CLOUD NATIVE & KUBERNETES AI DAY NORTH AMERICA

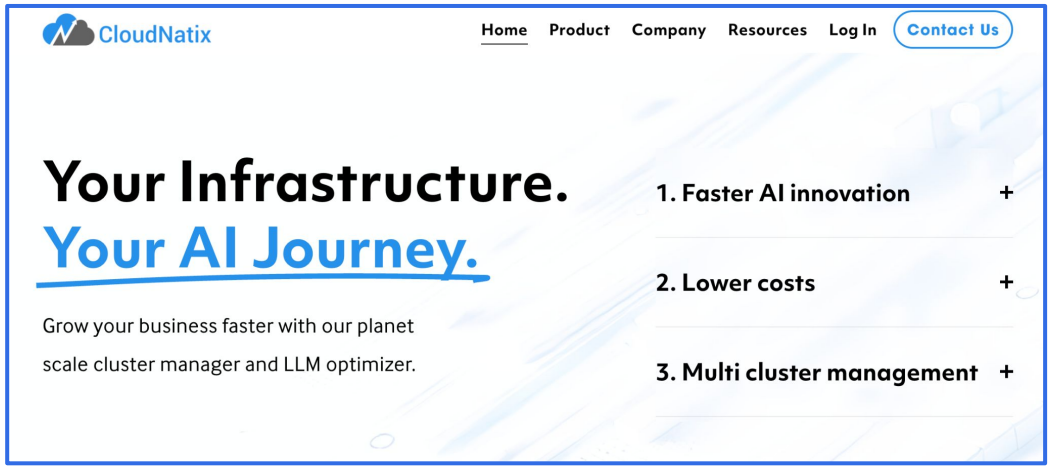
Transform Your Kubernetes Cluster Into a GenAI Platform

Get Ready-to-Use LLM APIs Today!

Kenji Kaneda

Self Introduction

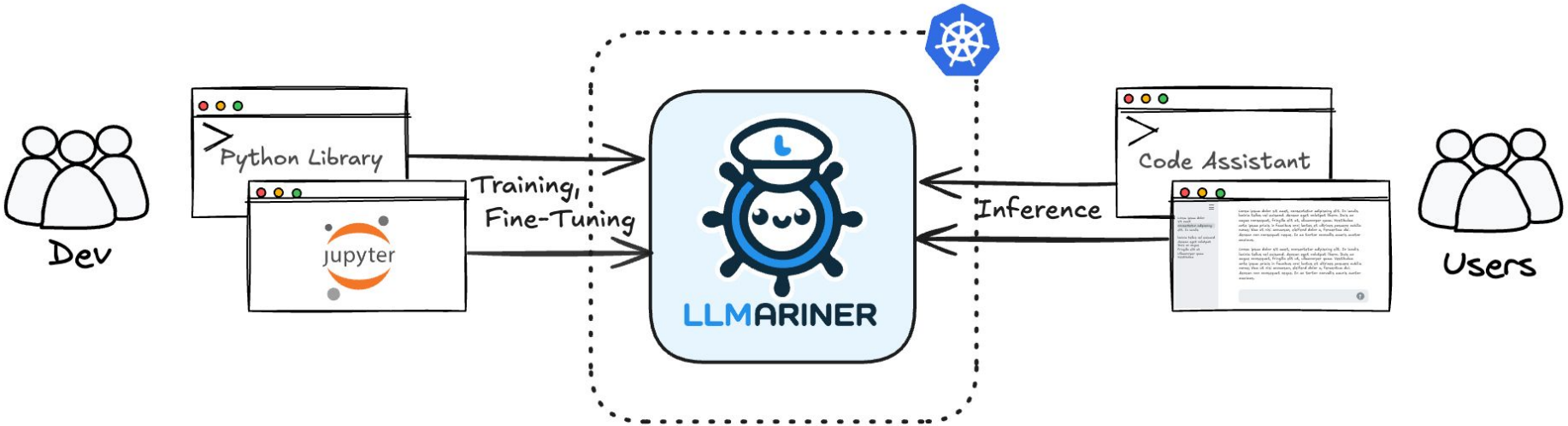
- Kenji Kaneda
  - Chief architect @ CloudNatix
  - Ex-Nvidia, Square, and Google



https://cloudnatix.com

LLMariner (= LLM + Mariner)

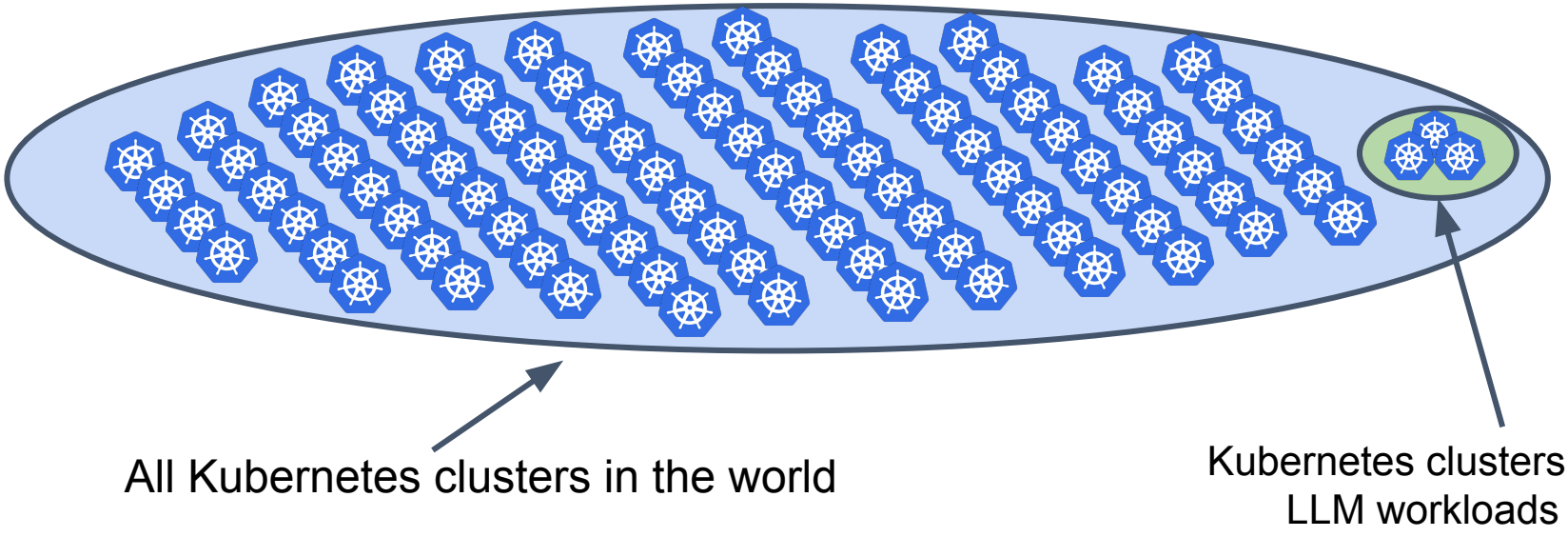
An open source platform for simplifying the management of generative AI workloads



https://llmariner.ai/

Why LLMariner?

Because not many people are running LLMs in K8s



Why Does K8s Lack LLMs?

Focus on this talk!

Non technical reasons

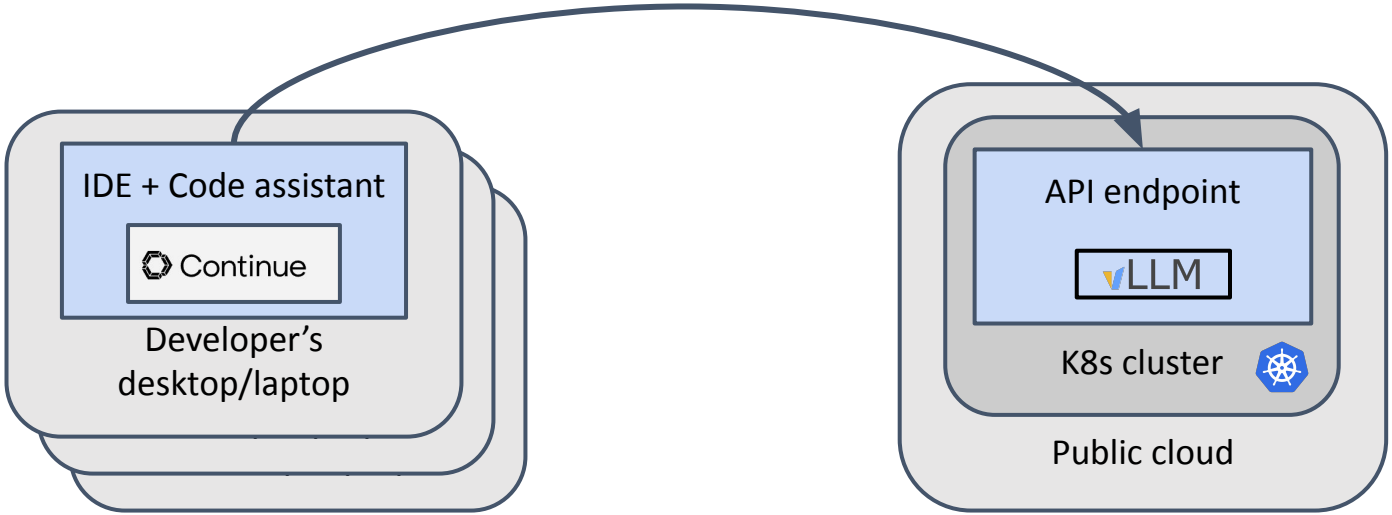
- No real use cases
- No need to host in K8s
- No budget for GPU
- …

Technical reasons

- Difficulty in satisfying the enterprise requirements on security, reliability, and efficiency

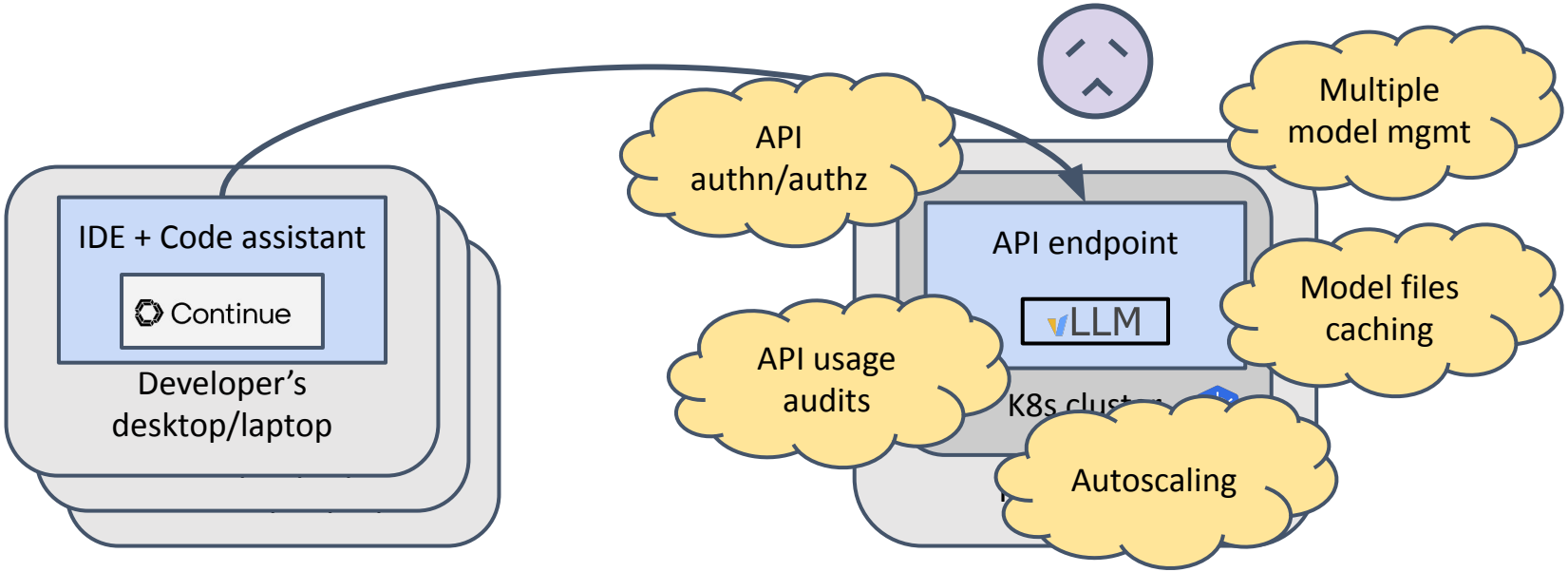
A Real-World Scenario

- Deploy LLMs in your K8s cluster to power a coding assistant
  - Continue (https://continue.dev) as a VS Code plugin
  - vLLM for serving inference requests



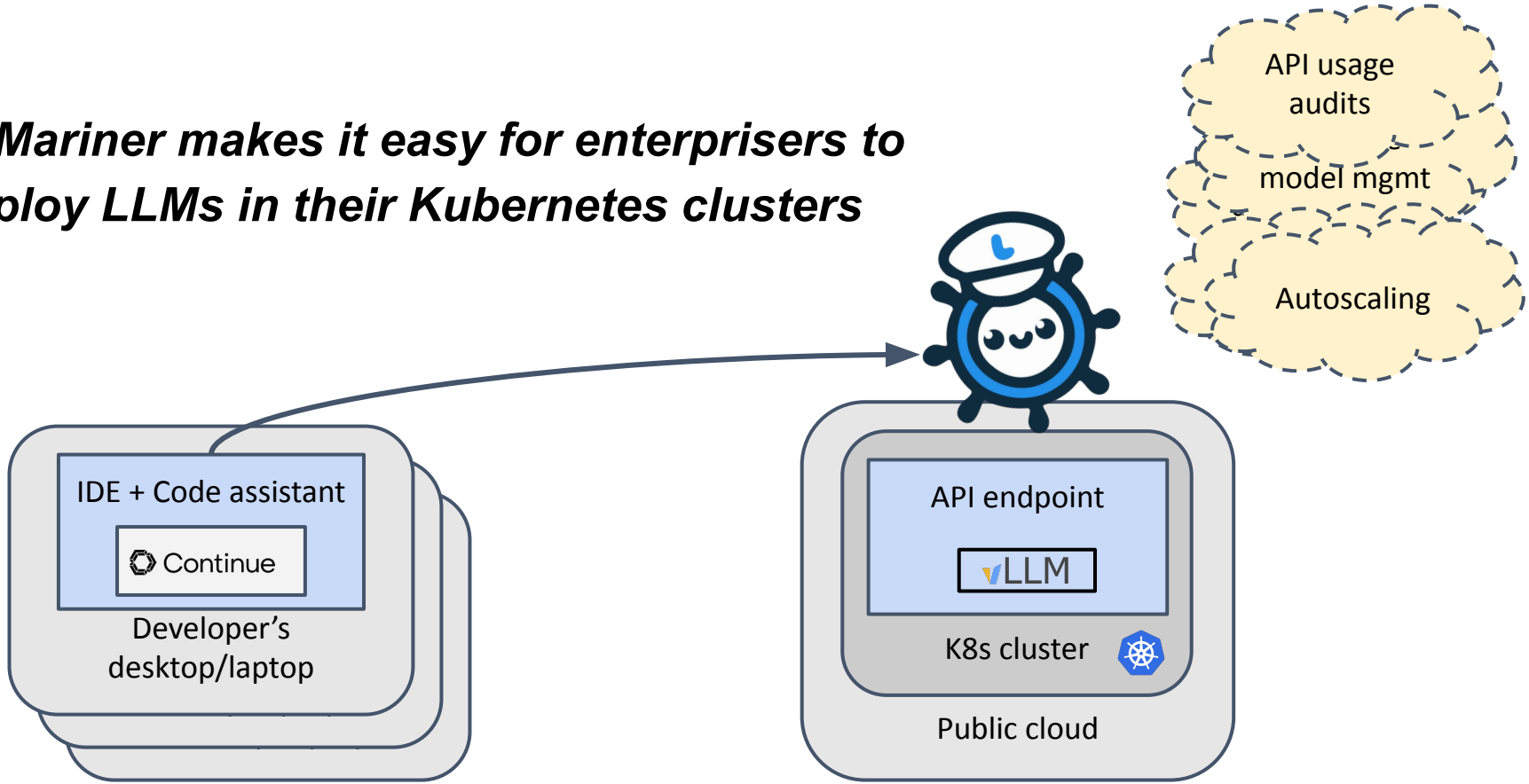
Enterprise Requirements Add Complexity

Hosting vLLM containers in the cluster is not sufficient to meet the enterprise requirements



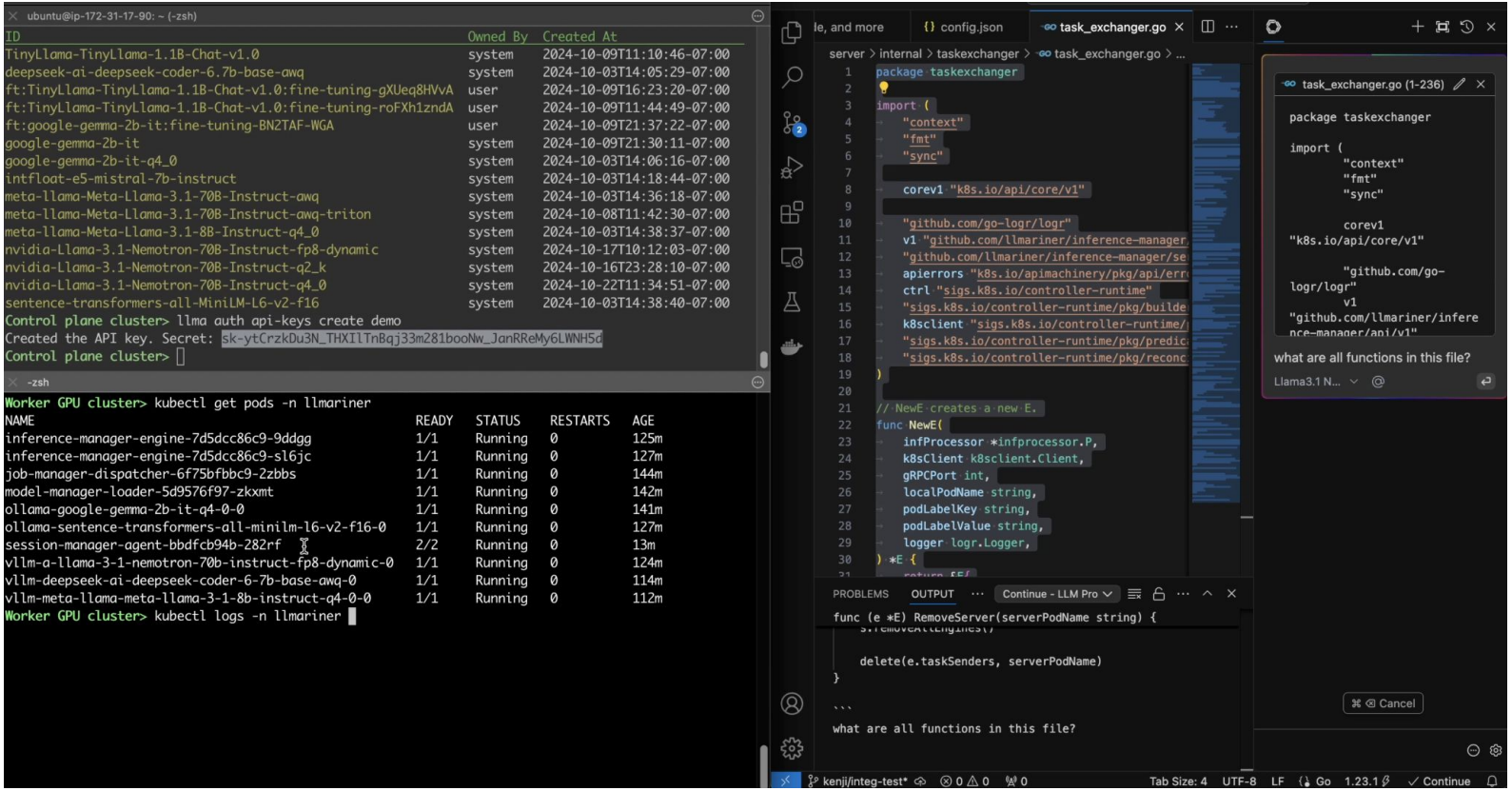
Breaking Barriers with LLMariner

LLMariner makes it easy for enterprisers to deploy LLMs in their Kubernetes clusters



Demo - Coding Assistant with LLMariner

CLOUD NATIVE & KUBERNETES AI DAY NORTH AMERICA



Link: https://vimeo.com/1024806457

LLMariner Features for AI/ML and Infra Teams

APIs for the AI/ML team

OpenAI-compatible API
(chat completion, embedding, RAG, fine-tuning, …)

Workbench with
Jupyter Notebooks

General purpose
training jobs

Model mgmt

- Open models
- Closed models owned by your org
- Fine-tuned models

Inference engine

Runtime mgmt
(e.g., autoscaling, routing)

- vLLM
- Nvidia Triton
- Ollama

GPU workloads mgmt

- Fine-tuning jobs
- Jupyter Notebooks
- Training jobs
- Kueue

Storage mgmt

- Files
- Vector DBs
  - Milvus

User mgmt

- Dex

API authn/authz

- API key mgmt
- Orgs & projects mgmt

API usage audits

Cluster federation

- Cluster mgmt
- Secure session mgmt

K8s cluster

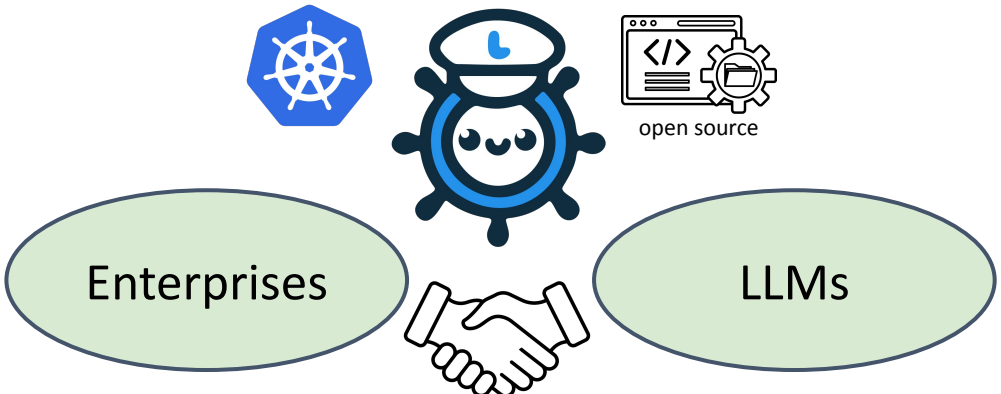
K8s cluster

K8s cluster

Opportunities with LLMariner

LLMariner opens various opportunities for enterprises

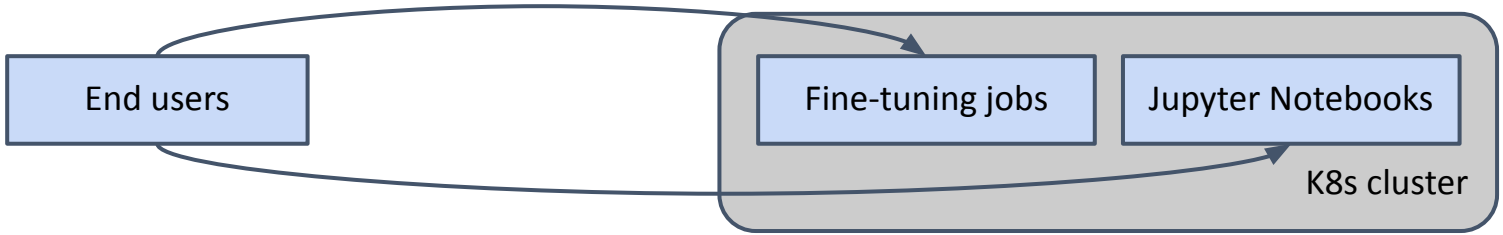
- Provides **full control** by hosting LLMs on your infrastructure
- Enables teams to catch up with the rapid AI/ML evolution by providing a foundation for **open source technologies**



Looking at One of the Features Closely

Management of fine-tuning jobs and Jupyter Notebooks

- View logs of fine-tunings jobs
- Exec into the containers of fine-tuning of jobs
- Open Jupyter Notebooks

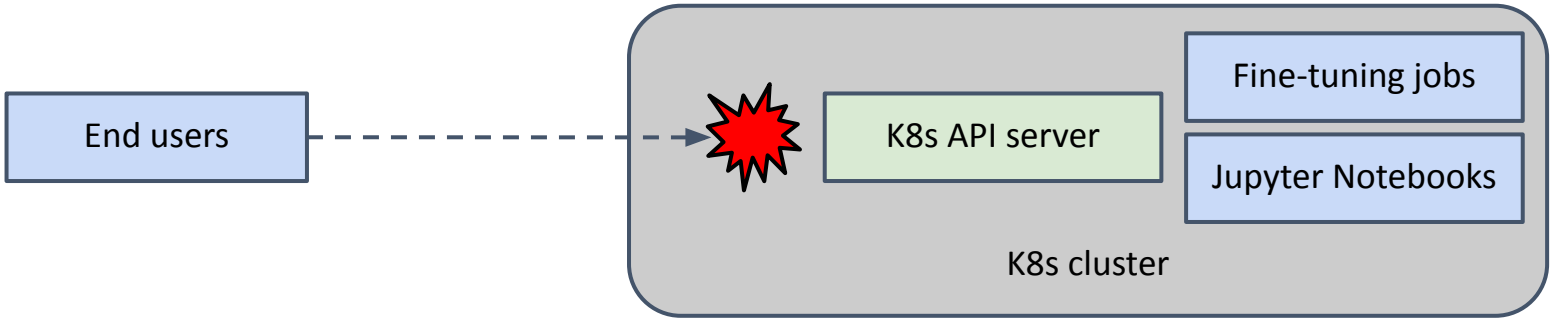


How should we implement?

- Provide direct access to the K8s API server (and use CRDs to track fine-tuning jobs and Jupyter Notebooks)?

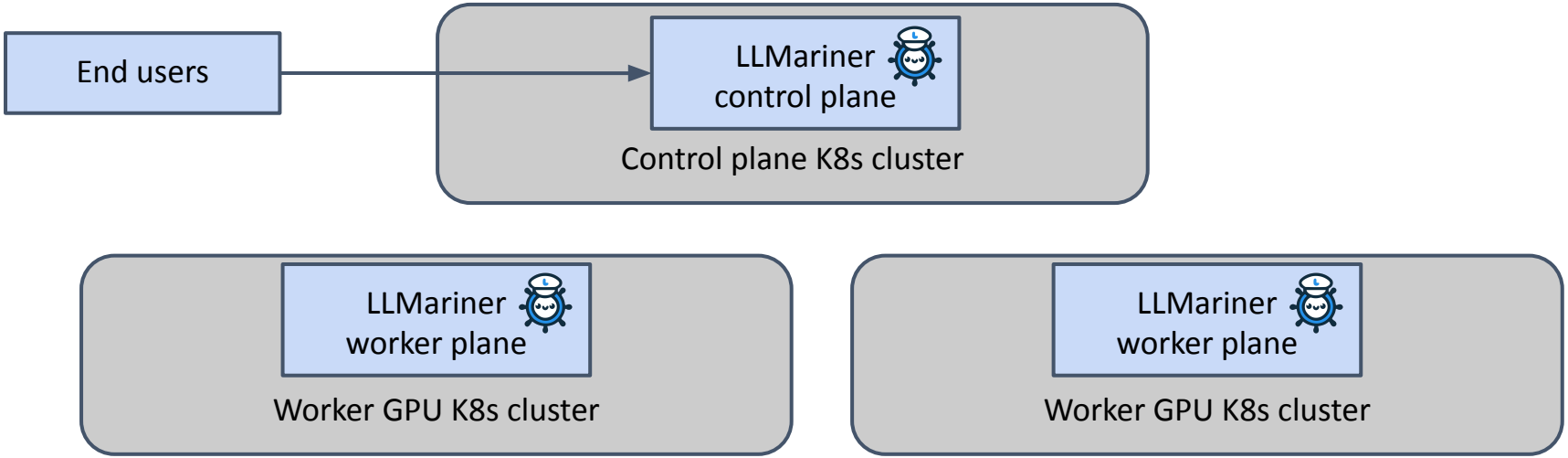
Potential Issues with Direct Access to K8s API Server and CRDs

- End users might lack:
  - Network connectivity to the K8s API server
  - Access privilege to the K8s API server
  - Knowledge/experience on K8s API



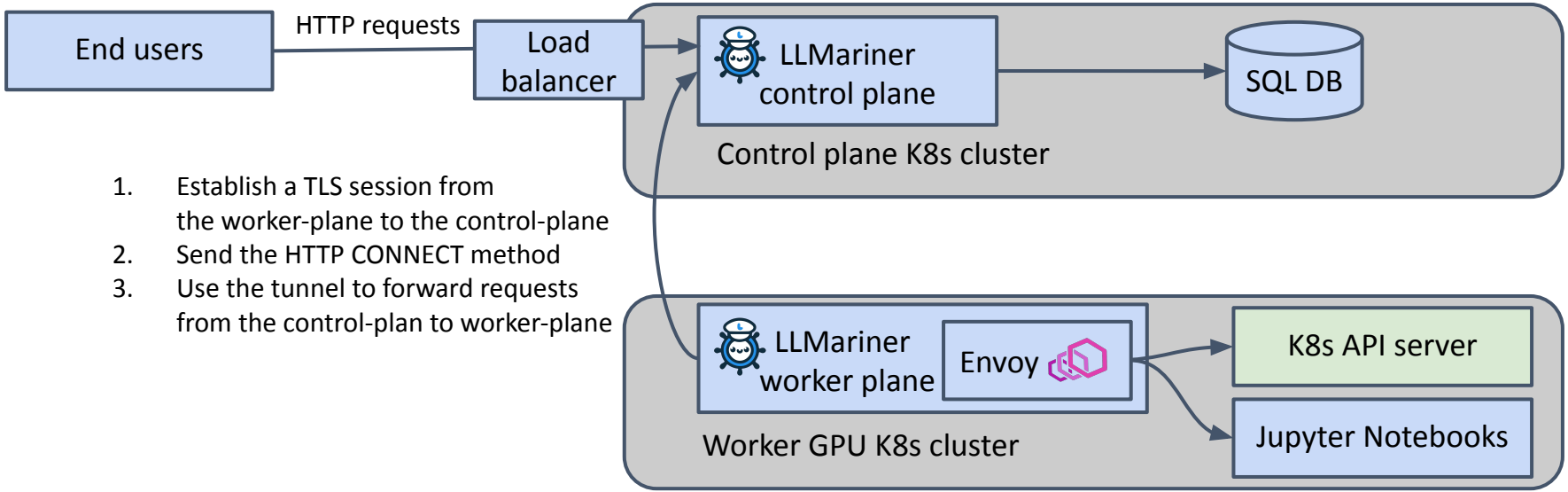
Potential Issues with Direct Access to K8s API Server and CRDs

- Complexity with supporting multi-cluster federation
  - E.g.) Deploy LLMariner control-plane and worker-plane separately
  - E.g.) Support multiple worker GPU clusters across multiple clouds



Our Approach on Fine-tuning Jobs and Jupyter Notebooks Management

- Enable the control plane to communicate to the worker plane without opening an incoming port
- Use an SQL database to track fine-tuning jobs and Jupyter Notebooks



Thank you!

LLMariner makes LLMs ready for enterprise

Please visit https://llmariner.ai to learn more!

Contact Information

CLOUD NATIVE & KUBERNETES AI DAY NORTH AMERICA

Email: kenji@cloudnatix.com