

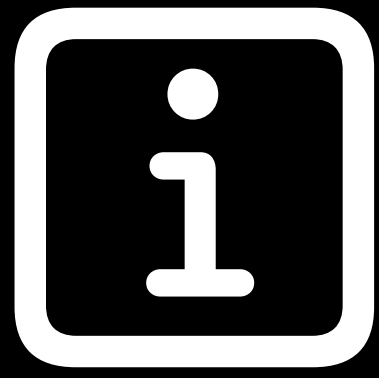


A Comprehensive Analysis on Kueue, Volcano and YuniKorn

2025.4.2 KubeCon London

Wei Huang, Apple

Shiming Zhang, DaoCloud



Overview

? Why Custom Schedulers?



What's missing in kube-scheduler?



Pod

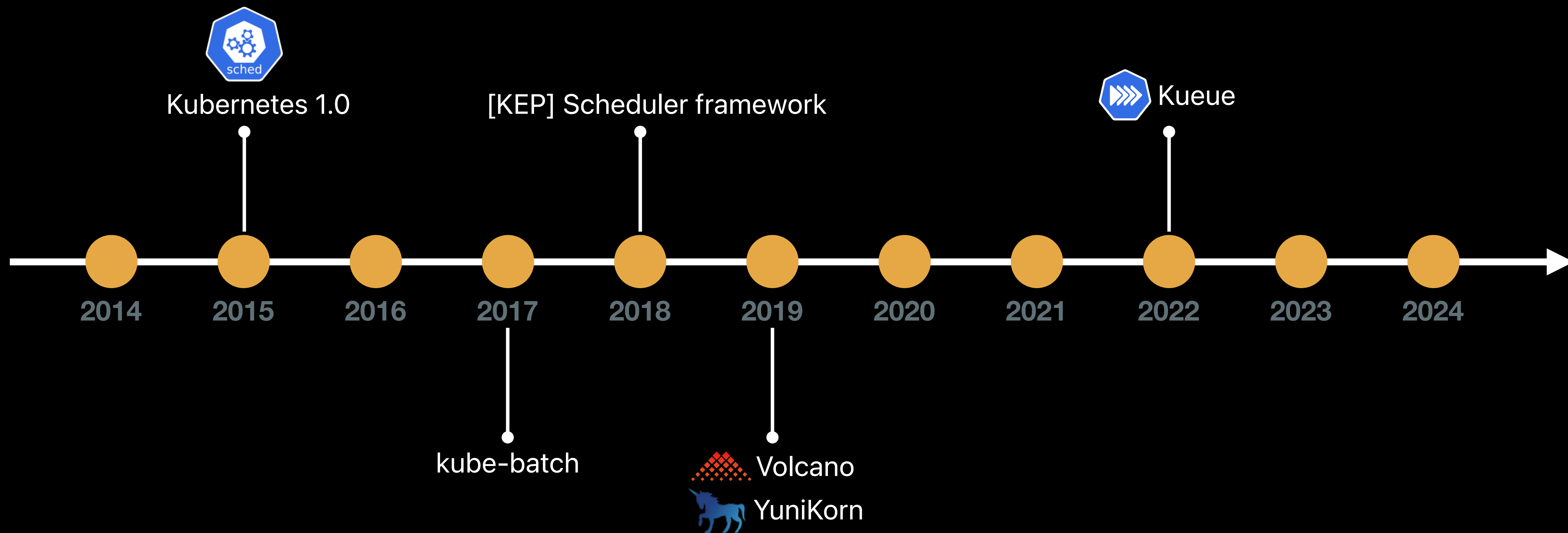


Quota

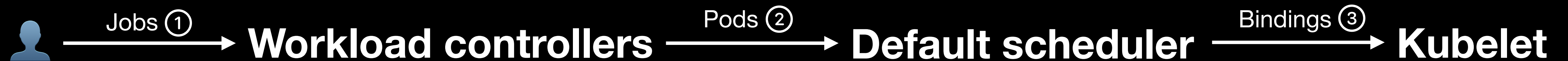


Queue

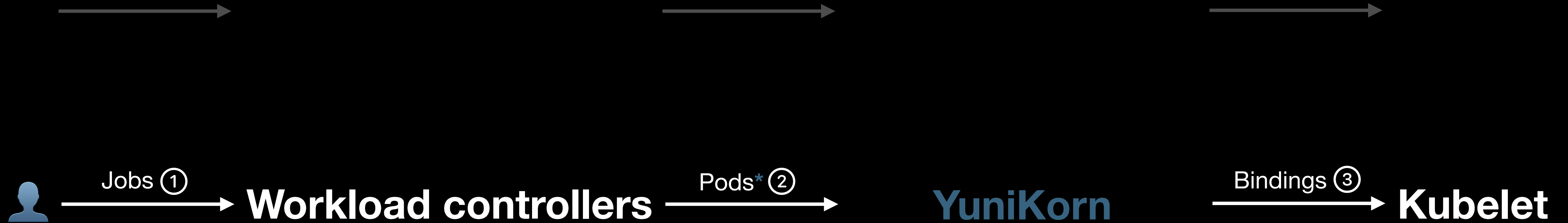
A Glimpse at the History



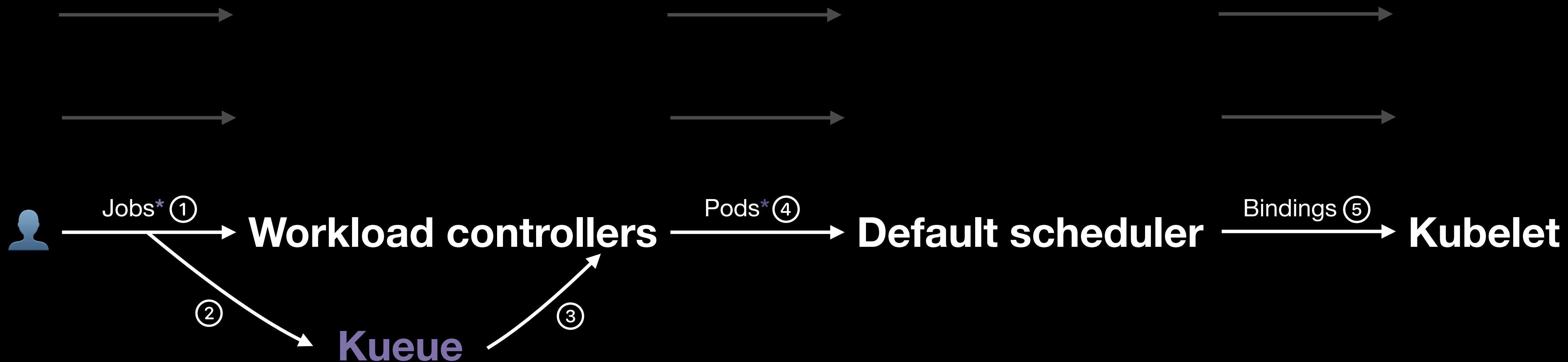
➡ High-level Workflows



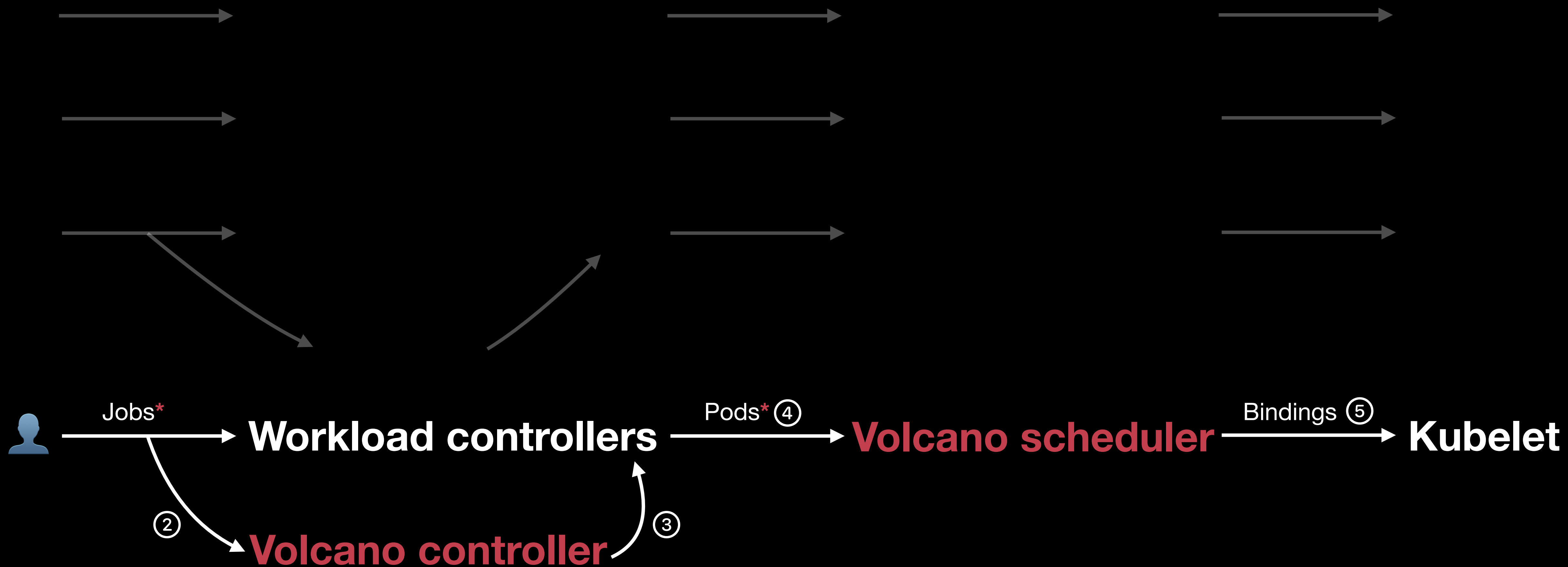
➡ High-level Workflows



➔ High-level Workflows

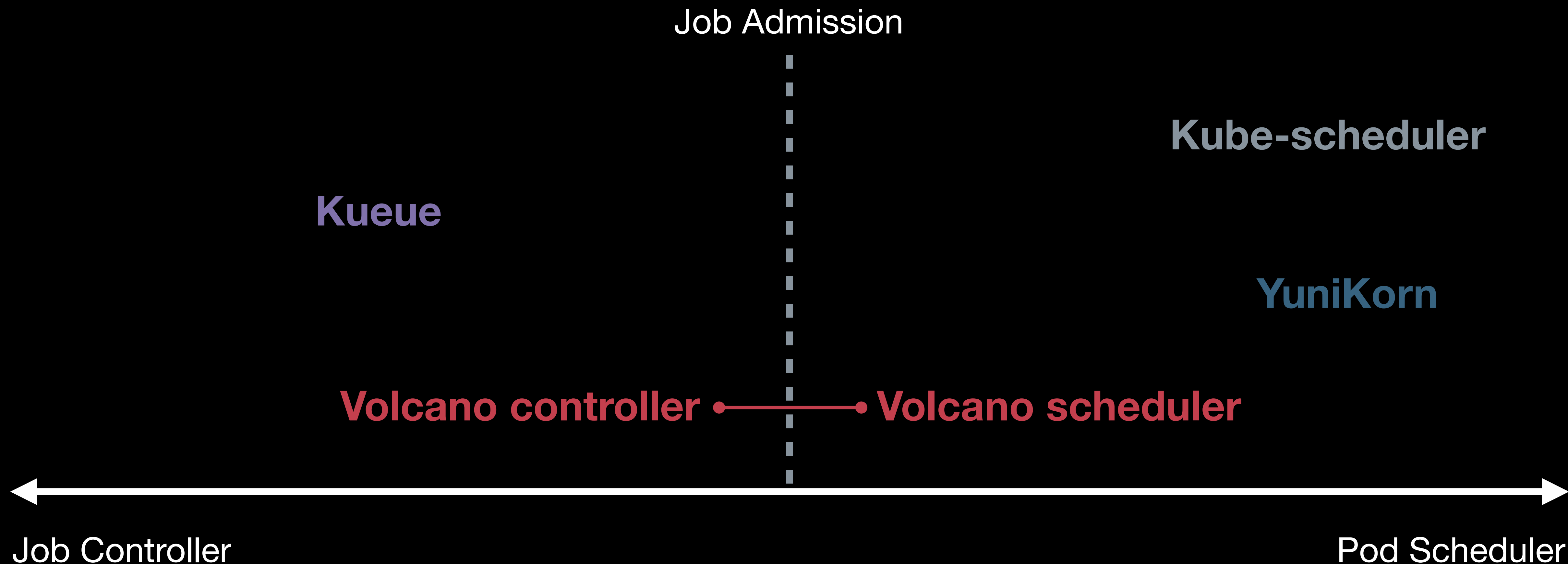


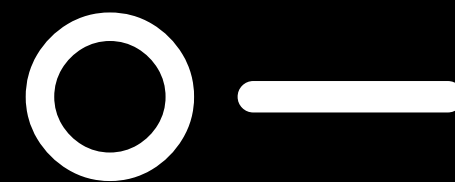
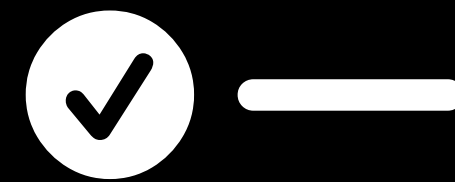
➡ High-level Workflows





Scheduler Spectrum

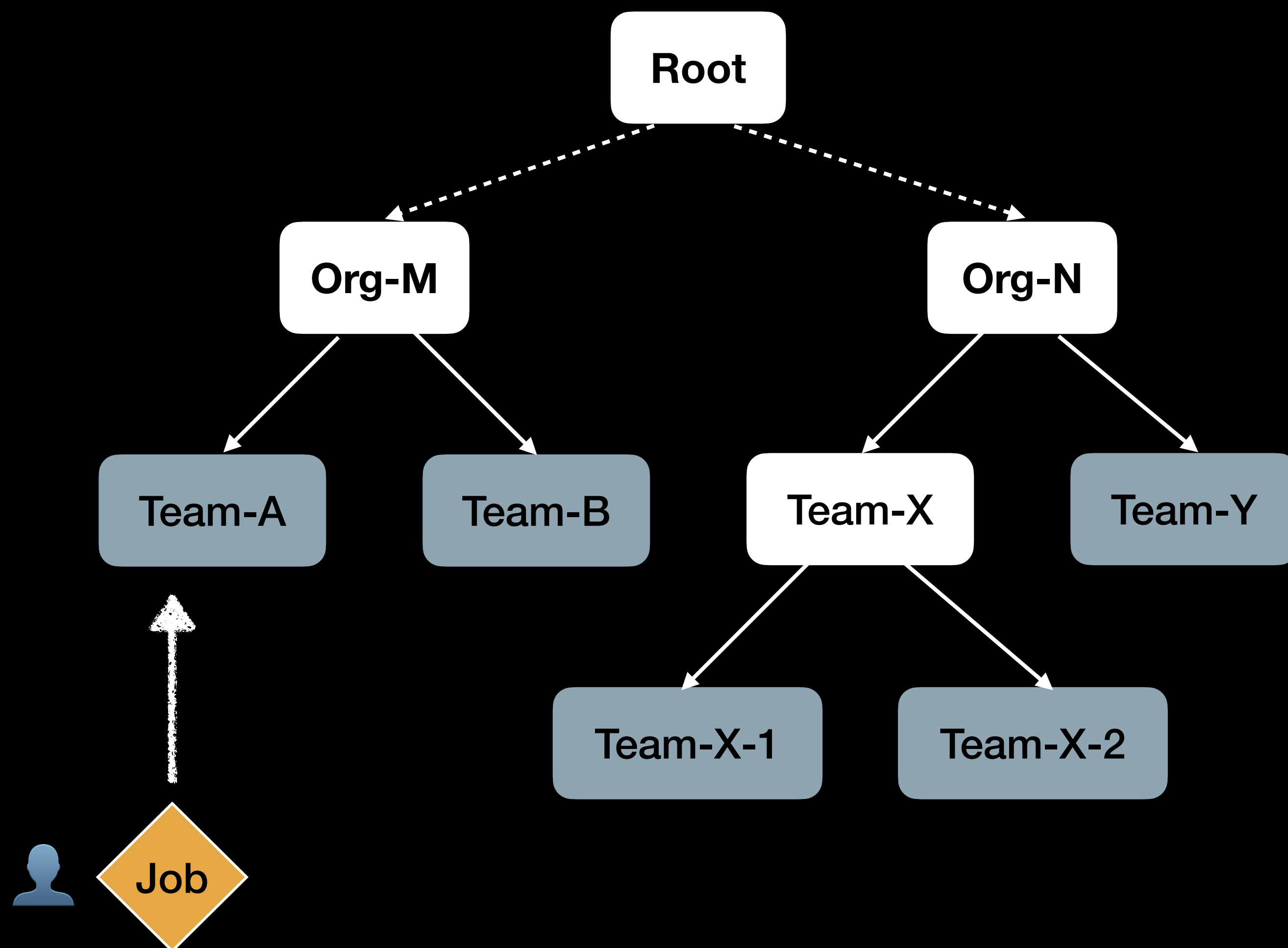
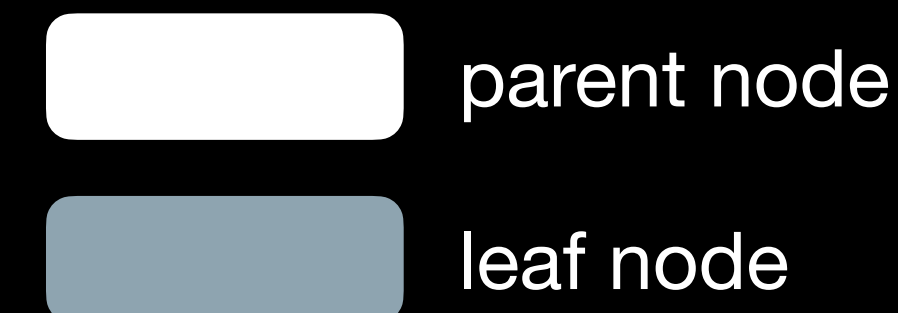




Comparison



Hierarchical Queues



Multi-level

User submits Jobs to leaf node



Hierarchical Queues (cont.)



Queue API & Config



Cohort & ClusterQueue CRD



Queue CRD



queues.yaml (in ConfigMap)



Tenant Jobs <=> Queue



LocalQueue & queue label



Queue spec field



Queue label



Job Submission AccessControl



namespaceSelector & admissionChecks



n/a



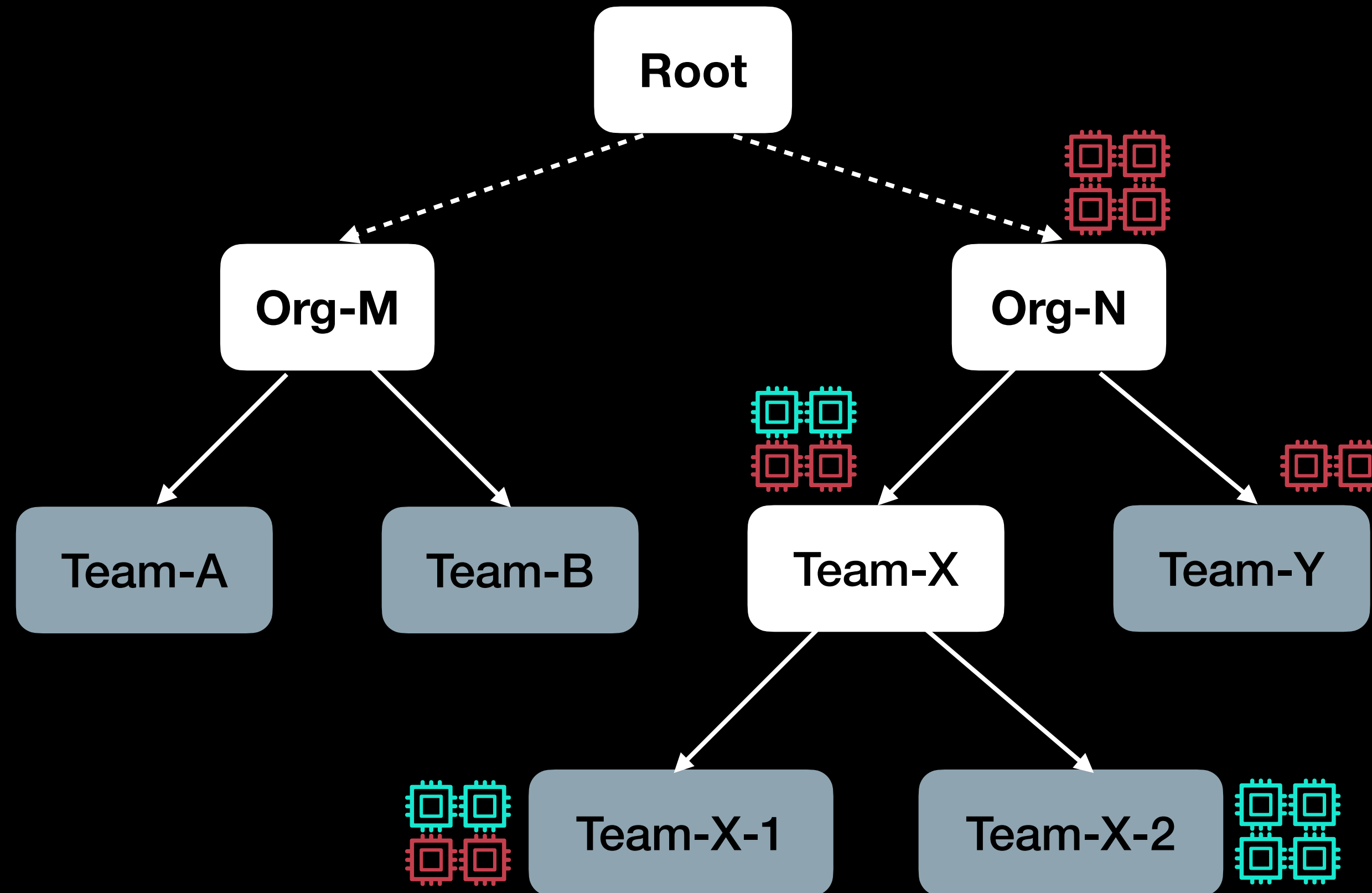
queues[*].submitacl (in ConfigMap)



Quota Management

 guaranteed

 best-efforts



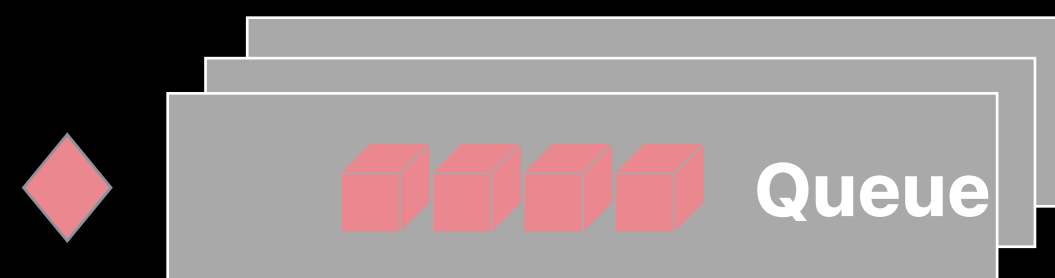
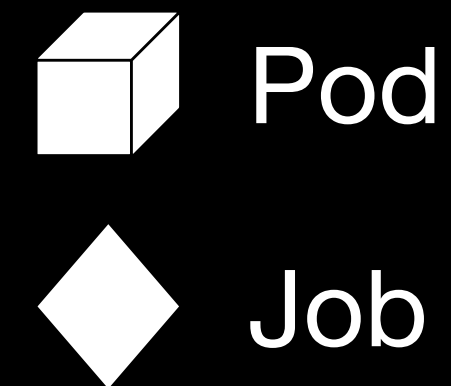
Σ Quota defined in queue

+/- Elastic quota

✨ Queue Based Features

- Lending / Borrowing limits
- Reservation / Backfill
- Preemption / Reclaim / Priority
 - [YuniKorn](#) supports inter-queue preemption only
 - [Kueue](#) preempts a Job
- Queue policy – FIFO / PriorityQueue
- Fair share

Gang-scheduling



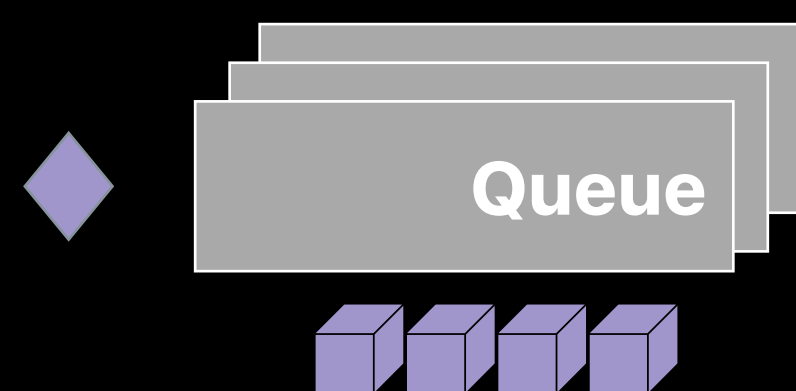
Kueue



Volcano



YuniKorn



waitForPodsReady





Ecosystem - Computing Framework



▼ Run Workloads

- Kubernetes Jobs
- Kubernetes CronJobs
- LeaderWorkerSet
- AppWrappers
- RayClusters
- RayJobs
- Deployment
- StatefulSet
- Plain Pods
- ▶ Kubeflow Jobs
 - Python
 - Jobsets
- ▶ Multi-Cluster
- ▼ External Frameworks
 - Custom Workload
 - Flux MiniClusters
 - Argo Workflow
 - Tekton Pipeline



Ecosystem

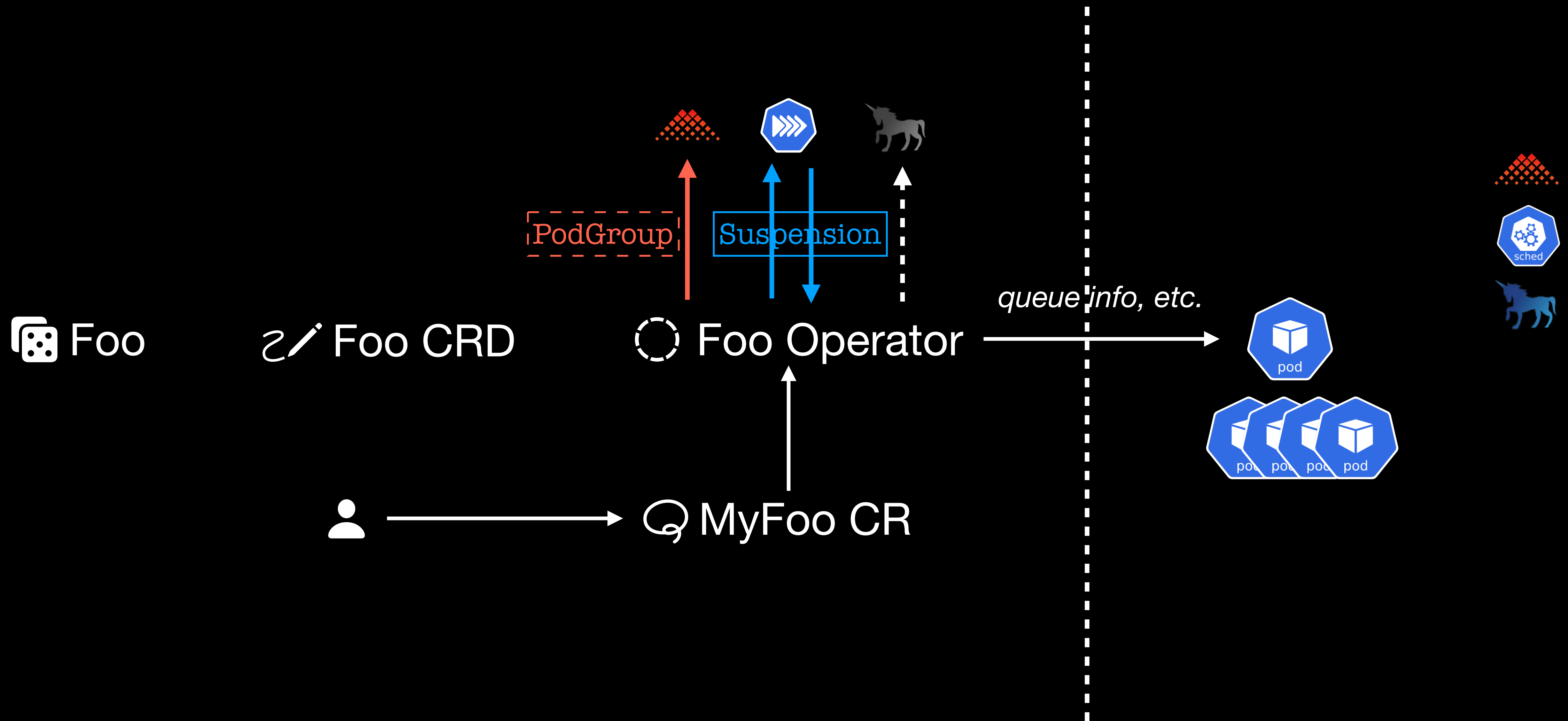
- Flink
- Kubeflow
- MindSpore
- MPI
- PaddlePaddle
- TensorFlow
- Spark
- Ray
- PyTorch
- Argo
- Horovod
- Mxnet
- KubeGene



Workloads

- Overview
- Run NVIDIA GPU Jobs
- Run Spark Jobs
- Run Flink Jobs
- Run TensorFlow Jobs
- Run MPI Jobs
- Run RayCluster
- Run RayJob
- Run RayService

Ecosystem - Computing Framework (cont.)



Ecosystem - Cluster AutoScaler

- kube-scheduler SDK is used by **Cluster AutoScaler**
- ProvisionRequest supported by **Kueue**

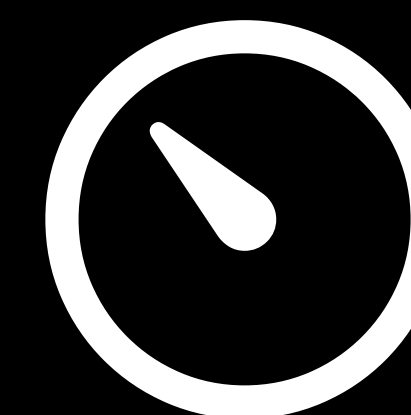


Multi-Cluster

- MultiKueue
- **Volcano global** (via Karmada)

Other Considerations

- Extensibility
- Debuggability & Operational Excellence
- Support & Maturity
- Community & Adoption
- Roadmap



Performance



Performance

← → ↻ 🔒 <https://kubernetes.io/docs/reference/command-line-tools-reference/kube-scheduler/>

kubernetes Documentation Kubernetes

🔍 Search this site

- Glossary
- ▶ API Overview
- ▶ API Access Control
- ▶ Well-Known Labels, Annotations and Taints
- ▶ Kubernetes API
- ▶ Instrumentation
- ▶ Kubernetes Issues and Security
- ▶ Node Reference Information
- ▶ Networking Reference
- ▶ Setup tools
- ▶ Command line tool (kubectl)
- ▼ Component tools
 - Feature Gates
 - Feature Gates (removed)
 - kubelet
 - kube-apiserver
 - kube-controller-manager
 - kube-proxy
 - kube-scheduler**
- ▶ Debug cluster

kube:WindowsGracefulNodeShutdown=true|false (ALPHA - default=false)
kube:WindowsHostNetwork=true|false (ALPHA - default=true)

-h, --help
help for kube-scheduler

--http2-max-streams-per-connection int
The limit that the server gives to clients for the maximum number of concurrent streams. Use go's default.

--kube-api-burst int32 Default: 100
DEPRECATED: burst to use while talking with kubernetes apiserver in --config.

--kube-api-content-type string Default: "application/vnd.kubernetes.protobuf"
DEPRECATED: content type of requests sent to apiserver. This parameter is ignored.

--kube-api-qps float Default: 50
DEPRECATED: QPS to use while talking with kubernetes apiserver in --config.

--kubeconfig string

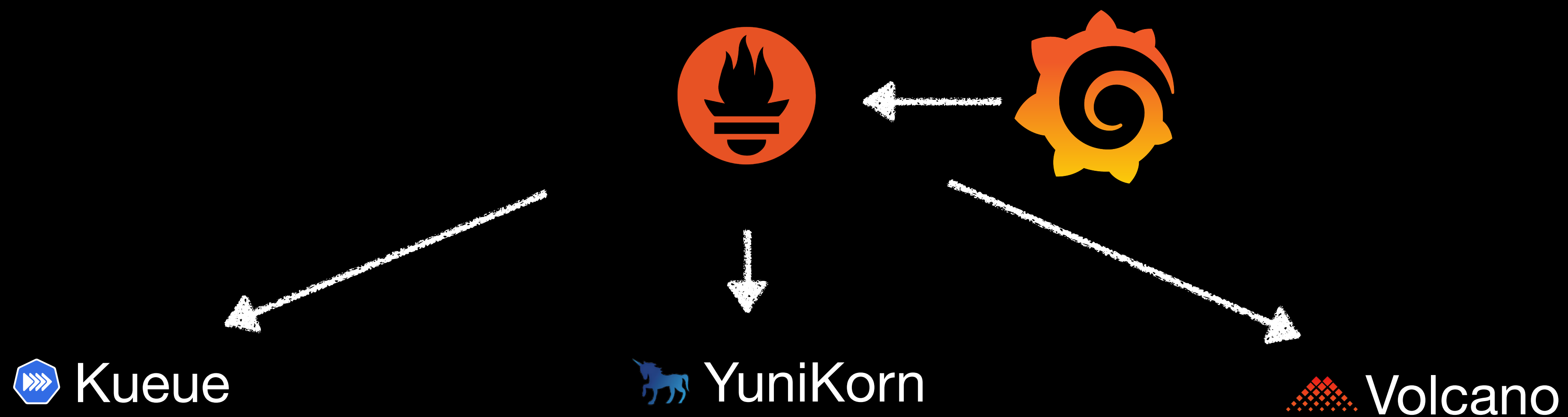
← ↻ 🔒 https://yunikorn.apache.org/docs/user_guide/service_config/#default-configmap

Apache YuniKorn Docs Roadmap Download Community ▼

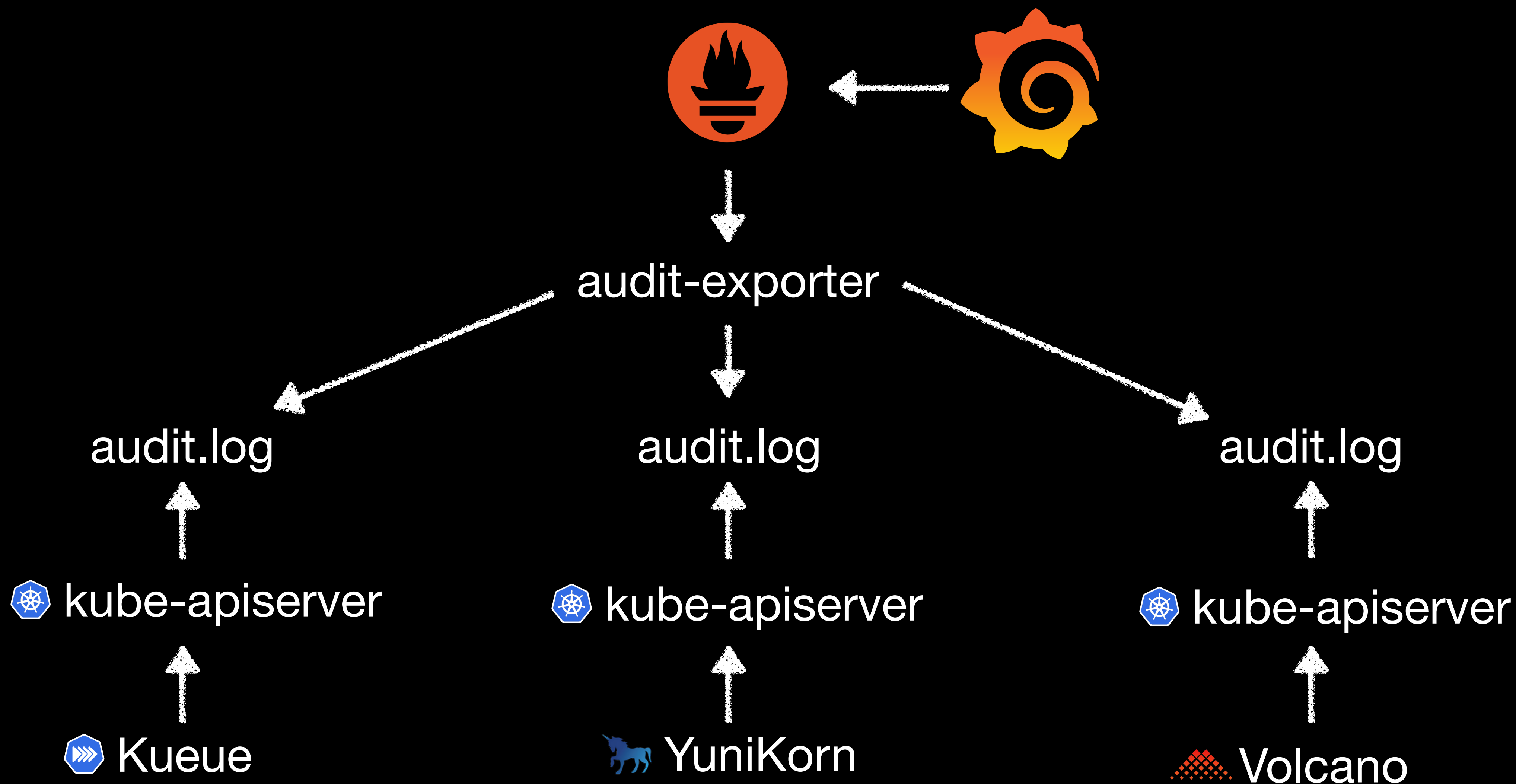
- Get Started >
- User Guide** ▼
- Deployment Modes
- Service Configuration**
- Partition and Queue Configuration
- App Placement Rules
- User & Group Resolution
- Sorting Policies
- App & Queue Priorities
- Preemption
- ACLs
- Resource Quota Management
- Gang Scheduling
- Labels and Annotations in YuniKorn

```
apiVersion: v1
kind: ConfigMap
metadata:
  name: yunikorn-configs
data:
  service.clusterId: "mycluster"
  service.policyGroup: "queues"
  service.schedulingInterval: "1s"
  service.volumeBindTimeout: "10m"
  service.eventChannelCapacity: "1048576"
  service.dispatchTimeout: "5m"
  service.disableGangScheduling: "false"
  service.enableConfigHotRefresh: "true"
  service.placeholderImage: "registry.k8s.io/pause:3.1"
  service.instanceTypeNodeLabelKey: "node.kubernetes.io/instance-type"
  health.checkInterval: "30s"
  log.level: "INFO"
  kubernetes.qps: "1000"
  kubernetes.burst: "1000"
  admissionController.webhook.amServiceName: "yunikorn-admission-controller"
  admissionController.webhook.schedulerServiceAddress: "yunikorn-scheduler:8080"
  admissionController.filtering.processNamespaces: ""
  admissionController.filtering.bypassNamespaces: "^kubernetes.default$"
  admissionController.filtering.labelNamespaces: ""
  admissionController.filtering.noLabelNamespaces: ""
```






How to Collect Metrics?



How to Collect Metrics?



Benchmark Environment

 Kubernetes 1.32.2		
Job Controller	Pod Scheduler	Node Simulator
 Kueue(0.10.3)	kube-scheduler	 KWOK
	Coscheduling(0.30.6 with a bugfix)	
kube-controller-manager	 YuniKorn(1.6.2)	
 Volcano(1.12.0-alpha.0)		

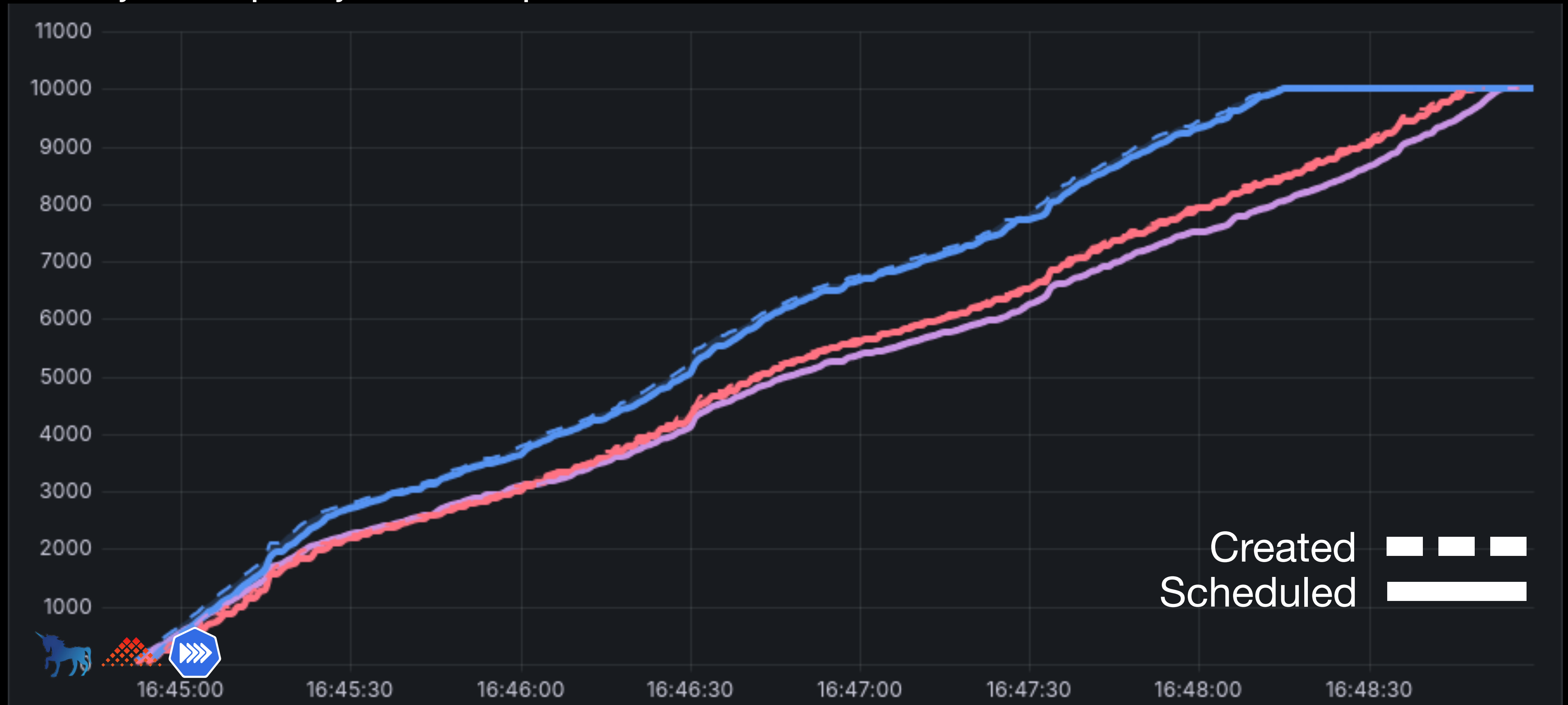
Benchmark Cases

Gang Scheduling	Jobs per Queue	Pods per Job	Total Pods
Off	10k	1	10k
Off	500	20	
Off	20	500	
Off	1	10k	
On	10k	1	
On	500	20	
On	20	500	
On	1	10k	



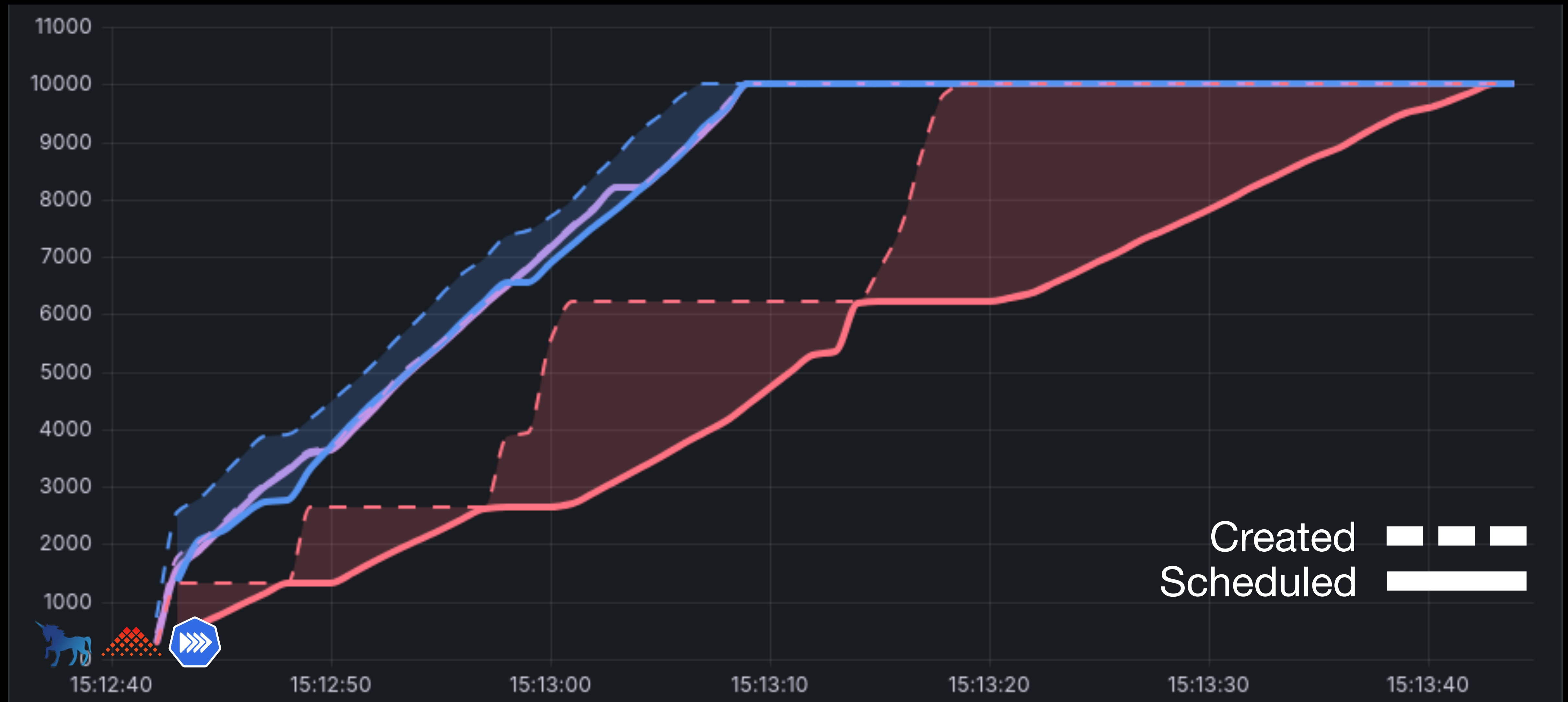
Job Submission

10k job * 1 pod/job = 10k pod, QPS: 1k, Node: 1k



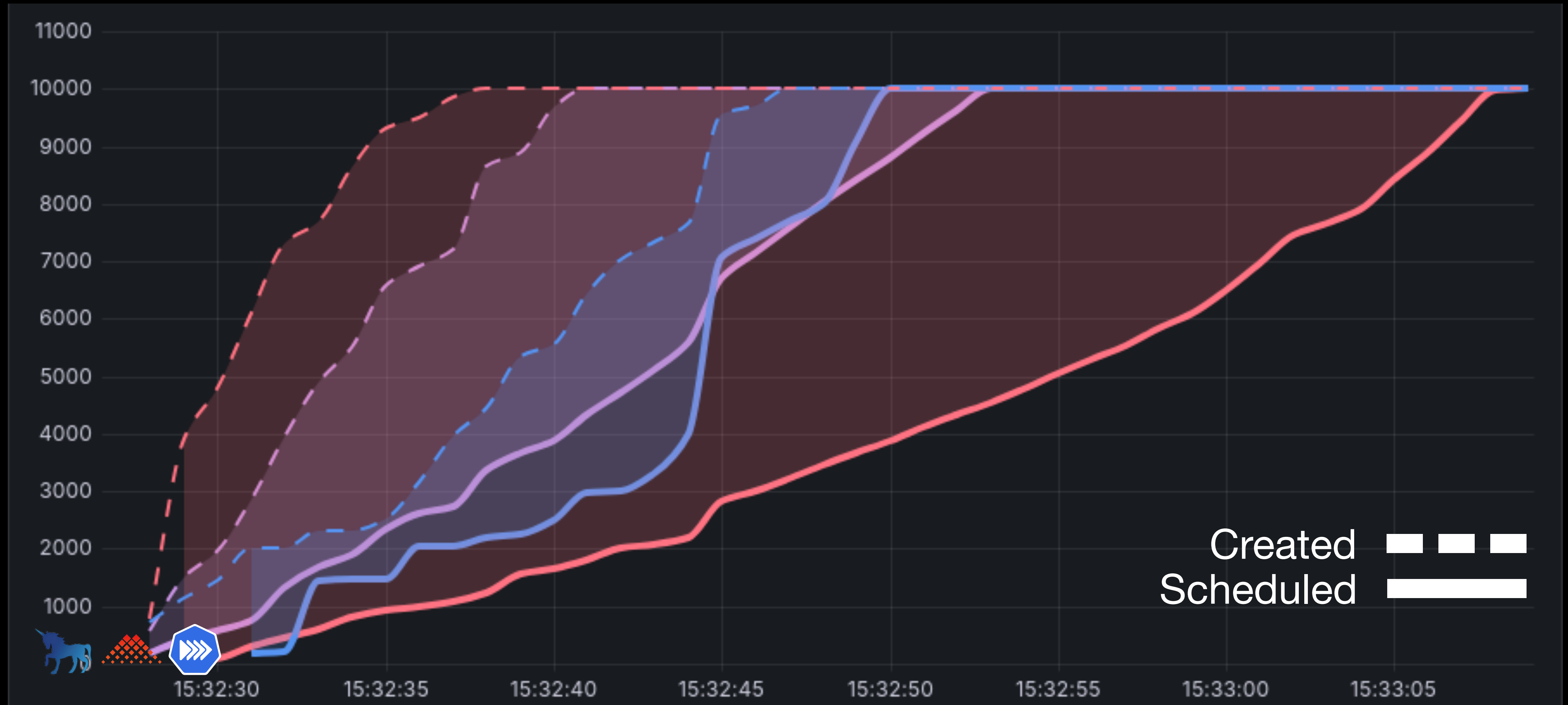
Job Submission

500 job * 20 pod/job = 10k pod, QPS: 1k, Node: 1k



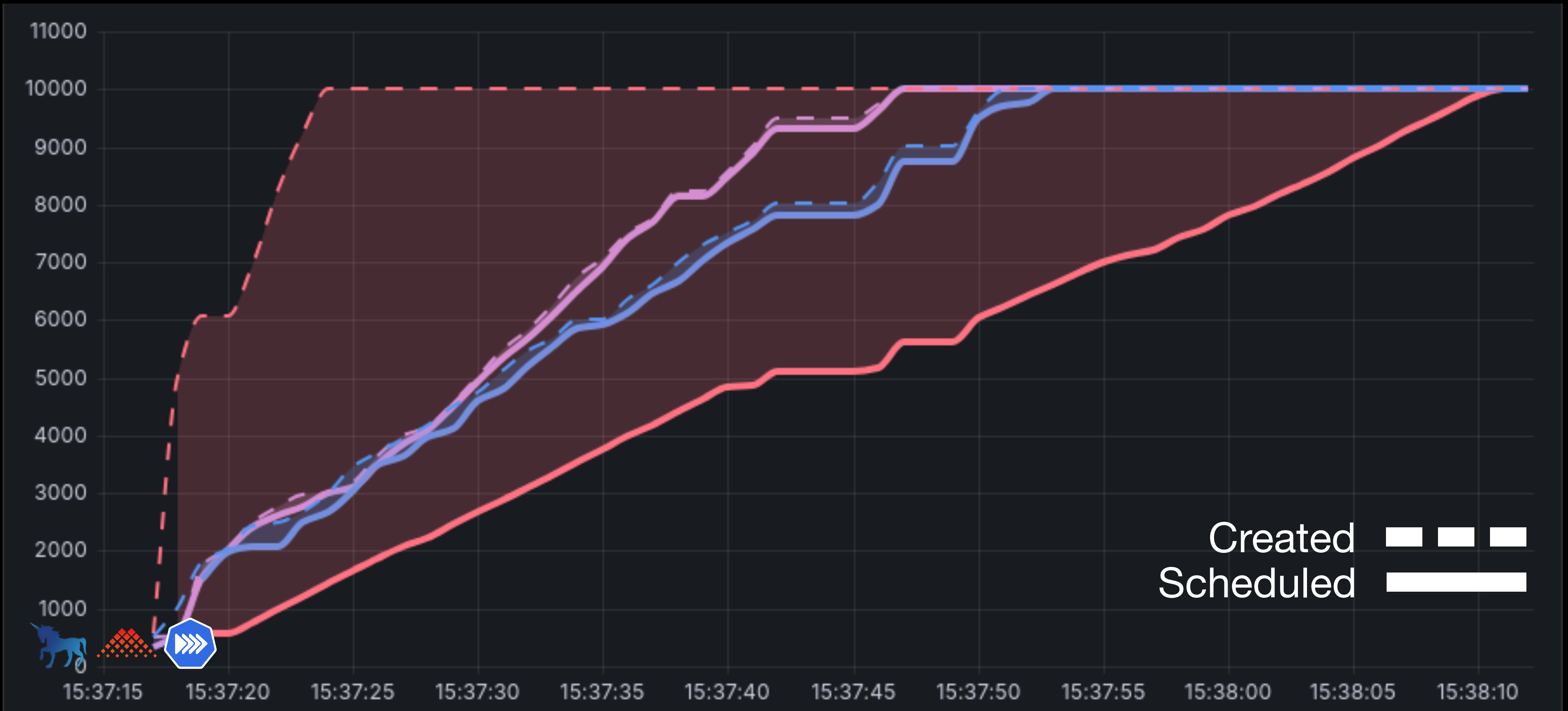
Job Submission

20 job * 500 pod/job = 10k pod, QPS: 1k, Node: 1k



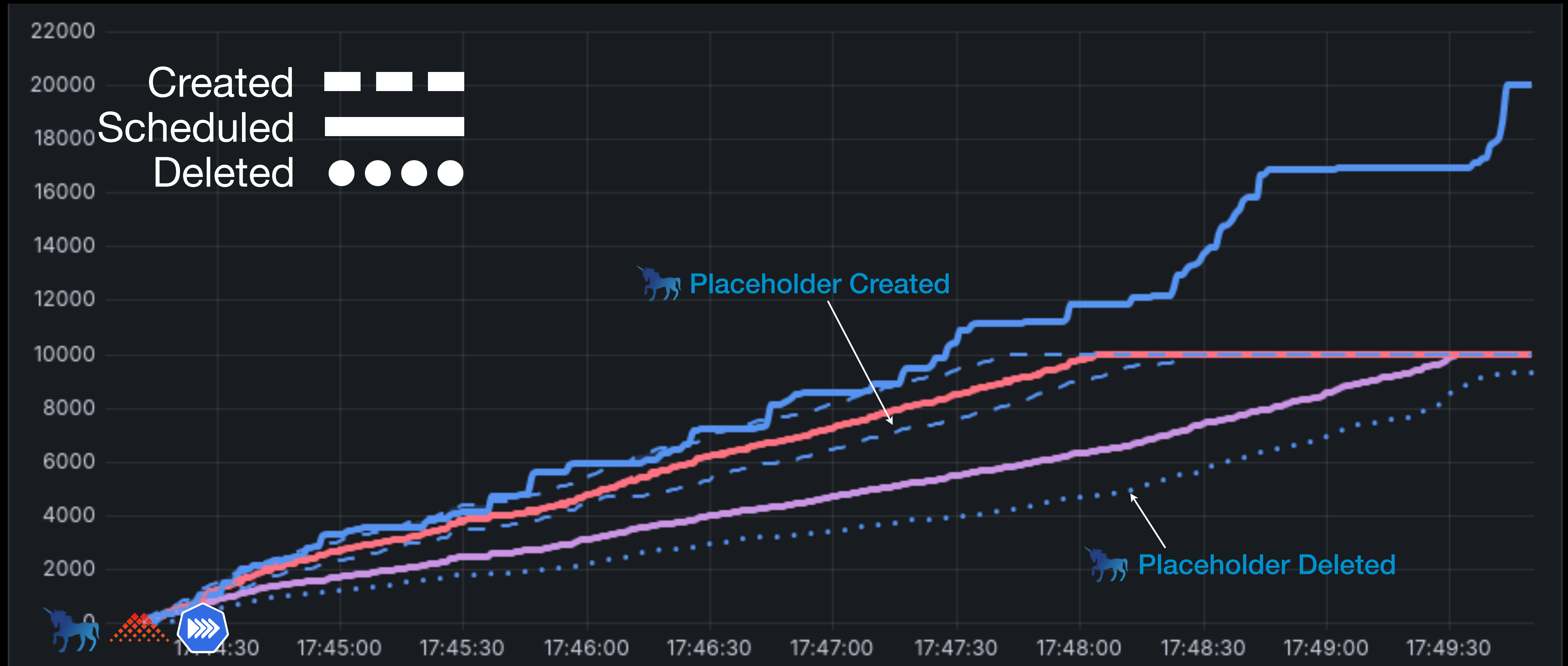
Job Submission

1 job * 10k pod/job = 10k pod, QPS: 1k, Node: 1k



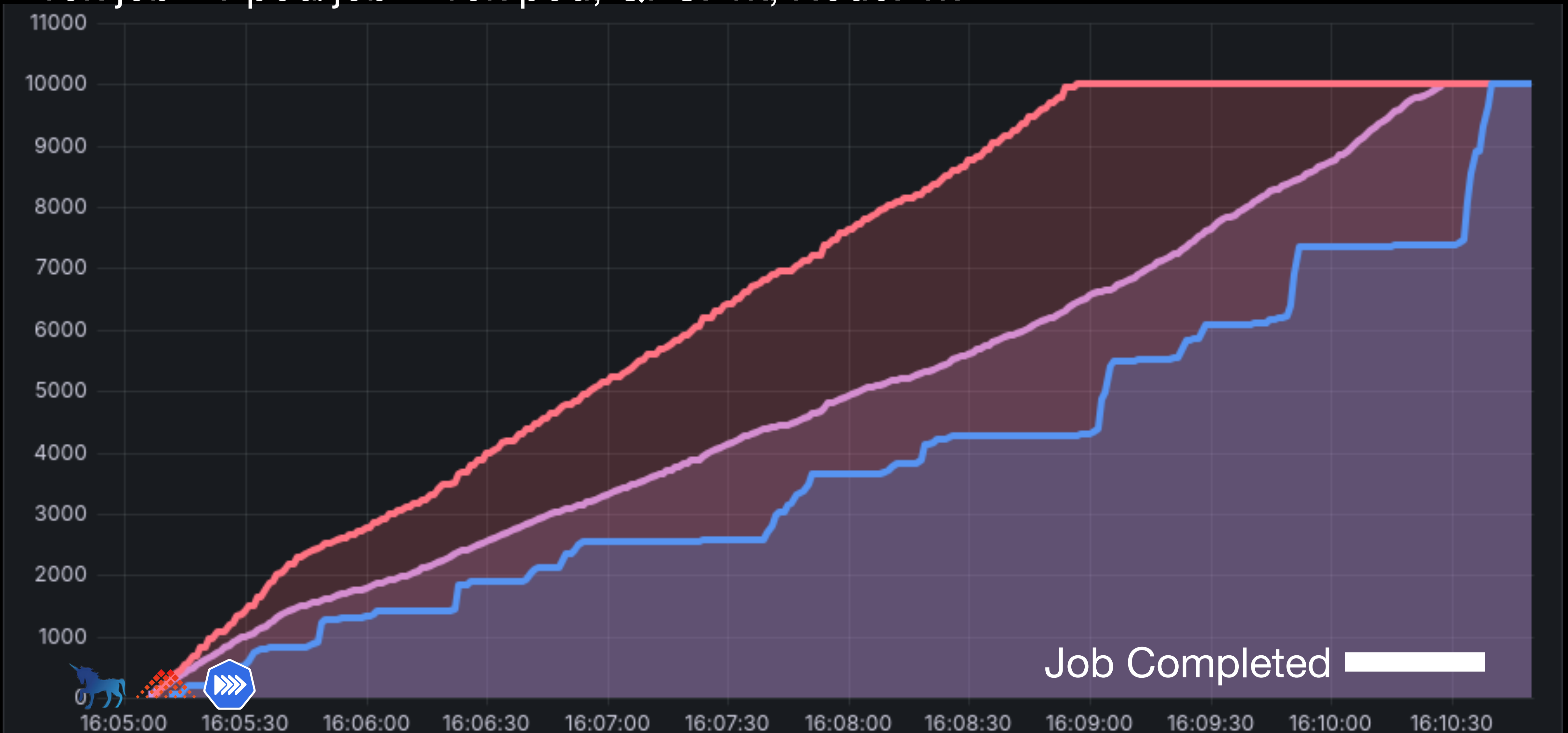
Job Submission with Gang Scheduling

10k job * 1 pod/job = 10k pod, QPS: 1k, Node: 1k



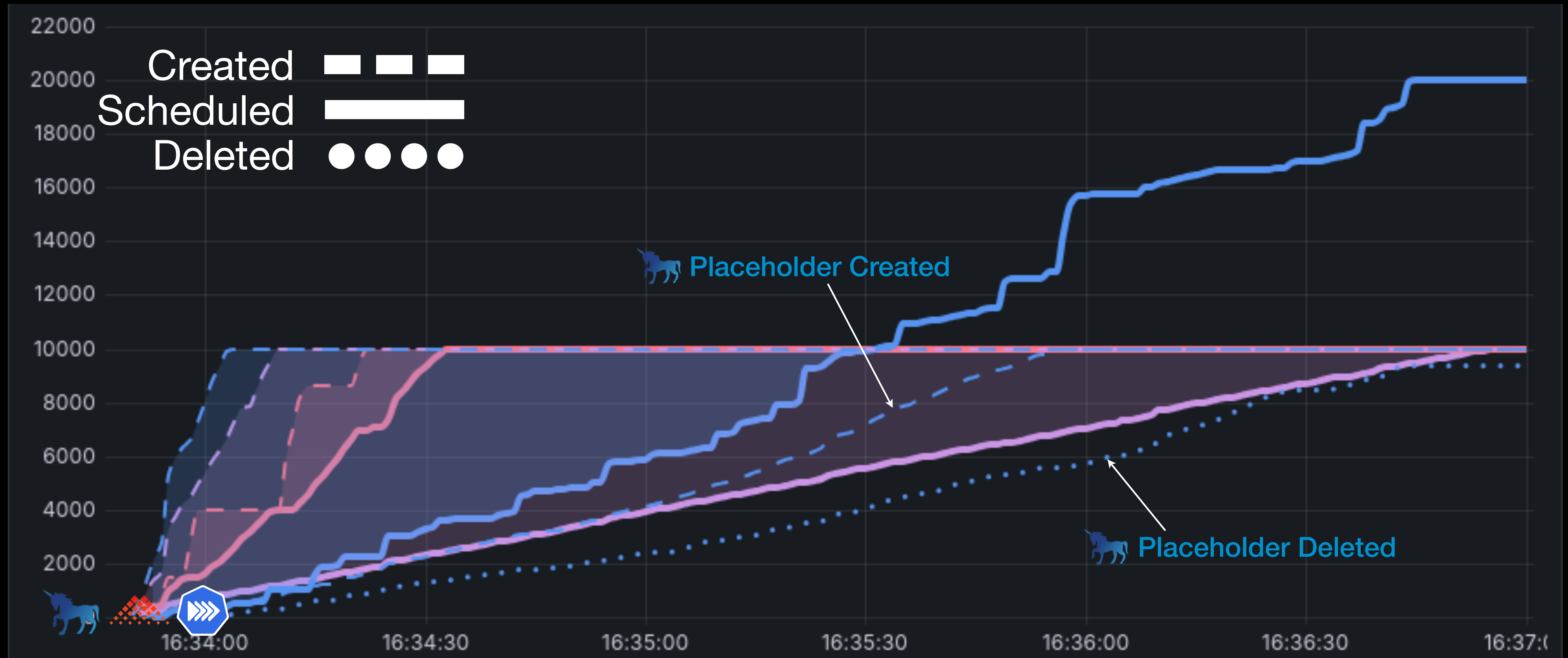
Job Submission with Gang Scheduling

10k job * 1 pod/job = 10k pod, QPS: 1k, Node: 1k



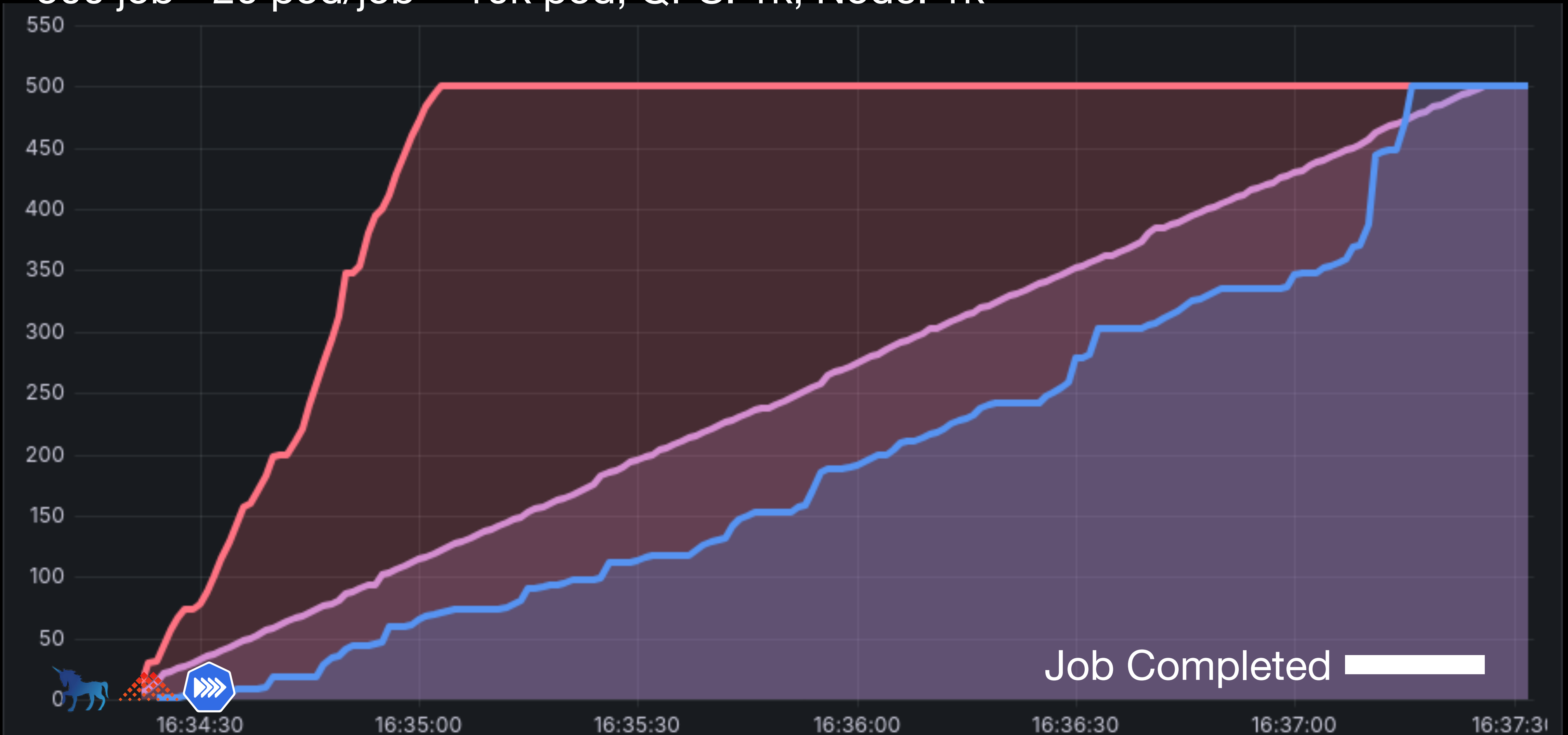
Job Submission with Gang Scheduling

500 job * 20 pod/job = 10k pod, QPS: 1k, Node: 1k



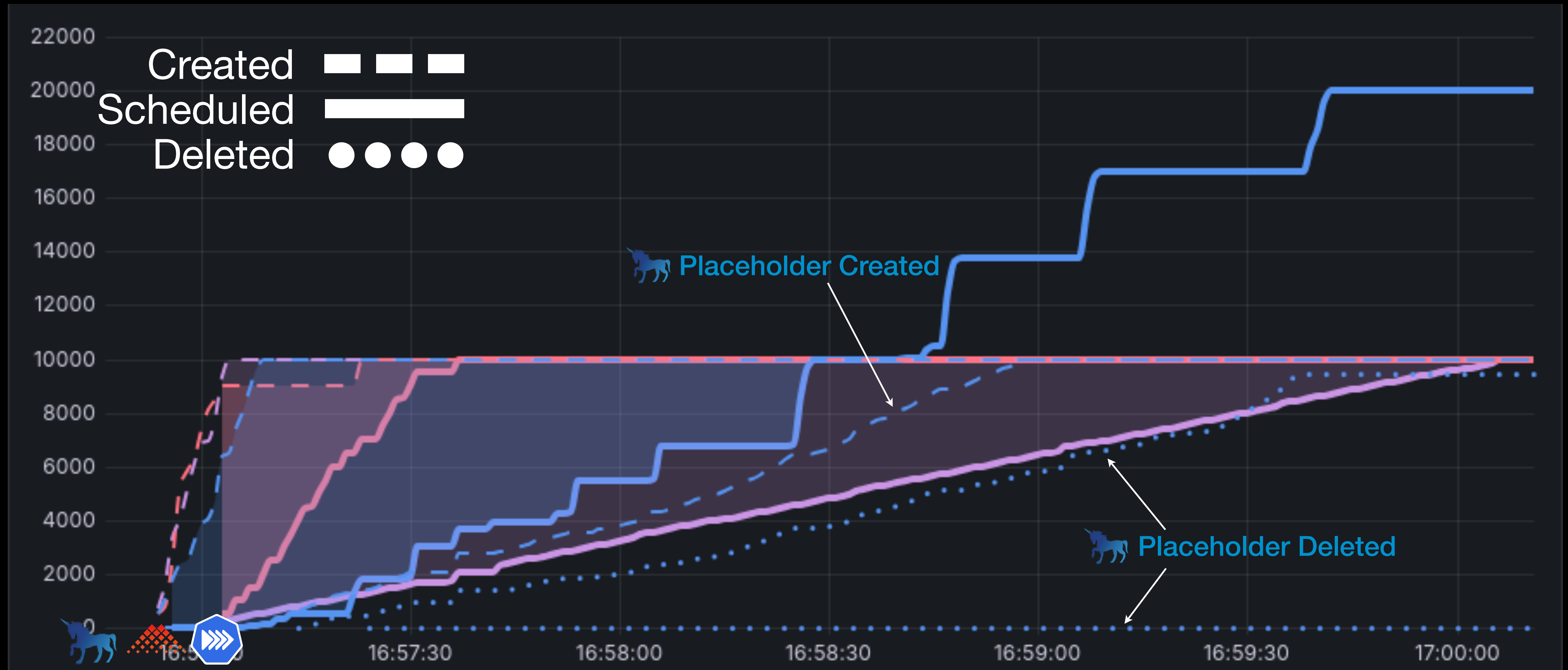
Job Submission with Gang Scheduling

500 job * 20 pod/job = 10k pod, QPS: 1k, Node: 1k



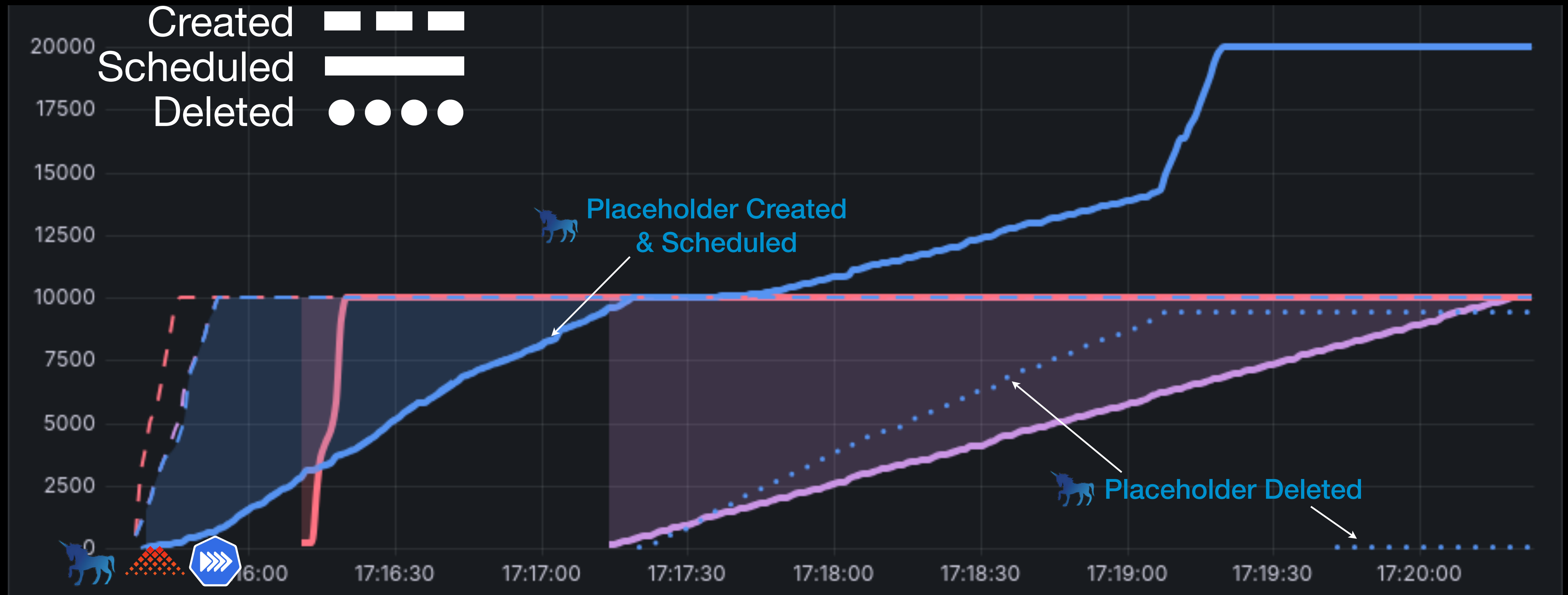
Job Submission with Gang Scheduling

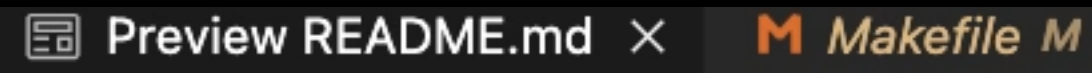
20 job * 500 pod/job = 10k pod, QPS: 1k, Node: 1k



Job Submission with Gang Scheduling

1 job * 10k pod/job = 10k pod, QPS: 1k, Node: 1k





> OPEN EDITORS

```
> base
```

```
> bin
```

- clusters

> kueue

> overview

> volcano

- > yunikorn

```
> GOPATH
```

> hack

```
> logs
```

```
> registry-data
```

> results

> schedulers

```
> test
```

```
> tmp
```

 .gitignore

! .yamlfmt.yaml

! audit-policy.yaml

 [go.mod](#)

≡ go.sum

 LICENSE

M Makefile

[i](#) README.md

M

M

> OUTLINE

> TIMELINE

- Kueue
- Volcano
- YuniKorn

Prerequisites and Setup

This project allows for quick deletion and recreation of the cluster, as it uses the kind local registry.

This approach also helps control variables during different scheduler tests and does not have to wait for the slow deletion process.

Dependencies

TERMINAL

PROBLEMS

OUTPUT

PORTS 5

>

>

>

✓ **TERMINAL**





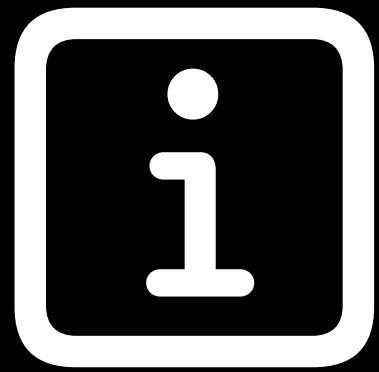

Performance Roadmap

- Align Features
 - Preemption / Priority
 - Restricted Quota
 - Lend / Borrow
 - Resource conflicts
 - ...
- Automatically run benchmarks for the latest versions
 - Github Actions / Github Pages
- More Scenarios
- ...

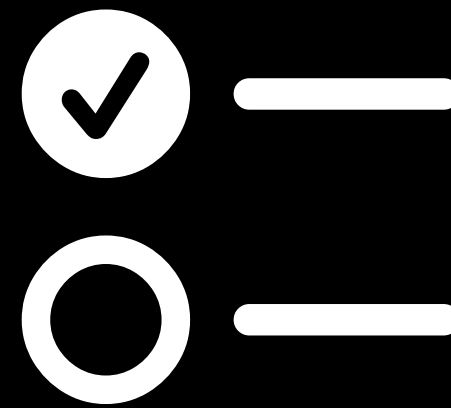
<https://github.com/wzshiming/kube-scheduling-perf>



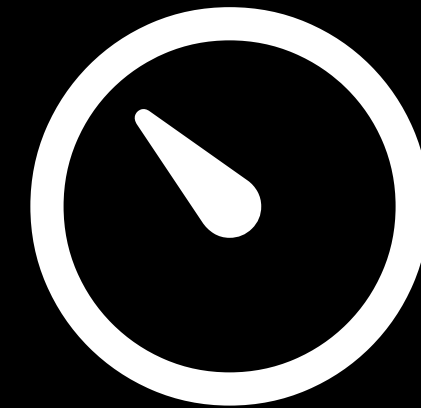
Wrap-up



Overview



Comparison



Performance

 **Thanks**

- **Weiwei Yang (YuniKorn)**
- **Yuki Iwai (Kueue)**
- **Xuzheng Chang (Volcano)**