Accelerate Your AI Journey with Confident, Scalable Infrastructure

Value statement

What if you could deploy high-performance AI infrastructure that is already validated by the industry leader? We provide a prescriptive blueprint endorsed by NVIDIA to help you operationalize AI workloads faster and with less risk.

Benefits

- **Deploy with confidence with NVIDIA's endorsement:** eliminate guesswork with a design endorsed by NVIDIA for Infrastructure Configuration and Spectrum-X, based on the NVIDIA ERA for 2-8-9-400
- **Maximize performance with 800G networking:** ensure lossless, low-latency GPU-to-GPU communication using Cisco Nexus 9000 Series Switches with 800GbE capabilities and Cisco® Silicon One® technology
- **Simplify operations at scale:** manage both backend and frontend fabrics centrally using Cisco Nexus Dashboard and Cisco Intersight, providing a "one fabric" approach that minimizes configuration errors
- **Scale modularly:** grow your AI clusters cost-effectively Using Scalable Units (SUs) of Cisco UCS C885A M8 Rack Servers, designed for high-density training and fine-tuning workloads

Overview

Artificial Intelligence (AI) and Machine Learning (ML)—specifically generative AI and HPC-scale training—are reshaping business innovation. However, the infrastructure required to support these workloads is fundamentally different from traditional enterprise applications. You need massive parallel GPU compute, ultra-low latency networking, and specialized data pipelines. Building this from scratch is complex, risky, and slow.

The **Cisco AI POD infrastructure for enterprises** solves this challenge. It is a fully validated, full-stack solution that integrates NVIDIA's accelerated computing stack with Cisco's 800G-ready networking, global-scale observability, and zero-trust security. Most importantly, this design is **endorsed by NVIDIA** as an Enterprise Reference Architecture (ERA). It provides you with modular building blocks to confidently deploy, manage, and scale AI clusters, ensuring your infrastructure is faster to deploy, easier to operate, and cost-effective to scale.

Figure 1. Accelerating AI innovation with Cisco and NVIDIA

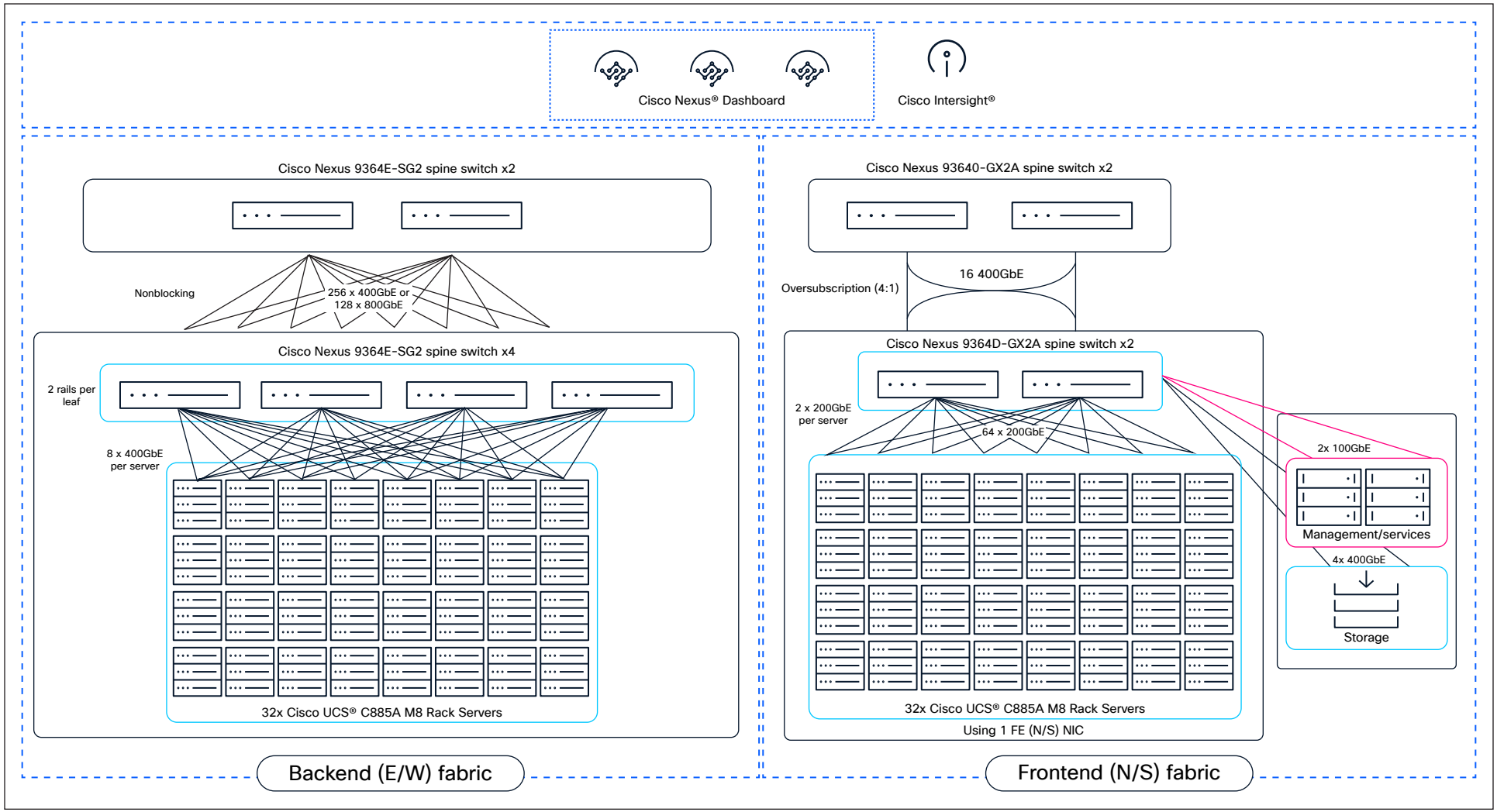

Figure 2. High-level solution topology for a scalable 256-GPU cluster using Cisco AI PODs

Trends and challenges

The infrastructure gap in the AI era

The rapid rise of agentic AI, physical AI, and generative AI has created a critical demand for specialized infrastructure. Traditional enterprise data centers are often ill-equipped to handle the thermal, power, and bandwidth requirements of these modern workloads.

- **Performance bottlenecks:** AI models require extensive matrix multiplications and parallel computations. Without a high-bandwidth, lossless network fabric, Job-Completion Times (JCTs) for training increase significantly, slowing down innovation.
- **Operational complexity:** Integrating GPUs, DPUs, storage, and networking into a cohesive system is difficult. Fragmented management tools lead to silos, security gaps, and inefficient resource utilization.
- **Scalability issues:** As AI initiatives move from pilot to production, enterprises struggle to scale their infrastructure linearly without incurring prohibitive costs or redesigning their network architecture.

You need an architecture that bridges this gap—one that brings together the best of accelerated computing with enterprise-grade networking and security.

How it works

The Cisco AI POD Infrastructure is a prescriptive reference architecture that combines best-in-class components into a unified solution.

Compute: The powerhouse

At the core is the **Cisco UCS C885A M8 Rack Server**, a high-performance 8RU system based on NVIDIA's HGX architecture. It features 8x NVIDIA HGX H200 GPUs interconnected through NVIDIA's NVLink for massive compute power. It utilizes NVIDIA BlueField-3 network adapters (SuperNICs) to ensure high-speed data movement.

Networking: The nervous system

The solution utilizes a dual-fabric design managed by **Cisco Nexus Dashboard:**

- **Backend fabric (east/west):** dedicated to GPU-to-GPU communication. This fabric uses Cisco Nexus 9364E-SG2 switches (800GbE) in a nonblocking, lossless spine-leaf topology to minimize latency and maximize training performance.
- **Frontend fabric (north/south):** handles management, storage, and user traffic. This fabricuses Cisco Nexus 9332D-GX2B or 9364D-GX2A switches (400GbE) to connect the cluster to the rest of the enterprise.

Management and orchestration

Cisco Intersight provides SaaS-based lifecycle management for the compute infrastructure, while **Red Hat OpenShift** serves as the container orchestration platform, ensuring a robust environment for your AI workloads.

NVIDIA Endorsement

This specific configuration is endorsed by NVIDIA under their Enterprise Reference Architecture (ERA) for the **2-8-9-400** design pattern (2 CPUs, 8 GPUs, 9 NICs, and 400Gbps bandwidth), guaranteeing that the system meets the rigorous standards required for NVIDIA AI Enterprise software.

Services

Accelerate value with Cisco Customer Experience (Cisco CX)

Cisco Customer Experience (Cisco CX) services can help you navigate the complexities of AI deployment. From advisory services to validate your architectural goals to implementation services that speed up the deployment of your Cisco AI PODs, our experts help ensure that your infrastructure is optimized for performance and aligned with your business outcomes.

Use cases

Industry	Use case description
Generative AI	Large Language Model (LLM) training: build and fine-tune massive foundational models requiring synchronized parallel processing across hundreds of GPUs
Healthcare	Medical imaging and drug discovery: accelerate the analysis of complex biological data and high-resolution imaging to speed up diagnosis and treatment development
Finance	Algorithmic trading and fraud detection: process vast datasets in real time to identify fraudulent patterns and execute high-frequency trades with ultra-low latency
Automotive	Autonomous driving simulation: train deep-learning neural networks using massive datasets to improve the safety and reliability of self-driving vehicle systems

“The infrastructure required to operationalize AI workloads is fundamentally different... AI infrastructure demands massive parallel GPU compute, ultra-low latency networking, and security embedded at every layer.”

“This architecture brings together NVIDIA’s accelerated computing stack with Cisco’s 800G-ready networking to make AI clusters faster to deploy, easier to operate, and more cost-effective to scale.”

Understanding the NVIDIA ERA endorsement

What is it?

NVIDIA Enterprise Reference Architecture (ERA) is a certification that validates specific infrastructure designs for optimal performance with NVIDIA AI software and hardware.

Why it matters?

The Cisco AI POD design described here is endorsed for the NVIDIA ERA 2-8-9-400 configuration. This means that Cisco UCS C885A M8 Rack Servers and Cisco Nexus 9000 Series Switches have been verified to provide the exact bandwidth, latency, and topology required to extract maximum performance from NVIDIA H200 GPUs. It assures you that your investment is built on a foundation certified by the leader in AI computing.

The Cisco advantage

Cisco is uniquely positioned to power the AI era by combining the industry's most robust networking portfolio with strategic partnerships. Unlike siloed solutions, Cisco Secure AI Factory with NVIDIA integrates silicon, systems, and software. With the Cisco AI POD, you get the raw power of NVIDIA GPUs combined with the operational simplicity of Cisco Intersight and Nexus Dashboard backed by a global support network that helps keep your business up and running.

Learn more

Start building your Cisco Secure AI Factory with NVIDIA

Ready to operationalize your AI strategy with infrastructure you can trust? Explore the detailed design of the Cisco AI POD and see how our partnership with NVIDIA can accelerate your innovation. For additional information, visit the Cisco AI Infrastructure page then contact your Cisco account manager or authorized partner.

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.

C22-5773501-00 02/26