Cisco AI POD for Training with Pure Storage

Benefits

- **Accelerate AI training:** high-performance GPU clusters paired with lossless, low-latency networking deliver faster model training and reduced Job Completion Times (JCTs).
- **Scale with confidence:** modular scale units (32/64/128 GPUs) that allow predictable, incremental expansion as AI initiatives grow.
- **Operational simplicity:** unified management through Cisco Nexus® Dashboard and Cisco Intersight® streamlines deployment, monitoring, automation, and day-2 operations.
- **High-performance data access:** Pure Storage FlashBlade//S provides ultra-fast, parallel file and object storage with NIVIDIA GPU Direct Storage (GDS) for efficient, data-intensive AI pipelines.
- **Enterprise reliability:** nondisruptive upgrades, integrated data protection, and full-stack observability ensure consistent performance and high availability for mission-critical AI workloads.

An integrated, high-performance AI-training platform built for the enterprise

The **Cisco AI POD** for Training with **Pure Storage** is a validated, full-stack architecture combining Cisco UCS® GPU servers, Cisco Nexus 9000 Series networking, and Pure Storage **FlashBlade//S** to deliver a scalable and predictable AI-training platform. Designed for enterprise data centers, this solution accelerates distributed training of large AI and ML models while simplifying operational complexity.

The architecture uses dedicated backend (east/west) and frontend (north/south) fabrics to ensure optimal GPU performance, high-speed data ingestion, and uninterrupted access to training datasets. Modular scale units provide a consistent way to expand GPU clusters as workload demands increase.

Learn more

- Cisco AI POD design guides
- Cisco UCS C885A M8 Rack Server
- Cisco UCS C845A M8 Rack Server
- Cisco Nexus 9000 Series Switches
- Pure Storage FlashBlade//S
- Cisco Intersight
- Cisco Nexus Dashboard

End-to-end infrastructure optimized for large-scale AI training

At the core of the solution are Cisco UCS C885A M8 and Cisco UCS C845A M8 rack servers—dense, high-bandwidth GPU platforms engineered for large-scale distributed training. These GPU nodes connect to a lossless backend fabric built on Cisco Nexus 9000 Series Switches, ensuring fast, deterministic communication between GPUs—one of the biggest factors in reducing training times.

Pure Storage FlashBlade//S delivers the high-performance data layer required for AI. Its flash-optimized file and object platform provides consistently low latency and parallel data access, while support for NVIDIA GPU Direct Storage (GDS) enables direct, accelerated data movement between storage and GPUs. This reduces overhead, keeps GPUs fully utilized, and eliminates training slowdowns caused by I/O bottlenecks.

Scaling is straightforward. Modular scale units—available in 32, 64, or 128 GPU configurations—allow organizations to expand capacity without redesigning their architecture. Unified management through Cisco Intersight and Cisco Nexus Dashboard simplifies deployment and day-2 operations, providing end-to-end visibility and automation across compute, network, and storage.

C45-5724101-00 02/26