Exascale by Co-Design Architecture

Michael Kagan, CTO

June 2016

Mellanox TECHNOLOGIES

Connect. Accelerate. Outperform.™

The Ever Growing Demand for Higher Performance

Performance Development

Terascale

Petascale

1st

"Roadrunner"

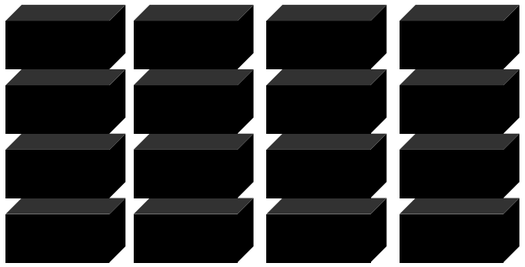
Exascale

OAK RIDGE National Laboratory
"Summit" System

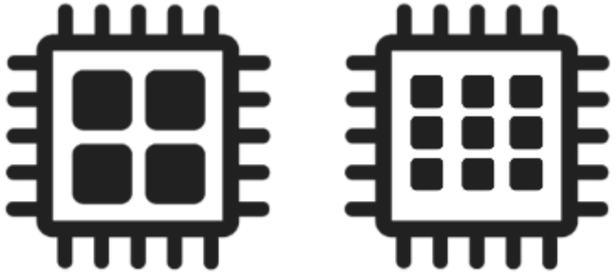
Lawrence Livermore National Laboratory
"Sierra" System

2000 2005 2010 2015 2020

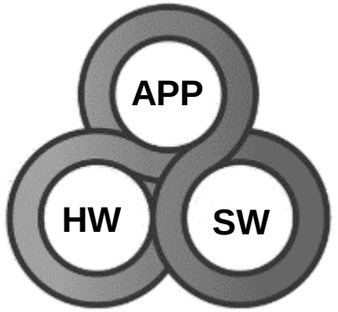
The Interconnect is the Enabling Technology



SMP to Clusters



Single-Core to Many-Core

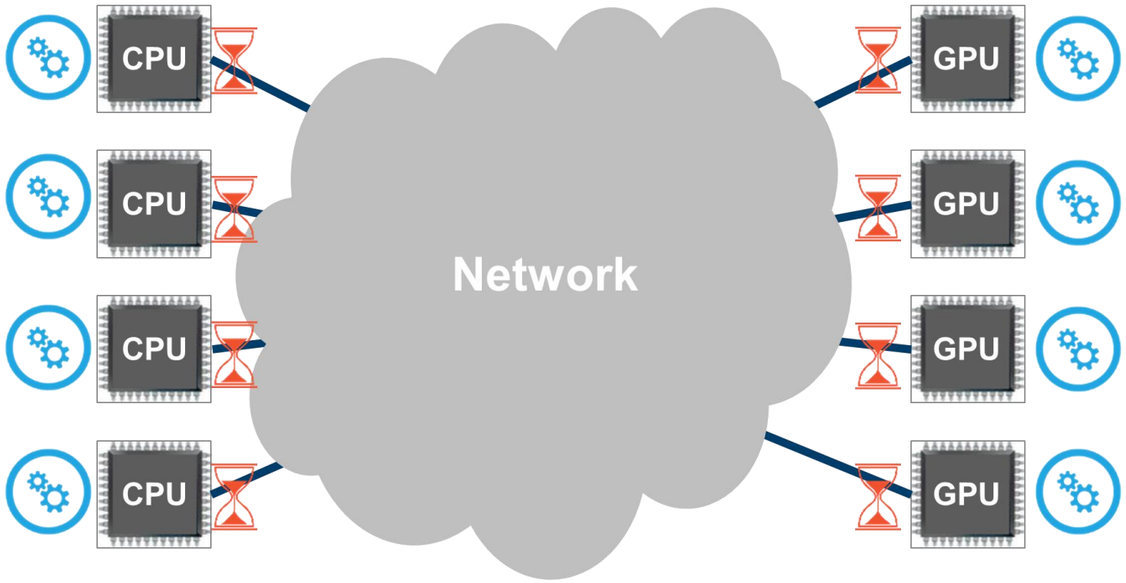


Application
Software
Hardware

Co-Design

The Intelligent Interconnect to Enable Exascale Performance

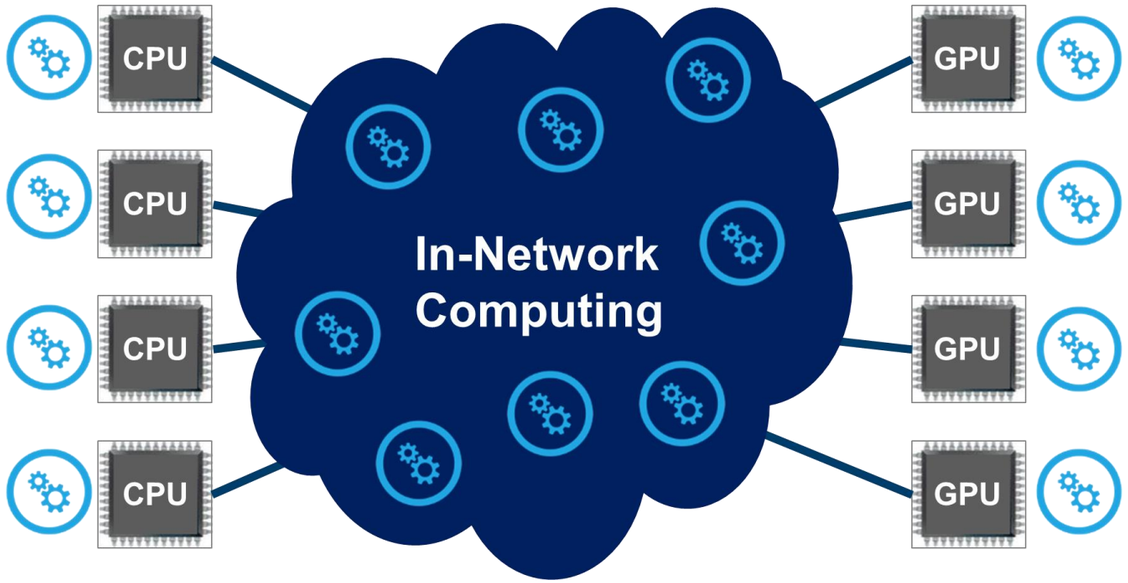
CPU-Centric



Limited to Main CPU Usage
Results in Performance Limitation

Must Wait for the Data
Creates Performance Bottlenecks

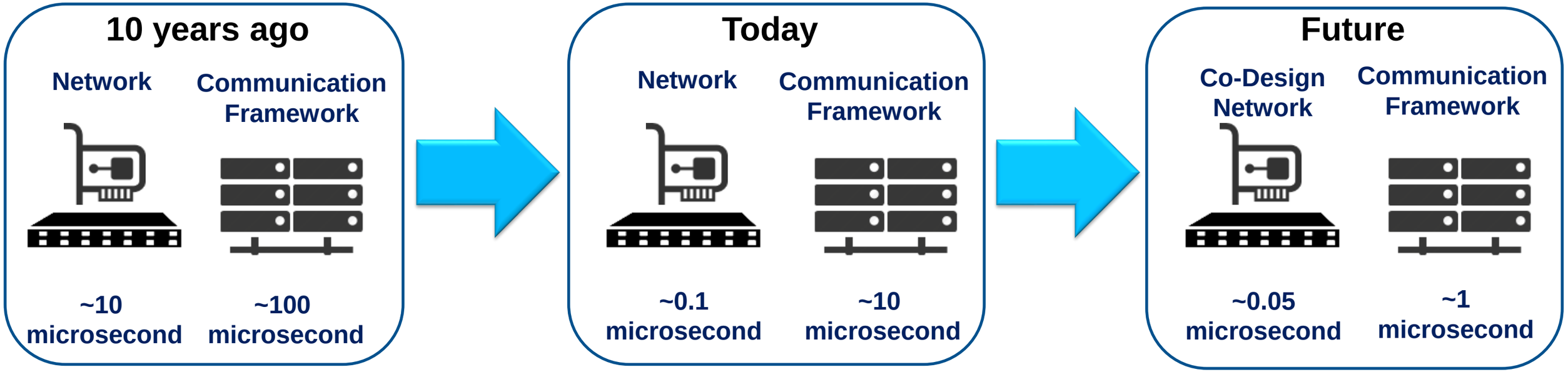
Co-Design



Creating Synergies
Enables Higher Performance and Scale

Work on The Data as it Moves
Enables Performance and Scale

Breaking the Application Latency Wall



- Today: Network device latencies are on the order of 100 nanoseconds
- Challenge: Enabling the next order of magnitude improvement in application performance
- Solution: Creating synergies between software and hardware – intelligent interconnect

Intelligent Interconnect Paves the Road to Exascale Performance

ISC'16: Introducing ConnectX-5 World's Smartest Adapter

State of the Smart

a new generation of co-processors emerges

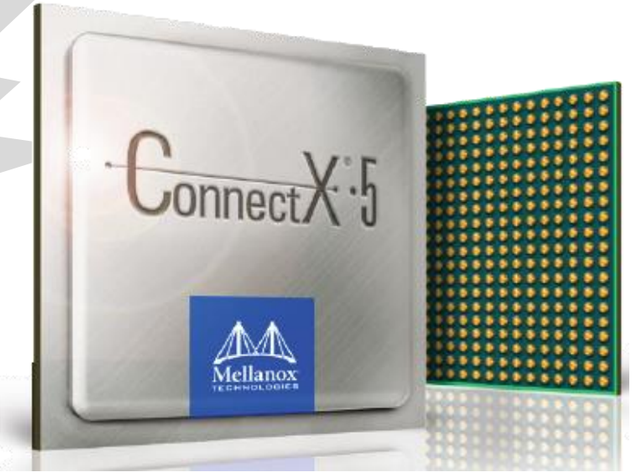
Mellanox Smart Interconnect

Switch-IB 2

SHArP

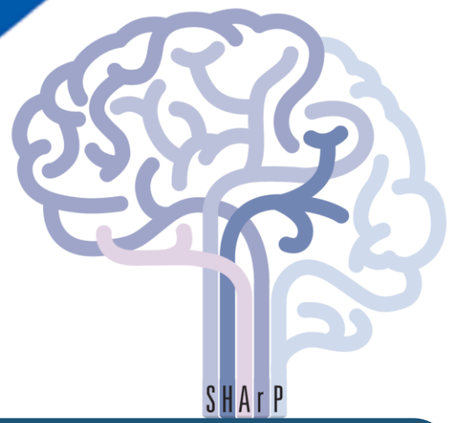
NEW!

ConnectX-5



ConnectX-5 EDR 100G Advantages

State of the Smart



Switch-IB™ 2 SHArP

SHArP Enables Switch-IB 2 to Manage and Execute MPI Operations in the Network

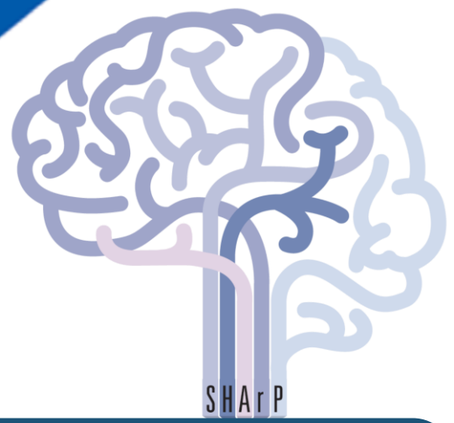
Switch-IB 2 Enables the Switch Network to Operate as a Co-Processor

Delivering 10X Performance Improvement for MPI and SHMEM/PAGS Communications

ConnectX-5 EDR 100G Advantages

State of the Smart

NEW!



Performance
- 100Gb/s Throughput
- 0.6usec Latency (end-to-end)
- 200M Messages per Second

Smart
- MPI Collectives in Hardware
- MPI Tag Matching in Hardware
- In-Network Memory

Platform
- PCIe Gen3 and Gen4
- Integrated PCIe Switch
- Advanced Dynamic Routing

Highest-Performance 100Gb/s Interconnect Solutions

Adapters

ConnectX®-5

100Gb/s Adapter, 0.6us latency

200 million messages per second

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

Switch

SwitchIB™-2

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

Throughput of 7.2Tb/s

7.02 Billion msg/sec (195M msg/sec/port)

Switch

Spectrum™

32 100GbE Ports, 64 25/50GbE Ports

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

Throughput of 6.4Tb/s

Interconnect

LinkX™

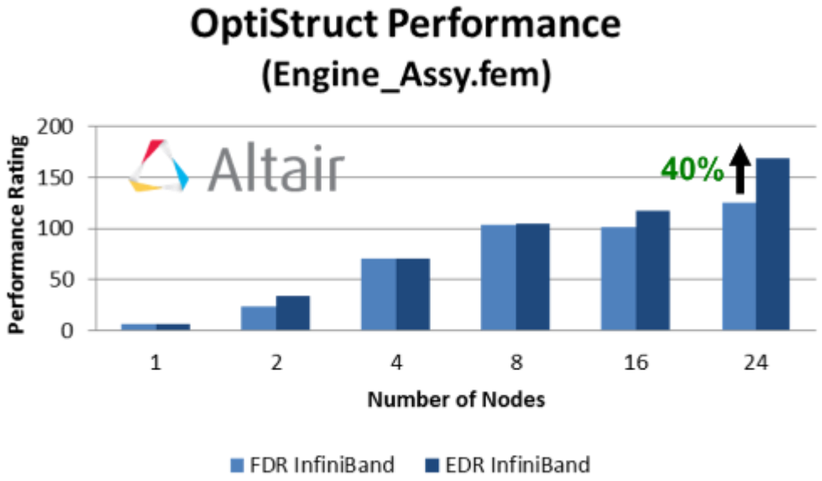
Transceivers

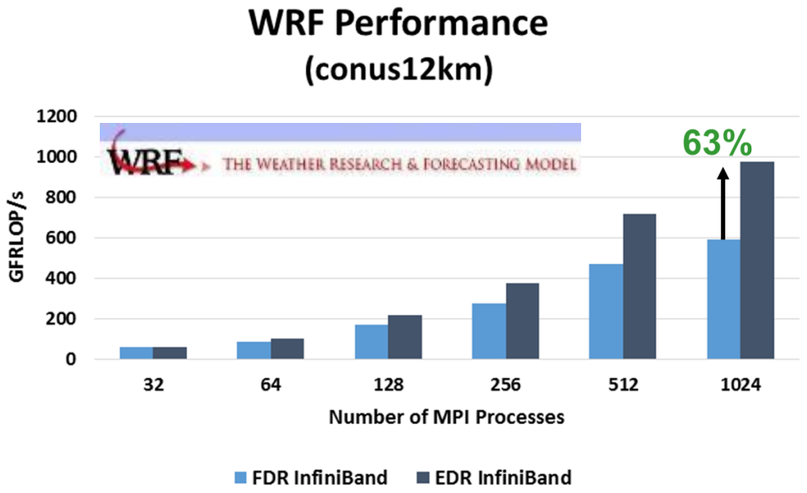
Active Optical and Copper Cables

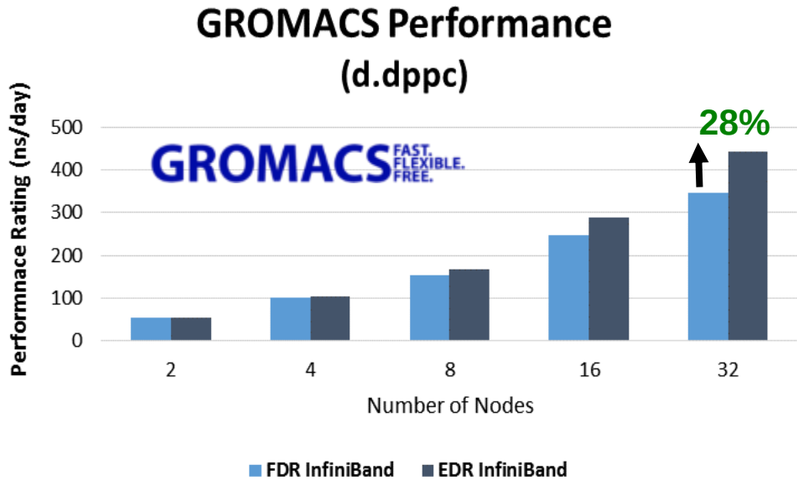
(10 / 25 / 40 / 50 / 56 / 100Gb/s)

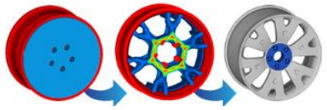
VCSELs, Silicon Photonics and Copper

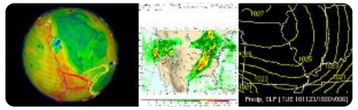
The Performance Advantage of EDR 100G InfiniBand (28-80%)

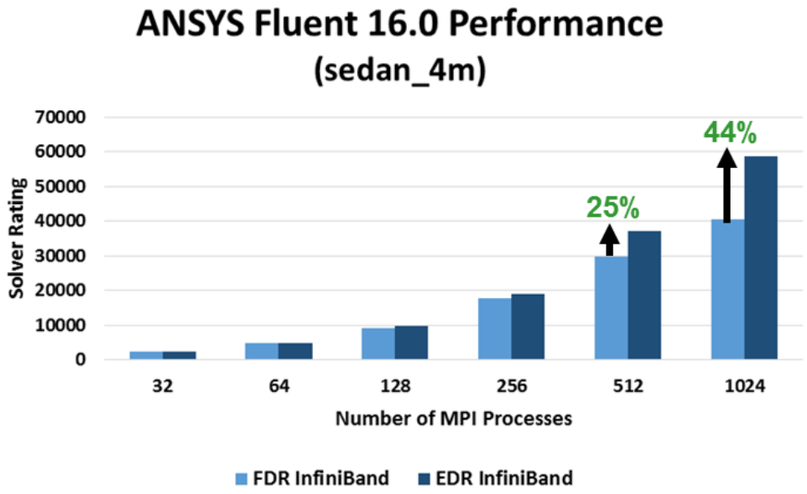


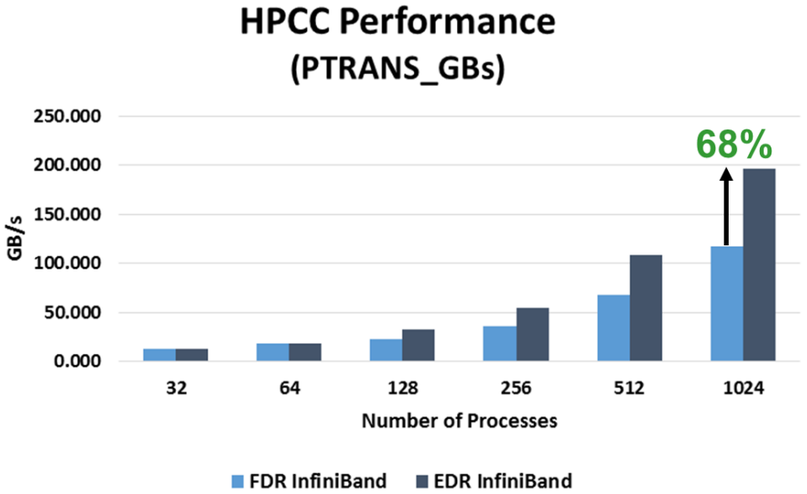


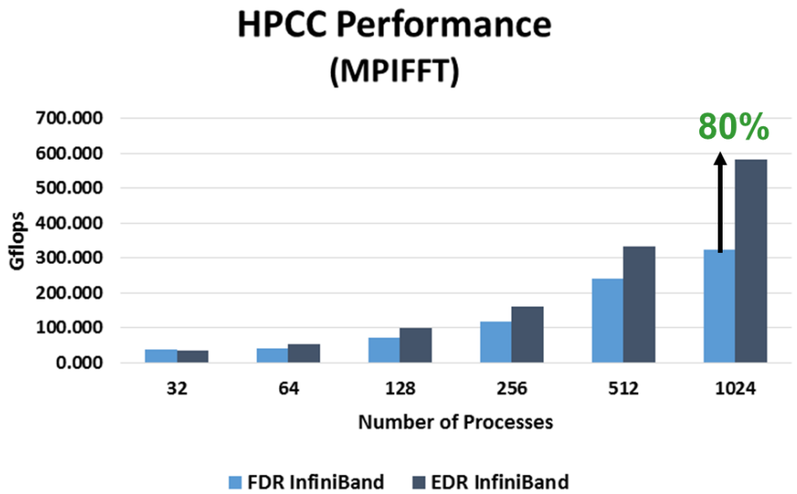






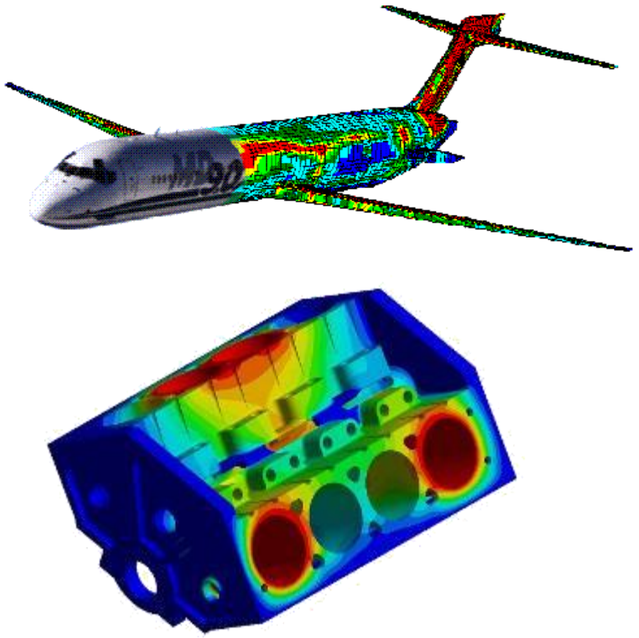


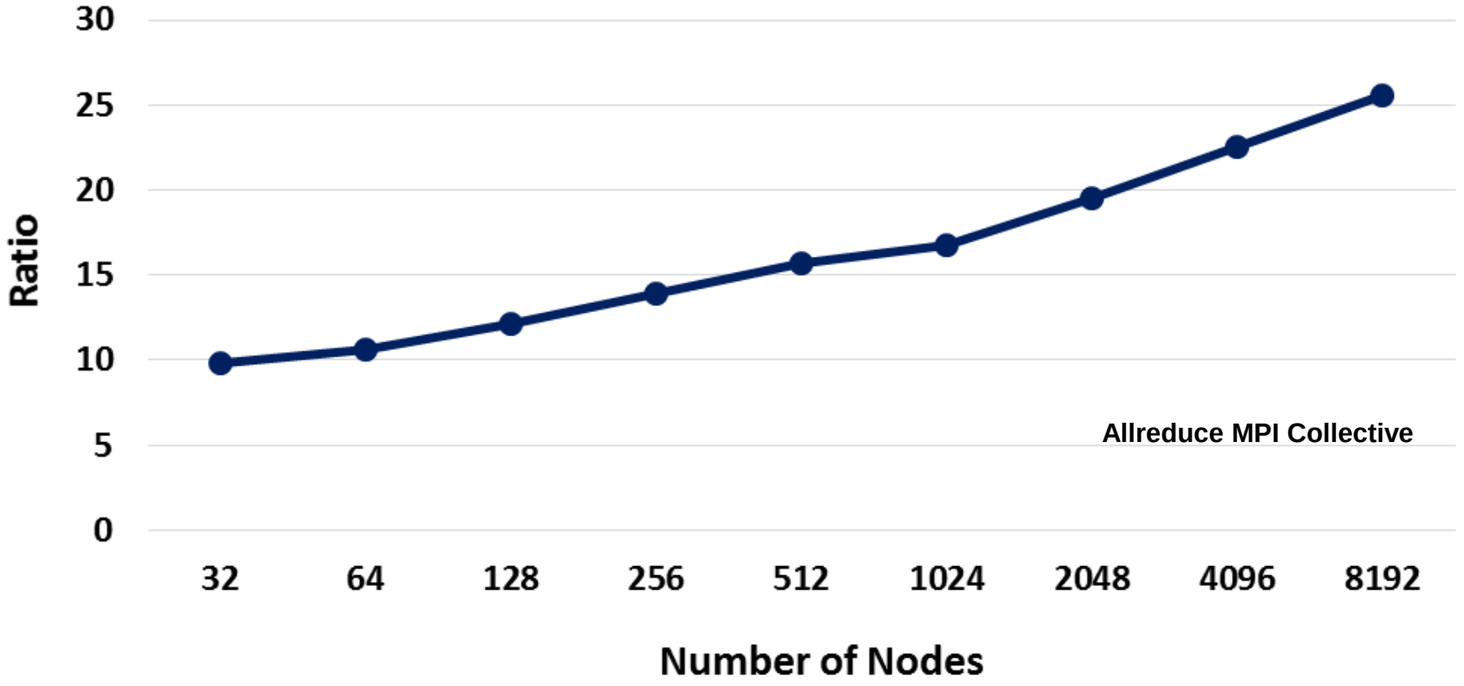




Switch-IB 2 SHArP Performance Advantage

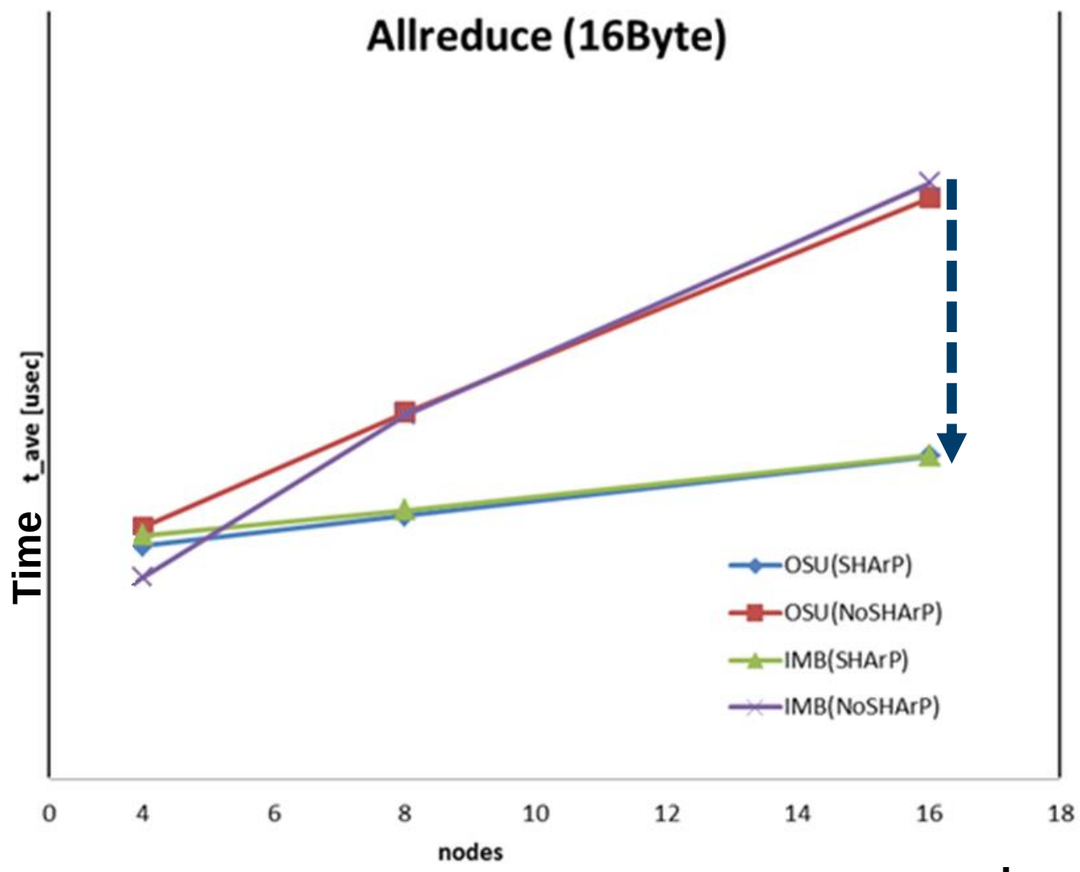
- MiniFE is a Finite Element mini-application
  - Implements kernels that represent implicit finite-element applications

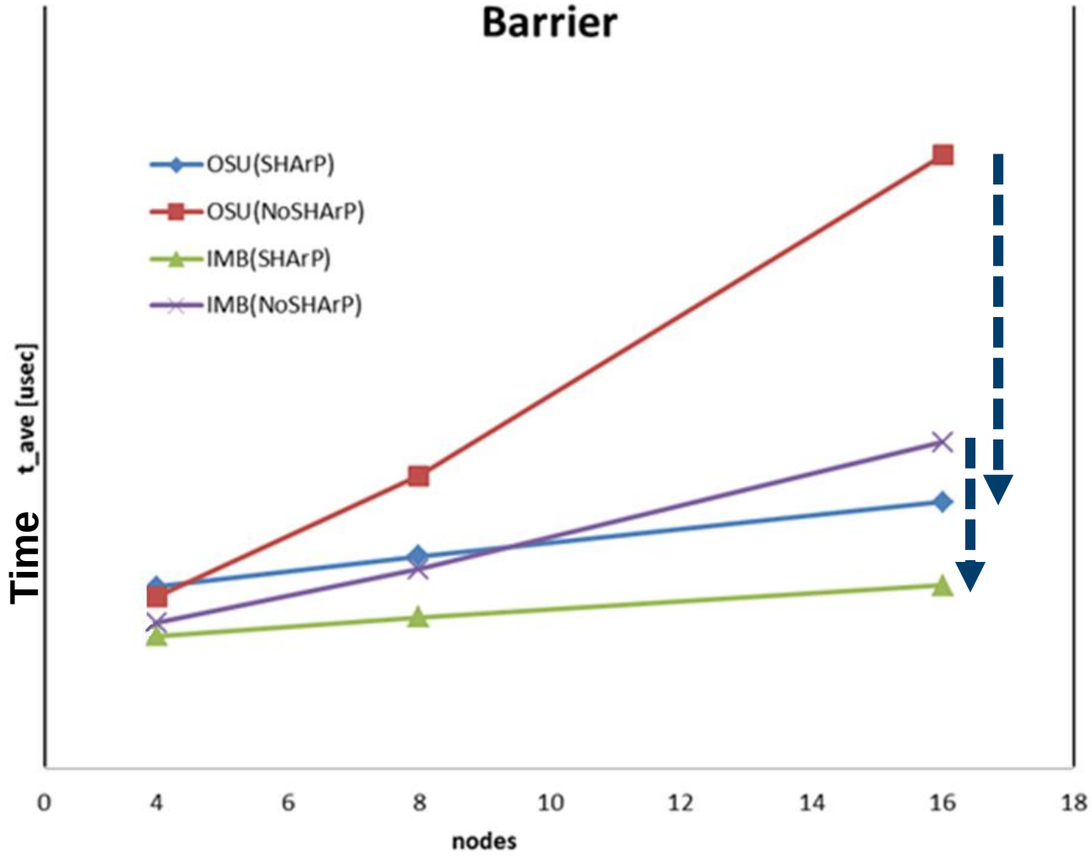




10X to 25X Performance Improvement

SHArP Performance Advantage with Intel Xeon Phi Knight Landing





Lower is better

OSU - OSU MPI benchmark; IMB – Intel MPI Benchmark

Maximizing KNL Performance – 50% Reduction in Run Time (Customer Results)

Innova ConnectX-4 Lx Programmable Adapter

Scalable, Efficient, High-Performance and Flexible Solution


Security


Cloud/Virtualization


Storage


High Performance Computing


Precision Time Synchronization

Innova

Networking + FPGA

Mellanox Acceleration Engines and FGPA Programmability On One Adapter

BlueField System-on-a-Chip (SoC) Solution

Storage

NVMe Flash Storage Arrays

Scale-Out Storage (NVMe over Fabric)



NFV

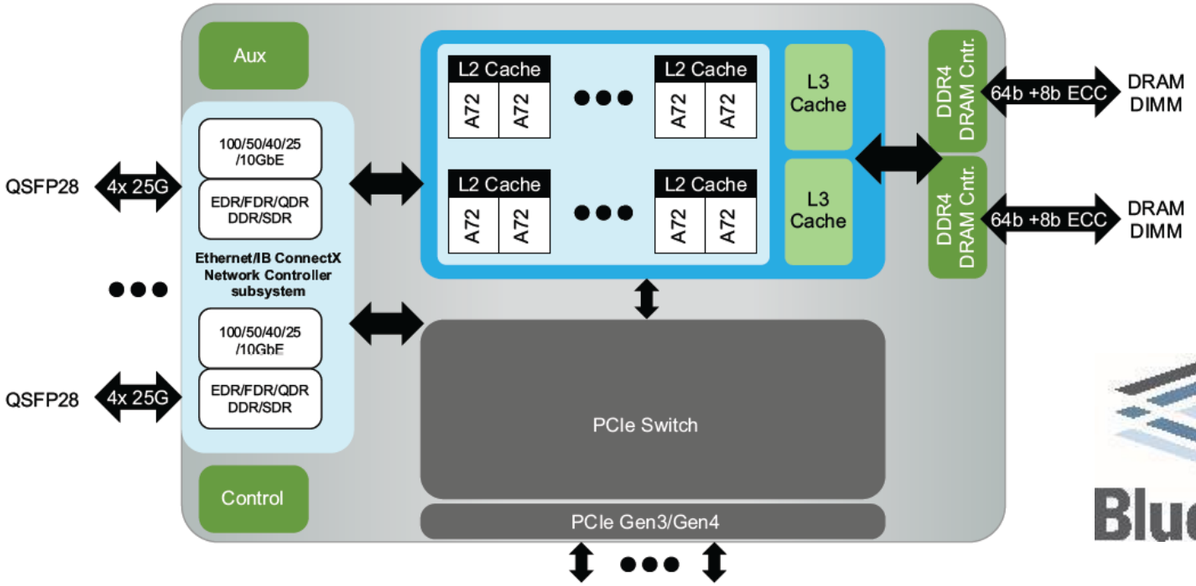
Accelerating & Virtualizing VNFs

Open vSwitch (OVS), SDN

Overlay networking offloads


- Integration of ConnectX5 + Multicore ARM
- State of the art capabilities
  - 10 / 25 / 40 / 50 / 100G Ethernet & InfiniBand
  - PCIe Gen3 / Gen4
  - Hardware acceleration offload
    - RDMA, RoCE, NVMeF, RAID
- Family of products
  - Range of ARM core counts and I/O ports/speeds
  - Price/Performance points

BlueField Architecture

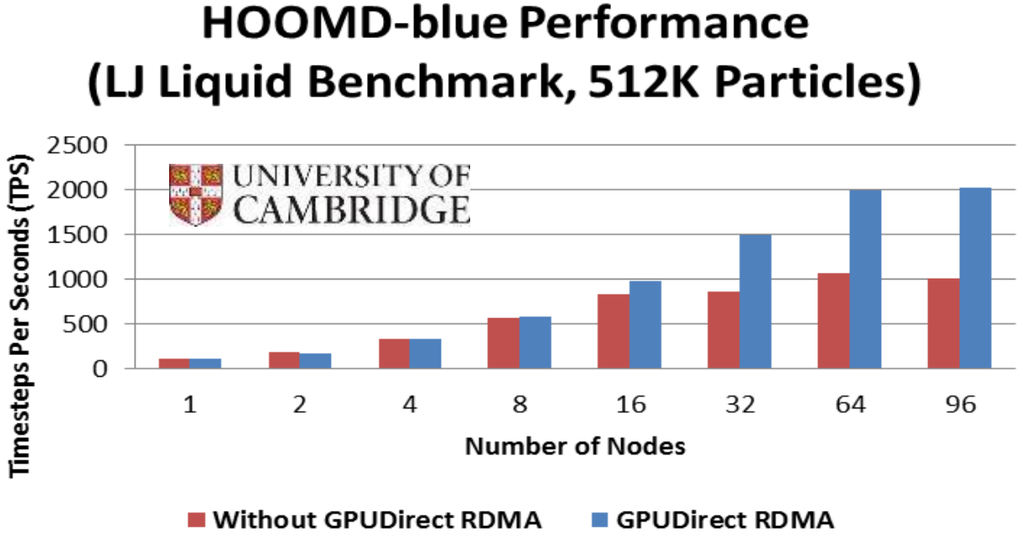


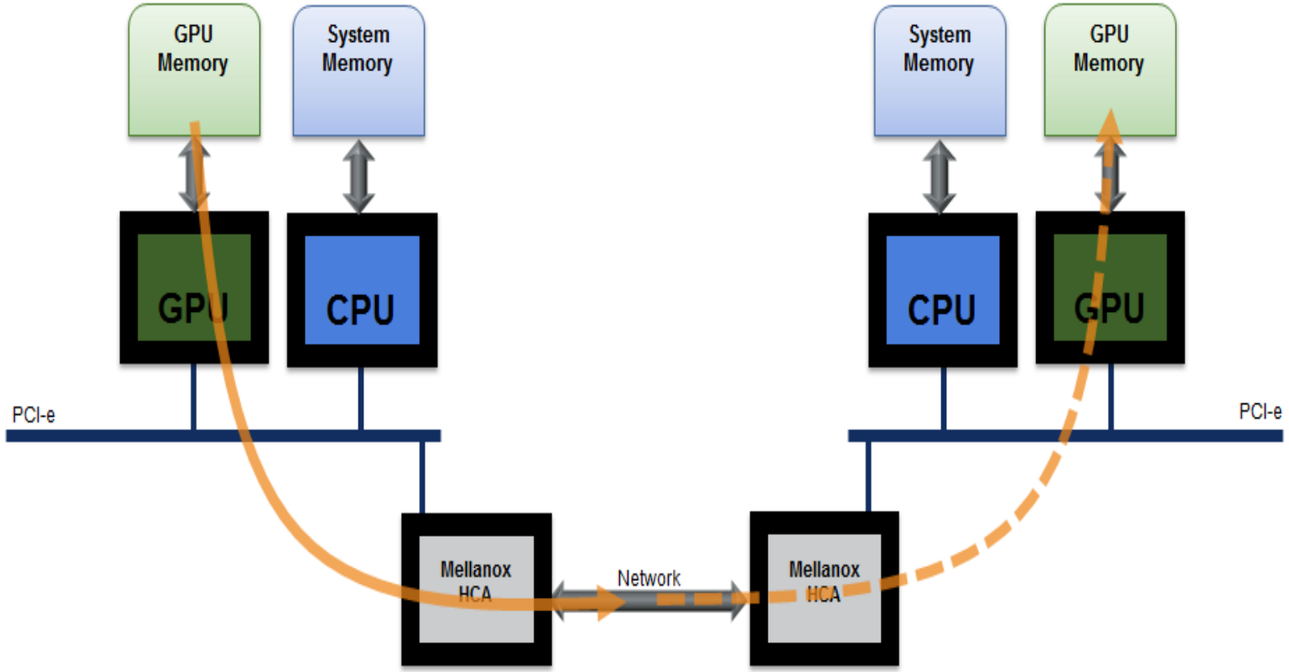
BlueField

GPUDirect RDMA Technology

Maximize Performance via Accelerator and GPU Offloads

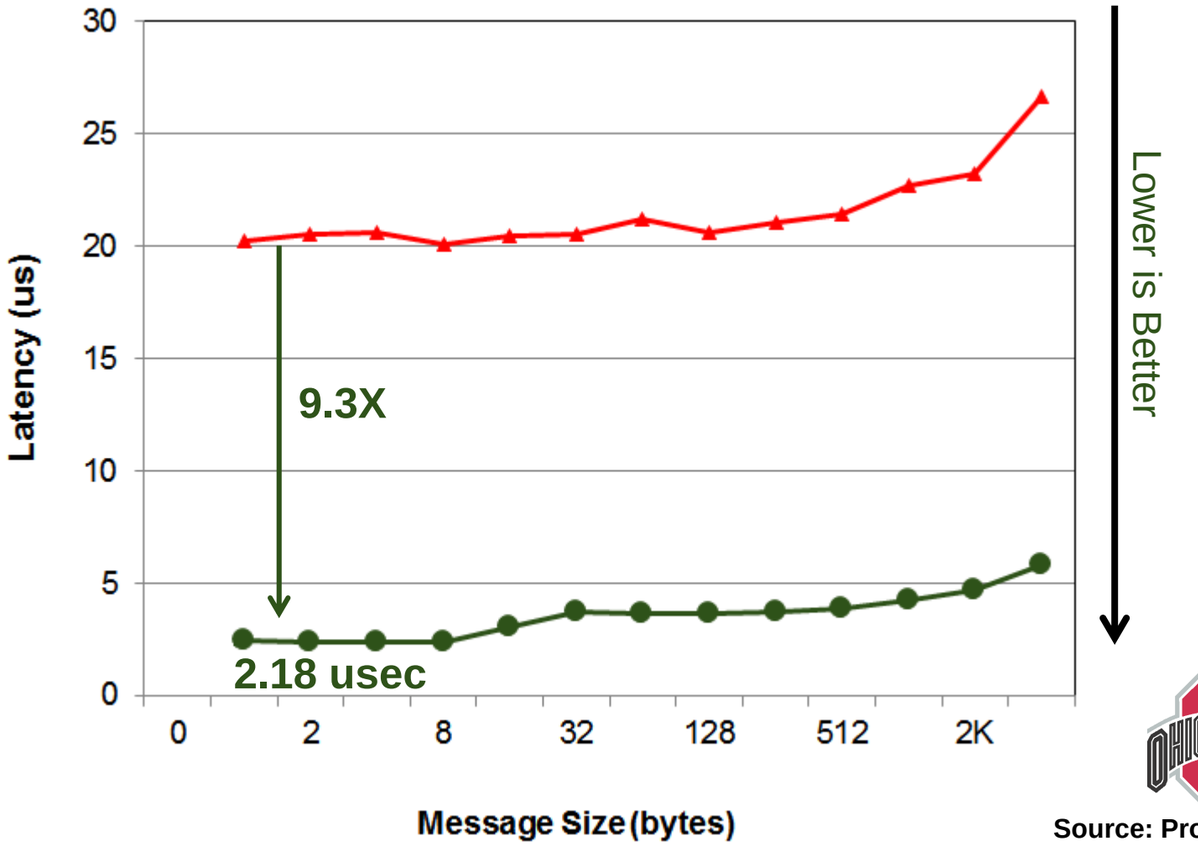
GPUs are Everywhere!





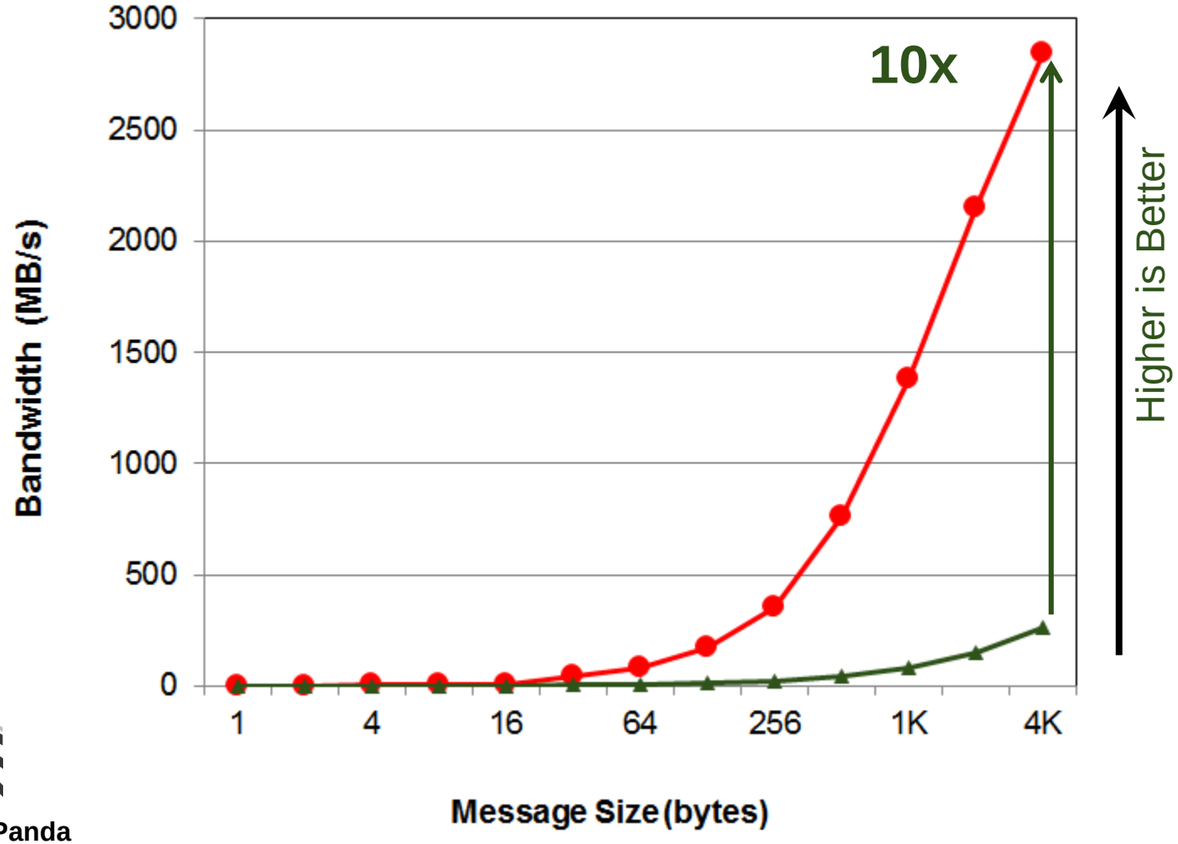
Performance of MPI with GPUDirect RDMA

GPU-GPU Internode MPI Latency



Source: Prof. DK Panda

GPU-GPU Internode MPI Bandwidth

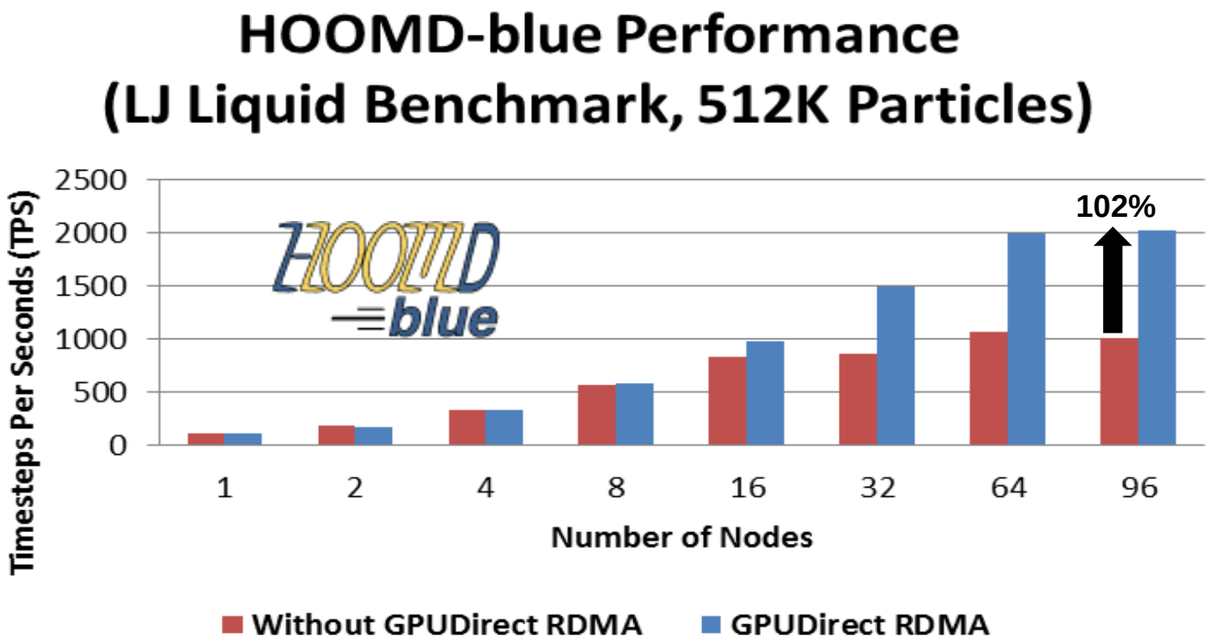


88% Lower Latency

10X Increase in Throughput

Mellanox GPUDirect RDMA Performance Advantage

- 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



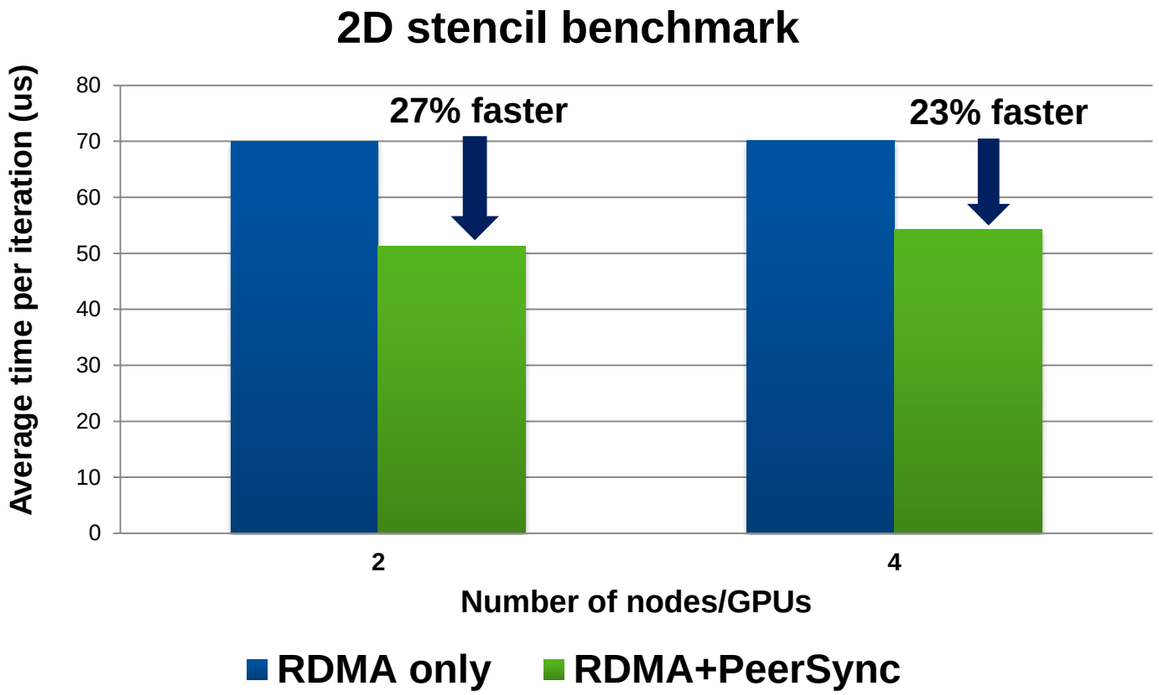
2X Application Performance!

GPUDirect Sync (GPUDirect 4.0)

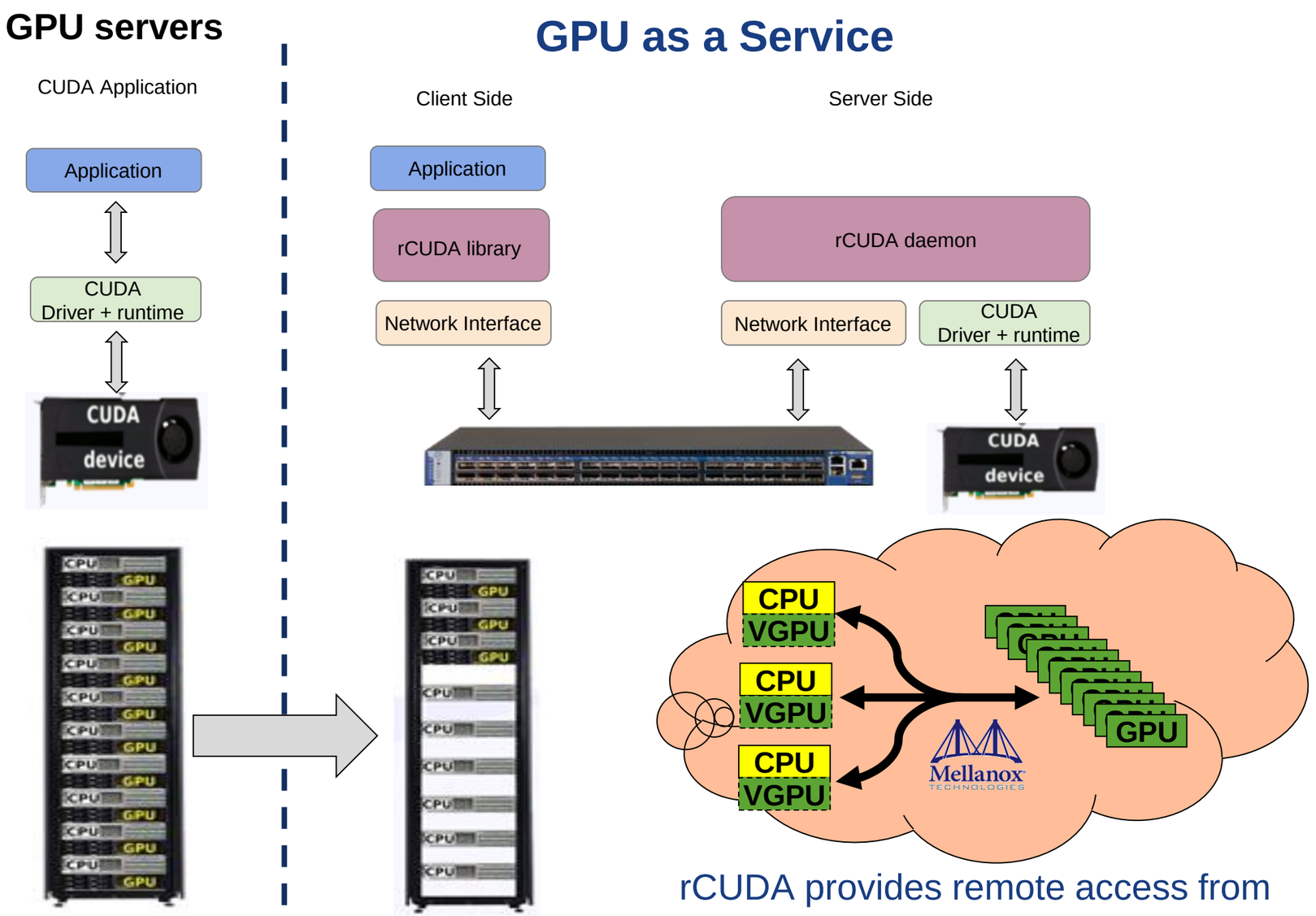
- GPUDirect RDMA (3.0) – direct data path between the GPU and Mellanox interconnect
  - Control path still uses the CPU
    - CPU prepares and queues communication tasks on GPU
    - GPU triggers communication on HCA
    - Mellanox HCA directly accesses GPU memory

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

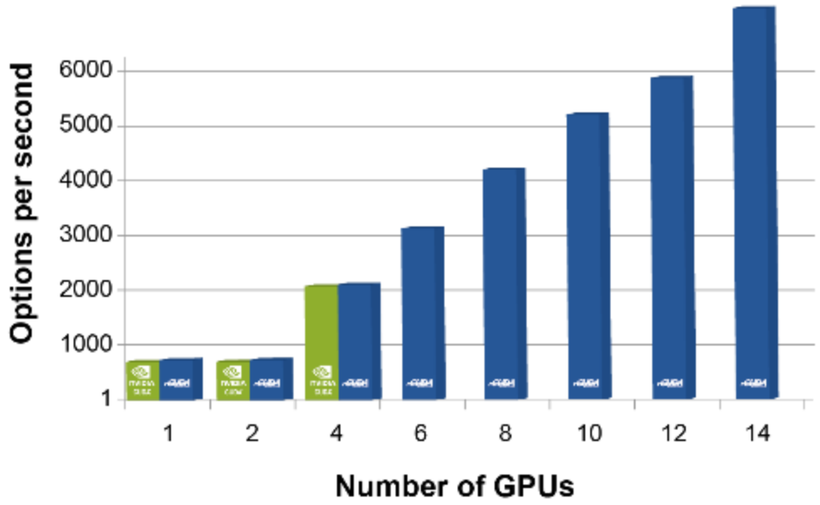
Maximum Performance For GPU Clusters

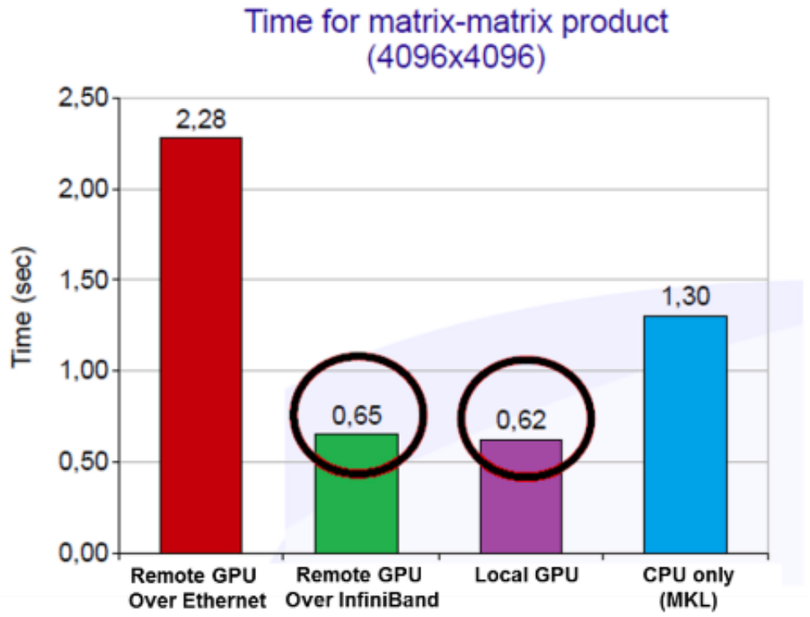


Remote GPU Access through rCUDA



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





Artificial Intelligence and Cognitive Computing

Advancing Technology to Affect Science, Business, and Society

By Enabling Critical and Timely Decision Making

Health Care, Business Integrity, Business Intelligence
Knowledge Discovery, Security, Customer Support and more

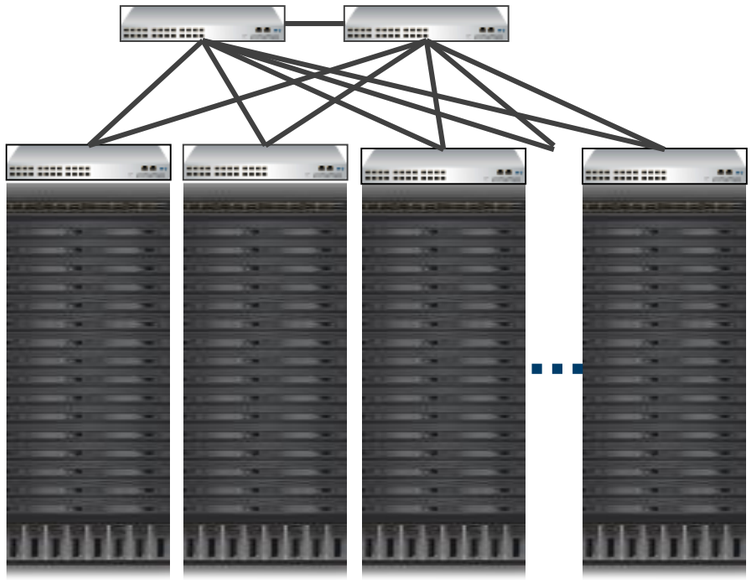
- Artificial Intelligence systems can understand massive amounts of information
  - Bridge gaps in our knowledge
  - Help to glean better insights
  - Accelerate discoveries and decisions with more confidence.

More Data → Faster Interconnect → Better Insight → Competitive Advantage

Deep Learning - Transforming Data to Intelligence

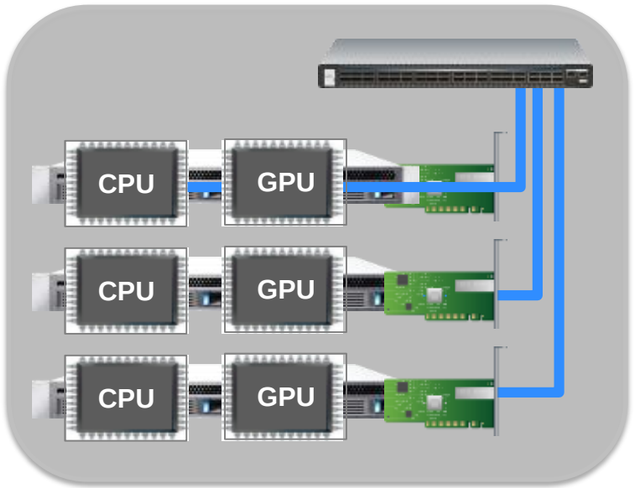
GPUDirect Enables Efficient Training Platform for Deep Neural Network

Fortune100 Web 2.0 Company



(a) Face (b) Body (c) Cat

THE OHIO STATE UNIVERSITY



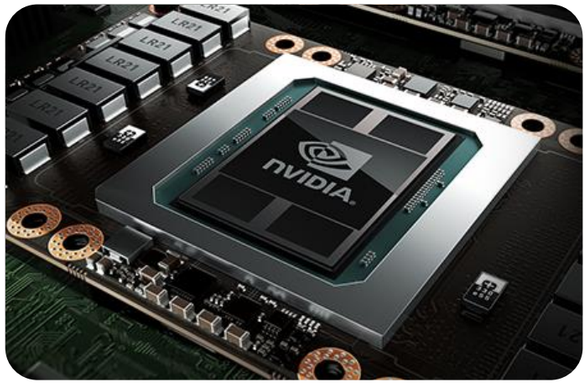
1K nodes (16K cores) for 1 week

3 Nodes with 3 GPUs for 3 days
Mellanox InfiniBand and GPU-Direct

NVIDIA DGX-1 Deep Learning Appliance

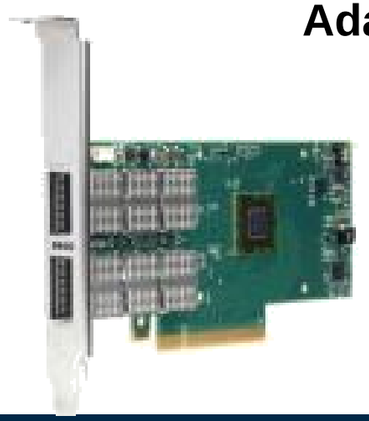
Deep Learning Supercomputer in a Box

8 x Pascal GPU (P100)

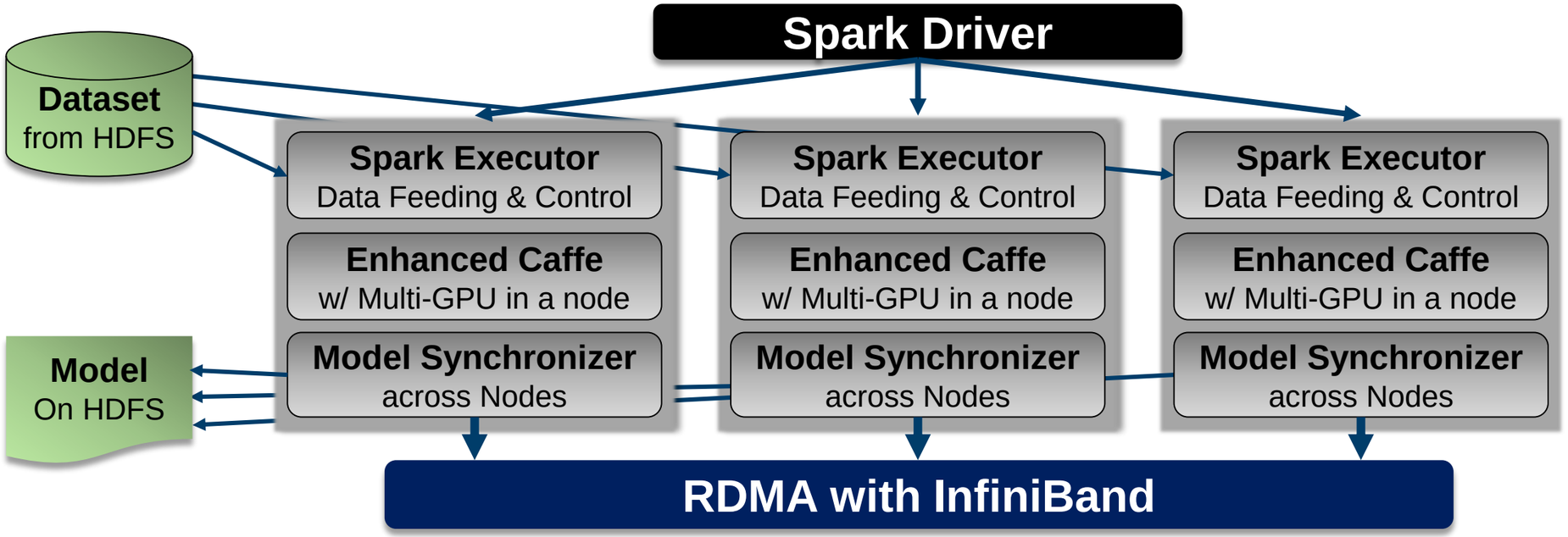


5.3TFlops
16nm FinFET
NVLINK

4 x ConnectX-4 EDR 100G InfiniBand Adapters



RDMA Accelerated Deep Learning (Hadoop)



flickr from YAHOO!

Caffe

Large Scale Distributed Deep Learning on Hadoop Clusters - Yahoo Big ML Team [link]

- RDMA enables Deep Learning with Caffe + Hadoop
- 18.7x Overall Speedup, 80% Accuracy , 10 hours of training
  - 4 servers with 8 GPUs and Mellanox InfiniBand

Enabling Advanced Predictive Analytics for Image Recognition