Building a full stack AI-ready data center

Architecture and best practice

Chris Gascoigne
Principal Solutions Engineer

Session ID: BRKCOM-2115

Cisco Webex App

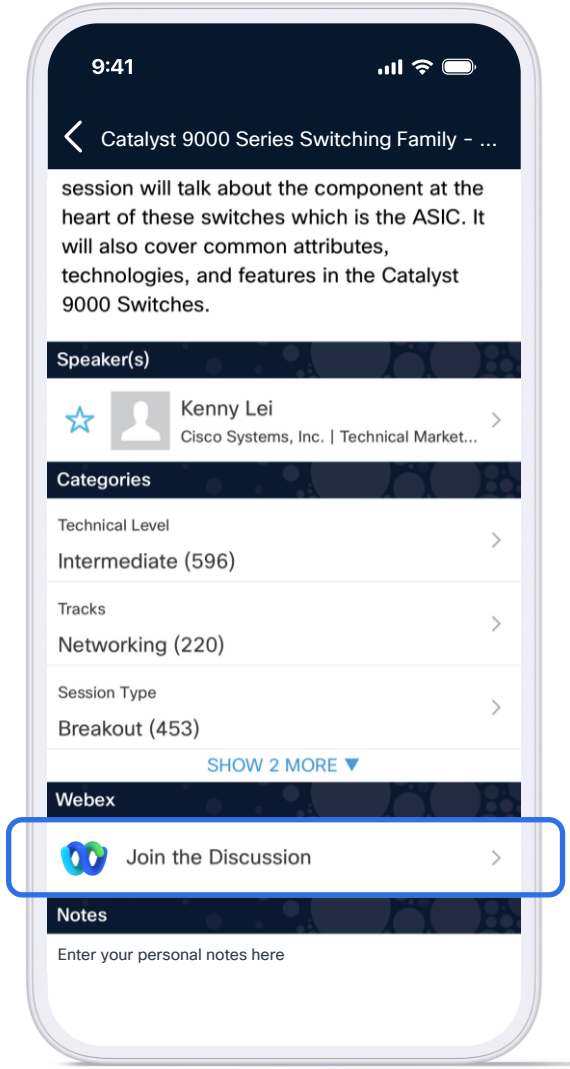
Questions?

Use Cisco Webex App to chat
with the speaker after the session

How

1. Find this session in the Cisco Live Mobile App
2. Click “Join the Discussion”
3. Install the Webex App or go directly to the Webex space
4. Enter messages/questions in the Webex space

Webex spaces will be moderated by the speaker until 14 November 2025.



https://ciscolive.ciscoevents.com/
ciscolivebot/**#BRKCOM-2115**

Agenda

01 Operationalizing AI

02 Model optimisation for compute

03 AI Networking Design

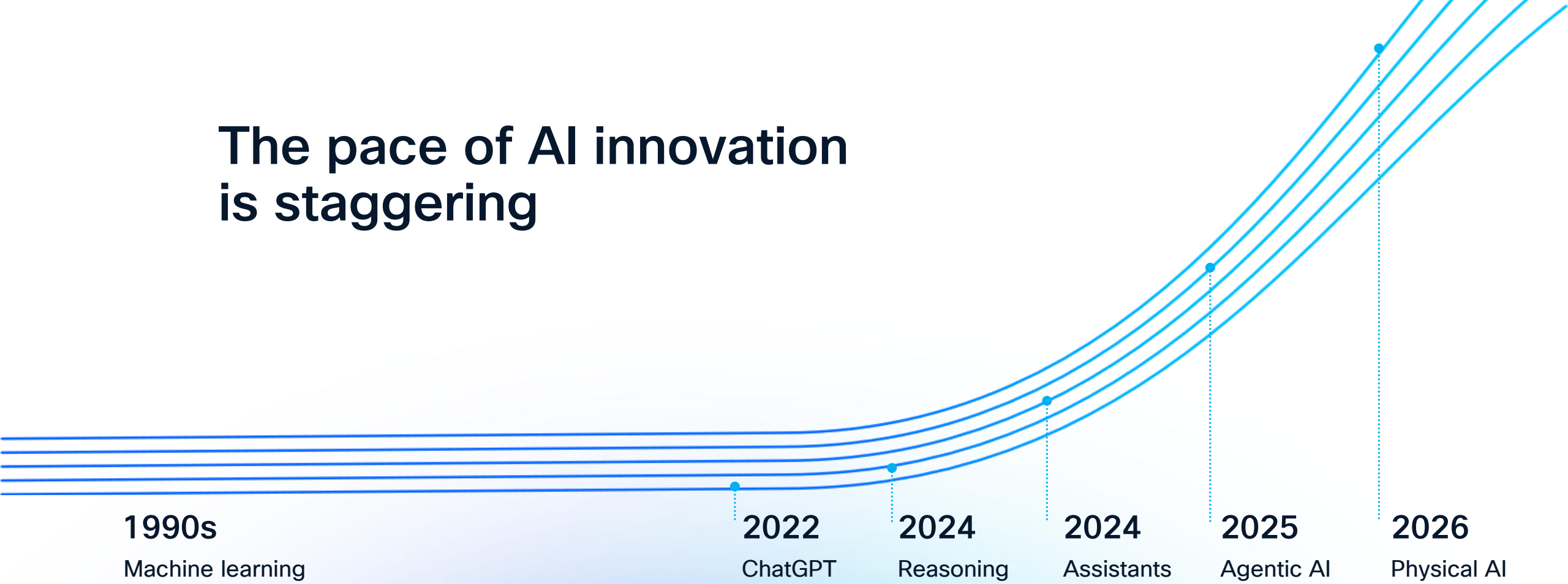
04 Minimizing Attack Surface

05 Secure AI Factory

Operationalizing AI

Determine consumption models and flexible design patterns that facilitate intelligent AI workload placement

The pace of AI innovation is staggering



architectural shifts

Inference-led repatriation

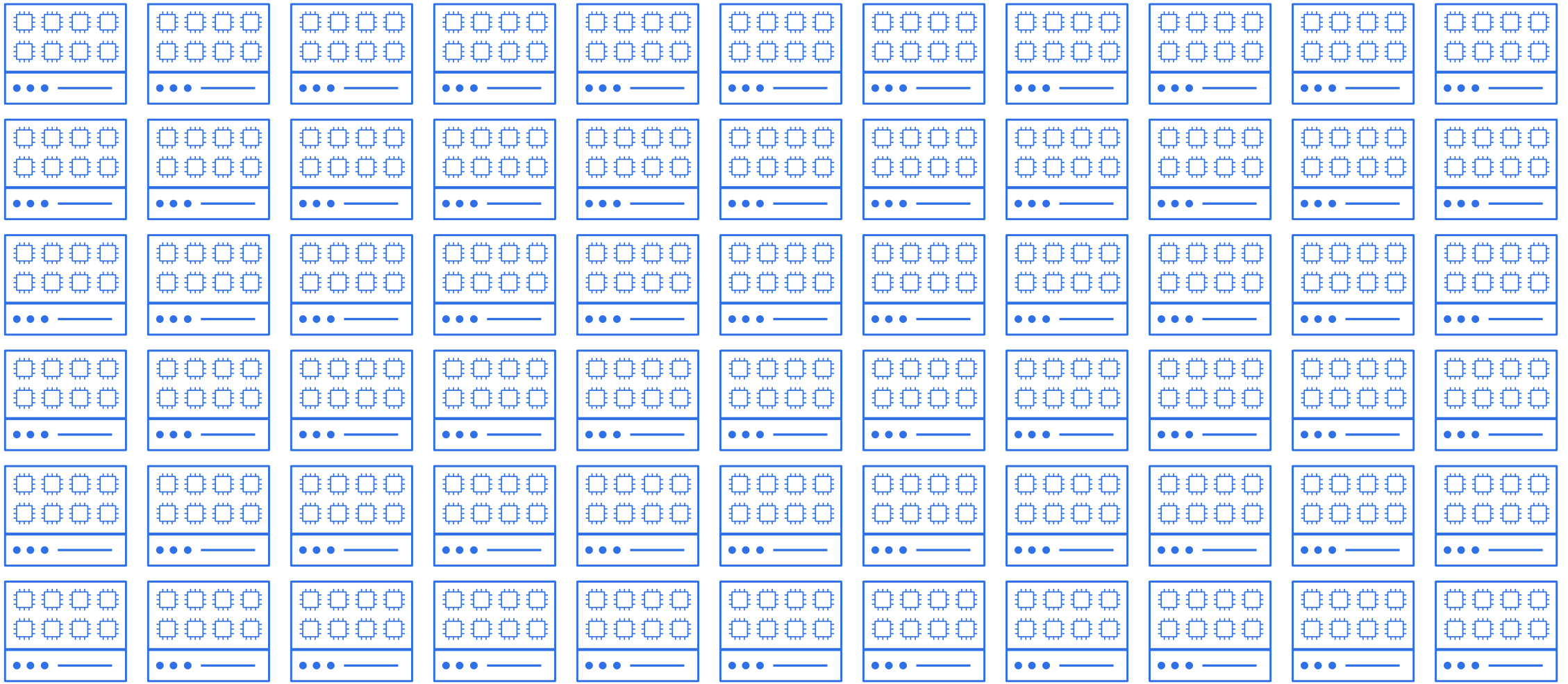
AI is network bound

Common substrate for AI security

Hyper-distributed security

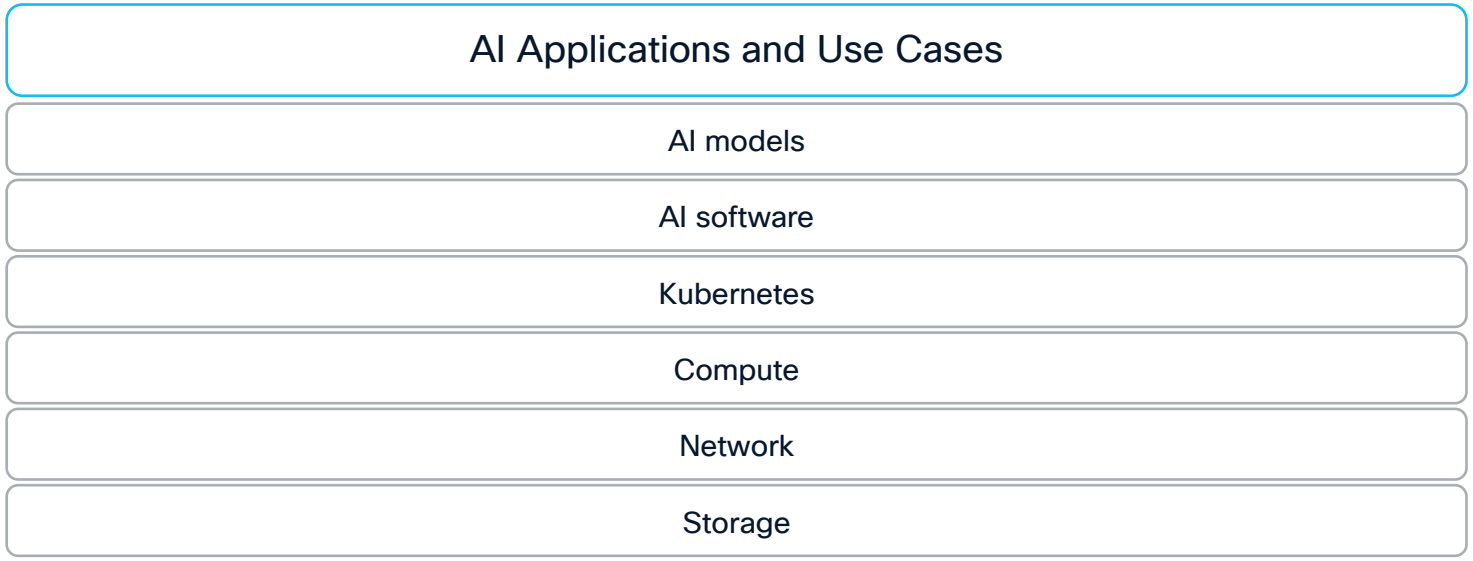
Customer choice in silicon

How many 8 GPU servers do you need?



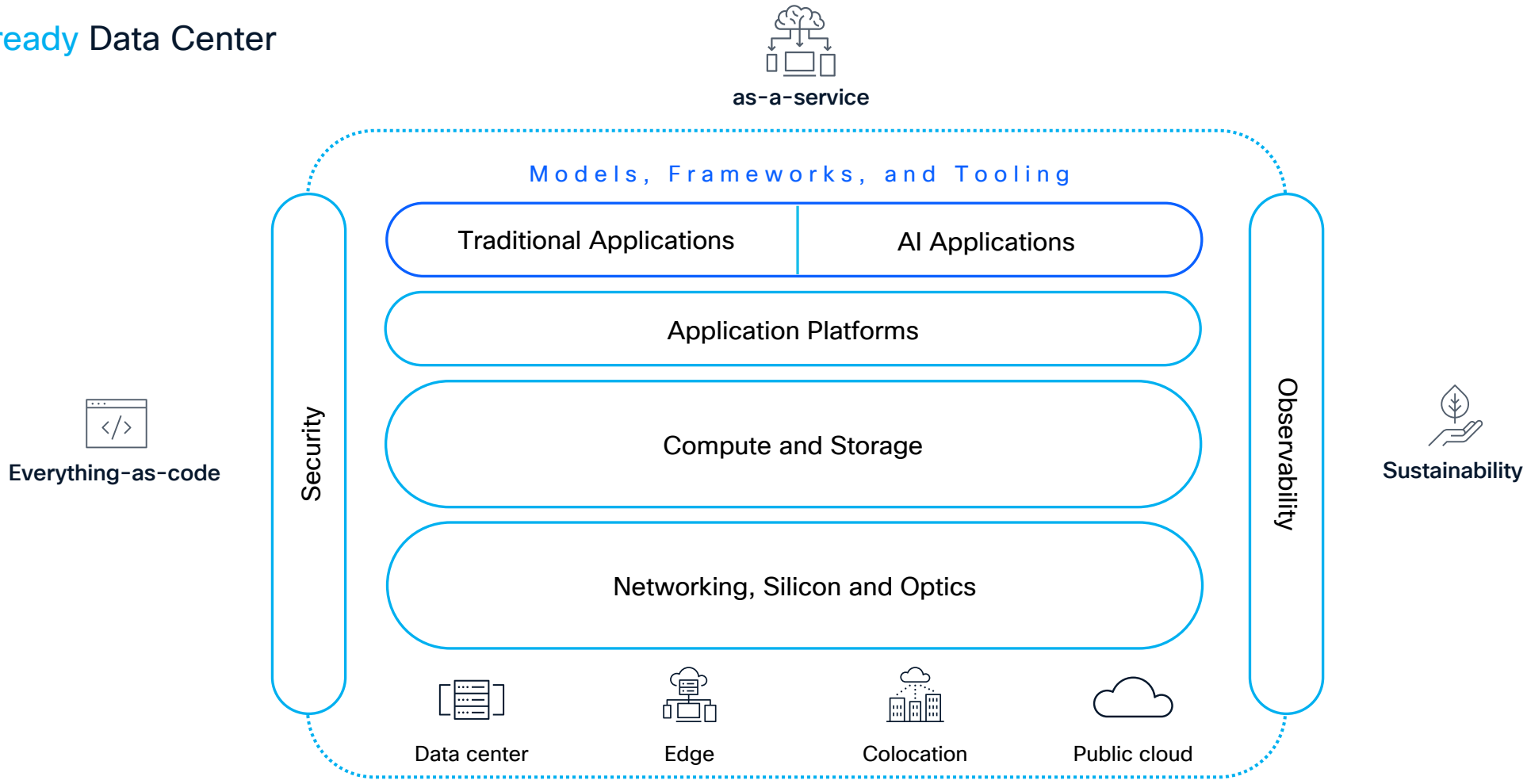
cisco

There’s a lot more to AI Infrastructure than GPUs

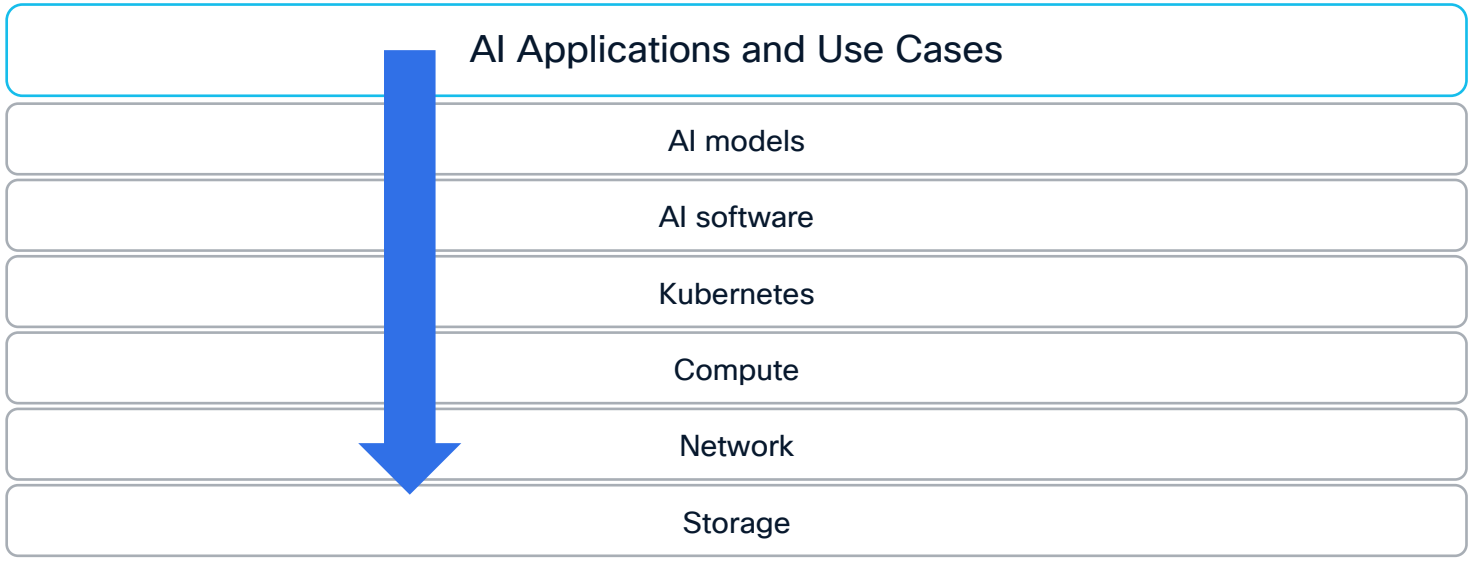


Shouldn’t think of AI as an island

Cisco AI-ready Data Center



Let’s take a top-down approach

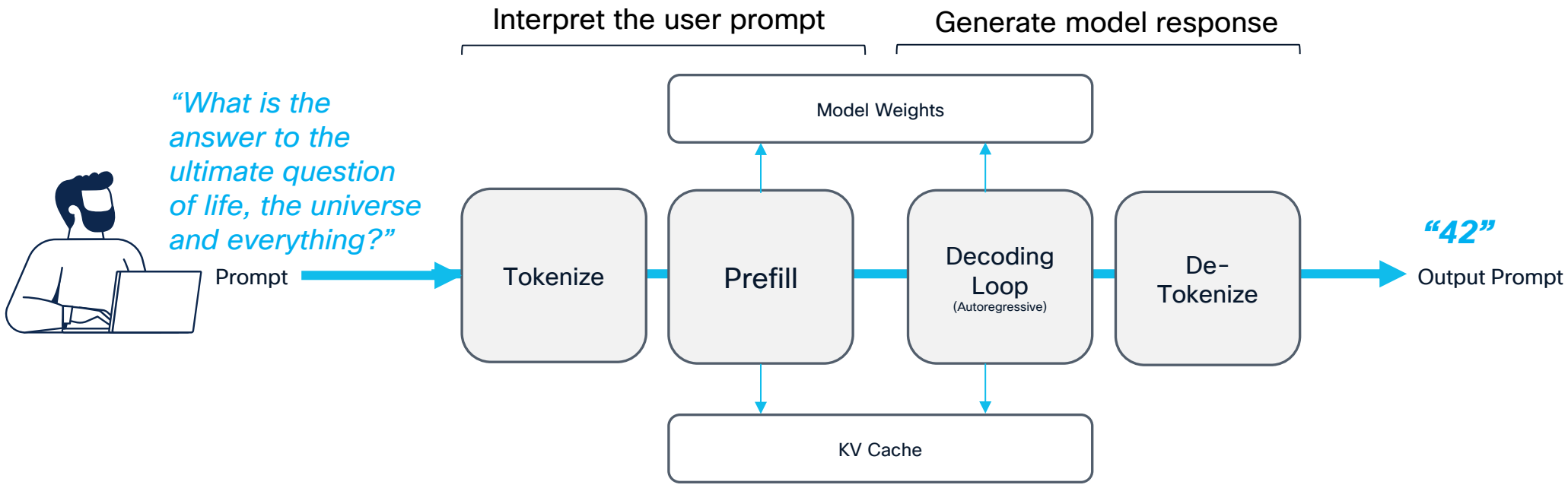


Model Optimization

Right-size models to available GPU hardware balancing speed, efficiency, and carbon footprint.

A day in the life of a prompt

Inference



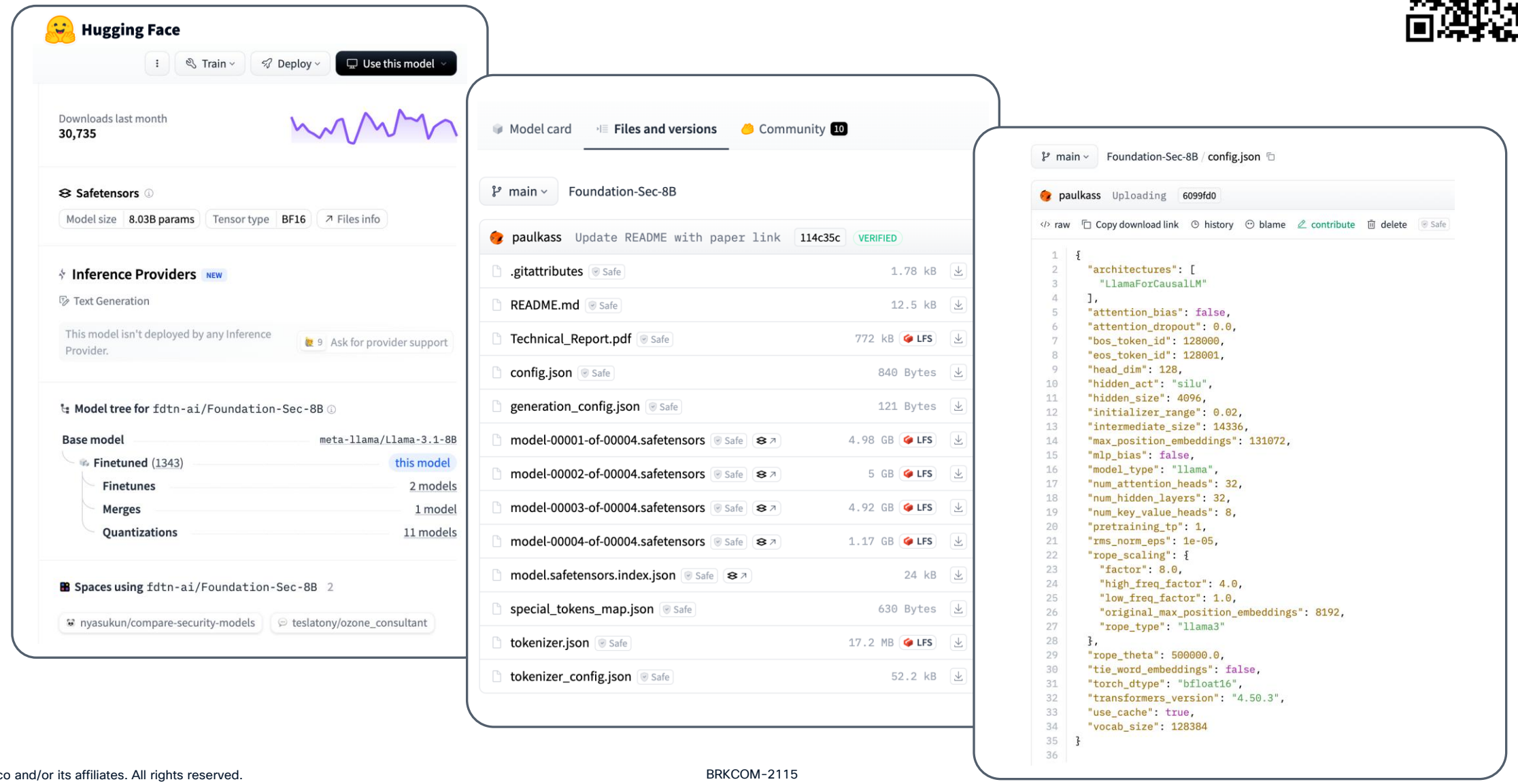
Longer prompts = longer prefill = larger KV Cache = longer time to first token (TTFT)

Each user request create a unique prefill, KV cache and decoding loop

Goal: Optimize for memory, performance and cost

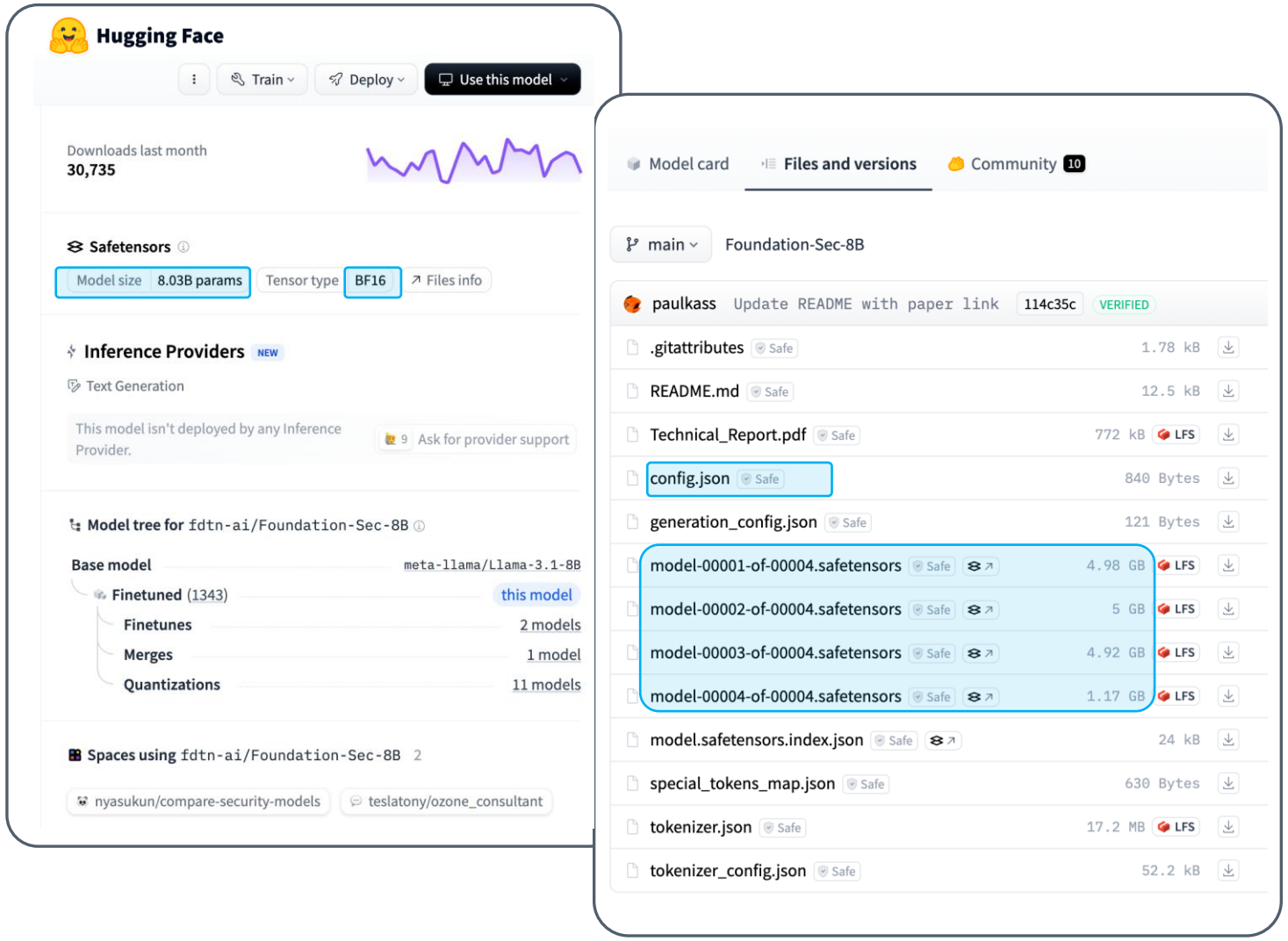
Model, Token and GPU Sizing

Inference



Model, Token and GPU Sizing

Inference



1

RAW Model Size

= # parameters * Model Precision

= 8B x 2

= **16 GB**

Precision	Size per Parameter
FP32 (float32)	4 bytes
FP16 (float16)	2 bytes
BF16	2 bytes
INT8	1 byte
FP8 (Hopper)	1 byte

2

Alternative method
RAW Model Size

= 4.98GB + 5GB + 4.92GB + 1.17GB

= **16.07 GB**

Model, Token and GPU Sizing

Inference

3

Per Token Footprint (stored in the KV Cache)

= (2 * precision * # layers * d_model) / GQA factor

= (2 * 2 * 32 * 4096) / 4

= **131 KB per stored token**

4

Total GPU Memory

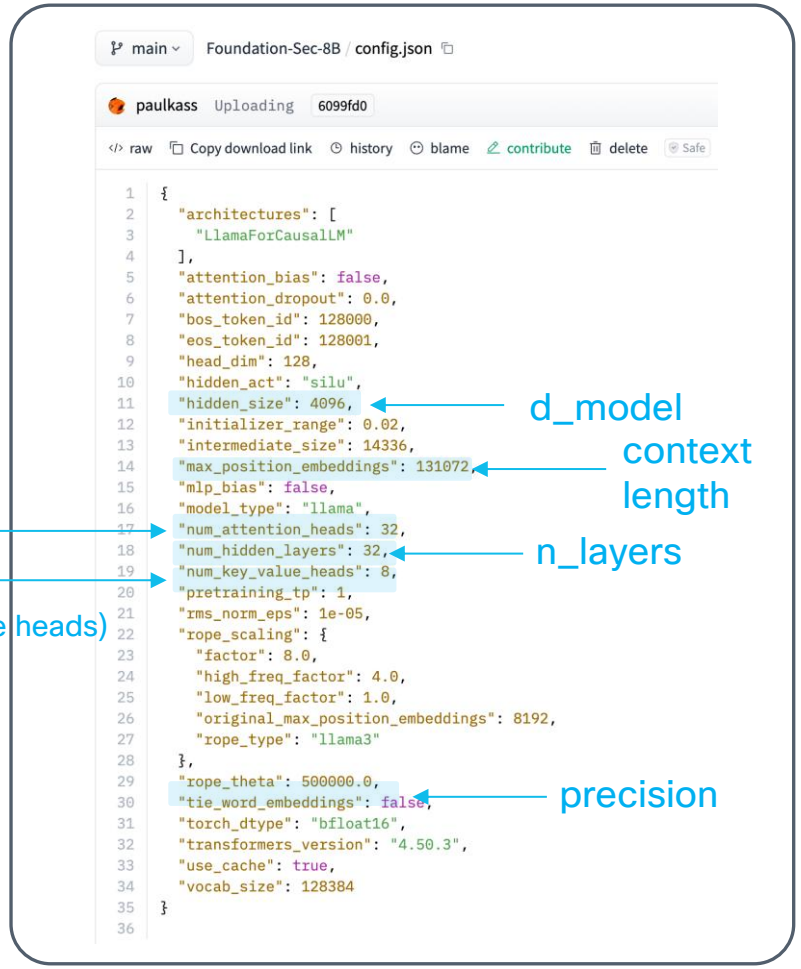
= Raw model size + **KV Cache**

= Raw model Size + (per token size * **context length**)

= 16GB + (131KB * 131072)

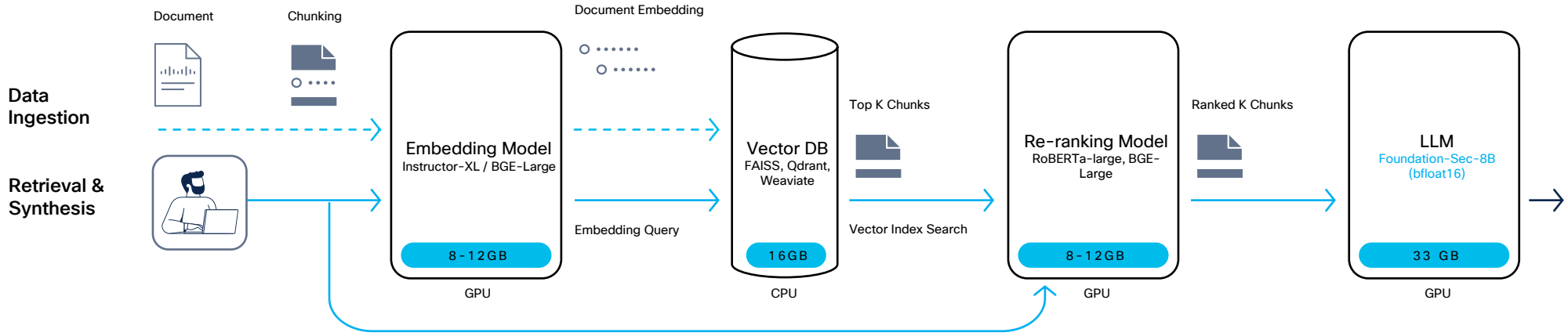
= 16GB + (17GB)

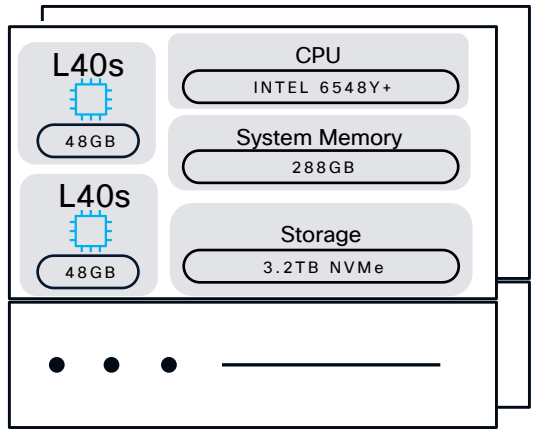
= **33 GB**



Model to GPU Sizing

Retrieval Augmented Generation (RAG) Inference Example





High Availability x 2

Model to GPU Sizing

KV Cache and Concurrency for Llama 3.1 8B

The KV cache grows with user concurrency as each user session requires its own separate cache.

Scaling options

1. Optimize Single GPU

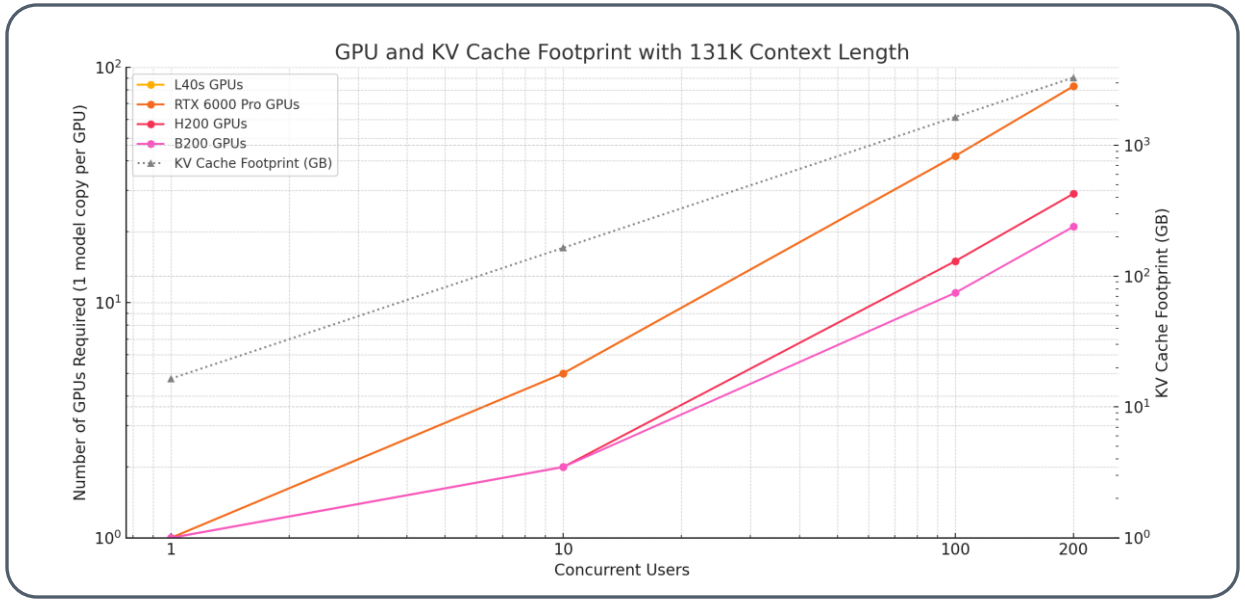
- Quantize KV Cache by dropping precision
- Reduce Context Length

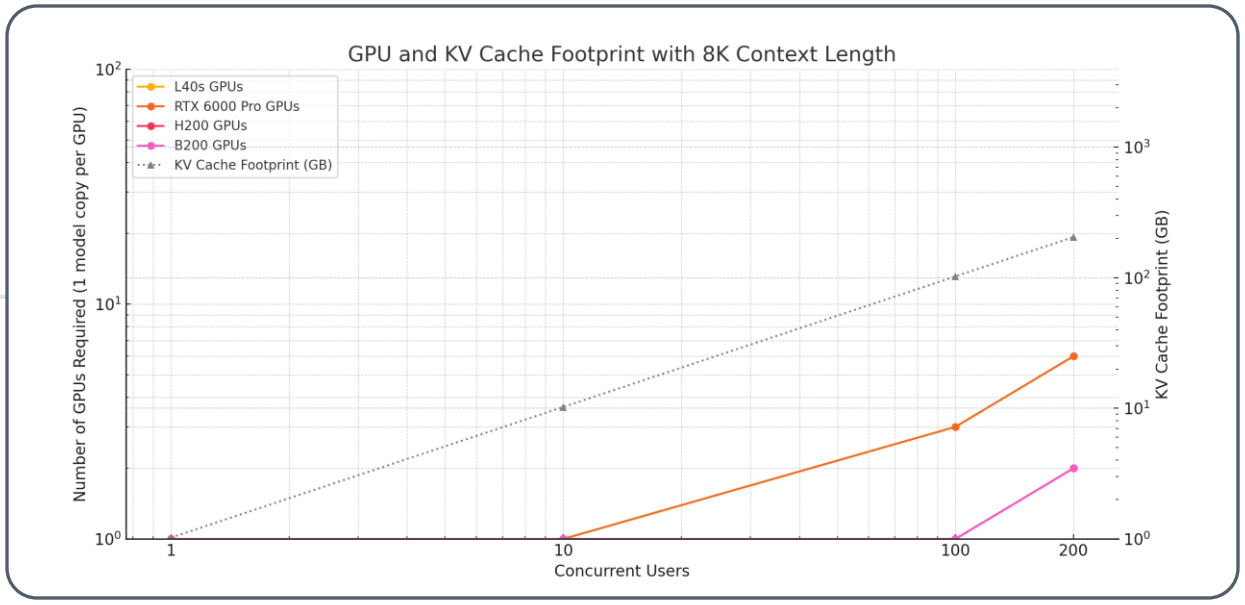
2. Horizontal Scale Out

- Deploy 1 x model per GPU and maximise KV Cache

GPU	Memory (GB)	Bandwidth (GB/S)	FP16 Tensor Core (TFLOP/s)
NVIDIA L40s (PCIe)	48	864	362
RTX Pro 6000 (PCIe)	96	960	1457
NVIDIA H100 (PCIe/SXM)	80	2000/3350	756/989
NVIDIA H200 (SXM)	141	4800	989
NVIDIA B100 (SXM)	192	8000	1800
NVIDIA B200 (SXM)	192	8000	2250

Llama 3.1 8B, 1 model copy per GPU, per token memory 131KB, GQA 4, Layers 32, Precision bloat16





Source: Chatgpt

Model Performance Metrics

Splunk Observability Cloud Dashboard

Real-Time Monitoring and Troubleshooting:

- Monitor infrastructure, applications, and user interfaces in real time at any scale.

Actionable Insights:

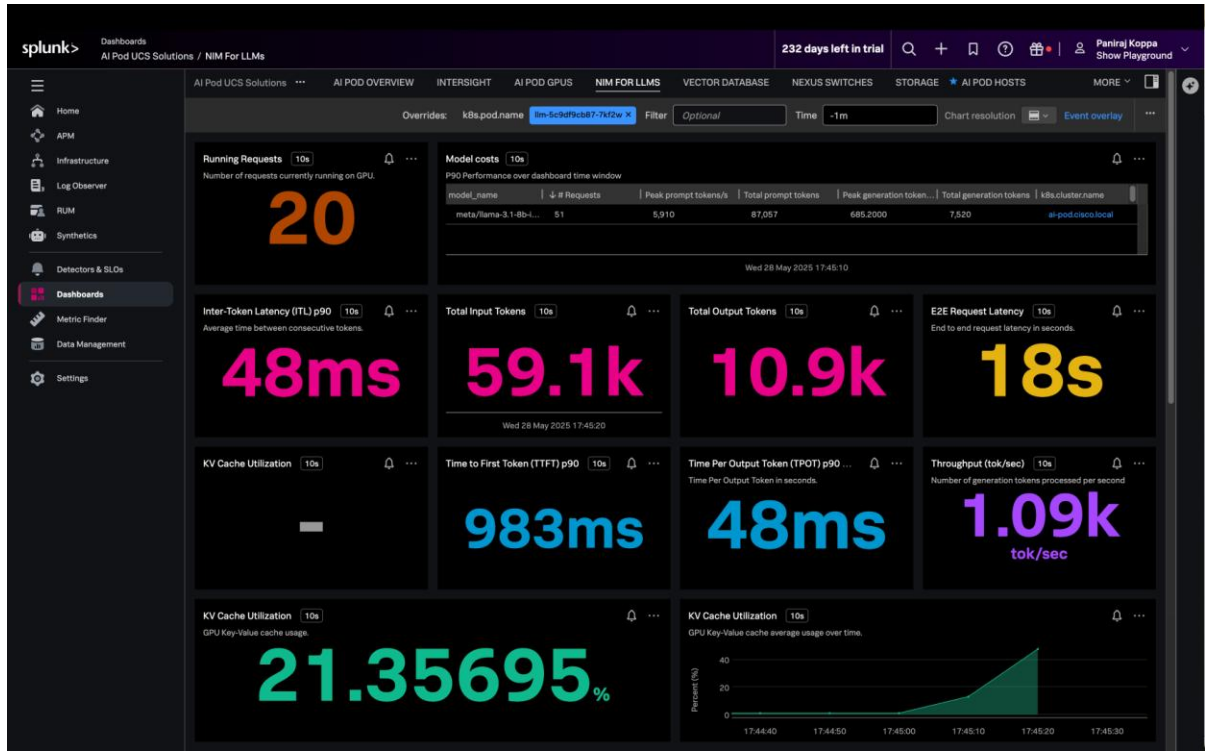
- Splunk Observability Cloud transforms metrics, traces, and logs into dashboards, alerts, and visualizations.

Efficient Data Ingestion:

- Use Splunk's OpenTelemetry Collector to ingest telemetry data seamlessly.

Telemetry Data Processing:

- The Collector simplifies receiving, processing, and exporting telemetry data.

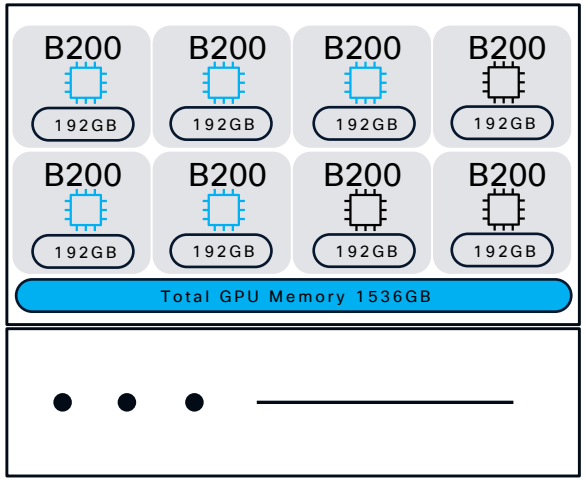


Model to GPU Sizing

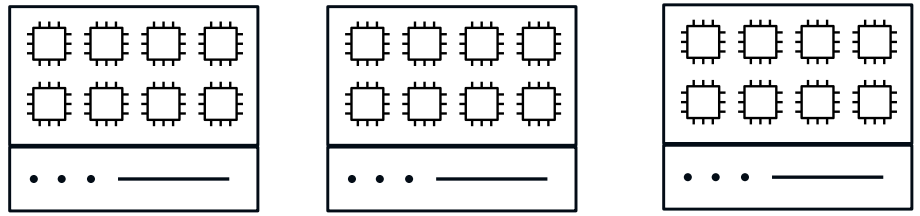
Model Parallelism

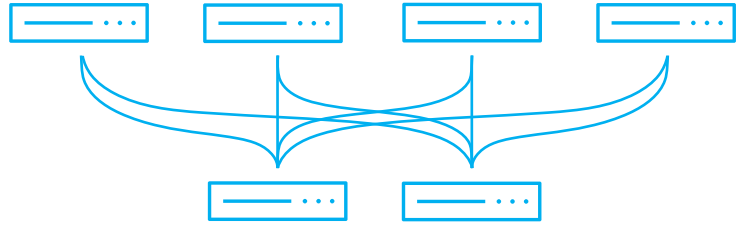
Llama 3.1 405B (bfloat16)

Precision	Estimated Model Size	With 20% Overhead
FP32	~1,620 GB	~1,944 GB
BF16/FP16	~810 GB	**~972 GB**
INT8	~405 GB	~486 GB
INT4	~202 GB	~243 GB



Parallelism	Split By	Best Use Case
Tensor	Tensors (weights)	Large models, compute-bound inference
Pipeline	Layers	Moderate models, latency-tolerant
Hybrid	Tensors + layers	Massive models like GPT-3/LLama 400B

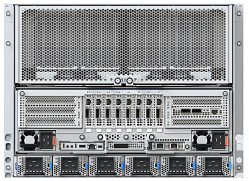




Intra-GPU Connectivity

Oversimplified view of UCS C885A and C845A based on NVIDIA HGX and MGX architectures

UCS C885A



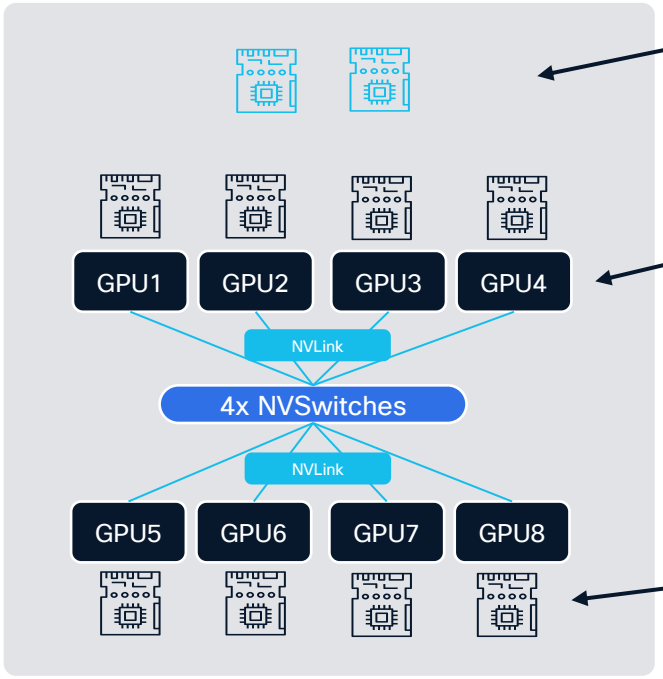
UCS C845A

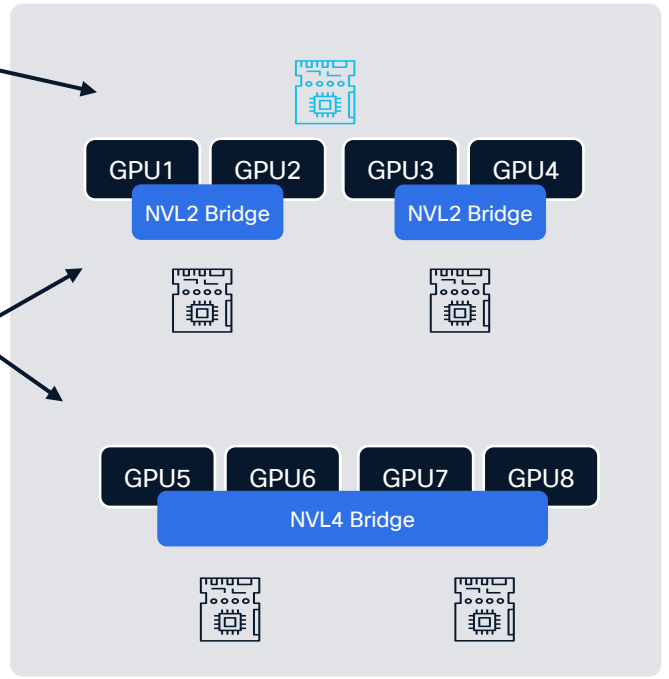
Dedicated NICs for north-south connectivity to front-end networks

Each GPU has multiple NVLink ports and connects to four non-blocking NVSwitches

2 or 4-Way Bridge configurations for intra-GPU connectivity.
Each GPU pair has also dedicated NICs GPU-to-GPU connectivity across nodes

Each GPU has dedicated NICs for GPU-to-GPU connectivity across nodes



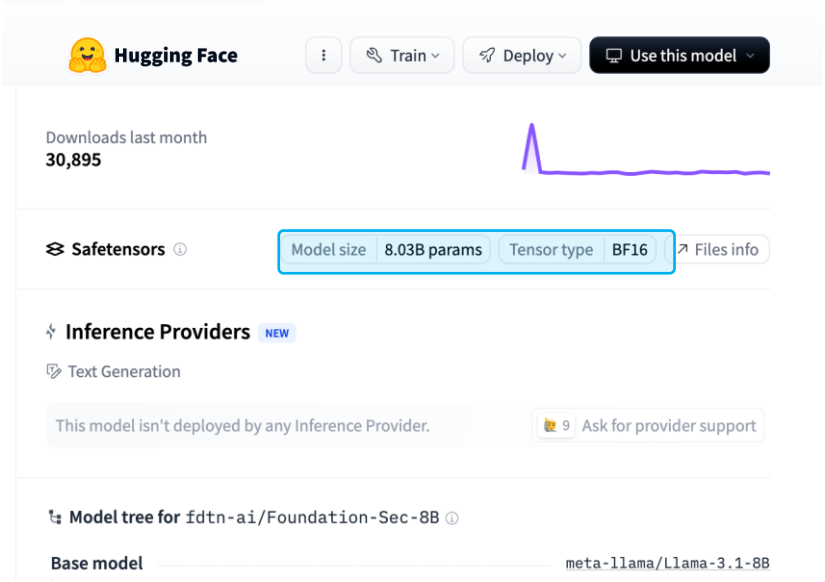


*CPU & PCIe switches omitted for brevity

Model to GPU Sizing

Raw Model Size (bytes)

= Number of Parameters × Size per Parameter



Precision	Size per Parameter
FP32 (float32)	4 bytes
FP16 (float16)	2 bytes
BF16	2 bytes
INT8	1 byte
FP8 (Hopper)	1 byte

= 8.03 Billion × 2 bytes = **16.06 GB**

Model weights only and not the full GPU memory requirement.

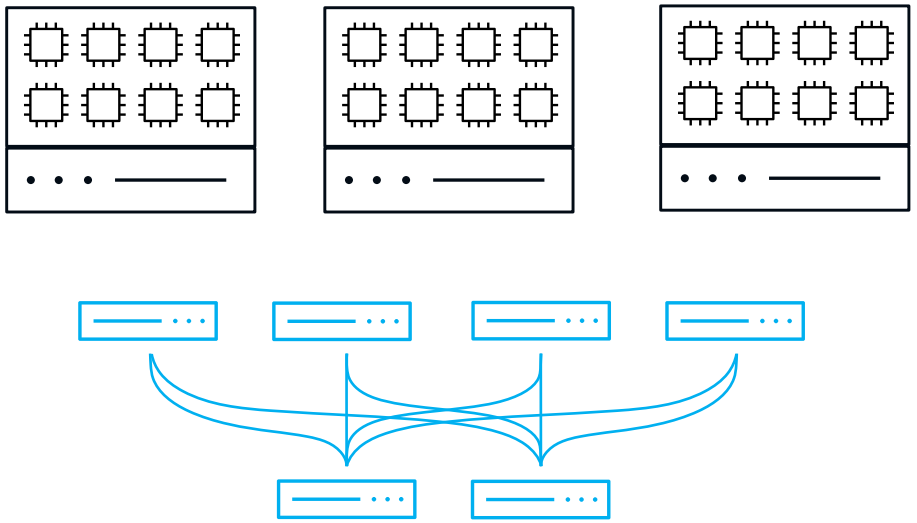
Training

Much higher memory usage due to backpropagation and optimizer state

Training Total Memory ≈ Raw size × (6–10)

Component	Estimate
FP16 Weights	16.06 GB
Activations (~2- 3× model size)	= 2.5 x 16.06 = 40 GB
Optimizer States (Adam, in FP32)	8.03B × 4 (bytes) × 2 (bytes) = **64.24 GB**
Total	**~120+ GB per GPU**

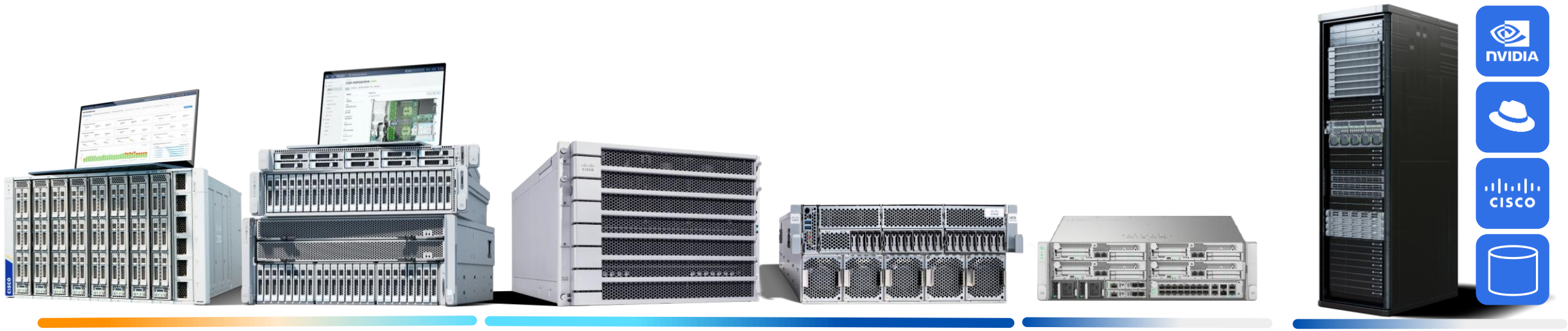
Single GPU Setup 1 x H200, 1 x B100/B200
Multi-GPU setup 2 x A100 80GB, 4 x A100 40GB, 2 x H100 80GB



Cisco AI Compute Portfolio

Unified approach to accelerated AI compute

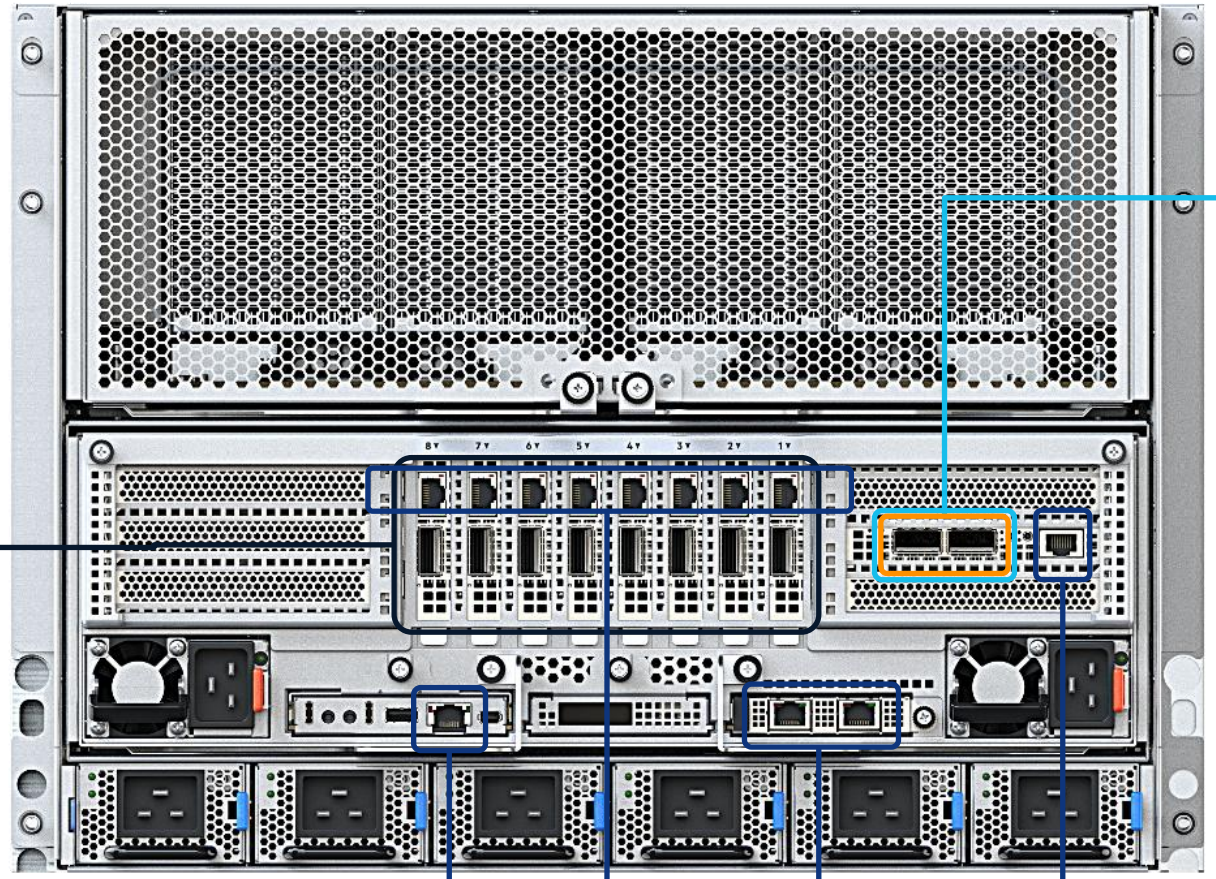
Validated solutions for AI with compute, network, storage, and software



GPU Accelerated | GPU Optimized | Unified Edge | AI POD

Build the model | Training | **Optimize the model | Fine-tuning and RAG** | **Use the model | Inferencing**

UCS C885a M8 External Connectivity



Front-End Network

Storage Network

May be converged to a shared physical network

Inter-GPU Backend Network

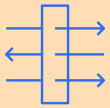
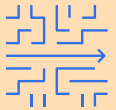
Management Network

8 x 400G GPU
2 x 400G CPU
3 x 1G BMC
2 x 10G Management

Network Design

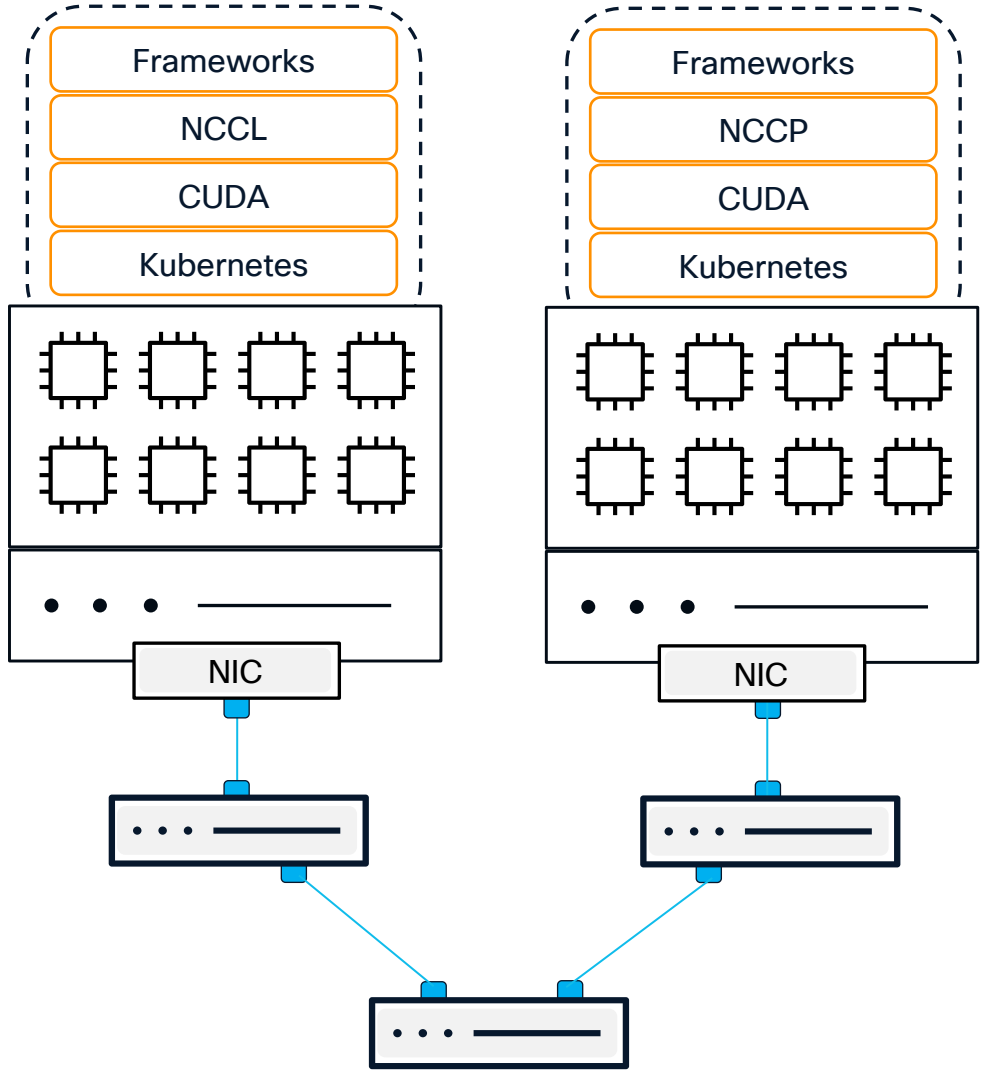
Engineer tightly integrated networking stacks to optimise training and inference

AI/ML Key Network Requirements

	1	2	3	4
Requirement	 Non-Blocking, High Bandwidth	 Lossless	 Low, Consistent Latency	 Operations
Driver	Data transferred during training increases at almost **6x** model size	**2%** packet loss can cause training to be **8x** slower	Increased network delay and variance degrade distributed training performance	AI Infrastructure is expensive, downtime and performance degradation are expensive
Solution	400/800G Dynamic Load Balancing Static Pinning ECMP with UDF	RoCEv2 DCQCN (PFC/ECN) Dedicated buffers	Optimised ASICs Rich QoS Intelligent Buffer Management	AI Fabric Automation AI Fabric Visibility Telemetry

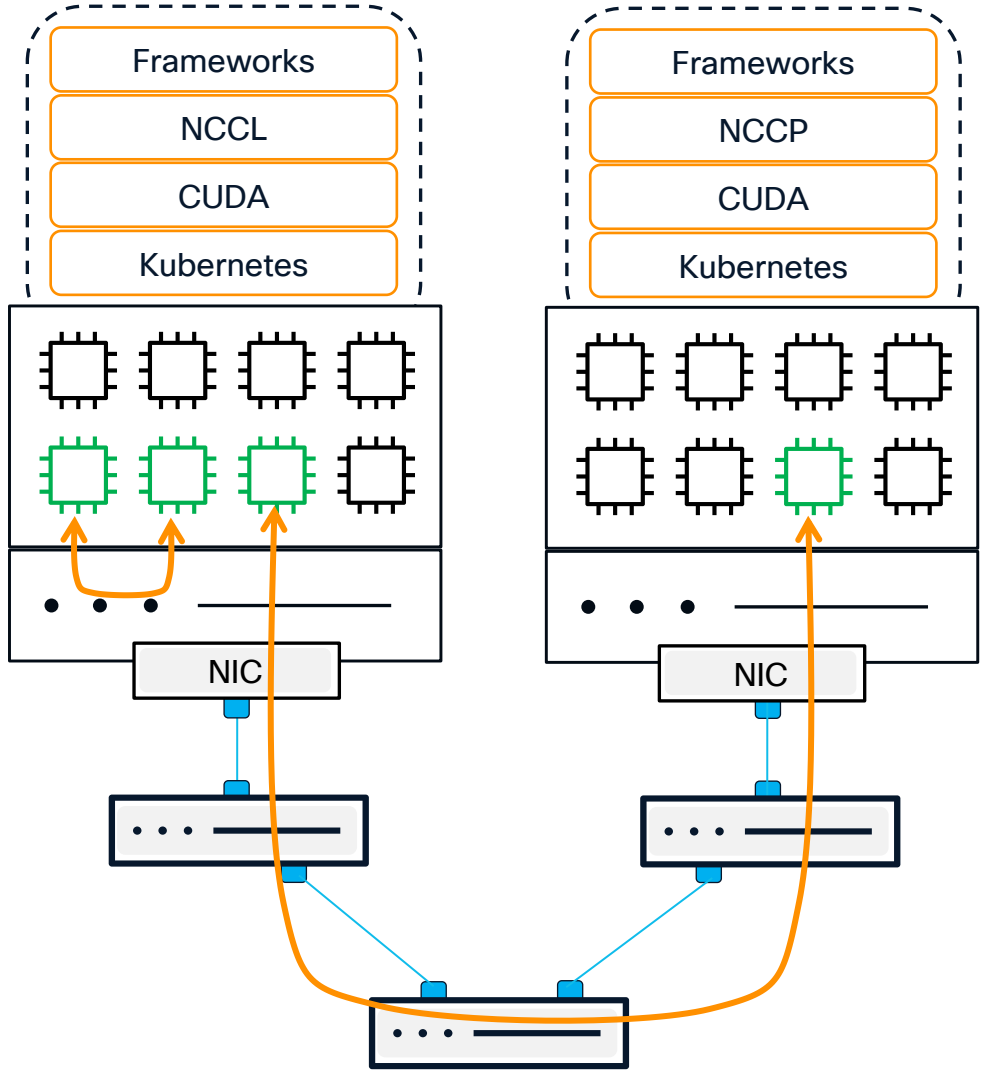
AI depends on tight integration between software and hardware

optimized for availability, performance and scale



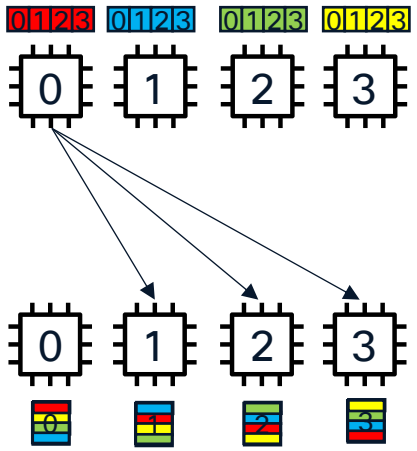
AI depends on tight integration between software and hardware

optimized for availability, performance and scale



Collective Operations

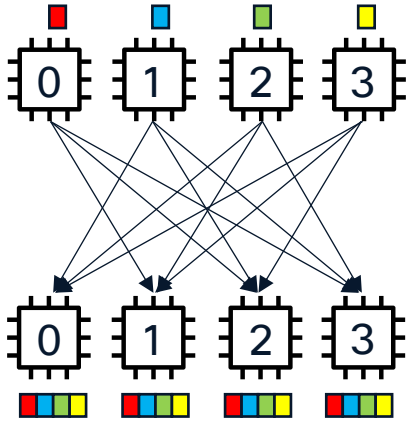
Reduce-Scatter



Data is reduced across GPUs and each gets a portion of the result

Optimized version of AllReduce

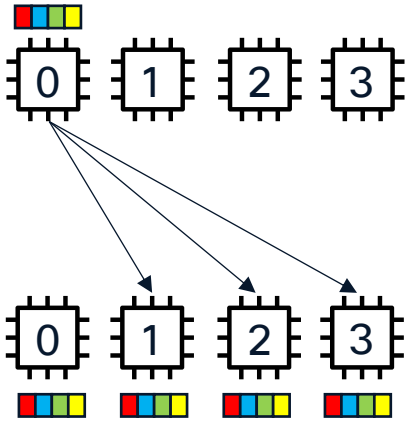
AllGather



Each GPU shares its data, and all GPUs receive the complete combined result

Sharing partial tensors or embeddings

Broadcast



One GPU sends data to all others

Sending model weights to all GPUs

AllReduce = Reduce-Scatter + AllGather

Two sides to a collective

Logical

All-to-All, All-Reduce, All-Gather, Reduce, Scatter, Gather, Broadcast and Reduce-scatter

Transport

Ring, Halving-doubling pattern, Binary trees, Flat-tree, K-Nomial Trees, Direct Algorithms

cisco

GPU to GPU across the Network

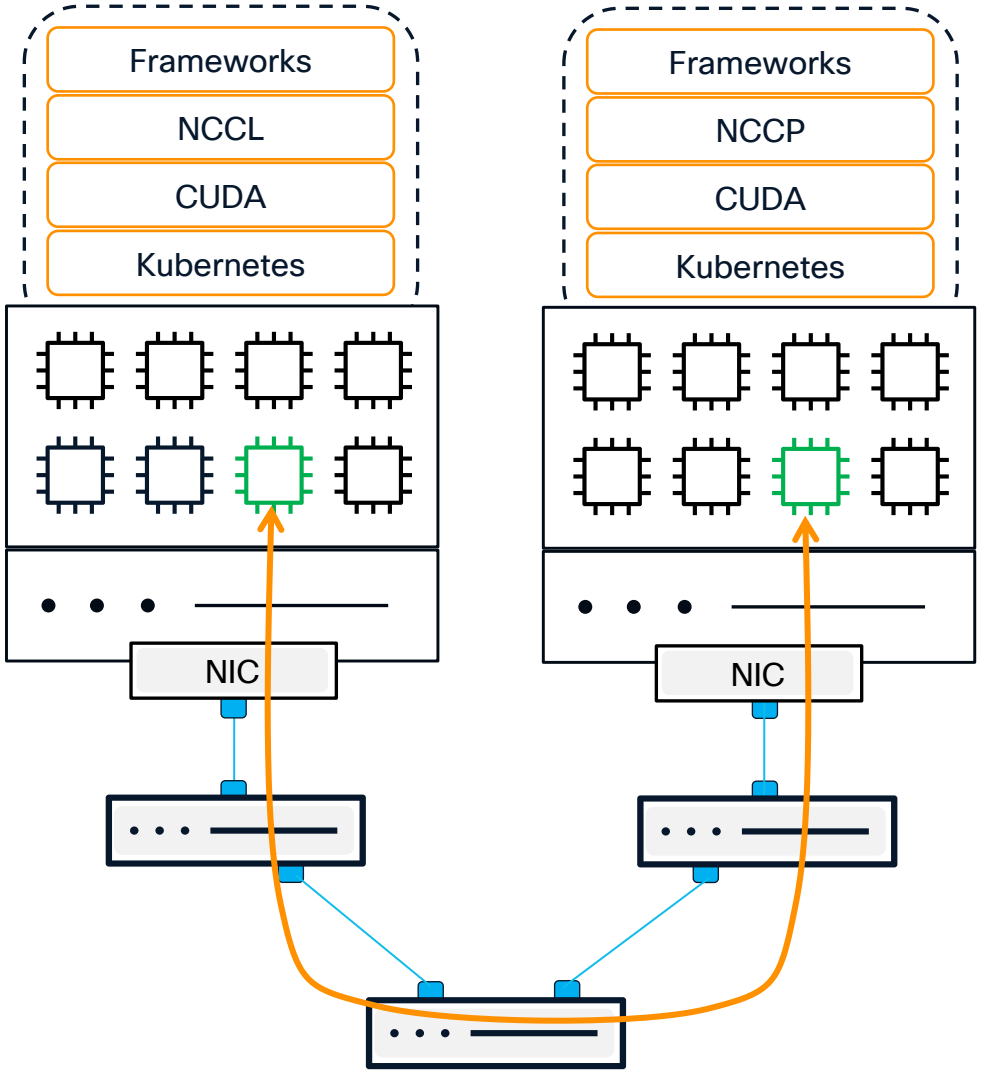
RDMA over Converged Ethernet (RoCEv2)

RDMA bypasses CPU to optimise performance using routable UDP. RoCEv2 is implemented on the server NIC.

Lossless Transport

UDP transport assured by **Explicit congestion notification** (ECN) and **Priority Flow Control** (PFC) to deliver a lossless fabric

RoCEv2 Packet

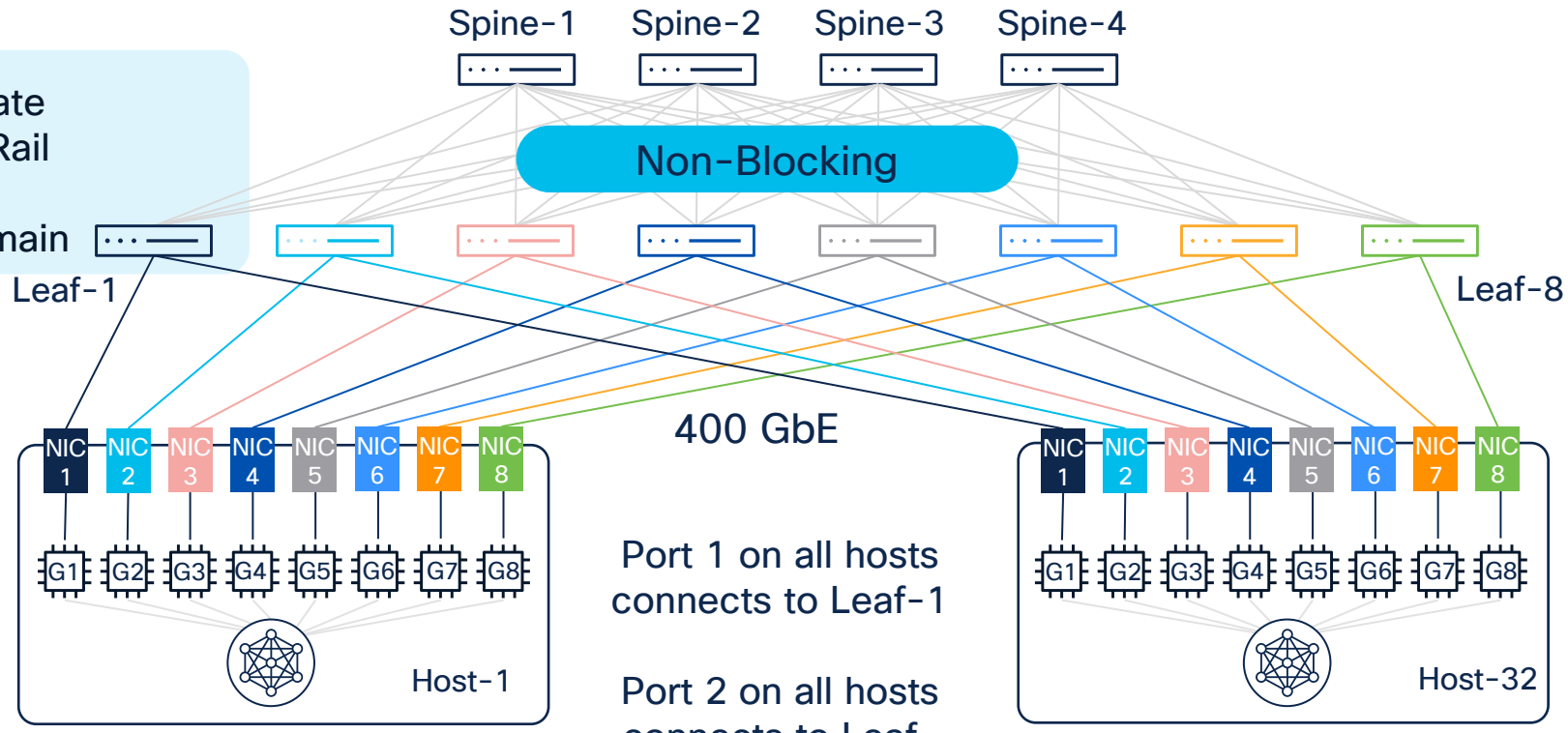


Backend Network

Rail Design

Crossing domains leverage the spine enabling All-to-All Collective

GPUs communicate within the same Rail allowing for local switching HB domain



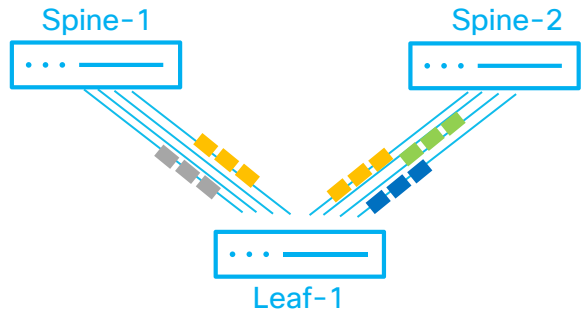
Port 1 on all hosts connects to Leaf-1

Port 2 on all hosts connects to Leaf-2, and so on

Collective operations overlay across the fabric to share or aggregate data.

Load Balancing Backend Network Links

Required when traffic traverses the spines



Dynamic Load Balancing (DLB)

ECMP

Default Equal Cost multi-path load balancing based on layer 4 UDP flow.

Flowlet

A flowlet is a **burst of packets** within a flow that belongs to the same connection.
Single long-lived flow (like an NCCL collective)split across **multiple paths** in the network **without packet reordering** and can be safely routed on a different path than previous or future flowlets

Per Packet

PPLB splits traffic **at the packet level**, not at the flow or flowlet level. That means **every individual packet** can be routed over a different link or path, based on current congestion or scheduling logic.

Packets can arrive out of order, so they need re-ordering on switches or NIC

IB Queue Pair

An **RDMA queue pair** (Send Queue, RX queue) has a unique **QP number**. Load balancing is based on packets with the same QP number in the same 5-tuple flow. All packets related to the same destination QP must arrive in order

Reference Architectures

Enterprise Reference Architecture (ERA)

- For enterprise, commercial and public sector customers
- Cluster of 32 – 256 GPUs

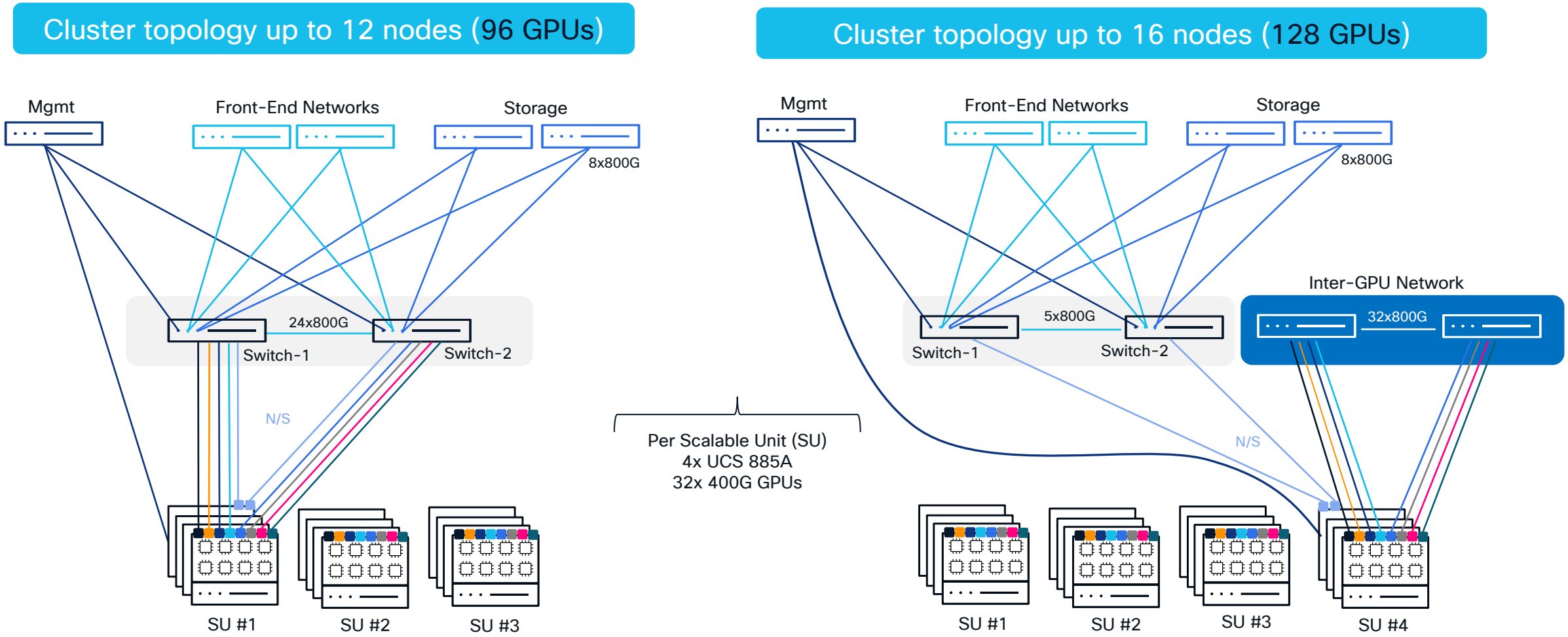
NVIIDA Cloud Partner (NCP)

- For cloud provider and web scale customers
- Clusters of 256 – 16,000+ GPUs

https://resources.nvidia.com/c/collection-9563de54?x=sTI8RF

Example for small clusters, but not limited to

Adheres to the NVIDIA Enterprise Reference Architecture (ERA)



Cisco AI Networking Reference Architectures

AI Infrastructure with Cisco Nexus 9000 Switches Data Sheet

Cisco Reference Architecture

Updated: March 18, 2025

Bias-Free Language

Table of Contents

- Introduction
- Hardware +
- Networking topologies +
- Storage architecture
- Software +
- Security
- Testing and certification
- Summary
- Appendix A – Compute server ...
- Appendix B – Control node ser...
- References

Save Translations Download Print

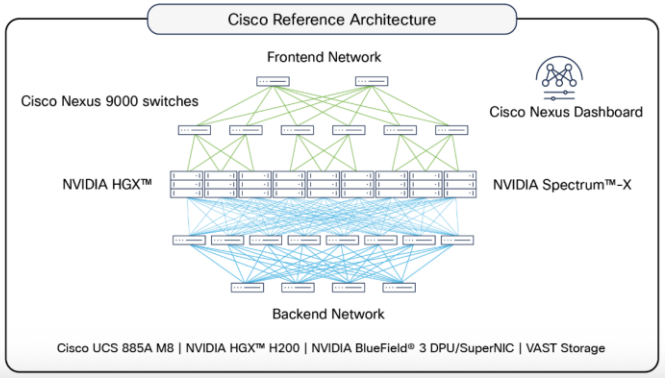
Featuring Cisco UCS C885A compute servers with NVIDIA HGX™ H200 and NVIDIA Spectrum™-X

Introduction

This Cisco Reference Architecture is based on Cisco Nexus 9000 switches for networking AI clusters managed by the on-premises Nexus Dashboard platform. It adheres to the NVIDIA Enterprise Reference Architecture for NVIDIA HGX™ H200 with NVIDIA Spectrum™-X networking.

Cisco Nexus 9000 s witches, powered by Cisco Silicon One and Cloud Scale architectures, provide high-speed, deterministic, low-latency, and power-efficient connectivity for AI and High-Performance Computing (HPC) workloads. With the availability of multiple form-factors, optics, and rich software features of the NX-OS operating system, Nexus 9000 switches provide a unified experience for backend, frontend, management, and storage networks (see Figure 1).

Cisco Nexus Dashboard is the operations and automation platform for managing the Nexus 9000 Switch-based fabrics. It complements the data-plane features of the Nexus 9000 switches by simplifying their configuration using built-in templates. It detects network health issues, such as congestion, bit errors, and traffic bursts in real time and automatically flags them as anomalies. These issues can be resolved faster using integrations with commonly used tools, such as ServiceNow and Ansible, allowing the networks of an AI cluster to be aligned with the existing workflows of an organization.



NVIDIA Certified Cisco Nexus Hyperfabric AI Enterprise Reference Architecture Data Sheet

Updated: May 14, 2025

Bias-Free Language

Table of Contents

- Introduction
- Hardware +
- Networking topologies +
- Storage architecture
- Software +
- Security
- Testing and certification
- Summary
- Appendix A – Compute server ...
- Appendix B – Control node ser...

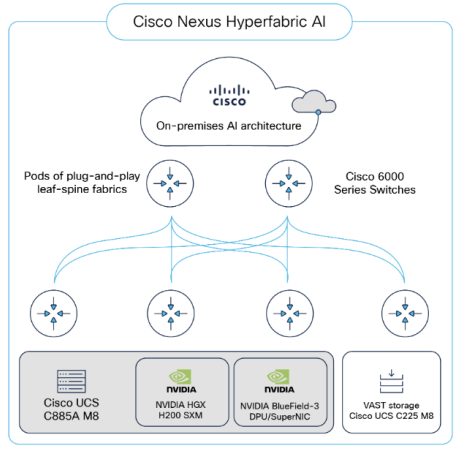
Save Translations Download Print

Cisco Nexus Hyperfabric AI Enterprise Reference Architecture certified by NVIDIA, featuring Cisco® cloud-managed AI/ML networking of Cisco UCS® C885A M8 Racker Servers with NVIDIA HGX™ H200 and NVIDIA Spectrum™-X.

Introduction

Cisco Nexus® Hyperfabric AI is an on-premises AI cluster that is managed by a cloud-hosted controller. It empowers and simplifies your AI initiatives and accelerates AI deployments with a comprehensive, integrated, cloud-managed solution. Cisco Nexus Hyperfabric AI Reference Architecture is based on Cisco Silicon One® switches and adheres to the NVIDIA Enterprise Reference Architecture (Enterprise RA) for NVIDIA HGX H200 and Spectrum-X.

Figure 1 shows the key components of the solution. The key hardware components used in the cluster are described in the next section.



Minimize Attack Surface

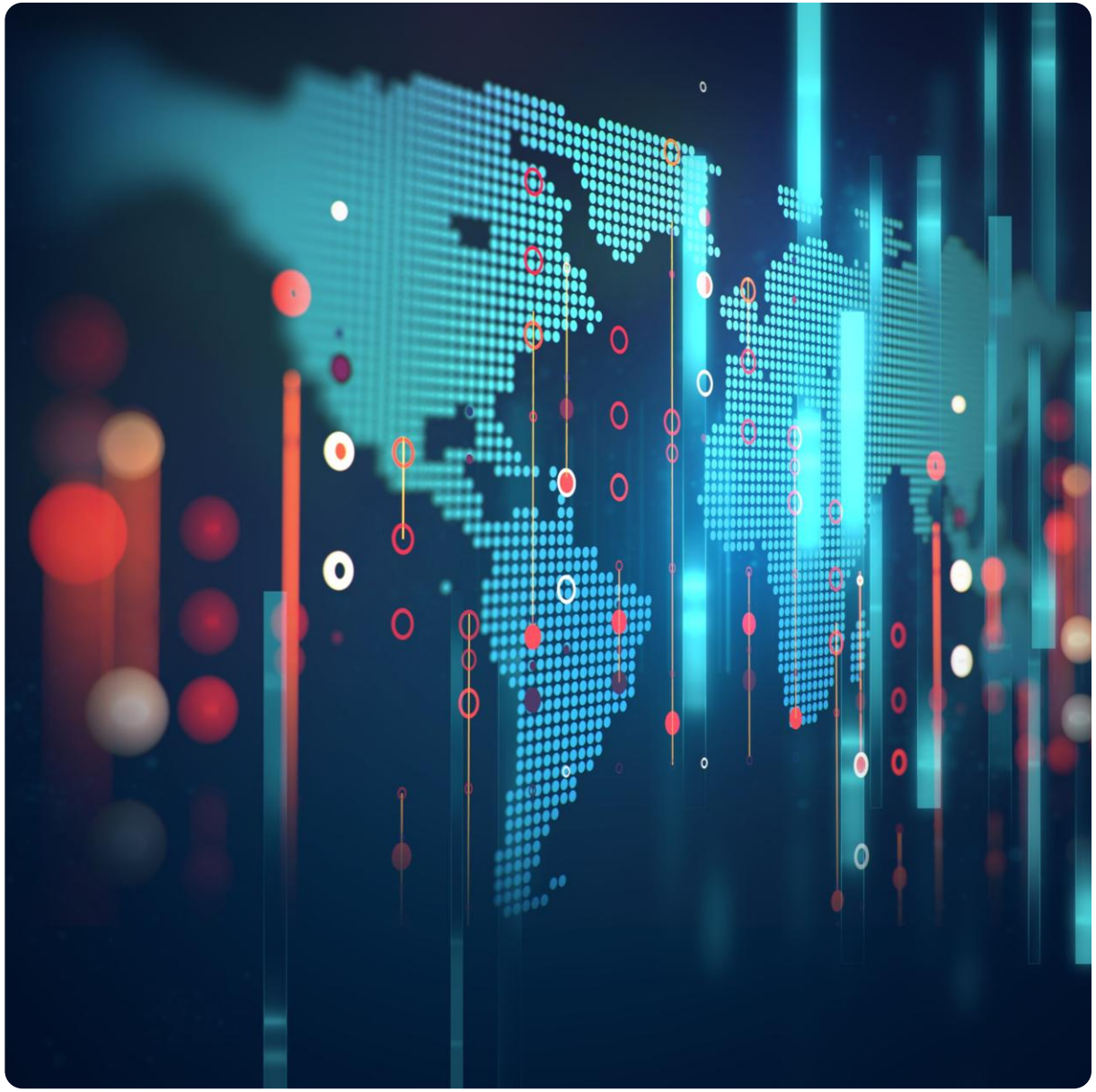
Embed security throughout the full stack from infrastructure to model to protect data, systems and users.

Security requires a cohesive and integrated toolset

Data Source risks include data poisoning, exfiltration and accidental leaks

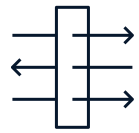
Training risks include open-source model contamination, insecure APIs and unknowingly consuming 3rd party IP

Inference risks include denial of service attacks, prompt injection and data mining



Hybrid Mesh Firewall

Security Cloud Control



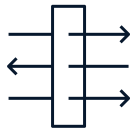
Secure Firewall

Secure Workload

Cilium

Hypershield

Smart Switch



3rd Party Firewall

Protecting from the infrastructure to the application layer

AI Defense: Develop, deploy, and run secure AI

Discover

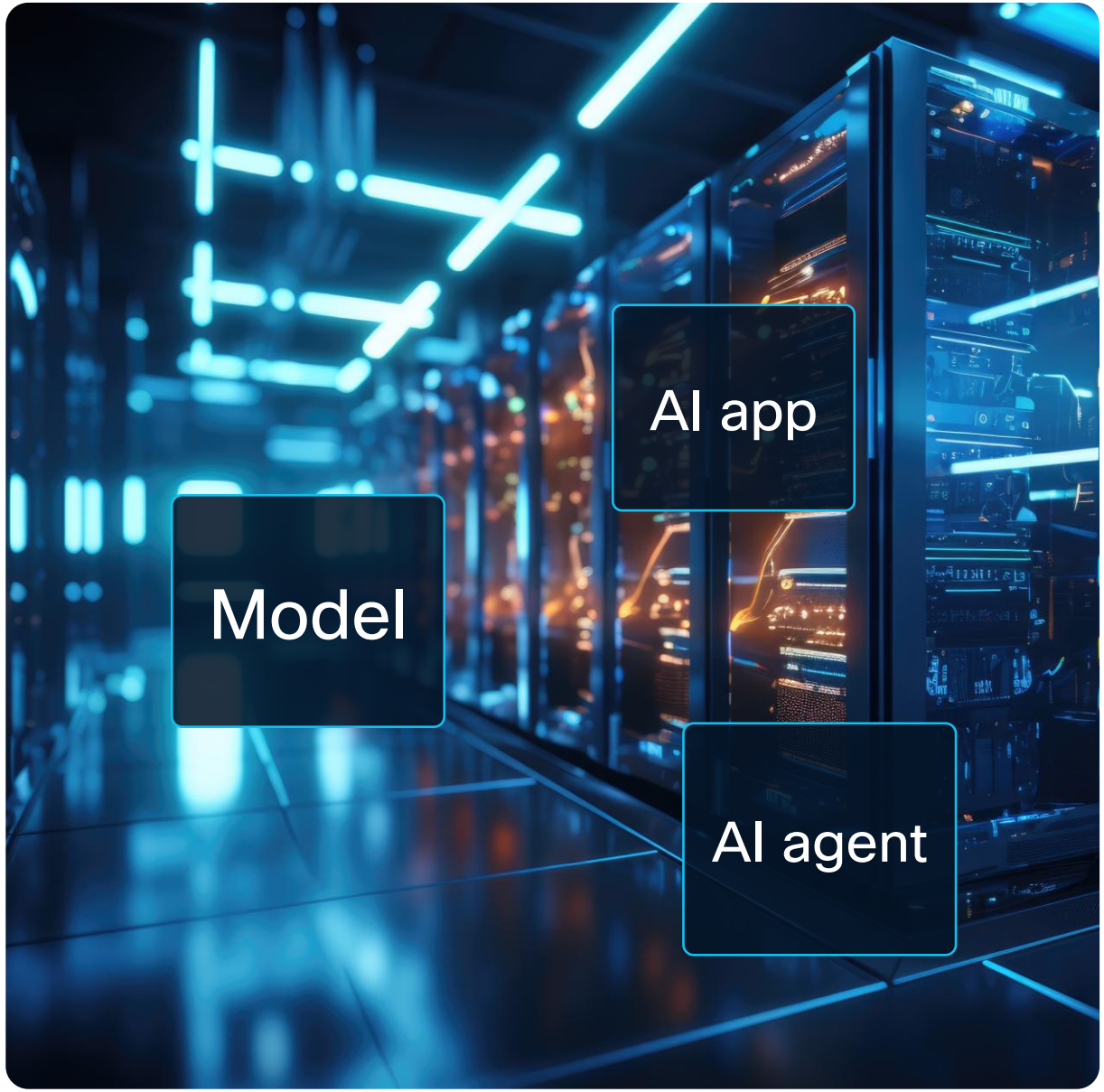
Uncover enterprise AI usage

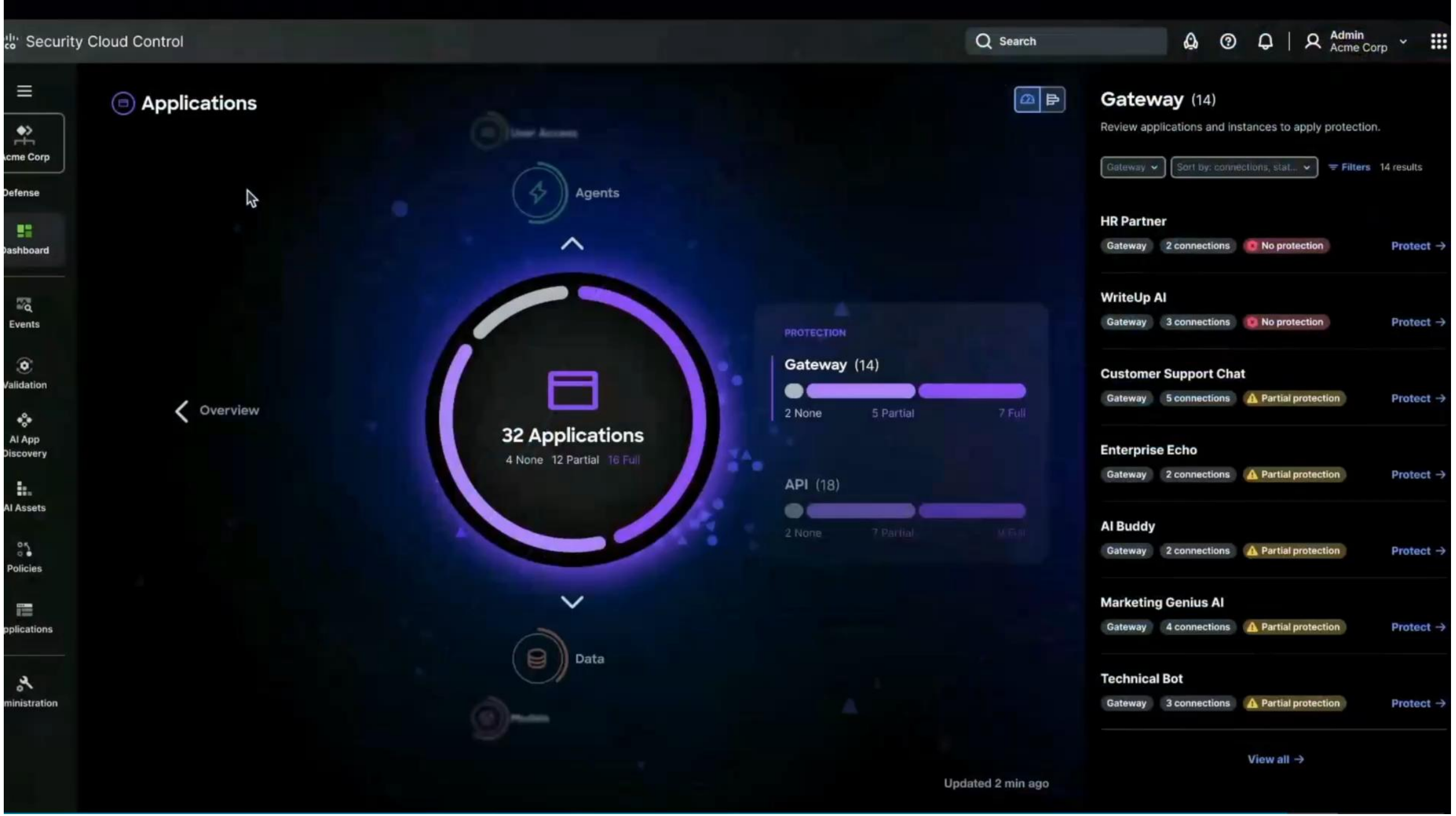
Detect

Test for risk, vulnerabilities, and adversarial attacks

Protect

Defend against runtime threats with guardrails



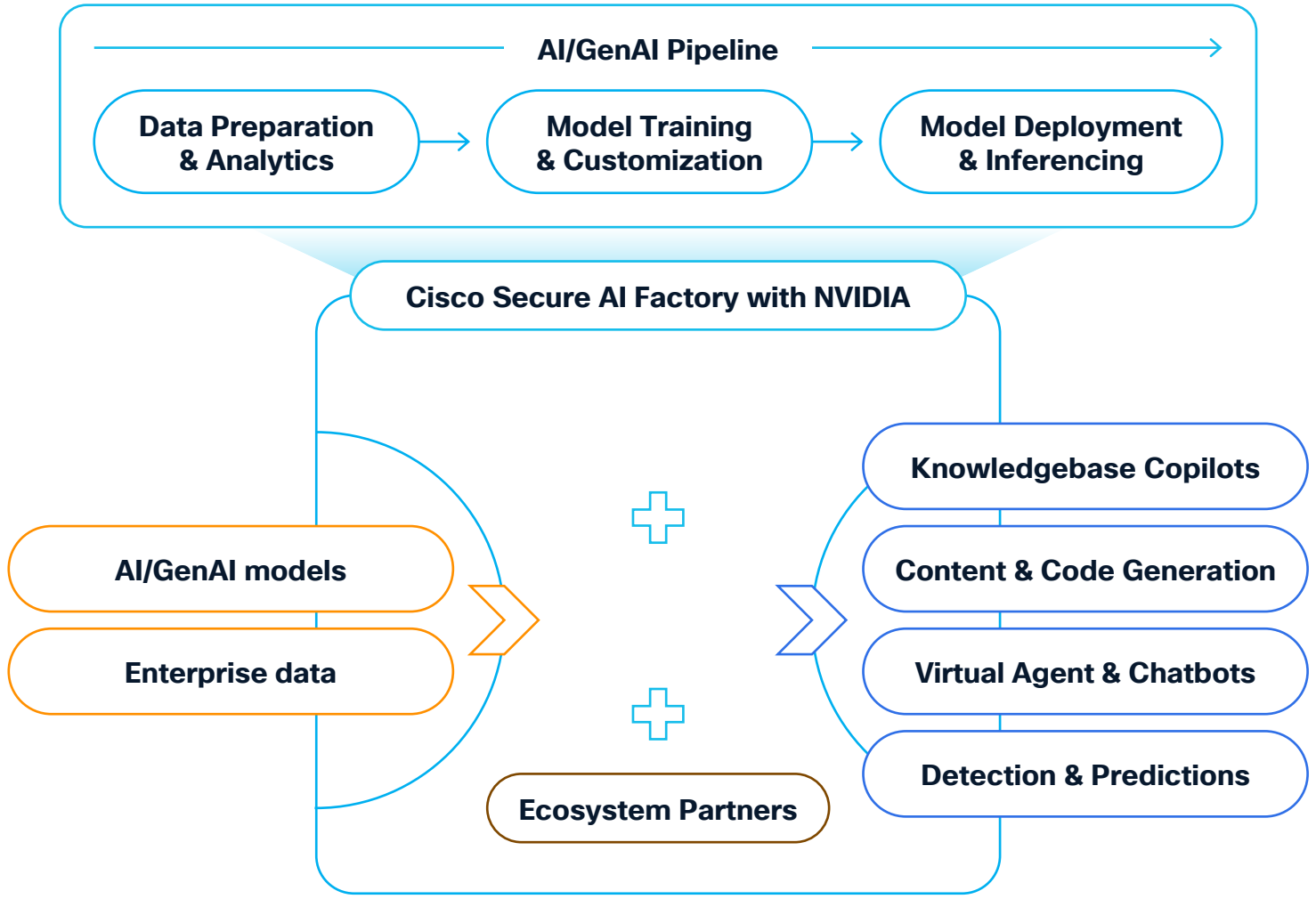


Bringing it all together

Cisco Secure AI Factory

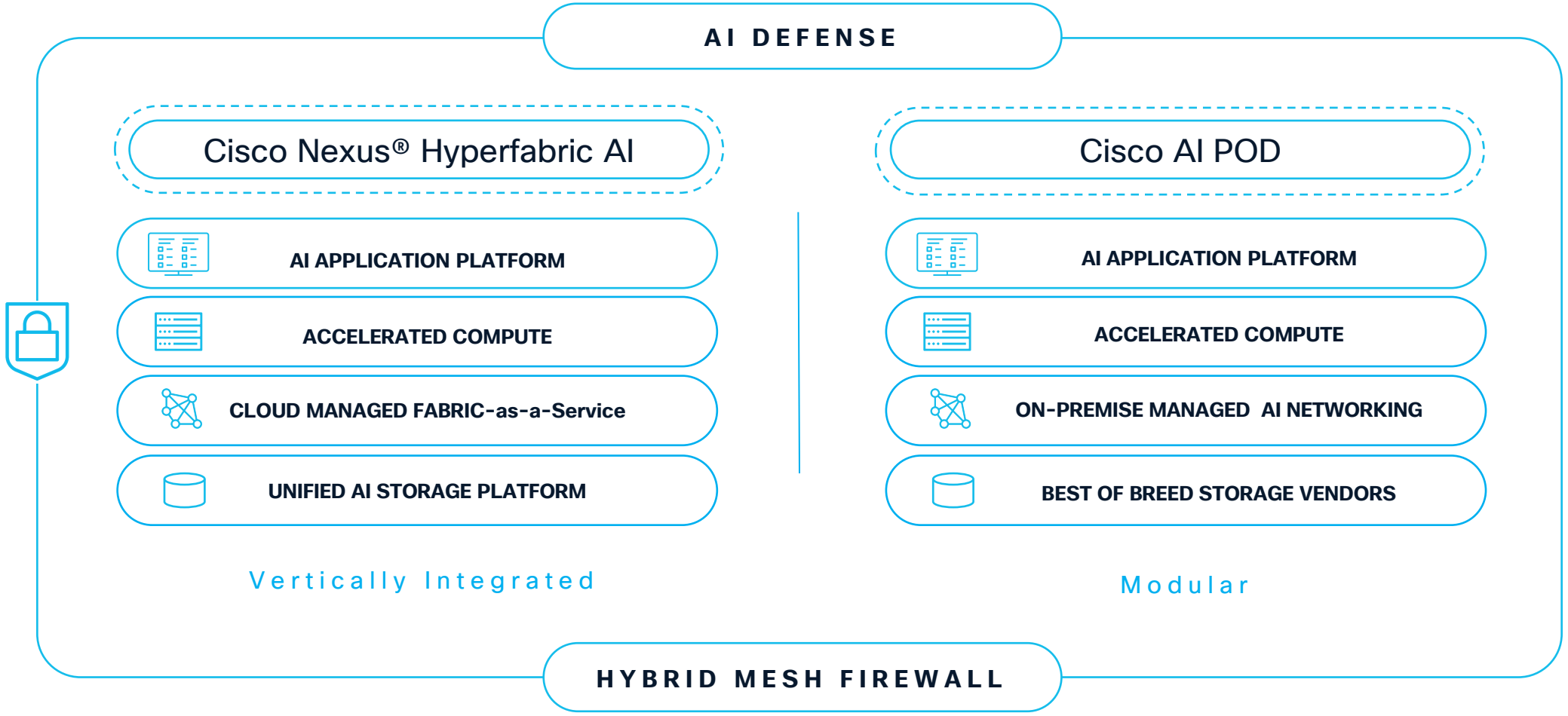
Accelerate the AI adoption for enterprises with an integrated AI infrastructure and software solution from Cisco, NVIDIA, and our strategic partners

- Security-first architecture
- High-performance, enterprise-proven networking, compute, storage, and AI software
- Pre-validated, with flexible deployment options: Vertically integrated or modular system



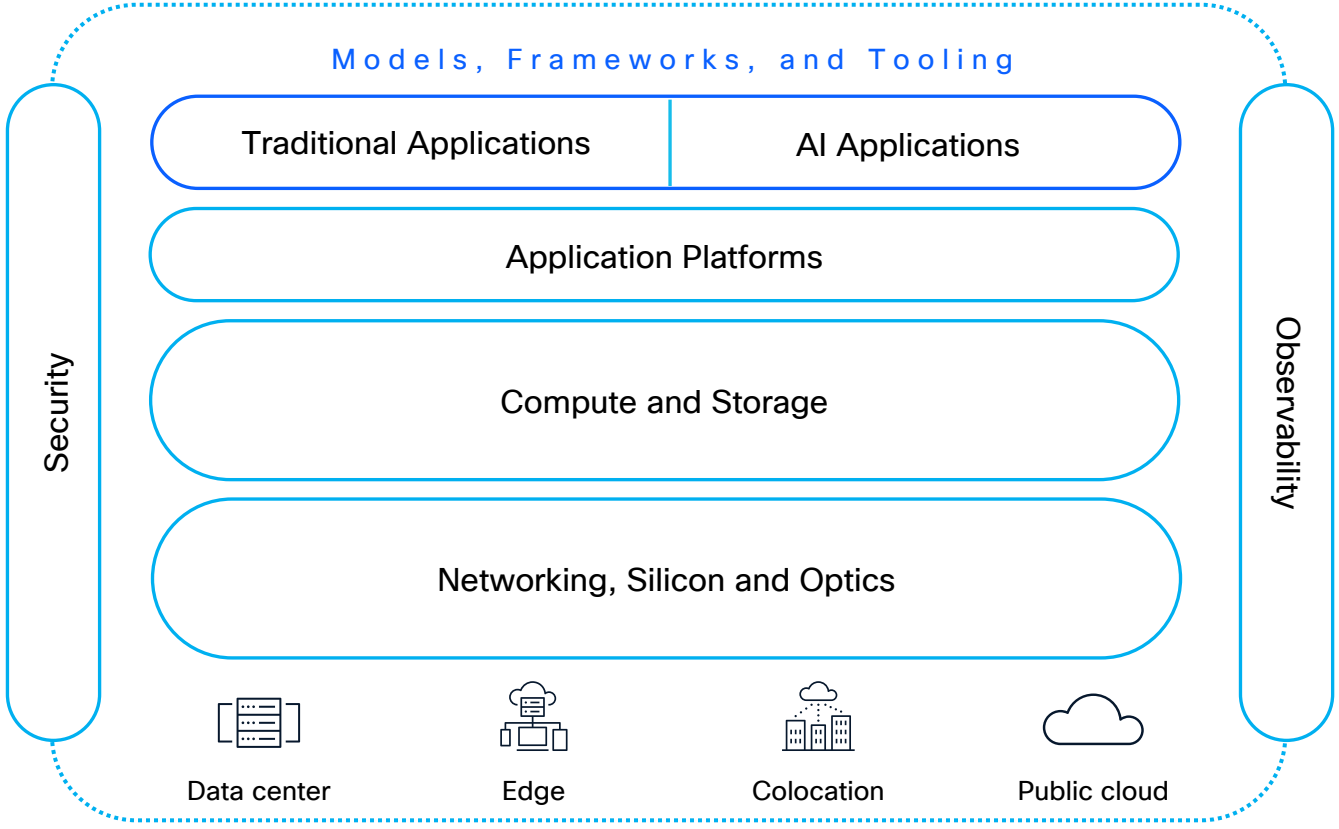
Cisco Secure AI Factory

Fabrication Options



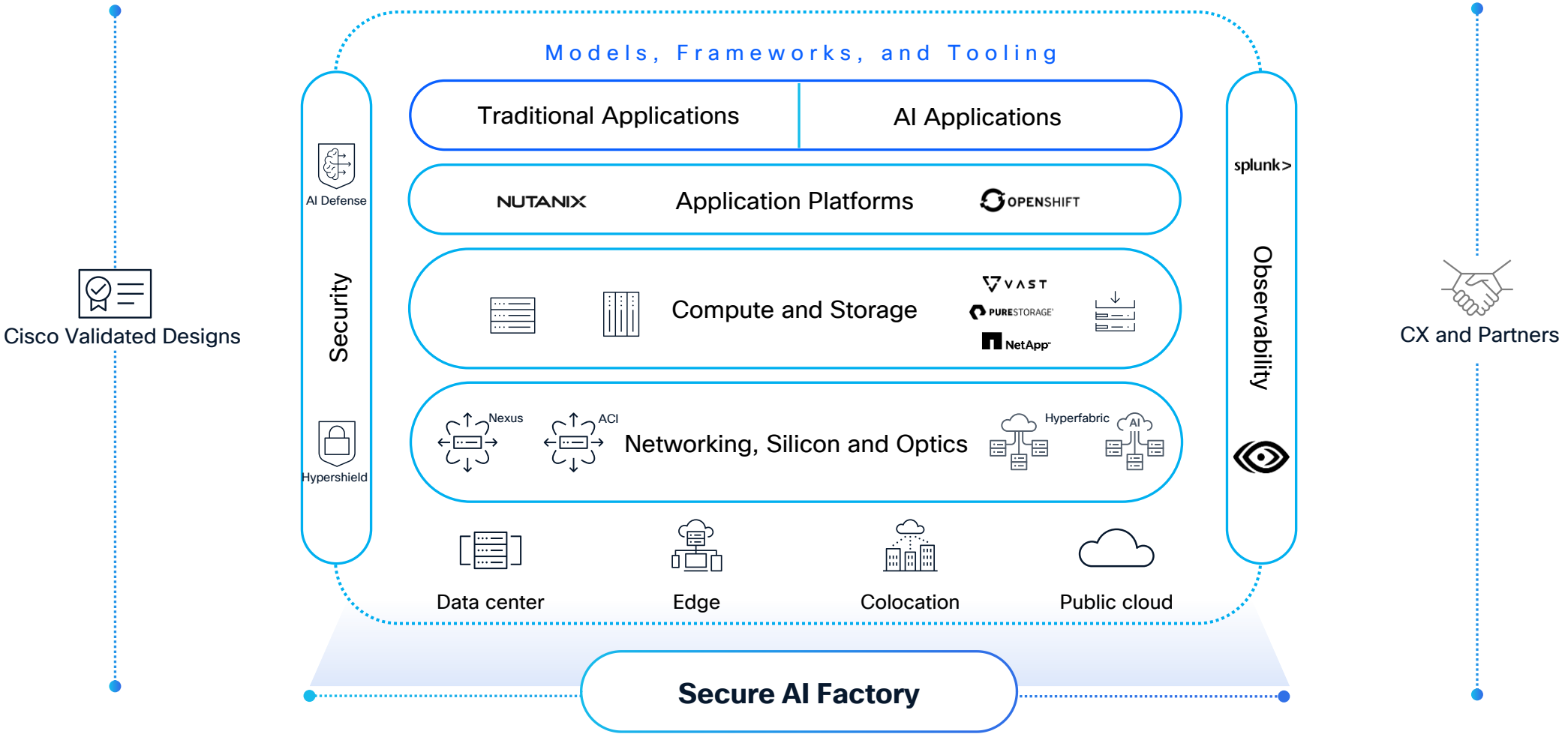
AI-ready data Center

Tightly integrated full stack



AI-ready data Center

Tightly integrated full stack



Only Cisco unifies networking, compute, security and observability to deliver the AI-ready data centers.

cisco

Complete your session evaluations

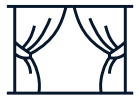
Complete a minimum of 4 session surveys and the Overall Event Survey to claim a Cisco Live T-Shirt.

Earn up to 800 points by completing all surveys and climb the Cisco Live Challenge leaderboard.

Level up and earn exclusive prizes!

Complete your surveys in the Cisco Live Events app.

Continue your education



Visit the Cisco Stand for related demos

Book your one-on-one Meet the Expert meeting

Attend the interactive education with Capture the Flag, and Walk-in Labs

Visit the On-Demand Library for more sessions at www.CiscoLive.com/on-demand

Contact me at: : cgascoig@cisco.com

Thank you


CISCO