FlexPod®

Simplify AI Infrastructure and Operations with FlexPod

Overview

Artificial Intelligence (AI) presents a myriad of opportunities across every industry, transforming traditional practices and making way for more efficient processes. AI workloads are imposing fresh demands on compute, network, and storage systems. AI algorithms require sophisticated compute-processing capabilities. The need for high-performance computing has spurred advancements in hardware architectures, leading to the development of specialized processors such as Graphics Processing Units (GPUs). The implications of AI on networking and storage are equally profound. Networking infrastructure plays a critical role in facilitating the seamless exchange of data between AI systems, particularly in distributed environments. As AI applications become more data-intensive, the demand for robust, low-latency networks intensifies. Simultaneously, the storage landscape is evolving to accommodate the vast datasets required for training sophisticated AI models. High-capacity, high-speed storage solutions are essential to meet the demands of AI applications. In summary, the opportunities AI creates extend beyond algorithmic advancements, influencing the very fabric of compute, networking, and storage technologies.

AI poses a big opportunity for organizations to extract more intelligence from their data throughout its lifecycle, make better decisions and make them faster, and introduce new capabilities in their offerings. FlexPod unites innovative compute, storage, networking, and management to help enterprise IT teams to simplify on a common platform, automate deployment, and secure AI infrastructure and workloads.

Together with a full partner ecosystem, a new roadmap of Cisco Validated Designs (reference architectures) for FlexPod for AI will help IT teams deploy accelerated compute and high-performance storage on proven solutions:

- NVIDIA AI Enterprise (NVAIE) platform and Red Hat OpenShift on FlexPod for Generative-AI Inferencing.
- SUSE Rancher with NVIDIA on FlexPod for AI Workloads.
- Scaling FlexPod with GPU-intensive applications.

FlexPod Datacenter for AI provides converged infrastructure that is optimized for general purpose AI/ML workloads in the IT operations. Building on the popular FlexPod Datacenter platform, the solution includes the Cisco UCS® X-Series Modular System with Cisco UCS X-Series compute nodes, Cisco Nexus® 9000 Series Switches, Cisco UCS 6500 Series Fabric Interconnects, and NetApp AFF A-Series and C-Series flash storage arrays with NetApp ONTAP data management software.

A deeper look at an industry-leading AI platform

AI is driving big changes in data-center technology, and FlexPod for AI is helping enterprise IT teams to simplify on a common platform, automate deployment, and secure AI infrastructure and workloads.

Mainstream your AI Infrastructure: Reduce the complexity of AI infrastructure at core and edge with cloud-based operations that deliver global visibility, consistency, and control. FlexPod for AI enables you to run AI and enterprise workloads side-by-side with a unified operational model, providing faster time to value, lower risk, better scalability, easier system upgrades, better reliability, improved staff productivity, faster scaling and a longer lifespan for the platform.

FlexPod offers an AI-ready infrastructure that combines:

- Cisco UCS X-Series Modular System with X-Fabric technology allows for flexible CPU/GPU ratios and cloud-based management for computing distributed anywhere across core and edge.
- Cisco Nexus delivers the high performance, throughput, and lossless fabrics needed for AI/ML workloads.
- High-performance storage systems from NetApp complete the picture with the scalability and efficiency that large, growing data sets demand. NetApp AI tools such as the DataOps Toolkit make it simple for developers, data scientists, DevOps engineers, and data engineers to perform various data management tasks within a Kubernetes cluster.

Automate IT deployments: Deploy validated solutions for AI to save time and reduce risk, and take advantage of playbooks that automate deployment of common AI models from the cloud.

FlexPod offers reference architectures, including Cisco Validated Designs (CVDs) and NetApp Validated Architectures (NVAs). These reference architectures are extensively tested and are comprehensive blueprints that guide the selection, scaling, wiring, deployment, scaling and provisioning of AI/ML workloads. They enable customers to reduce deployment risks, streamline operations, and enhance long-term scalability, resilience, and security for their infrastructure.

Benefits of FlexPod for AI:

- Simplifies AI infrastructure with tightly coupled resources and one-call support.
- Operationalizes AI deployments with validated designs and automation playbooks.
- Offers seamless integration of NVIDIA AI into VMware and Red Hat OpenShift for both Kubernetes and virtual machines.
- Enhances security with a holistic approach including secure separation, device hardening, microsegmentation, encryption, and zero-trust architecture for data protection and threat mitigation.

With automated IT deployment playbooks (Ansible scripts), customers can:

- Streamline deployment processes by minimizing human errors and accelerate time to market by automating repetitive tasks, freeing up valuable resources, and allowing teams to focus on more strategic initiatives.
- Ensure consistency and reliability by offering a consistent set of practices and guidelines so that each AI deployment is executed with the same level of precision.
- Enable scalability and adoption for expanding AI capabilities by providing framework that can be adjusted and modified to accommodate new technologies and methods.

Make your AI platforms secure and future ready: Keep AI infrastructure running and protected with proactive, automated resiliency and security capabilities.

FlexPod prioritizes trust and security through 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 Secure AI platform customers can:

- Enhance data security and privacy by safeguarding sensitive data against unauthorized access and cyber threats and at the same time adhere to data-protection standards or regulations such as GDPR, HIPAA, etc.
- Increase reliability and uptime, ensuring that the AI platform remains operational, which is crucial for businesses relying on continuous AI services.
- Enhance trust and reputation, because customers are more likely to engage in businesses that demonstrate a commitment for protecting their data.

AI has found applications across a diverse range of industries, transforming processes and creating innovative solutions. Some notable AI use cases include:

Conversational agents
Generative AI can power chatbots and virtual assistants capable of engaging in natural, context-aware conversations. This enhances customer support services, providing users with a more interactive and personalized experience.

Text generation
This can be useful when creating blog posts, articles, stories, poems, creative writing, novels, and even YouTube scripts or social media posts. By providing a few words or sentences as input, it will generate new and unique text based on its parameters and what it learns from its trained data.

Weather modelling
In weather forecasting, AI plays a crucial role in enhancing the accuracy and granularity of predictions.

Below are some of the AI solutions we have delivered with FlexPod.

Fine-tuning and inferencing with FlexPod for generative-AI use cases

Introducing the latest advancements in the FlexPod Datacenter solution, our enhanced design encompasses several key features to elevate performance and versatility. These advancements include the implementation of low-latency switching capabilities within FlexPod designs. The solution provides a versatile base platform, accommodating integration of compute and storage connectivity at speeds ranging from 10G to 100G. With the incorporation of vSphere 8.0, users benefit from advanced virtualization capabilities. The addition of NVIDIA GPUs to the FlexPod framework further amplifies processing power, enabling support for intensive workloads. This comprehensive solution extends its versatility by seamlessly integrating a Red Hat OpenShift Container Platform design with VMware vSphere.

Additionally, the FlexPod Datacenter solution offers an AI-inferencing platform equipped with a variety of models, providing organizations with a powerful toolset for artificial intelligence workloads. We have tested various models catering to text-generation and conversational-AI use cases such as Nemotron, Google FLAN, Defog SQLCoder, Llama 2, Stable Diffusion 2.0, etc.

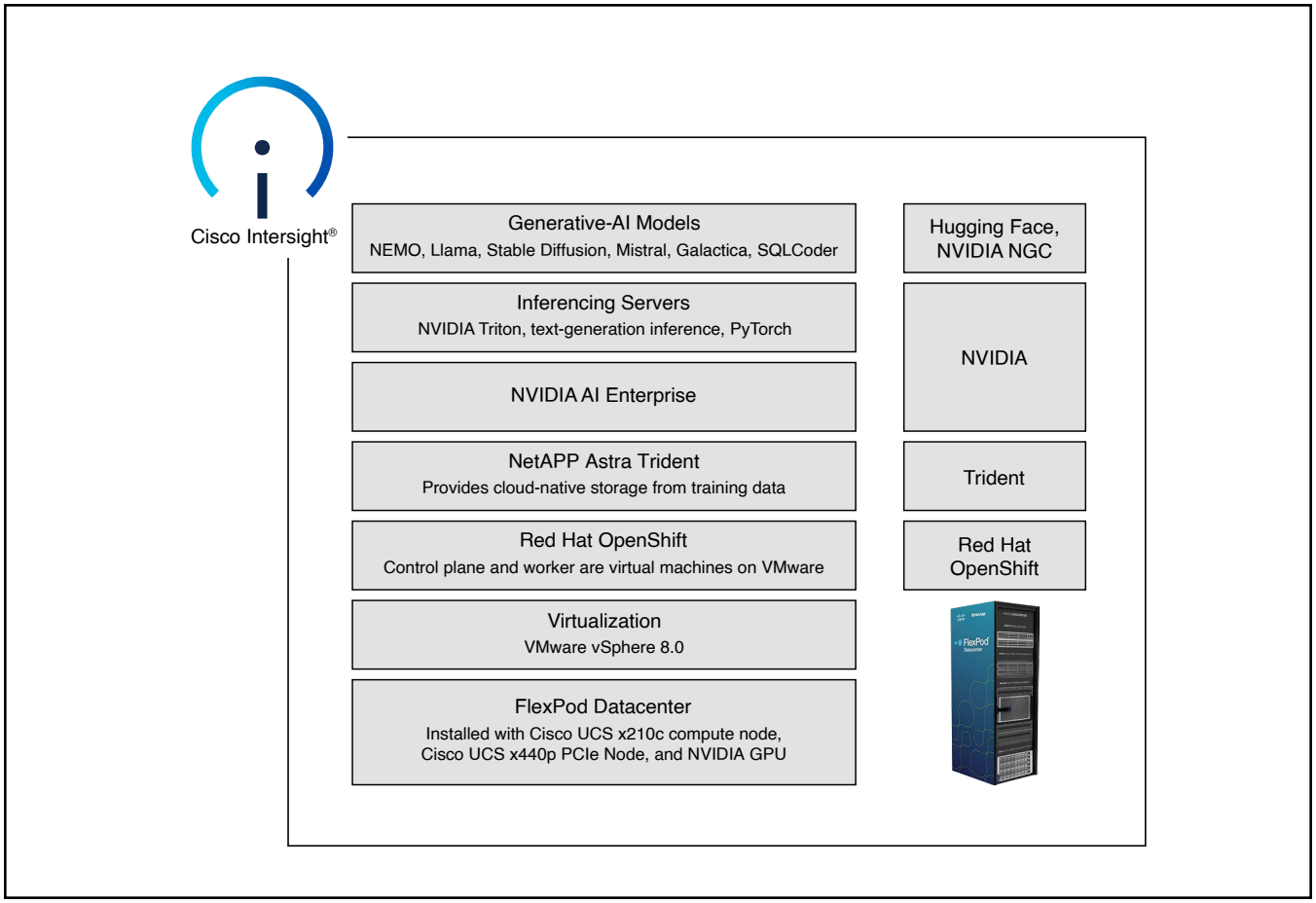

Figure 1. FlexPod for generative AI architecture

FlexPod with SUSE Rancher for AI Workloads

FlexPod Datacenter for SUSE Rancher Enterprise Container Management supports Kubernetes (K8s) workloads with high availability and server redundancy. Whether deploying SUSE Rancher Kubernetes Engine Government (RKE2) as a bare-metal cluster or virtualized on VMware vSphere or KVM, this solution operates seamlessly on Cisco UCS servers within FlexPod infrastructure.

In addition, the solution incorporates the powerful NVIDIA AI Enterprise software platform, delivering end-to-end AI capabilities securely. This platform accelerates the data-science pipeline and simplifies the development and deployment of production AI, covering diverse applications such as generative AI, computer vision, speech AI, and more. With support and testing for many models and development tools, this solution propels enterprises to the forefront of AI innovation while ensuring accessibility for every business.

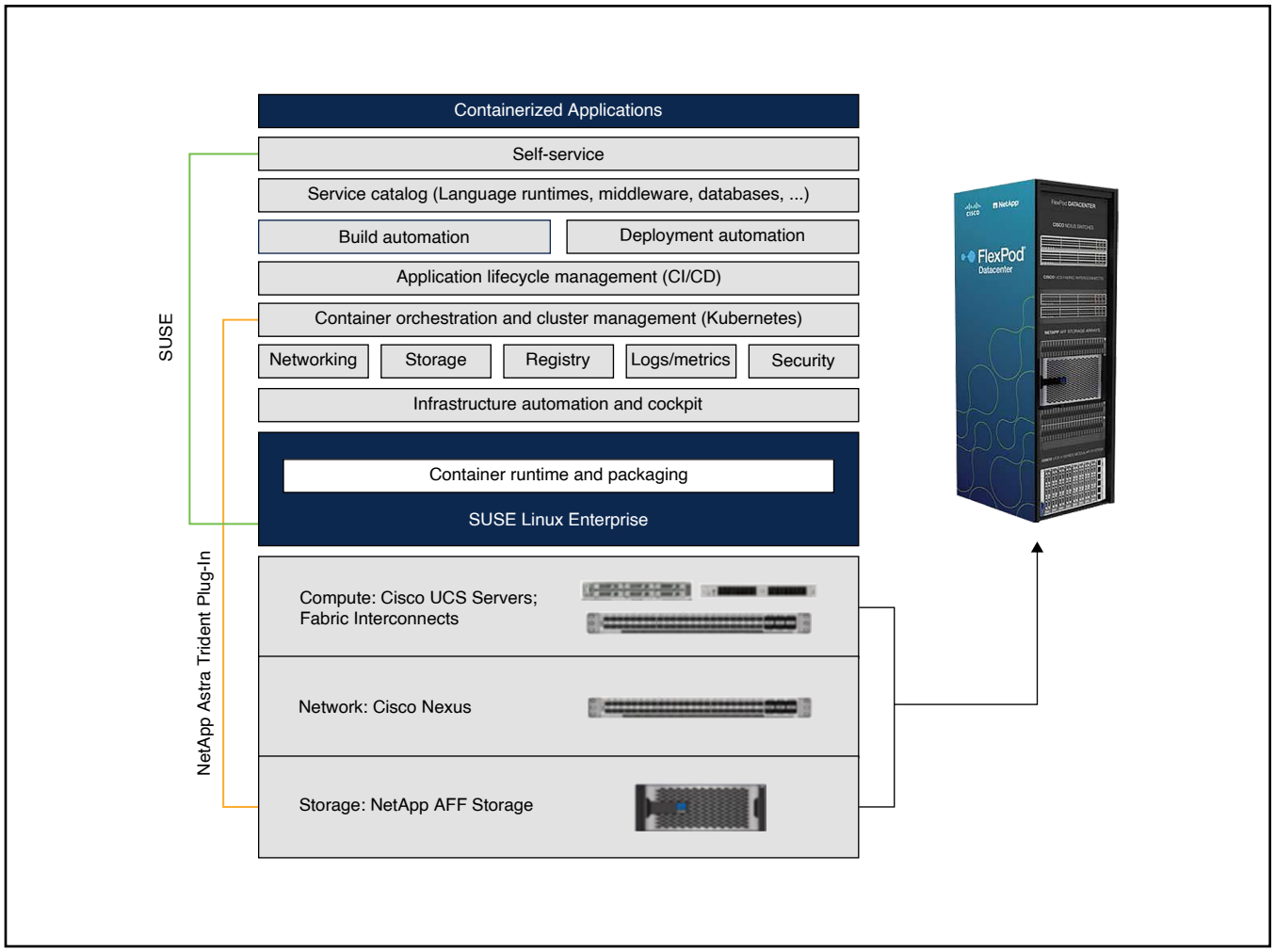

Figure 2. FlexPod with SUSE Rancher for AI workloads

Scaling FlexPod for GPU-intensive applications

This FlexPod for AI solution is carefully crafted to meet the ever-changing demands of AI workloads, providing a myriad of essential features for smooth integration and improved operational efficiency. It allows for easy deployment and management of general-purpose AI workloads, ensuring faster value realization and quick AI implementation. This solution emphasizes on securing AI infrastructure, protecting systems, data, and applications. Proven linear scalability ensures that FlexPod consistently offers peak performance across different dataset sizes.

Using this solution, you can benefit from centralized management and automation powered by Cisco Intersight®, reducing deployment times, optimizing resource utilization, and streamlining operations. FlexPod integrates seamlessly with the NVIDIA HPC-X Software Toolkit, validated to ensure optimal performance with technologies such as MPI, OpenACC, and UCX. Additionally, you can leverage NetApp tools such as the DataOps Toolkit, simplifying data management tasks for developers, data scientists, and data engineers.

This solution undergoes comprehensive testing for real-world workloads such as weather simulation, nuclear engineering and astrophysics, evaluating scalability across diverse datasets and application areas. Through rigorous testing, we provide valuable insights, comparing CPU-only performance with GPU-equipped systems to showcase the robust capabilities of FlexPod. Embrace the future of AI with FlexPod—a solution designed to elevate your AI infrastructure with reliability, efficiency, and scalability.

FlexPod with Red Hat OpenShift AI for MLOps

This FlexPod solution delivers an MLOps platform using Red Hat OpenShift AI for rapidly orchestrating and operationalizing AI models. Red Hat OpenShift AI provides an easy-to-use, integrated environment to experiment, train, and deploy 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), providing flexibility to innovate faster with an open-source approach. The solution incorporates DevOps practices for complete lifecycle management and pipeline automation, enabling you to manage multiple initiatives simultaneously, with ease, consistency, and scale. For example, with this solution, you can build a multi-step pipeline to automatically retrain and deploy an AI model as it receives new data. This continuous-update mechanism helps ensure the ongoing reliability and optimal performance of your model by adapting to ever-changing data.

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

- **Fraud detection**, such as analyzing credit-card transactions for potentially fraudulent activity.
- **Object detection**, such as detecting instances of semantic objects of a certain class (such as humans, buildings, or cars) in digital images and videos. Object detection is used in many different domains, including autonomous driving, video surveillance, and healthcare.

FlexPod AI for Retrieval-Augmented Generation (RAG)

The integration of Retrieval-Augmented Generation (RAG) with FlexPod AI reference architecture marks a transformative leap in AI infrastructure, providing unparalleled performance, scalability, and efficiency for sophisticated AI workloads. By adopting this robust combination, organizations can explore new frontiers in AI applications, fostering innovation and securing a competitive edge.

RAG is a cutting-edge approach in Natural Language Processing (NLP) that synergizes retrieval-based and generation-based methodologies. It utilizes an extensive corpus of data to retrieve pertinent information, which is then employed 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:

- Question-and-answer chatbots.
- Search augmentation.
- Knowledge engines: enable users to query their data efficiently.

This solution for FlexPod capitalizes on industry-leading components, including Cisco UCS servers, NetApp ONTAP Storage, NetApp Astra, Red Hat OpenShift, NVIDIA L40S GPU, NVIDIA AI Enterprise software, and NVIDIA Inference Microservices (NVIDIA NIM).

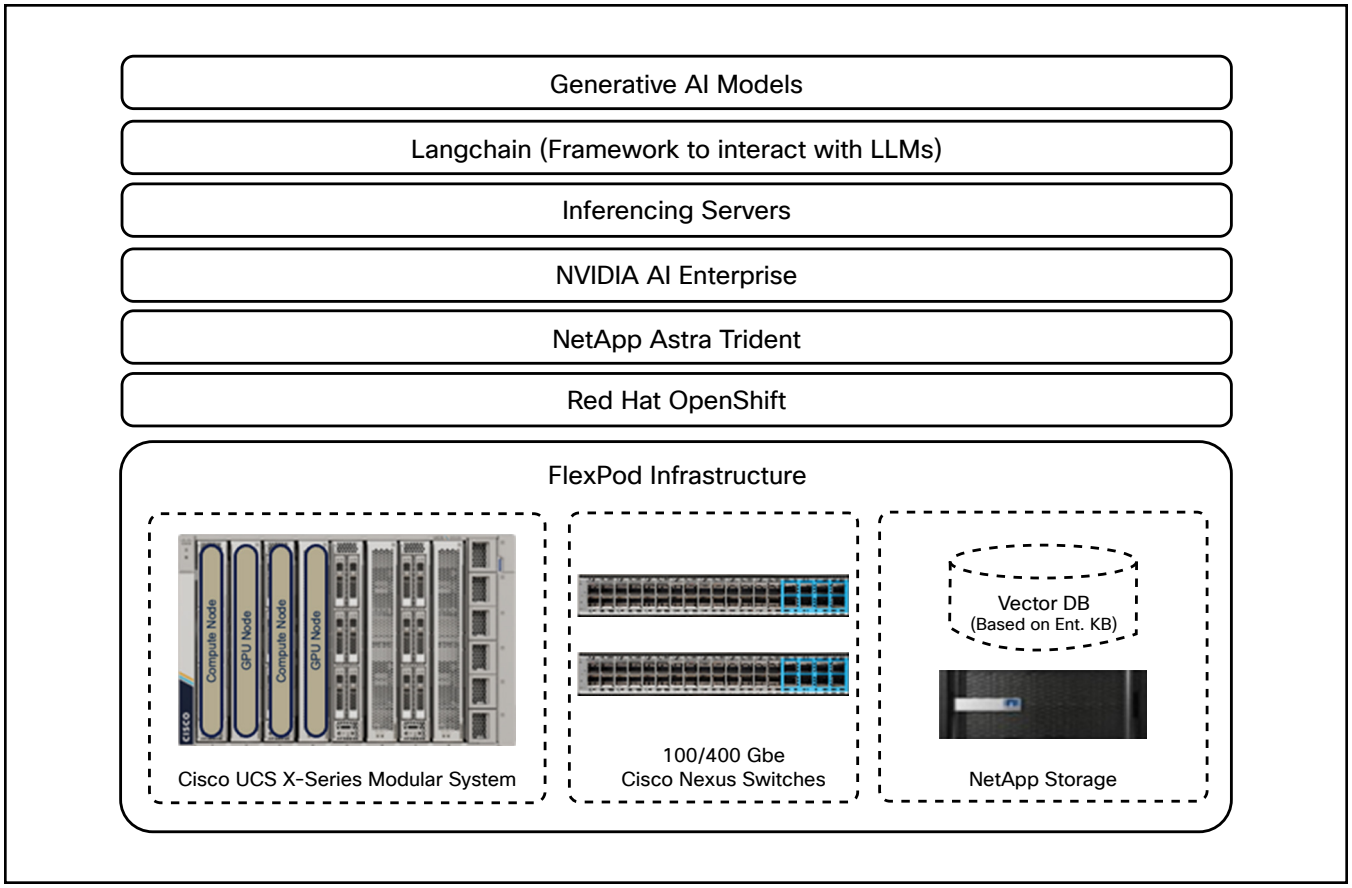

Figure 3. FlexPod for Retrieval-Augmented Generation (RAG)

FlexPod Datacenter for AI

In summary, the FlexPod Datacenter for AI is a comprehensive solution specifically designed for AI/ML environments. By seamlessly integrating GPU, compute, storage, and networking technologies into a unified platform, it empowers AI applications while concurrently enhancing operational efficiency and agility. Notably, this solution excels in reducing complexity, offering organizations a powerful and streamlined infrastructure to unlock the full potential of AI within their operations.

Learn more

- FlexPod Design Guides.
- FlexPod from Cisco and NetApp.
- NetApp – FlexPod: Where Converged Meets Cloud.

C22-4223021-01 08/24