



CLOUD NATIVE &
KUBERNETES

AI DAY

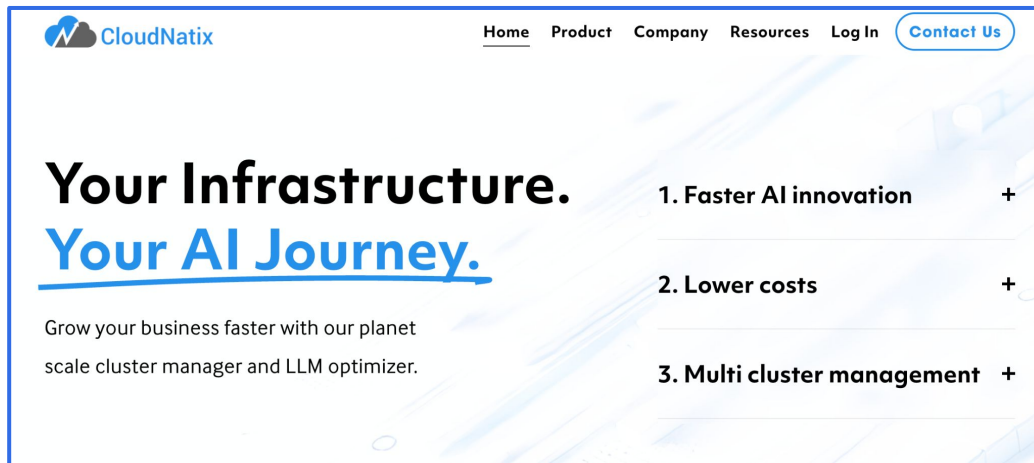
NORTH AMERICA

Transform Your Kubernetes Cluster Into a GenAI Platform

Get Ready-to-Use LLM APIs Today!

Kenji Kaneda

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



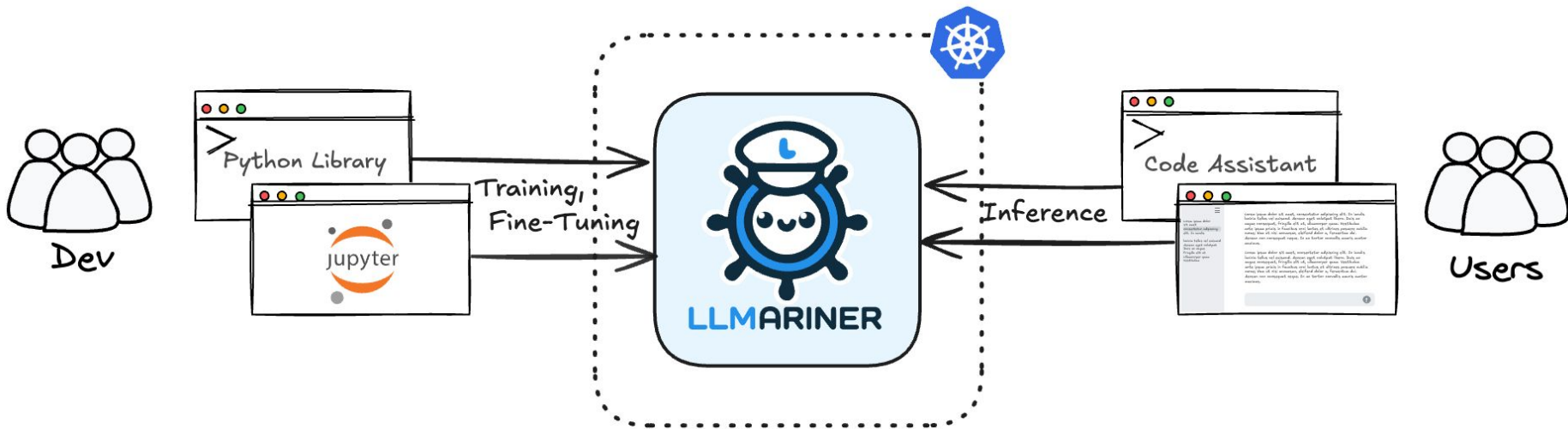
<https://cloudnatix.com>

LLMariner (= LLM + Mariner)



CLOUD NATIVE &
KUBERNETES
AI DAY
NORTH AMERICA

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



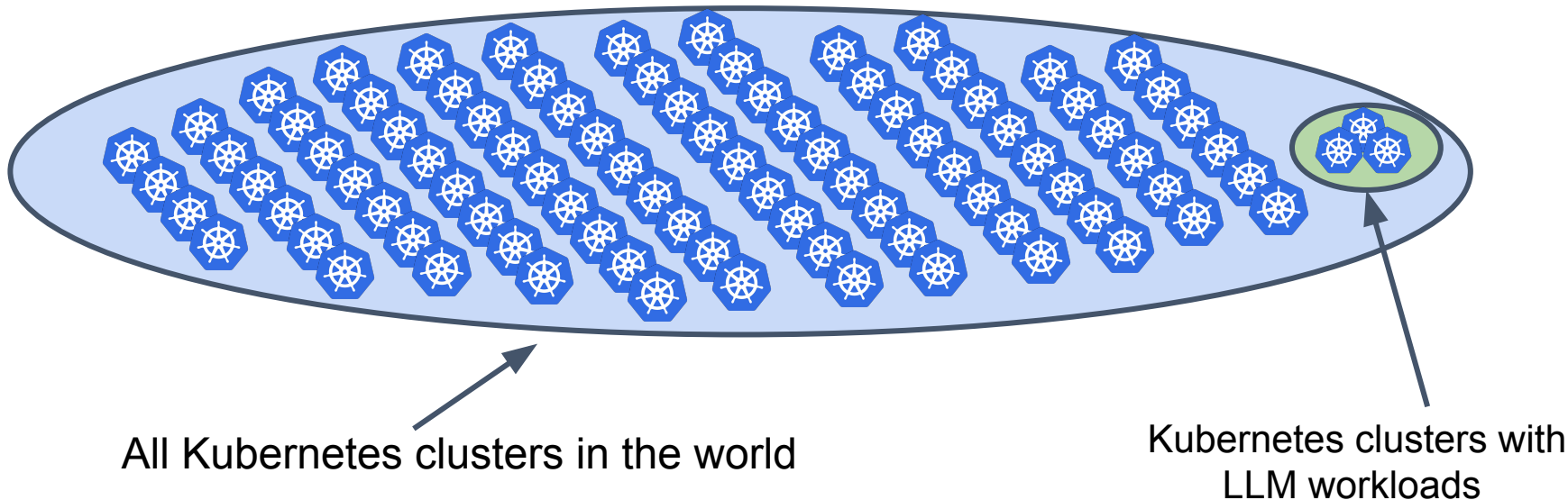
<https://llmariner.ai/>

Why LLMariner?



CLOUD NATIVE &
KUBERNETES
AI DAY
NORTH AMERICA

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

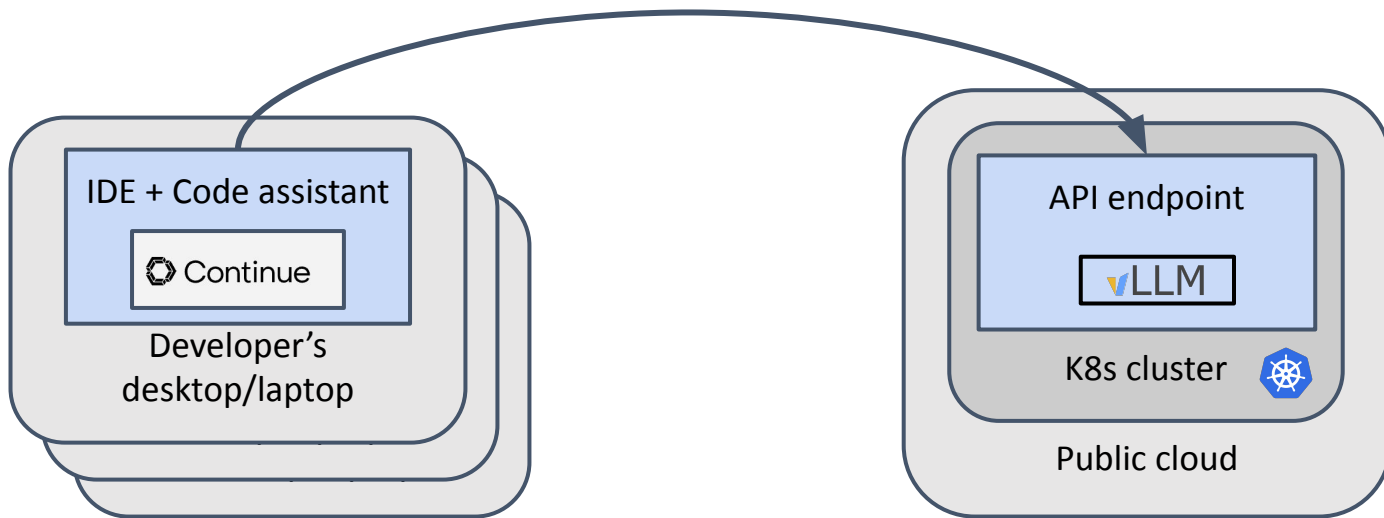
- Difficulty in satisfying the enterprise requirements on security, reliability, and efficiency

A Real-World Scenario



CLOUD NATIVE &
KUBERNETES
AI DAY
NORTH AMERICA

- 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

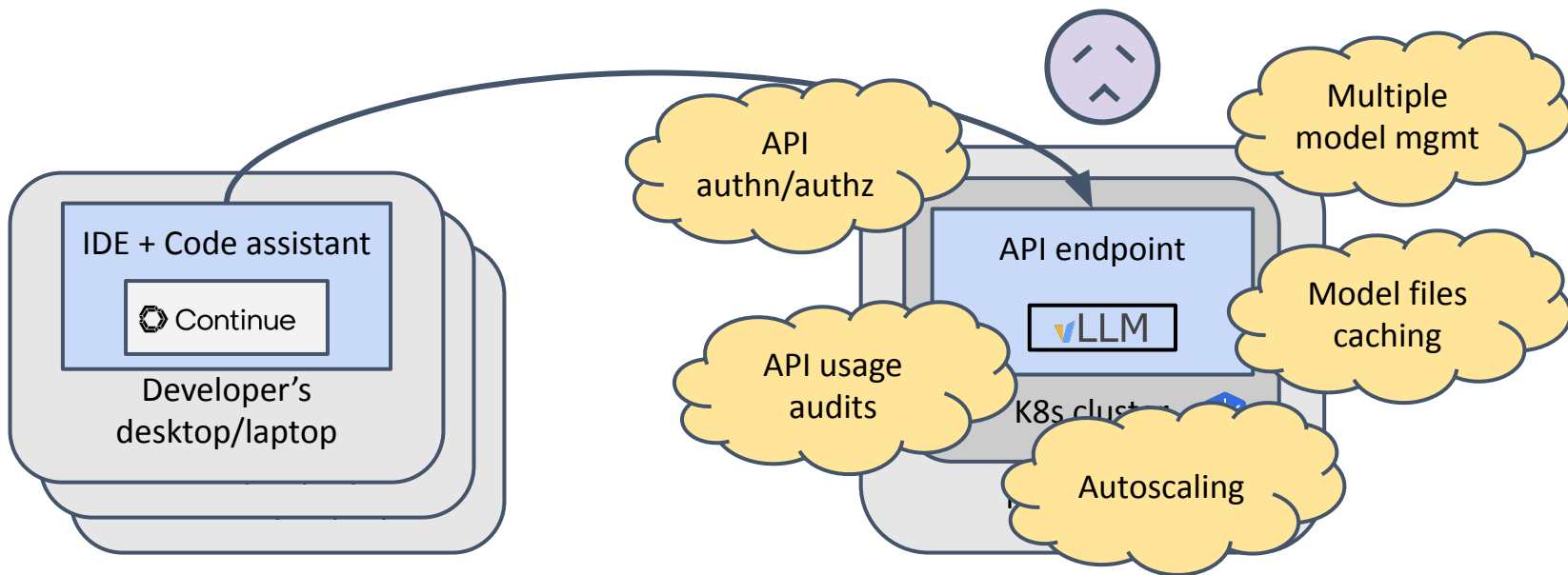


Enterprise Requirements Add Complexity



CLOUD NATIVE &
KUBERNETES
AI DAY
NORTH AMERICA

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

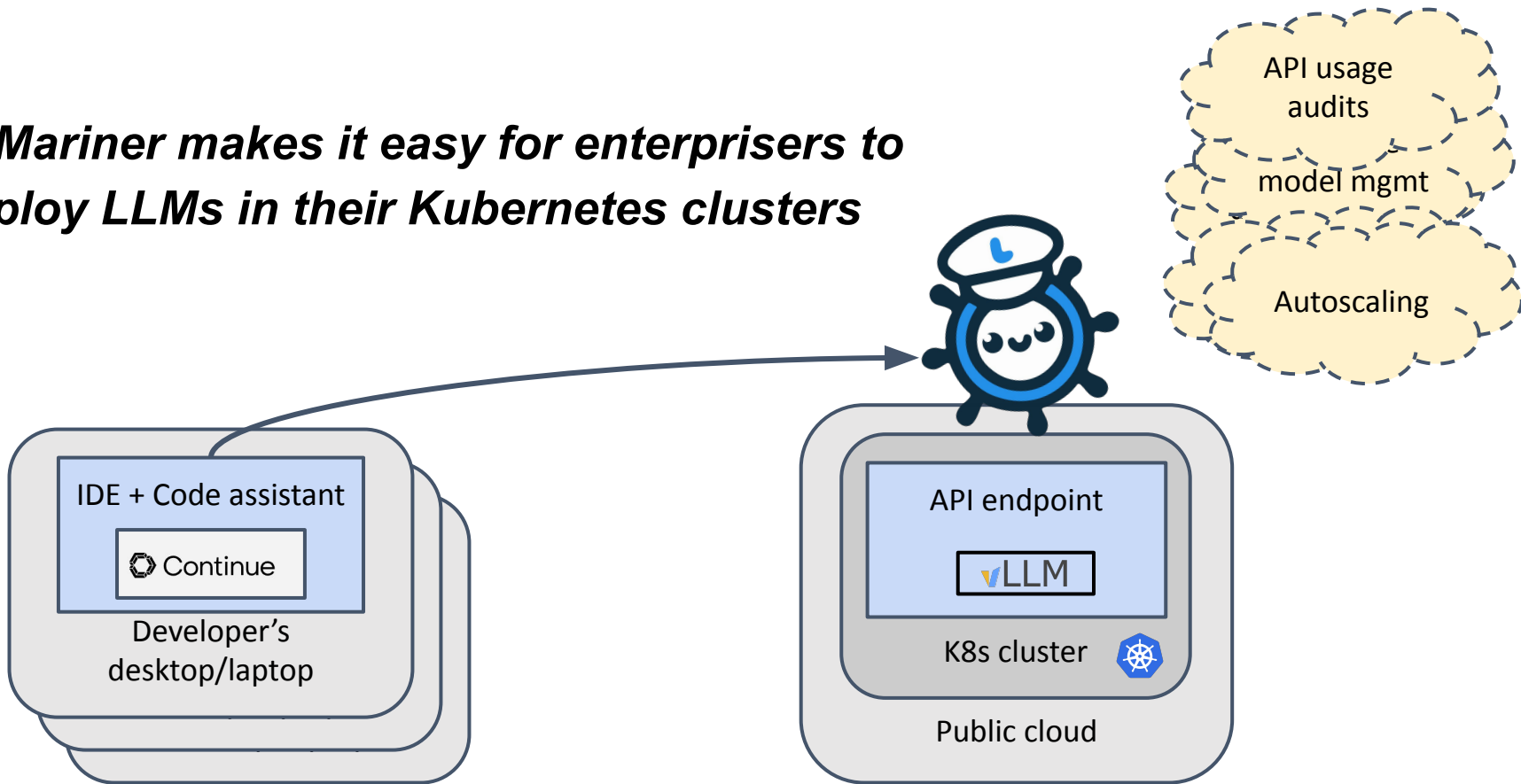


Breaking Barriers with LLMariner



CLOUD NATIVE &
KUBERNETES
AI DAY
NORTH AMERICA

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



Demo - Coding Assistant with LLMariner



CLOUD NATIVE &
KUBERNETES
AI DAY
NORTH AMERICA

```
ubuntu@ip-172-31-17-90: ~ (-zsh)
ID                               Owned By   Created At
TinyLlama-TinyLlama-1.1B-Chat-v1.0 system    2024-10-09T11:10:46-07:00
deepseek-ai-deepseek-coder-6.7b-base-awq system    2024-10-03T14:05:29-07:00
ft: TinyLlama-TinyLlama-1.1B-Chat-v1.0: fine-tuning-gXUeq8HVvA user      2024-10-09T16:23:20-07:00
ft: TinyLlama-TinyLlama-1.1B-Chat-v1.0: fine-tuning-roFXhizndA user      2024-10-09T11:44:49-07:00
ft: google-gemma-2b-it: fine-tuning-BN2TAF-WGA user      2024-10-09T21:37:22-07:00
google-gemma-2b-it              system    2024-10-09T21:30:11-07:00
google-gemma-2b-it-q4_0         system    2024-10-03T14:06:16-07:00
intfloat-e5-mistral-7b-instruct system    2024-10-03T14:18:44-07:00
meta-llama-Meta-Llama-3.1-70B-Instruct-awq system    2024-10-03T14:36:18-07:00
meta-llama-Meta-Llama-3.1-70B-Instruct-awq-triton system    2024-10-08T11:42:30-07:00
meta-llama-Meta-Llama-3.1-8B-Instruct-q4_0 system    2024-10-03T14:38:37-07:00
nvidia-llama-3.1-Nemotron-70B-Instruct-fp8-dynamic system    2024-10-17T10:12:03-07:00
nvidia-llama-3.1-Nemotron-70B-Instruct-q2_k system    2024-10-16T23:28:10-07:00
nvidia-llama-3.1-Nemotron-70B-Instruct-q4_0 system    2024-10-22T11:34:51-07:00
sentence-transformers-all-MiniLM-L6-v2-f16 system    2024-10-03T14:38:40-07:00

Control plane cluster> llama auth api-keys create demo
Created the API key. Secret: sk-ytCrzkDu3N_THXILtNbqj33m281boodNw_JanRReMy6LWNHsd
Control plane cluster>

Worker GPU cluster> kubectl get pods -n llmariner
NAME                                READY   STATUS    RESTARTS   AGE
inference-manager-engine-7d5dcc86c9-9ddgg 1/1     Running   0           125m
inference-manager-engine-7d5dcc86c9-s16jc 1/1     Running   0           127m
job-manager-dispatcher-6f75bfbb9-2zbbs 1/1     Running   0           144m
model-manager-loader-5d9576f97-zkxmt 1/1     Running   0           142m
ollama-google-gemma-2b-it-q4-0-0 1/1     Running   0           141m
ollama-sentence-transformers-all-minilm-l6-v2-f16-0 1/1     Running   0           127m
session-manager-agent-bbdfcb94b-282rf 2/2     Running   0           13m
vllm-a-llama-3.1-nemotron-70b-instruct-fp8-dynamic-0 1/1     Running   0           124m
vllm-deepseek-ai-deepseek-coder-6.7b-base-awq-0 1/1     Running   0           114m
vllm-meta-llama-meta-llama-3.1-8b-instruct-q4-0-0 1/1     Running   0           112m

Worker GPU cluster> kubectl logs -n llmariner
```

```
le, and more {} config.json task_exchanger.go x ...
server > internal > taskexchanger > task_exchanger.go > ...
1 package taskexchanger
2
3 import (
4     "context"
5     "fmt"
6     "sync"
7
8     corev1 "k8s.io/api/core/v1"
9
10    "github.com/go-logr/logr"
11
12    v1 "github.com/llmariner/inference-manager/se
13    "github.com/llmariner/inference-manager/se
14    apierrors "k8s.io/apimachinery/pkg/api/err
15    ctrl "sigs.k8s.io/controller-runtime"
16    "sigs.k8s.io/controller-runtime/pkg/builde
17    k8sclient "sigs.k8s.io/controller-runtime/pkg/client"
18    "sigs.k8s.io/controller-runtime/pkg/predic
19    "sigs.k8s.io/controller-runtime/pkg/reconc
20
21    // NewE creates a new E.
22    func NewE(
23        infProcessor *infprocessor.P,
24        k8sClient k8sclient.Client,
25        gRPCPort int,
26        localPodName string,
27        podLabelKey string,
28        podLabelValue string,
29        logger logr.Logger,
30    ) *E {
31        // ...
32    }
33
34 func (e *E) RemoveServer(serverPodName string) {
35     // ...
36     delete(e.taskSenders, serverPodName)
37 }
38
39 ...
40
41 what are all functions in this file?
```

```
task_exchanger.go (1-236) x
package taskexchanger

import (
    "context"
    "fmt"
    "sync"

    corev1 "k8s.io/api/core/v1"

    "github.com/go-logr/logr"
    v1 "github.com/llmariner/inference-manager/api/v1"
    "github.com/llmariner/inference-manager/api/v1"

    what are all functions in this file?
    Llama3.1.N... @
```

```
PROBLEMS OUTPUT ... Continue - LLM Pro
func (e *E) RemoveServer(serverPodName string) {
    // ...
    delete(e.taskSenders, serverPodName)
}

...

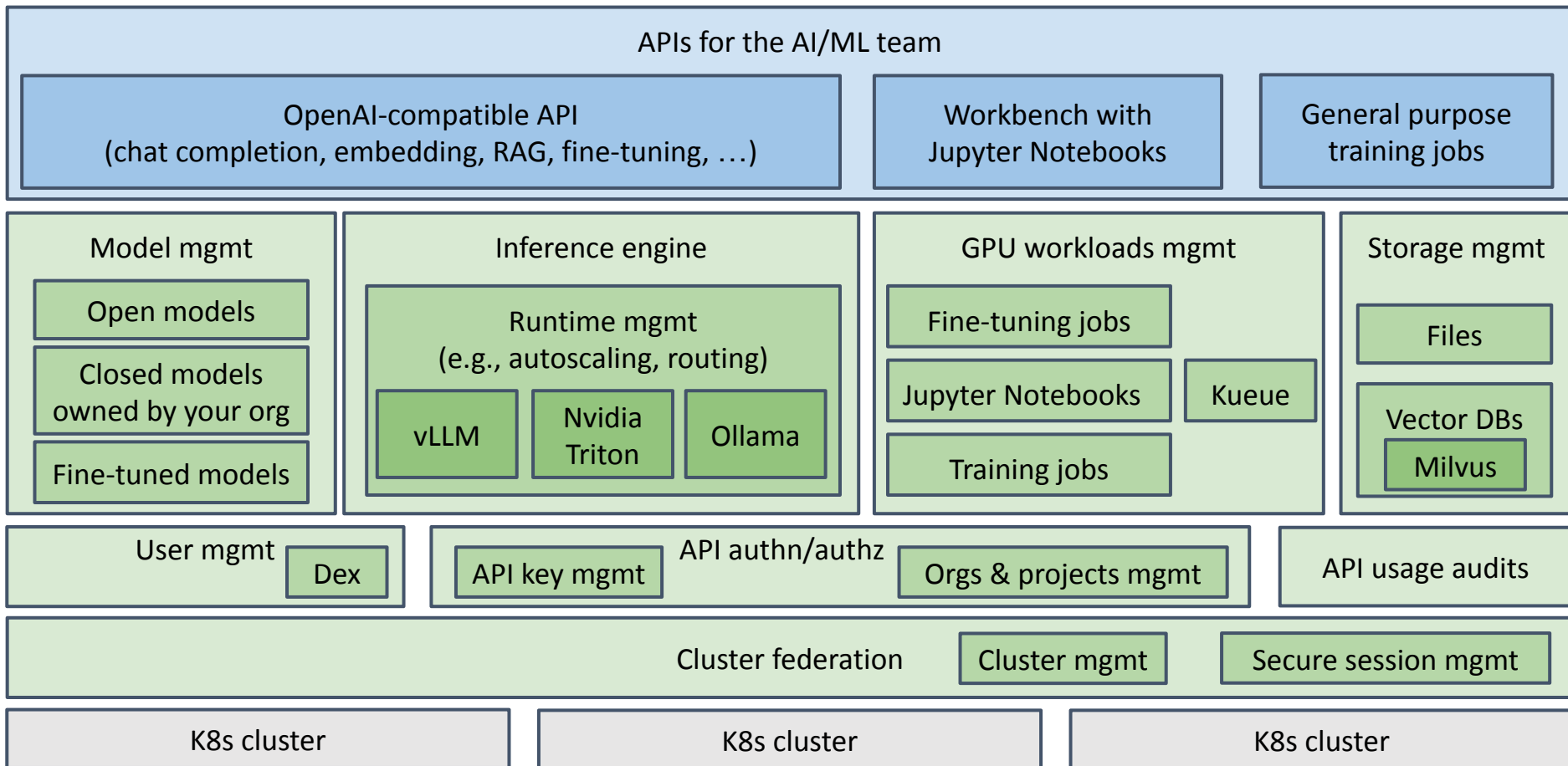
what are all functions in this file?
```

Link: <https://vimeo.com/1024806457>

LLMariner Features for AI/ML and Infra Teams

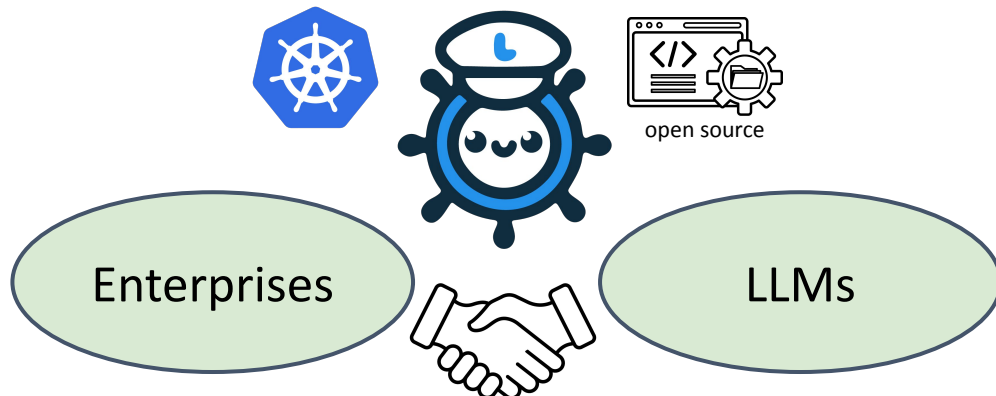


CLOUD NATIVE &
KUBERNETES
AI DAY
NORTH AMERICA



LLMariner opens various opportunities for enterprises

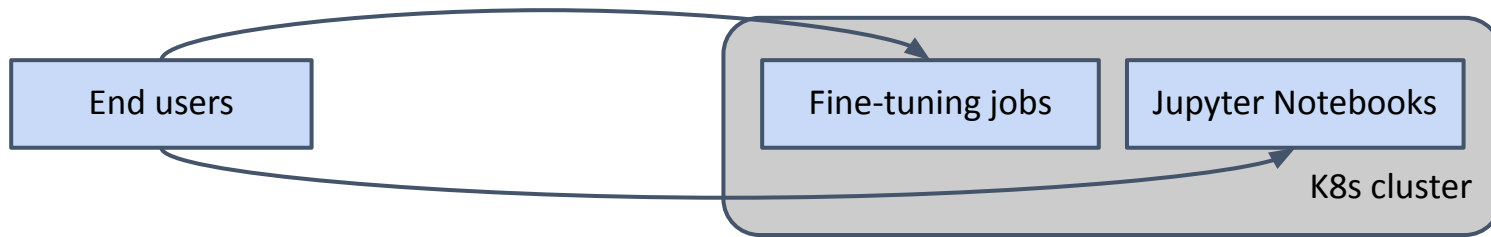
- Provides **full control** by hosting LLMs on your infrastructure
- Enables teams to catch up with the rapid AI/ML evolution by providing a foundation for **open source technologies**



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?

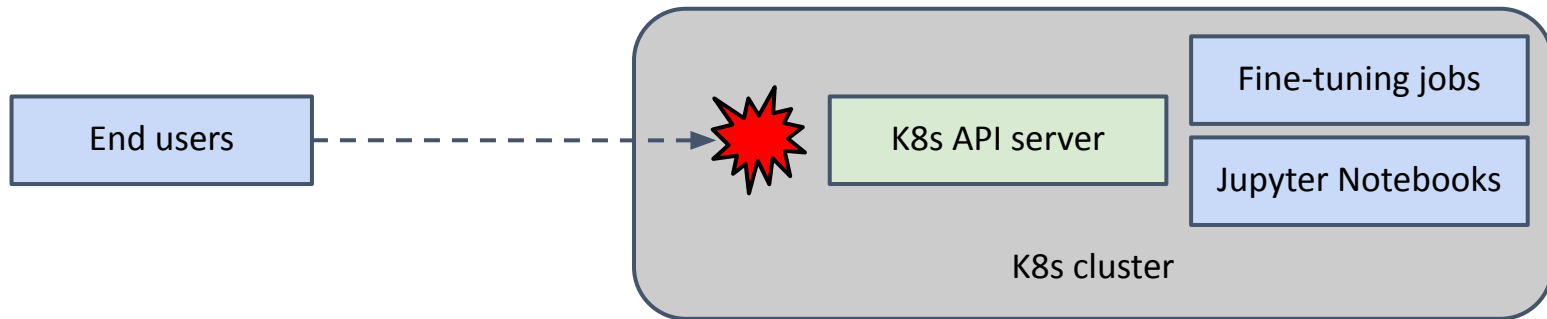
- Provide direct access to the K8s API server (and use CRDs to track fine-tuning jobs and Jupyter Notebooks)?

Potential Issues with Direct Access to K8s API Server and CRDs



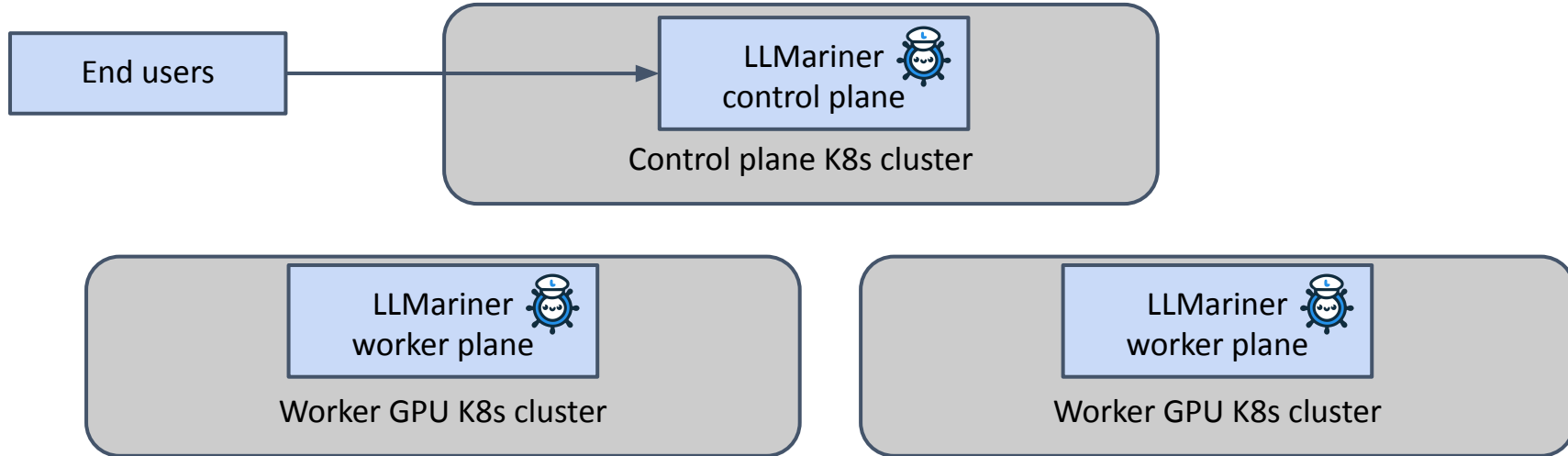
CLOUD NATIVE &
KUBERNETES
AI DAY
NORTH AMERICA

- 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

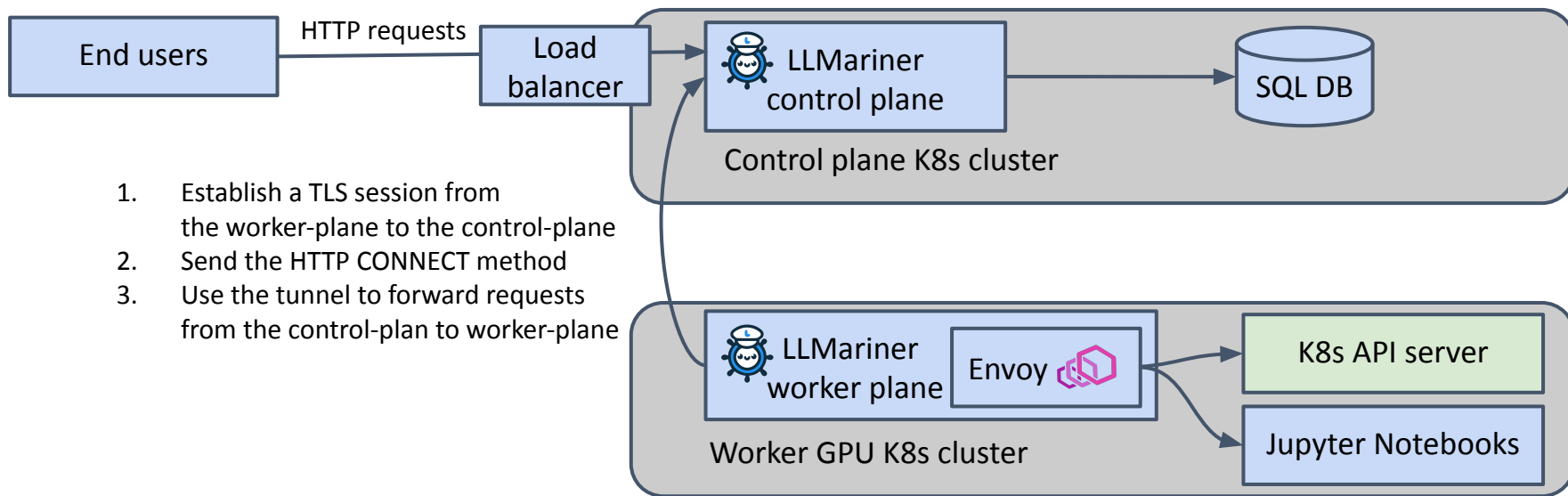


Our Approach on Fine-tuning Jobs and Jupyter Notebooks Management



CLOUD NATIVE &
KUBERNETES
AI DAY
NORTH AMERICA

- 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!



CLOUD NATIVE &
KUBERNETES
AI DAY
NORTH AMERICA

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

