Red Hat Summit

Connect

It's time to simplify AI infrastructure and operations

Stefano Gioia, EMEA Solution Eng., Cloud & AI
Cisco

Milan, 19 Nov 2024

cisco | Red Hat

Agenda

- Cisco AI-Ready Data Center
- The value of the Cisco & Red Hat Partnership
- Call to Action

By the end of this session…

- Become familiar with the Cisco AI-Ready Data Center
- Understand the different options available within AI PODs
- Know the value of the Cisco & Red Hat Solution

Cisco & Red Hat Go-to-Market Solutions

Backed by Cisco Validated Design and Solution Support

Application Platform Modernization

VM Options

AI Ready Infrastructure

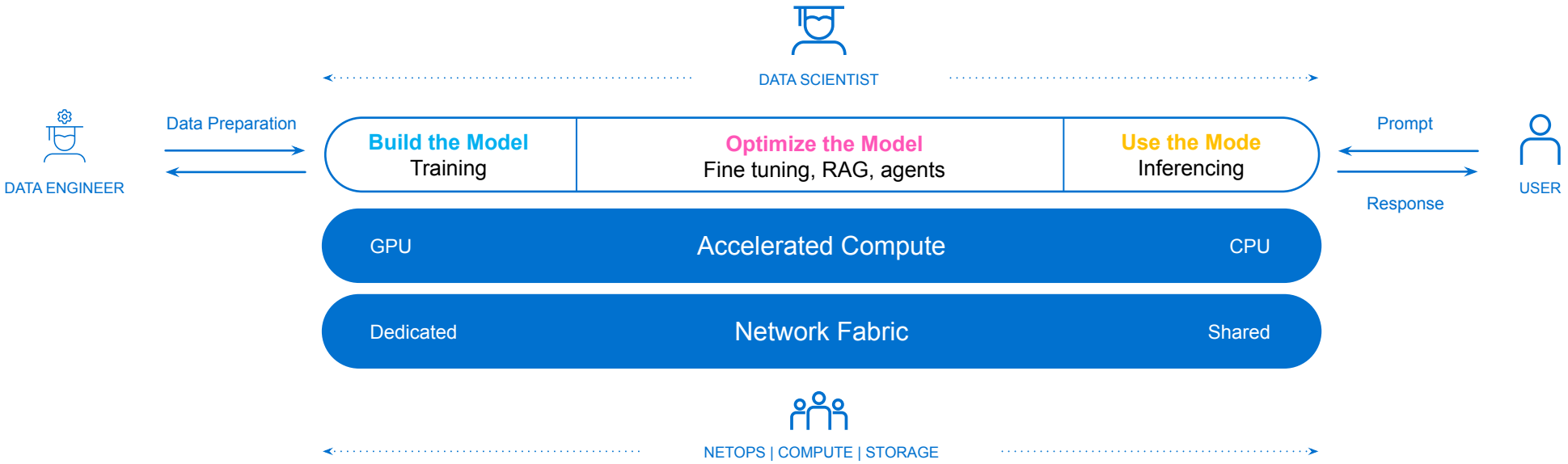
AI Ready

Edge Workload Platforms

Edge

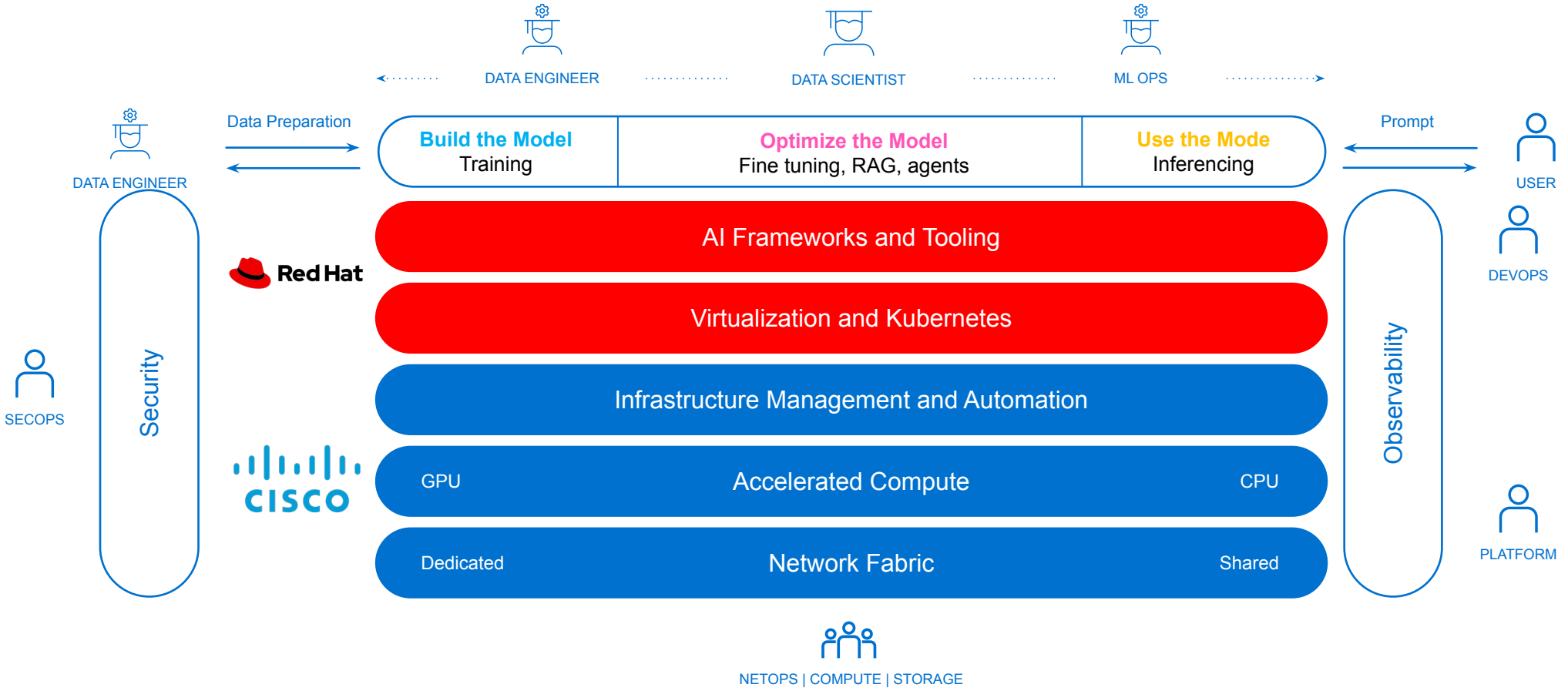
"It's time to simplify AI infrastructure and operations" – Cisco System

Generative AI – Full Stack Systems



"It's time to simplify AI infrastructure and operations" – Cisco System

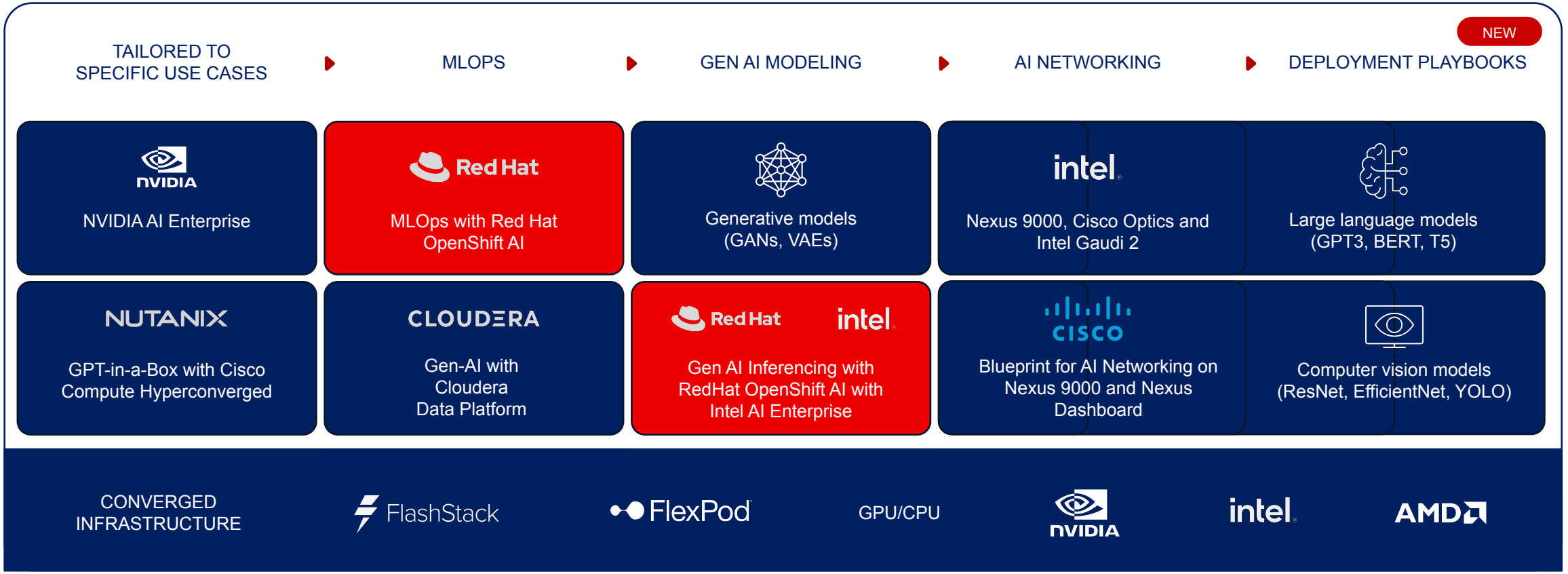
Generative AI – Full Stack Systems



"It's time to simplify AI infrastructure and operations" – Cisco System

Solutions to Simplify and Automate AI Infrastructure

Cisco Validated Designs for the Data Center



Every organization's AI approach and needs are different

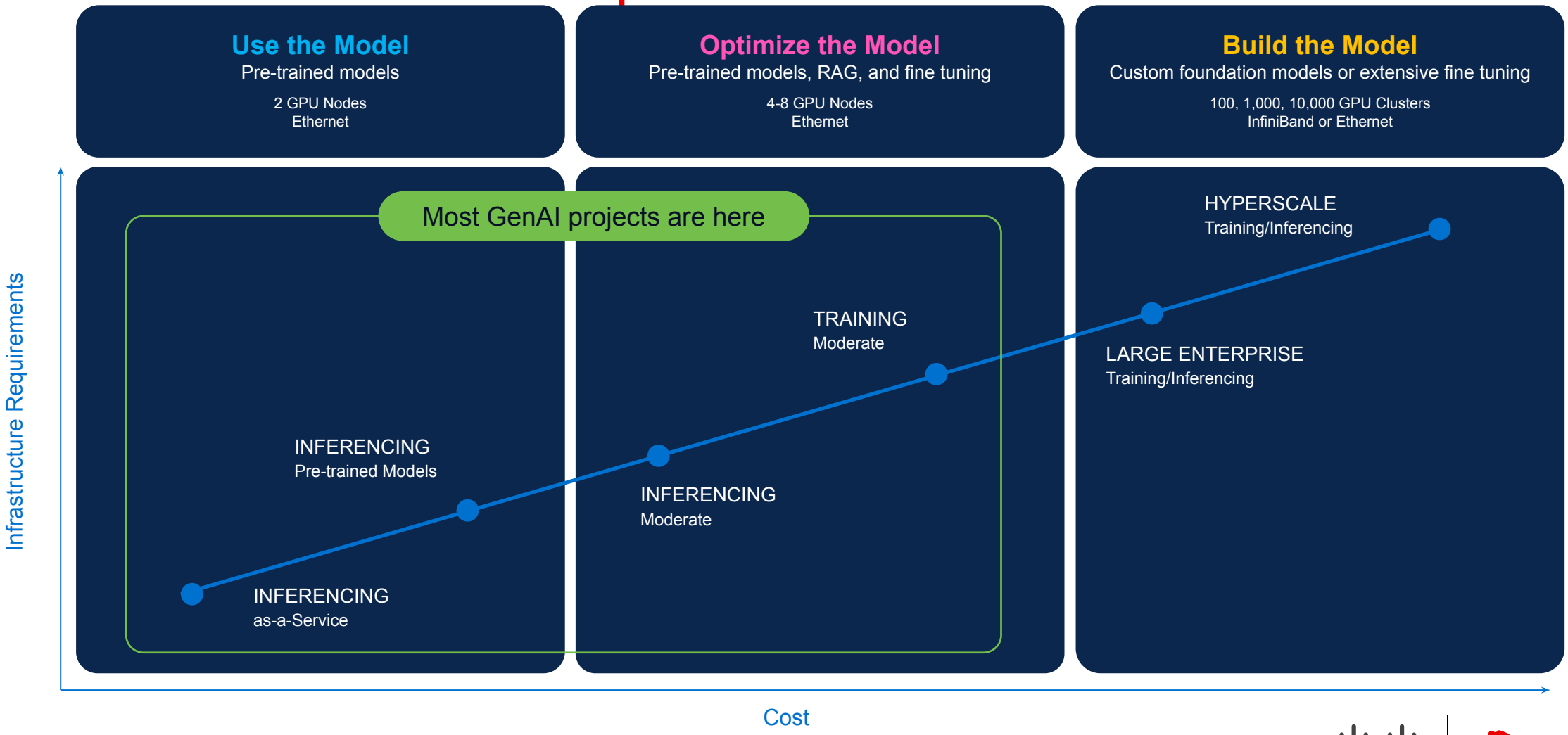
Build the model
Training

Optimize the model
Fine-tuning and RAG

Use the model
Inferencing

"It's time to simplify AI infrastructure and operations" – Cisco System

Generative AI is a Spectrum



"It's time to simplify AI infrastructure and operations" – Cisco System

Every organization's AI approach and needs are different

Build the model
Training

Optimize the model
Fine-tuning and RAG

Use the model
Inferencing

"It's time to simplify AI infrastructure and operations" – Cisco System

Introducing Cisco UCS C885A

Building high-density GPU servers to the Cisco UCS family and to Cisco's AI solution portfolio

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

UCS Accelerated

UCS C885A M8

Nvidia HGX with
8 Nvidia H100/H200 GPUs

AMD Mi300X

2 AMD 4th Gen
EPYC™ Processors

"It's time to simplify AI infrastructure and operations" – Cisco System

cisco | Red Hat

Every organization's AI approach and needs are different

Build the model
Training

Optimize the model
Fine-tuning and RAG

Use the model
Inferencing

Introducing Cisco AI PODS

Faster time to value with pre-configured bundles

Deploy AI with confidence

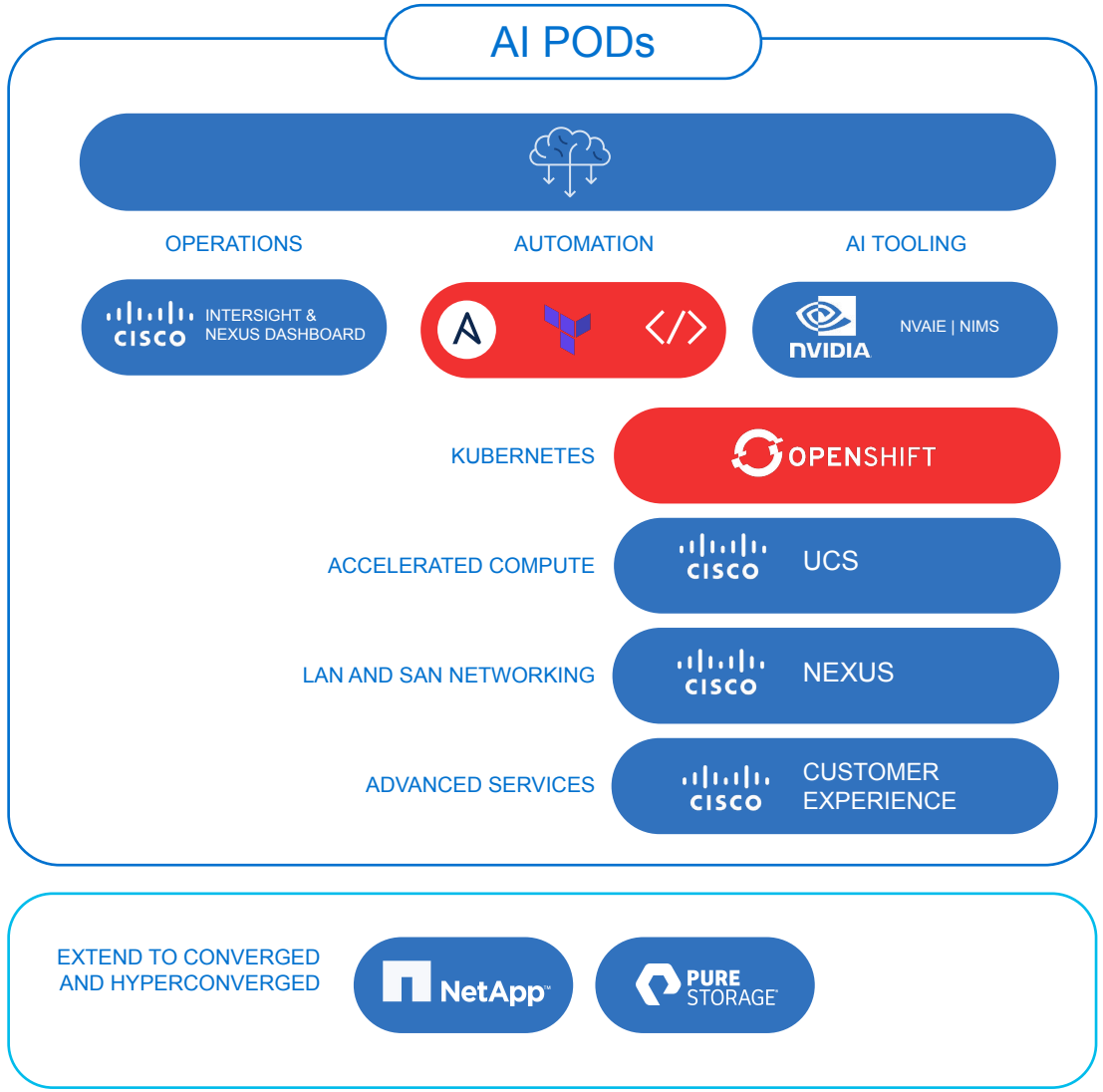
Orderable, validated AI-Ready infrastructure stacks

Fully supported stack including Cisco and 3rd party components

AI Advisor tool for configuration guidance

COMING SOON

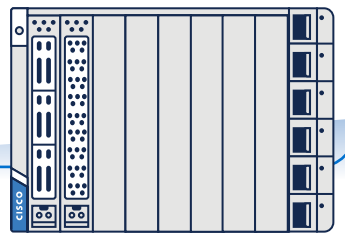
Cisco AI-Ready Infrastructure Stacks

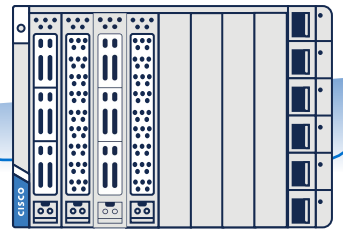


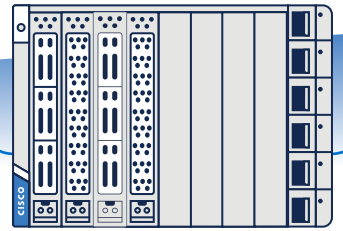
"It's time to simplify AI infrastructure and operations" – Cisco System

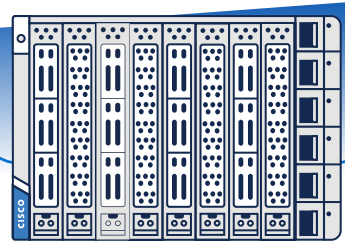
AI PODs Use Cases

	Data Center & Edge Inferencing	RAG Augmented Inferencing	Scale Up for High Performance	Scale Out for Large Deployments
Typical use case		Contextual Accuracy	Expanded Context	Multi-Model Deployments High Concurrency
Sizing Example	(Llama-2 7B GPT 2B)	(Llama-2 13B OPT 13B)	(Code Llama 34B Falcon 40B)	
Hardware specification	1x x210C compute node 2x Intel 5th Gen 6548Y+ 512 GB system memory 2x 1.6 NVMe drives 1x x440p PCIe 1x NVIDIA L40s	2x x210C compute nodes 4x Intel 5th Gen 6548Y+ 1 TB system memory 4x 1.6 NVMe drives 2x x440p PCIe 4x NVIDIA L40s	2x x210C compute nodes 4x Intel 5th Gen 6548Y+ 1 TB system memory 4x 1.6 NVMe drives 2x x440p PCIe 4x NVIDIA H100 NVL	4x x210C compute nodes 8x Intel 5th Gen 6548Y+ 4 TB system memory 8x 1.6 NVMe drives 4x x440p PCIe 8x NVIDIA L40s









Performance and Scale

"It's time to simplify AI infrastructure and operations" – Cisco System

Every organization's AI approach and needs are different

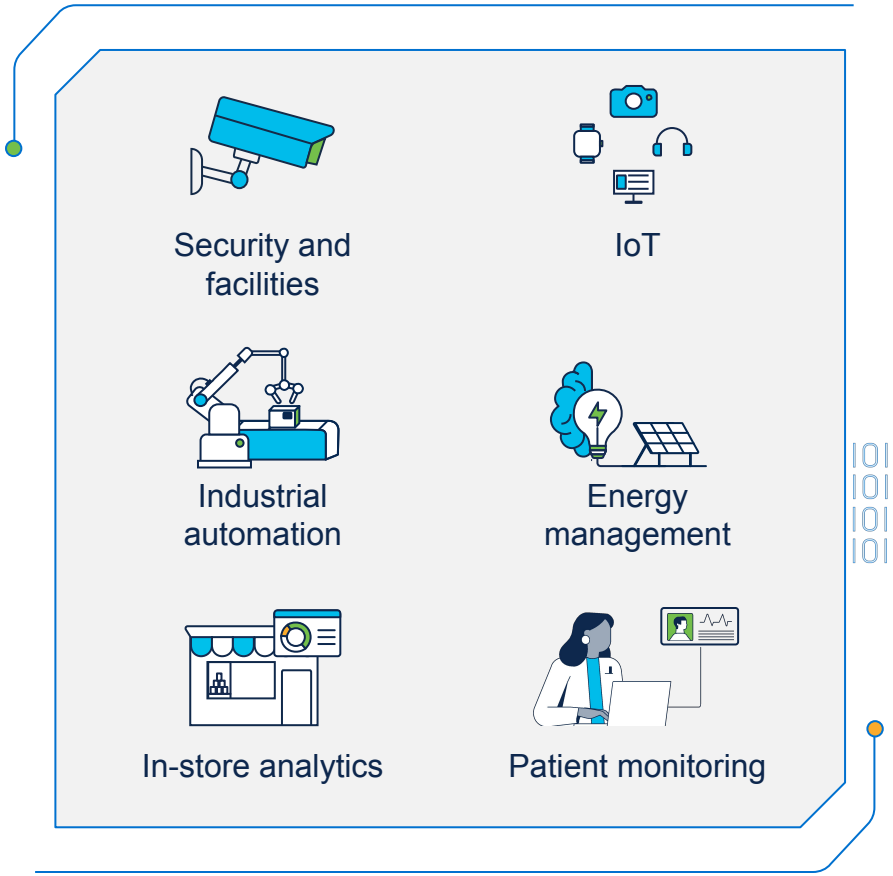
Build the model
Training

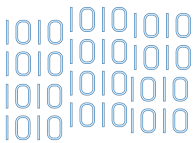
Optimize the model
Fine-tuning and RAG

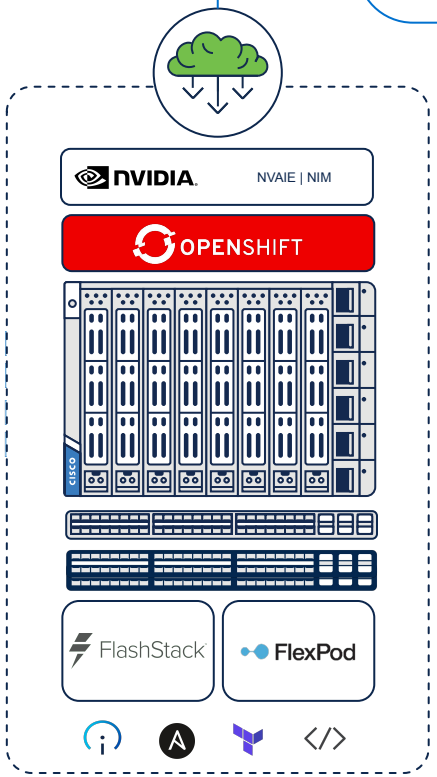
Use the model
Inferencing

"It's time to simplify AI infrastructure and operations" – Cisco System

Inferencing at the Edge







UCS X-Series Direct

UCSX Direct

- Reduce latency for real-time processing and responses
- Enhanced privacy and security by keeping data local
- Gain operational reliability even with intermittent connectivity outages
- Boost energy efficiency, especially for IoT devices

"It's time to simplify AI infrastructure and operations" – Cisco System

cisco | Red Hat

Cisco AI Pods Benefits

1 **Simplified Purchasing Experience**

2 **Seamless Deployment and Integration**

3 **Efficient and Scalable Operations**

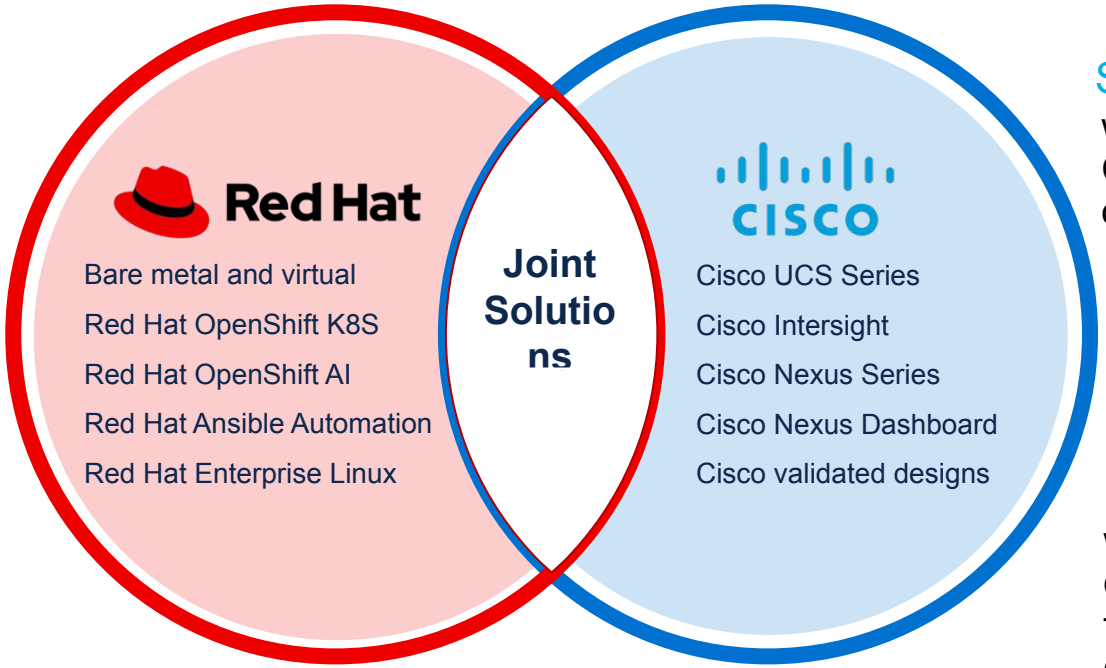
Cisco + RedHat Partnership

Open Cloud Infrastructure

platform built on open-source innovation

Accelerated time to value

with turnkey experience and integrated automation
For VMs and C



Simplified Operations and Support

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

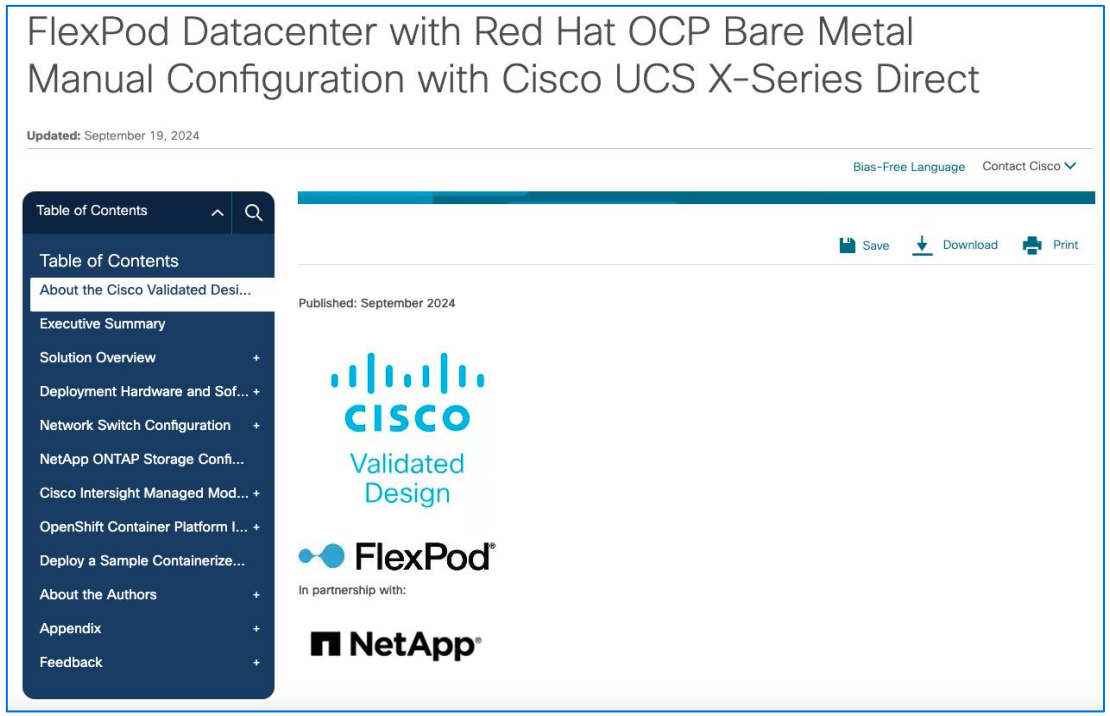
Reduced Risk

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

"It's time to simplify AI infrastructure and operations" – Cisco System

Cisco Validated Design with Red Hat



Red Hat OpenShift is now supported on bare metal UCS/UCSX servers! Cisco CVD's will no longer require VMware to remove layers and simplify our solutions.

Just released:

- Flexpod Datacenter with Red Hat OpenShift Bare Metal
- More OpenShift on bare metal CVD's coming soon:
- FlexPod with OpenShift Virtualization
- FlexPod with OpenShift AI
- FlashStack with OpenShift Virtualization
- FlashStack with OpenShift AI
- Datacenter Cisco Validated Design Center

"It's time to simplify AI infrastructure and operations" – Cisco System

Red Hat Summit

Connect

Q&A

cisco | Red Hat

Red Hat Summit

Connect

Thank you

cisco | Red Hat