Cisco AI PODs with FlexPod Datacenter

Value statement

What if your AI infrastructure could be deployed, scaled, and operated using a fully validated, modular design?

Cisco AI PODs with FlexPod Datacenter provide a pre-tested architecture combining GPU-optimized compute, high-speed networking, and NetApp ONTAP storage. With automation through Cisco Intersight and Nexus Dashboard, the solution delivers predictable performance and operational consistency for AI training and fine-tuning workloads.

Organizations deploying AI training and fine-tuning workloads require infrastructure that delivers high performance, predictable scalability, and end-to-end operational visibility. However, building an AI environment that integrates compute, networking, and storage components at scale is complex, especially when high-throughput datasets, distributed GPUs, and Kubernetes-based platforms are required.

Cisco AI PODs with FlexPod Datacenter provide a comprehensive, validated architecture that integrates:

- Cisco UCS® C885A M8 Rack Servers with 8× NVIDIA H200 GPUs
- Cisco Nexus® backend (east/west) and frontend (north/south) switching fabrics
- NetApp AFF A90 storage with NFS, RDMA, FlexGroup, and S3 capabilities
- Red Hat OpenShift for containerized workloads
- VIDIA Base Command Manager for workload provisioning
- Cisco Intersight® and Nexus Dashboard for deployment, monitoring, and lifecycle operations

The solution is designed for AI training and fine-tuning use cases, providing the high-performance data paths, scalable compute infrastructure, and validated storage configurations required for consistent AI workload performance. The design also enables modular expansion as workload demands grow, allowing compute, storage, or network capacity to be added independently.

Benefits

- **High-performance data access for AI training**
  NetApp AFF A90 with NFSv4.1, RDMA, and FlexGroup provides the high-throughput sequential reads required for loading large training datasets and writing checkpoints.
- **Scalable GPU infrastructure**
  Cisco UCS C885A M8 Rack servers with NVIDIA H200 GPUs connect through a fully non-blocking 400G backend fabric, enabling multi-node GPU training and east/west data movement at scale.
- **Modular, flexible expansion**
  The design supports independent scaling of compute, storage, and network bandwidth using a modular leaf-spine architecture.
- **Validated, repeatable deployment**
  The configuration follows Cisco Validated Design (CVD) principles and includes detailed steps for cabling, BIOS configuration, firmware updates, and ONTAP setup.
- **Centralized operations and automation**
  Cisco Intersight and Nexus Dashboard provide unified management and automation for Cisco UCS servers and switching fabrics, improving operational consistency.

Trends and challenges

Rising performance requirements for AI training

AI training and fine-tuning workloads require extremely high-throughput storage access and low-latency GPU-to-GPU communication. Large datasets must be repeatedly loaded into GPU memory, and distributed training requires reliable east/west networking across GPU nodes. Traditional infrastructure often introduces bottlenecks that limit scalability.

Operational complexity across compute, network, and storage

AI environments incorporate heterogeneous components—servers, high-speed fabrics, Kubernetes, GPU runtimes, and storage protocols such as NFS, RDMA, and S3. Managing firmware, provisioning, network VLANs, and storage namespaces across domains introduces operational overhead.

Need for validated, repeatable deployment models

Customers require validated reference architectures that remove guesswork. The deployment guide provides a fully validated configuration, including:

- VLAN layouts
- Cabling matrices
- BIOS and BMC settings
- ONTAP IPspaces, broadcast domains, and LIFs
- GPU server firmware upgrade processes
- Integration with NVIDIA Base Command Manager

This ensures predictable performance and reduces deployment risk.

How it works

Cisco AI PODs with FlexPod Datacenter combine GPU-accelerated compute, a dual high-speed fabric, and NetApp ONTAP storage into a unified architecture that is fully validated for AI training and fine-tuning workloads. The deployment guide outlines the complete configuration—from BIOS settings and cabling to storage namespaces and Kubernetes integration—ensuring that every component works together to deliver predictable performance and scalable throughput.

GPU compute infrastructure

At the heart of the solution are Cisco UCS C885A M8 Rack Servers, each equipped with eight NVIDIA H200 SXM GPUs. These servers are configured specifically for AI performance through detailed BIOS settings, including IOMMU enablement, CPU performance tuning, and PCIe optimization. The design begins with an initial BMC setup that establishes NTP synchronization, DNS configuration, and secure administrative access.

Firmware is a critical part of maintaining performance and compatibility in AI systems, and the guide provides explicit steps to update BIOS, BMC firmware, PCIe switch firmware, and NIC firmware. The document also identifies a key operational requirement: when using NVIDIA BlueField-3-based NICs for north/south connectivity, the DPU's internal CPU must be disabled to allow LACP port channels to function correctly. These instructions ensure that the GPU nodes are configured consistently and can be scaled horizontally without unpredictable behavior.

Key compute characteristics:

- Cisco UCS C885A M8 Rack Servers with 8× NVIDIA H200 SXM GPUs
- AI-tuned BIOS configurations for stability and throughput
- Structured firmware update workflow for all components
- BMC setup for NTP, DNS, and management
- NVIDIA BlueField-3 DPU CPU disablement for reliable LACP operations

High-speed networking architecture

The networking design uses two independent Cisco Nexus fabrics, each serving a distinct purpose. The backend (east/west) fabric, built with Cisco Nexus 9332D-GX2B leaf switches and Cisco Nexus 9364D-GX2A spine switches, operates at 400G and is dedicated to GPU-to-GPU communication—essential for multi-node training and distributed workload synchronization. This fabric ensures GPU nodes exchange data at maximum throughput with minimal latency.

The frontend (north/south) fabric connects compute, storage, and control-plane elements through 100G and 200G links. It carries all traffic between the cluster, OpenShift nodes, and ONTAP storage. The deployment guide includes complete cabling matrices, VLAN assignments for management, NFS, and S3 traffic, and NTP distribution through tenant VRFs. The fabrics can be provisioned and automated through Nexus Dashboard, enabling consistent rollout across large environments.

Key network characteristics:

- 400G backend fabric for east/west GPU communication
- 100G/200G frontend fabric for storage and cluster services
- Detailed port-by-port cabling guidance for predictable deployments
- VLANs for management, NFS, RDMA, and S3
- Automated fabric creation and lifecycle management through Nexus Dashboard

NetApp ONTAP AFF A90 storage

The storage subsystem uses NetApp AFF A90 controllers configured according to the validated design. AI training requires extremely high-throughput, low-latency access to large datasets, and the guide outlines how to meet those requirements using FlexGroup, NFSv4.1, RDMA, and NVIDIA GPUDirect Storage. The design describes how to create IPspaces, broadcast domains, VLANs, LIFs, and SVMs specifically for OpenShift and Ubuntu-based training workloads.

Storage resiliency is built into the configuration through clustered controllers, mirrored root volumes using SnapMirror LS sets, and a structured LIF layout for NFS, RDMA, S3, and management interfaces. ONTAP S3 is also configured for storing training artifacts, checkpoints, logs, and model files. These capabilities allow the storage layer to support both performance-intensive training loops and object-based workflows from NVIDIA Base Command Manager or OpenShift.

Key storage characteristics:

- AFF A90 HA pair configured for AI datasets
- FlexGroup for distributed, parallelized access
- NFSv3/v4.1, RDMA, and GPUDirect Storage support
- Tenant-specific IPspaces, broadcast domains, and VLANs
- LIFs for NFS, RDMA, S3, and management
- SnapMirror-based resiliency for root volumes

Kubernetes and AI runtime integration

Red Hat OpenShift 4.16 serves as the platform for containerized workloads, configuration VMs, and services such as DNS/DHCP when required. The deployment describes how the Cisco UCS X9508 Chassis and Cisco UCS X210c M8 Compute Nodes are integrated into the OpenShift cluster to host these services. NVIDIA Base Command Manager (BCM) provides the provisioning and orchestration layer for the GPU nodes, enabling SLURM-based scheduling and workload execution on Ubuntu 22.04 GPU images.

The BCM head node connects to the frontend fabric through bonded 100G links, ensuring sufficient bandwidth for image transfers and workload coordination. This integration allows the compute nodes to boot through PXE, receive NVIDIA-optimized software stacks, and participate in AI training clusters in a tightly controlled manner.

Key runtime characteristics:

- Red Hat OpenShift 4.16 for container and VM workloads
- NVIDIA BCM for GPU image provisioning and SLURM scheduling
- VIC-based 100G bonding for BCM head node connectivity
- Support for Ubuntu-based GPU runtime images
- Integration with storage and fabrics for end-to-end workflow continuity

Unified operations

Operations are centralized through Cisco Intersight and Nexus Dashboard, each responsible for a critical part of the environment. Intersight provides hardware inventory, monitoring, BMC/KVM access, and firmware management for the Cisco UCS C885A GPU servers and Cisco UCS X-Series nodes. Nexus Dashboard manages both backend and frontend fabrics, including fabric creation, switch provisioning, and continuous monitoring.

Together, these tools offer visibility across compute, networking, and storage components—allowing administrators to maintain consistent configurations, diagnose issues quickly, and scale the environment with confidence.

Key operations characteristics:

- Cisco Intersight for UCS monitoring, server lifecycle, and BMC/KVM access
- Cisco Nexus Dashboard for fabric automation and visibility
- Unified monitoring across compute, network, and storage
- Consistency through validated policies and automated deployment workflows

The FlexPod advantage

Cisco uniquely brings together compute, networking, and cloud-managed operations into a unified AI platform. With Cisco Intersight and Nexus Dashboard, customers manage AI clusters with the same visibility and control as their existing workloads. Deep integration with NetApp ONTAP ensures predictable data performance, consistent operations, and a validated path to scale AI training with confidence.

Learn more

- Cisco® AI infrastructure solutions
- FlexPod design guides
- NetApp ONTAP AI
- NVIDIA Base Command

Use cases

Use case	Description
AI training	High-throughput training requiring multi-node GPU clusters, fast sequential reads, and distributed east/west communication
AI fine-tuning	Updating models with domain-specific datasets using accelerated storage and GPU compute
Dataset preprocessing	Large-scale data preparation using container-based workflows on OpenShift
Model checkpointing	Frequent writes to high-speed ONTAP storage during training loops
S3-based model storage	Storage of artifacts, logs, and checkpoints using ONTAP S3 interfaces

Cisco Capital

Financing to help you achieve your objectives

Cisco Capital® can help you acquire the technology you need to achieve your objectives and stay competitive. We can help you reduce CapEx. Accelerate your growth. Optimize your investment dollars and ROI. Cisco Capital financing gives you flexibility in acquiring hardware, software, services, and complementary third-party equipment. And there's just one predictable payment. Cisco Capital is available in more than 100 countries. Learn more.

C22-5724051-00 02/26