Cisco Nexus Hyperfabric AI

© 2025 Cisco and/or its affiliates. All rights reserved.

Contents

© 2025 Cisco and/or its affiliates. All rights reserved.

Overview

To build on the robust and scalable AI infrastructure provided by Cisco Secure AI Factory with NVIDIA, the Cisco Nexus® Hyperfabric AI solution offers a vertically integrated approach that unifies network, compute, GPUs, and storage components. This integration streamlines deployment and management, delivering optimized performance and agility tailored to enterprise AI workloads, enabling organizations to accelerate innovation with a cohesive, end-to-end AI platform.

Cisco Nexus Hyperfabric is a cloud-managed, full-stack AI infrastructure solution designed to accelerate AI lifecycle management from design to scale. This solution delivers high throughput, low latency, and scalable AI infrastructure optimized for AI workloads. It supports two deployment models:

- **Hyperfabric AI:** A turnkey, vertically integrated, and NVIDIA ERA–certified full-stack AI infrastructure solution, consisting of networking, servers, GPUs, and storage. It is cloud-managed to simplify the entire lifecycle, from design to operation, for demanding AI enterprise workloads. This prescriptive solution optimizes for time to value of dedicated AI clusters from concept to production.
- **Bring Your Own AI (BYO AI) with Hyperfabric:** Hyperfabric supports a Bring Your Own AI (BYO AI) approach, allowing enterprises to integrate their own servers, GPUs, and storage with Cisco's cloud-managed network fabric for flexible, scalable AI infrastructure deployment.

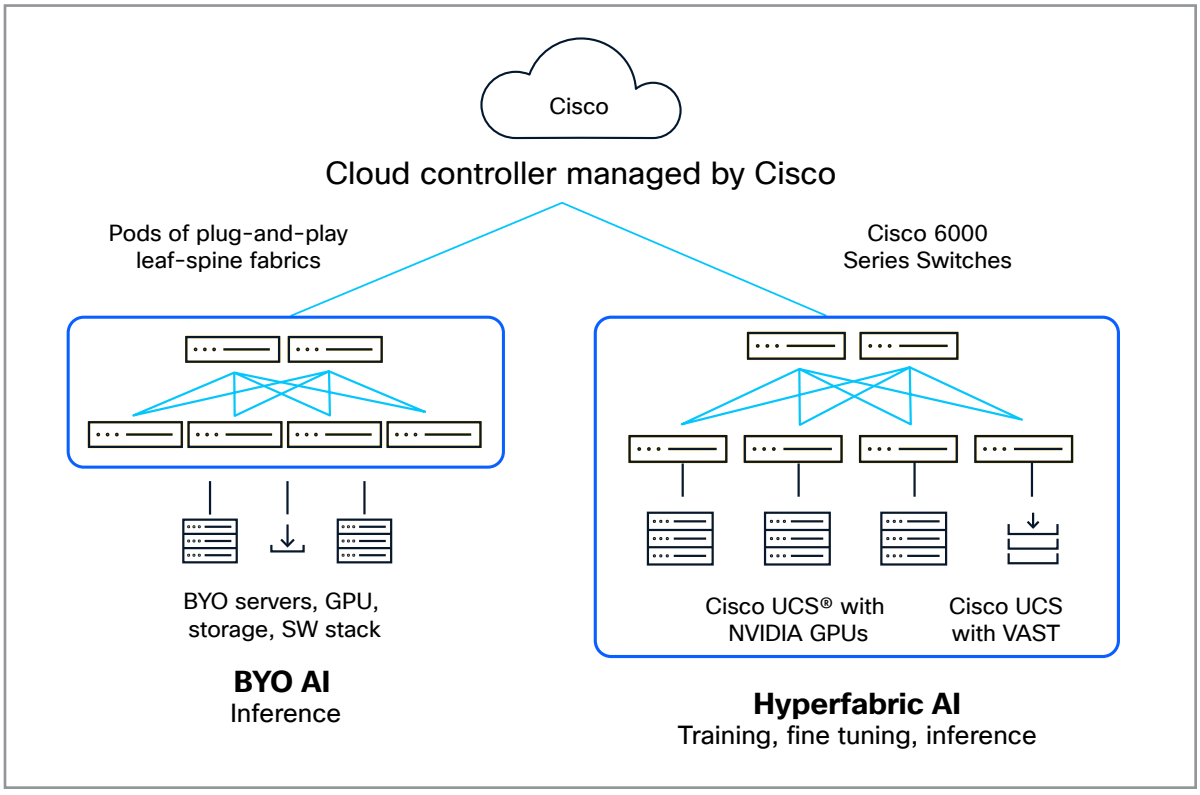

Figure 1. AI solutions with Hyperfabric

© 2025 Cisco and/or its affiliates. All rights reserved.

Cisco Nexus Hyperfabric AI

Cisco Nexus Hyperfabric AI is a turnkey, on-premises, AI-cluster solution managed by a cloud-hosted controller, designed to simplify and accelerate AI deployments for enterprises. This cloud-managed vertical stack solution consisting of purpose-built hardware, software, a cloud controller, day-2 operations, automation, and Cisco support that eliminates complexity. IT, application, and DevOps teams manage the full lifecycle of designing, ordering, deploying, validating, monitoring, and scaling fabrics without requiring deep networking or operational expertise. It integrates Cisco's high-performance networking hardware, including Cisco® Silicon One® switches and Cisco UCS C885A M8 Rack Servers with NVIDIA HGX, with a cloud-managed operational model. This solution offers preconfigured templates for AI clusters, enabling rapid design, deployment, and scaling of AI infrastructure with zero-touch provisioning and continuous monitoring. It supports a unified stack that includes networking, compute, GPUs and storage components, all managed through a cloud controller to provide full visibility and control. Cisco Nexus Hyperfabric AI is built to empower IT generalists, data scientists, and DevOps teams by delivering a simplified, scalable, and secure AI infrastructure that aligns with NVIDIA's Enterprise Reference Architecture, facilitating efficient AI workloads and innovation across industries.

Figure 2 shows the key components of the solution. The key hardware components used in the cluster are described in the next section.

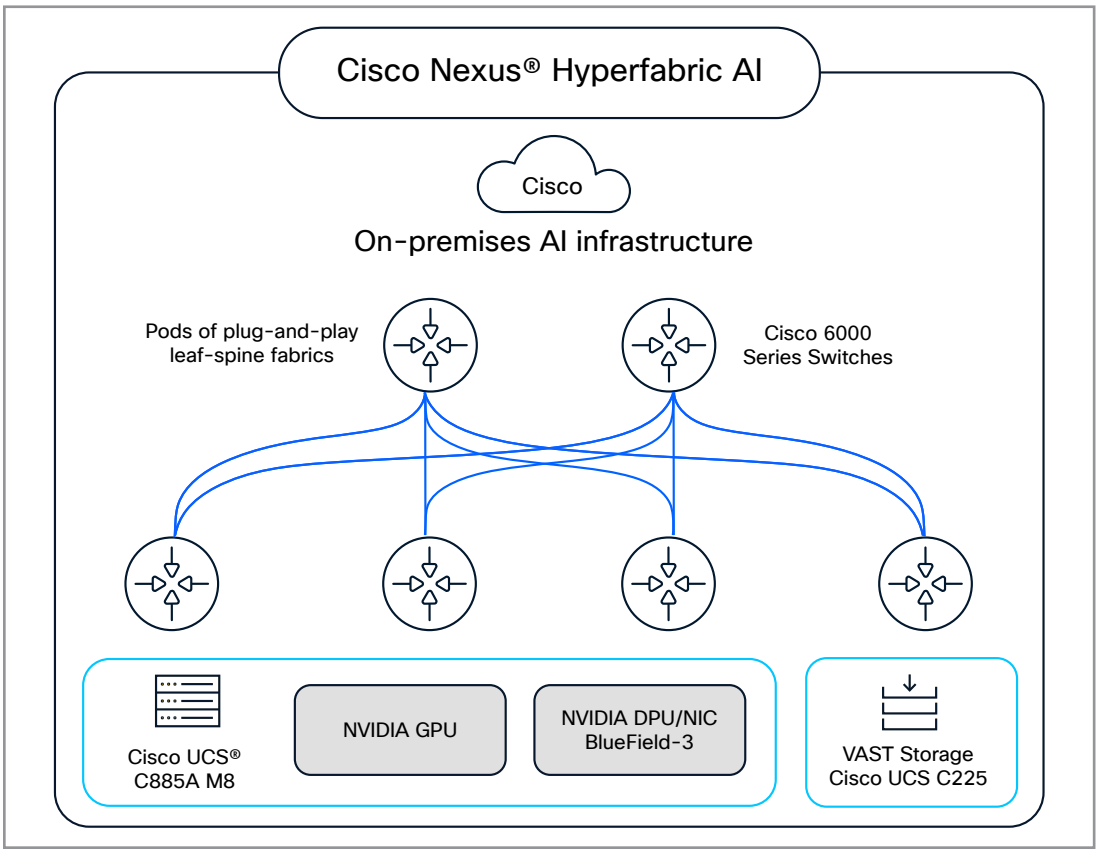

Figure 2. Cisco Nexus Hyperfabric AI

© 2025 Cisco and/or its affiliates. All rights reserved.

Key components

- Cisco Nexus 6000 Series Switches with Silicon One technology, including HF6100 variants.
- Cisco UCS C885A M8 Rack servers equipped with NVIDIA HGX H200 GPUs and BlueField-3 DPUs.
- Optional VAST data-storage platform integrated with Cisco UCS servers.
- Cisco optics and cables.
- Optional AI software stack including NVIDIA AI Enterprise (NVAIE).
- Onsite installation services by Cisco Customer Experience (CX).

Performance and features

- Certified by NVIDIA ERA for AI workloads, thus ensuring validated performance and scalability.
- High-throughput fabric with 800 GB/s bandwidth and ultra-low latency.
- Support for RDMA, NVMe-oF, and advanced telemetry.
- Cloud-managed lifecycle with zero-touch provisioning and continuous software and firmware updates.
- End-to-end assertion-based monitoring from network fabric to compute NICs and storage.
- Preconfigured leaf-and-spine pods for plug-and-play deployment.

Scalability

- Modular design supporting small, medium, and large AI clusters.
- Pre-configured designs for the entire infrastructure component suite – frontend, backend, rails, storage, and management.
- Independent scaling of network, compute, GPUs, and storage resources.
- Fully integrated turnkey solution with VAST storage or support for Bring Your Own Storage.

Use cases and benefits

- Large-scale enterprise AI model training, fine tuning, and inference.
- Dedicated AI/ML clusters with optimized efficiency and low latency.
- Faster time to value with replicable deployment models.
- Validated turnkey AI stack through Cisco, NVIDIA, and storage vendor partnership, ensuring optimized hardware and software integration.

© 2025 Cisco and/or its affiliates. All rights reserved.

Architecture and cloud management

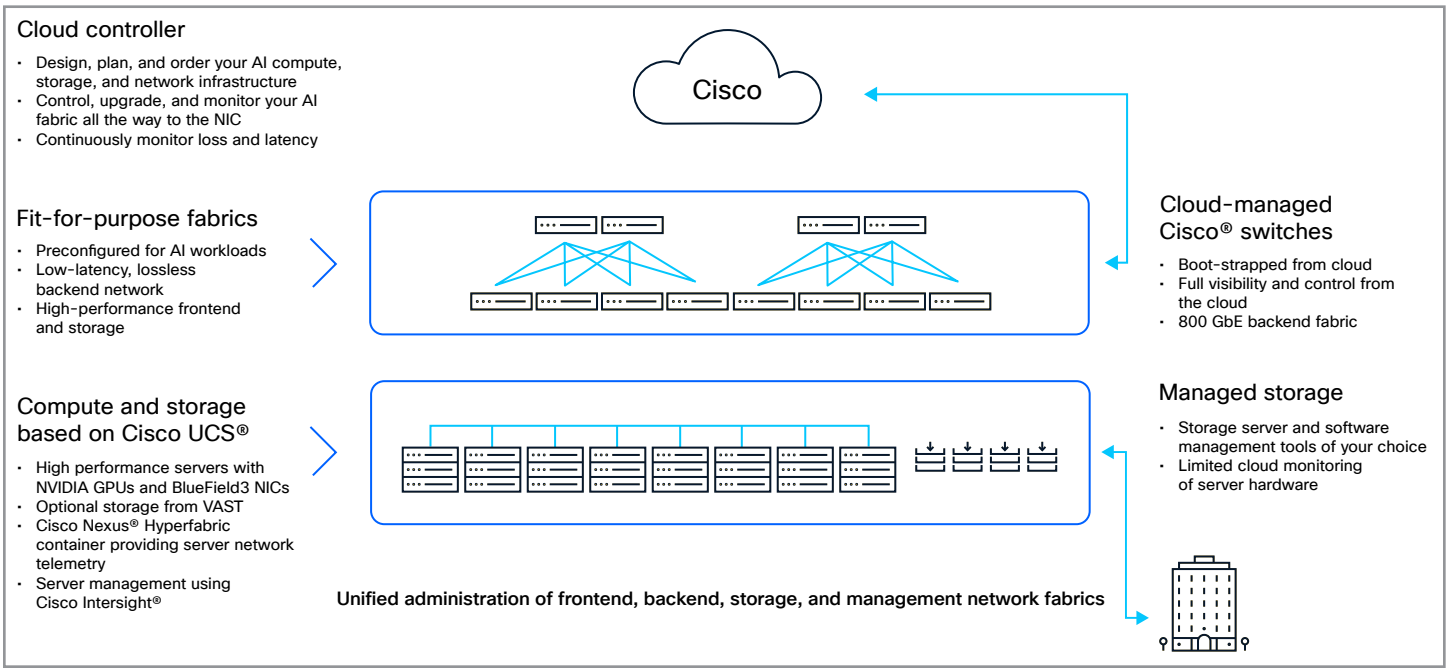

Figure 3. How Cisco Nexus Hyperfabric AI is built

The Cisco Nexus Hyperfabric AI offers a cloud-managed full-stack AI infrastructure that simplifies the entire lifecycle management process, including design, ordering, deployment, management, and scaling of AI compute, GPUs, storage, and network infrastructure. This solution features a cloud controller that enables comprehensive design, planning, ordering, operation, upgrade, and monitoring of the AI fabric down to the NIC level, with continuous monitoring of loss and latency. Switches, GPU clusters, and storage servers are bootstrapped and fully visible and controlled from the cloud, ensuring seamless management.

The infrastructure includes fit-for-purpose fabrics that are preconfigured specifically for AI workloads, providing a low-latency, lossless backend network alongside high-performance frontend and storage fabrics. Software and monitoring capabilities are integrated through a Hyperfabric container that delivers server network telemetry, unified administration of frontend, backend, storage, and management fabrics, as well as assertion-based fabric monitoring and real-time health checks.

Scalability and flexibility are key features of this architecture, supporting independent scaling of network, compute, and storage resources. Its modular design allows customers to bring their own storage or opt for fully integrated turnkey solutions, enabling tailored deployments that meet diverse AI workload requirements. This comprehensive approach facilitates fast, reliable AI cluster deployment with a replicable model that eases AI cluster management and future scalability.

© 2025 Cisco and/or its affiliates. All rights reserved.

Deployment experience

The guided installation process for Cisco Nexus Hyperfabric AI simplifies AI cluster deployments through real-time verification, cloud-managed lifecycle, continuous software updates, and a replicable deployment model. On day 0, users access the centralized cloud-managed controller portal at hyperfabric.cisco.com to plan and design the infrastructure using the designer tool, finalize the bill of materials including optics and cables, and begin installation planning based on the cabling plan. On day 1, upon equipment arrival, Cisco 6000 Series Switches, Cisco UCS servers with GPUs, and optional Cisco UCS servers with 3rd party storage software are connected and registered to the cloud controller. The mobile-friendly portal provides step-by-step deployment guidance and real-time topology validation to ensure that the physical fabric matches the design. From day 2 onward, lifecycle management is handled remotely through the cloud controller, enabling monitoring, upgrades, and collaboration. This process automates provisioning, guides on-site tasks, and allows remote management to be accessible even to non-expert staff, thereby streamlining deployment and operational management of AI infrastructure.

Bring Your Own AI (BYO AI) with Cisco Nexus Hyperfabric

Bring Your Own AI (BYO AI) with Cisco Nexus Hyperfabric is a cloud-managed networking fabric solution designed to simplify the lifecycle management of an AI-cluster network. It offers a flexible, modular architecture where enterprises can bring their own servers, GPUs, storage, and software stack while leveraging Cisco's high-performance Nexus Hyperfabric network fabric. The BYO AI with Hyperfabric provides enterprises the flexibility to maintain their preferred AI infrastructure servers, GPUs, and storage choices while benefiting from Cisco's optimized, secure, and cloud-managed high-performance AI networking and management capabilities.

For BYO AI with Cisco Nexus Hyperfabric, the solution supports mixed AI inference and non-AI use cases by enabling a multi-purpose data-center environment that can handle both AI workloads and traditional non-AI services. This approach allows enterprises to deploy inference workloads alongside other applications within the same infrastructure, providing flexibility and efficient resource utilization.

Components

- Cisco Nexus 6000 Series Switches with Silicon One technology.
- Customer-provided GPU servers, and storage.

Features

- Cloud-managed network fabric with advanced telemetry and assertion-based monitoring for real-time health and performance visibility.
- Supports multi-purpose data centers running AI and non-AI workloads enabling flexible infrastructure utilization.
- High-performance network fabric optimized for AI workloads with ultra-low latency and high throughput.

© 2025 Cisco and/or its affiliates. All rights reserved.

Management

- Network fabric managed through a cloud controller, providing centralized design, provisioning, monitoring, and upgrades.
- AI software, compute, and storage managed separately by the customer, thereby allowing tailored operational control.

Use cases

- Enterprises seeking easy-to-deploy, cloud-managed and operate network fabric.
- Multi-purpose data centers supporting inference, Retrieval-Augmented Generation (RAG), and fine tuning and other AI workloads alongside traditional applications.

Network switch details

- **HF6100-64ED:** 64x 100/200/400/800 GbE OSFP form factor.
- **HF6100-32D:** 32x 100/200/400 GbE QSFP56-DD form factor.
- **HF6100-60L4D:** 60x 10/25/50 GbE SFP56.

Fabric performance

- 800 GbE backend fabric leveraging Cisco's silicon photonics technology for ultra-high bandwidth and low latency.
- Linear scale-out performance with scaling, thereby ensuring consistent throughput and fabric efficiency as the network grows.

© 2025 Cisco and/or its affiliates. All rights reserved.

Licensing

A subscription entitlement is needed for every Cisco 6000 Series Switch that is deployed and used. Subscription entitlements may be initially purchased for three, five, or seven years and may be renewed. The subscription-feature entitlement tiers are based on fabric use cases. Currently, two packages are available: one (“Essentials”) for general-purpose fabrics for BYO AI and the second (“Premier”) for Hyperfabric AI. All the switches in a fabric must use the same entitlement tier; however, an organization may concurrently manage multiple fabrics that use different entitlement tiers.

Table 1. Essentials and Premier feature tiers

Features	Essentials (for BYO AI with Hyperfabric)	Premier (only for Cisco Nexus Hyperfabric AI)
Cisco support 8x5xNBD	Yes	Yes
Cloud controller	Yes	Yes
Designer (no purchase required)	Yes	Yes
Cloud-driven software upgrades	Yes	Yes
BOM generation	Yes	Yes
On-site deployment assistance	Yes	Yes
Plug-and-play deployment	Yes	Yes
Spine-and-leaf topologies	Yes	Yes
Mesh (spineless) topologies	Yes	Yes
EVPN/VXLAN underlay (opaque)	Yes	Yes
Static and BGP routing	Yes	Yes

© 2025 Cisco and/or its affiliates. All rights reserved.

Features	Essentials (for BYO AI with Hyperfabric)	Premier (only for Cisco Nexus Hyperfabric AI)
MLAG	Yes	Yes
Supports RDMA over Converged Ethernet v2 (RoCEv2)	Yes	Yes
Real-time cloud-accessed telemetry	Yes	Yes
IPv4 and IPv6	Yes	Yes
Assertion-based monitoring	Yes	Yes
Survivable data and local management plane	Yes	Yes
Hardware-based attestation and security	Yes	Yes
API for headless provisioning and monitoring	Yes	Yes
AI use-case support	Yes	Yes
Aligned to NVIDIA Enterprise Reference Architecture	No	Yes
NVIDIA Adaptive Routing on backend switches	No	Yes
Deploy AI-validated blueprints built into workflow	No	Yes
Automatically provisions lossless backend AI and storage networks	No	Yes

© 2025 Cisco and/or its affiliates. All rights reserved.

Features	Essentials (for BYO AI with Hyperfabric)	Premier (only for Cisco Nexus Hyperfabric AI)
Internal performance monitoring between switches and servers	No	Yes
Options for AI servers	Cisco UCS or Bring-Your-Own	Bundled Cisco UCS

Product sustainability

Information about Cisco's Environmental, Social, and Governance (ESG) initiatives and performance is provided in Cisco's CSR and sustainability reporting.

Table 2. Cisco environmental sustainability information

Sustainability topic		Reference
General	Information on product-material-content laws and regulations	Materials
	Information on electronic waste laws and regulations, including our products, batteries, and packaging	WEEE Compliance
	Information on product takeback and reuse program	Cisco Takeback and Reuse Program
	Sustainability inquiries	Contact: csr_inquiries@cisco.com
Material	Product packaging weight and materials	Contact: environment@cisco.com

© 2025 Cisco and/or its affiliates. All rights reserved.

Cisco and partner services

Cisco and partner services offer a wide range of services to help accelerate your success in connecting Cisco 6000 Series Switches to the Nexus Hyperfabric cloud controller. Our innovative service offerings are delivered through a unique combination of people, processes, tools, and partners and are focused on helping you increase operational efficiency and improve your network control. Cisco Nexus Hyperfabric solution provides proactive support with the Cisco SMARTnet® service to help you resolve mission-critical problems with direct access at any time to Cisco network experts and award-winning resources. Spanning the entire network lifecycle, our service offerings help increase investment protection, optimize network operations, support migration operations, and strengthen your IT expertise.

For more information, please visit www.cisco.com/go/services.

Cisco Capital

Flexible payment solutions to help you achieve your objectives

Cisco Capital makes it easier to get the right technology to achieve your objectives, enable business transformation and help you stay competitive. We can help you reduce the total cost of ownership, conserve capital, and accelerate growth. In more than 100 countries, our flexible payment solutions can help you acquire hardware, software, services and complementary third-party equipment in easy, predictable payments. Learn more.

For more information

Try before you buy

Anyone with a Cisco ID or CCO ID may log into the Cisco Nexus Hyperfabric cloud controller at hyperfabric.cisco.com to request an organization identifier; they can then begin building network fabric blueprints for free.

For more information, visit: cisco.com/go/nexus-hyperfabric-ai.

© 2025 Cisco and/or its affiliates. All rights reserved. Cisco and the Cisco logo are trademarks or registered trademarks of Cisco and/or its affiliates in the U.S. and other countries. To view a list of Cisco trademarks, go to this URL: www.cisco.com/go/trademarks. Third-party trademarks mentioned are the property of their respective owners. The use of the word partner does not imply a partnership relationship between Cisco and any other company. (1110R)

C78-5394501-00 09/25