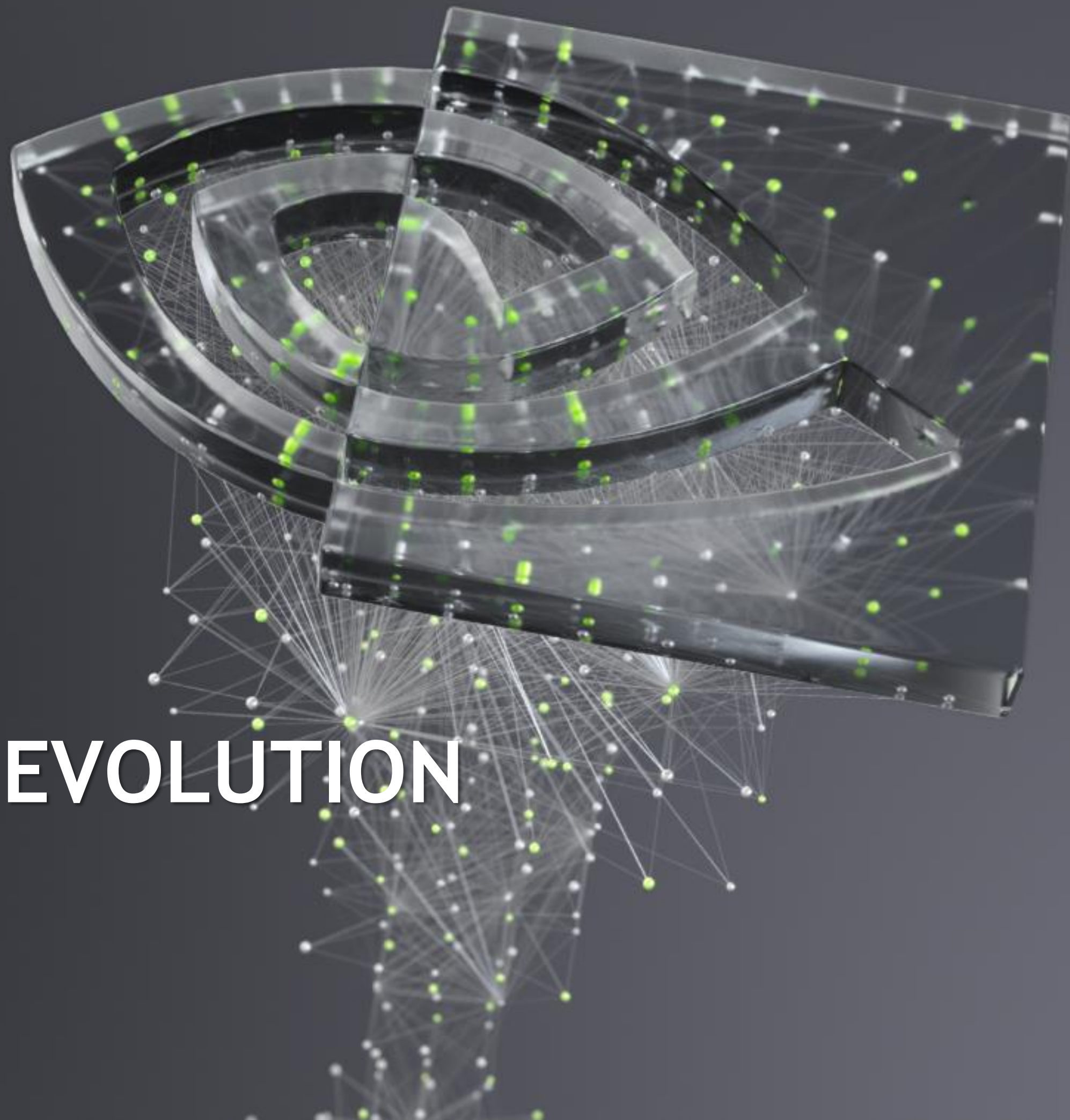




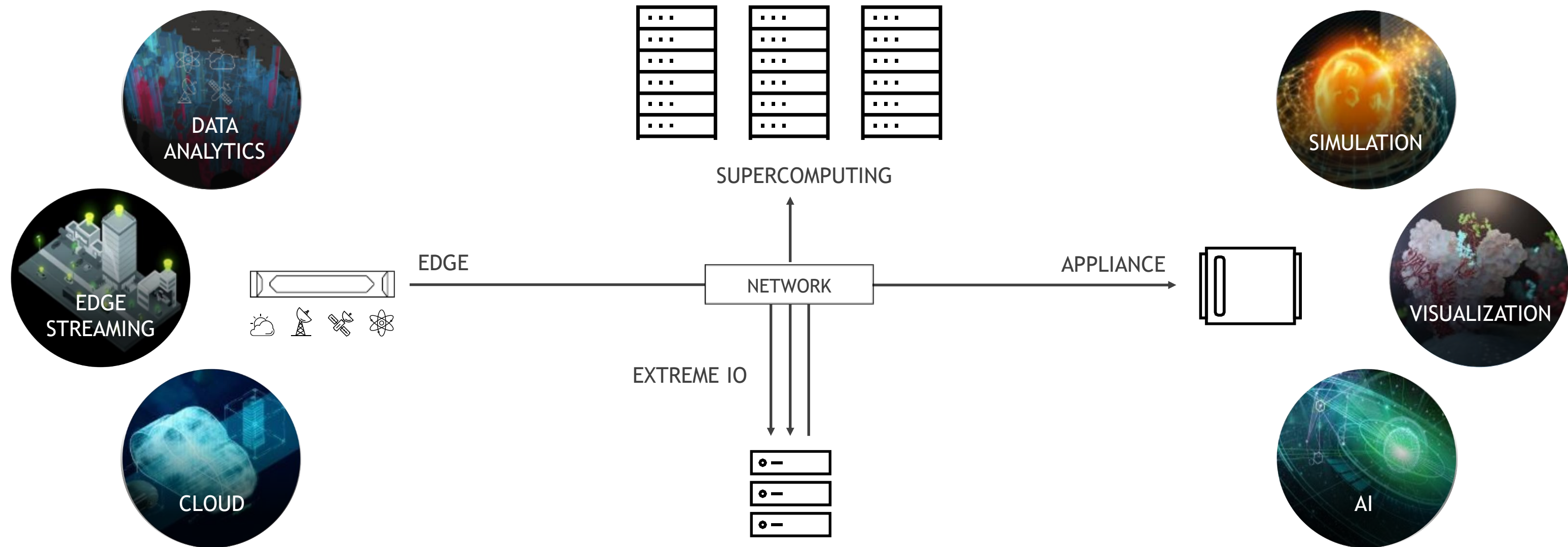
# HPC CONNECTIVITY EVOLUTION

## INFINIBAND NDR & DPU

Ramiro Alvarez - [ramiroa@nvidia.com](mailto:ramiroa@nvidia.com) - May 2021



# EXPANDING UNIVERSE OF SCIENTIFIC COMPUTING





# HDR 200G INFINIBAND ACCELERATES NEXT GENERATION HPC AND AI SUPERCOMPUTERS (EXAMPLES)



9 PFlops  
3K HDR Nodes  
Dragonfly+ Topology



19.5 PetaFLOPS  
2.5K HDR Nodes  
Dragonfly+ and Fat Tree



16 PFlops  
3K HDR Nodes  
Dragonfly+ Topology



8K HDR Nodes  
Dragonfly+ Topology



35.5 PFlops  
2K HDR Nodes  
Fat-Tree Topology



19.3 PFlops  
5.6K HDR Nodes  
Dragonfly+ Topology



63.5 PFlops  
4.5K HDR Nodes  
Fat-Tree Topology



HPC/AI Cloud  
HDR InfiniBand



**SDSC**  
SAN DIEGO SUPERCOMPUTER CENTER



**PURDUE**  
UNIVERSITY

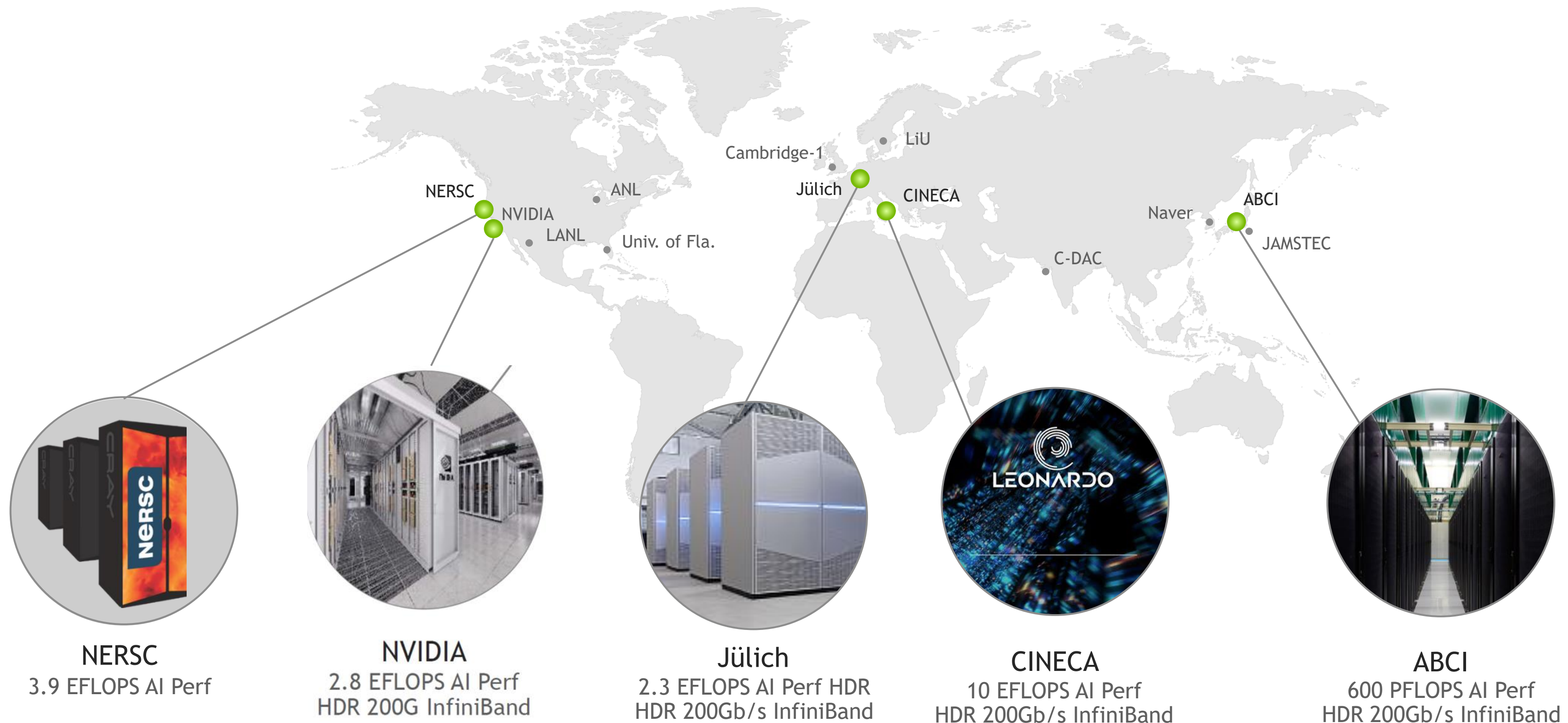
HDR InfiniBand  
Supercomputers



23.5 PFlops  
8K HDR Nodes  
Fat-Tree Topology



# NVIDIA PLATFORM POWERING THE EXASCALE AI SUPERCOMPUTERS

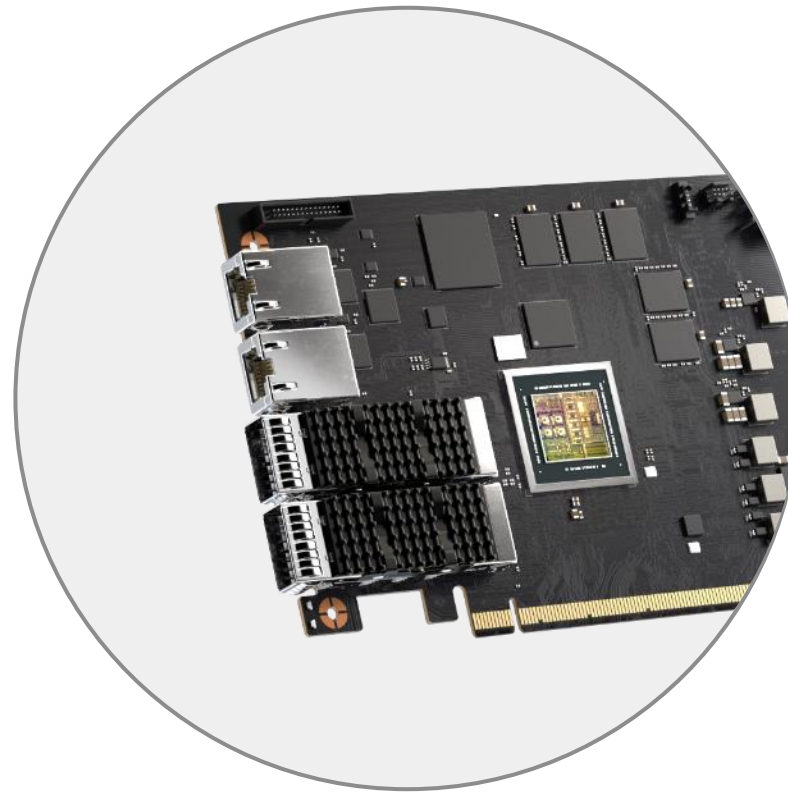


# MELLANOX NDR 400G INFINIBAND: NEXT-GENERATION INFINIBAND ARCHITECTURE



## Adapter

NDR 400G InfiniBand  
PCIe Gen4 and Gen5  
Programmable Datapath  
In-Network Computing



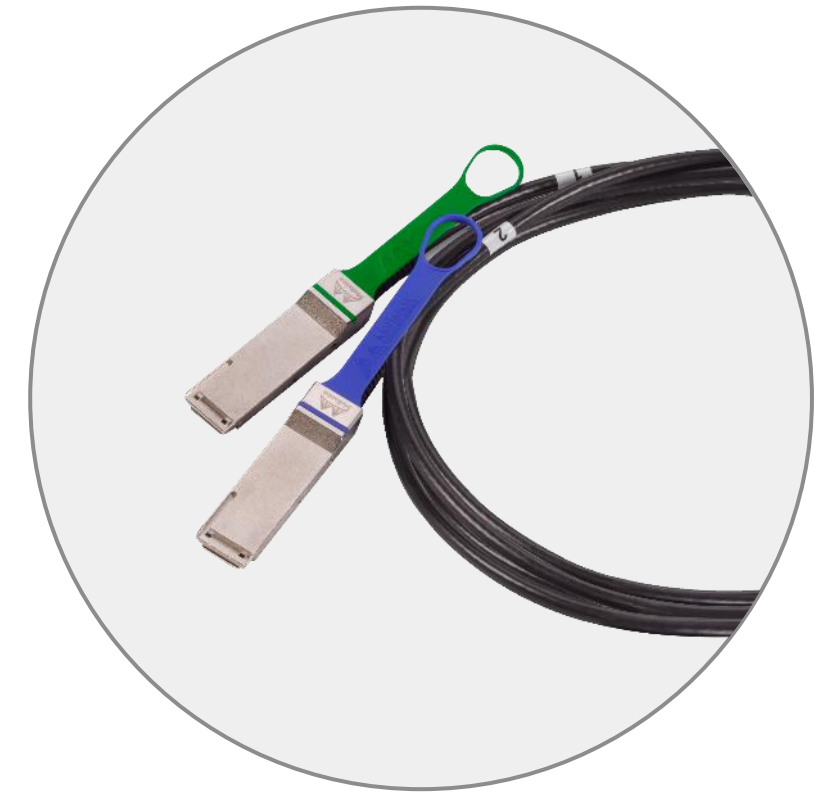
## DPU

NDR 400G InfiniBand with Arm Cores  
PCIe Gen4 and Gen5, DDR5  
AI Application Accelerators  
Programmable Datapath  
In-Network Computing



## Switch

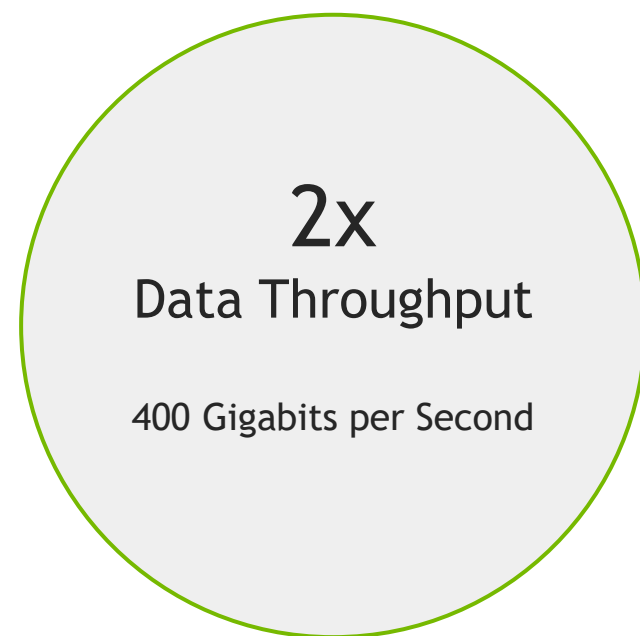
64-ports NDR 400G InfiniBand  
128-ports 200G NDR200  
In-Network Computing  
1U system: 64 x 400G, 128 x 200G  
Modular: 2048 x 400G, 4096 x 200G



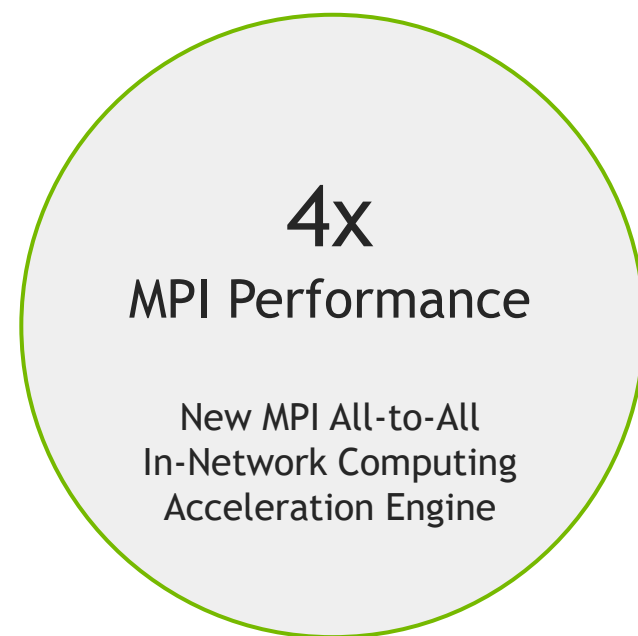
## Cable

Copper Cables  
Active Copper Cables  
Optical Transceivers

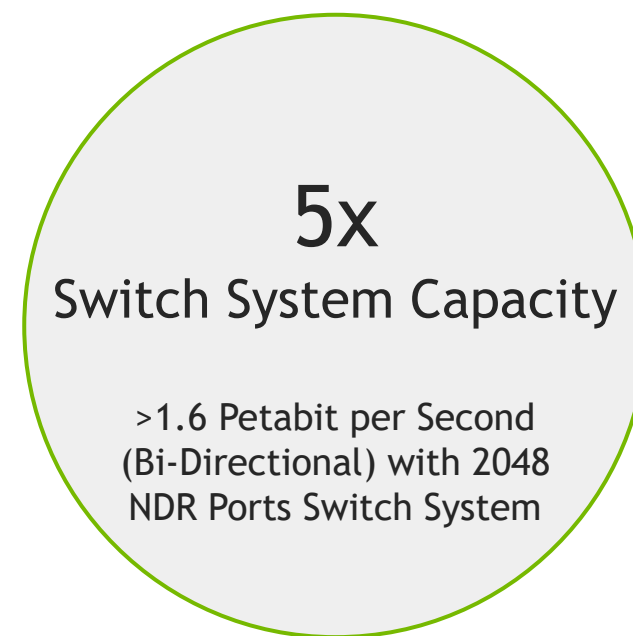
# MELLANOX NDR 400G INFINIBAND: INFINIBAND CONTINUES TO SET WORLD RECORDS



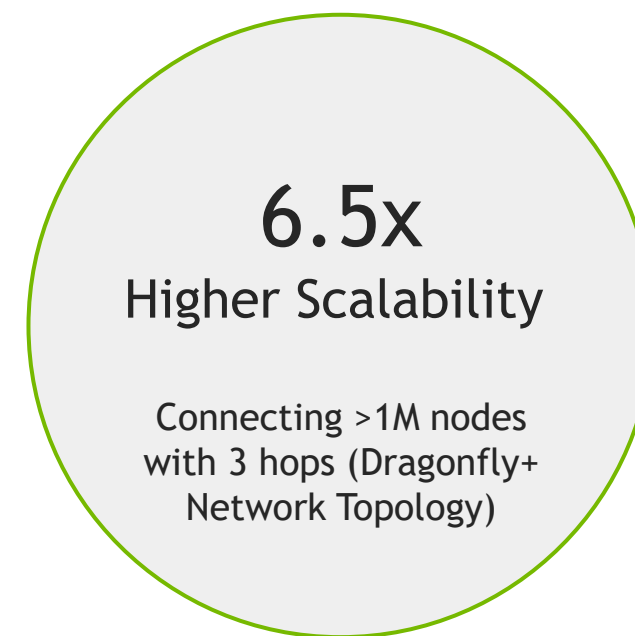
Data Speed



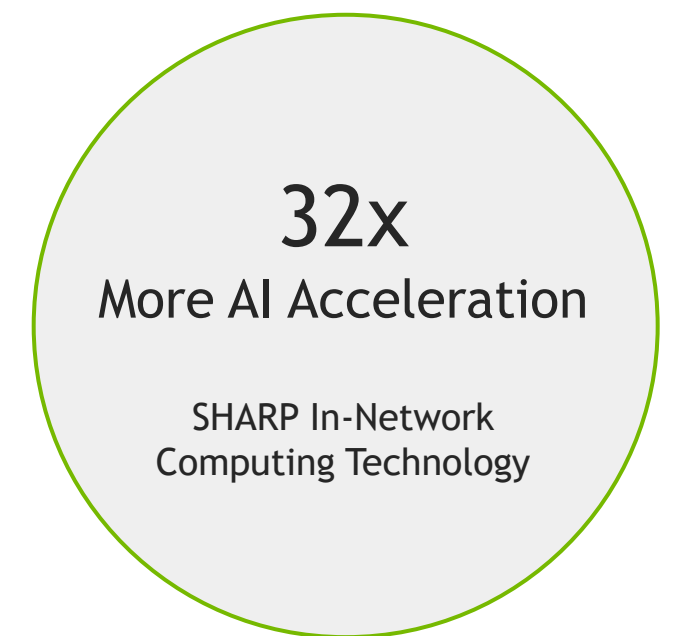
MPI Performance



Improved TCO



Exascale Ready



Accelerated Deep Learning



# NDR 400G INFINIBAND SOLUTIONS

400Gb/s NDR and 200Gb/s NDR200 InfiniBand

PCIe Gen5, PCIe Gen4, Multi-Host - up to 8 hosts

Switch: 64 ports of NDR (400G), 128 ports of NDR200 (200G)

Modular: 2048 ports of NDR (400G), 4096 ports of NDR200 (200G)

Modular: 1024 ports of NDR (400G), 2048 ports of NDR200 (200G)

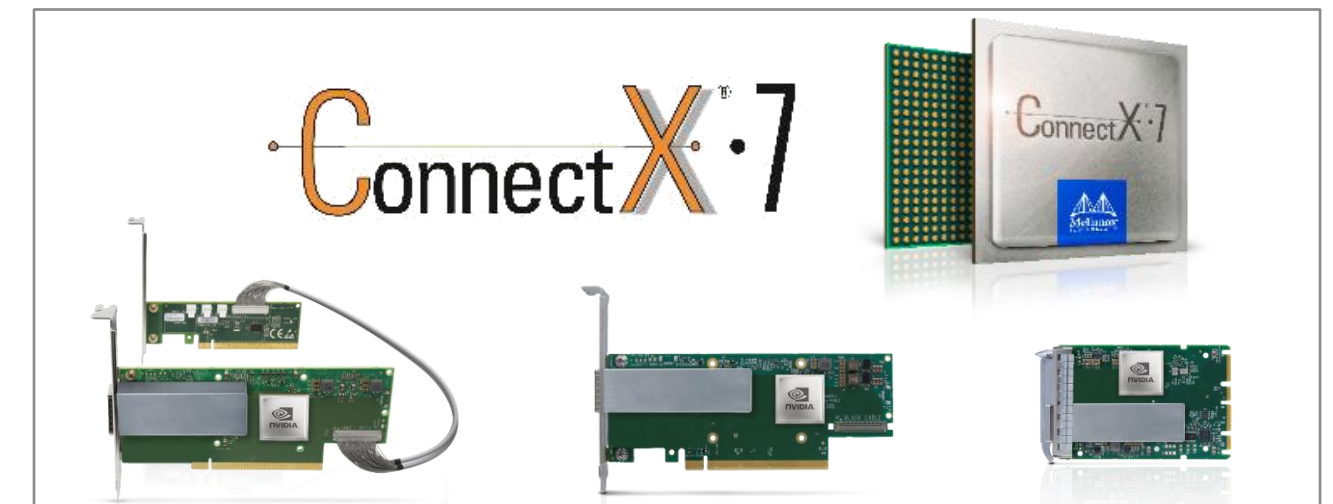
BlueField: 16 Arm cores, 24 PCIe lanes, 2.6GHz

Advanced adaptive routing and congestion control

SHARPy3 - low latency data reduction and streaming aggregation

MPI All-to-All hardware engine, MPI Tag Matching hardware engine

Enhanced NVMe-over-fabrics, Security accelerations and data compressions



# QUANTUM-2 SWITCH

## QM9700 and QM9790 Family of 1U Switches

64 ports of 400Gb/s (NDR)

128 ports of 200Gb/s (NDR200)

51.2Tb/s aggregate bandwidth

66.5 billion packets per second

SHARPV3 - low latency data reduction and streaming aggregation

Internally managed (QM9700), and externally managed (QM9790) SKUs

26'' depth, Air cooled

2 power supplies (1+1), hot swappable





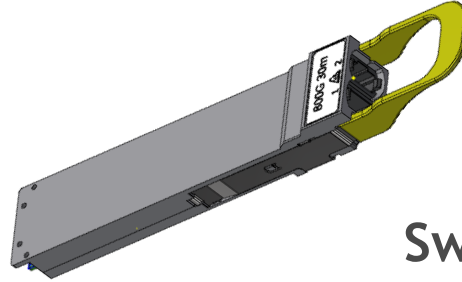
# NDR INFINIBAND CABLING OVERVIEW

## Switch

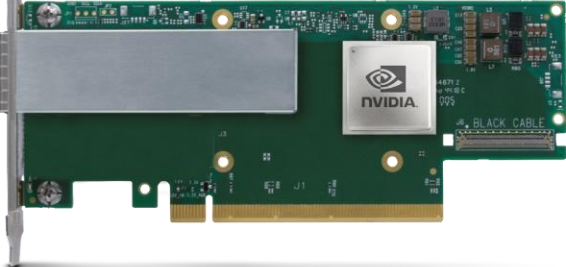
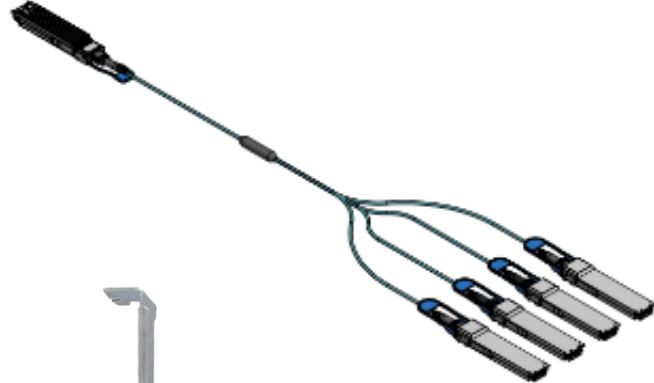
64 ports of NDR (4x100Gb/s PAM4)  
32 OSFP connectors - 2 ports per connector



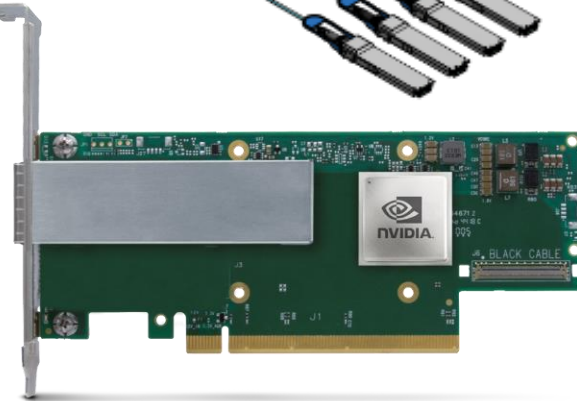
Switch-to-switch  
Optical cables  
Transceivers two MPOs



Switch-to-HCA  
OSFP to 2x OSFP  
or  
OSFP to 4x OSFP

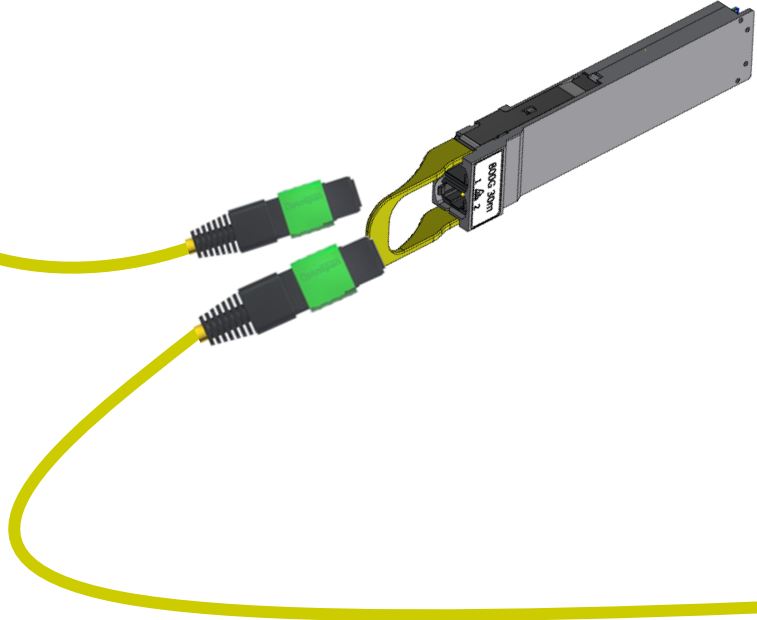


NDR Adapter



NDR200 Adapter

HCA  
NDR and NDR200 OSFP connectors



# NDR (4X100G PAM4) CONNECTIVITY

## Transceivers

Twin-port transceiver

Single-port NDR transceiver, Single-port NDR200 transceiver

MPO (for NDR) and split-MPO (for NDR200) offering

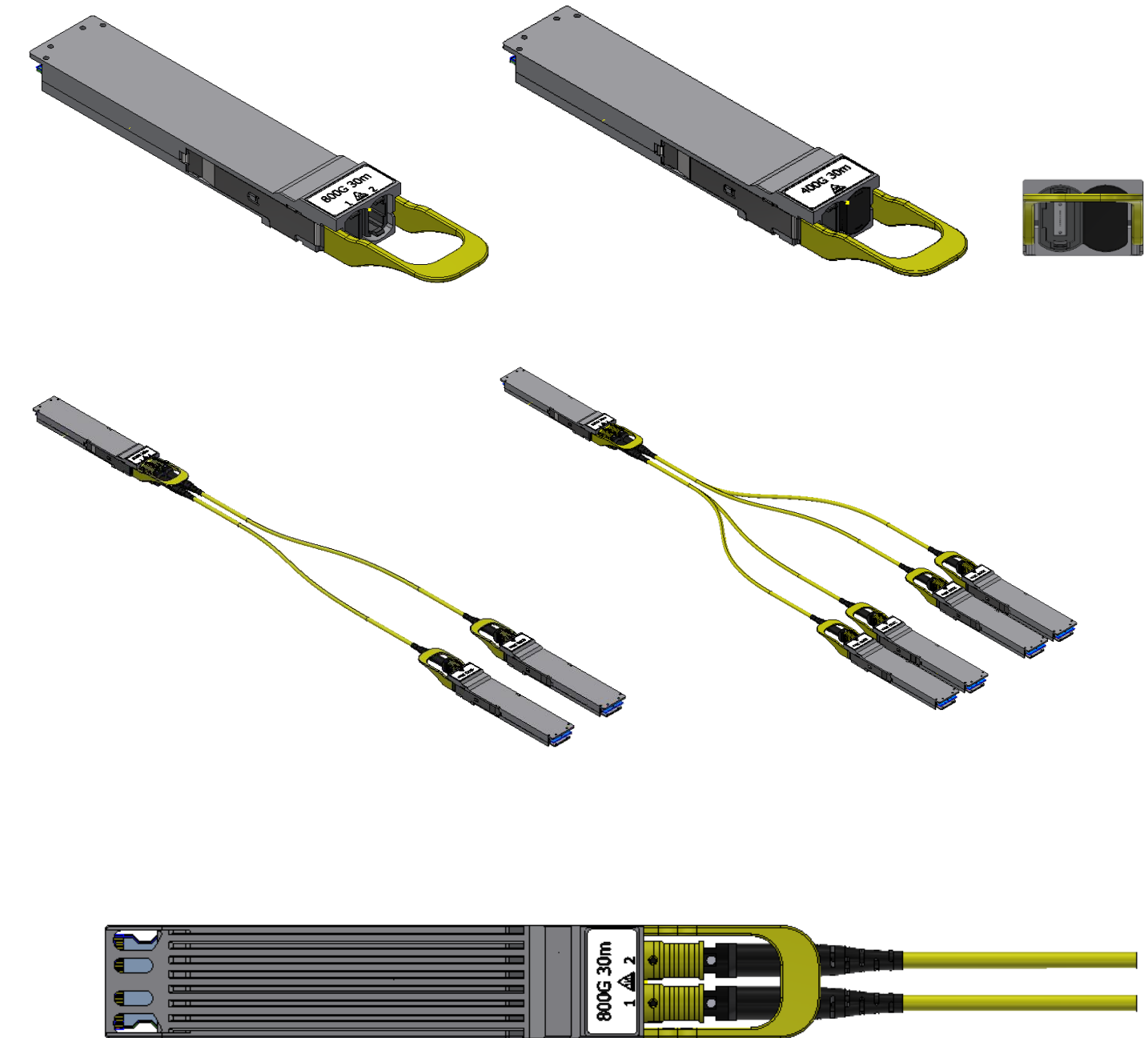
Single mode (Yellow)

Types of transceivers: up to 30m, up to 150m

17W for 2x NDR, 9W for NDR, 5W for NDR200

Finned OSFP - for air-cooled systems

Flat OSFP - for liquid-cooled port






*Pending U.S. patent application No. 16/750,632*



# NDR (4X100G PAM4) CONNECTIVITY



## Cables

	NDR Technology	
Switch	Air / Liquid cooled	Air / Liquid cooled
	Finned / Flat OSFP	Finned / Flat OSFP
	2x400 (2x NDR)	2x400 (4x NDR200)
	OSFP to 2xOSFP	OSFP to 4xOSFP
Form factor		
DAC - Copper up to 1.5m	✓	✓
ACC - Active copper up to 3m	✓	✓
HCA	NDR	NDR200
	OSFP	OSFP

	NDR Technology
Switch	Air / Liquid cooled
	Finned / Flat OSFP
	2x400 (2x NDR)
	OSFP to OSFP
Form factor	
DAC - Copper up to 0.5m	✓
ACC - Active copper up to 3m	✓
Switch	2x NDR
	OSFP

# NDR (4X100G PAM4) CONNECTIVITY

## Backward Compatibility

	Backward Compatibility		
	Air / Liquid cooled	Air / Liquid cooled	Air / Liquid cooled
	Finned / Flat OSFP	Finned / Flat OSFP	Finned / Flat OSFP
	2x200 (2x HDR)	2x200 (4x HDR100)	2x100 (2x EDR)
	OSFP to 2xQSFP	OSFP to 4xQSFP	OSFP to 2xQSFP
Switch			
Form factor			
DAC - Copper up to 1.5m	✓	✓	✓
ACC - Active copper	x	x	x
AOC - Optical cable up to 30m	✓	✓	✓
HCA or Switch	HDR	HDR100	EDR
	QSFP56	QSFP56	QSFP28



# CONNECTX-7 - 400G TO DATA-CENTRIC SOLUTIONS

400Gb/s ports using 100Gb/s SerDes

32 lanes of PCIe Gen5 (compatible with Gen4/Gen3)

PCIe switch and Multi-Host (up to 8 hosts) technology

400Gb/s (NDR) throughput

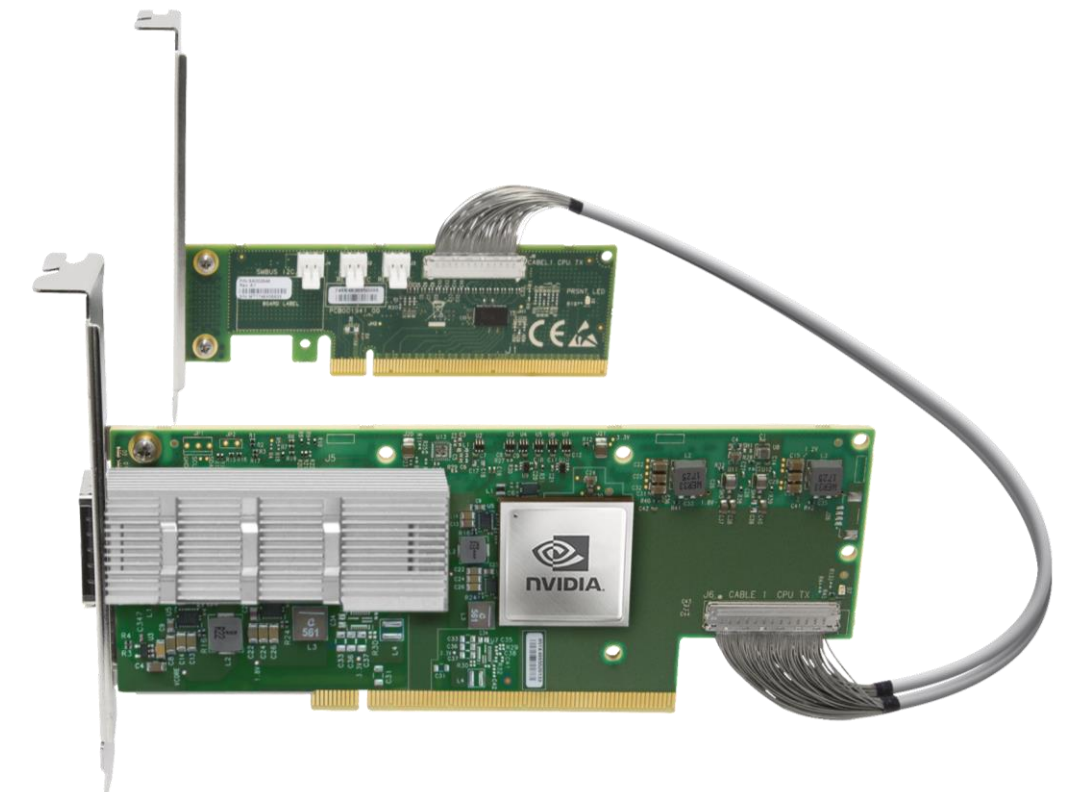
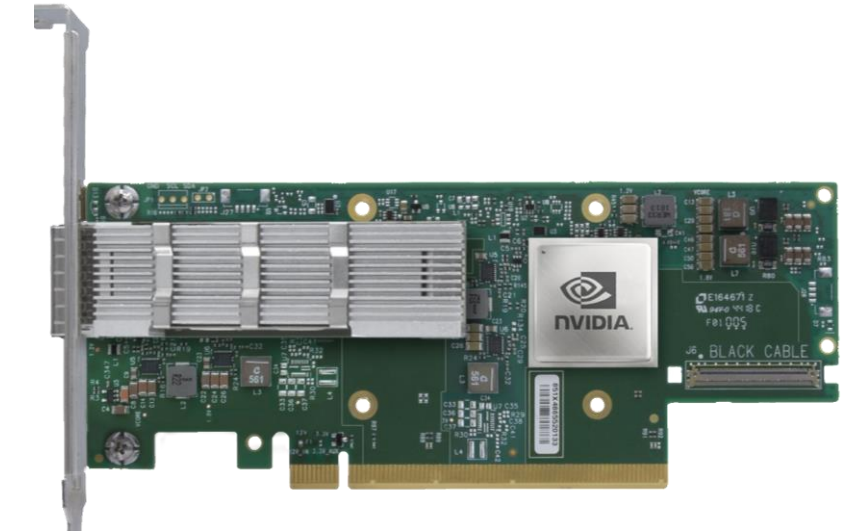
330-370M msg/sec rate

In-network computing

- MPI All-to-All hardware engine

- MPI Tag Matching hardware engine

- Programmable acceleration units



# MELLANOX SKYWAY™ INFINIBAND TO ETHERNET GATEWAY

100G EDR / 200G HDR InfiniBand to 100G and 200G Ethernet gateway

400G NDR / 800G XDR InfiniBand speeds ready

Eight EDR/HDR100/HDR InfiniBand ports to eight 100/200G Ethernet

Max throughput of 1.6 Terabit per second

High availability and load balancing

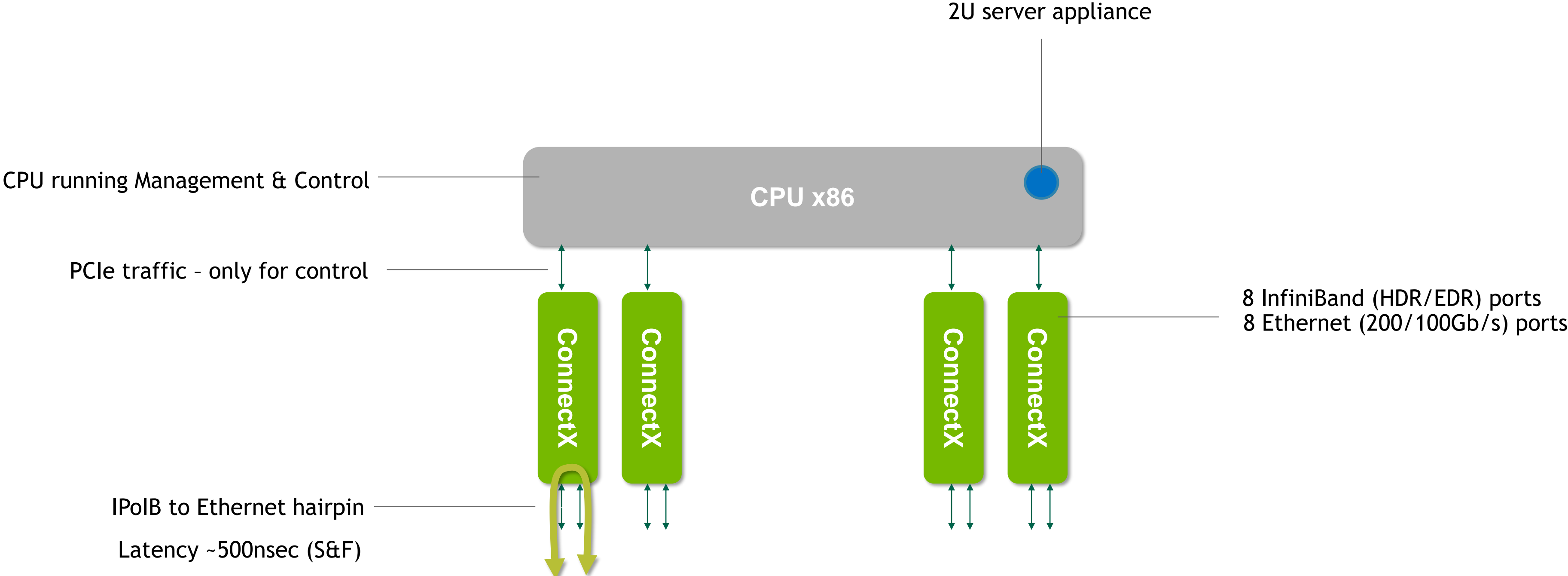
Mellanox Gateway operating system

Scalable and efficient

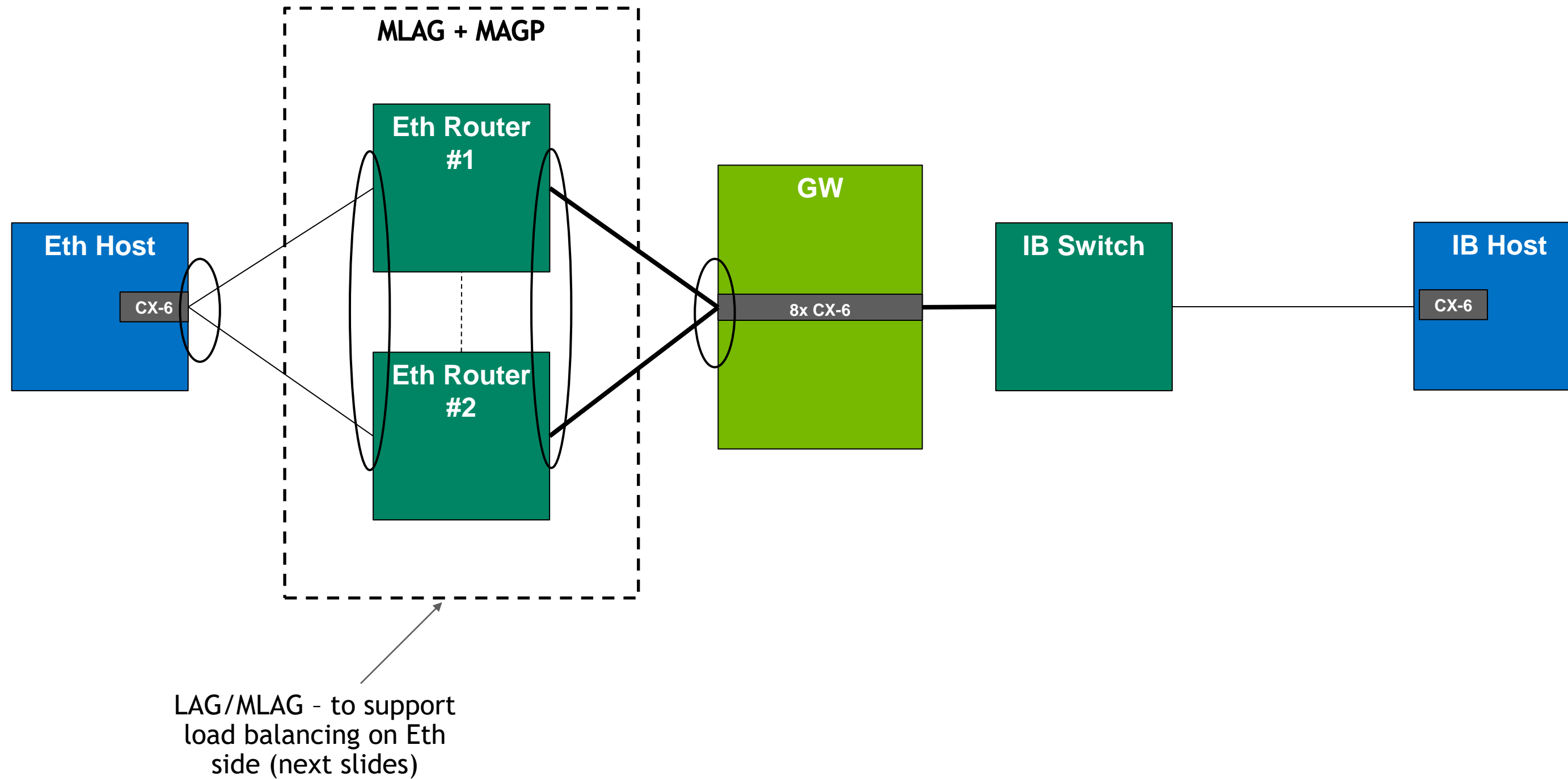




# GATEWAY SYSTEM ARCHITECTURE

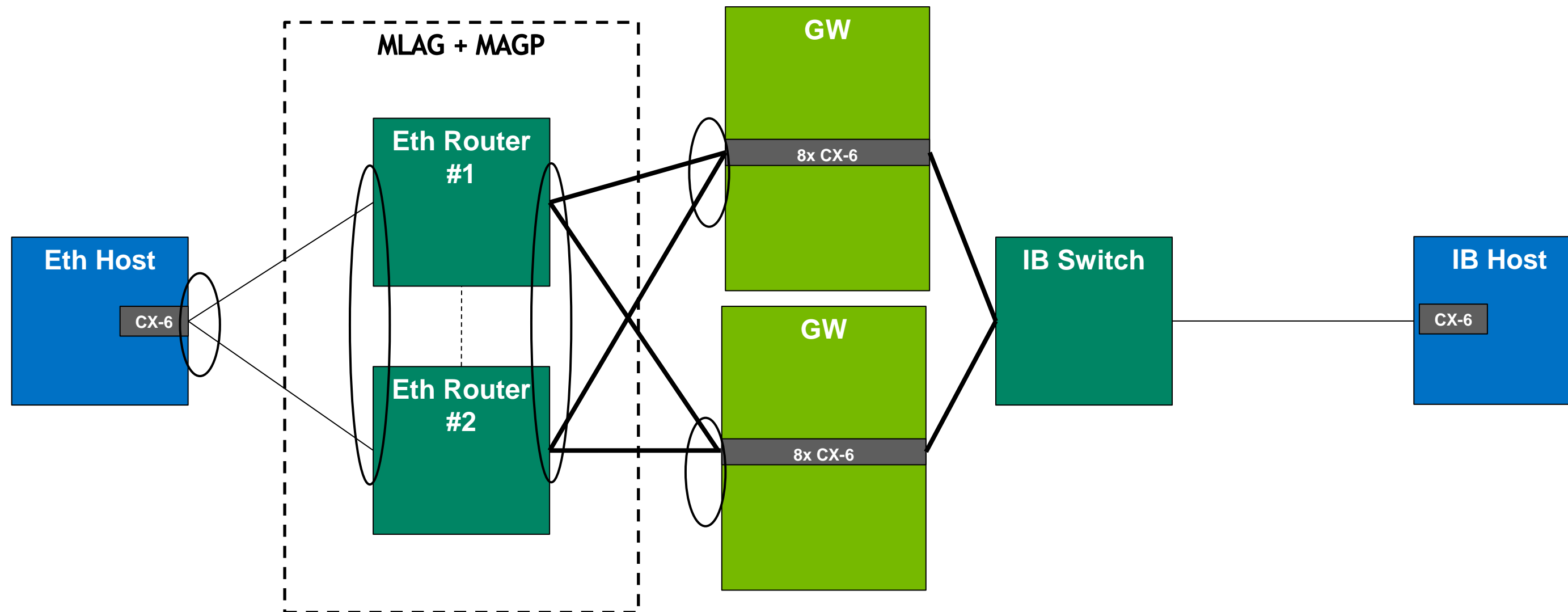


# BASIC TOPOLOGY - ONE GW





# BASIC TOPOLOGY - MULTIPLE GWS



# GATEWAY SYSTEM DETAILS

## Standard IP Gateway

Single InfiniBand network, single IP subnet

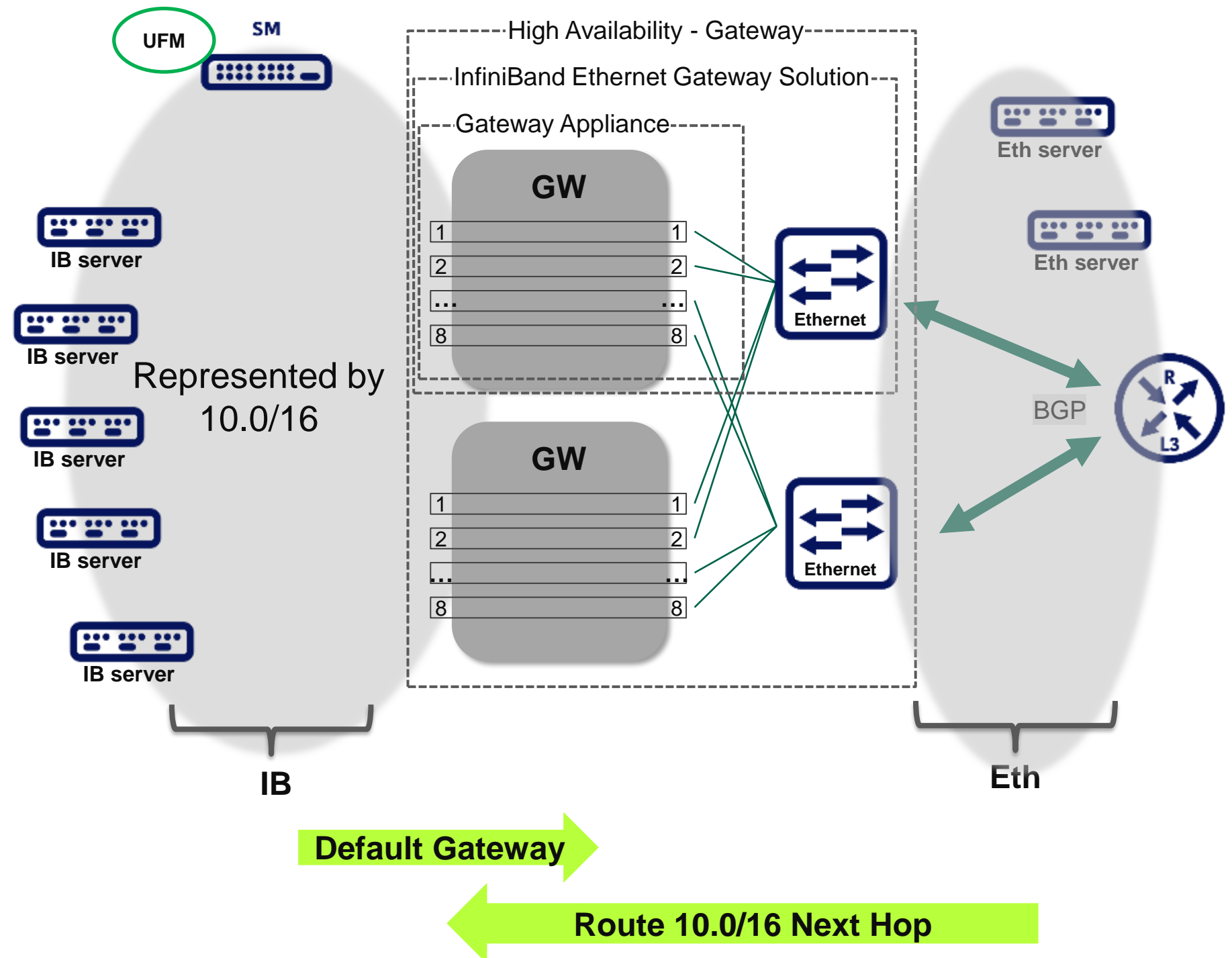
No use of Proxy ARP

IP Routing is supported via Ethernet switch

High availability and load balancing

CLI for appliance configuration

UFM for network configuration



# LOAD BALANCING & HIGH AVAILABILITY

## Standard IP Gateway

256 Gateway GIDs spread evenly among gateway IB ports

SM allocate LID per gateway GID

Standard ARP flow

ARP broadcast looking for GID of Gateway IP address

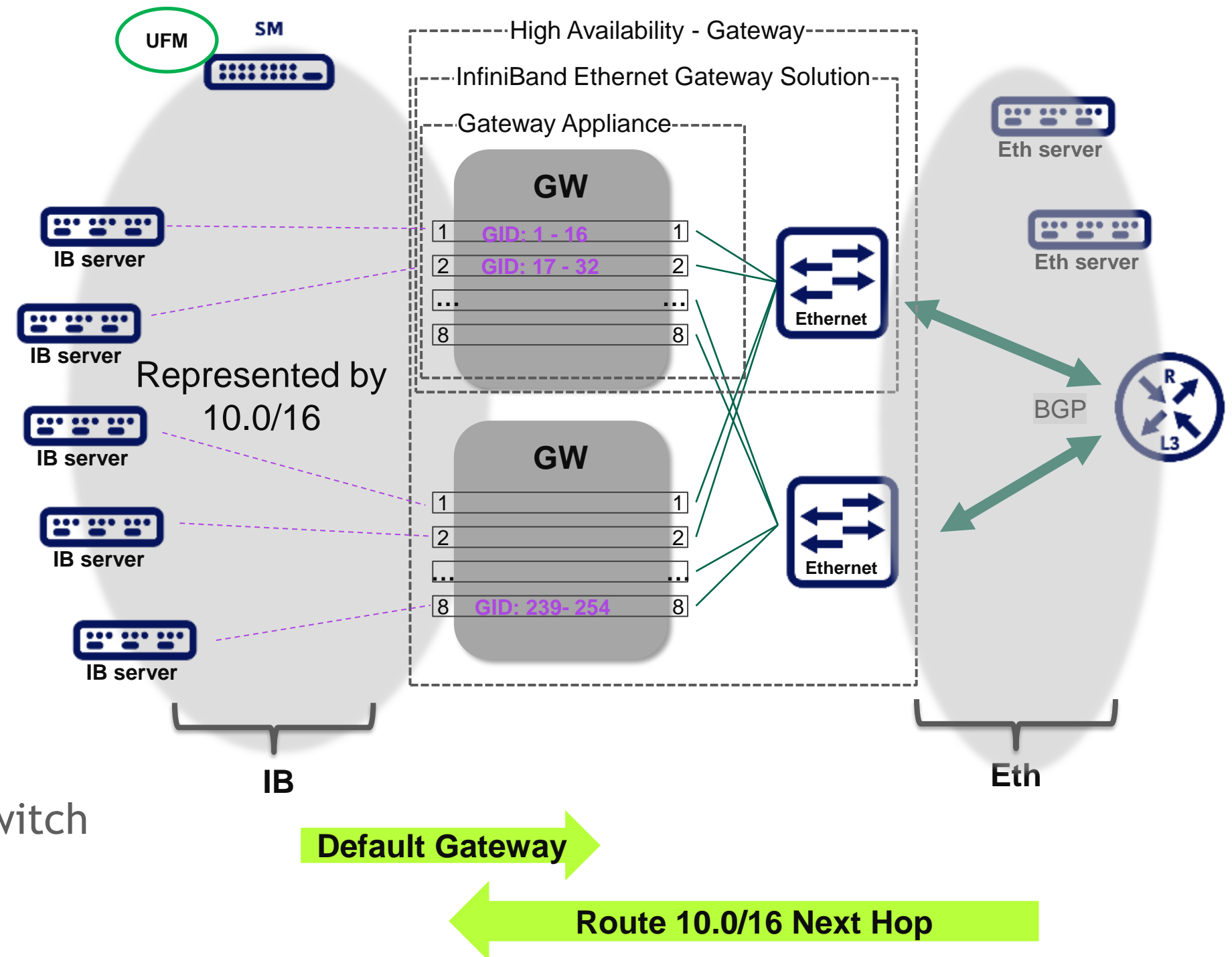
ARP reply with NEW-GID (based on originating host's IP)

Path query to determine LID

If an IB port fails, the Gateway GIDs are reassigned

Ethernet LAG among link between Gateway and Ethernet switch

LACP for Ethernet link selection



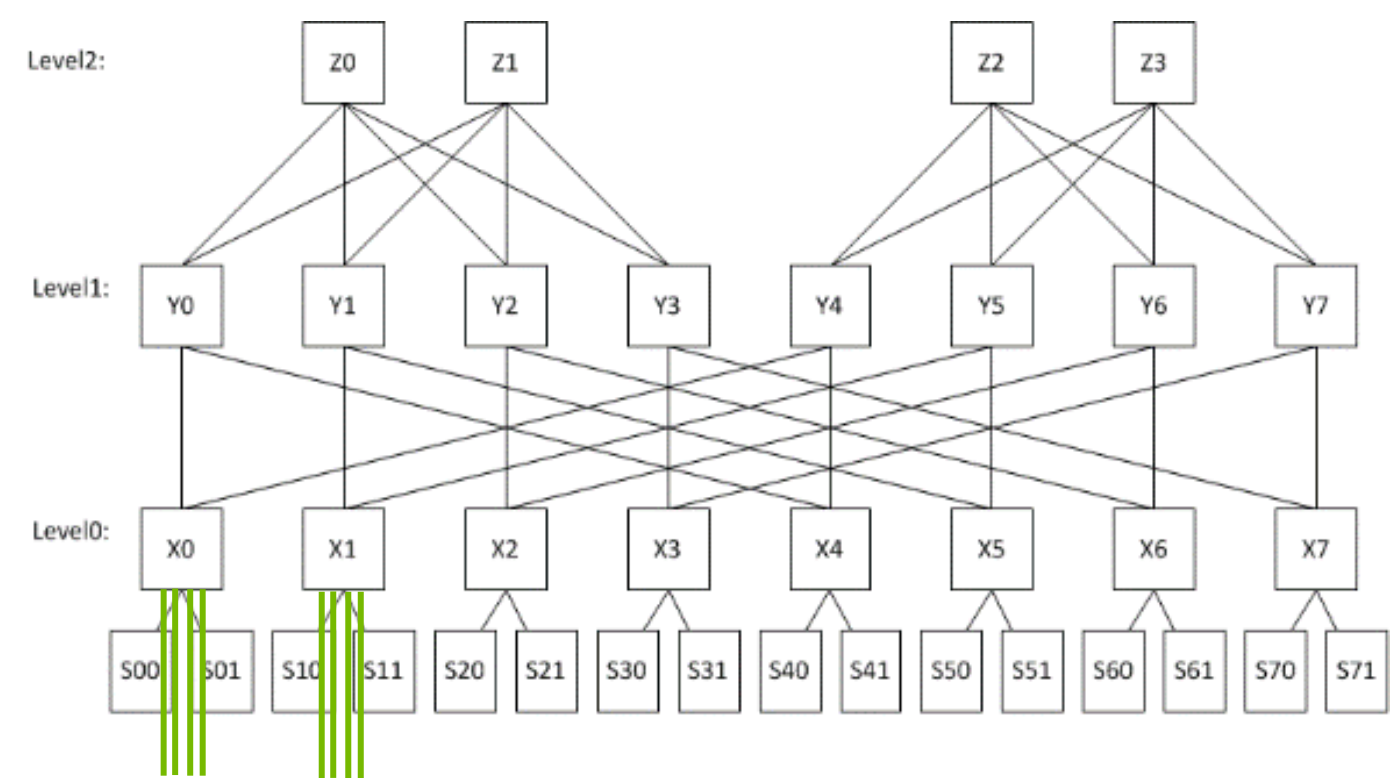


# GUIDELINES FOR CONNECTIVITY TO INFINIBAND NETWORK

Keep fair access from the InfiniBand nodes to the gateway

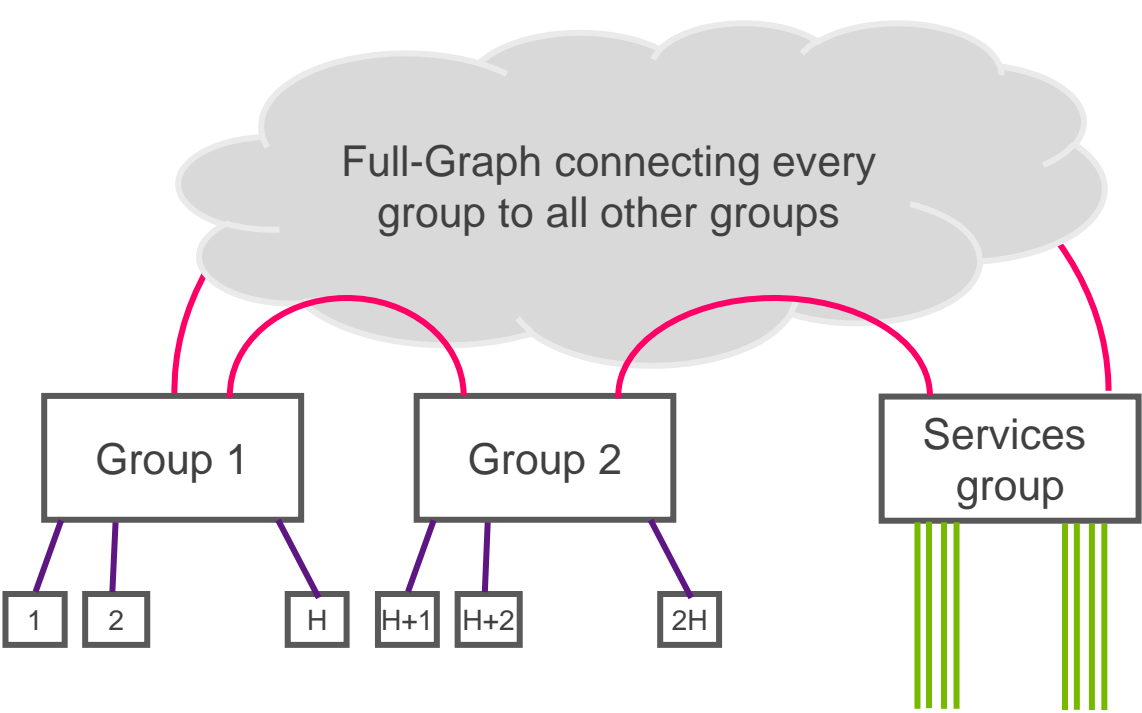
## Fat Tree

Connect to two separate leaves for high availability



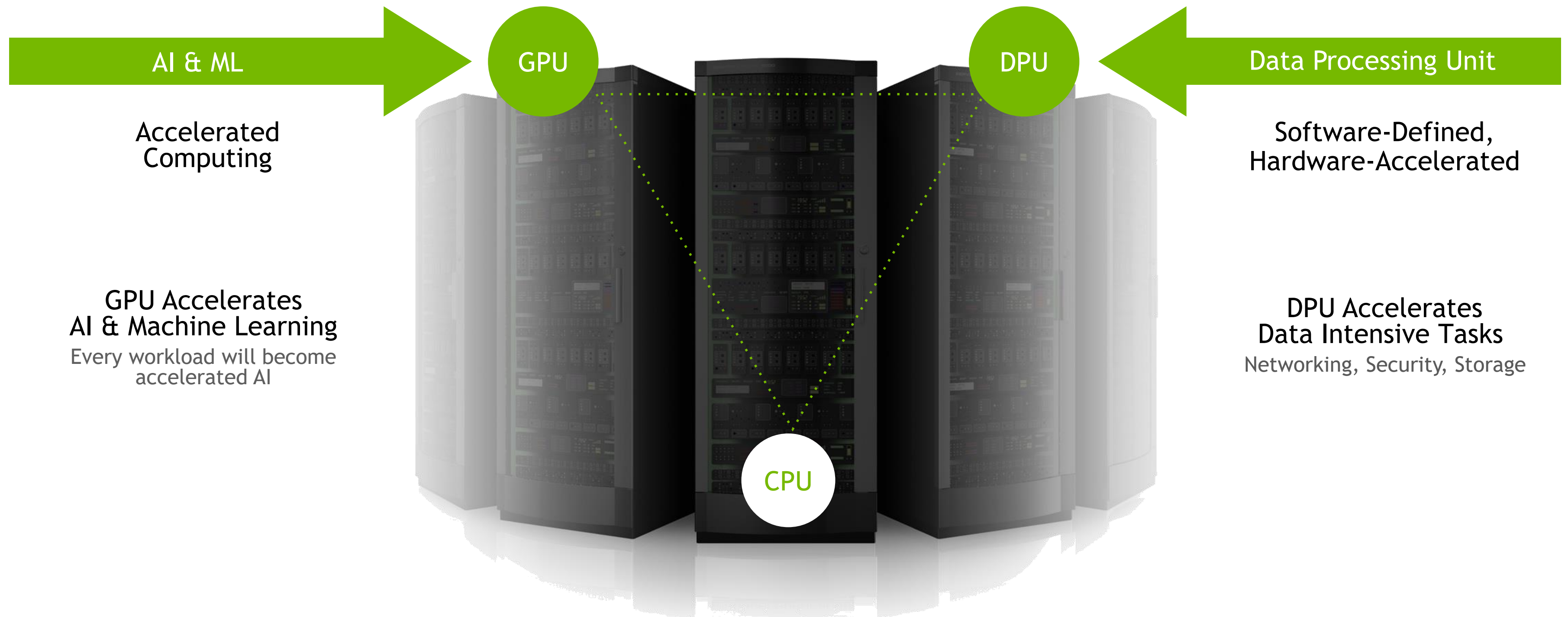
## DragonFly+

Connect to “services” island, where the storage is located



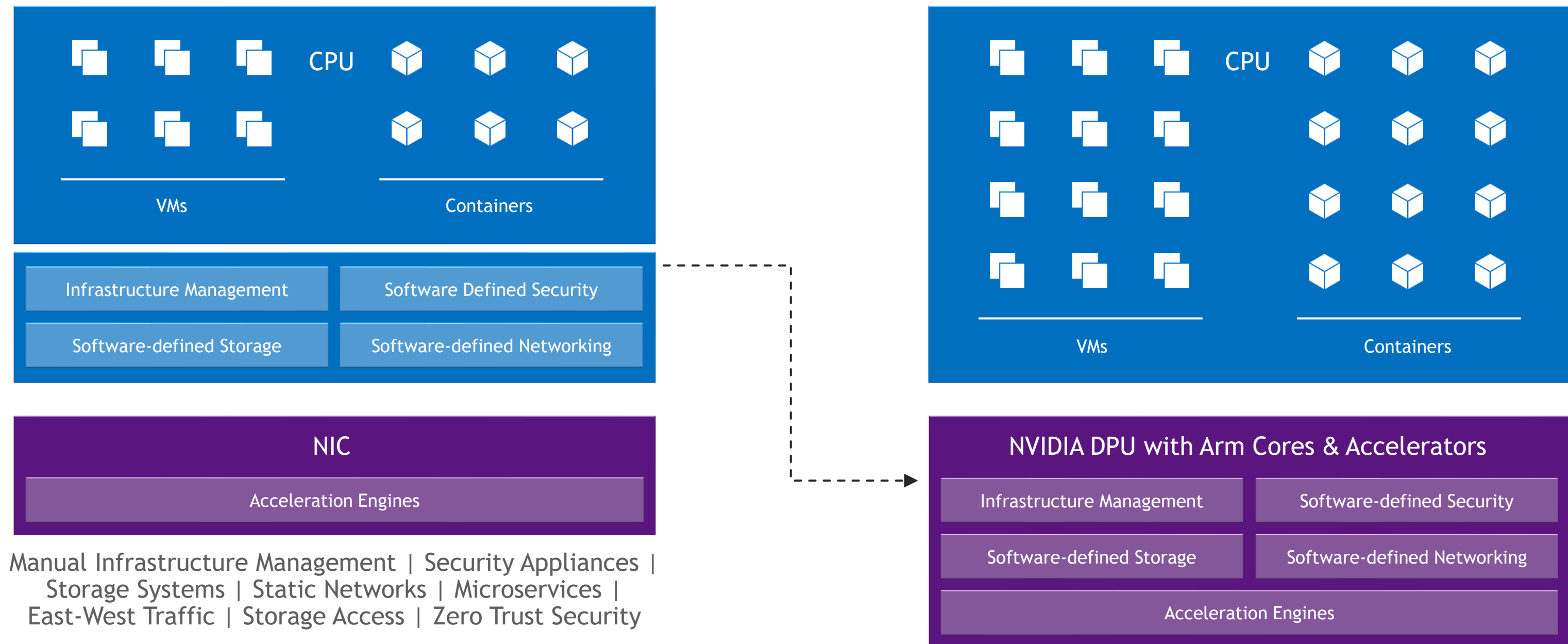
# REINVENTING THE DATA CENTER

The Data Center is the New Unit of Computing



# INTRODUCING THE DATA PROCESSING UNIT

Software-Defined, Hardware-Accelerated Data Center Infrastructure-on-a-Chip





# NVIDIA BLUEFIELD-2 DATA PROCESSING UNIT

## Data Center Infrastructure on a Chip

Up to 200Gb/s Ethernet and InfiniBand, PAM4/NRZ

ConnectX-6 Dx inside

8 Arm A72 CPUs subsystem - up to 2.75GHz

- 8MB L2 cache, 6MB L3 cache in 4 Tiles

- Fully coherent low-latency interconnect

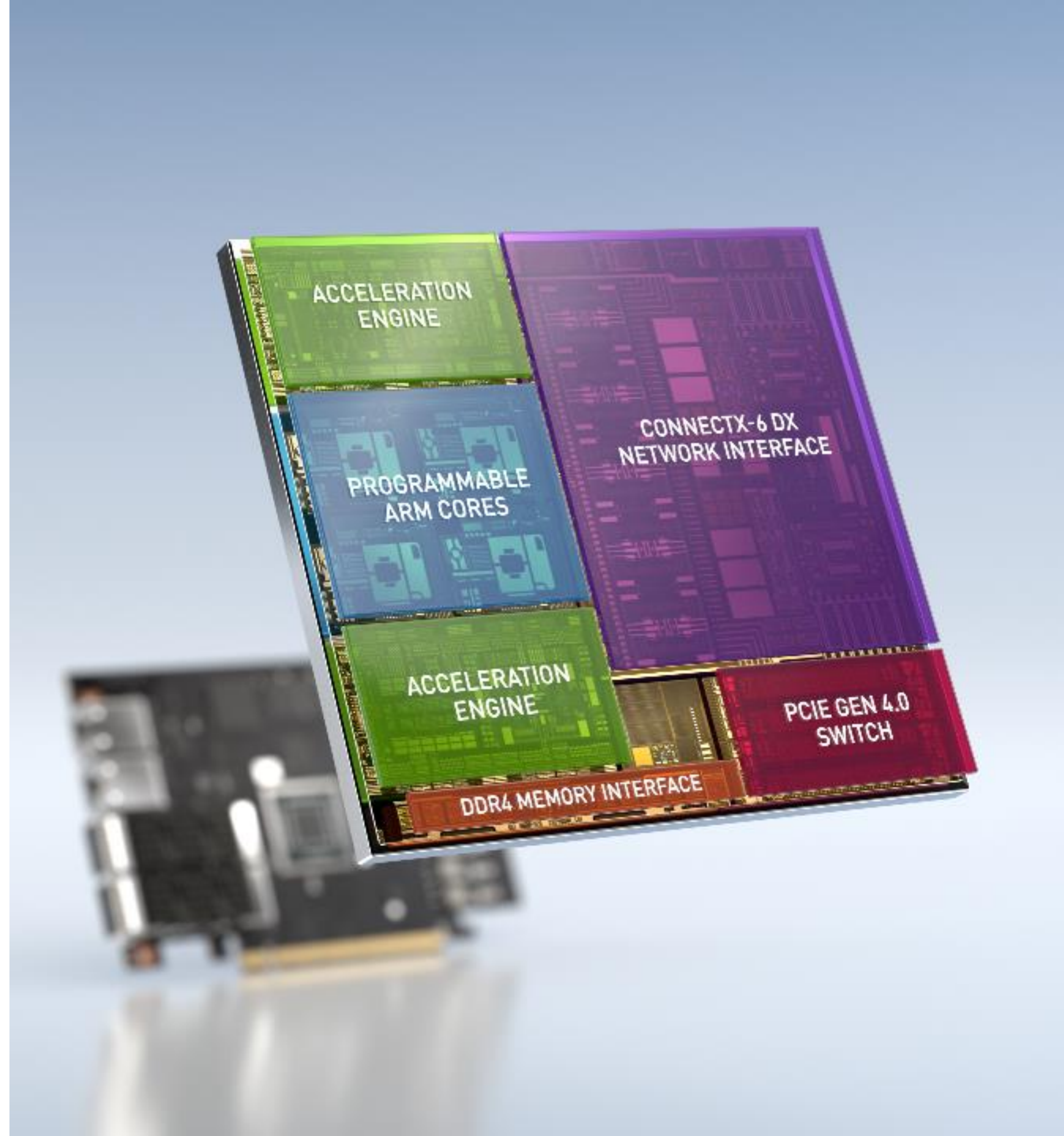
Integrated PCIe switch, 16x Gen4.0

- PCIe root complex or end point modes

Single DDR4 channel

1GbE Out-of-Band management port

Accelerated security, storage, networking



# BLUEFIELD-2X DATA PROCESSING UNIT

## AI-Powered DPU

200 Gb/s BlueField-2 augmented by Ampere GPU

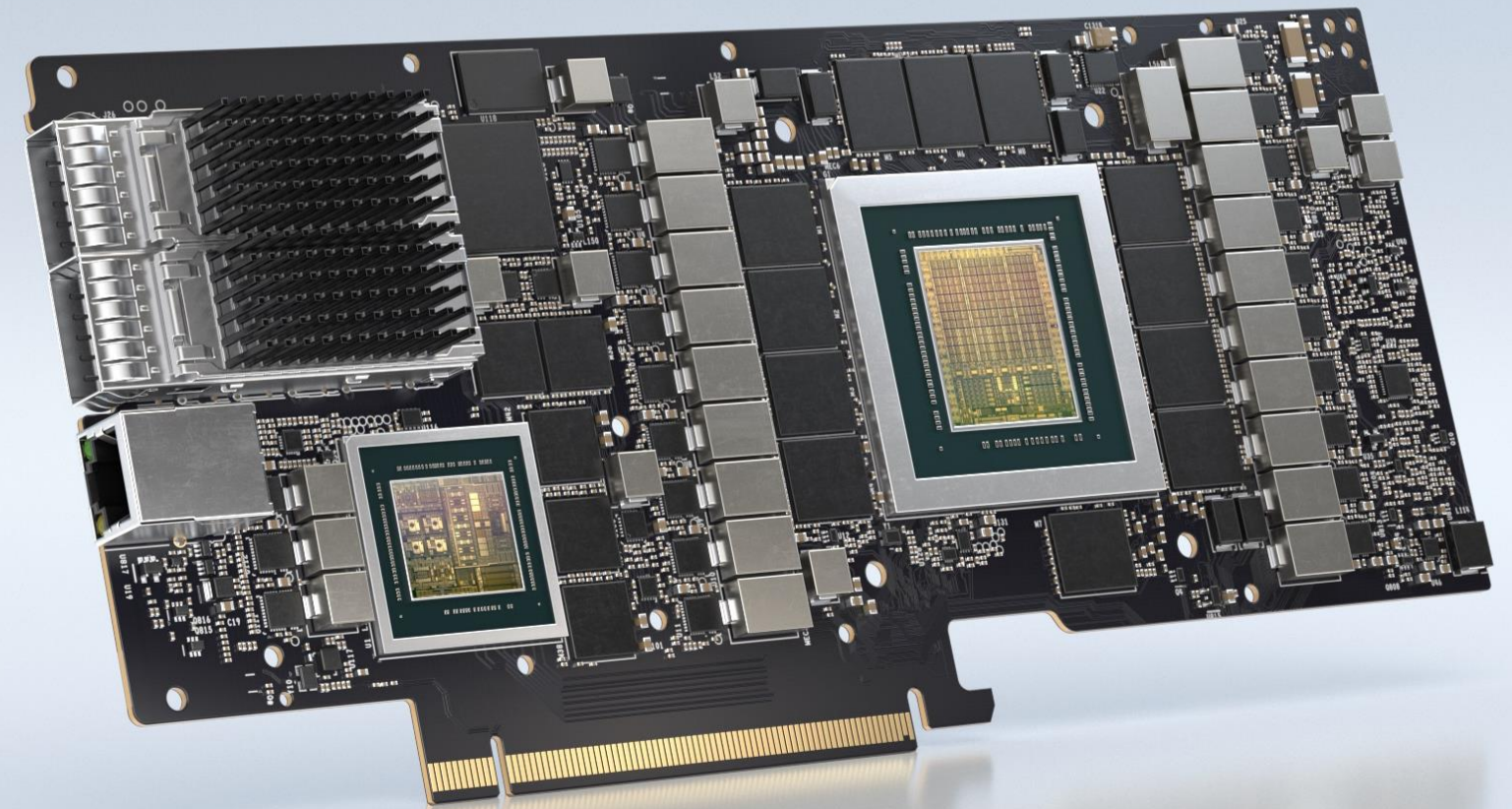
Enhanced the DPU with AI capabilities

Scale out computing performance with GPUDirect and CUDA

Tighter security across the PCIe bus

Apply AI to real time network traffic

- Anomaly detection & automated response
- Traffic shaping/steering
- Dynamic security orchestration





# NVIDIA DOCA

## Data-Center-Infrastructure-on-a-Chip Architecture



### COMMUNITY of DEVELOPERS

SDK for  
ecosystem partners,  
academia,  
community



### ACCELERATE TTM

Leverages open-source and  
industry standards (DPDK, P4);  
NGC-certified



### COMPETITIVE EDGE

Best performance;  
out-of-the-box experience;  
libraries with special  
capabilities



### LONG-TERM COMMITMENT

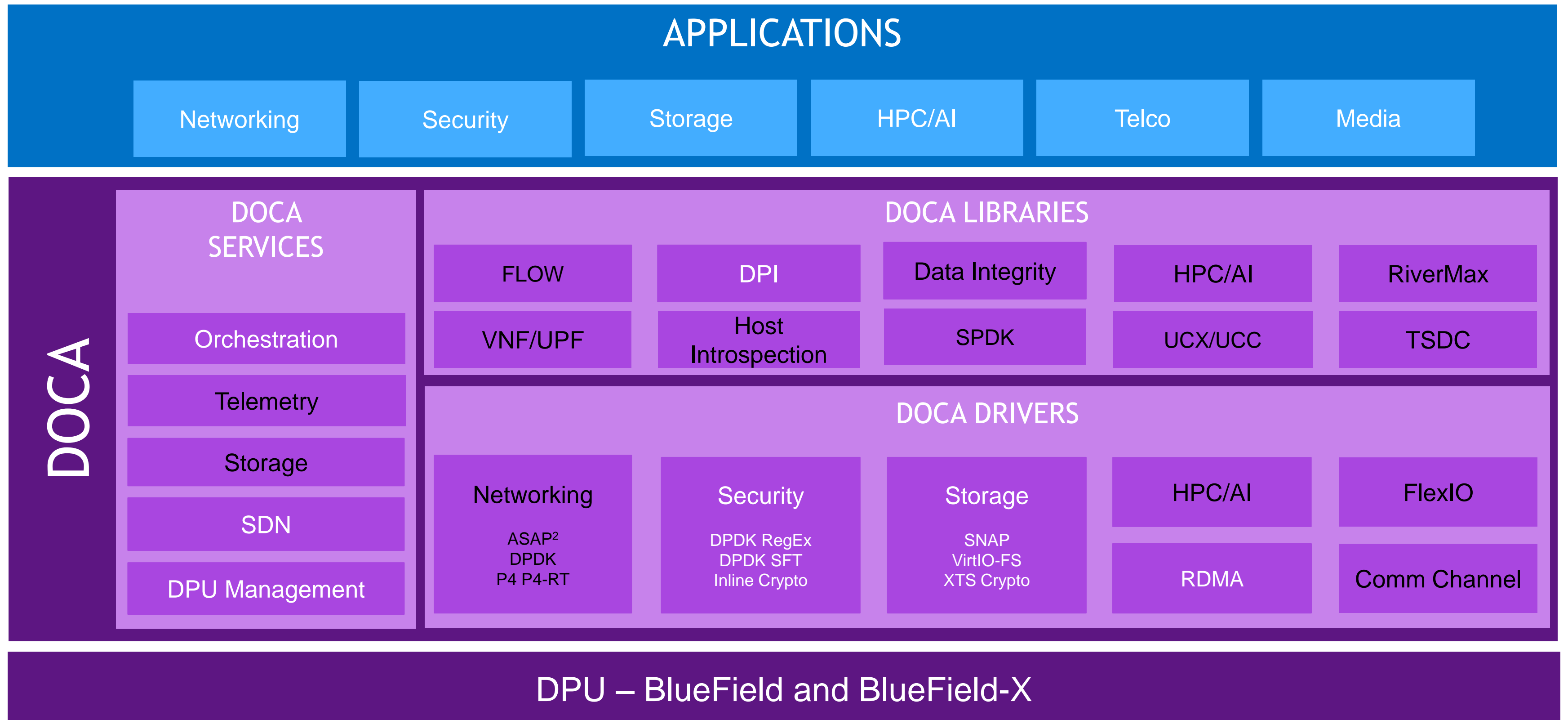
Backward and forward  
compatibility;  
consistency with  
performance improvements

DOCA is for DPUs what CUDA is for GPUs

<https://developer.nvidia.com/networking/doca>



# DOCA STACK



# ANNOUNCING NVIDIA BLUEFIELD-3 DPU

First 400Gb/s Data Processing Unit

22 Billion Transistors

400Gb/s Ethernet & InfiniBand Connectivity

Powerful CPU - 16x Arm A78 Cores

400Gb/s Crypto Acceleration

300 Equivalent x86 Cores

Offloads and Accelerates Data Center Infrastructure

Isolates Application from Control and Management Plane

DATA PATH ACCELERATOR

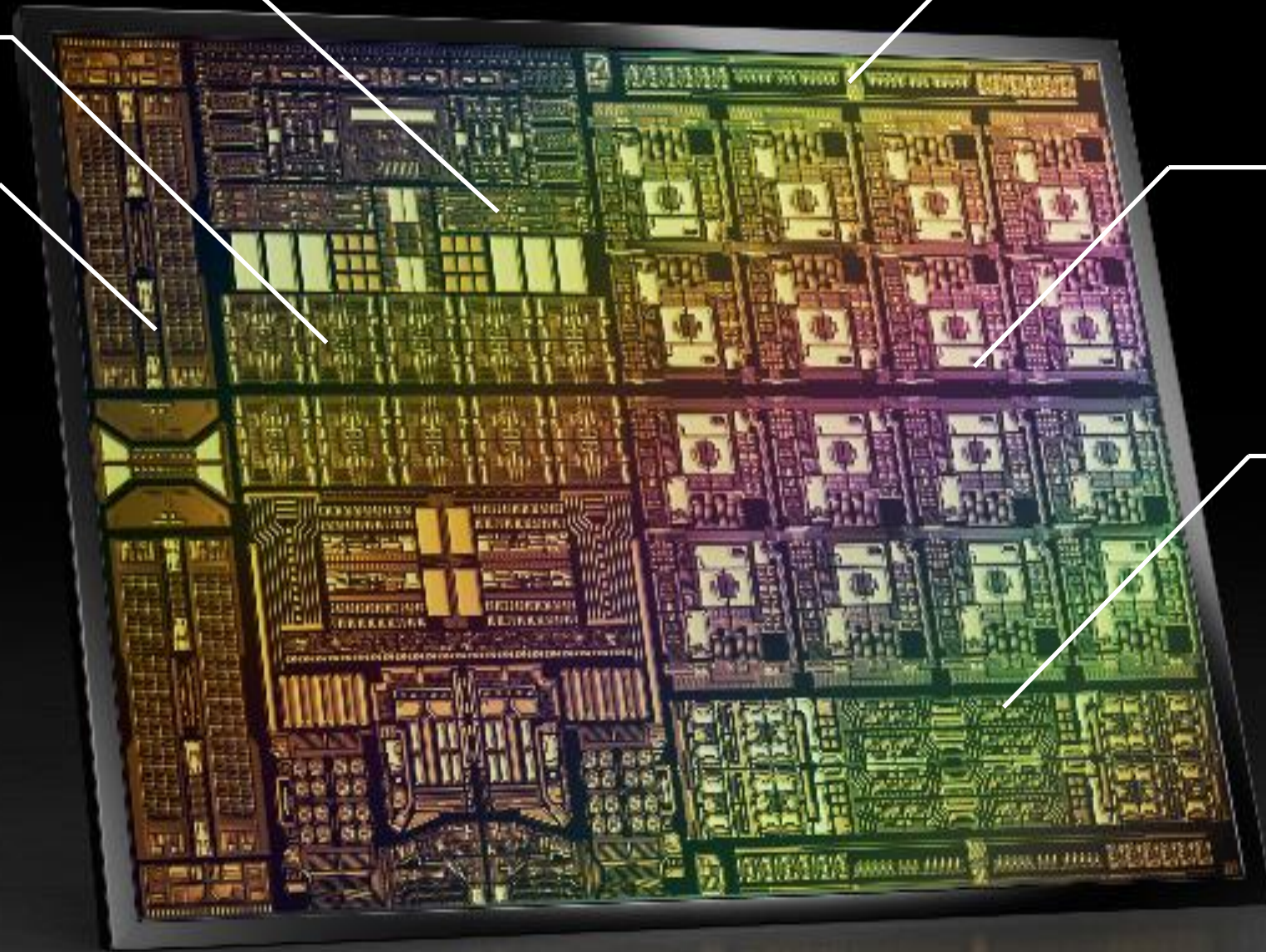
CONNECTX-7

PCIe GEN 5.0

DDR5 MEMORY INTERFACE

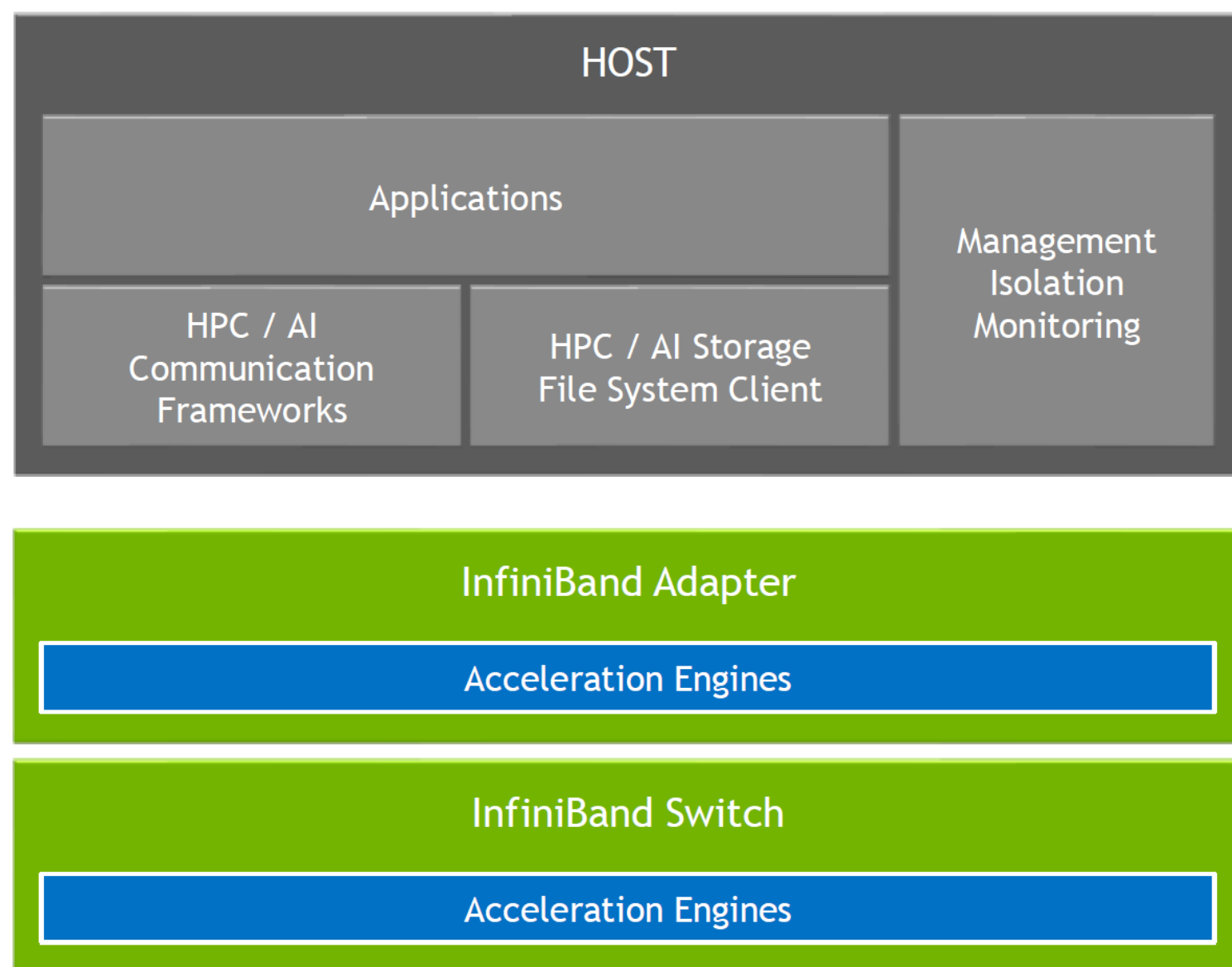
ARM CORES

ACCELERATION  
ENGINES

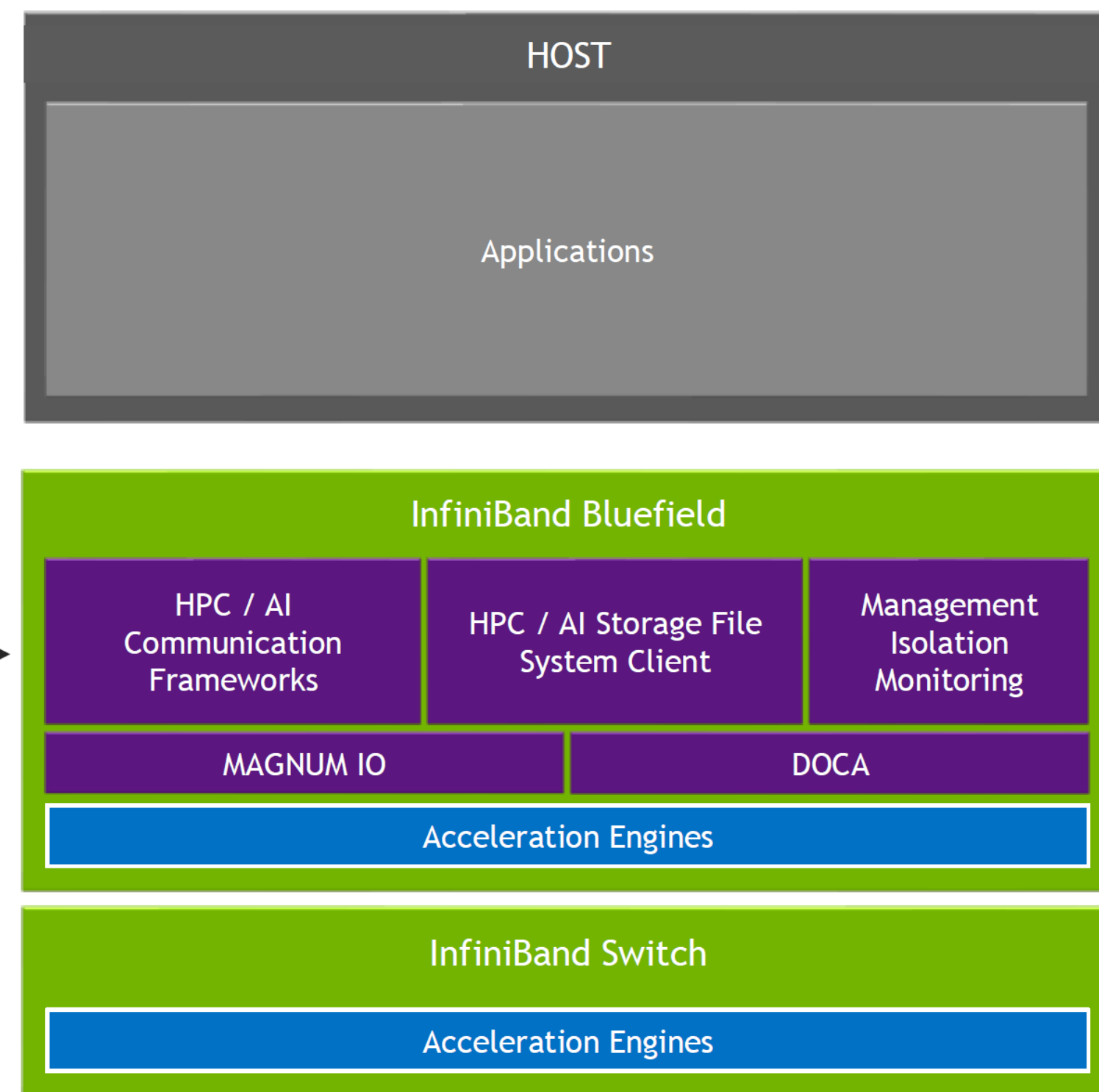


# CLOUD NATIVE SUPERCOMPUTING INFRASTRUCTURE

## TRADITIONAL SUPERCOMPUTING



## CLOUD NATIVE SUPERCOMPUTING





# MULTI TENANT ISOLATION

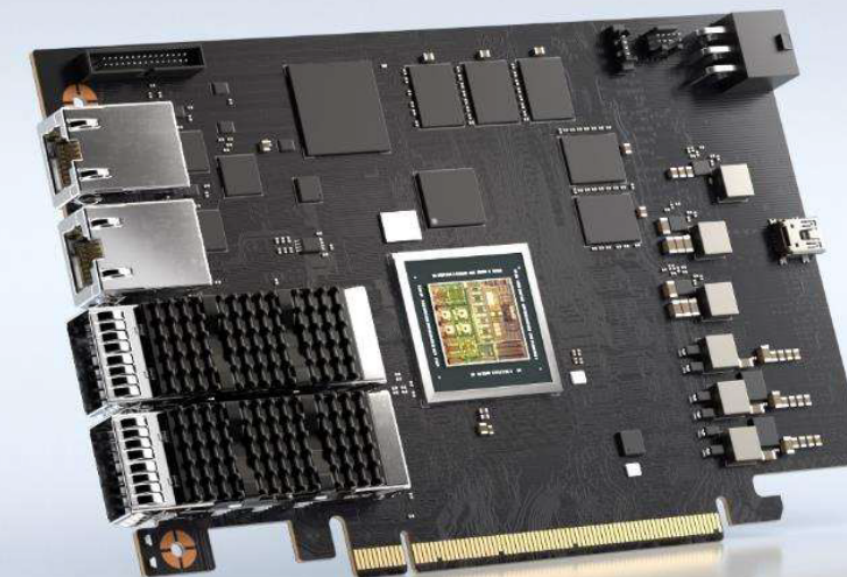
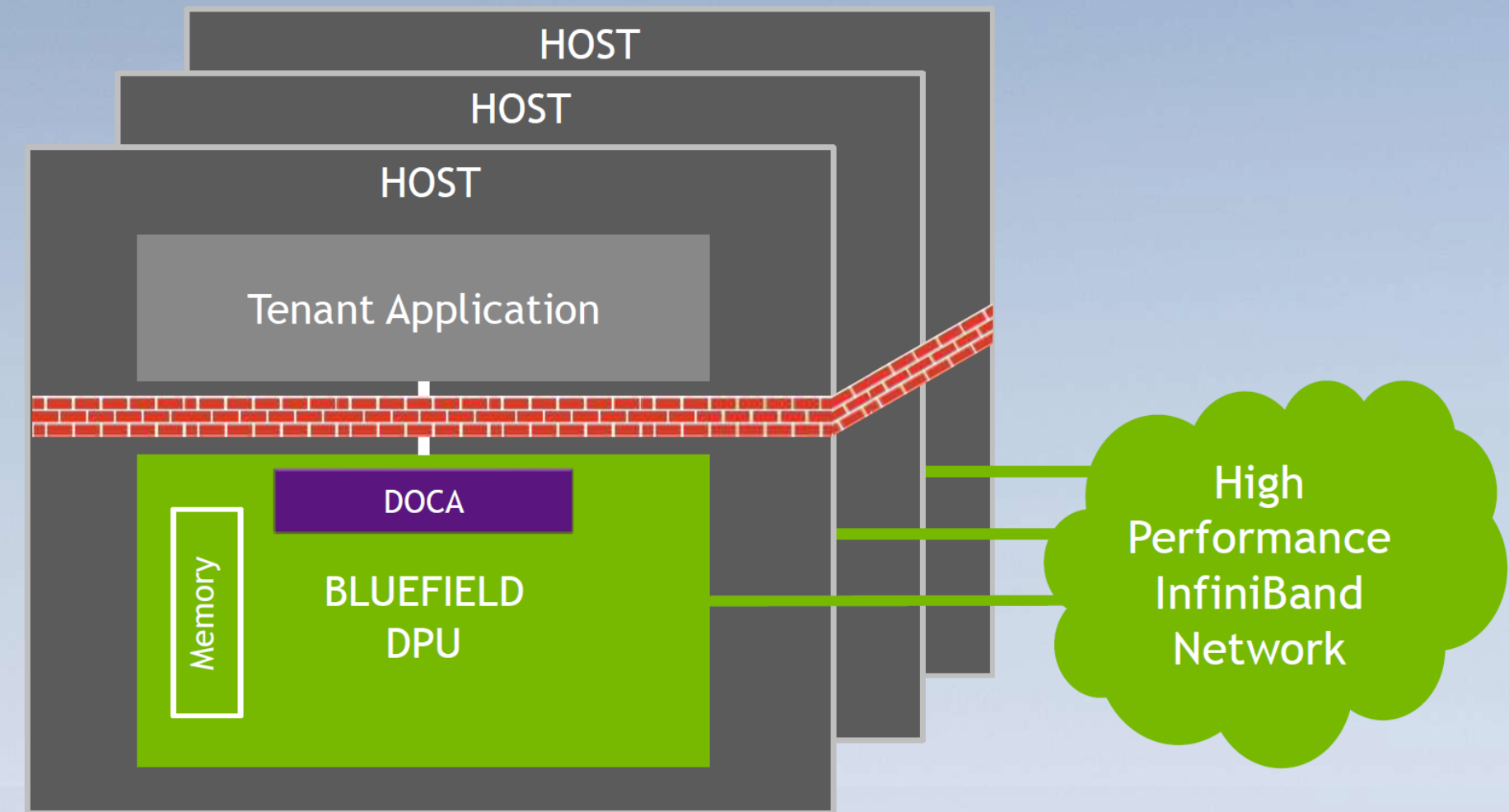
## Zero trust architecture

Secured Network infrastructure and configuration

Storage virtualization

Tenant Service Level Agreement (SLA)

32K concurrent isolated users on single subnet



# BROAD PARTNER ECOSYSTEM

Hybrid Cloud Compatibility | No Fork-Lift Upgrades | No Vendor Lock-In

CANONICAL



vmware®



FORTINET®



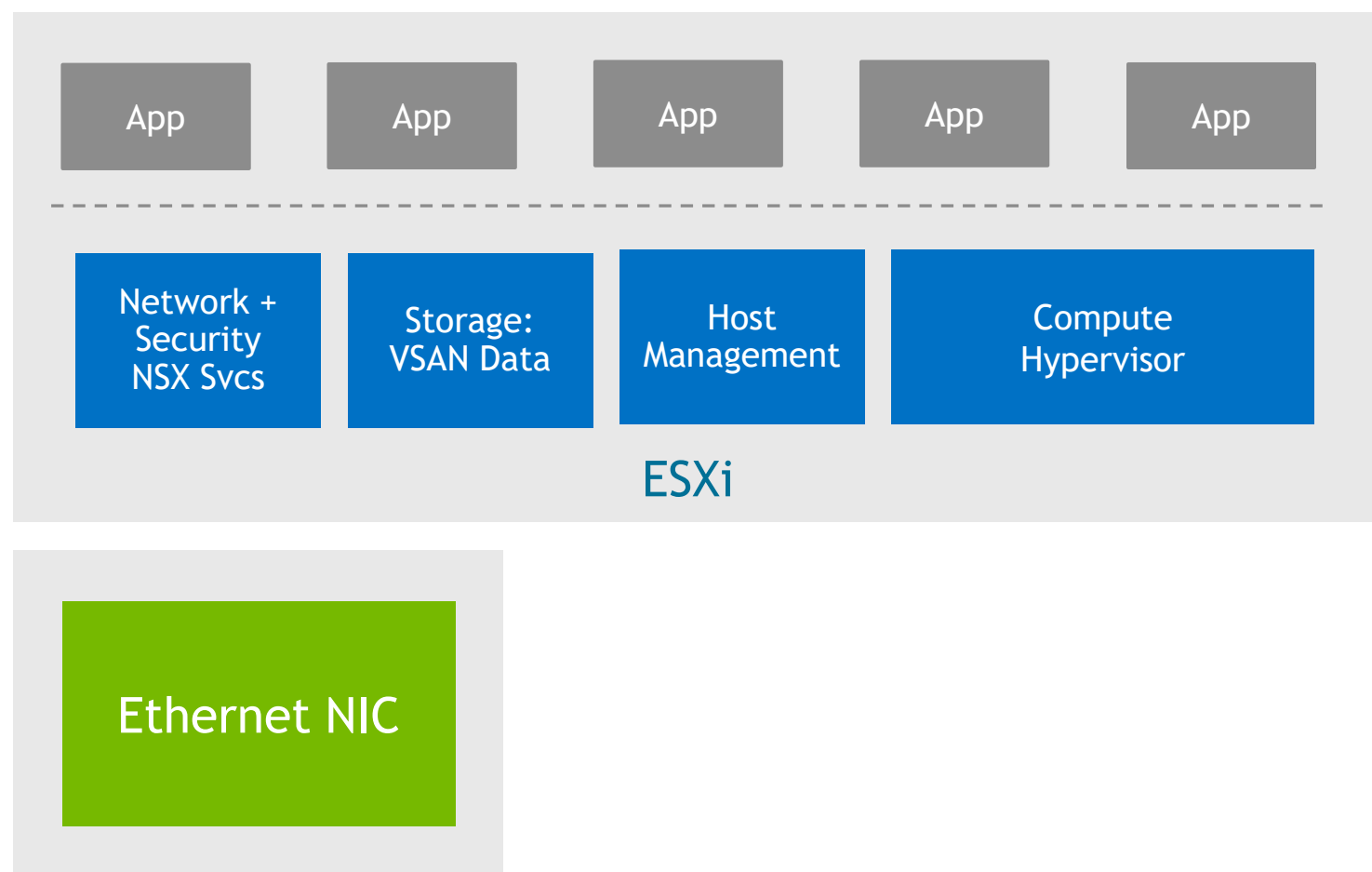
JUNIPER  
NETWORKS



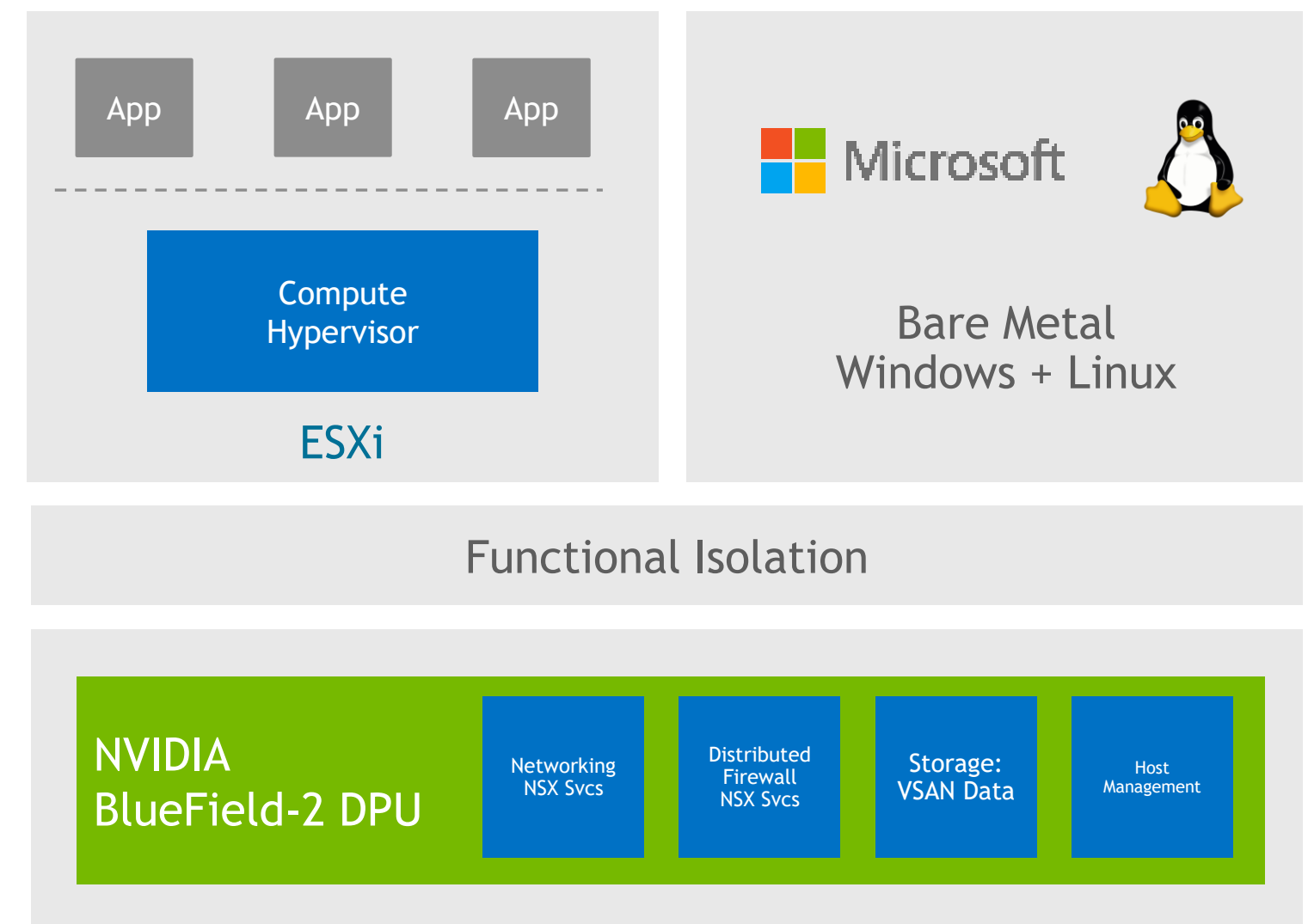
# PROJECT MONTEREY (VMWARE + NVIDIA)

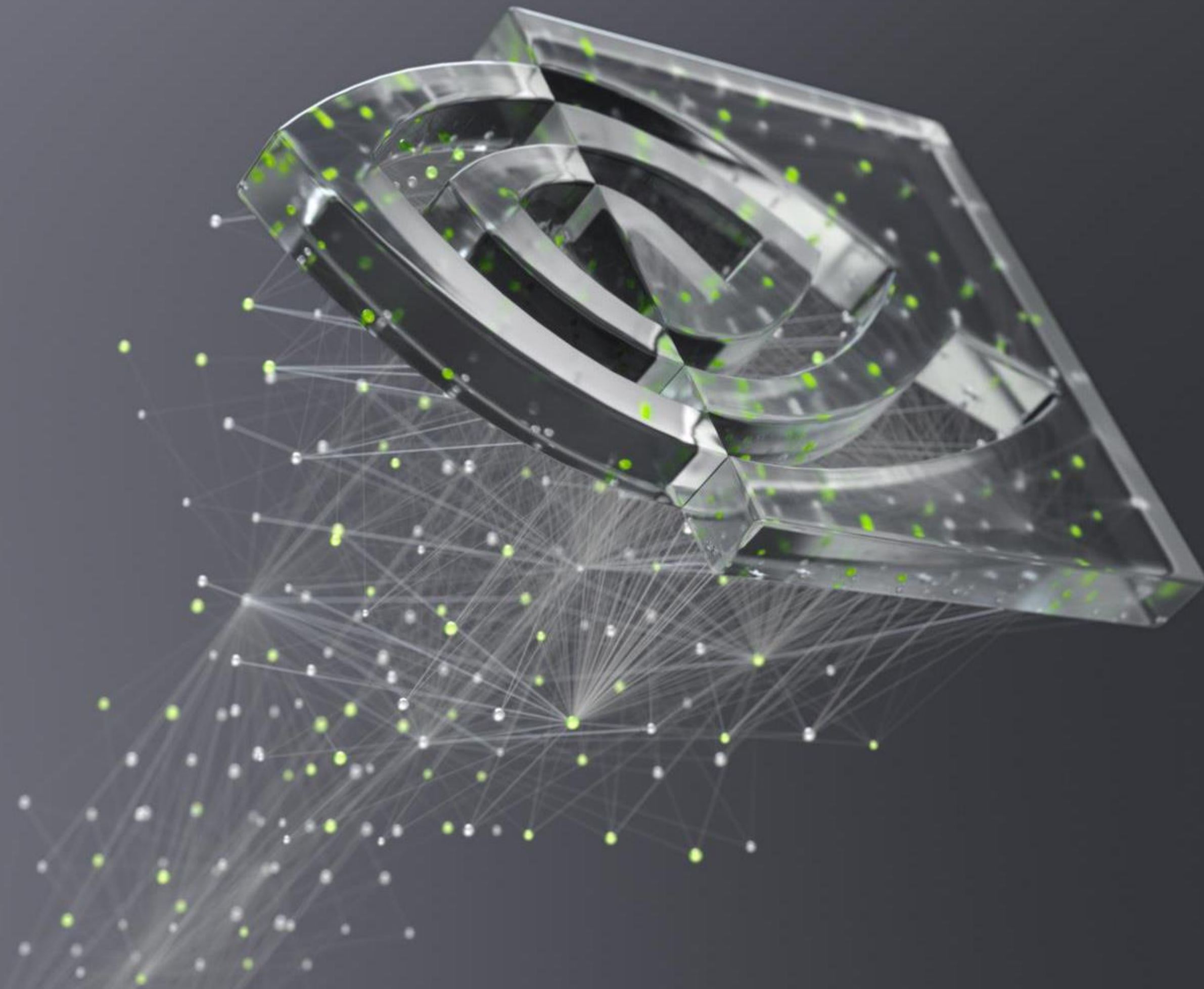
Run Modern Workloads Efficiently Over New Composable, Disaggregated Infrastructure

Today's Environment



Project Monterey

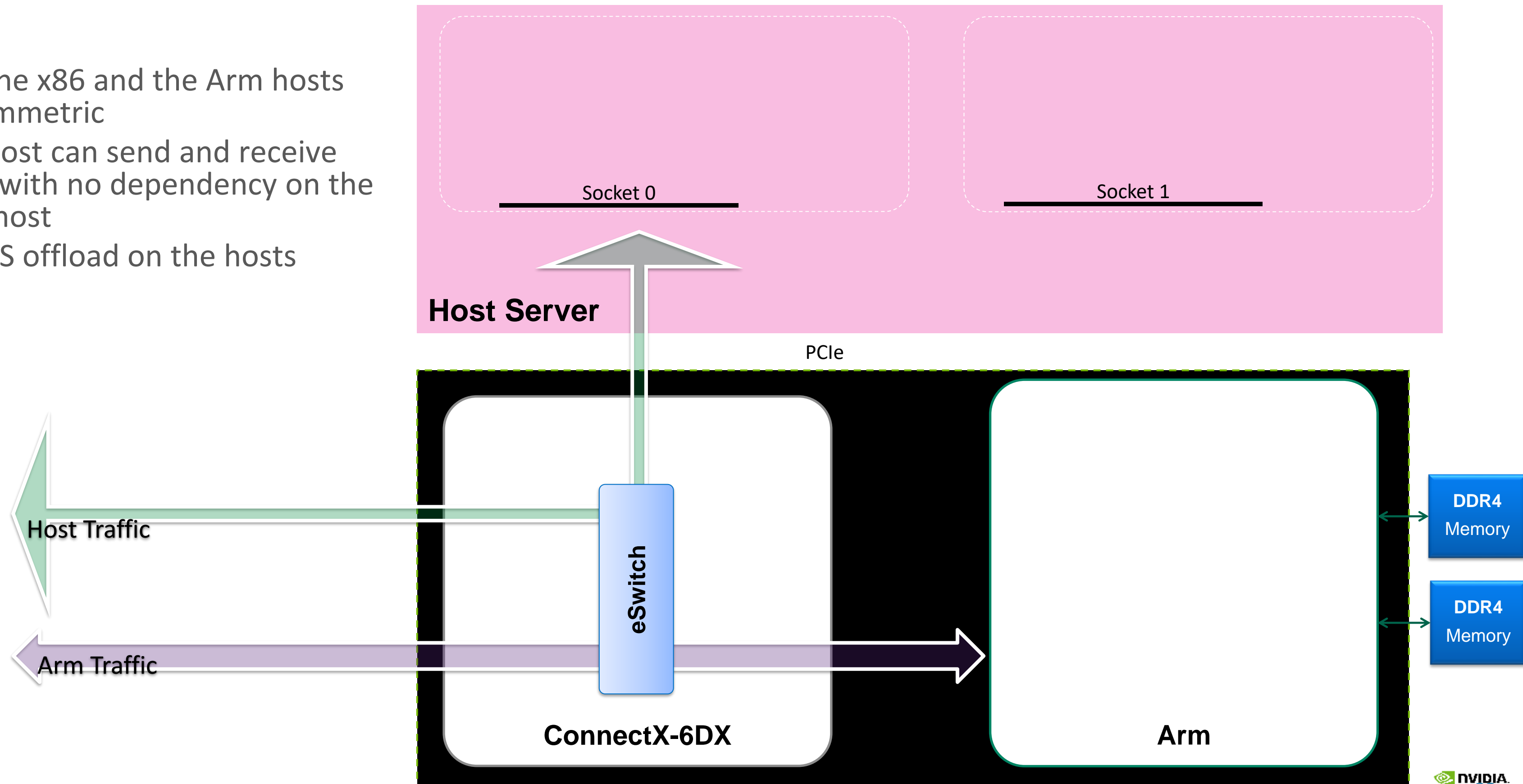






# Separated Hosts Mode (Default Configuration)

- Both the x86 and the Arm hosts are symmetric
- Each host can send and receive traffic with no dependency on the other host
- No OVS offload on the hosts



# Embedded Mode(Arm Switch Ownership)

- OVS (with ASAP<sup>2</sup>) runs on the Arm cores
- All host traffic is controlled by the vswitch
- But can be offloaded to eSwitch hardware.

