KubeCon | CloudNativeCon

North America 2024

Production AI at Scale

Cloudera's Journey Building a Robust Inference Platform

Zoram Thanga

Peter Ableda

Speaker Intro

Zoram Thanga

Principal Engineer

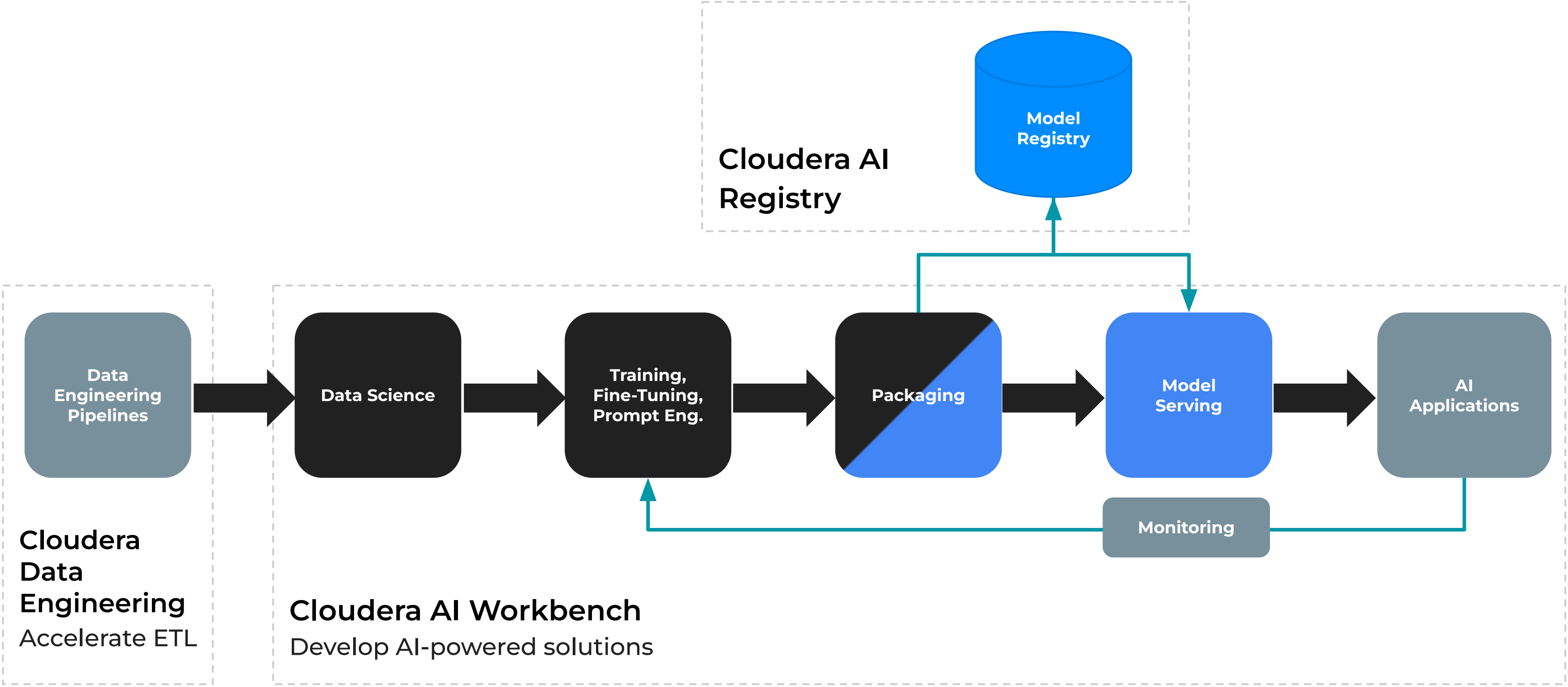
AI/ML Infrastructure

Peter Ableda

Director,
Product Management

Enterprise AI

Cloudera AI Pre-2024



Our Idea of a Robust Inference Platform

- Support for LLM and traditional model serving
- Use emerging Inference API standards (OpenAI, OIP)
- HA, fault tolerant, Zero-downtime upgrade
- High Scale, auto-scaling, scale-to-zero
- Security Controls, fine grained access control, Audit everything
- Monitoring for performance and drift
- Different operational, security, fault domain from dev
- Highly automatable (everything is an API)
- Highly customizable
- Run Anywhere (multi-cloud and on premises)
- Private deployment

Our Idea of a Robust Inference Platform

- Support for LLM and traditional model serving
- Use emerging Inference API standards (OpenAI, OIP)
- HA, fault tolerant, Zero-downtime upgrade
- High Scale, auto-scaling, scale-to-zero
- Security Controls, fine grained access control, Audit everything
- Monitoring for performance and drift
- Different operational, security, fault domain from dev
- Highly automatable (everything is an API)
- Highly customizable
- Run Anywhere (multi-cloud and on premises)
- Private deployment

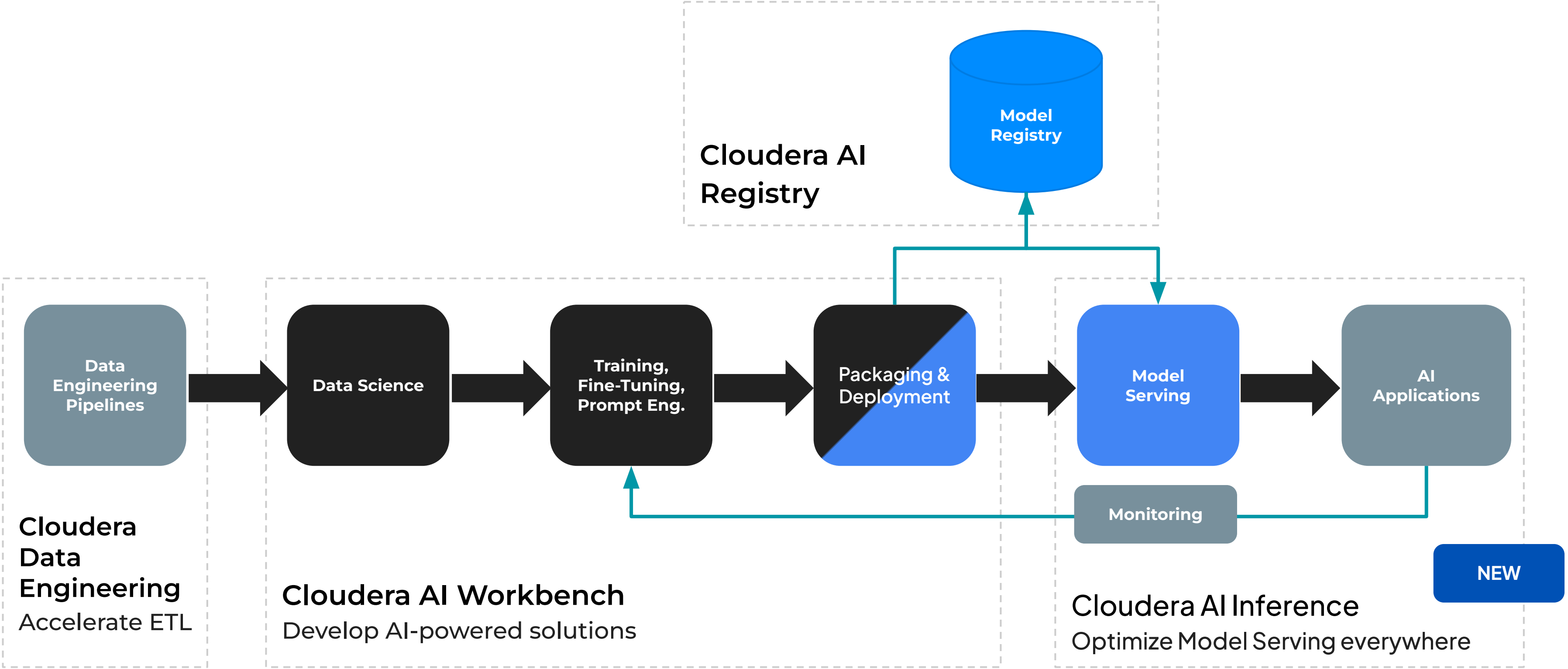
Our Idea of a Robust Inference Platform

- Support for LLM and traditional model serving
- Use emerging Inference API standards (OpenAI, OIP)
- HA, fault tolerant, Zero-downtime upgrade
- High Scale, auto-scaling, scale-to-zero
- Security Controls, fine grained access control, Audit everything
- Monitoring for performance and drift
- Different operational, security, fault domain from dev
- Highly automatable (everything is an API)
- Highly customizable
- Run Anywhere (multi-cloud and on premises)
- Private deployment

Our Idea of a Robust Inference Platform

- Support for LLM and traditional model serving
- Use emerging Inference API standards (OpenAI, OIP)
- HA, fault tolerant, Zero-downtime upgrade
- High Scale, auto-scaling, scale-to-zero
- Security Controls, fine grained access control, Audit everything
- Monitoring for performance and drift
- Different operational, security, fault domain from dev
- Highly automatable (everything is an API)
- Highly customizable
- Run Anywhere (multi-cloud and on premises)
- Private deployment

What We Need



Cloudera AI Registry

Cloudera Data Engineering
Accelerate ETL

Cloudera AI Workbench
Develop AI-powered solutions

Cloudera AI Inference
Optimize Model Serving everywhere

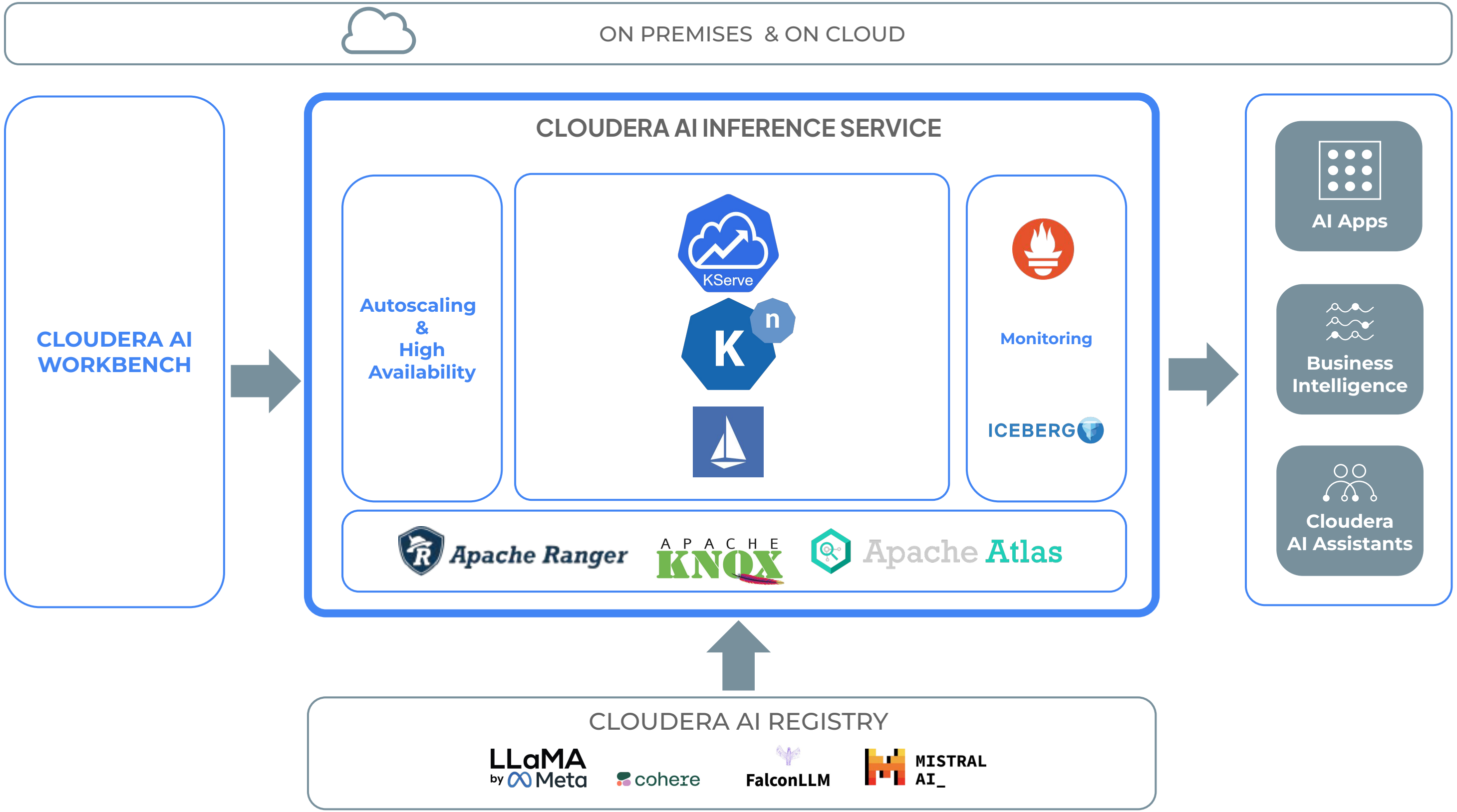
So...What Do We Do?

- Build everything from scratch?
  - Sure, we could with unlimited time and resources...
- Adopt an open source project
  - Build enterprise security + governance around it.
- But which one?
  - Seldon Core ❌
  - Yatai + Bento ML ❌
  - Ray Serve ❌
  - KServe ✅

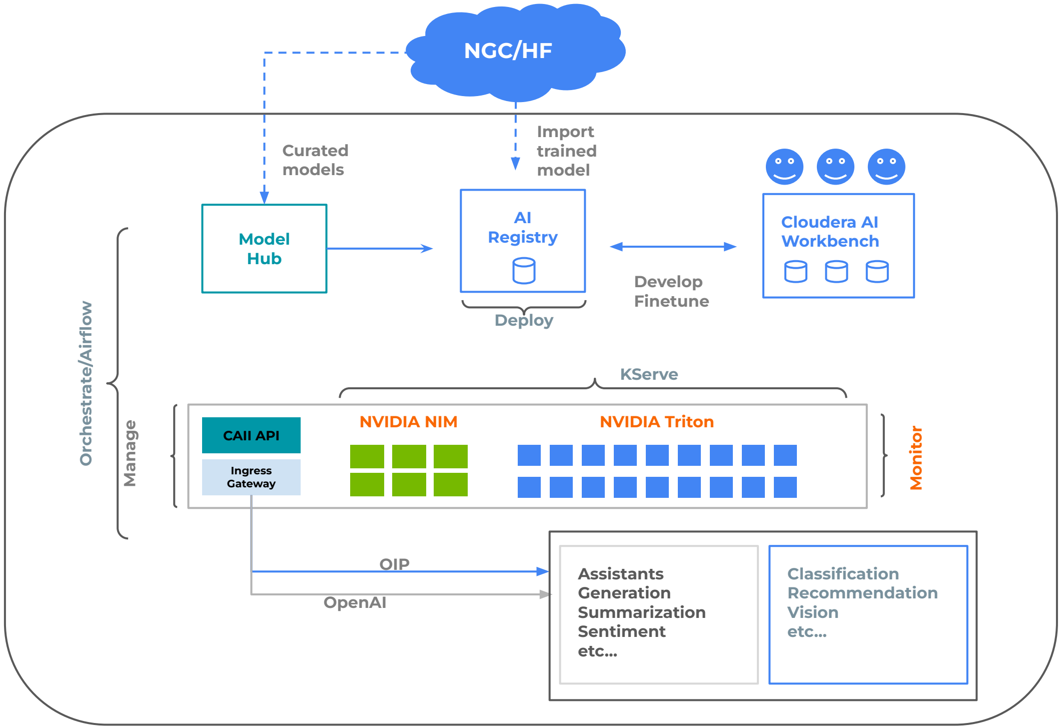
Why Did We Pick Kserve?

- Community and governance structure
- Technical reasons
  - Serverless
  - OIP
  - Multi-framework
  - Flexibility - custom runtimes
  - Monitoring, logging
  - Easy to security-fence

30,000 Feet

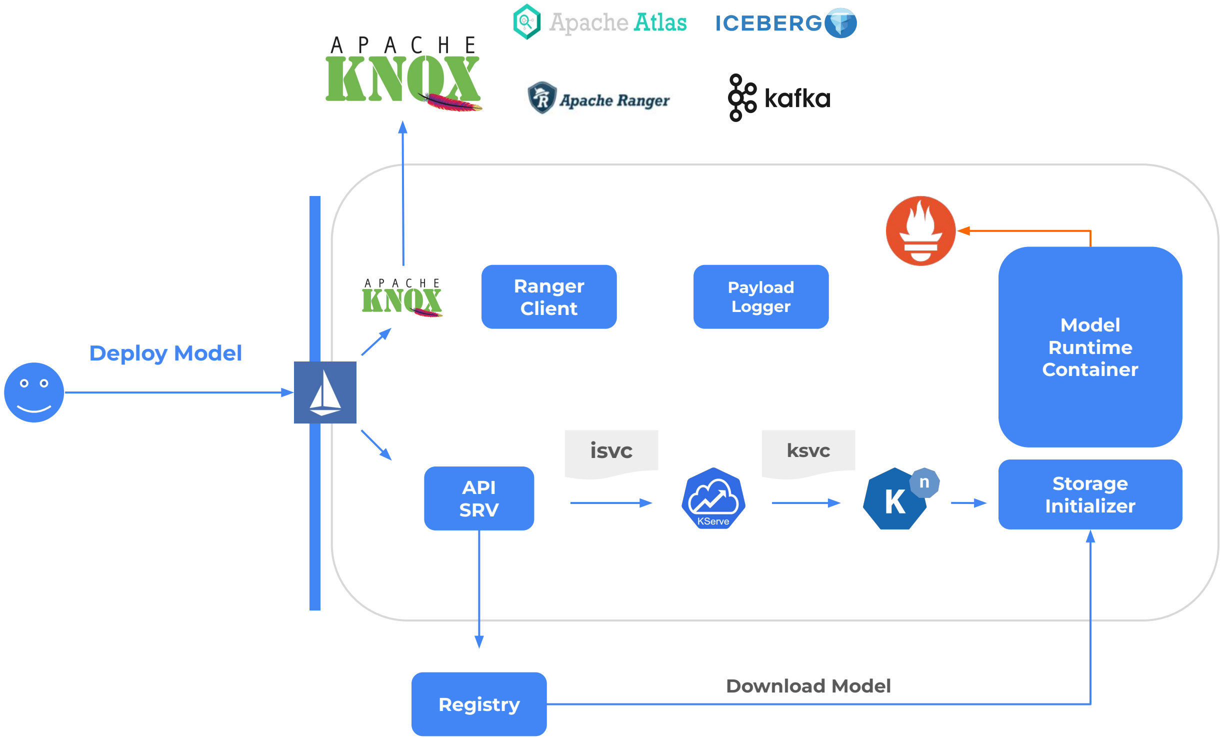

ON PREMISES & ON CLOUD
CLOUDERA AI WORKBENCH
CLOUDERA AI INFERENCE SERVICE
Autoscaling
&
High
Availability
KServe
K
n
Monitoring
ICEBERG
Apache Ranger
APACHE KNOX
Apache Atlas
AI Apps
Business Intelligence
Cloudera AI Assistants
CLOUDERA AI REGISTRY
LLaMA by Meta
cohere
FalconLLM
MISTRAL AI_

10,000 Feet

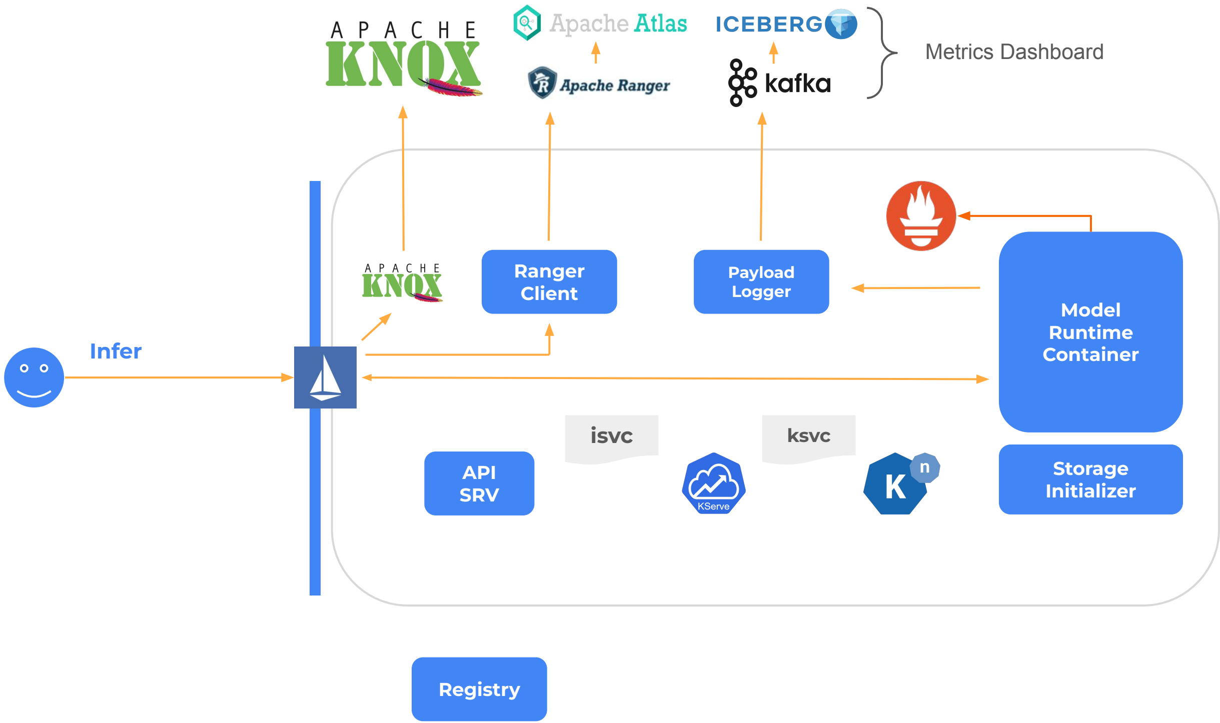


- 100s of endpoints
- 100s of replicas/endpoint
- heterogeneous GPU nodes
- heterogeneous CPU nodes
- autoscaling nodes
- autoscaling models

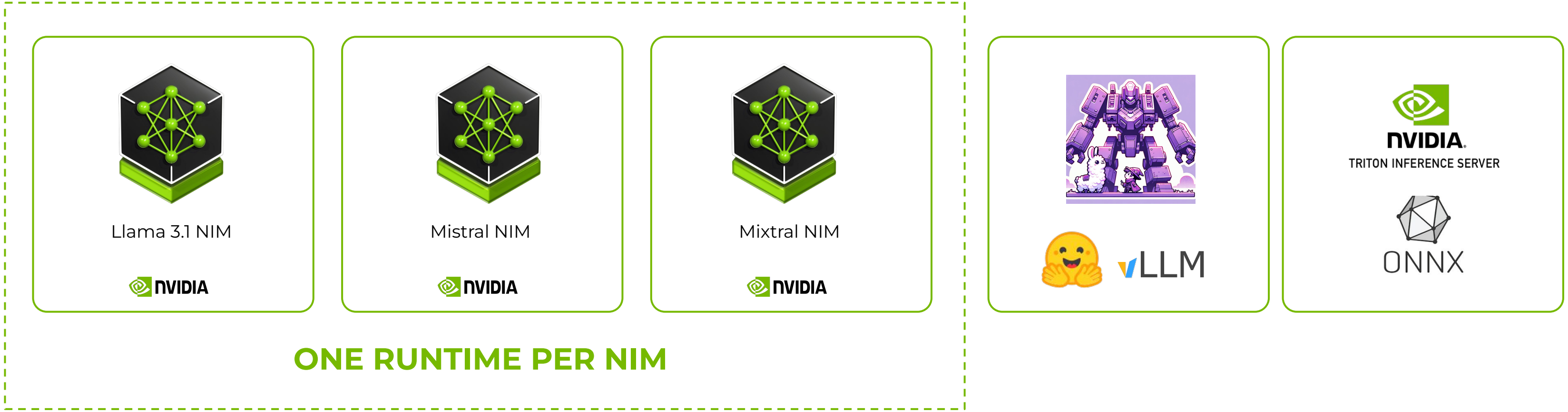
Control Flow



Inference Flow



Runtimes



Llama 3.1 NIM

NVIDIA

Mistral NIM

NVIDIA

Mixtral NIM

NVIDIA

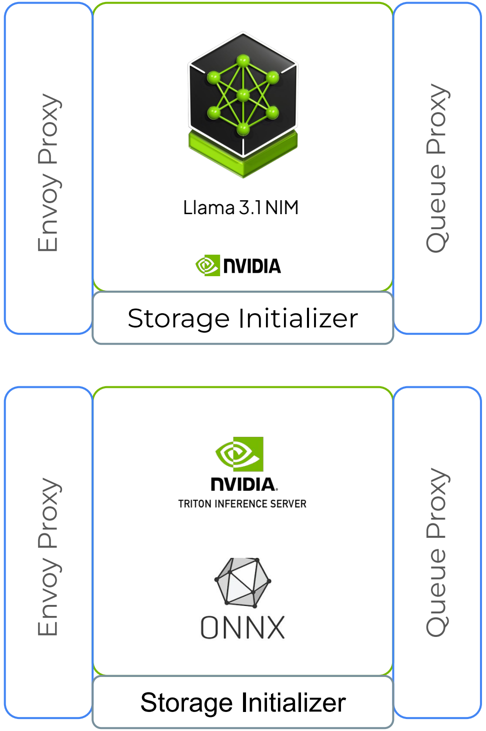
ONE RUNTIME PER NIM

vLLM

NVIDIA
TRITON INFERENCE SERVER

ONNX

Pod Level View



- Storage initializer
  - Downloads model artifacts
  - Some conversions
- Envoy proxy
  - Fine grained authorization (ext_authz) filter
- Queue proxy
  - For Knative serverless magic
- Main container runs model server

Key Takeaways

You can build a Robust Inference Platform:

1. **Run Anywhere**
   - Run on K8s to achieve multi-cloud and on-premises compatibility.
2. **Enterprise-Grade Scalability**
   - Start with KServe, supporting LLMs and traditional models with high availability and auto-scaling.
3. **Community-Driven Customizability**
   - Leverage KServe's open-source, community-driven ecosystem for flexibility and broad compatibility with multiple frameworks and custom runtimes.
4. **End-to-End Security**
   - Integrate with **Ranger, Knox**, and **Iceberg** for fine-grained access, monitoring, and i/o audit—critical for compliance.

"Our journey highlights the importance of choosing flexible, community-driven technologies and building security and scalability from day one — essential for enterprise AI at scale."

Thank You!

Q&A

Find us on LinkedIn:
Peter Ableda – Director, Product Management
Zoram Thanga – Principal Engineer, AI/ML Infrastructure

Give us feedback!