Interconnect Your Future

Smart Interconnect for Next Generation HPC Platforms

Gilad Shainer, August 2016, 4th Annual MVAPICH User Group (MUG) Meeting

Mellanox TECHNOLOGIES

Connect. Accelerate. Outperform.™

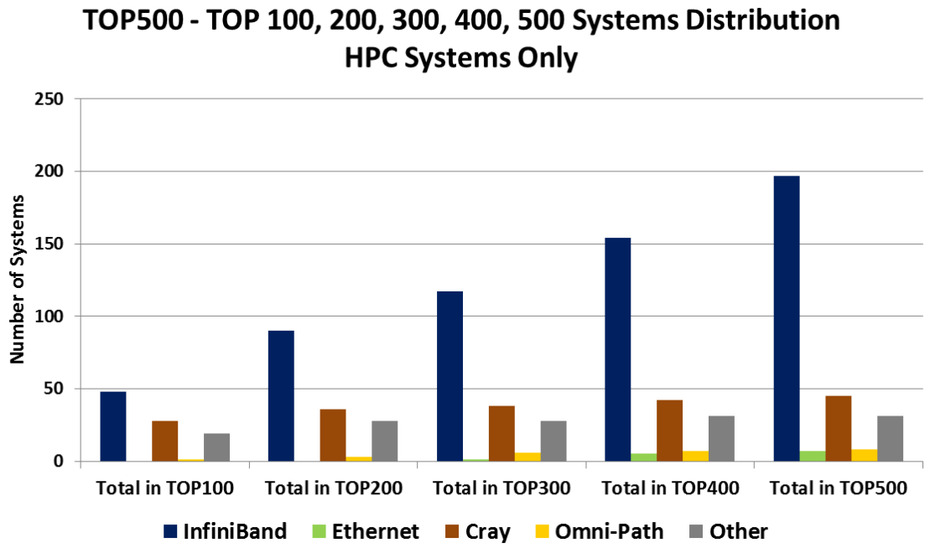
Mellanox Connects the World's Fastest Supercomputer

National Supercomputing Center in Wuxi, China

#1 on the TOP500 Supercomputing List

- 93 Petaflop performance, 3X higher versus #2 on the TOP500
- 41K nodes, 10 million cores, 256 cores per CPU
- Mellanox adapter and switch solutions

- The TOP500 list has evolved, includes HPC & Cloud / Web2.0 Hyperscale systems
- Mellanox connects 41.2% of overall TOP500 systems
- Mellanox connects 70.4% of the TOP500 HPC platforms
- Mellanox connects 46 Petascale systems, Nearly 50% of the total Petascale systems



InfiniBand is the Interconnect of Choice for HPC Compute and Storage Infrastructures

Accelerating HPC Leadership Systems

OAK RIDGE National Laboratory

“Summit” System

U.S. DEPARTMENT OF ENERGY

Lawrence Livermore National Laboratory

“Sierra” System

Proud to Pave the Path to Exascale

The Ever Growing Demand for Higher Performance

Performance Development

Terascale

Petascale

1st

"Roadrunner"

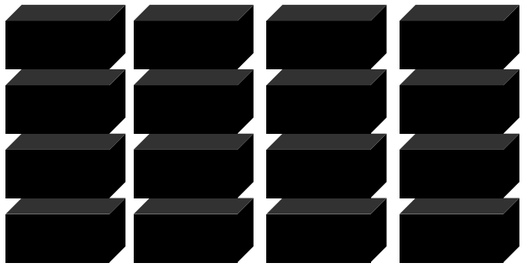
Exascale

OAK RIDGE National Laboratory
"Summit" System

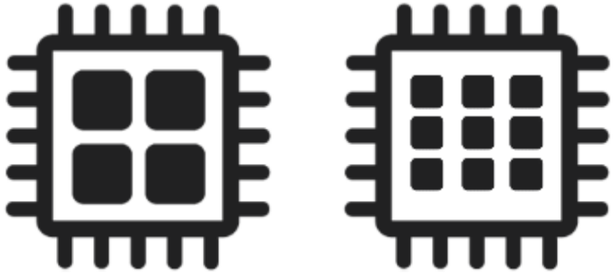
Lawrence Livermore National Laboratory
"Sierra" System

2000 2005 2010 2015 2020

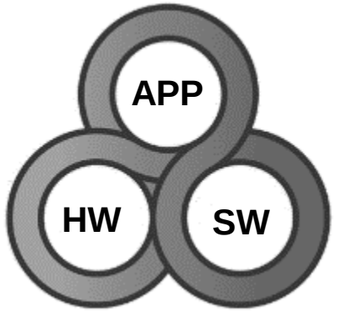
The Interconnect is the Enabling Technology



SMP to Clusters



Single-Core to Many-Core

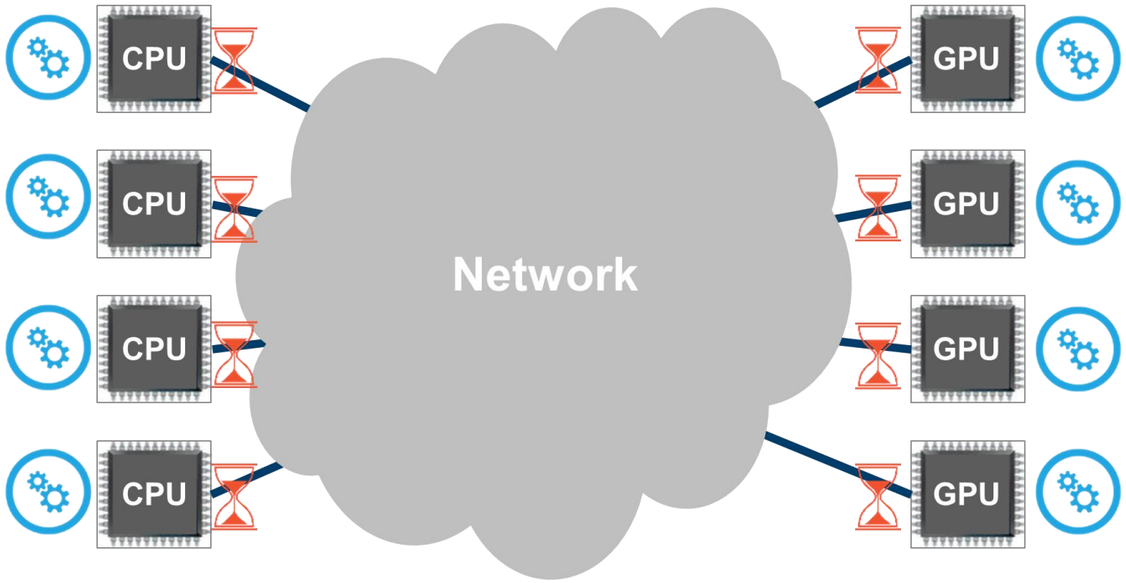


Application
Software
Hardware

Co-Design

The Intelligent Interconnect to Enable Exascale Performance

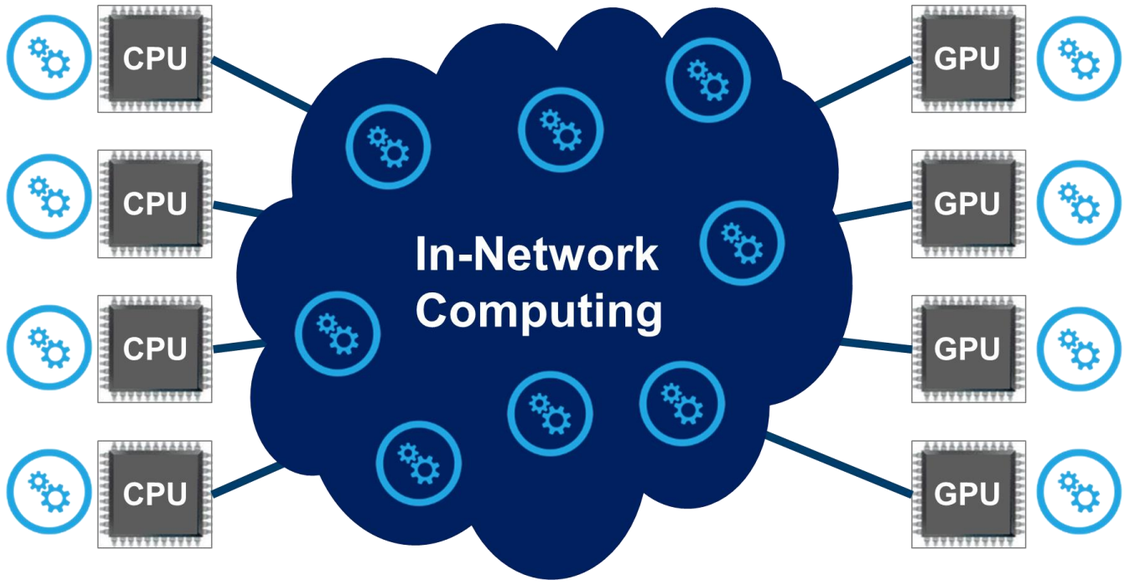
CPU-Centric



Limited to Main CPU Usage
Results in Performance Limitation

Must Wait for the Data
Creates Performance Bottlenecks

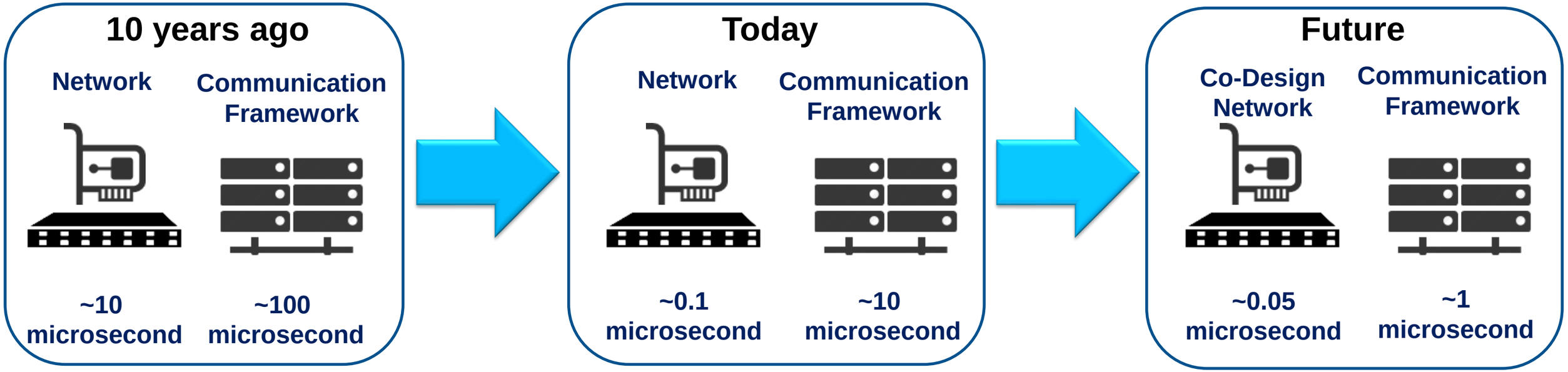
Co-Design



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

Mellanox Smart Interconnect Solutions: Switch-IB 2 and ConnectX-5

State of the Smart

a new generation of co-processors emerges

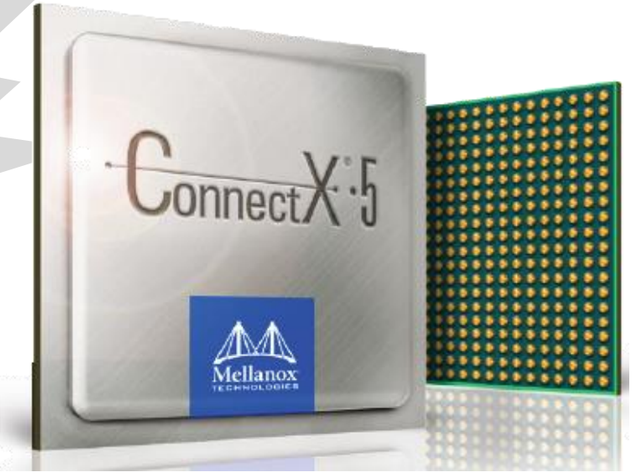
Mellanox Smart Interconnect

Switch-IB 2

SHArP

NEW!

ConnectX-5



Switch-IB 2 and ConnectX-5 Smart Interconnect Solutions

SHArP Enables Switch-IB 2 to Execute Data Aggregation / Reduction Operations in the Network

Barrier, Reduce, All-Reduce, Broadcast
Sum, Min, Max, Min-loc, max-loc, OR, XOR, AND
Integer and Floating-Point, 32 / 64 bit

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

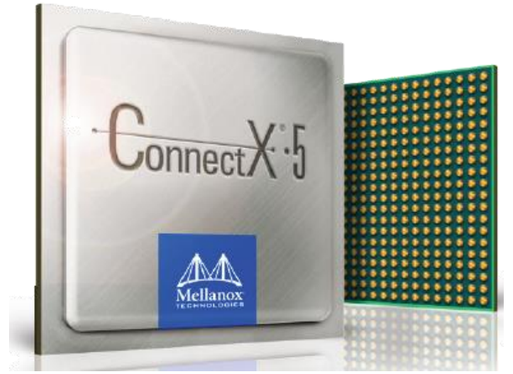
100Gb/s Throughput
0.6usec Latency (end-to-end)
200M Messages per Second

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

PCIe Gen3 and Gen4
Integrated PCIe Switch
Advanced Dynamic Routing

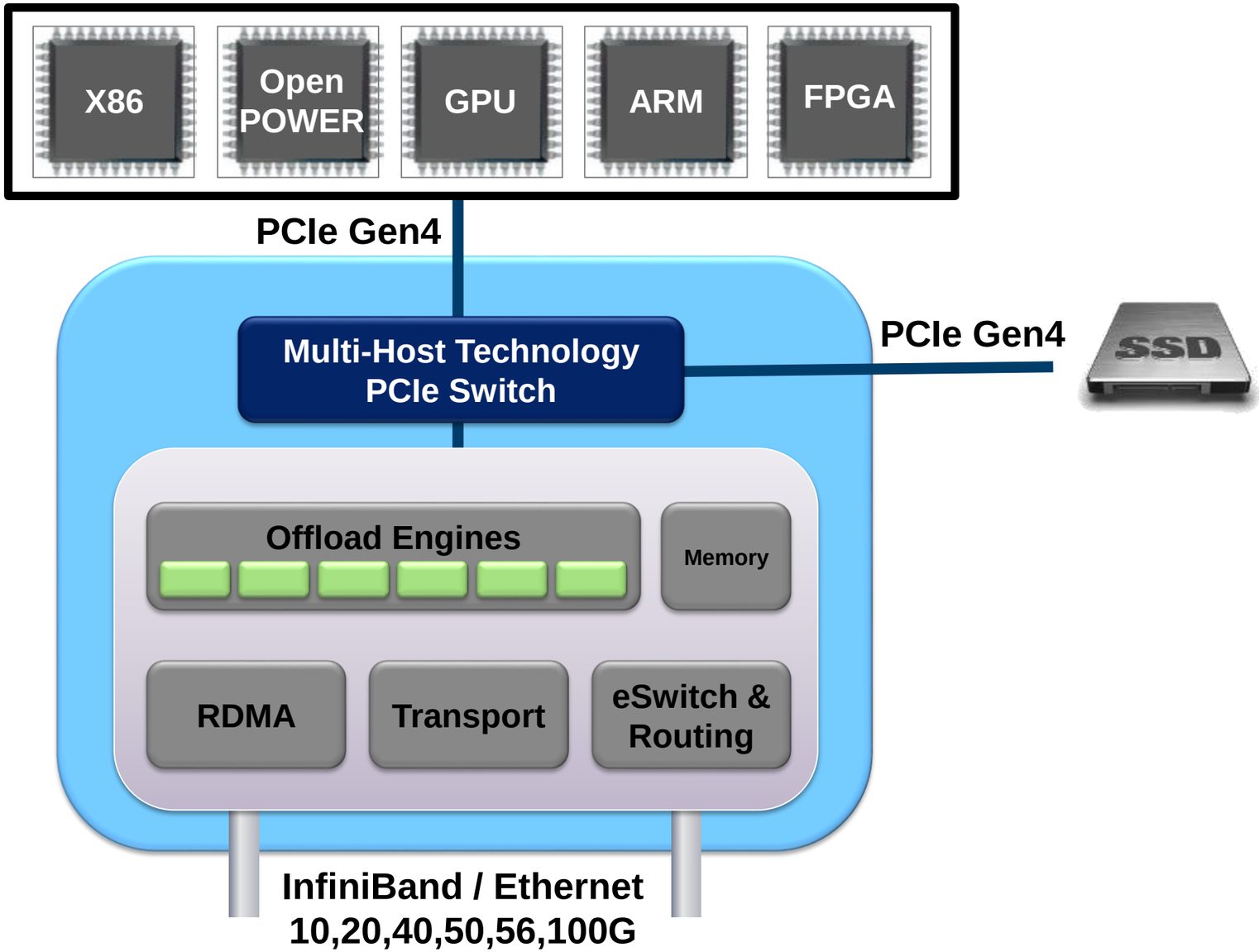
Switch-IB 2 SHArP

ConnectX-5



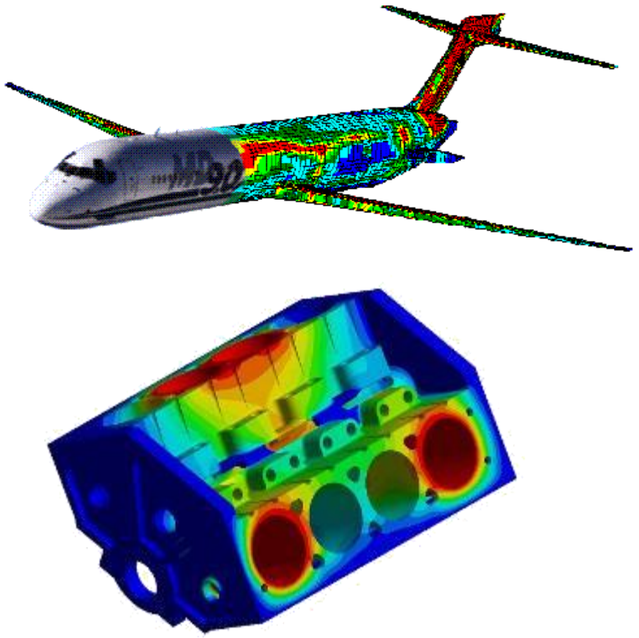
ConnectX-5 EDR 100G Architecture

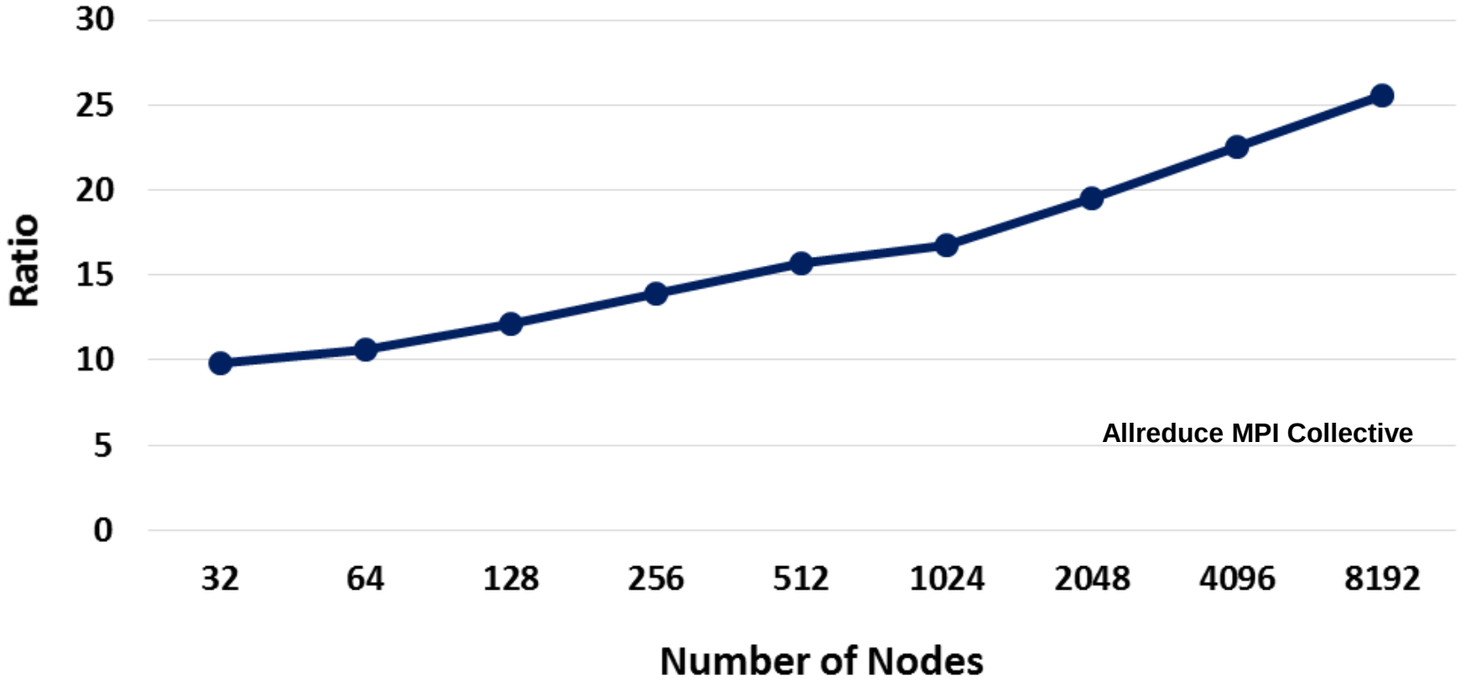
Breakthrough in Performance & Total Cost of Ownership!



Switch-IB 2 SHArP Performance Advantage

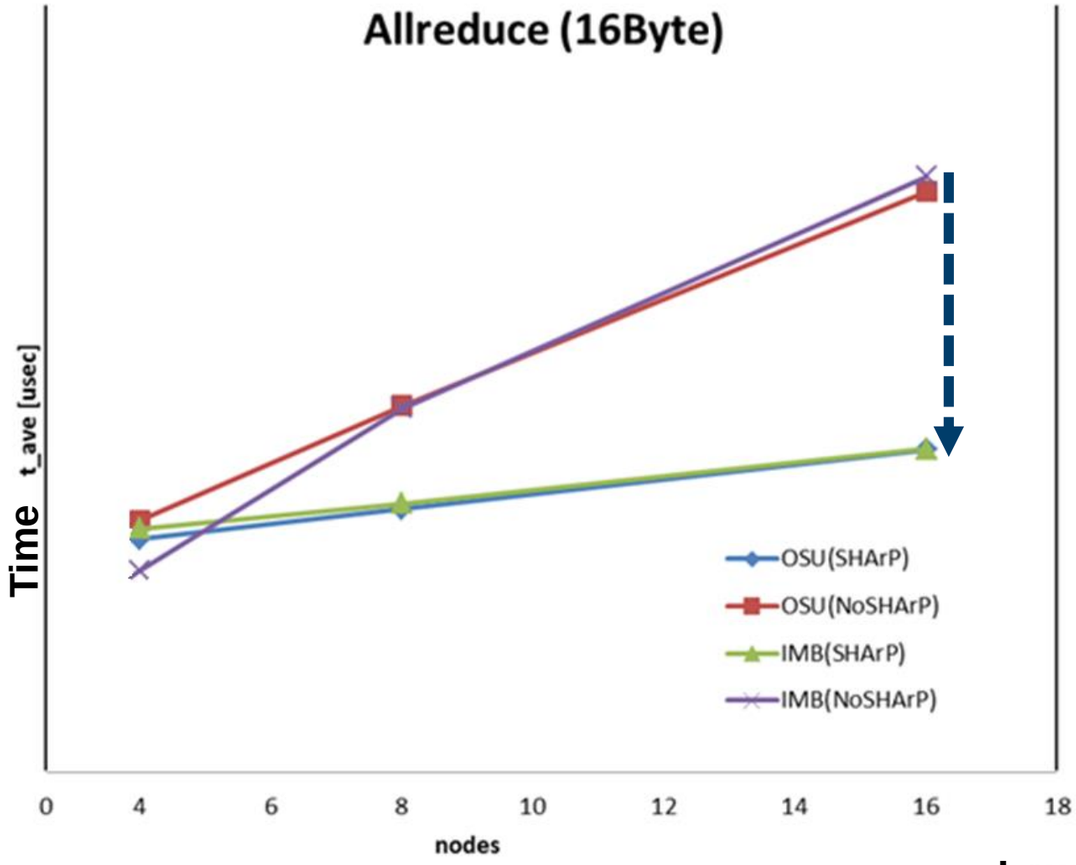
- MiniFE is a Finite Element mini-application
  - Implements kernels that represent implicit finite-element applications





10X to 25X Performance Improvement

SHArP Performance Advantage with Intel Xeon Phi Knight Landing



Lower is better

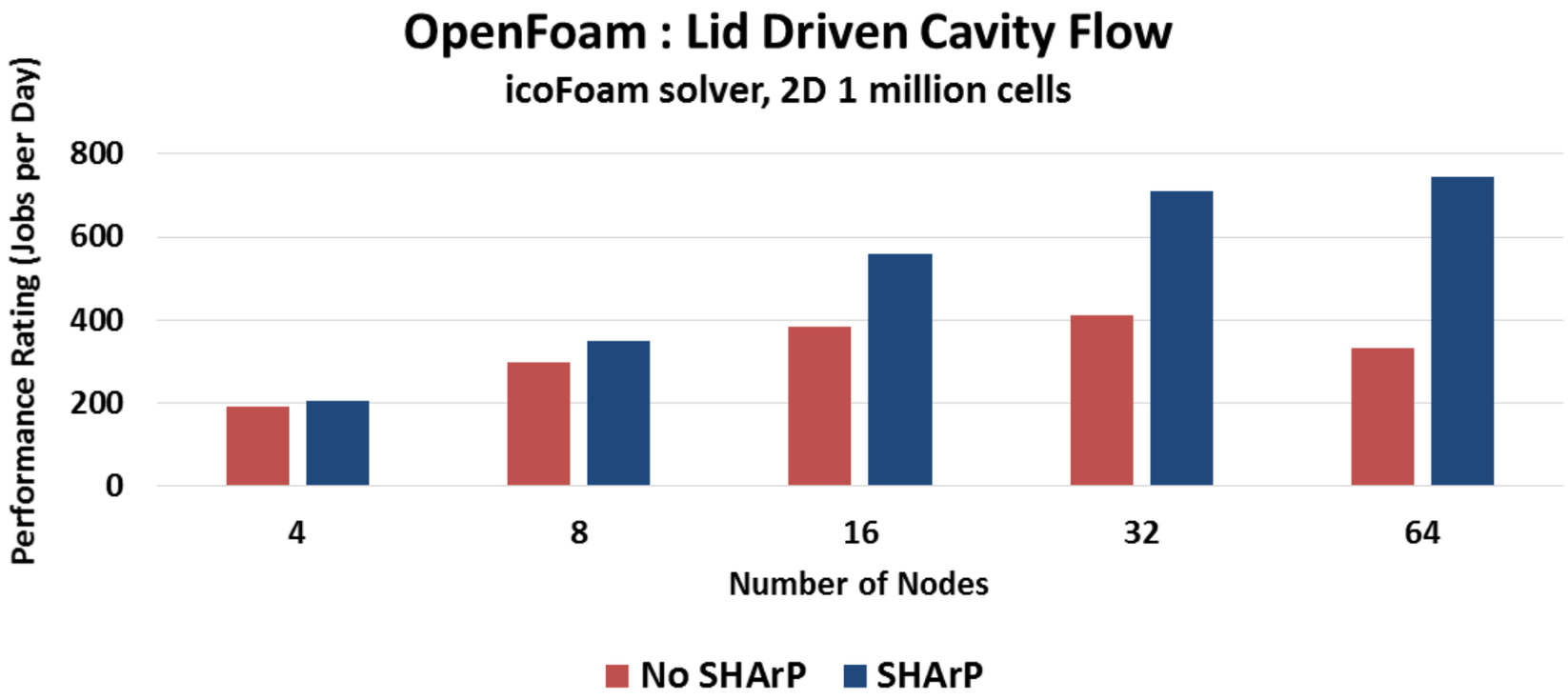
OSU - OSU MPI benchmark; IMB – Intel MPI Benchmark

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

SHArP Technology Performance

OpenFOAM

OpenFOAM is a popular computational fluid dynamics application



SHArP Delivers 2.2X Higher Performance

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

Storage

NVMe Flash Storage Arrays

Scale-Out Storage (NVMe over Fabric)



NFV

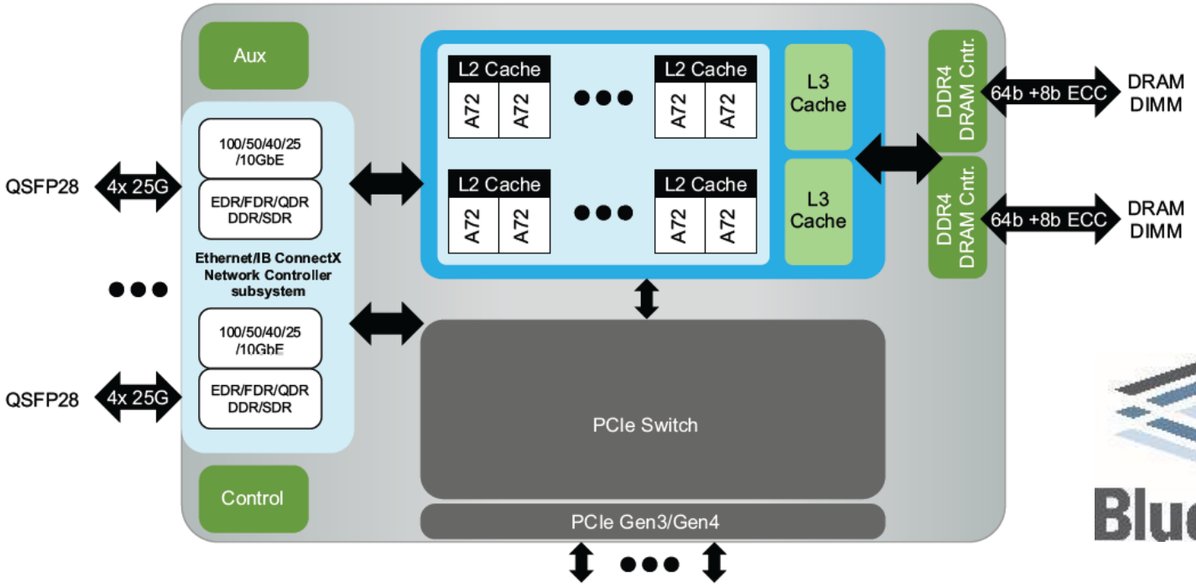
Accelerating & Virtualizing VNFs

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

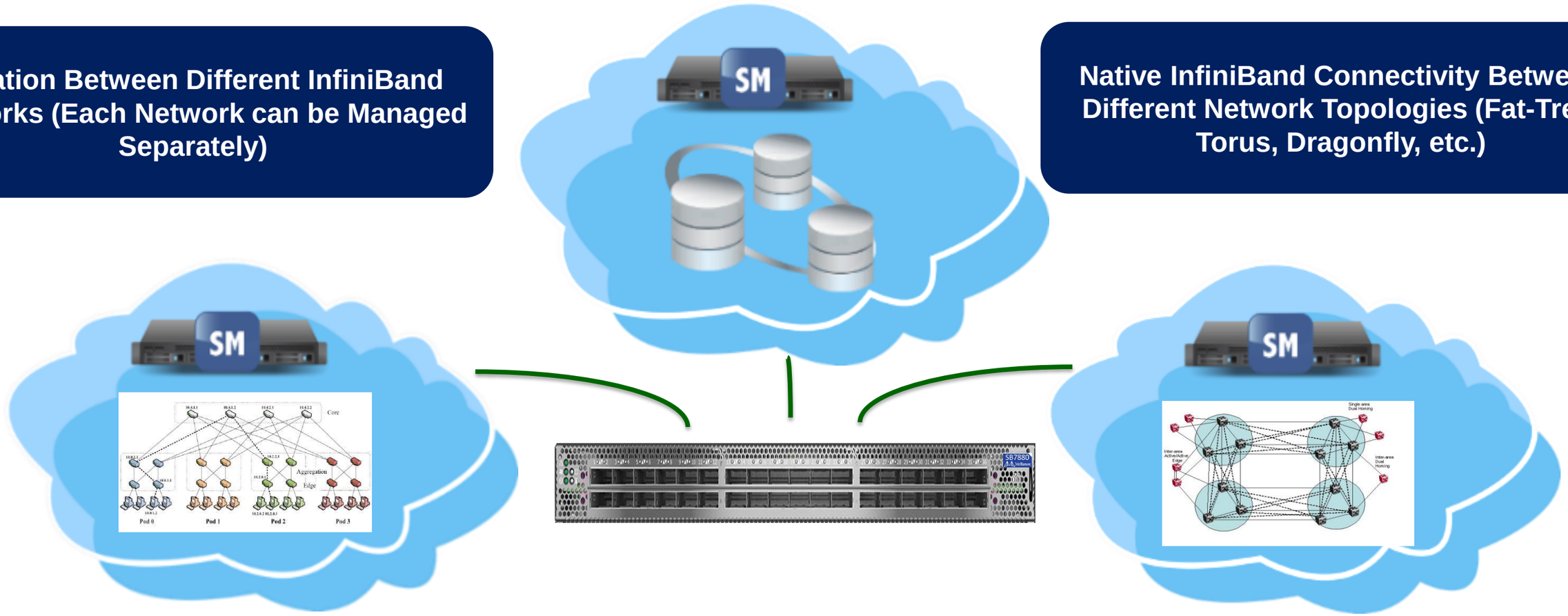


BlueField

InfiniBand Router Solutions

Isolation Between Different InfiniBand Networks (Each Network can be Managed Separately)

Native InfiniBand Connectivity Between Different Network Topologies (Fat-Tree, Torus, Dragonfly, etc.)



SB7780 Router 1U

Supports up to 6 Different Subnets

Technology Roadmap

Petascale

Exascale

#1 TOP500, 100Petaflop

OAK RIDGE National Laboratory
"Summit" System

Lawrence Livermore National Laboratory
"Sierra" System

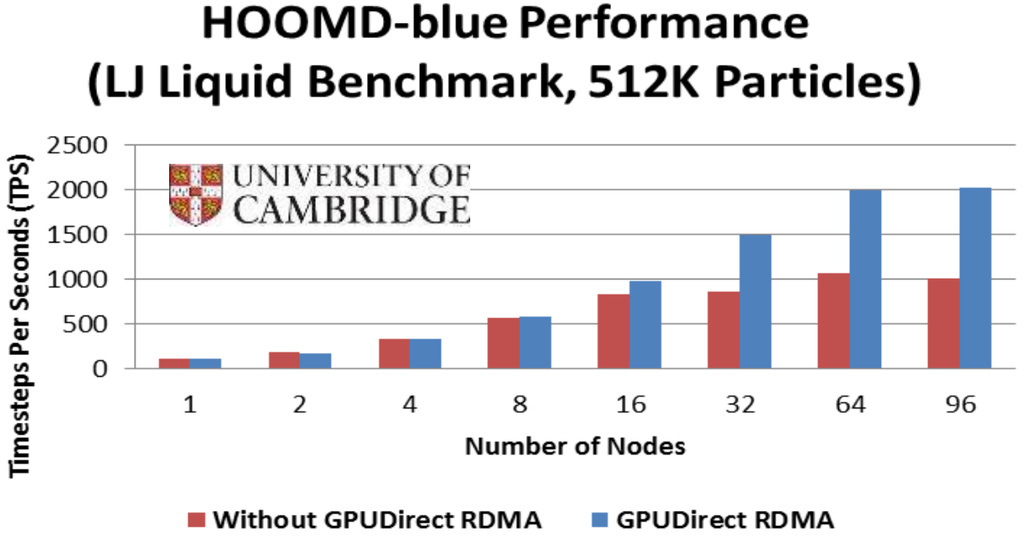
World-wide Programs

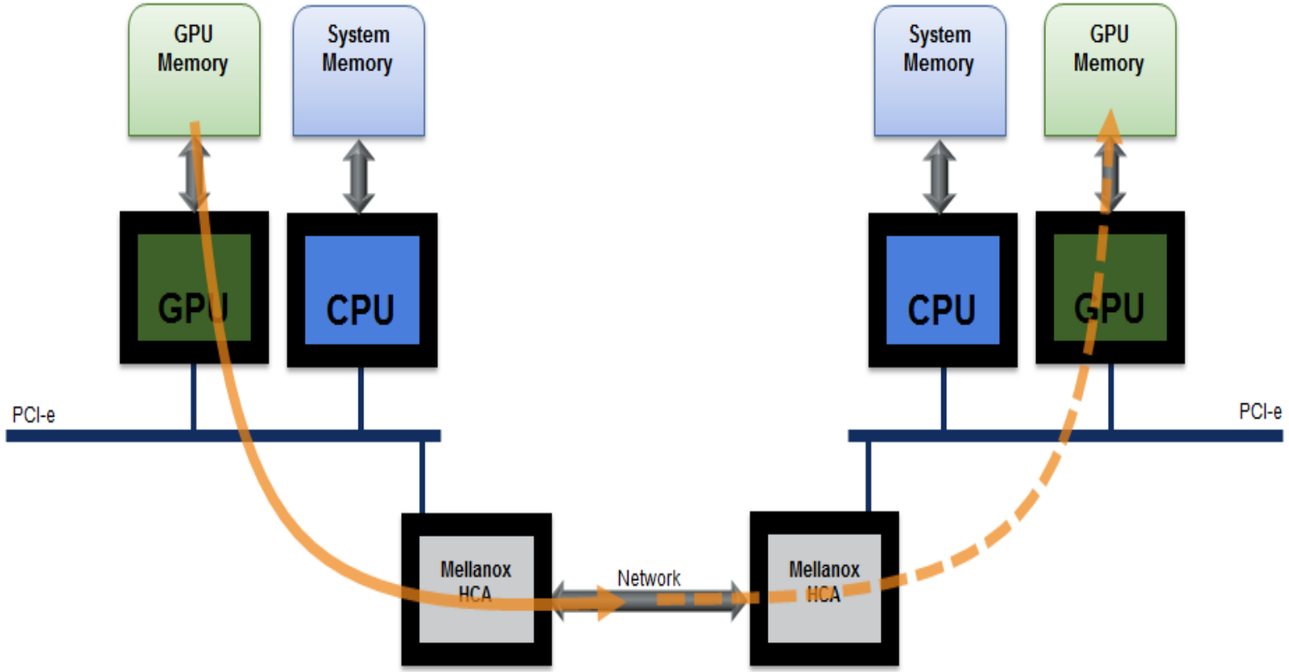
2016 **2019** **2021**

GPUDirect RDMA Technology

Maximize Performance via Accelerator and GPU Offloads

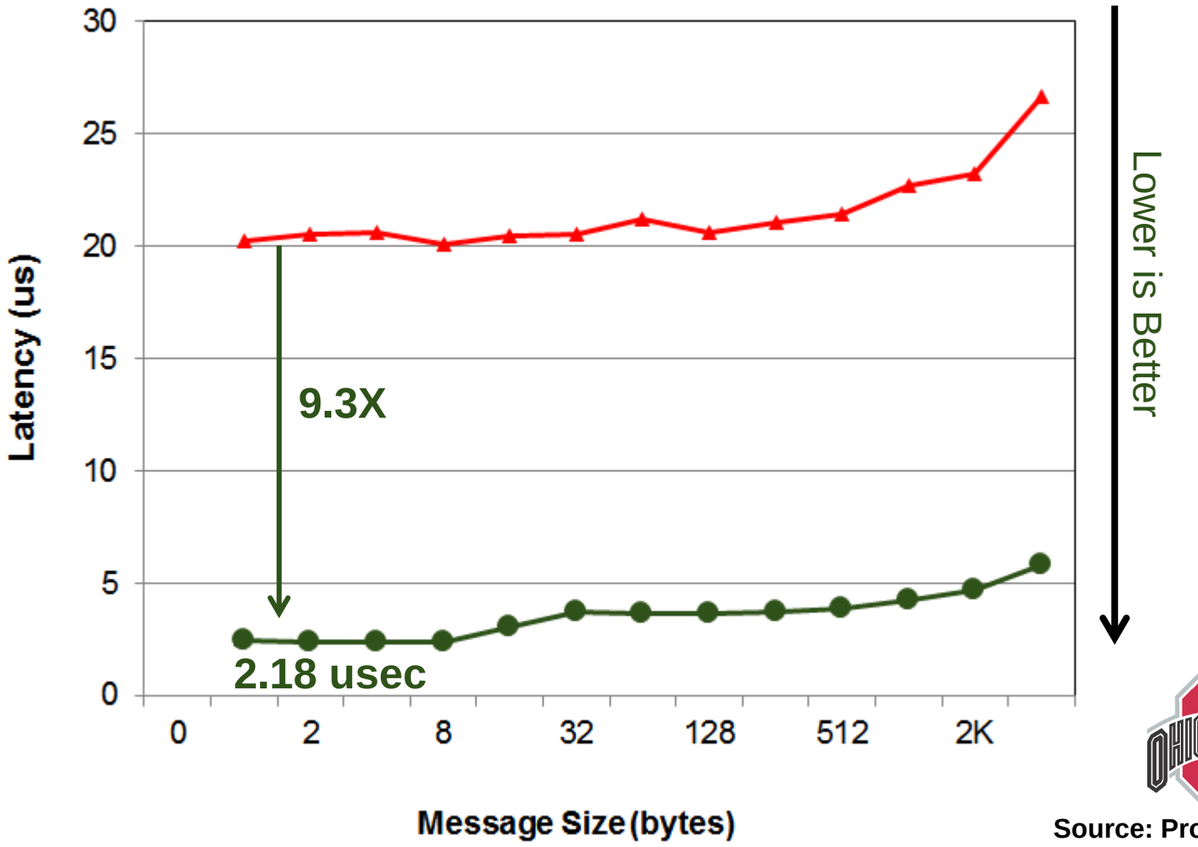
GPUs are Everywhere!



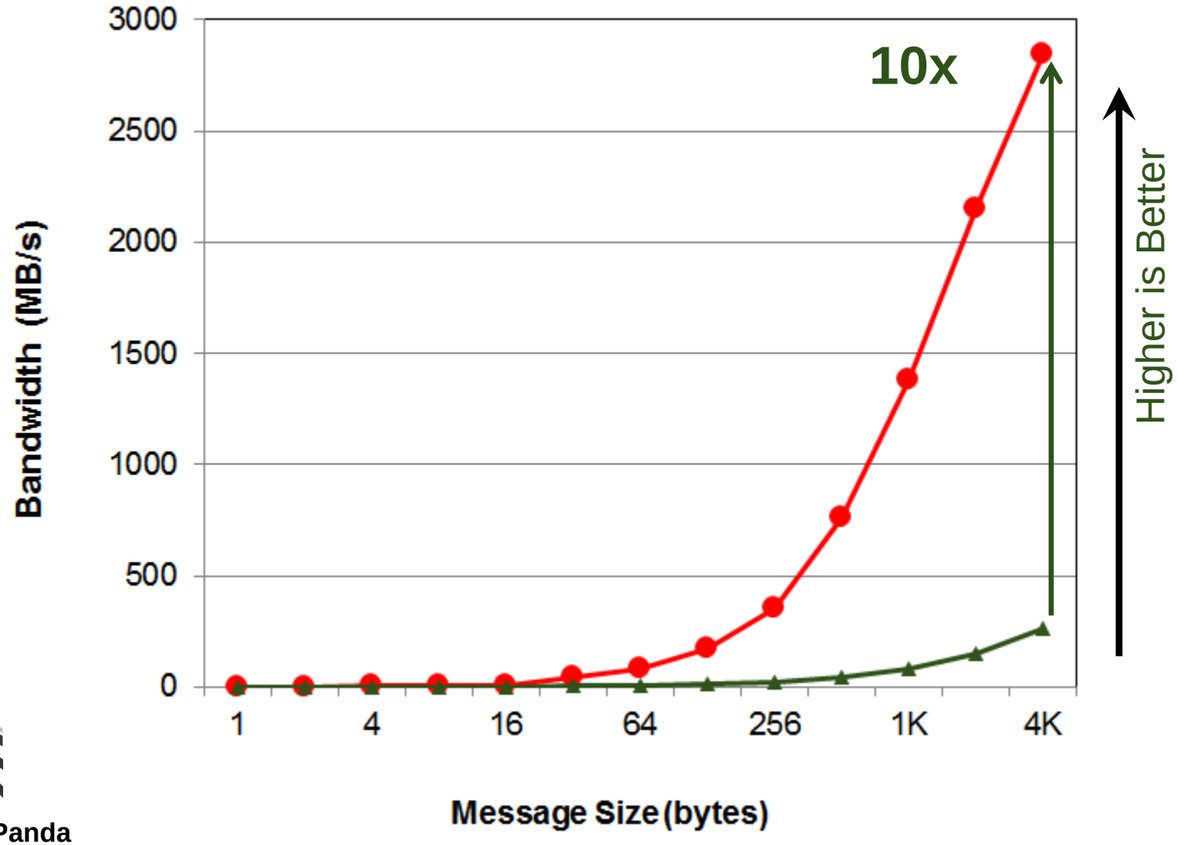


Performance of MPI with GPUDirect RDMA

GPU-GPU Internode MPI Latency



GPU-GPU Internode MPI Bandwidth



Source: Prof. DK Panda

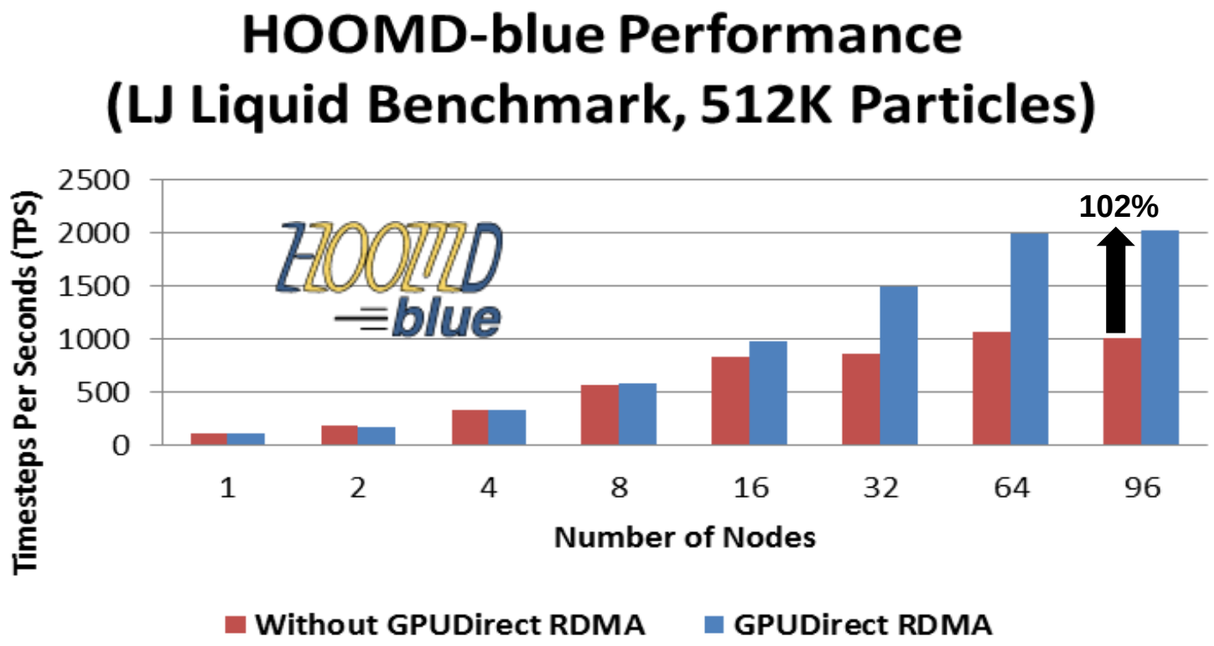
88% Lower Latency

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

UNIVERSITY OF CAMBRIDGE



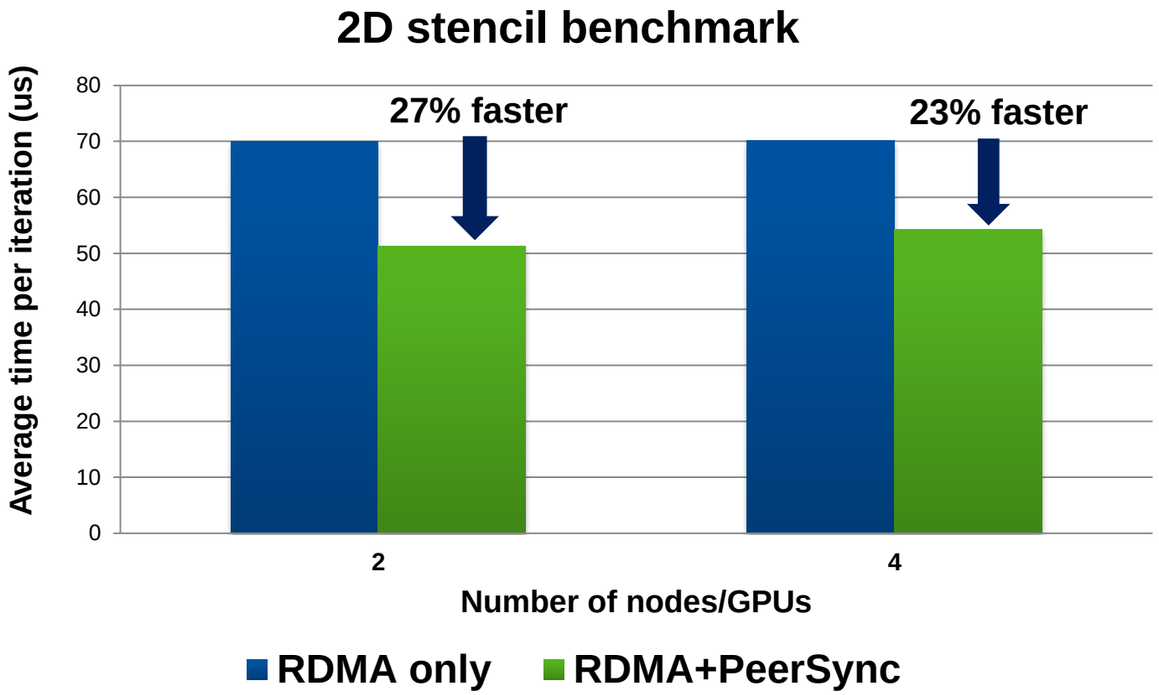
2X Application Performance!

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

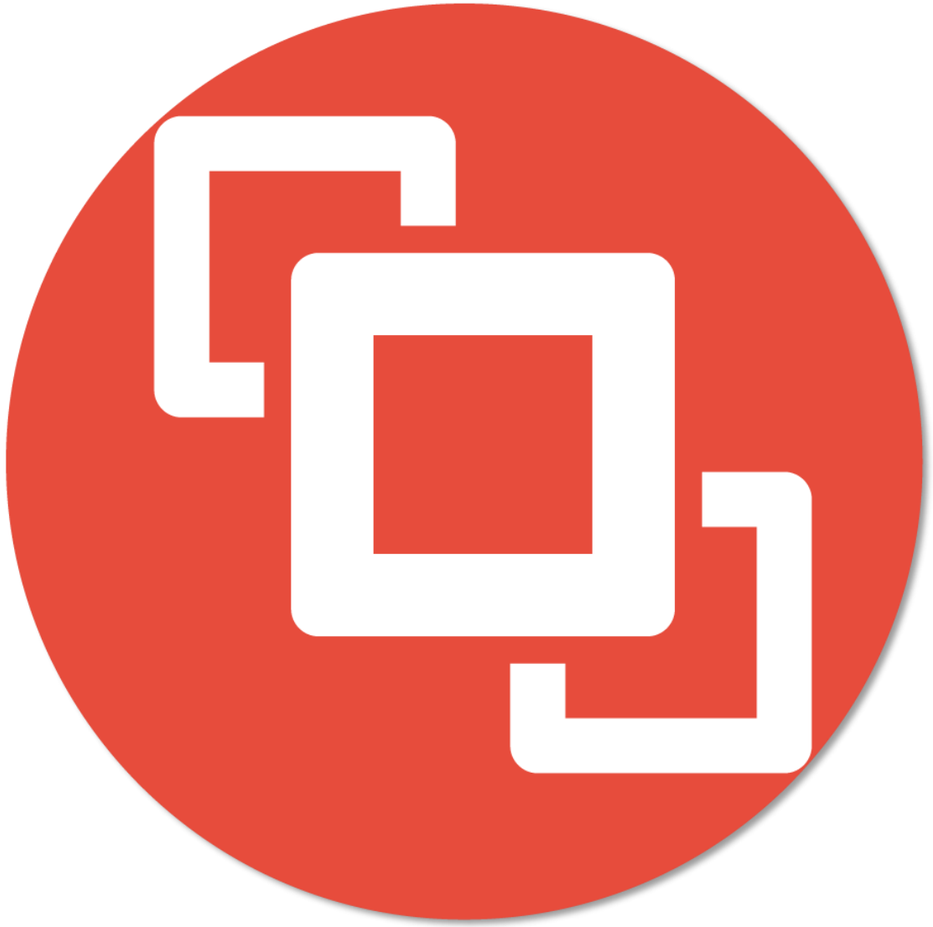
- GPUDirect Sync (GPUDirect 4.0)
  - Both data path and control path go directly between the GPU and the Mellanox interconnect

Maximum Performance For GPU Clusters



UCX

Unified Communication X



UCX

Unified Communication - X Framework

www.openucx.org
ucx-group@email.ornl.gov

Mellanox TECHNOLOGIES Connect. Accelerate. Outperform.™
OAK RIDGE National Laboratory
NVIDIA
IBM
ICL UT
PathScale High Performance Compilers

Background

MXM

- Developed by Mellanox Technologies
- HPC communication library for InfiniBand devices and shared memory
- Primary focus: MPI, PGAS

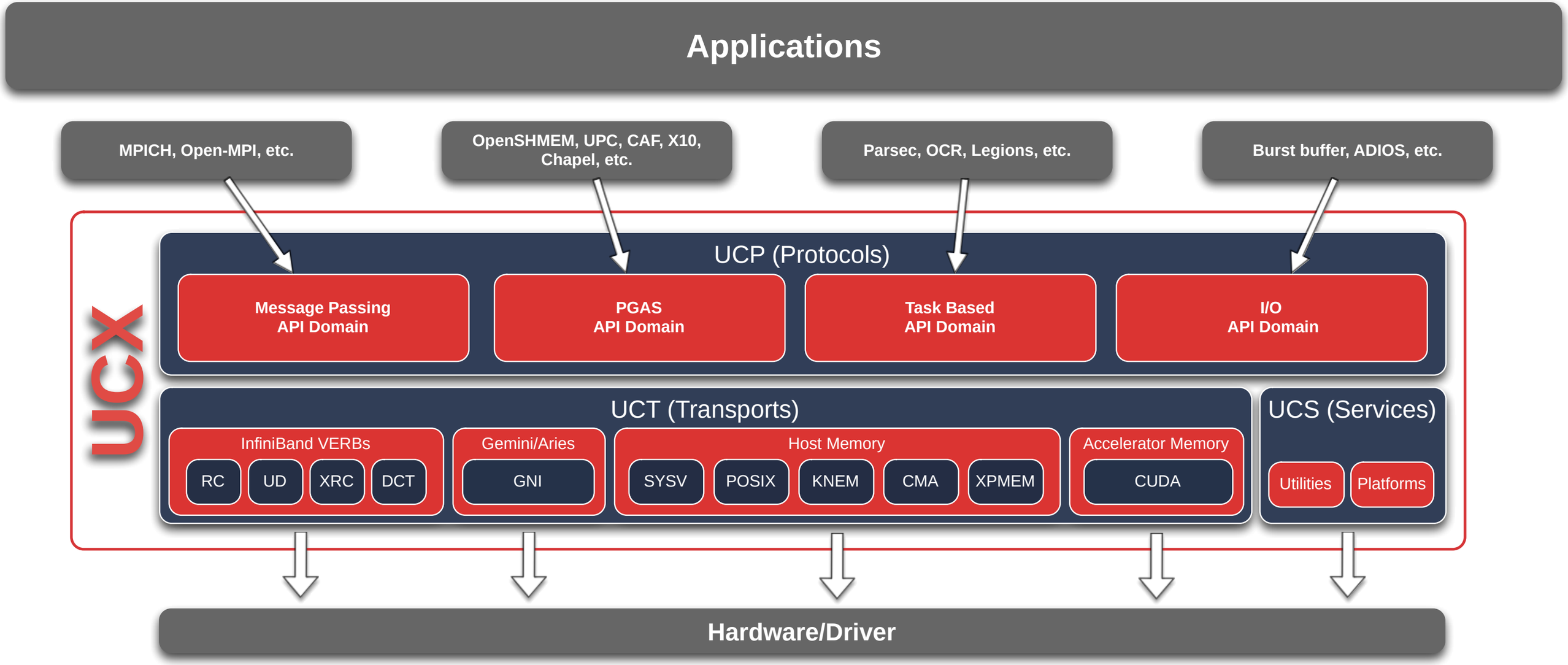
UCCS

- Developed by ORNL, UH, UTK
- Originally based on Open MPI BTL and OPAL layers
- HPC communication library for InfiniBand, Cray Gemini/Aries, and shared memory
- Primary focus: OpenSHMEM, PGAS
- Also supports: MPI

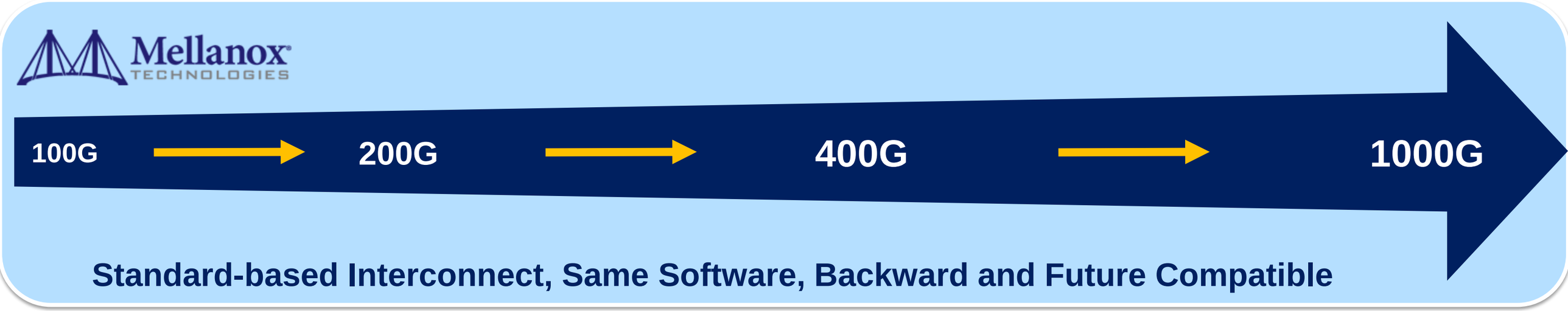
PAMI

- Developed by IBM on BG/Q, PERCS, IB VERBS
- Network devices and shared memory
- MPI, OpenSHMEM, PGAS, CHARM++, X10
- C++ components
- Aggressive multi-threading with contexts
- Active Messages
- Non-blocking collectives with hw accleration support

High-level Overview



Technology Roadmap



Petascale

Exascale

#1 TOP500, 100Petaflop

OAK RIDGE National Laboratory
"Summit" System

Lawrence Livermore National Laboratory
"Sierra" System

World-wide Programs

2016 **2019** **2021**

Thank You

Mellanox TECHNOLOGIES

Connect. Accelerate. Outperform.™