Executive Summary

Building the critical infrastructure for the AI economy—together

Organizations in every industry are rapidly evolving their strategies to harness the full potential of generative AI and, more recently, agentic AI. An explosion of investment is fueling efforts to meet the soaring demand for AI infrastructure to host, interconnect, and protect all aspects of AI systems, from model training, fine-tuning and inference workloads to complete agentic workflows. The AI industry has been pushing the boundaries of scale, with an initial focus on creating massive data centers and AI clusters to efficiently process the training of increasingly intelligent and diverse large language models (LLMs). While this buildout of foundational models continues unabated, there is a new wave of investment in the inference and agentic infrastructure required to enable distributed AI workflows across industries.

Infrastructure investments supporting evolving AI modalities continue to grow exponentially across all ecosystem participants (Figure 1), with global capital investment in data center IT equipment between 2025 and 2030 forecast to reach as much as $4.7 trillion.[1] This includes investments in servers, AI accelerators such as graphics processing units (GPUs) and tensor processing units (TPUs), networking, optics, and storage. While today, hyperscalers such as Microsoft, Google, Meta, Amazon, and Oracle are responsible for over 60% of the infrastructure investment, that share is forecast to decrease to 45% in 2035.[2] Other AI cloud providers like neocloud providers, sovereign cloud providers, communications service providers (CSPs), and colocation providers, together with enterprises themselves, are becoming responsible for a growing share of the investment, especially as inference and agentic AI move into the mainstream.

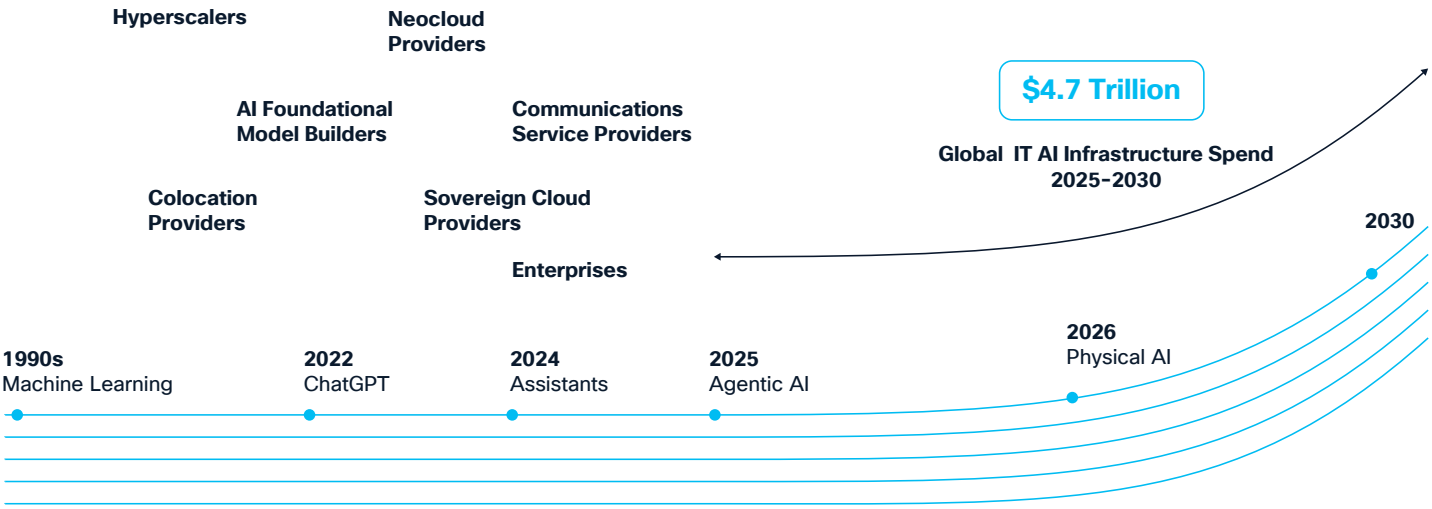

Figure 1: AI continues to drive huge investment across all AI cloud provider and enterprise segments.

A total global investment of $4.7T in data center IT infrastructure (incl. servers, accelerators, networking, optics, storage) is forecast between 2025–2030

1. The cost of compute: A $7 trillion race to scale data centers, McKinsey & Company, April 2025.

2. AI Infrastructure Survey, McKinsey & Company, June 2025.

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

Cisco is partnering with participants across all AI cloud provider segments to build efficient, high-performance, and resilient service offerings. We offer a broad and cutting-edge AI infrastructure portfolio, strong technology partnerships, deep go-to-market and financial support, alongside long-standing provider and enterprise relationships, and advanced services. By partnering with Cisco, AI cloud providers can accelerate time to market and growth, minimize risk, optimize efficiencies, and maximize return on investment (ROI).

Specifically, Cisco's extensive AI Infrastructure portfolio, both within and between AI data centers, delivers innovation, choice, and reliability for organizations delivering in-demand AI services at scale. It comprises AI-optimized infrastructure for uncompromised performance, robust security and compliance, and automated operations. By selecting the appropriate products, solutions, and reference architectures, AI cloud providers can support the full spectrum of AI services and consumption models demanded by a diverse customer base. They also gain the flexibility they need to quickly add new differentiated service models as market requirements shift.

Cisco works with a broad ecosystem of leading technology providers across software, AI accelerators (XPUs), storage, power, and cooling to address the comprehensive technology and business requirements of AI cloud providers. In particular, Cisco's expanded partnership with NVIDIA provides AI cloud providers with a powerful combination of Cisco and NVIDIA technology and reference architectures. With Cisco Powered Services, we collaborate on delivering the designs, operations, and systems providers need to optimize their offerings.

Cisco collaborates with AI cloud providers from every segment to help them achieve technological differentiation and sustainable go-to-market success for the diverse and evolving services they build and deliver.

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

Enabling the AI infrastructure ecosystem

As with any major technological shift, AI is generating vast opportunities across the entire ecosystem. To deliver value and effectively monetize AI in this rapidly evolving market, it's essential for organizations to continually innovate and stay ahead of transformational trends.

Key drivers for large-scale AI infrastructure buildouts

1. **Generative AI.** The race to train and deliver the most intelligent and competitive LLMs requires continued investment in the latest, most powerful accelerated compute clusters, which span one or more AI data centers. Between now and 2030, the number of foundational and specialized LLMs is expected to grow ten-fold, driven by industry-specific needs, state-sponsored AI programs, and fine-tuned enterprise models.
2. **Enterprise AI adoption.** Enterprises are determining their own AI strategies for LLMs, inference, and agentic workflows today. They are choosing whether to acquire AI infrastructure services, use colocation facilities, or, in some cases, upgrade their existing infrastructure to host the compute-heavy workloads at the core of their AI systems. Inferencing workloads are projected to surpass training to become the dominant workload in AI data centers,[3] driven by adoption of agentic AI by enterprises. This is causing AI cloud providers to reconsider their data center cluster design, location, and connectivity needs. Enterprises are evaluating hyperscalers, neoclouds, colocation providers and on-premises infrastructure options to support their AI needs. Many are choosing a hybrid model.
3. **Power constraints.** By the early 2030s, nearly 11,000 data centers will be in operation by AI cloud and colocation providers worldwide.[4] These data centers will demand huge power and cooling capacity, with specialized AI clusters requiring as much as 150 kW per rack today expected to grow to above 1 MW by the end of the decade. For this reason, many data centers are being built in remote locations where cheaper land and more sustainable power are more readily available, driving demanding new data center interconnectivity requirements.
4. **Sovereign data requirements.** As governments in many parts of the world tighten regulations to ensure private data stays within national boundaries or complies with cross-border data-transfer regulations, data sovereignty is emerging as a significant concern. This is especially true for industries that are considered "critical infrastructure," such as financial institutions, healthcare providers, public services, and energy networks. Policy and security demands and customer priorities require trust in more local AI data centers with clear and demonstrable control over access to systems and data.
5. **Edge AI.** There is a significant shift underway in data processing, moving from centralized data centers towards the on-premises edge. IDC projects that global spending on edge computing will reach nearly

> By 2030, inference will surpass training to become the dominant workload in AI data centers, representing more than half of all AI compute and roughly 30 to 40 percent of total data center demand.
>
> McKinsey & Company

3. The next big shifts in AI workloads and hyperscaler strategies, McKinsey & Company, December 2025.

4. AI Infrastructure: A new growth opportunity for telco operators, McKinsey & Company, February 2025.

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

$380 billion by 2028, driven by the need to process and analyze data locally.[5] For AI cloud providers, this trend underscores the growing importance of expanding highly centralized services out to edge environments to meet the changing demands of AI.

6. **Talent shortages.** There is a widely recognized global shortage of skilled AI professionals, including AI infrastructure specialists, data scientists, and AI/ML developers. This scarcity presents significant challenges for AI cloud providers like neoclouds and sovereign cloud providers, as they compete for a limited pool of experts. An in-house talent shortage could ultimately slow innovation and limit the ability of AI cloud providers to deliver the services customers need, with the reliability, security, and immediacy they expect.

In this climate, emerging AI cloud providers are challenged with choices that impact their business success, including:

- Identifying and building profitable, competitive, and differentiated services that will meet the needs of the market segment they are targeting, with the understanding that conditions and requirements are dynamic.
- Building a business model and go-to-market strategy that supports continued investment while delivering sustainable growth in this capital-intensive, fast-growing, and competitive market.
- Choosing partners that can address both their near-term and long-term technology, business, financing, and go-to-market needs.

Cisco delivers the most comprehensive and advanced set of secure AI infrastructure technologies and systems designed to enable participants in each segment of the AI economy (Figure 2) to innovate and succeed. Beyond technology, Cisco has programs in place to help address the financial and go-to-market constraints many organizations struggle with. Each participant has a unique set of priorities and challenges, making it imperative for technology partners like Cisco to be able to respond and adapt to their individual requirements.

Hyperscalers and neocloud providers

Cisco is helping AI cloud infrastructure providers build enormous clusters that support training and inferencing of compute-hungry LLMs. **Hyperscaler cloud providers,** including Microsoft, Google, Meta, and Amazon are estimated to have spent as much as $350

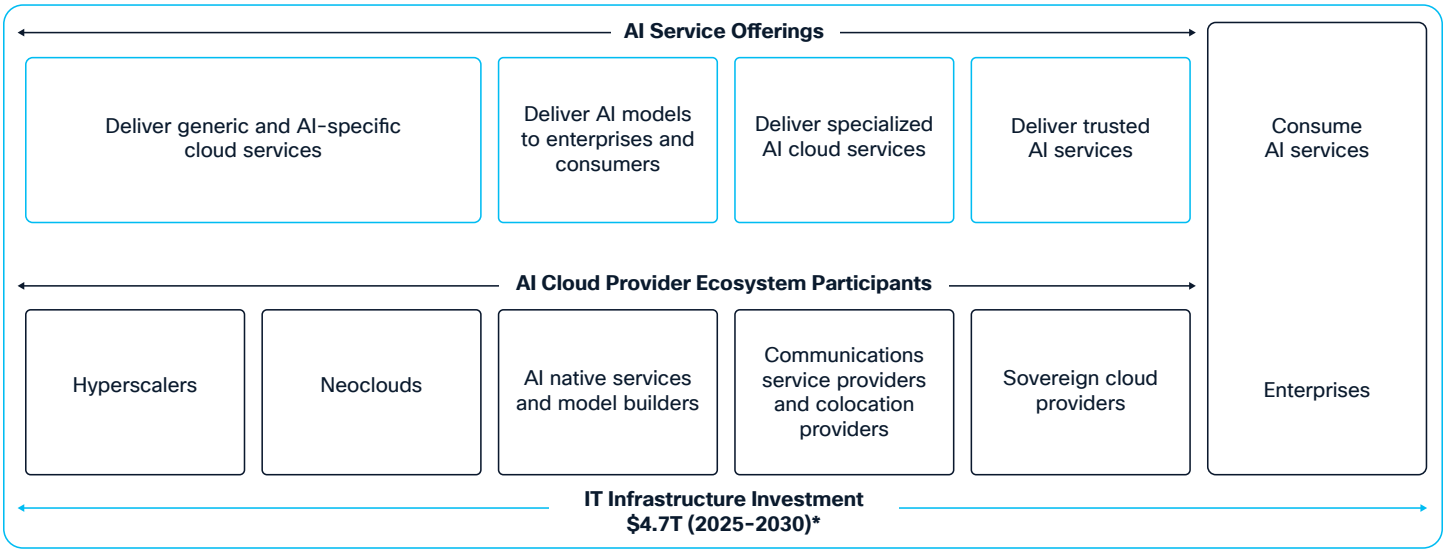

Figure 2: Cisco enables AI service offerings across all segments of the AI value chain.

5. Worldwide Edge Spending Guide, IDC, March 2025.

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

Billion on AI infrastructure in 2025.[6] Cisco's advanced network silicon, switching, and optics technologies are helping hyperscalers deploy their own custom-designed mass-scale AI clusters. In the first quarter of FY26 alone, Cisco announced orders of more than $1.3 billion in AI infrastructure from hyperscalers.

Likewise, **neocloud providers** are employing Cisco technologies and solutions to build specialized processing capacity that scales up to 100K graphics processing units (GPUs) and beyond. Their goal is to accelerate the readiness and resilience of their service offerings for training LLMs and providing inferencing services for their enterprise customers. Together with our technology partners, Cisco is uniquely positioned to deliver AI infrastructure built on validated reference designs. With Cisco, AI cloud providers like neoclouds have the flexibility to choose the reference architecture and operating environment that best aligns to their business models and technical requirements. Whether they want to adhere to the NVIDIA Enterprise Reference Architecture (ERA), the NVIDIA Cloud Partner (NCP) reference architecture, Cisco's Cloud Reference Architecture, or simply want to build their own, we provide the flexibility, validated designs, and support to accelerate deployment. Cisco's expanded partnership with NVIDIA results in close collaboration that lets AI cloud providers take advantage of NVIDIA Spectrum-X Technology integration and NVIDIA reference architecture compliance, together with the proven benefits of Cisco switching and operating systems.

AI cloud providers are looking to get the highest return on their infrastructure investments by accelerating time to service readiness and maximizing the operational efficiency, AI cluster performance and reliability of their infrastructures. With greater efficiency, performance, and reliability comes faster job completion times, better service levels, and increased revenue. In such a competitive and dynamic market, it's also crucial for providers to have the flexibility and choice of innovative technologies to create diverse consumption models and differentiated services that will attract and retain enterprise customers.

Communications service providers and colocation providers

Communications service providers (CSPs) have an opportunity to become pivotal players in the AI economy. This is especially apparent now, as the industry shifts its primary focus from capital-intensive training of large language models to widespread enterprise AI adoption and consumption of LLMs. The most natural route to revenue for these well-established telecommunications operators is via services that build upon existing telecom capabilities to deliver, optimize, and assure their customers' AI applications and workflows through intelligent networking services. For CSPs ready to expand into new territory, AI also creates a plethora of high-value service opportunities that include AI workload hosting, managed edge offerings, data sovereignty, monitoring, security services, and more.

The Cisco AI infrastructure portfolio is well designed to help CSPs leverage their existing and new assets to enable organizations to integrate AI into their business processes. CSPs can leverage regional data centers

> Cisco and NVIDIA are leading the way in defining the technologies that will power these AI-ready data centers in all their varieties, from emerging neoclouds, to global service providers, to enterprises, and beyond.
>
> Jeetu Patel, President and Chief Product Officer, Cisco, November 2025

6. Could 2025 Represent a Near-Term Peak in AI Capex?, Forbes, April 2025.

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

that are typically located closer to their customers to host AI infrastructure for a variety of purposes, including inference, sovereign data, and edge requirements. Learn more about AI opportunities for CSPs and how Cisco can help.

Meanwhile, **colocation providers** can use their state-of-the-art data centers to meet the challenging space, power, and cooling demands of AI workloads, and address the growing deficit in AI data center infrastructure. According to McKinsey, demand for AI-ready data center capacity will account for 70% of total data center demand by 2030 and is the main driver of a growing AI infrastructure deficit.[7] At the same time, colocation providers are providing high-performance network hubs that offer on-ramps to the growing array of AI cloud providers.

Cisco is well positioned to support colocation providers with hosting and building complete secure AI infrastructure stacks that can accelerate enterprise workload deployments. These stacks would often be too costly, time consuming or unviable to deploy in their customers' own data centers. In addition, Cisco can help colocation providers enhance their carrier-neutral interconnectivity hubs to ensure the performance, reliability, and security of their customers' AI workflows with direct, low-latency connectivity to hyperscalers and neoclouds.

Sovereign cloud providers

In many parts of the world, most notably Europe and parts of Asia, concerns about digital sovereignty for AI have risen to the top of the agenda. This is driving investments in AI infrastructure that must meet stringent security, compliance, performance, and operational requirements. As demands for localized trusted AI rise, regional **sovereign cloud providers** are offering AI services directly to organizations or alternatively through collaboration with hyperscalers and neocloud providers.

Cisco has built a Sovereign Cloud Operating Framework to meet the diverse sovereignty needs of these providers and their customers. This framework allows for secure and compliant cloud operations tailored to specific regulatory requirements. It ensures policy-driven governance and centralized management, fostering greater control and operational efficiency. The framework supports different sovereign cloud models depending on data classification requirements. For example, there are framework options for air-gapped, hybrid cloud, and multicloud models. Cisco integrates robust security into the model, providing zero trust and secure AI delivered through encryption, access control, and privacy-enhancing technologies.

Organizations concerned with data sovereignty also want to ensure that any data traversing the network outside the data center is routed intelligently and exclusively across compliant network routes and nodes that are deemed to be trusted. Cisco solutions facilitate compliance for AI traffic that leaves the data center by providing intelligent data center interconnect services that route traffic based on policy and deliver quantum-safe MACsec encryption.

Enterprises and public sector organizations

The majority of **enterprises and public sector organizations** rely on AI cloud providers and foundational model builders for a host of services, including model training and fine-tuning, foundational models and APIs, inferencing, and large-scale data processing. This allows them to focus their own efforts on applying AI to business processes, data, and differentiated use cases. A number of enterprises that have special requirements are building their own AI clusters that can scale to thousands of GPUs to support their training and inference needs. The Cisco AI infrastructure portfolio helps enterprises deploy and operate AI infrastructure at any scale in the most efficient and secure way. Cisco full-stack solutions comply with the NVIDIA Enterprise Reference Architecture (ERA) for those organizations that are building their clusters to the ERA specifications.

7. McKinsey AI expanding data center capacity report, October 2024.

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

Even those organizations building their own clusters will typically deploy a hybrid model with AI workloads and data distributed across private and public clouds. For enterprises using the services of AI cloud providers, choosing the right provider can be crucial. Provider services that are built on Cisco full-stack AI infrastructure and technologies enjoy the benefits of validated designs and deployments with the highest levels of reliability and security. They also offer more streamlined integration and on-boarding for those organizations with Cisco on-premises data center, interconnect, and security infrastructure.

AI infrastructure services delivery models

AI cloud providers are leveraging a variety of business models to bring AI infrastructure services to enterprises. Delivery and consumption models can be categorized into three main approaches that reflect diverse enterprise requirements for scalability, flexibility, cost, performance, and data locality. Individual providers will choose the Infrastructure as a Service (IaaS) and Platform as a Service (PaaS) approaches that best suit their business model and customer needs. The three main models are:

- **Reserved instances** (Dedicated or shared cloud AI IaaS): In this model, enterprises commit to using AI infrastructure within a public cloud provider's data center for a fixed term, typically one to three years. Reserved instances delivered on dedicated AI stacks or on shared infrastructure offer guaranteed capacity and promise better performance, and lower costs compared to on-demand pricing. This consumption model appeals to organizations with predictable, steady, and long-running AI workloads, such as large-scale model training, fine-tuning or sustained inference.
- **On-demand instances** (Shared public cloud AI IaaS): This model allows enterprise customers to consume AI-optimized compute resources, such as GPUs and tensor processing units (TPUs), from shared pools provided by a public cloud provider. Enterprises only pay for the resources they consume, typically on an hourly or minute-by-minute basis, without long-term commitments. This "pay-as-you-go" approach offers maximum flexibility and scalability, making it ideal for experimental workloads, development, testing, or bursty AI tasks where demand is not consistent. Providers deploy a cloud operating model that manages, scales, and secures the underlying multi-tenant infrastructure, allowing customers to focus on deploying and managing their AI software stack.
- **Serverless platforms** (Managed AI PaaS): In this model the cloud provider abstracts infrastructure management entirely, so that enterprises can consume AI resources and capabilities through managed platforms and APIs. The cloud provider handles provisioning, scaling, and lifecycle management, while customers pay based on usage metrics such as requests, tokens, or execution time. Managed AI PaaS is well suited for inference and application integration, especially where operational simplicity and elastic scaling are priorities.

While organizations can rely on centralized public cloud services, many are increasingly adopting hybrid approaches that span public cloud environments, on-premises or colocation data centers, and edge locations. Cloud providers may extend their IaaS and PaaS services to customer-owned hardware or edge devices, enabling enterprises and public sector organizations to retain control over sensitive data while benefiting from cloud-managed infrastructure for larger AI workloads. This allows the same cloud APIs and management frameworks to be used consistently across locations, even though the compute physically resides outside the public cloud. Balancing centralized AI with distributed AI capabilities closer to where data is generated or consumed can address specific customer requirements such as low-latency processing, data sovereignty, regulatory compliance, and reliability. This deployment approach is particularly valuable for customers that require real-time or near real-time decision-making, such as Industrial Internet of Things (IIoT), autonomous vehicles, or localized analytics.

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

Cisco infrastructure portfolio for AI cloud providers

Cisco helps organizations accelerate the buildout of highly scalable, ultra-high-performance and secure AI infrastructure to support their chosen business models. The Cisco AI infrastructure portfolio includes a complete set of cutting-edge AI infrastructure solutions and technologies (Figure 3) for:

- **AI data centers** including full stack and individual components spanning AI-optimized compute, fabric switching, storage, and optics for any scale of data center from mass-scale cloud data center to edge.
- **Data center interconnect and WAN** including coherent pluggable optics and optical systems, scale-across solutions, Routed Optical Networking, and wireless for all inter-data center connectivity requirements from the cloud to the edge.
- **Security and observability** including leading security and observability platforms and technologies integrated at every layer of the AI stack for end-to-end visibility and robust protection.
- **Agentic management platforms** including automation and assurance platforms for simplifying Day-0 resource provisioning, Day-1 configuration, and Day-N operations.

Together, these technologies enable providers to power, connect, protect and deliver generative AI (GenAI) models, including pre-training, fine-tuning, and inference, and support the deployment of their customers' agentic AI workflows. Providers can choose the systems, technologies, and consumption models most appropriate for their individual business requirements, whether they are deploying on-demand services, dedicated AI infrastructure, PaaS, hybrid/edge models, or any other model.

Figure 3: Cisco AI infrastructure portfolio for AI cloud providers.

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

Cisco portfolio key benefits

The following key benefits highlight how Cisco helps AI cloud providers build, secure, and operate high-performance AI services with confidence and efficiency:

Accelerated growth and ROI

Cisco empowers AI cloud providers to achieve faster market entry and maximize their ROI through:

- **Collaborative go-to-market approaches:** Cisco works with AI cloud providers to accelerate time to market and growth, minimize risk, optimize efficiencies, and maximize ROI. This includes collaborative go-to-market approaches that differentiate Cisco-powered AI infrastructure services and the co-creation of customized, ready-to-deploy secure AI infrastructure solutions tailored for enterprise customers.
- **Accelerated delivery with validated designs and reference architectures:** Cisco facilitates the accelerated delivery of AI services at scale by providing pre-configured validated designs that simplify deployment, leading to faster time to revenue and reduced risks. Cisco AI infrastructure designs help providers accelerate deployments by adhering to a choice of Cisco, NVIDIA, sovereign cloud and provider home-grown reference architectures.
- **Streamlining AI adoption with NVIDIA:** Cisco's close collaboration with NVIDIA helps organizations accelerate AI deployments, while maximizing operational efficiency, performance, and security for both enterprise and cloud provider environments. This collaboration includes NVIDIA Spectrum-X Technology integration on Cisco switches, Cisco security built into NVIDIA reference architectures, and Cisco cloud managed AI clusters, adhering the NVIDIA ERA.
- **Operational simplicity:** Cisco helps cloud providers dramatically simplify their operations by unifying network operations and tools across front-end and back-end fabrics and through on-premises or cloud-based management platforms that streamline service orchestration, day-to-day management, automated provisioning, and real-time observability.

Improved customer experiences

Cisco enhances the experience for AI cloud providers and their end-customers by enabling:

- **Flexibility for diverse consumption models:** Cisco offers AI cloud providers a rich AI Infrastructure portfolio of full-stack and component technologies to support the choice and flexibility required for a diverse set of customer requirements and consumption models.
- **Simplified and accelerated onboarding:** Cisco can help providers accelerate and simplify the onboarding, connectivity, and security for enterprise customers. Providers building their infrastructure on Cisco-powered AI infrastructure benefit from optimal reliability, security, and performance, and simpler integration for customers with existing Cisco on-premises data center, security, and interconnect infrastructure.
- **Reduced complexity through interoperability:** Cisco reduces complexity through seamless interoperability and collaboration with other leading technology providers in greenfield and existing environments. Cisco helps streamline integration and provides design and deployment guidance through partner-specific Cisco Validated Designs.

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

Service differentiation

Cisco enables AI cloud providers to differentiate their offerings in a competitive market through:

- **Robust security and observability:** Cisco security and observability, integrated at every layer of the AI stack, allows providers to differentiate with end-to-end visibility, enhanced reliability, and robust protection. For example, Cisco Secure AI Factory with NVIDIA enables safe enterprise AI with security and observability at all layers of the AI development and delivery stack, including AI models, workloads, and infrastructure.
- **Intelligent AI network fabric:** Cisco's integrated, hardware-accelerated intelligent AI fabric enables adaptive routing and congestion management that significantly outperforms traditional industry methods. It is purpose-built to boost the utilization of GPUs serving modern AI workloads by transparently handling diverse and changing network topologies and traffic patterns.
- **Support for secure multi-tenant environments:** Cisco's solutions natively support secure multi-tenant environments, to help ensure safe and efficient resource sharing, with complete isolation for customer resources.

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

Cisco Solutions for

AI data centers

Cisco partners with organizations delivering AI services to help accelerate and scale the deployment of their data center infrastructure with minimum risk and optimal operational simplicity. Cisco's extensive AI data center portfolio (Figure 4) offers organizations an unparalleled choice of technologies and solutions to support a diverse range of business and technical priorities. With Cisco, AI cloud providers have the flexibility to choose the reference architecture and operating environment that best meets their technical requirements. Whether they want to adhere to the NVIDIA Enterprise Reference Architecture (ERA), the NVIDIA Cloud Partner (NCP) reference architecture, Cisco's Cloud Reference Architecture, or prefer to build their own, Cisco provides the necessary flexibility and validated designs to accelerate deployment. Depending on their requirements, organizations can choose a customized, modular design approach, or an integrated stack that spans networking silicon, switching, compute, storage, optics, security, and observability.

Cisco AI data center solutions enable AI cloud providers to maximize GPU scale, performance, and efficiency and minimize job completion times, building on the industry's highest-performing, non-blocking Ethernet network infrastructure. By integrating its purpose-built AI security suite across every layer, Cisco enables organizations to operate securely and reliably, while at peak performance.

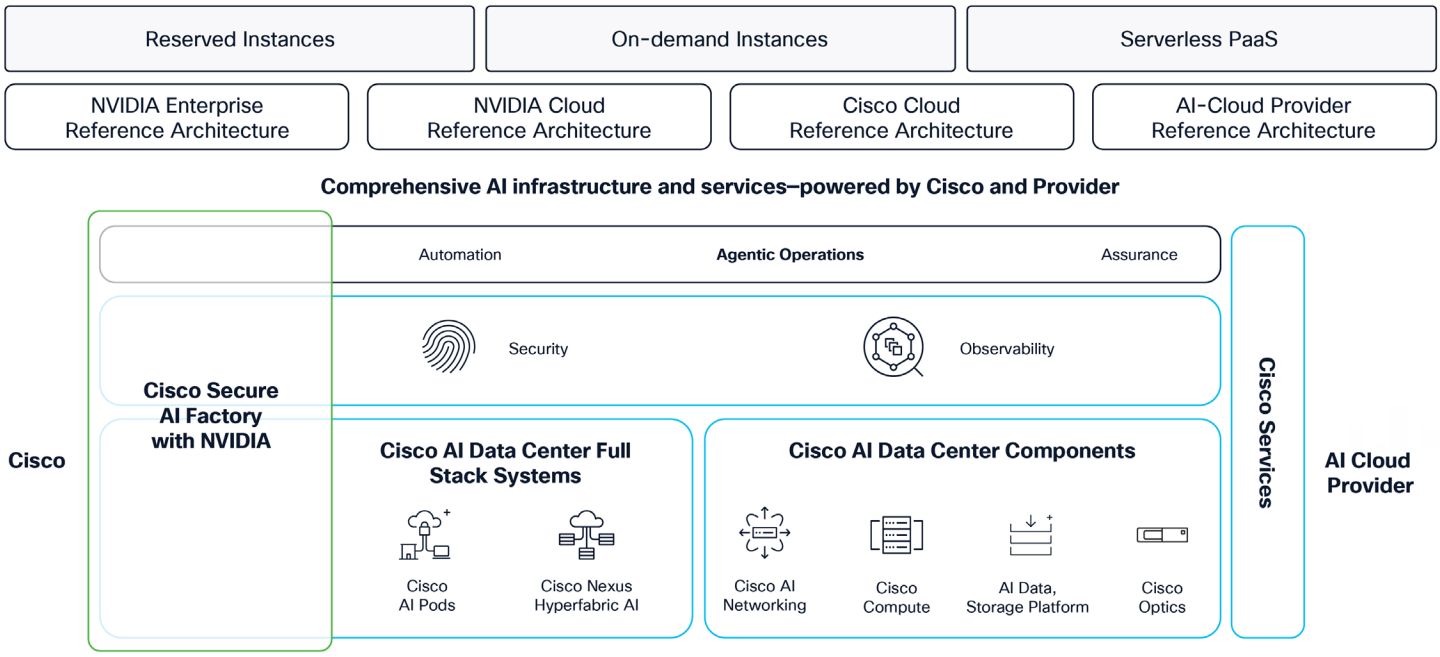

Figure 4: Cisco's AI data center portfolio provides choices aligned to business model and technology strategy.

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

Cisco full-stack AI infrastructure

Cisco is partnering with NVIDIA and other strategic technology leaders to empower organizations to implement, optimize, and secure their AI deployments. The Cisco Secure AI Factory with NVIDIA developed in collaboration with NVIDIA and other strategic ecosystem partners, is a modular reference design that delivers a secure, high-performance AI infrastructure that integrates networking, compute, security, AI software, and storage into a scalable full-stack system. It features built-in security and observability at every layer, superior networking, and seamless integration with the NVIDIA AI Enterprise software platform, specifically designed to streamline the development, deployment, and protection of AI workloads.

The Cisco Secure AI Factory with NVIDIA offers unparalleled deployment flexibility through pre-validated, modular stacks delivered via a Cisco AI-POD modular approach (Figure 5) or integrated cloud-managed turnkey stacks delivered by Cisco Nexus Hyperfabric AI.

Cisco AI-PODs offer a flexible, modular AI infrastructure, which is ideal for providers that require granular control and customization. Providers can choose specific components and integrate with their preferred tools, while also benefiting from Cisco's pre-validated architectures and advanced automation. AI-PODs are modular, secure, and pre-validated full-stack building blocks that integrate Cisco compute (Cisco UCS® servers), high-performance Cisco Nexus networking, and leading GPU/AI platforms into a unified, configurable system. This modular design enables providers to match the offering to the customer requirements and expand incrementally, with scalable units that support a broad range of GPUs, ensuring predictable performance and effortless expansion for diverse AI customers and applications.

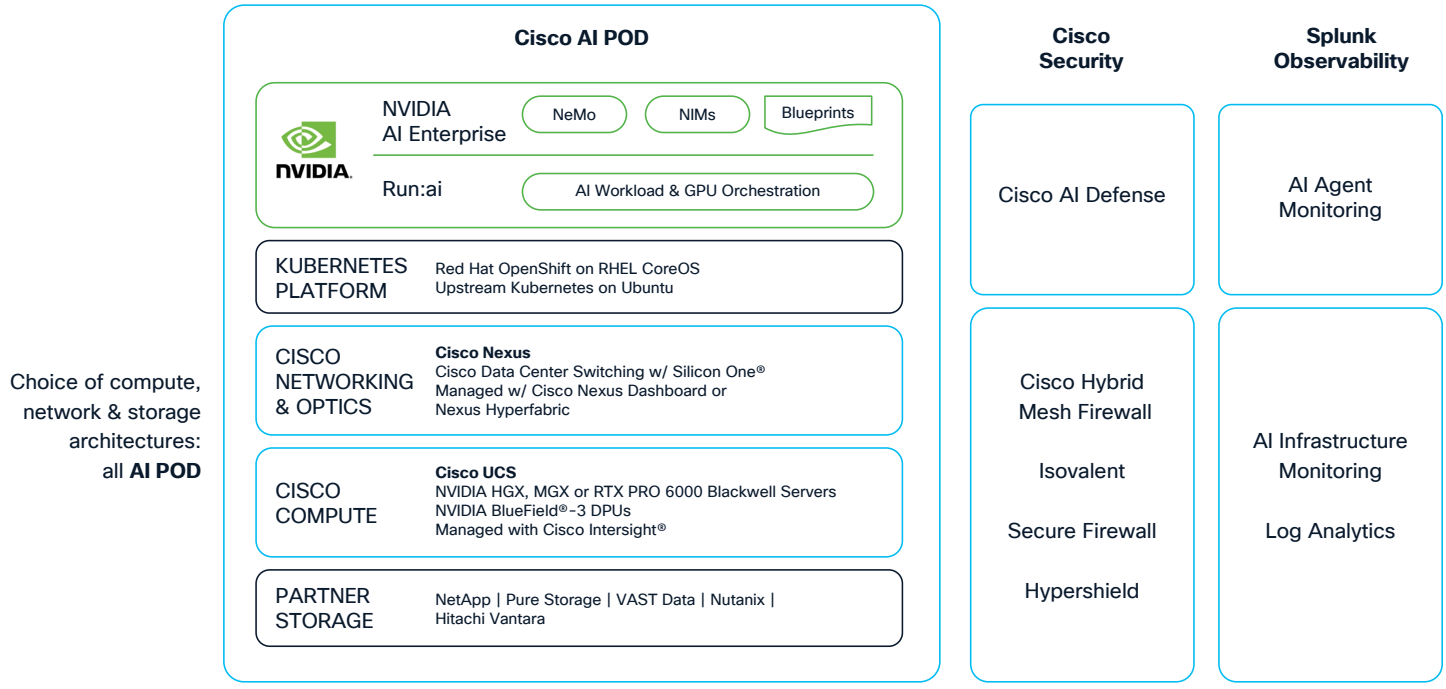

Figure 5: Cisco Secure AI Factory with NVIDIA modular system can be customized for specific needs.

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

AI PODs offer significant benefits:

- **Accelerated deployment:** Deploy AI-ready infrastructure in days, not months, leveraging Cisco Validated Designs (CVDs) and unified management tools that automate provisioning.
- **Superior performance and scalability:** Drive AI workload performance with cutting-edge GPUs and high-bandwidth, lossless Cisco Nexus 9000 Series networking.
- **Enterprise-grade security:** Uniquely embed security at every layer with integrated Cisco AI Defense, Cisco Hypershield, and Isovalent Enterprise Platform, proactively defending against AI-specific threats.
- **Operational simplicity:** Streamline day-to-day management with centralized control, automated workflows, and real-time observability through Cisco Intersight® and Cisco Nexus Dashboard.
- **Flexible and adaptable:** Support any AI workload —from training and fine-tuning to high-throughput inferencing and agentic AI—on a single, adaptable platform, with broad ecosystem compatibility.
- **AI data platform choice:** Cisco has partnered with storage leaders to deliver a choice of AI-optimized, high performance and scalable storage and data platforms. This includes Hitachi Vantara, NetApp, Nutanix NUS, Pure Storage and VAST Data. For example, Cisco, NVIDIA, and VAST Data have partnered to deliver a validated AI data platform designed specifically for large-scale AI. Built on Cisco UCS servers and Cisco AI PODs, integrated with VAST InsightEngine, this solution can form the data foundation of the Cisco Secure AI Factory with NVIDIA.

Simplified operations

To deploy and manage a high-scale AI cluster, a robust software stack with an automation-first approach is required. The use of controllers can dramatically simplify Day-0 resource provisioning, Day-1 configuration, and Day-N operationalization. The key software components involved in the reference architecture include:

- **Cisco Nexus Dashboard** is used to provision and manage the entire networking fabric, including the provisioning required for multi-tenants. It offers a unified platform that simplifies key services, such as network visibility and assurance, fabric automation, and operational simplicity. Operations teams can also manage the switches and overall networking via a broad set of open-source tools, integrating them with available model-driven programmability interfaces.
- **Cisco Intersight** is used to provision the Cisco UCS Servers and supports integration with other automation frameworks via RESTful APIs. Cloud partners or customers can also choose to use NVIDIA Base Command Manager or additional open-source or custom tools or frameworks via the management nodes for compute-node provisioning.
- The **Cisco EBox** storage controller is used for provisioning and managing the attached high-performance storage. Every tenant can be allocated a storage management URL, a user login, and a dashboard for configuration, monitoring, and overall data management. RESTful APIs are supported for integration with automation frameworks.

The Cisco Nexus Hyperfabric AI alternative to the AI-POD modular approach is for organizations seeking a simplified cloud-managed experience. It provides a vertically integrated full-stack solution. Designed for rapid deployment and streamlined operations, it integrates Cisco compute (NVIDIA-Certified Cisco UCS C885A M8 servers with NVIDIA HGX and DPUs), Silicon One-based Cisco 6000 series or Nexus 9000 series switches, Cisco optics, and VAST Data storage. This solution offers plug-and-play cloud management, unified observability across networking and compute, and adheres to NVIDIA Enterprise Reference Architectures, significantly accelerating time to value by reducing infrastructure management complexities.

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

To operationalize GPU as a Service, organizations need to expose their infrastructure through self-service, multitenant consumption models. Together with partners like Rafay, Cisco delivers a fully operational GPU cloud platform, one that enables SKU-based provisioning, GPU slicing, quota enforcement, and an AI workload catalog out of the box. Cisco provides the AI POD or Cisco Nexus Hyperfabric AI as the foundation, while Rafay transforms that foundation into a production-grade GPU cloud, to deliver a simple-to-use, flexible, and modular AI infrastructure and a seamless, self-service consumption experience.

Security-first architecture

Security is crucial for ensuring confidentiality, integrity, and high availability against adversarial attacks by implementing robust access controls and host and network isolation to prevent unauthorized access or manipulation. This is especially critical in a multitenant managed-service environment. Unlike other AI factories in the market, Cisco Secure AI Factory embeds security at every layer of the stack (AI models and associated software components, applications, workloads, infrastructure) to help securely develop and deliver trusted AI applications. For securing the application, **Cisco AI Defense,** integrated with NVIDIA AI, empowers the security and AI practitioner teams with comprehensive tools for robust testing and runtime security of LLMs and generative AI applications. It evaluates generative AI models against diverse security risks (e.g. data privacy, prompt injections) and safety risks (e.g. toxic behavior) without requiring application modifications. For securing the workloads and infrastructure, **Cisco Hybrid Mesh Firewall** delivers unified security management with consistent, pervasive policy enforcement across multiple control points. It includes the following Cisco products:

- **Cisco Isovalent** provides enhanced visibility into cloud-native interactions, enabling consistent policy definition and enforcement across software-defined networks.
- **Cisco Hypershield** protects against adversary lateral movement and proactively mitigates vulnerabilities without requiring patching, all managed through a single interface.
- **Cisco Secure Firewall** offers advanced threat protection at scale without compromising performance, featuring unified management across all firewall deployments.

For security operations, **Splunk Enterprise Security,** a threat detection and incident response platform, enables real-time detection, investigation, and response through powerful analytics, automation, and risk-based insights.

Observability at every layer

Observability is a foundational pillar of Cisco's AI cloud reference architecture, delivering continuous visibility, reliability, and high performance through intelligent tuning and scalable infrastructure. It streamlines debugging, strengthens security, and supports the trustworthiness and effectiveness of AI systems. Cisco Splunk® provides an industry-leading observability platform that enables cloud providers and their tenants to ingest massive volumes of telemetry and gain deep, actionable insights.

Testing and certification

Cisco validated designs ensure that the reference architectures based on Cisco technology are ready for scalable AI cluster deployments. These designs are based on system-level testing that considers all aspects of the management plane, control plane, and data plane, combining compute, storage, and networking. Benchmark test suites are run to evaluate performance and assist with tuning. These include HPC Benchmark, single and multi-hop IB PerfTest, NCCL collective communications tests, high-availability tests and MLCommons Training, and Inference benchmarks. Different elements and entities of the NVIDIA AI Enterprise ecosystem are also tested with use cases around model training, fine-tuning, inferencing, and Retrieval-Augmented Generation (RAG).

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

Cisco AI data center technology components

There are clear advantages to acquiring an integrated full-stack secure AI factory solution that's based on a validated design and offers a single source for services and support. However, some AI cloud providers may decide there are technical or business reasons to acquire the components separately and do the integration themselves. These customers can choose the components they need from our full stack secure AI factory, such as high-density AI servers, network switches and routers, and optics.

Cisco compute

High-density AI-optimized servers

Cisco dense GPU servers incorporate the latest XPU accelerators and are engineered to power the most demanding workloads across the entire AI lifecycle. This includes use cases such as model training, real time natural language processing, digital twins, and video generation. Meanwhile, Cisco's highly modular architectures, combined with high-performance CPUs and a broad range of XPU acceleration options, ensure efficient resource allocation and scalable and predictable performance for diverse AI requirements. **Cisco Intersight** streamlines and simplifies compute operations, providing a flexible, automated environment.

Cisco's comprehensive portfolio of advanced servers optimized for AI includes:

- The **Cisco UCS C880A M8 AI Rack Server,** powered by NVIDIA HGX B300 SXM GPUs and Intel Xeon 6 processors with P-cores, delivers a powerful combination of scalable performance, embedded security, and advanced lifecycle management with Cisco Intersight. It makes a reliable and powerful foundation, delivering the immense computational power needed for large-scale LLM training, model fine-tuning, and large model inferencing.
- The **Cisco UCS C885A M8 AI Rack Server,** a dense-GPU powerhouse based on the NVIDIA HGX platform, supports high-density configurations of NVIDIA H100 and H200 Tensor Core GPUs, and the cutting-edge NVIDIA Blackwell GPUs.
- The **Cisco UCS C845A M8 AI Rack Server,** built on the NVIDIA MGX reference architecture, provides a highly scalable, modular, and customizable platform for flexible configurations across a broad spectrum of AI use cases.
- **Cisco UCS X-Series Modular Systems,** including compute nodes like the X210c and X215c or X440p, offer a versatile and extensible platform for AI, supporting a blend of CPU (Intel Xeon Scalable and AMD EPYC processors) and GPU configurations.

Cisco compute platforms support a wide range of leading GPU technologies, including NVIDIA (H100, H200, RTX PRO 6000 Blackwell Server Edition, L40S, L4, A16) and AMD (MI300X, MI350X, MI210). Specifically, the NVIDIA RTX PRO 6000 Blackwell Server Edition GPU within Cisco UCS servers provides exceptional performance for a universal set of AI computing workloads.

The growing adoption of distributed AI models, with data processing, moving from centralized data centers towards the edge, requires compact, powerful modular edge compute systems. Cisco Unified Edge is a flexible AI-ready system that redefines edge computing, bringing all the strength of the data center to wherever data lives. It is a full-stack system that combines compute, networking, storage, and security, while pre-validated full-stack designs accelerate deployment without compromising performance.

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

Cisco AI networking

Advanced Ethernet switching for front-end and back-end AI networks

The rapid acceleration of AI and high-performance computing is driving unprecedented pressure on data center networks. Training and inference clusters are scaling from hundreds to tens of thousands of GPUs, requiring dense, high-radix fabrics that can move massive datasets and model parameters in real time.

An AI cluster typically has multiple networks—an inter-GPU back-end network, a front-end network, a storage network, and an out-of-band (OOB) management network (Figure 6). Specifically, the AI back-end cluster traffic is predominantly network bound, given the latency and congestion-sensitive GPU-to-GPU parallel communication. To keep GPUs and accelerators fully utilized, networks must now deliver aggregate capacities on the order of 100 Tbps per system. This requires high-radix, power-efficient switching capable of delivering 100 Tbps-class throughput per system, supporting next-generation port speeds (1.6Tbps) with minimal latency and loss. The fabric must support flexible topologies for both scale-up and scale-out multi-pod or scale across multi-site clusters, allowing seamless growth without frequent re-architecture. Integrated visibility and telemetry must operate at line rate to detect congestion, microbursts, or failures before they impact multi-million-dollar AI jobs.

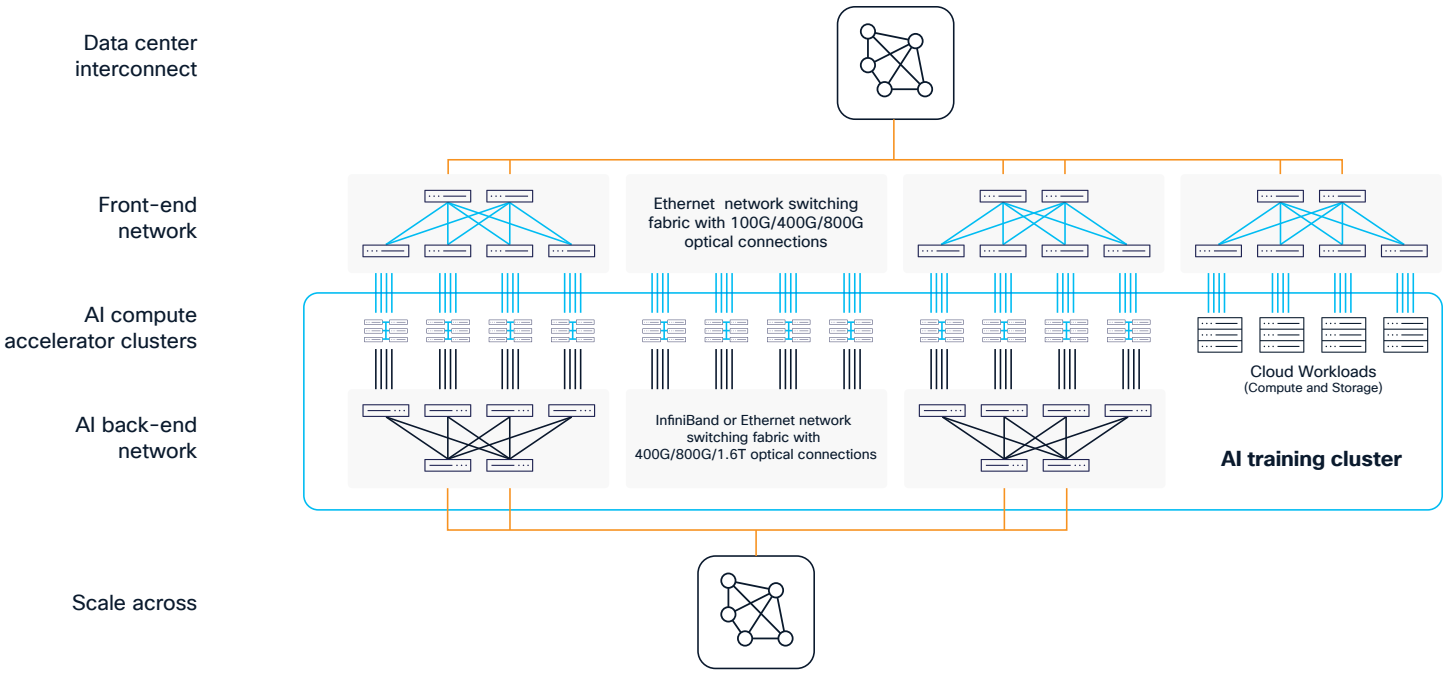

Figure 6: Cisco AI Networking—intelligent, unified AI Ethernet network fabric powered by Cisco Silicon One or NVIDIA Spectrum-X Ethernet switch silicon.

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

By 2027 it's widely expected that Ethernet will surpass Infiniband for back-end networks.

Dell'Oro

While InfiniBand has been the traditional back-end fabric of choice, Ethernet fabrics are proving to be highly effective and increasingly the protocol of choice for high-performance AI back-end networks. In fact, by 2027 it's widely expected that Ethernet will surpass Infiniband for back-end networks.[8] Some of the key reasons for the growing adoption of Ethernet for AI fabrics include:

- Multi-vendor support, operational tools, and consistency across front-end and back-end networks and network interface cards, based on industry standards.
- Comparable performance between enhanced Ethernet and Infiniband switching for demanding AI applications.
- The industry's proven record of rapidly increasing the total bandwidth and port bandwidth of Ethernet switches.
- The large number of ports that can be supported per system (compared to InfiniBand).
- The option to stretch Ethernet-based clusters between data centers using scale-across technology.
- Lower cost and better multi-job performance.

Cisco performance benchmarking demonstrates comparable performance across Ethernet and InfiniBand and shows that advanced Ethernet switches based on Cisco Silicon One and NVIDIA Spectrum-X silicon both tackle congestion and boost performance for demanding AI environments (Figure 7).[9]

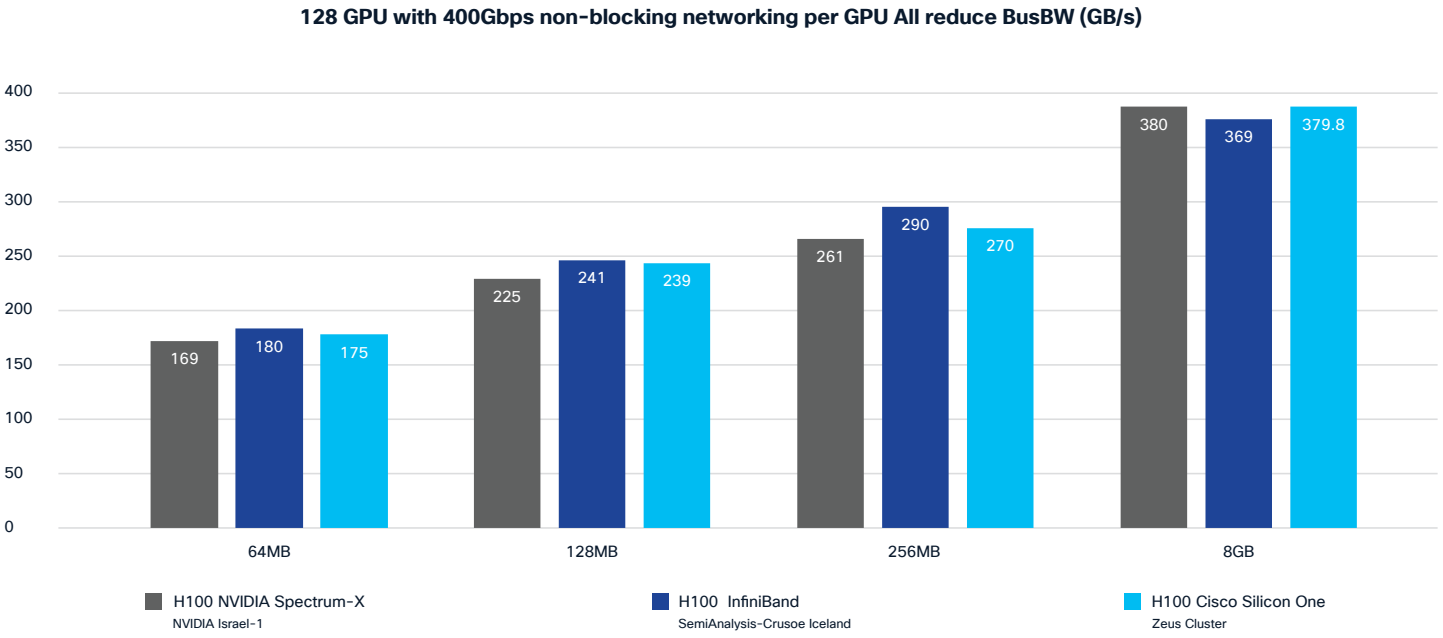

Figure 7: Cisco performance benchmarking—Ethernet and InfiniBand.

Learn more: Cisco AI Networking

8. Advanced Research Report on AI Networks for AI Workloads, Dell'Oro Group, July 2024.

9. Uncompromised Ethernet: Performance and Benchmarking for AI/ML Fabric, Cisco, April 2025.

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

Cisco Nexus 9000 series and Cisco 8000 series switches

Built on Cisco Silicon One G-series programmable ASICs, the Cisco Nexus 9300 series and Cisco 8100 series switches are engineered to meet the demands of mass-scale AI workloads. Support for high-density 400G, 800G and now 1.6Tbps Ethernet fabrics and a massive 512 radix makes them ideal for scalable, next-generation leaf-and-spine network designs. These switches use advanced load balancing, innovative congestion management, and flow control algorithms to improve job completion times. They also provide the low latency and telemetry required to meet the design and operational requirements of AI fabrics.

The Cisco Silicon One G200 supports the industry's highest radix switches, providing 512 x 100GE Ethernet ports on one device, which enables organizations to build a 32,000-GPU AI cluster with 400G interconnects and a two-layer network design. This configuration requires 50% less optics and 40% fewer switches than 256 radix switches with three-layer networks, dramatically reducing the environmental footprint of the AI cluster. Cisco Silicon One G200 also offers advanced, congestion-aware load balancing techniques to prevent traditional congestion events. Its advanced packet-spraying techniques minimize congestion hot spots in the network, while advanced hardware-based link-failure recovery delivers optimal performance across massive web-scale networks, even in the presence of faults.

More recently, Cisco has introduced the **Cisco Nexus 9364-SG3 and Cisco 813X series 1.6T switches** based on the groundbreaking **Cisco Silicon One G300,** a new class of 102.4 Tbps AI networking switching silicon designed to accelerate innovation and time to revenue in rapidly changing AI Infrastructure designs. The G300 delivers 102.4 Tbps throughput engineered for low latency and high-bandwidth back-end networking tailored for AI workloads. It provides intelligent load balancing, adaptive congestion control, traffic engineering, and advanced network telemetry—significantly outperforming traditional methods. The industry's largest shared packet buffer and advanced 224G SerDes in the G300 guarantee reliable GPU data delivery in bursty traffic environments and over longer distances. Advanced linear-pluggable optics (LPO) and Co-packaged Optics (CPO)-ready silicon and package designs support sustainable data center operations through power-efficient design.

The Cisco Nexus 9364-SG3 and Cisco 813X switches are available in both liquid and air-cooled configurations and use Cisco's advanced 1.6TE optics modules, silicon photonics and DSP technology to deliver reliable, high-bandwidth performance.

Cisco Nexus 9000 Series switches simplify the management and operation of AI data center and cloud fabrics with enhanced visibility and sustainability. By offering a full set of services from a single platform, Cisco Nexus Dashboard helps operators configure, operate, and analyze AI data centers and cloud networks using automation and analytics. For example, it provides templates designed to automate advanced configuration for AI Ethernet fabrics which reduce errors and accelerate the time to deploy.

Cisco has also partnered with NVIDIA to deliver the new Cisco Nexus 9100 switch with NVIDIA Spectrum-X switch silicon, which offers high-performance switching for AI-scale deployments. Cisco will offer an NVIDIA Cloud Partner (NCP) compliant reference architecture

The Cisco Nexus 9364-SG3 and Cisco 813X series 1.6T switches are based on the groundbreaking Cisco Silicon One G300, a new class of 102.4 Tbps AI networking switching silicon.

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

with the new Cisco Nexus 9100 Series Switches for AI cloud providers. This Cisco platform is the only third-party switching platform to achieve NCP Reference Architecture compliance. These switches combine the strengths of NVIDIA AI technology with Cisco reliability and streamlined management.

With the Nexus 9100 Series, customers also get a choice of operating systems. They can choose Cisco's proven NX-OS for reliability and operational consistency with automation and operations through Nexus Dashboard. Alternatively, they can select the SONiC open system network operating system, which is best suited for organizations that have the relevant technical resources and prefer open networking standards. The Cisco Nexus 9100 Series augments the broader Cisco Nexus portfolio (which includes switches powered by Cisco Silicon One and Cloud Scale silicon) by incorporating NVIDIA Spectrum-X Ethernet switch silicon. This enables organizations to unify back-end and front-end fabric operations under the Nexus operating model for simplified management with the fully featured Cisco Nexus Dashboard.

Learn more: Cisco Nexus 9000 Series Switches for AI

Cisco 8000 Series switches, powered by Cisco Silicon One, and SONiC open source deliver a wide range of speeds—from 100G to 400G, 800G, and now 1.6T—making them ideal for the most demanding AI clusters. By combining the high performance of Cisco Silicon One, the advanced switching capabilities of the Cisco 8000 Series switches, open-source software, and modular architecture, hyperscalers, neoclouds, and some large enterprises with deep technical resources can confidently build flexible, scalable, and sustainable network infrastructures.

Learn more: Cisco 8000 Series and SONiC

Both the Cisco Nexus 9300 and Cisco 8100 Series include Smart Switch models that integrate Data Processing Units (DPUs) into the switch. These programmable hardware accelerators offload processing and accelerate services such as encryption, firewall, load balancing and DDoS protection at scale, making the switches well suited for the scale requirements of AI fabrics. For example, this includes the hosting of distributed AI security delivered through Cisco Hypershield. Integrated DPUs remove the need for dedicated services appliances, reduce TCO, and deliver services in-line with the fabric for simplified management and better operational efficiency.

Learn more: Cisco Smart Switch

Cisco data center optics

Low power, high performance, scale, and reliability

Each new generation in XPU/GPU and network silicon introduces increased network interface and switch port speeds and the associated need for new optics technologies. As a result, the adoption of 400G and 800G client optics is fully in motion, while 1.6T optics are expected to satisfy the next cycle of AI advances.

There are key factors driving the development of new optical modules. Beyond the ongoing need to meet ever-faster data rates, these requirements include distance, physical size, the number of lanes aggregated

BlueSky Compute is excited to be one of the first neoclouds to deploy Cisco's N9100 Series Switches for the scale-out fabric in our AI training clusters—powering our vision to turn AI into ROI for the world's enterprises by building larger, more interconnected B300 clusters, faster. Cisco's NCP compliant reference architecture, based on N9100 Series Switches while maintaining the operational simplicity with Nexus Dashboard, is a game-changer.

Ian Hartley, CEO, BlueSky Compute

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

to deliver the target data rate, backward compatibility, and power and thermal management.

In addition, the reliability of optics technology is a critical success factor for AI deployments, since any failures can have severe operational and financial impacts. Communication breakdowns between GPUs, or increased latency and re-transmissions resulting from link flaps, result in delayed job completion times and GPU underutilization, which waste costly compute resources. An analysis from Meta found that a single slow GPU link or failed network connection can reduce cluster performance by up to 40%, resulting in drastically reduced efficiency of thousands of costly GPUs, as well as longer job completion times.[10] According to SemiAnalysis, in a large GPU cluster, even if a network link had a mean time to failure rate of 5 years, due to the high number of transceivers, it would only take 26 minutes for the first job failure.[11] This puts immense pressure on high-speed optics, which must operate under the intense conditions found in today's large clusters.

A single slow GPU link or failed network connection can reduce cluster performance by up to 40%.

Meta

Cisco data center optics innovation delivers on all the critical requirements for AI optics—low power, high performance, scale, and reliability. In addition to technology innovation with its silicon photonics platforms, Cisco's rigorous optics testing, robust monitoring, enhanced reliability, and additional performance margins are essential for every AI network and especially large training clusters. Cisco also strengthens supply chain resiliency for its in-house OSFP and QSFP optics by ensuring that there are multiple sources for the silicon photonics, DSPs, lasers, and manufacturing plants.

Cisco QSFP-DD and OSFP optics

There are two types of optical modules that have emerged for ultra high-speed client optics: OSFP and QSFP. Both optical module form factors support the high port density of switches required for AI use cases. Cisco supports both form factors to meet the broadest set of industry and use case requirements.

Cisco's latest OSFP 1.6T client optics pluggable transceiver modules deliver ultra-high bandwidth connections to support AI connectivity and next generation data center spine and leaf architectures up to 500m over single mode fiber as well as high density 2x800G, 4x400G, and 8x200G interconnects. Industry-leading performance and reliability are delivered though Cisco silicon photonics technology and advanced DSP-based diagnostics and monitoring.

The **Cisco OSFP 800G pluggable transceiver modules** are based on the OSFP specification commonly used in AI applications for both front-end and back-end networks, as well as other data center applications. It provides 800 Gigabit Ethernet (GE), 2x 400GE, 4x 200GE, and 8x 100GE connectivity options, complying with the OSFP configuration with integrated heat sink.

The **Cisco family of QSFP-DD modules** optimize thermal management with flexible heatsink designs and maximize port density with high-density interfaces to 100G and 800G modules. At 800G, the QSFP-DD800 transceiver provides the industry's highest bandwidth density of any transceiver module and backward compatibility with lower-speed QSFP pluggable modules and cables. The Cisco optics family also includes high-speed QSFP modules for AI server connectivity.

The adoption of the 102.4T switch cycle and beyond, together with high-density 800G and 1.6T pluggable

10. Special Requirements for Optics in AI Clusters - Highlights from a recent webinar, Lightcounting, August 2024.

11. 100,000 H100 Clusters: Power, Network Topology, Ethernet vs InfiniBand, Reliability, Failures, Checkpointing, SemiAnalysis, June 2024.

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

optical transceiver modules, will continue to pose challenges related to high power consumption. Cisco's Silicon One includes a world-class Serializer/Deserializer (SERDES) component that converts parallel data into serial data and vice versa. Its unique capabilities allow the use of lower power Linear Pluggable Optics (LPO), thereby reducing power consumption for a sustainable AI/ML cluster. Cisco is introducing 800G LPO on the latest Nexus 9000 and Cisco 8000 data center switching systems. Cisco 800G LPO delivers reliable, high-speed, energy-efficient connectivity by shifting power consuming DSP functionality from the pluggable optic to the Cisco Silicon One ASIC. Cisco LPO solutions reduce the power consumption of the optical module by 50% and lower overall power consumption of a fully populated switch by 30%. This approach enables reliable, cost-effective scaling for next-generation data center and AI networking.

Moving forward, Co-packaged Optics (CPO) also present a promising solution for optimizing power and cooling by integrating optical modules directly onto the switch ASIC substrate, reducing electrical reach and effectively addressing signal integrity issues. The Cisco Silicon One G300 is CPO-ready, demonstrating Cisco's commitment to emerging technologies such as CPO, that can further reduce the overall power consumption in large AI deployments.

Learn more: [The critical role of high-quality optics in AI data center networks]

WAN and data center interconnect

Resilient, secure, and high-performance inter-connectivity between data centers and edge environments is more important than ever, as AI training clusters span multiple data centers and GenAI

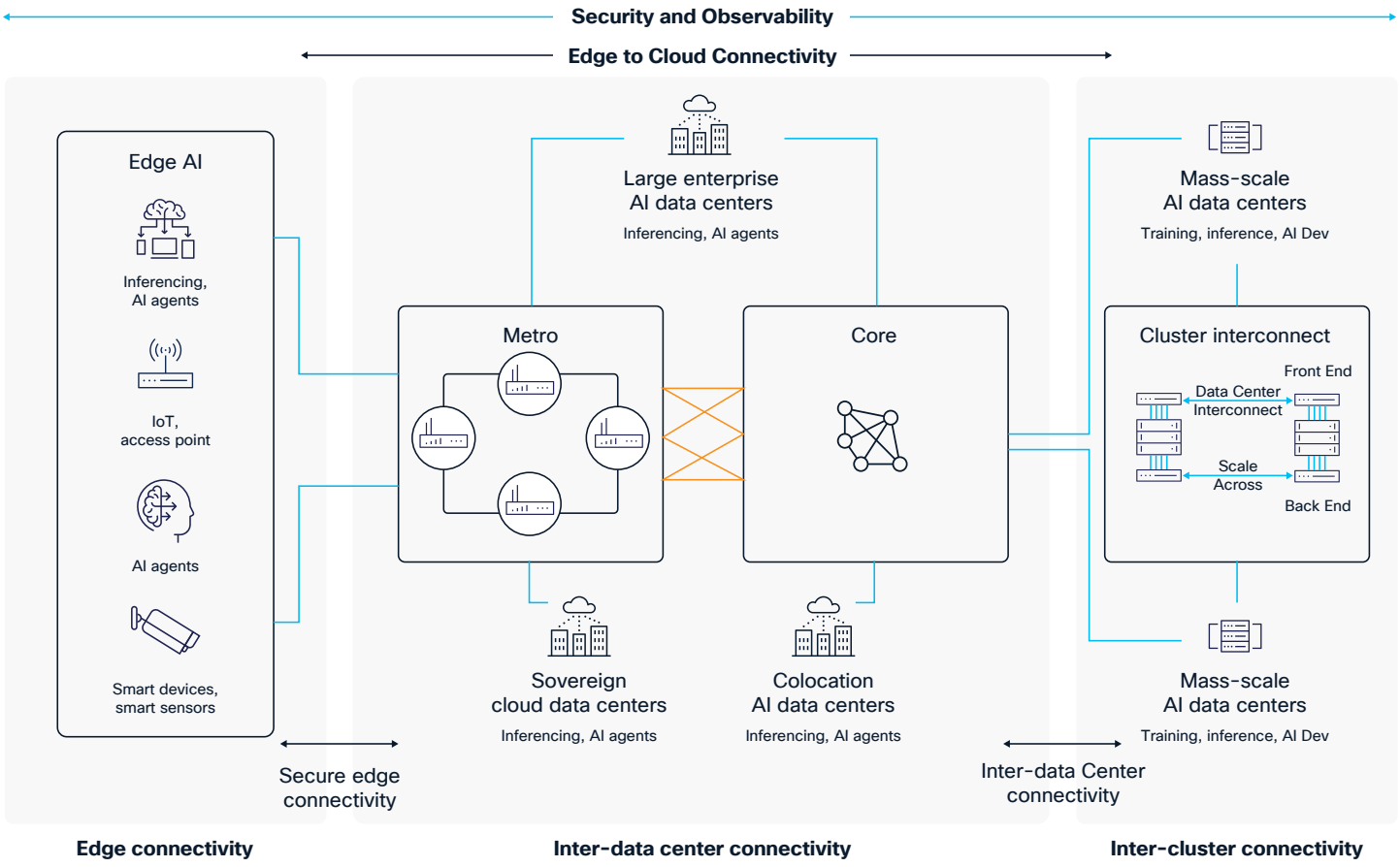

Figure 8: WAN data center interconnect–from edge to cloud.

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

workflows become increasingly distributed, collaborative and fundamental to the proper running of organizations. The growing need for data center interconnects (DCIs) for dispersed AI data centers is ramping up demand for high-speed optical networking and specialized deep buffer "scale across" routers. Cisco Coherent Pluggable Optics and Cisco Optical Networking Systems provide the reliable, power-efficient, high-bandwidth solutions providers can rely on for connecting two or more data center AI fabrics. For large-scale GPU training that requires connectivity between AI cluster back-ends, Cisco offers a new and innovative "scale-across" approach. With Cisco Silicon One-based deep buffer routers, providers can confidently extend customers' large AI training and inference GPU clusters across campus and metro data center locations.

These networks need to transport massive volumes of training data and interconnect users and agents with inference loads. Enterprises and providers are looking to evolve their networks to support AI connectivity, unlock new monetizable services, enhance security, and ensure resiliency—all while optimizing for operational and cost efficiency.

Cisco delivers a full portfolio of WAN and DCI products and solutions to ensure an uninterrupted and secure AI workflow across distributed data center and edge environments (Figure 8). The portfolio includes the following:

- **Cisco coherent pluggable optics and optical systems** to connect data centers over distances ranging from 2 km to 3000+ km over dark fiber.
- **Cisco scale-across solutions** with deep buffer routers for securely and reliably expanding AI clusters between data centers.
- **Cisco Agile Service Networking and Routed Optical Networks** that provide cloud-to-edge connectivity.

Cisco coherent pluggable optics and optical networking systems

Pluggable optics enable high-speed transmission, optimize power efficiency, and simplify network architecture. Cisco coherent pluggable optics contribute to the buildout of AI networks by:

- Enabling distributed AI training architectures across multiple physical sites to alleviate the challenges of power availability.
- Enabling data transfer, backup, replication, and normalization of petabyte training datasets when the source datasets and the AI training cluster are not co-located.
- Providing high-capacity bandwidth between AI data center front ends to address communications among AI agents and business support systems, as well as supporting movement of inference models.

Cisco's coherent pluggable optics family provides solutions for a wide range of data rates and of AI applications over intra-data center, metro, long-haul, or ultra long-haul reaches.

As data center interconnect requirements expand beyond the scale and distances addressed by coherent pluggable optics solutions, Cisco offers the Network Convergence System (NCS) series, a Dense Wavelength-Division Multiplexing (DWDM) line system. The Cisco NCS 1000 series models are optimized for a variety of use cases, distances and throughputs from 2 Tbps to 28 Tbps.

Cisco scale-across

For large-scale GPU training, power constraints and resiliency requirements are causing hyperscalers and neoclouds to embrace distributed AI clusters that span campus and metro regions. This traffic is extremely performance sensitive. Even minor packet loss can

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

severely impact job completion times. This requires deep-buffered routers that absorb bursts and prevent packet loss in addition to high-bandwidth capacity, energy efficiency, and robust security for connectivity between AI cluster back-ends. Cisco offers a new and innovative "scale-across" approach. The Cisco 8223 deep-buffer router is the world's first 51.2 Tbps deep-buffer fixed router, scaling from a single device to 13 Petabyte/sec two-tier clusters, to more than 3000 Petabytes/sec for three-tier clusters. Powered by a single Silicon One P200 processor, it delivers high performance and power efficiency in a compact 3RU footprint. Uncompromising security is delivered with line-rate encryption, integrated root of trust, authenticated software, and hardware analyzers.

Cisco Agile Services Networking and Routed Optical Networks

The Cisco Agile Services Networking architecture enables flexibility in the deployment of large-scale WAN infrastructures. This architecture features a highly efficient routing portfolio based on the Cisco 8000 platform and Cisco IOS XR, along with advanced technologies like segment routing SRv6 for scalability, programmability, and resiliency. When combined with Cisco innovations in Routed Optical Networking and AI-enabled operations across the service and device lifecycle, this architecture helps providers deploying large scale WANs to significantly improve services and lower total cost of ownership (TCO).

Cisco security and observability

As organizations continue to deploy ever more sophisticated, disaggregated agentic AI solutions, there will be an acute need for comprehensive security and insights into the performance of all elements of these business-critical systems.

The distribution of AI workflows and agents across on-premises, hybrid, and multicloud environments brings an increased risk of security threats, making it critical for organizations to ensure data integrity and confidentiality.

Cisco security for AI

The Cisco Security Cloud Control unified, AI-powered management platform provides centralized oversight and multi organization management that offers a single pane of glass for navigating across managed entities, subscriptions, and access controls. It reduces administrative overhead and streamlines customer onboarding and permission structures. Meanwhile, Cisco's Hybrid Mesh Firewall is a distributed security solution optimized to block advanced threats and protect against vulnerabilities, including those in AI models. It also enables zero-trust segmentation across data centers, clouds, and edge sites and uses network and workload identities to create micro-perimeters, applying segmentation and AI-driven threat protection at the app edge and within workloads. This hyper-distributed security enables organizations to define policy centrally and enforce it everywhere, across Cisco firewalls, Smart Switches, workload agents leveraging eBPF, and even third-party firewalls.

The Cisco 8223 deep-buffer router is the world's first 51.2 Tbps deep-buffer fixed router, purpose-built to connect geographically dispersed AI clusters.

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

Cisco Hypershield and the Isovalent platform, both based on the eBPF (extended Berkeley Packet Filter) technology, offer a modern approach to securing cloud-native environments. Cisco Hypershield delivers a unique distributed, AI-native Kubernetes-based security architecture, built specifically for AI workloads and designed to put security wherever it needs to be. The solution is designed to be AI-powered to automate the security policy lifecycle and security infrastructure upgrades.

The **Isovalent** platform provides zero-trust networking and lightweight, highly efficient network observability and security tools, tailor-made for Kubernetes and cloud environments. It embeds security at the kernel level to provide identity-based security, network segmentation, and traffic visibility with minimal overhead to better protect workloads and scale with seamless network policy enforcement.

In addition, Cisco Multicloud Defense offers multidirectional protection across any public or private cloud to block inbound attacks, lateral movement, and data exfiltration. In addition, Cisco has released Cisco AI Defense to help organizations securely develop and use AI by addressing "shadow AI" and ensuring that models and applications are safe and secure across the AI lifecycle. Cisco AI Defense runtime guardrails now integrates with NVIDIA NeMo Guardrails, a part of NVIDIA Enterprise software, bringing together two powerful solutions to maximize cybersecurity for AI deployments.

Another emerging threat source to consider is quantum computing. While quantum computers are not yet able to crack encryption keys, "Q-Day" is likely to arrive in the coming years. In the meantime, there's the risk of harvest now, decrypt later (HDNL). Cisco is working on ways to help organizations prepare for post-quantum cryptography and is developing products to protect against the threat.

Cisco observability for AI

As AI becomes increasingly central to running any organization, the ability to ensure end-to-end service levels of AI workflows becomes critical, including the performance of training or inferencing clusters within a data center, interactions over a WAN between inferencing and LLMs, or edge agents collaborating on an autonomous process.

With Splunk and AppDynamics, Cisco provides Observability for AI for unparalleled visibility into every layer of the environment, including GPU infrastructure, GenAI, LLM outputs, and AI applications. It allows organizations to monitor performance and directly link technical insights to business outcomes and scale AI across hybrid, on-prem, and cloud environments. Cisco Nexus Dashboard provides visibility and assurance that the fabrics within the data center are operating as required. Likewise, Cisco ThousandEyes provides visibility into cloud and Internet connectivity, helping providers better anticipate problems, accelerate their response, and deliver always-on digital experiences. With Cisco Provider Connectivity Assurance, providers can deliver the most comprehensive visibility for end-to-end network performance. Splunk IT Service Intelligence brings it all together with AI-driven incident prediction, detection, and resolution all from one place.

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

Cisco Silicon One

Advanced semiconductors are a foundational technology for delivering AI efficiently and at scale. They enable the ultra-high-speed processing, data movement and secure connectivity required to achieve reliable results in a timely and efficient manner.

For networking, Cisco Silicon One is the industry's most scalable switching and routing silicon architecture. This single architecture consists of five series that can be deployed across a broad range of networks, from back-end and front-end AI networks to scale across, data center interconnect, wide-area networks, and edge (Figure 9).

As AI networks evolve and new standards such as Ultra Ethernet emerge, organizations can continue to enhance their AI networks with the Silicon One programmable architecture.

The Cisco Silicon One Series provides organizations with a variety of unique benefits, including:

- An extensive range of capacity, scale, and interface options to meet the full spectrum of use cases.
- Advanced traffic management, load balancing, and fully shared packet buffers to ensure the lowest job completion times and maximum network efficiency.
- Deep visibility, analytics, and integrated performance monitoring (IPM) for unprecedented real-time insights into the performance of millions of parallel paths, which is a key requirement for both AI-training and AI-inference.

Learn more: Cisco Silicon One Series

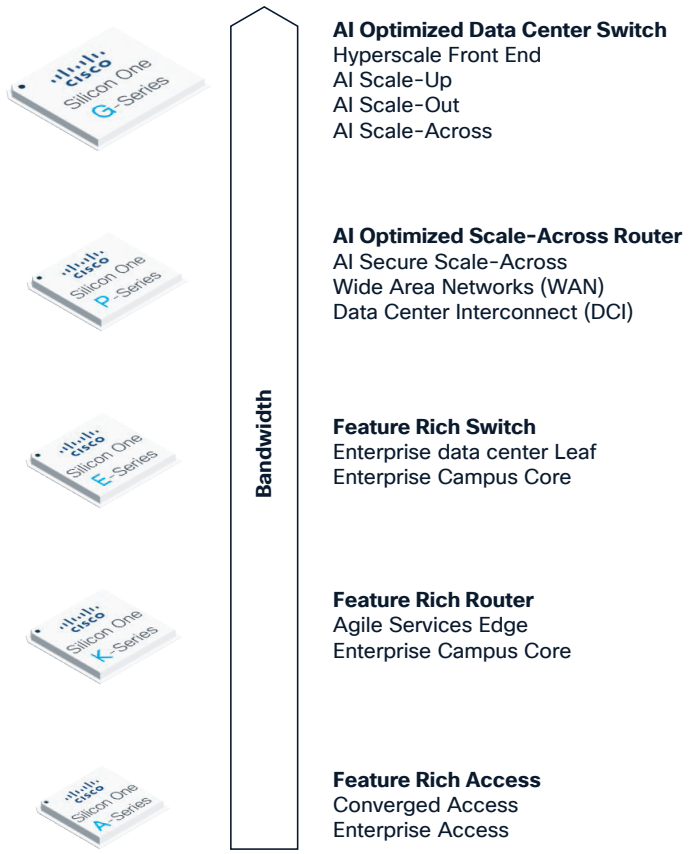

Figure 9: Cisco Silicon One—One Architecture, Five Series.

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

Cisco AI ecosystem

The rapid pace of AI innovation and the complexity of AI architectures make strong collaboration with technology partners essential. Cisco is continuously expanding its established partnerships and forging new ones to ensure customers have access to tried and tested technology stacks that seamlessly support their AI journeys. Cisco is building integrated solutions through strategic alliances with industry leaders including compute and accelerator partners, software partners, storage partners, liquid cooling partners, system integrators; and many others (Figure 10). This collaborative approach includes the creation of Cisco Validated Designs (CVDs), which provide pre-configured solution blueprints that simplify the integration of AI workloads into existing IT infrastructure, leading to faster deployment times and reduced risks associated with complex infrastructure configurations and deployments.

In particular, our partnership with NVIDIA continues to accelerate, empowering AI at scale with enhanced security, observability, and trustworthiness. Recent announcements highlight our deep collaboration, including the new Cisco Nexus 9100 switch, which integrates NVIDIA Spectrum-X switch silicon to deliver high-performance switching for AI-scale deployments.

Beyond our core technology partnerships, Cisco is also collaborating with leading liquid-cooling technology vendors, such as Vertiv, DeepCoolAI, Green Revolution Cooling, and Asperitas, to deliver state-of-the-art liquid cooling solutions as part of our fully tested AI infrastructure offerings.

Components			Technology Partners
Observability	Cisco		
Security	Cisco		
System Level Mgt, Orchestration*	Cisco	+	NVIDIA, Rafay, Red Hat
Compute, XPU	Cisco	+	NVIDIA, AMD, INTEL
Networking & Optics	Cisco		
Storage	Cisco	+	Hitachi Vantara, NetApp, Nutanix NUS, Pure Storage and VAST Data.
Power Distribution			Vertiv, DeepCoolAI
Cooling (RDHX, DLC, Immersion)			Vertiv, DeepCoolAI, Asperitas, GRC

Figure 10: Cisco ecosystem of AI technology leaders.

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

Why Cisco AI infrastructure

By collaborating with Cisco, AI cloud providers gain a differentiated advantage in the rapidly expanding AI market. Cisco's comprehensive AI infrastructure portfolio enables providers to accelerate AI service delivery at scale, while reducing time to revenue and maximizing return on investment. Our solutions support a wide range of business models, from on-demand instances to dedicated AI clouds, providing the flexibility needed to meet evolving customer demands.

The Cisco advantage is built on a complete, end to end network fabric stack spanning silicon, systems, optics, and operating systems, supported by a resilient global supply chain. This foundation is augmented by industry leading security and observability capabilities that reduce deployment risk and enable secure, multi tenant AI environments.

Beyond technology, Cisco enables provider success through collaborative go to market engagement and ready to deploy AI infrastructure bundles that speed adoption and drive sustainable growth. Cisco Customer Experience (CX) services and our global partner ecosystem deliver specialized AI practices that accelerate deployment and operations, leveraging decades of experience building the world's most scalable infrastructures. AI certifications and specializations further equip delivery partners with the expertise needed to support success in the AI era.

Conclusion

The rise of agentic AI represents a defining shift for the technology industry, enabling autonomous decision making, complex workflow orchestration, and step change improvements in productivity and innovation. Unlocking this potential requires purpose built AI infrastructure capable of supporting increasingly demanding and distributed workloads.

Cisco delivers the **critical infrastructure for the AI era.** We recognize that the AI economy spans diverse use cases, from hyperscale model training and sovereign clouds to distributed edge inference—and that no single architecture fits all. Cisco collaborates closely with AI cloud providers from all segments to deliver tailored solutions that address these varied requirements.

With a comprehensive portfolio, Cisco enables providers to deploy and scale AI infrastructure with confidence. By choosing Cisco, AI cloud providers can build efficient, high performance AI infrastructure faster—and position themselves for sustained leadership in the rapidly evolving AI economy.

Learn more about Cisco AI Infrastructure solutions.

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

© 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) P1104011821