Build for scalable AI with Red Hat and Cisco

According to IDC, AI is now a strategic necessity for competitive differentiation and long-term success.[1]

Confront the challenges of generative AI

The adoption of artificial intelligence (AI) is accelerating, providing organizations that use it with many important benefits.

Advances in traditional IT systems, including the incorporation of graphics processing units (GPUs) have made it possible to effectively run predictive AI workloads.

However, this acceleration is also exposing unforeseen challenges for AI adopters. The advent of generative AI (gen AI) has put an even greater strain on IT systems and software requirements. The data volume, velocity, and variability of gen AI at scale are all unmanageable for traditional infrastructure, resulting in fragmented application stacks and lock-in issues.

Unfamiliar, constantly-evolving gen AI software can also increase operational complexity, creating skills gaps and disconnect between teams. Managing AI-enabled application workloads can become more difficult as a result.

Use of AI can also pose security and data compliance hurdles for organizations to overcome. It is important to safeguard applications and data from emerging threats and model bias, and maintain compliance with ethics and privacy regulations.

In order to counter these challenges, a robust, comprehensive solution for gen AI and machine learning (ML) is required.

IDC predicts that by 2026, **75%** of large enterprises will rely on AI-infused processes and worldwide spending on AI-centric systems will pass **US$300 billion**.[2]

Simplify AI/ML infrastructure and operations with Cisco and Red Hat

Cisco is a leader in data infrastructure with a history of providing successful technology solutions for enterprise datacenters. By combining Cisco's hardware and software architectures with open source solutions from Red Hat, organizations are better suited to overcome the challenges of predictive and generative AI/ML.

Together, Cisco and Red Hat provide a proven and validated full-stack architecture to streamline, automate, and scale operations for AI/ML, helping organizations to:

- Train, tune, serve, monitor, and manage AI/ML models.
- Build an infrastructure for AI that is scalable and sustainable into the future.
- Reduce complexity and keep infrastructure and operations connected.
- Simplify development, deployment, and management across hybrid cloud environments.

f facebook.com/redhatinc
𝕏 twitter.com/RedHat
in linkedin.com/company/red-hat

[1] IDC Perspective. "AI Partner Ecosystem and Roles." Document # US51966624, April 2024. (Registration required)
[2] IDC Study. "IDC FutureScape: Worldwide Artificial Intelligence and Automation 2023 Predictions." Document # US49748122, Oct. 2022. (Registration required)

Why Red Hat and Cisco?

- Validated architectures tested by both Cisco and Red Hat.
- Automated Red Hat OpenShift deployment.
- Cisco UCS stateless configuration management.
- Cisco Solution Support for a single point of contact.

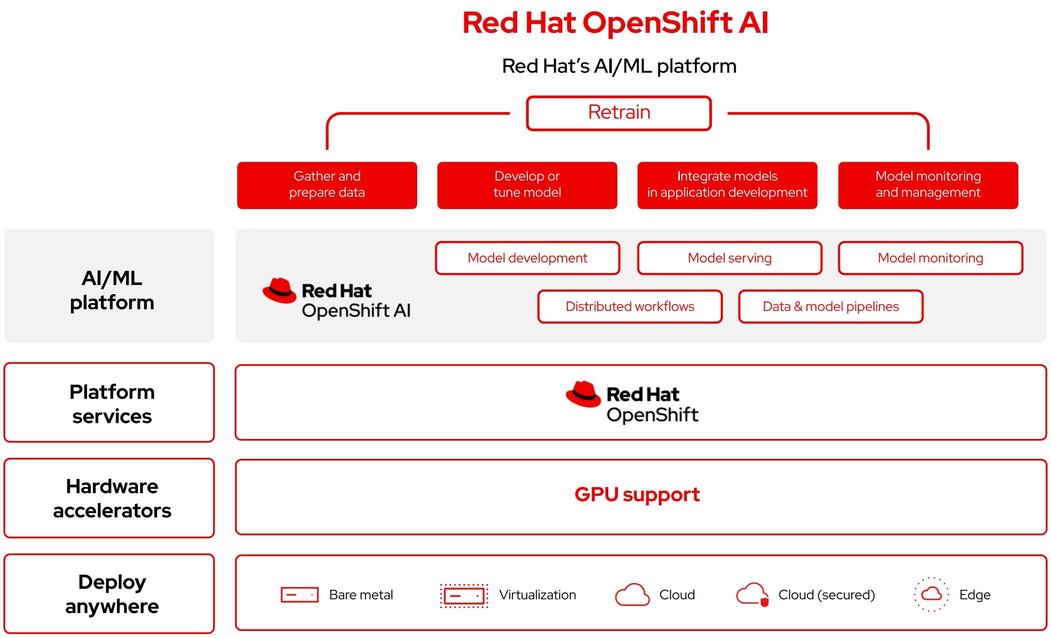

Figure 1. Red Hat OpenShift AI is a flexible, scalable AI/ML platform that allows enterprises to create and deliver AI-enabled applications at scale across hybrid cloud environments.

Discover joint solutions from Red Hat and Cisco

Cisco Validated Designs

CVD solutions reduce the complexity of AI/ML deployments by providing validated solutions that save time and lower risk with expert guidance.

Cisco Validated Designs (CVDs) are tested, documented blueprints that can help organizations successfully create, deploy, and scale new workloads and technologies. These guides provide validated reference architecture that organizations can use as a starting point for their rollouts.

Organizations can use CVDs to mitigate risk with proven, tested architectures for standardized, repeatable deployments. CVDs address common use cases and enterprise IT priorities to ensure that different products and technologies integrate and work together effectively and efficiently. CVD validation ensures predictable and reliable enterprise deployments.

Using CVDs together with automated playbooks for AI infrastructure can help organizations simplify and accelerate deployment.

Use CVDs to architect a solution powered by Red Hat OpenShift AI

Red Hat® OpenShift® AI is a flexible, scalable AI and ML platform that helps enterprises to create and deliver AI-powered applications at scale across hybrid cloud environments.

Built using open source technologies, Red Hat OpenShift AI provides trusted, operationally consistent capabilities for teams to experiment, serve models, and deliver innovative applications.

Discover CVDs for AI/ML that incorporate Red Hat OpenShift or Red Hat OpenShift AI into their architecture, including:

- FlashStack for AI: MLOps using Red Hat OpenShift AI.
- UCS X-Series with 5th Gen Intel Xeon Scalable Processors on Red Hat OpenShift AI.
- FlashStack for generative AI inferencing design guide.
- FlexPod datacenter with generative AI inferencing.

Cisco Intersight, Red Hat OpenShift, and Red Hat Ansible Automation Platform

Cisco Intersight is an IT operations platform that provides a guided and automated process for system setup and configuration anywhere at scale. It includes back-end integration with Cisco's Hardware Compatibility List and other operational tools for proactive return material authorization (RMA), uploading of log files, and other features for simplifying server management at scale.

The integration of Cisco Intersight and Red Hat OpenShift allows for automated Cisco Unified Computing System (UCS) bare-metal configuration, provisioning, and installation. This helps organizations simplify management and maintain servers anywhere, from datacenters to edge locations worldwide. UCS is an integrated computing infrastructure that brings together compute, networking, and storage in a single system to power your applications.

With Red Hat Ansible® Automation Platform, organizations also gain access to Red Hat Ansible Certified Content Collections for Cisco Intersight, and Red Hat OpenShift solutions. These Ansible Content Collections, managed and supported by Cisco and Red Hat, can be used to automate the deployment of CVDs, and modules and roles are provided for common operational tasks.

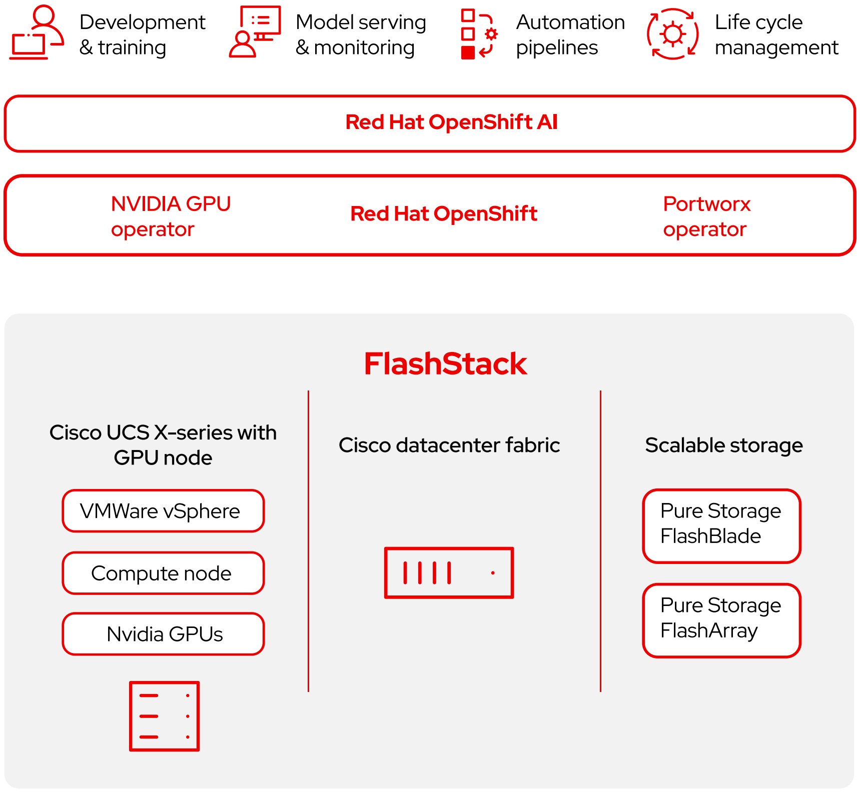

Figure 2. A high-level architecture of the FlashStack AI MLOps solution using Red Hat OpenShift AI.

Cisco Solution Support

Cisco Solution Support provides a single point of contact for technical support, across Red Hat and Cisco infrastructure stacks.

With support from experts who focus on hardware and software from Cisco and its partners, including Red Hat, organizations can benefit from multiproduct and multivendor issue resolution across their deployment environment.

Simplify AI with help from Red Hat and Cisco

Together, Red Hat and Cisco provide a blueprint for simplified delivery of predicative and generative AI/ML models. With Cisco's full-stack architecture working in tandem with Red Hat's powerful, open source containerization and automation solutions, organizations can simplify, optimize, and scale their operations for AI/ML and gen AI adoption and build for the future of AI.

Learn more about Red Hat and Cisco

Discover Red Hat's portfolio of open source AI solutions.

Find out more about Red Hat's partnership with Cisco.

> *"Cisco and Red Hat are working together to help organizations realize the value of AI through improved operational efficiencies, increased productivity, and faster time to market."*
>
> **Steven Huels**
> General Manager,
> Artificial Intelligence
> Business, Red Hat.[3]

> *"By integrating Cisco UCS and Intersight with Red Hat OpenShift, customers can benefit from enhanced performance for demanding AI workloads backed by a reliable containerized environment for simplified application development and management."*
>
> **Tushar Katarki**
> Senior Director, Product
> Management, Hybrid Cloud
> Platforms, Red Hat.[4]

[3] Foster, Jeremy. "Time to simplify: A fresh look at infrastructure and operations for artificial intelligence." Cisco.com, 7 Nov. 2023.
[4] Brannon, Todd. "Operational innovations for AI and cloud-native workloads from Cisco and Red Hat." Cisco.com, 2 May 2024.

About Red Hat

Red Hat helps customers standardize across environments, develop cloud-native applications, and integrate, automate, secure, and manage complex environments with award-winning support, training, and consulting services.

North America	Europe, Middle East, and Africa	Asia Pacific	Latin America
1 888 REDHAT1 www.redhat.com	00800 7334 2835 europe@redhat.com	+65 6490 4200 apac@redhat.com	+54 11 4329 7300 info-latam@redhat.com

facebook.com/redhatinc
twitter.com/RedHat
linkedin.com/company/red-hat