CISCO Connect

Compute Infrastructure for the AI Era

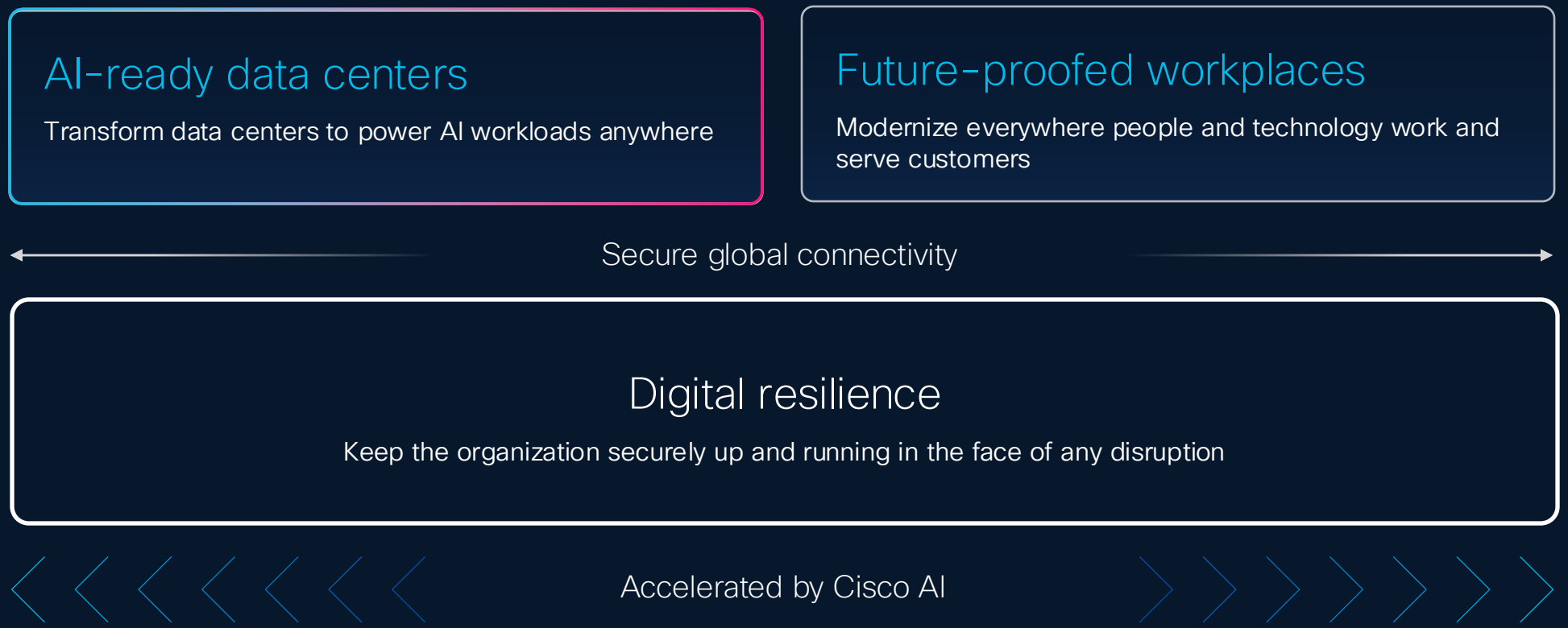
William J. Roberts, Solutions Engineer – C+AI, Cisco Systems, Inc.

April 7, 2026

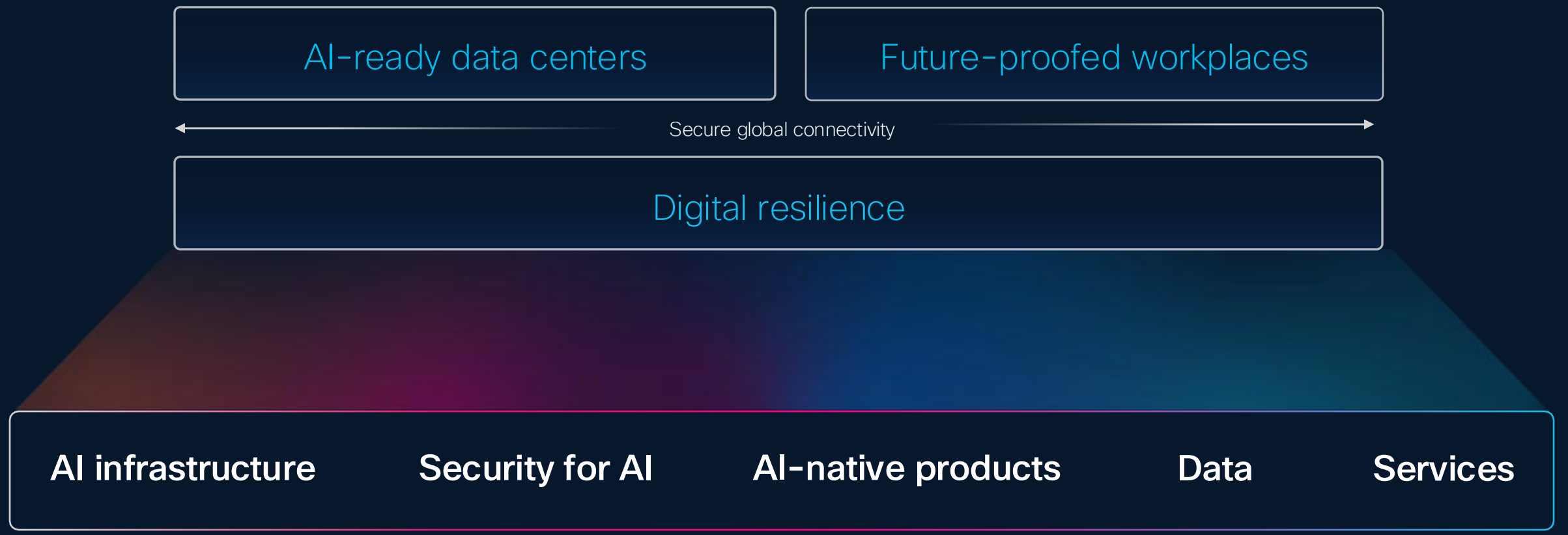
Agenda

1. AI Basics Abridged
2. Full Stack AI Infrastructure
3. Addressing AI Security
4. Dense Compute Systems for AI
5. Emerging Trend: Disaggregated Inference

Cisco powers how people and technology work together across the physical and digital worlds



Cisco AI: Accelerating Outcomes

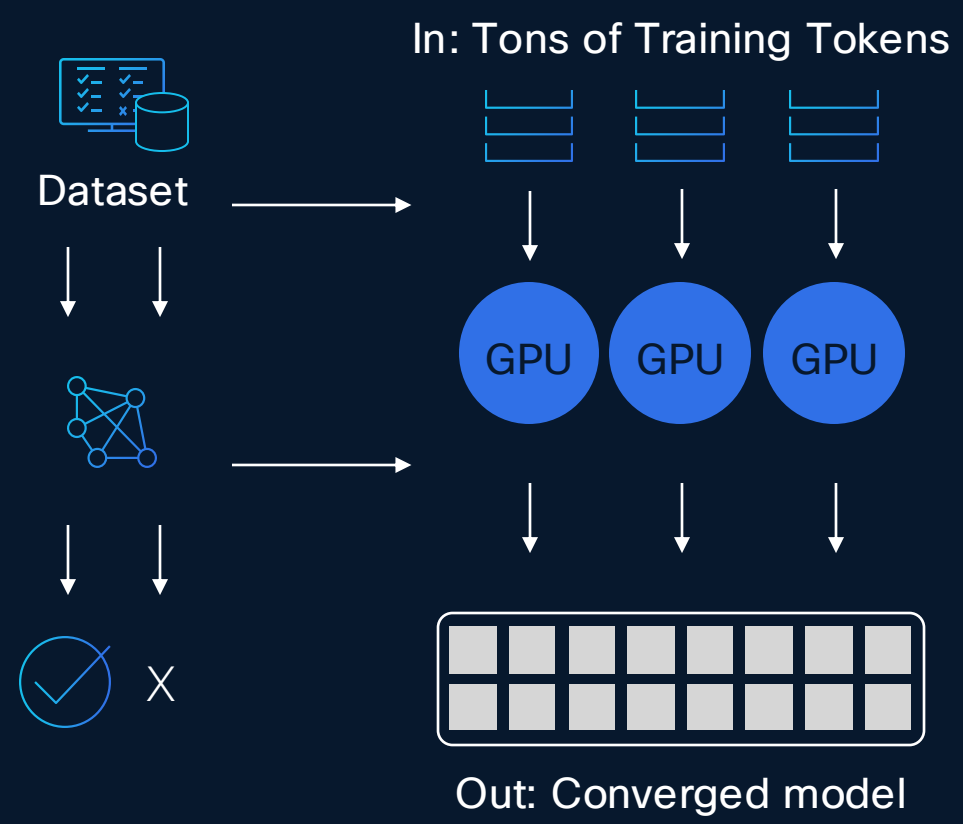


AI Workload Demands on Infrastructure

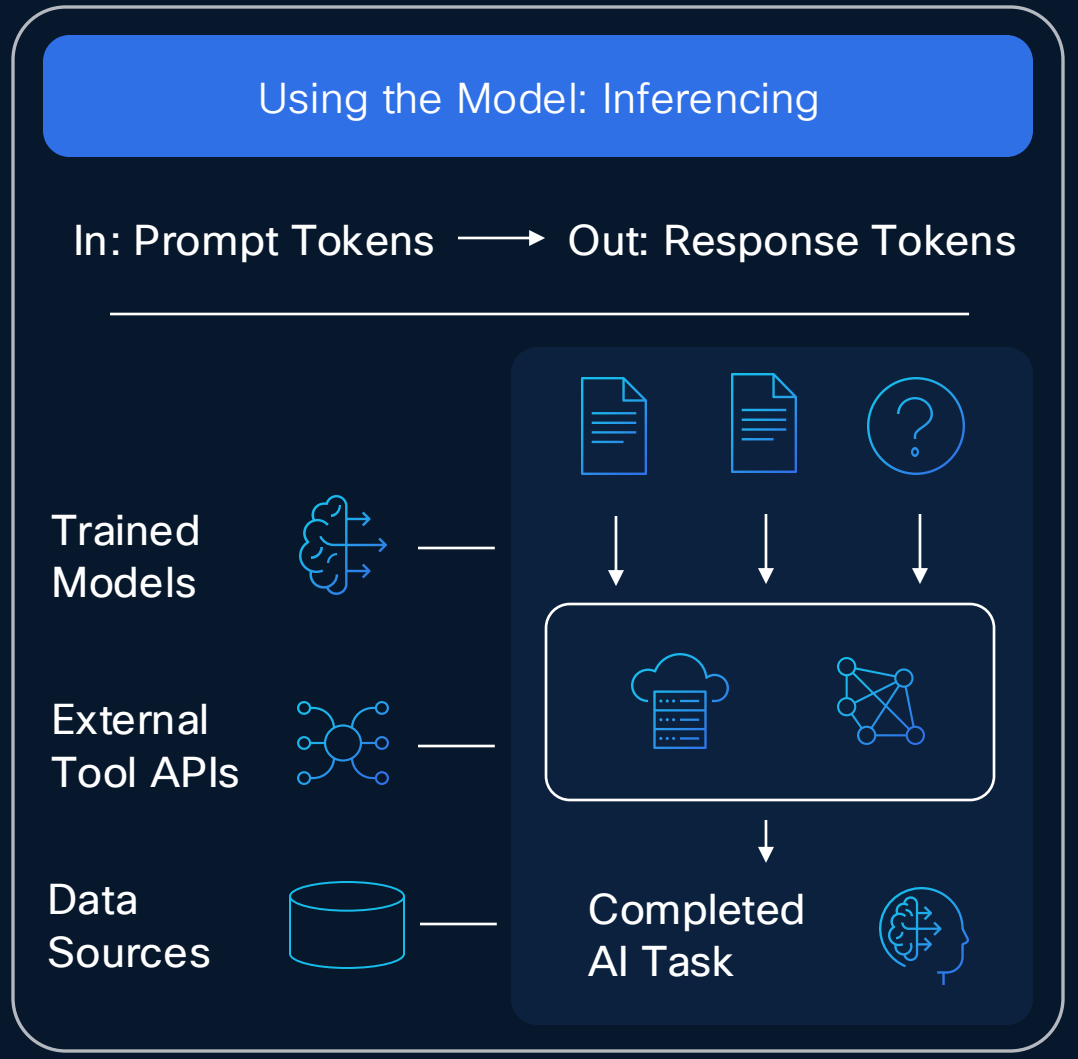
An Abridged Review

AI Workload Types & Components

Building the Model: Training and Fine-tuning



Using the Model: Inferencing



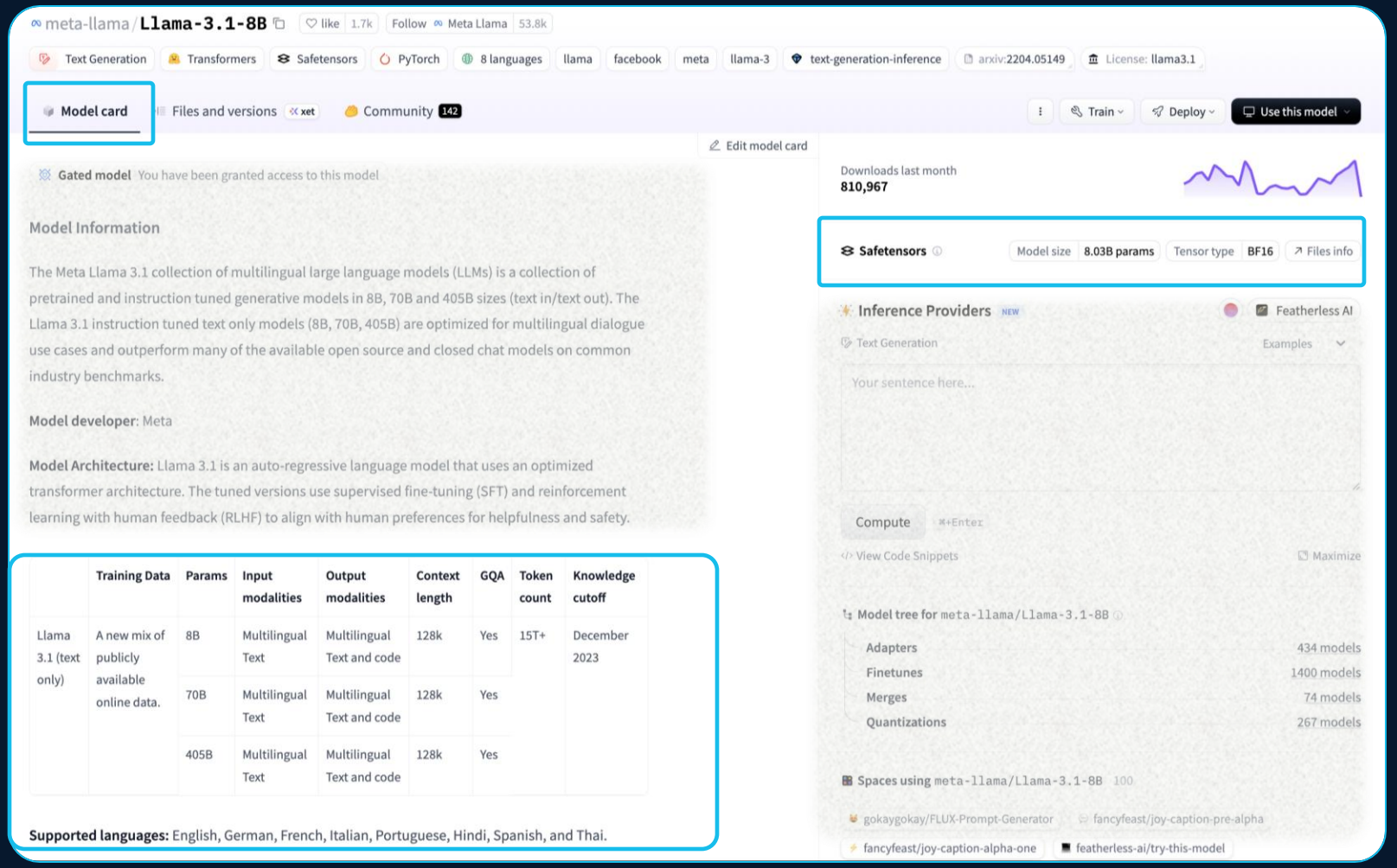
Model Card – an LLM's "Data Sheet"

Model Size: parameter count

Default Precision: bytes per parameter

Context Length: max tokens per single request stored in...

KV Cache: the model's "working memory"

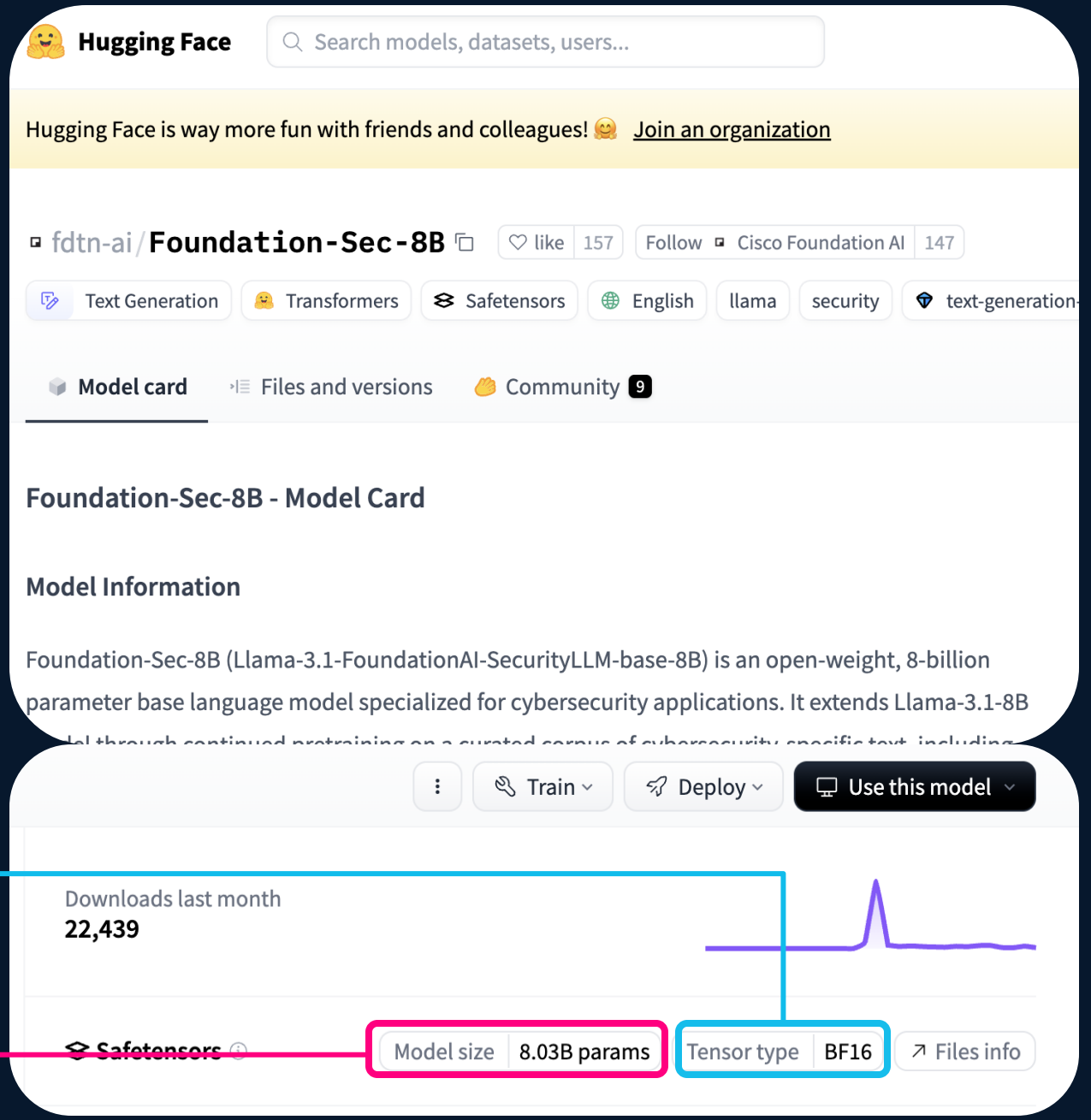


Source: https://huggingface.co/meta-llama/Llama-3.1-8B

Calculating Model Memory

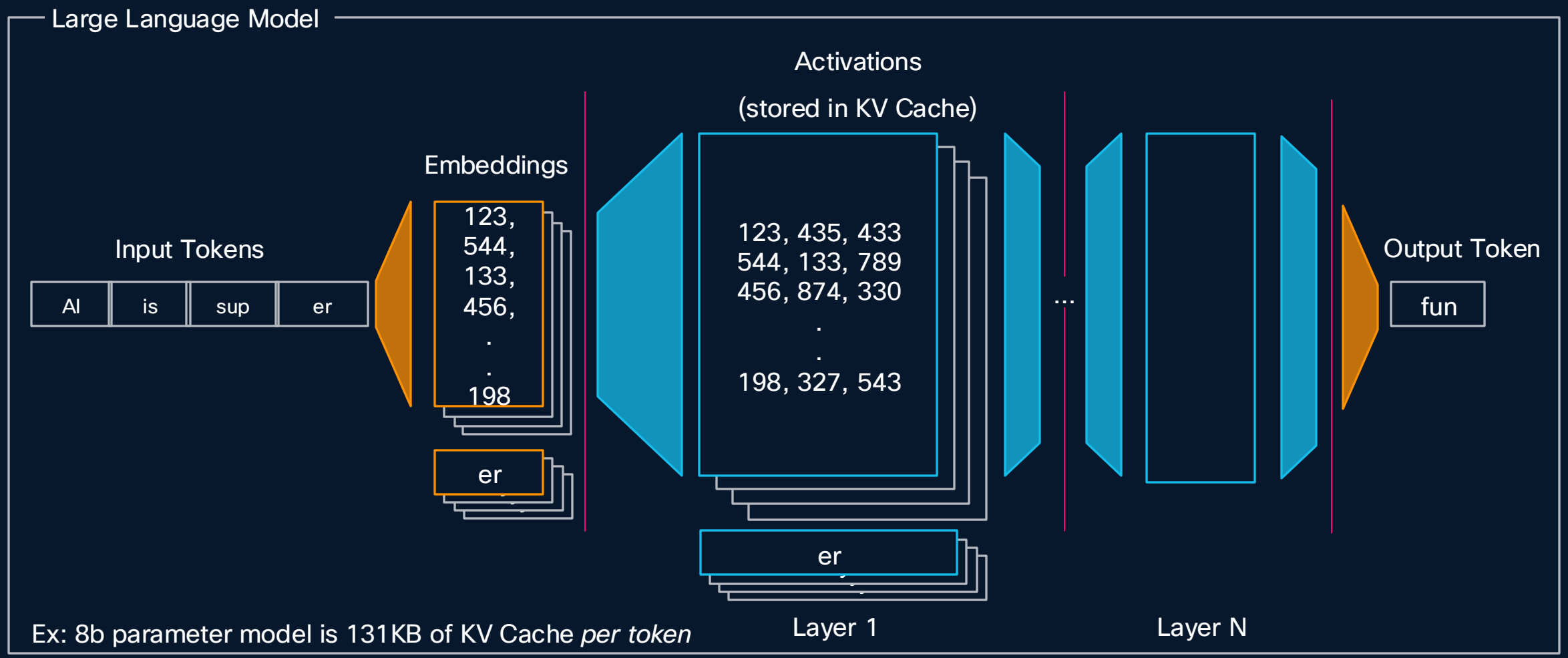
Parameters * precision = size

8 billion * 2 bytes = 16GB

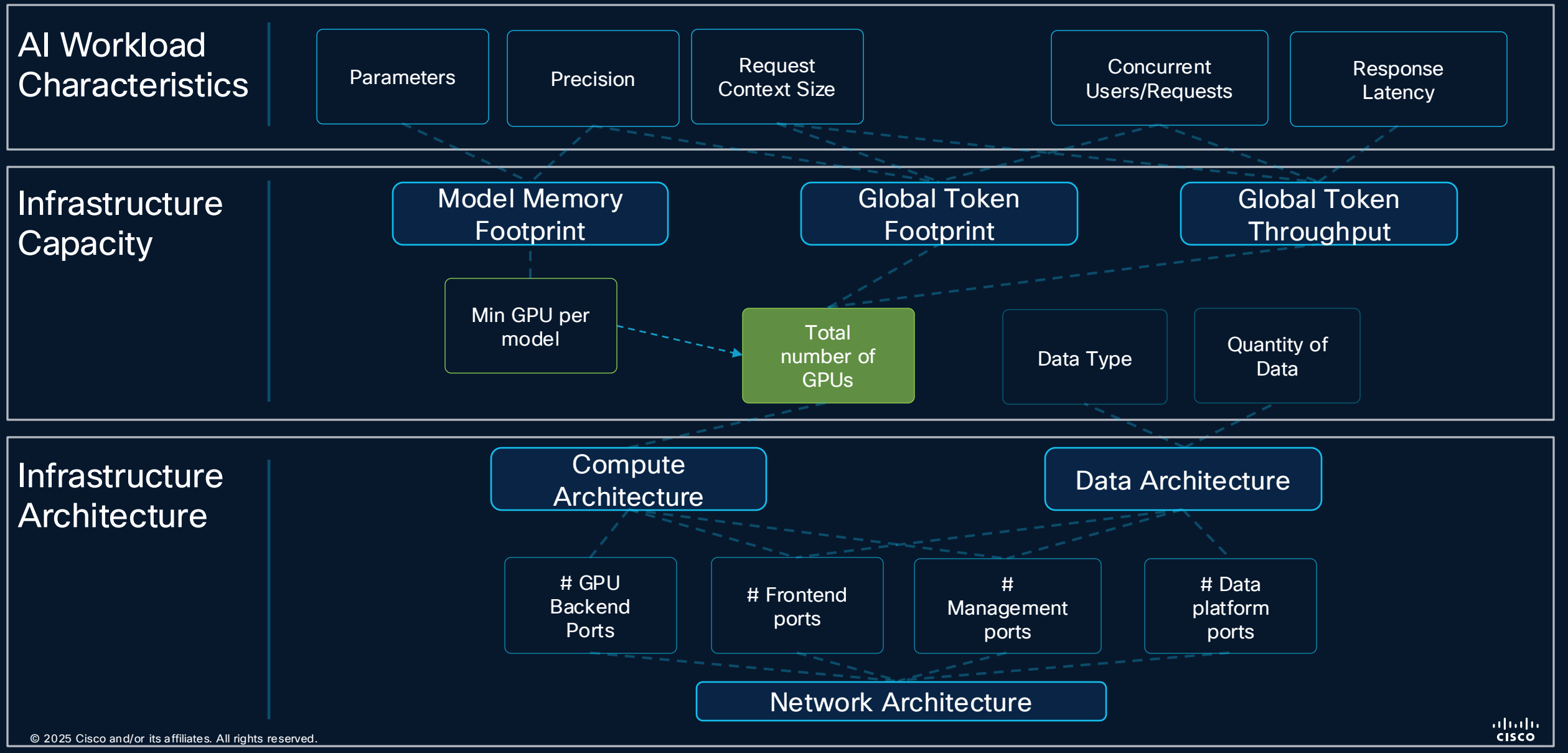


Why AI Takes Up So Much Memory

Data Expansion From Tokens, Embeddings, Activations, and KV Cache



Mapping AI Inference Workload Design to Capacity Requirements



Full Stack AI Infrastructure

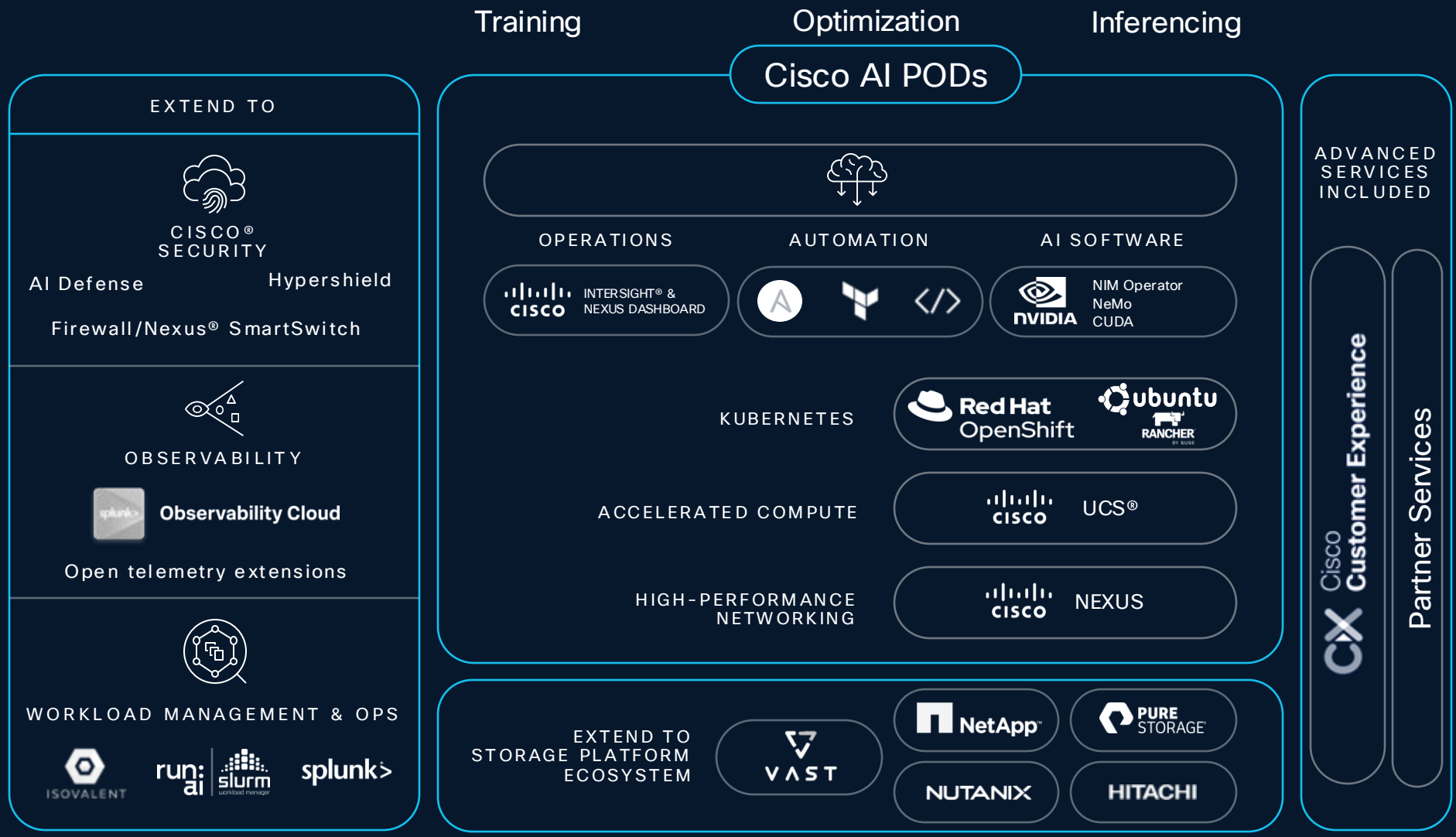
Cisco AI PODs

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

Deploy AI with confidence
Cisco CVD, NVIDIA ERA

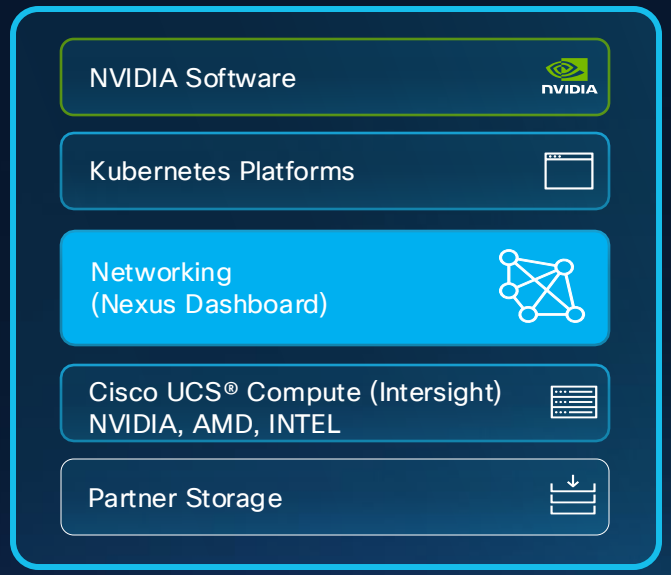
Fully supported stack including Cisco and 3rd party components
Cisco CX Success Track

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



Cisco AI PODs: Flexible Operating Models

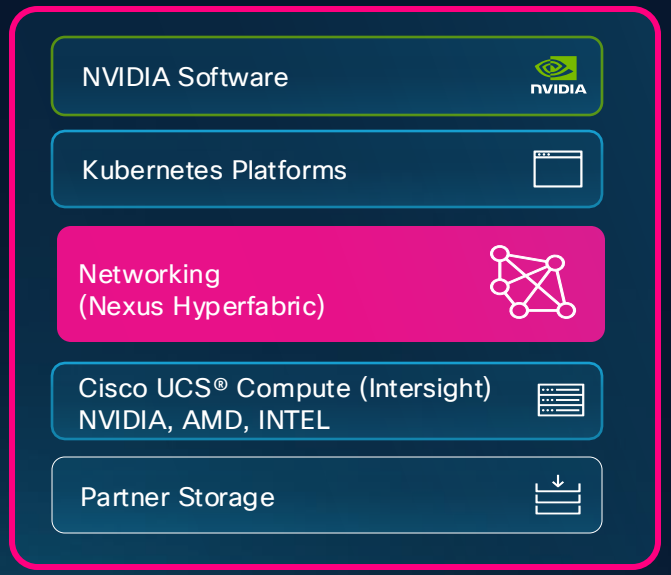
AI POD w/ On-prem management



Modular, pre-validated infrastructure:

- Full stack, buy & deploy
- Nexus Dashboard: On-prem networking management

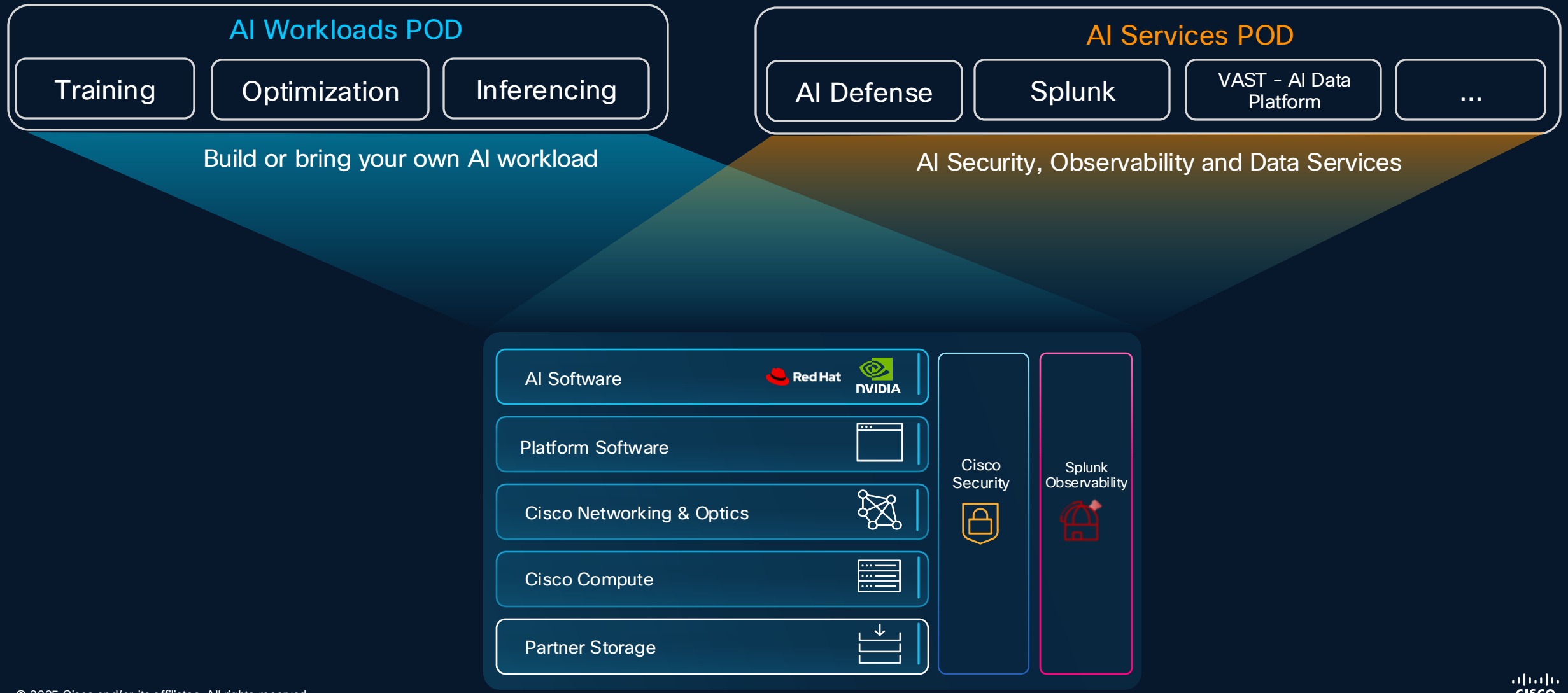
AI POD w/ Cloud management



Turnkey infrastructure:

- Full stack, buy & deploy
- Nexus Hyperfabric: Cloud-managed Networking
- Nexus Hyperfabric AI: Cloud-managed physical infrastructure

Deployable Workloads



Splunk Observability Cloud

For AI PODs

OpenTelemetry-native

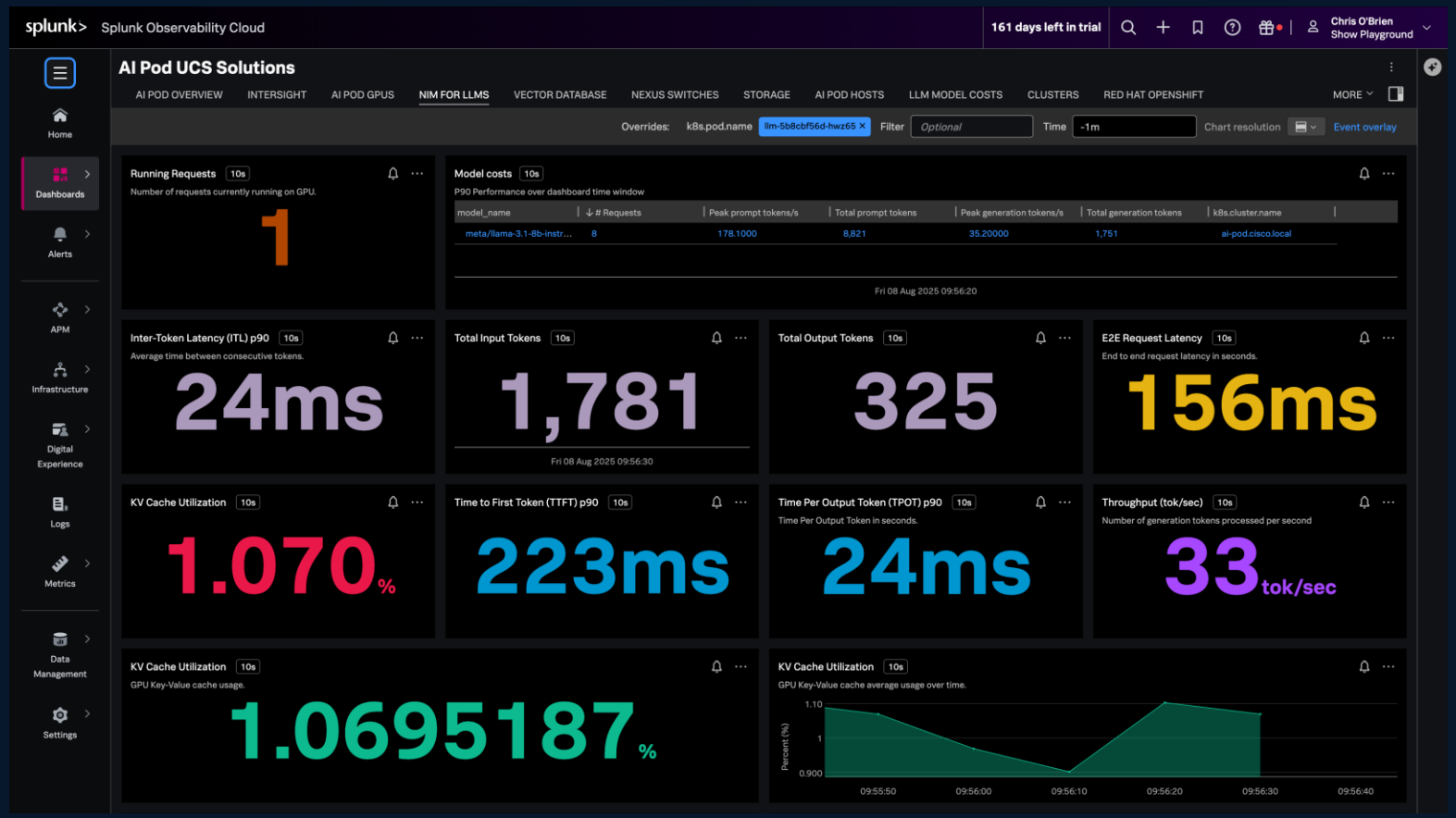
Own and control your data, avoid vendor lock-in and instrument only once on a common standard as you build new applications.

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.



Addressing AI Security

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

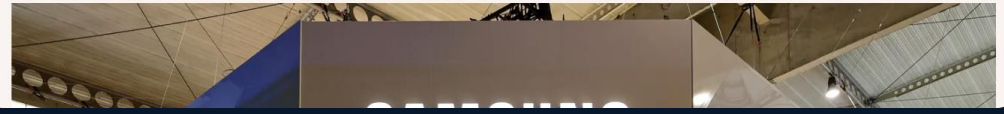
€35 million or 6–7% of global annual turnover for violation of EU AI Act

Downtime cost

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

TOPLINE

Samsung Electronics has banned the use of ChatGPT and other AI-powered chatbots by its employees, Bloomberg reported, becoming the latest company to crack down on the workplace use of AI services amid concerns about sensitive internal information being leaked on such platforms.



ars TECHNICA BIZ & IT TECH SCIENCE POLICY CARS GAMING & CULTURE STORE FORUMS

ADVENTURES IN 21ST-CENTURY HACKING —

AI-powered Bing Chat spills its secrets via prompt injection attack [Updated]

By asking "Sydney" to ignore previous instructions, it reveals its original directives.

BENJ EDWARDS - 2/10/2023, 11:11 AM

BBC

Home News Sport Business Innovation Culture Travel Earth Video Live

Airline held liable for its chatbot giving passenger bad advice - what this means for travellers

23 February 2024

By Maria Yagoda, Features correspondent

*see notes for sources

What Does the AI Threat Landscape Look Like?

OWASP

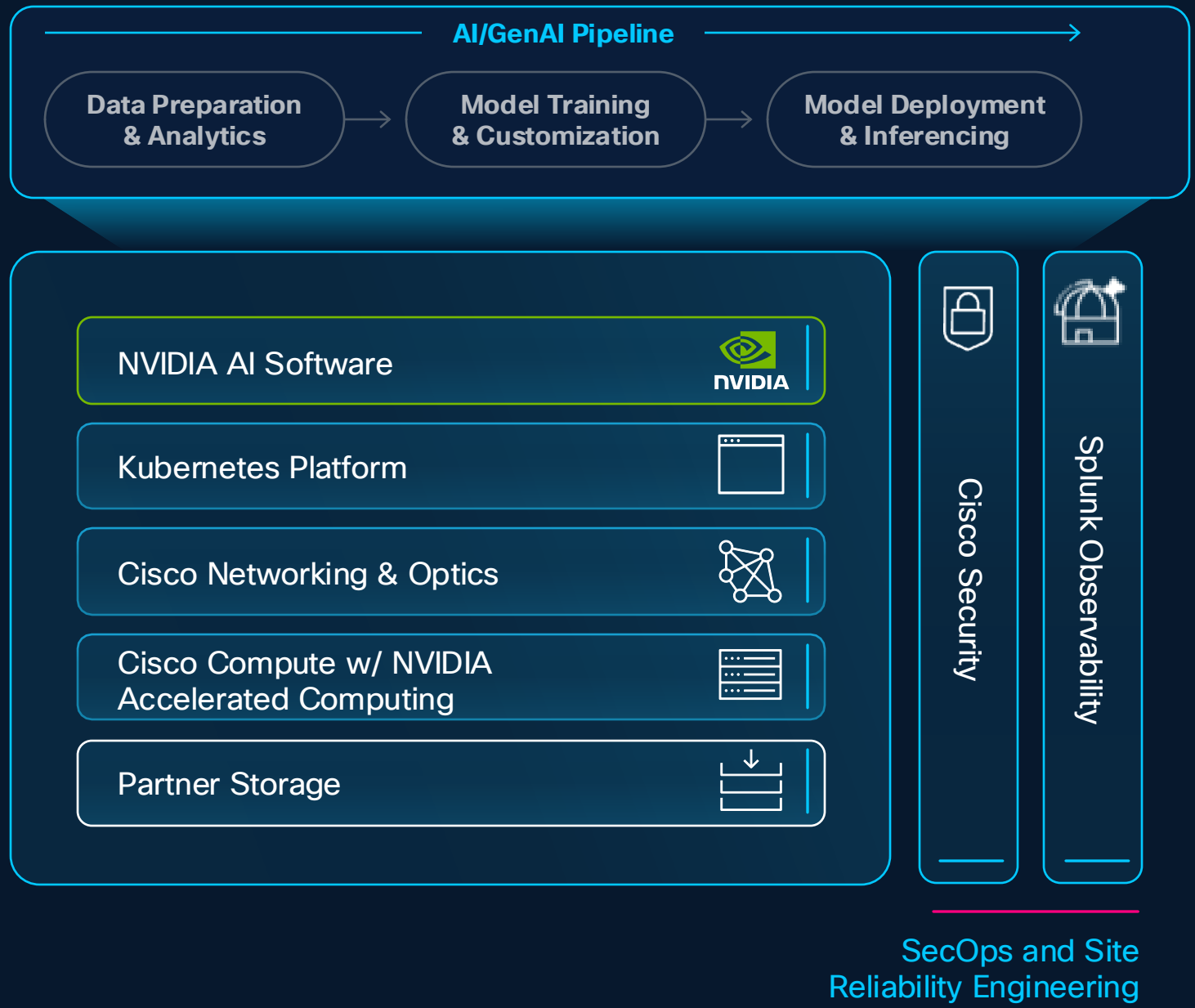
LLM01 Prompt Injection	LLM06 Excessive Agency
LLM02 Sensitive Information Disclosure	LLM07 System Prompt Leakage
LLM03 Supply Chain	LLM08 Vector and Embedding Weaknesses
LLM04 Model Denial of Service	LLM09 Misinformation
LLM05 Improper Output Handling	LLM10 Unbounded Consumption

MITRE ATLAS

Cisco Secure AI Factory with NVIDIA

Delivering Trusted AI Outcomes

A reference design with validated architectures to accelerate AI adoption for enterprises with integrated AI infrastructure and software solutions



Cisco Secure AI Factory with NVIDIA

Delivering Trusted AI Outcomes

Choice of compute, network & storage architectures:

all AI POD

Cisco AI POD

NVIDIA AI Enterprise: NeMo | NIMs | Blueprints

Run:ai: AI Workload & GPU Orchestration

KUBERNETES PLATFORM: Red Hat OpenShift on RHEL CoreOS; Upstream Kubernetes on Ubuntu

CISCO NETWORKING & OPTICS: Cisco Nexus; Cisco Data Center Switching w/ Silicon One®; Managed w/ Cisco Nexus Dashboard or Nexus Hyperfabric

CISCO COMPUTE: Cisco UCS; NVIDIA HGX, MGX or RTX PRO Servers; NVIDIA BlueField®-3 DPUs; Managed with Cisco Intersight®

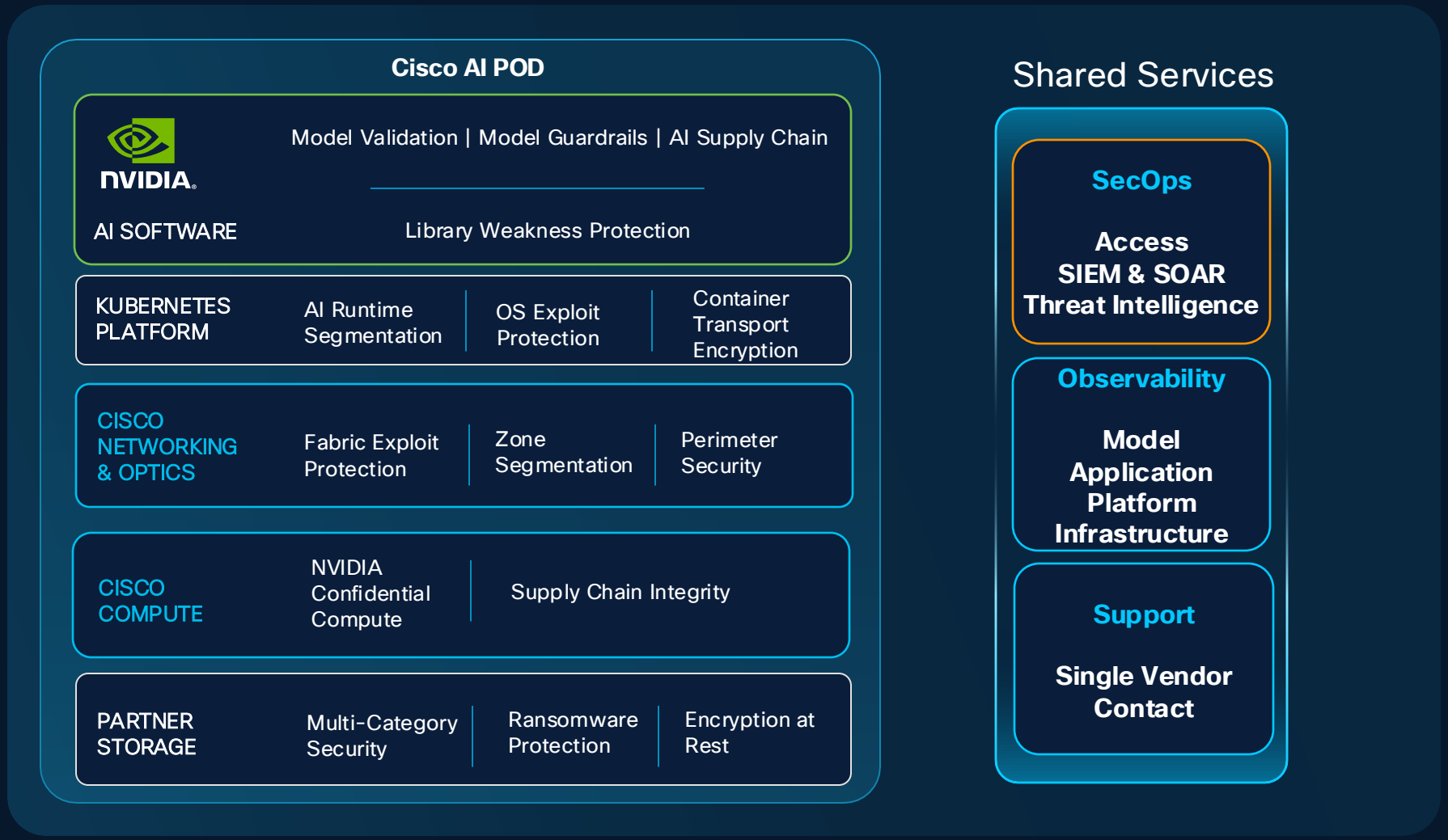
PARTNER STORAGE: NetApp | Pure Storage | VAST Data | Hitachi Vantara

Cisco Security	Splunk Observability Cloud
Cisco AI Defense	AI Application Monitoring
Cisco Hybrid Mesh Firewall Isovalent Secure Firewall Hypershield	Infrastructure Monitoring Log Analytics

Bringing observability to the Factory

Security Capabilities in Cisco Secure AI Factory with NVDIA

Delivering Trusted AI Outcomes



GenAI Security Architecture

Biggest Concerns

Loss of intellectual property

Service disruption

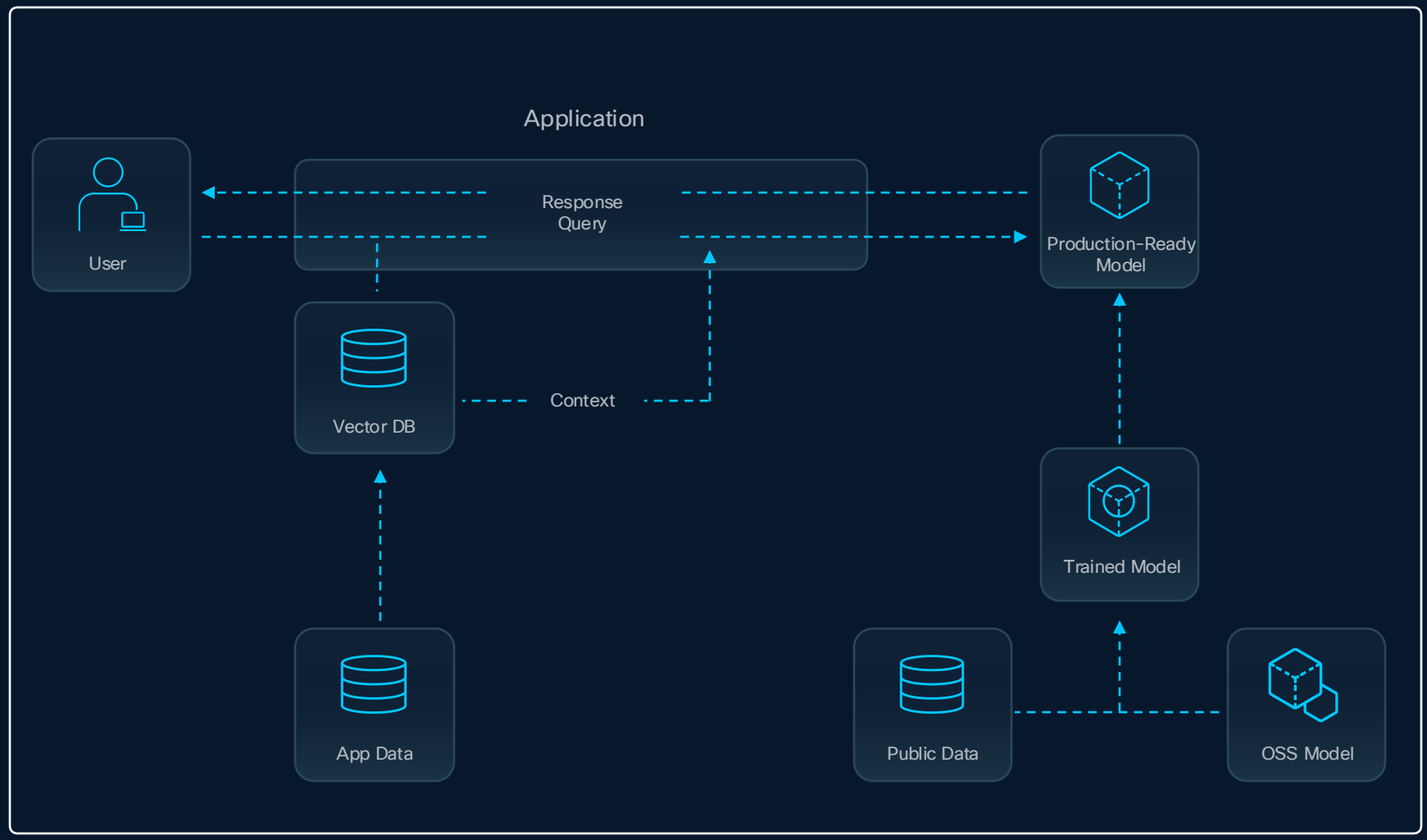
Top Risks

Exfiltration
Denial of Service
Privacy Attacks
Cost Harvesting
Toxicity
Data Poisoning
Model Backdoor

Prompt Injection
Data Extraction
Misalignment
Hallucinations
Indirect Injection
Factual Inconsistency

Capabilities Preventing Threats

- Model Validation
- Model Guardrails
- AI Supply Chain Validation
- Software Weakness & Exploit Protection
- Transport Encryption
- Fabric Exploit Protection
- Zone Segmentation
- Confidential Computing
- Perimeter Security



GenAI Security Architecture

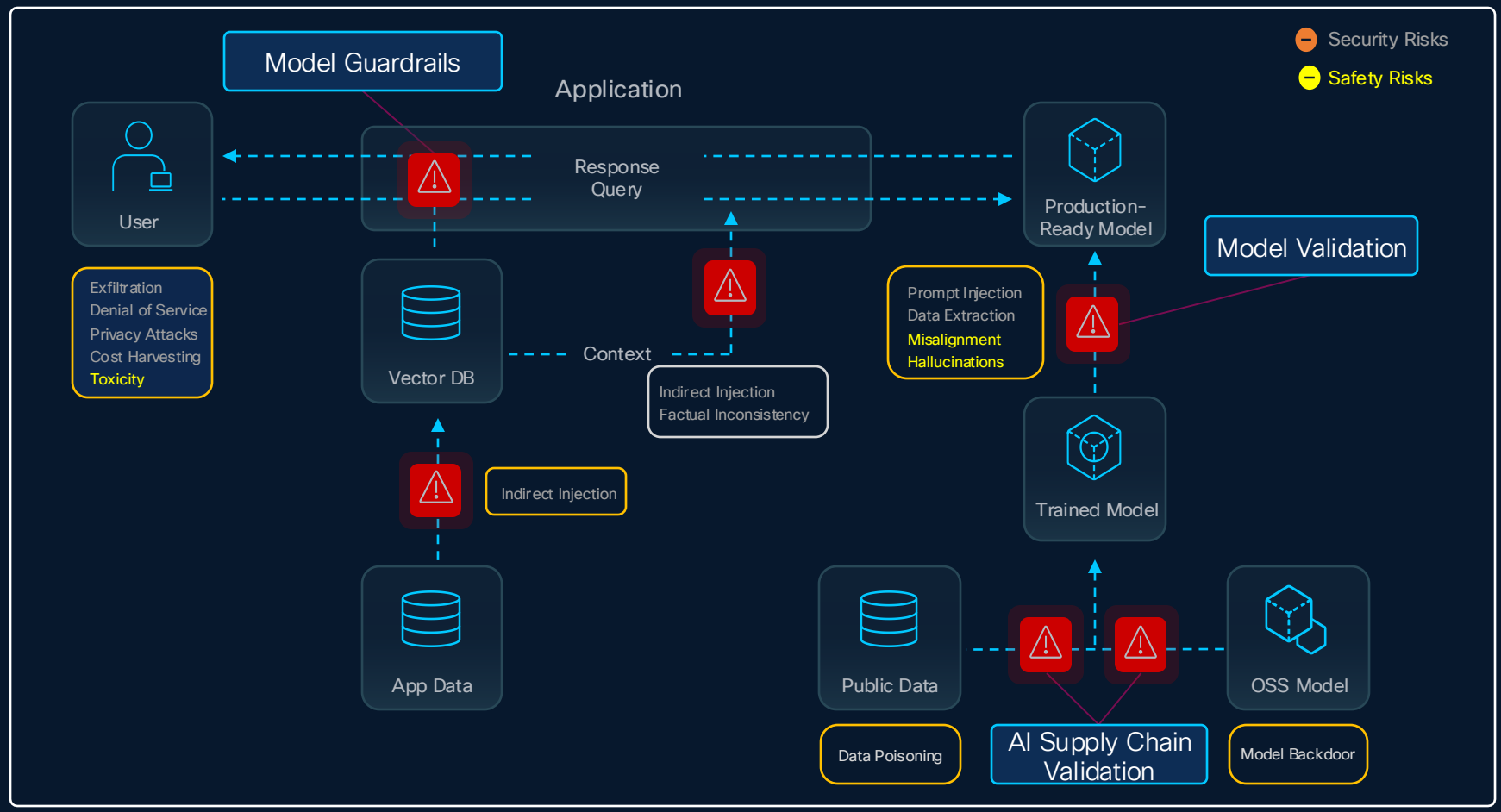
Biggest Concerns

Loss of intellectual property

Service disruption

Capabilities Preventing Threats

- Model Validation
- Model Guardrails
- AI Supply Chain Validation
- Software Weakness & Exploit Protection
- Transport Encryption
- Fabric Exploit Protection
- Zone Segmentation
- Confidential Computing
- Perimeter Security



GenAI Security Architecture

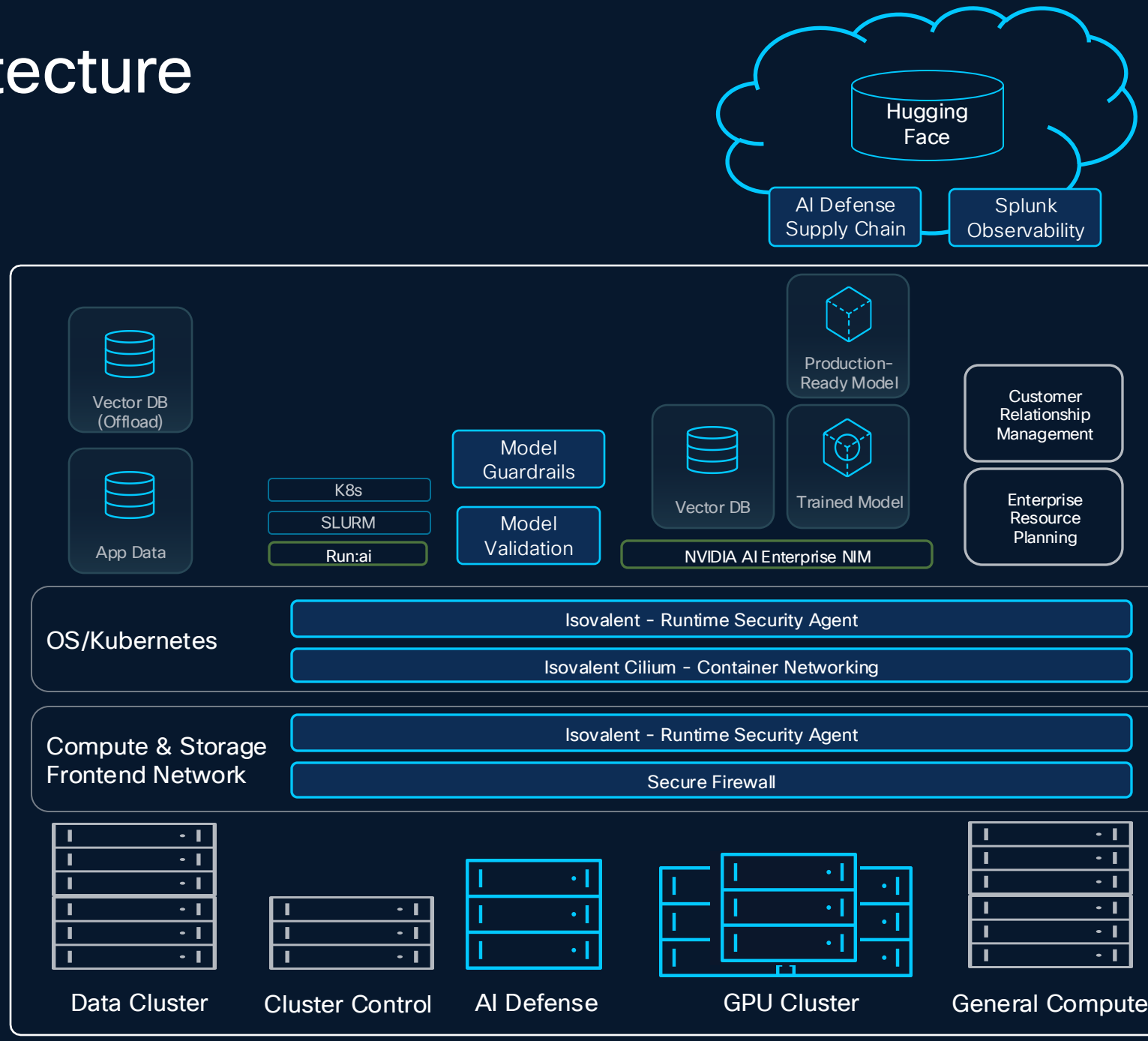
Biggest Concerns

Loss of intellectual property

Service disruption

Capabilities Preventing Threats

- Model Validation
- Model Guardrails
- AI Supply Chain Validation
- Software Weakness & Exploit Protection
- Transport Encryption
- Fabric Exploit Protection
- Zone Segmentation
- Confidential Computing
- Perimeter Security



Cisco UCS

“Show me the metal”

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

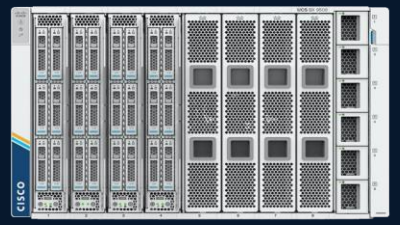
Full-stack AI with compute and networking

Cisco UCS Compute Portfolio

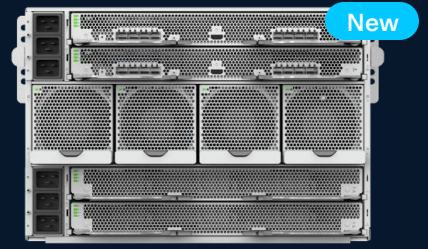
MAINSTREAM ENTERPRISE SERVERS

UCS X-Series
X9508 Chassis

IFM Module



UCS X-Series Direct



UCS x580p M8

UCS X210c M7

UCS X210c M8

UCS X215c M8

UCS X410c M7

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 C240 M6S
14 SSD/HDD Media drive

UCS C240 M6N
14 NVMe Media Drive

UCS C220 M8E3S
16 EDSFF E3.S1T

UCS C220 M8S
10 HDD/SSD/NVMe

UCS C220 M7N
10 NVMe

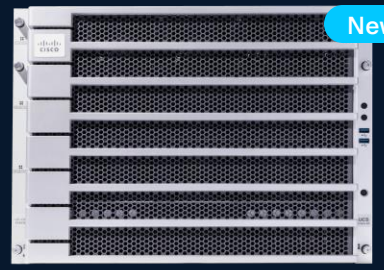
UCS C245 M8SX
28 HDD/SDD

UCS C225 M8S
10 HDD/SSD

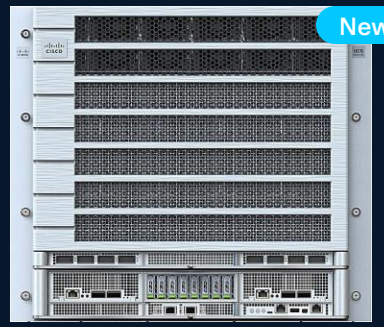
UCS C225 M8N
10 NVMe

AI SERVERS

UCS C885A M8
8RU Dense GPU Server



UCS C880A M8
10RU Dense GPU Server



UCS C845A M8
4RU MGX Server

High-Density Blackwell GPU Server for in Cisco UCS Family

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 in Cisco UCS Family

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

cisco

Flexible, Modular AI Servers

"Start small and scale up" with AI

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

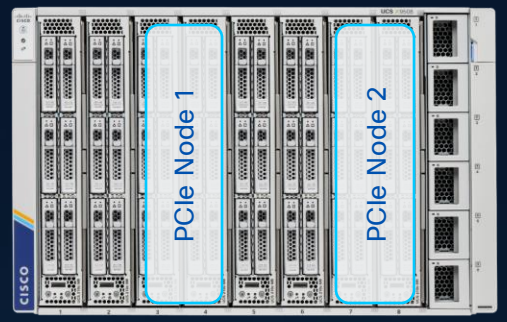
Dual Wide PCIe Node and Switched X-Fabric PCIe Gen5

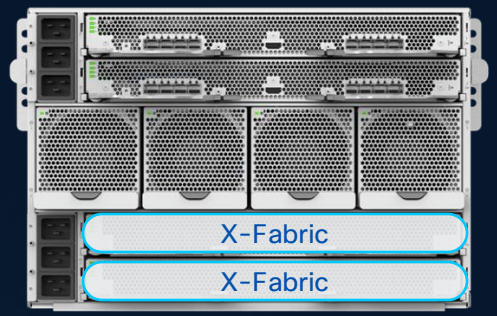
Solution for customer who needs higher GPU density

Supports wide range of workloads

Intersight managed solution

Competitive differentiation with X-Fabric and X-Series

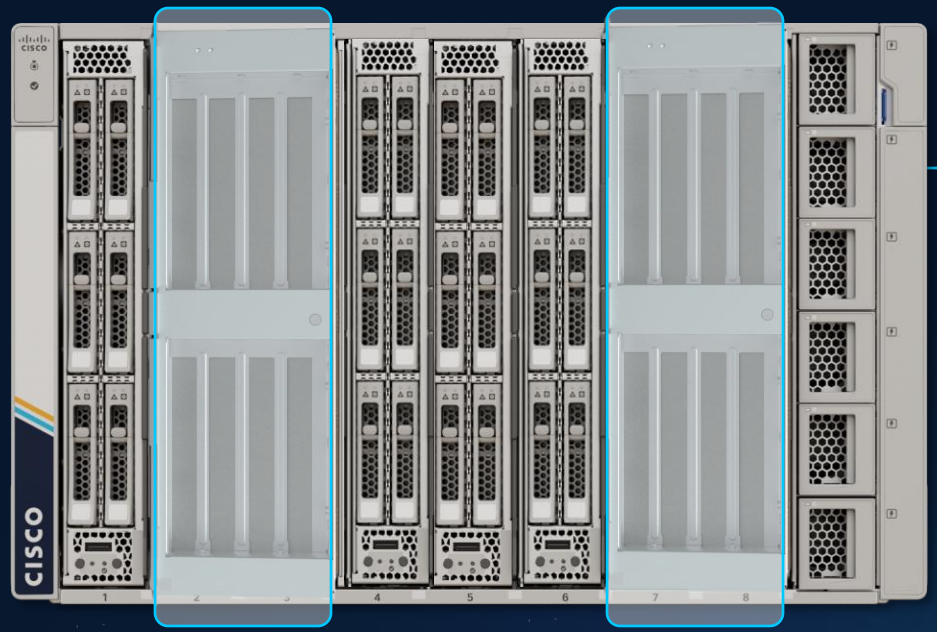




UCS X-Fabric Technology with PCIe Node

- 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

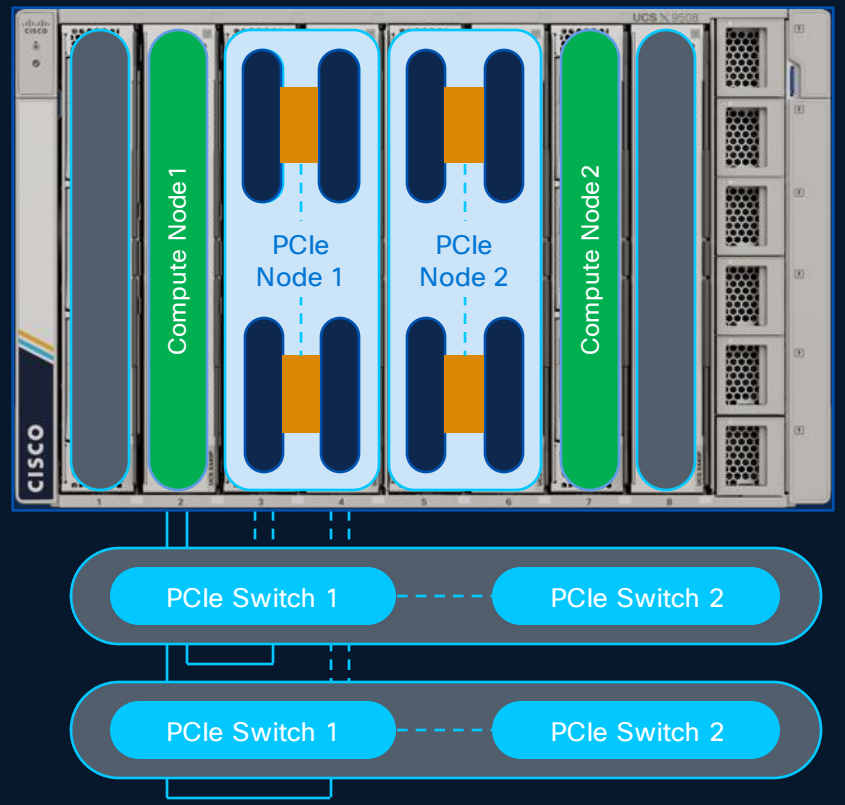
X580p PCIe Node

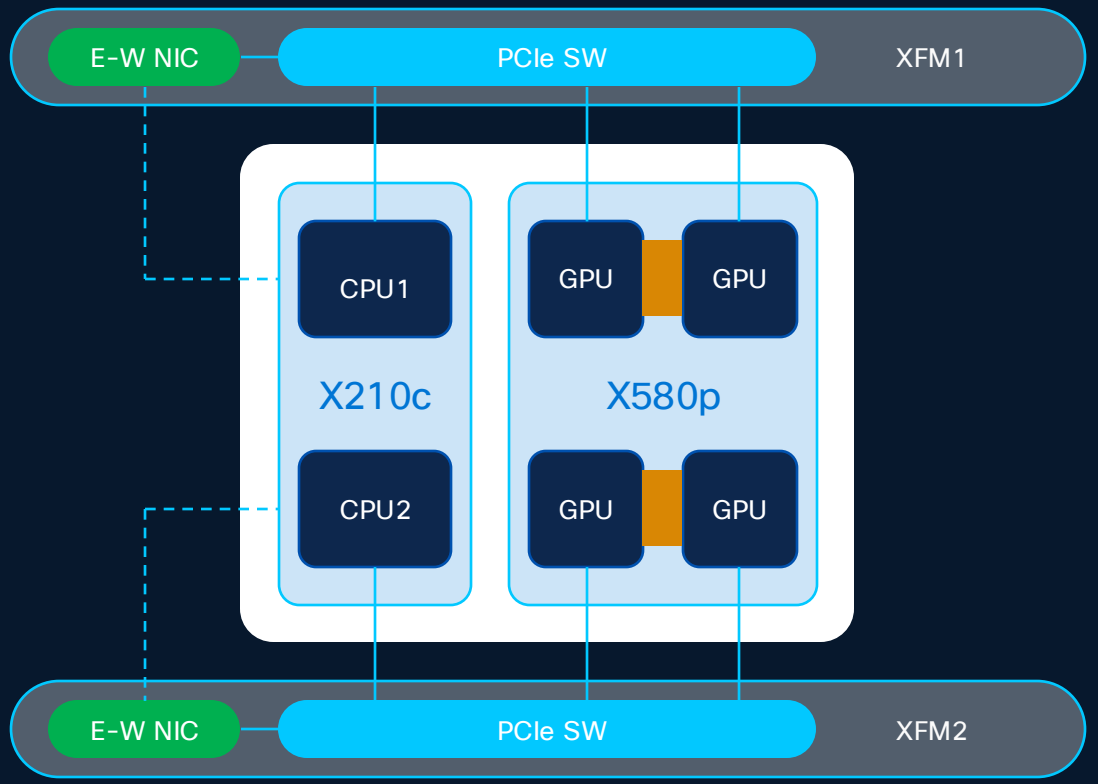


- Double wide PCIe node for 4x FHFL GPU and PCIe G5 GPU support
  - Nvidia H200-NVL, RTX PRO 6000 & L40S
- Support multiple vendors: Nvidia, AMD*/Intel*
- NVLink bridge support
- Support up to 600W FHFL GPU
- Managed PCIe node with BMC support
- Policy based GPU management
- Ability to share GPUs across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

* AMD & Intel GPUs support will be post FCS & TBD

Intersight based allocation: 4x GPU

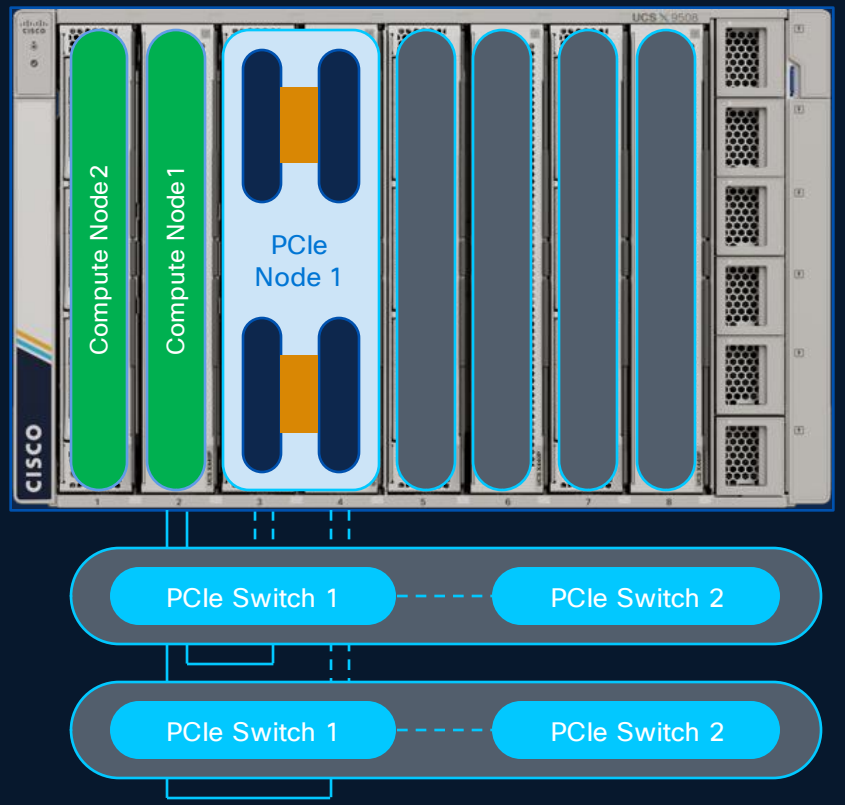




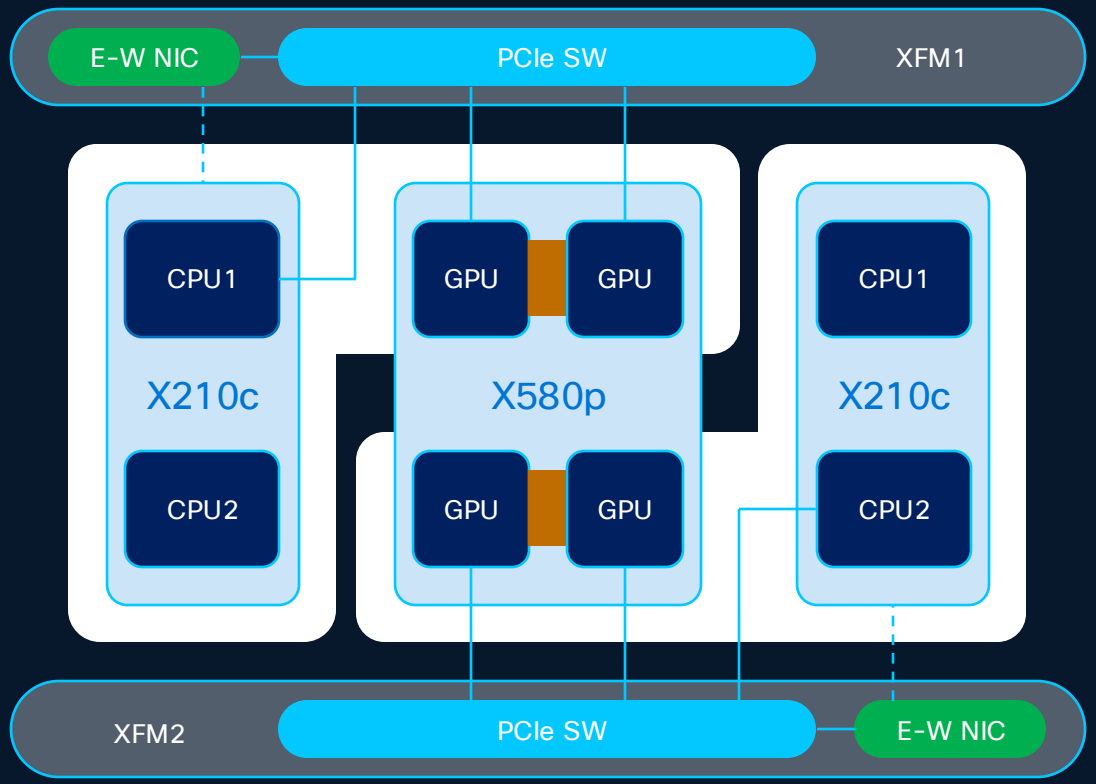
- Support of up to 4 GPUs per compute node
- PCIe Nodes supported in Slots 1-4 and 5-8
- Compute Nodes connectivity to PCIe node on same side of the chassis, ex:
  - Compute Node in Slot 1/2 ←→ PCIe Node in Slot 3-4
  - Compute Node in Slot 5/6 ←→ PCIe Node in Slot 7-8

- Intersight assignment of PCIe Node/GPUs to a Compute node
- Intersight assign NIC within the profile
- Allocation of necessary PCIe lanes on the switch

Intersight Based Allocation: 2x GPU

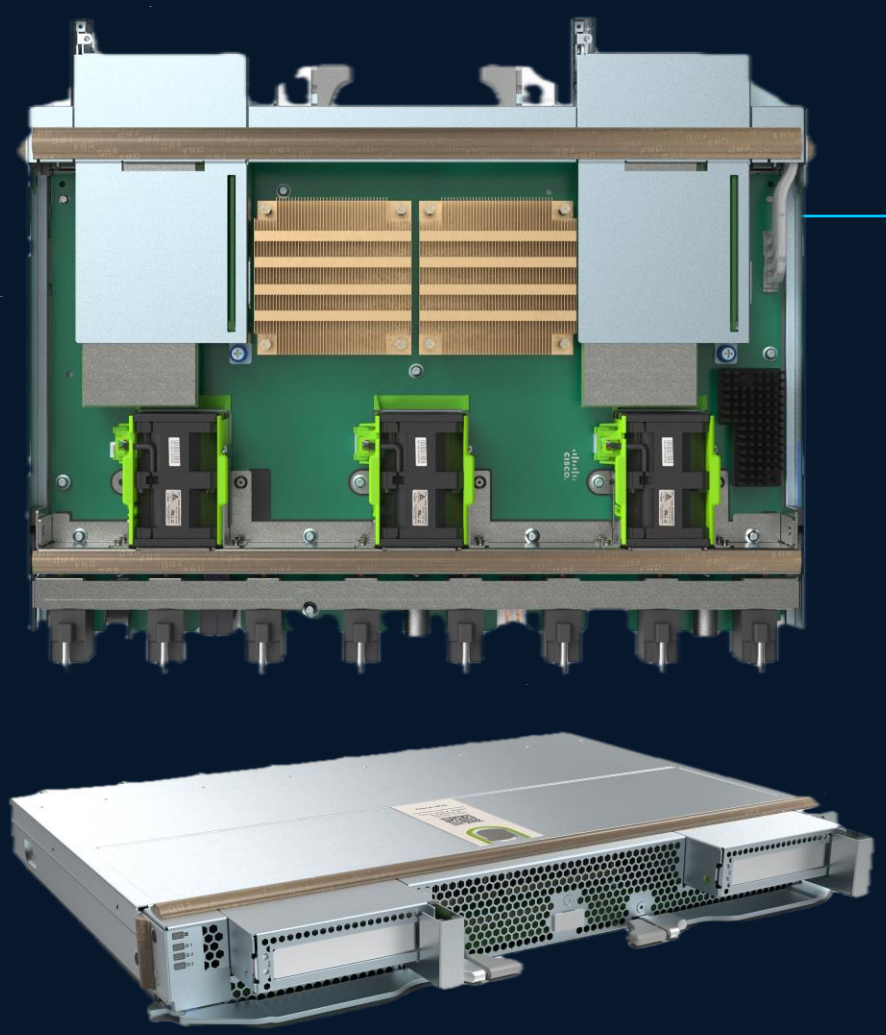


- Support of up to 2 GPUs per compute node
- PCIe Nodes supported in Slots 1-4 (Zone1) and 5-8 (Zone2)
- Compute Nodes connectivity to PCIe node on same side of the chassis, ex:
  - Compute Node in Slot 1/2 ←→ PCIe Node in Slot 3-4
  - Compute Node in Slot 5/6 ←→ PCIe Node in Slot 7-8



- Intersight assignment of PCIe Node/GPUs to a Compute node
- Intersight assign NIC within the profile
- Allocation of necessary PCIe lanes on the switch

UCS 9516 X-Fabric



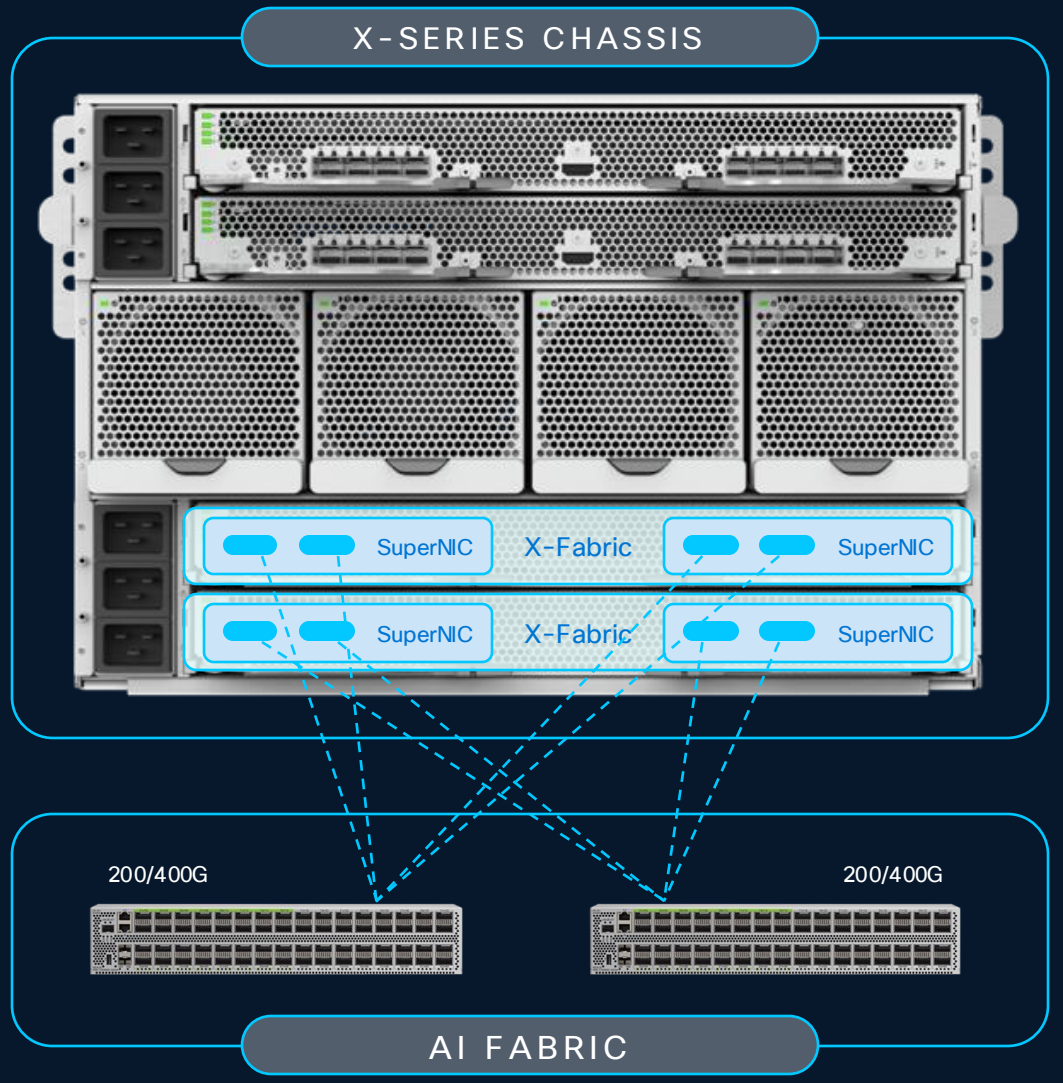
- PCIe Gen5 Switching
- 2x CEM Slots to support HHHL NIC cards
  - ConnectX 7 (2x 200GB & 1x 400G)
- Managed XFM Modules with BMC support
- GPU Direct Support over RDMA
- GPU Backend(East-West Traffic) network support

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

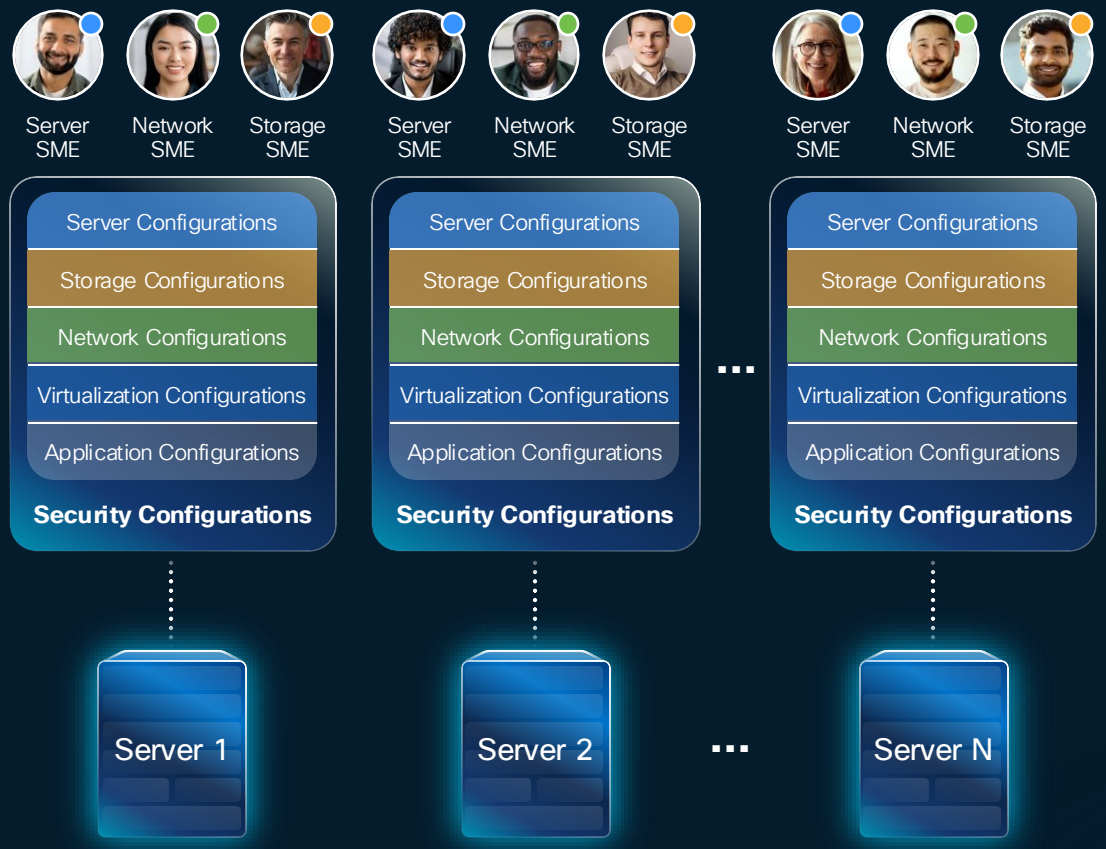


Unified Operations

Cisco Software-Defined Computing

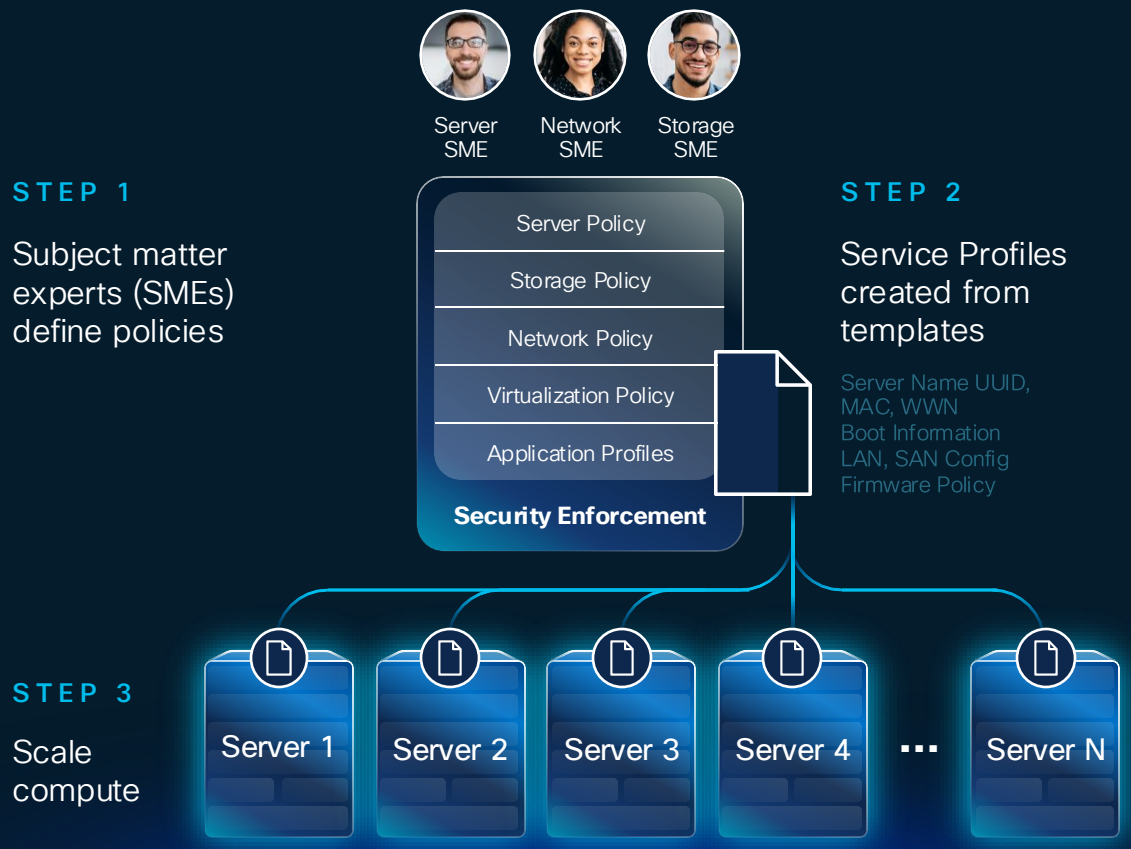
Conventional Approach

Massive complexity and time to scale compute



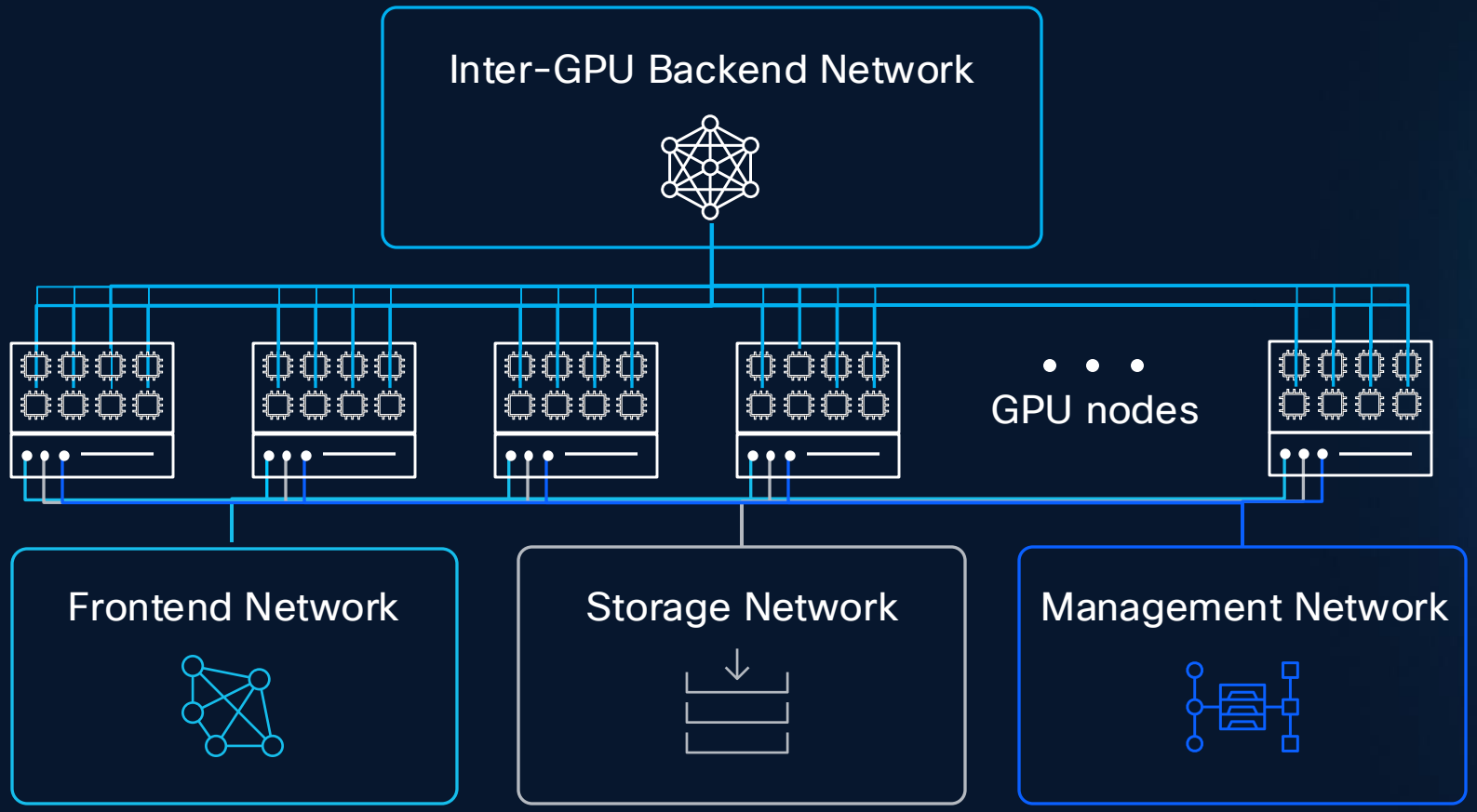
Cisco Computing System

State-less to State-full compute in minutes

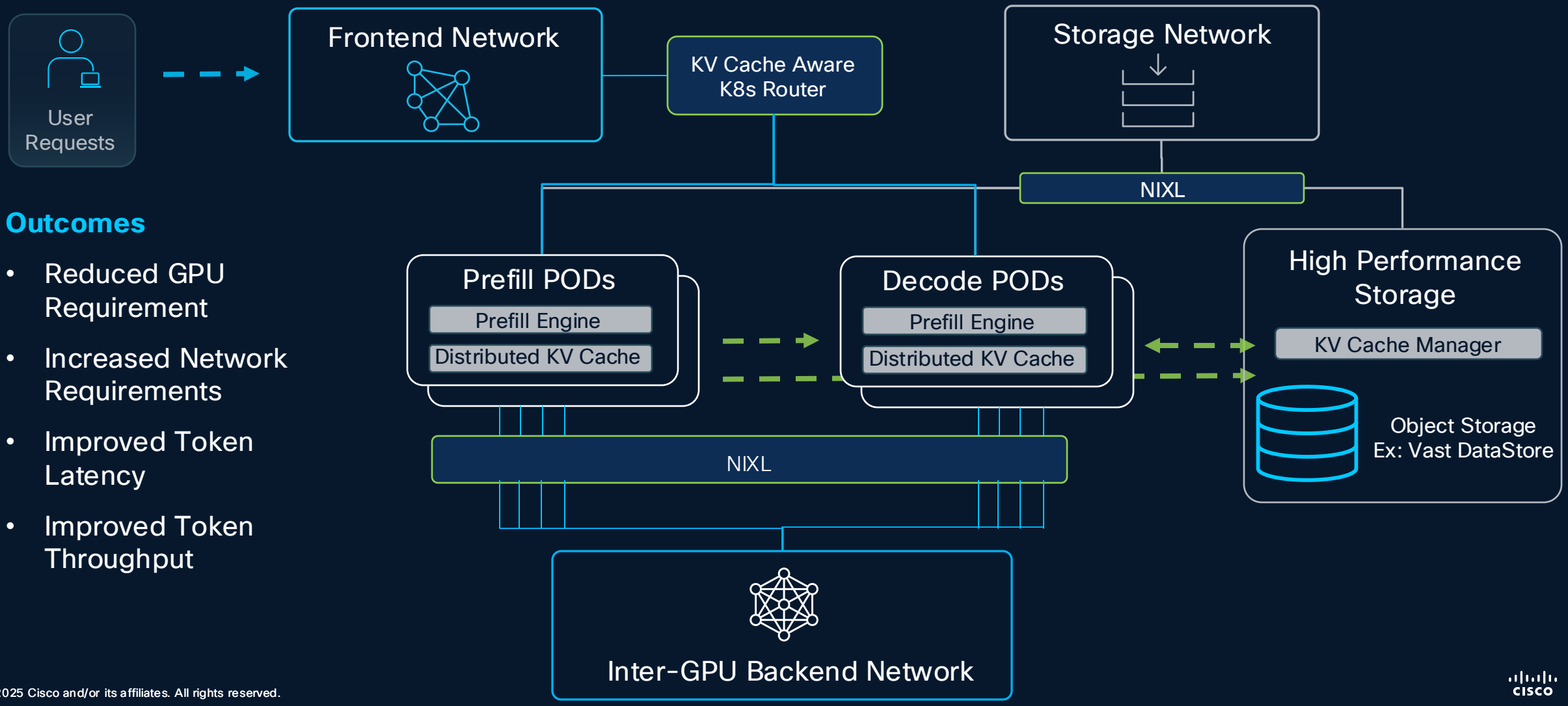


Connecting Computing Platforms for AI

AI Networking



Disaggregated Inference
With NVIDIA Inference Transfer Library (NIXL)



Outcomes

- Reduced GPU Requirement
- Increased Network Requirements
- Improved Token Latency
- Improved Token Throughput

Cisco Nexus High Density 800G & 400G Fixed Switches

Nexus 9300 64-port 800G Switch

512-wide radix	Fully shared packet buffer	Advanced load balancing	Low Latency

N9364E-SG2-Q or N9364E-SG2-O

Compact 2RU 51.2T Switch

G200 ASIC (5nm) | 100G SerDes | 256MB packet buffer

64 800G ports | Up to 128 line-rate 400G ports (2x400G breakout)

Choice of QSFP-DD800 or OSFP ports

Cisco NXOS spine and AI/ML spine/leaf capable

Ultra Ethernet Consortium MEMBER

Nexus 9332D-GX2B
32p 400G
8p MACsec/CloudSec

Nexus 9364D-GX2A
64p 400G
16p MACsec/CloudSec

ACI Leaf, ACI Spine, and NX-OS
25.6T, 19.2T, and 12.8T 400G switches
120MB smart buffer

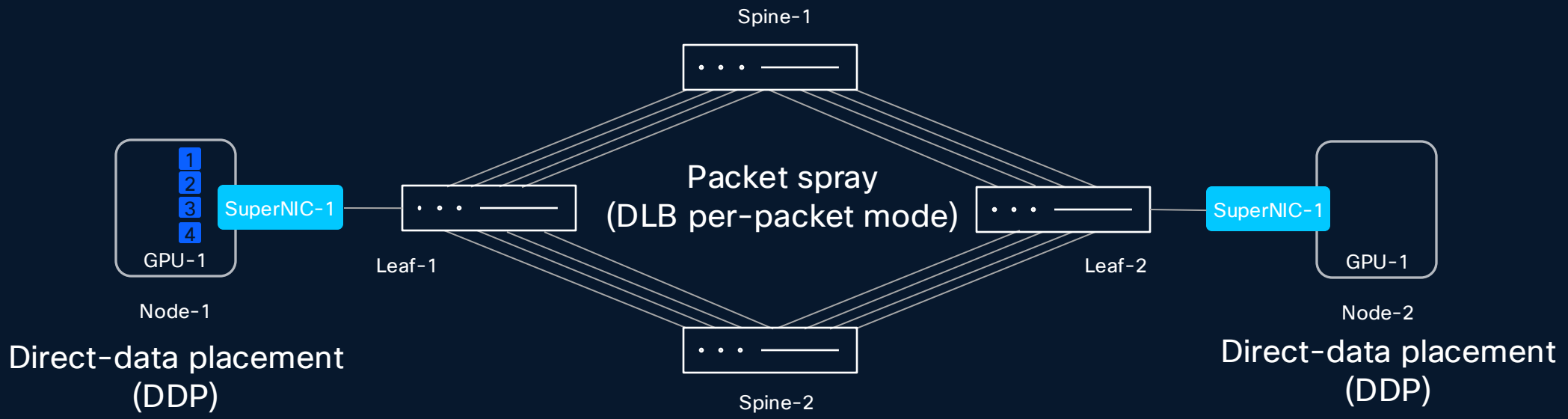
Security
MACsec and CloudSec

Telemetry
FT, FTE, SSX, INT-XD

cisco

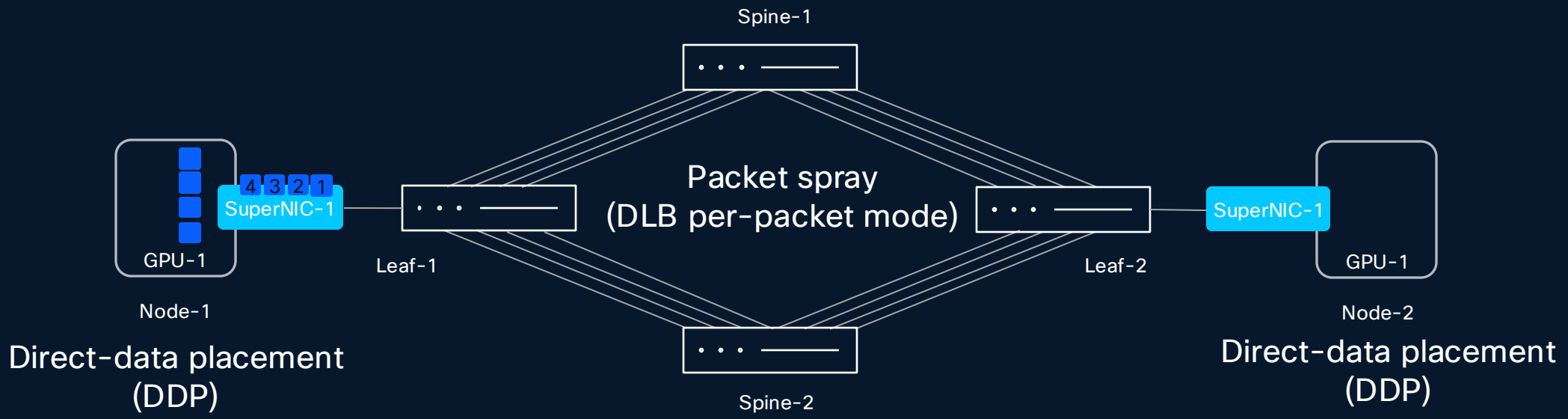
DLB Per-Packet Mode (Packet Spray) and DDP

Bringing Spectrum-X to Cisco Nexus



DLB Per-Packet Mode (Packet Spray) and DDP

Bringing Spectrum-X to Cisco Nexus



Resources to Learn More

Cisco Compute

View on cisco.com/go/ucs

AI-Ready Infrastructure

View on cisco.com

Isovalent Enterprise Platform

Visit Isovalent.com (now part of Cisco)

Cisco Compute YouTube channel

View on youtube.com

Blogs

Visit blogs.cisco.com/datacenter

Online community

Visit the Data Center and Cloud online community

Thank you