
KubeCon
CloudNativeCon
THE LINUX FOUNDATION
OPEN SOURCE SUMMIT
AI_dev
Open Source GenAI & ML Summit
China 2024

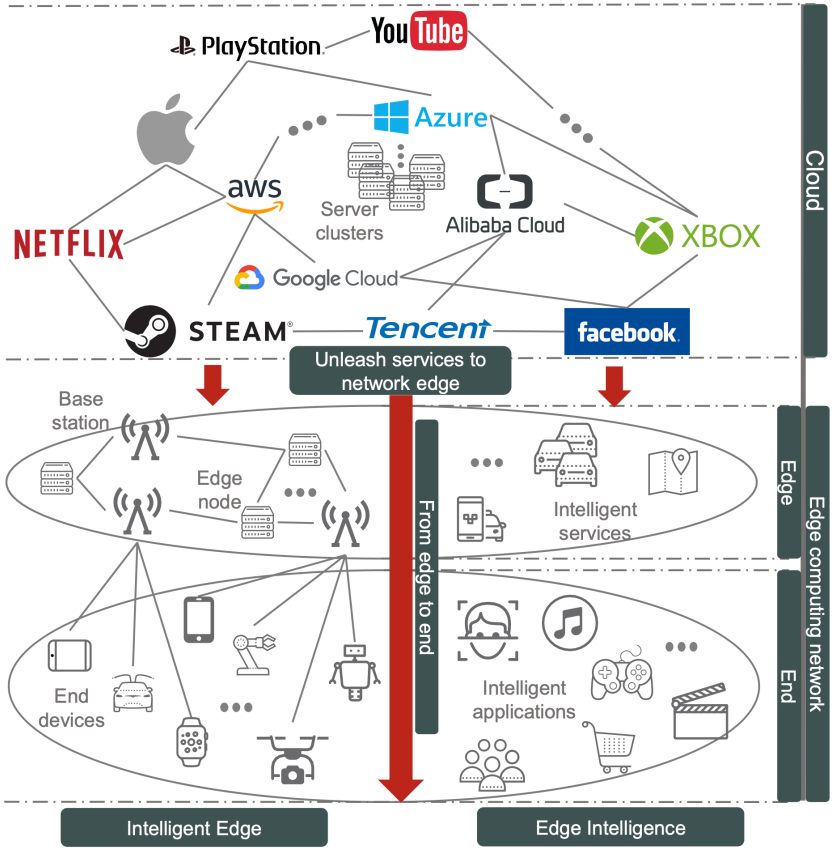

KubeCon
CloudNativeCon
THE LINUX FOUNDATION
OPEN SOURCE SUMMIT
AI_dev
Open Source GenAI & ML Summit
China 2024
OpenYurt & Dragonfly: Enhancing Efficient Distribution of LLMs in Cloud-Edge Collaborative
Linbo He, Alibaba cloud
Jim Ma, Ant Group

Agenda

- Edge Computing and AI
- OpenYurt
- Dragonfly
- Practice

Edge Computing and AI

- The evolution of large models is rapid, and edge AI will be the next frontier.
- Cloud-edge collaborative AI computing power deployment to address diversified challenges.
- The efficiency issue on the edge side is prominent, and agile development is key point.



Refer: https://arxiv.org/pdf/1907.08349.pdf

OpenYurt and Dragonfly

OpenYurt, as the industry's first non-intrusive open-source project for edge computing cloud-native platforms, can address challenges in various aspects such as edge autonomy, edge networking, and edge storage.

At the same time, in terms of model distribution, Dragonfly currently supports accelerated distribution of models for various applications.

The collaboration between OpenYurt and Dragonfly can provide efficient and lightweight deployment of AI applications.

Challenges

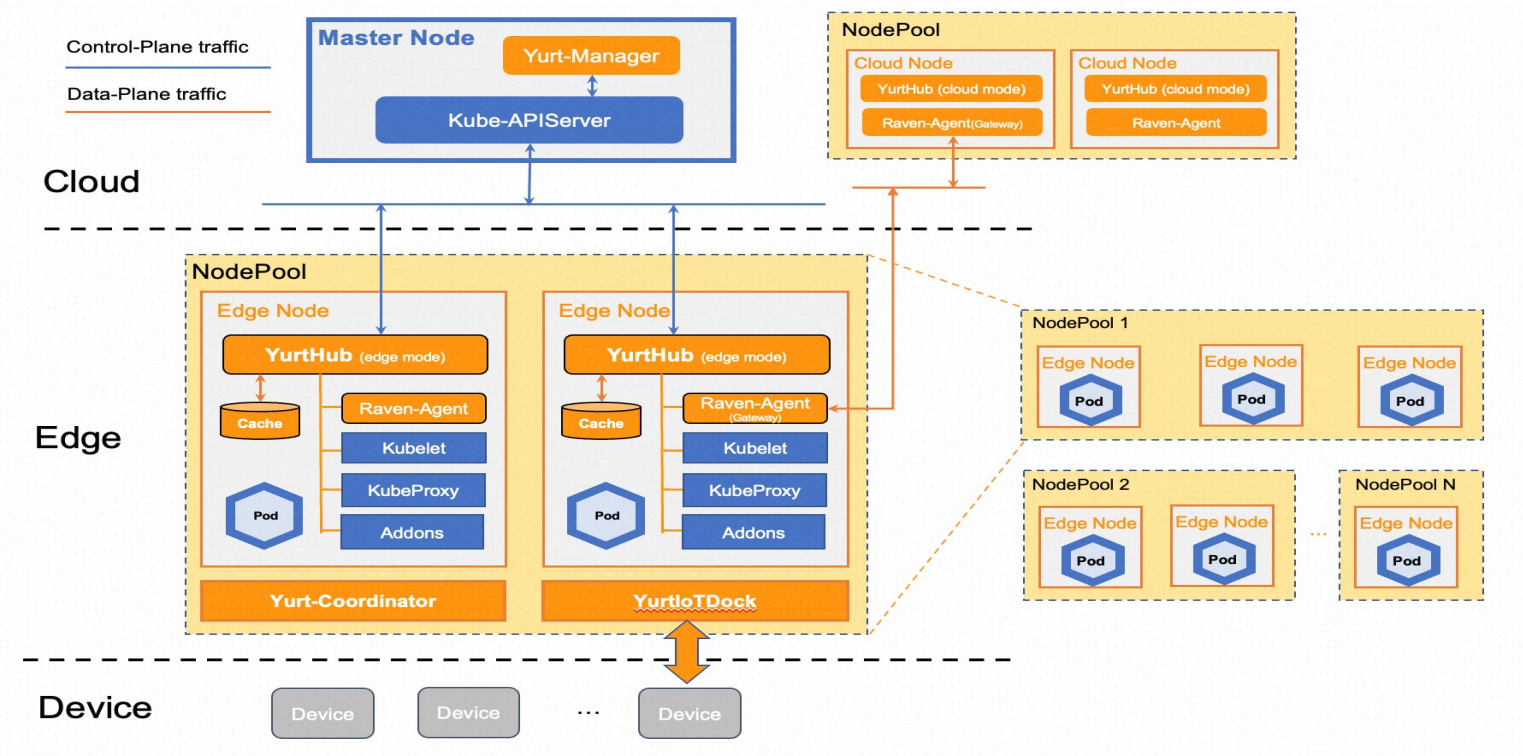
Distribute LLM to edge nodes in multiple regions, The challenges are faced by OpenYurt and Dragonfly as following:

- OpenYurt
  - distribute workloads to multiple regions with different configurations.
  - how to expose LLM service in multiple regions?
  - how to reduce control-plane bandwidth in large scale LLM application management?

- Dragonfly
  - how to load LLM images in multiple regions from cloud efficiently?
  - how to integrate deployment of OpenYurt and Dragonfly?

What's OpenYurt?

- OpenYurt is the industry's first edge computing platform that requires no modifications to the Kubernetes system.
- Through the control-plane located in the cloud, it centrally manages massive edge resources(such as CDN sites) across various locations.
- OpenYurt helps users to easily complete large-scale application deployment, operation, maintenance.



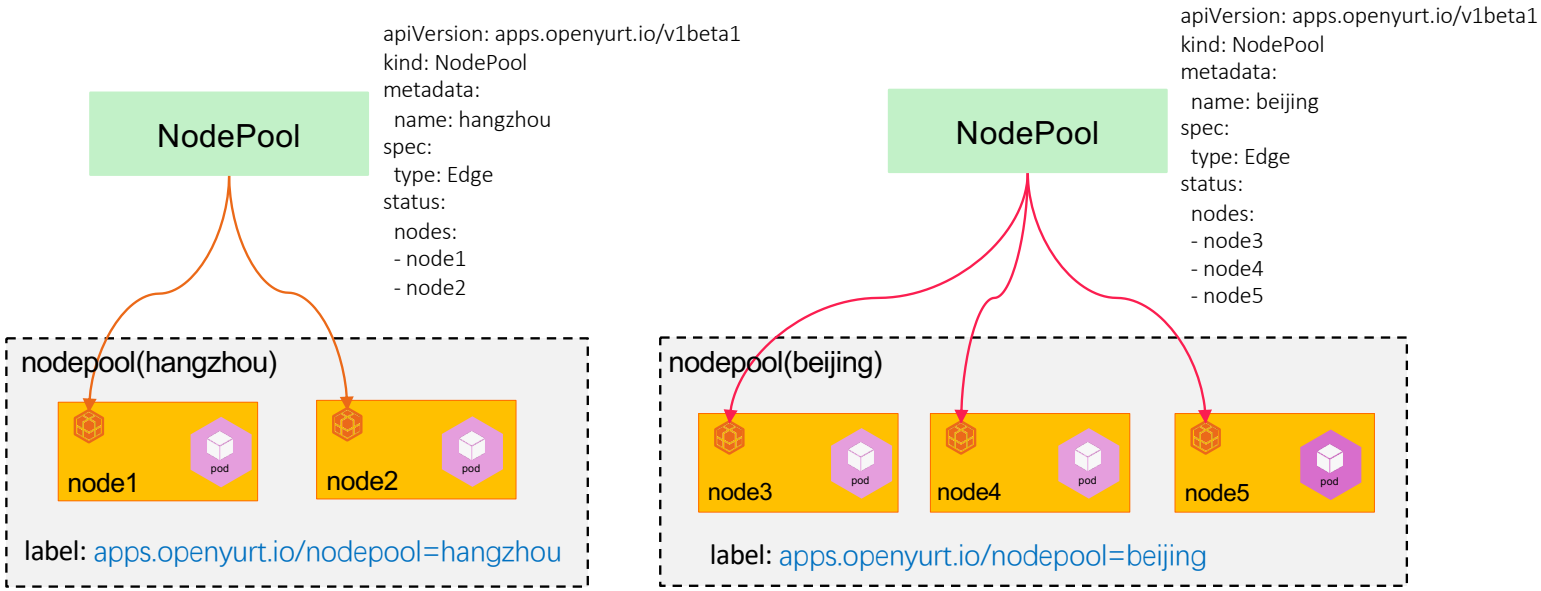
Key Features

➢ Key Features:

- Powerful edge autonomy capability: Ensures edge applications can provide continuous service 24/7.
- Multi-Region workload management: support multi-regional workloads with diverse configurations, along with regional application rollout, unified traffic access, and bandwidth reduction.
- Cross Region network communication capability, while ensuring compatibility with the native Container Network Interface(CNI)
- Cloud-native edge device management: Integrate new version of EdgeX Foundry automatically.

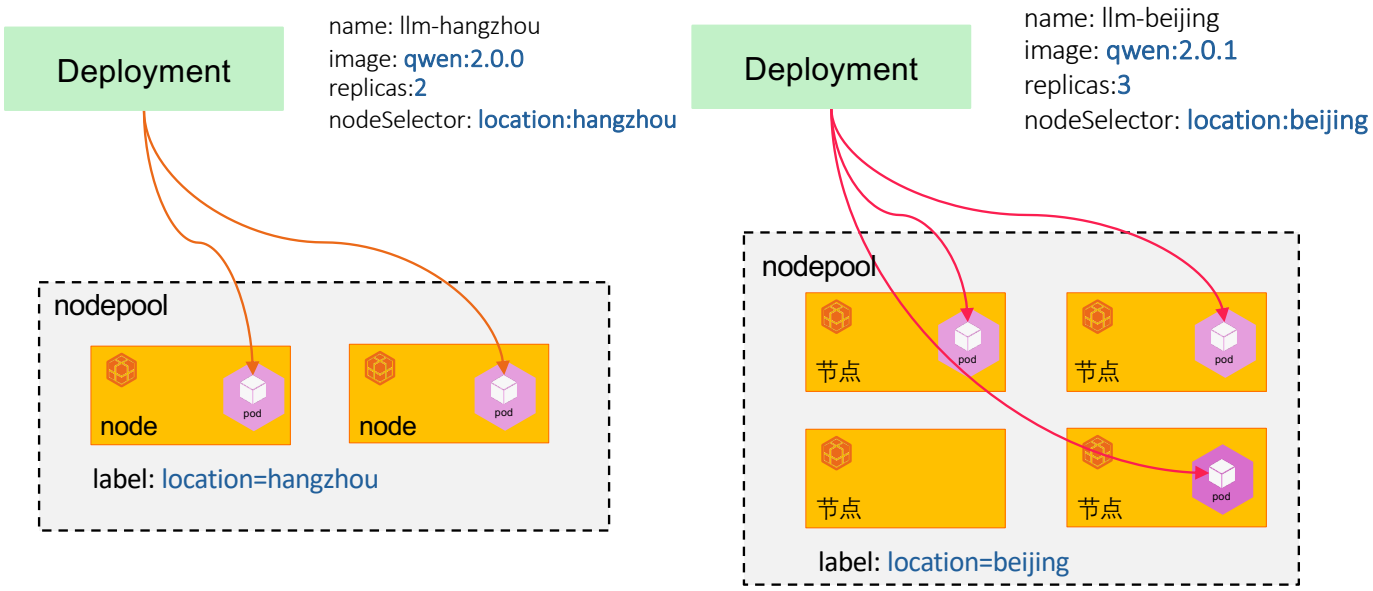
Multi-Region Nodes Management

- A new CRD named NodePool is added for managing nodes(like labels, annotations, taints of node) in different regions.
- NodePool is the first citizen in OpenYurt, If a node is joined into nodepool, apps.openyurt.io/nodepool=pool-name label is added for the node.



Multi-Region Workload Management

In native Kubernetes, use multiple deployments to deploy an application on nodes in different regions as following figure.

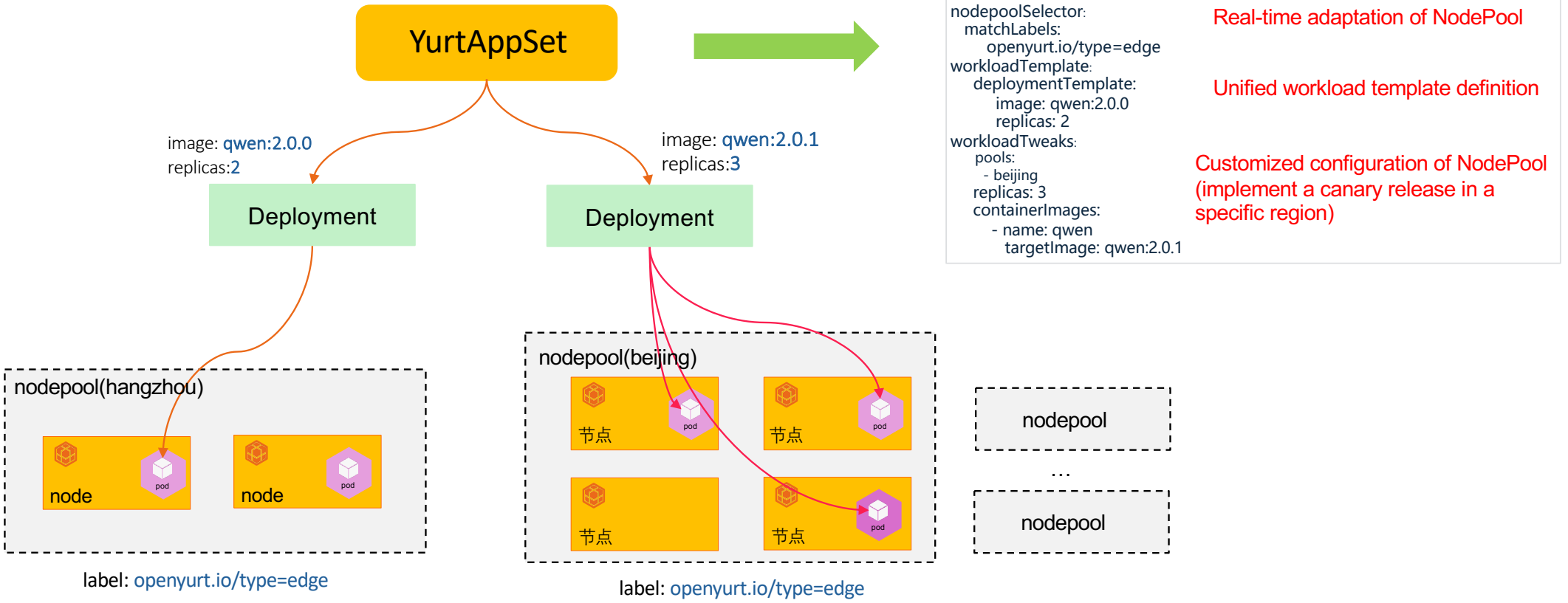


➢ Challenges:

- Complex application maintenance: You need to manually distinguish and maintain Deployments in different regions.
- Redundant application configurations: The configurations of Deployments in different regions are highly similar, resulting in complex and error-prone configuration management.

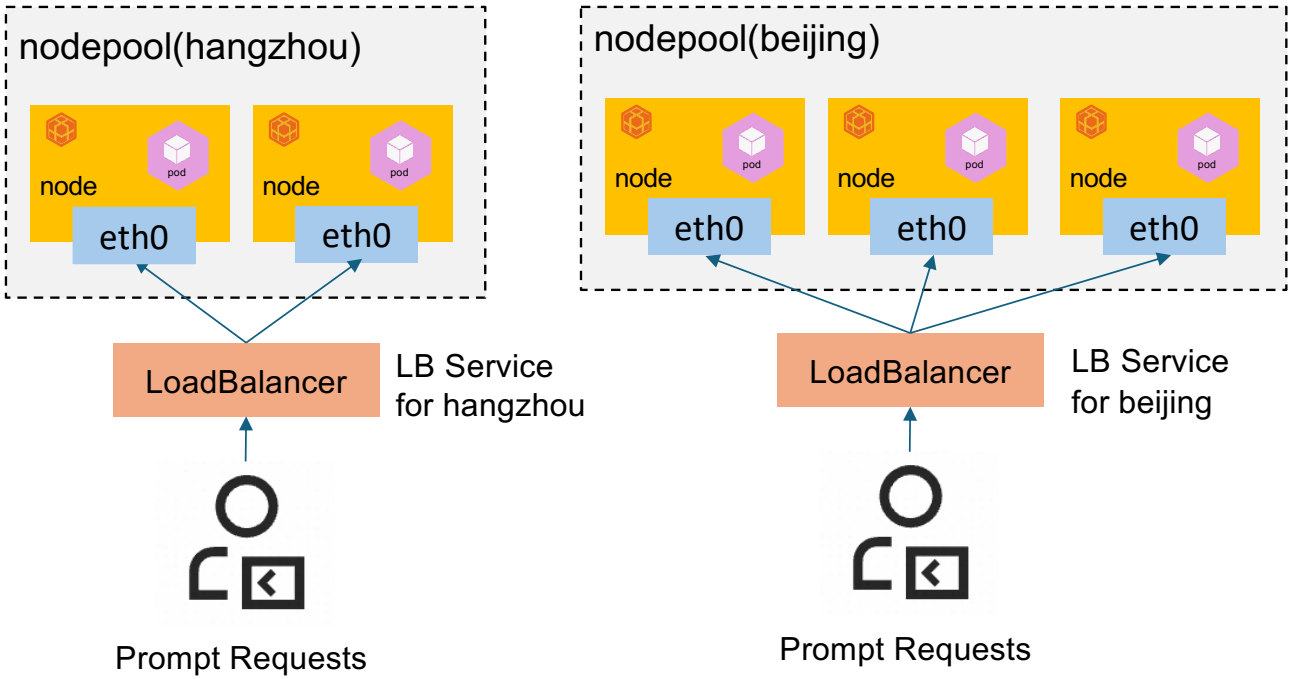
Multi-Region Workload Management

In OpenYurt, CRD YurtAppSet is provided to reduce the complexity of distributed deployment in multiple regions. users can use YurtAppSet to create, update, and delete multiple Deployments in a centralized manner.



Multi-Region LoadBalancer Management

In native Kubernetes, A individual LB service with different label selector should be created for each nodepool, then pod labels in each nodepool should be managed carefully in order to adapt the correct LB service.



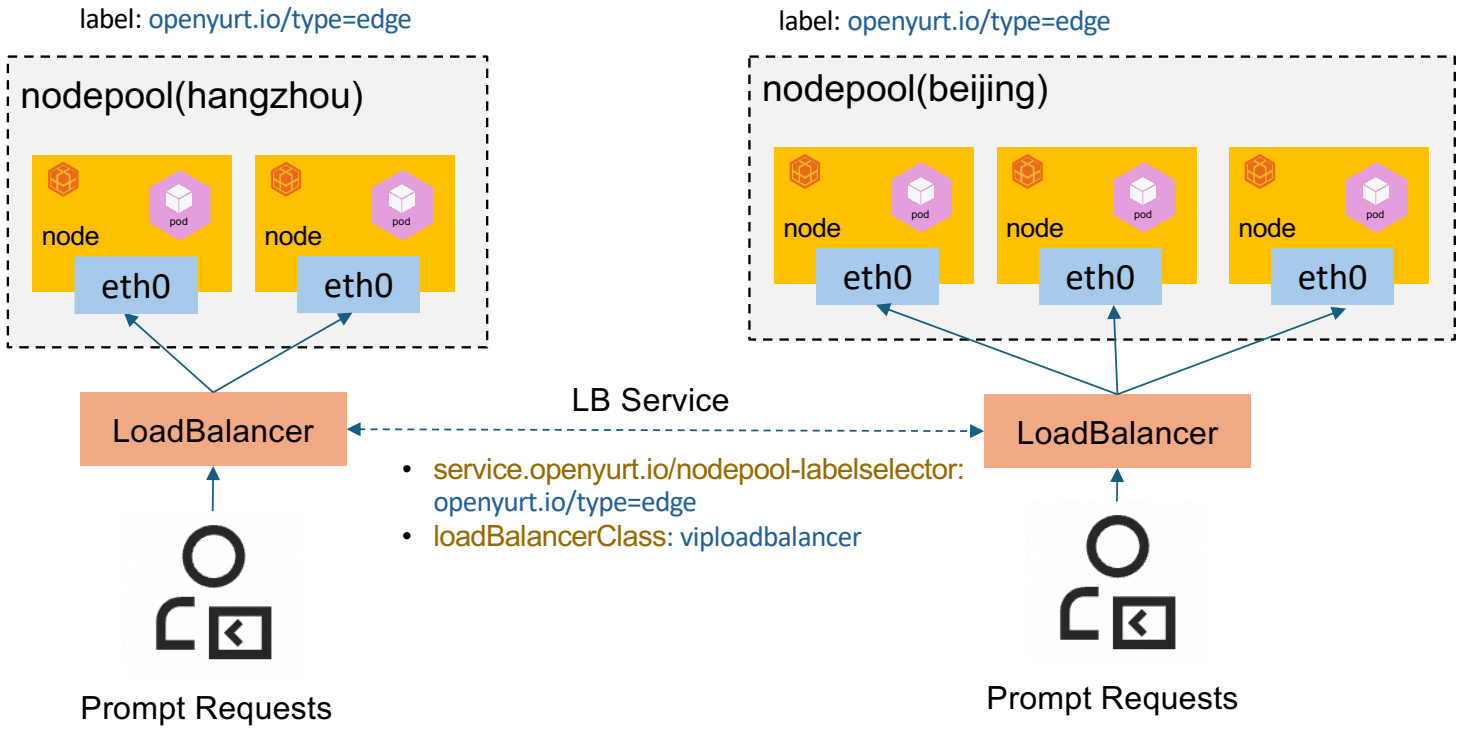
➢ Challenges:

- Multiple LB services Management: You need to make sure each LB service match the corresponding pods.
- Dynamic new/delete LB services: Pods managed by YurtAppSet are automatically created/deleted based on the creation/deletion of the nodepool. But you need to create/delete LB service manually.

Multi-Region LoadBalancer Management

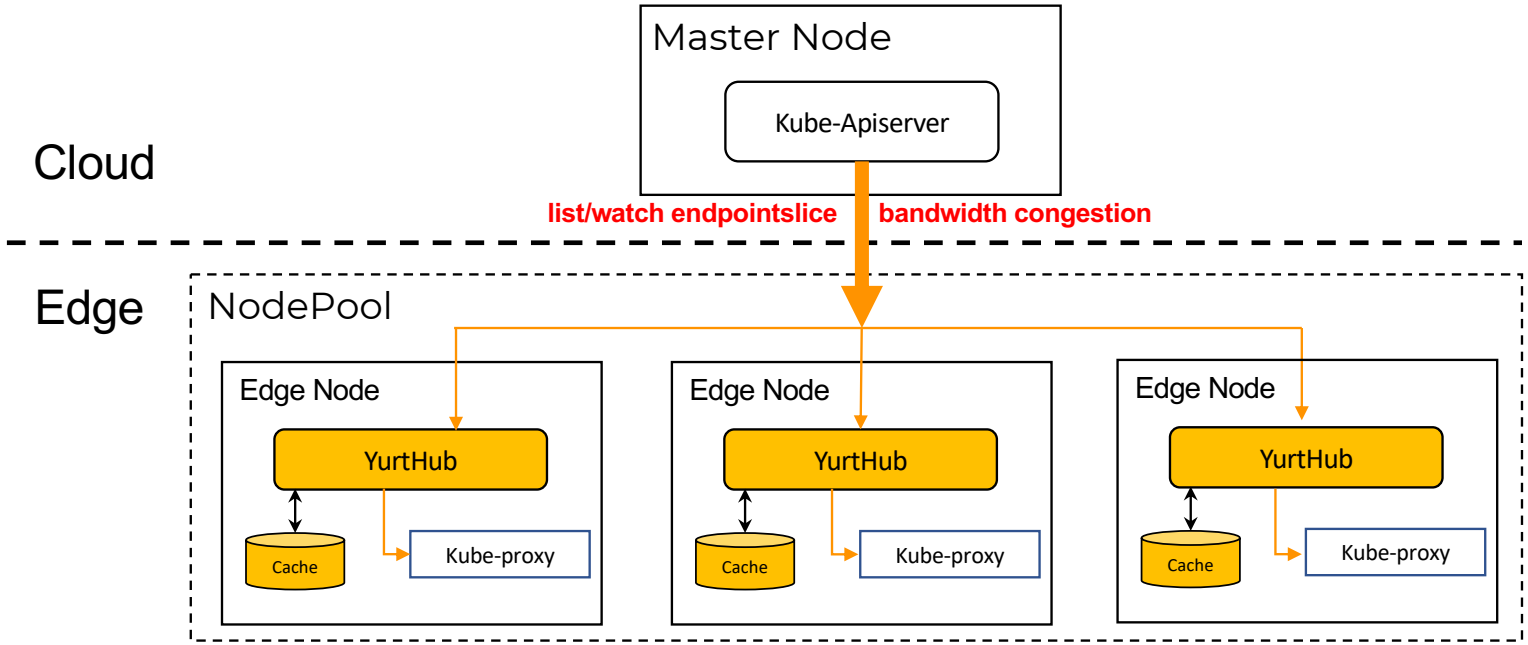
OpenYurt will provide a feature called LoadBalancerSet to dynamically scale the LoadBalancer service to multiple node pools and work with YurtAppSet.

- Load Balancer service with the annotation service.openyurt.io/nodepool-labelselector will be recognized as LoadBalancerSet service and load balancers will be created for all node pools matching this selector.
- viploadbalancer class will be supported to provide keepalived and virtual IP (VIP) based LoadBalancer access points on edge nodes.



Control-Plane Bandwidth Management

- Our tests revealed that during application upgrades, the outbound bandwidth of the cloud can easily reach its maximum capacity.
- This is primarily due to the large-scale creation and deletion of applications, which leads to frequent changes in EndpointSlice resources. Moreover, every change in EndpointSlice must be pushed to each edge node, inevitably increasing the demand for bandwidth.

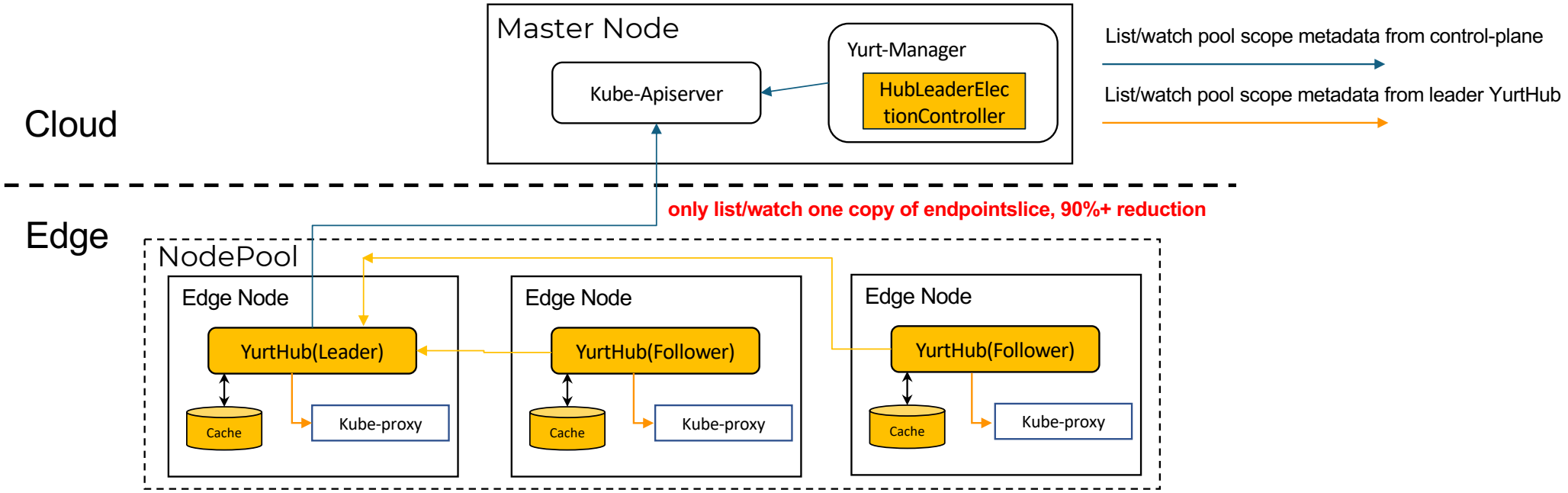


- Pool Scope Metadata:
The metadata from cloud of each node in the NodePool is the same, such as endpointslice. The main cause for bandwidth congestion.

- Node Scope Metadata:
The metadata of each node in the NodePool is unique, such as pods, secrets, etc.

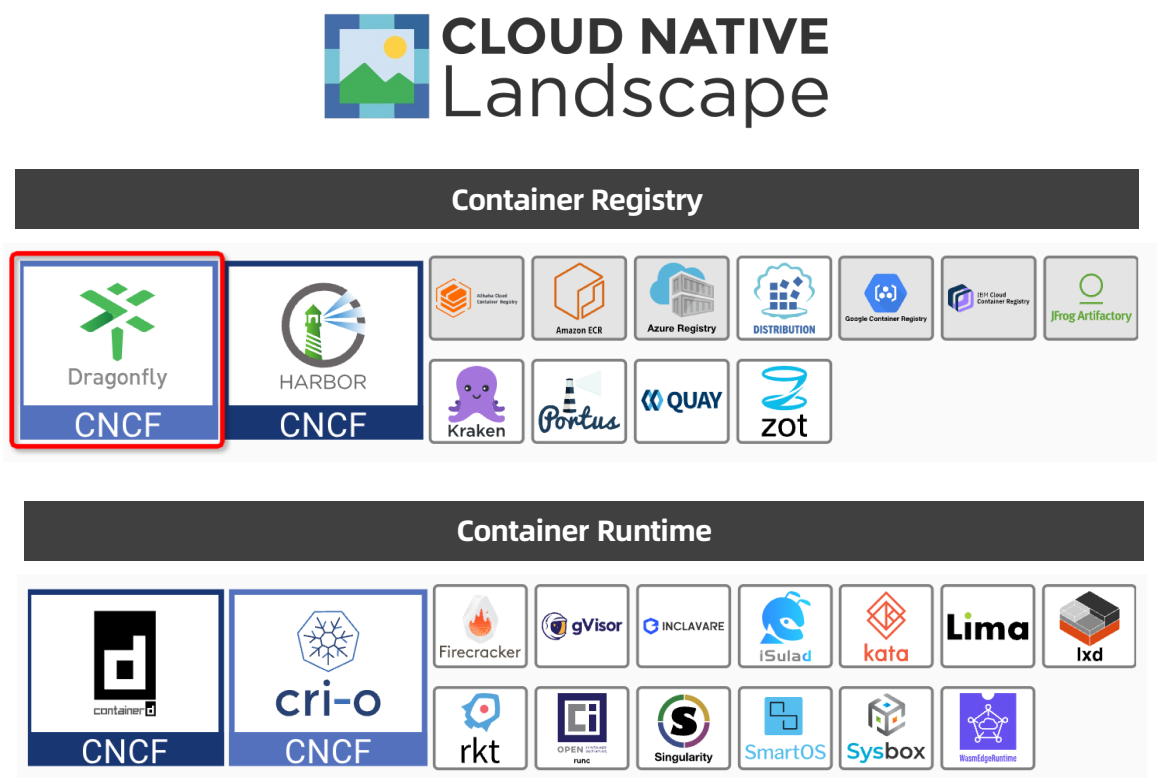
Control-Plane Bandwidth Management

- In the nodepool, all nodes read the same metadata, such as EndpointSlice, which is referred to as pool scope metadata.
- A new controller named HubLeaderElection is used for selecting the Leader Yurthub in the nodepool. There are several election policies(such as random, mark, etc.) are supported.
- Leader YurtHub will read the pool scope metadata from the cloud and store them in local cache.
- Requests from clients(like kube-proxy) go through YurtHub component, and follower YurtHub redirects these requests to leader Yurthub, eliminating the use of public network bandwidth.



Dragonfly Introduction

KubeCon CloudNativeCon China 2024 | THE LINUX FOUNDATION OPEN SOURCE SUMMIT | AI_dev



What is Dragonfly?

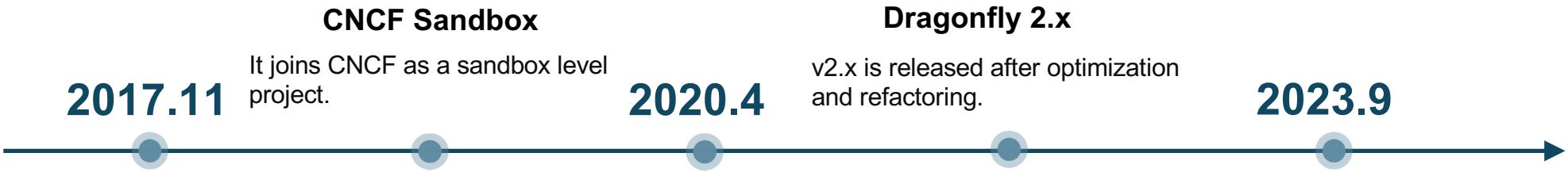
Provide efficient, stable and secure file distribution and image acceleration based on p2p technology to be the best practice and standard solution in cloud native architectures.

It is hosted by the Cloud Native Computing Foundation(CNCF)as an ***Incubating*** Level Project.

There are more than ***100*** contributors, and maintainers come from ***Ant Group***, ***Alibaba Group***, ***ByteDance***, ***Intel***, ***Baidu Group***, ***Zhipu AI*** and ***Dalian University of Technology***.

Public cloud users include ***Alibaba Cloud(Aliyun)***, ***Google Cloud Platform (GCP)***, ***Volcano Engine***, ***Baidu AI Cloud***, etc.

Dragonfly Story



Dragonfly 1.x

It is open source from Alibaba Group.

CNCF Sandbox

It joins CNCF as a sandbox level project.

CNCF Incubation

It promoted to a CNCF incubation level project.

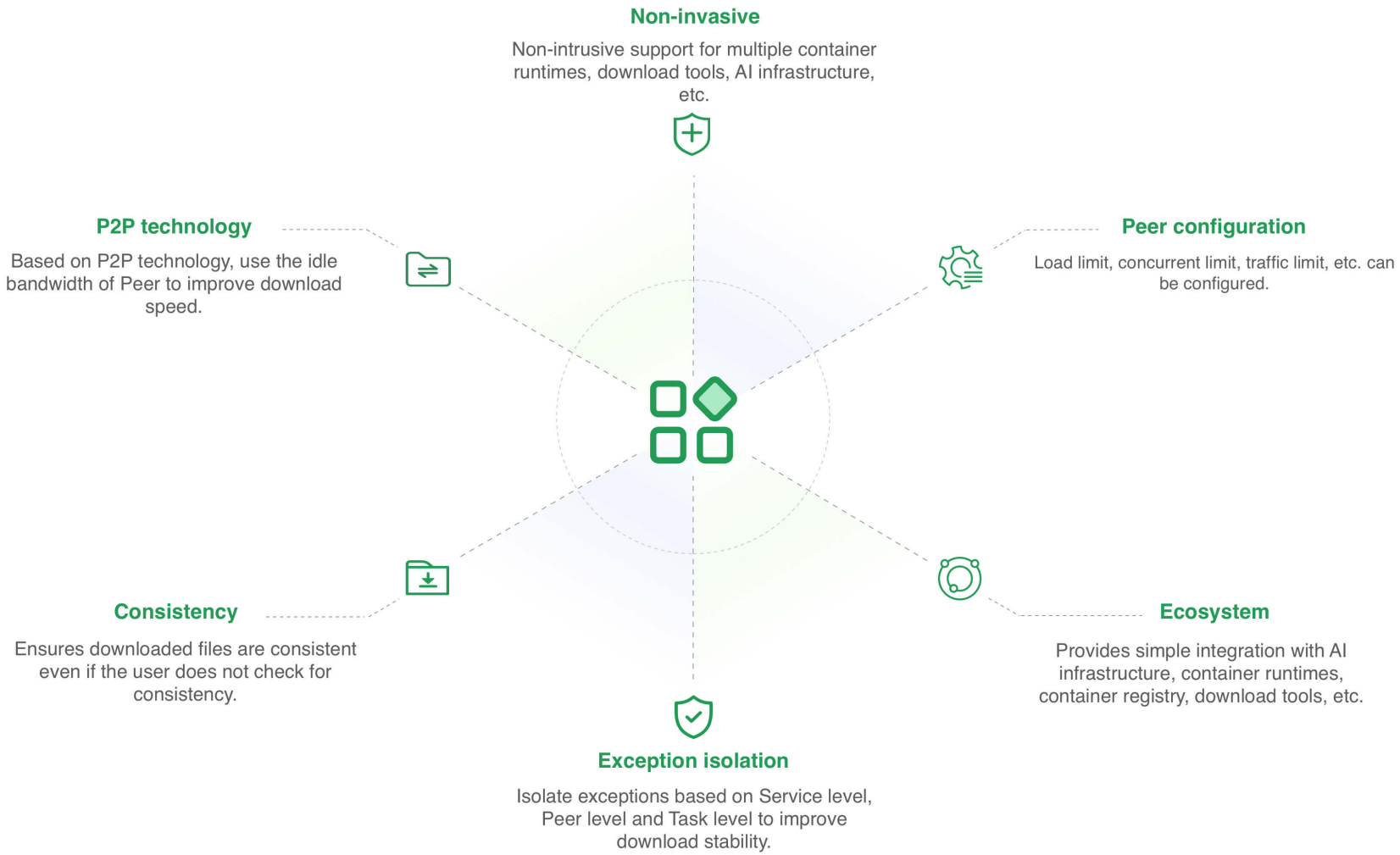
Dragonfly 2.x

v2.x is released after optimization and refactoring.

CNCF Graduation Proposal

It submits the formal proposal for dragonfly graduation.

Dragonfly Features



Dragonfly Features Cont.

Image Acceleration

Dragonfly supports various container clients such as *containerd*, *Docker*, *cri-o*, *ORAS*, etc. It provides three solutions for image acceleration.

The first solution is to use Dragonfly to distribute images based on P2P technology, which is suitable for *large-scale cluster*.

The second solution is to use Dragonfly and Nydus to distribute accelerated images, which is suitable for *large-scale cluster* and *faster container launching*.

The third solution is to use Nydus to distribute accelerated images, which is suitable for *faster container launching*.

File Distribution

Dragonfly supports large-scale file distribution and uses P2P technology to *eliminate the impact of origin bandwidth limitations*.

It supports file distribution of protocols including *HTTP*, *HDFS*, etc. Additionally, it also supports different object storage protocols includes *S3*, *OSS*, *OBS*, etc.

Add *Dfstore* to expand the file distribution capability, it can depend on different types of object storage, such as S3, OSS, OBS, etc. to *provide stable object storage capabilities*.

AI Infrastructure

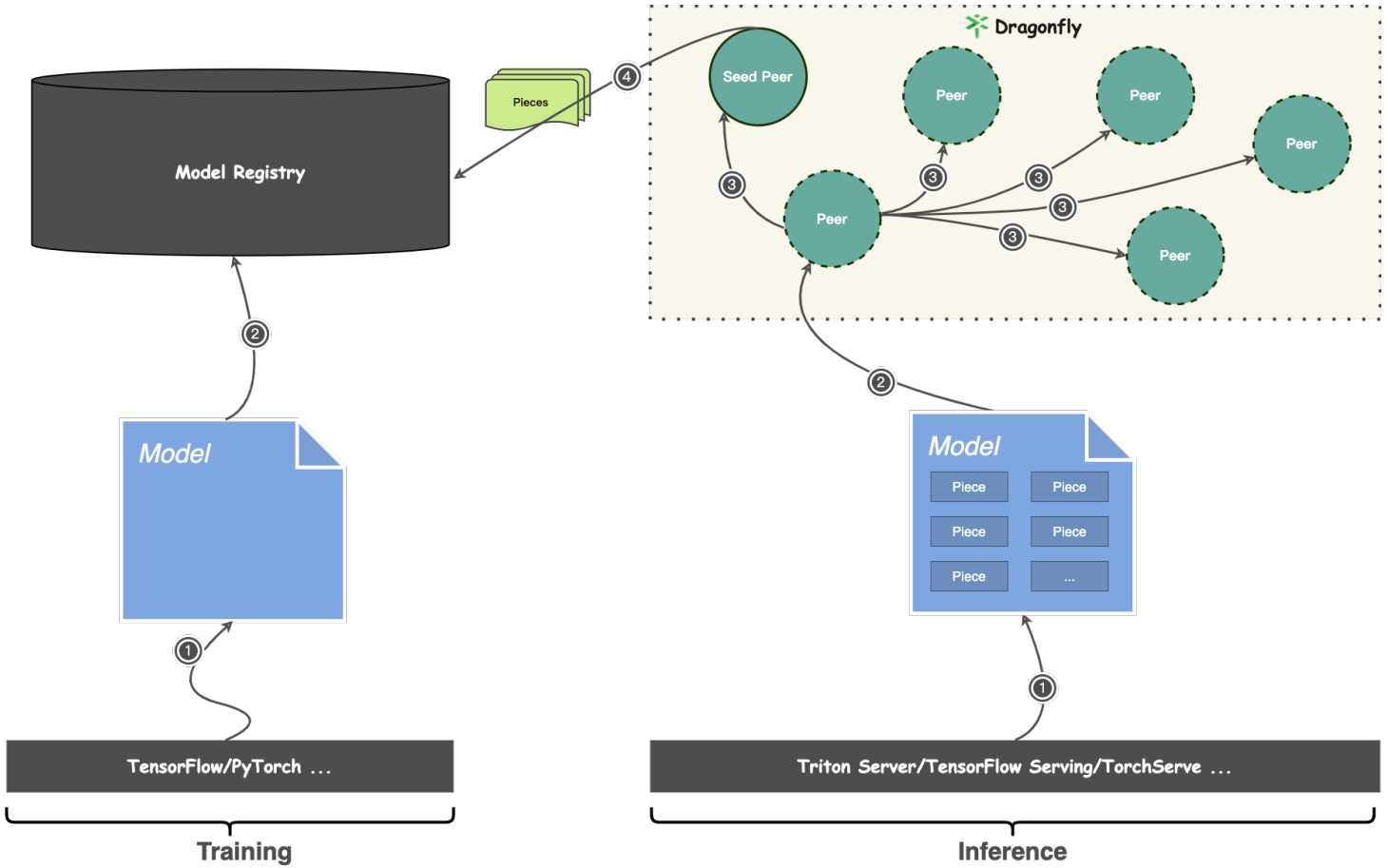
Dragonfly supports distributing data during *AI training* and *AI inference*.

In the AI inference, Dragonfly supports *Triton Server* and *TorchServe* to use Dragonfly distribution model.

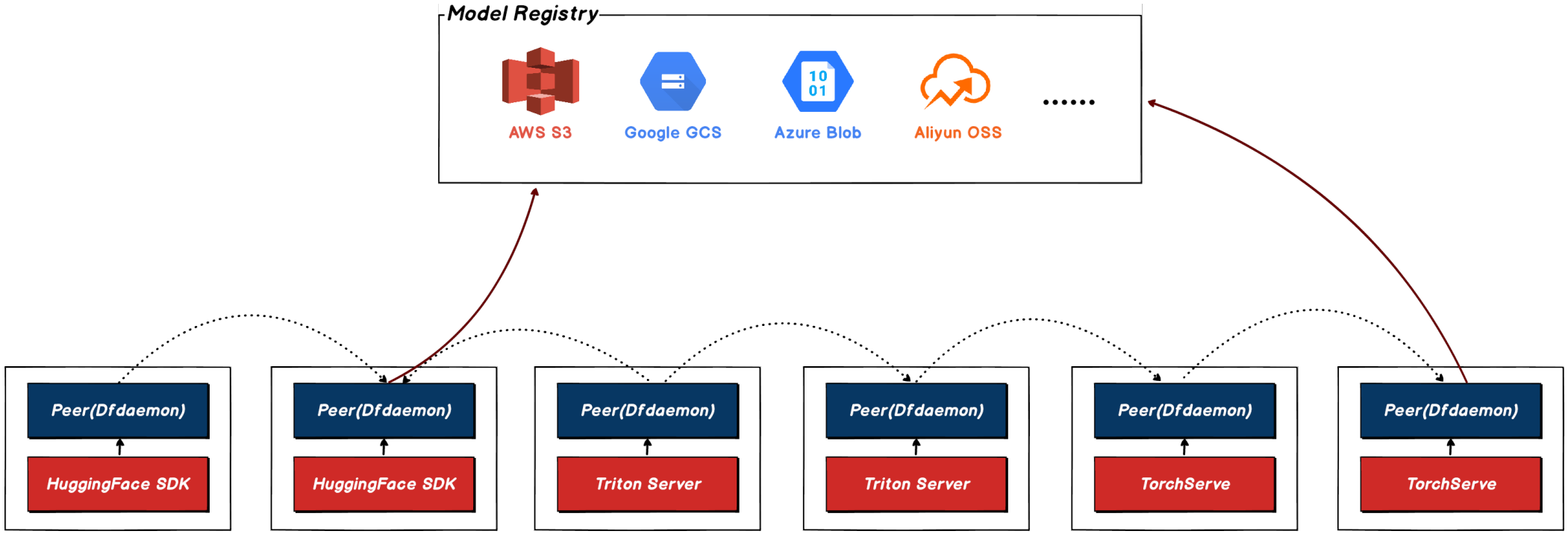
In the AI training, *Fluid* downloads dataset through Dragonfly when running based on *JuiceFS*.

Supports downloading models and datasets from *Hugging Face Hub by* SDK through Dragonfly HTTP proxy.

Dragonfly & AI Inference



Dragonfly & AI Inference Cont.



Dragonfly & Hugging Face

KubeCon | CloudNativeCon | THE LINUX FOUNDATION OPEN SOURCE SUMMIT | AI_dev
China 2024

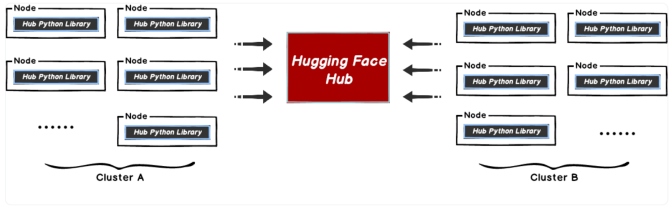
Hugging Face accelerates distribution of models and datasets based on Dragonfly

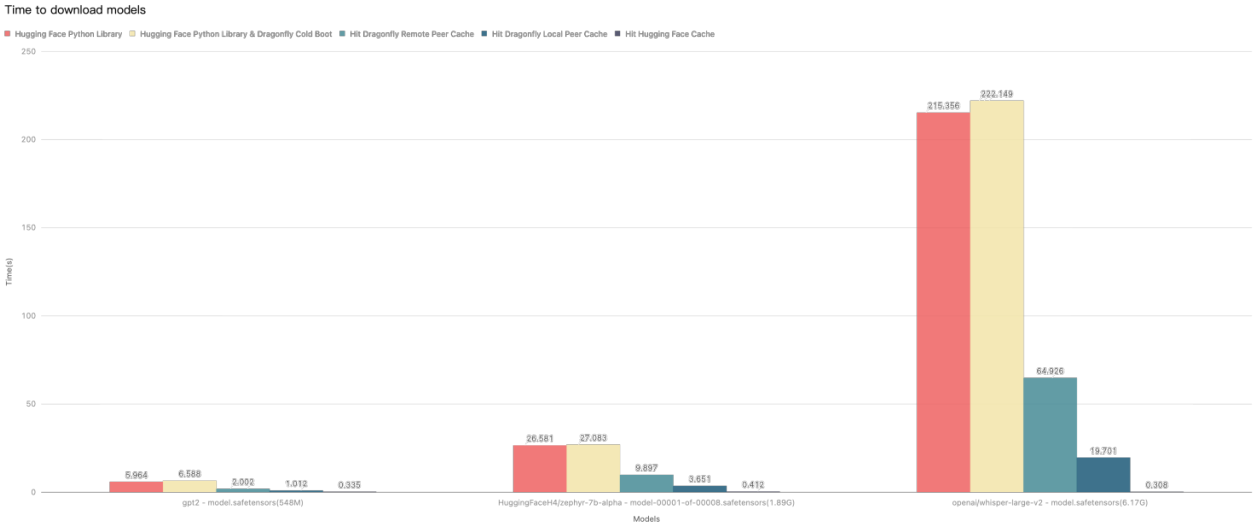
Community blog post Published November 8, 2023

Edit article

gaius-qi
Gaius Qi

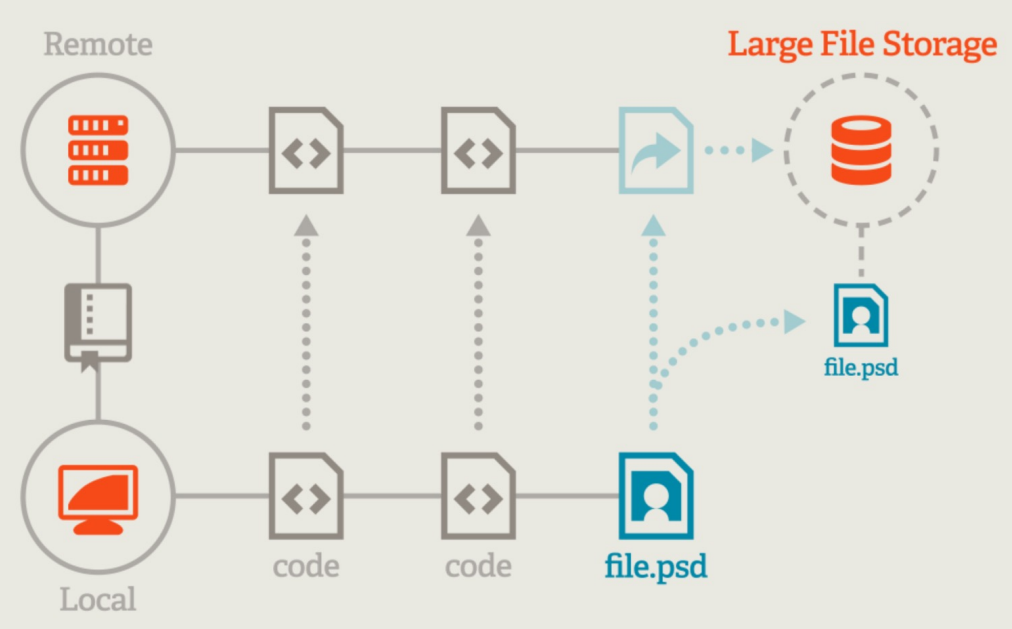
This document will help you experience how to use dragonfly with hugging face. During the downloading of datasets or models, the file size is large and there are many services downloading the files at the same time. The bandwidth of the storage will reach the limit and the download will be slow.

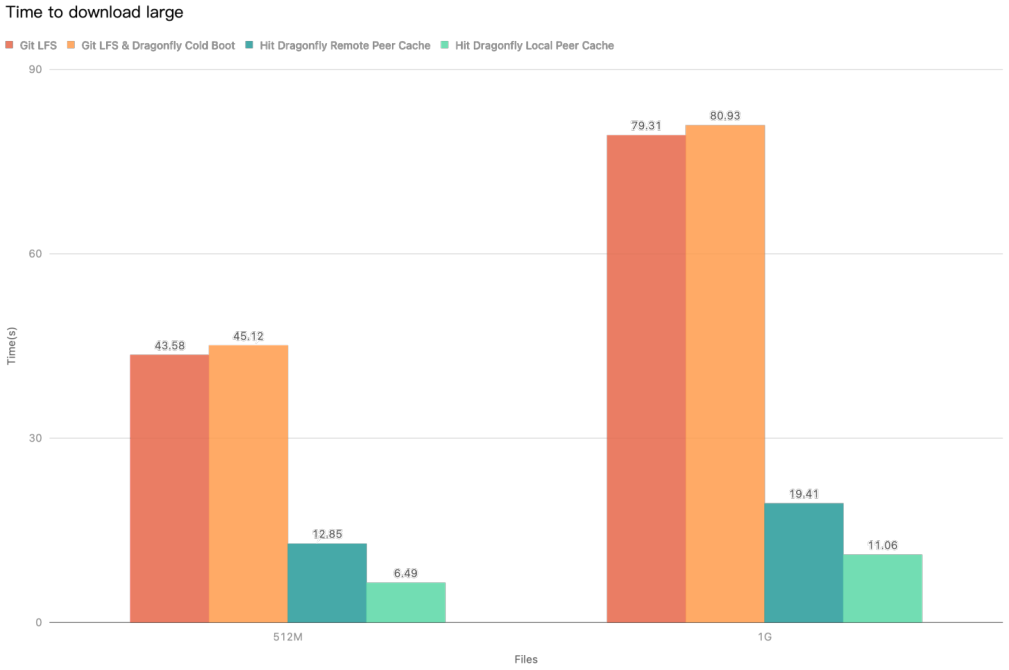




Refer: https://d7y.io/docs/next/operations/integrations/hugging-face/

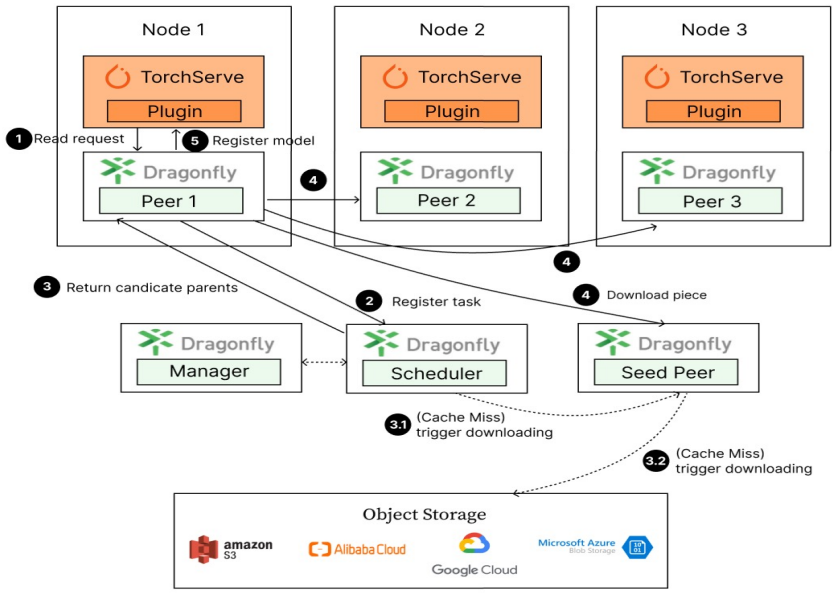
Dragonfly & Git LFS

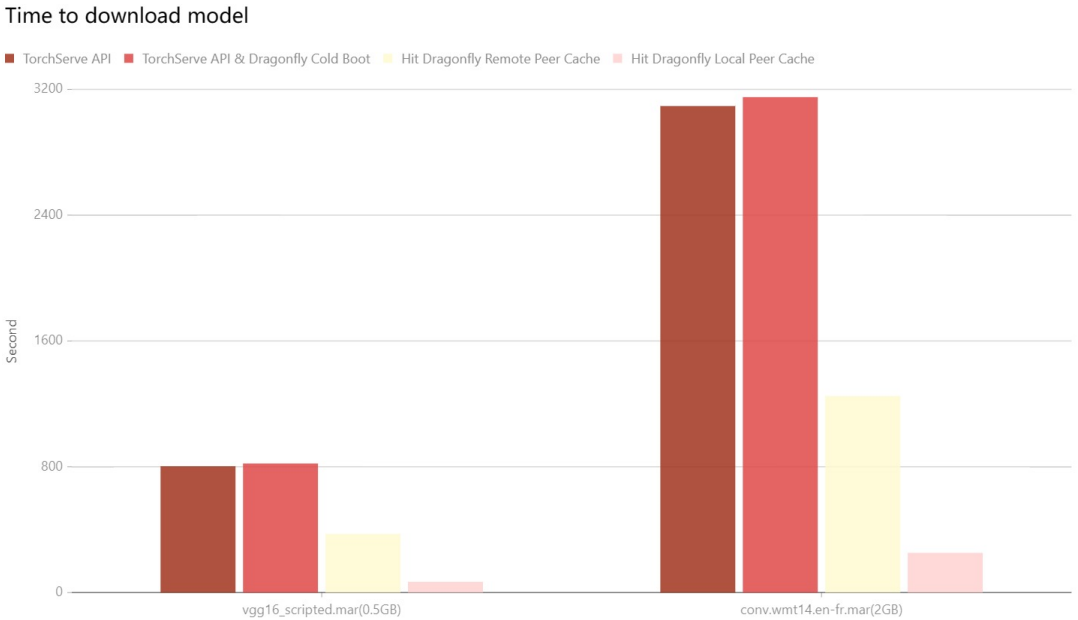




Refer: https://d7y.io/docs/next/operations/integrations/git-lfs/

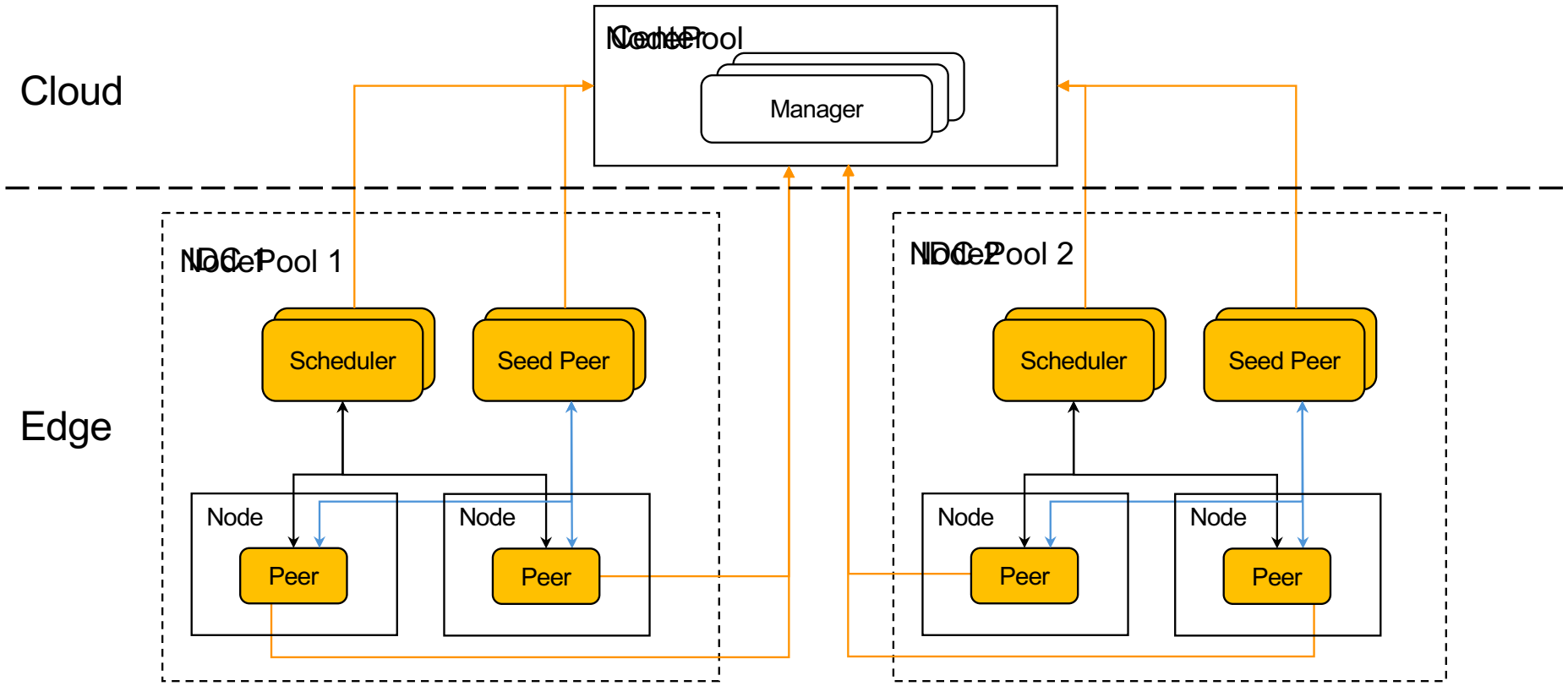
Dragonfly & PyTorch





Refer: https://d7y.io/docs/next/operations/integrations/torchserve/

Dragonfly & OpenYurt Practice




KubeCon
CloudNativeCon
THE LINUX FOUNDATION
OPEN SOURCE SUMMIT
AI_dev
Open Source GenAI & ML Summit
China 2024
THANKS

KubeCon | CloudNativeCon

THE LINUX FOUNDATION OPEN SOURCE SUMMIT

AI_dev Open Source GenAI & ML Summit

China 2024

Q&A

OpenYurt Community Group(2群)
410人

此二维码365天内有效（2025年08月21日前）

钉钉扫一扫群二维码，立即加入群聊

Dragonfly DevGroup 2
184人

扫一扫群二维码，立刻加入该群。


KubeCon
CloudNativeCon
THE LINUX FOUNDATION
OPEN SOURCE SUMMIT
AI_dev
Open Source GenAI & ML Summit
China 2024