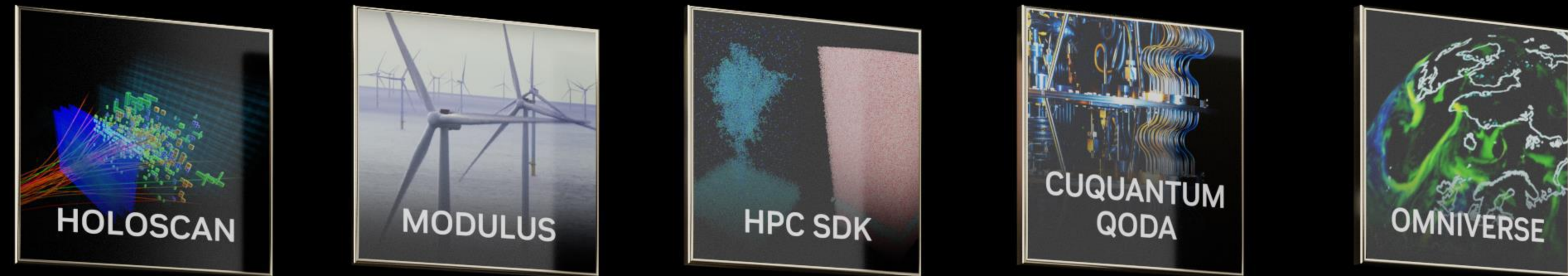




NVIDIA HPC Networking Platform

May 2023

NVIDIA Full Stack HPC Platform



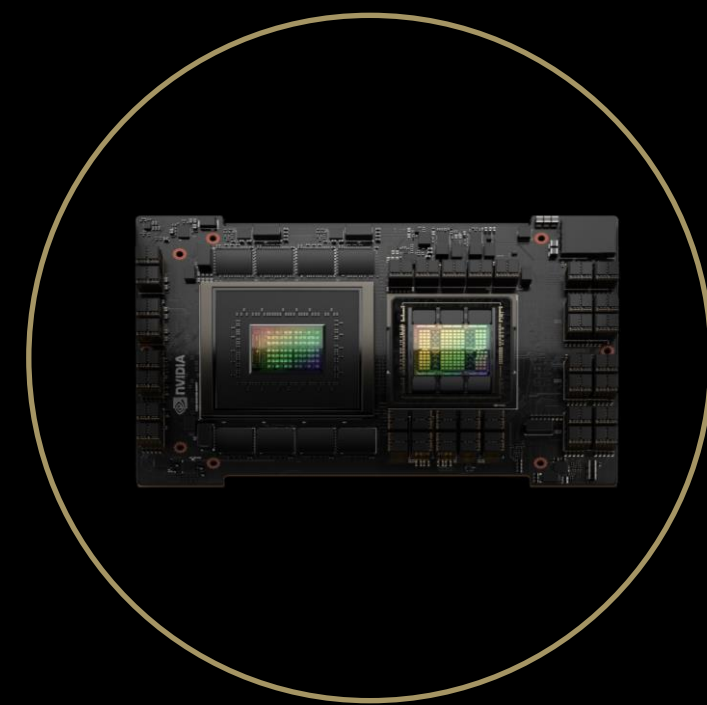
NVIDIA HPC



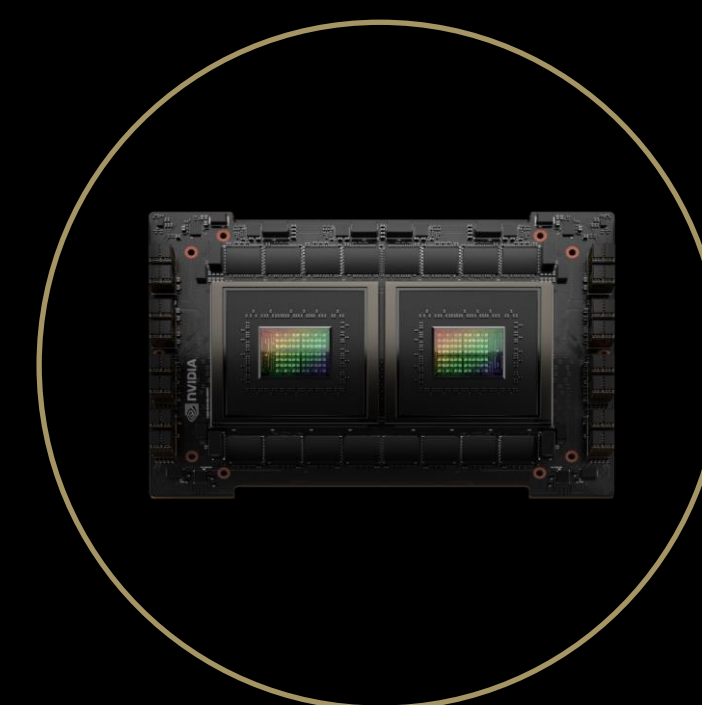
NVIDIA AI



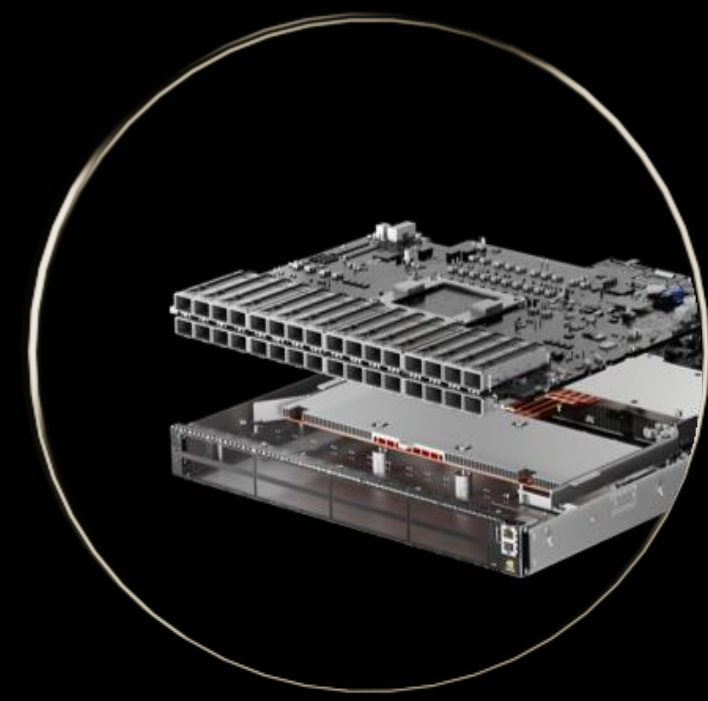
NVIDIA Omniverse



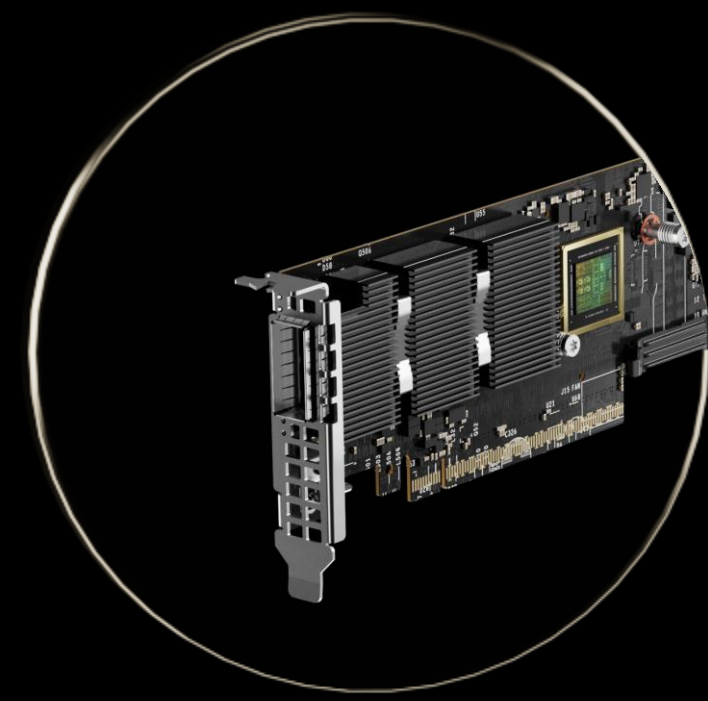
Grace Hopper



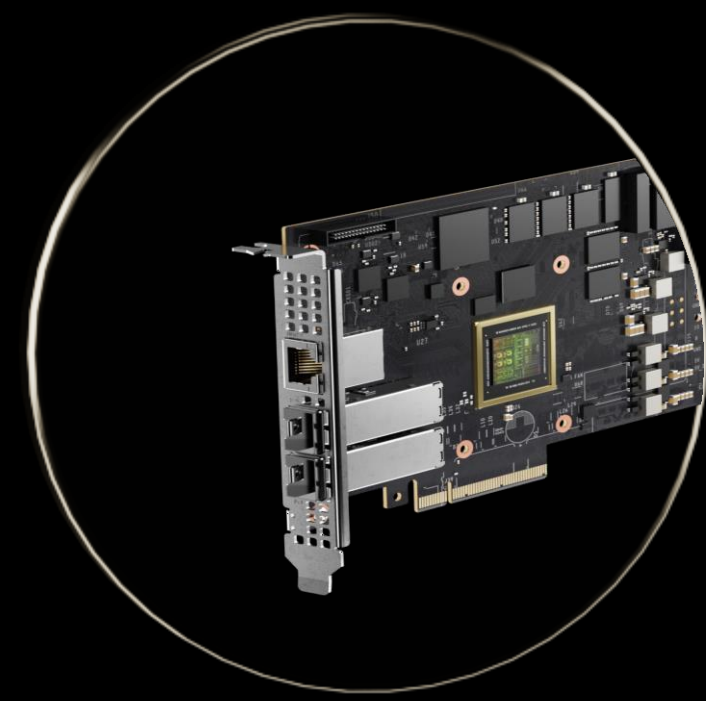
Grace Superchip



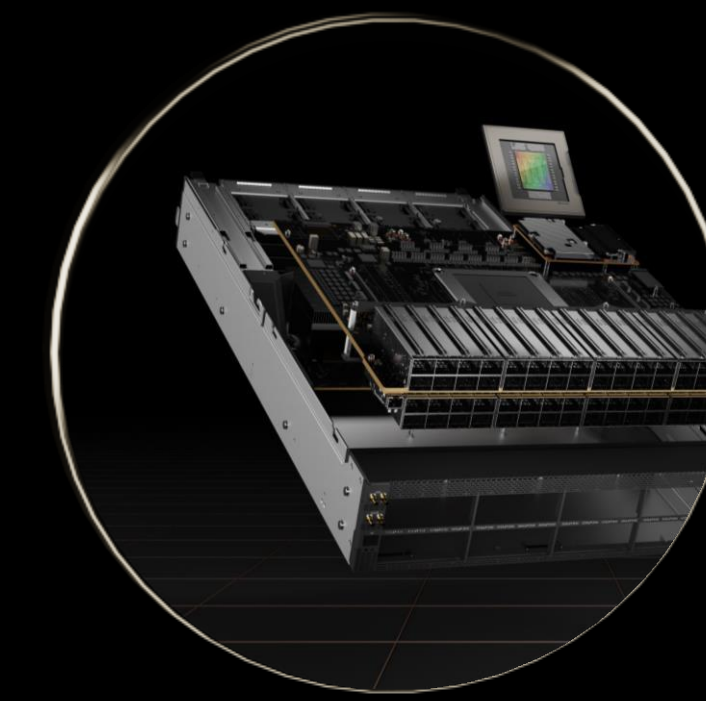
QUANTUM-2 INFINIBAND SWITCH



CONNECTX-7 SMARTNIC



BLUEFIELD-3 DPU



SPECTRUM-4 ETHERNET SWITCH



MANAGEMENT

In-Network Computing Accelerated Supercomputing

Software-Defined, Hardware-Accelerated, InfiniBand Network

Networking

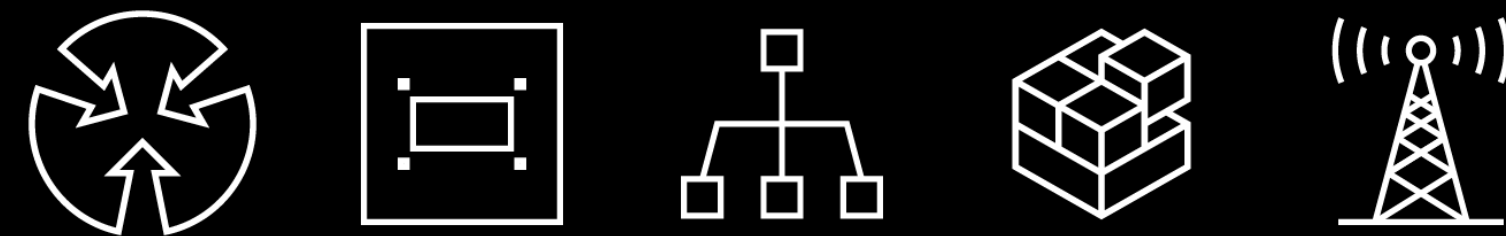
End-to-End	High Throughput	Extremely Low Latency	High Message Rate
	RDMA	GPUDirect RDMA	GPUDirect Storage
	Adaptive Routing	Congestion Control	Smart Topologies

In-Network Computing

Adapter/DPU	All-to-All	MPI Tag Matching	Data Reductions (SHARP)	Switch
	Programmable Datapath Accelerator	Data processing units (Arm cores)	Self Healing Network	
End-to-End	Security / Isolation			End-to-End

BlueField Data Processing Unit

SOFTWARE DEFINED NETWORKING



SOFTWARE DEFINED SECURITY



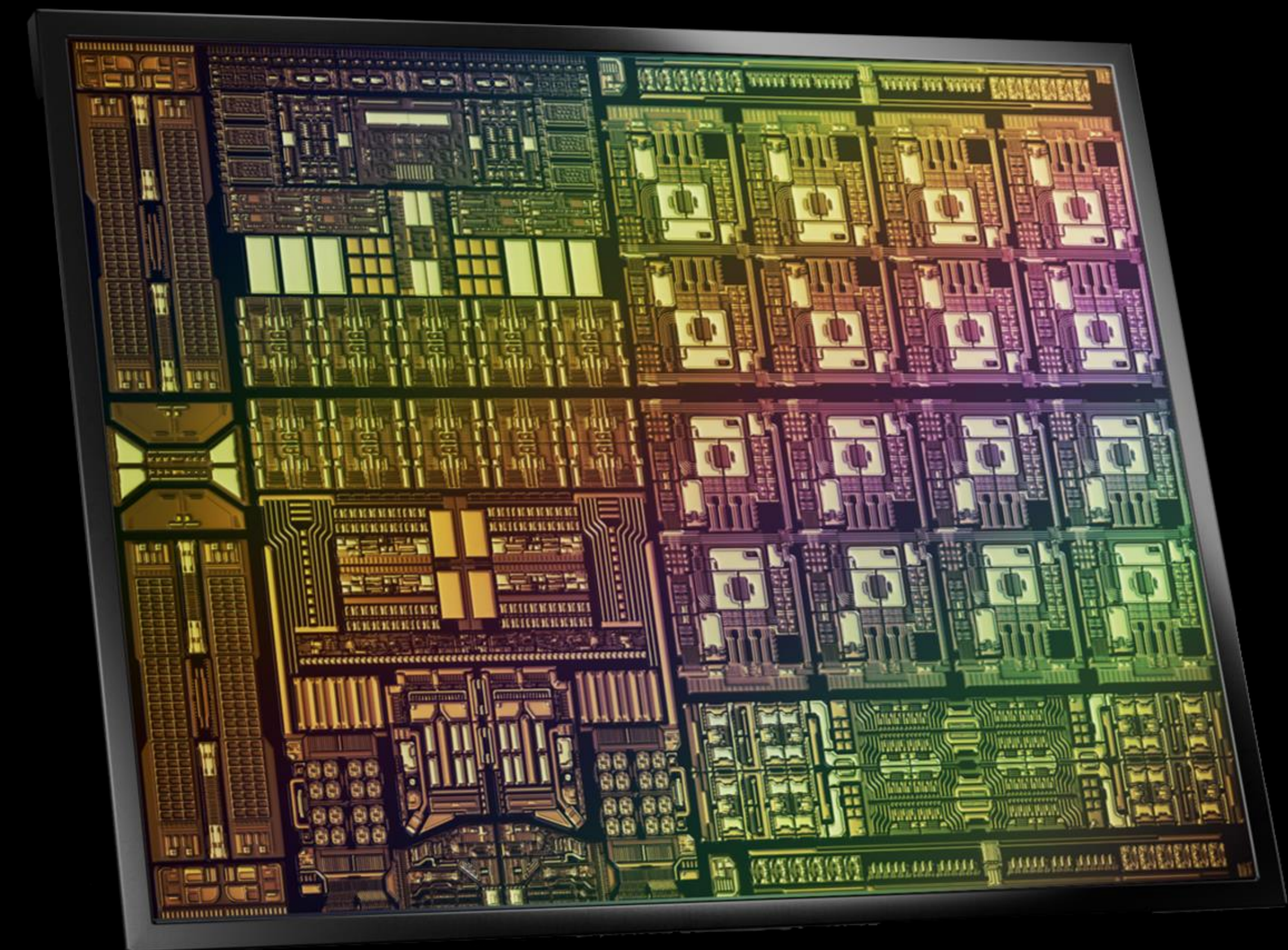
SOFTWARE DEFINED STORAGE



Infrastructure Services



BlueField Infrastructure Compute Platform



Data Center on a Chip

16 Arm 64-Bit Cores

16 Core / 256 Threads Datapath Accelerator

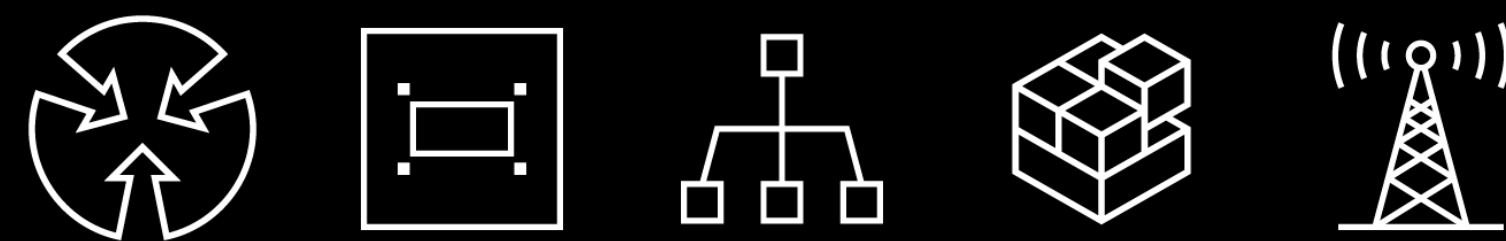
ConnectX InfiniBand / Ethernet

DDR memory interface

PCIe switch

The New Computing Platform for the Data Center Infrastructure

SOFTWARE DEFINED NETWORKING



SOFTWARE DEFINED SECURITY



SOFTWARE DEFINED STORAGE



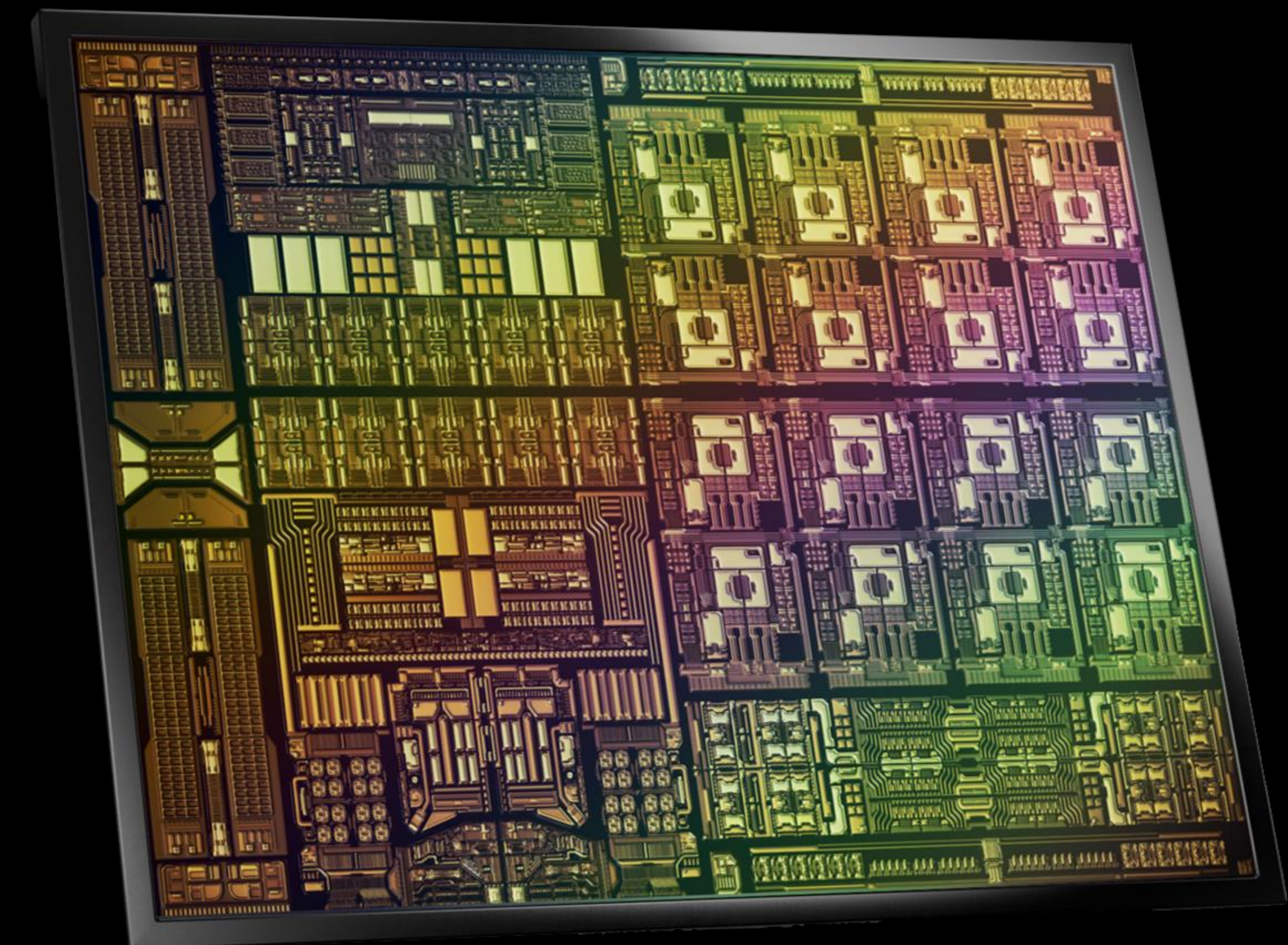
Infrastructure Services

	BlueField-2	BlueField-3
--	-------------	-------------

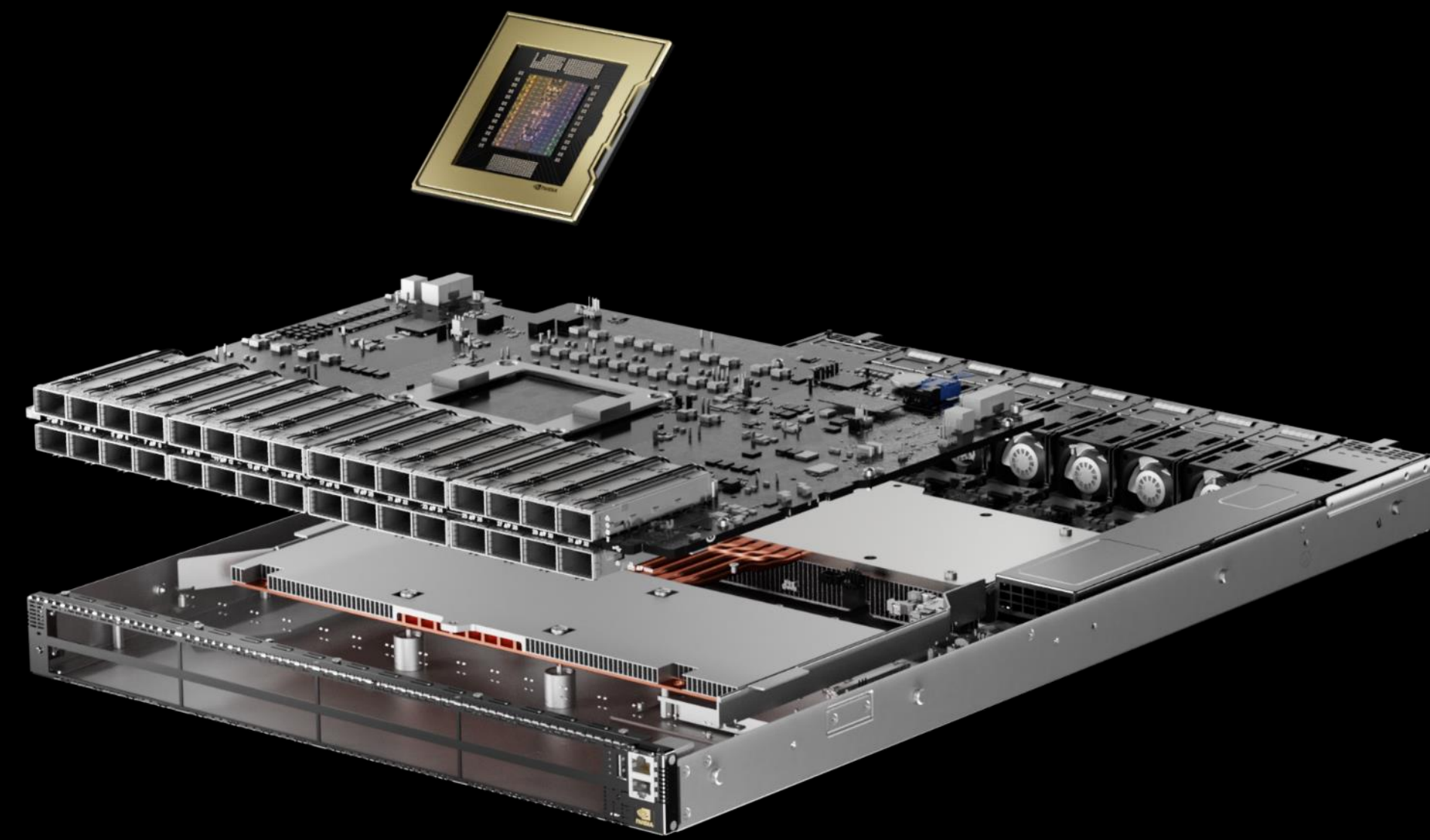
Network Bandwidth	200Gb/s	400Gb/s
RDMA msg rate	215Mpps	370Mpps
Compute	SPECINT2K17: 9.8	SPECINT2K17: 42
Memory Bandwidth	17GB/s	80GB/s



BlueField Infrastructure Compute Platform



NVIDIA Quantum-2 400G In-Network Computing



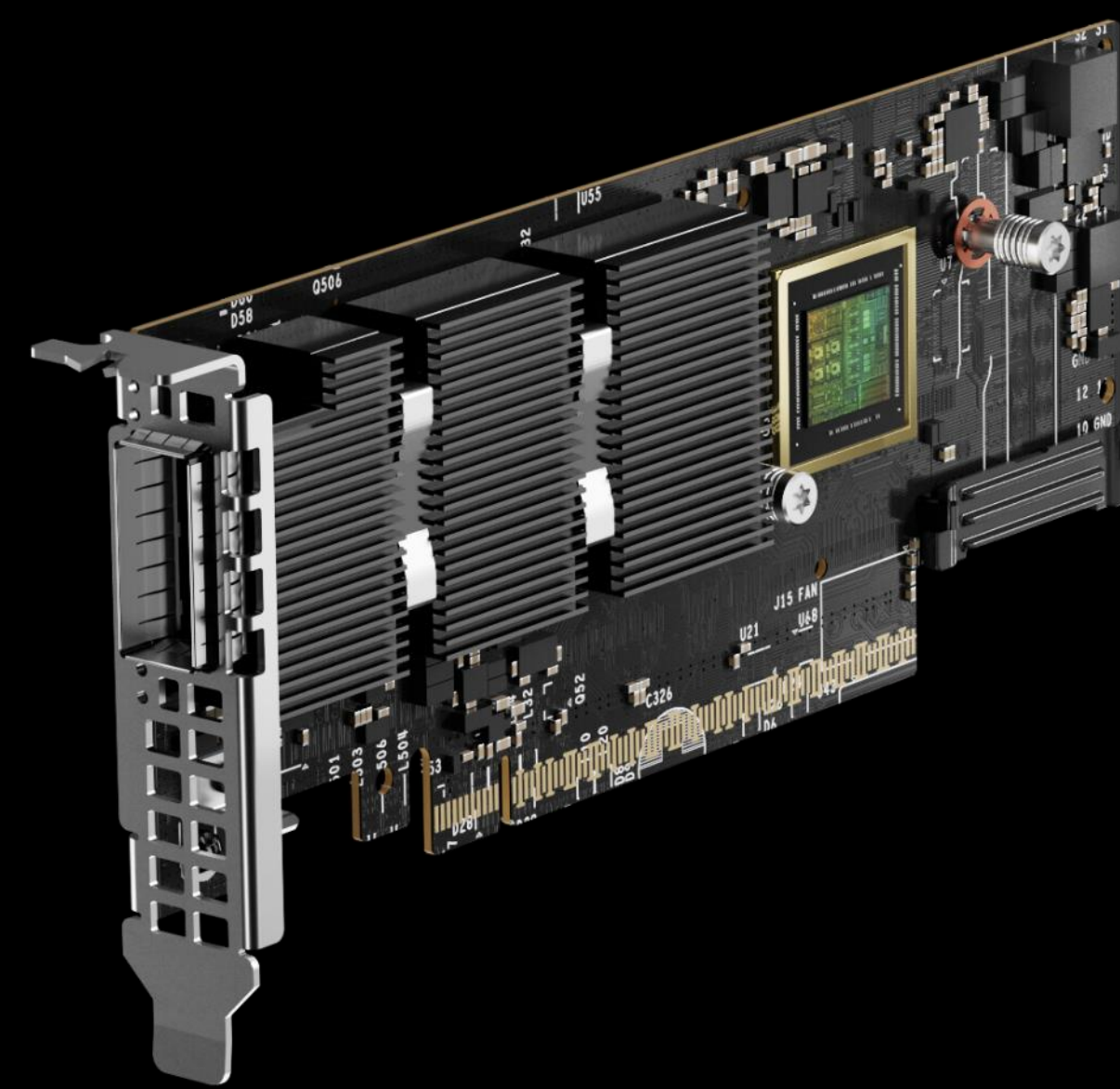
QUANTUM-2 SWITCH

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

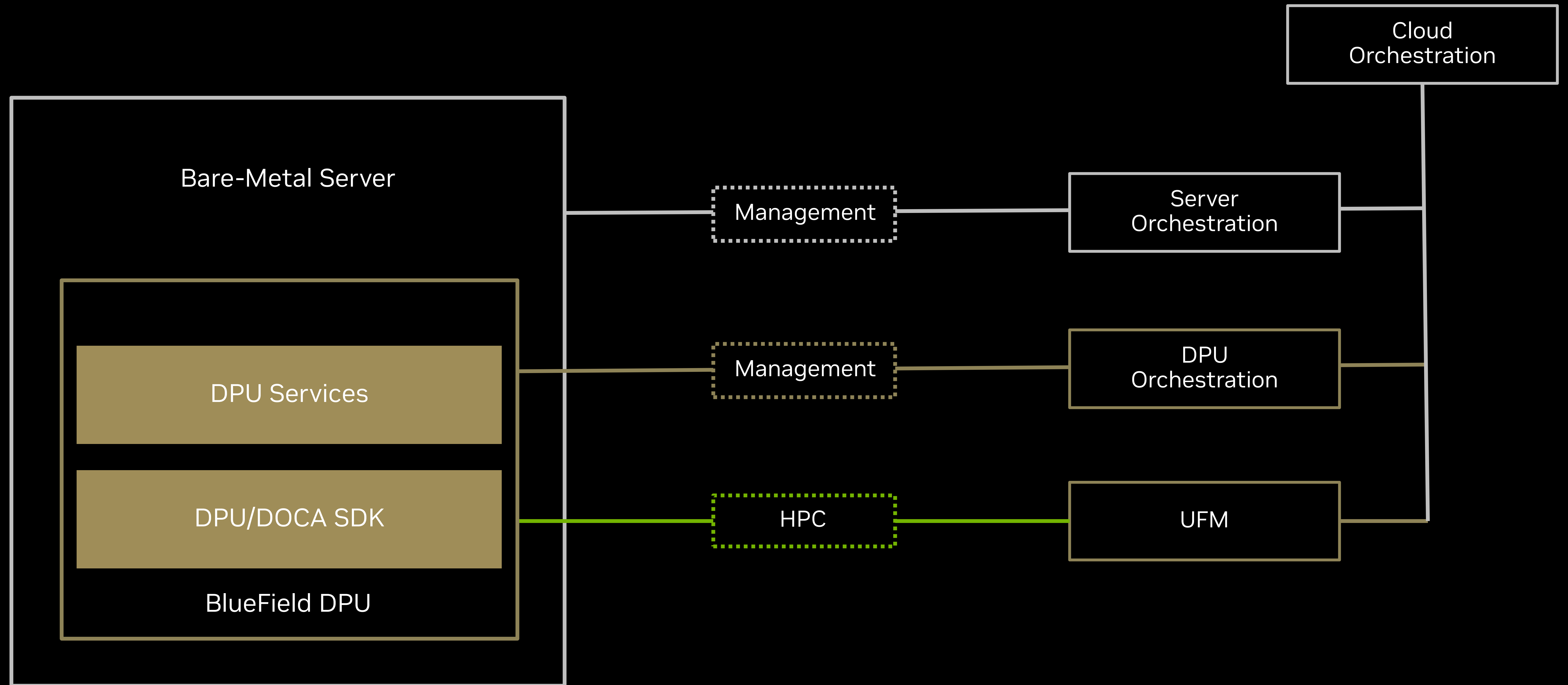
Hardware-Based RDMA / GPUDirect

MPI and NCCL Accelerations

Computational Storage

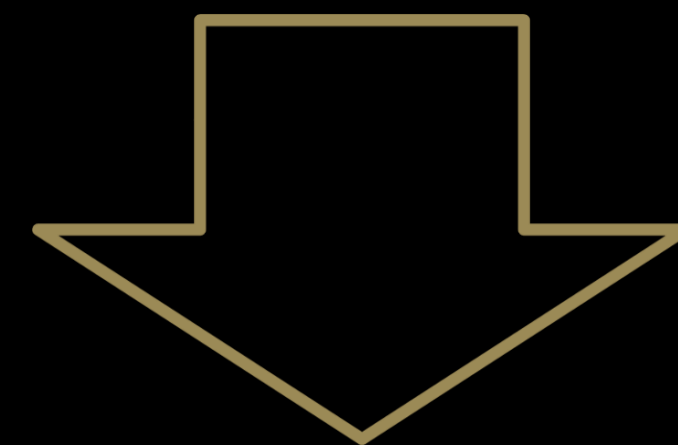
Security Engines

Delivering Cloud Native Supercomputing



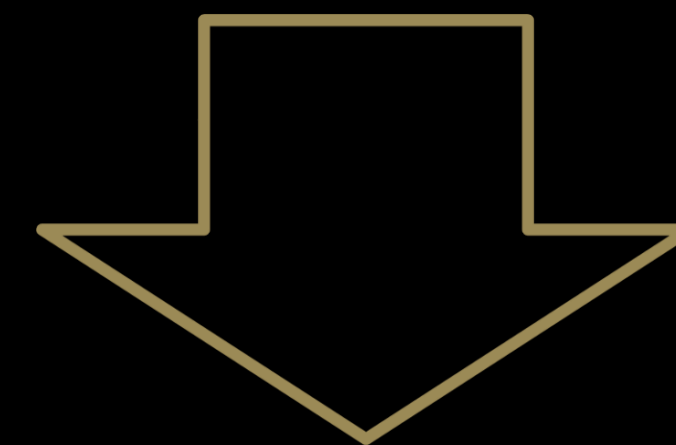
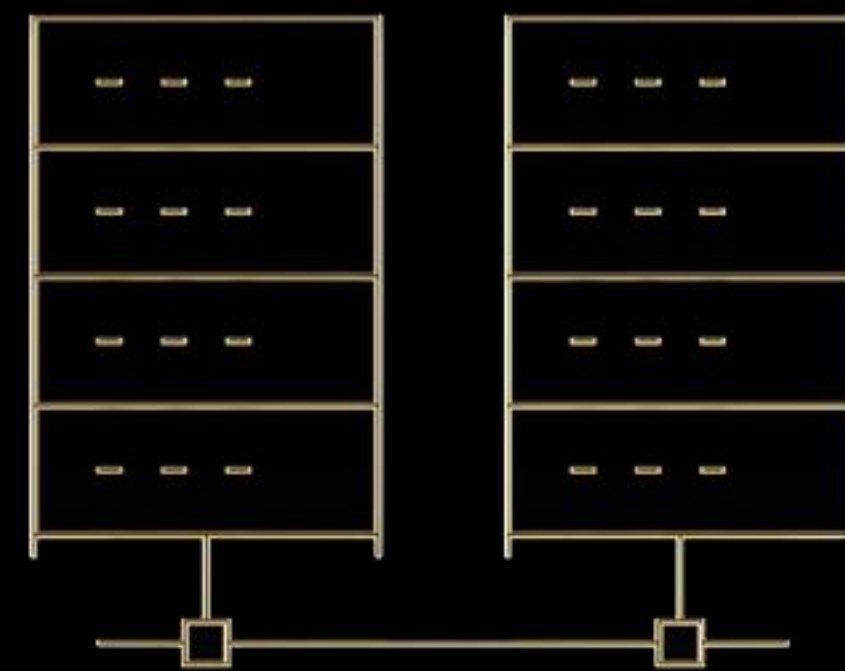
HPC Performance Bottlenecks

Overlapping



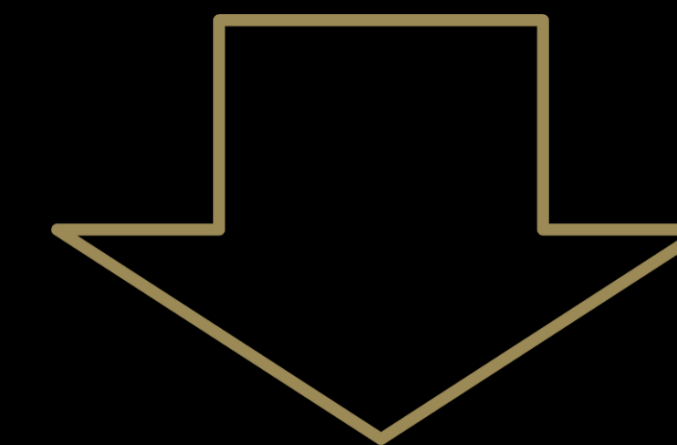
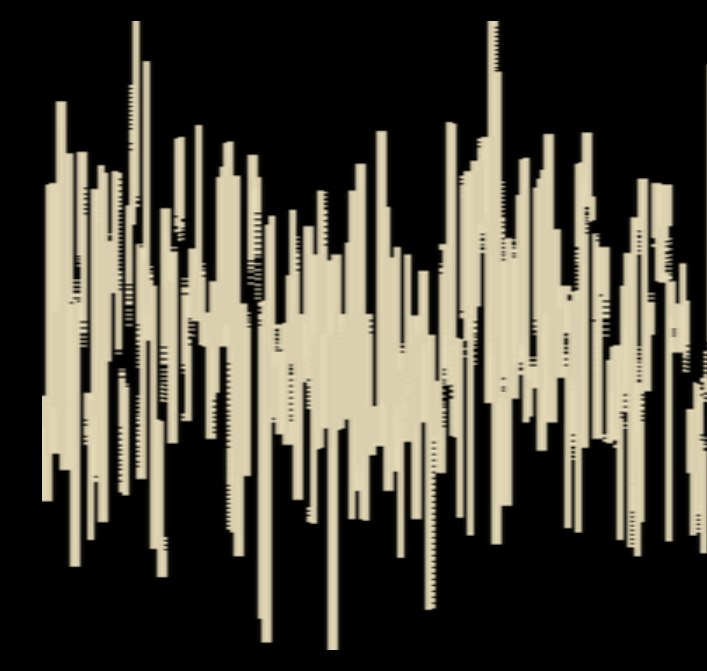
In-Network Computing
Asynchronous Progress
(Compute – Communication Overlap)

Load Imbalanced



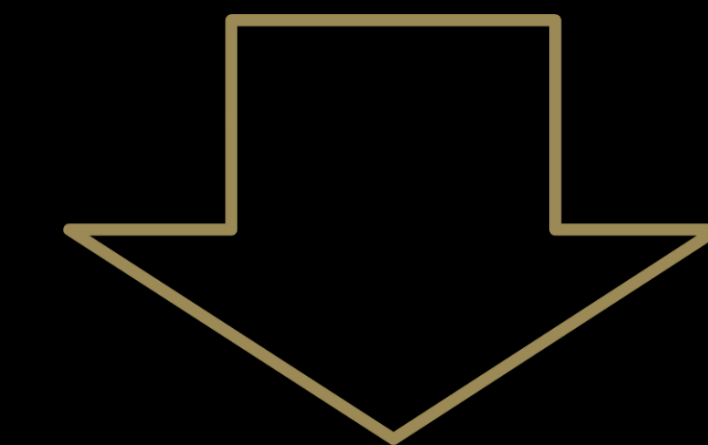
In-Network Computing
and DPU Synchronization

Jitter



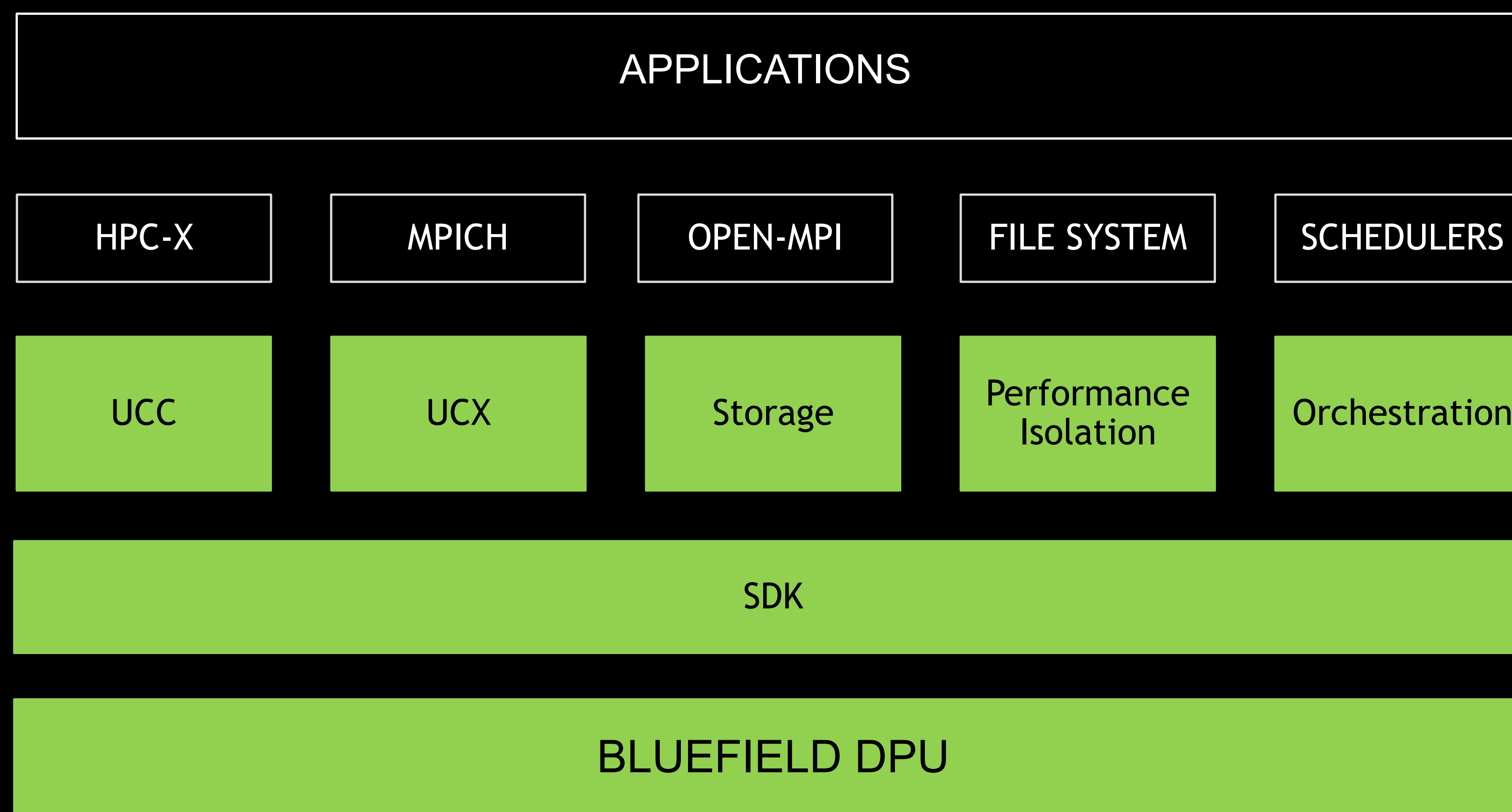
In-Network Computing
Infrastructure Processing

Multi-Job Performance



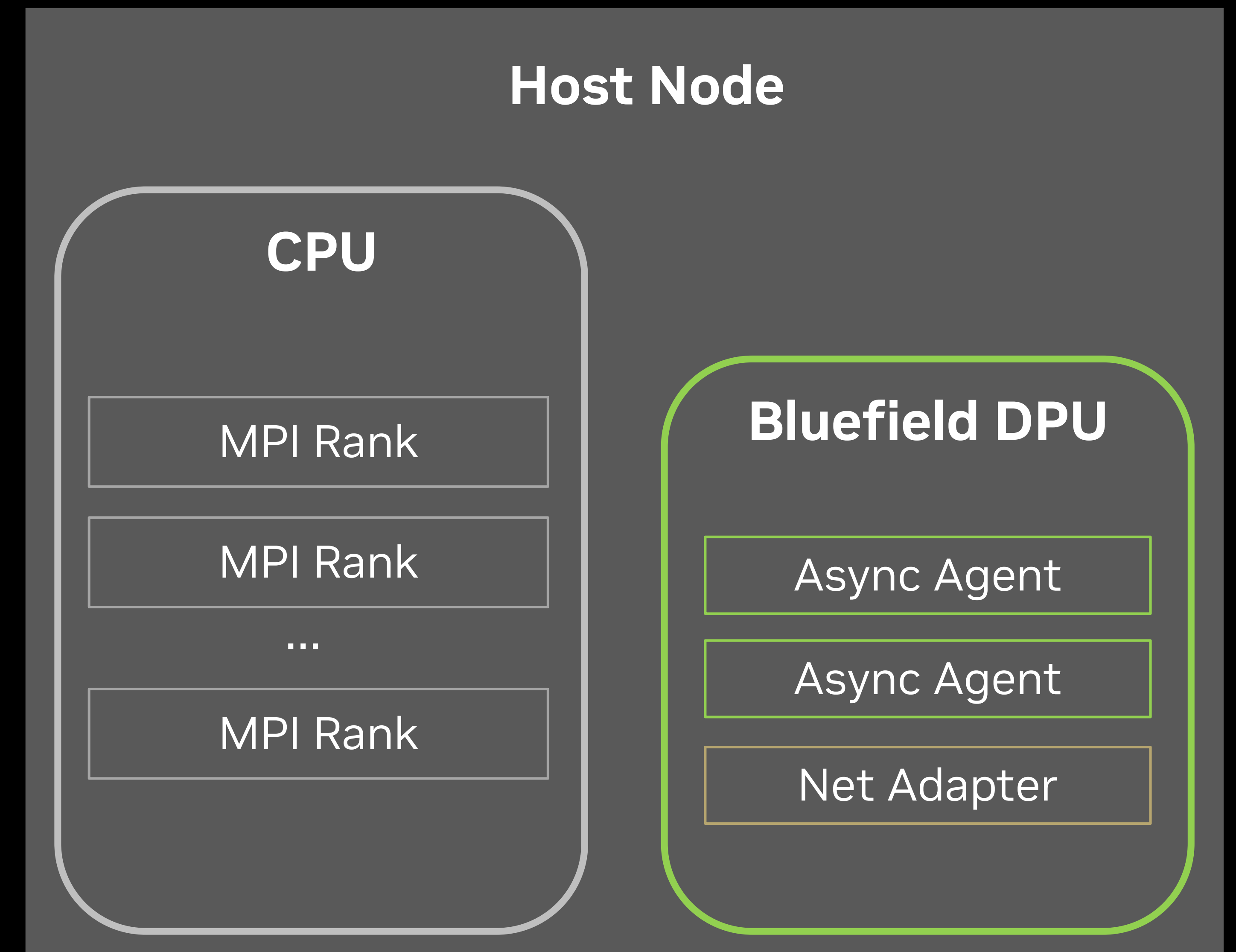
Adaptive Performance
Isolation

Accelerating HPC Applications with DPU/DOCA Services



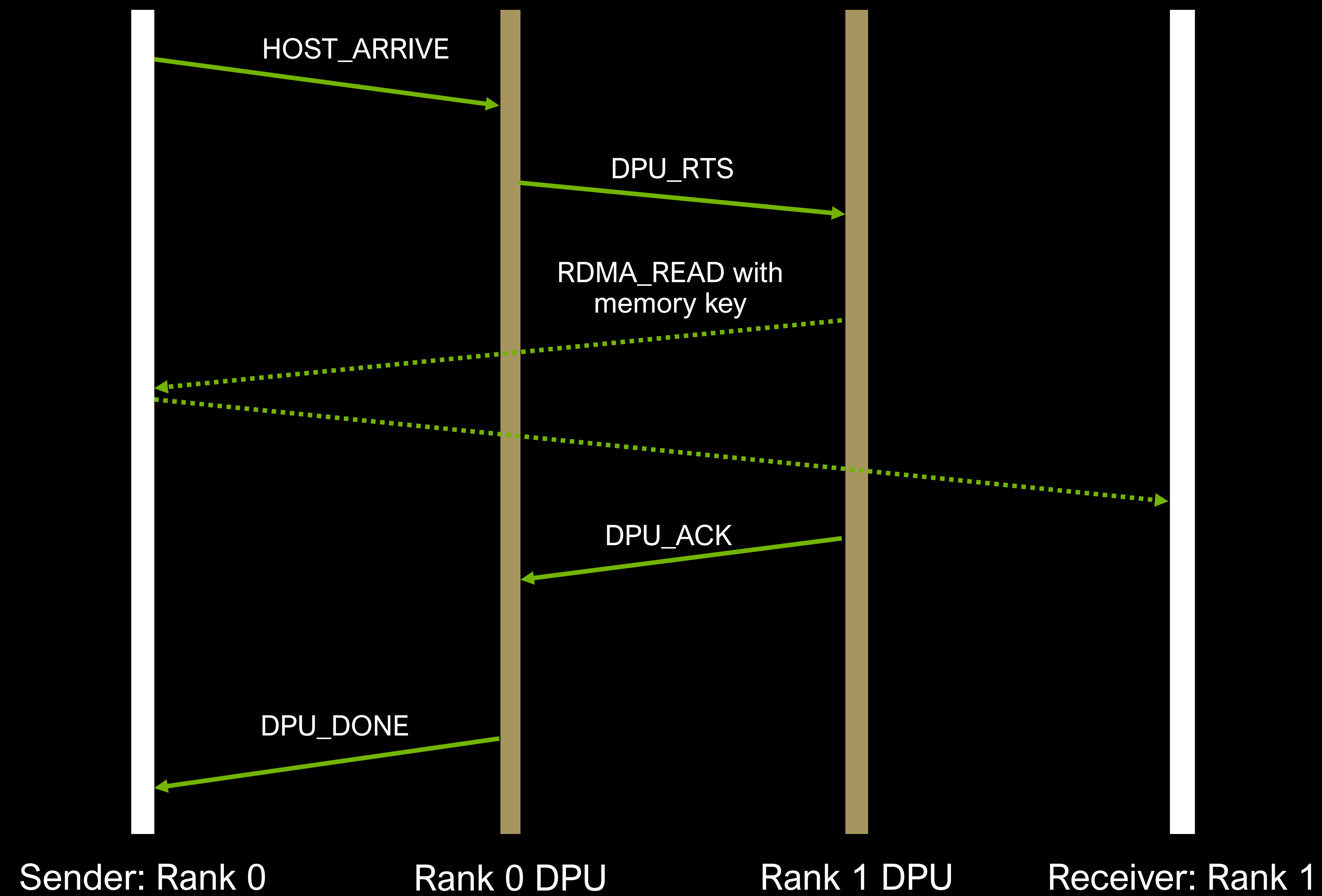
High Level System Components from Software's Perspective

- 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 initiate remote / local RDMA operations
- DPU memory is involved if the data originates from it



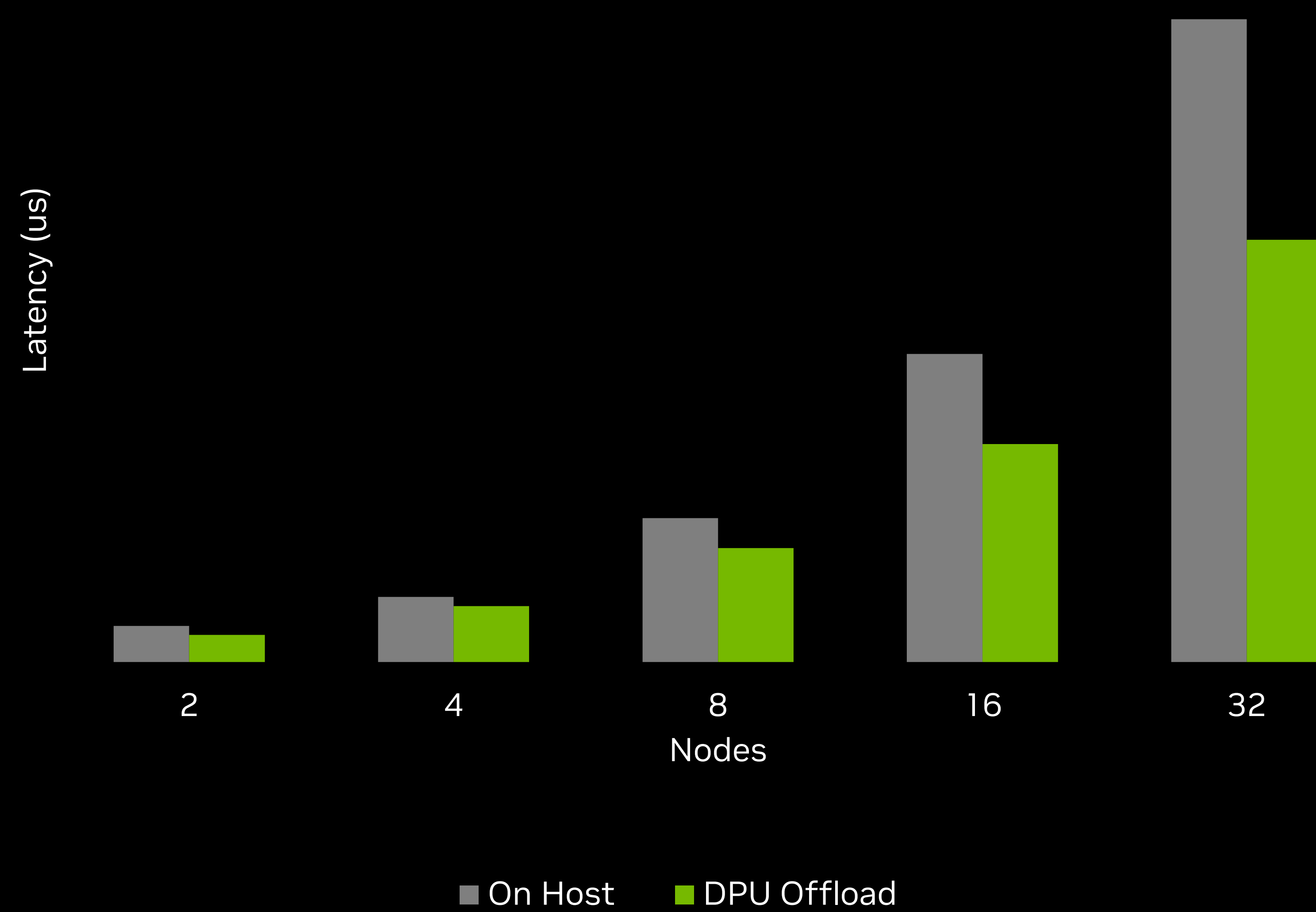
Offloading and Accelerating Data Exchange Example

An Element of Collective Algorithm

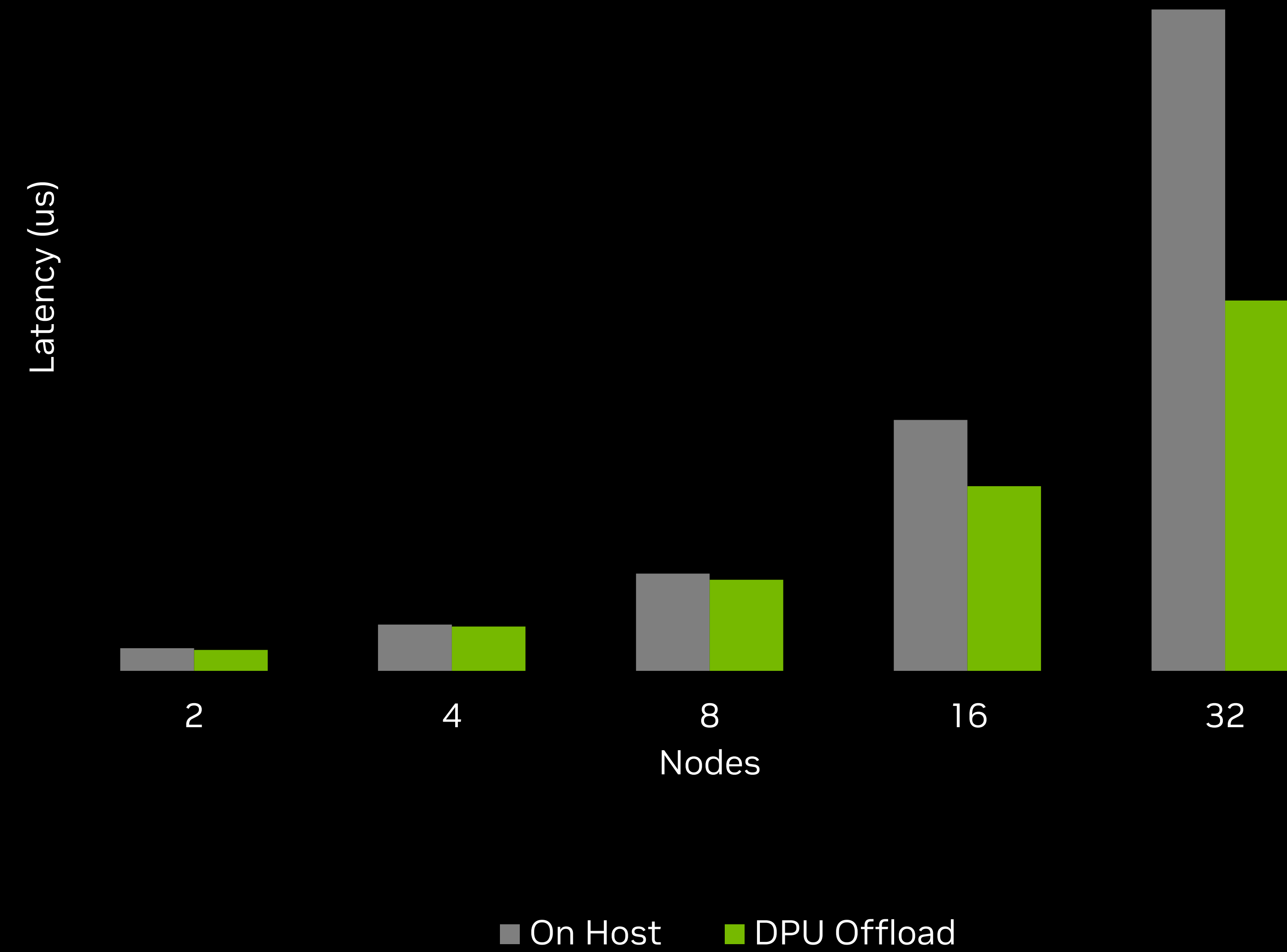


Alltoallv Latency

OSU Alltoallv 1 PPN, Size = 128 KB

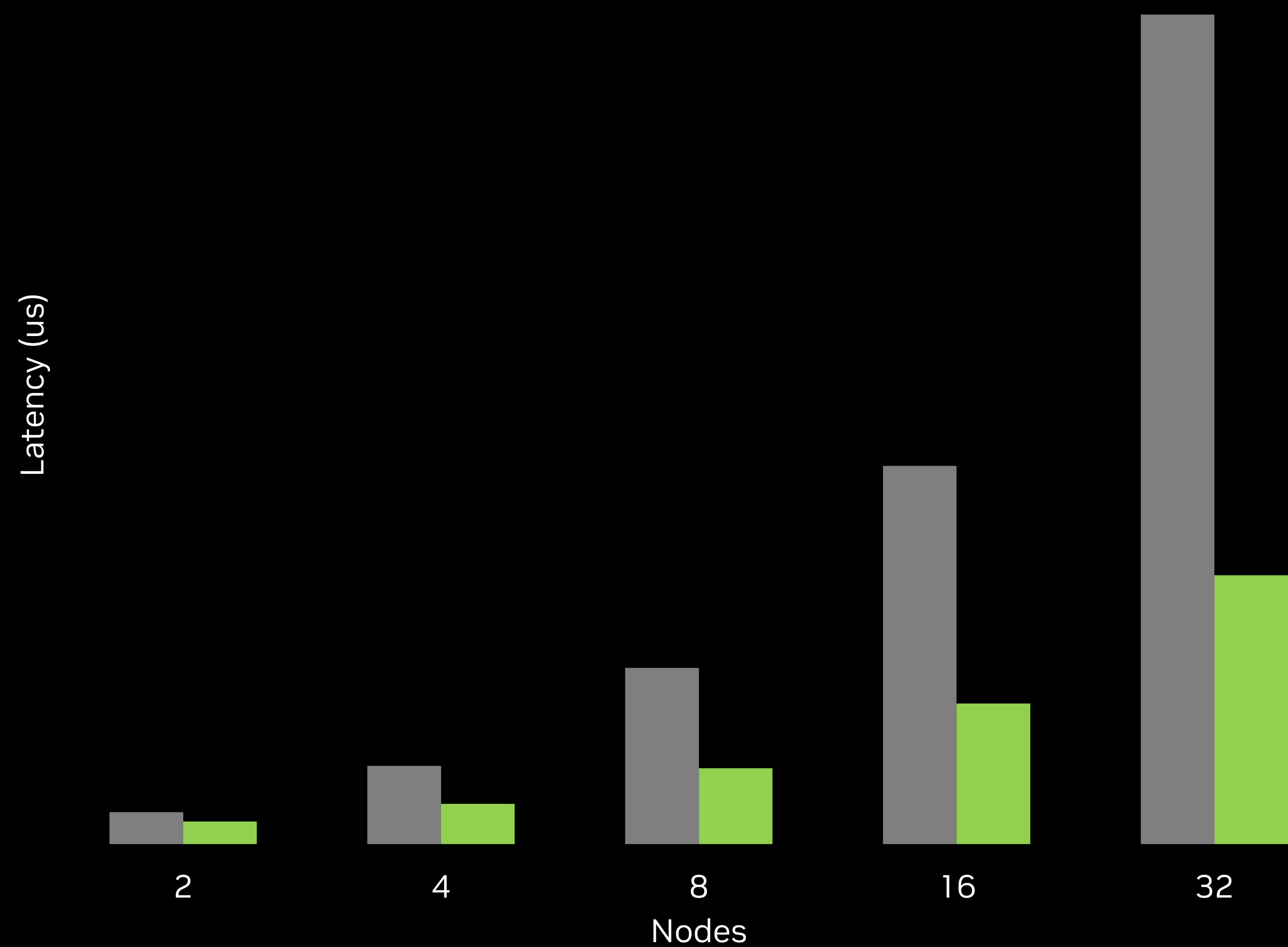


OSU Alltoallv 32 (full) PPN, Size = 128 KB

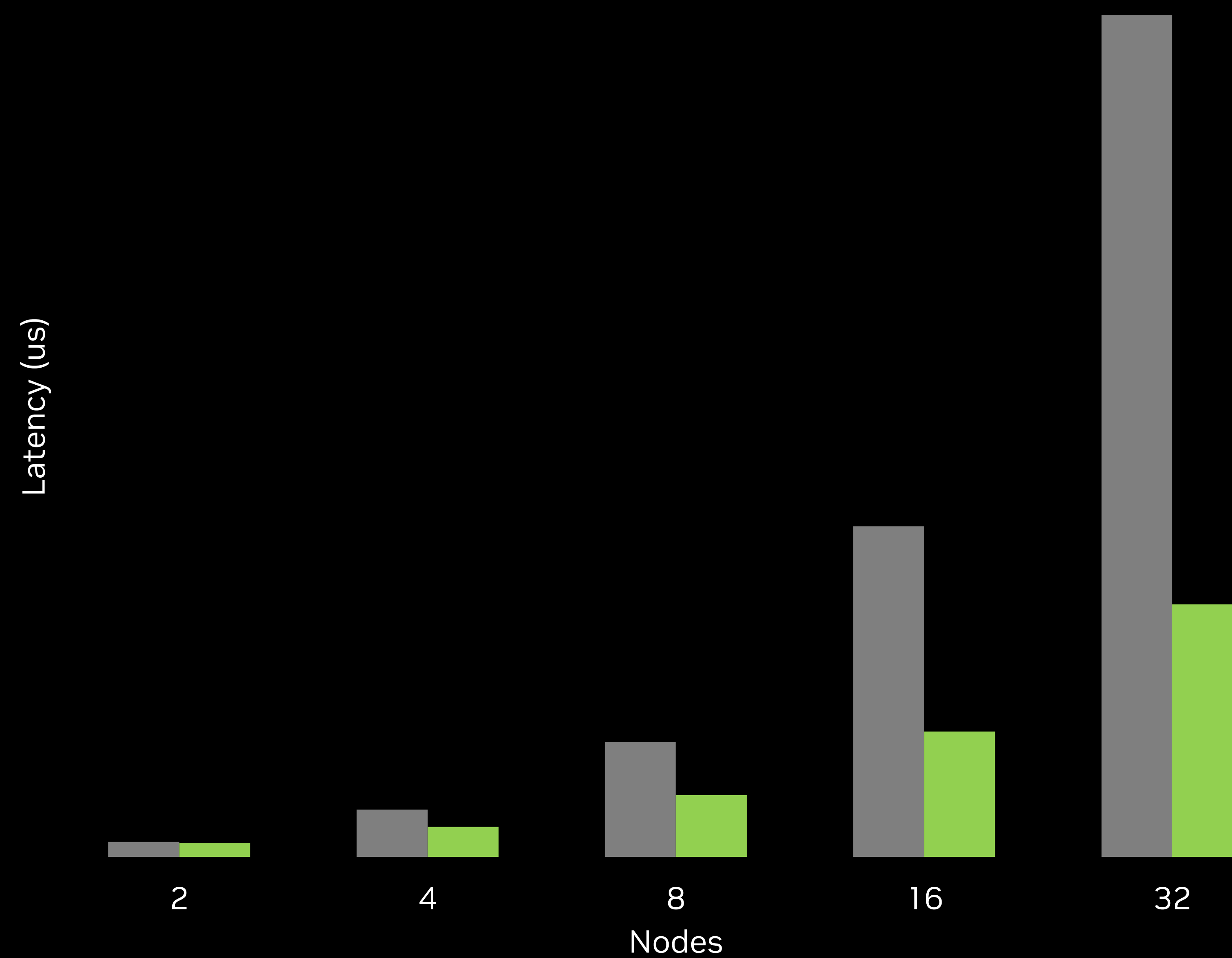


iAlltoallv latency

OSU ialltoallv 1 PPN, Size = 128 KB



OSU ialltoallv 32 (full) PPN, Size = 128 KB

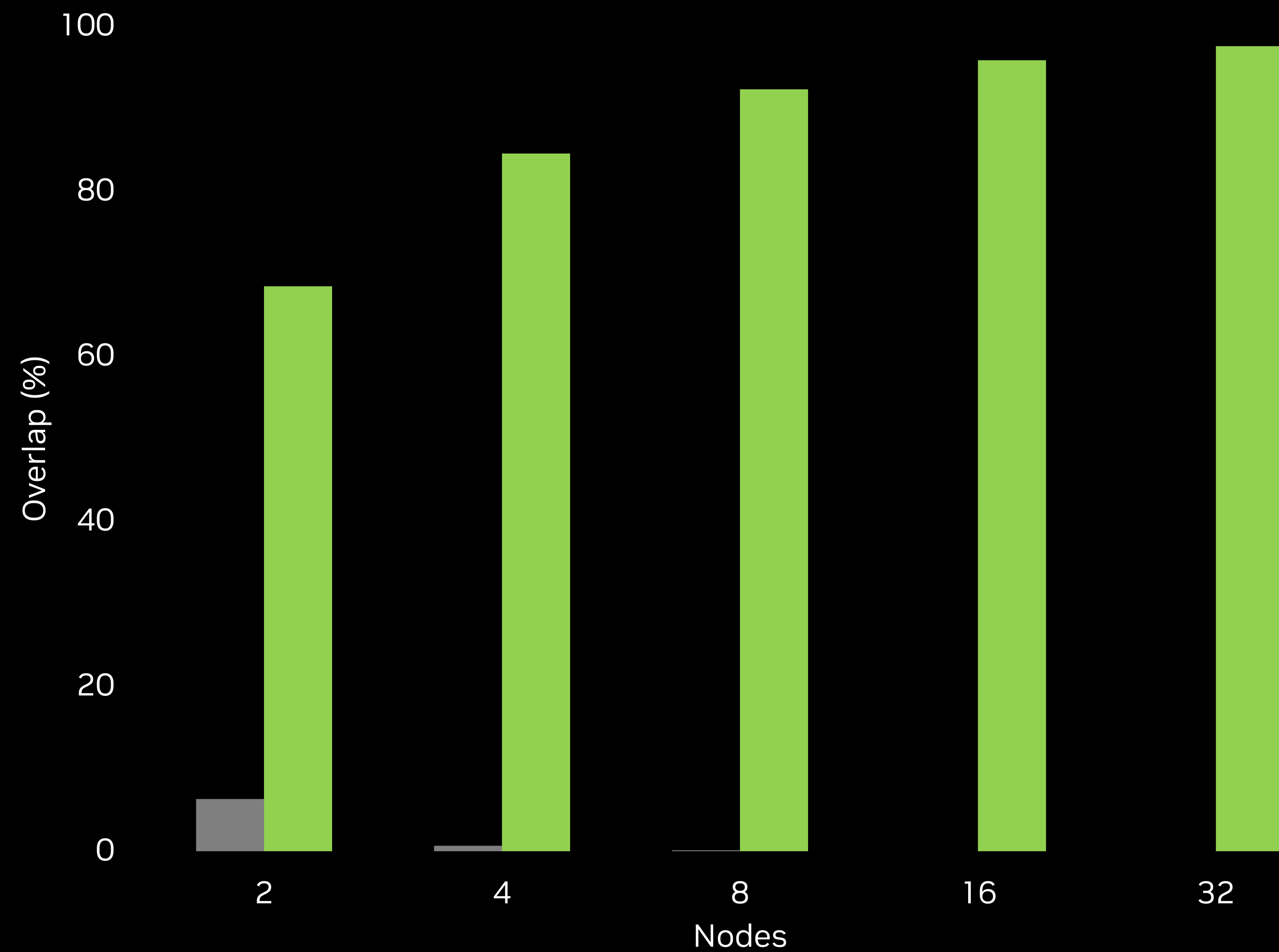


■ On Host ■ DPU Offload

■ On Host ■ DPU Offload

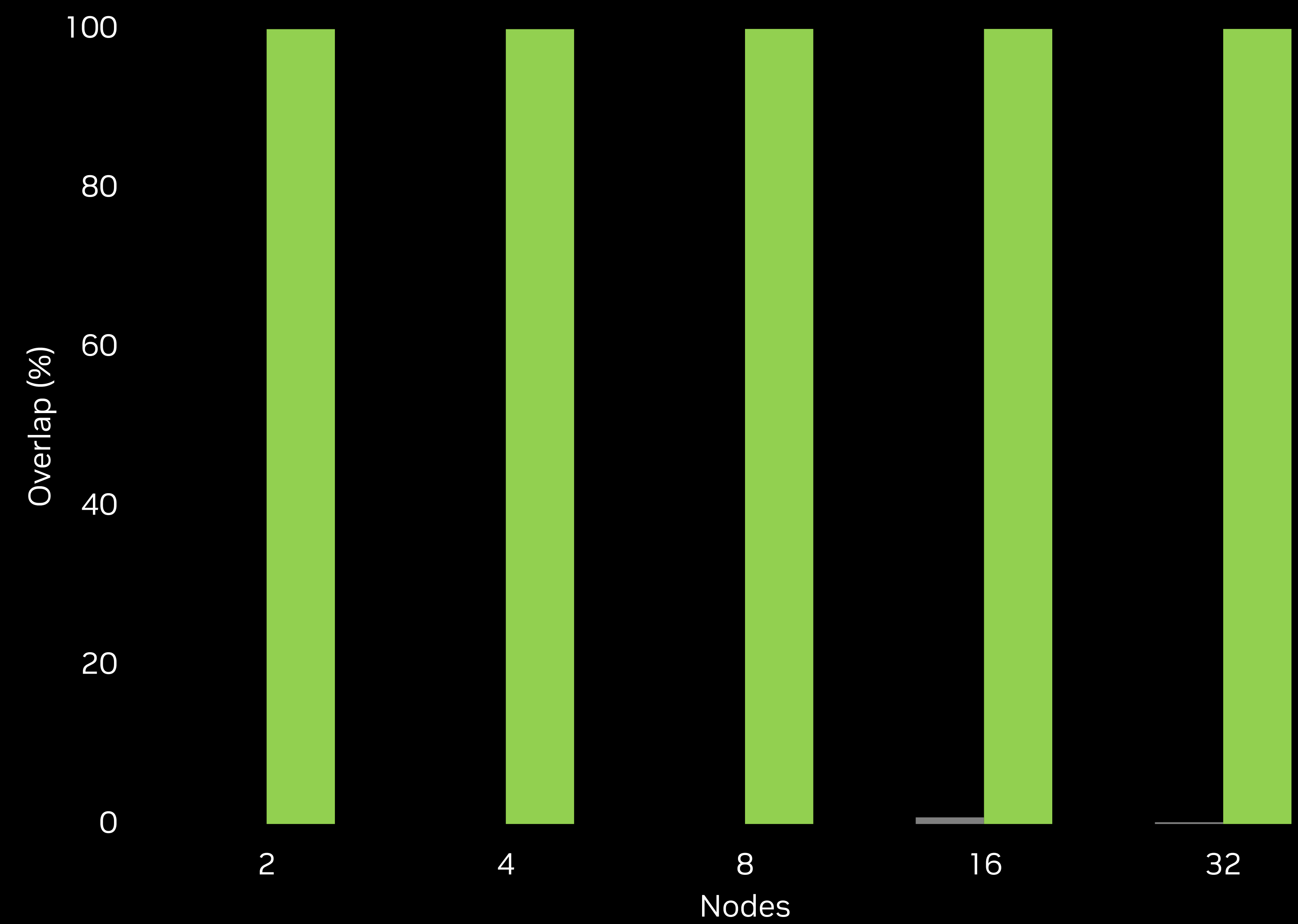
iAlltoallv compute/communication overlap

OSU Ialltoallv 1 PPN, Size = 128 KB



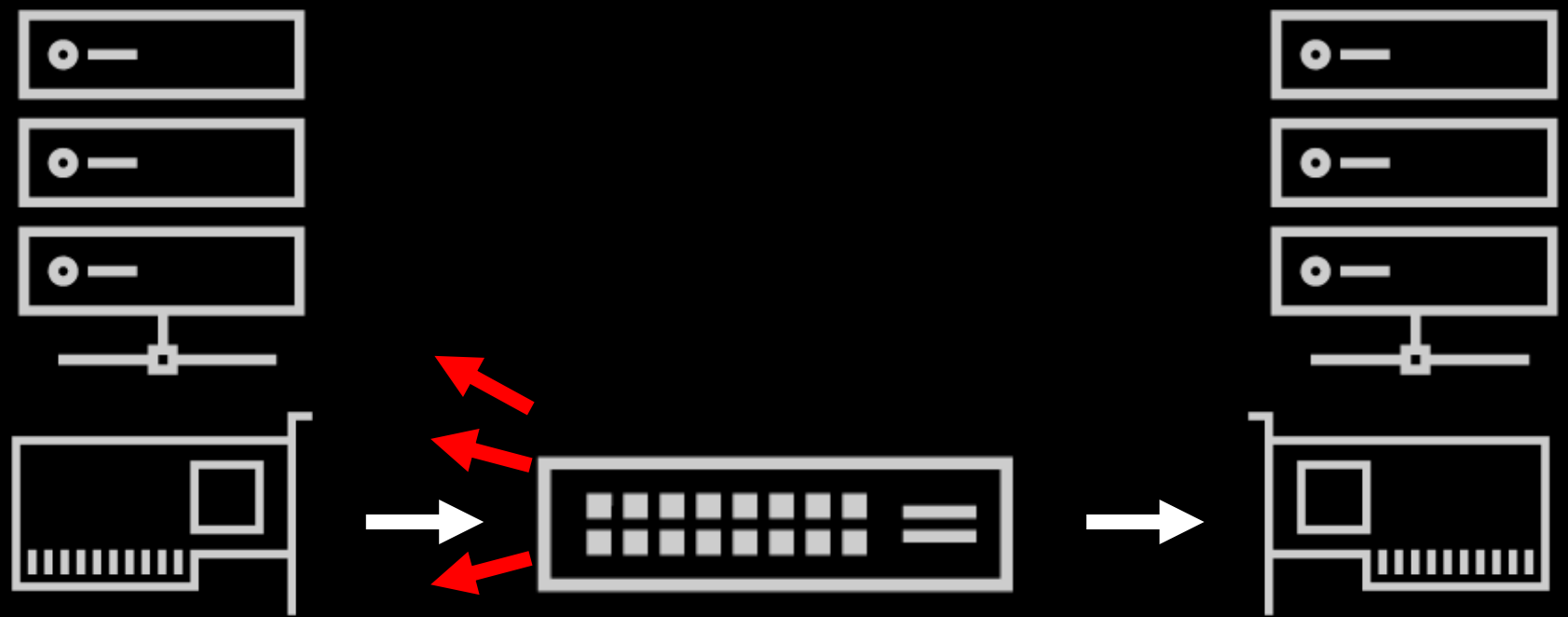
■ On Host ■ DPU Offload

OSU Ialltoallv 32 (full) PPN, Size = 128 KB

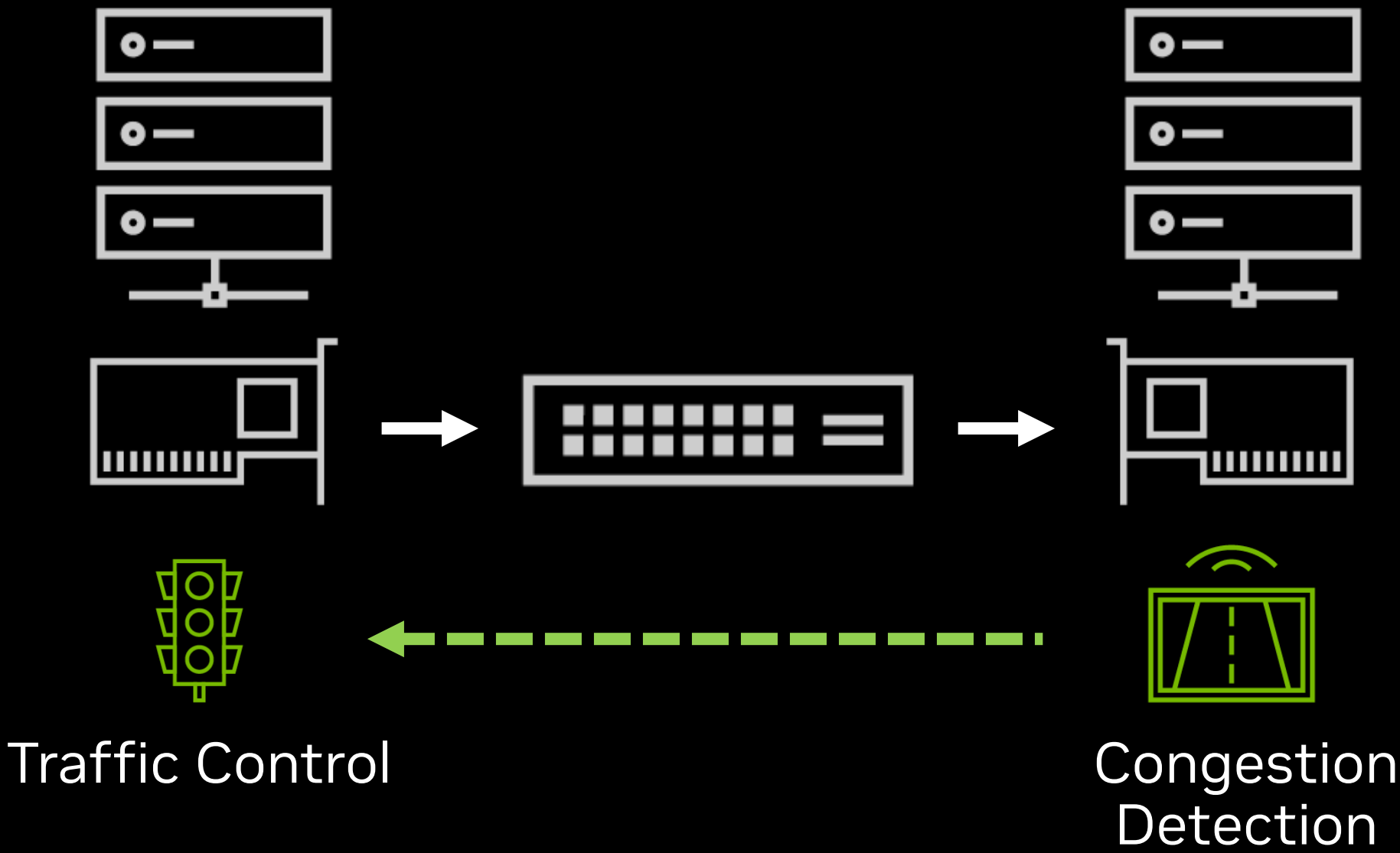


■ On Host ■ DPU Offload

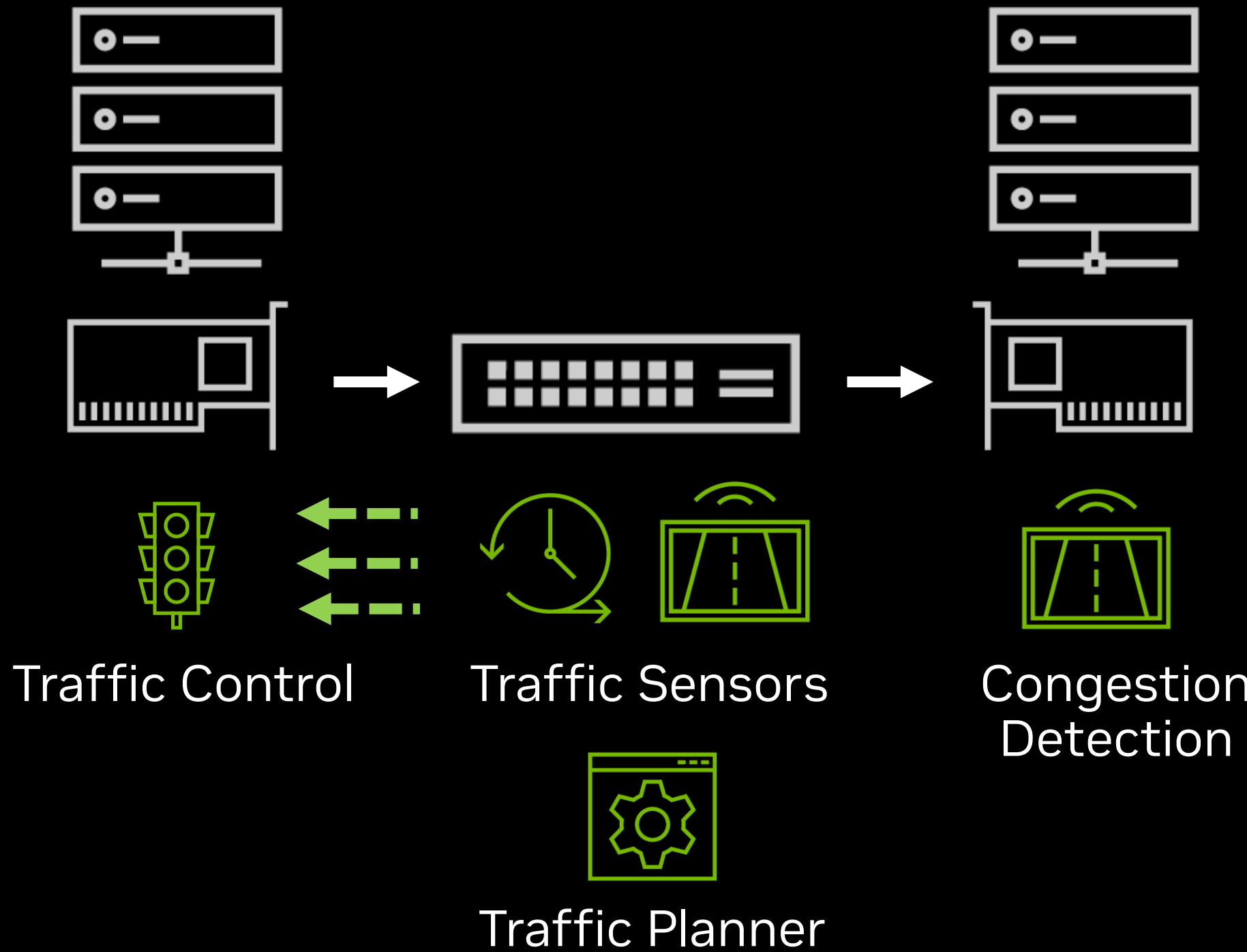
Congestion Control Evolution



Congestion Spreading

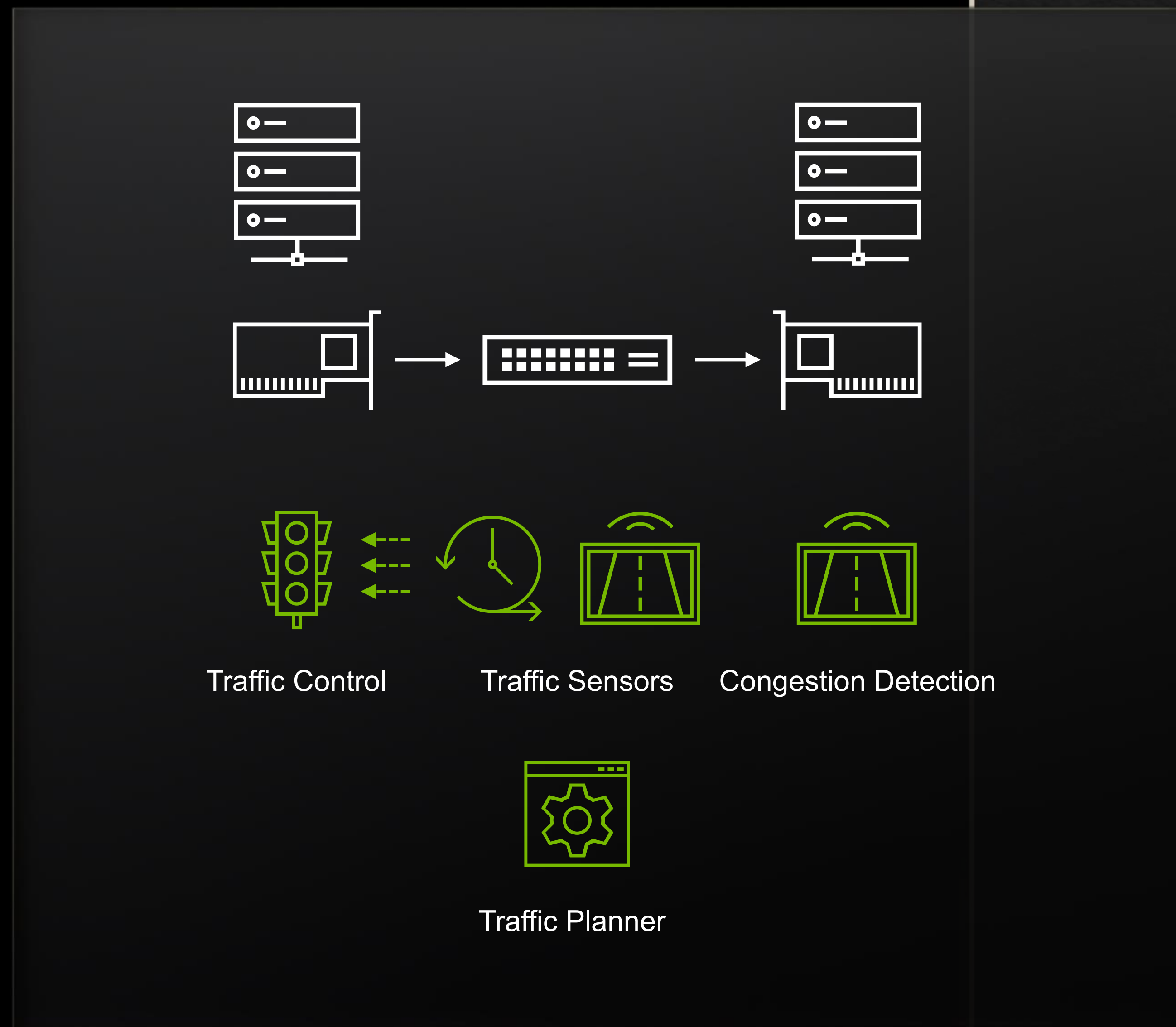


Congestion Detectors

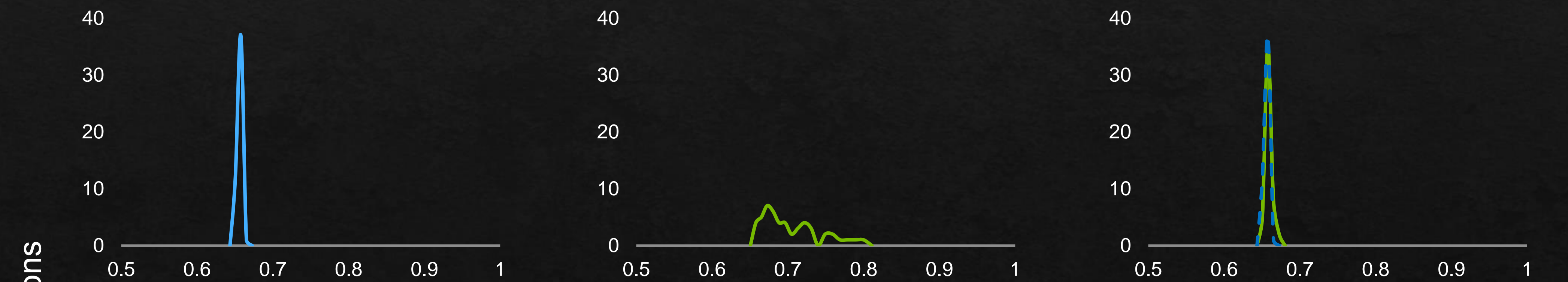


Telemetry Data
Time Sensors
Traffic Planners

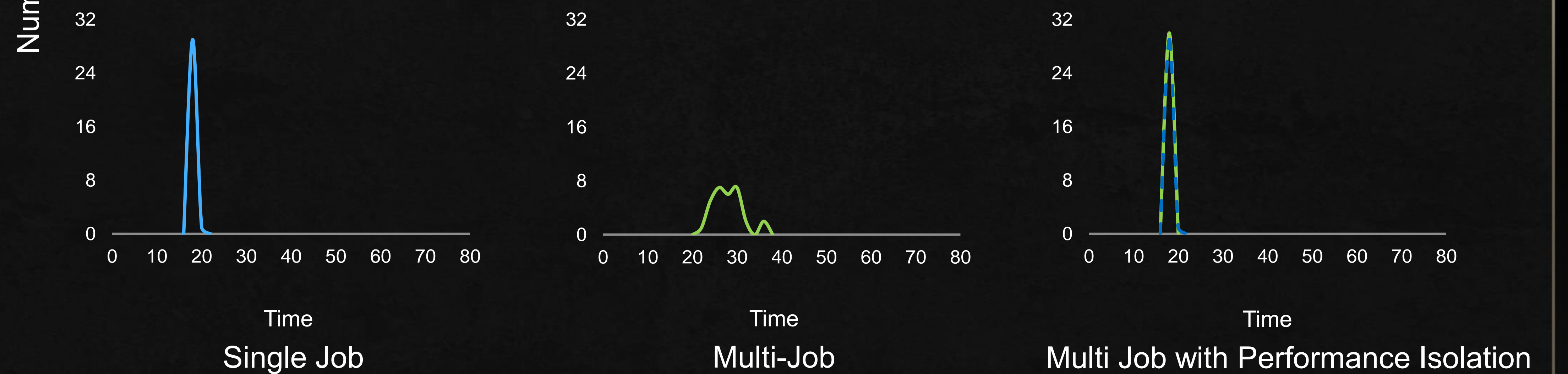
Performance Isolation



MOLECULAR DYNAMICS (LAMMPS)



PARALLEL ALGEBRAIC MULTI-GRID SOLVER (AMG)



NVIDIA Full Stack HPC Platform

