FlexPod®

Secure AI Factory with FlexPod AI

The rise of generative AI

Generative AI has seen exponential growth in recent years, with large language models (LLMs), diffusion models, and foundation models leading the charge. These technologies are enabling unprecedented capabilities in natural language understanding, image generation, code completion, and more.

LLMs like GPT, Claude, and Gemini are demonstrating capabilities that were once thought to be impossible, while diffusion models like Stable Diffusion are revolutionizing creative workflows. And foundation models serve as versatile platforms for multiple downstream tasks.

However, deploying these models in enterprise environments comes with challenges, such as:

Compute infrastructure. AI workloads require massive parallel processing capabilities, often needing specialized GPUs such as NVIDIA L40S or H100.

Security. Ensuring the security of AI systems is critical, as they introduce unique attack surfaces ranging from training data pipelines to inference endpoints, which require dedicated security strategies to prevent data breaches, model poisoning, and adversarial attacks.

Scalability. As models grow, scaling compute, networking, and storage infrastructure without bottlenecks is complex.

Vendor lock-in. Proprietary software licensing can lock enterprises into inflexible terms, affecting agility and cost efficiency.

The need for security in AI

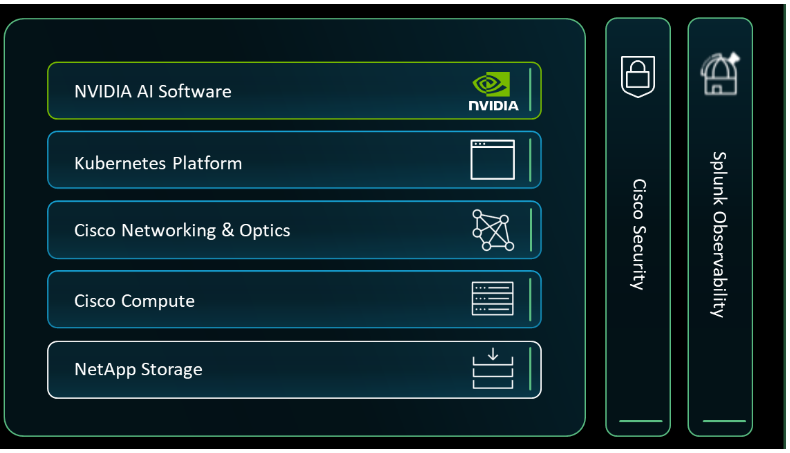
AI systems now sit at the heart of mission-critical operations

What is it?

Reference architecture

Validated solutions and turnkey offerings

Differentiated with Security and Observability



across industries such as finance, healthcare, manufacturing, and government. Their ability to process sensitive data, make autonomous decisions, and influence large-scale outcomes makes them highly attractive targets for cyberthreats. Unlike traditional IT systems, AI introduces unique attack surfaces, ranging from training data pipelines to inference endpoints, which require dedicated security strategies.

AI also presents several risks, such as:

Data breaches. Training datasets often contain proprietary business information, regulated personal data, or intellectual property. A breach not only risks legal and compliance violations but can also lead to loss of competitive advantage.

Model poisoning. Attackers can insert malicious data during training to subtly or drastically alter model behavior. This act can result in biased decisions, misinformation generation, or deliberate system sabotage.

Prompt injection and adversarial attacks. In generative AI systems, crafted prompts or inputs can manipulate the model into producing harmful or unintended outputs. Similarly, adversarial examples—specially modified inputs—can cause models to misclassify or to misinterpret critical information.

Model theft and reverse engineering. Sophisticated adversaries can attempt to replicate or to steal proprietary AI models through repeated queries, undermining the organization’s investment in training and R&D.

Pipeline and supply chain attacks. AI development pipelines rely on multiple open-source libraries, cloud APIs, and data sources. Compromising any element in this chain can introduce vulnerabilities into the deployed AI systems.

Introduction to Secure AI Factory

Secure AI Factory represents a new paradigm in enterprise AI infrastructure—one that tightly integrates compute, networking, and storage with security as a foundational element rather than as an afterthought. It provides a standardized, validated, and secure architecture to help enterprises accelerate AI deployment while

Streamline AI Deployment

Simple

Start small and expand with a pre- configured, pre-validated, turnkey solution that is fast, easy to deploy, and cost effective.

Secure Enterprise Data

Secure

Retain peace of mind knowing your data and investments are both protected.

Scalable Solution for Growth

Scalable

Deploy a high-performance solution with the flexibility to scale alongside the needs of your business.

safeguarding critical data, intellectual property, and operational integrity.

Strategic partnership: Cisco, NVIDIA, and NetApp

Secure AI Factory takes advantage of the combined strengths of three technology leaders:

Cisco brings best-in-class networking, Zero Trust security frameworks, and workload protection through innovations such as Cisco Secure AI Defense, Hypershield, and Hybrid Mesh Firewall.

NVIDIA provides the AI acceleration foundation, including GPUs such as L40S and H100, NVIDIA AI Enterprise software, and optimized frameworks for LLMs, generative AI, and deep learning.

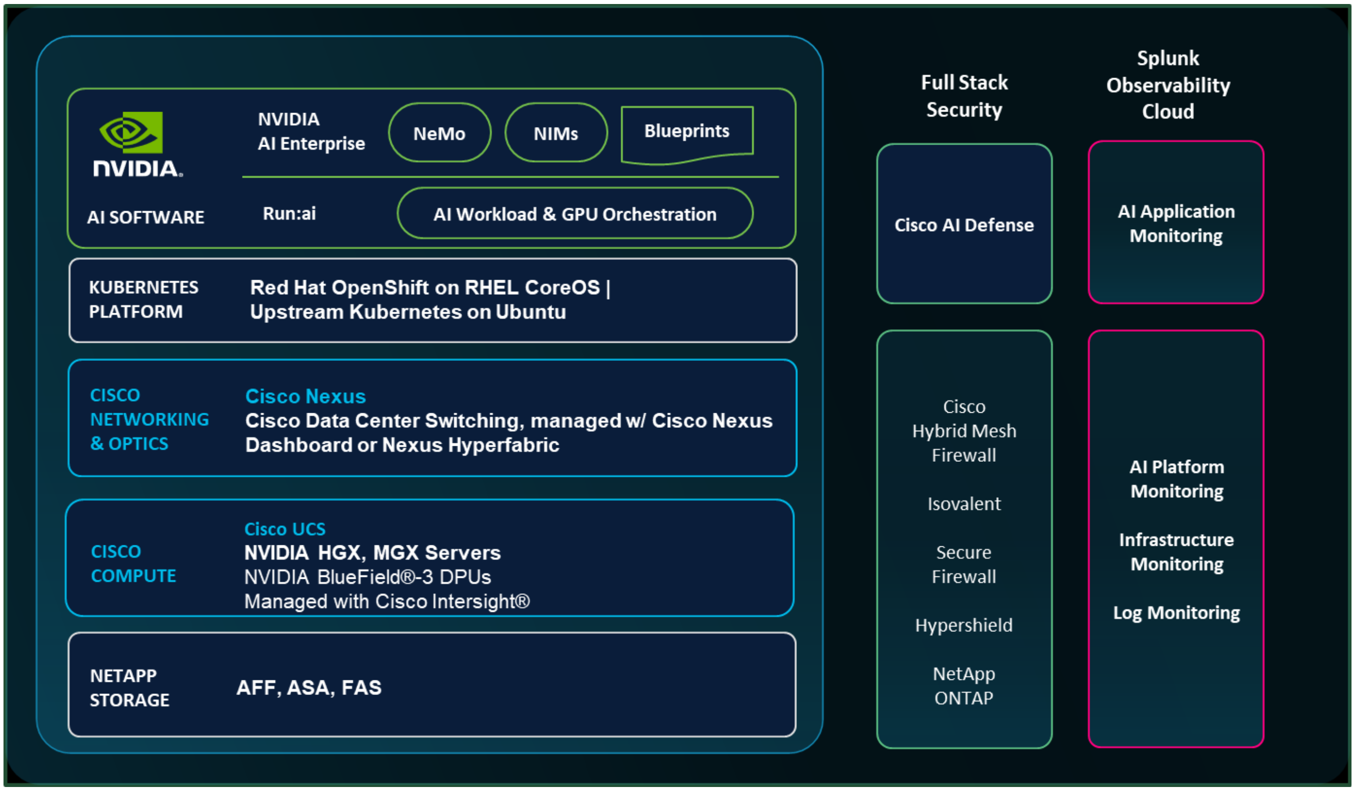
NetApp delivers AI-ready data management and storage, with high-performance NetApp® AFF A-Series systems, data orchestration through NetApp Trident™ software, and ransomware protection through BlueXP.

With this collaboration, every layer—from the GPU compute nodes, to the network fabric, to the storage and data pipelines—is validated, secure, and performance-optimized.

FlexPod AI introduction

At the core of Secure AI Factory lies FlexPod® AI, a cutting-edge platform that serves as the backbone for building and scaling secure, enterprise-grade AI capabilities in an organization's data center. Co-engineered by Cisco and NetApp, FlexPod AI is purpose-built to overcome the unique challenges of modern AI workloads, providing the performance, reliability, and flexibility required to unlock the full potential of AI. With its robust architecture, FlexPod AI seamless integrates with the Secure AI Factory framework, enabling organizations to accelerate AI adoption while maintaining strict security, governance, and operational standards.

When augmented with NVIDIA GPUs and NVIDIA AI Enterprise software, FlexPod AI becomes a turnkey solution that not only simplifies deployment but also


NVIDIA
AI SOFTWARE
NVIDIA AI Enterprise
NeMo
NIMs
Blueprints
Run:ai
AI Workload & GPU Orchestration
KUBERNETES PLATFORM
Red Hat OpenShift on RHEL CoreOS | Upstream Kubernetes on Ubuntu
CISCO NETWORKING & OPTICS
Cisco Nexus
Cisco Data Center Switching, managed w/ Cisco Nexus Dashboard or Nexus Hyperfabric
CISCO COMPUTE
Cisco UCS
NVIDIA HGX, MGX Servers
NVIDIA BlueField®-3 DPUs
Managed with Cisco Intersight®
NETAPP STORAGE
AFF, ASA, FAS
Full Stack Security
Cisco AI Defense
Cisco Hybrid Mesh Firewall
Isovalent
Secure Firewall
Hypershield
NetApp ONTAP
Splunk Observability Cloud
AI Application Monitoring
AI Platform Monitoring
Infrastructure Monitoring
Log Monitoring

delivers the scalability and efficiency needed to power AI innovation at any scale. By anchoring Secure AI Factory on FlexPod AI, organizations gain a solid foundation to propel transformative outcomes—from data preparation and model training to inference and deployment—all within a secure, enterprise-ready environment.

FlexPod AI integrates:

Cisco UCS compute platforms for high-density, GPU-optimized processing.

Cisco Nexus networking for ultralow-latency and high-bandwidth data flows.

NetApp AFF all-flash storage for AI-ready, high-performance data pipelines.

NVIDIA GPUs (such as L40S and H200) and **AI Enterprise software** for accelerated model training, fine-tuning, and inference.

FlexPod AI and Secure AI Factory with NVIDIA

Workload Security

Cisco Secure AI Defense is designed to address the emerging and unique security challenges of AI workloads. Unlike traditional applications, AI models and pipelines are vulnerable to threats such as model poisoning, prompt injection, and adversarial attacks.

Cisco Hypershield is an advanced security layer that delivers deep run-time protection for workloads, with a special focus on east-west traffic security inside data centers and cloud environments.

Isovalent Secure Firewall provides container-native, microsegmentation security for Kubernetes and cloud-native AI workloads with Extended Berkeley Packet Filter (eBPF)-powered firewalling.

Infrastructure Security

Cisco Hybrid Mesh Firewall is a next-generation, distributed security solution that's purpose-built for multisite and hybrid cloud environments, making it an ideal fit for securing AI workloads that span on-premises data centers, edge locations, and public clouds.

Data Security

With the **Zero Trust architecture with FlexPod AI**, no user, device, or workload is trusted by default, even if they are inside the enterprise network. Security is based on continuous authentication, strict access controls, and policy-driven authorization. FlexPod AI is delivered with OS- and firmware-level security hardening to reduce the attack surface from the outset. This feature includes secure boot and firmware validation, patch management and vulnerability remediation, and configuration baselines. Ransomware poses a significant risk to AI environments, which rely on large volumes of valuable training data. FlexPod AI integrates multilayered ransomware defense, including anomaly detection, immutable backups, and rapid recovery. NetApp AFF A-Series systems are the high-performance storage foundation for FlexPod AI, purpose-built for AI workloads. Benefits include: Ultralow latency, Parallel I/O optimization, End-to-end NVMe.

In conclusion, Secure AI Factory with FlexPod AI is designed to accelerate AI adoption by providing a robust, secure, and performance-optimized infrastructure. This comprehensive solution overcomes the unique challenges of deploying AI at scale, enabling enterprises to innovate rapidly while maintaining rigorous security and compliance standards.