AI Infrastructure for the Agentic Era

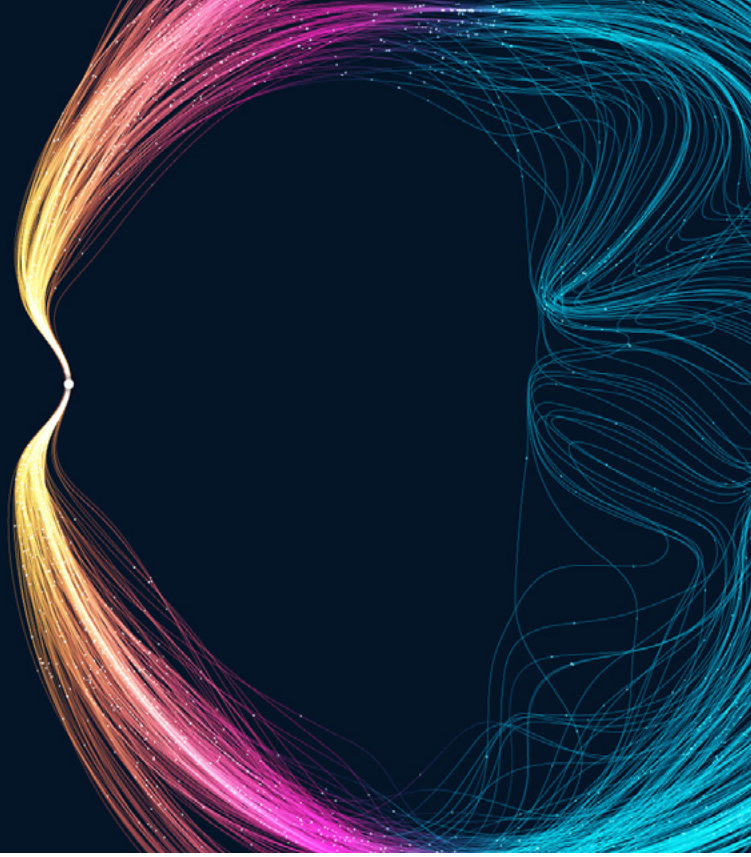

Table of Contents

The emergence of agentic AI

We are entering the "Internet of Agents" era. Today, our workforces are entirely human. Tomorrow, they will expand exponentially to include a variety of non-human AI "workers"—including apps, agents, robots, and even humanoids. Agentic AI introduces a world where connected AI agents and people will work together to orchestrate all manner of complex workflows. This will result in massive gains in productivity and capacity, with considerable shared benefits for organizations, workers, and society alike.

The evolution to agentic AI has been rapid and resolute. When OpenAI released an early demo of ChatGPT-3 on November 30, 2022, the generative AI (GenAI) chatbot quickly went viral. Within five days, ChatGPT had introduced one million users to its powerful ability to create original content on demand. This marked the beginning of the Cambrian-like explosion of GenAI applications.

Multimodal GenAI is a natural extension to GenAI. In the real world, people encounter and comprehend information through a combination of different modalities, such as text, audio, visual, and sensing. Multimodal GenAI replicates this process by combining and analyzing different types of data inputs to generate more robust results. A shift is now underway with the development of natively multimodal GenAI models such as OpenAI's GPT-4o. Gartner estimates that by 2027, 40% of GenAI solutions will be multimodal, up from 1% in 2023.[1]

As shown in Figure 1, agentic AI represents the next frontier in this evolution, providing autonomous decision-making, planning, and adaptive execution to complete multi-step processes. By 2028, at least 15% of day-to-day business decisions will be made autonomously through agentic AI machine workflows for rule-based tasks, up from 0% in 2024, according to Gartner.[2]

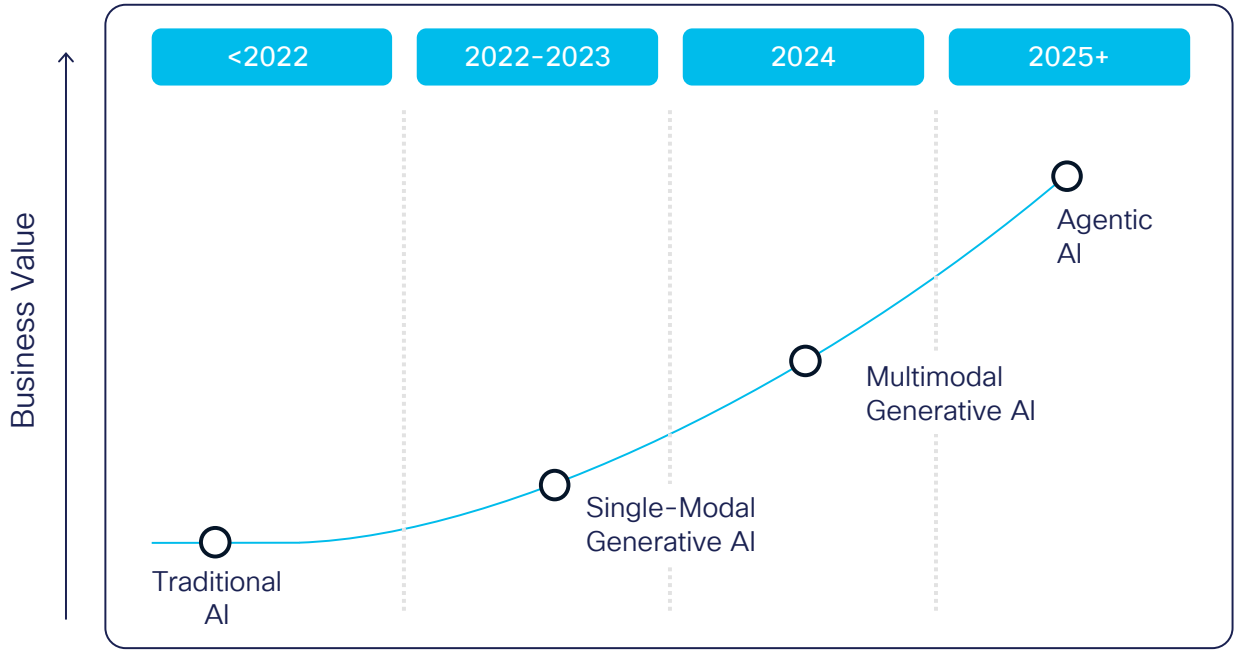

Figure 1. The evolution to agentic AI is driving greater business value

These AI agents, powered by large language models (LLMs) and equipped with increasingly advanced reasoning capabilities, will be able to discover and learn individually. Then, over time, AI agents will collaborate with each other by forming chains of operations, which will enable them to automate business functions. This new "Internet of Agents" era is defined by Cisco as an "open, interoperable internet for agent-agent and agent-human quantum-safe communication" that enables secure agent collaboration while preserving organizational autonomy.[3]

The agentic AI lifecycle

The agentic AI lifecycle builds on current AI LLM and inferencing models (Figure 2). The foundational element for agentic AI architecture is the pre-training of large datasets that create general purpose foundational and frontier LLMs. These LLMs are fine-tuned based on additional domain-specific data to create domain-specific or job-specific LLMs.

Enterprises use retrieval-augmented generation (RAG) and other inferencing techniques that augment contextual information from proprietary databases to generate more accurate and relevant responses that extend beyond the scope of the original model.

Agentic AI workflows provide the ability to make autonomous business decisions and can greatly benefit from collaboration with connected agents working across multiple devices, types, and locations. The powerful ability to combine the intelligence of agents with specialized knowledge to make decisions and act in real time has the potential to drive substantial business impact across every aspect of the organization.

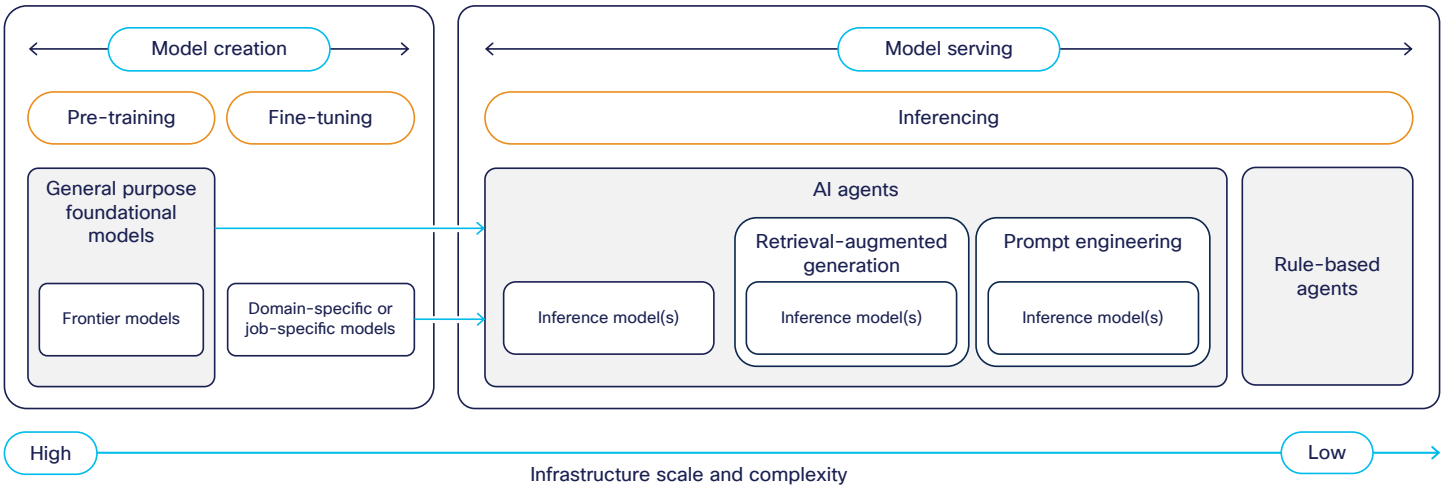

Figure 2. Agentic AI lifecycle builds on current models

Infrastructure innovations driving AI forward

New and rapid innovations across the complete AI infrastructure stack (Figure 3) are making the evolution to agentic AI a reality. Innovations in accelerated compute are leading to the densification of racks, which provides the level of performance required to support all aspects of the agentic AI lifecycle. The increasing speed and scale of accelerators like graphics processing units (GPUs) is driving dramatic performance and scale improvements in scale-up and scale-out networks, spurring demand for more advanced network switching silicon and higher data-rate optics.

Denser and faster infrastructure is also driving the need for innovations in direct-to-chip liquid cooling to significantly reduce the energy required for cooling. At the same time, the evolution to agentic AI is resulting in increasingly complex workflows between distributed agents and LLMs, requiring fast, reliable, secure wide area network (WAN) connectivity from the cloud to the edge and everywhere in between. Importantly, these distributed workflows also require innovative new approaches to making the infrastructure, workloads, and data more secure and assured.

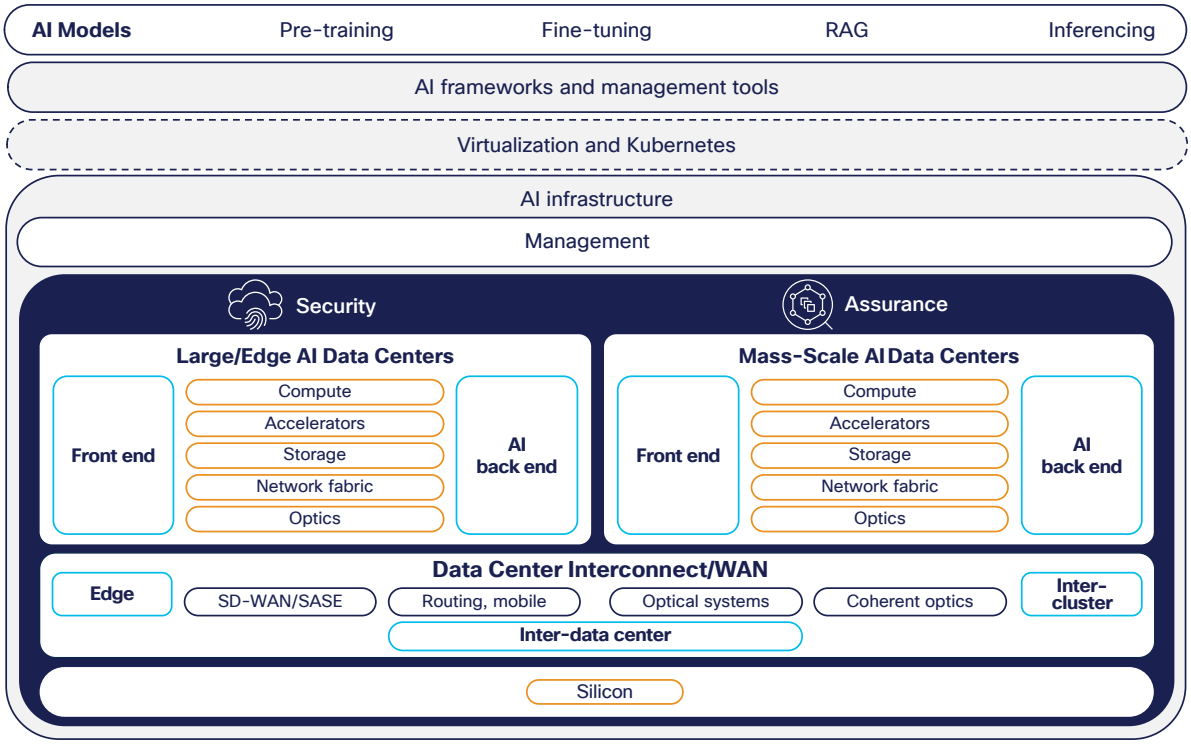

Figure 3. AI evolution requires infrastructure innovation across the full stack

Explosion of demand for AI infrastructure

Agentic AI is proving to be a major driver for data center infrastructure build-outs and expansion. According to McKinsey & Company, global data center capacity demands could more than triple by 2030, with 70% of that growth being driven by AI workloads.[4] This demand is driving the emergence of new mass-scale data center hubs, where the availability of sustainable power and cooling are the main selection criteria. In addition to these large investments being made in data centers to run LLMs, LLMs are creating demand for the private and public cloud AI infrastructure required to run inferencing models.

Gartner expects that the use of GenAI models will influence over 90% of organizations to pursue hybrid cloud environments through 2027, in turn creating demand for data center interconnect (DCI) networks that transport data between distributed, hybrid and cloud data center architectures.[5] Meanwhile, the approaching tsunami of distributed AI agents at the edge is causing infrastructure investments across the cloud-to-edge continuum to ratchet up. By 2026, at least 50% of edge computing deployments will involve machine learning, compared to just 5% in 2022.[6]

The infrastructure requirements for providing AI services vary significantly by provider type. **Hyperscaler cloud providers** such as Microsoft, Google, Meta, and Amazon are making significant investments to scale data center AI clusters and even extend clusters across metro areas. More recently, so-called **"neocloud" GPU-as-a-service providers** and **foundational model providers** are emerging to rival the AI cloud infrastructure scale of hyperscalers to address the growing LLM and inferencing demands.[7] At the same time, as **enterprise and public sector** organizations ramp up their AI initiatives, they are choosing to meet their inference modelling needs through private, public, or hybrid cloud approaches.

Meanwhile, the data center infrastructure and internetworking services of **colocation providers** and **communications service providers (CSP)** are in high demand due to the enormous space, power, and cooling demands of AI workloads, as well as WAN connectivity across data centers and regions. This demand is being propelled by the growing global deficit in modern AI-ready data center facilities to host enterprise inference models, and also by data sovereignty requirements that are creating the need for regional "sovereign clouds" to address specific industry or governmental regulatory and compliance needs. At the same time, the emergence of distributed agentic architectures with demanding latency, security, and regulatory requirements is driving new intelligent internet and private backbone architectures.

From training scale to test-time scale

The AI industry continues to explore the limits of scale and opportunities to improve the accuracy and usefulness of foundational models. Pre-training and post-training (i.e., fine-tuning) were the initial phases where scale was being exploited. Now there is a transformational shift toward test-time scale. This involves scaling compute and network resources during the inference phase (or test time) based on business criteria.

A good example of the benefits of test-time scale is OpenAI's o3 breakthrough model, which achieved unprecedented results in adaptability and generalization in the ARC-AGI-1 benchmark.[8] The implications of this shift to multi-step test-time computing are significant and will multiply the requirements for infrastructure to create substantially more tokens while returning results within a desired time. According to Gartner, more than 80% of workload accelerators deployed in data centers will be used to execute AI inference workloads by 2028.[9] Additionally, agentic AI means that multiple AI agents will work together to solve harder and harder problems.

AI lifecycle infrastructure considerations

The comprehensive agentic AI lifecycle view is supported by four key architectures:

- Mass-scale cloud AI data centers
- Large-scale AI data centers
- Edge AI data centers
- Interconnecting wide area networks (WANs)

Each of these architectures has an important part to play in delivering on generative and agentic AI, and each has unique requirements. The complete lifecycle depends on each of the participants in the AI ecosystem contributing their part. This ecosystem includes hyperscalers, neocloud and model providers, enterprise and public sector organizations, CSPs, and colocation providers. Figure 4 shows the primary AI infrastructure requirements for each.

Ecosystem Participant	Enterprise and Public Sector	Communications Service Providers	Colocation Service Providers	AI Model Builders	Neocloud Providers	Hyperscalers and Tier 2 Cloud Providers
Primary AI Focus	• Use generative and agentic AI solutions building on large language models to improve business processes	• Offer connectivity services to power agentic AI • Offer colocation services and sovereign clouds	• Offer AI data center infrastructure services and interconnect services	• Build general purpose frontier and foundational models	• Offer cloud-delivered AI infrastructure services such as GPU-as-a-service	
Agentic AI Phase	Fine-Tuning and Inferencing			Pre-training, Fine-Tuning, and Inferencing		
Infrastructure Focus	**Large-Scale and Edge AI Data Centers**			**Mass-Scale AI Data Centers**		
	Data Center Interconnect/WAN					
Primary Infrastructure Requirements	• Small to large clusters: 8 GPUs to 5k+ GPUs • Simplified deployment and operations • Integrated infrastructure stack* • End-to-end security • AI workflow service levels	• Cloud-to-edge secure AI connectivity • Automation and assurance	• Power efficiency and cooling • Multi-tenancy security • Cloud interconnect	• Mass-scale clusters: 5K-100K+ GPUs • Infrastructure performance efficiency • Multi-tenancy security and reliability • Power efficiency		

* Integrated stack combines compute/accelerator, networking, optics, and storage for a complete AI cluster

Figure 4. AI Infrastructure requirements depend on the specific needs of the ecosystem participant

Mass-scale AI data centers

Mass-scale AI data center infrastructures for pre-training and fine-tuning LLMs differ in scale and architecture from traditional data centers. This is especially the case in the deployment of back-end AI training clusters that scale and optimize performance by distributing training processes across many thousands of GPUs in parallel (Figure 5).

Today, pre-training and fine-tuning for most LLM models happens in hyperscaler or neocloud environments due to the massive scale and investment required. Hyperscalers, Tier 2 cloud providers, and some large neocloud providers buy components and use open-source solutions to build high-density architectures. These architectures are fully optimized and automated and feature custom management solutions which address scalability and performance and reduce the total cost of ownership (TCO).

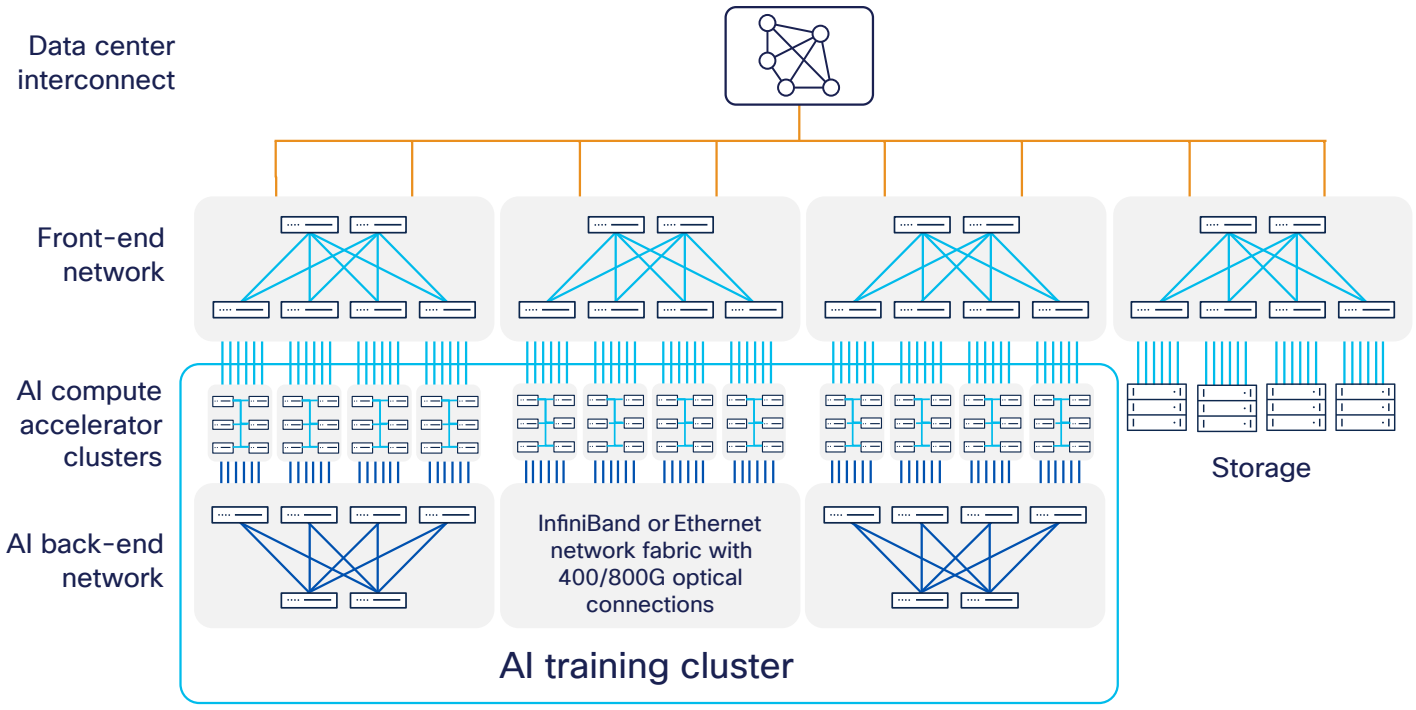

Figure 5. The AI-ready data center architecture and technology stack differ significantly from traditional data centers

Large-scale AI data centers

Today, organizations that are focusing their AI build initiatives on creating their own inferencing capabilities typically need to invest less in infrastructure than those building their own LLMs. However, as inferencing techniques become more powerful and increasingly use iterative reasoning techniques, such as test-time scale, the infrastructure investments required to support them will increase too. Likewise, enterprises are increasingly aware of the need to protect their AI applications, data, and infrastructure from attack.

Enterprises that are building out their own models can do so in their own data centers, in colocation facilities, or in hyperscaler and neocloud facilities. Large enterprises, CSPs, and colocation providers often choose to adopt vertically integrated technology stacks and build high-density architectures to align with scalability and time-to-market requirements.

Edge AI data centers

As inferencing models become more advanced, multiple forces are pushing toward more distributed and regionalized inference closer to the edge; among them are power and space density, user scale, data regulation and sovereignty, and over time, network performance and latency. Edge AI deployments distribute smaller-scale compute and networking resources closer to the data, end users, and devices. This reduces the strain on centralized cloud and data center infrastructures, addresses data privacy concerns, and allows for more real-time data processing and decision-making. At the same time, the distributed nature of data and GPU resources at the edge requires additional attention to security.

WAN interconnect for AI

While much of the AI infrastructure attention has been on data center infrastructure requirements, as AI workflows become more distributed there is a growing need for different classes of connectivity between cloud, enterprise, and edge environments. The emergence of distributed AI agents and iterative run-time inferencing models will further expand the requirements for interconnectivity across the cloud-edge continuum.

For enterprises, this means choosing affordable, scalable networking services or self-owned secure networking solutions that address the run-time nature of distributed AI-enabled processes and the need for data integrity, authenticity, confidentiality, and compliance. Organizations need to choose solutions that can secure their AI WAN traffic everywhere through cloud-based security and segmentation and assure AI performance with end-to-end visibility and the optimization of AI flows across owned and unowned networks.

To help meet these interconnectivity requirements, CSPs will need a converged transport network spanning access and metro to unify what were traditionally layered silos. In addition, CSPs will need to implement architectural changes that allow for the extension of services more deeply into the metro to reduce latency, increase protection, and ensure regulatory compliance of specific flows. End-to-end security that goes beyond traditional protocols, like IPsec and MACsec, will also be essential for addressing future threats, such as those posed by quantum computing.

There are three primary interconnectivity use cases, as shown in Figure 6:

- **Inter-cluster connectivity**
  Running high-density and power-hungry AI workloads often necessitates the use of multiple data centers within a metro area. The goal is to create larger GPU resource pools by extending AI clusters using multiple highly secure, ultra-high-bandwidth optical connections.
- **Inter-data center connectivity**
  Reliable high-speed connectivity is required between all data centers that contribute to the AI lifecycle, whether for the data lifecycle, the agentic AI lifecycle, or both. Large volumes of data are collected, consolidated, cleaned, normalized, and duplicated across data centers. In addition, foundation models need to be distributed to further fine-tune the model to create domain-specific models and address inference requirements.

  As inference models become more advanced and iterative, the performance requirements of inter-data center connectivity become increasingly important.
- **Edge AI connectivity**
  Edge AI deployments require secure, real-time connectivity to ensure that any remote inferencing models and agents running on an edge data center or an IoT or end-user device have consistent, secure connectivity to central data centers. And as agentic AI increasingly requires agents to collaborate, optimized end-to-end network service levels and security are also critical requirements for edge AI connectivity.

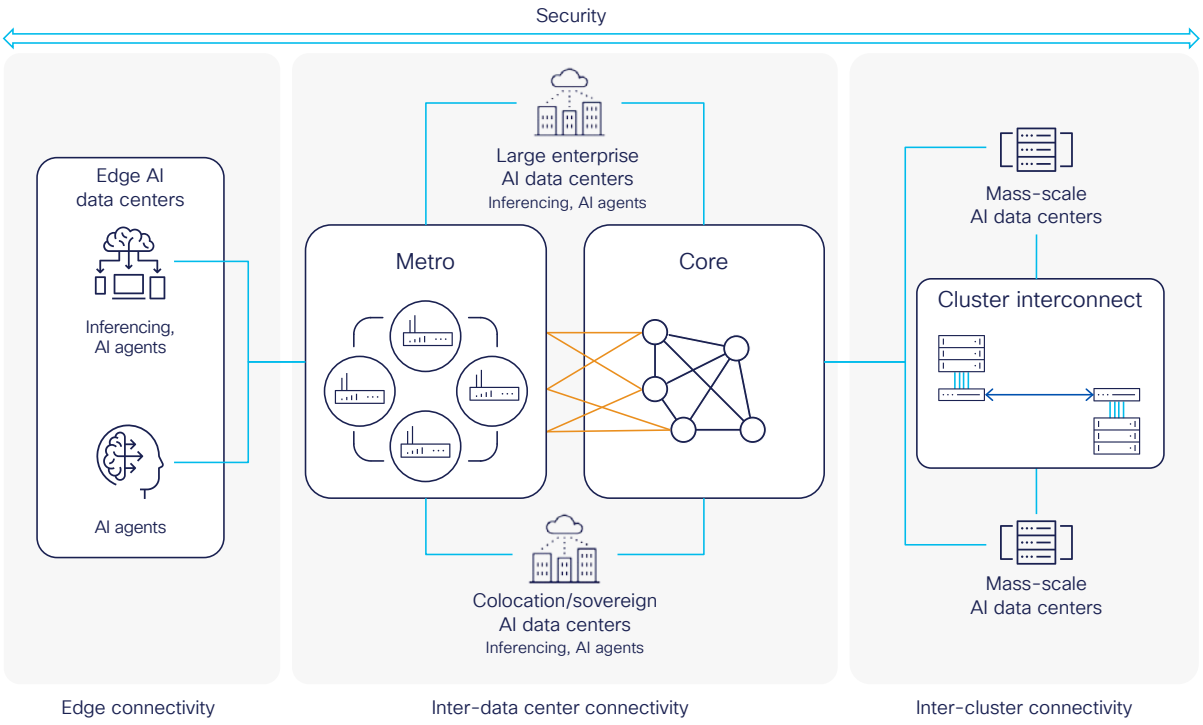

Figure 6. AI network connectivity across the cloud-to-edge continuum

Cisco is meeting the diverse infrastructure needs of the AI lifecycle

As organizations move forward on their journey through generative, multimodal, and now agentic AI, they need well-suited solutions that minimize risk and cost, maximize speed and efficiency, improve security, and address regulatory compliance.

The time is now. According to the 2024 McKinsey Global survey, 75% of respondents believe AI will usher in a disruptive change in their industry, and those organizations that lag in AI adoption will risk becoming irrelevant.[10] And yet, according to the Cisco 2024 AI Readiness Index, only 13% of companies are ready to leverage AI-powered technologies to their full potential today, while 85% of respondents say they have less than 18 months to deploy an AI strategy before they will see negative business effects.[11]

Cisco is revolutionizing how infrastructure and data connect and protect organizations in the AI era. Our AI infrastructure innovations and solutions are helping organizations power and secure AI across the continuum of highly varied needs from hyperscalers, neoclouds, and enterprises to service providers and colocation providers.

Our solutions build on technologies and products that range from silicon to complete AI systems and span networking, compute, optics, data center interconnect systems, security, and observability. Organizations can choose to adopt these technologies as validated full-stack solutions with the option for simplified, integrated management or as components that can be used to build an optimized full stack. In either case, organizations deploying their own AI infrastructure can choose to use proven Cisco Validated Designs (CVDs) and industry-leading AI reference architectures to accelerate and streamline successful deployment.

The fast pace of AI innovation and architectures requires strong collaboration between technology partners to keep the evolving needs of organizations at the forefront. Cisco is expanding on established technology partnerships and building new ones to ensure customers have tried-and-tested technology stacks and timely access to the most advanced AI technologies, including accelerated compute, networking, storage, software platforms, and liquid cooling, among others. Cisco has established a deep partnership with NVIDIA to help organizations accelerate and secure their AI initiatives by delivering a secure AI factory.

Cisco continues to innovate and support the full spectrum of organizations and use cases that make up the AI lifecycle ecosystem. Whether it's training in the cloud, inferencing in the enterprise data center, or agents at the edge, Cisco solutions span everything inside and between the largest to smallest cloud, data center, and edge environments.

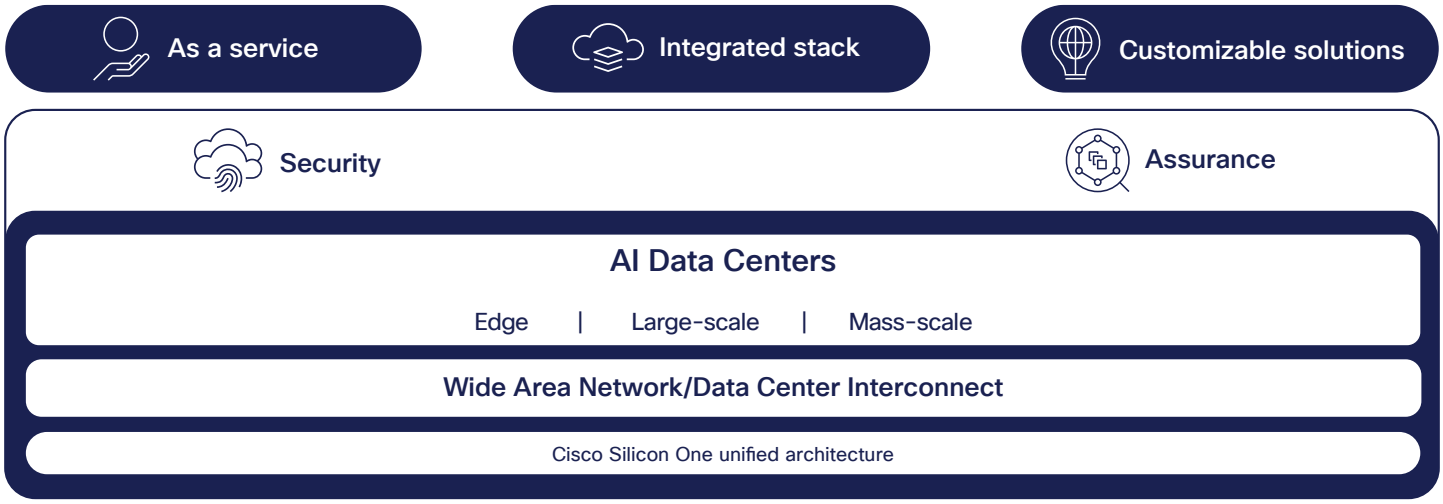

Figure 7. Cisco addresses AI infrastructure needs both within and between data centers at all scales and with a choice of deployment models

The Cisco AI infrastructure portfolio (Figure 7) provides organizations a choice of solutions and deployment models from cloud to core to edge that best meet specific business and technical workload requirements. It also delivers the resilience needed to operate securely at scale and at peak performance within data centers, between them, and across the entire connected landscape. It offers choices and innovations that help accelerate time to deployment, value, and full realization of the business opportunities unlocked by AI.

Cisco Silicon One

Cisco Silicon One provides a single cutting-edge semiconductor architecture that can be deployed across a broad range of networks, from back-end and front-end networks to data center interconnect and WANs. As AI workloads and networks evolve, organizations can continue to evolve their AI infrastructure with the Silicon One programmable architecture.

Specifically, the Cisco Silicon One G200 powers ultra-high-performance Ethernet-based AI back-end networks that require advanced congestion avoidance and fault recovery techniques to ensure maximum GPU efficiency and low job completion times.

G200-based switches support 800G port speeds, and at 512, the industry's highest radix. In comparison to a 256-radix switch, this reduces a large AI cluster from a three-layer network to a two-layer network, requiring 50% less optics and 40% fewer switches, which drastically reduces the environmental footprint and latency of the AI cluster.[12]

Cisco is partnering with NVIDIA to enable Cisco Silicon One-based switches to be coupled with NVIDIA SuperNICs to become part of the NVIDIA Spectrum™-X Ethernet networking platform. Cisco is the only partner silicon included in NVIDIA Spectrum-X.

AI-ready data centers

Cisco helps organizations efficiently deploy and operate the best-performing Ethernet-connected data centers at edge, regional, and centralized locations to support both AI and traditional workloads. We provide the building blocks of silicon, networking, compute, optics, and security, as well as fully integrated modular and scalable stacks (Figure 8) that are purpose-built for AI workloads.

This allows organizations to either adopt an integrated AI full stack from Cisco or build their own AI stacks based on Cisco and partner technologies, depending on their business and technology priorities.

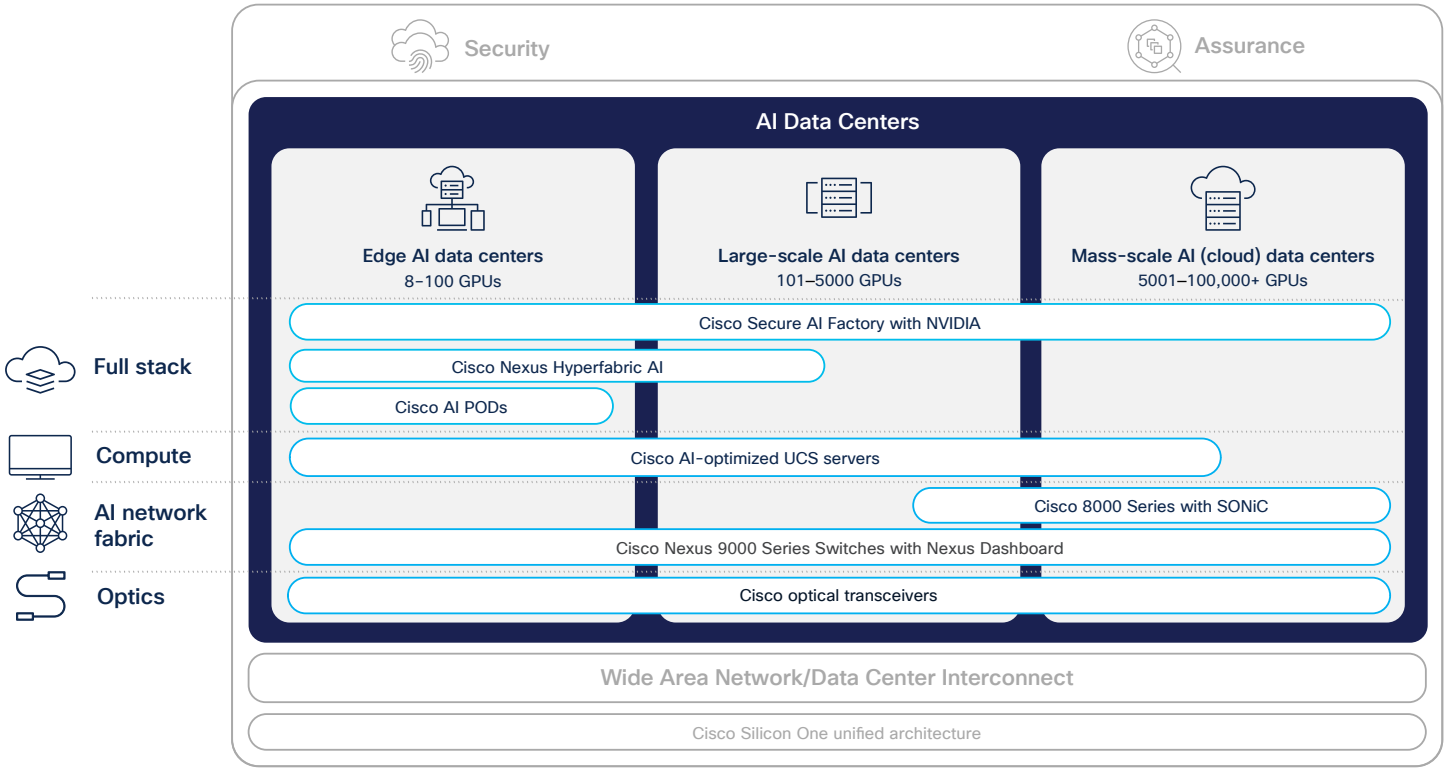

Figure 8. The Cisco AI data center portfolio delivers choice and innovation across mass-scale, large, and edge deployments

Cisco AI network fabrics

Cisco Nexus 9000 Series Switches running the widely adopted Nexus Operating System (NXOS) or application-centric infrastructure (ACI) and Cisco 8000 Series platforms running Software for Open Networking in the Cloud (SONiC) are engineered to meet the most stringent front-end and back-end networking demands of large-scale AI workloads. Specifically, Nexus 9000 and Cisco 8000 systems built on the Cisco Silicon One G200 programmable ASIC supporting high-density 400G and 800G 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 help improve job completion times (JCTs). They also provide the low latency and telemetry needed to meet the design and operational requirements of AI fabrics.

Cisco data center optics

Cisco optics innovation delivers on the key requirements for AI optics—low power, performance, scale, and reliability. In addition to technology innovation with silicon photonics platforms, Cisco's rigorous optics testing, robust monitoring, enhanced reliability, and additional performance margins are essential for AI networks. Cisco enables a broad set of AI use cases by supporting the two module form factors widely used to connect clients and switches with high port density and optimized thermal management. The Cisco OSFP 800G transceiver modules are based on the OSFP specification commonly used in AI applications, while the Cisco family of QSFP-DD modules maximize port density for 100G and 800G with backward compatibility to lower-speed QSFP modules.

Meanwhile, the Cisco optics family also includes high-speed QSFP modules for AI server connectivity. Cisco is also innovating terabit speed pluggable optics and contributing to new advances such as linear pluggable optics (LPO) capabilities that can help improve AI fabric switch power efficiencies.

Cisco compute

Cisco AI-optimized servers are designed for demanding use cases like AI fine-tuning and inferencing, among others. With their future-ready, highly modular architecture and their blend of high-performance CPUs and optional GPU acceleration, Cisco servers deliver efficient resource allocation for diverse workloads. As an example, the Cisco Unified Computing System (Cisco UCS) C885A M8 Rack Server is a dense-GPU server designed to deliver scalable accelerated compute capabilities to address the most demanding AI workloads. For environments that require flexible configurations across a broad range of AI use cases, the Cisco UCS C845A M8 Rack Server, based on NVIDIA MGX reference architecture, provides a highly scalable, modular, and customizable platform. With software-defined automation and streamlined management, the Cisco Intersight cloud operations platform provides Cisco servers with a simplified and flexible operational environment.

Cisco full-stack architecture—Cisco Nexus Hyperfabric AI

Cisco Nexus Hyperfabric AI is a vertically integrated AI cluster designed to accelerate and simplify deployments with plug-and-play cloud management. With Cisco Nexus Hyperfabric AI clusters, organizations are able to streamline the entire infrastructure lifecycle process by deploying an innovative full stack that integrates Cisco compute with NVIDIA GPUs and DPUs, Silicon One-based Cisco 6000 Series Switches, Cisco optics, and VAST Data storage. The reference architecture includes NVIDIA AI Enterprise deployed and supported on NVIDIA-certified Cisco UCS C885A M8 Rack Servers and adheres to the NVIDIA Enterprise Reference Architecture (Enterprise RA) for NVIDIA HGX™ and Spectrum-X.

Cisco full-stack architecture—Cisco Secure AI Factory with NVIDIA

Cisco is partnering with NVIDIA to empower organizations to implement, optimize, and secure AI deployments. Cisco Secure AI Factory with NVIDIA is a secure and high-performance AI infrastructure that integrates security, networking, compute, AI software, and storage into a scalable full-stack system. Featuring built-in security at every layer, superior networking, and seamless integration with the NVIDIA AI Enterprise software platform, it is purposefully designed to enable organizations to streamline the development, deployment, and protection of AI workloads.

Recognizing the unique paths organizations take in their AI journey, Cisco Secure AI Factory with NVIDIA offers deployment flexibility by offering a complete vertically integrated solution or a flexible modular AI architecture tailored to support a broad range of needs. The vertically integrated option integrates the Cisco AI security portfolio into the Cisco Nexus Hyperfabric AI full-stack solution.

Cisco full-stack architecture for inferencing—Cisco AI PODs

Cisco helps organizations get their inferencing environments up and running with less time, effort, and human error with Cisco AI PODs. With validated configurations, rapid and consistent refreshes, and a single support model, Cisco AI PODs help mitigate the complexities of AI integration and deployment and deliver a secure and scalable path from initial deployment to support for the most advanced applications.

WAN and data center interconnect

Cisco delivers a full portfolio of WAN and Data Center Interconnect (DCI) products and solutions to ensure an uninterrupted and secure AI workflow across distributed data center and edge environments.

These solutions are designed to provide exceptionally reliable, secure, and intelligent connectivity for all three use cases: inter-cluster connectivity, any-to-any data center connectivity, and edge AI connectivity.

Cisco coherent pluggable optics

The Cisco family of coherent pluggable optics plays a significant role in enabling connectivity between data centers and AI clusters. These optical transceivers promote high-speed transmission, optimize power efficiency, and simplify network architecture and operations. They provide solutions for a wide range of data rates and AI applications over local intra-data center, metro, long-haul, and ultra-long-haul reaches.

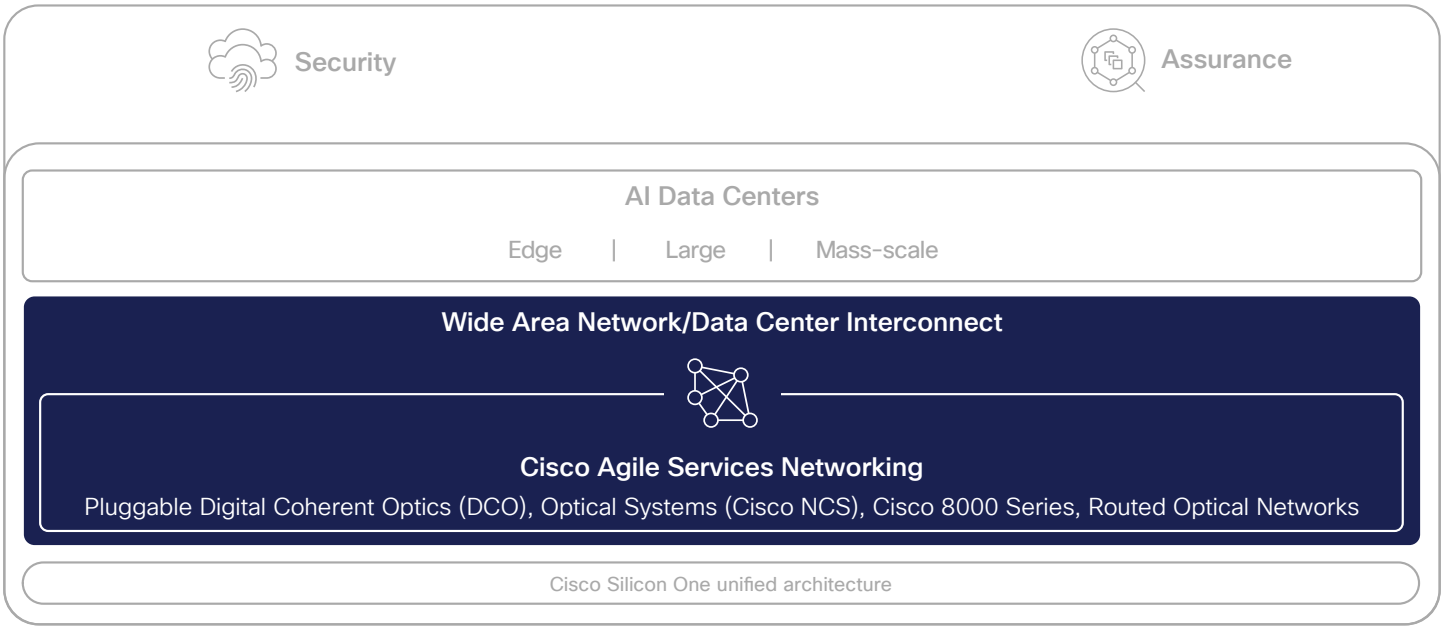

Figure 9. Cisco Agile Services Networking delivers secure and flexible connectivity across the complete agentic AI workflow

Cisco optical systems

As data center interconnect requirements expand beyond the scale and distance addressed by coherent pluggable optics solutions, Cisco offers the Network Convergence System (NCS), a dense wavelength-division multiplexing (DWDM) line system. The Cisco NCS 1000 Series are optimized to efficiently interconnect multiple AI data centers across a variety of use cases, distances, and throughputs ranging from 2 Tbps to 28 Tbps.

Cisco SD-WAN

Cisco SD-WAN provides organizations with a robust framework that can support the deployment and management of agentic AI workflows combined with other enterprise and cloud traffic by simplifying and securing reliable connectivity between all locations. Cisco SD-WAN optimizes network resources and routes traffic intelligently so that AI workflows can function effectively. The integration of Cisco SD-WAN and Cisco Secure Access capabilities delivers a cloud-native, integrated secure access service edge (SASE) approach that protects AI data and workflows across the cloud-to-edge continuum and delivers network and security policy management and enforcement consistently across all environments.

Cisco Agile Services Networking and Cisco Routed Optical Networking

The Cisco Agile Services Networking architecture enables flexibility in the deployment of large-scale 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 Routed Optical Networking innovations, this architecture helps service providers significantly lower TCO.

With the Cisco Agile Services Networking architecture, service providers and organizations with large-scale networking needs can optimize the delivery of assured networking to customers and users. A single operating system and unified management across both routing and optical layers simplifies operations. Additionally, Agile Services Networking integrates AI-enabled assurance and automation with the Cisco Provider Connectivity Assurance and Crosswork Network Automation suites, enabling seamless and intelligent network operations.

Security and assurance

As AI becomes central to the running of almost every organization, the ability to protect and ensure the end-to-end service levels of AI workflows becomes critical. The distribution of AI workflows and agents across on-premises, hybrid, and multicloud environments attracts the increased risk of security threats and the need to ensure data integrity and confidentiality at every touch point.

Cisco security and assurance solutions (Figure 10) can help organizations deliver AI-driven outcomes consistently across the complete workflow with confidence and lower risk. That includes protecting the data and maintaining the performance of training or inferencing clusters within data centers, interactions between inferencing and training models over a WAN, or edge agents that are continually collaborating on autonomous processes.

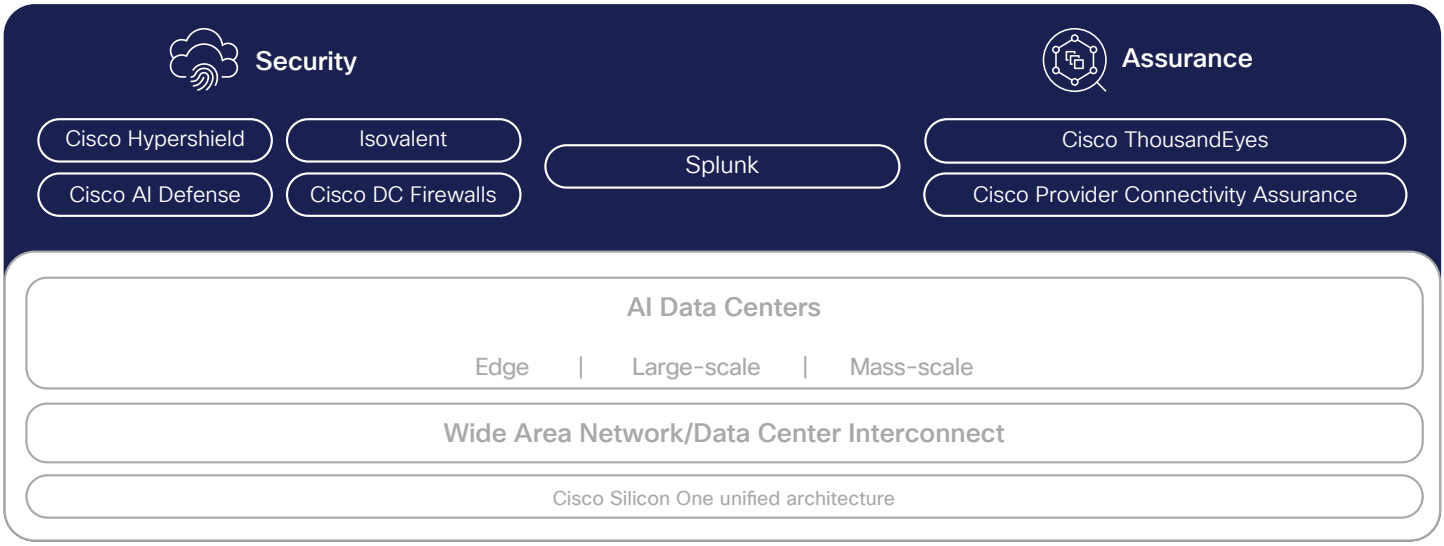

Figure 10. Cisco security and assurance solutions deliver comprehensive digital resilience for AI workflows

With the Cisco Cloud Protection Suite, organizations can dynamically protect applications and users wherever they are located. By automating segmentation and using consistent zero-trust policies with an advanced AI-enabled security architecture, organizations can protect modern applications and AI-scale data centers across public, private, and hybrid clouds.

The suite includes Cisco Hypershield and the Isovalent platform, both based on the extended Berkeley Packet Filter (eBPF) technology that offers a modern approach to securing cloud-native environments.

Cisco Hypershield delivers a unique, distributed, AI-native security architecture based on Kubernetes that is 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 security policy lifecycle and security infrastructure upgrades. The Isovalent platform provides zero-trust networking and lightweight, highly efficient network observability and security tools, all tailor-made for Kubernetes and cloud environments.

AI workflows are becoming increasingly distributed and complex, requiring visibility and assurance at every stage to consistently ensure the required service levels. This is the case for the performance of training or inferencing clusters within a data center, interactions over a WAN between inferencing and LLMs, or for edge agents that are collaborating on an autonomous process.

Cisco Splunk Observability Cloud delivers real-time correlated visibility and insights across the complete infrastructure to ensure the fastest time to troubleshoot and remediate issues. It integrates, aggregates, and correlates data from a broad set of Cisco and third-party monitoring and assurance tools, including Cisco Nexus Dashboard, Cisco Provider Connectivity Assurance, and Cisco ThousandEyes.

Conclusion

No single architecture will satisfy all requirements of the variety of participants in the agentic AI ecosystem. Mass-scale AI data center architectures required by hyperscalers and neocloud providers for pre-training and fine-tuning LLMs will differ substantially from those required by enterprises for inferencing at the edge. And these in turn will differ from the converged, intelligent transport network spanning access and metro and a highly scalable, future-proof IP core required by communication service providers.

Cisco is helping organizations build their AI capabilities at all scales. Combinations of our Cisco 8000, Silicon One, optics, and optical systems are being deployed by five of the largest hyperscalers in their mass-scale back-end training networks. Our complete portfolio of Cisco-developed AI infrastructure technologies, from silicon to full-stack systems, is designed to help AI ecosystem participants thrive in the agentic AI era by delivering innovative solutions that meet the unique demands of complex and resource-intensive AI workflows. These cutting-edge solutions span vertically integrated full-stack systems, high-performance silicon, optics, compute, networking, and software to meet AI needs across the full spectrum of organizations and use cases.

Explore how the Cisco AI infrastructure portfolio and mass-scale AI infrastructure solutions can help your organization build efficient, high-performance AI infrastructure faster while minimizing risk and maximizing ROI.

Notes

1 Gartner Predicts 40% of Generative AI Solutions Will Be Multimodal By 2027, Gartner, September 9, 2024.

2 Gartner Identifies the Top 10 Strategic Technology Trends for 2025, Gartner, October 21, 2024.

3 Building a scalable foundation for AI's future, Outshift, January 22, 2025.

4 AI power: Expanding data center capacity to meet growing demand, McKinsey & Company, October 19, 2024.

5 Gartner Forecasts Worldwide Public Cloud End-User Spending to Total $723 Billion in 2025, Gartner, November 19, 2024.

6 Hype Cycle for Edge Computing, 2024, Gartner, July 15, 2024.

7 AI Neocloud Playbook and Anatomy, SemiAnalysis, October 3, 2024.

8 OpenAI o3 Breakthrough High Score on ARC-AGI-Pub, ARC Prize, December 20, 2024.

9 Forecast Analysis: AI Semiconductors, Worldwide, Gartner, May 6, 2024.

10 The state of AI in 2023: Generative AI's breakout year, McKinsey & Company, August 1, 2023.

11 Cisco 2024 AI Readiness Index, Cisco, 2024.

12 Cisco Silicon One Breaks the 51.2 Tbps Barrier, June 20, 2023.