
AWS
re:Invent
DECEMBER 2 – 6, 2024 | LAS VEGAS, NV

KUB313

Architecture patterns for MLOps on Amazon EKS

Re Alvarez Parmar
(he/him)
Principal Solutions Architect
Amazon Web Services

Nirmal Mehta
(he/him)
Principal Solutions Architect
Amazon Web Services

Agenda

What is MLOps?

Why is it better with Amazon EKS?

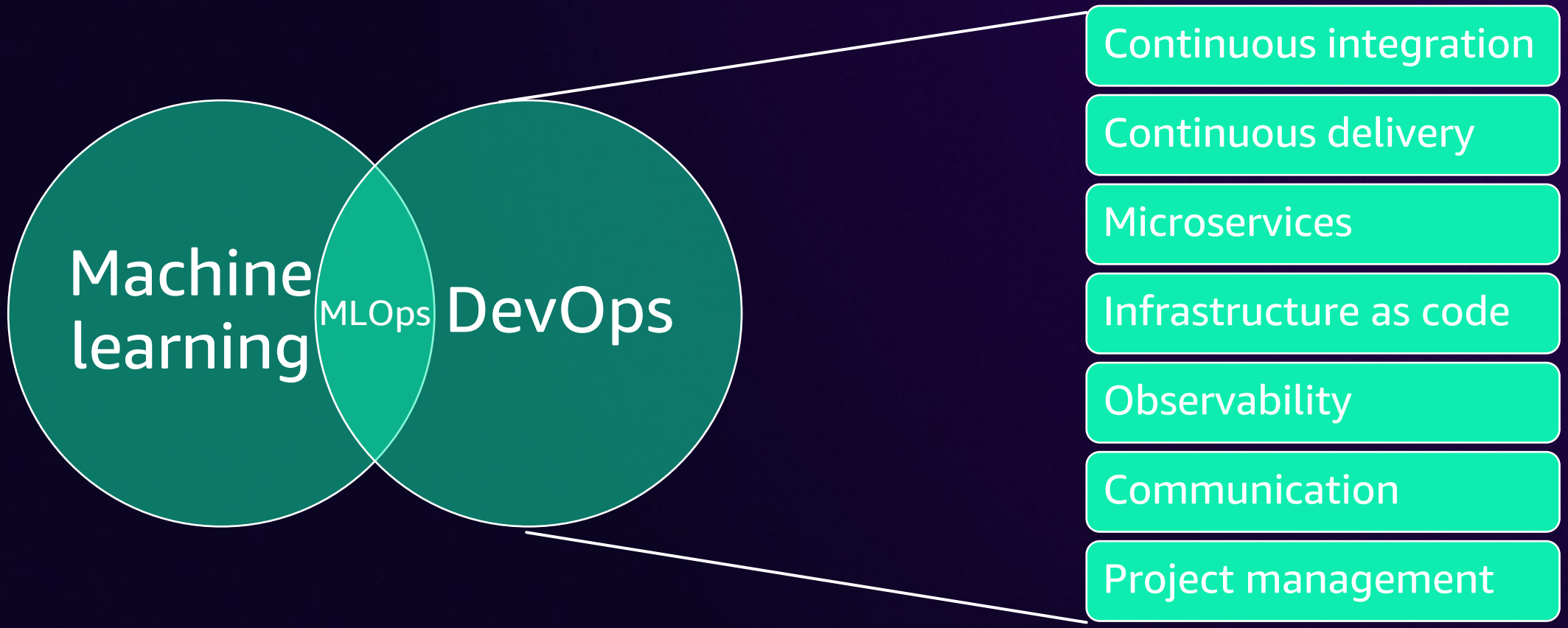
Distributed training patterns on Amazon EKS

What does ML let me do that was previously impossible?

How do I do it faster?

– Tech community

What is MLOps?



Containerization and ML

Easier dependency management

Controlled deployments

Simplified scaling

Packaging standardization

Kubernetes makes it faster to scale, deploy, and test models

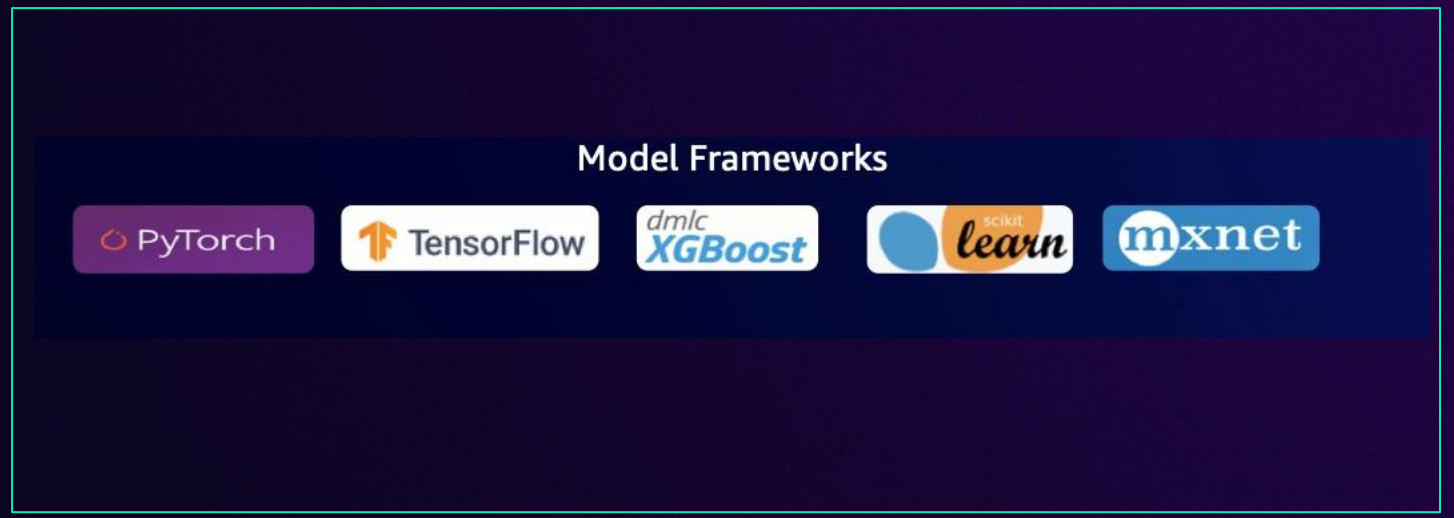
MLOps options in AWS

Fully managed
Easy to get started
Less operational overhead

Open source-based
Fine control over resources
Build your own stack

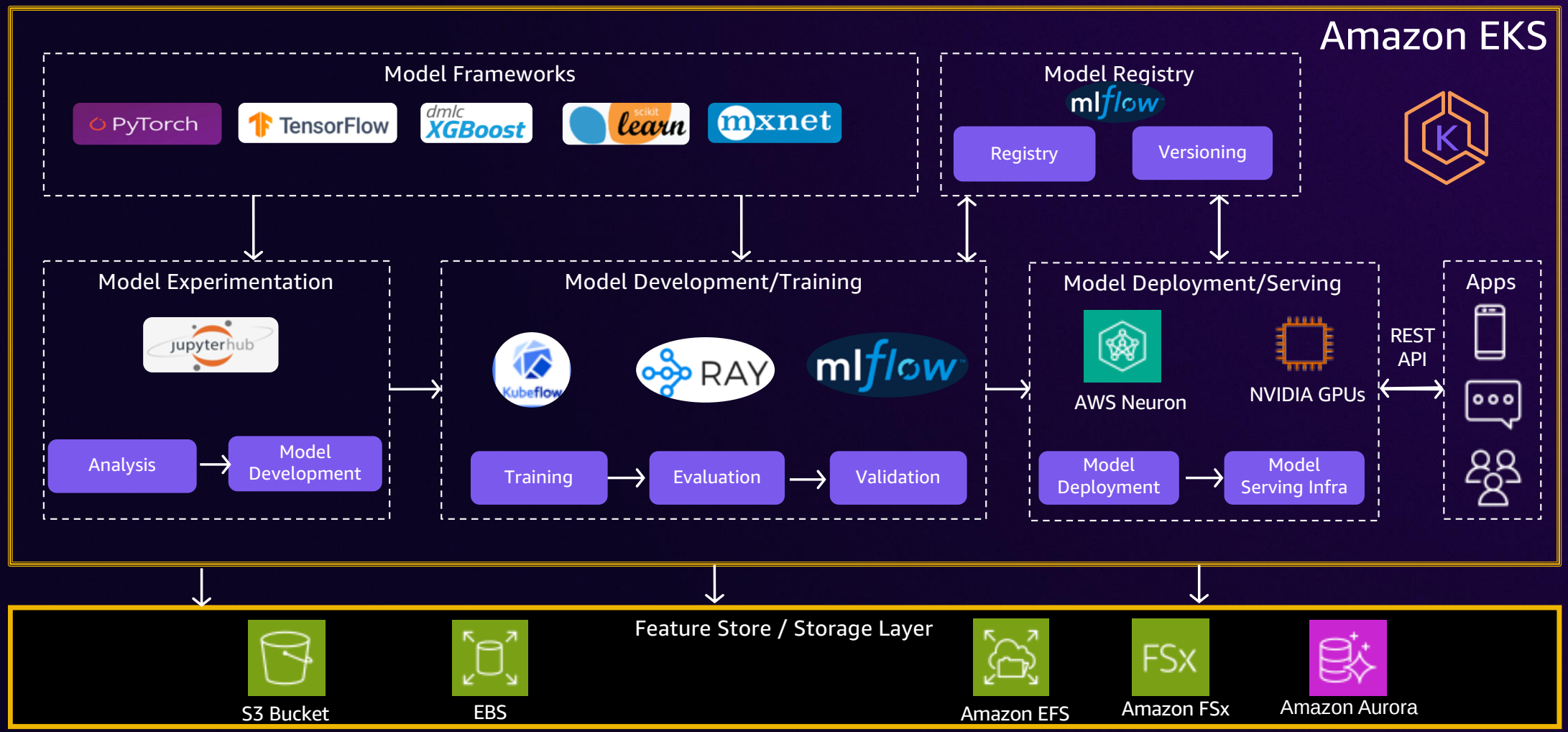
Enterprise MLOps

ML scientists

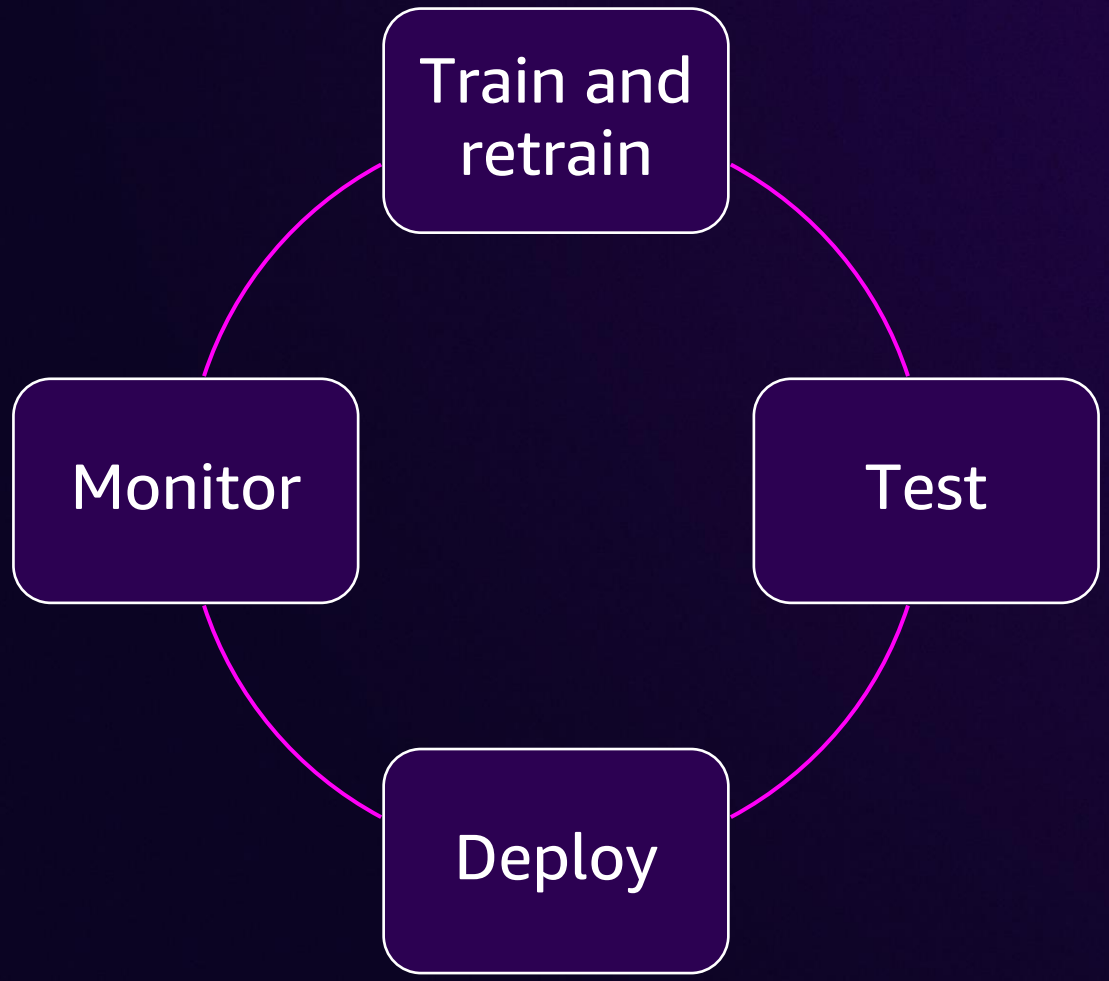


Platform engineers

OSS ecosystem on EKS



Machine learning (ML) feedback loop



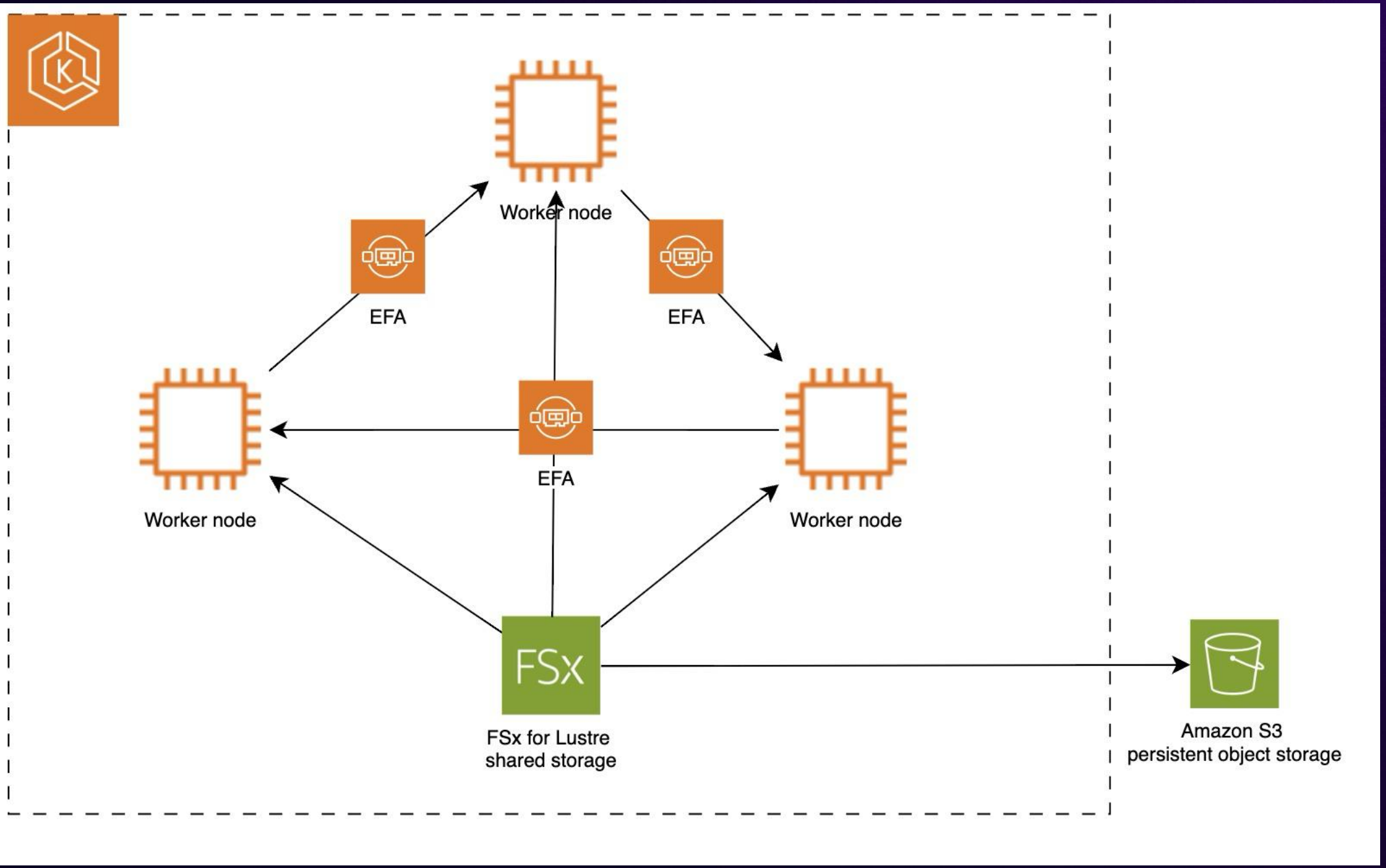
Training

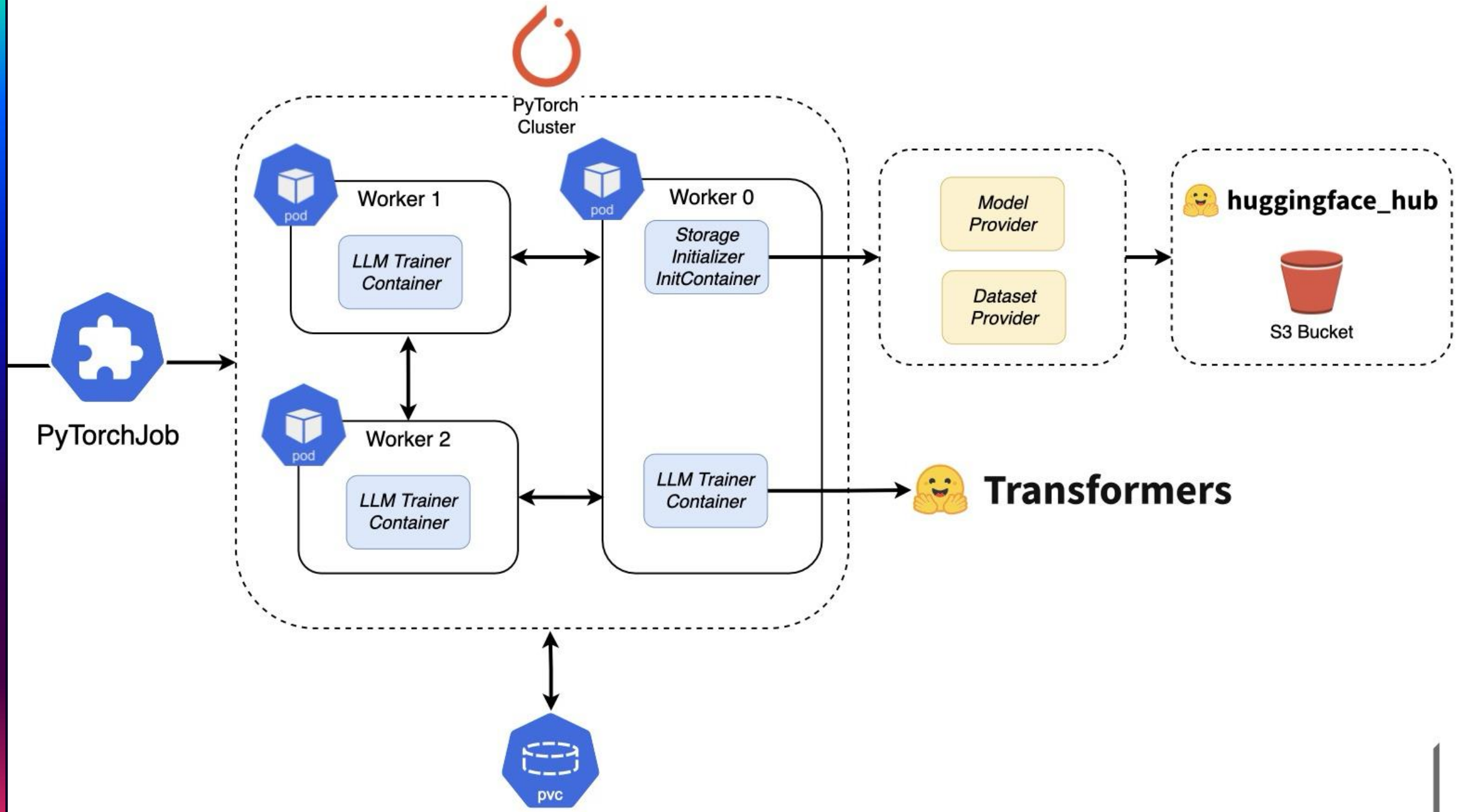
Distributed training architecture

Model parallelization and data parallelization

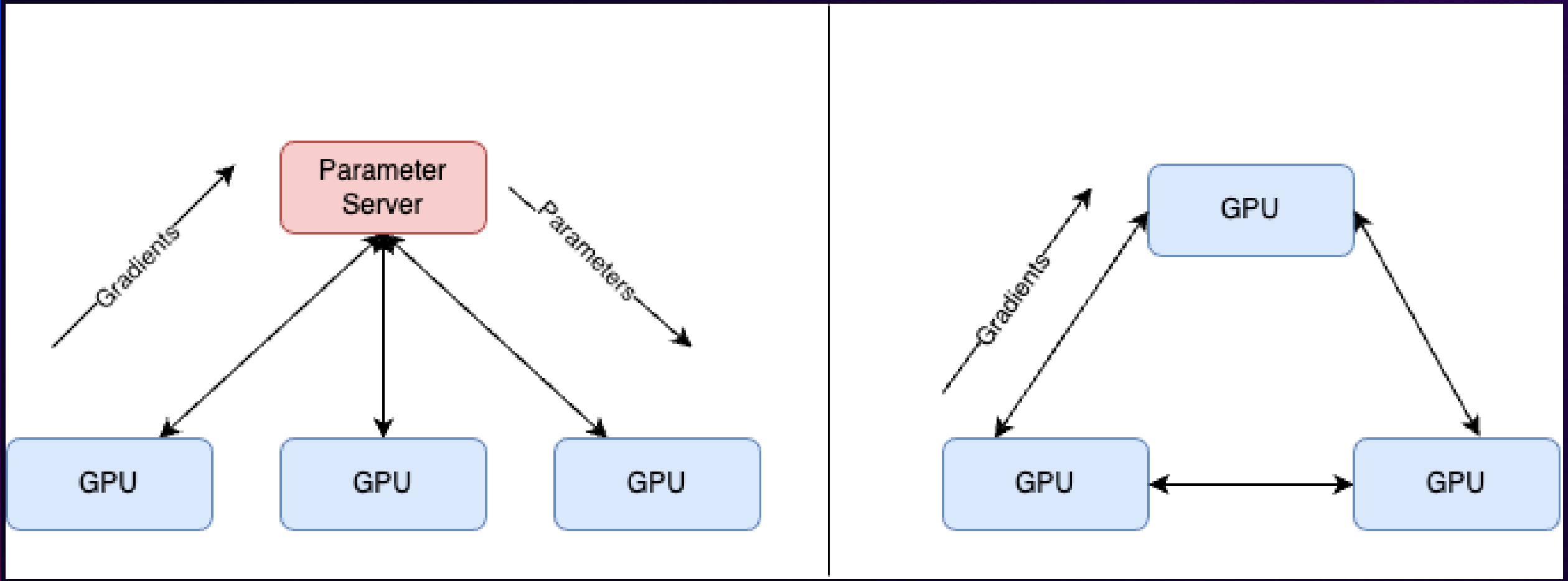
Parameter server versus collective communications

Distributed training with Ray

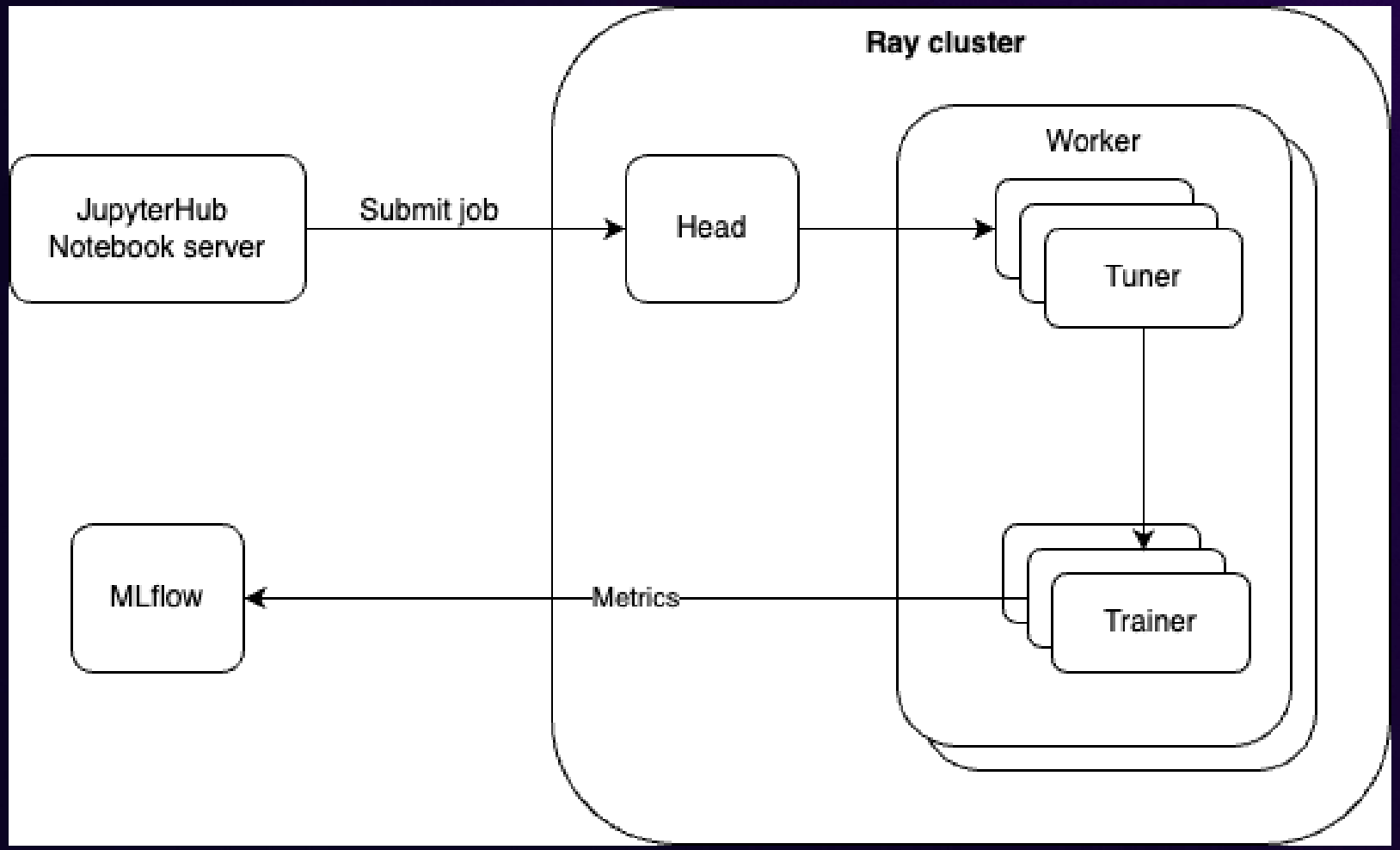



PyTorch
Cluster
Worker 1
pod
LLM Trainer
Container
Worker 0
pod
Storage
Initializer
InitContainer
PyTorchJob
Worker 2
pod
LLM Trainer
Container
LLM Trainer
Container
pvc
Model
Provider
Dataset
Provider
huggingface_hub
S3 Bucket
Transformers

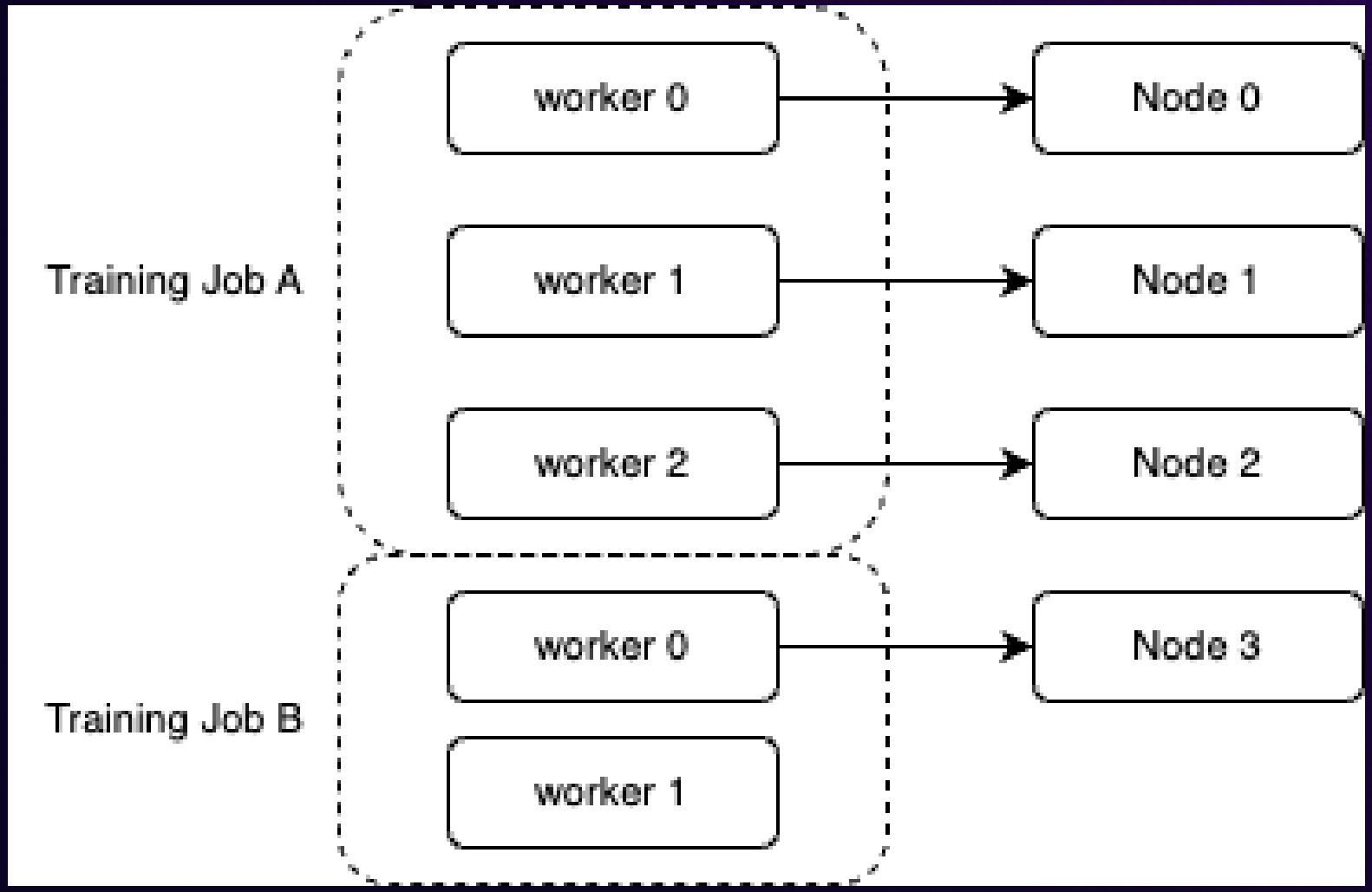
Parameter server vs. collective communications



Distributed HPO with Ray



Gang scheduling



Call for action

Data on EKS (github.com/awslabs/data-on-eks)

EKS Workshop (archive.eksworkshop.com/)

DoEKS Github repository (github.com/aws-samples/aws-do-eks)

Check out these other sessions

KUB 314 – High-Performance Generative AI on Amazon EKS

KUB 405 – Amazon EKS as data platform for analytics

KUB 316 – Deploy optimized inference pipelines on Amazon EKS

KUB 320 – Building modern data processing pipelines on Amazon EKS

KUB 403 – High-performance LLM inference scaling on Amazon EKS

KUB 401 – Workshop: Generative AI with Data on Amazon EKS (DoEKS)

Continue your Amazon EKS learning

Learn at your own pace

Take the **Amazon EKS Workshop** to expand your EKS skills

Increase your knowledge

Use our **Best Practices Guide** to build your Kubernetes knowledge

Earn Amazon EKS badge

Demonstrate your knowledge by achieving **digital badges**

https://github.com/aws-samples/reinvent24

Session resources

Thank you!

Please complete the session survey in the mobile app

Re Alvarez Parmar
reparmar@amazon.com

Nirmal Mehta
nkmehta@amazon.com