

INFN

**National Institute for Nuclear Physics
Italy**



Kubernetes technologies for ML-based solutions

Mauro Gattari (AC)
DSI/DataCloud - mgattari@infn.it

Workshop
Computing@CNS5
10/2024

Agenda

Intro: Kubernetes

Inference: KServe

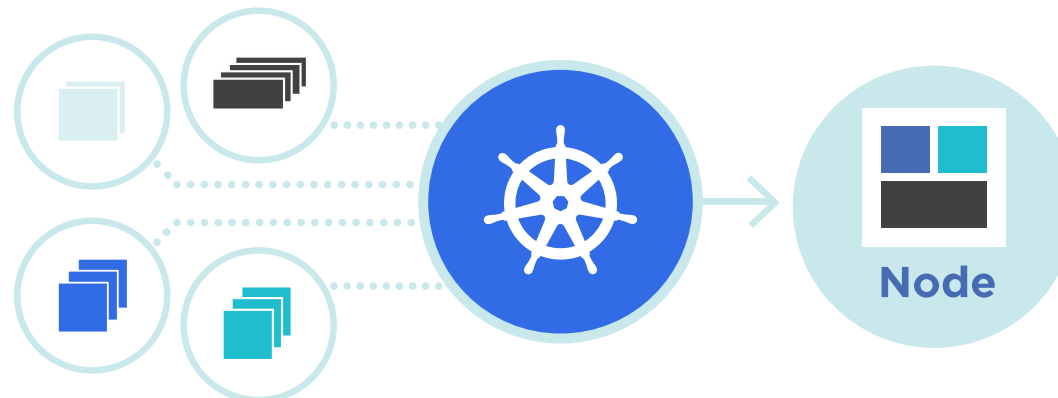
Streaming and Offloading

Distributed Training: Kubeflow Training Operator

INFN Cloud Integration

Kubernetes

- **Open source** technology (since 2014) for running **containerized** applications **at scale**.
- Kubernetes manage containers in a cluster, providing features such as:
 - **Service Discovery**: networking layer that allows containers to find and communicate with each other.
 - **Load Balancing**: distributing traffic between containers.
 - **Scaling**: automatically scaling the number of running containers based on resources utilization.
 - **Self-Healing**: monitoring and restarting of failed containers.
 - **Automated Rollouts and Rollbacks**: deploying new versions of applications and reverting to old versions in case of failure.



- Kubernetes-based **open source** platform for Machine Learning **models serving** at scale (first release 2019).
 - Serving a single model can require multiple GPUs across different nodes. E.g., Meta's **Llama 405B** requires 11 NVidia H100 80GB GPUs (in 16-bit half-precision).
- KServe supports **multiple ML frameworks**, including TensorFlow, PyTorch, XGBoost, Scikit-Learn, and others.
 - This allows you to serve models built with different frameworks using a unified platform.
- Provide numerous features, e.g.:
 - Uniform APIs across ML frameworks.
 - Handle model versioning.
 - Scaling to zero and scaling based on concurrent requests.
 - Batching of incoming requests.
 - Canary rollouts.



KServe - Use Cases

Text Classification

Framework: **PyTorch**

Task: infer the INFN structure name given an author's affiliation string:

- "INFN Frascati Natl Labs, I-00044 Frascati, Roma" -> **LNF**
- "INFN Sez, Lab Nazl Frascati, Rome" -> **LNF**
- "Ist Nazl Fis Nucl, LNF, Via E Fermi, Roma" -> **LNF**
- "INFN Bari, Dept Phys, Bari, Italy" -> **BA**
- "INFN Natl Inst Nucl Phys, Bari Div, Bari" -> **BA**

HEP Analysis

Frameworks: **TensorFlow/Scikit-learn**

- *ttH* analysis in the boosted, all-hadronic final states

This model discriminates $t\bar{t}H(\bar{b}b)$ events with all-jets final state, where at least one of the jets of the final state is a boosted jet, and where the Higgs boson decays in a pair of well resolved jets identified as a result of the hadronization of bottom quarks.

- The Higgs boson ML challenge:

This model allows to face the Higgs boson machine learning challenge organized by a small group of ATLAS physicists and data scientists, hosted by Kaggle in 2014.

GenAI

Serving of open-source **LLMs** to implement a ChatBot configured to answer questions (text generation) about a private knowledge base (RAG – Retrieval Augmented Generation)



How can I help you today?

Ask question...



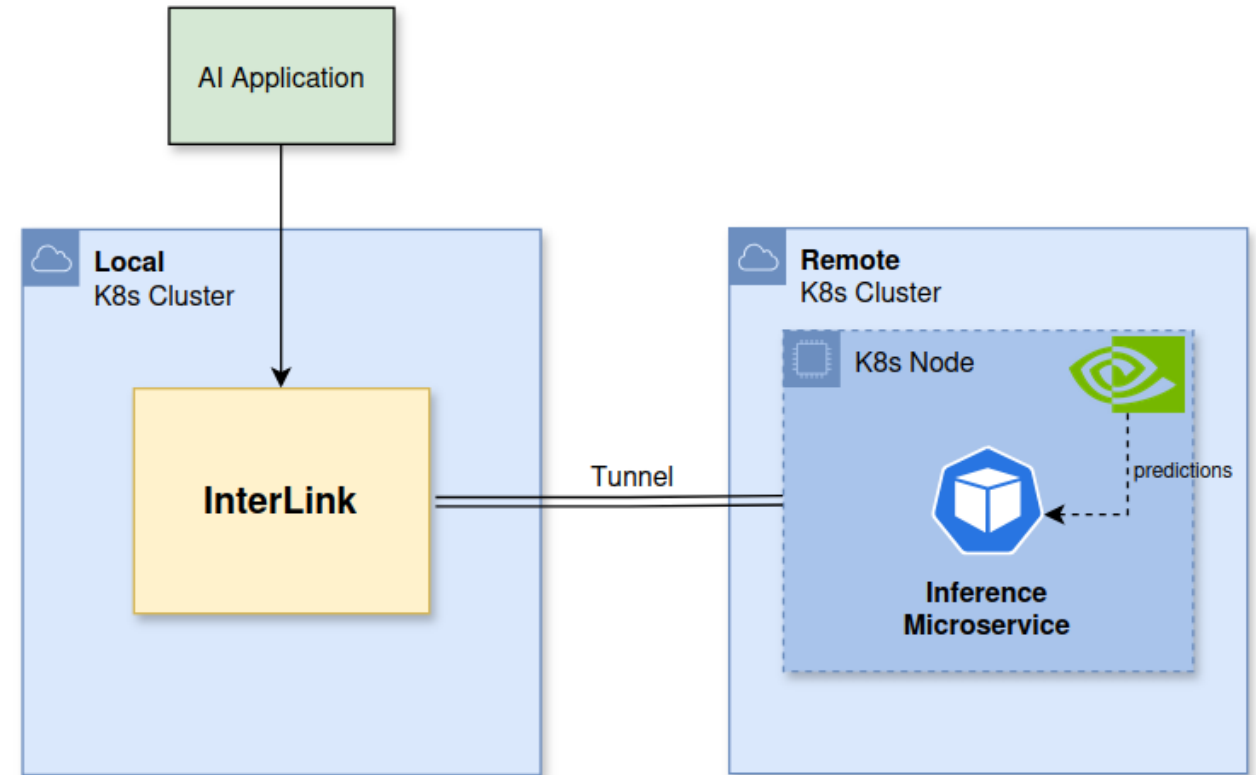
KServe + Offloading

(Work In Progress)

- **InterTwin:** project funded by the EU for the development of an open source platform, called Digital Twin Engine (DTE), to handle "digital twins" of selected scientific communities.
- **InterLink:** transparent offloading of resources to heterogeneous computing providers

Inference Microservice:

- specify requirements, e.g., n GPUs;
- resources may not be available on local cluster;
- service can be opportunistically offloaded to a remote cluster where resources are available.



KServe + Kafka

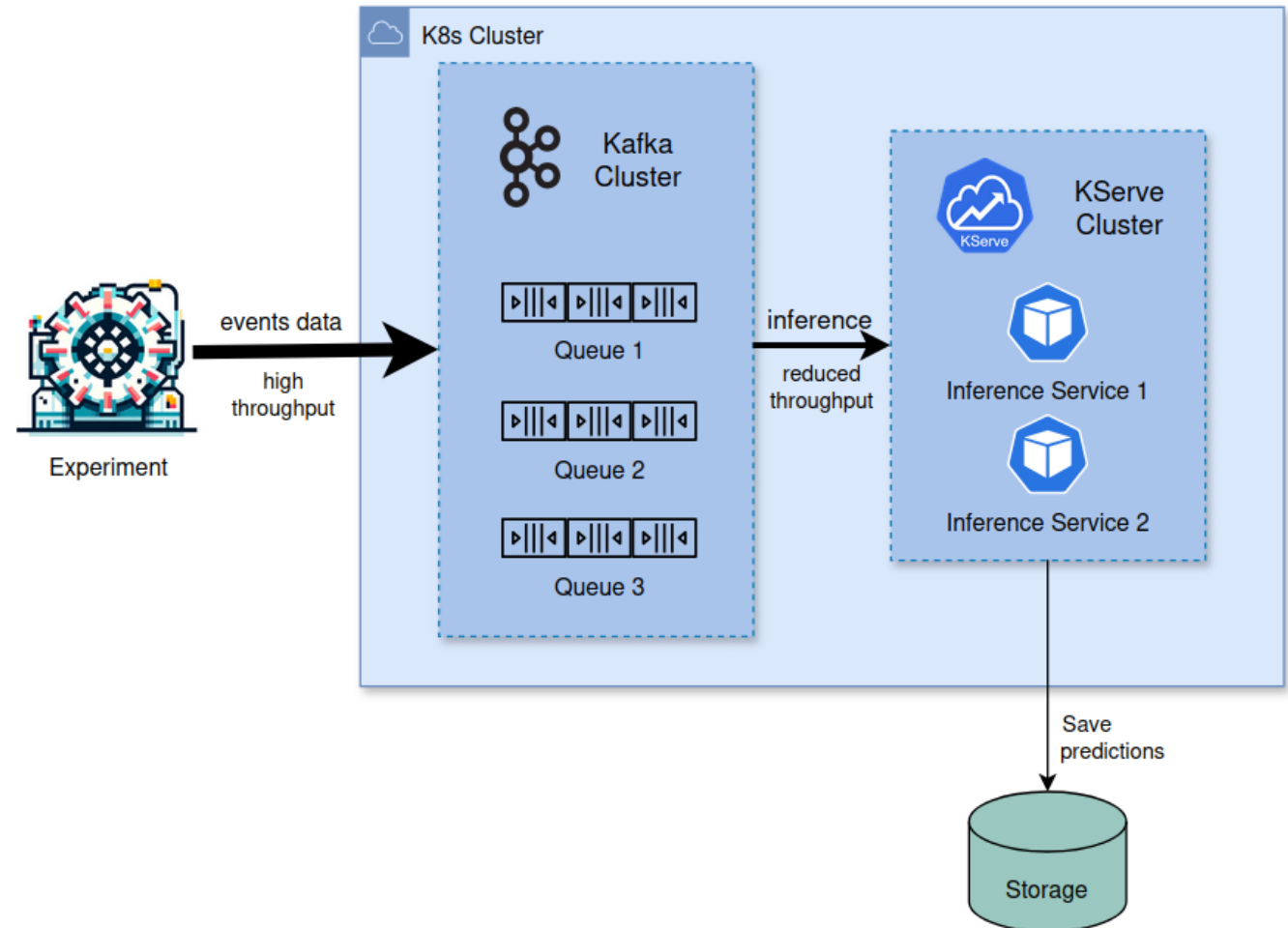
(Natively supported)

Apache Kafka

- **Open source** distributed **event streaming** platform designed to handle large volumes of real-time data feeds

Use Case:

- a running experiment produces a stream of events;
- we setup a data ingestion pipeline through which events are delivered to Kafka topics (queues);
- KServe natively integrates with Kafka: fetch data from queues and invoke Inference Services to get predictions;
- Inference is slow, but data is not lost, as it's stored in Kafka topics waiting to be processed.



Kubeflow Training Operator (KTO)

- Kubernetes-based **open source** platform for **distributed training/fine-tuning** of Machine Learning models **at scale** (first major release 2021).
- KTO supports **multiple ML frameworks**, including TensorFlow, PyTorch, XGBoost, MPI, and others.
 - KTO offers a uniform API across ML frameworks to submit your training jobs.
- Provide numerous features, e.g.:
 - Automated deployment of training jobs across nodes.
 - Handle monitoring and fault-tolerance.
 - Scalability: easily scale model training from single machine to large-scale distributed Kubernetes cluster.



Use Case

Text Classification

Framework: **PyTorch**

Task: infer the INFN structure name given an author's affiliation string

Training dataset:

- ~6k **positive** samples
 - "INFN Frascati Natl Labs, I-00044 Frascati, Roma" -> **LNF**
 - "INFN Bari, Dept Phys, Bari, Italy" -> **BA**
- ~6k **negative** samples
 - "Univ Siena, Dipartimento Fis, Pisa, Italy" -> **[Unknown]**
- dataset augmented to **~400k samples** by adding "smart" typos

Training evaluation:

- **97% accuracy** on test set

Cluster configurations – thanks to **AI_INF**N for resources:

- **1 x NVidia T4** - single node training: ~2 hours training
- **1 x NVidia T4 x 2 nodes** - two nodes training: ~5 hours training
- **2 x NVidia T4** - single node training: ~1.5 hours training (25% saved time)

What's next

INFN Cloud integration



- **INFN Cloud** is the INFN cloud computing infrastructure.
- The infrastructure is based on a core **backbone** connecting the large data centers of CNAF and Bari, and on a set of loosely coupled distributed and **federated sites** connected to the backbone.
- Backbone sites are high speed connected and host the INFN Cloud core services.
- Federated clouds: Cloud@CNAF, CloudVeneto, Cloud@ReCaSBari, Cloud-CT, Cloud-IBISCO-Na. Coming soon: LNGS, Milano, HTC in Tier-2s, HPC bubbles.

- The **INFN Cloud Dashboard** allows users to:
 - access centralized services;
 - Instantiate PaaS services, e.g., Virtual Machines, Docker Compose, etc.

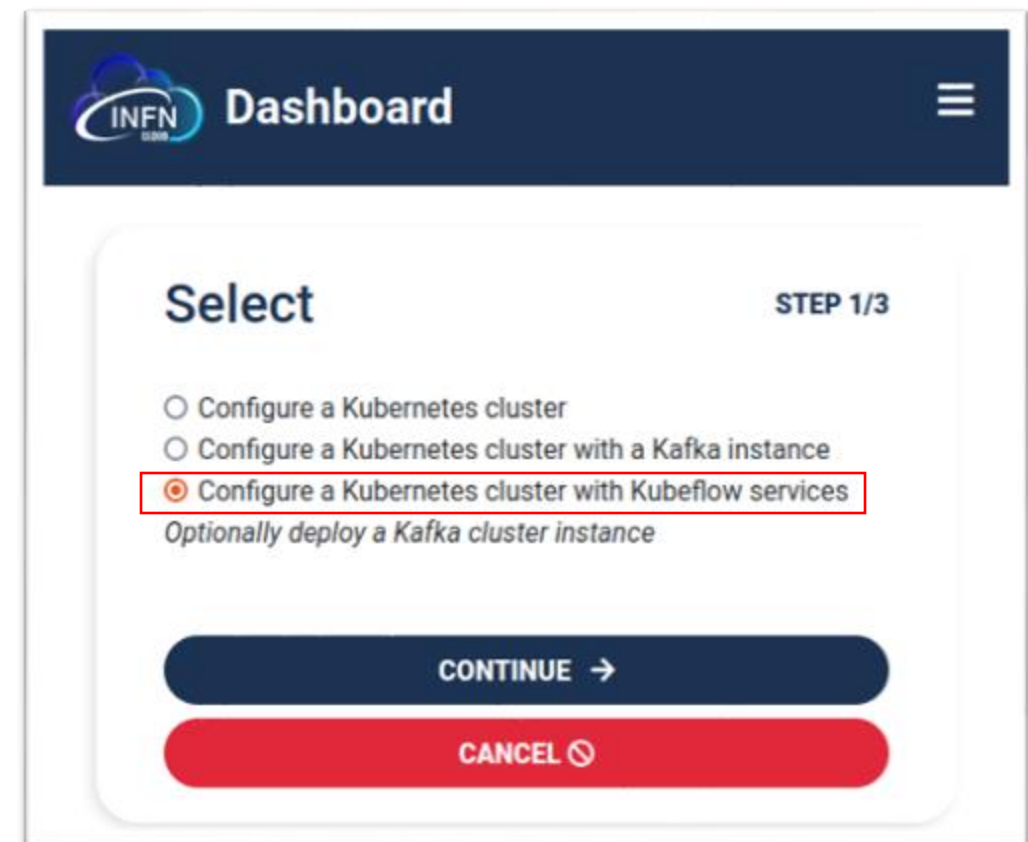
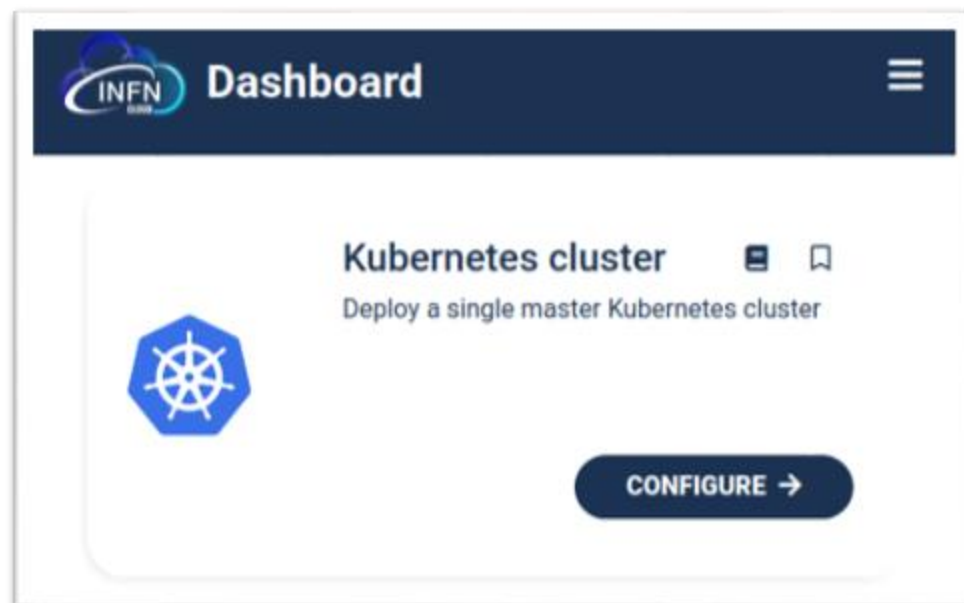


What's next

INFN Cloud integration

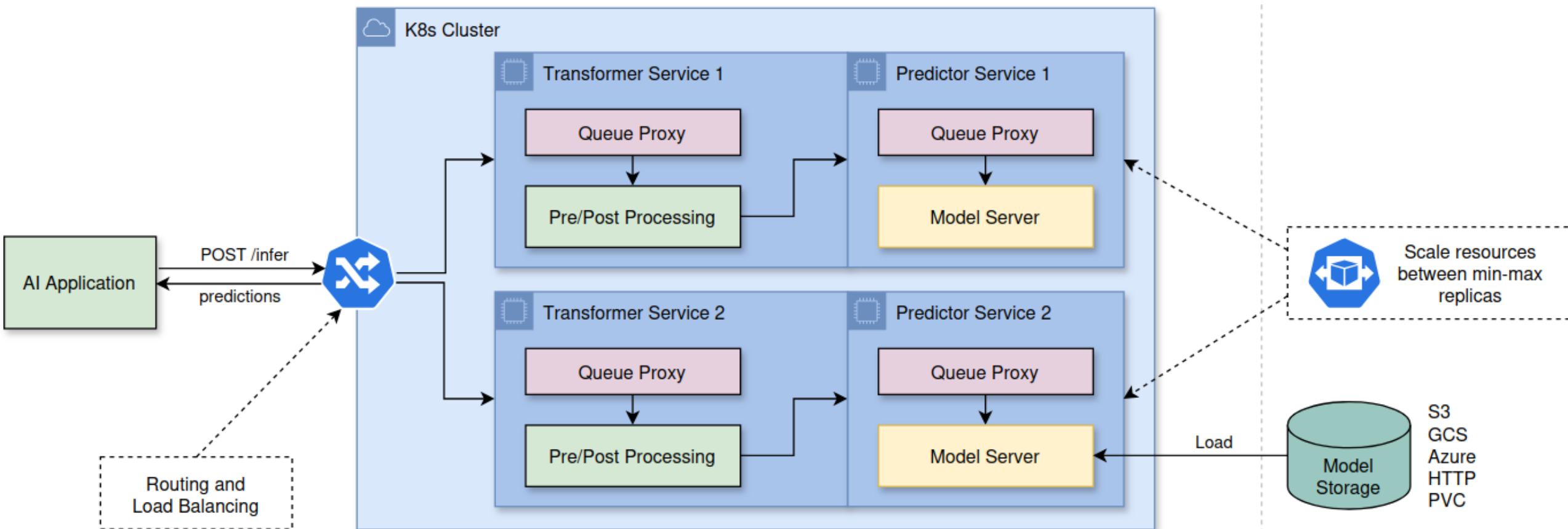


- Kubernetes belongs to INFN Cloud PaaS portfolio
- What to do:
 - package KServe + KTO in a suitable "installer"
 - add option to configure and install these services on top of a Kubernetes cluster



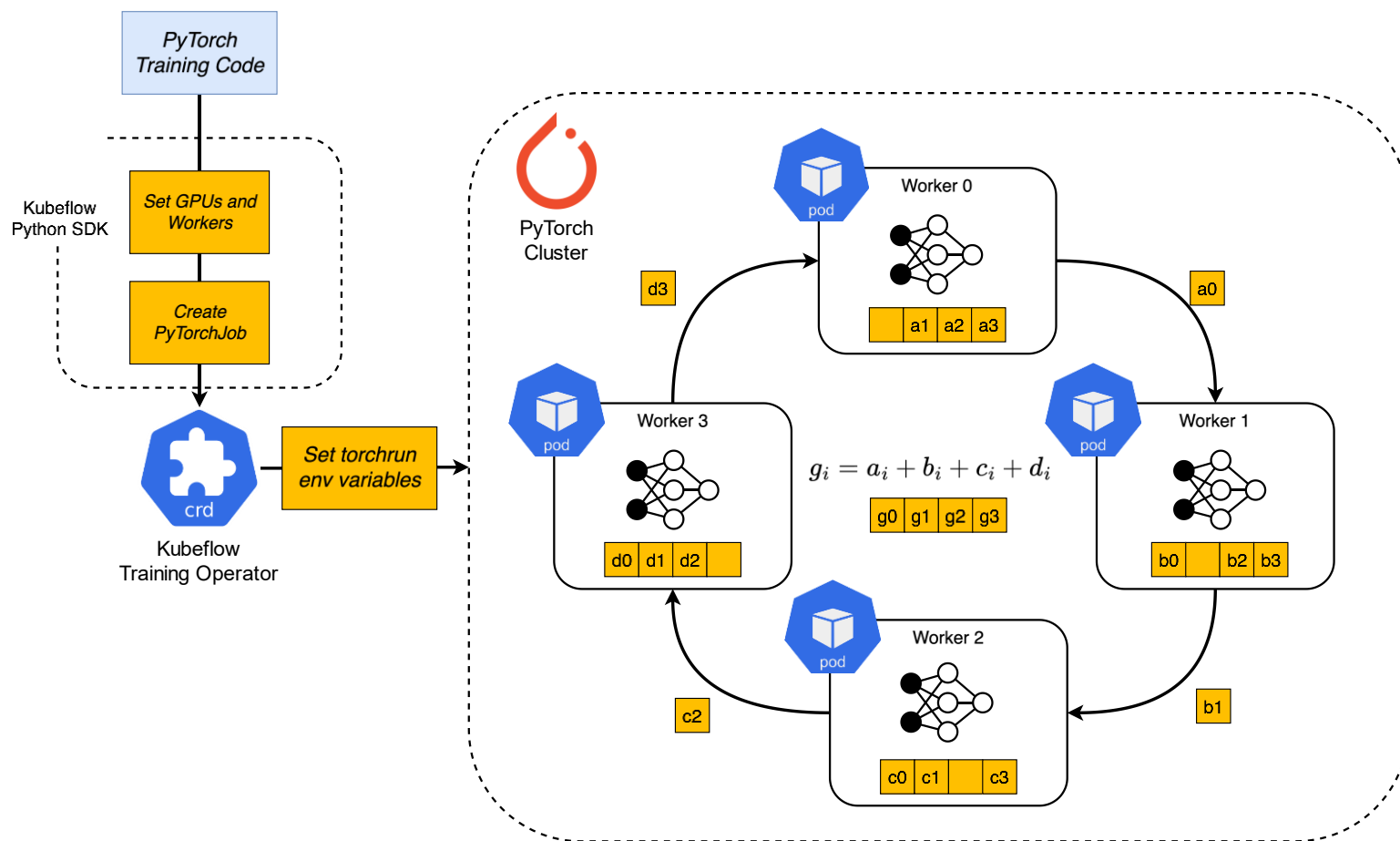
Thank You

KServe Architecture



PyTorch Distributed Training

- You write the training code and submit the training job to KTO
- KTO creates PyTorch workers and enables communication among them for the **ring all-reduce** algorithm.



KServe

Kubeflow Dashboard



Home

Notebooks

TensorBoards

Volumes

Katib Experiments

KServe Endpoints

Pipelines

Pipelines

Experiments

Runs

Recurring Runs

Artifacts

Executions

Manage Contributors

GitHub

Documentation

build version - dev_local

mlaas2 (owner)

Endpoint details

sklearn-iris

OVERVIEWDETAILSLOGSEVENTSYAML

URL external

http://sklearn-iris.mlaas2.infn.it

URL internal

http://sklearn-iris.mlaas2.svc.cluster.local

Component

predictor

Storage URI

gs://kfserving-examples/models/sklearn/1.0/model

Predictor

sklearn

Runtime

SKLearn ModelServer

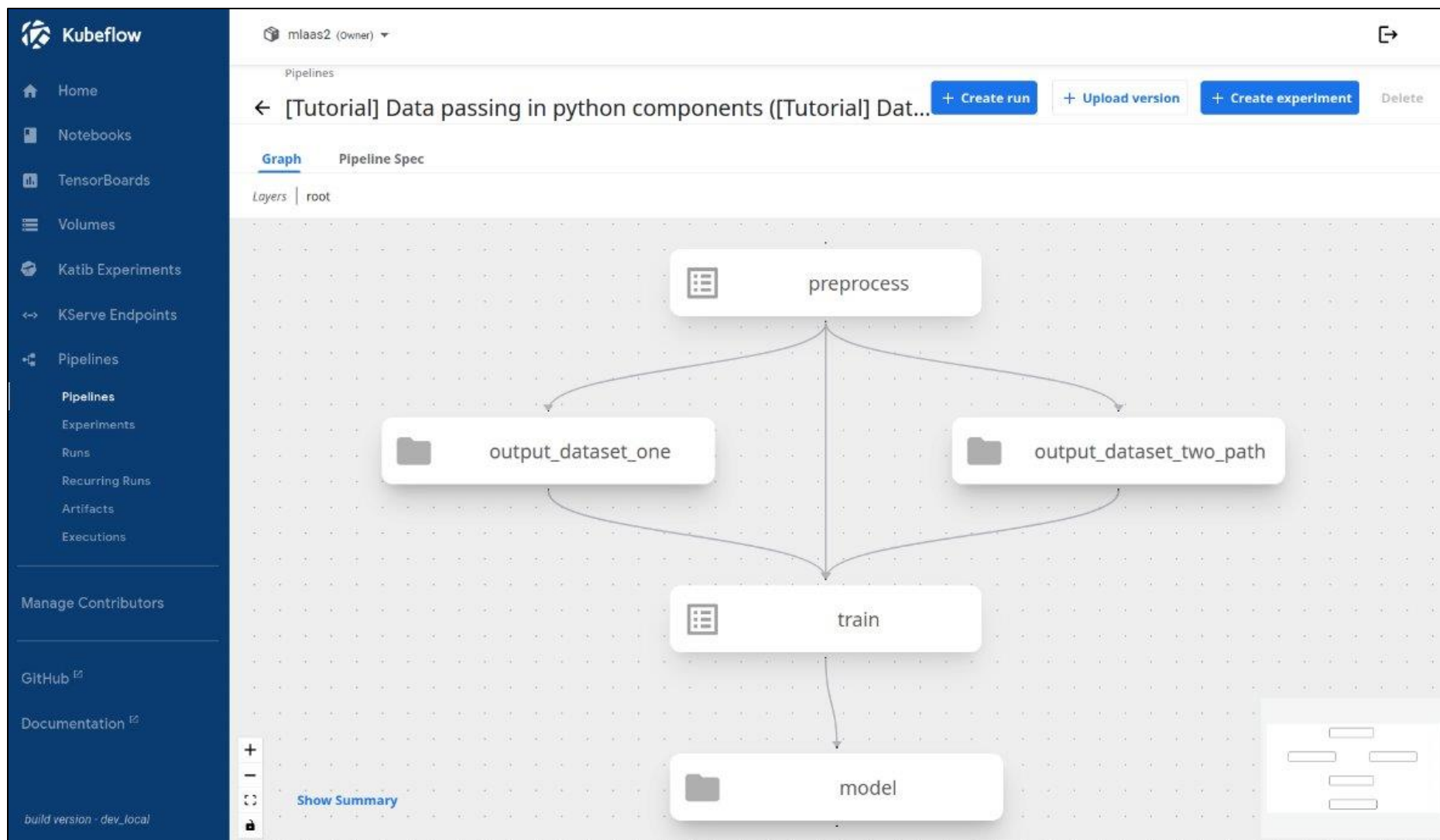
InferenceService Conditions

Filter Enter property name or value

Status	Type	Last Transition Time ↓	Reason	Message
✓	IngressReady	3 days ago		
✓	LatestDeploymentReady	3 days ago		
✓	PredictorConfigurationReady	3 days ago		
✓	PredictorReady	3 days ago		
✓	PredictorRouteReady	3 days ago		

ML Pipelines

Kubeflow Dashboard



Home

[Home](#)

Models

[Catalog](#)[Chat](#)[Inference](#)[Train](#)

Settings

[Settings](#)

CATALOG

 Iris Classifier**Objective:** classification**Format:** sklearn**Host:** http://localhost:4200/iris (KServe/v1)

Scikit-learn model trained with the Iris dataset. This dataset has three output class: Iris Setosa, Iris Versicolour, and Iris Virginica.

 the Higgs boson ML challenge**Objective:** classification**Format:** sklearn**Host:** http://localhost:4200/hep-2 (KServe/v1)

This challenge focuses on one particular decay topology of the Higgs boson among the many possible ones: events $H \rightarrow \tau\tau$ where one tau decays into an electron or a muon and two

 $t\bar{t}H(bb)$ analysis**Objective:** classification**Format:** tensorflow**Host:** http://localhost:4200/hep (KServe/v1)

This model discriminates $t\bar{t}H(5b)$ events with all-jets final state, where at least one of the jets of the final state is a boosted jet, and where the Higgs boson decays in a pair of well resolved jets

 TinyLlama/TinyLlama-1.1B-Chat-v1.0**Objective:** text-generation**Format:** ModelServer**Host:** http://131.154.98.72:30080 (KServe/v2)

TinyLlama is pretrained 1.1B Llama model on 3 trillion tokens. TinyLlama adopts the same architecture and tokenizer as Llama 2. Besides, TinyLlama is compact with only 1.1B parameters.

 google/gemma-2b-it (localhost)**Objective:** text-generation**Format:** ModelServer**Host:** http://localhost:8080 (KServe/v2)

Gemma is a family of lightweight, state-of-the-art open models from Google, built from the same research and technology used to create the Gemini models. They are text-to-text, decoder-only large

 mistralai/Mixtral-8x7B-Instruct-v0.1**Objective:** text-generation**Format:** ModelServer**Host:** http://131.154.98.96:30080 (KServe/v2)

The Mixtral-8x7B Large Language Model (LLM) is a pretrained generative Sparse Mixture of Experts.

 unsloth/llama-3-70b-Instruct-bnb-4bit**Objective:** text-generation**Format:** ModelServer**Host:** http://131.154.98.96:30080 (KServe/v2)

Meta developed and released the Meta Llama 3 family of large language models (LLMs), a collection of pretrained and instruction tuned generative text models in 8 and 70B sizes. The

 unsloth/llama-3-8b-Instruct-bnb-4bit**Objective:** text-generation**Format:** ModelServer**Host:** http://131.154.98.72:30080 (KServe/v2)

Meta developed and released the Meta Llama 3 family of large language models (LLMs), a collection of pretrained and instruction tuned generative text models in 8 and 70B sizes. The

[+ Add](#)

Home

Home

Models

Catalog

Chat

Inference

Train

Settings

Settings

INFERENCE

Model Server: *ttH(bb)* analysis ✓ ✕

Settings

Model selector

Model Server ●

ttH(bb) analysis (KServe/v1) ▼

INPUT Send ↗

Events data

```
1 {
2   "instances": [
3     [0.19563319290790765, 0.8628343629750731, 0.20469675544301077, 0.5979233885840486, 0.5624403641089002, 0.4966360687831127, 0.
9971232134875923, 0.9641571184466814, 0.016140890033353065, 0.012650983007060364, 0.044779417127065256, 0.04623121102305415, 0.
027365998536597175, 0.004034759345313149, 0.07267125173331217, 0.4668850294559054, 0.10894376909915392, 0.044679238817932156, 1.0, 0.
9496461903795053, 0.9982200258458502, 0.5, 0.0, 0.0, 0.35235498377667923, 0.6612158851740676, 0.6065199265679636, 0.3931907707503391, 0.
37482121050755157],
4     [0.08704080381772918, 0.6462195461371039, 0.6251125365793502, 0.10531701202713299, 0.5783607282924024, 0.5257032073478767, 0.
7715301792601783, 0.8643515820911014, 0.0572575048444786, 0.024066481779652534, 0.05685699616254509, 0.058440006176756924, 0.
08445073800102749, 0.00918989915016802, 0.010650522091712461, 0.39854084316539823, 0.49500531223932435, 0.006548080688324688, 1.0, 0.
9995915433503754, 0.9915019438570389, 1.0, 0.3333333333333333, 0.0, 0.3873084709413965, 0.5798505283787048, 0.4003676756011203, 0.
31962079085319367, 0.2719980271120939]
5   ]
6 }
```

OUTPUT

Predictions from model

```
▾ predictions:
0: 0.110319018
1: 0.403853774
```

Home

Home

Models

Catalog

Chat

Inference

Train

Settings

Settings

INFERENCE

Model Server: the Higgs boson ML challenge ✓ ✕

Settings

Model selector

Model Server ●

the Higgs boson ML challenge (KServe/v1) ▼

INPUT Send ↗

Events data

```
1 {
2   "instances": [
3     [0.05163852,0.69959853,-0.19880623,-0.87887007,-0.09081739,1.03192017,-0.76172901,-0.72999068,0.79123704,0.99892949,-0.66461696,1.
4     22560102,-1.20479845,0.15843245,0.65179506,0.51384626,-0.43892379,-1.20784558,-1.00179211,-0.24120595],
5     [-0.12257931,-0.32612476,-0.0762696,0.24577648,-0.72471509,-0.39736371,-0.75149261,0.15967885,-0.90869178,1.26367869,0.75337113,-0.
6     89864782,-0.64996439,-0.42788395,-0.87945491,0.45659669,-0.55930819,-0.27914992,1.04440205,-0.65468282]
```

OUTPUT

Predictions from model

```
predictions:
  0: 0
  1: 1
```