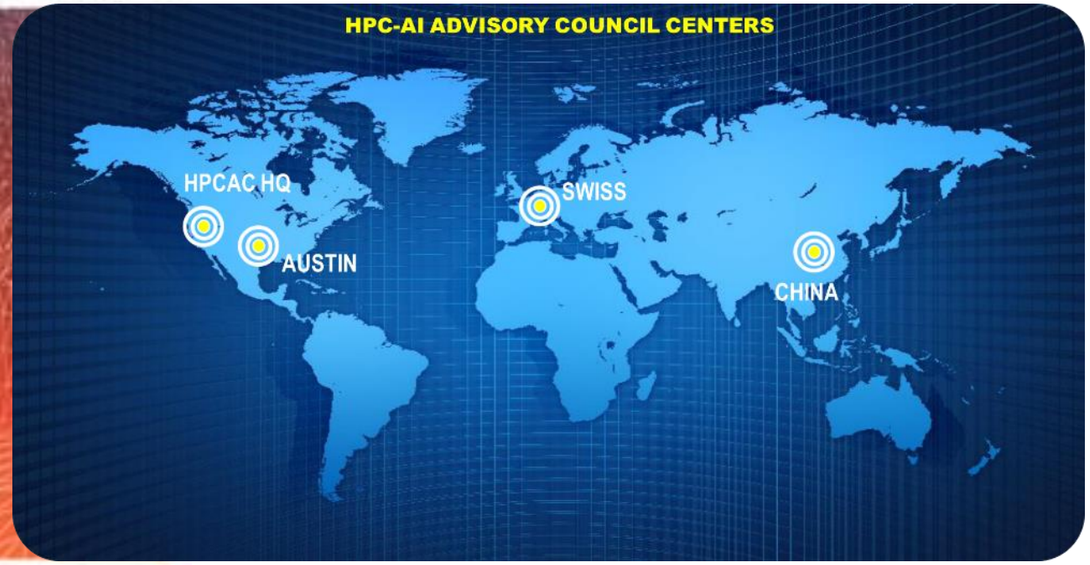
HPC-AI ADVISORY COUNCIL
NETWORK OF EXPERTISE

HPC-AI Advisory Council Introduction

Qingchun Song, Chair Of APAC

The HPC-AI Advisory Council

- Worldwide HPC and AI community organization, established in 2008
- More than 450 member companies / universities / research centers
- Bridges the gap between HPC and AI usage and its potential
- Provides best practices, education, technology demonstrations, development center
- Explores future technologies and future developments



HPC-AI Advisory Council Cluster Center

- The Council operates a cluster center with 14 clusters available for operation
- Providing free of charge access to variety of compute, network and storage technologies
- For more information: http://hpcadvisorycouncil.com/cluster_center.php

- Daytona_X AMD 8-node cluster
- Dual Socket AMD EPYC 7742 64-Core Processor @ 2.25GHz
- Mellanox ConnectX-6 HDR 200Gb/s InfiniBand/Ethernet
- Mellanox HDR Quantum Switch QM7800 40-Port 200Gb/s HDR InfiniBand
- Memory: 256GB DDR4 2666MHz RDIMMs per node
- Lustre Storage, NFS

- Dell C6400 32-node cluster
- Dual Socket Intel(R) Xeon(R) Gold 6148 CPU @ 2.40GHz
- Mellanox ConnectX-6 HDR100 100Gb/s InfiniBand/VPI adapters
- Mellanox HDR Quantum Switch QM7800 40-Port 200Gb/s HDR InfiniBand
- Memory: 192GB DDR4 2666MHz RDIMMs per node
- Lustre Storage, NFS

- Supermicro SYS-6029U-TR4 / Foxconn Groot 1A42USF00-600-G 32-node cluster
- Dual Socket Intel(R) Xeon(R) Gold 6138 CPU @ 2.00GHz
- Mellanox ConnectX-6 HDR/HDR100 200/100Gb/s InfiniBand/VPI adapters with Socket Direct
- Mellanox HDR Quantum Switch QM7800 40-Port 200Gb/s HDR InfiniBand
- Memory: 192GB DDR4 2666MHz RDIMMs per node
- 1TB 7.2K RPM SSD 2.5" hard drive per node

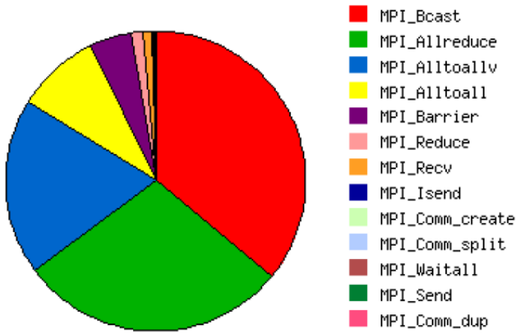
- IBM S822LC POWER8 8-node cluster
- Dual Socket IBM POWER8 10-core CPUs @ 2.86 GHz
- Mellanox ConnectX-4 EDR 100Gb/s InfiniBand adapters
- Mellanox Switch-IB SB7700 36-Port 100Gb/s EDR InfiniBand switch
- Memory: 256GB DDR3 PC3-14900 RDIMMs per node
- 1TB 7.2K RPM 6.0 Gb/s SATA 2.5" hard drive per node
- GPU: NVIDIA Kepler K80 GPUs

Multiple Application Best Practices Published

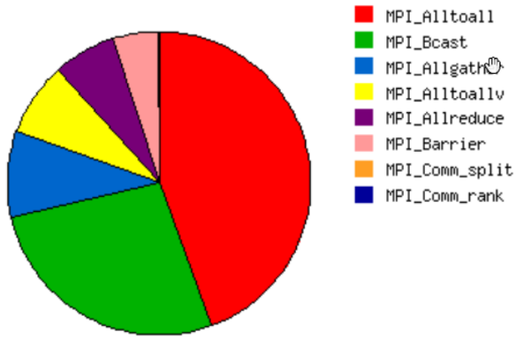
HPC-AI ADVISORY COUNCIL
NETWORK OF EXPERTISE

Abaqus	ABySS	AcuSolve	Amber	AMG	AMR
B_EFF	BiFrost	BQCD	BSMBench	CAM-SE	CASTEP
CCSM	CESM	ChaNGa	CFX	COSMO	CP2K
CPMD	Dacapo	Desmond	DL-POLY	Eclipse	FLOW-3D
Fluent	GADGET	Graph500	GRID	GROMACS	Himeno
HIT3D	HOOMD	HPCC	HPCG	HYCOM	ICON
Lattice	LAMMPS	LS-DYNA	MetaComp	miniFE	MILC
MSC	MR-Bayes	MM5	MPQC	NAMD	Nekbone
NEMO	NEMO5	NWChem	Octopus	OpenAtom	OpenFOAM
OpenMX	OptiStruct	PARATEC	PFA	PFLOTRAN	Quantum
RADIOSS	RFD	SNAP	SPECFEM3D	STAR	VASP
VPS	WRF				

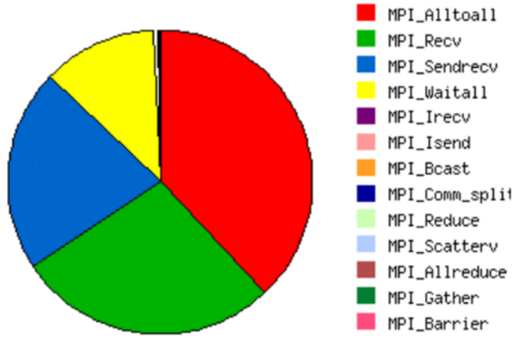
HPC Application Profiling – MPI Operations



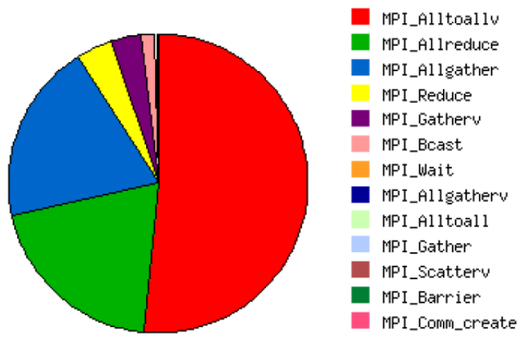
35%
of application's time is MPI communications



37%
of application's time is MPI communications



20%
of application's time is MPI communications



68%
of application's time is MPI communications

b-initio VASP ienna ackage imulation

GFS Global Forecast System

Global Forecast System

GROMACS FAST. FLEXIBLE. FREE.

GSI Gridpoint Statistical Interpolation

Special Interest Subgroups

- HPC|Scale Subgroup
- HPC|Cloud Subgroup
- HPC|Works Subgroup
- HPC|Storage Subgroup
- HPC|AI Subgroup
- HPC|Music Subgroup

HPC-AI Advisory Council 2021/2022 Activities

- **Conferences**
  - 2nd Annual Japan Conference, February 16, 2022
  - HPC-AI Track Of SCAsia22, March 1, 2022
  - 12th Annual Swiss Conference, March 1-4, 2022
  - 13th Annual Stanford California Conference, April 11-14, 2022
  - 6th HPC-AI Advisory Council Australia Conference
  - 14th Annual China Conference Under HPC China
  - 2022 China SC

- **University Competitions**
  - 5th Annual APAC HPC-AI Competition – March 2022
  - 11th Annual ISC Germany Student Cluster Competition – June 2022
  - 10th Annual APAC RDMA Programming Workshop and Competition – July 2022

- **For more information**
  - www.hpcadvisorycouncil.com
  - info@hpcadvisorycouncil.com

Global Competition – ISC Students Cluster Competition

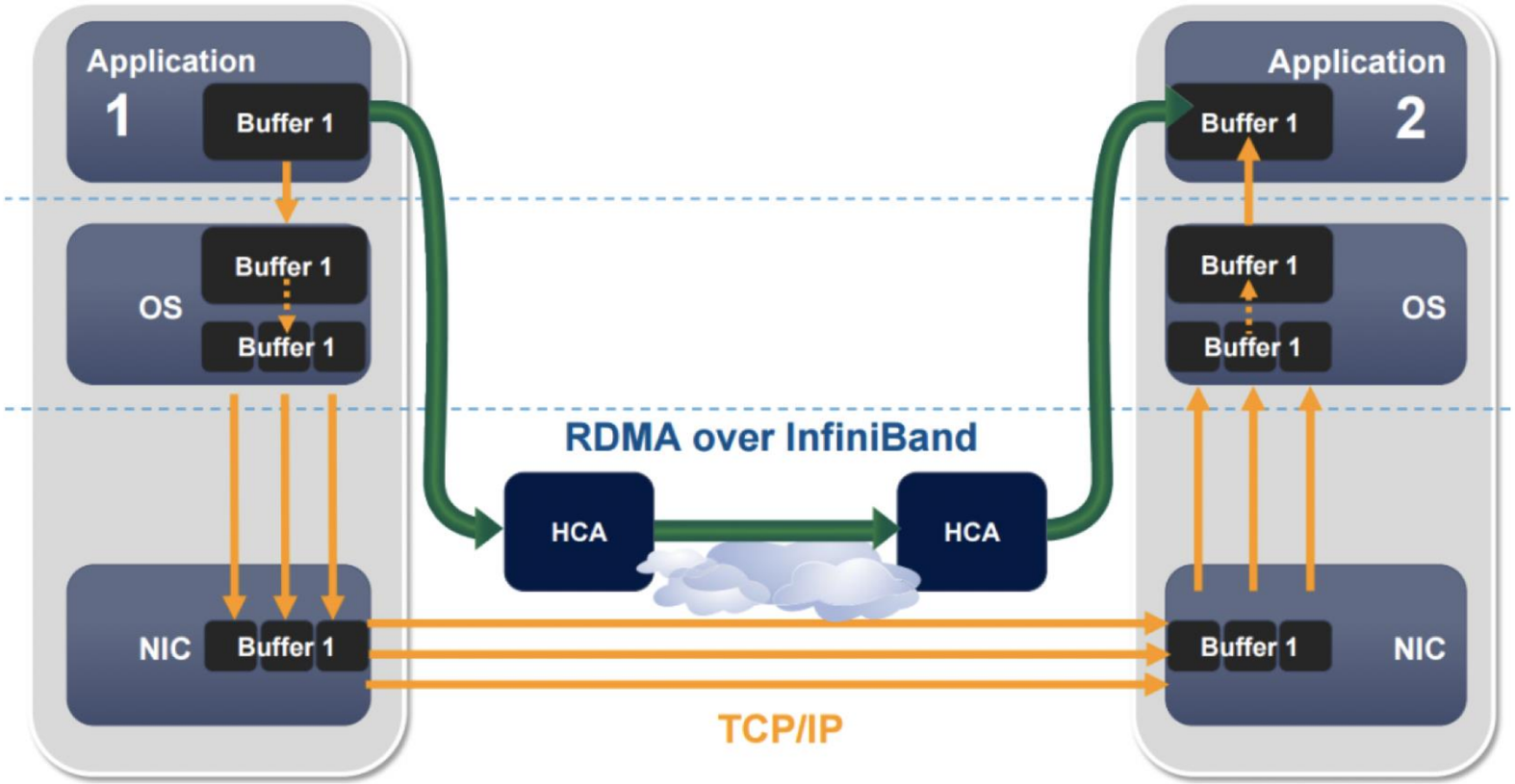
- **Micro Benchmarks**
  - HPC Challenge
  - High Performance LINPACK (HPL)
- **HPC Applications**
  - WRF
  - GPAW
  - MetaHipMer 2.0
  - LAMMPS
  - Coding Challenge

APAC Competition – HPC-AI Competition

- Part 1 – Artificial Intelligence – DLRM (Deep Learning Recommendation Model)
- Part 2 – High-Performance Computing – GROMACS (GROningen MAchine for Chemical Simulations)

APAC Competition - RDMA Programming Competition

- https://www.hpcadvisorycouncil.com/events/2021/rdma/



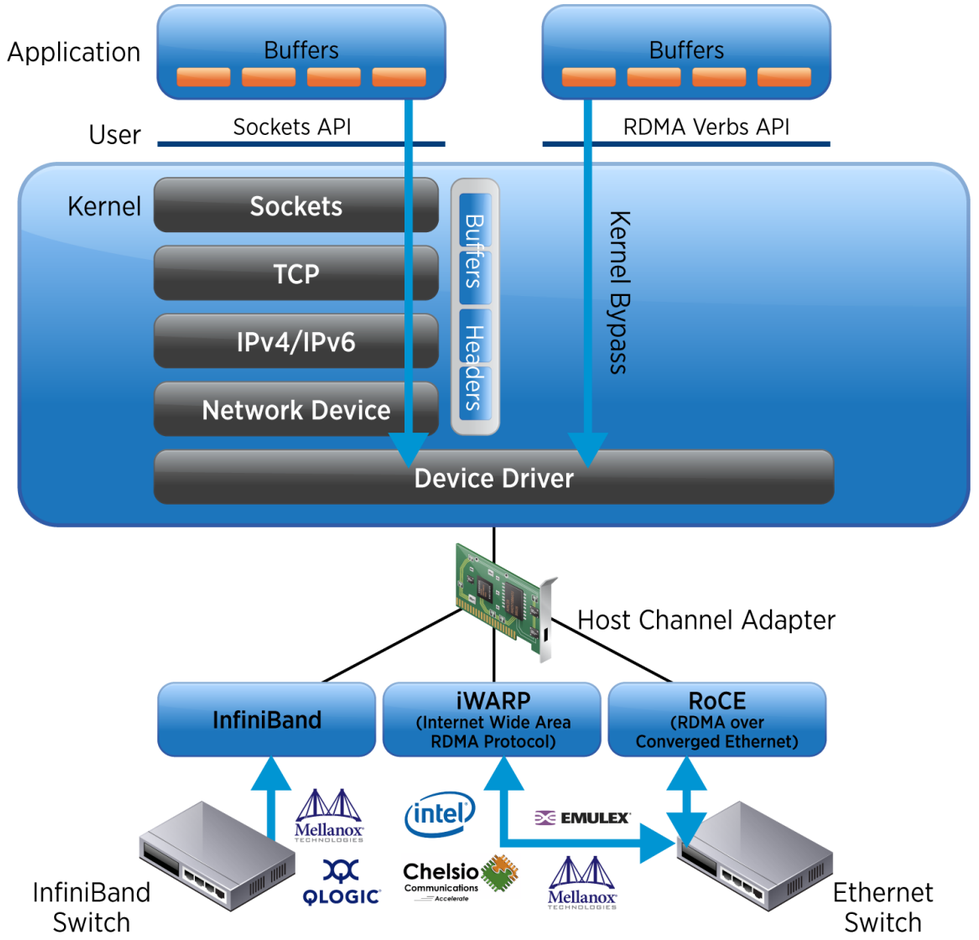
2022 APAC HPC-AI Competition

- https://www.hpcadvisorycouncil.com/events/2022/APAC-AI-HPC/
- https://www.hpcadvisorycouncil.com/events/2022/APAC-AI-HPC/register.php
- Qingchun@hpcadvisorycouncil.com
- Pengzhi@hpcadvisorycouncil..com

Advanced Network Technologies

OS Bypass & RDMA

- **The basic idea: cut the middleman!**
- **Zero Copies in traditional networking isn't "true"**
  - Buffering MUST occur between kernel and application
  - Communication buffer (kernel) to application buffer
- **Basic working principles:**
  - RDMA traffic sent directly to NIC without interrupting CPU
  - A remote memory region registers with the NIC first
  - NIC records virtual to physical page mappings.
  - When NIC receives RDMA request, it performs a Direct Memory Access into memory and returns the data to client.
  - Kernel bypass on both sides of traffic



RDMA Standards

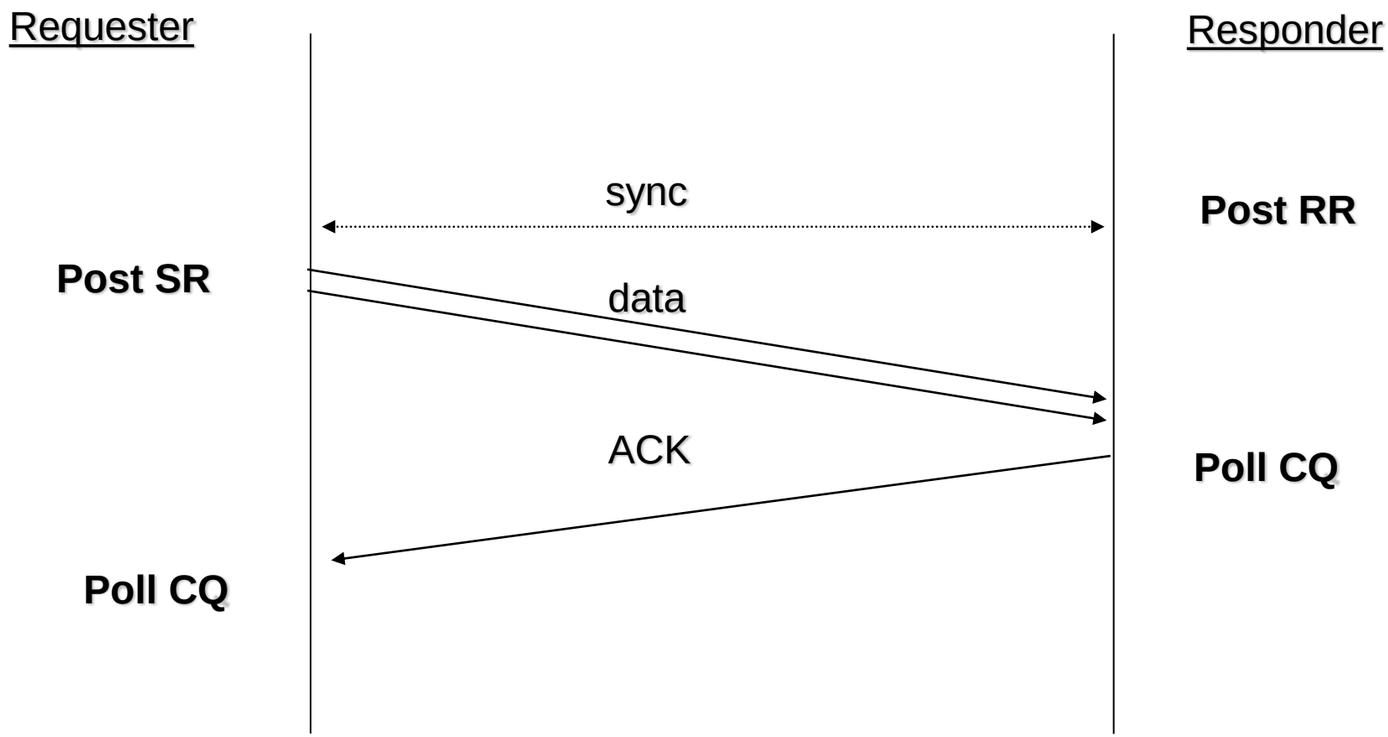
- **RDMA traditionally used in InfiniBand Networks**
  - Used extensively in HPC machines (Supercomputers)
  - Expensive, requires specialized hardware (physical network and NIC)
  - 200Gb/s standard

- **RoCE: RDMA done over Ethernet instead of InfiniBand**
  - (RDMA over Converged Ethernet)
  - Still requires specialized hardware
  - Cheaper because need only specialized NICs
  - 200Gb/s (and maybe 60Gb/s)
  - RoCE seems to scale worse

- **iWARP**
  - RDMA over TCP
  - Once again, cheaper; only needs specialized NICs

RDMA Opcodes: Send

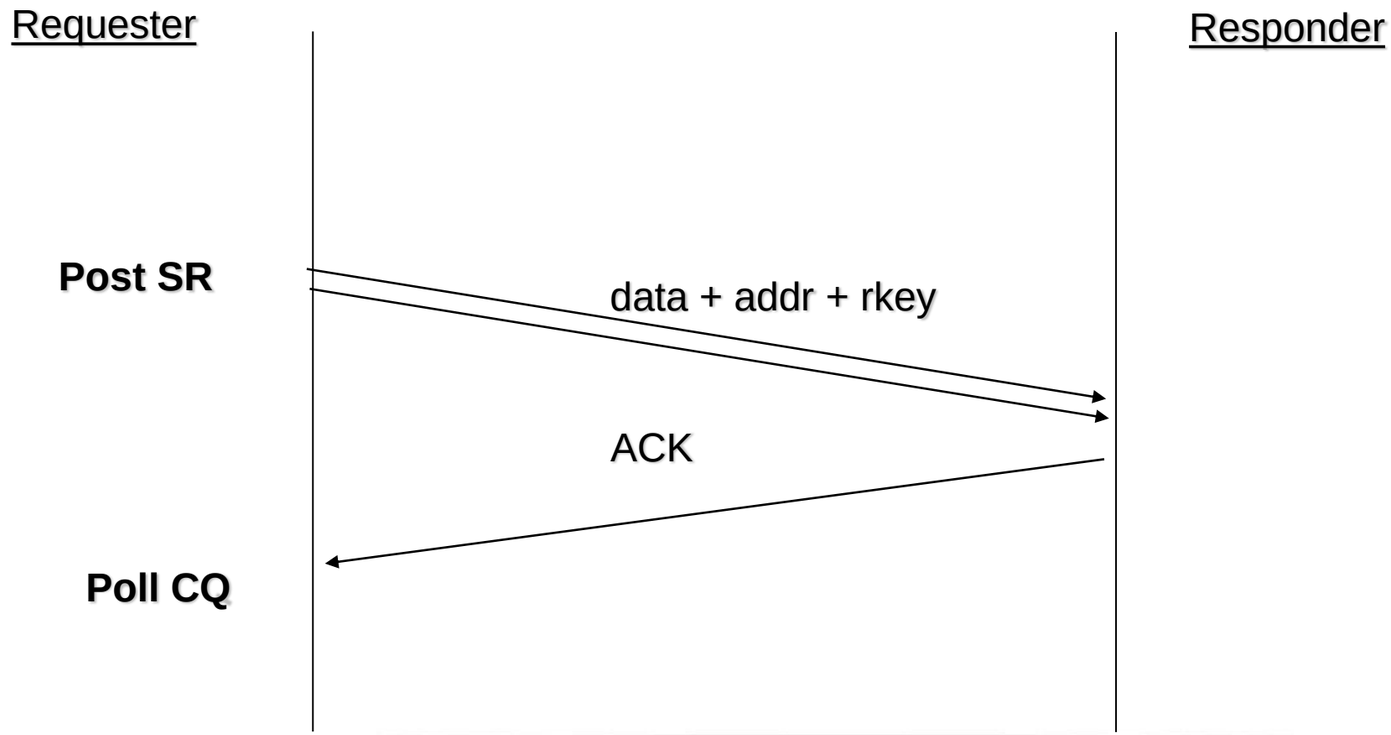
- **The responder Post Receive Requests (before data is received)**
- **The requester Post Send Request**
  - Only data is sent over the wire
- **ACK is sent only in reliable transport types**



15

RDMA Opcodes: RDMA Write

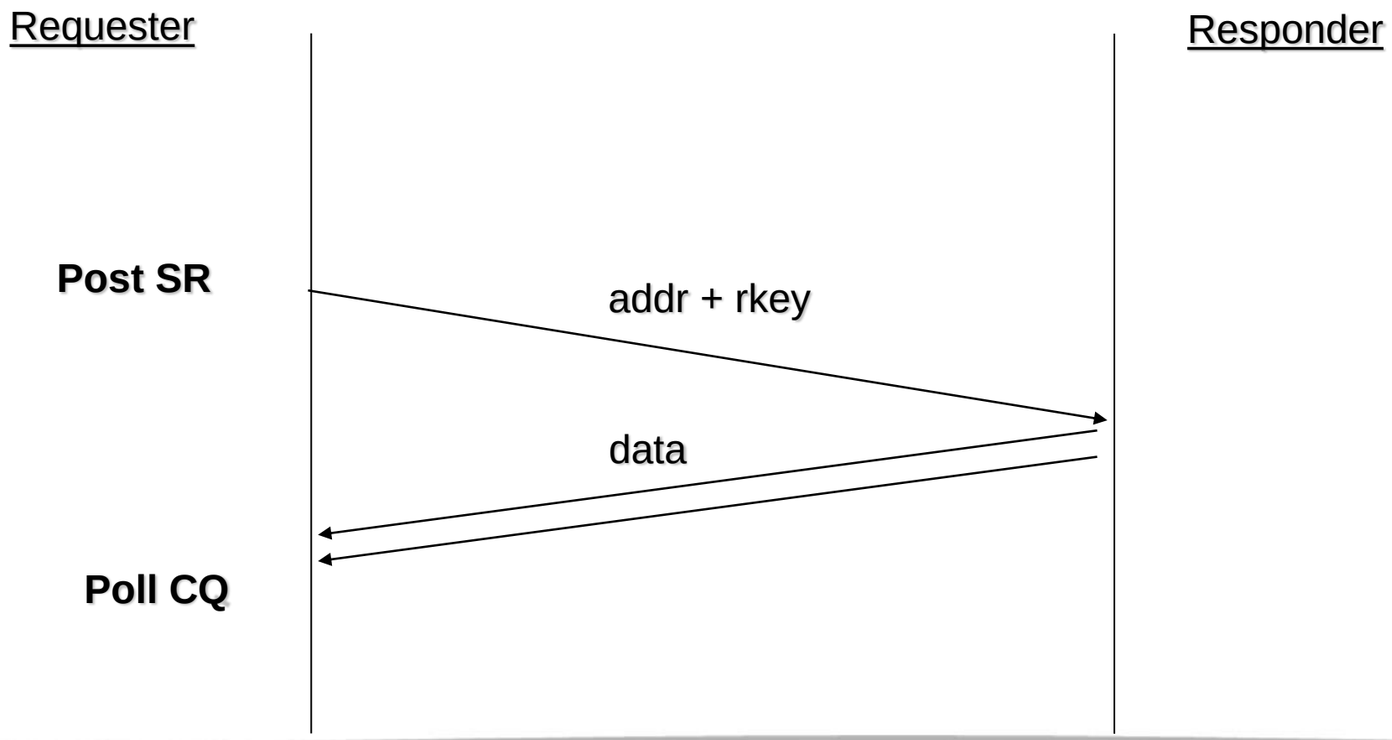
- **The requester Post Send Request**
  - Data and remote memory attributes are sent
  - Responder is passive
  - Immediate data can be used to consume RRs at the responder side
- **ACK is sent only in reliable transport types**



16

RDMA Opcodes: RDMA Read

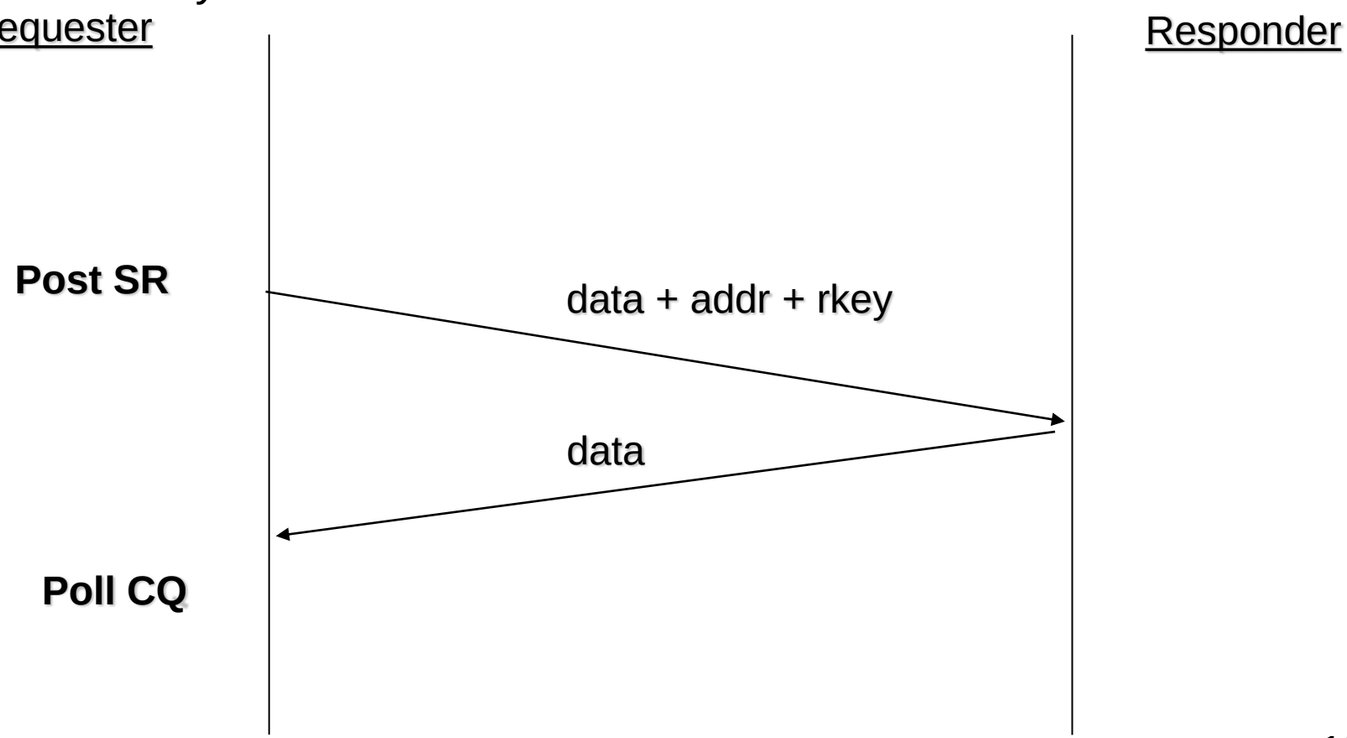
- **The requester Post Send Request**
  - Data and remote memory attributes are sent
  - Responder is passive
- **Data is sent from the responder**
  - Available only in reliable transport types



17

RDMA Opcodes: Atomic

- **The requester Post Send Request**
  - Data and remote memory attributes are sent
  - Responder is passive
- **Original data is sent from the responder**
  - Read-modify-write is performed in responder's memory
  - Available only in reliable transport types

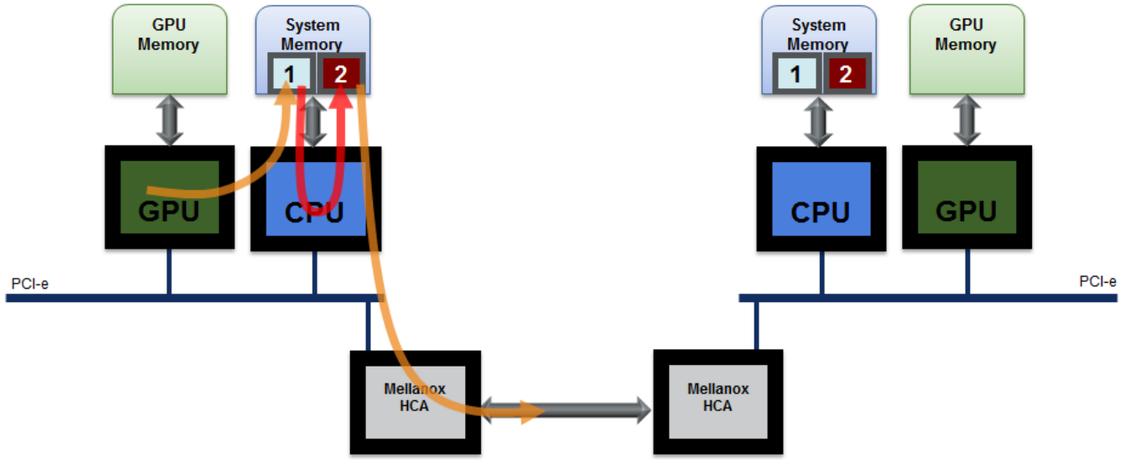


18

Evolution of GPUDirect RDMA

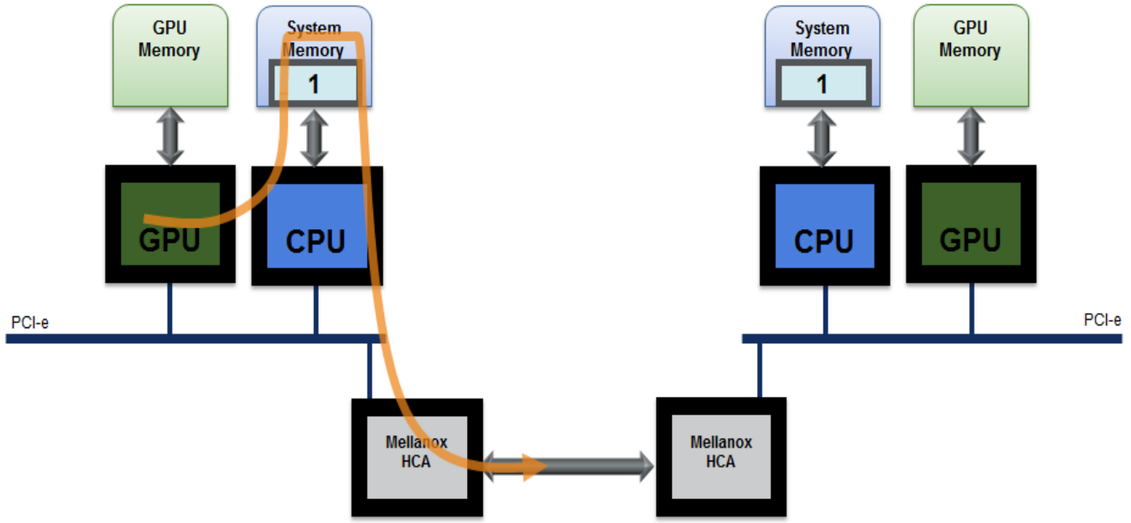
Before GPUDirect

- Network and third-party device drivers, did not share buffers, and needed to make a redundant copy in host memory.



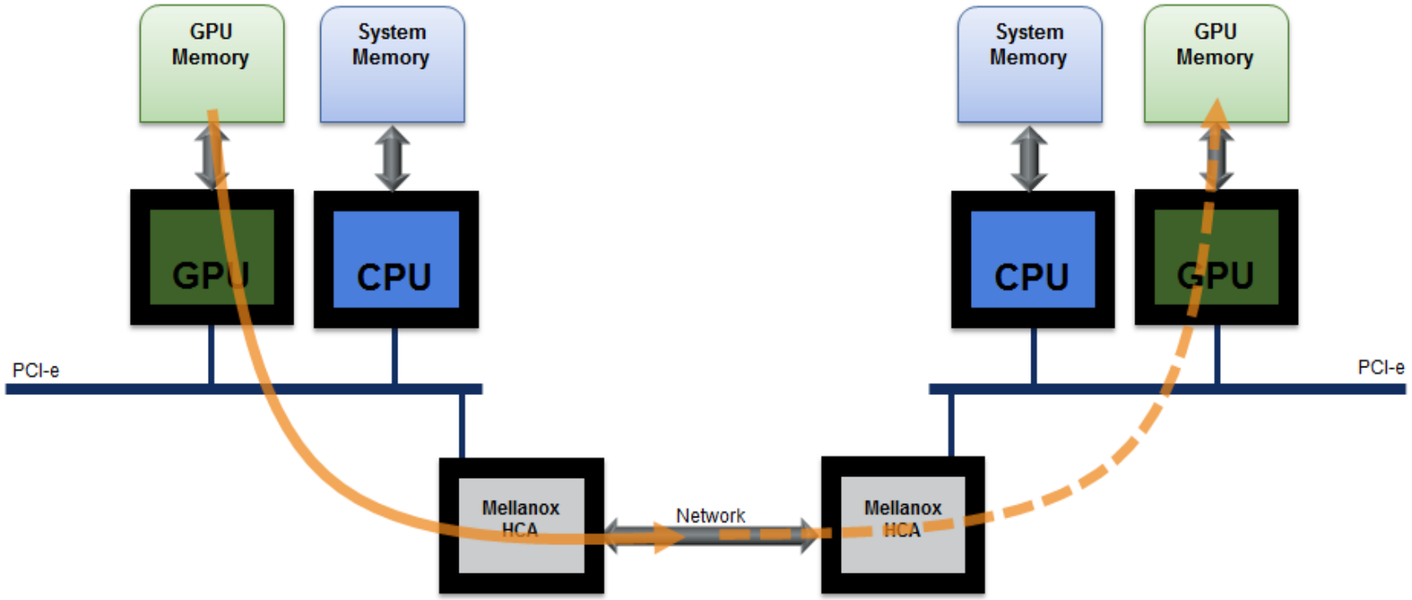
With GPUDirect Shared Host Memory Pages

- The network and GPU can share "pinned" (page-locked) buffers, eliminating the need to make a redundant copy in host memory.



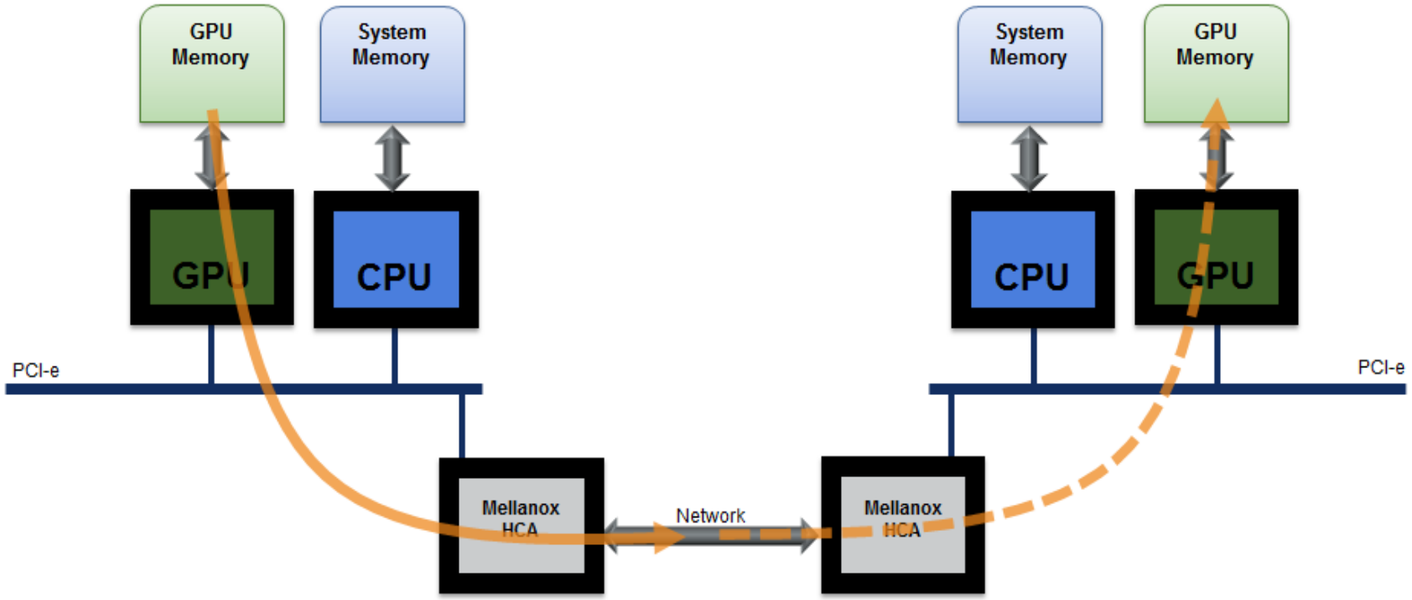
GPUDirect RDMA

- Eliminates CPU bandwidth and latency bottlenecks
- Uses remote direct memory access (RDMA) transfers between GPUs
- Resulting in significantly improved MPISendRecv() efficiency between GPUs in remote nodes

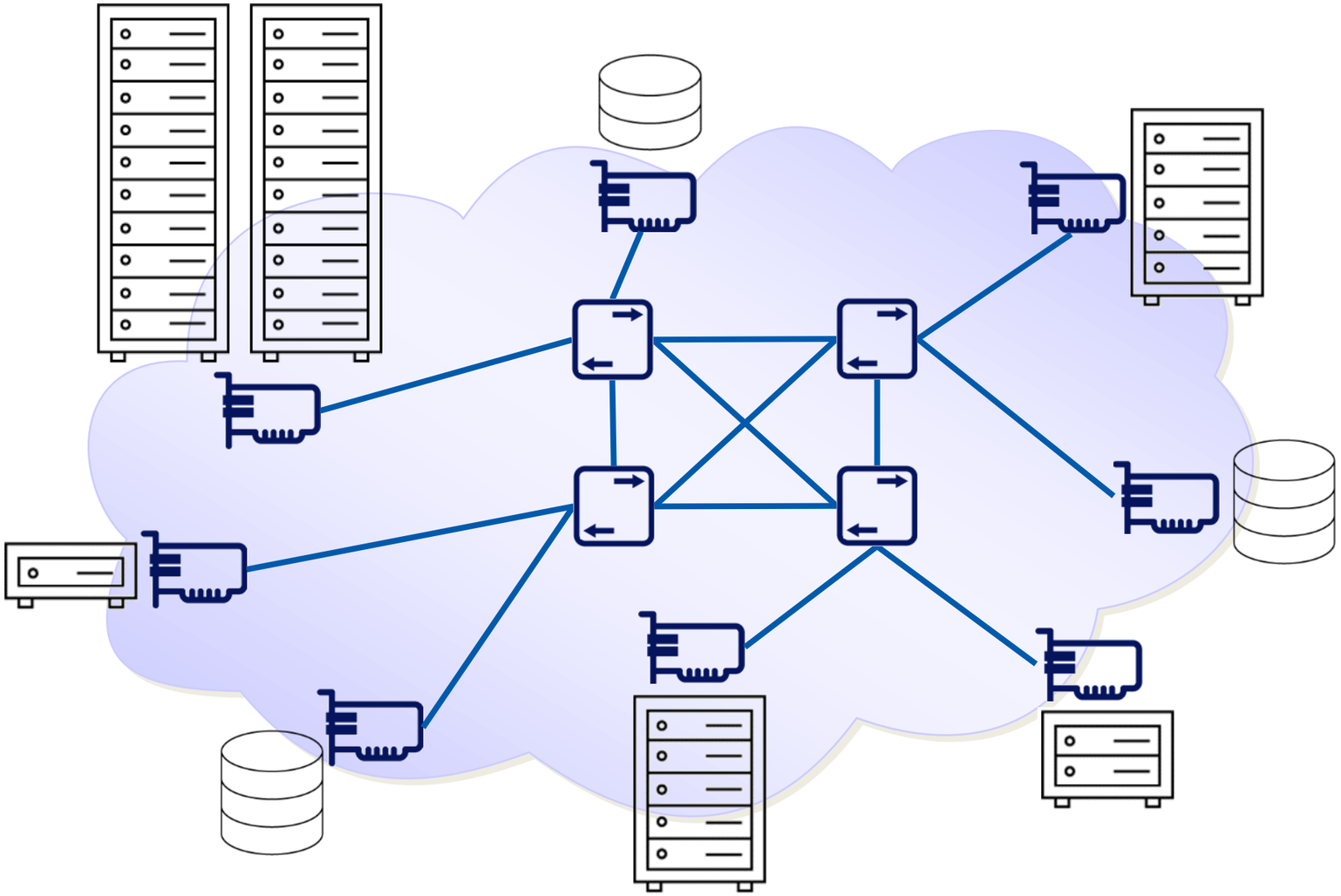


GPUDirect (3.0) RDMA

- Eliminates CPU bandwidth and latency bottlenecks
- Uses remote direct memory access (RDMA) transfers between GPUs
- Resulting in significantly improved MPISendRecv() efficiency between GPUs in remote nodes



Traditional HPC Data Center

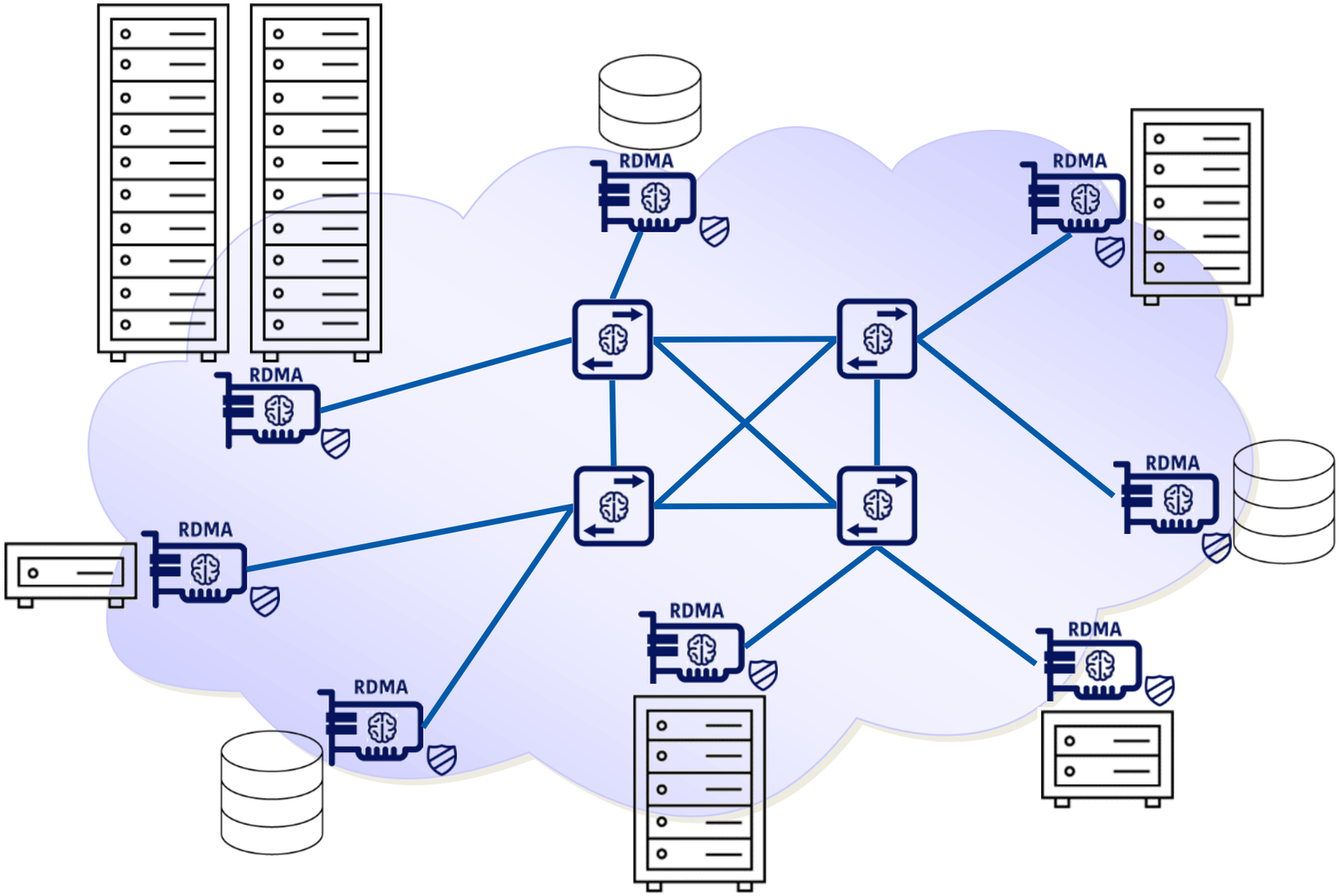


Cloud Native Supercomputing

In-Network Computing

Infrastructure Services

Computational Storage



Thank You