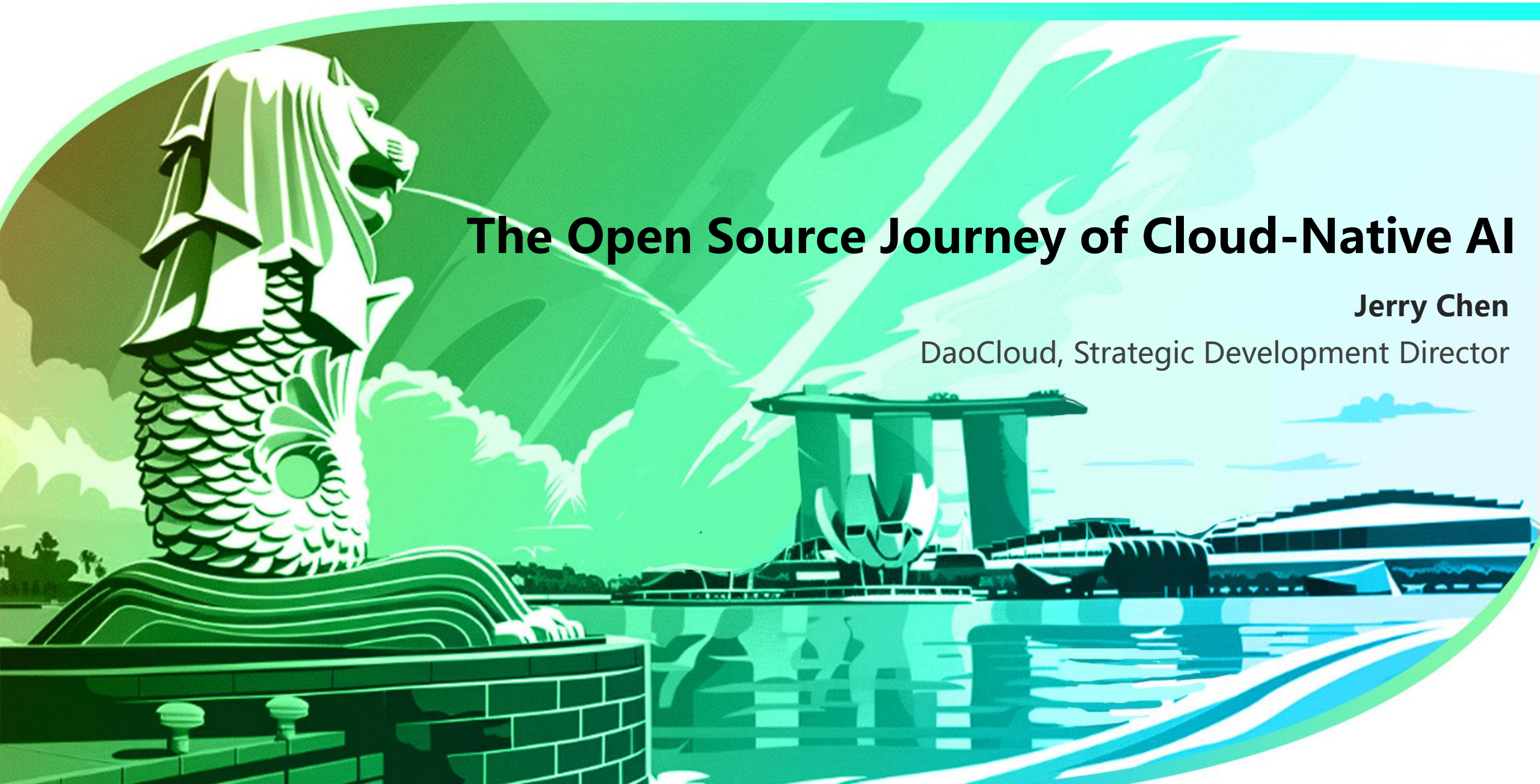
DaoCloud BYTEDANCE OPEN SOURCE Shopee WasmEdgeRuntime

Cloud Native AI Meetup

10th July, 2024

18:30-20:30 | Singapore



The Open Source Journey of Cloud-Native AI

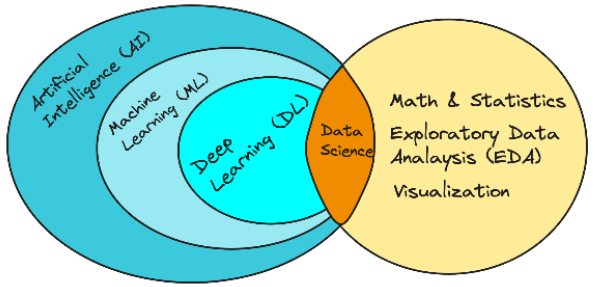
Jerry Chen

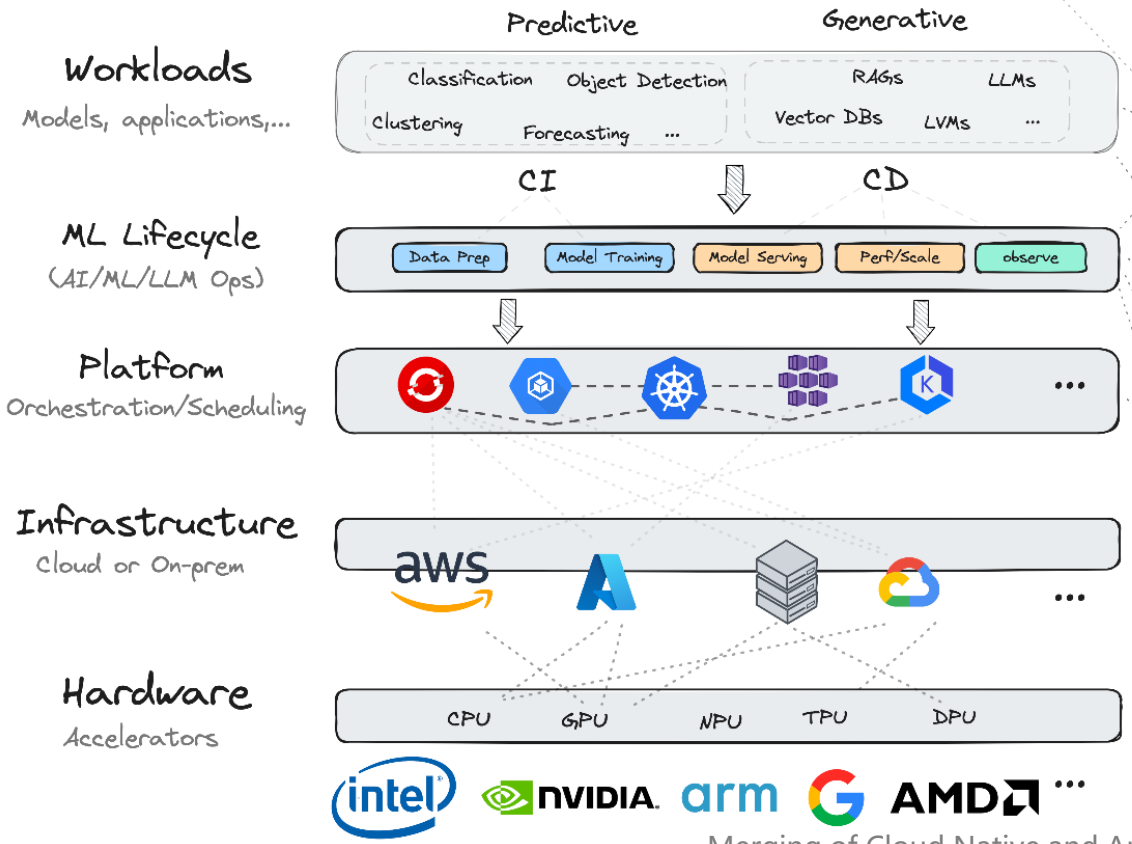
DaoCloud, Strategic Development Director

Content

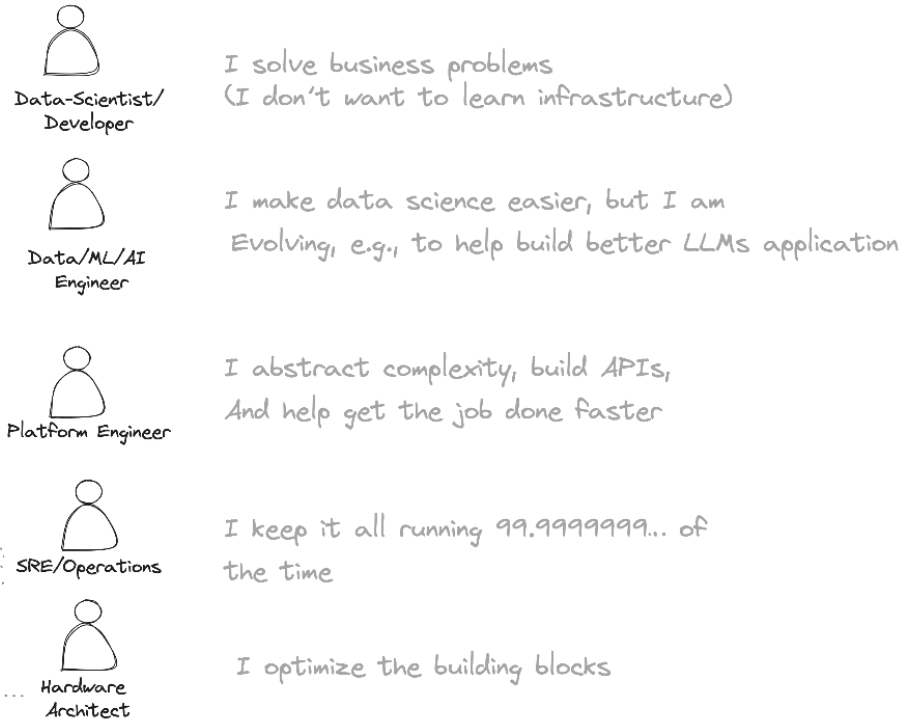
Part 01
Introduction to Cloud-Native AI and DaoCloud

Relationship between Cloud Native and Cloud Native AI





- AI > ML > Deep Learning. AI is the application of ML and deep learning to simulate human intelligence and problem-solving capabilities.
- The ML lifecycle involves a collaborative effort among various disciplines and roles to ensure ML models' successful development, deployment, and maintenance.



Merging of Cloud Native and Artificial Intelligence

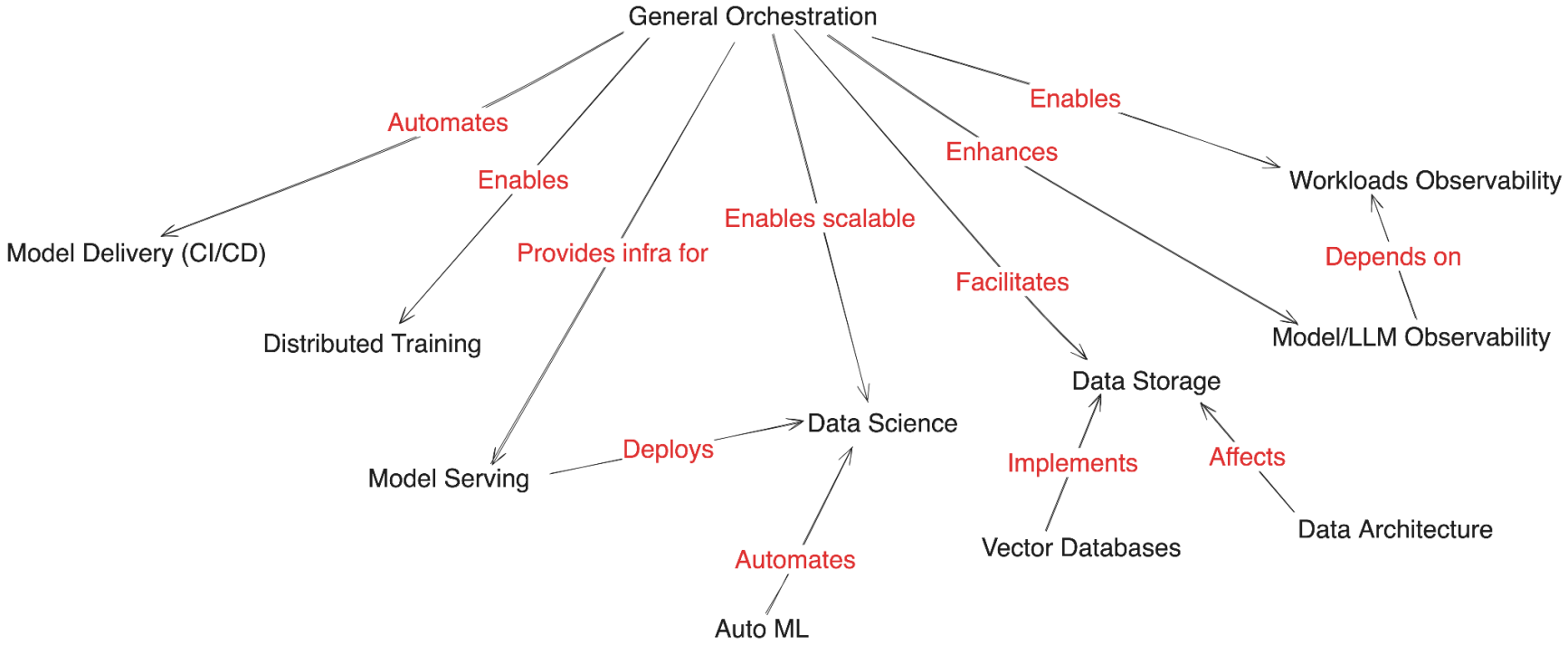
Challenges For AI

A sample set of areas where predictive and generative AI have distinct needs across computing, networking, and storage.

Challenges/Need	Generative AI	Predictive AI
Computational Power	Extremely high. Requires specialized hardware.	Moderate to high. General-purpose hardware can suffice.
Data Volume and Diversity	Massive, diverse datasets for training.	Specific historical data for prediction.
Model Training and Fine-tuning	Complex, iterative training with specialized compute.	Moderate training.
Scalability and Elasticity	Highly scalable and elastic infrastructure (variable and intensive computational demands)	Scalability is necessary but lower elasticity demands. Batch processing or event-driven tasks.
Storage and Throughput	High-performance storage with excellent throughput. Diverse data types. Requires high throughput and low-latency access to data.	Efficient storage with moderate throughput. It focuses more on data analysis and less on data generation; data is mostly structured.
Networking	High bandwidth and low latency for data transfer and model synchronization (e.g., during distributed training).	Consistent and reliable connectivity for data access.

What is Cloud Native AI?

- Cloud Native Artificial Intelligence (CNAI) refers to approaches and patterns for building and deploying AI applications and workloads using the principles of Cloud Native. Enabling repeatable and scalable AI-focused workflows allows AI practitioners to focus on their domain.
- CNAI solutions address challenges AI application scientists, developers, and deployers face in developing, deploying, running, scaling, and monitoring AI workloads on cloud infrastructure.
- By leveraging the underlying cloud infrastructure’ s computing (e.g., CPUs and GPUs), network, and storage capabilities, as well as providing isolation and controlled sharing mechanisms, it accelerates AI application performance and reduces costs.



DaoCloud's Mission in Cloud-Native AI

- **Focus:** Founded at the end of 2014. An innovative leader in cloud-native and application modernization. Dedicated to developing an open Cloud OS that empowers enterprises to seamlessly achieve digital transformation.
- **Mission:** Empowering businesses with innovative cloud-native solutions and driving advancements in AI through open-source technologies.
- **Vision:** To be a global leader in cloud-native AI, transforming industries with cutting-edge, scalable, and sustainable solutions.

#3 Kubernetes	#2 ContainerD	#2 Istio	#1 Kubespray
#2 Karmada	#5 Docker	#8 Prometheus	#6 Helm
#4 CoreDNS	#2 KubeEdge	#1 Kubeadm *	#2 Metal LB
#3 CNI *	#9 Harbor	#14 etcd *	#20 Cilium *

CNCF ambassadors


Roby Chen


Iceber Gu

Kubernetes & Istio steering committees


Paco Xu


Kebe Liu

Cloud-Native AI Working Group


Michael Yao

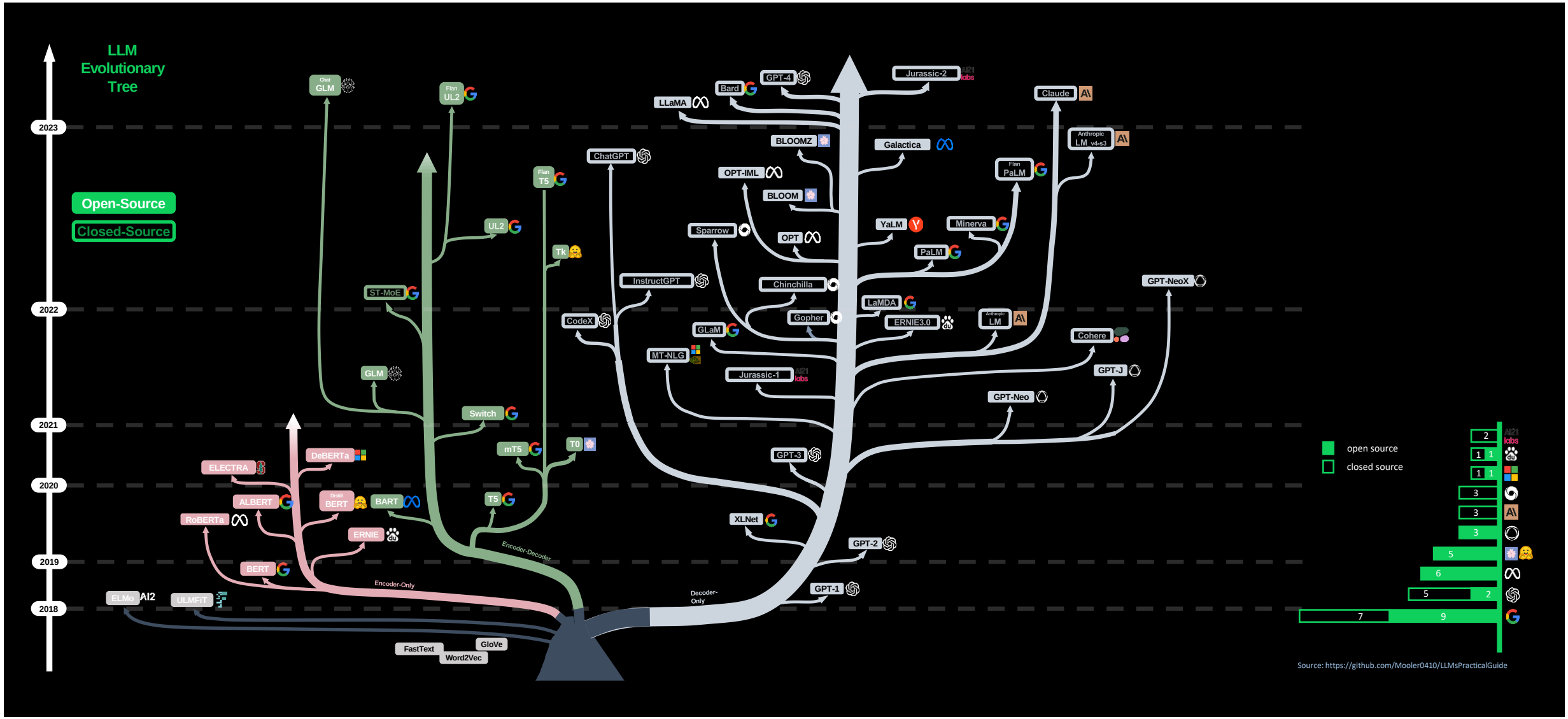

Peter Pan

LINUX FOUNDATION AUTHORIZED TRAINING PARTNER

eBPF LF AI & DATA OpenTelemetry

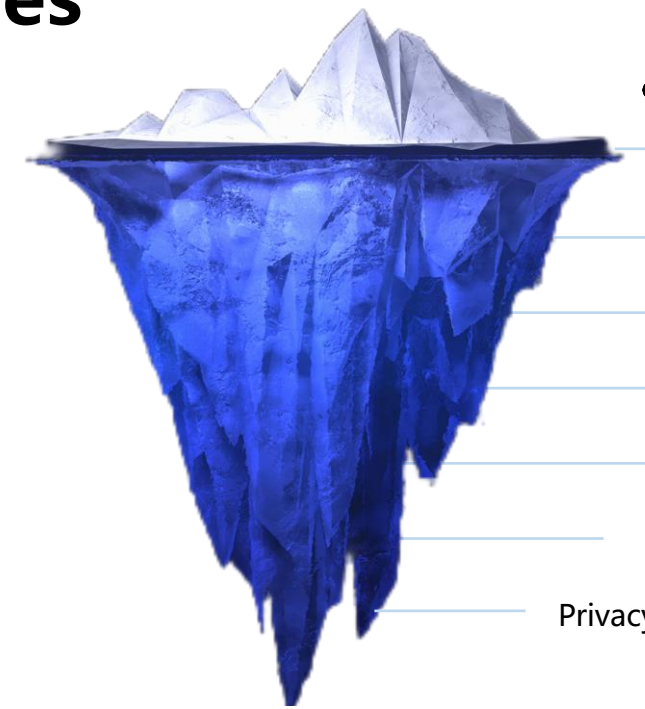
Part 02
Overcoming Challenges with Open Source Innovations

Breakthrough in GenAI, rapid development of LM



AGI tech advances rapidly, industry implementation faces numerous challenges

- GPT-3 175B has 175 billion model parameters and costs about US$1.4 million to train once.
- For larger LLM models, training costs range from $2 million to $12 million
- ChatGPT-175B requires approximately 375-625 8-card A100 servers for training. If you are willing to wait a month, 150-200 8-card units are enough. The total GPU resource consumption per training is 35,000 card days



Large model scale, high computational complexity

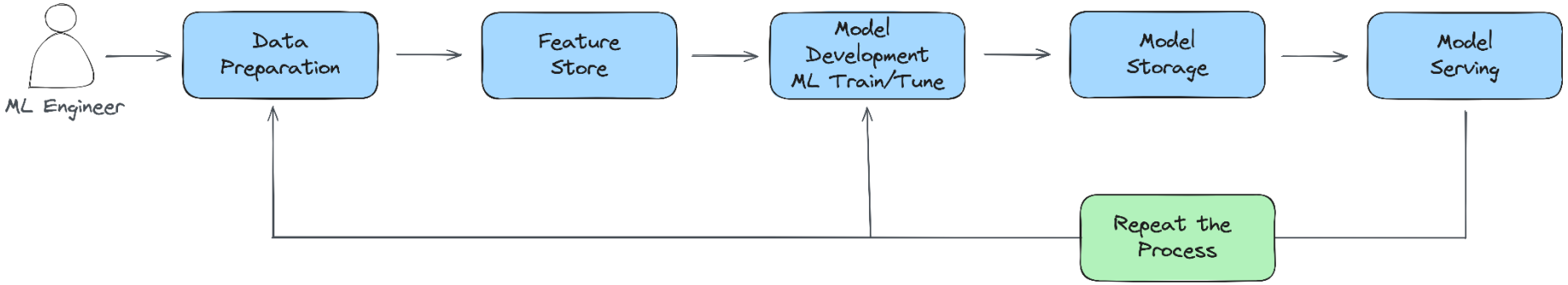
Requires managing heterogeneous devices

Resource allocation and task scheduling

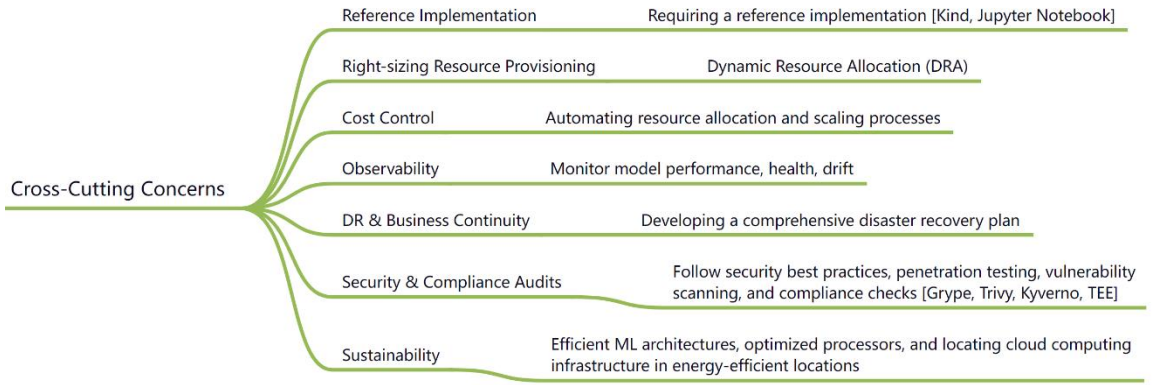
Specialized workforce and long-term cost investment

Network and Data Storage Bottlenecks

Privacy security, data risks



Challenges For Cloud Native AI



DaoCloud's Approach to Cloud-Native AI Challenges

- Computing Power Scheduling + GPU Virtualization - **HAMi**
- Platform Engineering & Scalability
- Heterogeneous Architecture Unified Management

Computing Power Resource Pool

High-Performance Network

- Kubernetes with RDMA
- Infiniband/RoCE with SR-IOV
- RoCE with macvlan/iplan
- Infiniband with overlay
- Infiniband with IPoIB
- **Spiderpool** Enhanced RDMA
- Flexible Networking

Key Technology

- Distributed Architecture - Ray
- Distributed Scheduling - **Kueue**
- K8s Without Kubelet - **KWOK**

Computational Task Scheduling

Data Storage Acceleration

- **HwameiStor** - Local Storage
- Distributed Cache Acceleration

Cloud-Native intelligent computing infrastructure capabilities

Compute efficiency optimization capabilities

AI task scheduling capabilities

Model storage capabilities

Compute network bottlenecks solution

Machine learning engineering efficiency optimization

Integration capabilities of large models with business

GPU scheduling management capabilities

Platform monitoring observability capabilities

Part 03
Highlighting Key Open-Source Projects by DaoCloud

Overview of Key Projects

Kueue
Job Queueing Controller

KWOK
Large-scale Cluster Simulation

Clusterpedia
Multi-cluster Resources Search

Kubean
Cluster Lifecycle Management

HwameiStor
Local Storage Acceleration

HAMi
vGPU Management

Spiderpool
High-speed Network

DataTunerX
LLM Fine-tuning Solution

DaoCloud has 100+ developers and has initiated a total of 15 CNCF open-source projects, including 6 CNCF Sandbox projects, 8 CNCF Landscape projects, and 1 eBPF Landscape project.
Note: As of June 2024, the CNCF Foundation has incubated a total of 179 projects worldwide.

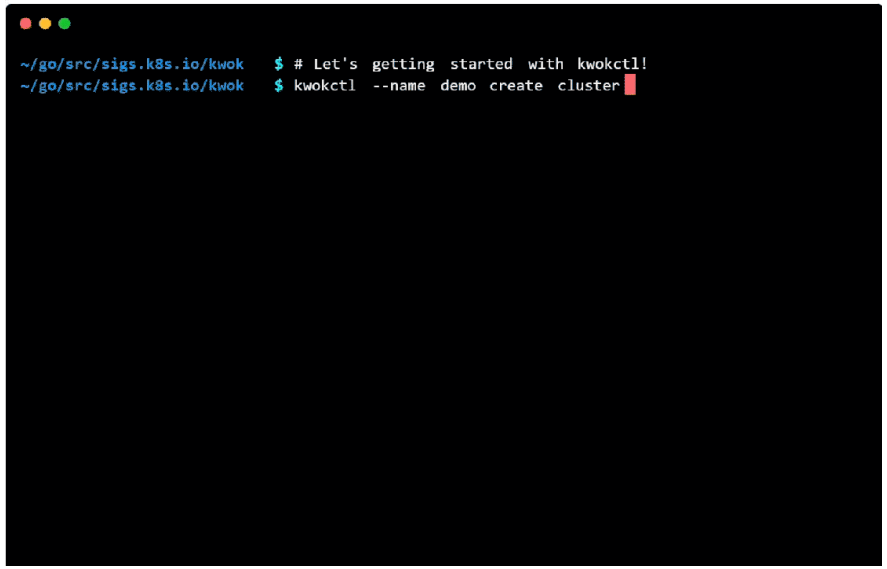
CNCF Landscape Project - KWOK

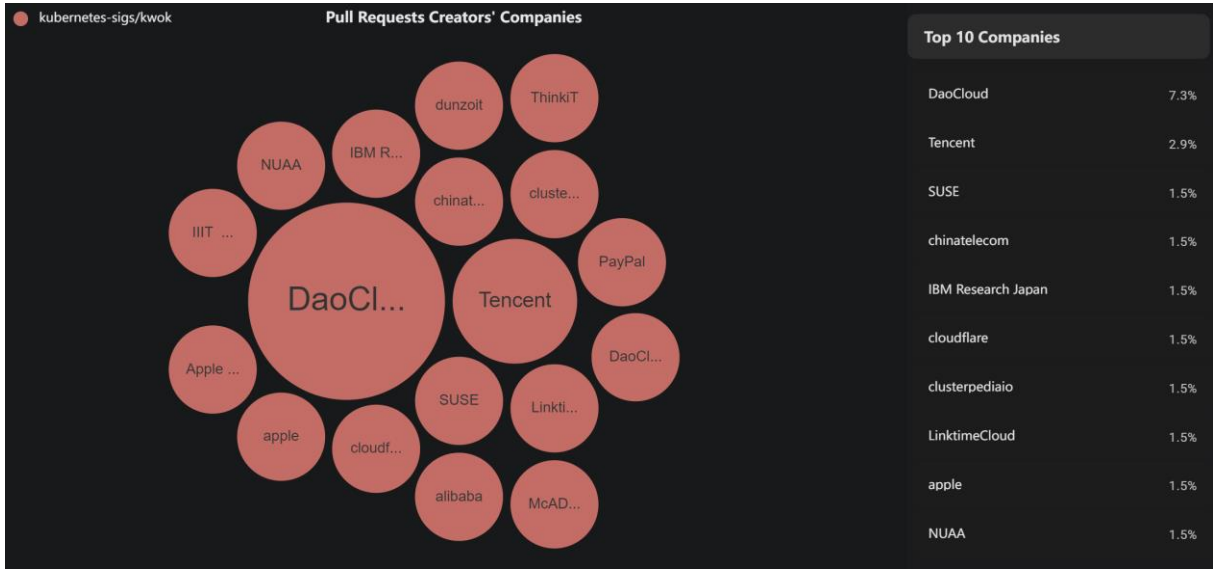
Intro: KWOK (Kubernetes WithOut Kubelet) is an open source project independently developed by DaoCloud and managed by the Kubernetes-SIG community. KWOK can build a cluster of thousands of nodes in seconds. In this scenario, all simulated nodes behave like real nodes.

Repo: https://github.com/kubernetes-sigs/kwok

Trend: Stars 2308, Commits 1442, Contributors 61

Adoption: Kubernetes and Karmada have integrated KWOK for end-to-end testing. KWOK is widely adopted by companies such as Tencent, Huawei, Red Hat, ByteDance, Microsoft, IBM, and VMware. For example, Tencent's Xingchen Computing Power Platform uses KWOK to set up large-scale cluster simulation environments, and Azure Container Networking uses KWOK for large-scale cluster testing.





CNCF Landscape Project - Kueue

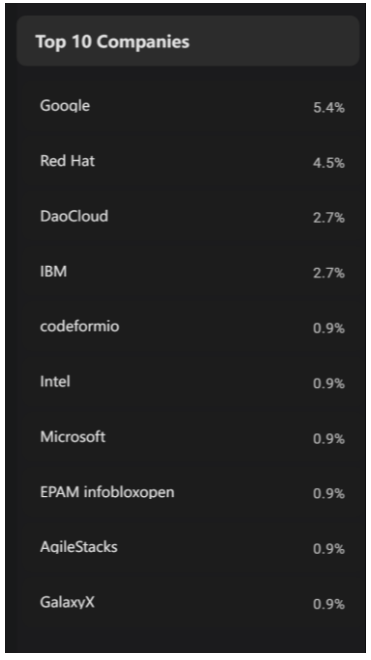
Intro: Cloud-native job queueing system for batch, HPC, AI/ML, and similar applications in a Kubernetes cluster. Implement fair scheduling, affinity, group scheduling, compact scheduling and other scheduling algorithms to cope with different computing power scenarios.

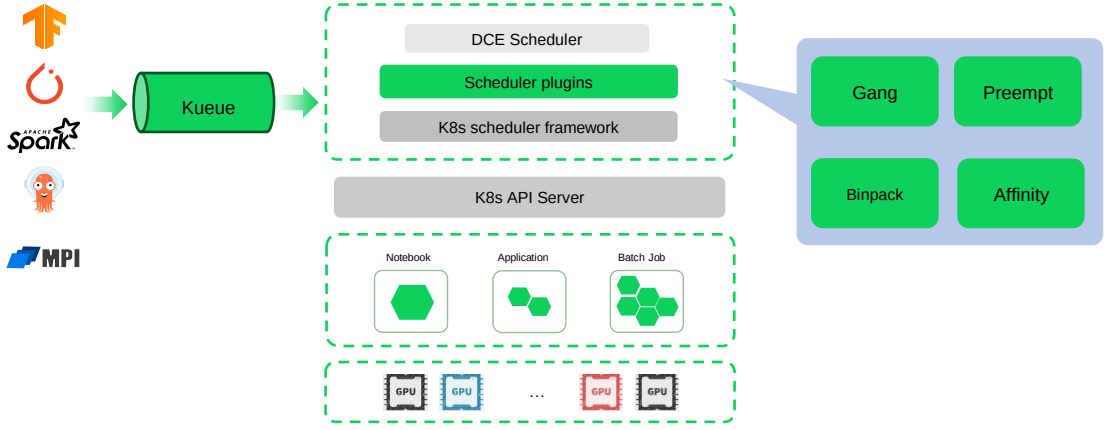
Repo: https://github.com/kubernetes-sigs/kueue

Trend: Stars 1178, Commits 2299, Contributors 108

Adoption: Kueue is a project co-developed in depth by DaoCloud and Google teams. This collaboration leverages the expertise and innovation of both organizations to advance the capabilities of Kubernetes job scheduling. Kueue is widely adopted by companies such as DaoCloud, Shopee, Google Cloud.

Organization	Type	Description	Integrations
CyberAgent, Inc.	End User	On-premise ML Platform	batch/job kubeflow.org/mpijob
DaoCloud, Inc.	End User	Part of the AI Platform for managing all kinds of Jobs.	batch/job RayJob ...
WattIQ, Inc.	End User	SaaS/IoT product	batch/job RayJob
Horizon, Inc.	End User	AI training platform	batch/job ...
FAR AI	End User	AI alignment research nonprofit	batch/job
Shopee, Inc.	End User	Training/batch inference/data processes in AI platform test env	Customized job RayJob ...
Mondoo, Inc.	End User	Helps power Mondoo's hosted security scanner	batch/job
Google Cloud	Provider	Part of kit for training ML workloads on TPUs	JobSet
Onna Technologies, Inc	End User	Unstructured Data Management Platform	batch/job





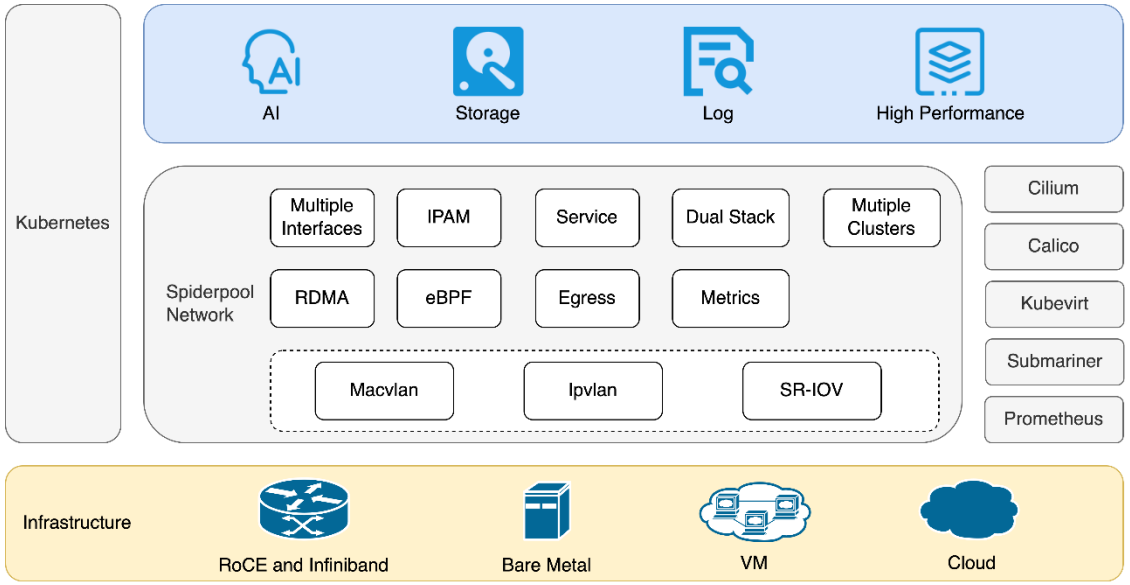
CNCF Sandbox Project - Spiderpool

Intro: The underlay and RDMA network solution of the Kubernetes, for bare metal, VM and public cloud. It enhances the capabilities of Macvlan CNI, ipvlan CNI, SR-IOV CNI, fulfills various networking needs. Spiderpool delivers exceptional network performance, particularly benefiting network I/O-intensive and low-latency applications like storage, middleware, and AI.

Repo: https://github.com/spidernet-io/spiderpool

Trend: Stars 533, Commits 5335, Contributors 34

Adoption: Spiderpool is widely adopted by companies such as China Mobile, vivo. Spiderpool solves vivo's pain points and is implemented in its AI production cluster. A Chinese securities uses it for ultra-fast trading applications.



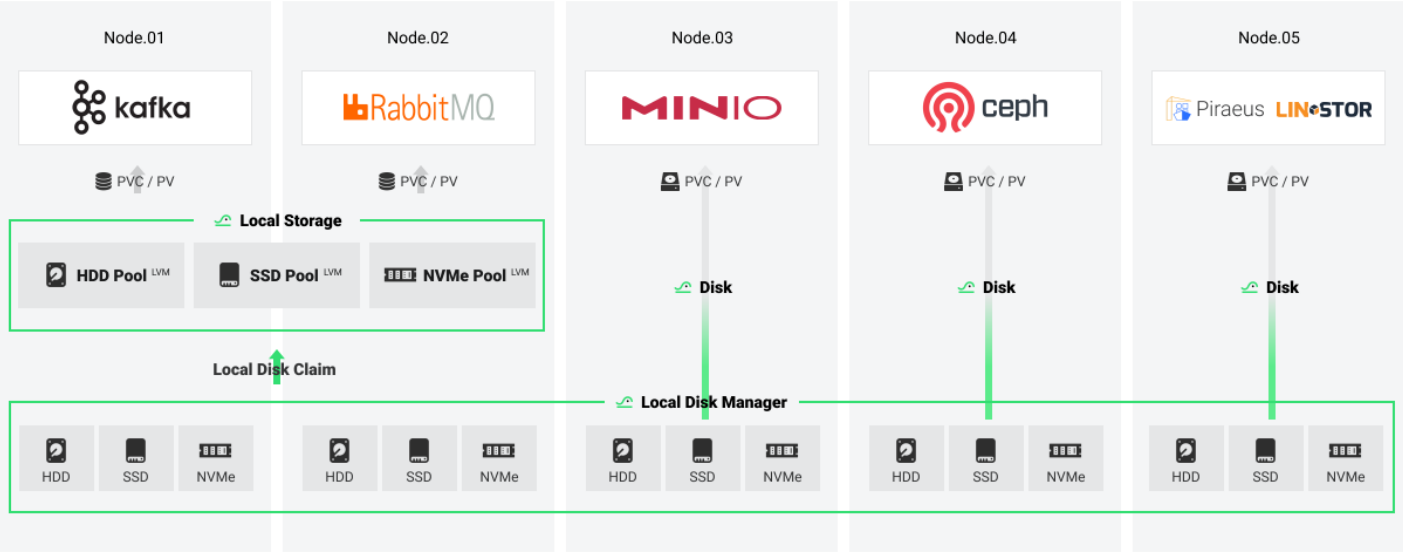
CNCF Sandbox Project - HwameiStor

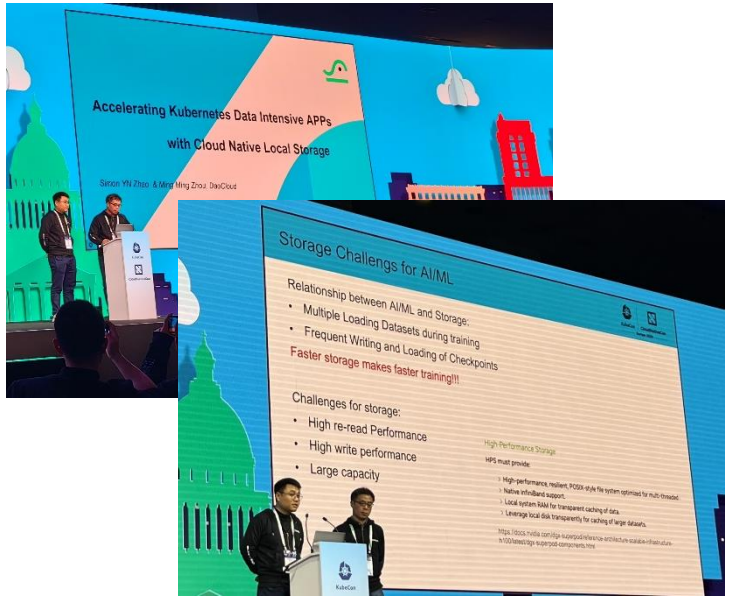
Intro: Hwameistor is an HA local storage system for cloud-native stateful workloads. HwameiStor creates a local storage resource pool for centrally managing all disks such as HDD, SSD, and NVMe. It uses the CSI to provide distributed services with local volumes. It is lightweight, and cost-efficient that can replace expensive traditional SAN storage.

Repo: https://github.com/hwameistor/hwameistor

Trend: Stars 530, Commits 2165, Contributors 41

Adoption: Hwameistor has been widely used in DaoCloud's customer projects.





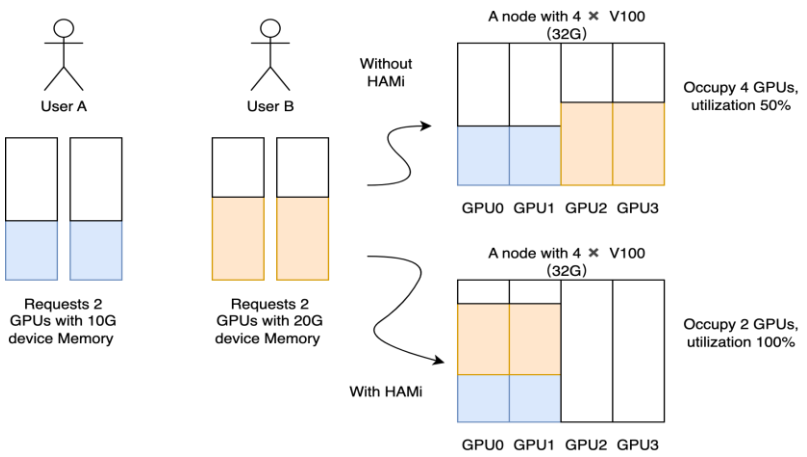
CNCF Landscape Project - HAMi

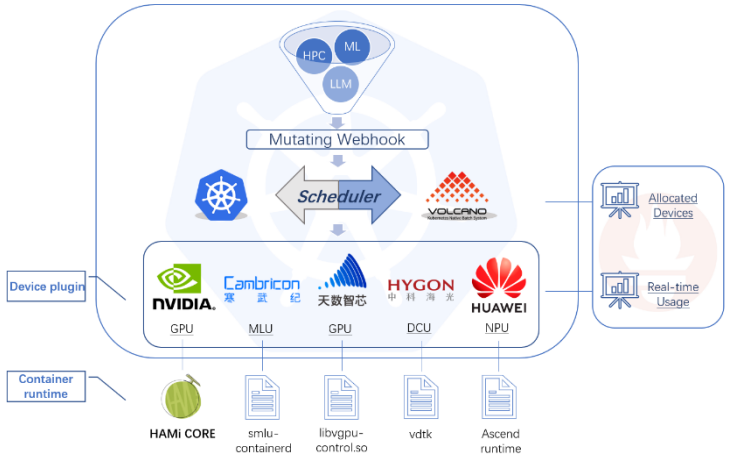
Intro: Heterogeneous AI Computing Virtualization Middleware (HAMi), is an "all-in-one" tool designed to manage Heterogeneous AI Computing Devices in Kubernetes cluster. Scenarios: Device sharing (or device virtualization) on Kubernetes, pods need to be allocated with specific device memory usage or device cores. Need to balance GPU usage in a cluster with multiple GPU nodes. Low utilization of device memory and computing units, such as running 10 TensorFlow servings on one GPU. Situations that require a large number of small GPUs.

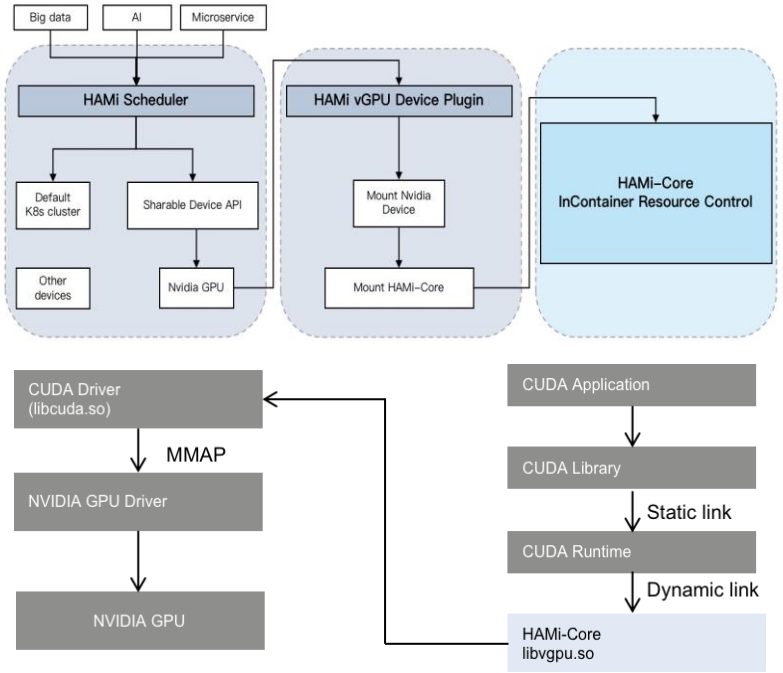
Repo: https://github.com/Project-HAMi/HAMi

Trend: Stars 455, Commits 724, Contributors 22

Adoption: HAMi was co-founded by 4paradigm and DaoCloud. 10K+ downloads, 40+ adopters, aleady support Nvidia, Cambricon, Hygon, Huawei ASCEND.







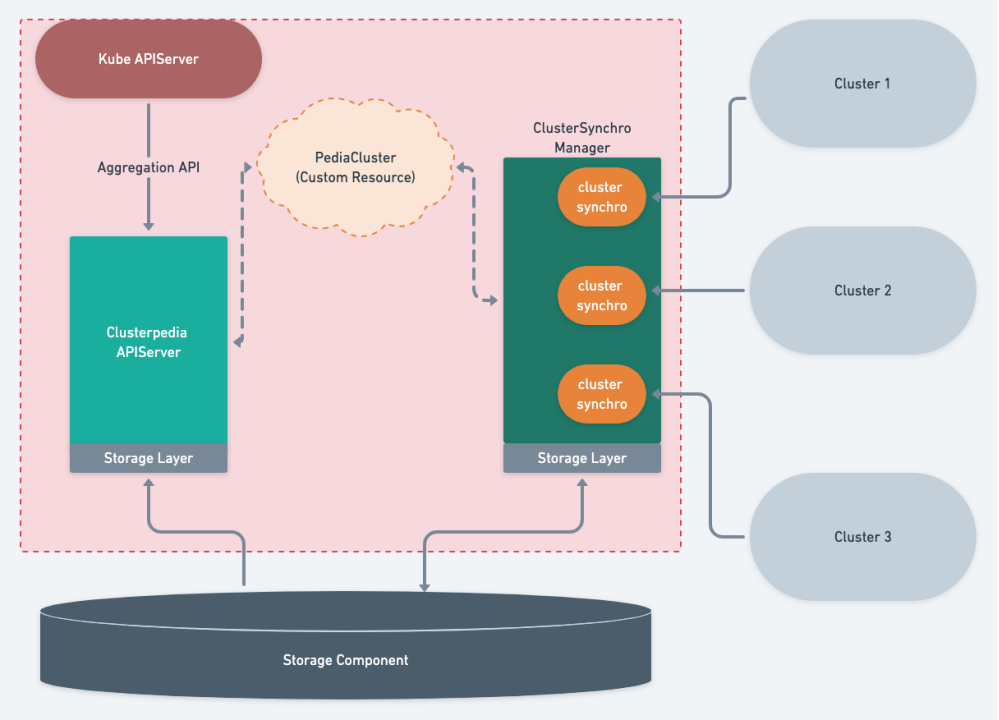
CNCF Sandbox Project - Clusterpedia

Intro: Clusterpedia is used for complex resources search across multiple clusters, support simultaneous search of a single kind of resource or multiple kinds of resources existing in multiple clusters. Clusterpedia can synchronize resources with multiple clusters and provide more powerful search features on the basis of compatibility with Kubernetes OpenAPIs.

Repo: https://github.com/clusterpedia-io/clusterpedia

Trend: Stars 776, Commits 857, Contributors 50

Adoption: China Mobile Cloud uses Clusterpedia to build a multi-Kubernetes cluster management platform.




```
$ kubectl --cluster clusterpedia get deployments -n kube-system
CLUSTER     NAME                      READY   UP-TO-DATE   AVAILABLE   AGE
cluster-1   coredns                   2/2     2            2           68d
cluster-2   calico-kube-controllers   1/1     1            1           64d
cluster-2   coredns                   2/2     2            2           64d
```

```
$ kubectl --cluster cluster-1 get deployments -A
NAMESPACE                          CLUSTER     NAME                                          READY
calico-apiserver                   cluster-1   calico-apiserver                              1/1
calico-system                      cluster-1   calico-kube-controllers                       1/1
calico-system                      cluster-1   calico-typha                                  1/1
capi-system                        cluster-1   capi-controller-manager                       1/1
capi-kubeadm-bootstrap-system      cluster-1   capi-kubeadm-bootstrap-controller-manager     1/1
capi-kubeadm-control-plane-system  cluster-1   capi-kubeadm-control-plane-controller-manager 1/1
capv-system                        cluster-1   capv-controller-manager                       1/1
cert-manager                       cluster-1   cert-manager                                  1/1
cert-manager                       cluster-1   cert-manager-cainjector                       1/1
cert-manager                       cluster-1   cert-manager-webhook                          1/1
clusterpedia-system                cluster-1   clusterpedia-apiserver                        1/1
clusterpedia-system                cluster-1   clusterpedia-clustersynchro-manager           1/1
clusterpedia-system                cluster-1   clusterpedia-internalstorage-mysql            1/1
kube-system                        cluster-1   coredns                                       2/2
tigera-operator                    cluster-1   tigera-operator                               1/1
```

CNCF Sandbox Project - Kubean

Intro: Kubean is a production-ready cluster lifecycle management toolchain based on kubespray and other cluster LCM engine. Deploy Kubean and manage Kubernetes clusters' robust lifecycles through declarative APIs. Supports multi-architecture delivery including AMD, ARM.

Repo: https://github.com/kubean-io/kubean

Trend: Stars 442, Commits 2346, Contributors 47

Adoption: Kubean has been widely used in DaoCloud customer projects for multi-cluster lifecycle management.

```
→ kubean git:(quick-start) ✗ helm repo add kubean-io https://kubean-io.github.io/kubean-helm-chart/
"kubean-io" already exists with the same configuration, skipping
→ kubean git:(quick-start) ✗
→ kubean git:(quick-start) ✗ helm install kubean kubean-io/kubean --create-namespace -n kubean-system
NAME: kubean
LAST DEPLOYED: Fri Jan 26 19:33:13 2024
NAMESPACE: kubean-system
STATUS: deployed
REVISION: 1
TEST SUITE: None
NOTES:
Thank you for installing kubean.

Chart Information:
    Chart Name: kubean
    Chart Description: A Helm chart for kubean

Release Information:
    Release Name: kubean
    Release Namespace: kubean-system

To learn more about the release, try:

  $ helm status kubean -n kubean-system
  $ helm get all kubean -n kubean-system

Documention: https://github.com/kubean-io/kubean/blob/main/README.md
→ kubean git:(quick-start) ✗
→ kubean git:(quick-start) ✗ kubectl get pods -n kubean-system
NAME                              READY   STATUS    RESTARTS   AGE
kubean-869484f6cc-xsn76           1/1     Running   0          7s
kubean-admission-59cc54dfcc-8gmhx   1/1     Running   0          7s
→ kubean git:(quick-start) ✗
→ kubean git:(quick-start) ✗ kubectl apply -f examples/install/1.minimal
configmap/mini-hosts-conf created
configmap/mini-vars-conf created
cluster.kubean.io/cluster-mini created
clusteroperation.kubean.io/cluster-mini-install-ops created
→ kubean git:(quick-start) ✗ kubectl get job -n kubean-system
NAME                                COMPLETIONS   DURATION   AGE
kubean-cluster-mini-install-ops-job   0/1           5s         5s
→ kubean git:(quick-start) ✗ exit
```

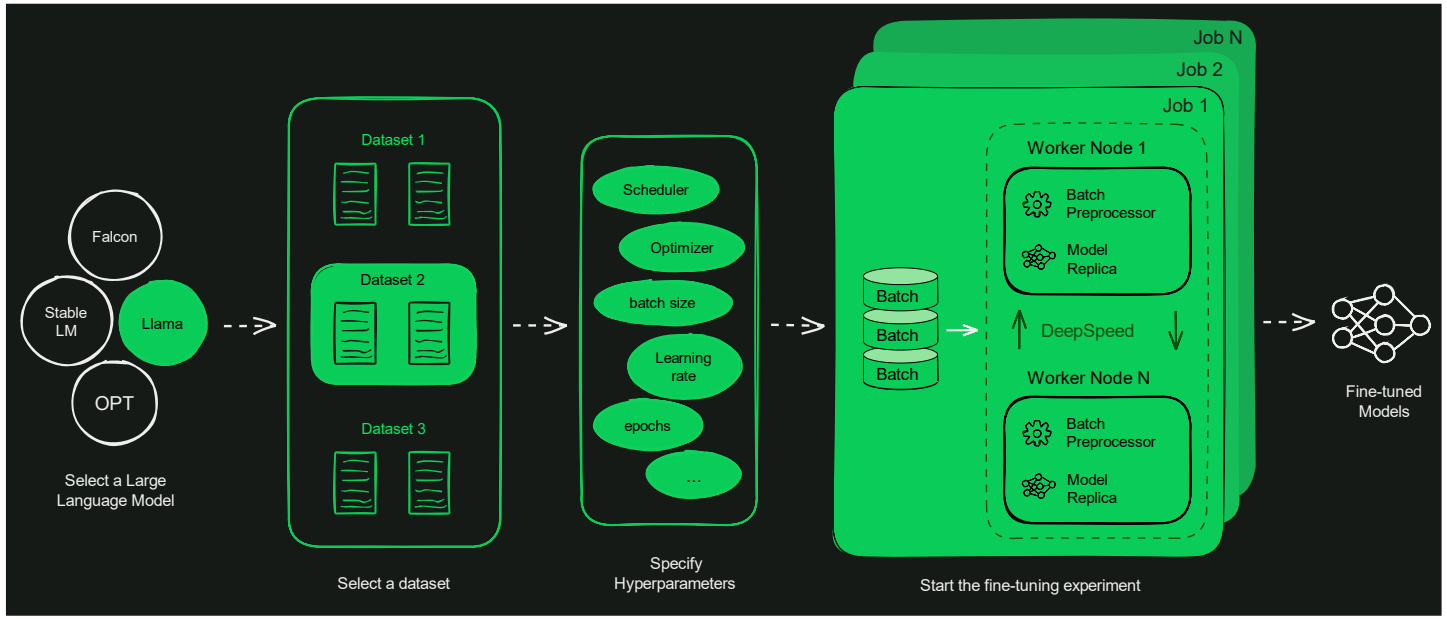
Project - DataTunerX

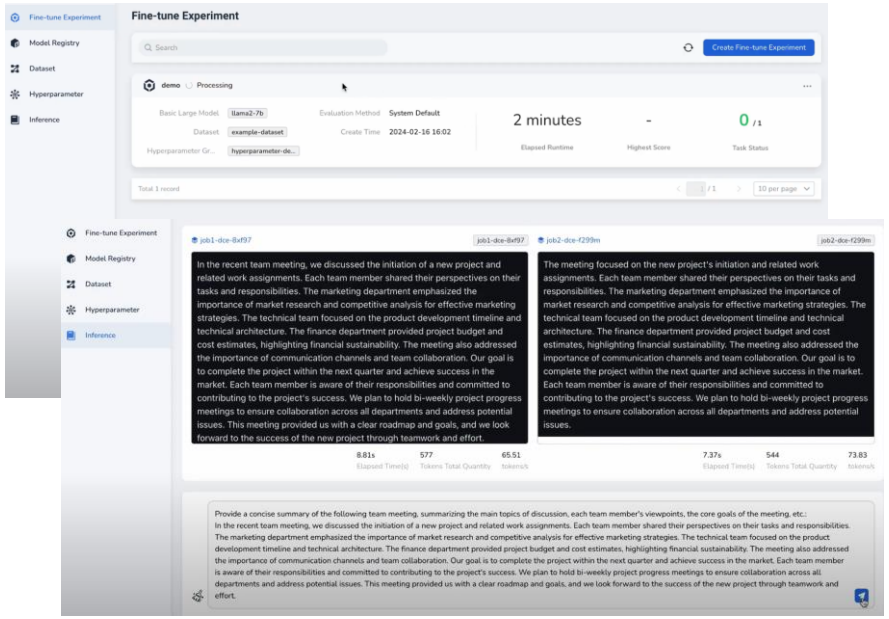
Intro: Large language model fine-tuning capabilities based on cloud native and distributed computing. DataTunerX (DTX) is designed as a cloud-native solution integrated with distributed computing frameworks. Leveraging scalable GPU resources, it's a platform built for efficient fine-tuning LLMs with a focus on practical utility.

Repo: https://github.com/DataTunerX/datatunerx

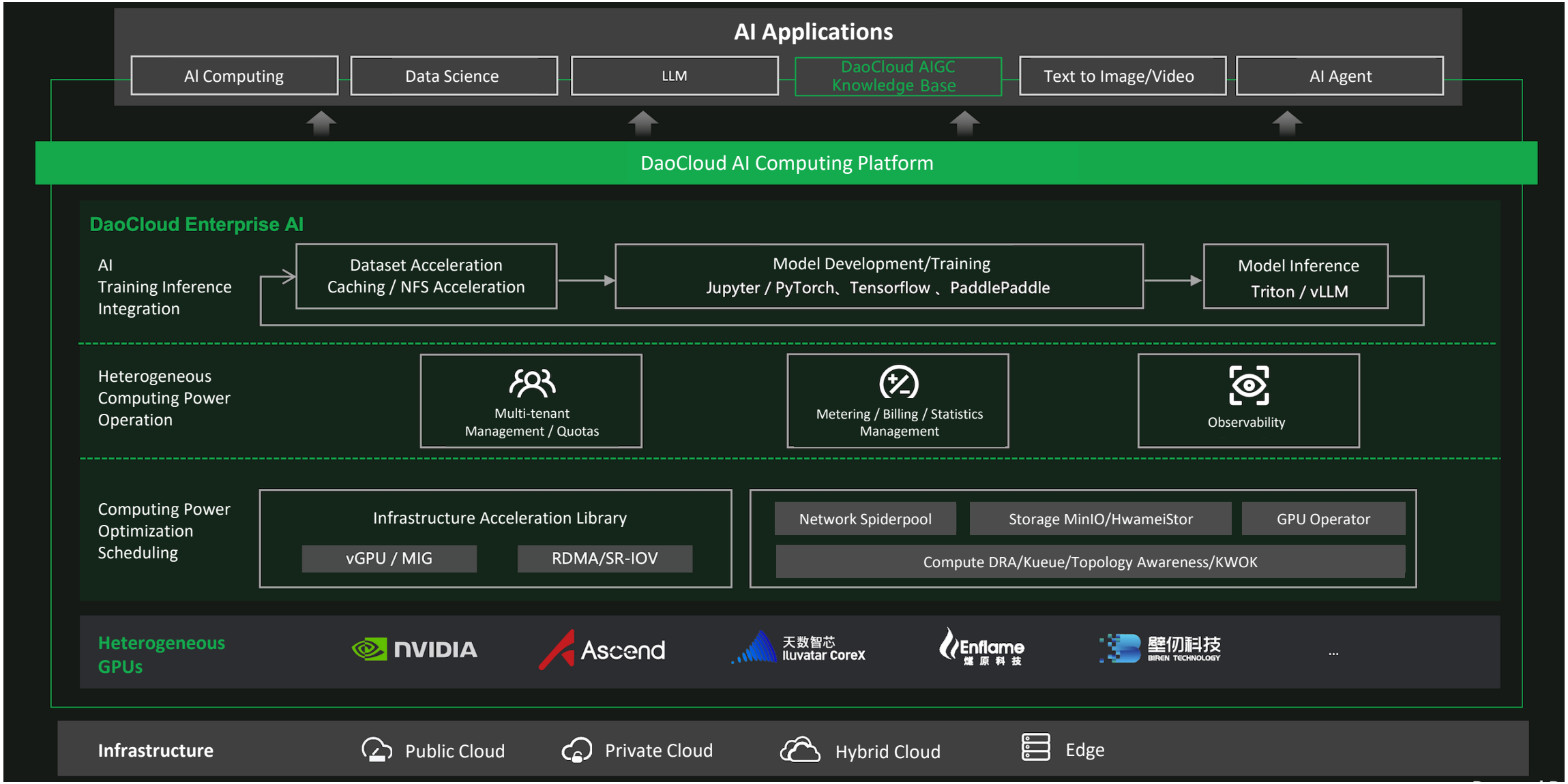
Trend: Stars 84, Commits 61, Contributors 3

Adoption: DataTunerX is currently one of the Model Suites of DaoCloud D.RUN advanced solutions.

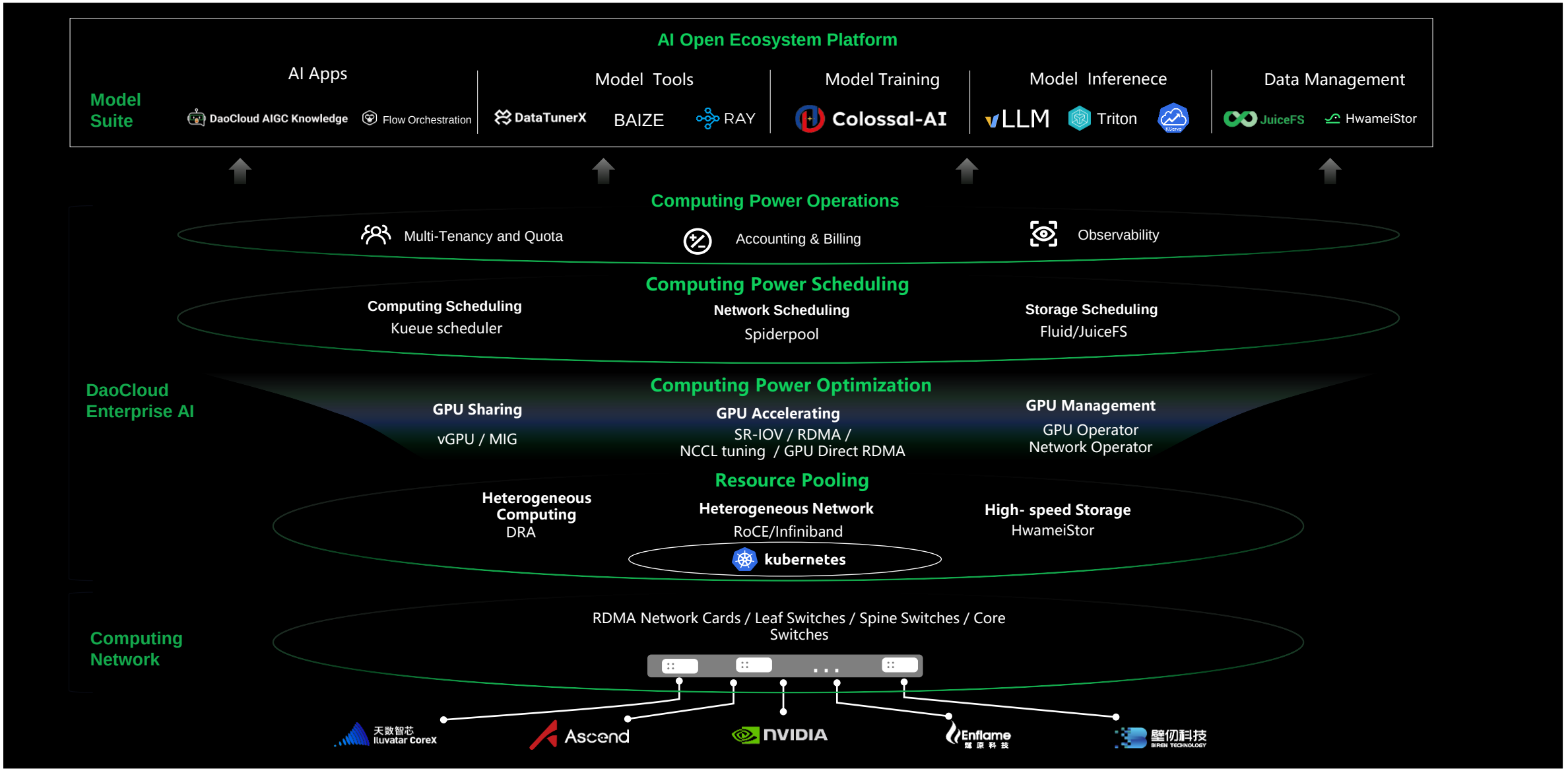




Advanced Solutions - DCE BAIZE

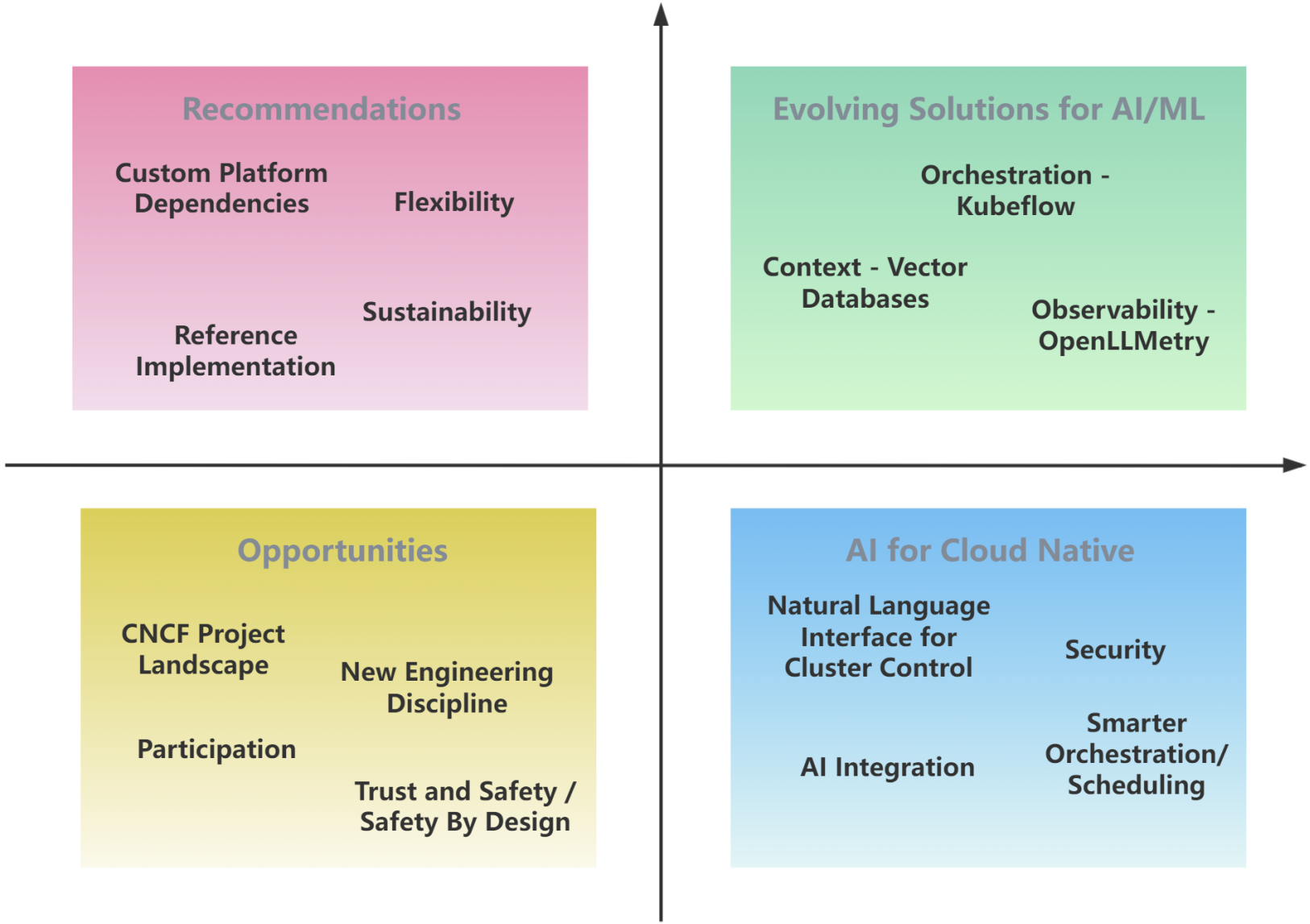


Advanced Solutions - D.RUN



Part 04
Future Directions in Cloud-Native AI

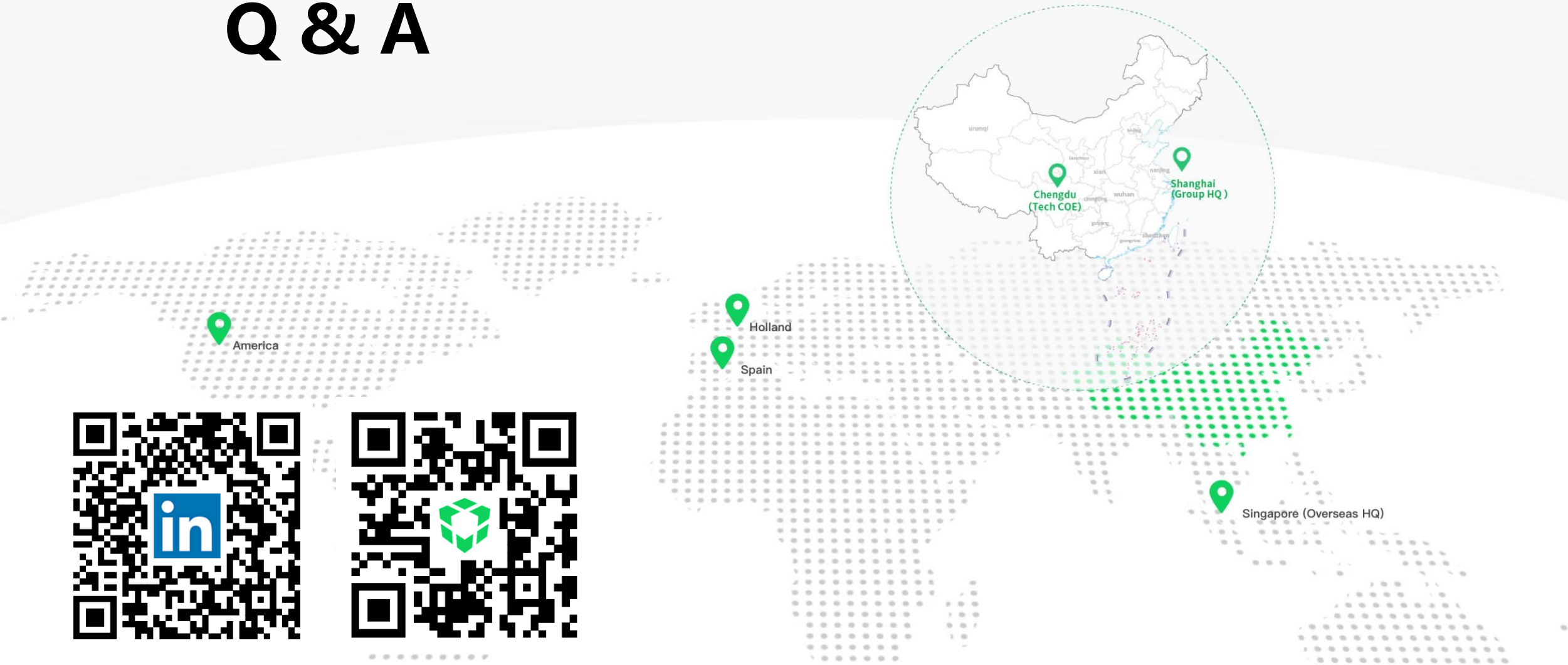
Future Directions and Innovations



Thanks.

DaoCloud BYTEDANCE OPEN SOURCE Shopee WasmEdgeRuntime

Q & A



Unlocking the Potential of Multi-Cloud Kubernetes with KubeAdmiral

Gary Liu, ByteDance

KubeAdmiral

Speaker

Gary is a software engineer from the Orchestration and Scheduling team @ ByteDance, focusing on developing large-scale cloud-native infrastructure.

Part 01

Kubernetes @ ByteDance

kubernetes

- An open-source container orchestration platform that automates the deployment, scaling, and management of containerized applications.
- Launched in 2014, Kubernetes has seen widespread adoption and has become the de facto container orchestration solution.

Why is Kubernetes so popular?

01

Automated Deployments

Kubernetes allows you to define how your application should be deployed and what resources it needs. It then automates the process of actually creating and managing the necessary containers and infrastructure.

02

Feature-completeness

Kubernetes comes with many features out-of-the box, including service discovery, load balancing, and configuration management.

03

Declarative nature

Applications are managed with declarative configurations, making it easy to reason about and achieve the desired state of the system.

Kubernetes Architecture

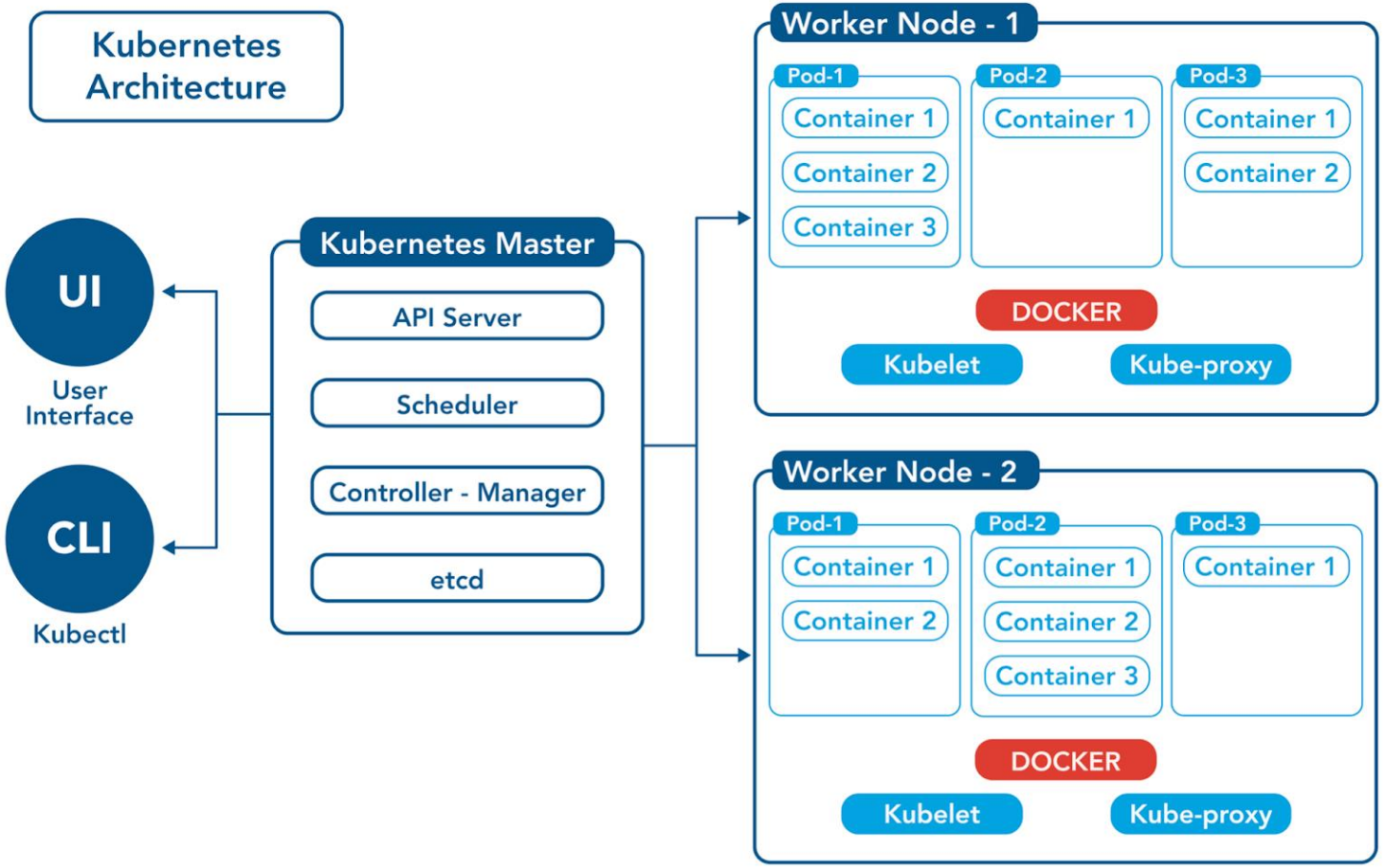
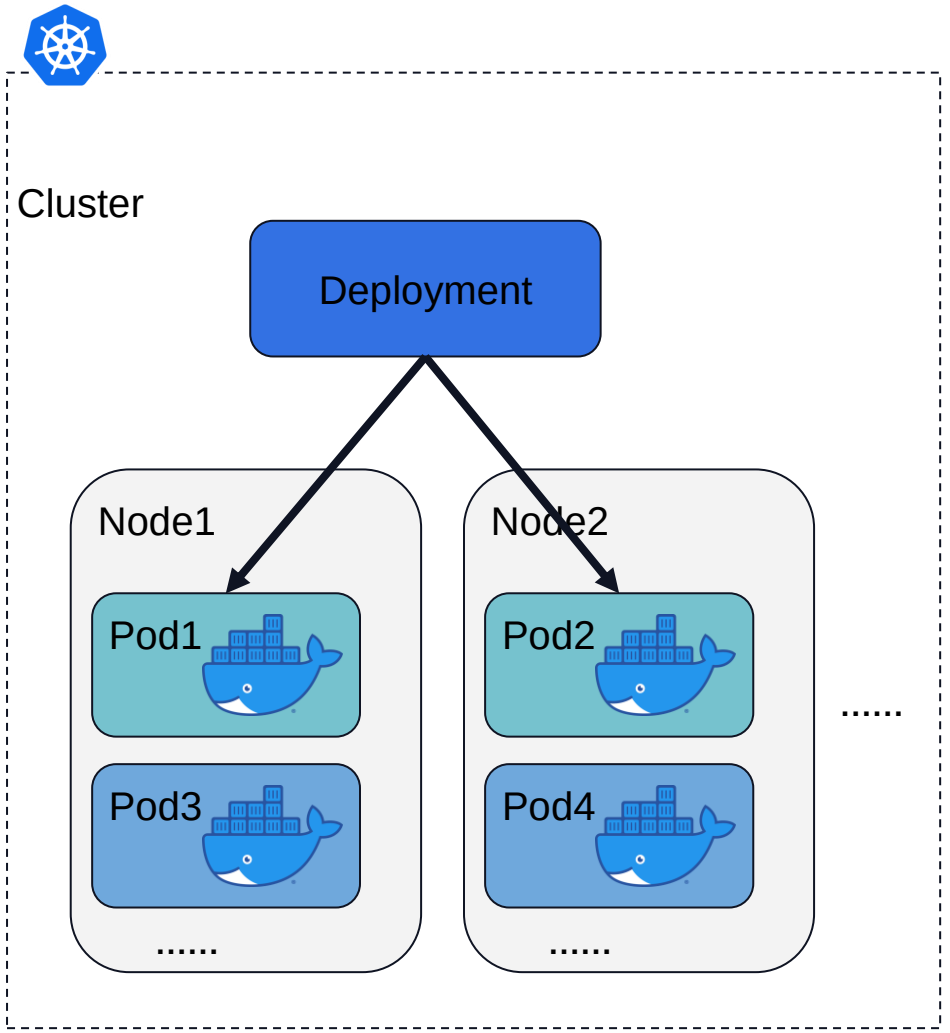


Image credit: https://www.opsramp.com/wp-content/uploads/2022/07/Kubernetes-Architecture.png

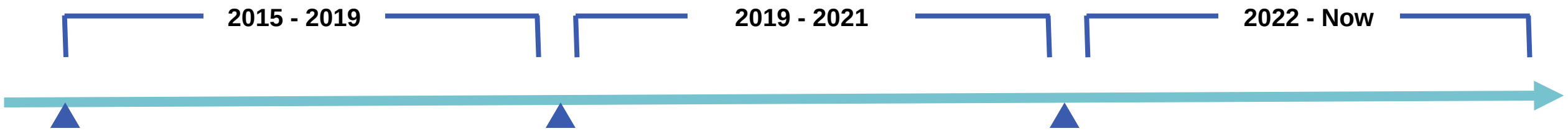
Deploying an Application Using Kubernetes

```yaml
apiVersion: apps/v1
kind: Deployment
metadata:
 name: nginx-deployment
spec:
 replicas: 2
 template:
   spec:
     containers:
     - name: nginx
       image: nginx:1.14.2
       ports:
       - containerPort: 80
```

```
kubectl create -f nginx-deploy.yaml
```



The Cloud Native Journey @ ByteDance



Basic capabilities

Built microservice platform with Kubernetes

Scale and resource efficiency

Adopted Kubernetes Federation based on KubeFed V2 for microservice orchestration

Open source and diversification

Next-Gen Kubernetes Federation with KubeAdmiral

Part 02

Multi-Cluster & Multi-Cloud Kubernetes

Why Multi-Cluster and Multi-Cloud?

01

Scale Limit

The officially recommended number of nodes in a single cluster is 5000. As businesses grow, 5000 nodes is no longer sufficient and multiple clusters have to be used.

02

Fault tolerance

Having applications spread across multiple clusters provides fault tolerance when a single cluster fails.

03

Locality

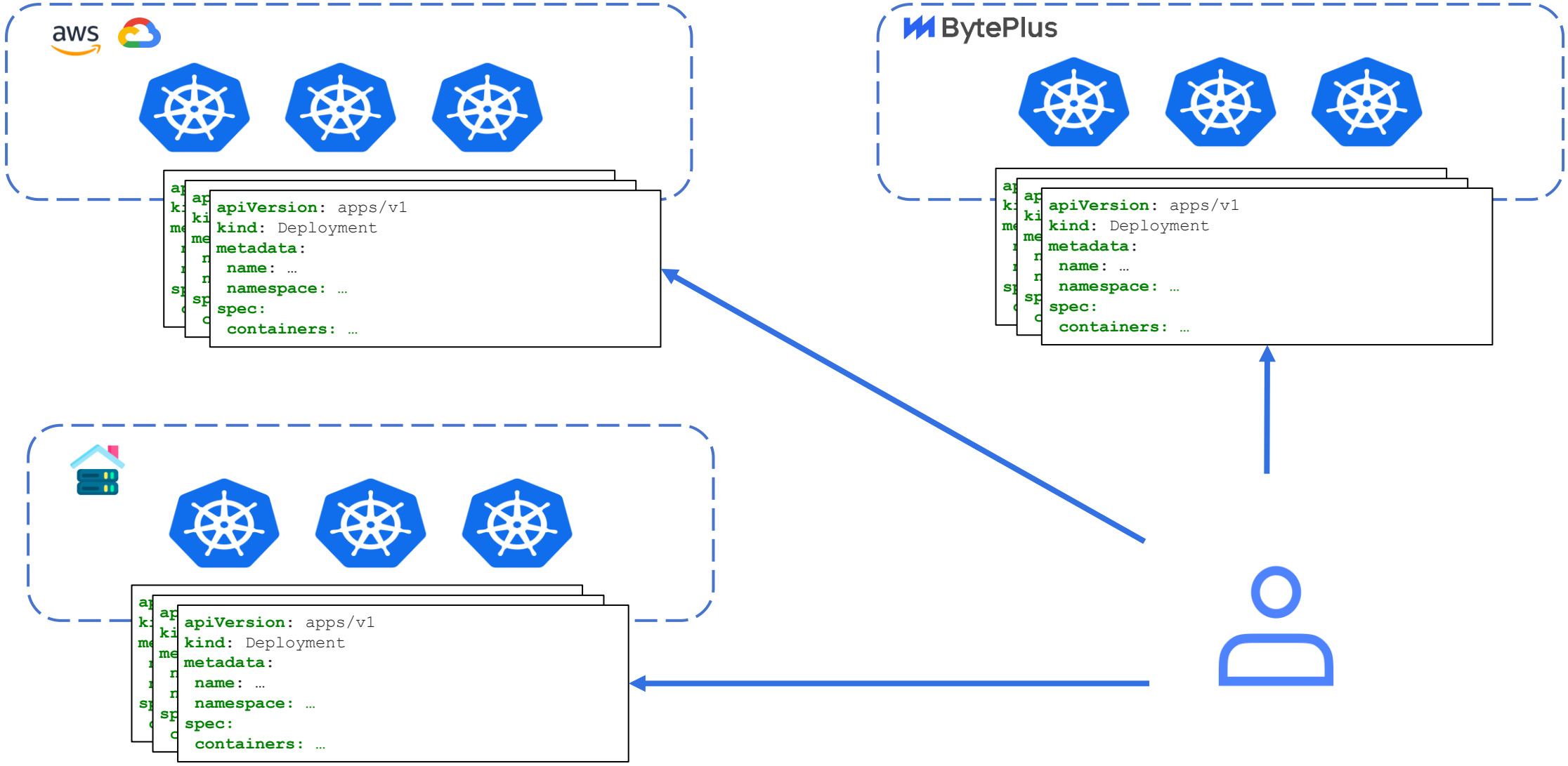
Having different clusters located closer to users can help to reduce latency of applications.

04

Avoid vendor lock-in

By staying agnostic to a particular cloud provider, we retain the flexibility to choose the best services, pricing and features from multiple providers.

Typical Multi-Cloud, Multi-Cluster Setup



… but it comes with many challenges

01

Mental Overhead

With multiple clusters, users must go through the additional trouble of choosing the right cluster to deploy to.

02

Operational Overhead

Users have to repeat their deployment for each cluster, which is burdensome and can lead to inconsistencies across clusters.

03

Resource Inefficiency

The resource utilisation across clusters is not uniform, resulting in low resource efficiency.

Part 03

KubeAdmiral - Next-Gen Kubernetes Cluster Federation

KubeAdmiral to the Rescue

KubeAdmiral is the next-generation Kubernetes cluster federation engine aimed at addressing the aforementioned pain points.

KubeAdmiral was initially based on the open source project KubeFed V2, and has been used at ByteDance since 2019.

KubeAdmiral

https://kubeadmiral.io

KubeAdmiral Manages Mission-Critical Workloads

dozens of clusters

microservices, batch jobs, serverless…

Supporting hundreds of millions of active users

300,000+ microservices

30,000+ deployments per day

Multi-Cloud Kubernetes, Minus the Headache

01

Overcome the limit of single-cluster Kubernetes

Run >5000 nodes in a single resource pool and benefit from cross-cluster rebalance and fault-tolerance.

02

Easily migrate from single-cluster Kubernetes

KubeAdmiral is compatible with the native Kubernetes API; you can use KubeAdmiral just like using single-cluster Kubernetes.

03

Configurable and extensible

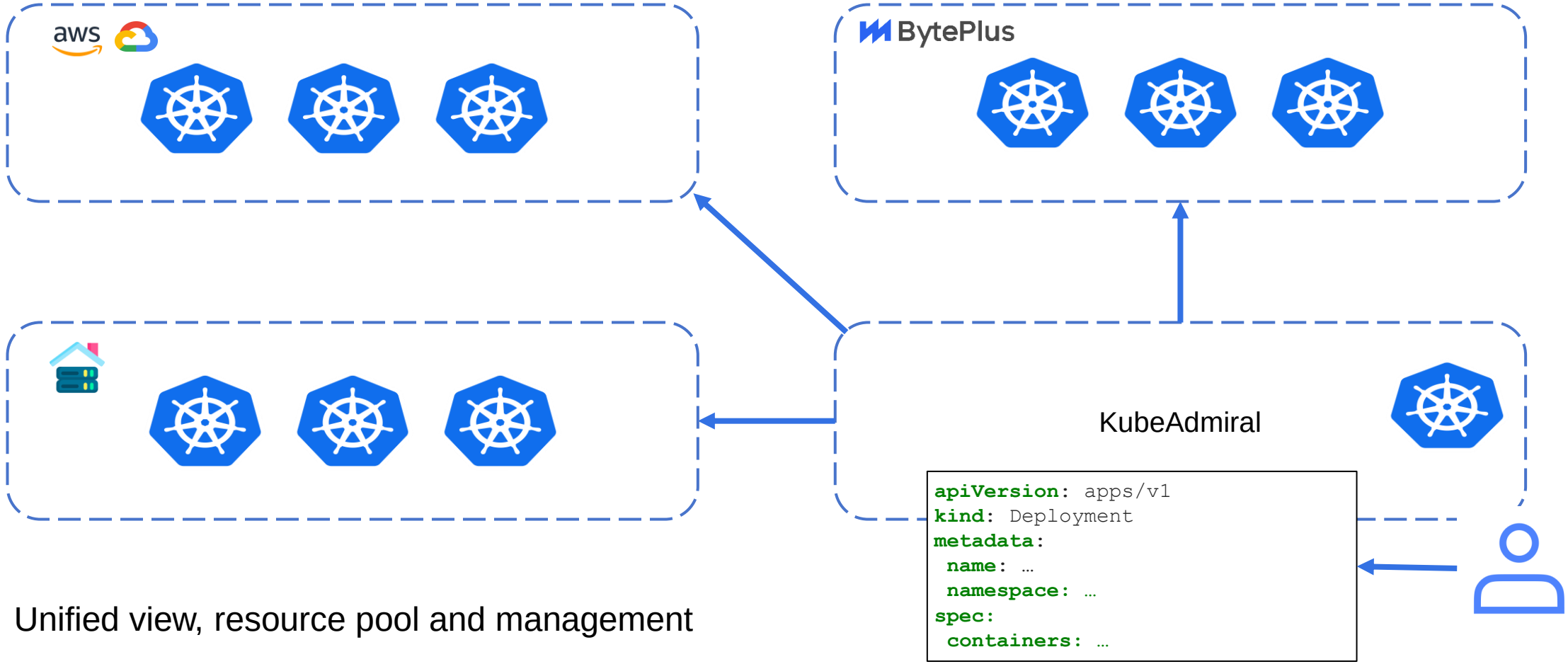
From scheduling policy to resource propagation, KubeAdmiral exposes a lot of knobs for you to turn. KubeAdmiral can be extended to enable new behaviours.

04

Extremely Performant

KubeAdmiral is battle-tested at the ByteDance production scale.

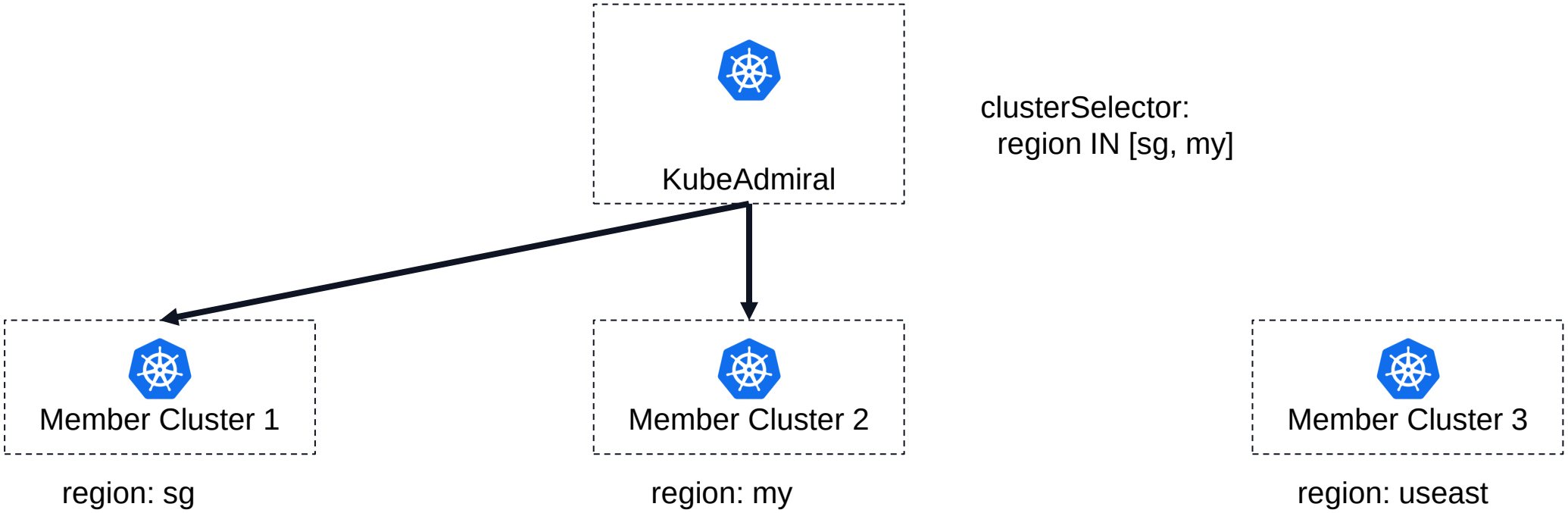
Use Case 1: Cross-Cluster Workload Management



Unified view, resource pool and management

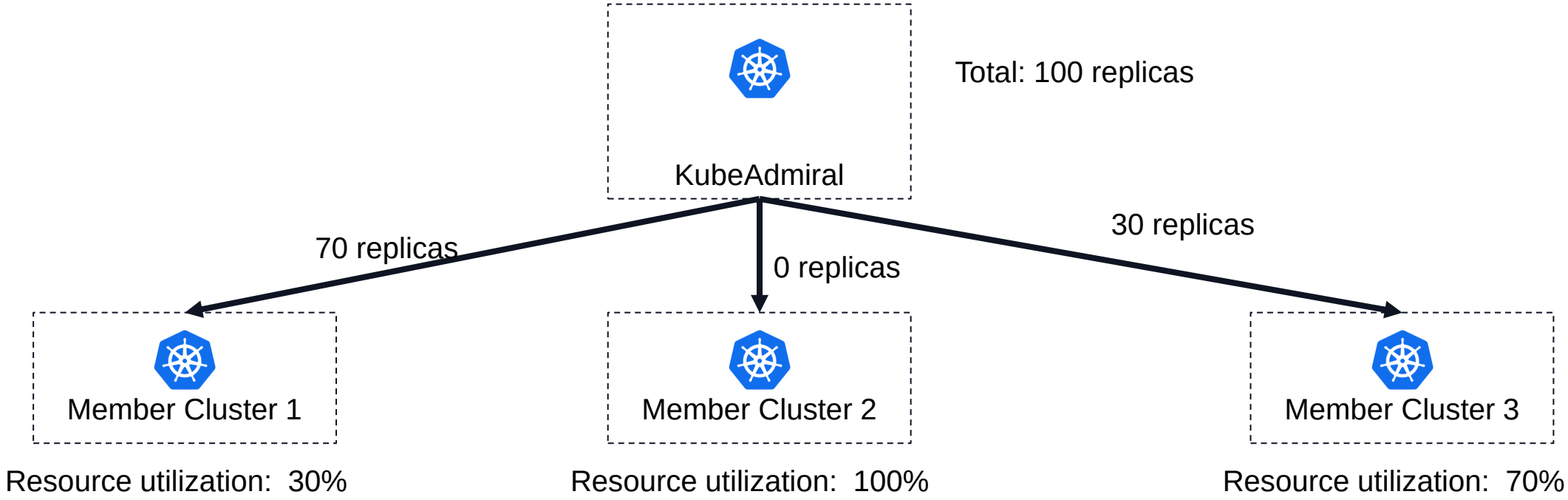
Use Case 2: Flexible Cluster-Level Scheduling

KubeAdmiral can schedule workloads to clusters with specific characteristics based on the workload’s requirements.



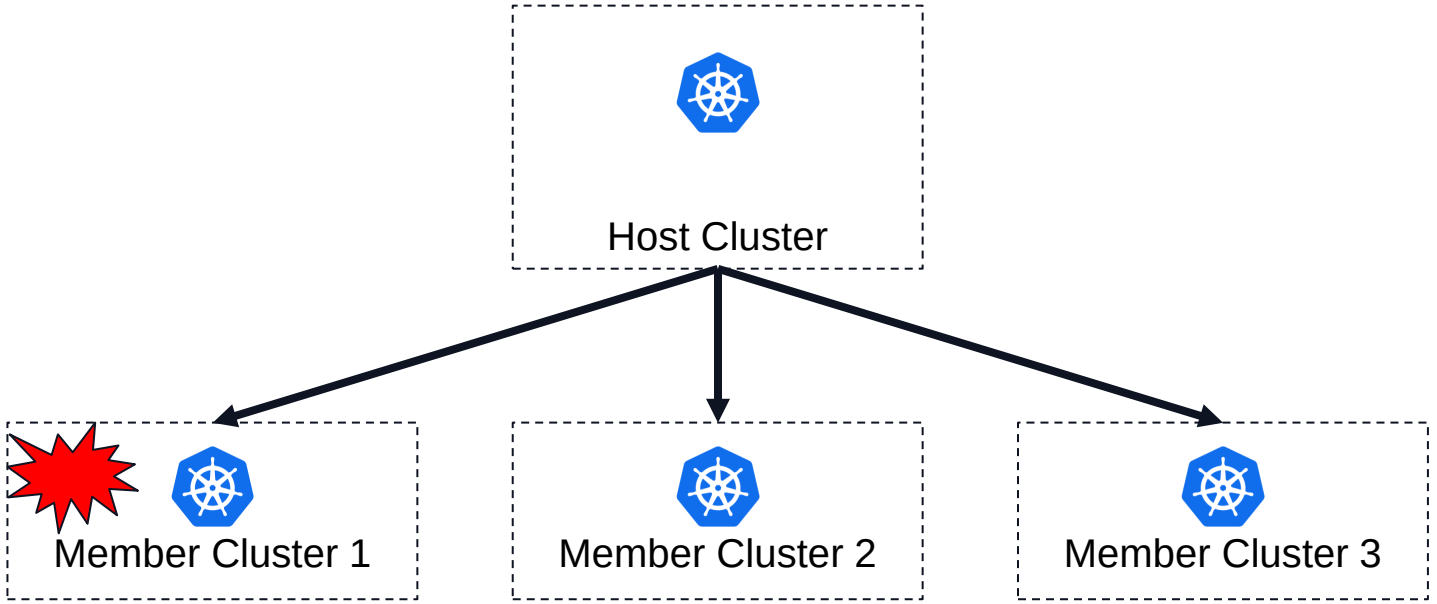
Use Case 3: Dynamic Replica Distribution

Dynamic replica distribution based on cluster resource utilization for higher resource efficiency (95%-98% at ByteDance).



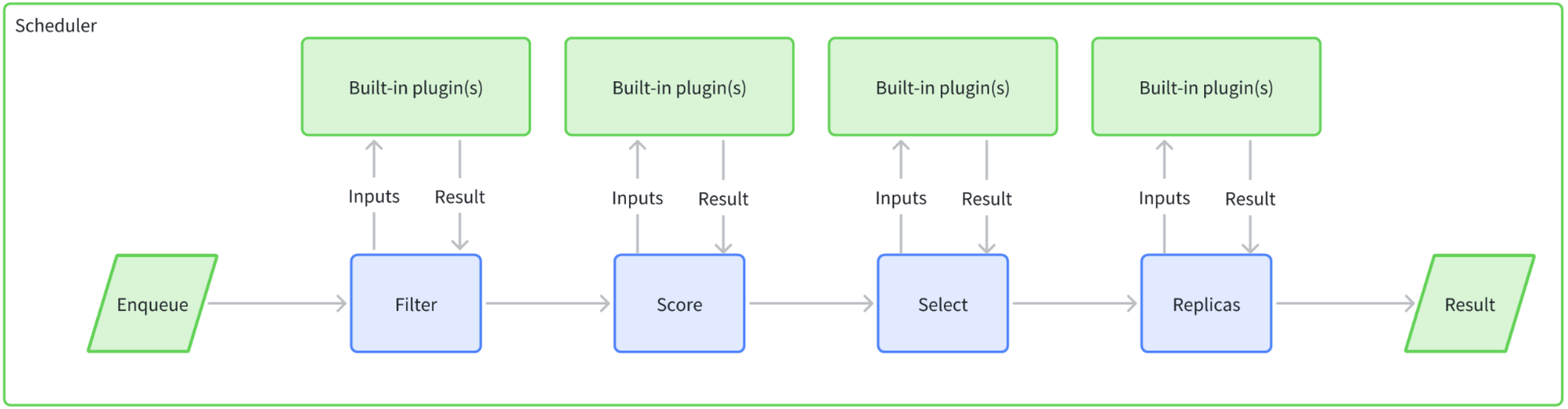
Use Case 4: Cross-Cluster Failover

Workloads in unhealthy clusters can be migrated to healthy clusters.



Use Case 5: Extending the Scheduler with Custom Logic

The KubeAdmiral scheduler can be extended with plugin to filter and score clusters based on custom criteria, e.g. geographical topology, tenancy, QoS, resource utilization, etc.



Advanced Features

Yet to be open-sourced:

- Respecting deployment rolling update strategy across multiple clusters.
- Protection against accidental pod deletion.
- Sharding for orchestrating even larger cluster federations.

Part 04

KubeWharf - The Enhanced Kubernetes Toolkit

Other KubeWharf Projects

KubeBrain

High-performance metadata system for Kubernetes; etcd drop-in replacement

KubeGateway

Traffic governance for kube-apiserver

KubeZoo

Light-weight Kubernetes multi-tenancy gateway

Katalyst

Resource efficiency toolkit for Kubernetes

Gödel Scheduler

An unified scheduler for online and offline tasks

Kelemetry

Global control plane tracing for Kubernetes

Thanks!

KubeWharf

https://kubewharf.io

KubeAdmiral

https://kubeadmiral.io

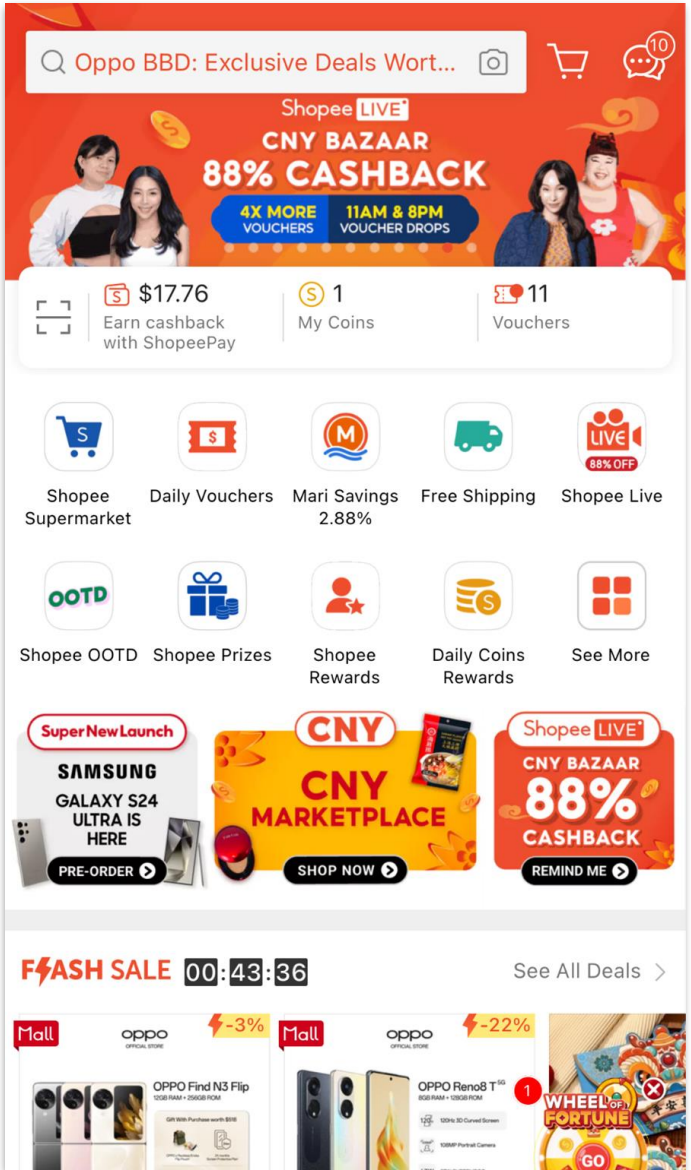
WeChat Official Account

Squeeze Your K8S

Adopting Time-Series Forecasting in FinOps Practices

Who we are

Shopee



Who we are

Leading e-commerce platform in Southeast Asia, Taiwan and Brazil

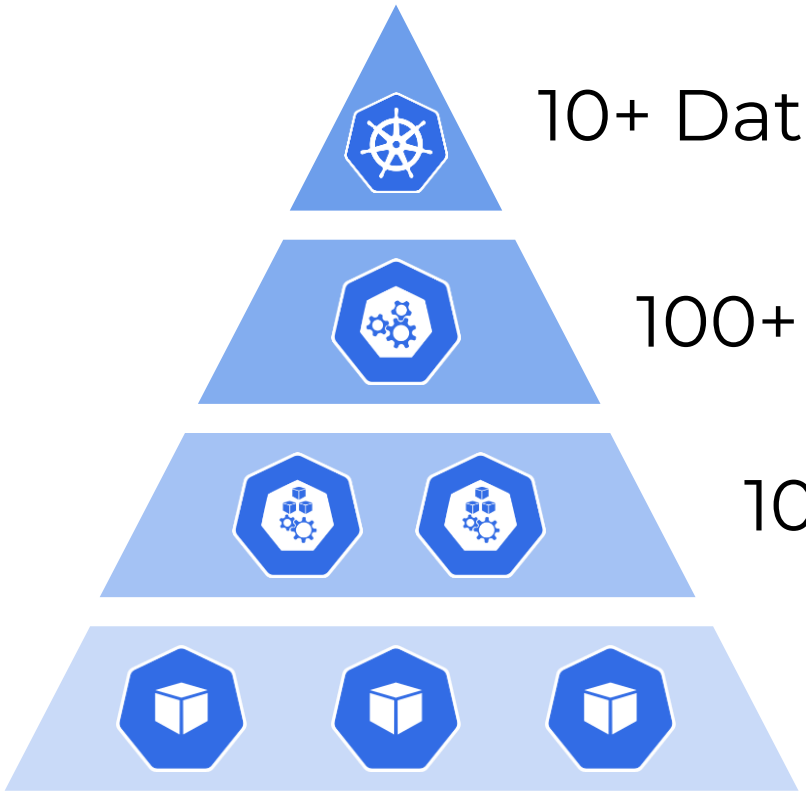
#1 Shopping App in Southeast Asia and Taiwan

By average Monthly Active Users and total time spent in-app

#1 Shopping App in Brazil

By average Monthly Active Users and total time spent in-app

Kubernetes in Shopee



10+ Data Centers

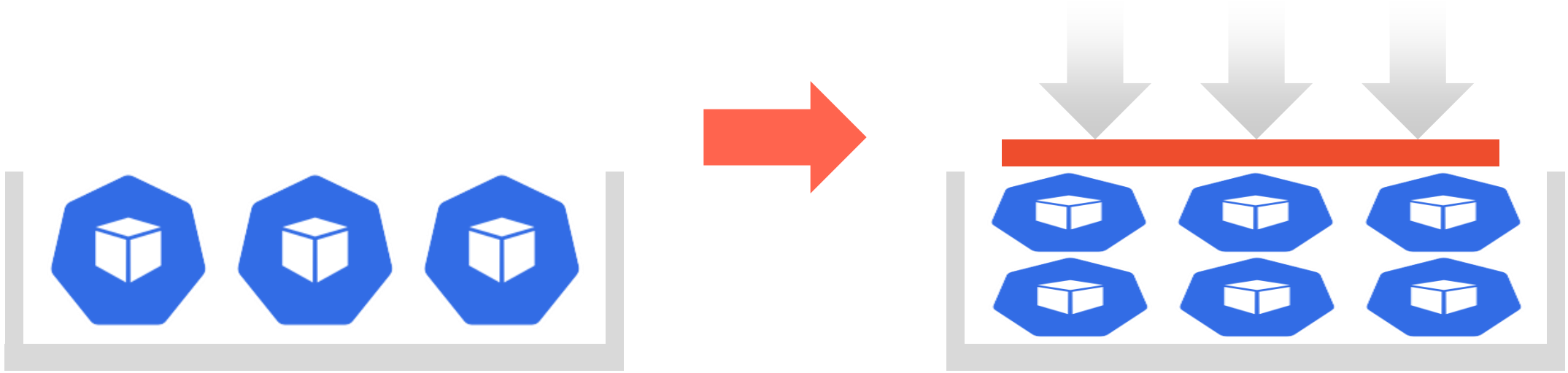
100+ Clusters

10,000+ Nodes

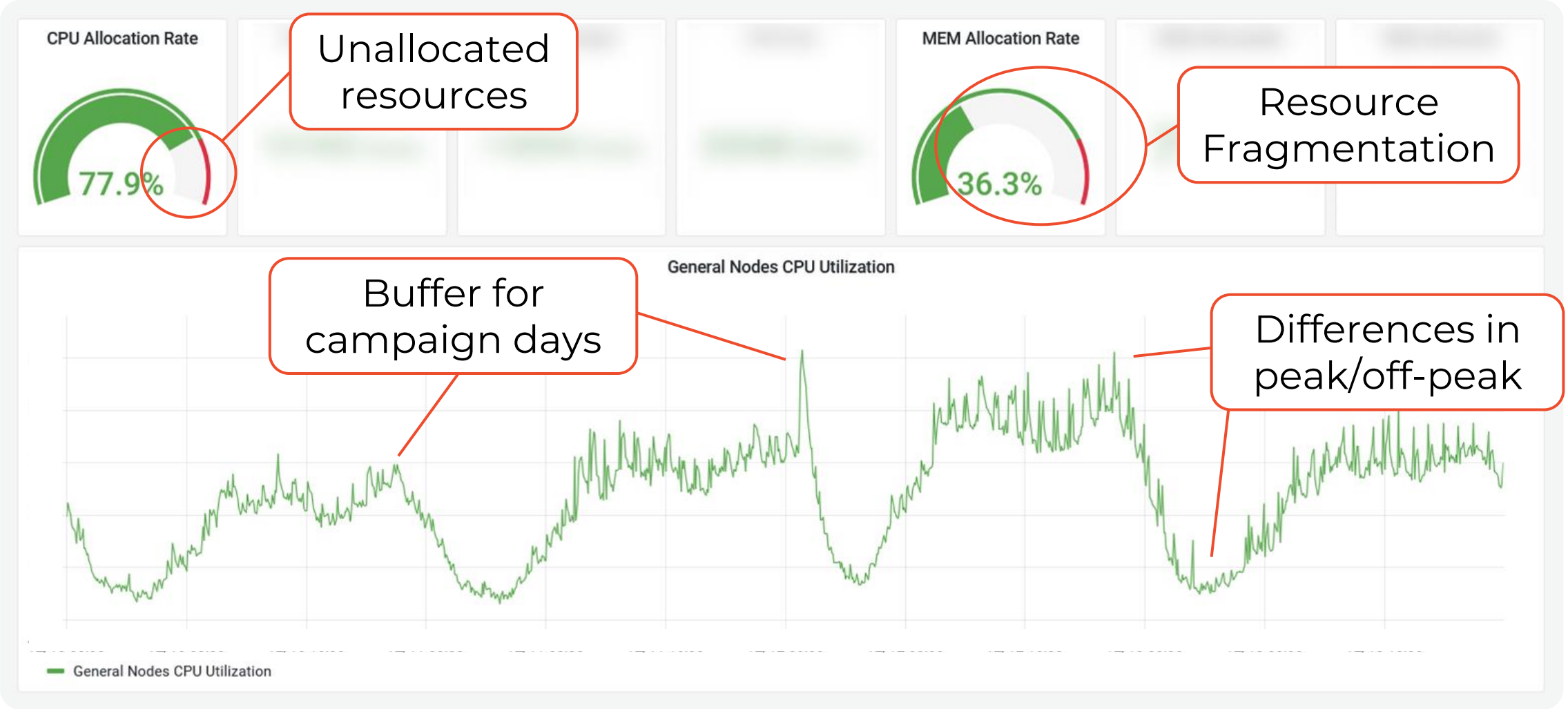
100,000+ Pods

FinOps: Minimizing Resource Costs

Improve efficiency of existing resources to support more workloads

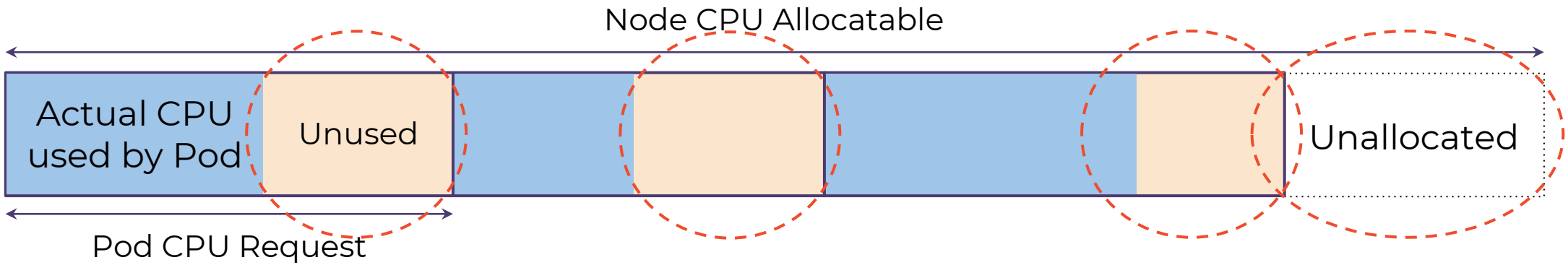


Sources of Resource Wastage



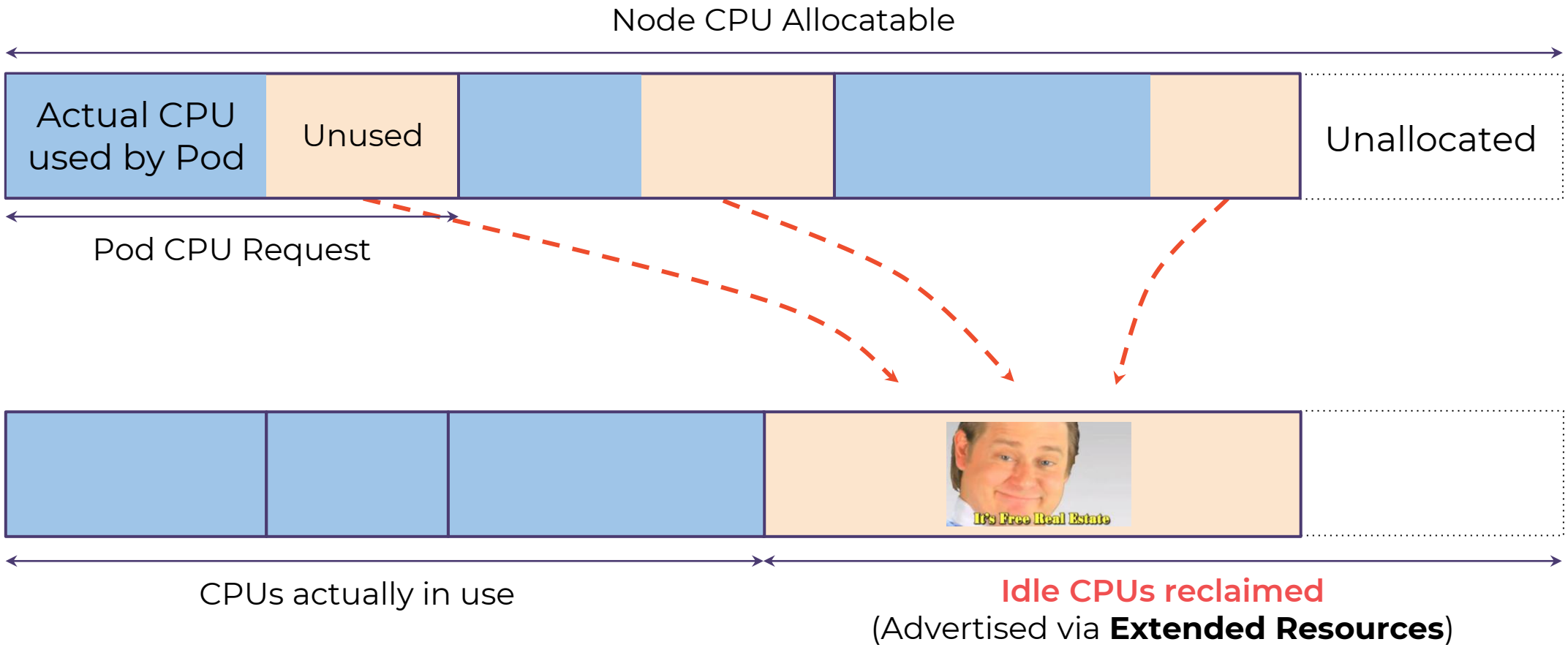
Maximizing use of Physical Resources

What can we do with idle resources?



Maximizing use of Physical Resources

Run Batch services to consume idle resources!



Introducing Batch Services

Batch Services

Big Data, Transcoding

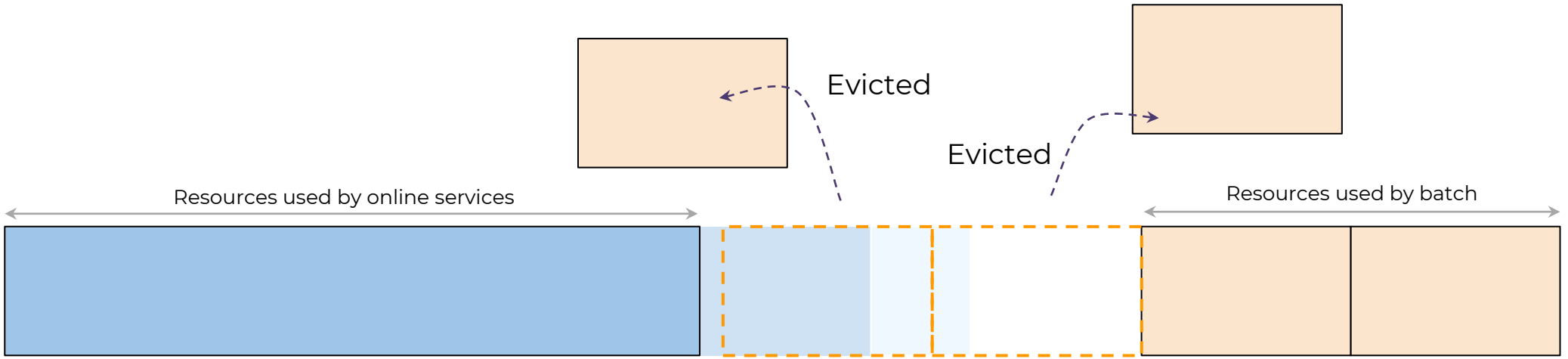
- Tolerates eviction
- Tolerates throttling
- 90% Availability

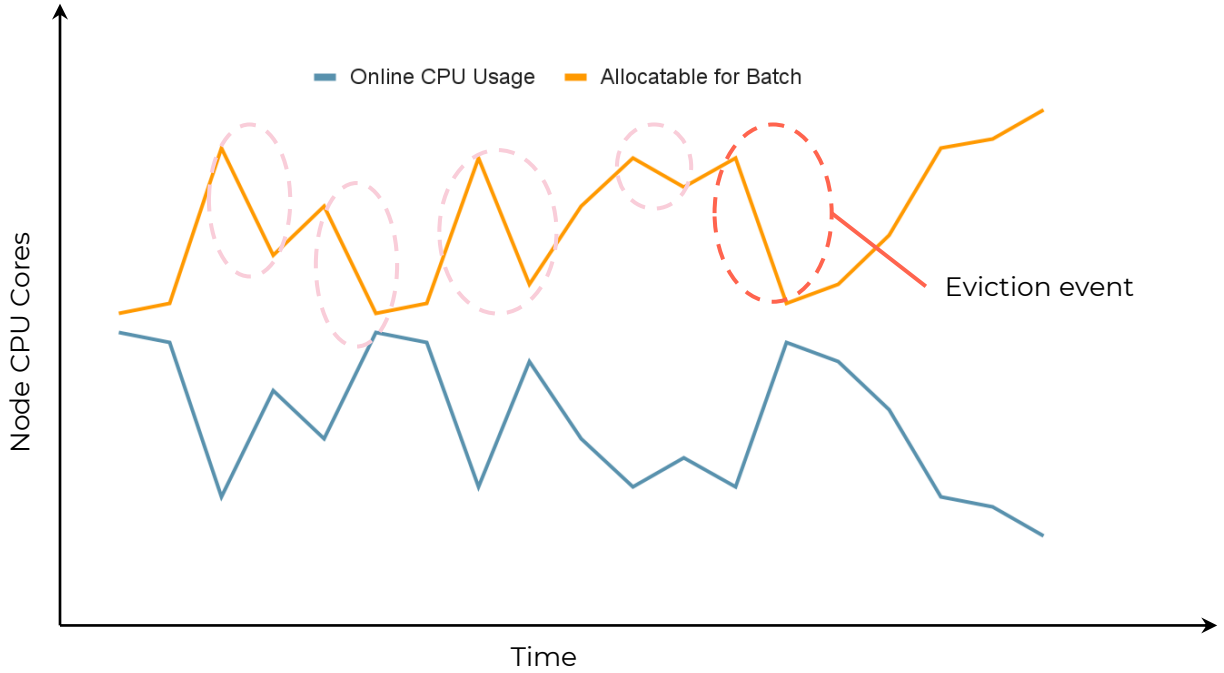
Online Services

Web servers, Databases

- User facing
- Latency SLA (< 300ms)
- 99.9...% Availability

Challenges with Batch Resources



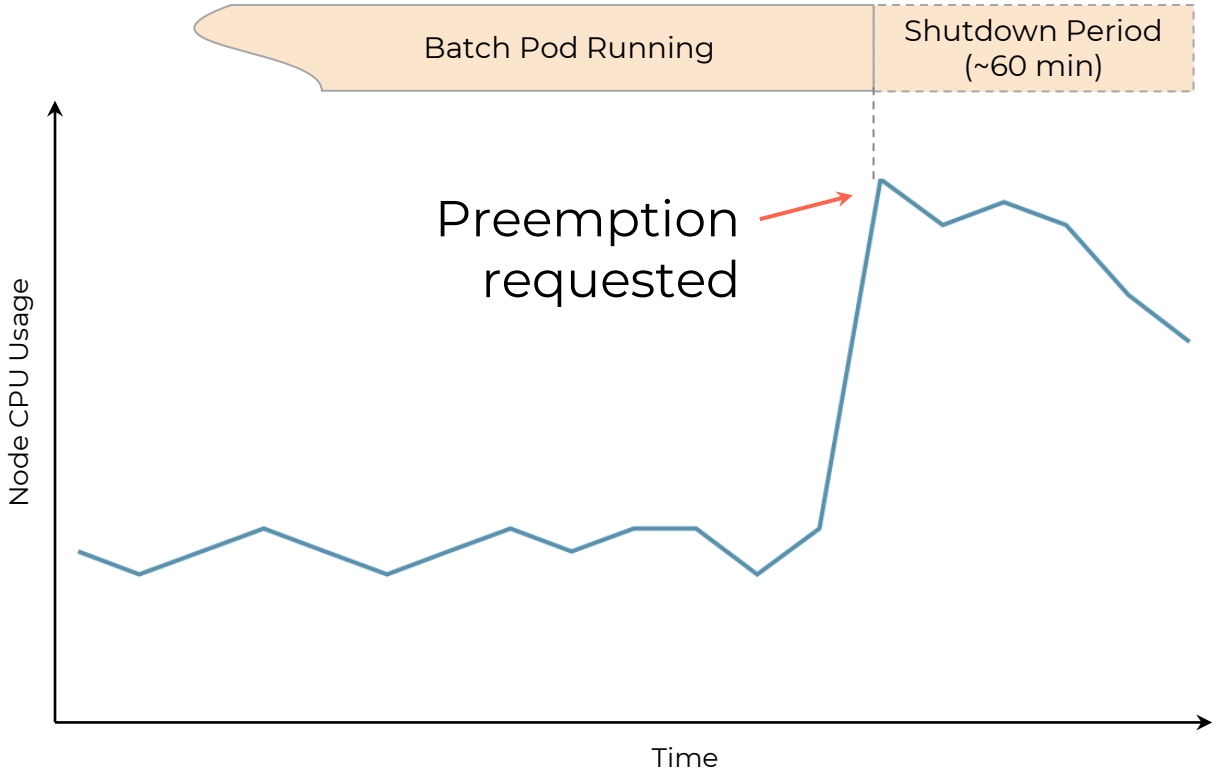


Problem

If Online pod resource usage oscillates, Batch pods gets evicted frequently...

Eviction Cost

Some workloads require grace period for termination



Problems

- Batch job restarts from beginning if forced killed
- Batch job starves if not killed

Practical Forecasting

Using Forecasts to Solve Allocation Challenges

Types of Forecasting

Short-Term

- Focuses on trend
- Higher accuracy and precision due to recency
- Detect anomalies in real-time

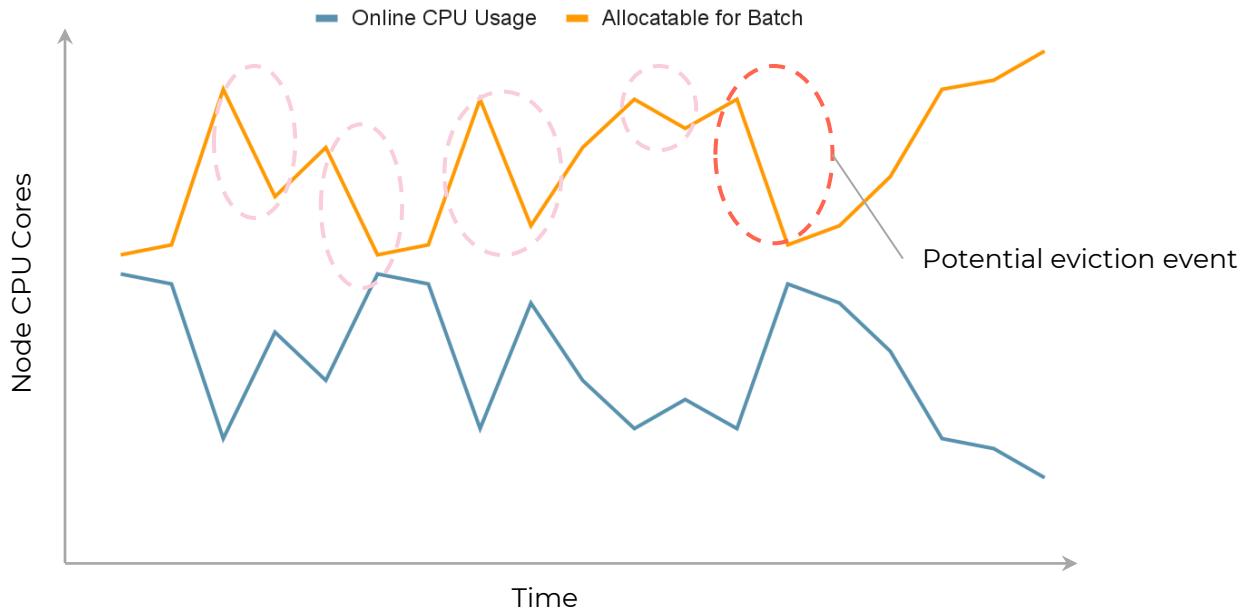
Avoid flapping of allocatable Batch resources

Long-Term

- Focuses on patterns (cyclical, seasonal)
- Lower precision
- Longer prediction window

Predict if pod can be placed without eviction later

Short-Term Forecasting

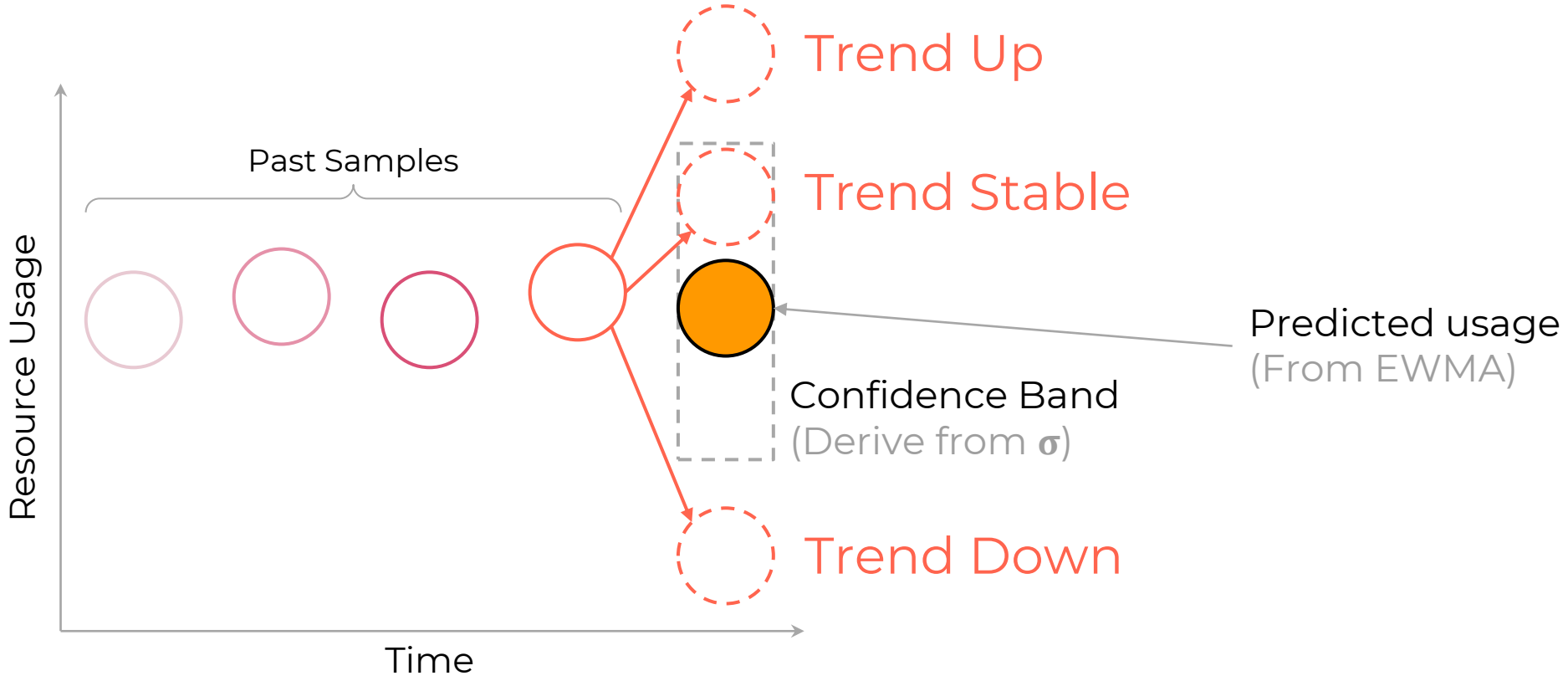


Goals

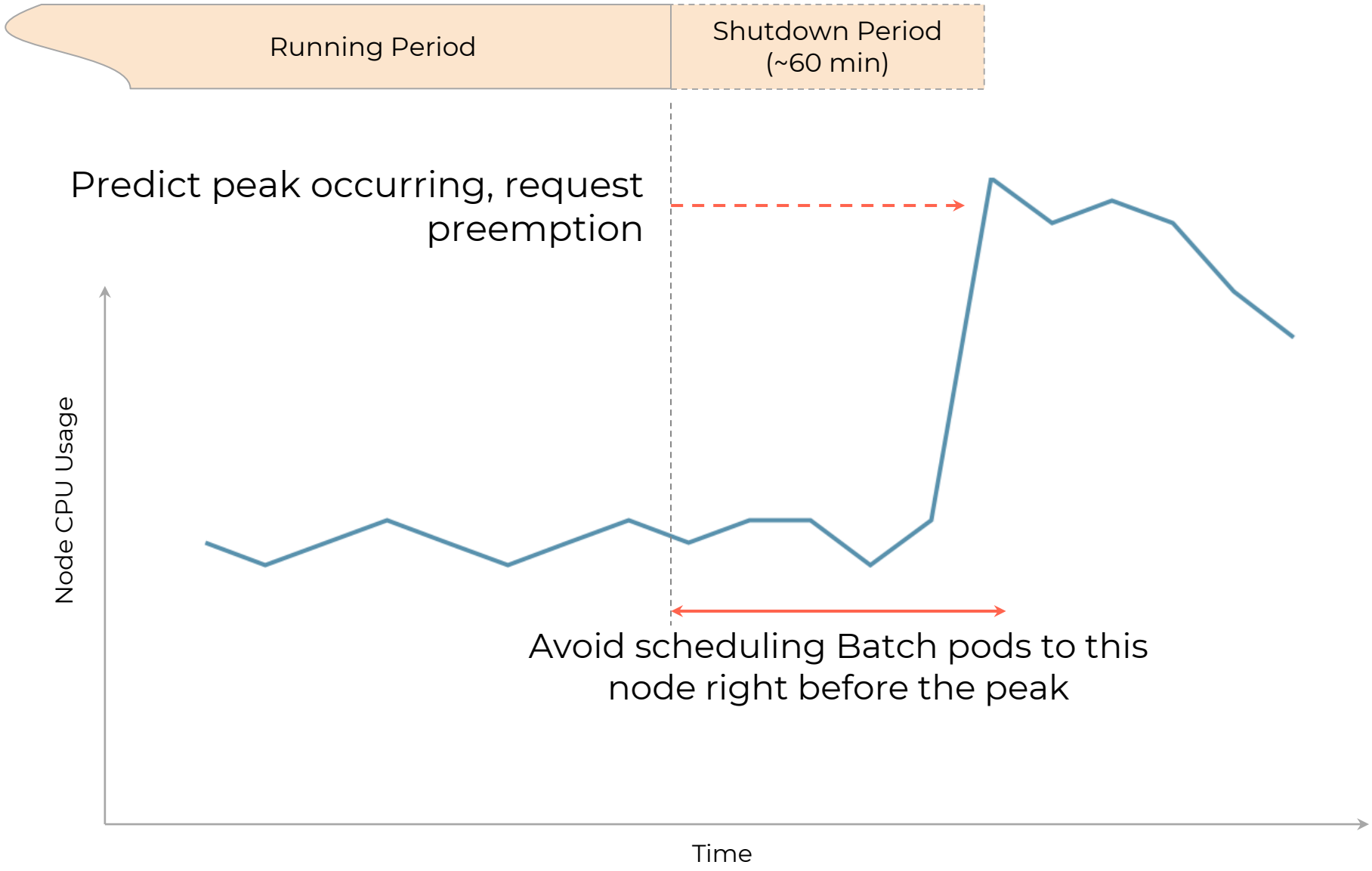
- Reduce number of evictions
- React quickly to sudden reductions in resources

Anomaly Detection

- Exponentially-Weighted Moving Average reacts quickly to recent changes in data
- Limit changes of allocatable resources that arise from noisy data

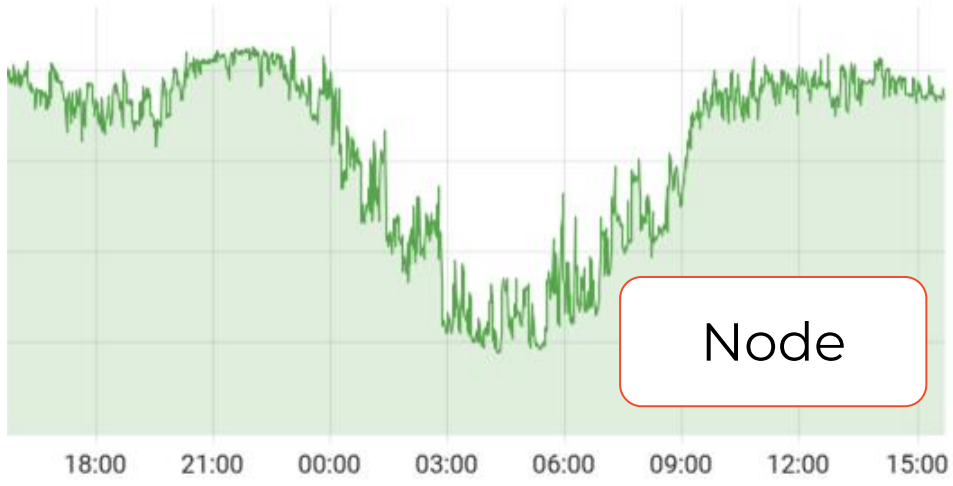


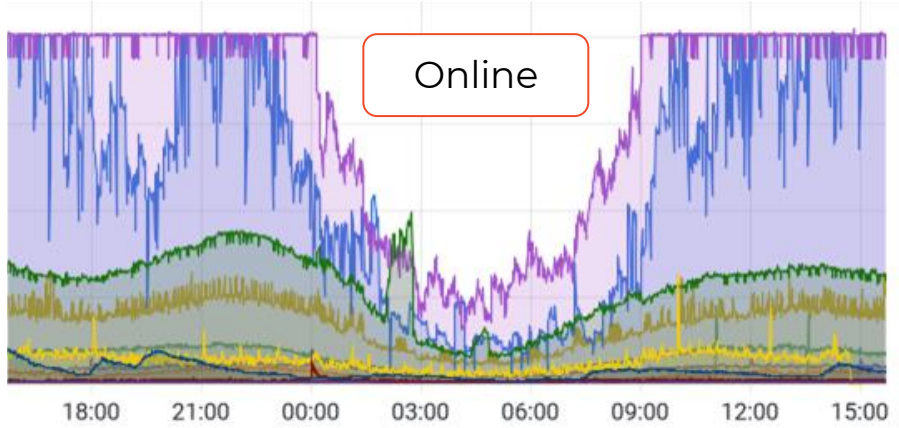
Graceful Shutdowns with Long-Term Forecasts

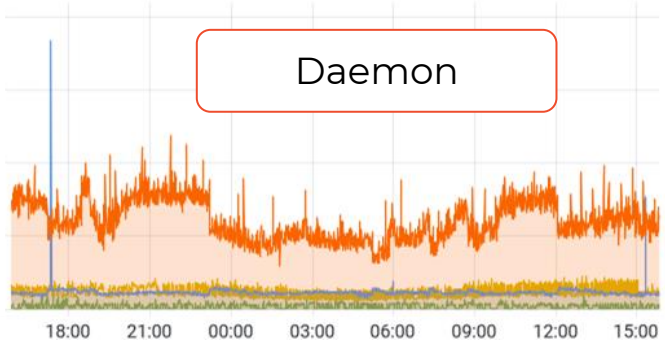


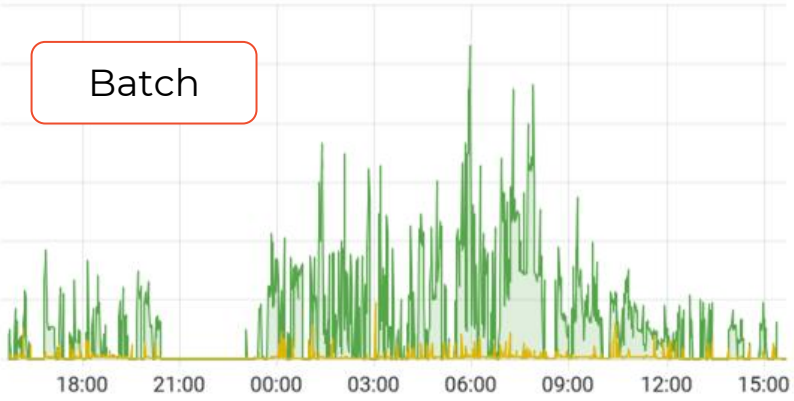
Breaking Down Node Utilization

Decompose components to forecast each part independently



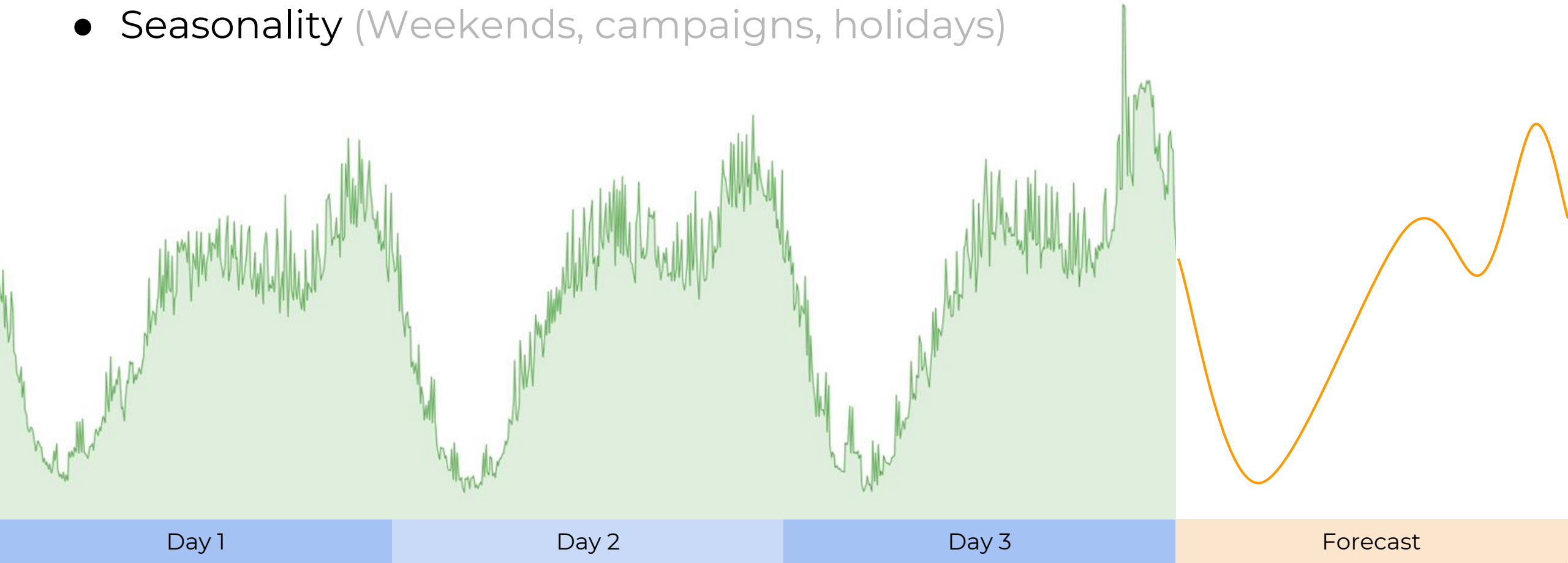




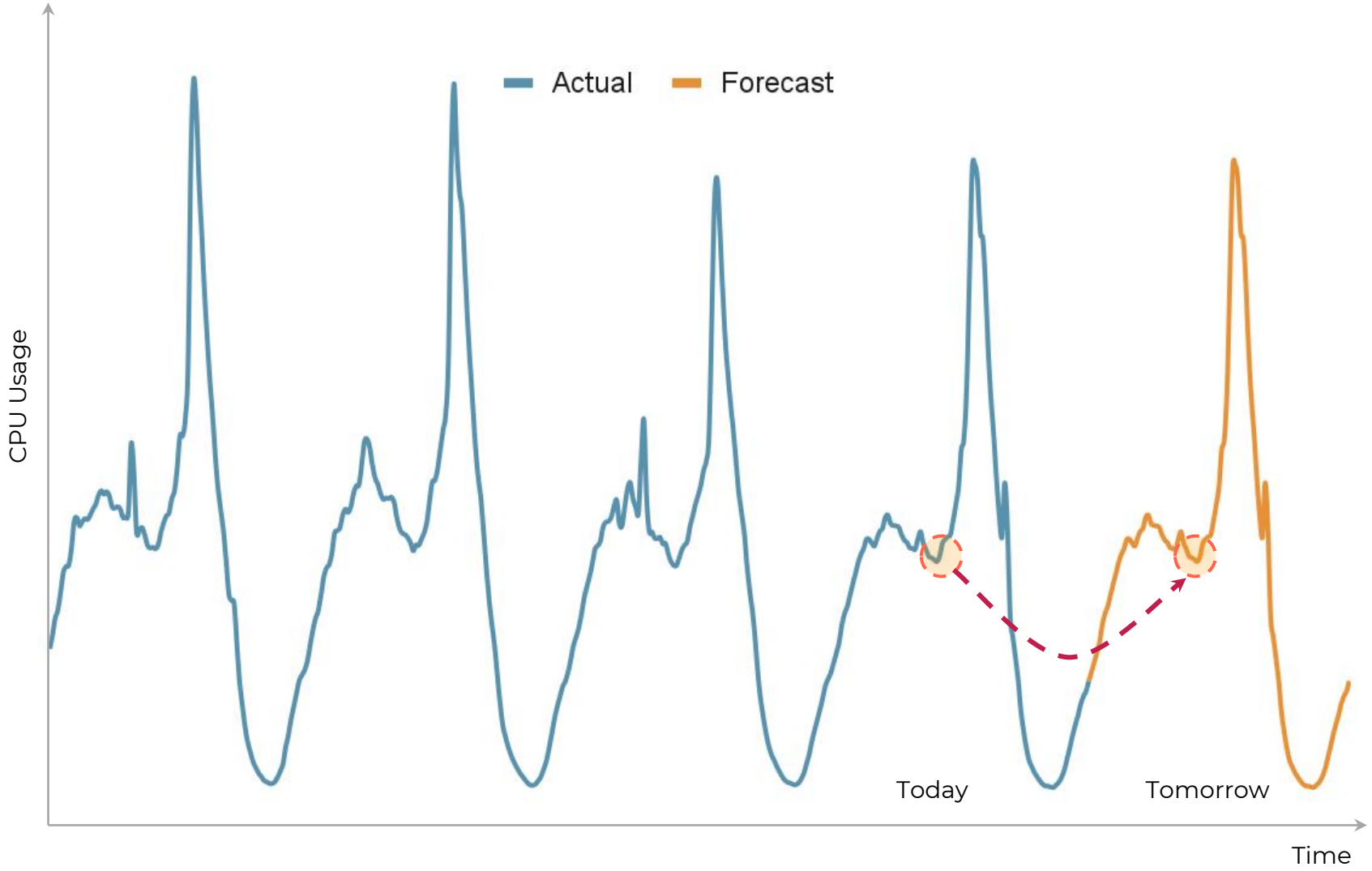


CPU Utilization Patterns of Services

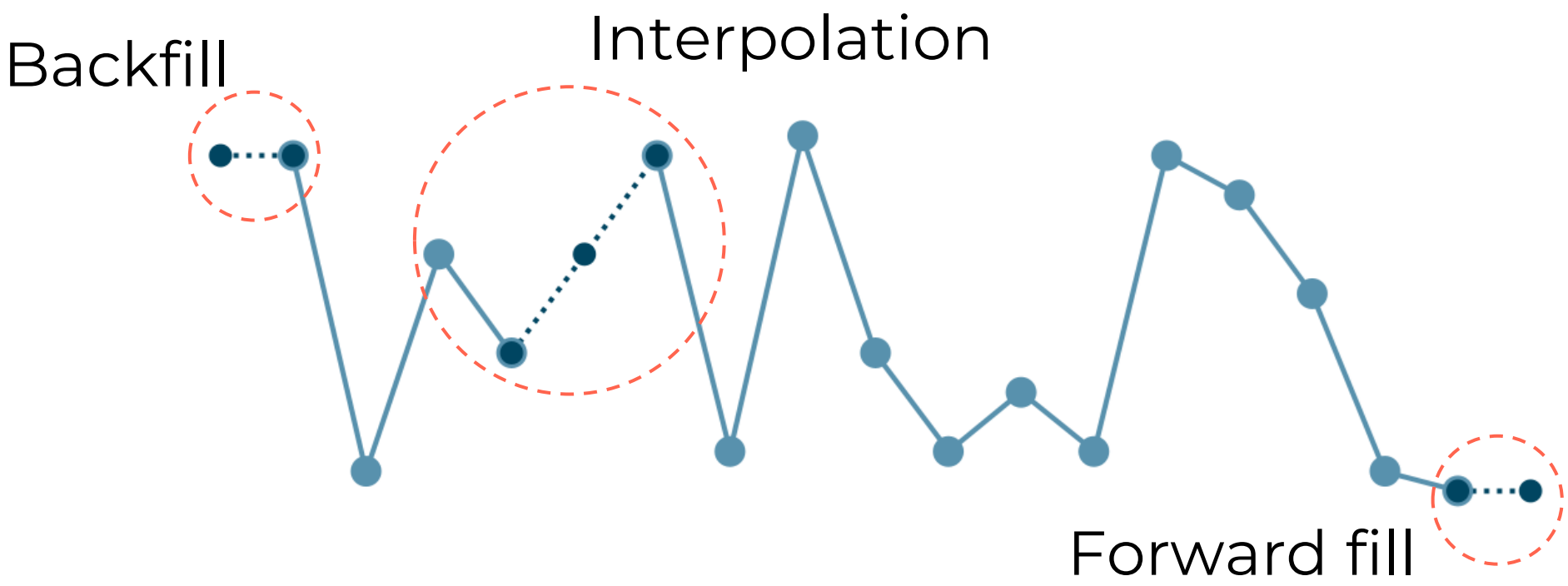
- Periodic (Concurrent users, periodic functions)
- Multiple peaks (Lunch, evening, time zones)
- Seasonality (Weekends, campaigns, holidays)



Naive Model for Forecasting

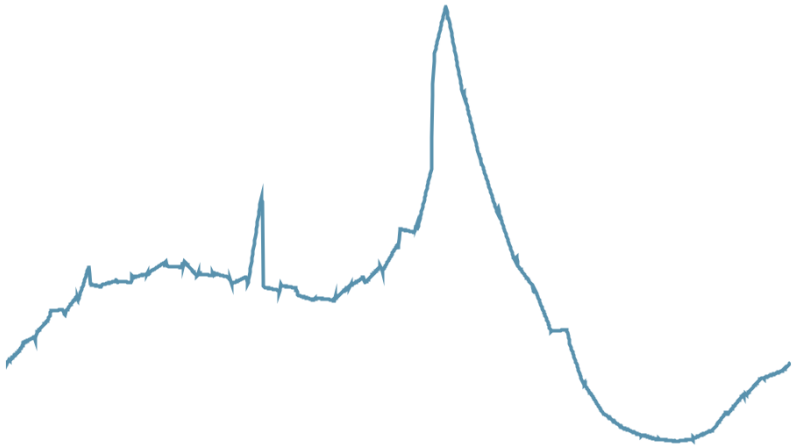


Filling Data Gaps

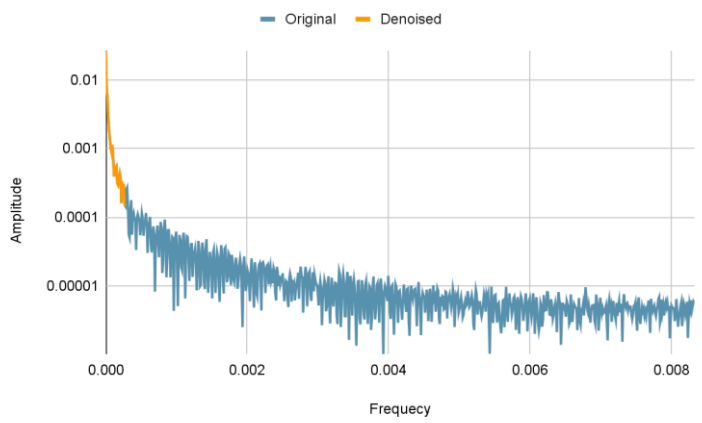


Fourier Transform

Denoise and extrapolate periodic data

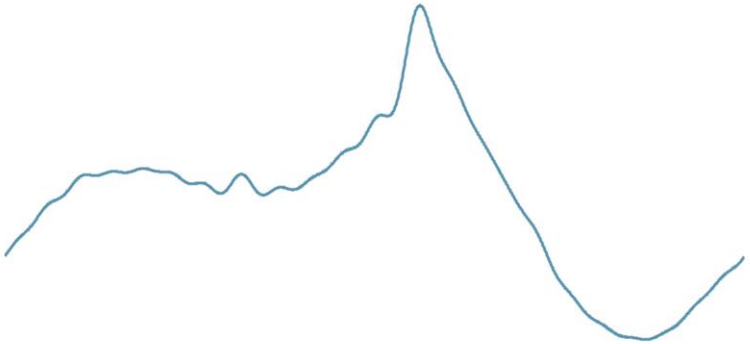


Convert to frequency domain

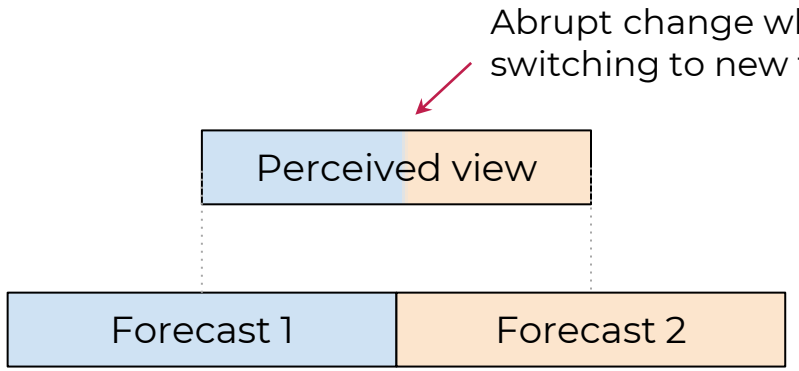


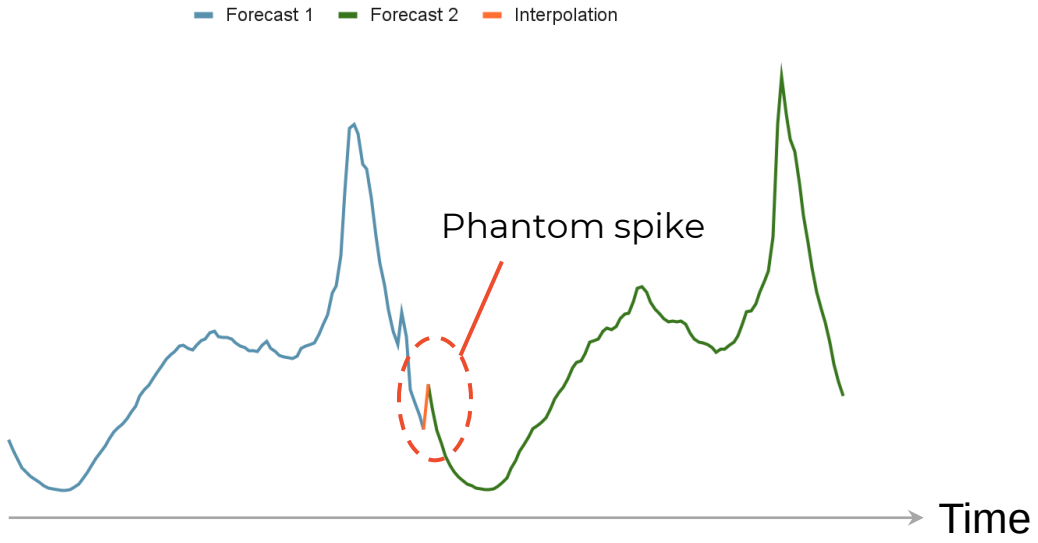
Remove low amplitudes

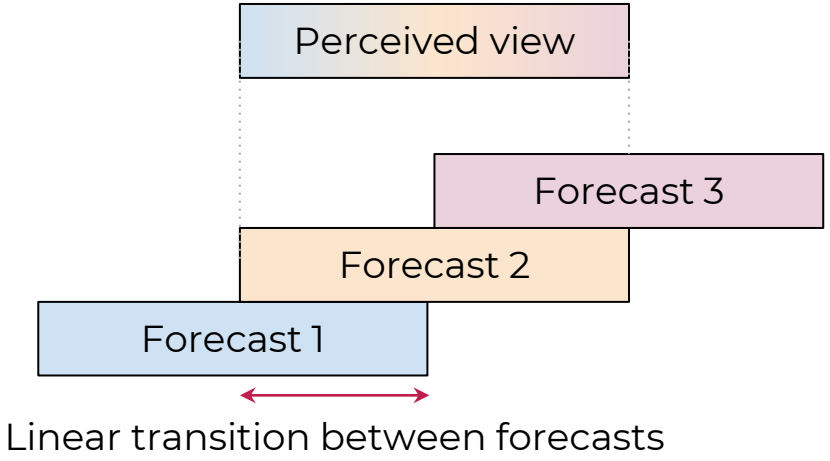
Convert back to time domain

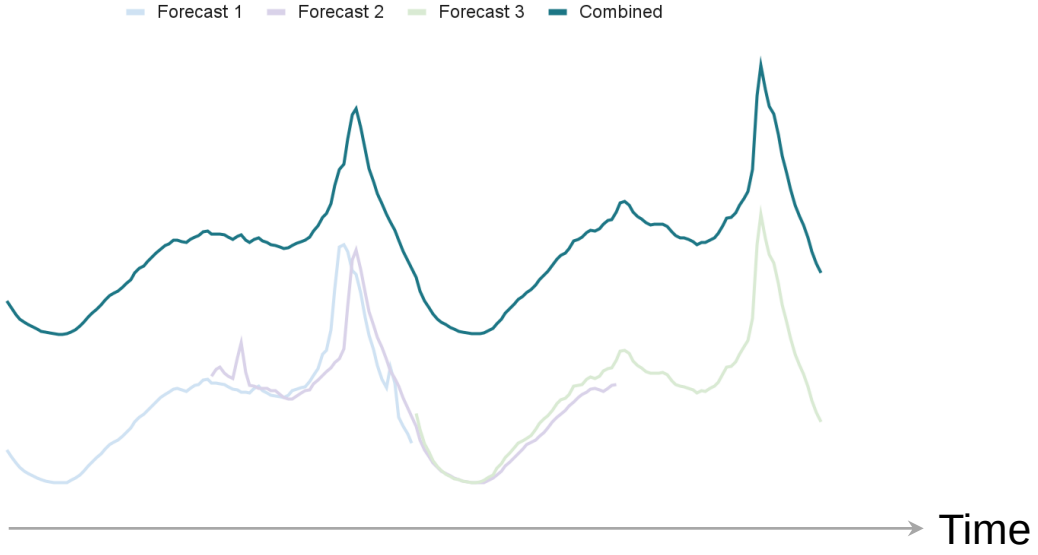


Bridging Forecasts of Different Periods

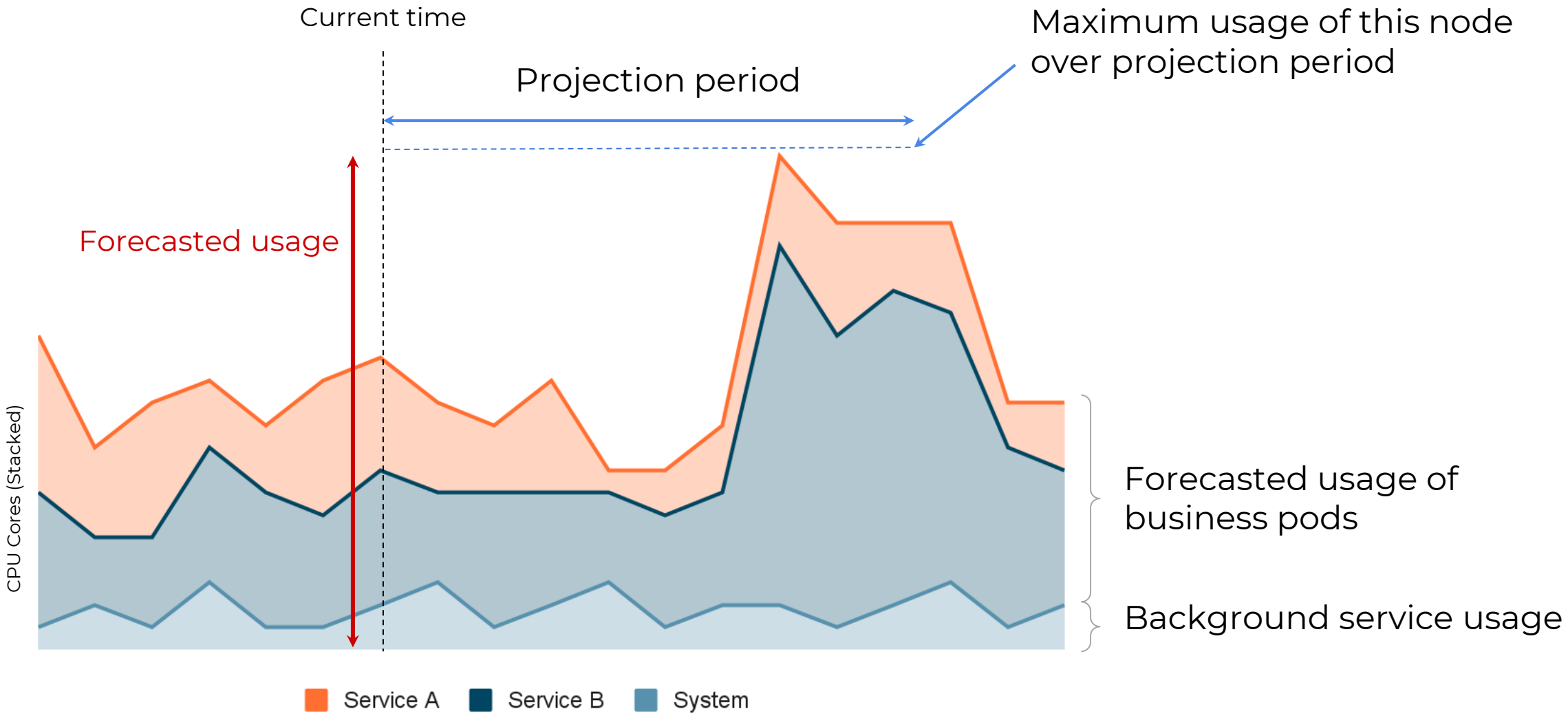




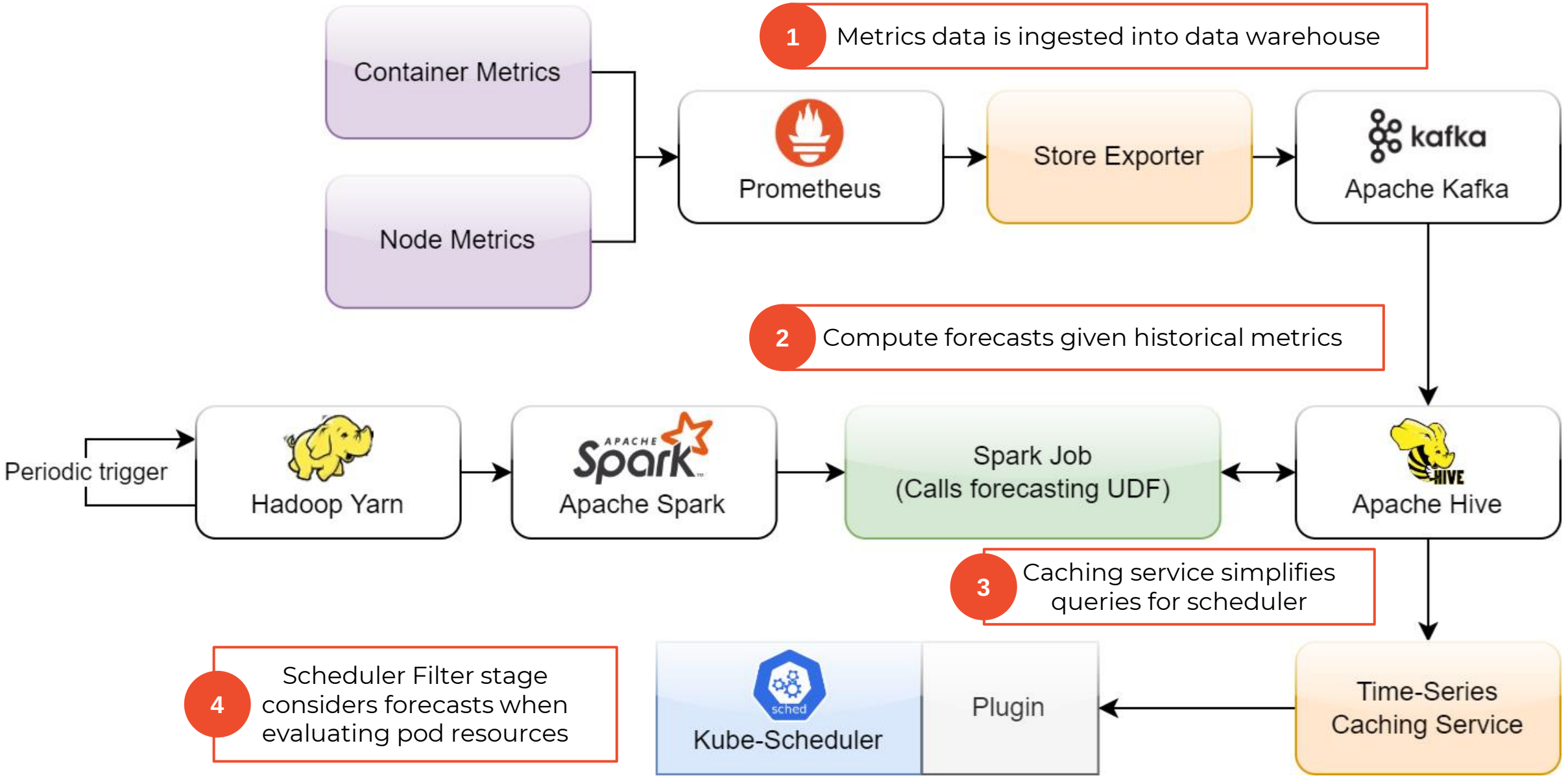




Using Forecasts in Scheduling



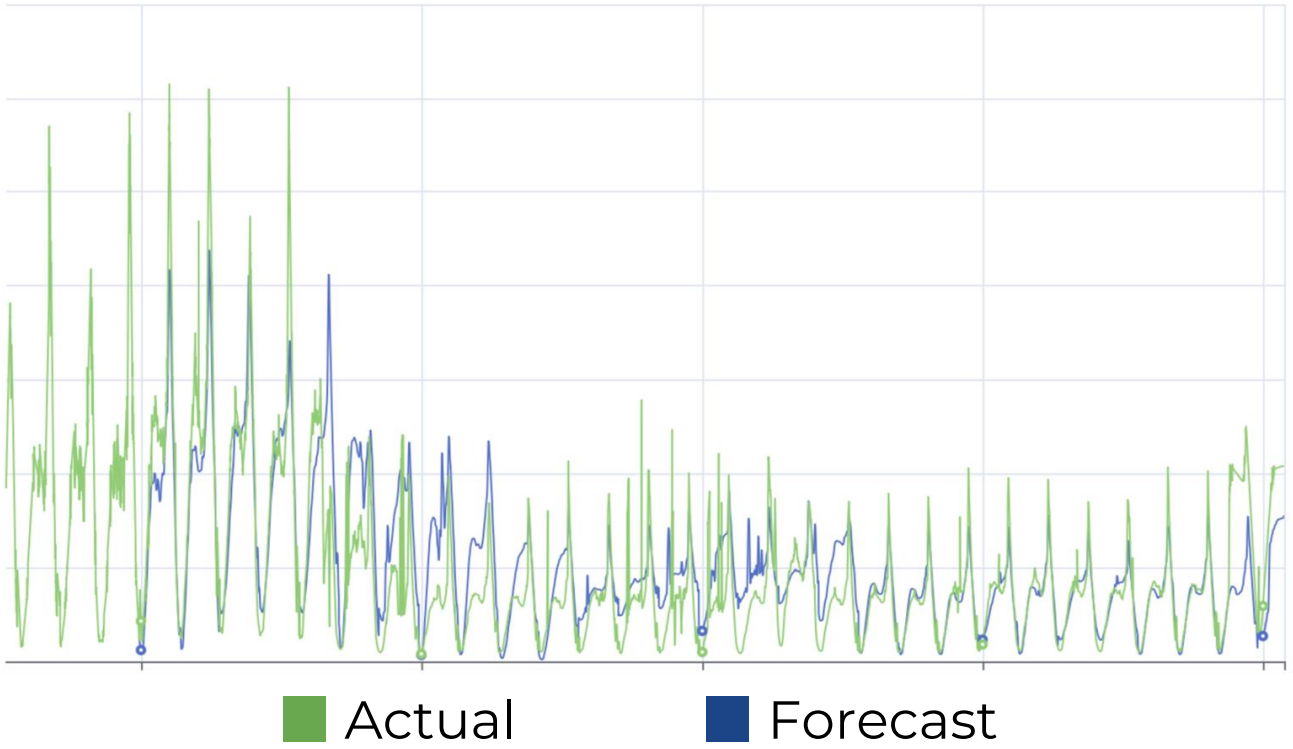
A Data-Driven Architecture with K8S



Better Forecasts?

Advanced Models to Improve Forecasting

Limitations of Naive Model



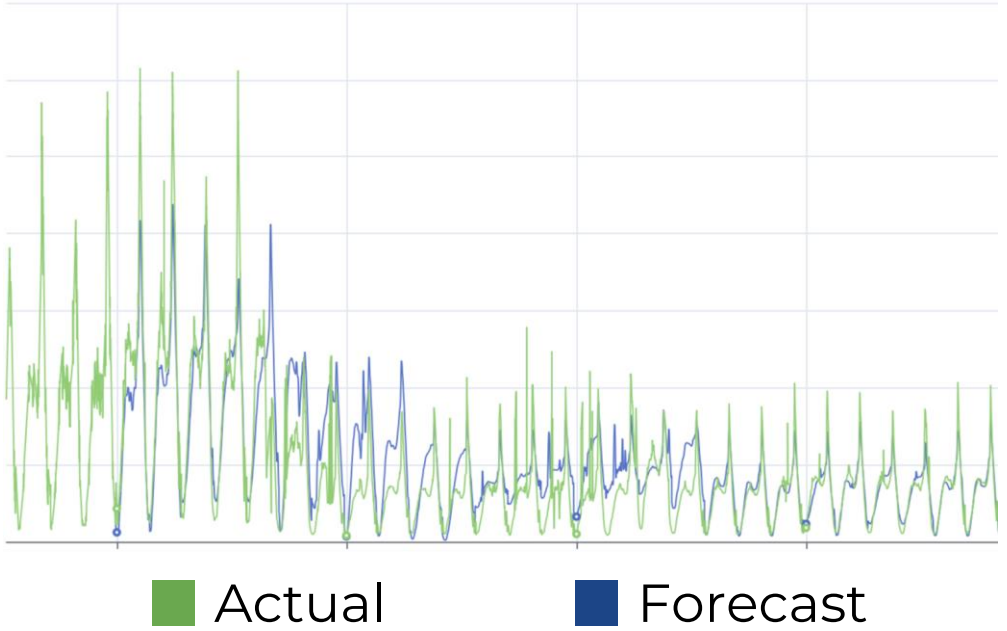
Does not react to recent changes in trend quickly

Does not handle seasonality of long durations

Cannot be pre-trained

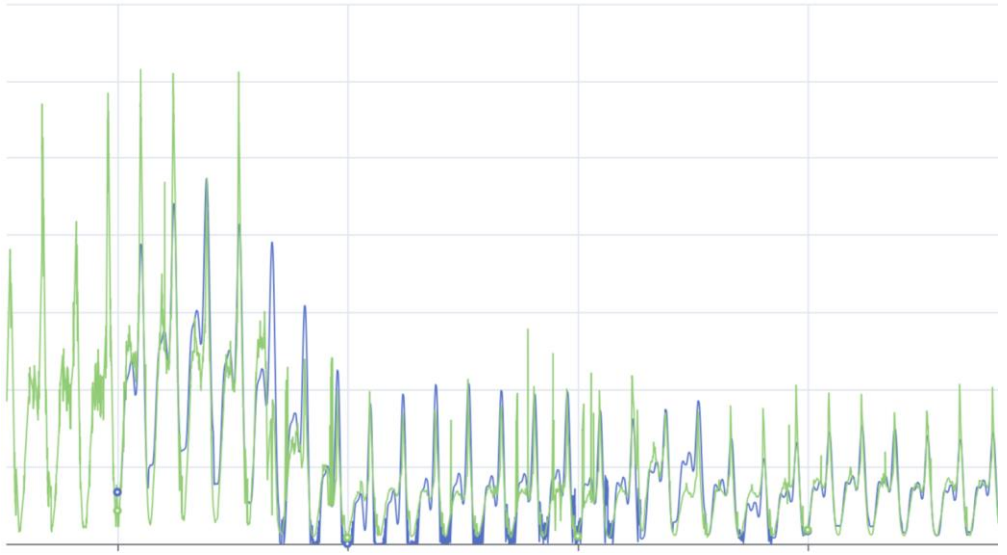
Statistical Models for Forecasting

Naive FFT



PROPHET

https://github.com/facebook/prophet



Machine Learning is Popular Now

NLinear

DeepAR

NBeats

GRU

Linear Regression

Chronos

RNN

Transformers

CNN

LSTM

iTransformer

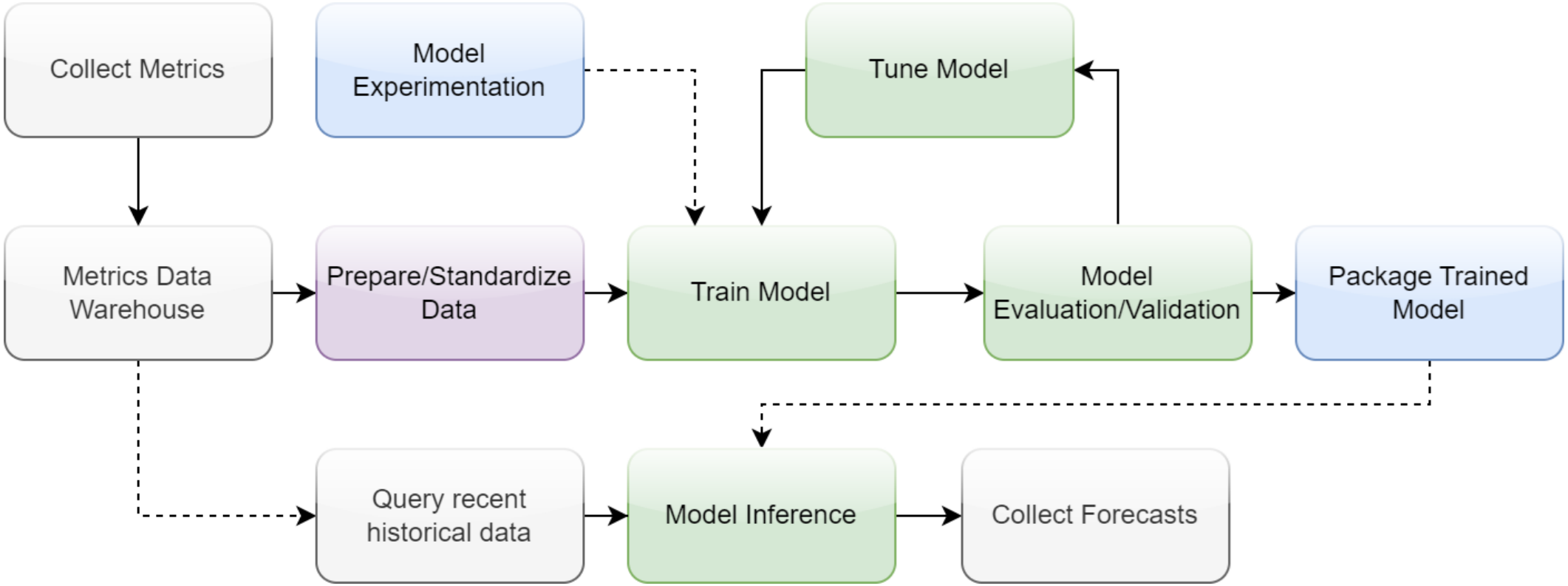
GPT

LLM

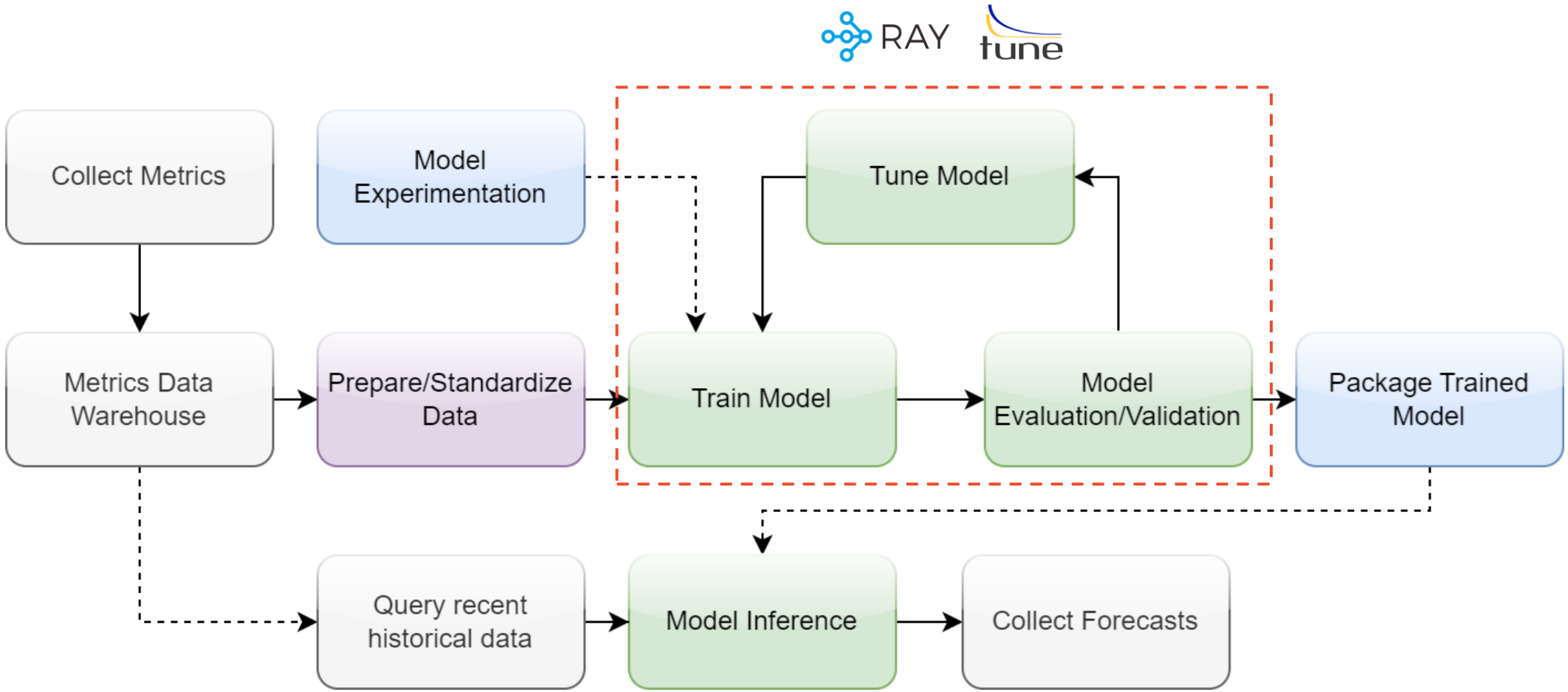
TimesFM

Autoformer

Training Machine-Learning Models



Model Tuning



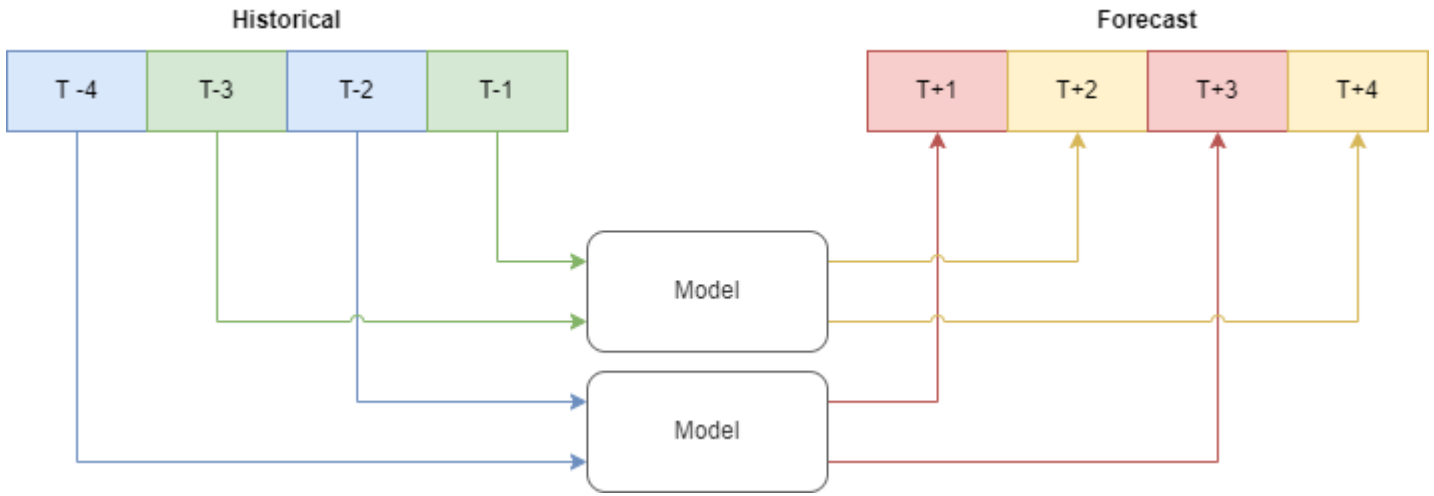
Reducing Model Inference Cost

Large models take significant memory and compute resources

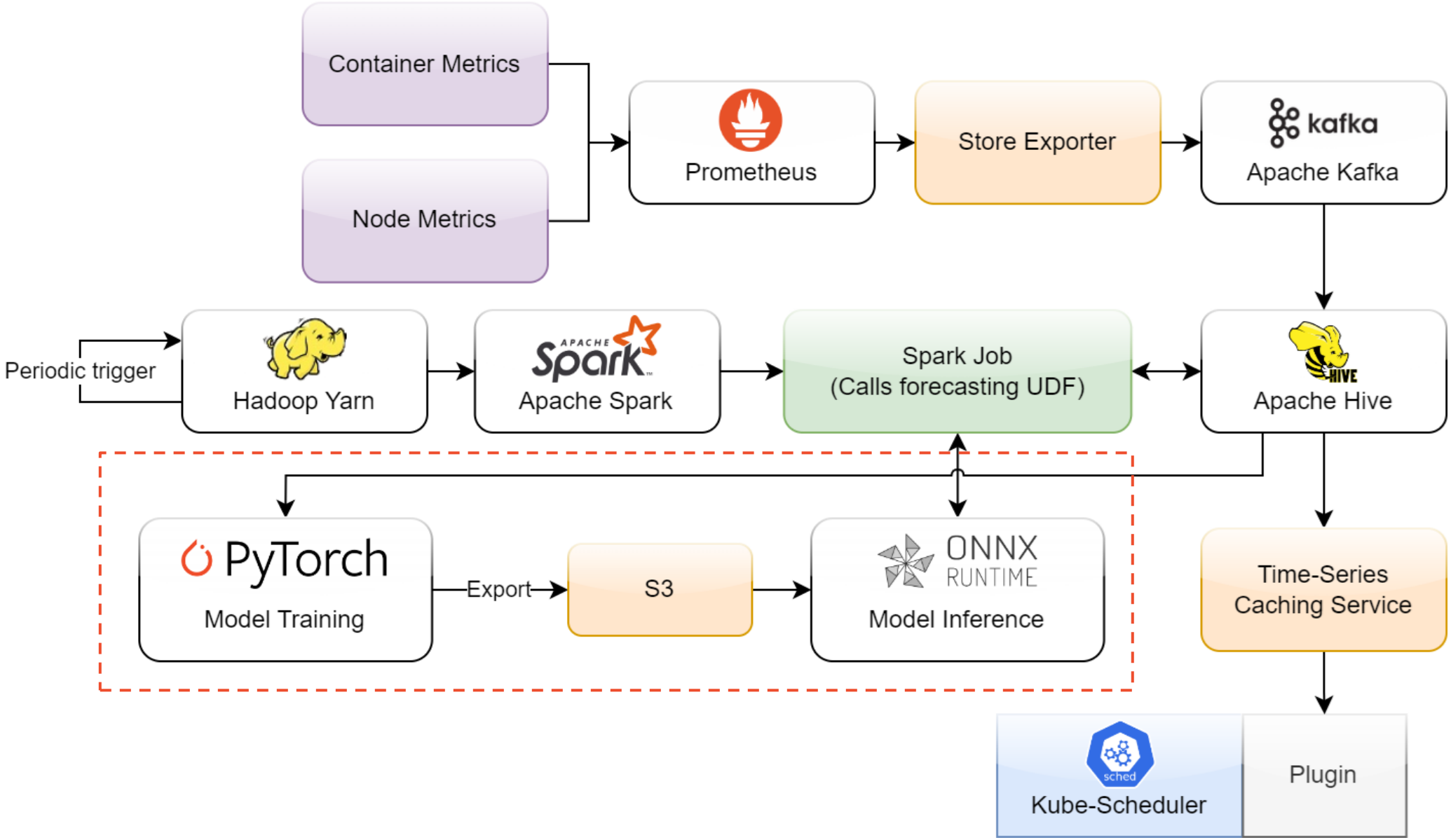
Expensive to run!

Optimization methods

- **Quantization** → Use lower precision (e.g. bfloat16)
- **Strided window** → Use multiple smaller model invocations for wider forecast

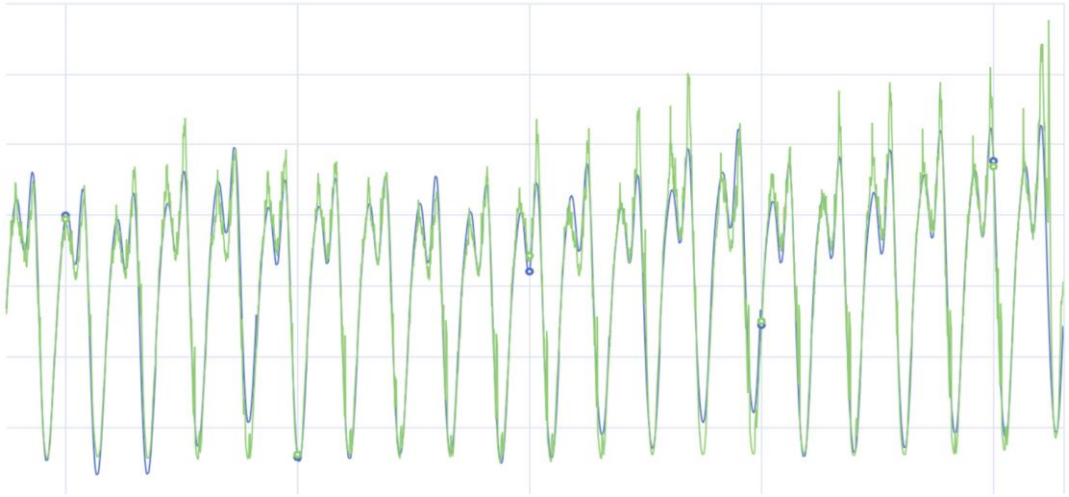


Forecasting with Machine-Learning Models



Comparison between Models

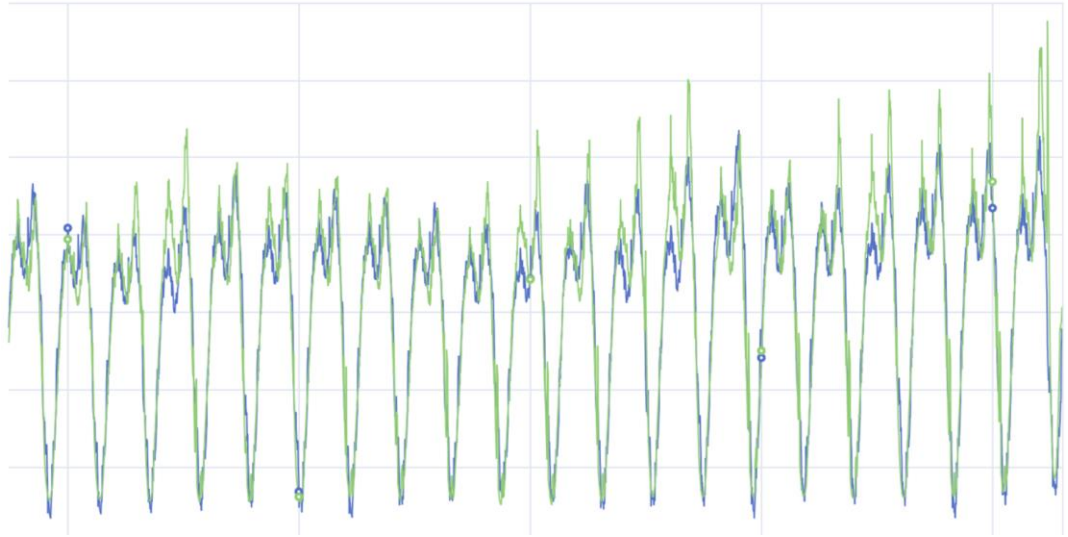
PROPHET



PyTorch

+

NLinear Model



Statistical or Machine Learning?

	Naive	Prophet	Machine Learning
Training/Fit Cost	None	Seconds	(Many) Minutes
Inference Cost (per series)	Seconds	Minutes	Seconds
CPU Time for forecasting thousands of series	Low	High (Each series inferred separately)	Moderate (Pre-trained model can infer multiple series)
Infrastructural Cost	Low	Moderate (For bulk inference within a reasonable timeframe)	High (For training within a reasonable timeframe)
Forecast accuracy	Moderate	High	High

Going Further

Using Forecasting to Improve Resource Density

The Traditional Approach of Overselling

Doing 2x CPU oversell

```
apiVersion: v1
kind: Pod
metadata:
  name: service
spec:
  containers:
  - name: service
    image: shopee
    resources:
      limits:
        memory: "1Gi"
        cpu: "2"
      requests:
        memory: "1Gi"
        cpu: "1"
```

Expectation...

- Set resource requests and limits based on what is needed
- Resource requests factor in peak/campaign periods

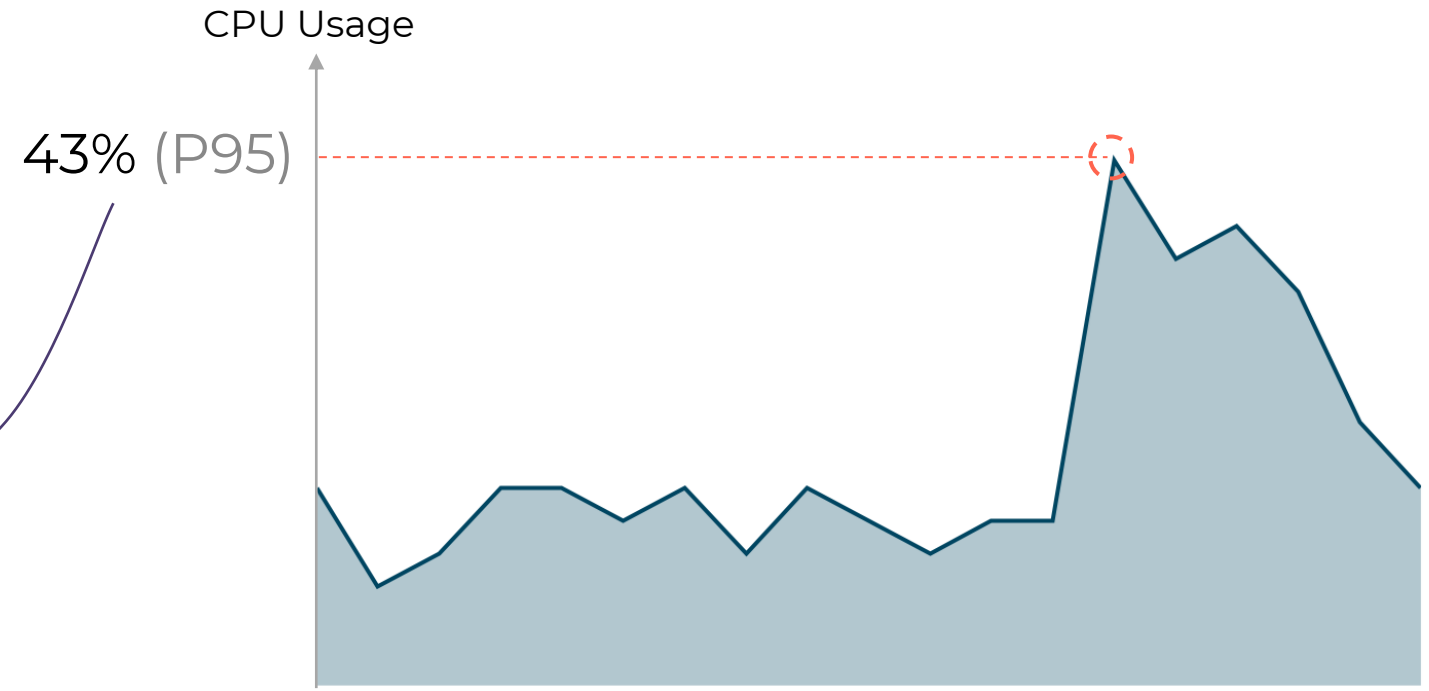
Reality...

- Resources **overprovisioned**
- Resources **not revised frequently**

Automating Oversell of Business Services

Infer oversell ratio from forecasted usage

```
    limits:
      memory: "1Gi"
      cpu: "2000m"
    requests:
      memory: "1Gi"
-     cpu: "1000m"
+     cpu: "860m"
```

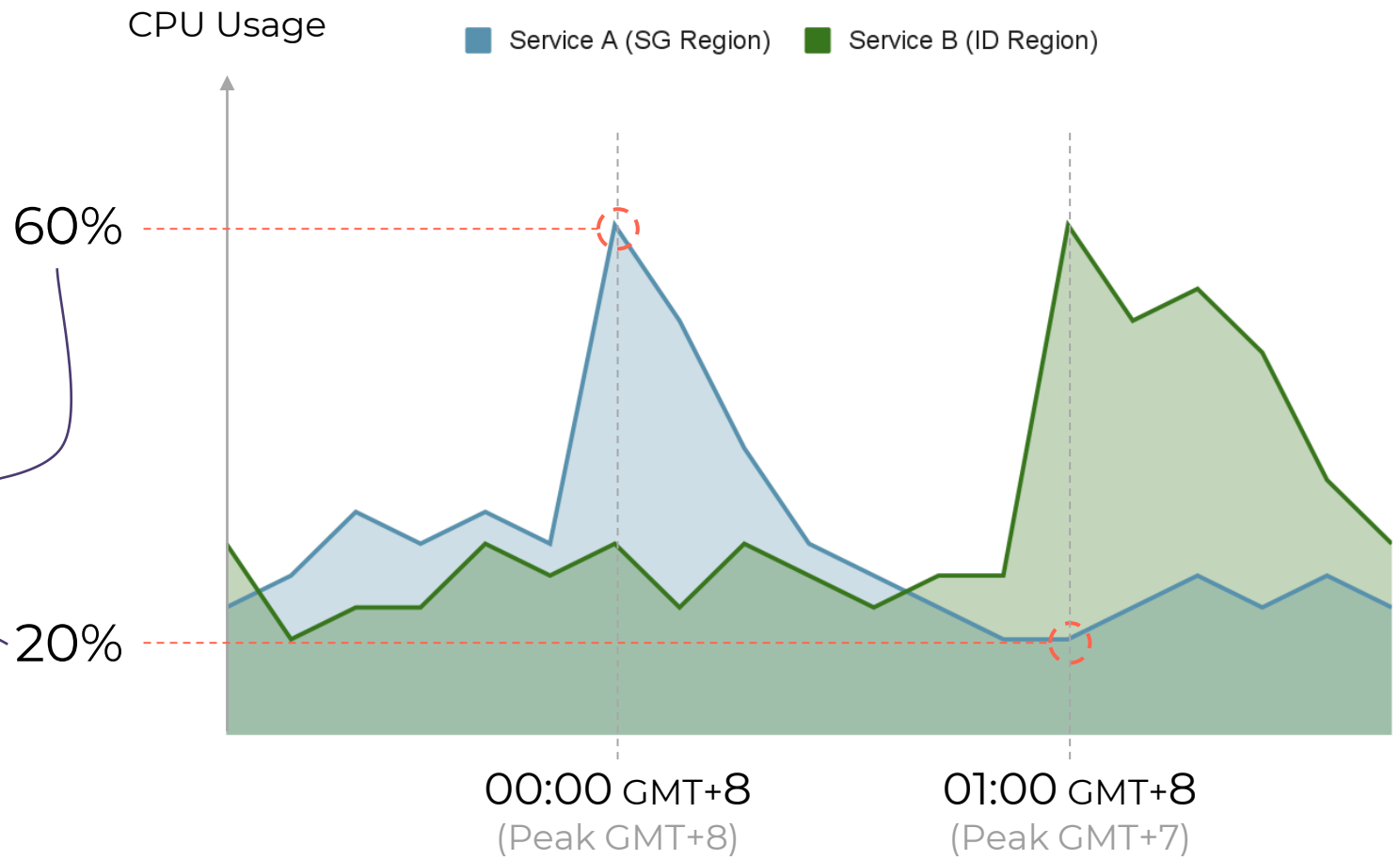


Oversell Across Time

Overlap services catering for different time zones

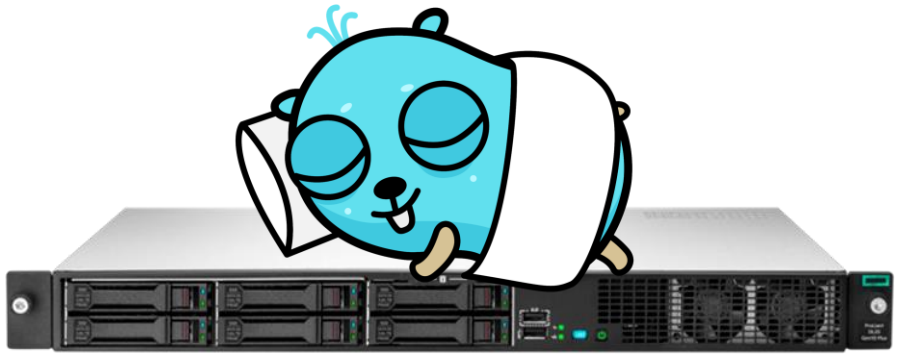
Service A

```
      requests:
    memory: "1Gi"
+   shopee.com/usage-1600: "1200"
+   shopee.com/usage-1700: "400"
```



Saving Electrical Costs

Put CPU cores to low-power state on idle machines



C-State C0

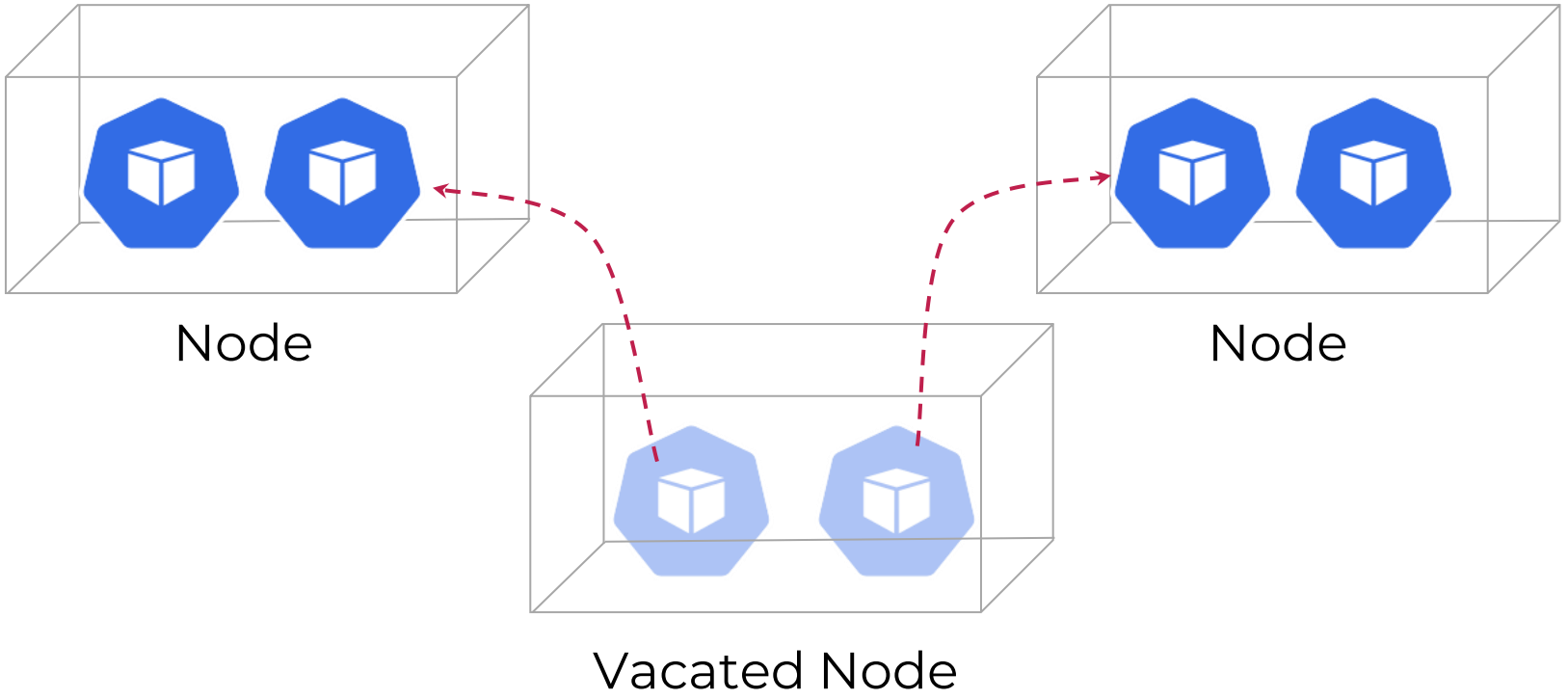
C-State C1

Power savings up to **60W** per machine

No business pods should be on node in power-saving state!

Maximizing Node Density

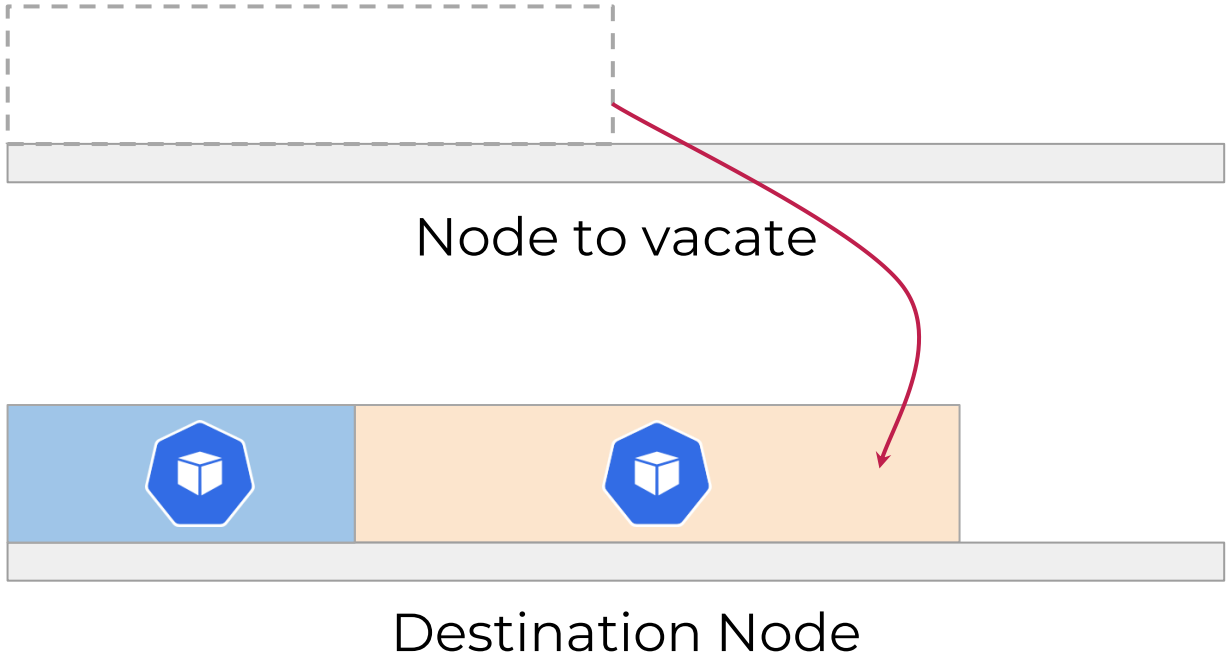
Redistribute pods to vacate and compact nodes



Scheduling Simulation for Node Selection

Pick node with lowest allocated ratio and iterate through pods on the node

1. Simulate scheduling of pod to other nodes to validate vacating this node is feasible
2. Reserve resources on target nodes
3. Evict and relaunch pod
4. Repeat until node is vacated

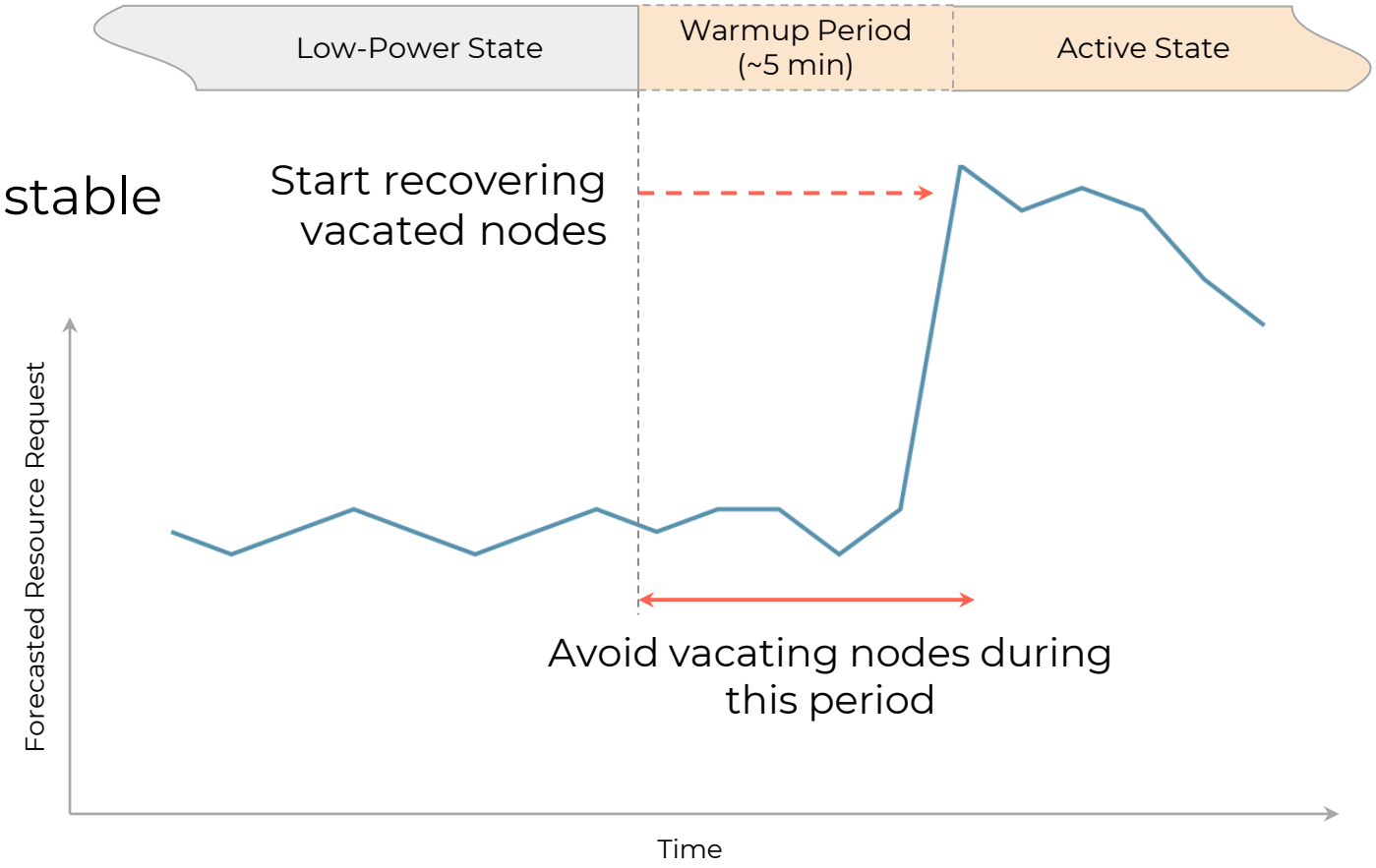


Avoiding Starvation via Forecasting

Scale Up Triggers

- Deployments
- SRE operation
- **HPA**
- **VPA**

Forecastable



Takeaways

- Historical metrics can be exploited to increase density of resource utilization
- Machine learning is viable for forecasting metrics
- Forecasts allows for earlier reaction to changes that take time to take effect
  → New solutions become possible

Thank You!

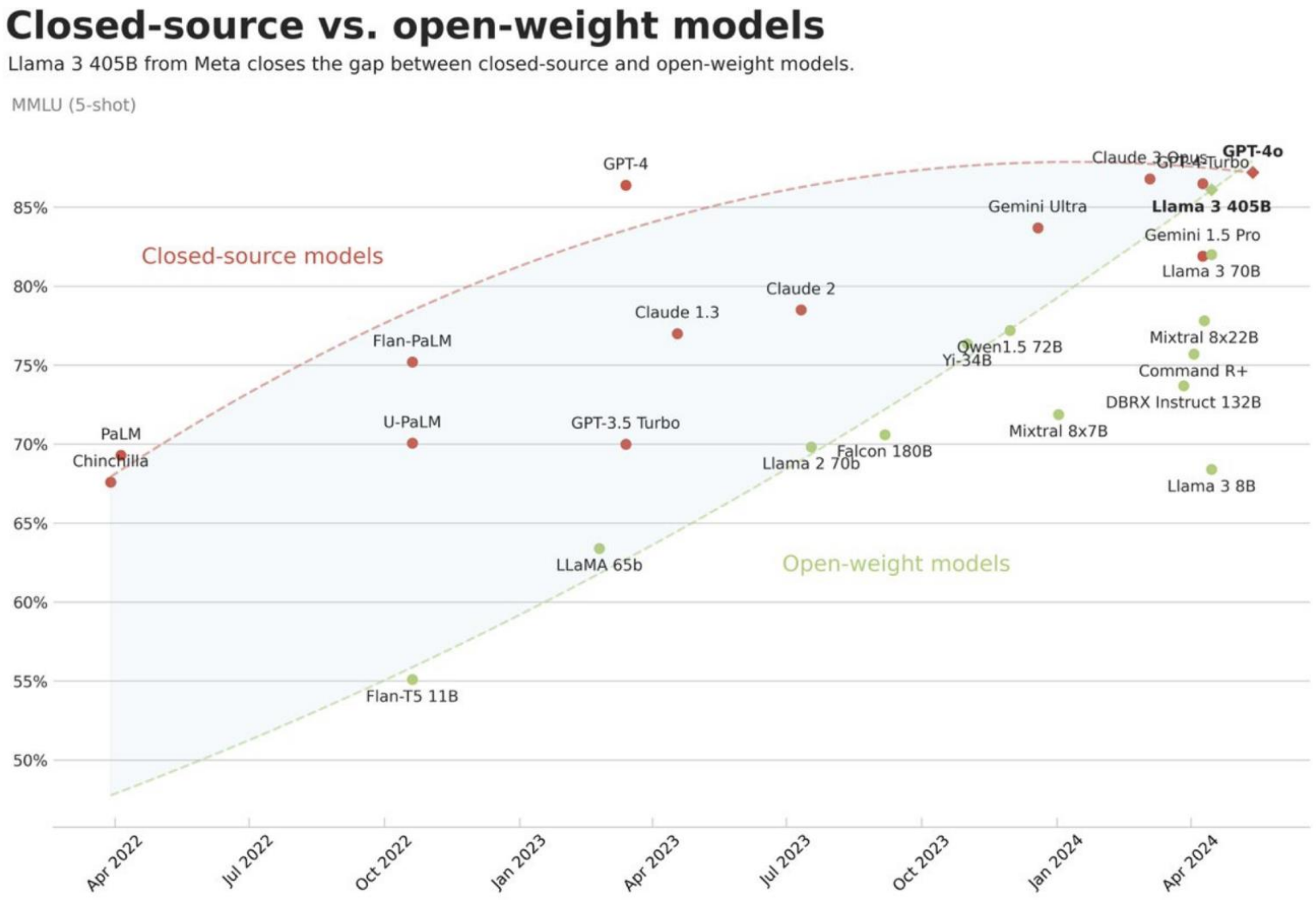
We are hiring!


DaoCloud
BYTEDANCE OPEN SOURCE
Shopee
WasmEdgeRuntime
Cloud Native AI Meetup
10th July, 2024
18:30-20:30 | Singapore
Portable AI / LLM Inference across CPUs/GPUs
Alan Poon QA engineer Redfin Tech

Agenda

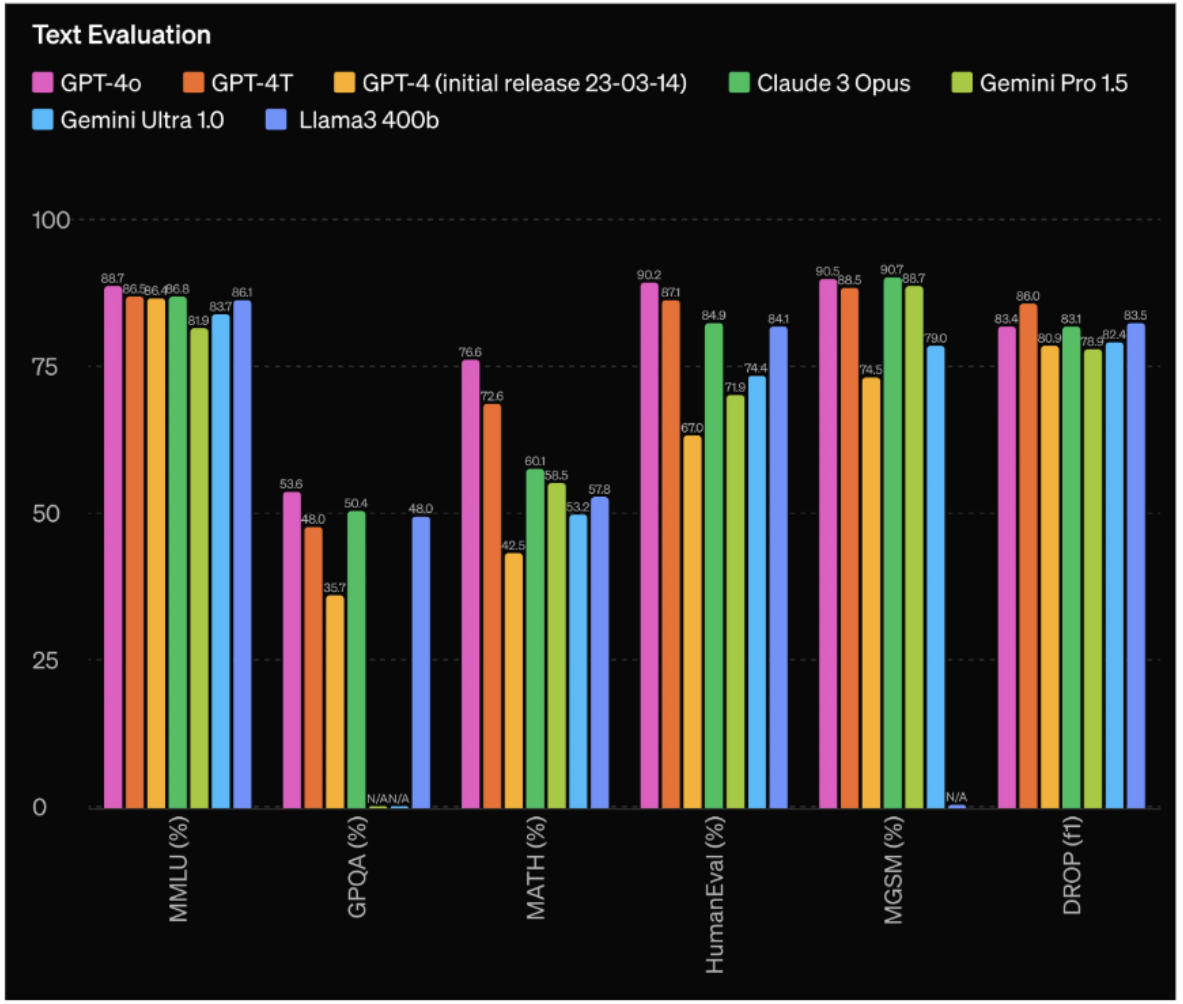
- Why Wasm-based LLM Inference
- Demo: Build a RAG based LLM applications
- Bonus: Use container tools to manage the LLM workloads

Why use open source LLMs



Why use open source LLMs

	Meta Llama 3 70B	Gemini Pro 1.5 Published	Claude 3 Sonnet Published
MMLU 5-shot	**82.0**	81.9	79.0
GPQA 0-shot	39.5	**41.5** CoT	38.5 CoT
HumanEval 0-shot	**81.7**	71.9	73.0
GSM-8K 8-shot, CoT	**93.0**	91.7 11-shot	92.3 0-shot
MATH 4-shot, CoT	50.4	**58.5** Minerva prompt	40.5

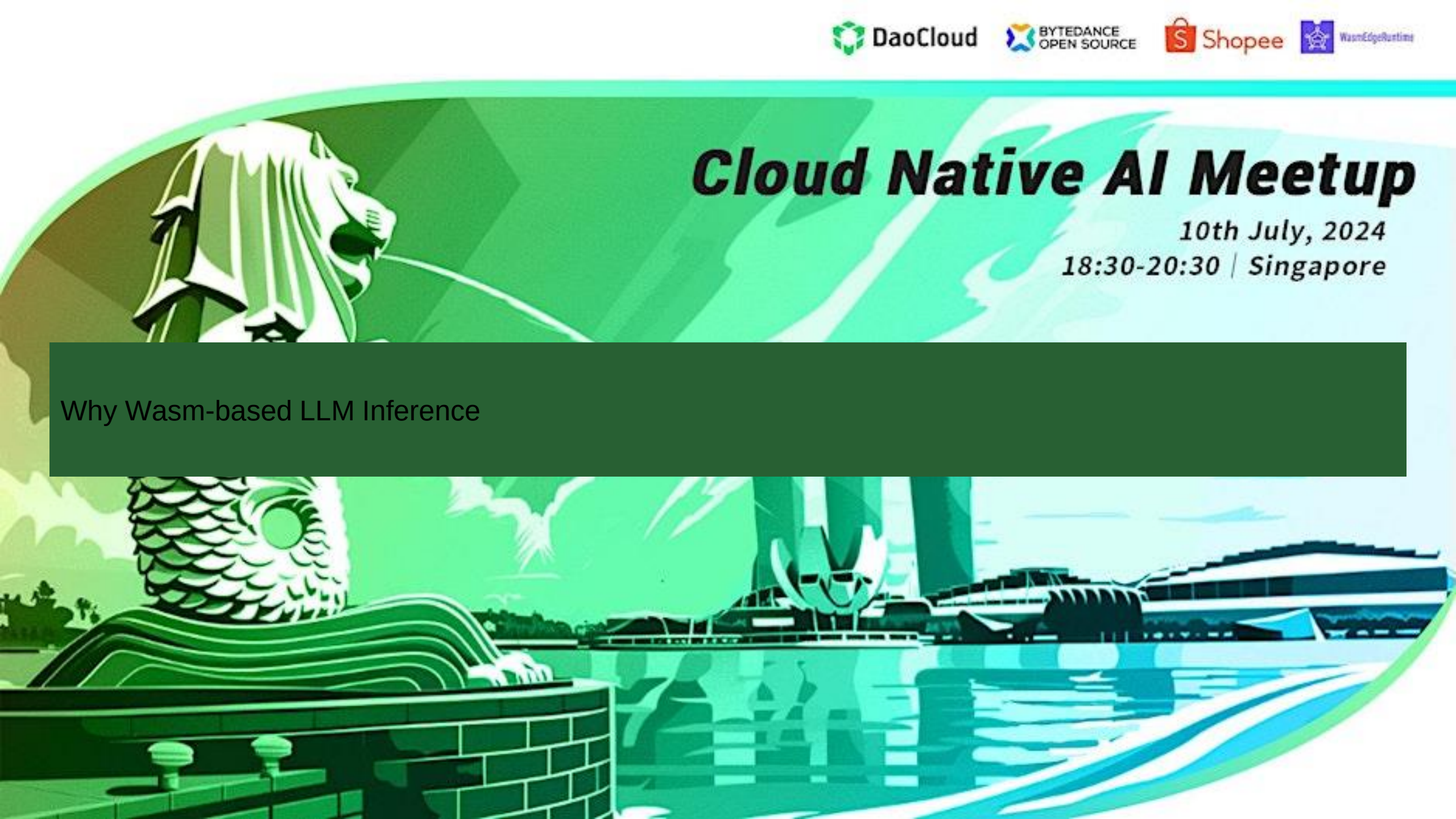


Open source LLM is as good as OpenAI, but with

Privacy

Have full control of your data

Less cost


DaoCloud
BYTEDANCE OPEN SOURCE
Shopee
WasmEdgeRuntime
Cloud Native AI Meetup
10th July, 2024
18:30-20:30 | Singapore
Why Wasm-based LLM Inference

Cloud Native AI Meetup

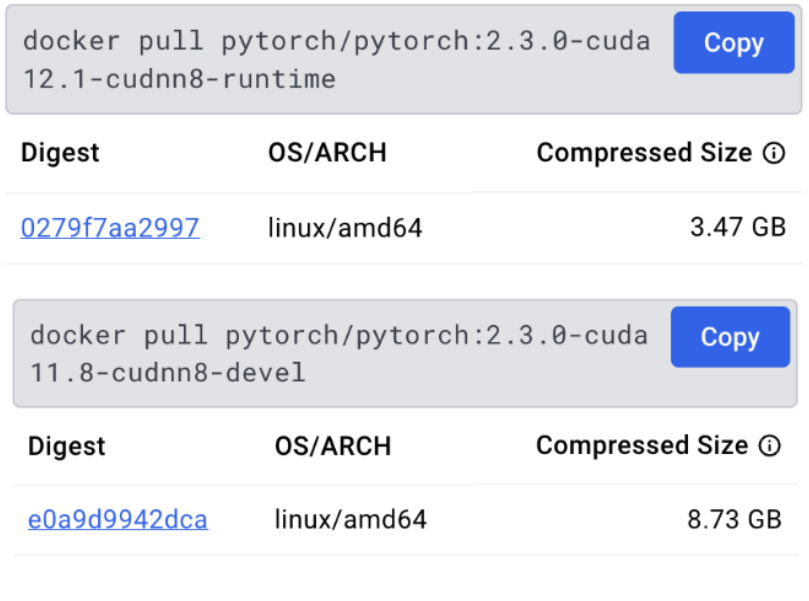
Why use Wasm for LLM workloads?

Very lightweight and fast

- Entire runtime + app is less than 30MB
- Runs well on Raspberry Pi and Jetson devices
- Full native GPU and hardware accelerator support

Cross platform across a wide range devices and drivers. Runs at native GPU speed

- Nvidia CUDA
- Apple M chips with metal
- Advanced CPUs
- ARM NPUs



Introduce LlamaEdge

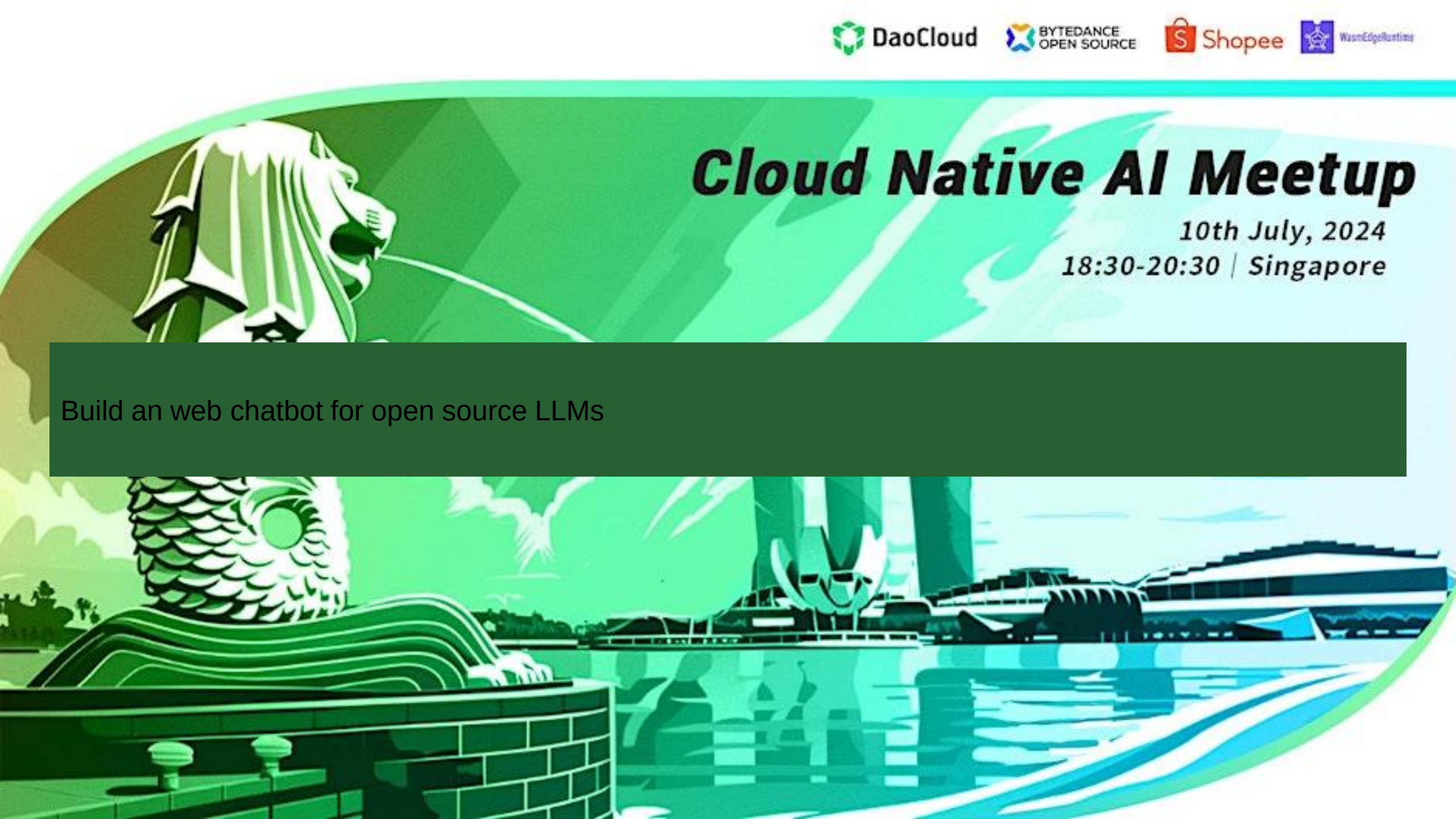
- A single portable and deployable app
    - Improve efficiency
    - Simplify development and deployment workflow
    - Improve security
- No Python dependency, based on llama.cpp and WasmEdge
- Use Rust to extend LlamaEdge components!
- Supports a wide range of LLMs, VLMs, MoE models on Hugging face out of the box
- Single command to install and run as an unprivileged user – no daemon, no sudo
- Can be managed and orchestrated directly by container tools and k8s
- OpenAI API compatible
    - i.e., highly integrated OpenAI Assistant API

https://github.com/LlamaEdge/LlamaEdge

LlamaEdge is a developer platform

- Use PyTorch / llama.cpp to fine tune a model
- Use LangChain / LlamaIndex to create the knowledge base or vector collection
- Use LlamaEdge to run the service!

https://github.com/LlamaEdge/LlamaEdge/


DaoCloud
BYTEDANCE OPEN SOURCE
Shopee
WasmEdgeRuntime
Cloud Native AI Meetup
10th July, 2024
18:30-20:30 | Singapore
Build an web chatbot for open source LLMs

Run an open sourced model on you own device

bash <(curl -sSfL 'https://raw.githubusercontent.com/LlamaEdge/LlamaEdge/main/run-llm.sh') --interactive

Run an open sourced model on you own device

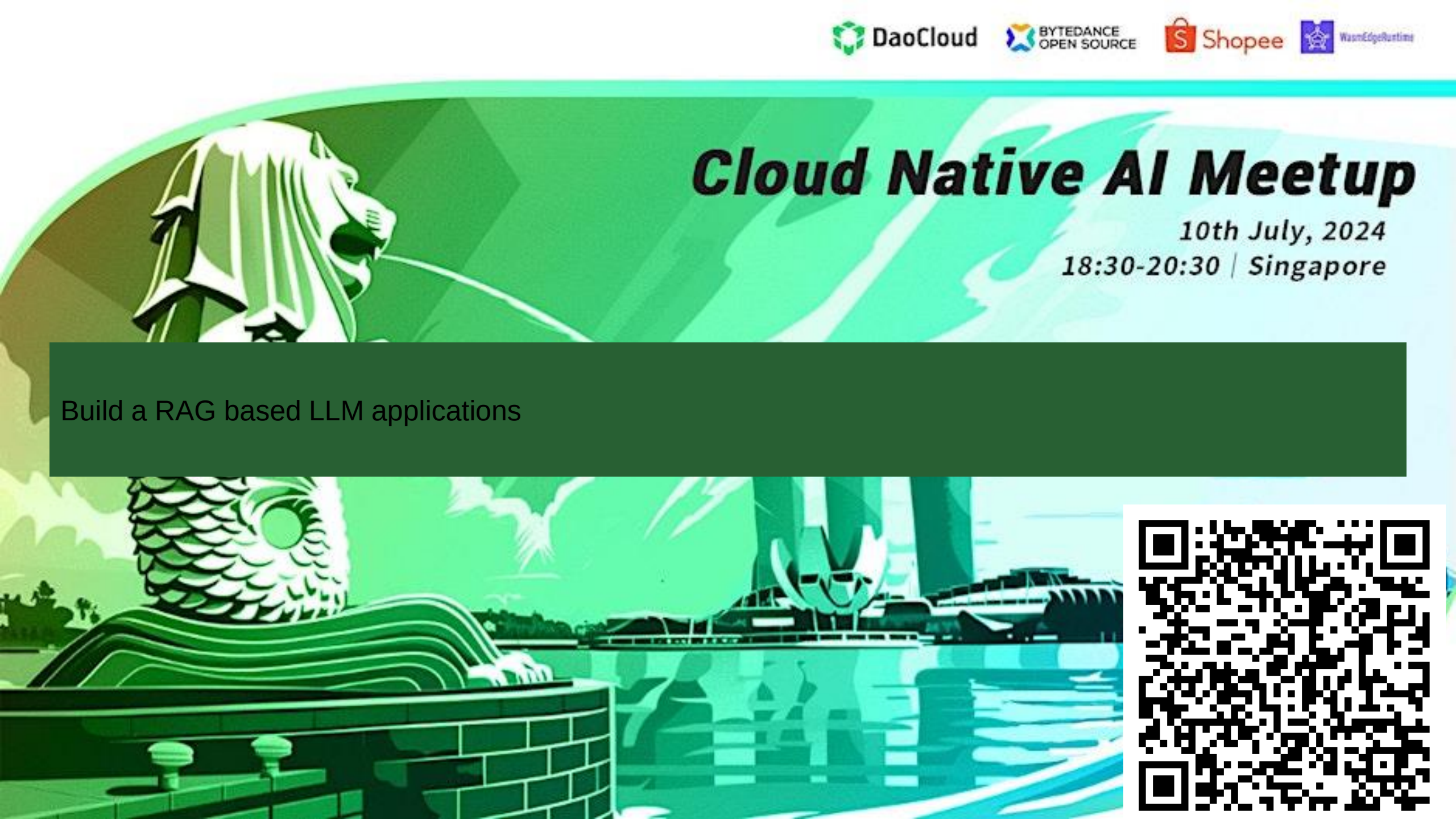
- Install the WasmEdge runtime
- Download the model
- Download the portable Wasm file for API server
- Run the API server for the model

Don't like the UI?

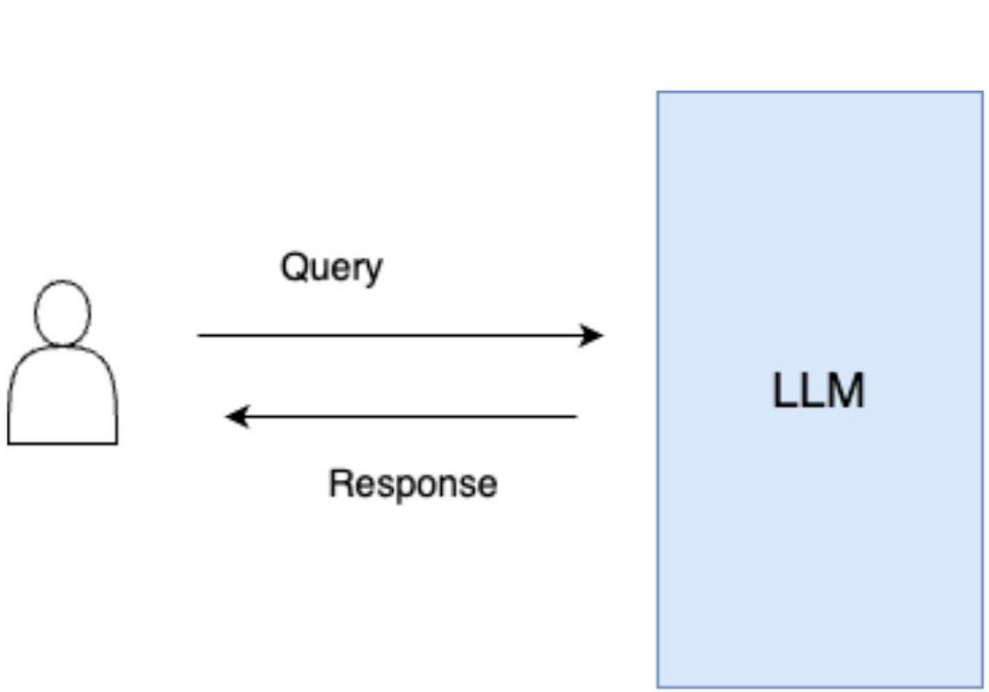
LlamaEdge is an openAI-compatible API server.

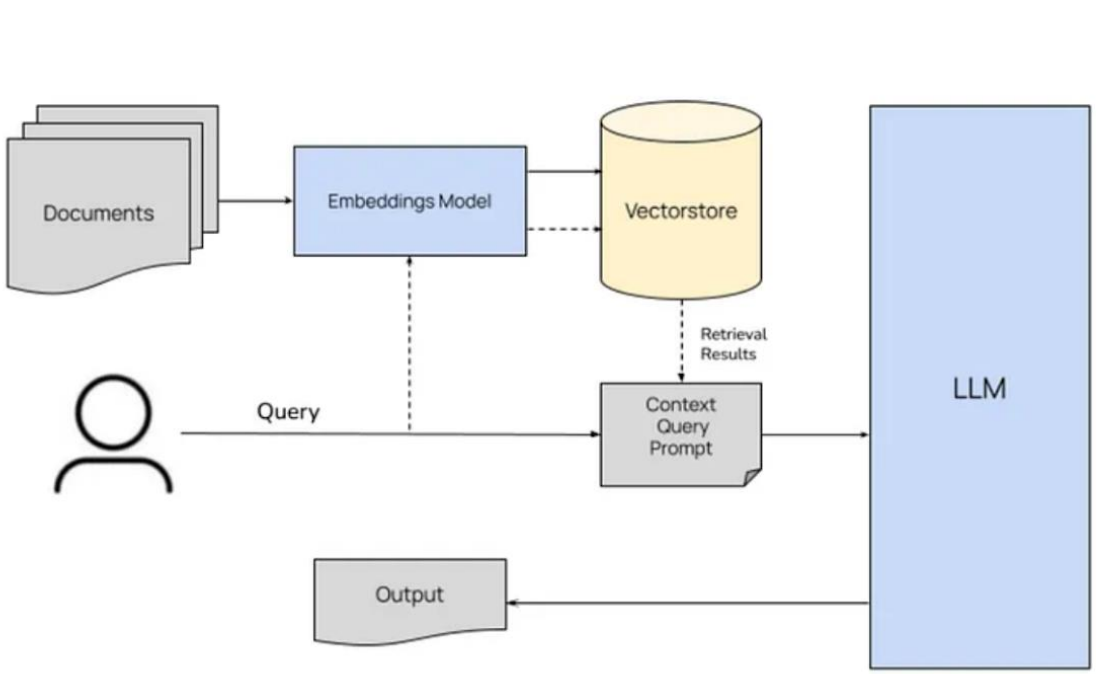
Try to integrate LlamaEdge with a more fancy WebUI!

a=


DaoCloud
BYTEDANCE OPEN SOURCE
Shopee
WasmEdgeRuntime
Cloud Native AI Meetup
10th July, 2024
18:30-20:30 | Singapore
Build a RAG based LLM applications

What is RAG?


Query
Response
LLM


Documents
Embeddings Model
Vectorstore
Retrieval Results
Query
Context Query Prompt
LLM
Output

What is RAG

Retrieval Augmented Generation (RAG) is a framework to solve the hallucination problem of LLM by attaching domain specific knowledge.

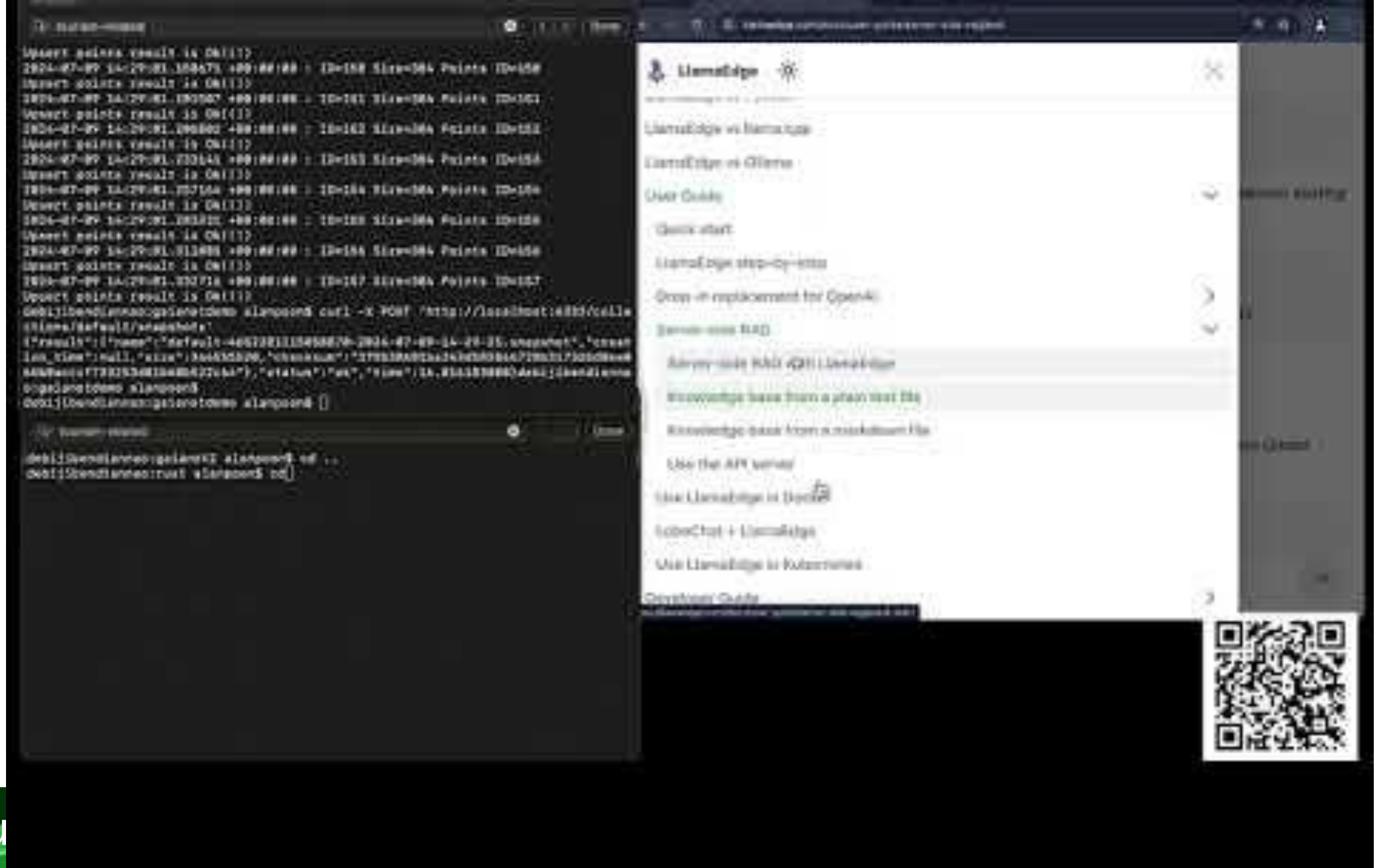
- A regular LLM
- An embedding model
- A vector DB
- An API server application

GaiaNet

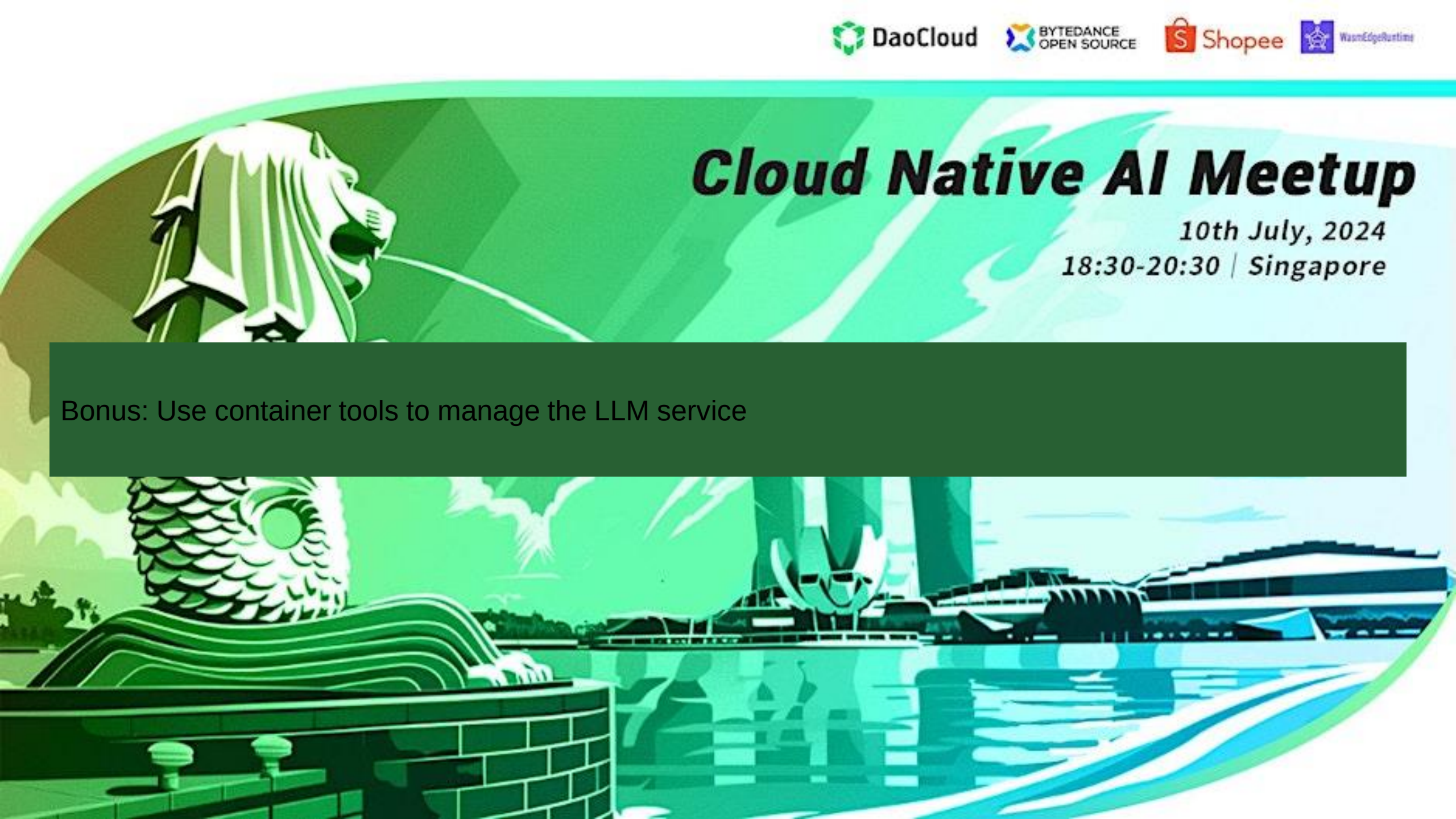
GaiaNet is a developer tool to build RAG-based application, which is build on top of LlamaEdge.

- WasmEdge as the LLM runtime
- Qdrant as the vector DB
- Support all open source LLMs and embedding models
- Customize the system prompt

Demo: Build a RAG application with a Paris guide

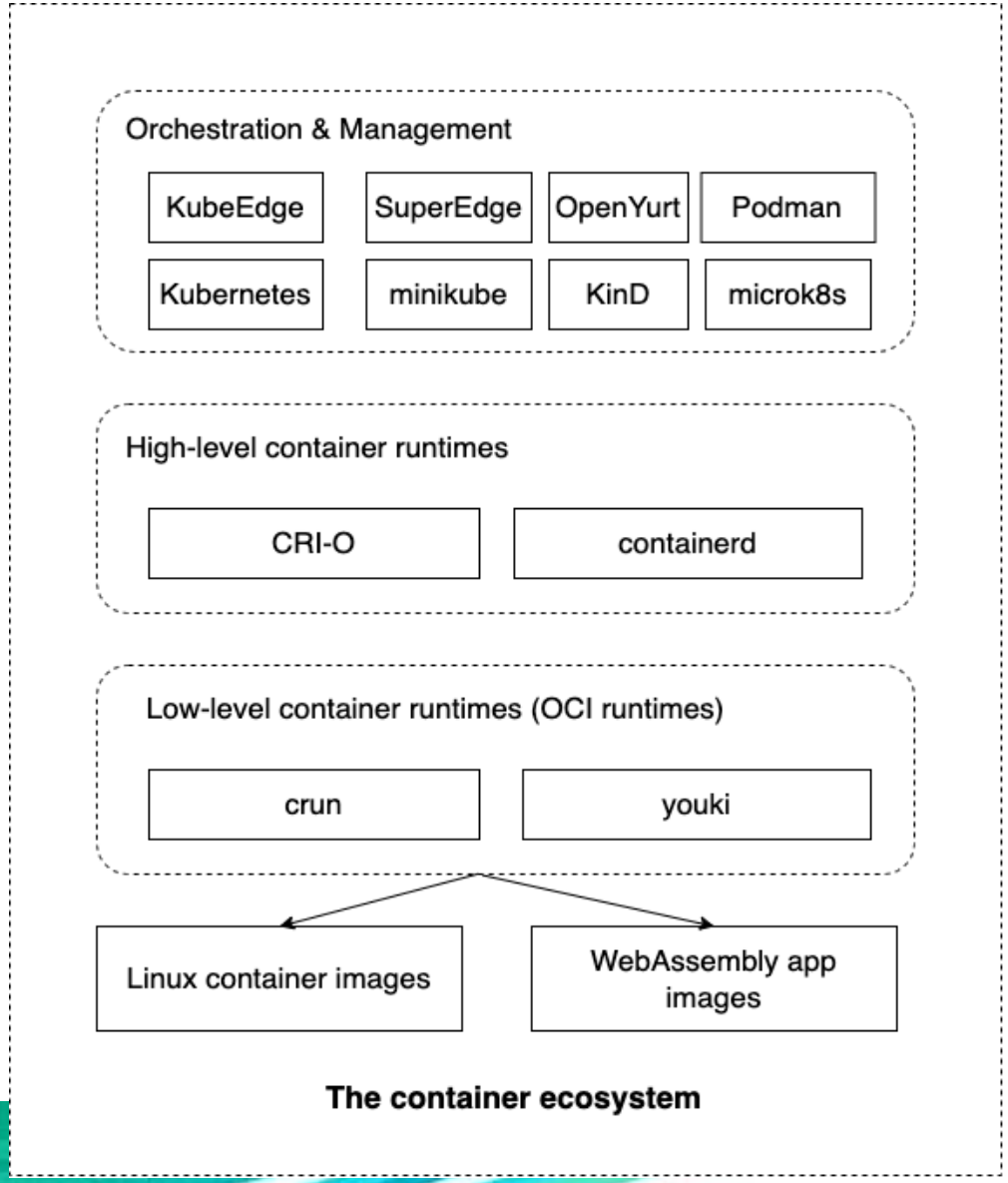


Clou


DaoCloud
BYTEDANCE OPEN SOURCE
Shopee
WasmEdgeRuntime
Cloud Native AI Meetup
10th July, 2024
18:30-20:30 | Singapore
Bonus: Use container tools to manage the LLM service

Portable and lightweight LLM workloads

K8s can copy and deploy the lightweight wasm file on any device with any GPU in the cloud and on the edge.



The container ecosystem

Resources

- LlamaEdge: https://github.com/LlamaEdge/LlamaEdge
- LlamaEdge docs: https://llamaedge.com/docs/
- GaiaNet docs: https://docs.gaianet.ai
- Container tools: https://wasmedge.org/docs/category/deploy-wasmedge-apps-in-kubernetes


DaoCloud
BYTEDANCE OPEN SOURCE
Shopee
WasmEdgeRuntime
Cloud Native AI Meetup
10th July, 2024
18:30-20:30 | Singapore
Thank you!

Mastering Microservices with Open Source Technologies

Thursday, July 18, 2024

6:00 PM to 9:00 PM GMT+8

Scan the QR code to register

- CloudWeGo Introduction: An overview of CloudWeGo's open-source middleware for building and governing microservices.
- Make RPC Faster than Fast: A detailed look at the performance optimization journey of Kitex to improve RPC speed.
- Key to Efficiency: Insights into scaffold code generation and project engineering practices with cwgo to enhaance development efficiency.
- Practical Implementation of Multi-Runtime Architecture: Practical aspects of implementing a multi-runtime architecture using Layotto
- Explore the Next Stop of Microservice Architecture: Examination of Koupleless, a modular development framework and operation scheduling system
- Building an Efficient Service Mesh: Technical details on Merbridge's innovations in eBPF implementation and Istio Ambient for building an efficient service mesh.