# Securely Deploy and Operate HGX/MGX AI Data Centers

Sungta Tsai, SA | Aug 2023.

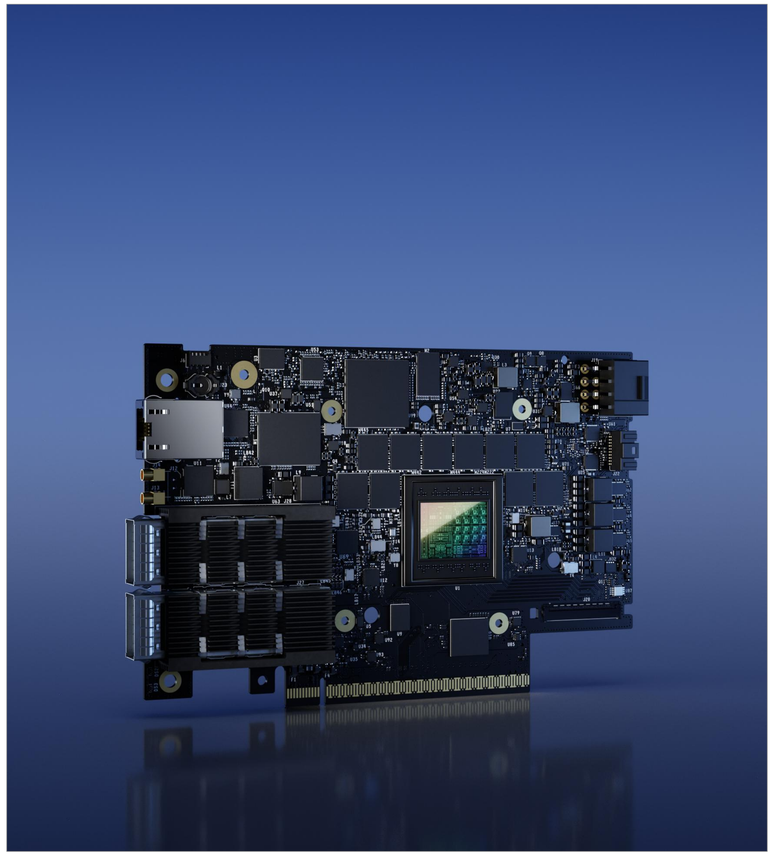
# Securely Deploy and Operate AI Data Centers

## Powered by NVIDIA BlueField

### Elastic GPU Computing
Rapid 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 GPU / MGX Grace Hopper
The world's most advanced enterprise AI infrastructure

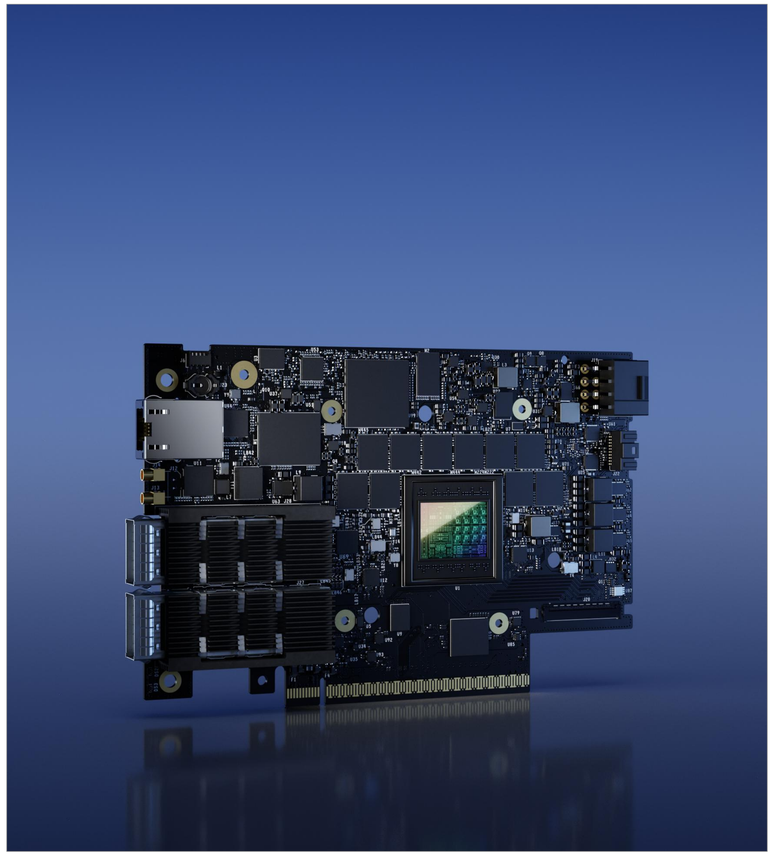
# Securely Deploy and Operate AI Data Centers

## Powered by NVIDIA BlueField

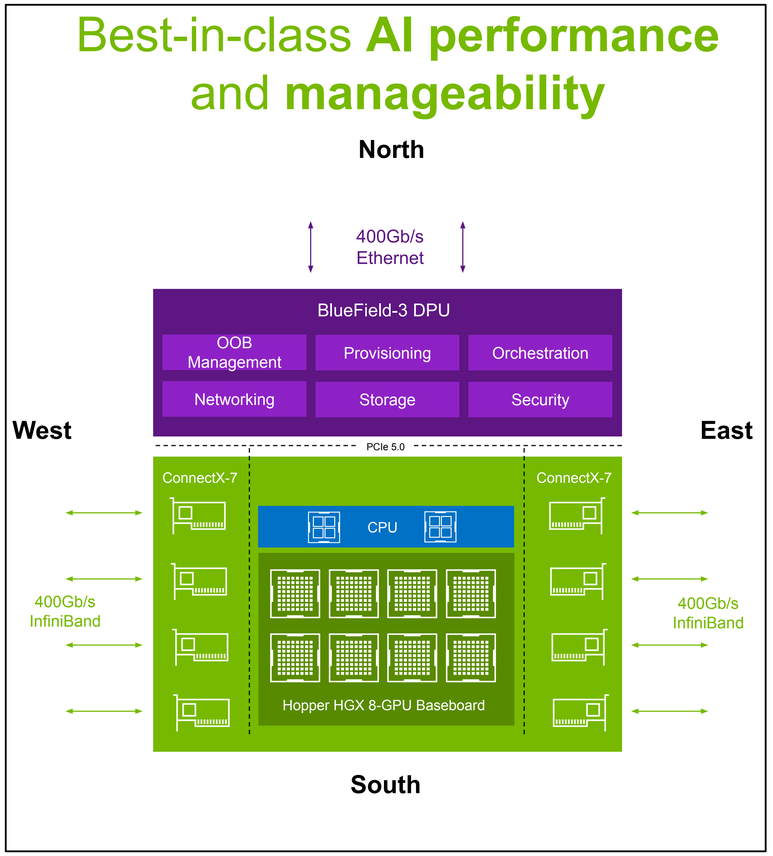
### Elastic GPU Computing
Rapid provisioning, fungible GPU compute and limitless scaling

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

# NVIDIA BlueField-3 Overview

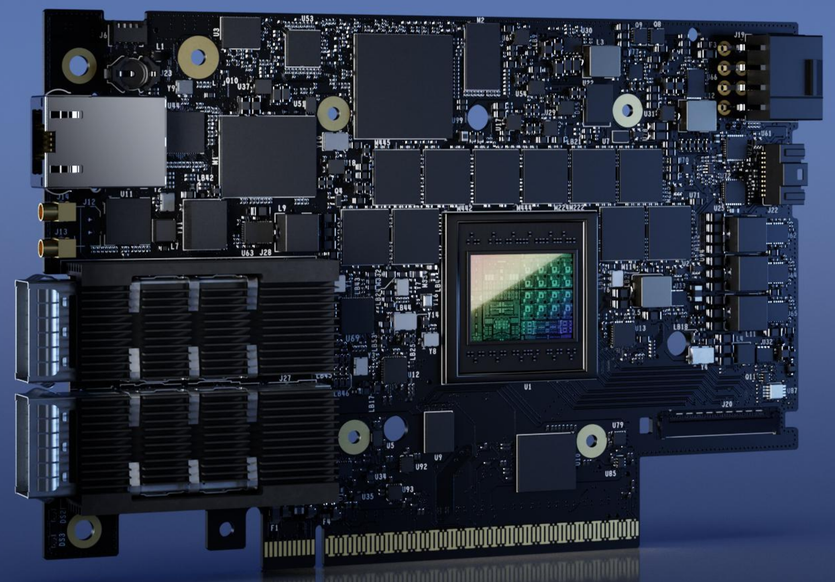
## 400Gb/s Infrastructure Compute Platform

### 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



Infrastructure Compute Platform

Offload | Accelerate | Isolate

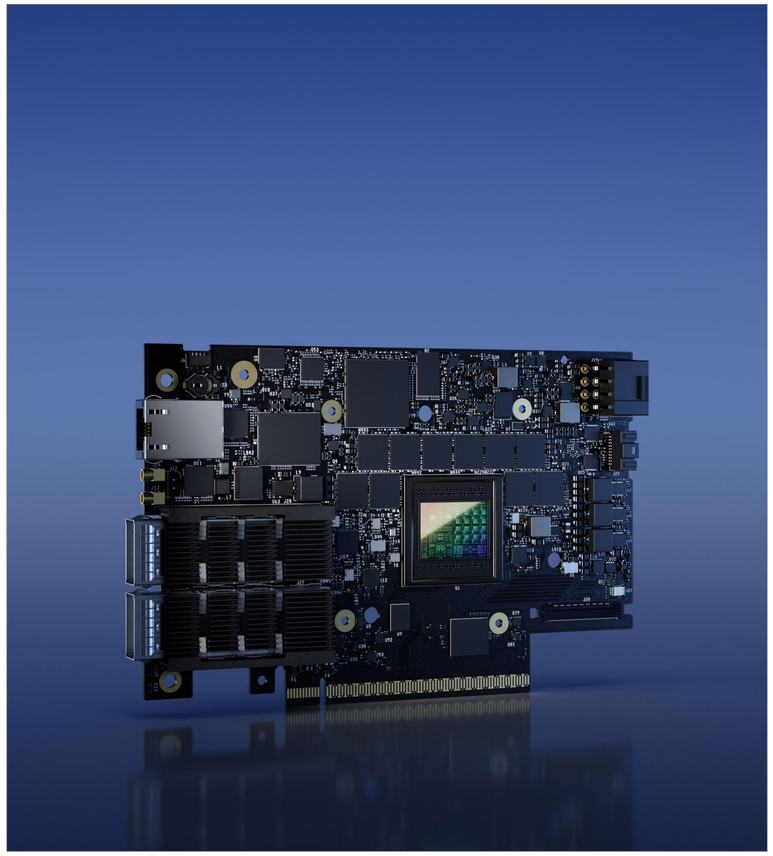
# Securely Deploy and Operate AI Data Centers

## Powered by NVIDIA BlueField

### Elastic GPU Computing
Rapid 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

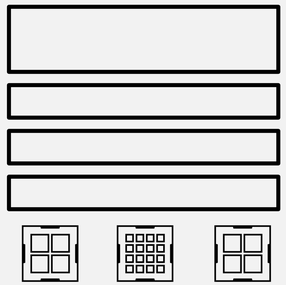


NVIDIA BlueField-3 DPU
400Gb/s Infrastructure compute platform

NVIDIA HGX H100 GPU / MGX Grace Hopper
The world's most advanced enterprise AI infrastructure

# Organizations Struggle to Operationalize Generative AI

Building accelerated AI data centers is an incredibly complex and challenging task

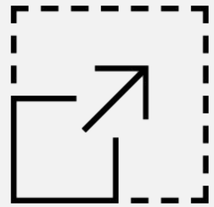


### Infrastructure Complexity

Accelerated computing is a full-stack challenge

### Massive Scale

LLMs and Generative AI can scale up to tens of thousands of GPUs

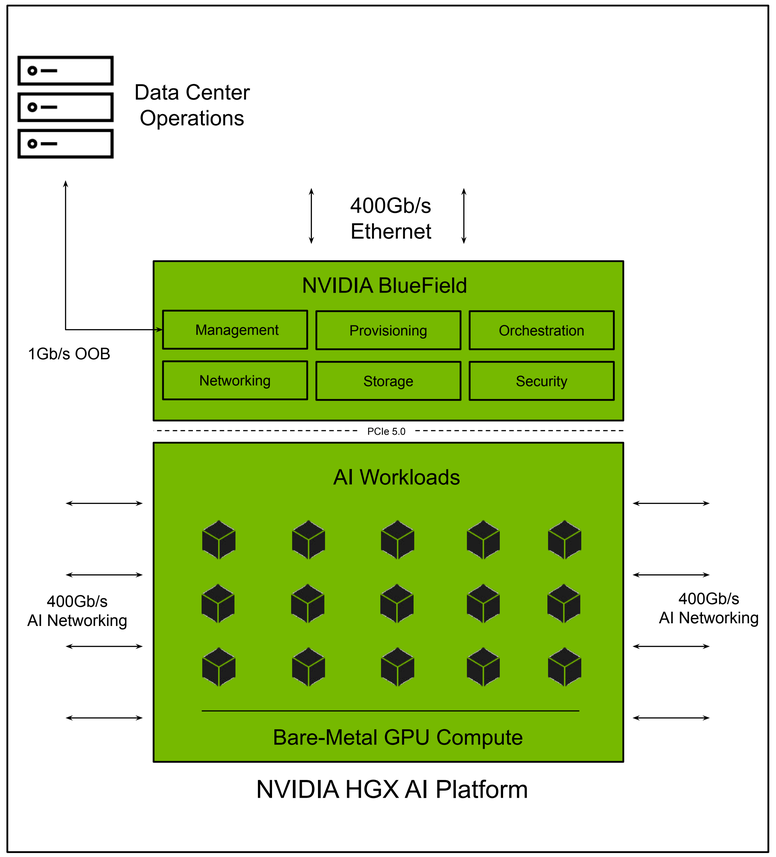


### Fluctuating Demand

AI workloads are transient; often resulting in over or under provisioning of resources

# NVIDIA BlueField Accelerates Time-to-Market for Generative AI

## Operationalize an AI data center and launch apps in days, not months

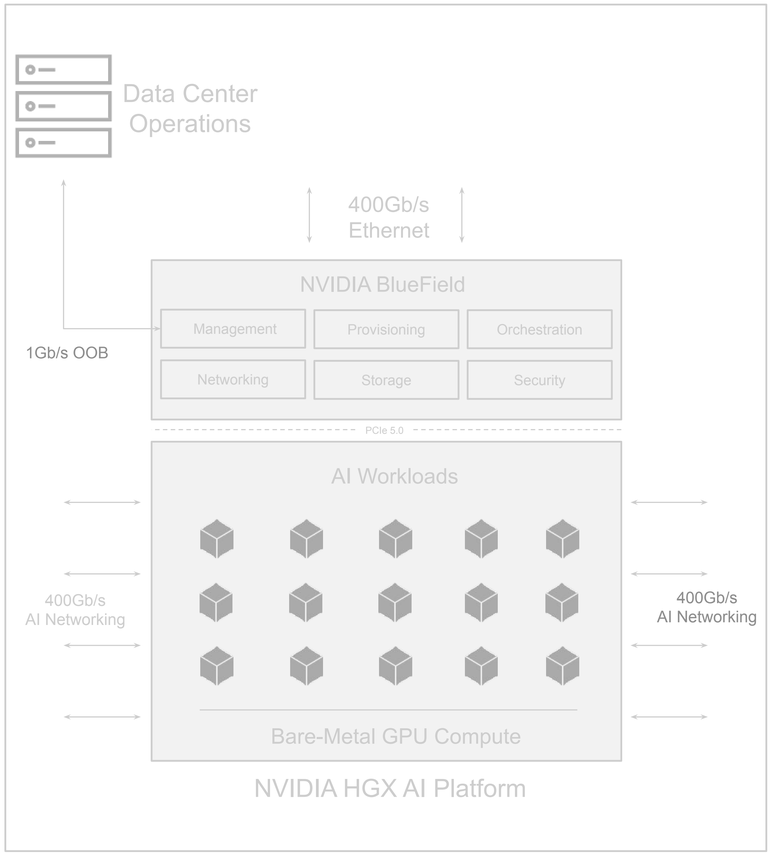



### Rapid Provisioning

Speed-up lifecycle operations from bring-up to production

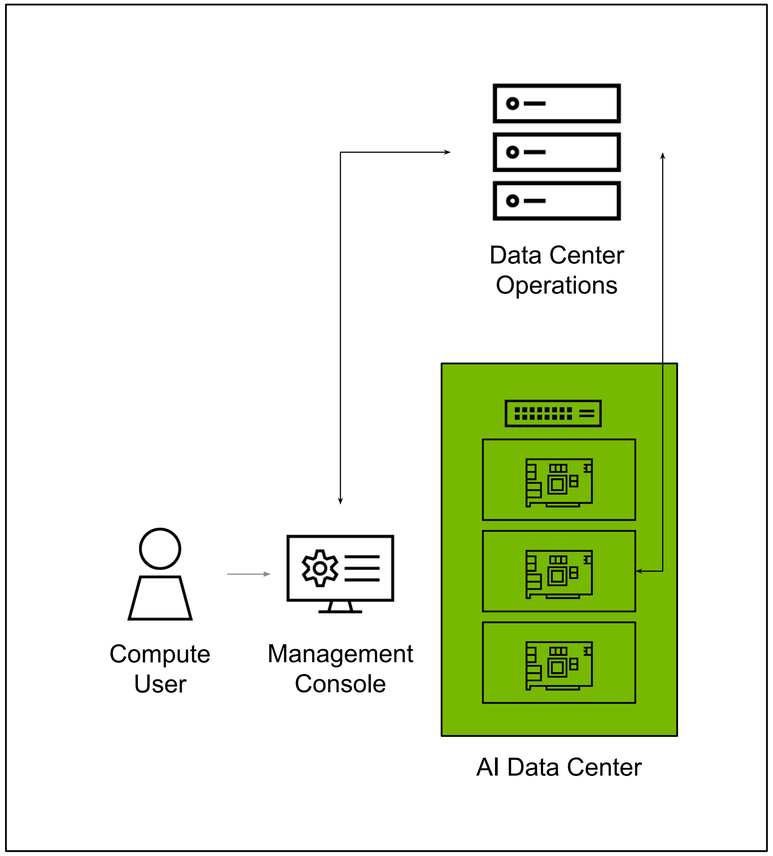
# NVIDIA BlueField Accelerates Time-to-Market for Generative AI

## Repurpose data center infrastructure and deploy new workloads in hours, not weeks




### Rapid Provisioning

Speed-up operations from bring-up to production

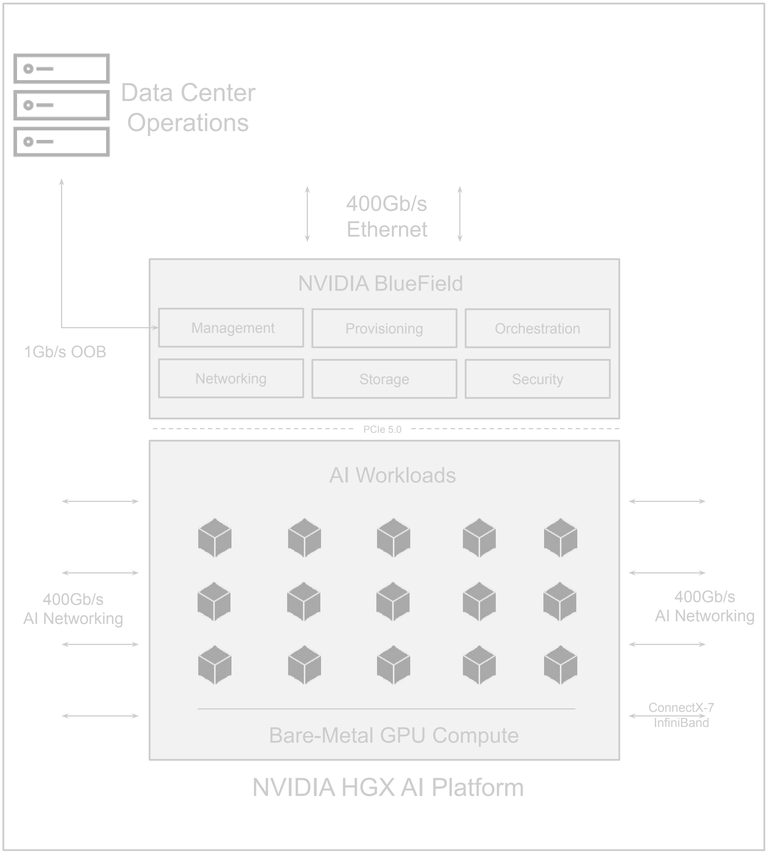



### Elastic, Fungible Capacity

Dynamically repurpose and allocate resources to users

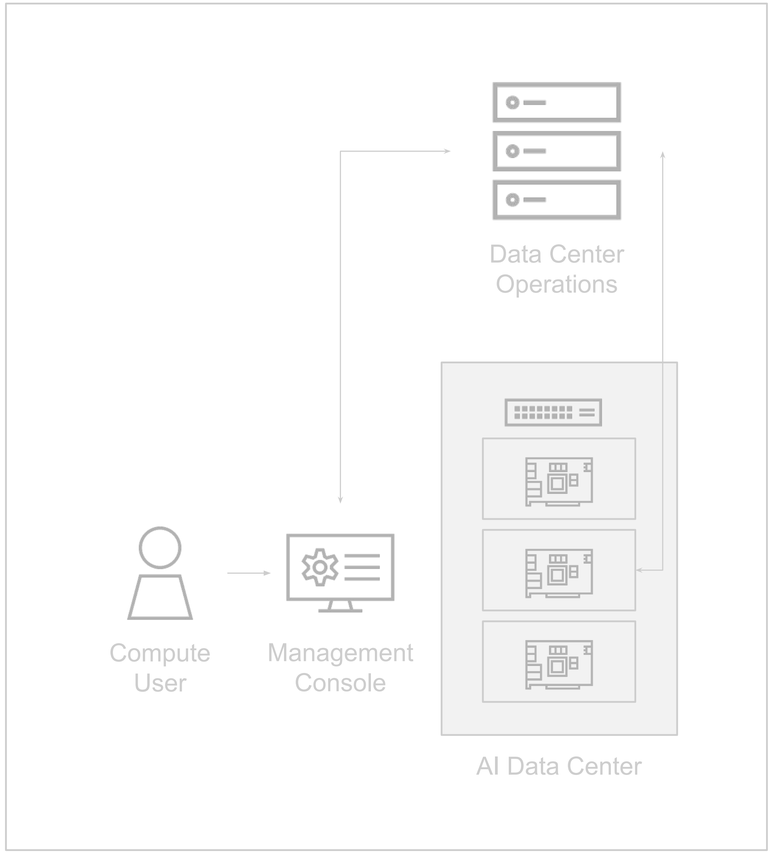
# NVIDIA BlueField Accelerates Time-to-Market for Generative AI

## Scale effectively from dozens to tens of thousands of GPUs



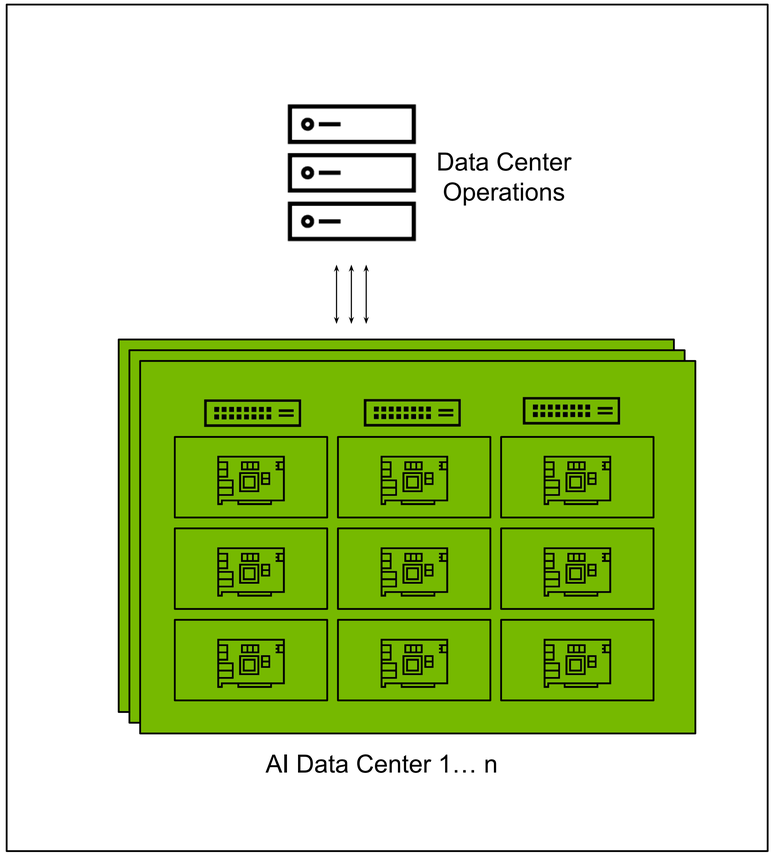

### Rapid Provisioning

Speed-up operations from bring-up to production




### Elastic, Fungible Capacity

Dynamically repurpose and allocate resources to users

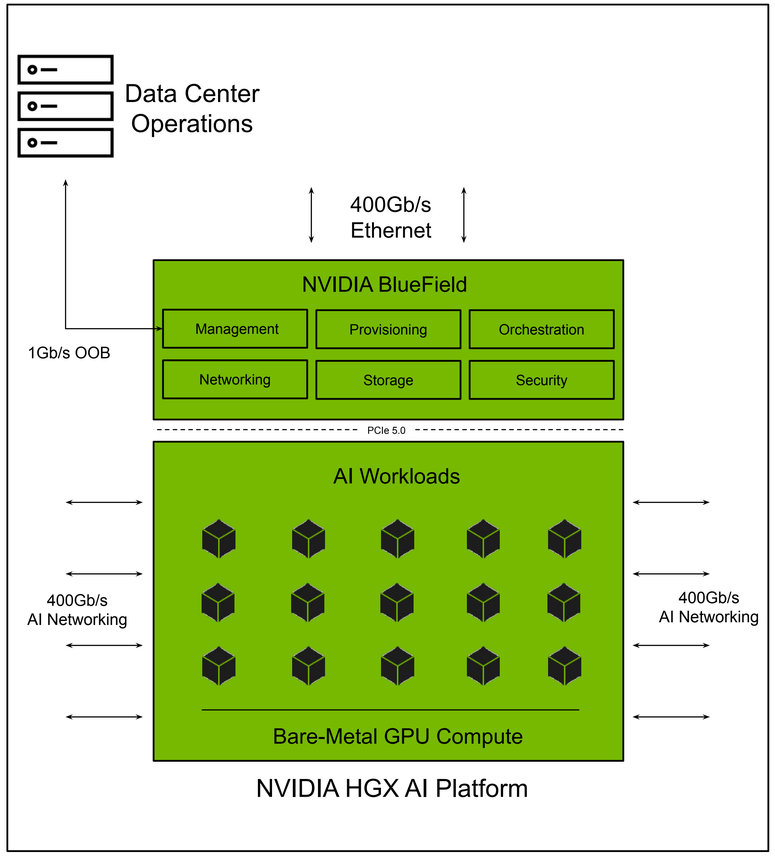



### Limitless Scaling

Scale with confidence through a robust, fault-tolerant design

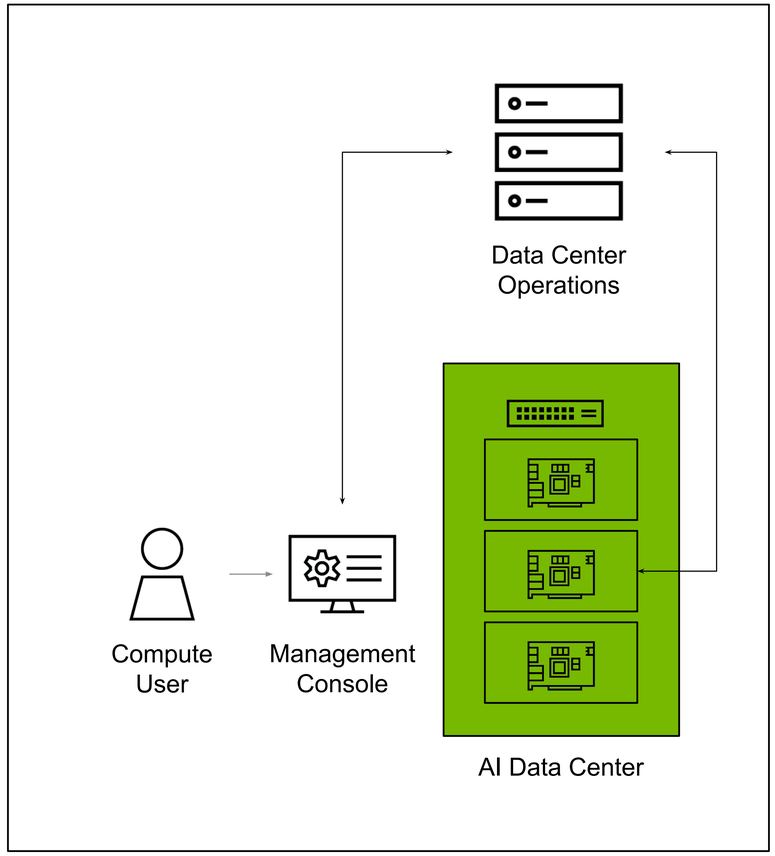
# NVIDIA BlueField Accelerates Time-to-Market for Generative AI

## BlueField streamlines AI data center deployment and operations at every scale



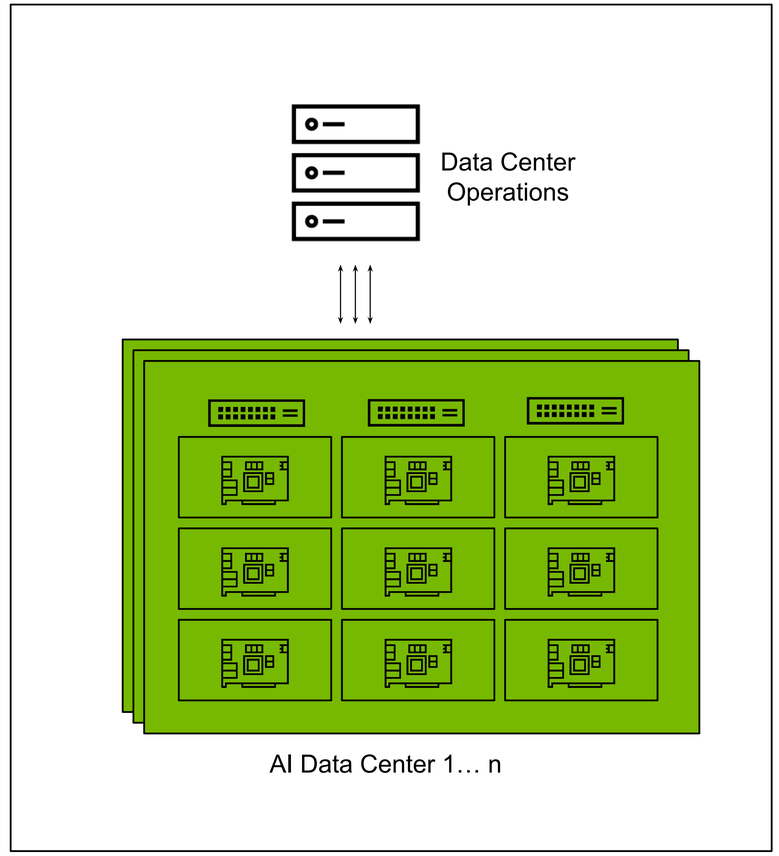

### Rapid Provisioning

Speed-up operations from bring-up to production




### Elastic, Fungible Capacity

Dynamically repurpose and allocate resources to users




### Limitless Scaling

Scale with confidence through a robust, fault-tolerant design

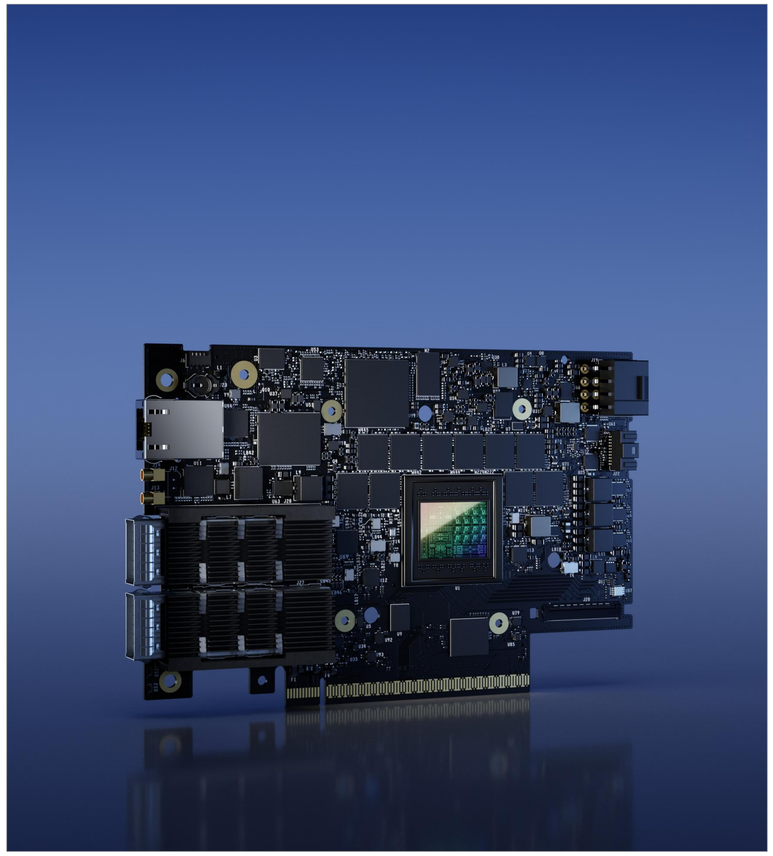
# Securely Deploy and Operate AI Data Centers

## Powered by NVIDIA BlueField

### Elastic GPU Computing
Rapid 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

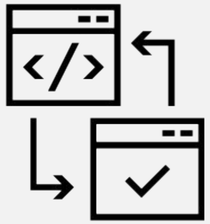
NVIDIA HGX H100 GPU / MGX Grace Hopper
The world's most advanced enterprise AI infrastructure

# Navigating Security Risks in Modern AI Data Centers

External attackers aren’t the only threats organizations need to consider in their cybersecurity planning

Insider threat incidents have risen 44% from 2020, with average cost of $15M per incident

Source: Ponemon



Costs of software supply chain attacks could exceed $46B this year, and almost $81B by 2026

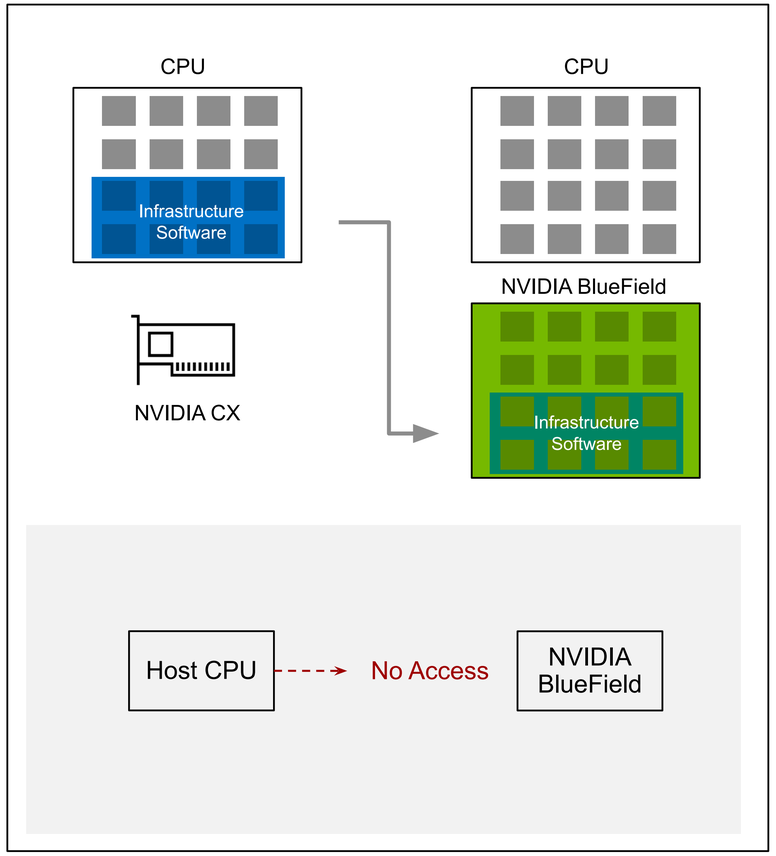
Source: Juniper

The global average cost of a data breach increased to $4.35M in 2022, the highest in last 17 years

Source: IBM

# NVIDIA BlueField Creates Zero-Trust AI Data Centers

## Isolate the data center control-plane from application workloads

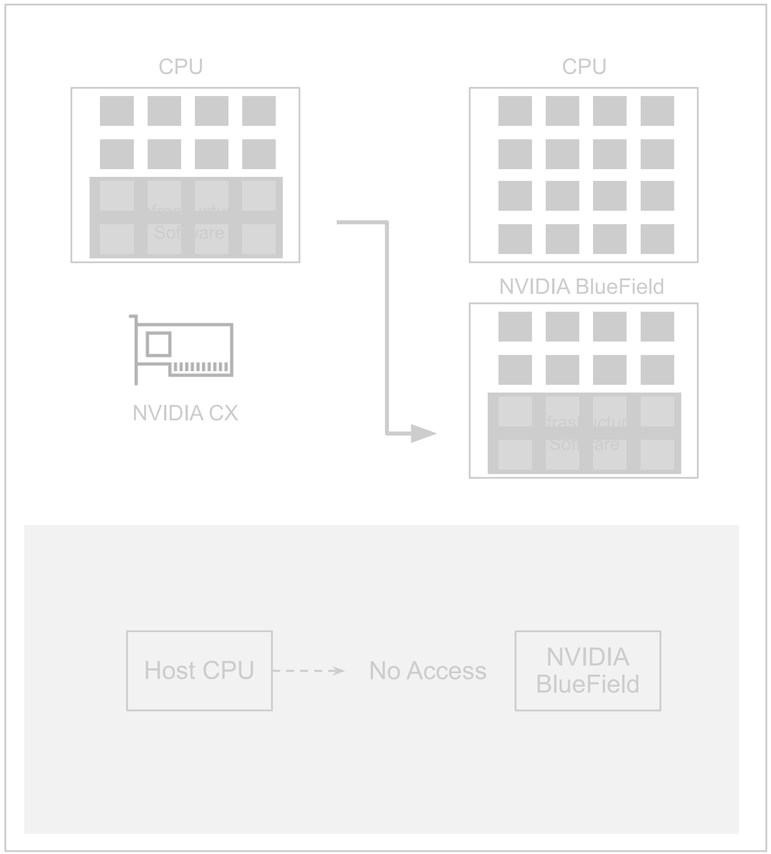



Zero-Trust Architecture

Host is untrusted, cannot access BlueField

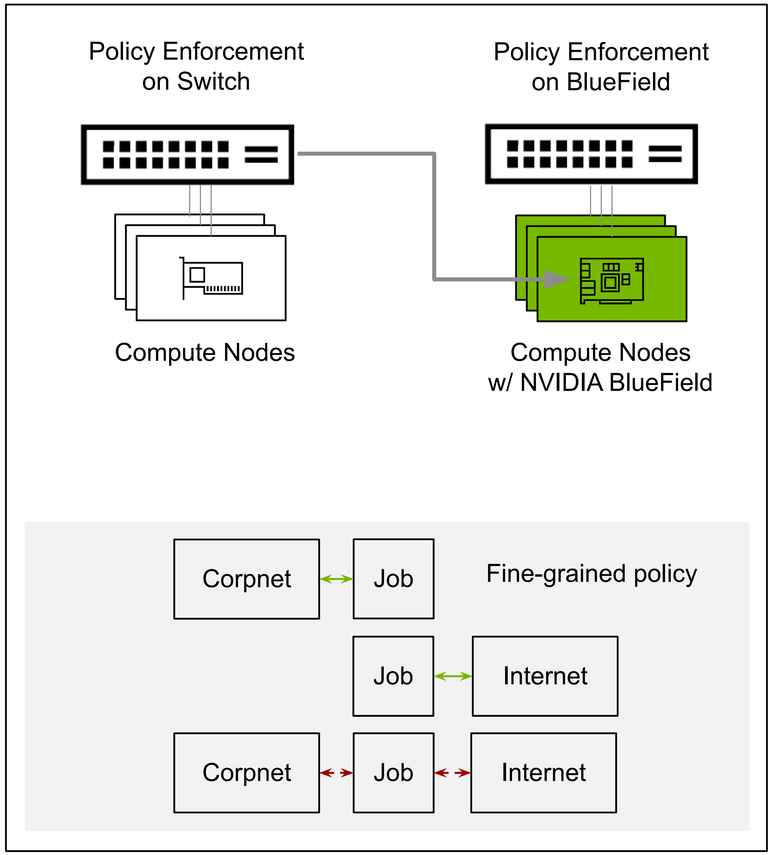
# NVIDIA BlueField Creates Zero-Trust AI Data Centers

Enforce a resilient, fine-grained security policy down to every node




Zero-Trust Architecture

Host is untrusted, cannot access BlueField

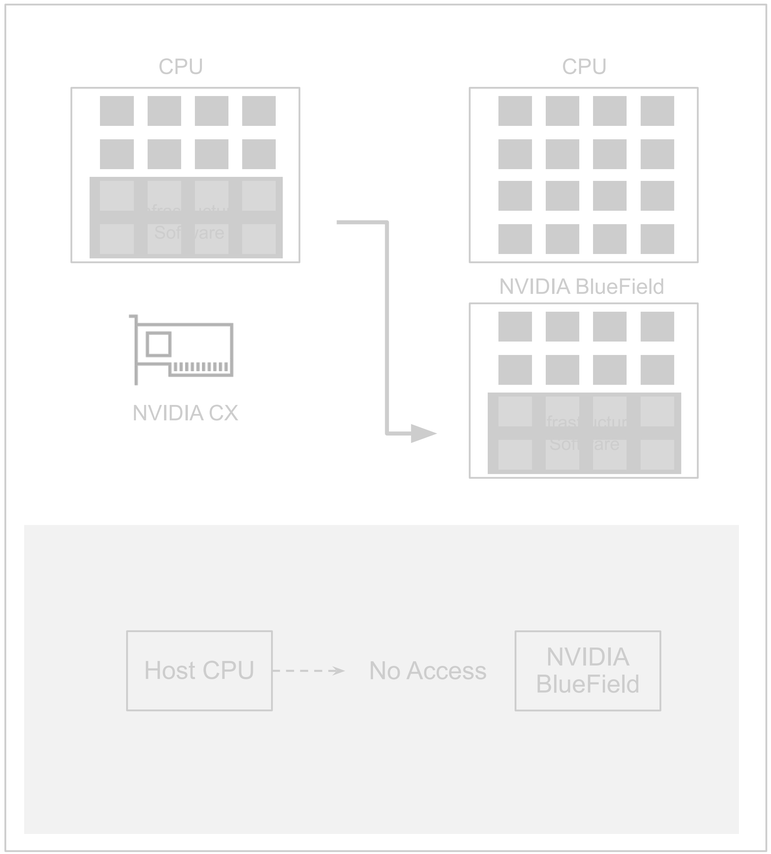



Distributed, Fine-Grained Security

Policy enforcement on BlueField

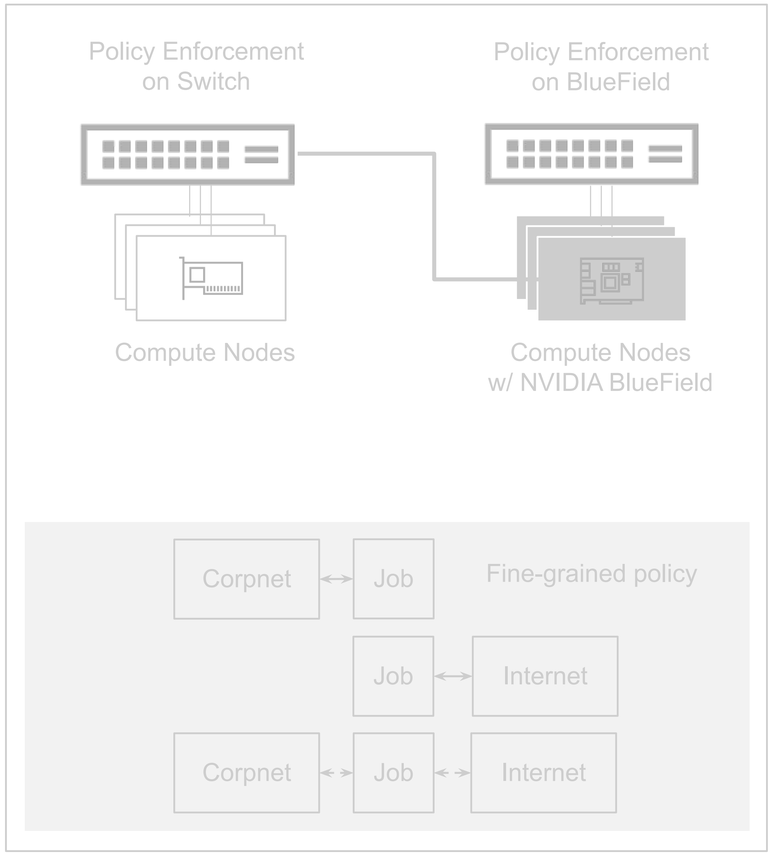
# NVIDIA BlueField Creates Zero-Trust AI Data Centers

## Enhance data security posture with another layer of protection



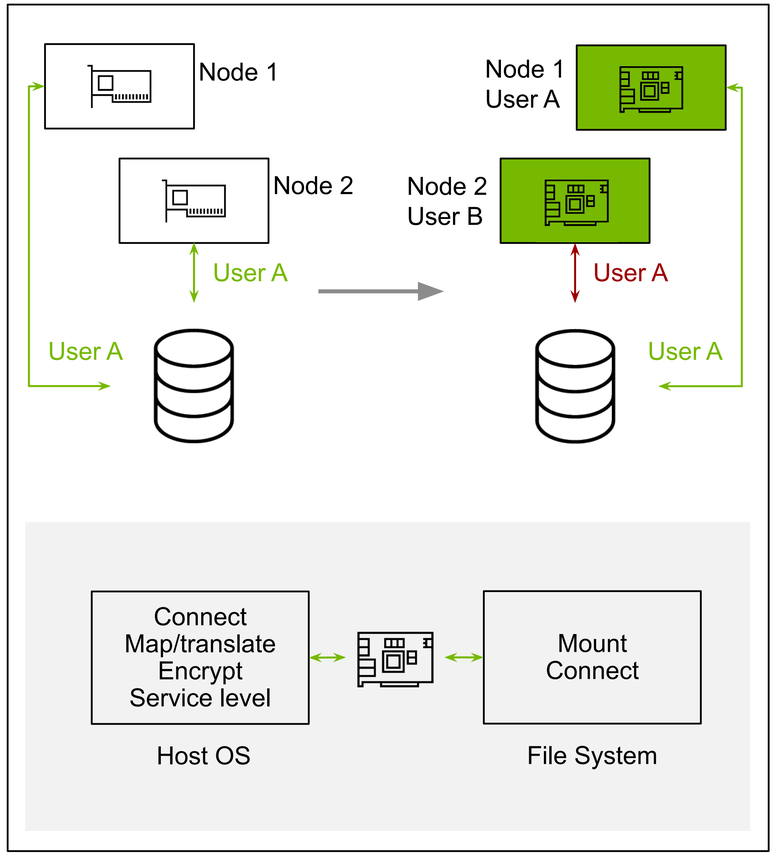

### Zero-Trust Architecture

Host is untrusted, cannot access BlueField




### Distributed, Fine-Grained Security

Policy enforcement on BlueField

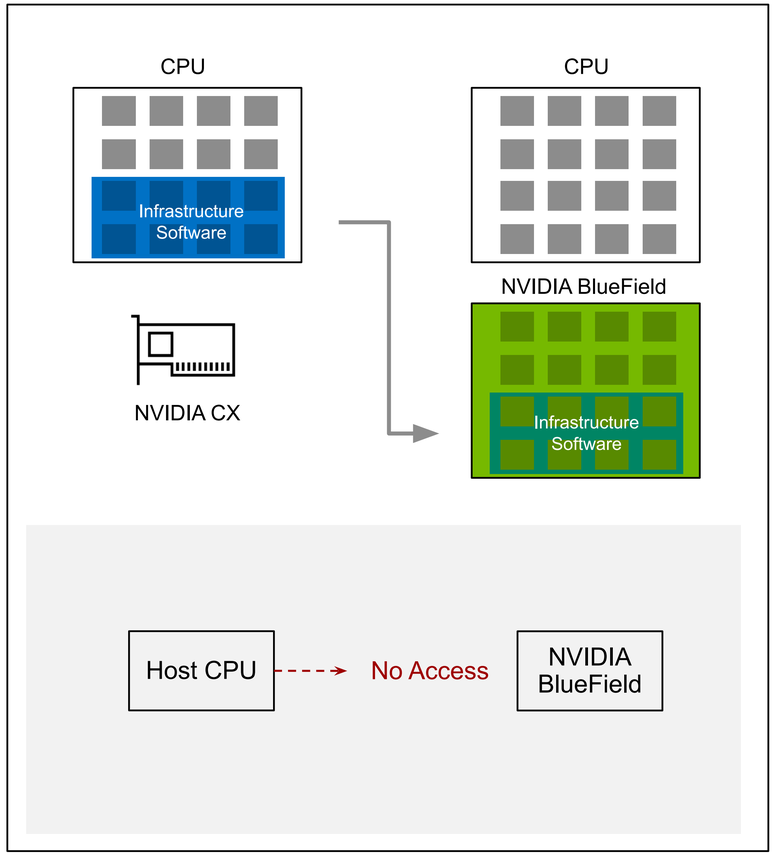



### Data Security

Move trusted software to BlueField

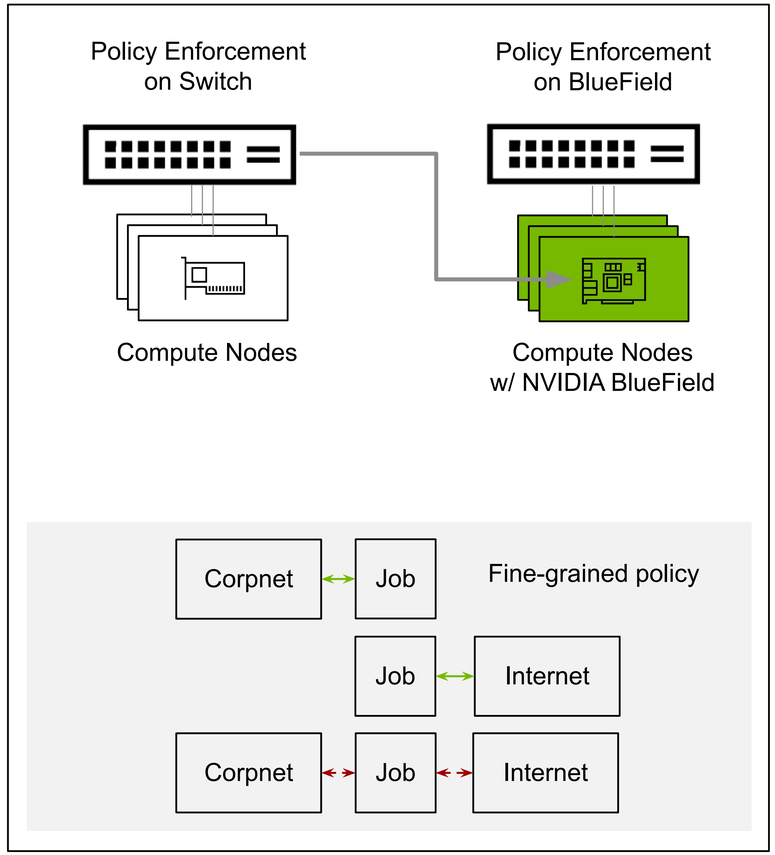
# NVIDIA BlueField Creates Zero-Trust AI Data Centers

BlueField enables zero-trust, fine-grained security from the ground up



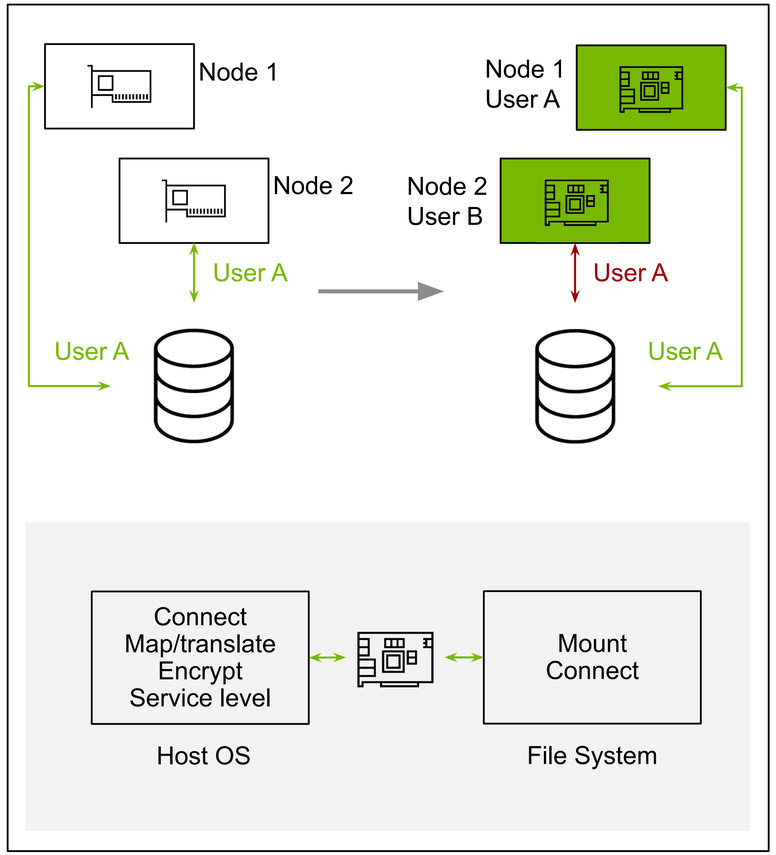

## Zero-Trust Architecture

Host is untrusted, cannot access BlueField




## Distributed, Fine-Grained Security

Policy enforcement on BlueField




## Data Security

Move trusted software to BlueField

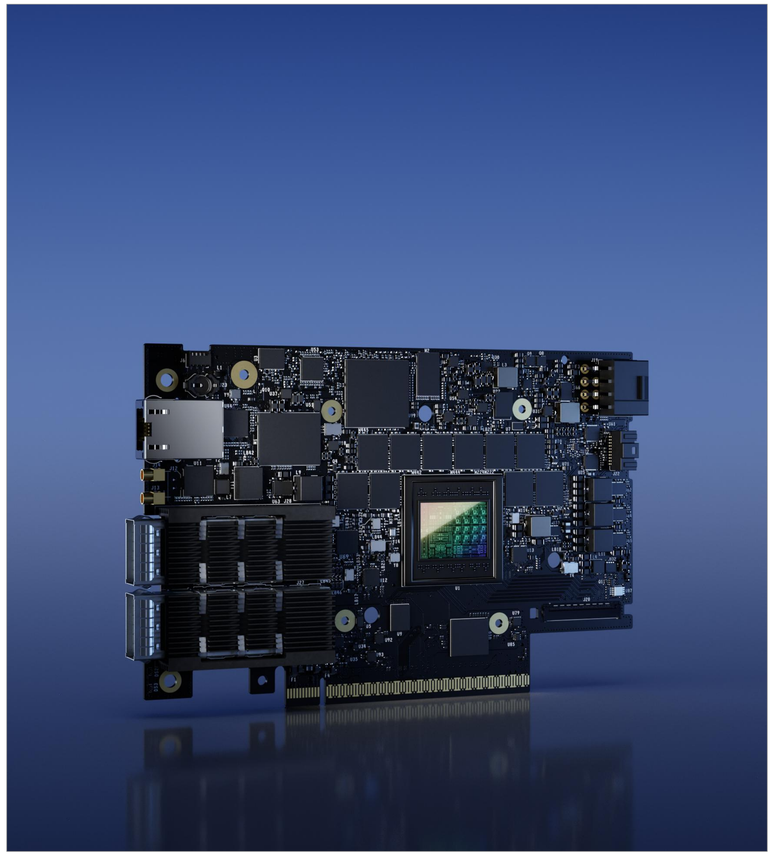
# Securely Deploy and Operate AI Data Centers

## Powered by NVIDIA BlueField

### Elastic GPU Computing
Rapid 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 data management services for AI workloads



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

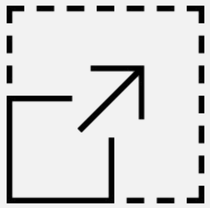
NVIDIA HGX H100 GPU / MGX Grace Hopper
The world's most advanced enterprise AI infrastructure

# Tackling Data Complexities in AI Data Centers

## Traditional storage technologies not equipped to support Gen AI and LLM training

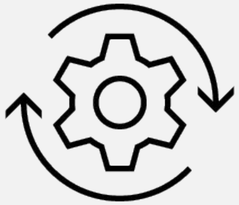
### Inadequate Data Performance

Software-defined storage (SDS) can be a bottleneck for AI workloads



### Limited Scaling

Local storage is resource bound and doesn’t scale effectively

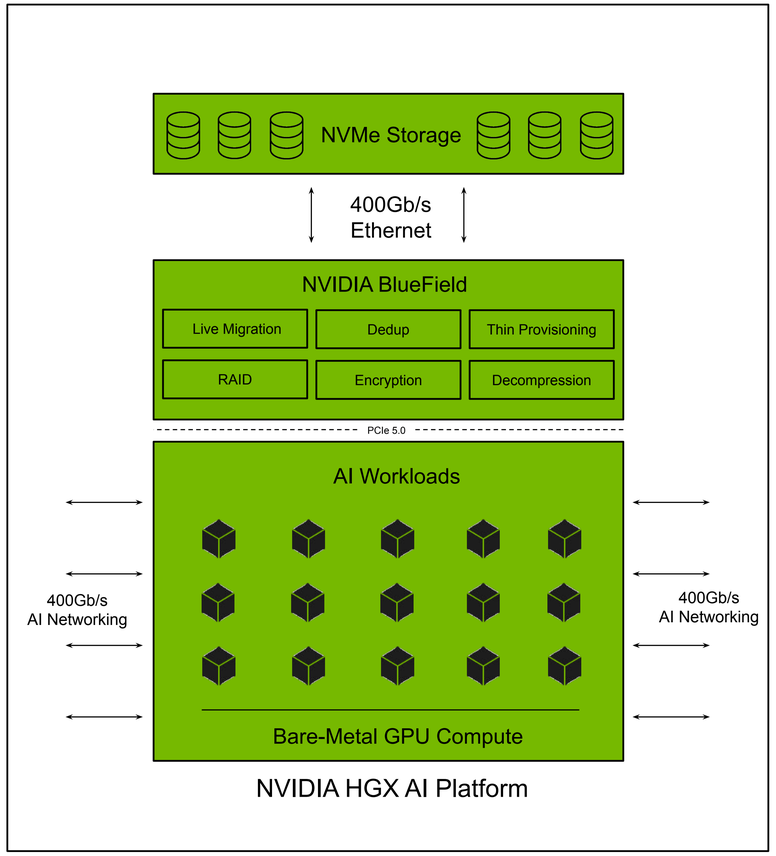


### Harder to Manage and Protect

Managing and protecting local storage is cumbersome

# NVIDIA BlueField Streamlines Data Management for AI

## Accelerate GPU compute access to cloud data store with performance exceeding 10 million IOPs

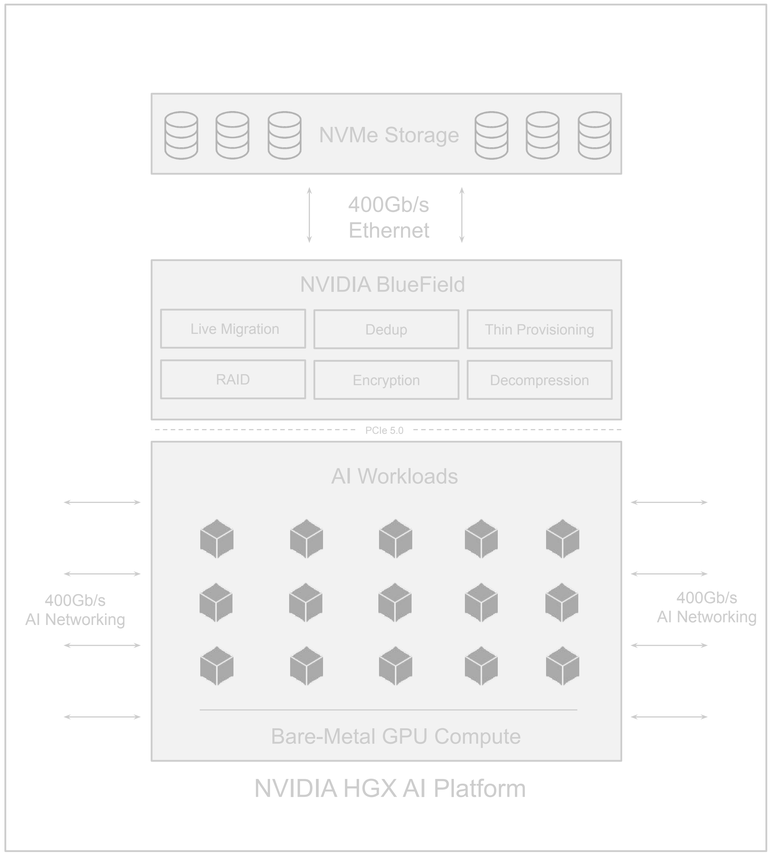



### Cloud Storage Acceleration

Software-defined, composable storage with performance higher than local storage

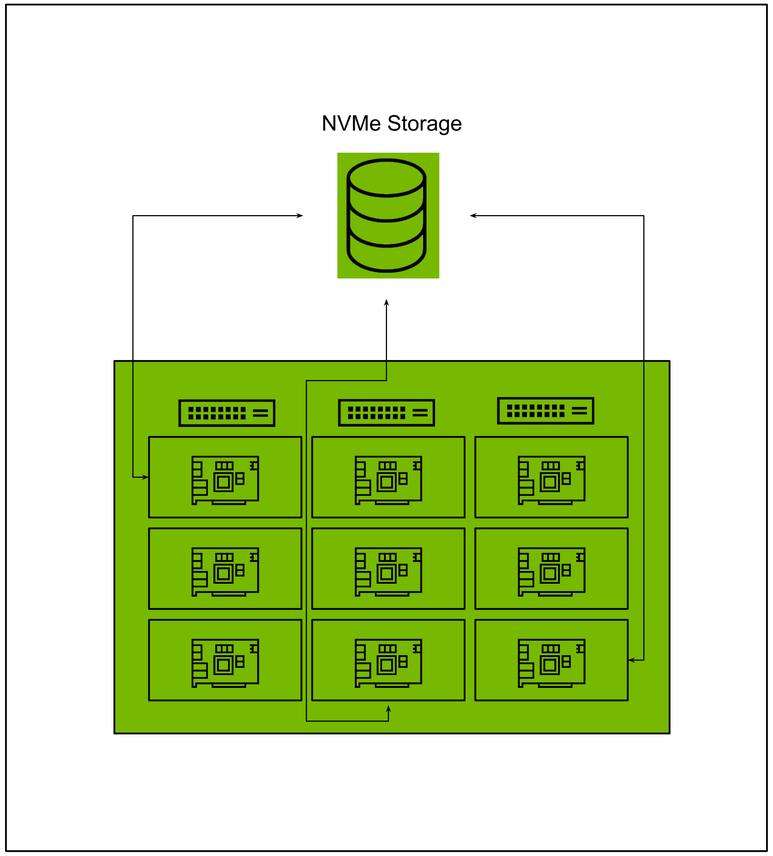
# NVIDIA BlueField Streamlines Data Management for AI

## Unlock limitless scalability and operational flexibility for your cloud data store




### Cloud Storage Acceleration

Software-defined, composable storage with performance higher than local storage

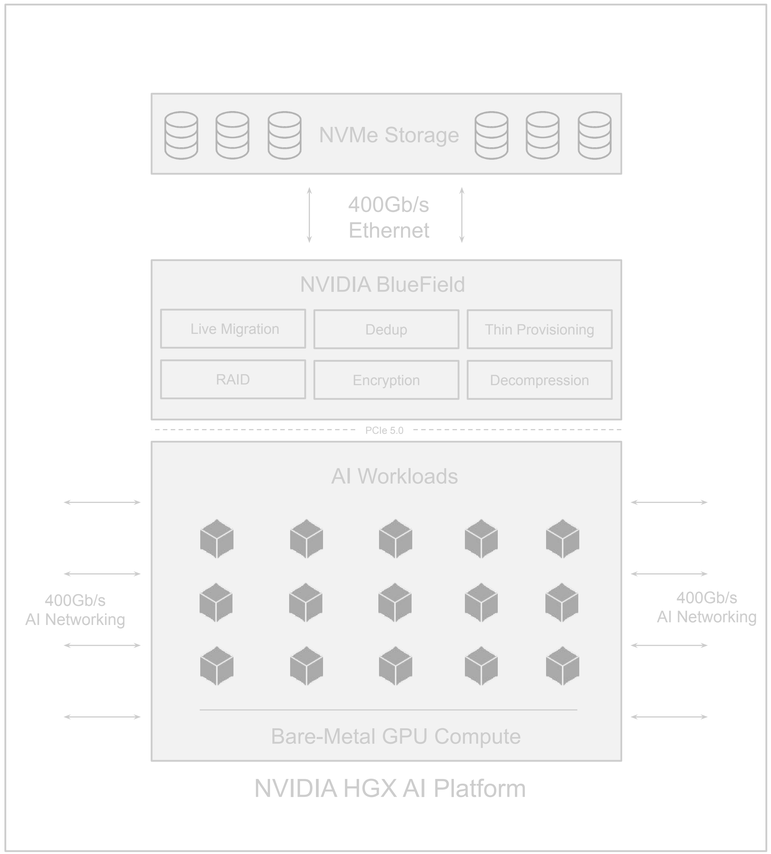



### Efficient Data Operations

Attach volumes to workloads in seconds

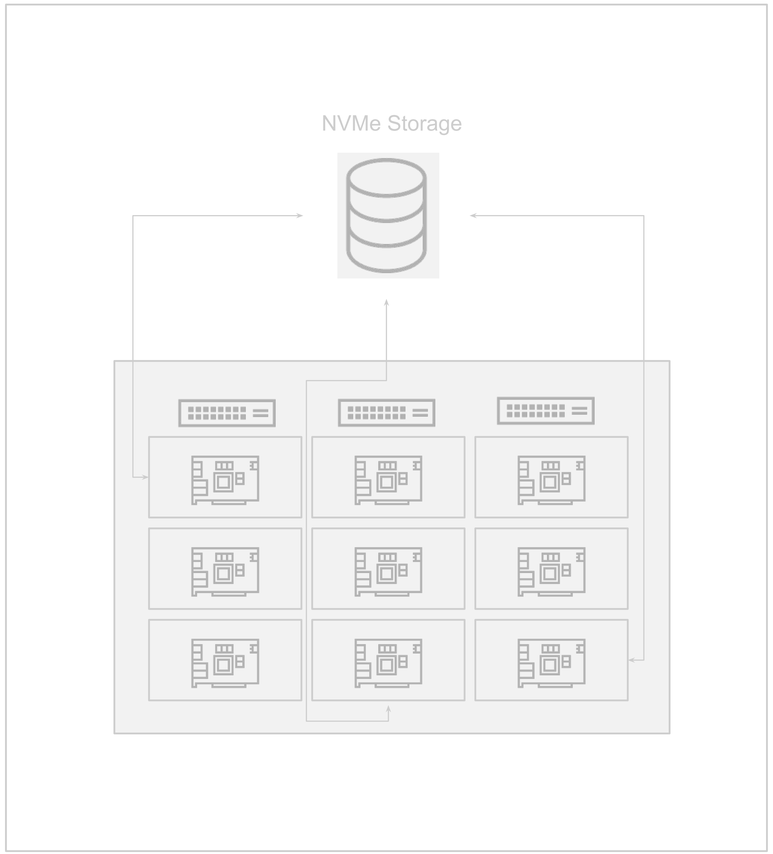
# NVIDIA BlueField Streamlines Data Management for AI

## Enhance data security, integrity, and protection



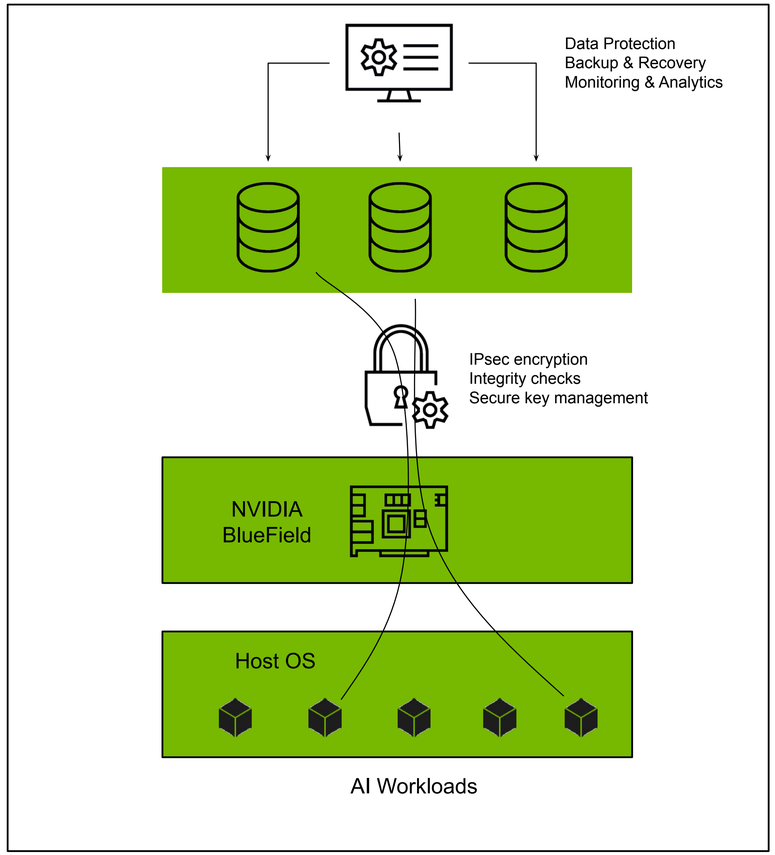

### Cloud Storage Acceleration

Software-defined, composable storage with performance higher than local storage




### Efficient Data Operations

Attach volumes to workloads in seconds

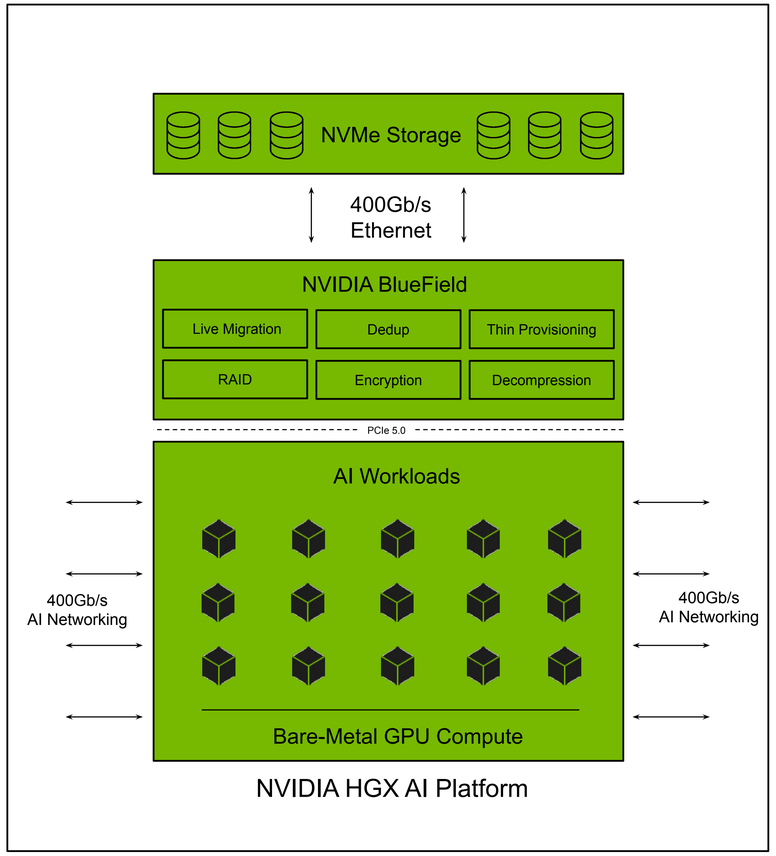



### Secure Data Fabric

Robust, zero-trust security to protect AI data from various threats

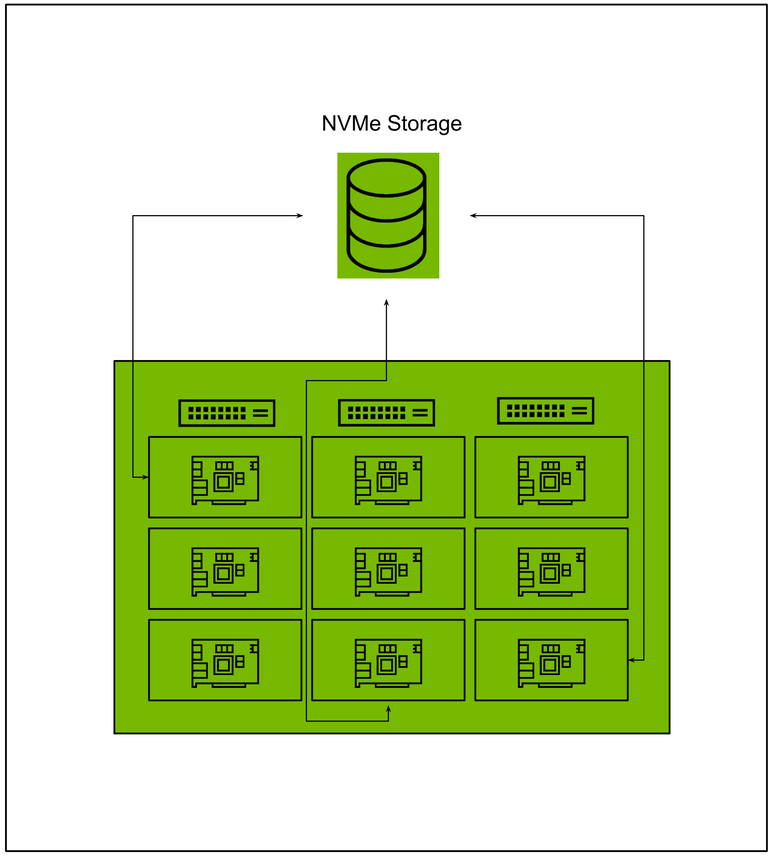
# NVIDIA BlueField Streamlines Data Management for AI

## BlueField enables high-performance and secure data platform for AI data centers



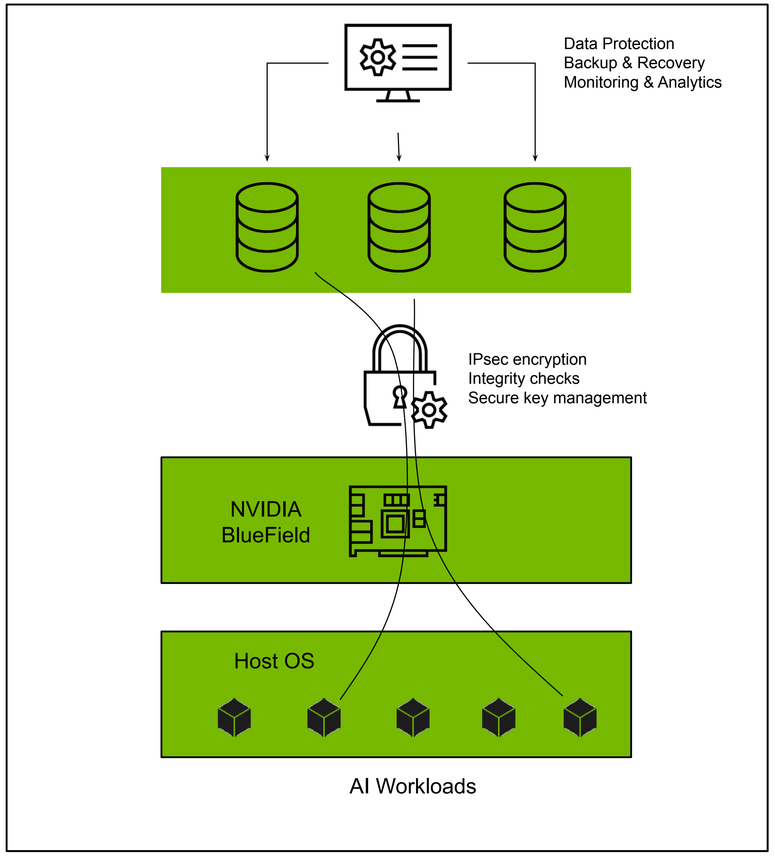

### Cloud Storage Acceleration

Software-defined, composable storage with performance higher than local storage




### Efficient Data Operations

Attach volumes to workloads in seconds




### Secure Data Fabric

Robust, zero-trust security to protect AI data from various threats

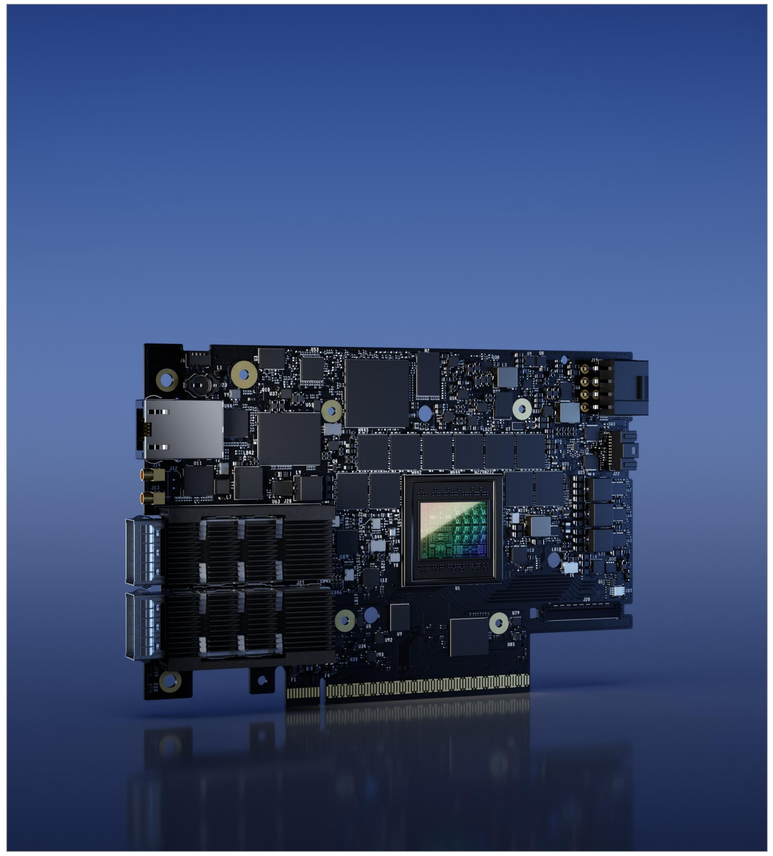
# Securely Deploy and Operate AI Data Centers

## Powered by NVIDIA BlueField

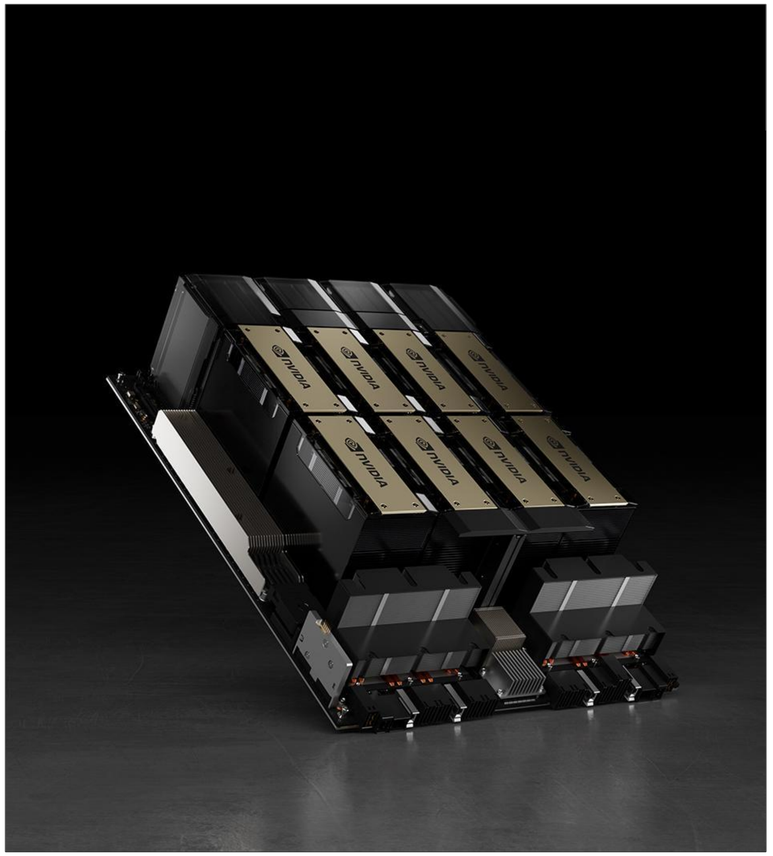
### Elastic GPU Computing
Rapid provisioning, fungible GPU compute and limitless scaling

### 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



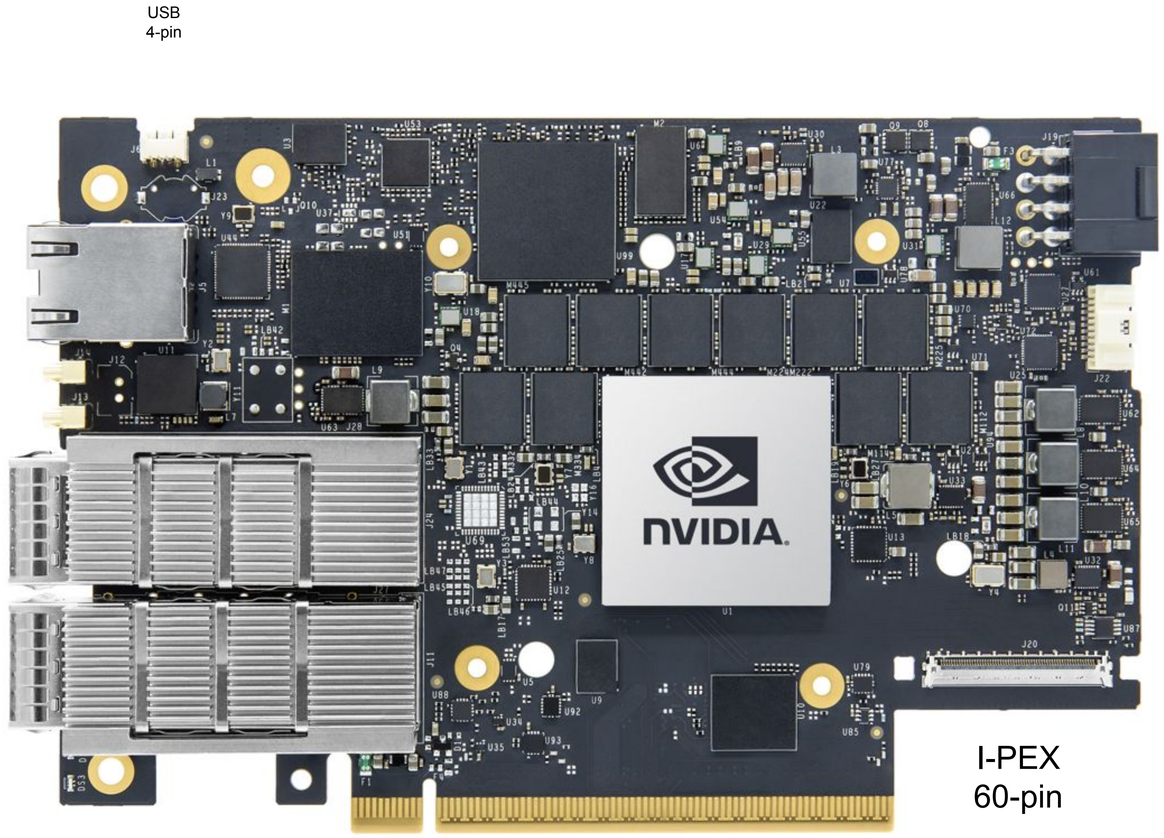
NVIDIA BlueField-3 DPU
400Gb/s Infrastructure compute platform



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

# NVIDIA B3220 Platform for HGX N-S

| Model | B3220 |
| --- | --- |
| Network Speed | 2 x 200Gb/s |
| Arm Cores | 16 x  ArmA78 @2.2GHz |
| Host Interfaces | Gen5 x16 + x16 |
| Memory | 32GB DDR5 |



# NVIDIA DOCA Stack

## Comprehensive Acceleration SDK, Compilers, Services, and Tools

Containerized services

**Applications**

Networking | Security | Storage | HPC/AI | Telco | Media

**DOCA**

**DOCA Services**

- Orchestration
- Telemetry
- Firefly
- SDN/P4 RT/HBN
- DPU Management

**DOCA Libs**

| | | | | |
|---|---|---|---|---|
| Crypto | App Shield | Rivermax | HPC/AI | GPUNetIO |
| Flow<br>Gateway, Firewall | DPI | Storage | Comm Channel | DPA |

**DOCA Drivers**

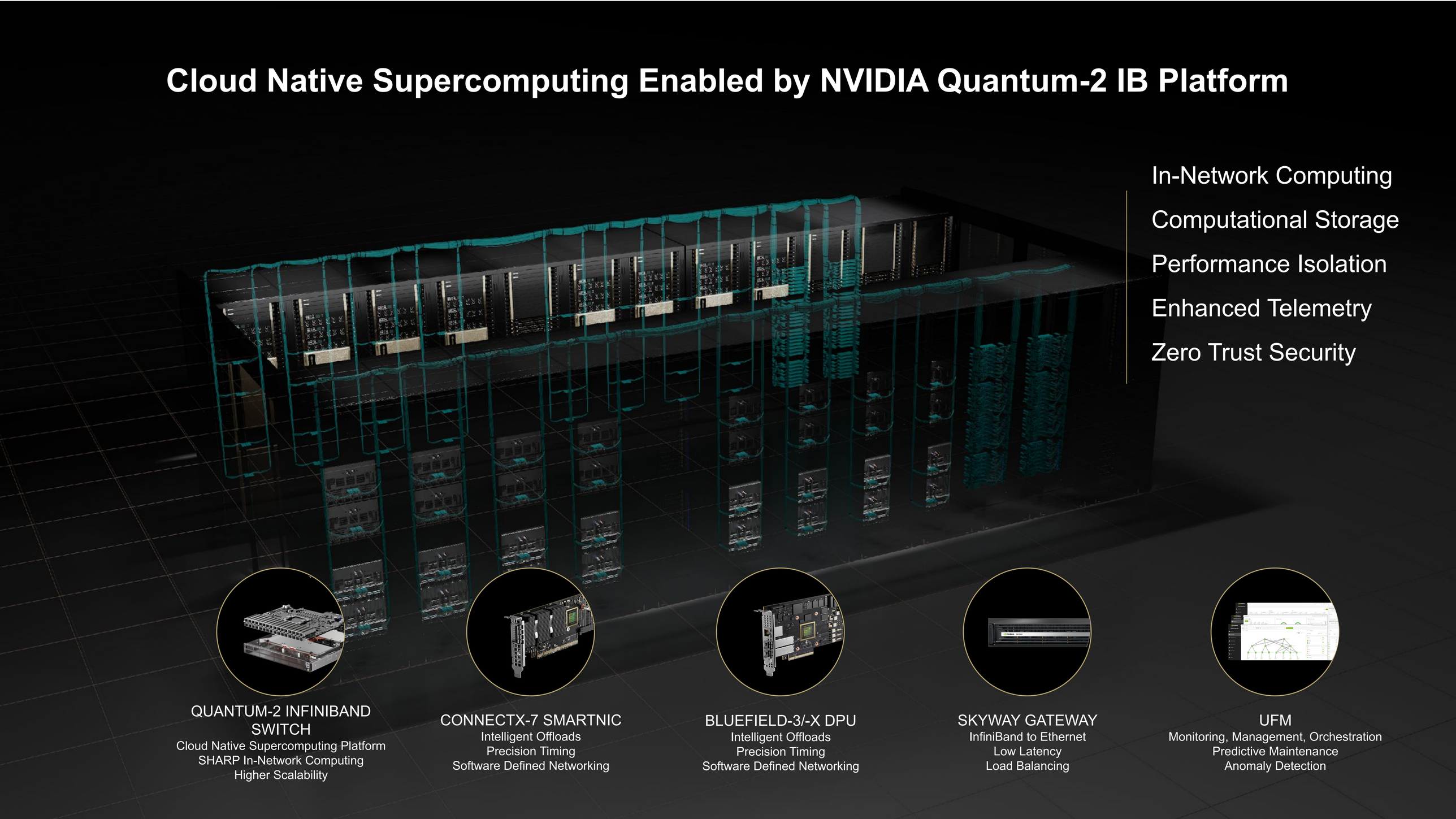
| | | | | |
|---|---|---|---|---|
| Networking<br>ASAP$^2$<br>DPDK<br>XLIO | Security<br>DPDK RegEx<br>DPDK SFT<br>Inline Crypto | Storage<br>SPDK SNAP<br>Erasure Coding<br>XTS Crypto | UCX<br>RDMA | UCC<br>FlexIO |

Secure, performance accelerated

Open foundation


NVIDIA BlueField-3 DPU

# Cloud Native Supercomputing Enabled by NVIDIA Quantum-2 IB Platform



In-Network Computing

Computational Storage

Performance Isolation

Enhanced Telemetry

Zero Trust Security



QUANTUM-2 INFINIBAND SWITCH

Cloud Native Supercomputing Platform
SHARP In-Network Computing
Higher Scalability



CONNECTX-7 SMARTNIC

Intelligent Offloads
Precision Timing
Software Defined Networking



BLUEFIELD-3/-X DPU

Intelligent Offloads
Precision Timing
Software Defined Networking



SKYWAY GATEWAY

InfiniBand to Ethernet
Low Latency
Load Balancing



UFM

Monitoring, Management, Orchestration
Predictive Maintenance
Anomaly Detection

# Thank You

# HGX H100 and BlueField-3 for Gen AI

## Sales Kit

| | |
|---|---|
| 1 | BlueField Powers NVIDIA-Accelerated AI Systems (Genius Hub)<br>• Positions BlueField-3 for HGX N-S<br>• CSPs, OEM/ODM, multi-tenant CRISPs |
| 2 | NVIDIA Spectrum-X and HGX H100 Accelerate AI Clouds (Genius Hub)<br>• Positions Spectrum-X with BlueField-3 for HGX E-W (Ethernet)<br>• CSPs, OEM/ODM, multi-tenant CRISPs |
| 3 | Securely Deploy and Operate HGX AI Data Centers (Genius Hub)<br>• Positions BlueField-3 for HGX N-S in single-tenant environments<br>• Single-tenant CRISPs/Enterprise |

NVIDIA

# Securely Deploy and Operate HGX AI Data Centers

## Presentation Outline

- Strategy:
  - Present data center challenges and discuss how BlueField can help address them
  - Lead with AI data center vs. cloud.
  - Position Forge-like capabilities without talking of Forge
- Key problem statements and BlueField value props:
  - DC Operations Problem (1): Organizations struggle to operationalize generative AI
  - Value Prop (1): NVIDIA BlueField accelerates time-to-market for generative AI

---

  - Security Problem (2): Navigating security risks in modern AI data centers
  - Security Value Prop (2): NVIDIA BlueField Creates Zero-Trust AI Data Centers

---

  - Data Problem (3): Tackling Data Complexities in AI Data Centers
  - Data Value Prop (3): NVIDIA BlueField Streamlines Data Management for AI