KubeCon

CloudNativeCon

THE LINUX FOUNDATION

OPEN SOURCE SUMMIT

AI_dev

Open Source GenAI & ML Summit

China 2024

KubeCon | CloudNativeCon

THE LINUX FOUNDATION OPEN SOURCE SUMMIT

AI_dev Open Source GenAI & ML Summit

China 2024

Large Scale and Reliability Testing in Kubernetes using KWOK

Shiming Zhang, DaoCloud

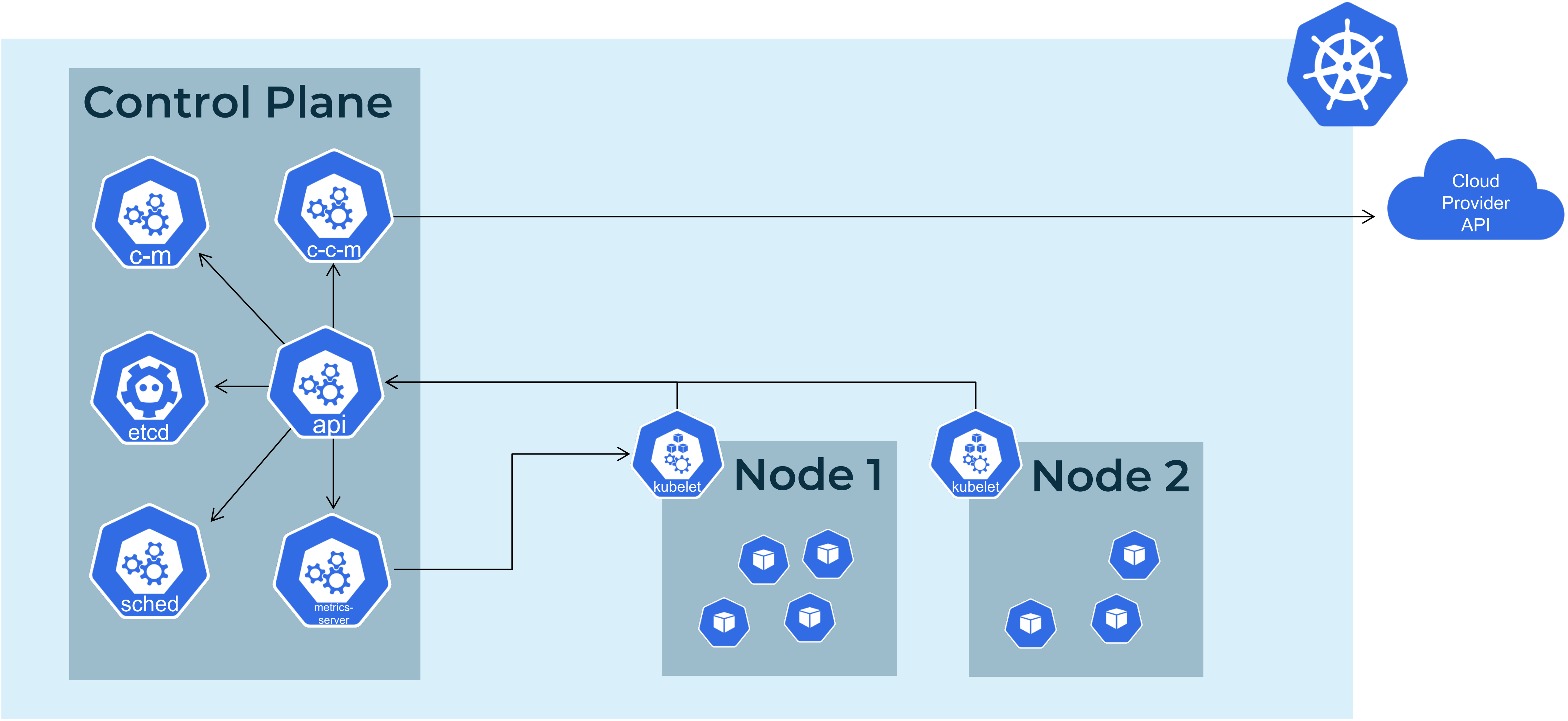
Yuan Chen, NVIDIA

Outline

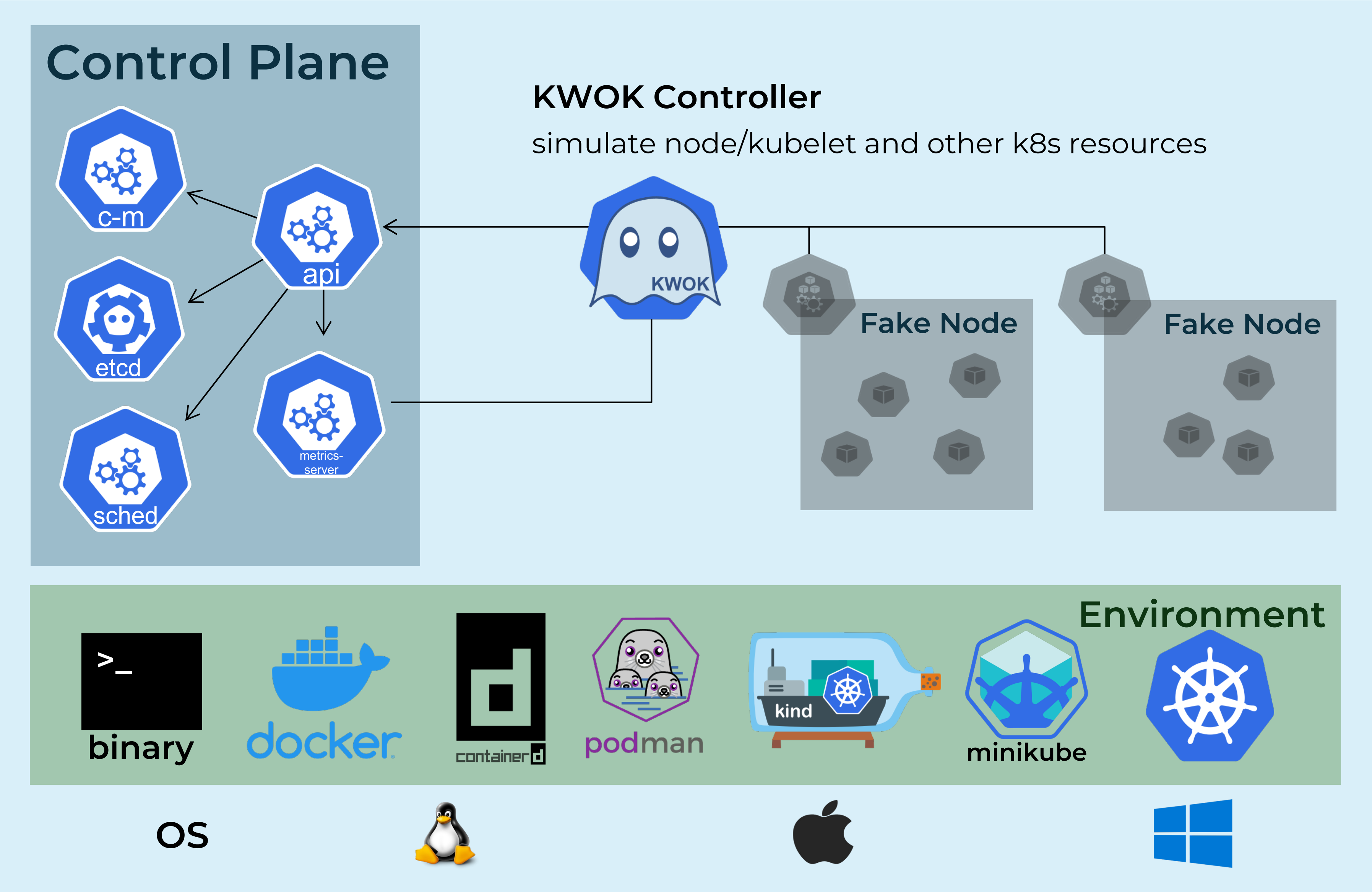
- KWOK overview and demo
- Fault injection for reliability testing and demo
- Summary

KWOK Overview

Kubernetes Cluster



KWOK: Kubernetes WithOut Kubelet



Control Plane

KWOK Controller

simulate node/kubelet and other k8s resources

Fake Node

Fake Node

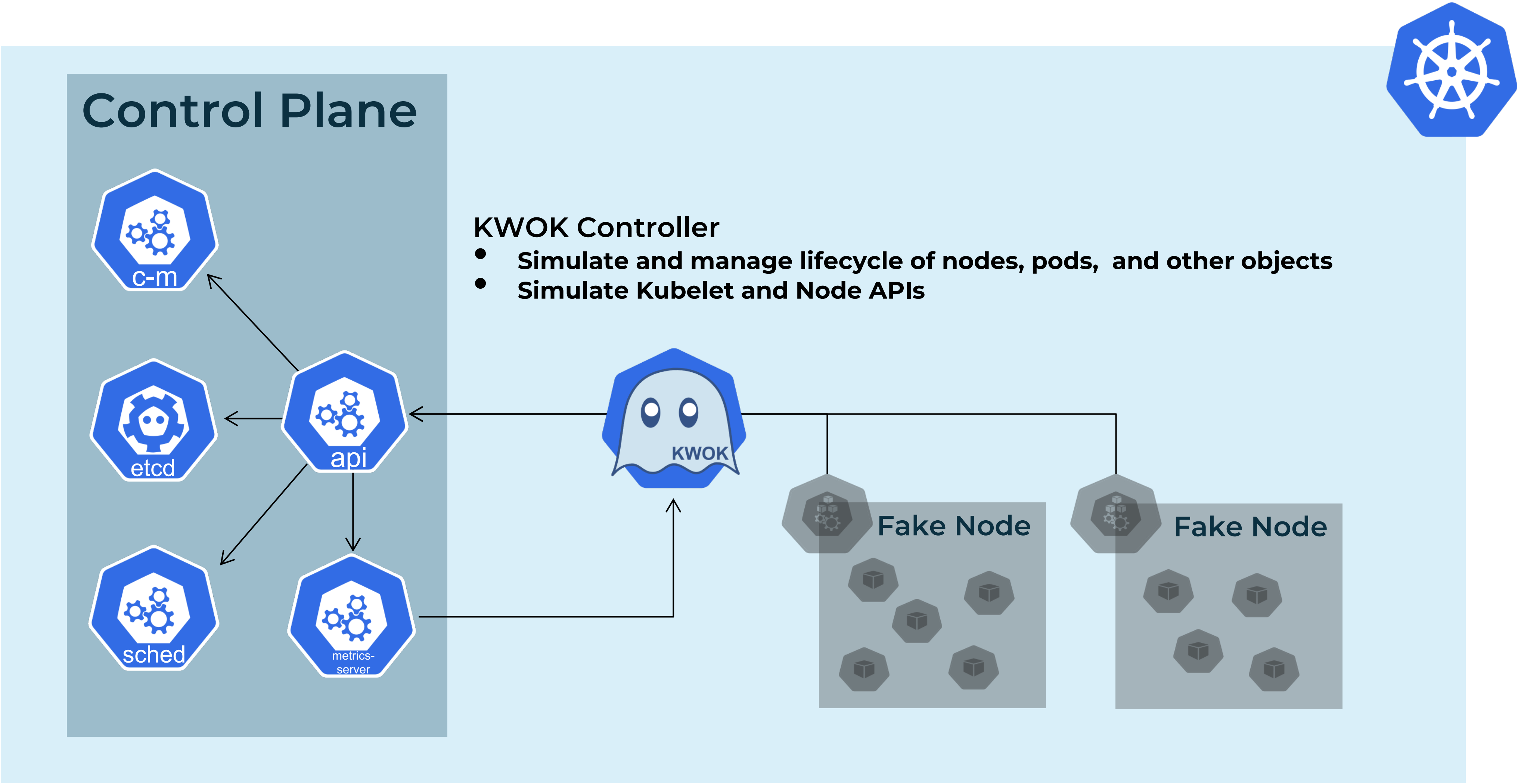
Environment

OS

kwokctl

command line tool

KWOK Controller

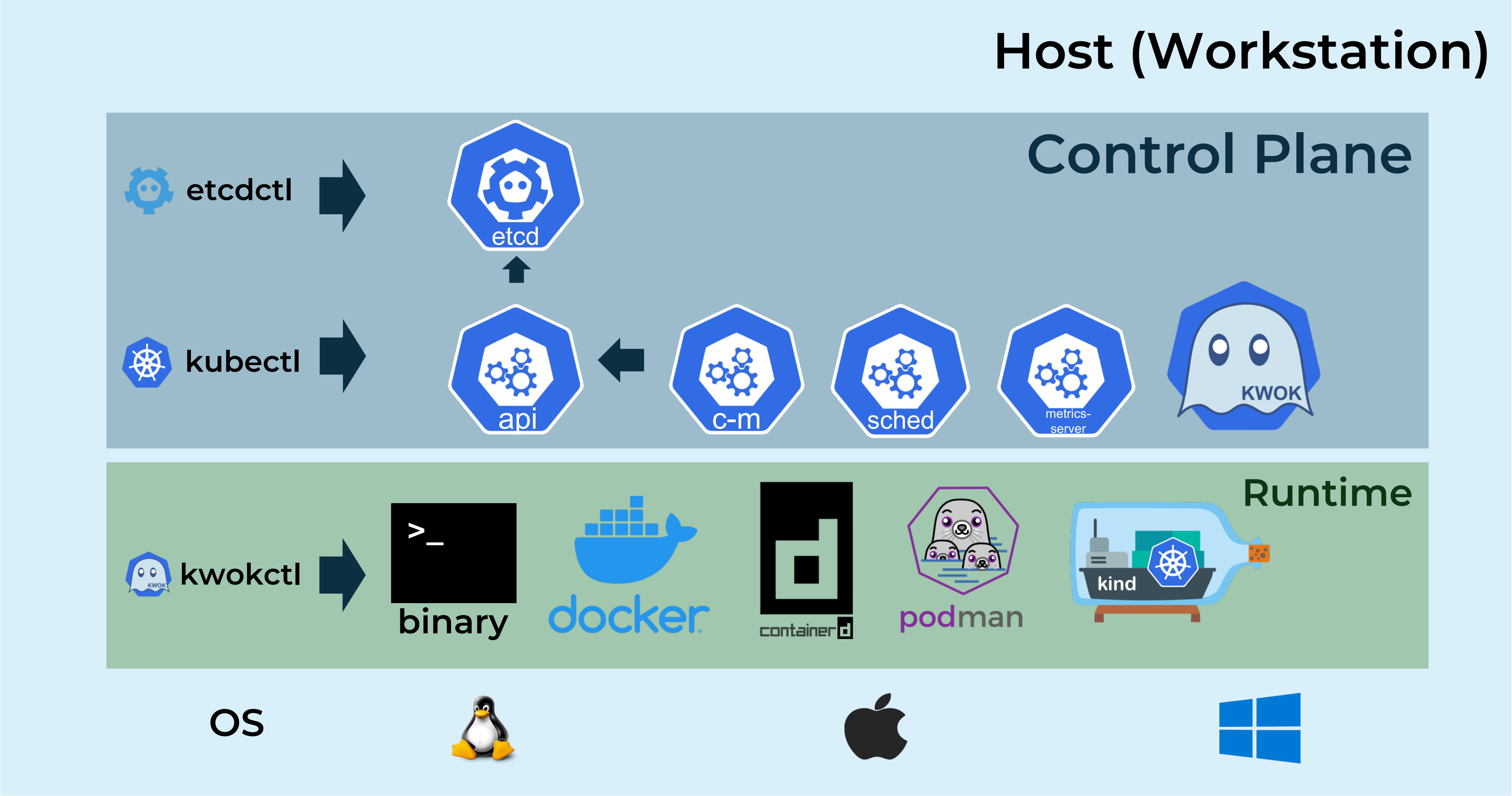


KWOK Controller

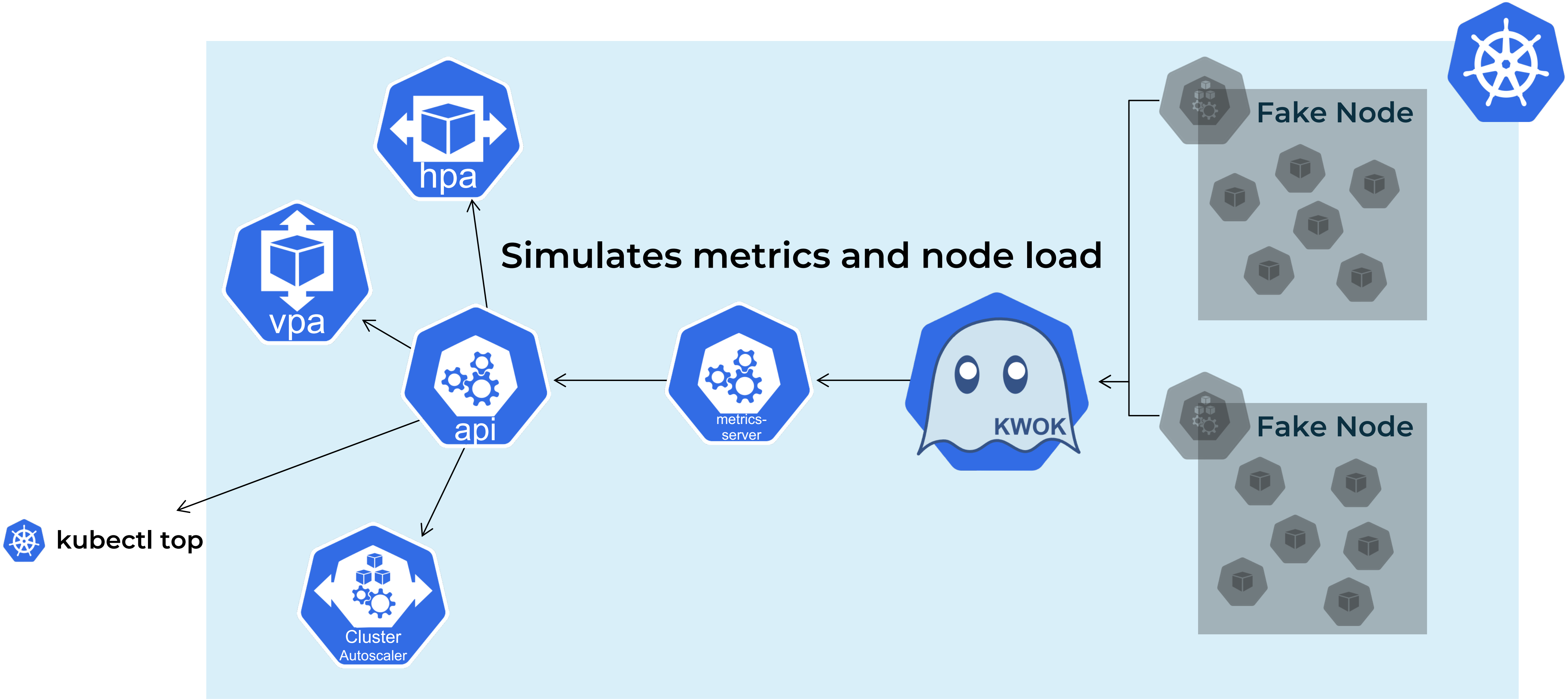
- Simulate and manage lifecycle of nodes, pods, and other objects
- Simulate Kubelet and Node APIs

kwokctl

A command line tool for cluster creation and management

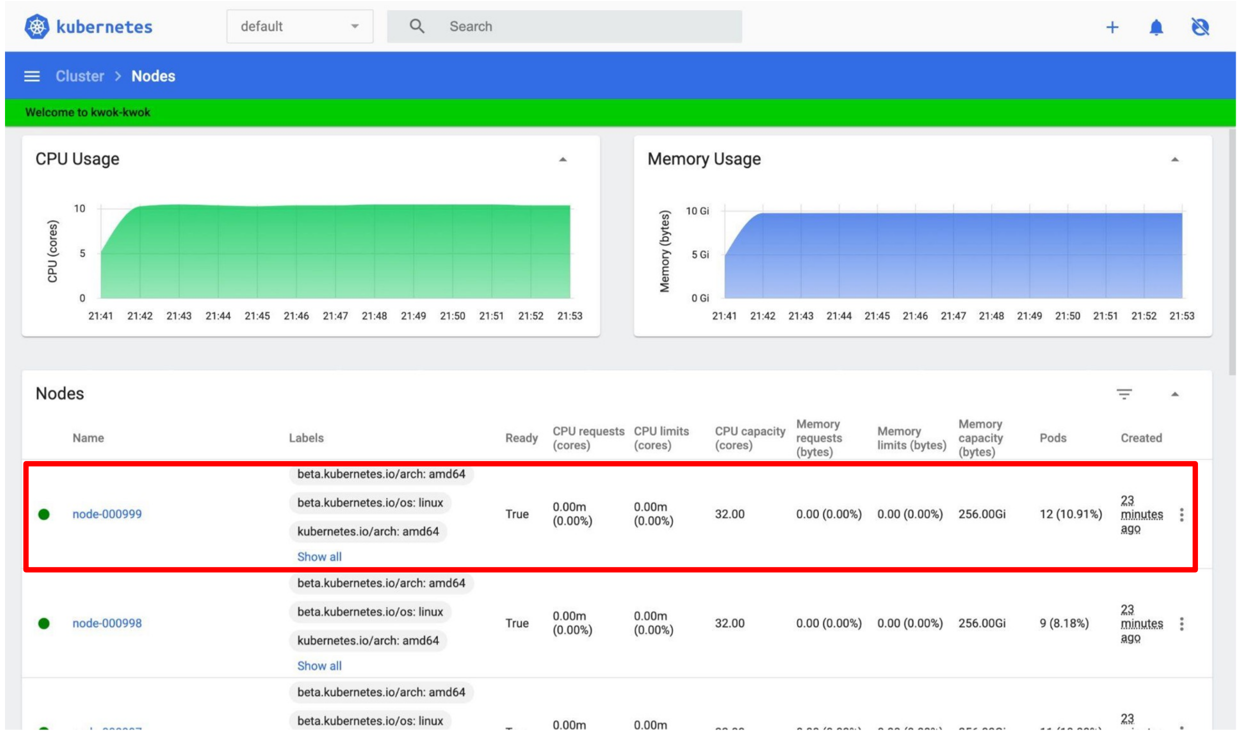


KWOK: Simulate Node Utilization



Simulates metrics and node load

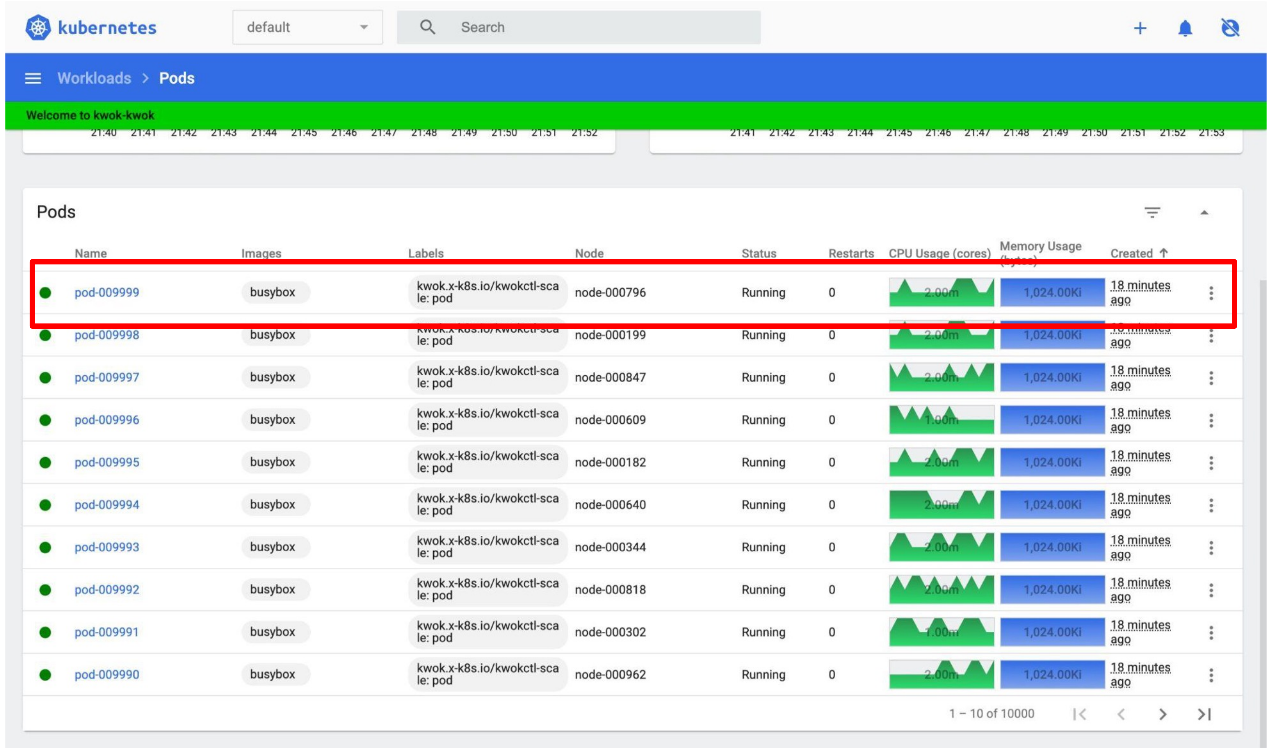
KWOK: Create Large Scale Clusters



Nodes

	Name	Labels	Ready	CPU requests (cores)	CPU limits (cores)	CPU capacity (cores)	Memory requests (bytes)	Memory limits (bytes)	Memory capacity (bytes)	Pods	Created
●	node-000999	beta.kubernetes.io/arch: amd64 beta.kubernetes.io/os: linux kubernetes.io/arch: amd64 Show all	True	0.00m (0.00%)	0.00m (0.00%)	32.00	0.00 (0.00%)	0.00 (0.00%)	256.00Gi	12 (10.91%)	23 minutes ago
●	node-000998	beta.kubernetes.io/arch: amd64 beta.kubernetes.io/os: linux kubernetes.io/arch: amd64 Show all	True	0.00m (0.00%)	0.00m (0.00%)	32.00	0.00 (0.00%)	0.00 (0.00%)	256.00Gi	9 (8.18%)	23 minutes ago
●		beta.kubernetes.io/arch: amd64 beta.kubernetes.io/os: linux		0.00m	0.00m						23

1K Nodes



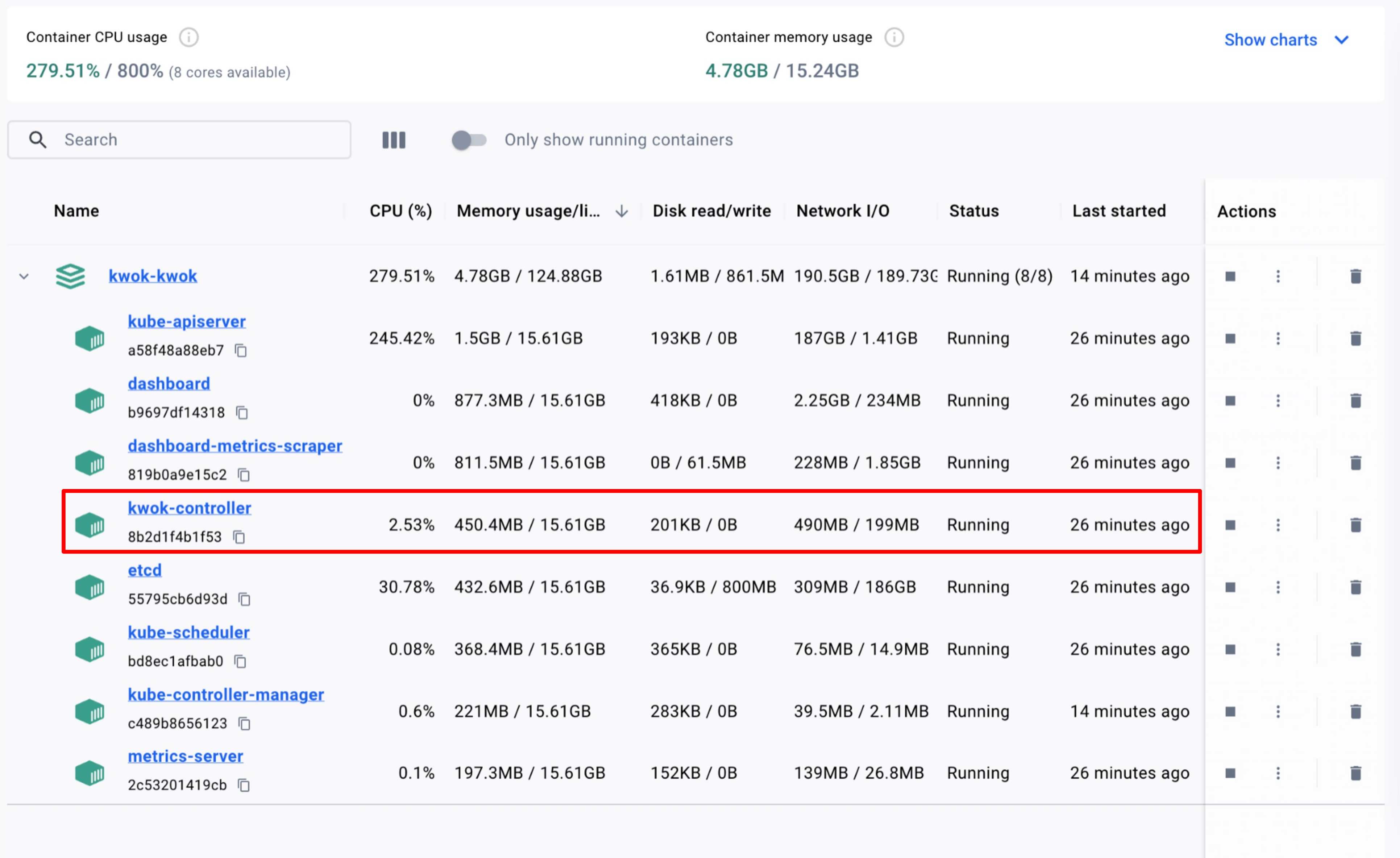
Pods

	Name	Images	Labels	Node	Status	Restarts	CPU Usage (cores)	Memory Usage (bytes)	Created ↑
●	pod-009999	busybox	kwok.x-k8s.io/kwokctl-scale: pod	node-000796	Running	0	2.00m	1,024.00Ki	18 minutes ago
●	pod-009998	busybox	kwok.x-k8s.io/kwokctl-scale: pod	node-000199	Running	0	2.00m	1,024.00Ki	18 minutes ago
●	pod-009997	busybox	kwok.x-k8s.io/kwokctl-scale: pod	node-000847	Running	0	2.00m	1,024.00Ki	18 minutes ago
●	pod-009996	busybox	kwok.x-k8s.io/kwokctl-scale: pod	node-000609	Running	0	1.00m	1,024.00Ki	18 minutes ago
●	pod-009995	busybox	kwok.x-k8s.io/kwokctl-scale: pod	node-000182	Running	0	2.00m	1,024.00Ki	18 minutes ago
●	pod-009994	busybox	kwok.x-k8s.io/kwokctl-scale: pod	node-000640	Running	0	2.00m	1,024.00Ki	18 minutes ago
●	pod-009993	busybox	kwok.x-k8s.io/kwokctl-scale: pod	node-000344	Running	0	2.00m	1,024.00Ki	18 minutes ago
●	pod-009992	busybox	kwok.x-k8s.io/kwokctl-scale: pod	node-000818	Running	0	2.00m	1,024.00Ki	18 minutes ago
●	pod-009991	busybox	kwok.x-k8s.io/kwokctl-scale: pod	node-000302	Running	0	1.00m	1,024.00Ki	18 minutes ago
●	pod-009990	busybox	kwok.x-k8s.io/kwokctl-scale: pod	node-000962	Running	0	2.00m	1,024.00Ki	18 minutes ago

1 – 10 of 10000

10K Pods

KWOK: Use Low Resource



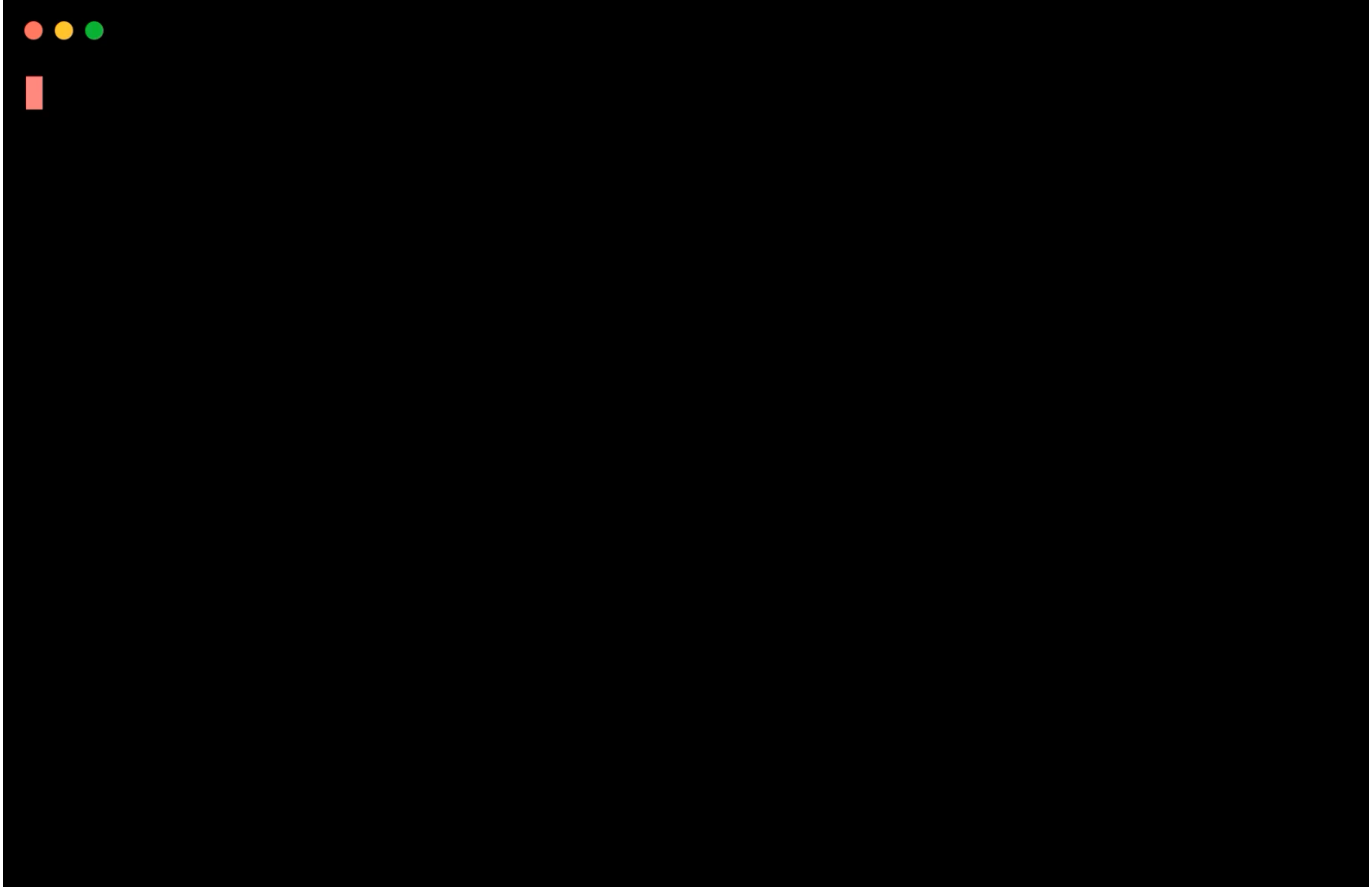
Container CPU usage ⓘ

279.51% / 800% (8 cores available)

Container memory usage ⓘ

4.78GB / 15.24GB

Name	CPU (%)	Memory usage/li... ↓	Disk read/write	Network I/O	Status	Last started	Actions
kwok-kwok	279.51%	4.78GB / 124.88GB	1.61MB / 861.5M	190.5GB / 189.73G	Running (8/8)	14 minutes ago	■ ⋮ 🗑
kube-apiserver a58f48a88eb7	245.42%	1.5GB / 15.61GB	193KB / 0B	187GB / 1.41GB	Running	26 minutes ago	■ ⋮ 🗑
dashboard b9697df14318	0%	877.3MB / 15.61GB	418KB / 0B	2.25GB / 234MB	Running	26 minutes ago	■ ⋮ 🗑
dashboard-metrics-scraper 819b0a9e15c2	0%	811.5MB / 15.61GB	0B / 61.5MB	228MB / 1.85GB	Running	26 minutes ago	■ ⋮ 🗑
kwok-controller 8b2d1f4b1f53	2.53%	450.4MB / 15.61GB	201KB / 0B	490MB / 199MB	Running	26 minutes ago	■ ⋮ 🗑
etcd 55795cb6d93d	30.78%	432.6MB / 15.61GB	36.9KB / 800MB	309MB / 186GB	Running	26 minutes ago	■ ⋮ 🗑
kube-scheduler bd8ec1afbab0	0.08%	368.4MB / 15.61GB	365KB / 0B	76.5MB / 14.9MB	Running	26 minutes ago	■ ⋮ 🗑
kube-controller-manager c489b8656123	0.6%	221MB / 15.61GB	283KB / 0B	39.5MB / 2.11MB	Running	14 minutes ago	■ ⋮ 🗑
metrics-server 2c53201419cb	0.1%	197.3MB / 15.61GB	152KB / 0B	139MB / 26.8MB	Running	26 minutes ago	■ ⋮ 🗑



KWOK Summary

KWOK is a toolkit for creating and managing large scale Kubernetes clusters with fake nodes using minimum resources

kwok controller: core component

- Simulate lifecycle of nodes, pods, and other Kubernetes objects
- Simulate nodes and Kubelet APIs
- Simulate node utilization via Kubelet metrics

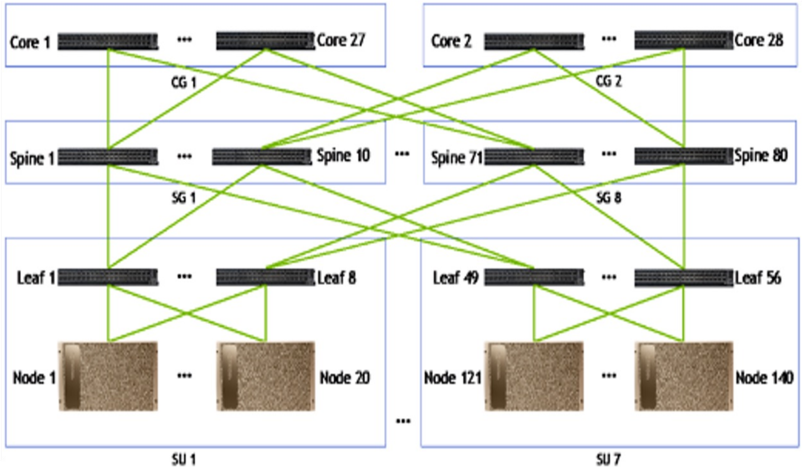
Kwokctl: a series of command line tools

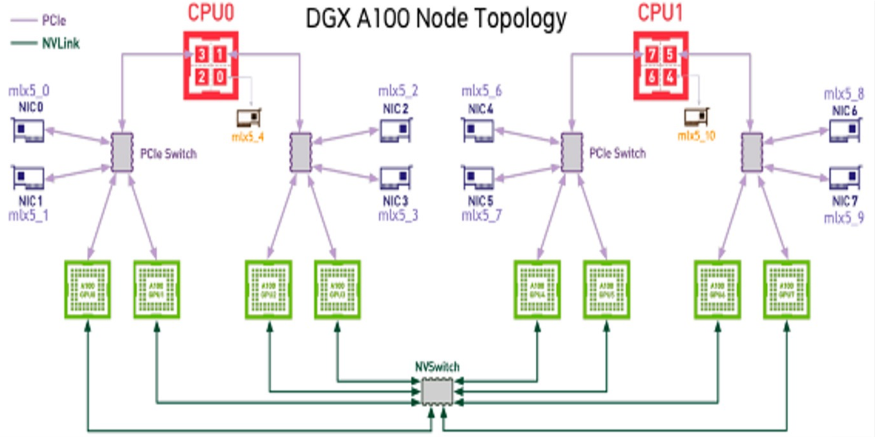
- Create and manage kwok clusters
- Dump/restore cluster snapshot

Failure Injection and Reliability Testing

Large Scale Kubernetes GPU Clusters

Hardware Architecture and Topology





NVLINK + GPUDirect RDMA

NUMA binding

Multi-level EW switching fabric

Rack + spine

Switch hierarchy

Network topology

Software Stacks &Components

Host-level Components

nvidia-container-toolkit

nvidia-gpu-driver

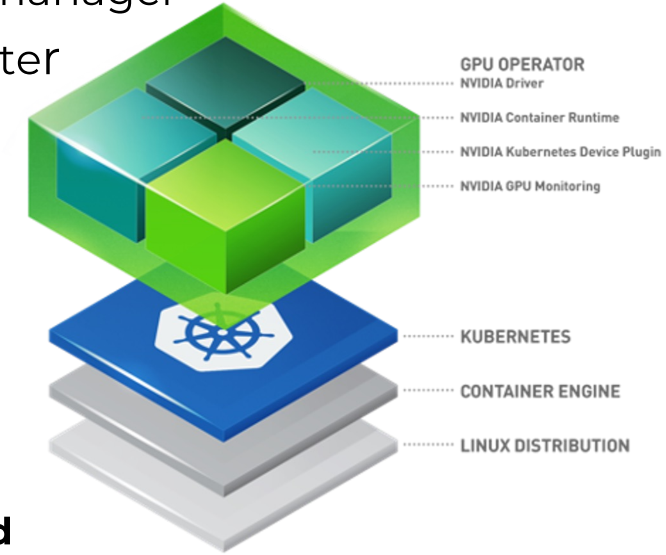
Kubernetes Components

k8s-device-plugin

gpu-feature-discovery

nvidia-mig-manager

dcgm-exporter



Source: Accelerating AI Workloads with GPUs in Kubernetes - Kevin Klues, Distinguished Engineer & Sanjay Chatterjee, Engineering Manager, NVIDIA, Keynote at KubeCon 2024 EU.

Failures in GPU Clusters

Errors/failures are the **New "Normal"**

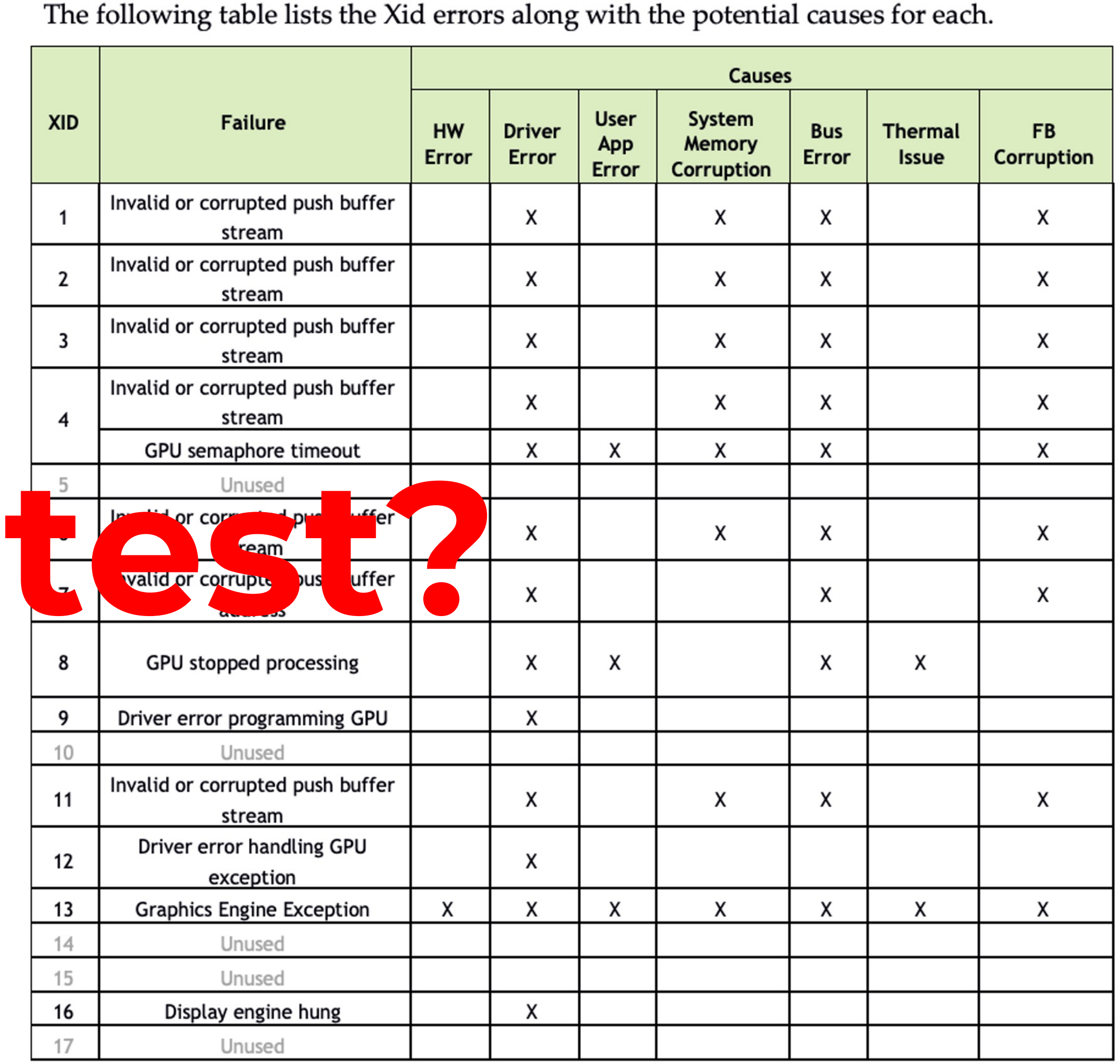
- Hardware faults: GPU, network interface, interconnect
- Software errors: driver/firmware/controllers

How to test?

Failures are costly

- Re-run a training job from scratch

Fault-tolerance is critical



The following table lists the Xid errors along with the potential causes for each.

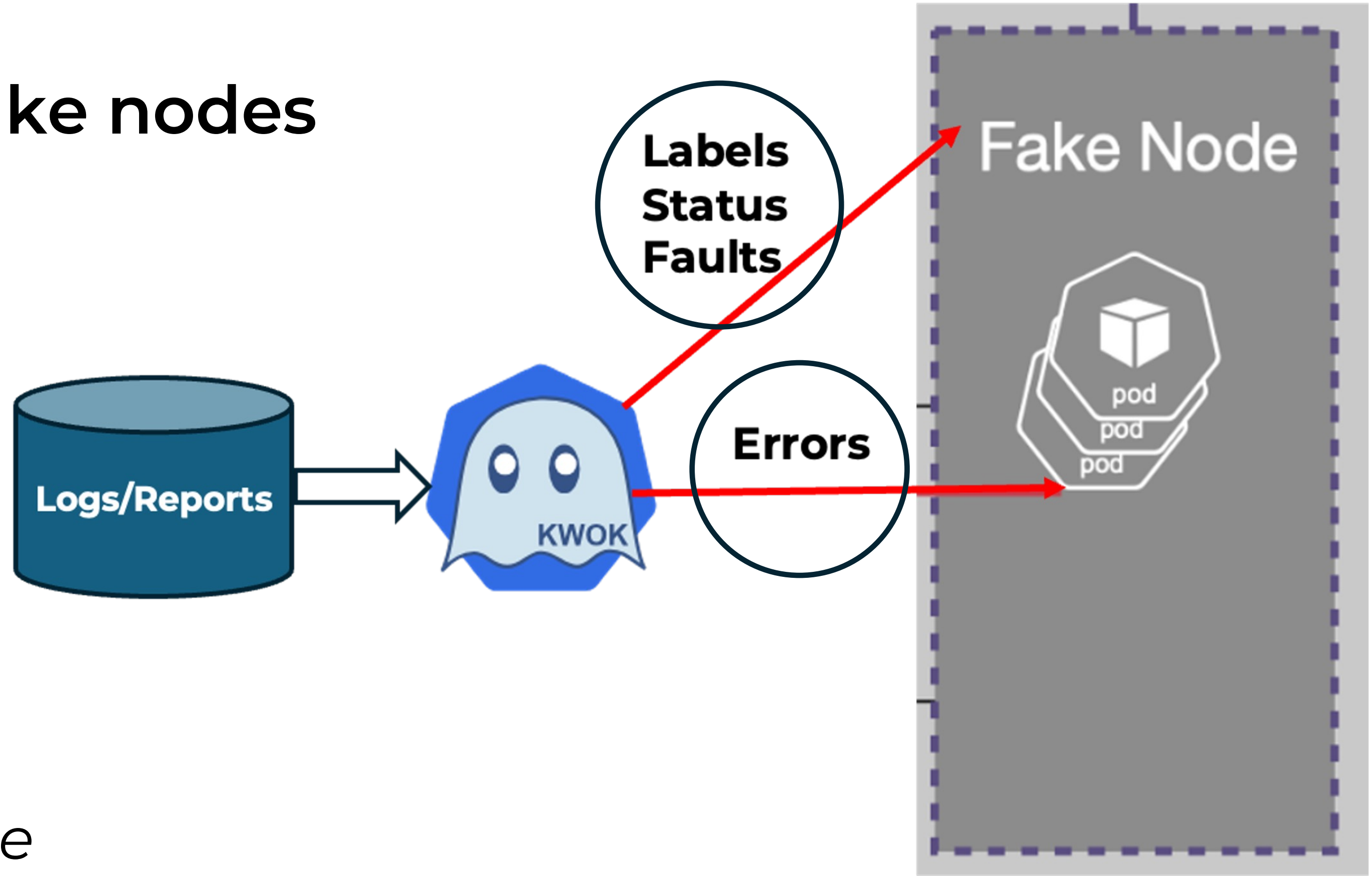
XID	Failure	Causes						
		HW Error	Driver Error	User App Error	System Memory Corruption	Bus Error	Thermal Issue	FB Corruption
1	Invalid or corrupted push buffer stream		X		X	X		X
2	Invalid or corrupted push buffer stream		X		X	X		X
3	Invalid or corrupted push buffer stream		X		X	X		X
4	Invalid or corrupted push buffer stream		X		X	X		X
	GPU semaphore timeout		X	X	X	X		X
5	Unused							
6	Invalid or corrupted push buffer stream		X		X	X		X
7	Invalid or corrupted push buffer address		X			X		X
8	GPU stopped processing		X	X		X	X	
9	Driver error programming GPU		X					
10	Unused							
11	Invalid or corrupted push buffer stream		X		X	X		X
12	Driver error handling GPU exception		X					
13	Graphics Engine Exception	X	X	X	X	X	X	X
14	Unused							
15	Unused							
16	Display engine hung		X					
17	Unused							

https://docs.nvidia.com/deploy/pdf/XID_Errors.pdf

KWOK: Fault and Error Injection

Simulate failures

- **Inject conditions/errors to fake nodes**
  - Taints
  - Labels and annotations
  - Status/conditions



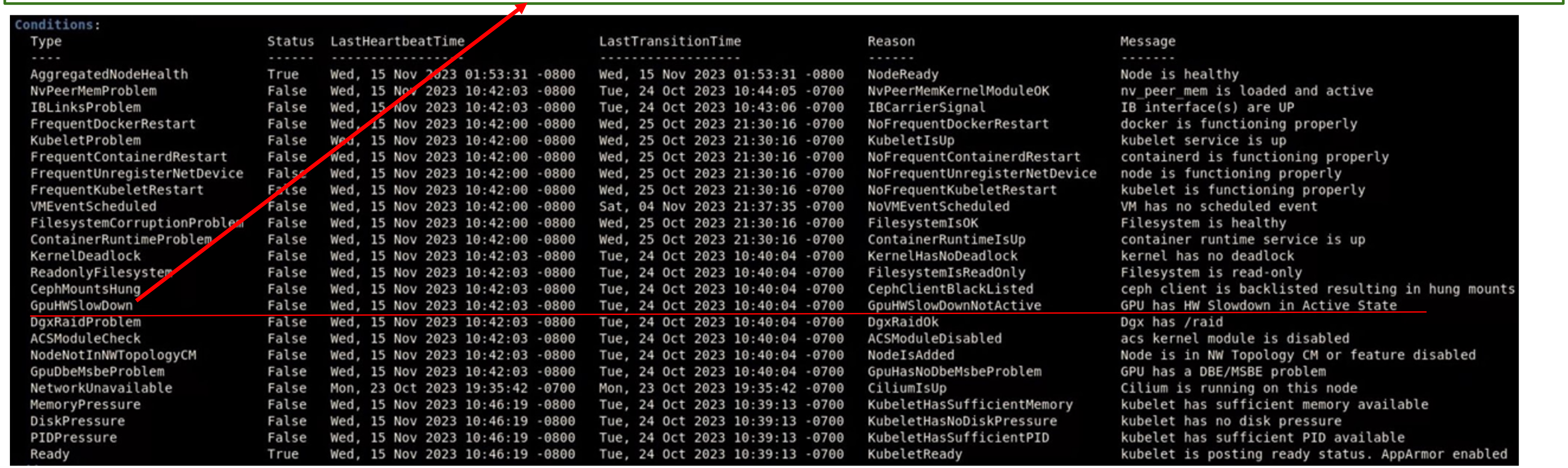
- **Inject faults to pods**
  - Initial and app. containers
  - Custom faults : *exitCode, failureReason, FailureMessage*

Node Fault Injection

Simulate node issues by injecting node conditions

- **Node Problem Detector (NPD)**: hardware (GPU, mem, disk), kernel, container runtime issues
- **DCGM Health Check:** GPU health on the node reported by NVIDIA DCGM tool APIs

Type: GpuHWSlowDown, Status: False, Reason: GpuHWSlowDownNotActive,
Message: GPU has HW Slowdown in Active State

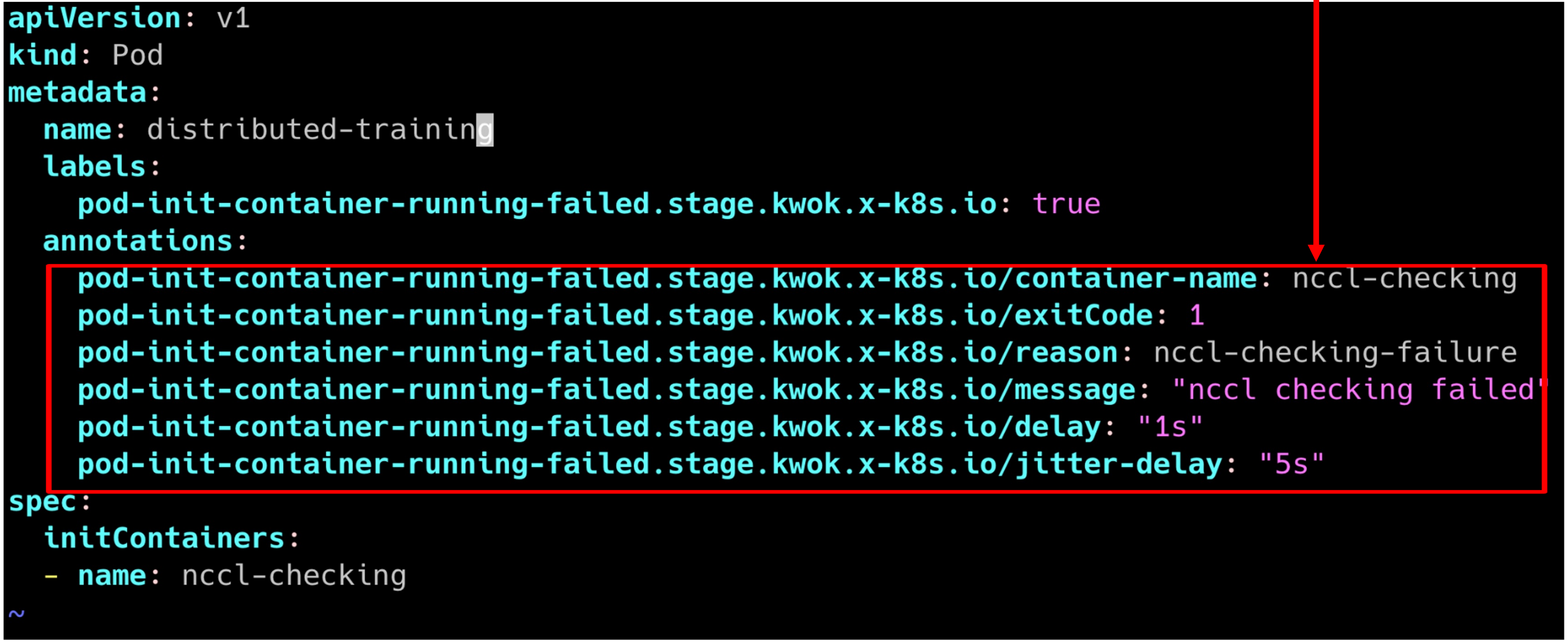


```
Conditions:
  Type                          Status  LastHeartbeatTime                 LastTransitionTime                Reason                          Message
  ----                          ------  -----------------                 ------------------                ------                          -------
  AggregatedNodeHealth          True    Wed, 15 Nov 2023 01:53:31 -0800   Wed, 15 Nov 2023 01:53:31 -0800   NodeReady                       Node is healthy
  NvPeerMemProblem              False   Wed, 15 Nov 2023 10:42:03 -0800   Tue, 24 Oct 2023 10:44:05 -0700   NvPeerMemKernelModuleOK         nv_peer_mem is loaded and active
  IBLinksProblem                False   Wed, 15 Nov 2023 10:42:03 -0800   Tue, 24 Oct 2023 10:43:06 -0700   IBCarrierSignal                 IB interface(s) are UP
  FrequentDockerRestart         False   Wed, 15 Nov 2023 10:42:00 -0800   Wed, 25 Oct 2023 21:30:16 -0700   NoFrequentDockerRestart         docker is functioning properly
  KubeletProblem                False   Wed, 15 Nov 2023 10:42:00 -0800   Wed, 25 Oct 2023 21:30:16 -0700   KubeletIsUp                     kubelet service is up
  FrequentContainerdRestart     False   Wed, 15 Nov 2023 10:42:00 -0800   Wed, 25 Oct 2023 21:30:16 -0700   NoFrequentContainerdRestart     containerd is functioning properly
  FrequentUnregisterNetDevice   False   Wed, 15 Nov 2023 10:42:00 -0800   Wed, 25 Oct 2023 21:30:16 -0700   NoFrequentUnregisterNetDevice   node is functioning properly
  FrequentKubeletRestart        False   Wed, 15 Nov 2023 10:42:00 -0800   Wed, 25 Oct 2023 21:30:16 -0700   NoFrequentKubeletRestart        kubelet is functioning properly
  VMEventScheduled              False   Wed, 15 Nov 2023 10:42:00 -0800   Sat, 04 Nov 2023 21:37:35 -0700   NoVMEventScheduled              VM has no scheduled event
  FilesystemCorruptionProblem   False   Wed, 15 Nov 2023 10:42:00 -0800   Wed, 25 Oct 2023 21:30:16 -0700   FilesystemIsOK                  Filesystem is healthy
  ContainerRuntimeProblem       False   Wed, 15 Nov 2023 10:42:00 -0800   Wed, 25 Oct 2023 21:30:16 -0700   ContainerRuntimeIsUp            container runtime service is up
  KernelDeadlock                False   Wed, 15 Nov 2023 10:42:03 -0800   Tue, 24 Oct 2023 10:40:04 -0700   KernelHasNoDeadlock             kernel has no deadlock
  ReadonlyFilesystem            False   Wed, 15 Nov 2023 10:42:03 -0800   Tue, 24 Oct 2023 10:40:04 -0700   FilesystemIsReadOnly            Filesystem is read-only
  CephMountsHung                False   Wed, 15 Nov 2023 10:42:03 -0800   Tue, 24 Oct 2023 10:40:04 -0700   CephClientBlackListed           ceph client is backlisted resulting in hung mounts
  GpuHWSlowDown                 False   Wed, 15 Nov 2023 10:42:03 -0800   Tue, 24 Oct 2023 10:40:04 -0700   GpuHWSlowDownNotActive          GPU has HW Slowdown in Active State
  DgxRaidProblem                False   Wed, 15 Nov 2023 10:42:03 -0800   Tue, 24 Oct 2023 10:40:04 -0700   DgxRaidOk                       Dgx has /raid
  ACSModuleCheck                False   Wed, 15 Nov 2023 10:42:03 -0800   Tue, 24 Oct 2023 10:40:04 -0700   ACSModuleDisabled               acs kernel module is disabled
  NodeNotInNWTopologyCM         False   Wed, 15 Nov 2023 10:42:03 -0800   Tue, 24 Oct 2023 10:40:04 -0700   NodeIsAdded                     Node is in NW Topology CM or feature disabled
  GpuDbeMsbeProblem             False   Wed, 15 Nov 2023 10:42:03 -0800   Tue, 24 Oct 2023 10:40:04 -0700   GpuHasNoDbeMsbeProblem          GPU has a DBE/MSBE problem
  NetworkUnavailable            False   Mon, 23 Oct 2023 19:35:42 -0700   Mon, 23 Oct 2023 19:35:42 -0700   CiliumIsUp                      Cilium is running on this node
  MemoryPressure                False   Wed, 15 Nov 2023 10:46:19 -0800   Tue, 24 Oct 2023 10:39:13 -0700   KubeletHasSufficientMemory      kubelet has sufficient memory available
  DiskPressure                  False   Wed, 15 Nov 2023 10:46:19 -0800   Tue, 24 Oct 2023 10:39:13 -0700   KubeletHasNoDiskPressure        kubelet has no disk pressure
  PIDPressure                   False   Wed, 15 Nov 2023 10:46:19 -0800   Tue, 24 Oct 2023 10:39:13 -0700   KubeletHasSufficientPID         kubelet has sufficient PID available
  Ready                         True    Wed, 15 Nov 2023 10:46:19 -0800   Tue, 24 Oct 2023 10:39:13 -0700   KubeletReady                    kubelet is posting ready status. AppArmor enabled
```

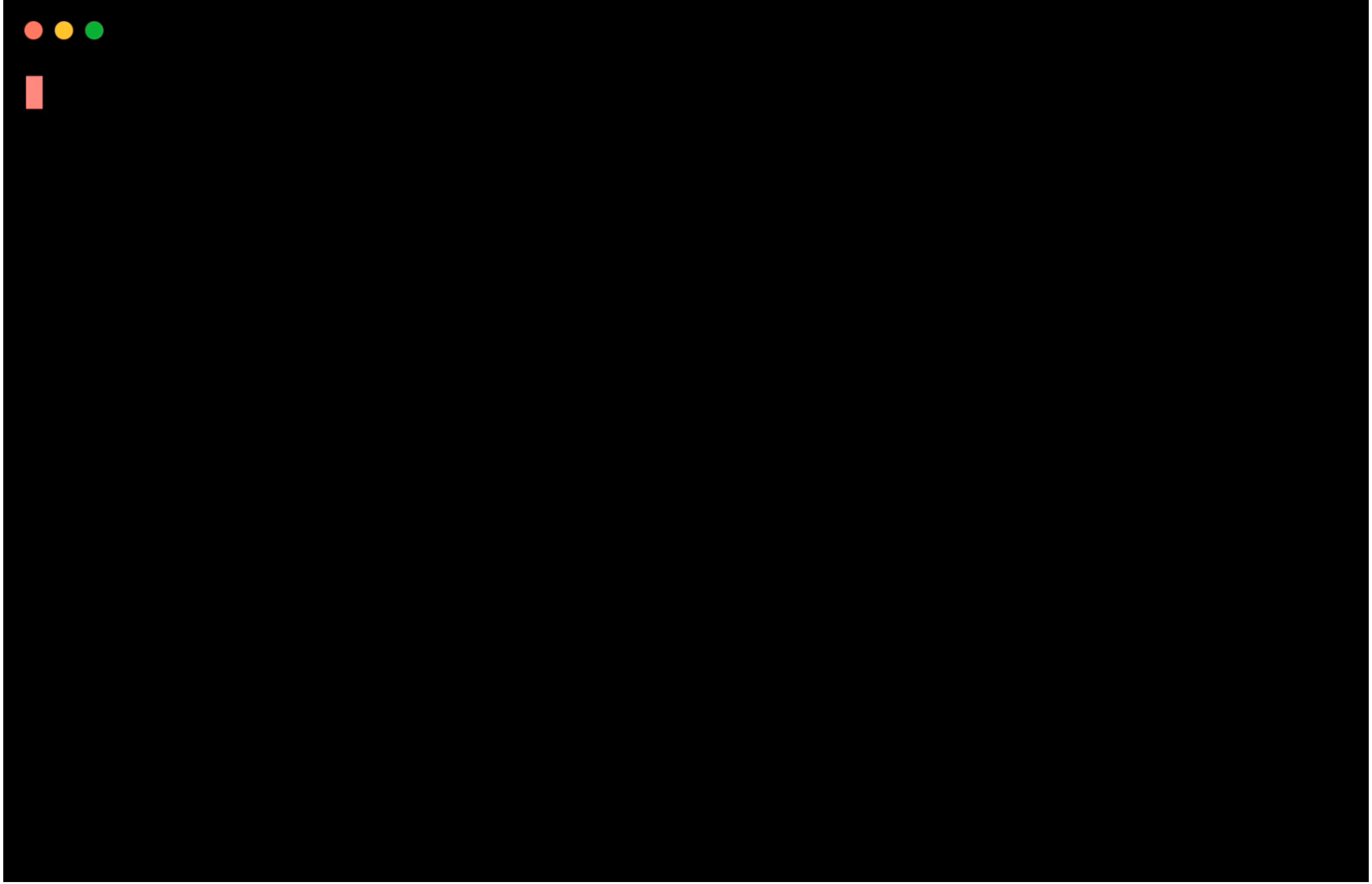
Pod Fault Injection

Inject errors to **initContainer** to simulate preflight check failures: e.g., NCCL check, prolog-check, etc.

Custom fault: container, exitCode, message, reason, delay



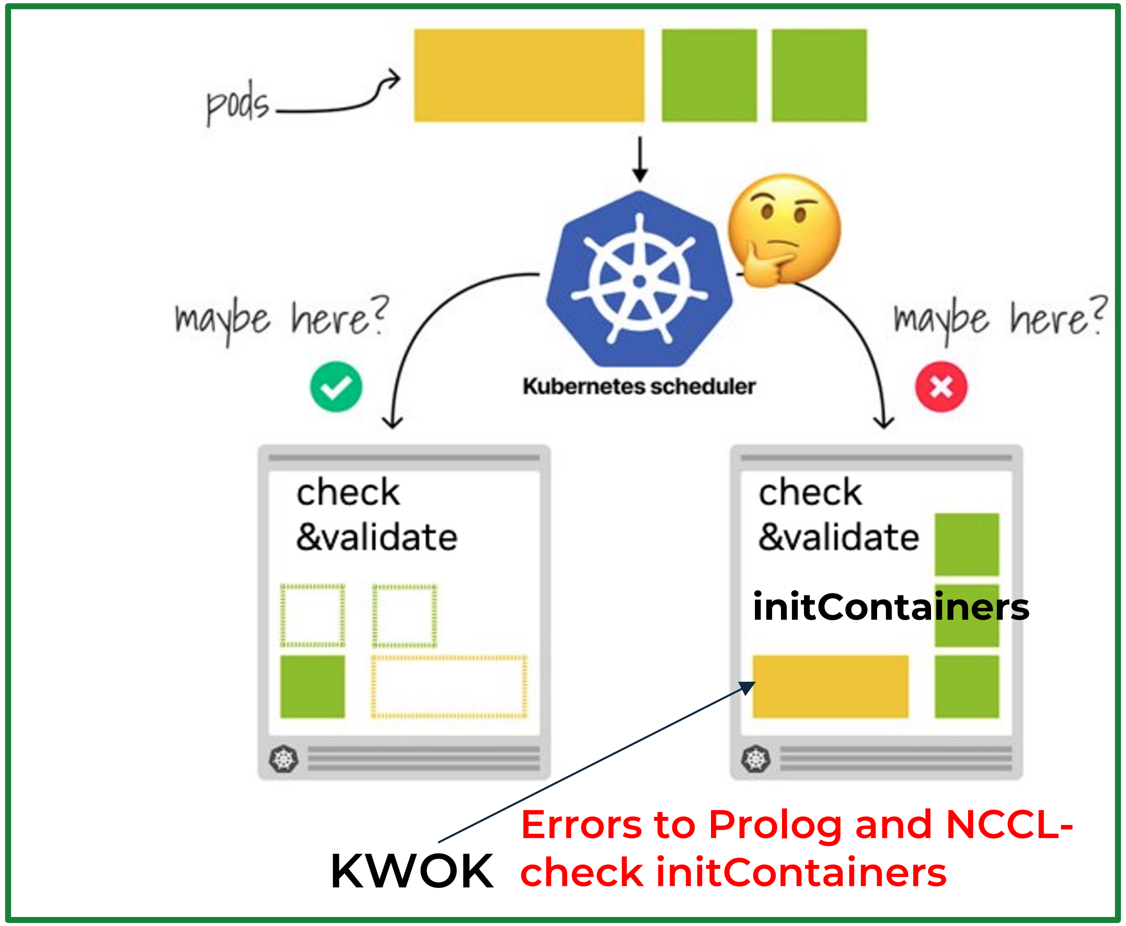
```yaml
apiVersion: v1
kind: Pod
metadata:
  name: distributed-training
  labels:
    pod-init-container-running-failed.stage.kwok.x-k8s.io: true
  annotations:
    pod-init-container-running-failed.stage.kwok.x-k8s.io/container-name: nccl-checking
    pod-init-container-running-failed.stage.kwok.x-k8s.io/exitCode: 1
    pod-init-container-running-failed.stage.kwok.x-k8s.io/reason: nccl-checking-failure
    pod-init-container-running-failed.stage.kwok.x-k8s.io/message: "nccl checking failed"
    pod-init-container-running-failed.stage.kwok.x-k8s.io/delay: "1s"
    pod-init-container-running-failed.stage.kwok.x-k8s.io/jitter-delay: "5s"
spec:
  initContainers:
  - name: nccl-checking
~
```



Use Case: Testing and Evaluating Fault-tolerant Job Scheduling

Proactive Fault-tolerant Scheduling

Preflight check to avoid scheduling jobs on problematic nodes

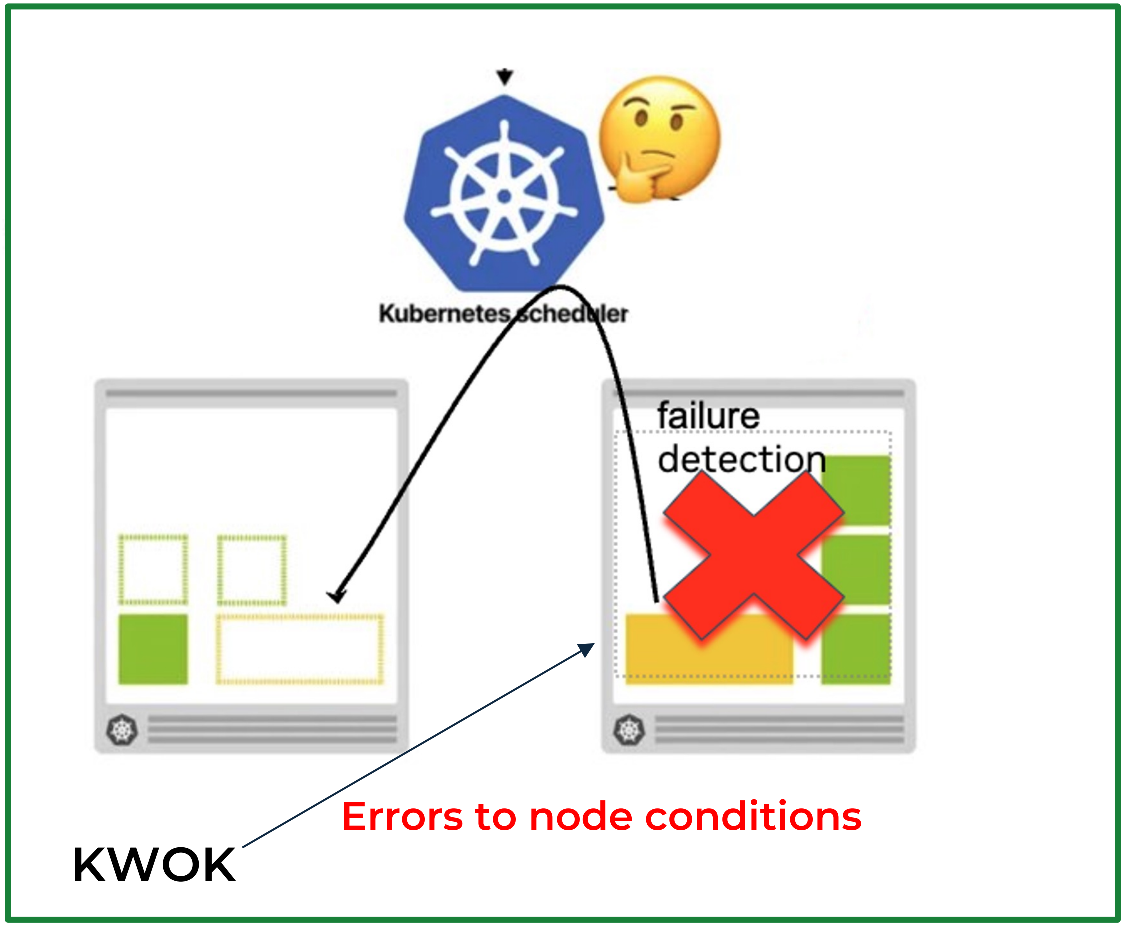


KWOK

Errors to Prolog and NCCL-check initContainers

Reactive Fault-tolerant Scheduling

Detect fault and take corrective actions



KWOK

Errors to node conditions

Source: Fault-tolerance Scheduling. Sanjay Chatterjee, Arpit Singh, Abhijit Paithankar, NVIDIA.

Summary

KWOK Use Cases and Adoption

DaoCloud

Multi-cluster Testing

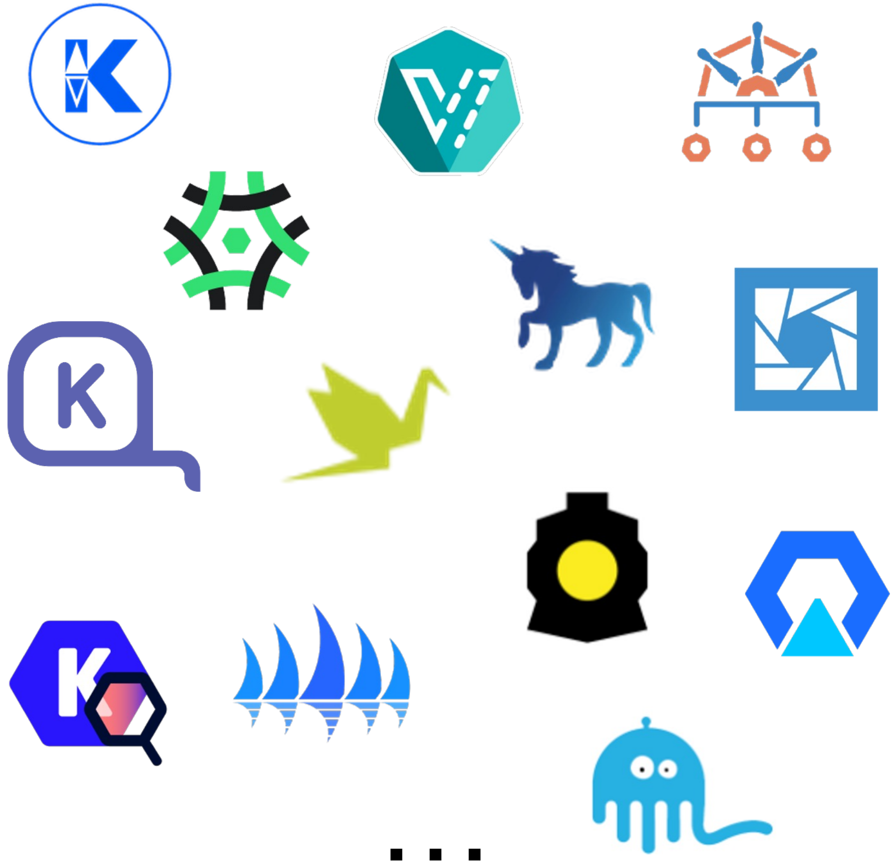
- ClusterPedia: search Kubernetes resources across multi-clusters
- DCE 5: private cloud management platform
- **Large-scale** cluster stress testing
- ...

NVIDIA

Testing in GPU Clusters

- Knavigator: NVIDIA Kubernetes testing framework
- Testing of **fault-tolerant job scheduling**
- Comparison and evaluation of scheduling systems for AI/ML
  - K8s
  - Slurm
  - Volcano
  - Kueue
  - ...

Related Open Source Projects



Summary

KWOW is **a power tool** for large scale Kubernetes testing at a low cost.

KWOK provides **support of failure injection and simulation** for testing.

What's next?

- **GPU nodes and clusters for AI/ML workloads**
  - Simulate node operators: e.g., fake GPU operator
- **Failure and reliability testing**
  - Simulate and integrate different GPU faults and errors
  - Integrate data from failure monitoring, such as DCGM, Node Problem Detector
- Advanced kwok-operator
  - Manage **multiple kwoks** to simulate larger clusters
  - Manage creation and deletion of any resources

References

KWOK

- Project:: https://kwok.sigs.k8s.io/
- GitHub:: https://kwok.sigs.k8s.io/docs/adopters/
- Demos: https://github.com/kubernetes-sigs/kwok/tree/main/demo
- Related talks:
  - Shiming Zhang & Hao Liang, 深入研究：KWOK | Deep Dive: KWOK
  - Sara Kokkila-Schumacher & Vishakha Ramani Best Practices: Improving Batch Scheduling Performance at Scale Using MCAD and KWOK
  - Wei Huang & Weiwei Yang, Revolutionizing Kube Scalability Testing with KWOK
  - Dejan Zele Pejchev, Scaling the Heights: Simulating Very Large Kubernetes Clusters with KWOK

Knavigator

- GitHub: https://github.com/NVIDIA/knavigator

Projects that use KWOK (Adopters)

- https://github.com/kubernetes-sigs/kube-scheduler-simulator
- https://github.com/kubernetes-sigs/e2e-framework
- https://github.com/kubernetes-sigs/karpenter
- https://github.com/kubernetes/autoscaler
- https://github.com/capi-samples/cluster-api-provider-kwok
- https://github.com/kyverno/kyverno
- https://github.com/kubevirt/kubevirt
- https://github.com/NVIDIA/knavigator
- https://github.com/apache/yunikorn-k8shim
- https://github.com/Azure/azure-container-networking
- https://github.com/project-codeflare/multi-cluster-app-dispatcher
- https://github.com/openshift-psap/topsail
- https://github.com/kubescape/kwok-bench
- https://github.com/acrlabs/simkube
- https://github.com/run-ai/fake-gpu-operator
- https://github.com/kubeovn/kube-ovn
- https://github.com/nuodb/terraform-provider-nuodbaas
- https://github.com/vladimirvivien/ktop
- https://github.com/headlamp-k8s/headlamp
- https://github.com/turbonomic/kubeturbo
- https://github.com/kubewharf/kubeadmira
- https://github.com/clusterpedia-io/clusterpedia
- ...

Acknowledgements

DaoCloud

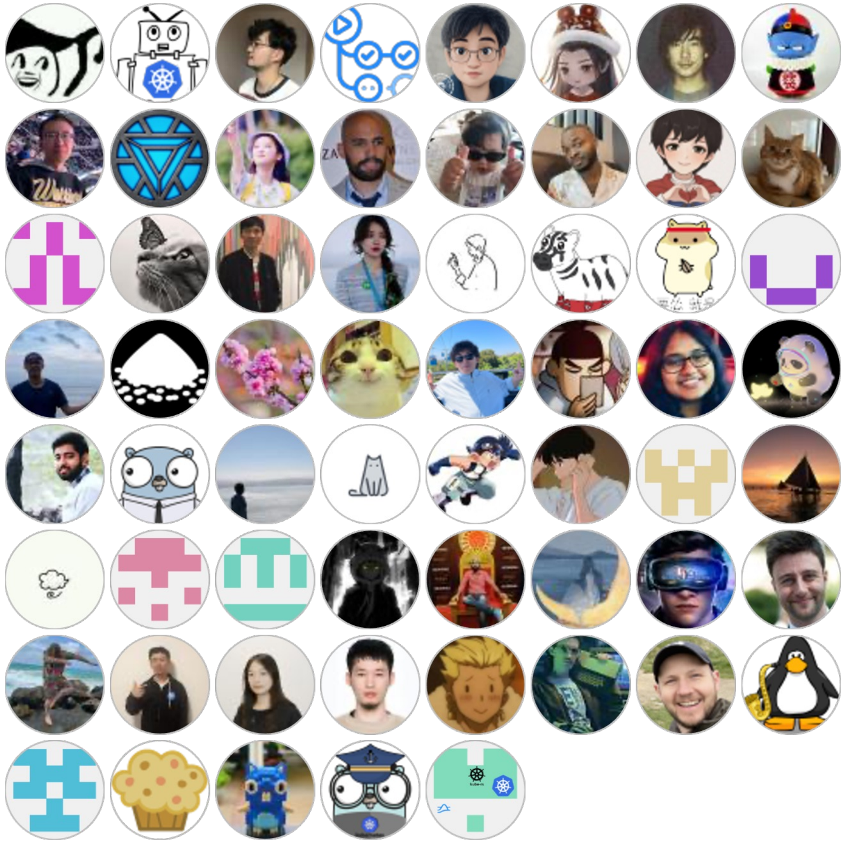
Paco Xu
Peter Pan
Kante Yin
Max Zhu
Mengjiao Liu
Michael Yao
Yang Xiao

Kay Yan
Iceber Gu
Carlory Fan
Chauncey Jiang
York Chen
Wenjie Song
Minjie Huang

NVIDIA

Dmitry Shmulevich
Kevin Klues
Sanjay Chatterjee
Brian Blitzer
Adam Tetelman
Rob Esker
Arpit Singh
Abhijit Paithankar
Carlos Arango Gutierrez

KWOK Contributors



Thank you!