

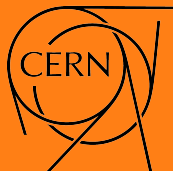


KubeCon



CloudNativeCon

Europe 2026



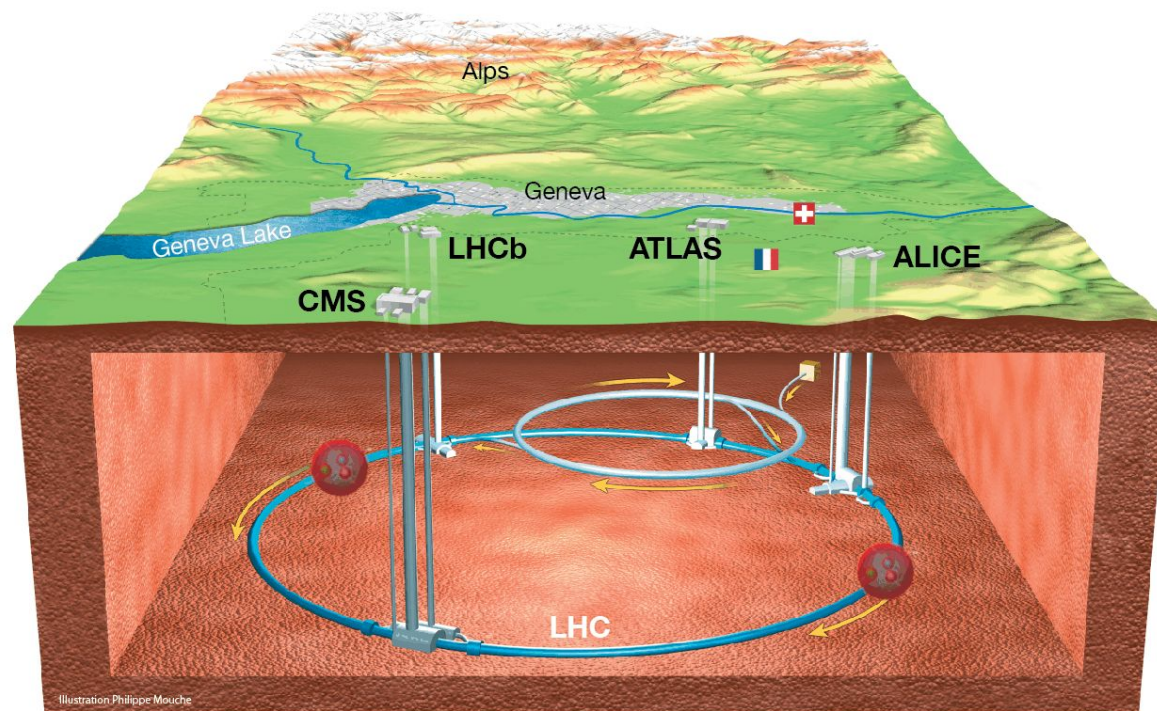
NextGen

#KubeCon #CloudNativeCon

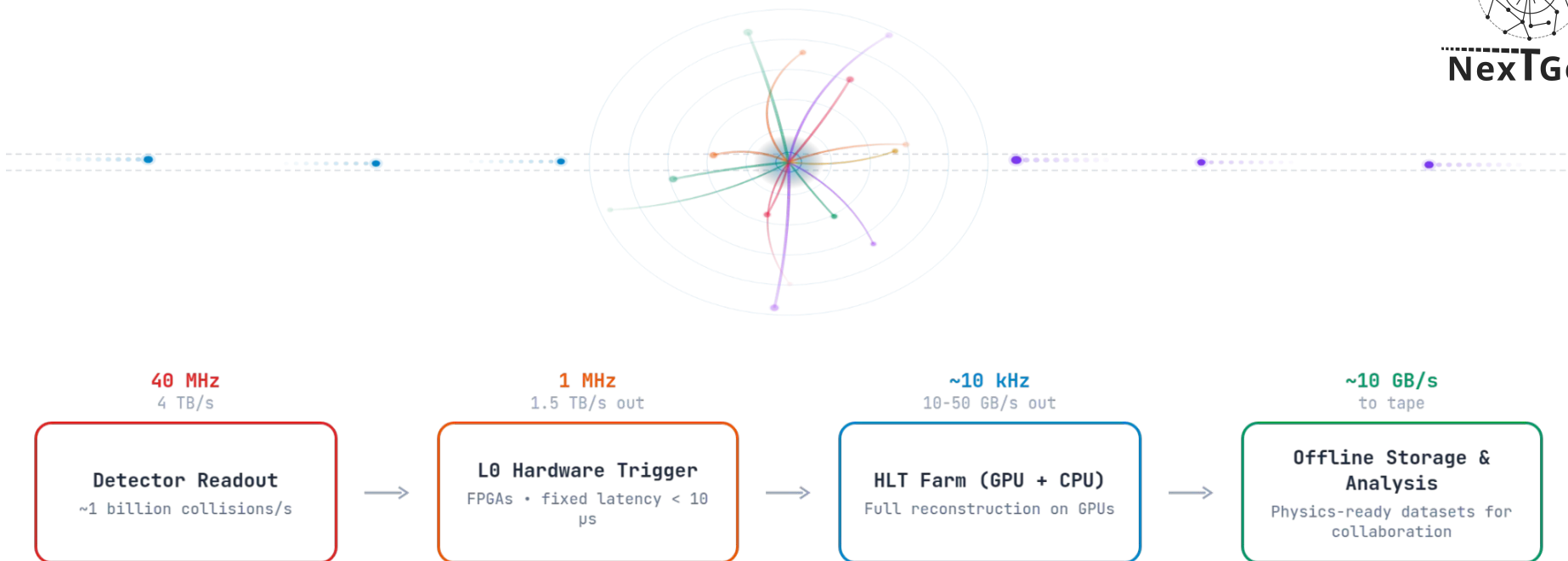
The Shell Awakens: Cloud Native Workflows for Particle Physicists

Raulian Chiorescu & Hannes Hansen, CERN





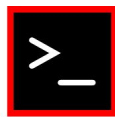
Next Generation Triggers



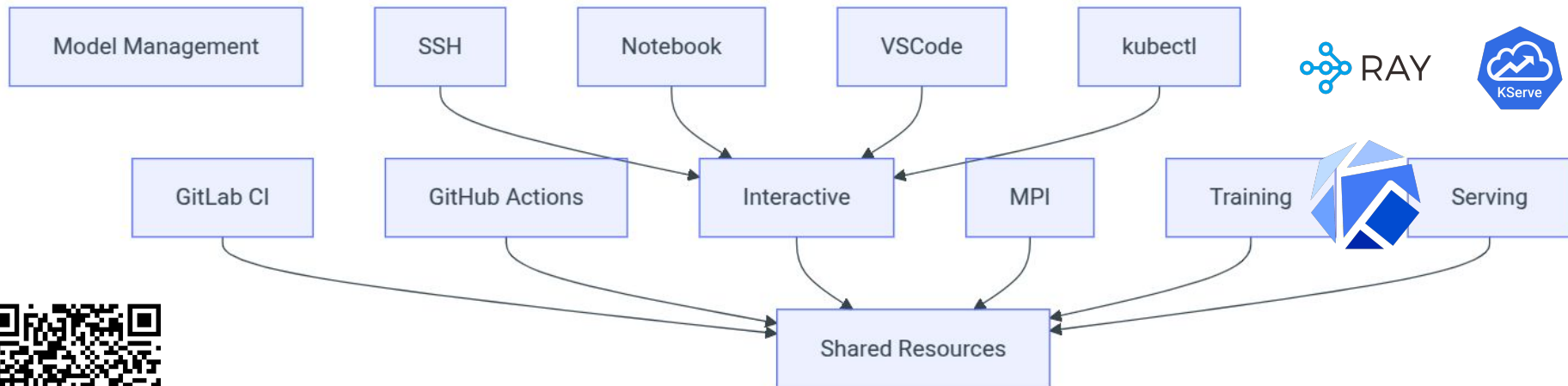


Architecture

mlflow™



ContainerSSH
Launch containers on demand



Find out more!



Compute Resources



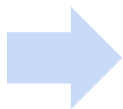
KubeCon



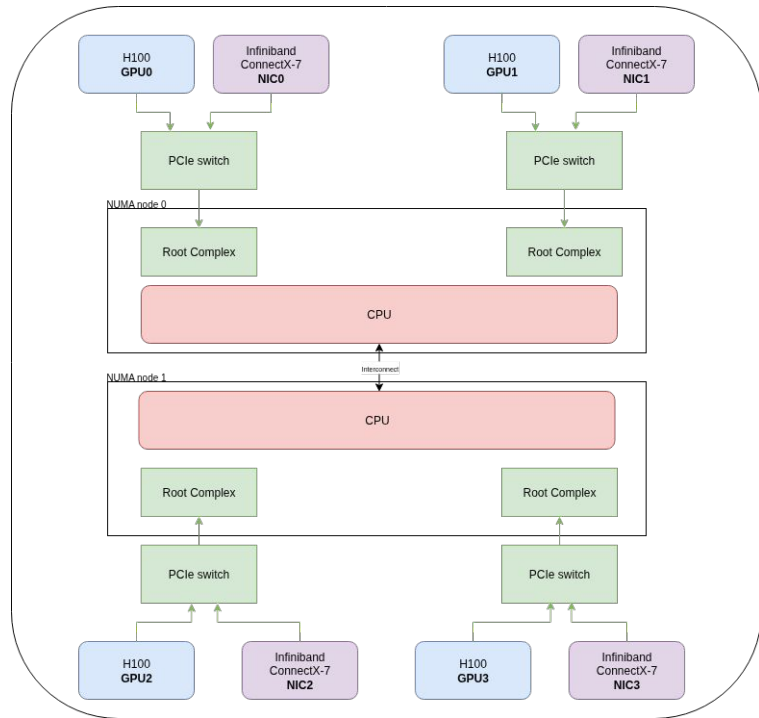
CloudNativeCon

Europe 2026

- **NVIDIA GPU** (H100, H200, L40S)
- **AMD GPU** (Mi300X, RadeonPro W7900)
- **CPU** (high memory & core & frequency)
- **Node Interconnects** (IB & RoCEv2)
- **FPGAs**



Heterogeneous Cluster
Different Topologies



H100 SXM with Infiniband Interconnect

Storage



KubeCon



CloudNativeCon

Europe 2026

	Local Node	Local Cluster	External
Deployment	<i>emptydir</i> Volume	Longhorn	CEPH
Use Case	Fast Access to temporary data, e.g. compilation	Shared access between workloads, e.g. multi node training or MPI	Long-term storage for output data, e.g. analysis output

Users don't always know when to use what!



LONGHORN

- We provide curated environments:
 - CUDA/ROCM libraries
 - Preinstalled tooling (ML libraries & physics software)
 - Images are prepulled on all nodes to **reduce waiting time** for users
- Advanced users build their own images on top of our base images
- Access to local scratch space as well as shared storage systems
 - \$HOME directory is set to CERN storage systems
- CPU & Memory & Local scratch space are preset (scaled with the number of requested GPUs)

Interactive Access



KubeCon



CloudNativeCon

Europe 2026

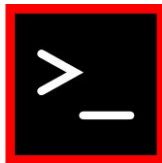


```
kubectl exec -it session -- bash
```

- SSH access via **ContainerSSH**
- Pod name is used as the identifier
- OIDC support - no need to copy public keys
- *sudo* capabilities
- Culling of idle sessions



```
ssh session@cern.ch  
$ id  
$ uid=123002(johndoe)  
$ sudo dnf install -y clang
```



ContainerSSH
Launch containers on demand

Interactive Access



KubeCon



CloudNativeCon

Europe 2026

```
apiVersion: v1
kind: Pod
metadata:
  name: session-kubecan
  labels:
    home: "afs"
    mount-eos: "true"
    mount-cvmfs: "true"
    mount-afs: "true"
spec:
  containers:
  - name: container
    image: registry.cern.ch/ngt/lxplus-like:9
    command: ["/usr/bin/tini", "--", "sleep", "infinity"]
    resources:
      limits:
        nvidia.com/gpu: 1
  nodeSelector:
    nvidia.com/gpu.product: NVIDIA-H100-NVL
```

kubectl create -f session.yaml

Name



JupyterLab

An interactive development environment for notebooks, code, and data. Ideal for prototyping and experimentation.

1

VisualStudio Code

A lightweight but powerful source code editor, redefined and optimized for building and debugging modern web and cloud applications.

2

RStudio

An integrated development environment for R, a programming language for statistical computing and graphics.

Custom Notebook



CPU / RAM ?

Resources will be set automatically based on the number and type of GPUs.

Minimum CPU

0.5

Minimum Memory Gi

1

Advanced Options

GPUs

Number of GPUs

1

GPU

NVIDIA H100 NVL (94GB)

Performance Optimizations



KubeCon

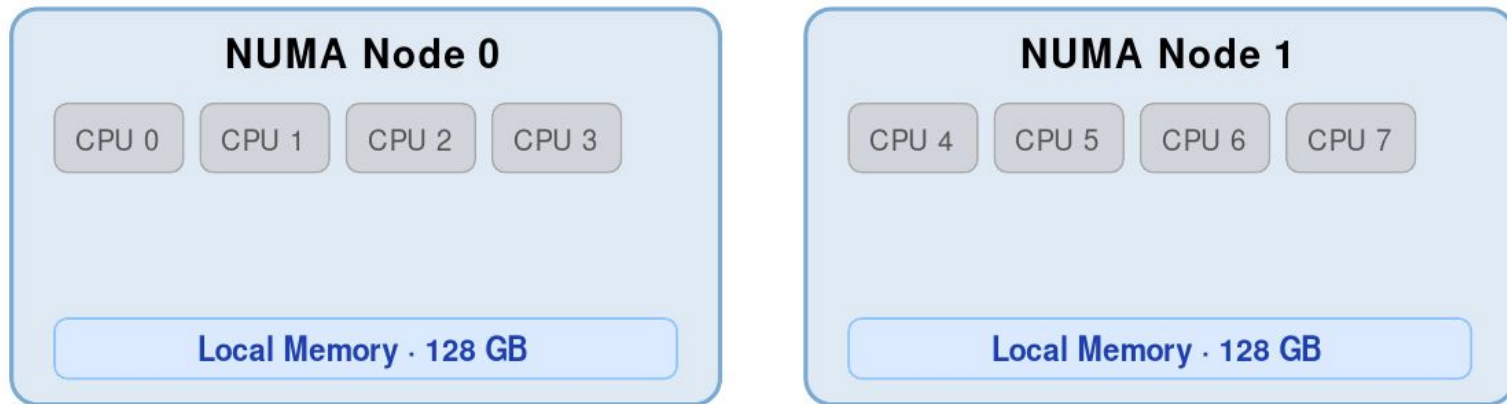


CloudNativeCon

Europe 2026

By default, container processes spread across CPU cores and NUMA nodes
This causes **higher memory latency**

Without Topology Awareness

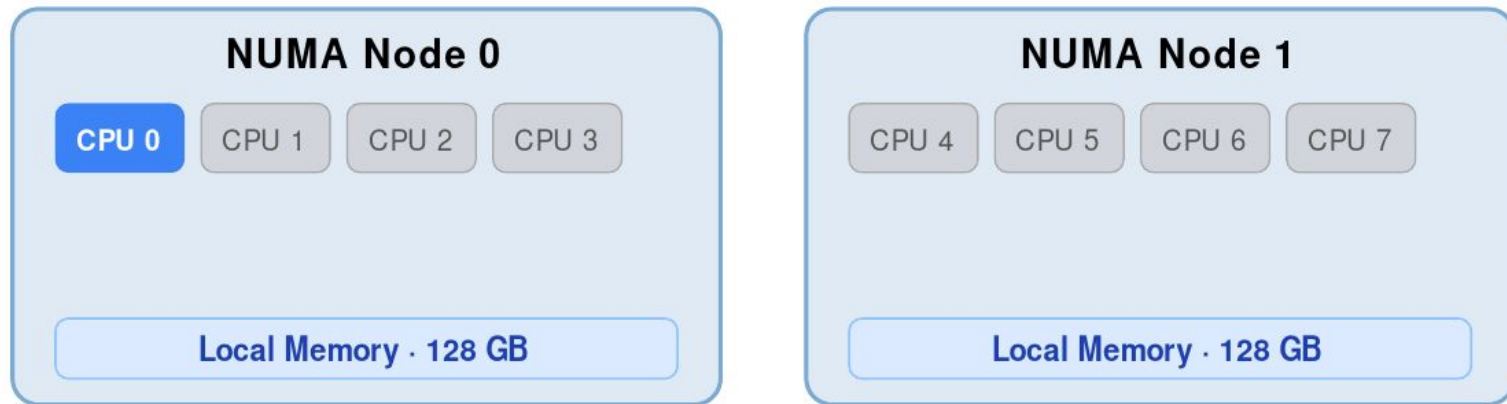


Pod requests 2 CPUs...

Performance Optimizations

By default, container processes spread across CPU cores and NUMA nodes
This causes **higher memory latency**

Without Topology Awareness



Performance Optimizations



KubeCon

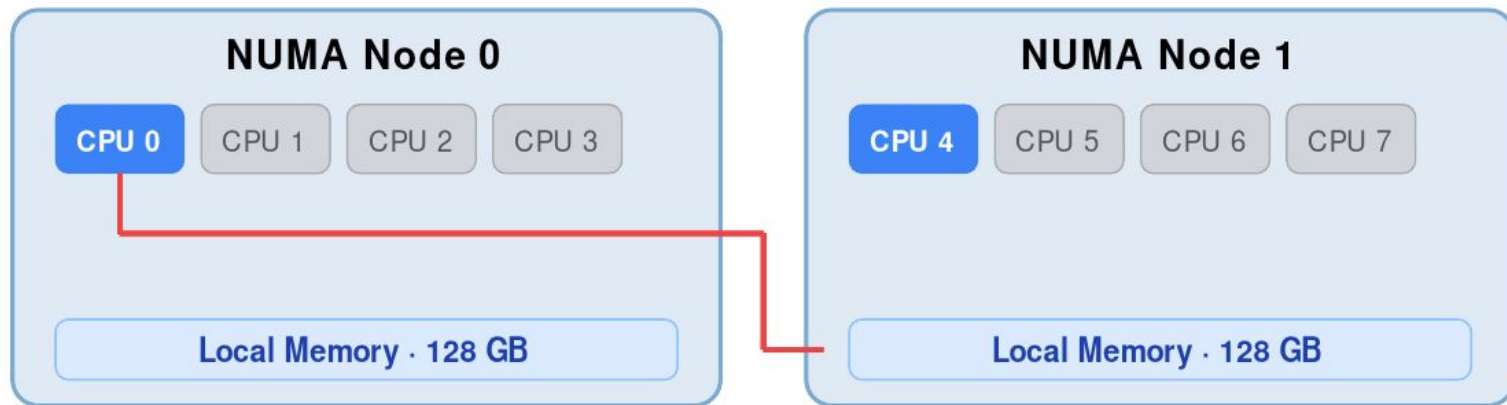


CloudNativeCon

Europe 2026

By default, container processes spread across CPU cores and NUMA nodes
This causes **higher memory latency**

Without Topology Awareness



cross-NUMA memory access

Performance Optimizations



KubeCon

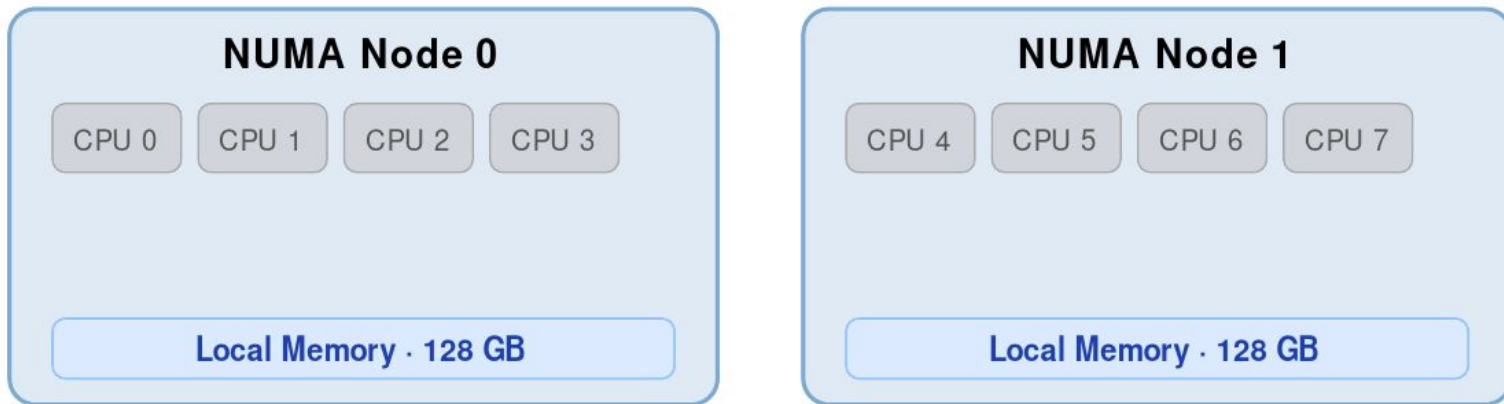


CloudNativeCon

Europe 2026

By default, container processes spread across CPU cores and NUMA nodes
This causes **higher memory latency**

With Topology Manager

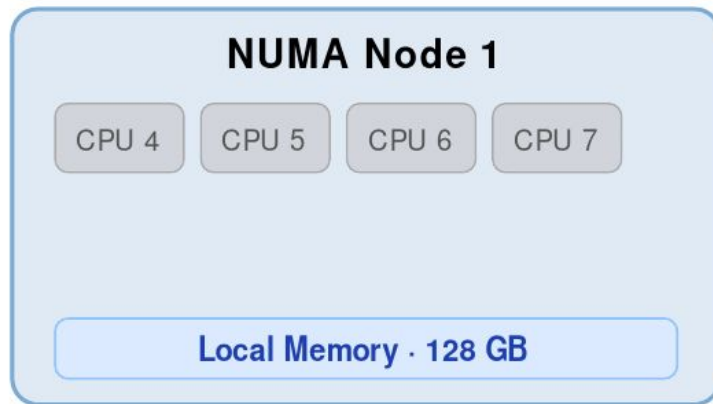
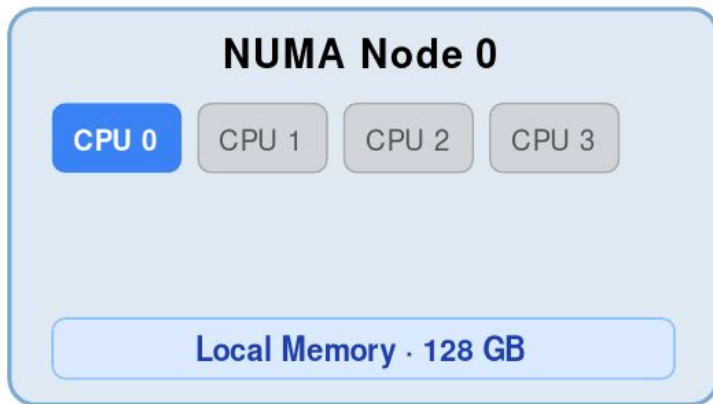


Pod requests 2 CPUs...

Performance Optimizations

By default, container processes spread across CPU cores and NUMA nodes
This causes **higher memory latency**

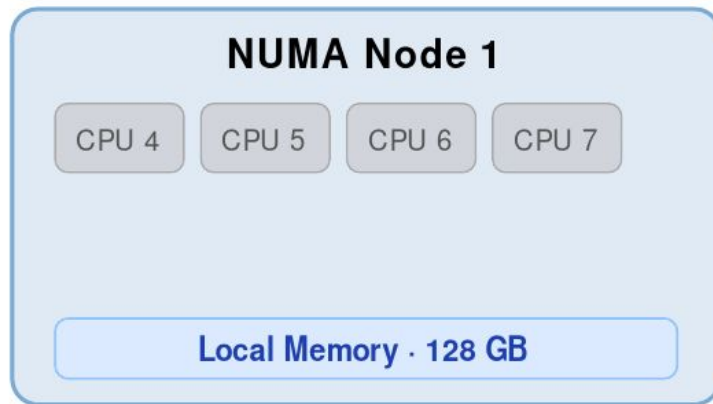
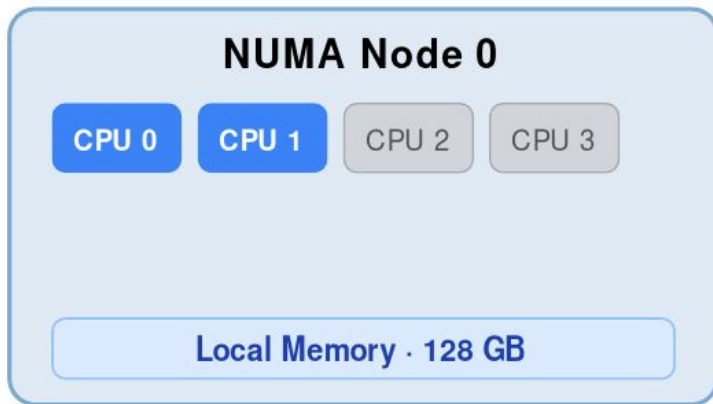
With Topology Manager



Performance Optimizations

By default, container processes spread across CPU cores and NUMA nodes
This causes **higher memory latency**

With Topology Manager



Performance Optimizations



KubeCon

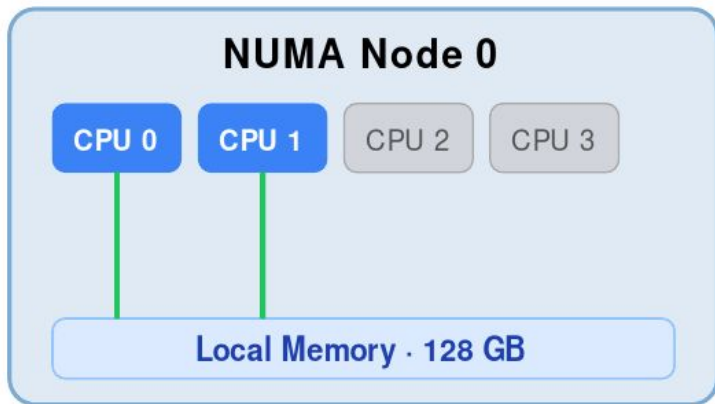


CloudNativeCon

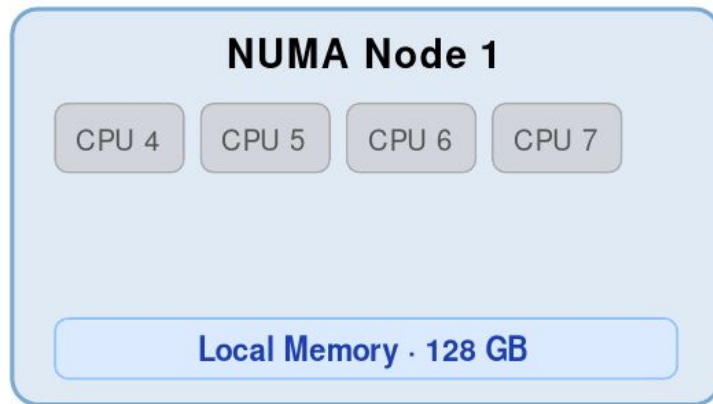
Europe 2026

By default, container processes spread across CPU cores and NUMA nodes
This causes **higher memory latency**

With Topology Manager



local memory access



Machine Learning

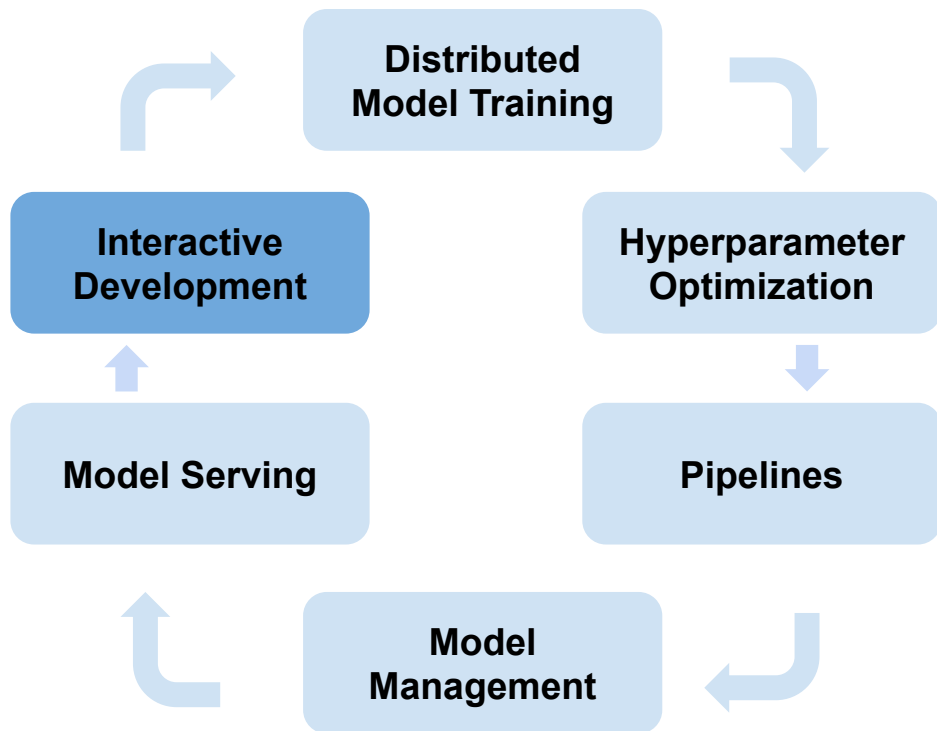


KubeCon



CloudNativeCon

Europe 2026



- Jupyter Notebooks via KubeFlow
- Curated ML environments
 - PyTorch
 - Tensorflow
 - SciPy
- SSH-based access
- Attachment of local VS Code IDE to a running session



Machine Learning



KubeCon



CloudNativeCon

Europe 2026

```
Untitled1.ipynb +
[ ]: import torch
import torch.nn as nn
import torch.optim as optim

# 1. Define the Model
class SimpleModel(nn.Module):
    def __init__(self, input_size, output_size):
        super(SimpleModel, self).__init__()
        self.fc = nn.Linear(input_size, output_size)

    def forward(self, x):
        return self.fc(x)
```

```
ssh torch-session@ngt
$ conda create -n ml-env
$ conda activate ml-env
$ tmux
$ python train.py
```

Machine Learning

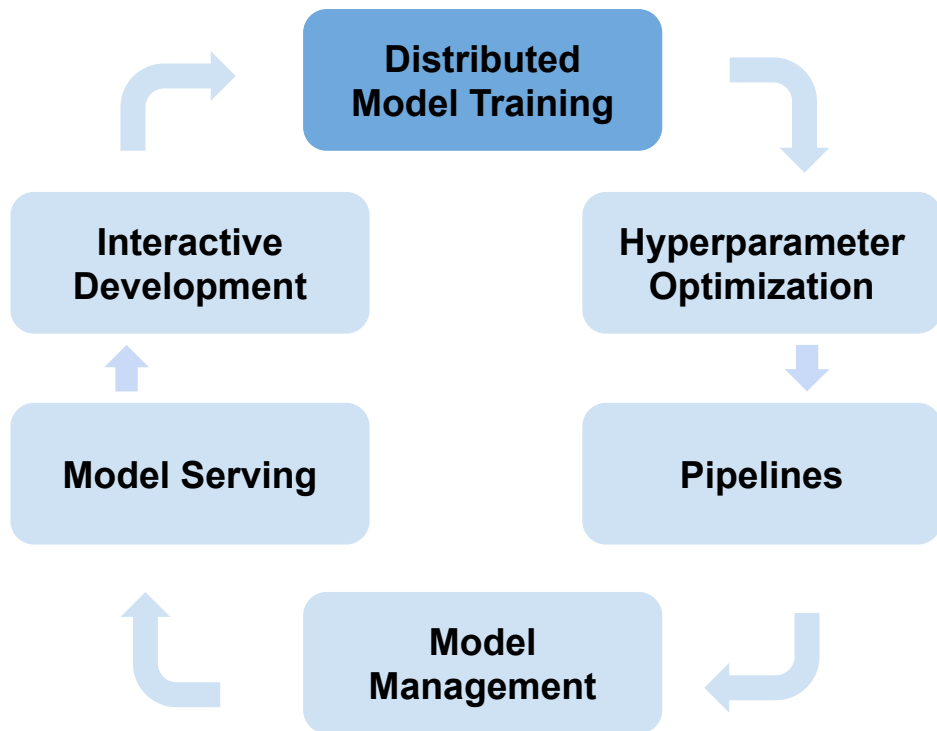


KubeCon



CloudNativeCon

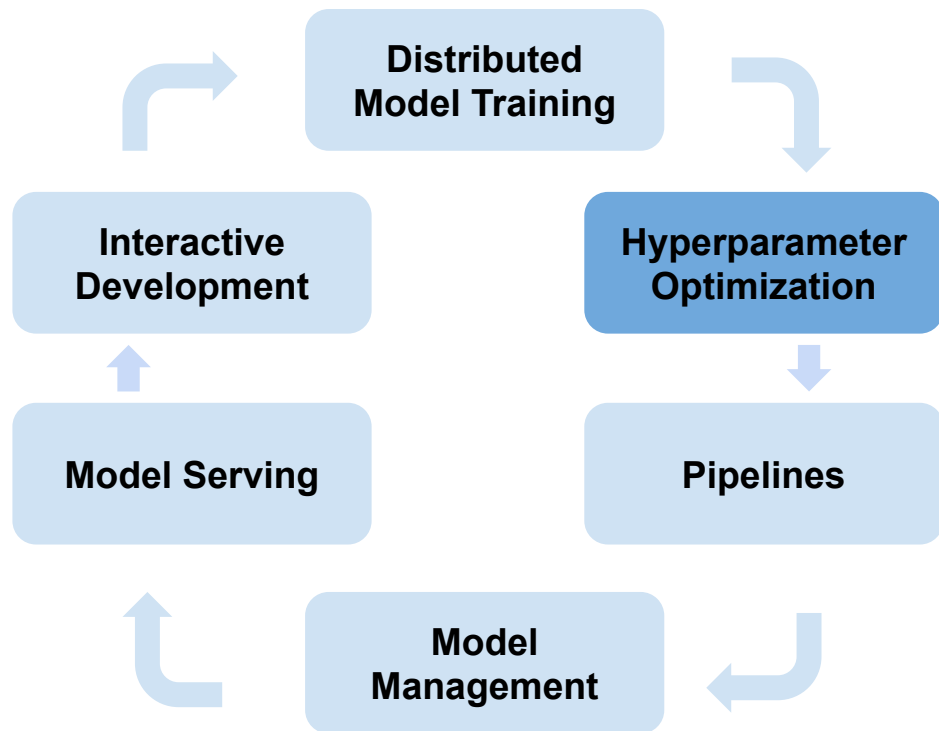
Europe 2026



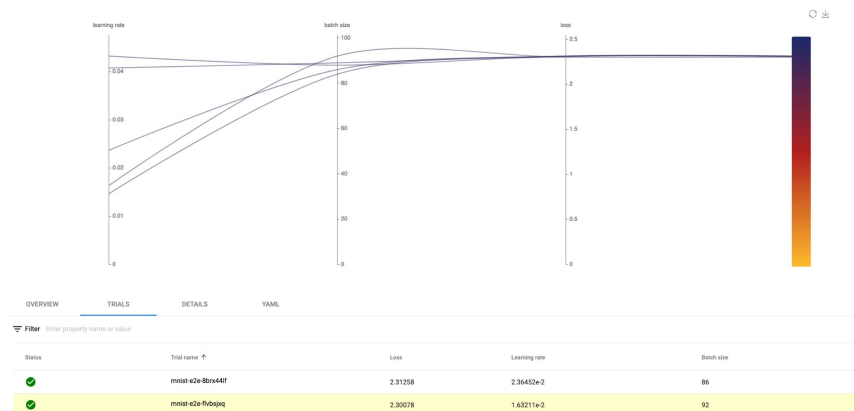
- Job Submission
 - KubeFlow (*TrainingJob*)
 - Ray (*RayJob*)
- Interactive with KubeFlow Trainer & Ray Train SDKs



Machine Learning



- KubeFlow Katib
- Ray Tune



Machine Learning

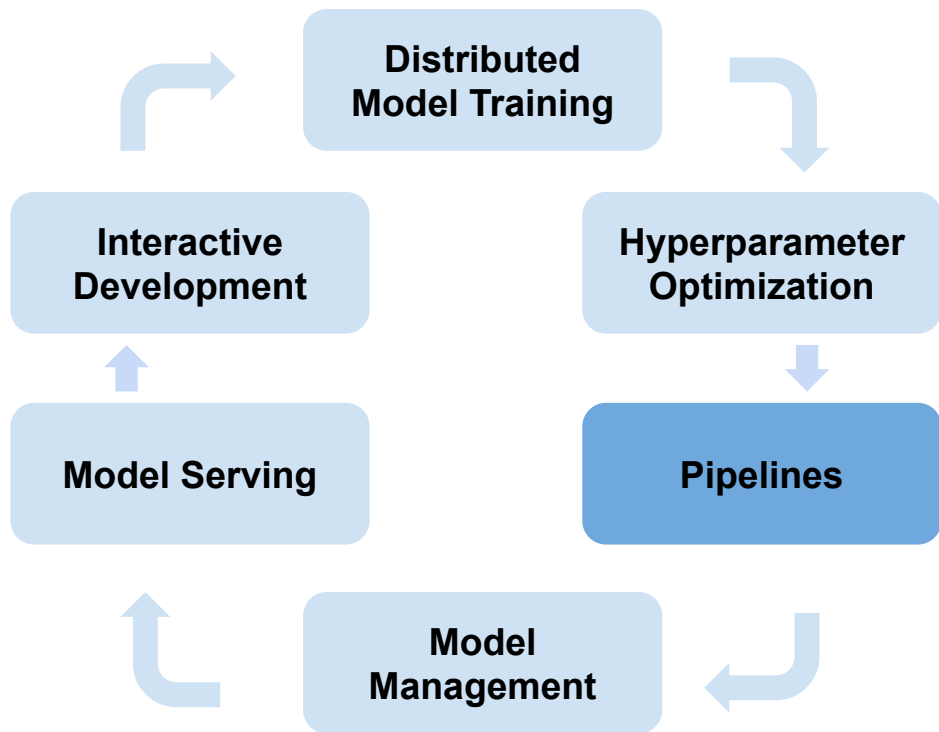


KubeCon

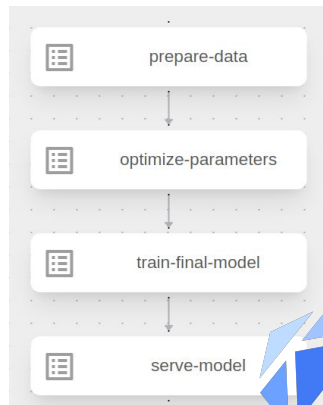


CloudNativeCon

Europe 2026



- KubeFlow pipelines for orchestration of more complex workflows
- Prebuild pipelines for model quantization and FPGA synthesis



Machine Learning

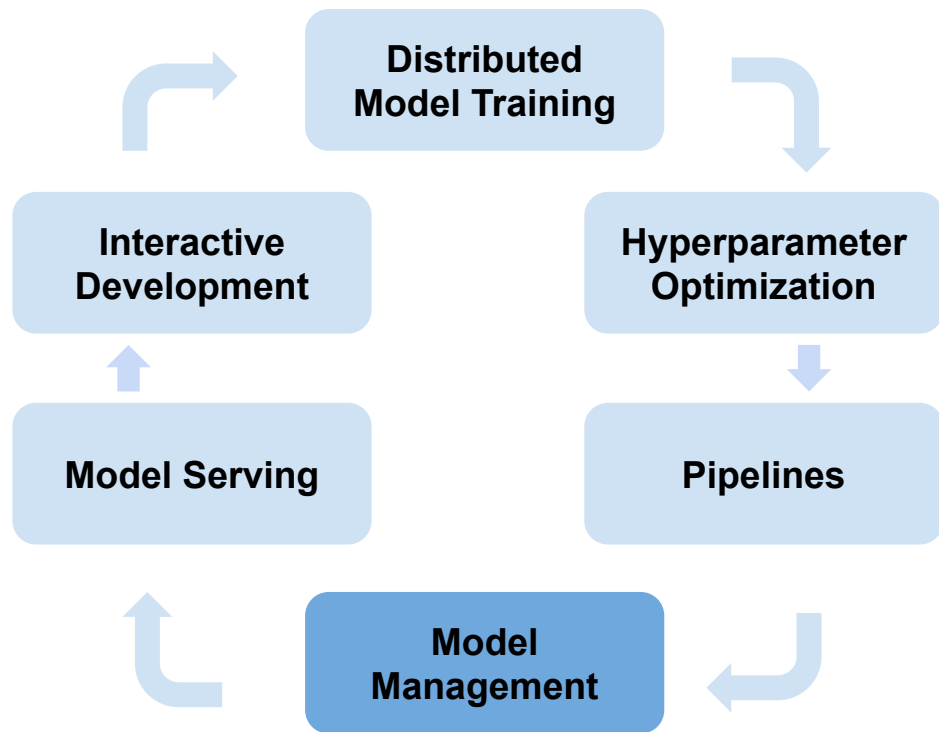


KubeCon

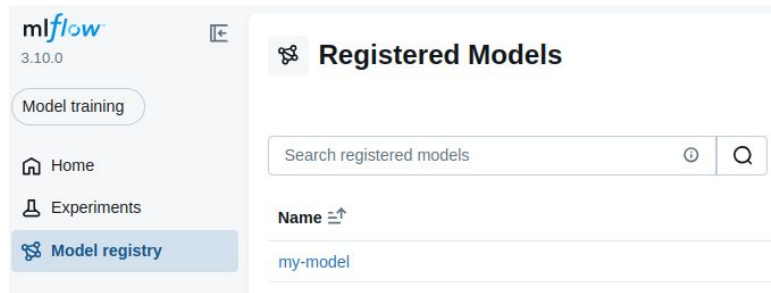


CloudNativeCon

Europe 2026



- MIFlow for model management and experiment tracking
- Integrated into our pipelines



Machine Learning

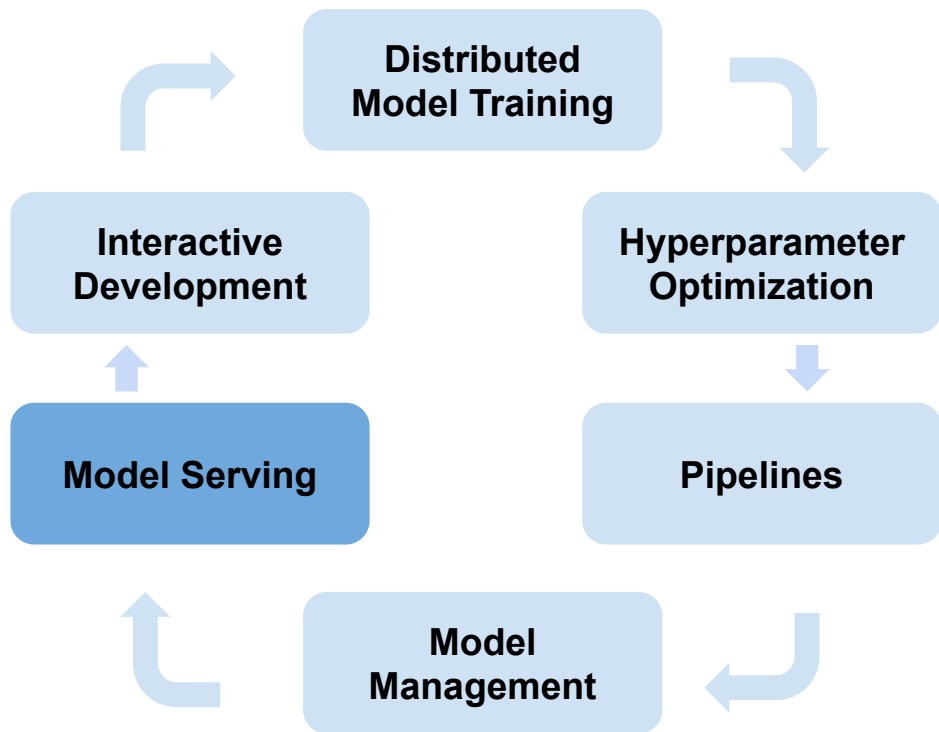


KubeCon



CloudNativeCon

Europe 2026

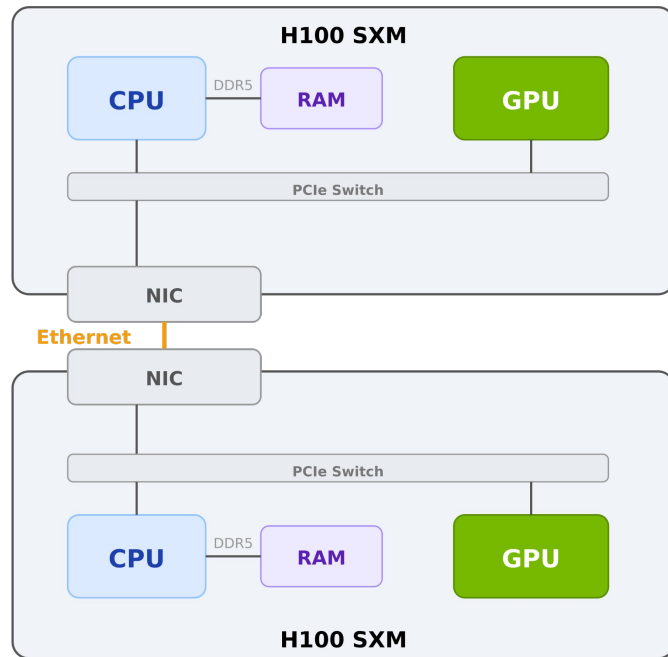


- KubeFlow & KServe
- Multi-Model Serving
- Zero Scaling
- Custom Serving Runtimes



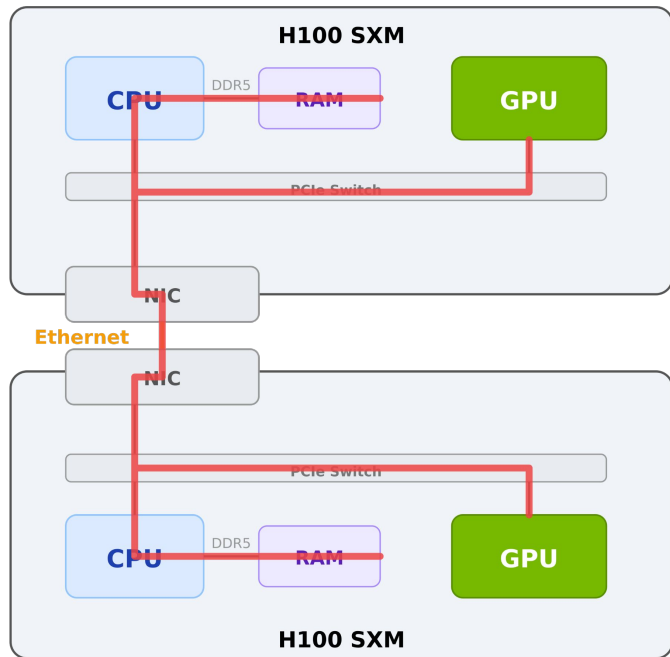
- Traditionally, GPU inter node communication involves staging through host memory and CPU involvement
- Using Ethernet interconnect
- Not suitable for latency sensitive workloads

Traditional: GPU-to-GPU



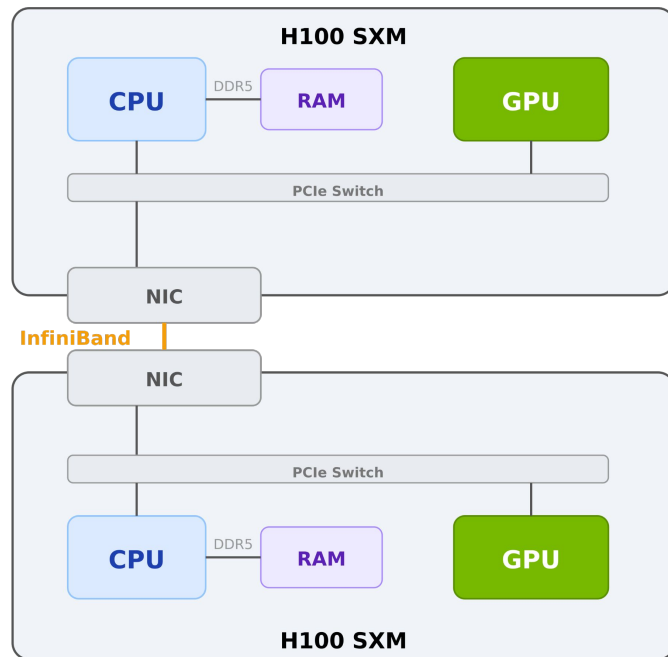
- Traditionally, GPU inter node communication involves staging through host memory and CPU involvement
- Using Ethernet interconnect
- Not suitable for latency sensitive workloads

Traditional: GPU-to-GPU



- Low latency networking through **Infiniband** and **RoCEv2**
- NVIDIA Network Operator and the AMD/NVIDIA GPU Operators
- Exposure of NIC devices to the Pods

GPUDirect RDMA



- GPUDirect Remote Direct Memory Access (RDMA) allows **direct communication** between the GPUs via the network interfaces

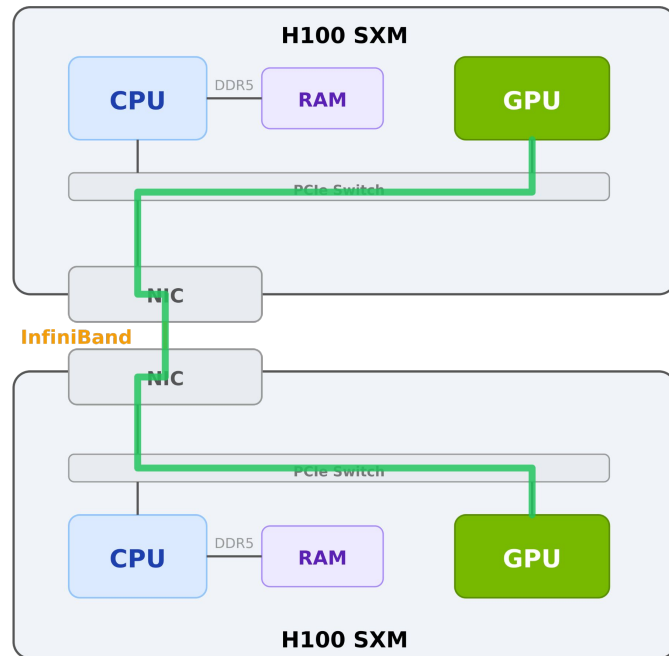
MPI
(OpenMPI & MPICH)

UCX

Infiniband/RoCE Drivers

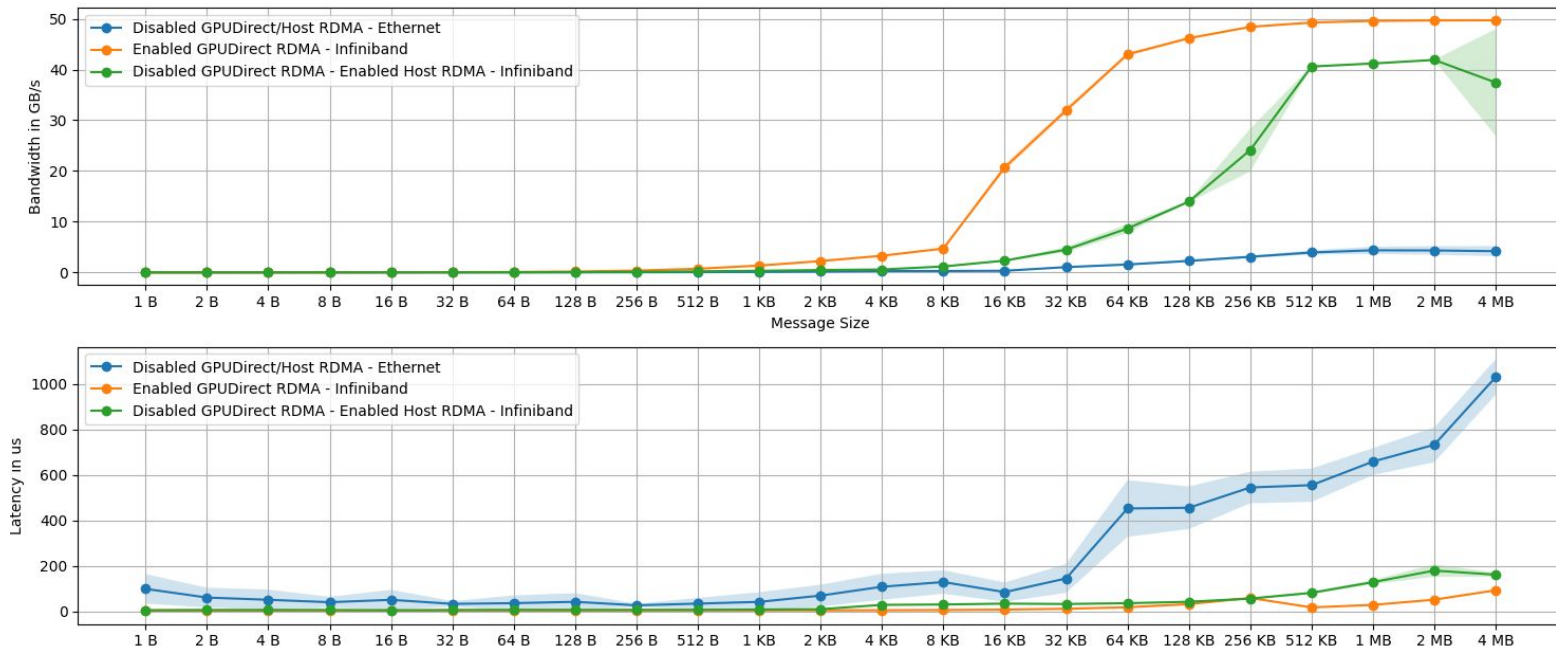
dma-buf

GPUDirect RDMA



Standard Benchmarks + User Workloads are used for Validation

OSU Benchmark - GPU2GPU - NVIDIA-H100-80GB-HBM3 - 2 Nodes/1 GPU



Workload Offloading

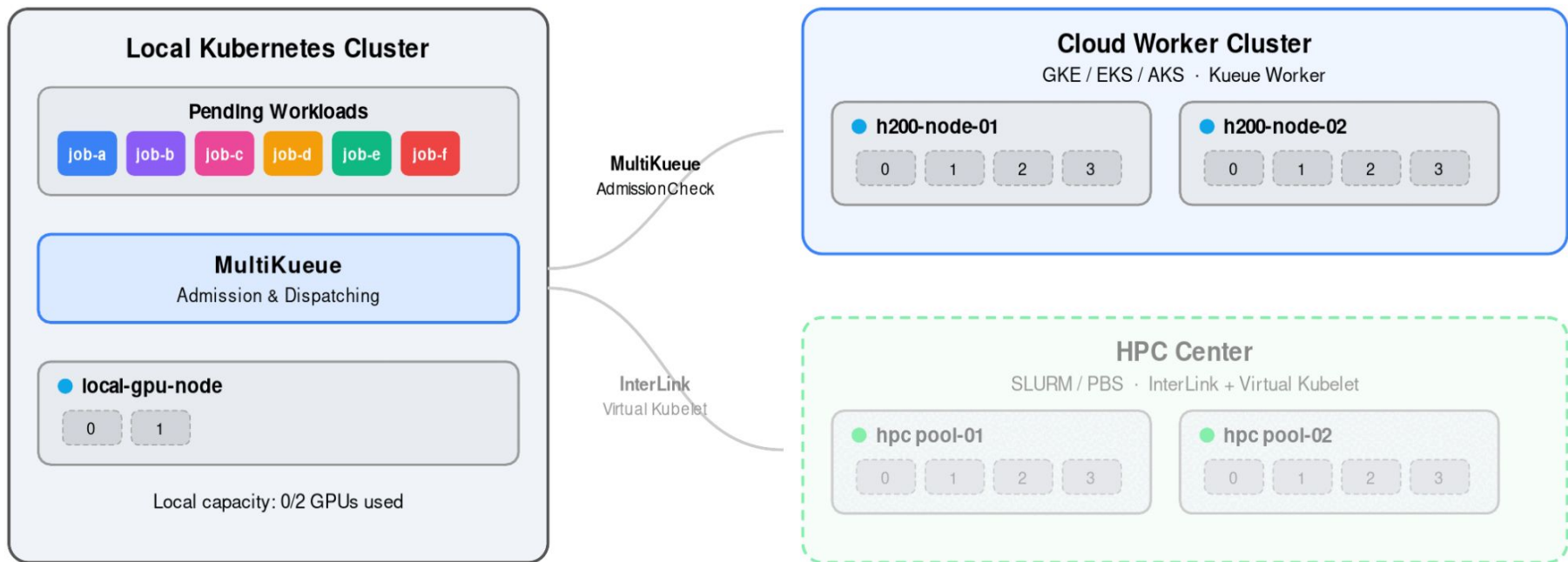


KubeCon



CloudNativeCon

Europe 2026



Workload Offloading

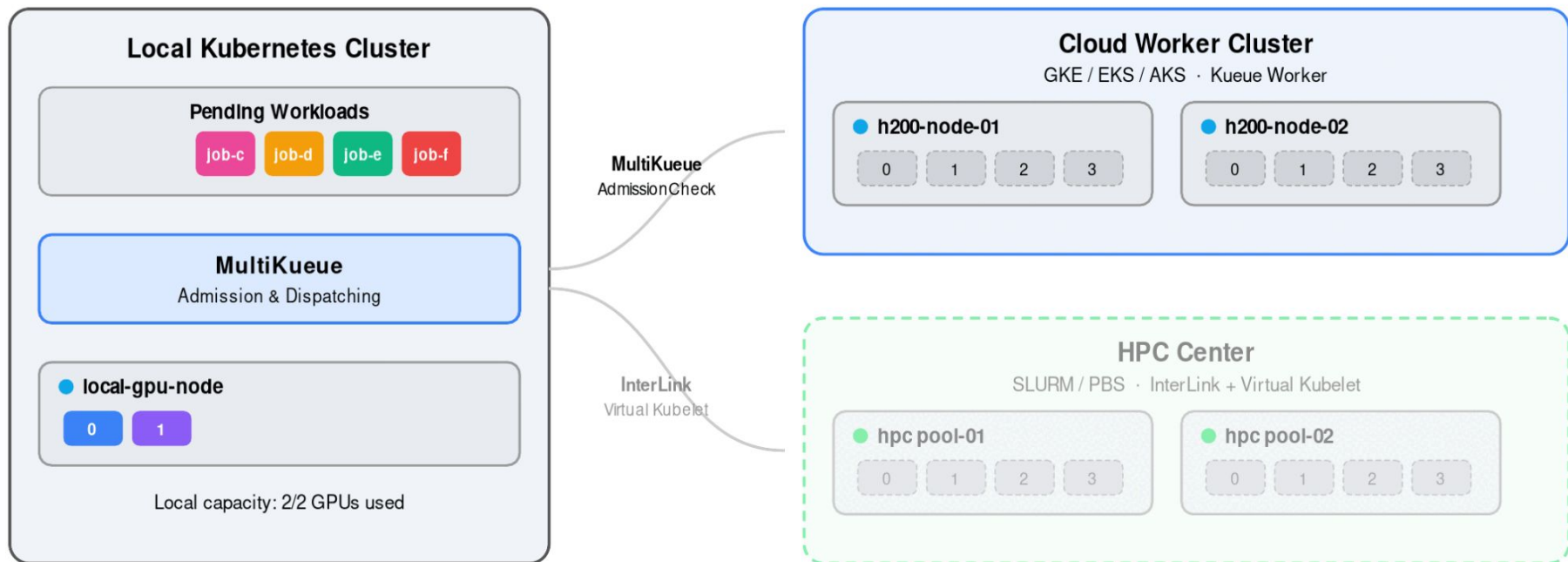


KubeCon



CloudNativeCon

Europe 2026



Workload Offloading

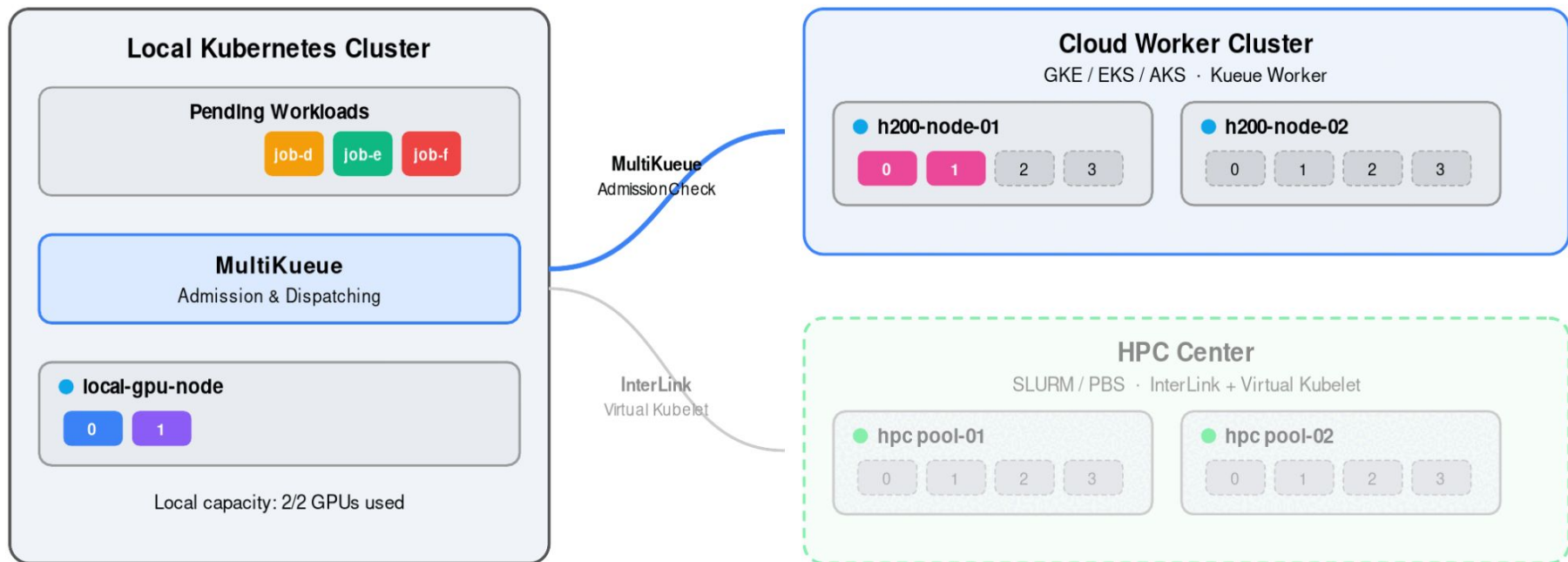


KubeCon



CloudNativeCon

Europe 2026



Workload Offloading

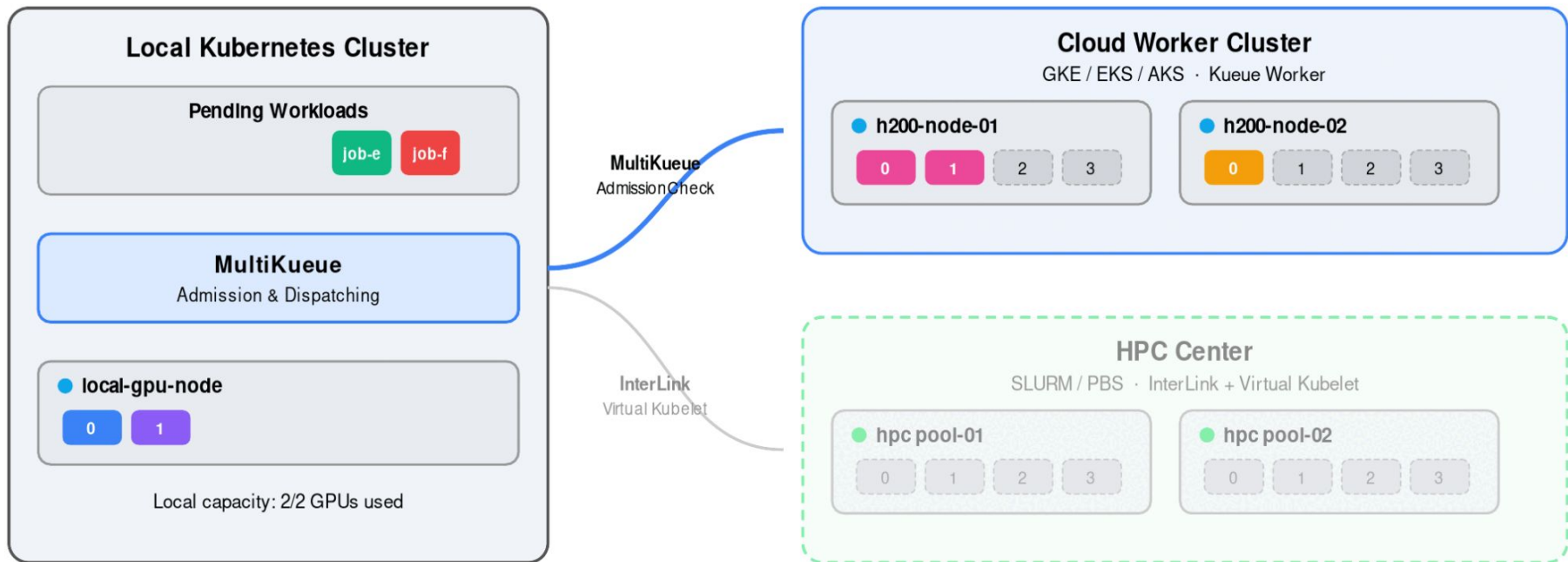


KubeCon



CloudNativeCon

Europe 2026



Workload Offloading

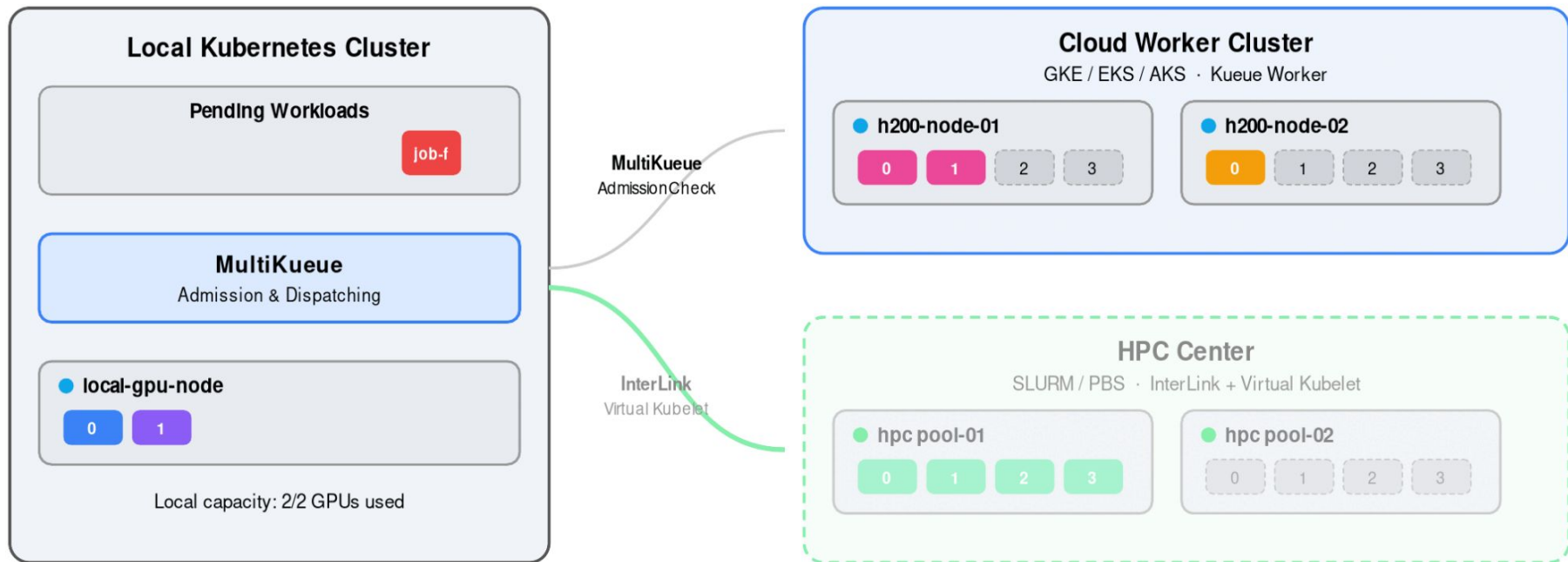


KubeCon



CloudNativeCon

Europe 2026



Workload Offloading

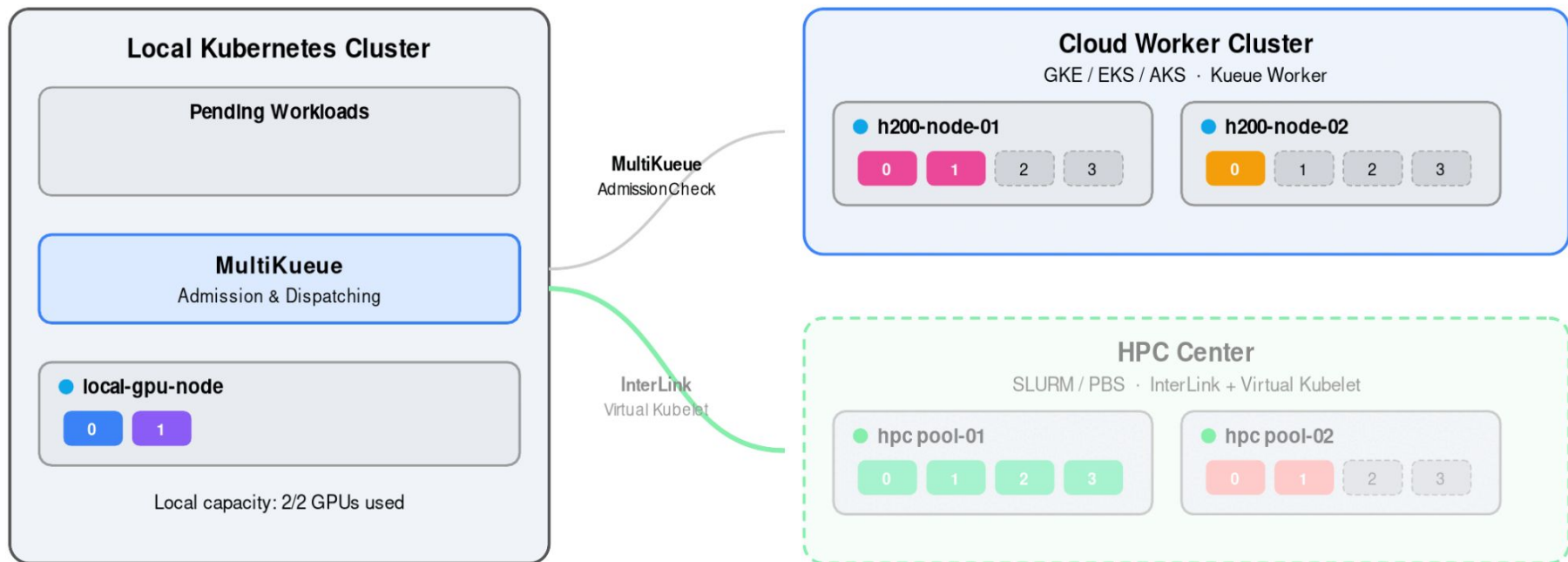


KubeCon



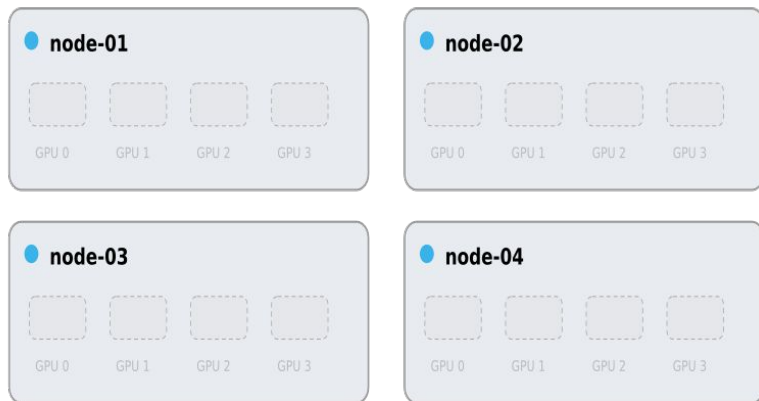
CloudNativeCon

Europe 2026

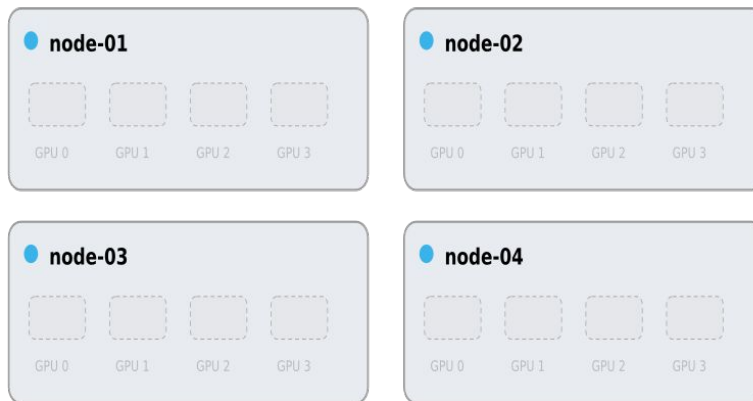


Binpack scheduling is used to allow **full node access** to users

LeastAllocated (default)

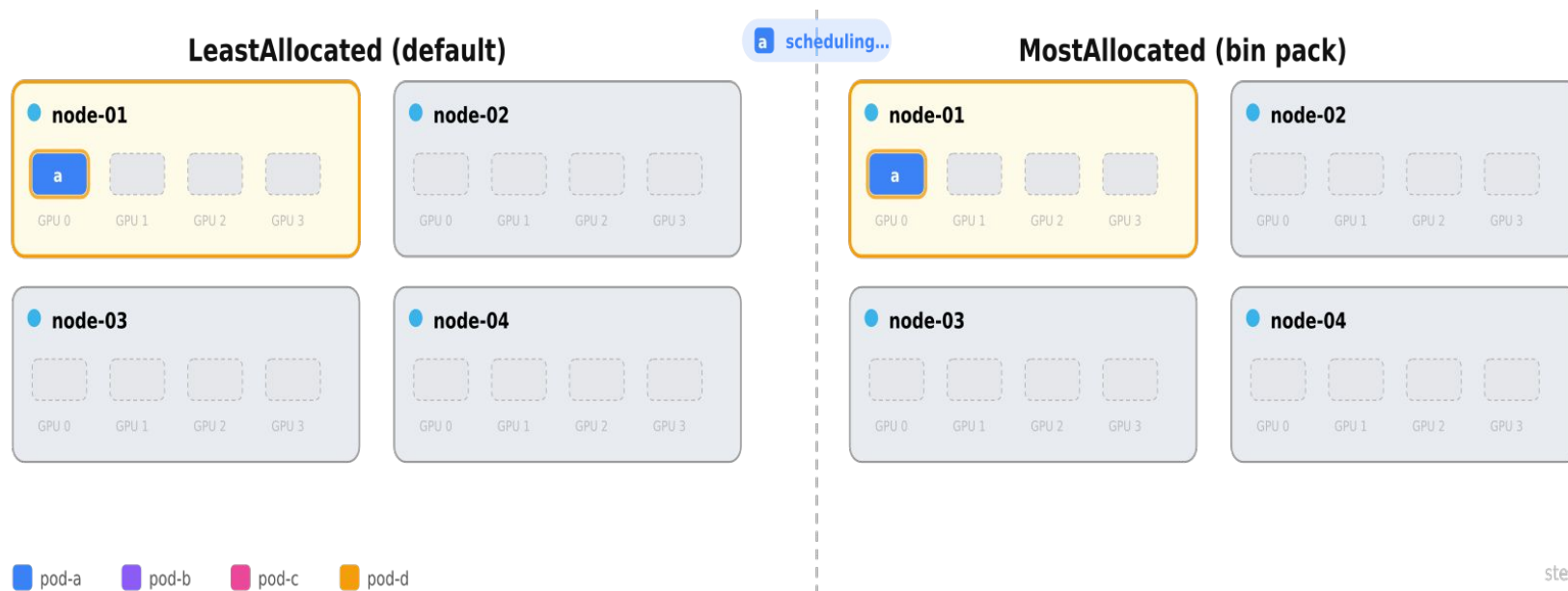


MostAllocated (bin pack)

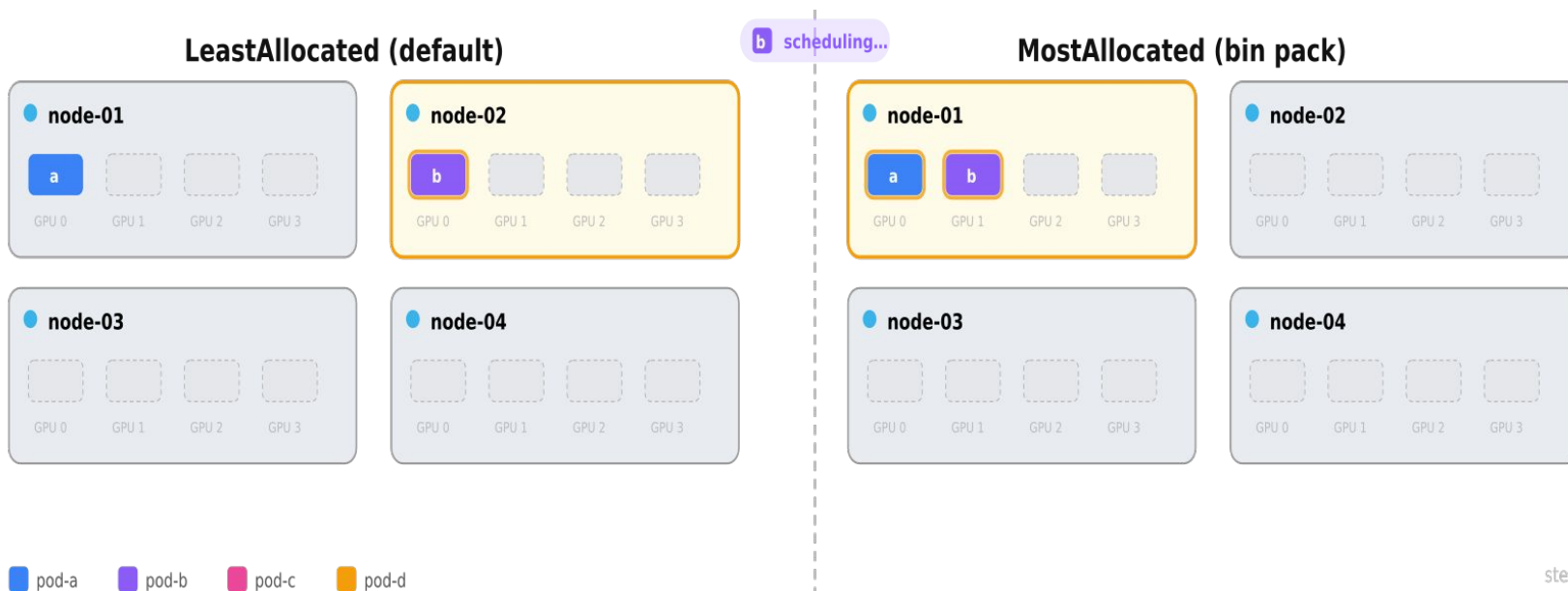


Scheduling

Binpack scheduling is used to allow **full node access** to users



Binpack scheduling is used to allow **full node access** to users



Scheduling



KubeCon

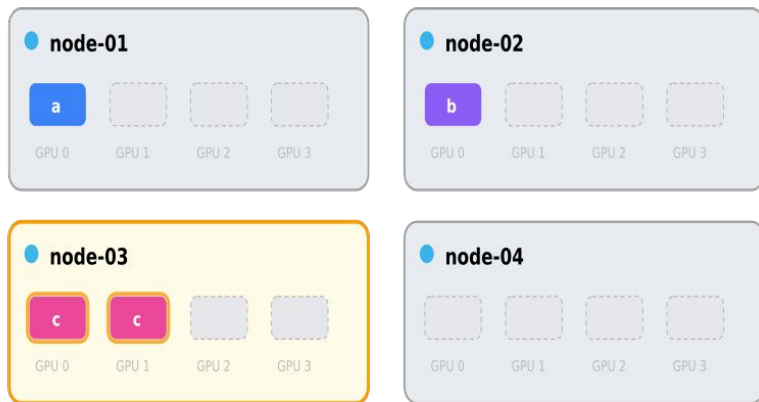


CloudNativeCon

Europe 2026

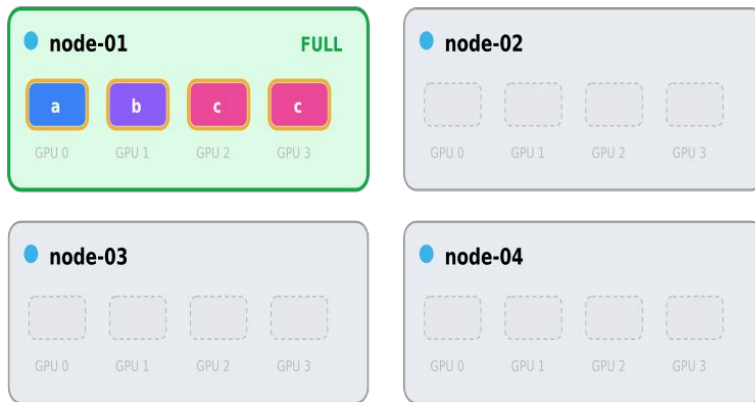
Binpack scheduling is used to allow **full node access** to users

LeastAllocated (default)



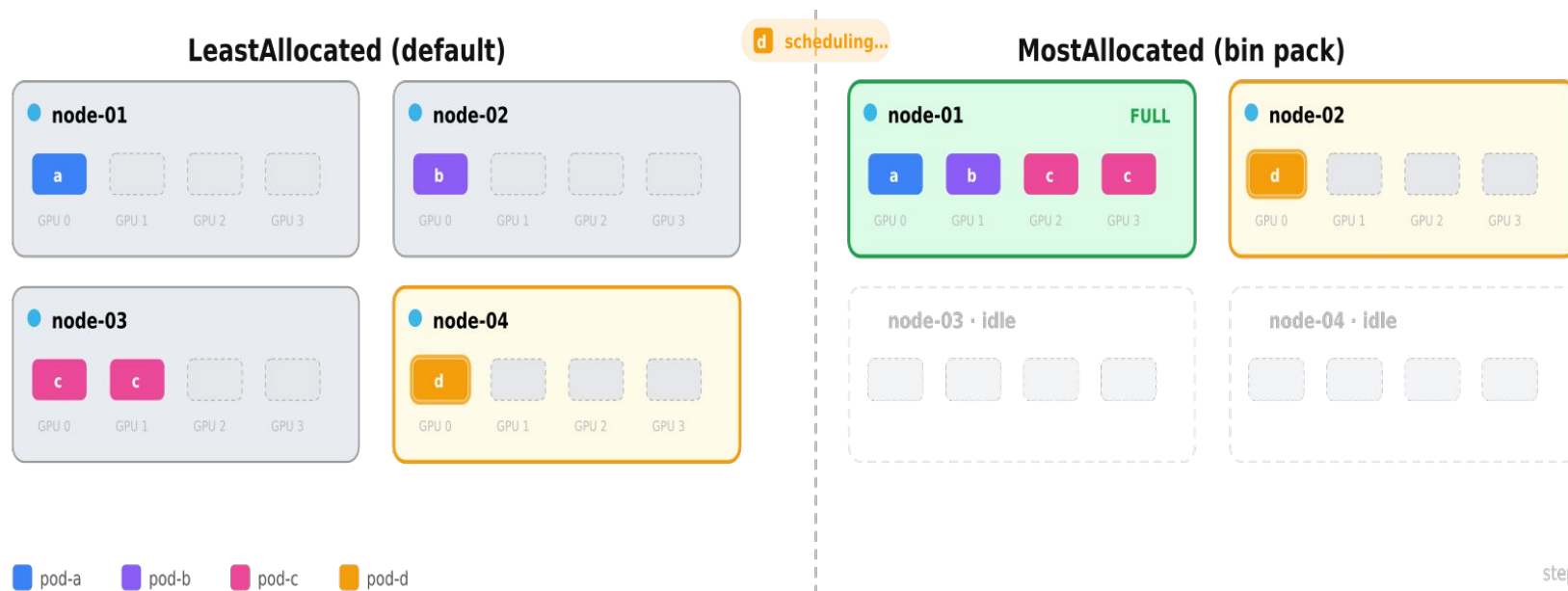
c scheduling...

MostAllocated (bin pack)



Scheduling

Binpack scheduling is used to allow **full node access** to users



Observability

We monitor overall cluster usage



Grafana

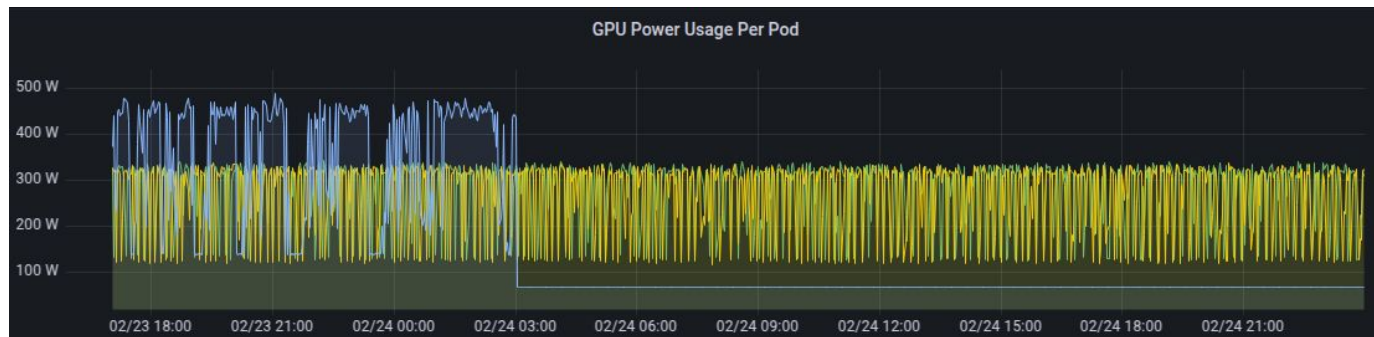
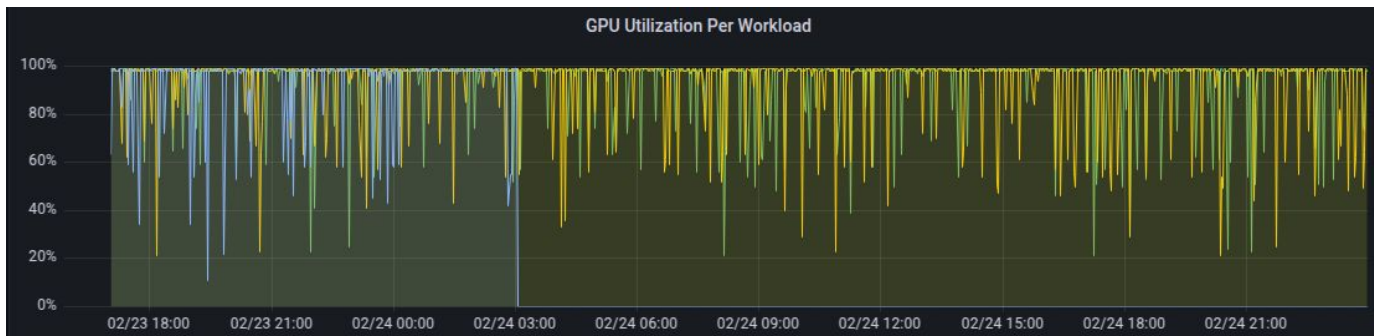


Prometheus



Find out more!

Detailed workload usage to give **insights to users**
(Power consumption, GPU/CPU/Memory usage)



Find out more!

[illegible]

```
apiVersion: v1
kind: Pod
metadata:
  name: session-1
  labels:
    home: "eos"
    mount-eos: "true"
    rdma: "true"
    nic-vendor: "broadcom"
spec:
  containers:
    - name: container
      image: registry.cern.ch/ngt/lxplus-like:9
      command: ["/usr/bin/tini", "--", "sleep", "infinity"]
```

Kyverno Mutation and Validation Policies

User configuration via labels

The success of the platform depends on acceptance by the users



Expose familiar interfaces



Provide known environments



Iterate quickly & showcase new features!

The success of the platform depends on acceptance by the users



Contribute features upstream alongside local patches



Constantly exchange with your users



Cloud native is ready for scientific computing and AI/ML!

Challenges and Outlook



KubeCon



CloudNativeCon

Europe 2026

- Multi-cluster Scheduling
- Low-latency networking
 - SR-IOV, DRANET
- GPU failure detection
- Bursting To HPC Resources
- DRA & Automated Partitioning
- Checkpoint & Restore
- Automated benchmarking on different hardware configurations



Checkpoint &
Restore WG

Contact Us



KubeCon



CloudNativeCon

Europe 2026

Reference Architecture



Email Us



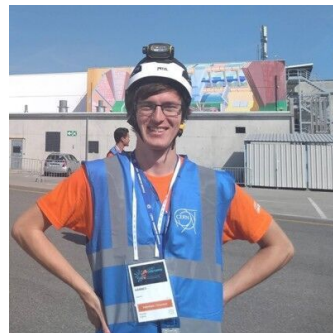
Raulian-Ionut Chiorescu



Slack



Hannes Hansen



Slack



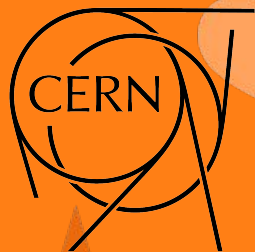


KubeCon



CloudNativeCon

Europe 2026



NextGen

