FlexPod®

Simplify AI infrastructure and operations with FlexPod

The challenge: Moving AI from pilot to production

AI poses a significant opportunity for organizations to extract more intelligence from their data throughout its lifecycle, make better decisions faster, and introduce new capabilities into their offerings. However, many enterprises struggle to move AI from pilot into production as new demands on their compute, network, and storage systems require consistent high performance and high-capacity infrastructure. As a result, the impact of AI infrastructure extends beyond algorithms, reshaping the foundational requirements of compute, networking, and storage technologies. With every stalled pilot or fragmented stack, AI failures can increase operational debt and business risk.

Unlike fragmented approaches, FlexPod®, co-developed by NetApp® and Cisco, delivers a secure, unified, and validated architecture to help accelerate AI adoption while reducing business risk. It unites innovative compute, storage, networking, and management to streamline operations. With FlexPod, enterprise IT teams can simplify operations on a common platform and automate deployment of AI workloads.

Together with a robust partner ecosystem and pre-tested reference architectures such as Cisco Validated Designs (CVDs) and NetApp Validated Architectures (NVAs), FlexPod for AI helps enterprises deploy accelerated compute and high-performance storage on proven solutions.

A production-ready AI platform: The FlexPod advantage

FlexPod AI empowers enterprise IT teams to reduce operational complexity and take control by unlocking the value of their own data.

Simplify and unify AI operations: By streamlining from core to edge, enterprises can achieve global visibility, consistency, and control at scale.

- Run AI and enterprise workloads side-by-side with a unified operational model that simplifies system upgrades and improves reliability to support faster time to market and a longer platform lifespan.
- FlexPod offers an AI-ready infrastructure that combines Cisco UCS compute systems, Cisco Nexus networking, and high-performance storage from NetApp for:
  - Flexible CPU/GPU ratios and cloud-based management across core and edge.
  - High performance, throughput, and lossless fabrics needed for AI/ML workloads.
  - Efficiency and scalability to support massive datasets.

Automate and accelerate IT deployments:
Take advantage of Ansible playbooks to automate deployment and operations, enabling enterprises to save time and reduce risk associated with manual tasks.

With these playbooks, customers can:

- Streamline processes to minimize human errors, accelerate time to market, and allow teams to focus on more strategic initiatives.
- Ensure consistency and reliability for each deployment with a consistent set of practices and guidelines.
- Enable scalability and expanded AI adoption with an adjustable framework that can be modified to accommodate new technologies and methods.

Secure, future-ready AI infrastructure:
Keep infrastructure running smoothly with measures such as device hardening, microsegmentation, least-privilege access, and a secure value chain. It goes further by encrypting data in transit and at rest, and employs a robust zero-trust architecture, including multi-admin verification and multifactor authentication, to thwart malicious actors.

With the FlexPod AI platform customers can:

- Enhance data security and privacy by safeguarding sensitive data against unauthorized access and cyber threats while adhering to data-protection standards and regulations such as GDPR and HIPAA.
- Increase reliability and uptime, ensuring that the AI platform remains operational, which is crucial for businesses relying on continuous AI services.

Scaling with confidence: Purpose-built for Enterprise AI

FlexPod AI is purpose-built to meet the dynamic demands of AI workloads, providing essential features for smooth integration and improved operational efficiency. It allows for easier deployment and management of general-purpose AI workloads, drives faster value realization, and accelerates AI implementation. Proven linear scalability ensures that FlexPod consistently offers peak performance across different dataset sizes.

Powered by Cisco Intersight®, the NVIDIA HPC-X Software Toolkit, and NetApp tools such as the DataOps Toolkit, organizations can leverage FlexPod AI to optimize resource utilization, ensure optimal performance, and simplify data management tasks for users.

Benefits of FlexPod for AI:

- **Simplify AI infrastructure** with tightly coupled resources, one-call support, and seamless integration of NVIDIA AI into VMware and Red Hat OpenShift for both Kubernetes and virtual machines.
- **Operationalize AI deployments** with validated designs and automation playbooks.
- **Mitigate risk and protect data** with enhanced security features including device hardening, microsegmentation, encryption, and zero-trust architecture.

Through rigorous testing for real-world workloads, FlexPod AI provides valuable insights, comparing CPU-only performance with GPU-equipped systems to showcase its robust capabilities to usher in the future of AI with reliability, efficiency, and scalability.

Enterprise AI at scale: Bring AI to existing data

FlexPod AI with AFX, AIDE, and Secure AI Factory is designed for organizations that want to accelerate AI adoption without moving or rearchitecting their large data environments. This solution brings AI directly to where enterprise data already resides, enabling organizations to prepare, secure, and operationalize data at scale. By integrating NetApp AFX storage, AIDE data management capabilities, and Secure AI Factory principles within a lab validated architecture, FlexPod provides a consistent, production ready platform for AI pipelines that demand performance, governance, and reliability.

Accessible AI at the right scale

FlexPod AI Mini brings the power of FlexPod to environments that require simplicity, speed, and efficiency at a smaller scale. Designed especially for edge and remote environments, it provides a simple cost-effective path to AI inferencing and retrieval-augmented generation (RAG). With all the benefits of FlexPod AI, this solution is a practical entry-point for organizations looking to use data without adding unnecessary risk or sprawl.

Accelerate Generative-AI use cases: From fine-tuning to inferencing

The FlexPod Datacenter with Generative AI Inferencing offers an AI inferencing platform that equips organizations with a powerful toolset for AI workloads. It delivers low-latency, high performance and enhanced versatility that integrates compute and storage connectivity at speeds ranging from 10G to 100G. The addition of NVIDIA GPUs further amplifies processing power, enabling support for data-intensive workloads. This comprehensive solution extends its versatility by seamlessly integrating a Red Hat OpenShift Container Platform design with VMware vSphere.

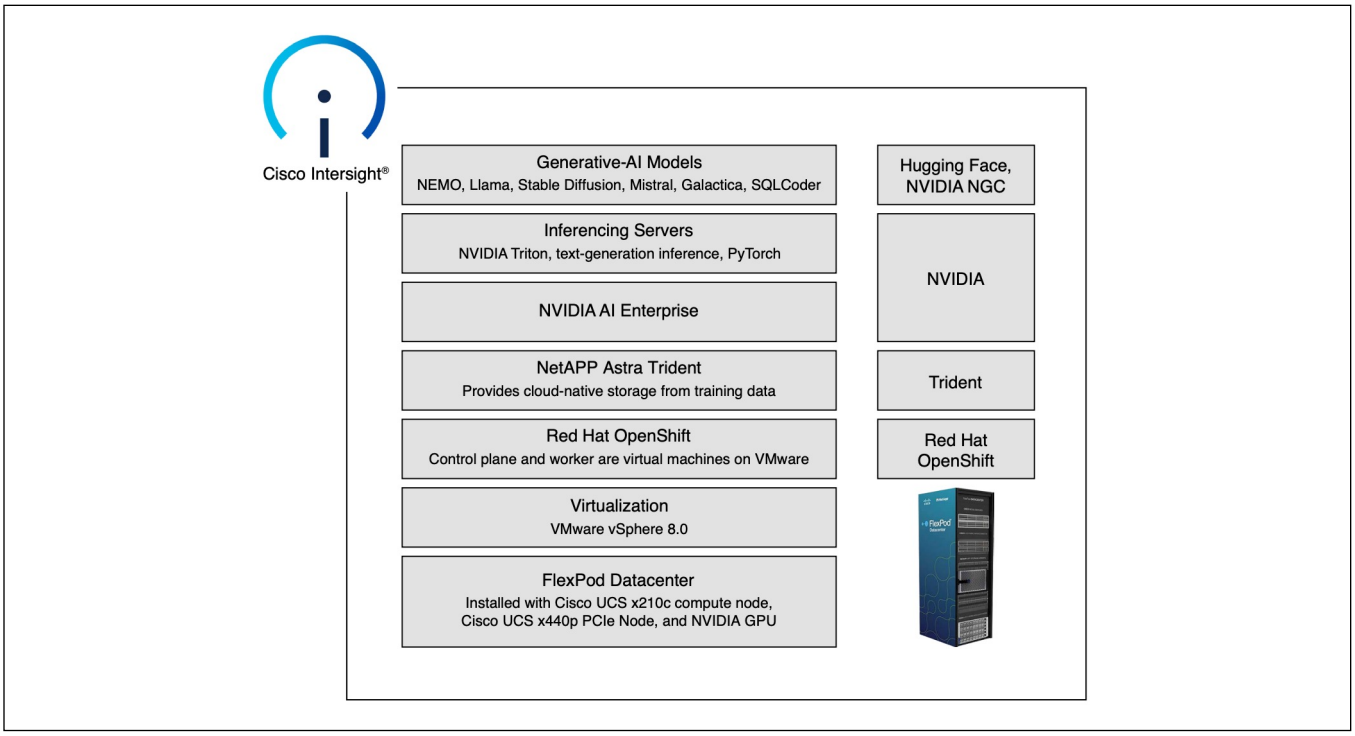

Figure 1. FlexPod for generative AI architecture

Container-ready AI at scale

FlexPod with SUSE Rancher Enterprise Container Management supports Kubernetes (K8s) workloads with high availability and server redundancy to make them more resilient, reliable, and scalable. Cisco UCS servers operate seamlessly within the FlexPod infrastructure, whether deploying SUSE Rancher Kubernetes Engine Government (RKE2) as a bare-metal cluster or virtualized on VMware vSphere or Kernel-based Virtual Machine (KVM). In addition, it incorporates the NVIDIA AI Enterprise software platform to securely deliver end-to-end AI capabilities. This platform helps accelerate the data science pipeline, covering diverse applications such as Generative AI, computer vision, and speech AI. With support and testing for a variety of models and development tools, this solution propels enterprises to the forefront of AI innovation while ensuring accessibility for every business.

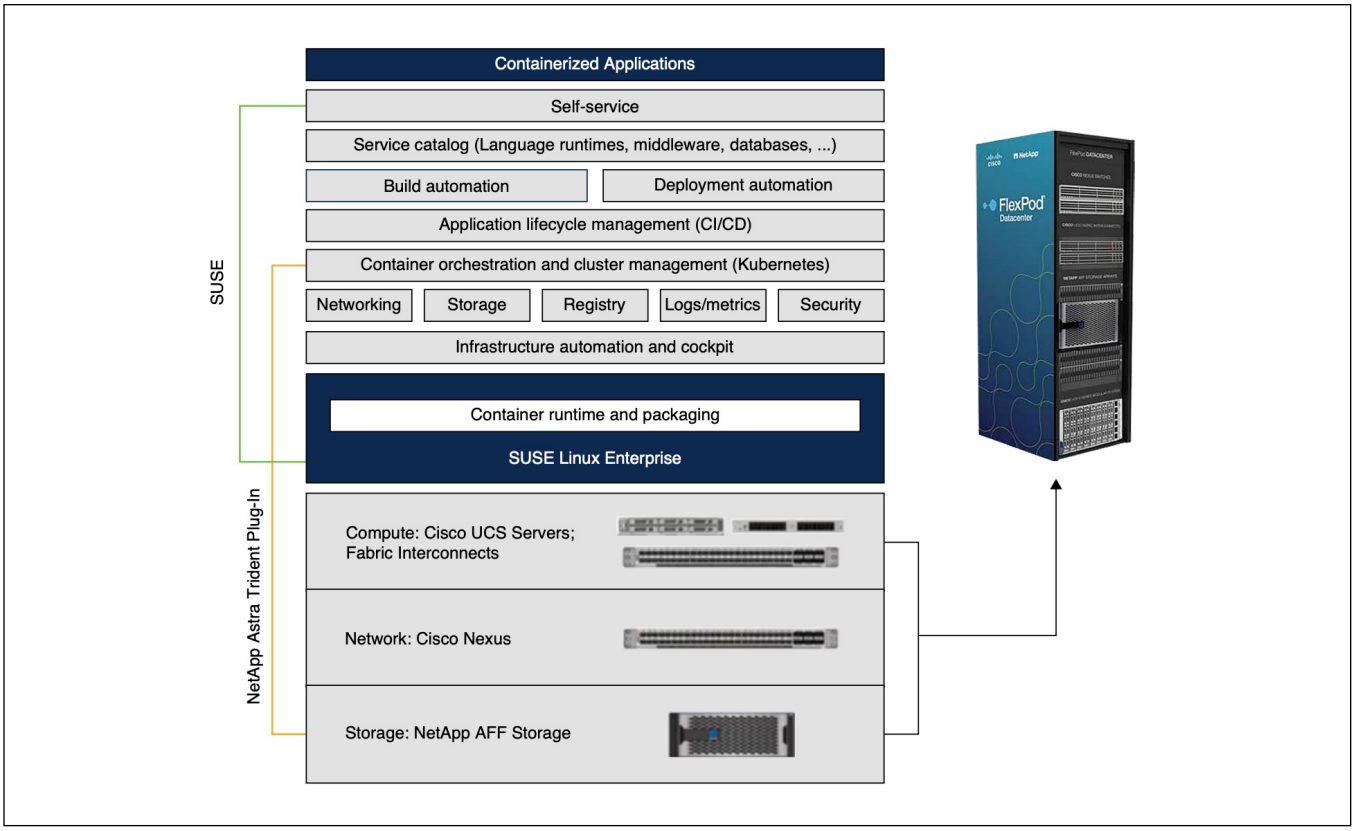

Figure 2. FlexPod with SUSE Rancher for AI workloads

Operationalize AI models with MLOps

Combining the proven capabilities of Red Hat OpenShift AI and Red Hat OpenShift, FlexPod with Red Hat OpenShift AI for MLOps helps accelerate AI pipelines and promotes intelligent application delivery to help operationalize AI use cases such as:

- **Fraud detection**, including analyzing credit-card transactions for potentially fraudulent activity.
- **Object detection**, including detecting instances of semantic objects such as humans, buildings, or cars in digital images and videos.

Ideal for experimenting, training, and deploying AI models for inferencing, it supports a broad range of custom and built-in tools, frameworks, and model-serving options (including PyTorch, TensorFlow, and NVIDIA Triton), to innovate faster with an open-source approach.

Reduce hallucinations with FlexPod AI for Retrieval-Augmented Generation (RAG)

RAG synergizes retrieval-based and generation-based methodologies to retrieve pertinent information, which is used to generate more precise and contextually appropriate responses. This hybrid model significantly enhances AI application performance by minimizing hallucinations and boosting the relevance of generated content.

Use cases for RAG include:

- Question-and-answer chatbots
- Search augmentation
- Knowledge engines

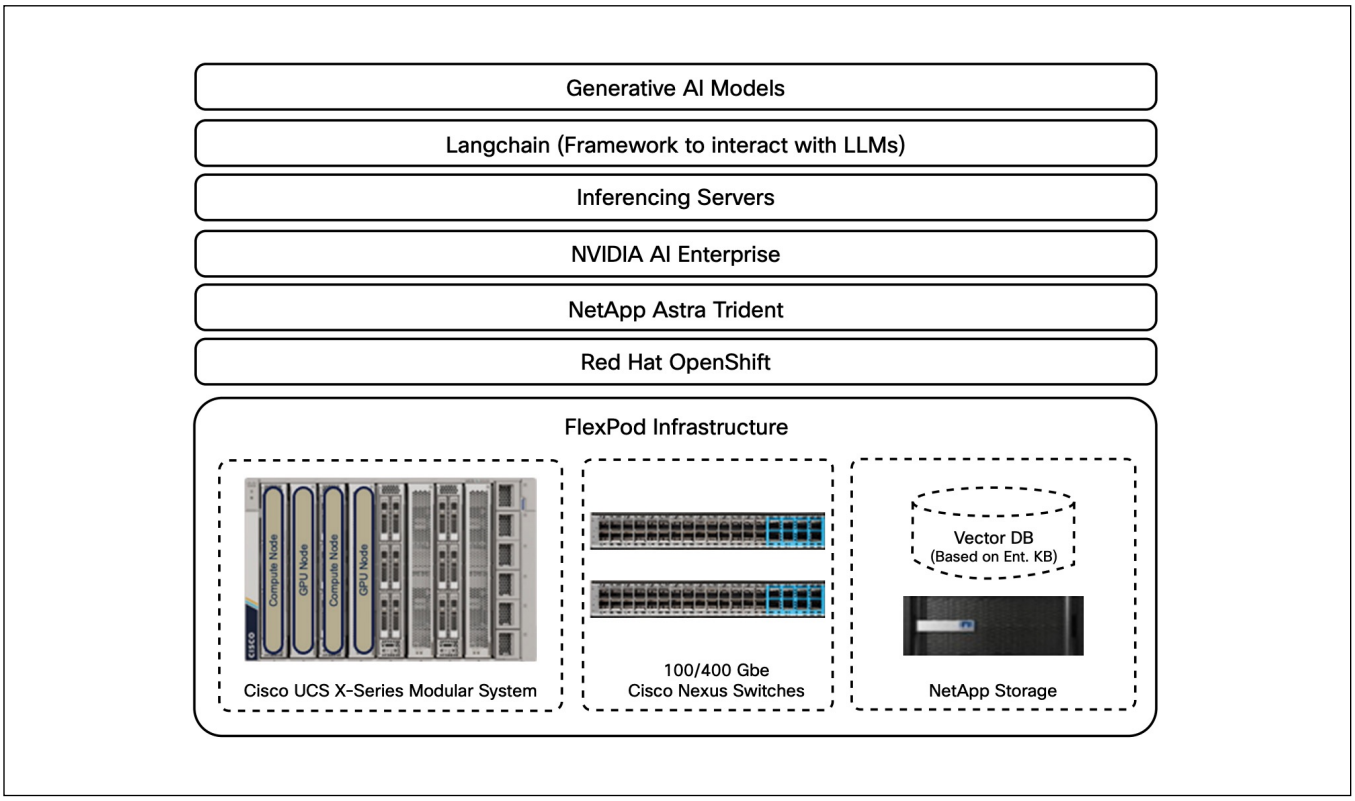

Figure 3. FlexPod for Retrieval-Augmented Generation (RAG)

FlexPod AI: Your trusted foundation for Enterprise AI

AI success depends on consistent and flexible operations. FlexPod AI removes uncertainty by delivering pre-tested, AI-ready validated solutions that enable organizations to move from pilot to production more efficiently. Unlock the full potential of your data with a unified, scalable, and secure AI infrastructure. Whether you are fine-tuning Generative AI models or deploying real-time inferencing, FlexPod accelerates your AI journey so you can move forward with confidence.

Ready to streamline your AI operations? Explore FlexPod Design Guides or contact a FlexPod expert today.