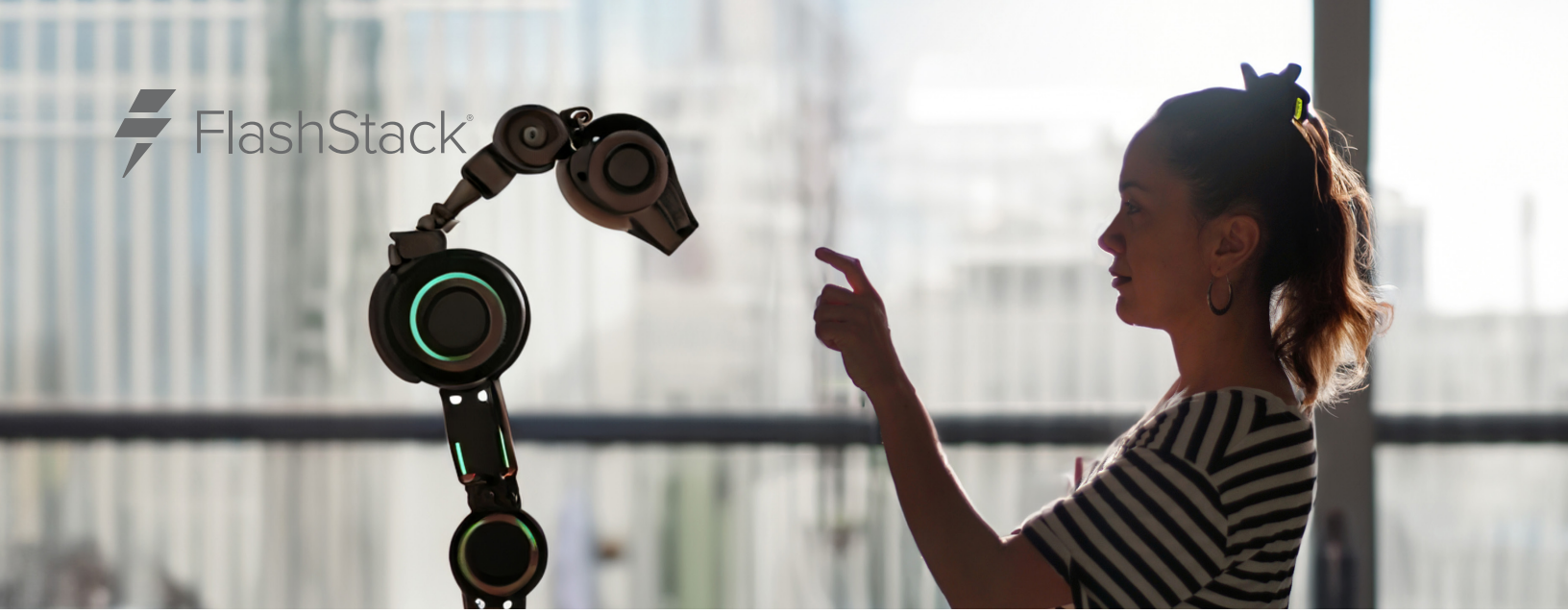


SOLUTION BRIEF

Simplify AI Infrastructure and Operations

FlashStack AI solutions for generative AI and MLOps

Imagine having the performance and capacity to bring all of your AI workloads into the same domain, helping to simplify, secure, and create a more sustainable environment. From training to inferencing, you have the support of leading AI tools and models on a platform that simplifies AI infrastructure and operations.

Highlights

- Gain an AI and MLOps platform that is simple, sustainable, and secure
- Support your entire AI/ML pipeline
- Reduce complexity
- Simplify deployment and management

FlashStack Is the Answer

This FlashStack® solution's high-density, high-performance architecture, designed to power demanding applications, is ideally suited to satisfy the need for efficient AI deployments at scale. Together with a full partner ecosystem, new AI-focused Cisco Validated Designs for FlashStack can help IT teams deploy accelerated computing, networking, and high-performance storage on a proven converged infrastructure solution. FlashStack solutions consist of of Cisco UCS® X-Series Modular systems, Cisco Nexus® switches, and Pure Storage FlashArray™ and FlashBlade™ storage systems. All are managed with the Cisco Intersight® IT operations platform.

Simple

Integrating new infrastructure for AI workloads can be a daunting task. It can require significant expenses and increased management commitments. The demanding prerequisites of vast and diverse AI workloads can strain conventional IT compute, network, and storage architectures, pushing them to their limits. FlashStack is proven infrastructure that reduces workload deployment

SEAL Sustainable Product of the Year Award for 2023

The Cisco UCS X-Series Modular System has achieved sustainable product of the year status as a system that is purpose built for a sustainable future. Cisco designed the X-Series with best-in-class energy efficiency in mind, helping balance performance needs with new sustainability demands on today's data centers. In a scenario where UCS X-Series is replacing 64 previous-generation servers:

- Customers can use 3.3x less hardware overall, saving precious rack space.
- They can also reduce almost 100,000 kilowatt hours (KwH) of energy, or the equivalent of powering 10 residential homes for a full year.
- The result is that we can help save almost 40 tons of carbon dioxide (tCO2e) emissions per year.

complexity with cloud-based operations that deliver global visibility, consistency, and control with the Cisco Intersight platform. We help you operationalize and automate deployment of AI/ML with Cisco Validated Designs to save time and reduce cost and risk. The Cisco UCS X-Series platform can support a wide range of inferencing engines to meet your needs today and scale into the future. Cisco Nexus enterprise networking delivers the high-throughput, low-latency, lossless Ethernet fabrics needed for AI/ML workloads. Pure Storage FlashBlade and FlashArray storage systems radically simplify managing the volume, velocity, and variety of data needed to create reliable models.

Sustainable

AI/ML typically requires high energy and cooling. FlashStack was redesigned to be the most energy-efficient infrastructure available, according to Enterprise Strategy Group (ESG). In fact, the Cisco UCS X-Series won the SEAL Sustainable Product of the Year 2023 Award (see sidebar). Its modular design enables you to buy what you need now, and grow as needed. This includes the ability to easily upgrade to the latest technologies by adding or swapping nodes and modules.

Secure

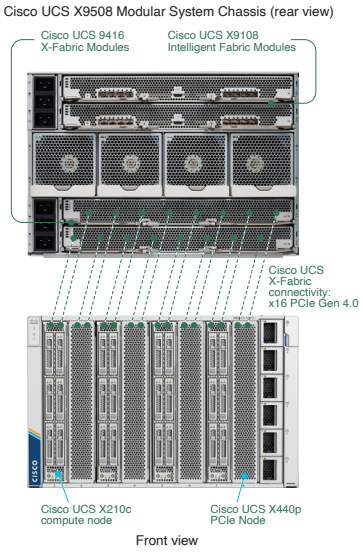
Make your AI platforms future ready and secure, to keep your infrastructure running and protected against threats (such as ransomware), with proactive, automated resiliency and security capabilities. Cisco UCS has inherent security features such as anti-tamper and anti-counterfeit protection. The Cisco Intersight platform delivers security advisories and proactive notifications, granular role-based access control (RBAC), and authorization. The FlashArray and FlashBlade storage systems come with always-on data protection through unlimited Pure Storage SafeMode™ immutable data snapshots. These layered data protection and cybersecurity features help to reduce risk.

FlashStack AI Solutions

FlashStack for AI-validated designs specifies a full stack of integrated hardware and software solutions you can use to accelerate your AI/ML efforts at scale. The FlashStack solutions provide versatile support for a wide range of AI tools and frameworks, accommodating diverse workloads while offering a choice between different GPU and CPU options. This flexibility empowers you to deploy AI applications in accordance with your specific requirements and preferences. Additionally, as AI deployments now commonly use cloud-native architectures, the solutions are built on the industry-leading container platform—Red Hat OpenShift with Portworx® by Pure Storage, the leading container storage and data-management platform. Portworx can easily be integrated into FlashStack infrastructure to simplify management of persistent storage for containerized workloads on Pure Storage FlashBlade and

Cisco UCS X-Fabric Technology Enables Up to 6 GPUs per Compute Node

Combined with up to two half-width GPU accelerators onboard each compute node, you can configure up to four additional half-width or two full-width GPUs using a Cisco UCS X440p PCIe Node, as shown below.



Cisco UCS X-Fabric Technology

- Connects compute nodes to PCIe nodes using PCIe Gen 4 connectivity
- Brings GPU acceleration to AI/ML applications
- No backplane or cables means easy upgrades to new GPU technologies

FlashArray systems. Furthermore, automated deployment of the entire infrastructure with Ansible playbooks can help accelerate your deployment and reduce the time to value.

Generative AI Inferencing

FlashStack enables enterprises to quickly design and deploy a generative AI inferencing solution (Figure 1). Our comprehensive testing and validation address the complexities of deploying and serving generative AI models in production. The solution covers a spectrum of inferencing servers such as NVIDIA Triton, Hugging Face Text Generation Inference, and PyTorch, giving you the flexibility to choose a model-serving option based on your needs and objectives. Notably, we tested model deployment in the NVIDIA Triton Inference server with TensorRT-LLM, which is a new library for compiling and optimizing large-language models (LLMs) for inference. Together, TensorRT-LLM and Triton Inference Server provide a toolkit for optimizing, deploying, and running LLMs efficiently. Additionally, we validated models including Stable Diffusion, Llama2, NeMo GPT, BLOOM, Mistral, Galactica, and SQLCoder. This helps you confidently operationalize commonly used AI models to accelerate time to value. These models demonstrate how Generative AI can transform a variety of use cases such as:

- **Natural language generation**—text-generation tasks such as creating blogs, articles, and reports; dialogue generation, summarization, and content creation for marketing and advertising purposes.

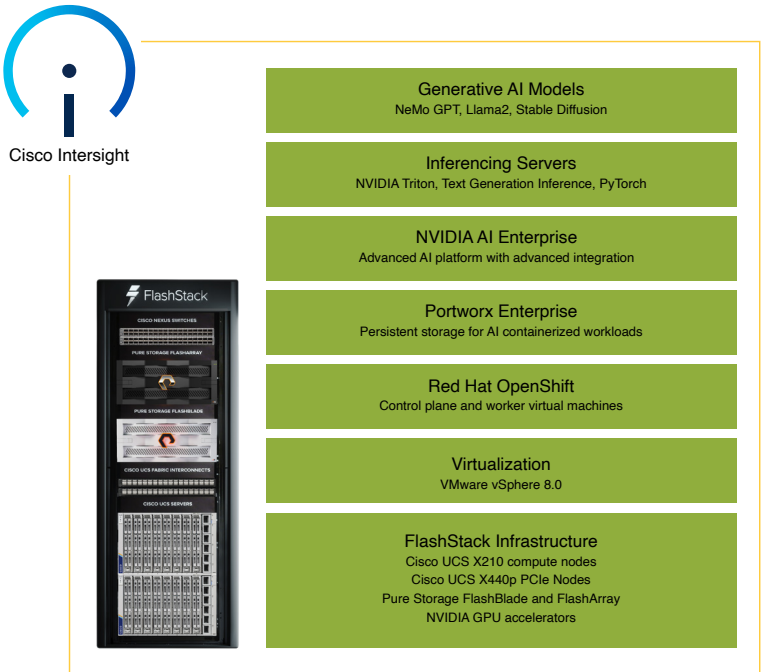


Figure 1. Architecture for Generative AI Inferencing

- **Chatbots and virtual assistants**—conversational agents, chatbots, and virtual assistants that generate natural-language responses based on user queries or instructions
- **Create digital art**—paintings and illustrations; compose original melodies or entire musical compositions
- **Personalized recommendations**—for products, movies, music, or content based on user preferences, behavior, and historical data
- **Code generation**—autocomplete code by suggesting or completing code snippets for programmers, or generate new code based on high-level descriptions or requirements
- **Data augmentation**—synthesize data samples to augment existing datasets, increasing the diversity and size of training data sets for machine-learning models

MLOps with Red Hat OpenShift AI

This FlashStack solution (Figure 2) 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, Intel OpenVino, 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. We validated models including Stable Diffusion, YOLOv8, and Keras, which can help you 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.

FlashStack®

SOLUTION BRIEF

IDC Business Value Survey of FlashStack Production Users

446%
5-year return on investment

72%
reduced total cost of operations

8 months
payback on investment.

Source: IDC document #US47408621. All IDC research is © 2021 by IDC. All rights reserved. All IDC materials are licensed with IDC's permission and in no way does the use or publication of IDC research indicate IDC's endorsement of Pure Storage's or Cisco's products or strategies.

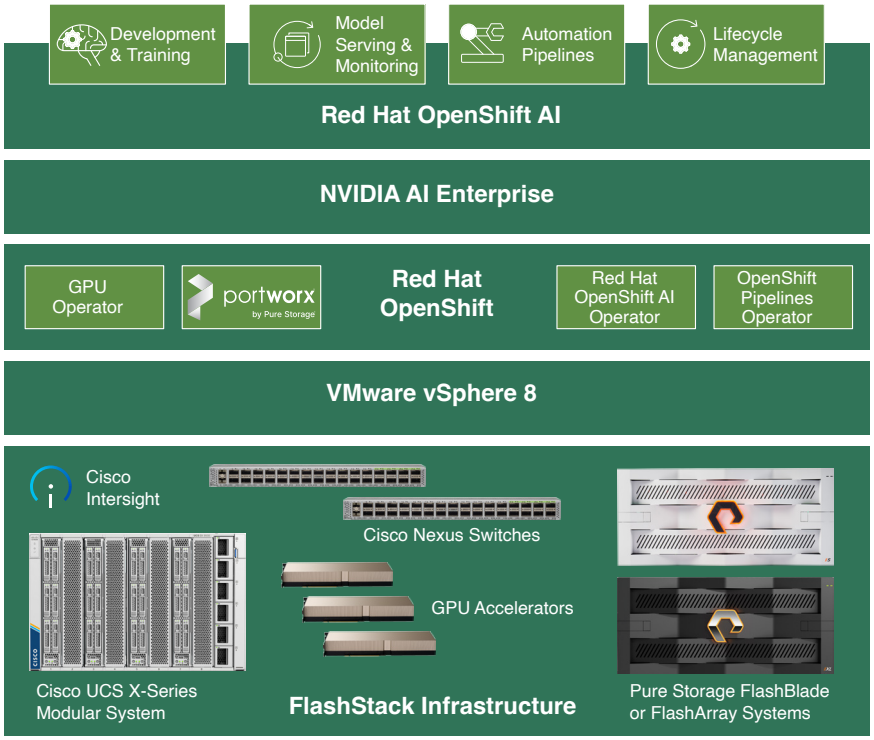


Figure 2. Architecture for MLOps using Red Hat OpenShift AI

Why Flashstack for AI?

FlashStack for AI delivers radical simplicity, unparalleled sustainability, and defense-in-depth security, and is made up of world-class compute, storage, and networking infrastructure. Powered by Cisco Intersight and Ansible automation playbooks, FlashStack delivers an AI-ready, prevalidated platform to get you up and running quickly and performing with peak efficiency at scale.

Learn More

- FlashStack for AI
- FlashStack for Generative AI Inferencing
- FlashStack for AI: MLOps using Red Hat OpenShift AI

flashstack@purestorage.com | www.cisco.com/go/flashstack | www.flashstack.com

CISCO PURESTORAGE®