In-Network Computing

Paving the Road to Exascale

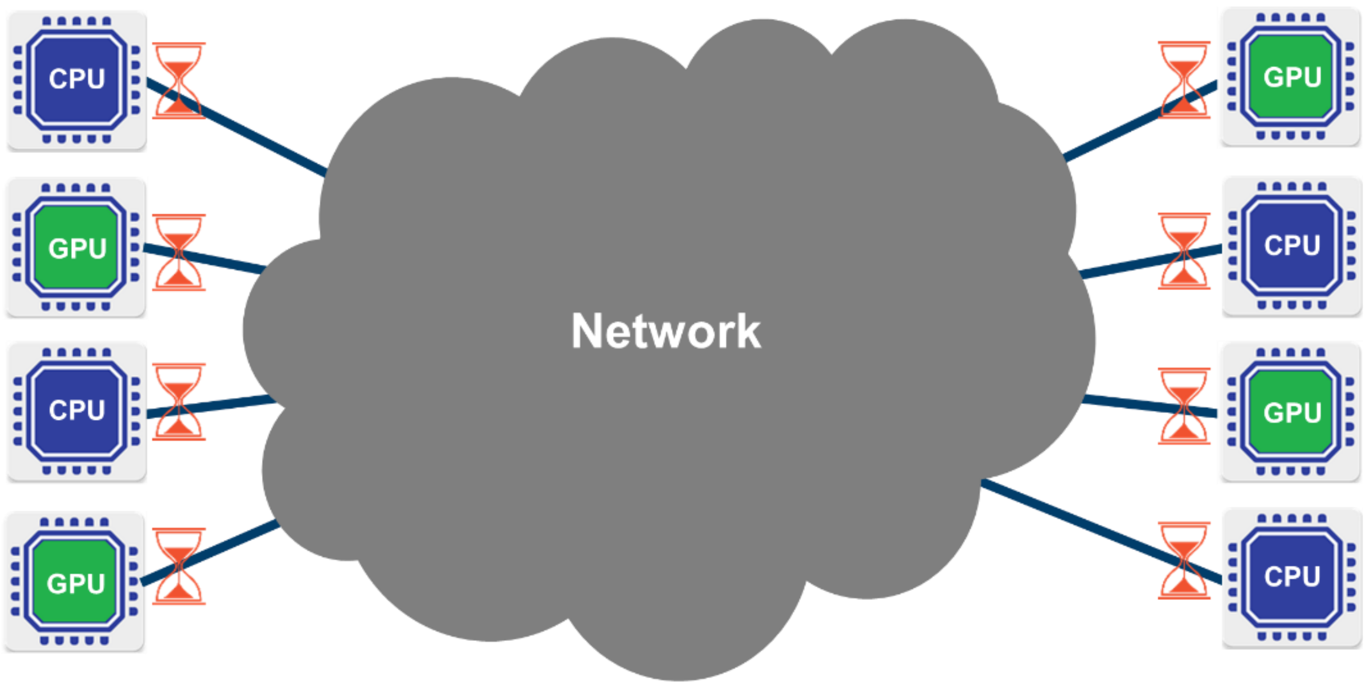
June 2017

Mellanox TECHNOLOGIES

Connect. Accelerate. Outperform.™

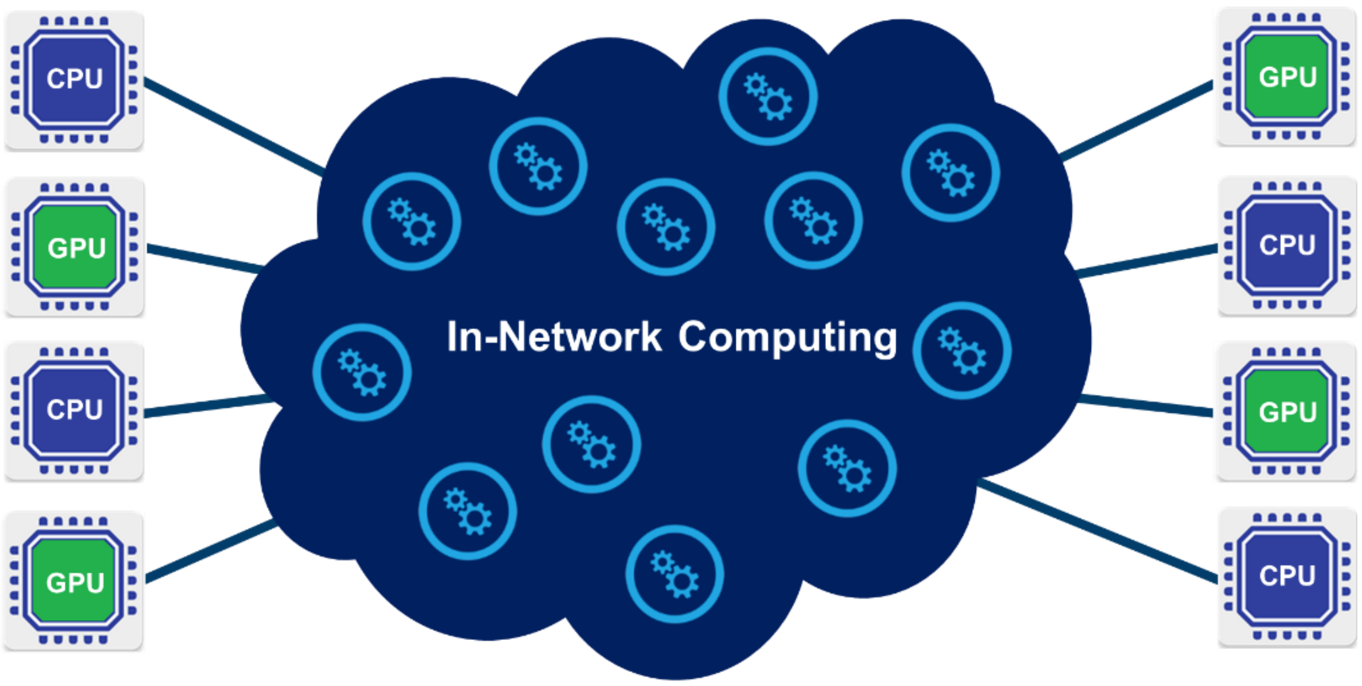
Exponential Data Growth – The Need for Intelligent and Faster Interconnect

CPU-Centric (Onload)



Must Wait for the Data
Creates Performance Bottlenecks

Data-Centric (Offload)

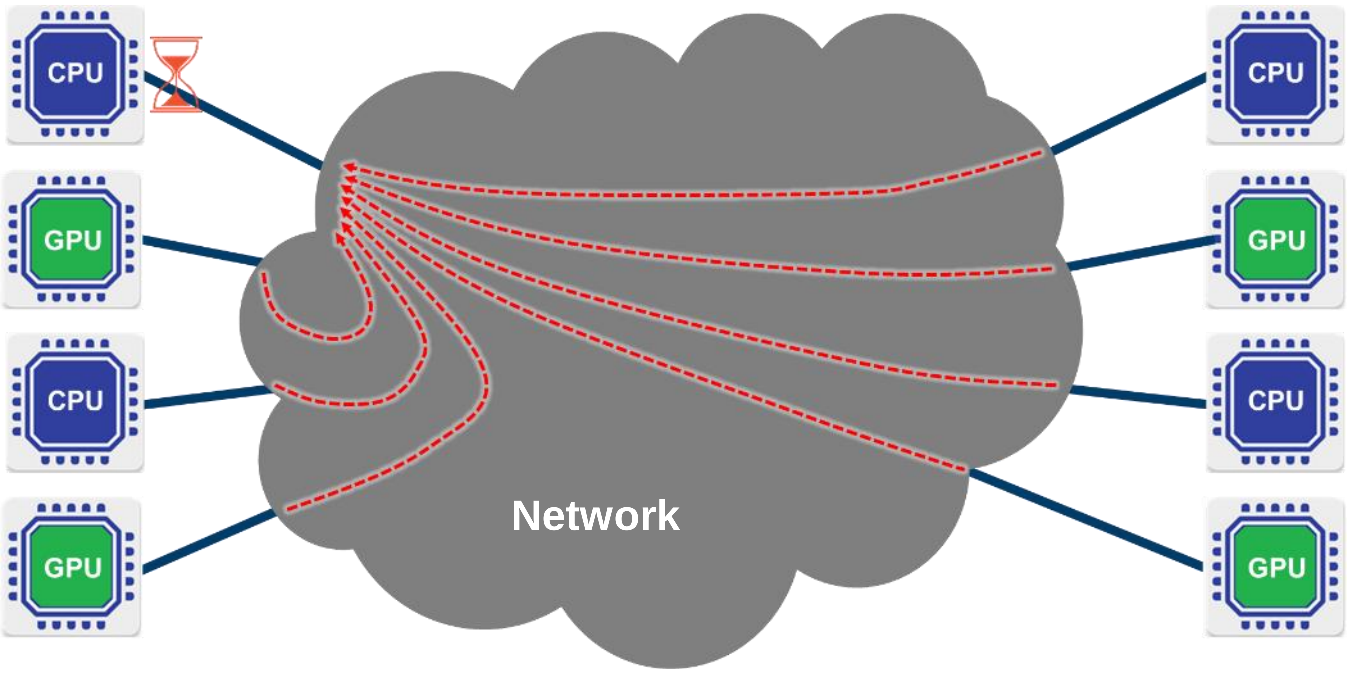


Analyze Data as it Moves!

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

Data Centric Architecture to Overcome Latency Bottlenecks

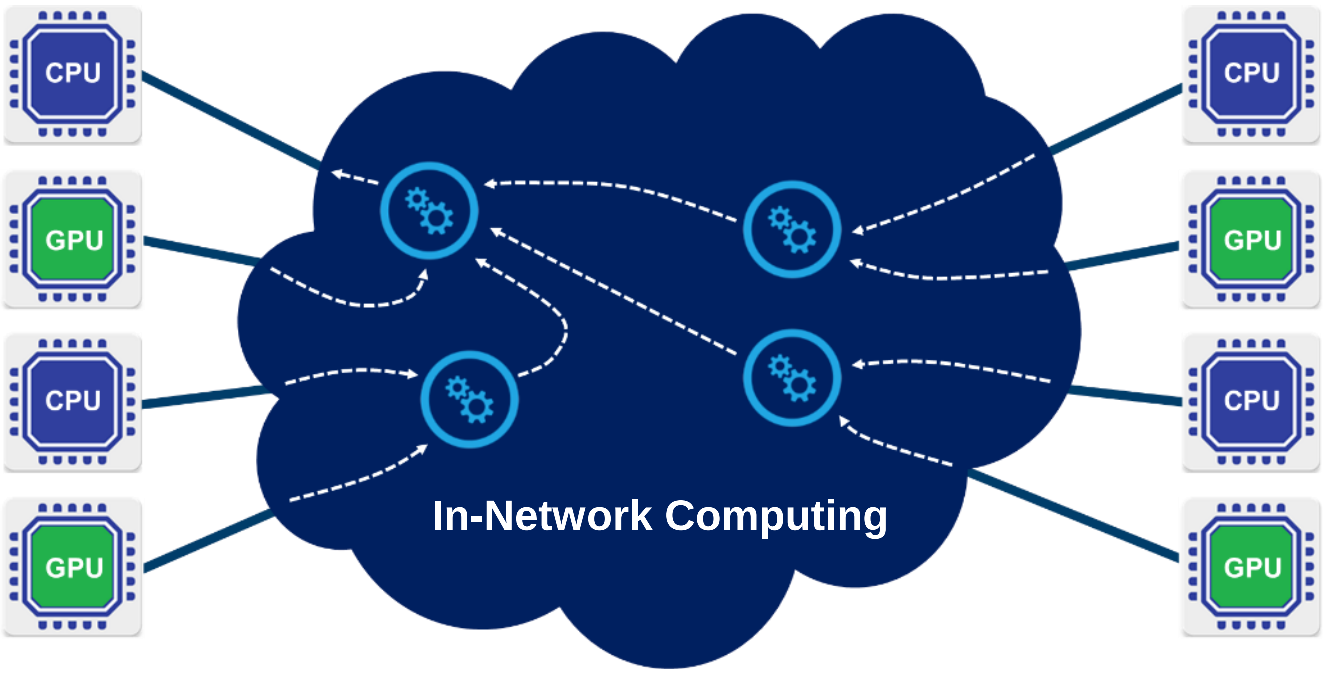
CPU-Centric (Onload)



HPC / Machine Learning

Communications Latencies of 30-40us

Data-Centric (Offload)

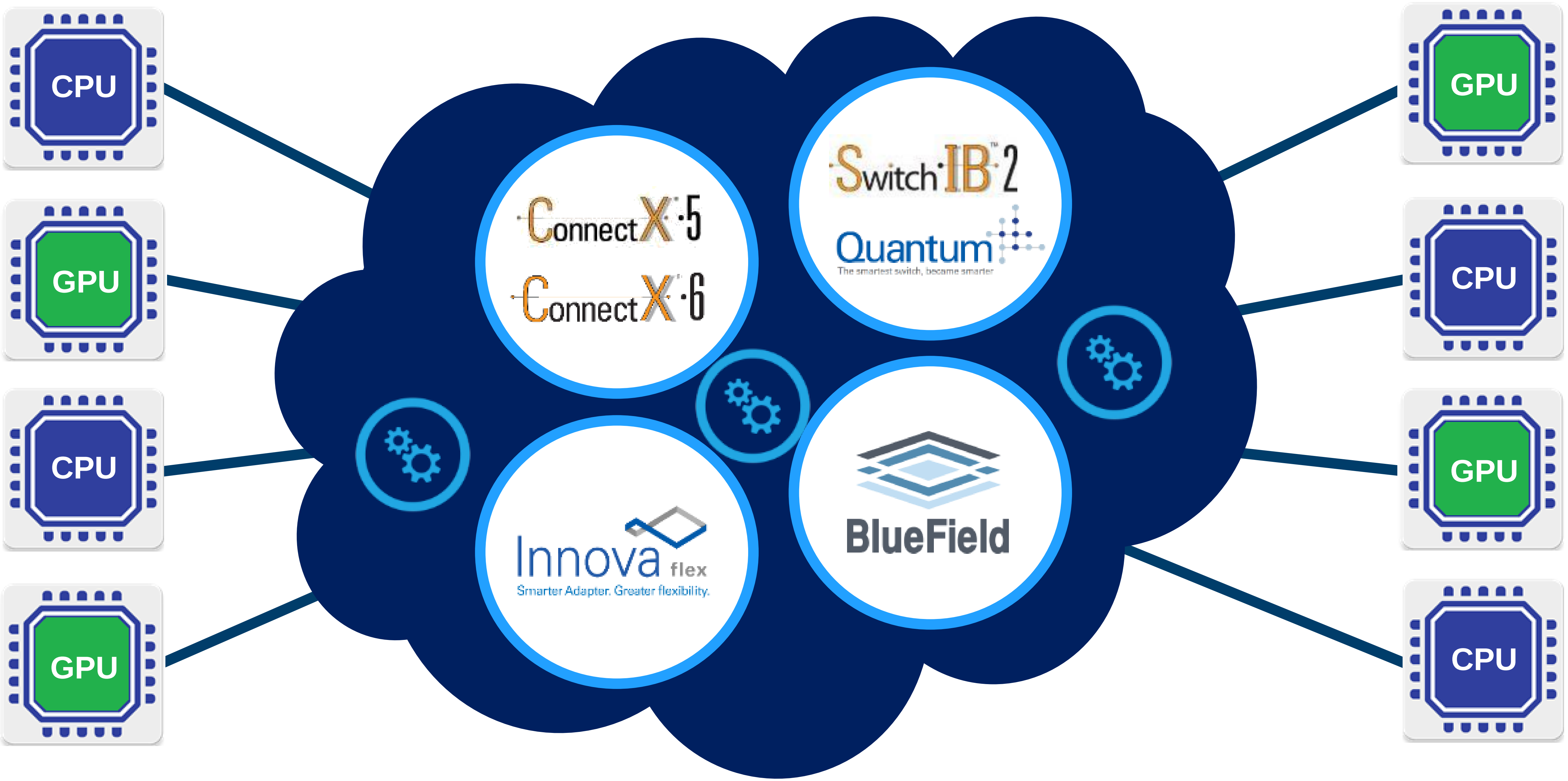


HPC / Machine Learning

Communications Latencies of 3-4us

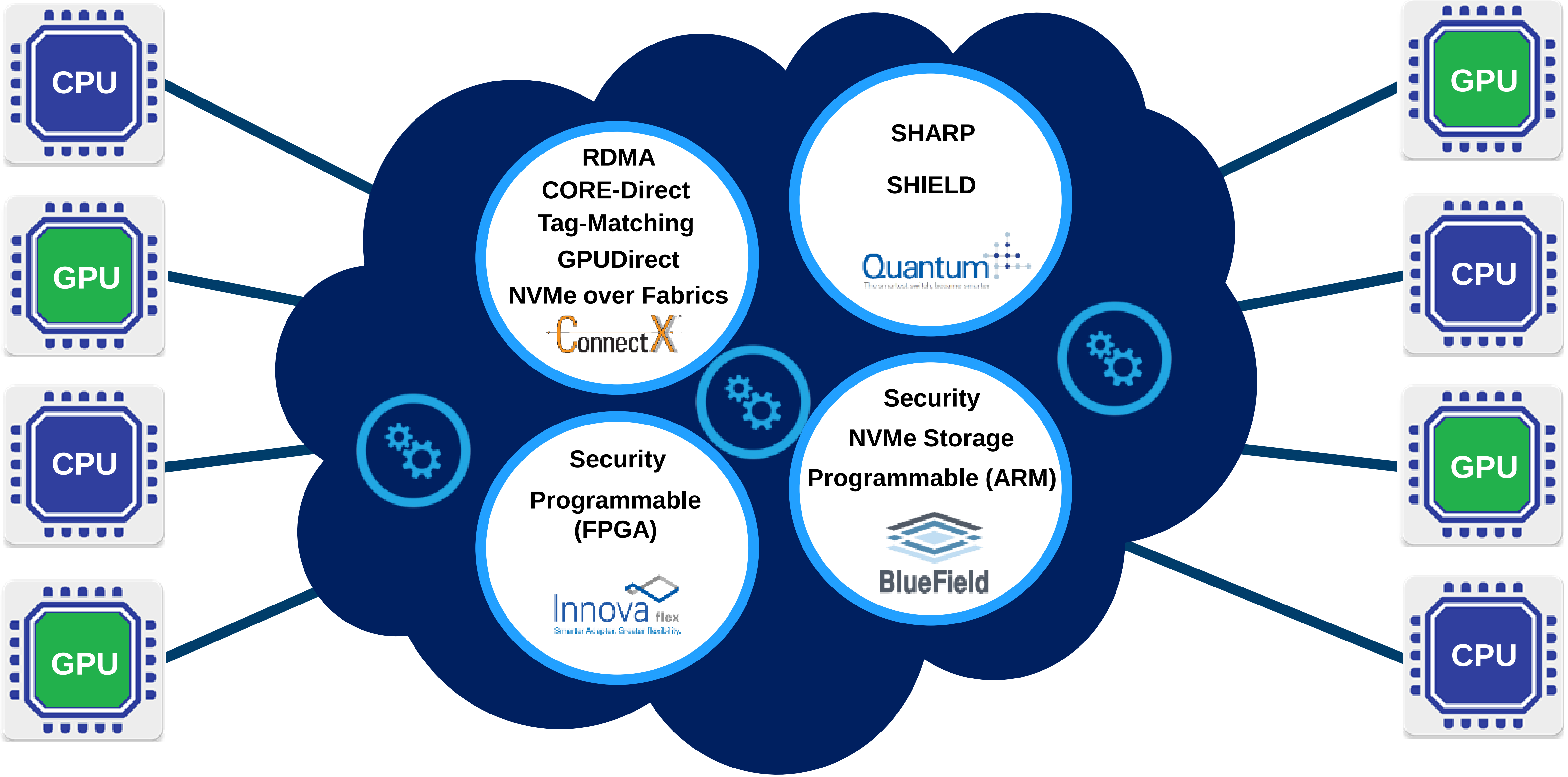
Intelligent Interconnect Paves the Road to Exascale Performance

In-Network Computing to Enable Data-Centric Data Center



In-Network Computing Key for Highest Return on Investment

In-Network Computing to Enable Data-Centric Data Centers



In-Network Computing Key for Highest Return on Investment

Scalable Hierarchical Aggregation and Reduction Protocol (SHARP)

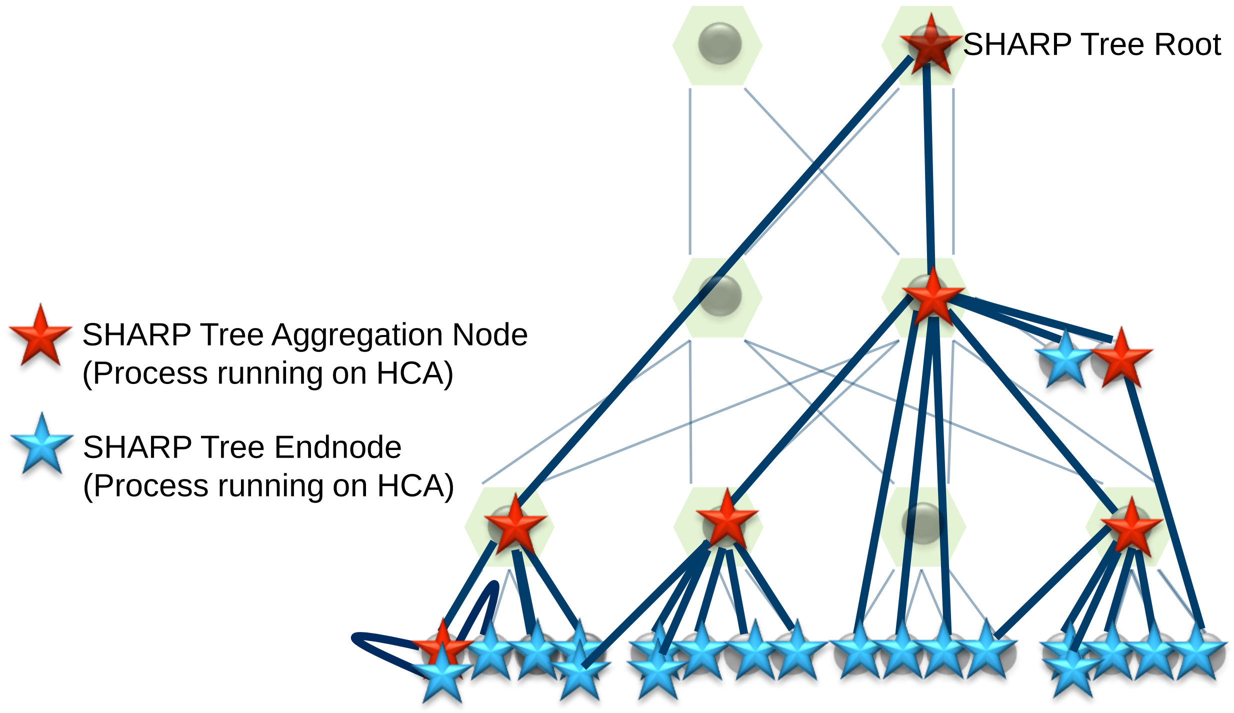
- Reliable Scalable General Purpose Primitive
  - In-network Tree based aggregation mechanism
  - Large number of groups
  - Multiple simultaneous outstanding operations

- 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, 16/32/64/128 bits

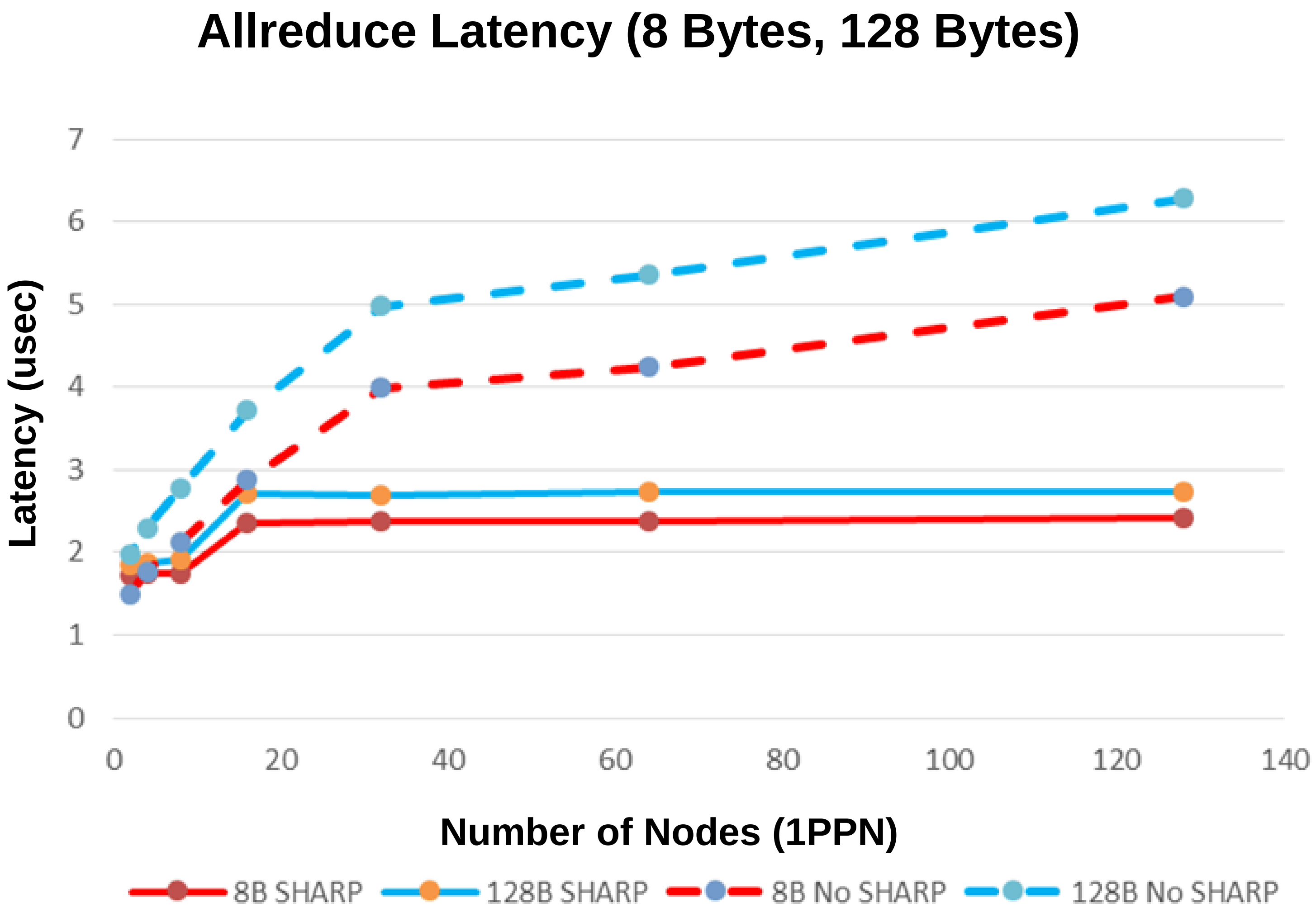
SHARP

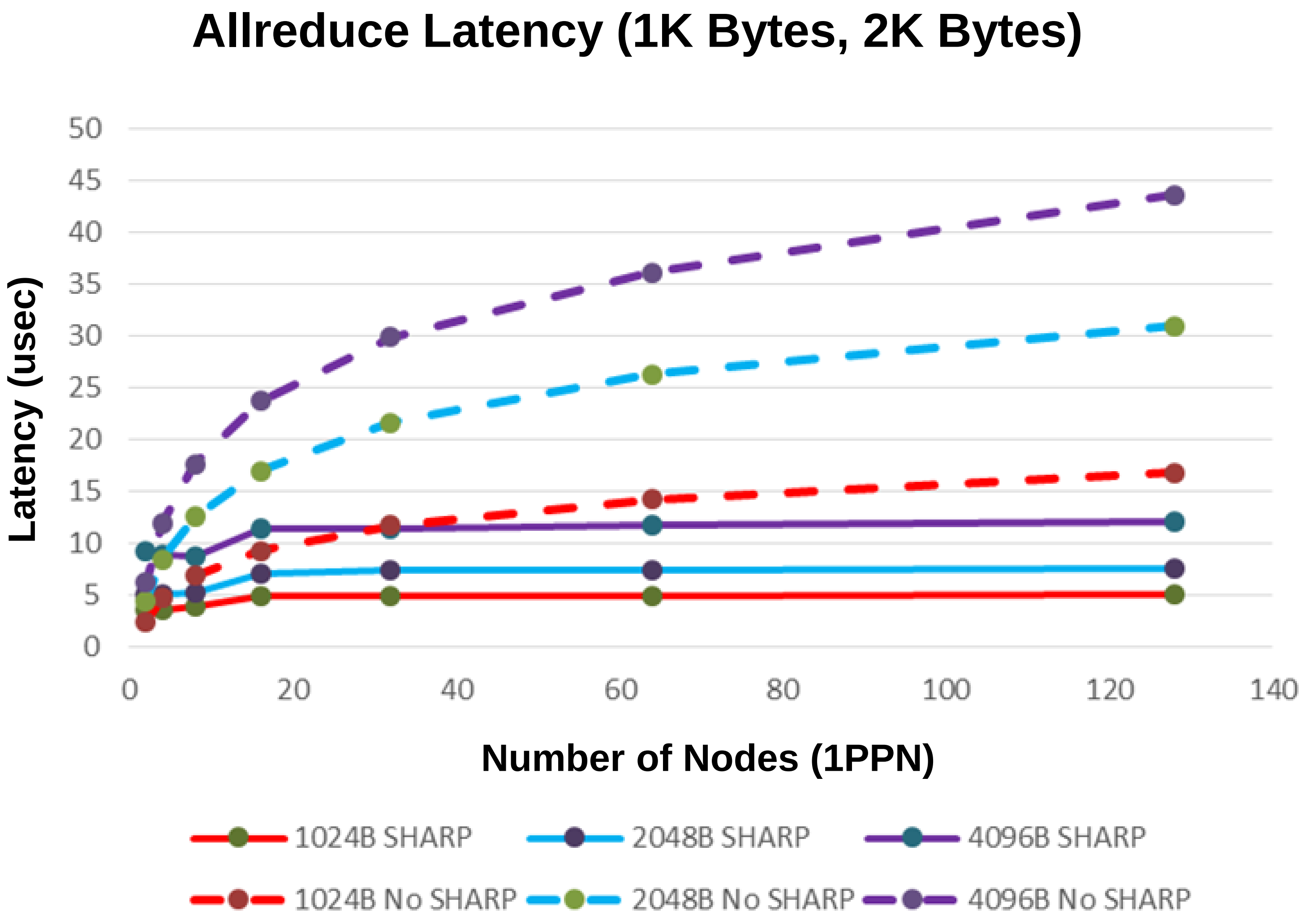
Scalable Hierarchical
Aggregation and
Reduction Protocol



SHArP Tree

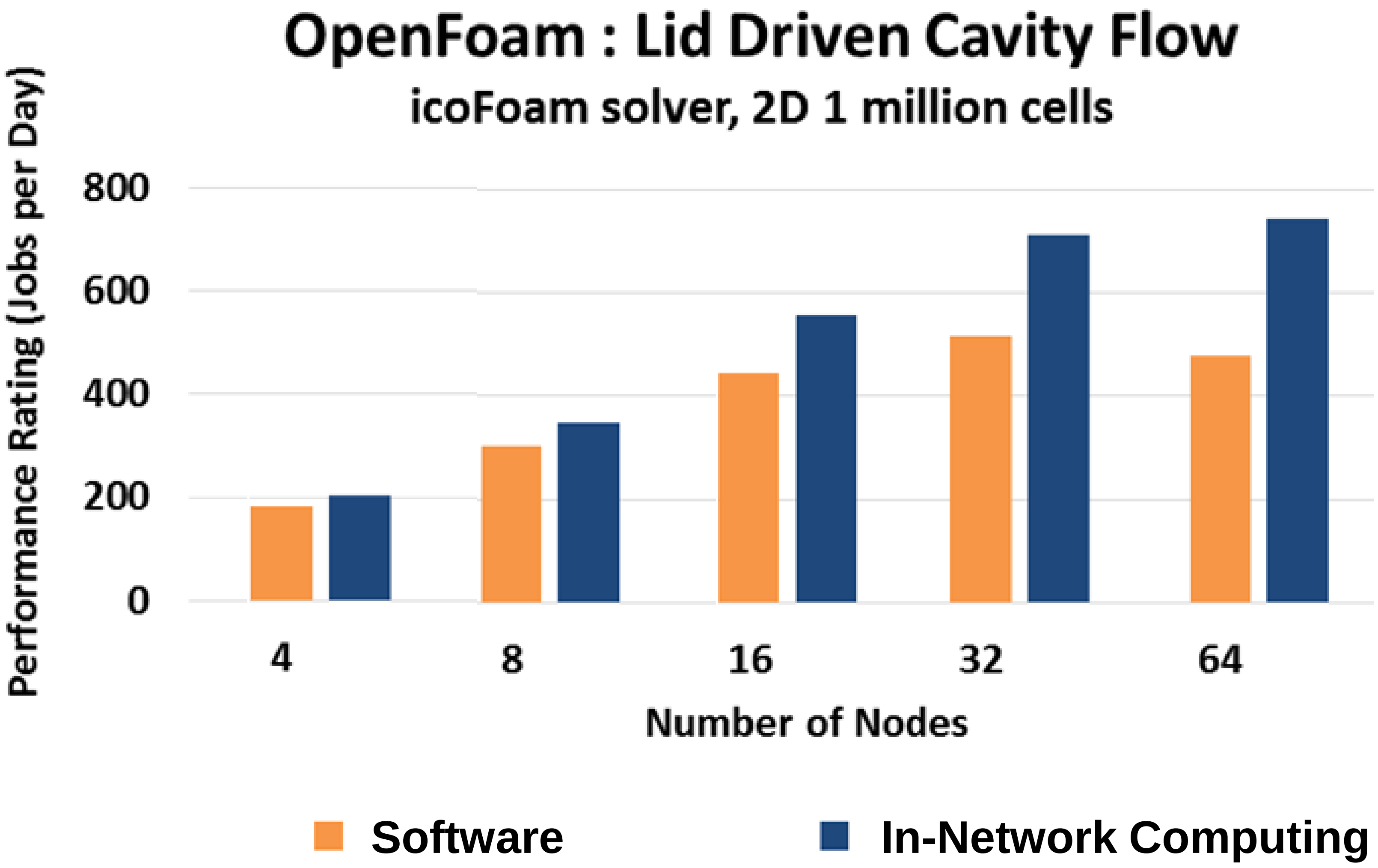
Allreduce Performance





OpenFOAM Application Performance

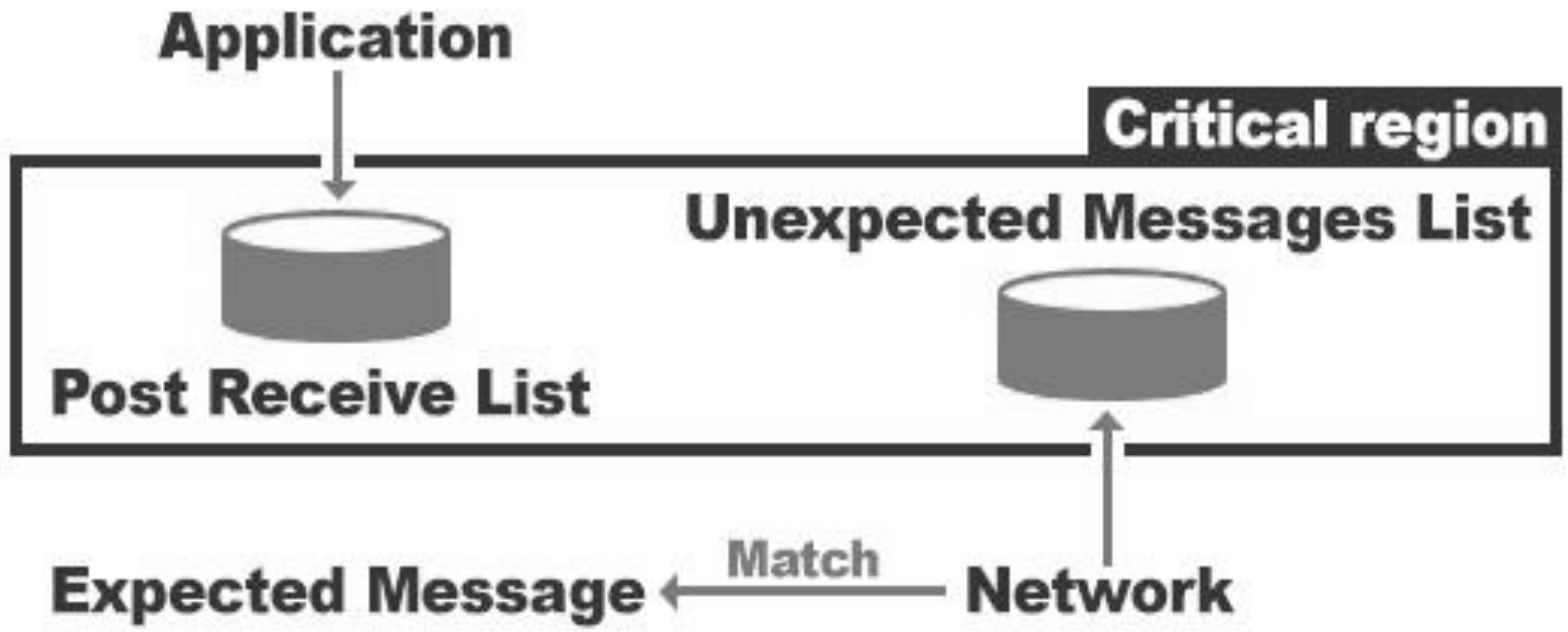
OpenFOAM is a popular computational fluid dynamics application



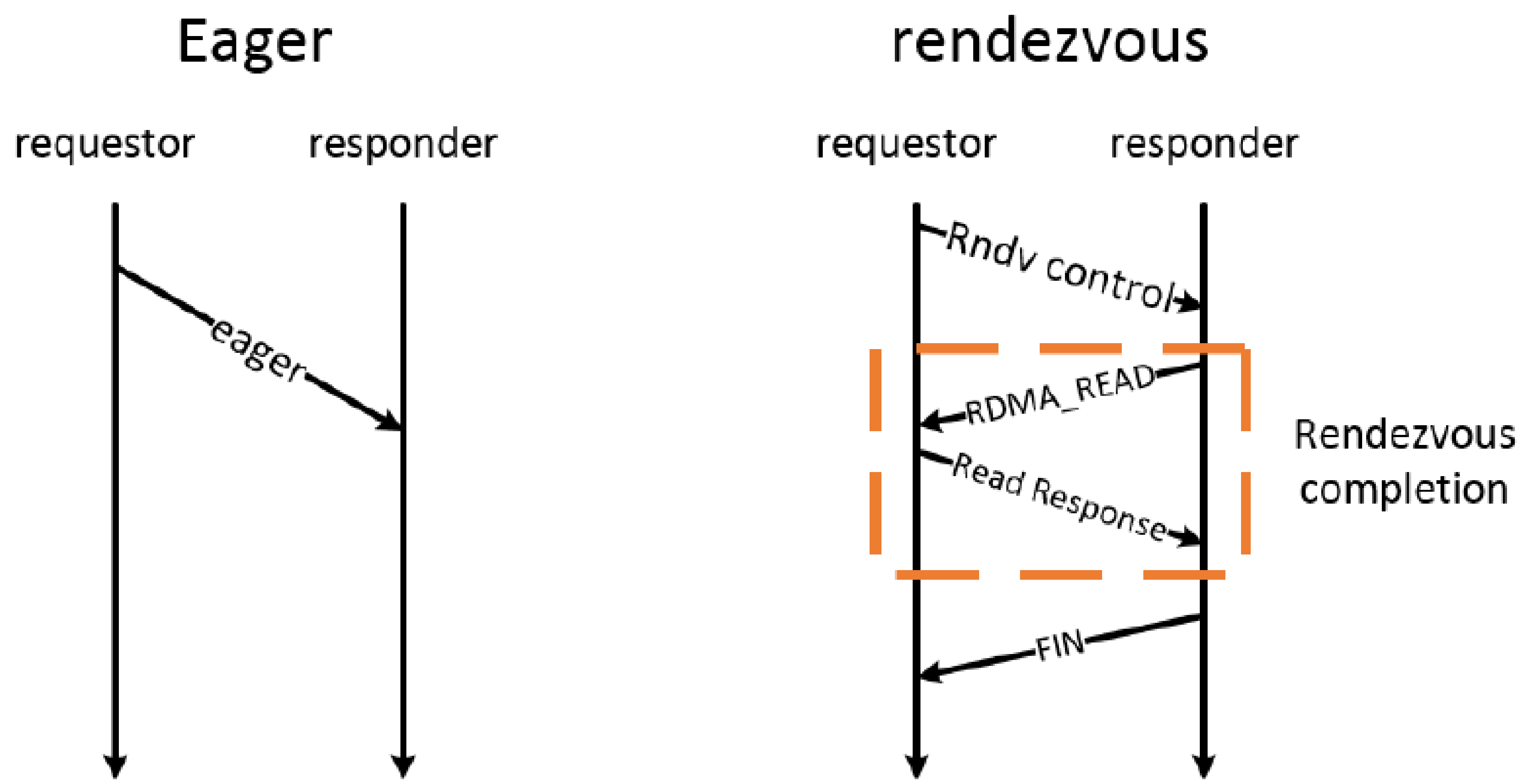
OpenFoam : Lid Driven Cavity Flow

icoFoam solver, 2D 1 million cells

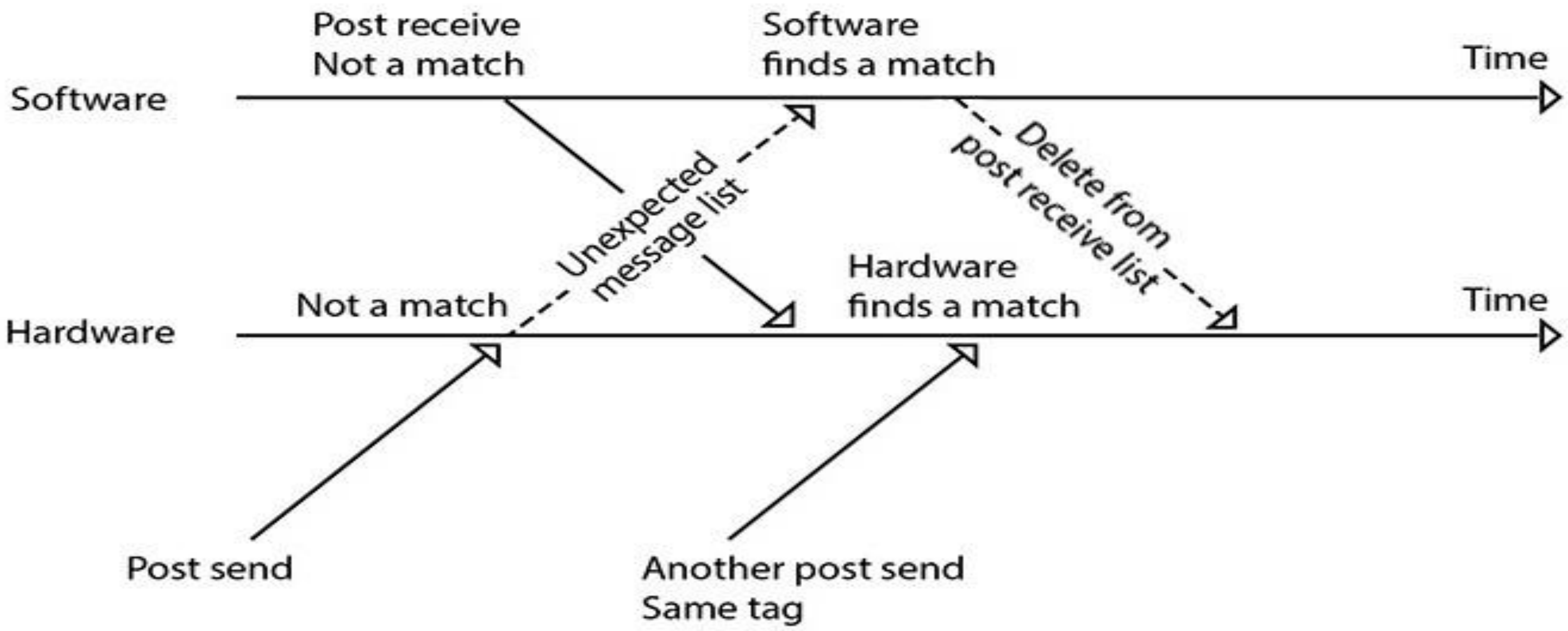
Tag Matching Logistics



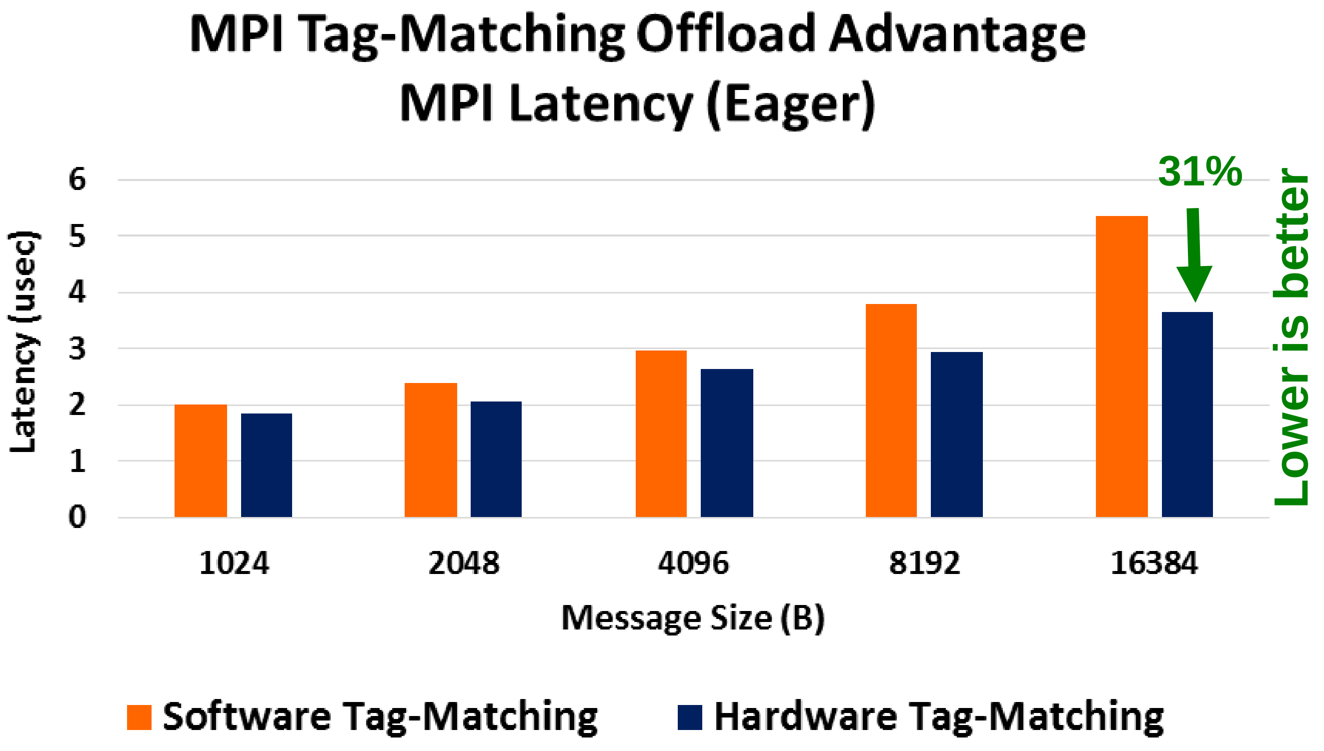
Tag Matching – Common Implementation Protocols

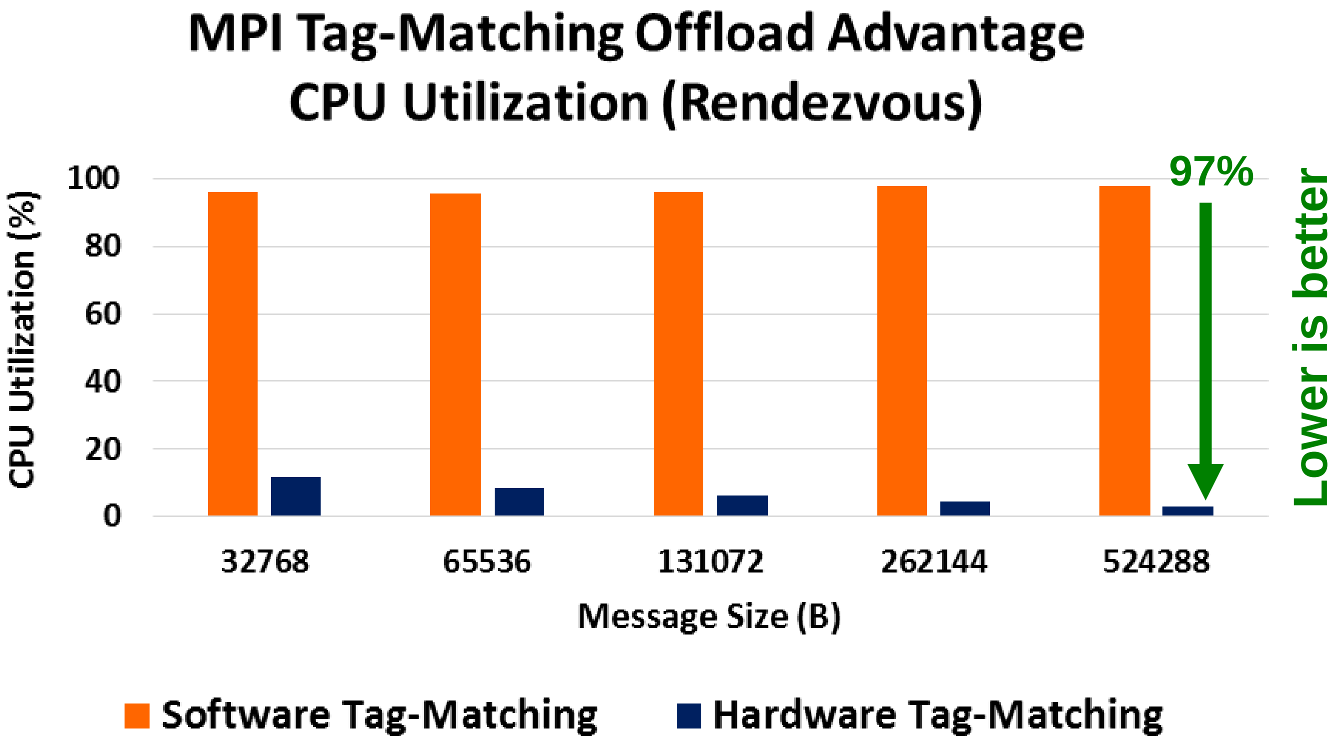


Tag Matching – Hardware Implementation Overview



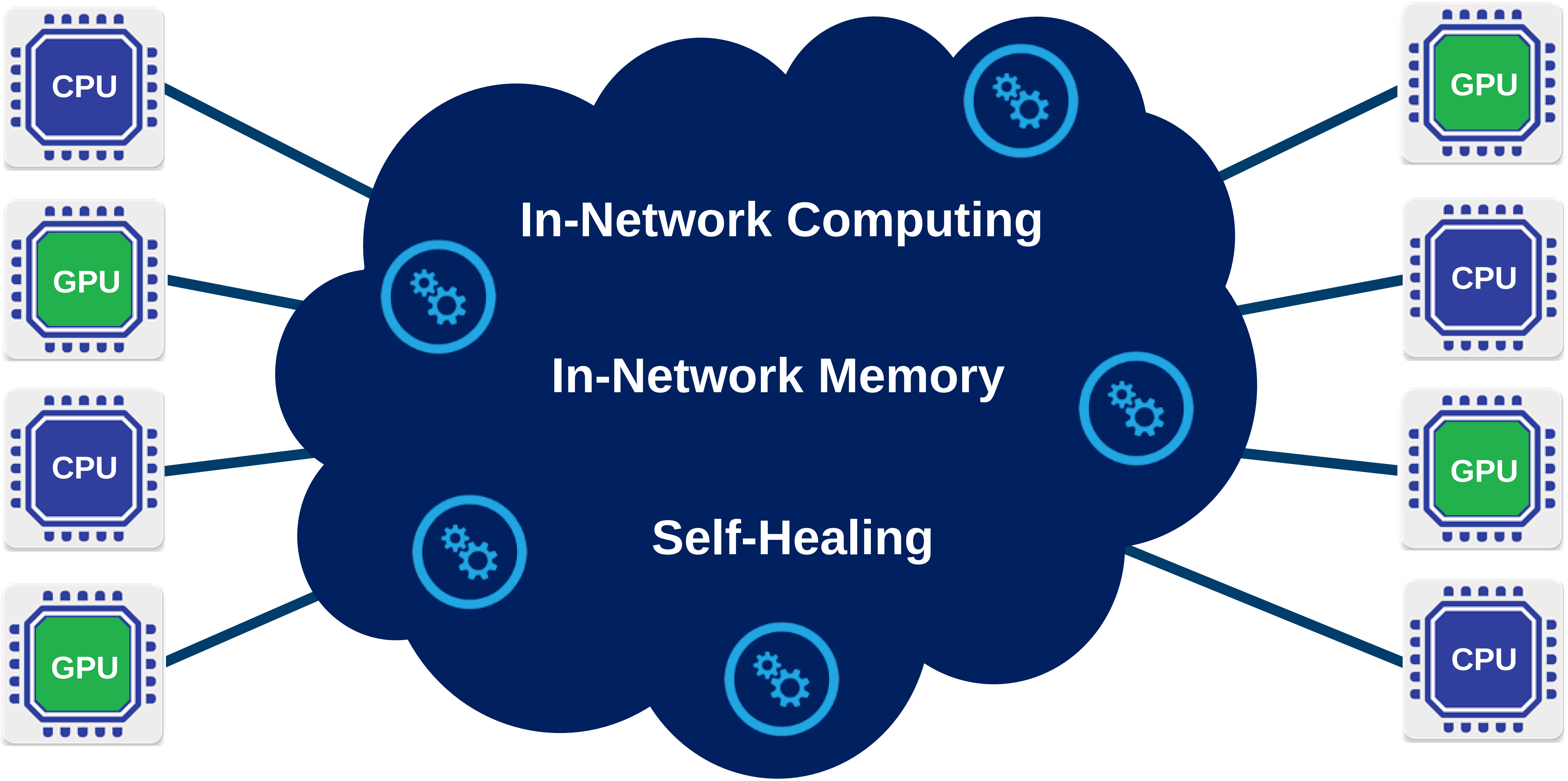
MPI Tag-Matching Offload Advantages





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

In-Network Computing to Enable Data-Centric Data Center



In-Network Computing

In-Network Memory

Self-Healing

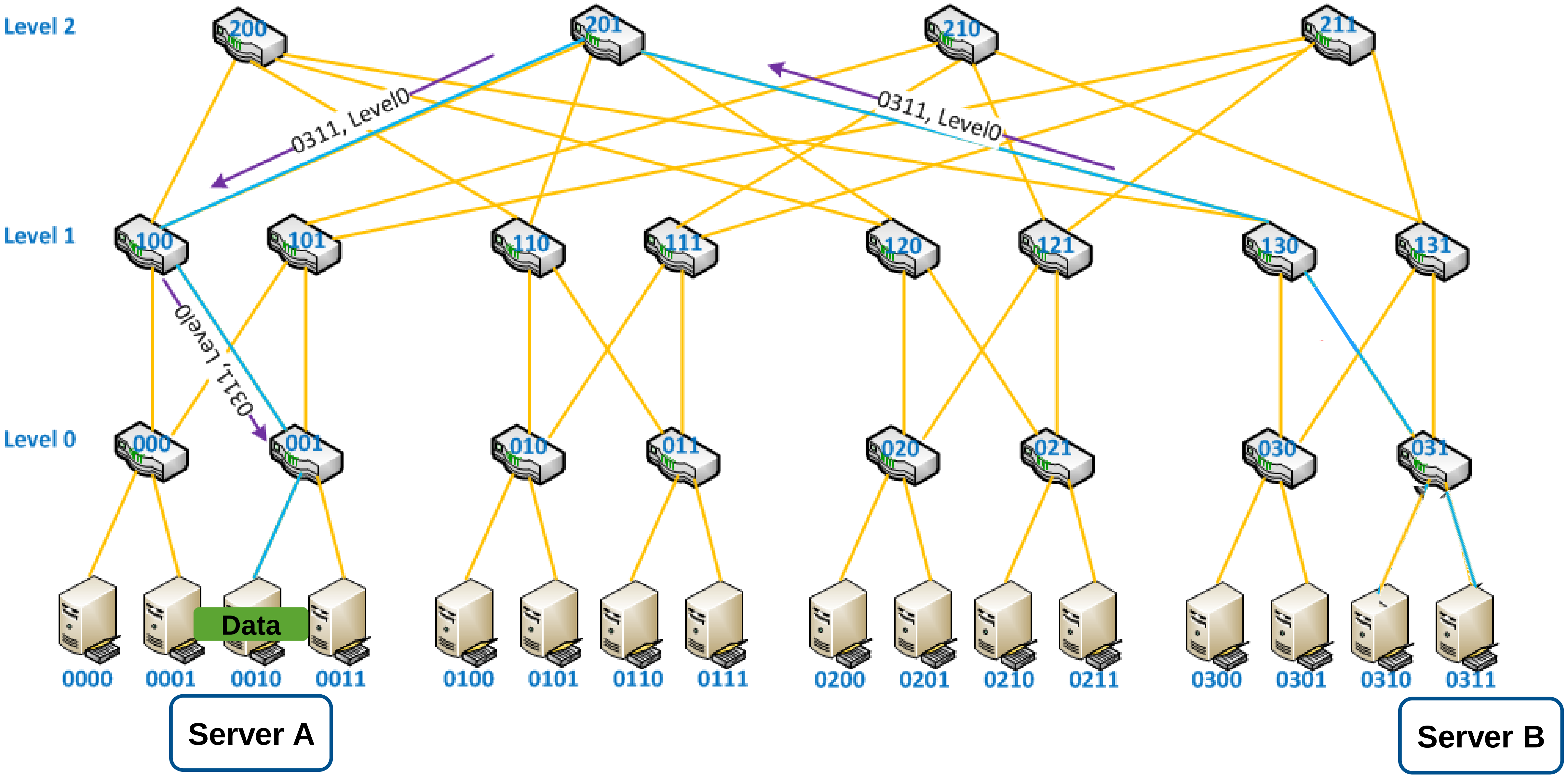
In-Network Computing Key for Highest Return on Investment

The SHIELD Self Healing Interconnect Technology

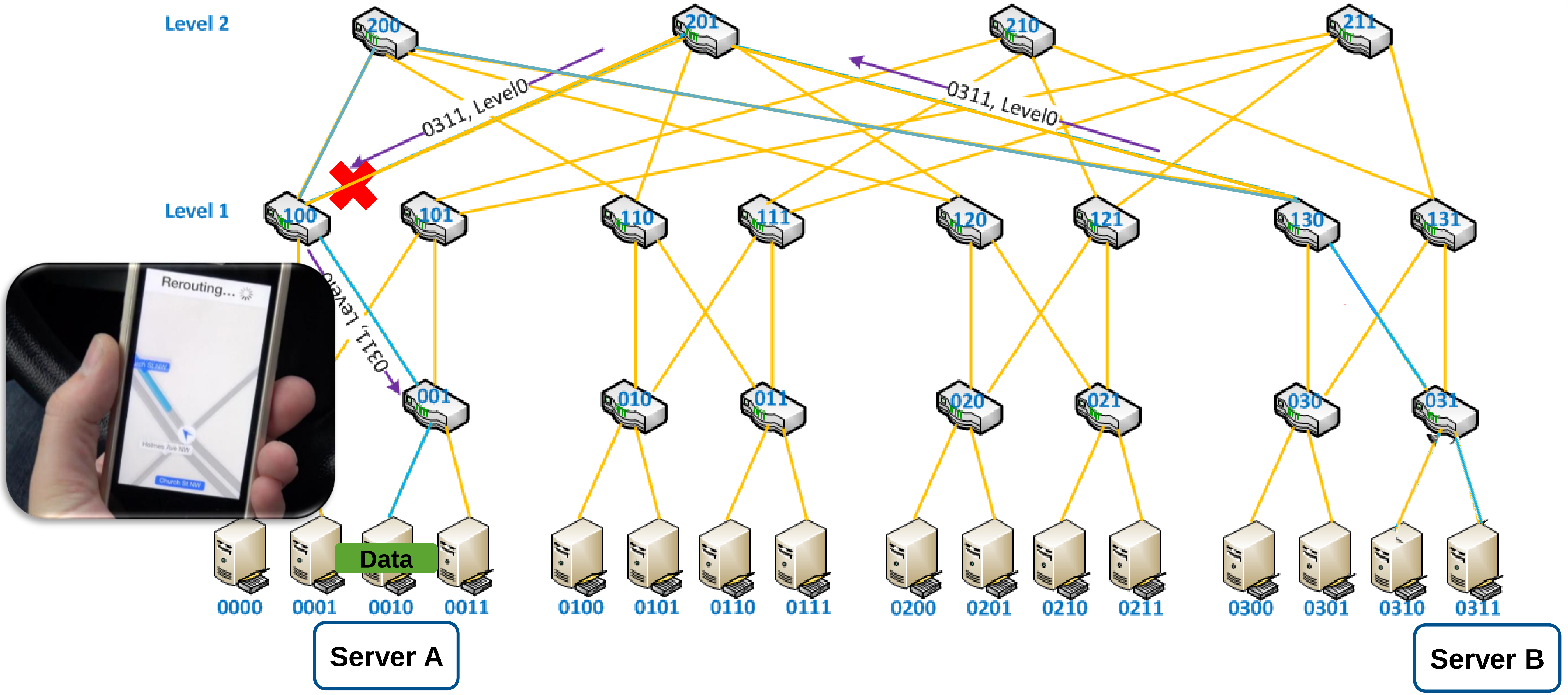
- Software-based solutions for network failures create long delays: 5-30 seconds for 1K to 10K node clusters
- During software-based recovery time, data can be lost, applications can fail
- Adaptive Routing creates further issues (failing links may act as “black holes”)
- Mellanox SHIELD technology is an innovative hardware-based solution
- SHIELD technology enables the generation of Self-Healing Interconnect
- The ability to overcome network failures by the network intelligent devices
- Accelerates network recovery time by 5000X
- Highly scalable and available for EDR and HDR solutions and beyond

Self-Healing Network Enables Unbreakable Data Centers

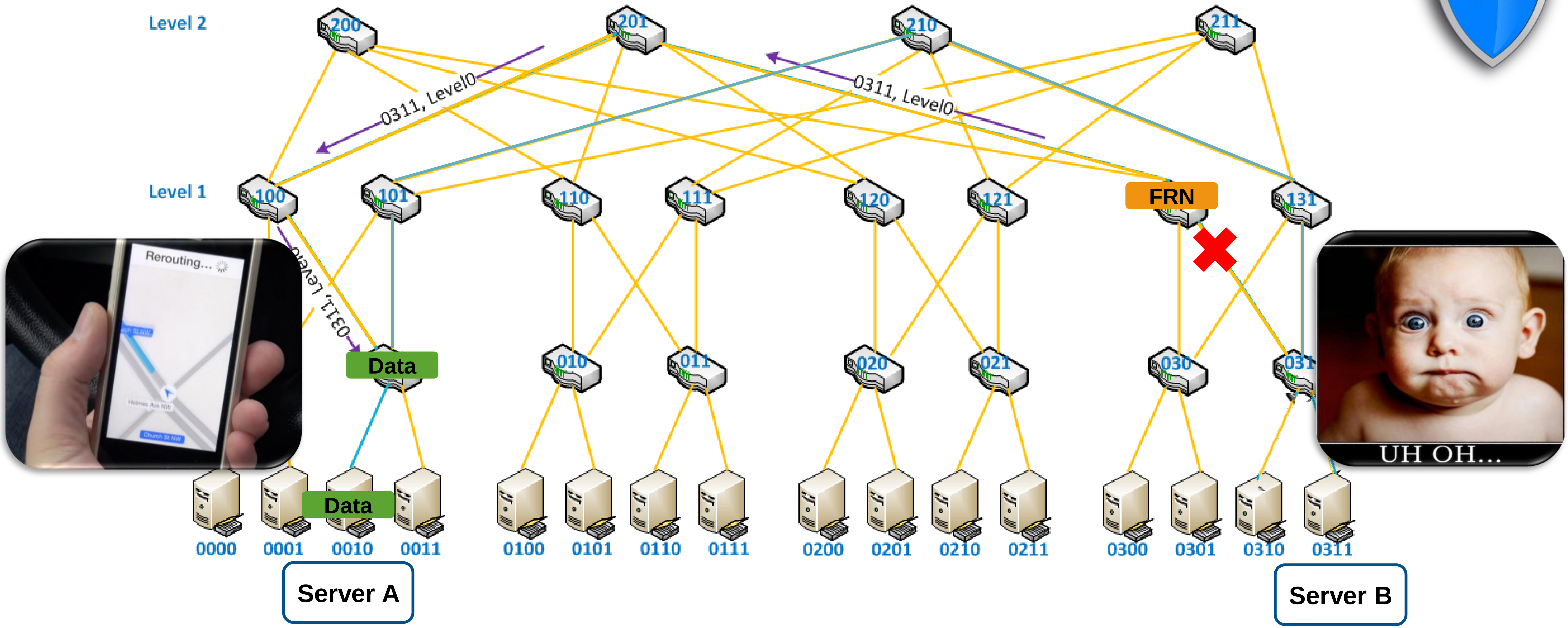
Consider a Flow From A to B



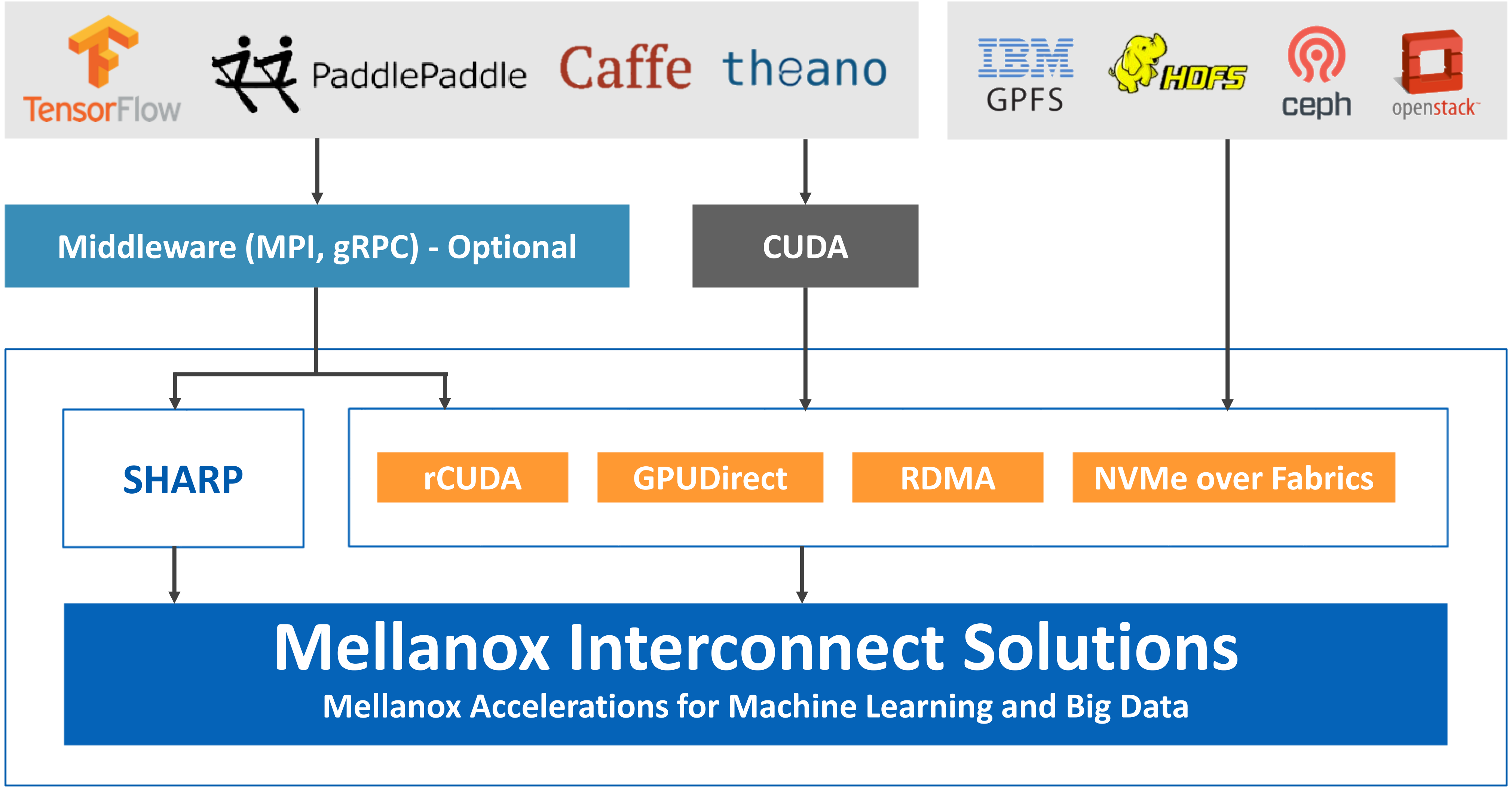
The Simple Case: Local Fix



The Remote Case: Using FRN's (Fault Recovery Notifications)



In-Network Computing Enables Deep Learning Frameworks



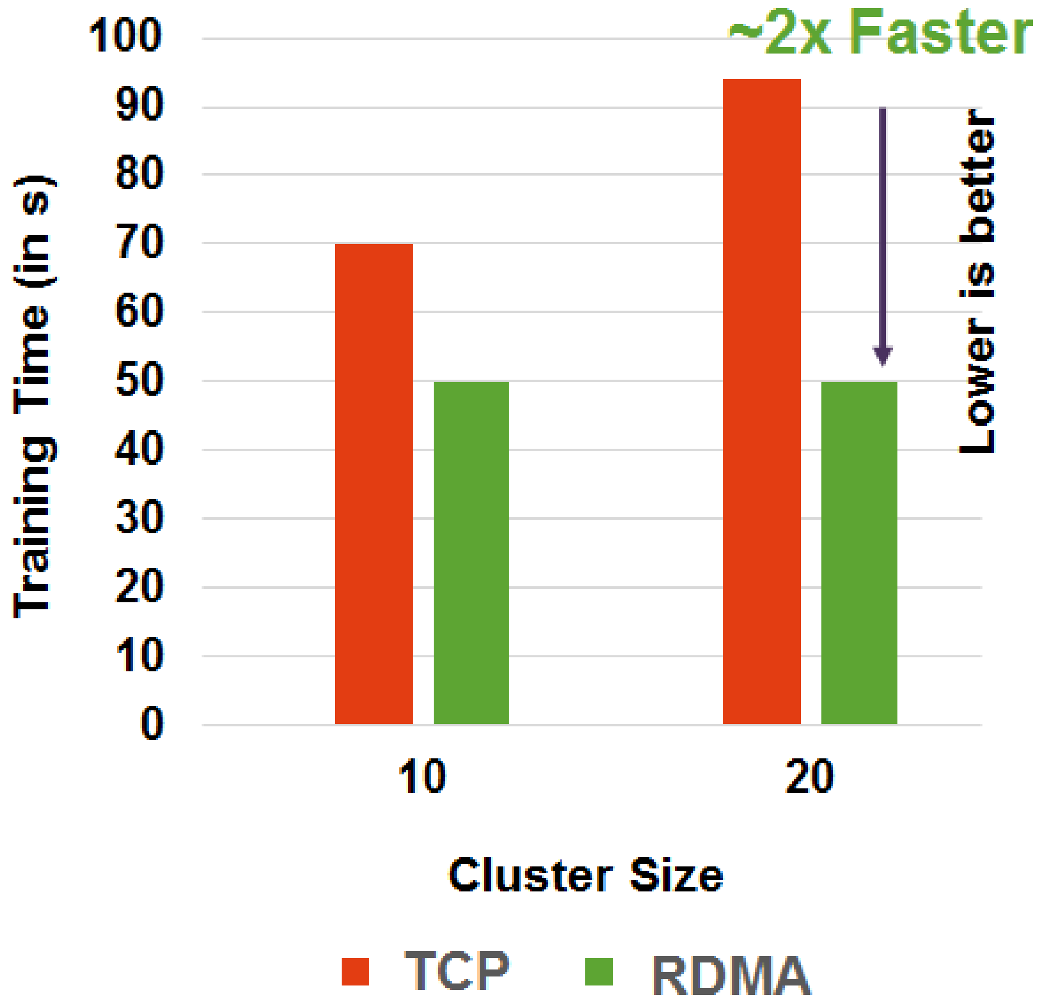
TensorFlow
PaddlePaddle
Caffe
theano
IBM GPFS
HDFS
ceph
openstack
Middleware (MPI, gRPC) - Optional
CUDA
SHARP
rCUDA
GPUDirect
RDMA
NVMe over Fabrics
Mellanox Interconnect Solutions
Mellanox Accelerations for Machine Learning and Big Data

2X Acceleration for Baidu

- Machine Learning Software from Baidu
  - Usage: word prediction, translation, image processing

- RDMA (GPUDirect) speeds training
  - Lowers latency, increases throughput
  - More cores for training
  - Even better results with optimized RDMA

PaddlePaddle



~2X Acceleration for Paddle Training with RDMA

The Ever Growing Demand for Higher Performance

Performance Development

Terascale

Petascale

1st

“Roadrunner”

Exascale

OAK RIDGE National Laboratory

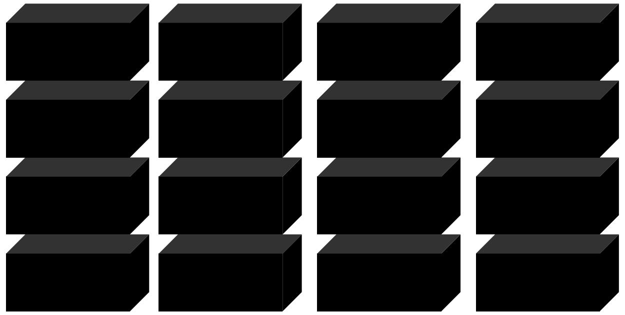
“Summit” System

Lawrence Livermore National Laboratory

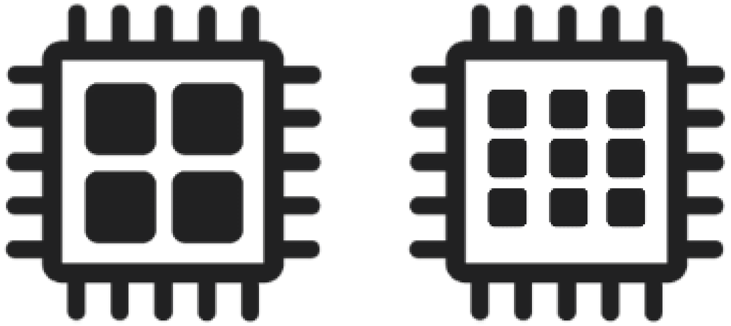
“Sierra” System

2000 2005 2010 2015 2020

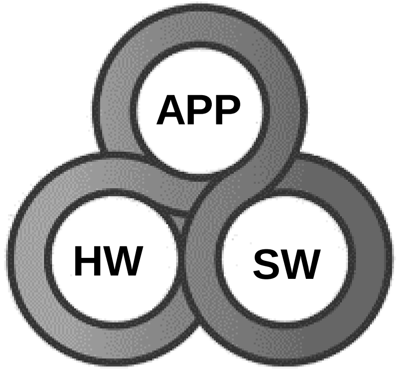
The Interconnect is the Enabling Technology



SMP to Clusters



Single-Core to Many-Core



Application

Software

Hardware

Co-Design

Thank You

Mellanox TECHNOLOGIES

Connect. Accelerate. Outperform.™