GO BEYOND

#CiscoLiveAPJC

UCS X-Series: Blurring the Line Between Rack and Blade for Modern Applications

Ravi Mishra
Director, UCS Product Management
@ravmishr

BRKCOM-3618

Cisco Webex App

Questions?

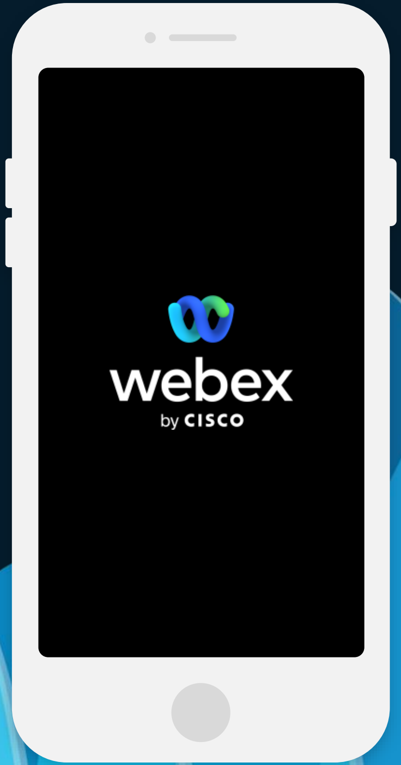
Use Cisco Webex App to chat
with the speaker after the session

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 November 15, 2024.

https://ciscolive.ciscoevents.com/
ciscolivebot/#BRKCOM-3618



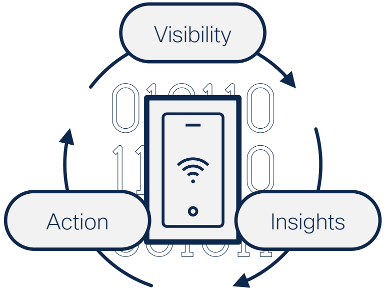
Agenda

- Challenges
- Solution
- Deep dive
- Application Examples

Cisco UCS

Architectural foundation for next gen Infrastructure

Future-proof platform architecture

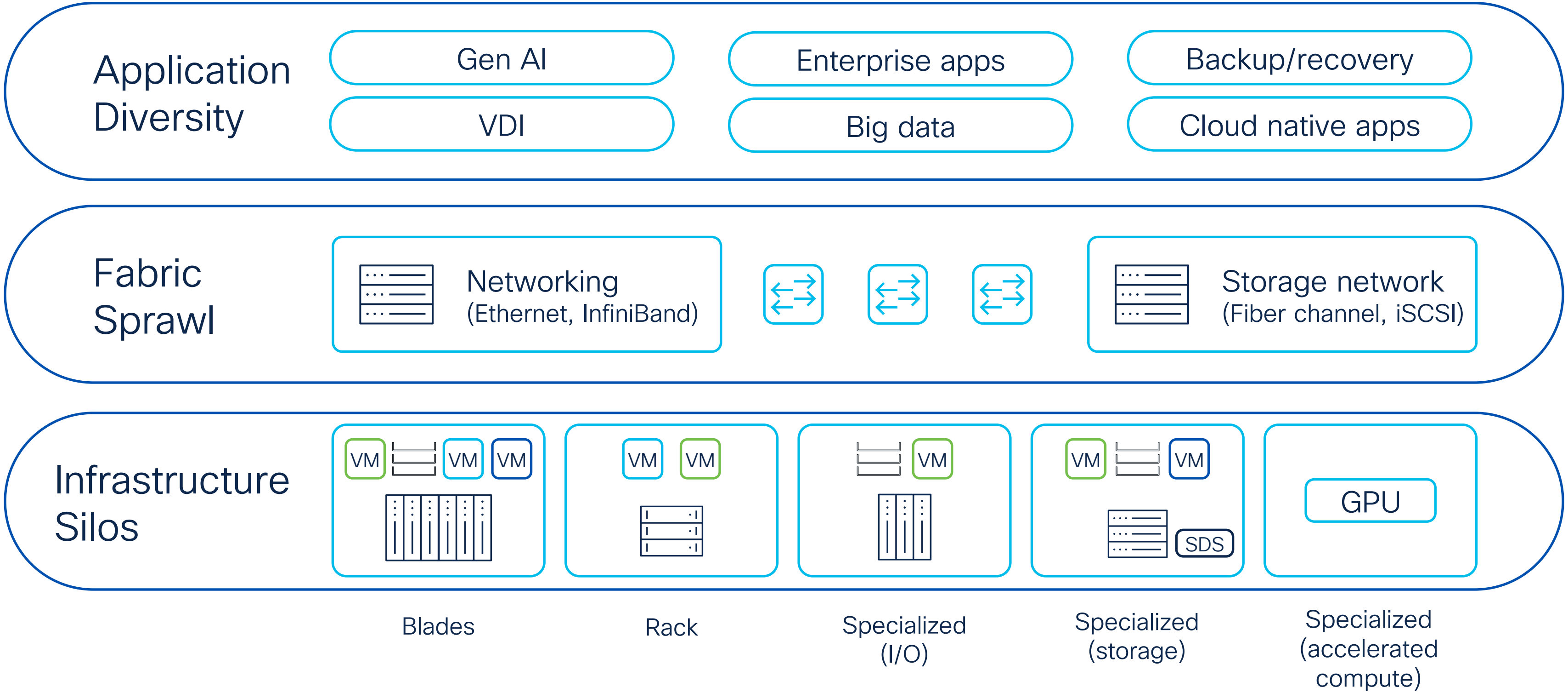


Cloud operating model

Optimized for traditional and AI Apps

AI Ready | Simple | Sustainable | Secure | Automation Centric

Architectural Silos Drive complexity



Application Diversity

Gen AI

VDI

Enterprise apps

Big data

Backup/recovery

Cloud native apps

Fabric Sprawl

Networking (Ethernet, InfiniBand)

Storage network (Fiber channel, iSCSI)

Infrastructure Silos

Blades

Rack

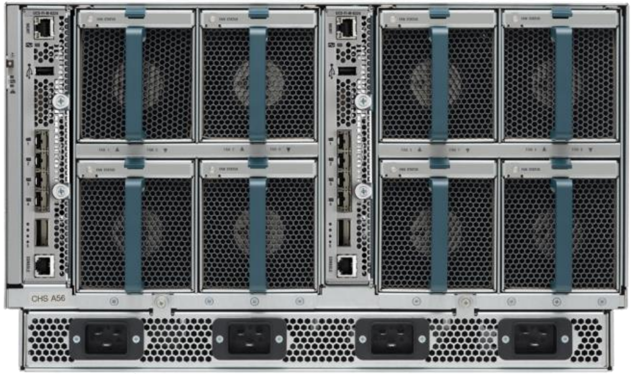
Specialized (I/O)

Specialized (storage)

Specialized (accelerated compute)

Cisco UCS: 5108 Blade Chassis

Celebrating 15+ years in service: First Customer Ship: June 2009



Building Block for Scalable and Flexible Datacenter Architectures

- Compact 6RU mount into industry standard 19-inch racks
- Up to 8 2S half-width blades or 4 4S full-width blades
- Standard front-to-back cooling
- No Chassis Management or blade switching
- Multi Generational Stateless Compute: M1 to M6 Compute Platforms
- Cisco Single Connect with Cisco VICs unified LAN, SAN and Management into one link

Industry leading platform for Converged Infrastructure

Workload Agnostic: Bare-Metal or Virtualized

Lower TCO, Reduced Cabling and Simplified Management

UCS: 2 Socket Blade servers

Modularity in compute with multiple generations

B200 M1/M2

Up to two Intel Xeon 5600

12 DIMMs, up to 192GB

Up to 20 Gbps I/O

Up to 600GB Internal Storage

B200 M3

Up to two Intel Xeon E5-2600 v3 (Max: 24 Cores Total)

24 DIMMs, up to 768 GB

Up to 80 Gbps I/O

Up to 2TB Internal Storage

B200 M4

Up to two Intel Xeon E5-2600 v3/v4 (Max: 44 Cores Total)

24 DIMMs, up to 1.5TB

Up to 80 Gbps I/O

Up to 3.2TB Internal Storage

Up to 1 GPU NVIDIA M6

Up to 2 SD Card

B200 M5

Up to two Intel Xeon Processor Scalable Family (Max: 56 Cores Total)

24 DIMMs, up to 9TB

Up to 80 Gbps I/O

Up to 15TB Internal Storage

Up to 2 GPU NVIDIA P6

Up to 2 SD/M.2 Cards

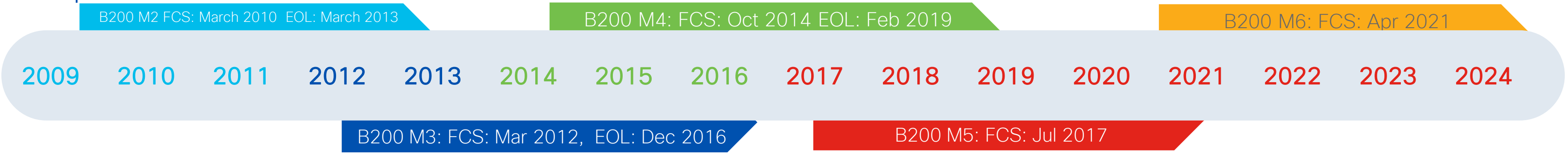
B200 M6

Up to two Intel Whitley Ice Lake CPUs (Max: 80 Cores Total)

32 DIMMs, up to 10TB

Up to 80 Gbps I/O (VIC 1400)

Up to 15 TB Internal Storage



B200 M2 FCS: March 2010 EOL: March 2013

B200 M4: FCS: Oct 2014 EOL: Feb 2019

B200 M6: FCS: Apr 2021

B200 M3: FCS: Mar 2012, EOL: Dec 2016

B200 M5: FCS: Jul 2017

B200 M1/M2

B200 M3

B200 M4

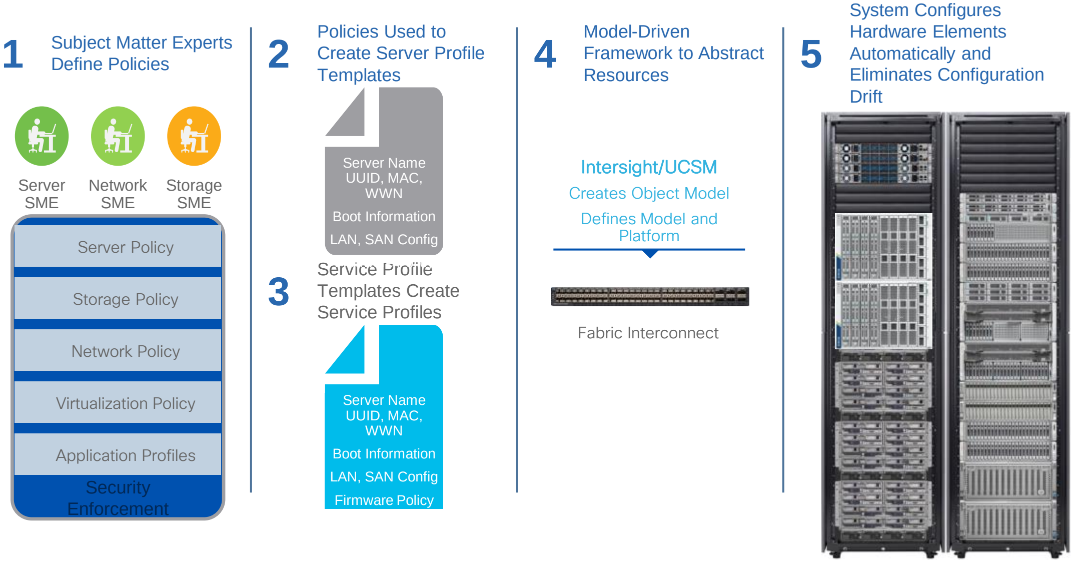
B200 M5

B200 M6

#1 Blade for Virtualization, Converged Infrastructure and Enterprise Applications

UCS: Policy and Model-Driven

- Reduces management costs with standardization
- Eliminates human error and missed steps
- Faster time to deploy and scale
- Enhanced security enforcement with centralized authorizations
  - Prevents local configuration changes and tampering



1 Subject Matter Experts Define Policies

Server SME

Network SME

Storage SME

Server Policy

Storage Policy

Network Policy

Virtualization Policy

Application Profiles

Security Enforcement

2 Policies Used to Create Server Profile Templates

Server Name
UUID, MAC, WWN
Boot Information
LAN, SAN Config

3 Service Profile Templates Create Service Profiles

Server Name
UUID, MAC, WWN
Boot Information
LAN, SAN Config
Firmware Policy

4 Model-Driven Framework to Abstract Resources

Intersight/UCSM

Creates Object Model

Defines Model and Platform

Fabric Interconnect

5 System Configures Hardware Elements Automatically and Eliminates Configuration Drift

UCS: Impact of Unified Fabric Design

Conventional Approach

Silos of multiple ethernet and SAN fabrics and adapters

48 LEGACY RACK SERVERS:

- 48 BMC management cables
- 96 DP ethernet cables
- 96 Fabric connect cables
- 384 Optics (4 ethernet, 4 fabric connects)

Massive complexity at scale

UCS

Cisco Unified Fabric and 100G reduces adapter, cabling, network and storage port needs

48 UCS X-SERIES SERVERS:

- 24 Chassis 100G cables
- 48 Chassis optics

88% ↓

reduction in cables and optics

66% ↓

reduction in cost of cabling components

Simplified, optimized, and automated

UCS X-Series leadership

Computing Purpose-Built for a Sustainable, AI-Powered Future

Cisco UCS X-Series with 4th & 5th Gen Intel® Xeon® Scalable processors and AMD 4th & 5th Gen EPYC™ processors

© 2024 Cisco and/or its affiliates. All rights reserved. Cisco Confidential

- Fastest ramping modular system in the industry
- Eliminating silos with UCS® X-Fabric
- Flexible, sustainable, easily upgradeable solution
- High-performance, smaller footprint
- Powers modern apps and traditional workloads
- Cisco's most energy-efficient server chassis

Silverlinings

Enterprise Strategy Group by TechTarget
TECHNICAL VALIDATION

SEAL
AWARD WINNER

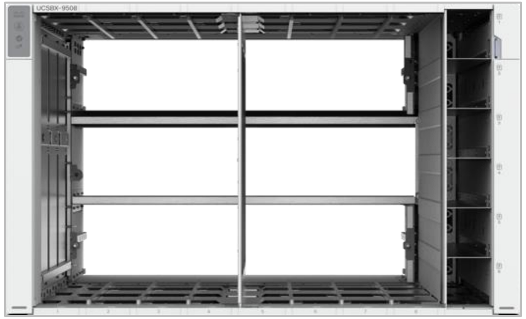
THE CHANNEL CO.
CRN
TECH INNOVATORS
WINNER
2023

THE CHANNEL CO.
CRN
Products of the YEAR
2023

UCS X9508 Modular Chassis

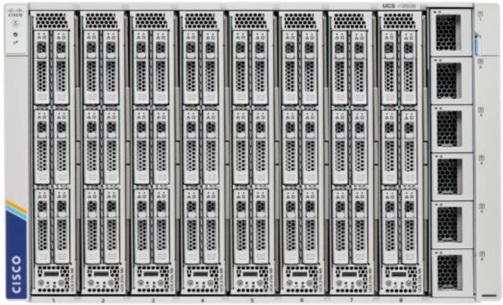
Chassis

7 RU



8 Nodes

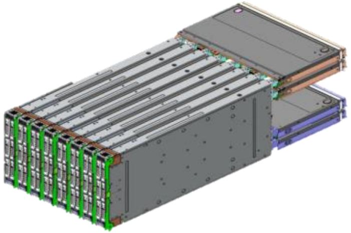
- Compute Node
- PCIe Node for GPU
- 6x 2800W PSU



Intelligent Fabric Modules

Flexible X-Fabric

Backplane-less design for fabric flexibility



Power and Cooling Innovation!

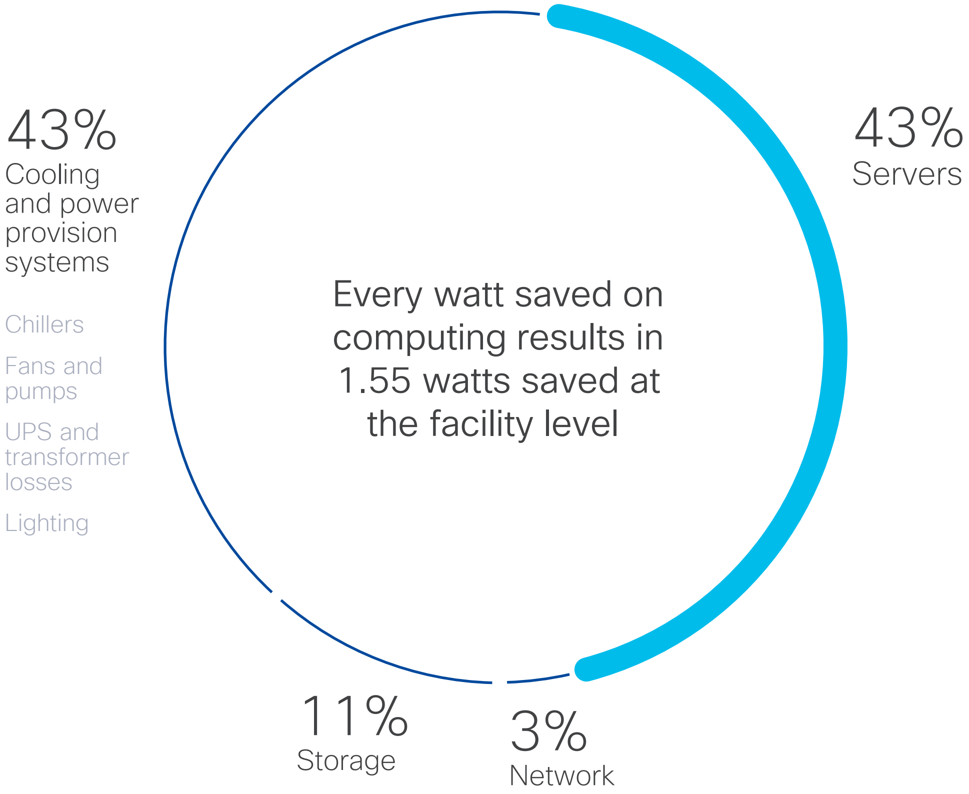
Cisco's approach to sustainability

Efficient data centers are an important sustainability opportunity

Today's data center accounts for **50X the power consumption** of a typical consumer office building.

This is expected to further increase with the adoption of high-performance computing and modern applications.

Source: akcp.com



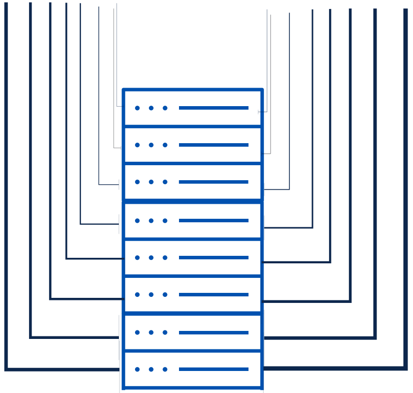
Every watt saved on computing results in 1.55 watts saved at the facility level

Impact of Cisco Power Efficiency Design

Powering a chassis v rack servers

Conventional approach

Rack servers requiring Dual PSUs per server for power and redundancy

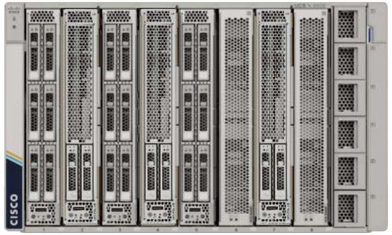


Rack Servers requiring 2 PSUs and 2 cords per server

→

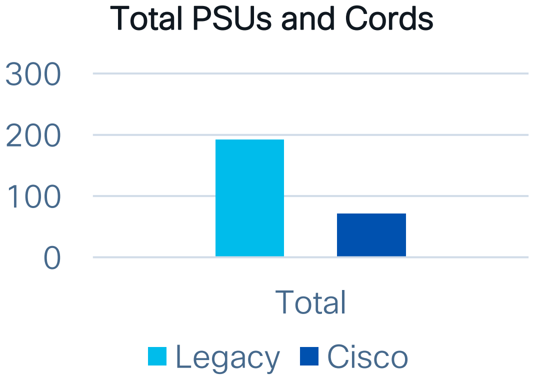
X-Series Power Efficiency

Purpose built for efficient transfer of power, redundancy, and materials reduction



- Efficient 54V Power Distribution
- Reduced Materials with fewer PSUs and cables
- 6 PSUs and 6 cords per chassis

X-Series vs Competitive Rack 48 Server comparison:



192 PSUs and cords

vs

Only 72 for X-Series

Energy Reduction Through Modernization

Case Study: Large Financial Services Organization

By replacing previous generation servers with UCS X-Series, a typical Cisco customer can expect:

70% reduction in total footprint

49% reduction in total power consumption

More modernization benefits for customers

90% ↓ reduction in hardware operating costs

72% ↓ reduction in hardware maintenance costs

75% ↓ reduction in recurring software support costs

X-series portfolio

COMPUTE

X210c Compute Node

- 2-Socket, single slot servers
- Two Generations: M6 and M7
- Intel 3rd Gen (Ice Lake) and 4th Gen (Sapphire Rapids) and 5th Gen (Granite Rapids) Xeon CPUs

X410c Compute Node

- 4-Socket, dual slot servers
- Intel 4th Gen Xeon CPU
- Up to 64 DDR5 DIMMs

X215c Compute Node

- 2-Socket, single slot servers
- M8 with AMD 4th gen EPYC CPU

FABRIC

4th and 5th Gen FI

- 25/100G ports
- Unified ports: Up to 16x 32G FC ports (6536)
- Supports VIC 1400, 14000 and 15000 series

UCS X-Series Direct

- Scale at the edge with X-series advantage for 1-16 servers

25/100G IFM

- 8 x 25/100G connectivity

4th and 5th Gen VIC

- 25/100G connectivity for both blades and racks

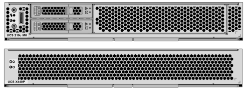
X-FABRIC AND PCIE NODE

X-Fabric

- Based on native PCIe Gen 4
- Provides GPU acceleration to enterprise application
- No backplane or cables = Easy upgrades

GPU Node and Front Mezz GPUs

- Nvidia A16, Nvidia L40, Nvidia L4 and Nvidia H100 GPUs today in various configurations



AMD Compute

UCS X215 M8

UCS C225 M8

UCS C245 M8

A complete rack and blade portfolio

Now with AMD’s game-changing high-performance 5th generation EPYC™ processors

Get high-performance servers customized to your unique use case, performance, and sustainability needs

AMD EPYC™ 700x and 900x series roadmap update

March 2023 Genoa

Genoa
4th Gen EPYC™

SP5 Platform

Zen4 96 cores, 5nm

~360W,

12CH DDR5-4800

160L PCIe5.0/64L CXL

High perf per socket and per core

CQ3'2023 Bergamo

Bergamo
4th Gen EPYC™

Genoa + highest thread density architecture

128 Zen4c cores

~360w

Optimized for Perf/watt

CQ3'2023 Genoa-X

Genoa-X
4th Gen EPYC™

Genoa + large L3 cache

Up to 1.15GB L3

~360w

Designed for technical computing

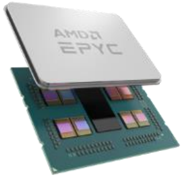
CQ4'2024 Turin

Turin
5th Gen EPYC™

128 ~400W

12CH DDR5-6000

CXL 2.0



Cisco UCS M8

EPYC 9004 to 9005 stack evolution

- EPYC 9005 “Turin”
- Elevates peak core count
- Streamlines OPN stack
- Ensures gen/gen transition for essential OPNs at iso TDP range
- Introduces memory optimized OPNs
- Re-introduces 8c and 16c options for lower TDP ranges
- Introduces 64C ‘Host Node’ CPU for GPU/AI

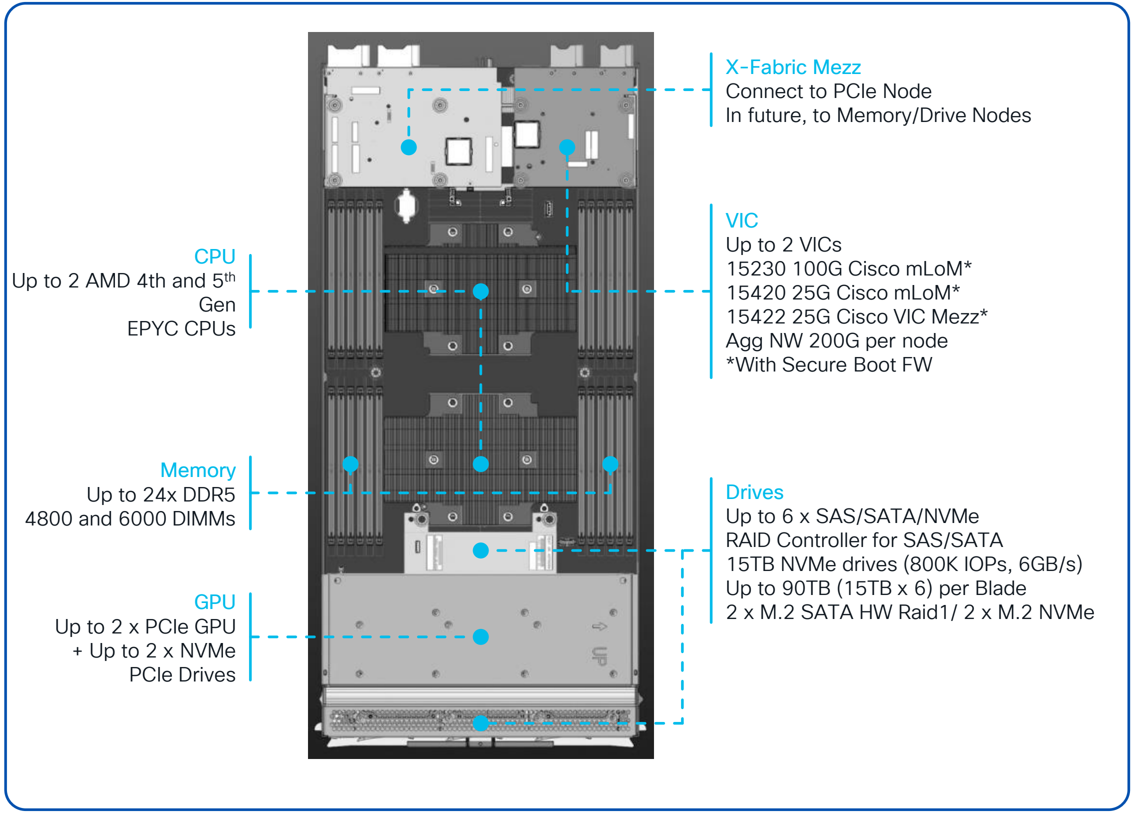
EPYC 9004 “Genoa” / “Bergamo”

128 cores	360W	9754
112 cores	340W	9734
96 cores	400W 360W	9684X 9654/P
84 cores	290W	9634
64 cores	360W 280W	9554/P 9534
48 cores	360W 290W	9474F 9454/P
32 cores	320W 320W 280W 210W	9384X 9374F 9354/P 9334
24 cores	320W 200W 200W	9274F 9254 9224
16 cores	320W 320W 200W	9184X 9174F 9124

EPYC 9005 “Turin”

160 cores	400W	9845
144 cores	400W	9825
128 cores	400W	9745
96 cores	400W 320W	9655/P 9645
72 cores	400W	9565
64 cores	320-400W 360W 300W	9575F 9555/P 9535
48 cores	360W 300W	9475F 9455/P
36 cores	300W	9365
32 cores	320W 280W 210W	9375F 9355/P 9335
24 cores	320W 200W	9275F 9255
16 cores	320W 200W 155W	9175F 9135 9115
8 cores	155W	9015

UCS X215c M8 2S Compute Node: Key features



CPU
Up to 2 AMD 4th and 5th Gen
EPYC CPUs

Memory
Up to 24x DDR5
4800 and 6000 DIMMs

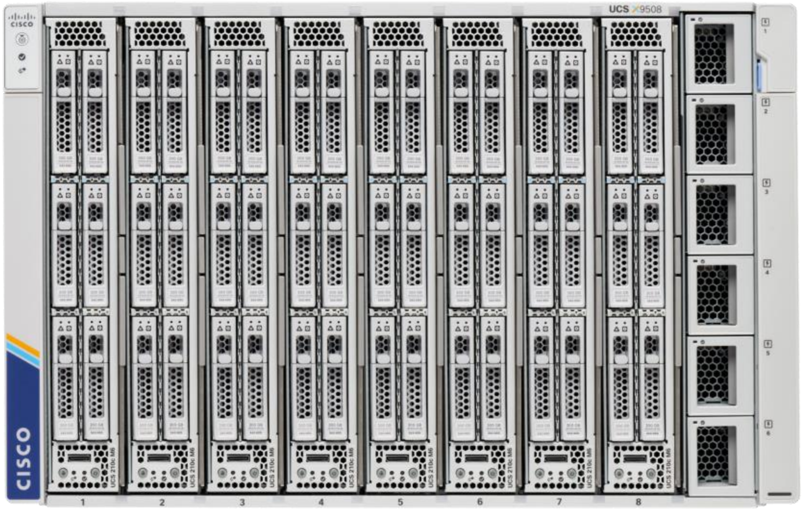
GPU
Up to 2 x PCIe GPU
+ Up to 2 x NVMe
PCIe Drives

X-Fabric Mezz
Connect to PCIe Node
In future, to Memory/Drive Nodes

VIC
Up to 2 VICs
15230 100G Cisco mLoM*
15420 25G Cisco mLoM*
15422 25G Cisco VIC Mezz*
Agg NW 200G per node
*With Secure Boot FW

Drives
Up to 6 x SAS/SATA/NVMe
RAID Controller for SAS/SATA
15TB NVMe drives (800K IOPs, 6GB/s)
Up to 90TB (15TB x 6) per Blade
2 x M.2 SATA HW Raid1/ 2 x M.2 NVMe

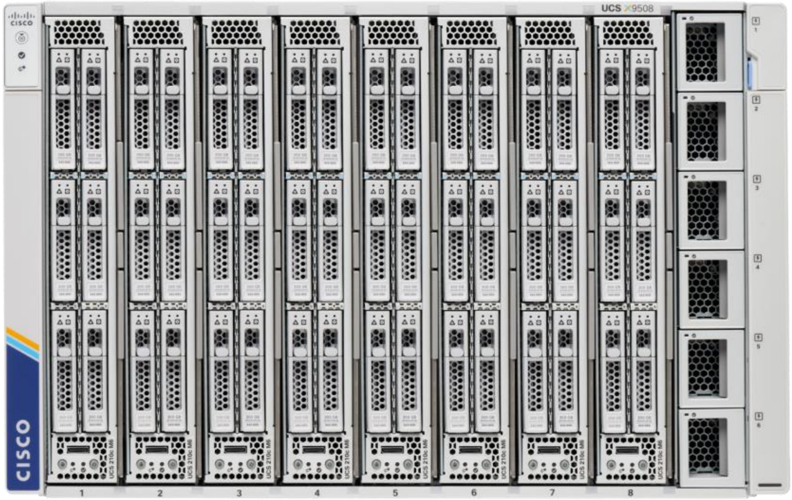
UCS X-Series with Intel M7



Up to 1024

Cores
per Chassis
(M6/M7)

UCS X-Series with AMD M8



Up to 2560

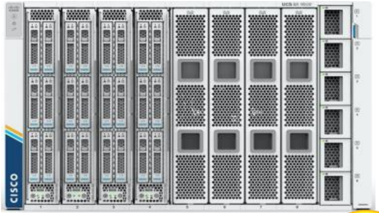
Cores
per Chassis
(M8)

Cisco UCS Compute Portfolio

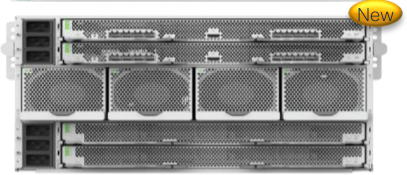
Mainstream Enterprise Servers

UCS X-Series
X9508 Chassis

IFM Module



UCS X-Series
Direct



UCS X210c M7

UCS X210c M6

UCS X410c M7

UCS B200 M6

UCS X215c M8

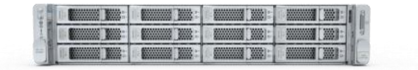
UCS C240 M7SX
28 HDD/SDD/NVMe

UCS C240 M7SN
28 NVMe

UCS C240 M6S
14 SSD/HDD
Media drive

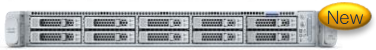
UCS C240 M6N
14 NVMe
Media Drive

UCS C240 M6L
16 LFF + 4 SFF



UCS C220 M7S
10 HDD/SSD/NVMe

UCS C220 M7N
10 NVMe



UCS C245 M8SX
28 HDD/SSD/NVMe

UCS C225 M8S
10 HDD/SSD/NVMe

UCS C225 M6N
10 NVMe

Specialized Servers

UCS C885A
8RU Dense GPU
Server

UCS S3260
60 LFF Drives
Storage

Cisco UCS X-Series Direct

Cisco unified edge use cases

SASE Edge

Connecting and securing users and devices

- Secure Gateway
- DNS
- Cloud Firewall
- VPN
- DLP
- Zero Trust
- CASB
- Cloud OnRamp
- SD-WAN

App Edge

Protecting and accelerating applications and data

Application Networking and Security

- Runtime App Self Protection
- Web App Firewall
- DDoS
- Bot Detect

Application and Content Delivery

- Caching
- DNS
- Content Opt.
- Load Balancing

Enterprise Edge

Infra platforms optimized for distributed apps and data

- Video Analytics
- DCaaS
- Industrial IoT
- Retail
- AI/ML
- AR/VR

Carrier Edge

Use cases for Service Provider Edge assets

- Connected Cars
- Private LTE
- Localization
- 5G
- Content and Media Delivery
- Enterprise Services

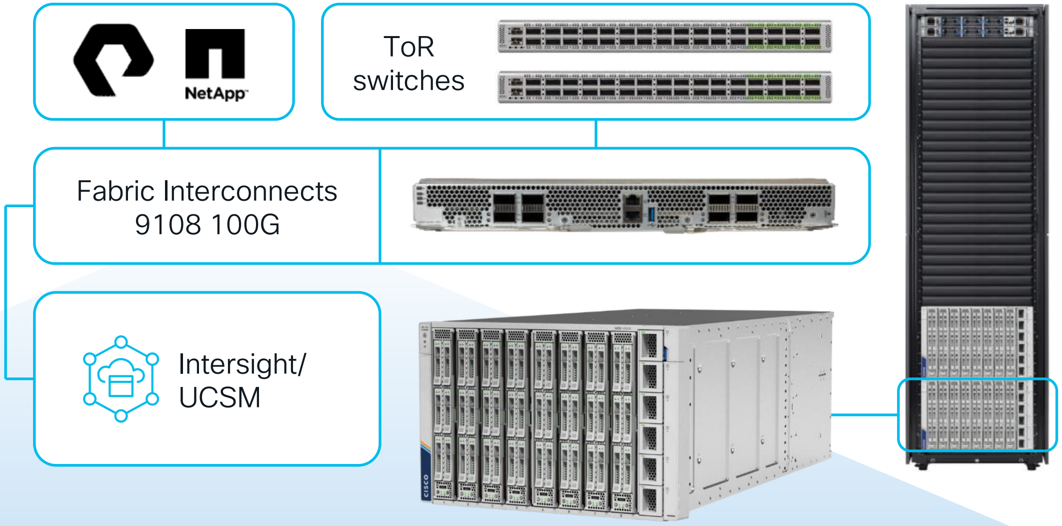
Target and Focus

Modernize at the edge

New edge solution for
hybrid cloud infrastructure

Flexible and powerful infrastructure to handle all workloads:

- Compute
- Networking (1/10/25/40/100G)
- Storage (8/16/32G)
- GPU
- Same chassis, compute nodes, fans, power supplies



Up to 68% CapEx savings

Up to 64% better performance

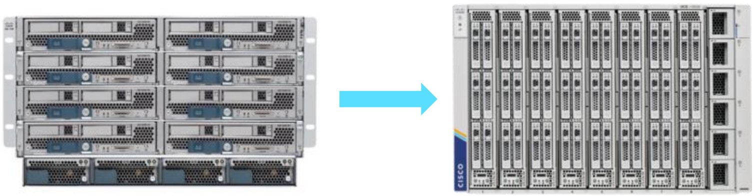
Up to 49% lower power

Up to 50% more sustainable*

* ESG: The Role of Cisco UCS X-Series in Fulfilling Sustainability Objectives: July 2023

UCS X-SERIES DIRECT

Customer Demand



Any Size - Any Location

Datacenter Remote Branch
Remote Branch+ Far Edge
Small DC
Far Edge+ Enterprise
Micro DC Public
Medium DC
Commercial

Any Workload

SAP

FlexPod
A Cisco and NetApp Solution

vmware
Horizon View

VMware vSphere 8

FlashStack

Microsoft SQL Server

RED HAT OPENSHIFT
Container Platform

VMware vSAN

Customer Savings

Sustainability through integrated FI

CapEx and OpEx savings through Consolidation improvement

Cisco UCS Fabric

UCS 4th and 5th generation Fabric

	4th gen	5th gen
Fabric Interconnect	 FI 6454 FI 64108	 FI 6536 UCSX-S9108-100G
IOM/IFM/ FEX	 IOM 2408 IFM X-9108 25G FEX 93180YC-FX3	 IOM 2408 IFM X-9108 25G FEX 93180YC-FX3 IOM 2304 IFM X-9108 100G
Virtual Interface Card	 VIC 1400 / 14000 Series (10G/25G/40G/100G)	 VIC 15000 Series (10/25/40/50/100/200G)

□ EOS in Nov-CY24

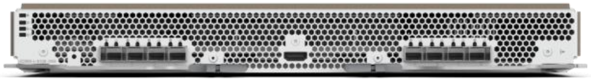
Fabric Interconnect– 25G and 100G

	6454	64108	6536
Total Ports	54	108	36
Ethernet only Ports	28x 10/25G, 4x 1/10/25G and 6x 40G	72x 10/25G, 8x 1/10/25G and 12x 40G	32(10*/25/40/100G)
Unified Ports	16x 10/25G Ethernet or 4/8/16G FC	16x 10/25G Ethernet or 4/8/16G FC	4x100G (10*/25/40/100G) Ethernet or 16x 8/16/32G FC with breakout
1G Ports	Port 45-48	Port 89-96	Port 9 and 10
IFM	25G	25G	25G and 100G

CISCO Live!

*Uplink and Appliance port only

Intelligent Fabric Module (IFM) – 25G and 100G

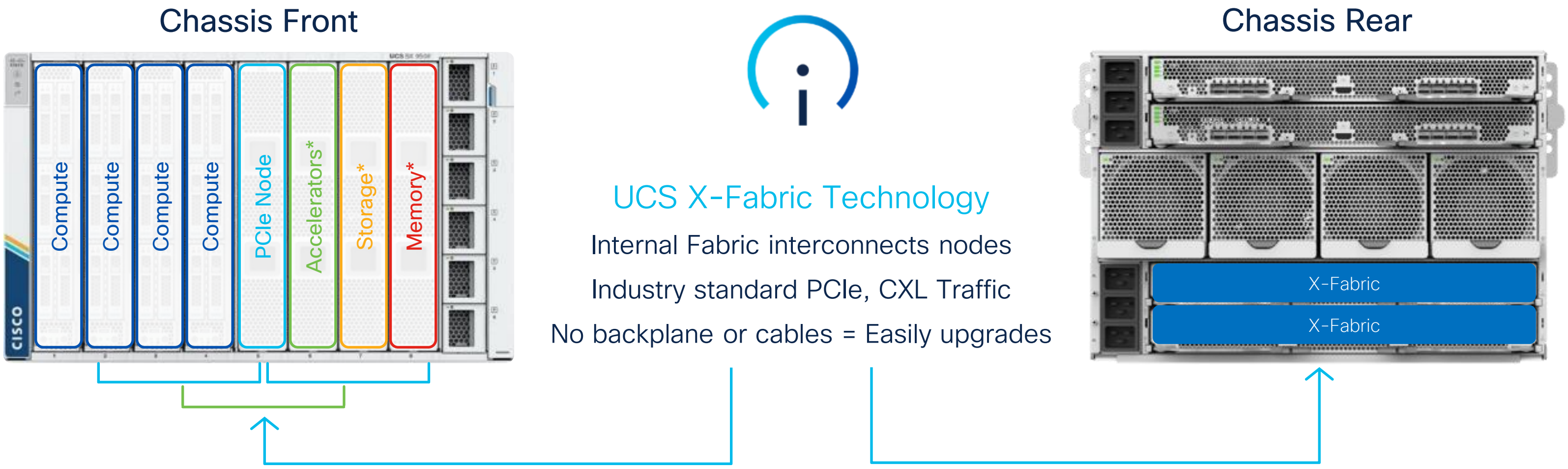


	IFM 9108-25G	IFM 9108-100G
Fabric Interconnect	6454, 64108, 6536	6536
VIC	15231,14425,+14825	15231,14425, +14825
Network Interface (NIF) Ports	8 x 25G (port-channel)	8 x 100G (port-channel)
Host Interface (HIF) Ports	32 x 25G	8 x 100G or 32 x 25G
Oversubscription	4 : 1	1 : 1

X-Fabric to Support the Future

UCS X-Fabric Technology

Shape infrastructure resources to applications



UCS X-Fabric Technology

Internal Fabric interconnects nodes

Industry standard PCIe, CXL Traffic

No backplane or cables = Easily upgrades

* Roadmap items – subject to change

Run modern apps by assembling modules into systems

Simple policy-based definition of resources from Intersight

Engineered to accommodate future technologies

UCS X-Fabric technology and PCIe Nodes with GPU

PCIe node supports up to:

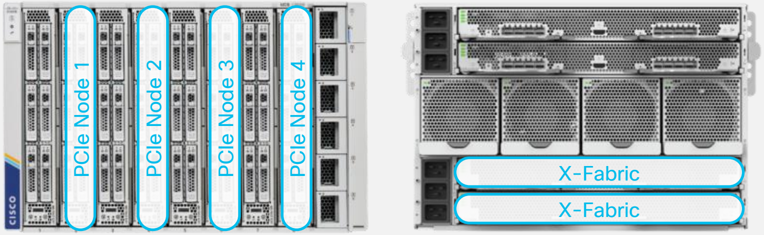
4x
Intel Data Center
GPU Flex 140

2x
Intel Data Center
GPU Flex 170

2x
Nvidia A16

2x
Nvidia H100
Nvidia L40
Nvidia L40S
Nvidia A40
Nvidia A100
Nvidia Nvidia H100 NVL

4x
Nvidia T4
Nvidia L4



X-Fabric decouples the lifecycles of CPU, GPU, memory, storage and fabrics, providing a perpetual architecture that efficiently brings you the latest innovations.

- ✓ Cloud-powered composability with Cisco Intersight
- ✓ Flexible GPU acceleration across server nodes
- ✓ No backplane or cables = easily upgrades

UCSX GPU front mezz

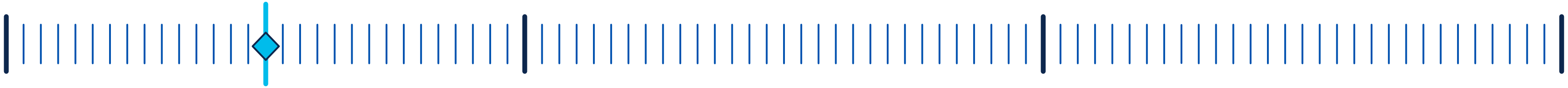
Run more apps in less space

- High-density form factor supports a wide range of workloads
- Two Nvidia T4, L4, Intel Flex140 GPUs for AI inferencing, data analytics, and graphics

Cloud operated with Cisco Intersight

Cisco AI Ready Datacenter

Every organization’s AI approach and needs are different



Build the model
Training

Optimize the model
Fine-tuning and RAG

Use the model
Inferencing

Bringing 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

Orderable Now

UCS Accelerated

UCS C885A M8

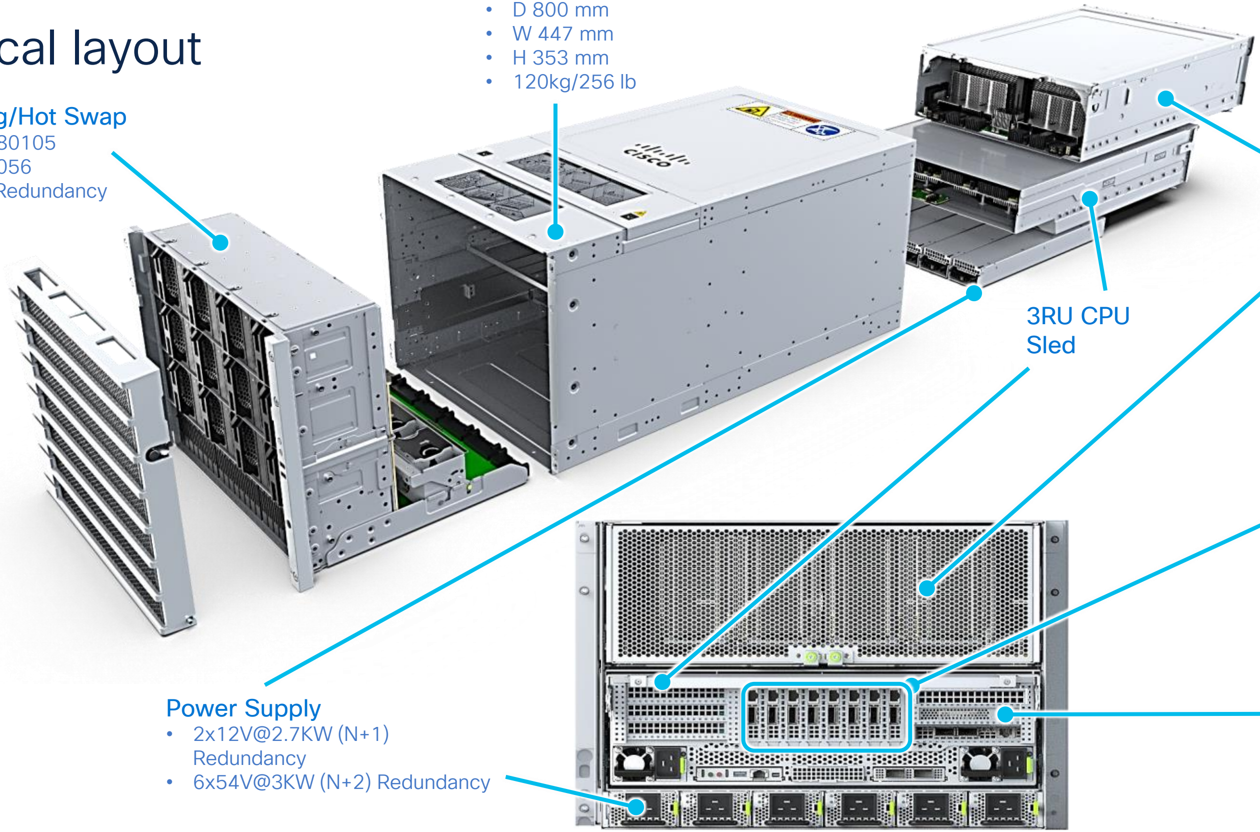
Nvidia HGX with
8 Nvidia H100/H200 GPUs

AMD Mi300X

2 AMD 4th Gen
EPYC™ Processors

C885A M8

Phvsical layout



8RU 19" Chassis
- D 800 mm
- W 447 mm
- H 353 mm
- 120kg/256 lb

Cooling/Hot Swap
- 12 x 80105
- 4 x 6056
- N+1 Redundancy

4RU GPU Sled

HGX/UBB Options (8xGPU)
- H100, H200
- MI300X

3RU CPU Sled

PCIe E-W Options
- 8 x CX7 400G
- 8 x BF3 B3140H (non crypto)

PCIe N-S
- 1 x BF3 B3220 (non-crypto)

Power Supply
- 2x12V@2.7KW (N+1) Redundancy
- 6x54V@3KW (N+2) Redundancy

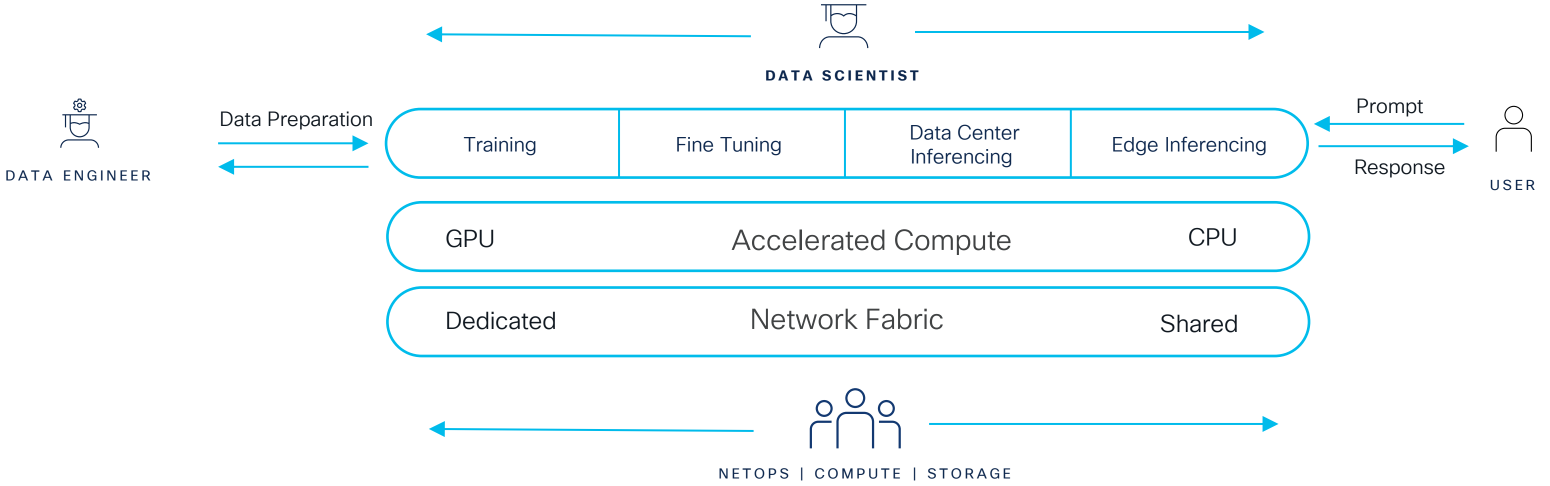
C885A M8

Product specifications

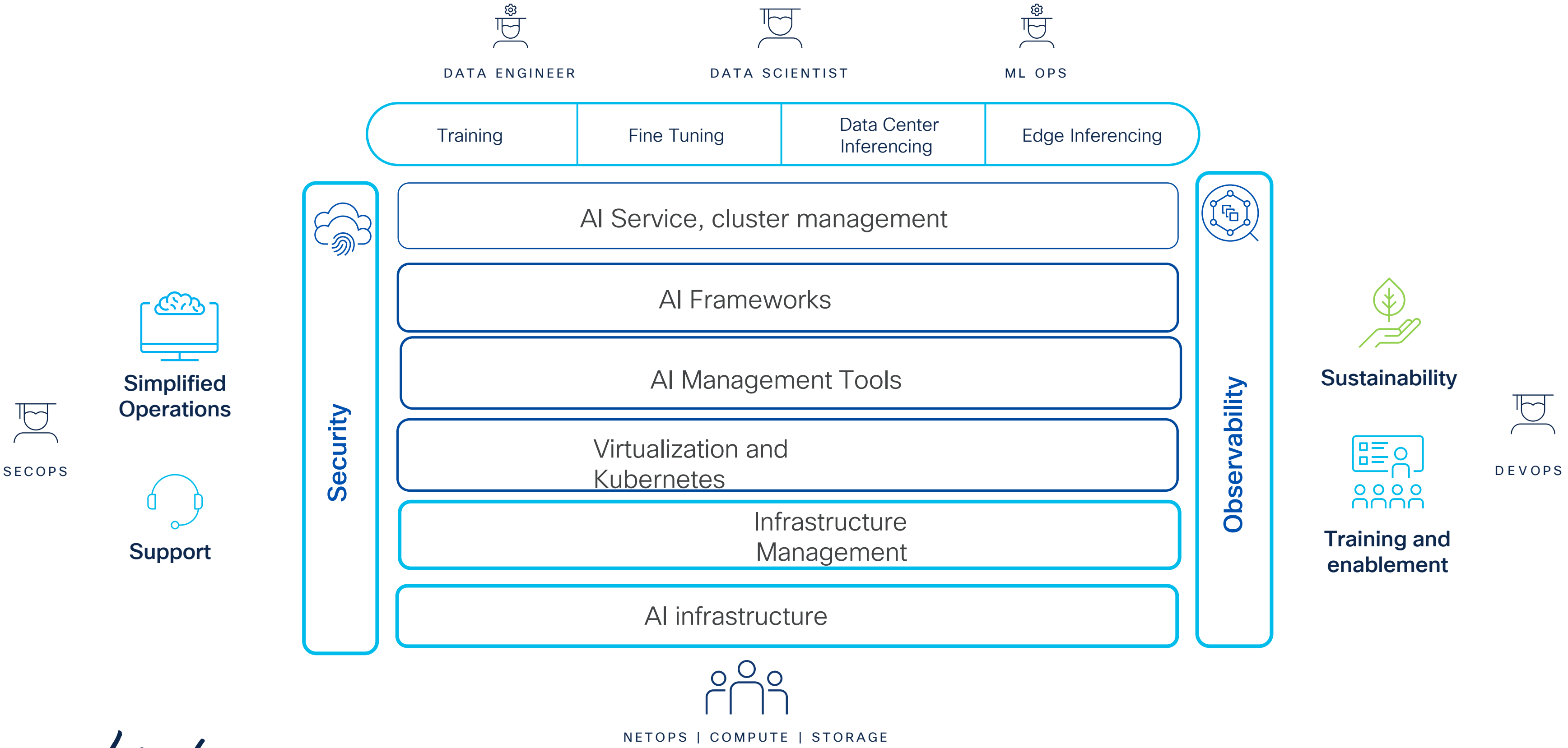
Form Factor	• HGX 8U 19" EIA Rack
Compute + Memory	• 2 AMD EPYC 4th (Genoa) or 5th (Turin) Gen CPUs • 24 DDR5 RDIMMs, up to 6,000 MT/S
Storage	• 1 PCIe3 x4 M.2 NVMe (boot device) • 16 PCIe5 x4 2.5" U.2 NVMe SSD (data cache)
GPU	• 8 H100 700W or 8 H200 700W or 8 B200A 700W • 8 MI300X 750W
Network Cards	• 8 PCIe5 x16 HHHL for E-W NIC ConnectX-7, BF3 B3140H • 5 PCIe5 x16 FHHL for N-S NIC BF3 B3220, B3240 (max 2) • 2 OCP 3.0 SFF
Cooling	• 12 80105 hot-swappable (N+1) fans for system cooling • 4 6056 fans for SSD cooling
Front IO	• 2 USB 2.0, 1 ID BTN, 1 Power Button
Rear IO	• 1 USB 3.0 A, 1 USB 3.0 C, mDP, 1 ID BTN, 1 Power Button, 1 USB 2.0 C (for debugging), 1 RJ45 (mgmt.)
Power Supply	• 6 54V 3kW (N+2) MCRPS and 2 12V 2.7kW CRPS, N+1 redundancy

Cisco AI Infrastructure POD

Generative AI – Infrastructure System (DS view)



Generative AI – Full Stack System



DATA ENGINEER

DATA SCIENTIST

ML OPS

Training | Fine Tuning | Data Center Inferencing | Edge Inferencing

Security

AI Service, cluster management

AI Frameworks

AI Management Tools

Virtualization and Kubernetes

Infrastructure Management

AI infrastructure

Observability

SECOPS

Simplified Operations

Support

Sustainability

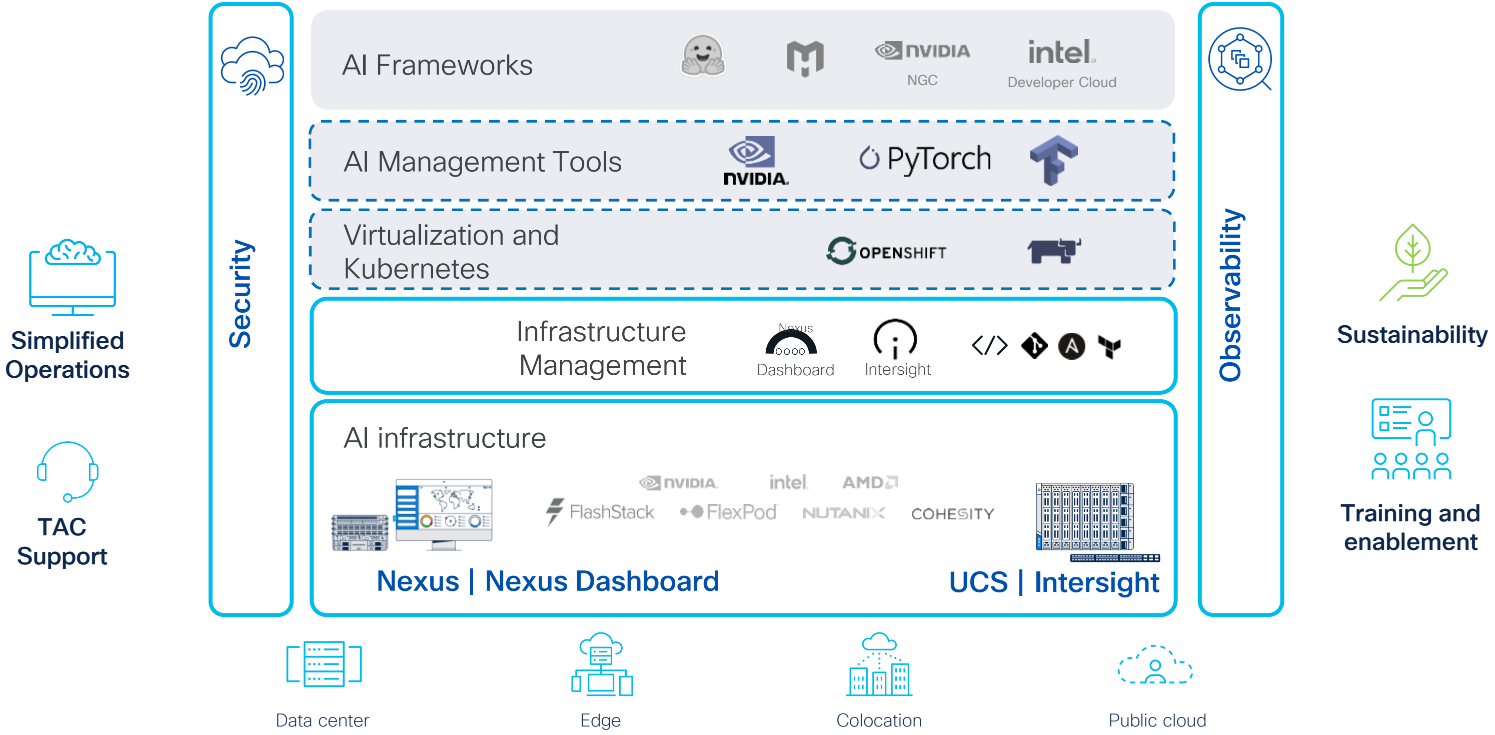
Training and enablement

DEVOPS

NETOPS | COMPUTE | STORAGE

AI-Ready Infrastructure Stack

Easy



Simplified Operations

TAC Support

Security

AI Frameworks

NVIDIA NGC

intel Developer Cloud

AI Management Tools

NVIDIA

PyTorch

Virtualization and Kubernetes

OPENSHIFT

Infrastructure Management

Nexus Dashboard

Intersight

AI infrastructure

NVIDIA intel AMD

FlashStack FlexPod NUTANIX COHESITY

Nexus | Nexus Dashboard

UCS | Intersight

Observability

Sustainability

Training and enablement

Data center

Edge

Colocation

Public cloud

Simplified Orderability

AI PODs

Faster time to value with pre-configured bundles

ORDERABLE NOVEMBER

Deploy AI with confidence

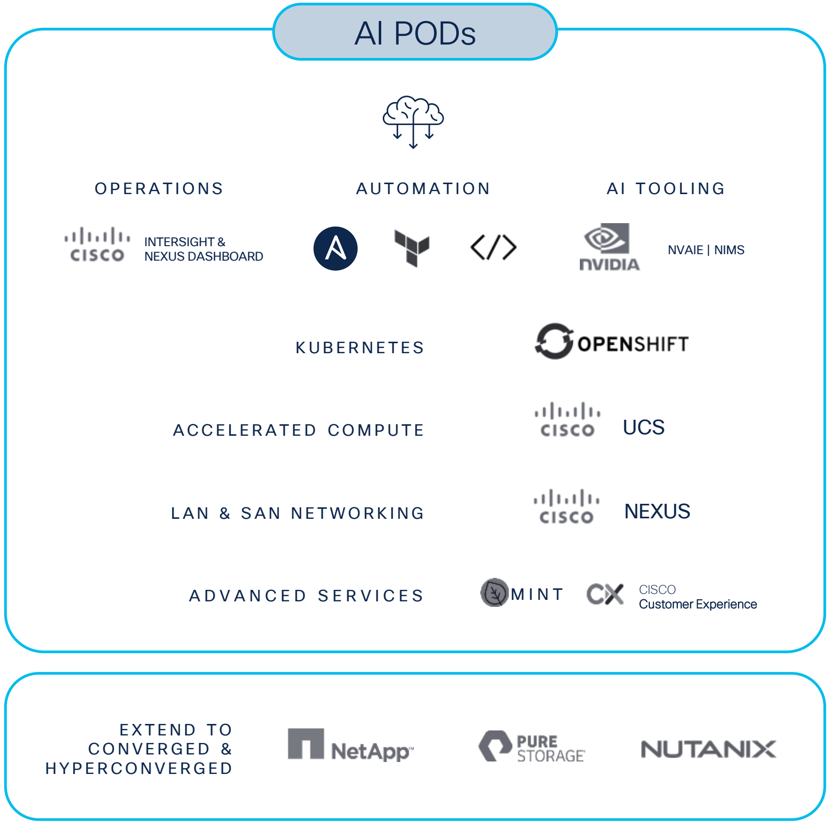
Orderable, validated AI-ready infrastructure stacks

Fully supported stack including Cisco and 3rd party components

AI Advisor tool for configuration guidance

COMING SOON

Cisco AI-Ready Infrastructure Stacks



AI PODs

OPERATIONS
- cisco INTERSIGHT & NEXUS DASHBOARD

AUTOMATION

AI TOOLING
- NVIDIA NVAIE | NIMS

KUBERNETES: OPENSHIFT

ACCELERATED COMPUTE: cisco UCS

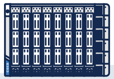
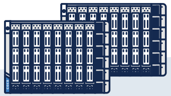
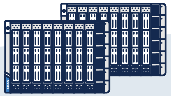
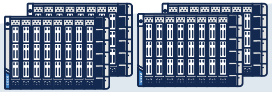
LAN & SAN NETWORKING: cisco NEXUS

ADVANCED SERVICES: MINT, CX CISCO Customer Experience

EXTEND TO CONVERGED & HYPERCONVERGED: NetApp, PURE STORAGE, NUTANIX

Infrastructure Modernization for AI

AI PODs

Typical use case	Data Center and Edge Inference	RAG Augmented Inference Contextual Accuracy	Scale Up for High Performance Inference Expanded Context	Scale Out for Large Deployments Multi-model deployments High Concurrency	Roadmap
Sizing	**(Llama-2 7B GPT 2B)**	**(Llama-2 13B OPT 13B)**	**(Code Llama 34B Falcon 40B)**		
PID	UCSX-AI-EDGE	UCSX-AI-RAG	UCSX-AI-LARGERAG	UCSX-AI-LARGEINF	
Pod Specification	1x X210C compute node 2x Intel 5th Gen 6548Y+ 512 GB System Memory 2x 1.6NVMe drives 1x X440p PCIe 1x Nvidia L40S	2x X210C compute nodes 4x Intel 5th Gen 6548Y+ 1 TB System Memory 4x 1.6NVMe drives 2x X440p PCIe 4x Nvidia L40s	2x X210C compute nodes 4x Intel 5th Gen 6548Y+ 1 TB System Memory 4x 1.6NVMe drives 2x X440p PCIe 4x Nvidia H100 NVL	4x X210C compute nodes 8x Intel 5th Gen 6548Y+ 4 TB System Memory 8x 1.6NVMe drives 4x X440p PCIe 8x Nvidia L40s	
					

Performance and Scale

AI Solutions

Integrated, validated solutions on proven platforms

Simplify and Automate AI Infrastructure

Curated automation playbooks to get started

1 Cisco Validated Designs for Simplified AI Infrastructure

EXPANDED ROADMAP

NVIDIA	Red Hat	OpenAI	CLOUDERA
NVIDIA AI Enterprise	Red Hat OpenShift AI	GPT-in-a-box on Nutanix Hyperconverged	Gen-AI with Cloudera Data Platform

AMD FlashStack FlexPod intel. NUTANIX

2 Deployment-ready playbooks for common AI models

NEW!

Large Language Models (GPT3, BERT, T5)	Computer Vision Models (ResNet, EfficientNet, YOLO)	Generative Models (GANs, VAEs)

NVIDIA NGC

intel. Developer Cloud

FlashStack for Generative AI Inferencing

Foundational Architecture for Gen AI

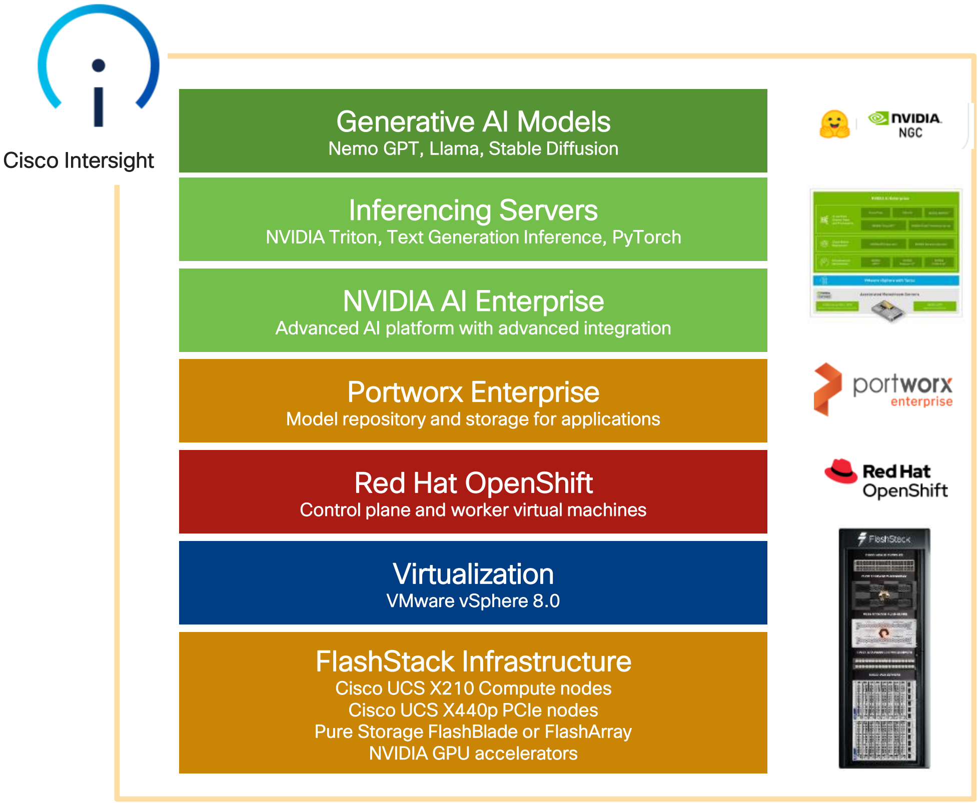
- Validated NVIDIA NeMo Inference with TensorRT-LLM that accelerates inference performance of LLMs on NVIDIA GPUs
- Validated models using Text Generation Inference server from Hugging Face
- Metrics dashboard for insights into infrastructure, cluster and GPU performance and behavior

Accelerate Model Deployment

- Extensive breadth of validation of AI models such as GPT, Stable Diffusion and Llama 2 LLMs with diverse model serving options
- Automated deployment with Ansible playbook

Consistent Performance

- Consistent average latency and Throughput
- Better price to performance ratio



Cisco Intersight

Generative AI Models
Nemo GPT, Llama, Stable Diffusion

Inferencing Servers
NVIDIA Triton, Text Generation Inference, PyTorch

NVIDIA AI Enterprise
Advanced AI platform with advanced integration

Portworx Enterprise
Model repository and storage for applications

Red Hat OpenShift
Control plane and worker virtual machines

Virtualization
VMware vSphere 8.0

FlashStack Infrastructure
Cisco UCS X210 Compute nodes
Cisco UCS X440p PCIe nodes
Pure Storage FlashBlade or FlashArray
NVIDIA GPU accelerators

FlexPod for Generative AI Inferencing

Optimized for AI

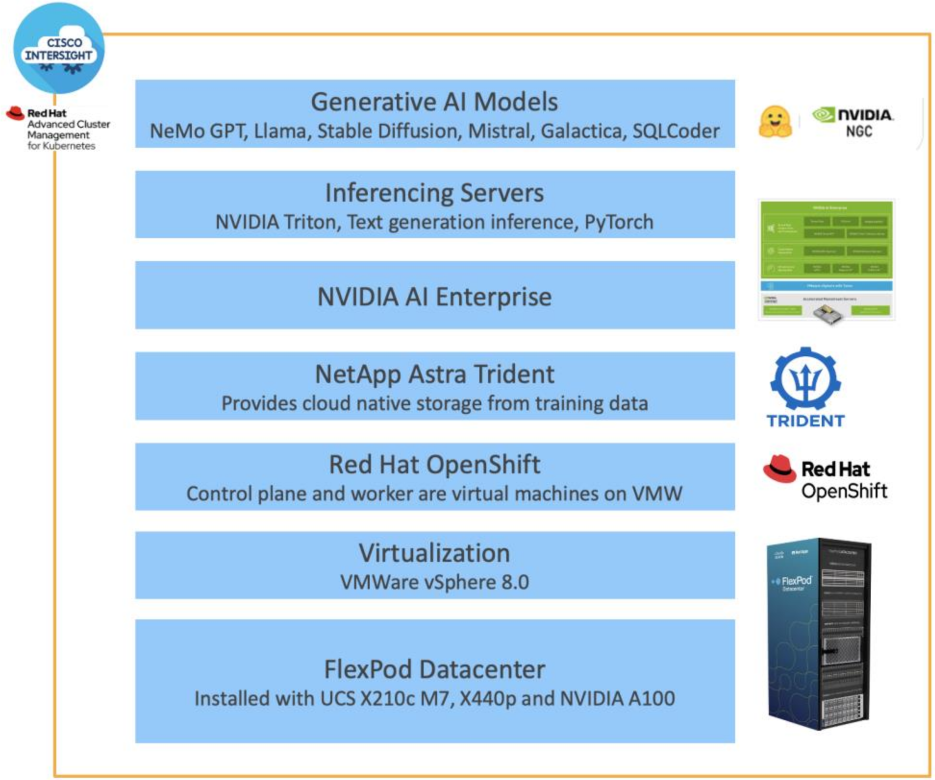
- Comprehensive suite of AI tools and frameworks with NVIDIA AI Enterprise that support optimization for NVIDIA GPU
- Validated NVIDIA NeMo with TRT-LLM that accelerates inference performance of LLMs on NVIDIA GPUs
- Metrics dashboard for insights into cluster and GPU performance and behavior

Accelerated Deployment

- Deployment validation of popular Inferencing Servers and AI models such as Stable Diffusion and Llama 2 LLMs with diverse model serving options
- Automated deployment with Ansible playbook

AI at Scale

- Scale discretely with future-ready and modular design



Generative AI Models
NeMo GPT, Llama, Stable Diffusion, Mistral, Galactica, SQLCoder

Inferencing Servers
NVIDIA Triton, Text generation inference, PyTorch

NVIDIA AI Enterprise

NetApp Astra Trident
Provides cloud native storage from training data

Red Hat OpenShift
Control plane and worker are virtual machines on VMW

Virtualization
VMWare vSphere 8.0

FlexPod Datacenter
Installed with UCS X210c M7, X440p and NVIDIA A100

cisco Live!

MLOps using Red Hat OpenShift AI on FlashStack

Operationalize AI

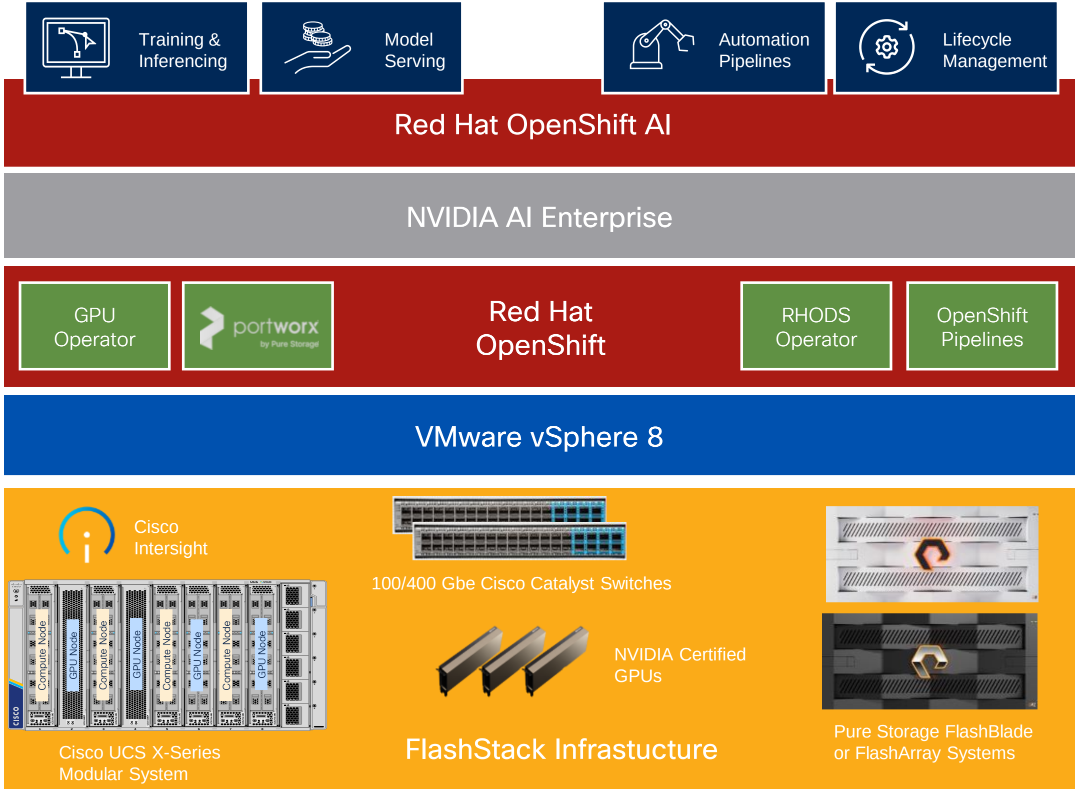
- Accelerate delivery of AI/ML models and applications seamlessly and consistently
- Automated deployment with Ansible playbook

Pre-Integrated with AI/ML Tools

- Compatible with TensorFlow, PyTorch, Jupyter notebook images, Data Hub etc, OpenVino.
- Innovate faster with an open-source approach.

Scalable

- Develop, deploy and manage multiple AI model initiatives consistently, and across datacenter, edge, and public clouds
- Scale discretely with future-ready and modular design



FlexPod with SUSE Rancher for AI workloads

Optimized for AI

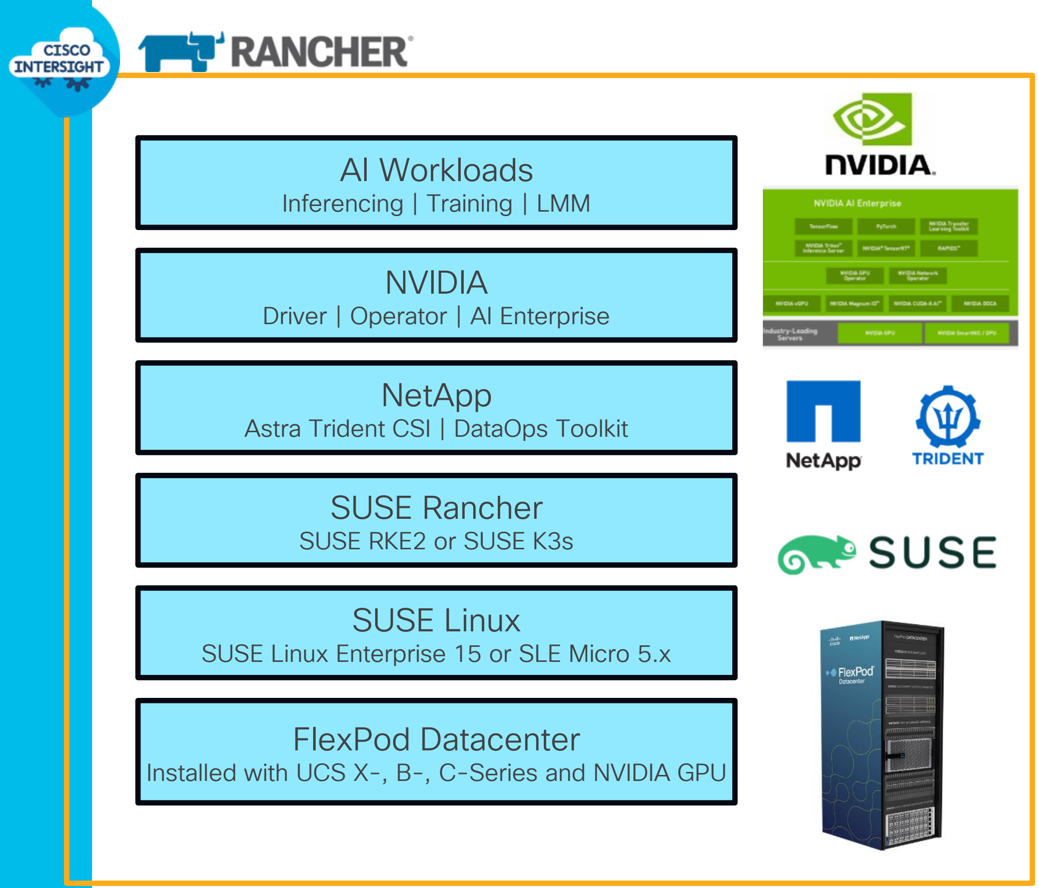
- Comprehensive suite of AI tools and frameworks with NVIDIA AI Enterprise that support optimization for NVIDIA GPU
- Validated different deployment options with SUSE Linux Enterprise (SLE), SUSE SLE Micro, RKE2 and K3s together with NVIDIA GPUs
- Grafana dashboard for insights into cluster and GPU performance and behavior

Accelerated Deployment

- Deployment validation of AI enabled platform with containerized NVIDIA software stack.
- Ready for many different AI workloads
- Automated deployment with Ansible playbook

AI at Scale

- Scale discretely with future-ready and modular design



AI Workloads
Inferencing | Training | LMM

NVIDIA
Driver | Operator | AI Enterprise

NetApp
Astra Trident CSI | DataOps Toolkit

SUSE Rancher
SUSE RKE2 or SUSE K3s

SUSE Linux
SUSE Linux Enterprise 15 or SLE Micro 5.x

FlexPod Datacenter
Installed with UCS X-, B-, C-Series and NVIDIA GPU

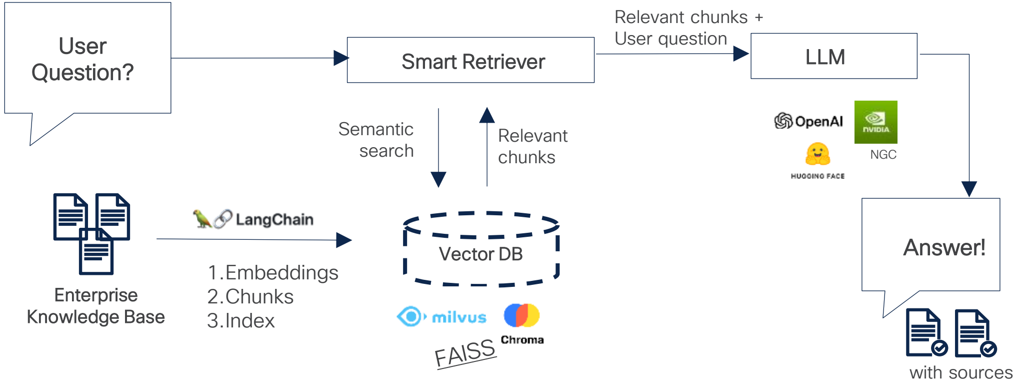
Retrieval-Augmented Generation (RAG)

LLM CHALLENGES

Presenting out-of-date or generic information when the user expects a specific, current response

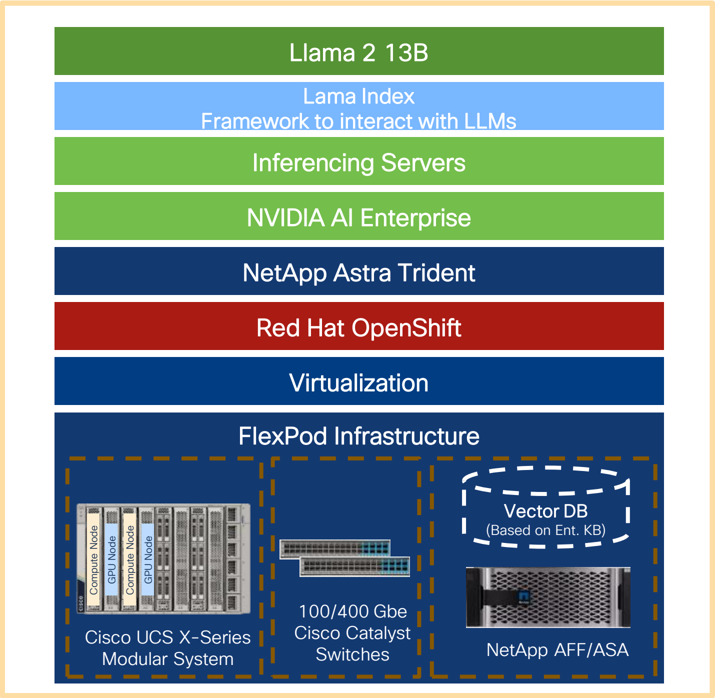
Need for frequent model training with new data, leading to increased expenses

"Hallucinations" - generating plausible sounding but ultimately inaccurate statements, e.g., when creating a response from non-authoritative sources



EXAMPLE USE CASES

- Conversational agents
- Content generation
- Personalized recommendation engines
- Real-time event commentary



Source: IDC Worldwide Gen AI 2024 Predictions

Cisco UCS X-Series for HCI

Cisco Hyperconverged HCI-X Solution

Cisco Compute Hyperconverged X-series chassis, X210c M7 All-NVMe Nodes (Intel 4th Gen and 5th Gen CPUs) and X-series Direct

Cisco Compute Hyperconverged X210c M7 All NVMe Nodes (Intel 4th and 5th Gen CPUs)

- Up to 6 x 1.9TB, 3.8TB, 7.6TB or 15.3TB NVMe disks
- 5th Gen mLOM @ 4x 25Gbps or 2x 100Gbps
- Dual M.2 SATA SSDs with HW RAID for boot
- Mandatory 2 CPUs and up to 8TB Memory
- GPUs via X440p PCIe node: L4, L40S, H100-80, A16.
- Intersight Managed Mode (IMM) support
- 1-node, 2-node, and 3-node+ clusters
- Support for ESX and AHV

Cisco Intersight

Nutanix Prism Central

HCI-X Hardware:

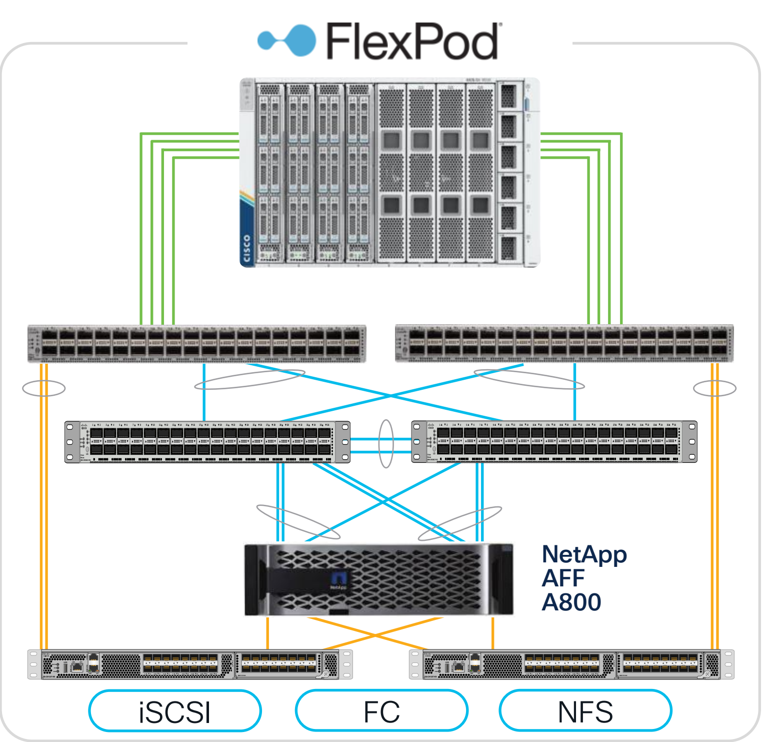
- HCI X-series system
- X210c M7 (4th and 5th Gen Emerald Rapids)
- X-series Direct

Intersight Managed Mode (IMM):

Day 0 - Cluster deployment with Foundation Central (FC)
Day 2 - Cluster expansion and Integrated firmware upgrades using LCM

Cisco UCS X-Series with FlexPod and FlashStack

Converged Infrastructure with X-Series



Youtube Video Link

Cisco UCSX-9508 Chassis

Cisco UCSX 9108-100G

Cisco UCS X210c M6 with
Cisco UCS VIC 15231

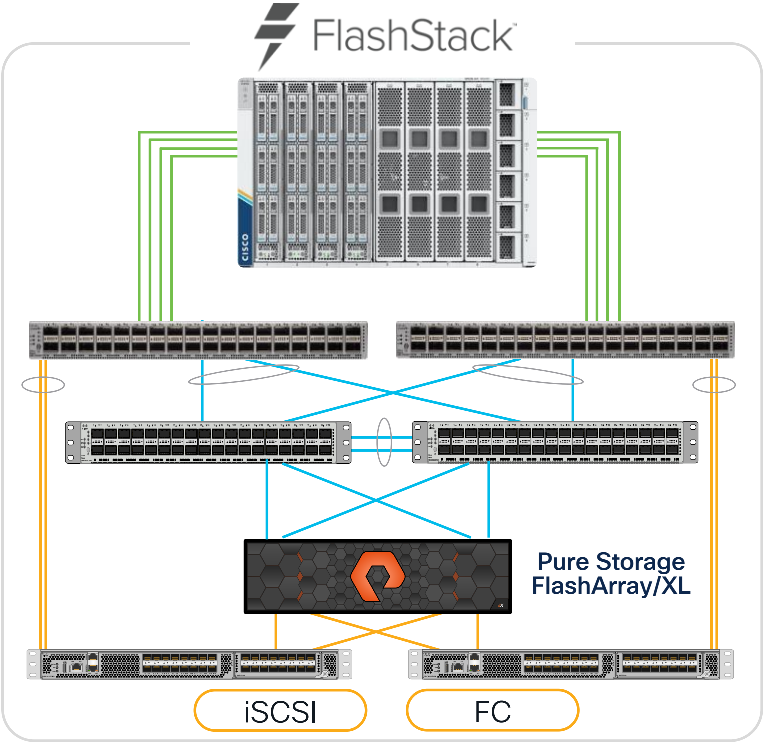
**Cisco UCS 6536
Fabric Interconnect**

Cisco Nexus 9336C-FX2

🟩 100 GE

🟦 100 GE

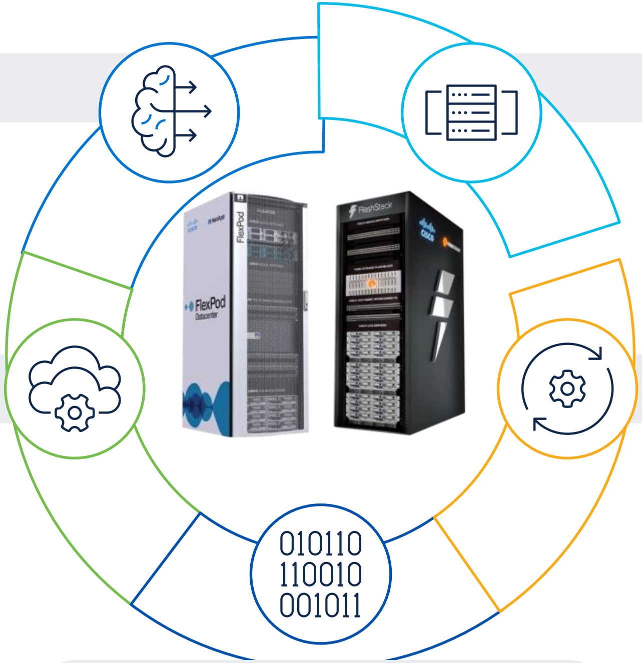
🟧 32Gbps FC



Youtube Video Link

Cisco Compute solution portfolio

Full Stack Solutions delivering best-in-class value to our customers



Mission critical workloads

SAP Microsoft CITRIX

ORACLE vmware Epic

AI and cloud-native ready

Red Hat OpenShift NVIDIA.

RANCHER BY SUSE OpenAI

PyTorch

Hybrid-cloud solutions

Google Cloud Microsoft Azure

aws

Automation and tools

ANSIBLE GitHub

HashiCorp Terraform

UCS sizer

Power calculator

TCO tool

Benchmarks

Data intensive solutions

cloudera splunk>

rubrik veeam

SCALITY COHESITY

cisco Live!

Keynote Deep Dives

Wednesday
10:30am –11:30am

Experiences Amplified: How AI Can Fuel Better Employee and Customer Experiences

Level 1
Room 106

Smart, Secure, Seamless: Transforming Experiences with Next-Generation Networking

Level 2
Room 204

Harness a Bold New Era: Transform Data Centre and Service Provider Connectivity

Level 2
Room 203

Securing User to Application and Everything in Between

Level 2
Melbourne Room 2

Unlocking Digital Resilience through Unified Observability

The HUB
Centre Stage

#CiscoLiveAPJC

Complete Your Session Evaluations

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

Complete your surveys in the **Cisco Live mobile 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 DevNet, Capture the Flag, and Walk-in Labs
- Visit the On-Demand Library for more sessions at www.CiscoLive.com/on-demand

Thank you

cisco Live!

GO **BEYOND**

#CiscoLiveAPJC