



KubeCon



CloudNativeCon

North America 2025

Simplifying Advanced AI Model Serving on Kubernetes Using Helm Charts

Ajay Vohra, Amazon
Caron Zhang, Apple

#KubeCon #CloudNativeCon



Agenda

- Kubernetes AI model serving stack
- Key challenges
- Proposed solution
- Two advanced examples:
 - Ray Serve + vLLM + Ray Cluster
 - LeaderWorkerSet + Triton Inference Server + vLLM + Ray Cluster + HPA

Kubernetes AI model serving stack



KubeCon



CloudNativeCon

North America 2025

Kubernetes AI model serving stack separation of concerns



AI/ML
Engineer /
Application
Developer

Inference servers, backends, models,
AI/ML frameworks

**Focus
of this talk**

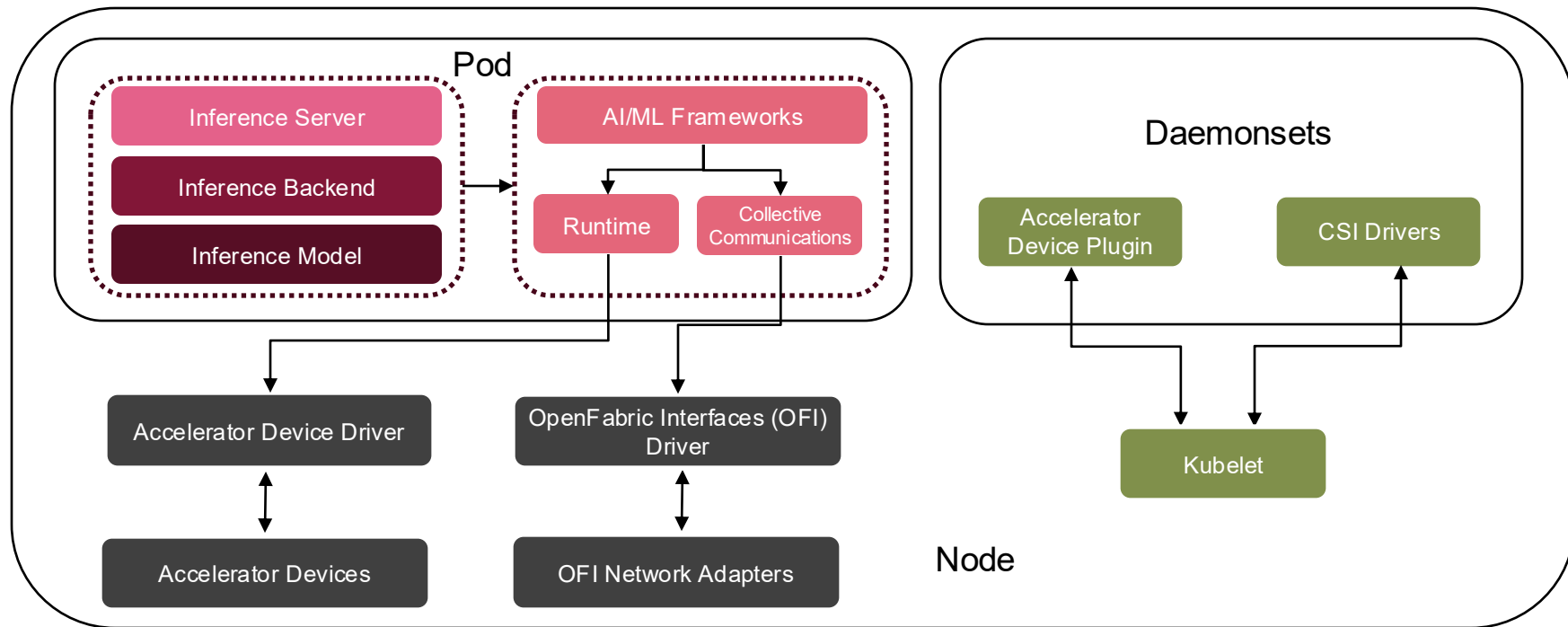
Platform
Engineering/
SRE

Kubernetes (k8s) operations and reliability

Infrastructure
Engineering

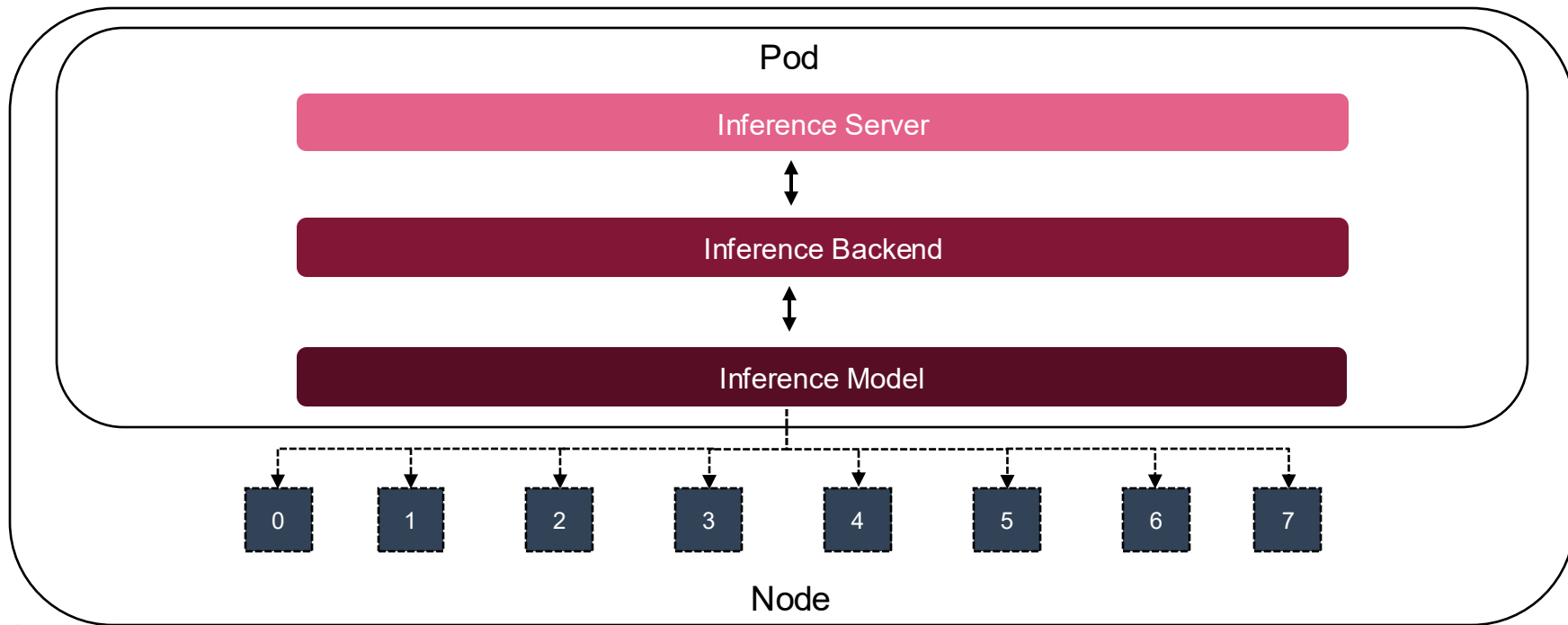
Machines, accelerators and networking

Kubernetes AI model serving generic stack



Single-node inference

Tensor parallel degree = 8



KubeCon

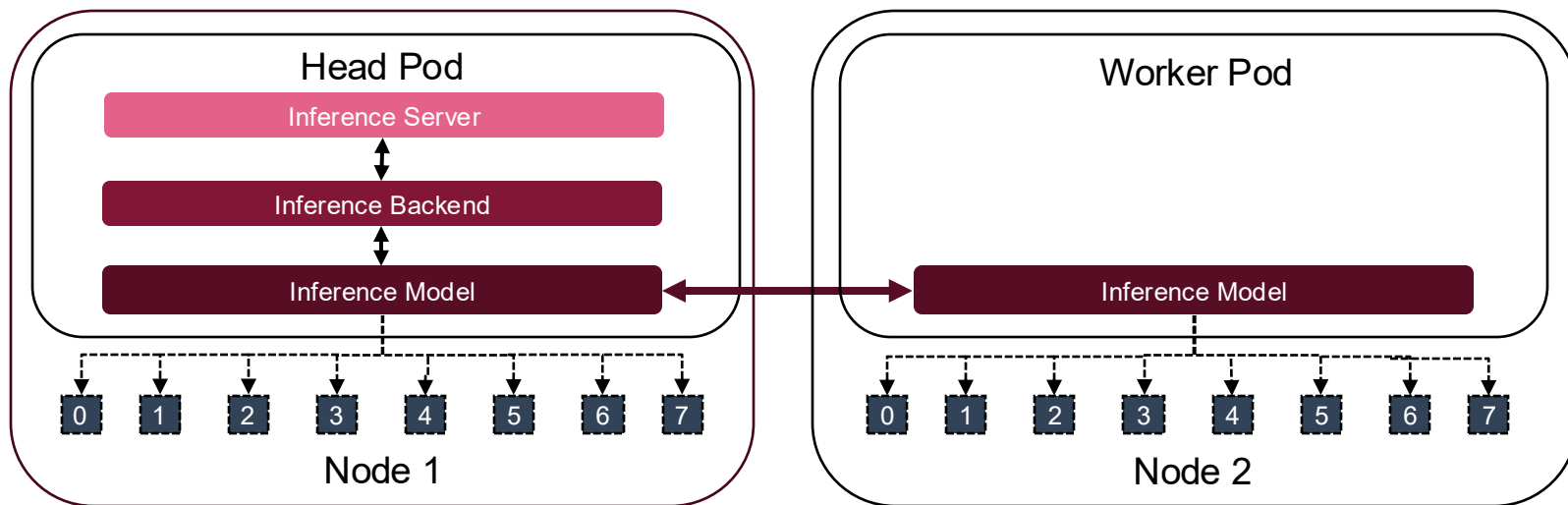


CloudNativeCon

North America 2025

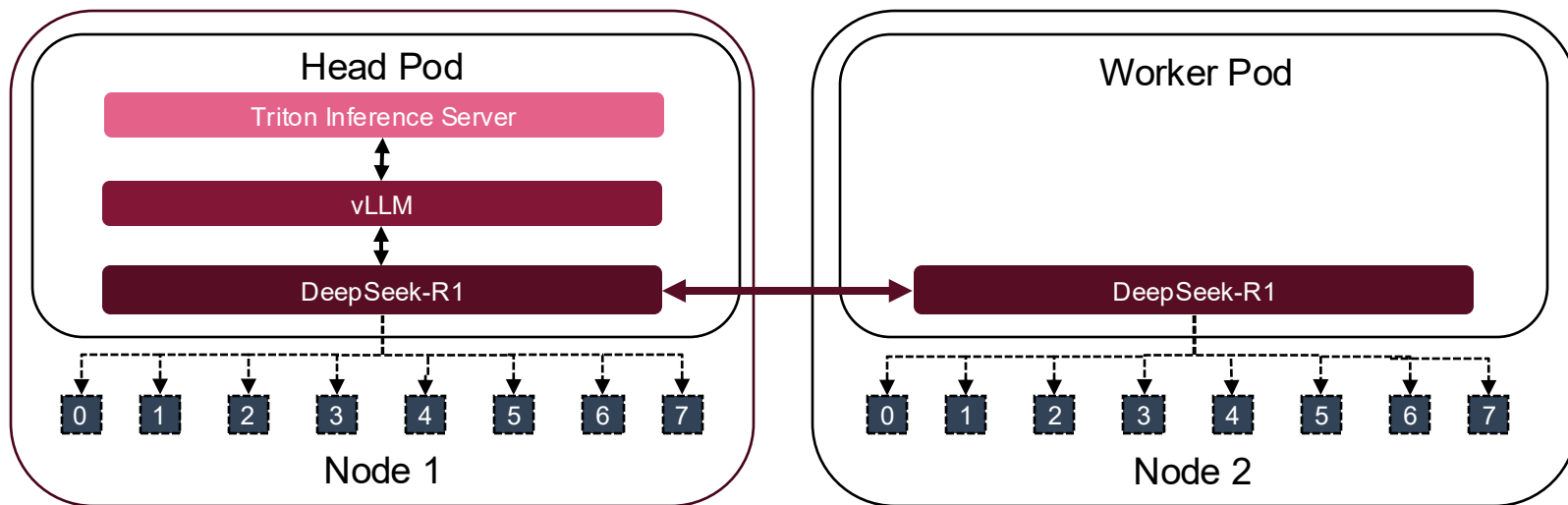
Multi-node inference

Tensor parallel degree = 8, Pipeline parallel degree = 2

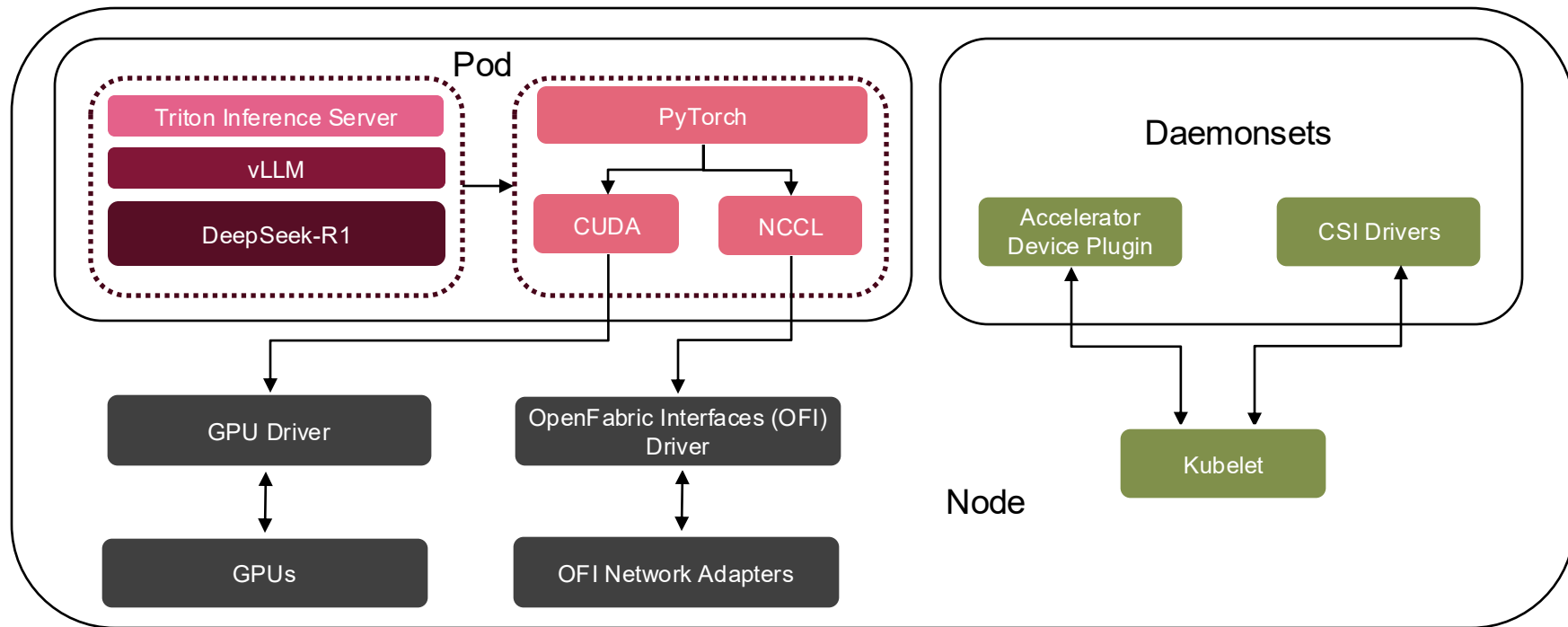


Multi-node inference of DeepSeek-R1

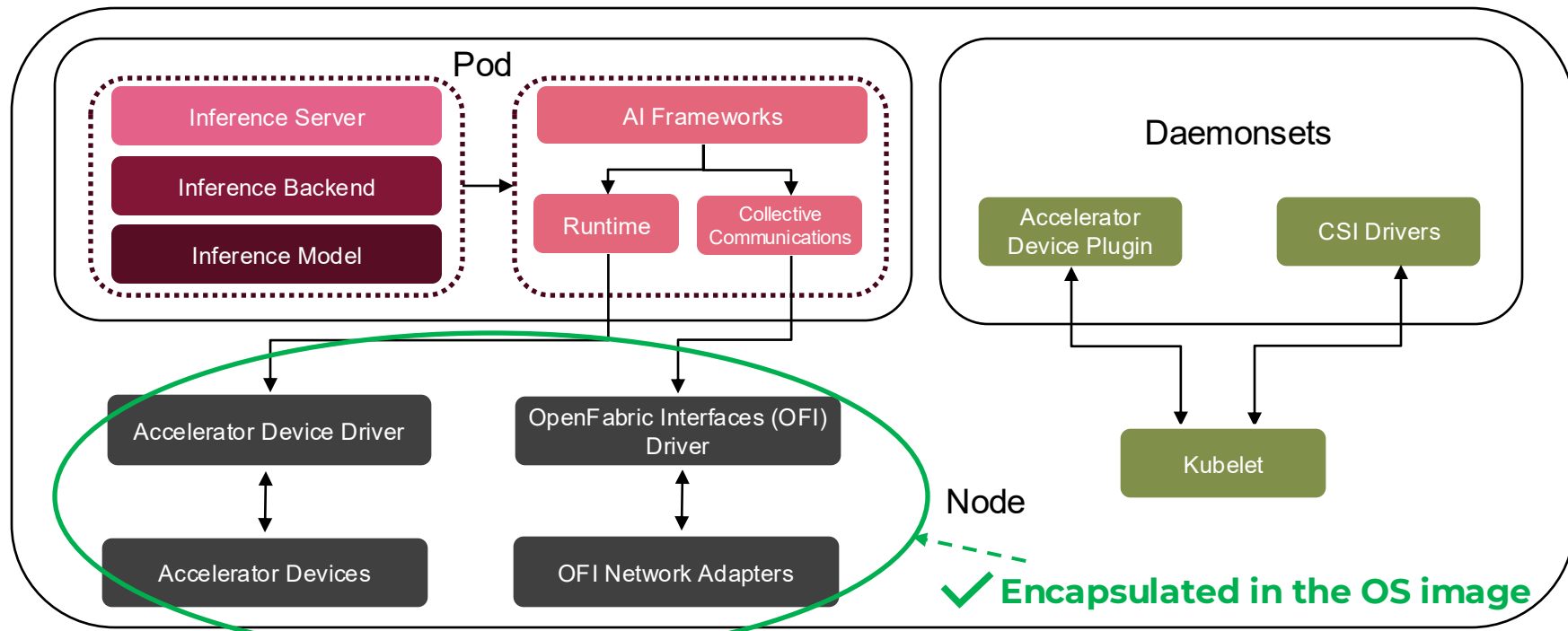
Tensor parallel degree = 8, Pipeline parallel degree = 2



Kubernetes DeepSeek-R1 model serving stack

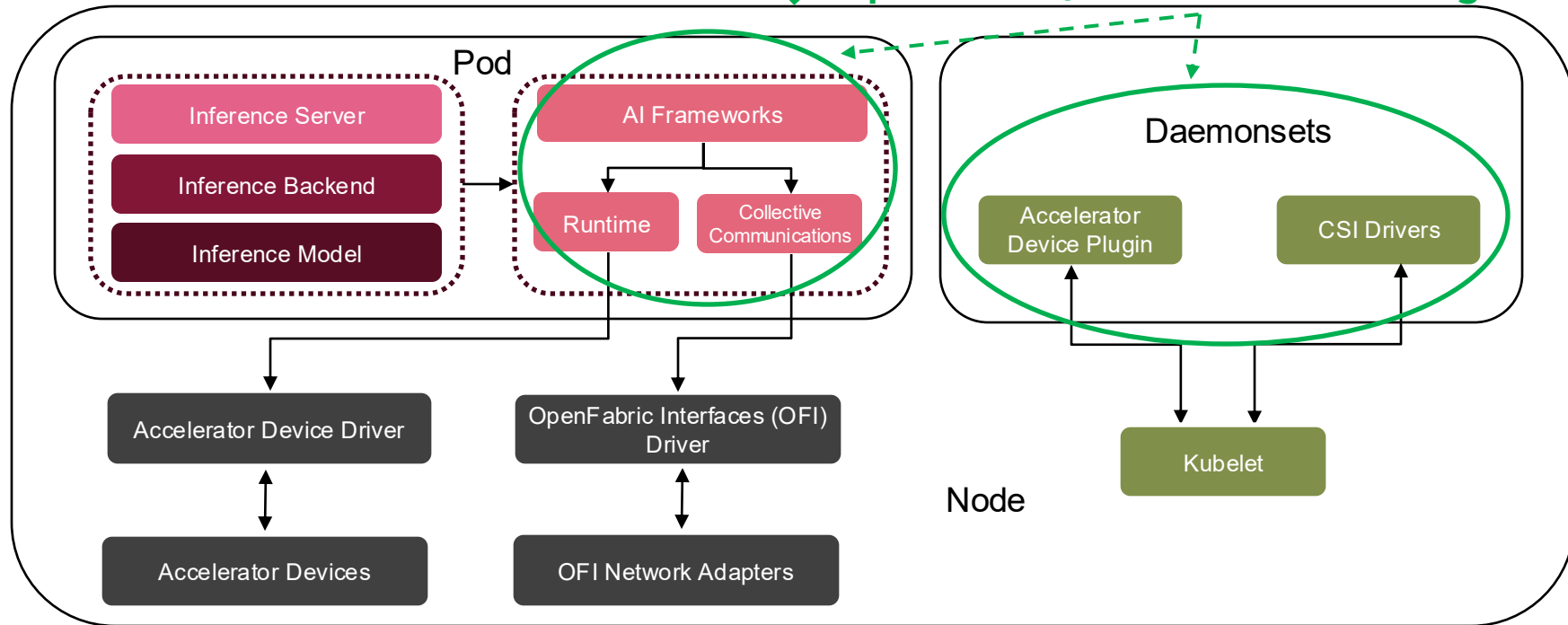


Kubernetes AI model serving generic stack



Kubernetes AI model serving generic stack

✓ Open-source/vendor container images



Key challenges



KubeCon

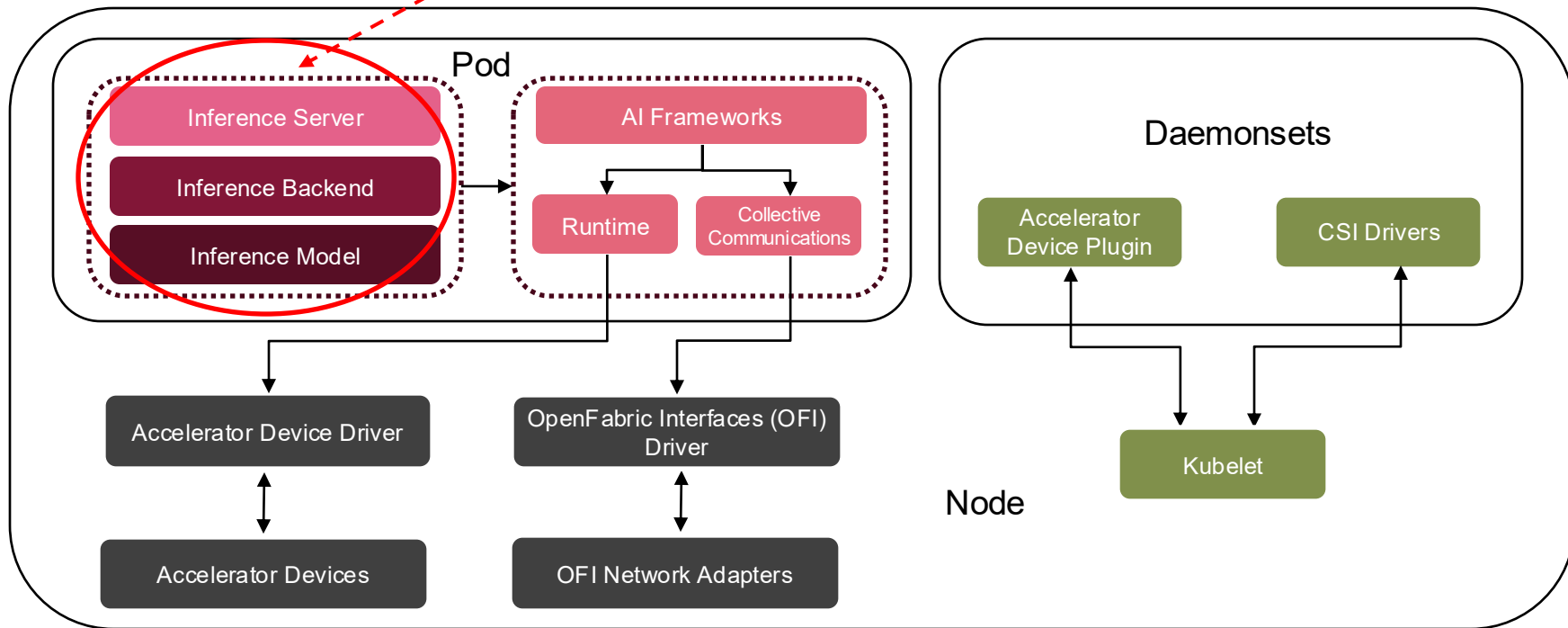


CloudNativeCon

North America 2025

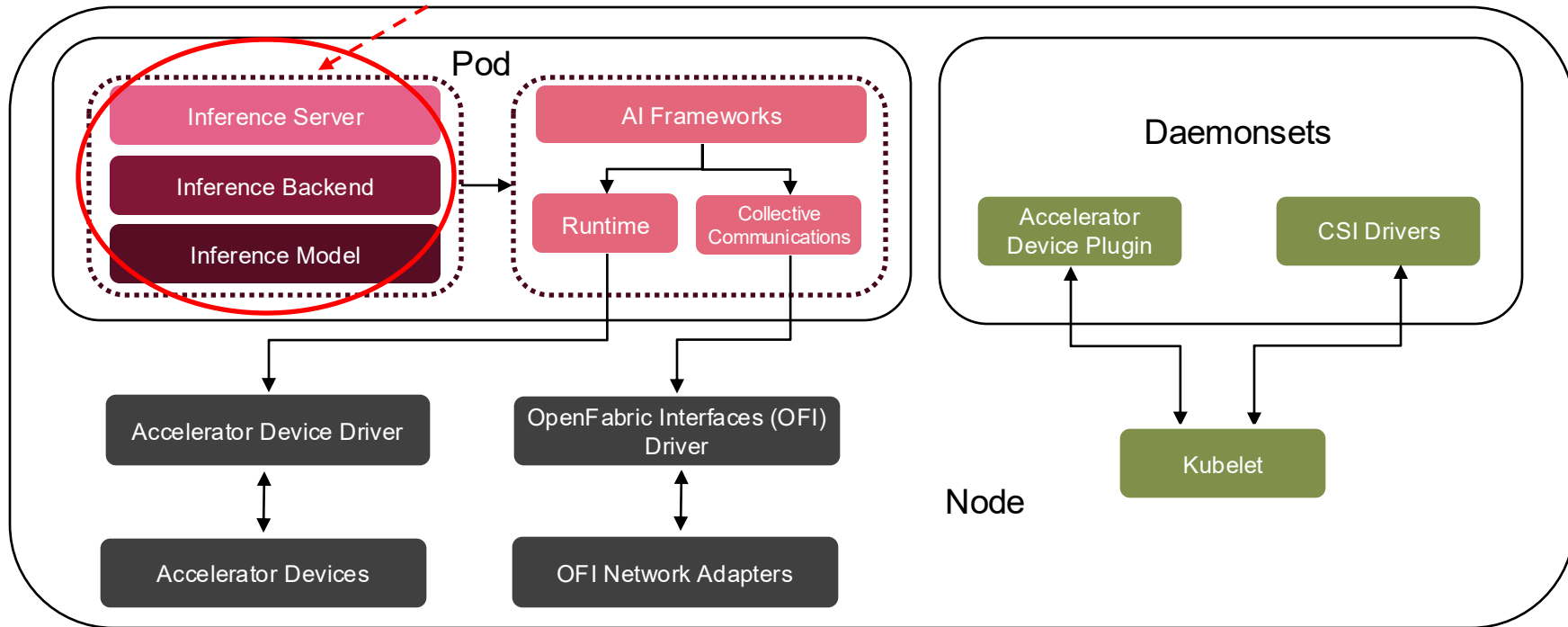
Kubernetes AI model serving generic stack

✗ Putting configs in a container image => slows down rapid development

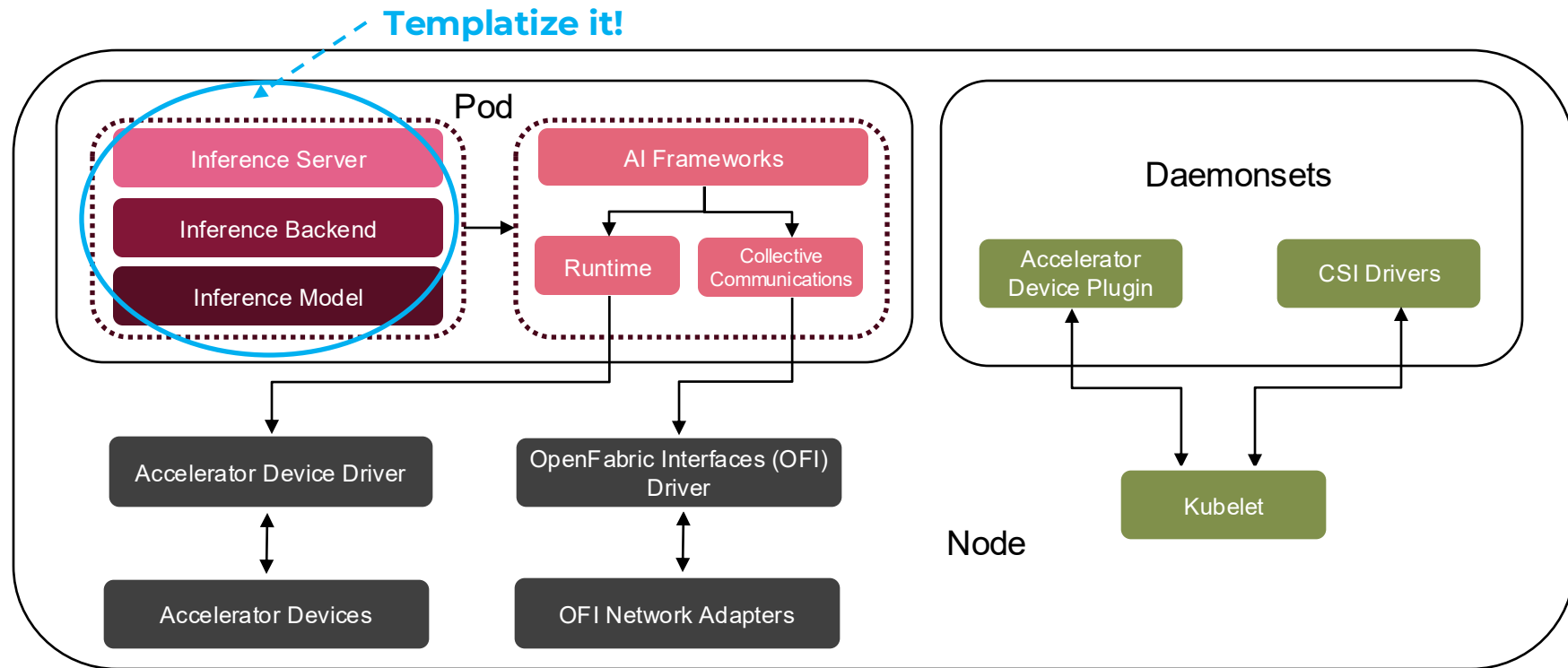


Kubernetes AI model serving generic stack

✗ Externalizing configs on a shared file-system => challenges reproducibility



Kubernetes AI model serving generic stack



Proposed solution



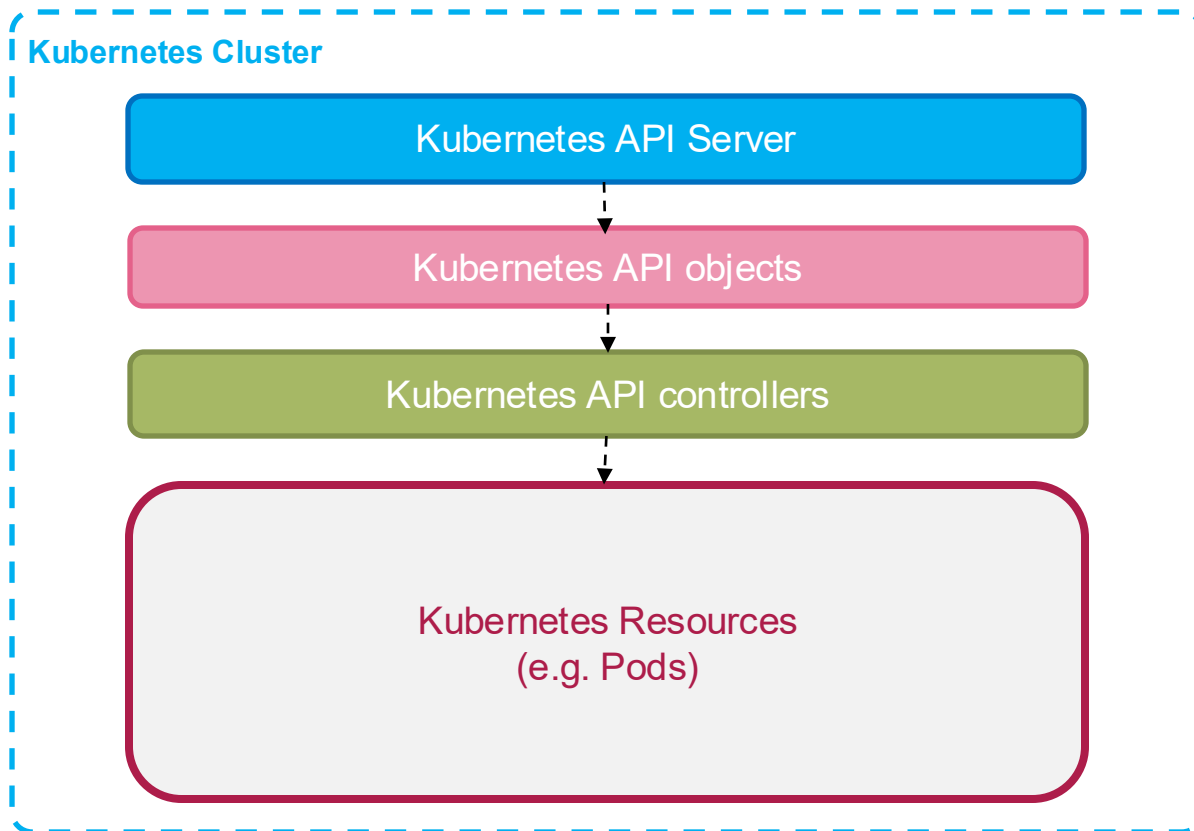
KubeCon



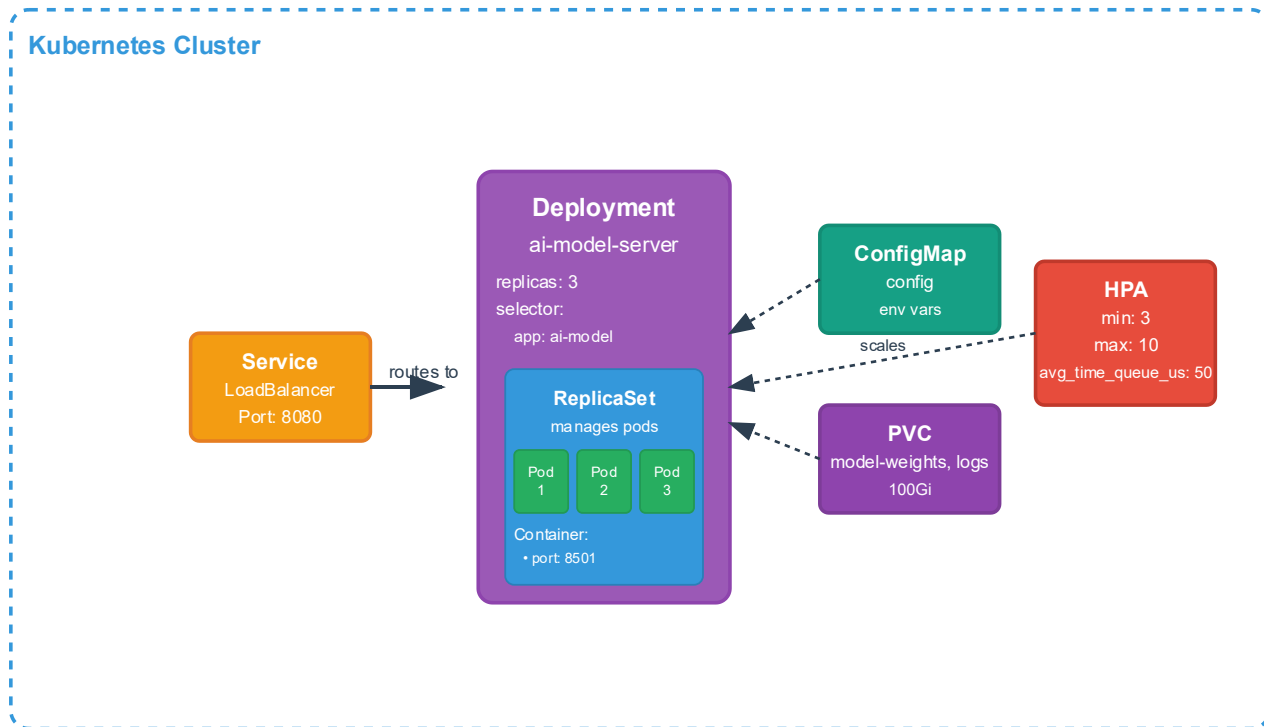
CloudNativeCon

North America 2025

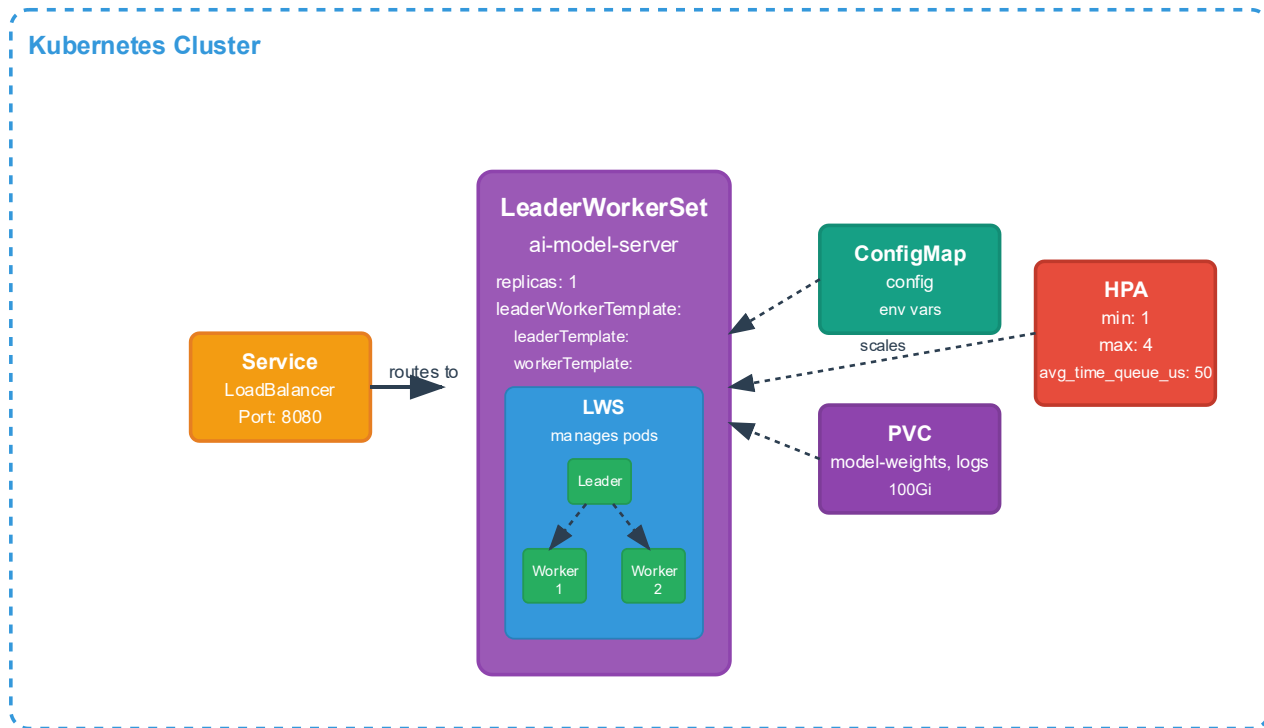
Kubernetes API recap



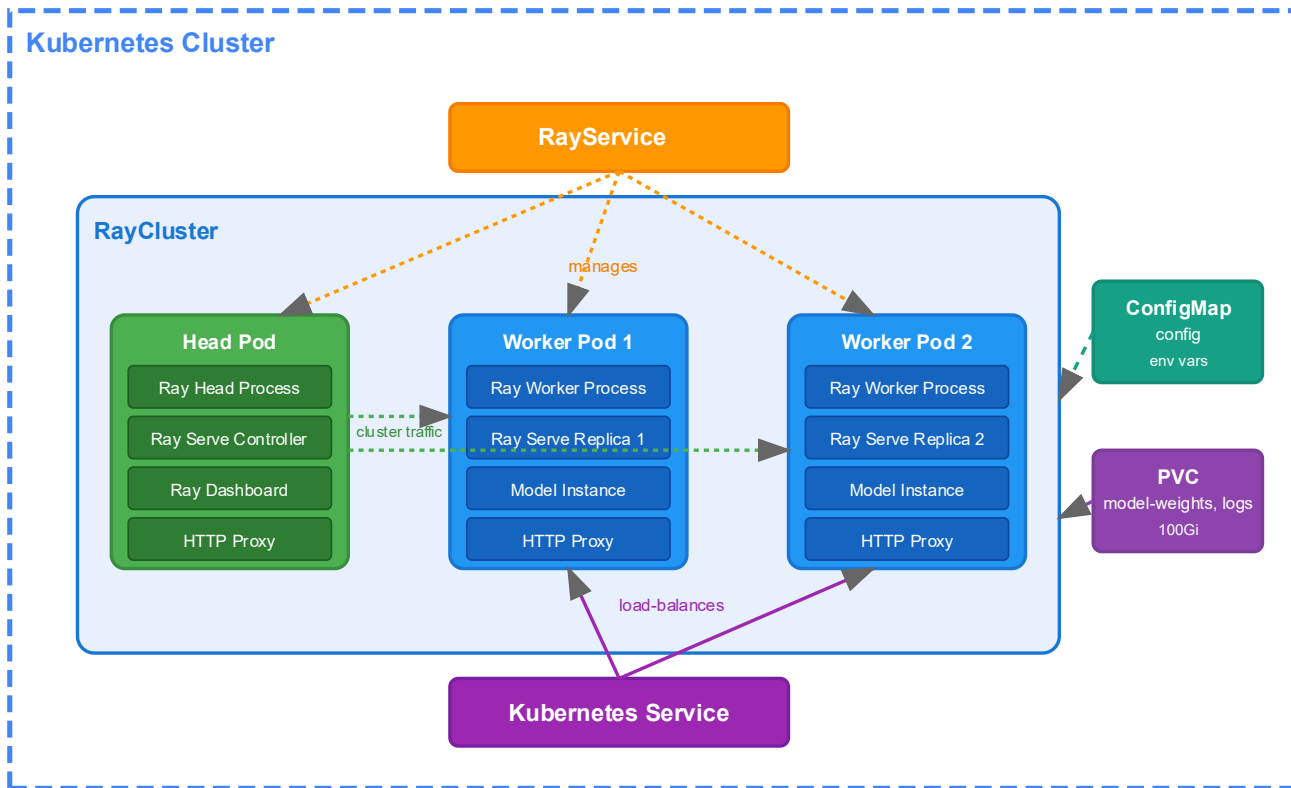
Kubernetes Deployment API for AI model serving



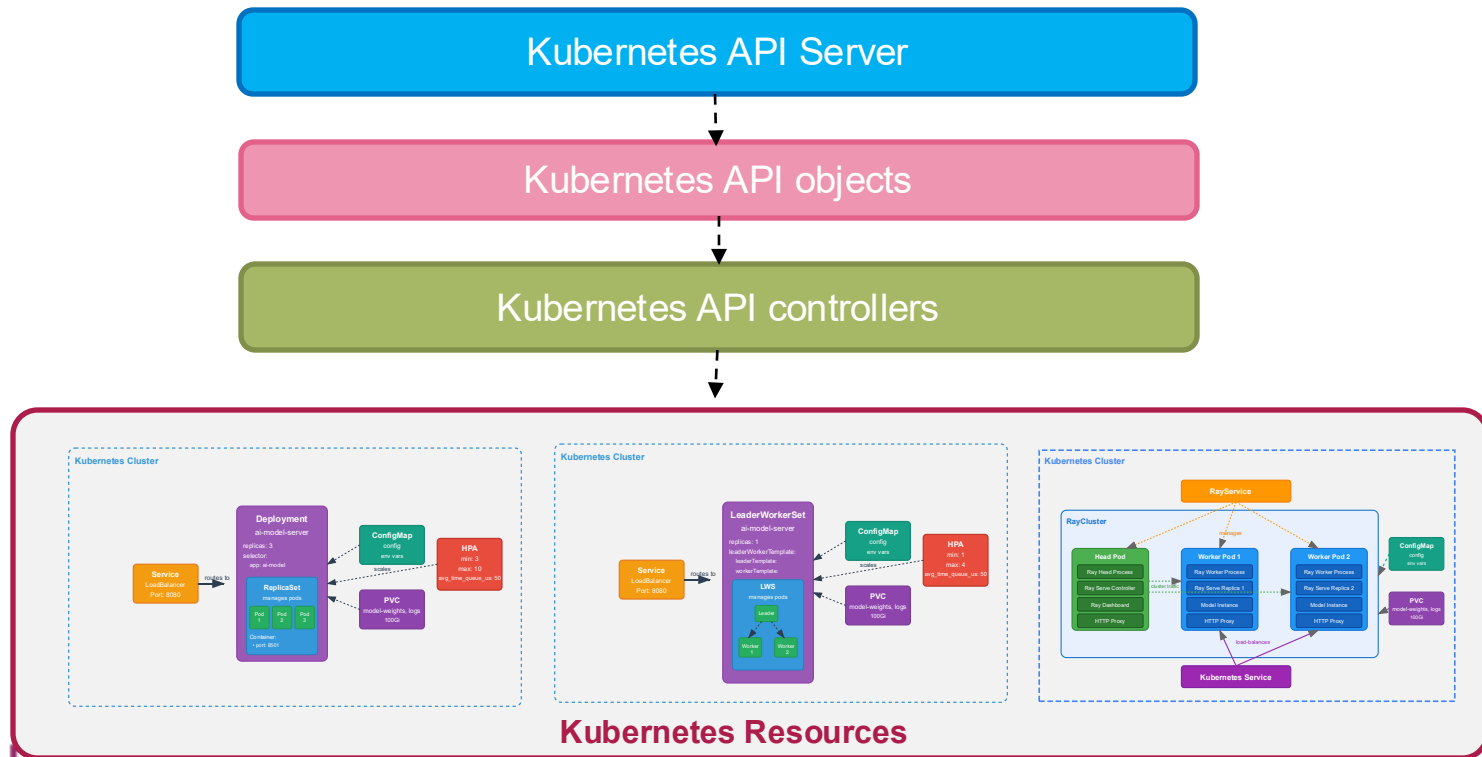
Kubernetes LeaderWorkerSet API for AI model serving



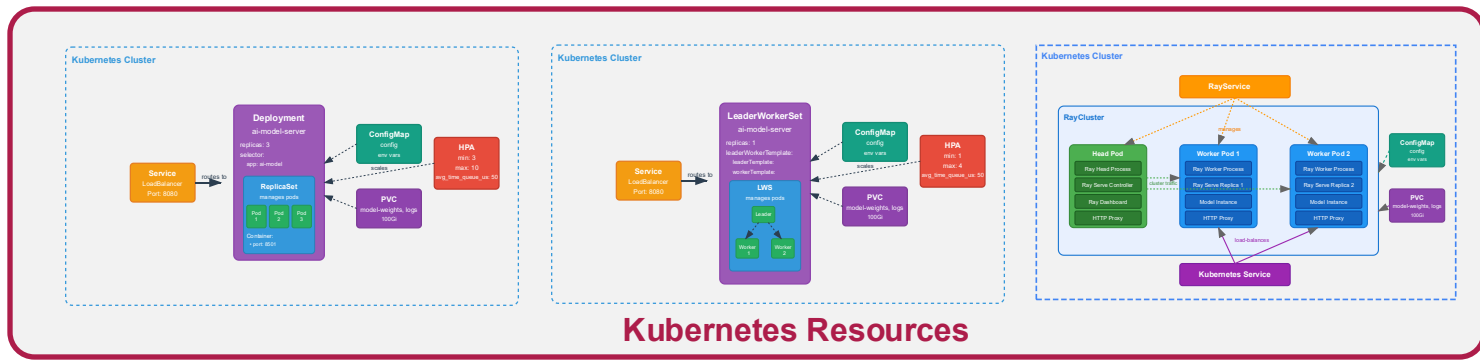
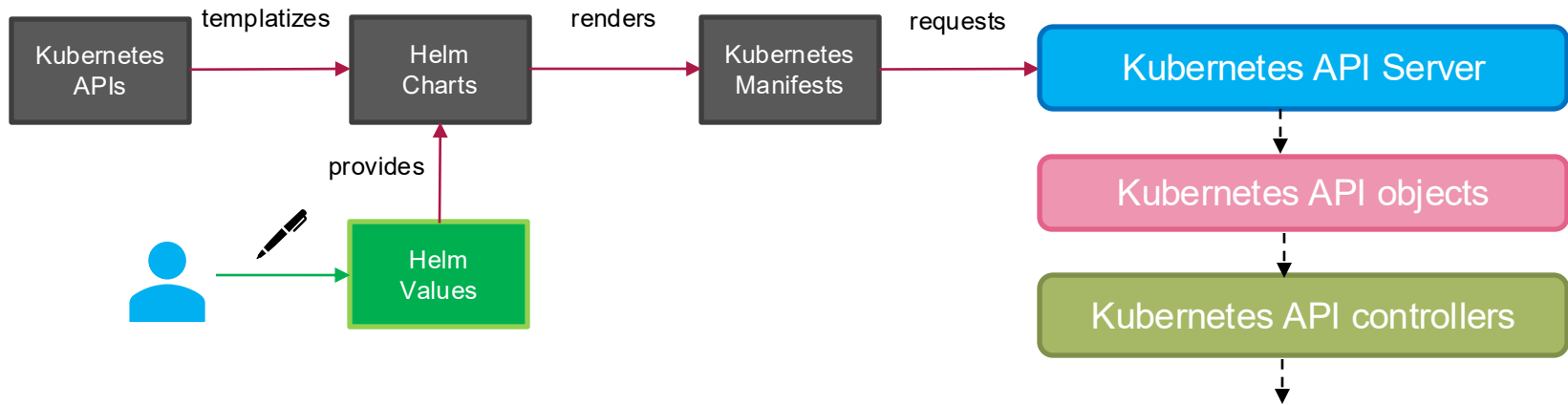
Kubernetes RayService API for AI model serving



Kubernetes APIs for AI model serving stacks



Templatize Kubernetes APIs in Helm charts



Two advanced examples: Multi-node LLM inference



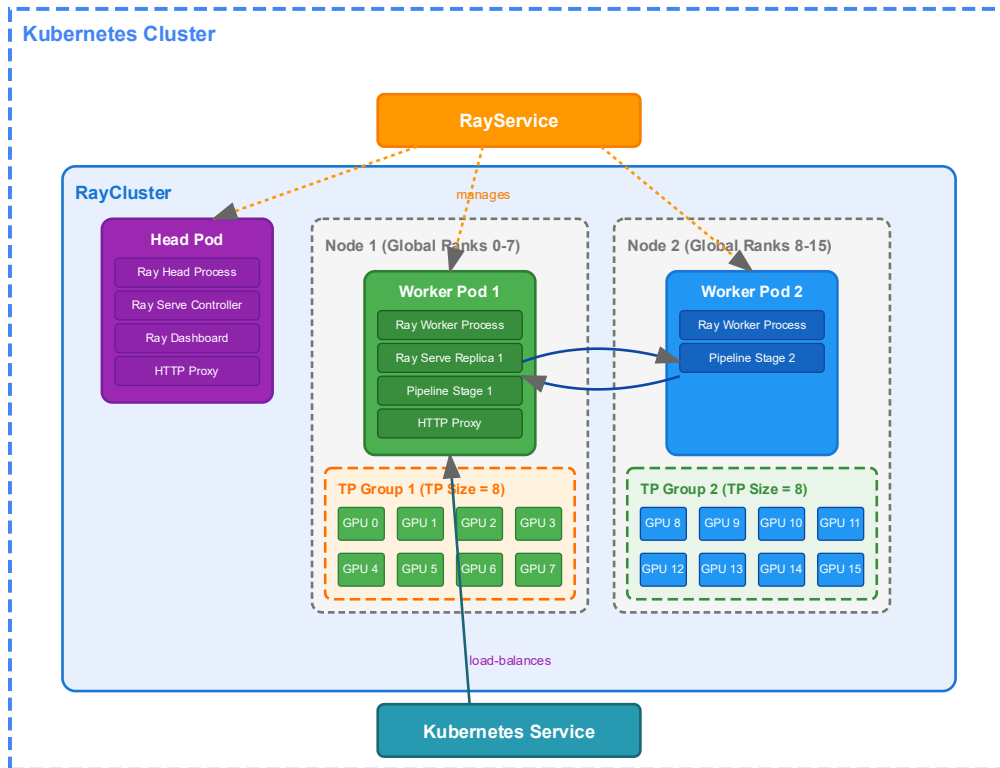
KubeCon



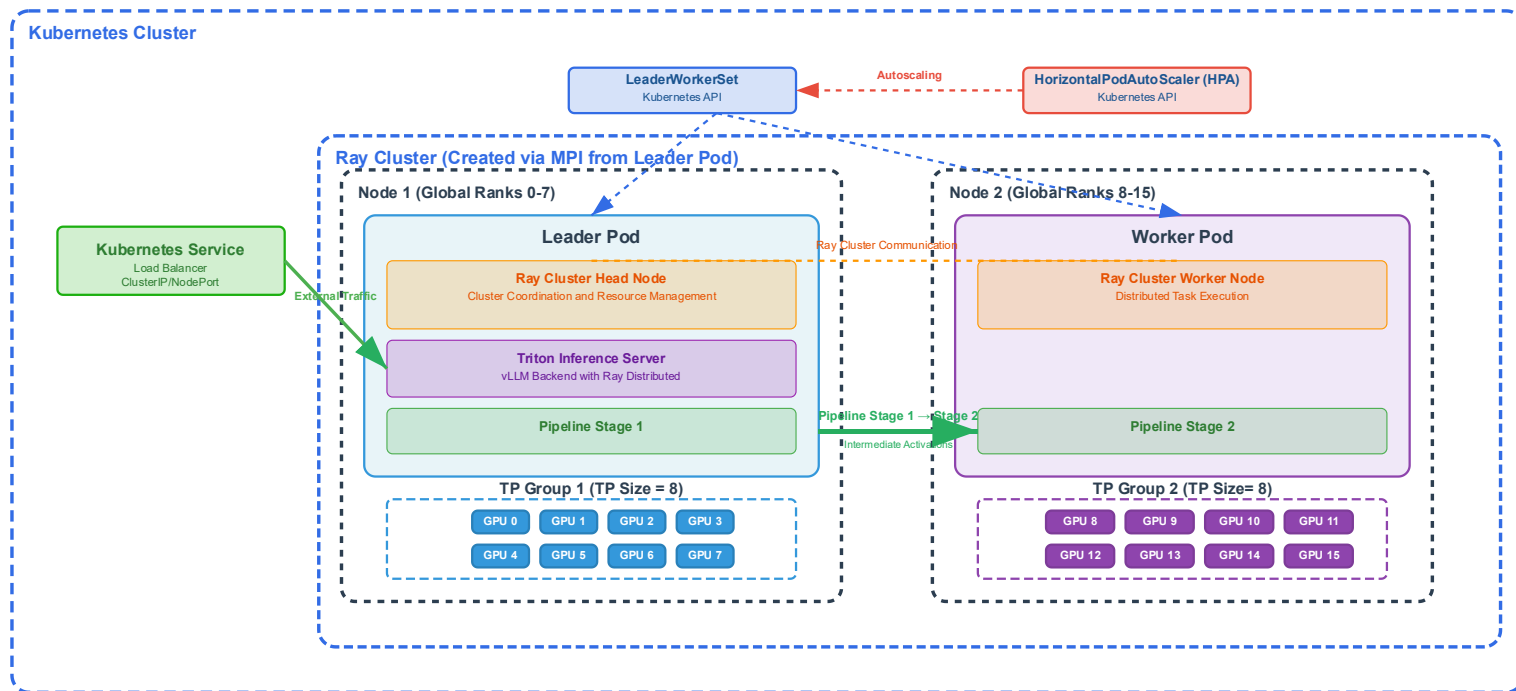
CloudNativeCon

North America 2025

Ray Serve + vLLM + Ray Cluster



LeaderWorkerSet + Triton + vLLM + Ray Cluster + HPA



GitHub Repository



<https://github.com/aws-samples/amazon-eks-machine-learning-with-terraform-and-kubeflow>



KubeCon



CloudNativeCon

— North America 2025 —

