


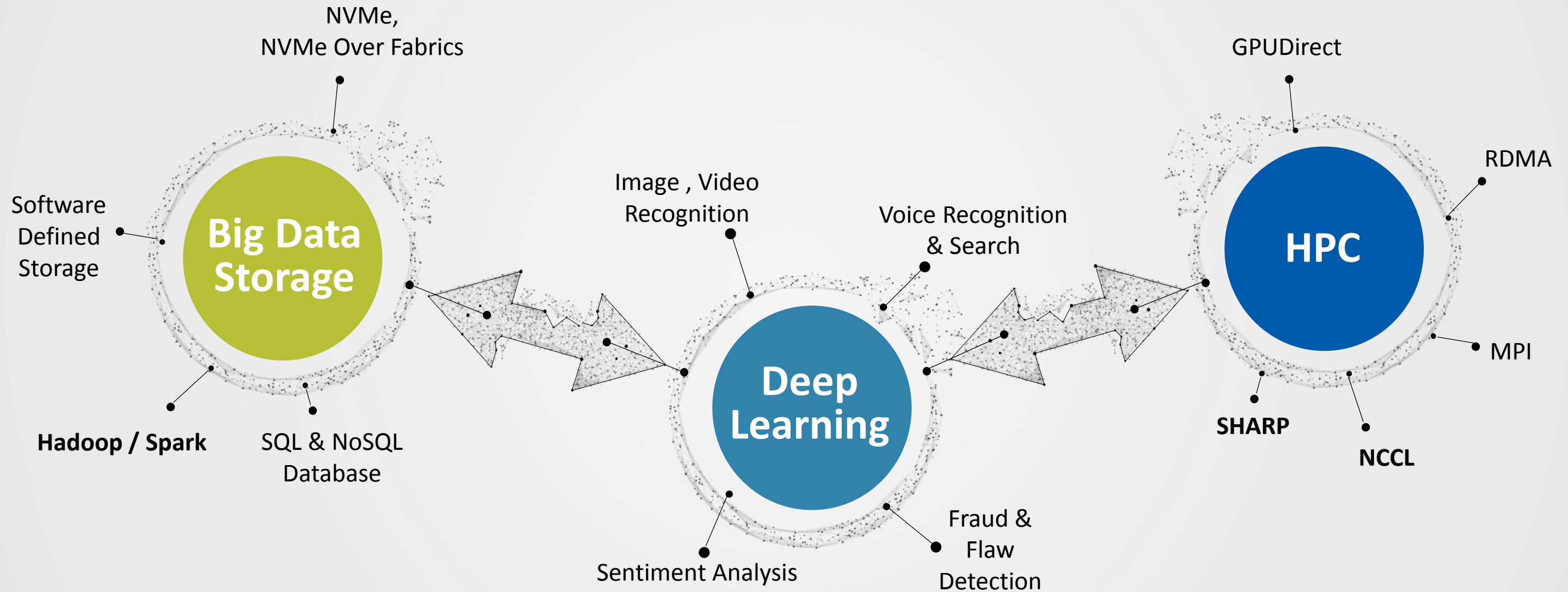


In-Network Computing Enables Next Generation HPC

February 2018



Same Interconnect Technology Enables a Variety of Applications



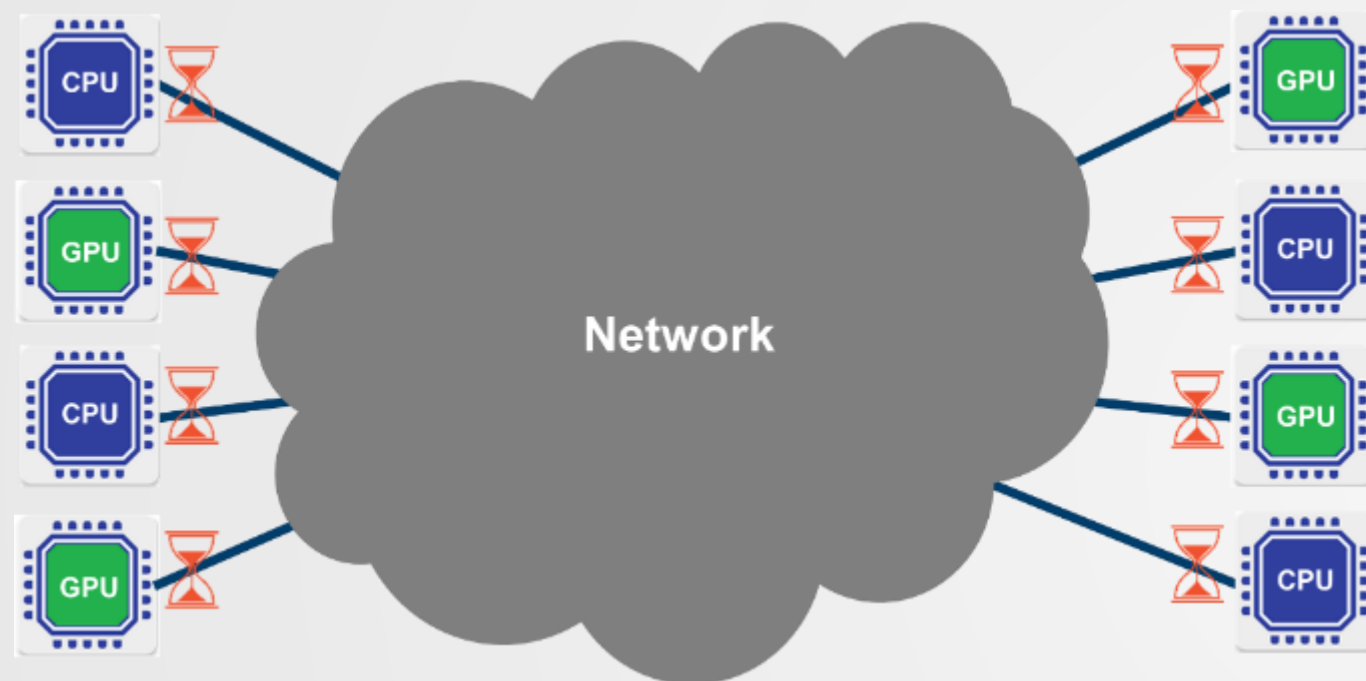
Challenges For Accelerated Computing

- Reduce application time to solution
- Increase overall system efficiency
- Reduce costs
- Opportunities from the network
 - Increase the performance of the communication stack
 - Decrease data motion
 - Processes data that is moving through the network
 - More efficient communication stack
 - Share the computational load
 - Accelerate post-processing



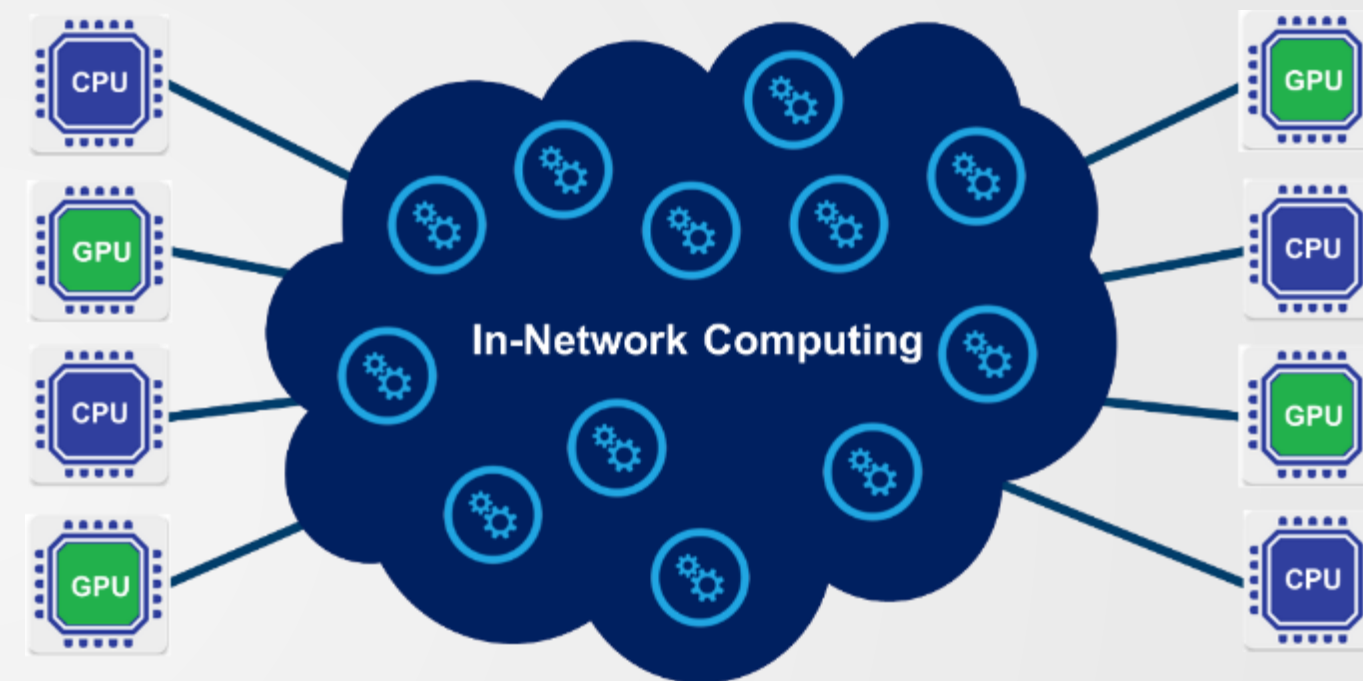
The Need for Intelligent and Faster Interconnect

CPU-Centric (Onload)



Must Wait for the Data
Creates Performance Bottlenecks

Data-Centric (Offload)

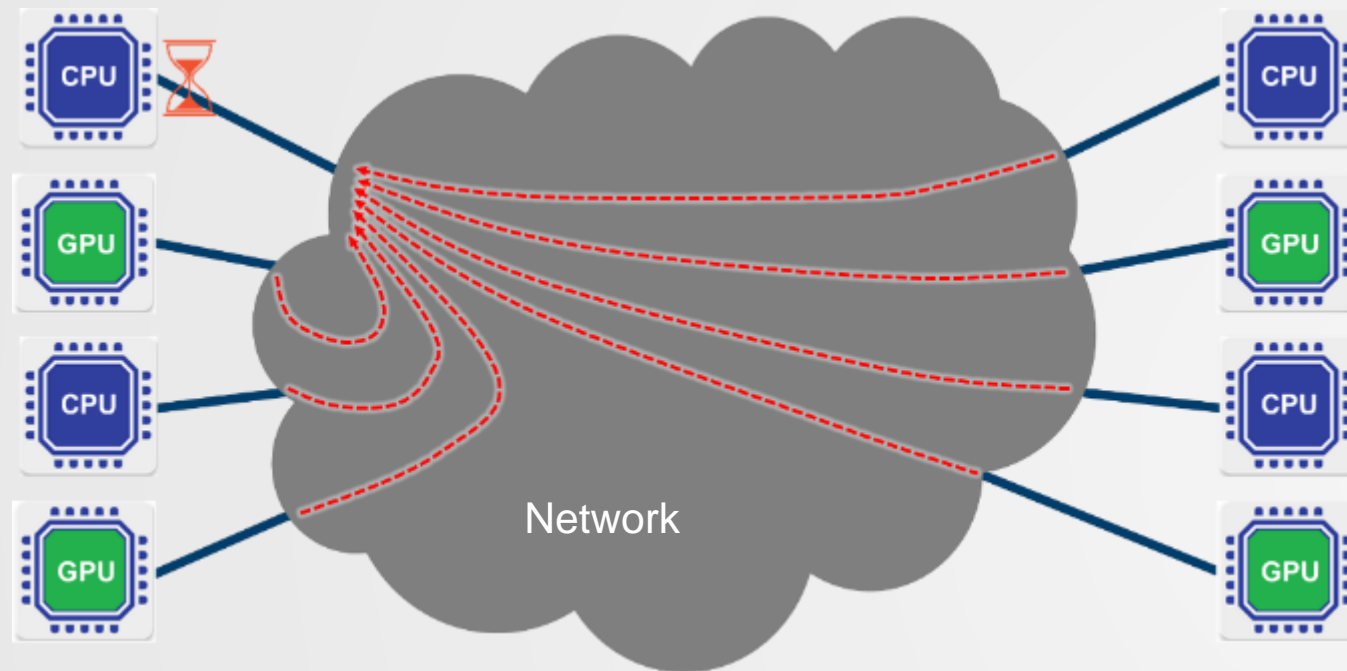


Analyze Data as it Moves!

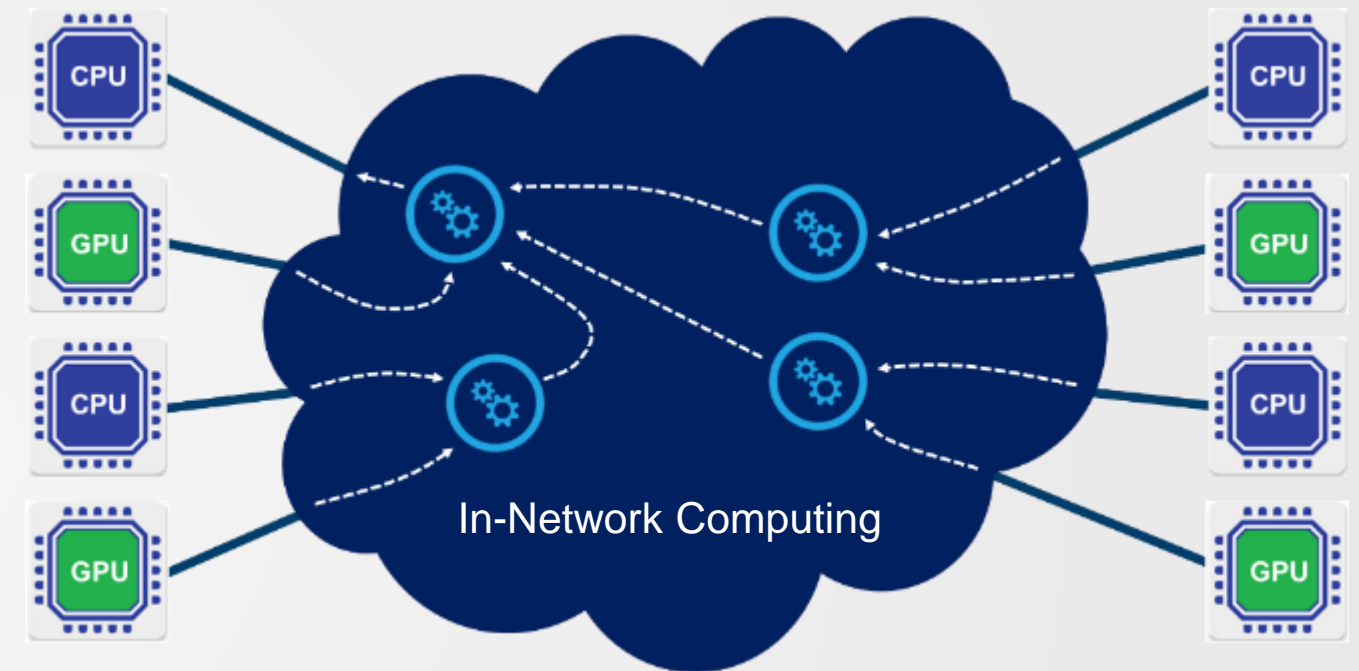
Faster Data Speeds and In-Network Computing Enable Higher Performance and Scale

Data Centric Architecture to Overcome Latency Bottlenecks

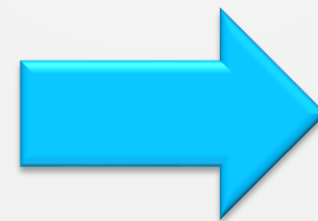
CPU-Centric (Onload)



Data-Centric (Offload)



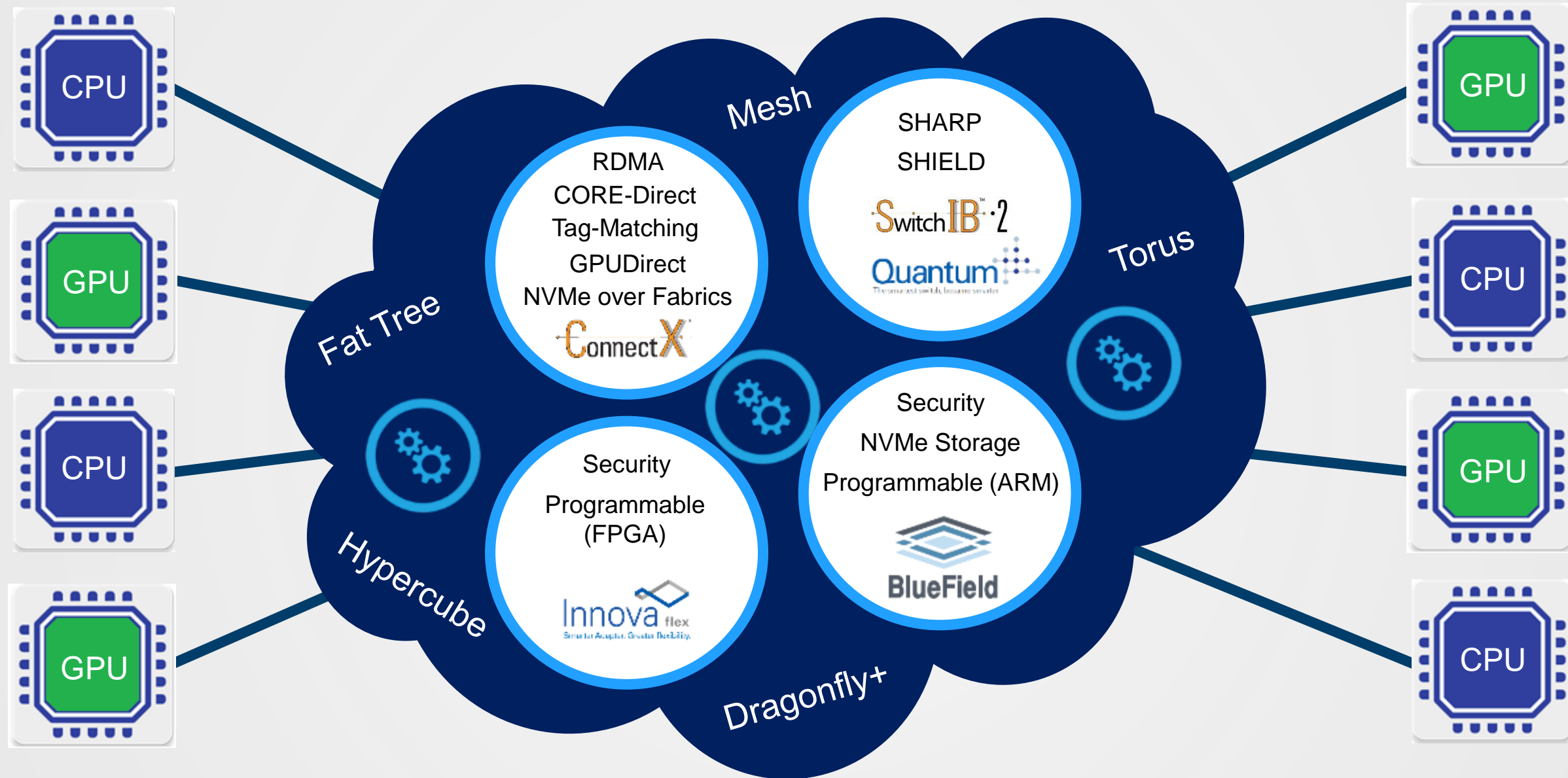
Communications Latencies of 30-40us



Communications Latencies of 3-4us

Intelligent Interconnect Paves the Road to Exascale Performance

In-Network Computing to Enable Data-Centric Data Centers



In-Network Computing Key for Highest Return on Investment

In-Network Computing Delivers Accelerated Performance



In-Network
Computing



10X

Performance Acceleration
Critical for HPC and Machine Learning Applications



In-Network
Computing



35X

Performance Acceleration
Delivers Highest Application Performance

GPUDirect™ RDMA
GPU Acceleration Technology



10X

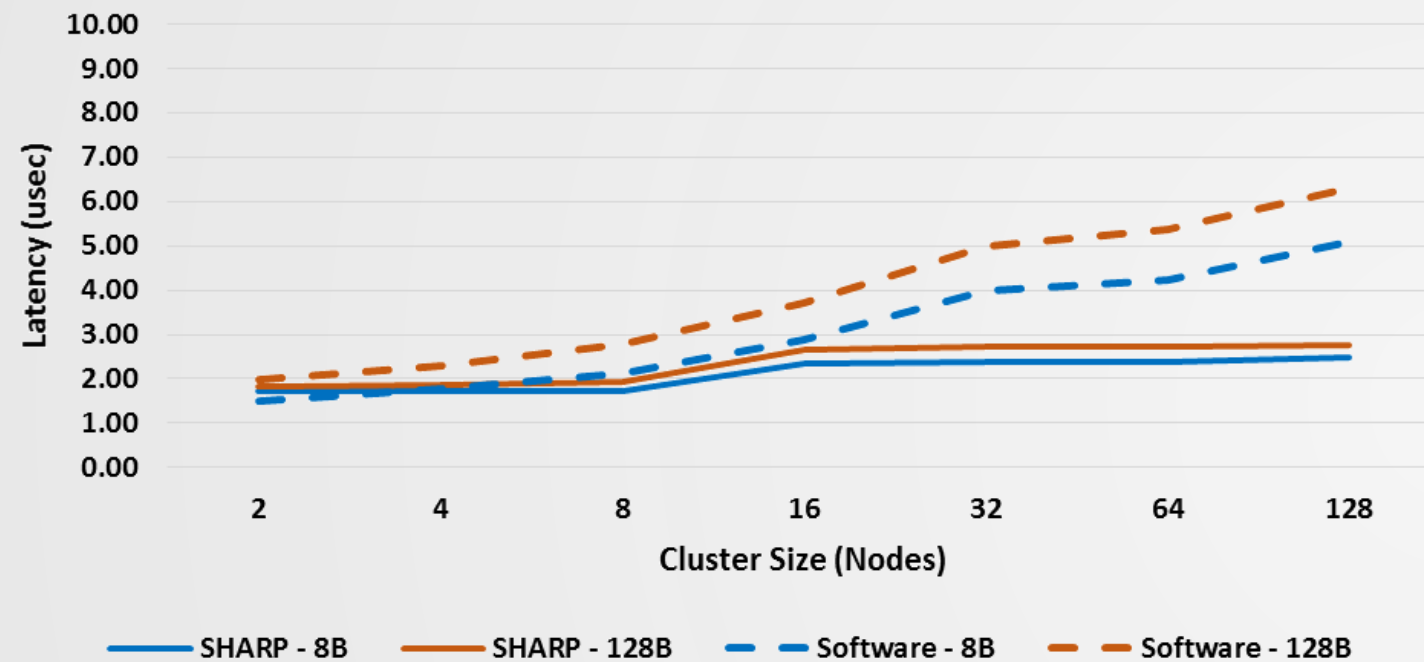
Performance Acceleration
Critical for HPC and Machine Learning Applications

SHARP

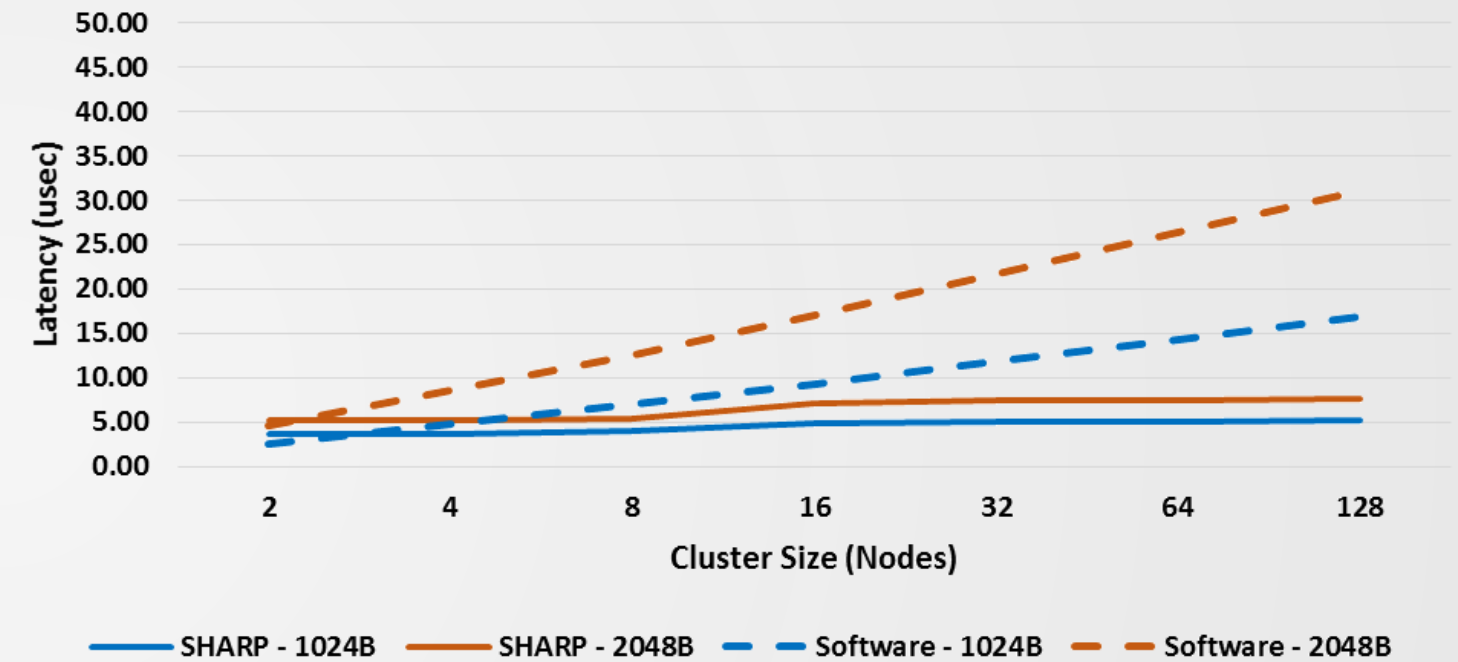


SHARP AllReduce Performance Advantages (120 Nodes)

Allreduce Latency



Allreduce Latency

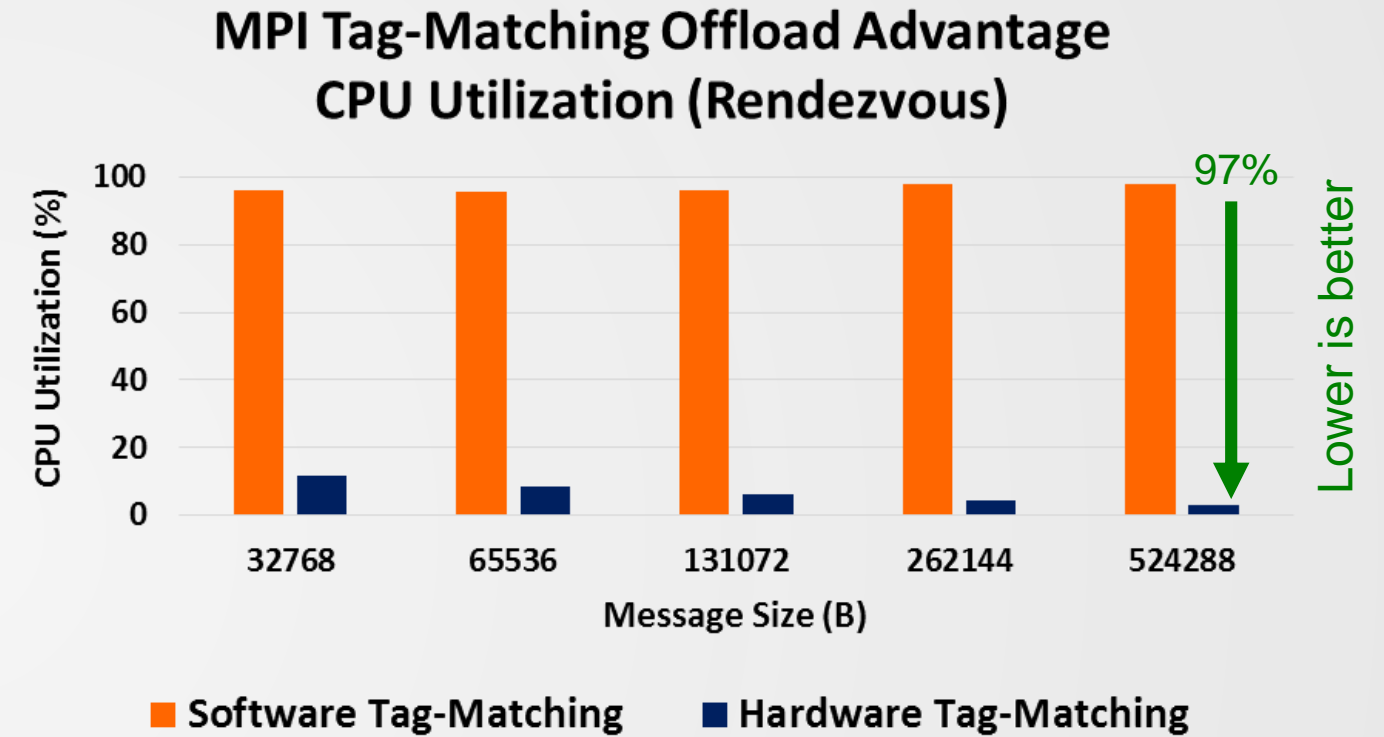
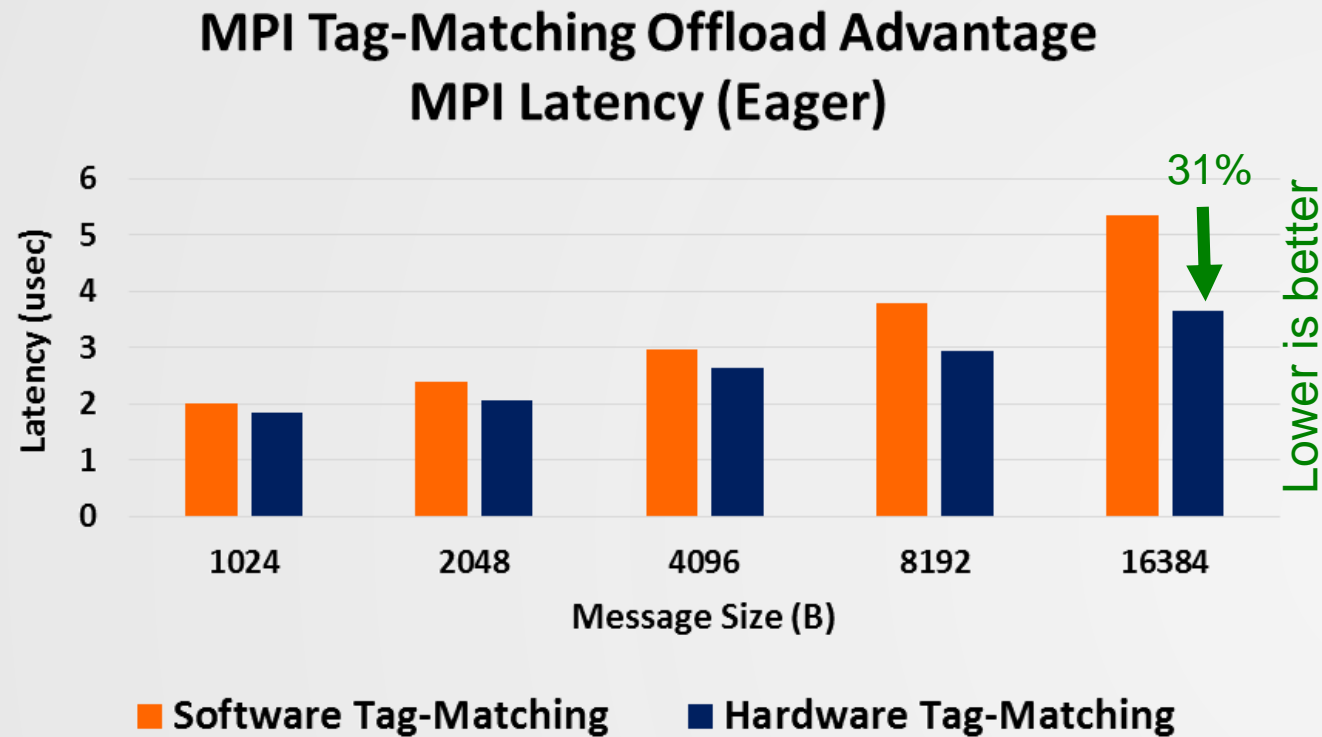


SHARP enables 75% Reduction in Latency
Providing Scalable Flat Latency

MPI Tag Matching



MPI Tag-Matching Offload Advantages



- 31% lower latency and 97% lower CPU utilization for MPI operations
- Performance comparisons based on ConnectX-5

Mellanox In-Network Computing Technology Deliver Highest Performance

GPUDirect – Accelerator Support



Performance of MPI with GPUDirect RDMA

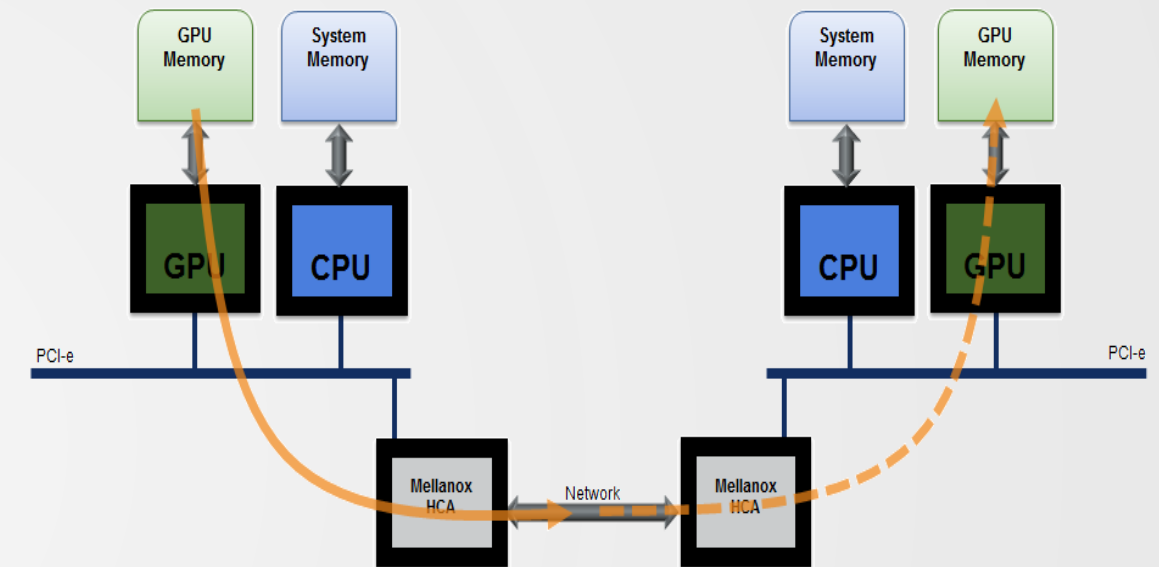
10X

Higher Performance
with GPUDirect™

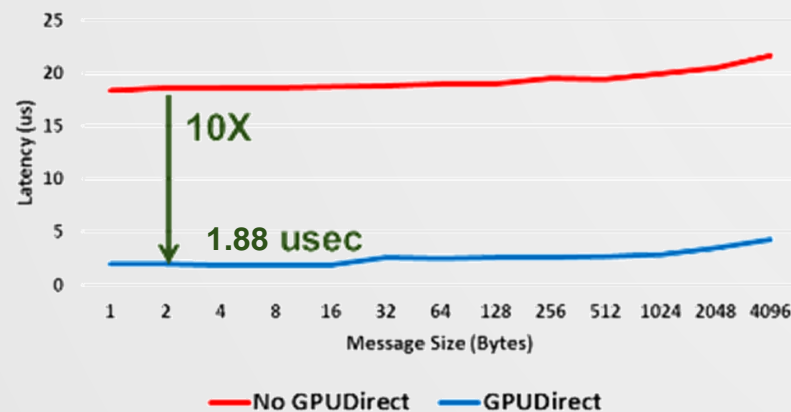


Source: Prof. DK Panda

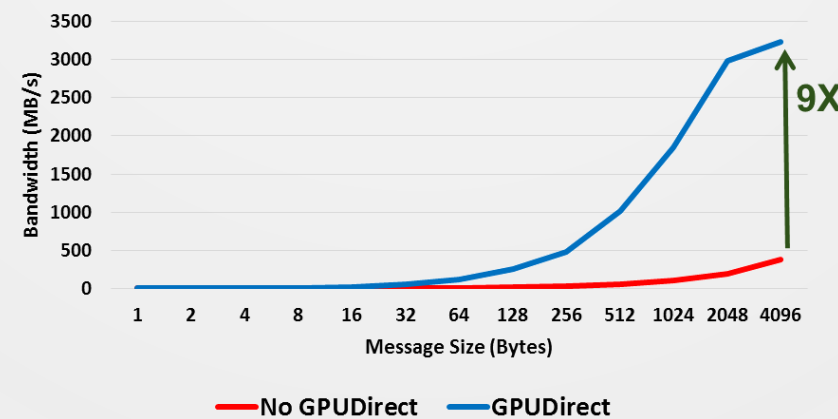
GPUDirect™ RDMA, GPUDirect™ ASYNC



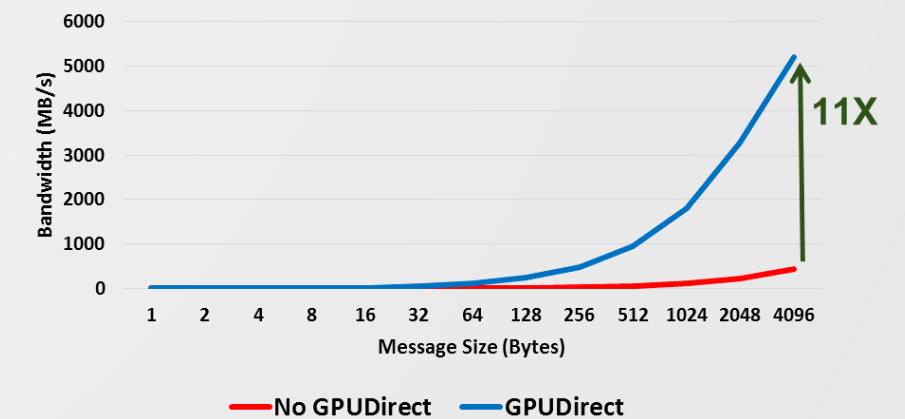
GPU-InfiniBand-GPU Latency



GPU-InfiniBand-GPU Throughput (Uni-Dir)



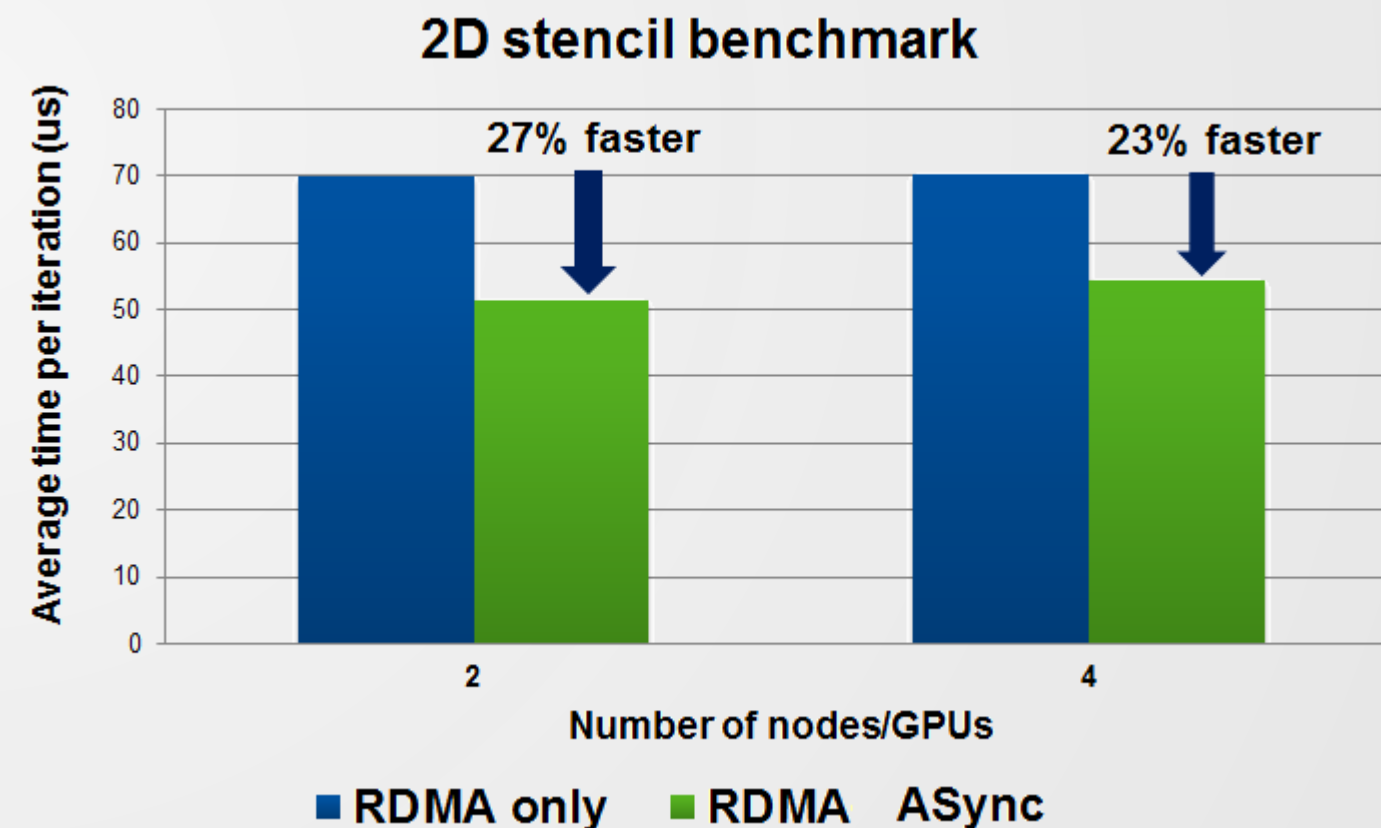
GPU-InfiniBand-GPU Throughput (Bi-Dir)



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





Thank You