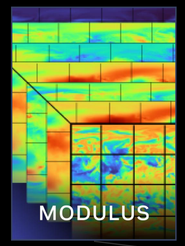
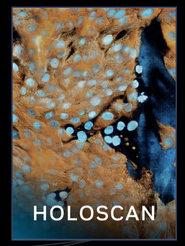
A Deep Dive into DPU Computing – Addressing HPC Performance Bottlenecks

Oct 2023

APPLICATION FRAMEWORKS





PLATFORM

NVIDIA AI

NVIDIA OMNIVERSE

ACCELERATION LIBRARIES

RTX
CUDA-X
CUDA

SYSTEM SOFTWARE

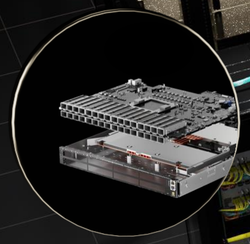
Magnum IO | DOCA | Base Command | Forge

HARDWARE

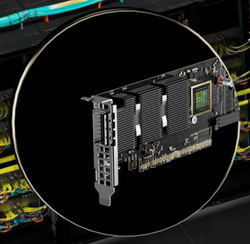
RTX DGX HGX EGX OVX AGX MLNX

GPU CPU DPU NIC SWITCH SOC

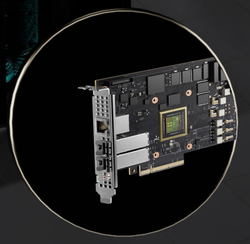
NVIDIA Networking



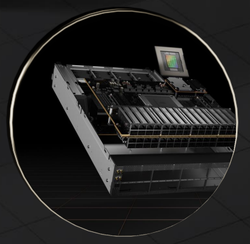
QUANTUM-2 INFINIBAND SWITCH



CONNECTX-7 SMARTNIC



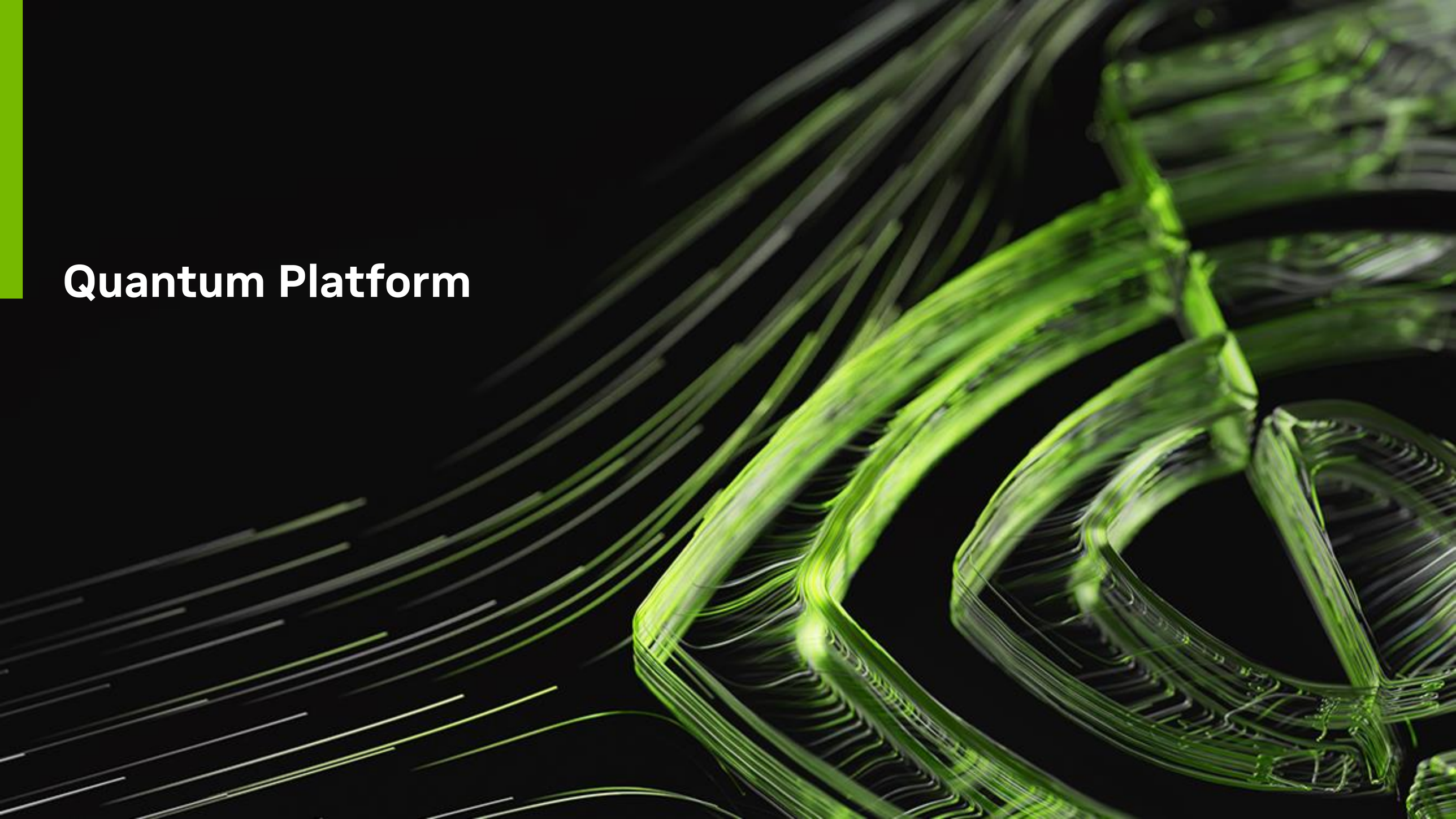
BLUEFIELD-3 DPU



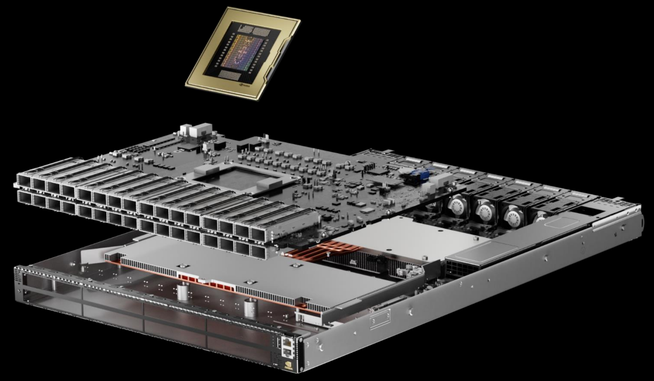
SPECTRUM-4 ETHERNET SWITCH

MANAGEMENT

Quantum Platform



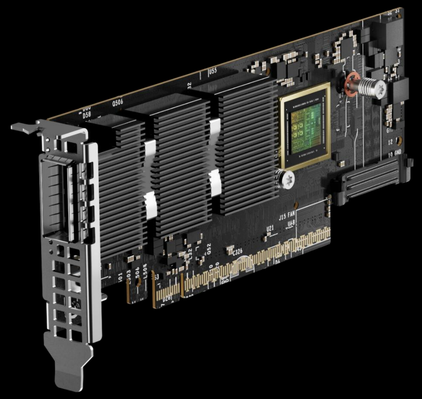
NVIDIA Quantum-2 400G In-Network Computing



QUANTUM-2 SWITCH

64-Ports of 400 Gbps or 128-Ports of 200 Gbps

SHARPv3 Data Reductions



CONNECTX-7 INFINIBAND

16 Core / 256 Threads Datapath Accelerator

Full Transport Offload and Telemetry

Hardware-Based RDMA / GPUDirect

MPI Tag Matching and All-to-All

Hardware Gather/Scatter

BLUEFIELD-3 INFINIBAND

16 Arm 64-Bit Cores

16 Core / 256 Threads Datapath Accelerator

ConnectX Networking

DDR memory interface

MPI and NCCL Accelerations

Computational Storage

Security Engines

BlueField Data Processing Unit

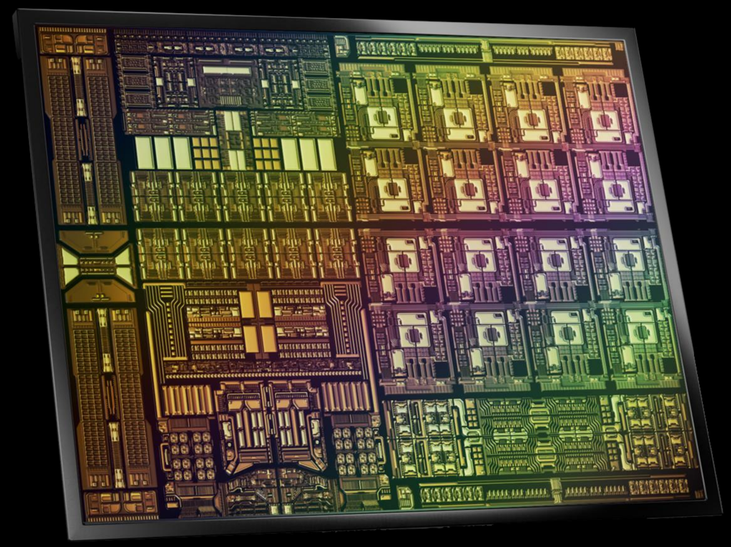
SOFTWARE DEFINED NETWORKING

SOFTWARE DEFINED SECURITY

SOFTWARE DEFINED STORAGE

Infrastructure Services

BlueField Infrastructure
Compute Platform



Data Center on a Chip

16 Arm 64-Bit Cores

16 Core / 256 Threads Datapath Accelerator

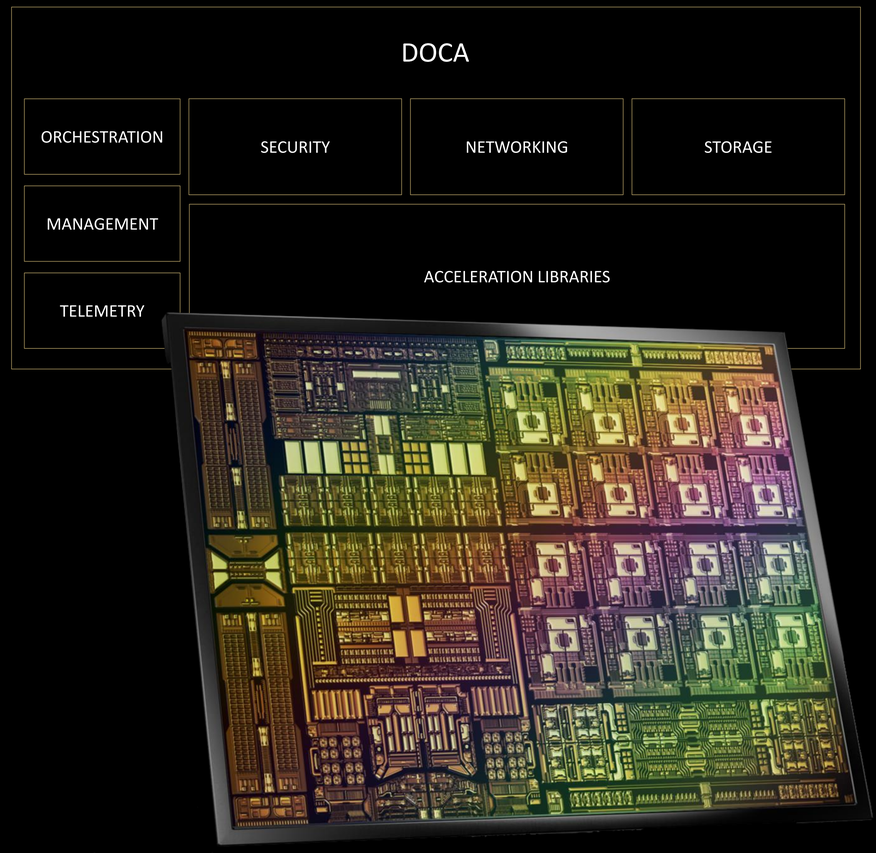
ConnectX InfiniBand / Ethernet

DDR memory interface

PCIe switch

BlueField System on a Chip

	BlueField-2	BlueField-3
Network Bandwidth	200Gb/s	400Gb/s
RDMA max msg rate	215Mpps	370Mpps
Compute Cores	8	16
Compute	SPECINT2K17: 9.8	SPECINT2K17: 42
Memory Bandwidth	17GB/s	80GB/s
NVMe SNAP	5.4M IOPs @ 4KB	10M IOPS @ 4KB

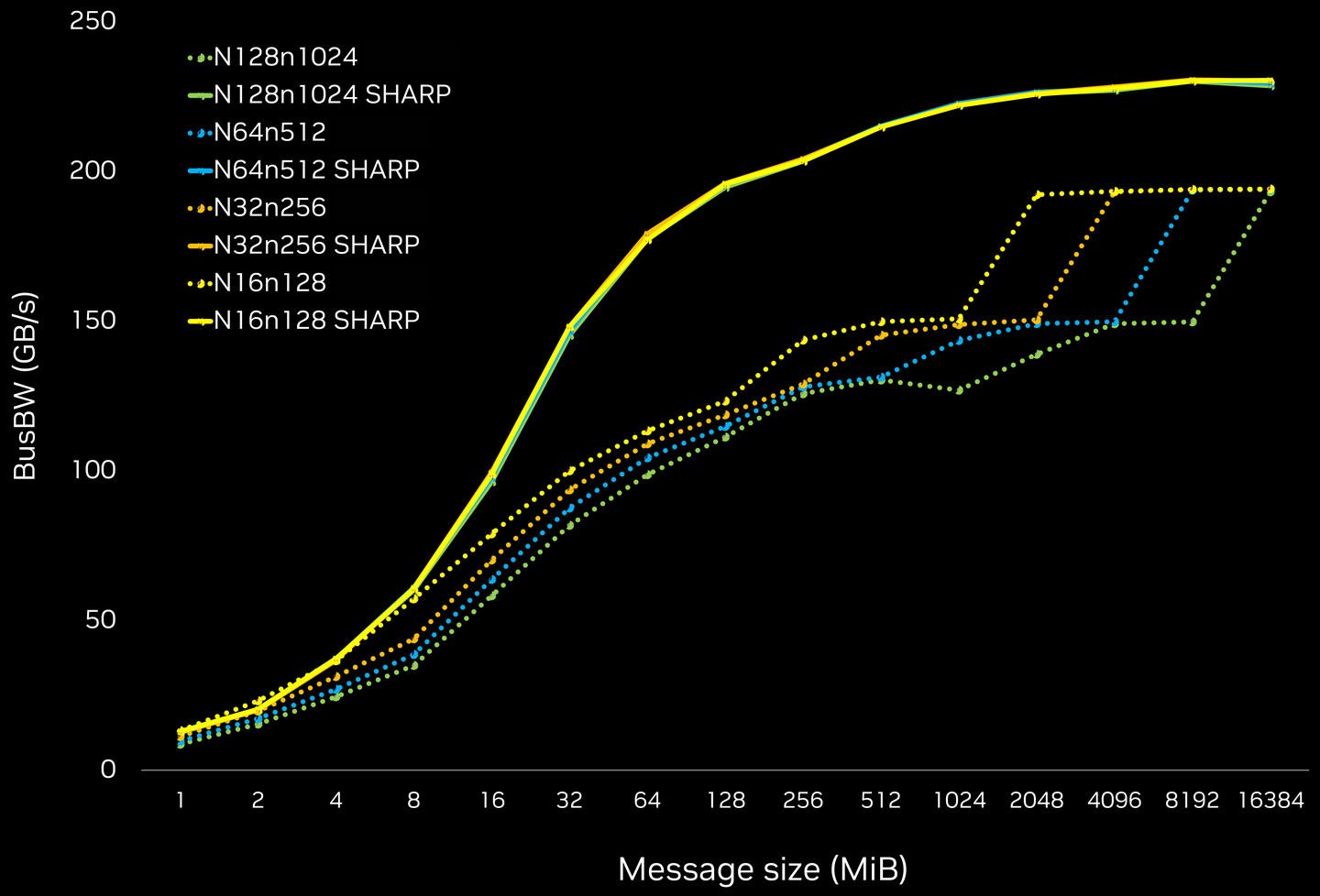


NVIDIA Quantum-2 InfiniBand Switch

	Quantum-2
Network Speed	400Gb/s
Network Protocols	InfiniBand
In-Network Computing	SHARPv3
Radix	Basic system: 64
2 switch Fat Tree Size	2048 ports
3 switch Fat Tree Size	65,536 ports
Switch Latency	210ns
Connectivity	DAC / Transceivers

NCCL All Reduce Performance with InfiniBand SHARP

InfiniBand SHARP maintains high bandwidth on large scale reaching up to 2x advantage

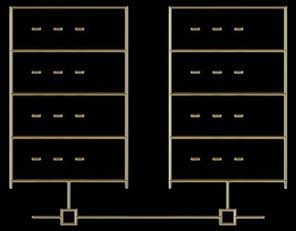


In-Network Computing to Solve Performance Bottlenecks

Overlapping

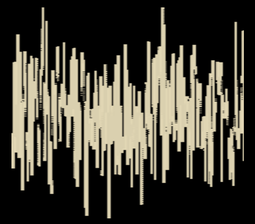
In-Network Computing
Asynchronous Progress
(Compute – Communication Overlap)

Load Imbalanced



In-Network Computing
and DPU Synchronization

Jitter



In-Network Computing
Infrastructure Processing

Multi-Job Performance

Adaptive Performance
Isolation

BlueField Design Considerations

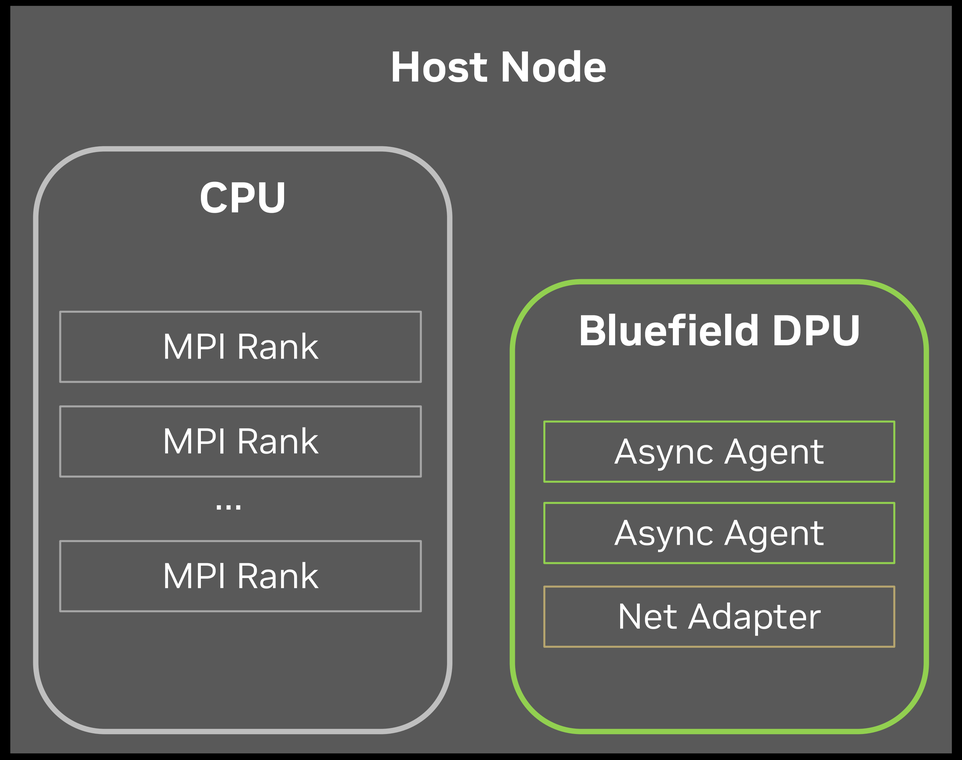
- Asynchronous with respect to the compute engines
- At least one order of magnitude less compute capabilities than the compute complex
  - Selective as to how much work to provide, so as not to become the bottleneck
  - Requires work sharing
- DPU cores may be less powerful computationally with respect to the host compute engines
- DPU have targeted acceleration engines
- Host and DPU need to be “in sync”
- Network access
  - Source/destination of network traffic
  - Can post network requests on behalf of memory locations that are host-resident
  - Agnostic to they type of compute host
- BlueField enhancements
    - Work requests can be posted on behalf of memory that is host-resident – Cross-GVMI memory keys
    - Some optimized data paths between the host and the BlueField – GGA

BlueField Design Considerations – Cont’ed

- Possess memory bandwidth independent of that of the host
  - Selectively use this memory resource to supplement what is available in the compute complex – not an all or none proposition
- Can’t do any better than saturate the network BW – need to do just enough to saturate the network

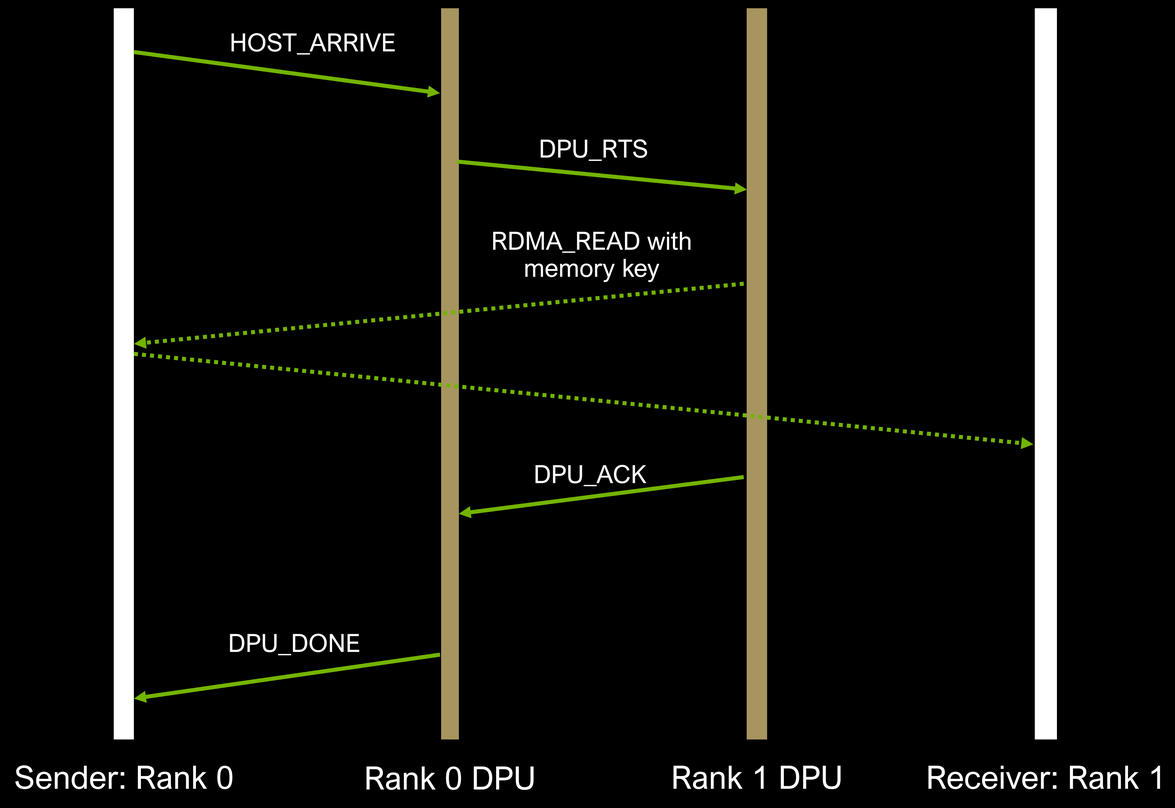
High Level System Components from Software's Perspective

- Host paired with local DPU
- Local DPU runs service processes (SP)
  - Each local user process (such as MPI process) has a service process that it is pair with
  - Each service process serves multiple local processes
  - Algorithm is split between host and DPU
    - Blocking and nonblocking may have different split
- Hosts and SP's may communicate with other hosts and/or SP's
- Cross-GVMI (XGVMI) - The DPU can initiates RDMA operations on behalf of host resident memory
  - DPU memory is involved only if the data originates from or is targeted to DPU memory

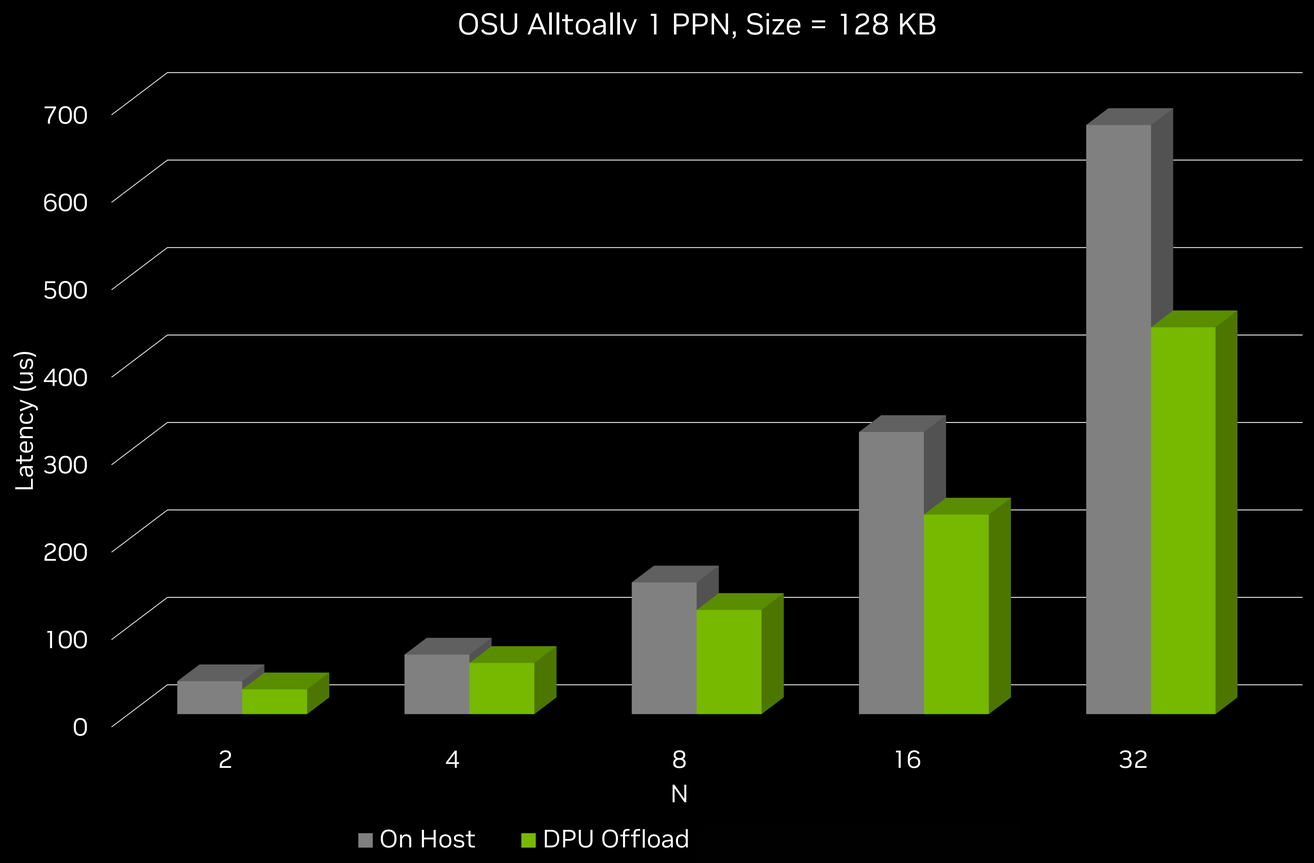


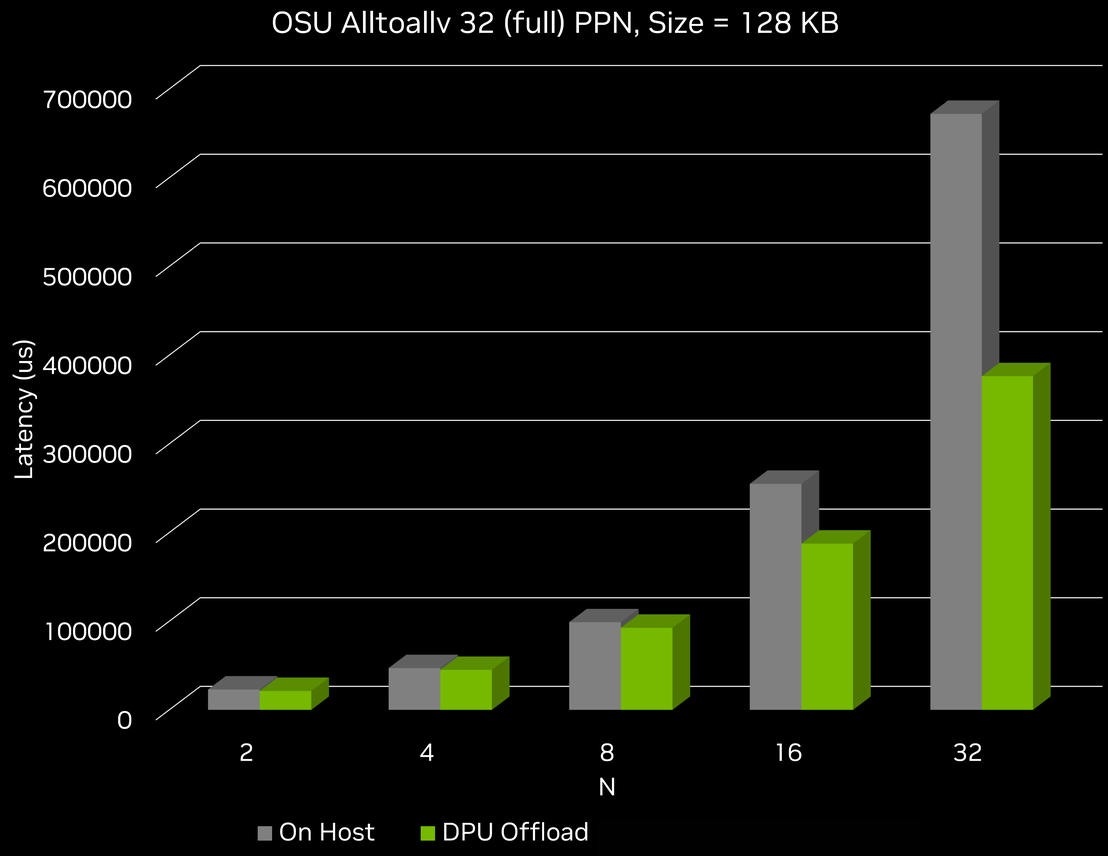
Offloading and Accelerating Data Exchange Example

An Element of Collective Algorithm

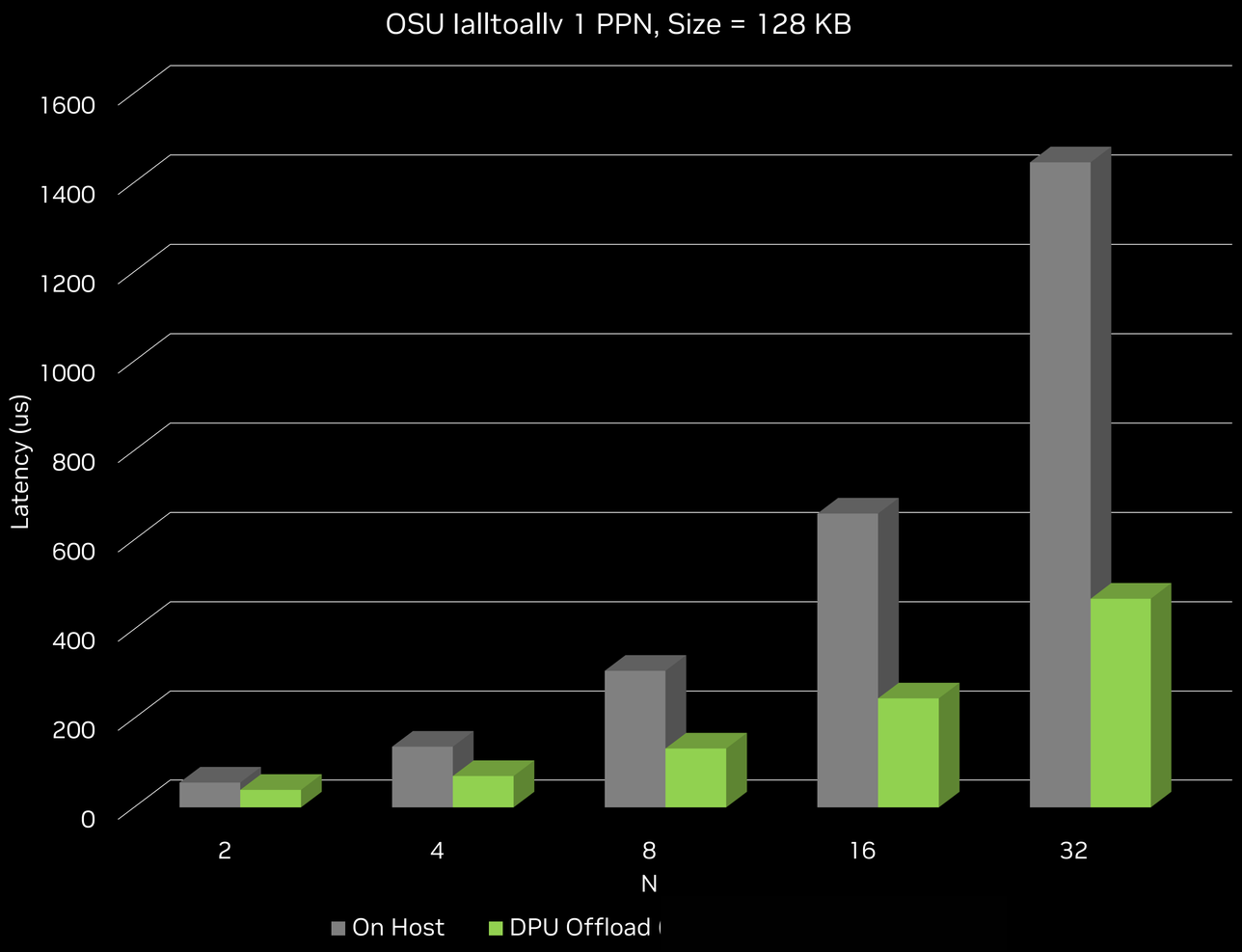


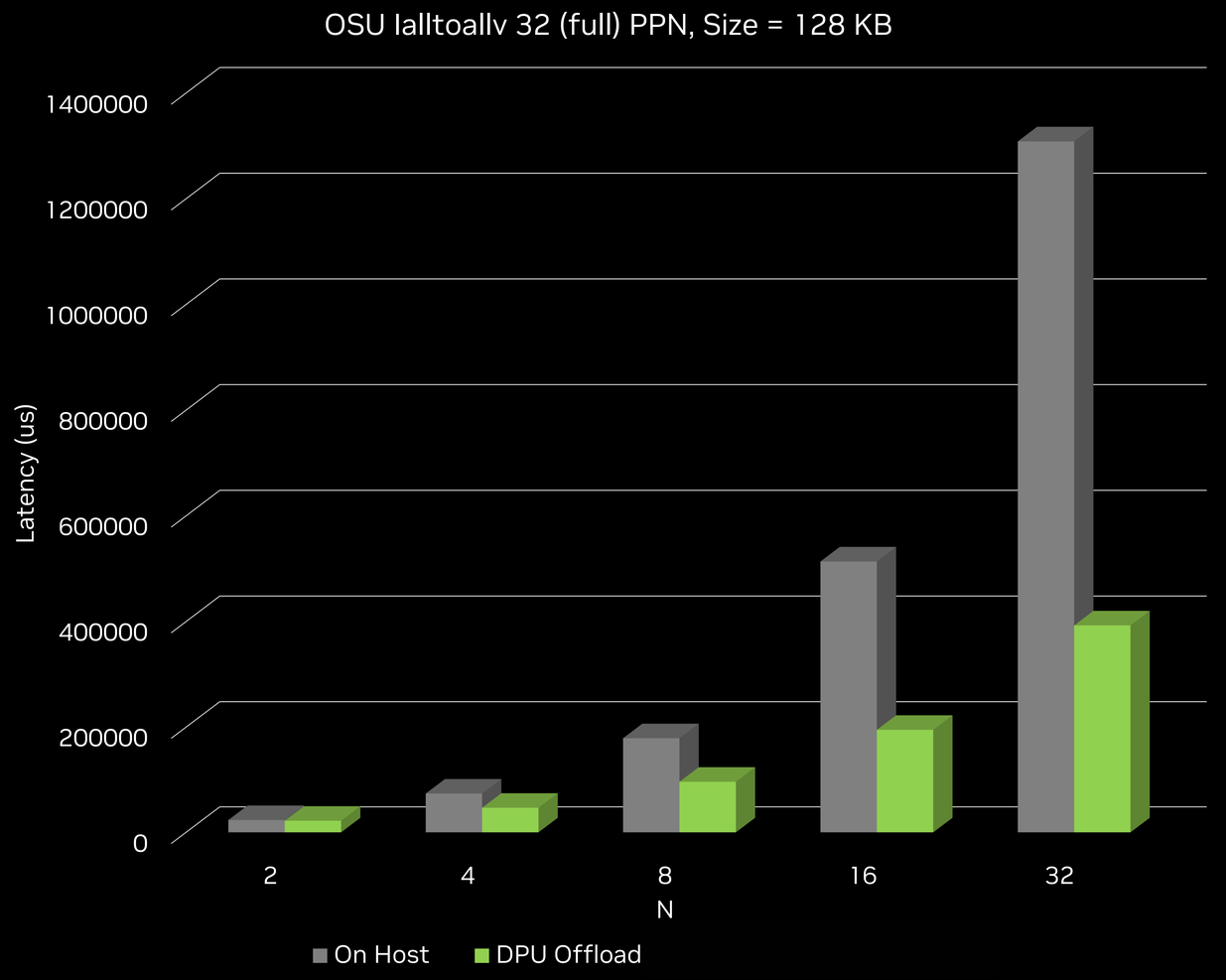
Alltoallv Latency



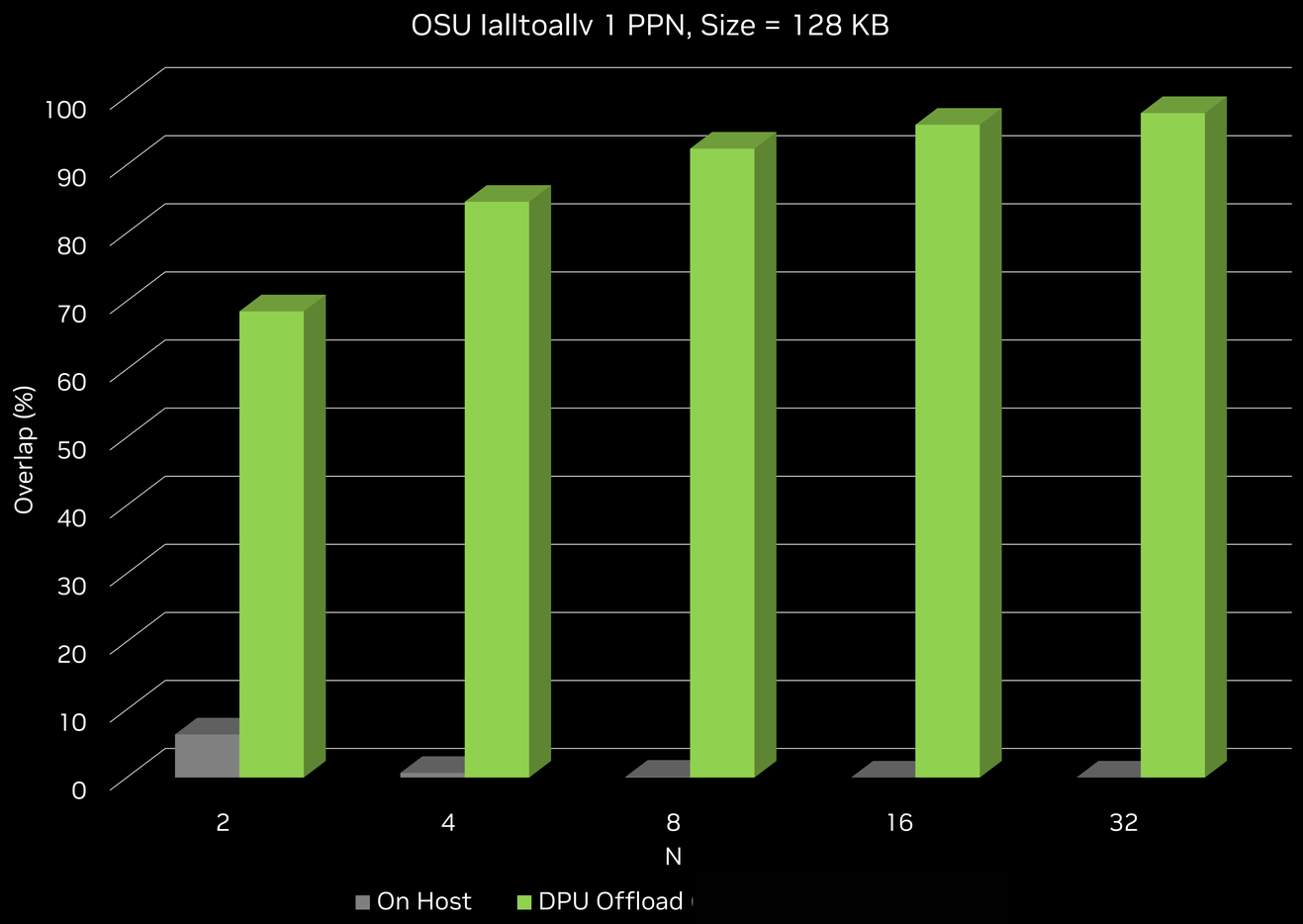


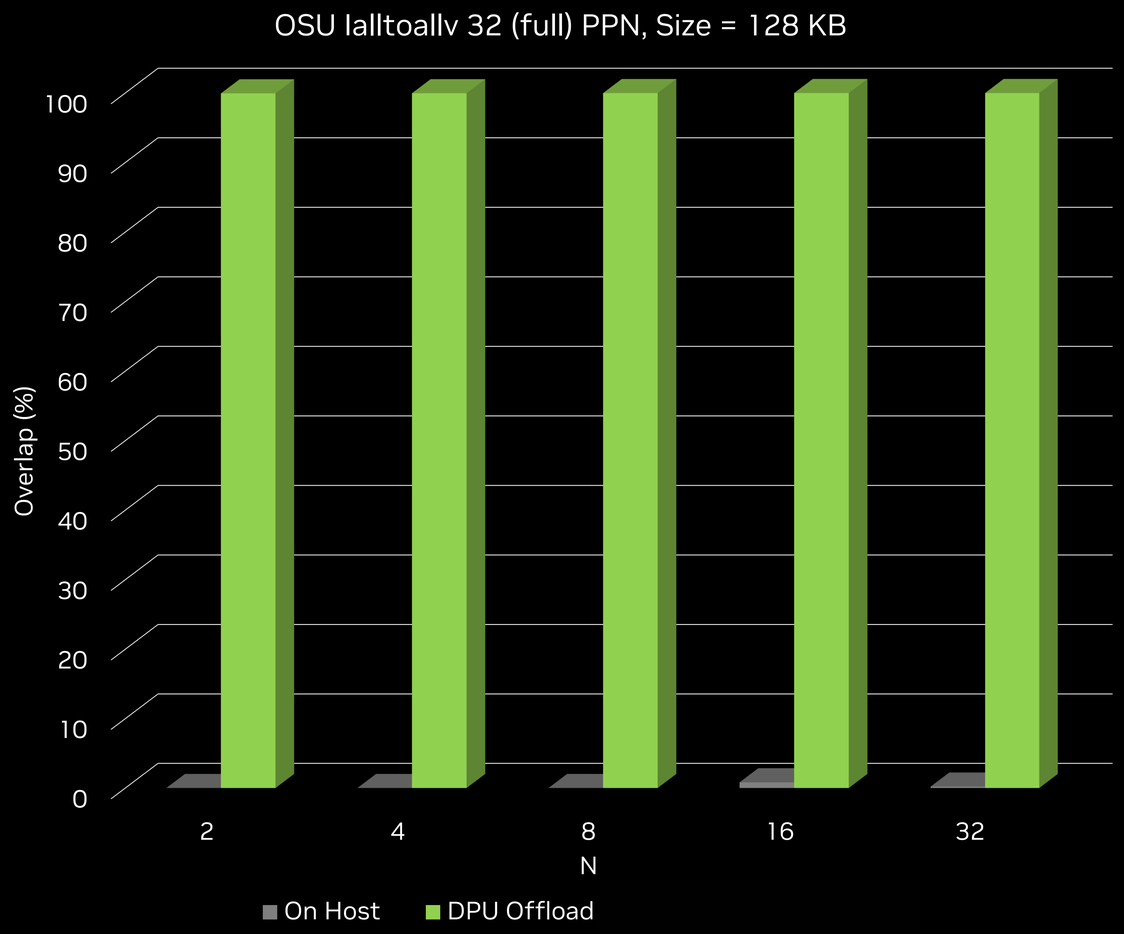
iAlltoallv latency



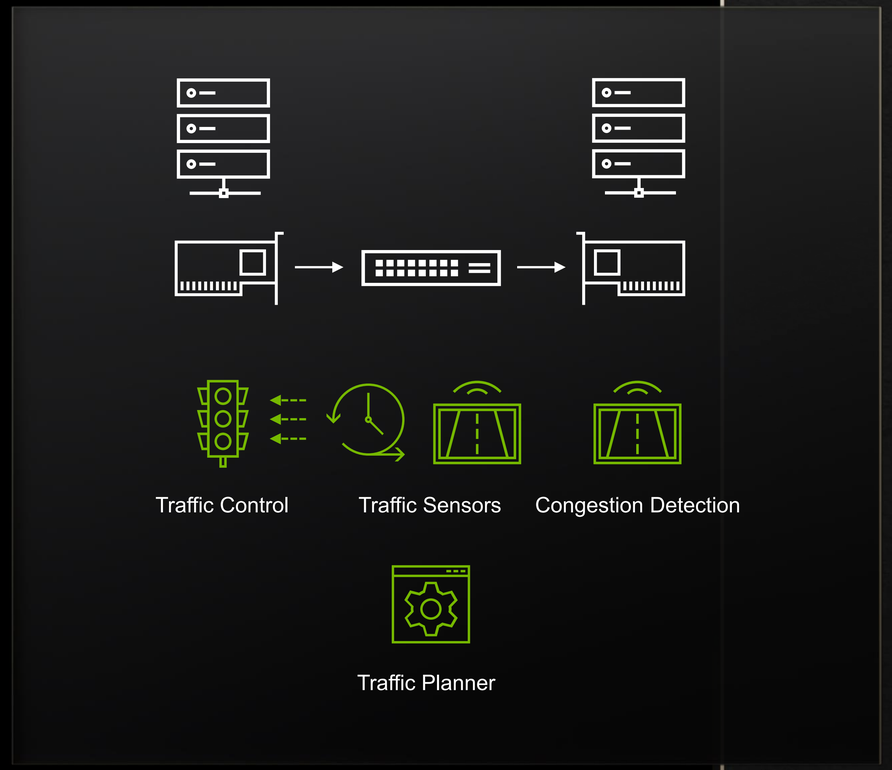


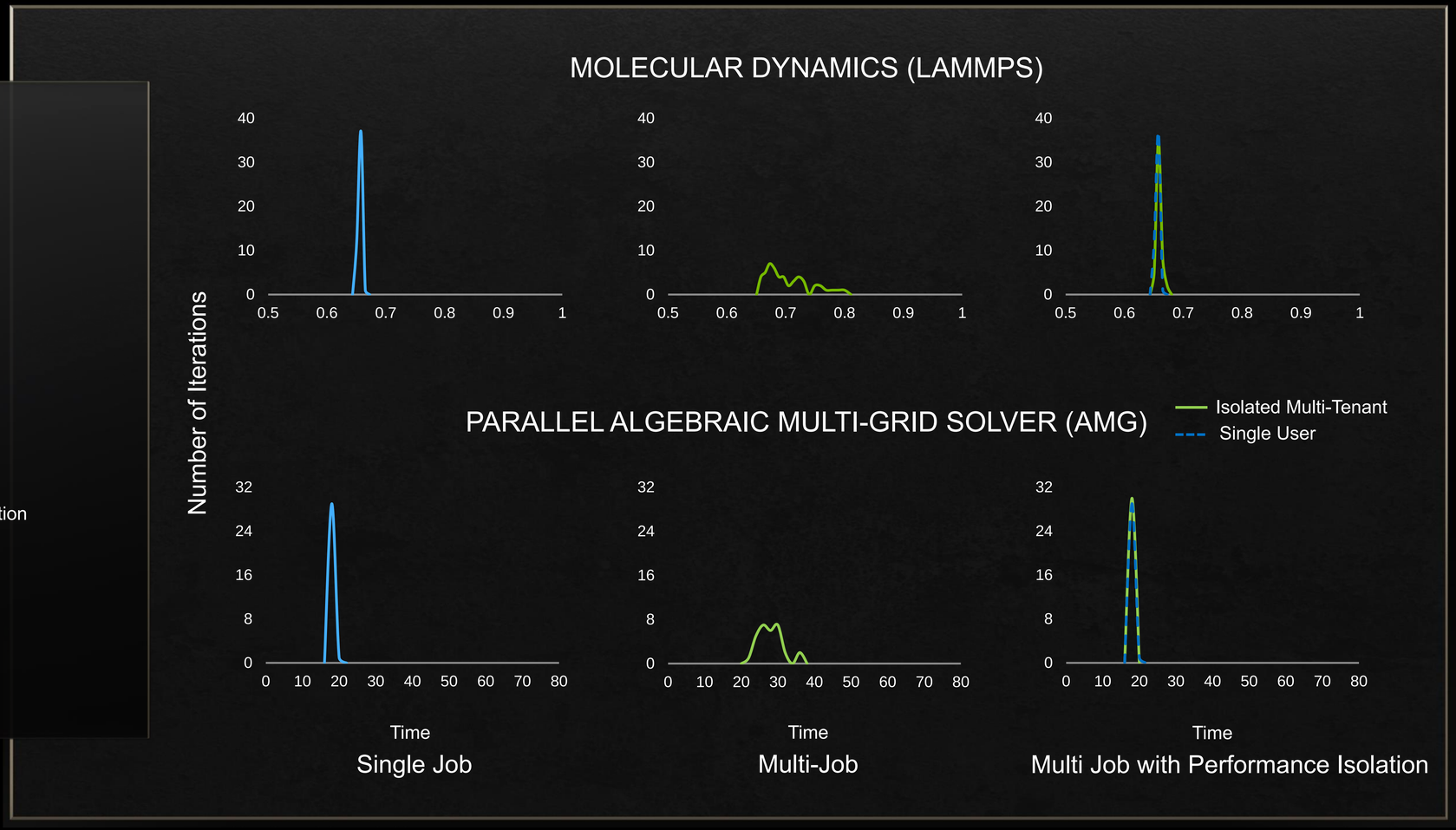
iAlltoallv compute/communication overlap





Performance Isolation






NVIDIA