AI-Ready Data Centers – Introducing Hyperfabric and AI Pods

Qiese Dides - SE, Cloud & AI Infrastructure

Ken Dieter - SE, Cloud & AI Infrastructure

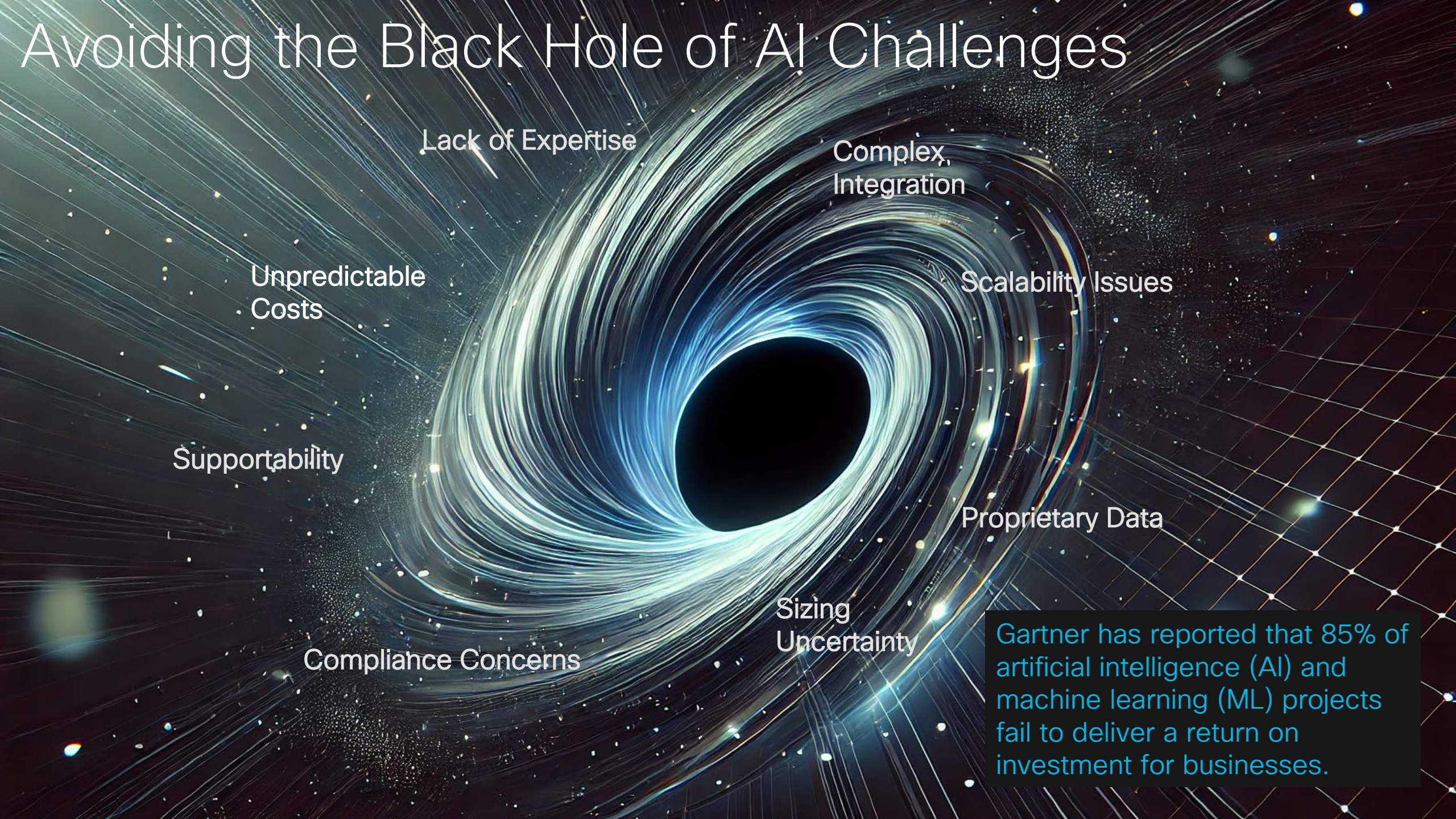
December 3rd, 2024

Disclaimer

Some of the features described herein remain in varying stages of development and will be offered on a when-and-if-available basis.

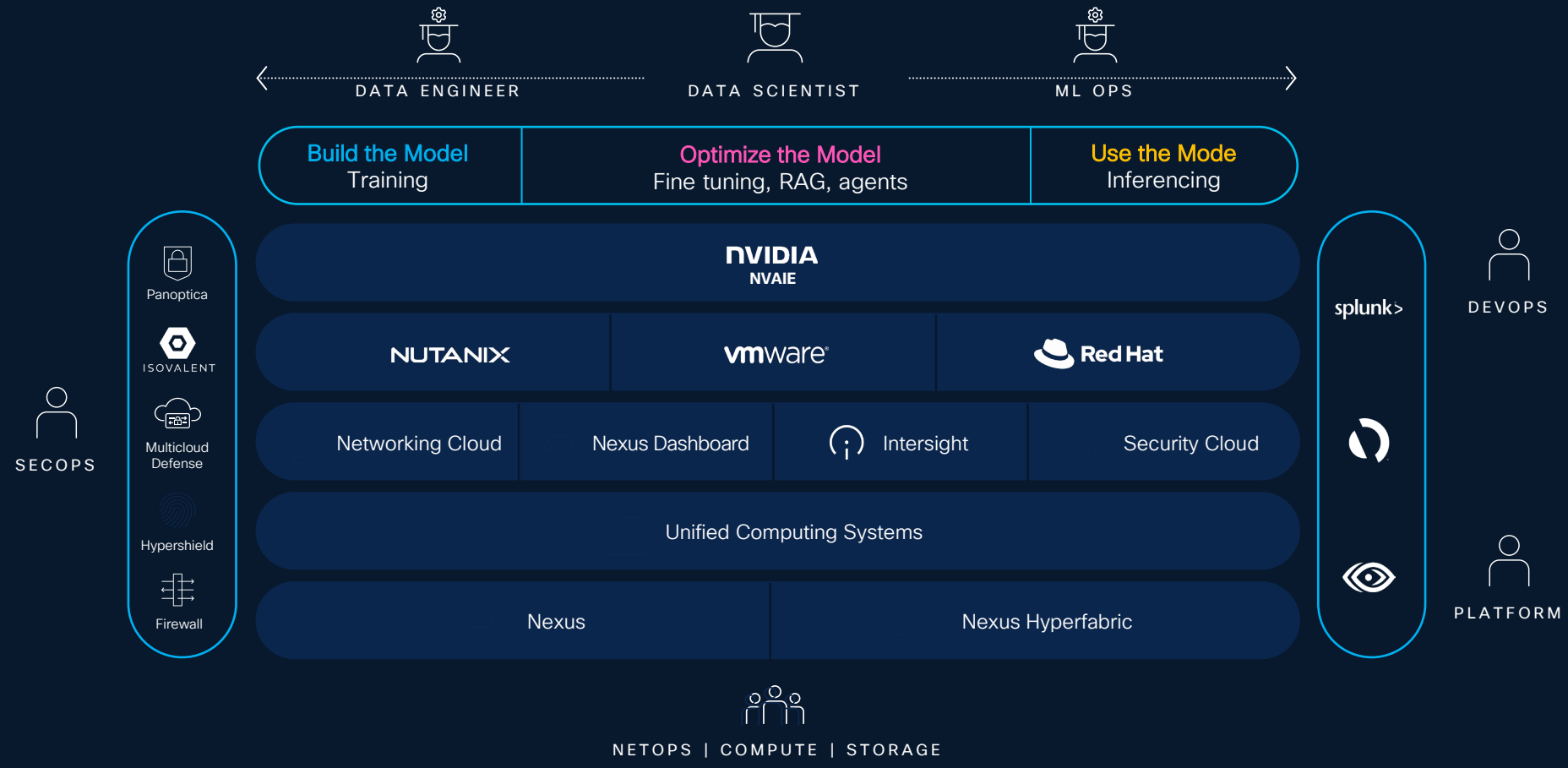
This roadmap is subject to change at the sole discretion of Cisco, and Cisco will have no liability for delay in the delivery or failure to deliver any of the products or features set forth in this presentation.

Avoiding the Black Hole of AI Challenges

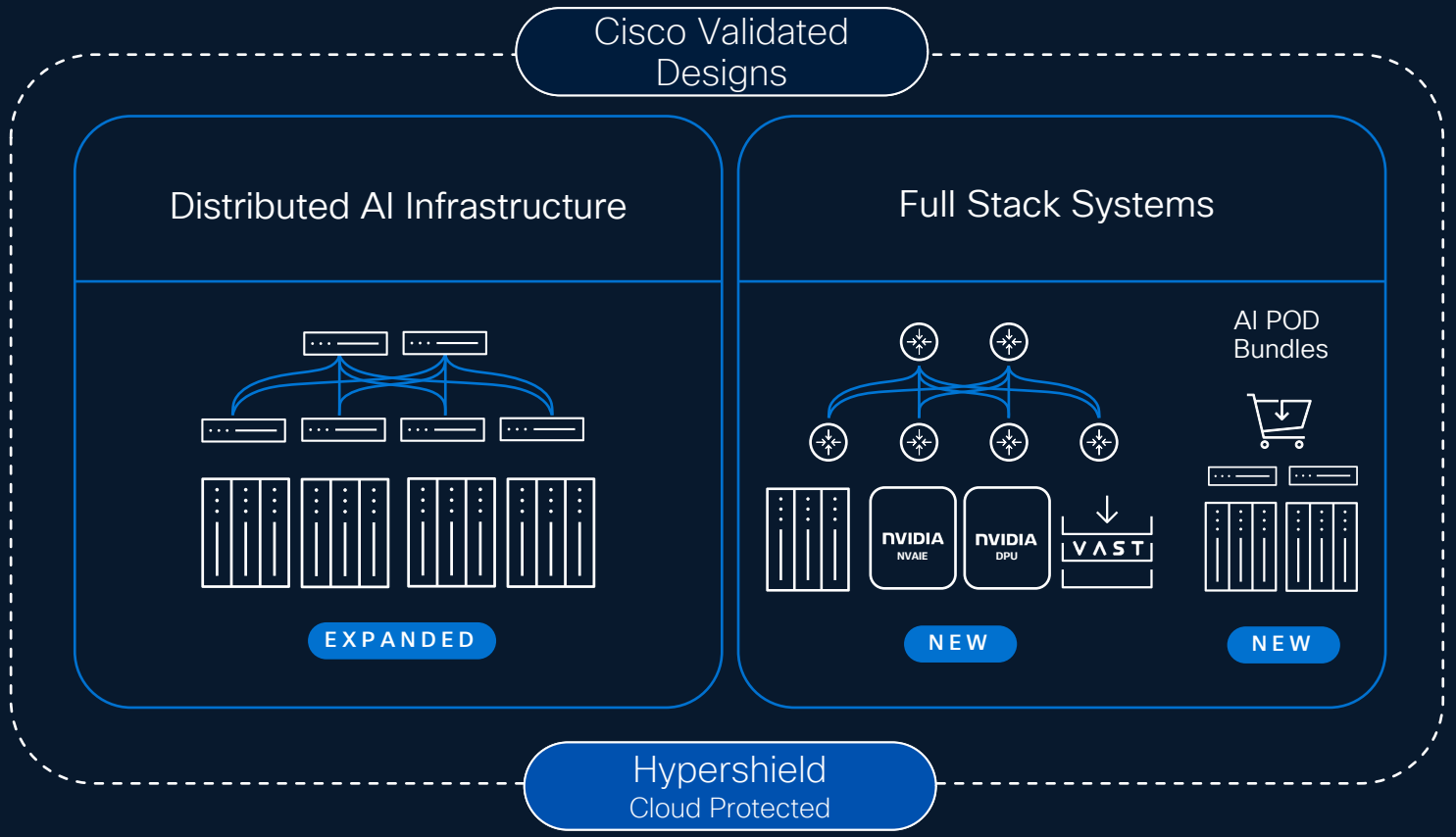


Gartner has reported that 85% of artificial intelligence (AI) and machine learning (ML) projects fail to deliver a return on investment for businesses.

Generative AI – Full Stack System



Cisco AI Ready Data Center



Cisco Validated Designs for the Data Center

Solutions to Simplify and Automate AI Infrastructure

TAILORED TO SPECIFIC USE CASES	MLOPS	GEN AI MODELING	AI NETWORKING	NEW DEPLOYMENT PLAYBOOKS
NVIDIA AI Enterprise	MLOps with Red Hat OpenShift AI	Generative models (GANs, VAEs)	Nexus 9000, Cisco Optics and Intel Gaudi 2	Large language models (GPT3, BERT, T5)
GPT-in-a-Box with Cisco Compute Hyperconverged	Gen-AI with Cloudera Data Platform	Gen AI Inferencing with RedHat OpenShift AI with Intel AI Enterprise	Blueprint for AI Networking on Nexus 9000 and Nexus Dashboard	Computer vision models (ResNet, EfficientNet, YOLO)

CONVERGED INFRASTRUCTURE	GPU/CPU
FlashStack, FlexPod	NVIDIA, intel, AMD

AI PODs

Simplified Orderability

Faster time to value with pre-configured bundles

ORDERABLE TODAY!!

Deploy AI with confidence

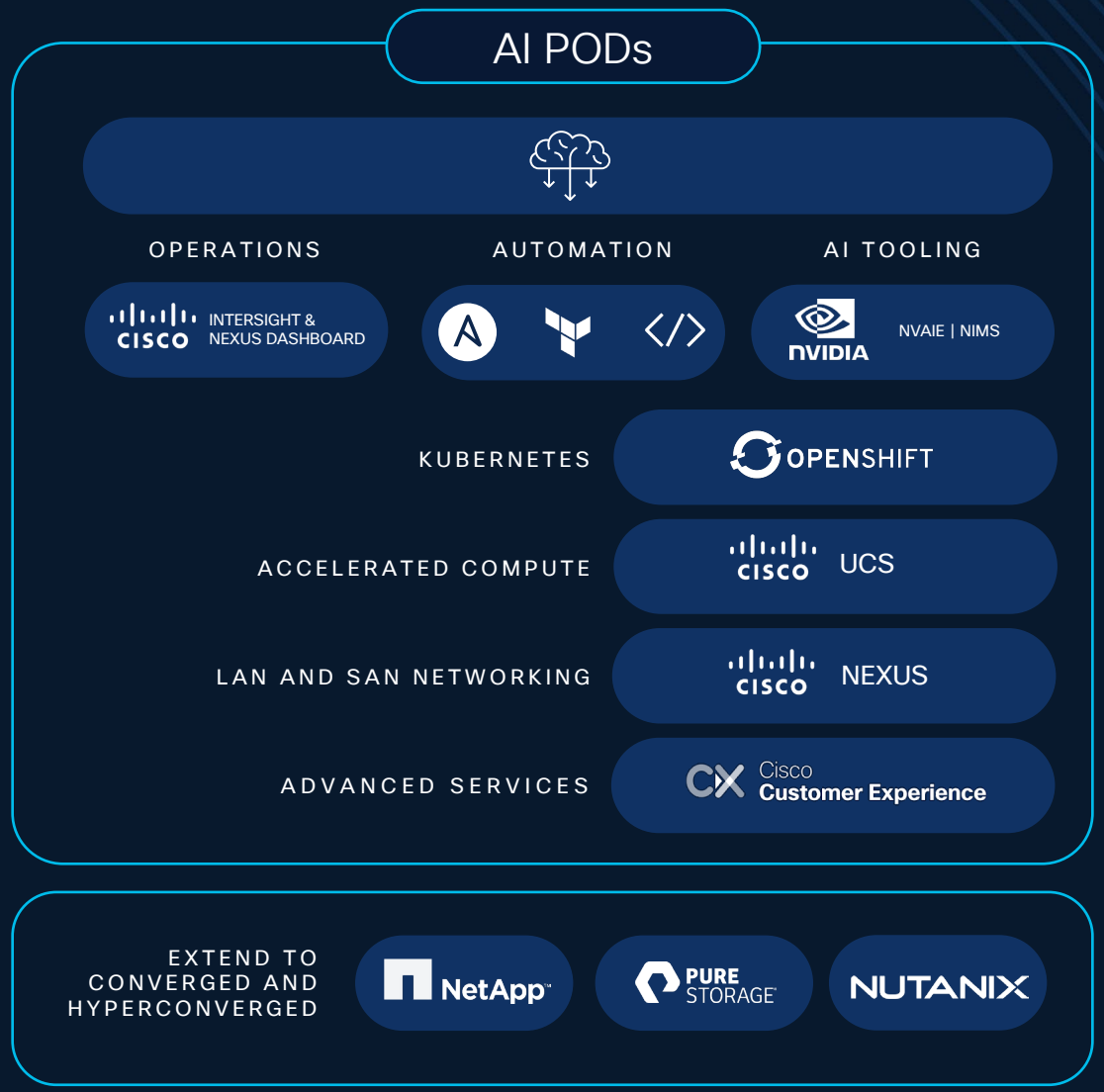
Orderable, validated AI-Ready infrastructure stacks

Fully supported stack including Cisco and 3rd party components

AI Advisor tool for configuration guidance

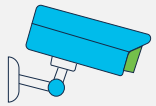
COMING SOON

Cisco AI-Ready Infrastructure Stacks



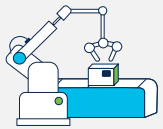
Edge inferencing provides several benefits

Gartner predicts that by 2025, 75% of enterprise-generated data will be created and processed at the edge.



Security and facilities

IoT

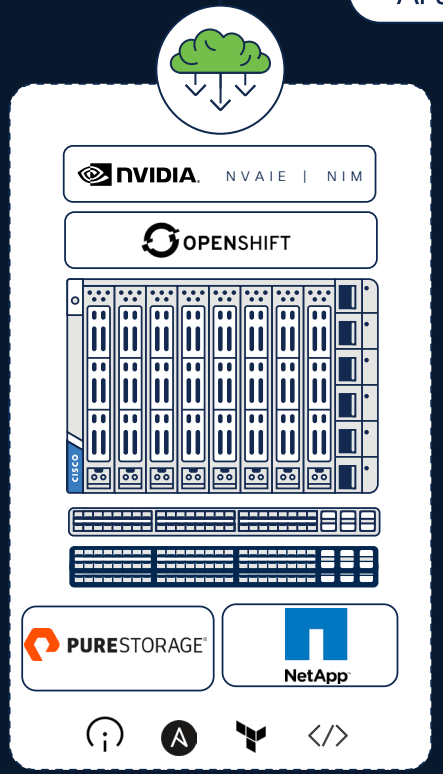


Industrial automation

Energy management

In-store analytics

Patient monitoring



UCS X-Series 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

Cisco AI PODs

Typical use case	Edge Inferencing (7B-13B Parameter)	RAG Augmented Inferencing (13B-40B+ Parameter)	Large-Scale RAG Augmented Inferencing	Scale-Out Inferencing Cluster (Inferencing Multiple Models)
Hardware specification	**Small** 1x X210C compute node • 2x Intel 5th Gen 6548Y+ • 512 GB System Memory • 5x 1.6 TB NVMe drives • 1x X440p PCIe • 1x NVIDIA L40S	**Medium** 2x X210C compute nodes • 4x Intel 5th Gen 6548Y+ • 1 TB System Memory • 10x 1.6 TB NVMe drives • 2x X440p PCIe • 4x NVIDIA L40S	**Large** 2x X210C compute nodes • 4x Intel 5th Gen 6548Y+ • 1 TB System Memory • 10x 1.6 TB NVMe drives • 2x X440p PCIe • 4x NVIDIA H100 NVL	**Scale-Out** 4x X210C compute nodes • 8x Intel 5th Gen 6548Y+ • 1.5 TB System Memory • 12x 1.9 TB NVMe drives • 4x X440p PCIe • 8x NVIDIA L40S

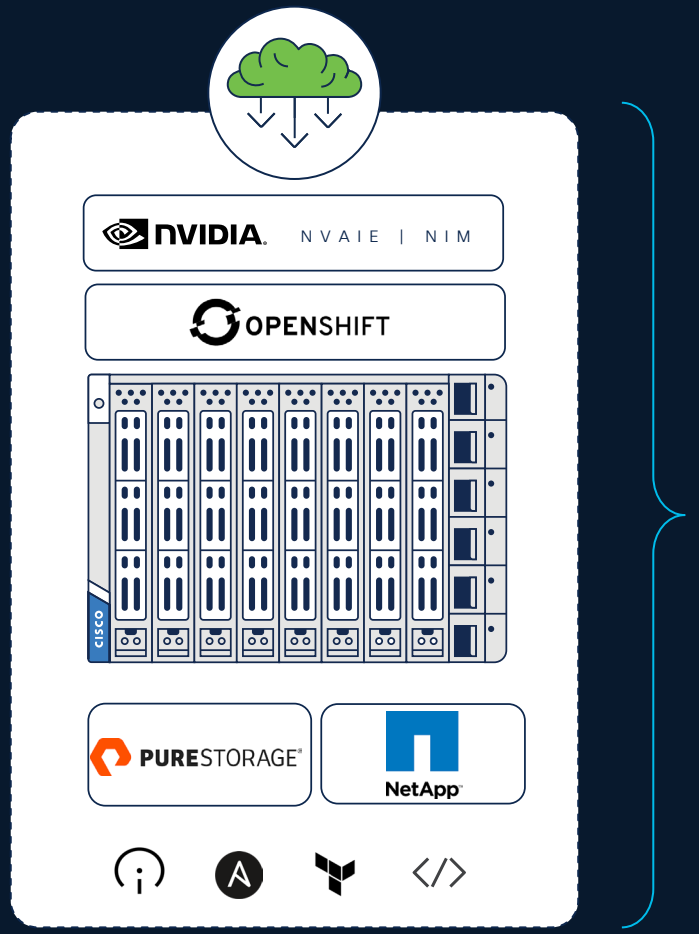
Performance and Scale

Inferencing Suite

Cisco AI PODs

Inferencing Suite

- Large language models
- AI tooling
- Kubernetes
- Accelerated compute
- Converged infrastructure
- Automation



Edge, RAG, Large-Scale RAG, Scale Out Inferencing Cluster

Components

Chassis

- Cisco® UCS X-Series Direct with 9108 100G Fabric Interconnect
  - X-Fabric Modules **(*FREE!)**

Modular nodes

- (1,2,2,4) x210c Compute Node
  - (2) Intel® 5th Gen 6548Y+ CPU
  - (512 GB, 1 TB, 1TB, 1.5 TB) of System Memory
  - (5,10,10,12) 1.6 TB High-Performance NVMe drives
- (1,2,2,4) x440P PCIe Node **(*FREE!)**
  - (1,4,8,4) L40S, H100 GPU

Software

- Cisco Intersight® Advantage
- NVIDIA NVAIE Essentials Subscription
- RedHat OpenShift Container Platform License

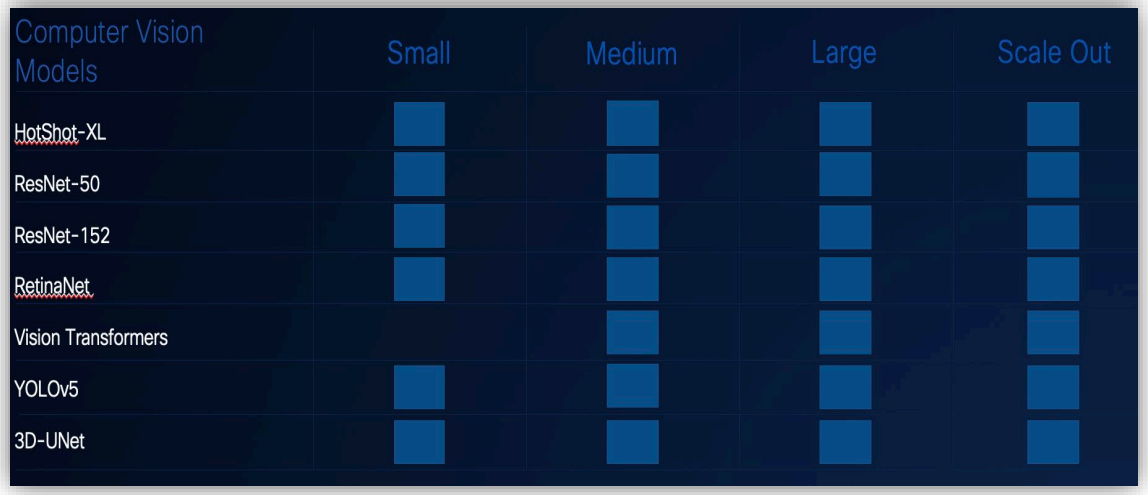
Services

- *AI Project Quick-Start Service

Add-on

- CI Storage solutions in partnership with Pure and NetApp

Compatibility Matrix



Supported
Projected

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.

Compute AI Portfolio

Address AI workloads with visibility, consistency, and control

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

Build the model
Training

Optimize the model
Fine-tuning and RAG

Use the model
Inferencing

UCS Dense GPU Servers

UCS Blade (w/GPU Expansion) and Rack

Dense compute for demanding AI

Full stack AI with compute and networking

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

8x NVIDIA HGX H100 or H200 GPUs or

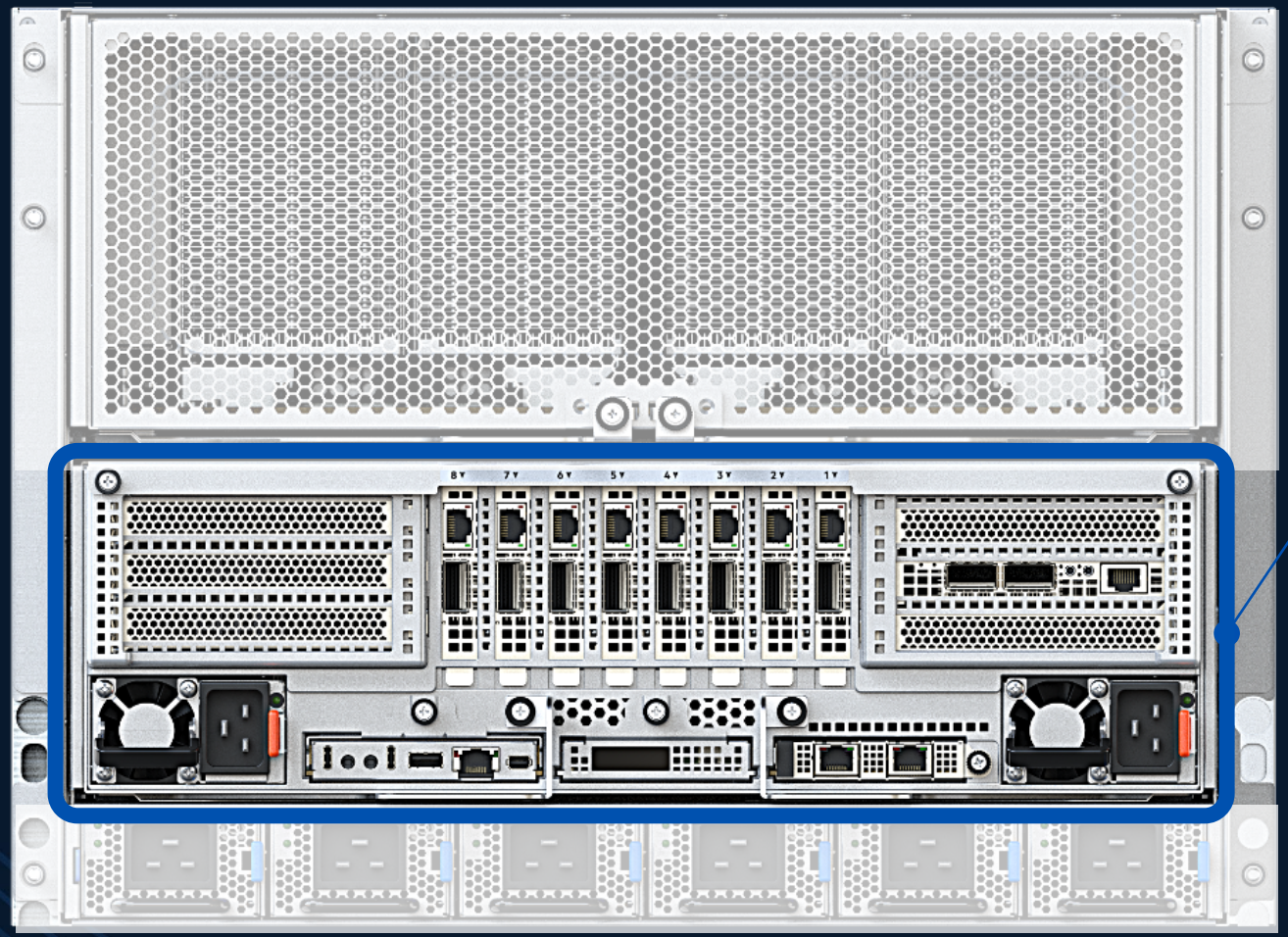
8x AMD MI300X GPUs

2x AMD 4th Gen or 5th Gen EPYC™ Processors

UCS C885A M8 Modular Sled Design

NVIDIA HGX Architecture

UCS C885A M8 CPU Tray



CPU Tray

CPU & Memory

2x
AMD 9554 (Genoa) CPUs
64 cores & up to 3.75GHz
360W/CPU

or

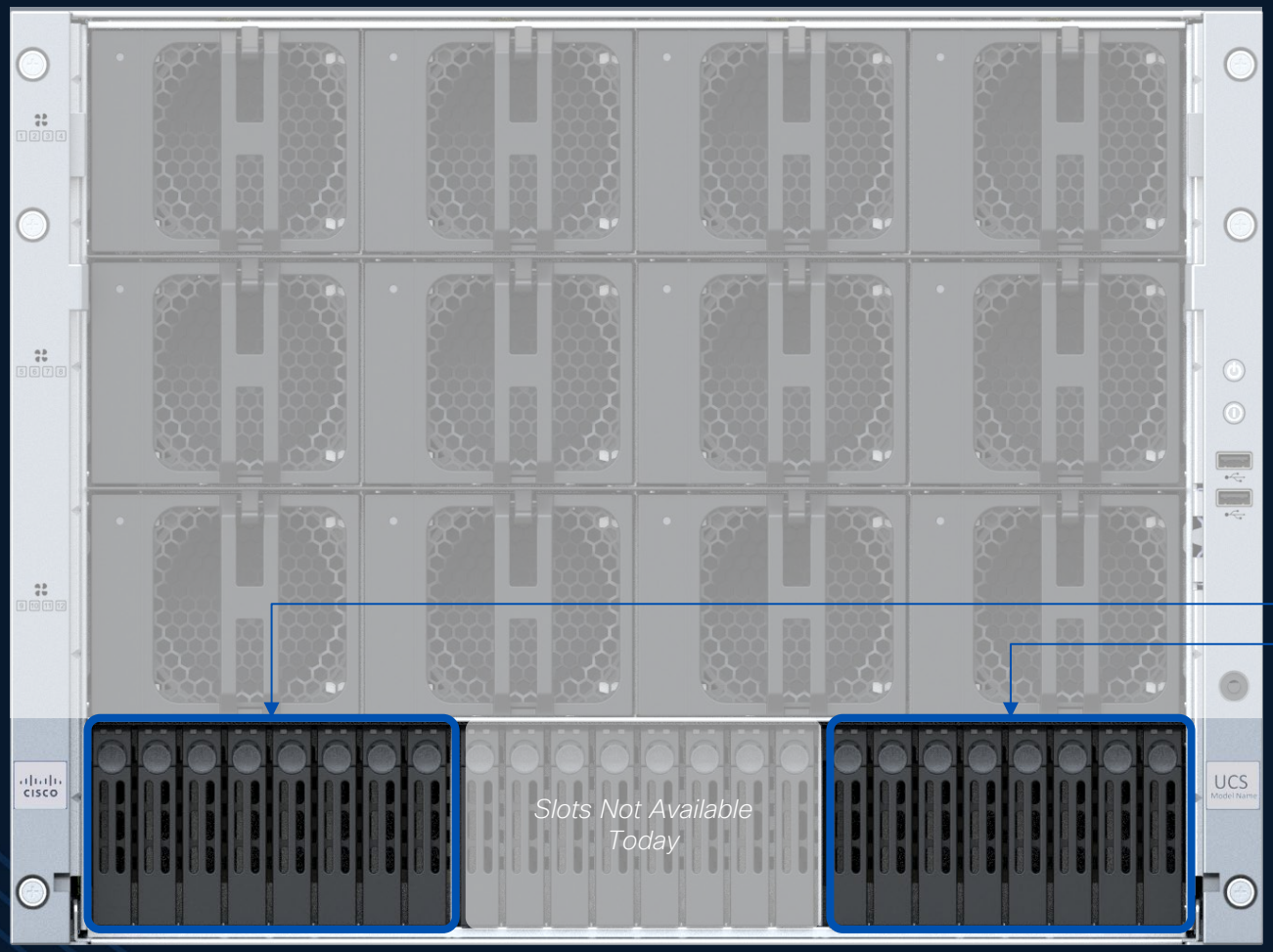
2x
AMD 9575F (Turin) CPUs
64 cores & up to 5GHz
400W/CPU

24x
96GB
DDR5 RDIMMs
Up to 6,000 MT/S

*128GB DIMM **option for some fixed configs** coming **soon***

Server Rear View

UCS C885A M8 Local Storage



1x

Internal M.2 NVMe Drive

Use-Case: Boot

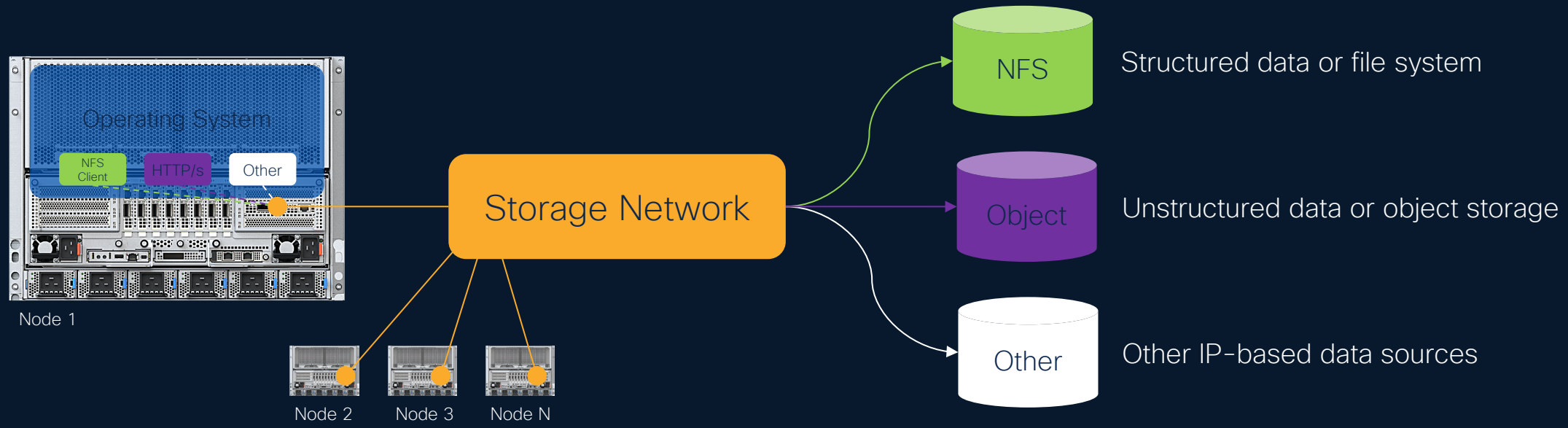
16x

External 2.5” U.2 NVMe SSD Drives

Server Front View

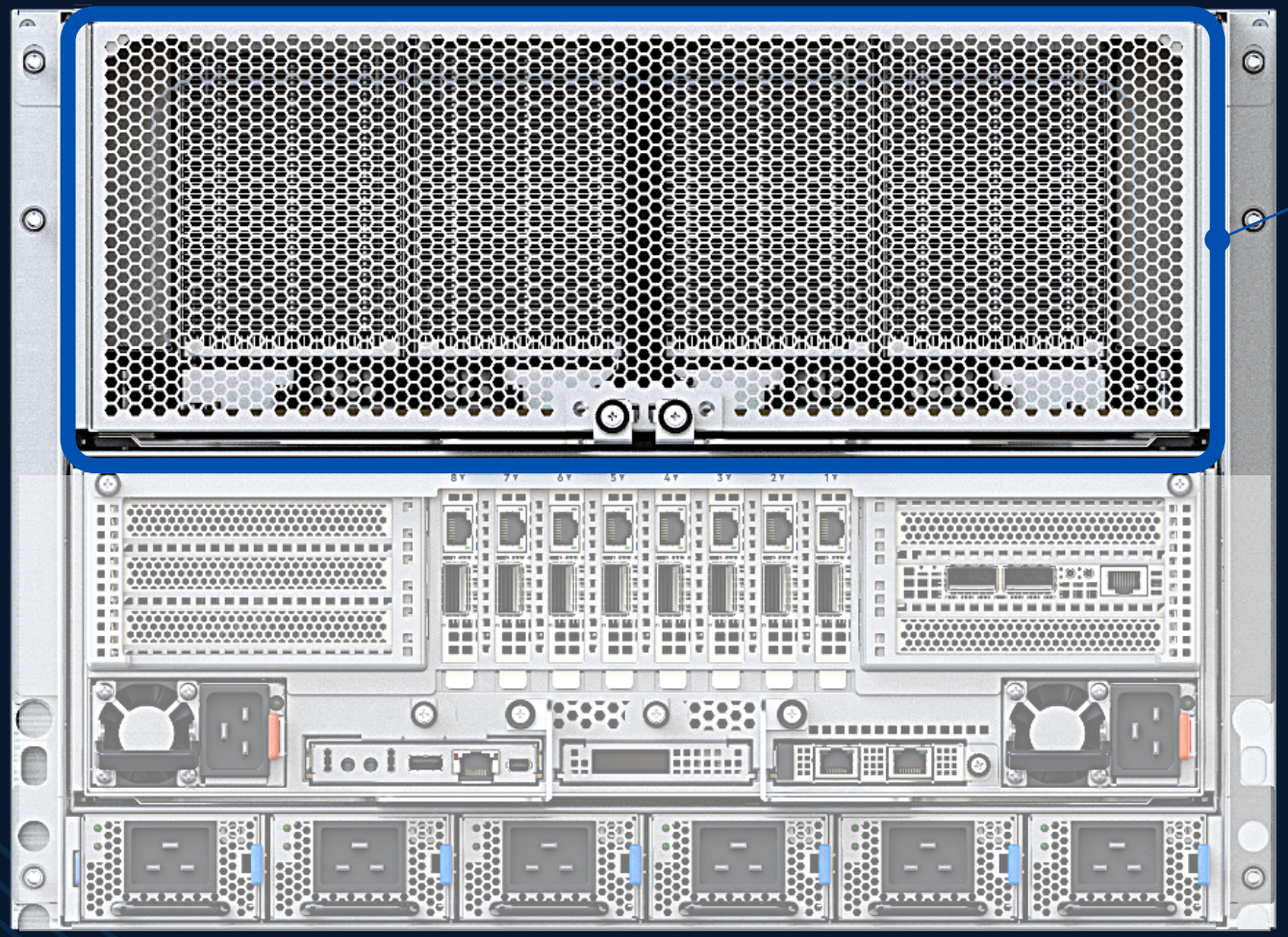
UCS C885A M8 External Storage Support

Common examples for AI/ML training use-cases include NFS and Object storage



Note: Most customers leverage a dedicated storage network, but storage traffic can be aggregated with other north-south networks as well

UCS C885A M8 GPU Tray



Server Rear View

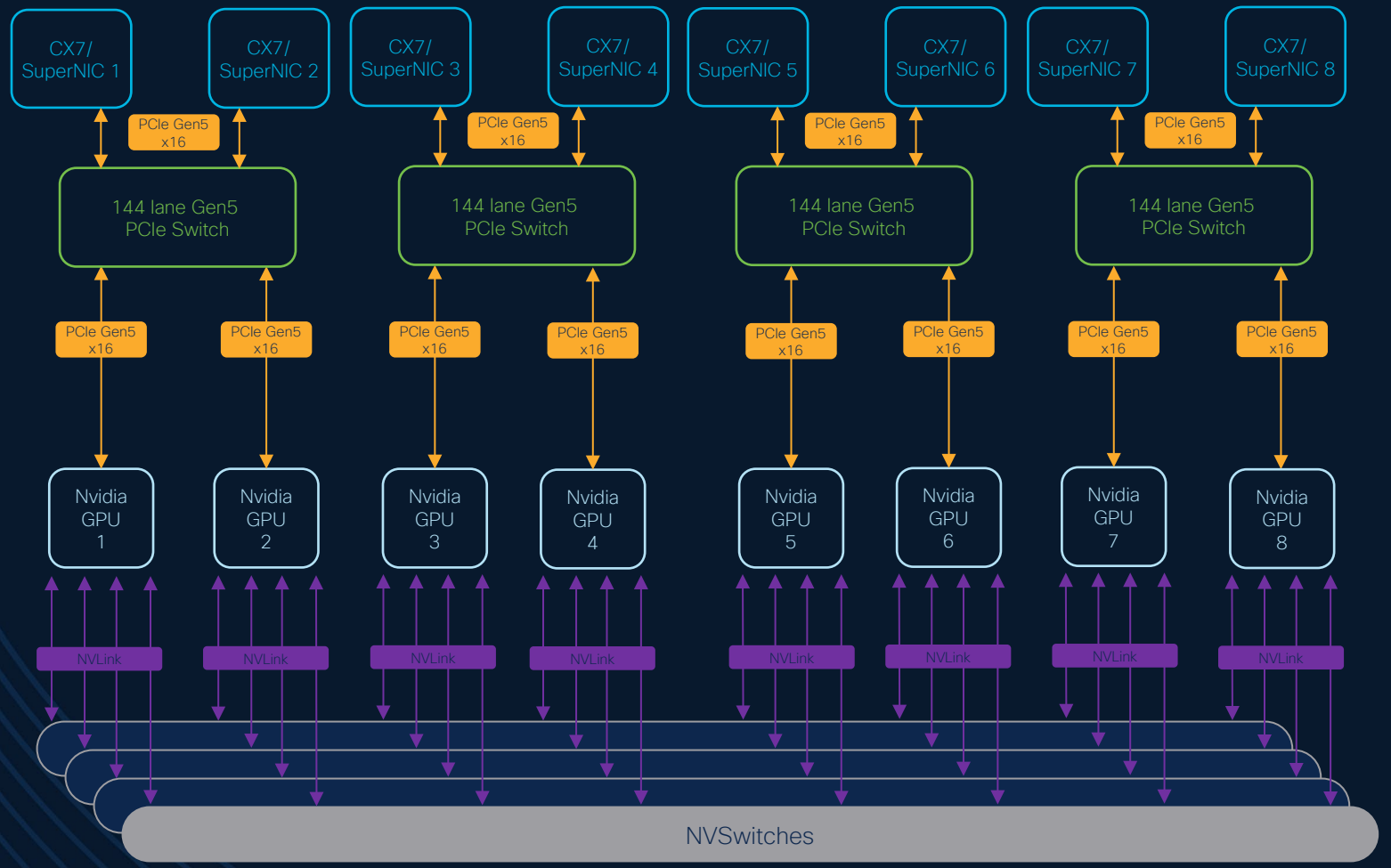
GPU Tray

8x
Nvidia
H100, H200 or B200A GPUs
700W/GPU

or

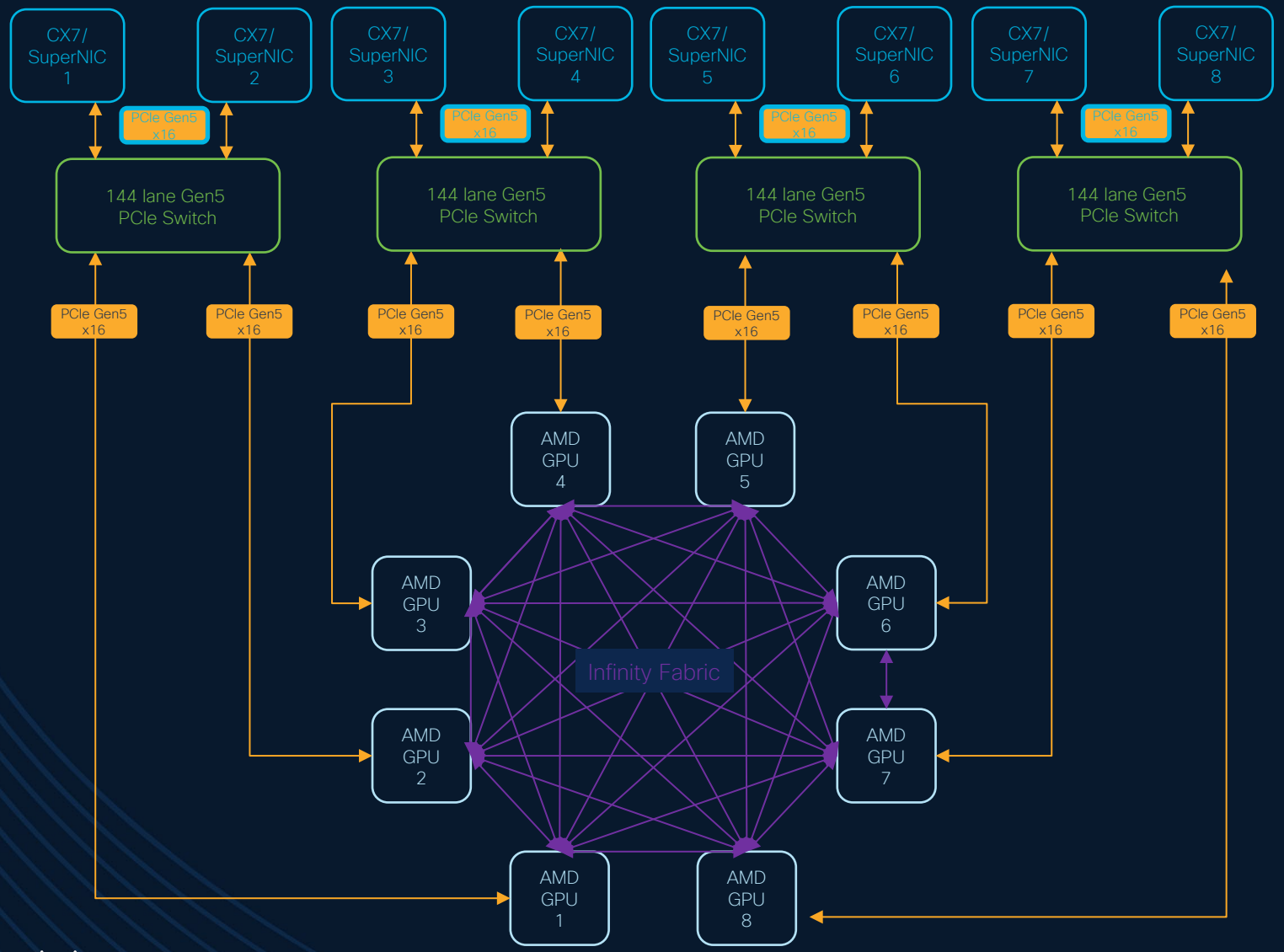
8x
AMD
MI300X GPUs
750W/GPU

UCS C885A – Nvidia GPU Connectivity



- 8x Nvidia H100/H200 SXM5 Tensor Core GPUs
- Each H100/H200 GPU has multiple NVLink ports and connects to all four NVSwitches
- 4 x fully non-blocking NVSwitches that connect all 8 GPUs
- NVLink bidirectional speed of 900GB/s between any pair of GPUs in the same node
- Each H100/H200 GPU also has a dedicated NIC/SuperNIC connected via PCIe Gen5 x16 for GPU-to-GPU connectivity across nodes

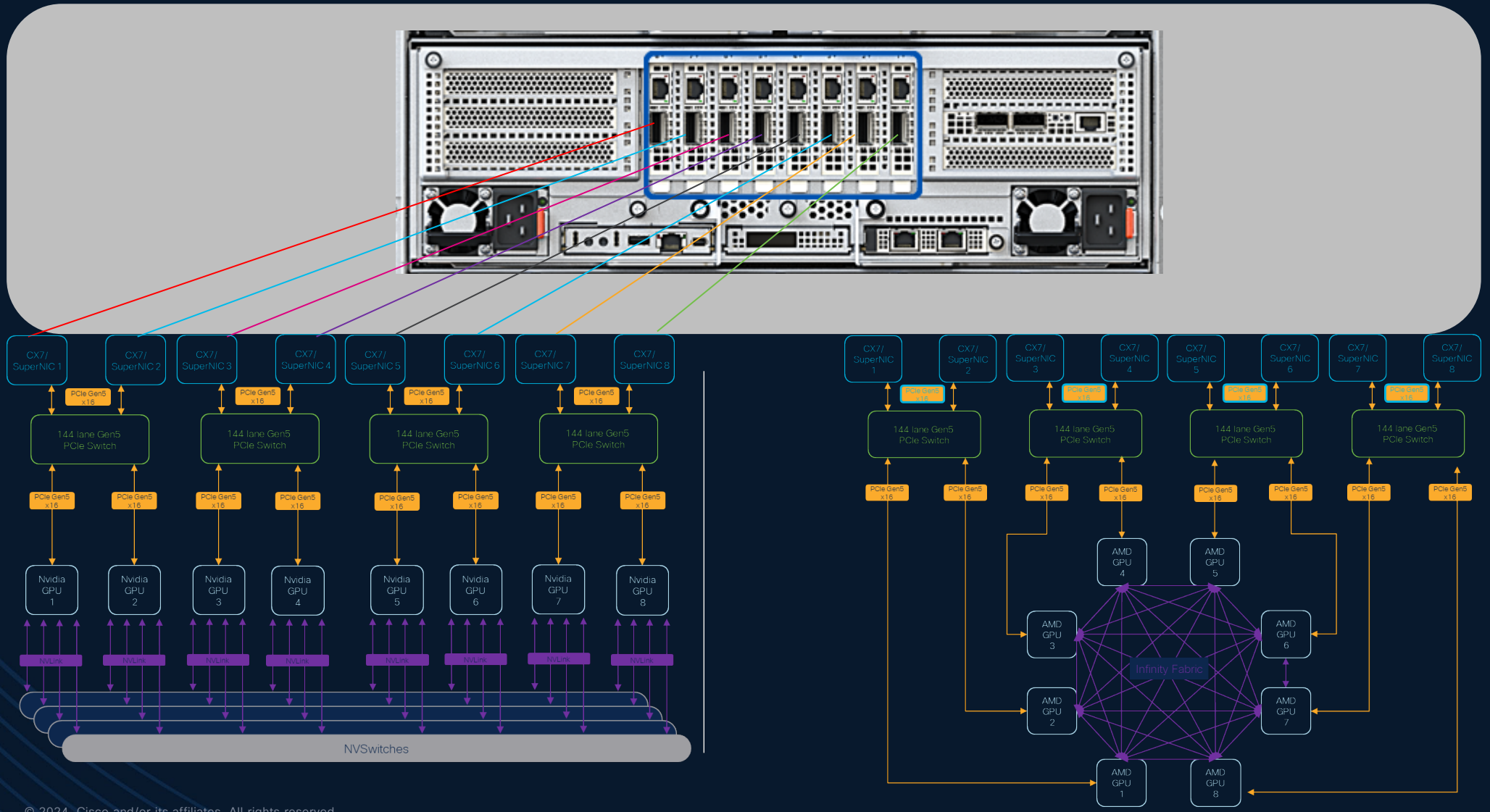
UCS C885A AMD GPU Connectivity



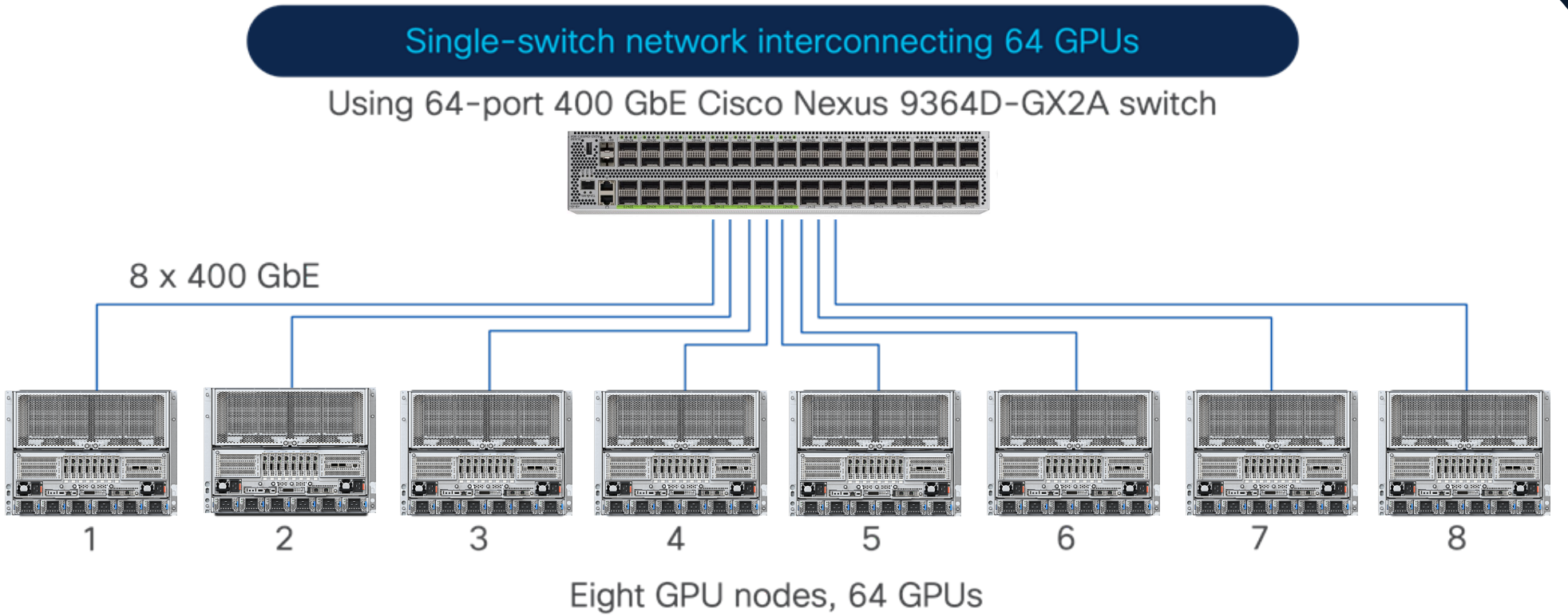
- 8x AMD MI300X OAM GPUs
- Each MI300X GPU has direct access to all other GPUs in full mesh topology over AMD Infinity Fabric mesh
- Bi-directional connectivity between each of the GPUs at over 128GB/s
- Each MI300XGPU also has a dedicated NIC/SuperNIC connected via PCIe Gen5 x16 for GPU-to-GPU connectivity across nodes

Cisco Networking Solutions for AI

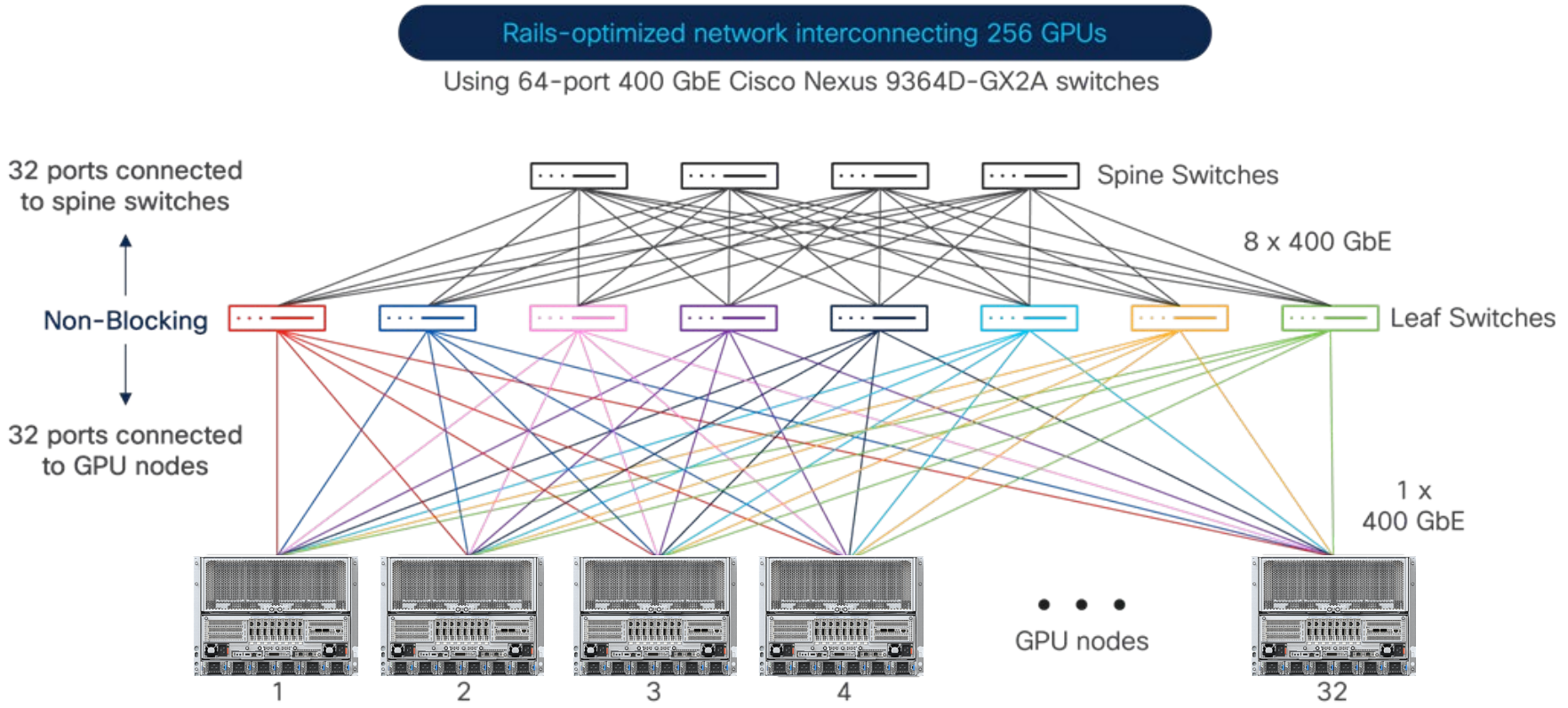
GPU-GPU Backend Network



Sample Inter-GPU Backend Network



Rails-optimized GPU Backend Fabric 256 GPUs

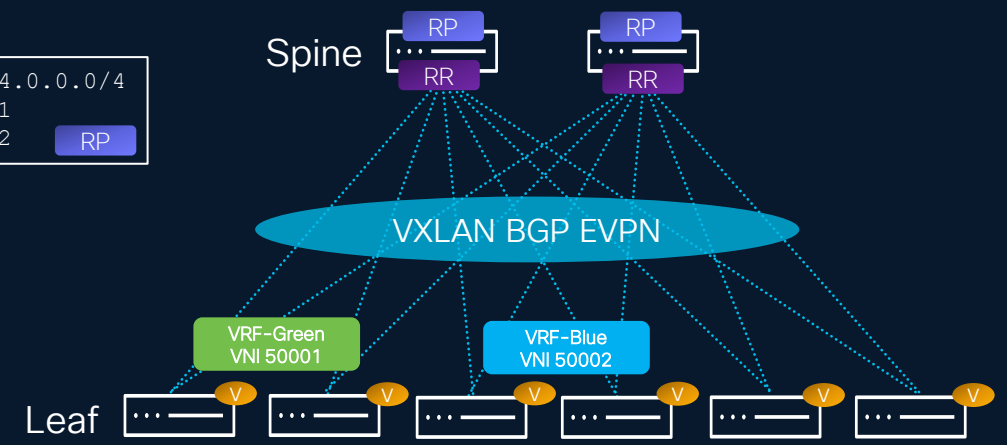


Network Configuration Snippet

```
ip pim rp-address 10.237.1.1 group-list 224.0.0.0/4
ip pim anycast-rp 10.237.1.1 10.255.255.101
ip pim anycast-rp 10.237.1.1 10.255.255.102
```
RP

```
router bgp 65001
  router-id 10.1.0.5
neighbor 10.1.0.1
    remote-as 65001
    update-source loopback0
    address-family l2vpn evpn
      send-community
      send-community extended
      route-reflector-client
```
RR

```
router bgp 65001
  router-id 10.1.0.4
  neighbor 10.1.0.5
    remote-as 65001
    update-source loopback0
    address-family l2vpn evpn
      send-community
      send-community extended
  vrf VRF-Green
    address-family ipv4 unicast
      advertise l2vpn evpn
    address-family ipv6 unicast
      advertise l2vpn evpn
  vrf VRF-Blue
    address-family ipv4 unicast
      advertise l2vpn evpn
    address-family ipv6 unicast
      advertise l2vpn evpn
```
V




```
ip pim rp-address 10.237.1.1 group-list 224.0.0.0/4

router ospf 1
interface Ethernet1/50
  mtu 9216
  ip pim sparse-mode
  ip address 192.168.1.10/31
  ip router ospf 1 area 0.0.0.0
```
PIM Multicast
V

```
vrf context VRF-Green
  vni 50001
  rd auto
  address-family ipv4 unicast
    route-target both auto
    route-target both auto evpn
evpn
  vni 3010 l2
    rd auto
    route-target both auto
```
Layer 3 VNI
V

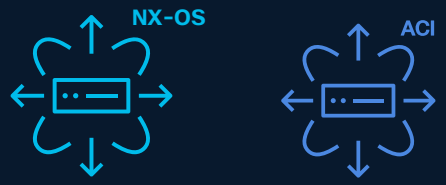
```
vrf context VRF-Blue
  vni 50002
  rd auto
  address-family ipv4 unicast
    route-target both auto
    route-target both auto evpn
evpn
  vni 3011 l2
    rd auto
    route-target both auto
```
V

```
Vlan 10
 vn-segment 3010
Vlan 11
 vn-segment 3011

Vlan 1000
 Layer 3 VNI
 vn-segment 50001
Vlan 2000
 Layer 3 VNI
 vn-segment 50002

interface Vlan10
  no shutdown
  vrf member VRF-Green
  ip address 192.168.10.254/24 tag 12345
  ipv6 address 2001::1/64 tag 12345
  fabric forwarding mode anycast-gateway

interface Vlan11
  no shutdown
  vrf member VRF-Blue
  ip address 192.168.11.254/24 tag 12345
  ipv6 address 2002::1/64 tag 12345
  fabric forwarding mode anycast-gateway

interface nve1
  source-interface loopback0
  host-reachability protocol bgp
  member vni 3010
    mcast-group 239.1.1.1
  member vni 3011
    mcast-group 239.1.1.1
  member vni 50001 associate-vrf
  member vni 50002 associate-vrf
fabric forwarding anycast-gateway-mac 0002.0002.0002
```
V
Layer 2 VNI
Layer 3 VNI
Map L2VNI to NVE
Associate L3VNI to NVE

Cisco Data Center Networking Portfolio

Nexus Dashboard

On-Premises Delivered



Powered by Nexus 9000 Series

Day 2 Ops | Visibility | Analytics | Troubleshooting | Compliance

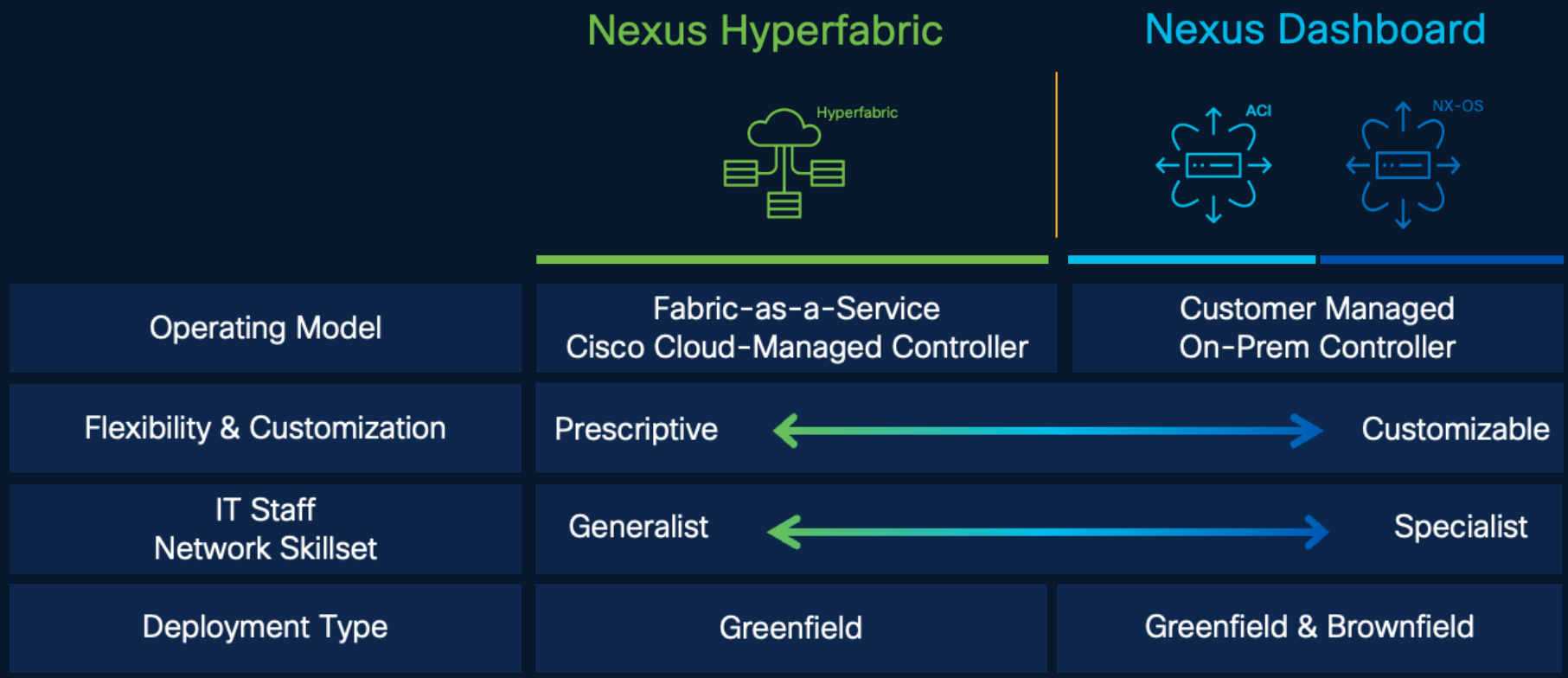
Nexus Hyperfabric

Cloud Delivered

Powered by Cisco 6000 Series

Delivering networking at speed from the cloud

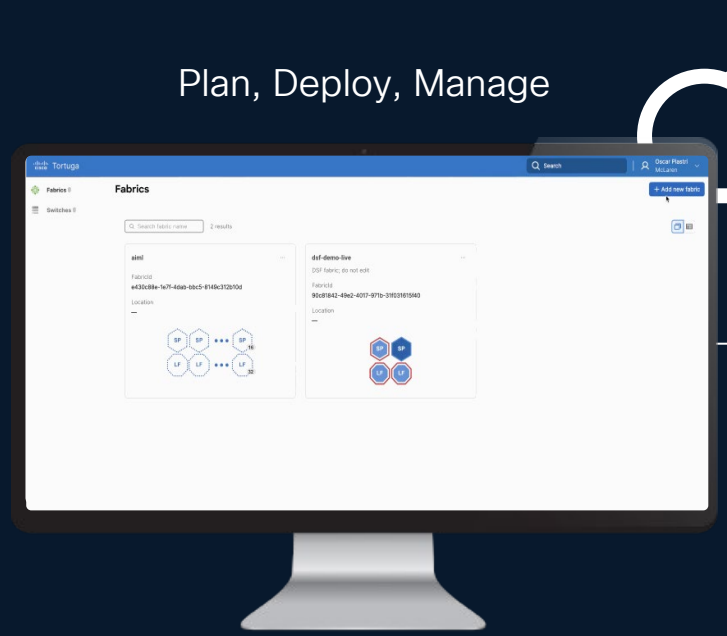
Cisco Data Center Networking Portfolio



Greenfield: new fabrics not being managed by Nexus Dashboard

How it Works

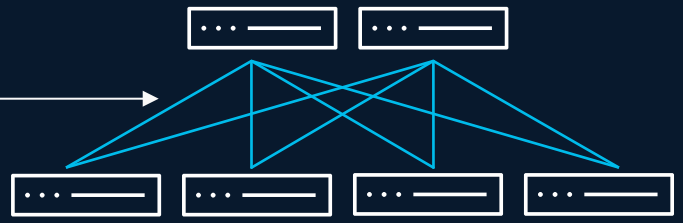
Plan, Deploy, Manage



https://hyperfabric.cisco.com

API

Cisco 6000 Series Switches



Plug-and-Play DC Fabrics
Self-Discovery / Standards-based
Always-on Telemetry
Assertions-based Monitoring

Purpose-built for predictable outcomes optimized for ease of use

Complete Lifecycle Experience

Day 0
Plan with Fabric Designer

Order
Configure network while equipment ships

Day 1
Plug in and Register to Fabric

Validate
Ensure Deployment matches Design

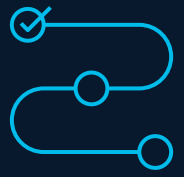
Day n
Monitor, Upgrade & Collaborate from **Anywhere**

Cloud-Managed Controller

- Scalable, globally distributed multi-tenant cloud service
- GUI, Mobile, and API access
- **Helping Hands App** for Smart Remote Hands visibility

Cloud-Managed Switch

- Cisco 6000 Series
- Boots from Cloud
- Full visibility & control from the cloud



Simplifying Operations

Design, deploy, and operate on-premises fabrics located anywhere

Eliminating Complexity

Easy enough for IT generalists, application, and DevOps teams

Through a shared responsibility model, Cisco manages the underlying configuration

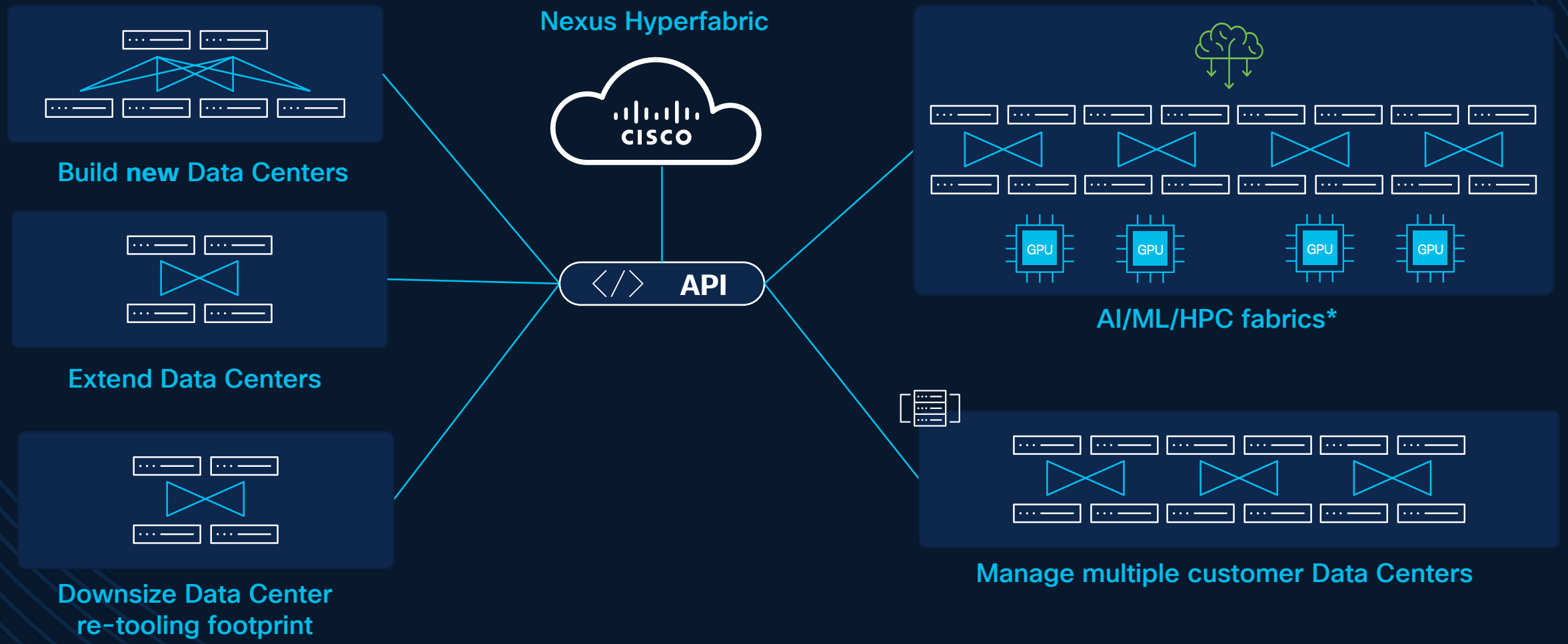
Integrated Vertical Stack

Outcome driven by a purpose-built vertical stack: hardware, software, cloud management, day 2 operations, and support

Use Cases & Architectures

Use Cases

Single global UI / API endpoint for all owned fabrics



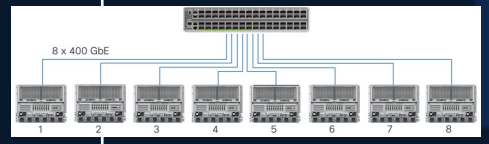
*Nexus Hyperfabric AI cluster Post 1.0 Release

Flexible Architectures

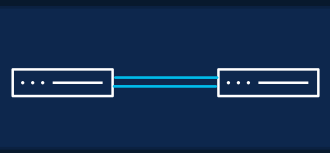
Deploy any fabric anywhere

Mesh / Spine-less Fabrics

"A Fabric of One"



2-Switch Fabric

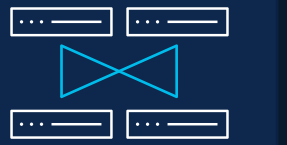


3-Switch Fabric

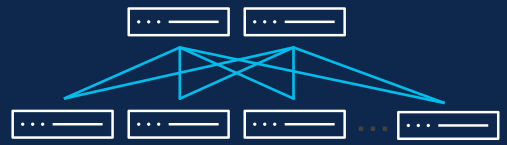
4-Switch Fabric

5-Switch Fabric

2 Spine, 2 Leaf

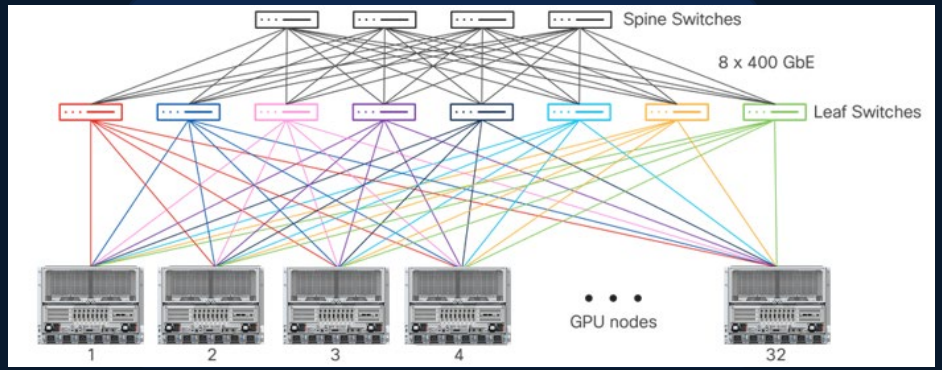


2- or 4-way Spine, 2-32 Leaf



Leaf-Spine DC Fabrics

8- or 16-way Spine, 32 Leaf



AI and HPC Fabric Pods

Cisco 6000 Switches

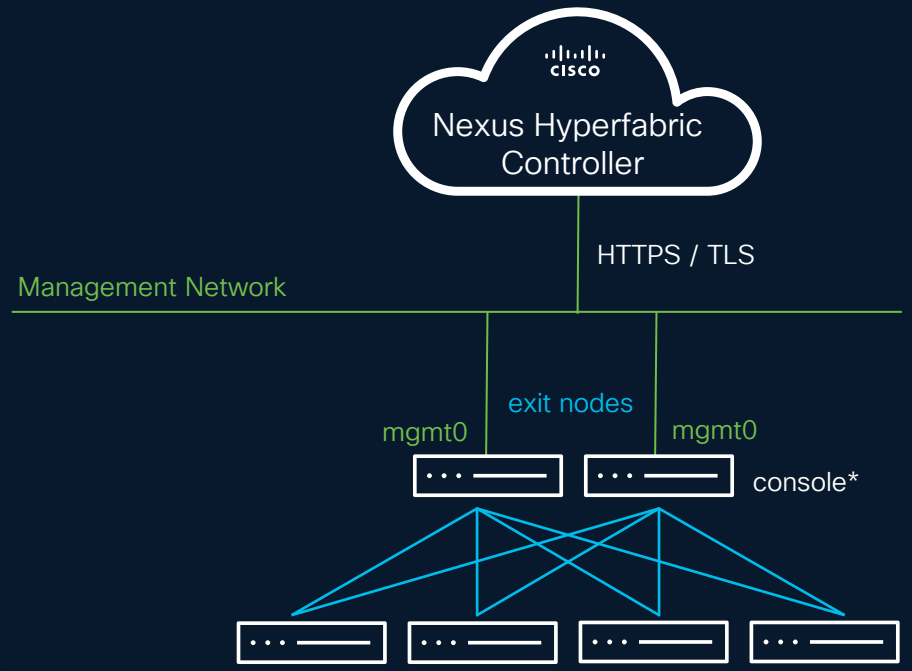
Leaf: HF6100-60L4D

- 4x 100/400GbE QSFP56-DD (16x 100G breakout)
- 60x 10/25/50GbE SFP56

Spine/Leaf: HF6100-32D

- 32x 100/400GbE QSFP56-DD
- 128x 100GbE via 400:100 breakout

Initial Setup and Cloud Connectivity



✓ Connect one or more exit nodes using mgmt0 for cloud access and authentication

✓ Register ("claim") switches online within their organization
- Grab claim token on switches via console
- Apply token on Nexus Hyperfabric controller
- Use "adjacent switch claim" for the remaining switches

Alternative: auto-claim via USB key with pre-list of serial numbers

✓ Assign ("bind") each physical switch to its logical blueprint

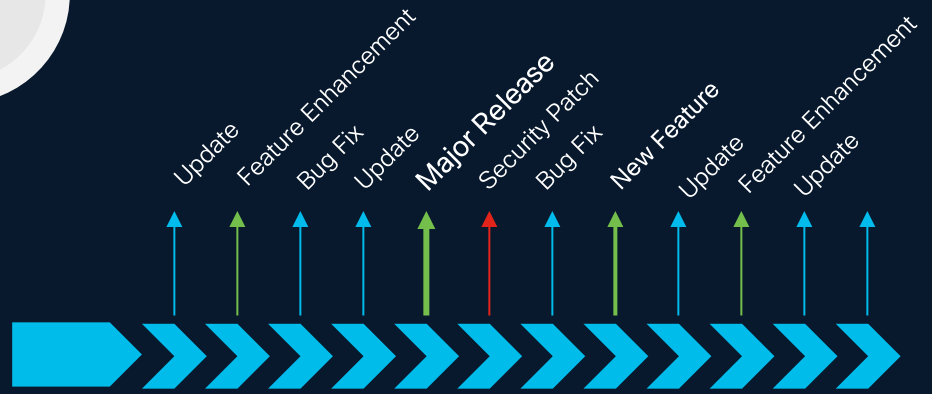
*For IP, proxy, DNS, and token configuration

Software Lifecycle Management

Cloud SaaS controller:

Continuous delivery model: always up-to-date

- Continuous delivery of new features and software updates to the production cloud service.
- No user testing or software maintenance required



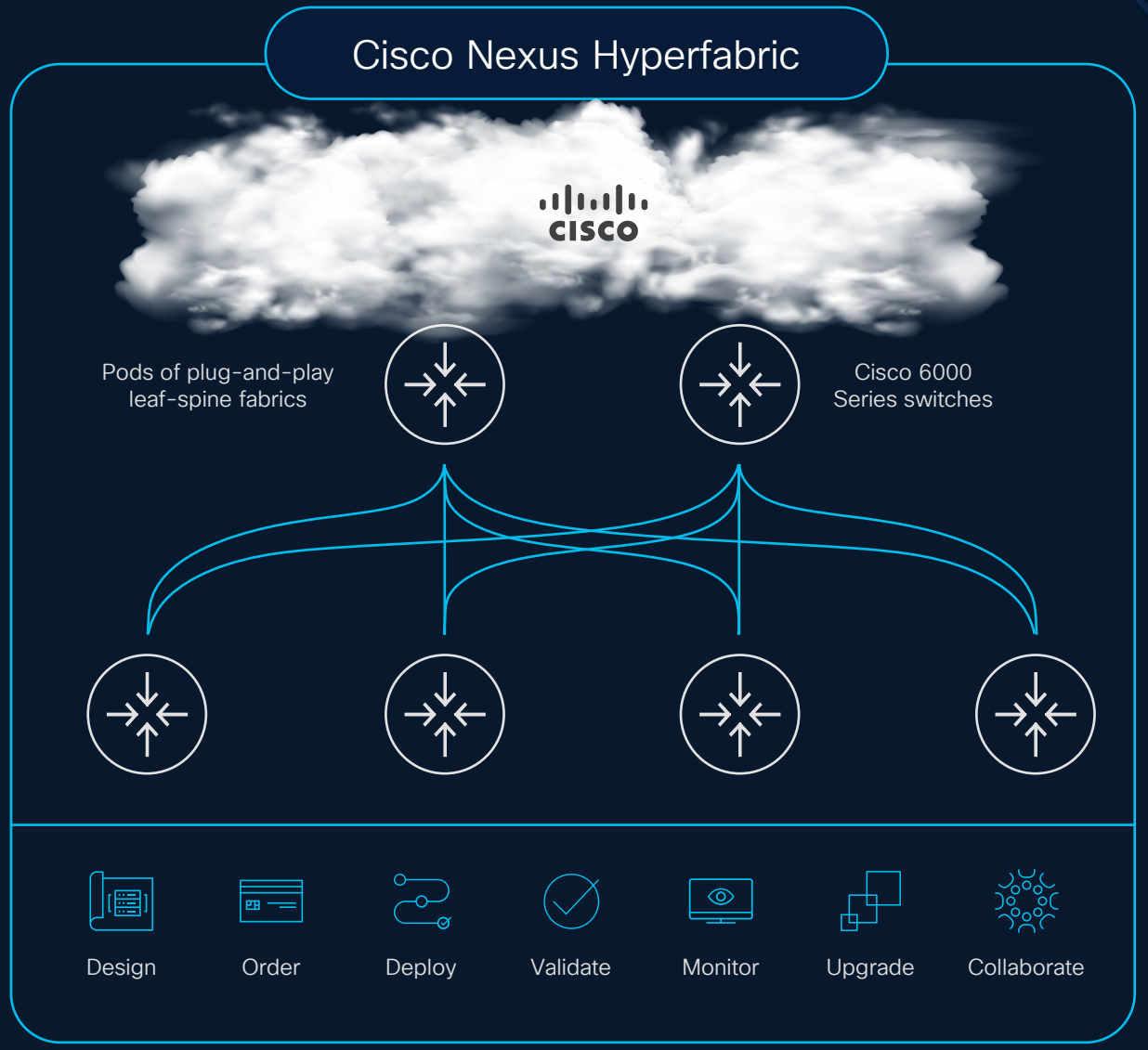
On-prem switch software

Cloud-delivered Software Upgrades: User-Driven Update Schedule

- Schedule firmware updates
- Software rollback support
- Intelligent sequencing of fabric upgrades

Cisco Nexus Hyperfabric

- Design, deploy and operate on-premises fabrics located anywhere
- Easy enough for IT generalists, application and DevOps teams
- Outcome driven by a purpose-built vertical stack



Hyperfabric AI

- AI Fabrics based on Cisco 6000 switches
- Cloud-Managed Controller
- Cisco UCS servers with high GPU density
- Optics

- AI Enterprise software
- NIM inference microservices
- GPUs starting with the H200 NVL
- BlueField DPU (N/S) and SuperNIC (E/W)

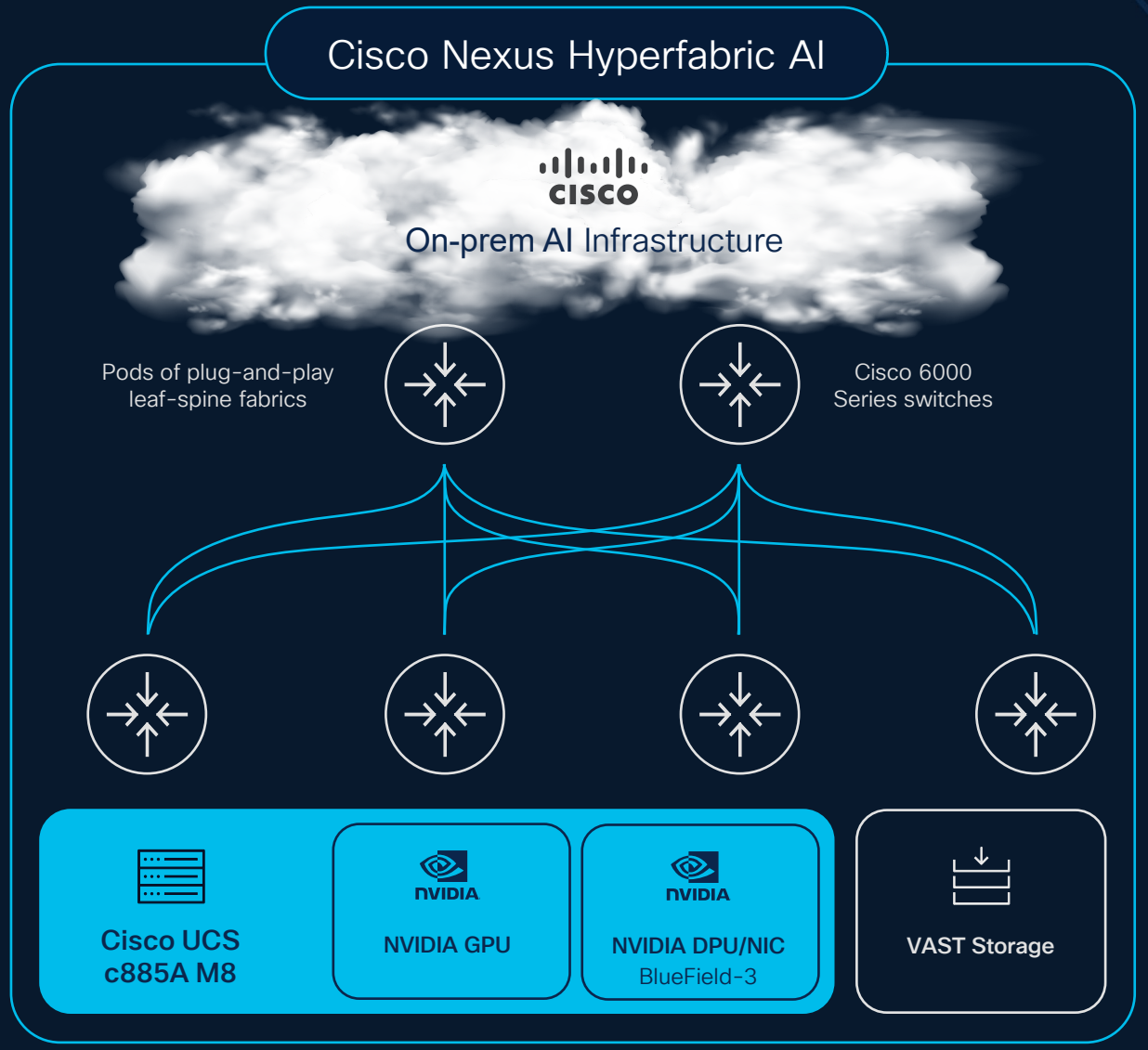
- Optional UCS to host VAST storage
- Unified storage, database, and a data-driven function engine built for AI

Based on NVIDIA HGX and/or MGX architectures (in progress)

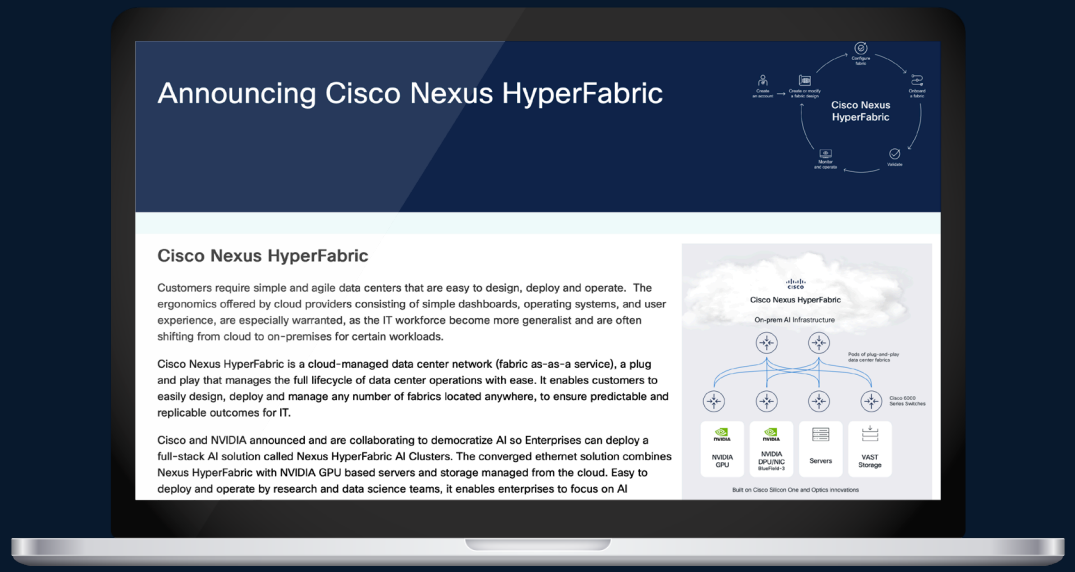
ORDERABLE 2025

Cisco Nexus Hyperfabric AI

- High-performance Ethernet
- Cloud-managed operations
- Unified stack including NVAIE
- AI-native operational model
- Democratize AI infrastructure
- Visibility into full stack AI

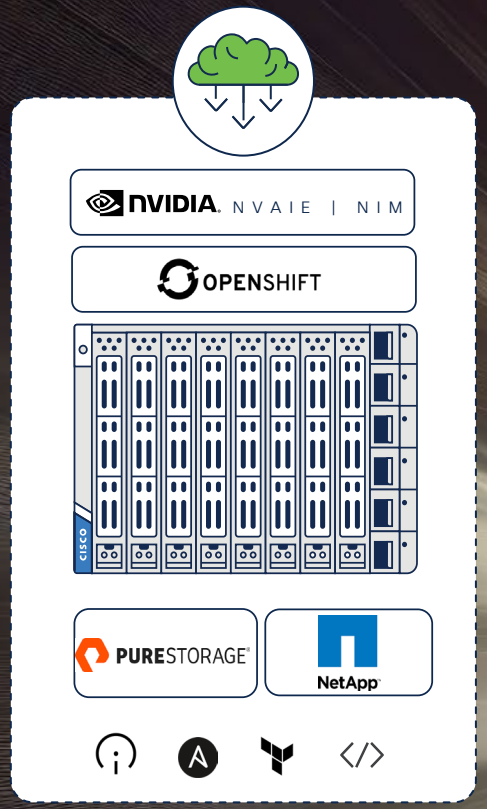


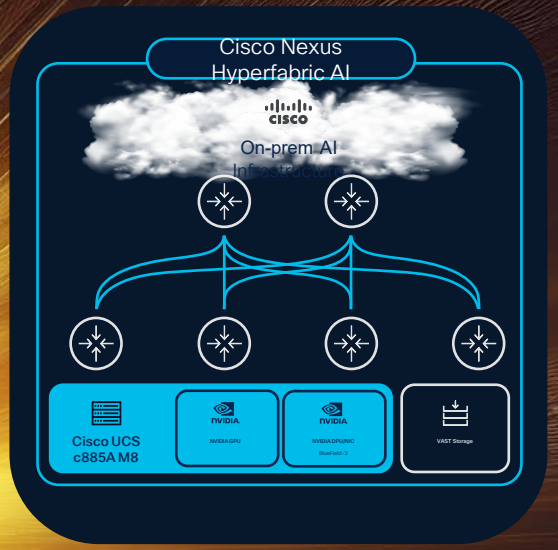
Additional Resources

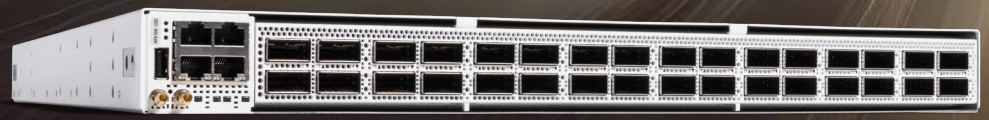


[Cisco.com - Cisco Nexus Hyperfabric]

From Chaos to Clarity: Cisco AI Ready Data Center







Q & A


CISCO