The bridge to possible

Building an Enterprise-Class AI/ML Infrastructure for MLOps

Using Cisco UCS, NVIDIA GPUs, and Red Hat OpenShift AI

Archana Sharma, Technical Marketing Engineer

BRKCOM-2018

Cisco Webex App

Questions?

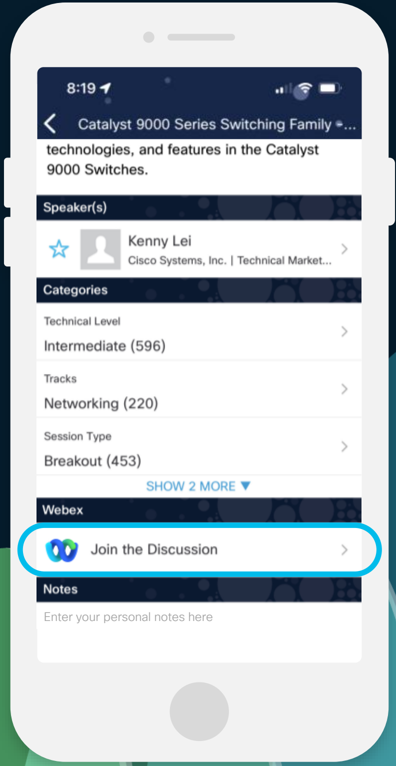
Use Cisco Webex App to chat
with the speaker after the session

How

1. Find this session in the Cisco Live Mobile App
2. Click “Join the Discussion”
3. Install the Webex App or go directly to the Webex space
4. Enter messages/questions in the Webex space

Webex spaces will be moderated by the speaker until June 7, 2024.

https://ciscolive.ciscoevents.com/
ciscolivebot/#BRKCOM-2018



Agenda

- MLOps
- Infrastructure Considerations
- Building AI/ML Infrastructure
- Demo (offline)
- Wrap-up

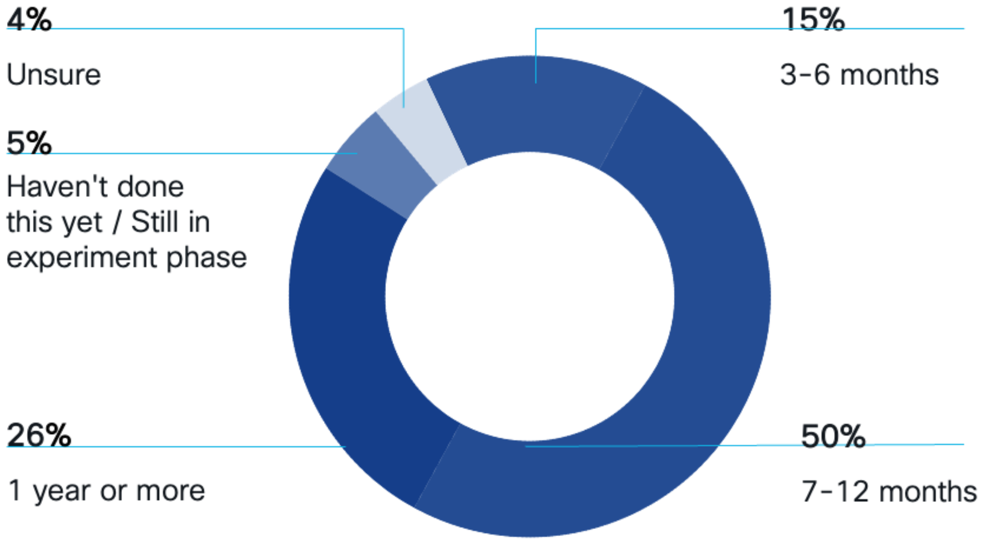
MLOps

Why MLOps?

Operationalizing AI is challenging

What is the average AI/ML timeline from idea to operationalizing the model?

Half of respondents (50%) say their average AI/ML timeline from idea to operationalizing the model is 7-12 months.



Gartner estimates, on average, 54% of AI projects make it from pilot to production

Source: Gartner Peer Insights, Open Source AI for Enterprise survey, 2023
Source: https://www.gartner.com/en/newsroom/press-releases/2022-08-22-gartner-survey-reveals-80-percent-of-executives-think-automation-can-be-applied-to-any-business-decision

Model Delivery Lifecycle

Streamline and scale using MLOps

- Scalability
- Governance
- Efficiency
- Reliability
- Adaptability

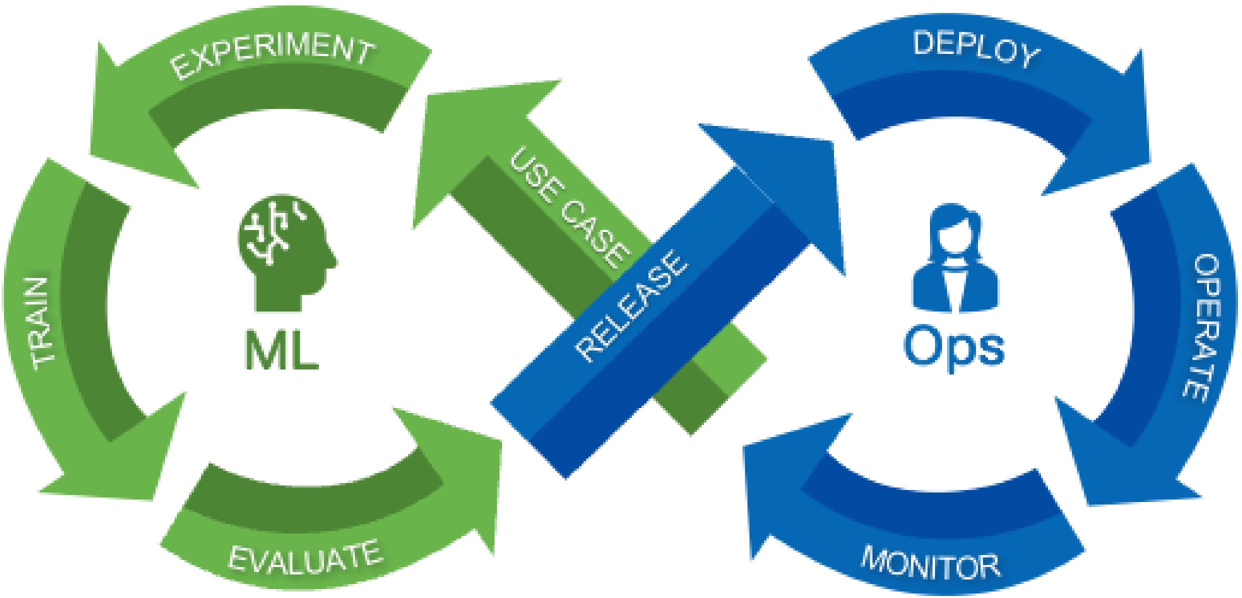
Iterate

Prepare Data	Experiment/Tune model	Serve and integrate with App	Monitor/Maintain model
Gathering & preparing data for AI	Apply scientific rigor to understand data and build/customize model	Model available for production inferencing	Track model quality, metrics and drift

Pace of AI/ML technology shifts require a strong foundation to adapt

What is MLOps?

Foundation for success



- Automation
- Version Control
- CI/CD
- Collaboration
- Accelerate
- Simplify

Infrastructure Considerations

Complementary Pillars of AI

Predictive AI

- Uses historical data to make statistical predictions on future outcomes
- Range of techniques from predictive analytics to ML and DL algorithms
- Fraud detection, risk assessment, anomaly detection, forecasting, recommendation systems, customer behavior prediction
- Delivering value today…and indispensable for organizations
- ~100M parameter range

Generative AI

- Generalizes patterns seen before to predict and generate multimodal content (ChatGPT, DALL-E)
- Transformative with unparalleled potential
- Popular model categories: Transformer models (GPT, BERT) and Stable Diffusion
- Large Language Models (LLMs) are significantly larger and resource-intensive than other ML models
- ~1B+ parameter range

Virtual Assistants

Chatbots

Use Cases

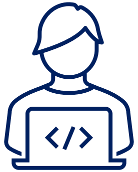
Range of applications and verticals

Application Use Cases

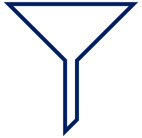
Marketing | Train/Teach | Customer Support | Cybersecurity

Legal | Drug Development | Medicine | Transportation

LLM Use Cases



Code Generation



Text Summarization

Question/ Answering

Content Generation

Text Translation

Speech Recognition

Large Language Models (LLMs)

Limitations for enterprise use

Hallucination	Can make stuff up, always has an answer
Sources	Where did the information come from ?
Outdated	Models maybe stale as quickly as it is released
Customize	Cannot personalize or use more current data
Update	Cannot edit the model to remove/change data

Training LLMs

Resource-Intensive and costly

Large Language Models are...

Pre-trained on a large corpus of publicly available unlabeled data

Training takes 1000s of GPUs over a span of months

Requires periodic re-training to stay up to date

GPT-3 Large – 175B parameters

- Training Set Tokens: 300B
- Vocabulary Size: ~50k
- Number of GPUs: 10k x V100
- Training Time: One Month

Llama – 65B parameters

- Training Set Tokens: ~1-1.3T
- Vocabulary Size: ~32k
- Number of GPUs: 2048 x A100
- Training Time: 21 Days

Building LLMs from scratch is cost-prohibitive for the average Enterprise

Use Foundational Models

Starting point for most Enterprises

Foundational models (FM)

Download

BERT
GPT
Llama
Mistral AI
Stable Diffusion
Cohere
Claude
BLOOM
...

Pre-trained, general-purpose models

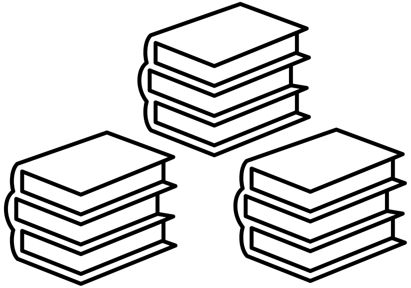
Customize or integrate directly for inferencing in enterprise applications

Model Size

- LLMs <100B
- Other Generative <1B
- Predictive <100M

LLMs lack domain knowledge

Limitations for enterprise use



Massive amount of general knowledge based on patterns seen during training

LLMs have broad knowledge but lack domain-specific knowledge

Customizing LLMs

To address LLM limitations

Adaptation Techniques

Fine-Tuning	Parameter Efficient Tuning

Changes what the model knows

Prompt Engineering	Retrieval Augmented Generation

Provides context as input to the model @Inference

Fine-tuning

- Adapts model for specific-tasks
- Updates model parameters
- Smaller dataset, less resources

Parameter-Efficient Fine Tuning

- Fine-tunes a subset of the model parameters
- Examples: LoRA, Prefix Tuning

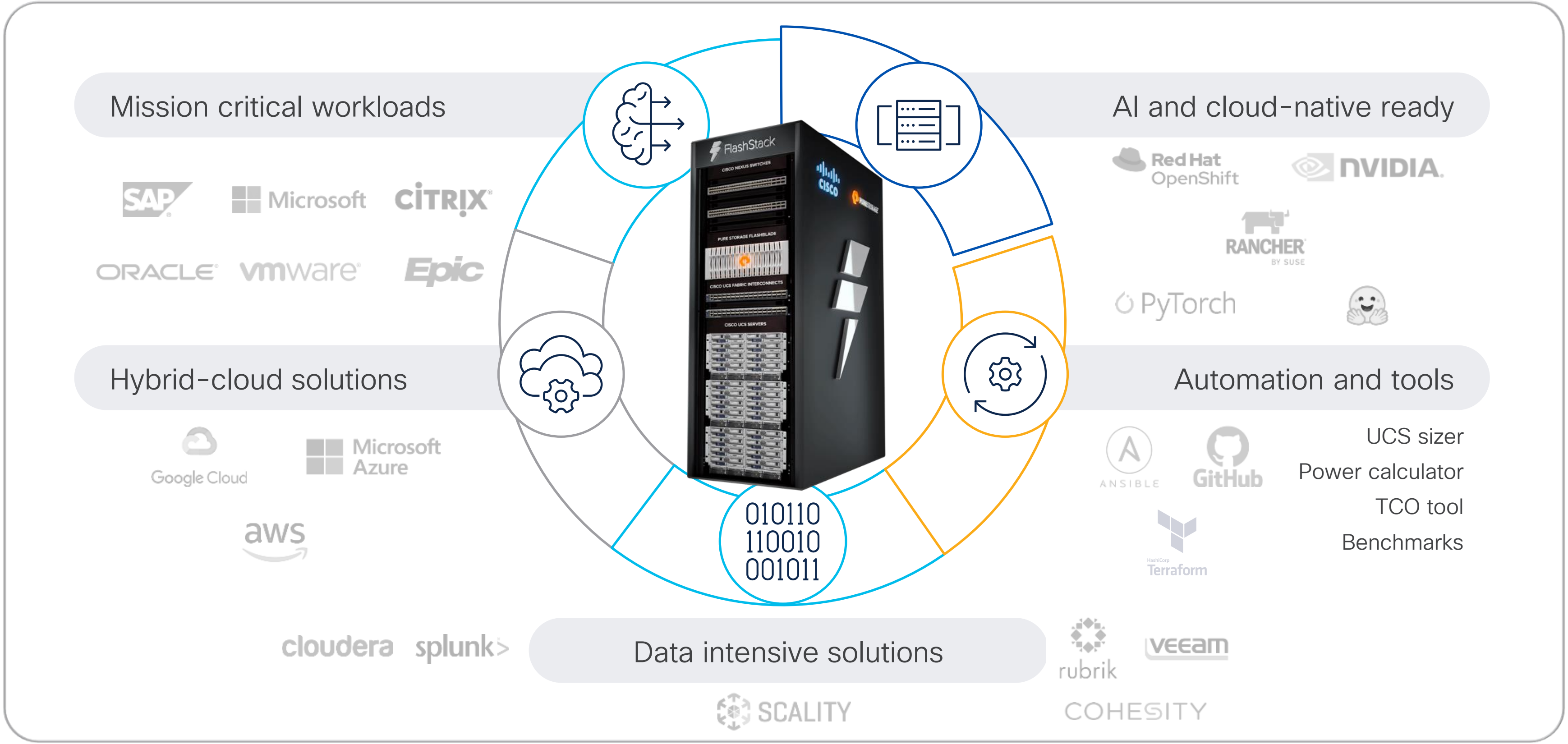
RAG, Prompt Engineering (PE)

- RAG: Add external data sources to LLM inferencing as context
- PE: Uses prompts for better output

Building an Enterprise-class AI/ML Infrastructure

Cisco Solution Portfolio

Full Stack Solutions delivering best-in-class value to our customers



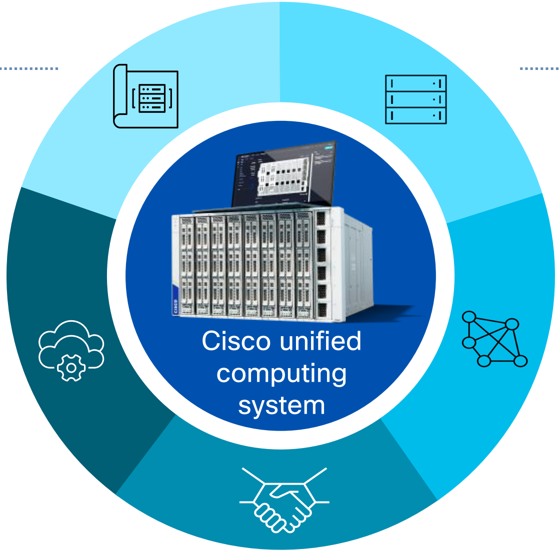
Cisco Validated Designs (CVD)

Accelerate

Ready to 'Go' solutions for faster time to value

Less risk

Reduce risk with tested architectures for standardized, repeatable deployments



Expert Guidance

CVDs provide everything from system designs to implementation guides, and ansible automation

Cisco TAC support

Single point of contact for solution. Cisco will coordinate with partners as needed to resolve issues

CVDs for AI/ML Infrastructure

1 Cisco Validated Designs for Simplified AI Infrastructure

- NVIDIA — NVIDIA AI Enterprise
- Red Hat — Red Hat OpenShift AI
- intel
- NUTANIX — GPT-in-a-box on Nutanix Hyperconverged
- CLOUDERA — Gen-AI with Cloudera Data Platform

AMD · FlashStack · FlexPod · NUTANIX

intel · NVIDIA NGC · intel Developer Cloud

2 Curated playbooks to automate base infra deployments

- Cisco Intersight
- ANSIBLE

Key Completed In progress Planned

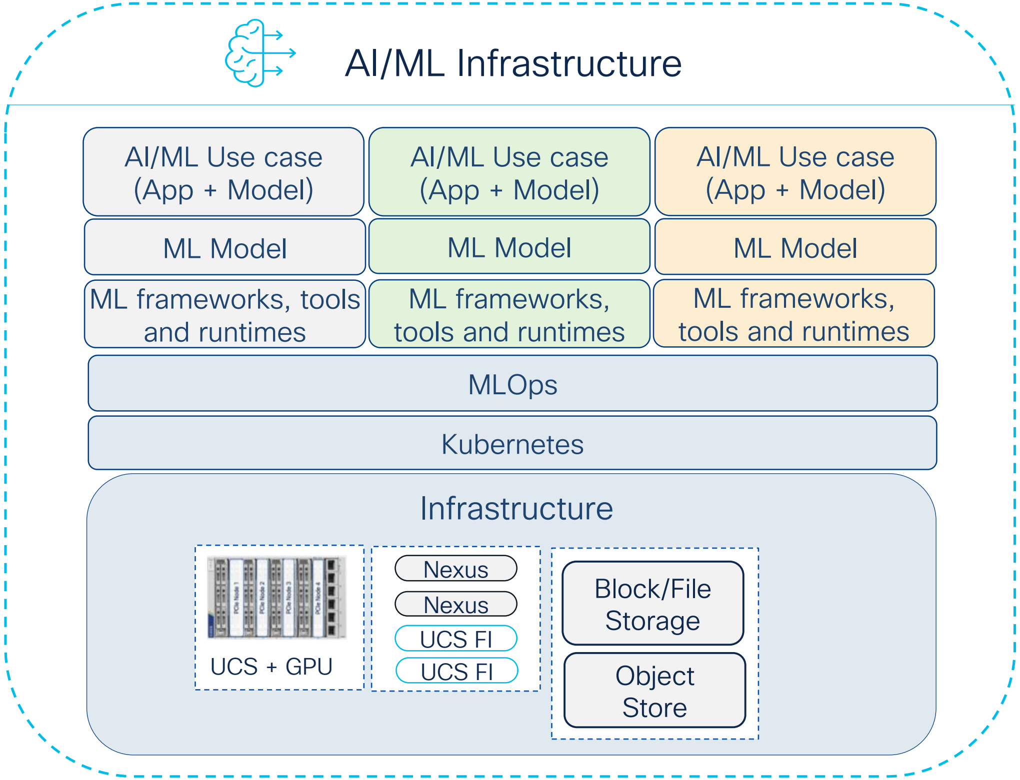
AI Solution Roadmap

CISCO VALIDATED DESIGN	DESCRIPTION	AVAILABILITY
Scaling FlexPod for GPU intensive Apps	Sizing guide for AI infrastructure leveraging real-life model simulations	
FlexPod with SUSE Rancher for AI Workloads	Foundational architecture for general-purpose AI deployments	
FlashStack with Red Hat OpenShift and NVIDIA AI Enterprise	Blueprint for deployment of Generative AI models for inferencing along with performance metrics	
FlexPod with Red Hat OpenShift and NVIDIA AI Enterprise		
FlashStack for MLOps using Red Hat OpenShift AI	Architecture to operationalize end-to-end AI workflow i.e., data prep, train, test & deploy, using Red Hat OpenShift AI	
FlexPod for MLOps using Red Hat OpenShift AI		Q3 CY24
Cisco UCS and Red Hat OpenShift AI with Intel AI Enterprise Platform		
AI Solution for the Enterprise with Cloudera Data Platform	Integrated architecture for AI combining data lake, compute farm & storage services	Q2 CY24
Retrieval-Augmented Generation (RAG) with Cisco Converged Infrastructures	Framework for enterprise-specific knowledge augmentation in LMMs	Q3 CY24
Intel AI Enterprise with Cisco Converged Infrastructures	AI deployment guide with Intel GPUs and Intel AI inferencing software suite	Q3 CY24
Generative Pre-trained Transformers (GPT) with Nutanix	AI-ready HCI architecture to fine-tune and deploy LLMs	Q2 CY24
Edge Inferencing Solution on UCS Edge Platform	Blueprint for deployment of AI models for inferencing in edge environments	TBD
Secure by Design – Confidential AI with FlexPod	Zero-trust framework for AI deployments	TBD

AI/ML Infrastructure

High-level Architecture

Generative AI and Predictive AI Use Cases



ML Infrastructure Design – Compute

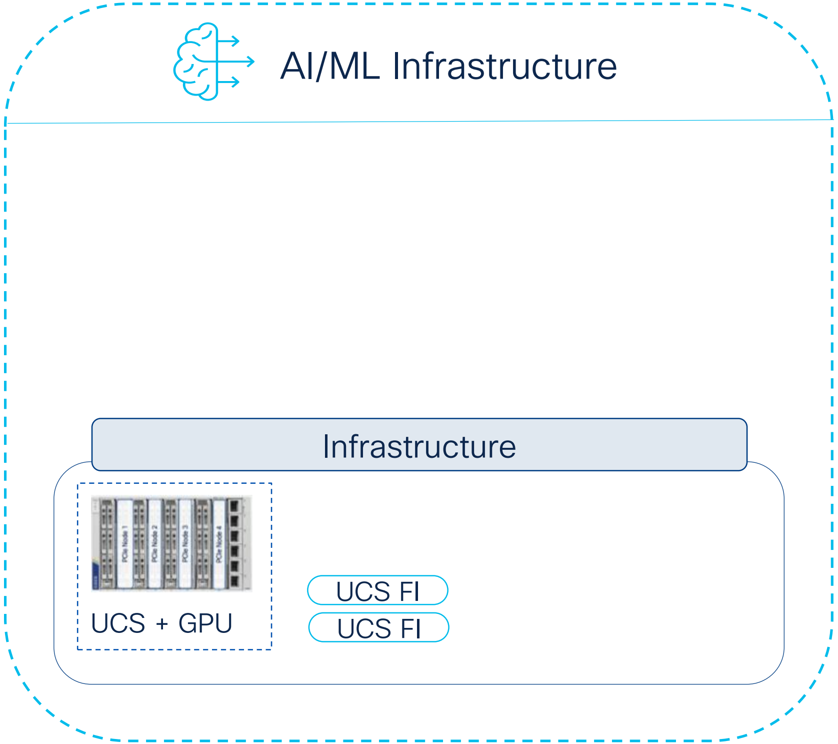
For inferencing, training/fine-tuning (smaller datasets), and other workloads

Cisco UCS

UCS rack and blade server providing a range of flexible and modular options including NVIDIA GPUs, Intel CPUs and AMD in the future

Cisco Intersight

SaaS platform enabling software-defined compute and cloud-based infrastructure management from data center to edge locations



Modular architecture

Ideal for AI component evolution

$49B

Global spending on data center construction by 2030

Investment preservation

- Convenience to upgrade or replace individual parts without overhauling the entire system
- Reduces cost and ensures that initial investments remain valuable over time

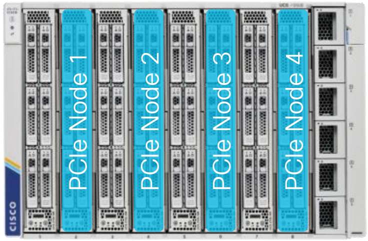
Multi-vendor support

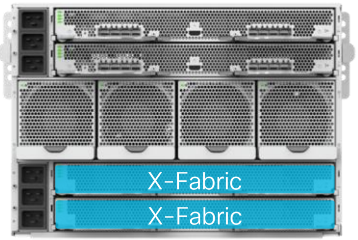
- Can select components from different vendors
- Best example is within CPU as you can move from Intel AMX to NVIDIA GPU A100 and then H100, or AMD in the future

Management & Upgradability

- Keep your technology stack current, adaptable, and competitive
- Cisco Intersight is a SaaS-based provides cloud-scale management from DC to edge

Modularity on X-Series





X-Series modular system decouples the lifecycles of CPU, GPU, memory, storage and fabrics – providing a perpetual architecture that efficiently brings you the latest innovations.

- Cloud-powered composability with Cisco Intersight
- Flexible GPU acceleration across server nodes
- No backplane or cables = easily upgrades

Flexible Acceleration

General Purpose Compute (CPU) ⟷ AI Acceleration (GPU)

1 Existing Applications

- Real-time video and audio streams
- Group chat
- Screen share
- Recording
- Generative AI for meeting summarization
- Inference for real-time transcription or translation

Example: AI-enabled Video Conferencing App

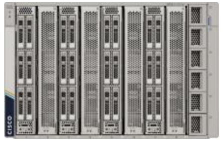
2 For diverse AI workloads

Data Ingest and Preparation	Edge Inferencing	Data Center Inferencing	Fine Tuning	Large Foundational Training
CPUs/GPUs	0-2	0-4	4-64	64-10K+

Cisco GPU Acceleration Options

Flexible Acceleration

X-Series



Up to 24x HHHL GPUs or
8x FHFL GPUs per X9508 chassis

X210c M6/M7 2S Blades
2x NVIDIA T4 (MEZZ)

X210c M7 2S Blade (Q2'24)
Intel Flex140 (MEZZ)

X210c M7 2S Blade (Q3'24)
NVIDIA L4 (MEZZ)

X215c M8 2S Blade (Q3'24)
NVIDIA L4 (MEZZ)

X440p + X210c M6/M7
4x NVIDIA T4 (M6 Only)
2x NVIDIA A16
2x NVIDIA A40
2 x NVIDIA A100-80

X440p + M7 (X210c & X410c)
2x NVIDIA H100-80
2x NVIDIA L40
4x NVIDIA L4
2x NVIDIA L40S

X440p + M7 (X210c & X410c)
4x Intel Flex140
2x Intel Flex170

Plan (Q3'24)

X440p + X210c M7
2x NVIDIA H100-NVL

X440p + X215c M8 AMD
2x NVIDIA H100-NVL
2x AMD MI210
4x NVIDIA L4
2x NVIDIA L40S
2x NVIDIA L40
2x NVIDIA A16

C-Series Rack Servers

C240 M6 INTEL
C245 M6 AMD

5x NVIDIA A10
3x NVIDIA A16
3x NVIDIA A30
3x NVIDIA A40
3x NVIDIA A100-80

8x NVIDIA L4
(C240 M6 only)

C240 M7 INTEL

3x NVIDIA A16
3x NVIDIA A30
3x NVIDIA A40
3x NVIDIA A100-80
2x NVIDIA H100-80
3x NVIDIA L40
8x NVIDIA L4
2x NVIDIA L40S
5x Intel Flex140
3x Intel Flex170

C220 M6 INTEL

3x NVIDIA T4
3x NVIDIA L4

C225 M6 AMD

3x NVIDIA T4

C220 M7 INTEL

3x NVIDIA L4
3x Intel FLex140

C245 M8 AMD

Plan (2H'24)

NVIDIA H100-80
NVIDIA L40S
NVIDIA L40
NVIDIA L4
NVIDIA H100-NVL
NVIDIA A16
AMD MI210

C225 M8 AMD

Plan (2H'24)

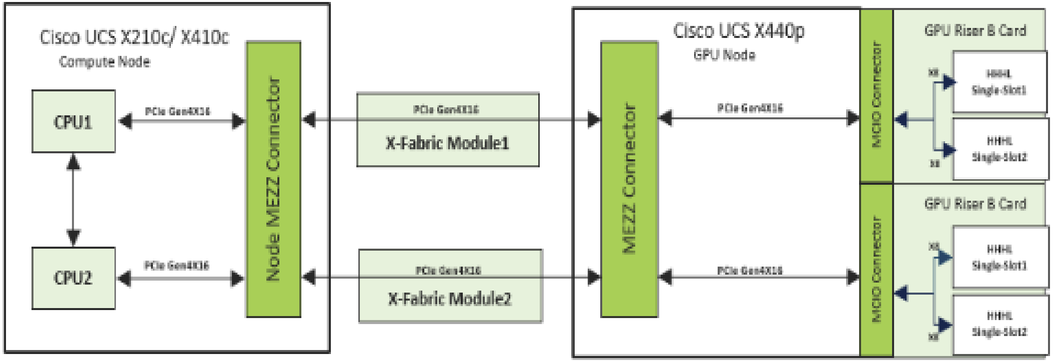
3x NVIDIA L4

Plans subject to change

Please Refer to the Server Specifications and HCL for detailed configuration support:
C-Series: https://www.cisco.com/c/en/us/support/servers-unified-computing/ucs-c-series-rack-servers/series.html#~tab-documents
X-Series: https://www.cisco.com/c/en/us/support/servers-unified-computing/ucs-x-series-modular-system/series.html#~tab-documents
UCS HCL: https://ucshcltool.cloudapps.cisco.com/public/

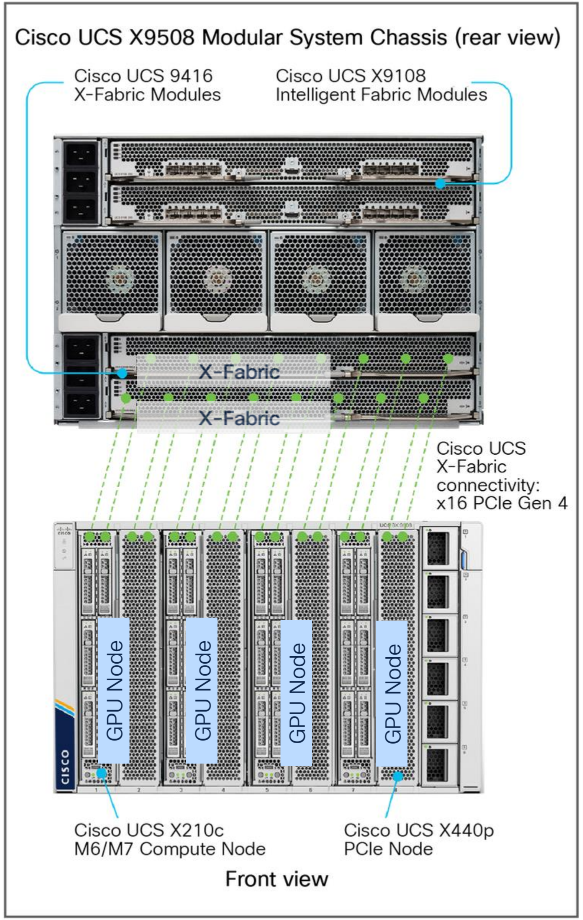
X-fabric + GPUs

- Each X440p is paired with a compute node in adjacent slot
- X-fabric provides PCIe Gen4 connectivity from server to GPU node (1:1 mapping)

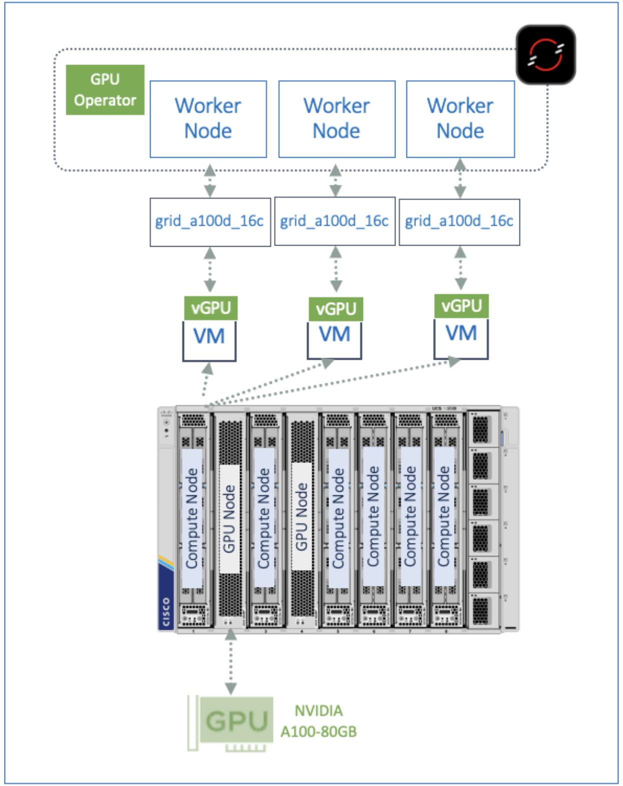


Scales up to:

960 cores per chassis	200 Gbps to compute node	~1600 Gbps aggregate bandwidth per chassis
24 GPUs per chassis	1-PB storage per chassis	



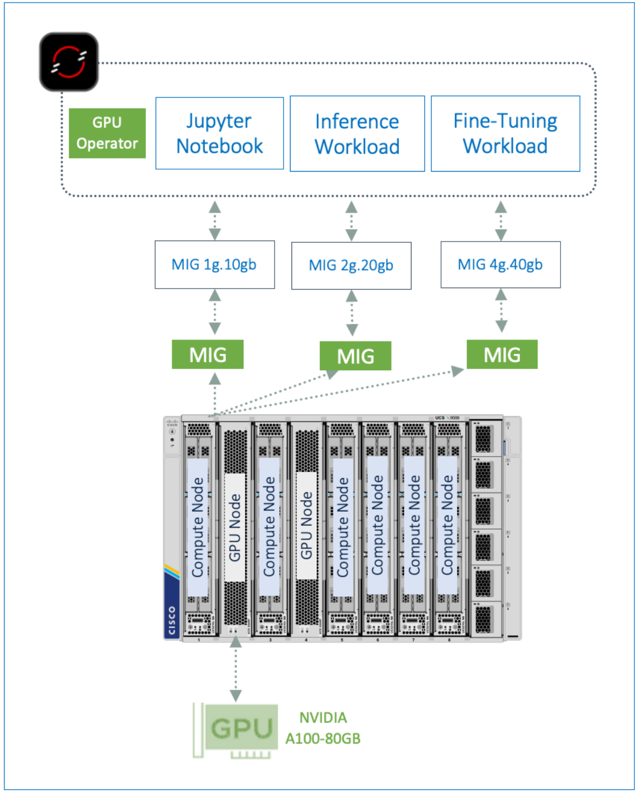
GPU Slicing – vGPU



NVIDIA vGPU Profile	Memory Buffer (MB)	Number of vGPUs per GPU
grid-a100d-80c	81920	1
grid-a100d-40c	40960	2
grid-a100d-20c	20480	4
grid-a100d-16c	16384	5
grid-a100d-10c	10240	8
grid-a100d-8c	8192	10
grid-a100d-4c	4096	20

- Memory isolation between instances but share compute
- Alternative to vGPU: GPU passthrough
- Can deploy Multi-Instance GPU (MIG) on vGPU instances

GPU Slicing – MIG

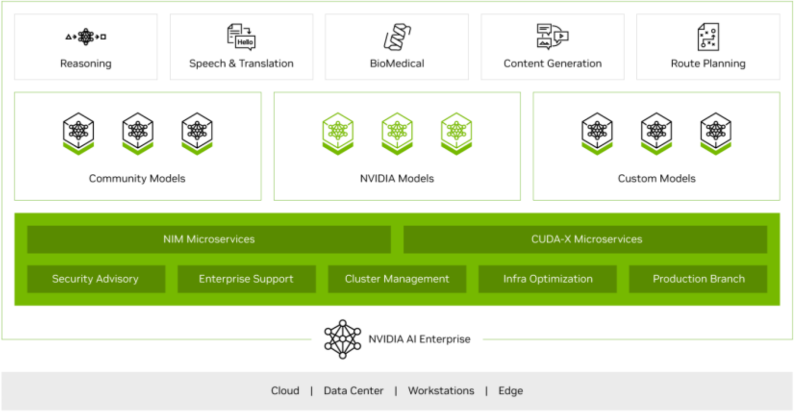


MIG Profile (A100-80)	GPU Instance - Memory (MB)	GPU Instance - SM Fraction	Number of GPU Instances	Compute Instances
MIG 7g.80gb	81920	7/7	1	7
MIG 4g.40gb	40960	4/7	1	4
MIG 3g.40gb	40960	3/7	2	3
MIG 2g.20gb	20480	2/7	3	2
MIG 1g.10gb	10240	1/7	7	1

- Multi-Instance GPI (MIG) - securely partitions up to 7 instances with isolation
- Can be further partitioned into compute instances
- Ampere (A100, H100) architecture onwards
- Bare-metal, VMs (GPU pass-through, vGPUs)

NVIDIA AI Enterprise (NVAIE)

NVIDIA GPU Licensing



- Required for all GPUs except for H100
- Enables support for features and services (NIM)
- Throttle GPU performance if not licensed
- Use any ML stack with NVIDIA GPUs

NVIDIA GPU Options for AI/ML workloads

Single GPU AI + HPC Secure Multi-Instance GPU

AI Training and Interference

HPC + Data Analytics

Confidential Compute MIG

Up to 2 GPUs per node
Up to 7 MIG Instances per node
Up to 8 vCPU cores per MIG

UCSC/X-GPU-H100-80
H100

350W | 80G | Gen5
2-slot FHFL

Fastest Universal AI + Graphics Text to Image/Video

Text to Image/Video AI
Multi-modal Generative AI

DL Training + Inference

Omniverse + Gen AI

Up to 2 GPUs per node
Fastest RT Graphics
Largest Render Models

UCSC/X-GPU-L40S
L40S

350W | 48G | Gen4
2-slot FHFL

Highest Perf Compute AI, HPC, Data Processing

DL Training

Scientific Research

Data Analytics

Fastest Compute, FP64
Up to 7 MIG instances

UCSC/X-GPU-A100-80
A100

300W | 80G | Gen4
2-slot FHFL

Sizing for Inferencing

LLM Inference – Estimating Memory

How much memory does my model need?

For a given precision: FP32, FP16, TF16...

- Model Memory

  Precision in Bytes x # of parameters (P)

Example: Llama2 – 13B parameters

- Model Memory:

  13 billion x 2Bytes/parameter = $\boxed{\text{26GB}}$

LLM Inference – Estimating Memory

How much memory does my model need?

For a given precision: FP32, FP16, TF16...

- Memory (Inference)

  Model Memory + ~20% overhead

Example: Llama2 – 13B parameters

- Memory (Inference):

  26GB + 20% overhead = 31.2GB

LLM Inference – Estimating Memory

How much memory does my model need?

For a given precision: FP32, FP16, TF16...

- Memory (Training)
  Model Memory
  + Optimiser Memory
  + Activation Memory
  + Gradient Memory

Example: Llama2 – 13B parameters

- Memory (Training):
  Model Memory (26GB)
  + Optimizer (4/6/12B/parameter * P)
  + Gradient (2/4B/parameter * P)
  + Activation ((2*P - 4*P) * Dataset (tokens))

Hugging Face Model Memory Calculator (Training & Inferencing):
https://huggingface.co/docs/accelerate/main/en/usage_guides/model_size_estimator

LLM Inference - GPU Estimation

Which GPU do I use?

Based on model memory, number of GPUs needed to load a 13B parameter model = any GPU with at least 32 GB

Similarly, a 70B parameter model, would require:
~2 A100-80 GPUs (168GB/80GB)

GPU Model	Memory (GB)	Memory Bandwidth (GB/s)	FP16 Tensor Core (TFLOP/s)
H100	80	2000	756
A100	80	1935	312
L40s	48	864	362
L4	24	300	121

LLM Inference Performance

How many GPUs do I need for inference?

Use Case	Model architecture	Context Length	GPU performance
▪ Determines model and minimum GPU ▪ CPU will also have an impact	▪ Impacts compute requirements per inference (TFLOPs)	▪ Will depend on the model ▪ Use average token size or vary token lengths in tests	▪ Will depend on its performance (TFLOPS) ▪ Use tests to verify performance

LLM Inferencing Performance

Objective and Subjective

Latency

- Time to first token
- Total Generation Time
- Time to second/next time

Throughput

- Requests per second dependent on concurrency and total generation time
- Tokens per second is the standard measure (> 30 per second)

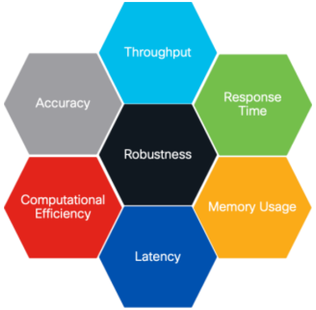
User experience – combination of low latency, throughput and accuracy

Prompt: What is Cisco UCS?

First Token

```
Cisco Unified Computing System (UCS) is a data center server computer
product line composed of computing hardware, virtualization support,
switching fabric, and management software. It was introduced by Cisco
Systems in 2009.
```

43 Output Tokens



LLM Inference - Methodology

How many GPUs do I need for inference?

For a given model and inferencing runtime, start with enough GPUs to load the model based on memory sizing

Vary concurrent inference requests and measure throughput and latency metrics for a given token length (context)

Vary batch sizes and measure throughput and latency - maximizes compute for non-RT use cases

Add a second GPU and repeat concurrent inference request and batch size tests (as needed)

Monitor GPU compute and memory utilization, along with inferencing performance, across all tests

Select a configuration that optimally balances latency, throughput and cost

Sample tool: https://github.com/openshift-psap/llm-load-test

Sample Benchmark Results – A100-80 GPU

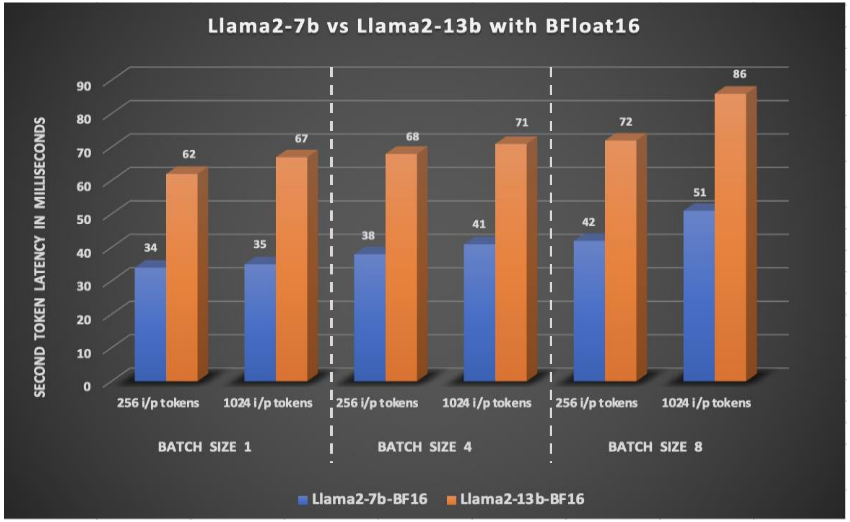
- Latency is higher as batch sizes increases
- For larger models
  - Latency is at least 2x higher
  - Throughput is at least 2x lower for larger models
- Latency is 2x or higher for larger models,

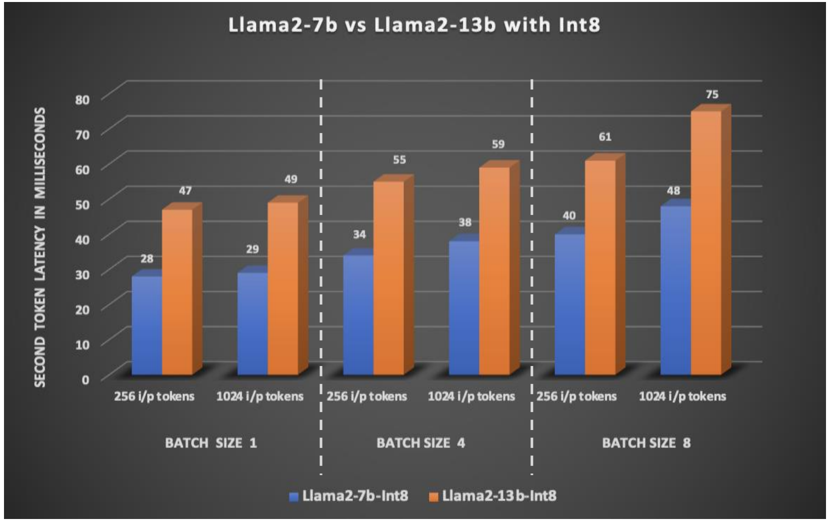
Inference performance from a user perspective needs to factor in the complete inferencing pipeline, including host cpu and memory

Model	Batch Size	Average Latency (ms)		Average Throughput (sentence/s)	
		1 GPU	2 GPUs	1 GPU	2 GPUs
Llama-2-7B-Chat	1	151.341	132.611	6.608	7.541
	2	156.135	143.724	12.809	13.916
	4	181.916	175.997	21.988	22.728
	8	231.947	254.829	34.491	31.394
Llama-2-13B-Chat	1	445.038	325.023	2.247	3.077
	2	464.125	357.096	4.309	5.601
	4	512.184	436.986	7.81	9.154
	8	604.336	551.75	13.238	14.499

Sample Benchmark Results – CPU

- Results for Intel's 5th Gen Xeon processor with built-in Intel AMX accelerator
- Results show before and after quantization
- Greater benefit with quantizing larger models, larger data size also improves accuracy and quality of output
- DeepSpeed enabled - optimization software for scaling and speeding up deep learning inference





***Hardware details** – Cisco UCS x210c M7 node, EMR CPU – 8568Y+ (48 cores), Memory – 1024GB, NVMe storage drive – 3.6TB
***Int8 –** Int8 is weight-only quantized (WOQ) to balance performance and accuracy

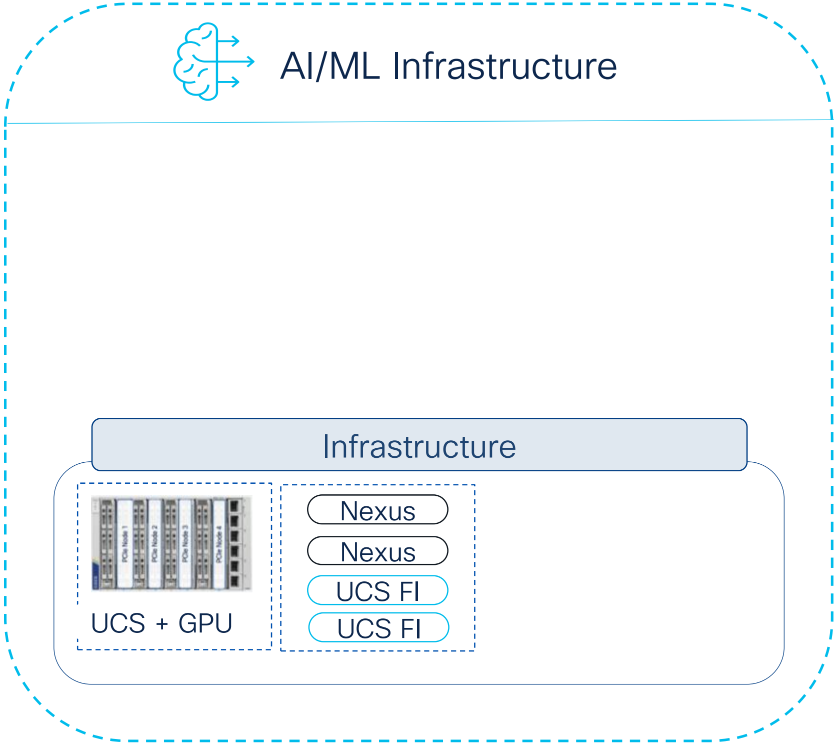
ML Infrastructure Design – Network

Cisco DC fabrics

Cisco ACI or VXLAN EVPN fabrics providing connectivity across top-of-racks that connect to compute and storage domains

Hyperscale Training Fabric

BGP and VXLAN EVPN based fabric, architected for dedicated training workloads



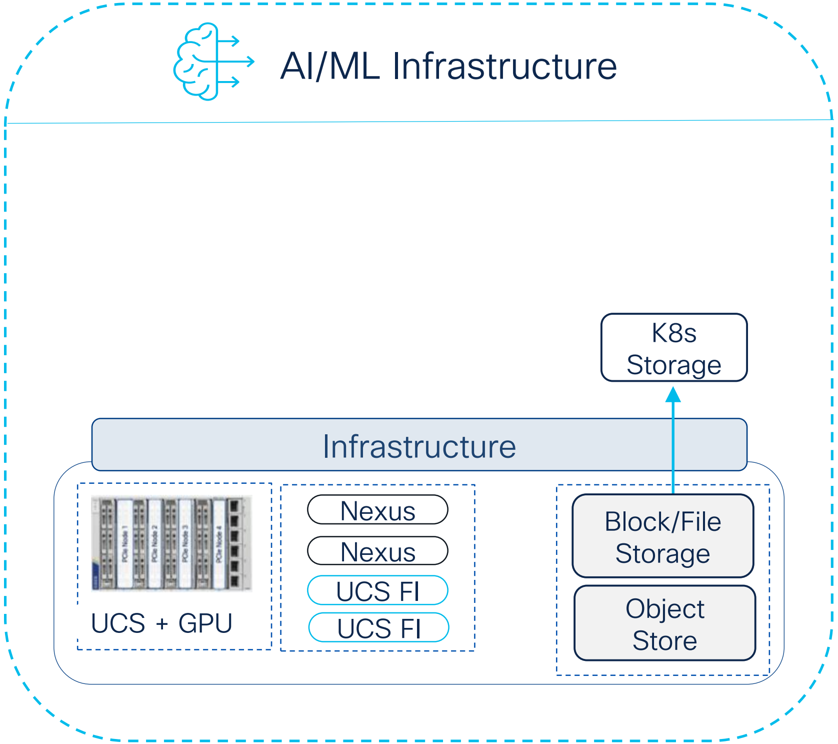
ML Infrastructure Design – Storage

Storage Partners

Range of eco-system enterprise storage partners including NetApp, Pure Storage, etc.

Local Storage

UCS-X system can support ~1 PB of local storage that can be leveraged using software-defined solutions such as Red Hat OpenShift Data Foundation, Nutanix for smaller efforts



Solution Components

MLOps for FlashStack AI using Red Hat OpenShift AI

On-Prem

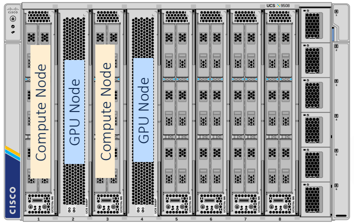
CISCO INTERSIGHT

Pure1

Cisco Nexus ®
93600CD-GX Switches

Cisco UCS 6536
Fabric Interconnects

Cisco UCS X9508
+ UCS X210c M7 Server
+ UCSX 9108 100G IFM
+ UCS VIC15231 Adapter
+ UCSX 9416 X-fabric
+ UCSX 440P PCIe Node



NVIDIA
A100-80GB GPU

Red Hat OpenShift AI

Red Hat OpenShift

NVIDIA
NVIDIA AI Enterprise

portworx by Pure Storage

Red Hat Ansible Automation Platform

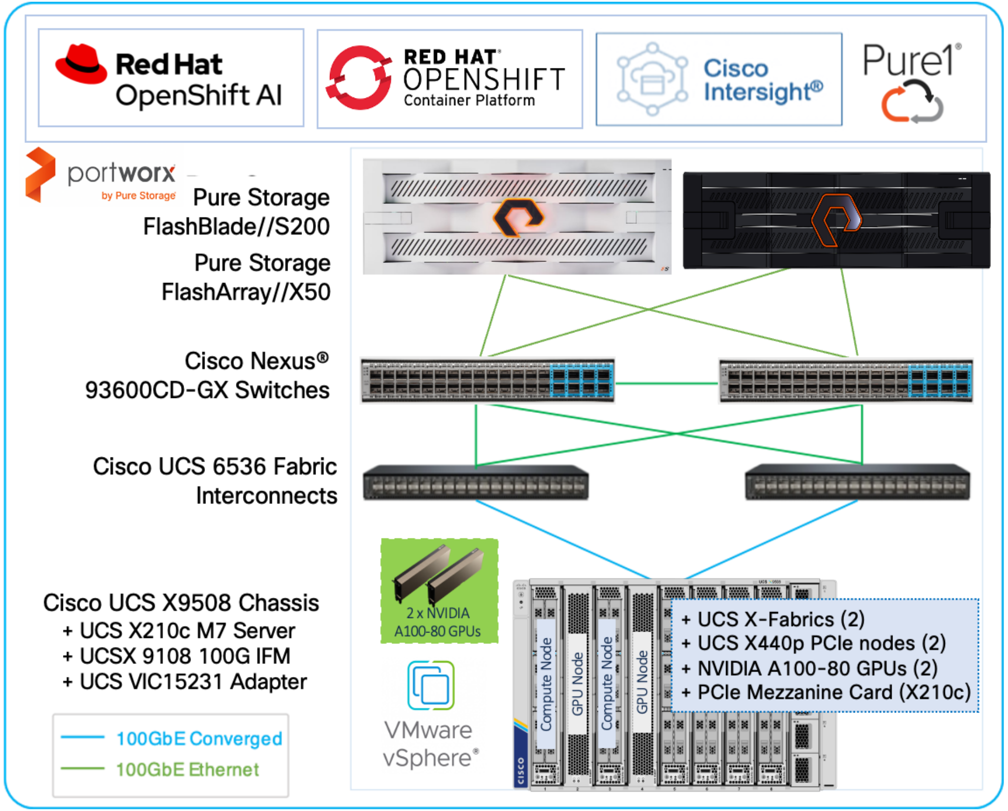
VMware vSphere®

Pure Storage
FlashArray//X50 R4

Pure Storage
FlashBlade//S200

Physical Topology

MLOps for FlashStack AI using Red Hat OpenShift AI



ML Infrastructure Design – K8s

Kubernetes

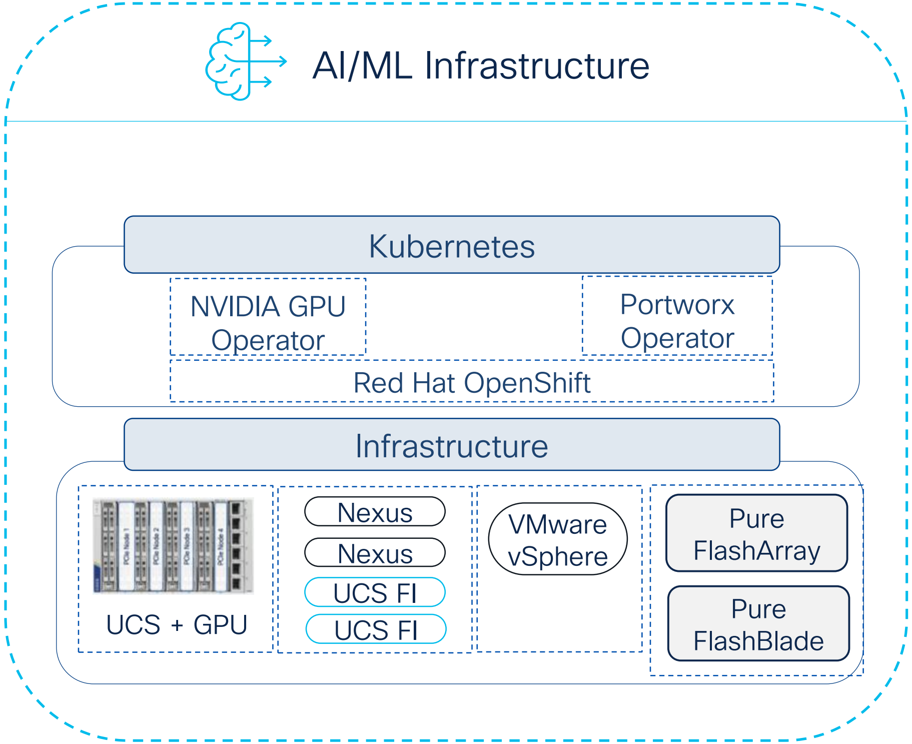
ML ecosystem has embraced containers for its portability, ease, and auto-scaling capabilities

K8s Operators

Operators provide a framework to add new capabilities to K8s including NVIDIA GPU and storage CSI operators

Virtualization

VMware vSphere enables GPU virtualization and mgmt. ease



OpenShift Operators

NVIDIA GPU Operator

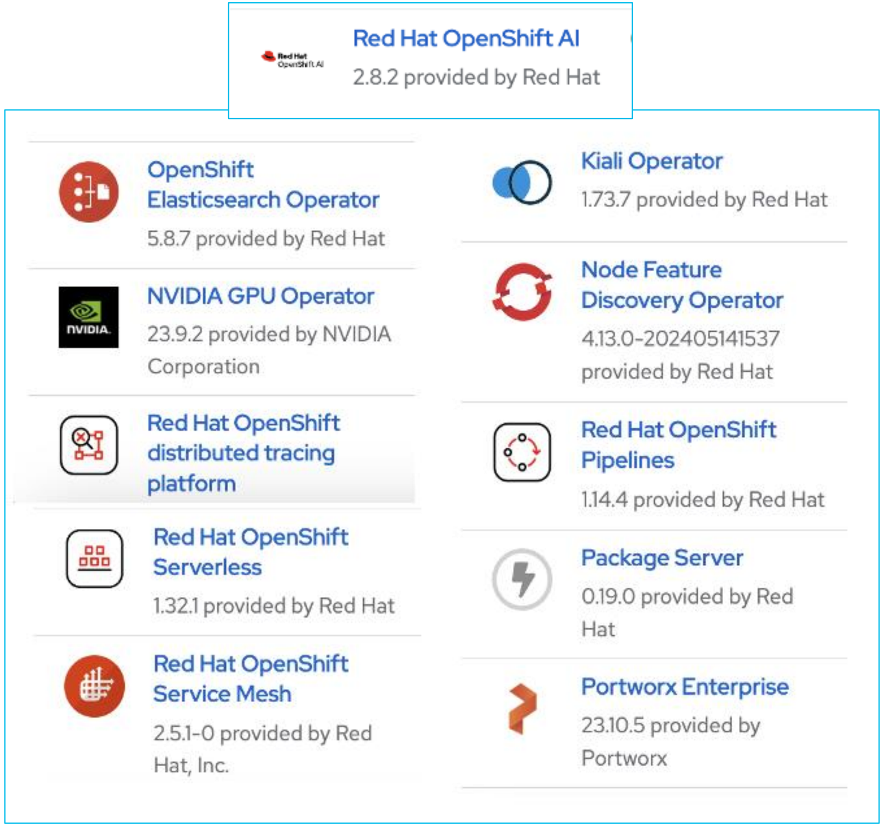
Automated the management of all NVIDIA software components required to use the GPU (drivers, DCGM, etc.)

Portworx Enterprise

Multi-cloud storage platform providing persistent storage with elastic scalability, with multiple storage backend options

Red Hat OpenShift AI

Provides a scalable foundation for AI/ML efforts to train, tune, serve, monitor and manage AI/ML experiments and models



Worker Node Considerations

- Add Taints/Tolerations

Tolerations on GPU workloads

```
tolerations:
  - key: nvidia/gpu
    operator: Exists
    effect: NoSchedule
```

Taints on worker nodes with GPU

```
taints:
  - key: nvidia/gpu
    effect: NoSchedule
```

- Worker nodes – Monitor CPU and memory and adjust as needed

K8s worker node

```
vCPUs: 16
RAM: 64GB
Storage: 500GB thin provisioned virtual disk
NIC: VMXNet3 connected to network
```

GPU Monitoring

Using nvidia-smi

GPU Burn Test

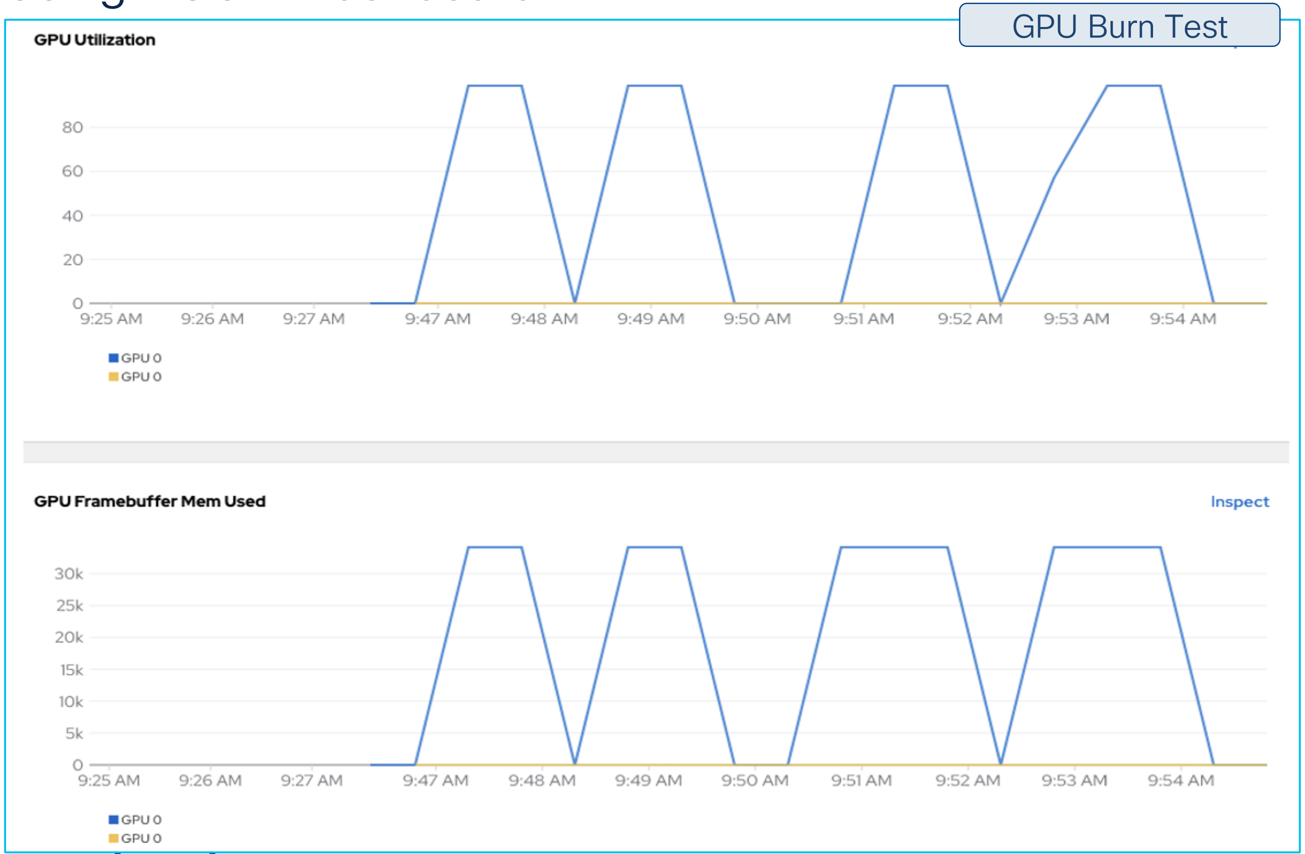
```
== CUDA ==
==========
CUDA Version 12.0.0
Container image Copyright (c) 2016-2023, NVIDIA CORPORATION & AFFILIATES. All
rights reserved.
....
GPU 0: GRID A100D-40C (UUID: GPU-ef5a53d2-34d3-11b2-99cb-146
Using compare file: compare.ptx
Burning for 60 seconds.
....
30.0%  proc'd: 128 (9171 Gflop/s)   errors: 0   temps: --
....
46.7%  proc'd: 256 (18593 Gflop/s)   errors: 0   temps: --
....
55.0%  proc'd: 384 (18567 Gflop/s)   errors: 0   temps: --
.....
63.3%  proc'd: 384 (18567 Gflop/s)   errors: 0   temps: --
....
71.7%  proc'd: 512 (18536 Gflop/s)   errors: 0   temps: --
....
80.0%  proc'd: 640 (18514 Gflop/s)   errors: 0   temps: --
....
90.0%  proc'd: 768 (18466 Gflop/s)   errors: 0   temps: --
....
100.0%  proc'd: 896 (18449 Gflop/s)   errors: 0   temps: --
....
Burning for 60 seconds.
Initialized device 0 with 40955 MB of memory (37077 MB avail
of it), using FLOATS
Results are 268435456 bytes each, thus performing 128 iterat
...
Tested 1 GPUs:
    GPU 0: OK
```

```
[administrator@FSV-AI-OCP-Installer OCP3]$ oc exec -it nvidia-driver-daemonset-413.92.202309261804-0-zshvt -- nvidia-smi
+-----------------------------------------------------------------------------+
| NVIDIA-SMI 525.60.13    Driver Version: 525.60.13    CUDA Version: 12.0     |
|-------------------------------+----------------------+----------------------+
| GPU  Name        Persistence-M| Bus-Id        Disp.A | Volatile Uncorr. ECC |
| Fan  Temp  Perf  Pwr:Usage/Cap|         Memory-Usage | GPU-Util  Compute M. |
|                               |                      |               MIG M. |
|===============================+======================+======================|
|   0  GRID A100D-40C      On   | 00000000:02:00.0 Off |                    0 |
| N/A   N/A    P0    N/A /  N/A |  34133MiB / 40960MiB |     99%      Default |
|                               |                      |             Disabled |
+-------------------------------+----------------------+----------------------+

+-----------------------------------------------------------------------------+
| Processes:                                                                  |
|  GPU   GI   CI        PID   Type   Process name                  GPU Memory |
|        ID   ID                                                   Usage      |
|=============================================================================|
|    0   N/A  N/A    425634      C   ./gpu_burn                     34069MiB  |
+-----------------------------------------------------------------------------+
```

GPU Monitoring

Using DCGM Dashboard



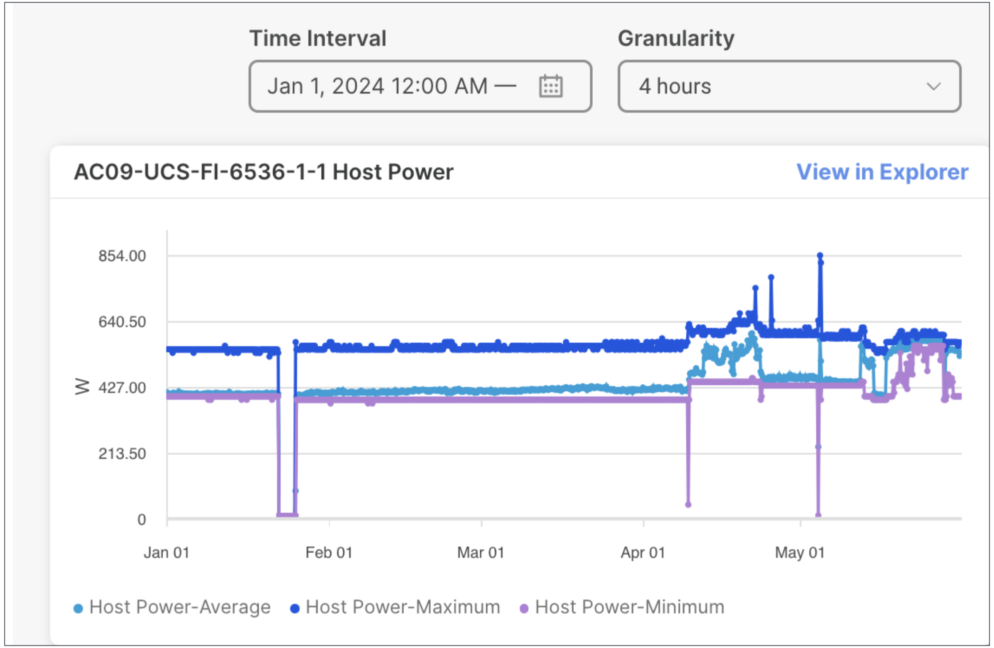
GPU Temperature
GPU Avg. Temp
GPU Power Usage
GPU Power Total
GPU SM Clocks
GPU Utilization
GPU Framebuffer Mem Used
Tensor Core Utilization

Server Power Consumption

Cisco Intersight

UCSX-210C-M7 server with 2-socket 4th-Gen Intel® Xeon® Gold 6430 processors and1 x A100-80 GPU

GPU Model	Power
H100	350W
A100	300W
L40s	350W
L4	72W



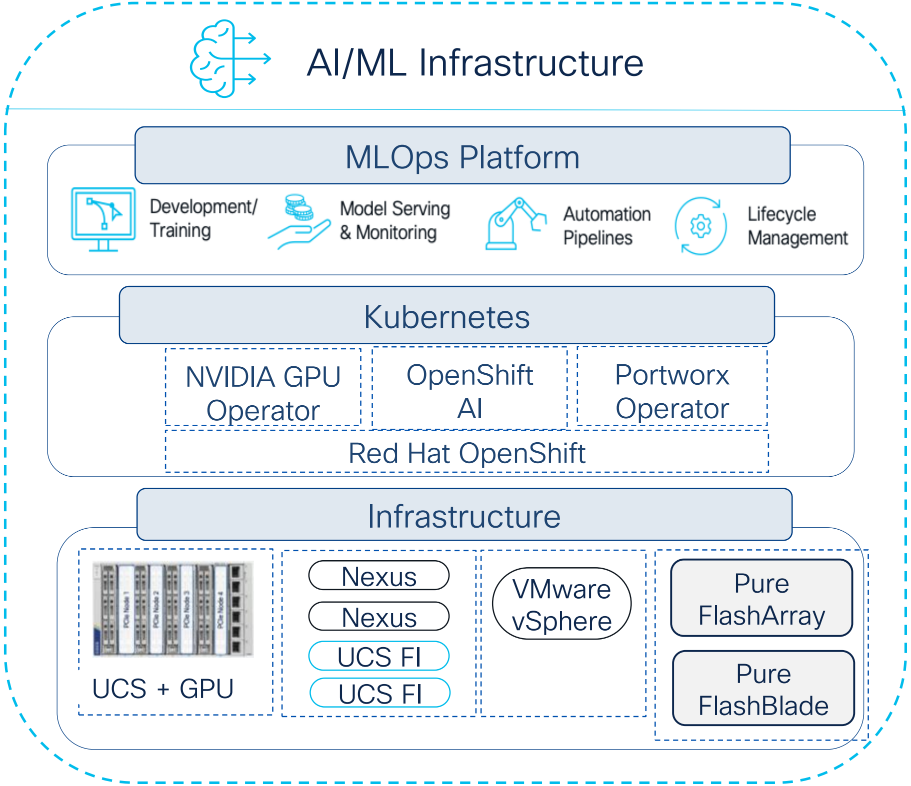
ML Infrastructure Design – MLOps

MLOps platform

OpenShift AI provides a scalable foundation for AI/ML efforts

K8s Operators

OpenShift AI operator is deployed to enable MLOps platform



Operationalizing AI/ML with Red Hat OpenShift AI

Red Hat OpenShift AI

Hybrid MLOps Platform

- An AI-focused platform that provides tools to train, tune, serve, monitor and manage AI/ML experiments and models.
- Collaborate within a common platform to bring IT, data science, and app dev teams together.
- Available as managed service or as self-managed on-site or in the cloud!
- Runs anywhere Red Hat Openshift does

Model development

Use core AI / ML libraries and frameworks including TensorFlow and PyTorch using Red Hat's notebook images or your own

Model serving & monitoring

Deploy models across any cloud, fully managed, and self-managed OpenShift and monitor their performance

Lifecycle Management

Create repeatable data science pipelines and integrate them with devops pipelines for delivery of models across your enterprise

Increased capabilities / collaboration

Create and share projects across teams. Combine Red Hat components, open-source software, and ISV certified software

Integrations

Retrain models

Gather and prepare data | Develop model | Integrate models in app dev | Model monitoring and management

Customer managed applications

Customer managed ISV software

elastic, Starburst, Pachyderm, mongoDB, crunchydata, Intel AI Analytics Toolkit, CLOUDERA, H2O.ai, redislabs, sas, watsonx, C3.ai, OpenVINO, CognitiveScale, avanseus, cnvrg.io

ISV managed cloud services

Starburst Galaxy, ANACONDA

Red Hat software and cloud services

Red Hat OpenShift AI, PyTorch, TensorFlow, jupyter

Model serving

Model monitoring

Data science pipelines

Red Hat AMQ Streams

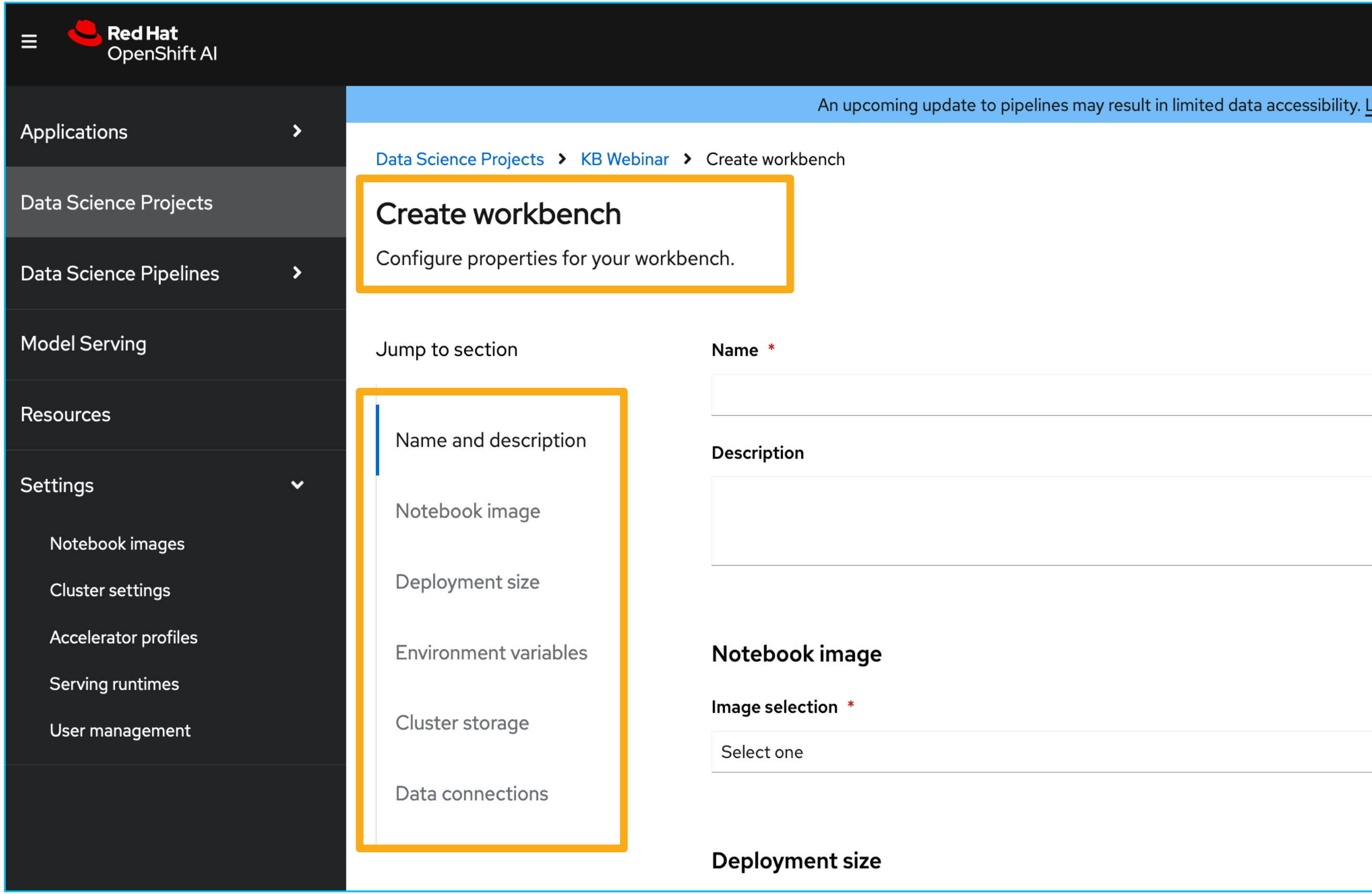
Red Hat on-premise and cloud platform

Open hybrid cloud platform: Red Hat OpenShift

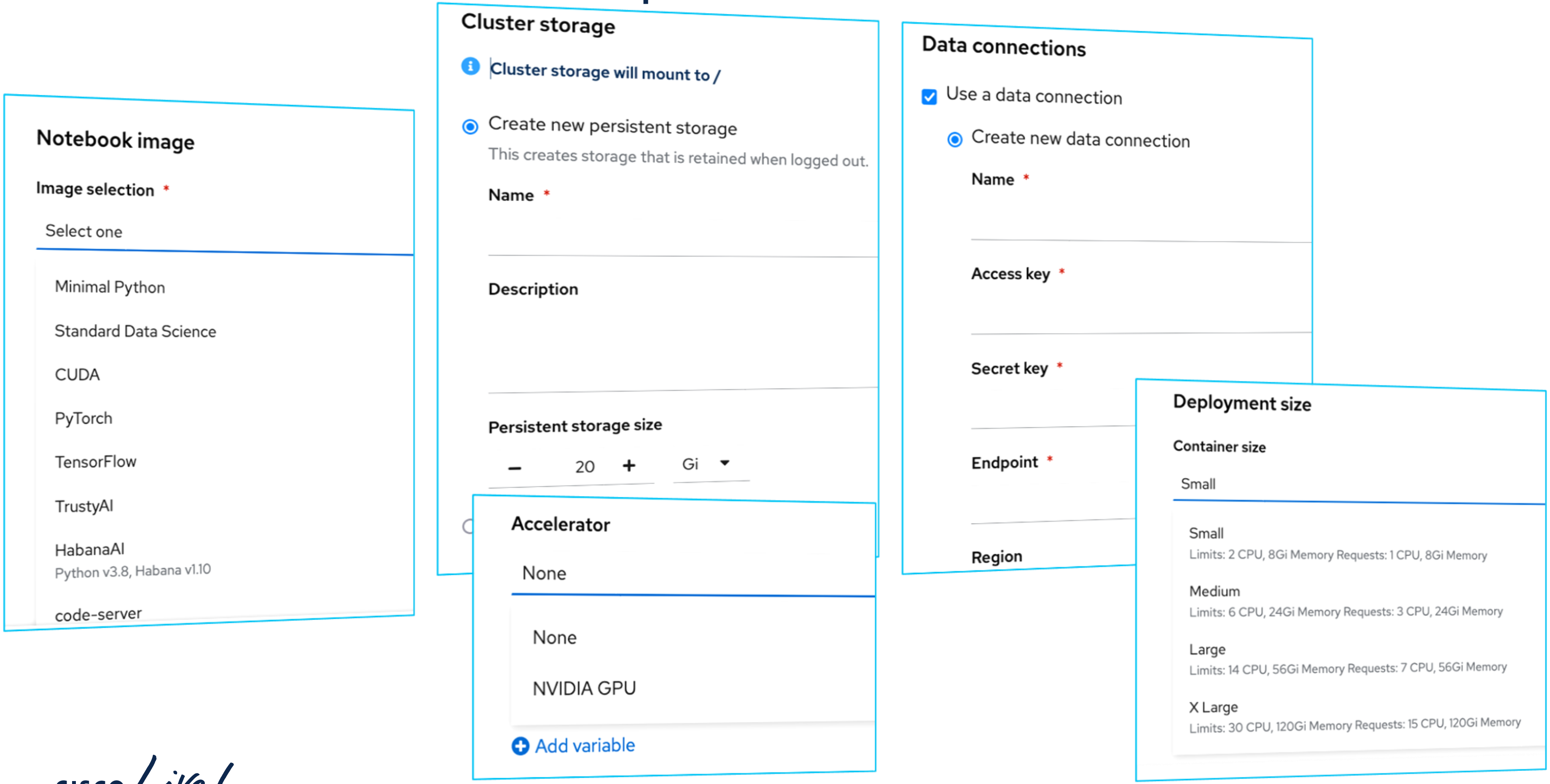
Accelerators: intel, NVIDIA

On-premise, cloud or edge infrastructure

Starting point for your AI/ML project



Workbench Infra Setup



Red Hat OpenShift AI

kube:admin

- Applications
- Data Science Projects
- Data Science Pipelines
- Model Serving
- Resources
- Settings
  - Notebook images
  - Cluster settings
  - Accelerator profiles
  - Serving runtimes
  - User management

An upcoming update to pipelines may result in limited data accessibility. Learn more

Data Science Projects > Cisco DC Demo: Fraud Detection-Intel

Cisco DC Demo: Fraud Detection-Intel

Starts your Jupyter notebook environment for development

Components | Permissions

Jump to section

- Workbenches
- Cluster storage
- Data connections
- Pipelines
- Models and model servers

Workbenches

Create workbench

Name	Notebook image	Container size	Status		
Demo-FD-Intel_WorkBench	TensorFlow	Medium	Running	Open	
Test-WB-1	Standard Data Science	Small	Stopped	Open	

Cluster storage

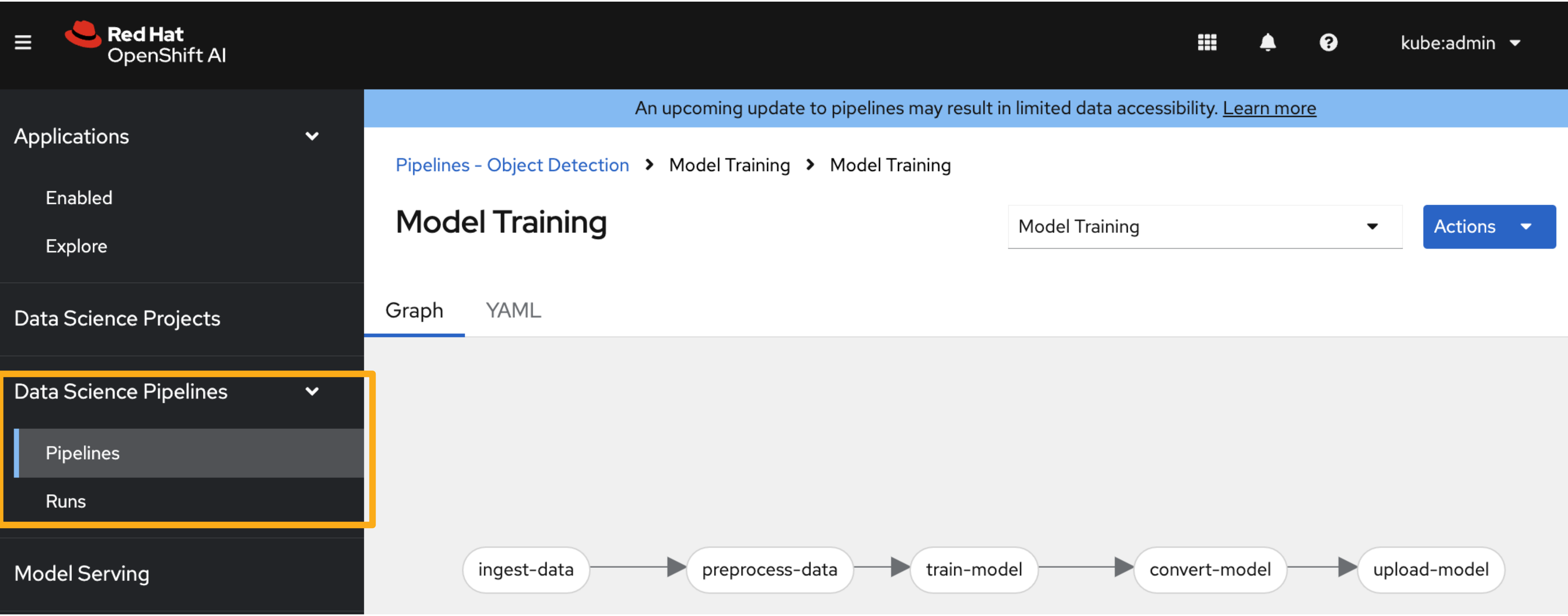
Add cluster storage

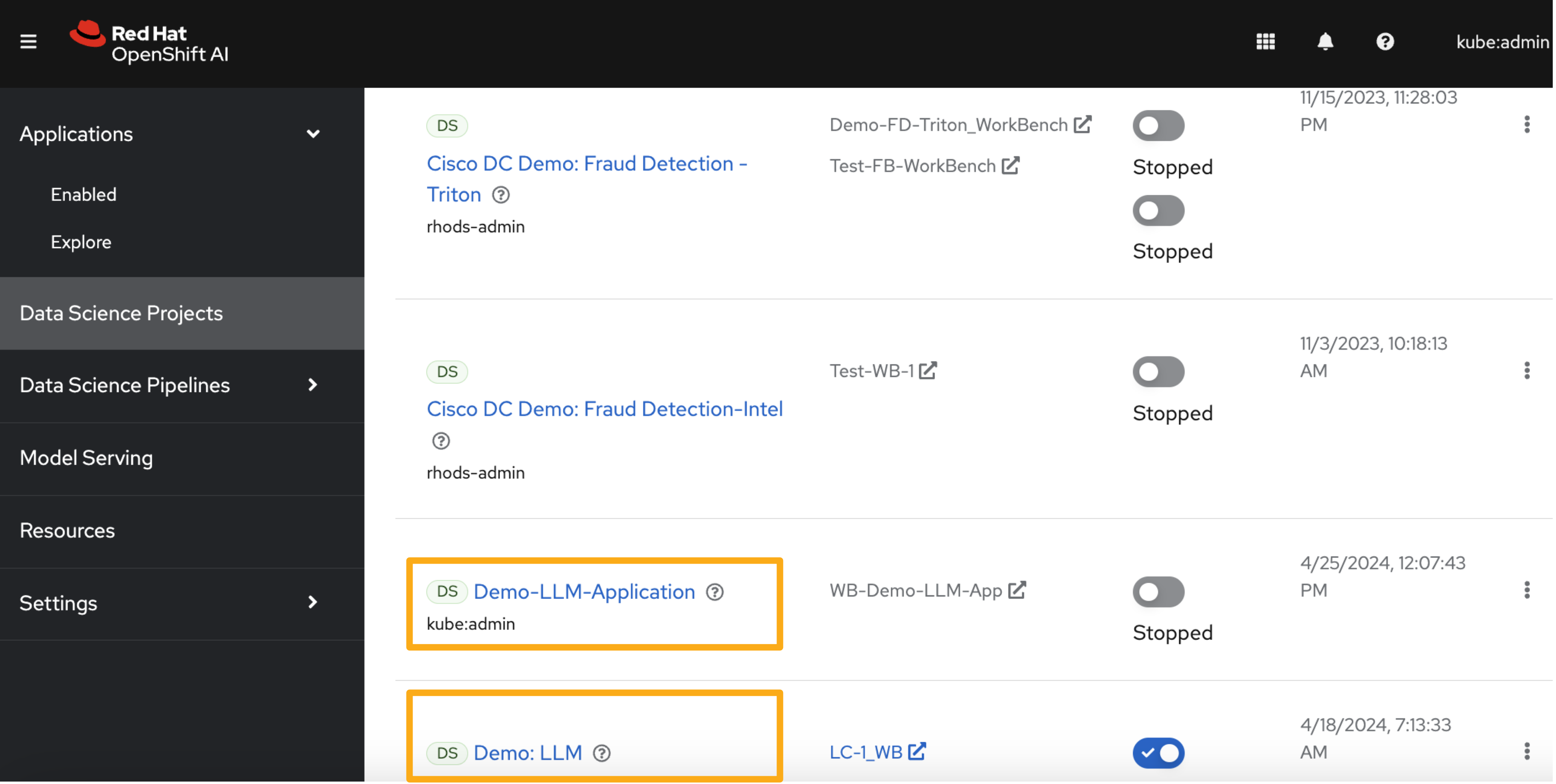
Name	Type	Connected workbenches
Demo-FD-Intel_WorkBench_PV	Persistent storage	Demo-FD-Intel_WorkBench
Test-WB-1	Persistent storage	Test-WB-1

Data connections

Add data connection

Name	Type	Connected workbenches
Model Storage - Pure FB-1	Object storage	No connections
Model Storage - Pure FB - 2	Object storage	Demo-FD-Intel_WorkBench
Pipeline Artifacts	Object storage	No connections


Red Hat
OpenShift AI
kube:admin
An upcoming update to pipelines may result in limited data accessibility. Learn more
Applications
Enabled
Explore
Data Science Projects
Data Science Pipelines
Pipelines
Runs
Model Serving
Pipelines - Object Detection > Model Training > Model Training
Model Training
Model Training
Actions
Graph
YAML
ingest-data
preprocess-data
train-model
convert-model
upload-model


Red Hat
OpenShift AI
kube:admin
Applications
Enabled
Explore
Data Science Projects
Data Science Pipelines
Model Serving
Resources
Settings
DS
Cisco DC Demo: Fraud Detection - Triton
rhods-admin
Demo-FD-Triton_WorkBench
Test-FB-WorkBench
Stopped
Stopped
11/15/2023, 11:28:03 PM
DS
Cisco DC Demo: Fraud Detection-Intel
rhods-admin
Test-WB-1
Stopped
11/3/2023, 10:18:13 AM
DS
Demo-LLM-Application
kube:admin
WB-Demo-LLM-App
Stopped
4/25/2024, 12:07:43 PM
DS
Demo: LLM
LC-1_WB
4/18/2024, 7:13:33 AM

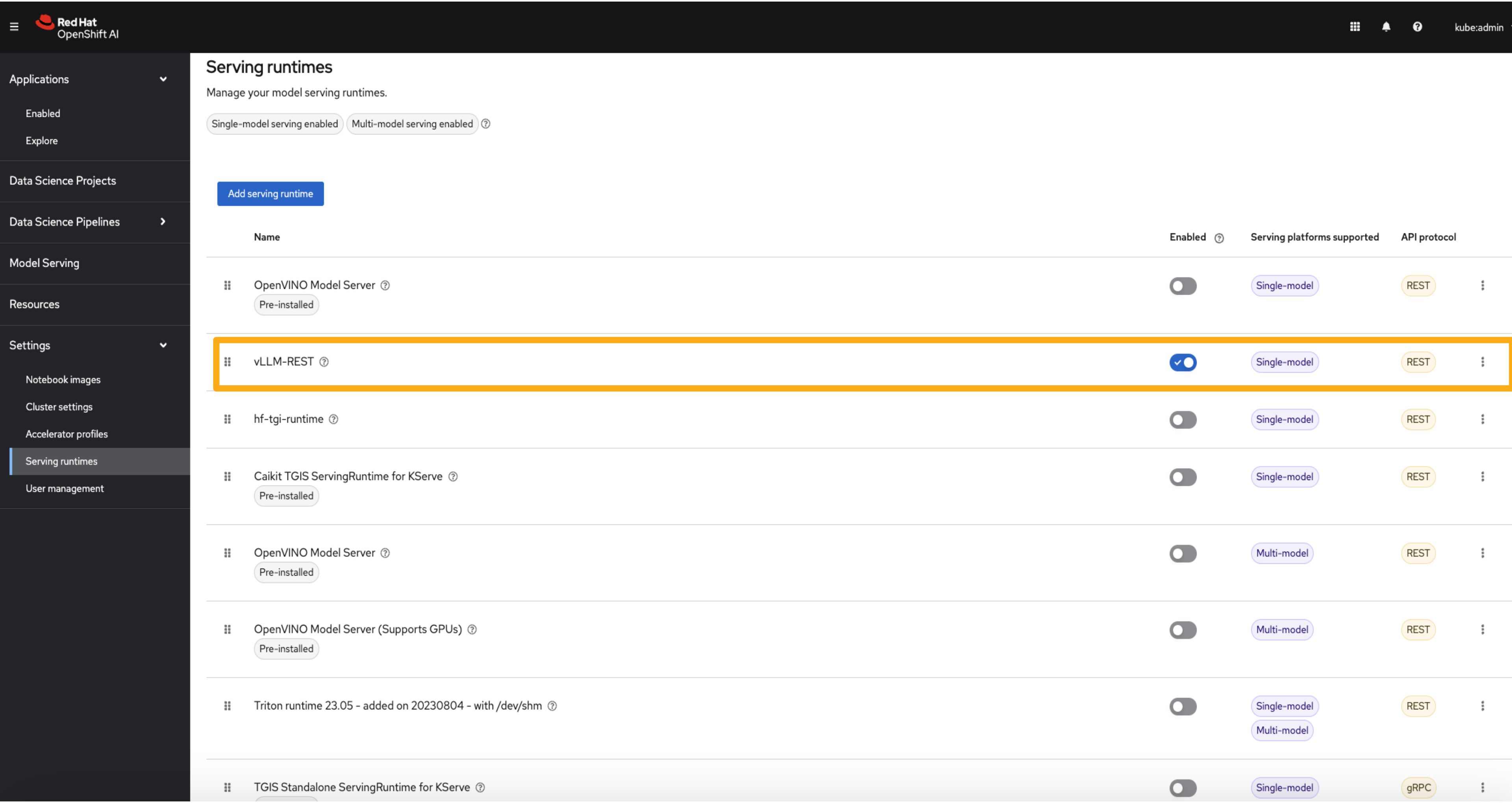
AI

kube:admin

s

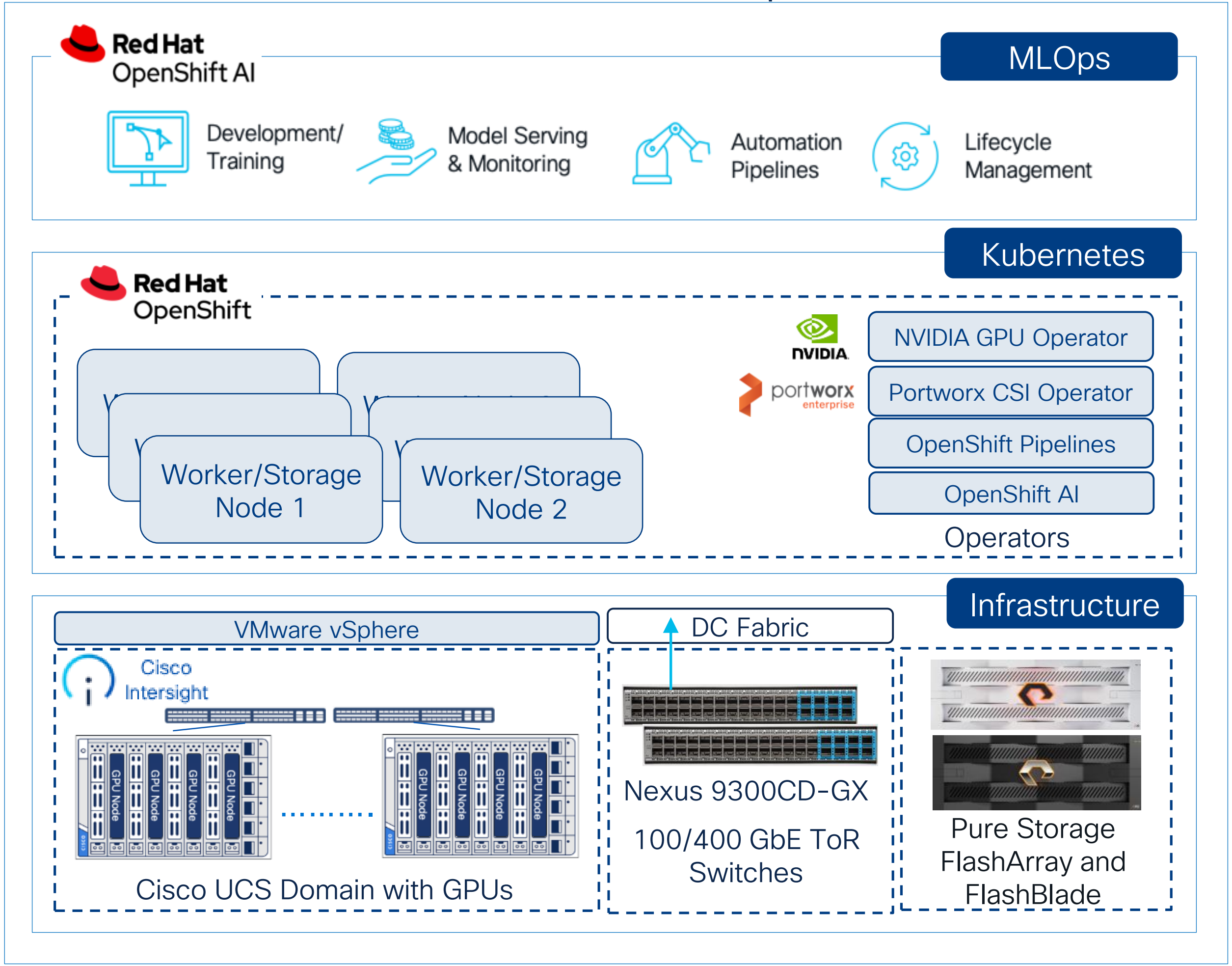
es

Model name	Project	Serving runtime	Inference endpoint	API protocol	Status
fraud	Cisco DC Demo: Fraud Detection-Intel Multi-model serving enabled	OpenVINO Model Server	Internal Service	REST	
Mistral-7B-Instruct	Demo: LLM Single-model serving enabled	vLLM-REST	https://...	REST	
yolo	Object Detection Multi-model serving enabled	OpenVINO Model Server	Internal Service	REST	
yolov5	Object Detection Multi-model serving enabled	OpenVINO Model Server	Internal Service	REST	


Red Hat
OpenShift AI
kube:admin
Applications
Enabled
Explore
Data Science Projects
Data Science Pipelines
Model Serving
Resources
Settings
Notebook images
Cluster settings
Accelerator profiles
Serving runtimes
User management
Serving runtimes
Manage your model serving runtimes.
Single-model serving enabled
Multi-model serving enabled
Add serving runtime
Name
Enabled
Serving platforms supported
API protocol
OpenVINO Model Server
Pre-installed
Single-model
REST
vLLM-REST
Single-model
REST
hf-tgi-runtime
Single-model
REST
Caikit TGIS ServingRuntime for KServe
Pre-installed
Single-model
REST
OpenVINO Model Server
Pre-installed
Multi-model
REST
OpenVINO Model Server (Supports GPUs)
Pre-installed
Multi-model
REST
Triton runtime 23.05 - added on 20230804 - with /dev/shm
Single-model
Multi-model
REST
TGIS Standalone ServingRuntime for KServe
Single-model
gRPC

AI/ML Ready Infrastructure

FlashStack for AI with Red Hat OpenShift AI



Enterprise AI/ML Platform

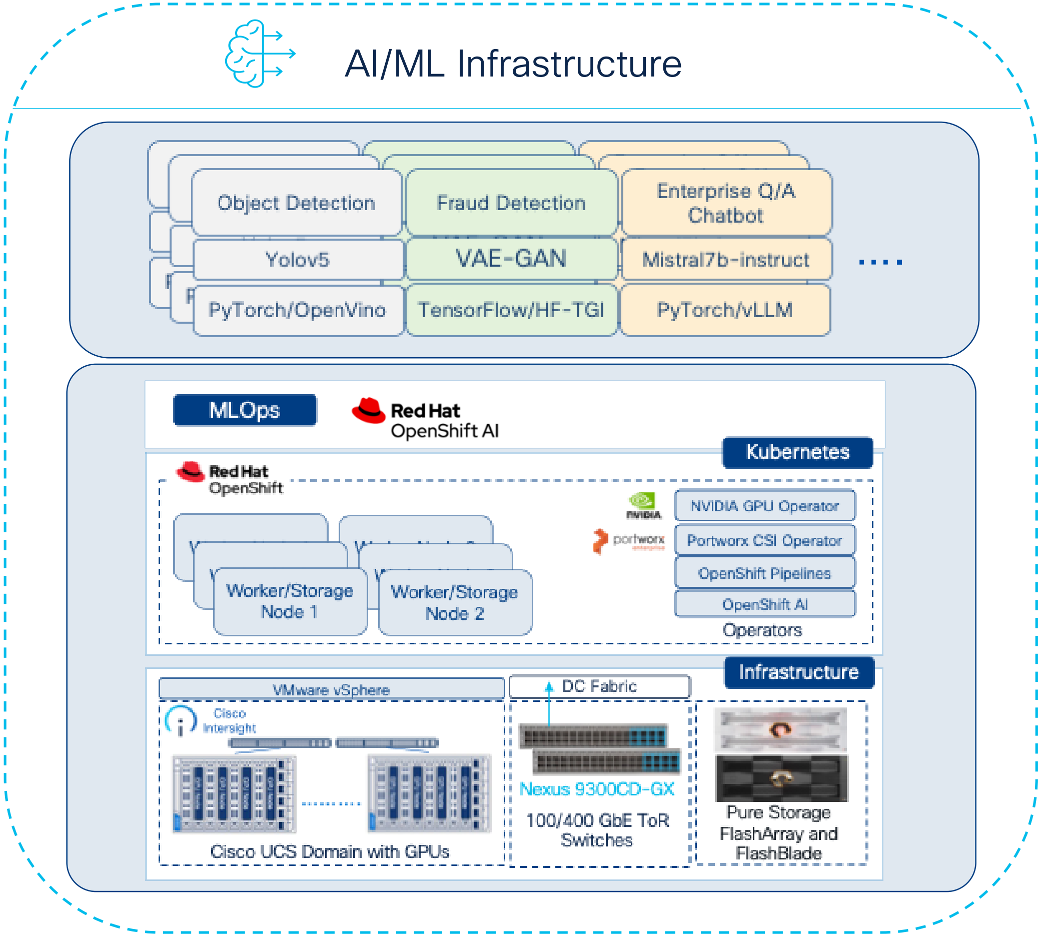
Scalable model delivery

Support multiple AI/ML efforts and use cases at scale with ease and consistency

Generative AI

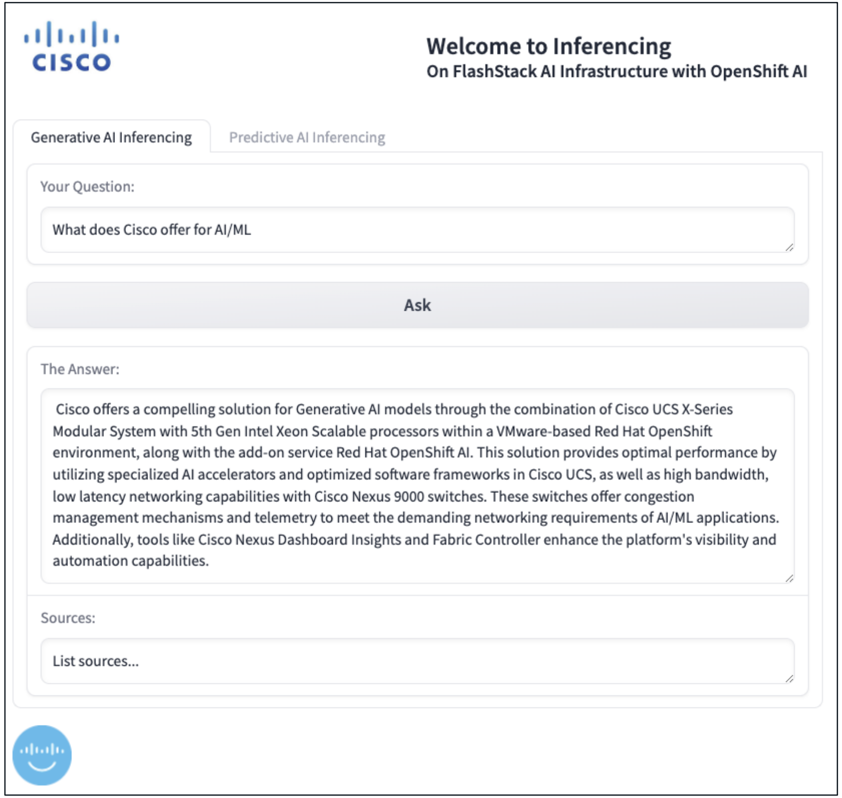
Predictive AI/Classis ML

- Customer & employee experience
- Language & code generation
- Recommendation systems



Demo - Q/A Chatbot using Enterprise knowledgebase

UI Frontend

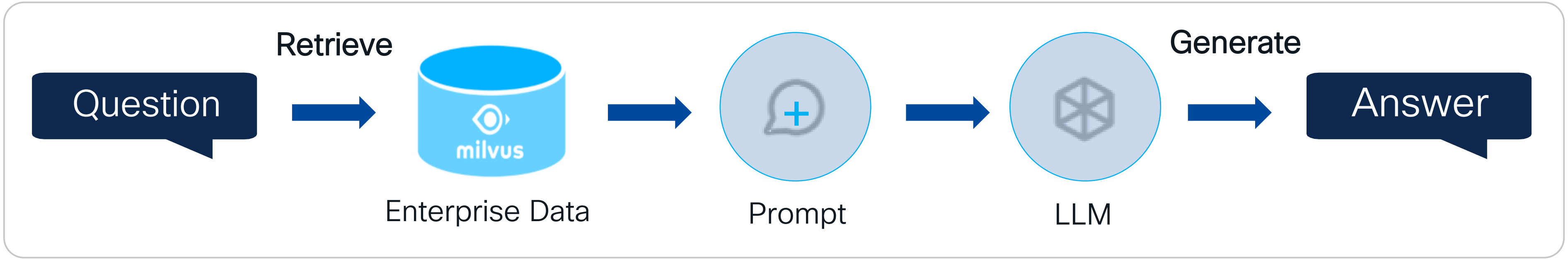


USE CASE COMPONENTS

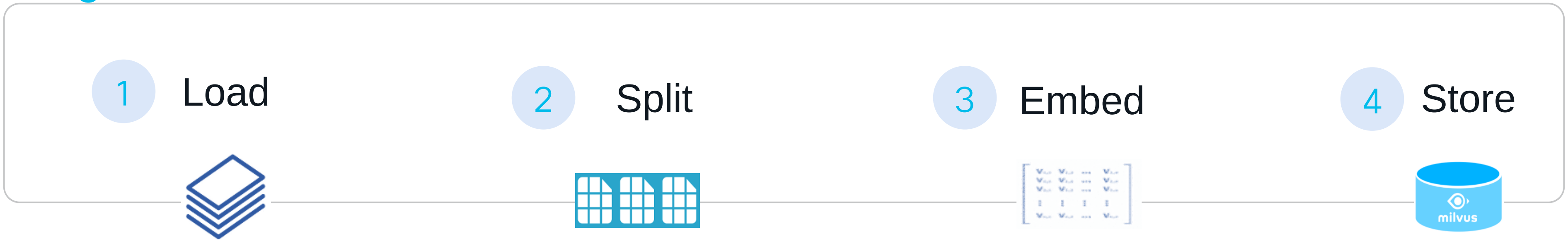
- AI-ready stack with Red Hat OpenShift
- MLOps (Red Hat OpenShift AI)
- NVIDIA GPU with 24GB of VRAM
- Large Language Model (LLM)
- Inferencing runtime (vLLM)
- Model Serving Platform (Kserve, Knative)
- ML Framework (PyTorch)
- Vector Store (Milvus)
- Embedding Model (NomicAI)
- Store and retrieval pipeline (LangChain)
- UI Engine (Gradio)

Demo Overview

Retrieval and Generation



Ingest and Store



Deployment Workflow

1

Deploy AI-ready infrastructure – FlashStack AI CVD

Cisco Intersight

ANSIBLE

FlashStack AI

Deploy Red Hat OpenShift and other resources

portworx by Pure Storage

Red Hat OpenShift AI

- Image Registry
- Pipelines Artifacts Repo
- Model Repo

3

Deploy two projects* in Openshift AI

- Model delivery pipeline
- Application inferencing pipeline**

* Workbenches/namespaces

** For demo purposes

4

Load LLM from Hugging Face and explore/evaluate

5

Save and upload model to Model Repo

6

Deploy model serving runtime

vLLM

7

Deploy LLM for inferencing

8

Deploy Vector Database for RAG

milvus

Attu open-source GUI

9

Ingest Enterprise data to vector database

Unstructured Data

milvus

10

Deploy Q/A Chatbot App using Enterprise Knowledge Base

LangChain

11

Deploy GUI front-end for Q/A Chatbot

gradio

12

Deploy Enterprise Q/A Chatbot for inferencing

Core

Edge

Wrap-up

Key Takeaways

Adopt MLOps to scale and accelerate AI/ML efforts with ease consistency

AI/ML workloads need a range of accelerators

Rapid pace of innovations – flexibility is key

Key Resources

Cisco MLOps CVD

- https://www.cisco.com/c/en/us/td/docs/unified_computing/ucs/UCS_CVDs/flashstack_ai_ml_ops.html

GitHub Repo

- https://github.com/ucs-compute-solutions/FlashStack-OpenShift-AI

Design Zone for AI Ready Infrastructure

- https://www.cisco.com/c/en/us/solutions/design-zone/ai-ready-infrastructure.html

Complete Your Session Evaluations

Complete a minimum of 4 session surveys and the Overall Event Survey to be entered in a drawing to **win 1 of 5 full conference passes** to Cisco Live 2025.

Earn 100 points per survey completed and compete on the Cisco Live Challenge leaderboard.

Level up and earn **exclusive prizes!**

Complete your surveys in the **Cisco Live mobile app.**

Continue your education

- Visit the Cisco Showcase for related demos
- Book your one-on-one Meet the Engineer meeting
- Attend the interactive education with DevNet, Capture the Flag, and Walk-in Labs
- Visit the On-Demand Library for more sessions at www.CiscoLive.com/on-demand

Contact me at: asharma@cisco.com

CISCO Live!

The bridge to possible

Thank you

#CiscoLive