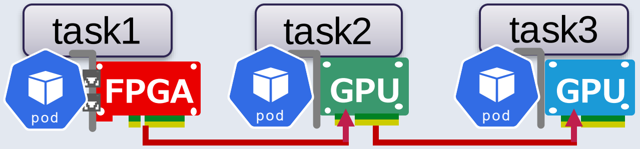
NTT
INTUIT

Effortlessly Build High-Performance AI/ML Pipelines With Accelerator Chaining and K8s Native Tech

KubeCon | CloudNativeCon
Europe 2025

Introduction

- Data Centers require higher performance while reducing power consumption
- In AI processing, accelerators are only used for certain tasks(e.g. Inference)
- It's important to leverage suitable accelerators for each AI processing task



Accelerator Chaining(*1)

- Preparing a processing infrastructure using accelerators is difficult,

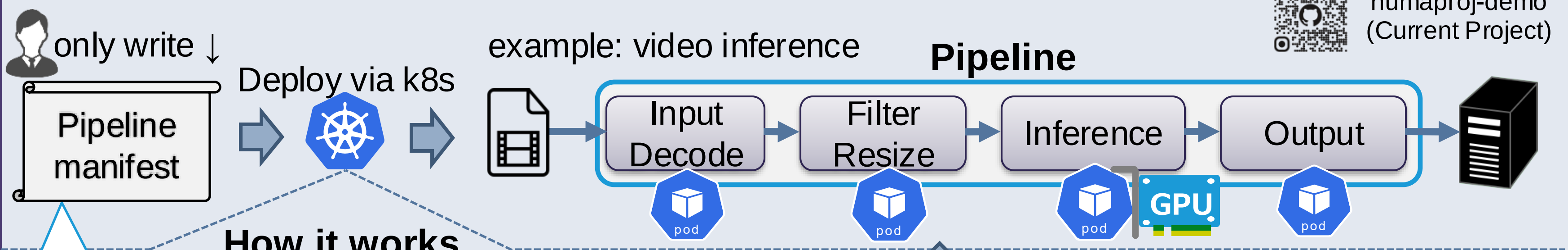
we provide a method to effortlessly build processing pipelines

PoC with K8s Native Tech (Numaflow and DRA)

(*1:)Open Kasugai (Previous Project)

numaproj-demo (Current Project)

This project is in progress to reimplement (*1:) the presentation at KubeCon EU 24 using OSS



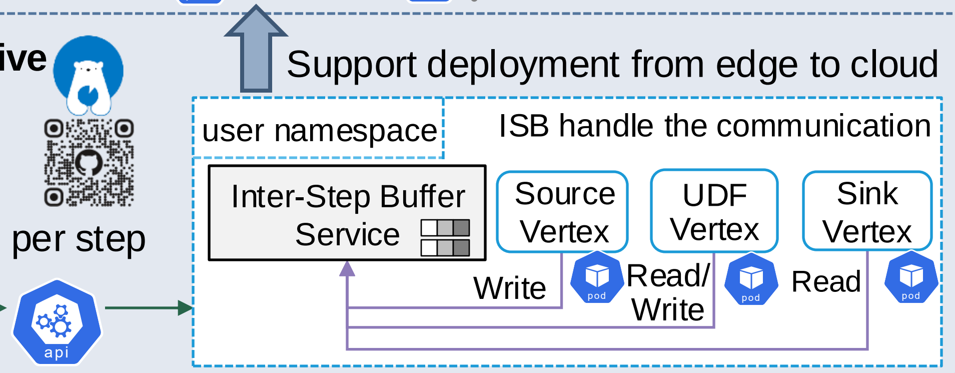
```
apiVersion: ...
kind: Pipeline
metadata:
    name: sample
spec:
    vertices:
      ...
    edges:
      ...
```

↑define each step and connection

How it works

Numaflow is **scalable K8s-native stream processing platform** that enables chaining multiple processing steps while allowing **Dynamic Resource Allocation** per step

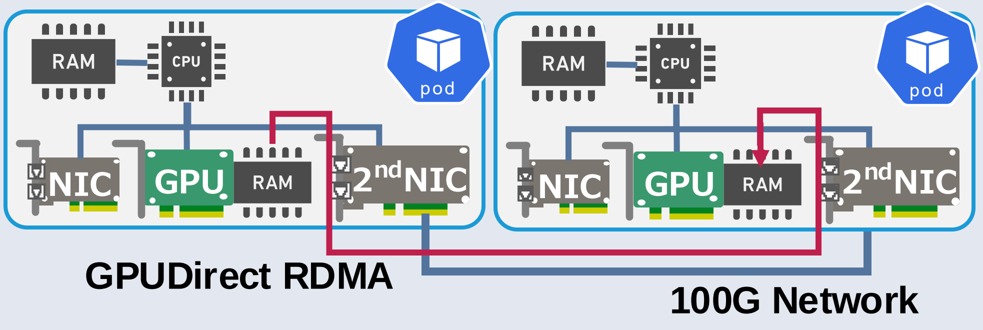
Pipeline Controller



api
Support deployment from edge to cloud
user namespace
ISB handle the communication
Inter-Step Buffer Service
Source Vertex
UDF Vertex
Sink Vertex
Write
Read/ Write
Read
pod

Future Work

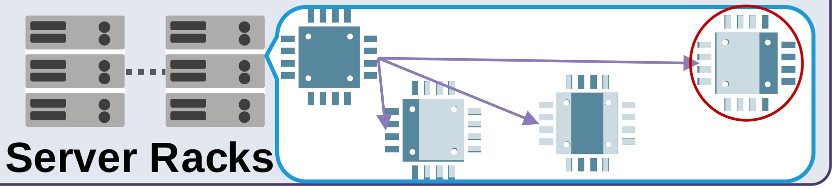
Direct Data Transfer for Accelerator Chaining



- Realize Data transfer bypassing CPU to utilize "GPUDirect RDMA" and "DRA"
- Our work towards enabling the assignment of the 2nd NIC, which is on the same PCIe bus as the GPU(*2), to the pod using DRA (= improve CNI driver for DRA)

New Scheduler Function

- As the inter-pod communication speed increases, the range of allocatable resources expands
- The scheduler allocates resources considering **resource efficiency** and App **latency constraint** while dividing resources **dynamically**



*2: The current GPUDirect only works when two devices share the same upstream PCI Express root complex