2023 OFA Virtual Workshop

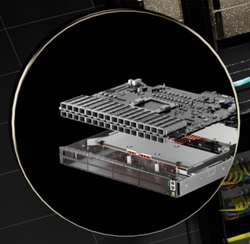
ACCELERATING SCIENTIFIC COMPUTING WORKLOADS WITH INFINIBAND DPUS

Richard Graham

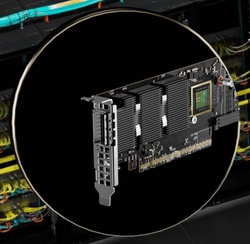
NVIDIA

NVIDIA CLOUD-NATIVE HPC PLATFORM

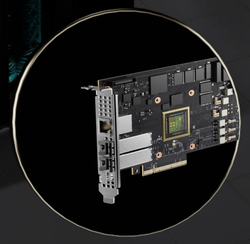
- In-Network Computing
- Zero Trust Security
- Performance Isolation
- Computational Storage
- Enhanced Telemetry



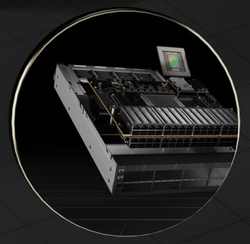
QUANTUM INFINIBAND SWITCH



CONNECTX SMARTNIC



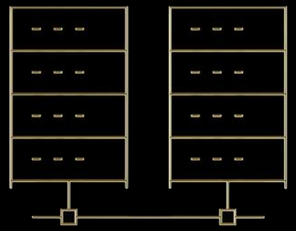
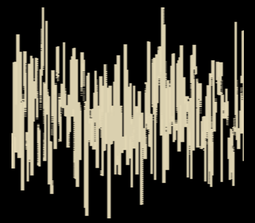
BLUEFIELD DPU



SPECTRUM ETHERNET SWITCH

UFM/NetQ

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

BlueField Data Processing Unit

SOFTWARE DEFINED NETWORKING

SOFTWARE DEFINED SECURITY

SOFTWARE DEFINED STORAGE

Infrastructure Services

BlueField Infrastructure
Compute Platform

Data Center on a Chip

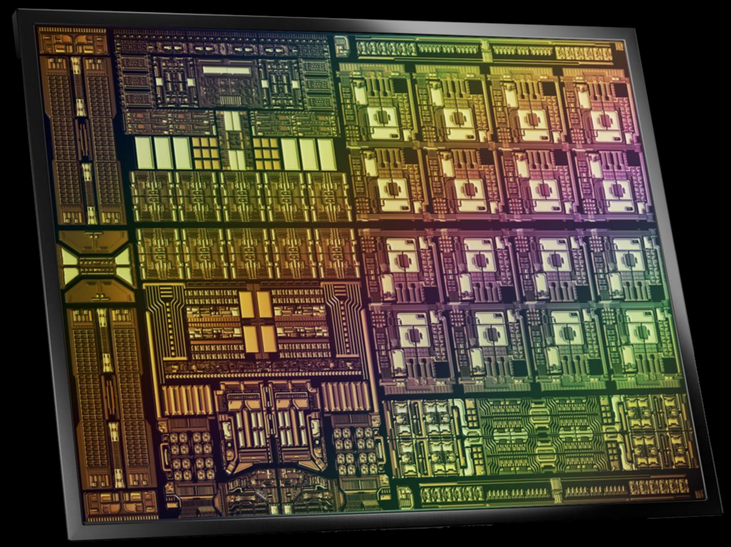
16 Arm 64-Bit Cores

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

PLATFORM INFRASTRUCTURE

STORAGE

CYBERSECURITY

COMPUTE

DOCA

Orchestration

Security

Networking

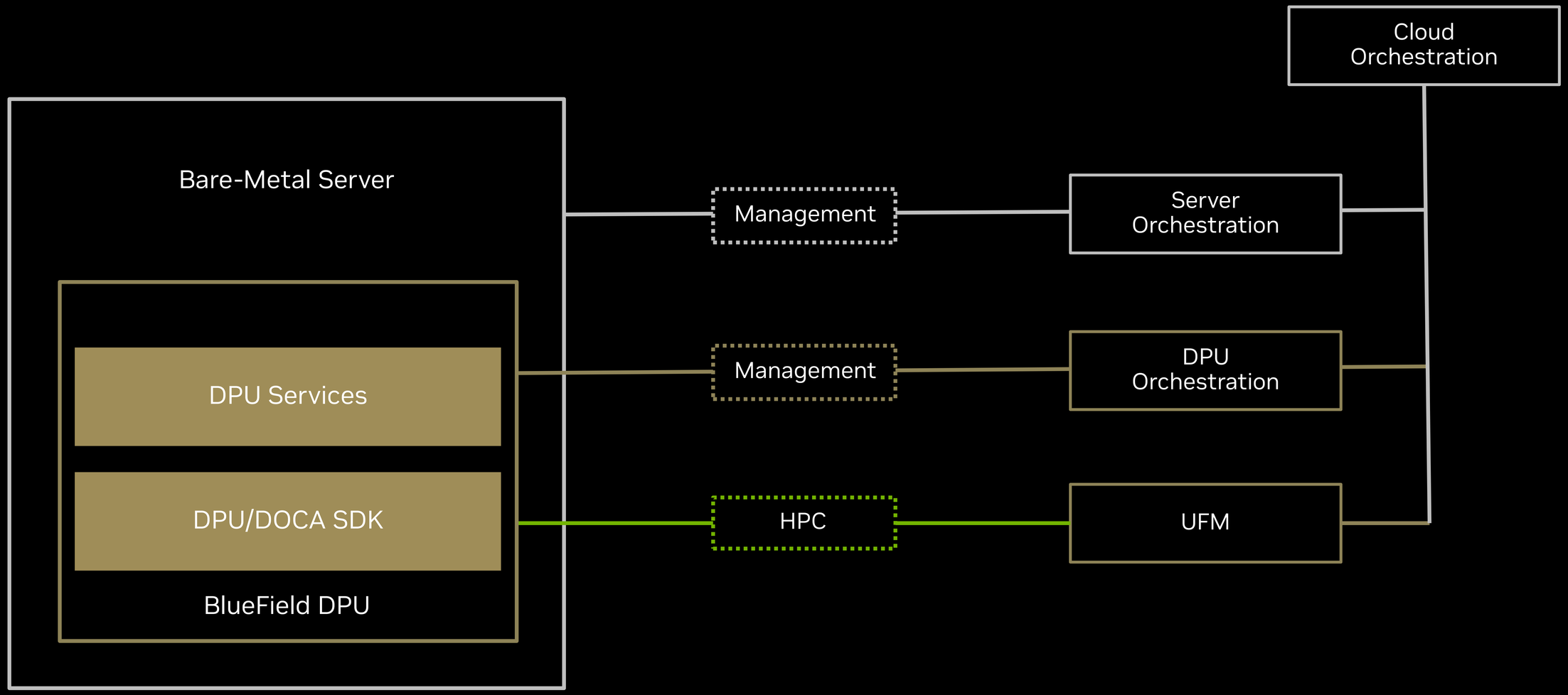
Storage

Management

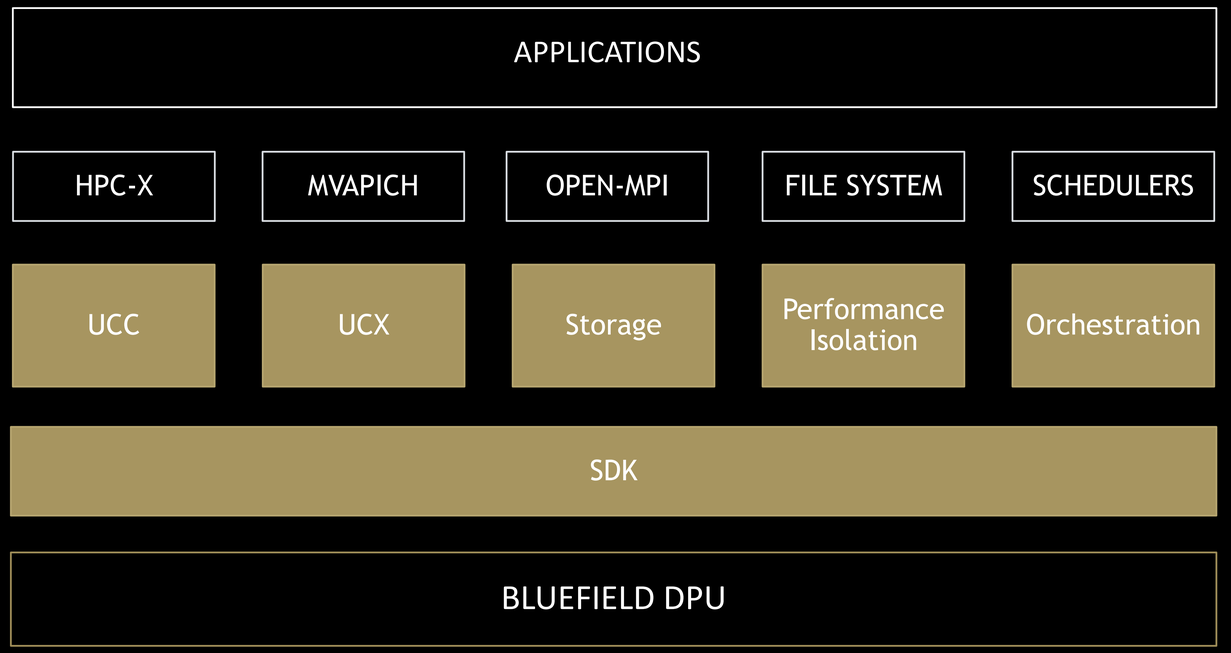
Acceleration Libraries

Telemetry

Delivering Cloud Native Supercomputing

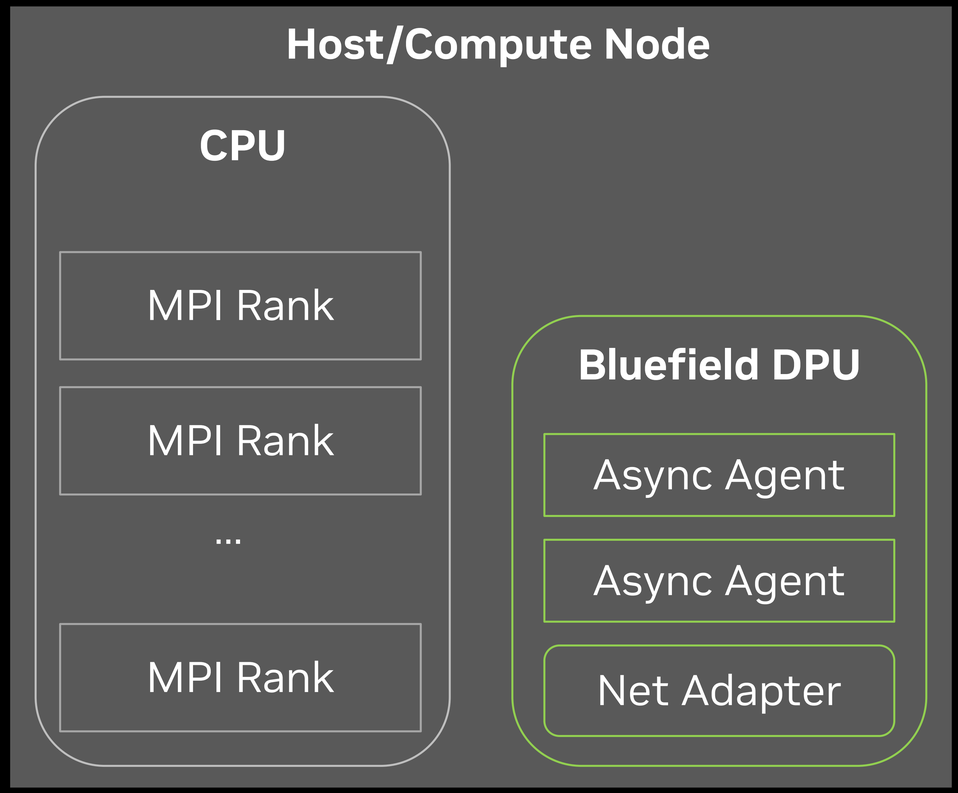


Accelerating HPC Applications with DPU/DOCA Services



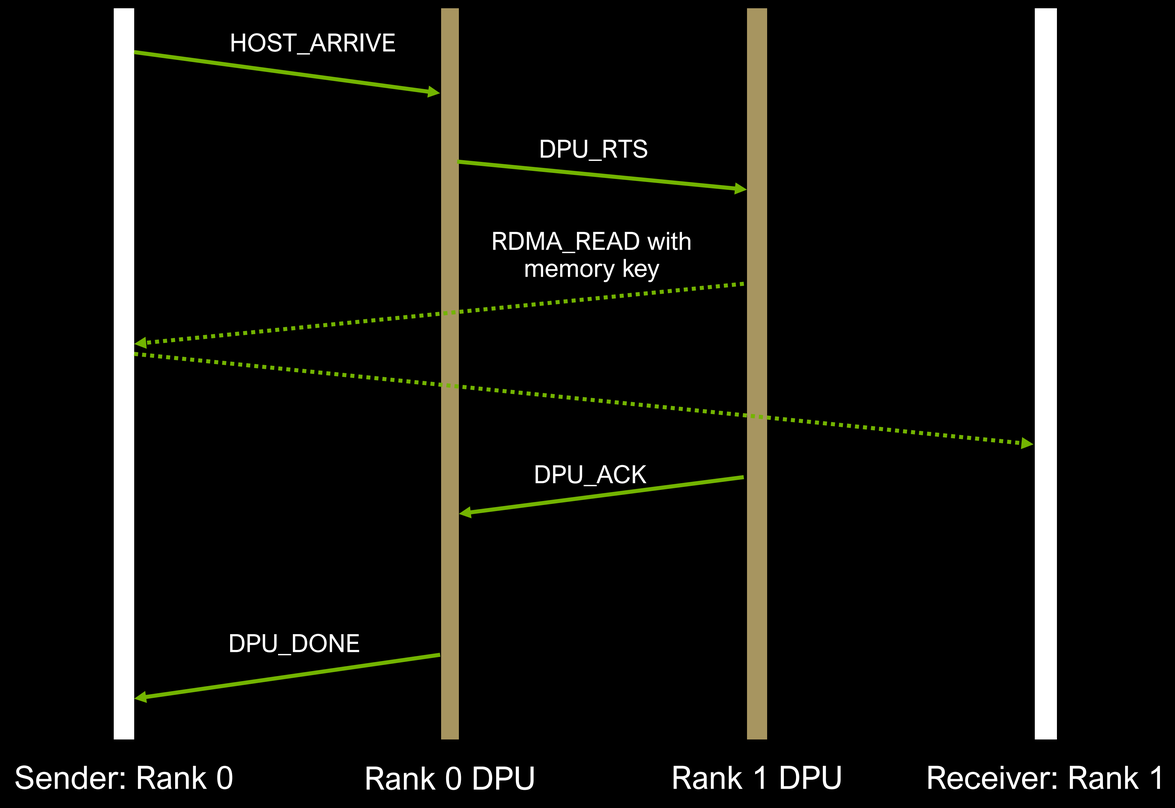
High Level System Components from Software's Perspective

- Host paired with local DPU
- Local DPU runs service processes (SP)
  - Each local user process (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

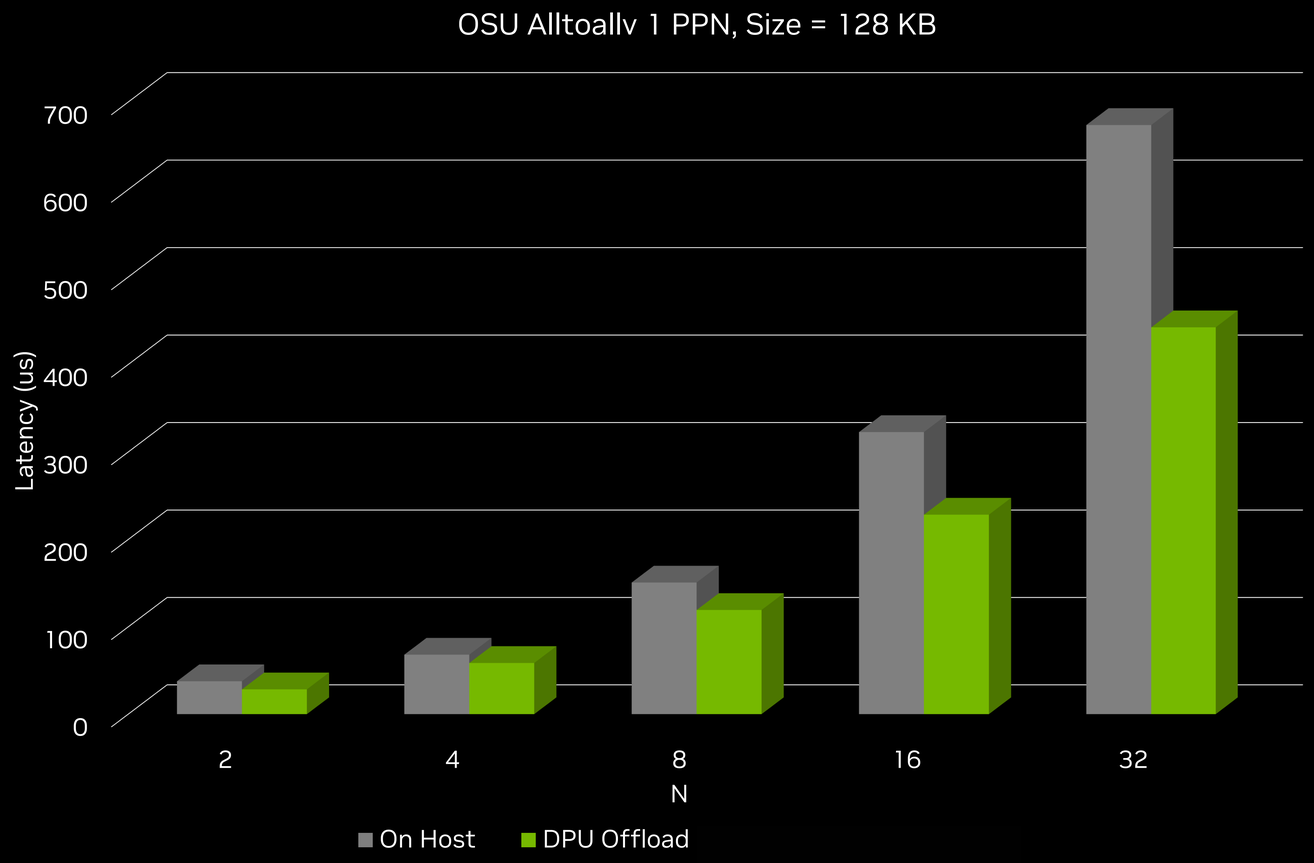


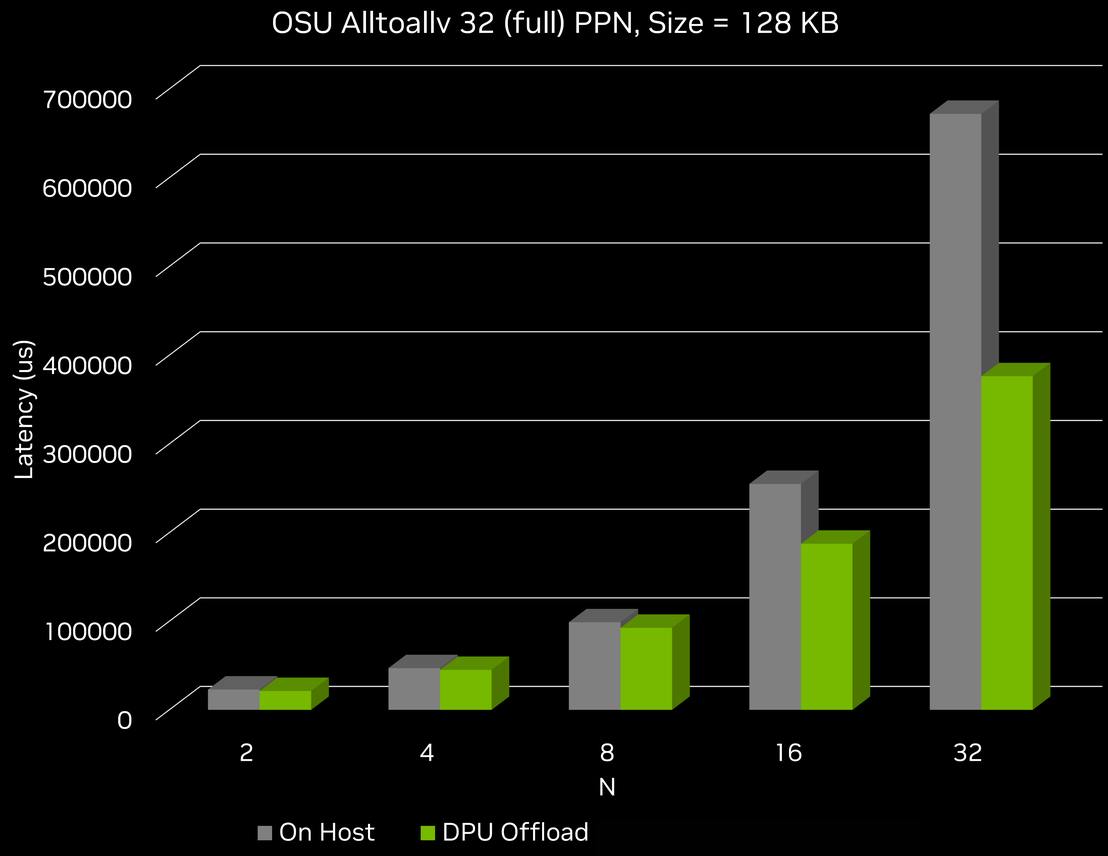
Offloading and Accelerating Data Exchange Example

An Element of Collective Algorithm

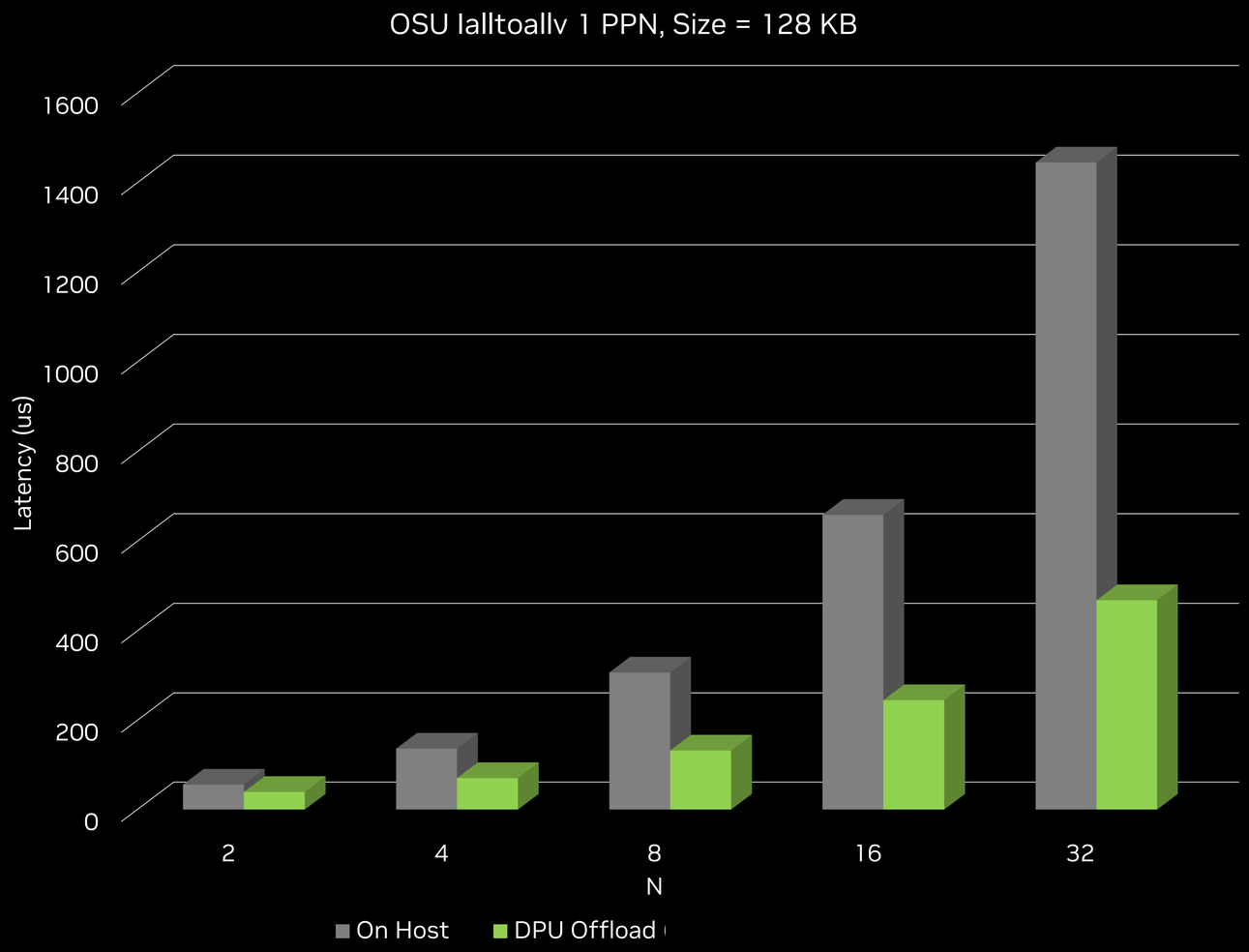


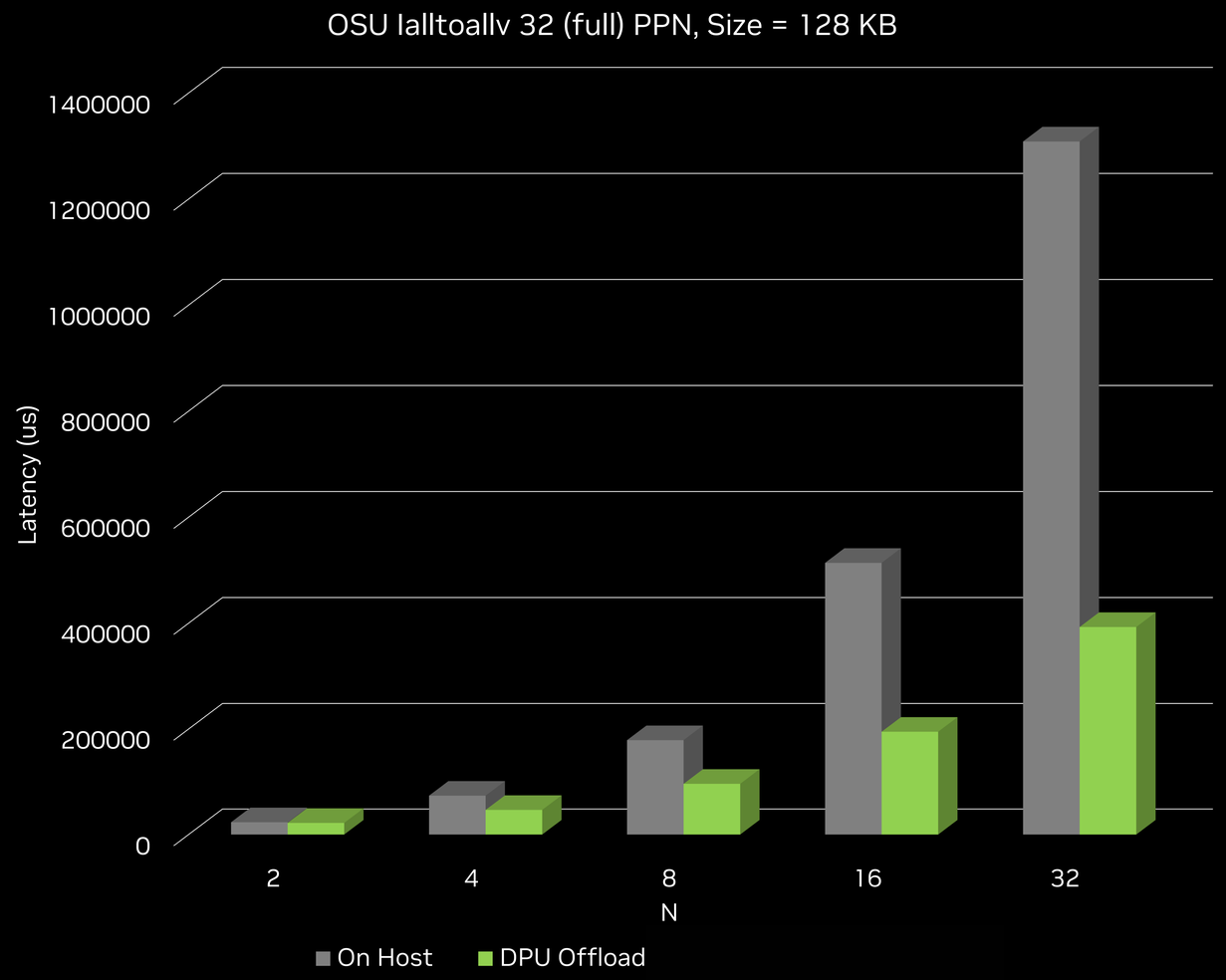
Alltoallv Latency



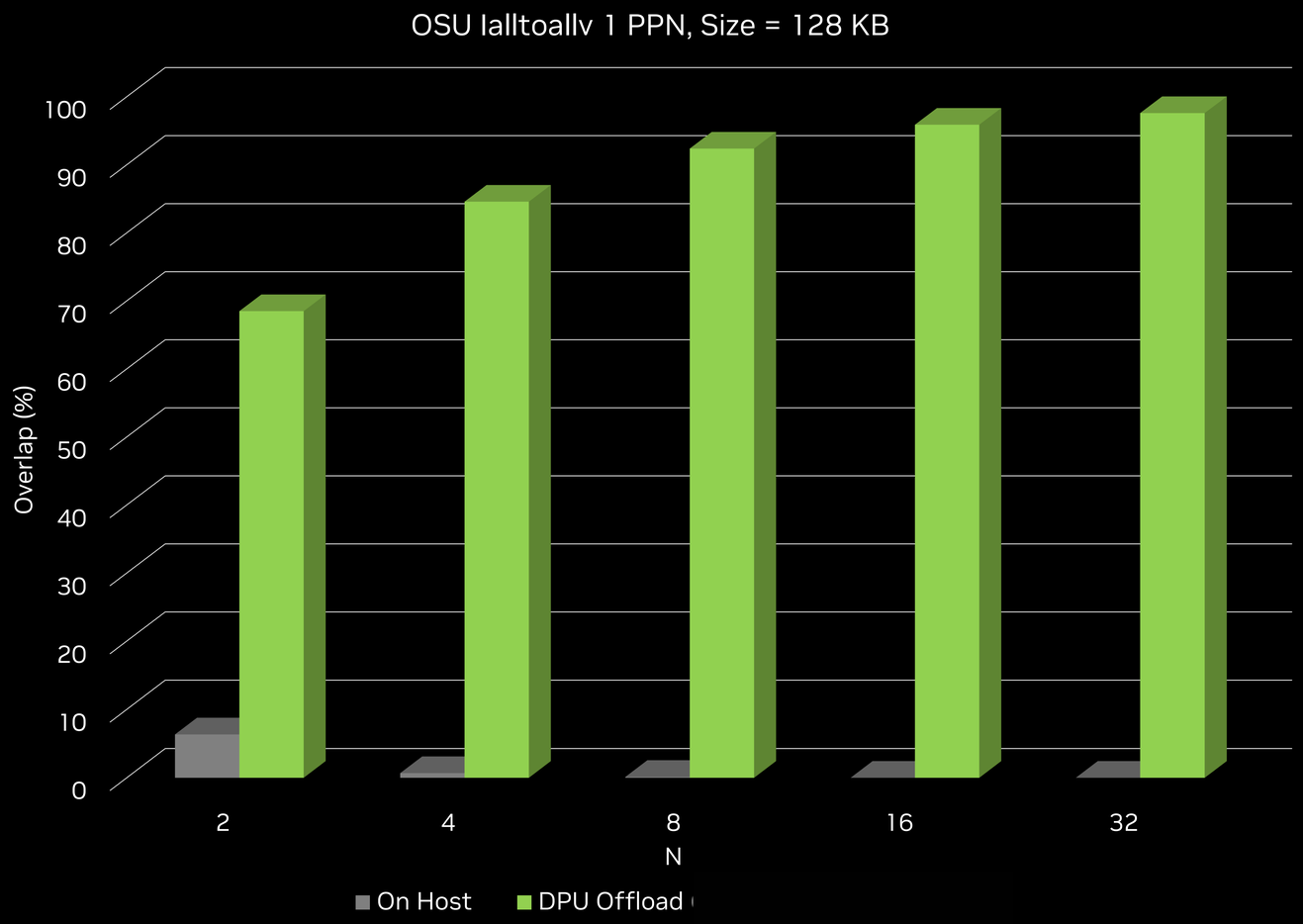


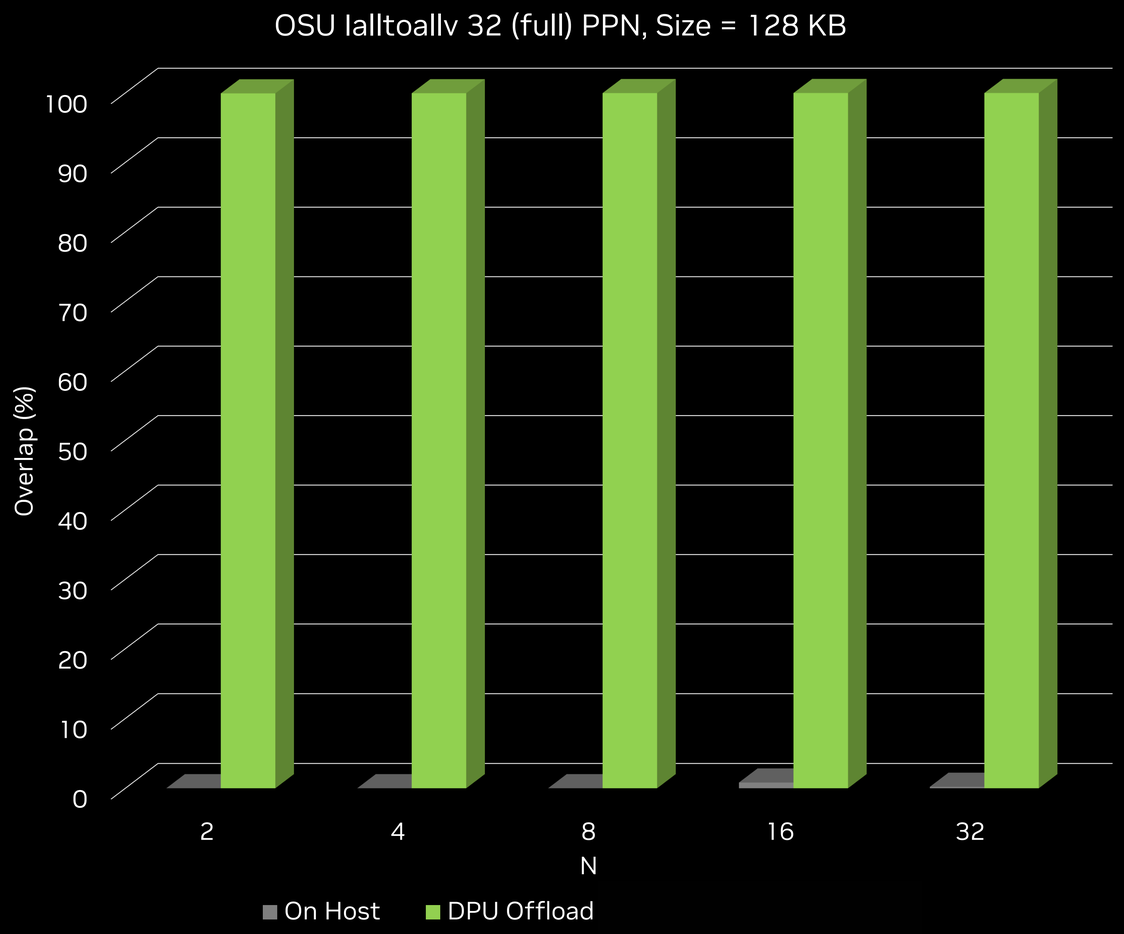
iAlltoallv latency





iAlltoallv compute/communication overlap

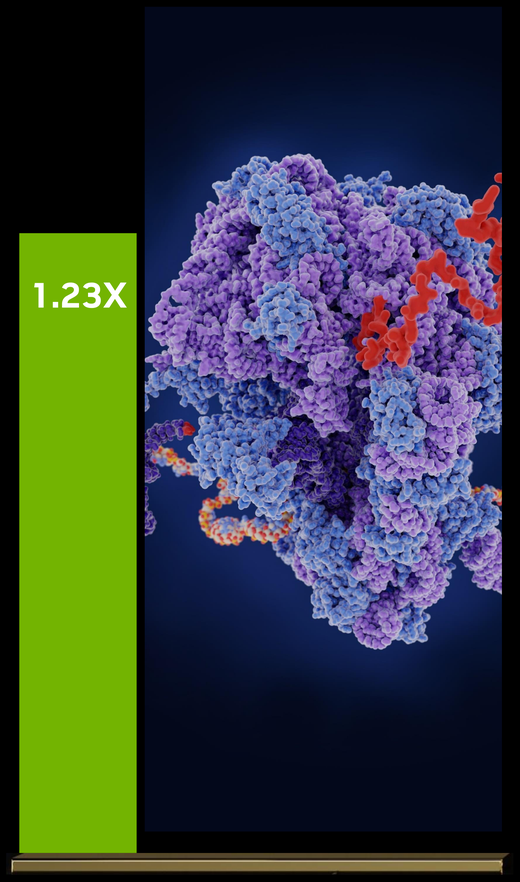




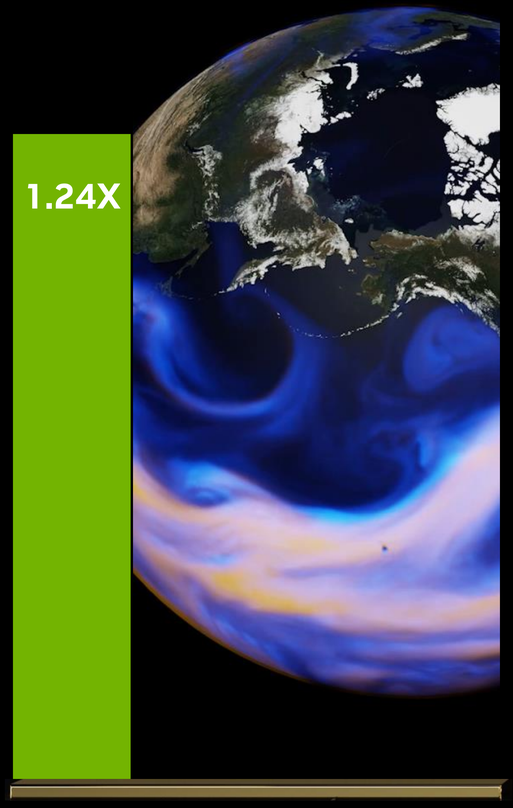
HIGHER APPLICATION PERFORMANCE

With BlueField DPU and Quantum InfiniBand In-Network Computing

Octopus (Physics / Chemistry)



GSAM (Weather)



FFT


NVIDIA