In-Network Computing

Paving the Road to Exascale

August 2018

Mellanox Accelerates Leading HPC and AI Systems

World’s Top 3 Supercomputers

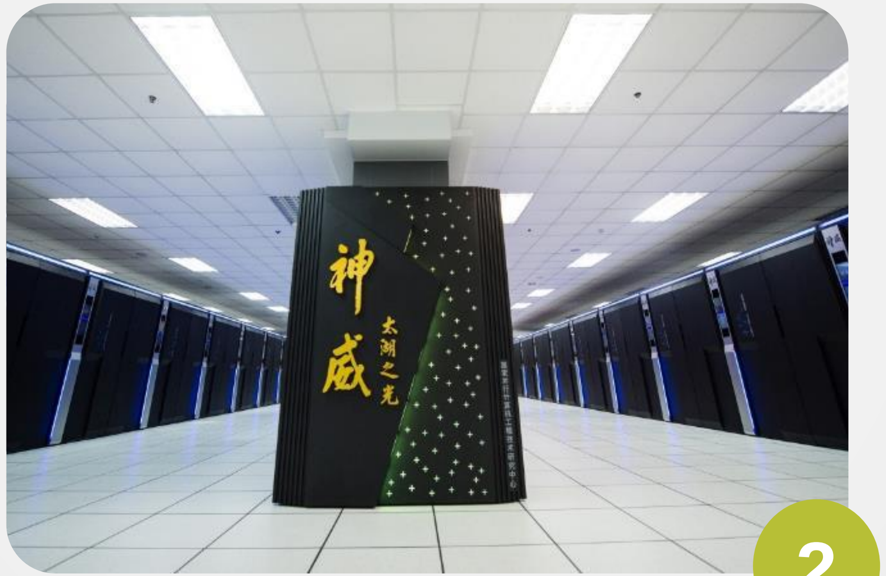
Oak Ridge National Laboratory
U.S. Department of Energy

1

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

TOP 500 The List.

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



2

Wuxi Supercomputing Center
Fastest Supercomputer in China
41K InfiniBand Nodes

TOP 500 The List.

Lawrence Livermore National Laboratory
U.S. Department of Energy

3

Sierra CORAL System
#2 USA Supercomputer
8.6K InfiniBand Nodes

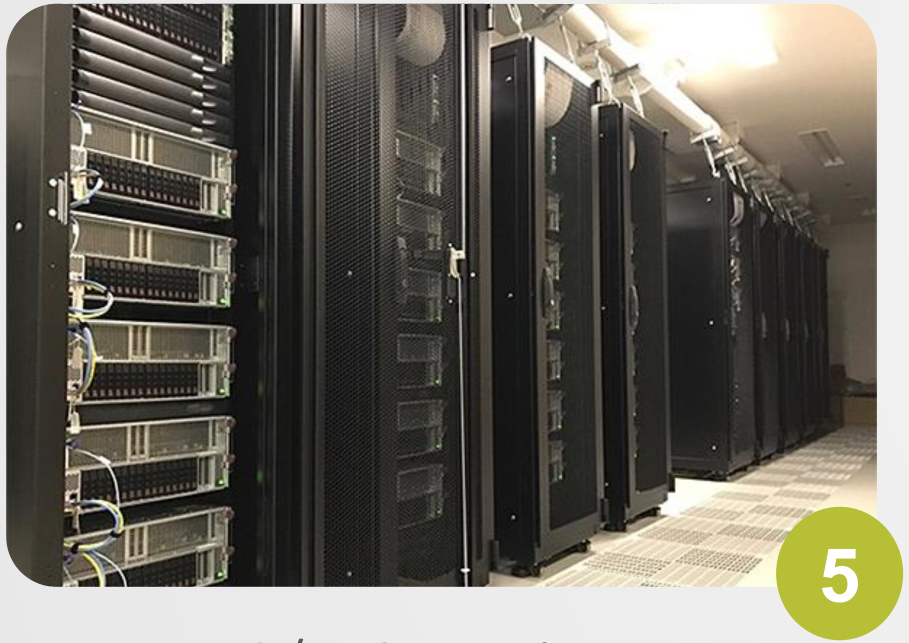
TOP 500 The List.

Mellanox Accelerates Leading HPC and AI Systems

(Examples)

AIST

NATIONAL INSTITUTE OF ADVANCED INDUSTRIAL SCIENCE AND TECHNOLOGY (AIST)



Fastest HPC / AI System in Japan
1.1K InfiniBand Nodes

TOP 500 The List.

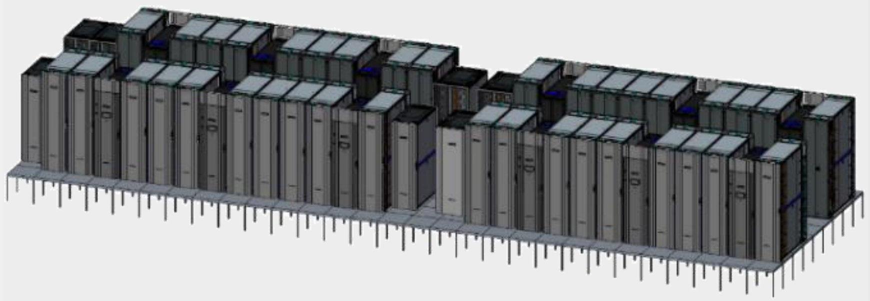
eni

The world's Fastest Industry Supercomputer
1.6K InfiniBand Nodes

TOP 500 The List.

Sandia National Laboratories

NNSA National Nuclear Security Administration



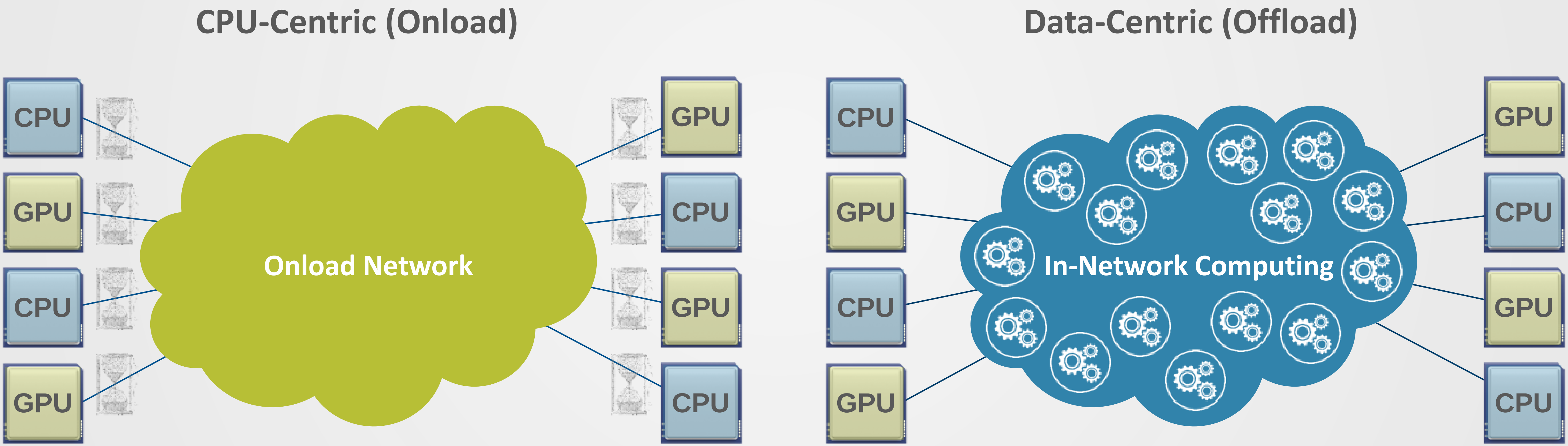
‘Astra’ Arm-Based Supercomputer
NNSA Vanguard Program
2.6K InfiniBand Nodes

To be Listed Nov’18 (TOP100)

The Need for Intelligent and Faster Interconnect

Faster Data Speeds and In-Network Computing

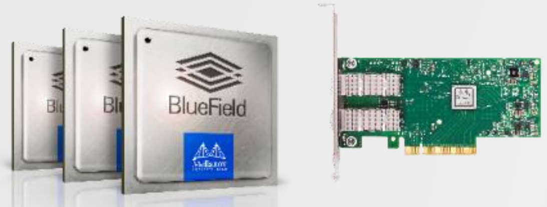
Enable Higher Performance and Scale



Must Wait for the Data
Creates Performance Bottlenecks

Analyze Data as it Moves!
Higher Performance and Scale

HPC and AI Needs the Most Intelligent Interconnect



SmartNIC

BlueField

System on a Chip

Innova ConnectX

Adapters

Spectrum Quantum

Switches

LinkX

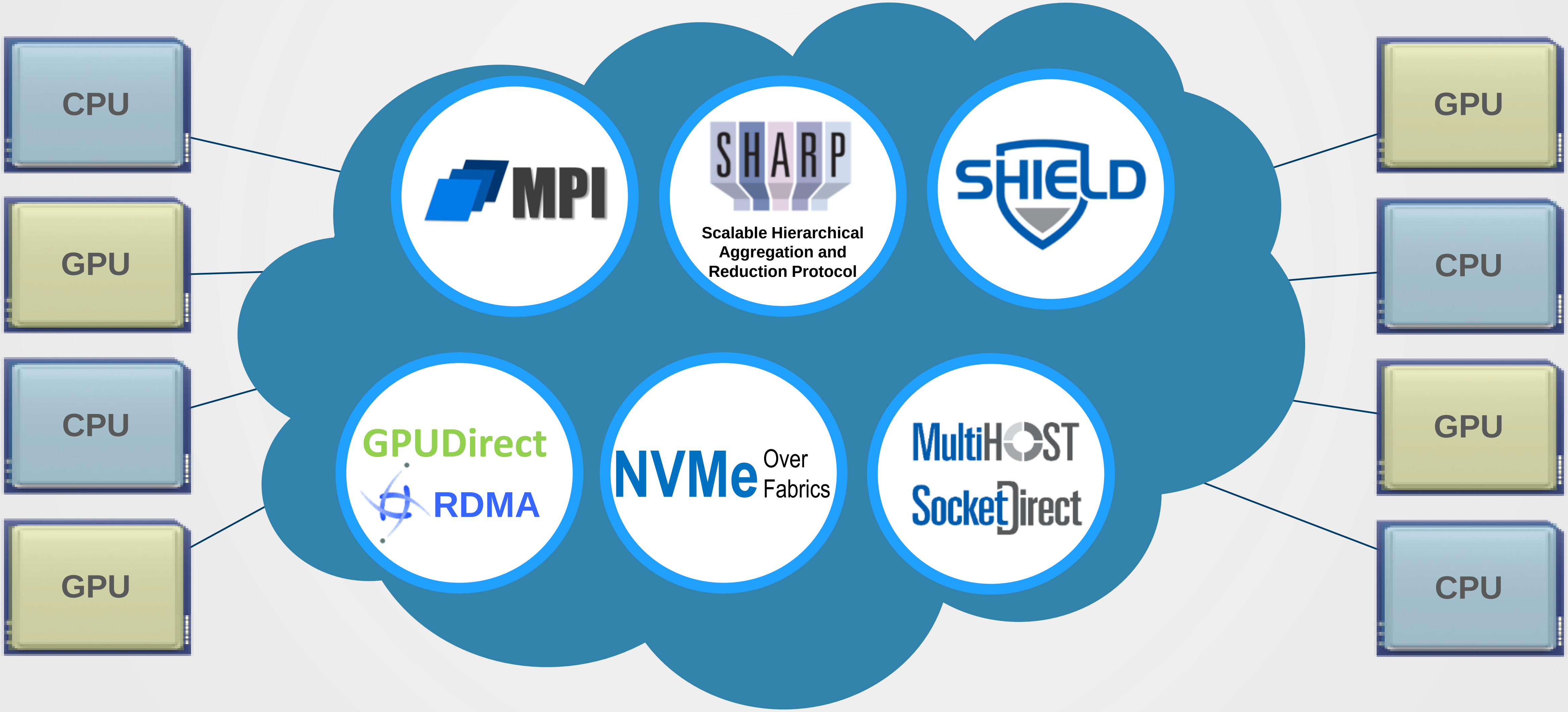
Cables & Transceivers

Higher Data Speeds

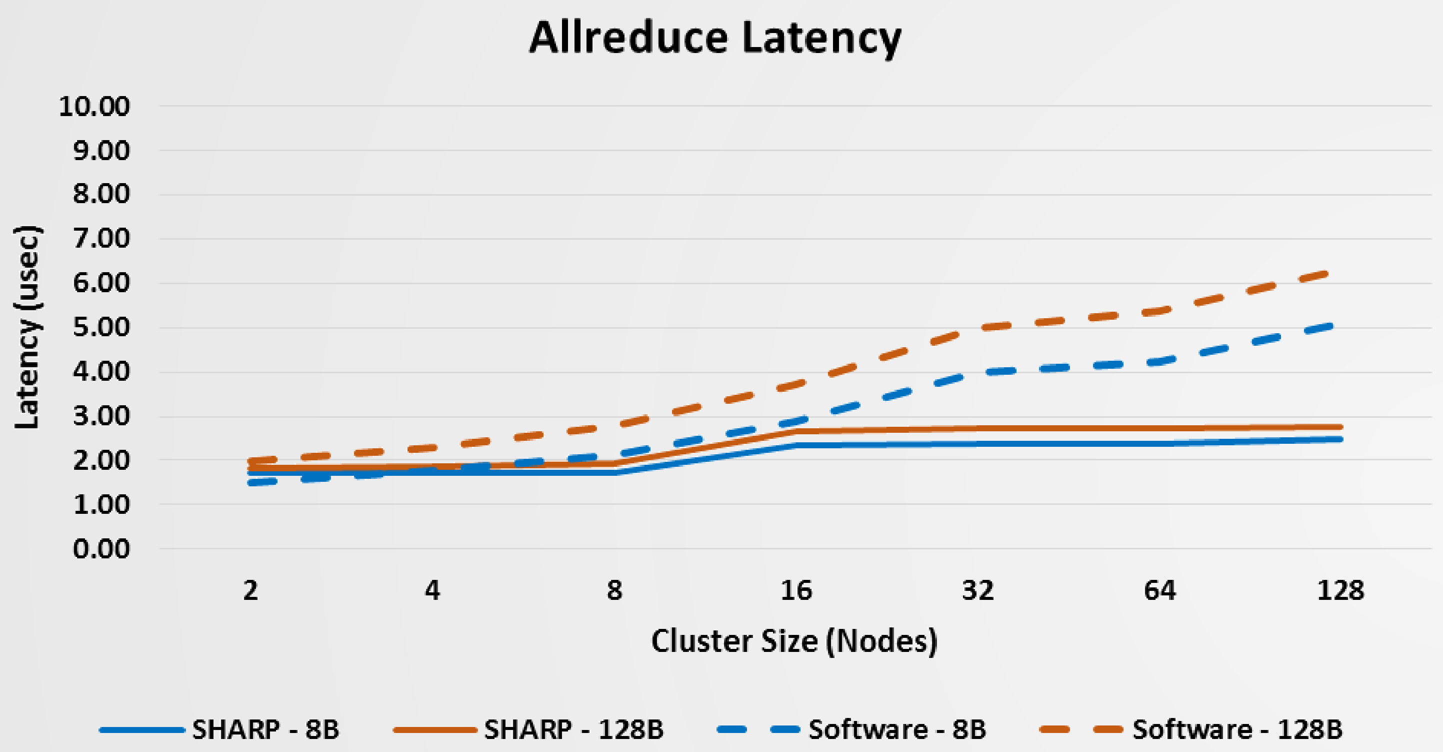
Faster Data Processing

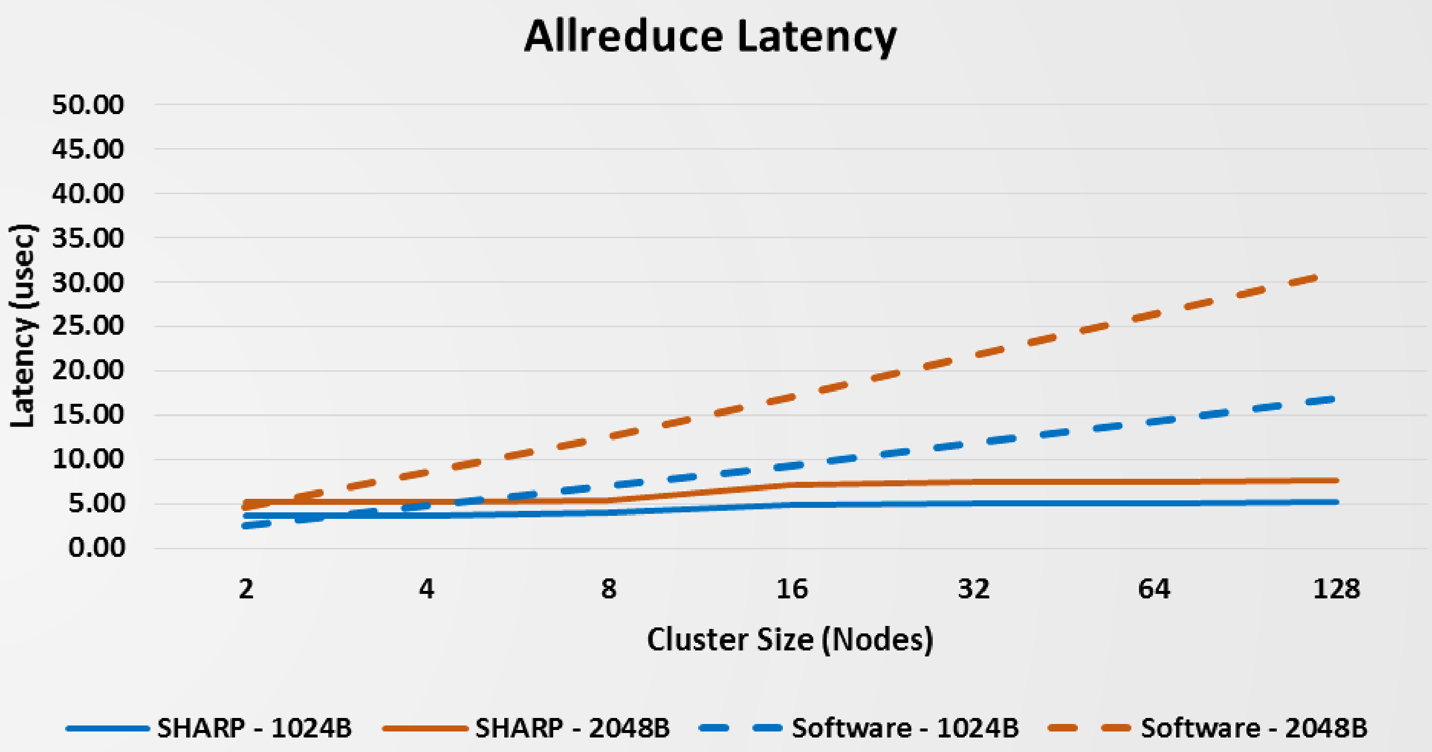
Better Data Security

In-Network Computing to Enable Data-Centric Data Centers



SHARP AllReduce Performance Advantages (128 Nodes)

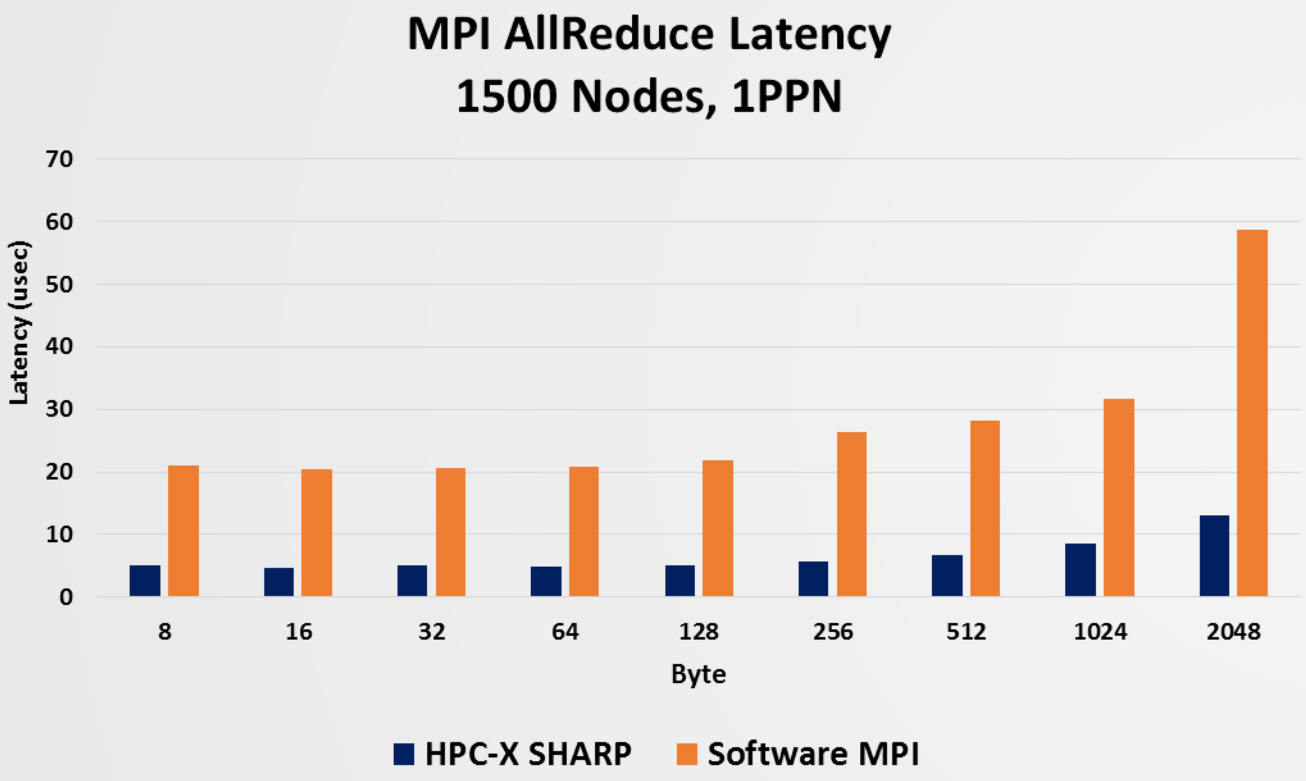


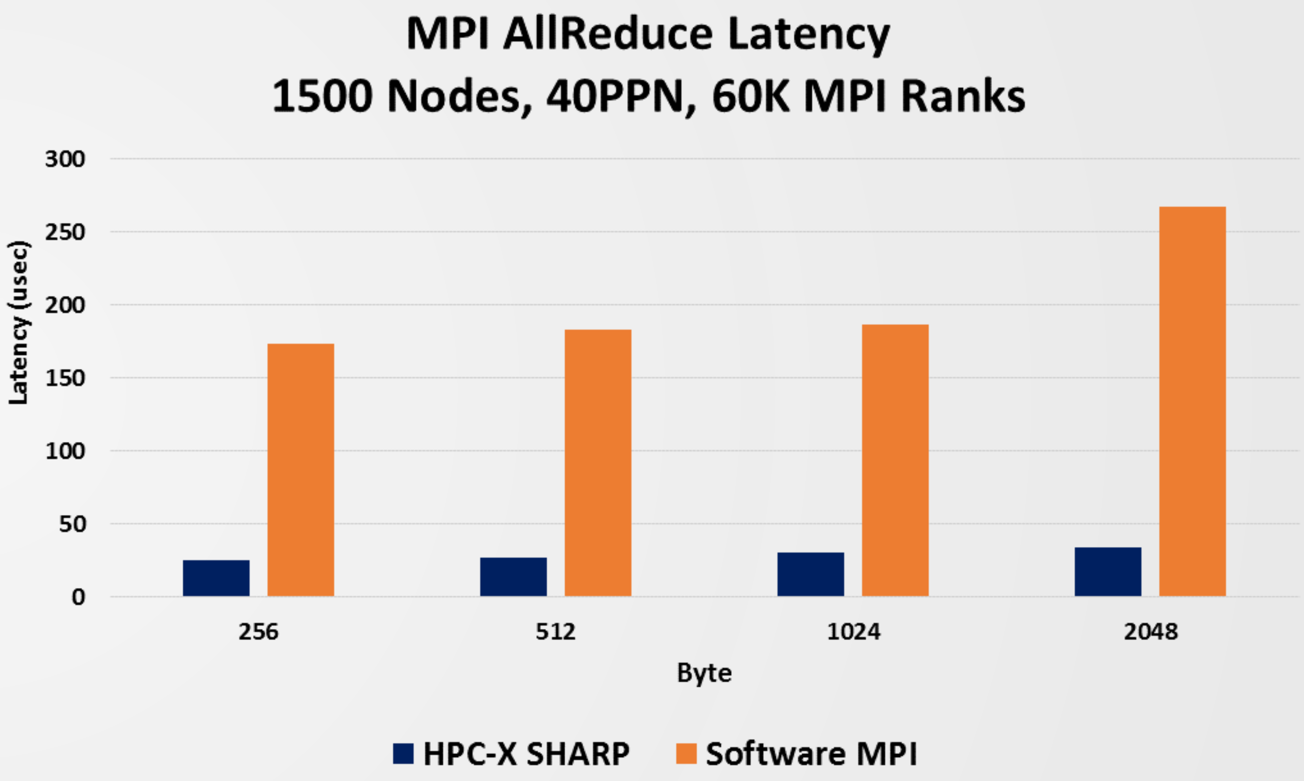


SHARP enables 75% Reduction in Latency
Providing Scalable Flat Latency

SHARP AllReduce Performance Advantages

1500 Nodes, 60K MPI Ranks, Dragonfly+ Topology





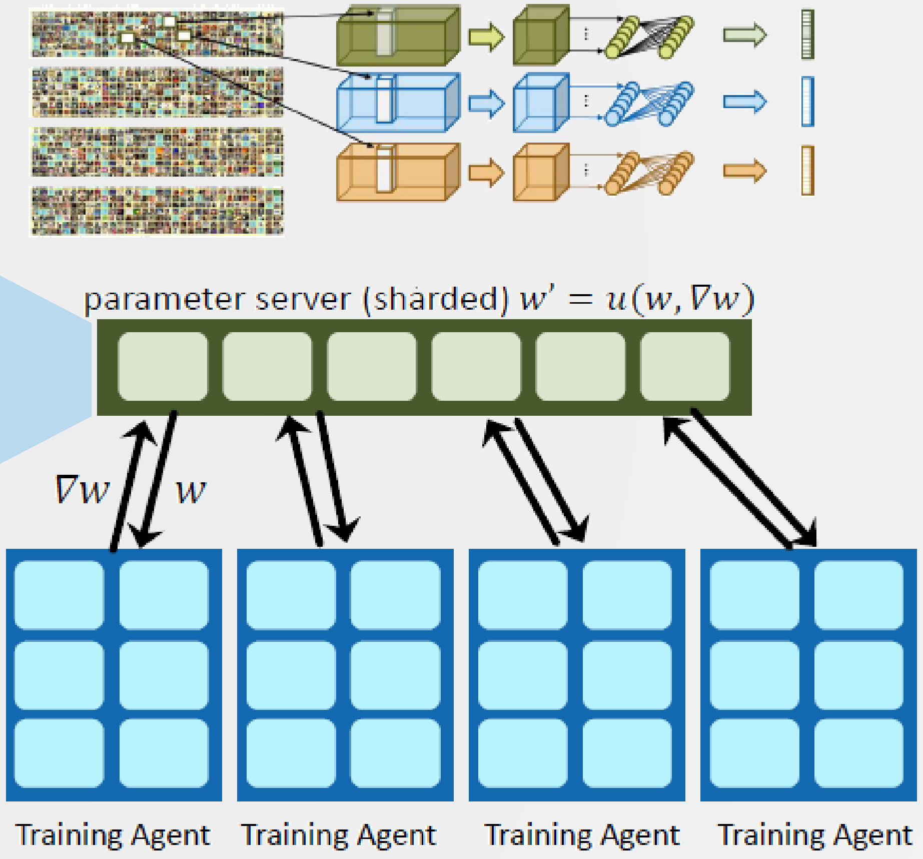
SHARP Enables Highest Performance

SHARP Accelerates AI Performance

The CPU in a parameter server
becomes the bottleneck

SHARP

Scalable Hierarchical
Aggregation and
Reduction Protocol



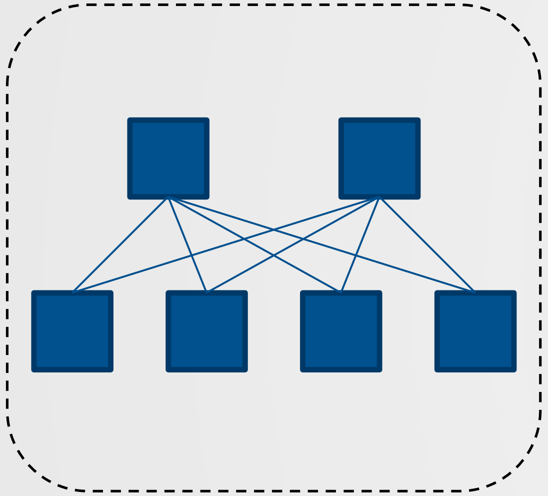
Performs the Gradient Averaging

Replaces all physical parameter servers

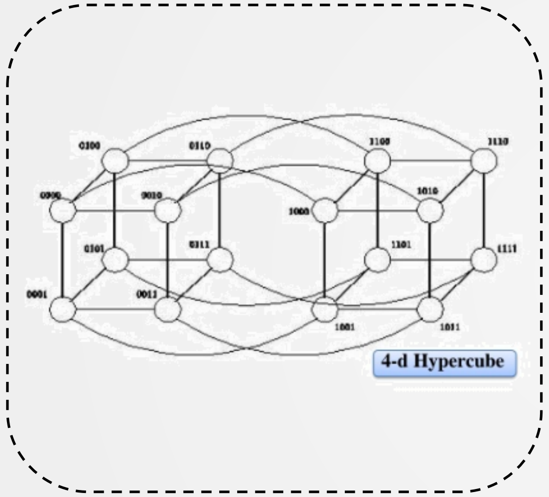
Accelerate AI Performance

Network Topologies

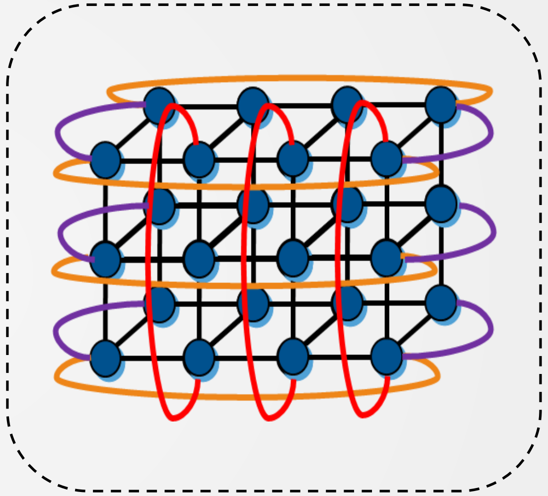
Supporting Variety of Topologies



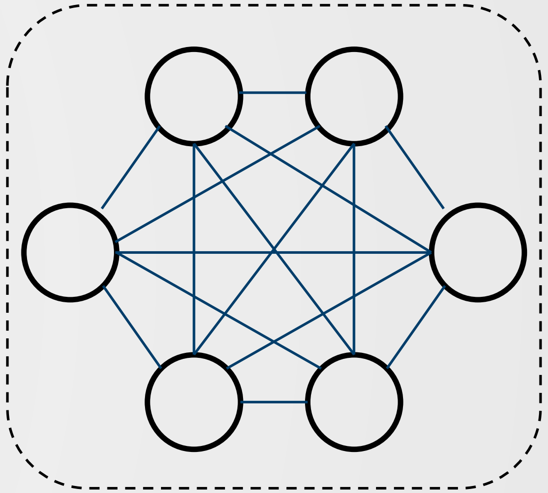
Fat Tree



Hypercube



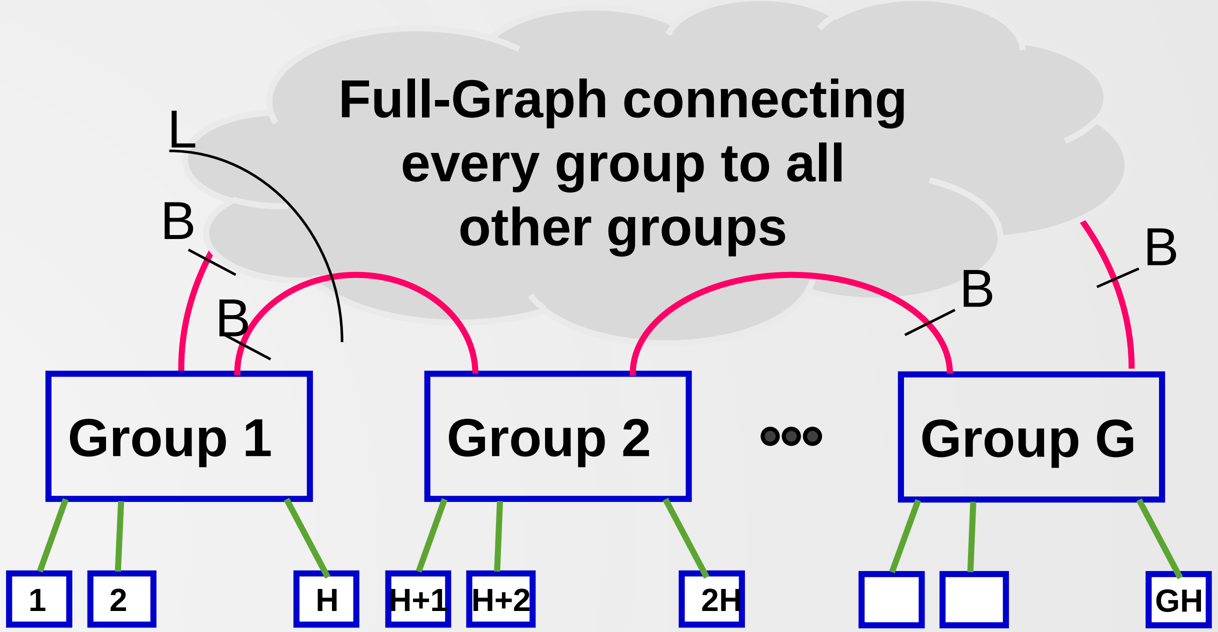
Torus



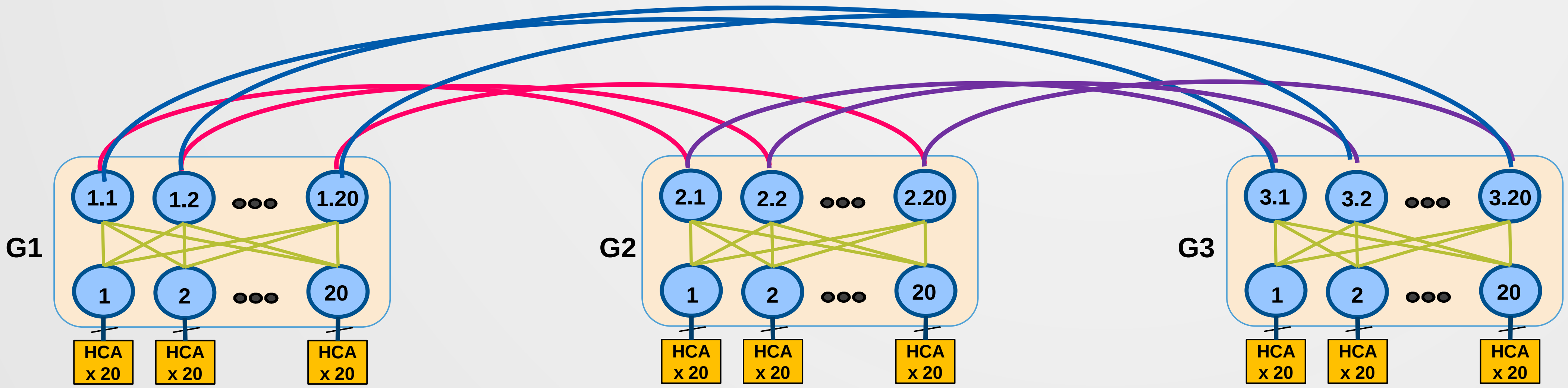
Dragonfly

Dragonfly+ Topology

- Several “groups”, connected using all to all links
- The topology inside each group can be any topology
- Reduce total cost of network (fewer long cables)
- Utilizes Adaptive Routing to for efficient operations
- Simplifies future system expansion



1200-Nodes Dragonfly+ Systems Example



ENABLER OF CO-DESIGN

UCF

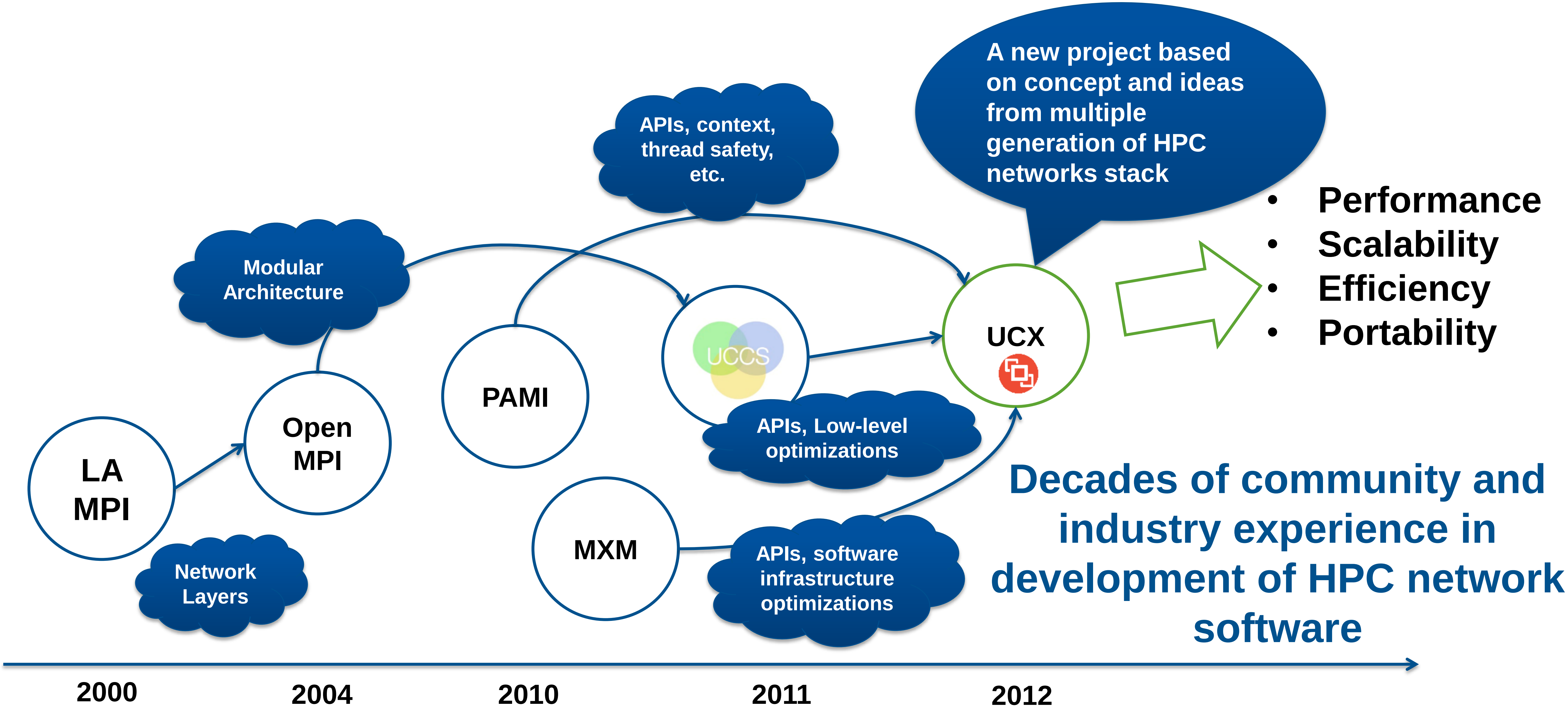
Unified Communication X (UCX)

August 2018

UCF Consortium

- Mission:
  - Collaboration between industry, laboratories, and academia to create production grade communication frameworks and open standards for data centric and high-performance applications
- Projects
  - UCX – Unified Communication X
  - Open RDMA
- Board members
  - **Jeff Kuehn**, UCF Chairman (Los Alamos National Laboratory)
  - **Gilad Shainer**, UCF President (Mellanox Technologies)
  - **Pavel Shamis**, UCF treasurer (ARM)
  - **Brad Benton**, Board Member (AMD)
  - **Duncan Poole**, Board Member (Nvidia)
  - **Pavan Balaji**, Board Member (Argonne National Laboratory)
  - **Sameh Sharkawi**, Board Member (IBM)
  - **Dhabaleswar K. (DK) Panda**, Board Member (Ohio State University)
  - **Steve Poole**, Board Member (Open Source Software Solutions)

UCX - History


A new project based on concept and ideas from multiple generation of HPC networks stack
APIs, context, thread safety, etc.
Modular Architecture
UCCS
UCX
PAMI
Open MPI
LA MPI
APIs, Low-level optimizations
Network Layers
MXM
APIs, software infrastructure optimizations
2000
2004
2010
2011
2012

- Performance
- Scalability
- Efficiency
- Portability

Decades of community and industry experience in development of HPC network software

UCX Framework Mission

- Collaboration between industry, laboratories, government (DoD, DoE), and academia
- Create open-source production grade communication framework for HPC applications
- Enable the highest performance through co-design of software-hardware interfaces

API

Exposes broad semantics that target data centric and HPC programming models and applications

Performance oriented

Optimization for low-software overheads in communication path allows near native-level performance

Production quality

Developed, maintained, tested, and used by industry and researcher community

Community driven

Collaboration between industry, laboratories, and academia

Research

The framework concepts and ideas are driven by research in academia, laboratories, and industry

Cross platform

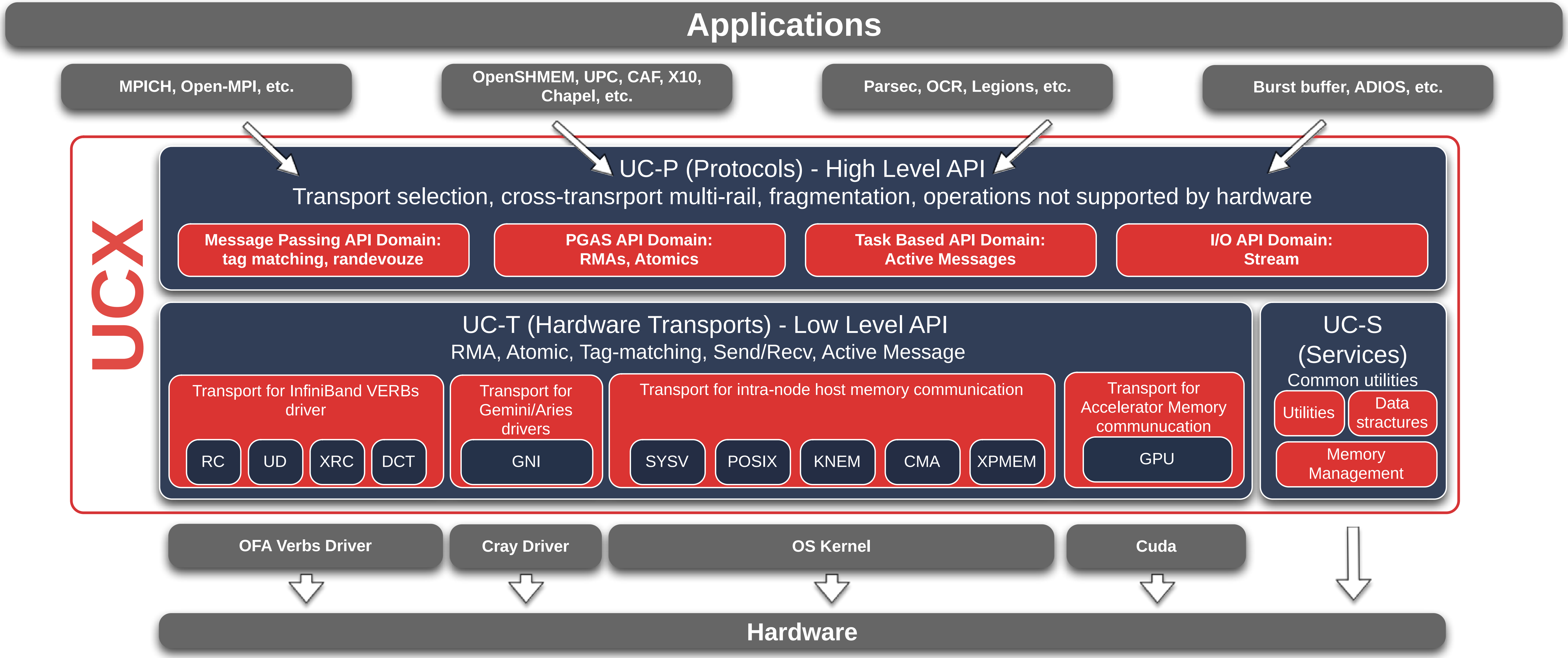
Support for Infiniband, Cray, various shared memory (x86-64, Power, ARMv8), GPUs

Co-design of Exascale Network APIs

UCX Framework

- UCX is a framework for network APIs and stacks
- UCX aims to unify the different network APIs, protocols and implementations into a single framework that is portable, efficient and functional
- UCX doesn’t focus on supporting a single programming model, instead it provides APIs and protocols that can be used to tailor the functionalities of a particular programming model efficiently
- When different programming paradigms and applications use UCX to implement their functionality, it increases their portability. As just implementing a small set of UCX APIs on top of a new hardware ensures that these applications can run seamlessly without having to implement it themselves

UCX High-level Overview



Applications

MPICH, Open-MPI, etc.

OpenSHMEM, UPC, CAF, X10, Chapel, etc.

Parsec, OCR, Legions, etc.

Burst buffer, ADIOS, etc.

UCX

UC-P (Protocols) - High Level API

Transport selection, cross-transrport multi-rail, fragmentation, operations not supported by hardware

Message Passing API Domain: tag matching, randevouze

PGAS API Domain: RMAs, Atomics

Task Based API Domain: Active Messages

I/O API Domain: Stream

UC-T (Hardware Transports) - Low Level API

RMA, Atomic, Tag-matching, Send/Recv, Active Message

Transport for InfiniBand VERBs driver

RC

UD

XRC

DCT

Transport for Gemini/Aries drivers

GNI

Transport for intra-node host memory communication

SYSV

POSIX

KNEM

CMA

XPMEM

Transport for Accelerator Memory communucation

GPU

UC-S (Services)

Common utilities

Utilities

Data structures

Memory Management

OFA Verbs Driver

Cray Driver

OS Kernel

Cuda

Hardware

Thank You