Cisco Secure AI Factory with NVIDIA

Cisco Secure AI Factory with NVIDIA is a scalable, high-performance, and secure AI infrastructure from Cisco, NVIDIA, and our strategic partners to help accelerate the adoption of AI for enterprises.

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

Benefits

- **Security-first architecture:** Enables safe enterprise AI. Cisco Secure AI Factory with NVIDIA integrates Cisco security solutions including Cisco AI Defense and Cisco Hypershield to provide comprehensive protection against cyber threats, ensuring data privacy and security across AI applications.
- **High-performance AI infrastructure:** Enables efficient model training, customization, and inferencing. Utilizing Cisco UCS® AI Servers and Cisco® switches, the Secure AI Factory with NVIDIA offers high-speed, low-latency networking and powerful computational capabilities, essential for demanding AI workloads.
- **Pre-validated AI infrastructure stack with flexible deployment options:** Improves data scientists' and developers' productivity. The Secure AI Factory with NVIDIA can be deployed as a vertically integrated stack, which simplifies deployments and improves operational efficiency.

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

Overview

In today's rapidly evolving technological landscape, enterprises face the dual challenge of ensuring robust security while harnessing the power of AI. Cisco Secure AI Factory with NVIDIA, in collaboration, offers a comprehensive solution that embeds security at every layer, ensuring high-performance and a scalable AI infrastructure.

- **Security-first AI infrastructure:** Unlike other AI solutions, Cisco Secure AI Factory with NVIDIA embeds security at every layer of the stack.
- **High-performance enterprise networking:** Cisco's market-leading high-performance networking that enterprises have trusted for 40 years is now even more critical for the success of AI initiatives.
- **Modular and flexible deployment:** Choose between an integrated AI stack or a modular system and solutions approach.
- **Enterprise-grade** scalability supports everything from data-engineering to AI model training, fine tuning, RAG, and inference workloads.
- **Seamless integration with NVIDIA AI Enterprise software:** Optimized for NVIDIA-powered AI workloads.

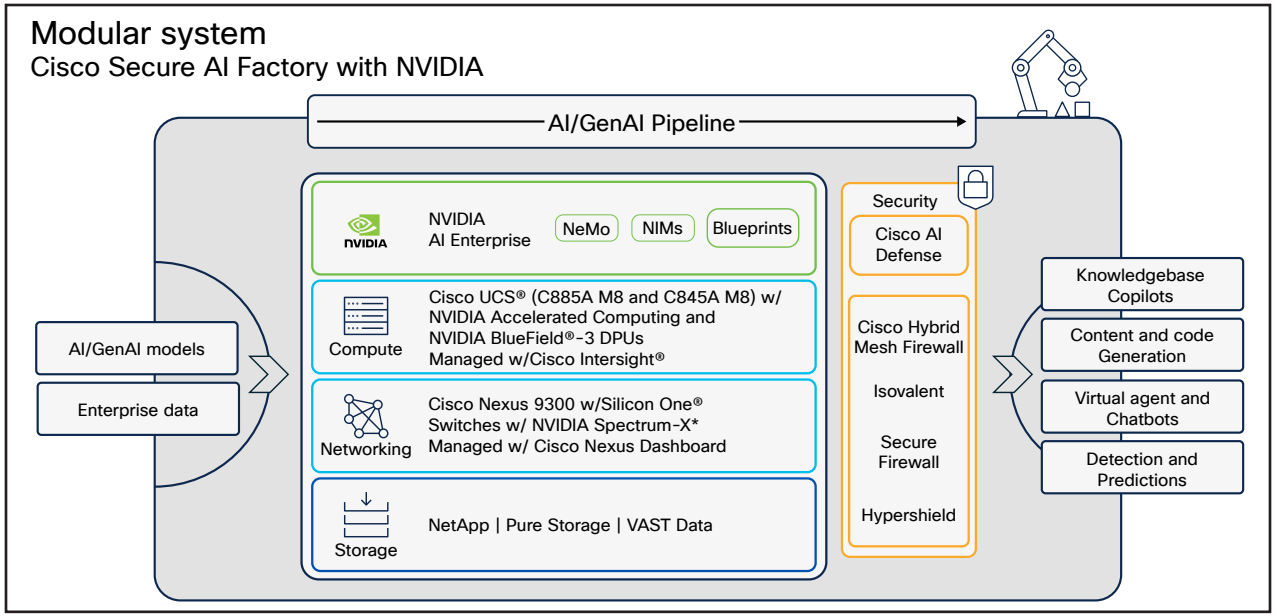

Figure 1. Cisco Secure AI Factory with NVIDIA modular system

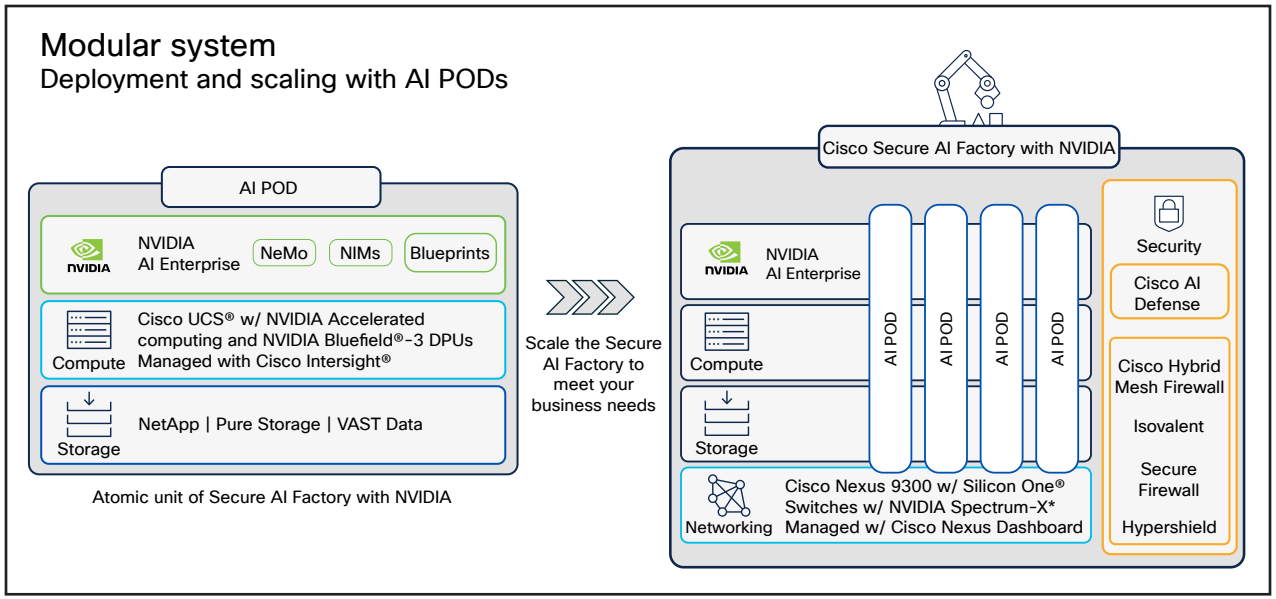

Figure 2. Cisco Secure AI Factory with NVIDIA modular system scaling with AI PODs

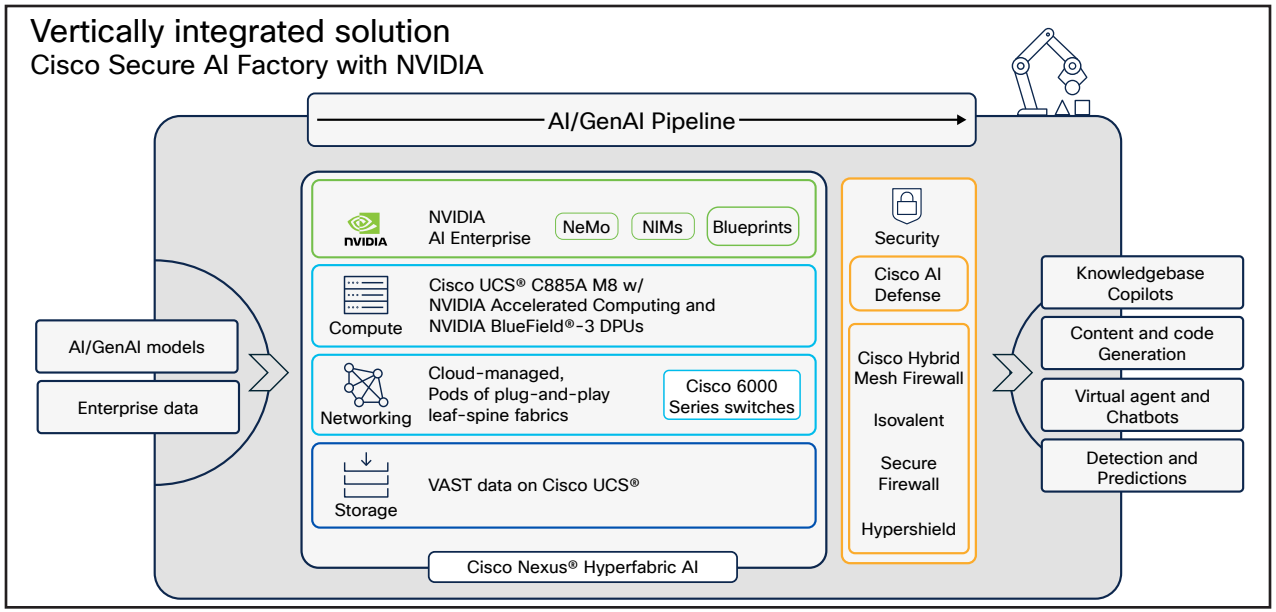

Figure 3. Cisco Secure AI Factory with NVIDIA vertically integrated solution

Market trends and opportunities

The AI market is experiencing significant growth, with enterprises increasingly adopting AI to enhance customer experiences and operational efficiency. According to IDC, the AI market is expected to surpass $749 billion by 2028, driven by advancements in AI/ML technologies and their applications across various industries. This growth presents opportunities for you to leverage AI for improved business outcomes, while also highlighting the need for secure AI deployments.

Challenges in AI deployment

- **AI security vulnerabilities:** AI models, frameworks, applications, data, and supporting infrastructure represent new surfaces for cyberattacks. Prompt injection, adversarial attacks, model poisoning and thefts, data leaks, unauthorized GPU access, etc., are some of the risks associated with AI projects.
- **Complex AI deployments:** Operationalizing a secure and scalable AI infrastructure stack for the ever-changing needs of AI pipelines is complex and expensive, and requires an integrated, pre-validated AI infrastructure that includes compute, networking, and storage working together as a single platform.

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

- **Networking performance bottlenecks:** Various phases of AI pipelines (for example, model training, checkpointing, fine-tuning, RAG, and inference) generate much GPU-to-GPU (east-west) and GPU-to-storage (north-south) traffic. Lack of enterprise-grade networking can delay delivery of desired technical and business outcomes.

How it works

The following products establish the foundation of Cisco Secure AI Factory with NVIDIA:

Security:

- Cisco Hybrid Mesh Firewall managed via Cisco Security Cloud Control (SaaS managed)
  - Isovalent – Network and runtime security on Kubernetes workloads
  - Cisco Secure Firewall – Advanced threat protections for encrypted and unencrypted traffic using NGFW (next generation firewall) technologies
  - Hypershield – Advanced segmentation using AI, and enforcement points within workloads
- AI Defense – Protects Gen AI LLMs and apps from cyber threats. Enables models discovery, threat assessment, and runtime protection

AI Infrastructure Deployment Options:

Option 1: Ready-to-deploy Infrastructure:

- Cisco Nexus Hyperfabric AI
  - **Management:** Cloud-based management
  - **Accelerated compute:**
    - Cisco UCS C885A M8 (NVIDIA HGX platform w/ NVIDIA H200 GPUs)
  - **Networking:**
    - Cisco 6000 Series swtiches with Cisco Silicon One, includes Cisco Optics family of optical modules to offer customer choice and deliver super high densities.
  - **Partner products:**
    - NVIDIA AI Enterprise Software Platform
    - Storage: VAST Data

Option 2: Build-your-own Infrastructure (Modular):

- Cisco products
  - **Accelerated Compute:**
    - Cisco UCS C885A M8 (NVIDIA HGX™ platform w/ NVIDIA HGX H100 and H200 GPUs) and Cisco UCS C845A M8 (NVIDIA MGX™ modular reference design w/ NVIDIA H100 NVL, H200 NVL, and L40S PCIe GPUs) managed by Cisco Intersight.
  - **Networking**:
    - Nexus 9K switches with Cisco Silicon One, managed by Nexus Dashboard
    - Future Cisco switches with NVIDIA Spectrum™-X silicon, managed by Nexus Dashboard
    - NVIDIA BlueField®-3 DPUs and SuperNICs, managed by Nexus Dashboard
- Partner products
  - NVIDIA AI Enterprise software platform
  - Storage: NetApp, Pure Storage, VAST Data

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

Cisco Capital

Financing to help you achieve your objectives

Cisco Capital® can help you acquire the technology you need to achieve your objectives and stay competitive. We can help you reduce CapEx. Accelerate your growth. Optimize your investment dollars and ROI. Cisco Capital financing gives you flexibility in acquiring hardware, software, services, and complementary third-party equipment. And there's just one predictable payment. Cisco Capital is available in more than 100 countries. Learn more.

Take the next step in AI security

Ensure your AI infrastructure is secure and efficient with Cisco Secure AI Factory with NVIDIA. For more information, contact your Cisco Services sales representative or Cisco authorized channel partner to learn more about deploying Cisco Secure AI Factory with NVIDIA.

Use cases

Name	Description
Knowledgebase copilots	• AI Assistants
Content and code generation	• Text | Images | Video | Code
Reporting and data analytics	• Summarize texts | Generate visualization
Language translation	• Multilingual real-time communication
Virtual agents and chatbots	• Specialized domain-specific chatbots
Detection and prediction	• Forecasts | Anomalies | Insights

The Cisco Advantage

Cisco Secure AI Factory with NVIDIA provides enterprises with a secure, scalable, and high-performance AI infrastructure that enables enterprises to deploy and manage AI workloads efficiently. Featuring built-in security, superior networking, and seamless integration with the NVIDIA AI Enterprise software platform, it stands as the definitive choice for innovation. Whether you require an integrated solution or a modular AI architecture, Cisco Secure AI Factory with NVIDIA delivers the most advanced and secure AI infrastructure, streamlining and expediting the secure development and deployment of next-generation AI applications.

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

C22-5040651-00 03/25