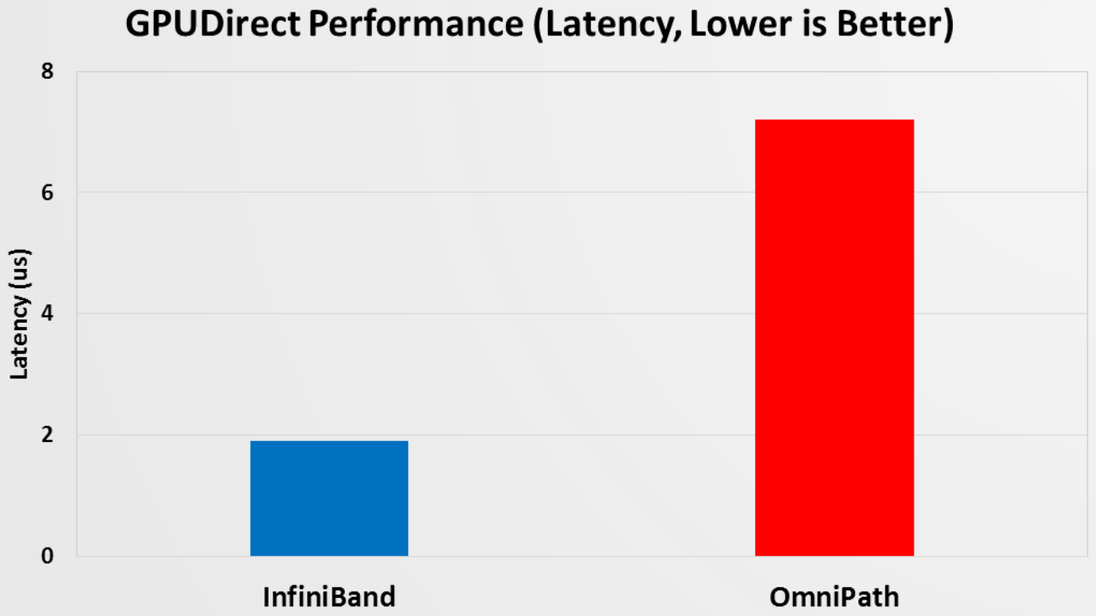


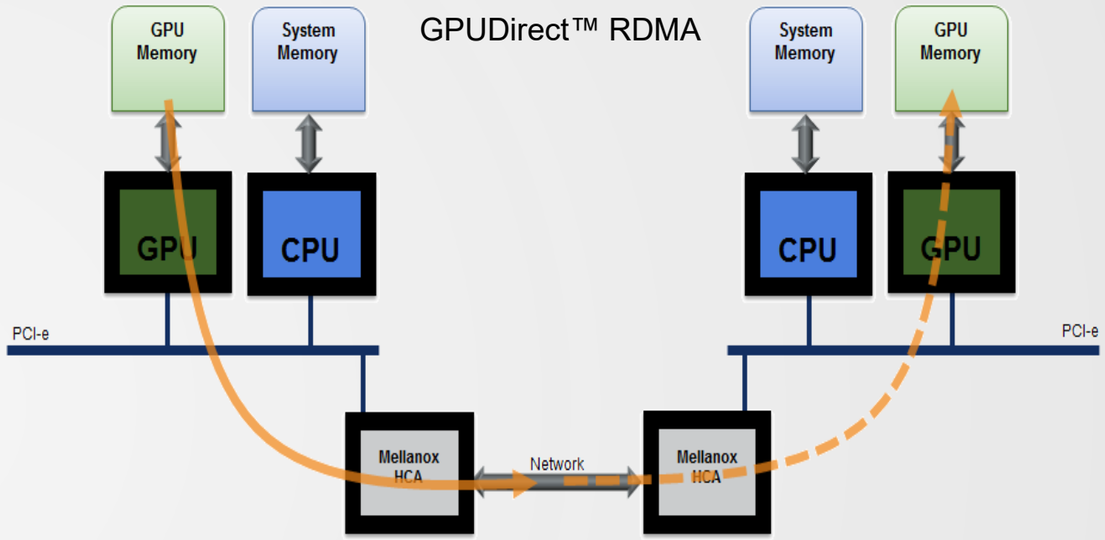


InfiniBand Enables Highest GPU Systems Performance

GPUDirect Enables Best GPU Performance

Requires Native RDMA Network





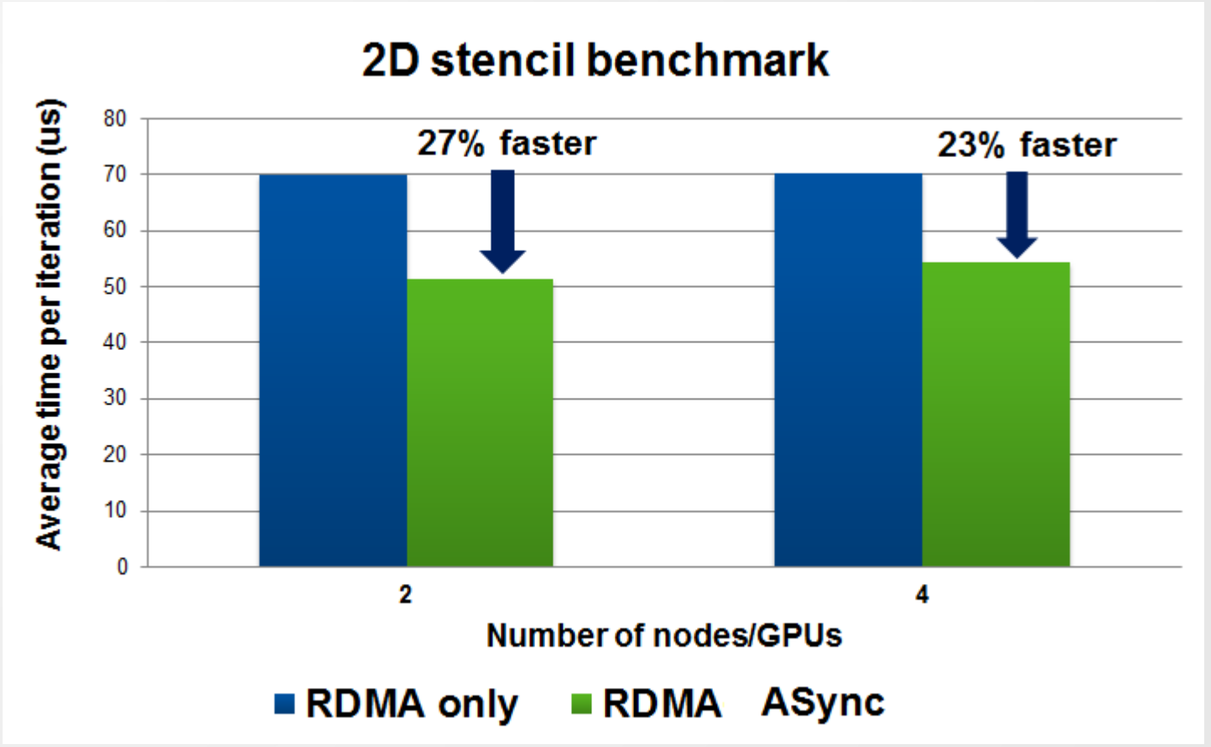
3.8X Better Latency versus Omni-Path

Omni-Path Has No Native RDMA Capabilities

GPUDirect™ ASYNC

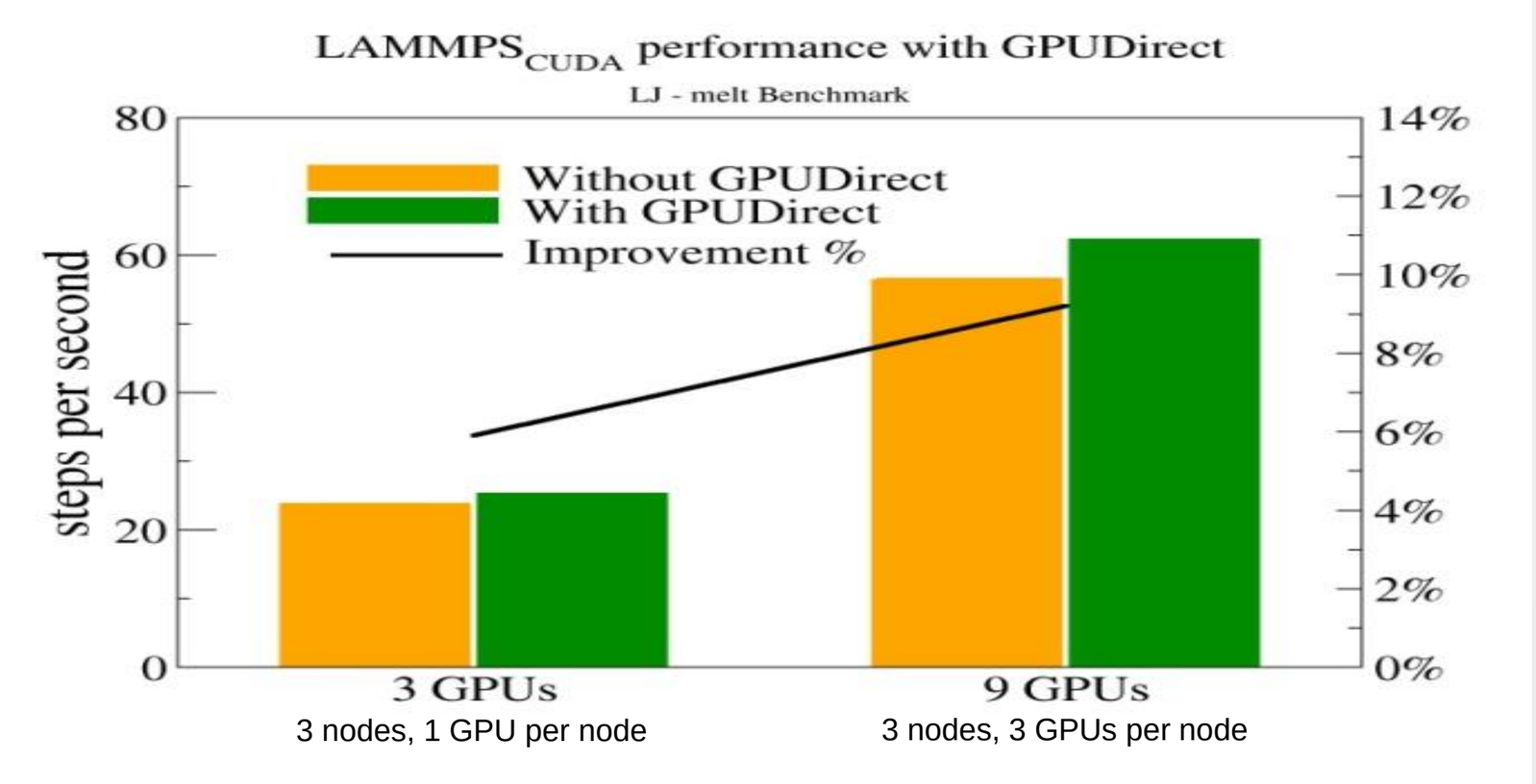
- GPUDirect RDMA (3.0) – direct data path between the GPU and Mellanox interconnect
  - Control path still uses the CPU
    - CPU prepares and queues communication tasks on GPU
    - GPU triggers communication on HCA
    - Mellanox HCA directly accesses GPU memory
- GPUDirect ASYNC (GPUDirect 4.0)
  - Both data path and control path go directly between the GPU and the Mellanox interconnect

Maximum Performance
For GPU Clusters



LAMMPS Performance with GPUDirect

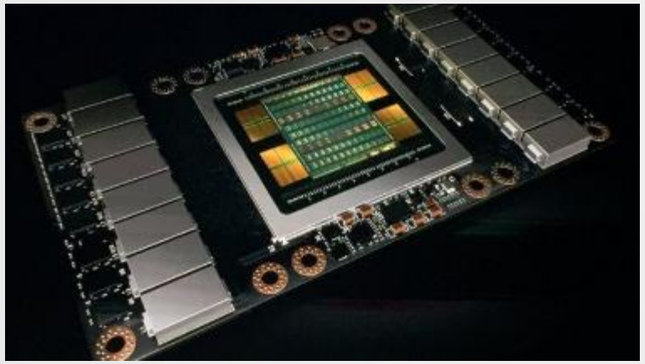
- GPUDirect accelerates LAMMPS by 10% at 3 nodes/9 GPUs
- Performance increases with cluster size - expect to reach 35% at larger size
- Performance increases with problem size



NVIDIA® DGX-1™ Deep Learning Server

NVIDIA® “SaturnV”

NVIDIA® Machine Learning Supercomputer
5280 Tesla V100 GPUs
660Pf (AI) with 660 DGX-1 nodes
80Pf (FP32) / 40Pf (FP64)



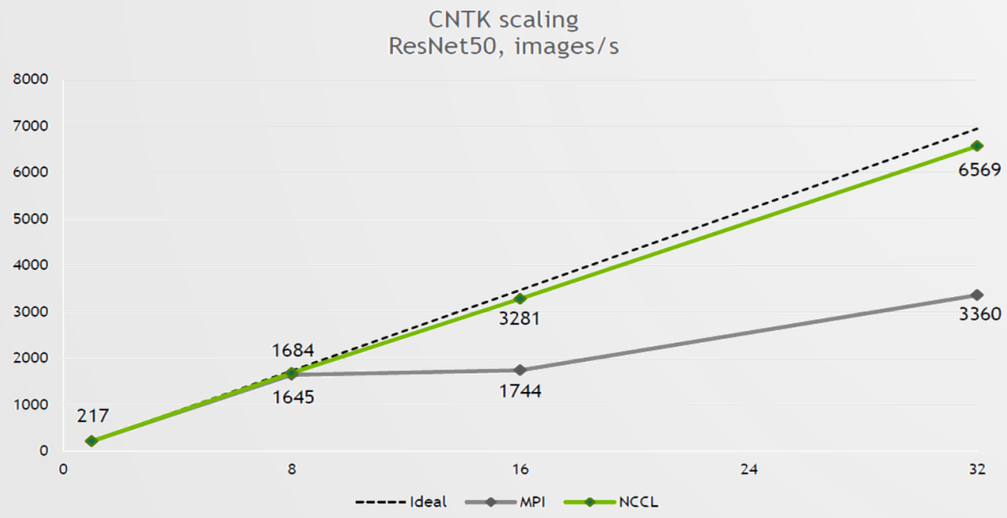
8 x NVIDIA Volta V100 GPUs

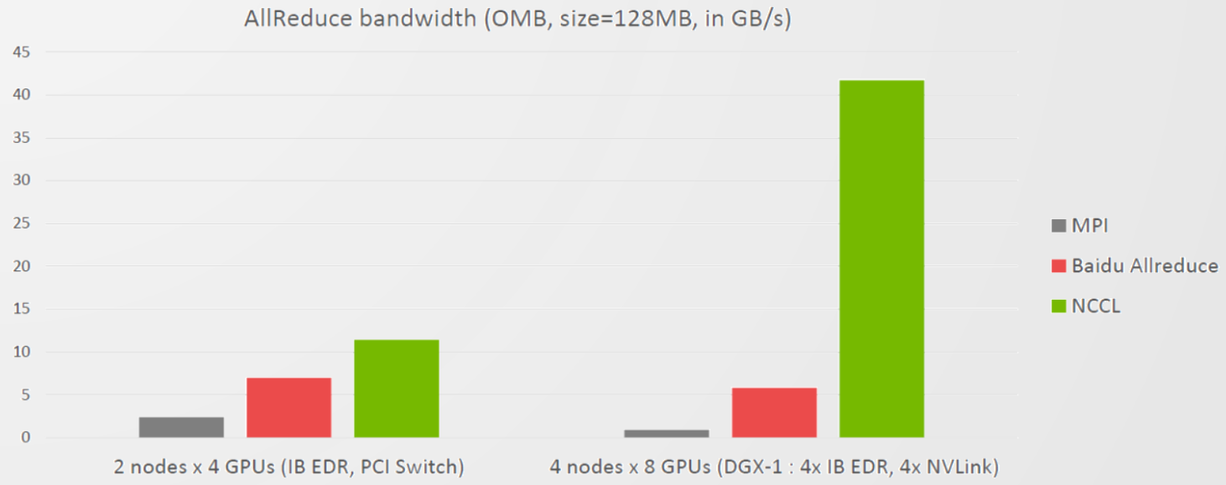
4x ConnectX-4 EDR 100G InfiniBand Adapters

Mellanox Accelerates NVIDIA NCCL 2.0

50% Performance Improvement

with NVIDIA® DGX-1 across
32 NVIDIA Tesla V100 GPUs
Using InfiniBand RDMA
and GPUDirect™ RDMA





Big Sur – An Open Artificial Intelligence Platform (Facebook)

- An OCP based, GPU Artificial Intelligence Platform
- Flexible Architecture supporting up to 8 GPUs
  - NVIDIA®, Intel®, AMD
- Mellanox Ethernet adapters
  - RDMA supported (RoCE)
- Accelerating Artificial Intelligence
  - Text Processing
  - Language Modeling
  - Computer Vision

https://code.facebook.com/posts/1687861518126048/facebook-to-open-source-ai-hardware-design/

Mellanox Intelligent Network for Intelligent Platform

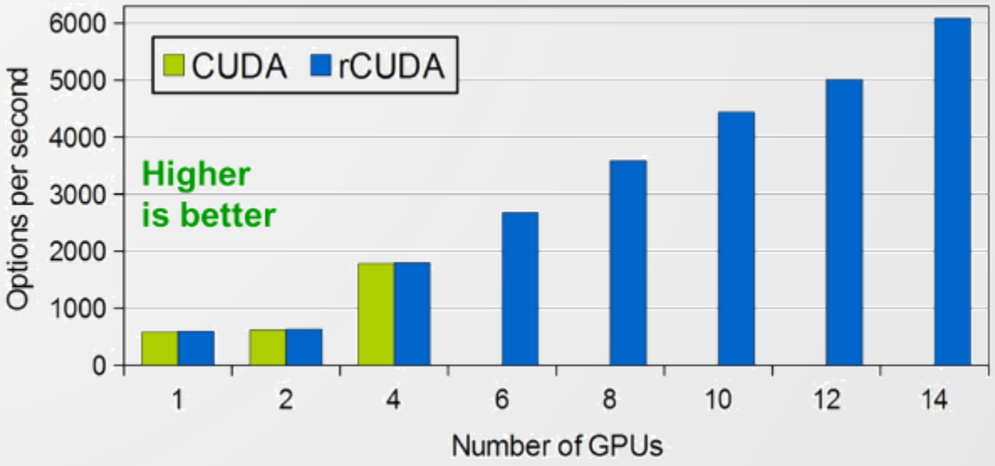
rCUDA Enables GPU Services Over Mellanox Interconnect

Moving from Physical Limitation
to Unlimited Scalability

rCUDA
remote CUDA

- Software that Enables Flexible use of GPUs
- Enable GPU services
- More GPUs available for a single application
- Enables GPU job migration
- Virtual machines can share GPUs

MonteCarlo Simulations



Mellanox Delivers Best Return on Investment for AI Platforms

60%	2.5X	Up to 95%	Up to 50%
Higher ROI	Better Performance	Scaling Efficiency	Savings on Capital & Operation Cost

Thank You