

# What Goes Into an AI Cluster

Nate Reid- Solutions Engineer, AI



*Any information provided in this document regarding future functionalities is for informational purposes only and is subject to change including ceasing any further development of such functionality. Many of these future functionalities remain in varying stages of development and will be offered on a when-and-if available basis, and Cisco makes no commitment as to the final delivery of any of such future functionalities. Cisco will have no liability for Cisco's failure to deliver any or all future functionalities and any such failure would not in any way imply the right to return any previously purchased Cisco products.*

# Agenda

- 01 AI Infrastructure Considerations
- 02 Cisco AI PODs
- 03 AI Workload Orchestration
- 04 Operations and Automation
- 05 MLOps
- 06 Cisco Secure AI Factory
- 07 Tying it all Together
- 08 Key Takeaways

# But First... a Quickish Level-Set on Some Terminology

- Token
- Vector
- Matrix
- Embedding
- Parameter
- Foundational Model
- Pre-training
- Fine-tuning
- Inferencing
- Agentic AI
- Guardrails

# Sizing Example: Llama 3.x 70B Full Training

$$\text{Required VRAM} \approx \underbrace{P * \text{Sum\_of\_Precision}}_{\text{Parameter memory}} + \underbrace{(\text{Bpe} * L(3\text{BSH} + 2\text{BS}(\text{Ff}*H) + 2(\text{BSI})))}_{\text{Activation memory}}$$

$$70\text{B} * 16 + (2 * 80(3 * 1 * 8192 * 8192 + 2 * 1 * 8192 * (8192 * .125) + 2 * 1 * 8192 * 28672))$$

Parameter Memory Term  $\approx 70\text{B} * 16 \approx 1.1\text{TB}$

Activation Memory Term  $\approx 2 * 80(200\text{MB} + 16\text{MB} + 470\text{MB}) \approx 110\text{GB}$

Additional  $\sim 10\%$  overhead buffer  $\approx 100\text{GB}$

**Total Memory  $\approx 1.3\text{TB}$**

P=Total\_parameters Bpe=Bytes\_per\_element Sum\_of\_Precision=Bytes\_per\_element+Gradients+Optimizer L=Layers B=Batch\_size H=Hidden\_size S=Sequence\_length  
I=Intermediate\_size Qh=Number\_query\_heads KVh=Number\_KV\_heads Ff=FlashAttention Factor

# Llama 3.3 70B Full Training – Key Takeaways

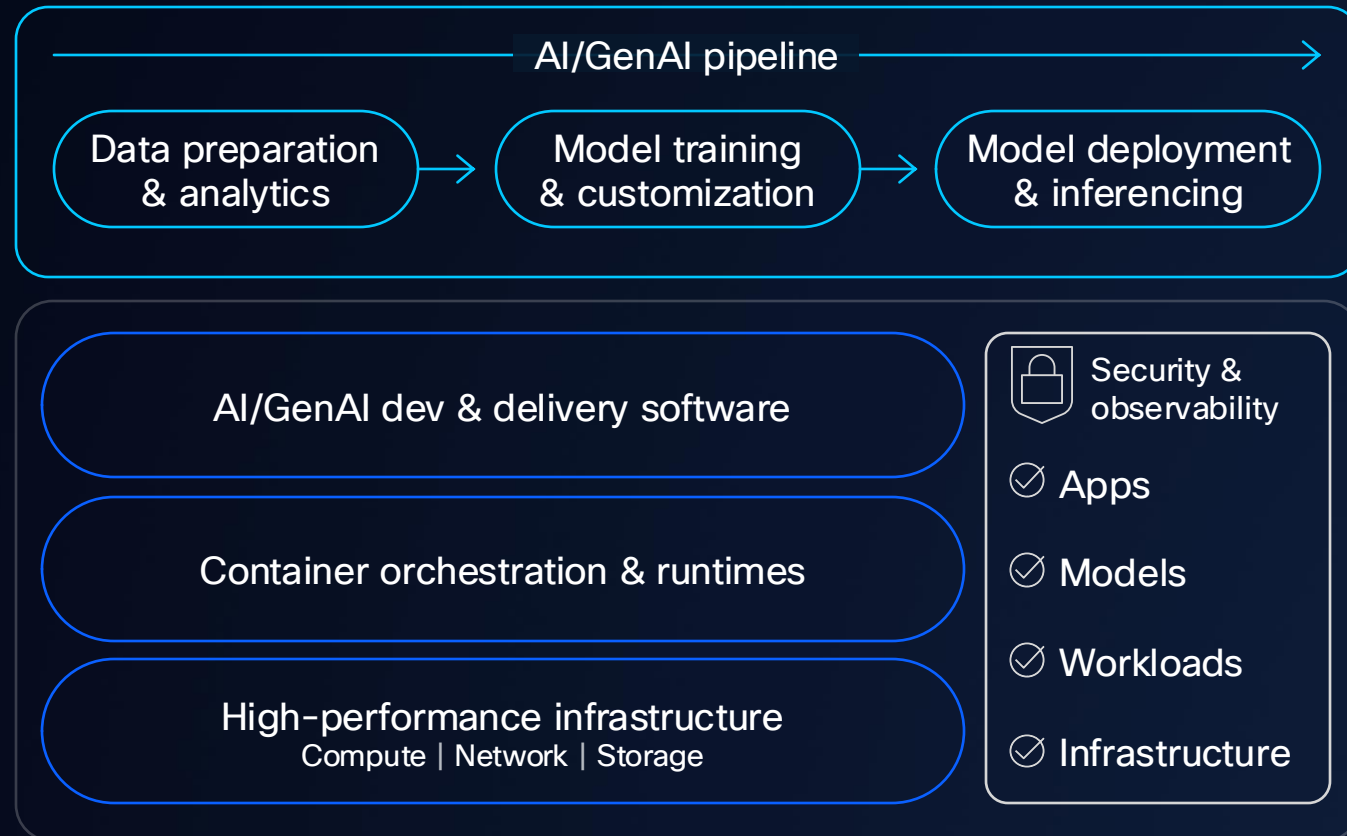
- 1.3TB represents the minimum memory required to train a single sequence of 8192 tokens at once.
- For reference, an H100 GPU holds 80GB VRAM.
  - 16 H100 GPUs minimum for a single sequence of 8192 tokens.
- Meta trained Llama 3.3 on 15 trillion tokens and eventually at a sequence length of 128K.
- Llama-3.x-70B consumed 7 million GPU hours on NVIDIA H100 GPUs.
- Meta built two 24,000 GPU clusters for Llama 3.x training.
- Many multi-GPU optimization strategies are applied at this level. Beyond the scope of this session.
- Key intuition:
  - Full model training at this level is not realistic for most organizations. This is why we have finetuning.

# Agenda

- 01 **AI Infrastructure Considerations**
- 02 Cisco AI PODs
- 03 AI Workload Orchestration
- 04 Operations and Automation
- 05 MLOps
- 06 Cisco Secure AI Factory
- 07 Tying it all Together
- 08 Key Takeaways

# Enterprises Need Capable Infrastructure to Operationalize AI

AI applications are different, and they are driving demand for new architectures, security mechanisms, and lifecycle management



**Faster time to business value**

**Mitigate risks**

**Simplify deployment**

**Reliability, availability, flexibility**

# Enterprise AI Infrastructure Requirements



## Customizing foundational models

Training LLMs from scratch is cost-prohibitive for the average enterprise



## Support multiple, smaller workloads

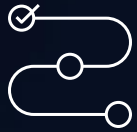
Enterprises can have many use cases spread across different LOBs, each using an LLM



## Integrate into existing data centers with ease

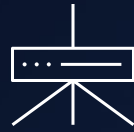
AI-enabled enterprise applications often need data, applications and other resources in existing data centers

# Enterprise AI Infrastructure Requirements



## Operational ease and consistency

Existing DC operational model to manage AI backend/frontend fabrics; simplify with fabric + server templates or IaC



## Consistency and simplicity at scale

Building block approach using a spine-leaf architecture to scale-out with consistency and predictability



## Fit-to-purpose AI infrastructure

Performant infrastructure without compromising on choice; established technologies

# Common AI Challenges



## Unclear business objectives & priorities

Unclear direction hinders cross team collaboration, creates confusion, and hampers acquisition of necessary skills



## Complex AI infrastructure deployment

Lack of high-performance infrastructure with integrated compute, network, storage, and AI software can stall AI projects



## Security vulnerabilities

AI models, frameworks, apps, and supporting infrastructure represent a new cyberattack surface



## Network performance challenges

Model training and inferencing can generate a lot of traffic, slowing networks. Multi-node training and inferencing demands significantly more from the network

# Cisco AI-Ready Data Center

Transform data centers to power AI workloads anywhere

Solving key AI Challenges...



Complexity of AI infrastructure deployments



Security vulnerabilities



Network performance bottlenecks

...by leveraging critical components:



Purpose-built integrated AI infrastructure



Full stack AI systems



E2E AI infrastructure operations



AI focused and equipped security

← Digital Resilience: security, reliability, and performance →

# Agenda

- 01 AI Infrastructure Considerations
- 02 Cisco AI PODs**
- 03 AI Workload Orchestration
- 04 Operations and Automation
- 05 MLOps
- 06 Cisco Secure AI Factory
- 07 Tying it all Together
- 08 Key Takeaways

# Cisco AI PODs

AI Practitioners  
/ MLOps

IT Infrastructure  
& Operations

NVIDIA AI Software



Kubernetes Platform



Cisco Networking & Optics



Cisco Compute



Partner Storage



# Cisco AI PODs

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

Deploy AI with confidence

**Cisco CVD, NVIDIA ERA**

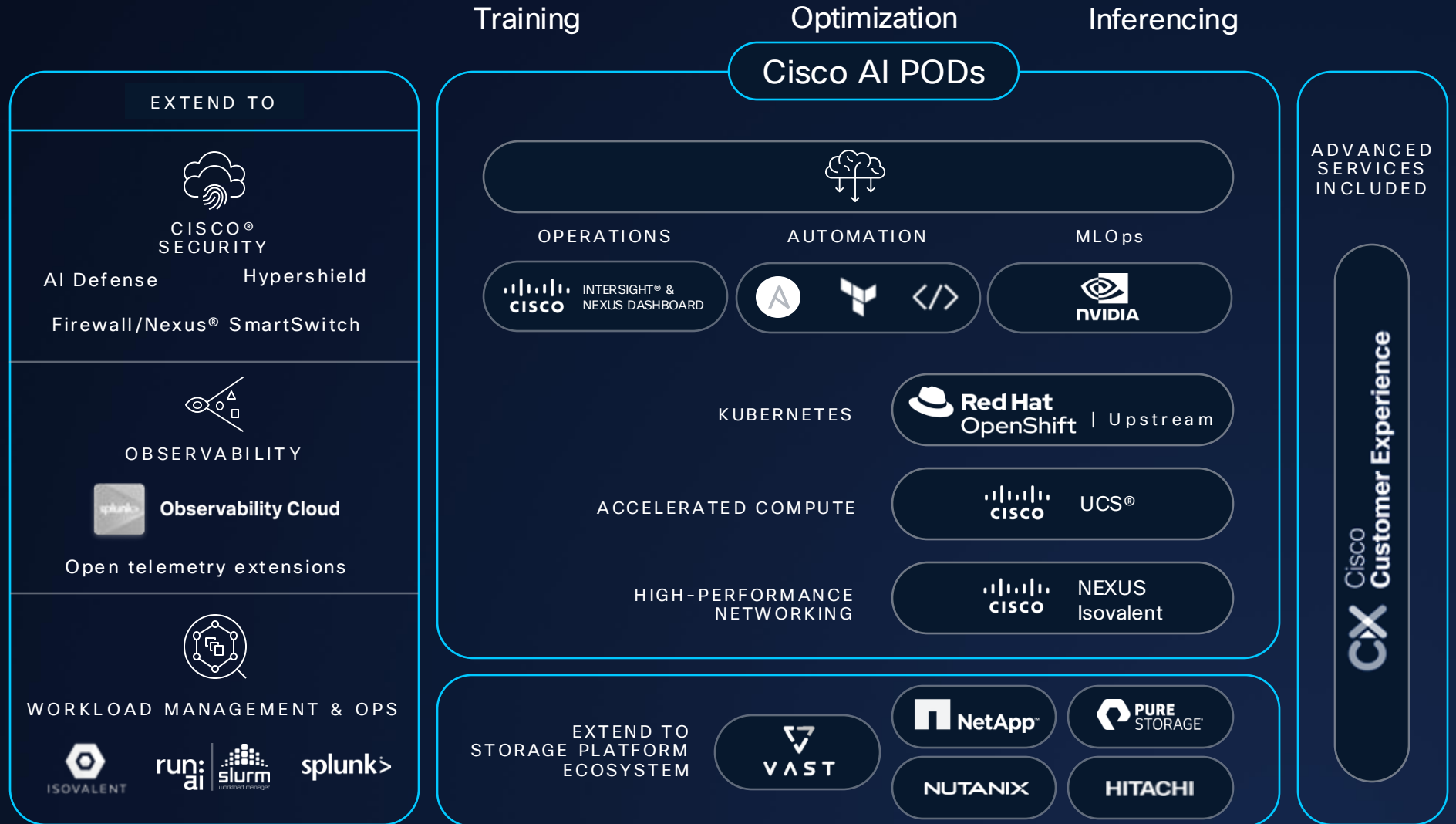
Fully supported stack including Cisco and 3<sup>rd</sup> party components

**Cisco CX Success Track**

Orderable, use case driven AI-Ready infrastructure stacks

**Inferencing.  
Optimization.  
Training.**

Incremental, atomic-level -or- fabric-based cluster scale



# Cisco Reference Architectures

## Aligned with NVIDIA Ready Architectures

### Foundational Cisco Reference Architectures

- NVIDIA AI Software
- NVIDIA certified servers (HGX, MGX, RTX Pro)
- AI Optimized Backend GPU Networking with Nexus, BlueField, & Spectrum-X
- Standard Scale Unit deployments

### Extended With

- Security & Observability solutions

AI Infrastructure with Cisco Nexus 9000 Switches  
Cisco Enterprise Reference Architecture

Updated: October 26, 2025

Table of Contents

Introduction

Featuring Cisco UCS® C885A M8 Rack Servers with NVIDIA HGX™ H200 and NVIDIA Spectrum™-X

Introduction

Cisco® Enterprise Reference Architecture (ERA) is based on Cisco Nexus™ 9000 Series Switches for networking AI clusters managed by the on-premises Cisco Nexus Dashboard platform. It adheres to the NVIDIA Enterprise Reference Architecture for NVIDIA HGX™ H200 with NVIDIA Spectrum™-X networking.

Cisco Nexus 9000 Series Switches, powered by Cisco Silicon One™ and Cisco Cloud Scale architecture, 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 Cisco NX-OS operating system, Nexus 9000 switches provide a consistent experience for frontend, storage, backend, and on-premises (OPE) management networks (see Figure 1).

Cisco Nexus Dashboard is the operations and automation platform for managing the Nexus 9000 switch-based fabric. It complements the data plane features of the Nexus 9000 switches by simplifying their configuration using built-in templates. It alerts network health issues, such as congestion, failures, and health issues, in real time and automatically fixes them as needed. 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.

Cisco Reference Architecture

link

Cisco Nexus 9000 Cloud Partner Reference Architecture

Updated: October 26, 2025

Table of Contents

Introduction

Featuring Networking Reference Architecture of Cisco UCS C885A M8 Rack Servers with NVIDIA HGX™ H200 and NVIDIA Spectrum™-X

Introduction

The Cisco Cloud Partner Reference Architecture (CPRA) is designed to be deployed with a high GPU scale, ranging from 1K to 32K GPUs, at large Cloud Service Providers (CSPs) and high-performance Super Computing Centers (SCCs) in order to solve the most computationally intensive problems without affecting user experience. The overall design supports multitenancy in order to maximize the use of deployed hardware and, if required, can be scaled to 64K GPUs. Enterprises looking to deploy AI clusters with GPU scale less than 1K should refer to Cisco Enterprise Reference Architecture (ERA) at this link.

The key technologies used in this CRA include:

- Cisco UCS® C885A Rack Servers with NVIDIA HGX™ H200 and Spectrum™-X E-Series
- Cisco® Silicon One™ and Cloud Scale NPV-based Nexus™ 9000 Series Switches combined with Cisco compute, networking, and storage controllers
- Cisco Optics and cables
- Cisco provisioning, observability and security frameworks

AI and HPC Applications

Tenant1 Tenant2 Tenant3

IoT Control Plane - Provisioning, Observability, Security

link

Cisco Nexus Hyperfabric AI Enterprise Reference Architecture  
Compliant with NVIDIA Enterprise Reference Architectures

Updated: October 23, 2025

Table of Contents

Introduction

Featuring Cisco® Hyperfabric AI Enterprise Reference Architecture compliant with NVIDIA Enterprise Reference Architectures, featuring Cisco® managed AI/ML networking of Cisco UCS® C885A M8 Rack 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-based controller. It empowers and simplifies your AI initiatives and accelerates AI deployments with a comprehensive, integrated, closed-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.

Cisco Nexus Hyperfabric AI

On-premises AI architecture

Pods of plug-and-play leaf-spine fabrics

Cisco 6000 Series Switches

link

Cisco Hyperfabric AI Cloud Partner Reference Architecture

Updated: October 26, 2025

Table of Contents

Introduction

Featuring Networking Reference Architecture of Cisco UCS C885A M8 Rack Servers with NVIDIA HGX™ H200 and NVIDIA Spectrum™-X

Introduction

The Cisco Cloud Partner Reference Architecture (CPRA) is designed to be deployed with a high GPU scale, ranging from 1K to 32K GPUs, at large Cloud Service Providers (CSPs) and high-performance Super Computing Centers (SCCs) in order to solve the most computationally intensive problems without affecting user experience. The overall design supports multitenancy in order to maximize the use of deployed hardware and, if required, can be scaled to 64K GPUs. Enterprises looking to deploy AI clusters with GPU scale less than 1K should refer to Cisco Enterprise Reference Architecture (ERA) at this link. The key technologies used in this CRA include:

- Cisco UCS® C885A M8 Rack Servers with NVIDIA HGX™ H200 and Spectrum™-X E-Series
- Cisco® Silicon One™ NPV-based Cisco Nexus Hyperfabric™ Switches combined with Cisco compute, networking, and storage controllers
- Cisco Optics and cables
- Cisco provisioning, observability and security frameworks

AI and HPC Applications

Tenant1 Tenant2 Tenant3

IoT Control Plane - Provisioning, Observability, Security

link

# Cisco CVDs for AI-Ready Infrastructure Automation

<https://www.cisco.com/c/en/us/solutions/design-zone/ai-ready-infrastructure.html>

Solutions / Cisco Validated Design Zone /


## Design guides for AI-ready infrastructure

Deliver AI-ready infrastructure everywhere—edge, cloud, data center

[Explore now](#) [Talk to an expert](#)


[Featured](#) [Communities](#) [Contact Cisco](#)

### Featured design guides




#### FlexPod for Accelerated RAG Pipeline

Deployment of Retrieval-Augmented Generation pipelines on FlexPod using NVIDIA Inference Microservices.




#### MLOps with Flashstack and Red Hat

Architecture to operationalize end-to-end AI workflow using Red Hat OpenShift AI.



#### Medical AI with Intel

Validated predictive AI with Intel AI accelerators.



#### AI/ML Networking

Build a modern, high-performance, lossless Ethernet fabric to address the stringent requirements of AI/ML workloads.

- Accelerated Deployment
- High Performance and Scalability
- Enterprise-Grade Security
- Operational Simplicity
- Reduced Deployment Risk and Optimized Performance
- Support and Reliability

# Agenda

- 01 AI Infrastructure Considerations
- 02 Cisco AI PODs
- 03 AI Workload Orchestration
- 04 Operations and Automation**
- 05 MLOps
- 06 Cisco Secure AI Factory
- 07 Tying it all Together
- 08 Key Takeaways

# Kubernetes



## Kubernetes Cluster



- Automation & orchestration – CI/CD
- Robust ecosystem of AI tooling
- GPU scheduling
- Portability
- Scalability
- Fault tolerance & reliability

# Kubernetes with Red Hat OpenShift

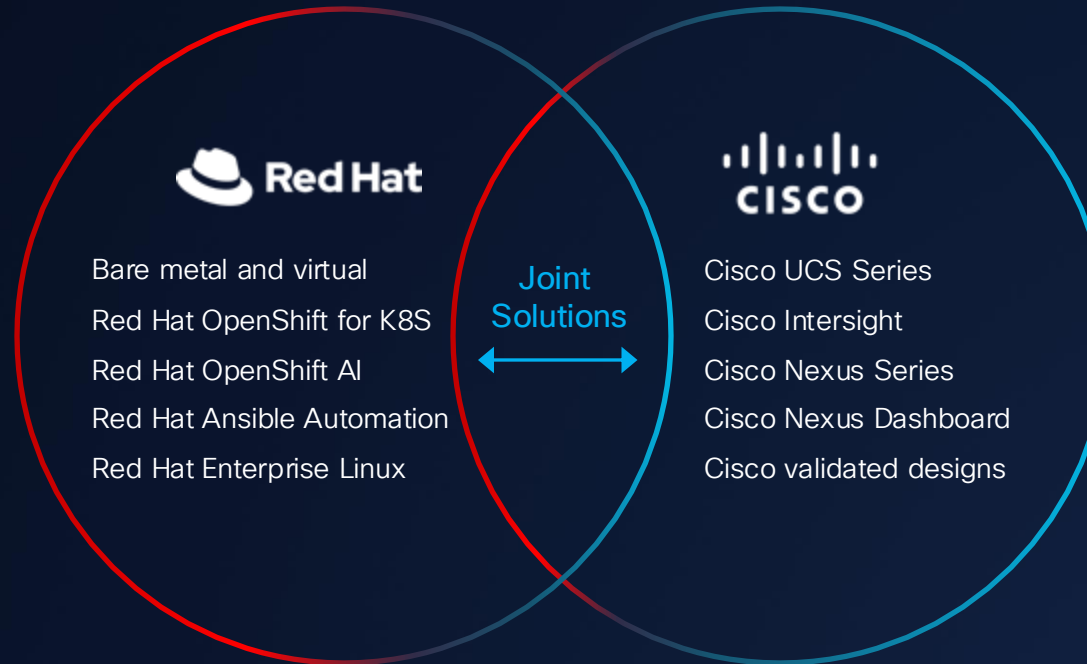
on-premises and cloud

## Open Cloud Infrastructure

platform built on open-source innovation

## Accelerated Time to value

with turnkey experience and integrated automation



## Simplified Operations and Support

with Cloud managed infrastructure and Cisco Solution Support across Red Hat on converged infrastructure stacks

## Reduced Risk

with Cisco Validated Designs, delivering tested architectures for standardized, repeatable deployments.

Operate across hybrid multicloud

More choice and flexibility

20+ Cisco Validated Designs

Consistent app dev experience

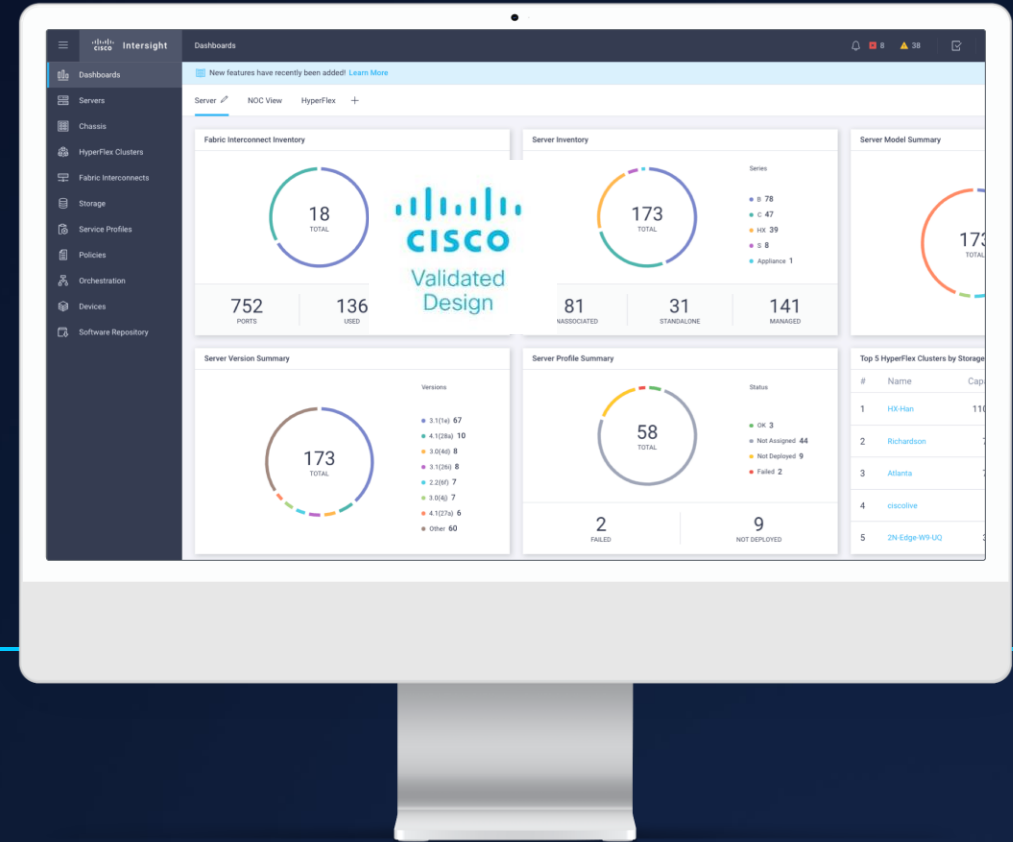
Increased sustainability

# Agenda

- 01 AI Infrastructure Considerations
- 02 Cisco AI PODs
- 03 AI Workload Orchestration
- 04 Operations and Automation**
- 05 MLOps
- 06 Cisco Secure AI Factory
- 07 Tying it all Together
- 08 Key Takeaways

# Cisco Intersight

## Unified Operating Model



## Intersight Dashboard

**Simplified** operation with AI-driven capabilities including Connected TAC, and Predictive Insight

**Automate** deployments, configuration, workflows, and day-0 to day-N tasks

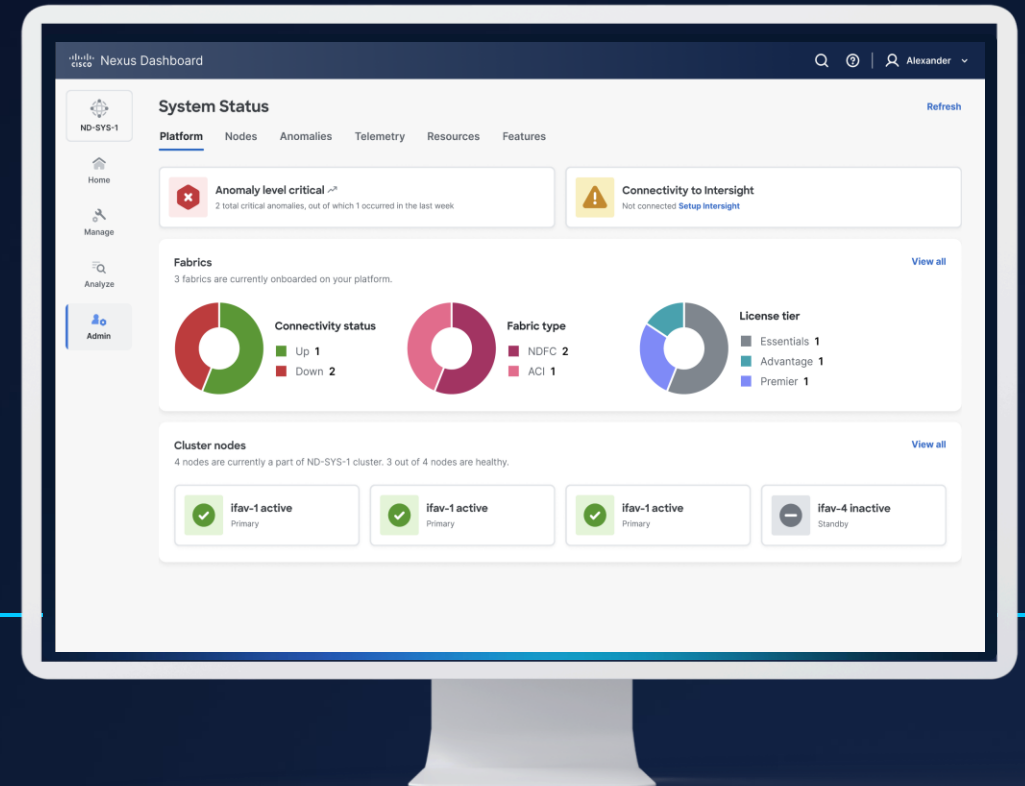
**Consistent** operational model globally, from DC to edge, at cloud scale

**Secure** operations with built-in advisories and continuous risk mitigation

# Cisco Nexus Dashboard

## Simplify Data Center Network Operations

Common policy across NX-OS and ACI fabrics



## Nexus Dashboard

**Configure, operate and analyze** your network from one place across data center networks

**Minimize** downtime through increased visibility and resolve problems with fix recommendations

**Track Power and Cooling** by surfacing your networks impact on KWh and CO2 emissions

**Accelerate** innovation with built in infrastructure-as-code and popular automation tooling integration

# Unified Operations with Cisco Nexus One

FULLY INTEGRATED STACK

## Nexus Dashboard

On premises

## Nexus Hyperfabric

Cloud management

SILICON

Cisco Silicon One  
NVIDIA Spectrum-X Ethernet

SYSTEMS

Cisco Switch  
Systems

OPTICS

Cisco Optics

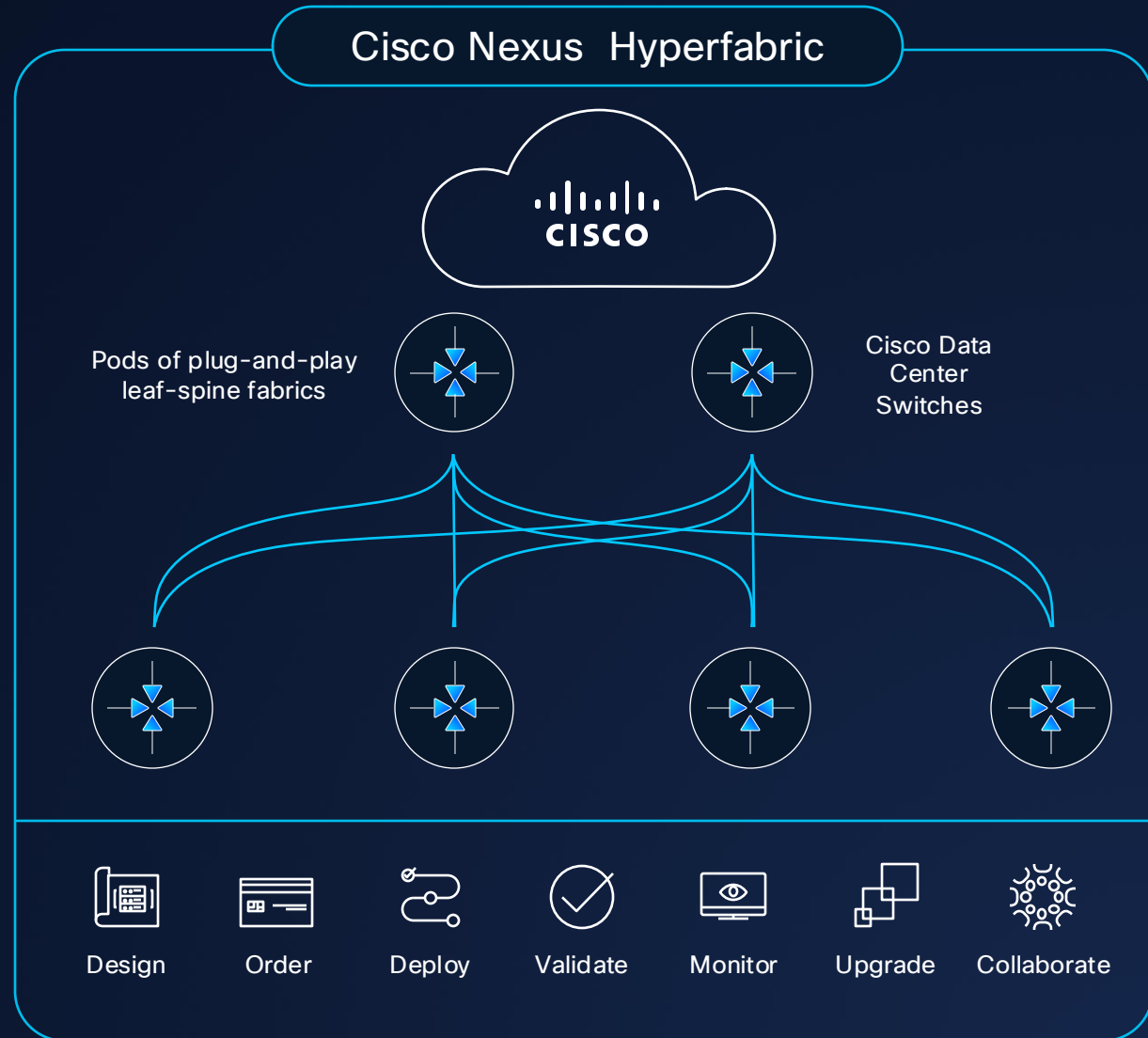
SOFTWARE

Cisco NX-OS  
Cisco ACI, SONiC

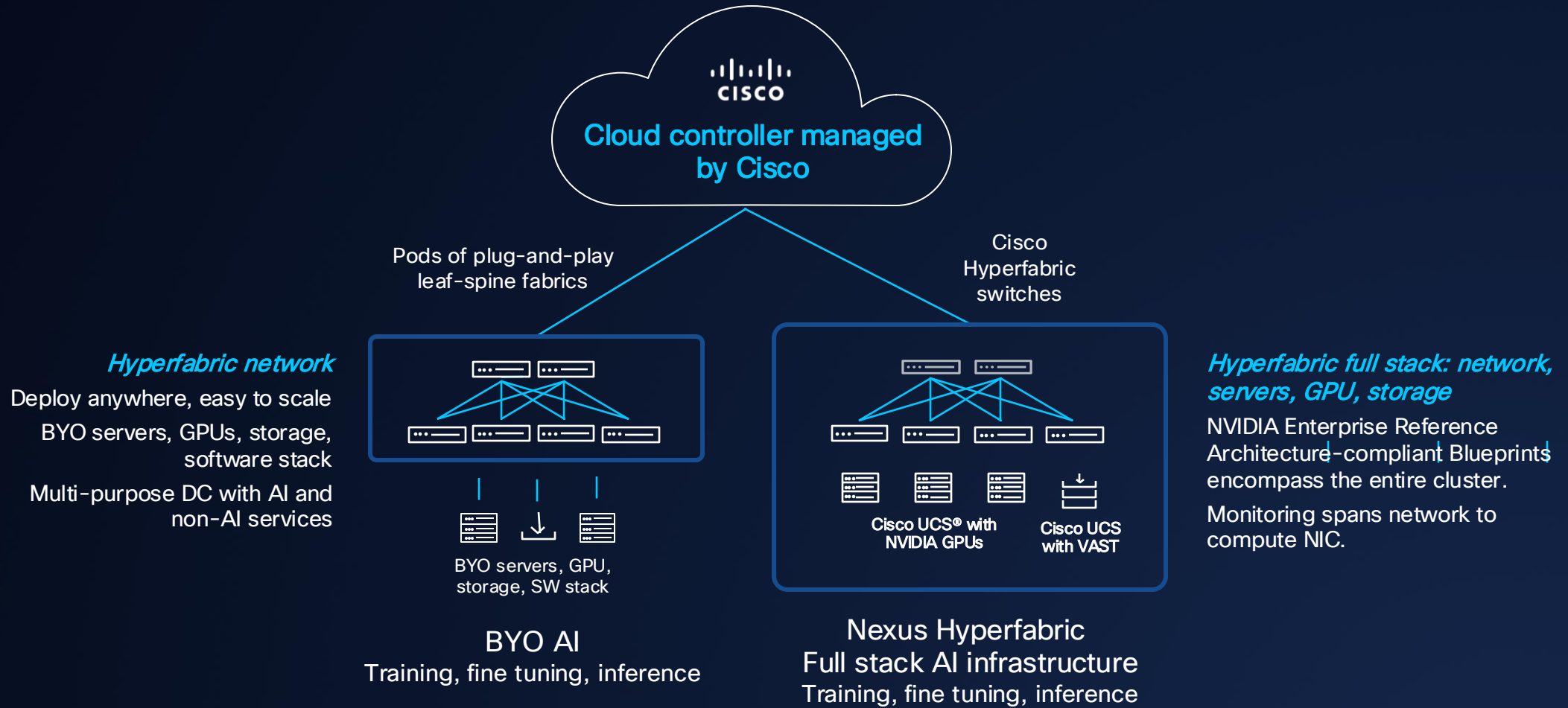
SECURITY AND OBSERVABILITY

# Cisco Nexus Hyperfabric

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



# AI Solutions with Cisco Nexus Hyperfabric



*Easy to design, deploy, operate, and scale High-performance Ethernet networks*

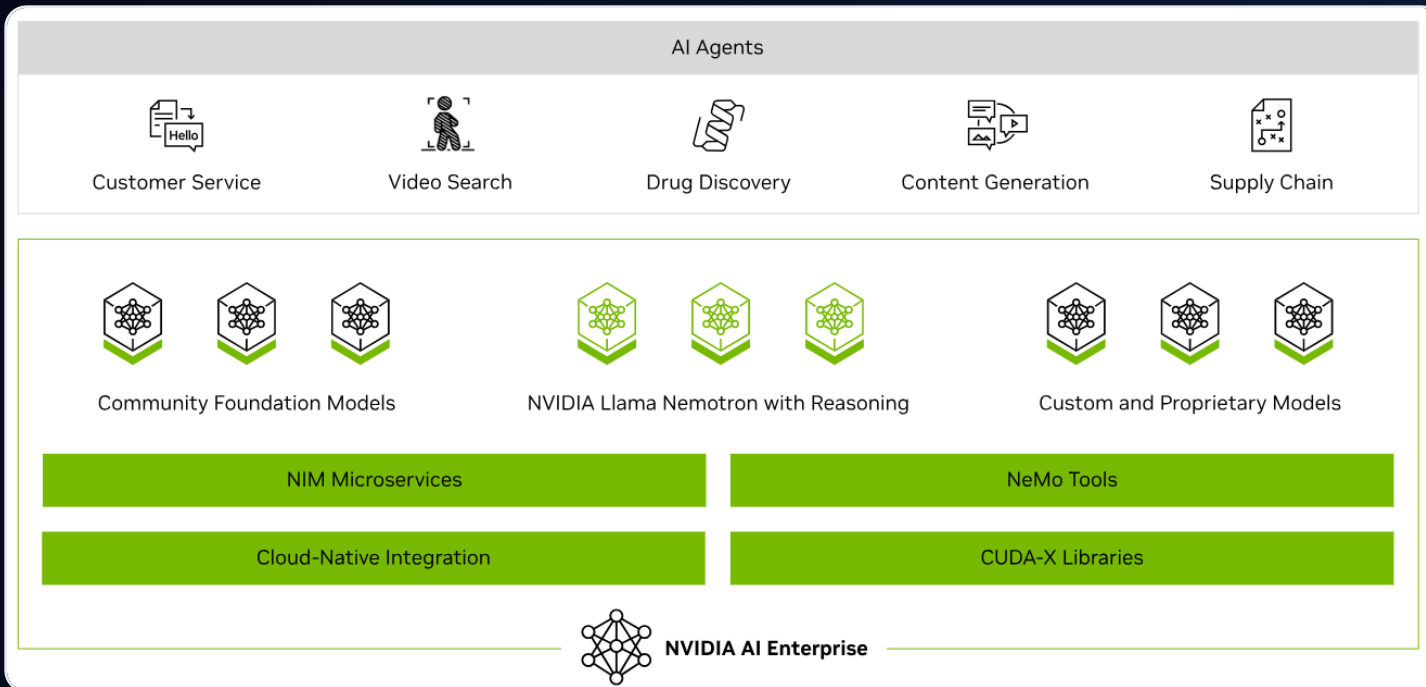
# Agenda

- 01 AI Infrastructure Considerations
- 02 Cisco AI PODs
- 03 AI Workload Orchestration
- 04 Operations and Automation
- 05 MLOps**
- 06 Cisco Secure AI Factory
- 07 Tying it all Together
- 08 Key Takeaways

# 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 workloads.

## 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  
Enterprise

NVIDIA  
Run:ai

NeMo

NIM

Blueprints

AI Workload & GPU Orchestration

# NVIDIA Run:ai

Software  
for AI



NVIDIA  
Enterprise

NVIDIA  
Run:ai

NeMo

NIM

Blueprints

AI Workload & GPU Orchestration

## Resource Management

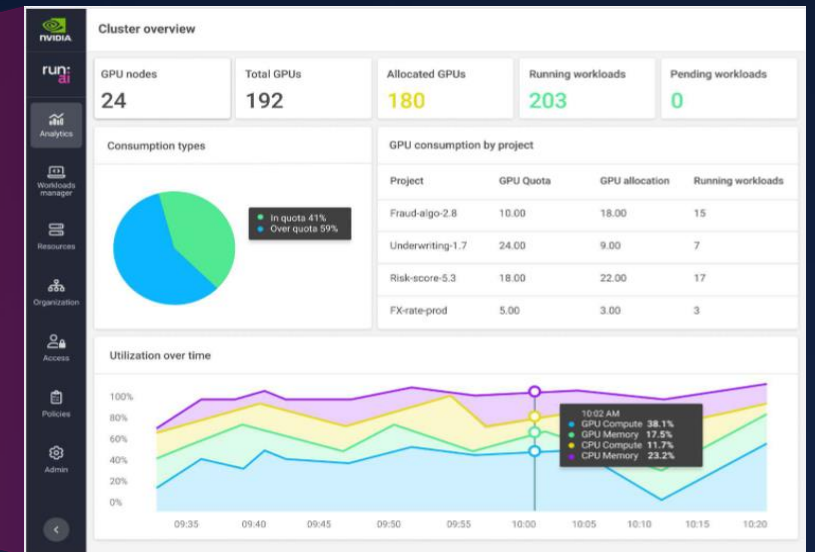
- Infrastructure Pooling
- Policy Engine

## 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.

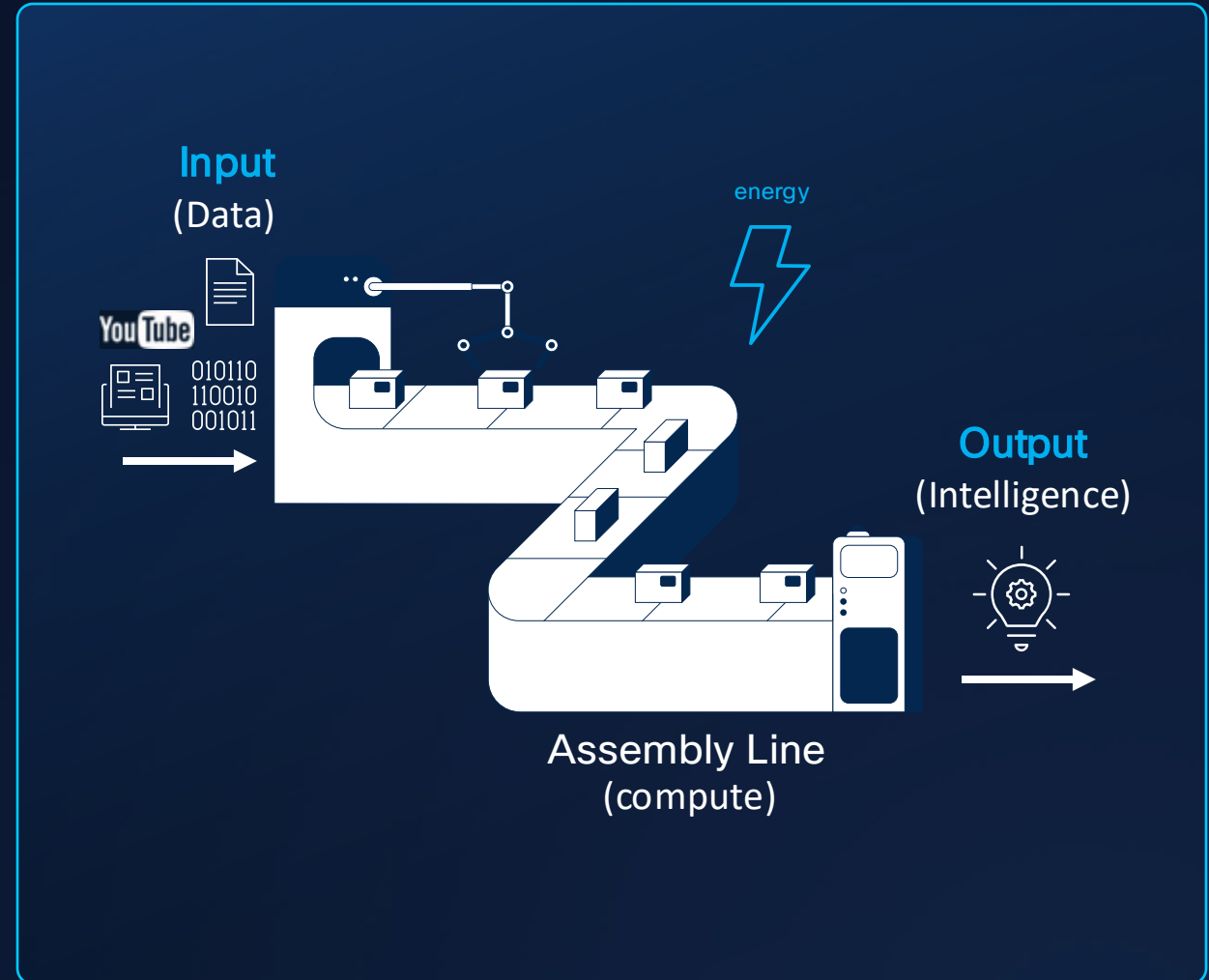
# Agenda

- 01 AI Infrastructure Considerations
- 02 Cisco AI PODs
- 03 AI Workload Orchestration
- 04 Operations and Automation
- 05 MLOps
- 06 Cisco Secure AI Factory**
- 07 Tying it all Together
- 08 Key Takeaways

# What is an AI Factory?

The processing plant for tokens

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



# Cisco Secure AI Factory With NVIDIA

ACCELERATE AI APPS DELIVERY →

Training

Optimization

Inferencing

A modular  
reference design

AI Software w/  
NVIDIA AI Enterprise



Kubernetes Platform



Cisco AI Networking



Cisco Compute w/  
NVIDIA Accelerated Computing



Partner Storage



Cisco Security



Splunk Observability



Data Center  
(Core)



Edge

# Agenda

- 01 AI Infrastructure Considerations
- 02 Cisco AI PODs
- 03 AI Workload Orchestration
- 04 Operations and Automation
- 05 MLOps
- 06 Cisco Secure AI Factory
- 07 Tying it all Together**
- 08 Key Takeaways

# Duration of Execution



# Agentic AI Use Case

## Scenario:

- Agentic workflow.
- Extremely long context.
- Latency sensitive response requirement.
- High prompt concurrency.
- Sensitive proprietary data.

## Problem:

- KV Cache is overrunning GPU compute and memory, causing inferencing crashes.
- Agentic workflows are lost mid-stream, causing expensive restarts.
- Agents drifting into problematic, highly-evasive prompting.

# Agentic AI Use Case

## Solution:

- Implement disaggregated inferencing and tiered storage class KV Cache.
- Apply guardrails.

## Requirements:

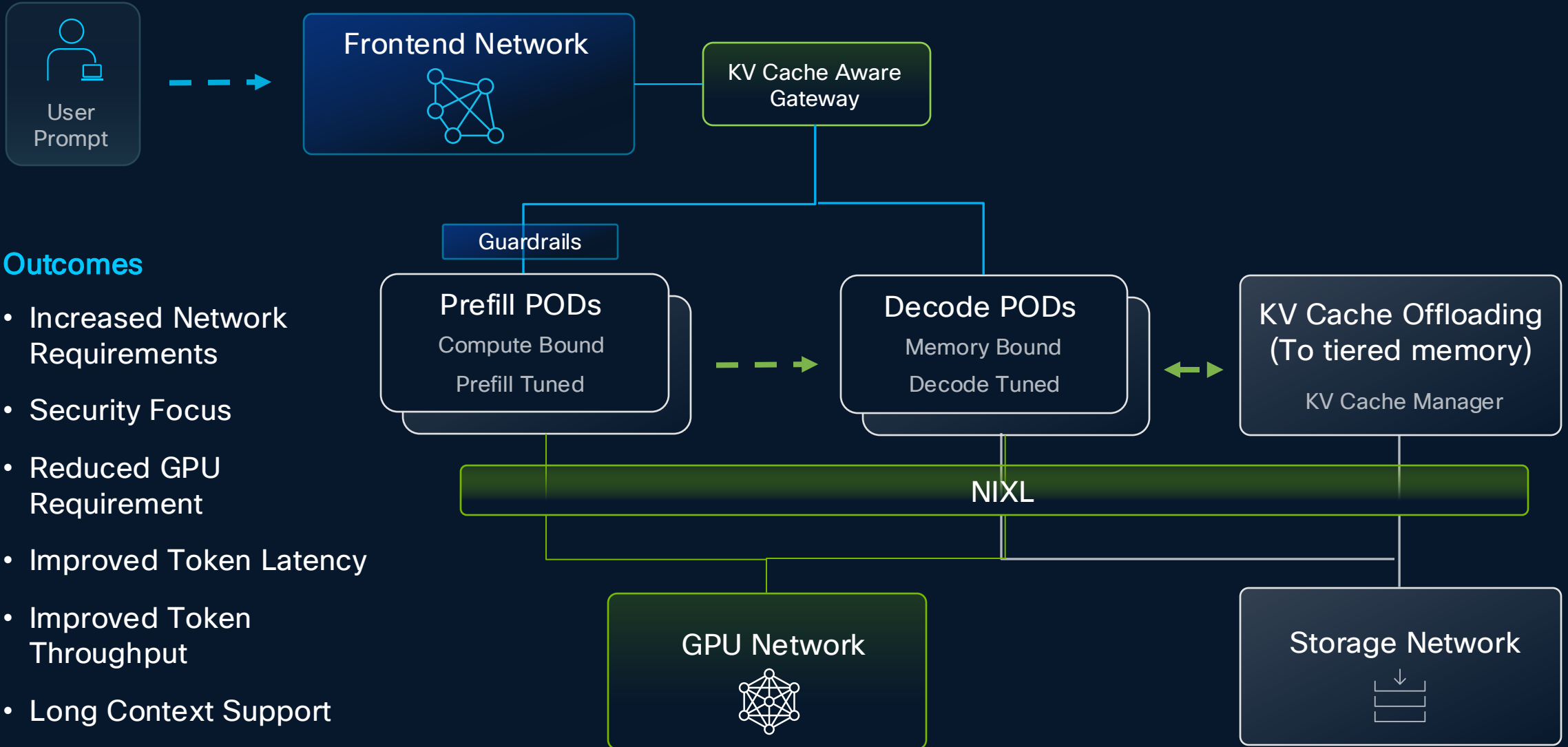
- Robust backend network implementation.
- Model serving and orchestration automation tooling.
- Multi-layer congestion and latency monitoring with autonomous re-routing
- Security in layers.

## Components:

- Cisco AI Defense
- Cisco Intersight
- Cisco Nexus Hyperfabric
- NVIDIA LLM Serving & GPU Orchestration
- Container Orchestration
- Isovalent Network Policy
- Splunk AI Observability

# Disaggregated Inference with Tiered KV Cache Offloading

With NVIDIA Inference Transfer Library (NIXL)



## Outcomes

- Increased Network Requirements
- Security Focus
- Reduced GPU Requirement
- Improved Token Latency
- Improved Token Throughput
- Long Context Support

# Agenda

- 01 AI Infrastructure Considerations
- 02 Cisco AI PODs
- 03 AI Workload Orchestration
- 04 Operations and Automation
- 05 MLOps
- 06 Cisco Secure AI Factory
- 07 Tying it all Together
- 08 Key Takeaways**

# Key Takeaways

- AI landscape continues to rapidly evolve.
- Tightly optimized network, compute, and storage is critical for AI workloads.
- From design through day-N operations; automation, design validation, and interoperable tooling are paramount.
- Extending existing infrastructure management expertise is best.
- Securing new AI-centric attack surfaces requires new approaches.
- Cisco's full-stack AI solutions enable organizations to:
  - Accelerate AI adoption with plug-and-play systems.
  - Automated lifecycle management.
  - Support any scale or use case.

**CISCO** Connect

Thank you



Making AI work for you.

