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

Milano, November 19th 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

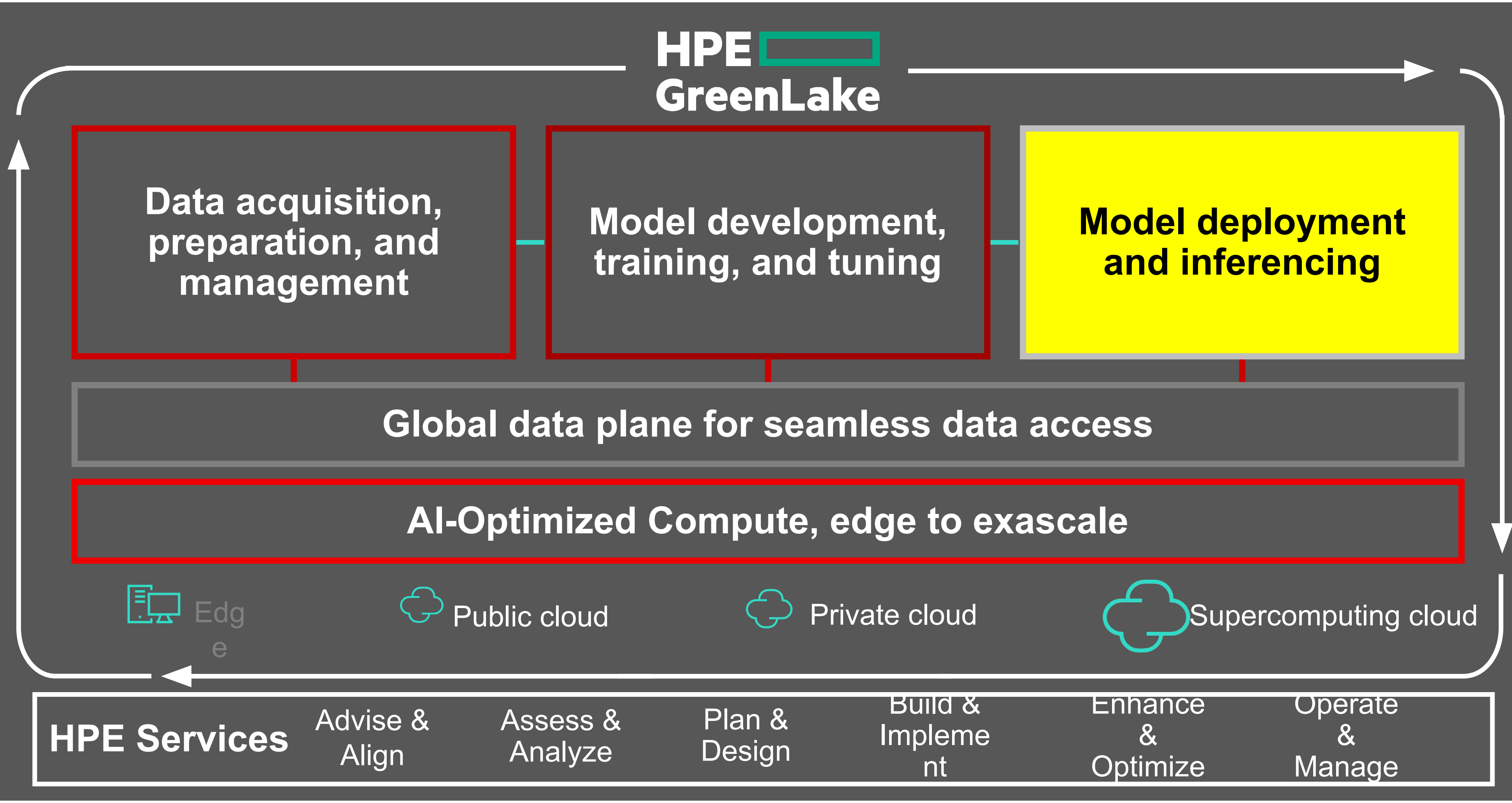
Secure your data from edge to cloud

Sustainability

as a catalyst

Hewlett Packard Enterprise | Red Hat

Unlock your competitive advantage with responsible AI at any scale



HPE GreenLake

Data acquisition, preparation, and management

Model development, training, and tuning

Model deployment and inferencing

Global data plane for seamless data access

AI-Optimized Compute, edge to exascale

Edge

Public cloud

Private cloud

Supercomputing cloud

HPE Services

Advise & Align

Assess & Analyze

Plan & Design

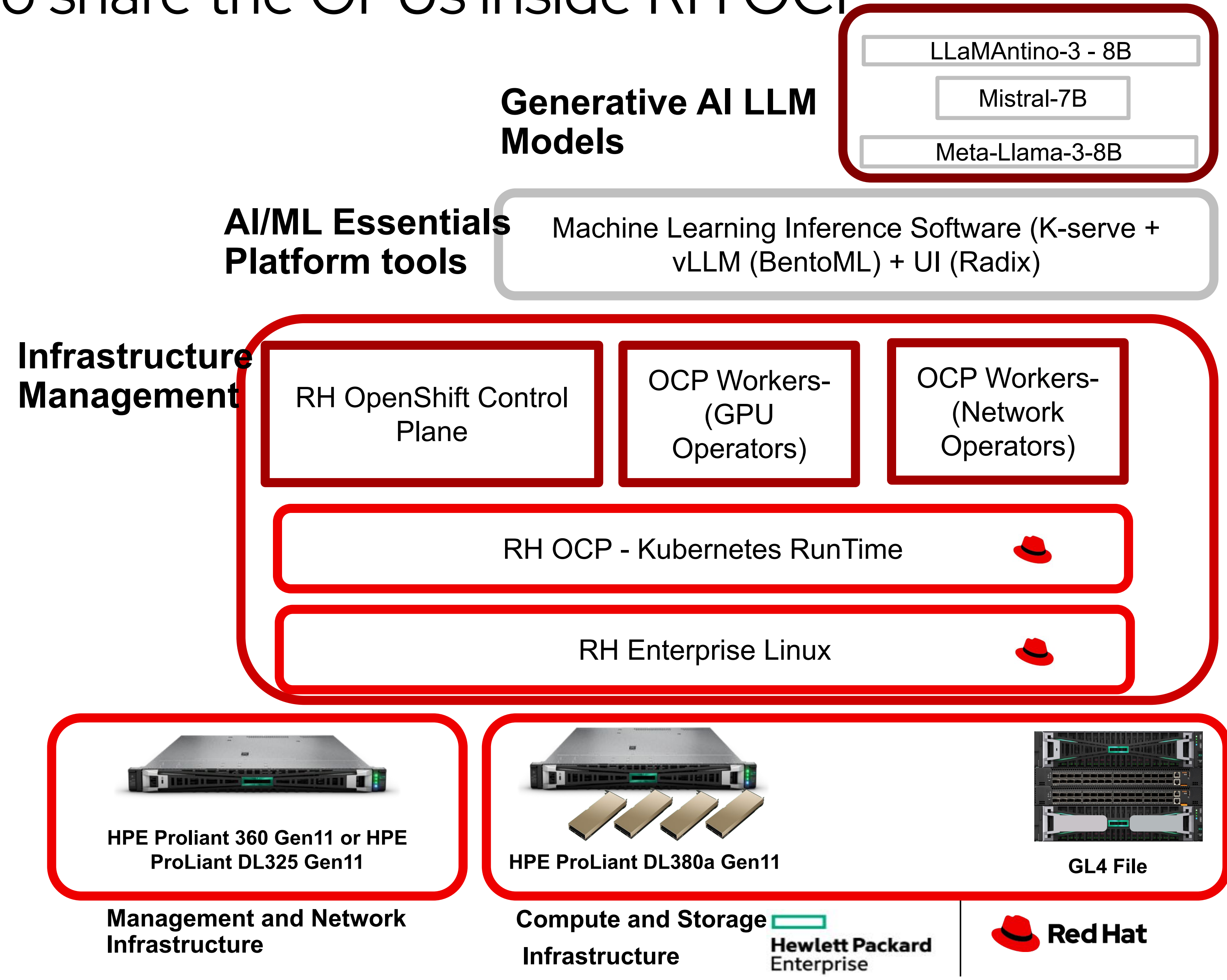
Build & Implement

Enhance & Optimize

Operate & Manage

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)



Generative AI LLM Models

LLaMAntino-3 - 8B

Mistral-7B

Meta-Llama-3-8B

AI/ML Essentials Platform tools

Machine Learning Inference Software (K-serve + vLLM (BentoML) + UI (Radix)

Infrastructure Management

RH OpenShift Control Plane

OCP Workers-(GPU Operators)

OCP Workers-(Network Operators)

RH OCP - Kubernetes RunTime

RH Enterprise Linux

HPE Proliant 360 Gen11 or HPE ProLiant DL325 Gen11

Management and Network Infrastructure

HPE ProLiant DL380a Gen11

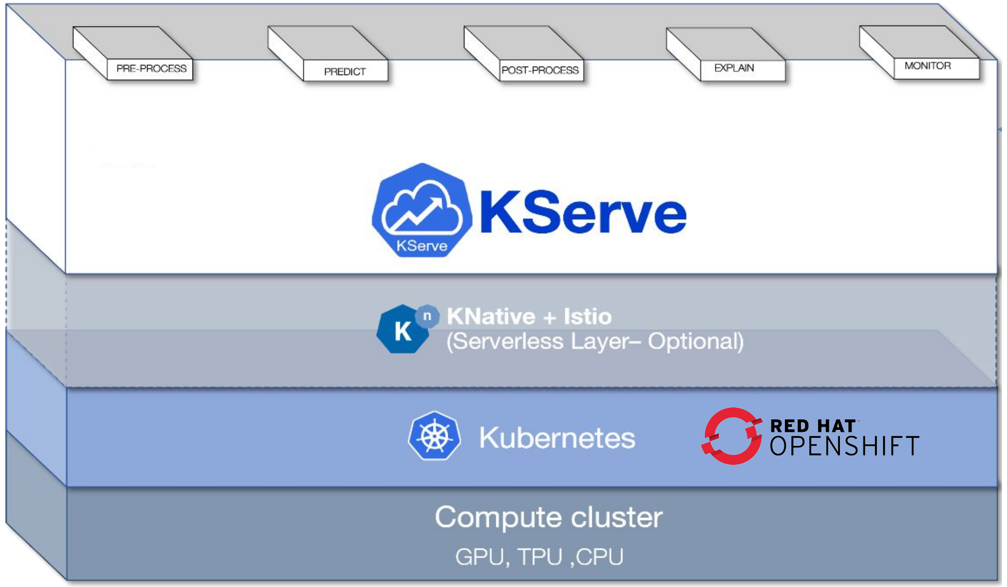
GL4 File

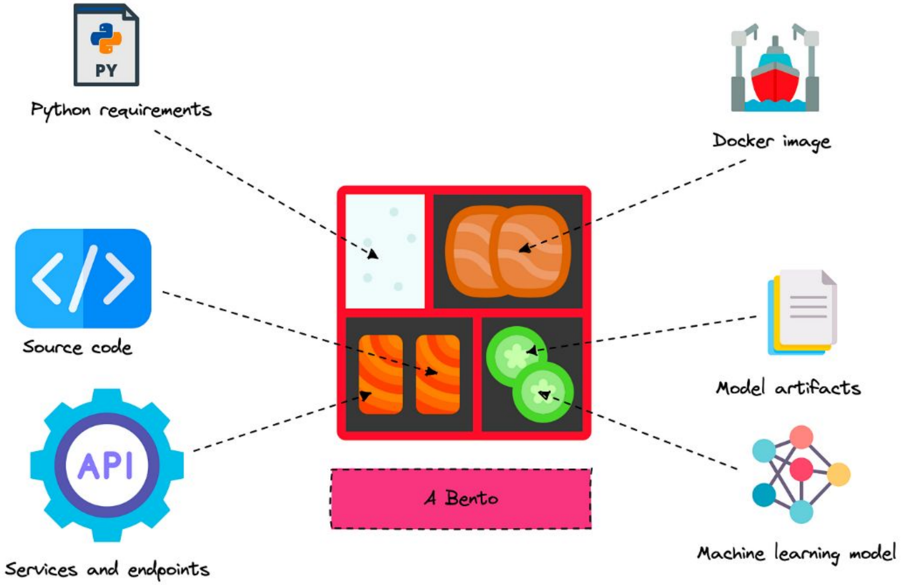
Compute and Storage Infrastructure

Hewlett Packard Enterprise

Red Hat

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

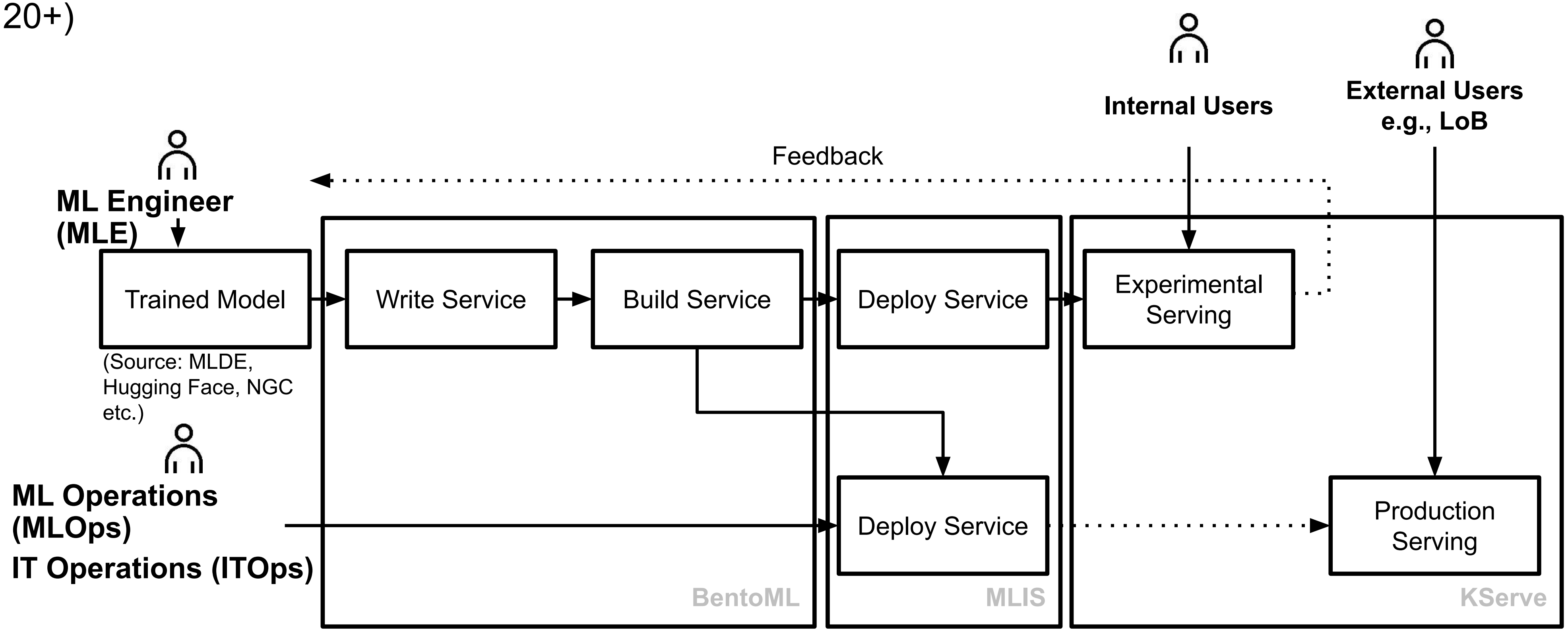
Hewlett Packard Enterprise | Red Hat

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

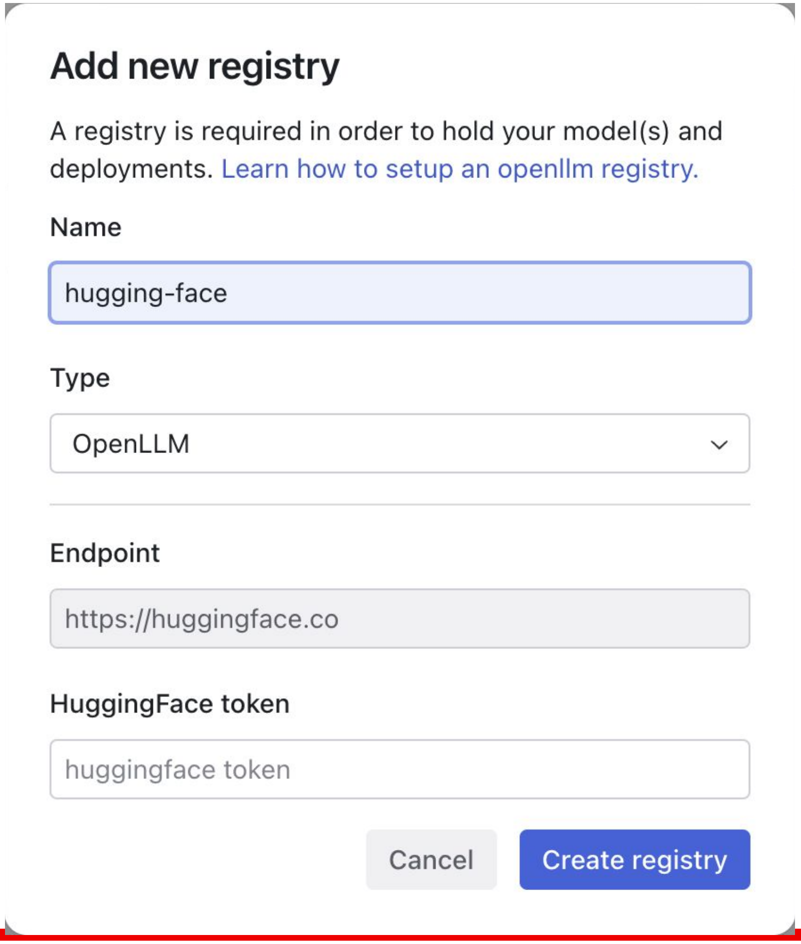
Configure Resources

Configure Scaling

Interact with your model

Connect to existing registries:

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



Add new registry

A registry is required in order to hold your model(s) and deployments. Learn how to setup an openllm registry.

Name: hugging-face

Type: OpenLLM

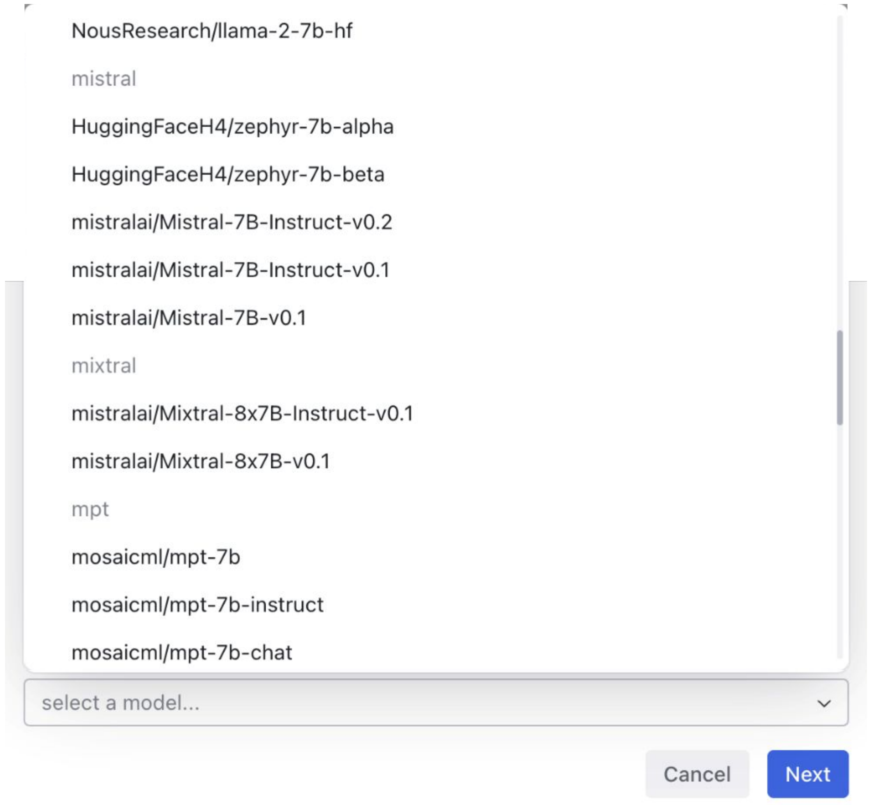
Endpoint: https://huggingface.co

HuggingFace token: huggingface token

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.



NousResearch/llama-2-7b-hf

mistral

HuggingFaceH4/zephyr-7b-alpha

HuggingFaceH4/zephyr-7b-beta

mistralai/Mistral-7B-Instruct-v0.2

mistralai/Mistral-7B-Instruct-v0.1

mistralai/Mistral-7B-v0.1

mixtral

mistralai/Mixtral-8x7B-Instruct-v0.1

mistralai/Mixtral-8x7B-v0.1

mpt

mosaicml/mpt-7b

mosaicml/mpt-7b-instruct

mosaicml/mpt-7b-chat

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

Connect to your model registry

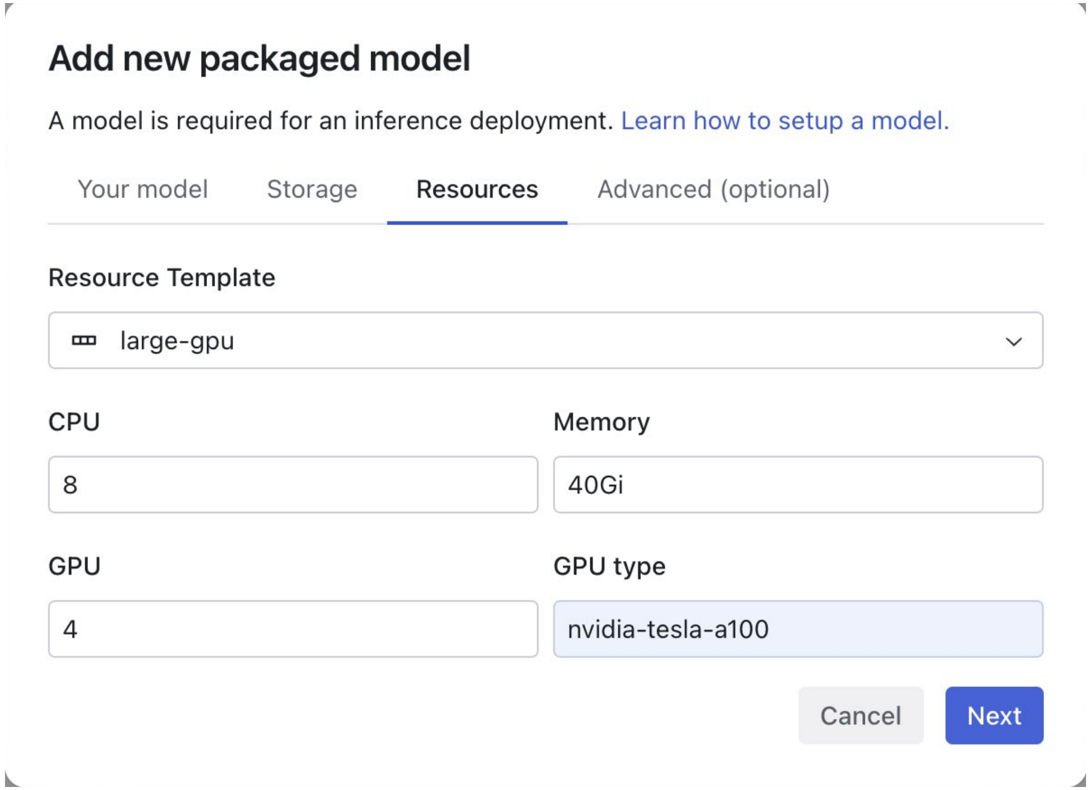
Select model

Configure Resources

Configure Scaling

Interact with your model

Easily configure resources in the UI.



Add new packaged model

A model is required for an inference deployment. Learn how to setup a model.

Resource Template: large-gpu

CPU: 8

Memory: 40Gi

GPU: 4

GPU type: nvidia-tesla-a100

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

Connect to your model registry

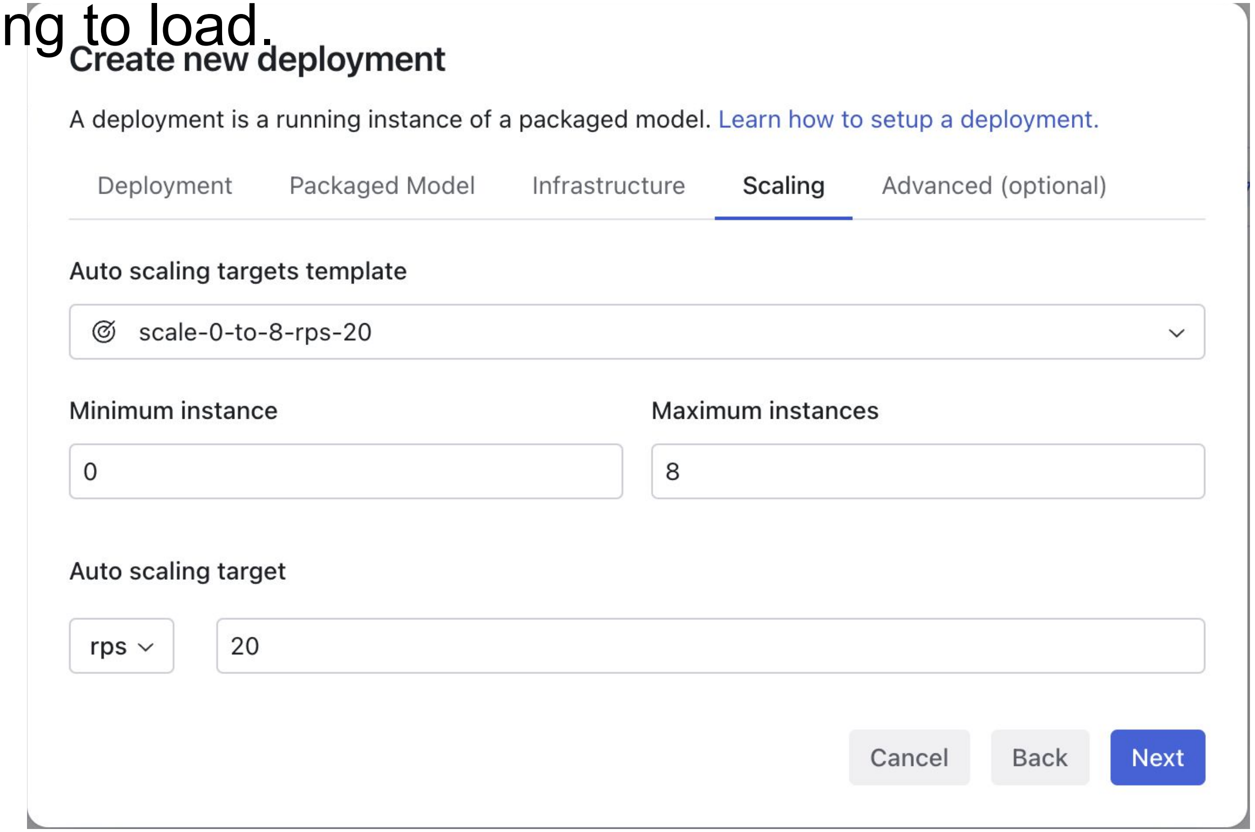
Select model

Configure Resources

Configure Scaling

Interact with your model

Set your deployment to scale according to load.



Create new deployment

A deployment is a running instance of a packaged model. Learn how to setup a deployment.

Auto scaling targets template: scale-0-to-8-rps-20

Minimum instance: 0

Maximum instances: 8

Auto scaling target: rps 20

Hewlett Packard Enterprise | Red Hat

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

Connect to your model registry

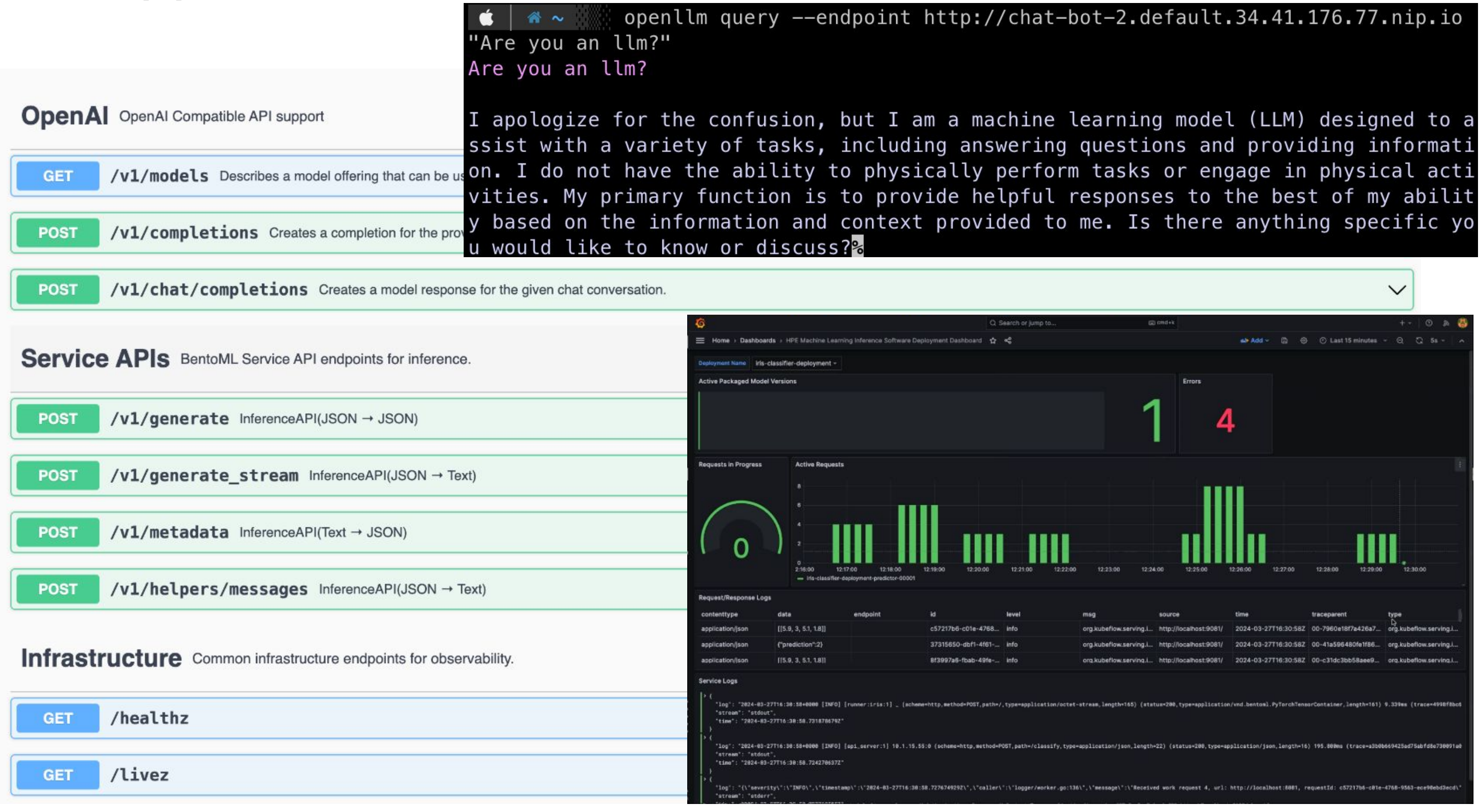
Select model

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 assist with a variety of tasks, including answering questions and providing information. I do not have the ability to physically perform tasks or engage in physical activities. My primary function is to provide helpful responses to the best of my ability based on the information and context provided to me. Is there anything specific you 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 prov
- 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`

Home › Dashboards › HPE Machine Learning Inference Software Deployment Dashboard

Deployment Name: iris-classifier-deployment

Active Packaged Model Versions: 1

Errors: 4

Requests in Progress: 0

Request/Response Logs

contenttype	data	endpoint	id	level	msg	source	time	traceparent	type
application/json	[[5.9, 3, 5.1, 1.8]]		c57217b6-c01e-4768...	info	org.kubeflow.serving.i...	http://localhost:9081/	2024-03-27T16:30:58Z	00-7960e18f7a426a7...	org.kubeflow.serving.i...
application/json	{"prediction":2}		37315650-dbf1-4f61-...	info	org.kubeflow.serving.i...	http://localhost:9081/	2024-03-27T16:30:58Z	00-41a596480fe1f86...	org.kubeflow.serving.i...
application/json	[[5.9, 3, 5.1, 1.8]]		8f3997a6-fbab-49fe-...	info	org.kubeflow.serving.i...	http://localhost:9081/	2024-03-27T16:30:58Z	00-c31dc3bb58aee9...	org.kubeflow.serving.i...

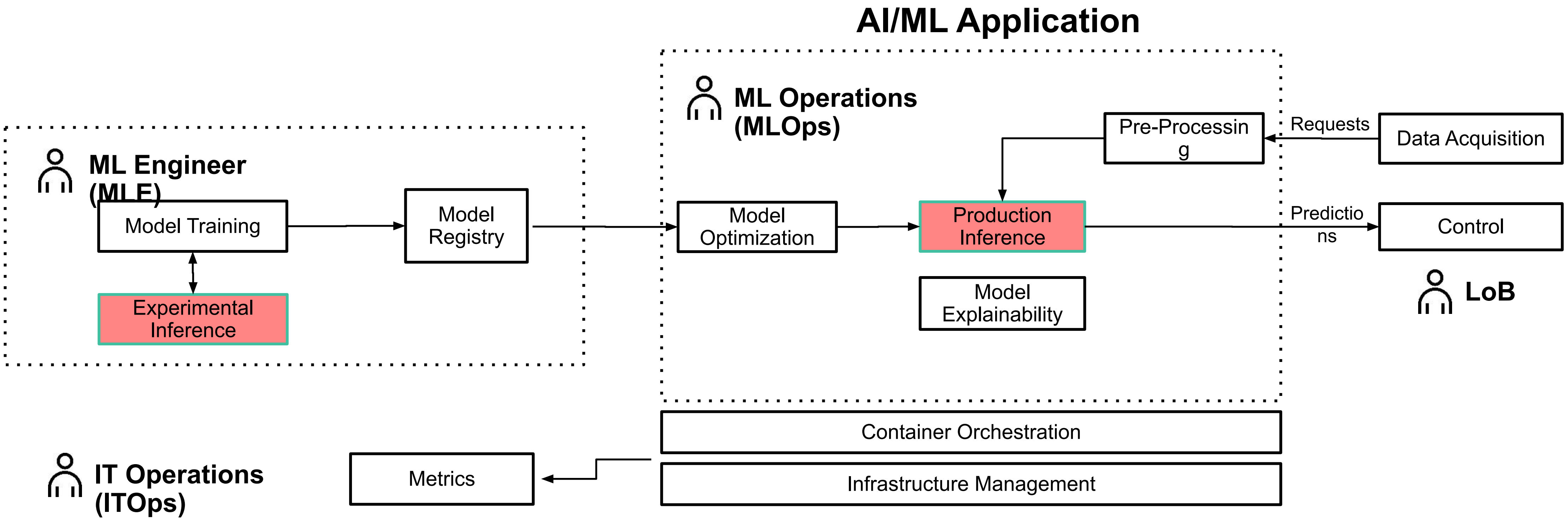
Service Logs

```
> {
    "log": "2024-03-27T16:30:58+0000 [INFO] [runner:iris:1] _ (scheme=http,method=POST,path=/,type=application/octet-stream,length=165) (status=200,type=application/vnd.bentoml.PyTorchTensorContainer,length=161) 9.339ms (trace=4998f8bc6
    "stream": "stdout",
    "time": "2024-03-27T16:30:58.731878679Z"
  }
> {
    "log": "2024-03-27T16:30:58+0000 [INFO] [api_server:1] 10.1.15.55:0 (scheme=http,method=POST,path=/classify,type=application/json,length=22) (status=200,type=application/json,length=16) 195.800ms (trace=a3b0b669425ad75abfd8e730091a0
    "stream": "stdout",
    "time": "2024-03-27T16:30:58.724270637Z"
  }
> {
    "log": "{\"severity\":\"INFO\",\"timestamp\":\"2024-03-27T16:30:58.727674929Z\",\"caller\":\"logger/worker.go:136\",\"message\":\"Received work request 4, url: http://localhost:8081, requestId: c57217b6-c01e-4768-9563-ece98ebd3ecd\"
    "stream": "stderr",
```

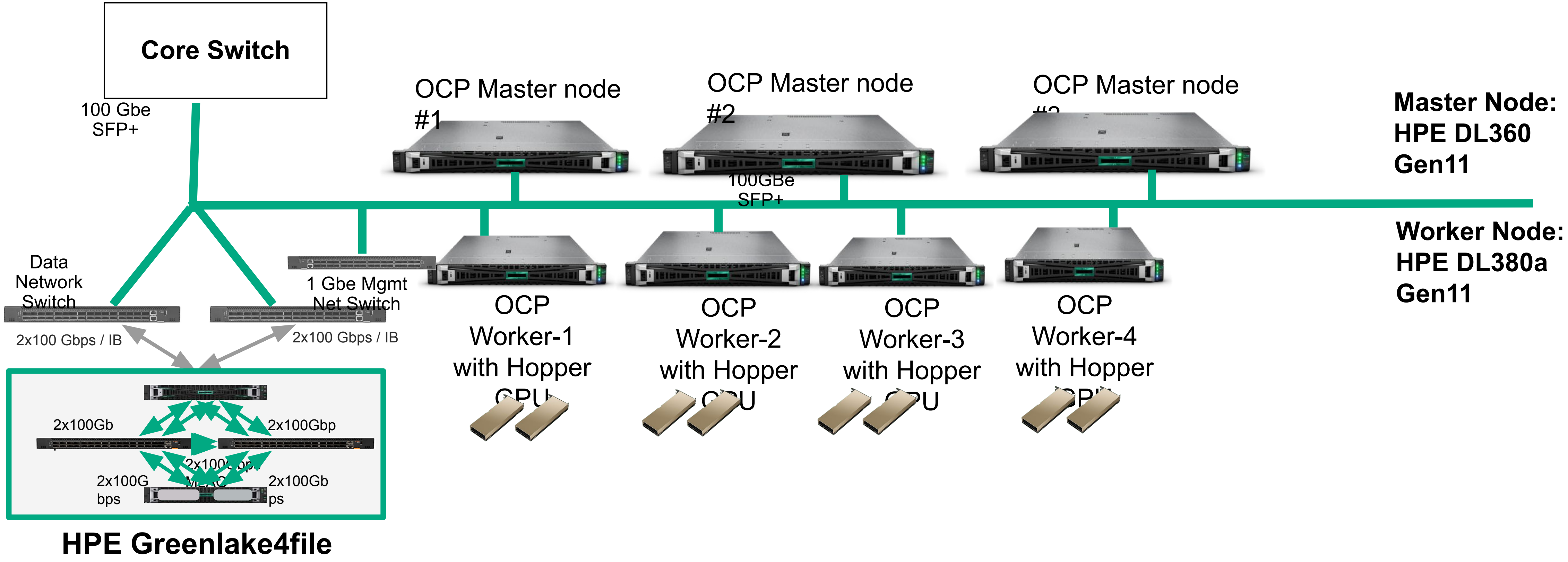
Hewlett Packard Enterprise

Red Hat

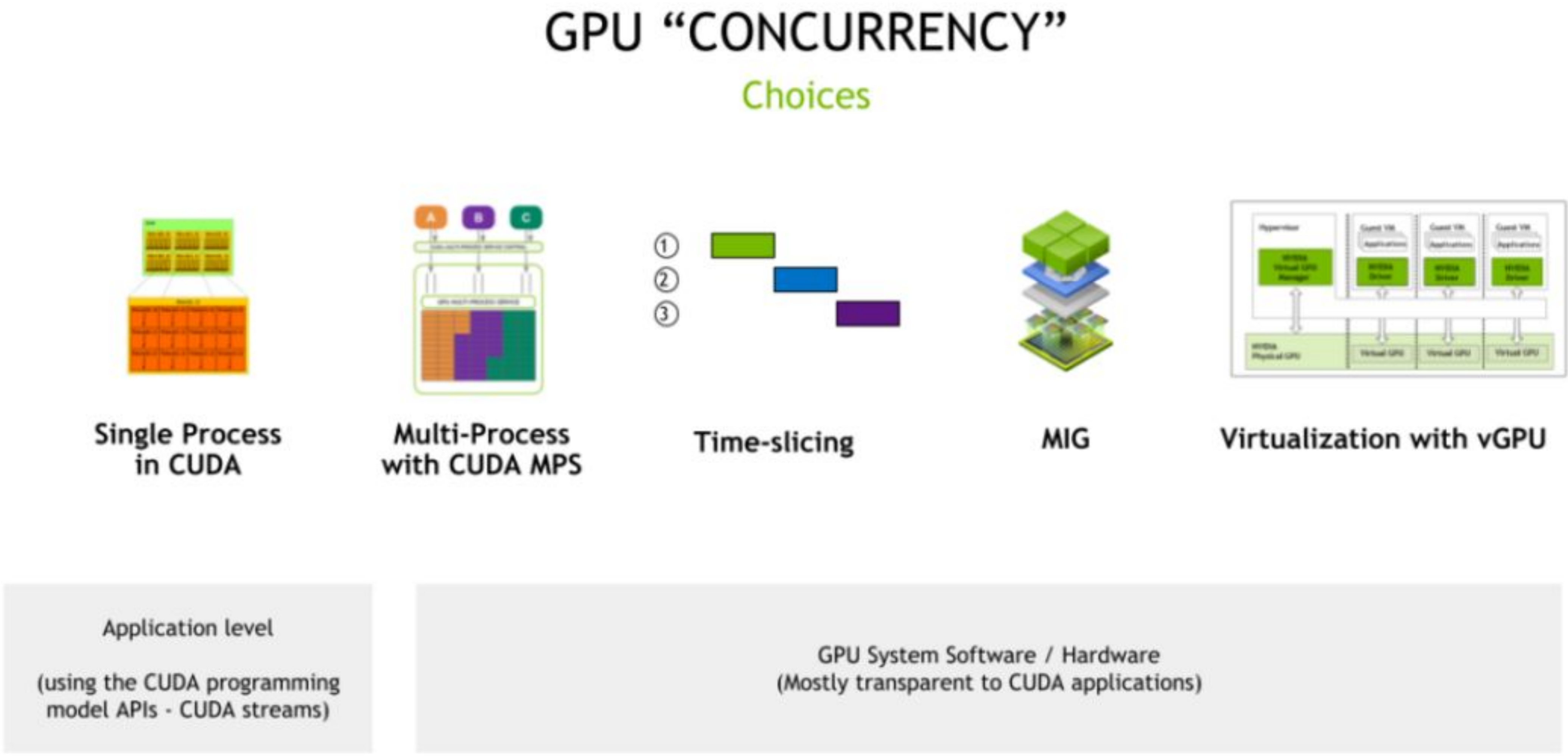
Deploying AI/ML Models into Production



Example of physical Architecture with RH Openshift



NVIDIA GPUs Concurrency choices



GPU “CONCURRENCY”

Choices

Single Process in CUDA

Multi-Process with CUDA MPS

Time-slicing

MIG

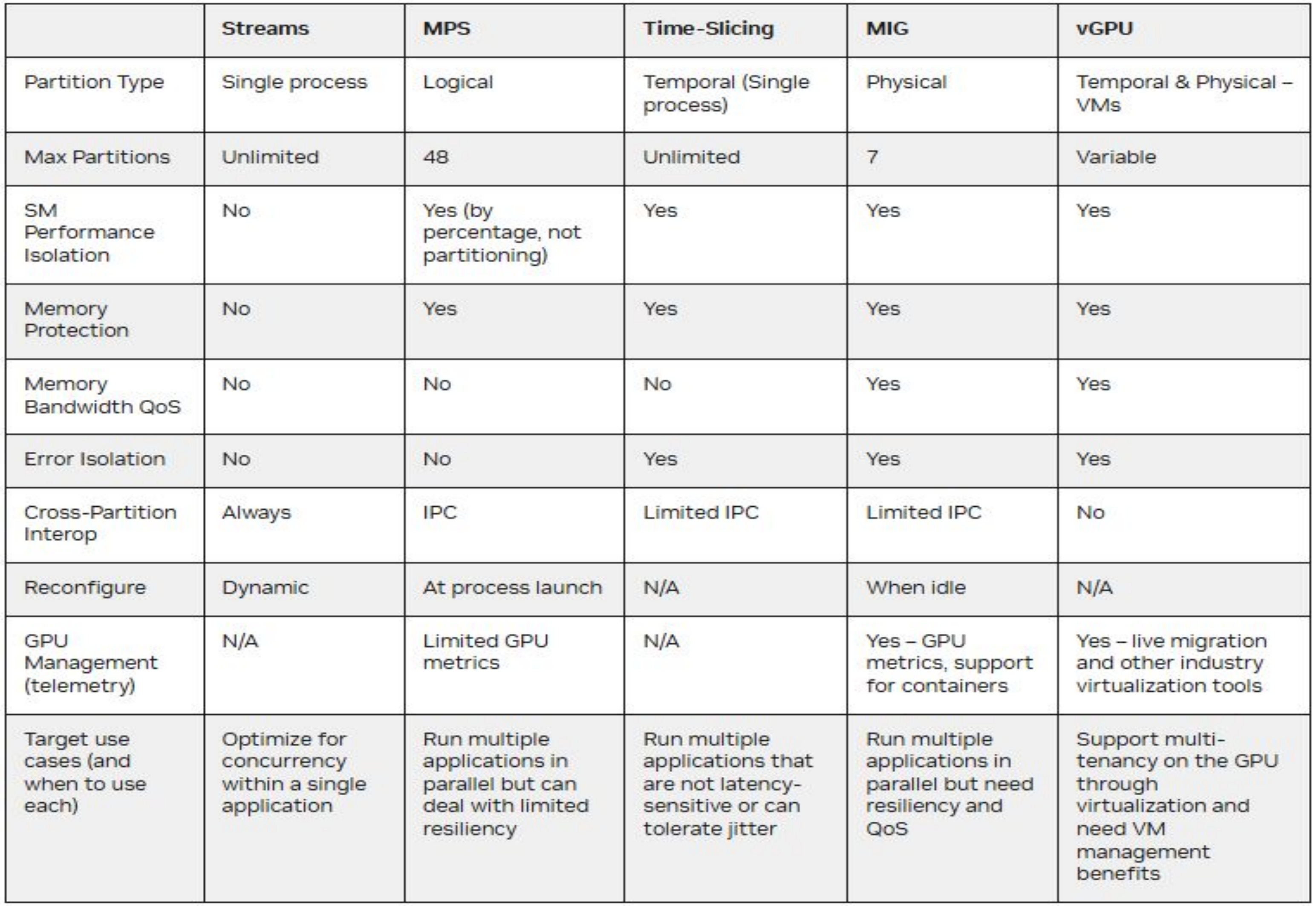
Virtualization with vGPU

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