



INFINIBAND NDR 400GB AND DPU FOR NEXT GEN HPC/AI ARCHITECTURES

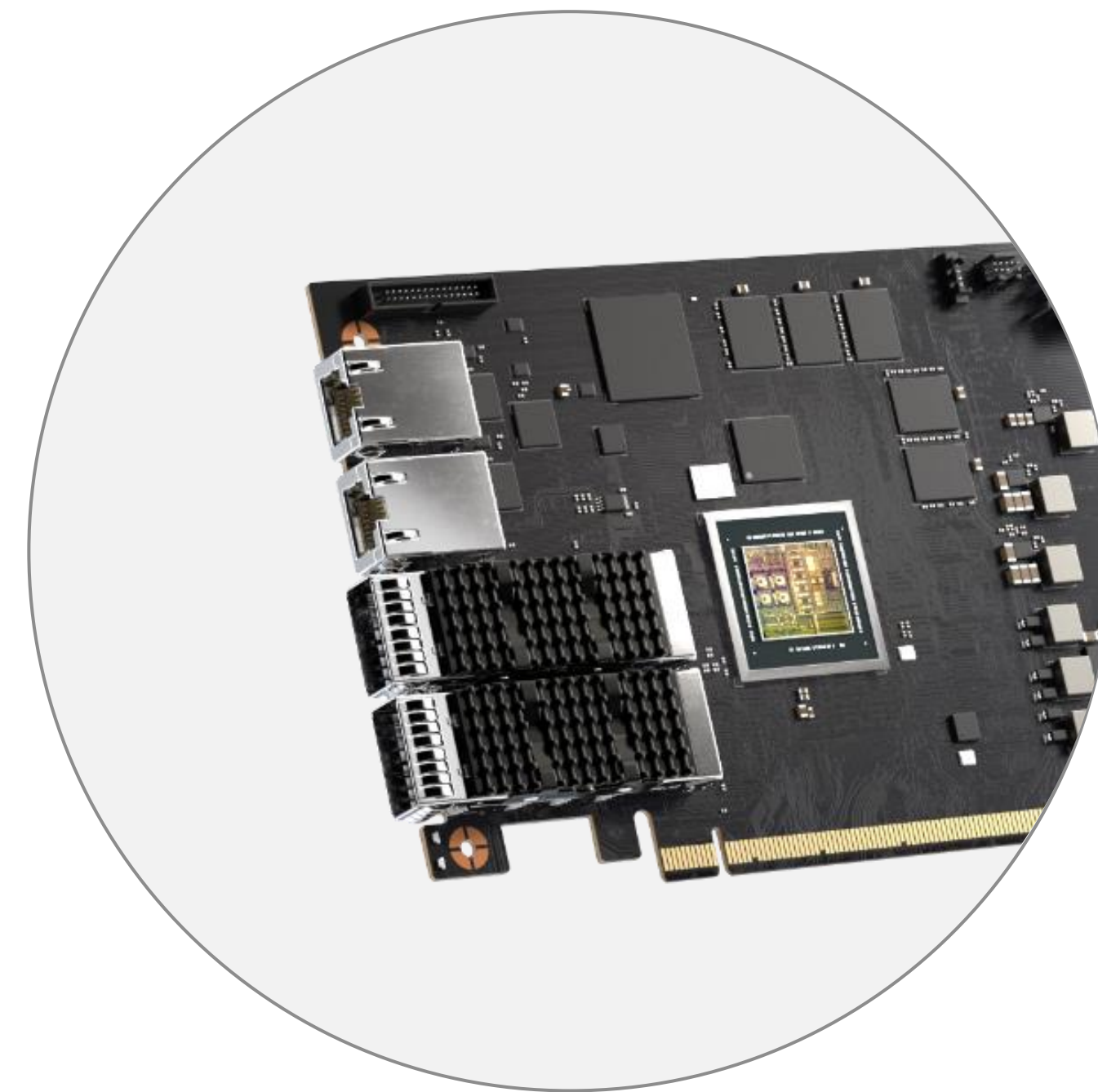
RAMIROA@NVIDIA.COM - APRIL 2021

NVIDIA QUANTUM-2 INFINIBAND PLATFORM



ConnectX-7 Adapter

400G InfiniBand
PCIe Gen5
Programmable Datapath
In-Network Computing



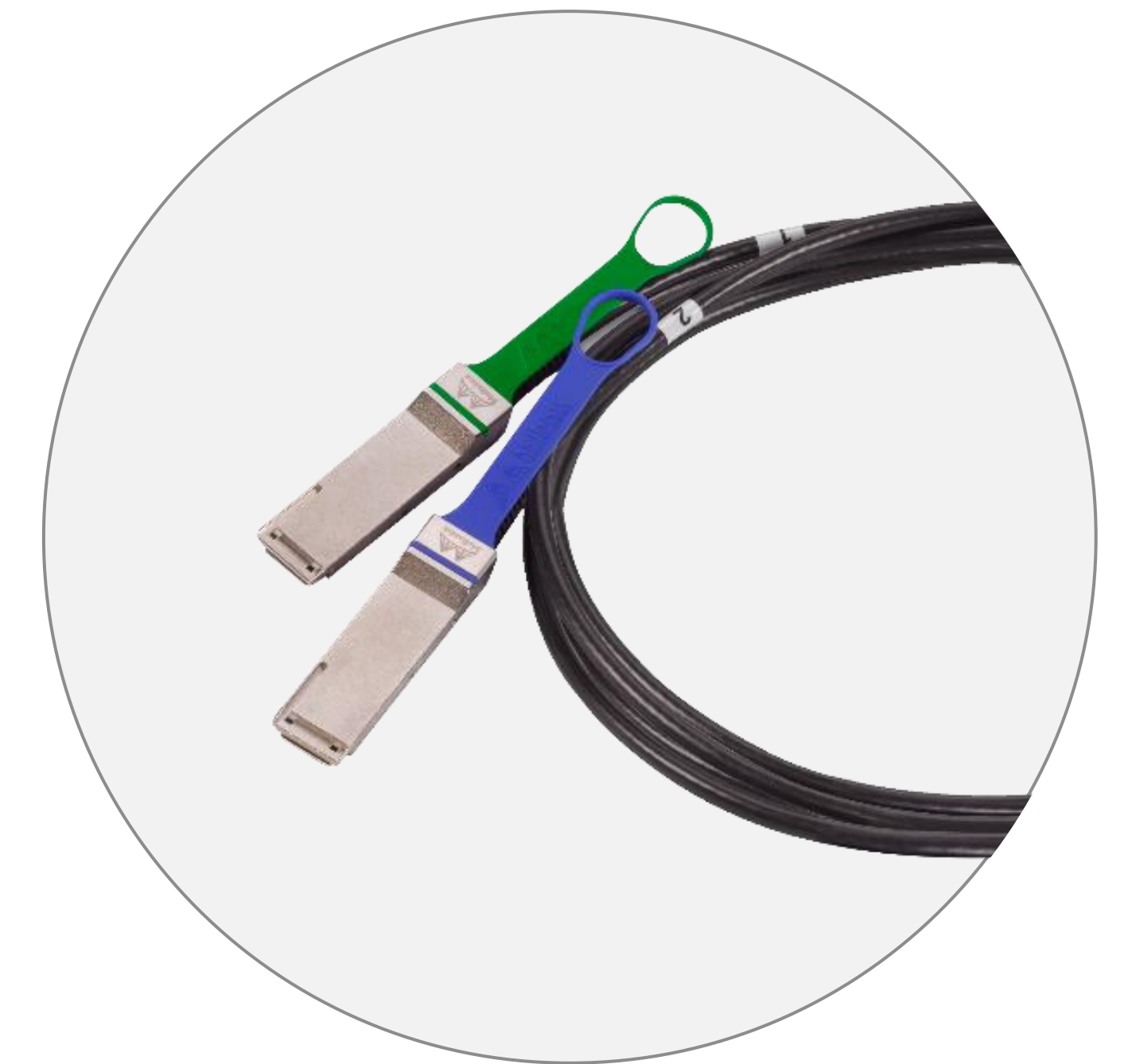
BlueField-3 DPU

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



Quantum-2 Switch

64-ports 400G InfiniBand
128-ports 200G
In-Network Computing



Cable

Copper Cables
Active Copper Cables
Optical Transceivers

NVIDIA QUANTUM NDR 400G INFINIBAND SYSTEMS

In-Network Computing Accelerated
Network for Cloud-Native
Supercomputing at Any Scale

2x

Data Throughput
400 Gigabits per Second

32x

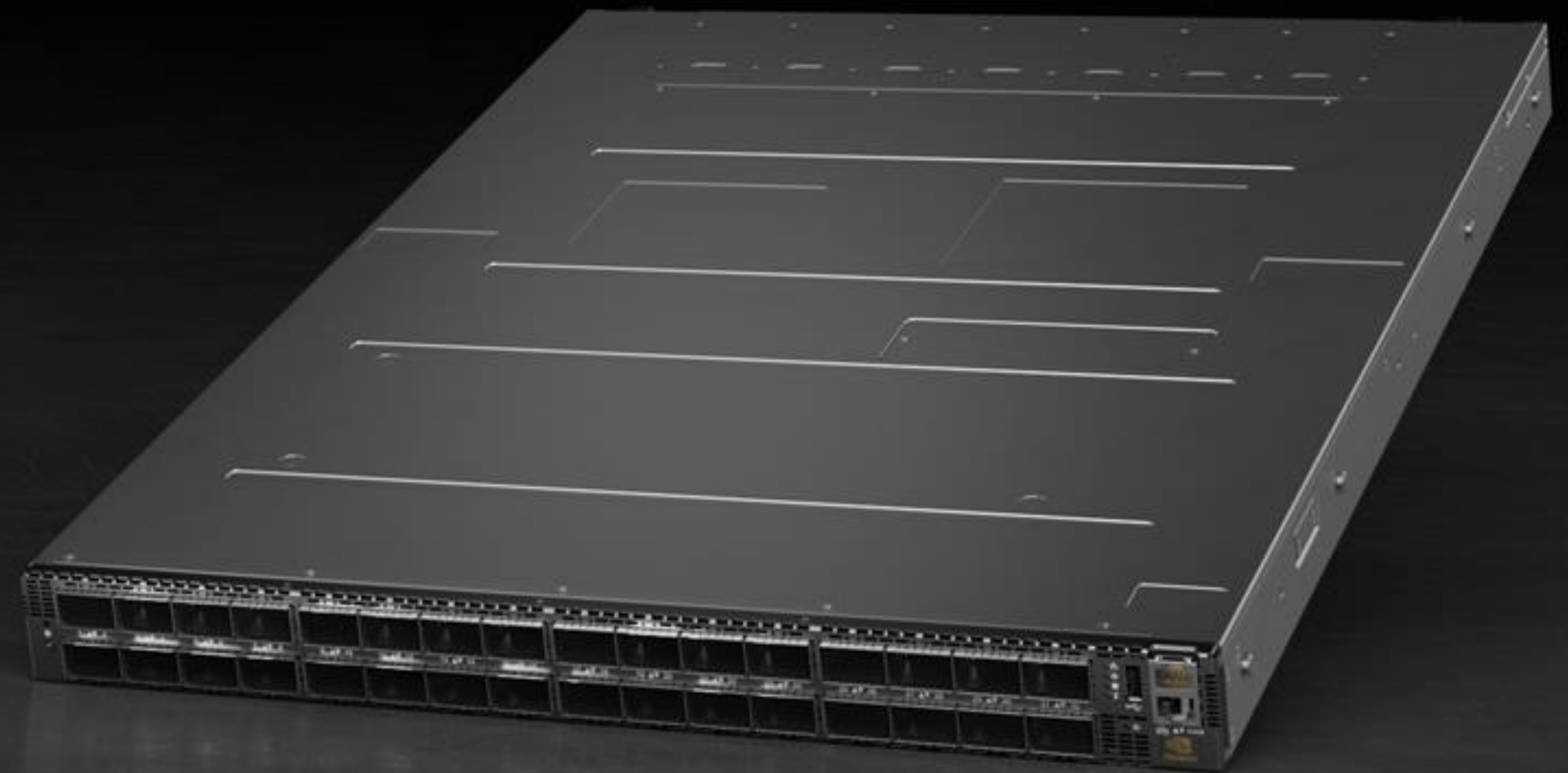
More AI Acceleration
SHARP In-Network Computing

6.5x

Higher Scalability
>1M nodes with DF+ 3 hops

5x

Switch System Capacity
>1.6 Petabit per Second



NVIDIA QUANTUM NDR 400G INFINIBAND SYSTEMS

In-Network Computing Accelerated
Network for Cloud-Native
Supercomputing at Any Scale

15%

Faster Deep Learning
Recommendations

17%

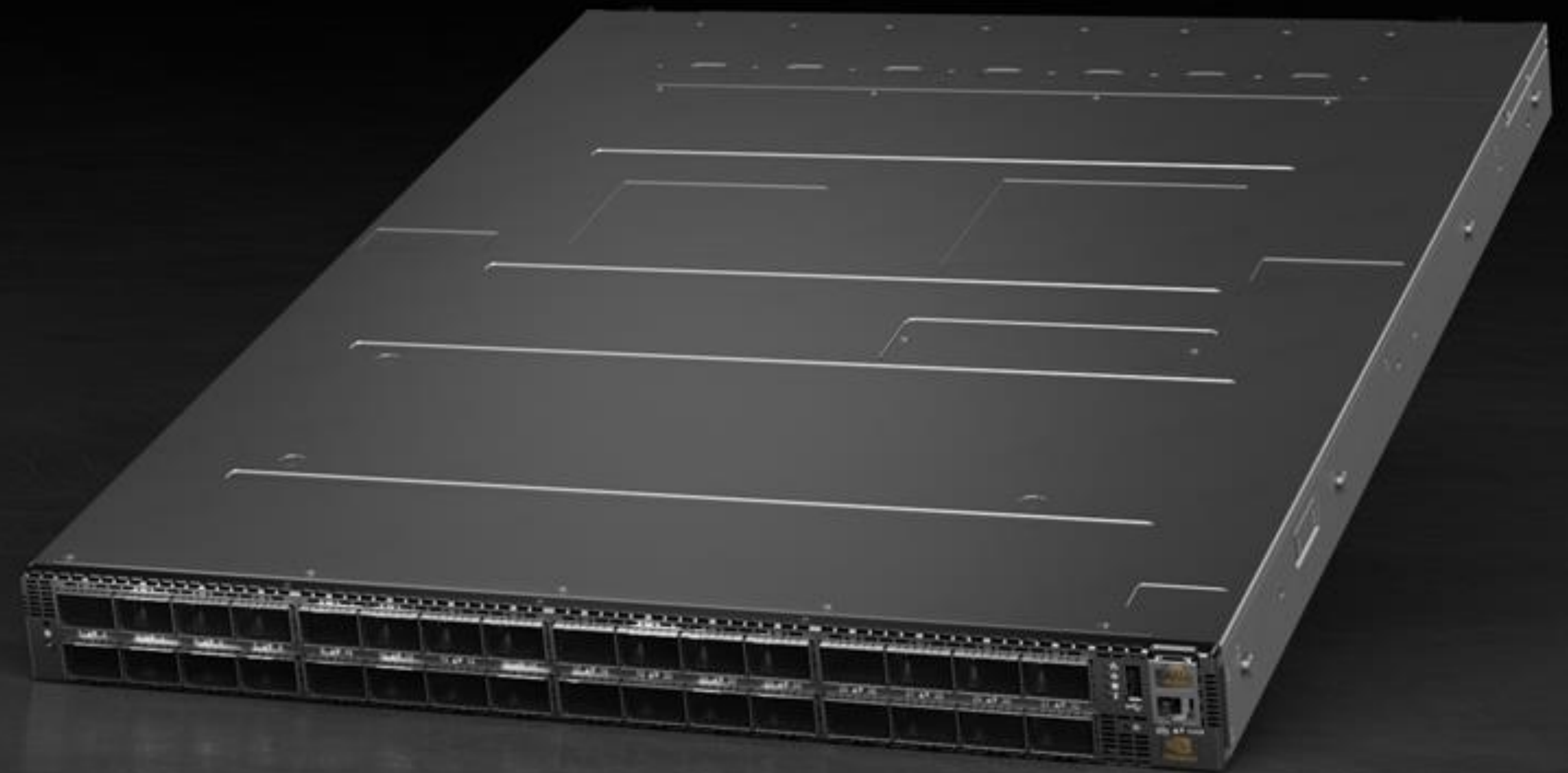
Faster Natural
Language Processing

15%

Faster Computational
Fluid Dynamics Simulations

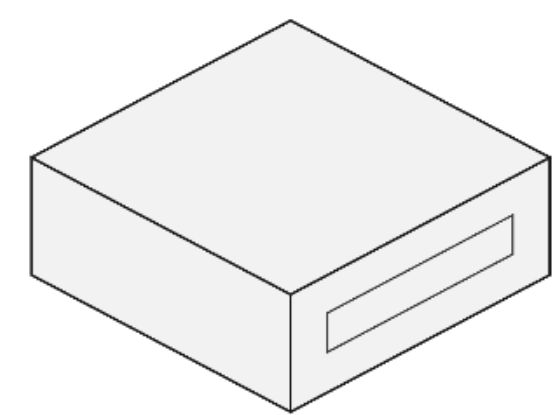
36%

Lower
Power Consumption

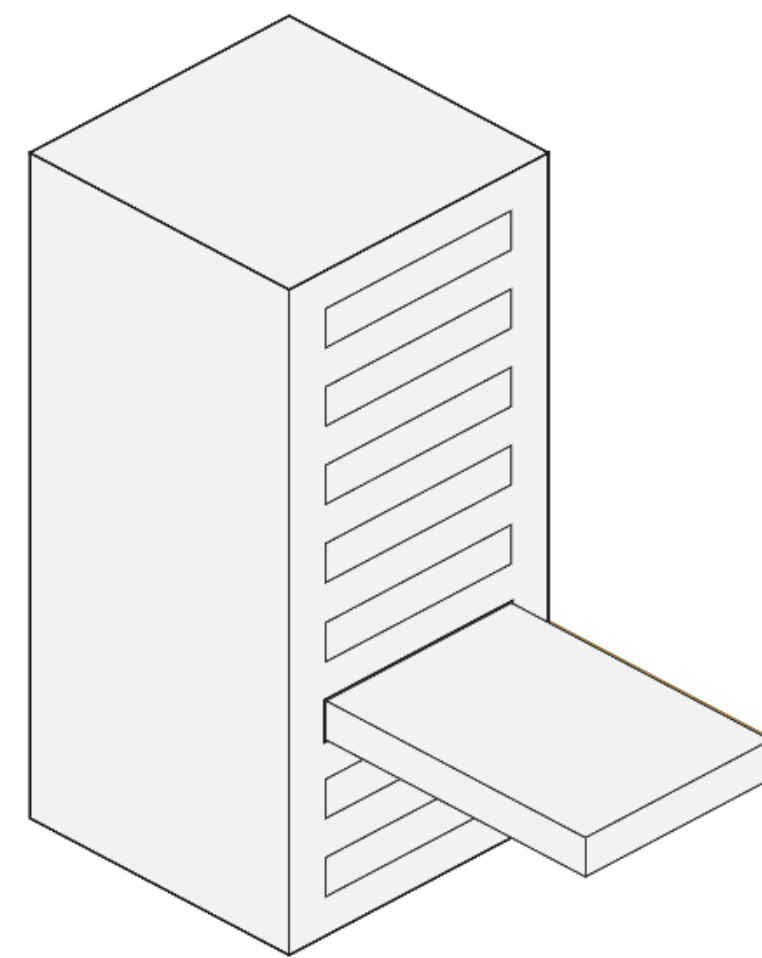


ANNOUNCING NVIDIA QUANTUM-2 INFINIBAND SWITCHES

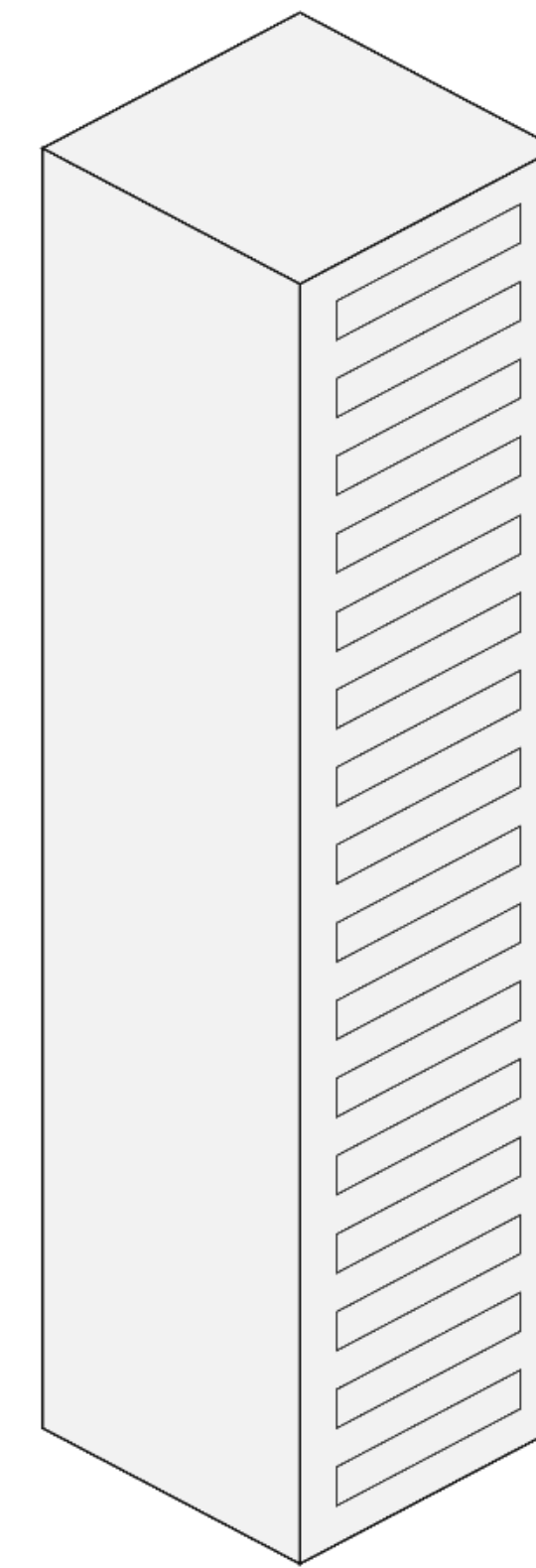
In-Network Computing Accelerates Cloud-Native Supercomputing at Any Scale



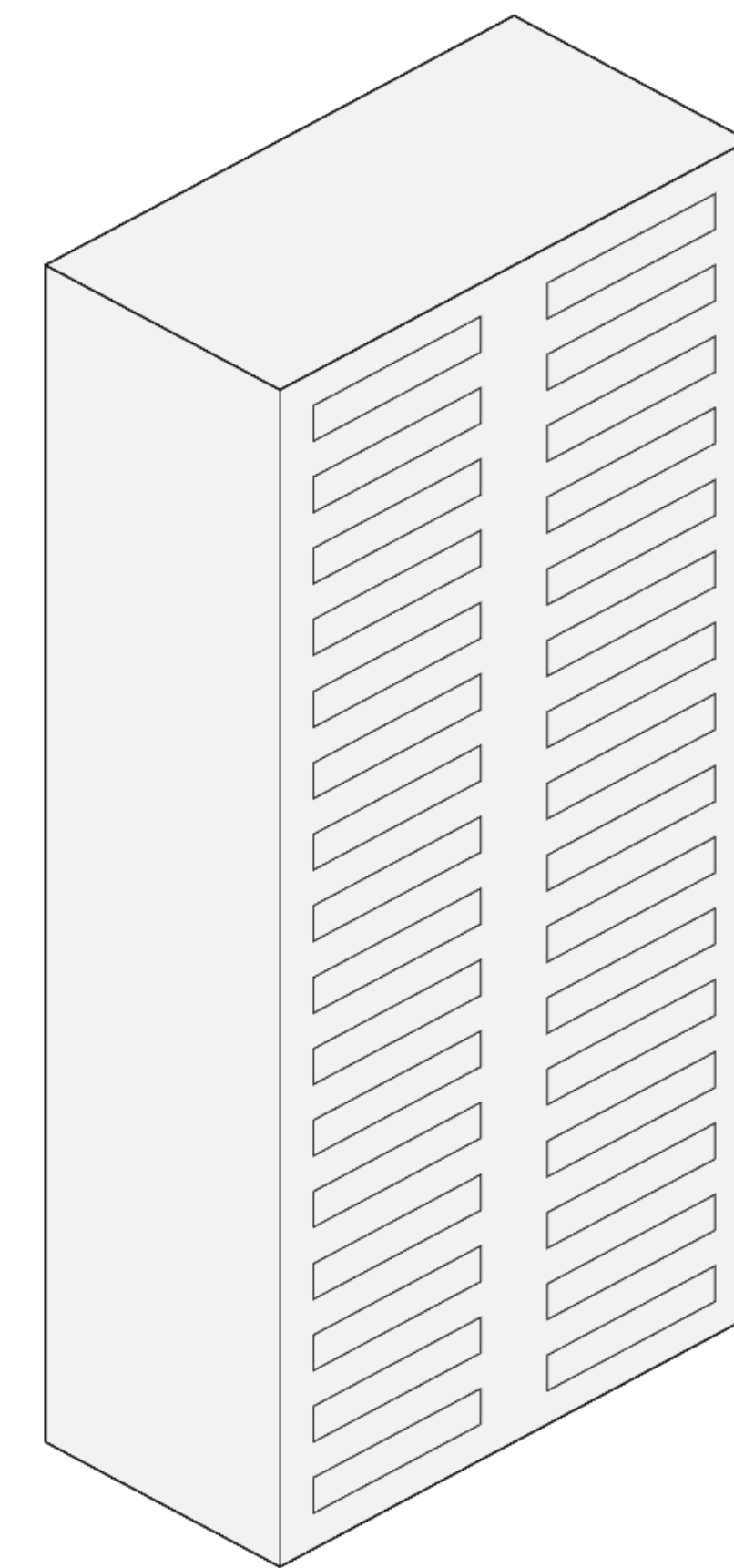
64 400Gb/s Ports
128 200Gb/s Ports



512 400Gb/s Ports
1024 200Gb/s Ports

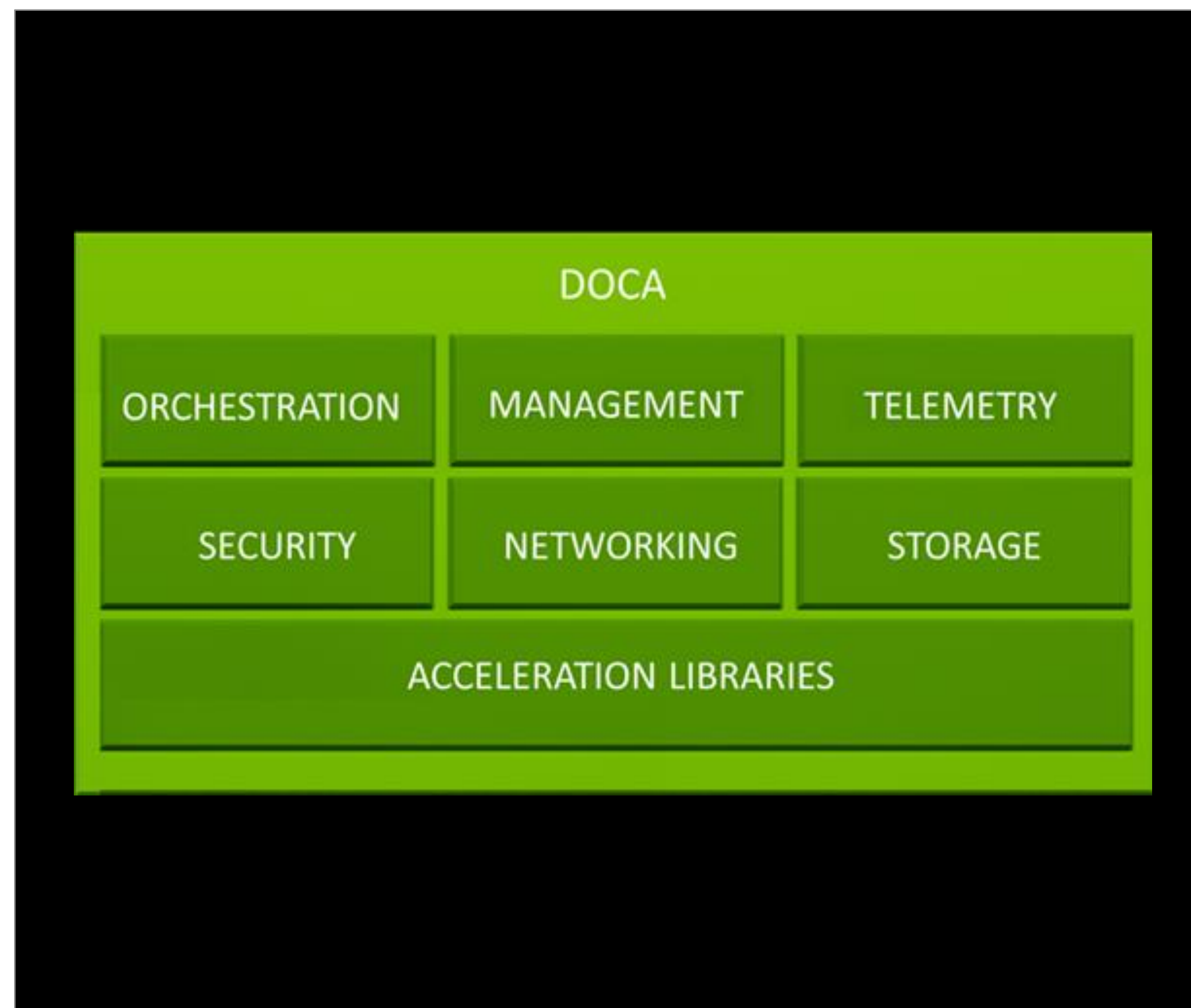


1024 400Gb/s Ports
2048 200Gb/s Ports

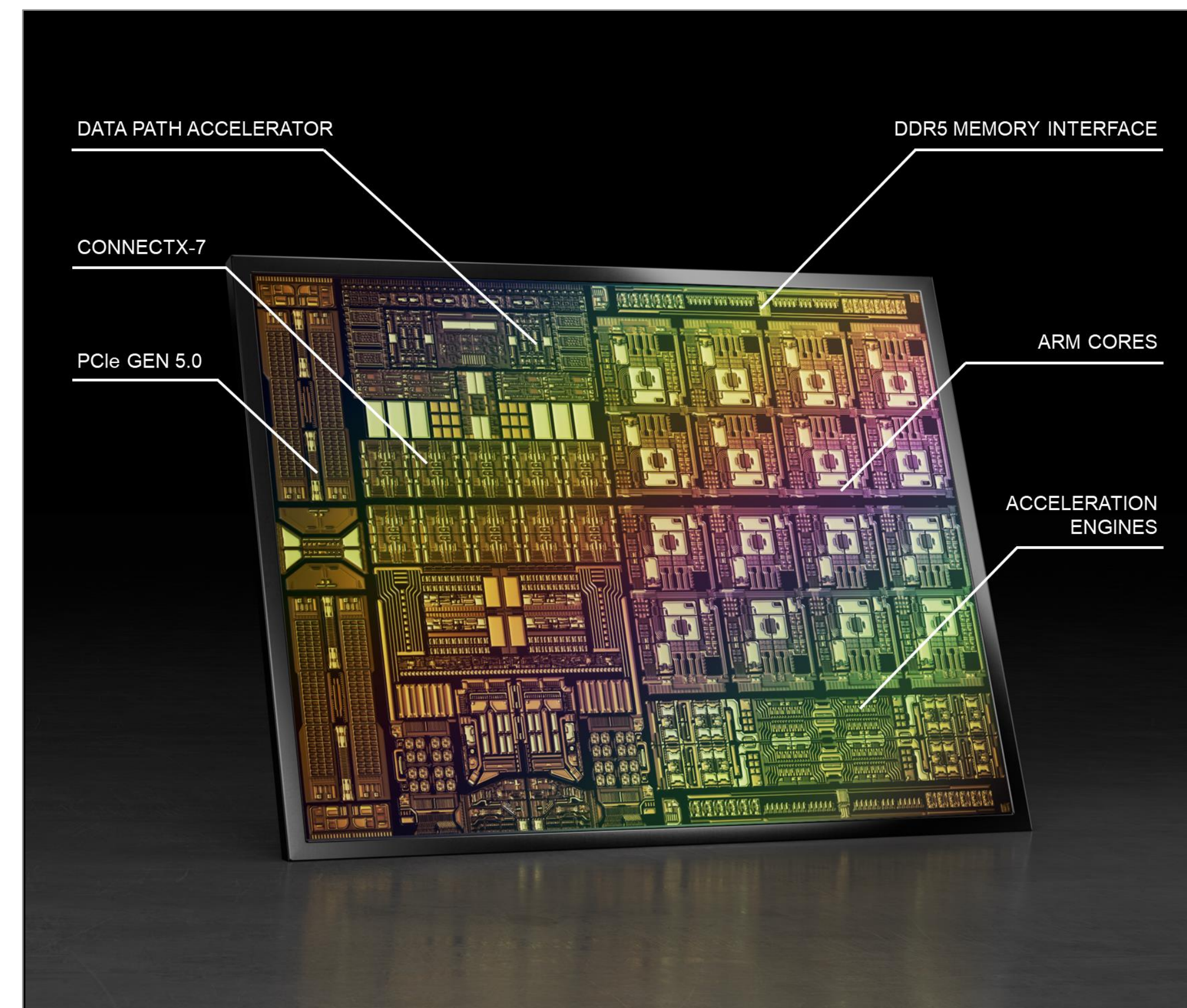


2048 400Gb/s Ports
4096 200Gb/s Ports

FROM SUPERCOMPUTERS TO SUPERCLOUDS: CLOUD-NATIVE SUPERCOMPUTERS



DOCA Enabling Growing
Partner Ecosystem



Bluefield-3 Next Generation 400G
Data Center Infra Processor



NVIDIA Quantum-2 400G InfiniBand
In-Network Computing Interconnect

QUANTUM-2 SWITCH

QM9700 and QM9790 Family of 1U Switches

64 ports of 400Gb/s (NDR) over 32 OSFP cages

128 ports of 200Gb/s (NDR200)

Secured Boot

51.2Tb/s aggregate bandwidth

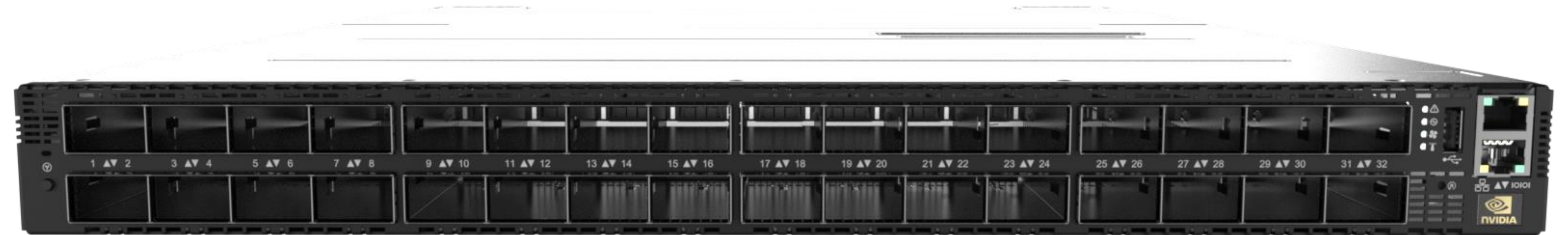
66.5 billion packets per second

SHARPV3 - low latency data reduction and streaming aggregation

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

26" depth, Air cooled

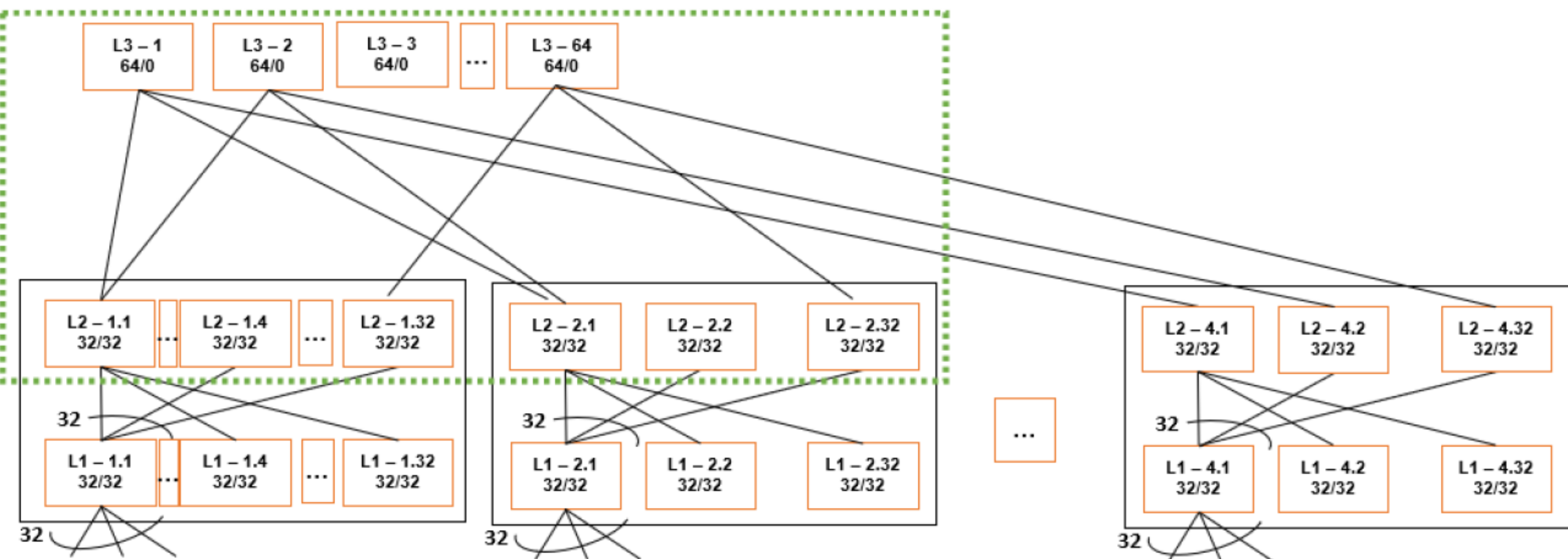
2 power supplies (1+1), hot swappable



QUANTUM-2 MODULAR SWITCHES

CS95xx Family of Modular Switches

- Single design for multiple configurations
- Leaf and Spine 2U modules (128 NDR ports over 64 OSFP cages, each)
- Network management system (UFM, Centralized NMS) to manage the entire modular system
- Liquid cooled solution by external AHX or CDU
- Internal interconnect (backplane)
 - Active copper OSFP based
 - Comes pre-routed as part of the chassis



	CS9500	CS9510	CS9520
Ports	2,048 NDR 4,096 NDR200	1,024 NDR 2,048 NDR200	512 NDR 1,024 NDR200
Modules	32 Leaf 16 Spine	16 Leaf 8 Spine	8 Leaf 4 Spine
Dimensions	Width: 3 racks size (2,100mm) Height: 48U (2,300mm) Depth: 1,200mm	Width: 1.5 racks size (1,200mm) Height: 48U (2,300mm) Depth: 1,200mm	Width: 1.5 racks size (1,200mm) Height: 48U (2,300mm) Depth: 1200mm
Max power	85KW	42.5KW	22KW
Weight	~3 tons	~1.5 tons	~1 ton



Picture shown is for illustration purpose only

400GB/S INFINIBAND CABLING OVERVIEW

Switch
64 ports of 400Gb/s (4x100Gb/s PAM4)
32 OSFP connectors – 2 ports per connector

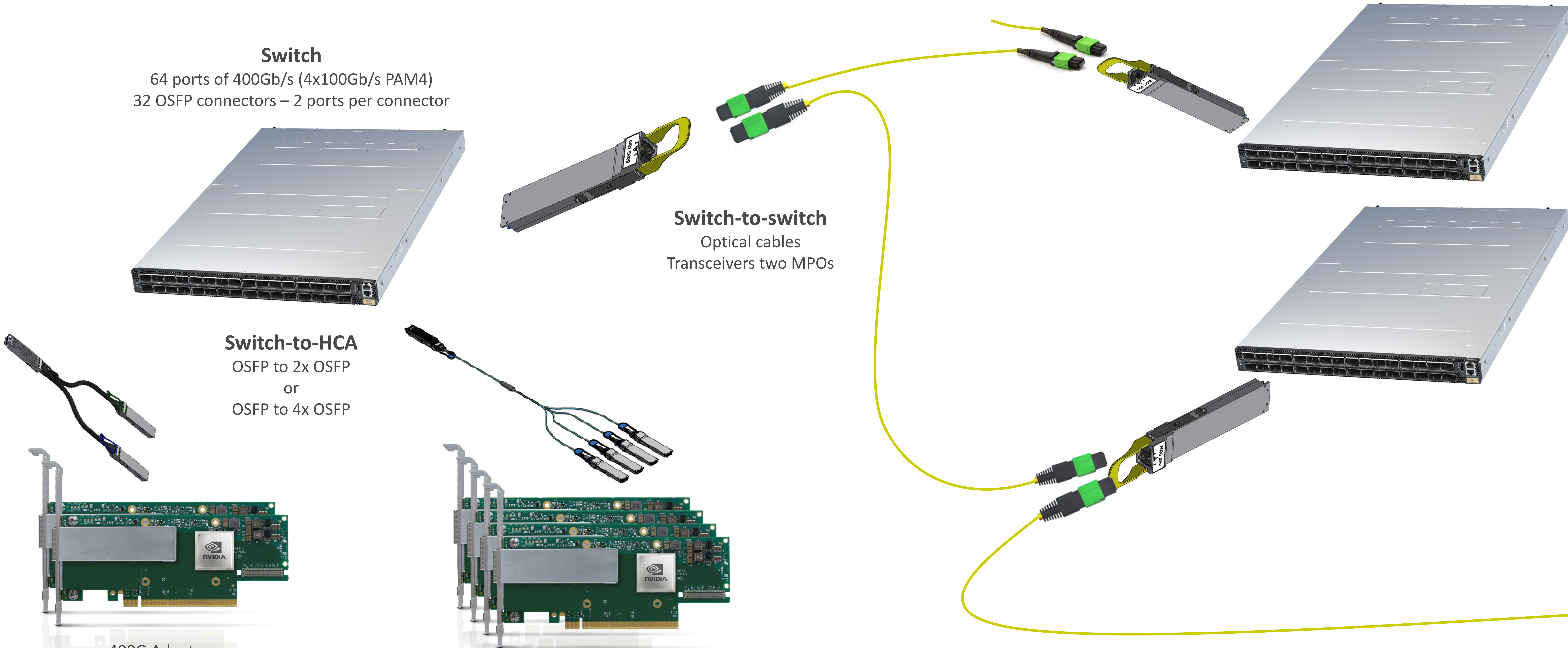
Switch-to-switch
Optical cables
Transceivers two MPOs

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

400G Adapter

HCA
400Gb/s and 200Gb/s OSFP connectors

200G Adapter



400GB/S(4X100G PAM4) CONNECTIVITY

Transceivers

Twin-port transceiver, Single-port NDR transceiver, Single-port 200Gb/s transceiver

Single Mode (Yellow pull-tab) and Multi Mode (Beige pull-tab)

MPO (for 400Gb/s) and split-MPO (for 200Gb/s) offering

Single Mode (Yellow cable jacket) and Multi Mode (Aqua cable jacket)

Transceivers' Power

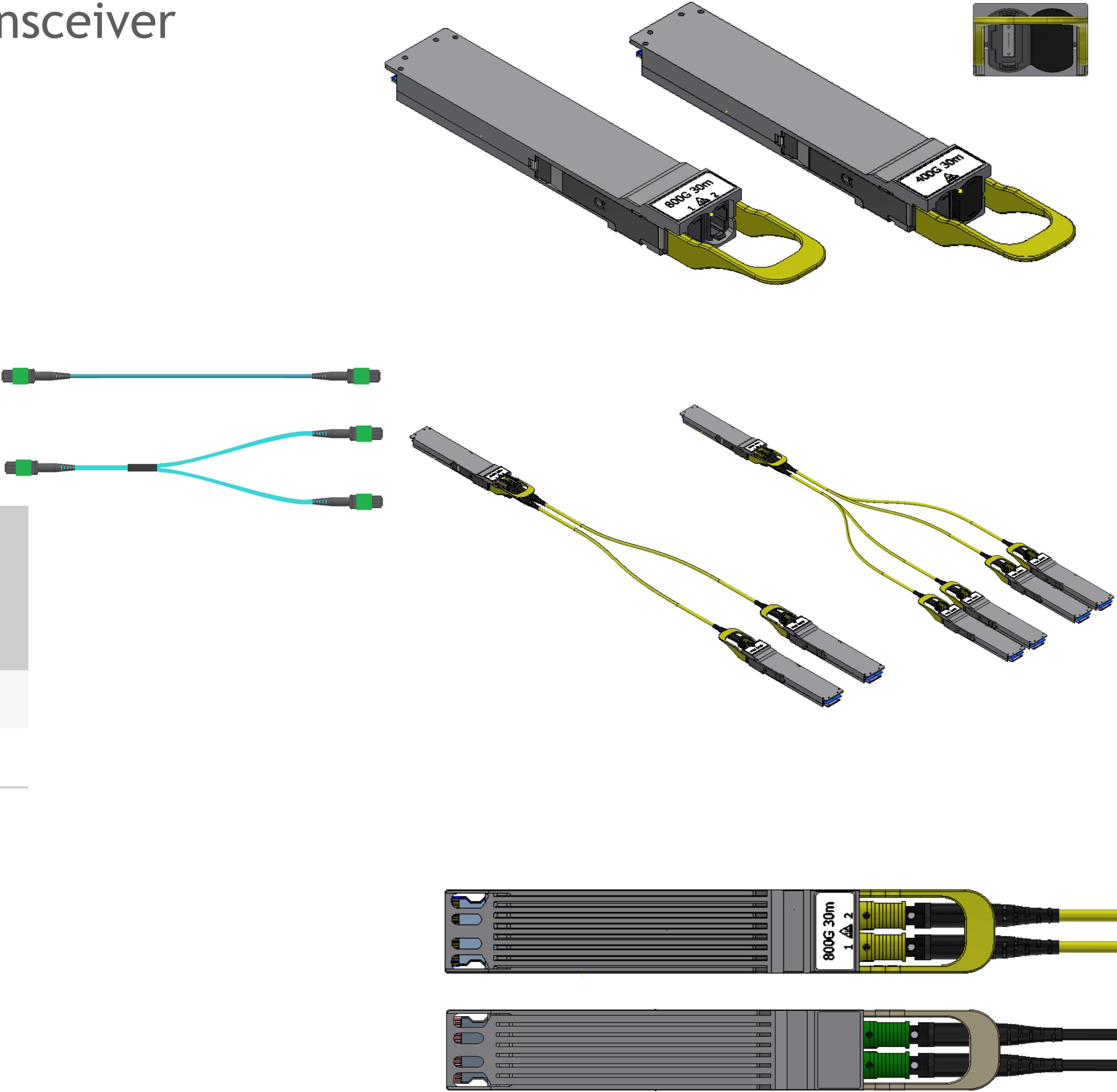
	Twin-port 400Gb/s	Single-port 400Gb/s for HCA	Single-port 200Gb/s for HCA
Single mode	17W	9W	5W
Multi Mode	15W	8W	5W

Types of transceivers

Up to 30m, up to 150m

Finned OSFP - for air-cooled systems

Flat OSFP - for liquid-cooled port, or HCAs with riding heatsink

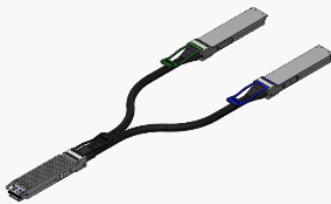
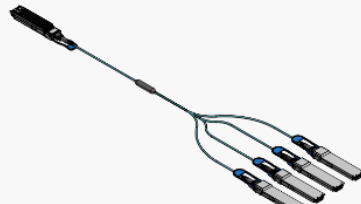



Pending U.S. patent application No. 16/750,632

400GB/S (4X100G PAM4) CONNECTIVITY

Cables

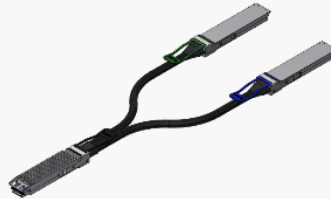
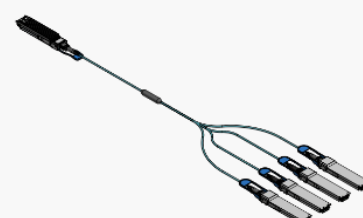
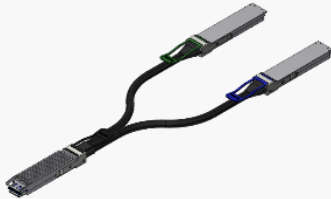


	NDR Technology	
Switch	Air / Liquid cooled	Air / Liquid cooled
	Finned / Flat OSFP	Finned / Flat OSFP
	2x400 (2x NDR)	2x400 (4x NDR200)
	OSFP to 2xOSFP	OSFP to 4xOSFP
Form factor		
DAC - Copper up to 3m	✓	✓
ACC - Active copper up to 5m	✓	✓
HCA	400Gb/s	200Gb/s
	OSFP	OSFP

	NDR Technology
Switch	Air / Liquid cooled
	Finned / Flat OSFP
	2x400 (2x NDR)
	OSFP to OSFP
Form factor	
DAC - Copper up to 2m	✓
ACC - Active copper up to 5m	✓
Switch	2x 400Gb/s
	OSFP

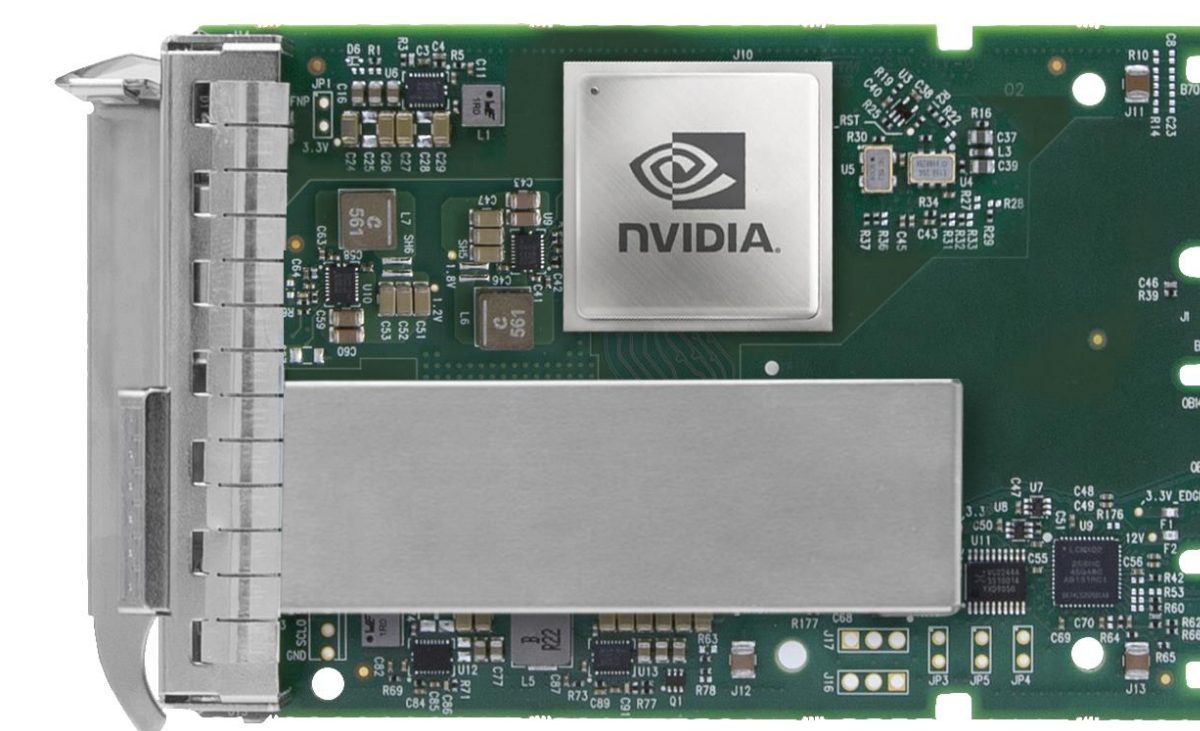
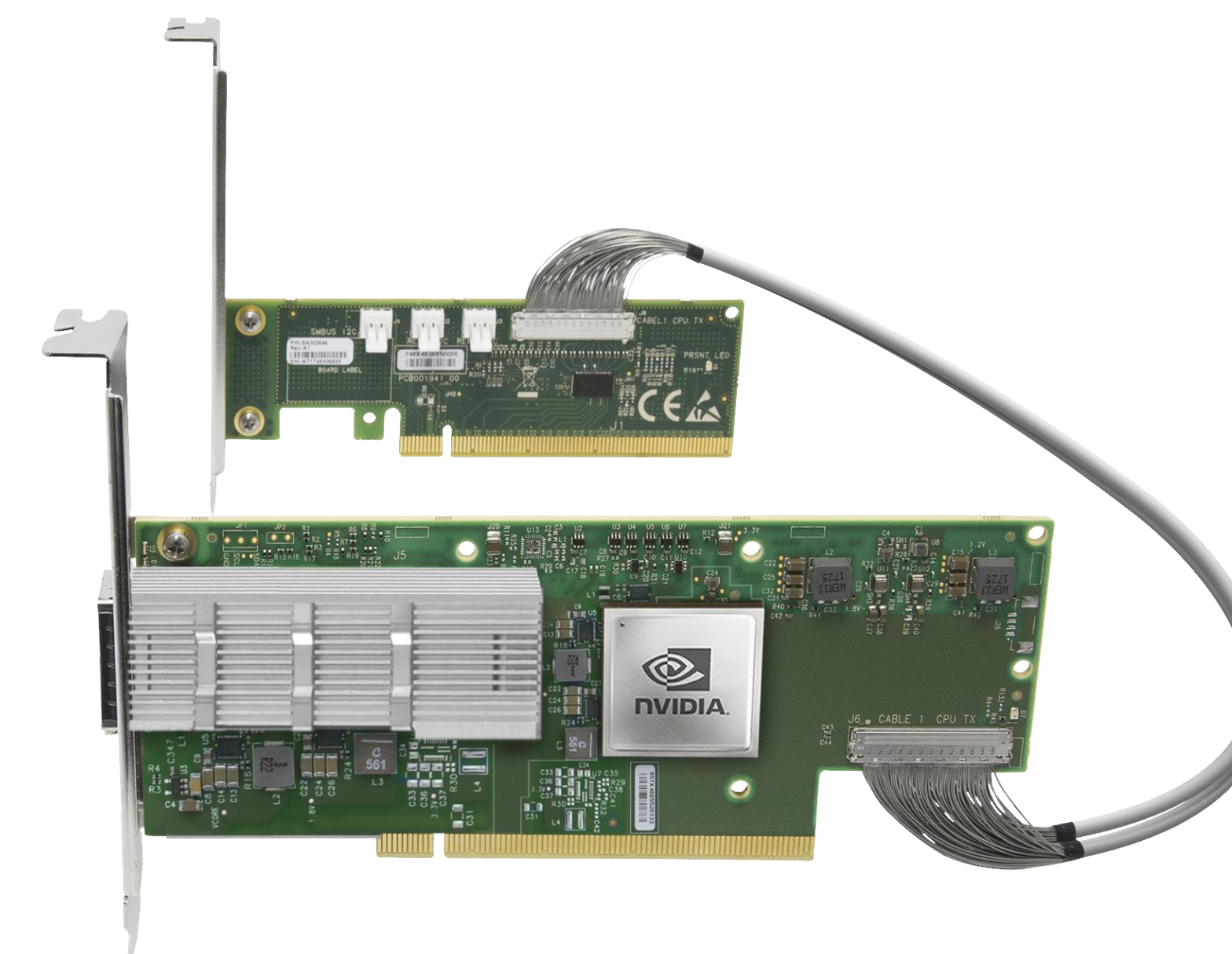
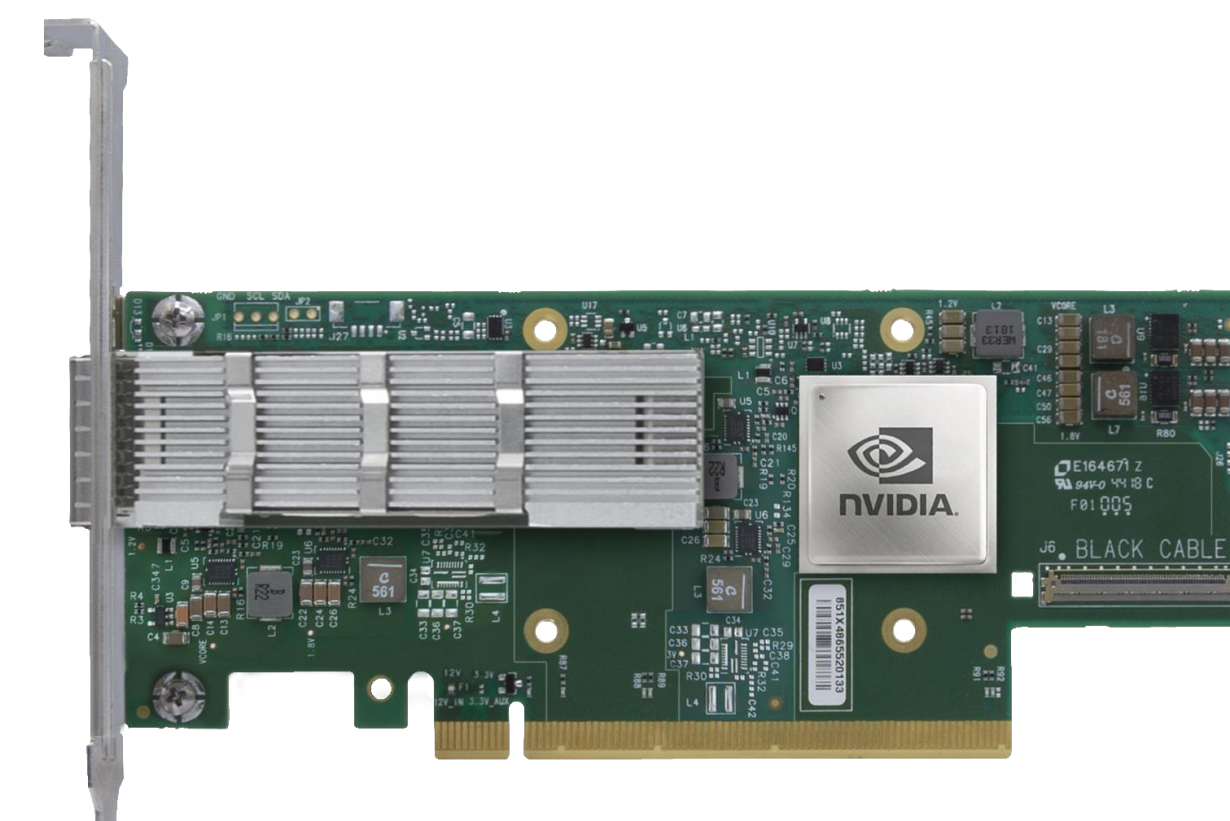
400GB/S (4X100G PAM4) CONNECTIVITY

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 2m	✓	✓	✓
ACC - Active copper	x	x	x
AOC - Optical cable up to 30m	✓	✓	✓
HCA or Switch	200Gb/s	100Gb/s	100Gb/s
	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



INFINIBAND GENERATIONS COMPARISON

Overview

	SwitchIB-2 & ConnectX-5	Quantum & ConnectX-6	Quantum-2 & ConnectX-7
Port speed	100G	200G	400G
Switch radix	36 ports 100Gb/s	40 ports 200Gb/s 80 ports 100Gb/s (100G)	64 ports 400Gb/s 128 ports 200Gb/s (200G)
SHARP		SHARPV2	SHARPV3
	SHARPV1	Small Message Reductions (LLT - Low Latency Transmission)	Small Message Reductions (LLT - Low Latency Transmission)
	Small Message Reductions (LLT - Low Latency Transmission)	Long Vector Reduction (SAT - Streaming Aggregation)	Long Vector Reduction (SAT - Streaming Aggregation)
		Support for 2 SAT flows per switch at the same time	Support for 64 SAT flows per switch at the same time

UFM PLATFORMS PORTFOLIO



UFM Telemetry
Real-Time Monitoring



UFM Enterprise
Management, Monitoring & Orchestration
(UFM Enterprise includes UFM Telemetry)



UFM Cyber-AI
Cyber Intelligence and Analytics
(UFM Cyber-AI includes UFM Enterprise)

UFM TELEMETRY PLATFORM

Real-Time Monitoring

- Network validation (adapters, switches, cables, transceivers) and connectivity checks
- System component validations
- Network performance tests
- Application tests
- Streaming of telemetry information into on-premises or cloud-based database
- Platform options: Docker container, software, or UFM Telemetry appliance



UFM ENTERPRISE PLATFORM

Management, Monitoring and Orchestration

- Includes all UFM Telemetry services
- Network setup, connectivity validation and secure cable management
- Automated network discovery and network provisioning
- Network telemetry and traffic monitoring, congestion discovery
- Performance, health and fault monitoring
- Centralized management for global software updates and advanced network configuration
- Job scheduler provisioning, integrated with Slurm and Platform LSF
- Network provisioning, integrated with OpenStack, Azure Cloud and VMware
- Advanced reporting, comprehensive REST APIs and rich web-based GUI
- Platform options: Docker container, software, or UFM Enterprise appliance



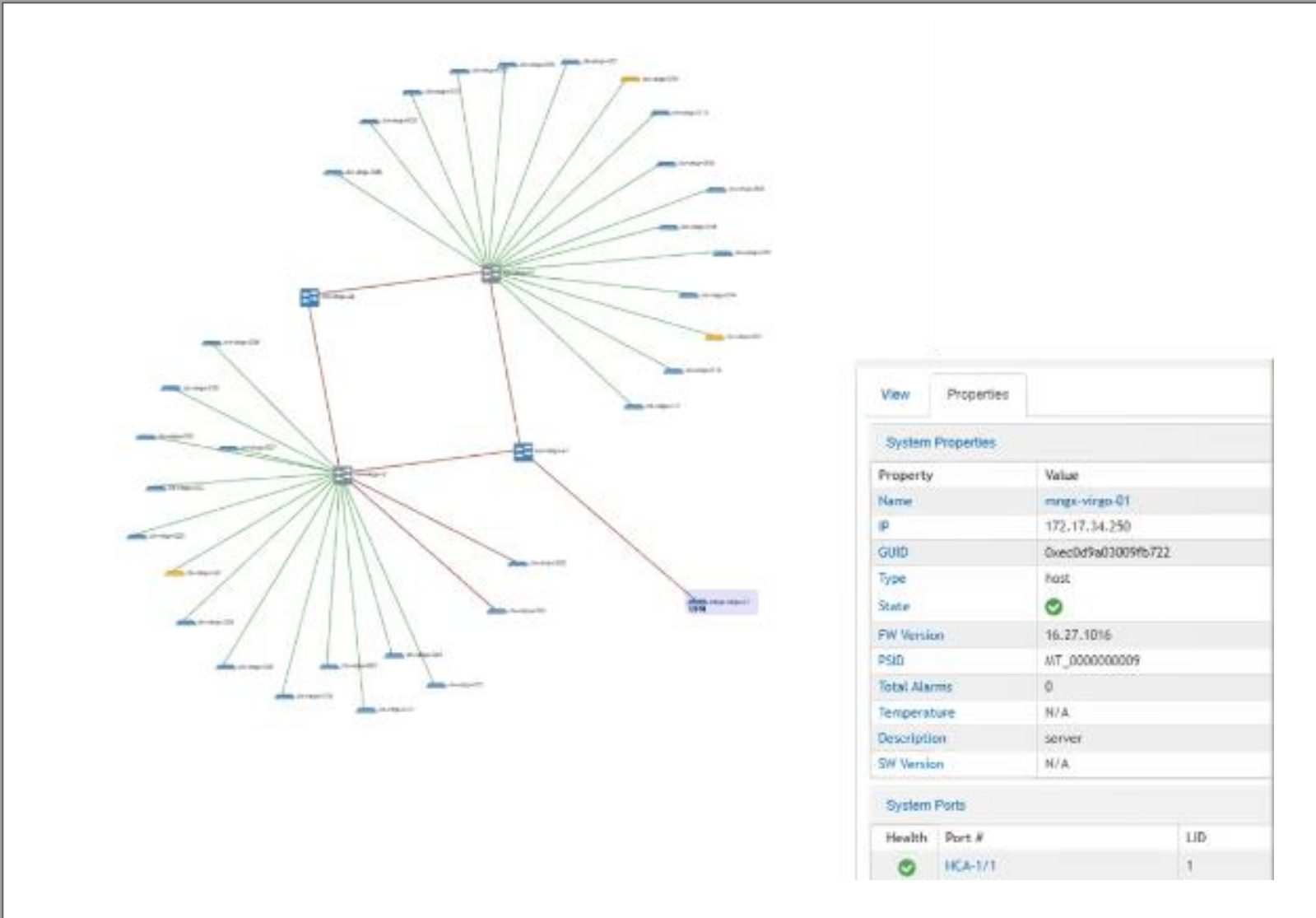
UFM CYBER-AI PLATFORM

Cyber Intelligence and Analytics

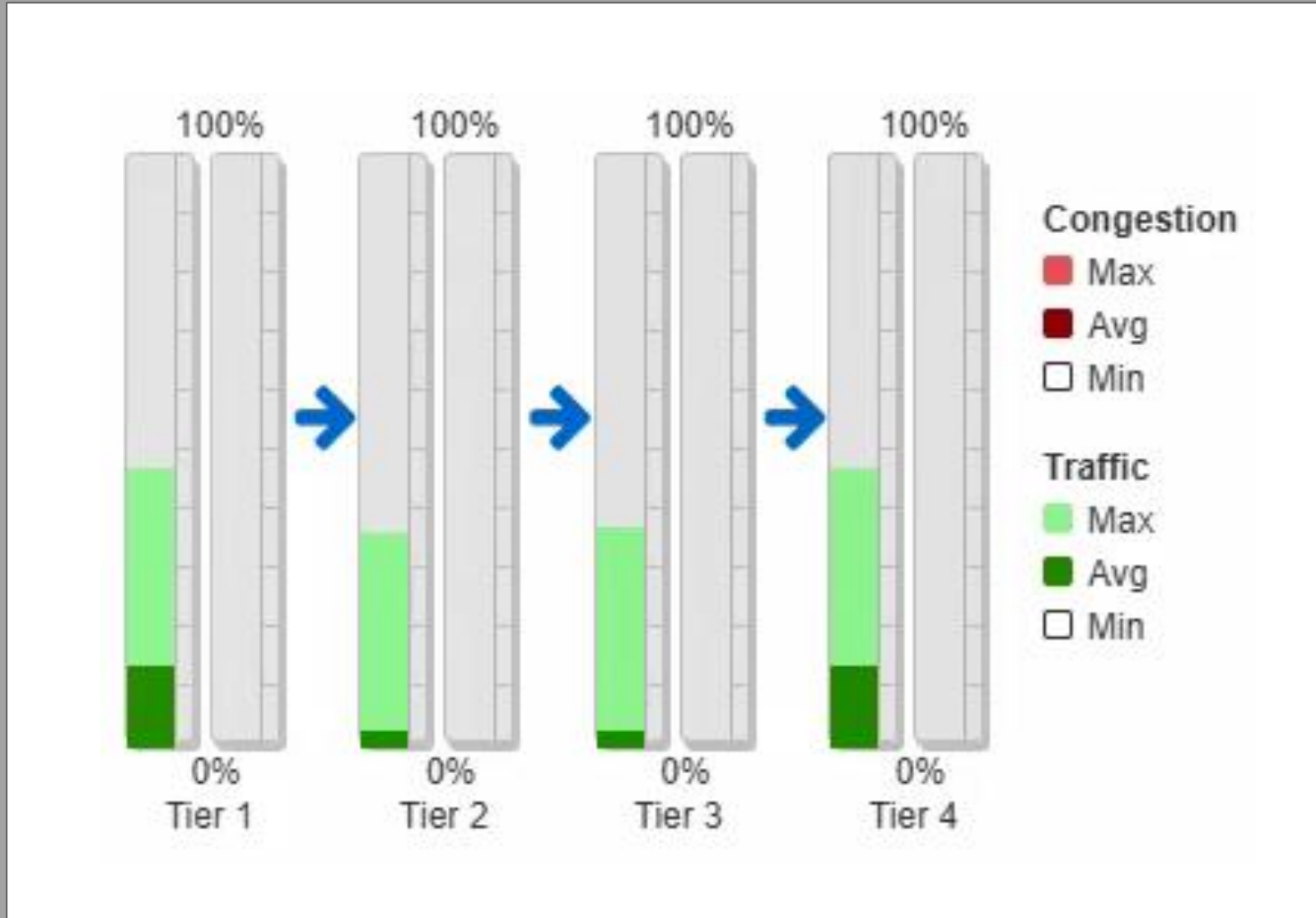
- Includes all UFM Enterprise services
- Learns system heartbeat, operation mode, condition, usage, workload network signatures
- Builds enhanced databased of telemetry information and discovers correlations
- Detects performance degradations, usage and profile changes over time
- Provides alerts of abnormal system and application behavior, and potential system failures
- System administrators can quickly detect and respond to potential security threats
- System administrators can efficiently plan and address future failures
- Predictability is optimized over time as system data is collected
- Performs corrective actions
- Platform: UFM Cyber-AI appliance



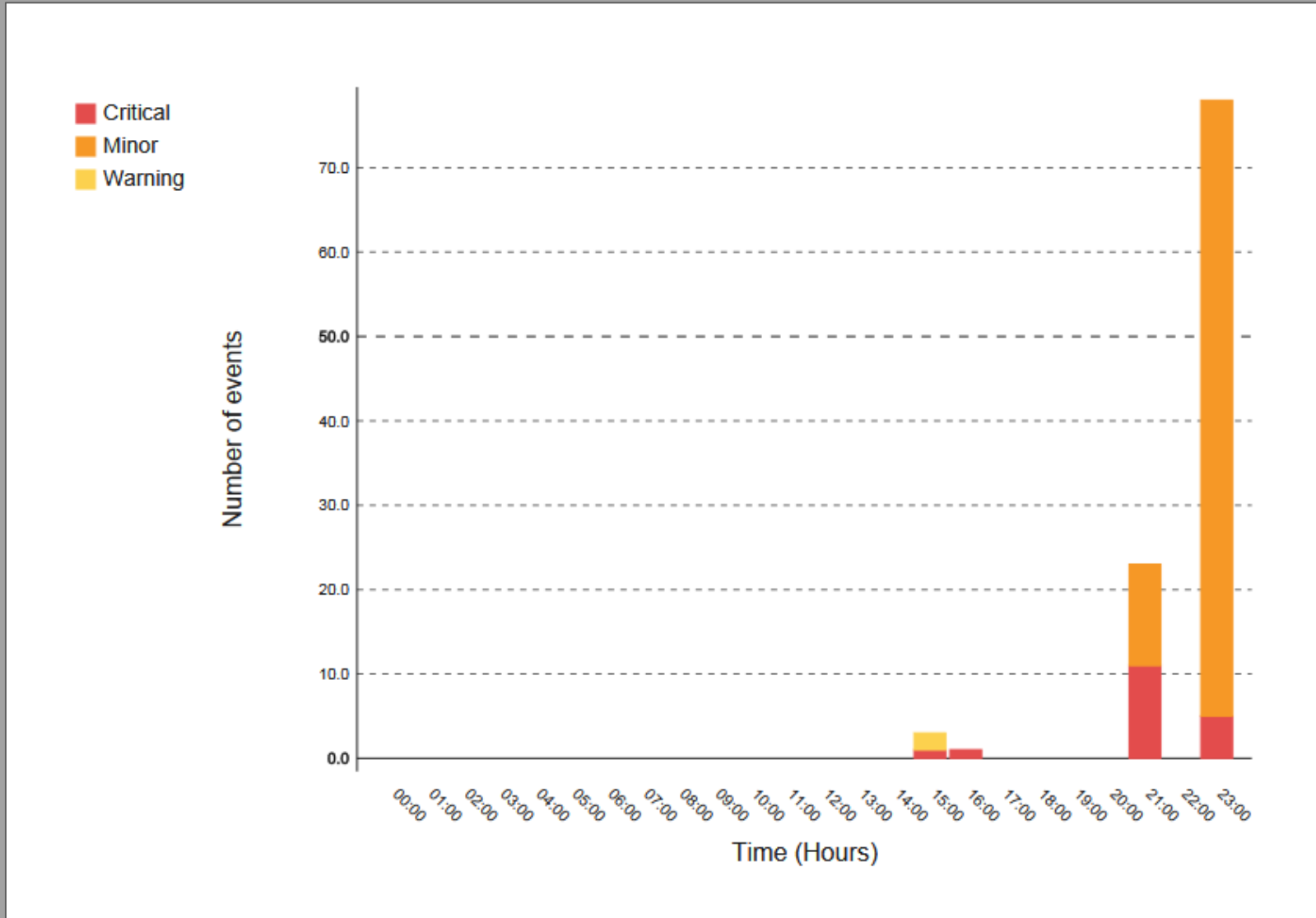
UFM (UNIFIED FABRIC MANAGER) DASHBOARD



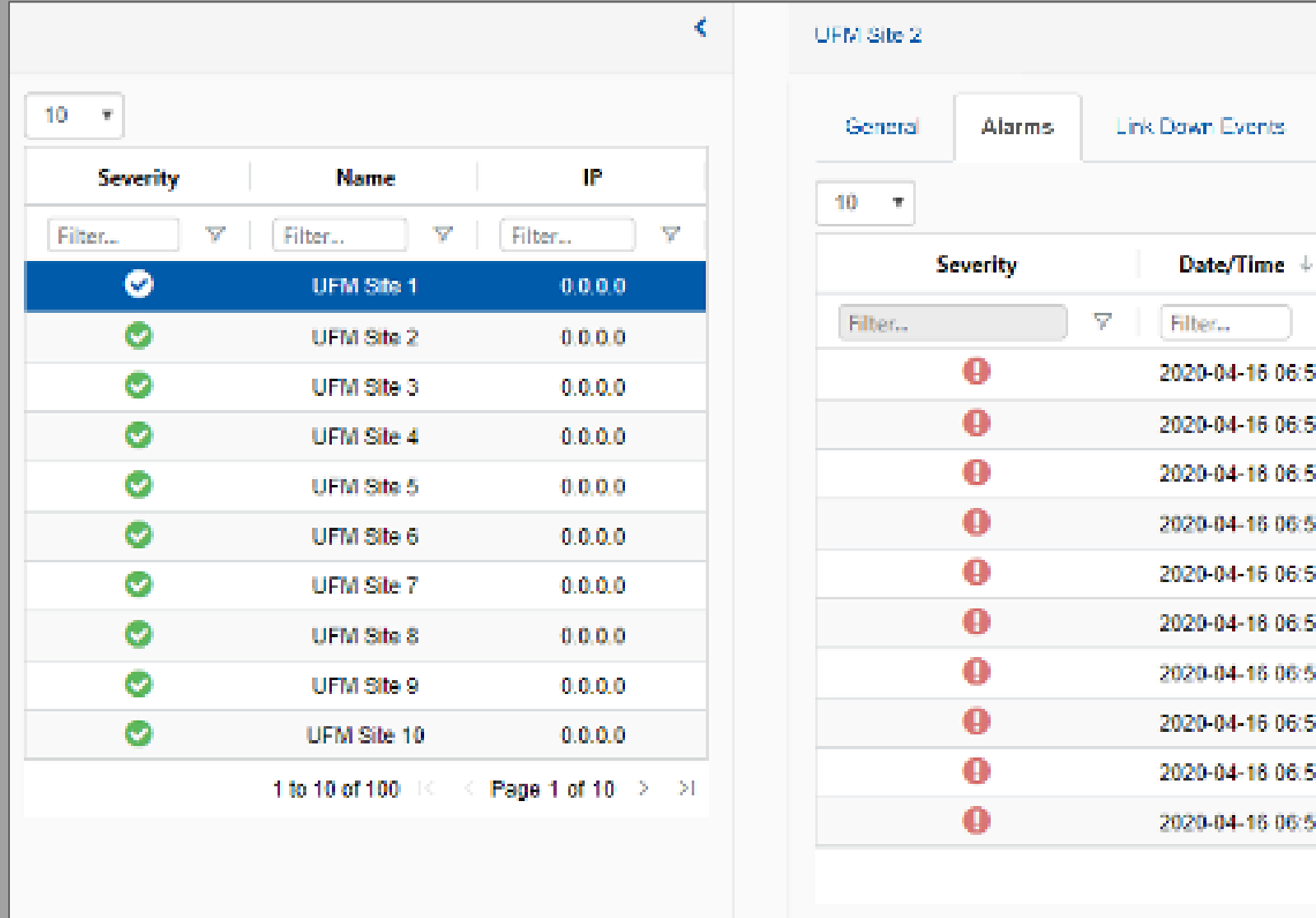
Network Validation



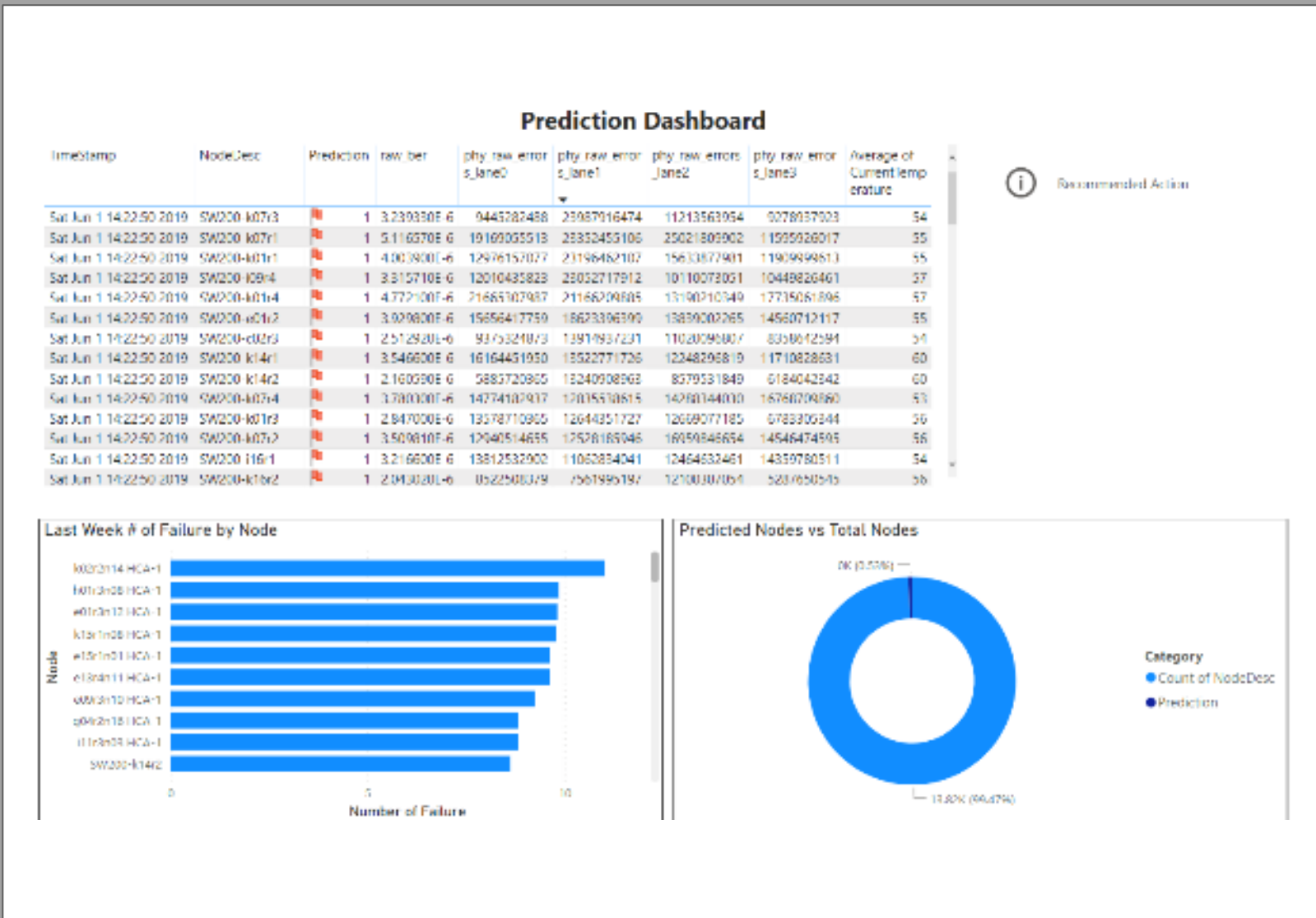
Congestion Mapping



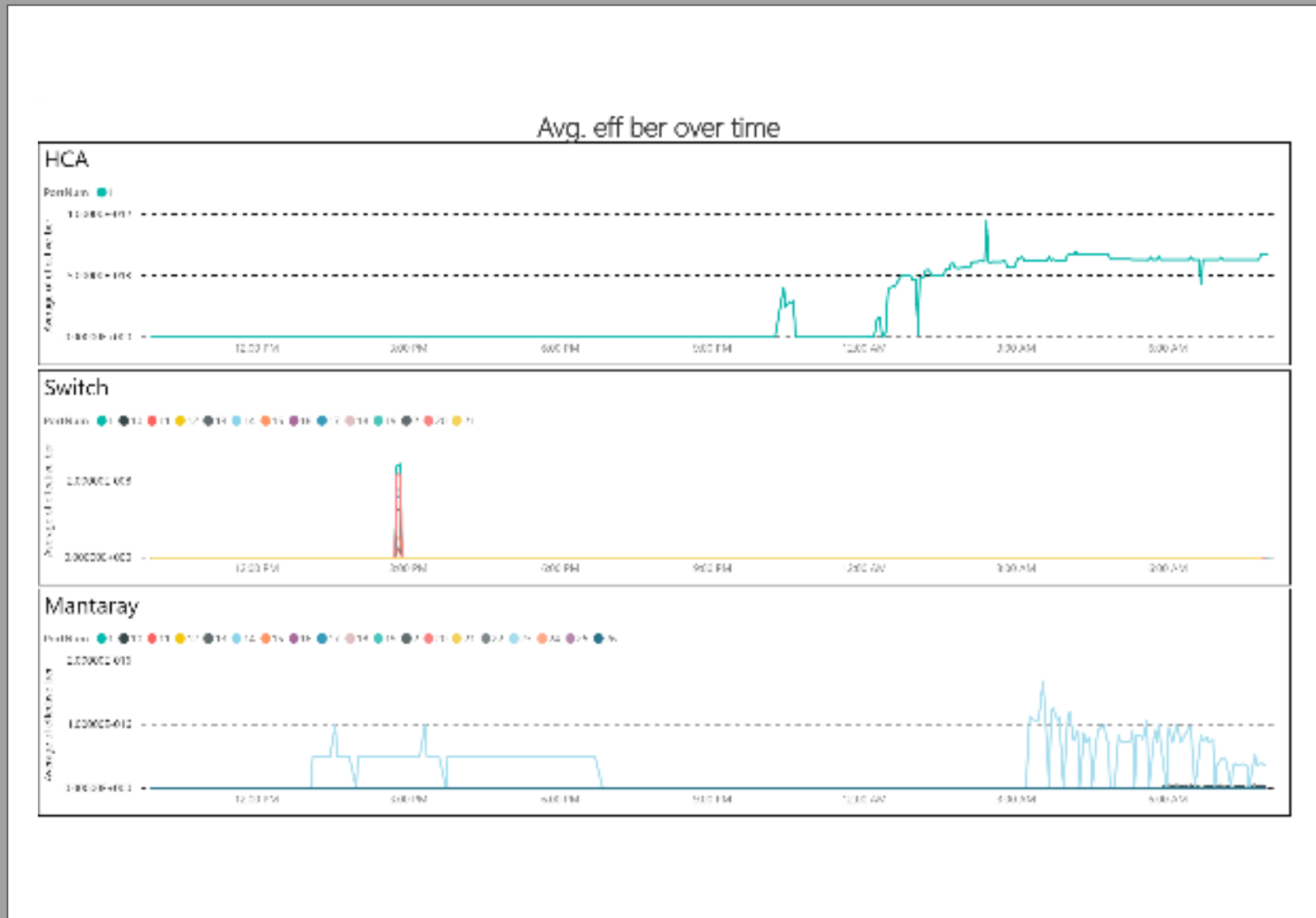
Health Reports



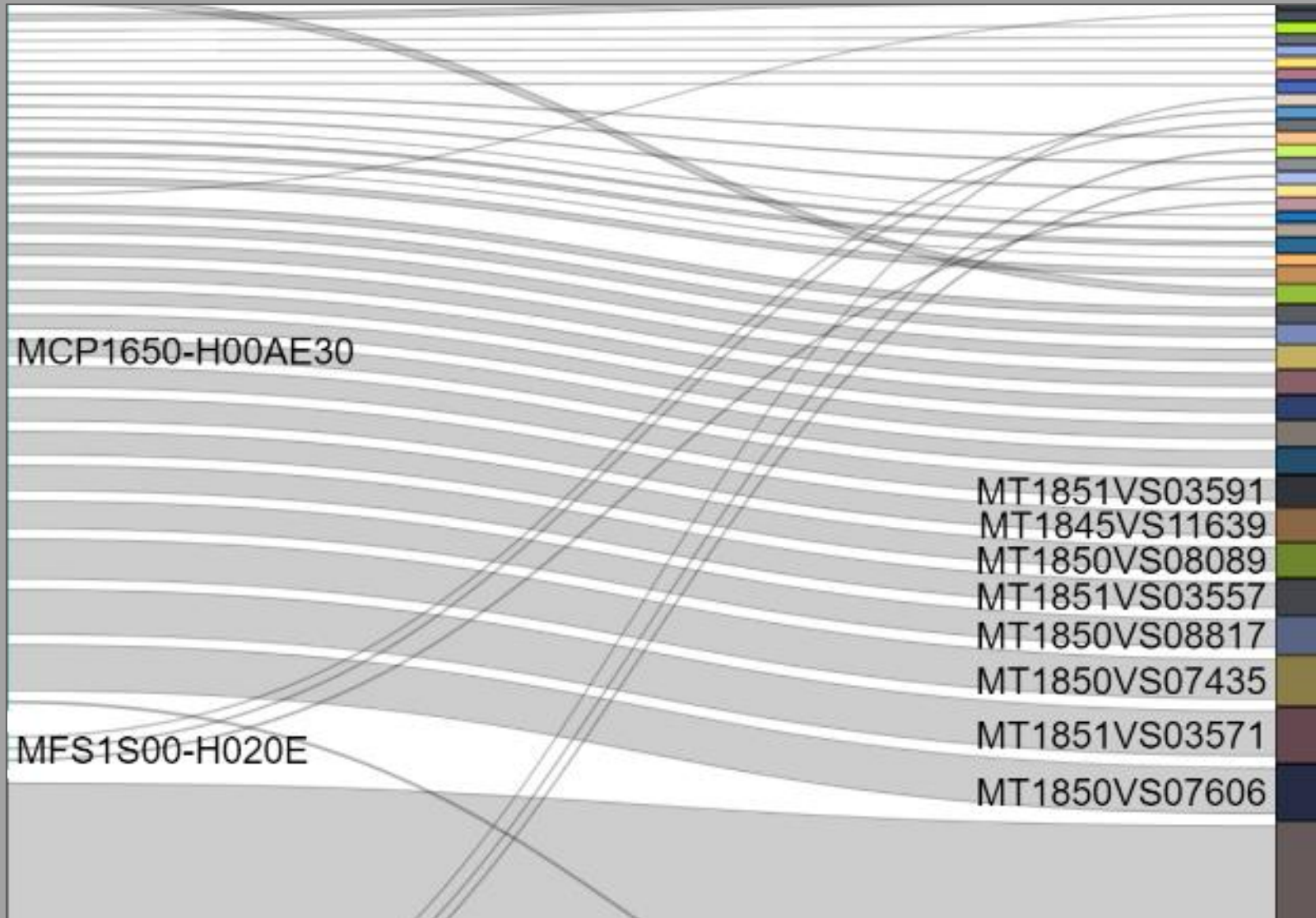
Inventory Mapping



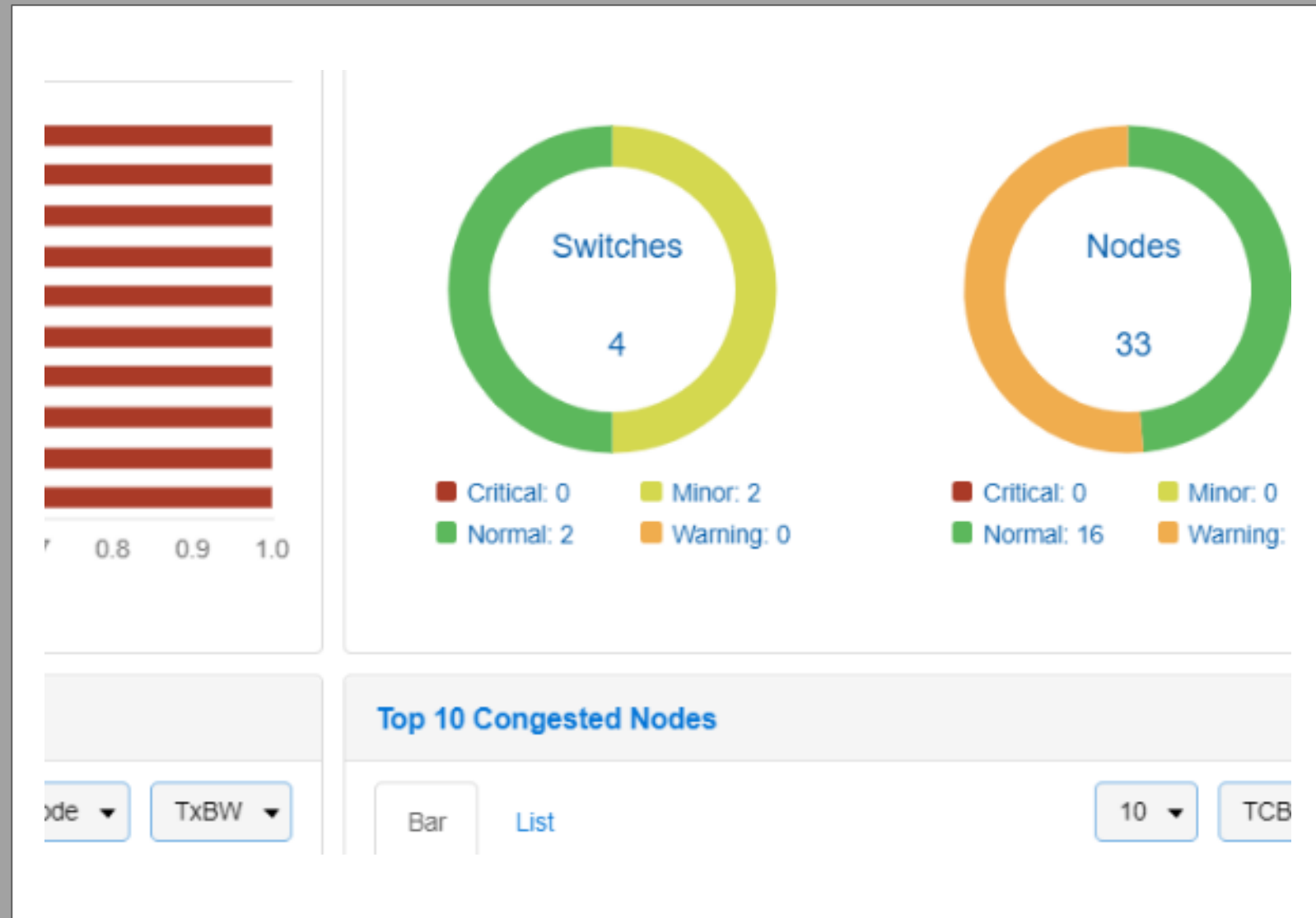
Prediction Dashboard



Real-Time Analysis




Performance Monitoring



Secure Cable Management

PREDICTION / ANOMALY


UFM Cyber-AI

Dashboard

General

Prediction Dashboard

Network Map

Managed Elements

Logical Elements

Events & Alarms

Telemetry

System Health

Jobs

Settings

Prediction Dashboard

Site Name: Local ?
Date: Last 6 hours

Suspicious Behavior

422 Network Alerts
11 Tenant/Application Alerts

Link Analysis

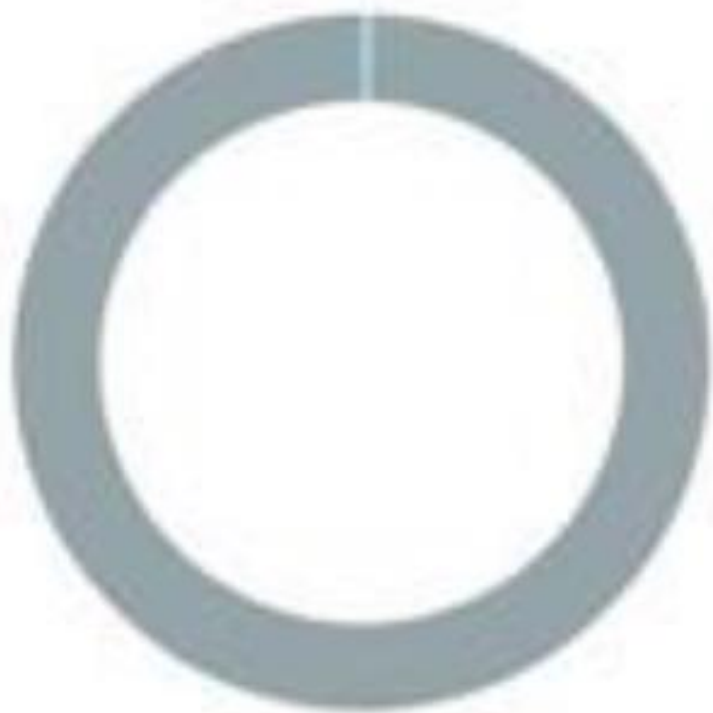
5 Link Failure Prediction
0 Link Anomaly
0 Cable Events

Top 10 nodes by Link Failure Indication

Number of Indications



k15r1n03 HCA-1	9
k11r2n03 HCA-1	4
b16r4n08 HCA-1	4
k11r2n03 HCA-1	2
k11r2n03 HCA-1	1

Anomaly Node vs Normal Node

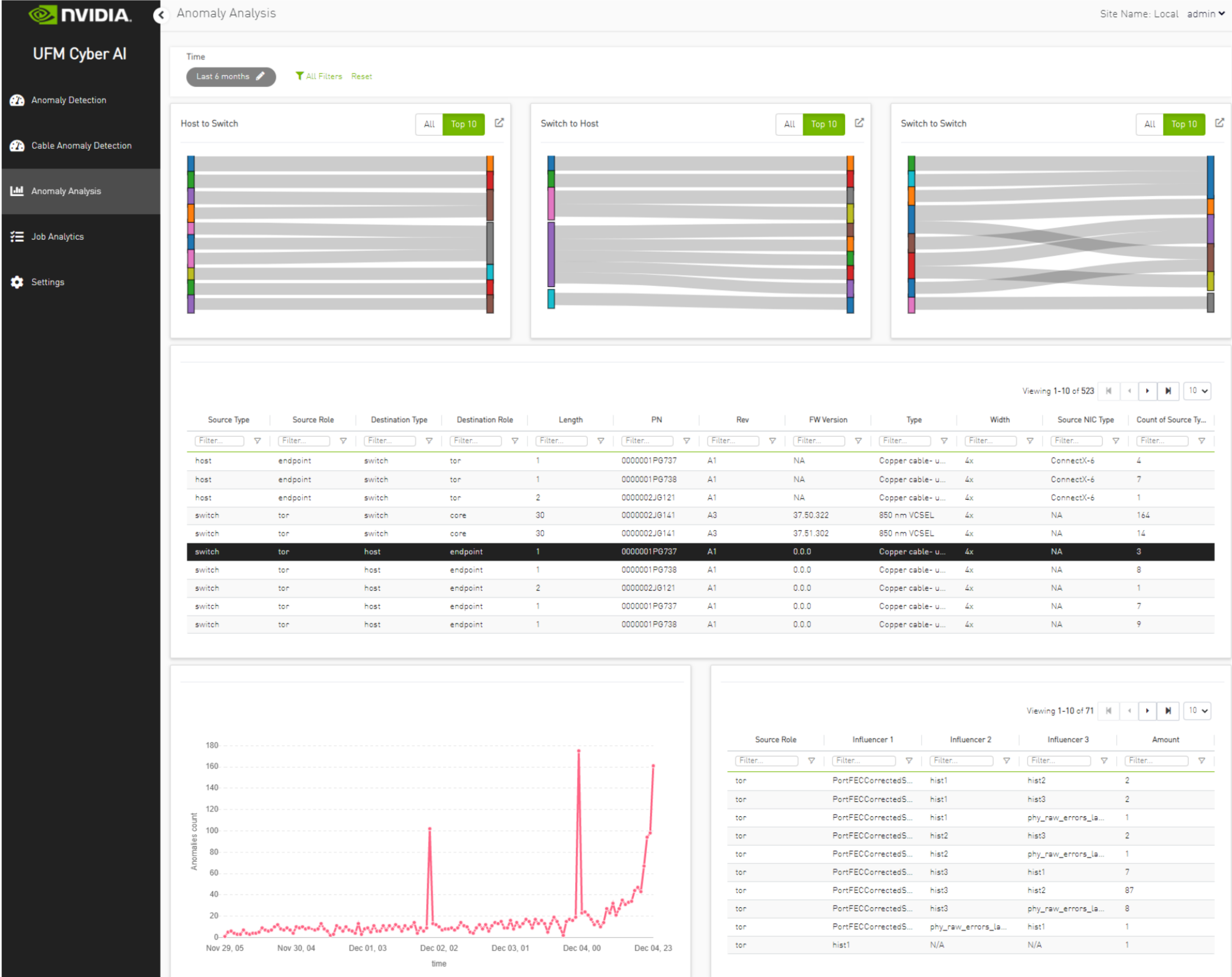


Anomaly Dashboard

10

Timestamp	Severity	Node ID	Port Number	Grade	Description	Recommended Action
Mon May 24 07:54:17 2021	Critical	k14r2n03 HCA-1	1	1	Link failure prediction detected on port k14r2n03 HCA-1	
Mon May 24 06:41:03 2021	Critical	k13r1n03 HCA-1	1	0	Link failure prediction detected on port k13r1n03 HCA-1	

NEW FEATURE - ANOMALY ANALYSIS

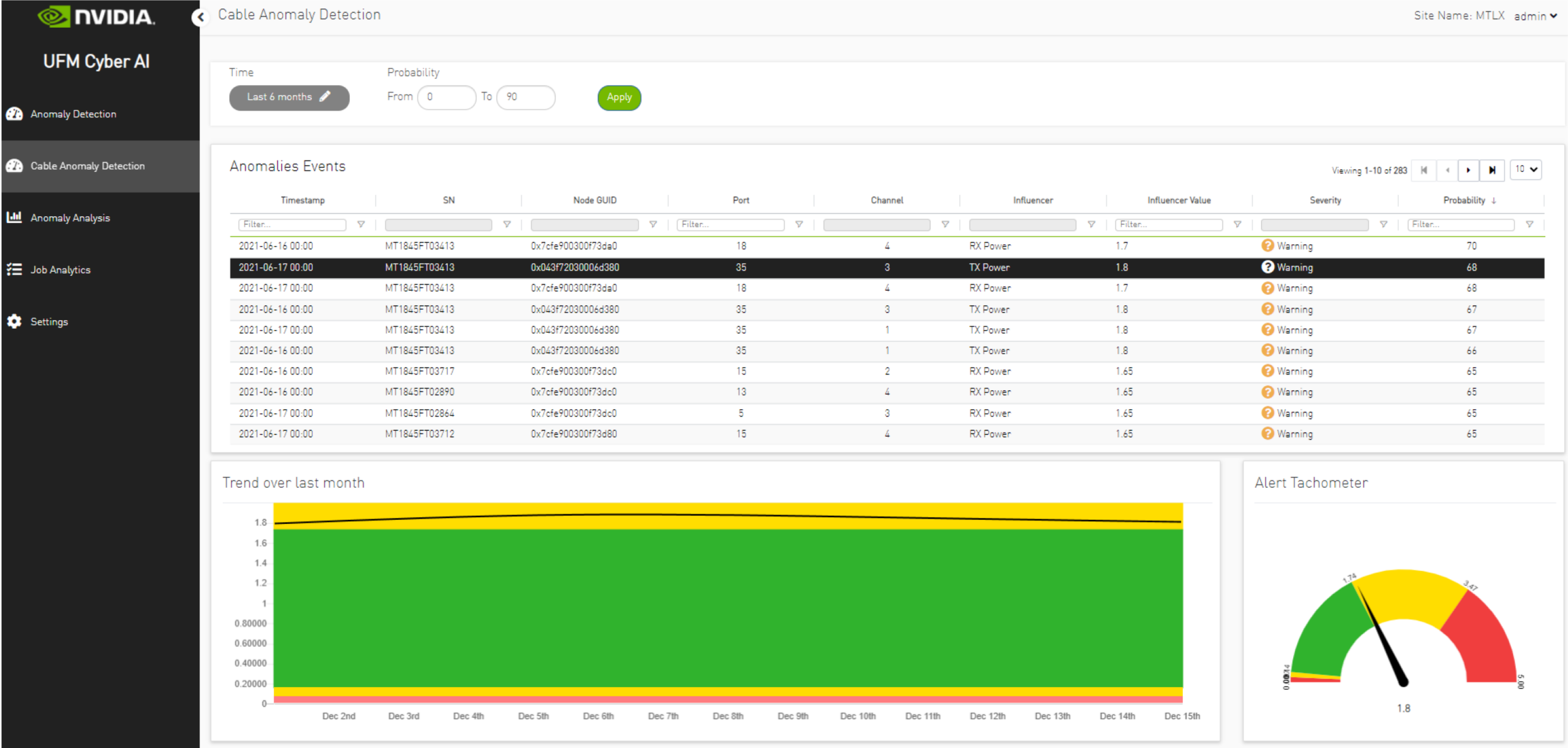


Anomalies per connection type:
Host to switch
Switch to Host
Switch to Switch

Filtered anomaly events
per selected pair

Anomalies over time
With a list of all influencers

NEW FEATURE - CABLE ANOMALY DETECTION

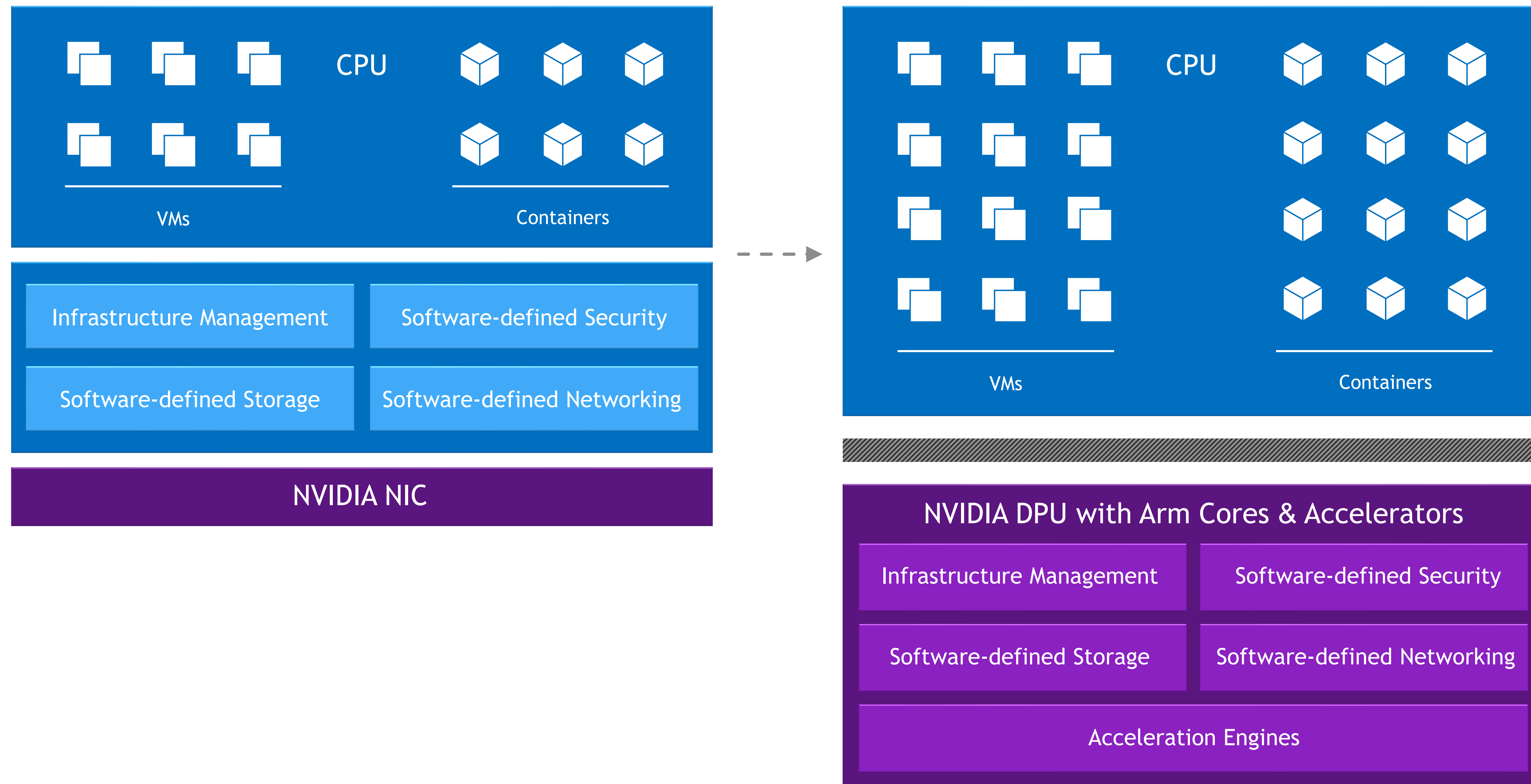


Cable Anomaly Events

TX/RX Power indication

INTRODUCING THE DATA PROCESSING UNIT

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

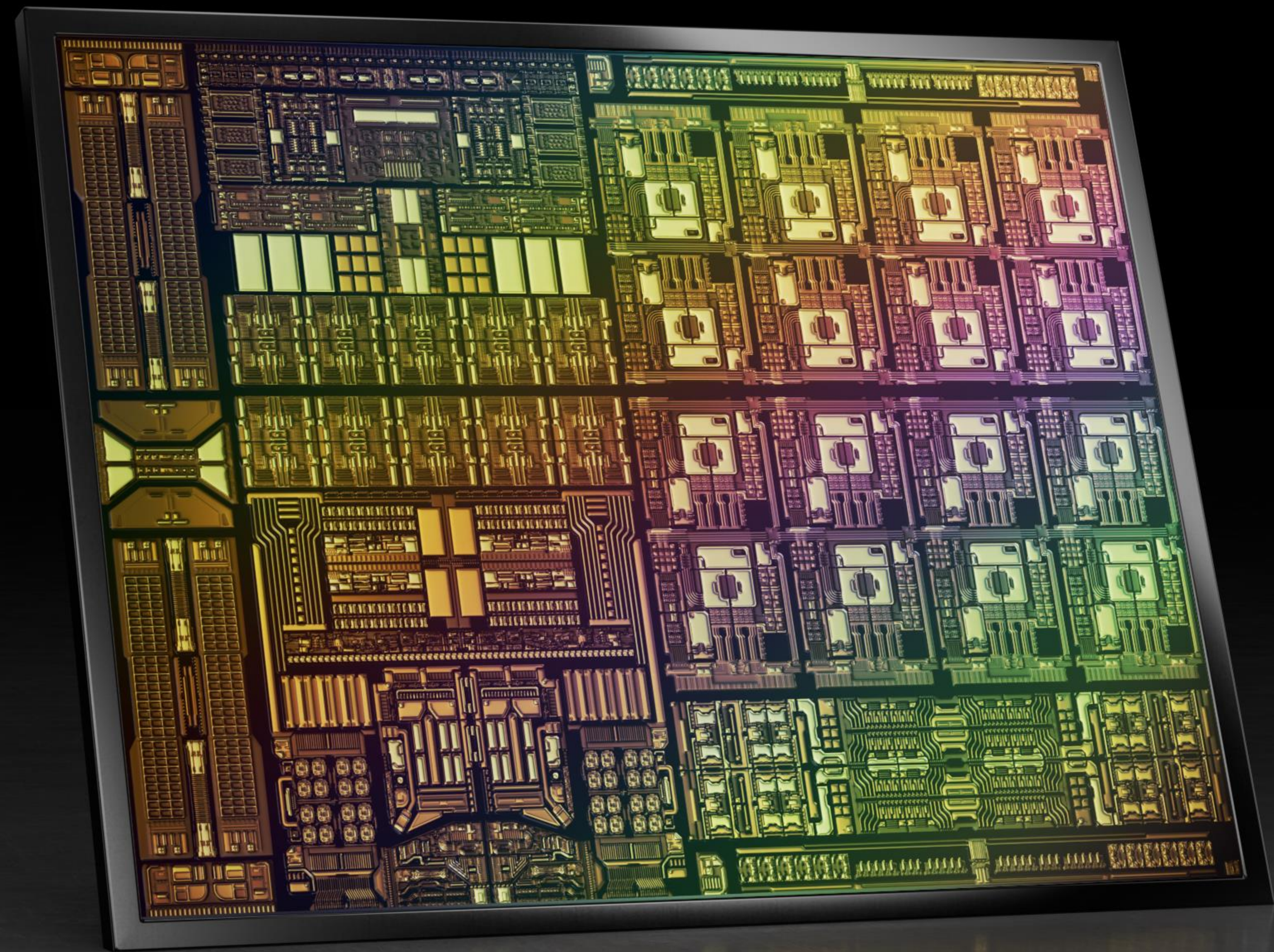


Software-Defined
Data Center Infrastructure on CPU

Software-Defined Hardware-Accelerated
Data Center Infrastructure on DPU

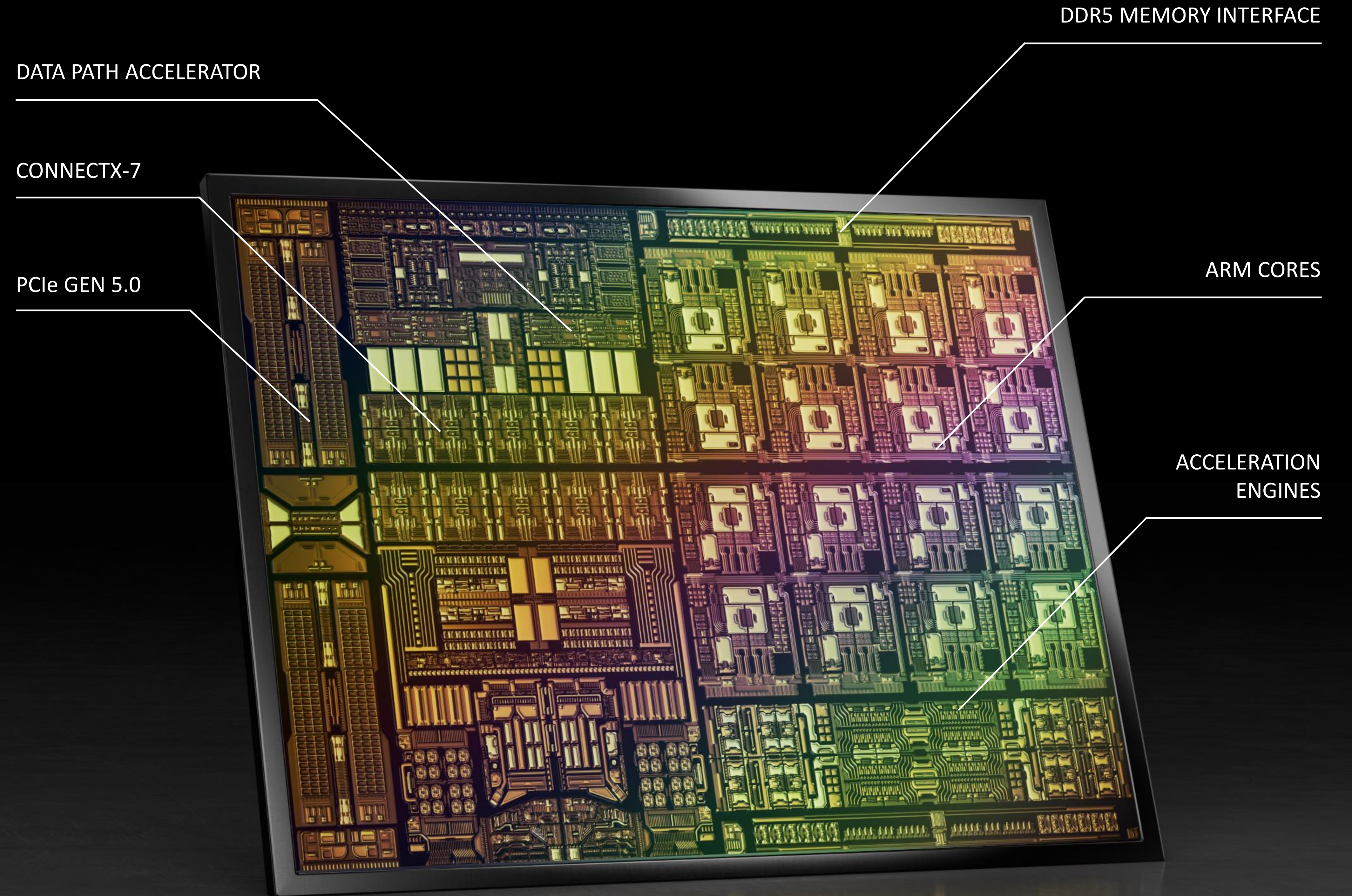
ANNOUNCING NVIDIA BLUEFIELD-3 DPU

- Offloads and Accelerates Data Center Infrastructure
- Isolates Application from Control and Management Plane
- Powerful CPU - 16x Arm A78 Cores
- Datapath Accelerator - 16x Cores, 256 Threads
- Process Networking, Storage, and Security at 400 Gbps



ANNOUNCING NVIDIA BLUEFIELD-3 DPU

- 22 Billion Transistors
- 400Gb/s Ethernet & InfiniBand Connectivity
- 400Gb/s Crypto Acceleration
- 18M IOP/s Elastic Block Storage
- 300 Equivalent x86 Cores



BLUEFIELD-3 PROGRAMMABLE ENGINES

ARM

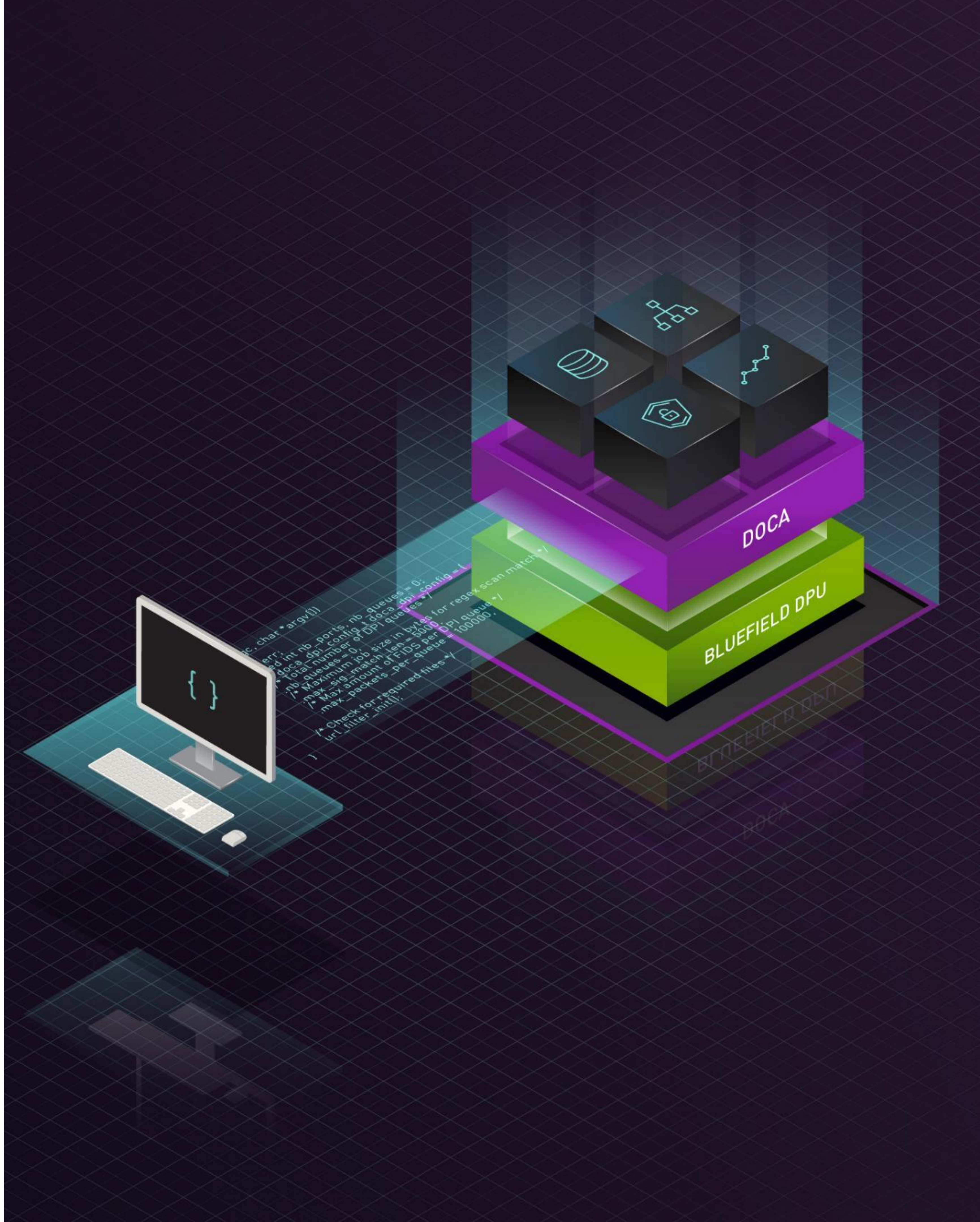
16 Arm A78 cores
Fully programmable OS
Apps/services, service chaining
Control Path / Slow Path

DATAPATH ACCELERATOR

16 cores, 256 threads
Programmability through DOCA
Heavy multi-threading application
acceleration

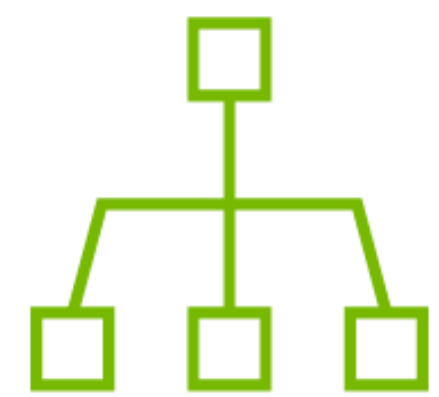
FLEXIBLE PIPELINE

Programmable packet processor
flow pipeline
Flow table based
Data Path



BLUEFIELD-3 ACCELERATION ENGINES

NETWORKING



- RDMA/RoCE
- Connection Tracking
- ASAP2
- SRIOV
- VirtIO-net - up to 80Mpps
- Time Sensitive Networking
- Timing Accuracy
- Streaming - Packet Pacing, FEC engine

STORAGE



- NVMe-OF - RDMA / TCP
- VirtIO-blk @18MIOPs, VirtIO-fs
- AES-XTS data-at-rest crypto
- Signature - CRC/T10DIF/SHA
- RAID - EC based
- Decompression - GZIP, ZLIB, LZ4

SECURITY



- IPsec, TLS, MACsec @400Gbs
- AES-GCM Bulk Crypto
- PKA@100K OP/s
- RegEx & DPI @50Gb/s
- Attestation
- TRNG

HPC / AI



- Communication libraries offload
- GPU Direct RDMA
- GPU Direct Storage

INDUSTRY-LEADING PERFORMANCE

Massive Advancements, Built for Cloud Scale

	BlueField-2	BlueField-3
Bandwidth	200Gb/s	400Gb/s
DPDK Max msg Rate	215Mpps	250Mpps
RDMA max msg rate	215Mpps	370Mpps
Compute	SPECINT2K17: 9.8	SPECINT2K17: 42
Memory BW	17GB/s	80GB/s
VirtIO Acceleration	40Mpps (*)	80Mpps (*)
Connections Per Second (CPS)	1.5M	6M
IPsec Acceleration	100Gb/s	400Gb/s
TLS Acceleration	200Gb/s	400Gb/s
MACsec Acceleration	X	400Gb/s
NVMe SNAP	5.4M IOPs @4K	10MIOPS @4K
NVMe TCP	2.1MIOPs	5MIOPs

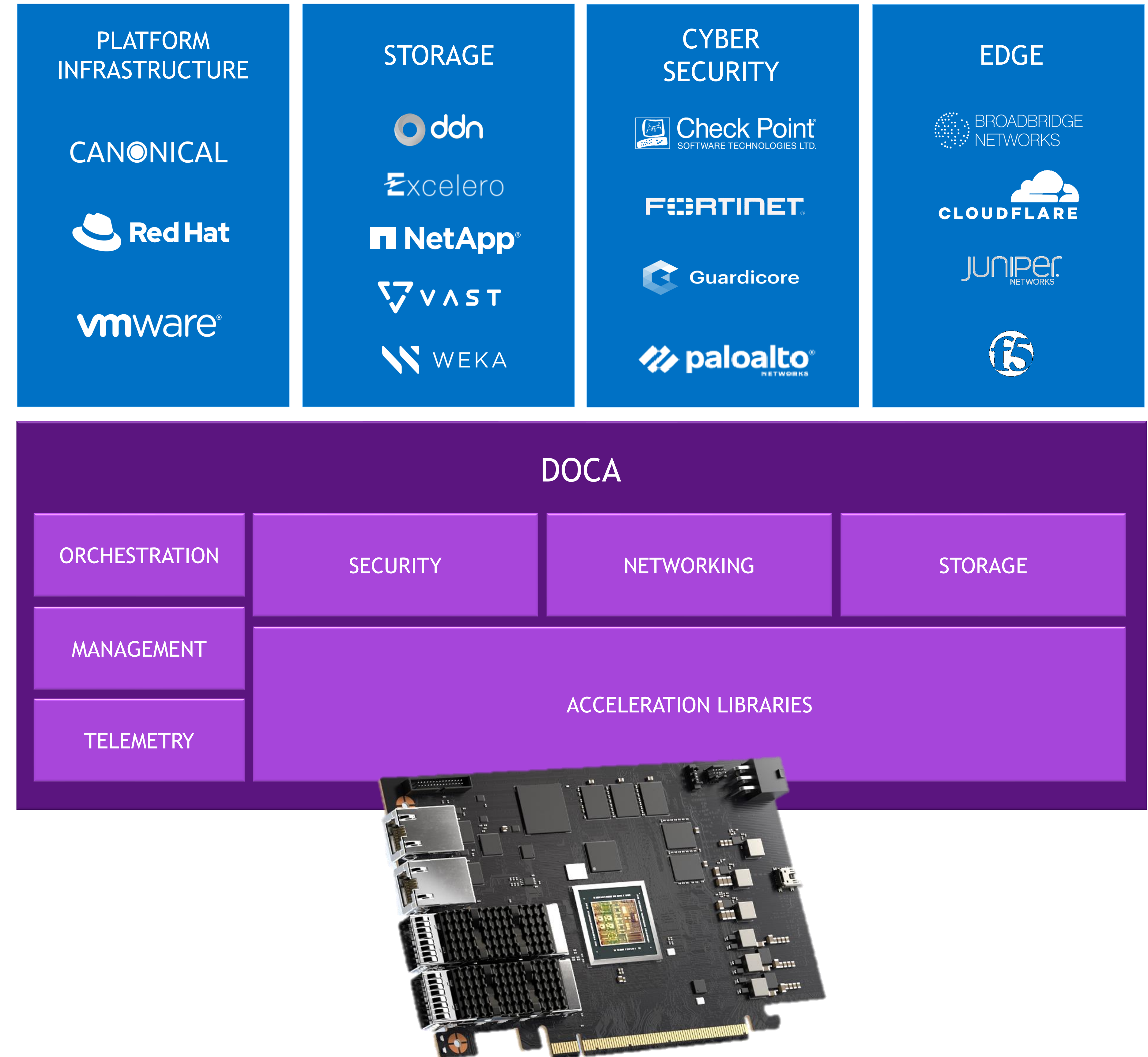
* Total packet rate for the sum of Tx and Rx



NVIDIA DOCA

Enabling Broad BlueField Partner Ecosystem

- Software Framework for BlueField DPUs
- Offload, Accelerate, and Isolate Infrastructure Processing
- Support for Hyperscale, Enterprise, Supercomputing and Hyperconverged Infrastructure
- Software Compatibility for Generations of BlueField DPUs
- DOCA is for DPUs what CUDA is for GPUs



BLUEFIELD ENABLES CLOUD-NATIVE SUPERCOMPUTING

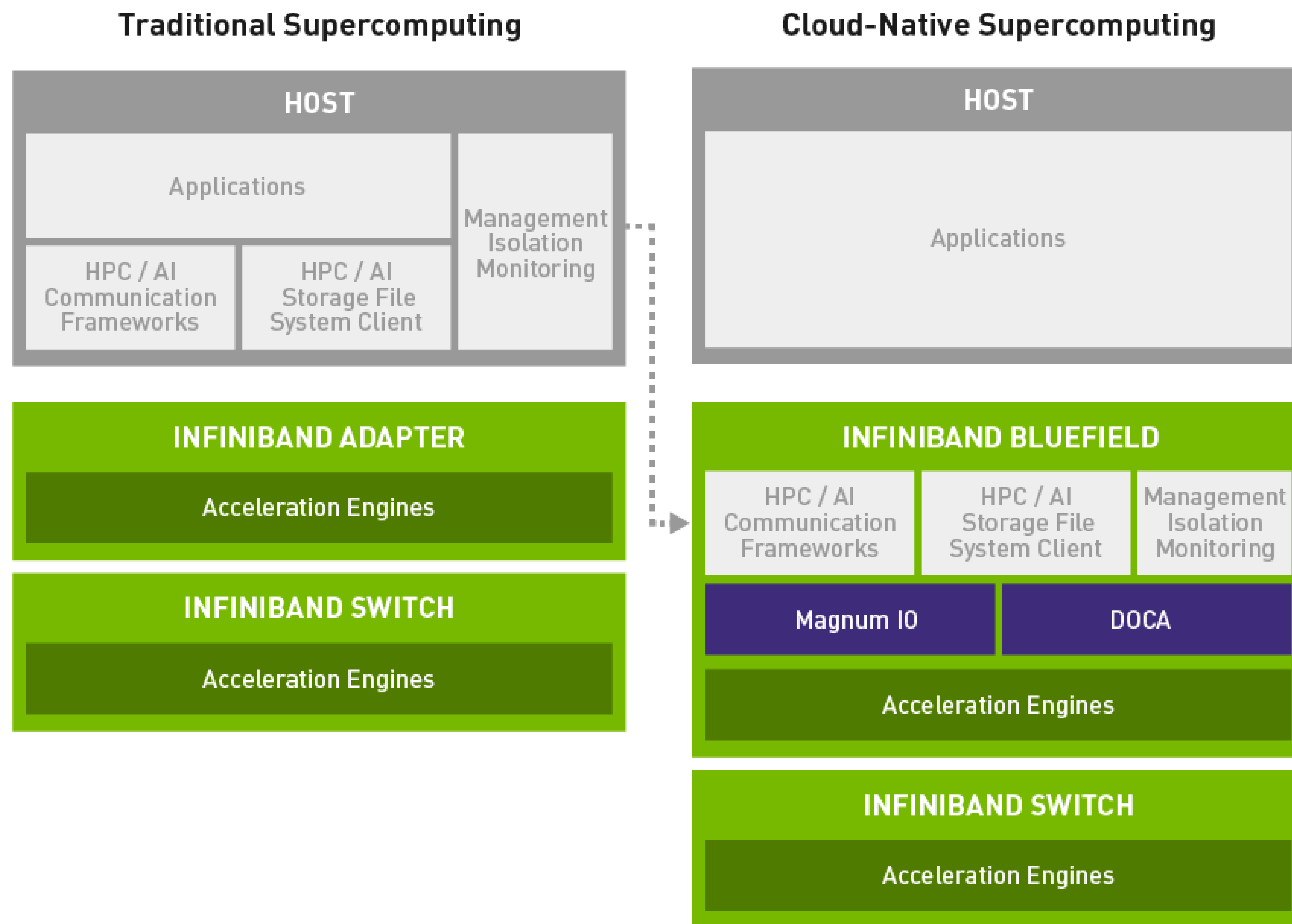
Multi-Tenancy with Zero-Trust Security

Collective offload with UCC accelerator

Smart MPI progression

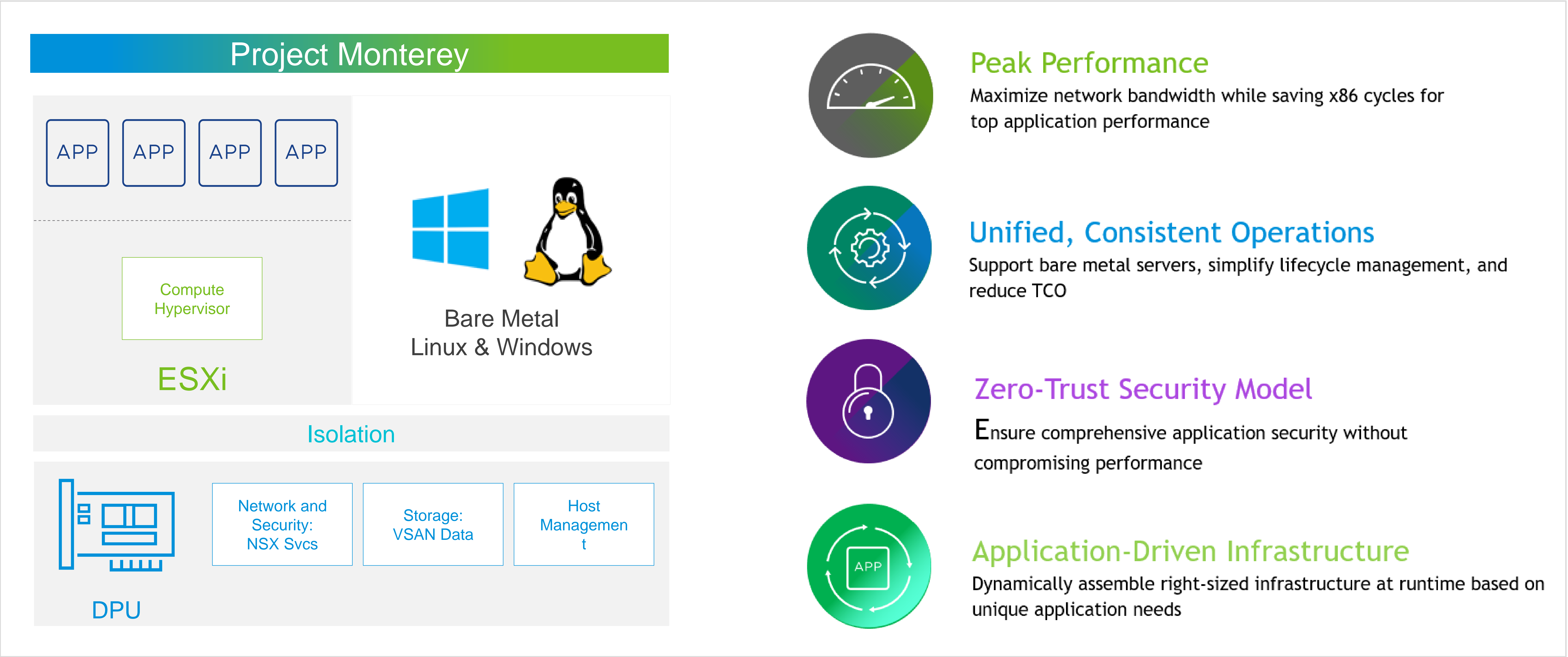
User-defined algorithms

1.4X higher application performance



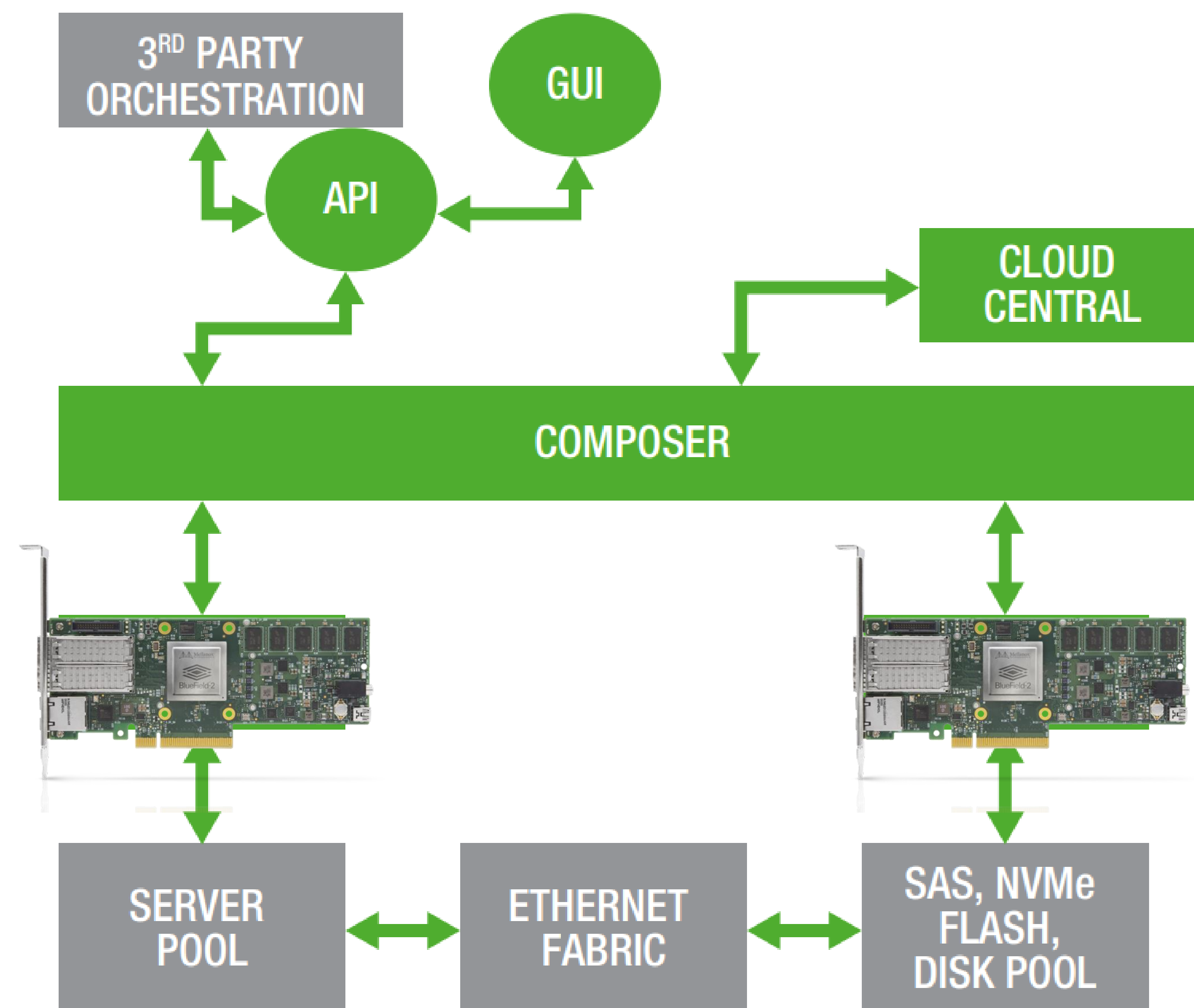
NVIDIA & VMWARE ENABLE HYBRID CLOUD ARCHITECTURE

Run Modern Workloads Efficiently Over New Composable, Disaggregated Infrastructure



FULLY AUTOMATED, SCALE-OUT NVME-OF STORAGE

High performance, continuously adaptable infrastructure for data-intensive applications



[DriveScale Blog](#)



Automated provisioning of networked storage to servers without using any host resources

Data analytics, ML/AI on any OS or hypervisor can now take advantage of scale-out NVMe storage

Instantly attach/detach data sets and storage, and replace failed components in seconds

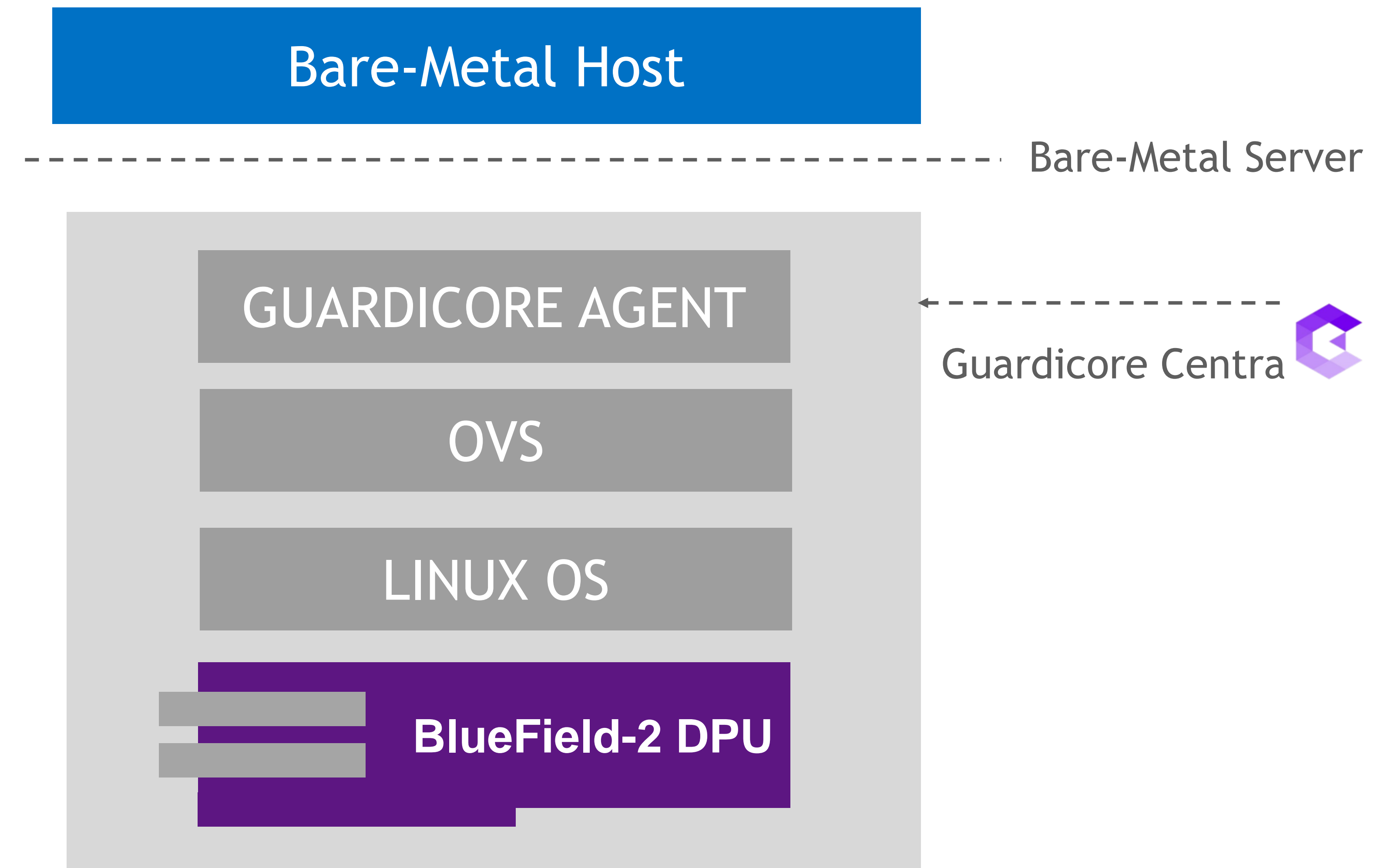
THE PATH TO AGENTLESS SEGMENTATION



Guardicore introduces complete network level visibility. Tracks connections and reports network events and their verdict

Guardicore enforcement policy is accelerate by the DPU hardware by offloading the segmentation rules

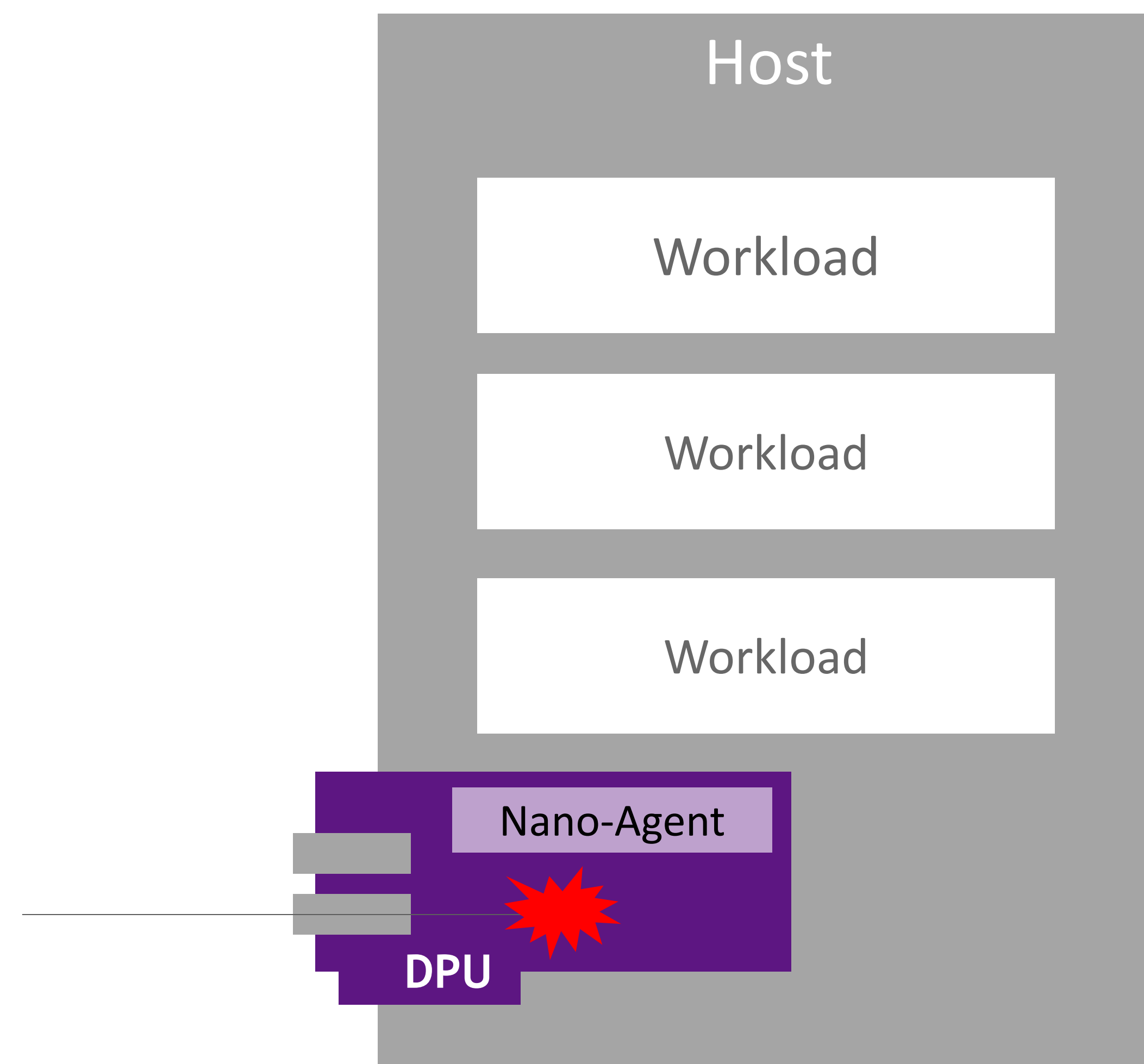
Guardicore agents running on BlueField cores in a separated trusted domain, enforce policies even on a compromised host



[Guardicore Blog](#)

CHECK POINT INFINITY CLOUD - DPU ACCELERATED

Protect, isolate and accelerate the cloud and edge



Check Point Infinity Nano-agents are deployed on the NVIDIA DPU to protect, isolate and accelerate the cloud and edge

The DPU & the Nano-agents enforce the distributed security policy created by the infinity centralized management

CheckPoint Tech & Nvidia are working together to accelerate security services with AI on every compute node

[CheckPoint Blog](#)

