Red Hat Summit

Connect

I servizi di inferenza per l'AI con HPE e Red Hat Openshift

Alessandro Moriondo, Hybrid Solutions Presales Manager
Hewlett Packard Enterprise - Italy

Rome, November 7th

2024

Hewlett Packard Enterprise | Red Hat

Agenda

HPE STRATEGY

HPE MACHINE LEARNING INFERENCE SOFTWARE

HPE | RED HAT OCP INFRASTRUCTURE BLUEPRINT

HPE | RED HAT OCP GPUS CONCURRENCY MODEL

Q&A

HPE Strategy – The key Pillars of Digital Enterprise

Edge
Connect your edge
Control and harness data to innovate at the edge

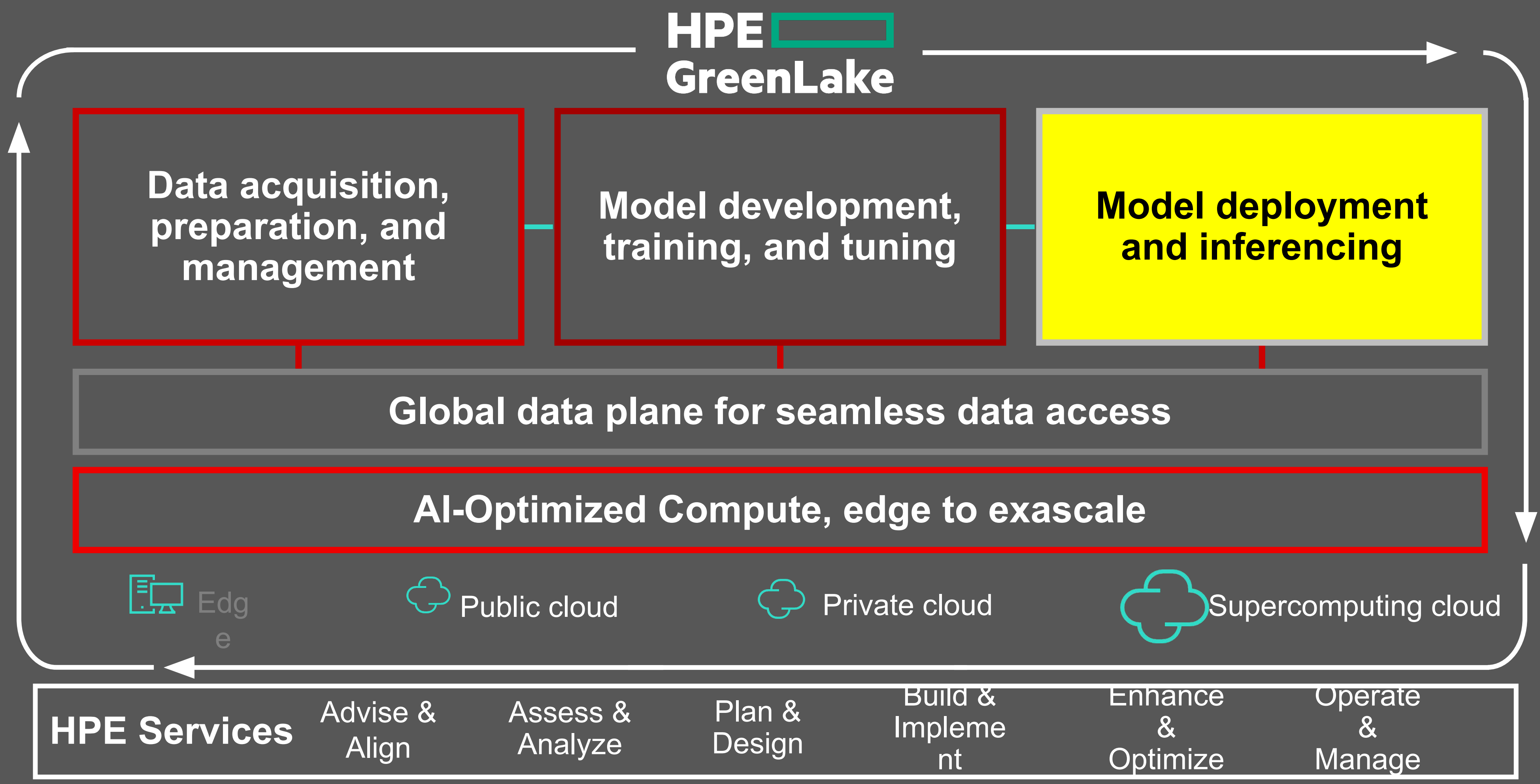
Data
Turn data into intelligence
Unify your data to make smart decisions

Cloud
Create your hybrid cloud
Achieve the cloud experience everywhere

Securit y
Secure your data
Secure your data from edge to cloud

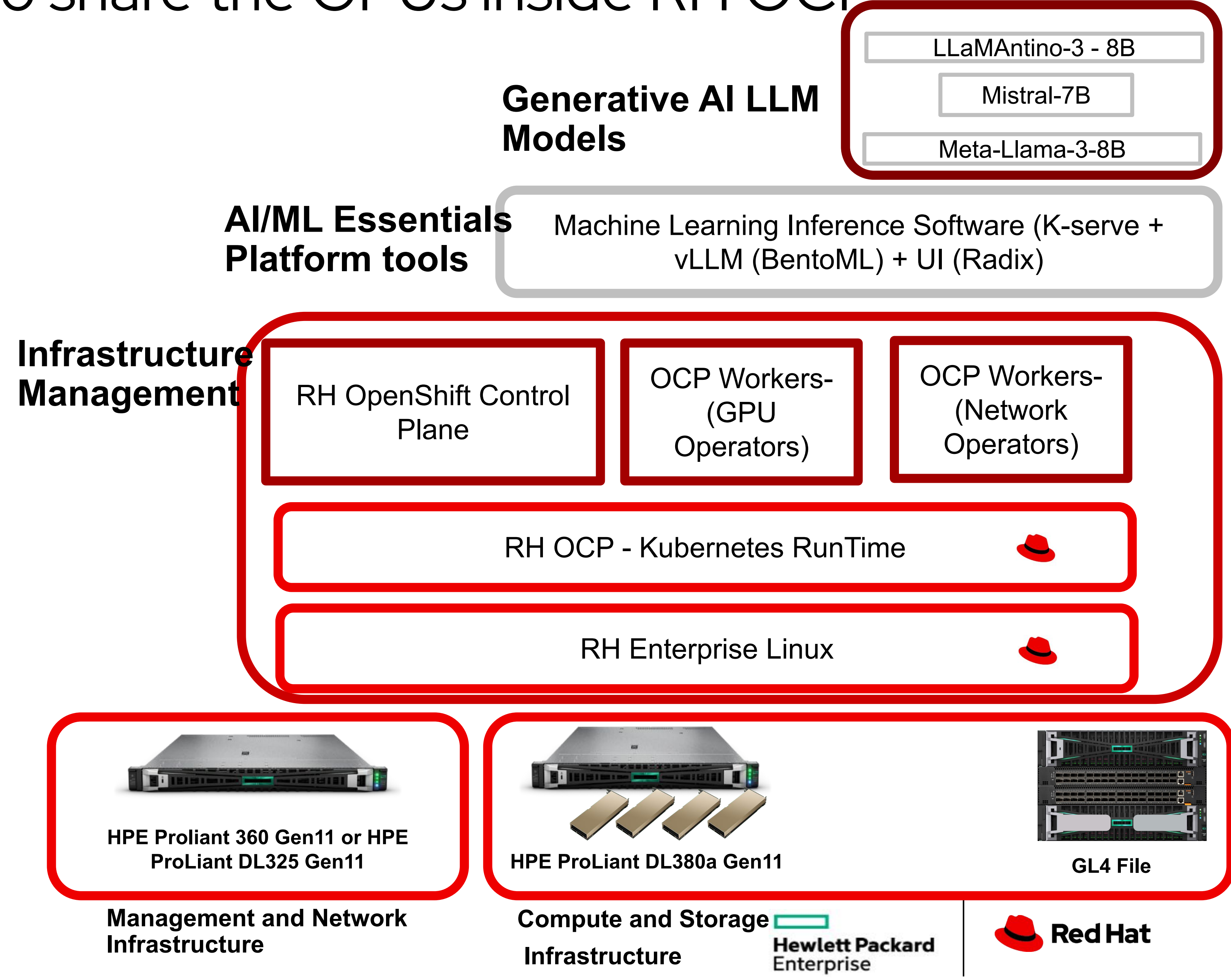
Sustainability
as a catalyst

Unlock your competitive advantage with responsible AI at any scale

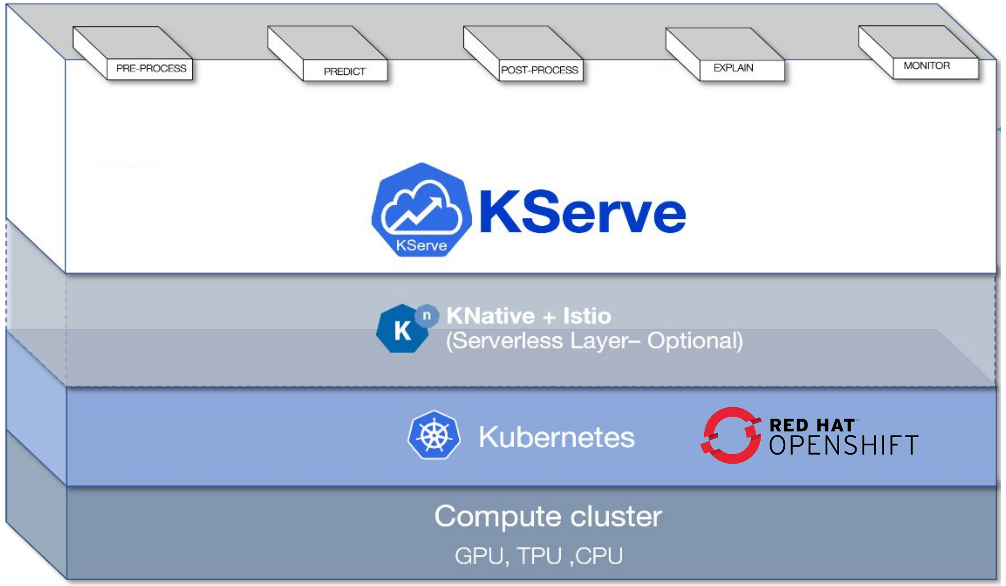


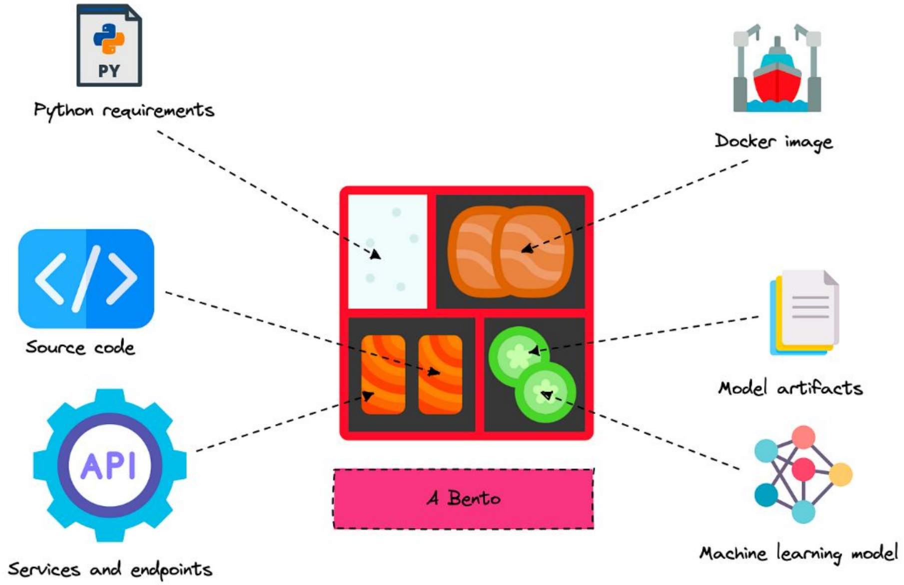
NVIDIA MIG to share the GPUs inside RH OCP

- **HW Compute stack** based on HPE Proliant Gen11 architecture
- **HW Network stack switch** at 100GB/s based on Aruba 8325 Switch for data and 6300 Switch for Mgmt at 1GB/s
- **HW Storage** based on latest GL4File NFS/S3 standard density rack
- **Gen AI models** choice either import from custom model or foundation models trained
- **AI/ML Platform sw tools**: choice of HPE MLIS (Machine Learning Inference Software)
- **Services:** HPE Deployment Inference Startup Service (DIY)



HPE MLIS Open source Components





KServe

- Kubernetes-based platform for deploying models at scale
- Autoscaling, canary rollouts, and batch inferencing capabilities

BentoML

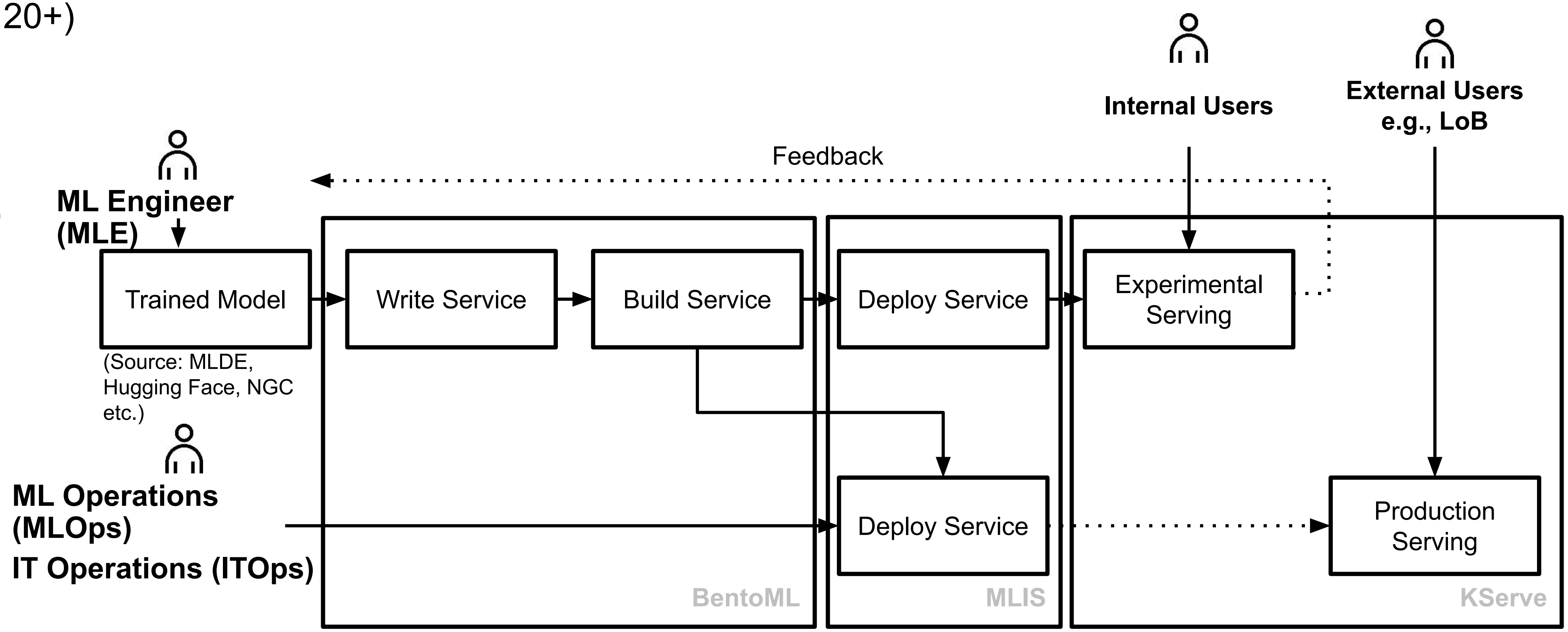
- SDK for standardizing model packaging for services
- Serving standards for REST interfaces, logging, metrics
- **OpenLLM** – Support for optimized LLM deployments
- **vLLM** tackles the bottleneck of slow LLM inference, optimizing performance

HPE Developed software Components

- **UI/UX**
  - Interface for managing and monitoring models, services, deployments, access tokens.
- **Security and authentication**
  - User management
  - Auth integration and access token management
- **Deployment APIs**
  - Reduce Kubernetes deployment friction
  - CLI and Python-native calls
- **Inferencing databases**
  - (Optionally) Capture data predictions
- **Integrated logging**
- **Metrics and Operations**
- **LLM deployment and support**

HPE MLIS Stack

- **Platform**
  - Kubernetes (v1.20+)
  - Helm (v3.0+)
  - Knative
  - Istio
- **Serving**
  - Kserve (v0.11+)
- **Services**
  - BentoML
  - OpenLLM
- **Logging**
  - Loki
- **Metrics**
  - Prometheus
  - Grafana
- **Security**
  - Dex
- **UI**
  - Radix



Hewlett Packard Enterprise | Red Hat

Inference Service Deployment for an LLM in 5 steps: 1) connect to a registry

Connect to your model registry

Select model

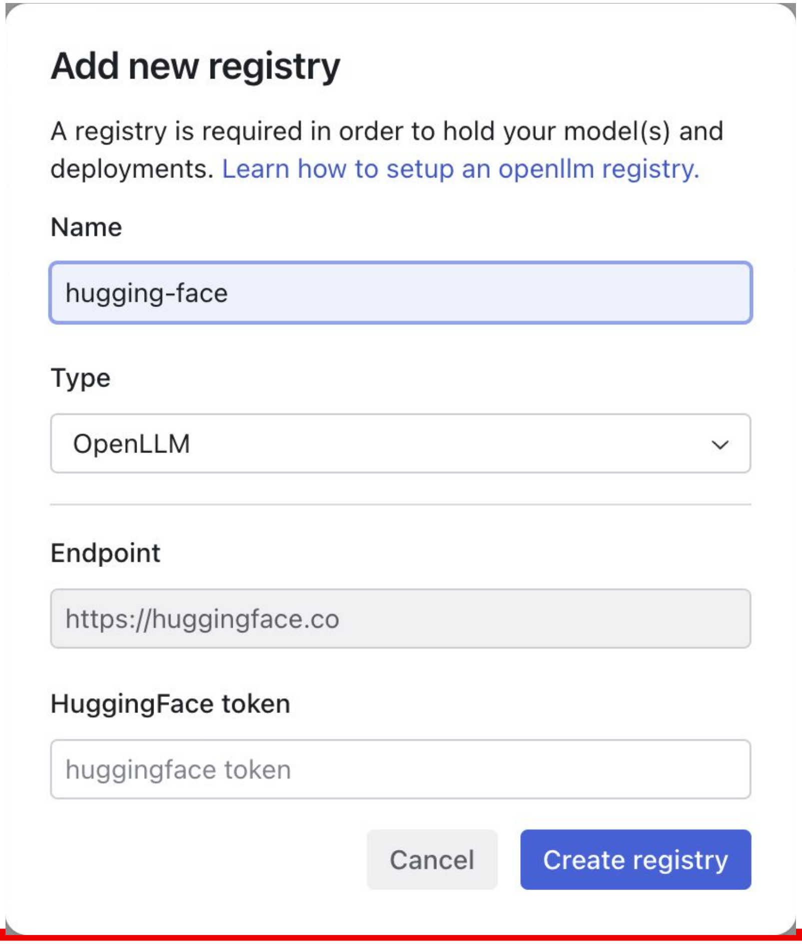
Configure Resources

Configure Scaling

Interact with your model

Connect to existing registries:

- NVIDIA NIM (NGC)
- OpenLLM (Hugging Face)
- AWS S3 Bucket
- Minio Registry



Hewlett Packard Enterprise | Red Hat

Inference Service Deployment for an LLM in 5 steps: 2) select model from a registry

Connect to your model registry

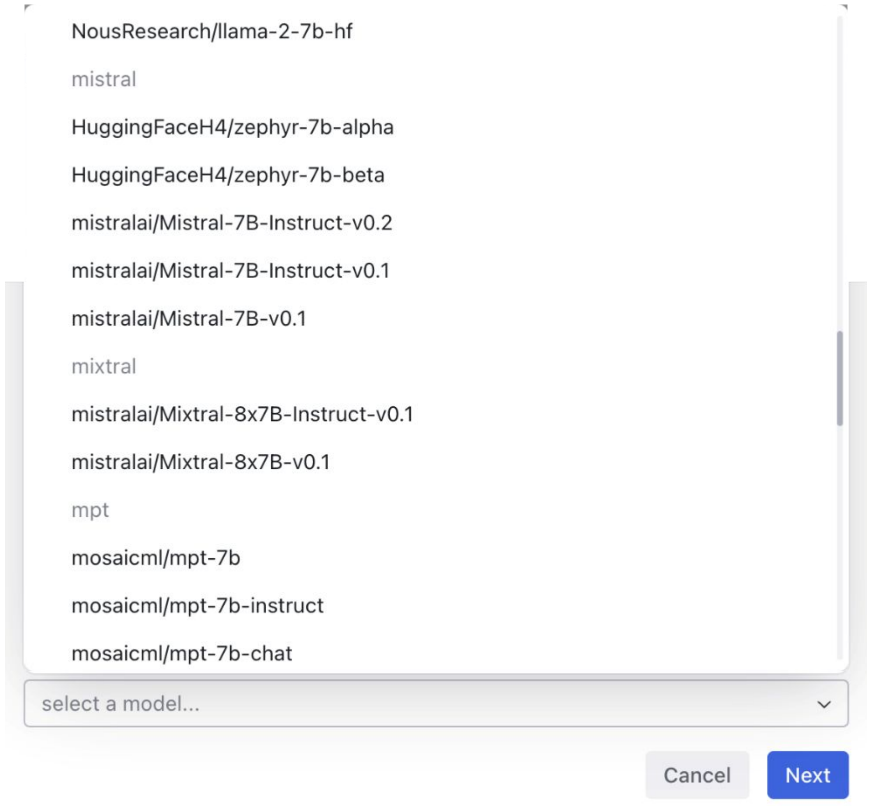
Select model

Configure Resources

Configure Scaling

Interact with your model

Select a model from the registry.



Inference Service Deployment for an LLM in 5 steps: 3) Configure model's resources

Connect to your model registry

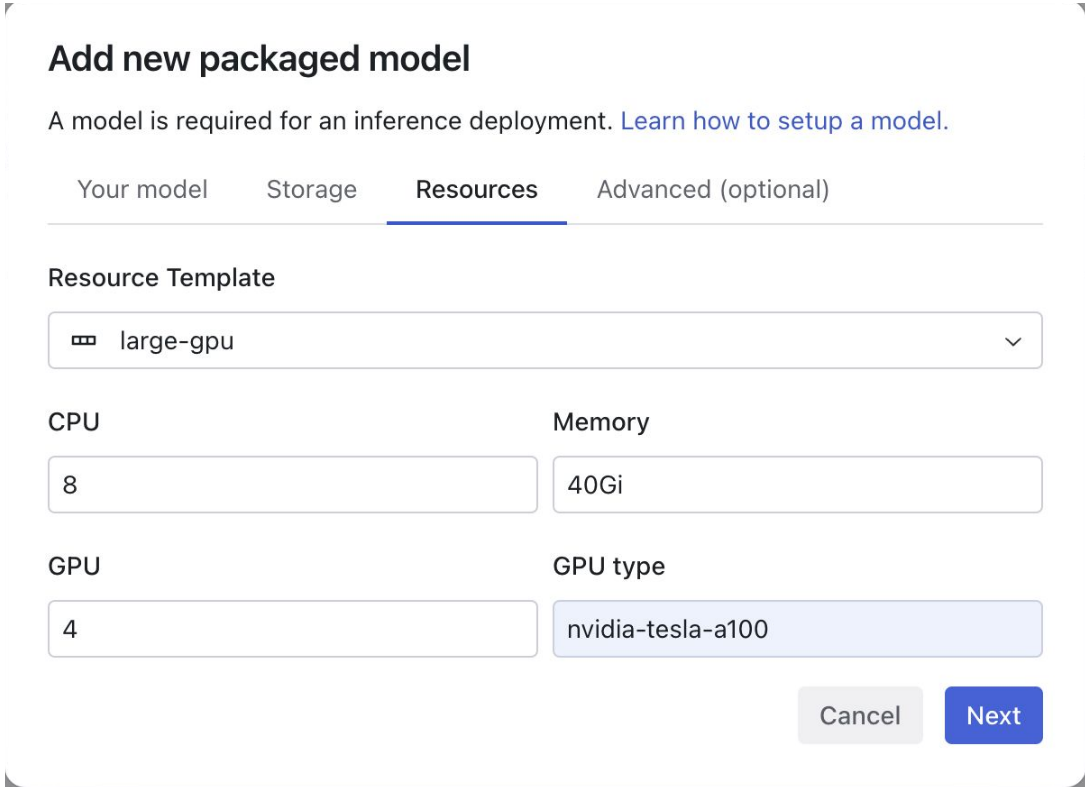
Select model

Configure Resources

Configure Scaling

Interact with your model

Easily configure resources in the UI.



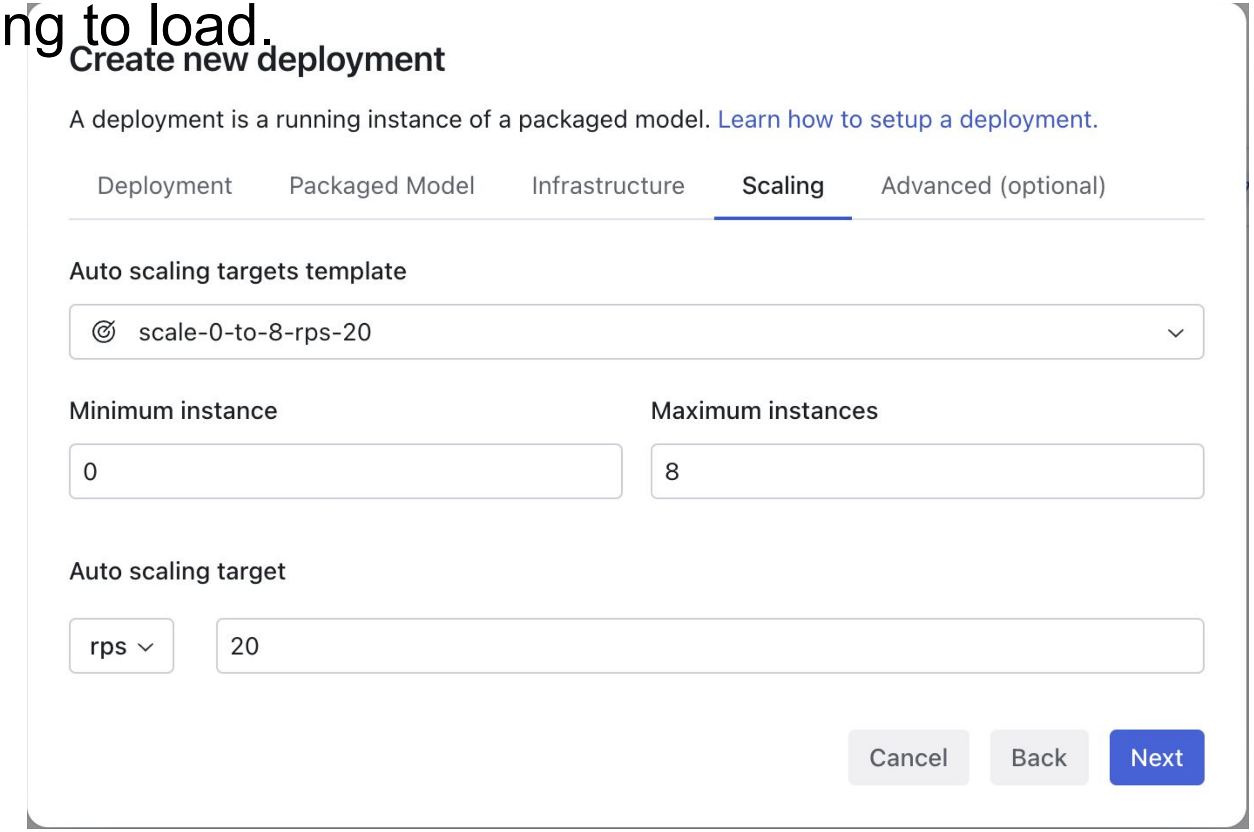
Hewlett Packard Enterprise

Red Hat

Inference Service Deployment for an LLM in 5 steps: 4) Set deployment scaling

- Connect to your model registry
- Select model
- Configure Resources
- **Configure Scaling**
- Interact with your model

Set your deployment to scale according to load.



Hewlett Packard Enterprise | Red Hat

Inference Service Deployment for an LLM in 5 steps: 5) Retrieve model predictions

Connect to your model registry

Select model

Configure Resources

Configure Scaling

Interact with your model

Retrieve model predictions through APIs, CLI, or applications.

```
openllm query --endpoint http://chat-bot-2.default.34.41.176.77.nip.io
"Are you an llm?"
Are you an llm?

I apologize for the confusion, but I am a machine learning model (LLM) designed to a
ssist with a variety of tasks, including answering questions and providing informati
on. I do not have the ability to physically perform tasks or engage in physical acti
vities. My primary function is to provide helpful responses to the best of my abilit
y based on the information and context provided to me. Is there anything specific yo
u would like to know or discuss?%
```

OpenAI OpenAI Compatible API support

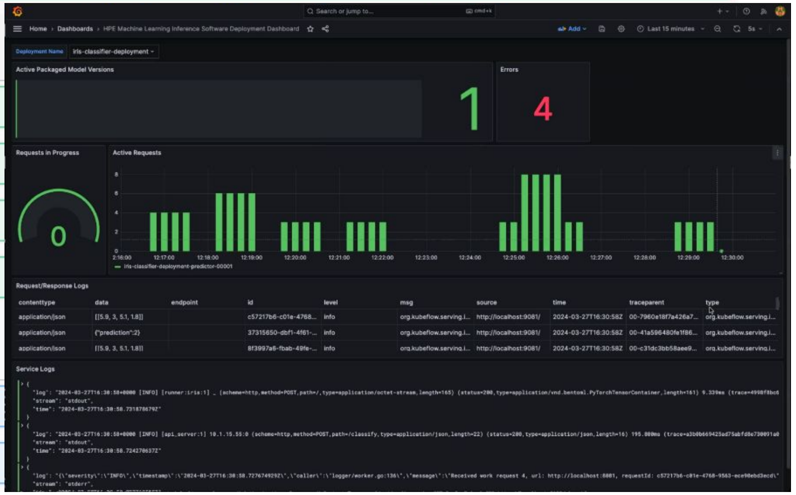
- GET `/v1/models` Describes a model offering that can be us
- POST `/v1/completions` Creates a completion for the pro
- POST `/v1/chat/completions` Creates a model response for the given chat conversation.

Service APIs BentoML Service API endpoints for inference.

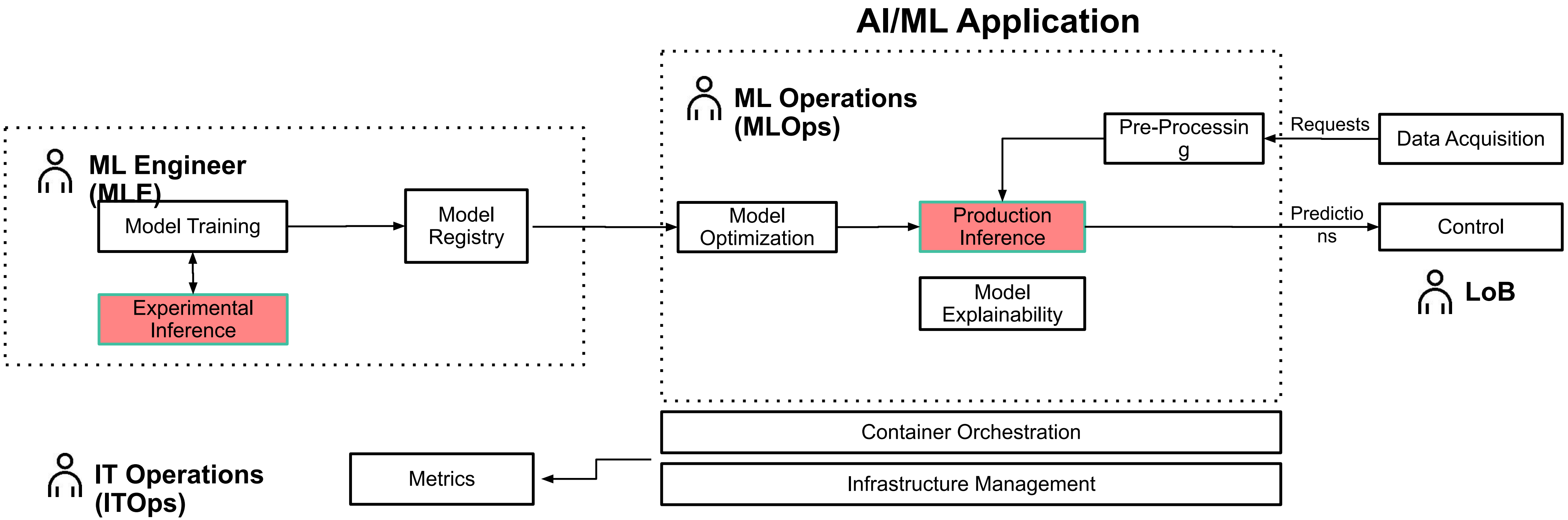
- POST `/v1/generate` InferenceAPI(JSON → JSON)
- POST `/v1/generate_stream` InferenceAPI(JSON → Text)
- POST `/v1/metadata` InferenceAPI(Text → JSON)
- POST `/v1/helpers/messages` InferenceAPI(JSON → Text)

Infrastructure Common infrastructure endpoints for observability.

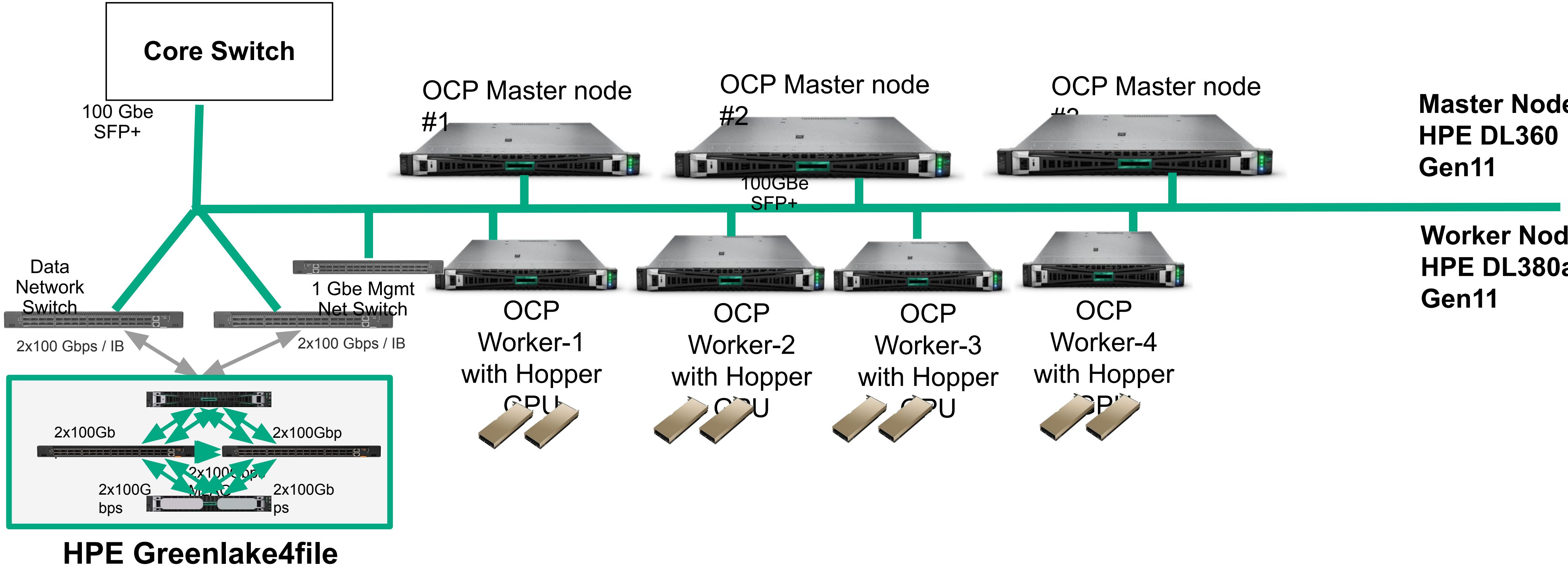
- GET `/healthz`
- GET `/livez`



Deploying AI/ML Models into Production



Example of physical Architecture with RH Openshift



Master Node: HPE DL360 Gen11

Worker Node: HPE DL380a Gen11

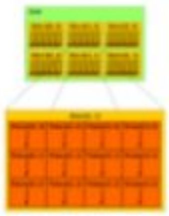
Hewlett Packard Enterprise

Red Hat

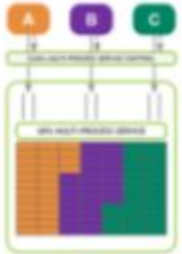
NVIDIA GPUs Concurrency choices

GPU “CONCURRENCY”

Choices



Single Process in CUDA



Multi-Process with CUDA MPS

Time-slicing

MIG

Virtualization with vGPU

Application level (using the CUDA programming model APIs - CUDA streams)	GPU System Software / Hardware (Mostly transparent to CUDA applications)

NVIDIA GPUs Concurrency choices

	Streams	MPS	Time-Slicing	MIG	vGPU
Partition Type	Single process	Logical	Temporal (Single process)	Physical	Temporal & Physical – VMs
Max Partitions	Unlimited	48	Unlimited	7	Variable
SM Performance Isolation	No	Yes (by percentage, not partitioning)	Yes	Yes	Yes
Memory Protection	No	Yes	Yes	Yes	Yes
Memory Bandwidth QoS	No	No	No	Yes	Yes
Error Isolation	No	No	Yes	Yes	Yes
Cross-Partition Interop	Always	IPC	Limited IPC	Limited IPC	No
Reconfigure	Dynamic	At process launch	N/A	When idle	N/A
GPU Management (telemetry)	N/A	Limited GPU metrics	N/A	Yes – GPU metrics, support for containers	Yes – live migration and other industry virtualization tools
Target use cases (and when to use each)	Optimize for concurrency within a single application	Run multiple applications in parallel but can deal with limited resiliency	Run multiple applications that are not latency-sensitive or can tolerate jitter	Run multiple applications in parallel but need resiliency and QoS	Support multi-tenancy on the GPU through virtualization and need VM management benefits

Red Hat Summit

Connect

Q&A

Hewlett Packard Enterprise | Red Hat

Red Hat Summit

Connect

Thank you

Hewlett Packard Enterprise | Red Hat