# AI Solutions with Cisco Nexus Hyperfabric

## (Formerly Nexus Hyperfabric AI)

Contents

Overview and definitions .....2

Deployment guidelines .....6

How to purchase .....9



© 2026 Cisco and/or its affiliates. All rights reserved.

# Overview and definitions

**Q: What is Cisco Nexus Hyperfabric™?**

**A:** Cisco Nexus® Hyperfabric is the cloud-managed Cisco Nexus One network operating model. It is a data-center network fabric solution delivered as a service. Using the Nexus Hyperfabric cloud controller managed by Cisco, customers easily design, deploy, and manage any number of fabrics located anywhere, spanning primary data centers, colocation facilities, and distributed data-center edge sites. It reinvents the IT operations lifecycle of the data center by simplifying every step of the process and ensures repeatable and predictable outcomes by IT-generalist, application, and DevOps teams. The vertical stack solution consists of purpose-built hardware, software, cloud management, day-2 automation, and Cisco support.

**Q: What is the key capability of the cloud-managed Nexus Hyperfabric network?**

**A:** It provides a flexible, cloud-managed Nexus Hyperfabric network that can connect to any AI server and storage platform, enabling simplified operations and high-performance networking tailored for AI workloads.

**Q: What are the two AI infrastructure options available with Nexus Hyperfabric?**

**A:** One option offers a cloud-managed Nexus Hyperfabric network designed to flexibly connect to any AI server and storage platform, providing simplified operations and high performance tailored for AI workloads. The other option provides a full-stack AI infrastructure solution that includes Cisco UCS servers with NVIDIA GPUs, partner storage, and networking components, all managed through a cloud controller for a turnkey AI infrastructure experience.

**Q: How does the cloud-managed Nexus Hyperfabric network option benefit AI deployments?**

**A:** This option delivers a flexible, modular Ethernet fabric architecture that supports scalable and multi-tenant AI deployments by seamlessly connecting diverse AI servers and storage platforms. It offers low-latency, lossless backend networking optimized for GPU clusters and AI workloads, with centralized cloud management for simplified lifecycle operations.

**Q: What advantages does the full-stack AI infrastructure solution provide?**

**A:** The full-stack solution offers a pre-validated, modular platform integrating compute, GPU acceleration, storage, and networking. It reduces deployment time from months to days, complies with NVIDIA Enterprise Reference Architecture (ERA) and NVIDIA Cloud Partner (NCP) Reference Architecture, and uses Cisco Intersight® for unified cloud-managed operations, enabling automated provisioning and consistent operational control.

**Q: How do these options simplify AI infrastructure lifecycle management?**

**A:** Both options leverage cloud-managed controllers for centralized design, deployment, monitoring, and upgrades. This streamlines operations from initial design through scaling and ongoing management, reducing complexity and operational overhead for IT and AI teams.

**Q: What is Cisco Nexus Hyperfabric's full stack AI Infrastructure option?**

**A:** Cisco Nexus Hyperfabric's full stack AI Infrastructure option is a cloud-managed, full-stack AI infrastructure solution delivered as hardware + software + service that is compliant with NVIDIA Enterprise Reference Architecture (ERA). Using a cloud controller managed by Cisco, customers easily design, deploy, and manage their on-premises network fabric, GPU servers, and storage servers. It reinvents the IT operations lifecycle of the AI infrastructure

© 2026 Cisco and/or its affiliates. All rights reserved.

in the data center by simplifying every step of the process, and ensures repeatable and predictable outcomes by IT-generalist, application, and DevOps teams. The vertical stack solution consists of purpose-built hardware, software, cloud management, day-2 automation, and Cisco support. Cisco Nexus Hyperfabric's full stack AI Infrastructure option is best suited for customers looking to build out their private cloud AI infrastructure.

**Q: Why is Cisco Nexus Hyperfabric AI transitioning to Cisco Nexus Hyperfabric with an optional full stack AI Infrastructure?**

**A:** Cisco Nexus Hyperfabric AI is now "Cisco Nexus Hyperfabric with a full stack AI infrastructure option" to make the role and capabilities of the platform clear and easy to understand and better reflect the actual offering. By keeping the name focused on "Nexus Hyperfabric," Cisco emphasizes that Cisco Nexus Hyperfabric is a powerful data-center network controller, not an AI bundle. Cisco Nexus Hyperfabric stands out for its unique ability to provide deep visibility across the data center with unique capabilities including visibility down to the server and GPU level within the data center, but its core function is to manage and orchestrate network infrastructure, regardless of the workloads running on top.

For organizations seeking a turnkey infrastructure specifically for AI workloads, Cisco offers a full stack AI infrastructure with Nexus Hyperfabric as an optional solution. The AI infrastructure bundles together the necessary servers, GPUs, and optional storage required for AI applications, simplifying procurement and deployment for customers with specialized needs. This clear distinction helps customers understand that, whereas Nexus Hyperfabric is essential for managing modern, high-performance data center networks—including those that support AI—the addition of the AI infrastructure transforms the solution into a ready-to-deploy AI infrastructure. This approach better reflects the actual offering.

**Q: What is the Cisco N9100 switch?**

**A:** The Cisco N9100 Series Switches are introduced as a new product in the Cisco Nexus Hyperfabric portfolio. The Cisco N9100 is a scale-out fabric switch powered by NVIDIA Spectrum-4 Ethernet technology, designed for high-performance AI data center environments. It supports 800 GbE connectivity and is compliant with the NVIDIA Cloud Partner (NCP) Reference Architecture for secure AI factory deployments.

**Q: What is Cisco Secure AI Factory?**

**A:** Cisco's Secure AI Factory, developed in collaboration with NVIDIA and strategic ecosystem partners, is a modular reference design that accelerates the delivery of trusted and transformative AI applications for enterprises. It combines high-performance AI infrastructure with built-in security, AI-powered monitoring, and advanced software tools to streamline the deployment and operation of AI and generative AI pipelines. This solution enables organizations to rapidly build and deploy predictive, generative, and reasoning-based AI applications, resulting in faster time to value, trusted AI outcomes, and new business opportunities through secure and scalable AI. The Secure AI Factory embeds security and observability at every layer of the stack, continuously monitoring and analyzing the security posture to provide advanced threat detection and incident response, distinguishing it from other AI infrastructure solutions.

The Secure AI Factory integrates Cisco's market-leading networking, accelerated compute, and scalable storage with AI orchestration and application software, forming a secure and efficient AI infrastructure foundation. Key security capabilities include Cisco AI Defense for runtime protection of generative AI applications, Cisco Hypershield™

© 2026 Cisco and/or its affiliates. All rights reserved.

for workload security, and Cisco® Hybrid Mesh Firewall for unified security management across multiple enforcement points. The solution offers flexible, modular deployment options, allowing enterprises to operationalize AI infrastructure at their own pace, either through pre-validated AI infrastructure stacks or by integrating Cisco and partner products independently. This comprehensive, security-first architecture supports enterprises in confidently navigating the evolving AI landscape with exceptional flexibility, enhanced security, and superior performance.

**Q: Why did we bring Cisco Nexus Hyperfabric's full stack AI Infrastructure option to market?**

**A:** The adoption of AI is increasingly viewed by organizations as a key enabler to drive innovation: "By 2027, 40% of enterprises will deploy GenAI network fabrics to enable cost and performance-optimized support for AI workloads in their own data centers" (IDC Perspective, March 2024). Our findings show us that 95% of customers are aware that AI will increase workloads, but only 17% are equipped to handle this increase, with 23% having limited or no capacity to meet the AI demand with current infrastructures. Given the significant business value and growth in AI adoption across enterprises, Cisco is working closely with NVIDIA and partners to ensure that customers can rapidly and reliably deploy AI wherever it is needed. This is in addition to simplifying the IT lifecycle enabling IT generalists, data science teams, and DevOps teams to easily design, deploy, and operate AI and non-AI data center fabrics. The solution is based on a converged Ethernet network so that organizations can easily support and scale by leveraging existing skills and processes.

**Q: What are the components of Cisco Nexus Hyperfabric's full stack AI Infrastructure option?**

**A:** Cisco Nexus Hyperfabric's full stack AI Infrastructure option consists of these primary components:

- Cisco Nexus Hyperfabric cloud controller: a scalable, globally distributed multitenant cloud service that is used to design, plan, control, upgrade, and monitor fabrics using a browser or APIs
- Cisco 6000 Series Switches, Cisco N9100 and selective 9300 Series Switches (coming soon) deployed with Cisco Nexus Hyperfabric's cloud-managed software, they connect to the cloud for centralized real-time visibility and control.
- Cisco UCS® NVIDIA GPU/DPU servers: Cisco UCS-C885A-M8-CN1 server that packs 8 NVIDIA H200 GPUs, with NVIDIA BlueField-3 DPU (3240H) and NVIDIA SuperNIC (3140H) connected using Cisco optics that can efficiently run training/inferencing/fine-tuning jobs. In addition, UCS 880 equipped with HGX B300 is coming soon.
- Cisco UCS-VAST storage (optional): Cisco UCS-C225-M8N-1P servers pack eight drives with each of the servers, with a cluster consisting of 11 servers that can be easily expanded. The specs start with 1 PB of storage that can be used for training, fine tuning, inferencing, Retrieval-Augmented Generation (RAG), and other data engineering work.
- NVIDIA AI Enterprise software (NVAIE): integration and access to the NVAIE software stack from day 1, allows access to different model-training software and data catalogues so that engineers and scientists can get started immediately. The Cisco Nexus Hyperfabric's full stack AI Infrastructure option is a single integrated solution from Cisco that can be designed, ordered, deployed, and operated as a cohesive solution.
- Cisco Customer Experience (Cisco CX) professional devices: professional services from Cisco CX including rack and stack.

© 2026 Cisco and/or its affiliates. All rights reserved.

**Q: What are the capabilities of Cisco Nexus Hyperfabric's full stack AI Infrastructure option?**

**A:** Cisco Nexus Hyperfabric's full stack AI Infrastructure option integrates these components into a single solution:

- Its cloud controller, operated by Cisco, serves as a single point for configuration, monitoring, and maintenance of all tenant customer fabrics, and provides health monitoring and visibility of GPU servers and storage servers, using real-time connection to switches and servers deployed either on premises or in colocation facilities.
- Network fabrics consist of cloud-managed Cisco 6000 Series Switches, Cisco N9100 and 9300 Series Switches (coming soon) that offer automated zero-touch provisioning, and a "helping hands assistant," which provides step-by-step cabling instructions combined with real-time verification.
- Network fabric spans the GPU servers as well as storage servers and can be set as a mesh or leaf-spine configuration. With RDMA over converged Ethernet (RoCE) Version 2, your investment in network hardware and expertise is protected.
- The GPU cluster consisting of Cisco UCS-C885A-M8-CN1 servers is highly scalable and flexible so that you can adapt the cluster size to your needs. In addition, the UCS 880 equipped with HGX B300 is coming soon.
- Optional storage cluster, consisting of Cisco UCS-C225-M8N-1P high-density storage servers built on the latest NVMe drives ensure fast Input/Output Operations Per Second (IOPS) that feed data to GPUs using Cisco optics so that you get the most value from your GPU investments.
- In a shared-responsibility model, automation and operations from Cisco support manage the cloud controller, the fabric underlay and overlay networks, and the software upgrade process, while customers maintain direct control of all interconnections to their applications, hosts, and the rest of their network.
- Assertion-based monitoring continuously verifies the availability and reliability of the fabric and connected resources and makes it easier to detect the root cause of any issue. A built-in designer helps customers construct a validated fabric design based on desired compute, storage, host and port capacity, oversubscription, and environmental considerations including cabling and power, and then creates an accurate Bill of Materials (BoM).
- Self-service fabric tenancies empower host and application teams to monitor and manage the fabric services they have been allocated, removing the need to depend on IT for most support services.

**Q: Which partnerships are part of Cisco Nexus Hyperfabric's full stack AI Infrastructure option solution?**

**A:** The Cisco Nexus Hyperfabric's full stack AI Infrastructure option cluster solution leverages the following partners:

- NVIDIA AI for Enterprise (NVAIE)
- NVIDIA for GPUs and Network Interface Cards (NICs) based on Hyperscale Graphics eXtension (HGX) reference architectures
- Optional VAST storage software on Cisco UCS storage servers

**Q: What is Cisco Nexus Hyperfabric's full stack AI Infrastructure option compliance with NVIDIA Enterprise Reference Architecture?**

**A:** Cisco Nexus Hyperfabric's full stack AI Infrastructure option is compliant with NVIDIA AI Enterprise Reference Architecture (ERA) to establish best practices and guidelines for building high-performance, scalable, and secure infrastructures in enterprise environments. This architecture integrates processors, GPUs, servers, networking devices, storage, management tools, and software into an NVIDIA-validated design. It provides customers with a vertically integrated solution, ensuring optimized performance and seamless scalability.

© 2026 Cisco and/or its affiliates. All rights reserved.

**Q: How will Cisco Nexus Hyperfabric's full stack AI Infrastructure option with NVIDIA ERA be managed?**

**A:** The primary management tool for ERA-compliant Cisco Nexus Hyperfabric's full stack AI Infrastructure option solutions is the Nexus Hyperfabric cloud controller.

- Nexus Hyperfabric cloud controller: provides assertion-based, high-level management and control across all three layers: network, GPUs, and storage
- Cisco Intersight® (purchased separately): provides detailed management of GPU servers and storage servers
- NVIDIA AI Enterprise software: provides AI/ML tools tailored for data scientists
- VAST Data (optional): performs data- and storage-management tasks. These management capabilities will include comprehensive fabric control and leverage NVIDIA's proprietary insights, directly integrated into Cisco Silicon One®, to enhance visibility into workload connectivity. The collaboration and integration between Cisco and NVIDIA products are expected to deepen further over time.

**Q: Why is Fabric as a Service so important? The hardware is all on premises.?**

**A:** AI infrastructure is very expensive. Acquisition costs are high, as are operational costs for space, power, and cooling, so customers need to see a fast return on their investment. This means many enterprise customers need to get their AI jobs running as fast as possible. Bringing up a new AI cluster will require the coordination of MLOps, NetOps, IT, facilities, and SecOps skills that might be in short supply after years of moving workloads to the cloud. Working with a trusted partner such as Cisco will enable customers to act quickly and with greater confidence in this rapidly evolving area. With Cisco Nexus Hyperfabric full stack AI Infrastructure option as a Fabric as a Service, it allows networking to be managed by teams that don't have deep networking skills to manage the network and focus more on innovation.

**Q: What about other storage partners? I don't use VAST.?**

**A:** VAST storage is our partner at launch. Customers are welcome to use their own storage vendor through the North/South (N/S) interface into the cluster; however, it will not be managed by Nexus Hyperfabric.

## Deployment guidelines

**Q: Is the Cisco 6000 Series Switch the same as the old Cisco Nexus 6000 Series Switch family?**

**A:** They are not the same. The old Cisco Nexus 6000 Series Switches have reached end-of-life and end-of-support and are not supported or compatible with Cisco Nexus Hyperfabric's full stack AI Infrastructure option.

**Q: Can Cisco Nexus Hyperfabric's full stack AI Infrastructure option manage Cisco Nexus 9000 Series Switches?**

**A:** The initial release of the Cisco Nexus Hyperfabric's full stack AI Infrastructure option does not support Cisco Nexus 9000 Series Switches. The new Cisco 6000 Series Switches are required. Support for the Cisco N9100 and N9300 Series Switches has been announced and are coming soon.

**Q: At a high level, how does Cisco Nexus Hyperfabric's full stack AI Infrastructure option work?**

**A:** Cisco Nexus Hyperfabric's full stack AI Infrastructure option works as a cloud-managed, turnkey solution that integrates networking, compute, GPUs, storage, and AI software into a vertically integrated platform

© 2026 Cisco and/or its affiliates. All rights reserved.

compliant with NVIDIA Enterprise Reference Architecture (ERA). Customers begin by logging into the Cisco Nexus Hyperfabric's full stack AI Infrastructure option designer, where they can start designing their AI infrastructure by selecting a pre-validated template that suits their requirements. Once these specifications are entered, the designer automatically creates a validated, complete fabric topology, including the necessary Cisco UCS compute servers, storage servers, optics, cabling, and switches. Cisco Nexus Hyperfabric's full stack AI Infrastructure option seamlessly integrates with Cisco's ordering tools to ensure there are no errors when converting the design into a bill of materials. The bill of materials also includes Cisco's day-0/1 installation and deployment services, as well as software subscription licenses for Cisco Nexus Hyperfabric, VAST storage software, and NVIDIA NVAIE. When the Cisco 6000 Series Switches, Cisco UCS Rack Servers, and Cisco UCS 225 M8 storage servers arrive on site, the Cisco Customer Experience (Cisco CX) team or partners will handle the rack and stack (day 0/1) to complete the installation and deployment process. Once deployed, the switches automatically connect to the cloud for claiming and provisioning by the cloud controller through zero-touch plug-and-play. This process quickly sets up a fully operational AI infrastructure, including the network fabric, in just minutes. Assertion-based monitoring continuously checks the availability and reliability of the AI infrastructure and its connected resources, immediately identifying the root cause of any issues. If changes are needed to adjust the capacity or design, customers can easily modify the in-progress design, approve the updates, and follow the same process again. The product offers guidance on all necessary physical changes to transition from the old topology to the new, including cabling adjustments, and automatically reconfigures itself.

**Q: What deployment use cases does the first release support?**

**A:** The initial release is designed to support the following use cases:

- Initial deployment of Gen AI infrastructure and Large Language Model (LLM), fine-tuning, and RAG-type deployments, supported on day 1.
- Limited Gen AI infrastructure expertise (that is, both depth and breadth) to build infrastructure expertise to support Gen AI workloads.
- Where capacity needs are unknown, the capability to scale as needs become known, rather than overbuilding upfront.
- Sharing of the infrastructure among multiple groups for training, fine-tuning, RAG, and data-engineering work.

**Q: Where is the Cisco Nexus Hyperfabric's full stack AI Infrastructure option cloud controller hosted?**

**A:** The Cisco Nexus Hyperfabric cloud controller is maintained by Cisco. The cloud controller includes global scalability, as well as multi-region redundancy, without any additional configuration for end users. The cloud controller is reachable through a single URL (https://hyperfabric.cisco.com/) and is used for all communications (including switch-management cloud connectivity, primary user interfaces, and RESTful API endpoints).

**Q: What are the plans for the cloud controller hosting outside the United States?**

**A:** Cisco plans to offer local cloud controllers hosted natively in Europe and Asia in future releases.

**Q: Is Cisco Nexus Hyperfabric's full stack AI Infrastructure option strictly cloud-based, or are there any plans for air-gapped environments?**

**A:** The cloud controller is managed by Cisco operations staff. There are no plans to hand off that responsibility to entities not owned or operated by Cisco.

© 2026 Cisco and/or its affiliates. All rights reserved.

**Q: What are the plans for government cloud compliance (under the Federal Risk and Authorization Management Program [FedRAMP])?**

**A:** Cisco is investigating offering local cloud controllers hosted in FedRAMP in a future release and will consider other government cloud-compliance environments.

**Q: How can I export logging to my own logging platforms?**

**A:** The Cisco Nexus Hyperfabric cloud controller will support external cloud-to-cloud integrations for delivering logging information through external Amazon S3 buckets to logging files in syslog format. In addition, the cloud controller will also allow API integration for querying internal states, alerts, and other telemetry data from the on-premises fabrics.

**Q: What are the bandwidth requirements for cloud connectivity?**

**A:** Target bandwidth per switch is less than 2Mb/sec at steady state. Certain types of real-time monitoring may increase upstream bandwidth when initiated by end users through the cloud controller UI.

**Q: How is real-time telemetry uplifted to the cloud controller?**

**A:** The Cisco 6000 Series Switches include a telemetry agent that establishes an outbound TLS connection to the cloud controller. This outbound Transport Layer Security (TLS) (TCP/443) session is designed to work with existing network security controls for standard web clients and provides all connectivity required for configuration, monitoring, and real-time telemetry for each switch managed by Cisco Nexus Hyperfabric's full stack AI Infrastructure option.

**Q: What happens if I lose cloud connectivity?**

**A:** The cloud connection from a deployed fabric to the cloud controller is used for management and telemetry purposes only. All stateful protocols are maintained independently on the switches residing on premises. Any fabric disconnected from the cloud will continue to operate normally (including local underlay and overlay fabric-management protocols, external peering protocols such as BGP, and all standard bridging functions). When disconnected from the cloud, a fabric will not be able to accept configuration updates, and telemetry will not be received by the cloud controller. The on-switch agent will intelligently buffer telemetry if disconnected and will upload this information and resynchronize any configuration updates once the fabric is reconnected to the cloud controller.

**Q: How can I interact with the Cisco Nexus Hyperfabric's full stack AI Infrastructure option API?**

**A:** The cloud controller implements a RESTful API using the same URL as standard management functions do. This API is extensively documented in an online API programming guide and is designed to be compatible with OpenAPI specifications. In addition, the Cisco Nexus Hyperfabric's full stack AI Infrastructure option will support providers for both Ansible and Terraform to help users exercise the API according to their existing management tooling.

**Q: Who will a customer contact for technical assistance support for the cluster?**

**A:** Cisco Technical Assistance Center (Cisco TAC)

**Q: What are the power requirements?**

**A:** In the range of 10-16kW/server per GPU server. Storage servers, optics, and switches will require additional power.

© 2026 Cisco and/or its affiliates. All rights reserved.

**Q: How are the components of Cisco Nexus Hyperfabric's full stack AI Infrastructure option connected?**

**A:** AI clusters are generally built around multiple logical network fabrics. The primary fabrics are distinct in their optimization based on which function they provide. Logically, there are multiple network fabrics that form the AI cluster network:

- A backend network interconnecting GPUs across the servers.
- A frontend network connecting the servers to the rest of the applications and end users.
- A storage network connecting the GPU servers to storage servers.
- A management network connecting to each switch and server in the network and used by Cisco Nexus Hyperfabric. These networks make up a single cluster, controlled by a single instance of Cisco Nexus Hyperfabric's full stack AI Infrastructure option.

**Q: Is the cluster air cooled?**

**A:** Yes. Future versions may require liquid cooling for large clusters and faster GPUs.

## How to purchase

**Q: When will Cisco Nexus Hyperfabric's full stack AI Infrastructure option be available?**

**A:** Cisco Nexus Hyperfabric's full stack AI Infrastructure option is currently available for order.

**Q: How do I buy Cisco Nexus Hyperfabric's full stack AI Infrastructure option?**

**A:** Cisco Nexus Hyperfabric's full stack AI Infrastructure option may be purchased from a certified reseller. Organizations able to purchase products directly from Cisco may also purchase the solution. In addition you can design your data center fabric before buying it at https://Hyperfabric.cisco.com.

**Q: How do I design a data center fabric before buying it?**

**A:** Anyone with a Cisco Connection Online (CCO) account or with an Identity Provider (IdP) tied to CCO may request and receive an organization tenancy in the cloud controller located at https://Hyperfabric.cisco.com.

Within the tenancy, customers may design blueprints that contain all the details needed to order, deploy, configure, and operate the fabric, including:

- Physical components including switches, optics, compute servers, storage servers, connectors, air-flow direction, power consumption, and a cabling plan
- Bill of materials for all Cisco and partner components that is the source of truth and integrated with Cisco Commerce to automate the process of obtaining an accurate quote
- Logical network including the entire network overlay and underlay and upstream route peering
- Host-port assignments to server infrastructure teams
- API integration to automate the provisioning of the blueprint Additional accounts may then be added to that tenancy so teams can collaborate. Once the equipment is deployed, it is automatically provisioned according to the blueprint.

**Q: Can Cisco Nexus Hyperfabric's full stack AI Infrastructure option be sold only by certified resellers?**

**A:** Yes, only Cisco Nexus Hyperfabric's full stack AI Infrastructure option—certified resellers may sell the product.

**Q: Where can I get more information?**

**A:** More information may be found at www.cisco.com/go/nexus-hyperfabric-ai.

Some products and features described are in development and offered on a when-and-if-available basis. Cisco reserves the right to change delivery timelines and will have no liability for any delays or failures to deliver.

© 2026 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)
C67-5040175-03 06/26