



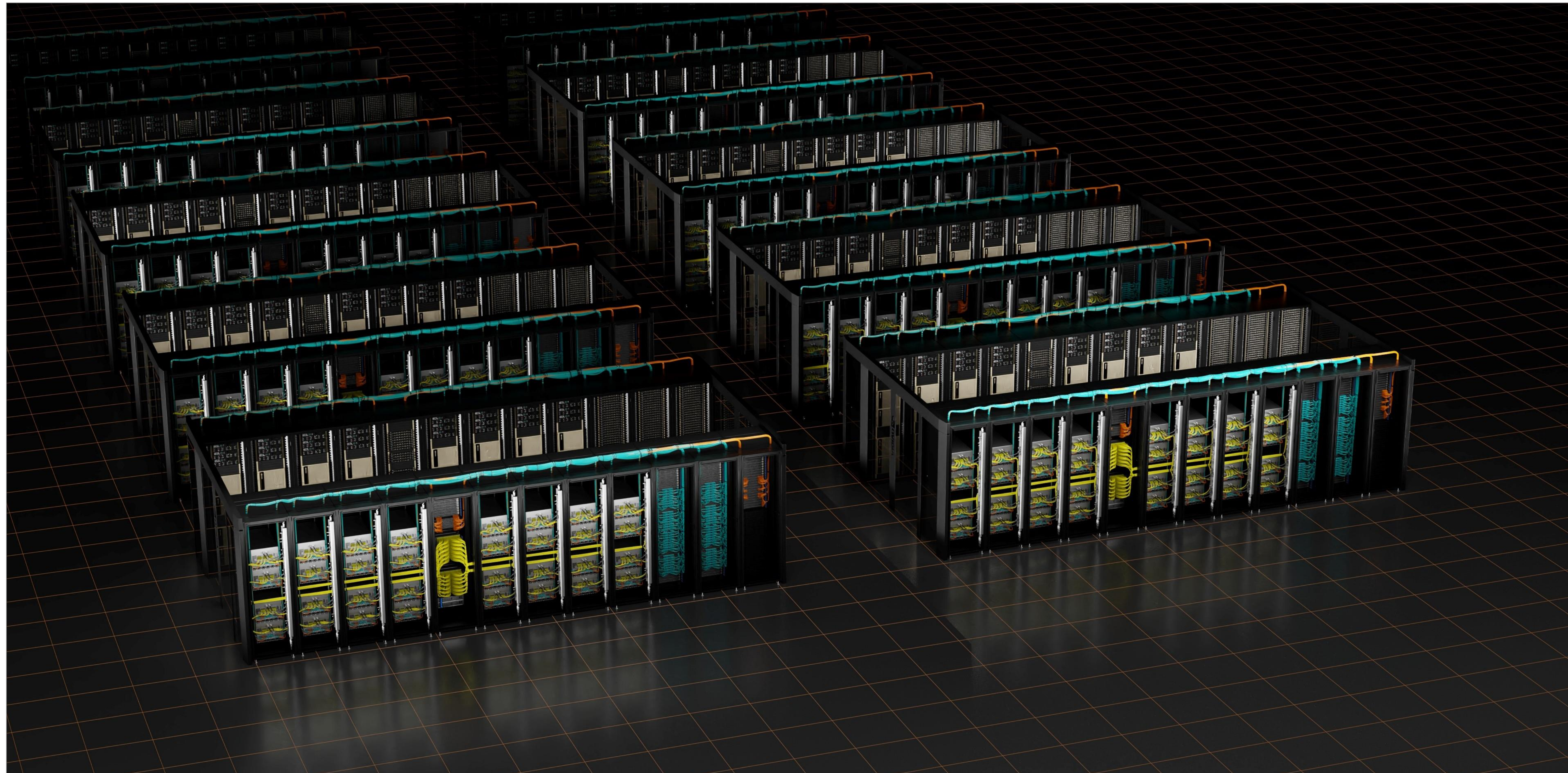
# **NVIDIA BlueField DPU Platforms**

Sebastian Kalcher, NVIDIA Solution Architect

October 2023

# Modern Data Centers are Becoming AI Factories

Producing Intelligence from Data



# Data Center Needs and Challenges

Traditional Infrastructure Not Equipped to Run Modern AI Applications



## Scale and Performance

End of Moore's Law  
Data Center Scale Computing  
Stringent Performance Requirements



## Efficient and Elastic

CPU Burden  
Resource Provisioning  
Data Centers are Power Limited

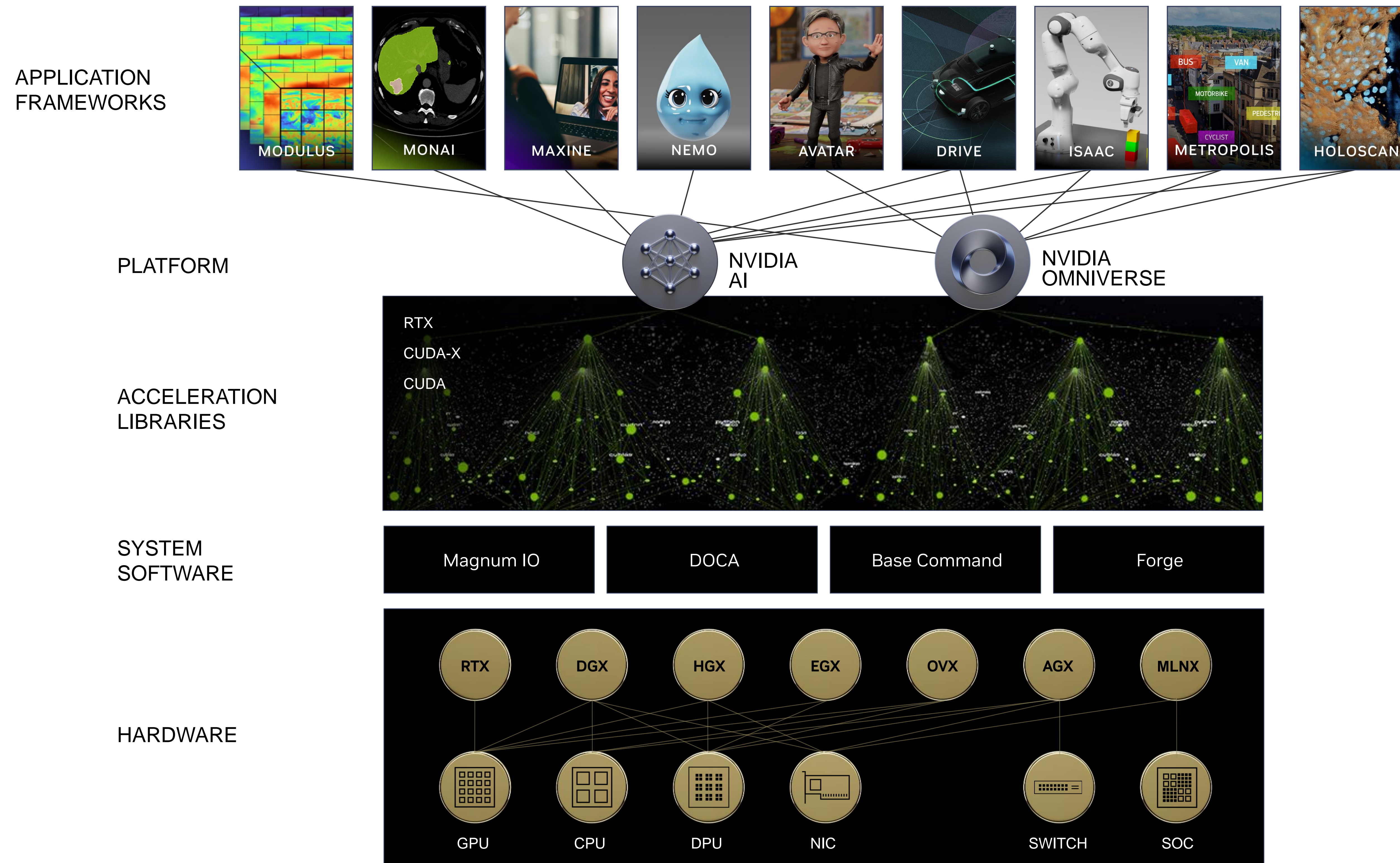


## Secure and Resilient Infrastructure

Multi-Tenant Environments  
Growing Cyber Threat Landscape  
Increased Attack Surface

# NVIDIA Accelerated Computing for Modern Data Centers

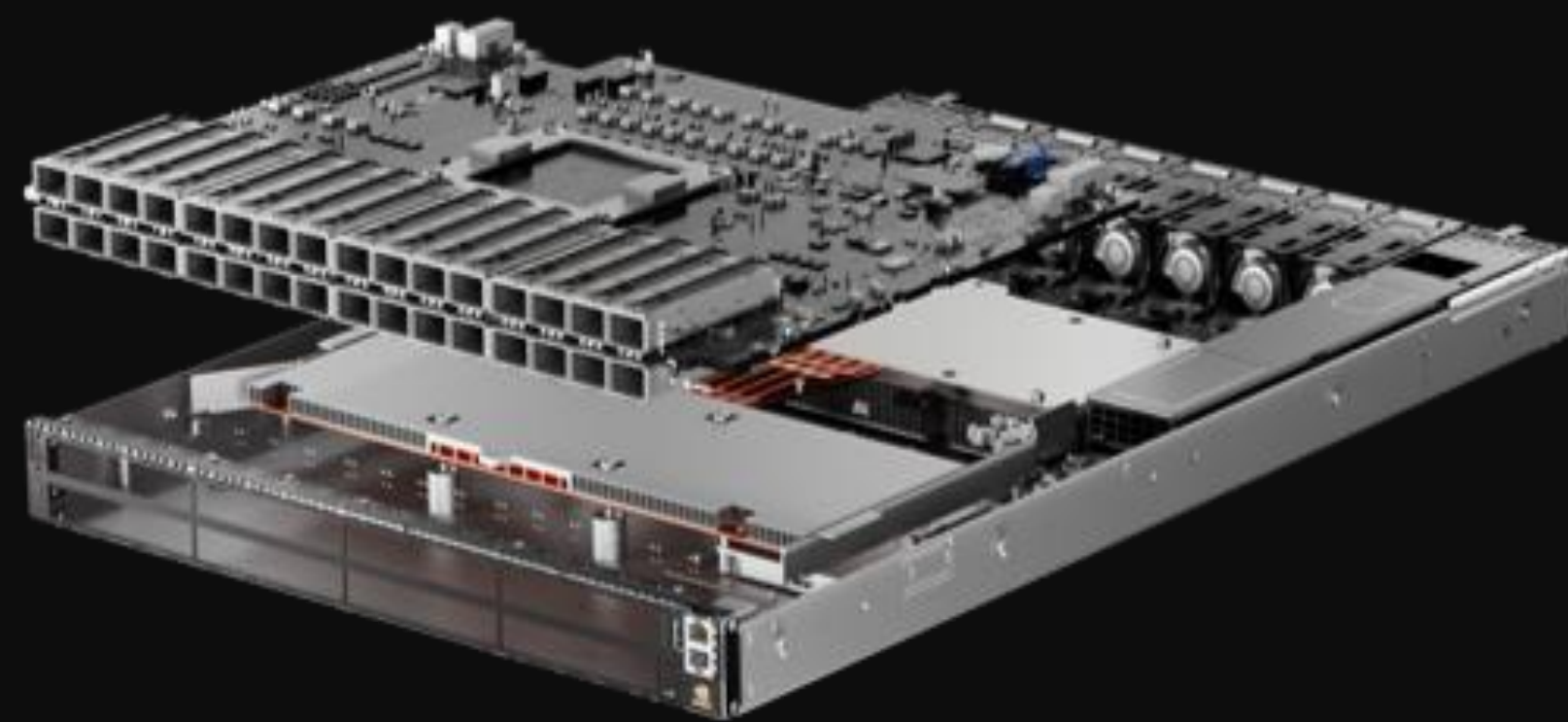
Accelerated Computing Services, Software and Systems Enabling New, Enhanced Business Models



# NVIDIA Networking Platforms

Accelerated Networking Solutions for the Era of AI

## Quantum-2 InfiniBand



### Extreme AI Performance

AI Factories and Cloud-Native Supercomputing

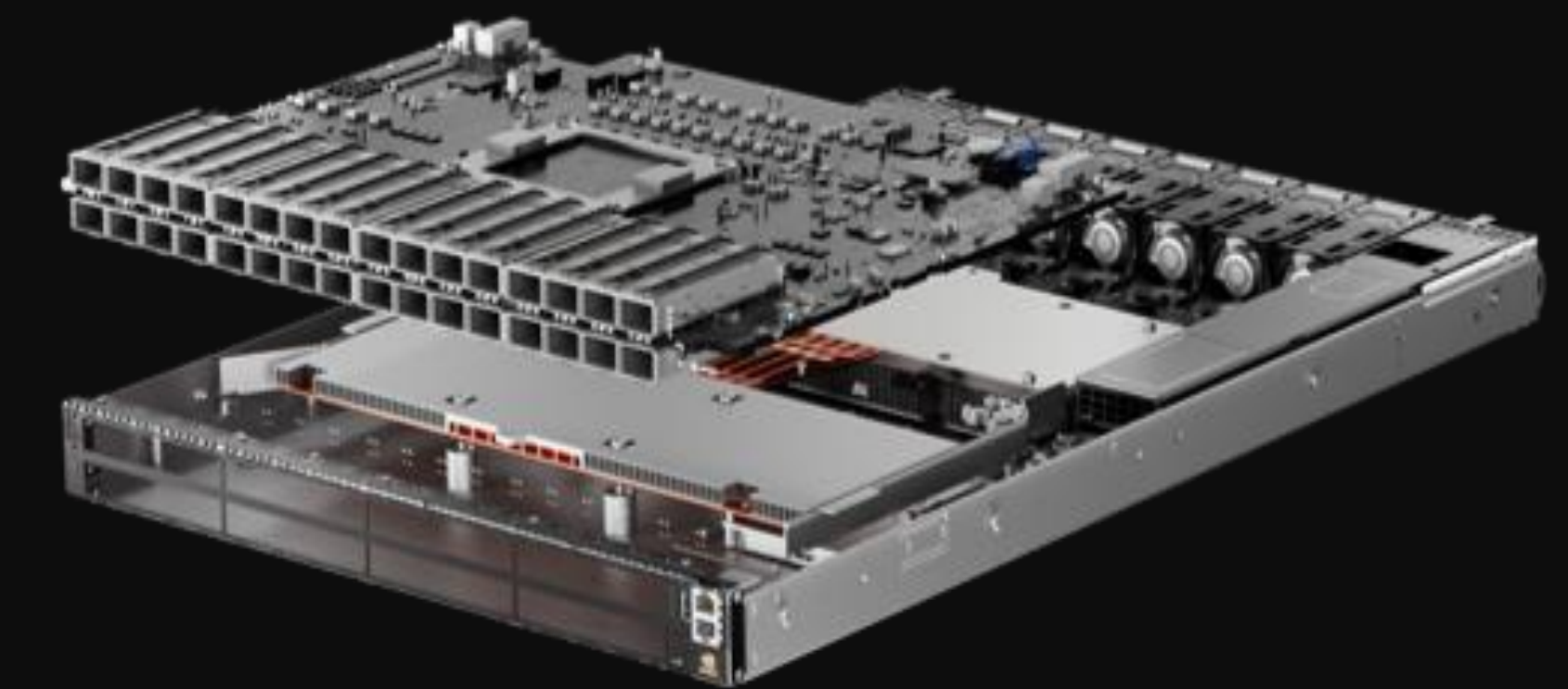
## BlueField-3 DPU



### Infrastructure Compute Platform

Offload, Accelerate, and Isolate Data Center Infrastructure

## Spectrum-4 Ethernet

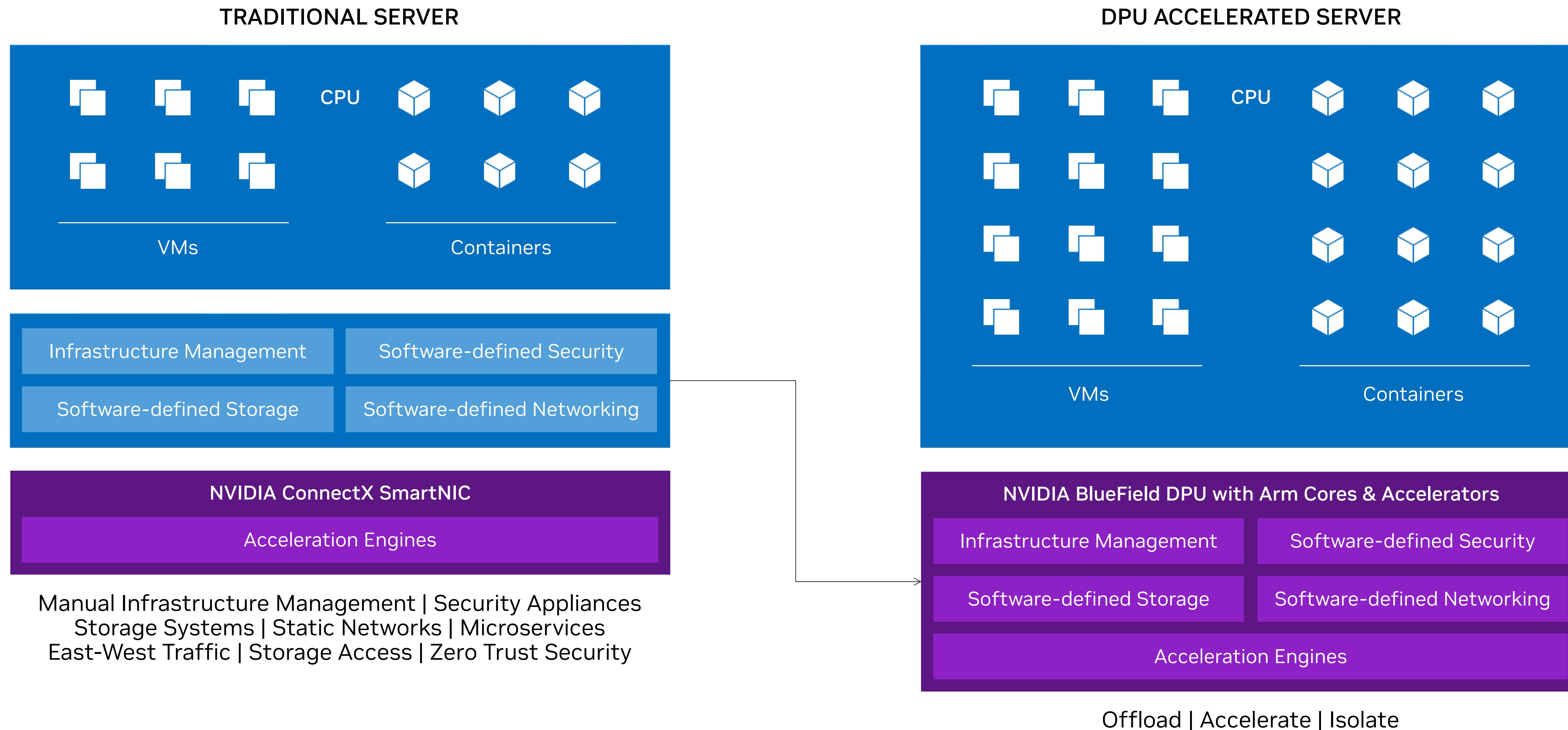


### Hyperscale Networking Platform

The Accelerated Cloud Fabric

# NVIDIA BlueField DPU Platform

Software-Defined, Hardware-Accelerated Infrastructure Compute Platform



# NVIDIA BlueField DPU Platform

Software-Defined, Hardware-Accelerated Infrastructure Compute Platform



## Accelerated Performance

Meet the most stringent performance requirements, run the most demanding workloads



## Cloud-Scale Efficiency

Free up x86 cores to business apps, achieve unprecedented scale and efficiency levels



## Robust Zero-Trust Security

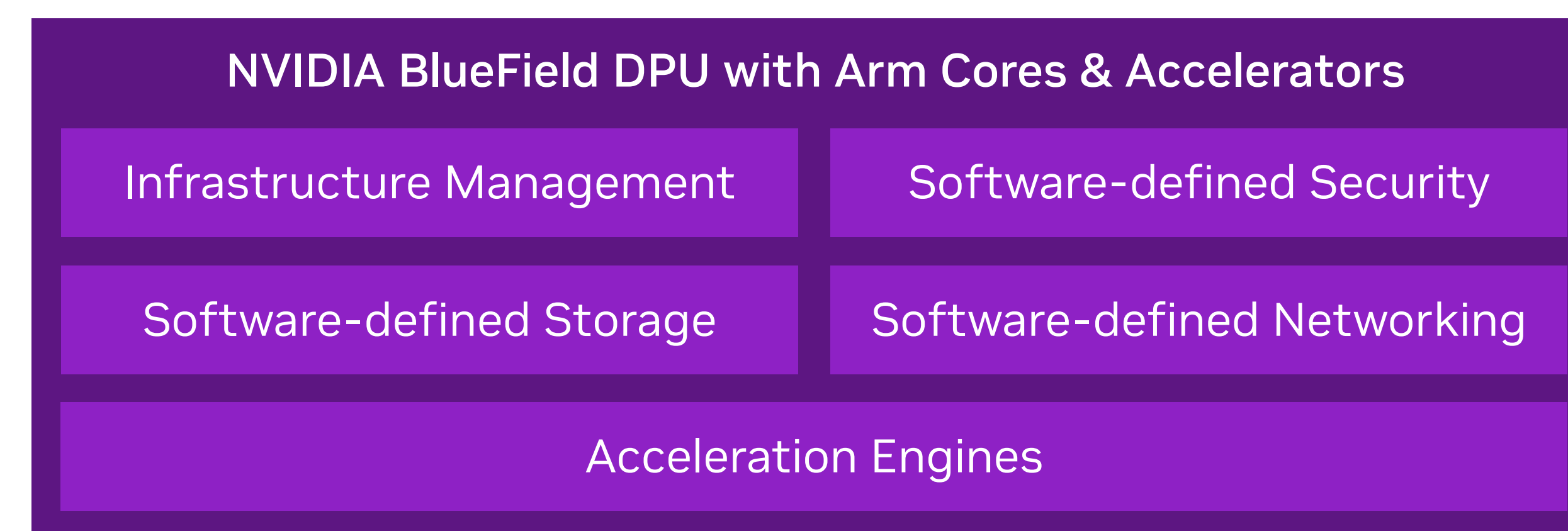
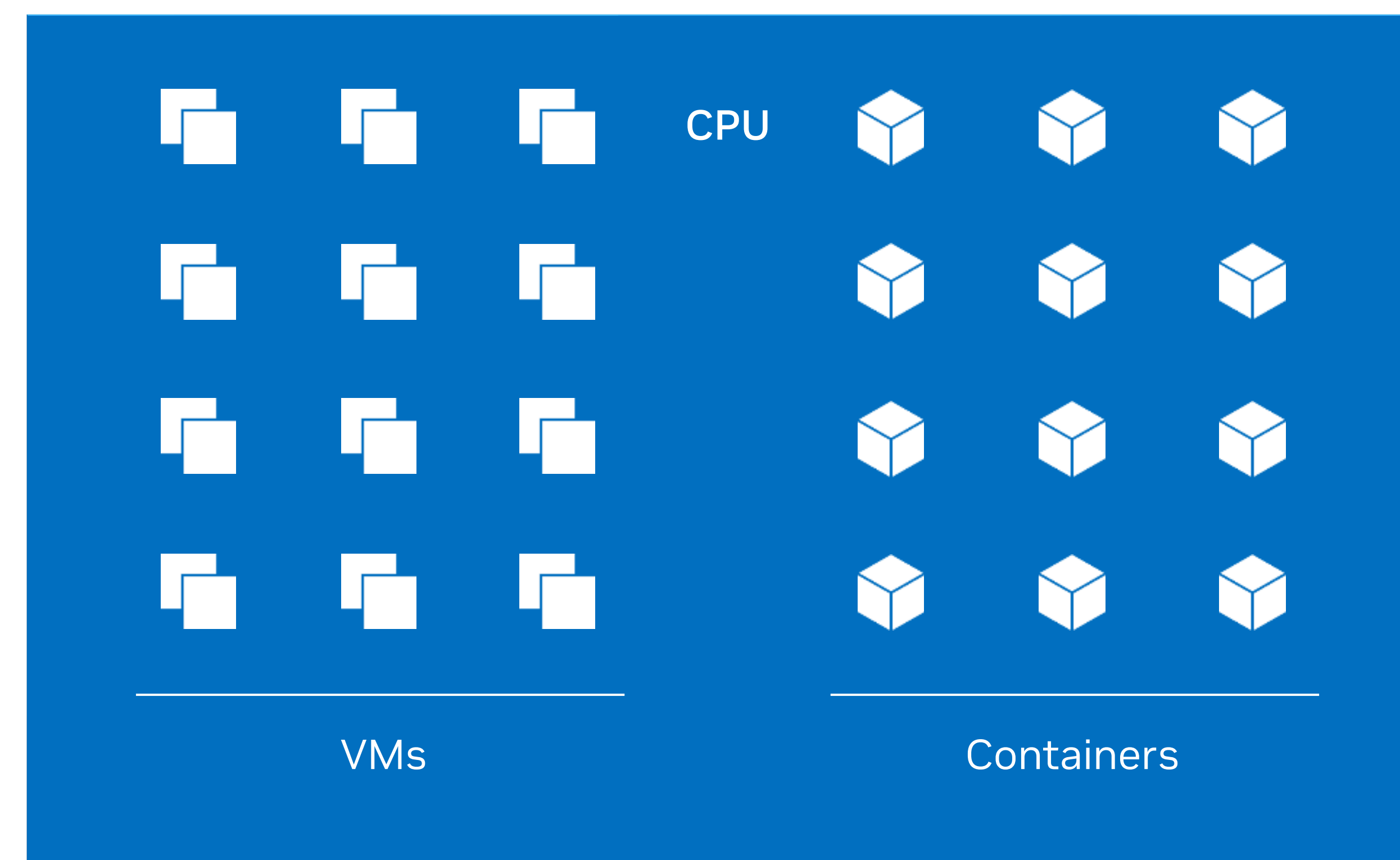
Ensure comprehensive data center security without compromising performance



## Programmable Infrastructure

Develop and run applications consistently with maximum performance

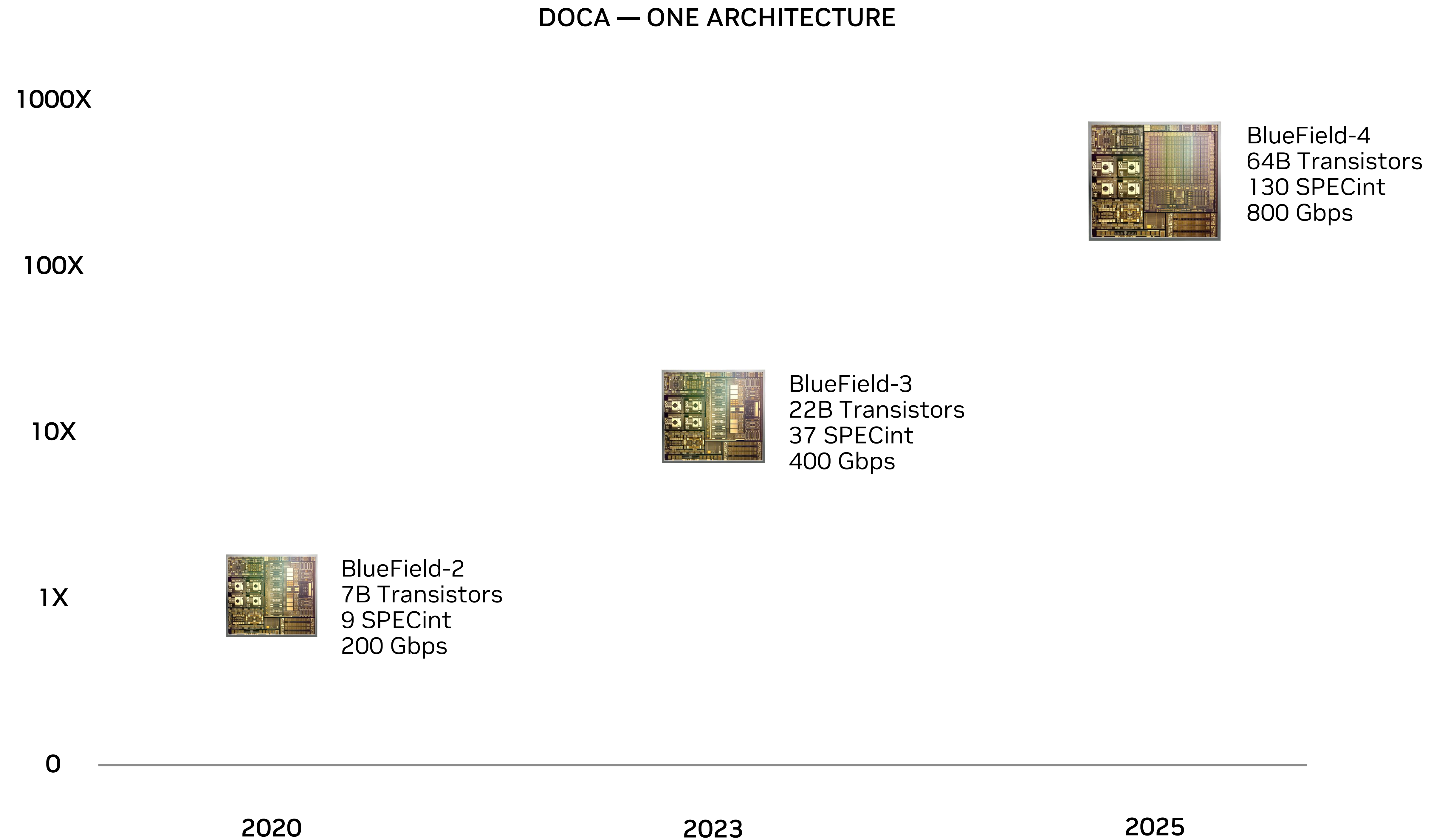
## DPU ACCELERATED SERVER



Offload | Accelerate | Isolate

# NVIDIA BlueField DPU Roadmap

Exponential Growth in Data Center Infrastructure Processing

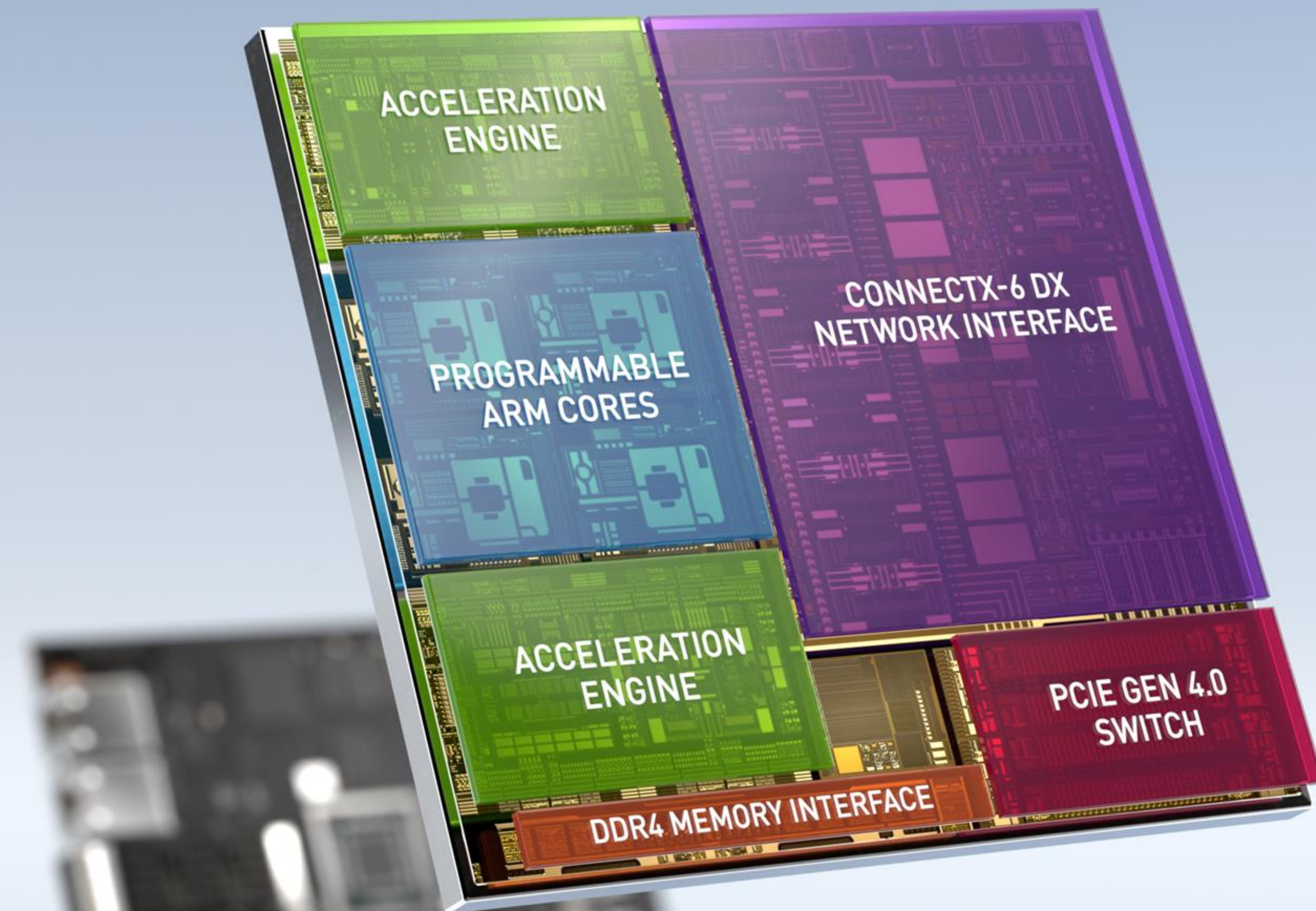




# NVIDIA BlueField-2 DPU

Data center infrastructure on a chip

- Combines 8 64-bit A72 Arm Cores, Acceleration Engines, and NVIDIA ConnectX-6 Dx NIC
- Accelerated Security: Isolation, Root of Trust, Crypto, Key Management, Regular Expression Engine
- Accelerated Networking: RDMA/RoCE, GPUDirect, SDN/NFV
- Accelerated Storage: NVMe-oF, Elastic Block Storage, Data Integrity, De-Dup, Compression



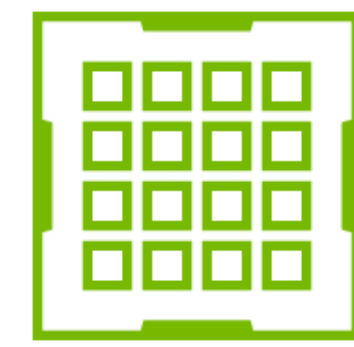
# NVIDIA BlueField-3 Overview

400Gb/s Infrastructure Compute Platform



## 400Gb Networking

RDMA/RoCE Accelerations  
SDN/NFV Accelerations  
Precision Timing



## Programmable Engines

16 x 64-bit A78 Arm Cores  
16 Hyperthreaded DPA Cores  
Accelerated Pipeline



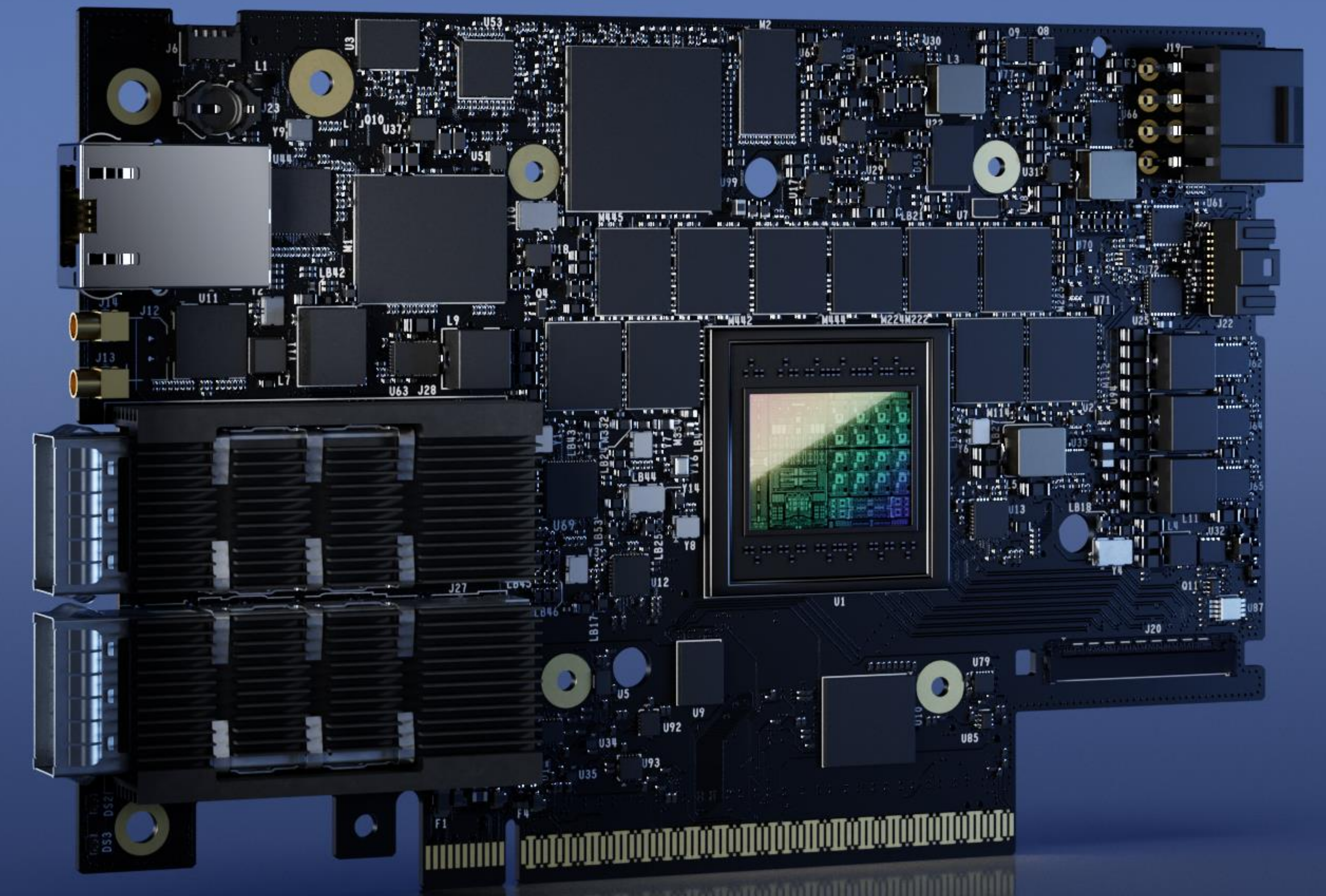
## Zero-Trust Security

Platform Security  
Crypto Accelerations  
Zero-Trust Infrastructure



## Composable Storage

Storage Disaggregation  
NVMe-oF, NVMe/TCP  
Storage Encryption



# NVIDIA BlueField-3 Overview

Massive Advancements, Built for Cloud Scale



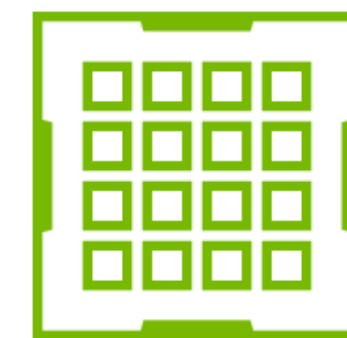
## 400Gb Networking

- 2X Network Bandwidth
- 2X Network Pipeline
- 4X Host Bandwidth



## Zero-Trust Security

- 4X IPsec Acceleration
- 2X TLS Acceleration
- New MACsec Acceleration



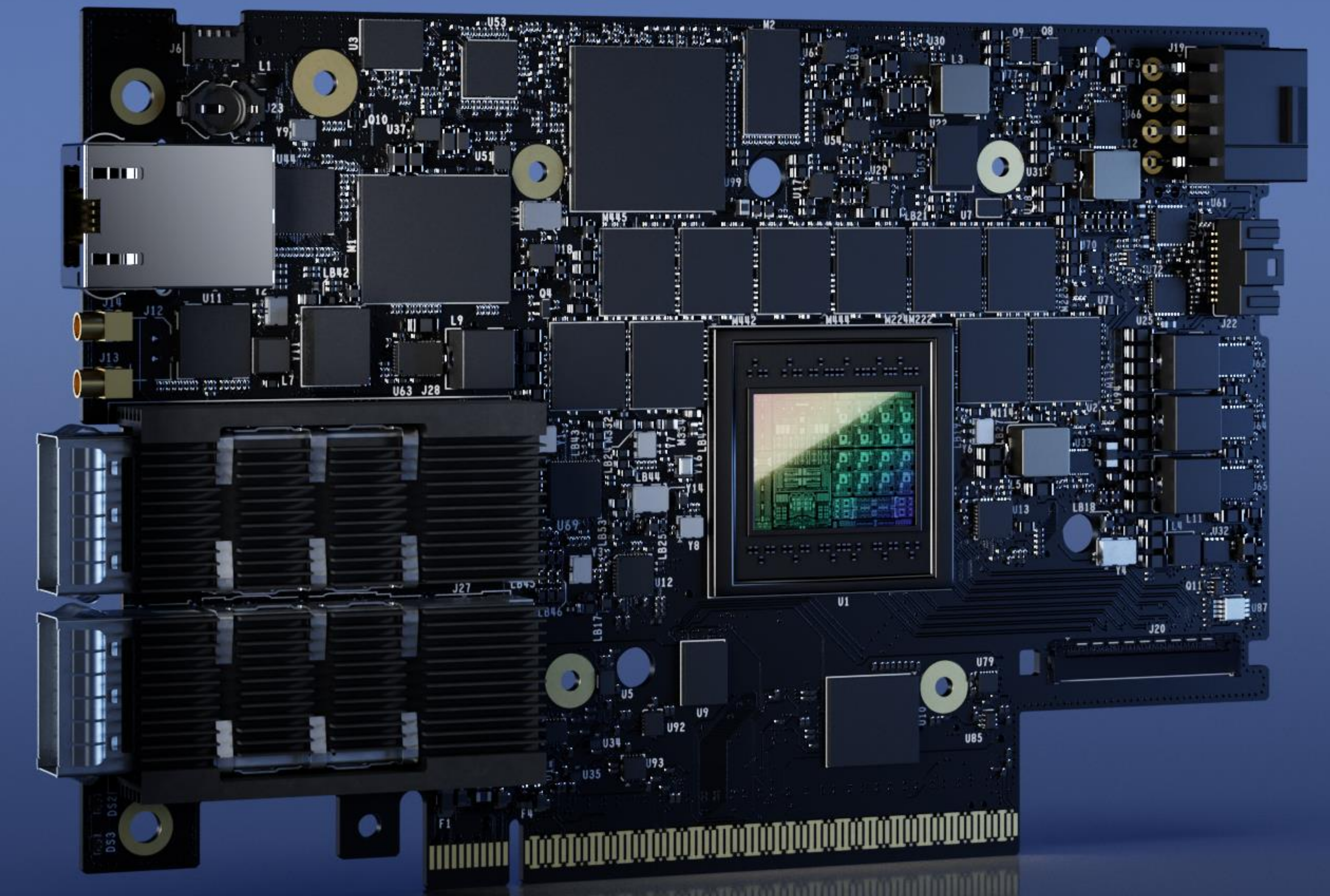
## Programmable Engines

- 4X Arm Compute
- 5X Memory
- New Datapath Accelerator



## Composable Storage

- 2X Storage IOPs
- 2X Storage Encryption
- New NVMe/TCP Acceleration

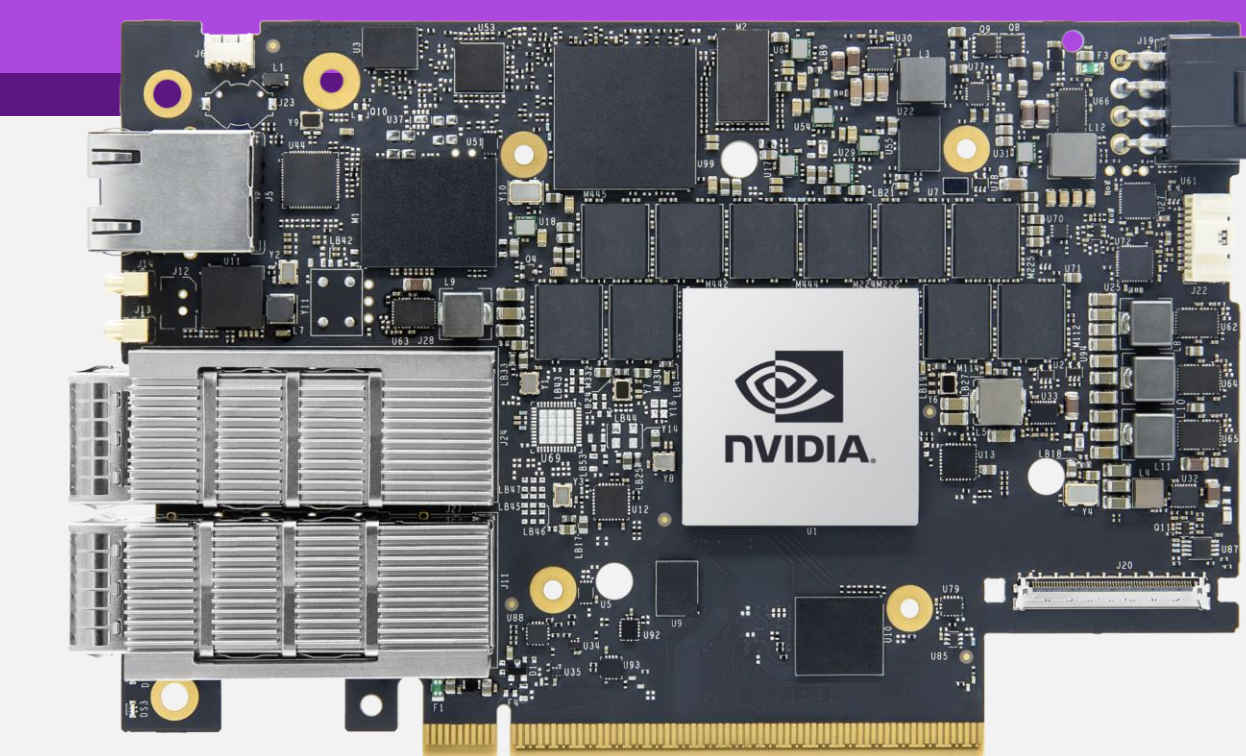
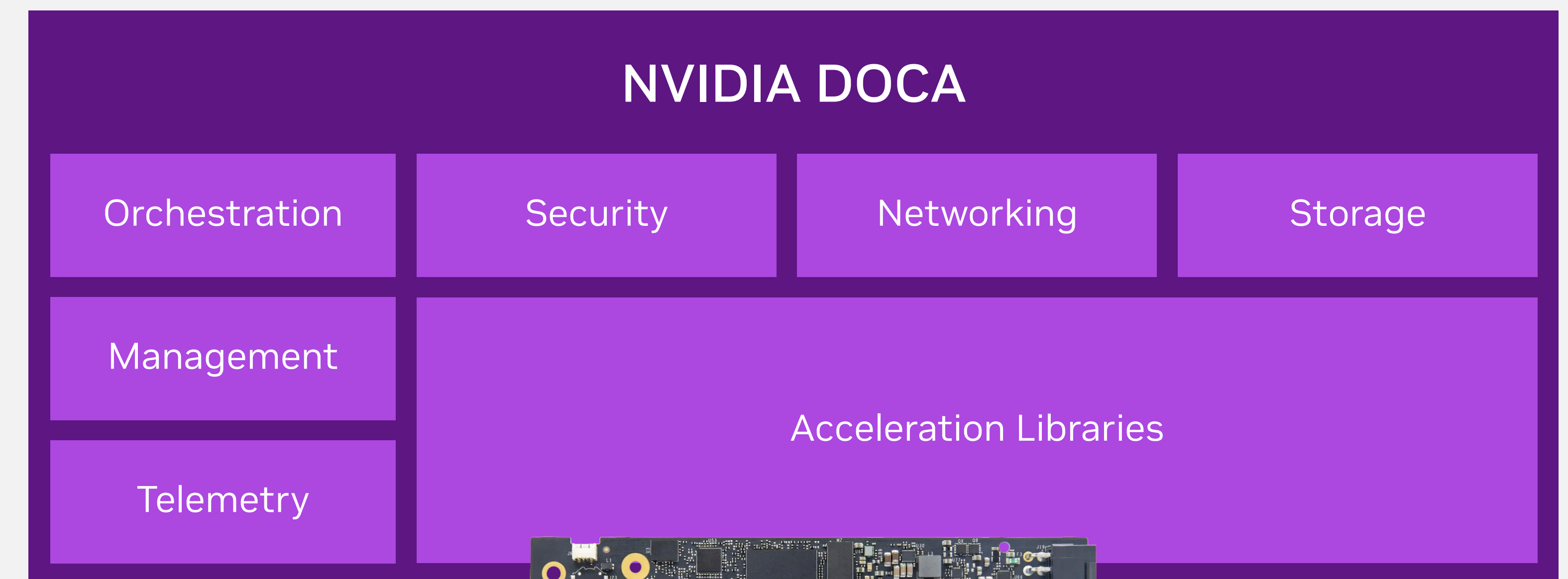


*\* Compared to previous BlueField generation*

# NVIDIA DOCA

## Comprehensive Acceleration SDK for BlueField DPUs

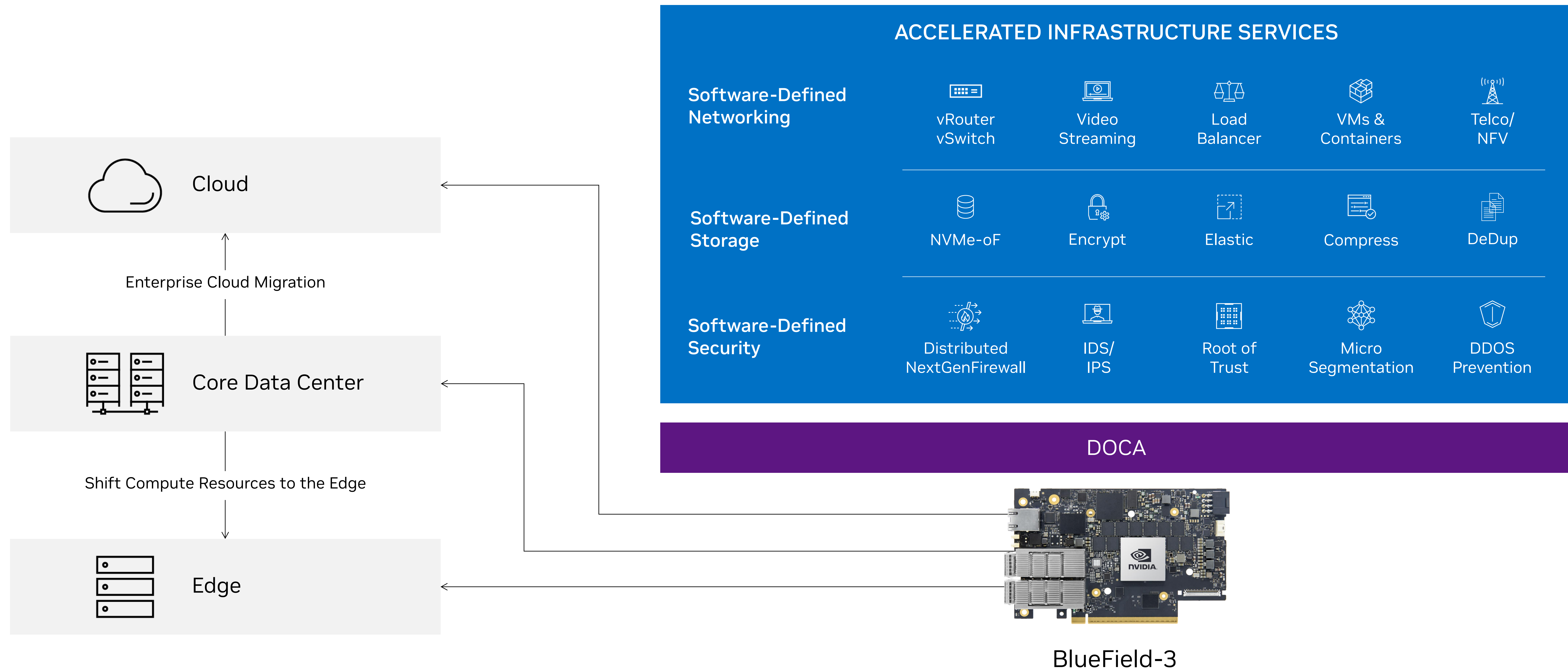
- Unified software framework for BlueField DPUs
- Offload, accelerate, and isolate infrastructure processing
- Support for hyperscale, enterprise, supercomputing and hyperconverged infrastructure
- Software compatibility for generations of BlueField DPUs
- Rich partner ecosystem



BlueField DPU

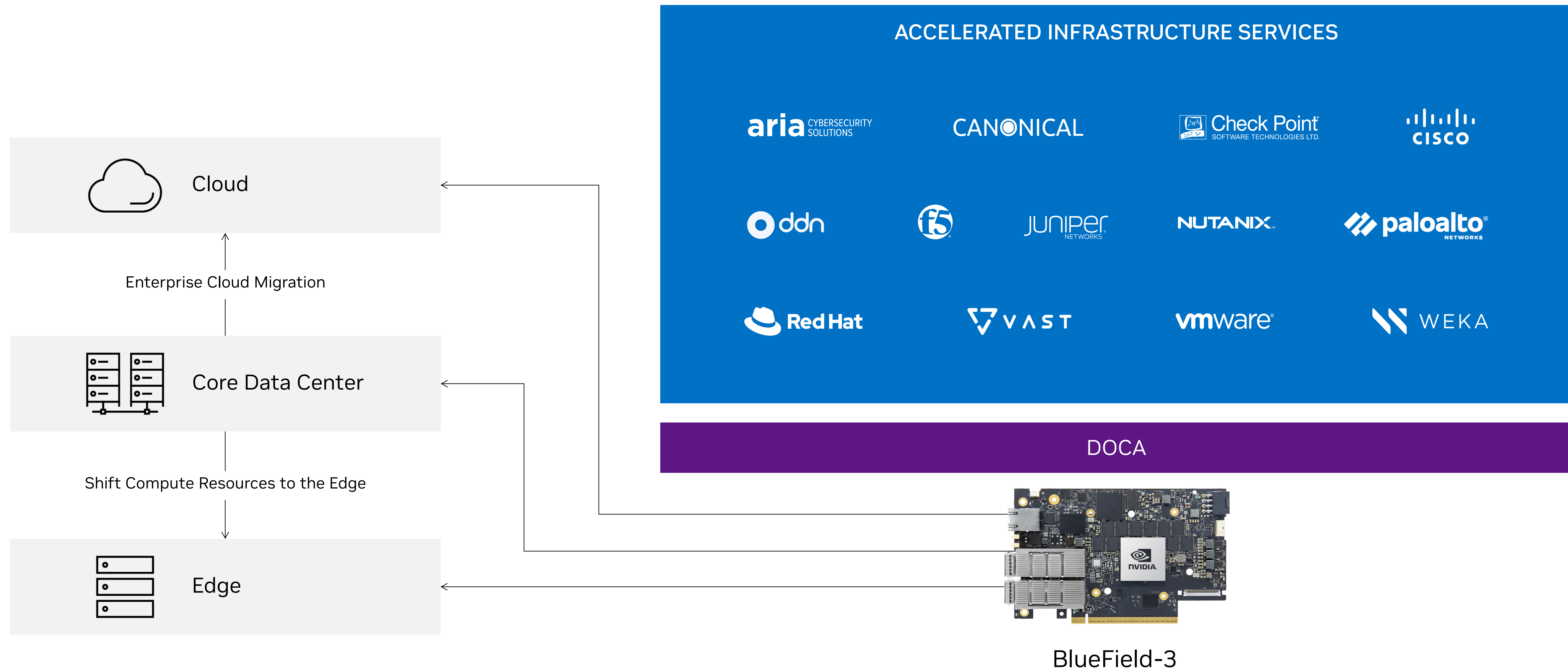
# BlueField is a Cloud Services Compute Platform

NVIDIA BlueField Accelerates Data Center Infrastructure Services from Cloud to Edge



# BlueField is a Cloud Services Compute Platform

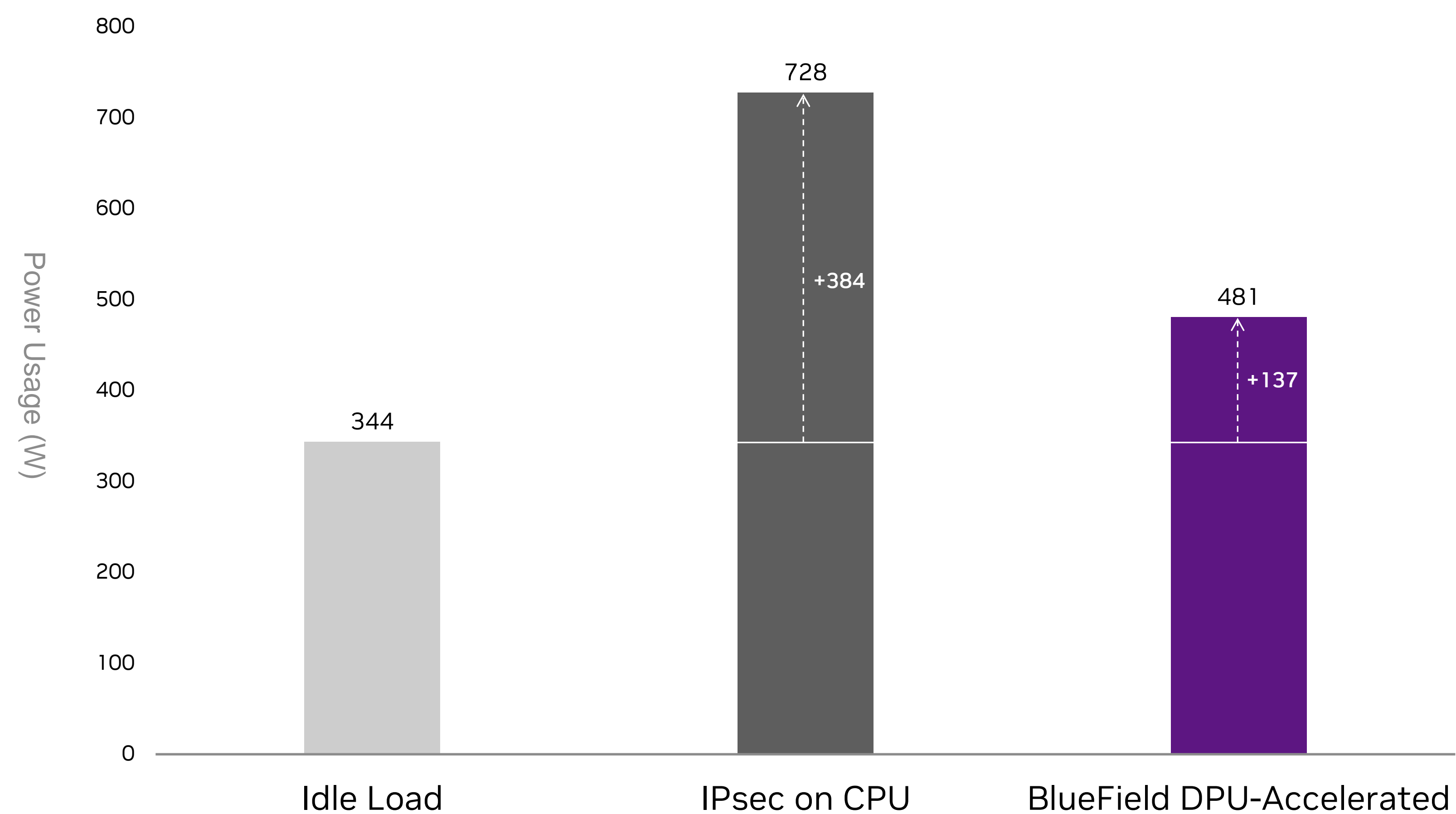
NVIDIA BlueField Accelerates Data Center Infrastructure Services from Cloud to Edge



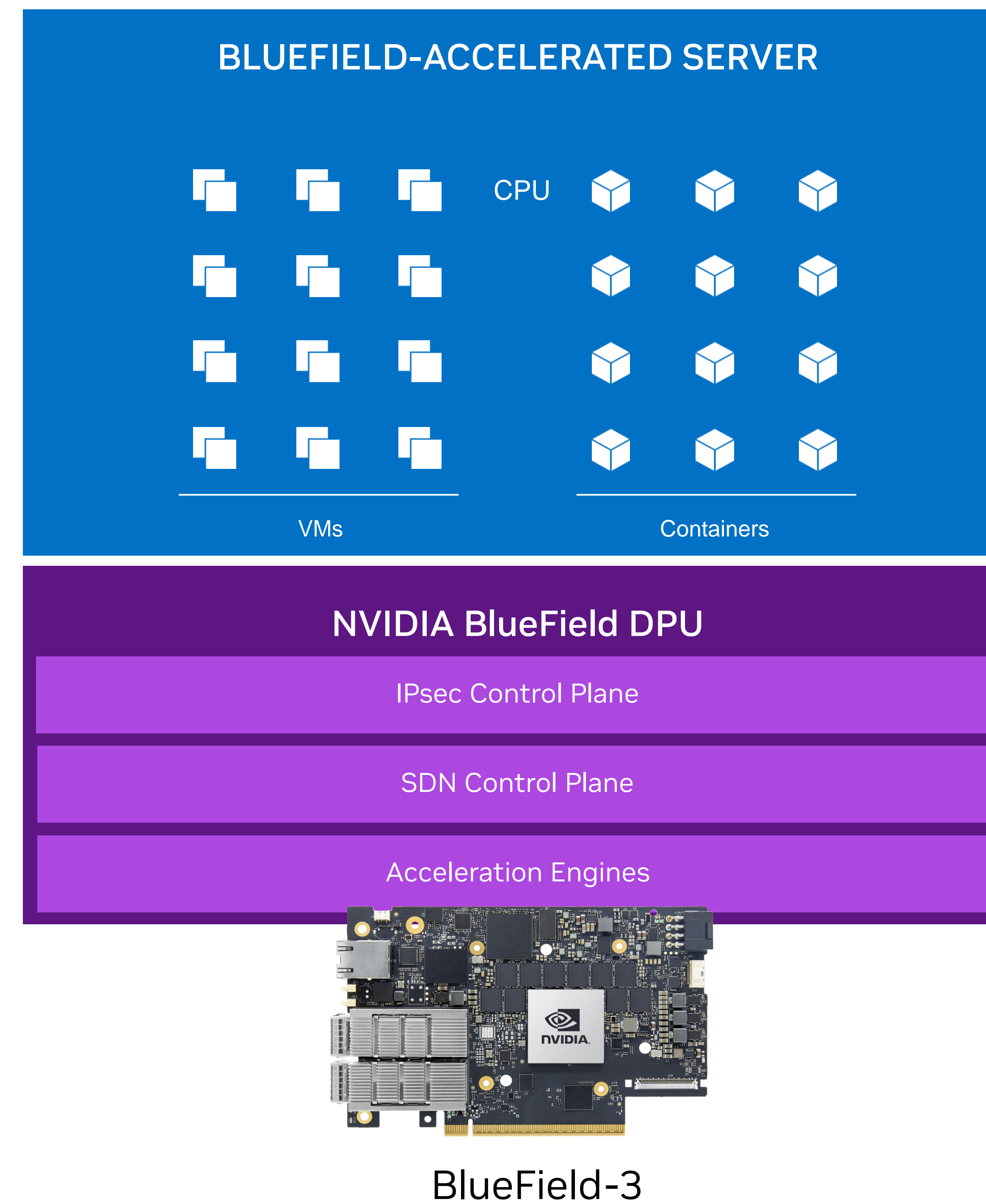
# Accelerated Computing is Sustainable Computing

BlueField-3 Enables Power-Efficient Cloud Data Centers

## 2.8X Better Performance/Watts



\* Compared to idle load power consumption, BlueField-2 test benchmarks



# NVIDIA BlueField Accelerates Infrastructure and Applications

Unprecedented Innovation for Modern Data Centers



## Cloud Computing

Bare-Metal | Virtualized | Containerized  
Private | Public | Hybrid Cloud



## Cybersecurity

Distributed Security | NGFW | Micro-segmentation



## HPC & AI

Scientific Computing | Accelerated DLRM



## Telco & Edge

Telco Cloud | CloudRAN | Edge Compute



## Data Storage

HCI | Elastic Block Storage | Instance Storage



## Media Streaming

Visual High Quality | 8K Video | CDN

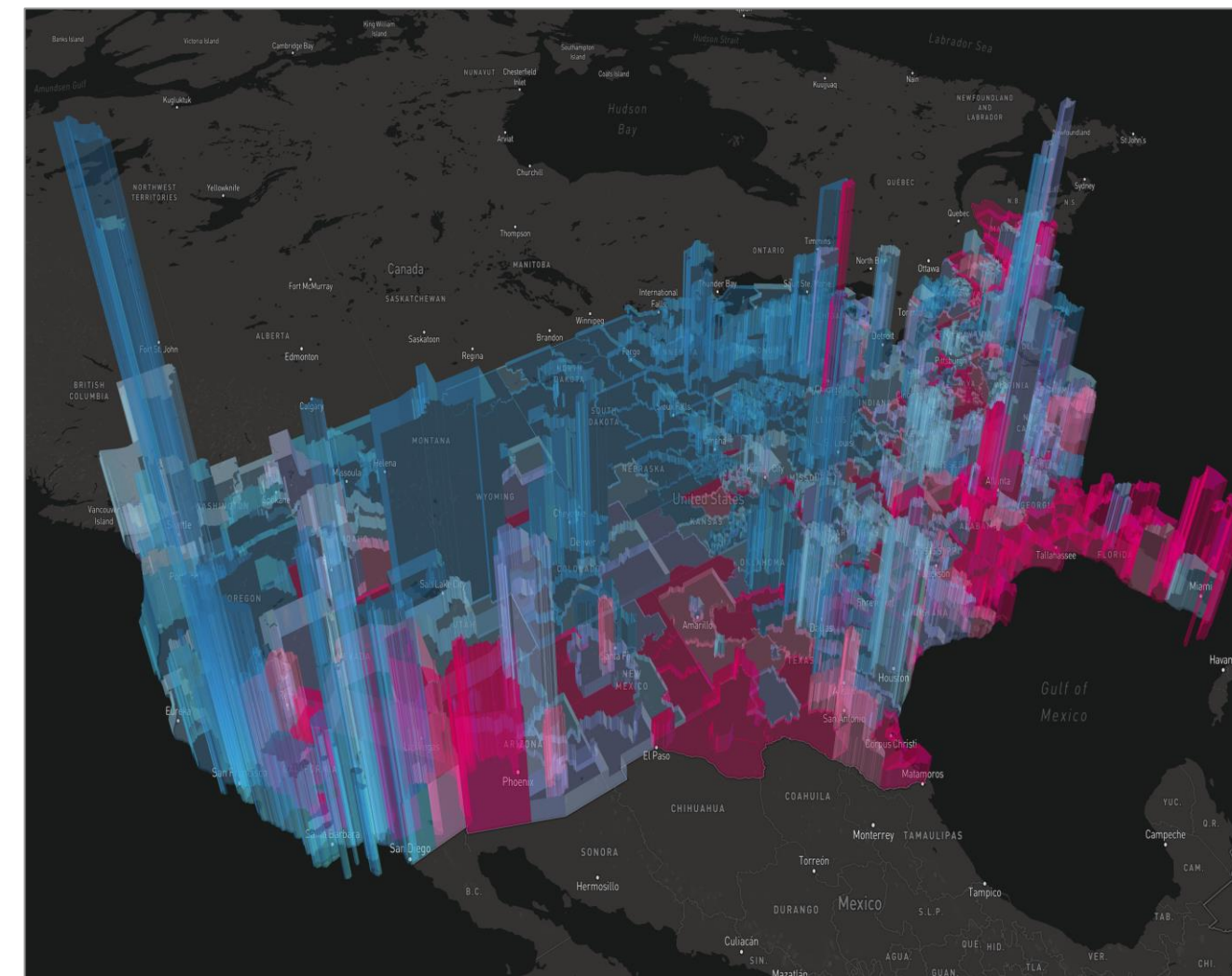


# BlueField Powers NVIDIA-Accelerated Computing Systems

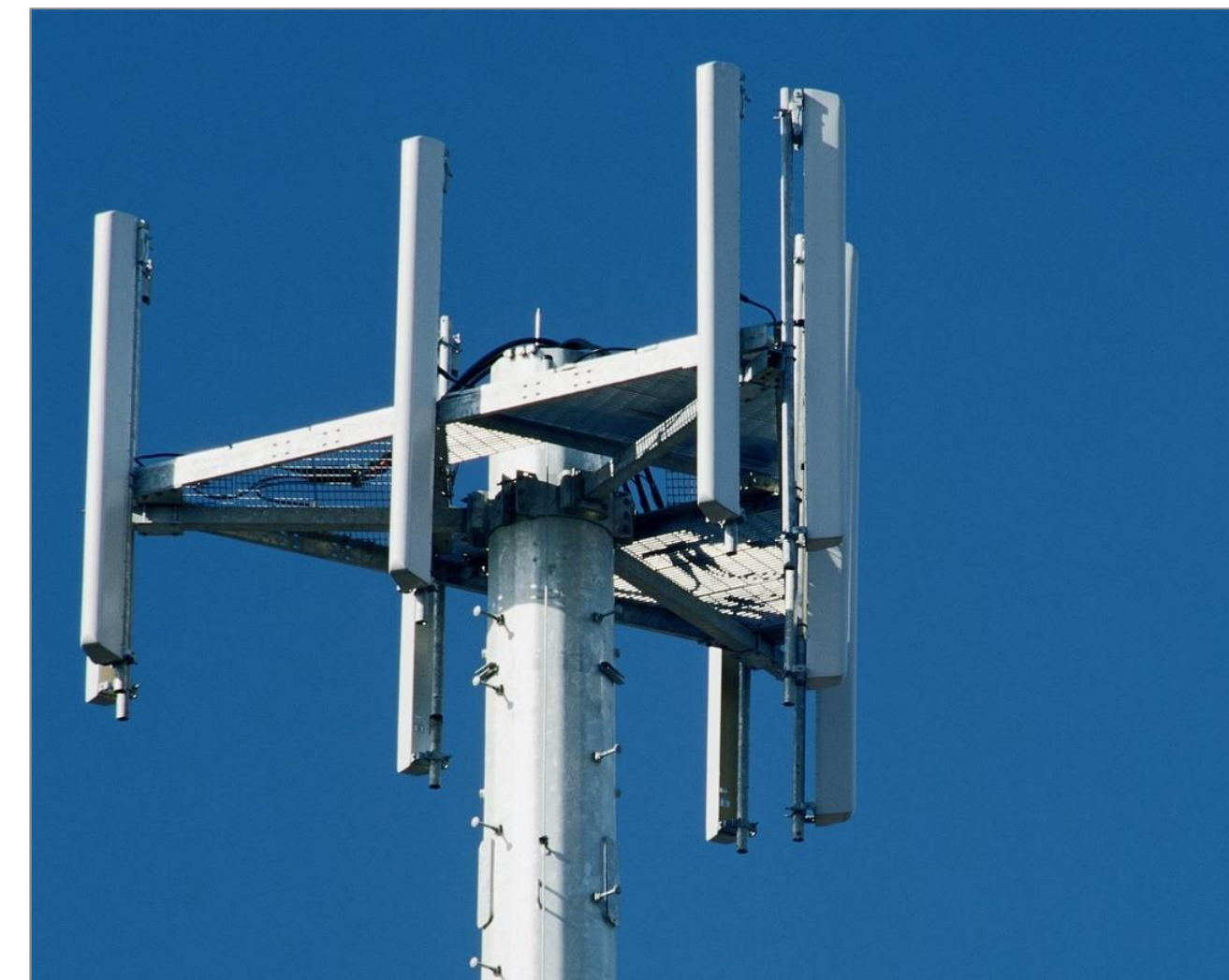
Full-Stack, Data Center-Scale, Multi-Domain Acceleration



Generative AI



Scientific Computing



5G Networks



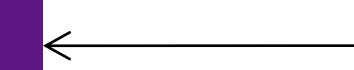
Distributed Database



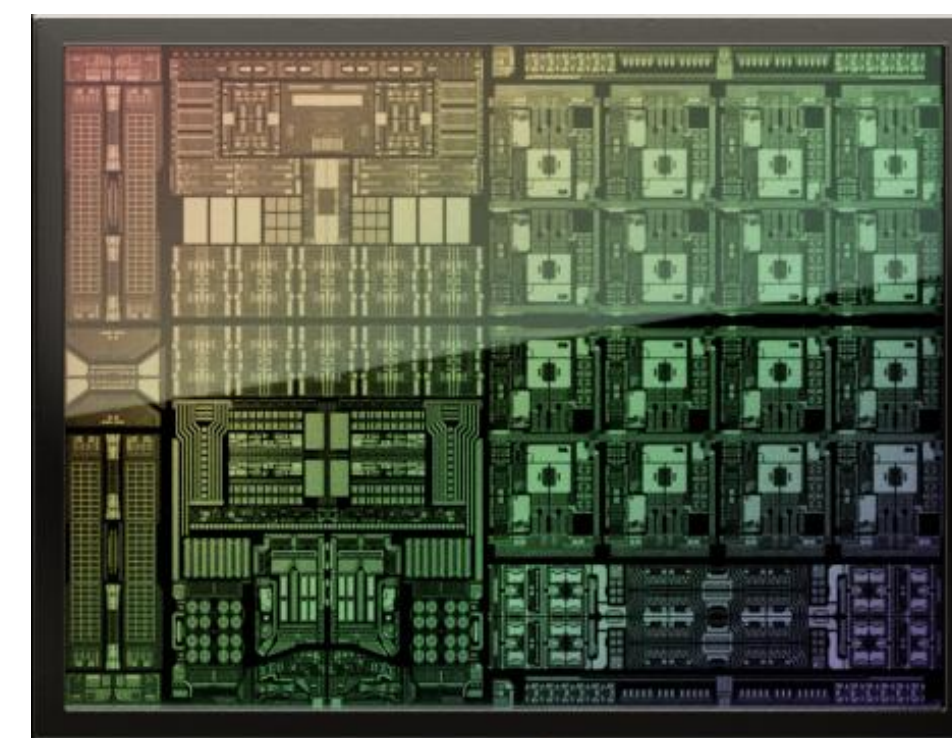
Internet Services



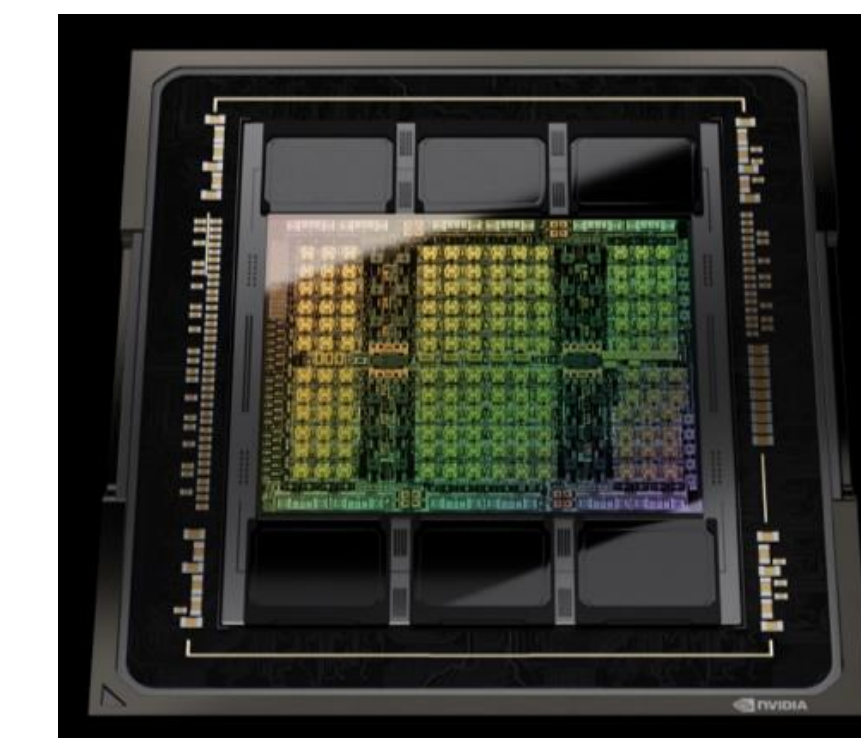
DOCA



CUDA



BlueField-3 DPU



H100 GPU



# Prominent Use-Cases

# Securely Deploy and Operate HGX AI Clouds

Powered by NVIDIA BlueField



## Cloud Network Acceleration

Software-driven VPC networking services at peak performance



## Elastic GPU Computing

Automated provisioning, fungible GPU compute, and limitless scaling



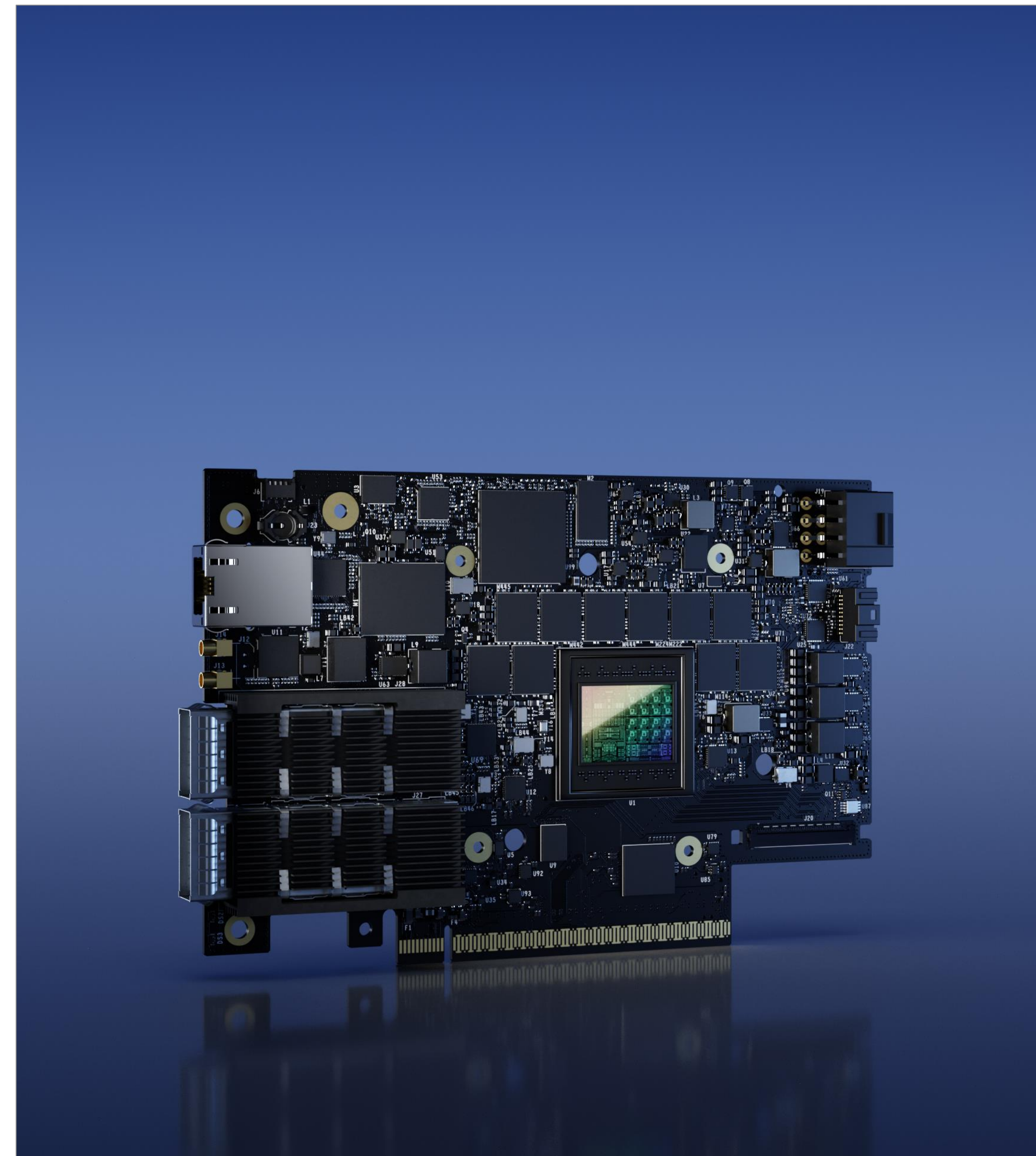
## Secure Infrastructure

Zero-trust, distributed, fine-grained security from the ground up

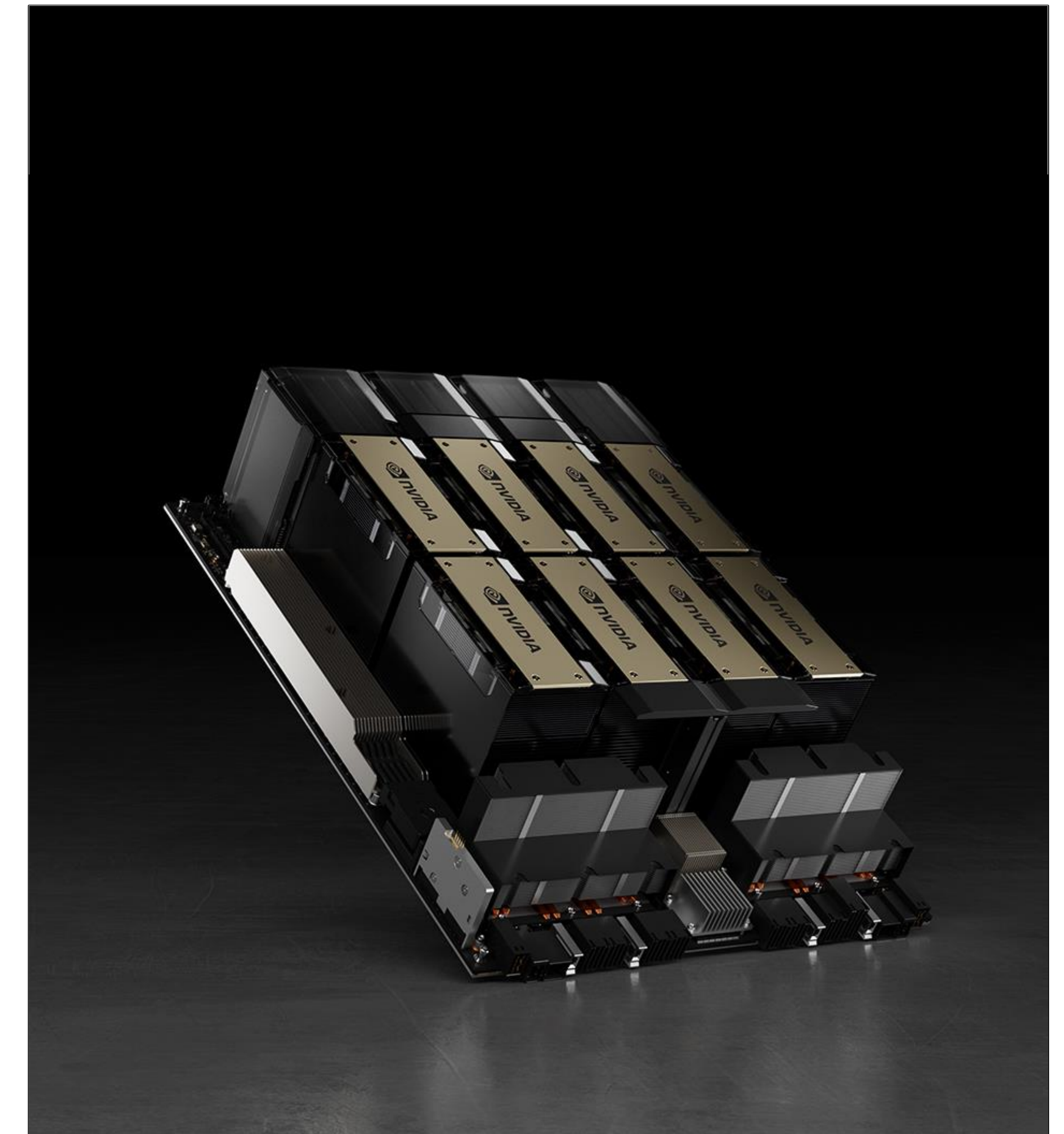


## Robust Data Platform

Blazing fast, scalable and robust data storage services for AI workloads



**NVIDIA BlueField-3 DPU**  
400Gb/s Infrastructure compute platform



**NVIDIA HGX H100 AI Supercomputer**  
The world's most advanced enterprise AI infrastructure

# Securely Deploy and Operate HGX AI Clouds

Highest AI Performance and Cloud Manageability with NVIDIA Quantum InfiniBand and BlueField



## Cloud Network Acceleration

Software-driven VPC networking services at peak performance



## Elastic GPU Computing

Automated provisioning, fungible GPU compute, and limitless scaling



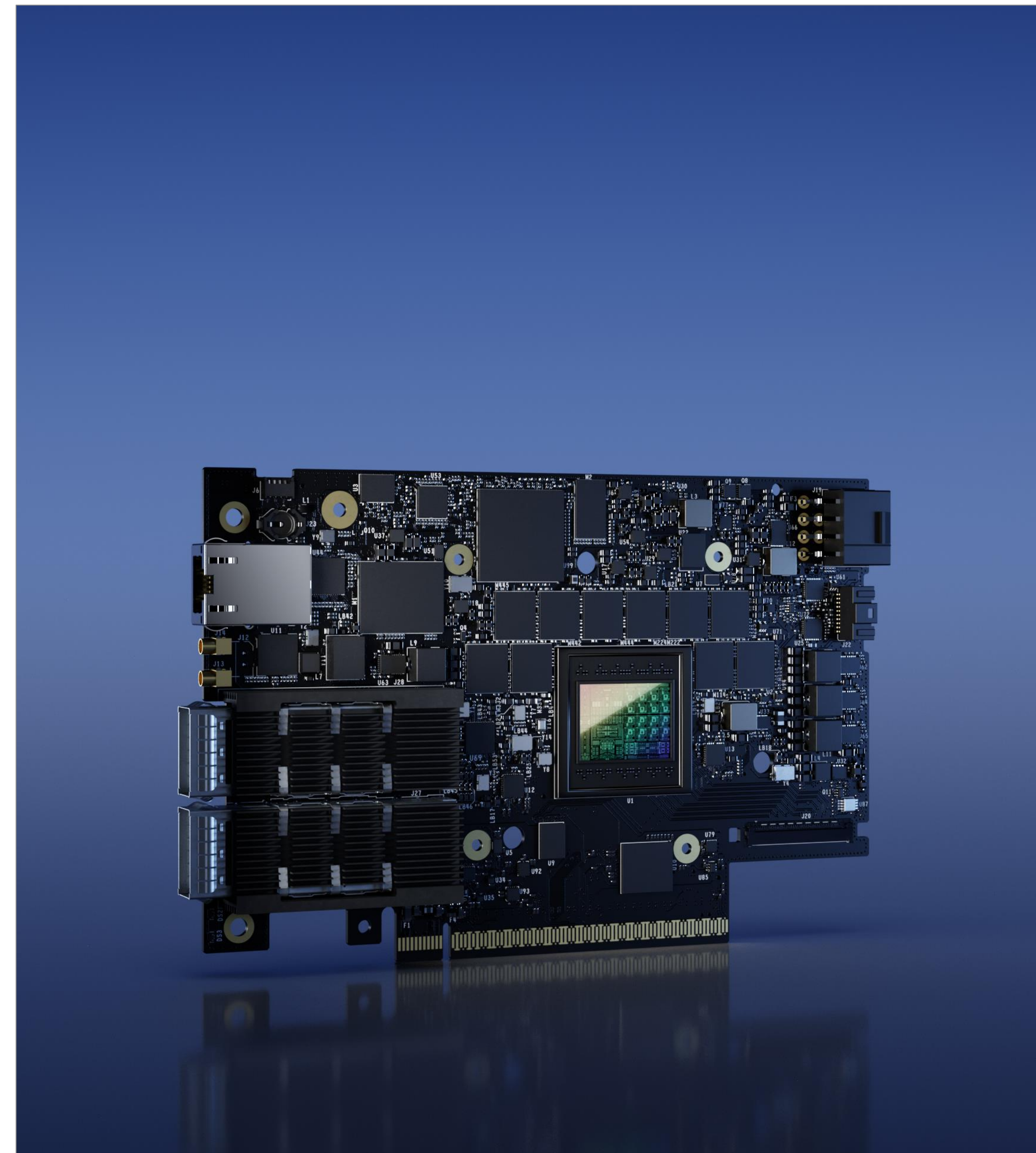
## Secure Infrastructure

Zero-trust, distributed, fine-grained security from the ground up

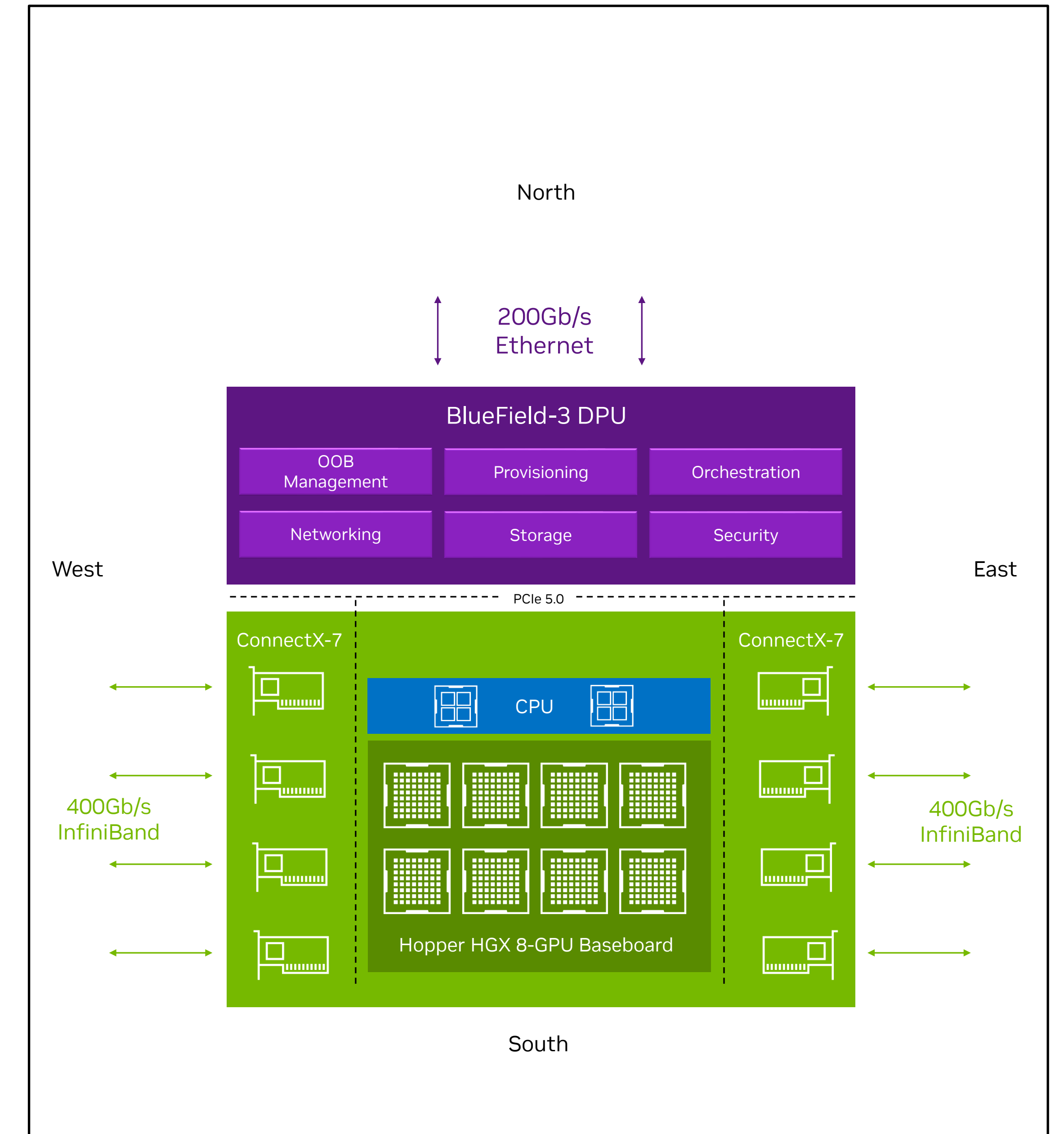


## Robust Data Platform

Blazing fast, scalable and robust data storage services for AI workloads



**NVIDIA BlueField-3 DPU**  
400Gb/s Infrastructure compute platform



**NVIDIA HGX H100 AI Supercomputer**  
AI services: 400Gb/s InfiniBand (East-West)  
Tenant networking: 200Gb/s Ethernet (North-South)

# HGX H100 and Spectrum-X Accelerate AI Clouds

Highest Ethernet AI Performance and Cloud Manageability with NVIDIA Spectrum-X



## Ethernet AI Performance Leadership

BlueField accelerates AI with adaptive routing, out-of-order packet handling, and congestion control



## Elastic GPU Computing

BlueField enables multi-tenant cloud computing and tenant isolation at massive scale



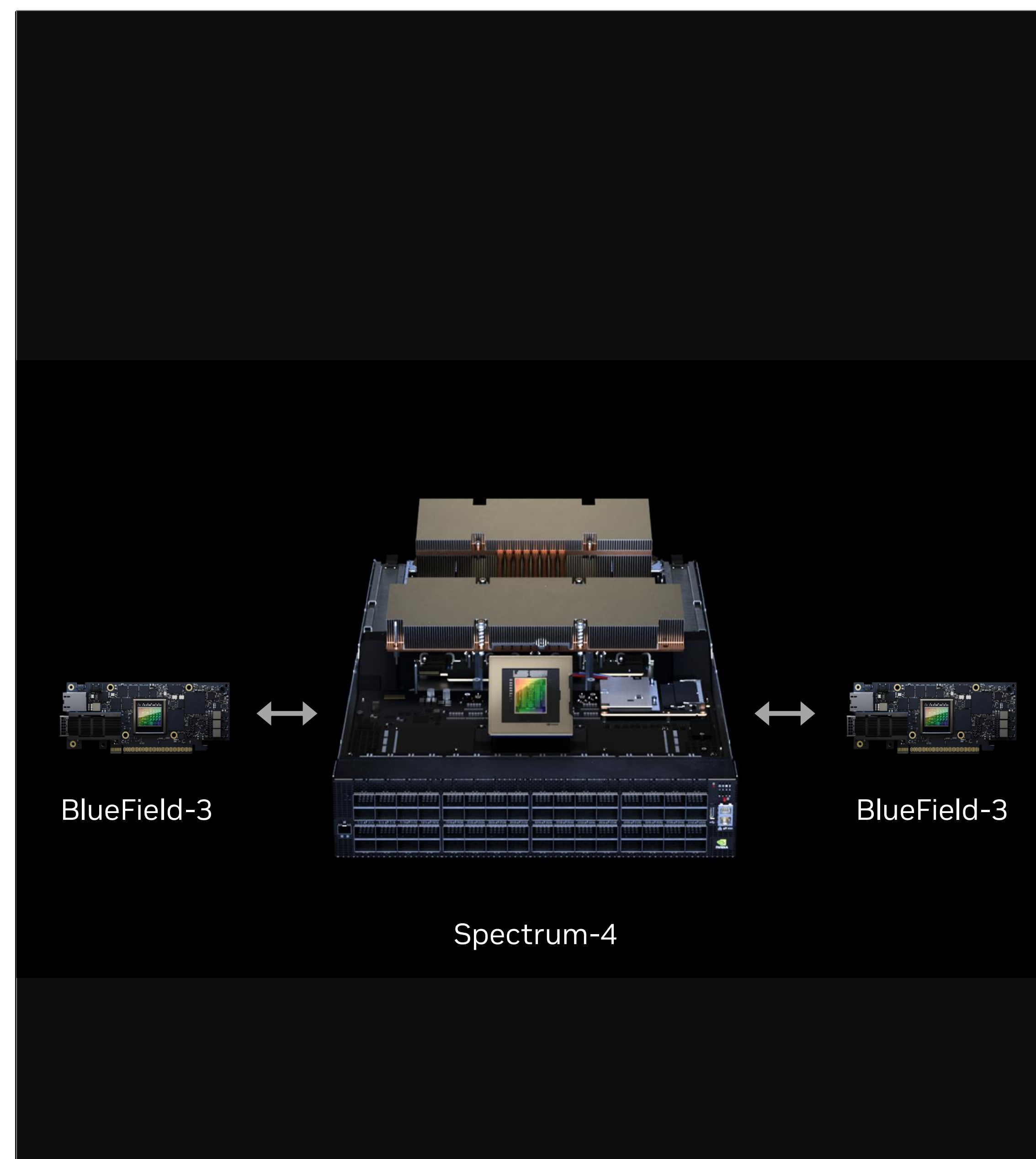
## Secure Infrastructure

BlueField creates a fully programmable compute infrastructure for AI



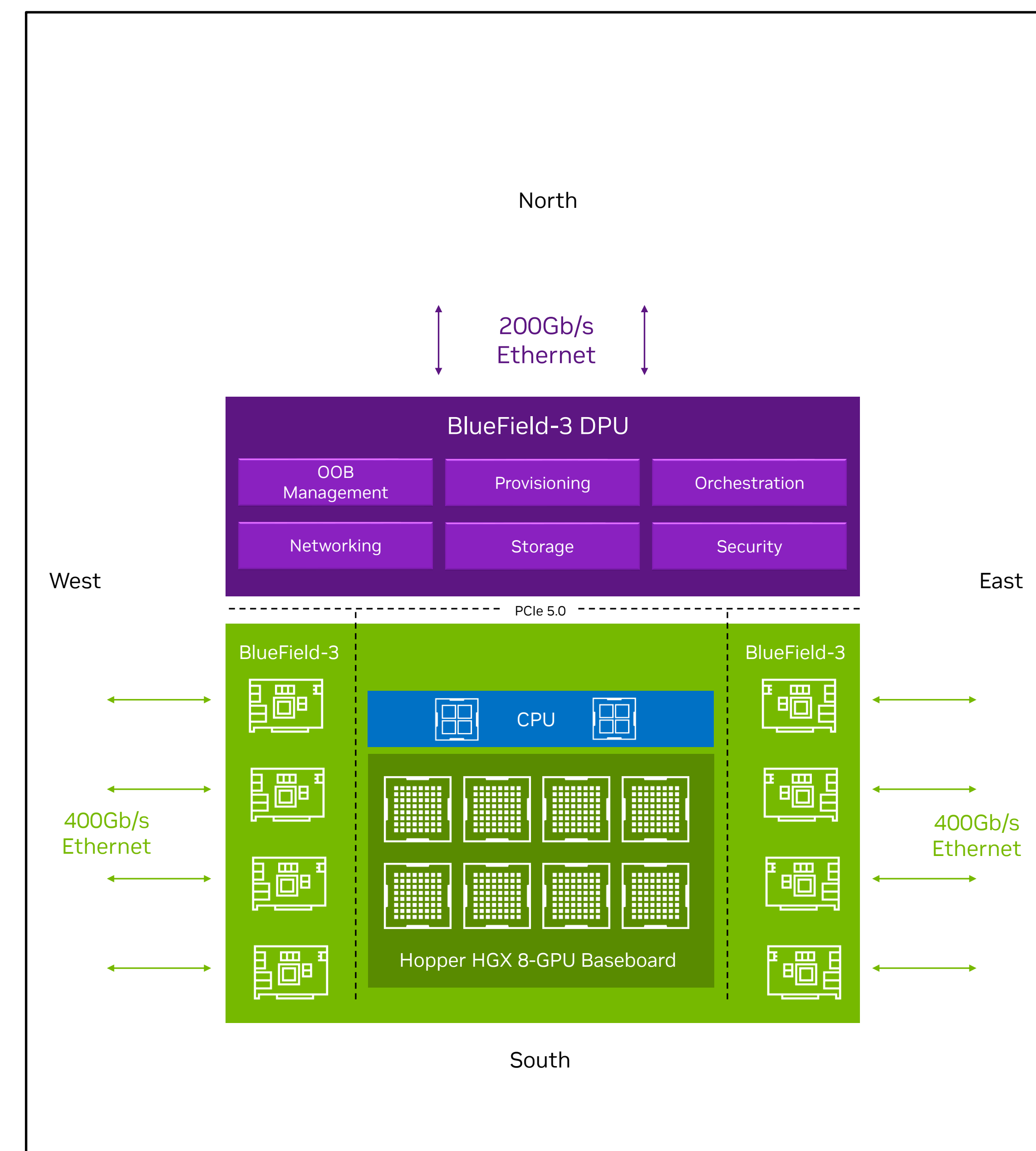
## Robust Data Platform

BlueField enables Ethernet AI cloud builders to make the most of NVIDIA GPUs, achieving new levels of efficiency and productivity



## NVIDIA Spectrum-X Platform

Purpose-built Ethernet Fabric for AI Clouds



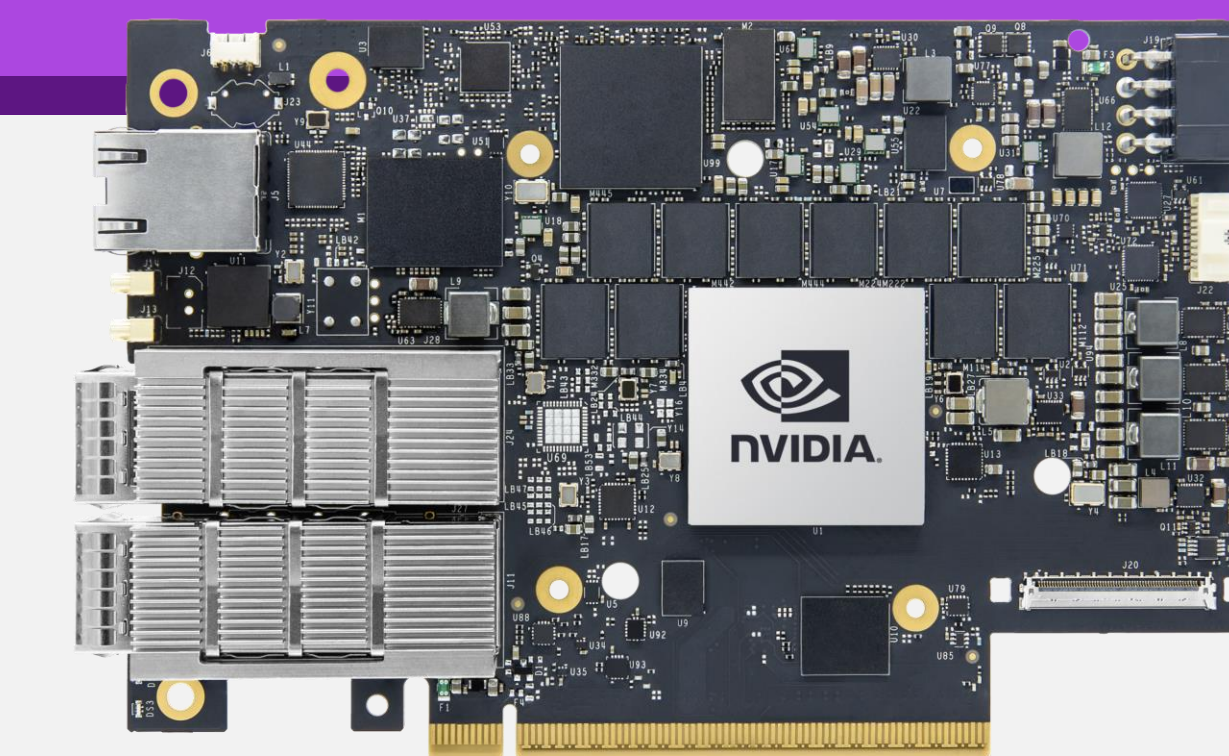
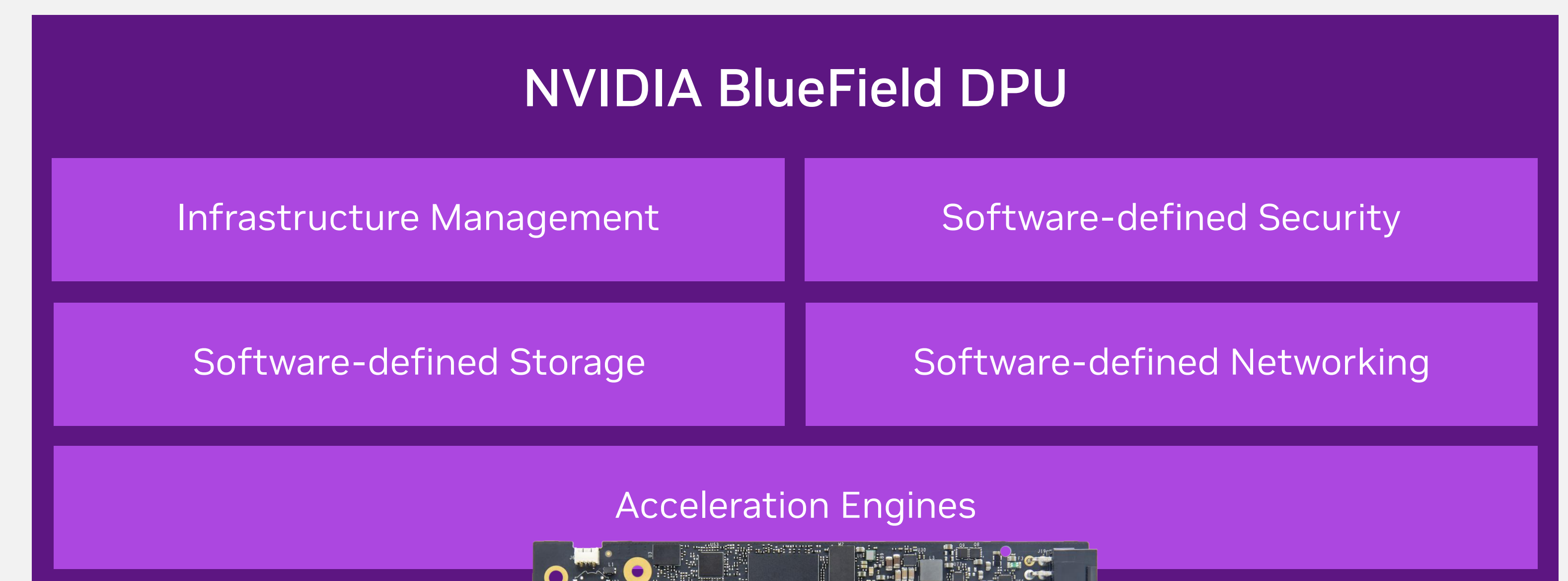
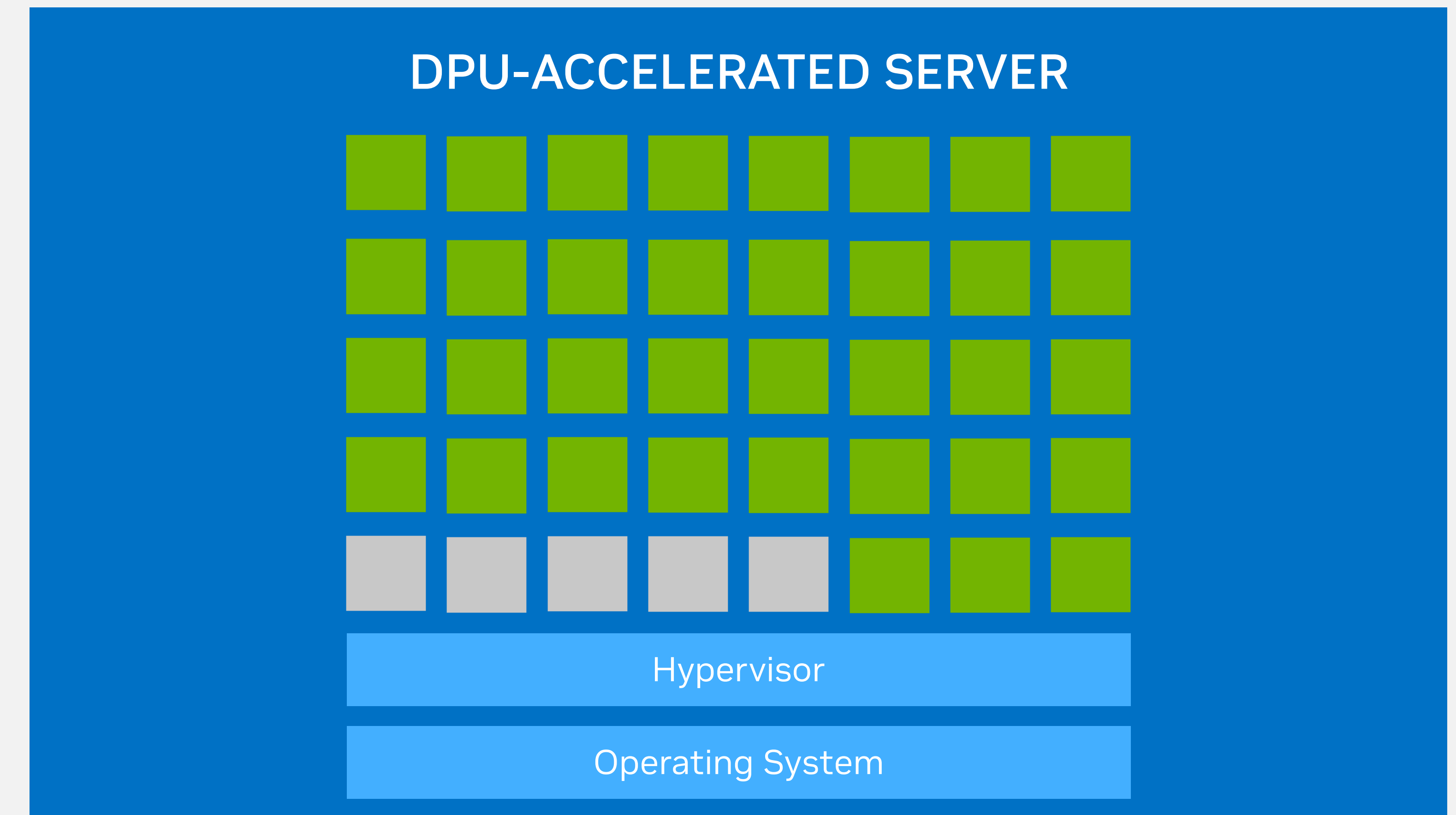
## NVIDIA HGX H100 AI Supercomputer

AI services: 400Gb/s Ethernet (East-West)  
Tenant networking: 200Gb/s Ethernet (North-South)



# Accelerate Cloud Computing

BlueField Powers Clouds to Host More Virtual Instances

- Up to 8X the number of virtual instance per node
- Generate revenues for the additional capacity
- Higher ROI and lower TCO for cloud data centers



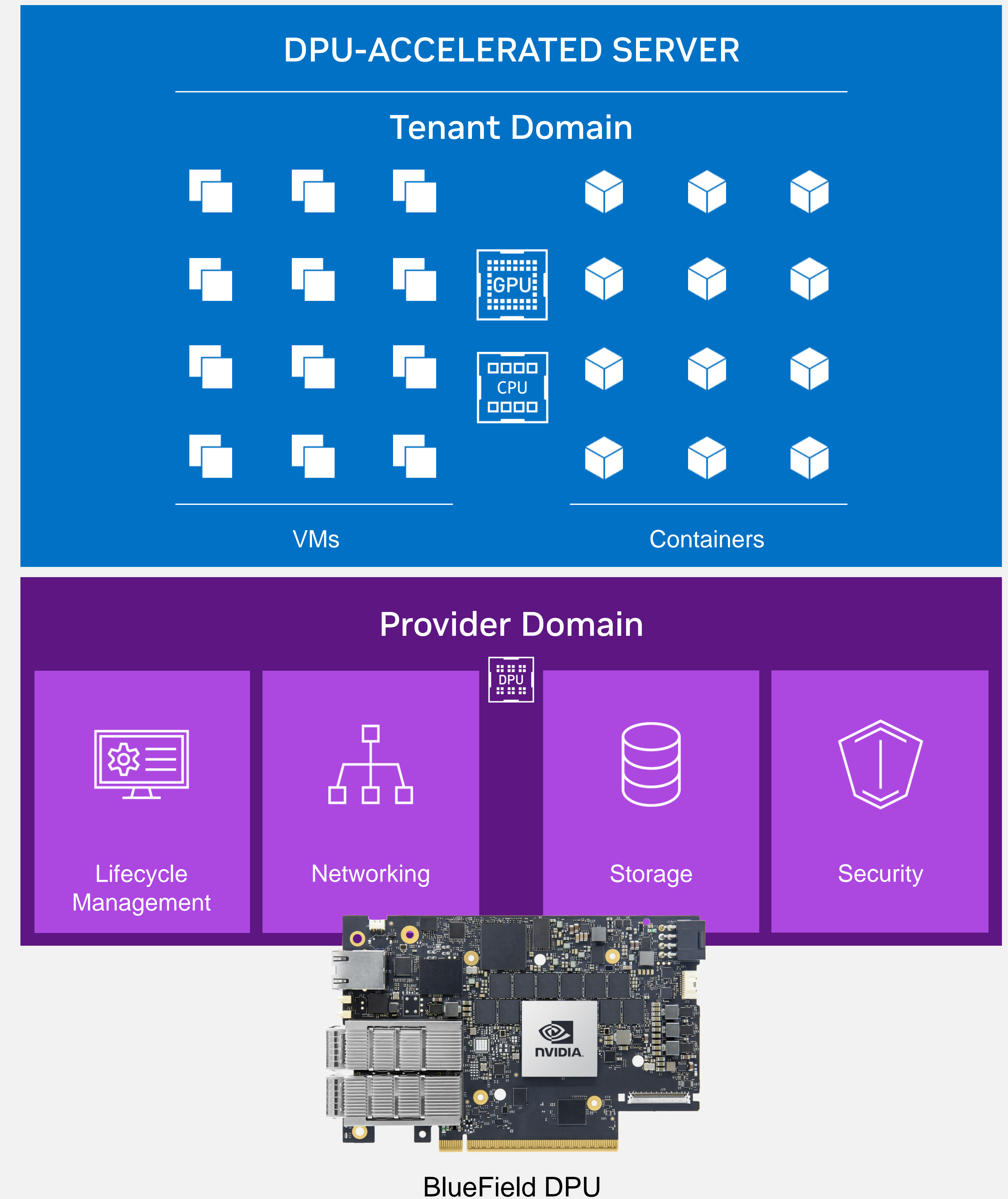
BlueField DPU

-  Virtual instance on BlueField-3
-  Virtual instance on BlueField-2

# Secure Cloud Computing

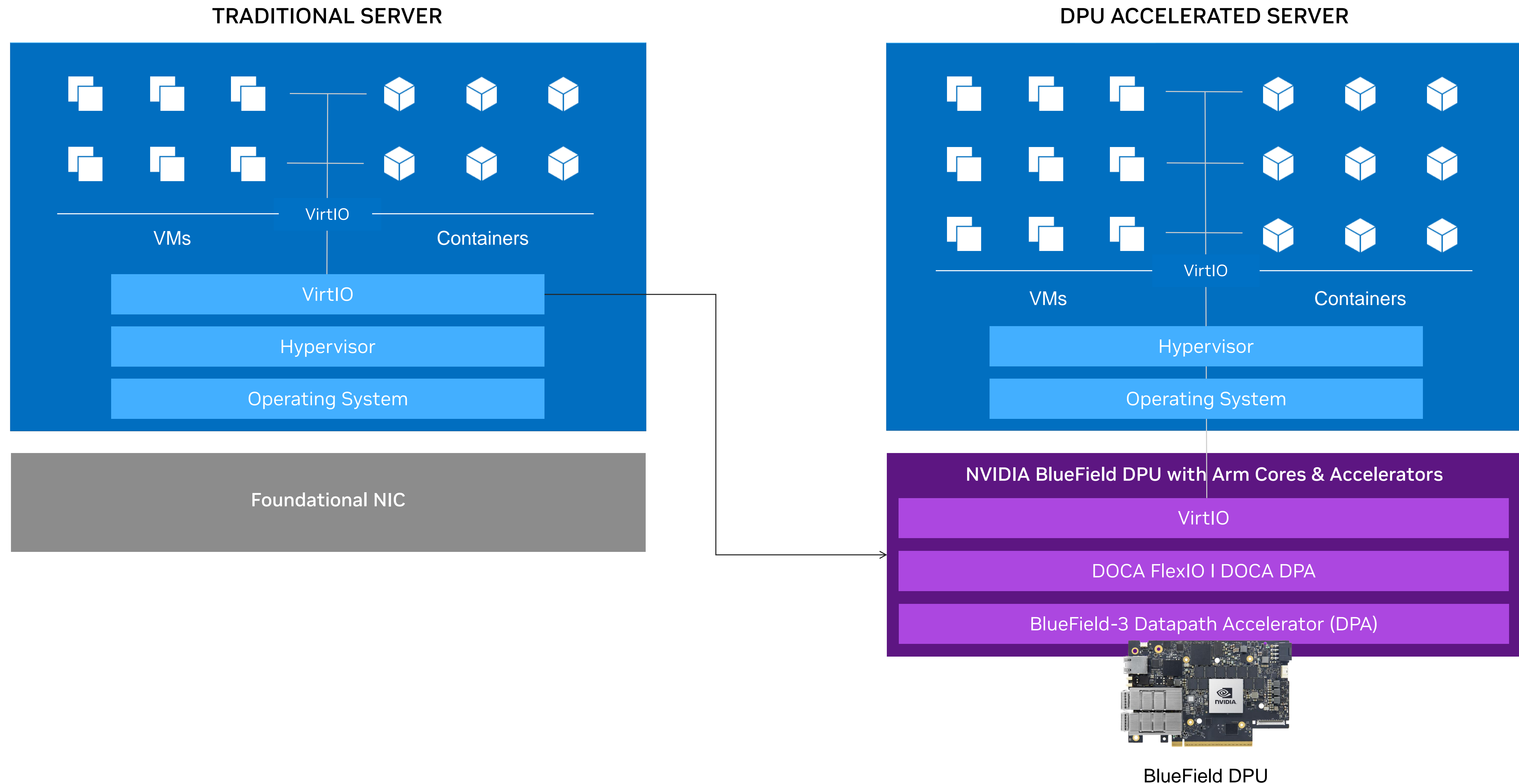
## BlueField-3 Provides a Secure Foundation for Cloud Compute Nodes

- Secure multi-tenant cloud
- Isolated data center control-plane
  - Tenant workloads run on the host
  - Infrastructure workloads run on BlueField DPUs
- Provisioning and lifecycle management through BlueField DPUs



# Accelerate Cloud Networking

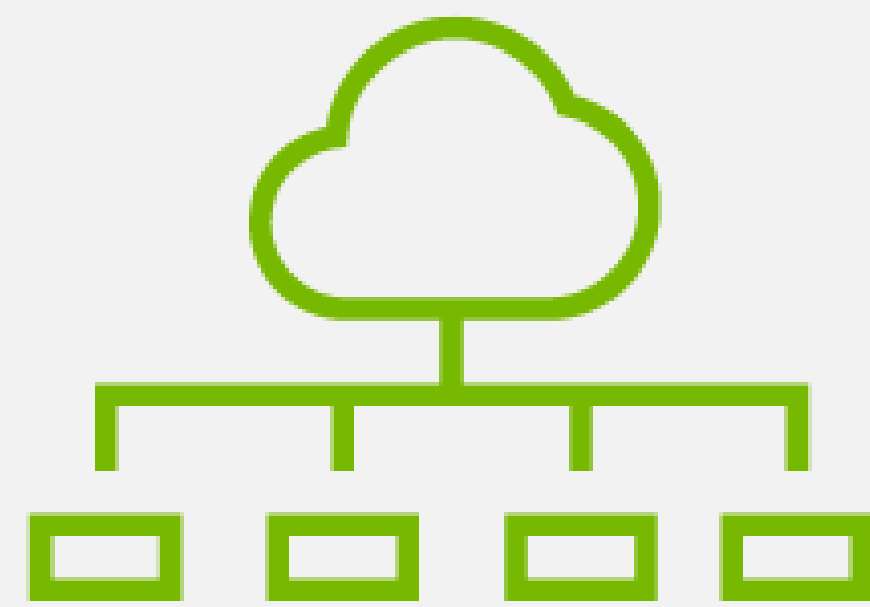
Powered by DOCA, BlueField-3 accelerate VirtIO network connectivity





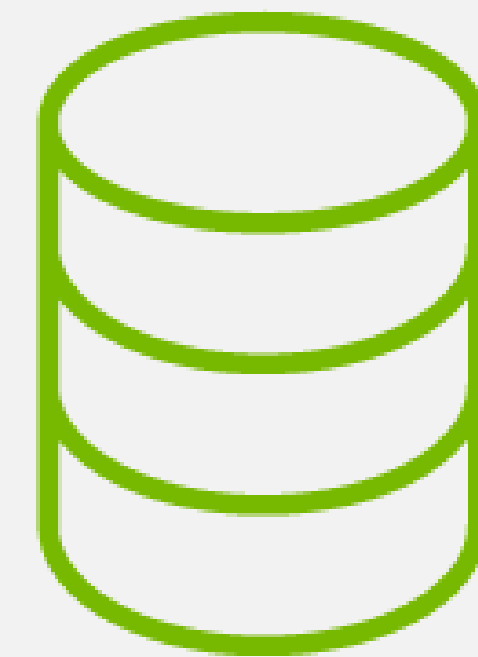
# NVIDIA BlueField Transforms Bare-Metal Clouds

From Hardware-Defined to Modern, Software-Defined, Hardware-Accelerated Cloud Infrastructure



## Software-Defined Networking

- Advanced SDN capabilities
- Full control and orchestration
- No driver installation



## Elastic Composable Storage

- Dynamic allocation of cloud storage
- High-performance storage access
- OS agnostic

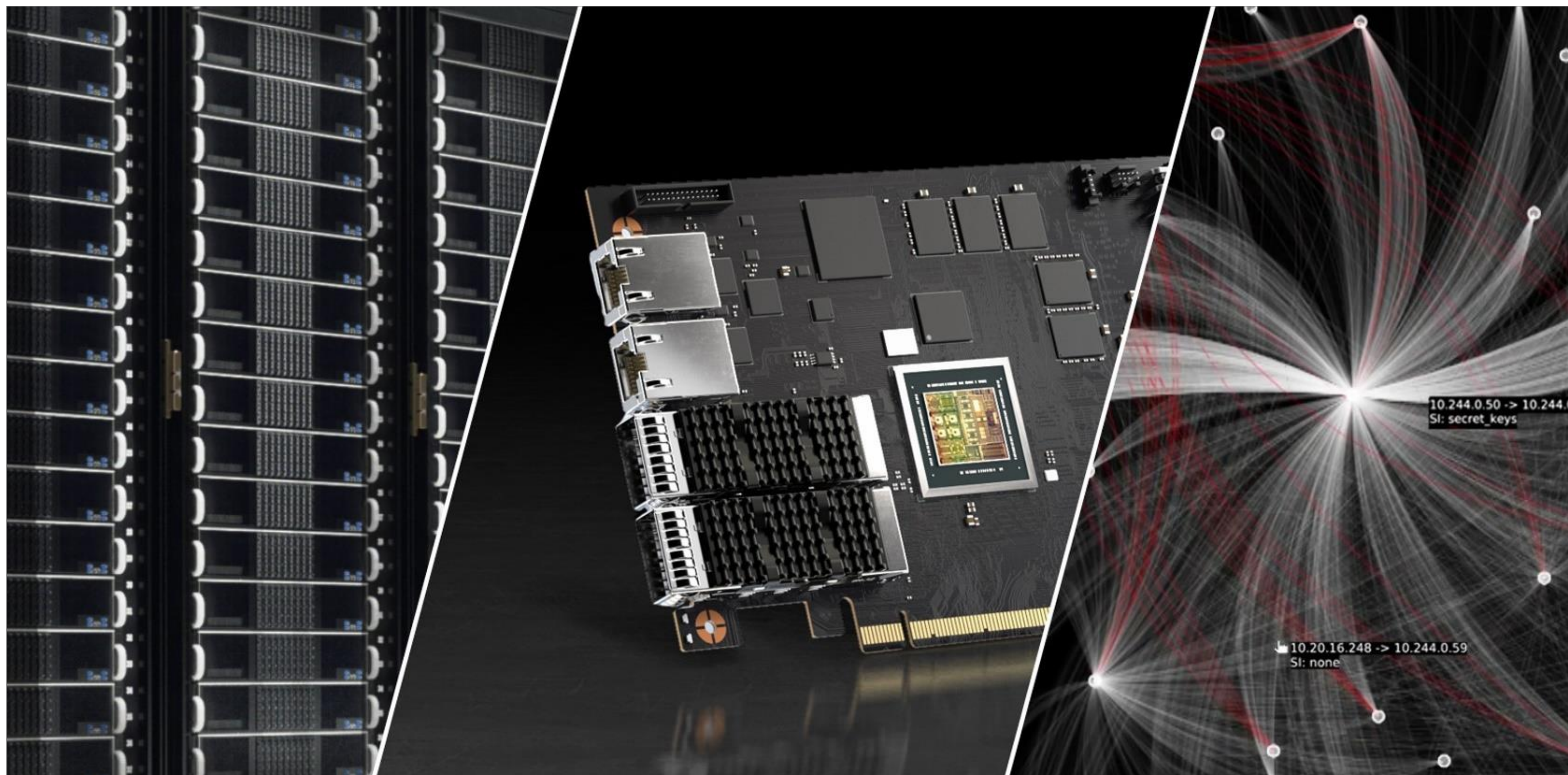


## Zero-Trust Security

- End-to-end encryption
- Isolated control-plane
- No agent is required

# Delivering AI-Ready Infrastructure for Enterprises

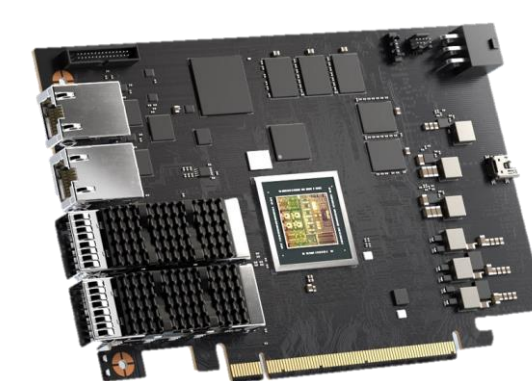
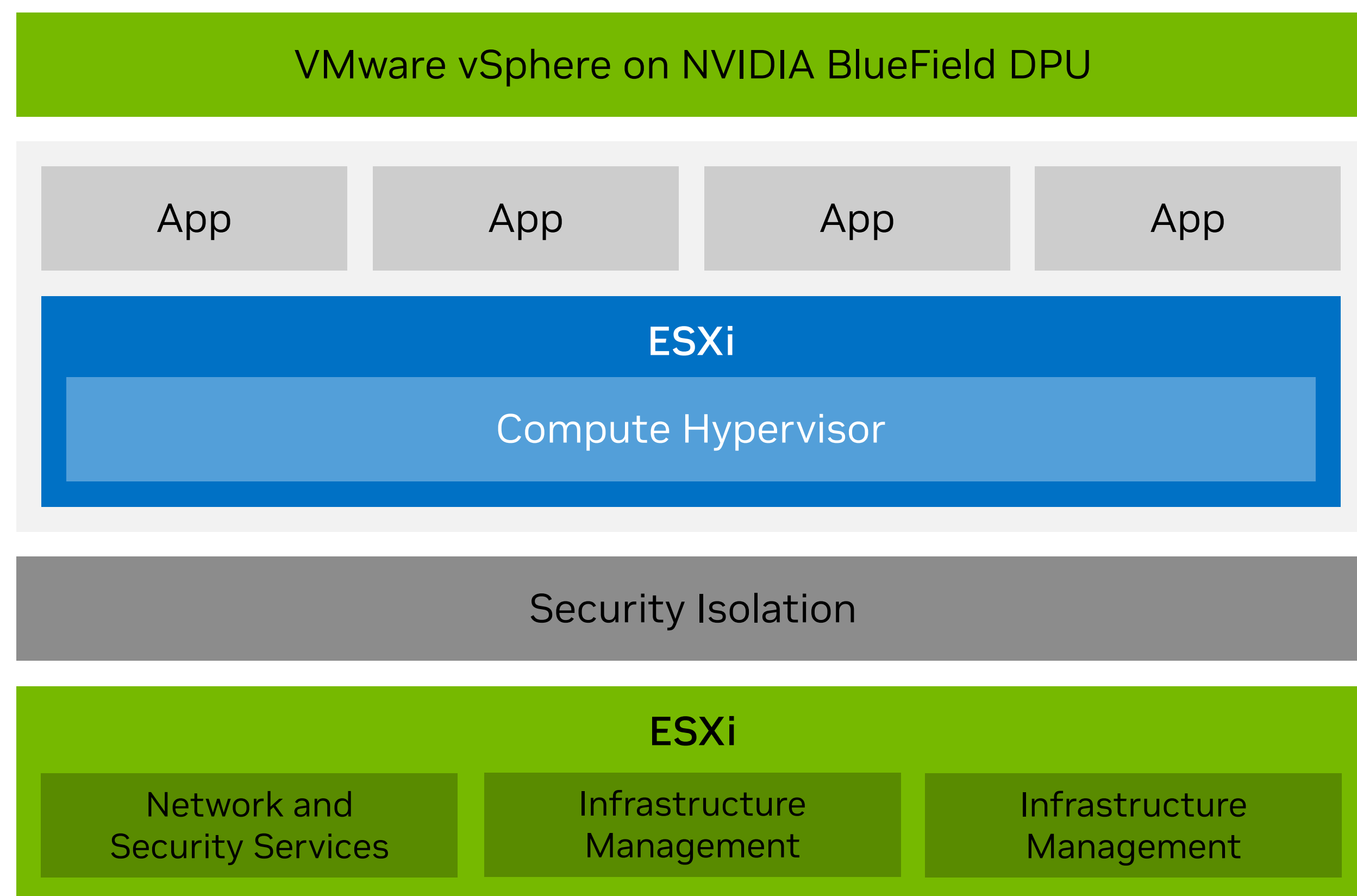
VMware vSphere Distributed Services Engine Powered by NVIDIA BlueField DPUs



Accelerate VMware Servers for Next-Generation Applications

# Delivering AI-Ready Infrastructure for Enterprises

VMware vSphere Distributed Services Engine Powered by NVIDIA BlueField DPUs

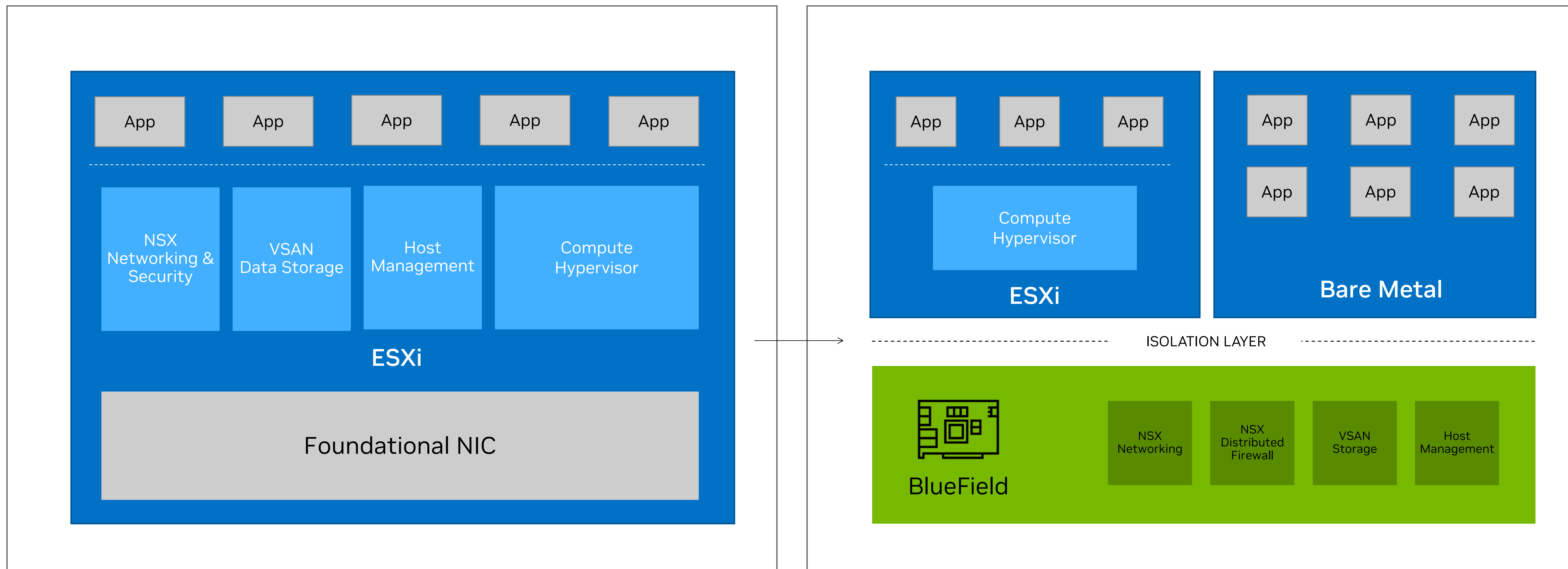


NVIDIA BlueField

- Single, secure, operating model across workload types and deployments
- Isolation of workload domain from the infrastructure domain
- Offload and Accelerate infrastructure service functions to DPUs

# Delivering AI-Ready Infrastructure for Enterprises

VMware vSphere Distributed Services Engine Powered by NVIDIA BlueField DPUs

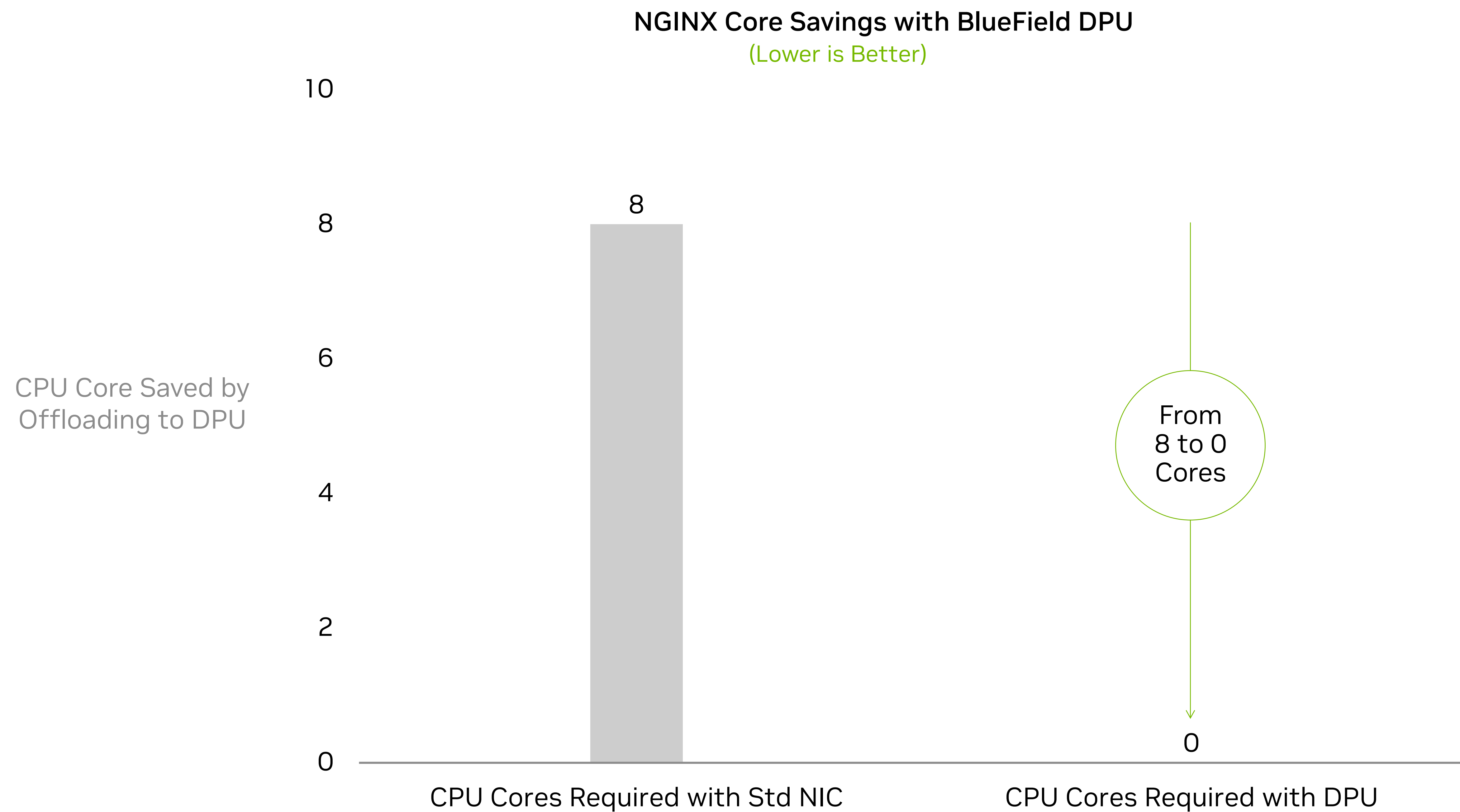


VMware Cloud Foundation

VMware Cloud Foundation w/ Distributed Services Engine

# Delivering AI-Ready Infrastructure for Enterprises

VMware vSphere Distributed Services Engine Powered by NVIDIA BlueField DPUs

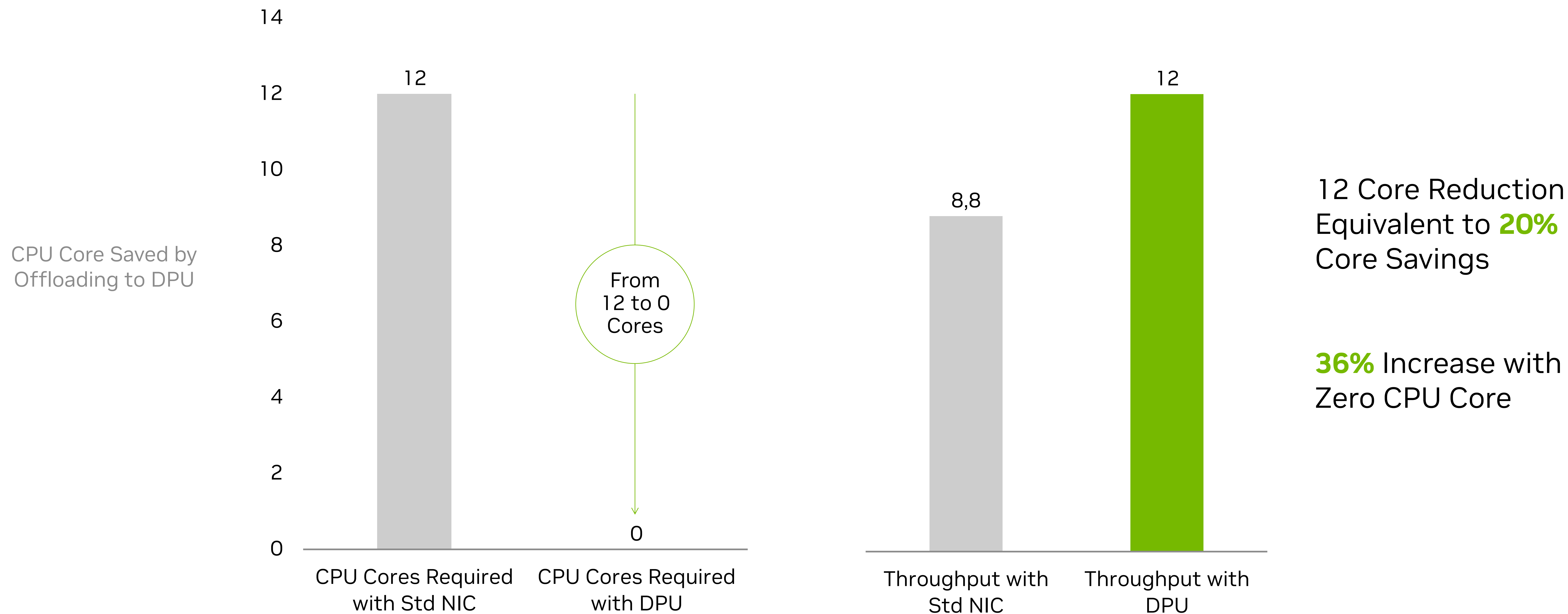


NGINX Webserver w/Stateful Distributed Firewall (1K Flow) 100G BF ~15Gbps

# Delivering AI-Ready Infrastructure for Enterprises

VMware vSphere Distributed Services Engine Powered by NVIDIA BlueField DPUs

Redis Core Savings and Throughput Gain with BlueField DPU



REDIS - In-memory key-value store, 36 Redis streams, 25G BF (Millions Transactions/sec)  
64 Core System

# Delivering AI-Ready Infrastructure for Enterprises

VMware vSphere Distributed Services Engine Powered by NVIDIA BlueField DPUs

22%

Improved Server Efficiency

5X

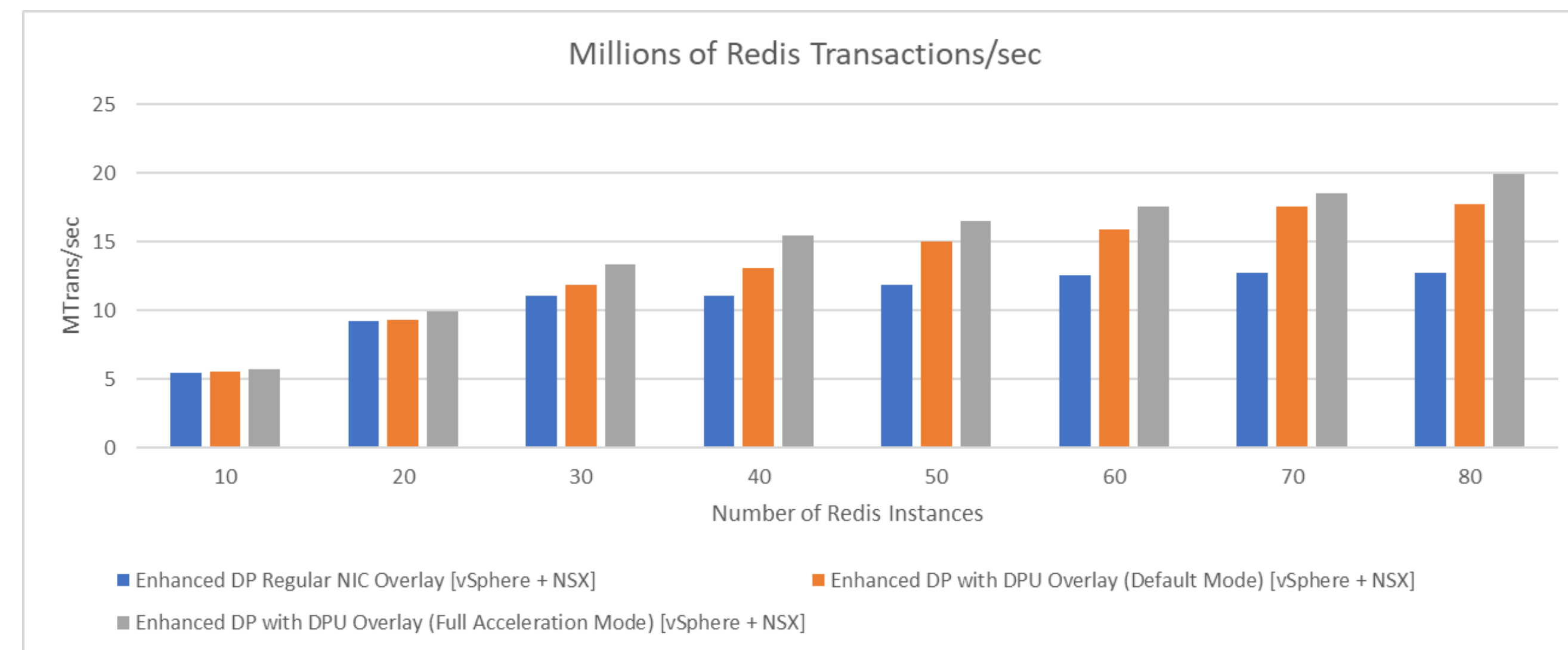
3-Year ROI over Std. NIC

- 780 servers w/BlueField DPUs equivalent to 1,000 servers with standard NIC
- TCO Savings of \$8,200/server
- \$1.8M in efficiency savings over 3 years

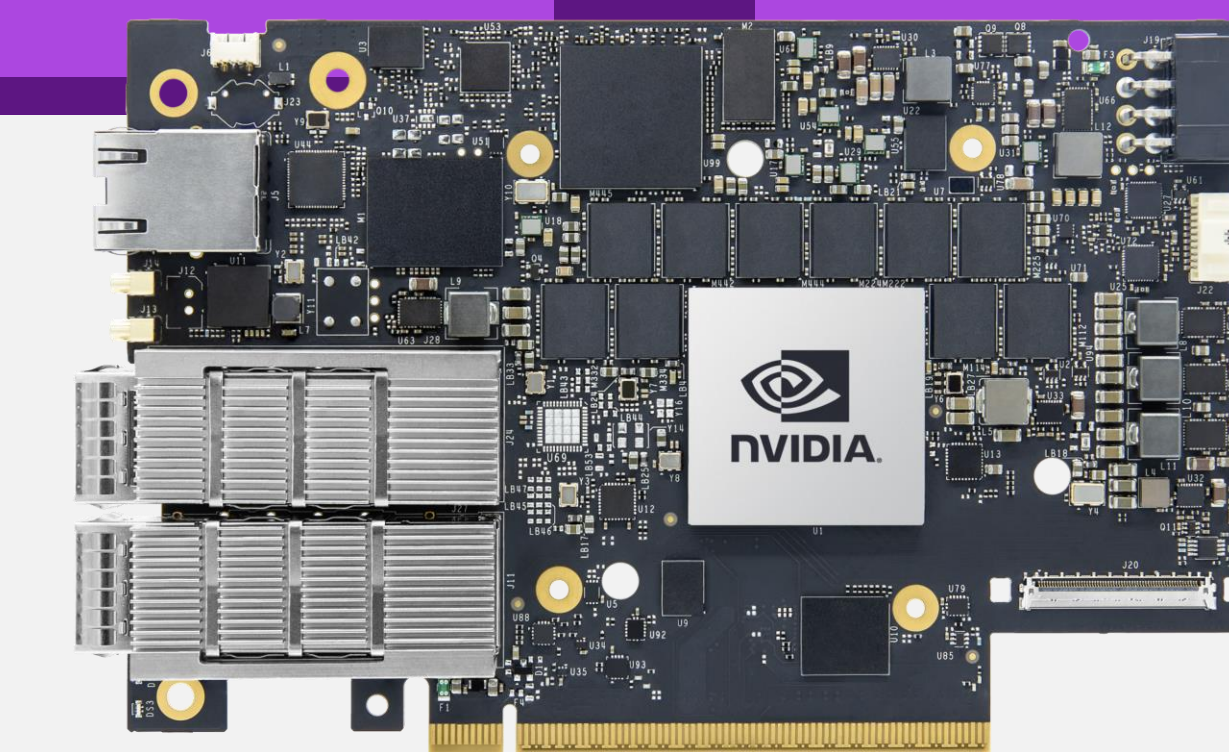
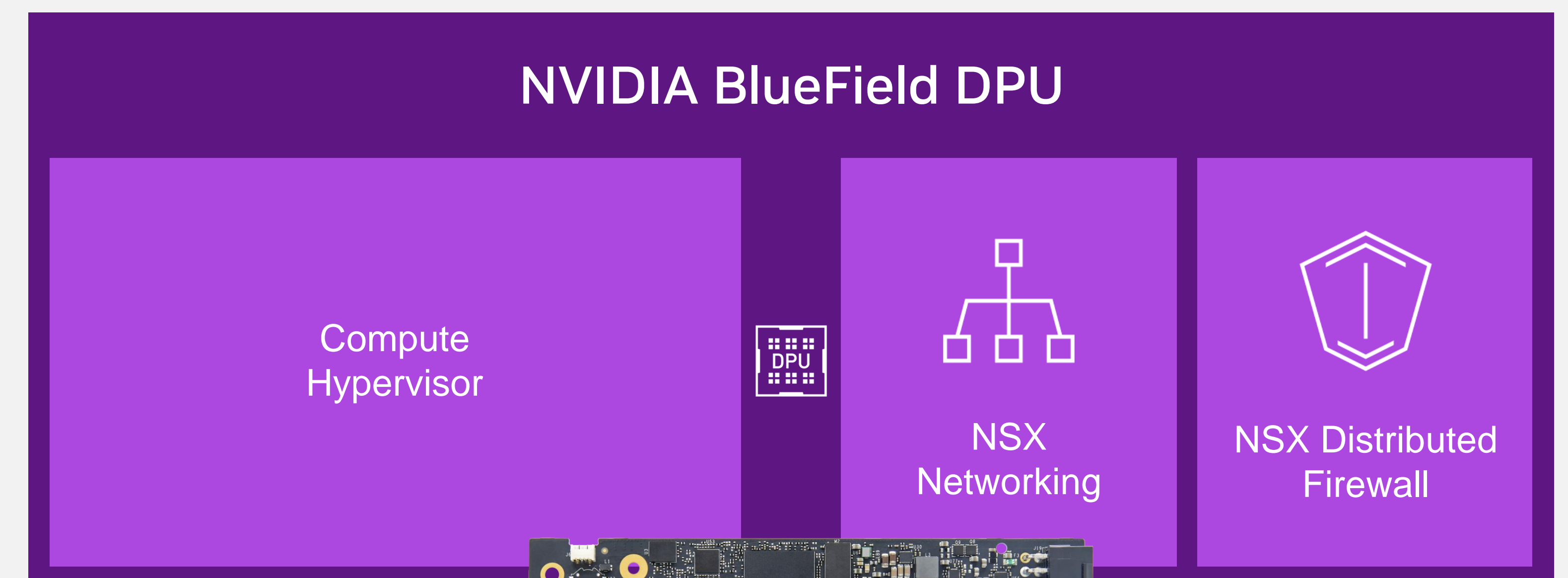
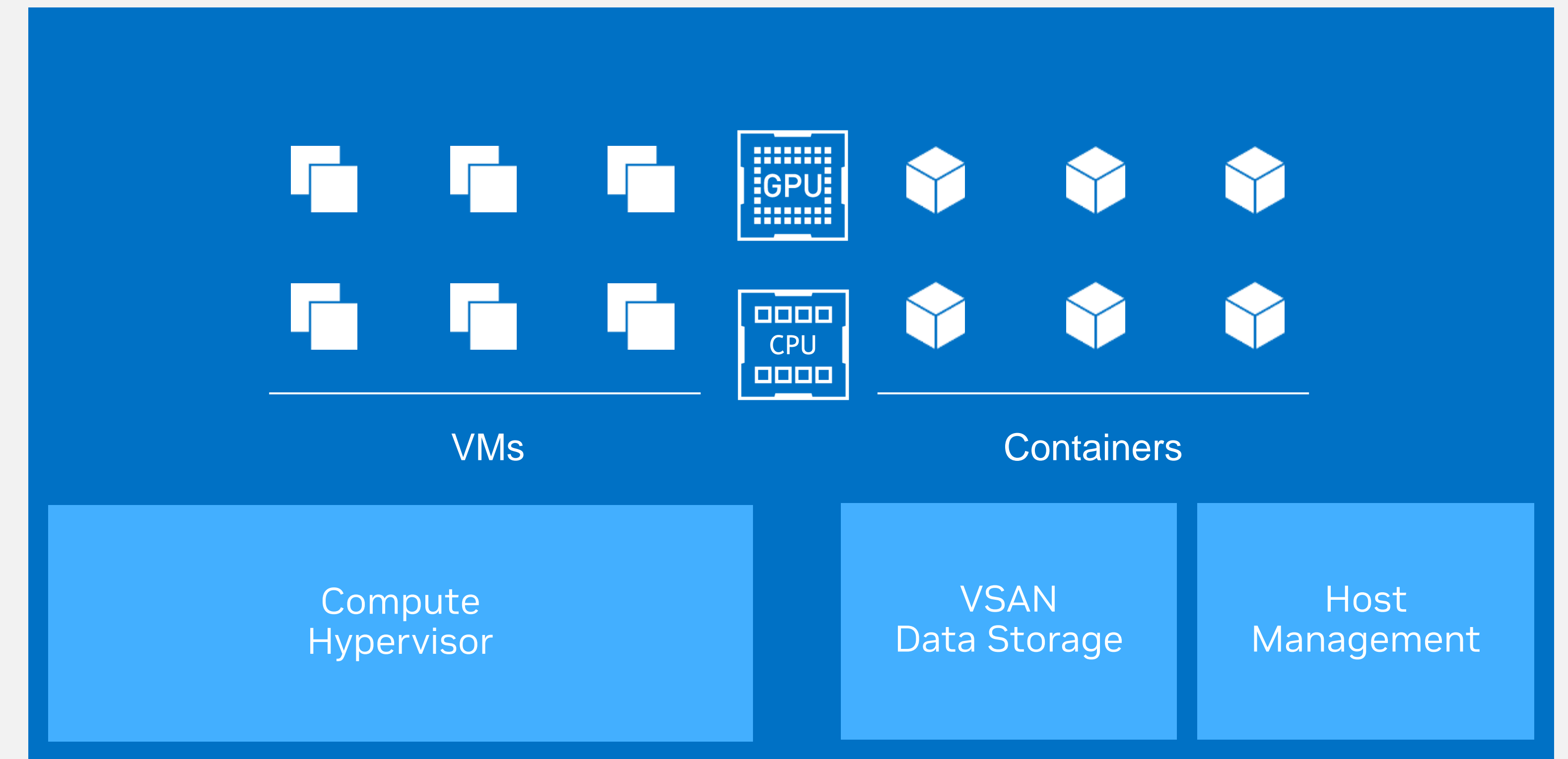
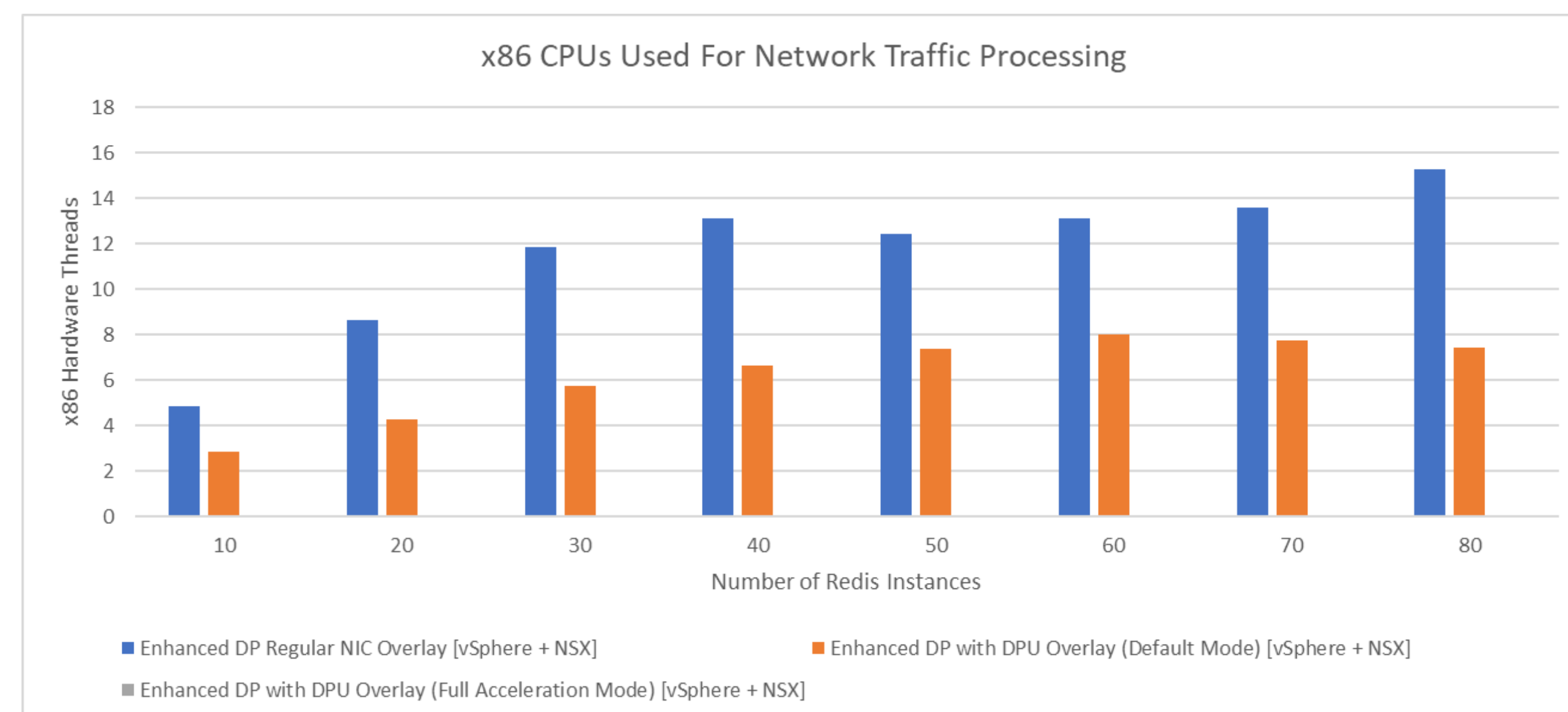
# Accelerate Enterprise Clouds

BlueField Accelerates VMware vSphere Virtualization Stack

## Up to 50% more Redis transactions/sec



## Zero CPU cores for VMware networking



BlueField DPU



# DOCA Zero-Trust Security Framework

Enabling the Secure, Accelerated Cloud

- Authenticates, measures, and secures the data center
- Platform security — hardware root of trust based authentication
- Implement security groups and access controls
- L2-L7 security services and policy enforcement
- Securing data at rest and in motion
- Enhanced visibility with comprehensive monitoring
- Supporting and expanding AI-based Cybersecurity



Software-Defined



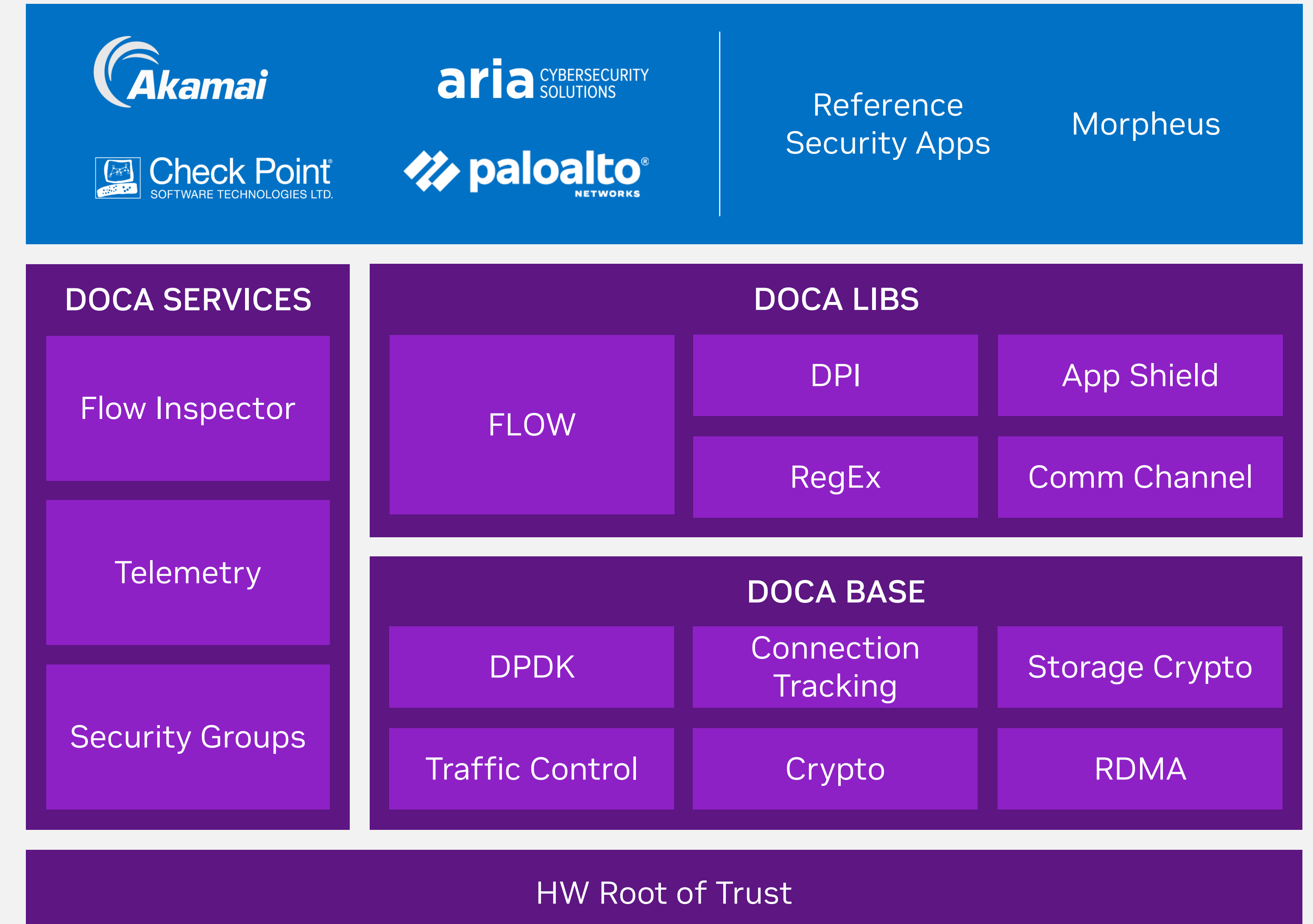
Scalable



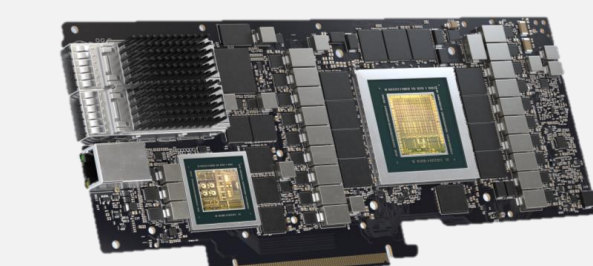
Robust



AI-Based Security



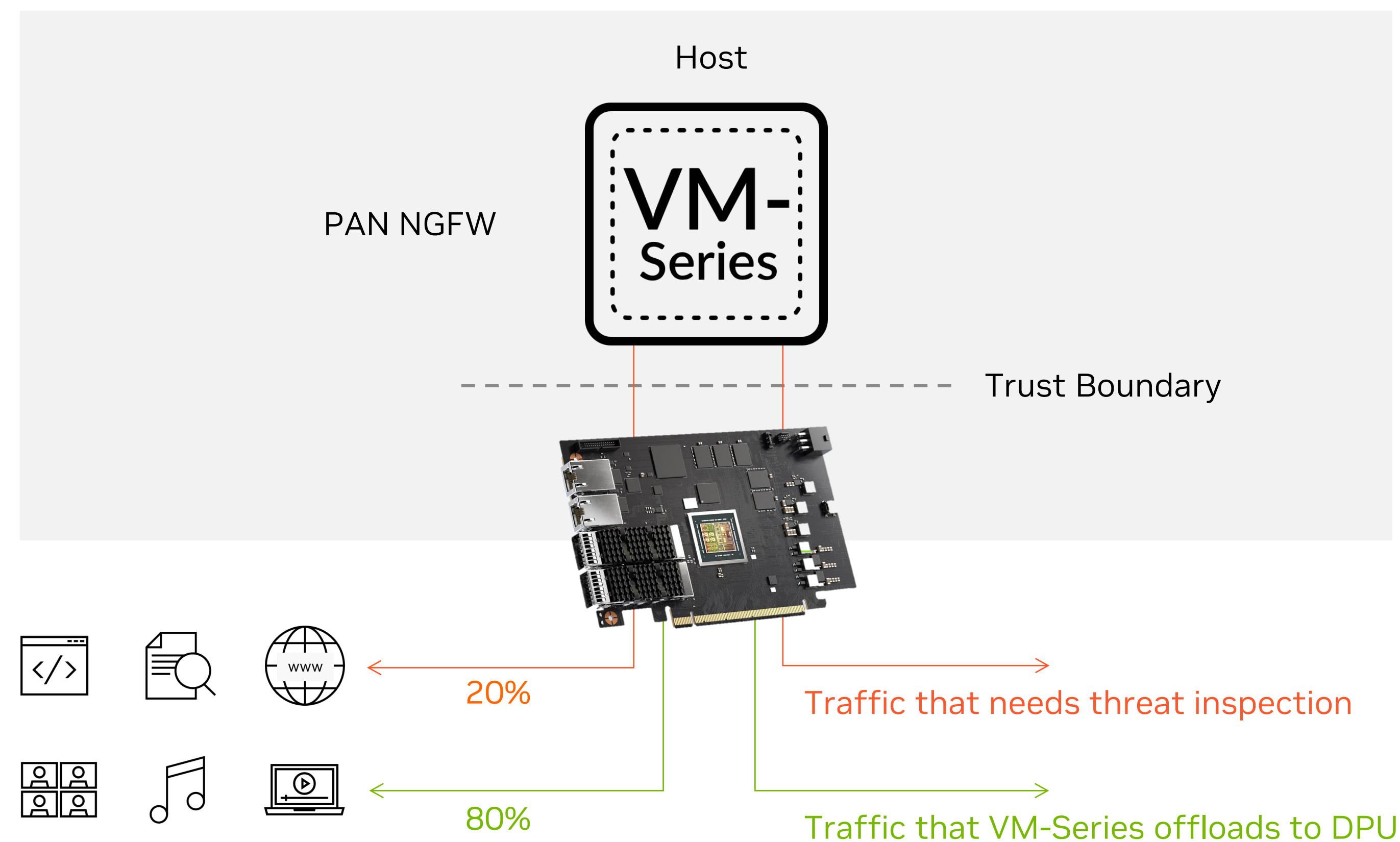
NVIDIA BlueField DPUs



NVIDIA Converged Accelerators

# Boosting Cyber Defenses

NVIDIA BlueField Offloads and Accelerates Palo Alto Networks' Virtual Next-Gen Firewall (NGFW)



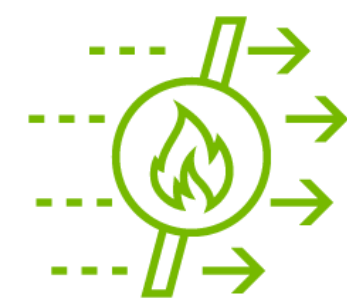
- Intelligent Traffic Offload (ITO) enables 5x acceleration
- Achieves near 100Gb/s network performance
- 80% traffic offloaded – video, audio and encrypted data
- Dynamic traffic management (offloaded/inspected)
- Flexible policy provisioning (automatic/manual)

# Boosting Cyber Defenses

NVIDIA BlueField Offloads and Accelerates Palo Alto Networks' Virtual Next-Gen Firewall (NGFW)



**Software-Defined, Hardware-Accelerated**  
5x throughput performance improvement



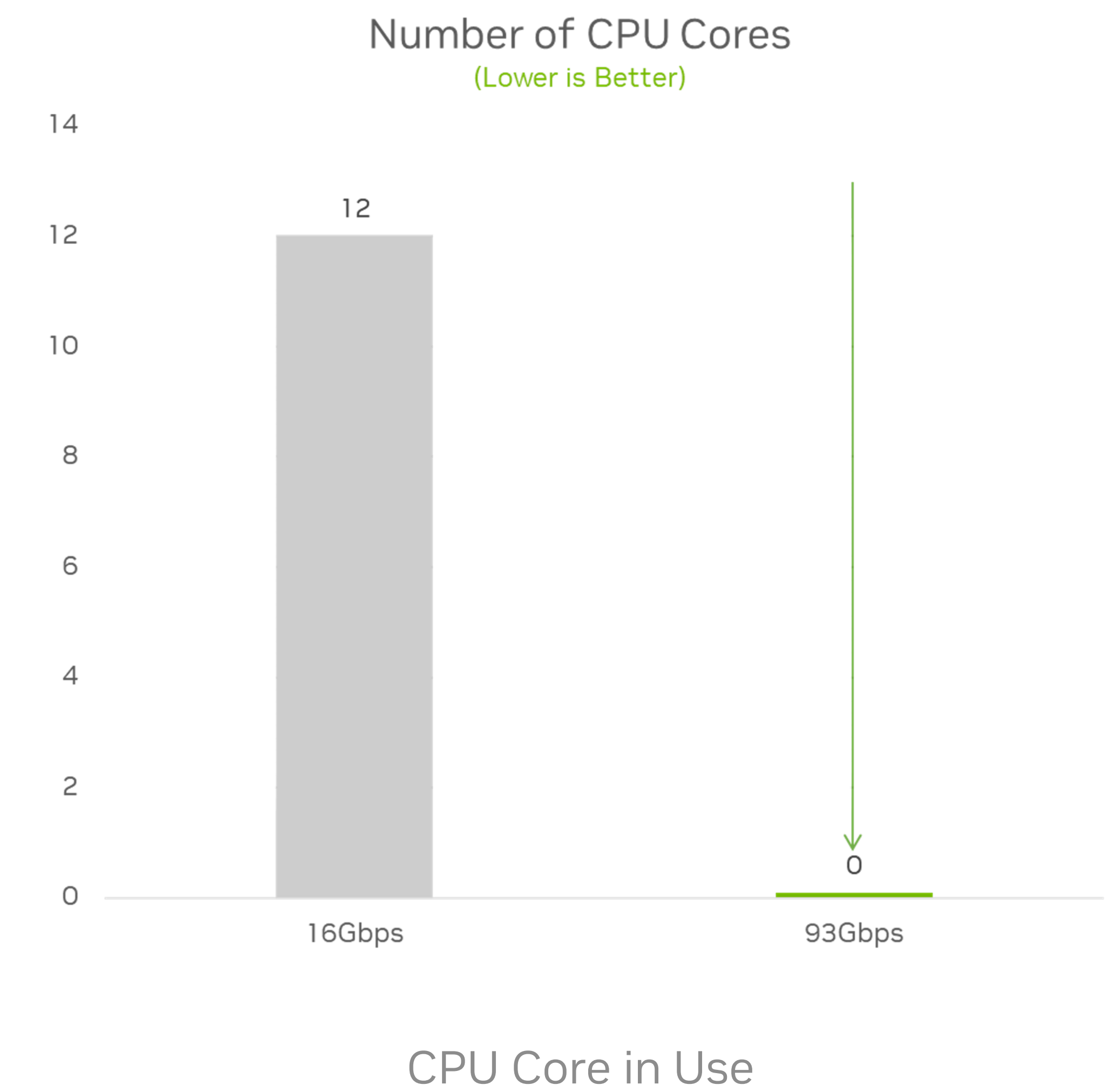
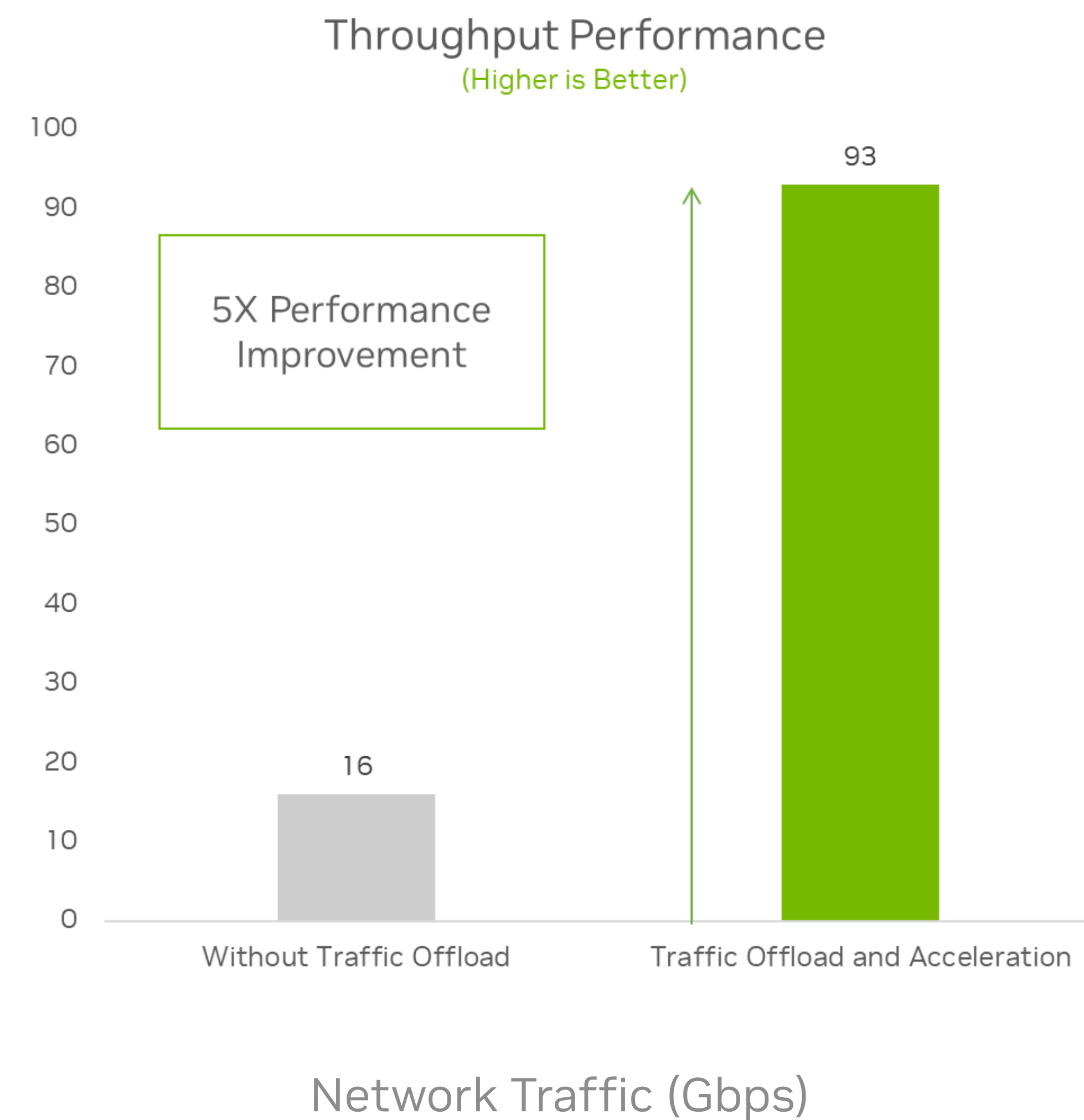
**Near-line Rate Performance**  
Protect networks at speeds up to 100Gb/s



**Incredible Efficiency**  
Drop CPU consumption from 12 to 0 cores

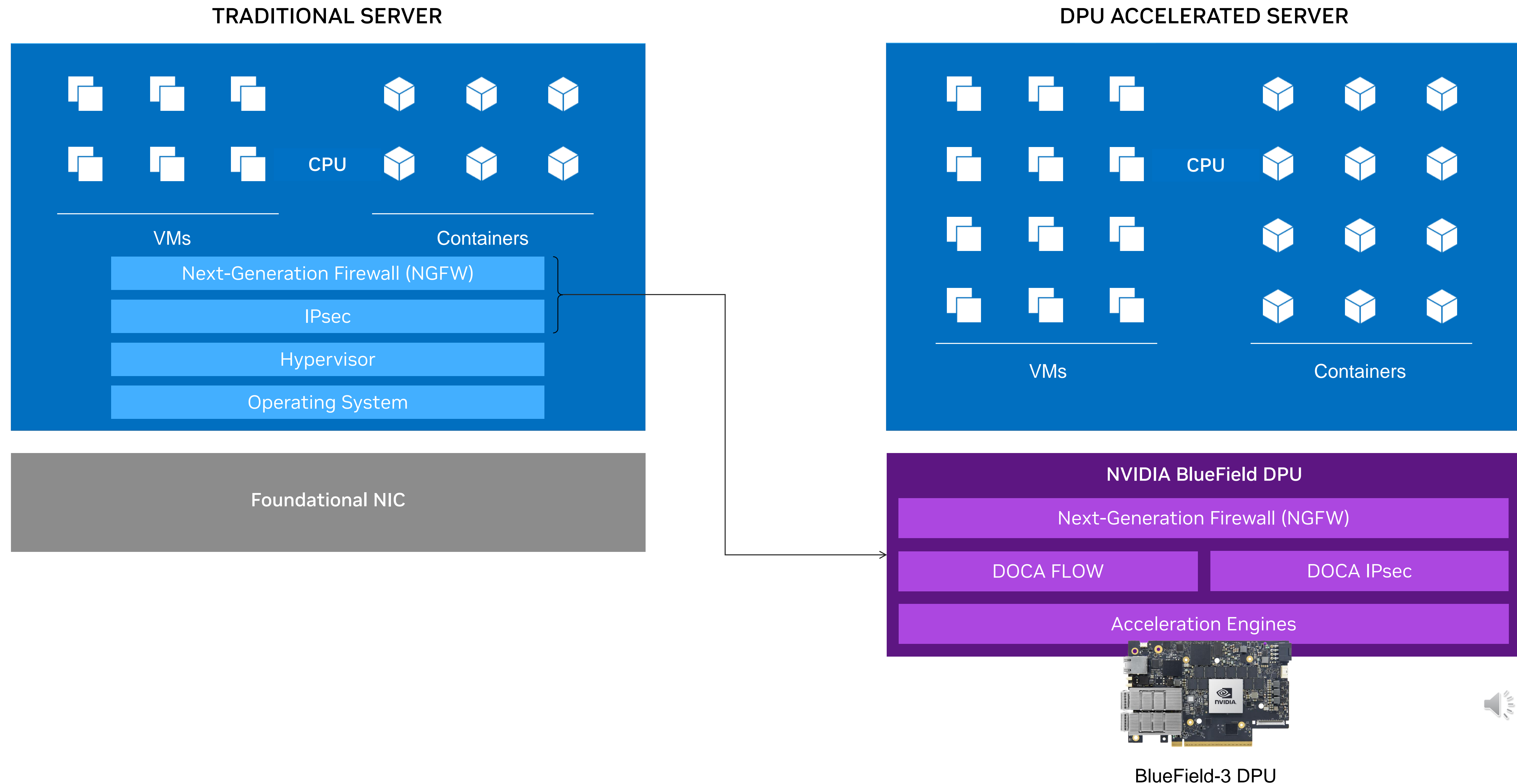


**Substantial TCO Savings**  
Up to 150% CapEx savings compared to legacy hardware



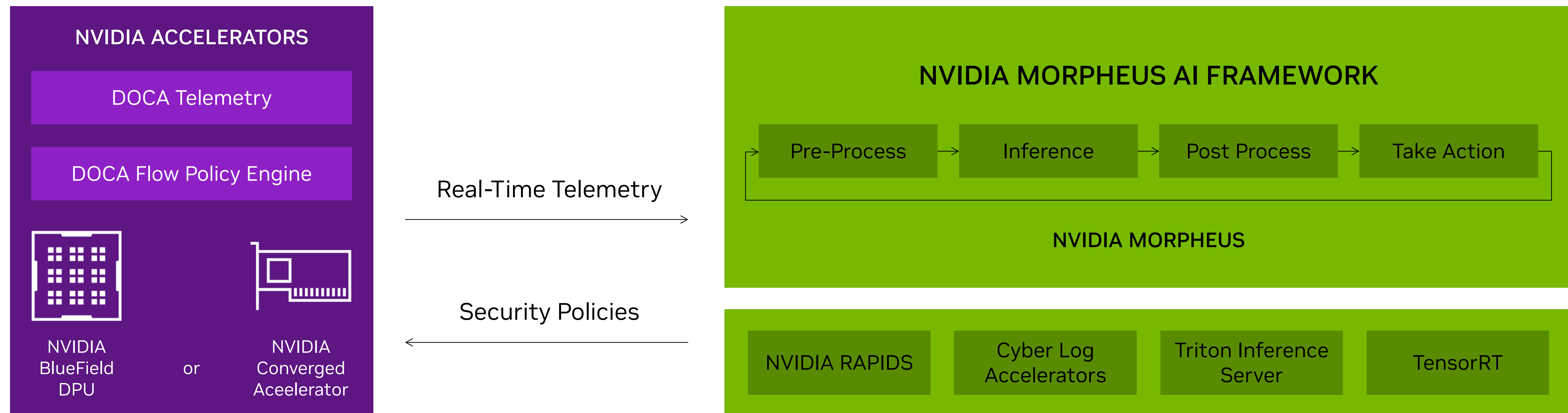
# Accelerate Virtual Next-Generation Firewalls

Powered by DOCA, BlueField-3 Offloads, Accelerates, and Isolates Virtual NGFW



# NVIDIA AI Cybersecurity Platform

NVIDIA BlueField Streams Network Telemetry into Morpheus in Real-Time



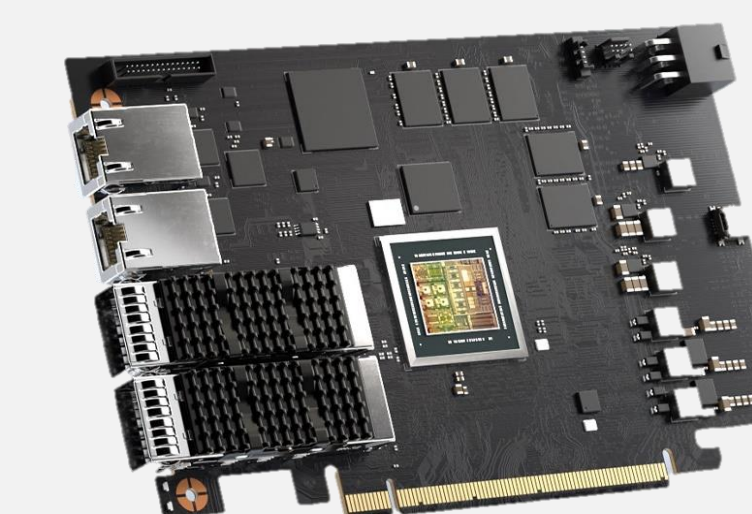
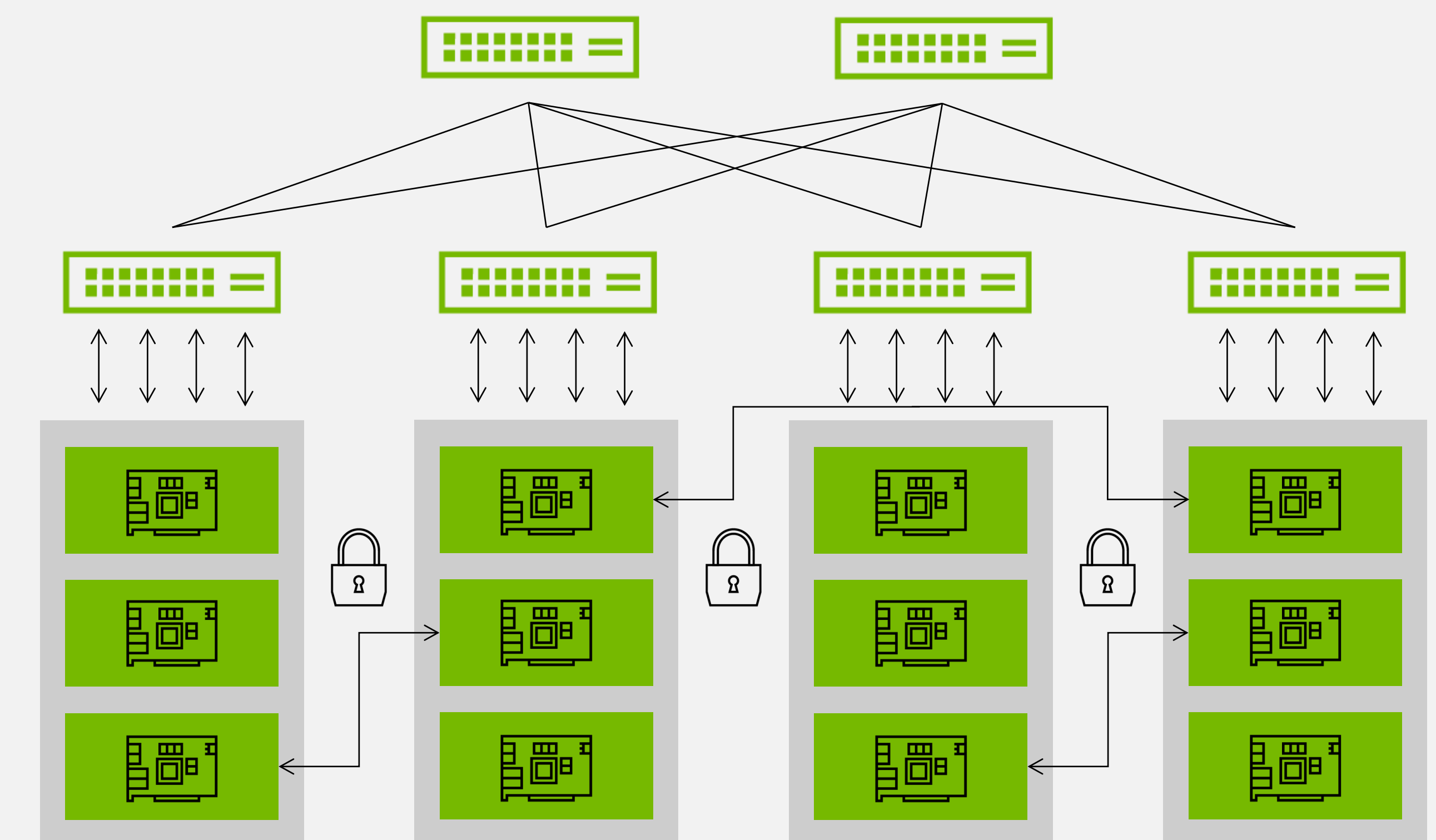
Performance, Isolation, Offloading

Automated, Real-Time Threat Detection at Scale

# Accelerating Cloud Networks

## NVIDIA DOCA Host-Based Networking Service

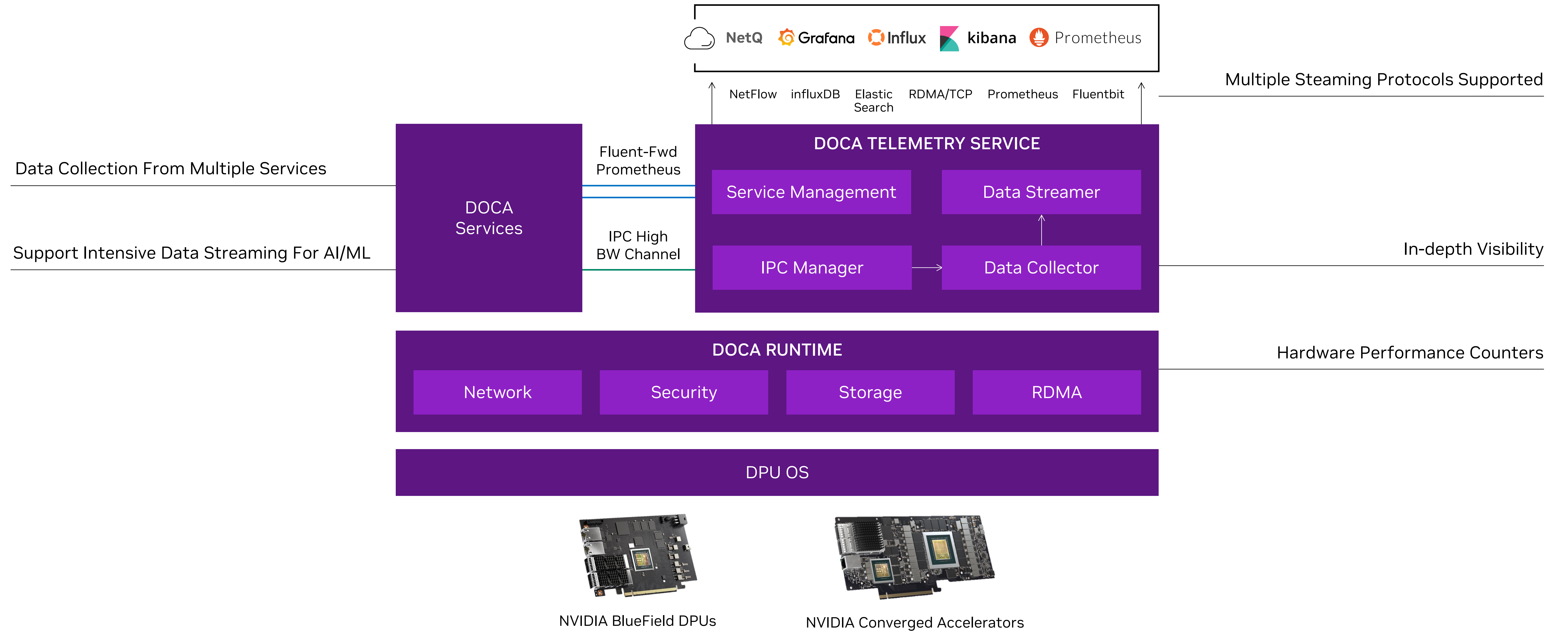
- Offload EVPN control-plane onto BlueField DPUs
- Streamline network operations, scale with confidence
- Line-speed networking and crypto performance
- Data center wide visibility with NVIDIA NetQ
- Advanced SDN programmability through DOCA



NVIDIA Cumulus on BlueField-2

# DOCA Telemetry

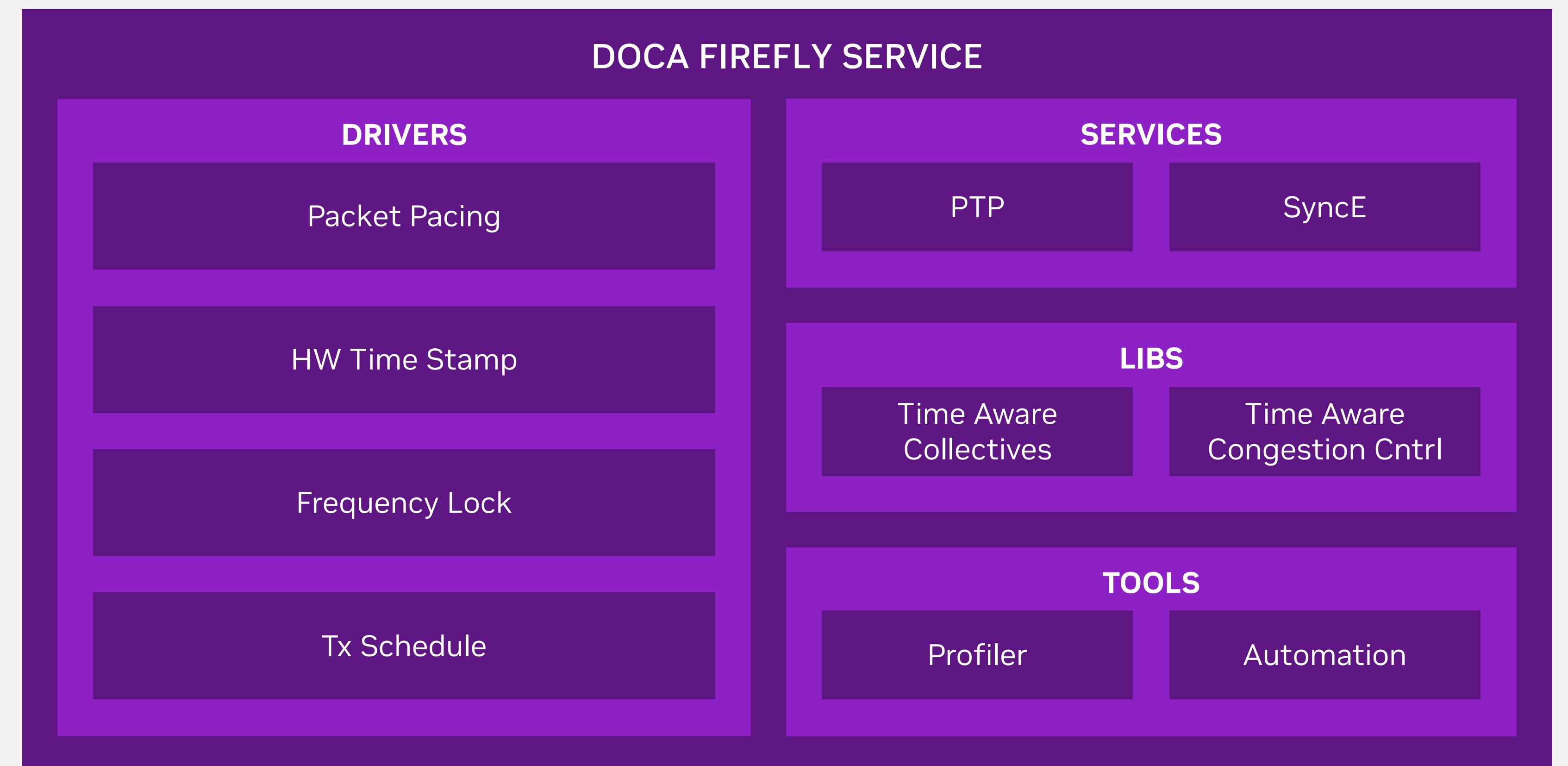
Complete Data Center Visibility



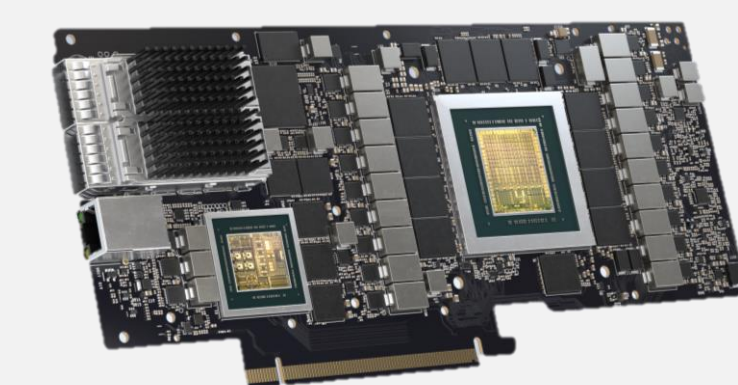
# DOCA Firefly

## Precision Time Synchronized Data Center Service

- Precision timing as a service for Data Center
- Time synchronous data center improves compute, storage and networking
- Time accelerated applications:
  - Globally synched DBs
  - Distributed Cache
  - Time aware congestion control



BlueField DPUs



Converged Accelerators



# NVIDIA Rivermax Accelerates Data Streaming Applications

Achieve Higher Application Performance and Infrastructure Efficiency



## Kernel Bypass

Transfer data between user space application's memory and the network interface



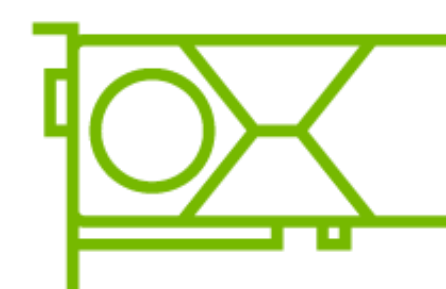
## CPU Efficient

Offload packet processing to the BlueField DPU hardware accelerators



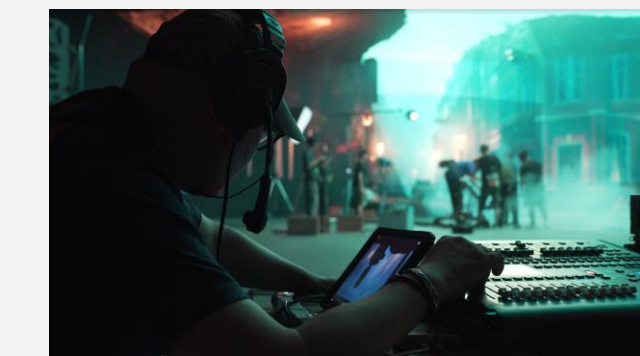
## Precision Timing

Synchronize clocks in real-time, hardware timestamp, packet pacing and scheduling

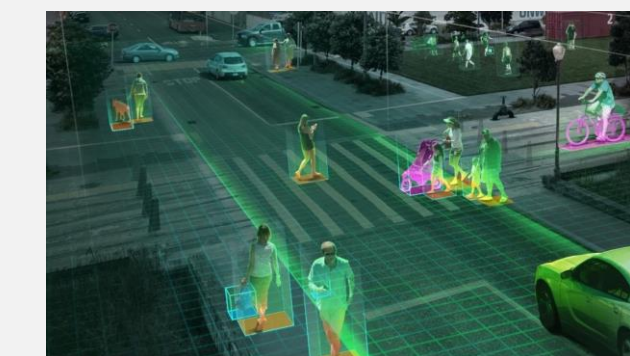


## GPUDirect

Deliver packets directly to the GPU memory



Production



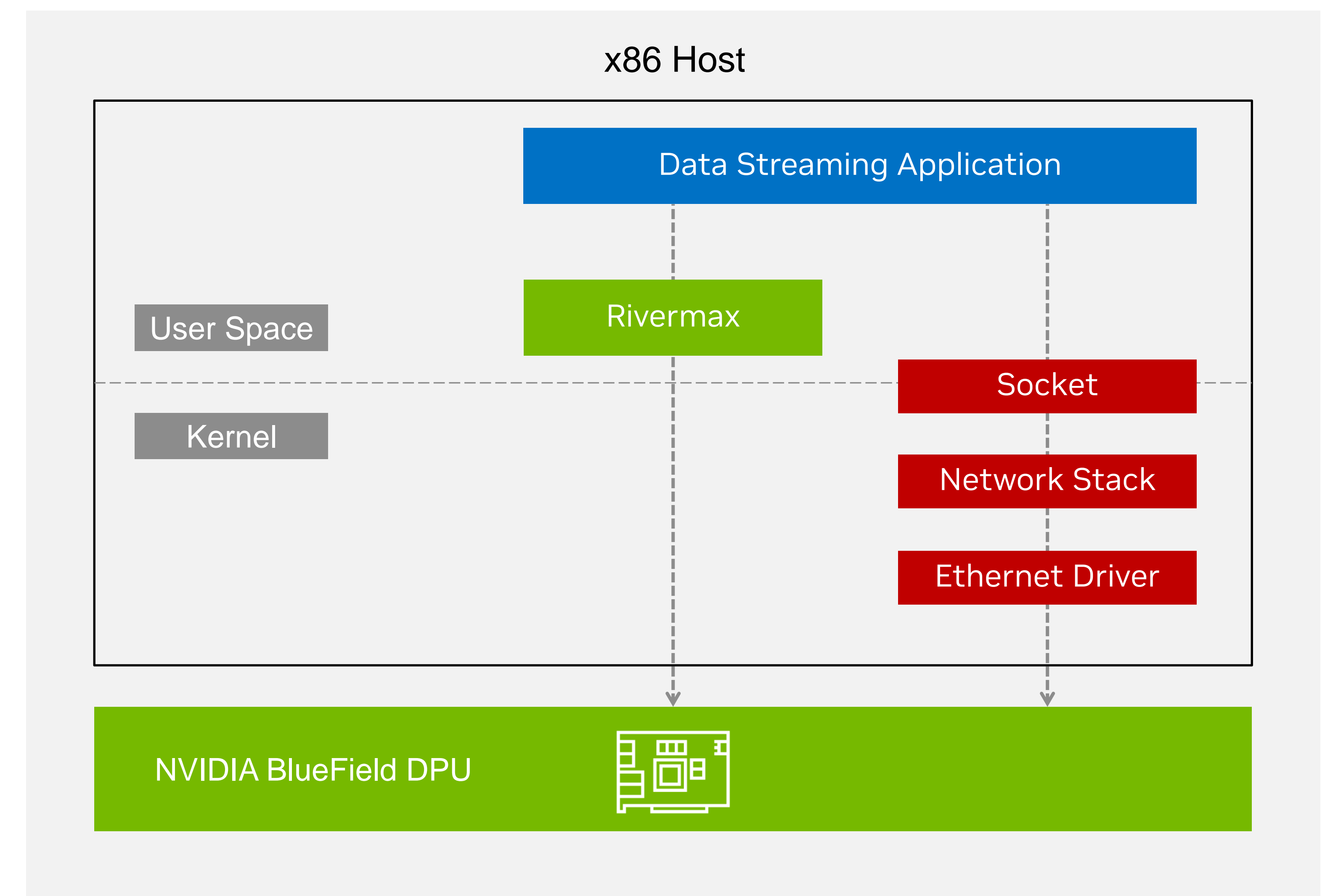
Smart City



Healthcare

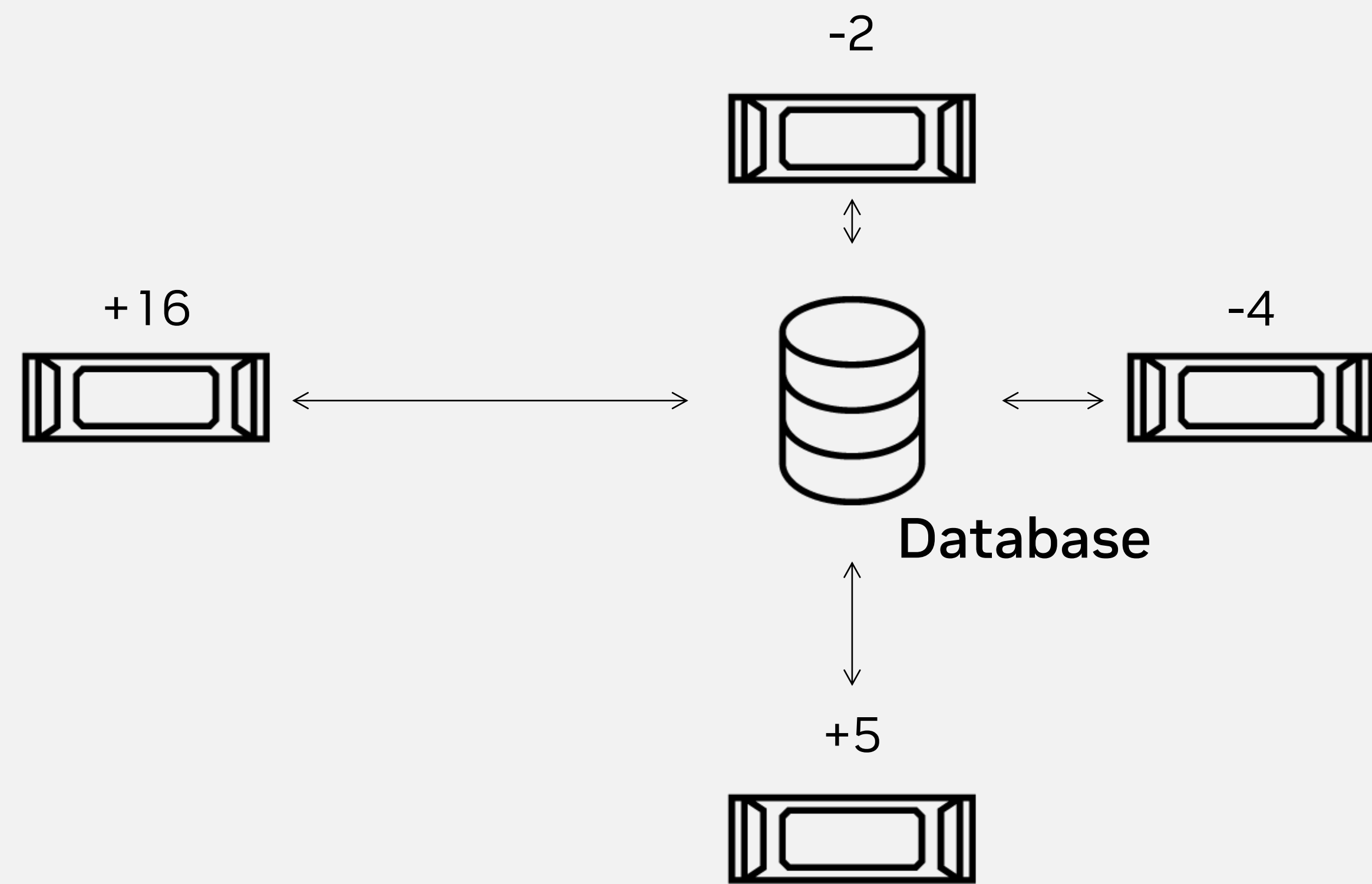


Live Broadcast

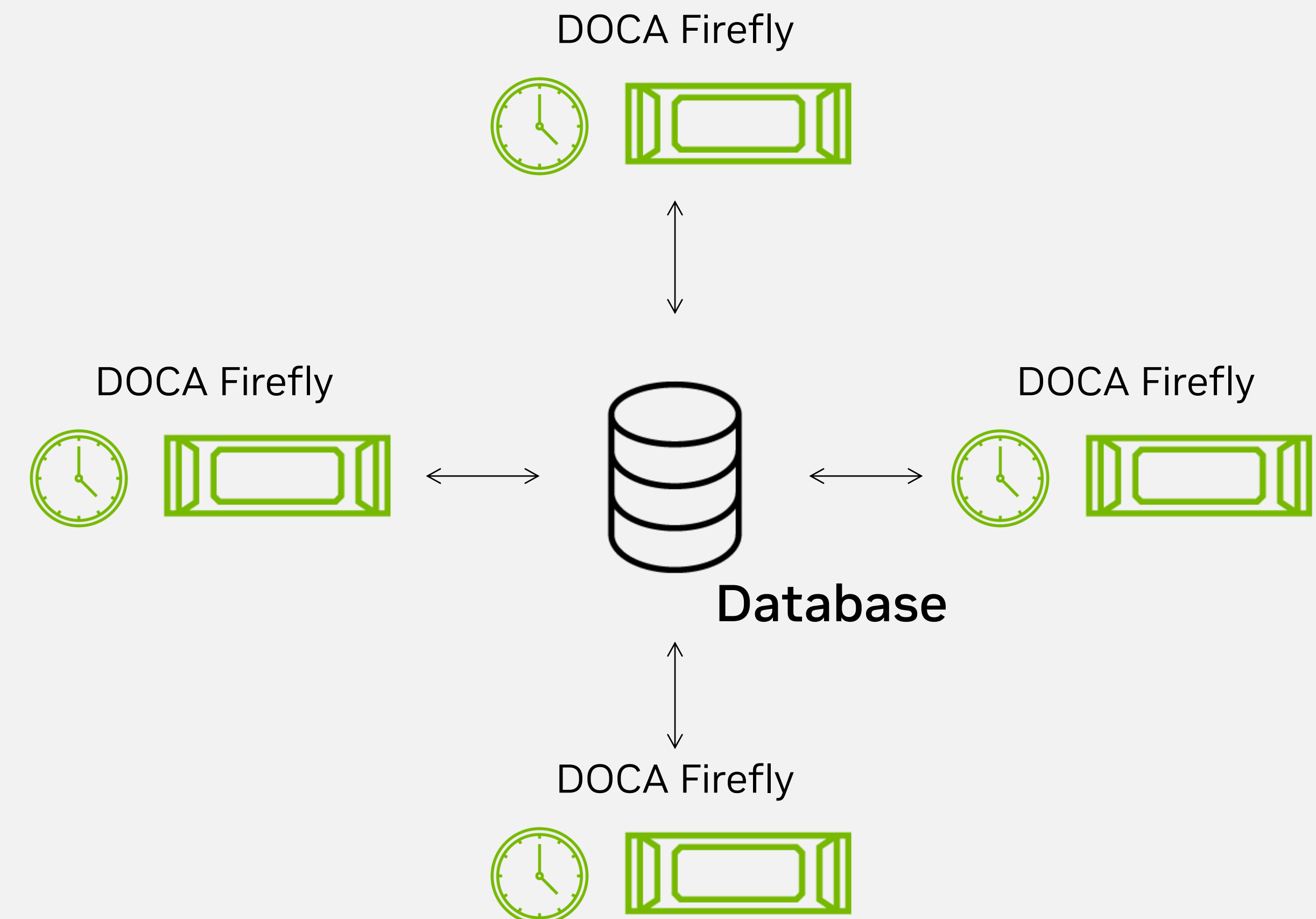


# Accelerating Scale-Out Databases

NVIDIA BlueField's Precision Timing Service Speeds-Up Database Performance



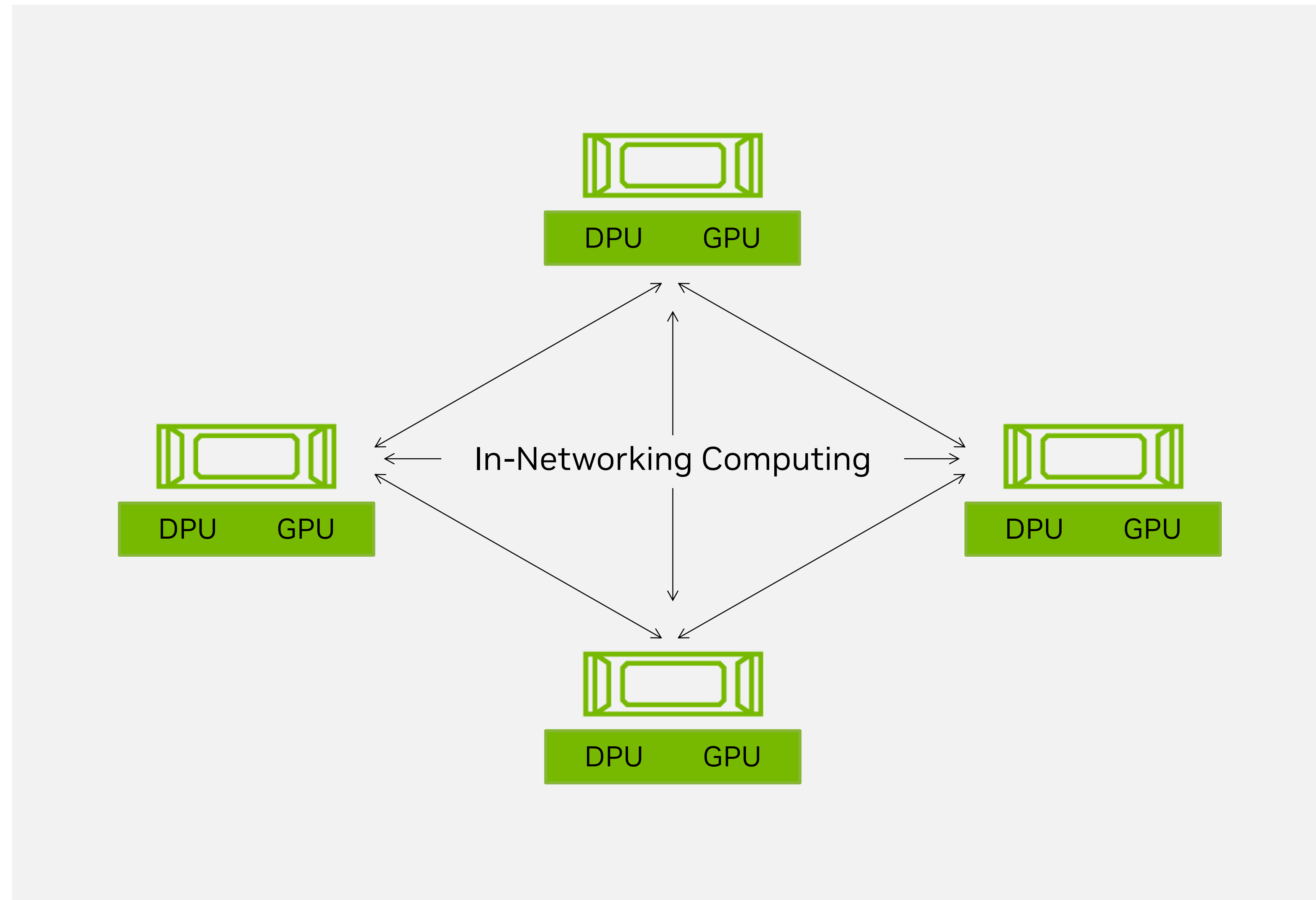
**NTP-Based Data Center**  
20 Millisecond Wait Time



**Precision Time Synchronized Data Center**  
50 Microsecond Wait Time, 400X Faster, 3X Database Performance

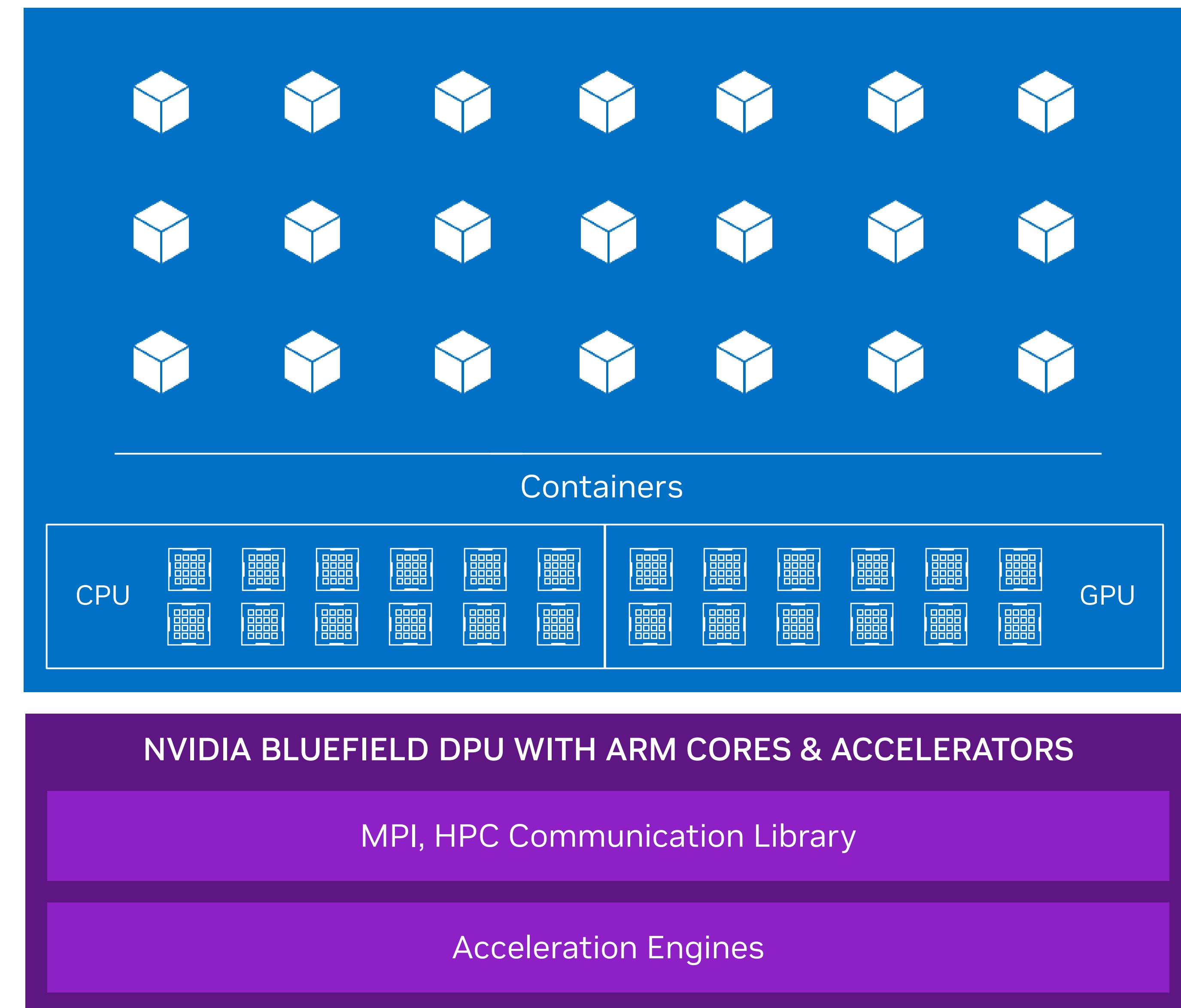
# Accelerating ML/AI Workloads

NVIDIA DPU and GPU Accelerate ML/AI and Scientific Computing Workloads



Accelerating MPI Operations

## DPU and GPU ACCELERATED SERVER



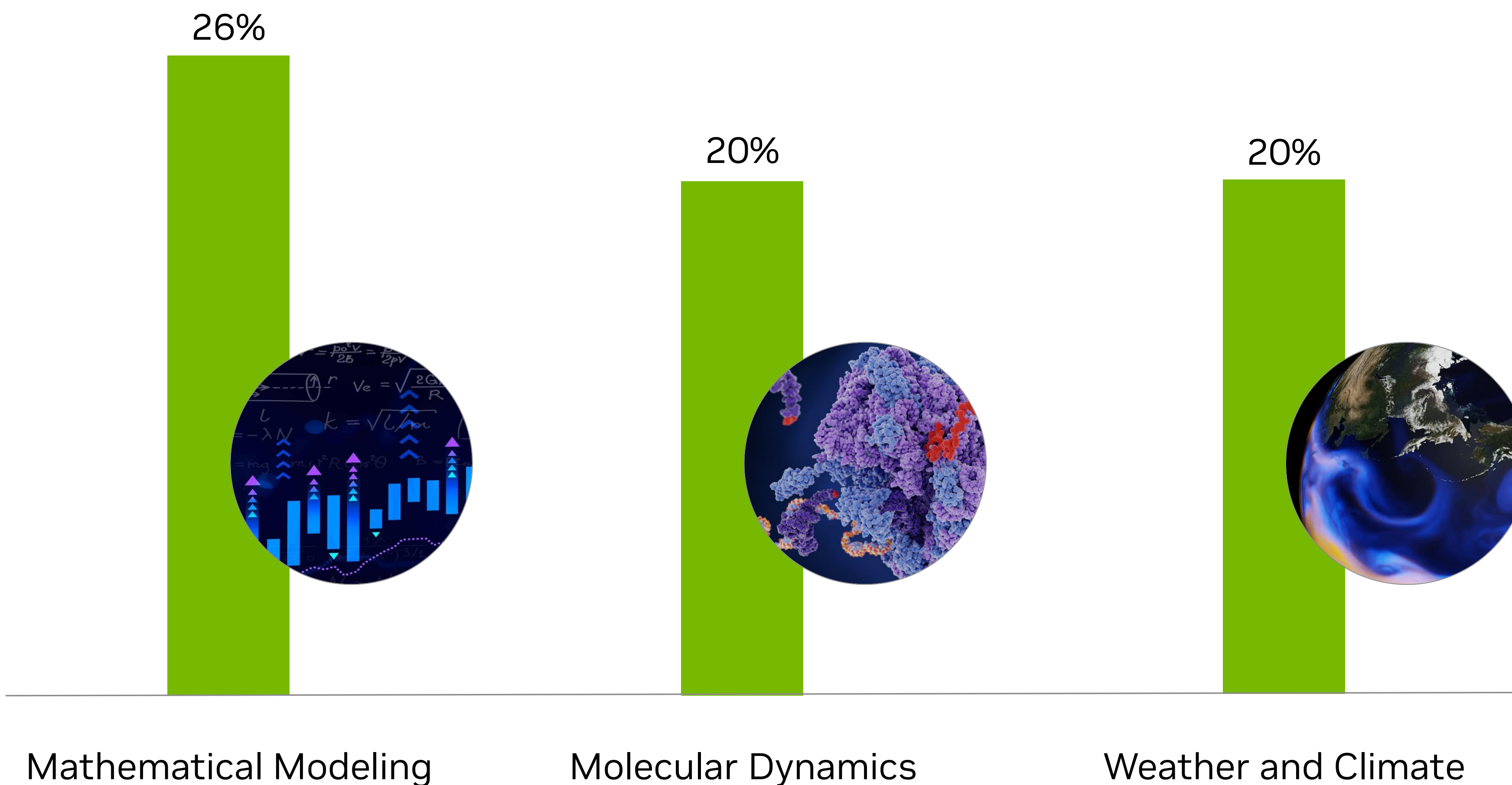
Higher GPU Utilization

Offload, Accelerate  
All-to-All, All-Reduce Operations

# Accelerating Scientific Computing Workloads

Ignite High-Performance Computing with NVIDIA BlueField and Quantum InfiniBand

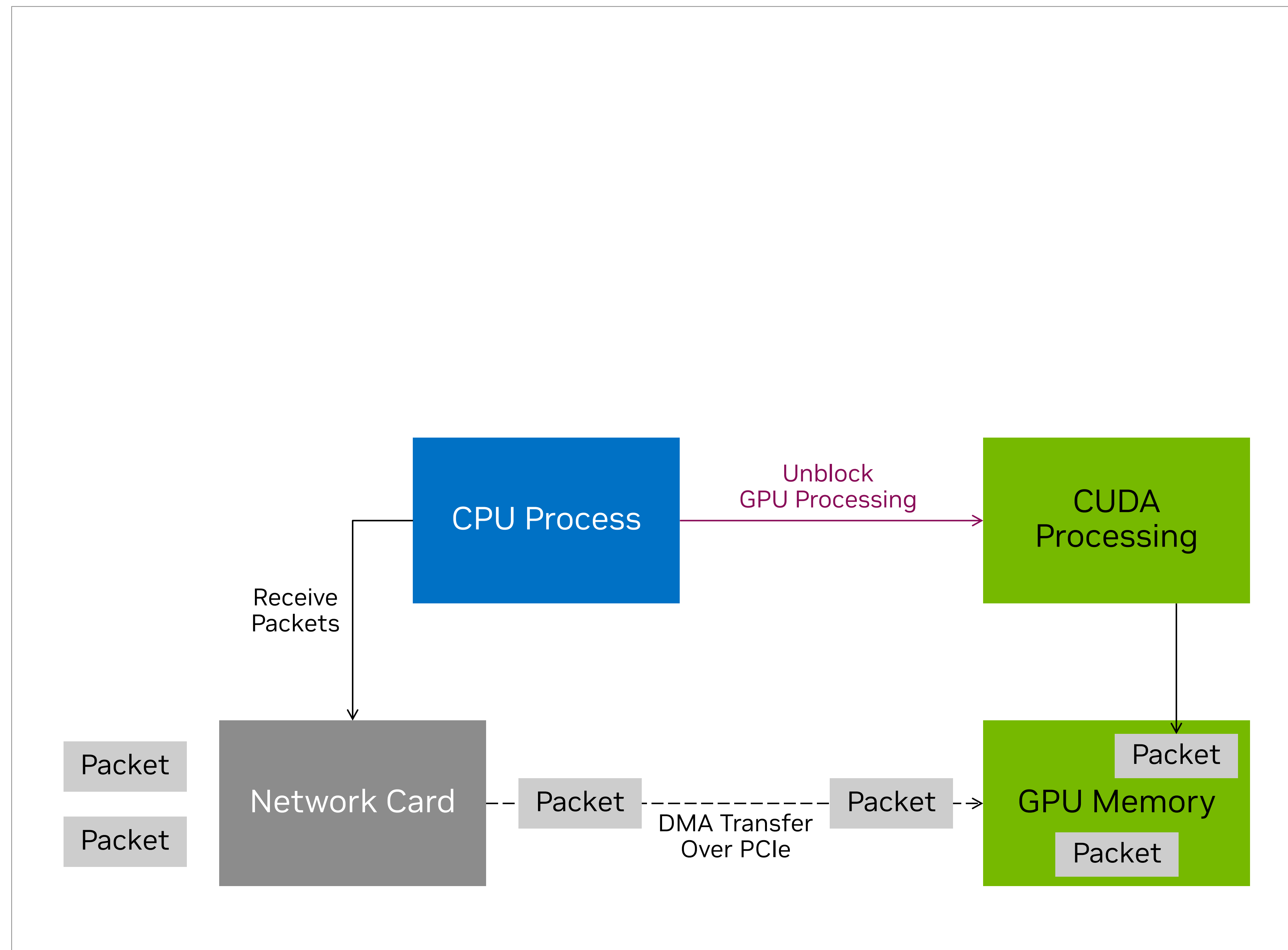
Application Performance Improvement



- Unleash application performance and system efficiency
- MPI performance acceleration
- Computational storage and advanced workloads
- Adaptive performance isolation

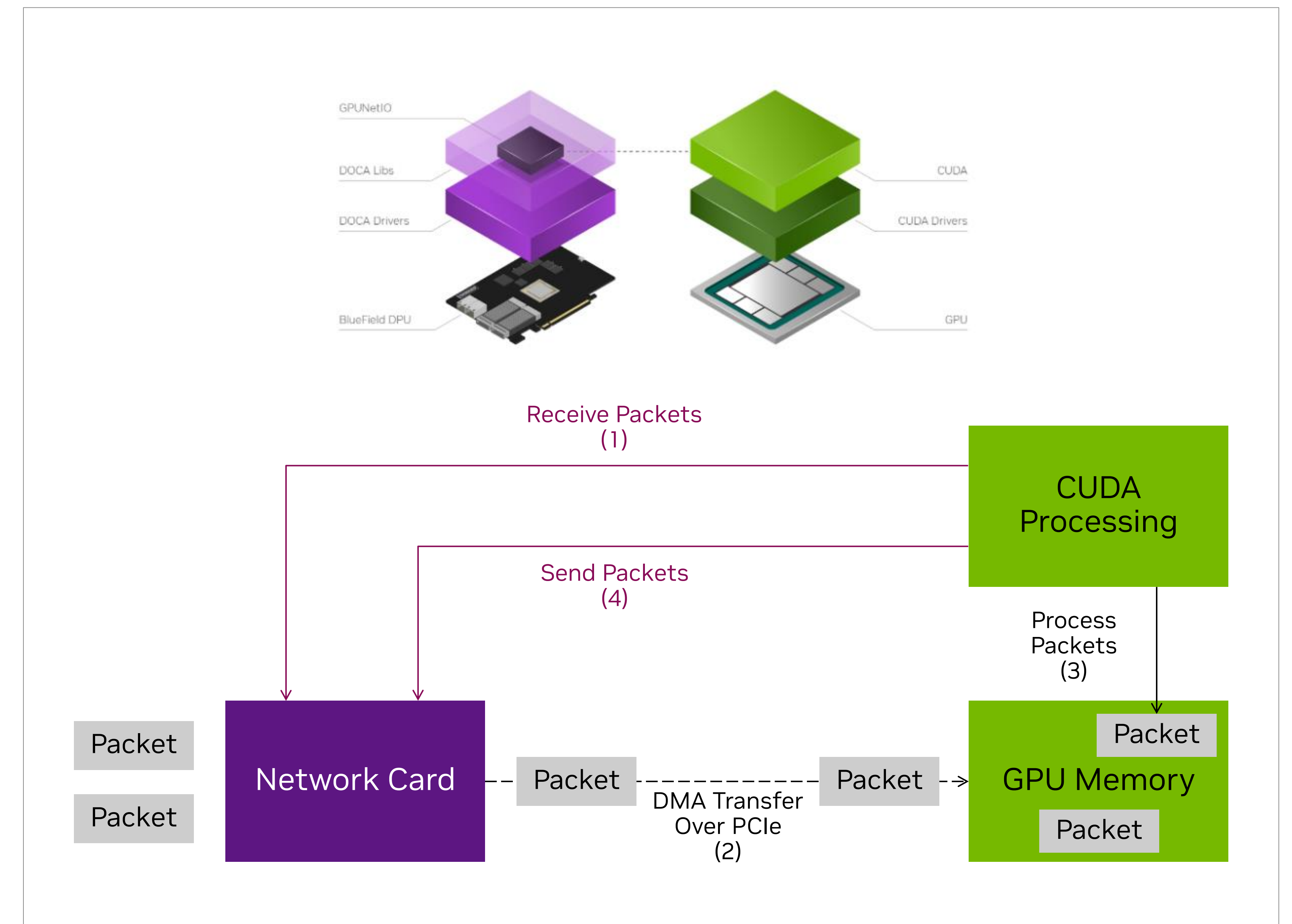
# Building Efficiency into GPU-Accelerated Workloads

DOCA GPUnetIO Removes CPU Bottleneck



## CPU-Centric Application

CPU orchestrating the GPU and network card work

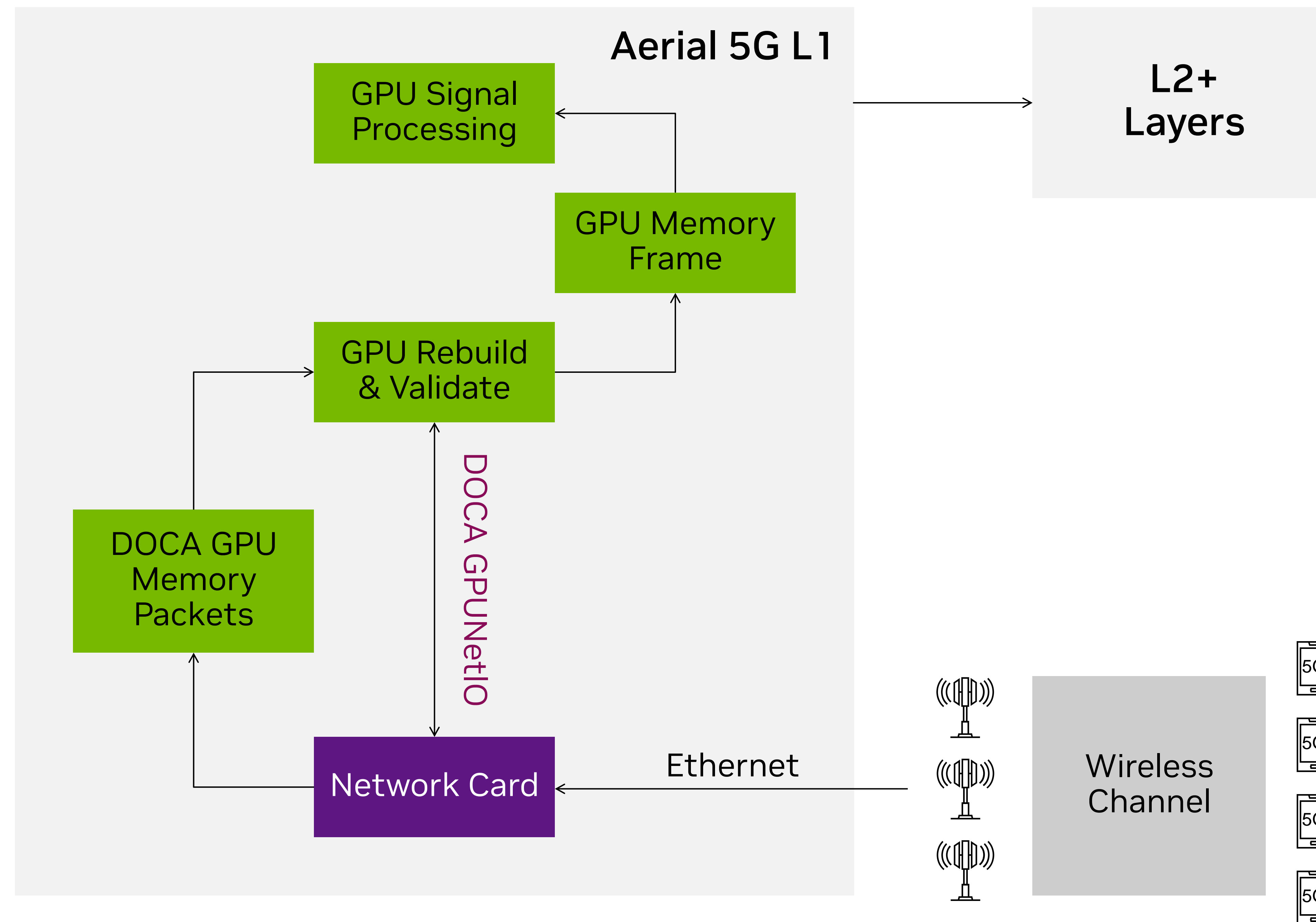


## GPU-Centric Application

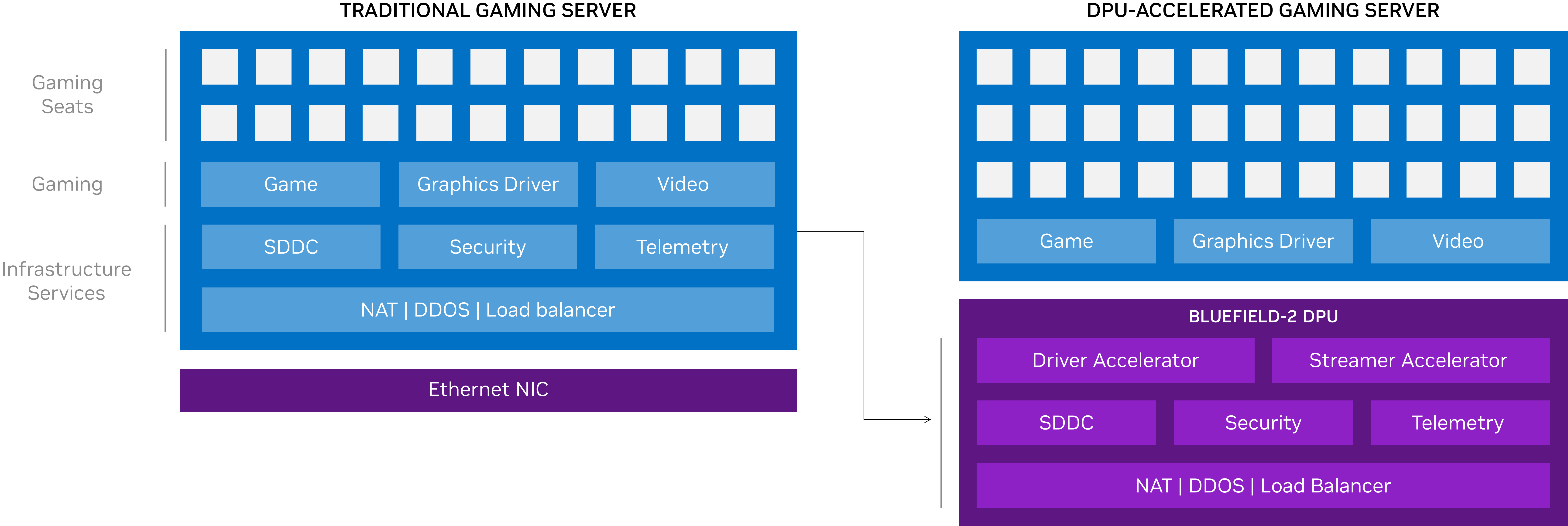
GPU controlling the network card and packet processing without the need of the CPU

# Building Efficiency Into GPU Communications for 5G Workloads

Aerial Application Framework for 5G Applications



# Securing and Accelerating Cloud Gaming Platforms



# Securing and Accelerating Cloud Gaming Platforms



## Enhanced Gaming Experience

Ensure delightful user experience while delivering consistent and predictable application performance



## More Concurrent Users

Scale the number of concurrent user per server by freeing up compute resources from infrastructure



## Secure Infrastructure

Protect data and assets in the cloud without compromising application performance



## Unified, Consistent Operations

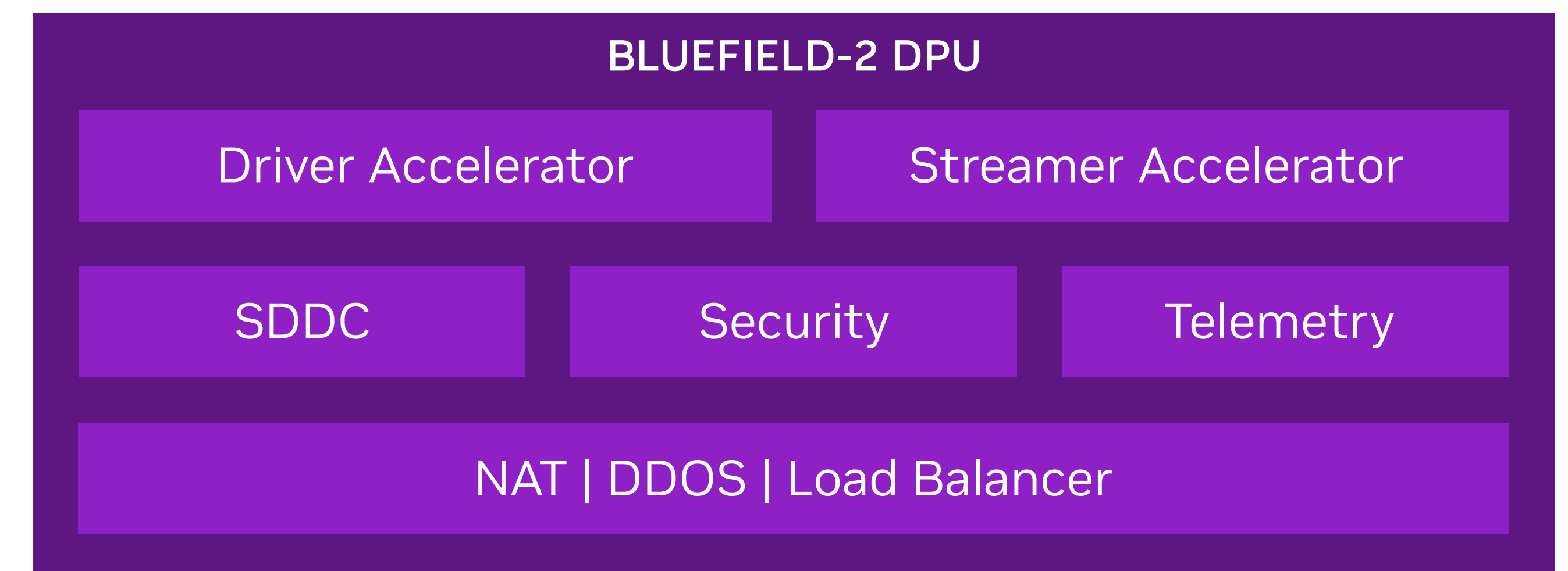
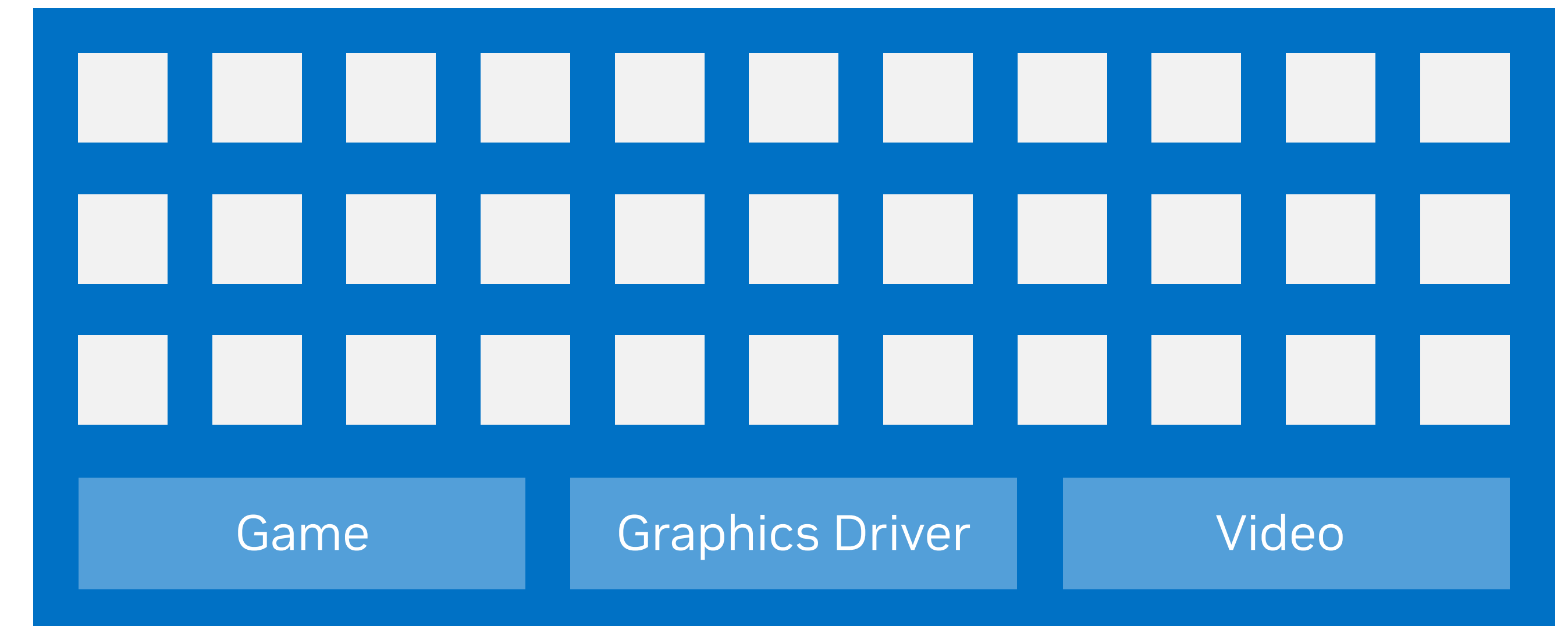
Run bare metal servers, simplify lifecycle management, and reduce TCO

Gaming Seats

Gaming

Infrastructure Services

### DPU-ACCELERATED GAMING SERVER





# NVIDIA BlueField DPU Platform

Software-Defined, Hardware-Accelerated Infrastructure Compute Platform



## Accelerated Performance

Meet the most stringent performance requirements, run the most demanding workloads



## Cloud-Scale Efficiency

Free up x86 cores to business apps, achieve unprecedented scale and efficiency levels



## Robust Zero-Trust Security

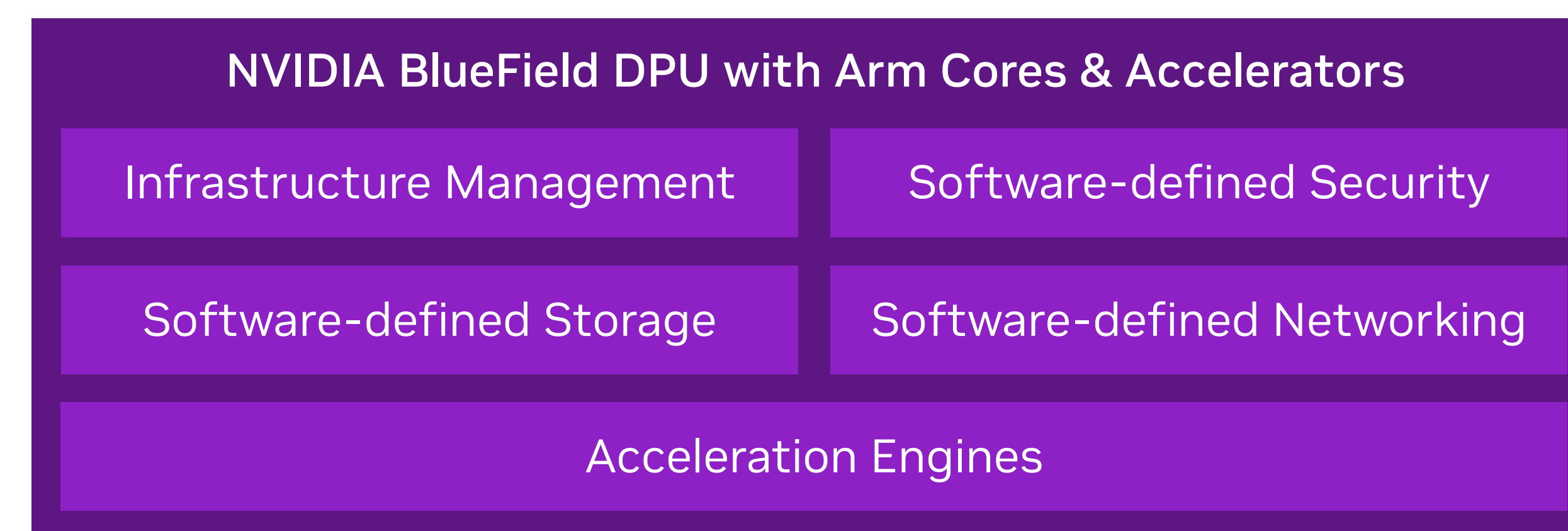
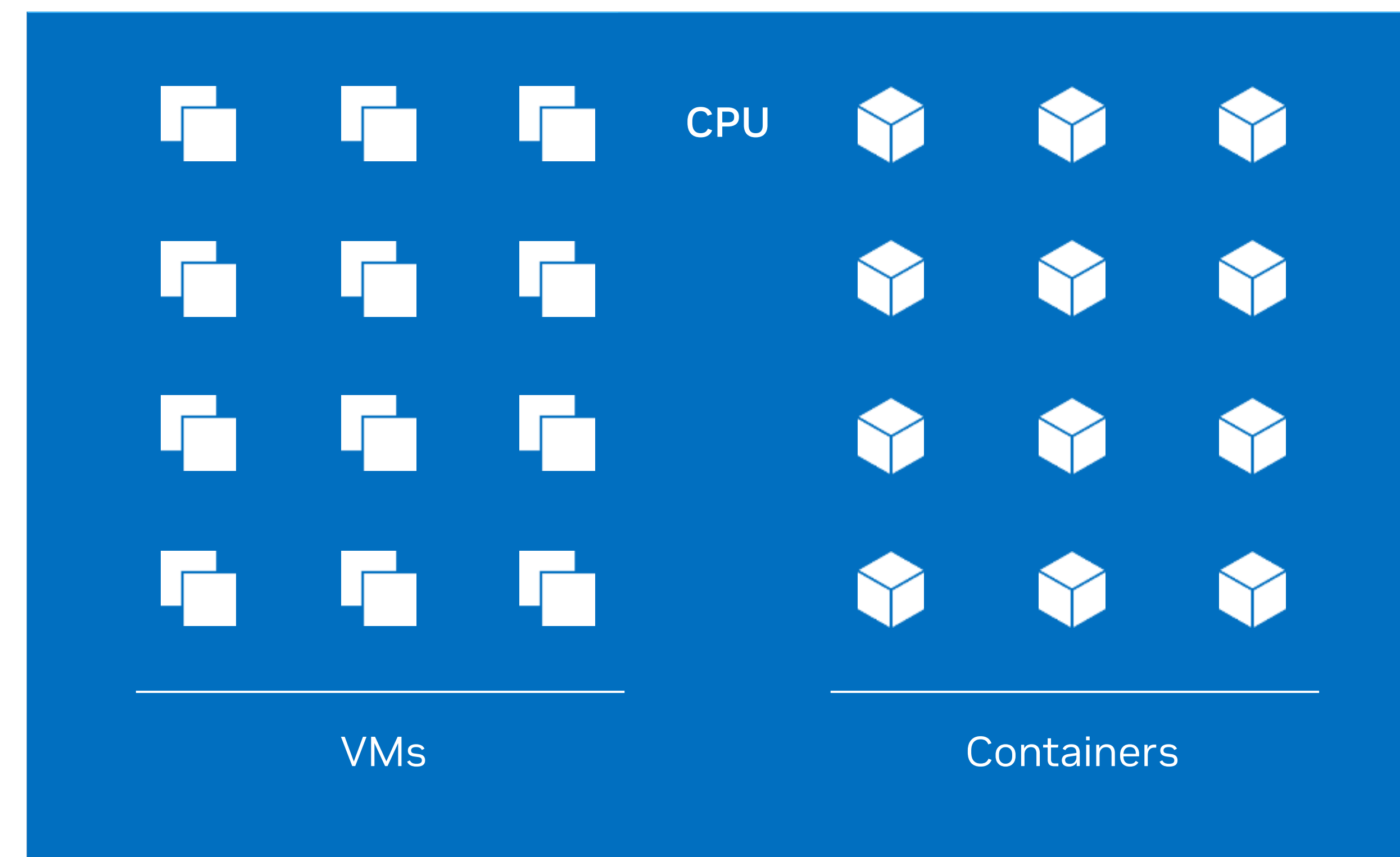
Ensure comprehensive data center security without compromising performance



## Programmable Infrastructure

Develop and run applications consistently with maximum performance

## DPU ACCELERATED SERVER



Offload | Accelerate | Isolate

# Transform the Data Center with NVIDIA BlueField DPUs

World's Most Advanced Computing Platform for Data Center Infrastructure



Questions?

