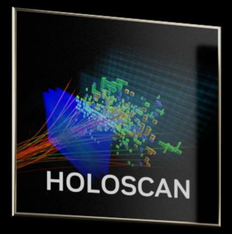
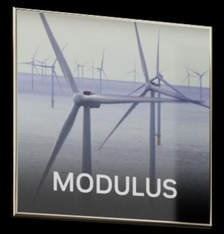
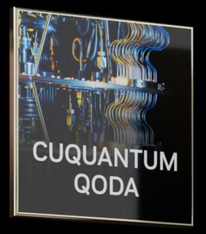
NVIDIA HPC Networking Platform

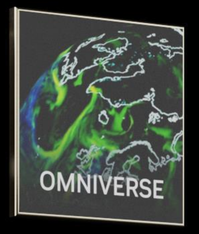
May 2023

NVIDIA Full Stack HPC Platform





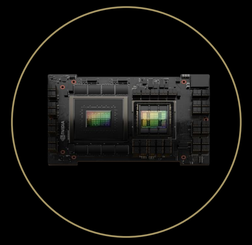




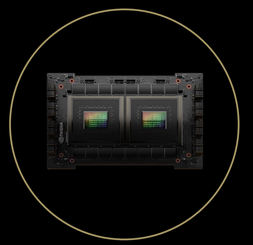
NVIDIA HPC

NVIDIA AI

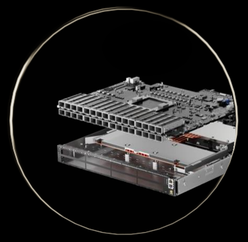
NVIDIA Omniverse



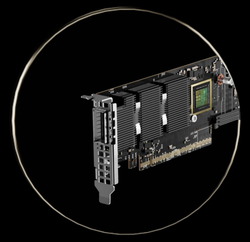
Grace Hopper



Grace Superchip

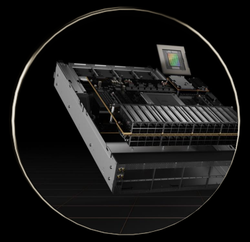


QUANTUM-2 INFINIBAND SWITCH



CONNECTX-7 SMARTNIC

BLUEFIELD-3 DPU



SPECTRUM-4 ETHERNET SWITCH

MANAGEMENT

BlueField Data Processing Unit

SOFTWARE DEFINED NETWORKING

SOFTWARE DEFINED SECURITY

SOFTWARE DEFINED STORAGE

Infrastructure Services

BlueField Infrastructure
Compute Platform

Data Center on a Chip

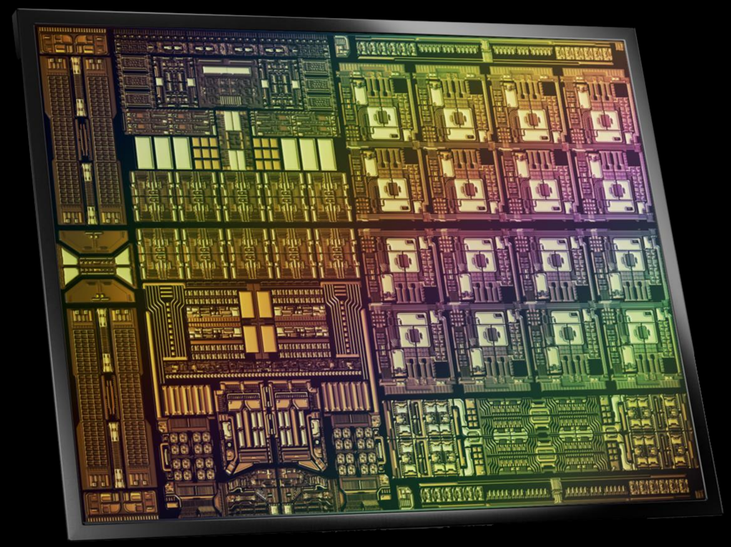
16 Arm 64-Bit Cores

16 Core / 256 Threads Datapath Accelerator

ConnectX InfiniBand / Ethernet

DDR memory interface

PCIe switch



The New Computing Platform for the Data Center Infrastructure

SOFTWARE DEFINED NETWORKING

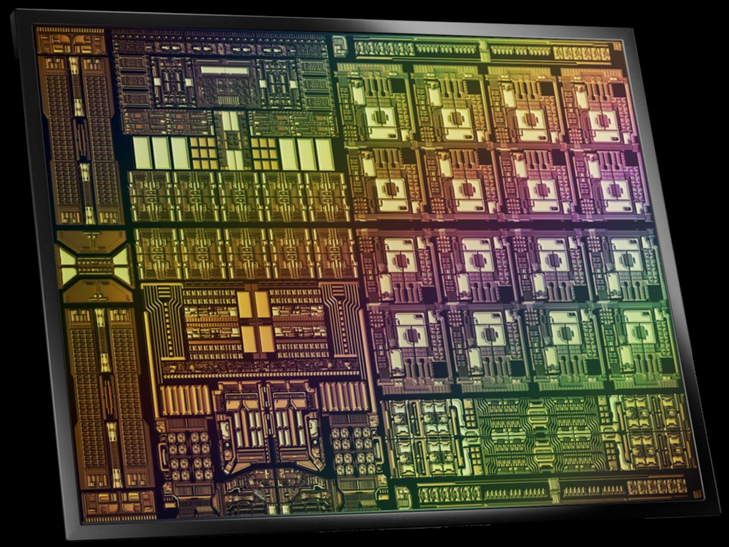
SOFTWARE DEFINED SECURITY

SOFTWARE DEFINED STORAGE

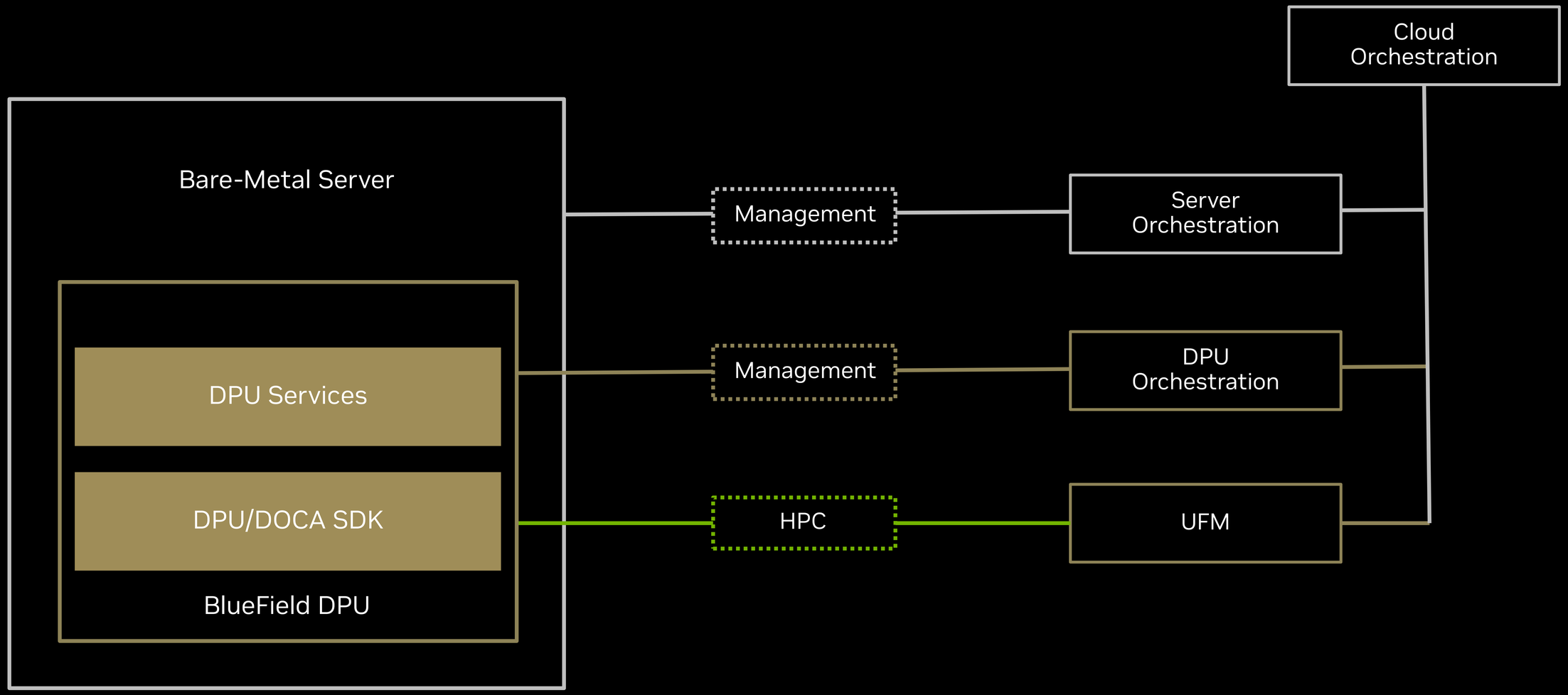
Infrastructure Services

	BlueField-2	BlueField-3
Network Bandwidth	200Gb/s	400Gb/s
RDMA msg rate	215Mpps	370Mpps
Compute	SPECINT2K17: 9.8	SPECINT2K17: 42
Memory Bandwidth	17GB/s	80GB/s

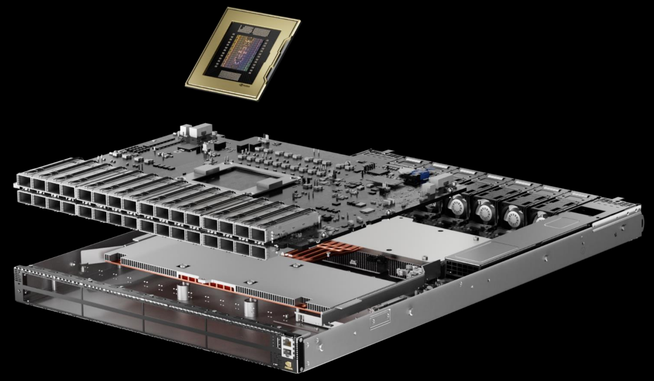
BlueField Infrastructure Compute Platform



Delivering Cloud Native Supercomputing



NVIDIA Quantum-2 400G In-Network Computing



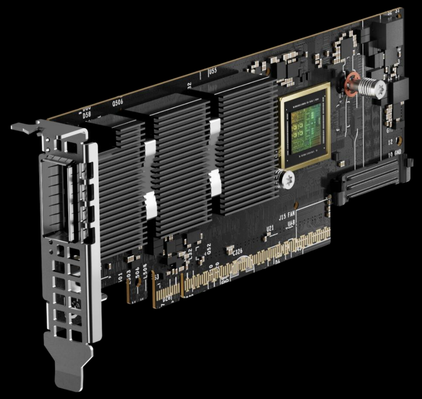
QUANTUM-2 SWITCH

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

SHARPv3 Small Message Data Reductions

SHARPv3 Large Message Data Reductions

32X More AI Acceleration Engines



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

BLUEFIELD-3 INFINIBAND

16 Arm 64-Bit Cores

16 Core / 256 Threads Datapath Accelerator

Full Transport Offload and Telemetry

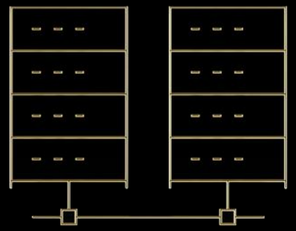
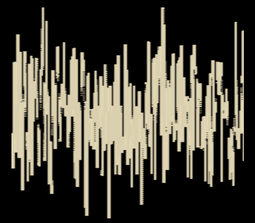
Hardware-Based RDMA / GPUDirect

MPI and NCCL Accelerations

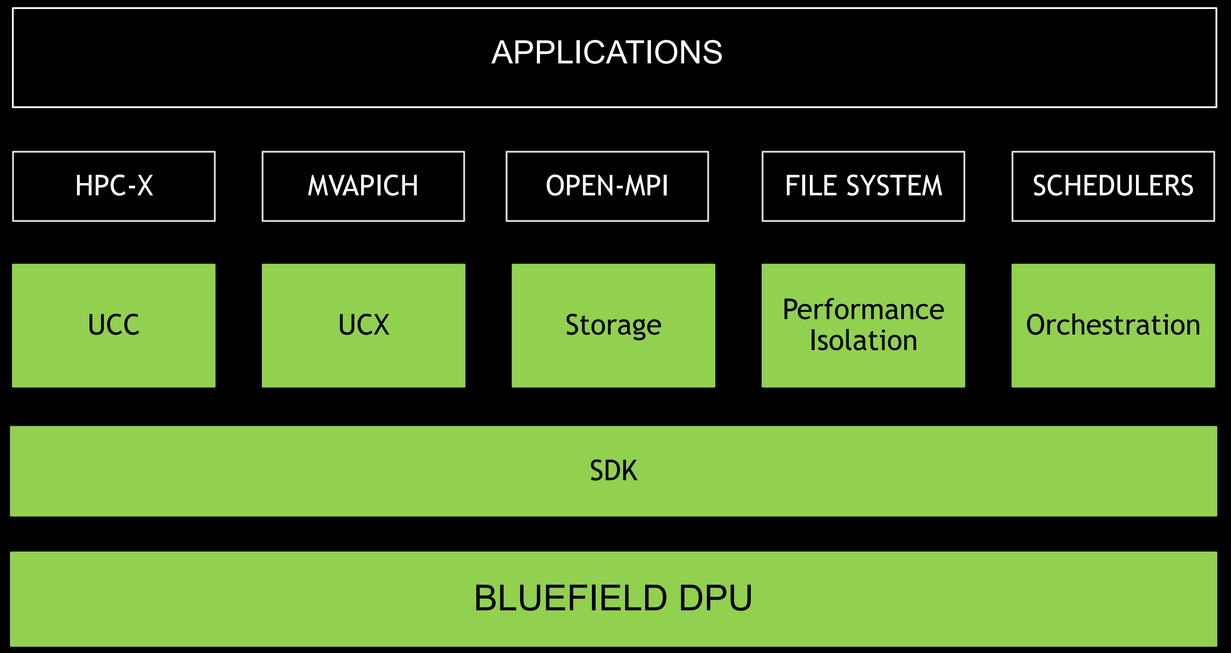
Computational Storage

Security Engines

HPC Performance Bottlenecks

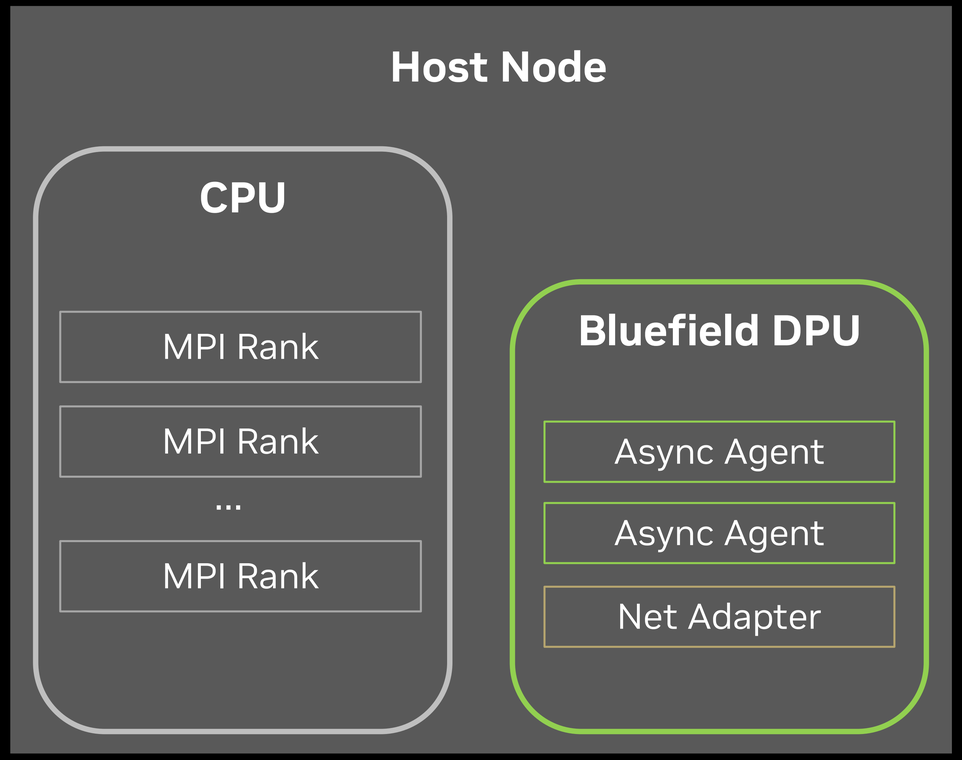
Overlapping	Load Imbalanced	Jitter	Multi-Job Performance
			
In-Network Computing Asynchronous Progress (Compute – Communication Overlap)	In-Network Computing and DPU Synchronization	In-Network Computing Infrastructure Processing	Adaptive Performance Isolation

Accelerating HPC Applications with DPU/DOCA Services



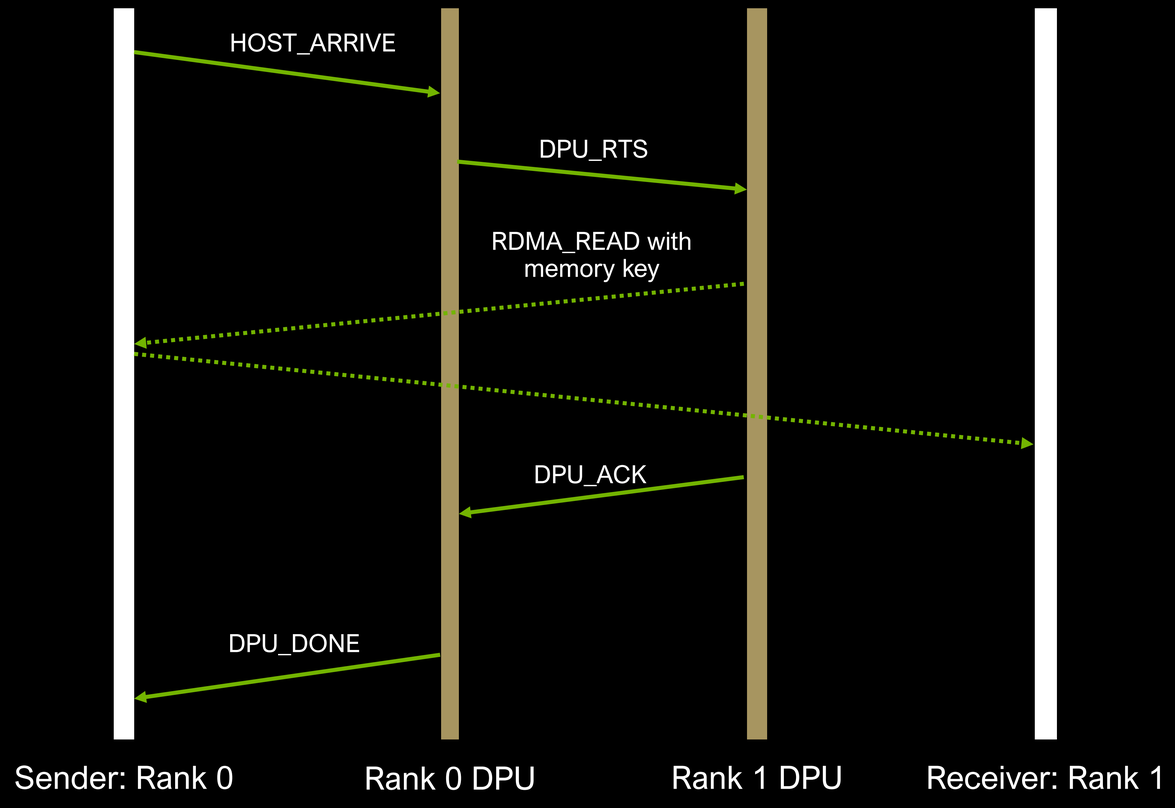
High Level System Components from Software's Perspective

- Host paired with local DPU
- Local DPU runs service processes (SP)
  - Each local user process (i.e. MPI process) has a service process
  - Each service process serves multiple local processes
  - Algorithm is split between host and DPU
- SP's may communicate with other hosts and/or SP's
- The DPU can initiates remote / local RDMA operations
- DPU memory is involved if the data originates from it

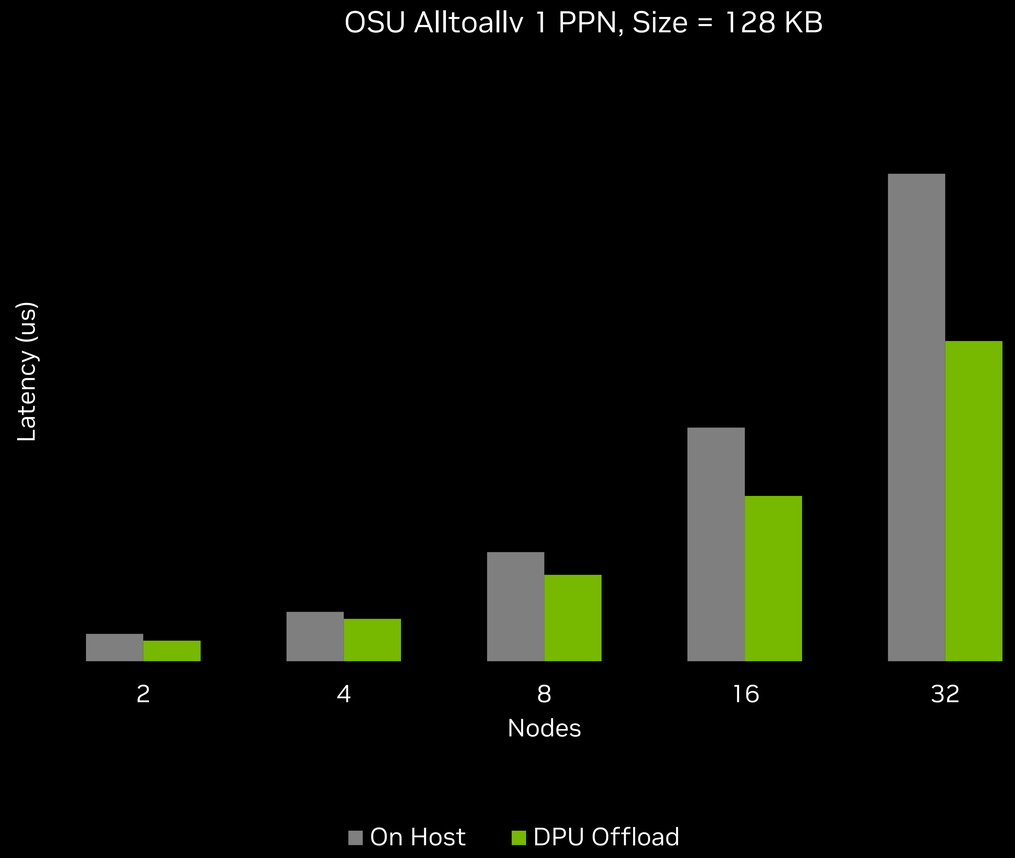


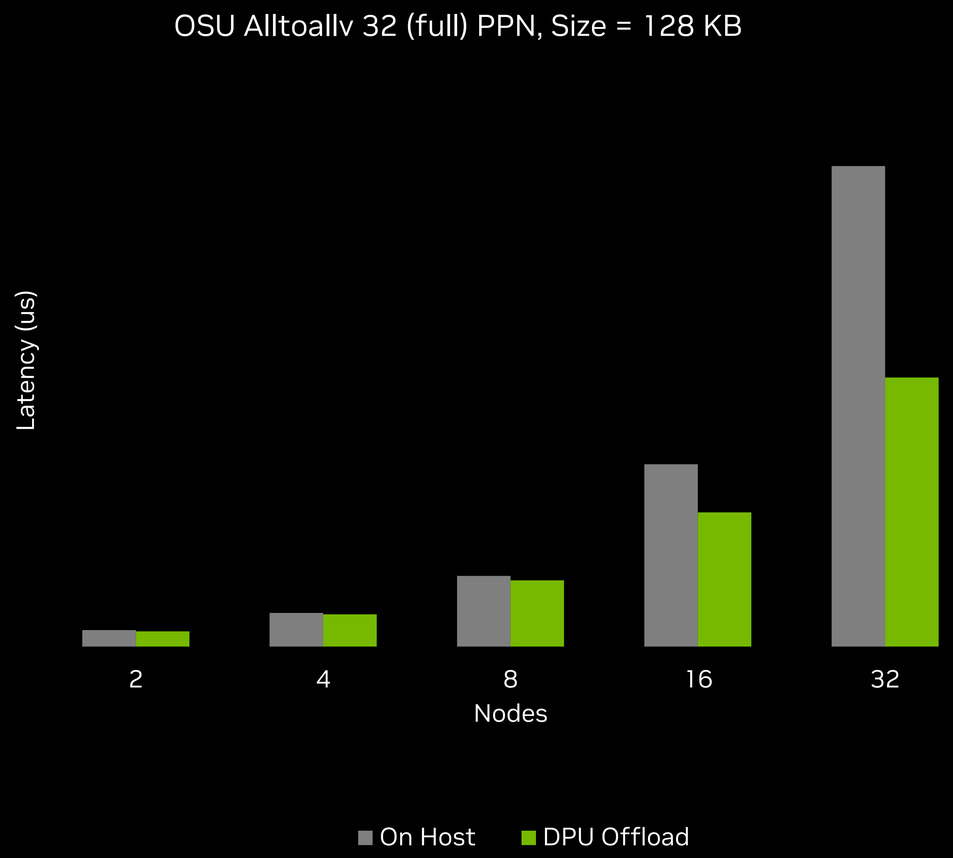
Offloading and Accelerating Data Exchange Example

An Element of Collective Algorithm

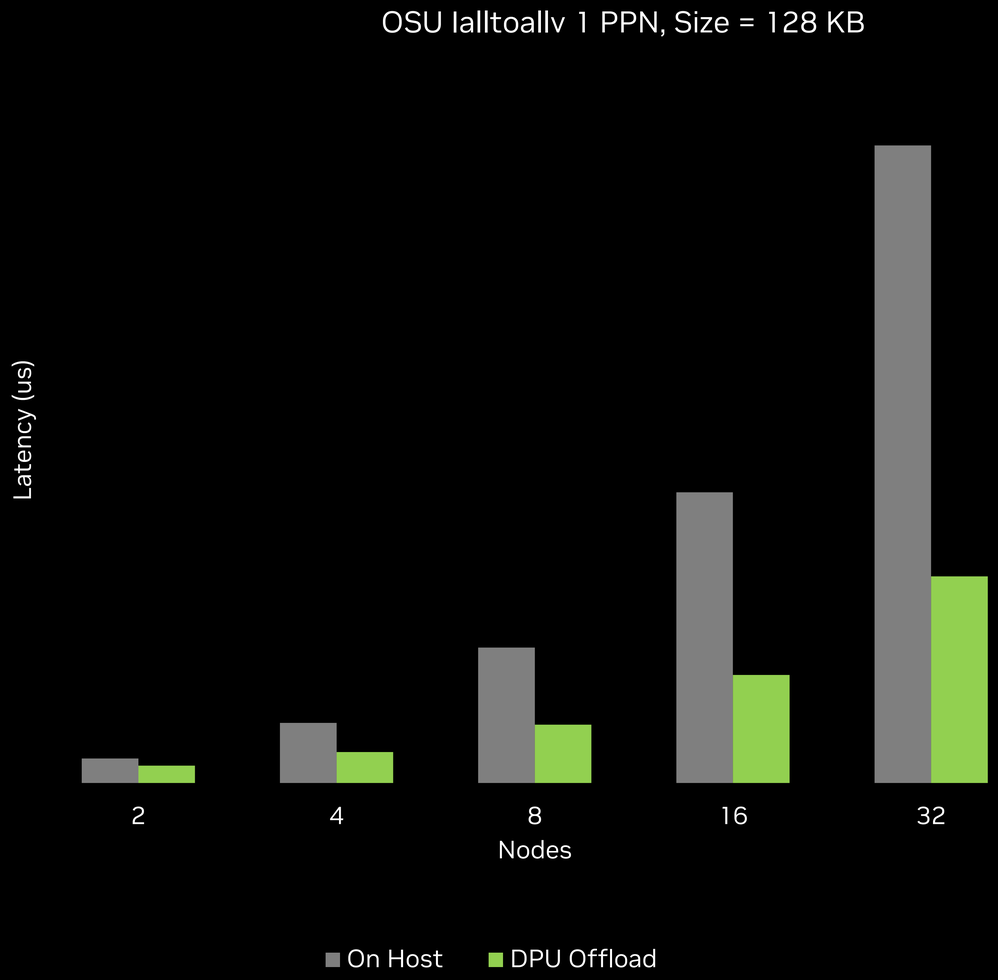


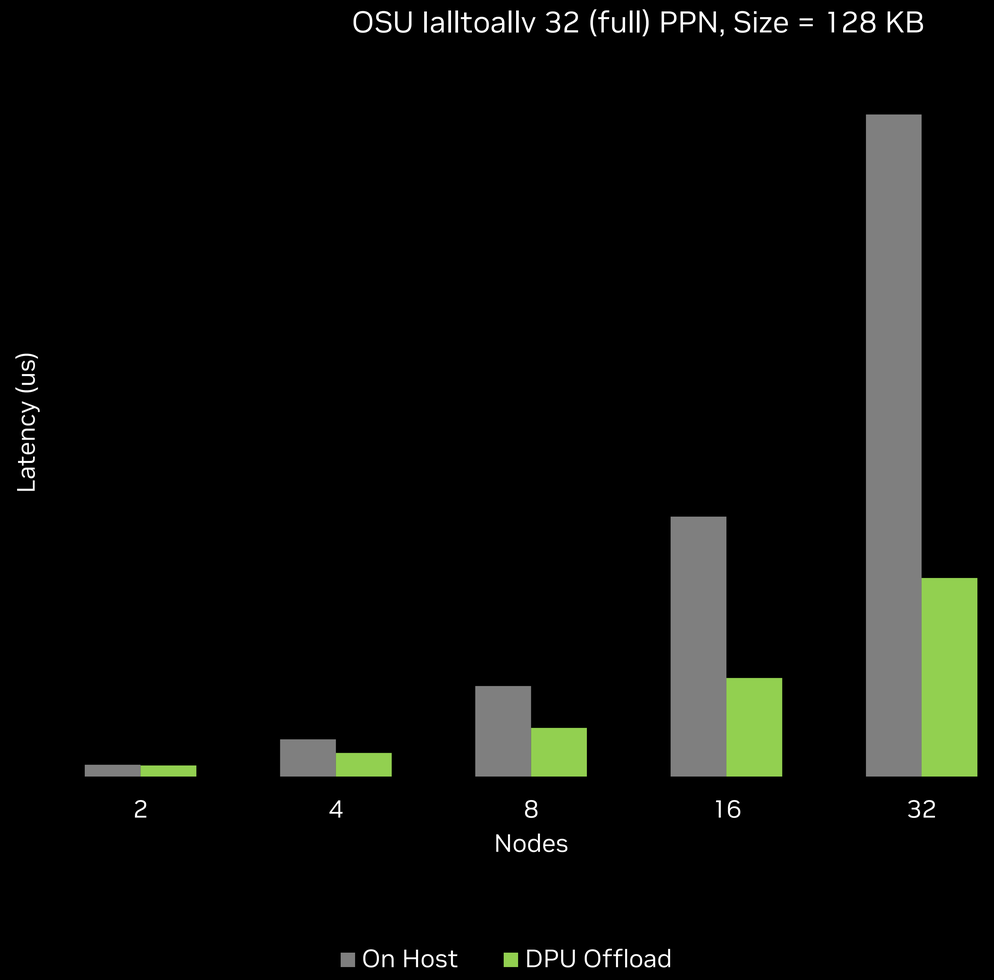
Alltoallv Latency



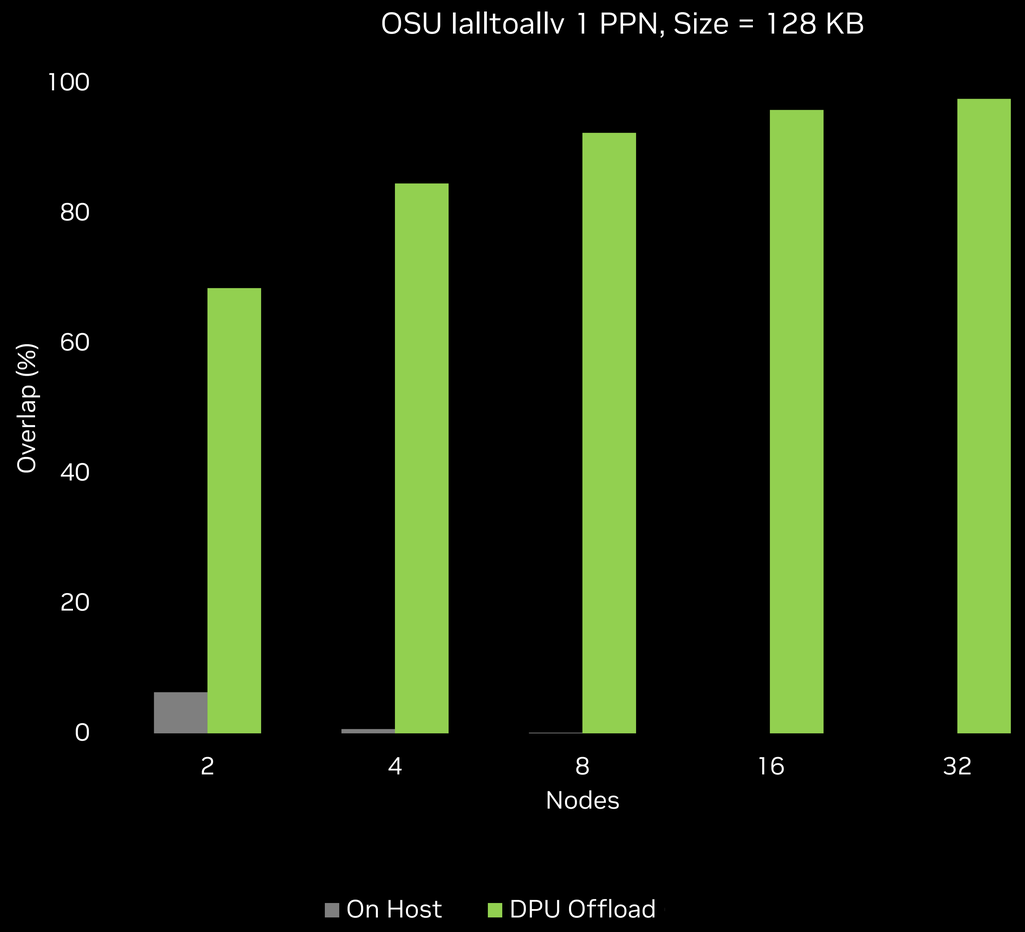


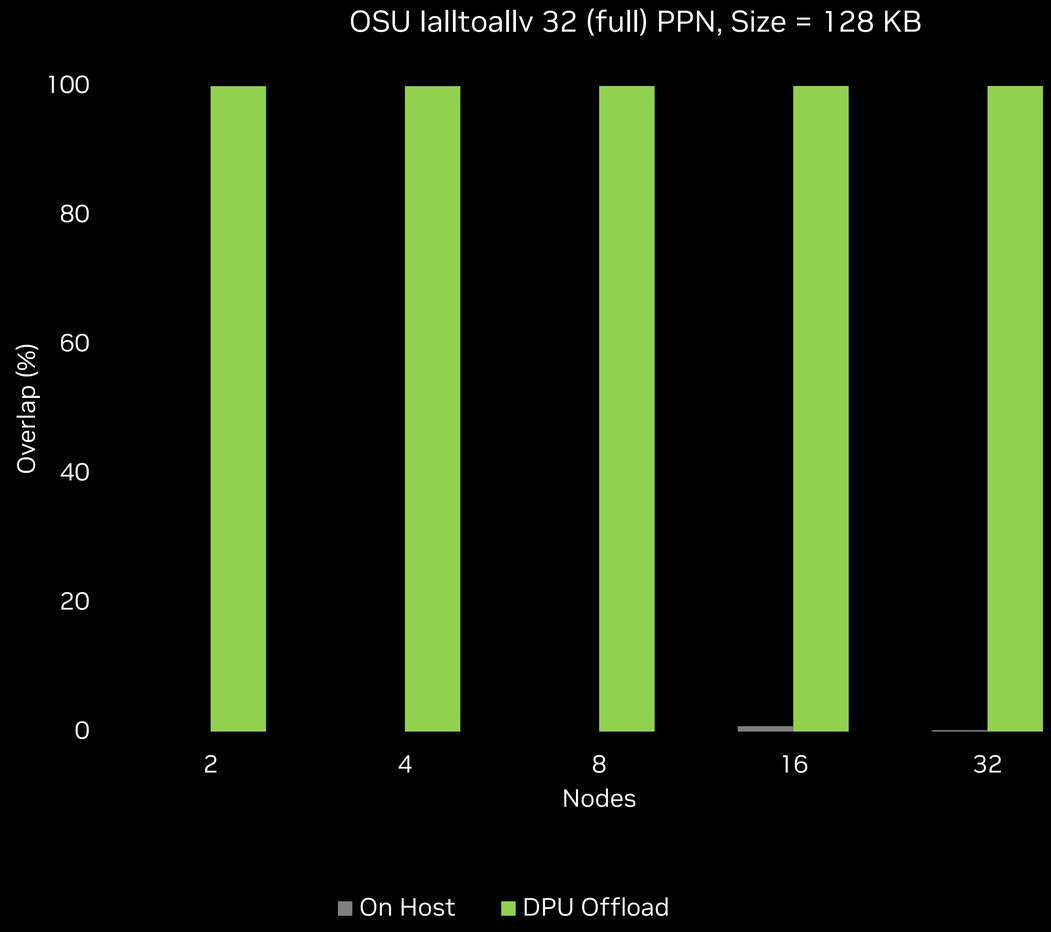
iAlltoallv latency



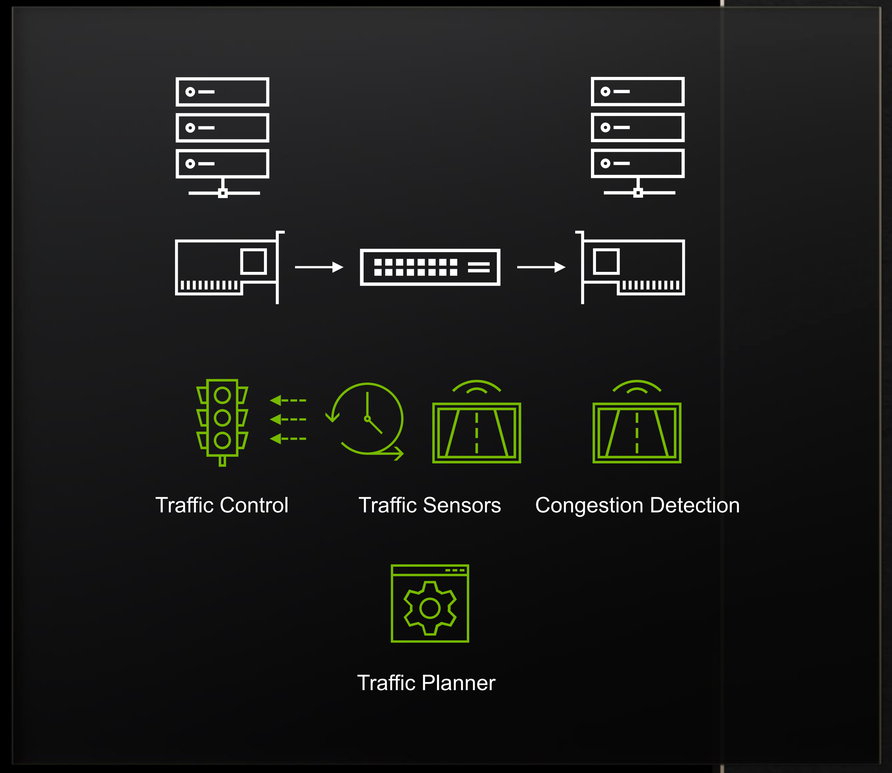


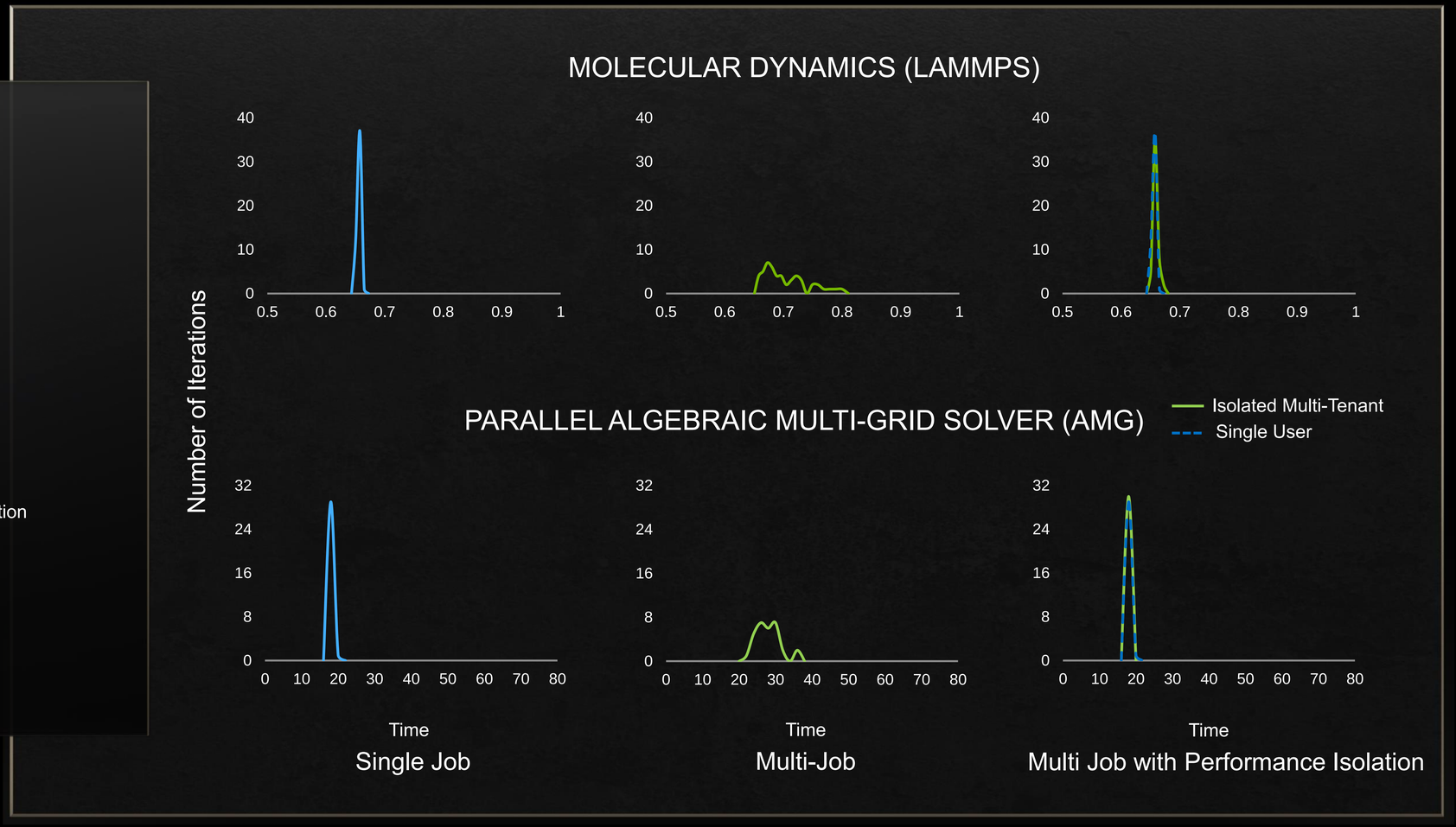
iAlltoallv compute/communication overlap





Performance Isolation





NVIDIA Full Stack HPC Platform

NVIDIA HPC | NVIDIA AI | NVIDIA Omniverse

GPU

In-Network Computing

CPU

Computational Storage

Performance Isolation

BlueField DPU

Enhanced Telemetry

Zero Trust Security


NVIDIA