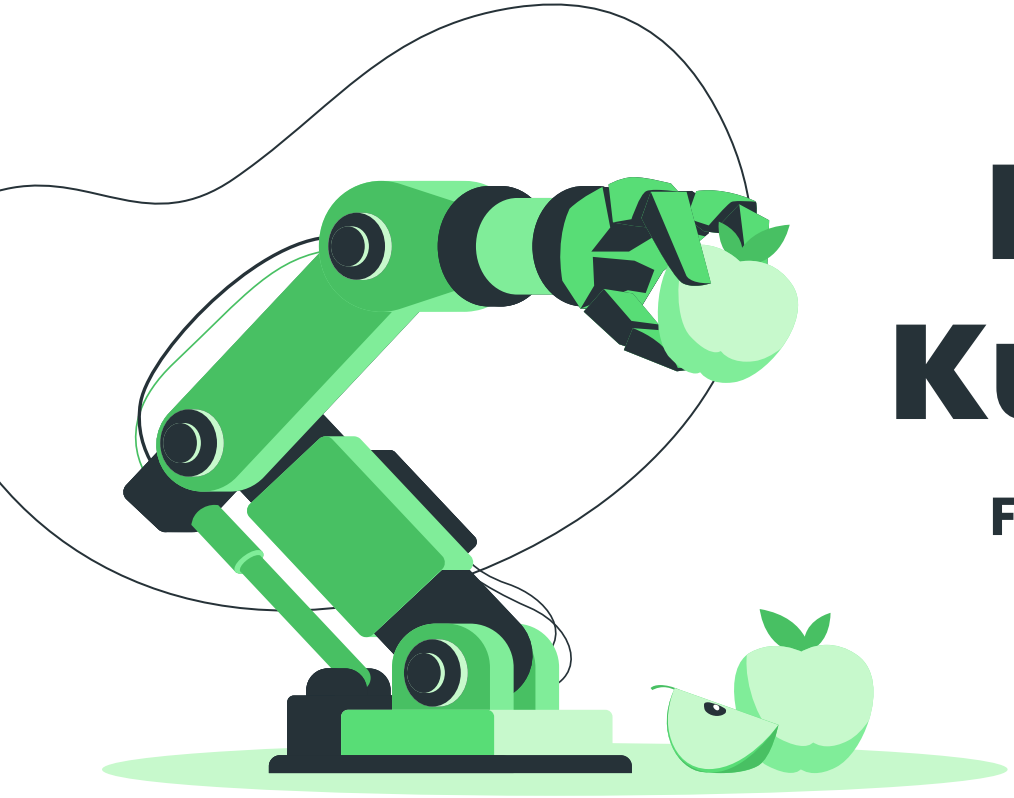


Herding LLaMAs on Kubernetes

For Sovereignty, Operations and Integration

Max Körbächer

Founder & Cloud Native Advisor @ Liquid Reply

With a focus on Platform Engineering, Internal Developer Platforms & Cloud Native Engineering

TAG ENVIRONMENTAL SUSTAINABILITY

- CNCF TAG Environmental Sustainability Initiator & Co-Chair
- CNCF Ambassador
- LF Europe Advisory Board
- Contributed 3y to the Kubernetes release team

maxkoerbaecher

mkoerbi

DIY or just buy?

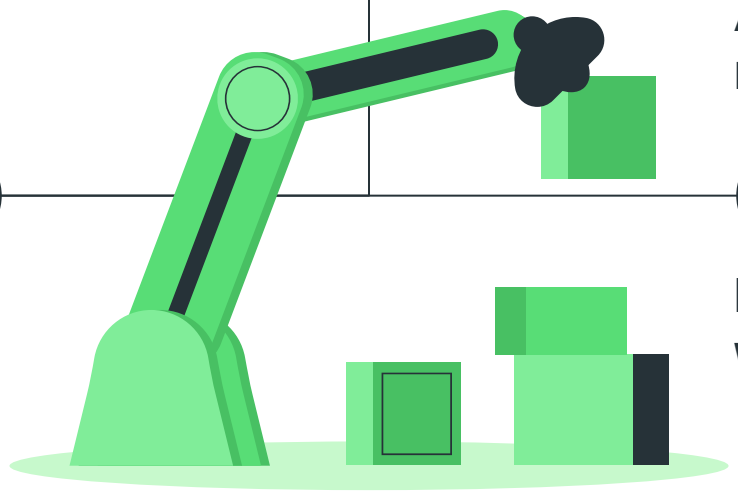
Buy-in ChatGPT & co

1

Extremely fast for prototyping

2

Zero to No expertise in AI/ML/LLM needed



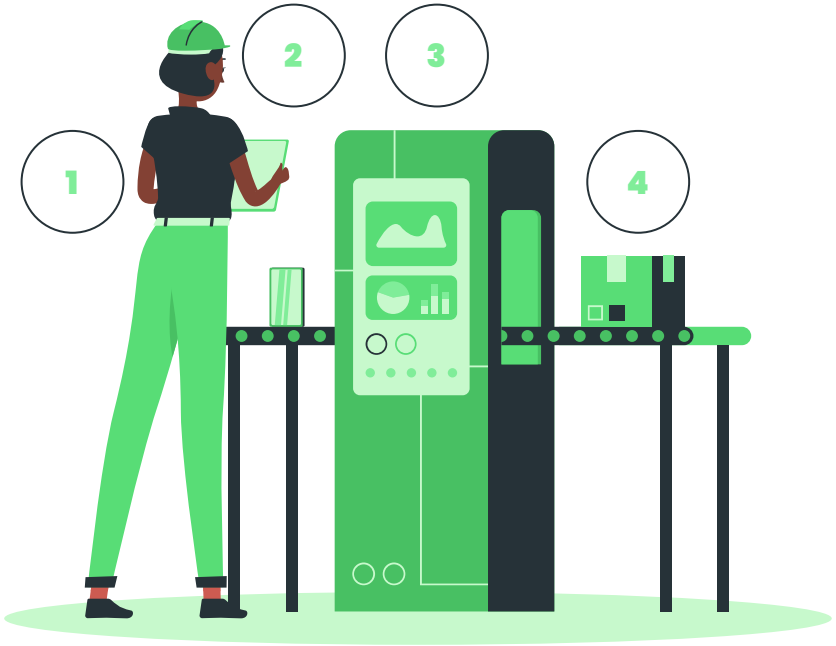
3

Provides you simple toolboxes and examples

4

Low costs to begin with

Self-hosting LLMs - the good

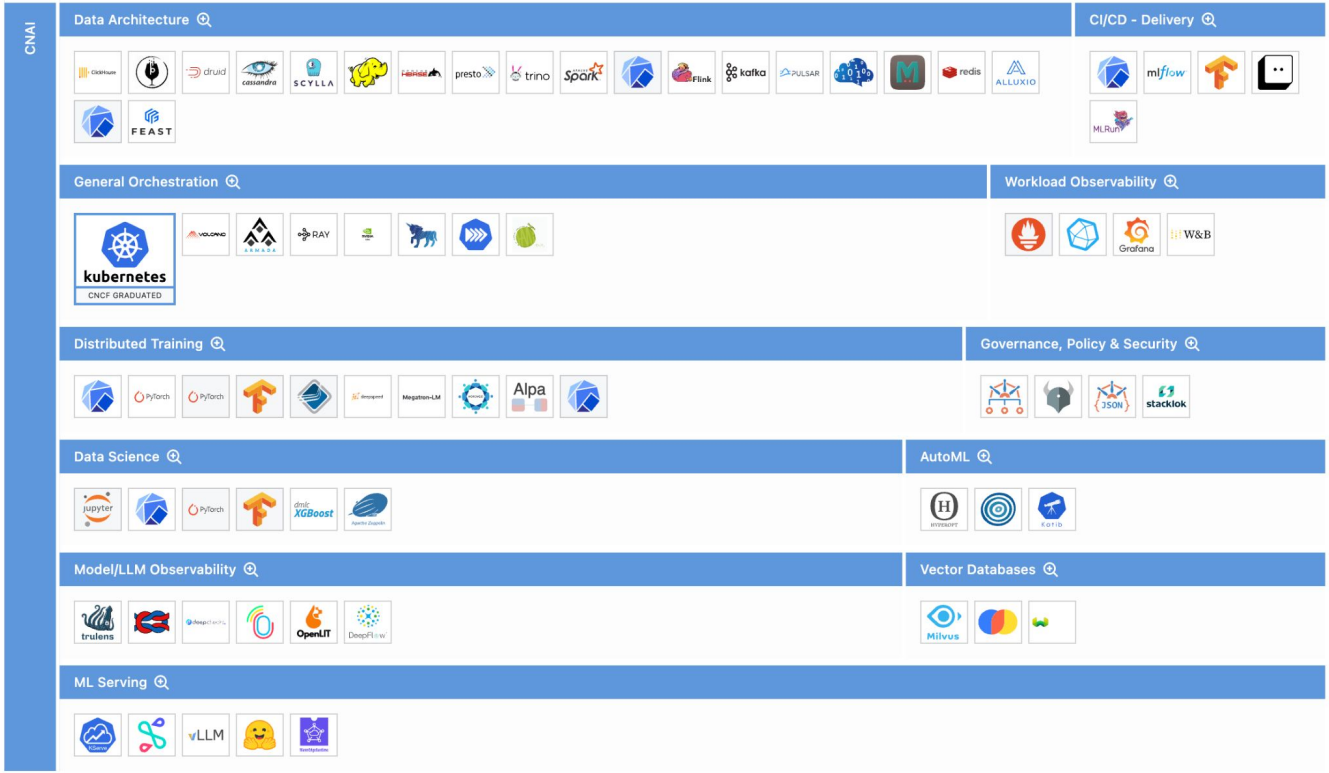


1. Avoid lock-in, enable specialized use cases
2. Greater customization possibilities
3. Enhanced Security, IP, Privacy, Compliance
4. Flatter cost curve on the long-run

Self-hosting LLMs - challenges

Model Size	Large sizes, issues in loading them fast, poor deployment speed, worse agility
Memory Requirements	Extreme memory requirements, good memory management is crucial
Compute Resources	Requires a lot in production/many end users, high operational costs
Latency	High inference latency due to their complexity
Scalability	Scaling GPUs or replace workload fast isn't trivial

Serving LLMs on K8s




virtual mac
kubernete


Expectation
Reality


How it started:
How it's going:

K8s benefits running AI w[illegible]oad

- **Portability** - allows to (more or less) r[illegible] anywhere
- **Scalability** - adjusts th[illegible]ed on its demand or adjust the workload to suit to[illegible]
- **Reso[illegible]** - Optimize utilization of resources
- **Ecosystem** - Feature rich and active ecosystem for tools, best practices and community

true BUT boring

Kubernetes is a platform to build platforms

Infrastructure Abstraction & Automation

Developer Platform (IDP)

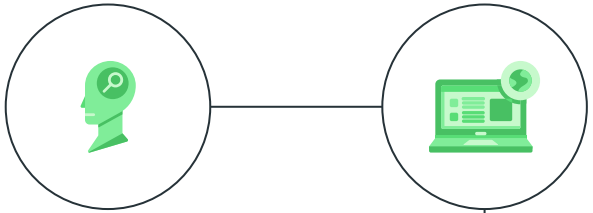
Edge Workload Handler (IoT)

AI/ML/LLM - the "new" kid on the block

K8s - Supporting Capabilities for LLMs

Observability

Track performance & health



CI/CD - "GitOps"

Adopt best practices for automation

Data Management

Utilize persistent & dynamic volumes

Security

Build on a rich ecosystem for protection and separation/isolation

Observability

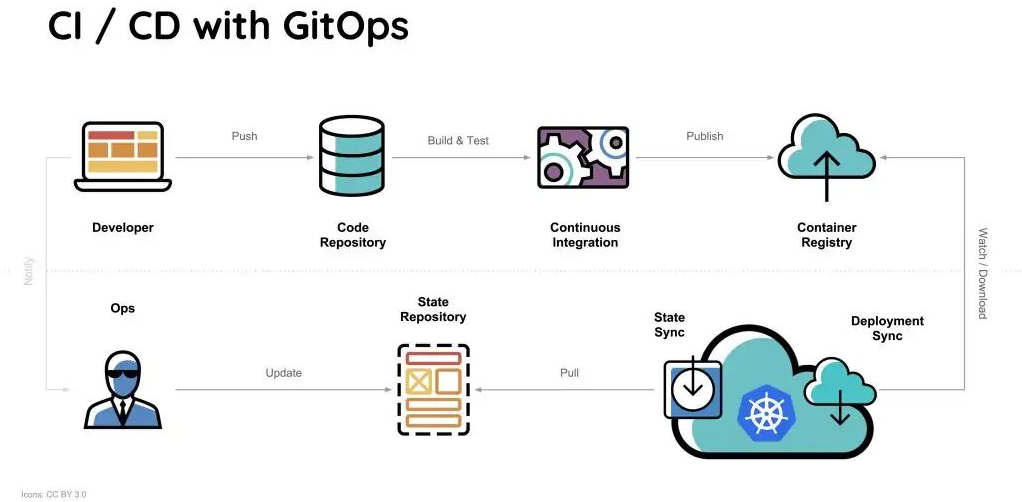
The K8s default provides a full transparency from infrastructure, through network, to the application layer.

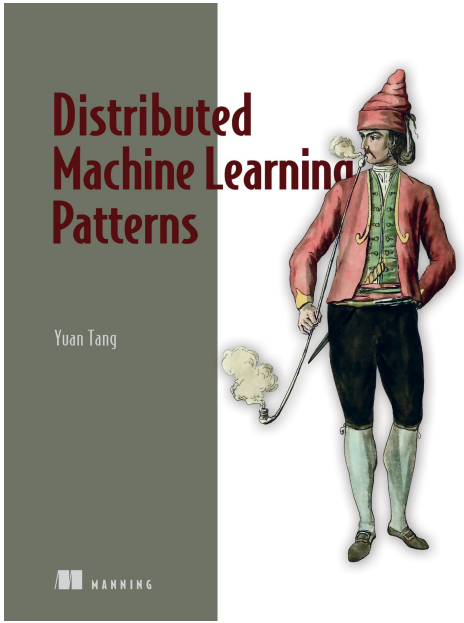
OpenTelemetry

OpenLLMetry

CI/CD – “GitOps”

CI/CD and GitOps provided the foundation for MLOps and LLMOps.





Source: Cloudflight

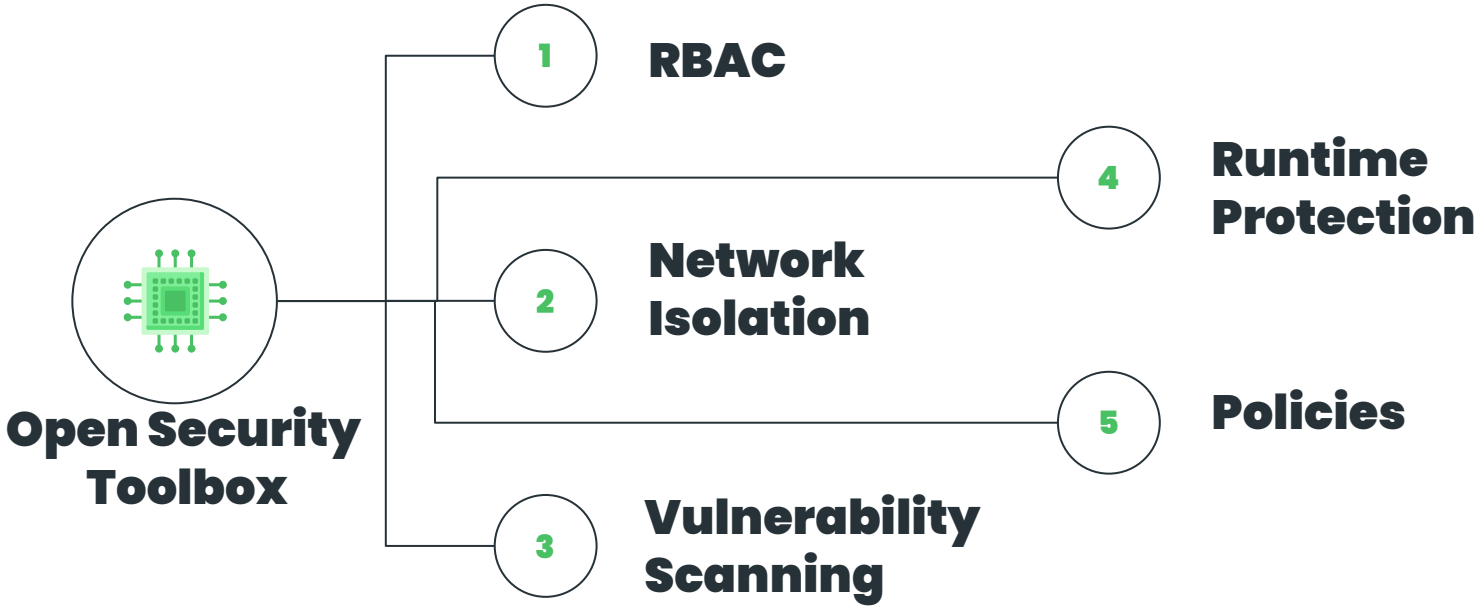
Data Management

No matter where, no matter what, K8s provides a unified approach of integrating with the storage.

For the end user it is always just a claim, how it is done is out of the sight.

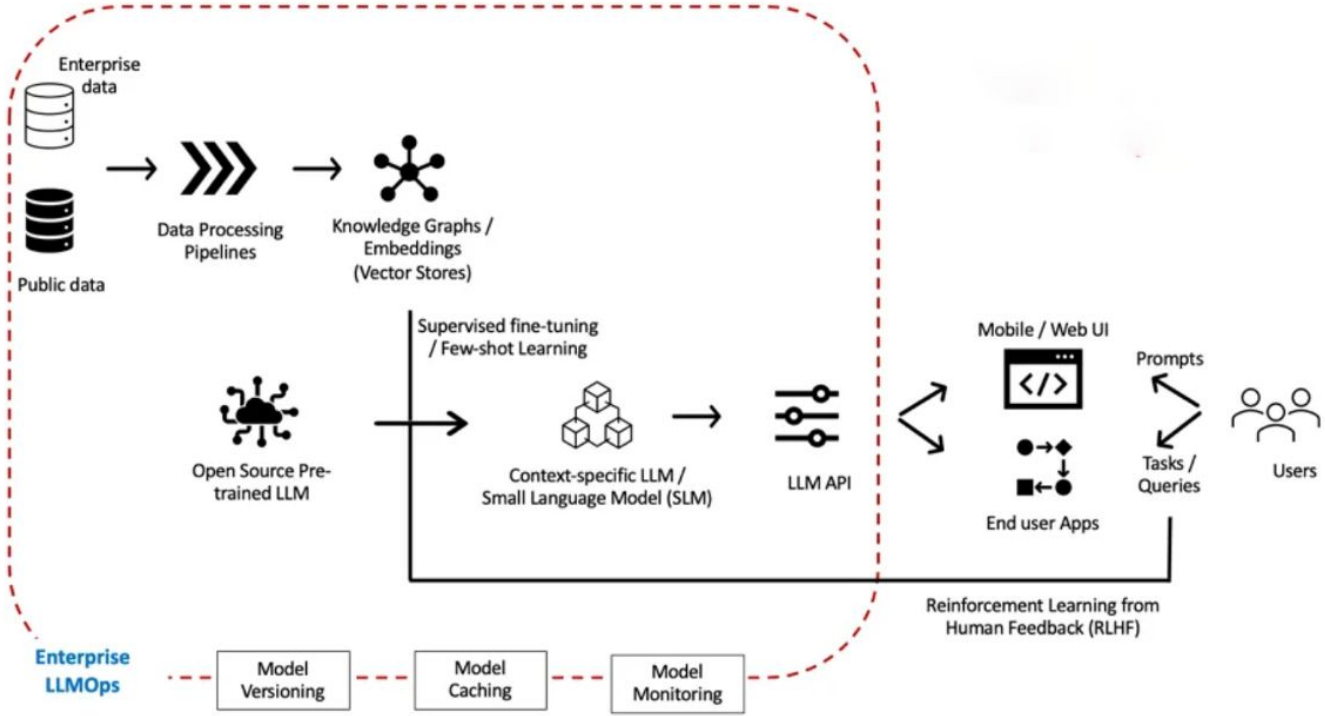
Re-claiming storage, backup of cluster and application state with the data and event-based triggers to adjust to certain metrics improve reliability drastically.

Security


1
RBAC
4
Runtime Protection
2
Network Isolation
5
Policies
3
Vulnerability Scanning
Open Security Toolbox

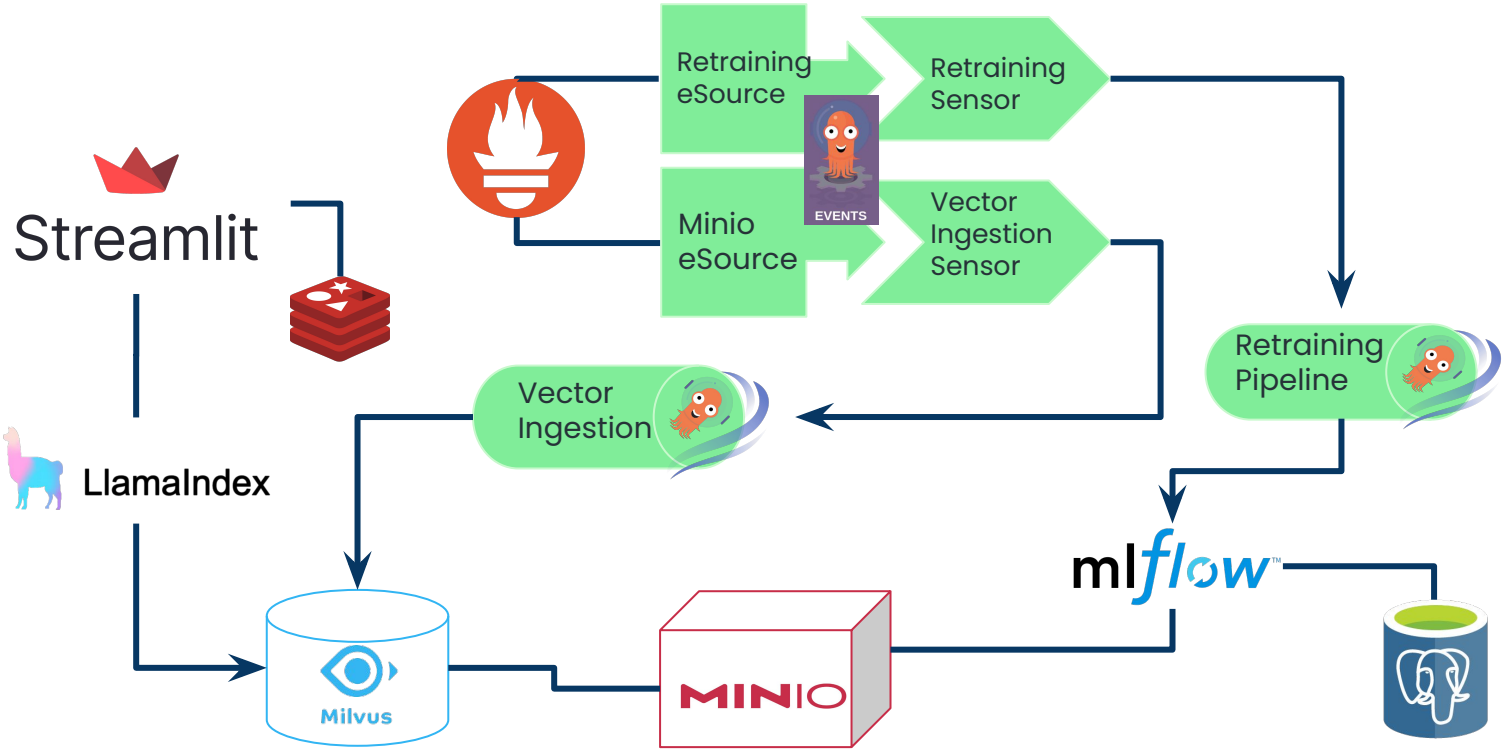
Let’s DIY

LLMOps Workflow

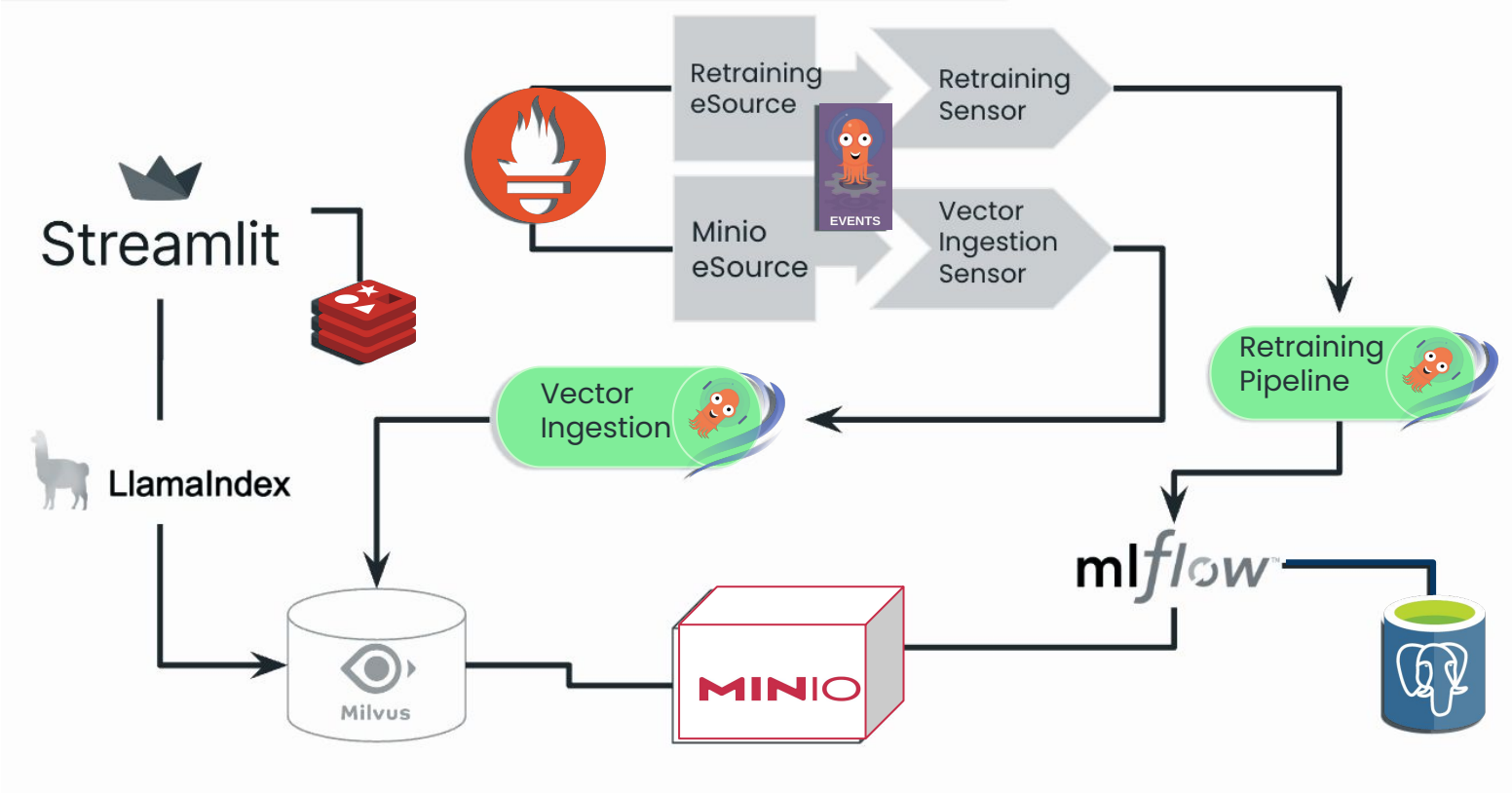


Source: Debmalya Biswas

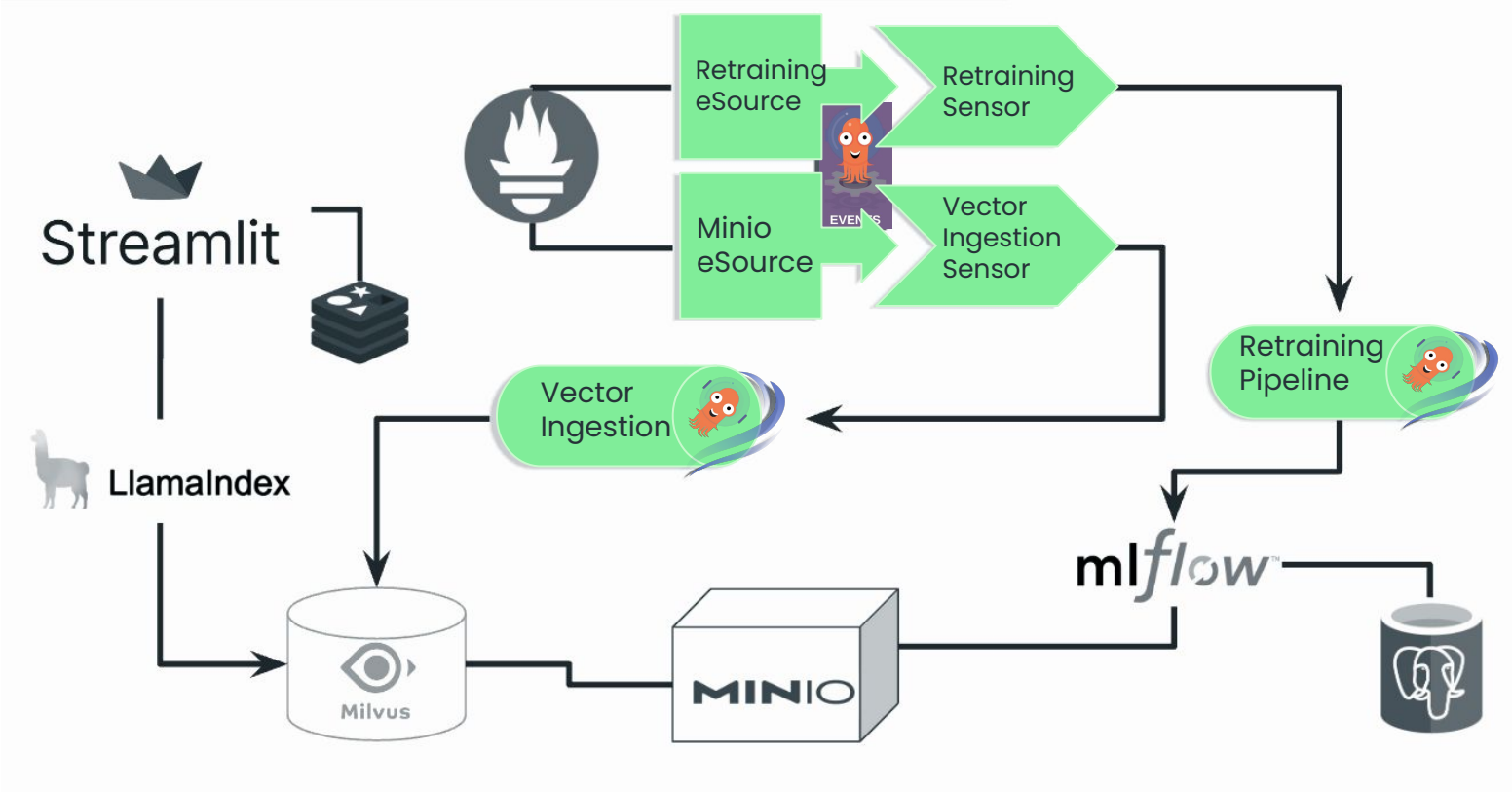
LLMOps example Architecture



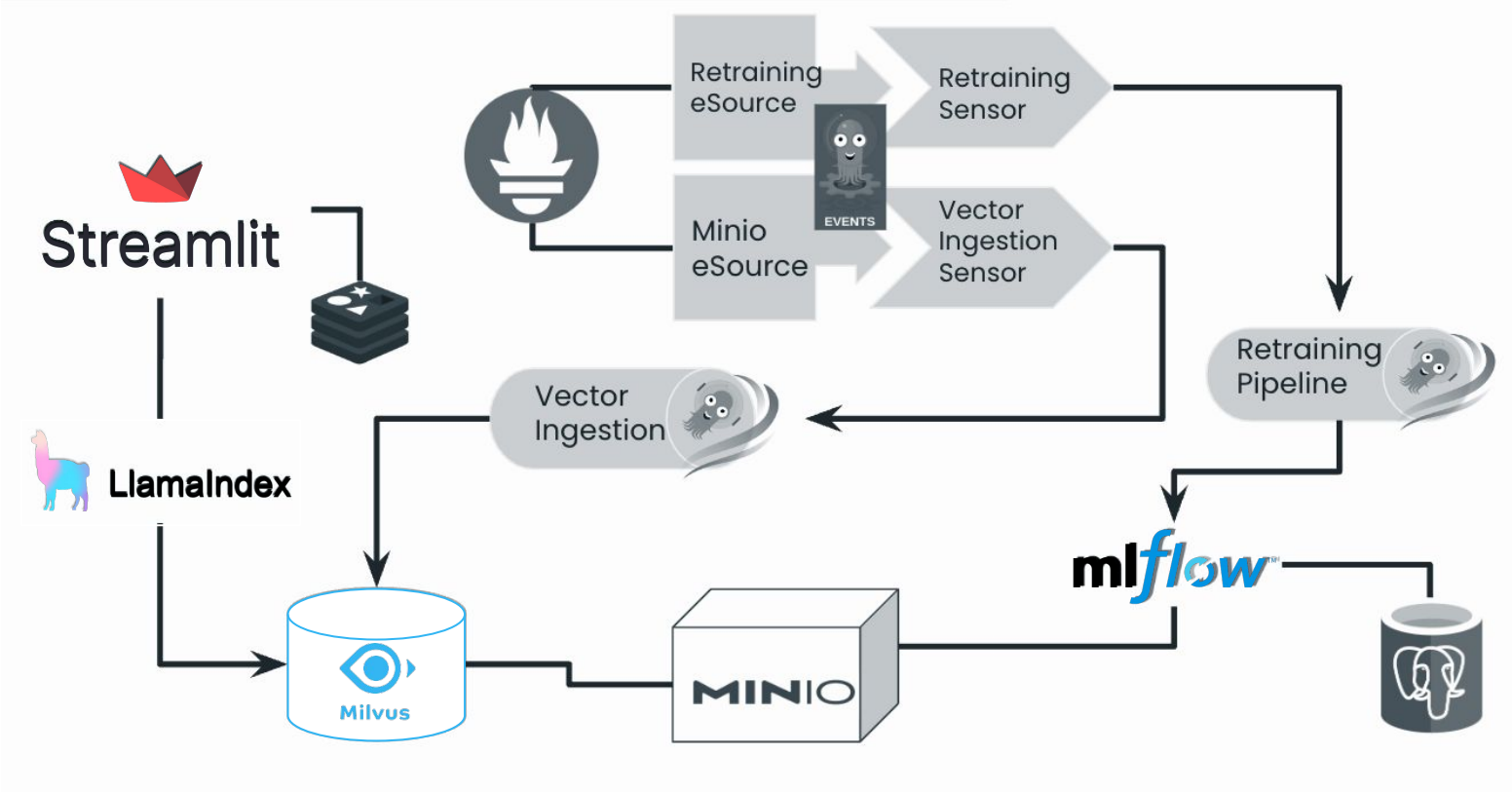
LLMOps example Architecture



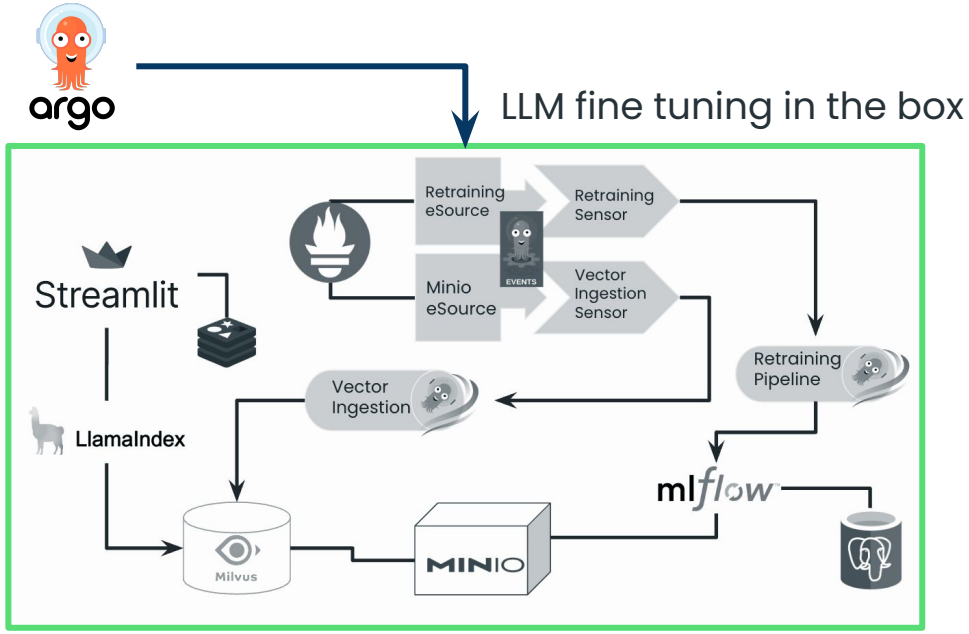
LLMOps example Architecture



LLMOps example Architecture



Kubernetes Empowerment



Kubernetes

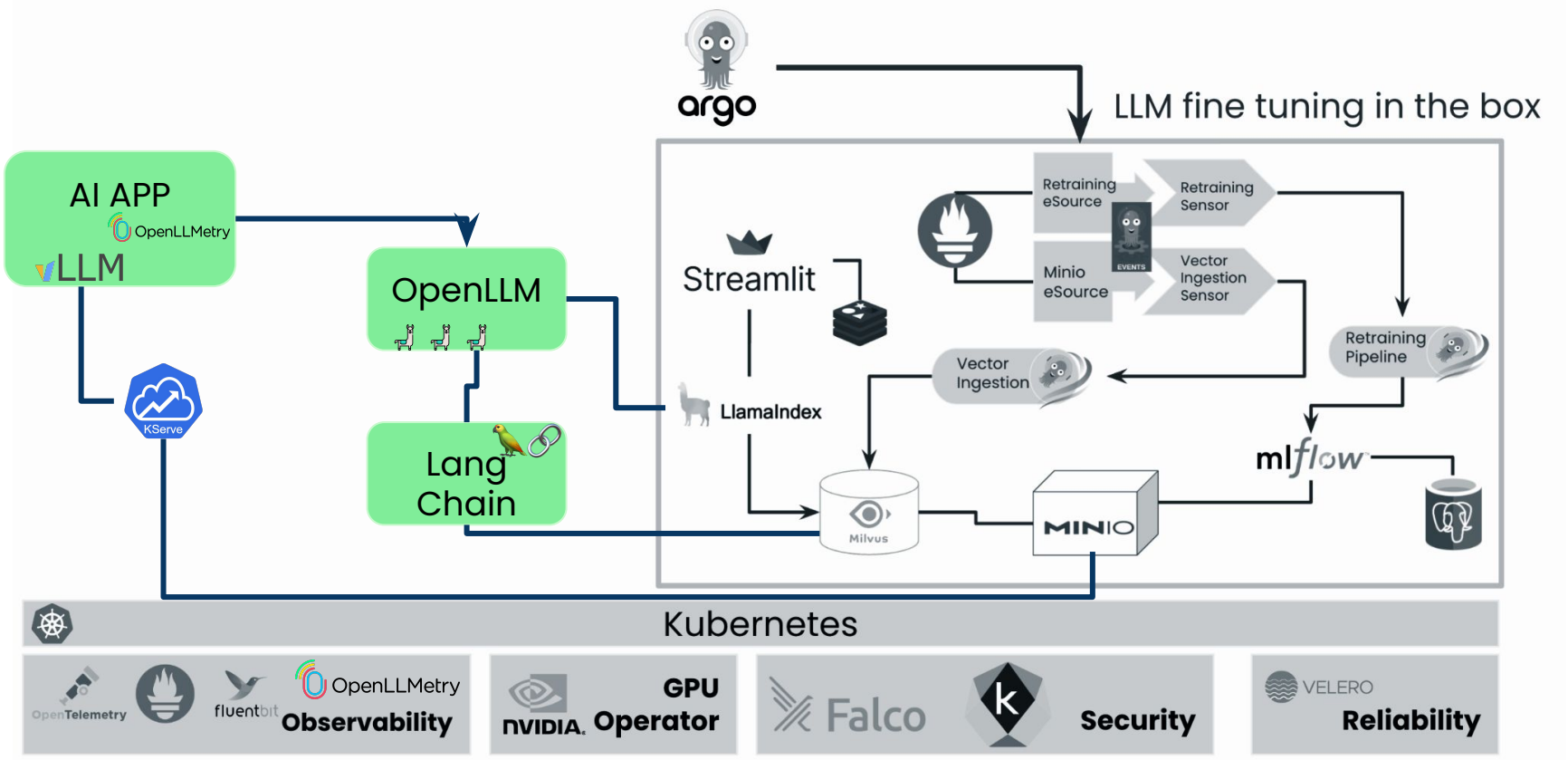
OpenTelemetry fluentbit **Observability**

NVIDIA **GPU Operator**

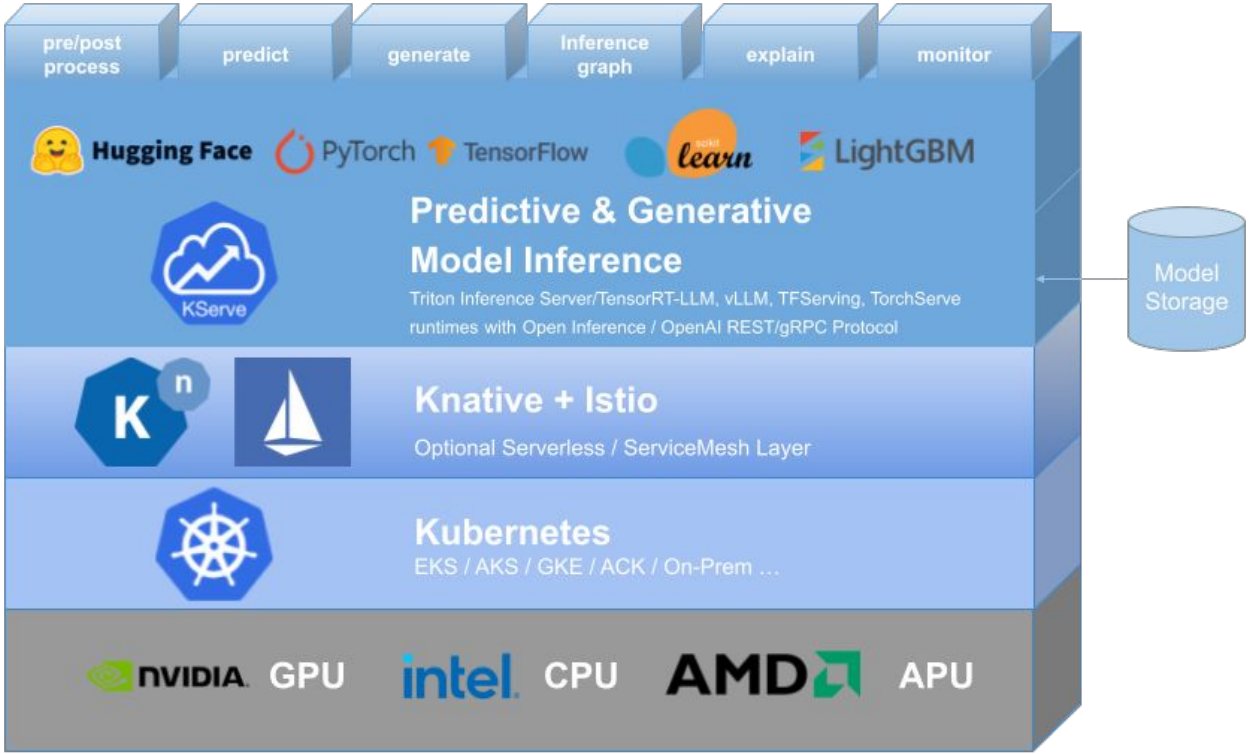
Falco **Security**

VELERO **Reliability**

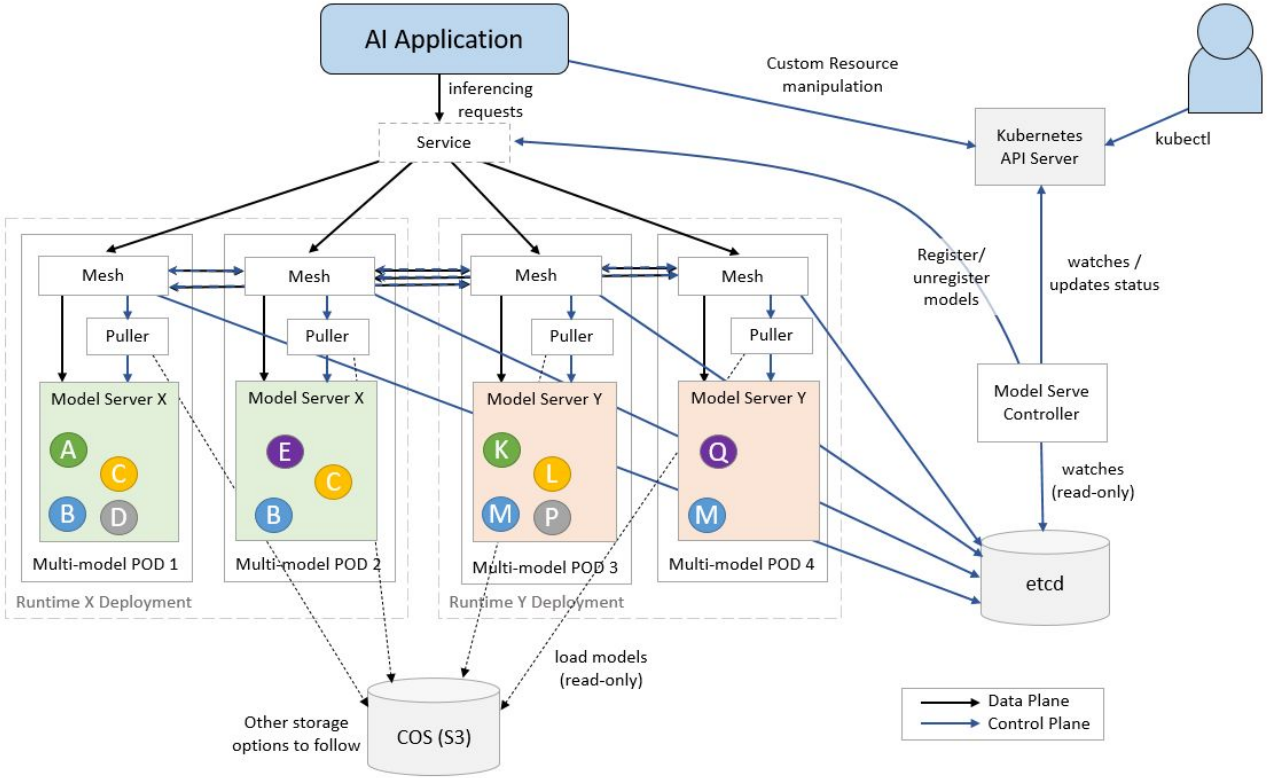
Using LLMs on K8s



10k feet view kserve



10k feet view kserve




An
Outlook

We (will) have various specialized LLMs to serve

That’s where K8s plays strong

Flexible provisioning & integrations

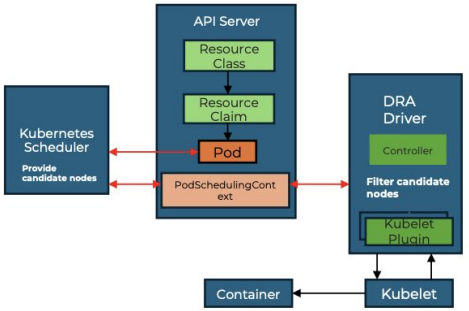
K8s support a very dynamic scaling and shifting of workload based on metrics.

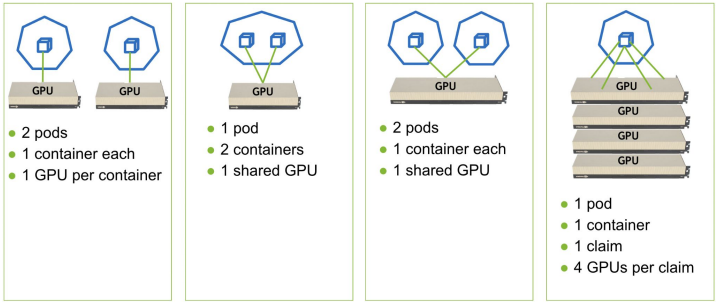
Adjust resources allocation - DRA

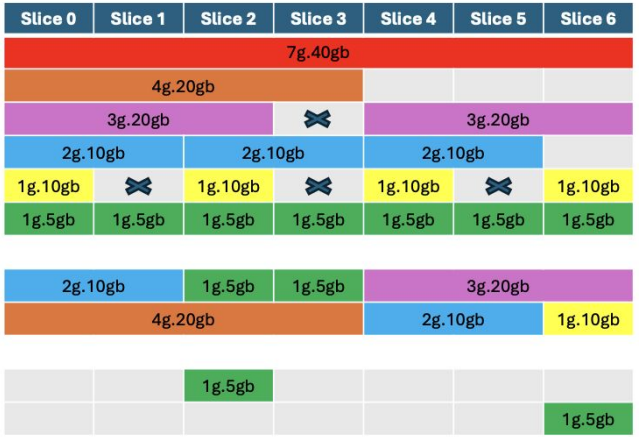
Dynamic Resource Allocation to improve GPU and memory utilization. -> better for the business

Dynamic Resouce Allocation

Dynamic Resource Allocation (DRA) allows users to create MIG slices







Example A100-40GB MIG layouts

Wasm & Inferencing

Problems

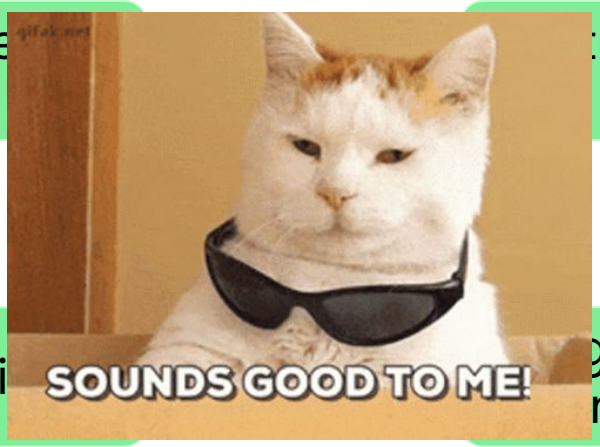
- Efficiency
- Language Dependencies
- Platform Dependencies

Power

- Minimum Resource Footprint
- Dev flexibility
- Secure
- Close to data

Combining DRA, kserve & WASM

Runs “anywhere
any scale



treme resource
efficient

Very cost effici

ghly dynamical
nd adjustable

Thank you!