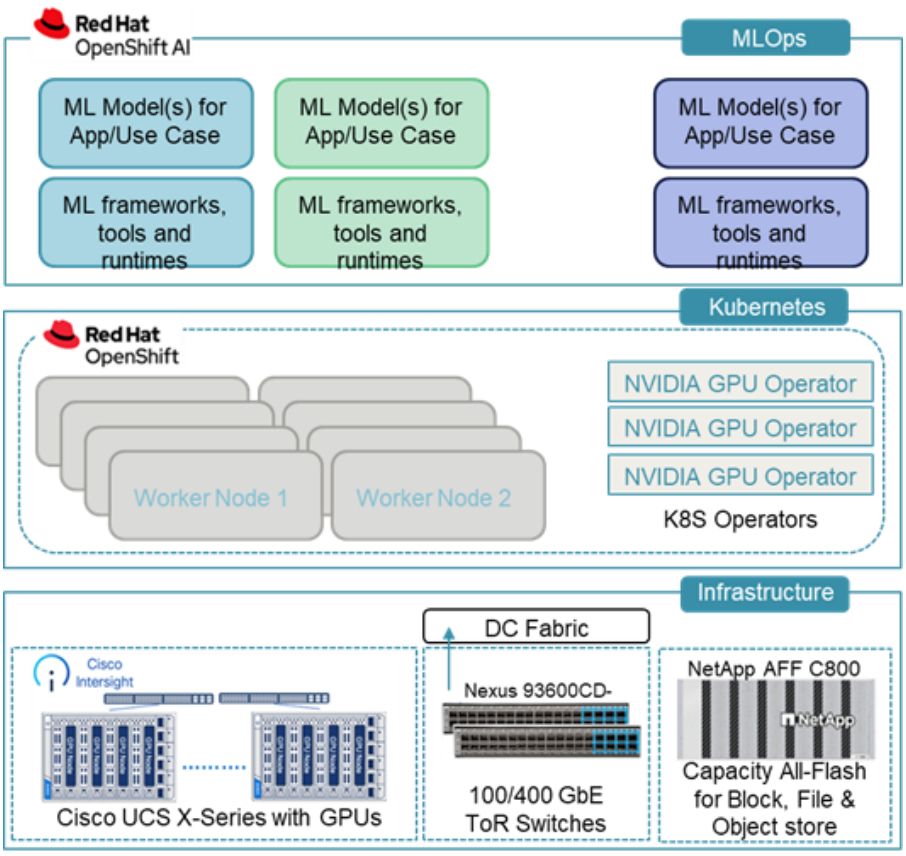
FlexPod®

FlexPod AI with MLOps using Red Hat OpenShift AI Solution Brief

FlexPod AI: A complete infrastructure stack with MLOps

Generative AI is one of the fastest growing markets in the technology industry. Cisco and NetApp have spent years optimizing FlexPod® designs to meet AI challenges. The solution described in the Cisco Validated Design (CVD) guide delivers a complete infrastructure stack with machine learning operations (MLOps) that enterprises can deploy to efficiently manage, accelerate, and scale multiple artificial and machine learning (AI/ML) initiatives from incubation to production. The solution uses Red Hat OpenShift AI as the MLOps platform running on FlexPod with a bare metal infrastructure. This FlexPod infrastructure includes Cisco UCS X-Series compute with X440p PCIe nodes, NetApp® storage, and Cisco Nexus switching with Red Hat OpenShift to support Enterprise AI/ML initiatives at scale.



Machine learning models

MLOps are a set of best practices to streamline and accelerate the delivery of machine learning models. The delivery of these ML models for production use or model serving is key to operationalizing AI so that enterprises can build and deliver ML-enabled applications.

Automation with MLOps

Automation is integral to MLOps to accelerate efforts that minimize technical debt, enabling enterprises to deliver and maintain models at scale. MLOps pipelines also need to

continuously retrain models to keep up with ever-changing data to ensure optimal model performance. MLOps bring consistency and efficiency to the model delivery process.

Red Hat OpenShift AI

Red Hat OpenShift AI serves as the MLOps platform to streamline, scale, and accelerate model delivery. It provides the development environment, tools, and frameworks that data scientists and machine learning teams need to build, deploy, and maintain AI/ML models in production. OpenShift AI streamlines the ML model delivery process from development to production deployment (model serving) with efficient lifecycle management and pipeline automation. From the OpenShift AI console, AI teams can select from a preintegrated, Red Hat-supported set of tools and technologies or custom components that are enterprise managed, providing the flexibility that teams need to innovate and operate efficiently.

Other key features of OpenShift AI include:

- **Collaborative workspaces.** OpenShift offers a collaborative workspace where teams can work together on one or more models in parallel.
- **Development environments.** ML teams can use Jupyter notebooks as a service using prebuilt images, common Python libraries, and open-source technologies such as TensorFlow and PyTorch to work on their models. Also, administrators can add customized environments for specific dependencies or for additional integrated development environments such as RStudio and VS Code.
- **Model serving at scale.** Multiple models can be served for integration into intelligent AI-enabled applications by using inferencing servers (for example, Intel OpenVINO, NVIDIA Triton) using GPU or CPU resources provided by the underlying OpenShift cluster without writing a custom API server. These models can be rebuilt, redeployed, and monitored by making changes to the source notebook.
- **Enhanced model serving** allows users to run predictive and GenAI on a single platform for multiple use cases, reducing costs and simplifying operations. Enhanced model serving enables out-of-the-box model serving for large language models and simplifies the surrounding user workflow.
- **Data science pipelines for GUI-based automation using OpenShift pipelines.** OpenShift AI leverages OpenShift pipelines to automate ML workflow by using an easy to drag-and-drop web UI as well as code-driven development of pipelines by using a Python SDK.
- **Model monitoring visualizations** for performance and operational metrics improve observability into how AI models are performing.
- **New accelerator profiles** enable administrators to configure different types of hardware accelerators available for model development and model-serving workflows, providing simple, self-service user access to the appropriate accelerator type for a specific workload.

To learn more, check out the FlexPod AI with MLOps using Red Hat OpenShift CVD in it's entirety.