

Red Hat
Summit

Connect

It's time to simplify AI infrastructure and operations

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

Rome, 7 Nov 2024



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

Helping our customer with Cisco Validated Designs that provide alternative virtualization technologies as they struggle with the current disruption in the virtualization marketplace

AI Ready Infrastructure



AI Ready

We provide AI-ready infrastructure based on Cisco Validate Designs, allowing our customers to quickly and confidently order, build, and deploy full-stack AI solutions.

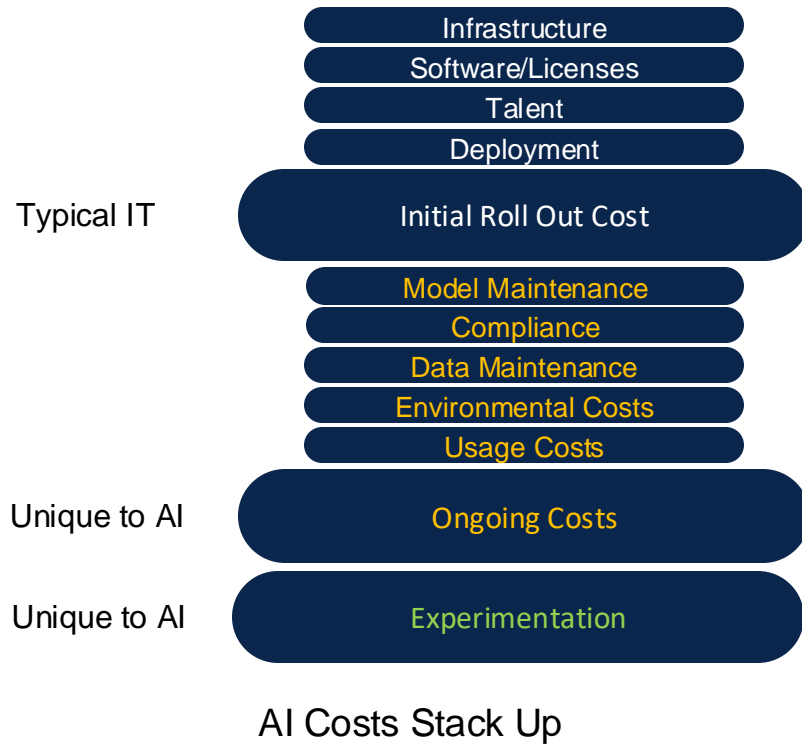
Edge Workload Platforms



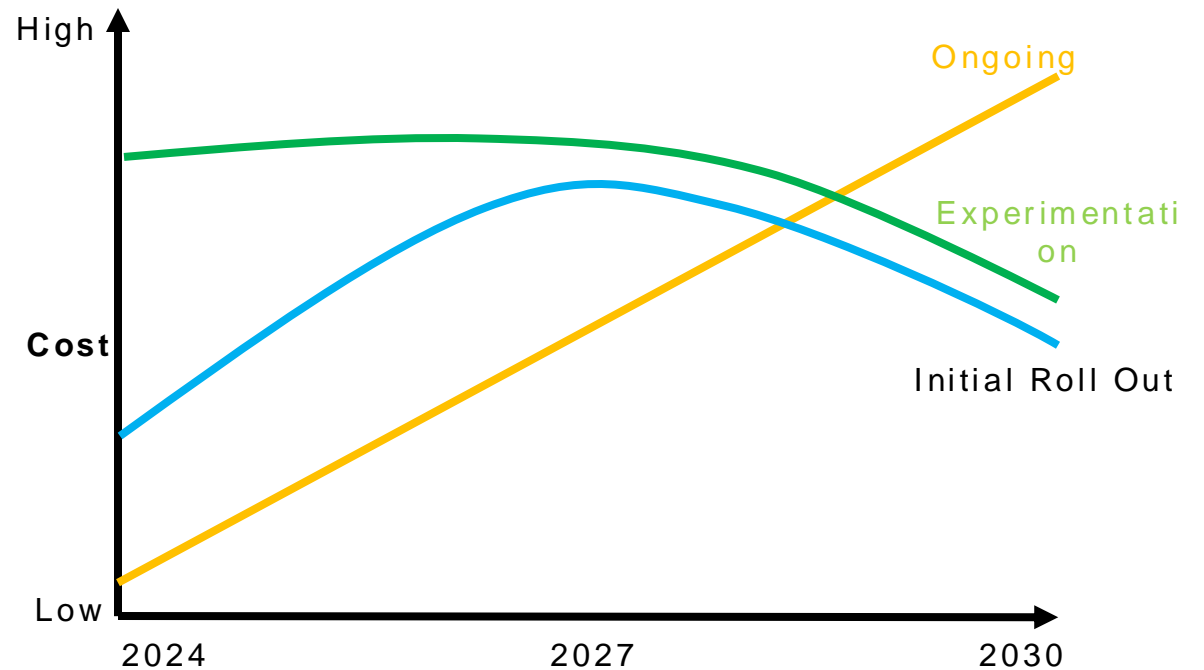
Edge

Harness the power of edge workloads with the new purpose-built edge infrastructure running the industry leading container platform. All managed at scale from the cloud

Calculating AI Cost

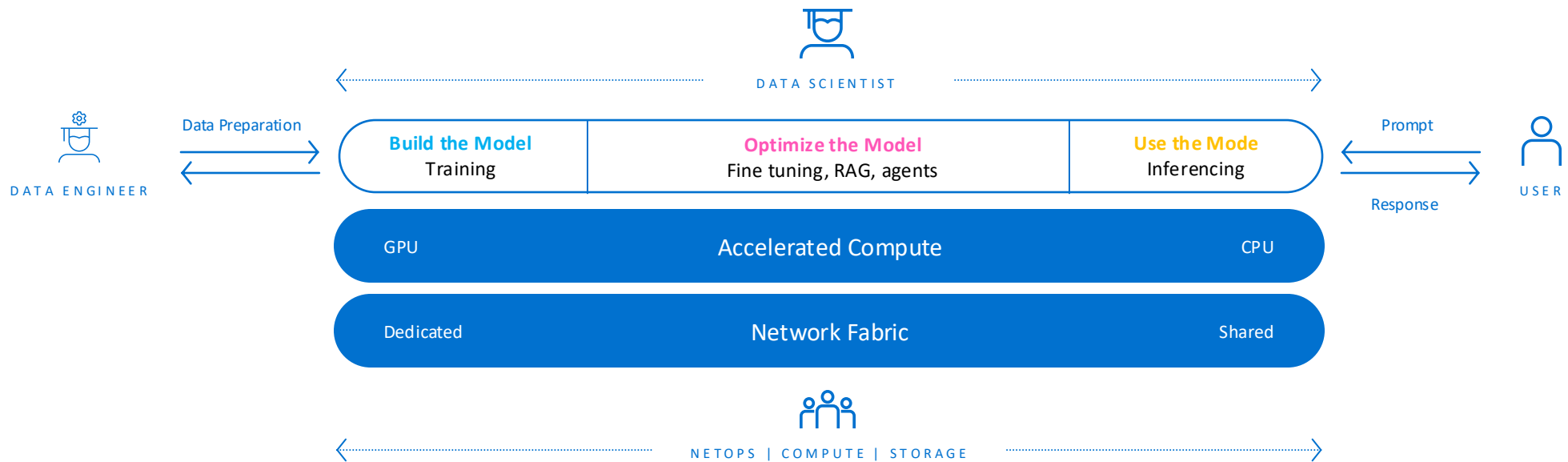


“Cost estimates can go awry by 500% - 1000%” - Gartner

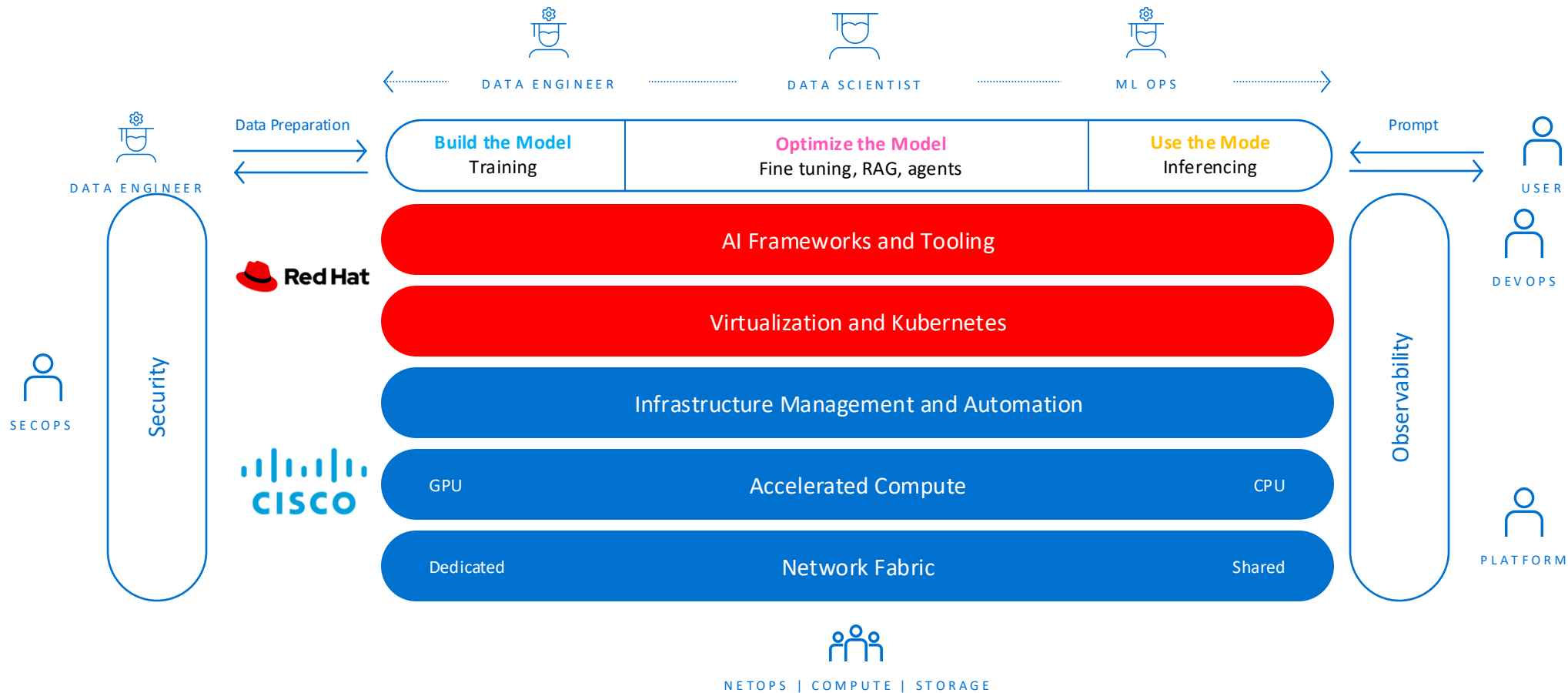


Source: AI Success depends on the CFO not IT | Gartner Finance Keynote

Generative AI – Full Stack Systems

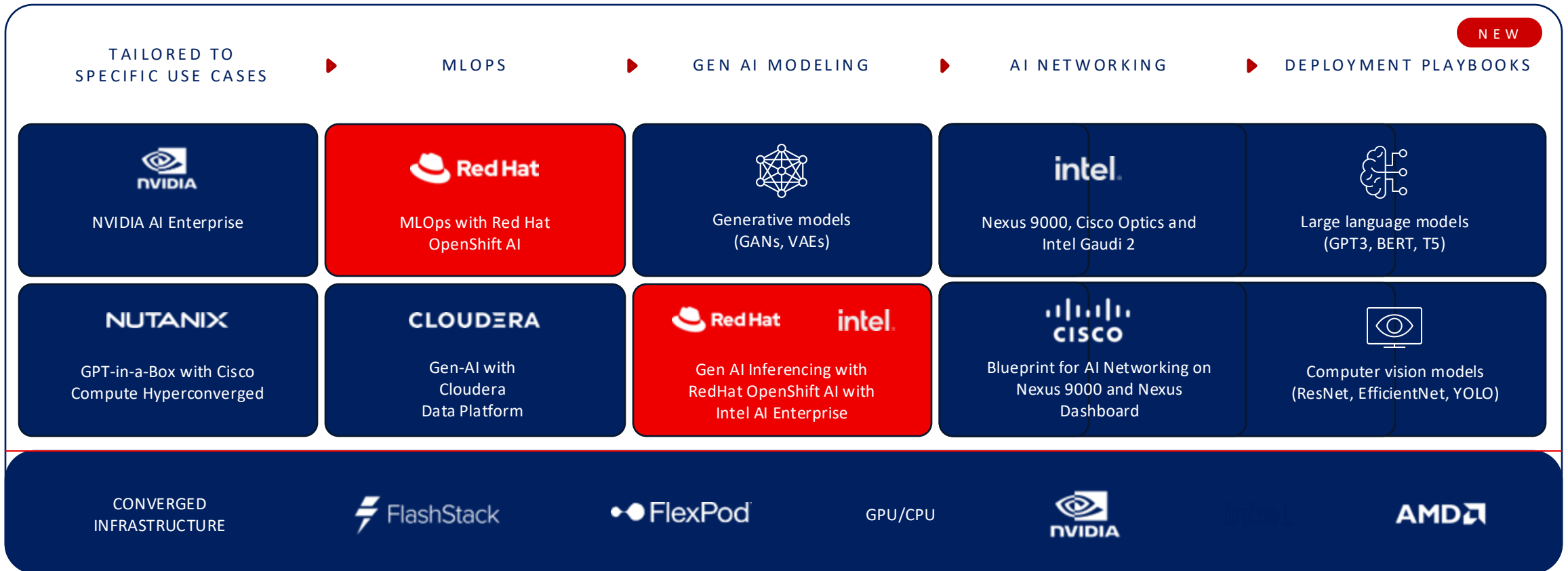


Generative AI – Full Stack Systems



Solutions to Simplify and Automate AI Infrastructure

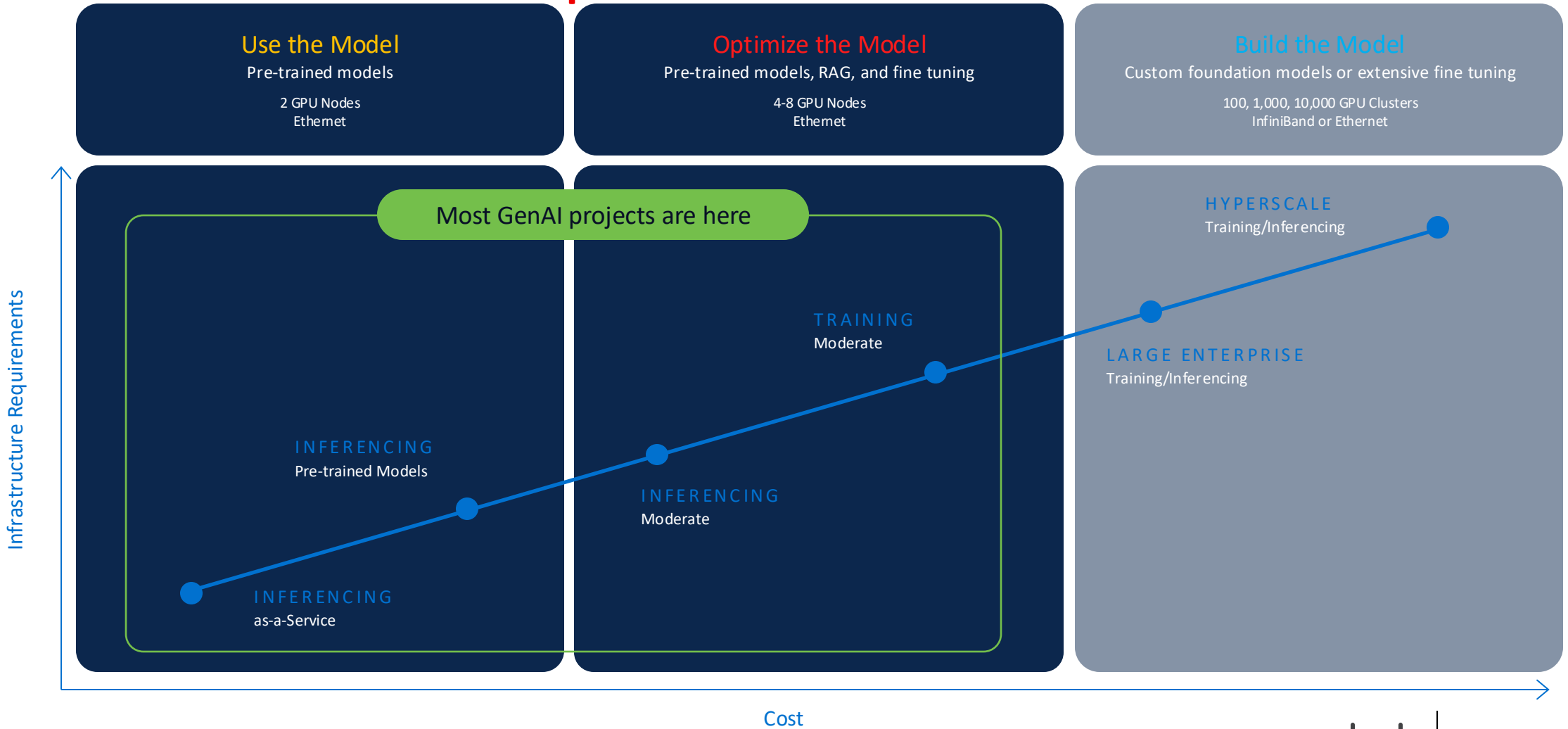
Cisco Validated Designs for the Data Center



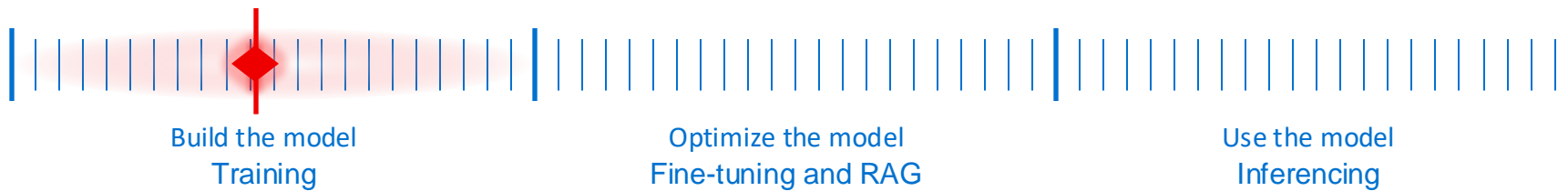
Every organization's AI approach and needs are different



Generative AI is a Spectrum



Every organization's AI approach and needs are different



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



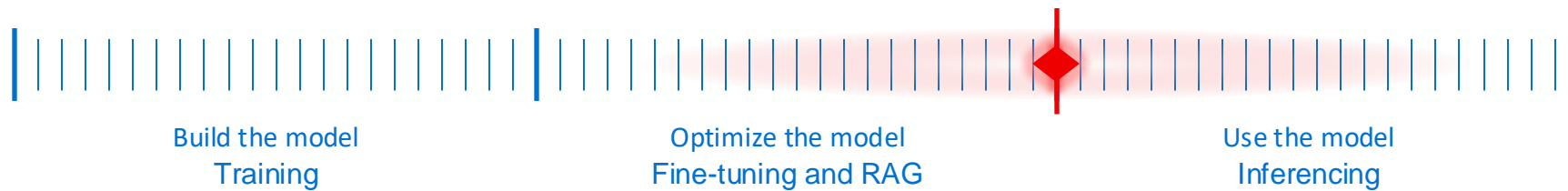
UCS Accelerated
UCS C885A M8

Nvidia HGX with
8 Nvidia H100/H200 GPUs

AMD Mi300X

2 AMD 4th Gen
EPYC™ Processors

Every organization's AI approach and needs are different



Introducing Cisco AI PODS

Faster time to value with pre-configured bundles

Deploy AI with confidence

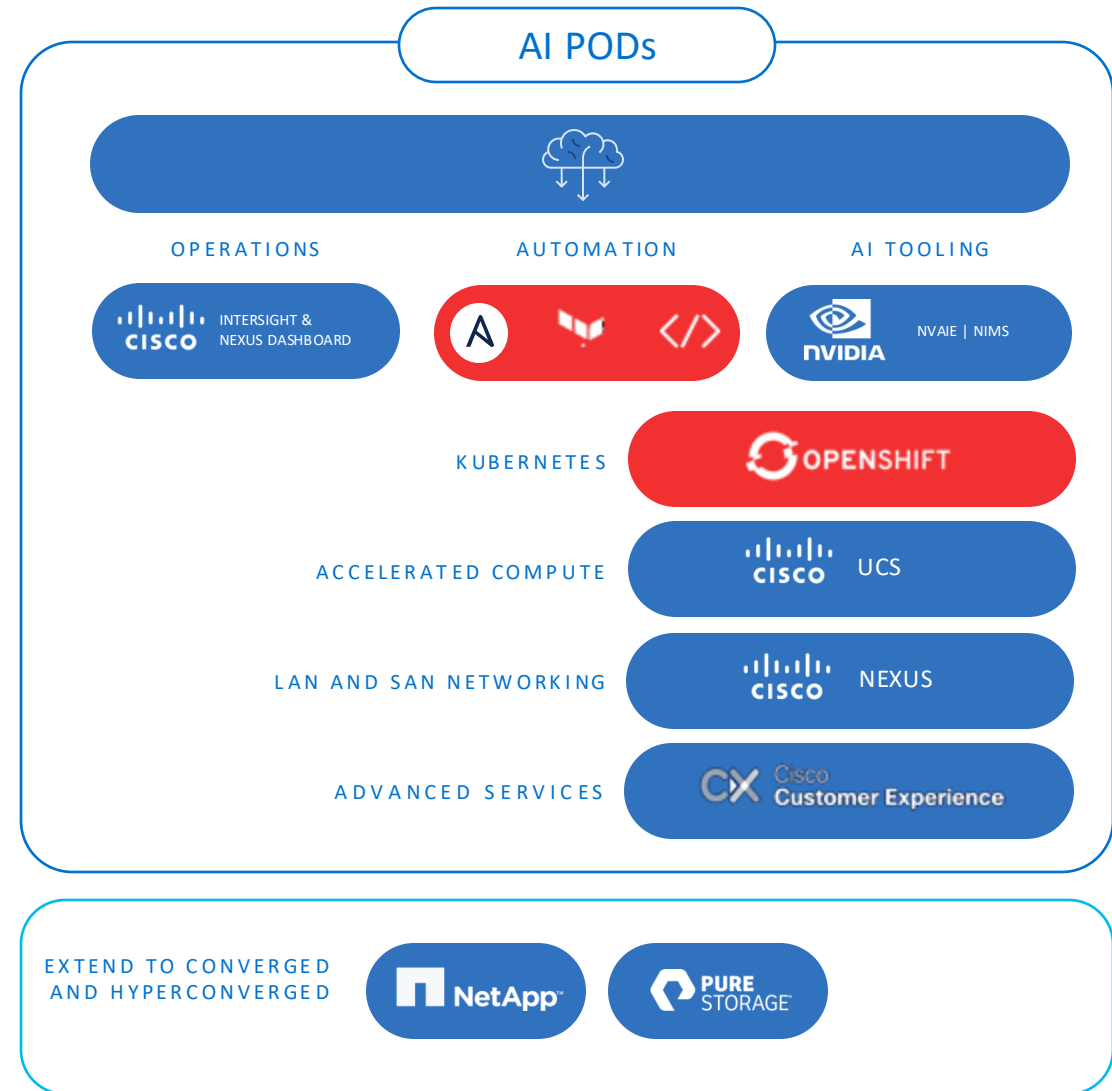
Orderable, validated AI-Ready infrastructure stacks

Fully supported stack including Cisco and 3rd party components

AI Advisor tool for configuration guidance

COMING SOON

Cisco AI-Ready Infrastructure Stacks



AI PODs T-Shirts

Typical use case

Data Center and Edge
Inferencing

RAG Augmented
Inferencing

Scale Up for High
Performance

Scale Out for
Large Deployments

Sizing example

(Llama-2 7B GPT 2B)

(Llama-2 13B OPT 13B)

(Code Llama 34B Falcon 40B)

Multi-Model Deployments
High Concurrency

PID

UCSX-AI-EDGE

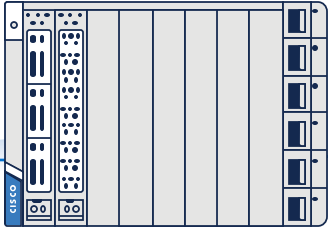
UCSX-AI-RAG

UCSX-AI-LARGERAG

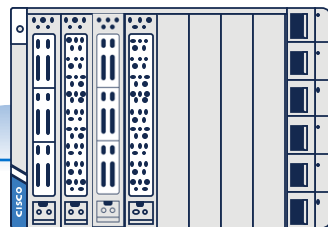
UCSX-AI-LARGEINF

POD specifications

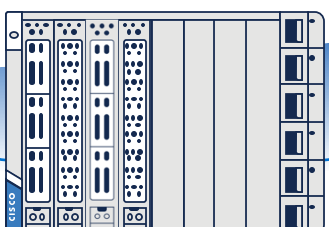
1x x210C compute node
2x Intel 5th Gen 6548Y+
512 GB system memory
2x 1.6 NVMe drives
1x x440p PCIe
1x NVIDIA L40s



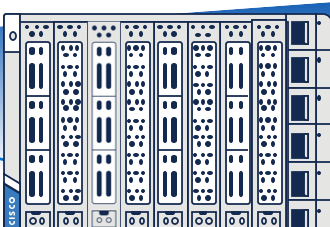
2x x210C compute nodes
4x Intel 5th Gen 6548Y+
1 TB system memory
4x 1.6 NVMe drives
2x x440p PCIe
4x NVIDIA L40s



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

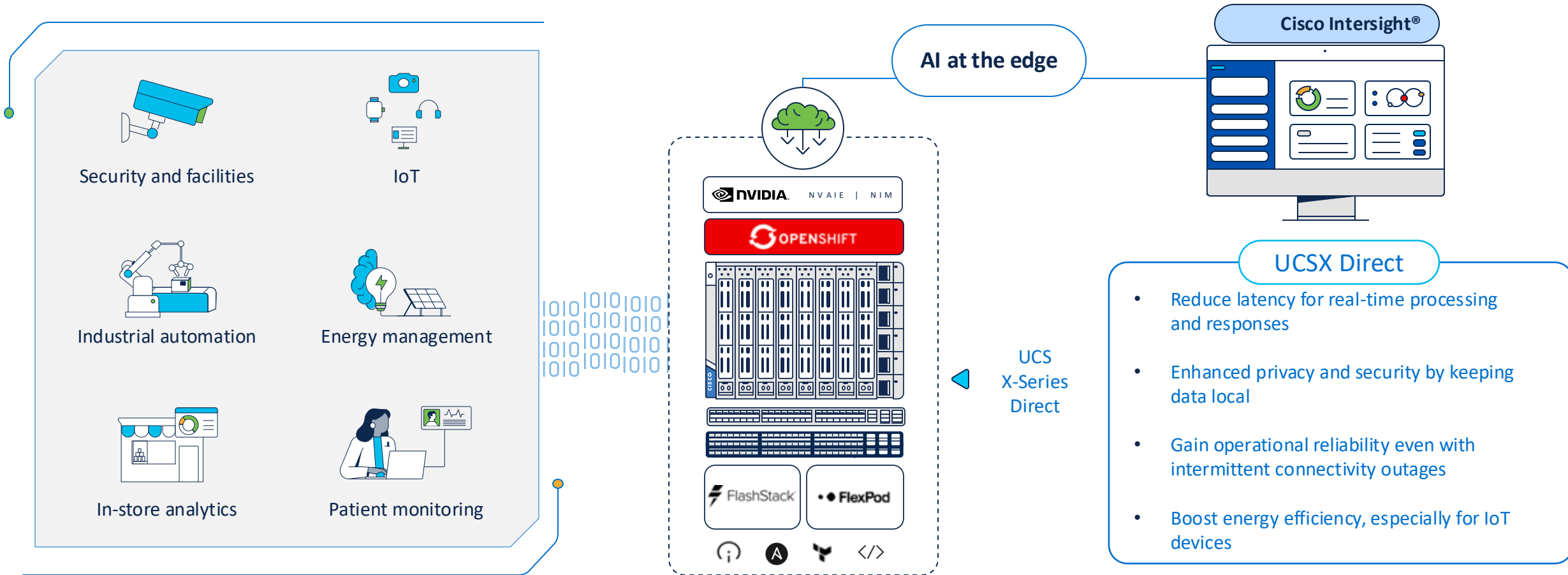


Performance and Scale

Every organization's AI approach and needs are different



Inferencing at the Edge



Cisco AI Pods Benefits

1 Simplified Purchasing Experience

Cisco AI Pods are designed to be easy to purchase for both sellers and customers. With pre-configured bundles based on validated designs, tailored to specific use cases, and available through trusted partners, the buying process is streamlined. These ready-to-deploy solutions come with clear AI infrastructure sizing and real-world examples to help customers choose the best fit for their needs.

2 Seamless Deployment and Integration

Deploying AI infrastructure is straightforward with Cisco AI Pods, thanks to their compatibility with existing storage, networking, and management systems. Pre-configured AI software, automated deployment tools, policy-based orchestration, and strong security measures ensure that AI solutions are deployed quickly and securely. Partnerships with ecosystem players further enhance deployment flexibility and support.

3 Efficient and Scalable Operations

Cisco AI Pods make ongoing operations easy by automating resource management and providing comprehensive monitoring and alert systems. These solutions are designed to integrate smoothly with existing IT environments, support DevOps and MLOps practices, and offer robust backup and recovery options. This approach ensures high performance, scalability, and cost-effective AI operations from initial deployment through to scaling and optimization.

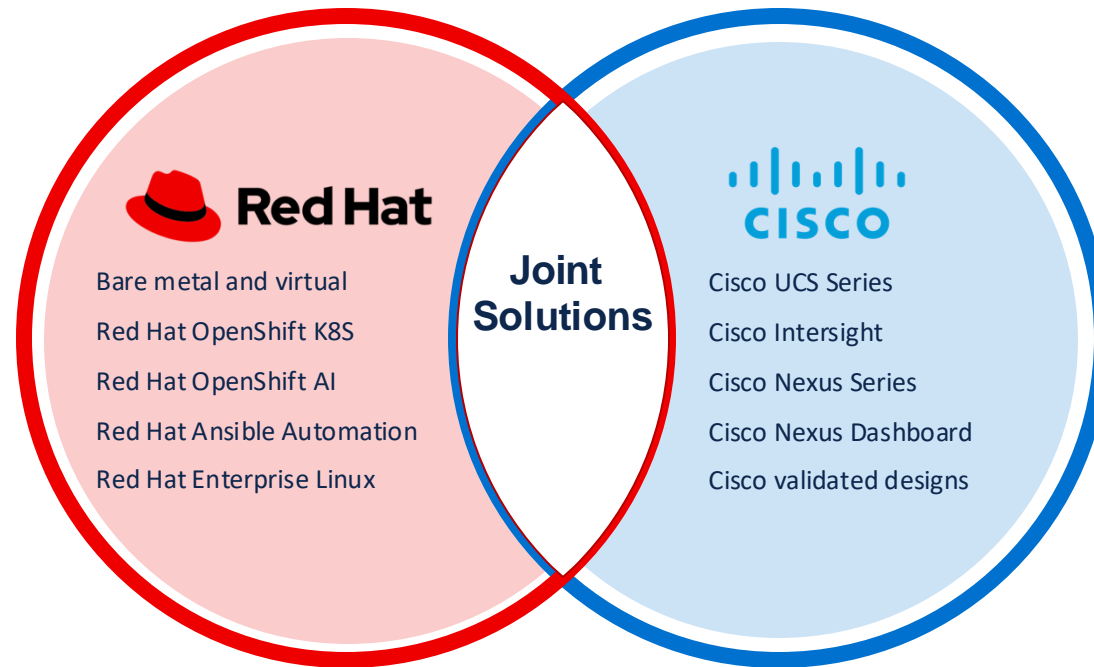
Cisco + RedHat Partnership

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

Cisco Validated Design with Red Hat



FlexPod Datacenter with Red Hat OCP Bare Metal Manual Configuration with Cisco UCS X-Series Direct

Updated: September 19, 2024

Bias-Free Language Contact Cisco

- Table of Contents
- Table of Contents
- About the Cisco Validated Desi...
- Executive Summary
- Solution Overview
- Deployment Hardware and Sof...
- Network Switch Configuration
- NetApp ONTAP Storage Conf...
- Cisco Intersight Managed Mod...
- OpenShift Container Platform I...
- Deploy a Sample Containerize...
- About the Authors
- Appendix
- Feedback

Published: September 2024



In partnership with:



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](#)





Connect

Q&A





Connect

Thank you

