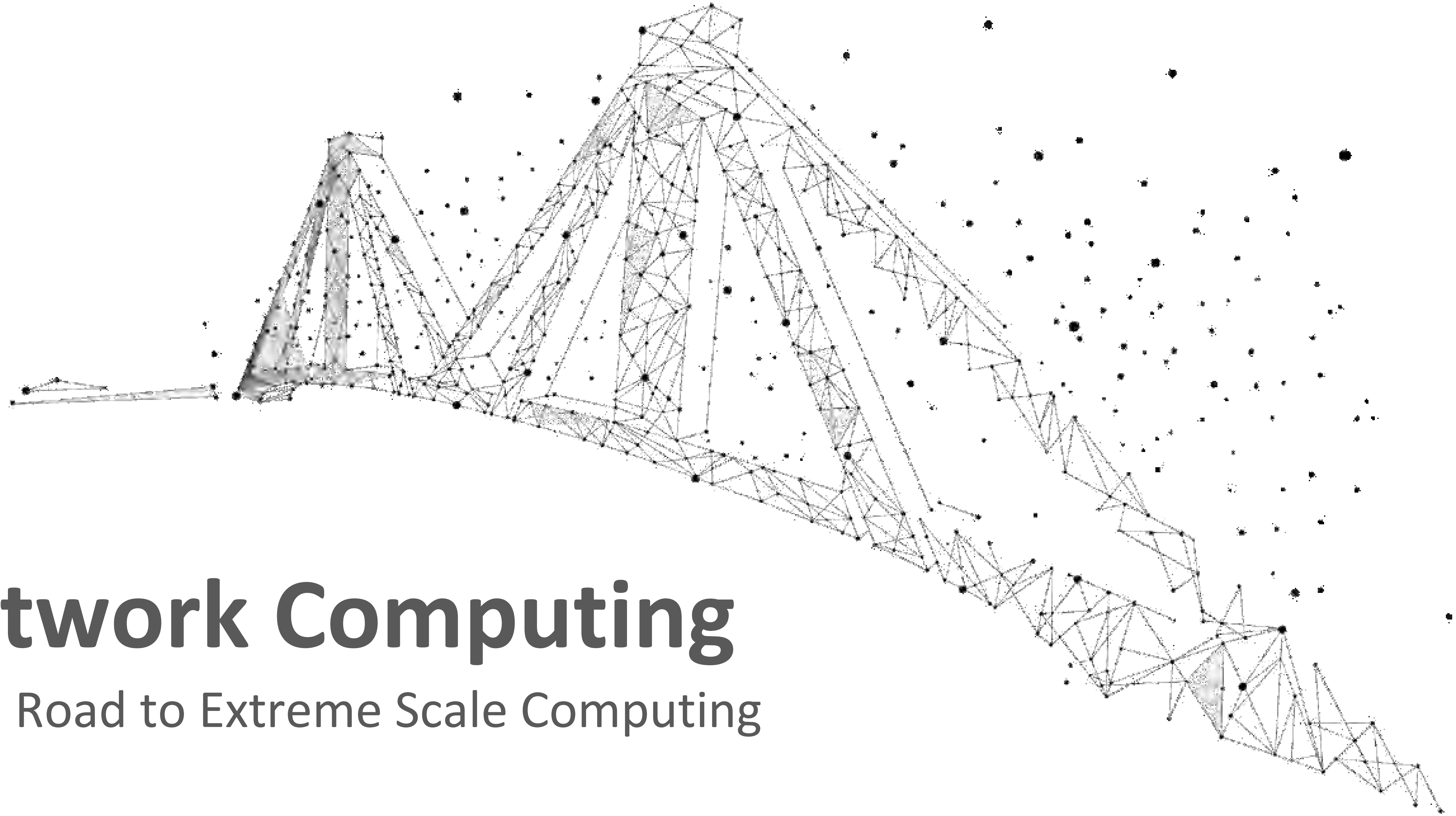


In-Network Computing

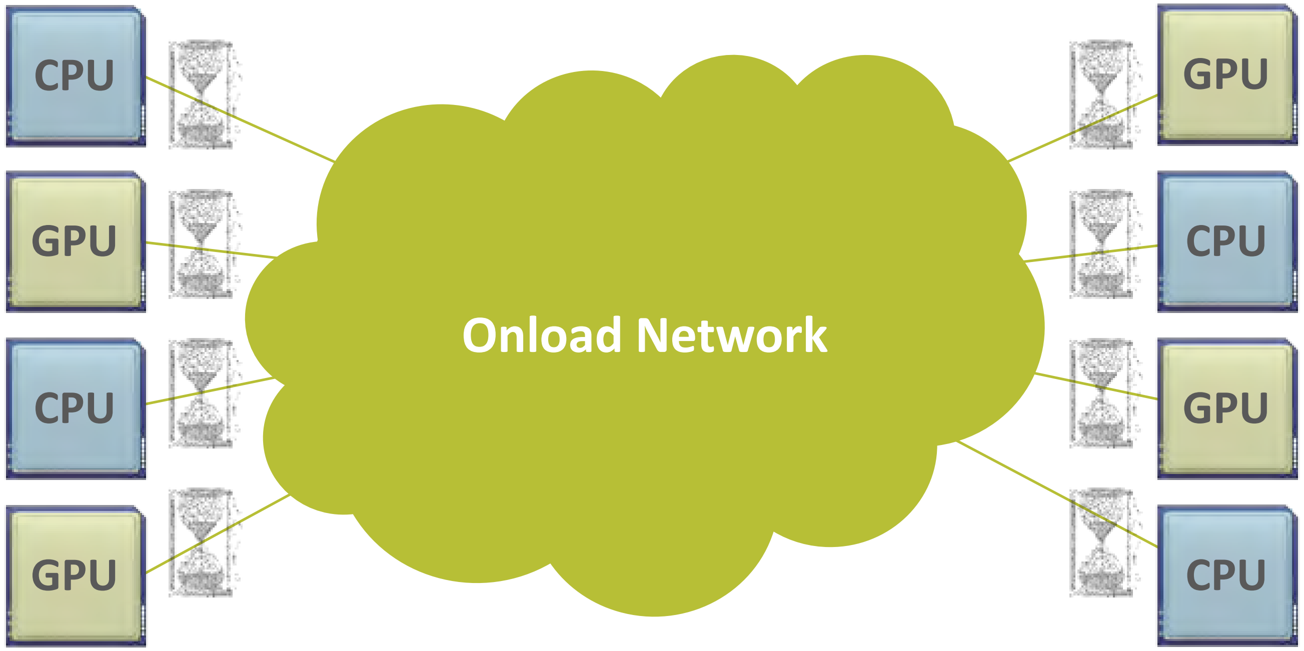
Paving the Road to Extreme Scale Computing

October 2019

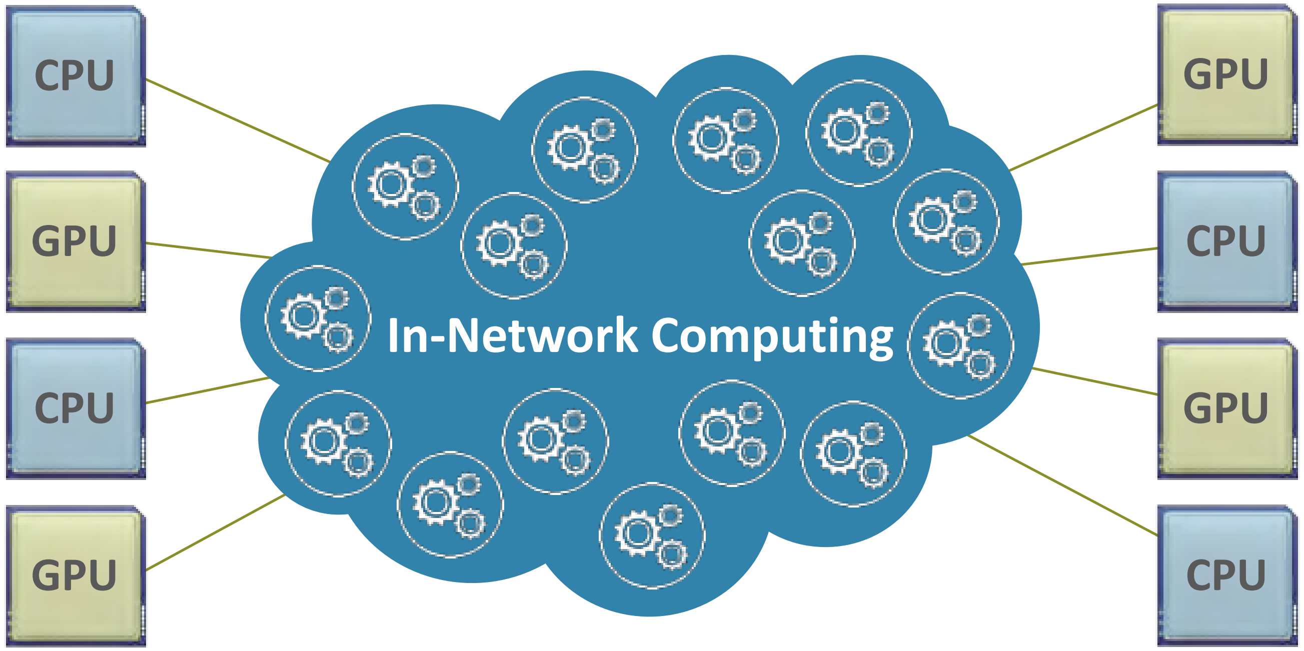
The Need for Intelligent and Faster Interconnect

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

CPU-Centric (Onload)



Data-Centric (Offload)



Must Wait for the Data
Creates Performance Bottlenecks

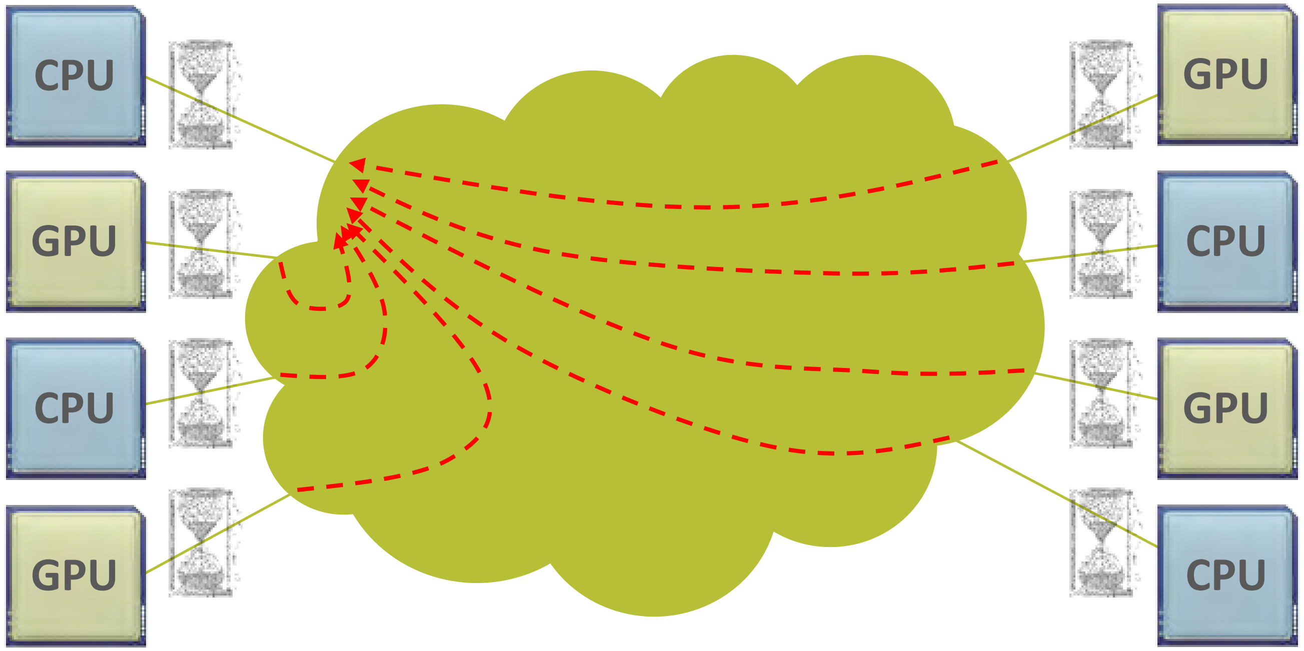
Analyze Data as it Moves
CPUs, CPU and IPUs (I/O Processing Units)
Higher Performance and Scale

© 2019 Mellanox Technologies | Confidential

Data Centric Architecture to Overcome Latency Bottlenecks

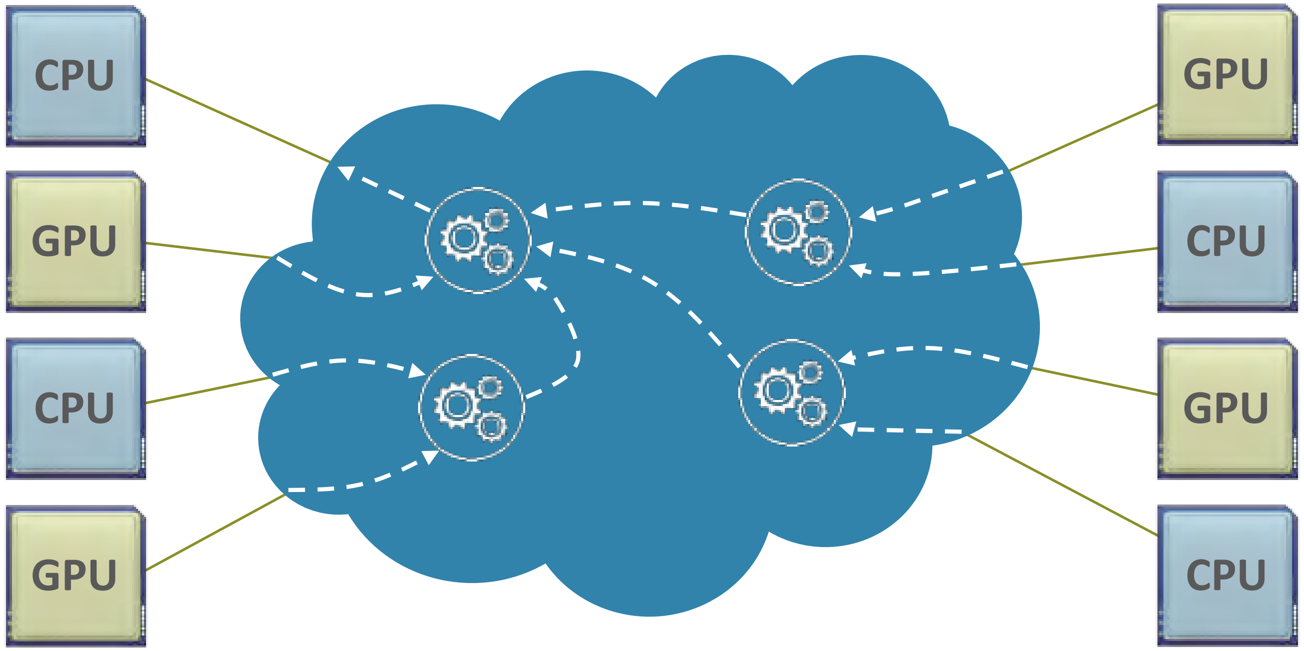
Intelligent Interconnect Paves the Road to Exascale Performance

CPU-Centric (Onload)



Communications Latencies
of 30-40us

Data-Centric (Offload)



Communications Latencies
of 3-4us

© 2019 Mellanox Technologies | Confidential

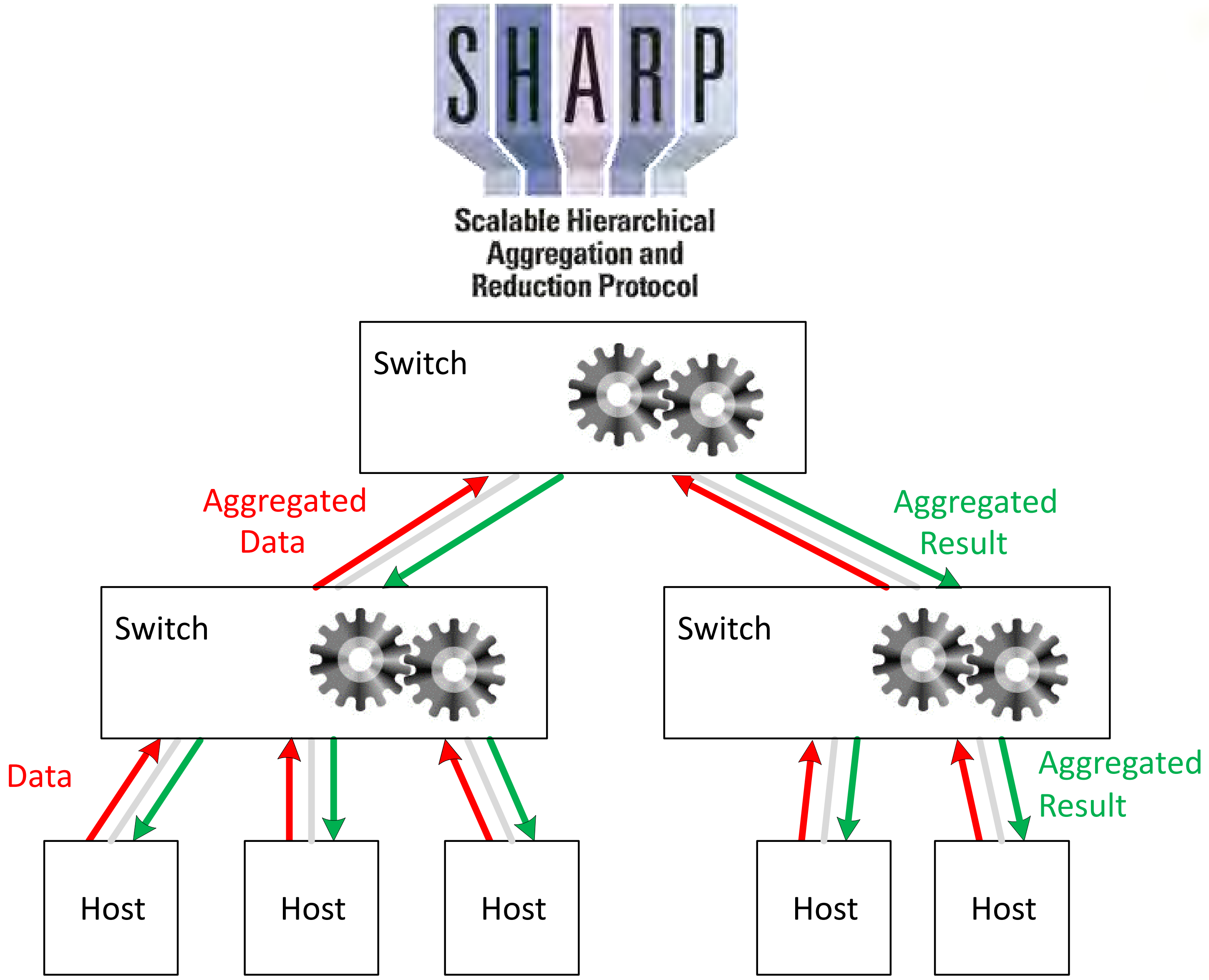
IPU Technologies: Scalable Hierarchical Aggregation and Reduction Protocol (SHARP)

Scalable Hierarchical Aggregation and Reduction Protocol (SHARP)

- Reliable Scalable General Purpose Primitive
  - In-network Tree based aggregation mechanism
  - Large number of groups
  - Multiple simultaneous outstanding operations

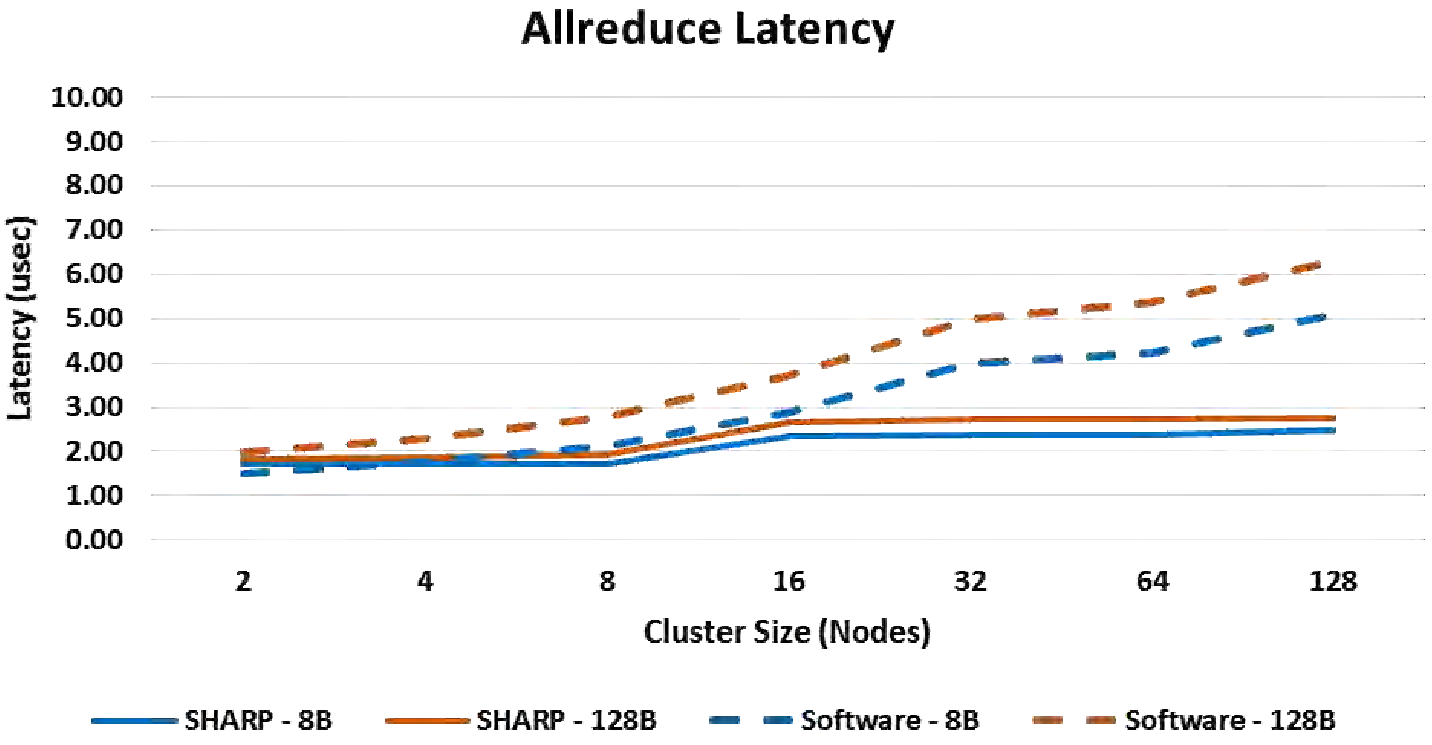
- Applicable to Multiple Use-cases
  - HPC Applications using MPI / SHMEM
  - Distributed Machine Learning applications

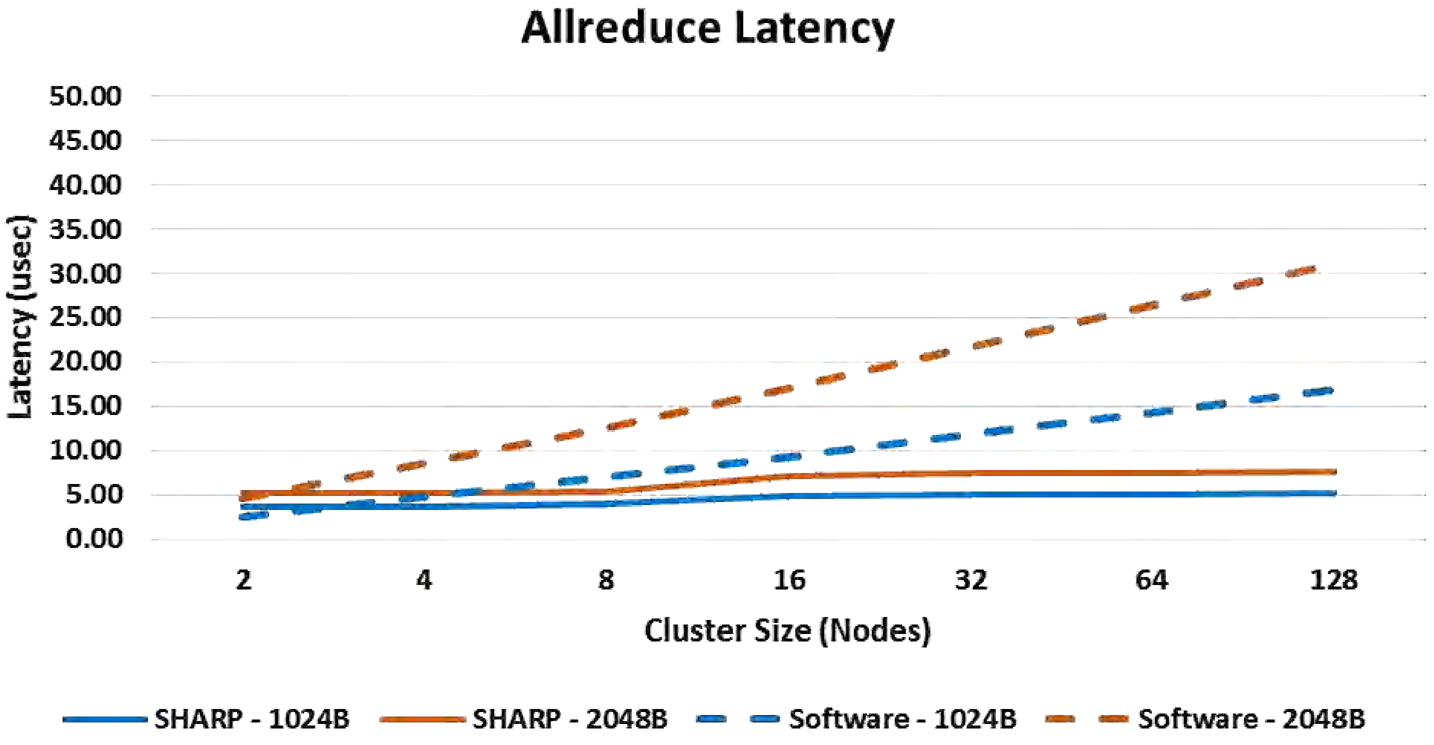
- Scalable High Performance Collective Offload
  - Barrier, Reduce, All-Reduce, Broadcast and more
  - Sum, Min, Max, Min-loc, max-loc, OR, XOR, AND
  - Integer and Floating-Point, 16/32/64 bits



© 2019 Mellanox Technologies | Confidential

SHARP AllReduce Performance Advantages (128 Nodes)

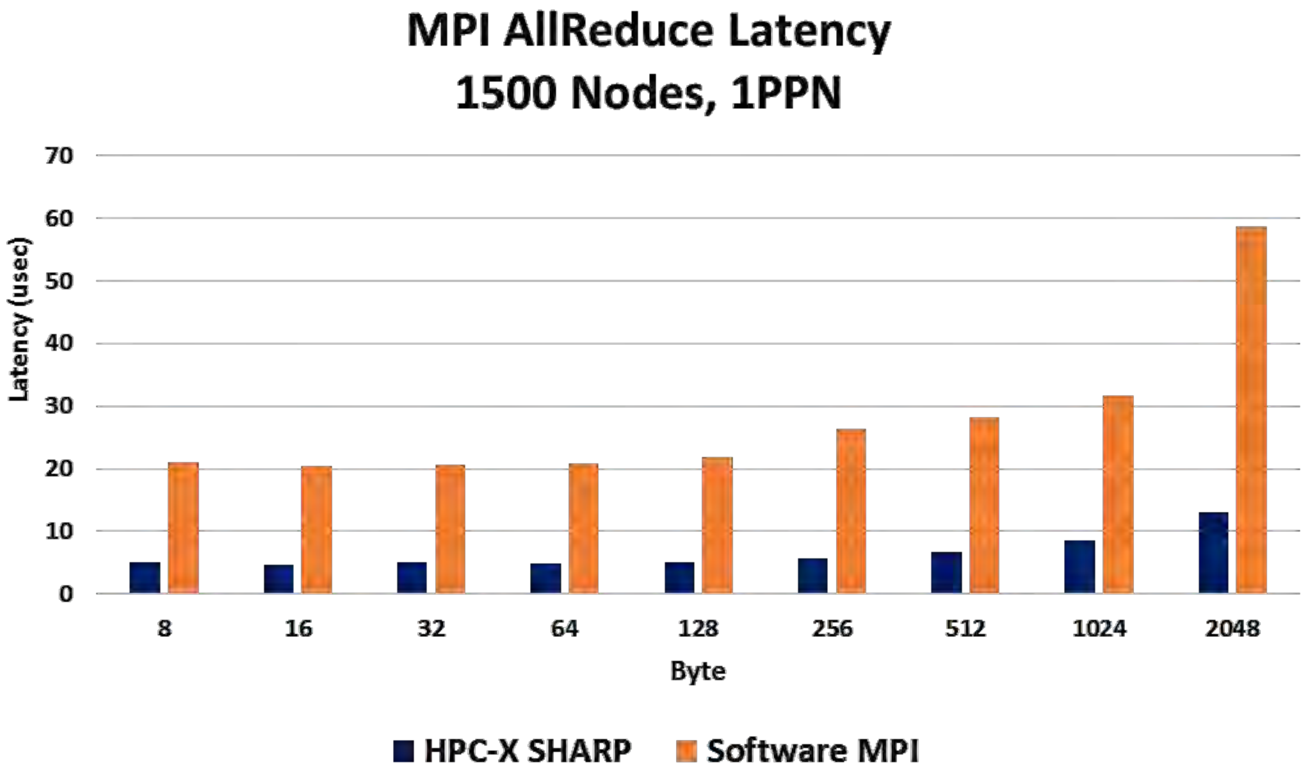


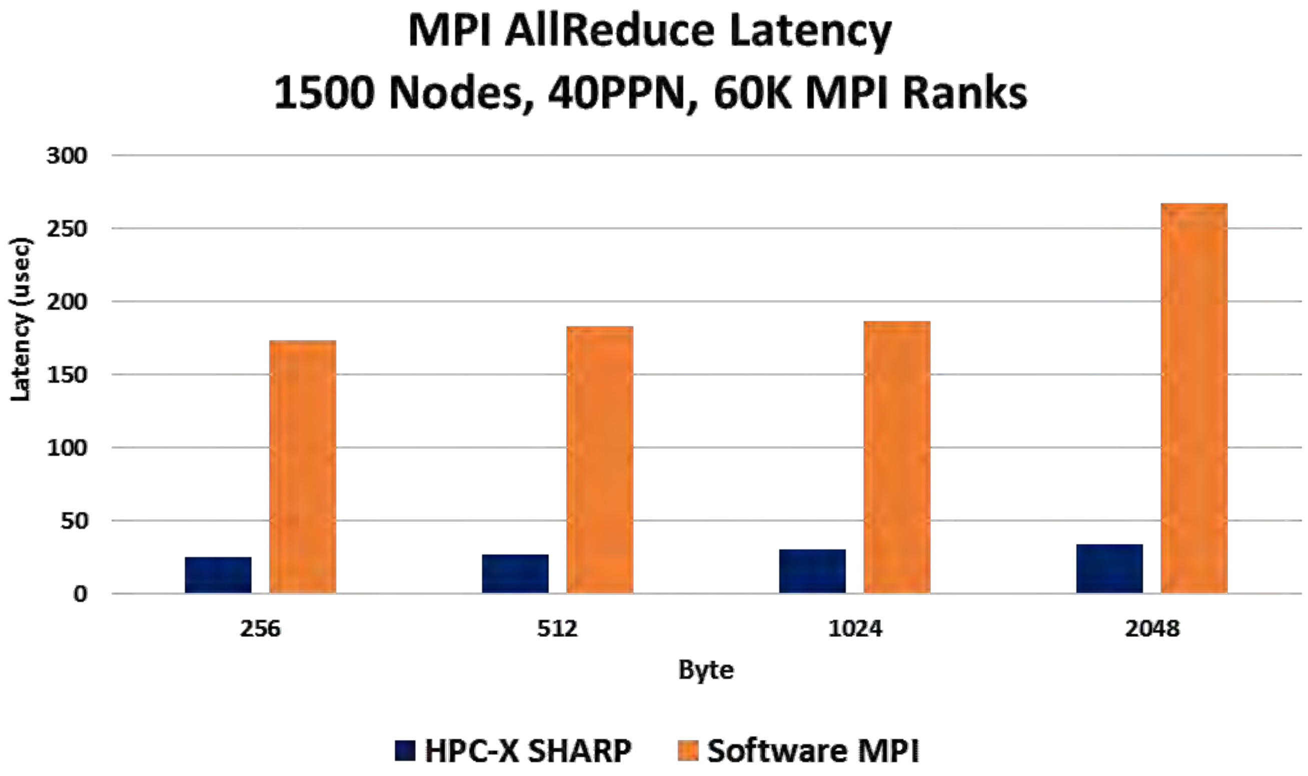


© 2019 Mellanox Technologies | Confidential

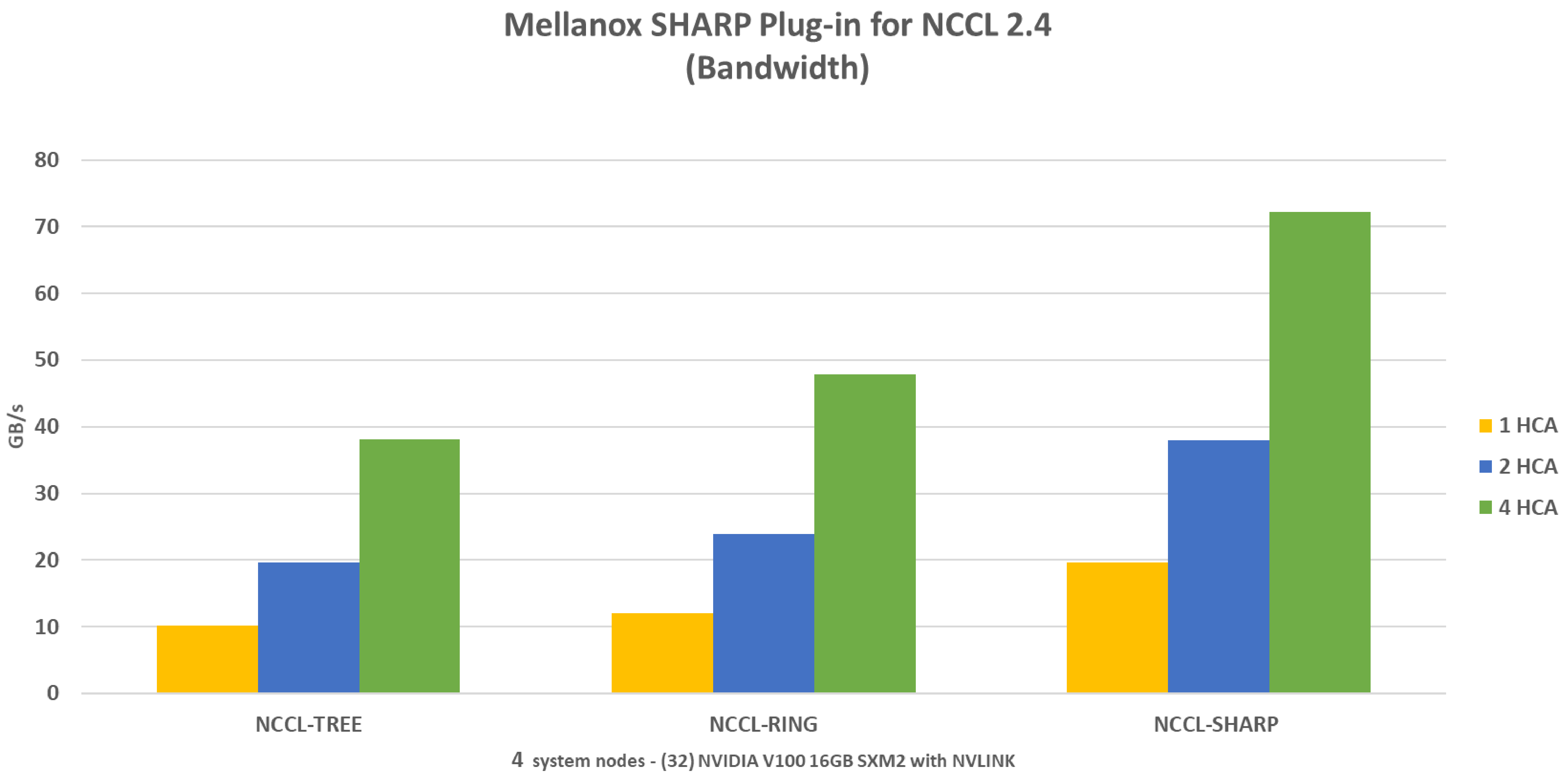
SHARP AllReduce Performance Advantages

1500 Nodes, 60K MPI Ranks, Dragonfly+ Topology (University of Toronto)



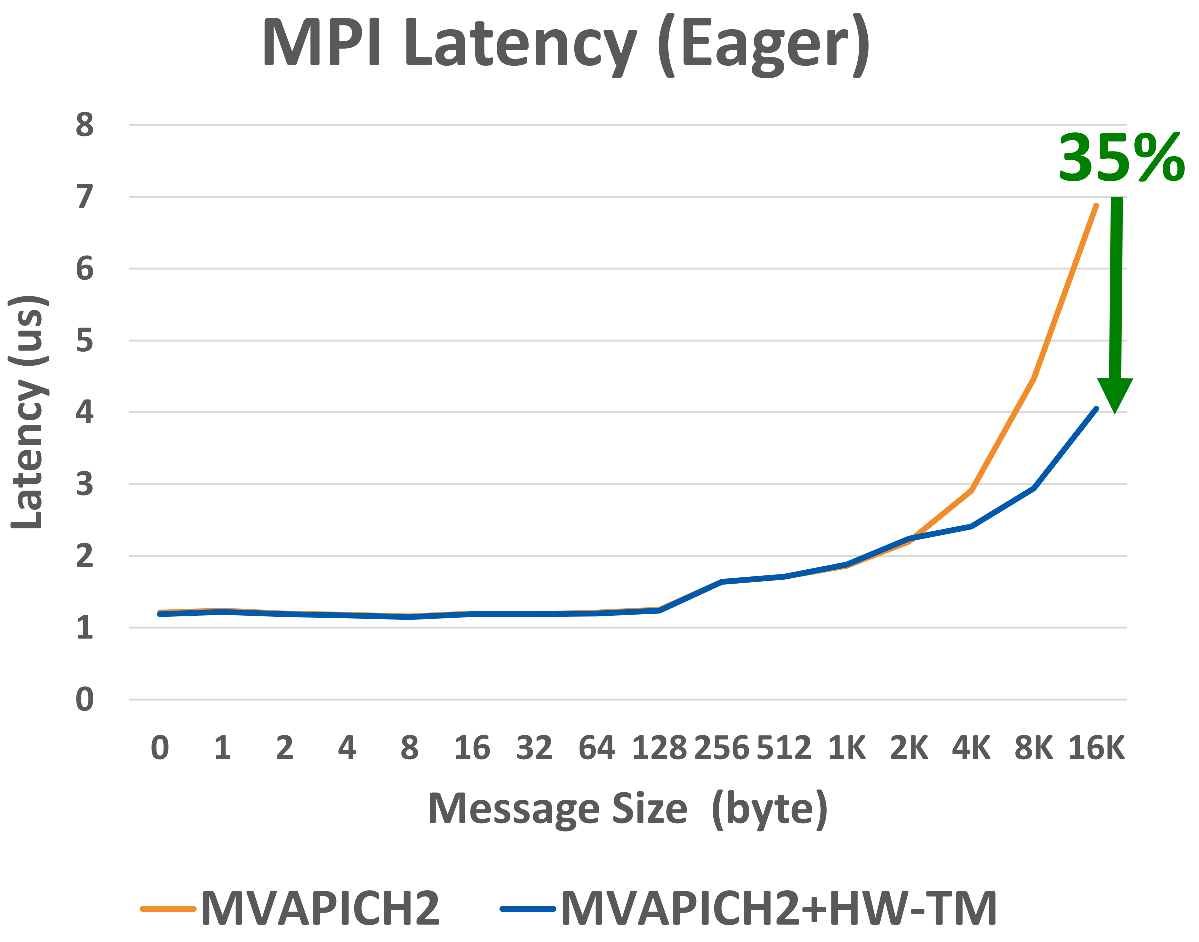


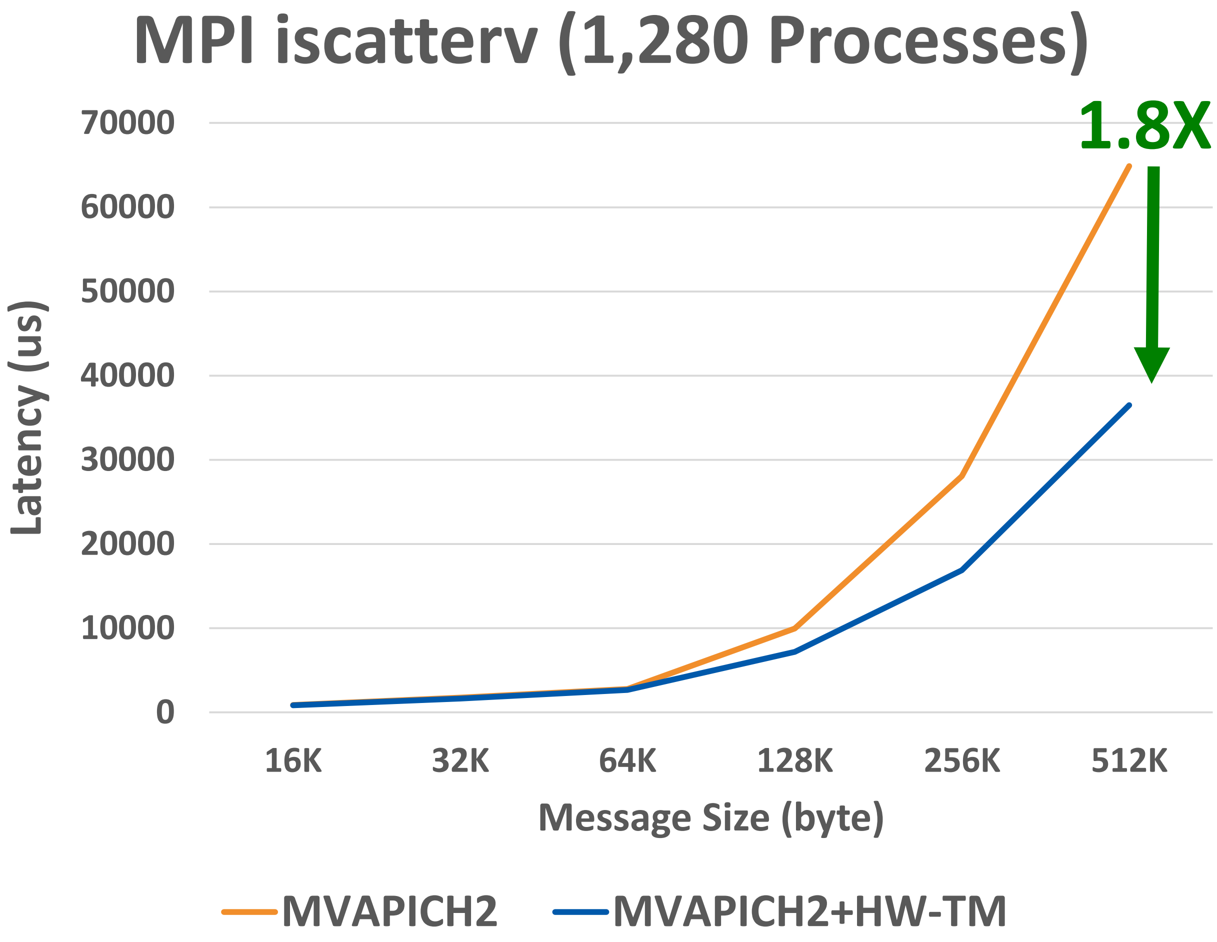
NCCL-SHARP Delivers Higeher Performance



IPU Technologies:
MPI Tag Matching Hardware Engine

Tag Matching Hardware Engine Performance Advantage

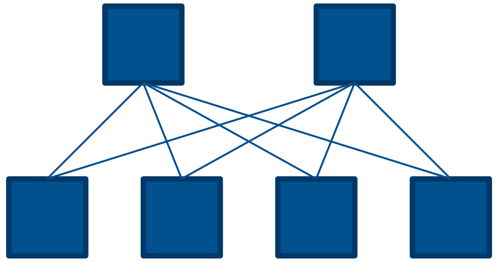




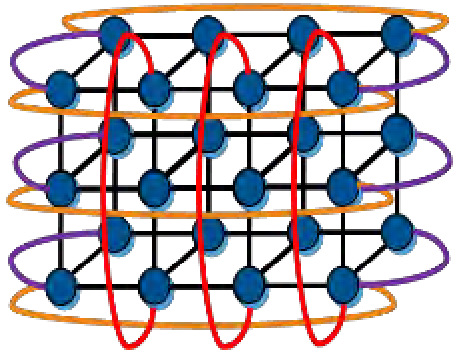
Courtesy of Dhabaleswar K. (DK) Panda
Ohio State University

Network Topologies
Leveraging Multi-Host Technology

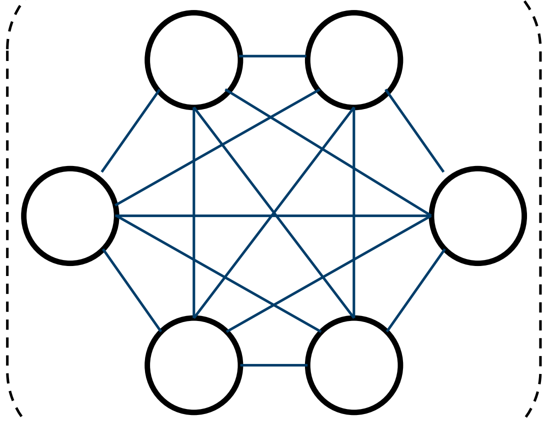
Supporting Variety of Topologies



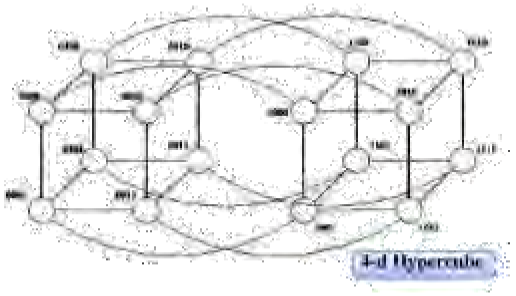
Fat Tree



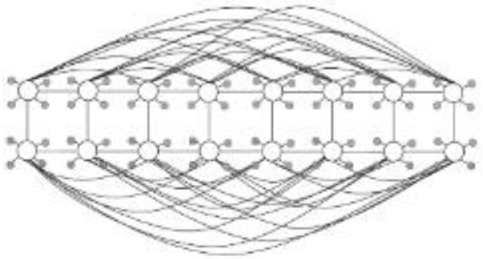
Torus



Dragonfly



Hypercube

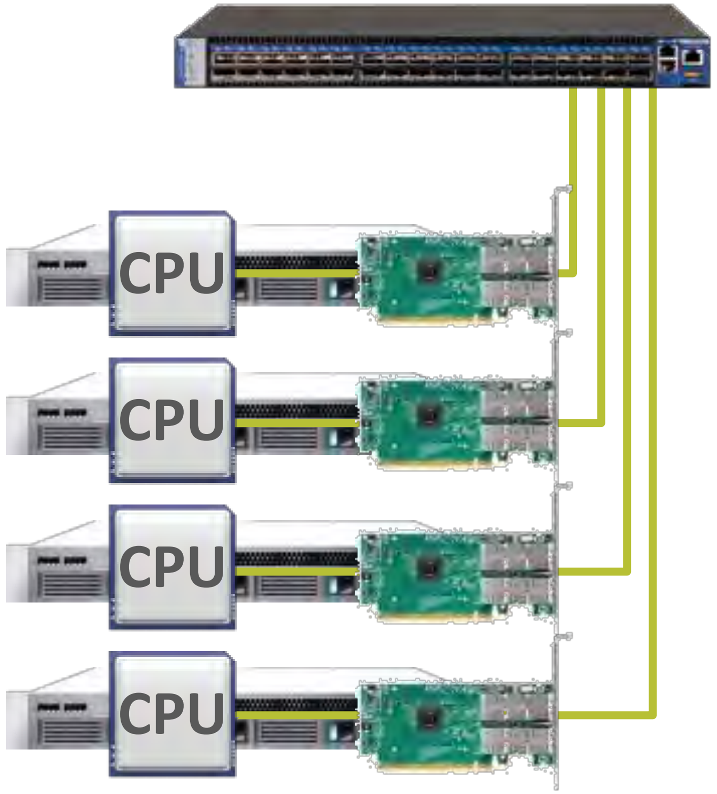


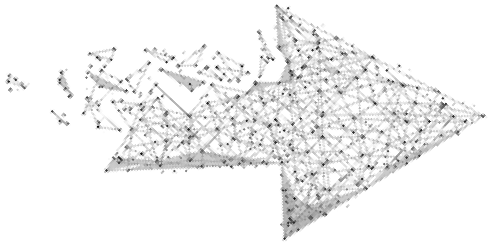
HyperX

Mellanox Multi-Host™ Technology

- Mellanox Multi Host® technology enables connecting multiple hosts into a single interconnect adapter
- By separating the ConnectX PCIe interface into multiple and independent PCIe interfaces
- Each interface is connected to a separate host with no performance degradation
- Increase datacenters performance while reducing CAPEX and OPEX

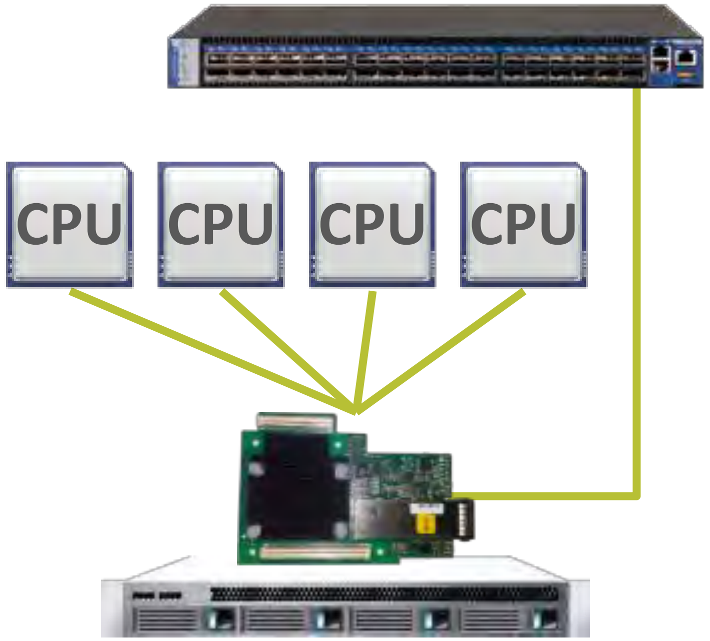
Traditional Design





Lower Connectivity Cost

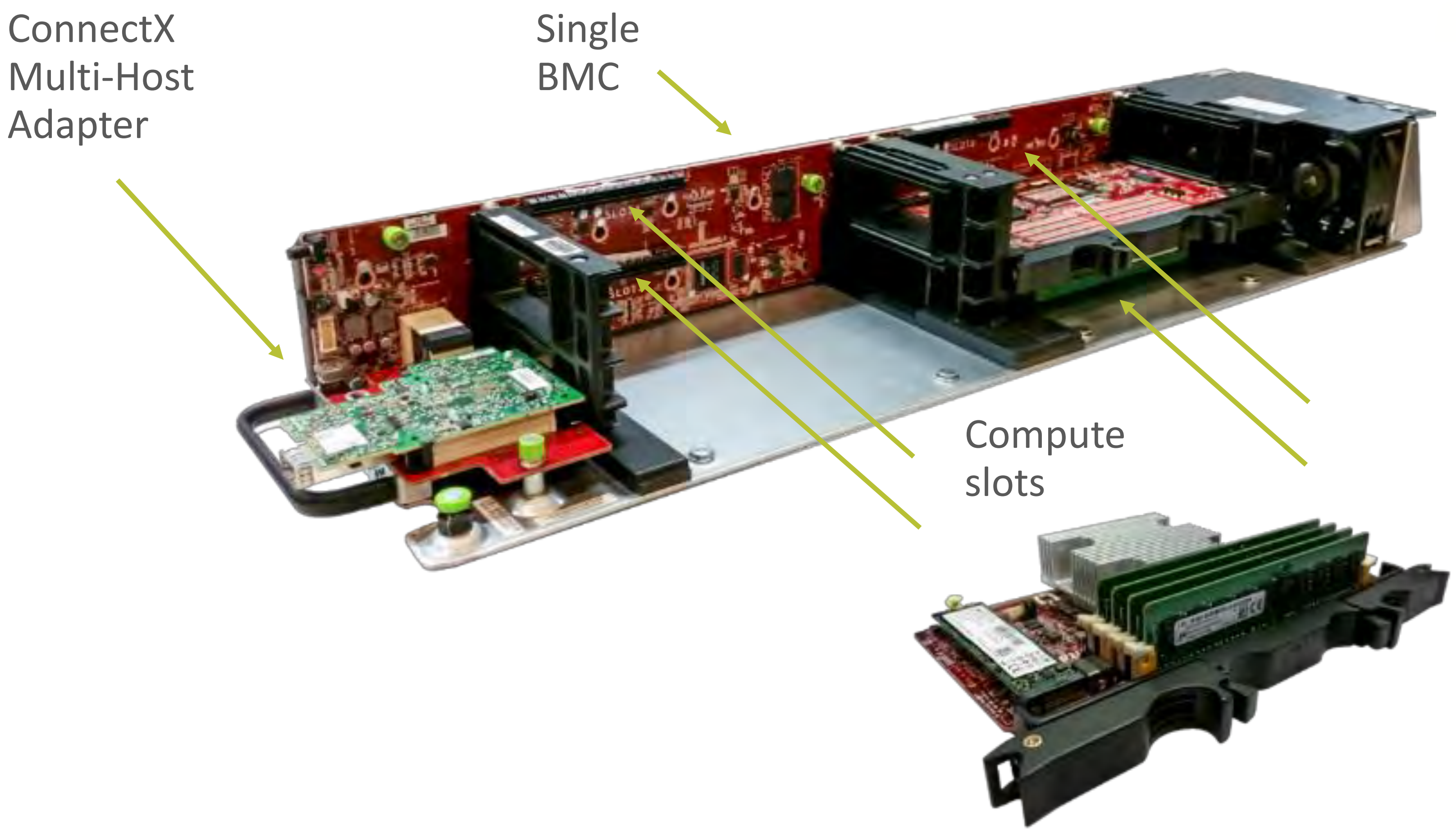
Multi-Host Technology



© 2019 Mellanox Technologies | Confidential

Facebook OCP Multi-Host Platform (Yosemite)

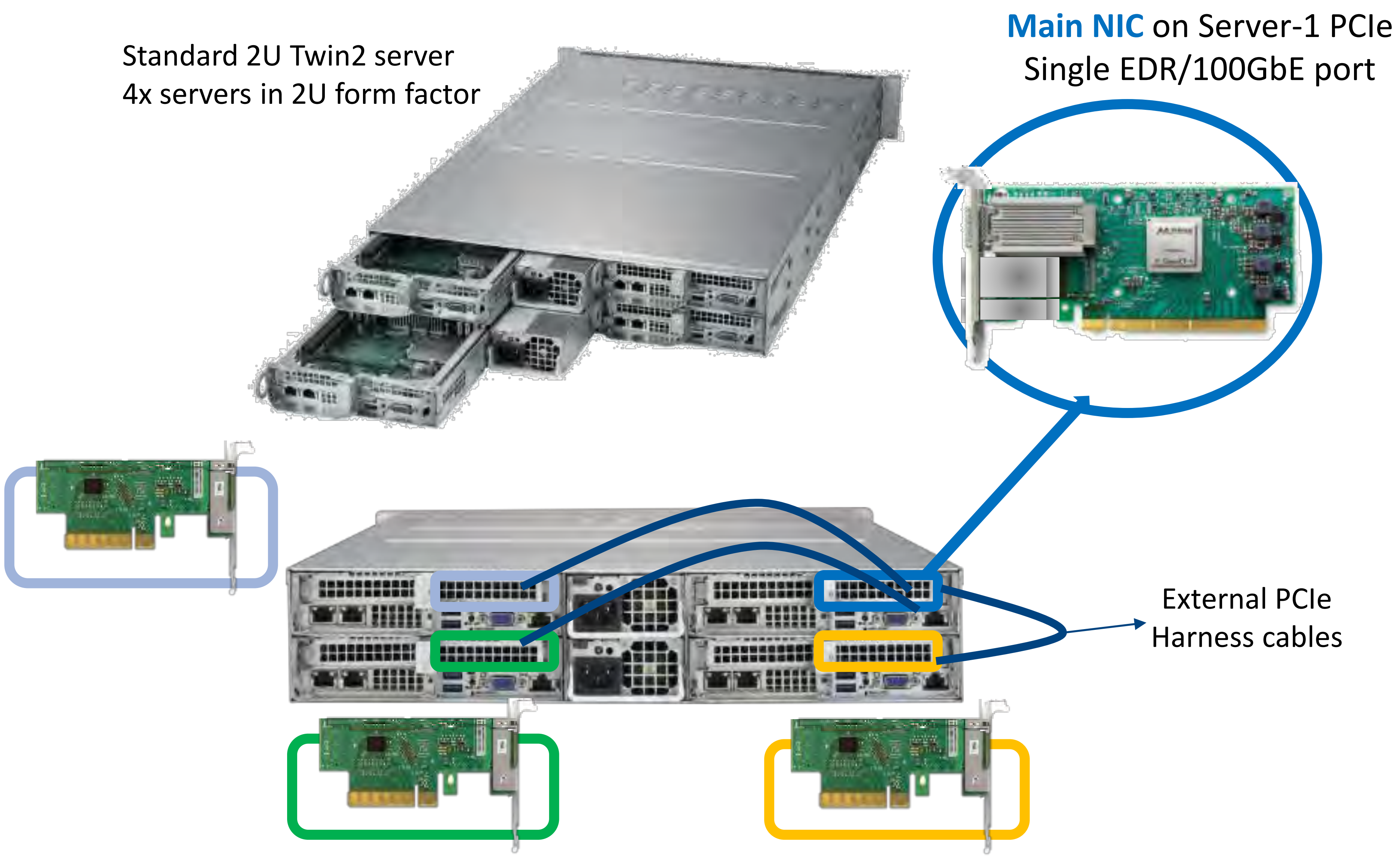
OPEN Compute Project


ConnectX Multi-Host Adapter

Single BMC

Compute slots

ConnectX External Multi Host

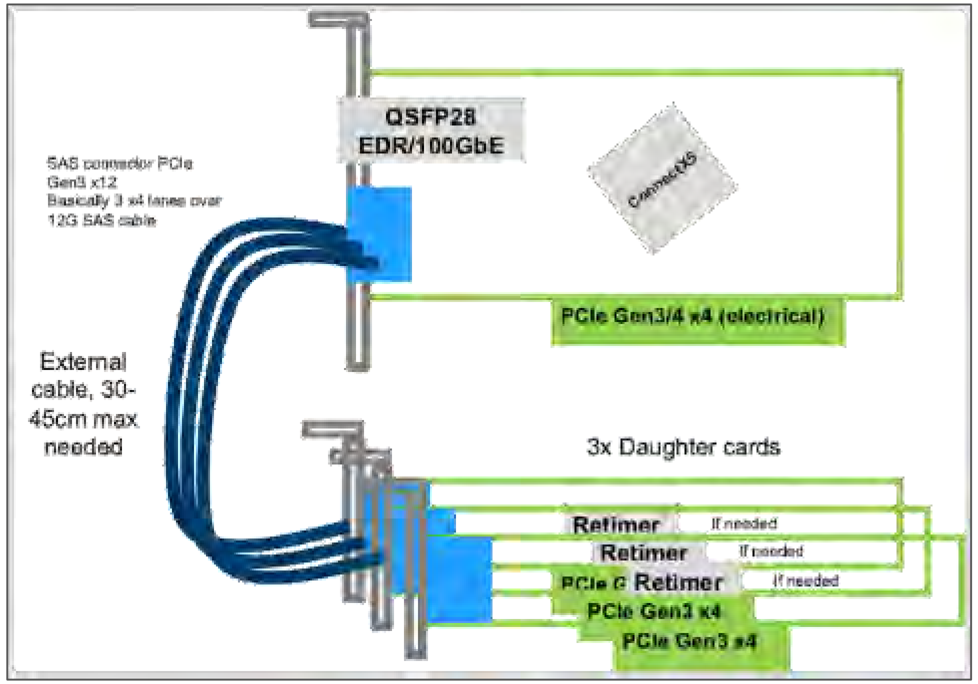


Standard 2U Twin2 server
4x servers in 2U form factor

Main NIC on Server-1 PCIe
Single EDR/100GbE port

External PCIe
Harness cables

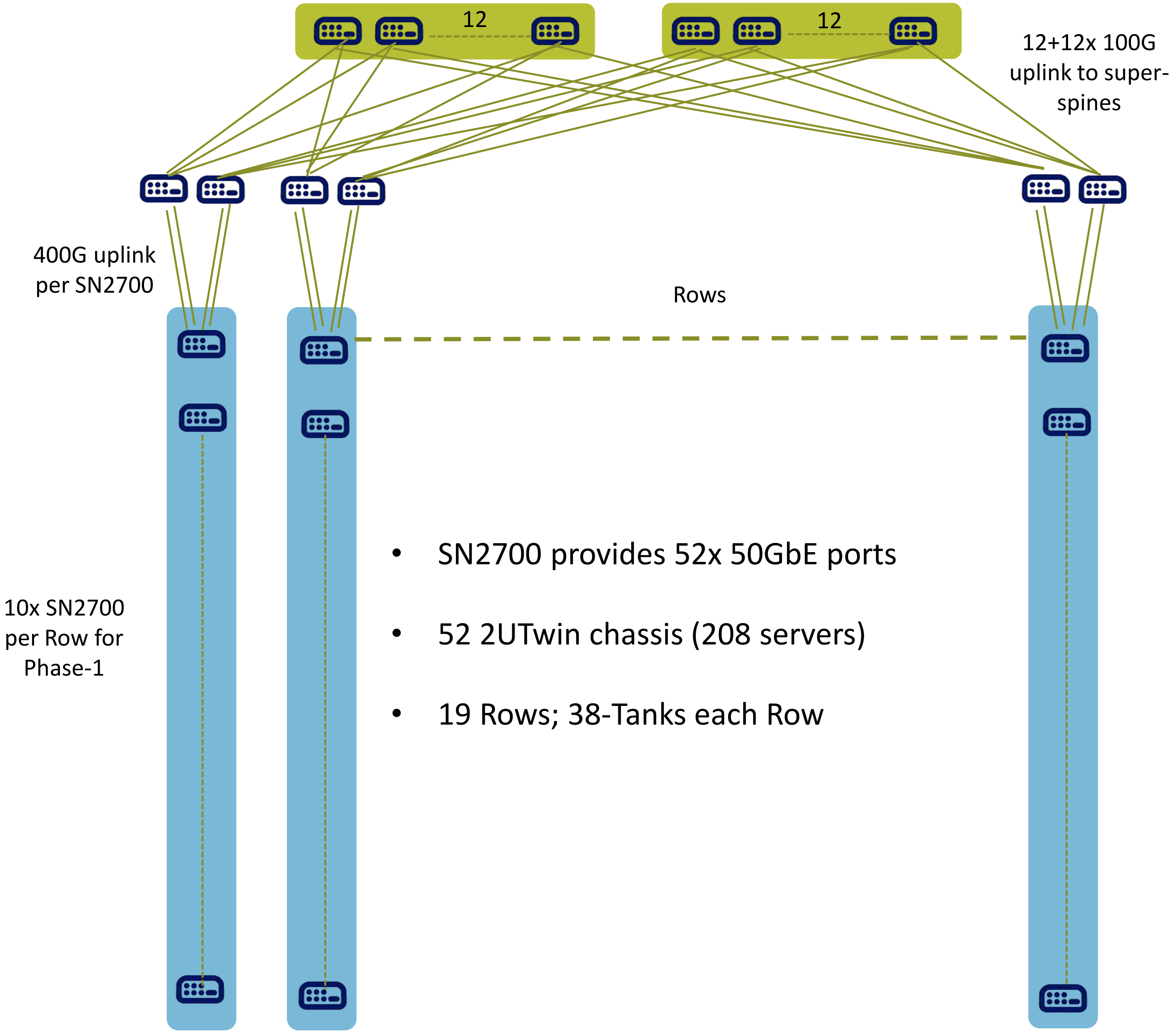
Auxiliary PCIe Extender cards on
remaining 3x servers



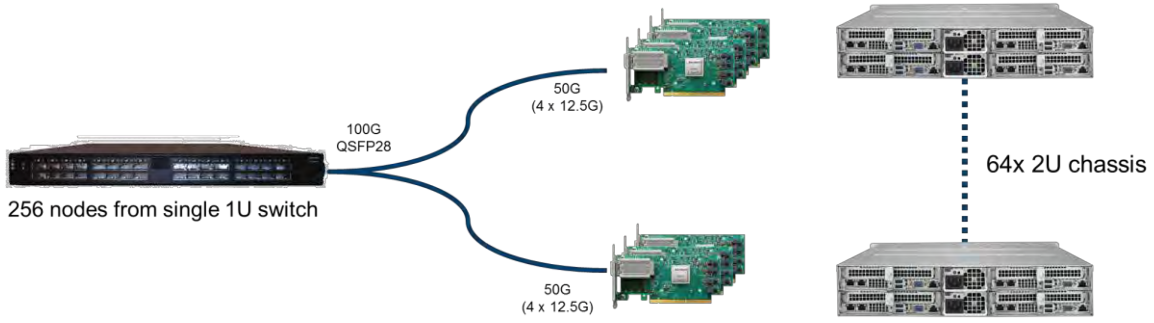
- CAPEX Savings
  - Switch ports, cables, cards
- OPEX Savings
  - Power, space and management savings

© 2019 Mellanox Technologies | Confidential

DownUnder Geo Multi-Host Network Topology


12
12
12+12x 100G uplink to super-spines
400G uplink per SN2700
Rows
10x SN2700 per Row for Phase-1

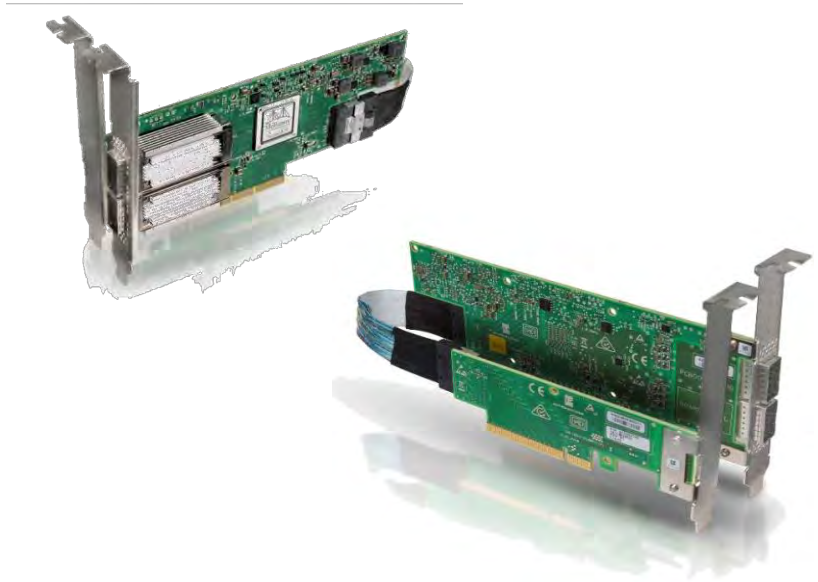
- SN2700 provides 52x 50GbE ports
- 52 2UTwin chassis (208 servers)
- 19 Rows; 38-Tanks each Row

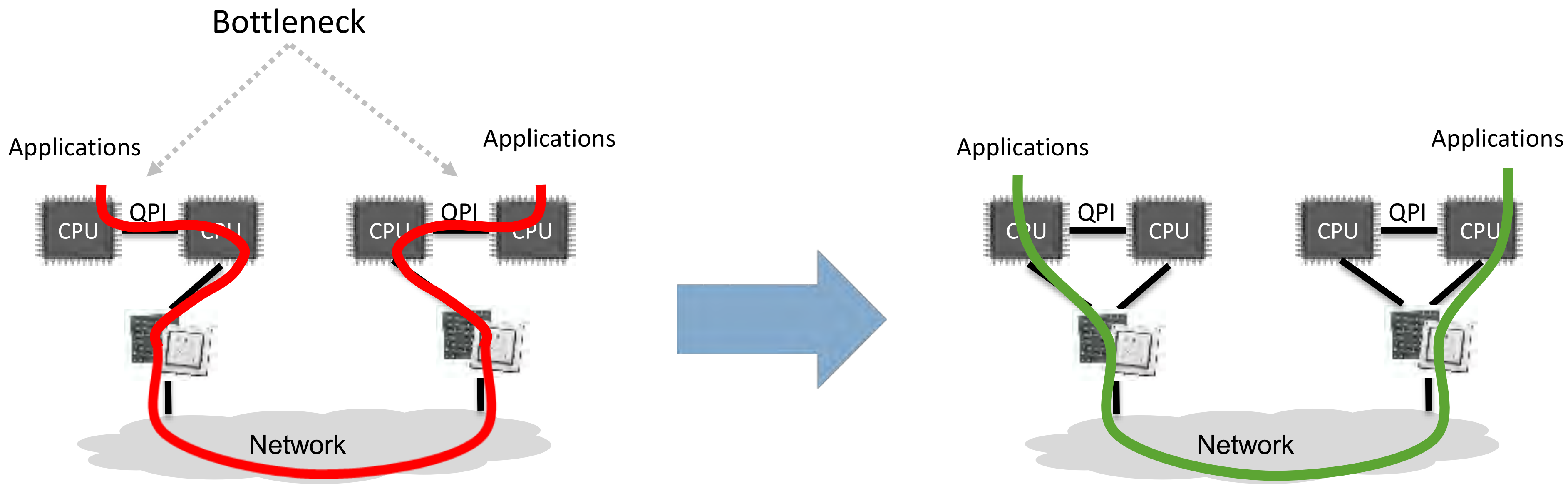


- 256 Node radix
- Embedded e-Switch within the NIC
- Can reach ~30Gb/s between the 4 servers
- A single server can peak at 30 Gb/s while other servers are not using the network
- Minimum 12.5G guaranteed to all servers simultaneously

Higher Server Performance with Socket Direct

- Overcomes CPU to CPU connectivity bottleneck
- Ensure optimal performance on both CPU sockets
- Enable GPUDirect Technology from both CPU PCI root Complex





© 2019 Mellanox Technologies | Confidential

HDR InfiniBand and

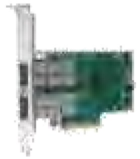
Highest-Performance 200Gb/s InfiniBand Solutions

Mellanox TECHNOLOGIES

Adapters

ConnectX-6

200Gb/s Adapter, 0.6us latency
215 million messages per second
(10 / 25 / 40 / 50 / 56 / 100 / 200Gb/s)



Switch

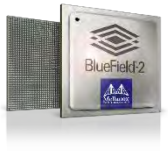
Mellanox Quantum

40 HDR (200Gb/s) InfiniBand Ports
80 HDR100 InfiniBand Ports
Throughput of 16Tb/s, <90ns Latency

SoC

BlueField-2

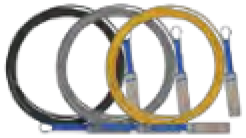
System on Chip and SmartNIC
Programmable adapter
Smart Offloads



Interconnect

LinkX™

Transceivers
Active Optical and Copper Cables
(10 / 25 / 40 / 50 / 56 / 100 / 200Gb/s)



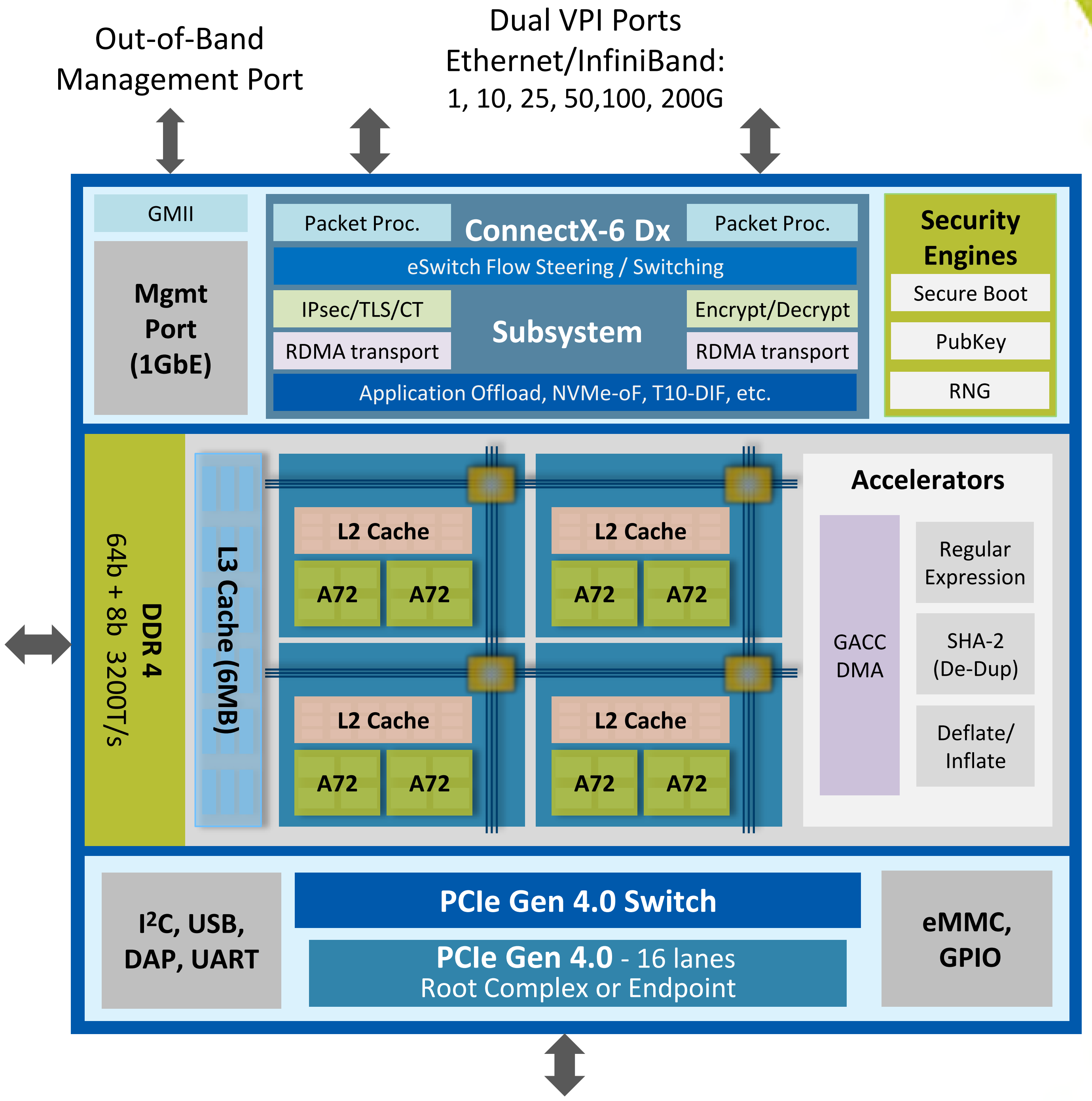
Software

HPC-X™

MPI, SHMEM/PGAS, UPC
For Commercial and Open Source Applications
Leverages Hardware Accelerations

BlueField-2 Block Diagram

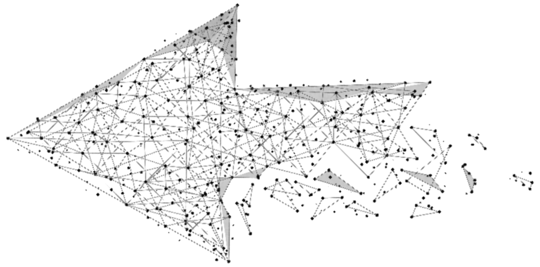
- Tile architecture running 8 x Arm ® A72 CPUs
  - SkyMesh™ coherent low-latency interconnect
  - 6MB L3 Last Level Cache
  - Arm frequency : 2GHz - 2.5GHz
- Up to 200Gb/s port bandwidth, InfiniBand or Ethernet
  - ConnectX-6 based
- Acceleration engines
  - ASAP2 switching and packet processing
  - NVMe SNAP™ storage emulation
  - IPsec/TLS data-in-motion and AES-XTS
  - Data-at-rest crypto accelerations
- Fully integrated PCIe switch
  - PCIe Gen3/4



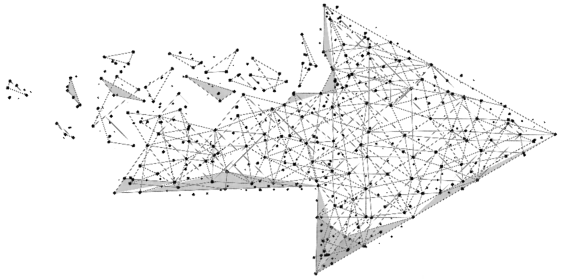
© 2019 Mellanox Technologies | Confidential

The New Architecture Vision:
Bring RDMA All the Way to the Edge

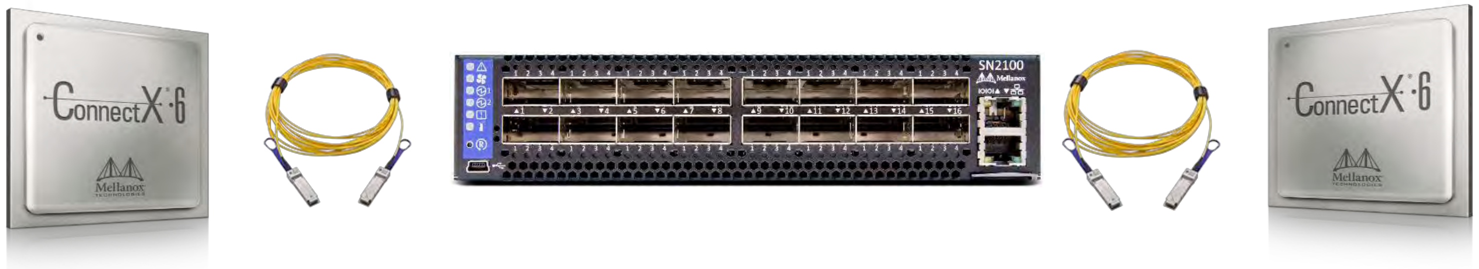
The Cloud



RDMA



The Edge



RDMA from IOT devices to the cloud

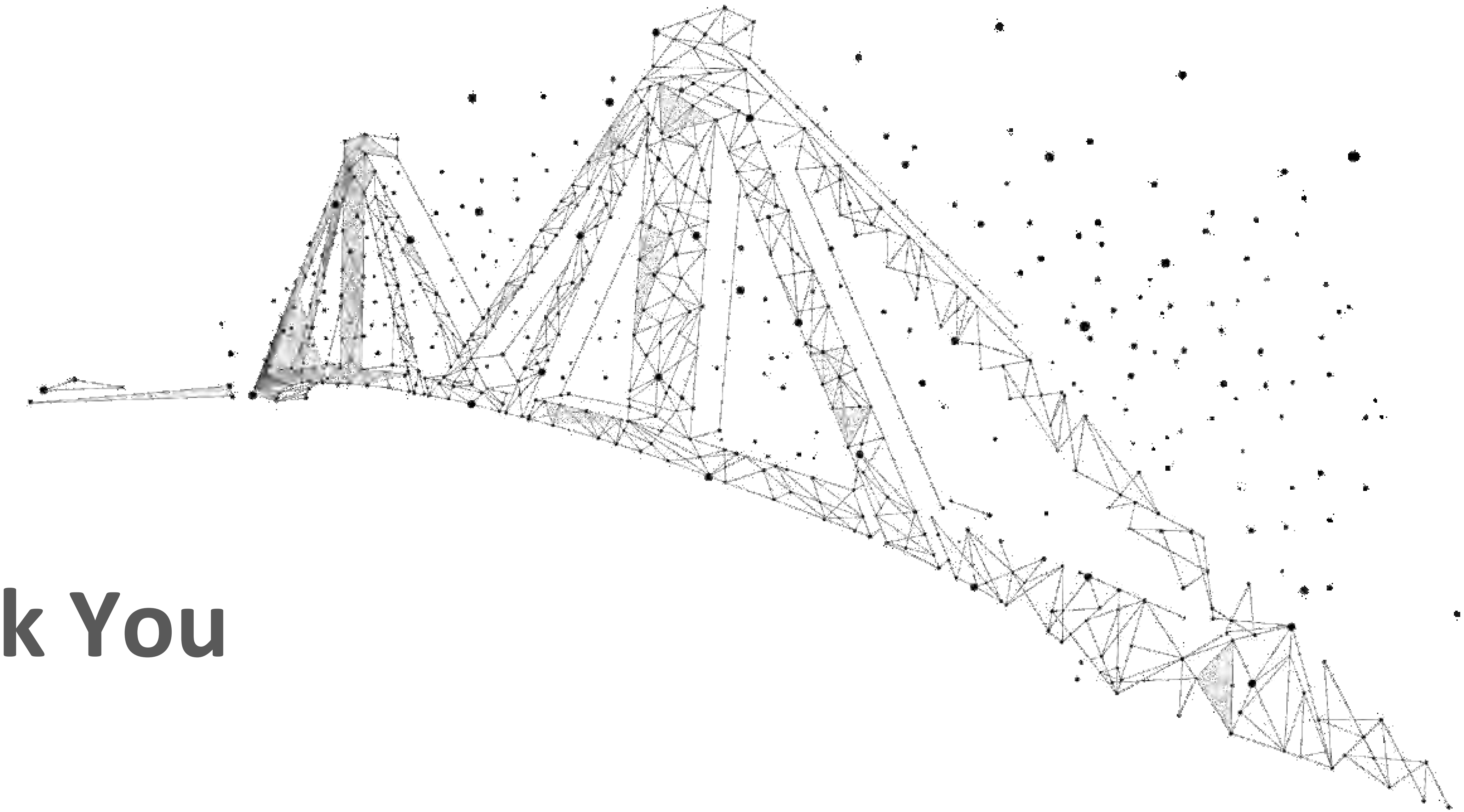
- Edge devices, autonomous cars, AI/ML appliances
- Use RDMA to move IOT data to cloud storage & processing

End-to-End Efficient RDMA Data Movement

From the Cloud back to consumer IOT devices

- Process data on the way
- Protect all customer data
- Move data Quickly, Efficiently & Securely

© 2019 Mellanox Technologies | Confidential



Thank You