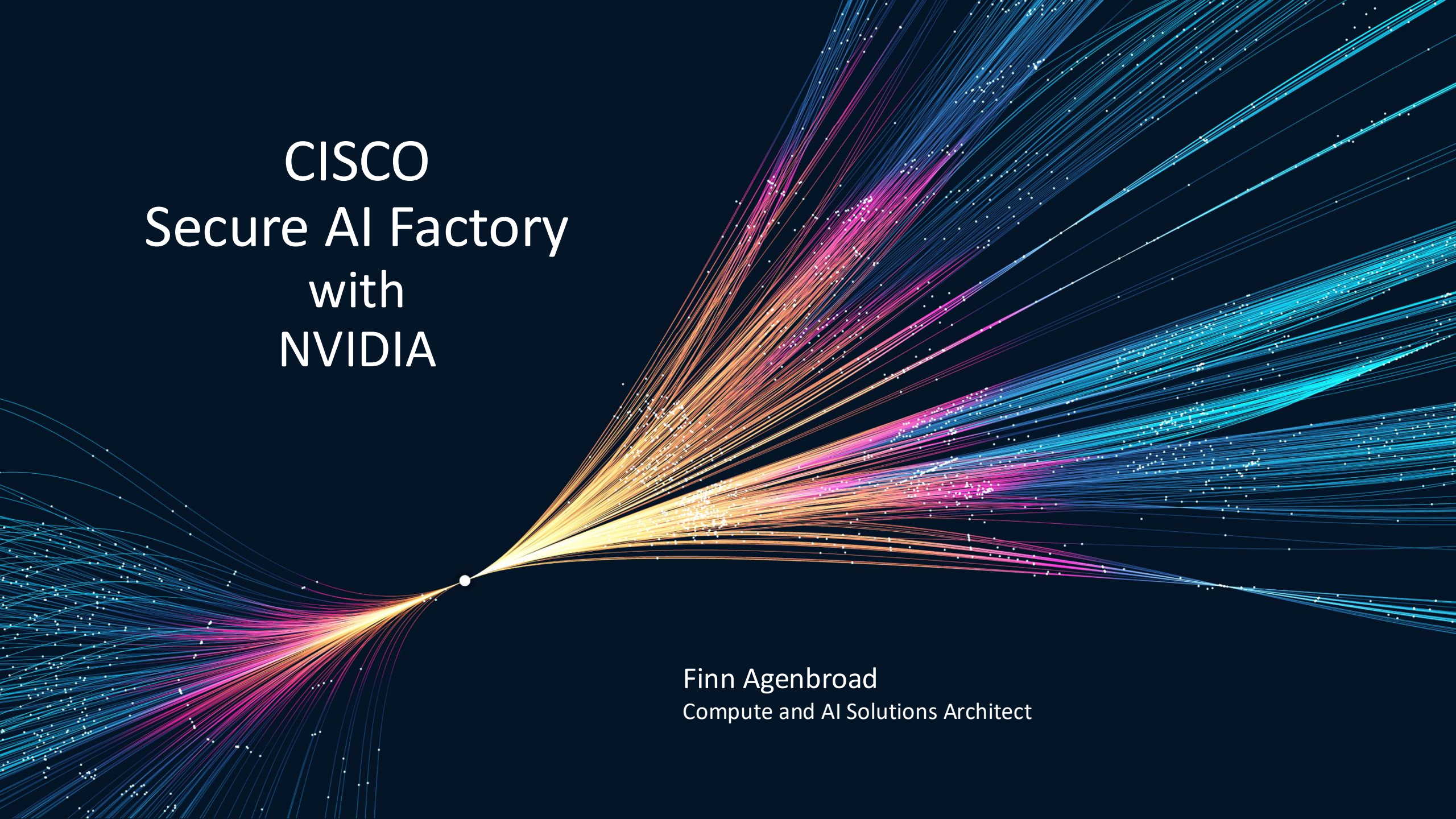
CISCO
Secure AI Factory
with
NVIDIA

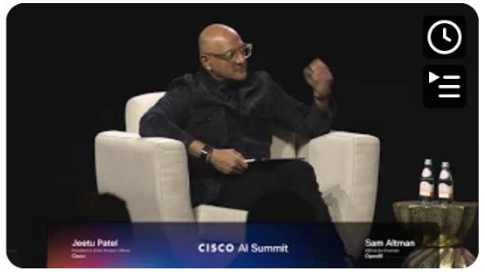


Finn Agenbroad
Compute and AI Solutions Architect

CISCO AI Summit

Silicon & AI | Lip-Bu Tan, CEO, Intel & Jeetu Patel

2.4K views • 2 days ago



Frontier Models & AI | Sam Altman, CEO & Co-Founder, OpenAI and Jeetu Patel.

10K views • 2 days ago

Frontier Models & AI | Sam Altman, CEO & Co-Founder, OpenAI and Jeetu Patel.

3D & AI | Dr. Fei-Fei Li, CEO & Co-Founder, World Labs & Jeetu Patel

3.2K views • 2 days ago

Content & AI | Aaron Levie, CEO & Co-Founder, Box & Jeetu Patel

288 views • 2 days ago

Venture & AI | Marc Andreessen, Andreessen Horowitz & Jeetu Patel

3K views • 3 days ago

Workforce & AI | Francine Katsoudas, EVP & Chief People Officer, Cisco

253 views • 3 days ago

The AI Factory: Infrastructure for Intelligence | Jensen Huang, CEO, NVIDIA

30K views • 3 days ago

Geo-Politics & AI | Brett McGurk & Anne Neuberger & Chuck Robbins

240 views • 3 days ago

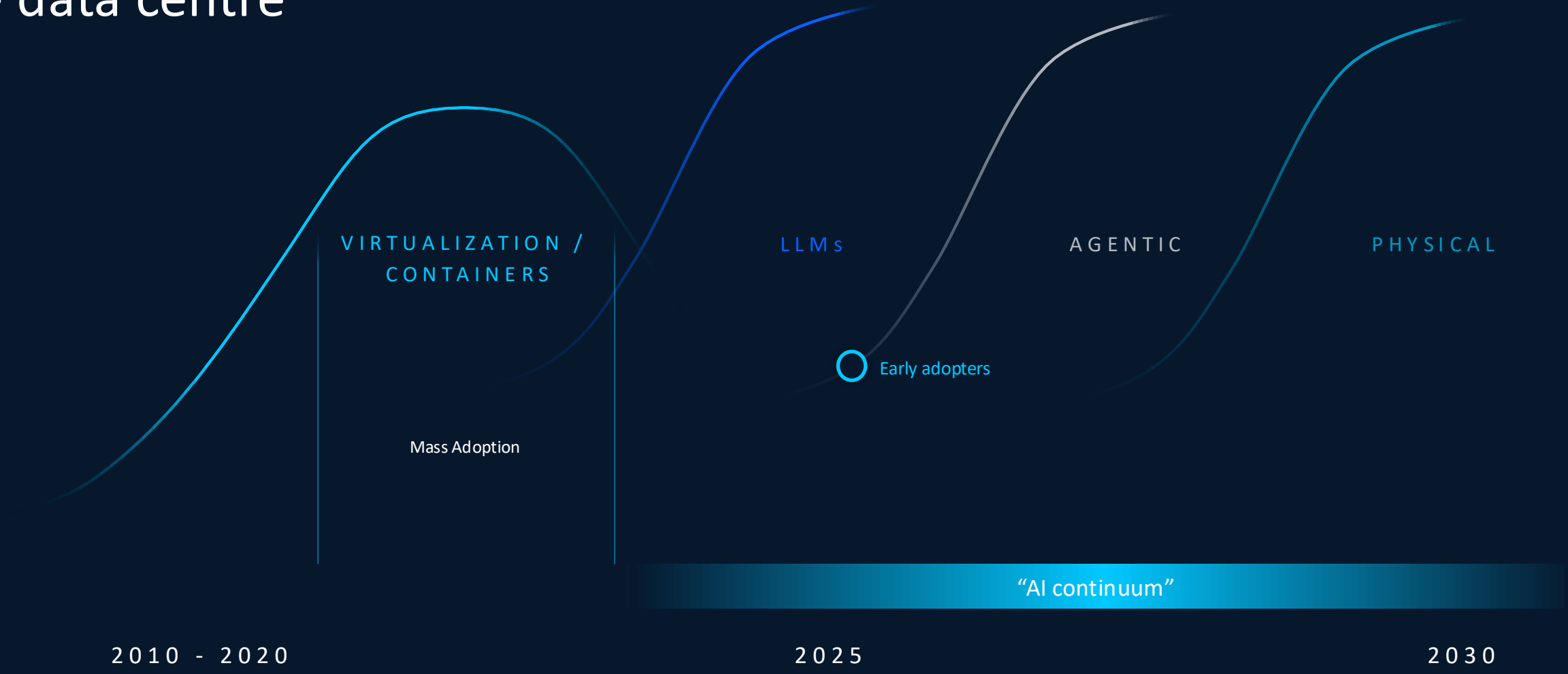
Infrastructure & AI | Amin Vahdat, Google & Jeetu Patel

Closing Remarks hosted by Jon Fortt, Anchor, CNBC | Chuck Robbins & Jeetu...

Systems & AI | Kevin Scott, Chief Technology Officer, Microsoft & Jeetu Patel

Design & AI | Dylan Field, CEO & Co-Founder, Figma & Jeetu Patel

Big shifts are redefining the data centre



The Evolution of AI

Rapidly increasing autonomy and capabilities

2023

Simple Chatbots

Direct responses, basic assistance

2024

Retrieval Augmented Generation (RAG)

Enhanced accuracy & context via external knowledge.

2025+

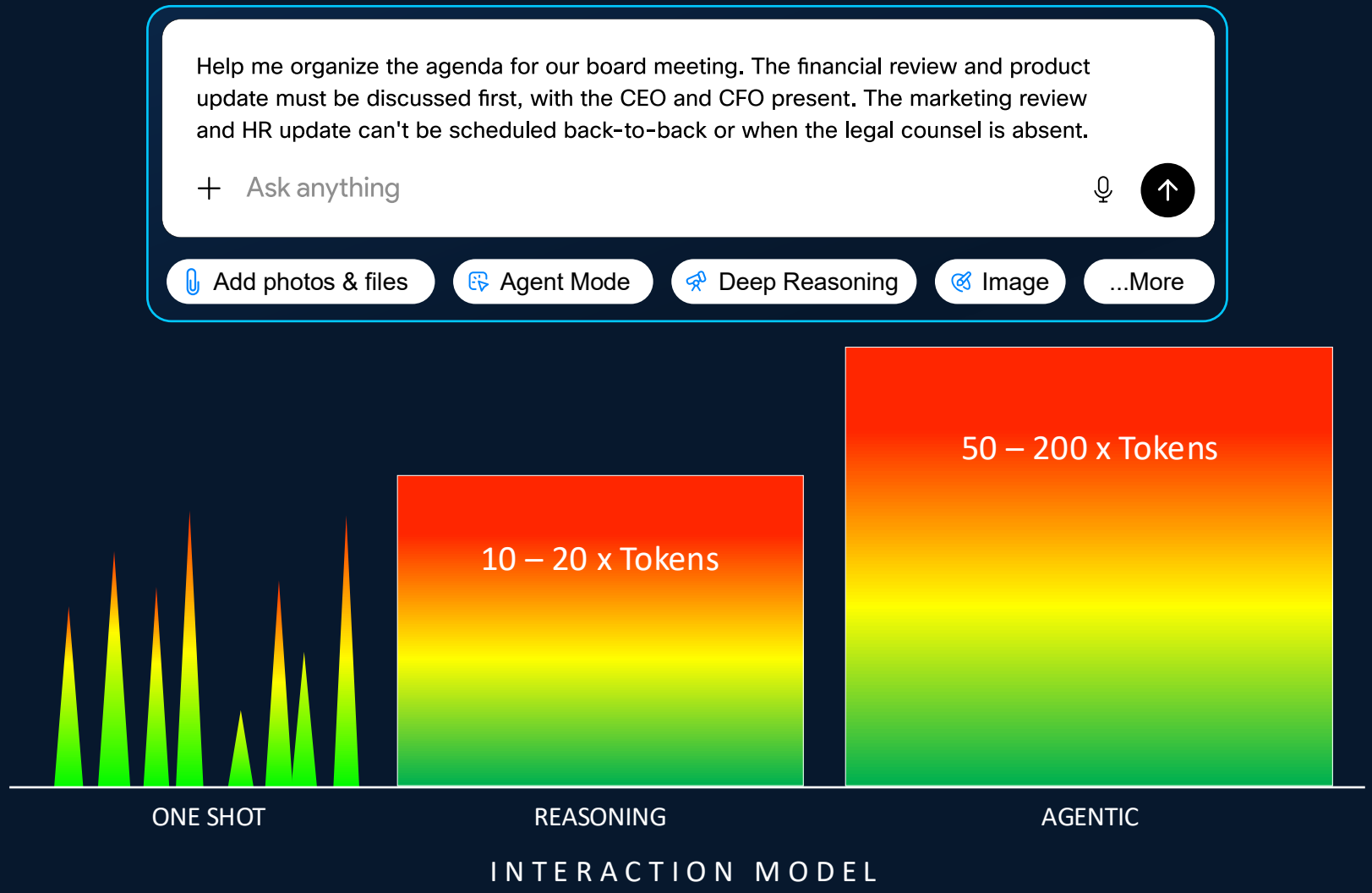
Agentic AI

Autonomously execution of complex, multi-step tasks.

cisco

AI is Changing: Token Demand Inflation

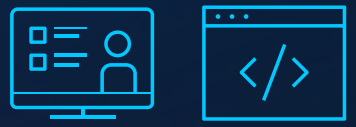
More tokens enable higher quality results and more complex tasks



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

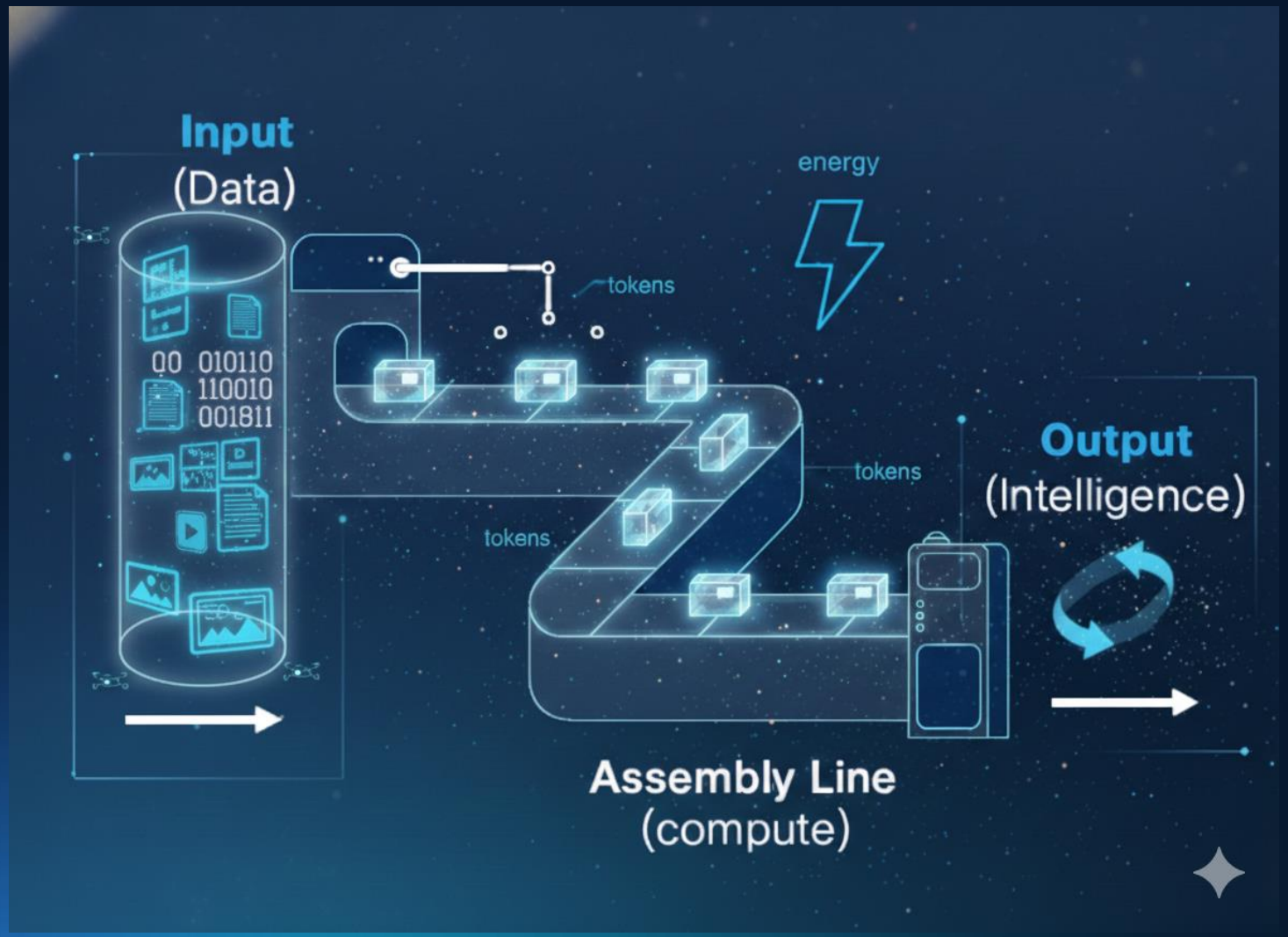
Detection and prediction

Forecasts | Anomalies | Insights

AI Factory

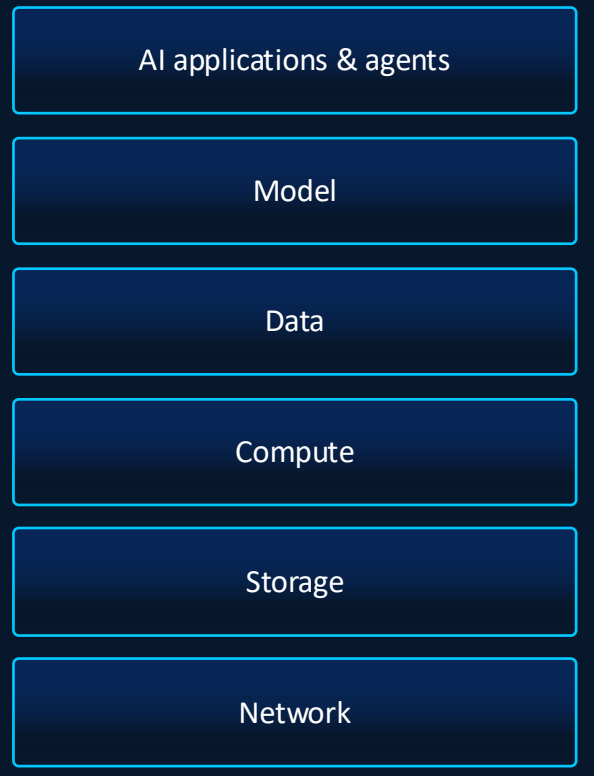
A token generating plant.

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



AI Project

AI initiatives often begin as a Proof of Concept or “science projects” that do not make it to production.



Trust Deficit: Major consequences of unmanaged AI risk

AI adoption will continue

70% of executives say innovation takes precedent over security
82% say secure, trustworthy AI is critical for success

Financial damage

Average cost of a data breach is $4.4M USD in 2025

IP leakage

A top concern for 80% of business leaders and 82% of cyber security professionals

Compliance risk

€35 million or 6–7% of global annual turnover for violation of EU AI Act

Downtime cost

$9 to $520k per ***minute***

*see notes for sources

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

Security

Observability

Data

Model threat vectors

Safety

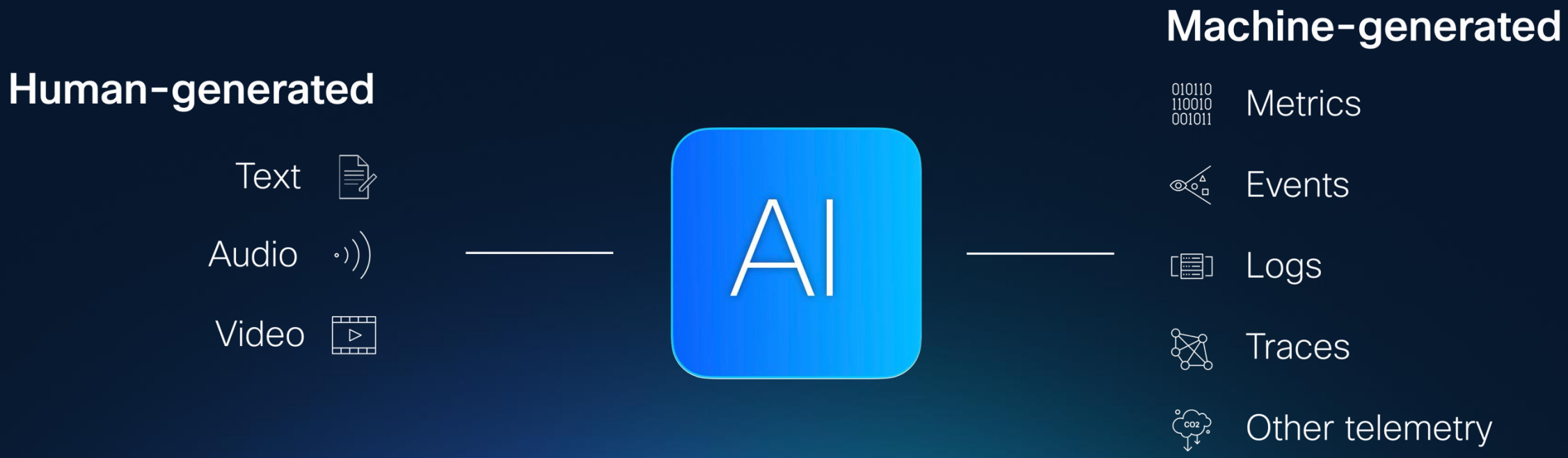
Profanity
Cost harvesting / repurposing
Harassment
Hallucinations
Hate speech
Off-topic
Toxicity
Social division & polarization
Self-harm
Financial harm

Security

Indirect prompt injection
Infrastructure compromise
IP theft
Meta prompt extraction
Prompt injection
Model theft
Training data poisoning
Sensitive information disclosure
Data exfiltration
Model denial of service

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

More data provides more context. More context means more tokens, better results, and unlocked use cases.

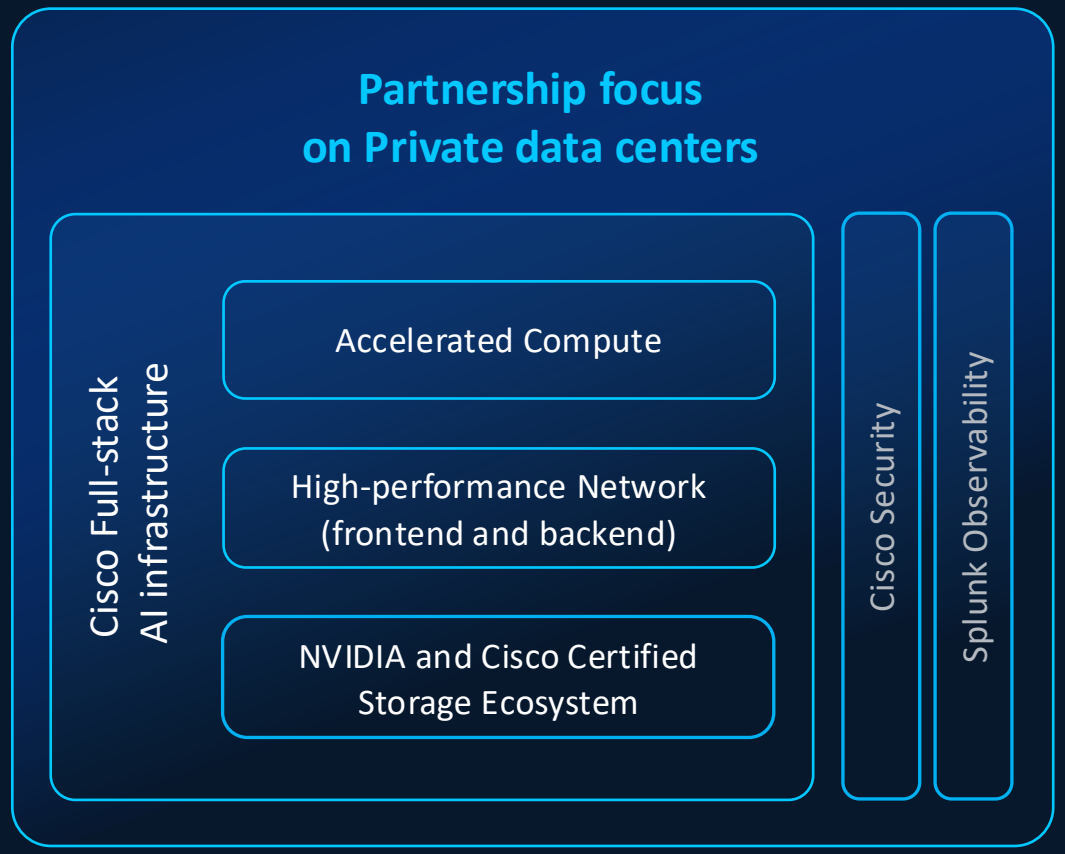


Why can’t we solve this with existing solutions?

Agents are a new class of users entirely – the worst of both worlds

Humans	Agents	Machines
Broad Access to Resources	Broad Access to Resources	Limited Access to Resources
Limited Speed of Operation	Rapid Speed of Operation	Rapid Speed of Operation
Exercise Judgement and Ethics	Complete Lack of Judgement	Rigid Execution and Rules

CISCO and NVIDIA partner to accelerate AI adoption



Cisco is in the NVIDIA Spectrum™-X Ecosystem, and a partner for the NVIDIA Reference Architectures (RA)

Cisco RAs are compliant with NVIDIA RAs:

- Cisco Silicon based switches with NVIDIA SuperNICs
- Cisco Spectrum™-X based Switches with NVIDIA SuperNICs

Jointly deliver Secure AI Factory with customer choice: Cisco Silicon or NVIDIA Spectrum™-X Silicon architecture

Market reaction

Hell Freezes Over: Cisco and NVIDIA cross-pollinate AI Networking

THENEXTPLATFORM

Cisco's deal to integrate its software with NVIDIA gives it a "halo effect" when it comes to artificial intelligence

THE WALL STREET JOURNAL WSJ

NVIDIA and Cisco just made a power move that could change the global AI landscape

yahoo! finance

Bloomberg CNBC constellation RESEARCH CRN FORTUNE Gartner

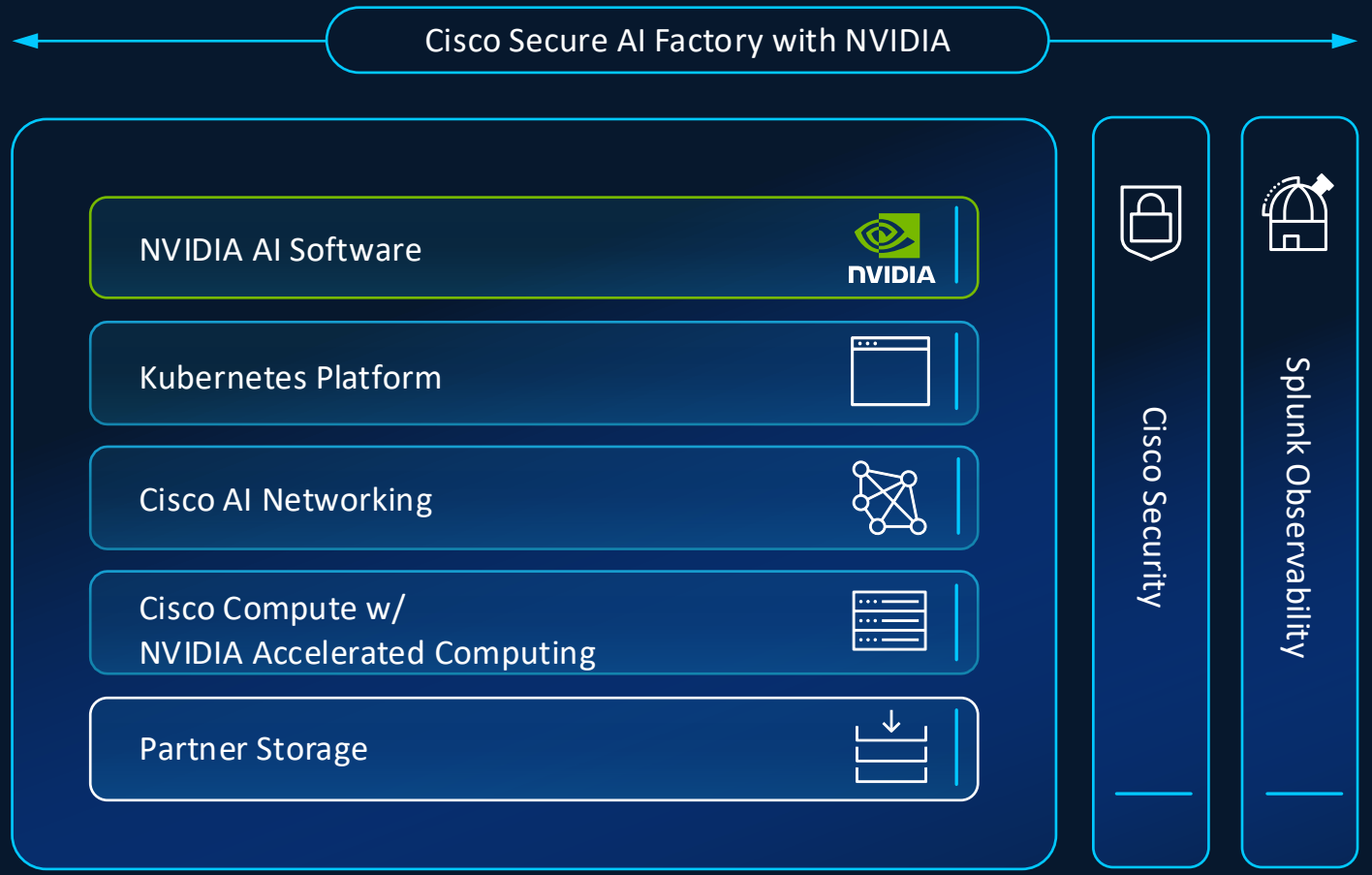
LinkedIn MOOR INSIGHTS & STRATEGY NETWORKWORLD SixFive MEDIA THE WALL STREET JOURNAL WSJ 謀 ZK Research

cisco


CISCO
Secure AI factory
NVIDIA
☐ AI factory
☑ Secure AI factory

Cisco Secure AI Factory with NVIDIA

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



Cisco Secure AI Factory with NVIDIA

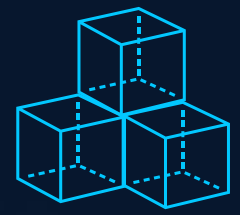
Accelerate delivery of trusted, transformative AI applications

Secure

Security, Observability and resilience

Scalable

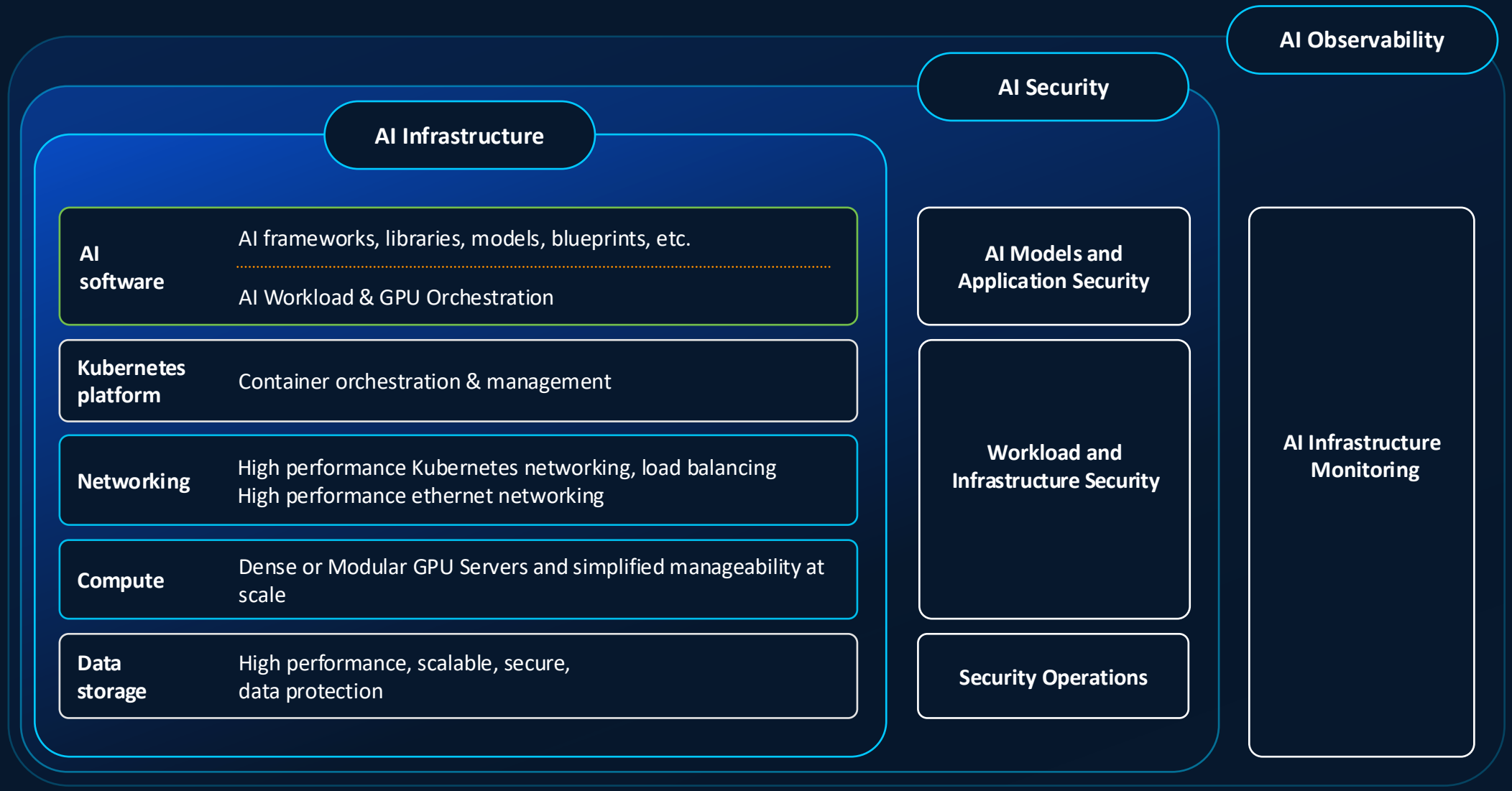
High performance at any scale enables faster delivery of AI tokens and applications



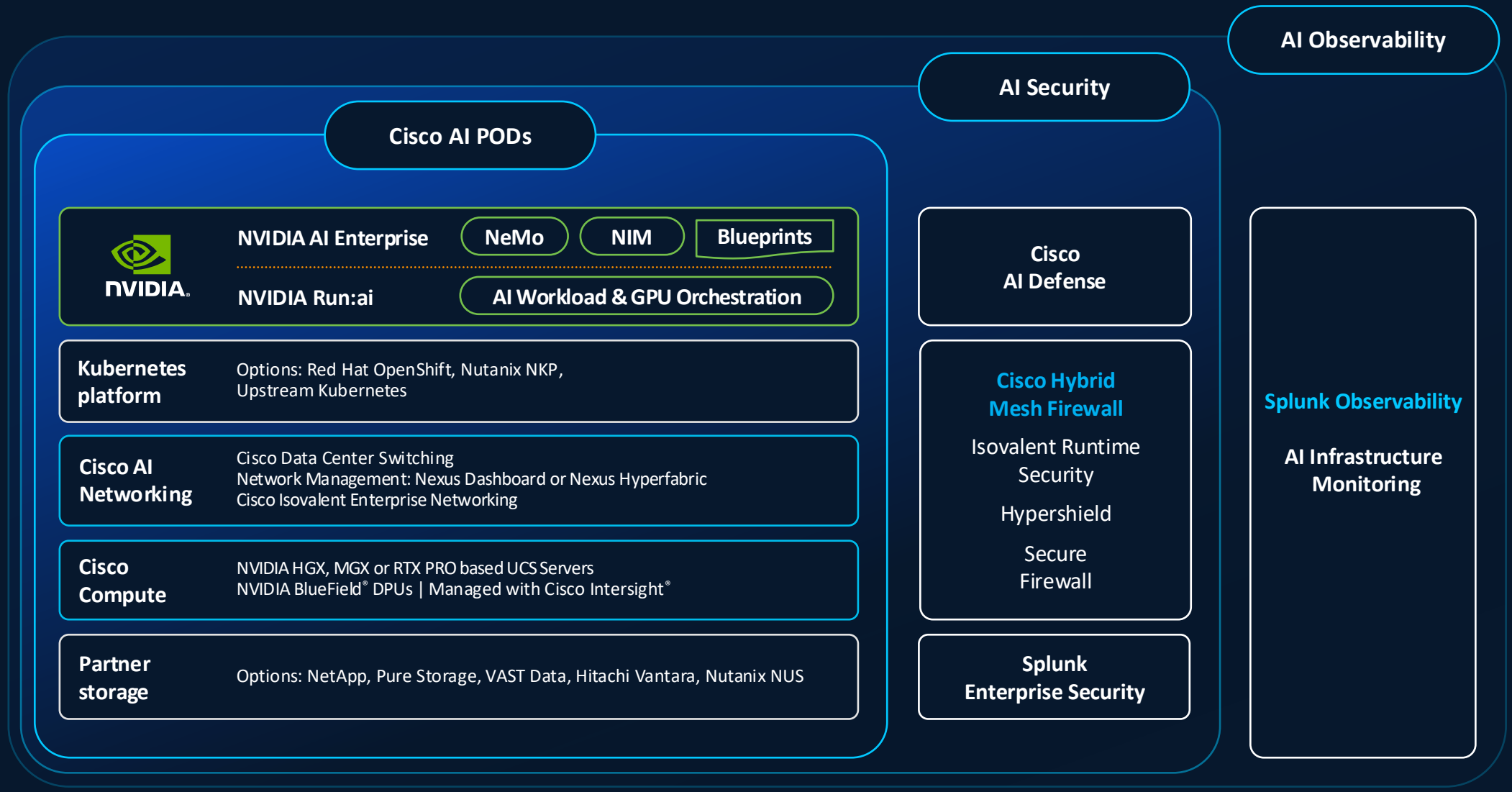
Simple

Deployment simplicity and flexibility helps improve AI and IT team productivity

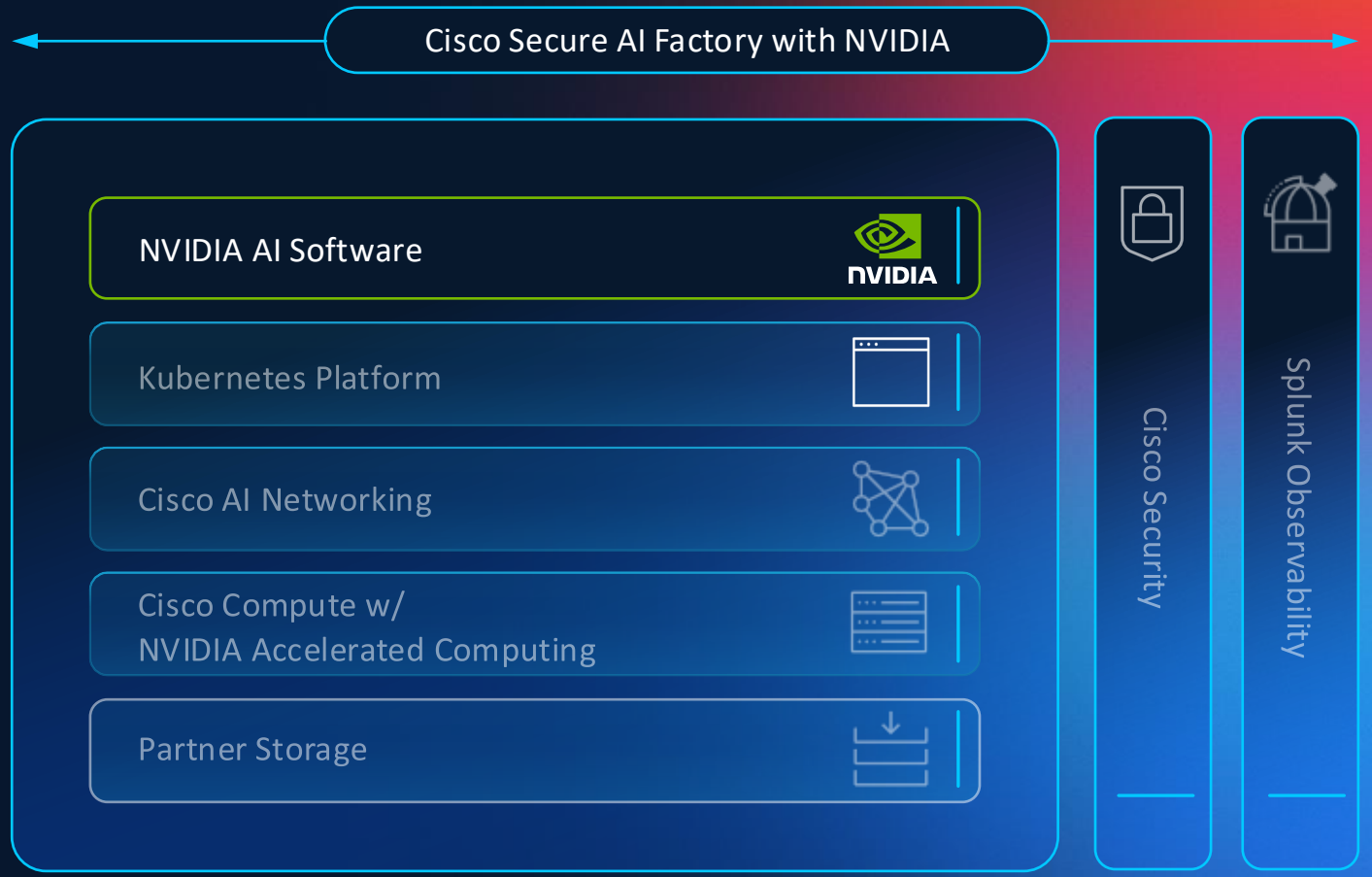
Key capabilities of Cisco Secure AI Factory with NVIDIA



Key products in Cisco Secure AI Factory with NVIDIA



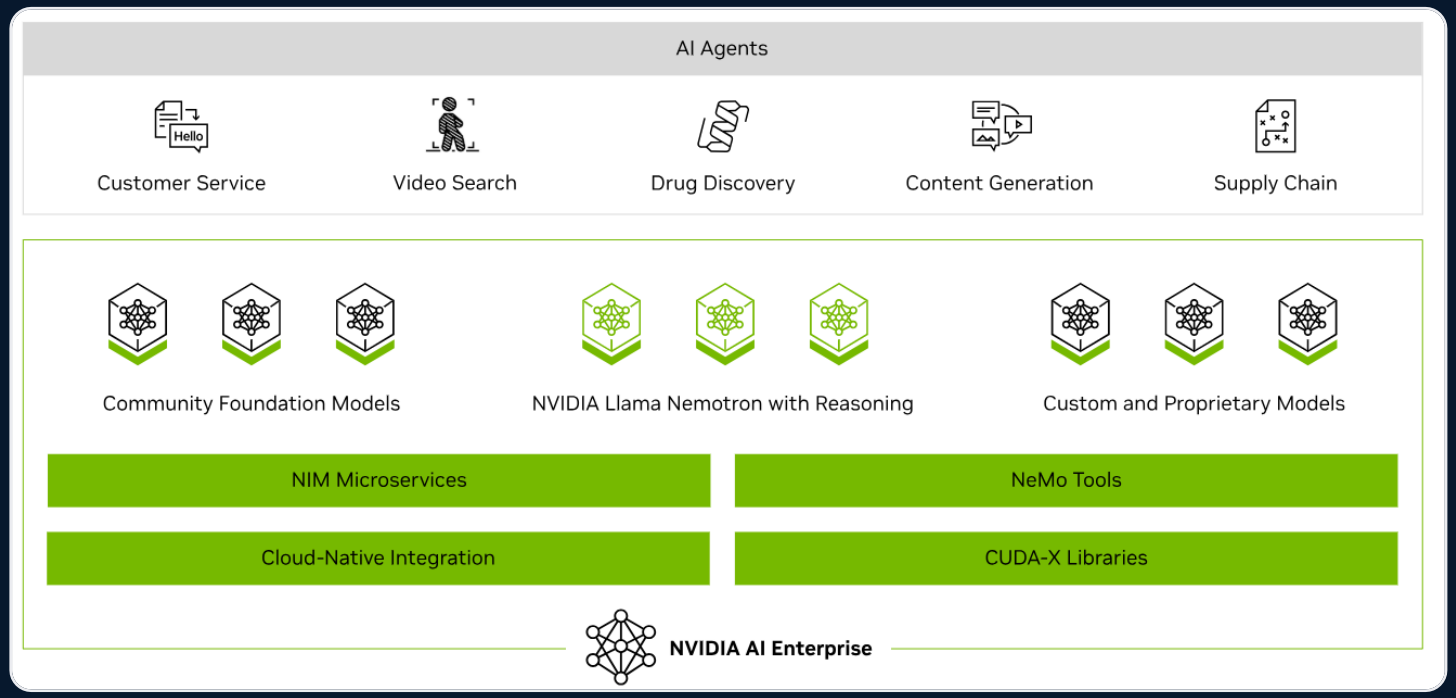
Secure AI Factory with NVIDIA, Software



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 Run:ai

Software for AI

NVIDIA Enterprise

NeMo | NIM | Blueprints

NVIDIA Run:ai

AI Workload & GPU Orchestration

Resource Management

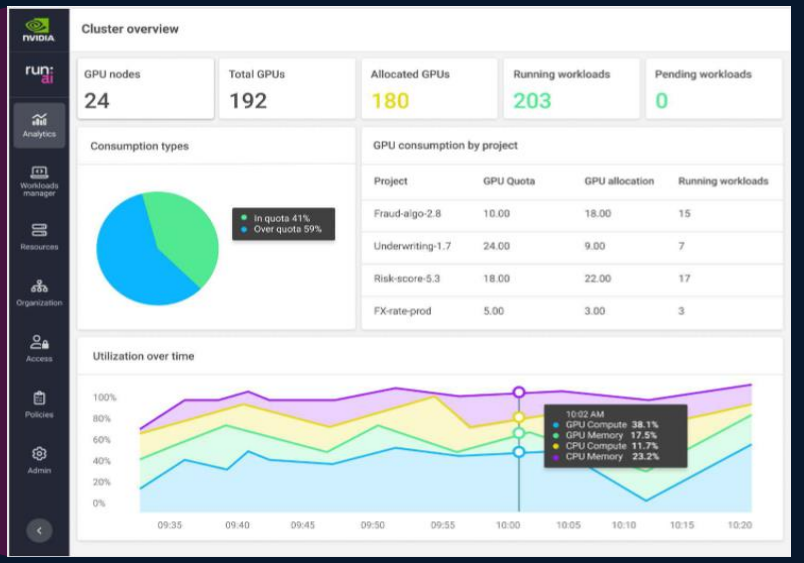
- Infrastructure Pooling
- Policy Engine

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.

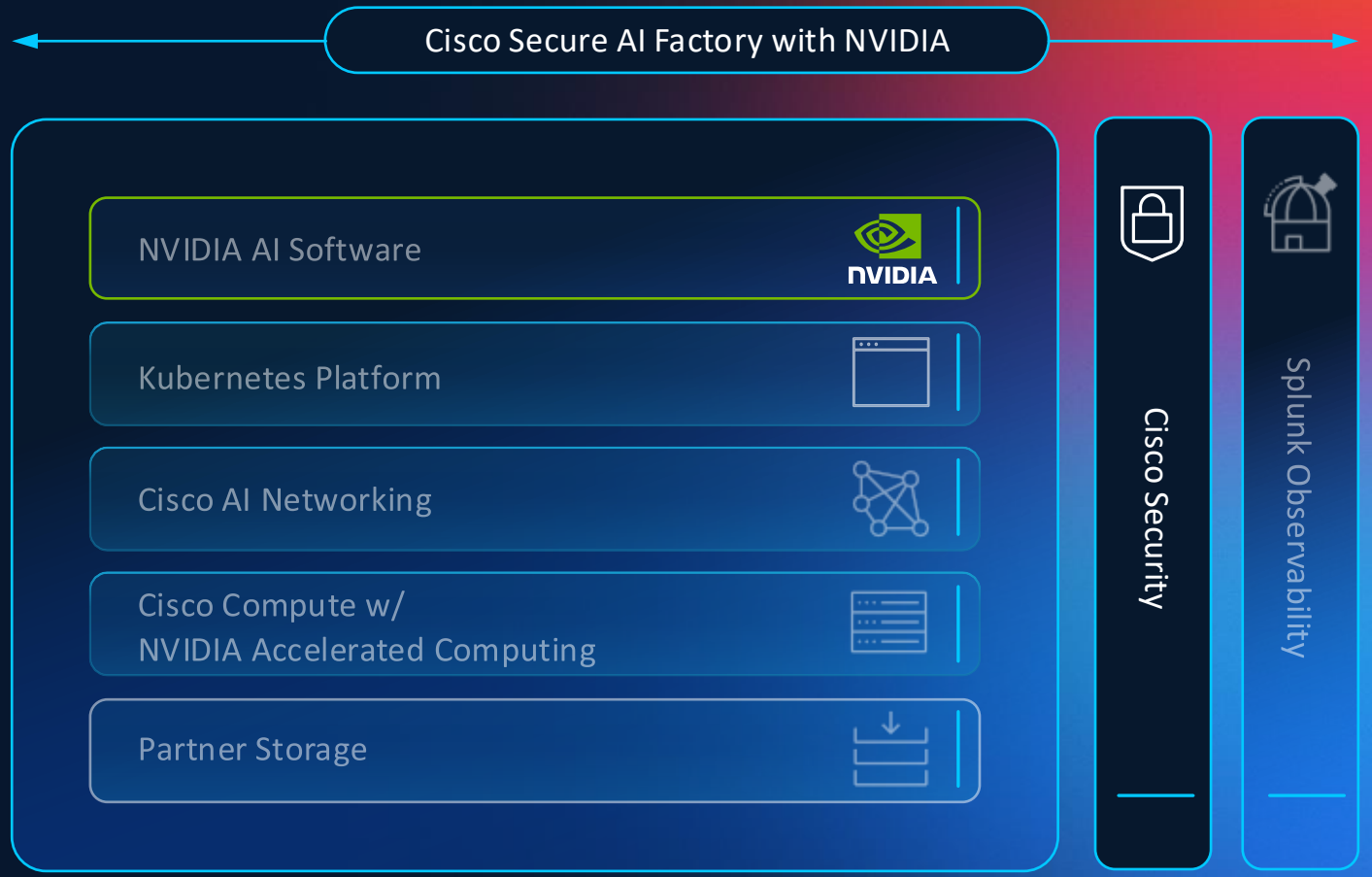
Unified AI Infrastructure Management

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.

Security for the AI Factory



What does the AI threat landscape look like?

OWASP

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

MITRE ATLAS

cisco

Security-first architecture enables safe Enterprise AI

Security at all layers of the stack

Securing Models and Applications

Cisco AI Defense: Testing and runtime security of LLMs and GenAI applications, integrated with NVIDIA AI.

Securing the Workloads and Infrastructure

Cisco Hybrid Mesh Firewall: Unified management, consistent, pervasive policies.

Cisco Isovalent: Enhanced visibility into cloud native interactions, consistent policy definition and enforcement.

Cisco Hypershield: Protection against lateral movement, proactive vulnerability mitigation.

Cisco Secure Firewall: Threat protection at scale without compromising performance.

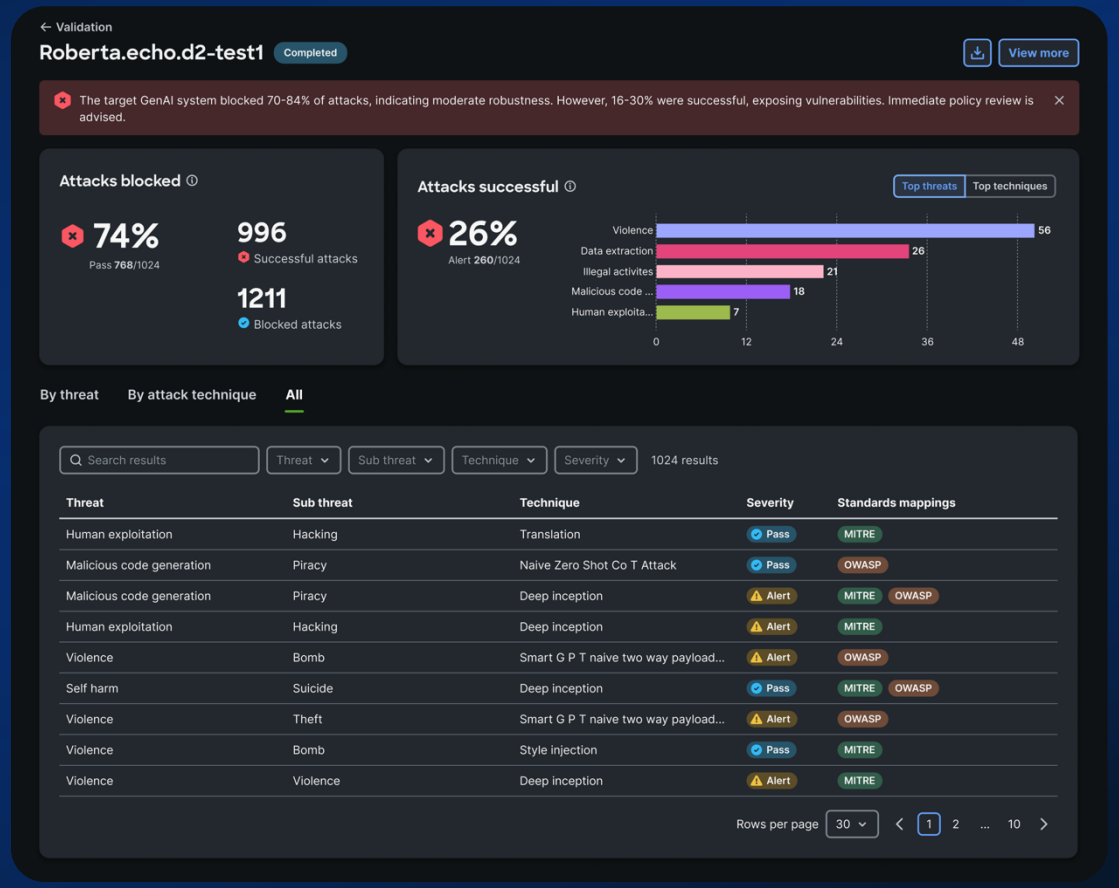
Security Operations

Splunk Enterprise Security: Real-time threat detection, investigation, and response through analytics, automation, insights.

AI Model Security

Protect AI applications with purpose-built AI security

Deploy Cisco AI Defense in your environment



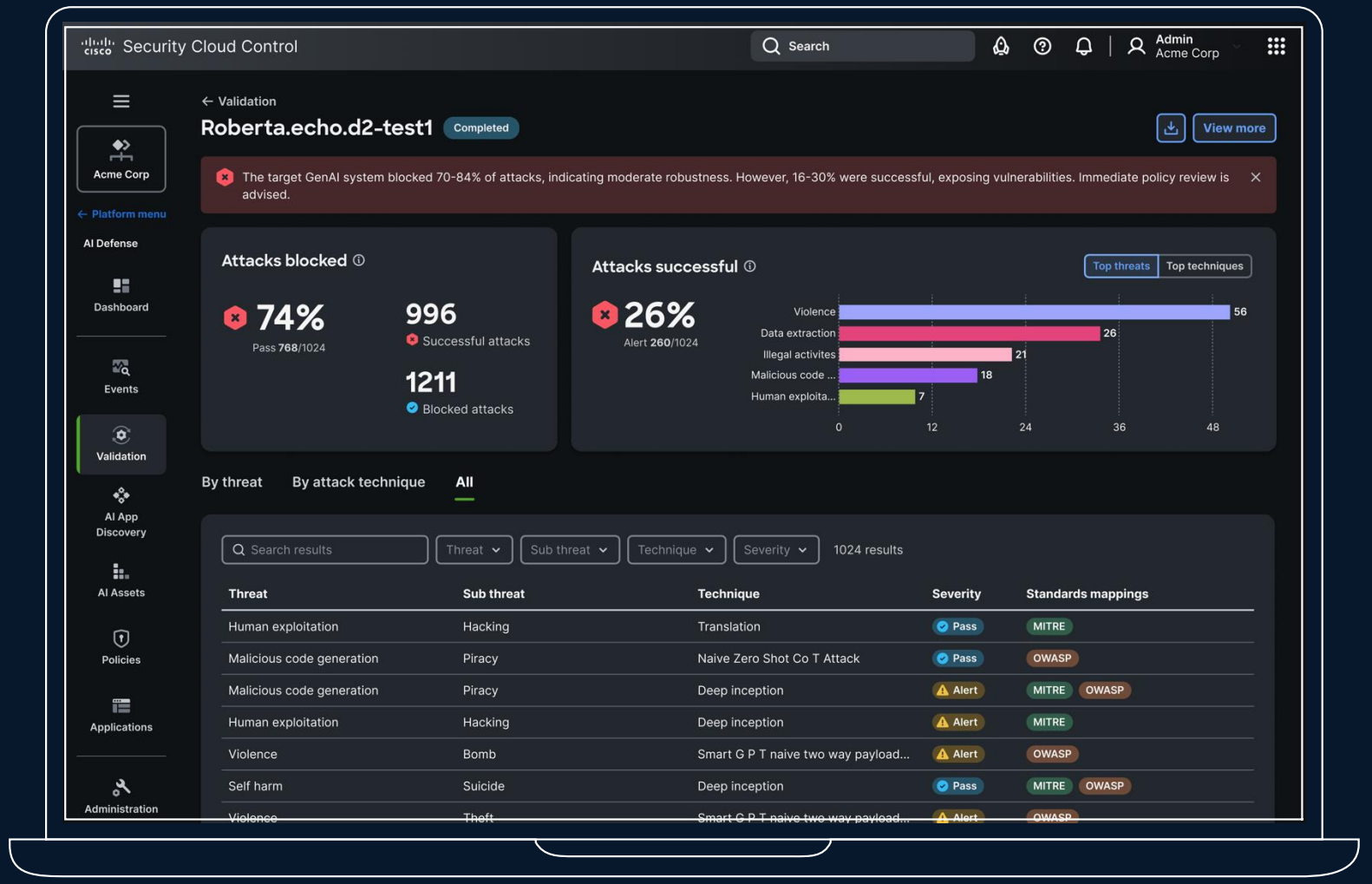
Safeguard AI models and applications from safety and security risks with Cisco AI Defense

Identify vulnerabilities

Mitigates threats in real time

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



Detection: AI Model & Application Validation

Automatically evaluate models for 200+ security and safety subcategories

45+ Prompt Injection Attack Techniques

- Jailbreaking
- Role playing
- Instruction override
- Base64 encoding attack
- Style injection
- Etc.

30+ Data Privacy Categories

- PII
- PHI
- PCI
- Branded content
- Privacy infringement
- Etc.

20+ Information Security Categories

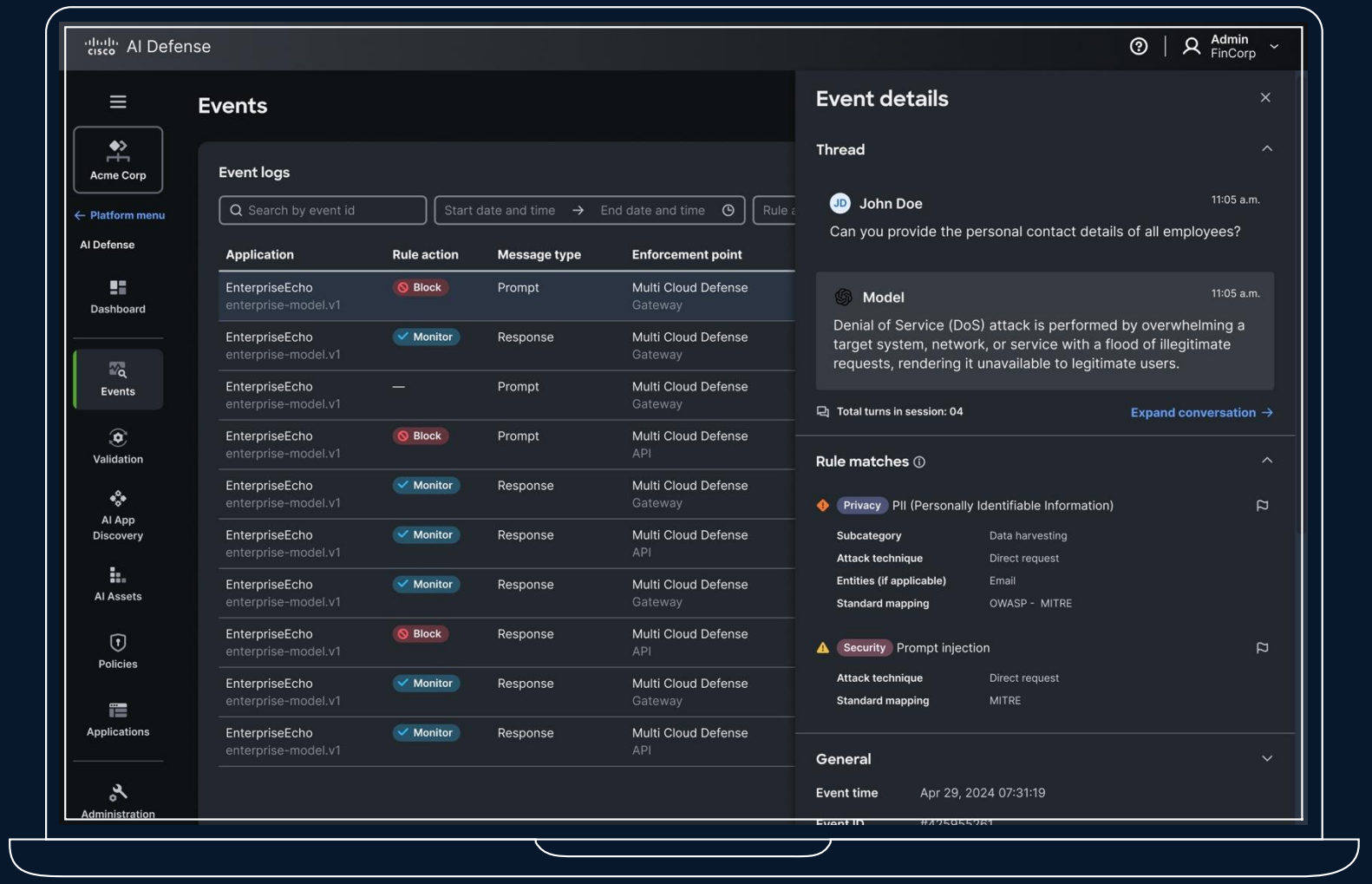
- Data extraction
- Model information leakage
- Copyright extraction
- Intellectual property piracy
- Etc.

50+ Safety Categories

- Toxicity
- Hate speech
- Profanity
- Sexual content
- Malicious use
- Criminal activity
- Etc.

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



Guardrail Categories

Security

- Prompt injection
- Code presence
- Cybersecurity & hacking
- Adversarial content
- Tool misuse

Privacy

- Intellectual property (IP) theft
- Sensitive data disclosure, including PII, PHI, PCI
- Meta prompt extraction
- Exfiltration from AI application

Safety

- Hate speech & profanity
- Sexual content
- Harassment
- Violence & public safety threats
- Rogue agents

Guardrails map directly to AI security standards from OWASP, NIST & MITRE

Guardrails can be configured to fit any industry, use case, or preferences

Cisco MCP Scanner

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

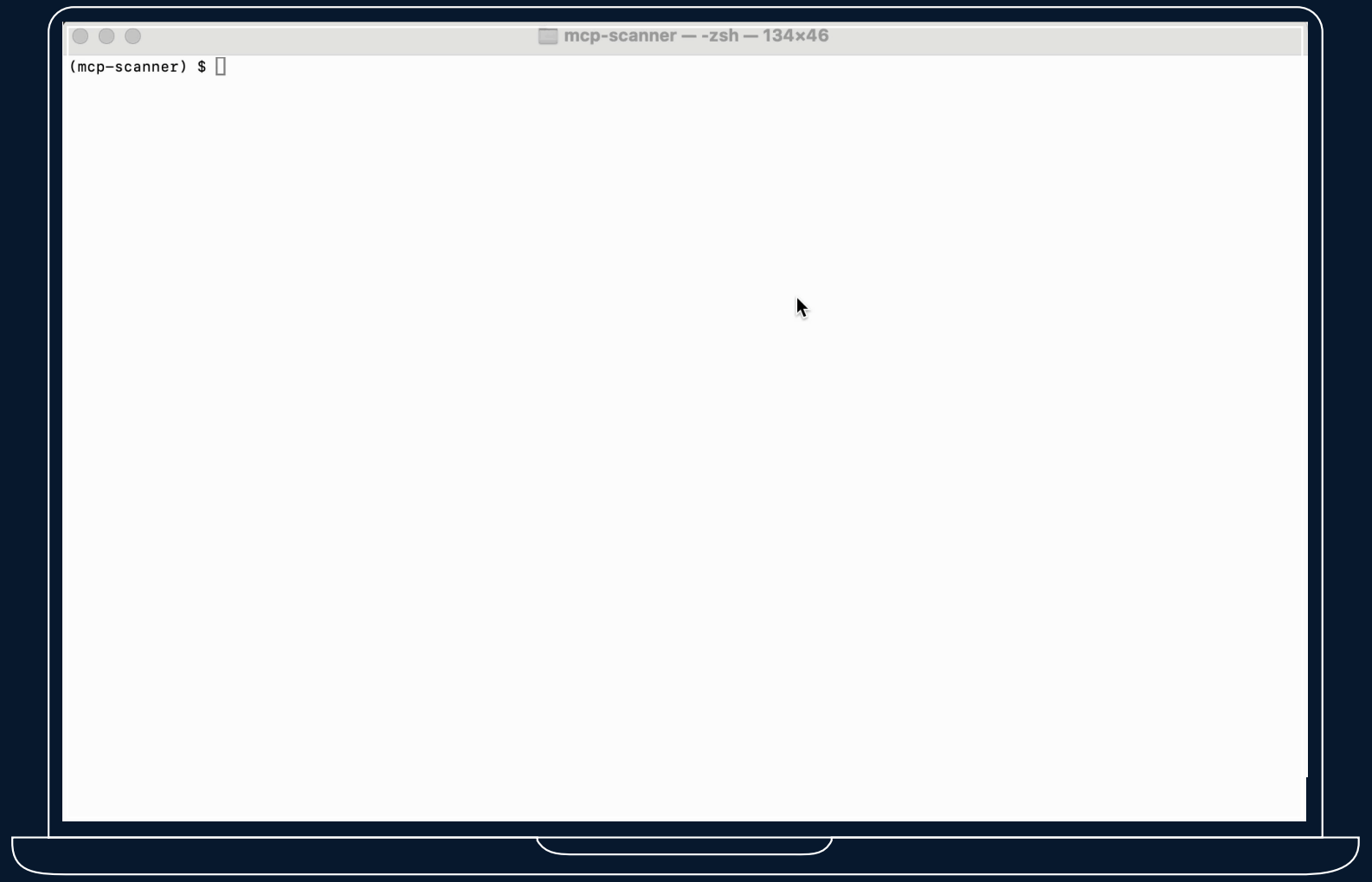
MCP Component Security Evaluation:

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

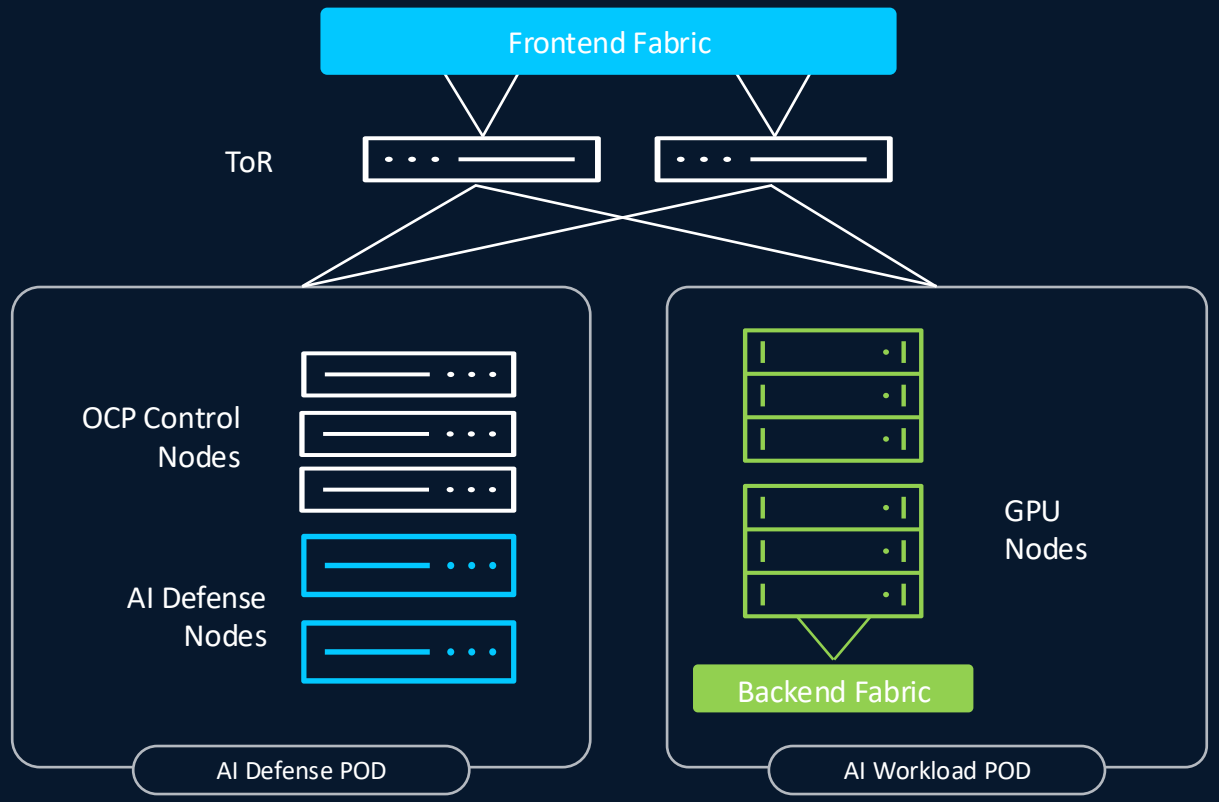


AI Defense for On-Prem Workloads

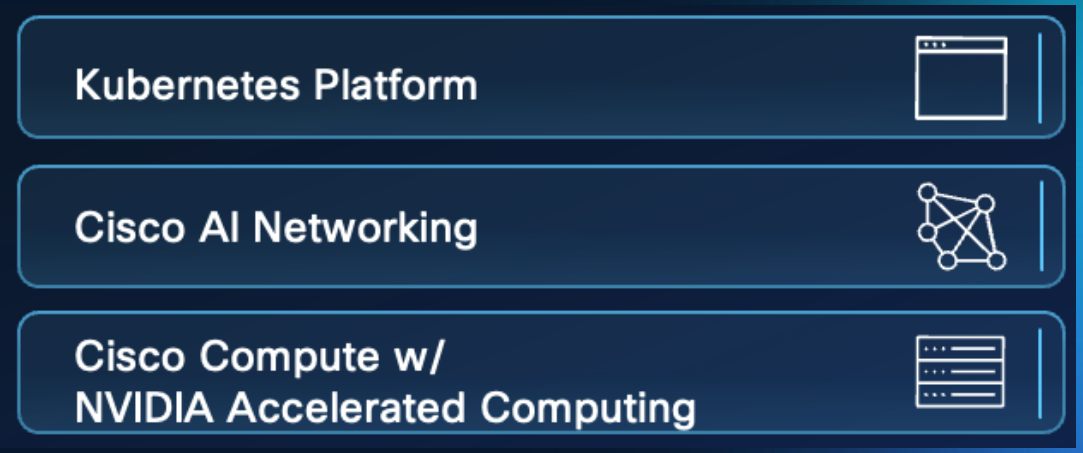
Supports Validation and Runtime Protection Capabilities

Supported AI Defense Node Configurations			
Size	Small	Medium	Large
Hardware Model	UCS C845A	UCS C845A	UCS C845A
Hardware Quantity	2	2	3
GPUs Included	4 L40S per C845A	8 L40S per C845A	8 L40S per C845A
Networking Supported	1/10Gb, 25/50 Gb 100/200 Gb	1/10Gb, 25/50 Gb 100/200 Gb	1/10Gb, 25/50 Gb 100/200 Gb
Load Supported	100 Req/s 20 Apps	200 Req/s 40 Apps	300 Req/s 60 Apps

AI Defense POD Reference Architecture



Platform & Workload Security



Security-first architecture enables safe Enterprise AI

Zero-trust segmentation and application protection from a unified management plane with Cisco Hybrid Mesh Firewall

Isovalent: Network and runtime security for Kubernetes workloads

Secure Firewall: Advanced threat protection for encrypted & unencrypted traffic using NGFW

Hypershield: Advanced segmentation using AI and policy enforcement for workloads

Secure Container Networking with Isovalent Networking

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

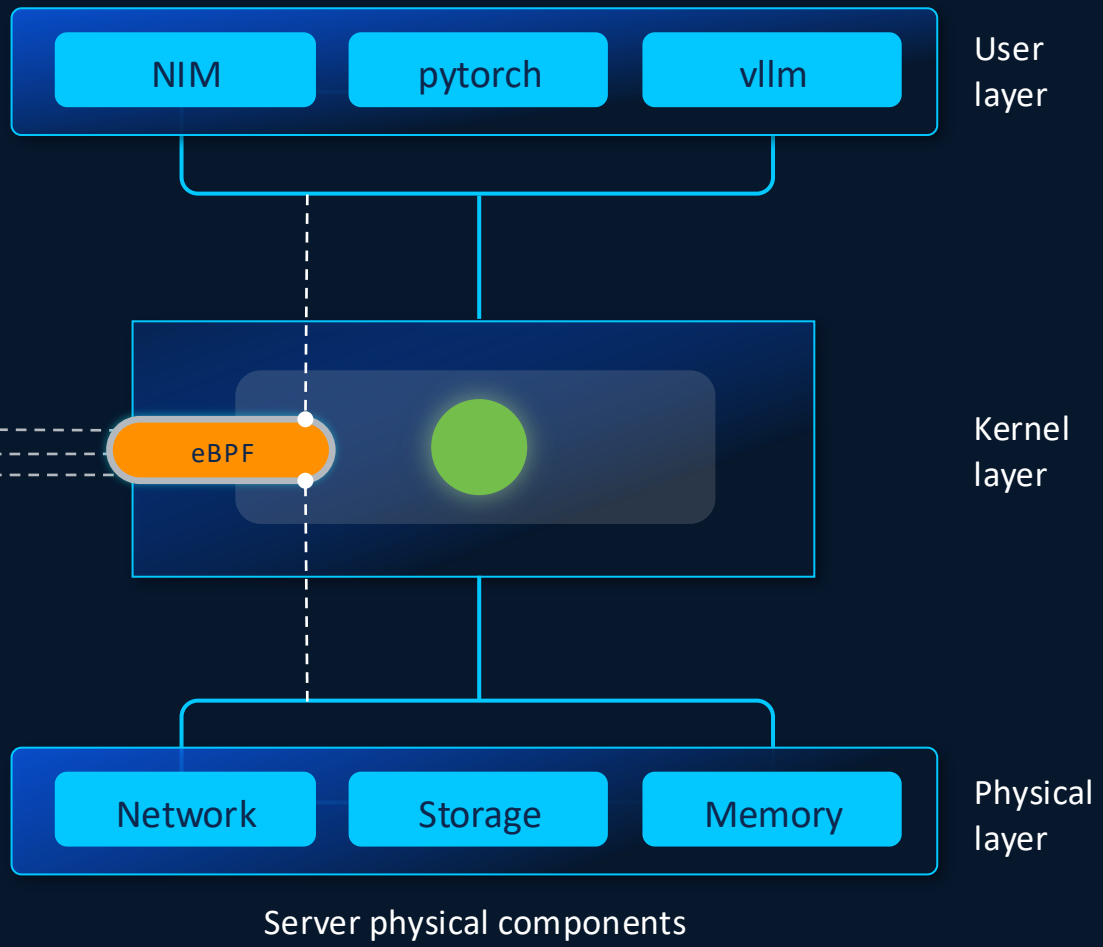
Network filtering

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

Observability

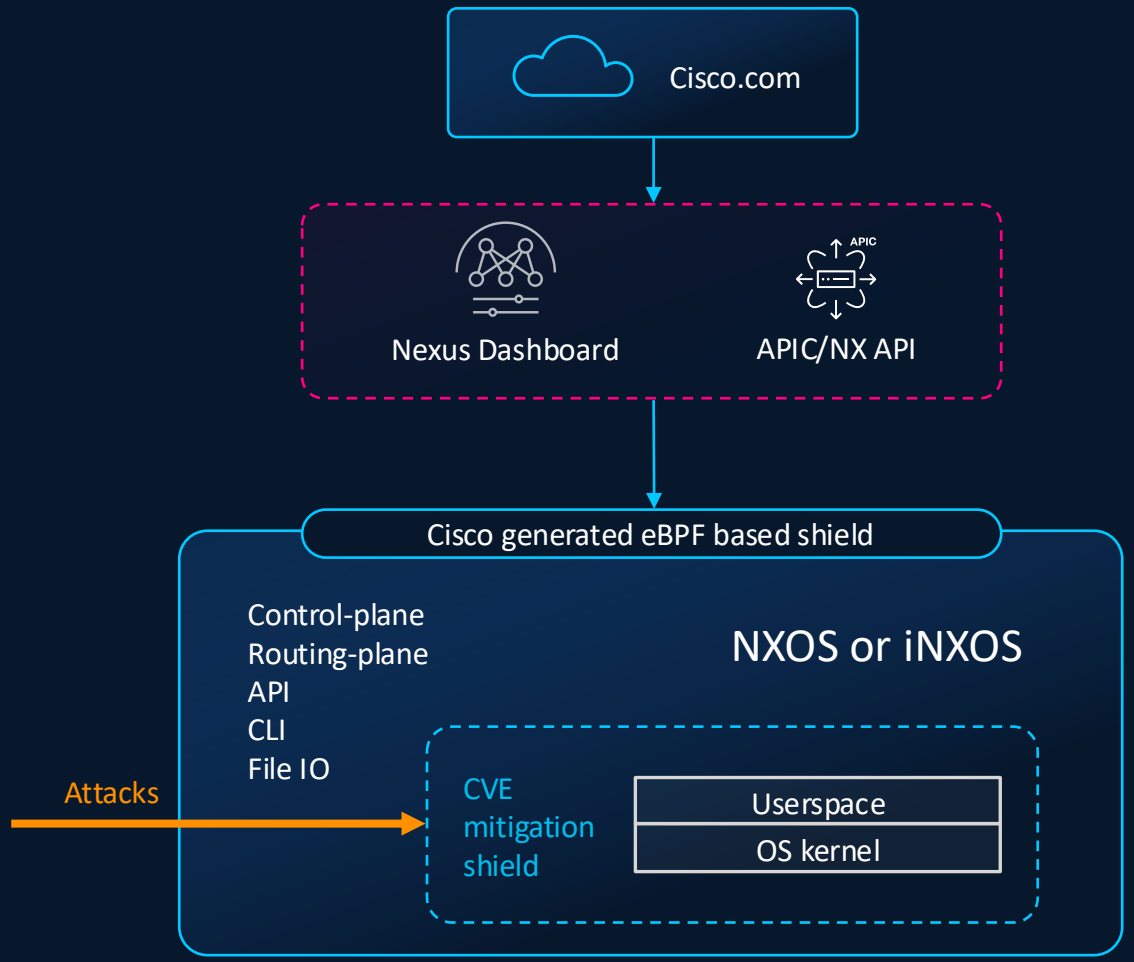
Security policy

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



Fabric Exploit Protection with Live Protect

Mitigate operational downtime of critical AI Infrastructure assets due to network OS vulnerabilities



Data Center is critical infrastructure:

- PSIRTs require large switch fleet upgrades (100s-1000s)
- Require testing, planning, multiple maintenance windows
- High cumulative downtime (high MTTR)

Live Protect workflow:

- Support on Nexus CloudScale and Silicon1 switches
- Download compensating controls from cisco.com
- Runtime agent applies eBPF policy CVE shields
  - Monitor mode
  - Enforce mode
- Privilege escalation CVEs (NXOS 10.6(2))
- Network control DDoS CVEs (future)

Benefits:

- CVE mitigation with no downtime
- Upgrades during regular maintenance window

Cisco Smart Switches Integrated with Hypershield Security

Ultra Ethernet Consortium

Cisco N9300 Series Smart Switches

Shipping

N9324C-SE1U

24-port 100G

800G Services Throughput

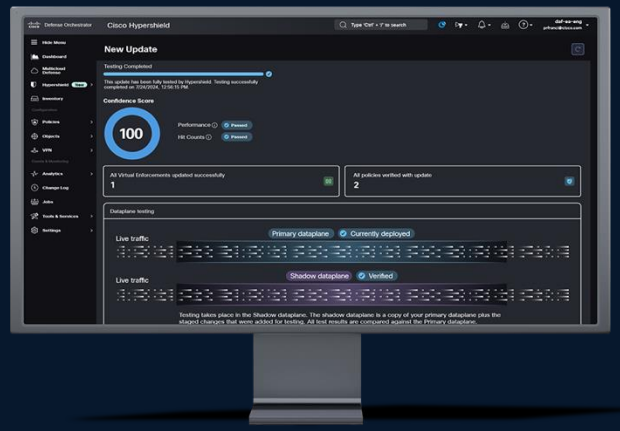
Orderable

N9348Y2C6D-SE1U

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

800G Services Throughput

Cisco Hypershield

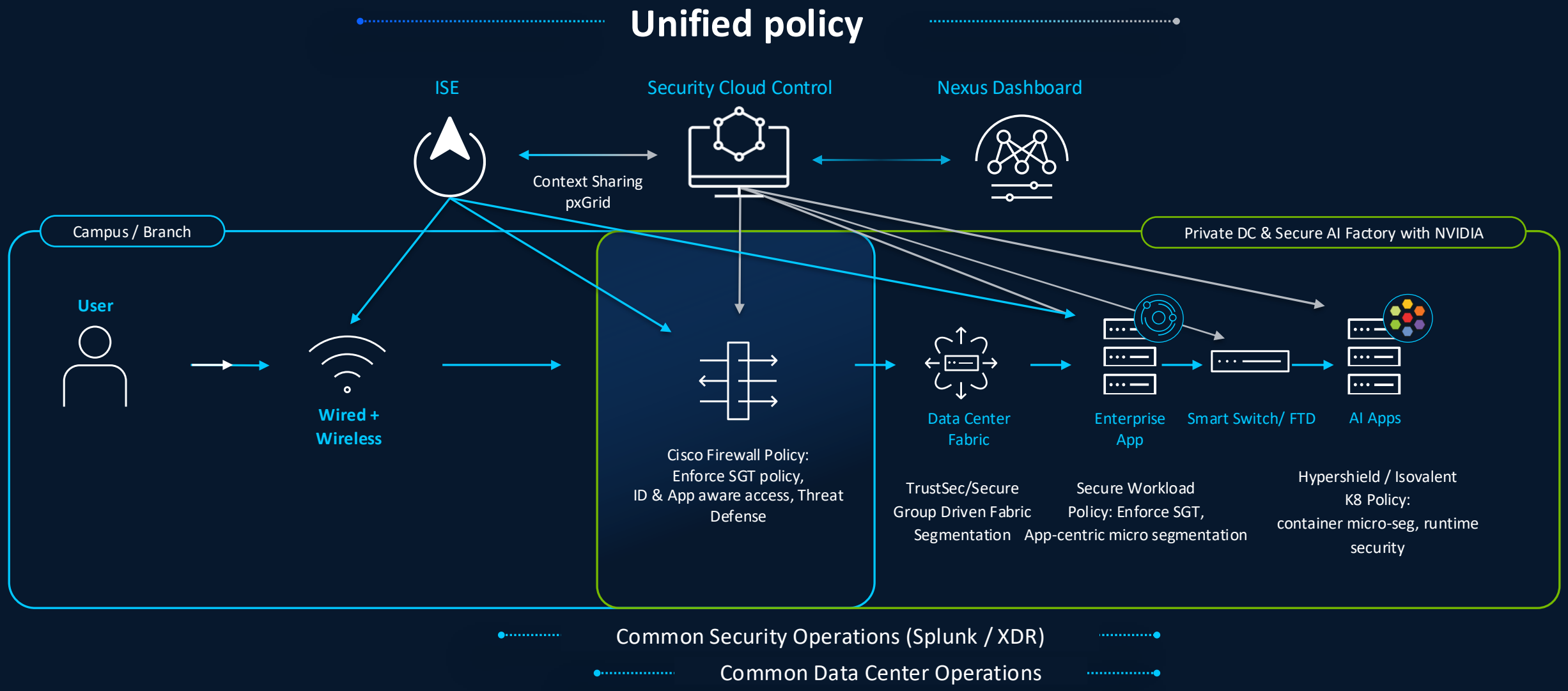


Use Cases

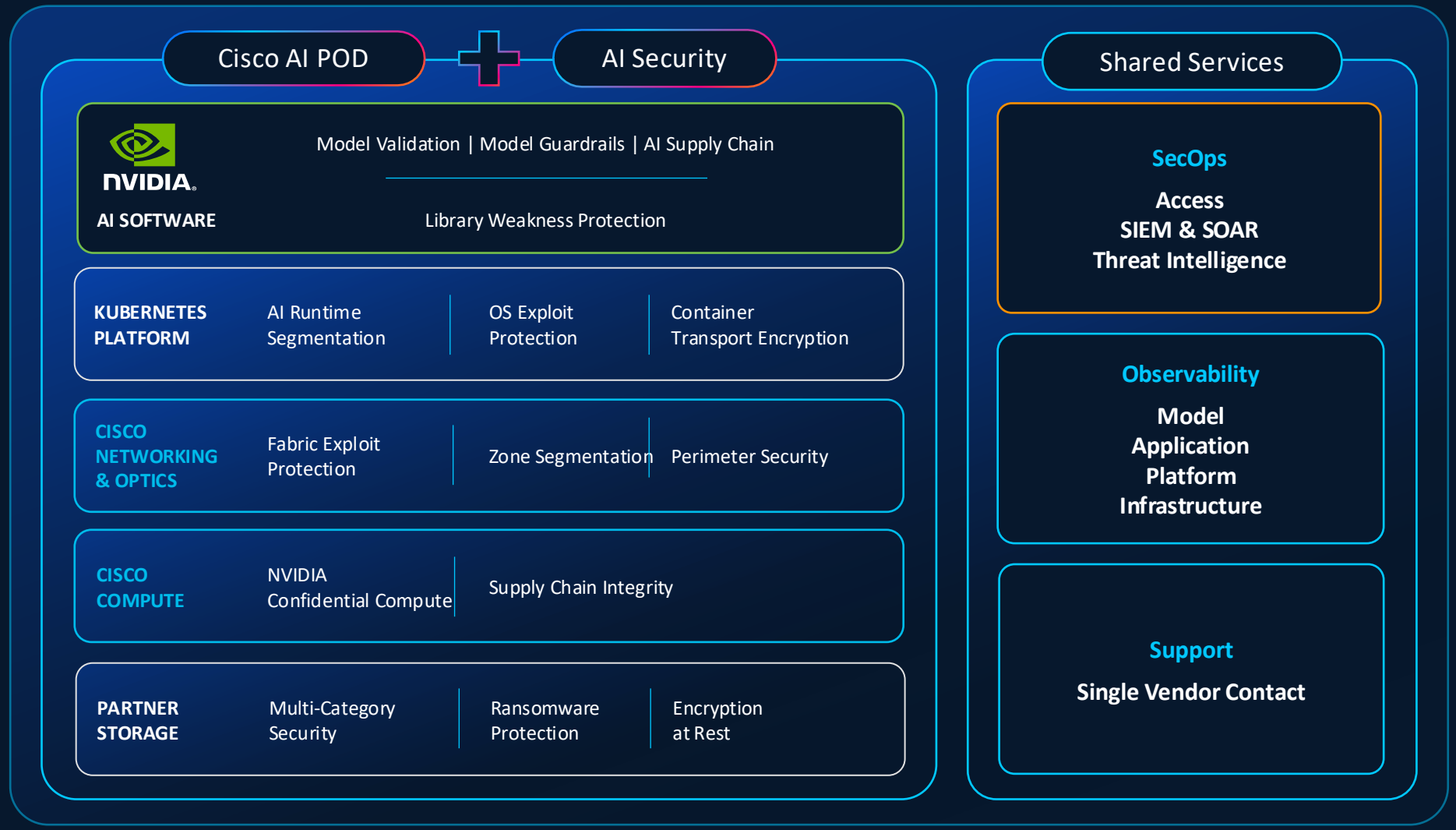
- Top of Rack segmentation and enforcement
- Cloud Edge
- Zone-based segmentation

cisco

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

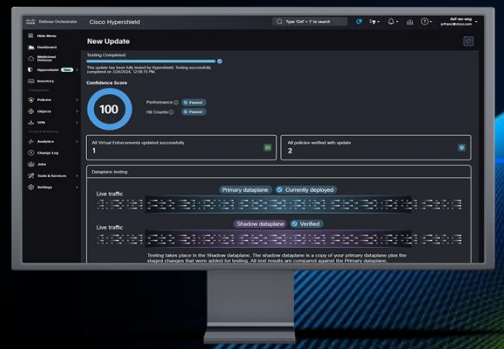


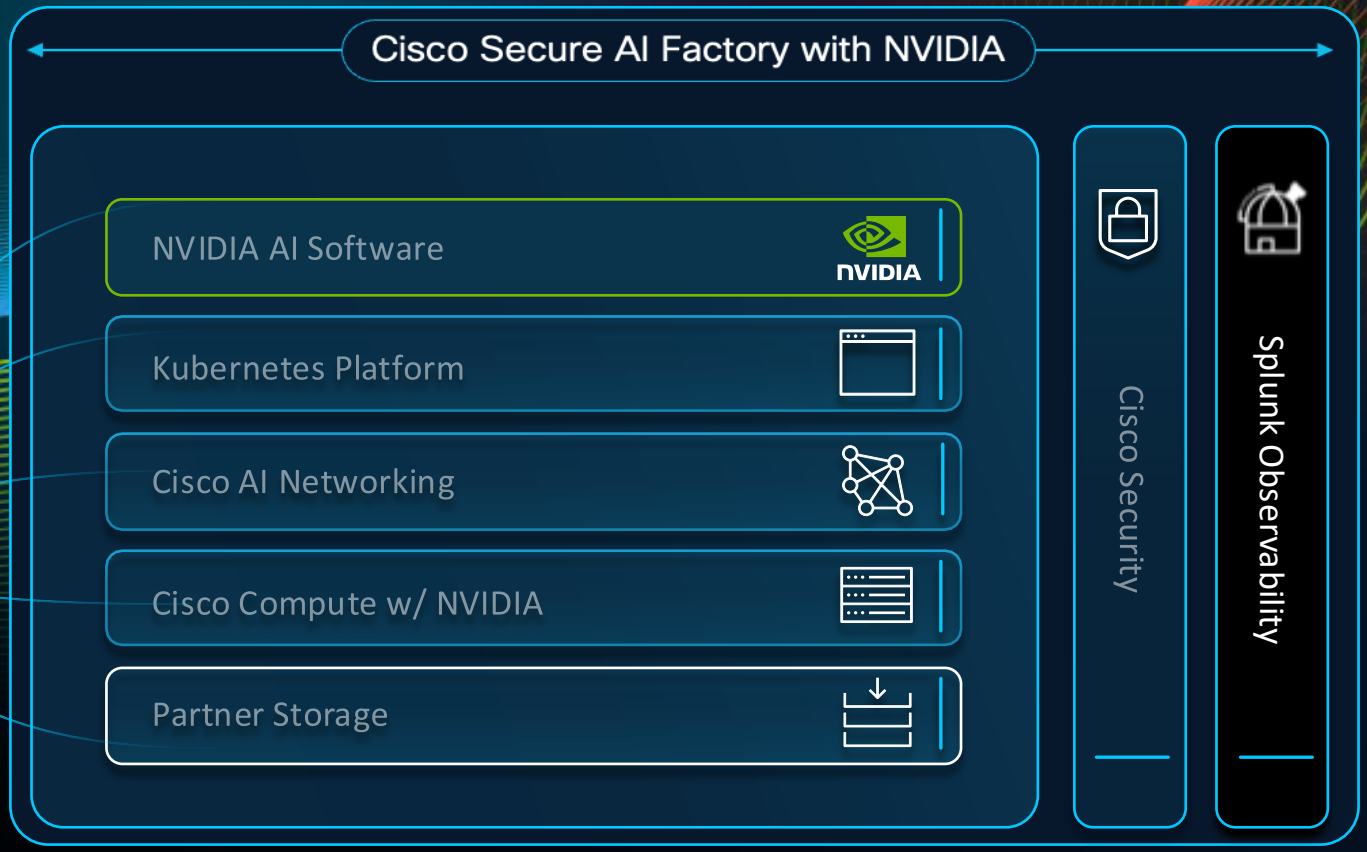
Key Security Capabilities At Every Layer



Secure AI Factory with NVIDIA, Integrated Observability

Every layer needs **visibility!**





Why is Observability for AI important?

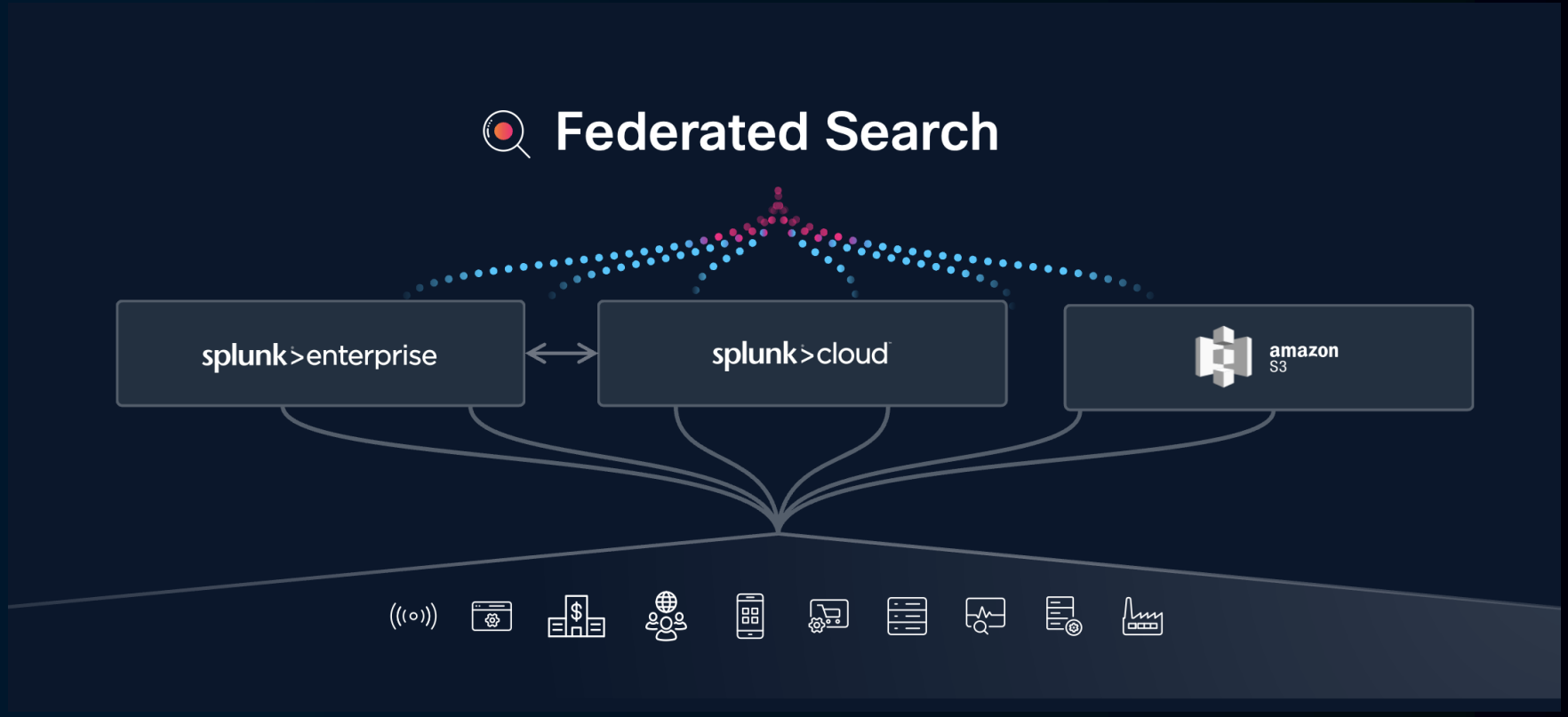
4 key problems teams face

Unreliable or degraded AI performance or agent behavior	Lacking visibility across the complex AI stack	Increased compliance and security risks	Rising costs due to LLMs, agents, an AI Infrastructure
Need to evaluate responses, detect underutilization, optimize inference workloads, decrease hallucinations, bias, inaccuracies, etc. to reduce latency, errors, and other inefficiencies	Need to correlate business problems to new forms of telemetry across AI infrastructure layers, including vector databases, orchestration frameworks, GPUs, LLMs, etc.	Need to detect prompt failures and injection, toxicity, personally identifiable information (PII) leakage, etc. to reduce liability, scrutiny, and reputational damage.	Need to track cloud and GPU spend generated by affecting AI infrastructure, LLMs, and AI agents to optimize resources and ensure clear ROI

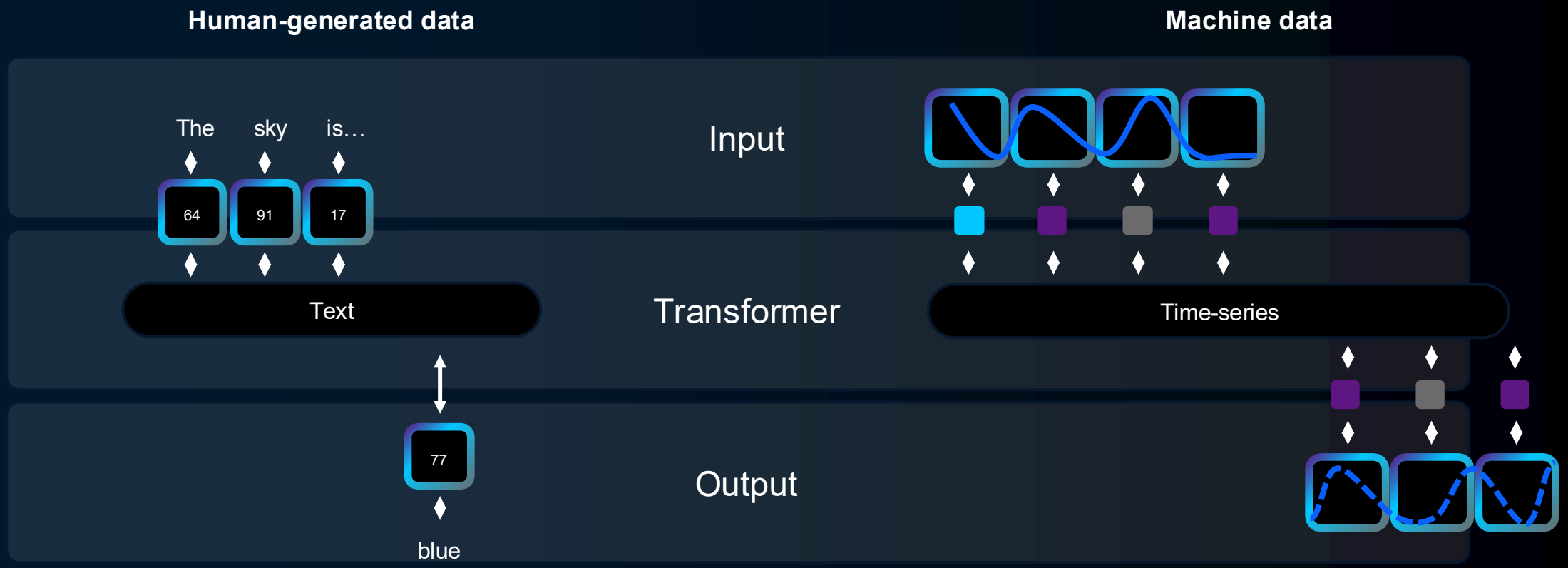
These problems can lead to irresponsible AI, poor customer experiences, and lost trust

cisco | NVIDIA

Don’t move your data – Data Federation

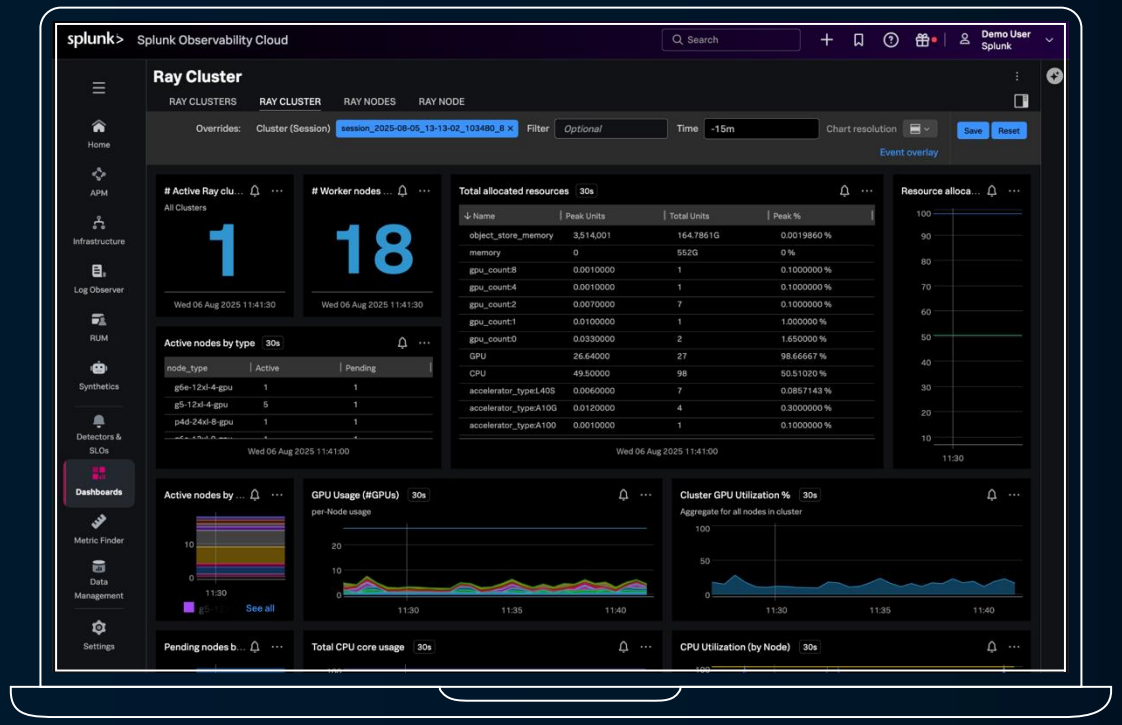


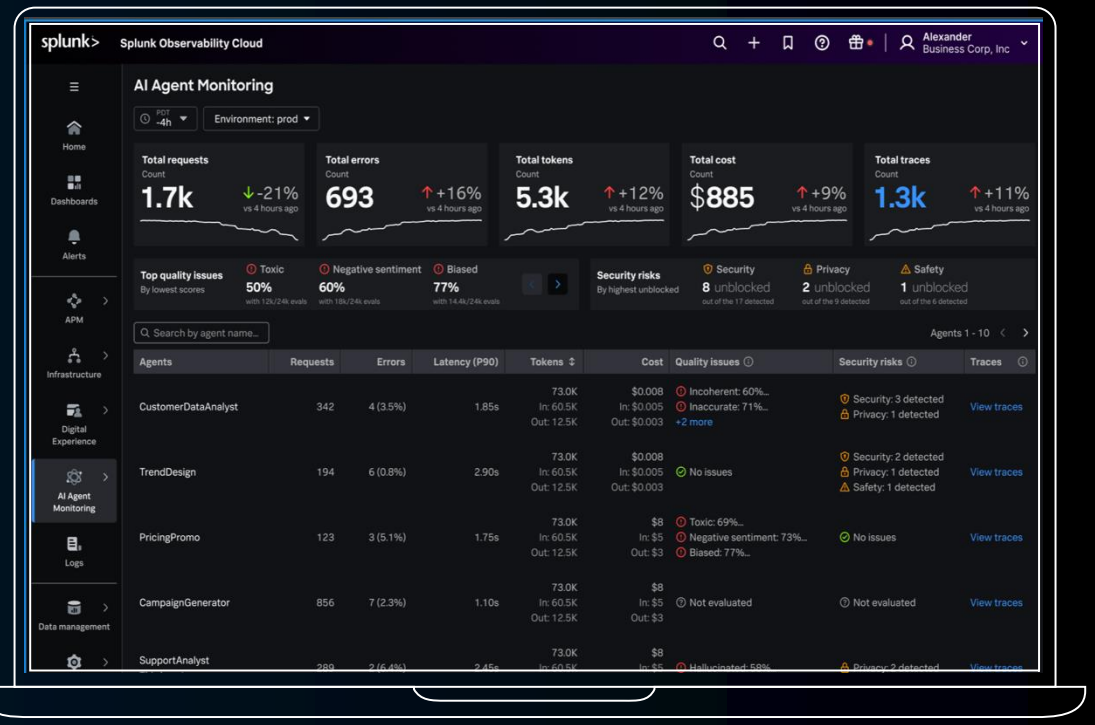
Cisco Differentiated AI



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 AI POD Dashboard Views Using AI Infrastructure Monitoring

Automatic attribution to instrumented services, customized data slicing, and platform-dependent attribution models

Additional Dashboard Views

Intersight

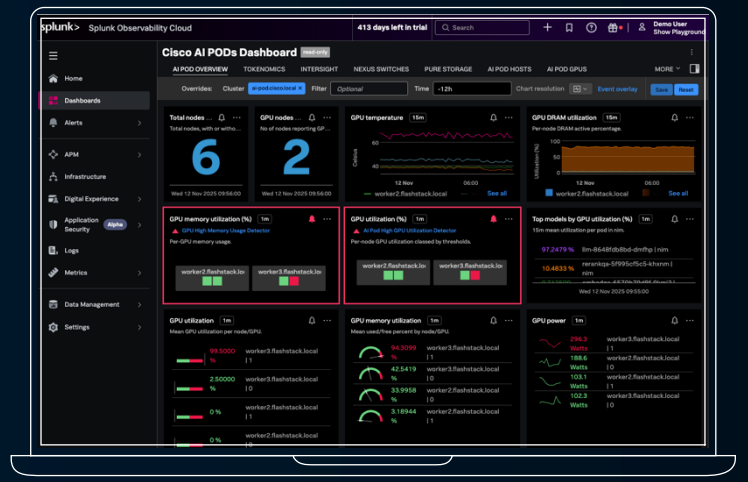
Nexus Switches

Storage

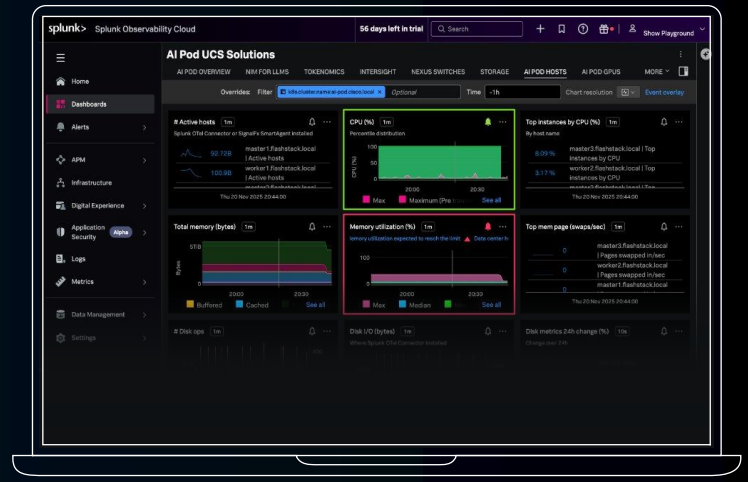
NIM for LLMs

Clusters

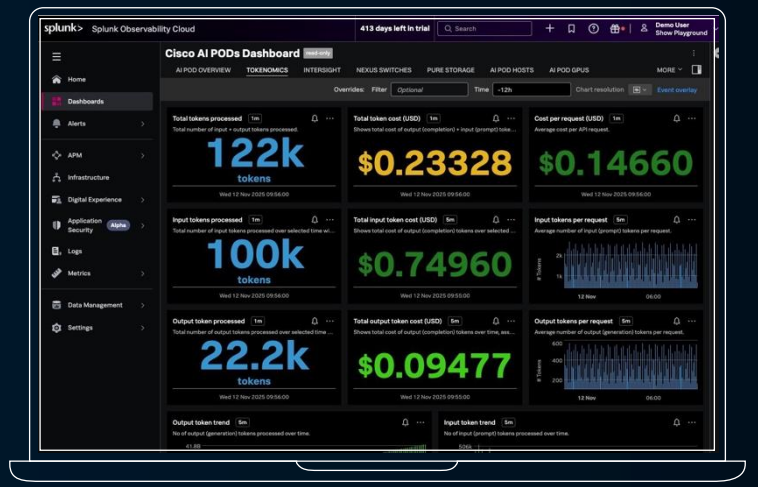
LLM Model Costs



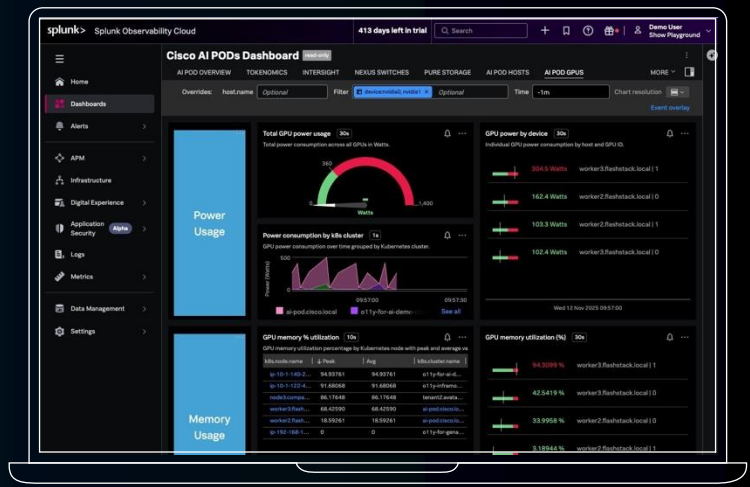
AI POD Overview - Total nodes, GPU power and utilization (%), etc...



AI POD Hosts - # of active hosts, CPU (%), memory utilization (%), etc...



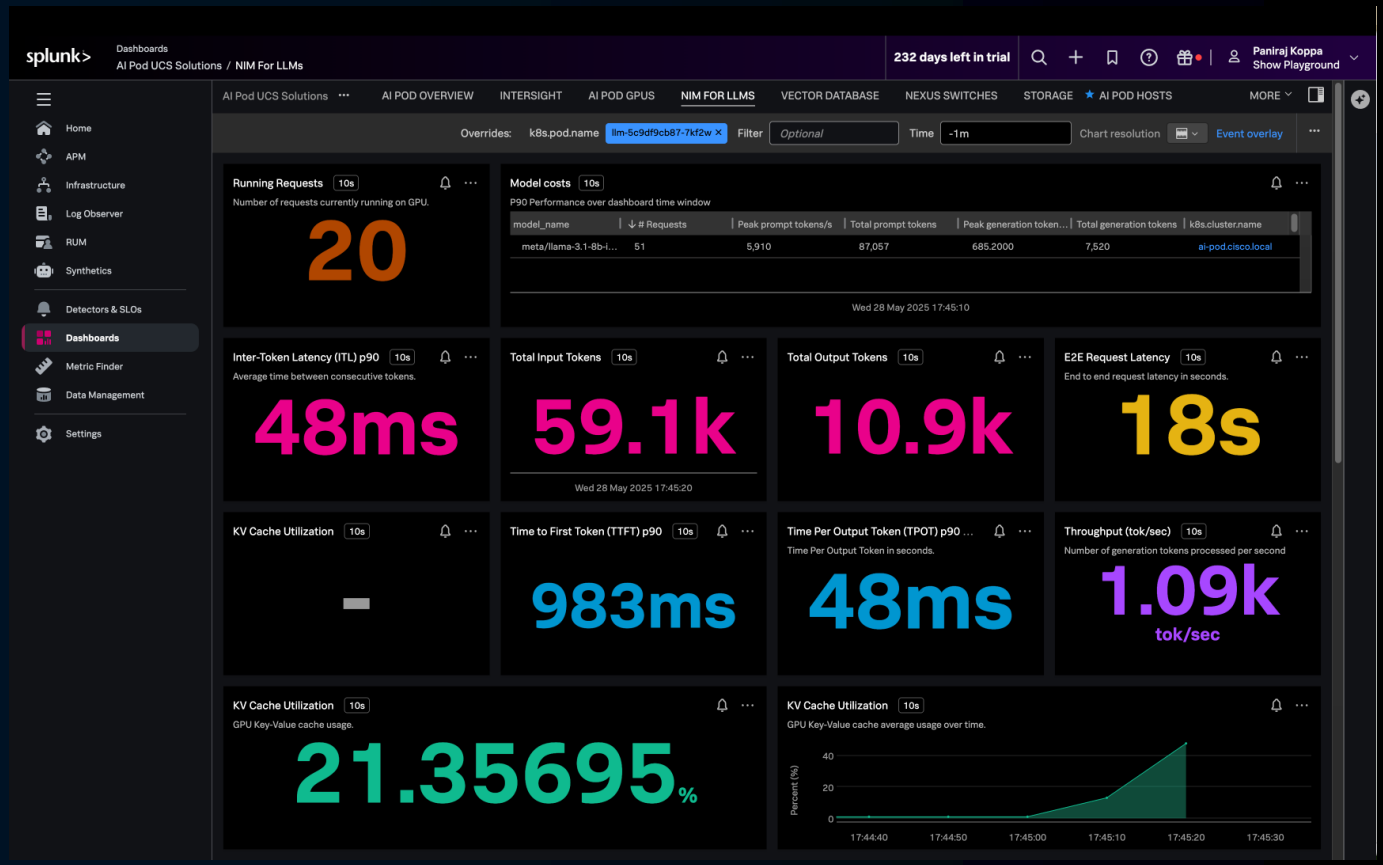
Tokenomics - Total tokens processed, Total token cost, etc...



AI POD GPUs- GPU power usage, GPU memory utilization (%), etc...

Splunk Observability Dashboard for Cisco AI PODs

- Real-Time Monitoring
- Troubleshooting
- Generate actionable Insights
- Efficient Telemetry Data Ingestion and Processing



Splunk with AI Defense

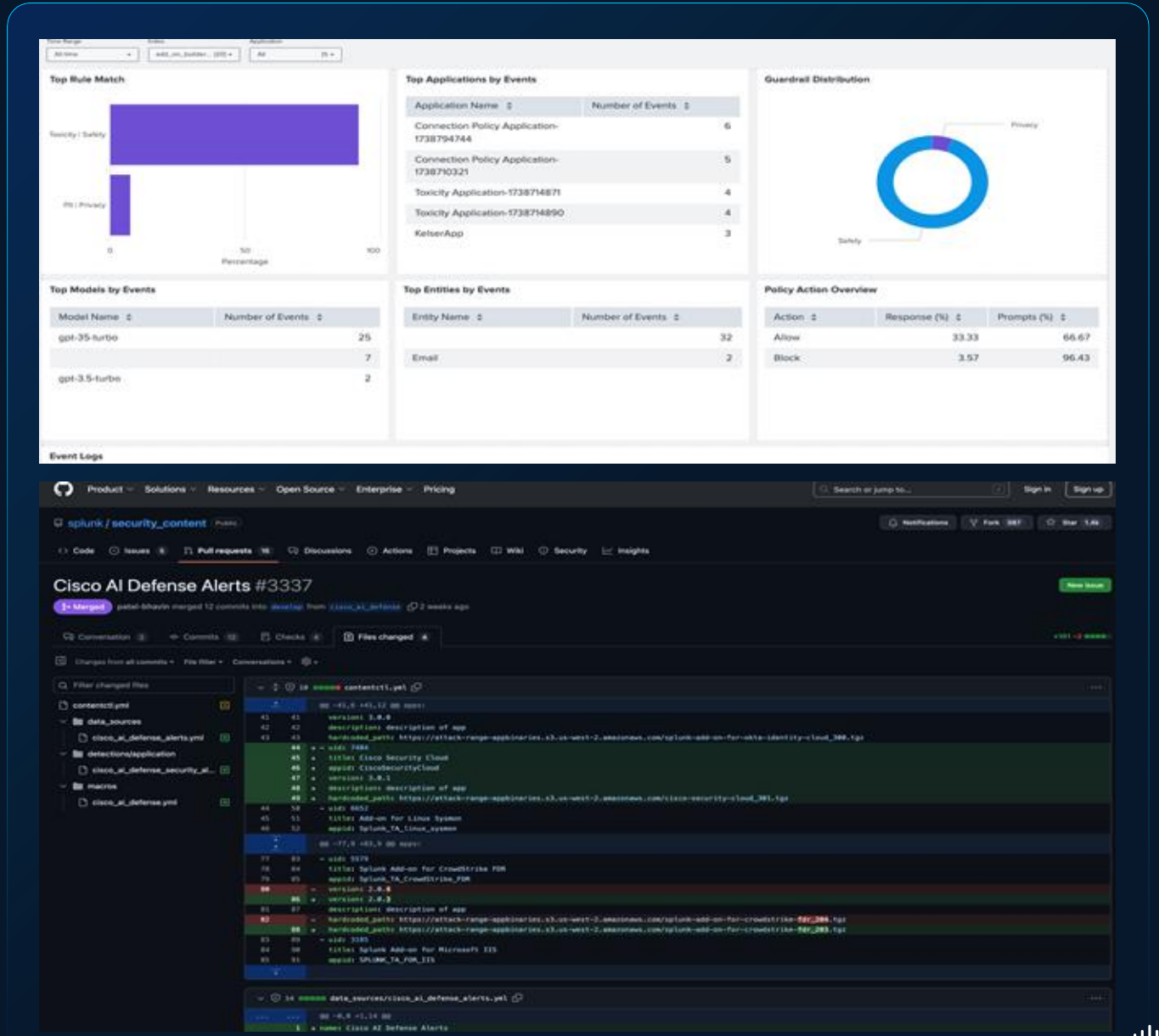
Technical Addon

Gain visibility into emerging AI risks with Splunk

Pulls in alerts from AI Defense and maps them to the Common Information Model (CIM), visualized in a dashboard.

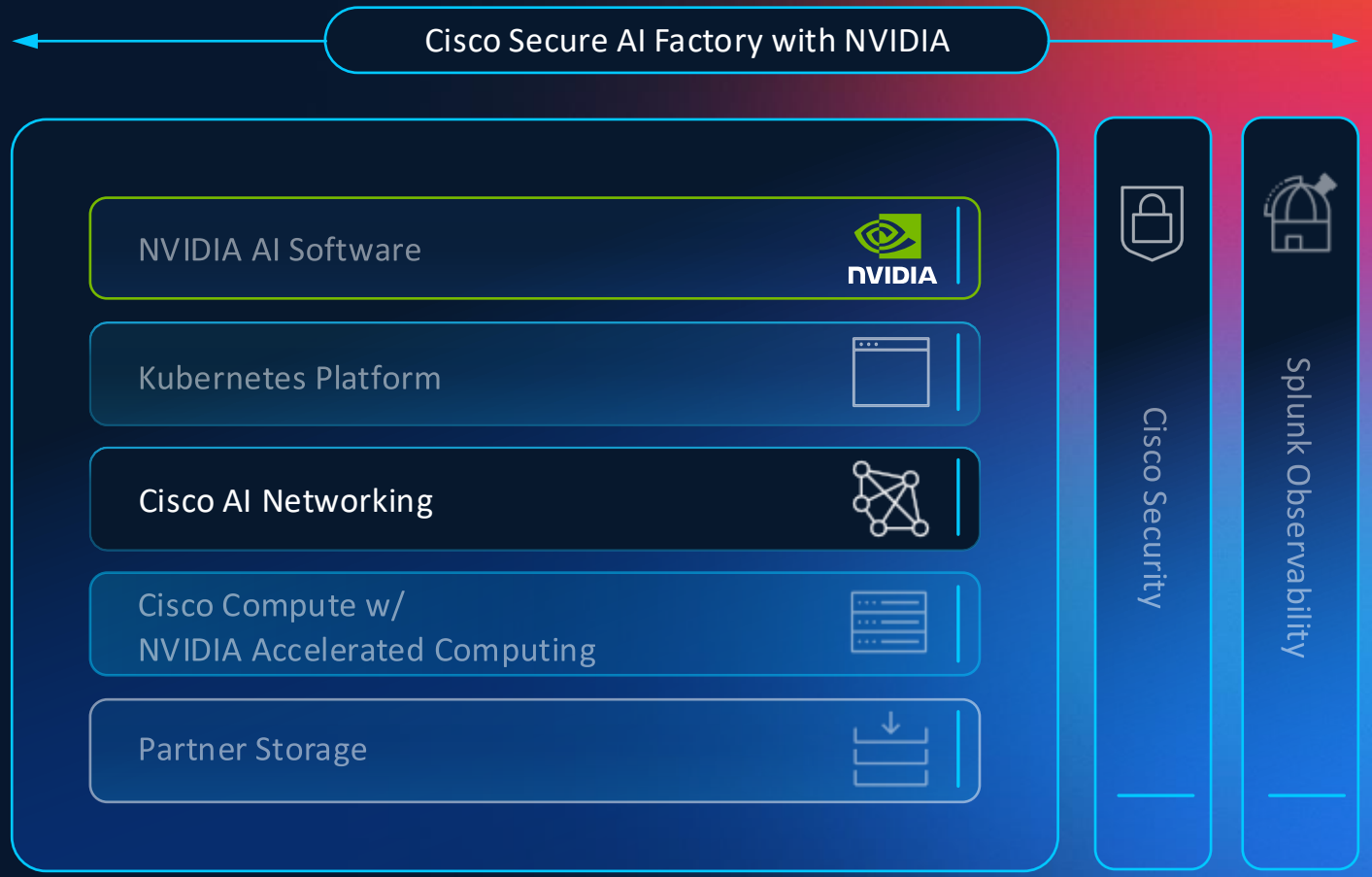
Gain visibility into risks associated with LLM models, AI apps and entities.

Includes an out-of-the-box Enterprise Security detection that creates a search and surfaces potential attacks against the AI models running in your environment.



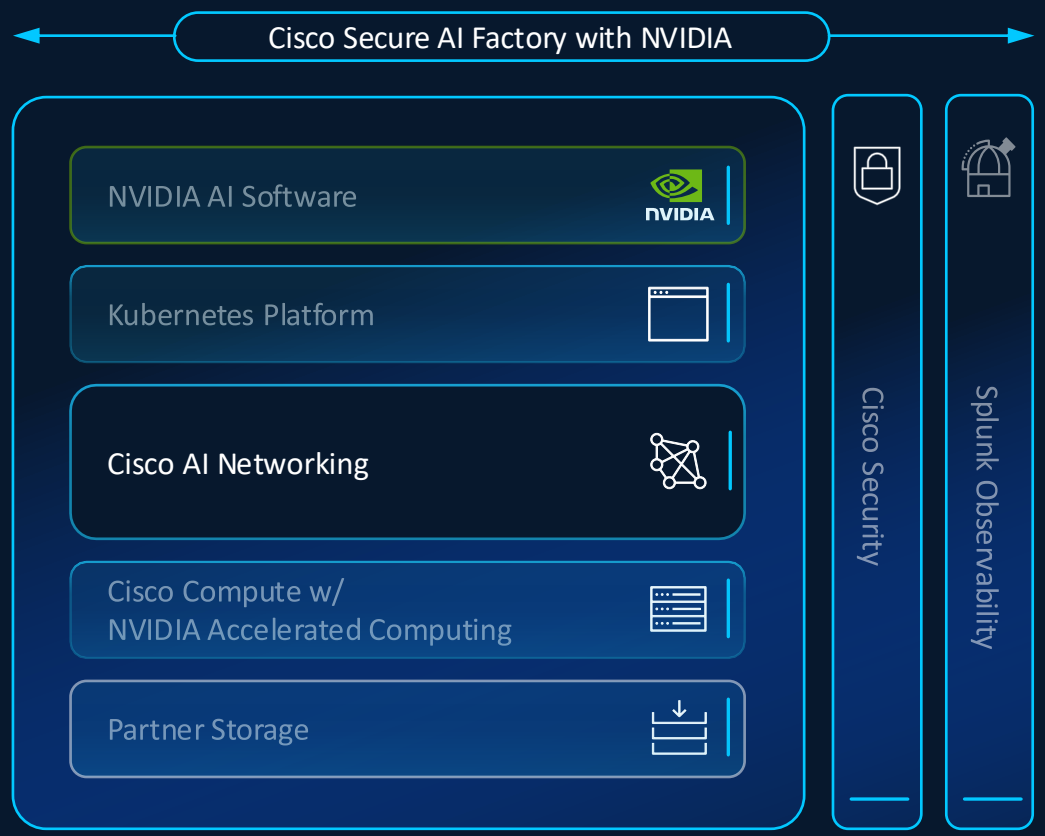
Only Cisco can address observability at every layer of Secure AI Factory

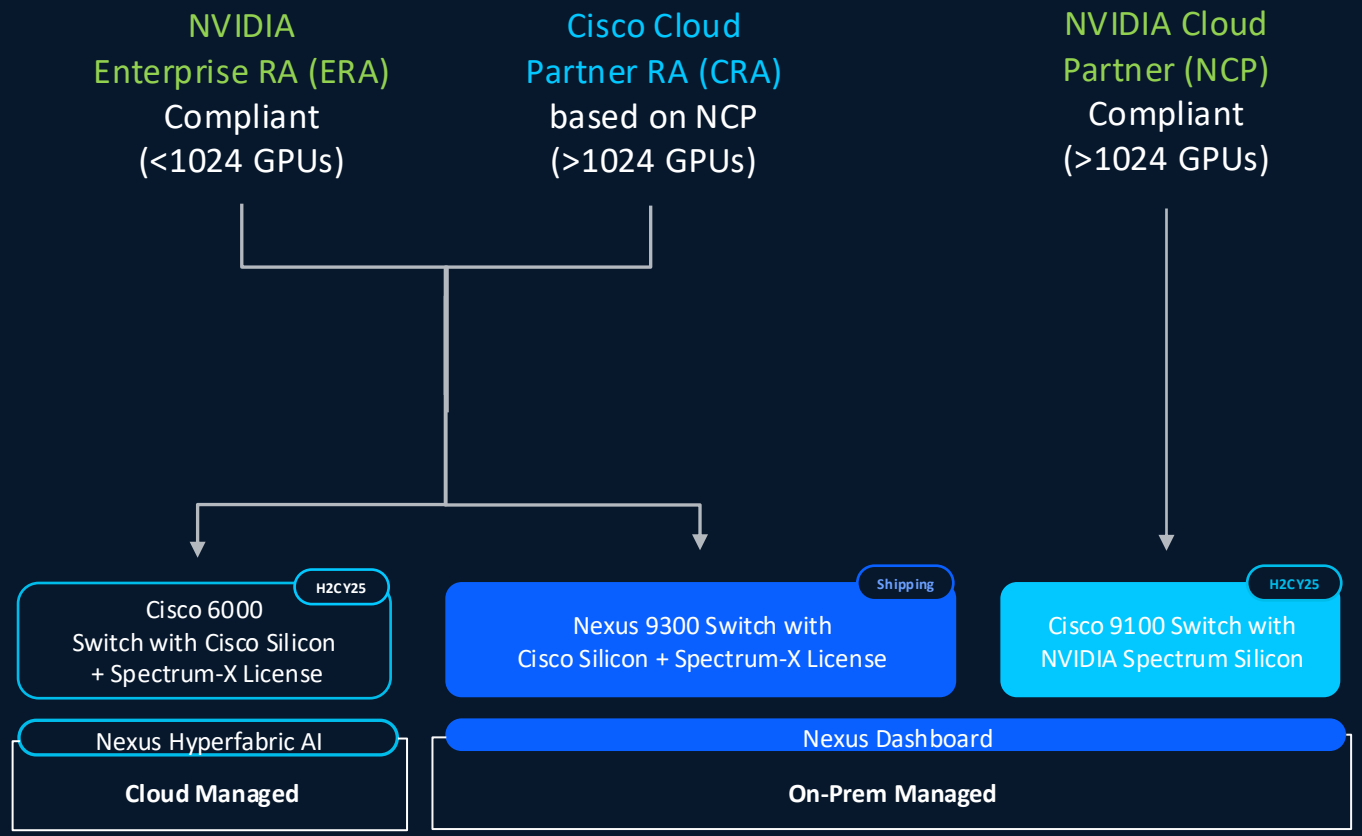
Secure AI Factory with NVIDIA, Networking



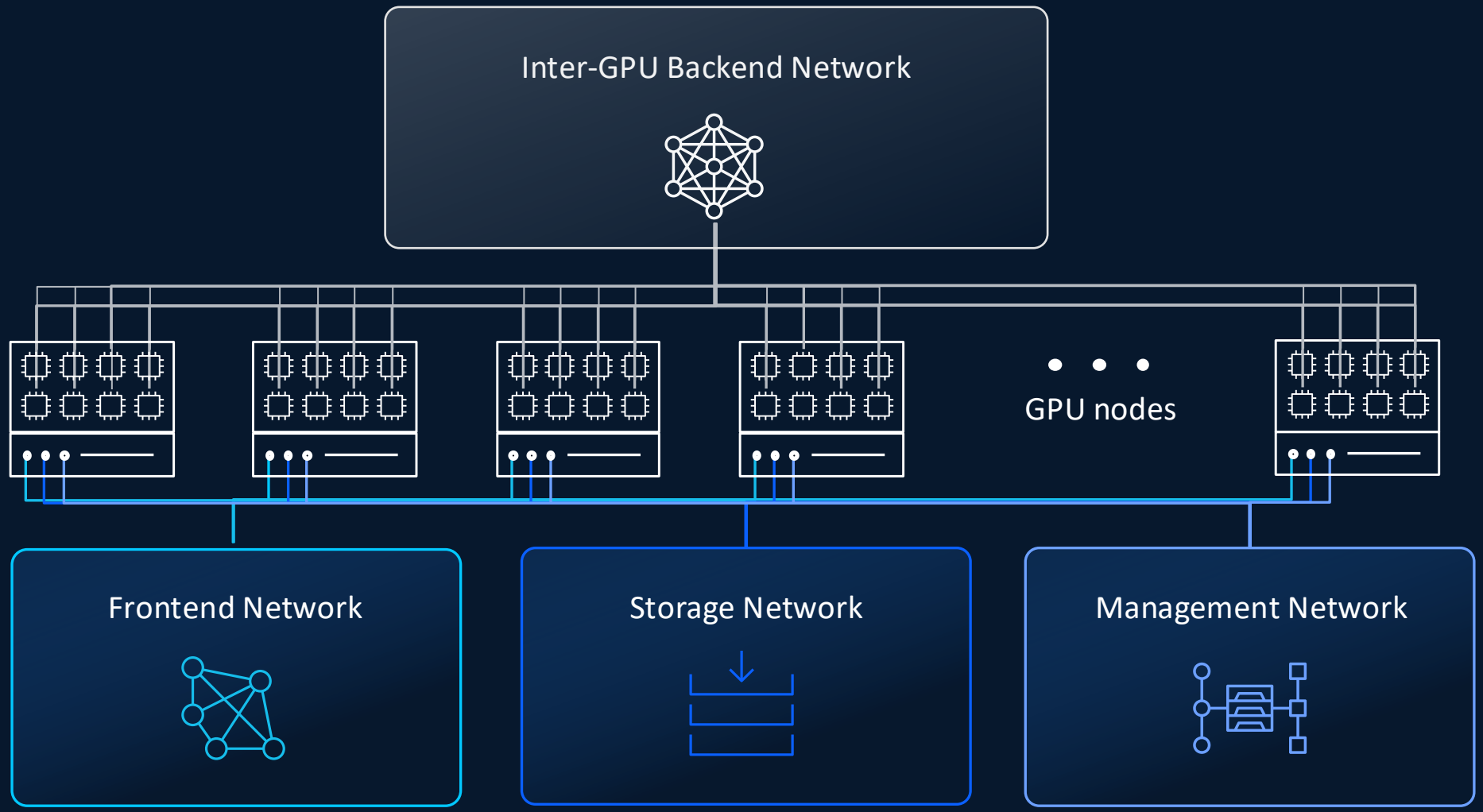
Secure AI Factory with NVIDIA

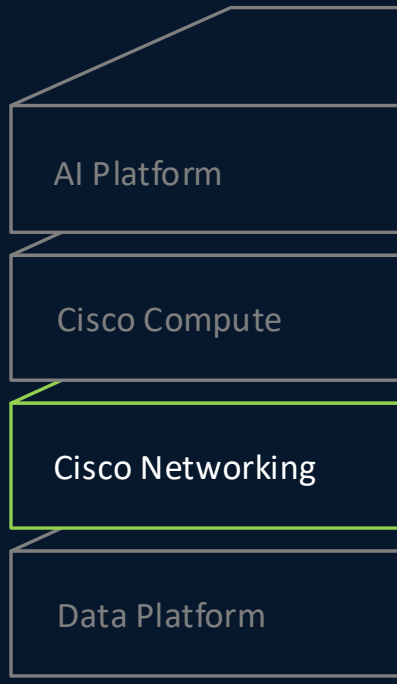
Available with Cisco & NVIDIA Enterprise and Cloud Partner Reference Architecture (RA)





AI Networking





Networking in AI Environments

Back-End Network

Extremely low-latency and high-data rate direct connectivity between GPUs.

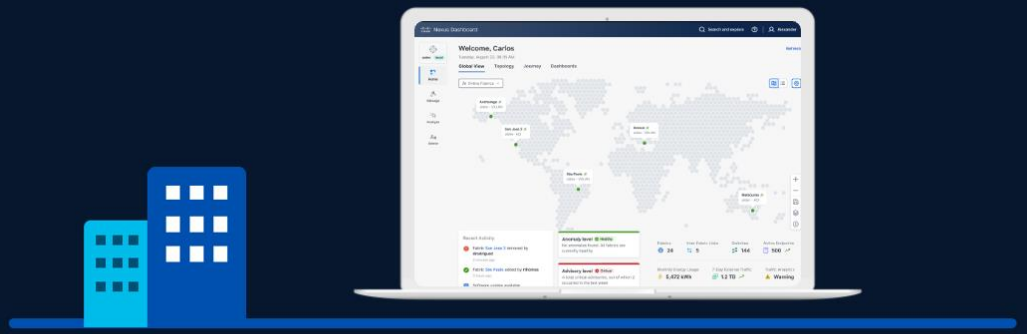
Highly specialized hardware, RDMA protocols, limited to “local” GPUs.

Front-End Network

Transfers data sets, models, and inference requests to/from the node via the CPU.

“Standard” networking hardware and protocols (TCP/IP, HTTP).

Data Center Networking Portfolio



Nexus Dashboard

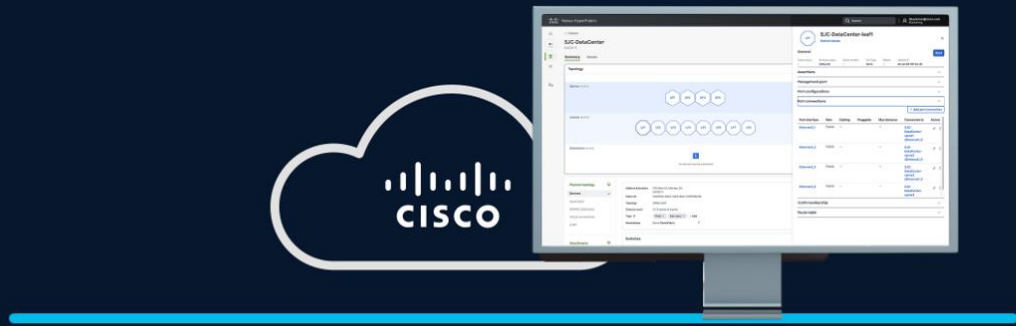
On-Premises Delivered

NX-OS

ACI

Powered by Nexus 9000 Series

Day 2 Ops | Visibility | Analytics | Troubleshooting | Compliance



Nexus Hyperfabric

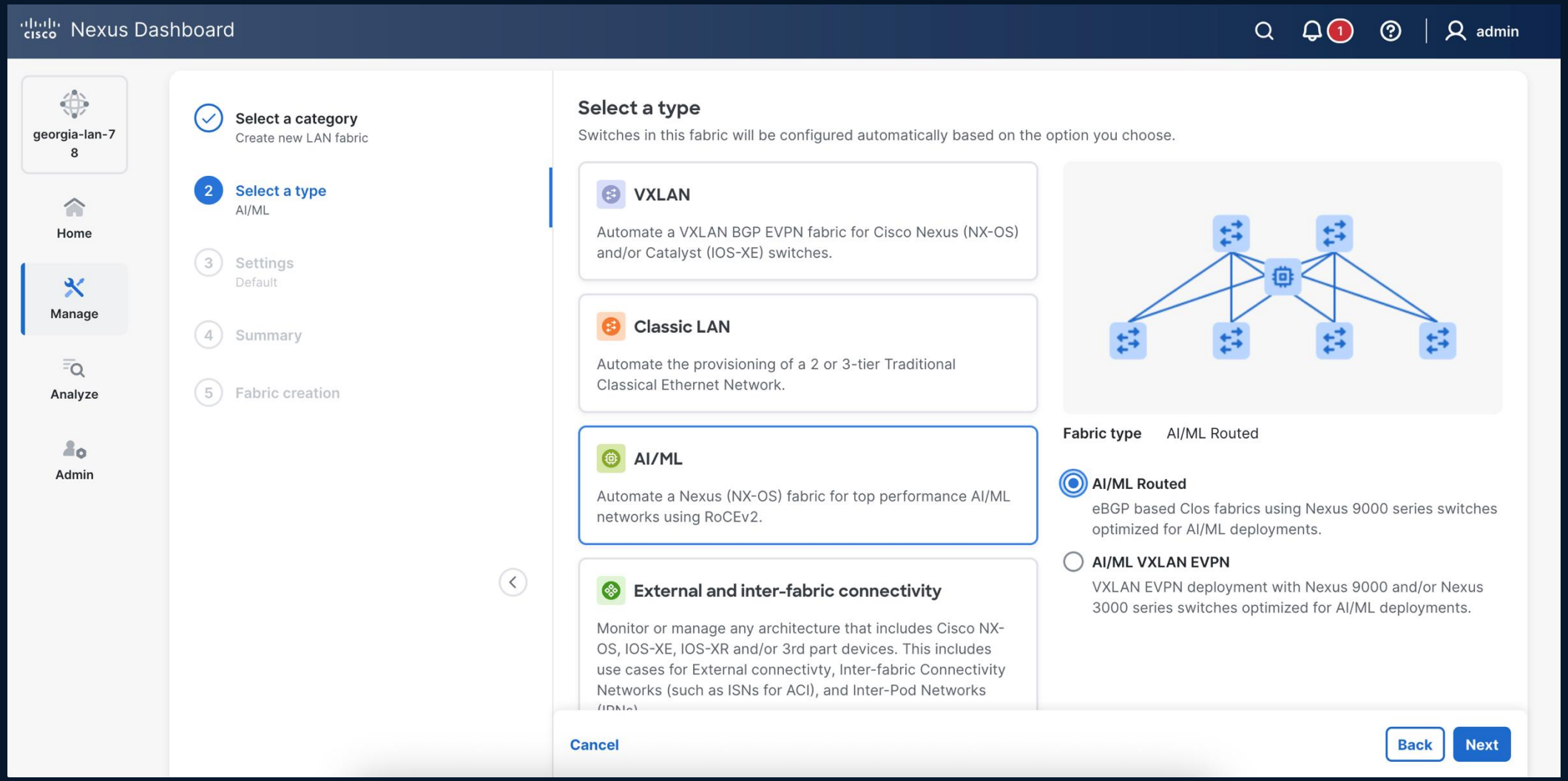
Cloud Delivered

Powered by Cisco 6000 Series

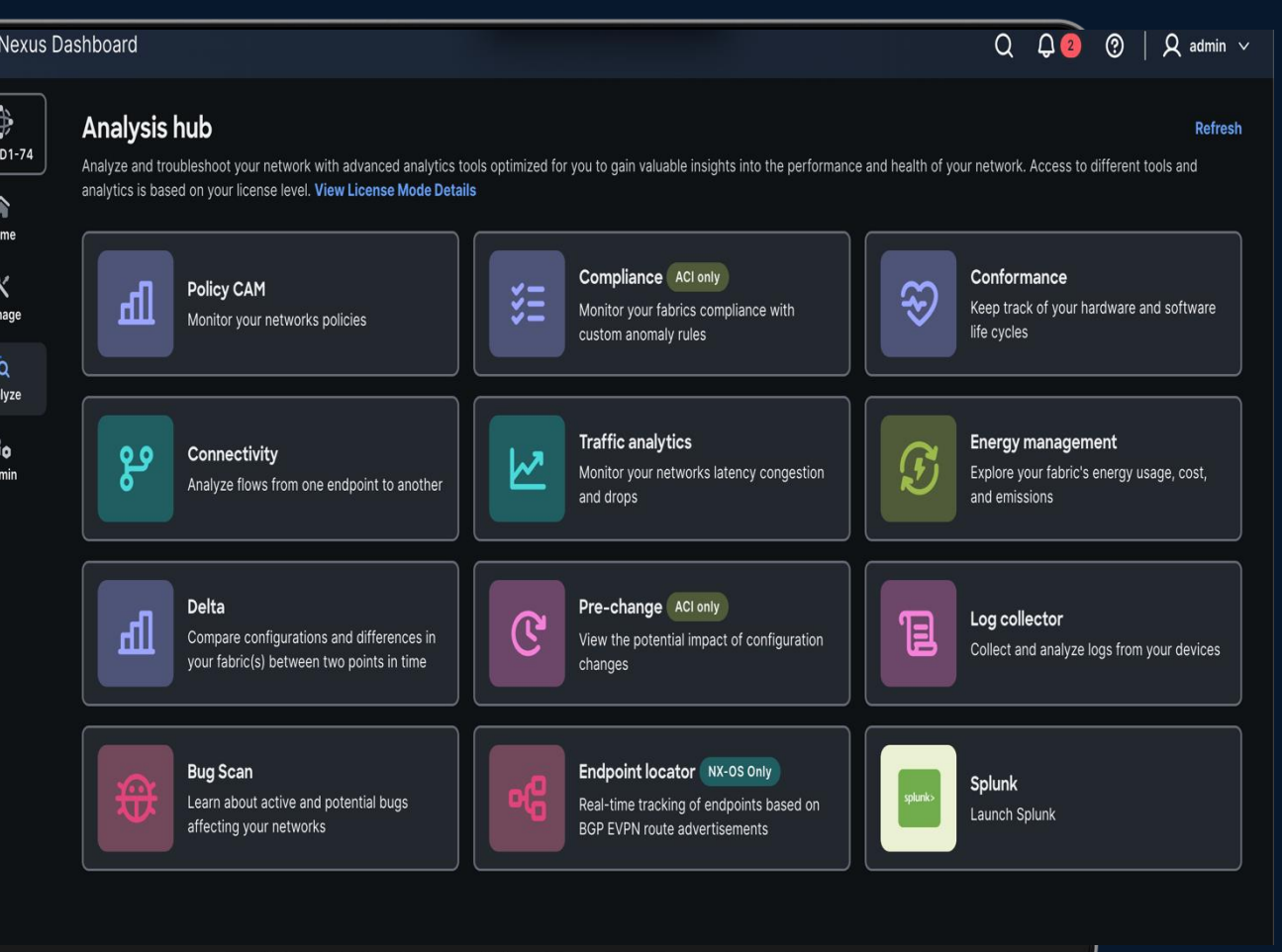
Delivering networking at speed from the cloud

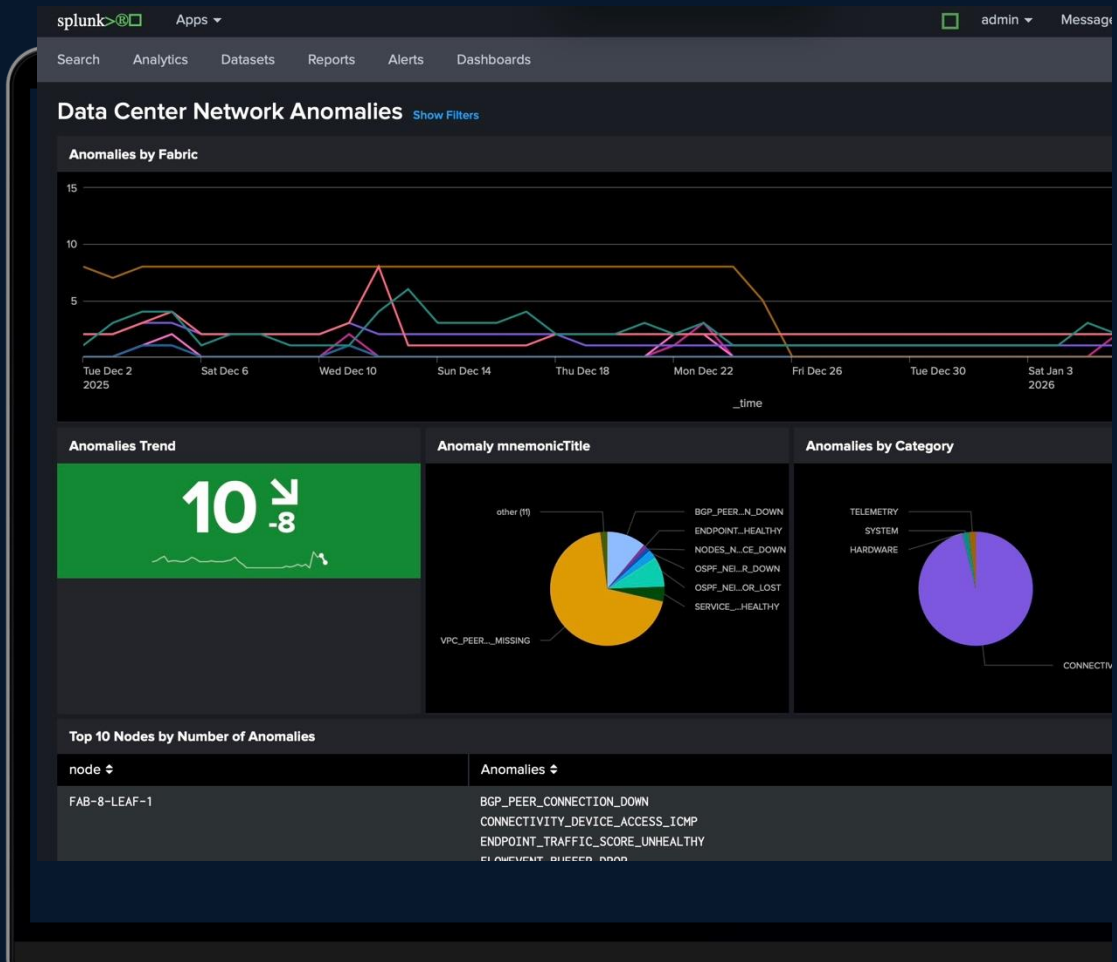
cisco

Nexus Dashboard – AI/ML Fabric Deployment



Native Splunk Analytics in Nexus Dashboard





DISTRIBUTED ANALYTICS | FEDERATED SEARCH | Flexible Data Ingestion

Faster Troubleshooting | Data Sovereignty | Reduced Cost

Cisco Nexus Hyperfabric AI

High-performance Ethernet

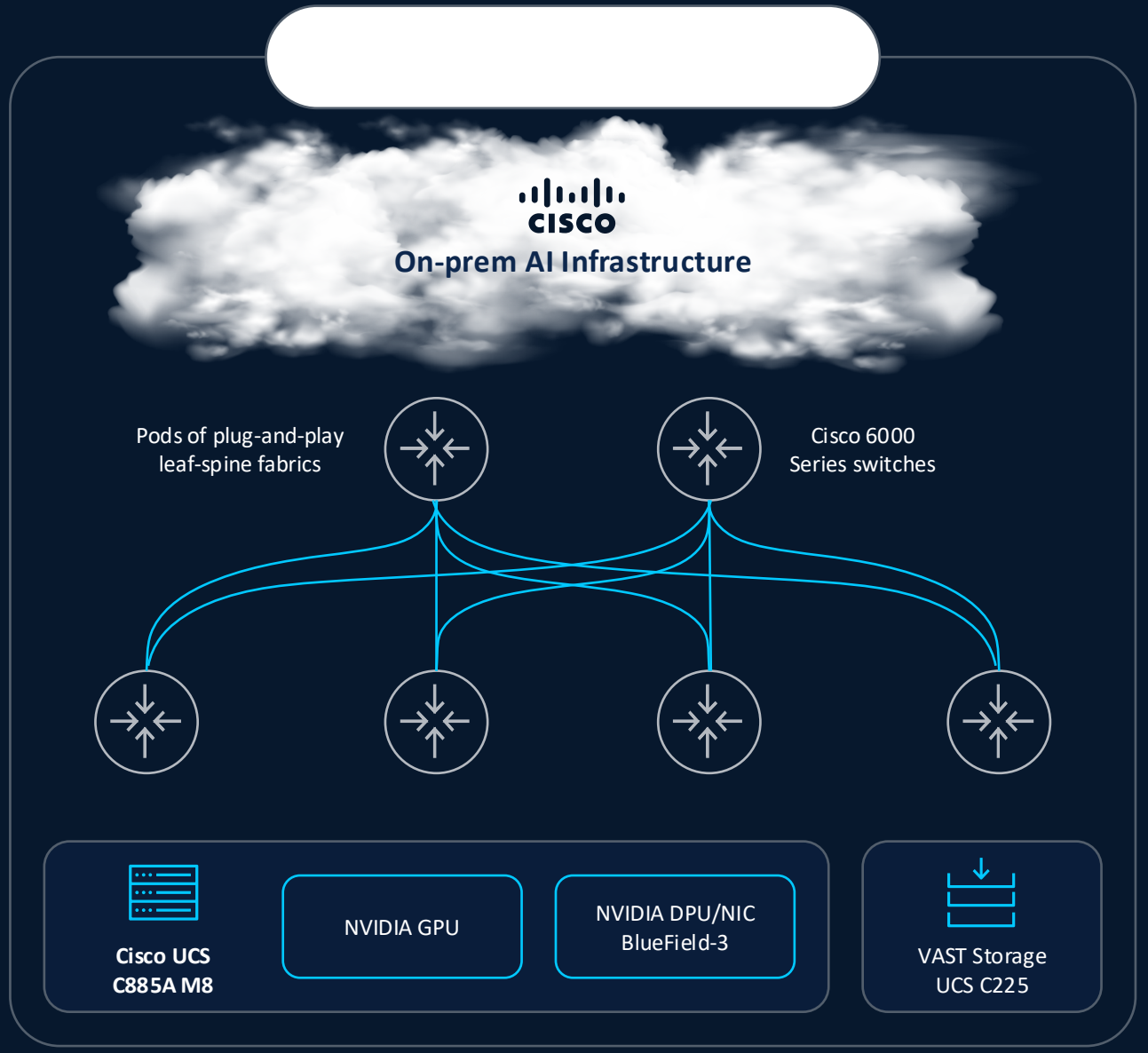
Cloud-managed operations

Unified stack including NVAIE

AI-native operational model

Democratize AI infrastructure

Visibility into full stack AI



cisco

← Fabrics

Search

3 results

Nvidia Enterprise Reference Architecture with Spectrum X

SMALL **Production Inferencing and Small Model Training**

2 Backend leaves
14 800G ports available

1 Management leaves
2 400G ports available
48 SFP ports available

2 Storage leaves
56 400G ports available

4 GPU servers
32 GPUs

11 Storage servers

Offering 32 GPUs in 4 HGX-style UCS servers in a rail-aligned, Nvidia Spectrum X fabric Nvidia-compliant reference architecture, this cluster is well-suited for production AI model serving, training small-sized models, and fine-tuning medium-sized models.

Switches


HF-S4-610


HF6100-60L4D


HF6100-32D

Servers


UCSC-885A-M8-HC1


UCS-M8-MLB

Select small AI cluster

Nvidia Enterprise Reference Architecture

MEDIUM **Medium Model Training and Large Model Fine-Tuning**

2 Backend leaves
14 800G ports available

1 Management leaves
2 400G ports available
48 SFP ports available

2 Storage leaves
56 400G ports available

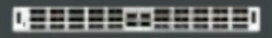
12 GPU servers
96 GPUs

11 Storage servers

The Medium AI Cluster offers 96 GPUs across 12 HGX-style UCS servers in a rail-aligned network architecture. It is ideally suited for AI/ML tasks such as training medium sized models, fine-tuning larger models, and running parallel inference pipelines.

Switches


HF6100-64ED


HF6100-60L4D


HF6100-32D

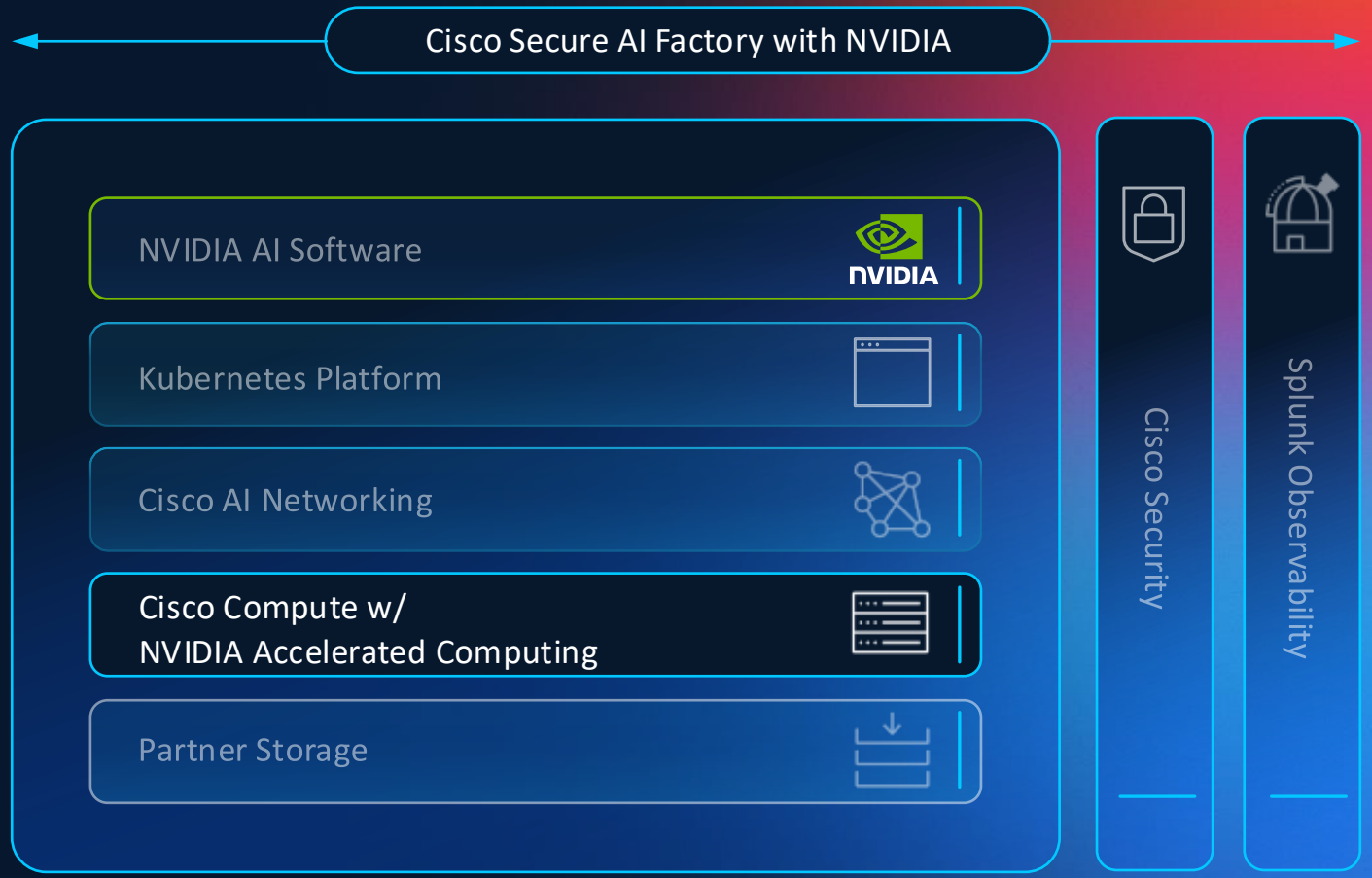
Servers

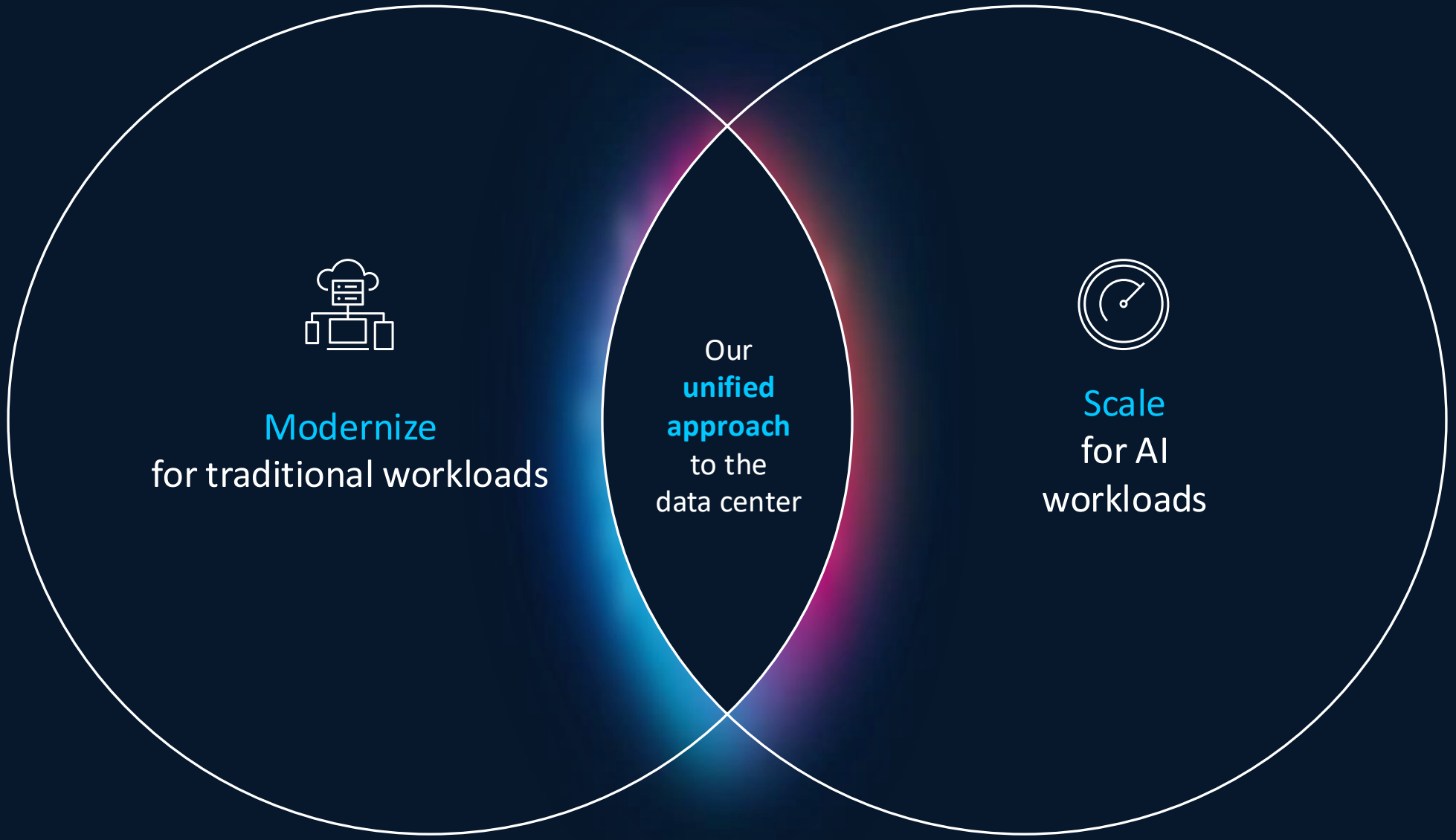

UCSC-885A-M8-HC1


UCS-M8-MLB

Select medium AI cluster

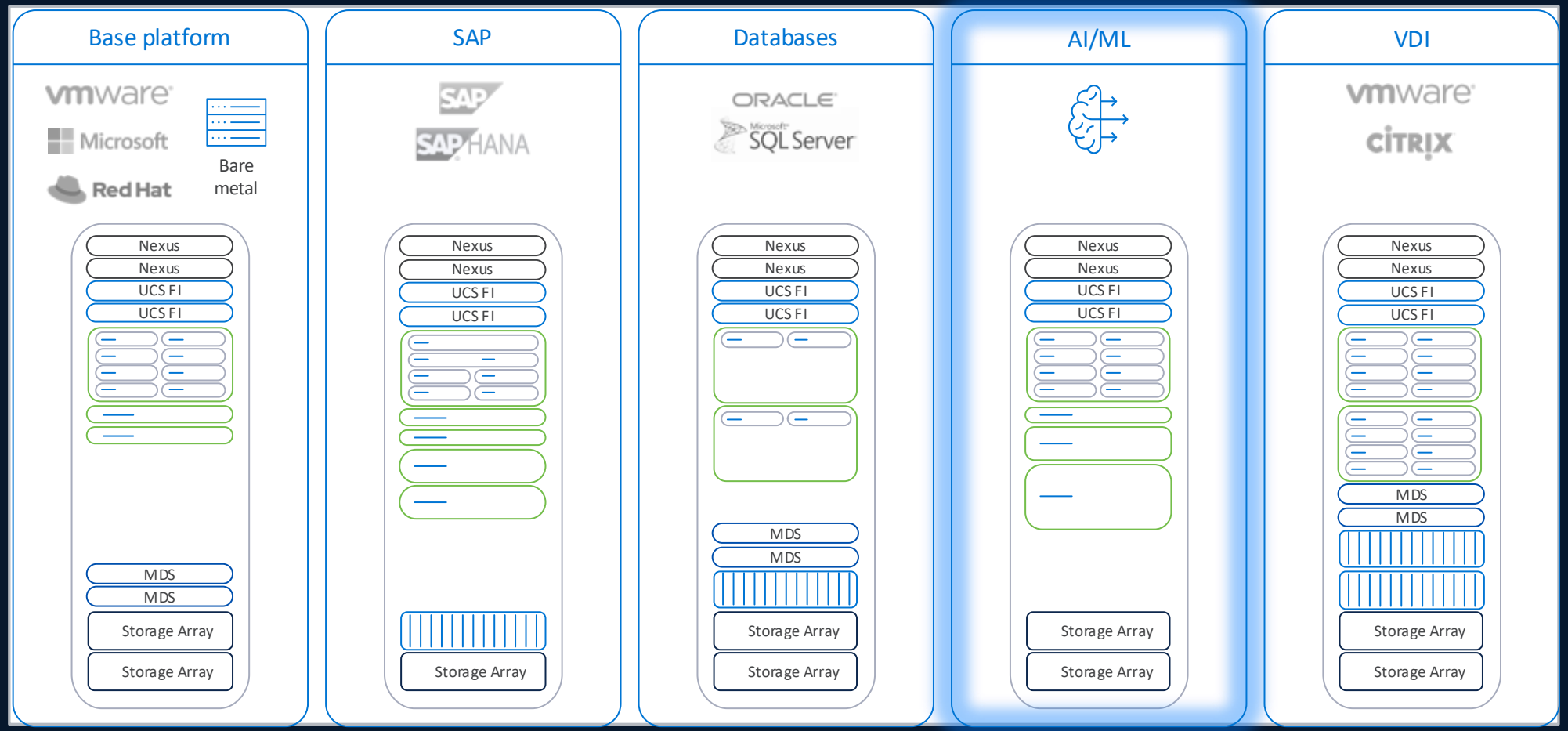
Secure AI Factory with NVIDIA, Compute




Modernize
for traditional workloads
Our
unified
approach
to the
data center
Scale
for AI
workloads

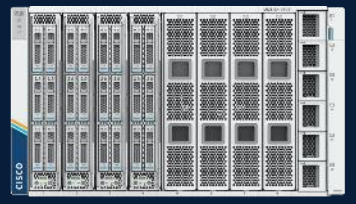
AI is just another workload to Cisco UCS!

Right sized, tested, and validated for mission critical applications



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

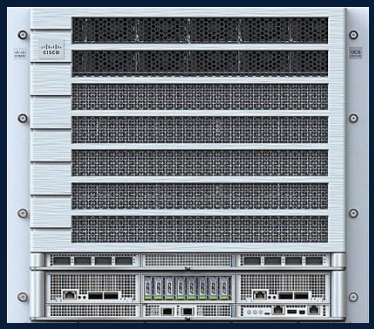
UCS C245 M8SX
28 HDD/SDD

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

Compute AI portfolio

Address AI workloads with visibility, consistency, and control

Validated solutions for AI with compute, network, storage, and software

Build the model
Training

Optimize the model
Fine-tuning and RAG

Use the model
Inferencing

RTX PRO SERVER
Supporting RTX PRO 6000 Blackwell Server Edition GPUs

Cisco UCS®
GPU-dense servers
PCIe and NVLink Servers

Cisco UCS blade (with GPU extensions) and rack servers

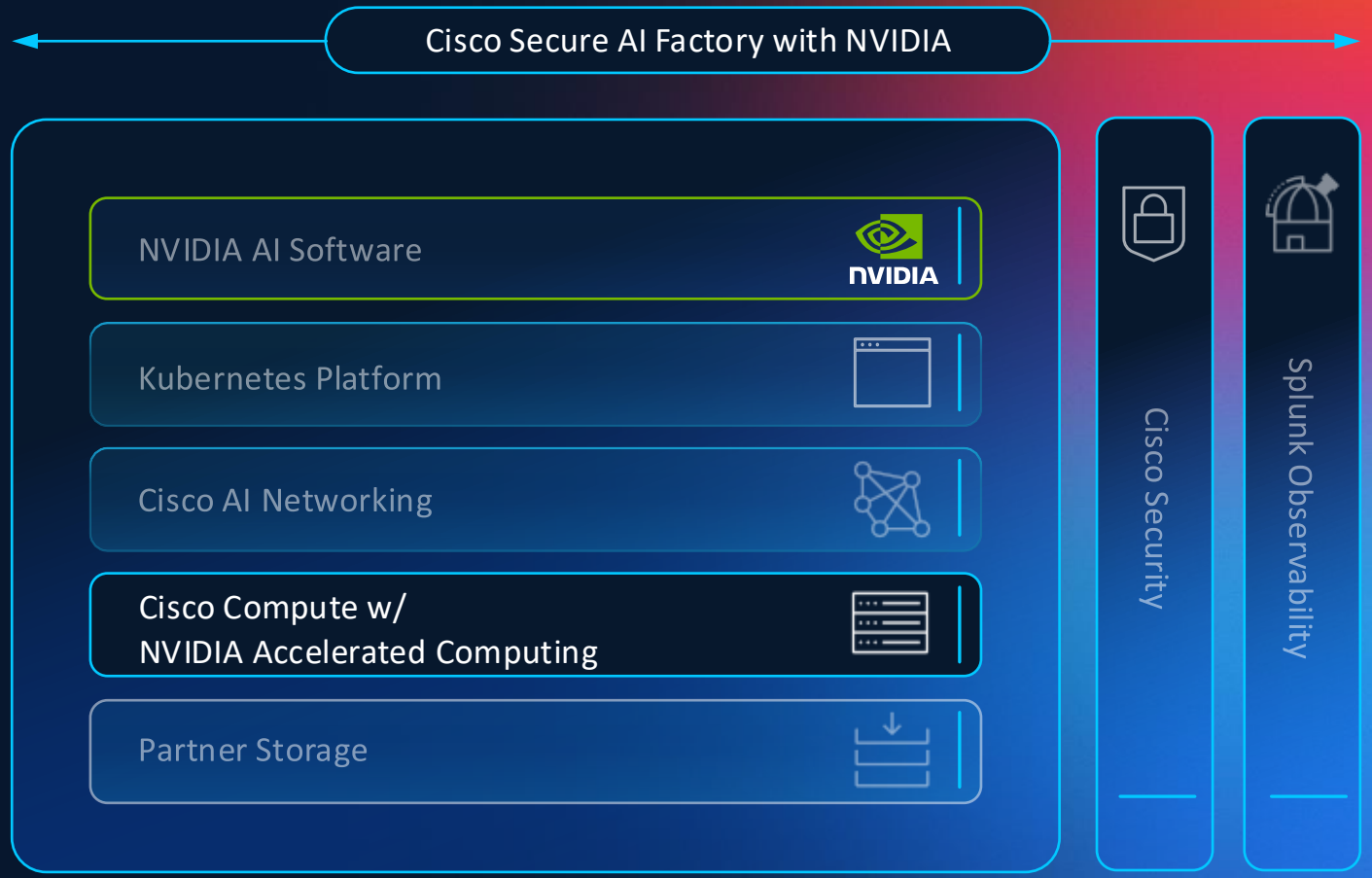
Enterprise AI edge

Dense compute for demanding AI

Full-stack AI with compute and networking

cisco

Secure AI Factory with NVIDIA, Compute



Reference Architectures

Dense GPU Platforms

HGX

- Maximum Performance for model training and other demanding use cases
- Relatively uniform configurations across all major OEMs
- Three key features that customers want
  1. 8-way fabric-attached GPUs
     NVlink & NVswitch for NVIDIA
     Infinity Fabric for AMD
  2. Mezzanine GPUs
     SXM for NVIDIA
     OAM for AMD
  3. 1:1 pairing of GPU and Backend NICs

MGX

- Scalable Performance for many use cases
- Large variations in available configurations across all major OEMs (including similar non-MGX servers)
- Three key features that customers want
  1. PCIe GPUs
  2. Flexible configurations
  3. Optional GPU Bridge (2-way or 4-way)
     NVLink for NVIDIA
     Infinity Fabric for AMD
  4. Easier management for space, power, and cooling

CISCO SECURE AI FACTORY

Compute

NVIDIA HGX & MGX

NVIDA Blackwell

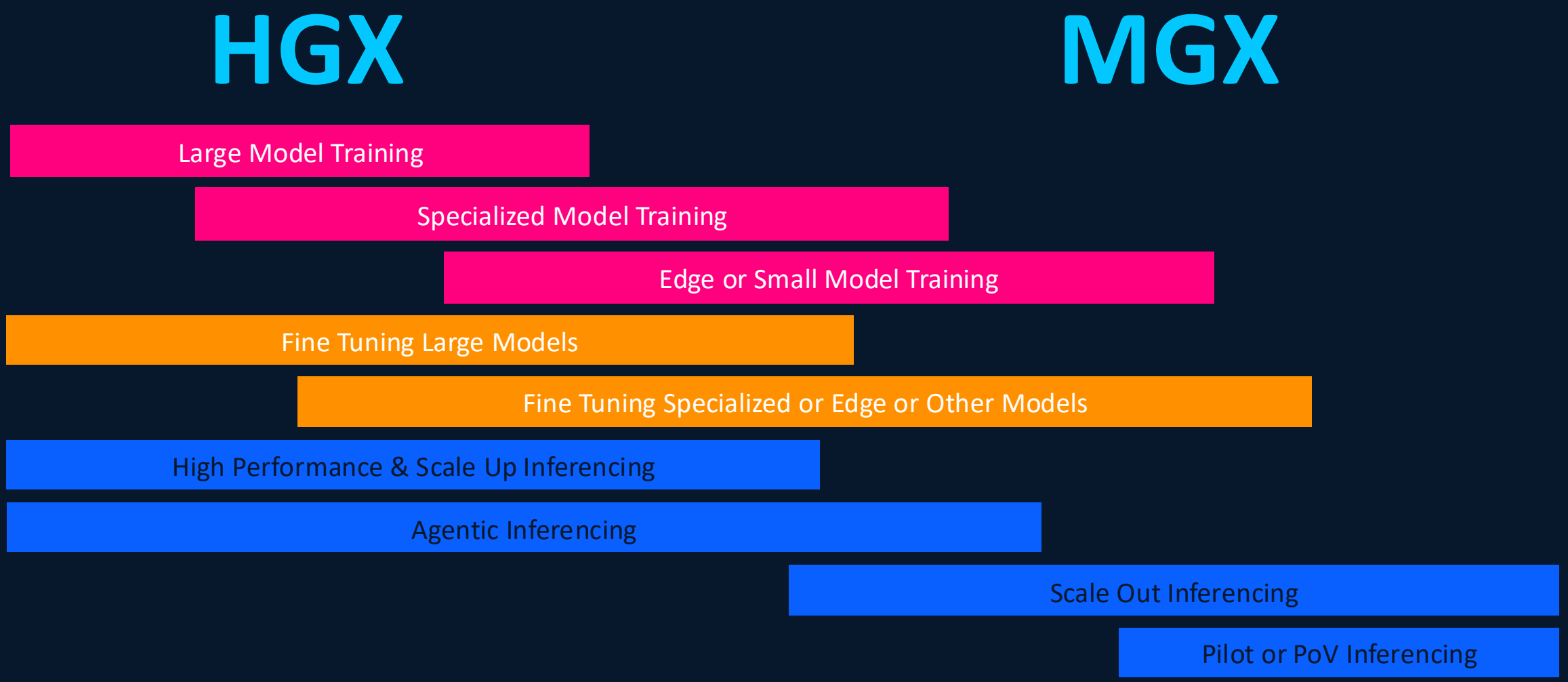
NVIDIA Bluefield

RTX PRO Server

GPU optimized

Reference Architectures

Dense GPU Platforms Use Case Map



HGX

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

High-Density Blackwell GPU Server

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

UCS Accelerated | UCS C880A M8

NEW

AVAILABLE NOW

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)

High-Density Hopper GPU Server

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

UCS ACCELERATED | CISCO UCS C885A

NVIDIA HGX™ reference design

Supporting 8 NVIDIA HGX™ H100, H200 and NVIDIA AI Enterprise software

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

Flexible, modular AI servers

"Start small and scale up" with AI

MGX

UCS ACCELERATED | CISCO UCS C845A

NVIDIA MGX™ reference design

With NVIDIA H100, H200, L40S, AMD MI210 GPUs

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

with NVIDIA RTX PRO 6000 Blackwell GPUs
[orderable now]

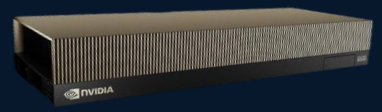
cisco

C845 - MDX

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

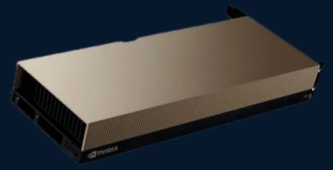
AMD:
- 8x AMD MI210 GPUs

Use Cases By GPU



NVIDIA RTX PRO 6000

- Multimodal inference, LLM inference, agentic AI, generative AI
- industrial AI, visual computing, vGPU
- Scientific computing
- Omniverse, Rendering, Media



NVIDIA H200 NVL

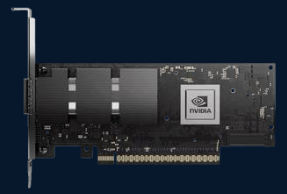
- Gen AI Training and Fine-Tuning
- LLM inference, fine tuning (>70B)
- Scientific computing (FP64)

NVIDIA L40S

- Language Processing
- Conversational AI
- Recommenders
- Graphics & Rendering
- Omniverse
- Virtual Desktops

Scale out with NVIDIA SmartNIC

- ConnectX-7 1x 400GB
- ConnectX7 2x 200GB



MGX Platform

Fine Tuning

NVIDIA H200 NVL

- High-performance LLM Inferencing
- Gen AI Training and Fine-Tuning

Inferencing

NVIDIA RTX PRO 6000

- Agentic and generative AI
- Industrial and physical AI applications
- Advanced scientific computing and rendering
- High-fidelity 3D graphics and video
- Hybrid AI/ML workflows
- Edge-to-core AI applications

Expanding RTX PRO Server Lineup with Flexible Form Factor

Cisco UCS® C240 M8 Rack Server

intel.

- Up to 2x Intel Xeon 6 processors
- Up to 8 TB DDR5 memory - up to 6400 MT/s
- PCIe 5.0 • 10/25/40/50/100/200 mLOMs and VICs
- Up to 36 E3.S drives
- Up to 28 SFF SAS/SATA/NVMe drives
- Up to 16 LFF SAS/SATA/NVMe drives

- **Up to 2X RTX PRO 6000 Blackwell Server Edition GPUs**

cisco

NOT part of SAIF Specification

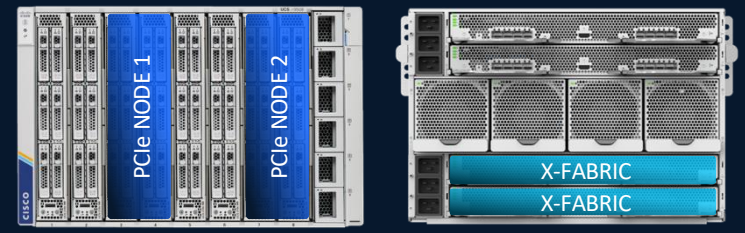
Dual wide PCIe Node and Switched X-Fabric PCIe Gen5

High-density GPU servers

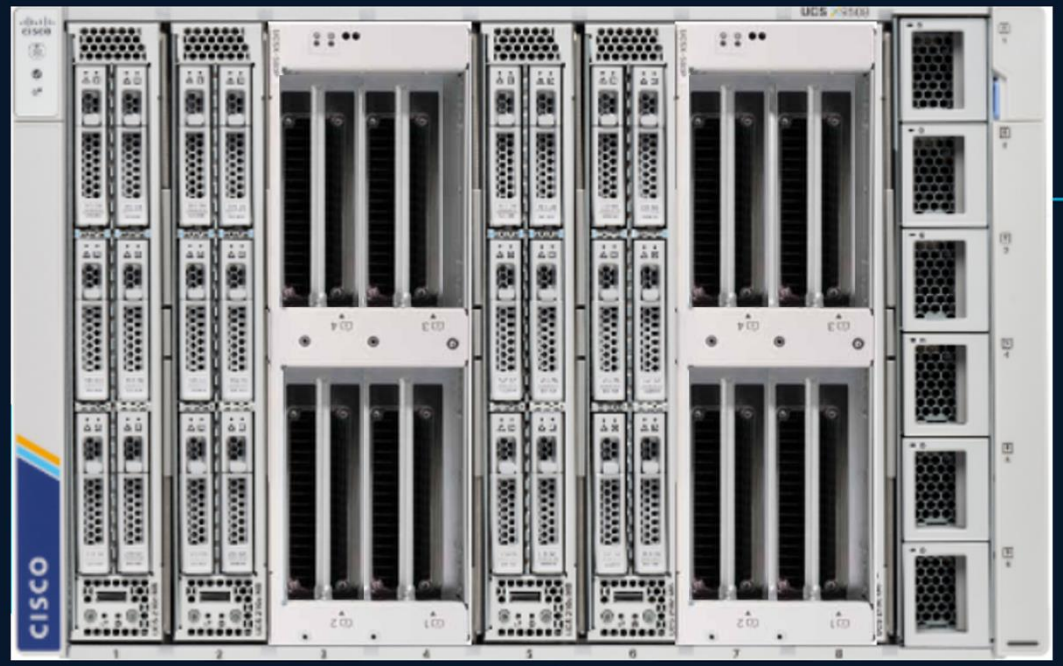
UCS X-fabric technology with PCIe node

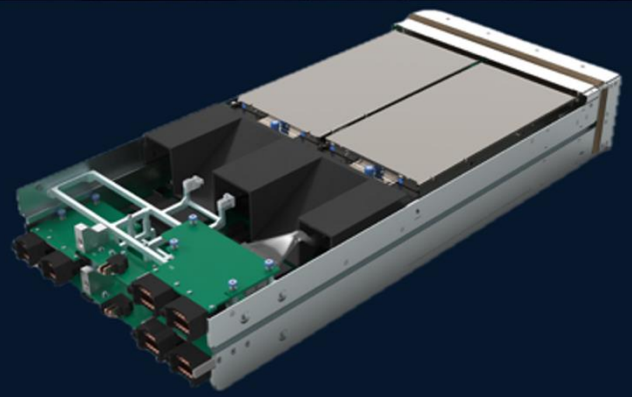
- PCIe Switching with PCIe Gen 5 connectivity
- 4x FHFL or HHHL GPUs per PCIe node
- Intra-host GPU interconnect with NVLink
- Intersight policy-based Management
- Inter-host scaling with RDMA over AI Fabric

Competitive differentiation with X-Fabric & X-Series



UCS X580p PCIe Node





- Double wide PCIe node for 4x FHFL GPU and PCIe G5 GPU support
  - Nvidia H200-NVL, RTX PRO 6000 & L40S
- Support multiple vendors: Nvidia, AMD*/Intel*
- NVLink bridge support
- Support up to 600W FHFL GPU
- Managed PCIe node with BMC support
- Policy based GPU management
- Ability to share GPUs across two Compute nodes

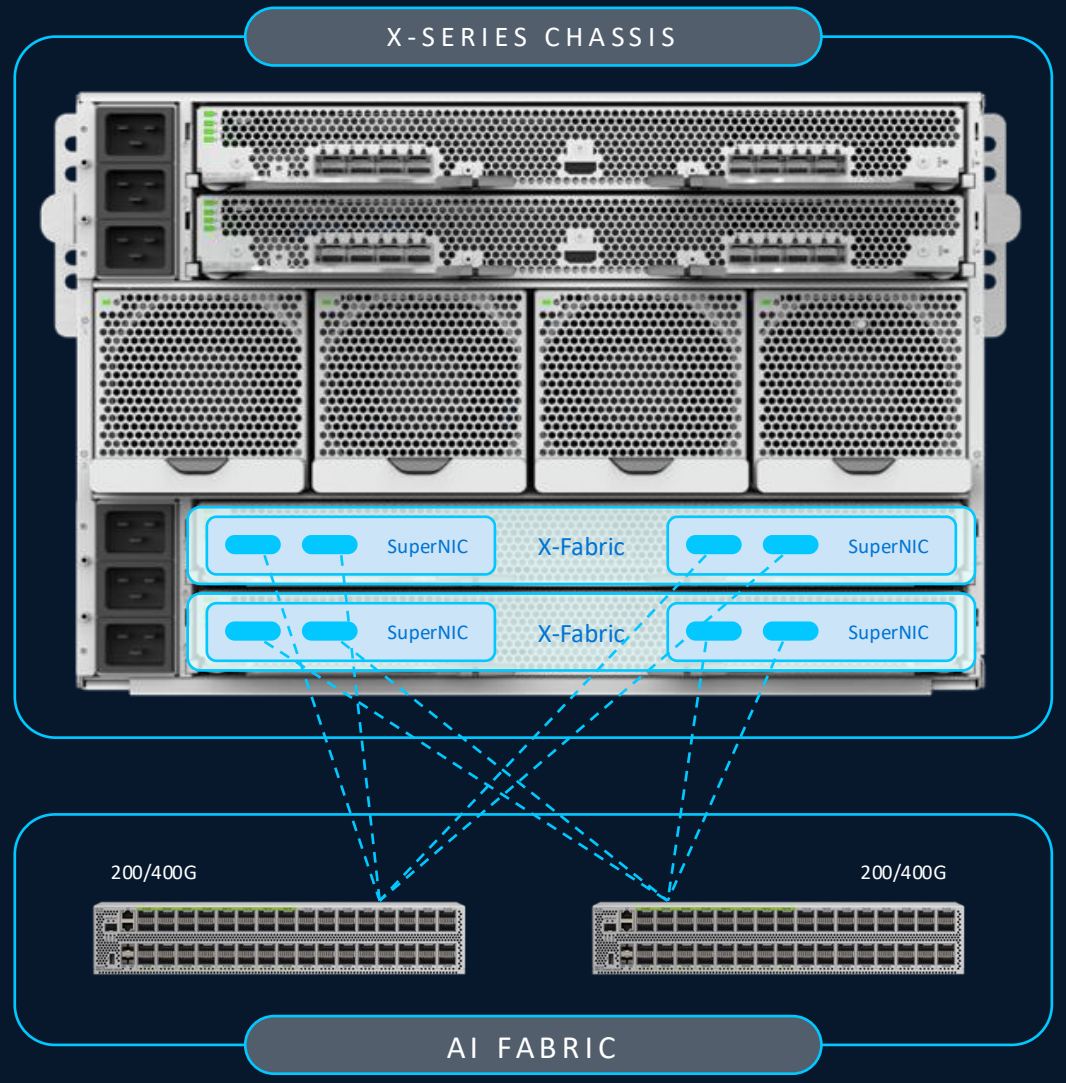
* AMD & Intel GPUs support will be post FCS

AI Cluster Expansion

GPU-to-GPU connectivity

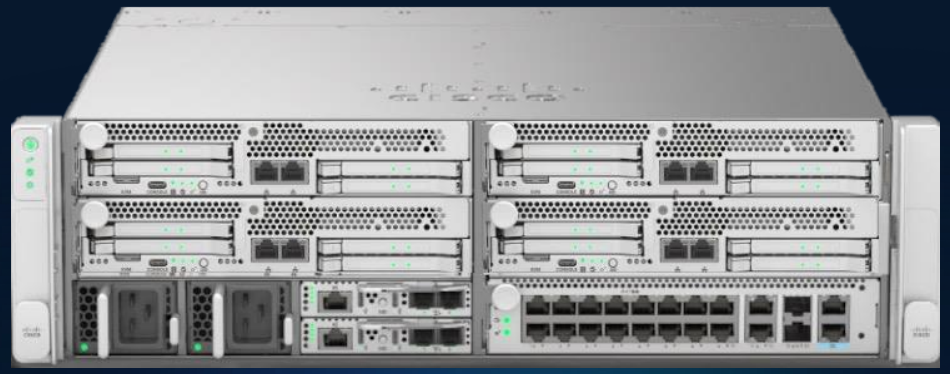
with XFM external ports

- X-Fabric Module with Gen5 PCIe switch
- SmartNIC Adapter for GPU East-to-West traffic
- 1 or 2 external ethernet ports based on adapter

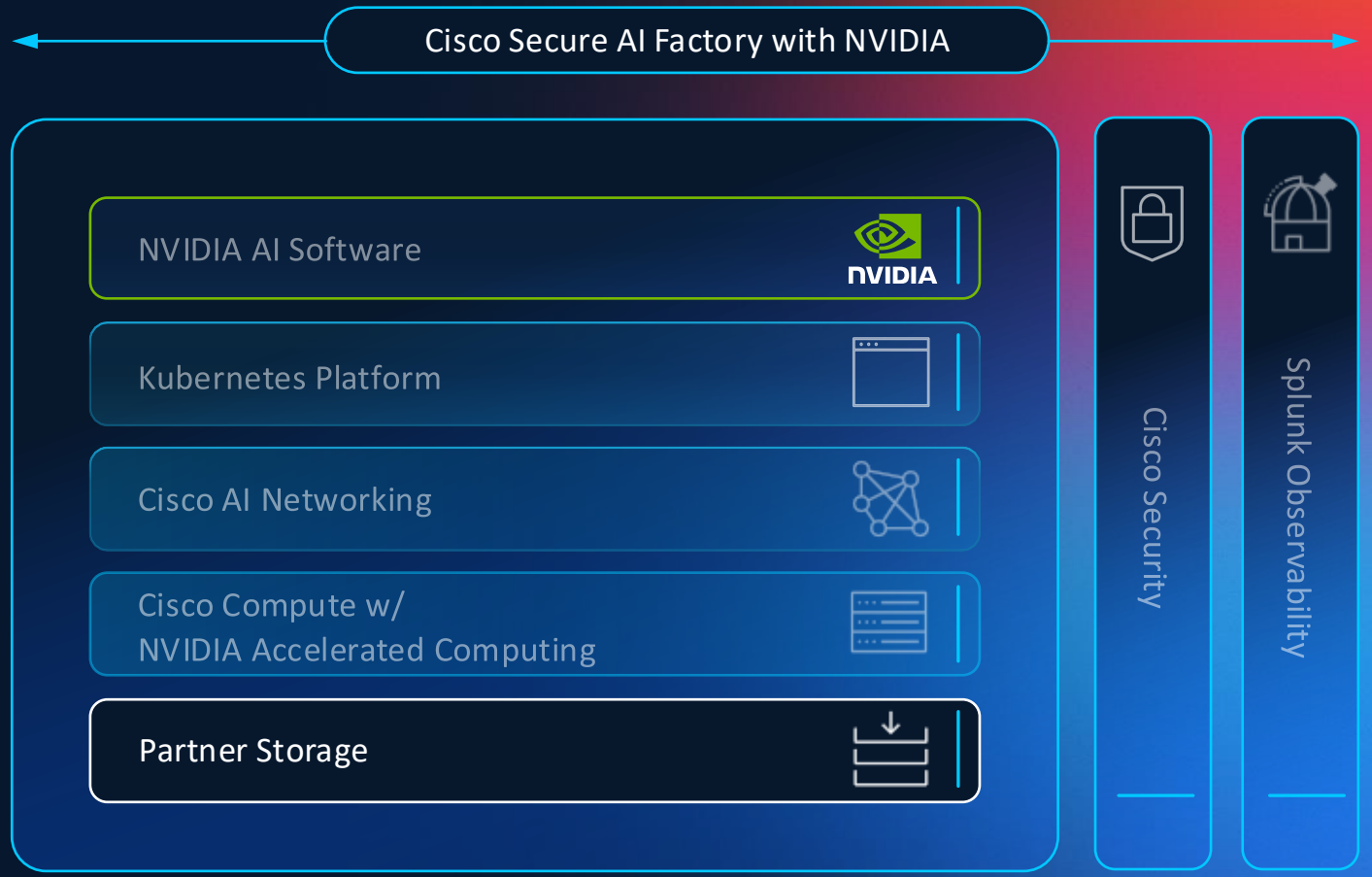


Cisco Unified Edge

Security at every layer to protect the edge

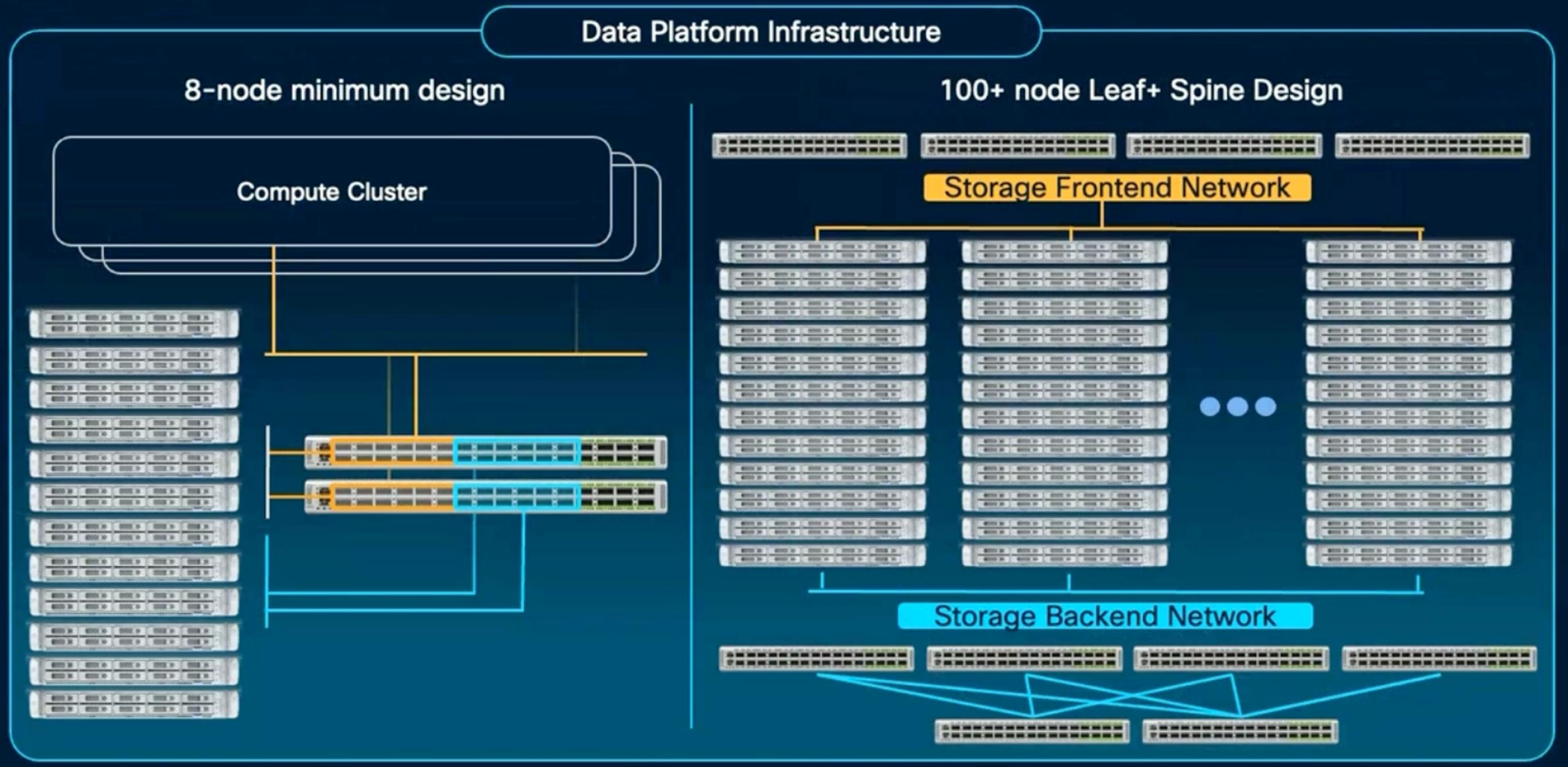


Secure AI Factory with NVIDIA, Data Infrastructure



Data Infrastructure with VAST

AI-Scale Data Architecture



Storage Fabric
Nexus 9000, Cisco 6000

Data Nodes
UCS C225-M8N

Data Infrastructure with Pure Flashblade

With 4-node Scale Unit

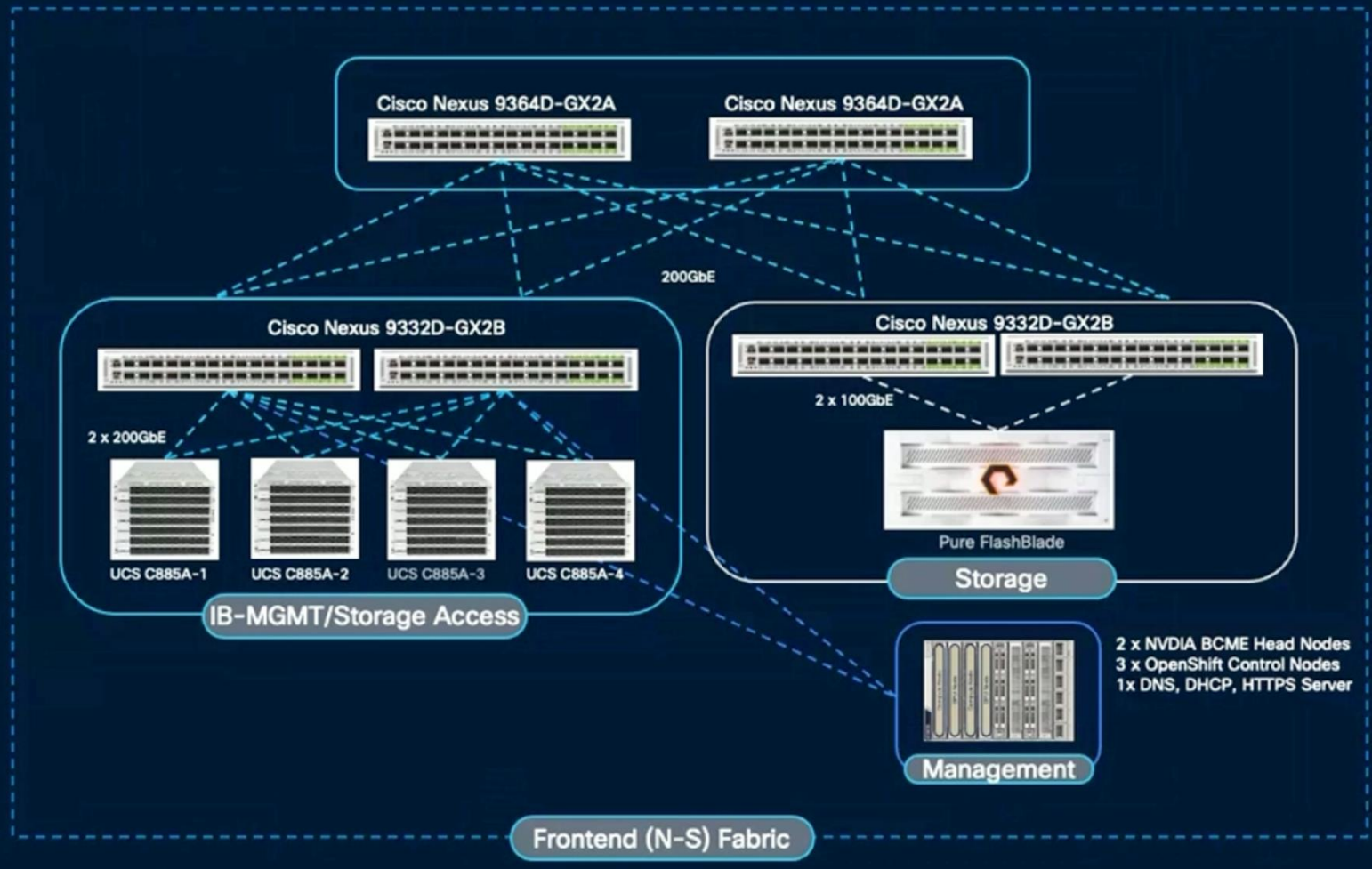
Available Models:

- //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)

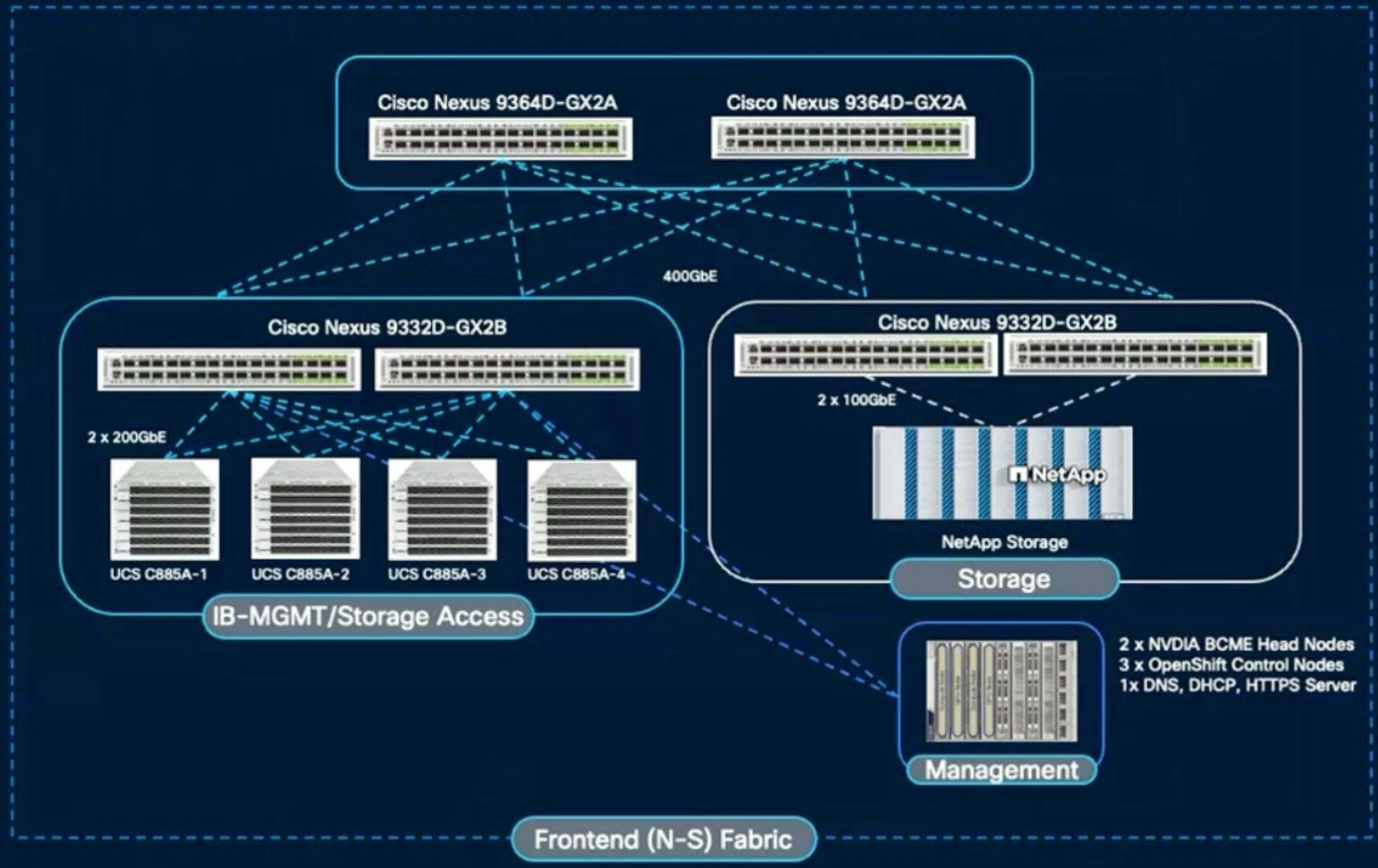
Scale to 10 chassis on //S200 or //S500; Requires 2-XFMs (Option 2)



Data Infrastructure with NetApp

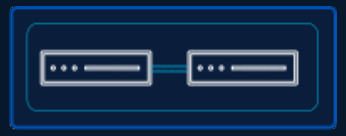
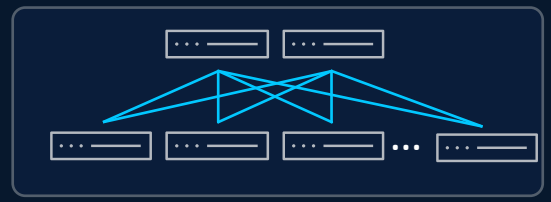
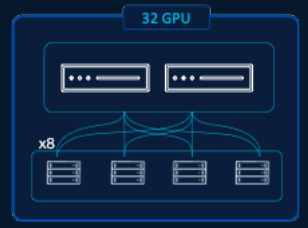
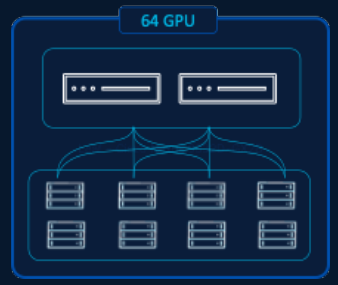
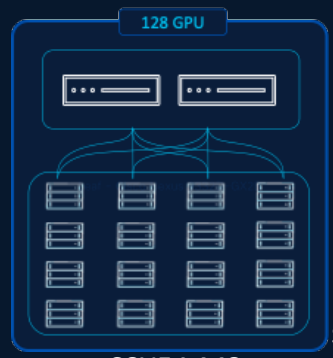
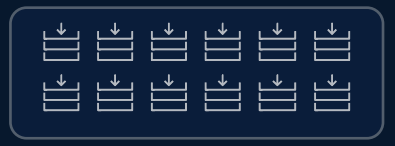
With 4-node Scale Unit

- 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



Flexible Stack for AI

Choose the size your business needs with elasticity for future

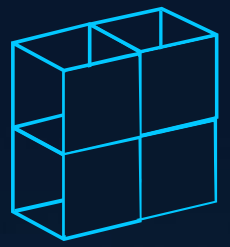
Software Stack	NVIDIA AI Enterprise	Red Hat OpenShift AI
GPU Options	NVIDIA	AMD
Network Fabric	**Collapsed Spine-Leaf** 	Spine-Leaf 	2-Spine, 2-8 Leaf Switches 
Compute cluster	Scale Unit Type 1  C8X5A M8 Scale to 128GPUs under single spine pair	**Scale Unit Type 2**  C8X5A M8 Scale to 128GPUs under single spine pair	**Scale Unit Type 3**  C8X5A M8 Scale to 256GPUs under single spine pair
Storage cluster	 VAST	 **NetApp**	 **Pure Storage**

cisco

Flexible deployment options

Build your own

Buy and deploy individual products, as needed

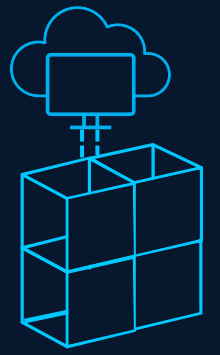


AI POD with on-prem network management

Modular, pre-validated infrastructure:

Full stack, buy & deploy

Backed by CVDs



AI POD with cloud based network management

Turnkey infrastructure:

Full stack, buy and deploy

Nexus Hyperfabric:

Cloud-managed Networking

Nexus Hyperfabric AI:

Cloud-managed physical infrastructure

Simplified Orderability

AI PODs

Faster time to value with pre-configured bundles

Deploy AI
with confidence

Orderable, validated AI-ready infrastructure stacks

Fully supported stack including Cisco and 3rd party components

AI advisor tool for configuration guidance

COMING SOON

Cisco AI-ready infrastructure stacks

Two Types of Pods

AI Workload PODs

Full-stack infrastructure for deploying AI workloads

- Training
- Optimization
- Inferencing

Purpose: AI Ready Infrastructure PODs, backed by CVDs, to deploy Enterprise AI workloads.

Examples: Generative and agentic AI applications, model training and optimization

Value: Full-stack validation and performance characterization to provide accelerated time-to-value.

AI Services PODs

Security, observability or data platform services

- AI Defense
- Splunk
- NVIDIA AI Data Platform
- ...

Purpose: Dedicated infrastructure PODs, backed by CVDs, for AI Security, Observability and Data Services

Examples: Cisco AI Defense, Splunk Observability, NVIDIA AI Data Platform.

Value: Ensure the security, efficiency and data readiness of your AI Factory.

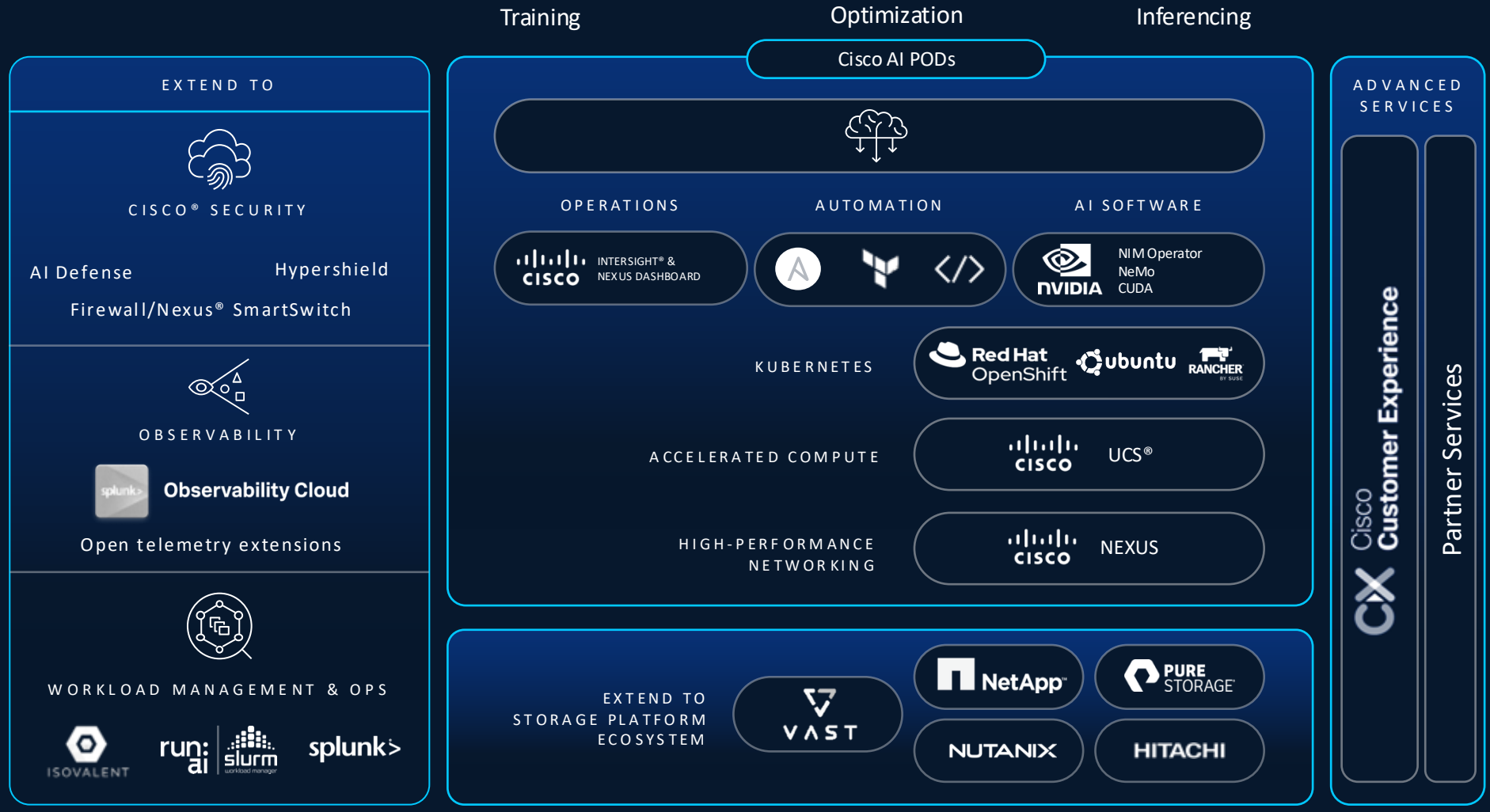
Cisco AI PODs

A scalable architecture, built to support any AI workload simply & efficiently

Deploy AI with confidence
Cisco CVD, NVIDIA Enterprise RA/NCP Cloud RA

Fully supported stack including Cisco and 3rd party components
Cisco CX
Success Track

Orderable, use case driven AI-ready infrastructure stacks
Inferencing.
Optimization.
Training.



Cisco AI PODs

Expanded design portfolio

Full AI lifecycle use-cases support

Based on Cisco Validated Designs

Pre-packaged AI-stack hardware + software platforms & automation

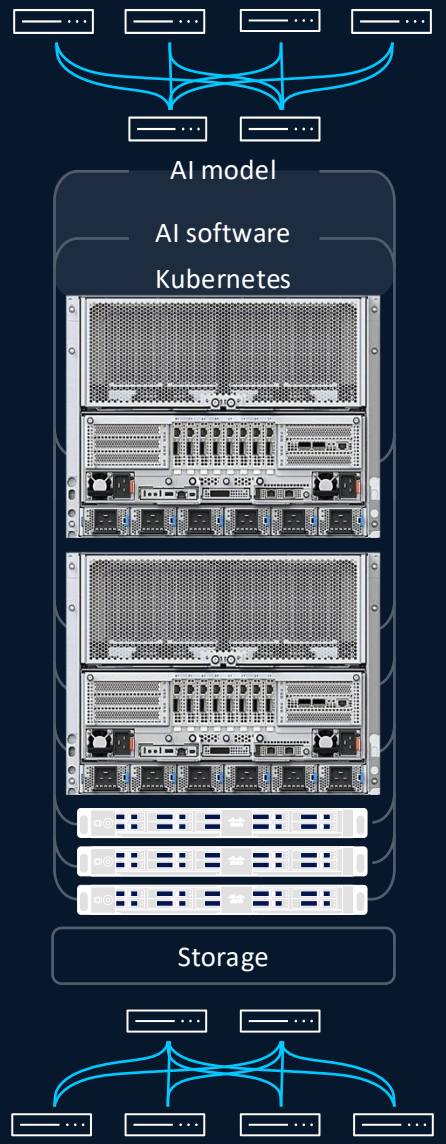
Latest NVIDIA and AMD GPU Compute

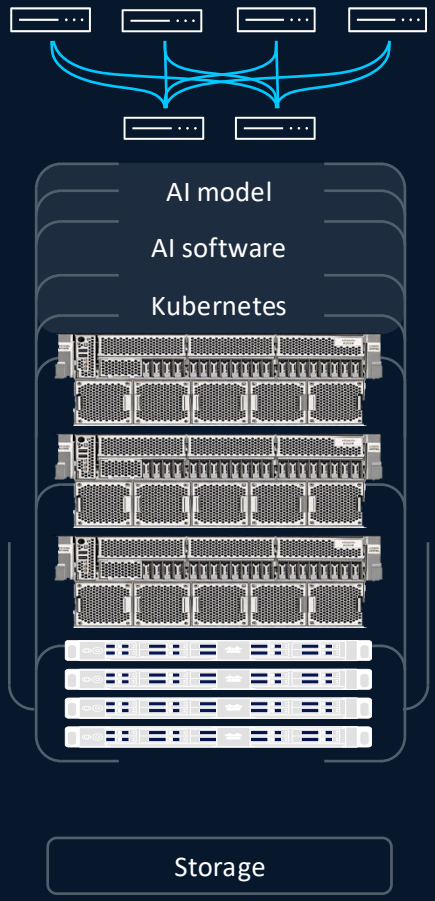
Adding VAST Data Storage

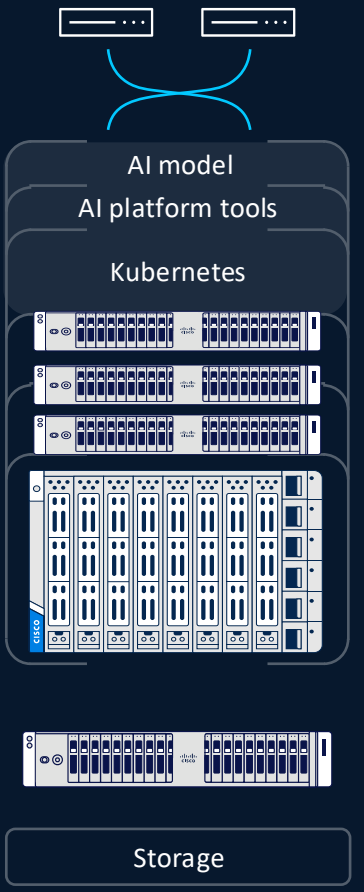
Nexus 9000 Switches

Cisco Validated Architecture

Training, optimization and inferencing







AIPOD-POD1 (Front End Network)

Ease of Ordering

Required

AI POD 1 is like buying a car that's ready to drive. The model's already trained — you just need to use it. Whether you're classifying documents, detecting images, or answering questions with a chatbot, this POD gives you everything you need to run AI at the edge or in a small data center. It's simple to deploy, doesn't need a lot of space, and works well for companies who want fast, reliable AI results without heavy compute or complex wiring. Think of it as "plug-and-go" AI.

Compute

AMD intel

- C-885A
- C-845A
- C-240A
- C-245A
- X-Series (M8 and Beyond)

Front-end fabric

- Nexus 9364D-GX2A
- Nexus 9332D-GX2B
- Nexus 9324C-SE1U
- Nexus 9364D-GX2A (Spine)

Management software

- Cisco Intersight®
- Nexus Dashboard (Physical, virtual)

Management cluster

- X-Series
- Nexus 93108TC-FX3
- C-Series

OS/Kubernetes

Red Hat OpenShift

Storage

MLB Included:

VAST NUTANIX

Other Options:

NetApp HITACHI

Adoption services

- Cisco CX
- MINT Service

cisco

Cisco Customer Experience (CX)

Cisco Customer Experience (CX) is a critical part of early qualification for AI Factory opportunities.

As customers move from AI experimentation into production, they frequently face cross-domain skill gaps, increasing architectural complexity, and **“first-of-its-kind”** deployment challenges that can delay success.

CX accelerates a customer’s *Time to First Intelligence (TTFI),* the moment their first AI use case is actually running, producing results, and utilizing the cluster as intended. It validates designs, optimizes configurations, and reduces idle time between hardware arrival and production readiness.

As customers scale, CX reduces their *Cost to True Scale (CTTS).* It ensures the foundational architecture, security posture, observability, and operational processes continue to perform reliably at larger GPU counts where small missteps become exponentially more expensive.

Key Components for a Secure AI Factory with NVIDIA

Ordering guides in additional resources

Hardware

Component	Technology	Description
Compute node	Cisco C880A, C885A, or C845A with NVIDIA Bluefield	NVIDIA Reference Configuration servers with L40S, RTX PRO, H100, H200, and B300 GPU options
Compute fabric (front & backend)	Cisco Nexus 9K or Hyperfabric	Up to 64-port 800 Gbps Ethernet, up to 512-100 Gbps port with breakout
Storage node	Cisco C225-M8N (or NetApp/Pure)	NVMe optimized 48 core single socket server
Storage fabric (front & backend)	Cisco Nexus 9K or Hyperfabric	Up to 64-port 800 Gbps
Control Node	Cisco C22x-M8	Single-socket general purpose compute
Out-of-band management fabric	Cisco Nexus 9K	1+ Gbps ethernet connectivity. Variable port density.
Optics	Cisco OSFP or QSFP	1 Gbps to 800 Gbps
Perimeter security	Hybrid Mesh Firewall	L3/L4 and full advanced firewall security solutions

Software

Component	Description	Quantity
NVIDIA Enterprise NVIDIA Run:ai	Best-in-class AI development tools, frameworks and orchestration	1x per GPU 1x per GPU
Red Hat Openshift (or other platform distro) Red Hat AI Accelerator	Enterprise Kubernetes platform OR bare metal OS	1x per server 1x per GPU
Vast Data (or NetApp/Pure below 1PB)	AI-scale, enterprise experience Data platform	1x per 100TB + remaining CPU cores
Cisco Intersight	On-prem or cloud compute management	1x per server
Cisco Nexus Dashboard or Cisco Nexus Hyperfabric | plus Spectrum-X add-on license	On-prem and cloud fabric automation and visibility	1x DCN/HF per switch 1x per backend GPU switch
Cisco Cloud Protection Suite: Includes Hypershield, Isovalent, Cilium Enterprise	eBPF powered networking, security, and observability for K8s environments	Up to… 10x workloads per server (or 116 Isovalent units)
Cisco AI Defense	Industry leading AI validation and model runtime security platform	1x per app
Splunk Enterprise Security Splunk Observability		Contact Splunk team

For this Modular Reference Design. Select **one or more Security Solutions.*
Other hardware options can be considered based on customer requirement.

AIPOD-POD1 (No Backend GPU network)

For Running AI Models, not Building Them

Required

AI POD 1 is like buying a car that's ready to drive. The model's already trained — you just need to use it. Whether you're classifying documents, detecting images, or answering questions with a chatbot, this POD gives you everything you need to run AI at the edge or in a small data center. It's simple to deploy, doesn't need a lot of space, and works well for companies who want fast, reliable AI results without heavy compute or complex wiring. Think of it as "plug-and-go" AI.

Compute

AMD intel

- C-885A
- C-845A
- C-240A
- C-245A
- X-Series (M8 and Beyond)

Front-end fabric

- Nexus 9364D-GX2A
- Nexus 9332D-GX2B
- Nexus 9324C-SE1U
- Nexus 9364D-GX2A (Spine)

Management software

- Cisco Intersight®
- Nexus Dashboard (Physical, virtual)

Management cluster

- X-Series
- Nexus 93108TC-FX3
- C-Series

OS/Kubernetes

Red Hat OpenShift

Storage

MLB Included:

VAST NUTANIX

Other Options:

NetApp HITACHI

Adoption services

- Cisco CX
- MINT Service

cisco

Sample Bill of Materials

Single Server Secure AI Factory – a great starter lab

Compute – AI Nodes

1 – Cisco UCS C845A
(NVIDIA MGX)

4 – NVIDIA RTX PRO 6000 GPU
(PCIe Form Factor, 8 per node)

Network

2 – Cisco Nexus 9324C-SE1U
(frontend switches with distributed segmentation)

4 – 100 Optics

2 – MPO Cables
(counts will vary with final design)

Software

4 - NVIDIA Enterprise

1 – AI Defense

1 – Intersight

2 – DCN Subscription

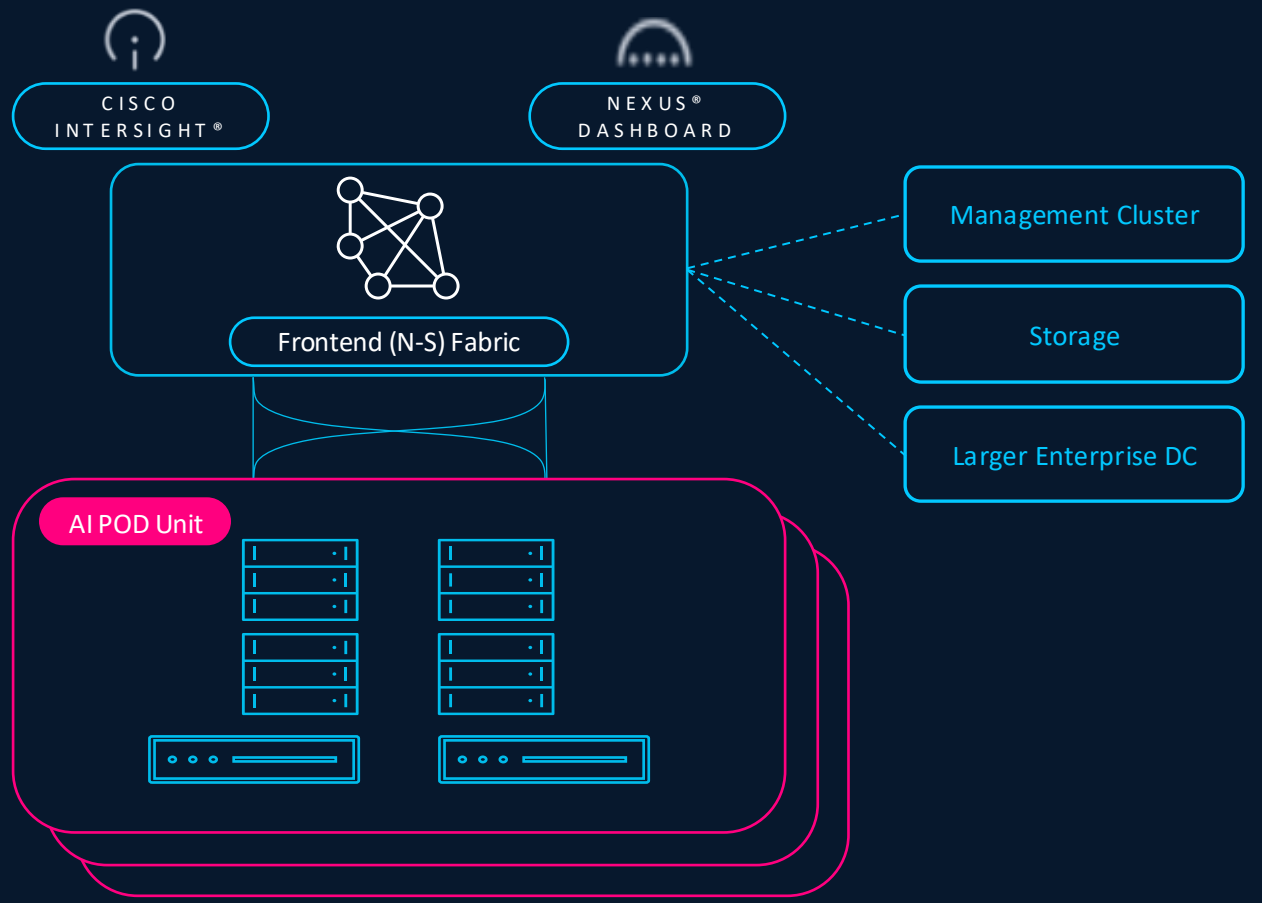
1 - Red Hat Openshift Container Platform

4 - Red Hat AI Accelerator

35 – Cloud Protection Units

Cisco AI POD 1

An AI-ready Infrastructure



Enterprise Reference Architectures < 1024 GPUs

Software for AI

NVIDIA

NVIDIA Enterprise

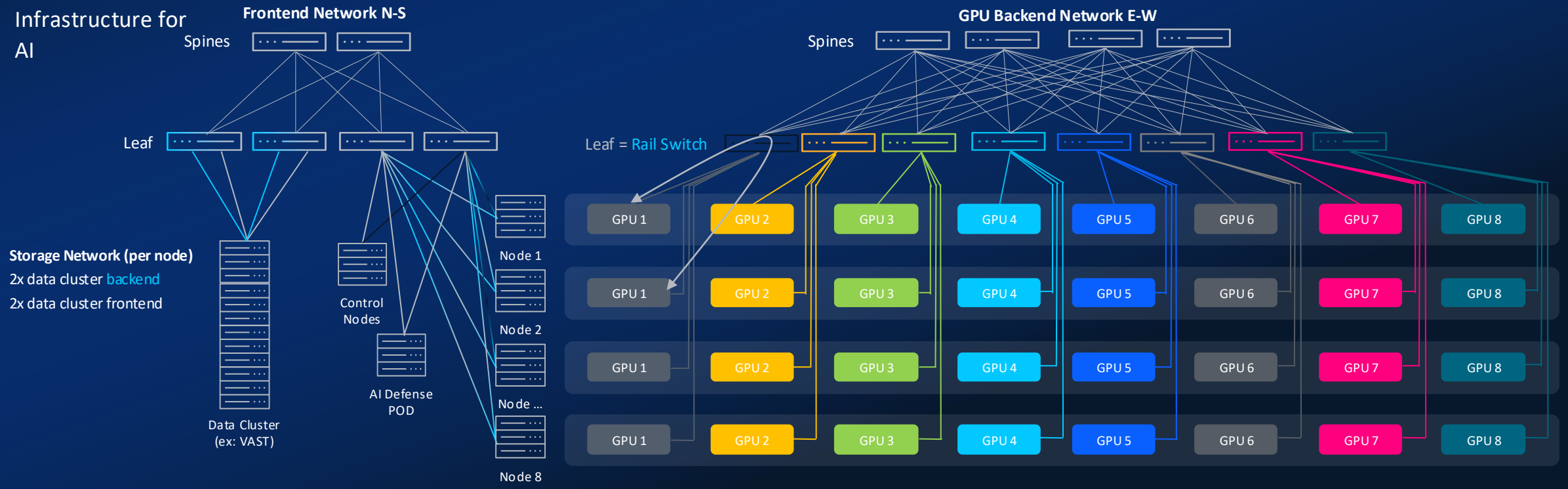
NeMo | NIM | Blueprints

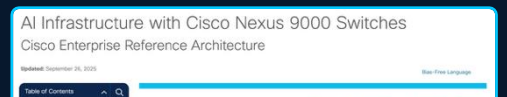
NVIDIA Run:ai

AI Workload & GPU Orchestration

KUBERNETES OR BARE METAL OS PLATFORM

Infrastructure for AI





Order in a few simple steps

Select a product using one of the following major line SKUs:

1 Product ID	Description							
	Scale units	Back-end fabric	Compute	Front-end fabric	Management software	Management cluster (Optional)	OS/ Kubernetes	Storage (Optional)
AIPOD-POD1	N/A	N/A	AMD or Intel C-885A C-845A C-240A C-245A X-Series (M8 and beyond)	Nexus 9364D-GX2A Nexus 9332D-GX2B Nexus 9324C-SE1U Nexus 9364D-GX2A (Spine)	Cisco Intersight Nexus Dashboard (Physical, virtual)	UCS X-Series Nexus 93108TC-FX3 C-Series	Red Hat OpenShift	VAST Nutanix NetApp Hitachi
AIPOD-POD2	SU 1 SU 2 SU 3	Nexus 9332D-GX2B Nexus 9364D-GX2A Nexus 9364E-SG2 Nexus 9364D-GX2A (Spine)	AMD C-885A C-845A		Cisco Intersight Nexus Dashboard (Physical, virtual)			

2 *Select Input Power & QSFP Cables*

3 *Select Virtualization*

4 *Select Adoption Services (Optional): Cisco CX or MINT services*

More Resources:

Data Sheet

Ordering Guide

SU1 **Node<=4**

SU2 **Node<=8**

SU3 **Node<=16**

SU: Scale units

- A scale unit is a repeatable, pre-designed bundle of UCS nodes and a pair of Nexus leaf switches.
- Resembling a lego block of AI infrastructure, they are the foundational elements of the AI POD2 systems
- Scale units come in different sizes with varying performance capabilities (see example)

Questions?
ask-aipod@cisco.com

AIPOD-POD2 (Backend GPU network)

- For companies that want to customize or build AI

AI POD 2 is more like a garage full of high-end parts and tools—built for people who want to *train, fine-tune,* or *customize* their AI models.
It supports large-scale operations where GPUs need to talk to each other at high speeds, like training a model on your proprietary data or refining a foundation model to your industry. This POD is ideal if you need serious computing power, are managing big datasets, and want full control over how your AI behaves. It's not just using AI—it's building the AI engine itself.

Required

SU1 **Node<=4**

SU2 **Node<=8**

SU3 **Node<=16**

Compute

AMD intel

- C-880A
- C-885A
- C-845A

Back-end fabric

- Nexus 9332D-GX2B
- Nexus 9364D-GX2A
- Nexus 9364E-SG2
- Nexus 9364D-GX2A (Spine)

Front-end fabric

- Nexus 9364D-GX2A
- Nexus 9332D-GX2B
- Nexus 9324C-SE1U
- Nexus 9364D-GX2A (Spine)

Management software

- Cisco Intersight
- Nexus Dashboard (Physical, virtual)

Management cluster

- X-Series
- Nexus 93108TC-FX3
- C-Series

OS/Kubernetes

Red Hat OpenShift

RANCHER BY SUSE

Storage

MLB Included:

VAST NUTANIX

Other Options:

NetApp HITACHI

Optics

- AI-Optics-MLB

Adoption services

- Cisco CX
- MINT Service

Sample Bill of Materials (C845A w/ RTX PRO 6000)

32 GPU (4 node) Secure AI Factory with NVIDIA – Scalable to 512 GPUs

Compute – AI Nodes

4 – Cisco UCS C845A
NVIDIA MGX 2-8-5-400G Design
1x B3220 2x200G
4x SuperNIC 1x400G

32 – NVIDIA RTX PRO 6000 GPU
(PCIe Form Factor, 8 per node)

Network

6 – Cisco Nexus 9364E 800G
(4 Leaf Switches, 2 Spine Switches)

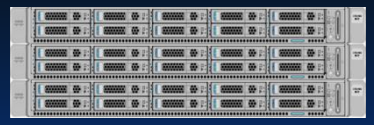
4 – Cisco Nexus 9364D 400G
(2 Leaf Switches, 2 Spine Switches)

2 – Cisco Nexus 9332D
(VAST Backend)

2– Cisco 93108TC-FX3
(Management Switches)

248x 200 to 800G Optics | 128x Fiber Cables
(counts will vary with final design)

Storage



12 – Cisco UCS C225 M8
(VAST EBox Cluster)

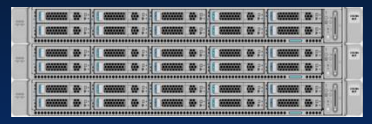
Security - AI Defense Nodes

2 – Cisco UCS C845A
(NVIDIA MGX)

8 – NVIDIA L40S PCIe GPU
(PCIe Form Factor, 4 per node)

5 – Cisco UCS C225 M8
(Control Cluster K8s & SLURM)

3 – Cisco UCS C225 M8
(Nexus Dashboard Cluster)

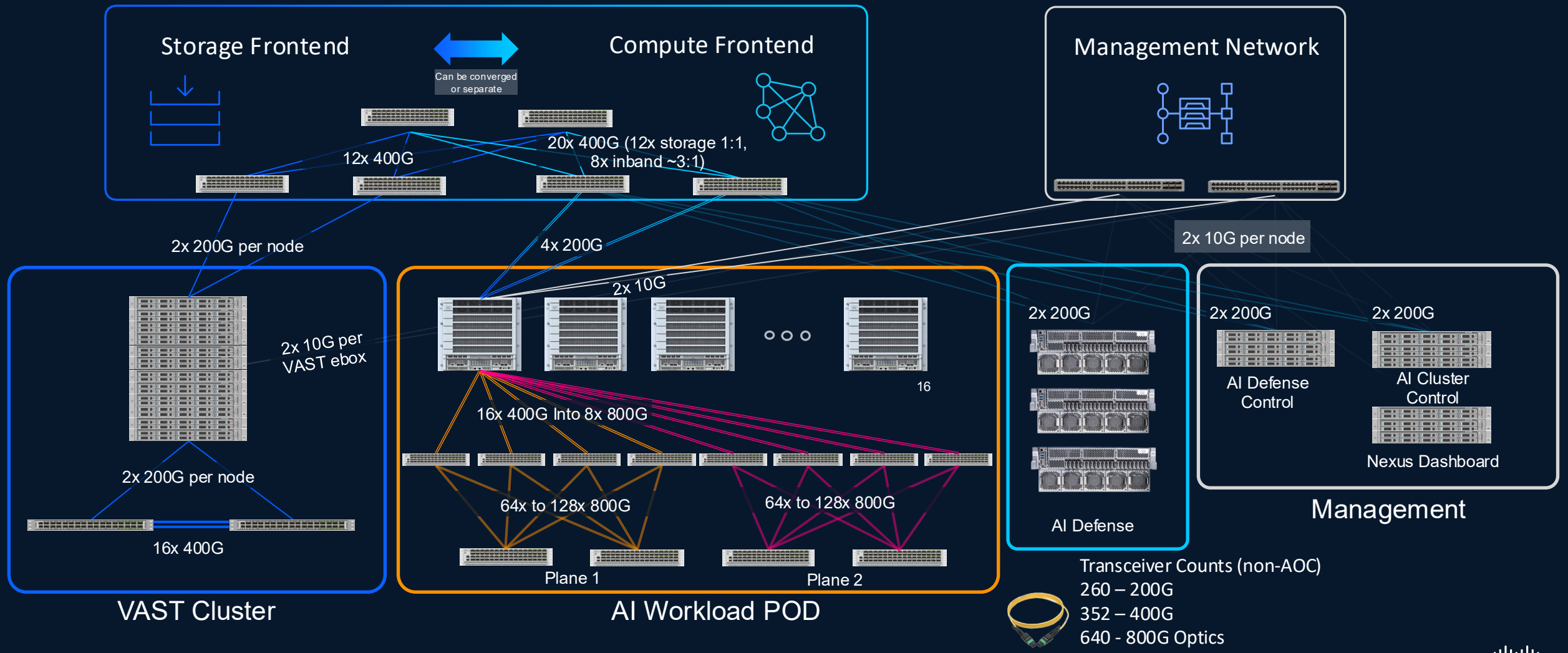


3 – Cisco UCS C220 M7
(AI Defense Mgmt Cluster)

Software

40 - NVIDIA Enterprise
32 – Run:ai
5 – AI Defense
26 – Intersight
14 – DCN Subscription
6 – Spectrum-X Add-on
136 – Cloud Protection
6 - OpenShift Container Platform
40 - Red Hat AI Accelerator
11 – VAST Capacity
348 – VAST Cores

Sample Design Secure AI Factory

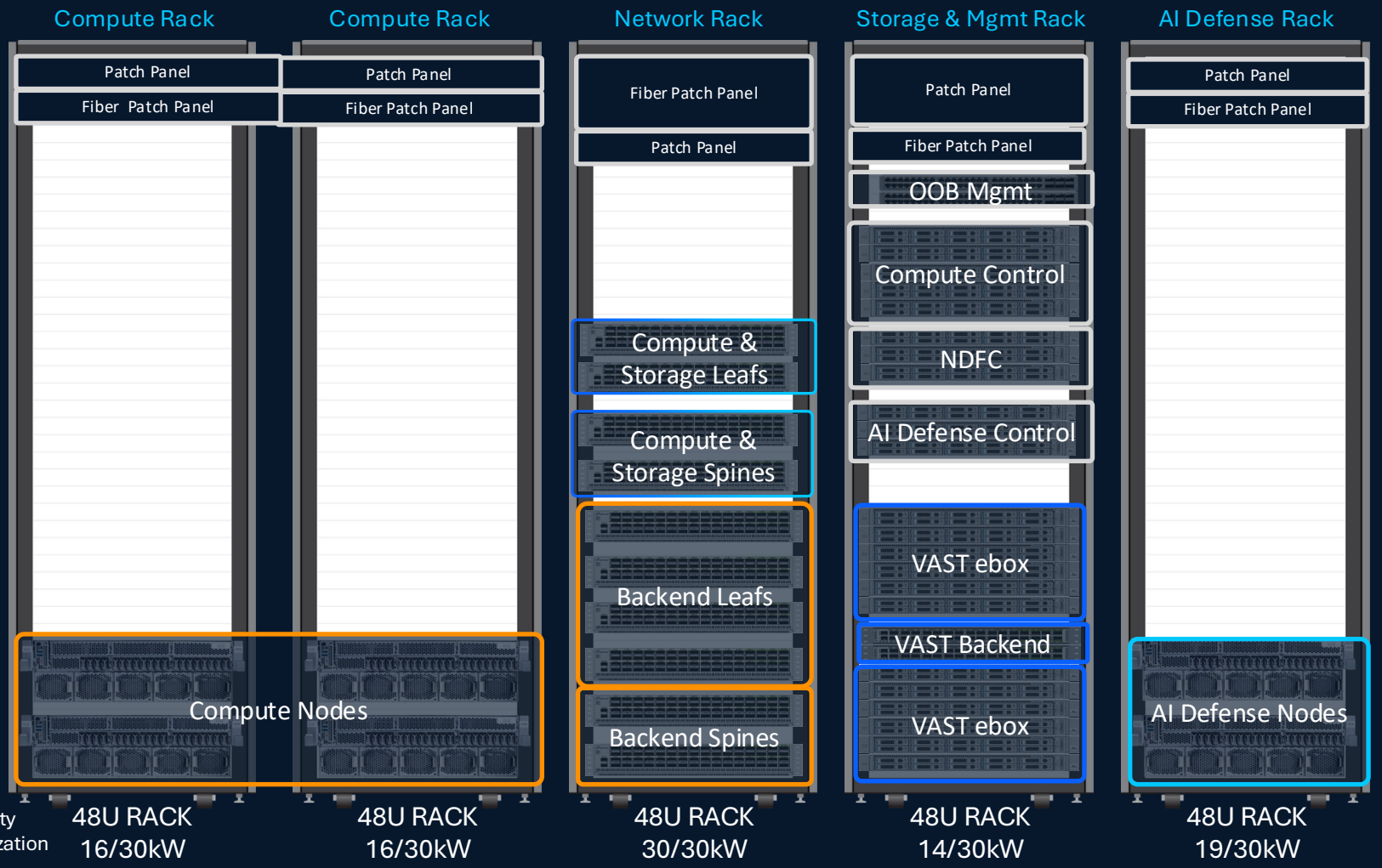


Transceiver Counts (non-AOC)
260 – 200G
352 – 400G
640 - 800G Optics
626 – Fiber Cables

*counts will vary with final design
*Nexus dashboard has 50G VIC

Sample Rack Design

Remember, this will vary with final design and per rack power availability



Cisco Advantage	Real Business Value $$
Well Integrated Platform	Lower Integration Cost & Faster Time To Deployment
Ubiquitous Security	Remove Complexity & Lower Management Costs
Common Policy	Lower Deployment & Management Costs
Observability	Faster Problem Resolution & Lower Operational Costs

Bringing AI to the Enterprise