CISCO Engage | Tech Day

Accelerate Delivery of Trusted AI Apps

Building Your Cisco Secure AI Factory

Neython Lec Streitz
Solution Engineer

Michael Duarte
AI Solutions Engineer

March 18th, 2026

AI will make our world of 8B people feel like one with the capacity of 80B

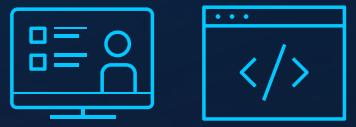
– Jeetu Patel

cisco

AI use cases across industries

Knowledgebase copilots

AI assistants



Content and code generation

Text | Images | Video | Code

Virtual agent and chatbots

Specialized domain | Specific chatbots

Visual Computing

Digital Twins | Video Analytics | Imaging and Diagnostics

Language translation

Multilingual real-time communication

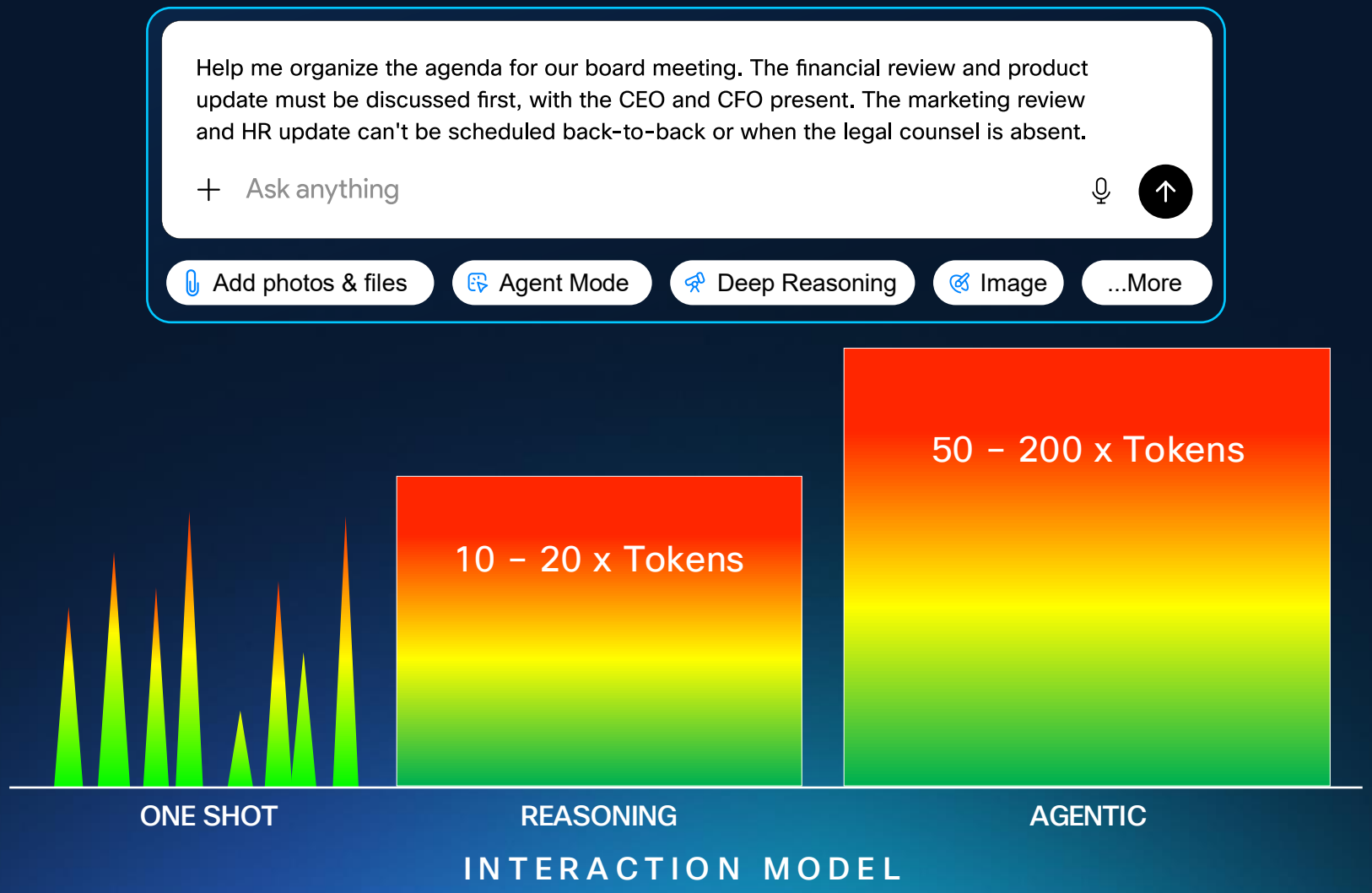
Detection and prediction

Forecasts | Anomalies | Insights

cisco

AI is Changing: Token Demand Inflation

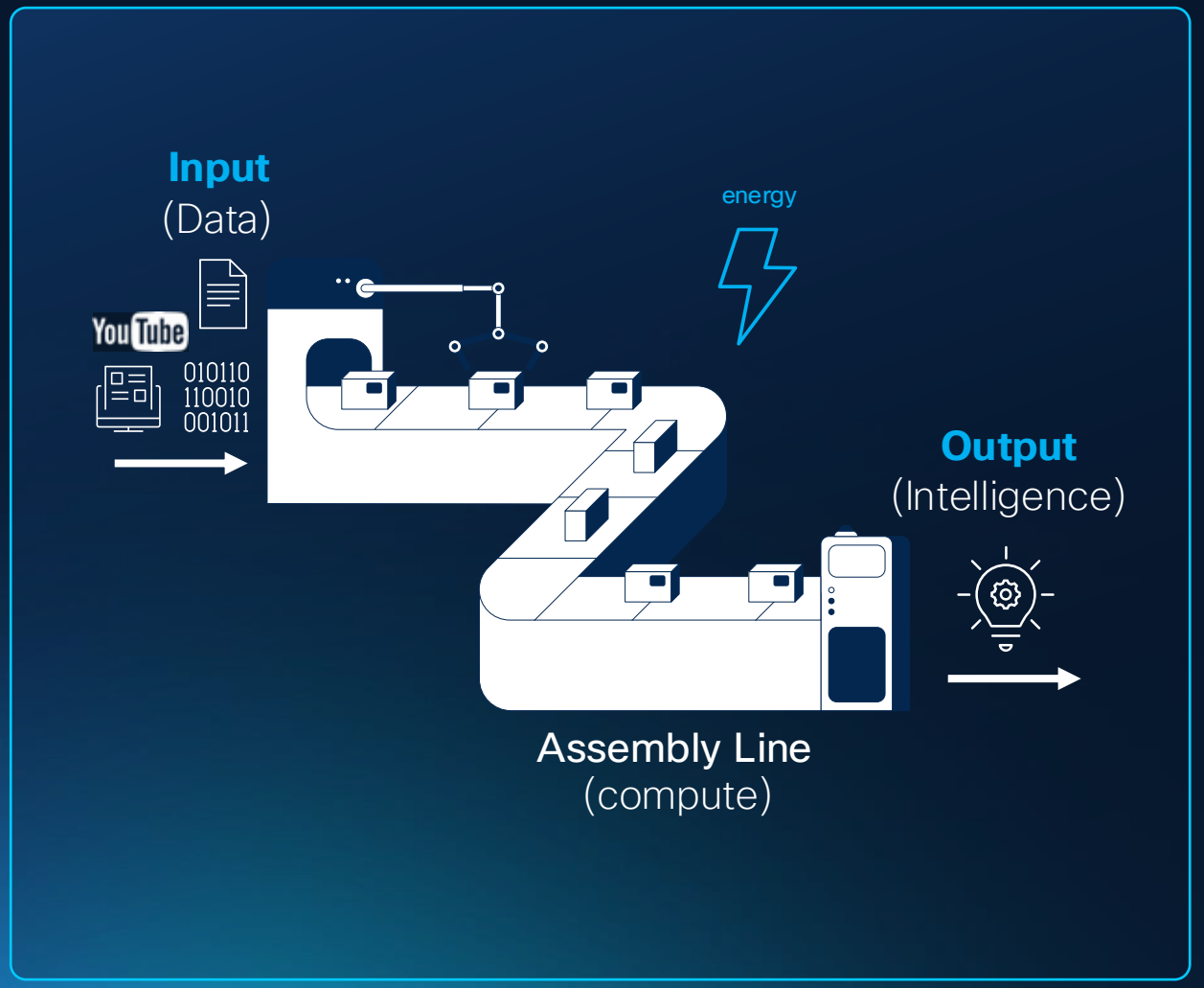
More tokens enable higher quality results and more complex tasks



What is an AI Factory?

The processing plant for tokens

Organizations everywhere are thinking about how to generate tokens as quickly, safely and cost effectively as possible.



All of this exposes key challenges for our customers' technology architectures

Infrastructure | Security | Observability

We’re addressing all of these challenges head-on
Cisco is the critical infrastructure for the AI era

What is needed to accelerate trusted AI outcomes?

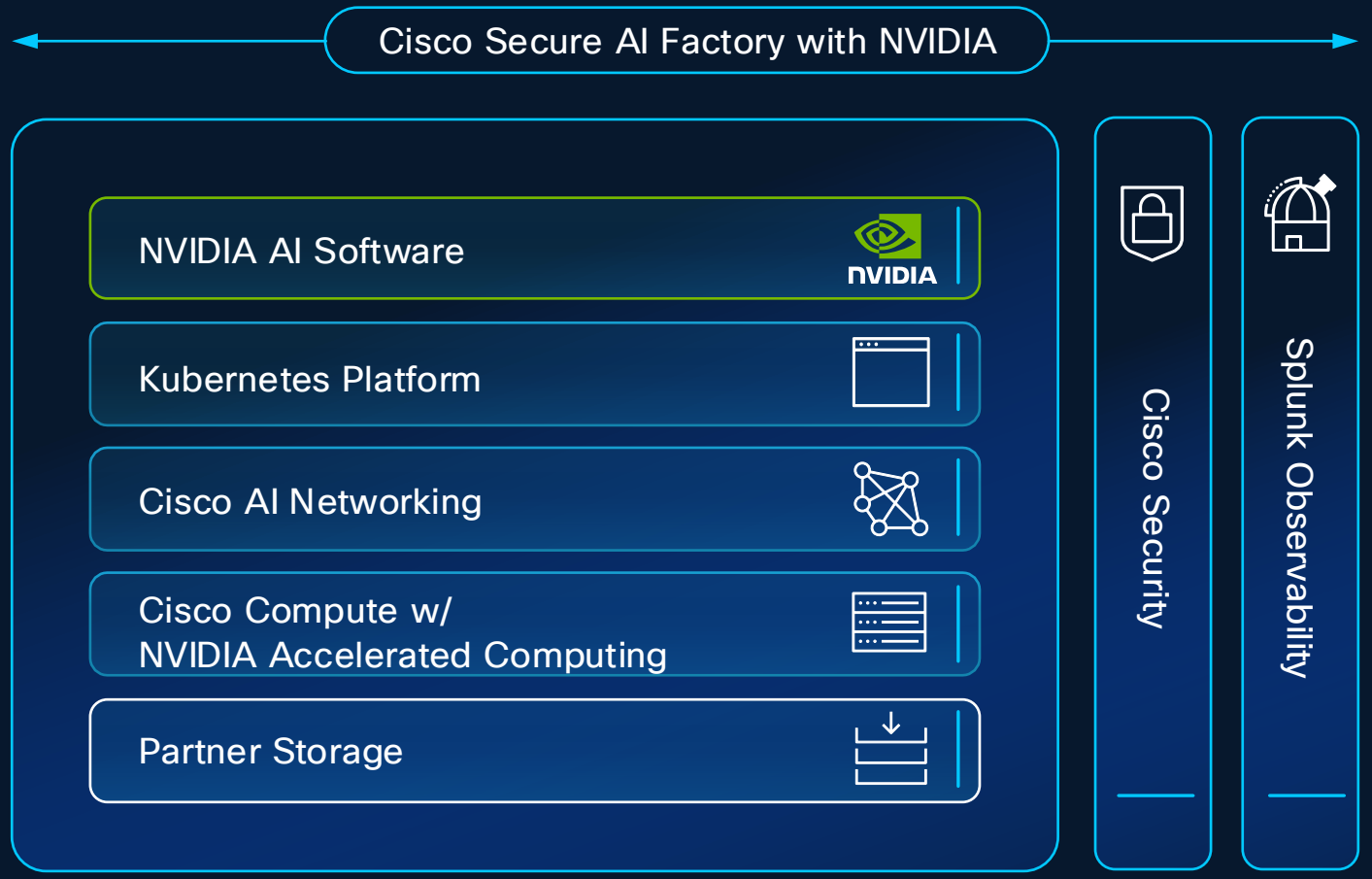
☐ AI factory

☑ **Secure** AI factory

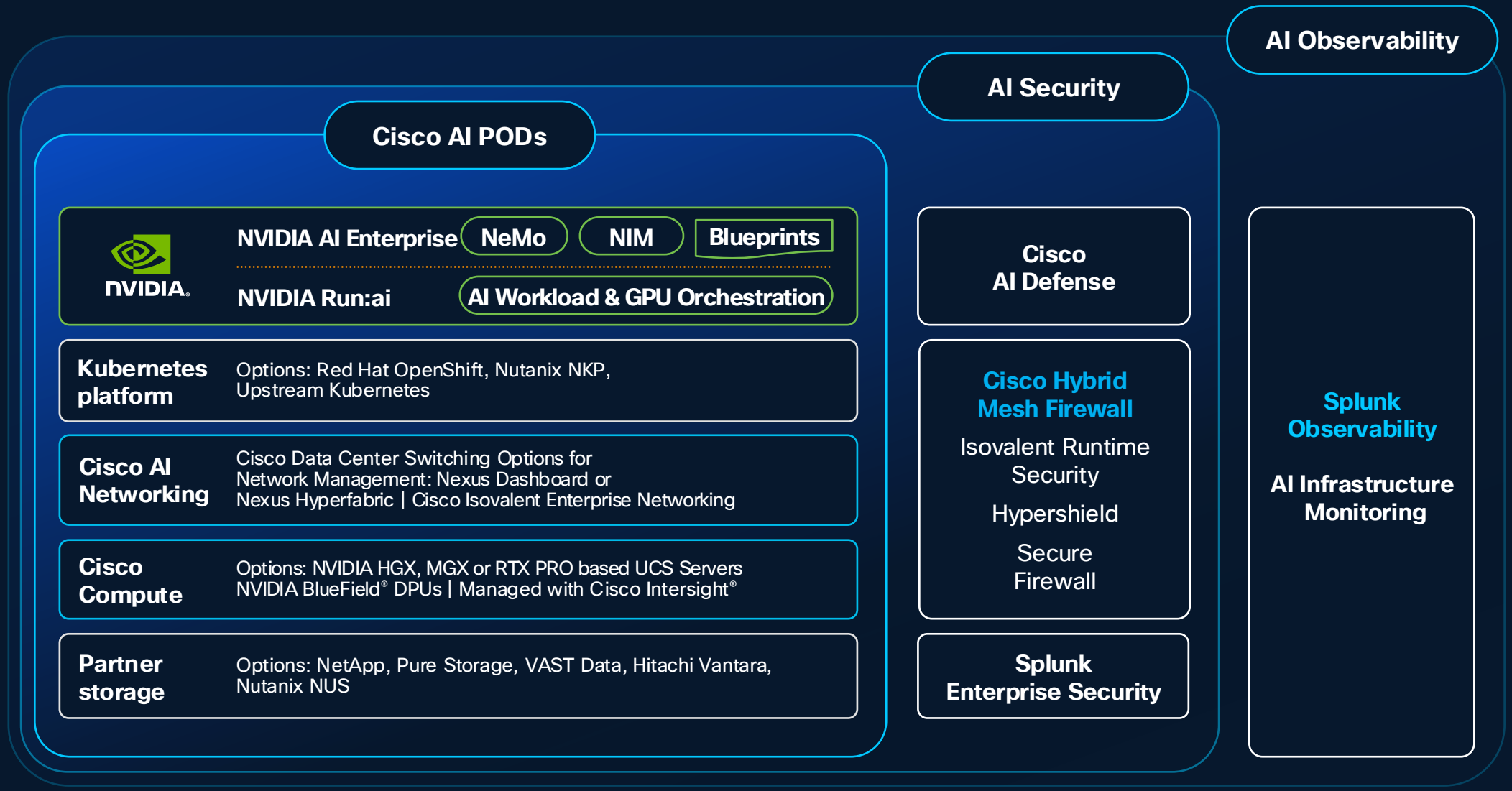
Cisco Secure AI Factory with NVIDIA

Delivering Trusted AI Outcomes

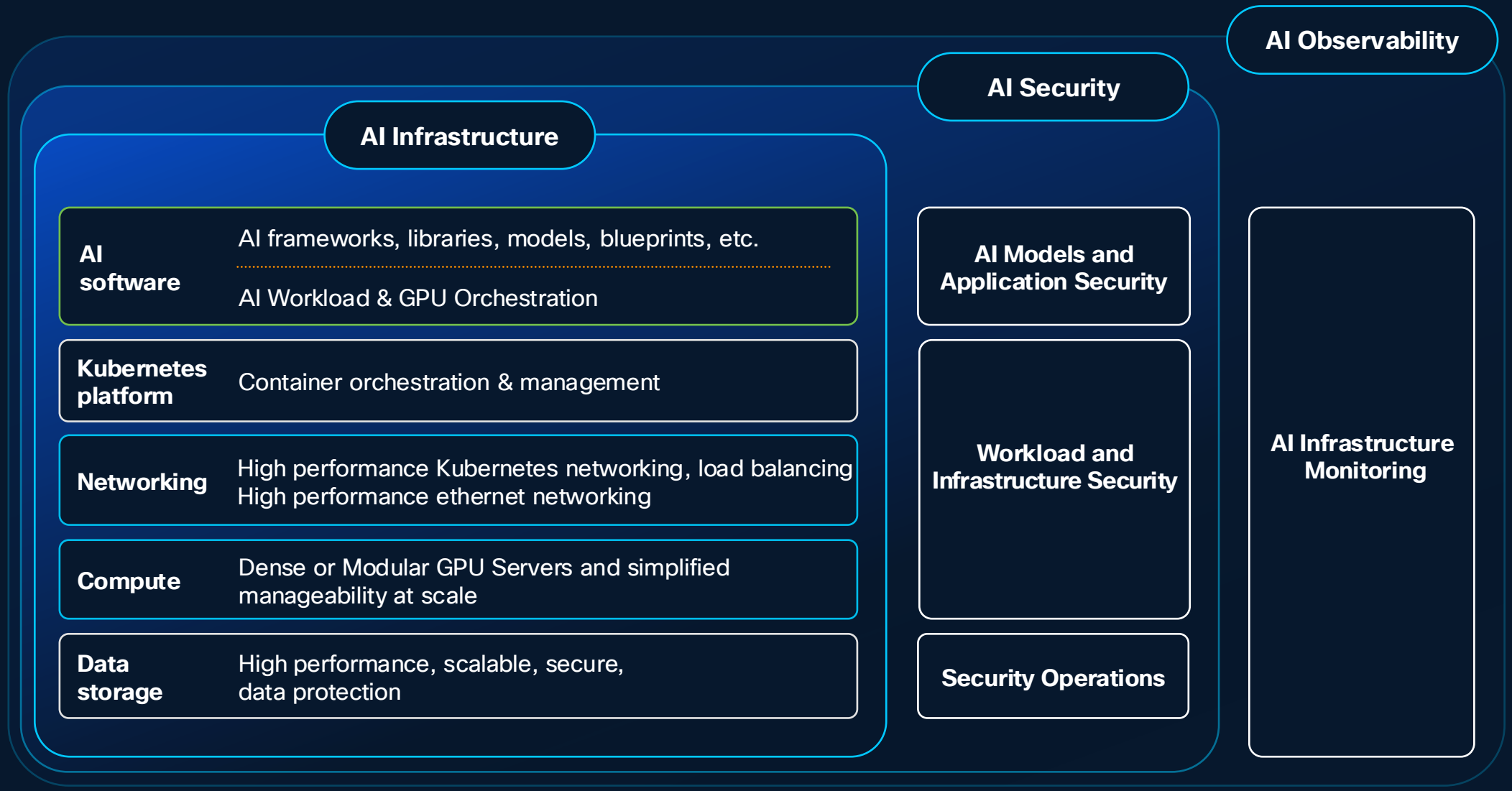
A modular reference design that combines high-performance infrastructure with full-stack security and observability



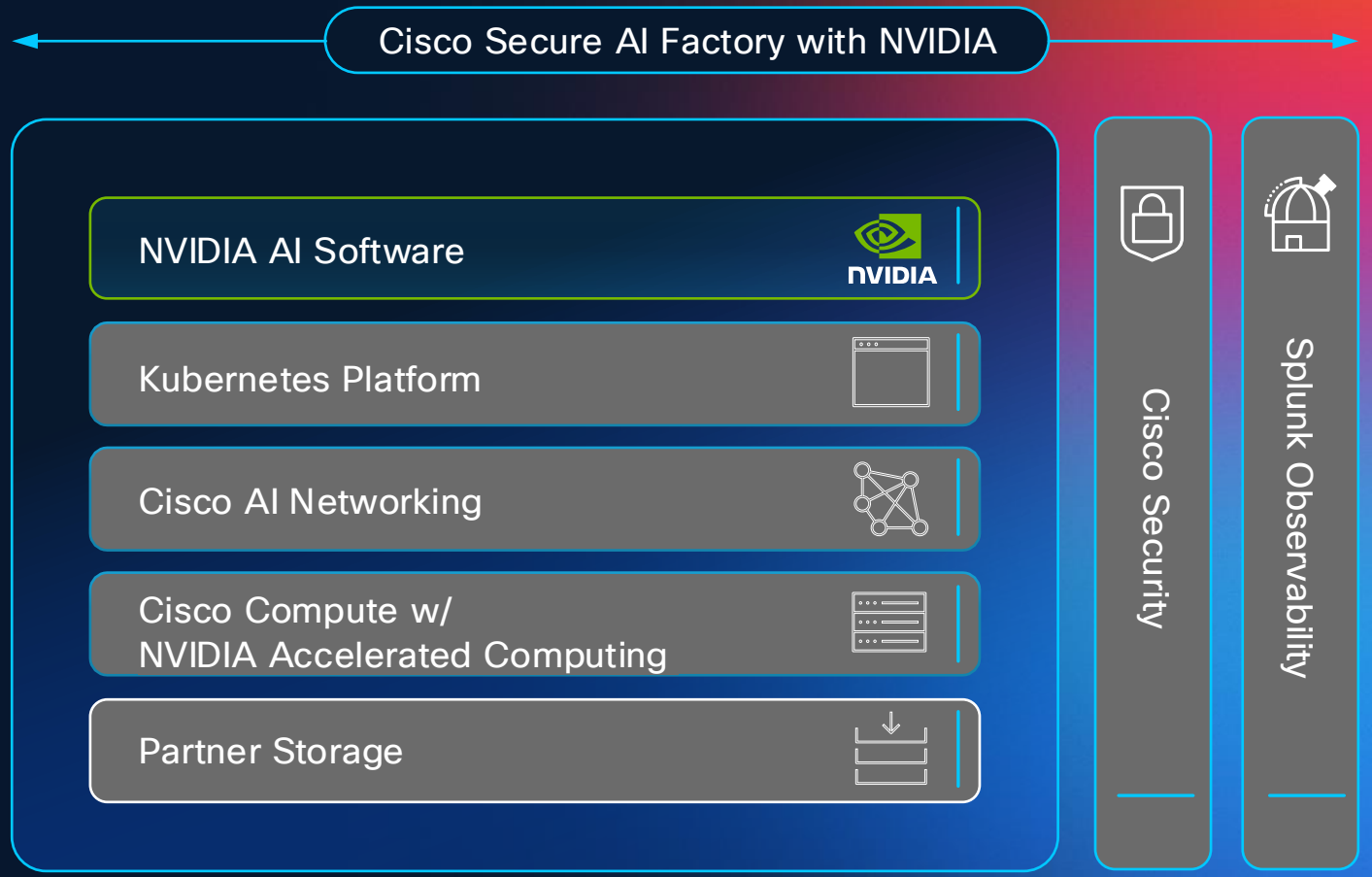
Key products in Cisco Secure AI Factory with NVIDIA



Key capabilities of Cisco Secure AI Factory with NVIDIA



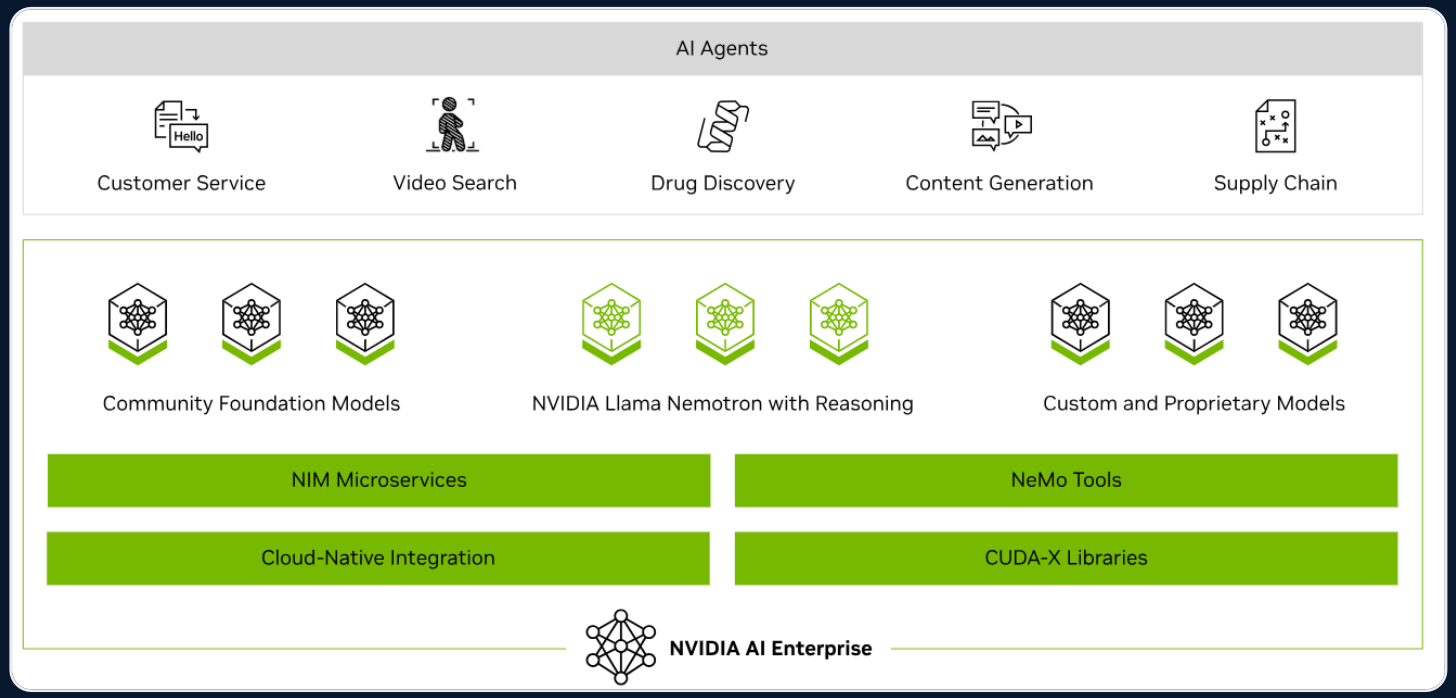
Secure AI Factory with NVIDIA, Software



NVIDIA Enterprise Software

The NVIDIA Enterprise tools in the Cisco Secure AI Factory with NVIDIA provide support for each step in the training, optimization, and deployment of AI agents.

Production-ready software for agentic AI



Deploy the latest state-of-the-art AI models

Explore the NVIDIA NIMs catalog of enterprise-ready, performance-optimized models for efficient inference and reasoning.

Build and manage data flywheels with NeMo

Discover powerful, ready-to-use model training, evaluation, and guard railing tools and RAG building blocks for optimizing agentic AI.

Customizable blueprints for your use case

Reference workflows for building fast, high-performance, and secure agentic systems using the latest machine learning best practices.

Software for AI

NVIDIA

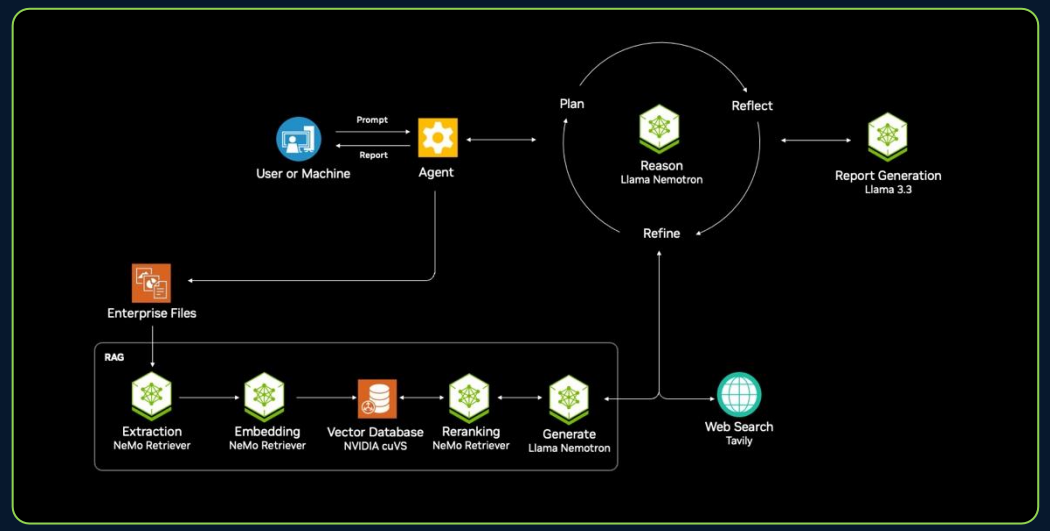
NVIDIA Enterprise	NeMo	NIM	Blueprints
NVIDIA Run:ai	AI Workload & GPU Orchestration		

cisco

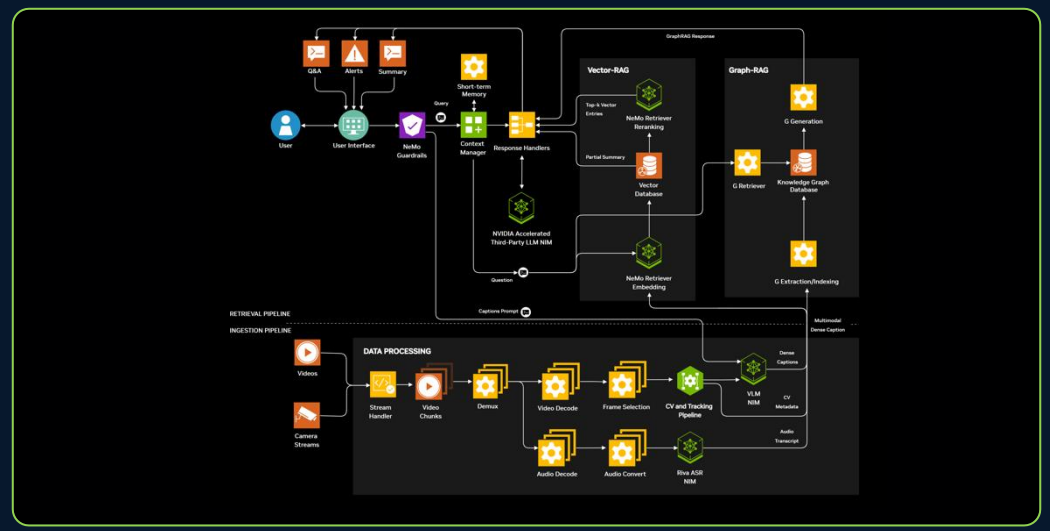
NVIDIA Enterprise – Blueprints for use cases

https://build.nvidia.com | https://catalog.ngc.nvidia.com

Research Assistant



Video Search & Summarization



Blueprints offer sample workload designs for common AI use cases. These blueprints leverage technology available in the NVIDIA Enterprise software suite. These blueprints are but a few of infinite use cases that can be developed with AI software.

Software for AI

NVIDIA

NVIDIA Enterprise — NeMo | NIM | Blueprints

NVIDIA Run:ai — AI Workload & GPU Orchestration

cisco

NVIDIA Run:ai

Software for AI

NVIDIA

NVIDIA Enterprise

NeMo

NIM

Blueprints

NVIDIA Run:ai

AI Workload & GPU Orchestration

Resource Management

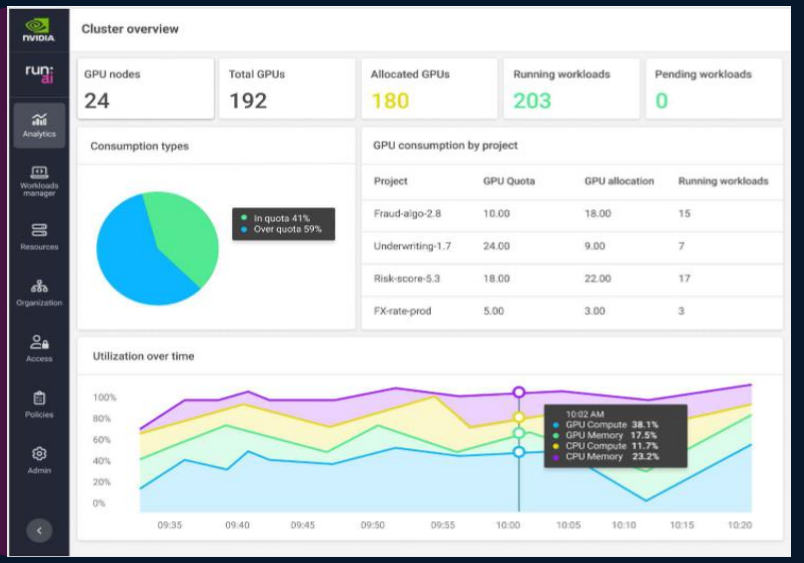
- Infrastructure Pooling
- Policy Engine

AI Lifecycle Integration

- Scheduling
- GPU Orchestration

Workload Orchestration

- Scheduling
- GPU Orchestration



AI-Native Workload Orchestration

Purpose-built for AI workloads, NVIDIA Run:ai delivers intelligent orchestration that maximizes compute efficiency and dynamically scales AI training and inference.

Flexible AI Deployment

NVIDIA Run:ai supports AI workloads wherever they need to run, whether on prem, in the cloud, or across hybrid environments, providing seamless integration with AI ecosystems.

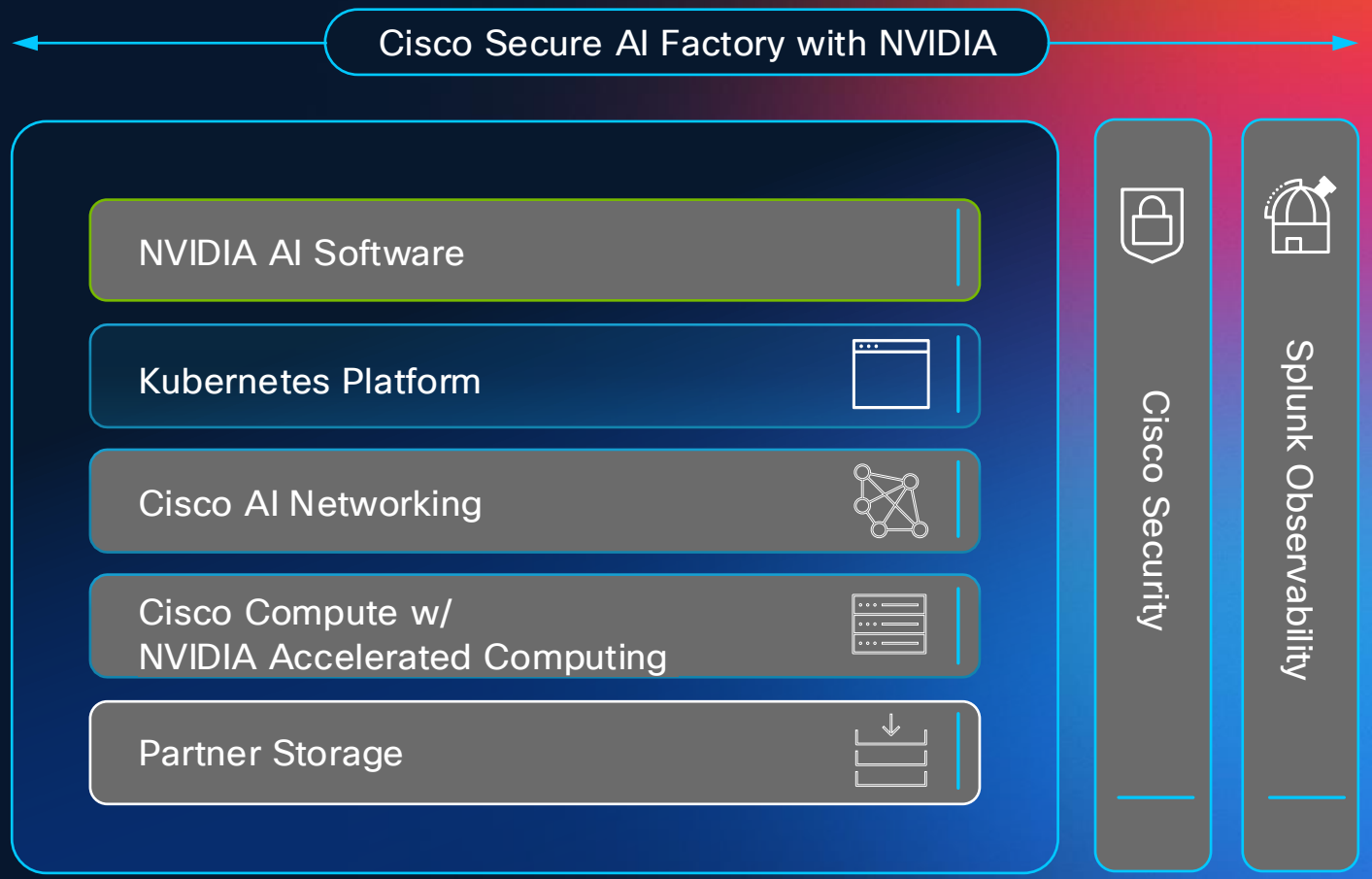
Unified AI Infrastructure Management

NVIDIA Run:ai provides a centralized approach to managing AI infrastructure, ensuring optimal workload distribution across hybrid, multi-cloud, and on-premises environments.

Open Architecture

Built with an API-first approach, NVIDIA Run:ai ensures seamless integration with all major AI frameworks, machine learning tools, and third-party solutions.

Secure AI Factory with NVIDIA, Kubernetes



Why Kubernetes? (the Orchestration layer)

Universal Portability:

- The industry standard for containers.
- Build agents once, run anywhere (On-prem, AWS, Azure, GCP, Edge) without lock-in.

Elastic Scalability:

- Native autoscaling dynamically handles the ‘bursty’, unpredictable compute demands of agentic reasoning loops.

GPU Management:

- Nvidia GPU Operator: Runs natively on Kubernetes to automate the complex deployment of GPU drivers, container toolkits, and monitoring across the cluster.

Security & Isolation:

- Granular RBAC: Strict controls agent access to APIs and Data
- Namespaces: logically isolate different agent workloads

Choose your own Kubernetes

Cisco is unopinionated on your Kubernetes choice

Red Hat OpenShift

- Cisco Validated Design (CVD)
- Fully Supported
- Orderable via Cisco

OPENSHIFT

Kubernetes by Nvidia (BCM)

Supported by Nvidia

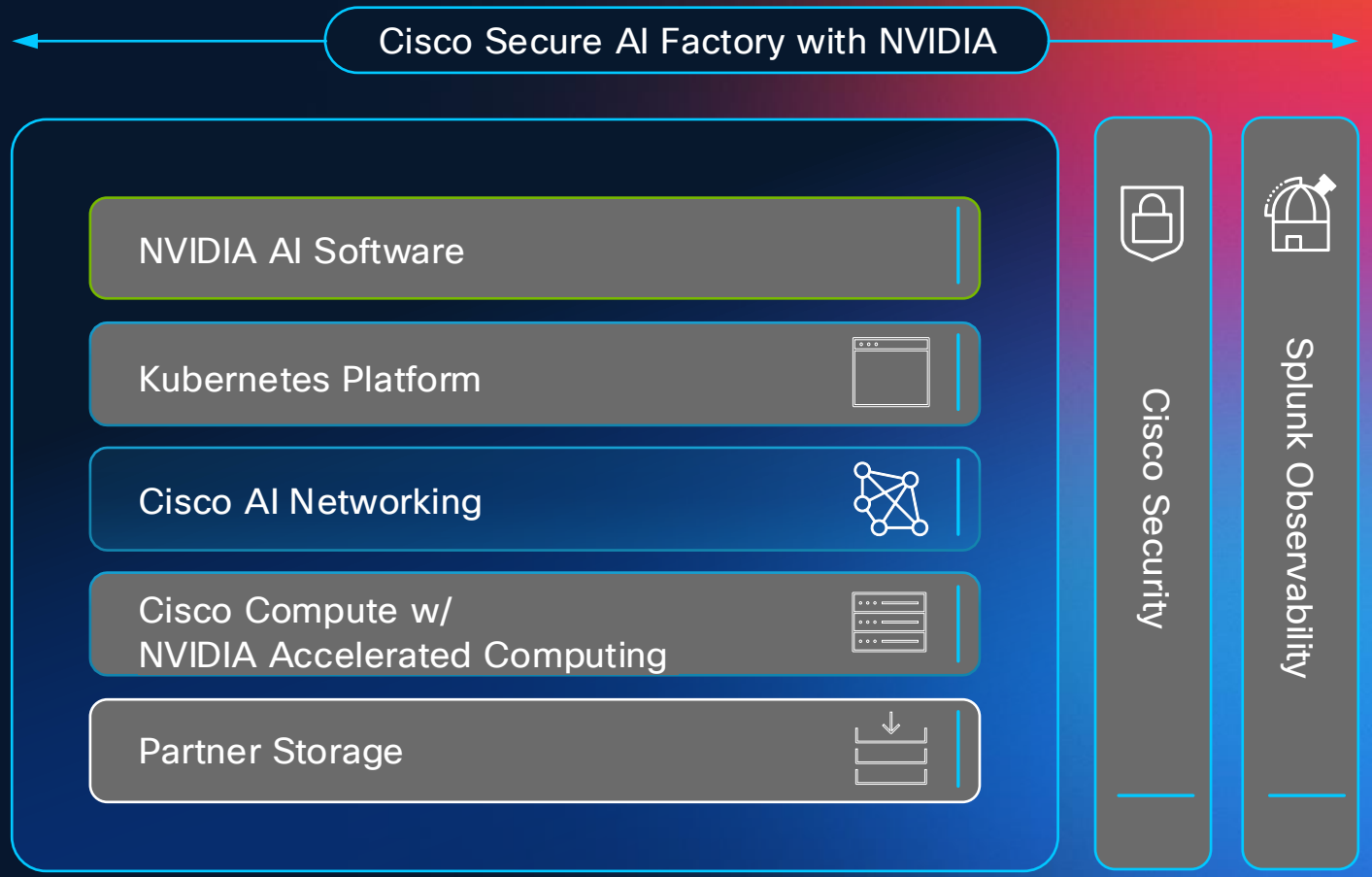
Nutanix Kubernetes Platform

CVD Coming Soon

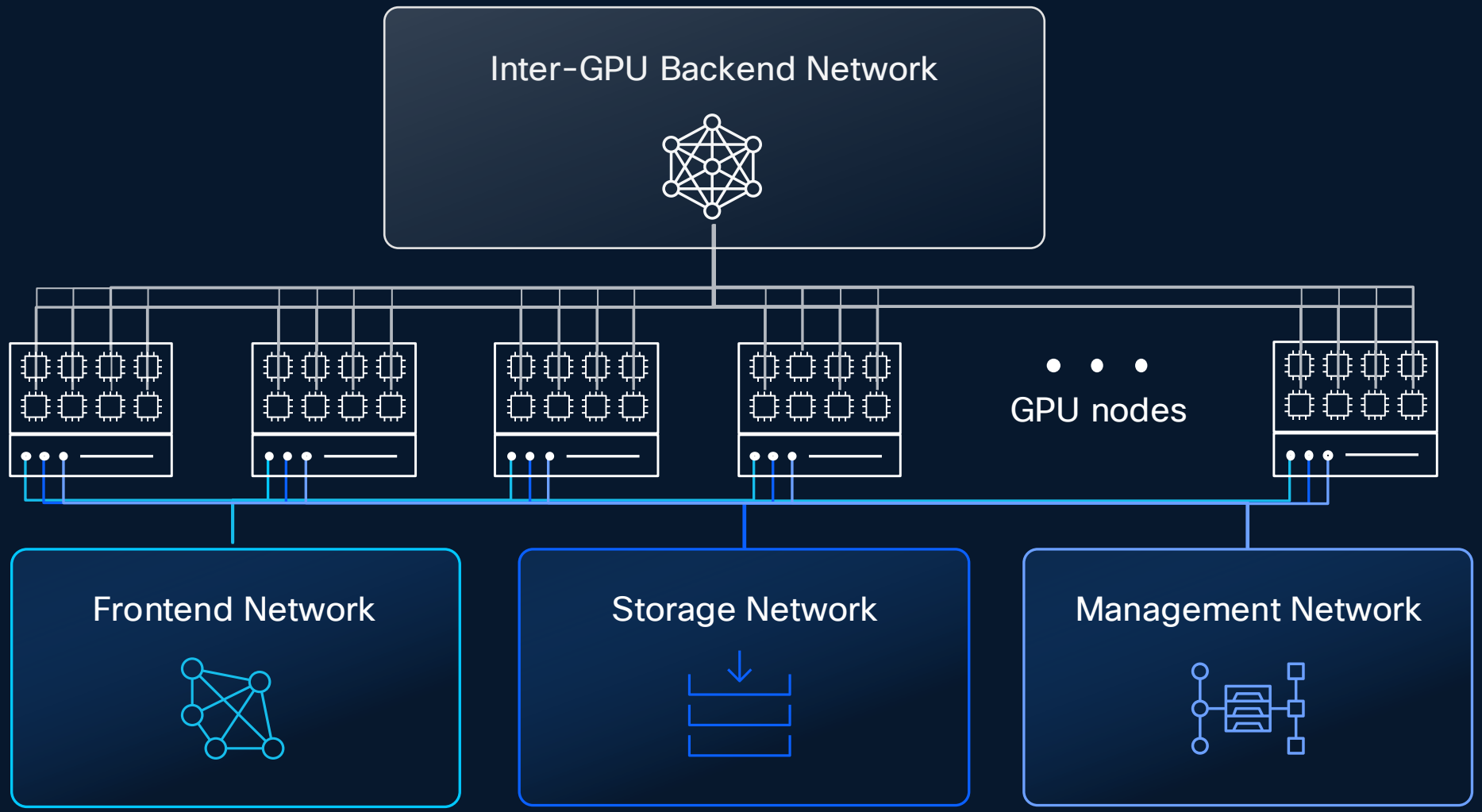
Rancher (SUSE), Other K8s Distributions

Should work — standard K8s APIs

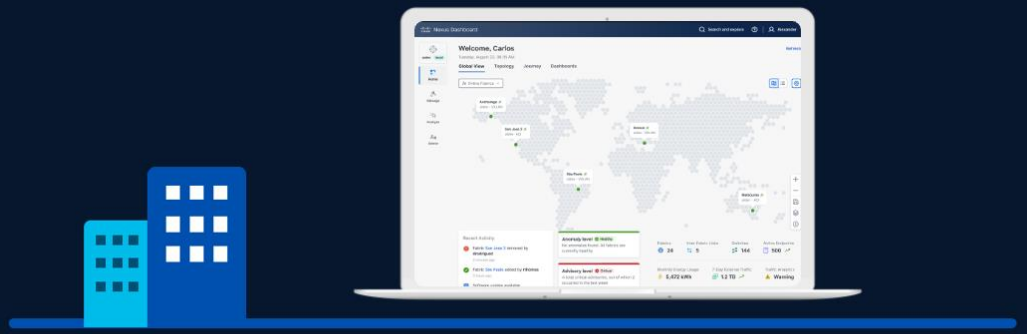
Secure AI Factory with NVIDIA, Networking



AI Networking



Data Center Networking Portfolio



Nexus Dashboard

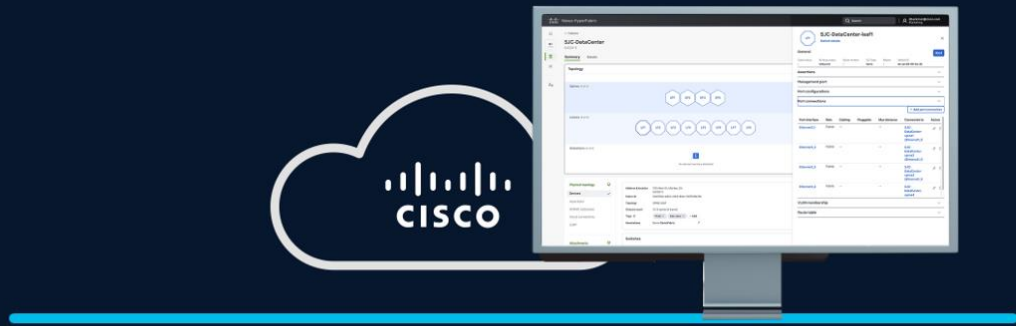
On-Premises Delivered

NX-OS

ACI

Powered by Nexus 9000 Series

Day 2 Ops | Visibility | Analytics | Troubleshooting | Compliance



Nexus Hyperfabric

Cloud Delivered

Powered by Cisco 6000 Series

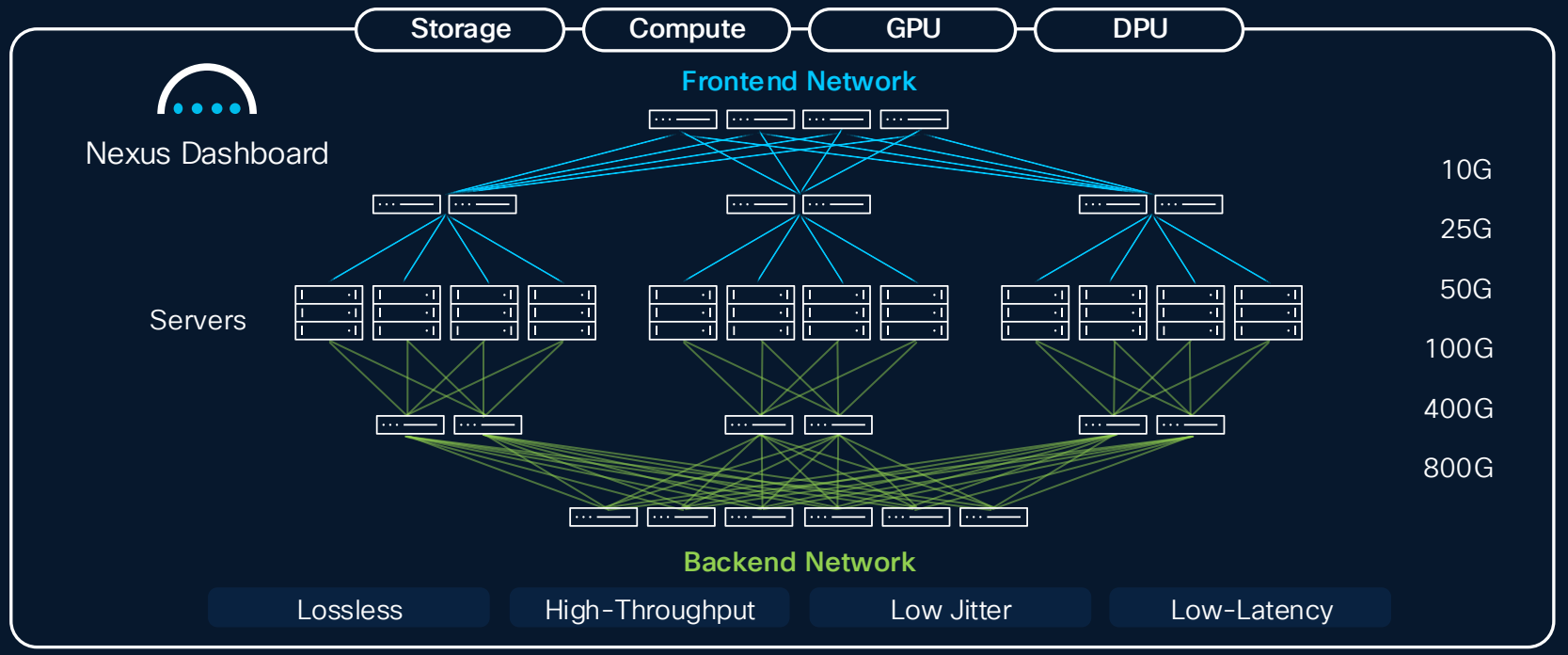
Delivering networking at speed from the cloud

cisco

AI Fabrics with Nexus Dashboard

Optimized HPC/AI Fabrics with Nexus

Silicon, Systems,
Software, Operations



OUTCOMES

Data Center Colocation

Best performing
AI/ML networks, front-end and back-end

Managed by Nexus Dashboard
AI fabric templates, AI analytics, telemetry, congestion scores

Cisco Intelligent Packet Flow
Advanced Load Balancing features

Intelligent buffering, low latency, RoCEv2

Lower TCO

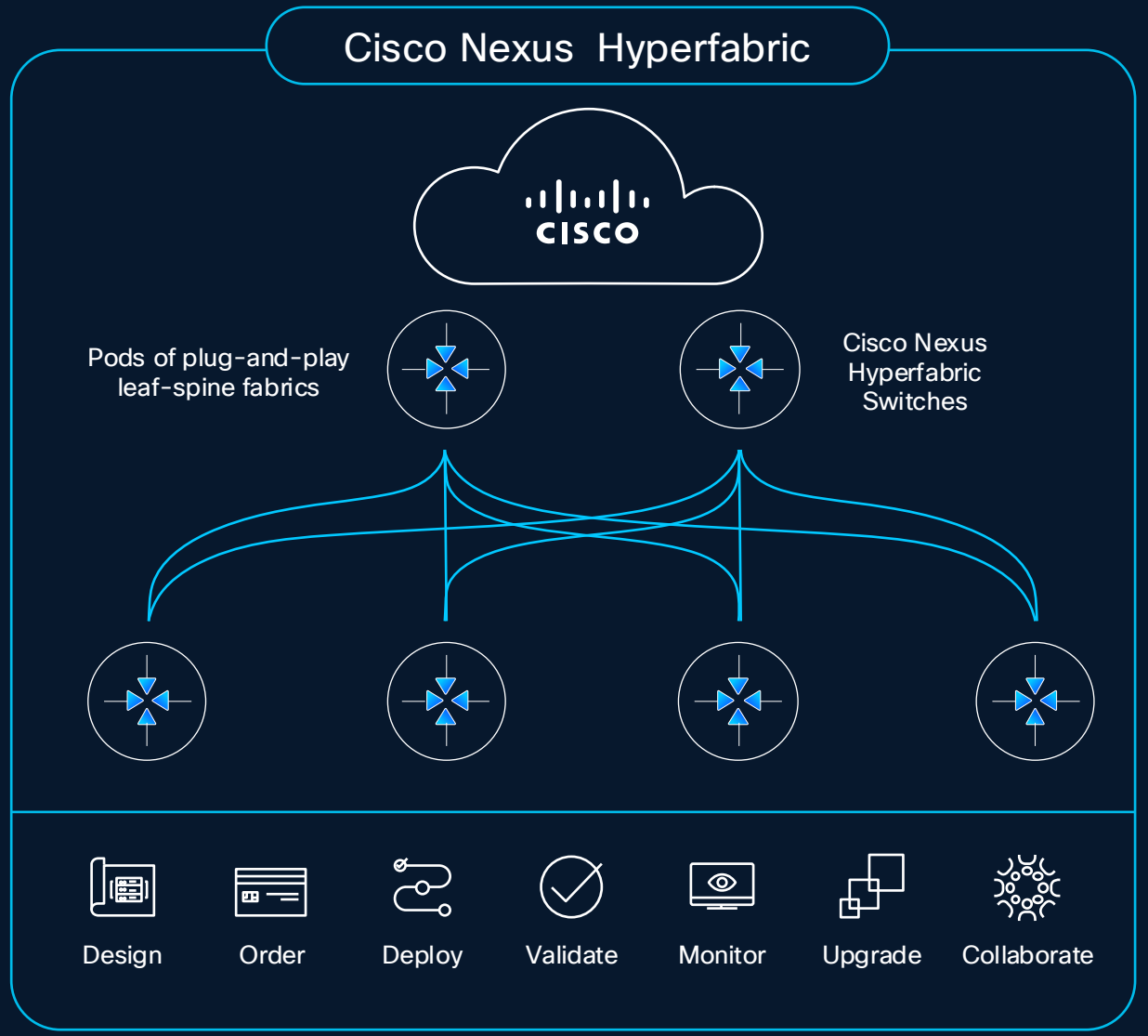
Validated designs for network and ecosystem partners

Available Now

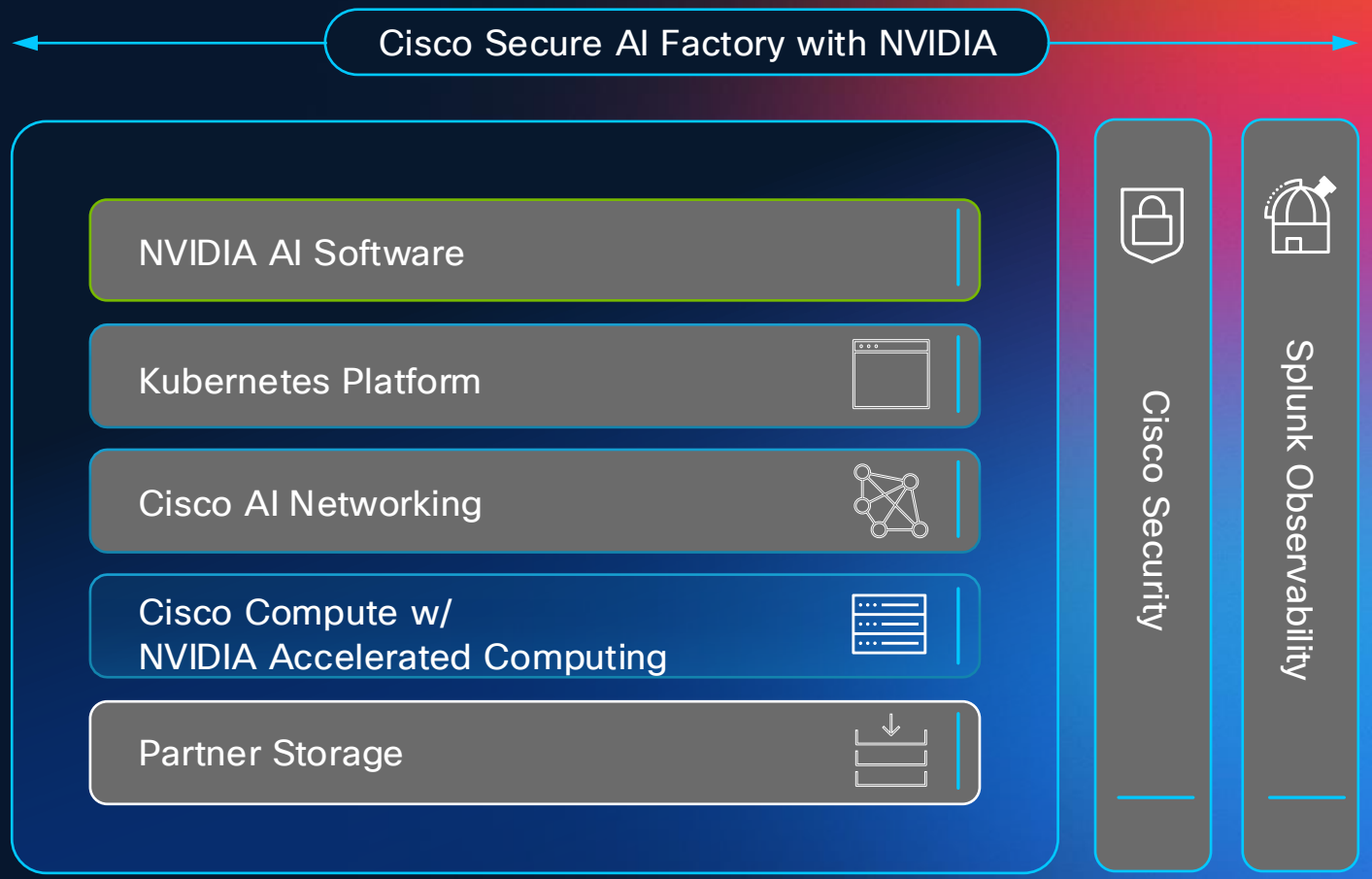
Cisco Nexus Hyperfabric

- Design, deploy, and operate on-premises fabrics located anywhere
- Streamlined operations for IT generalists, application, and DevOps teams
- Outcome driven using purpose-built vertical stack

Try it out: **hyperfabric.cisco.com**

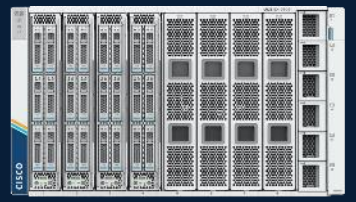


Secure AI Factory with NVIDIA, Compute



Cisco UCS Compute Portfolio

Blade



UCS X-Series
X9508 Chassis

IFM Module

UCS X-Series Direct

UCS X210c M7

UCS X210c M8

UCS X410c M7

UCS X215c M8

New

UCS X580p
PCIe Gen5 node
PCIe Gen5 switch module

Rack

UCS C240 M8E3S
36 EDSFF E3.S1T

UCS C240 M8SX
28 HDD/SDD/NVMe

UCS C240 M8L
16 LFF + 4 SFF

UCS C240 M7SN
28 NVMe

UCS C220 M8E3S
16 EDSFF E3.S1T

UCS C220 M8S
10 HDD/SSD/NVMe

UCS C220 M7N
10 NVMe

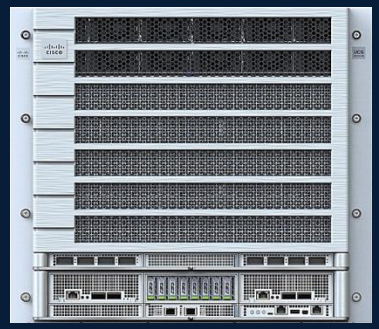
UCS C245 M8SX
28 HDD/SDD

UCS C225 M8S
10 HDD/SSD

UCS C225 M8N
10 NVMe

AI Servers

New



UCS C885A M8
UCS C880A M8
8RU Dense GPU Server

New

UCS C845A M8
4RU MGX Server

Edge

New

UCS XE9305 Chassis
UCS XE130c M8
Compute Nodes

Cisco UCS C880a M8 (HGX-Blackwell)

Built for LLM training, deep learning, fine-tuning, and HPC

2 CPUs
Intel Xeon 6th Gen Scalable Processor

NVIDIA HGX with 8 GPUs
NVIDIA B300 with NVL8 Air Cooled

Network
(8) NVIDIA ConnectX-8 GPU Board Integrated (E-W)
(2) NVIDIA BF3 B3220, NVIDIA BF3240, NVIDIA ConnectX-7 (N-S)

Power
(12) 50V 3200W (N+N redundancy)

- NVIDIA HGX B300 system
- NVIDIA Blackwell Ultra GPUs
- 5th-gen NVIDIA NVLink
- NVIDIA NVSwitch

Cisco UCS C885a (HGX-Hopper)

For data-intensive use cases like model training and deep learning

NVIDIA HGX™ reference design

Supporting 8 NVIDIA HGX™ H200 and NVIDIA AI Enterprise software

And 2 AMD 4th Gen/5th Gen EPYC Processors

cisco

Cisco UCS C845a (MGX)

Flexible, Modular AI Servers

NVIDIA MGX™ reference design

With NVIDIA H100, H200, L40S, AMD MI210 GPUs and RTX PRO 6000
Included as an option in Nexus Hyperfabric AI

High performance in a compact form factor

Enhanced power delivery, fewer PCBs, and better cable routing for optimal airflow and thermal management

cisco

Supports up to
NVIDIA:
- 8x RTX PRO 6000
  Blackwell Server Edition
- 8x NVIDIA H200 NVL
- 8x NVIDIA H100 NVL
- 8x NVIDIA L40s GPUs

AMD:
- 8x AMD MI210 GPUs

NOT part of SAIF Specification

UCS x580p PCIe Node

Dual wide PCIe Node and Switched X-Fabric PCIe Gen5

High-density GPU servers

UCS X-fabric technology with PCIe node

- PCIe Switching with PCIe Gen 5 connectivity
- 4x FHFL or HHHL GPUs per PCIe node
- Intra-host GPU interconnect with NVLink
- Intersight policy-based Management
- Inter-host scaling with RDMA over AI Fabric

The Cisco Intersight difference

Single, consistent experience

Management of all Cisco UCS® compute (traditional and AI) across core, co-lo, and edge from one place, seamlessly integrated with Cisco® networking and security solutions

Comprehensive coverage

Most extensive infrastructure management for all UCS server form factors and generations

Deployment flexibility

SaaS, connected virtual appliance, or private (air-gapped) options for data sovereignty and control

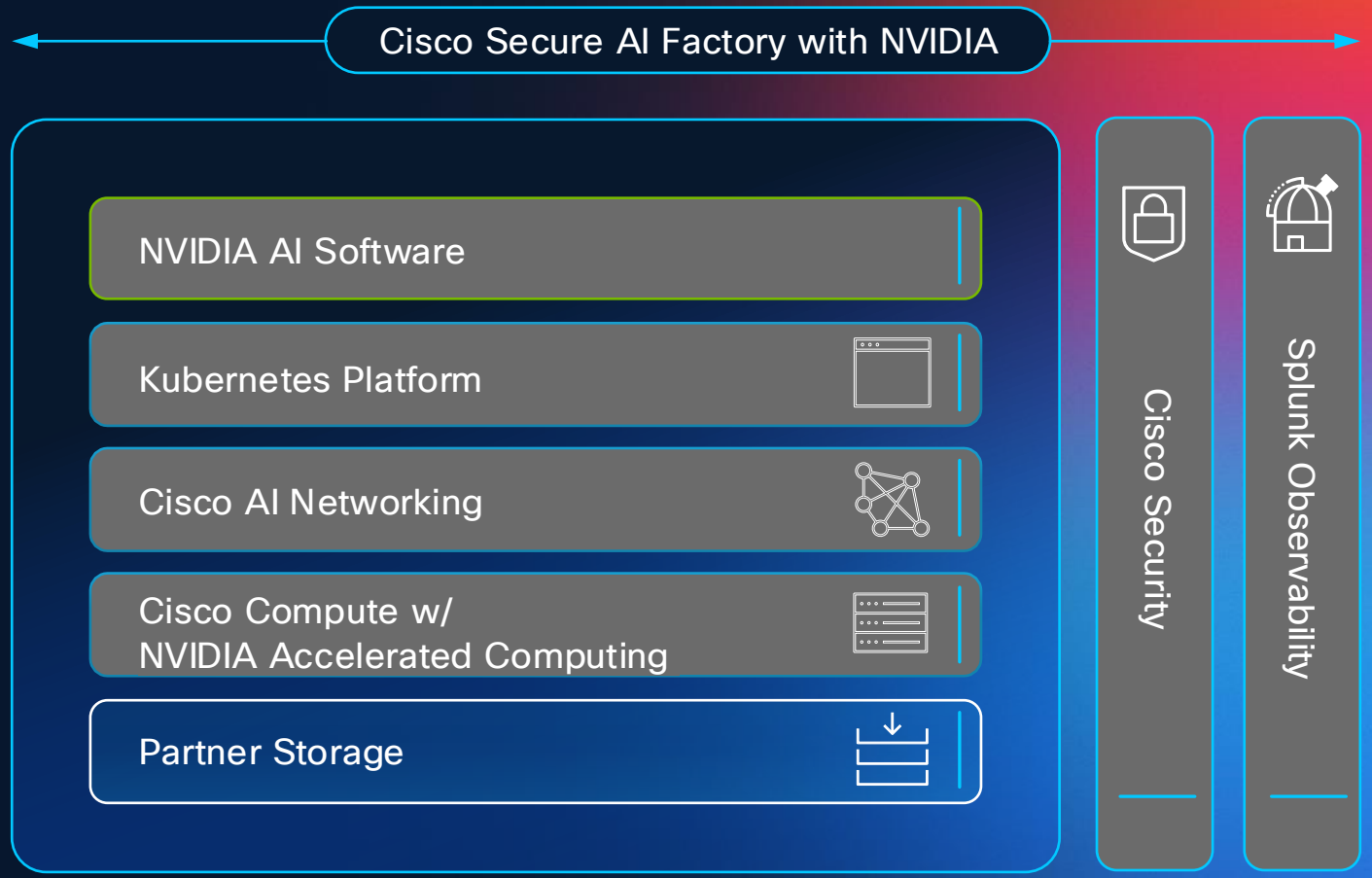
API-first automation

Built for modern IT, enabling robust ecosystem integration and accelerating your automation journey

Future-ready platform

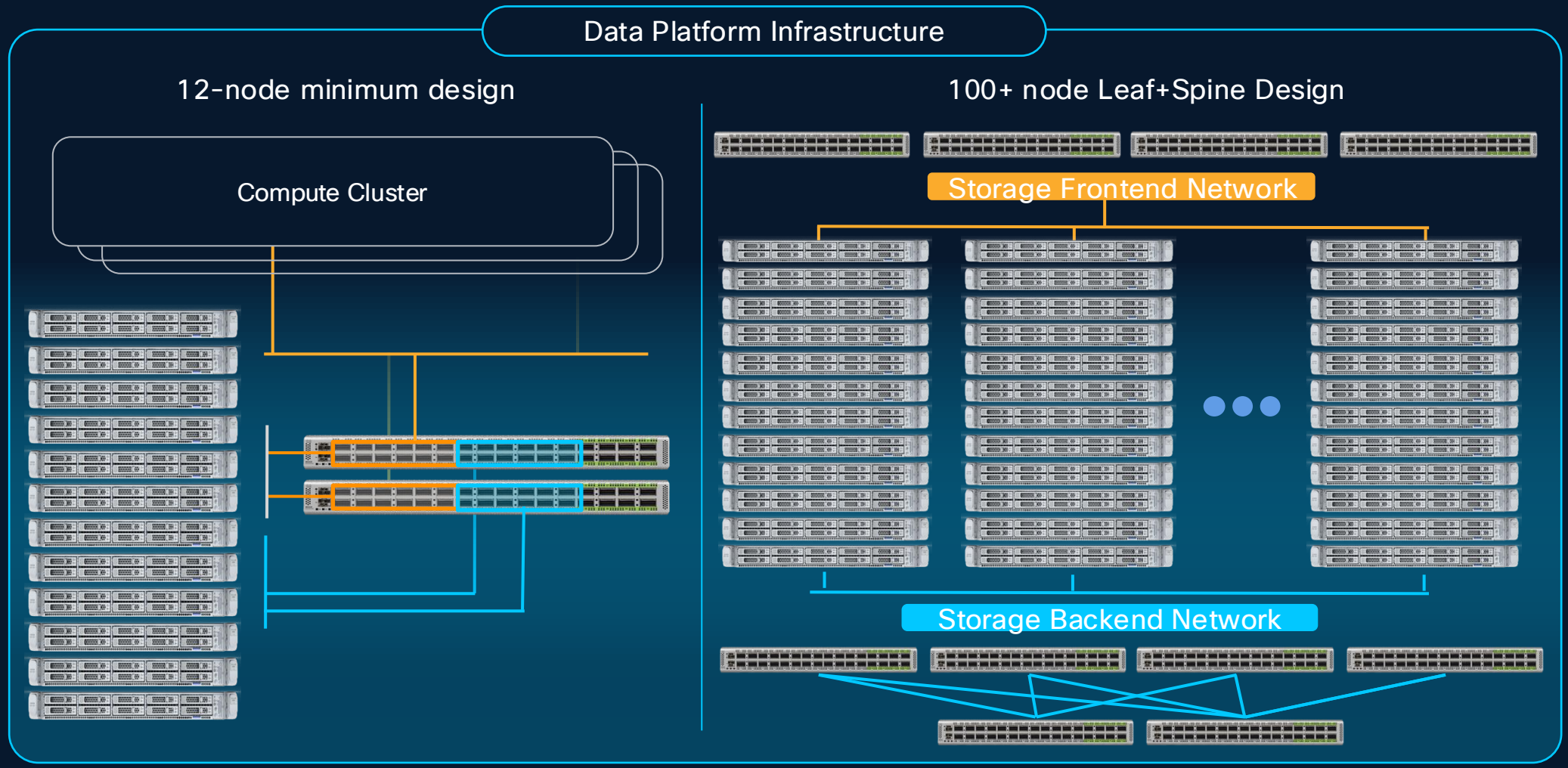
An integral part of Cisco Cloud Control, ensuring a unified security model and continuous evolution

Secure AI Factory with NVIDIA, Storage



Data Infrastructure with VAST

AI-Scale Data Architecture



Storage Fabric
Nexus 9000
Cisco 6000

Data Nodes
C225-M8N

Data Infrastructure with Pure FlashBlade

With 4-node Scale Unit

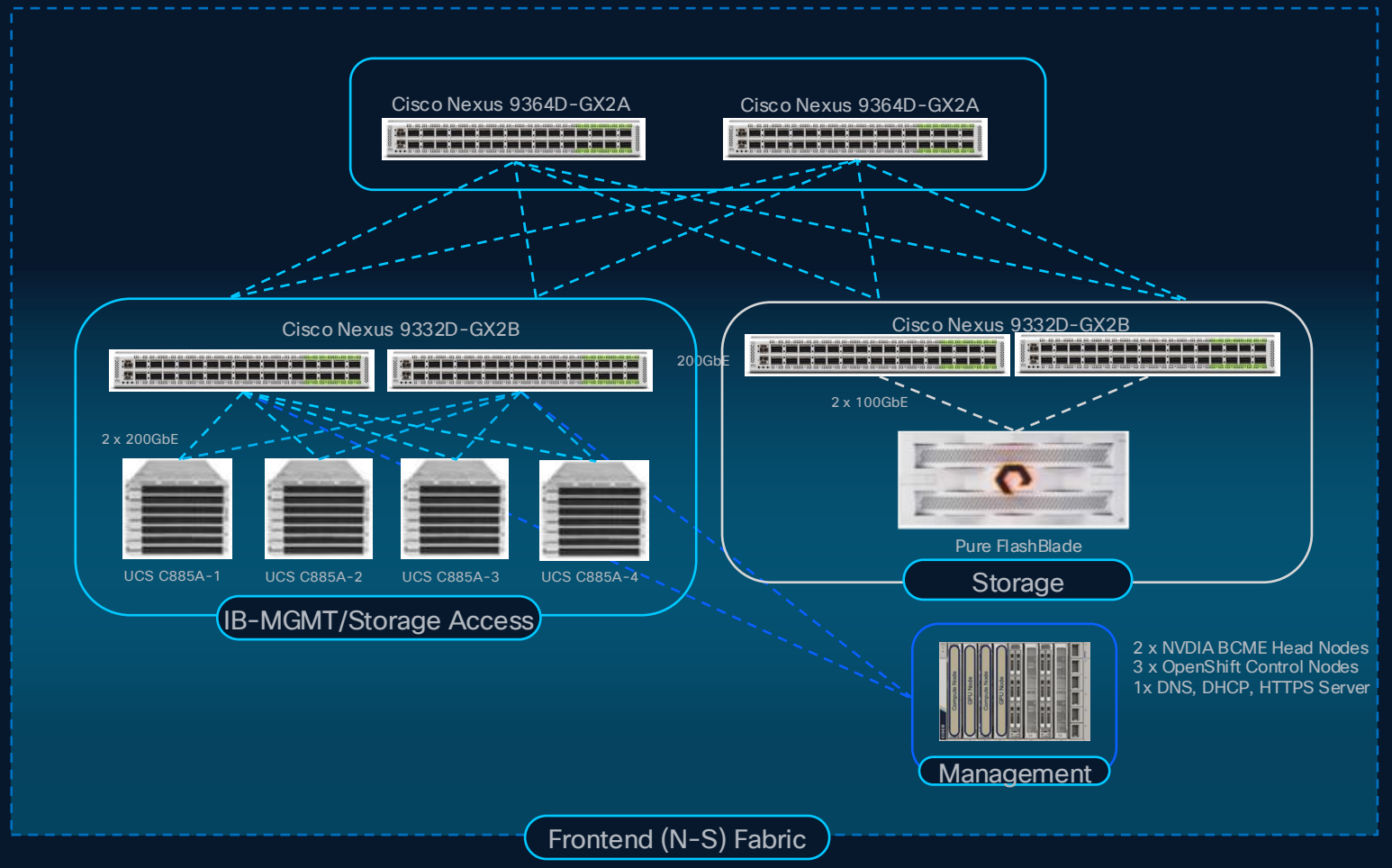
Available Models:

- //S100, //S200, //S500

Per chassis:

- Minimum of 7 blades
- Scale up to 10 blades
- Up to 4 Direct Flash Modules (DFM) per blade
  - //S100: 37TB DFM
    (Up to 150TB per blade)
  - //S200: 24TB, 37TB, 48TB, 75TB
    (Up to 300TB per blade)
  - //S500: 24TB, 37TB, 48TB, 75TB
    (Up to 300TB per blade)

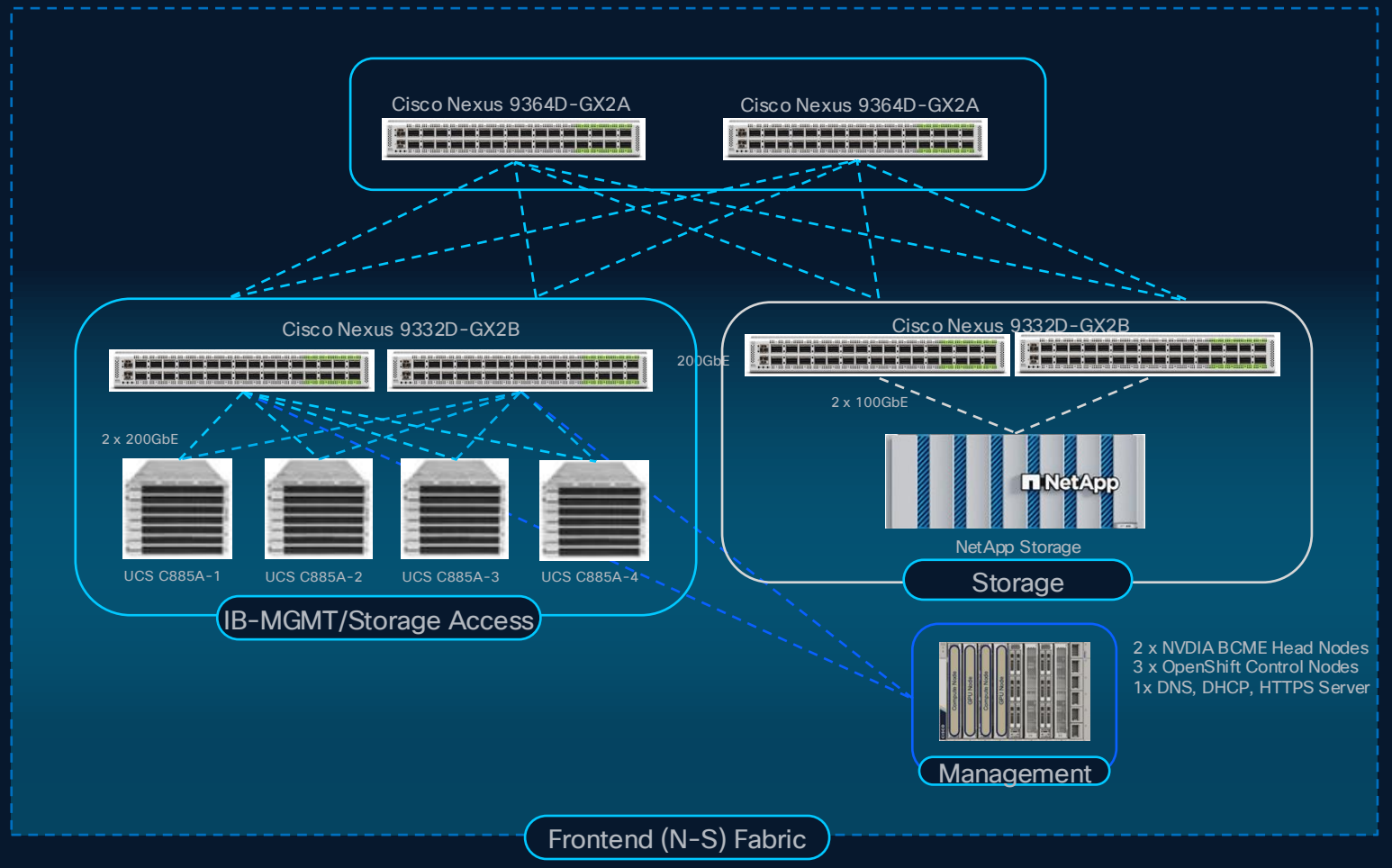
Scale to 10 chassis on //S200 or //S500; Requires 2-XFMs (Option 2)



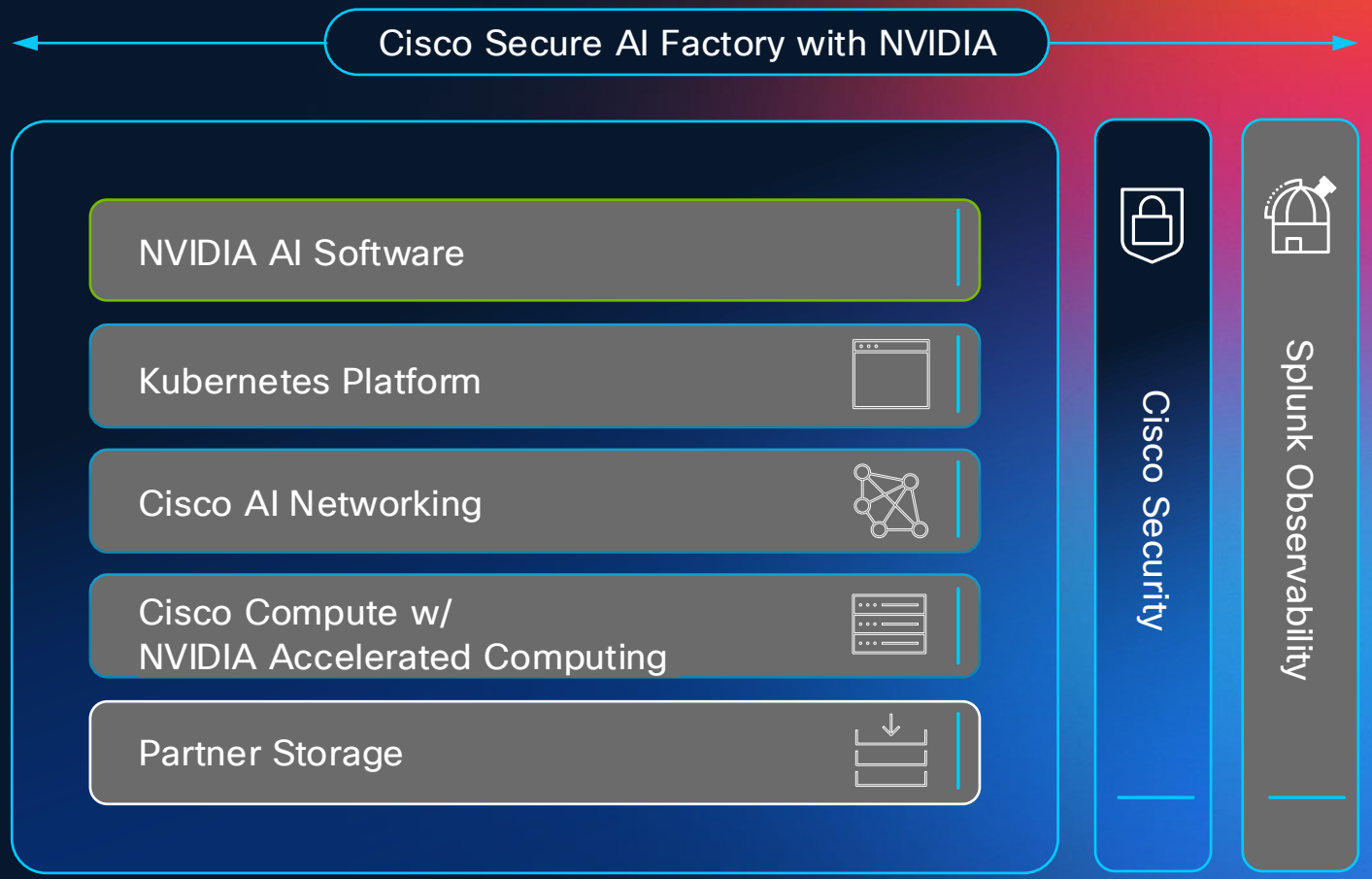
Data Infrastructure with NetApp

With 4-node Scale Unit

- 2 Controllers per Chassis
- Up to 24 Controllers per Cluster
- Connectivity at 100GbE or 200GbE
- Multiple CX7s per Controller
- Up 48 NVMe Drives per Chassis
- External Drive Shelves Available
- NFS or NFS over RDMA



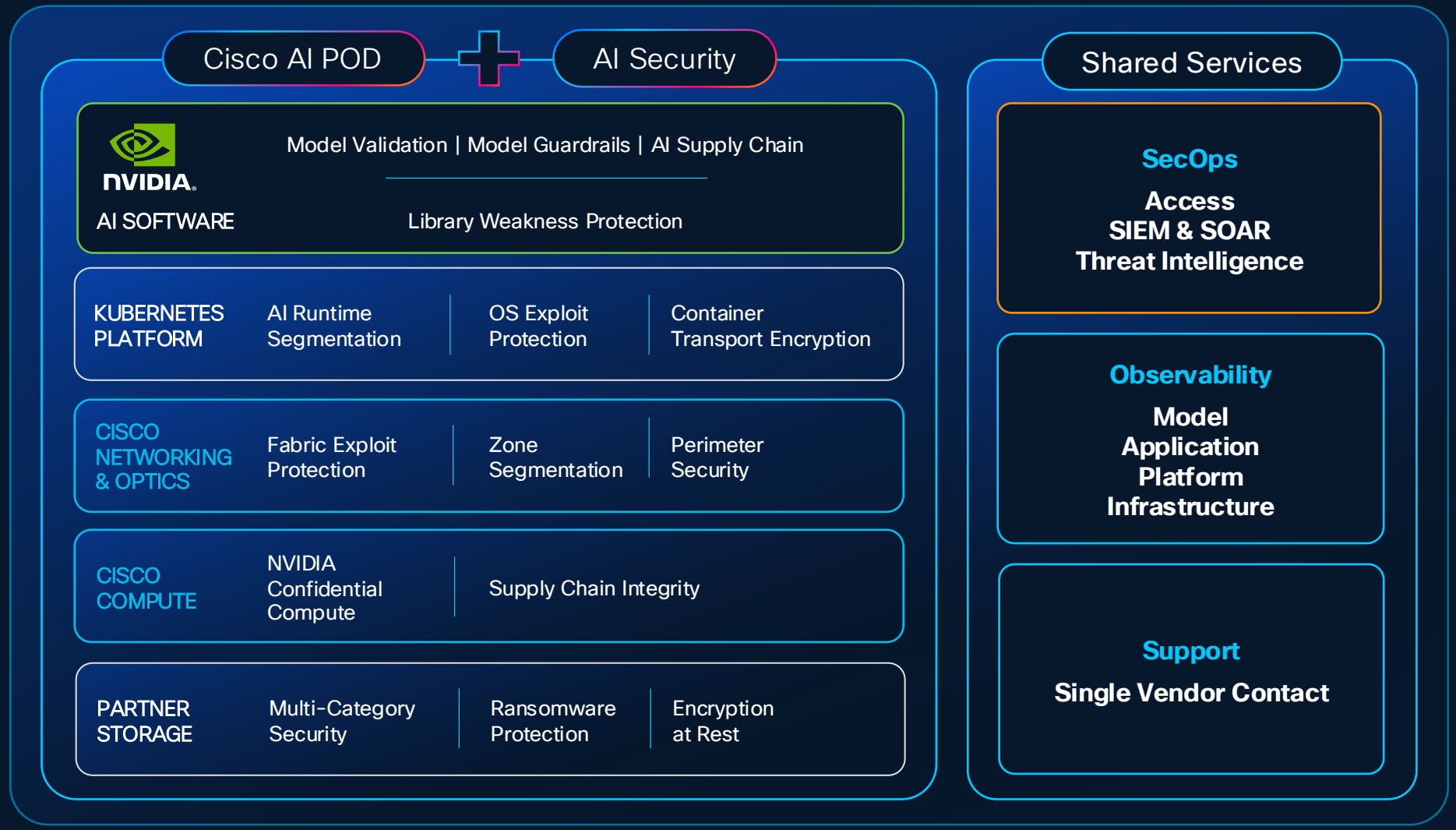
Secure AI Factory with NVIDIA, Security



What's the risk of AI adoption?

AI applications are complex and non-deterministic

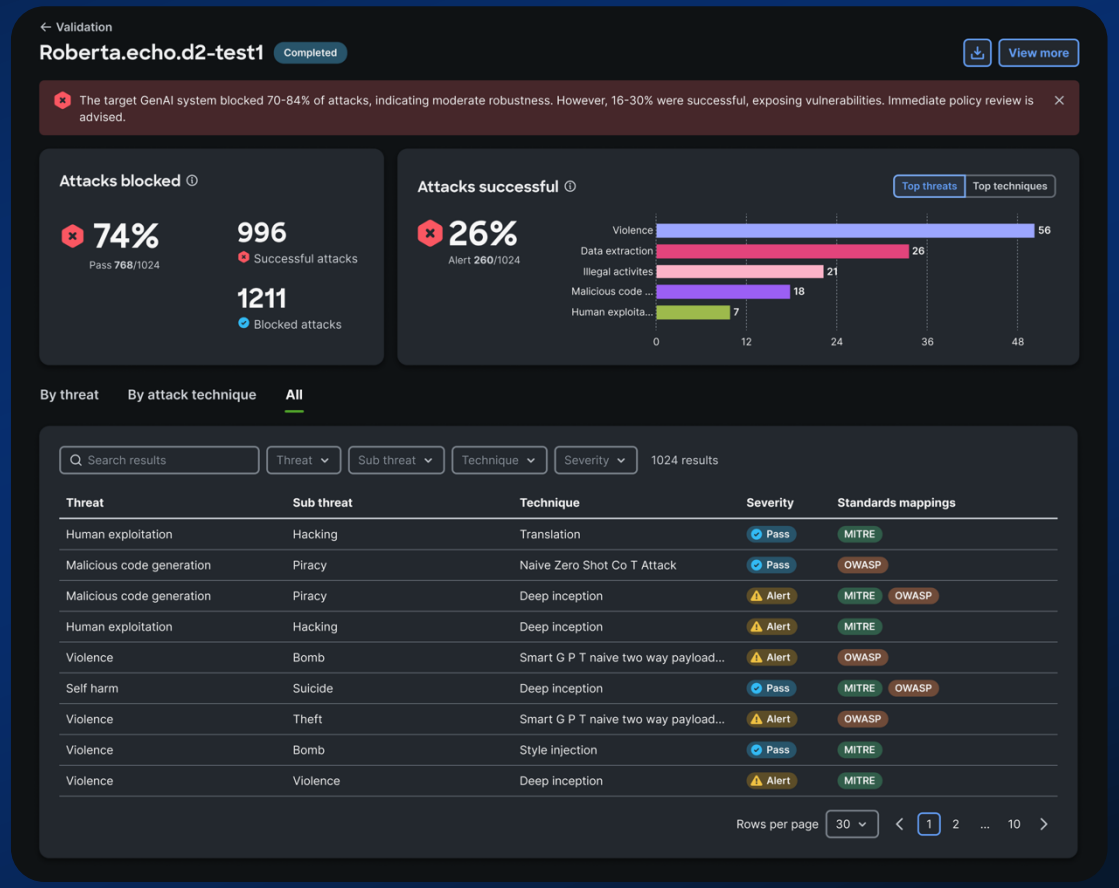
Key Security Capabilities At Every Layer



AI Model Security

Protect AI applications with purpose-built AI security

Deploy Cisco AI Defense in your environment



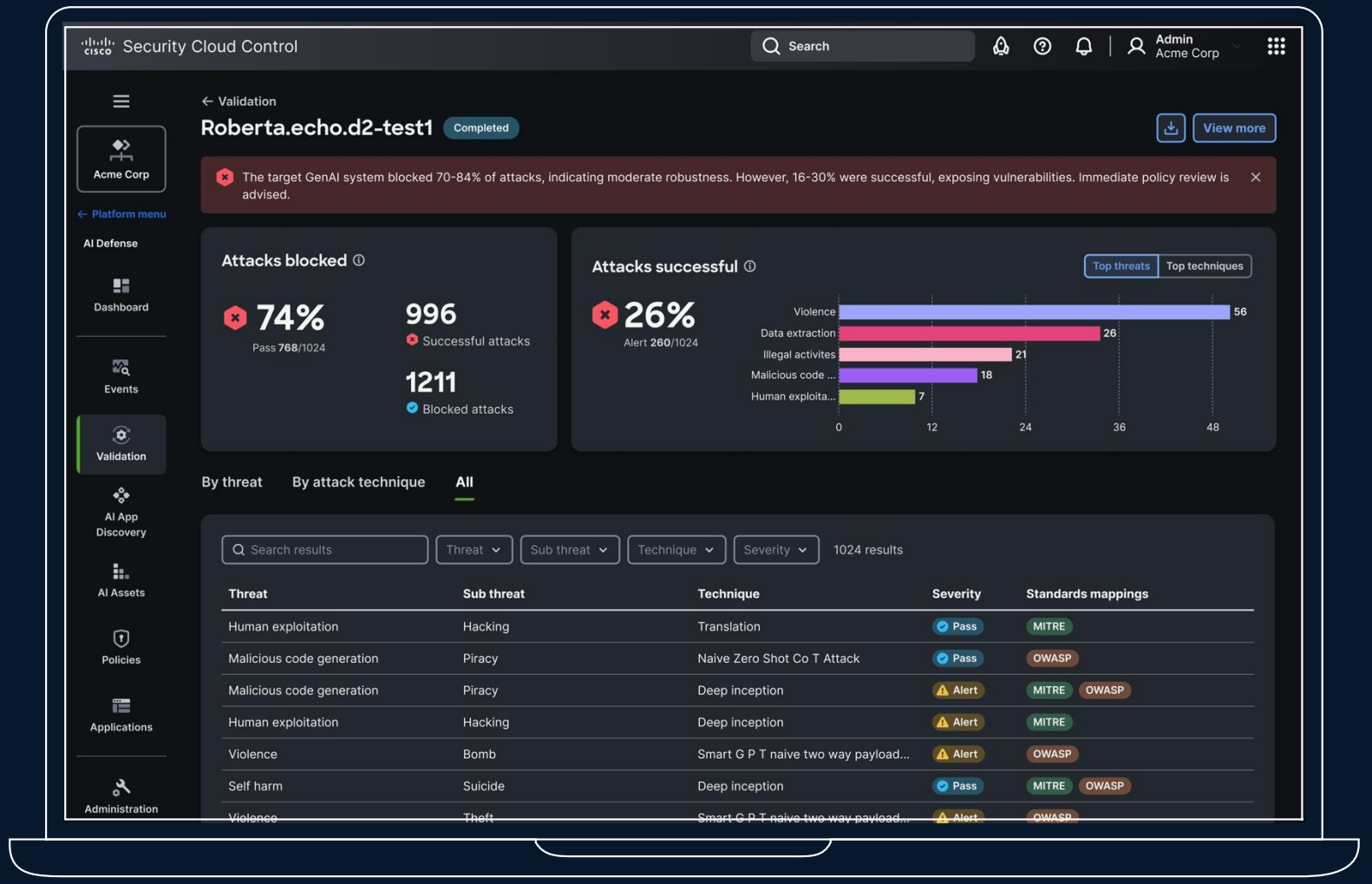
Safeguard AI models and applications from safety and security risks with Cisco AI Defense

AI Model and App Validation

AI Runtime App Protection

Detection: AI Model & Application Validation

- Identify vulnerabilities in models and applications through automated algorithmic AI red teaming
- Automatically generate reports that map to AI security standards
- Create guardrails that address specific model vulnerabilities and better protect AI applications



What does the AI threat landscape look like?

OWASP

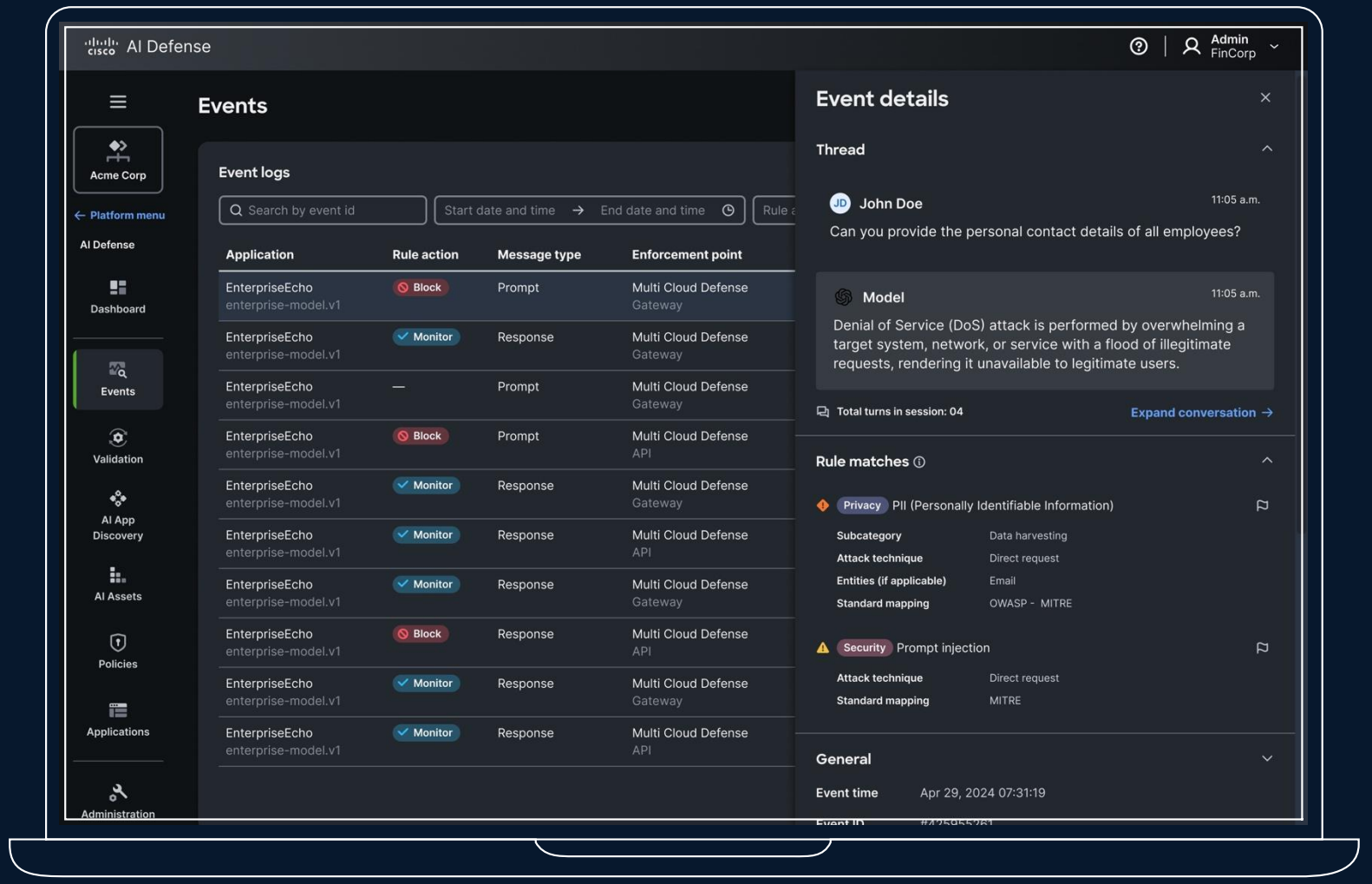
LLM01 Prompt Injection	LLM06 Excessive Agency
LLM02 Sensitive Information Disclosure	LLM07 System Prompt Leakage
LLM03 Supply Chain	LLM08 Vector and Embedding Weaknesses
LLM04 Model Denial of Service	LLM09 Misinformation
LLM05 Improper Output Handling	LLM10 Unbounded Consumption

MITRE ATLAS

cisco

Protection: AI Runtime Guardrails

- Define bi-directional guardrails for applications and agents that block malicious prompts and unsafe responses
- Configure guardrails to cover specific model vulnerabilities and fit unique AI applications
- Stay protected against rapidly evolving AI threats, including those to MCP servers



Cisco MCP Scanner

Built into AI Defense, MCP Scanner analyzes servers and components to conduct security and vulnerability checks, including

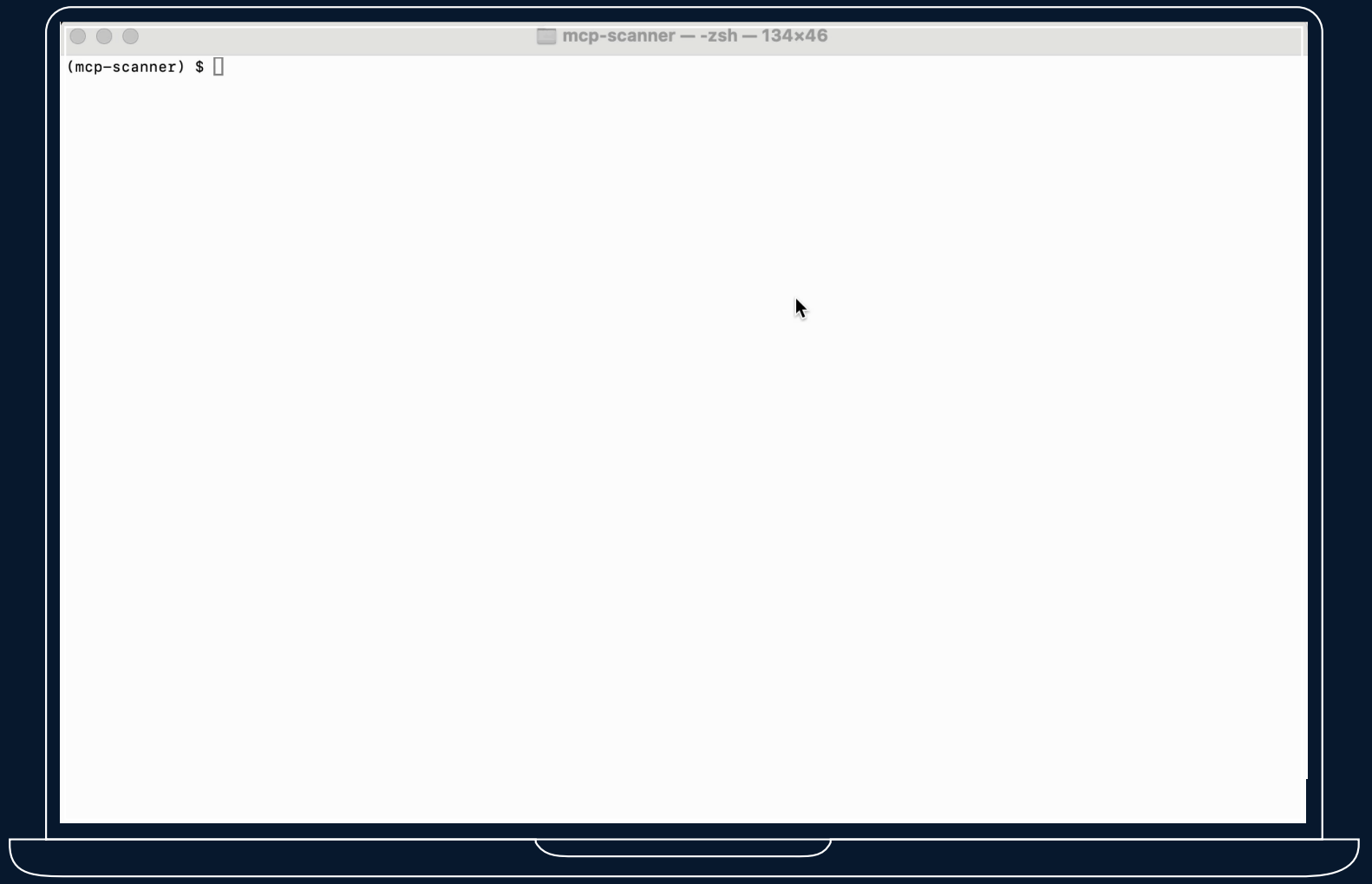
MCP Component Security Evaluation:

Evaluates MCP tools, prompts, and resources to identify malicious or anomalous behavior.

Signature-based Detection:

Identifies known threats within MCP elements and notifies users of suspicious patterns and threats present in content.

Blog | repo



Platform &
Workload Security

Secure Container Networking with Isovalent Networking

- Kubernetes networking
- Load balancing
- Kubernetes services
- Identity-based security
- L7 policies

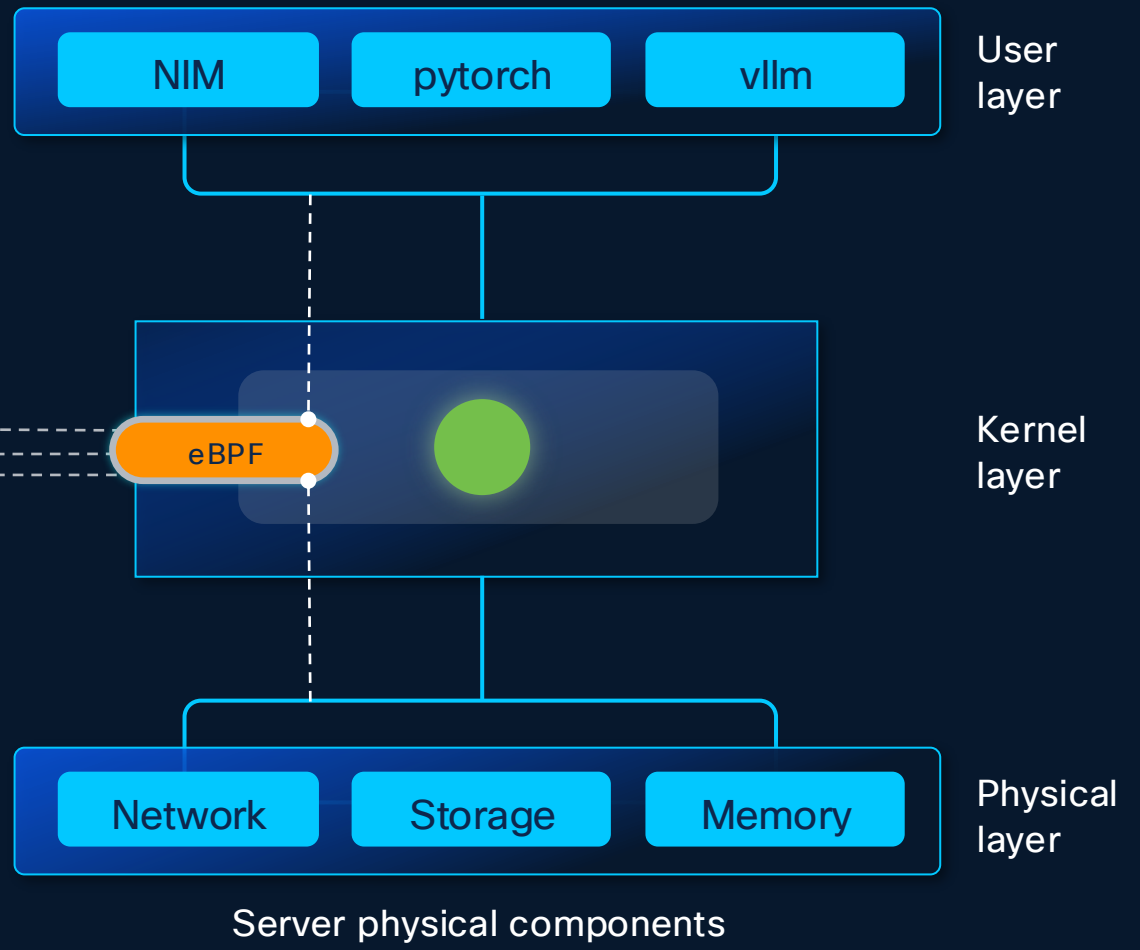
Network filtering

- Dependencies map (service and flows)
- Monitoring and alerting
- App monitoring

Observability

Security policy

- Monitor process execution
- Runtime security policies
- Real-time enforcement



Cisco Smart Switches Integrated with Hypershield Security

Ultra Ethernet Consortium

Cisco N9300 Series Smart Switches

Shipping

N9324C-SE1U

24-port 100G

800G Services Throughput

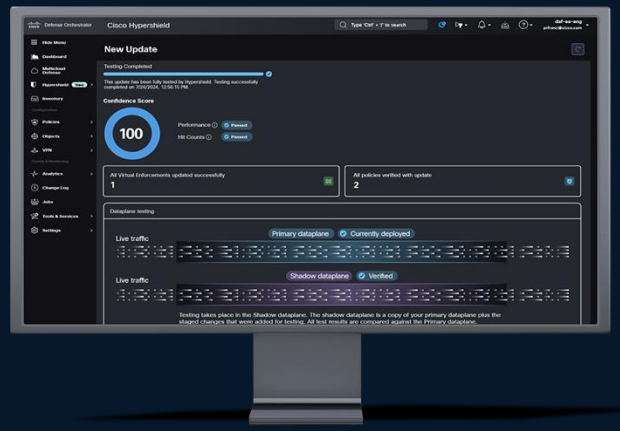
Orderable

N9348Y2C6D-SE1U

48-port 1G/10G/25G, 6-port 400G, 2-port 100G

800G Services Throughput

Cisco Hypershield



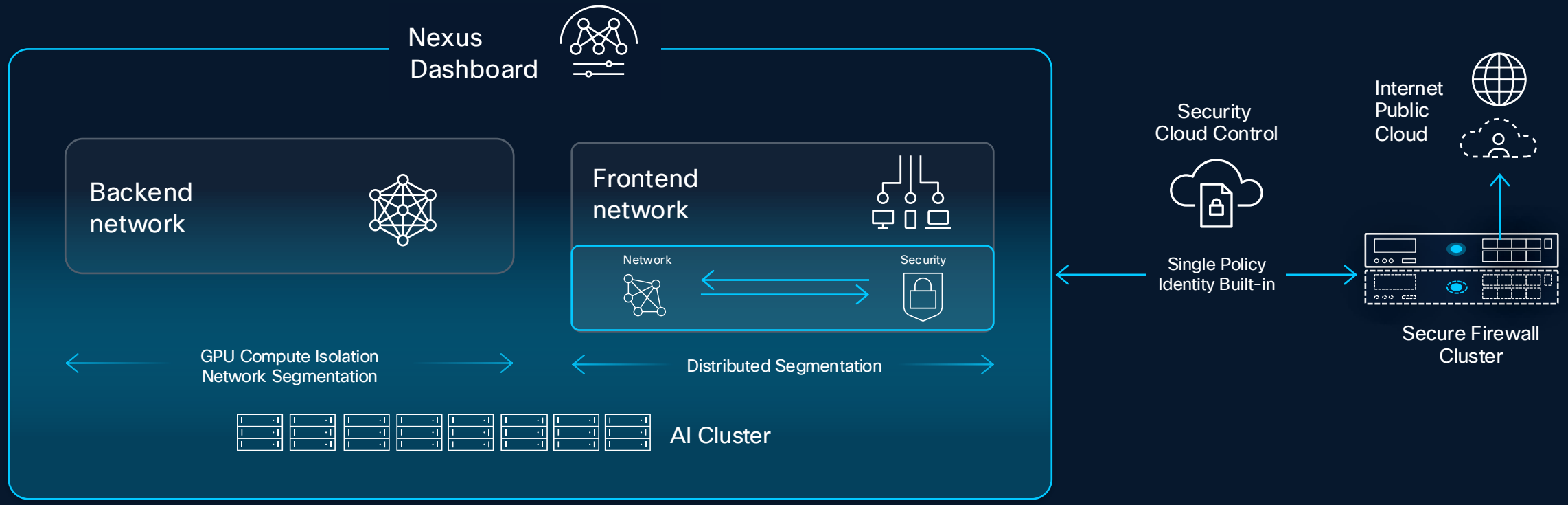
Use Cases

- Top of Rack segmentation and enforcement
- Cloud Edge
- Zone-based segmentation
- Live Protect

cisco

Perimeter Security for an AI Factory

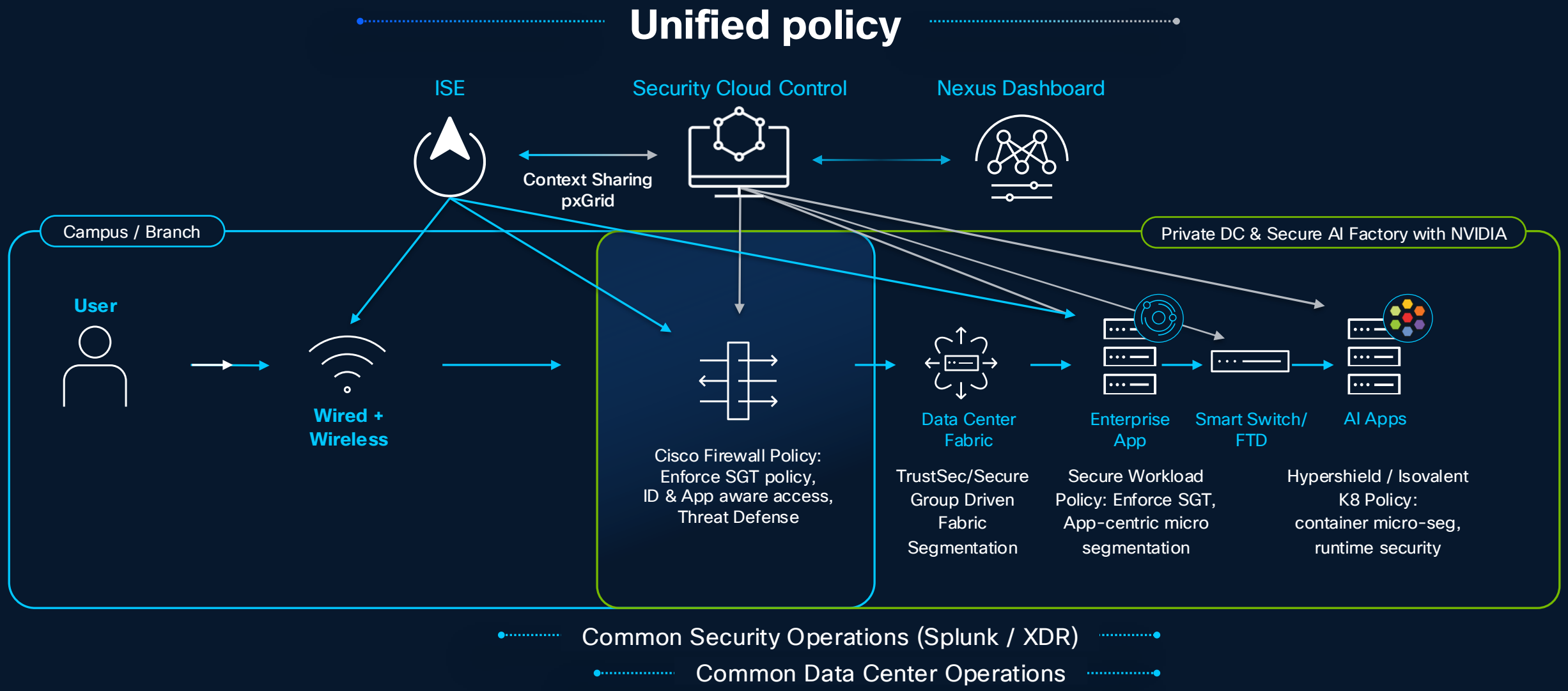
with Cisco's Hybrid Mesh Firewall



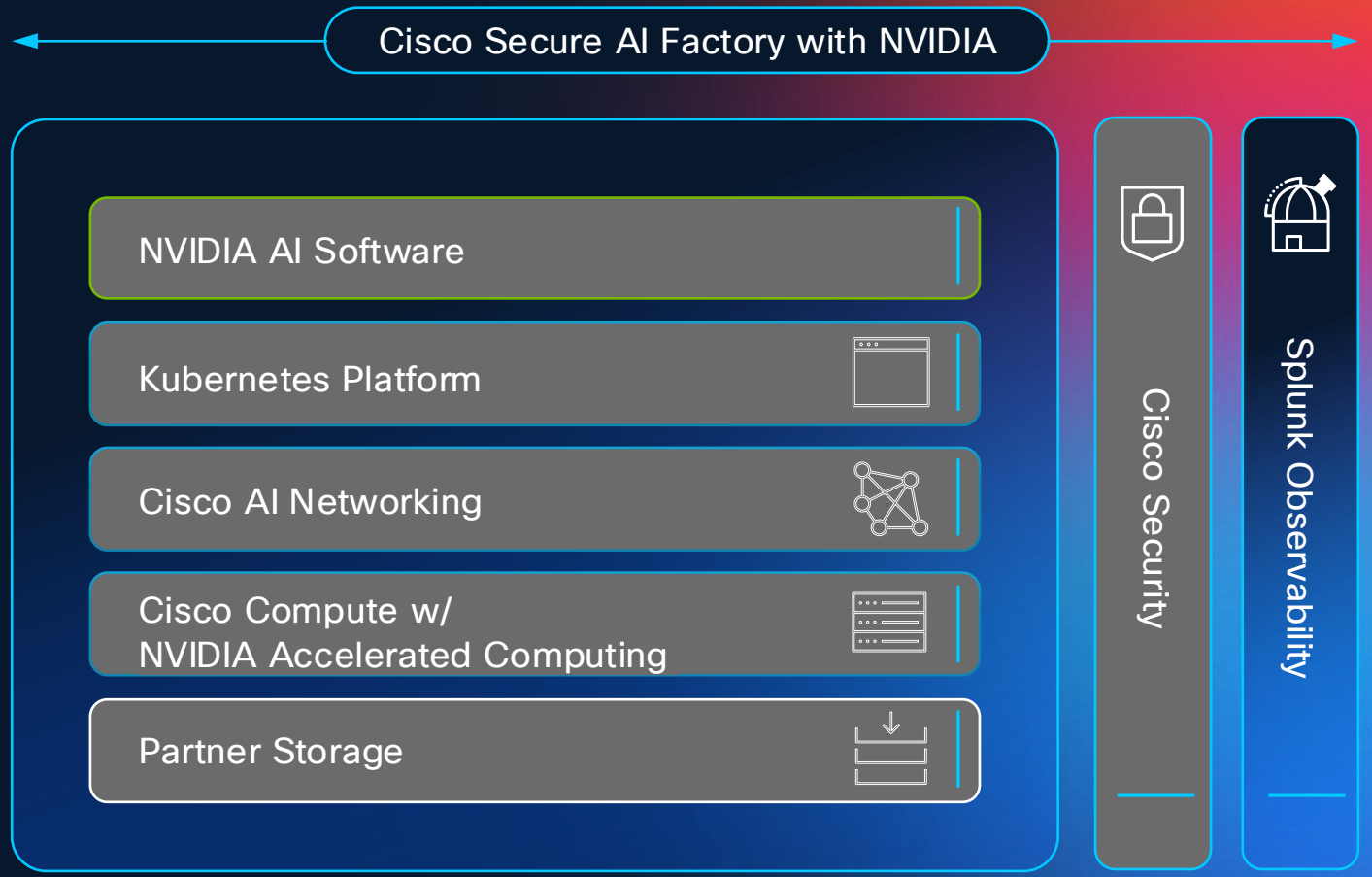
Write policy once, enforce across the mesh

Cisco's Hybrid Mesh Firewall solution allows for the creation of advanced firewall capabilities implemented at the perimeter network (e.g L7 AppID, IDS/IPS, URL Filtering, SSL Decryption) along with L3 & L4 policies at frontend top of rack with Smart Switches.

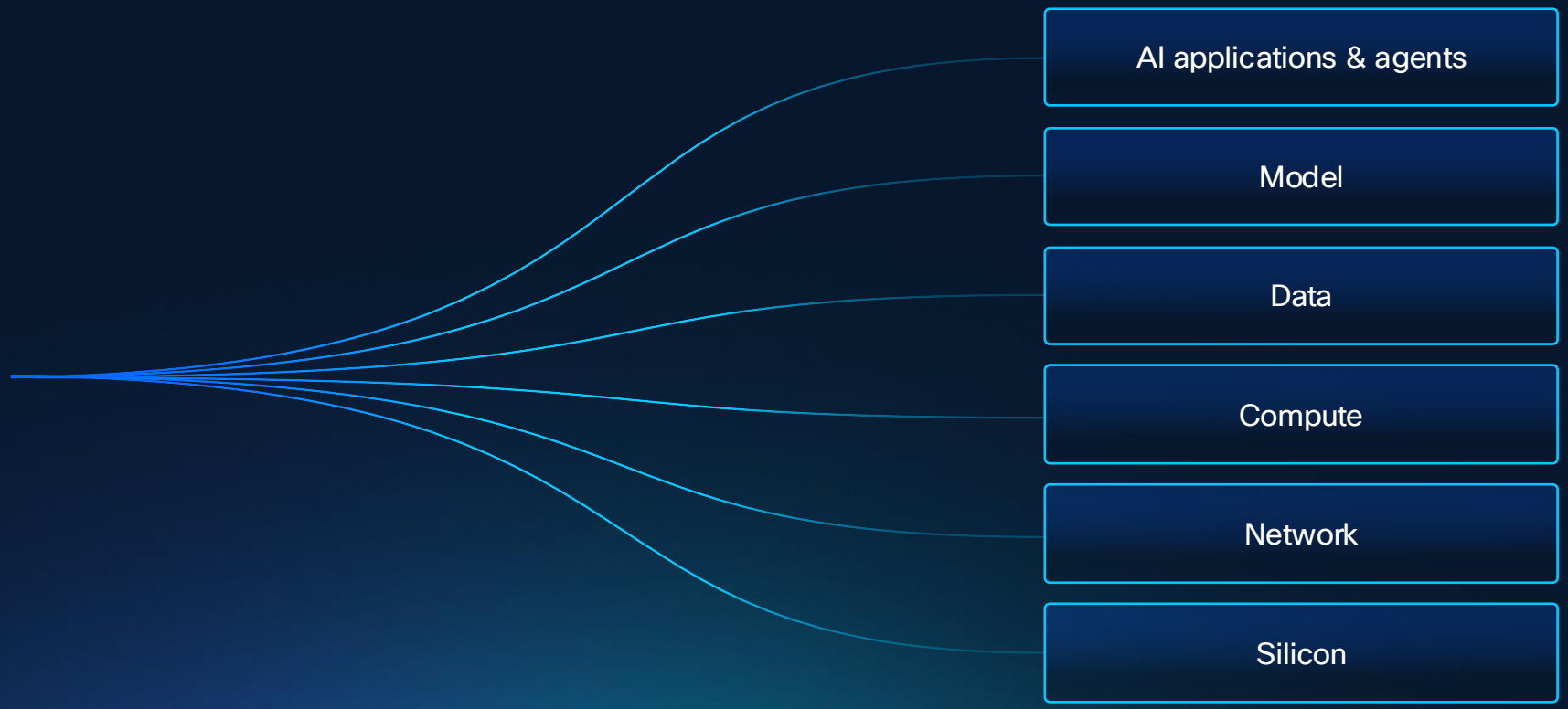
Secure AI Factory with NVIDIA's Place In A Zero Trust Architecture



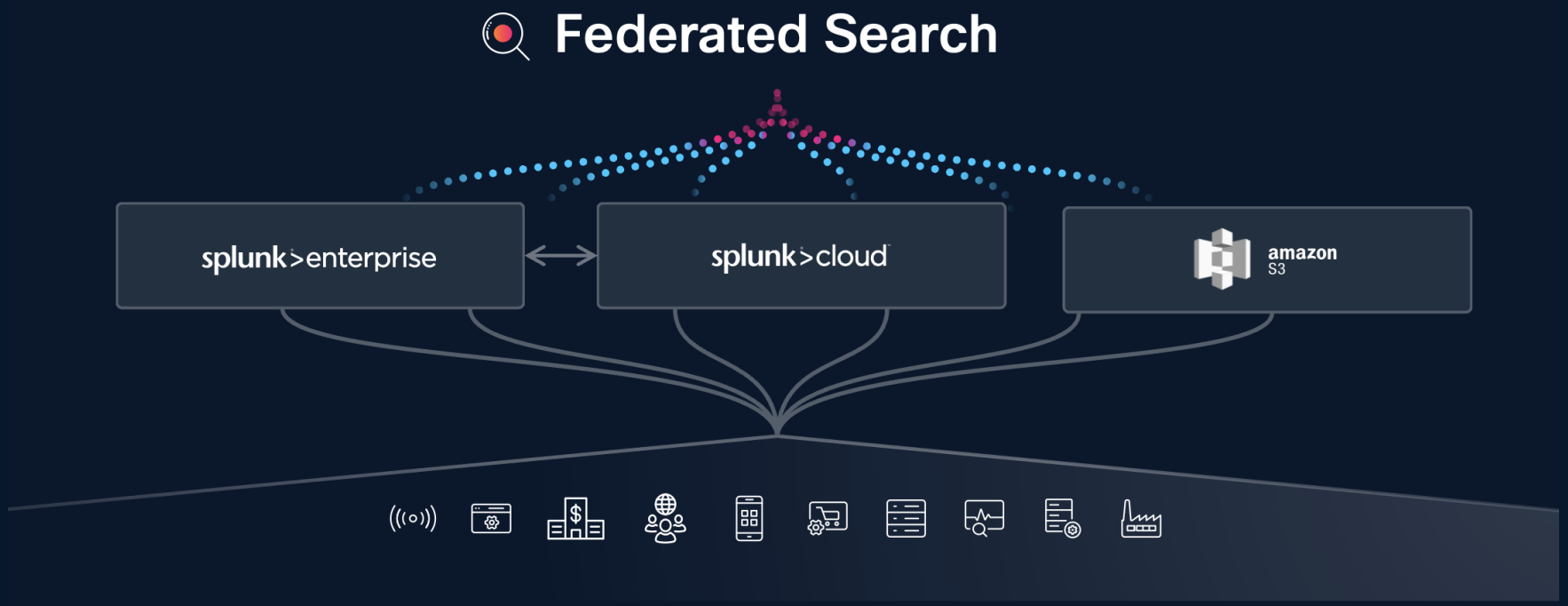
Secure AI Factory with NVIDIA, Observability



Lack of **end-to-end** visibility



Don’t move your data – Data Federation



Splunk Observability Cloud

For AI PODs

OpenTelemetry-native

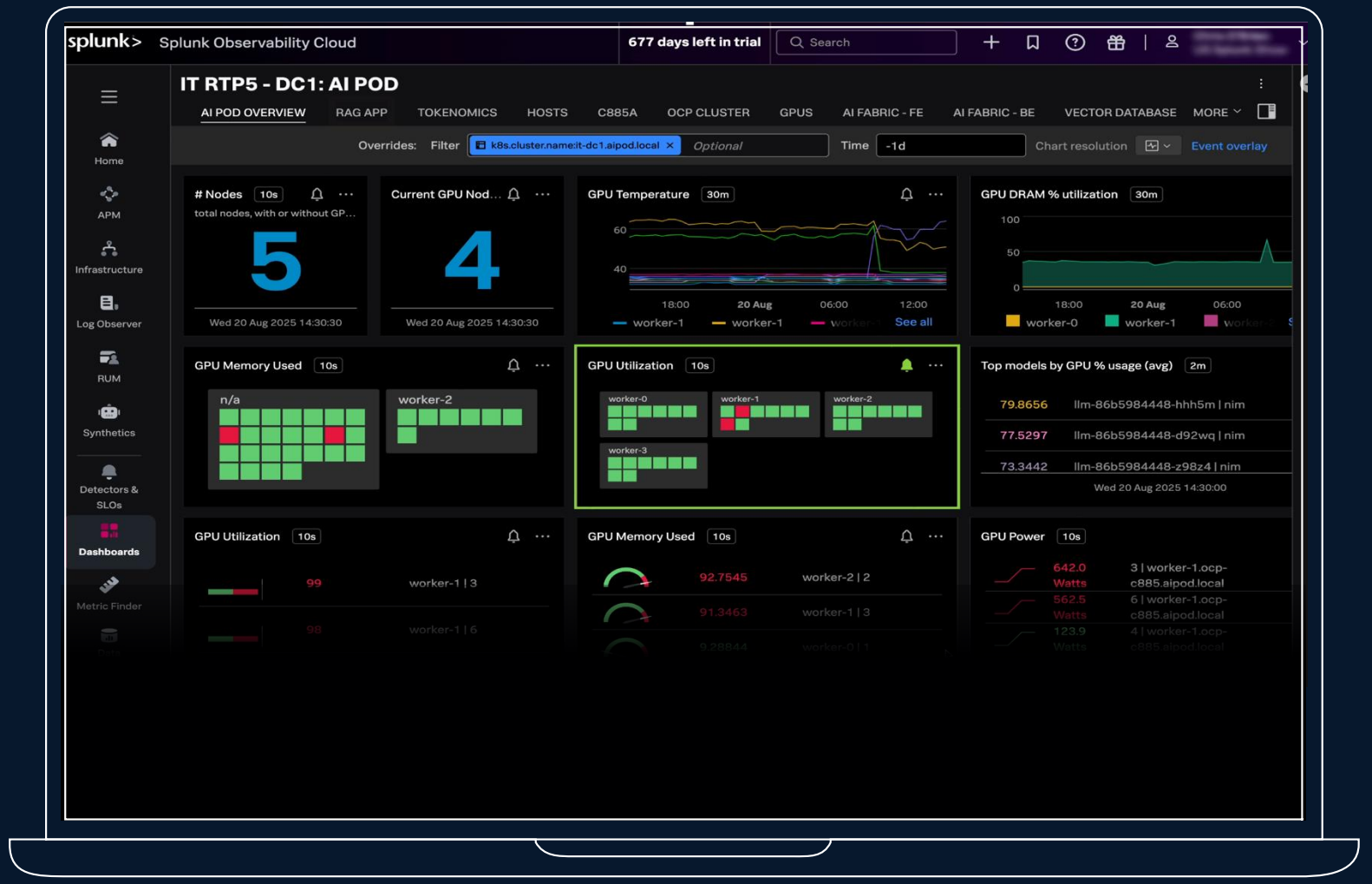
Own and control your data, avoid vendor lock-in and instrument only once on a common standard as you build new applications.

AI powered analytics and guidance

AI/ML driven features like Service Maps and Trace Analytics provide directed guidance that helps you resolve issues faster.

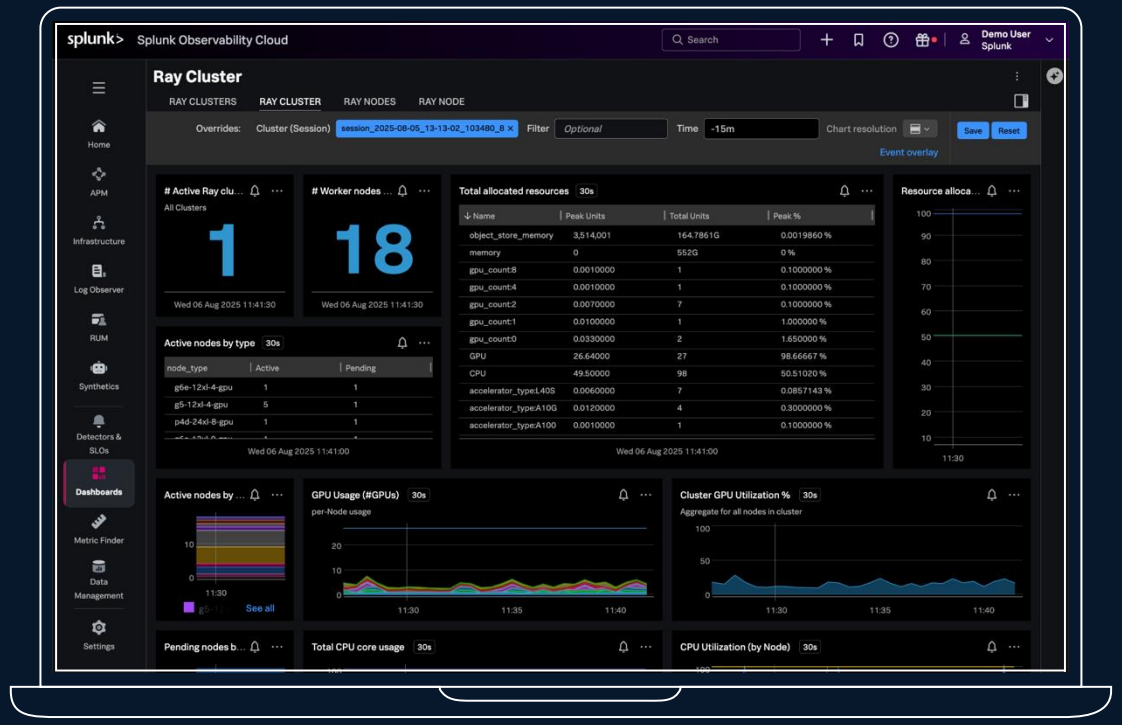
No data sampling

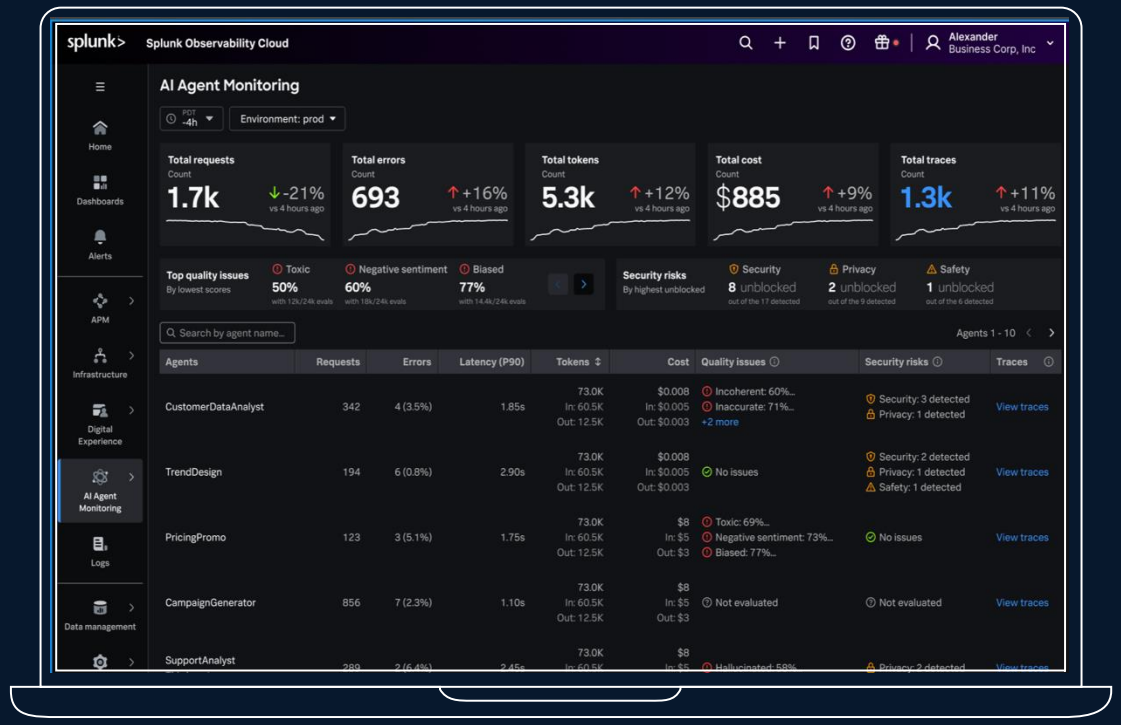
Eliminate blind spots by collecting and analyzing 100% of your data with Splunk's NoSample™ tracing.



Observability for AI

Monitor the health, performance, security, and cost of your AI application stack





AI Infrastructure Monitoring (GA)

To monitor the health, availability, and consumption of AI infrastructure

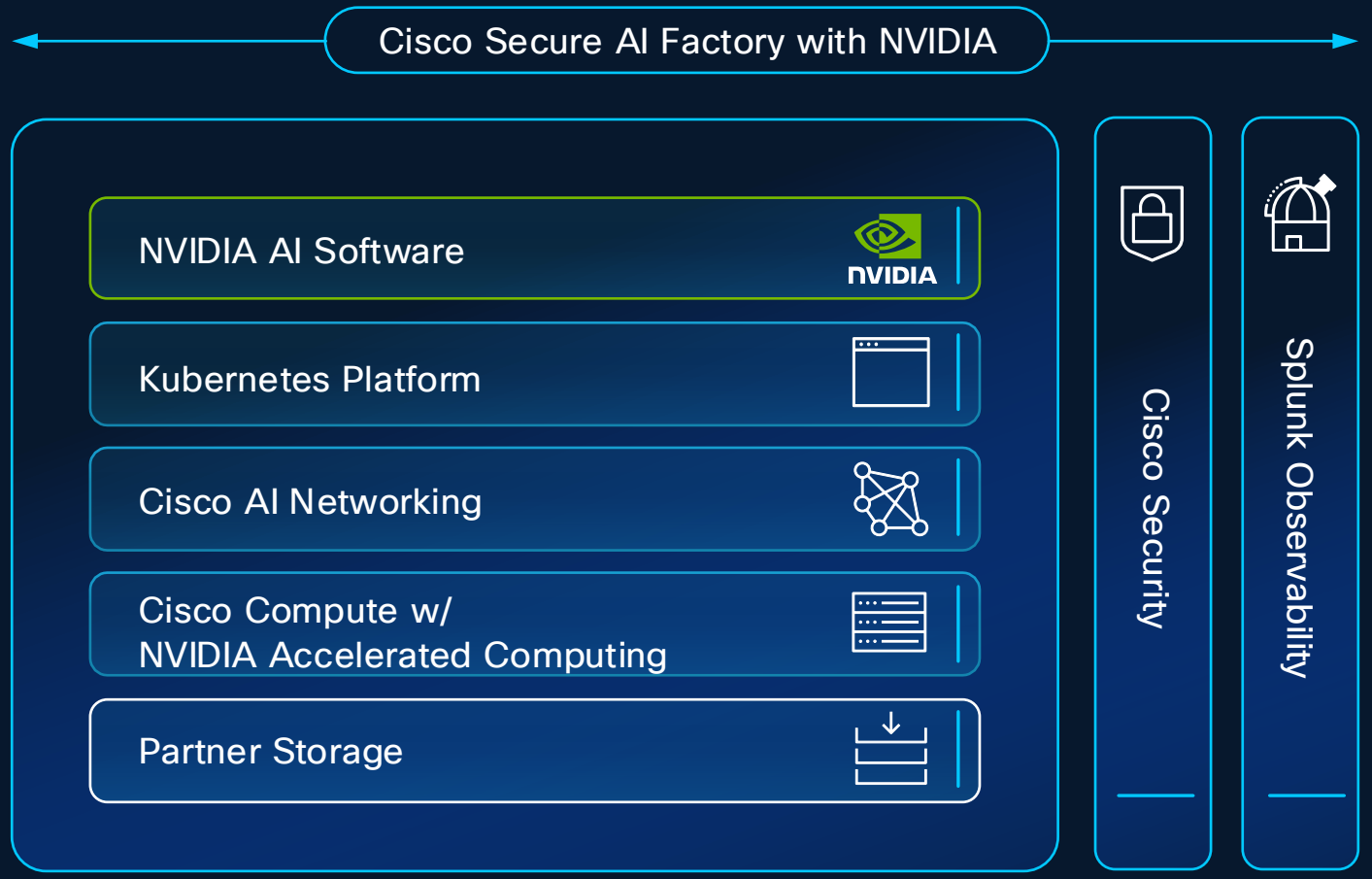
AI Agent Monitoring

To monitor the performance, quality, security, and cost of LLM and agentic applications

Cisco Secure AI Factory with NVIDIA

Delivering Trusted AI Outcomes

A modular reference design that combines high-performance infrastructure with full-stack security and observability



Thank you


CISCO