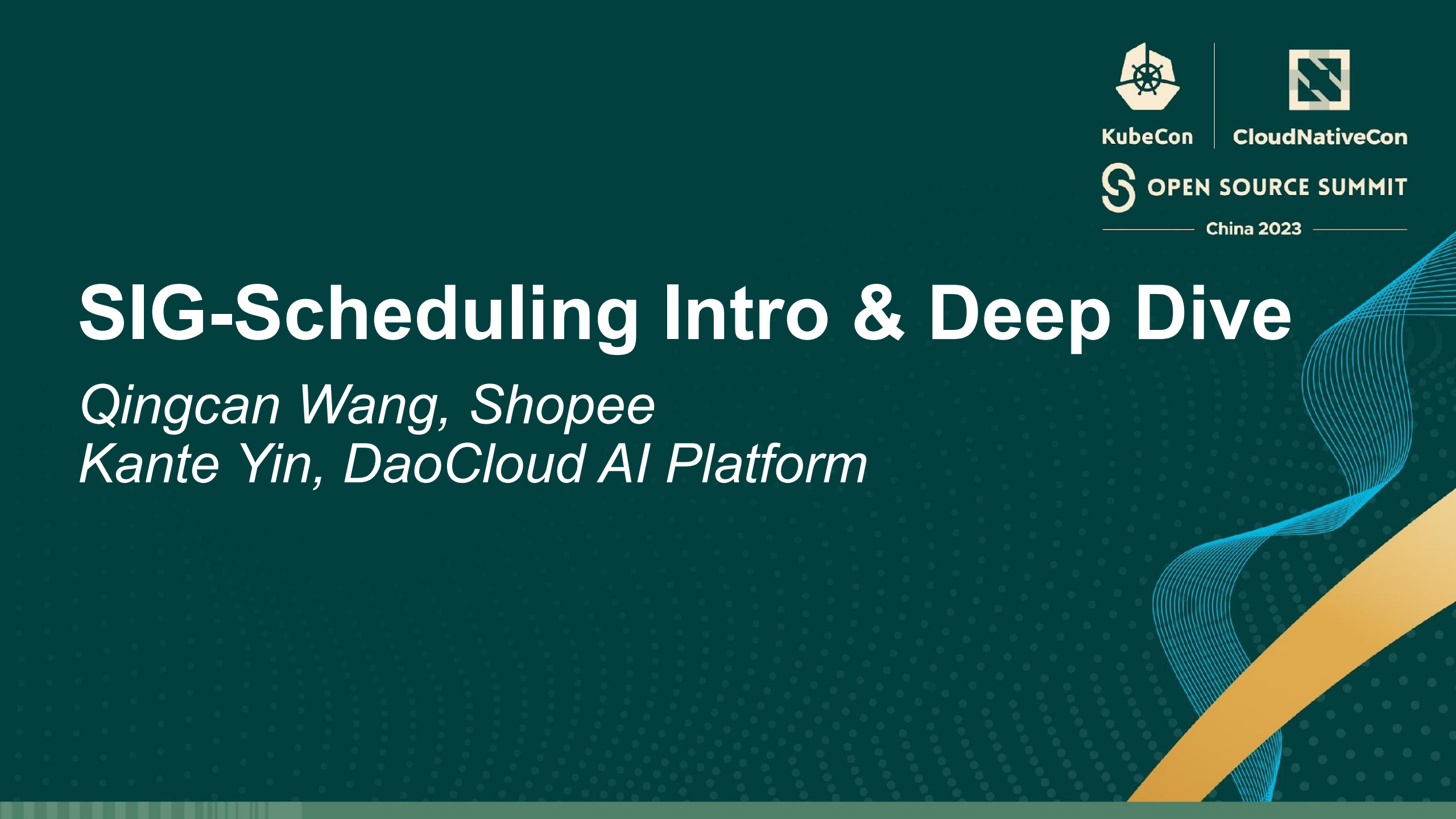

KubeCon
CloudNativeCon
OPEN SOURCE SUMMIT
China 2023



SIG-Scheduling Intro & Deep Dive

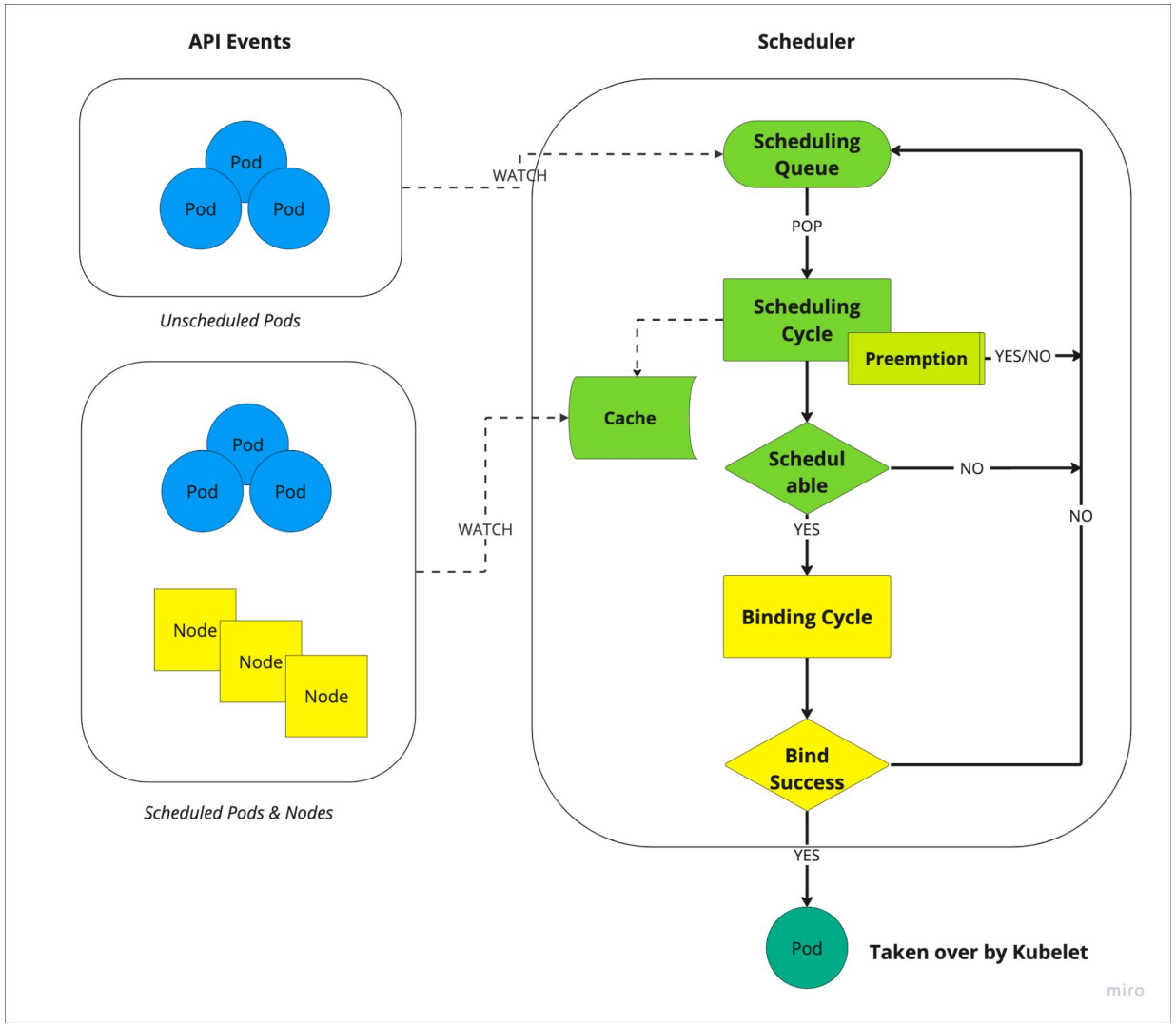
Qingcan Wang, Shopee
Kante Yin, DaoCloud AI Platform

Agenda

- Scheduler Overview
- Updates & Improvements
- Sub-projects Updates
- Join us
- Q & A

Scheduler Overview

Scheduler Overview

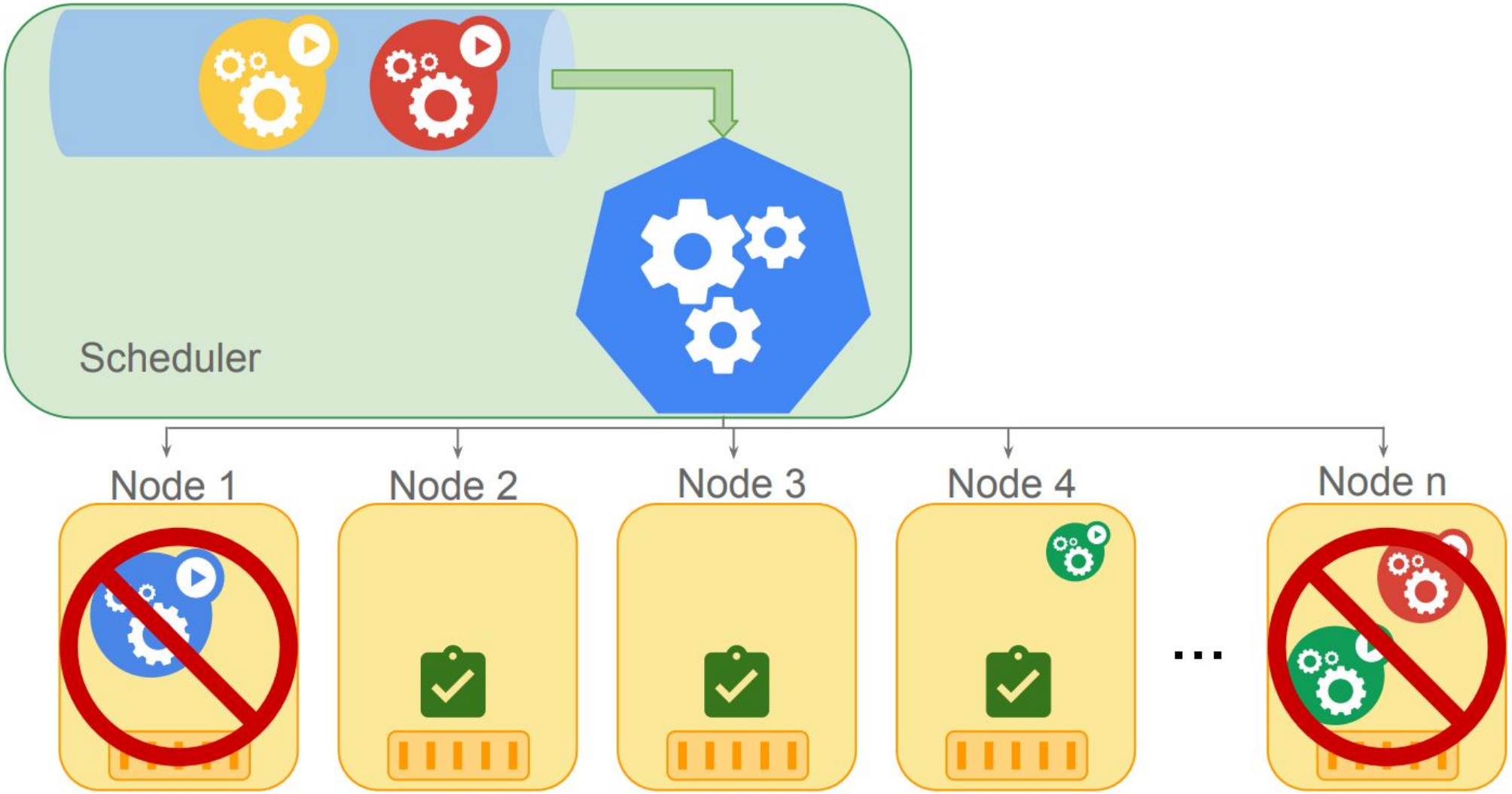


In Kubernetes, scheduling refers to making sure that Pods are matched to Nodes so that Kubelet can run them

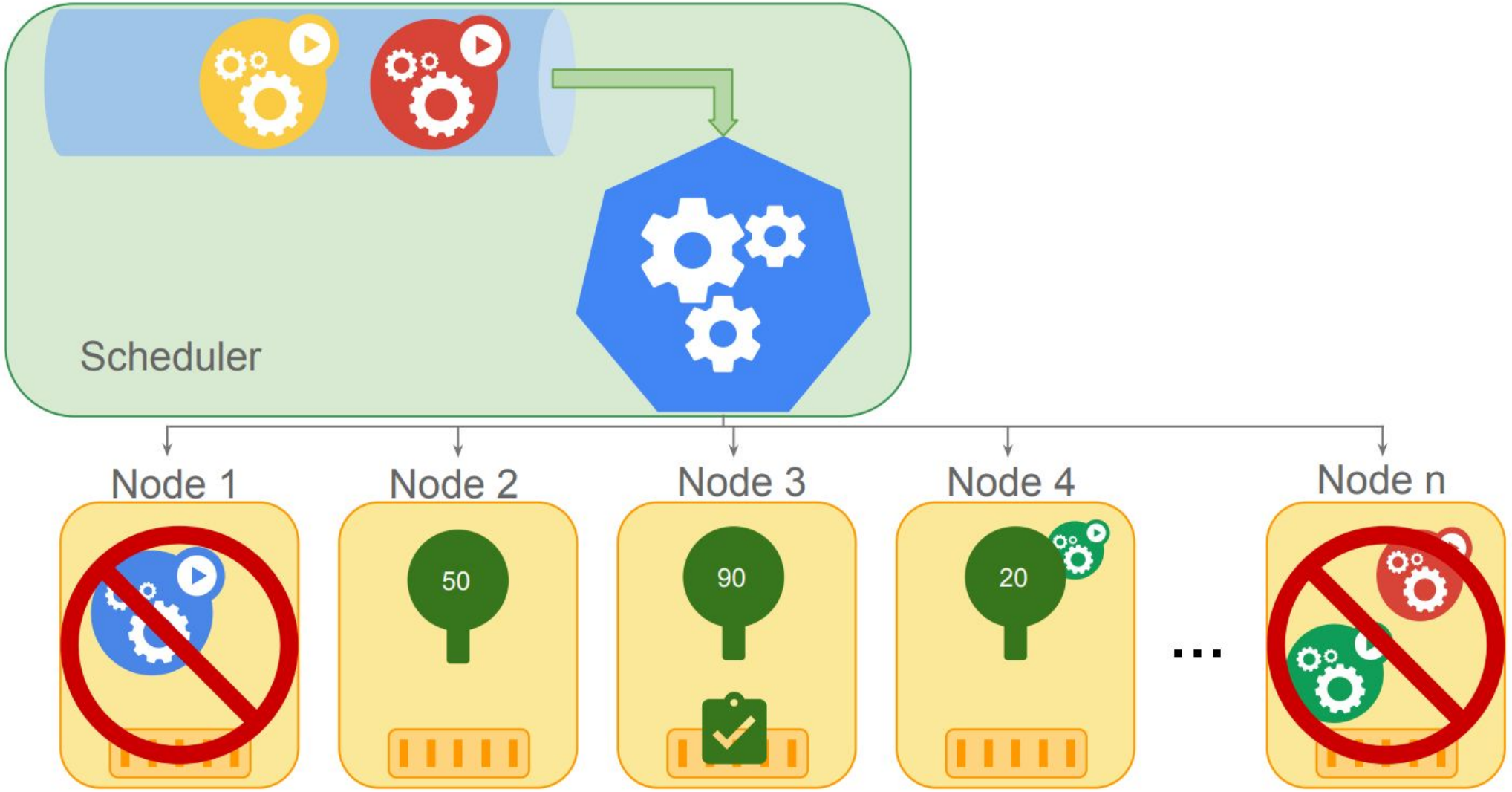
2-step operation

- Filtering: finds the set of Nodes where it's feasible to schedule the Pod
- Scoring: ranks the remaining nodes to choose the most suitable Pod placement

Filtering



Scoring

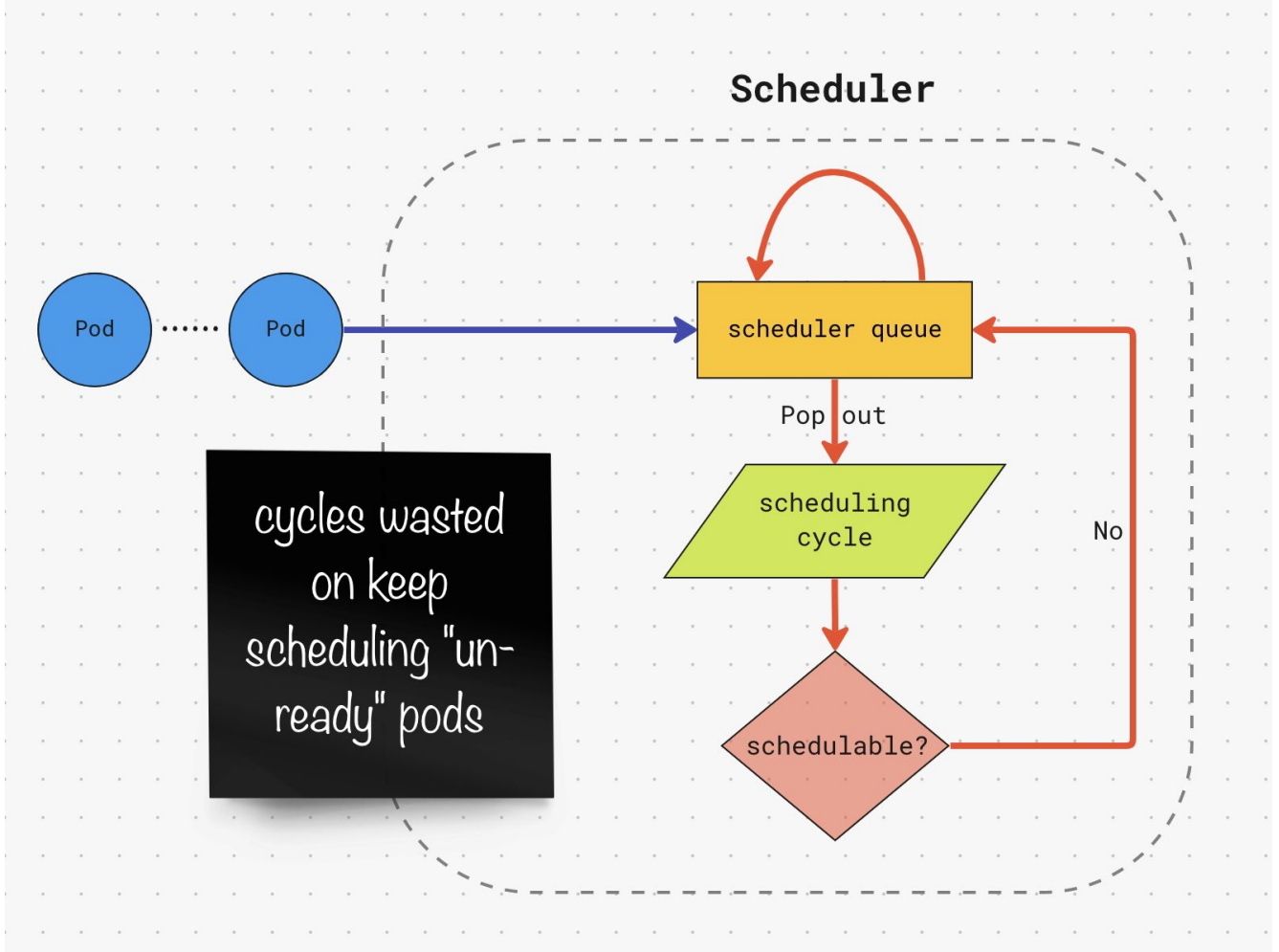


Recent Updates

– since v1.25

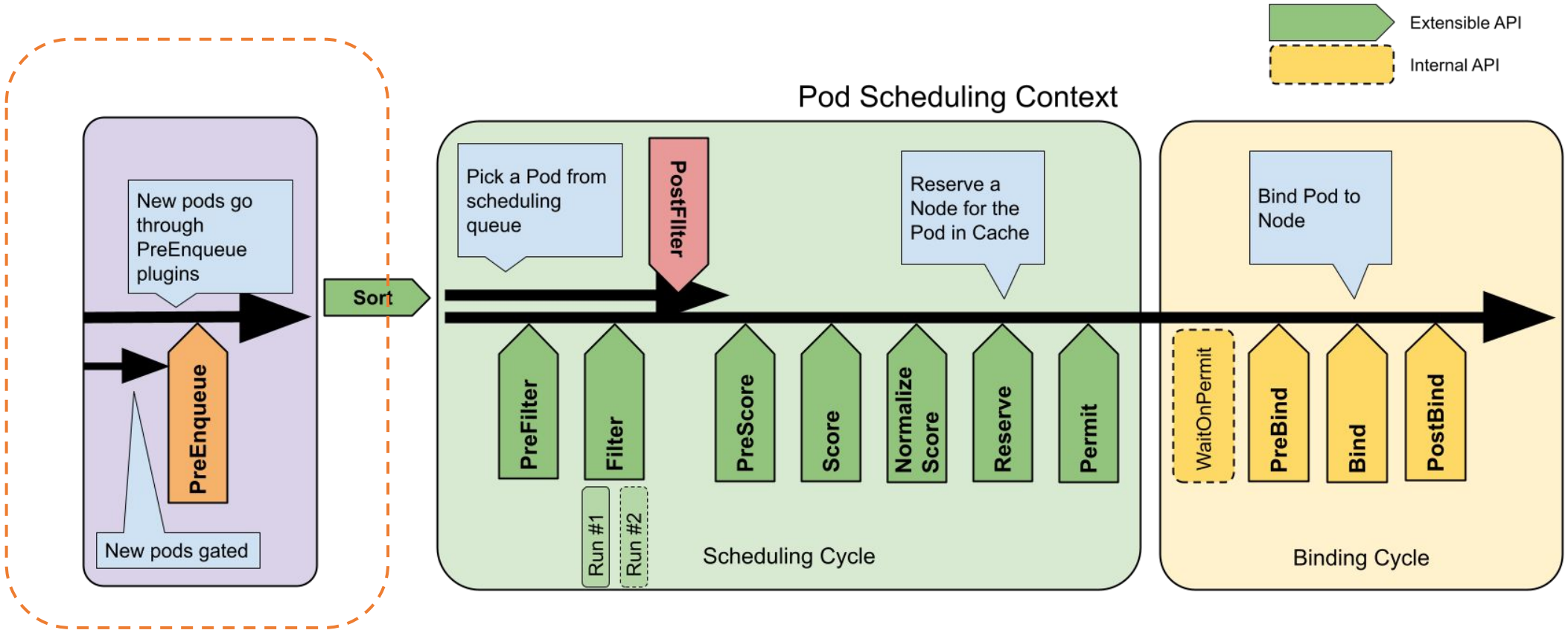
SchedulingGates Plugin

When pods not ready for scheduling,
e.g. miss essential resources.



SchedulingGates Plugin

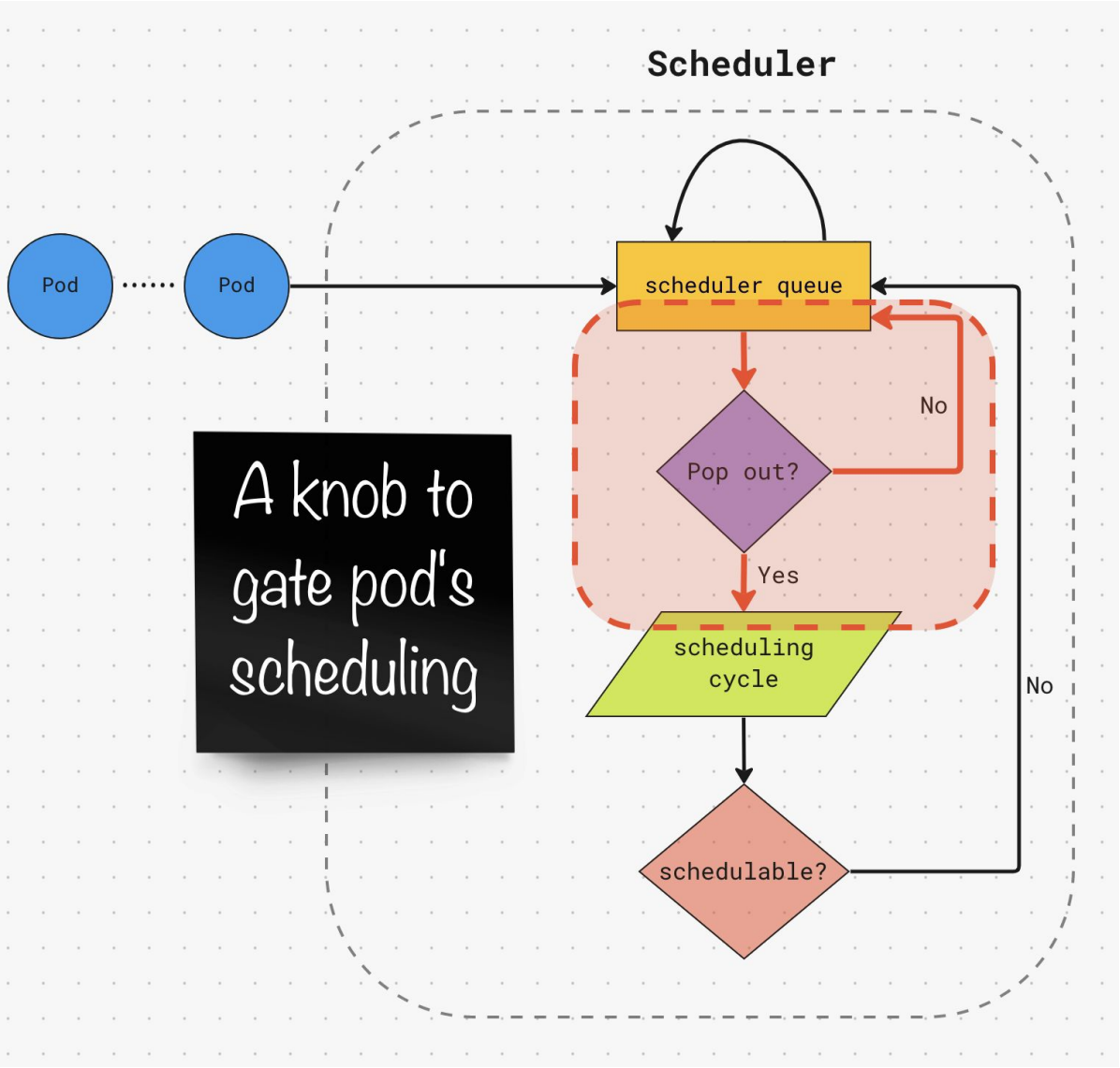
A new extension point **PreEnqueue** introduced in scheduler framework helps to gate pods entering into active queues.



SchedulingGates Plugin


```yaml
apiVersion: v1
kind: Pod
metadata:
  name: nginx
spec:
  schedulingGates:
    - name: example.com/foo
  containers:
  - name: nginx
    image: nginx:1.14.2
    ports:
    - containerPort: 80
```

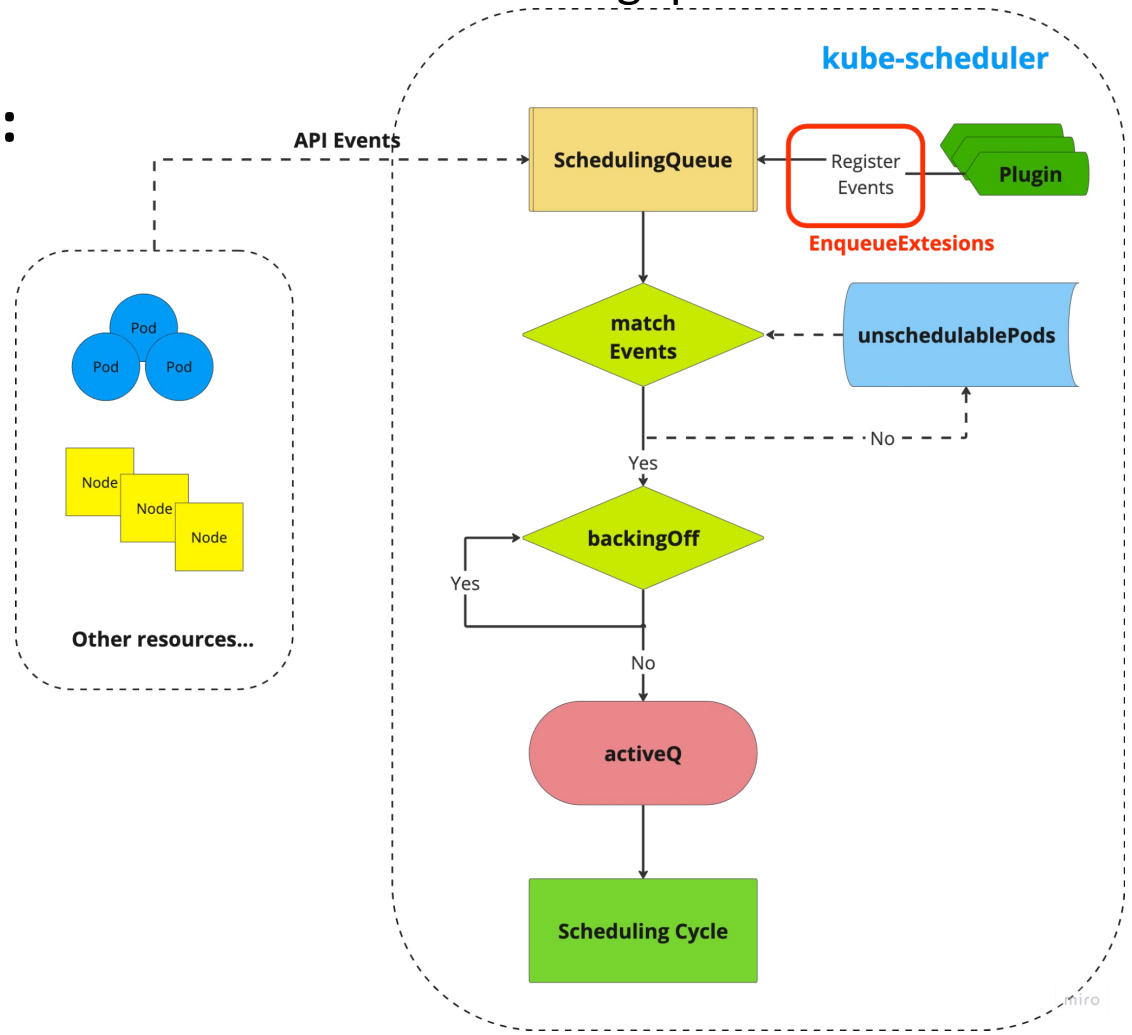
Inspired by PreEnqueue extension point.



Scheduler Framework

Break changes to **EnqueueExtensions**, which shapes functions to influence whether we should move unschedulable Pods to internal scheduling queues based on the cloud events.

Rescheduling Before:



Problem: roughly trigger the rescheduling cycle for unschedulable Pods, e.g. Pod rejected by *nodeAffinity plugin* doesn't pay for all the label updates.

Scheduler Framework


```
type EnqueueExtensions interface {
    EventsToRegister() []ClusterEvent
}
```

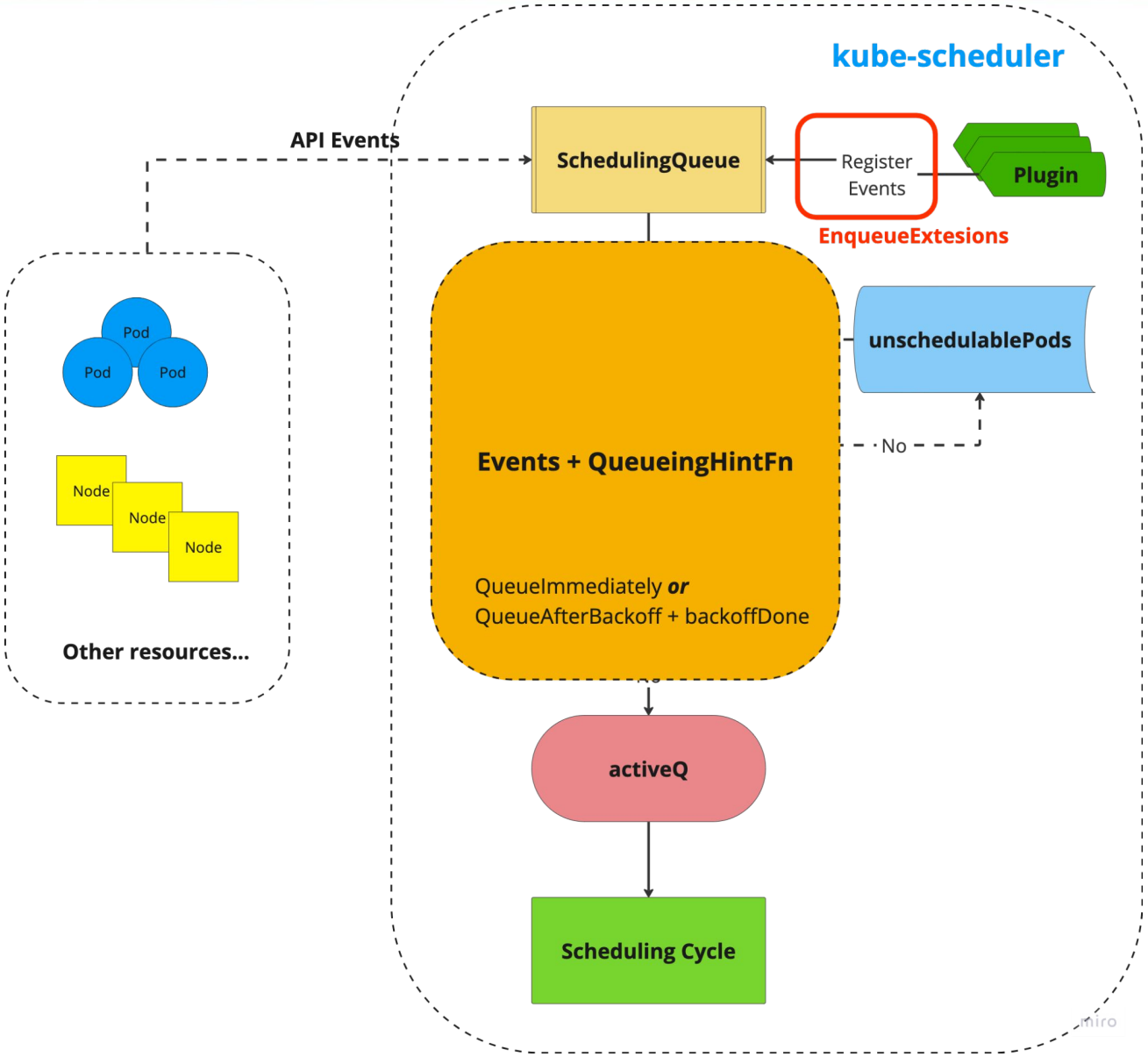
```
type EnqueueExtensions interface {
    Plugin
    EventsToRegister() []ClusterEventWithHint
}
type ClusterEventWithHint struct {
    Event ClusterEvent
    QueueingHintFn QueueingHintFn
}
```

```
type QueueingHintFn func(logger klog.Logger, pod *v1.Pod, oldObj, newObj interface{}) QueueingHint

const (
    QueueSkip QueueingHint = iota # still unschedulable, skip the scheduling cycle
    QueueAfterBackoff # maybe schedulable, we don't know, let's have a try
    QueueImmediately # a good chance it will be schedulable, let's go ahead
)
```

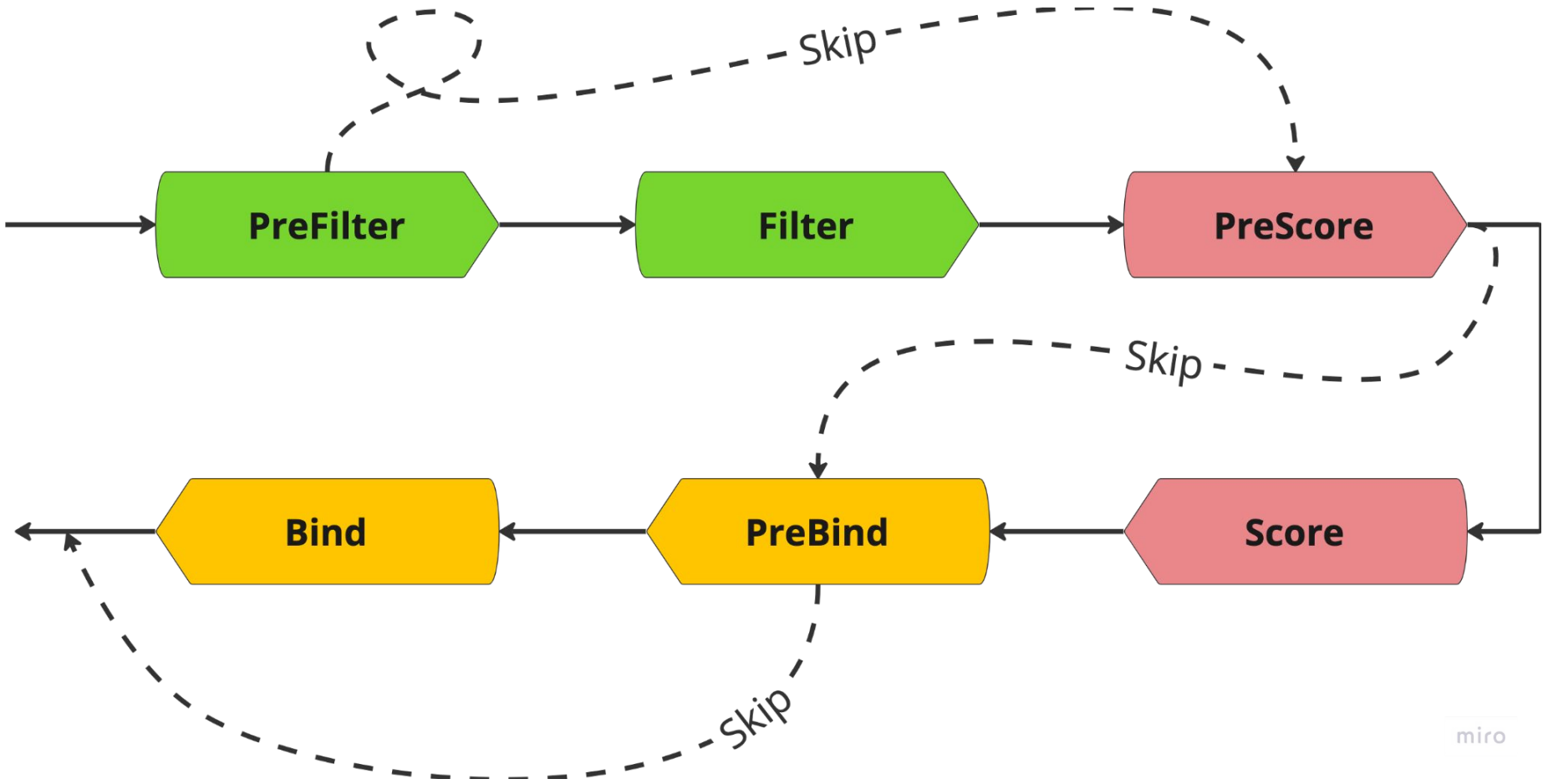
Scheduler Framework

We are now:



Scheduler Framework

Support **skip** operation in Filter & Score & Bind stage for the sake of performance



ComponentConfig

- ComponentConfig is stable now, v1 is available
- ComponentConfig v1beta2, v1beta3 is removed

```
apiVersion: kubescheduler.config.k8s.io/v1
kind: KubeSchedulerConfiguration
profiles:
  - plugins:
      score:
        disabled:
        - name: PodTopologySpread
        enabled:
        - name: MyCustomPluginA
          weight: 1
```

PodTopologySpread Plugin

- **minDomains(beta)**: define the minimum number of topology domains
- **nodeAffinityPolicy/nodeTaintsPolicy(beta)**: take taints/tolerations into consideration when calculating PodTopologySpread skew
- **matchLabelKeys(beta)**: respect podTopologySpread after rolling upgrades

```
topologySpreadConstraints:
  - maxSkew: <integer>
    minDomains: <integer> # optional; beta since v1.25
    topologyKey: <string>
    whenUnsatisfiable: <string>
    labelSelector: <object>
    matchLabelKeys: <list> # optional; beta since v1.27
    nodeAffinityPolicy: [Honor|Ignore] # optional; beta since v1.26
    nodeTaintsPolicy: [Honor|Ignore] # optional; beta since v1.26
```

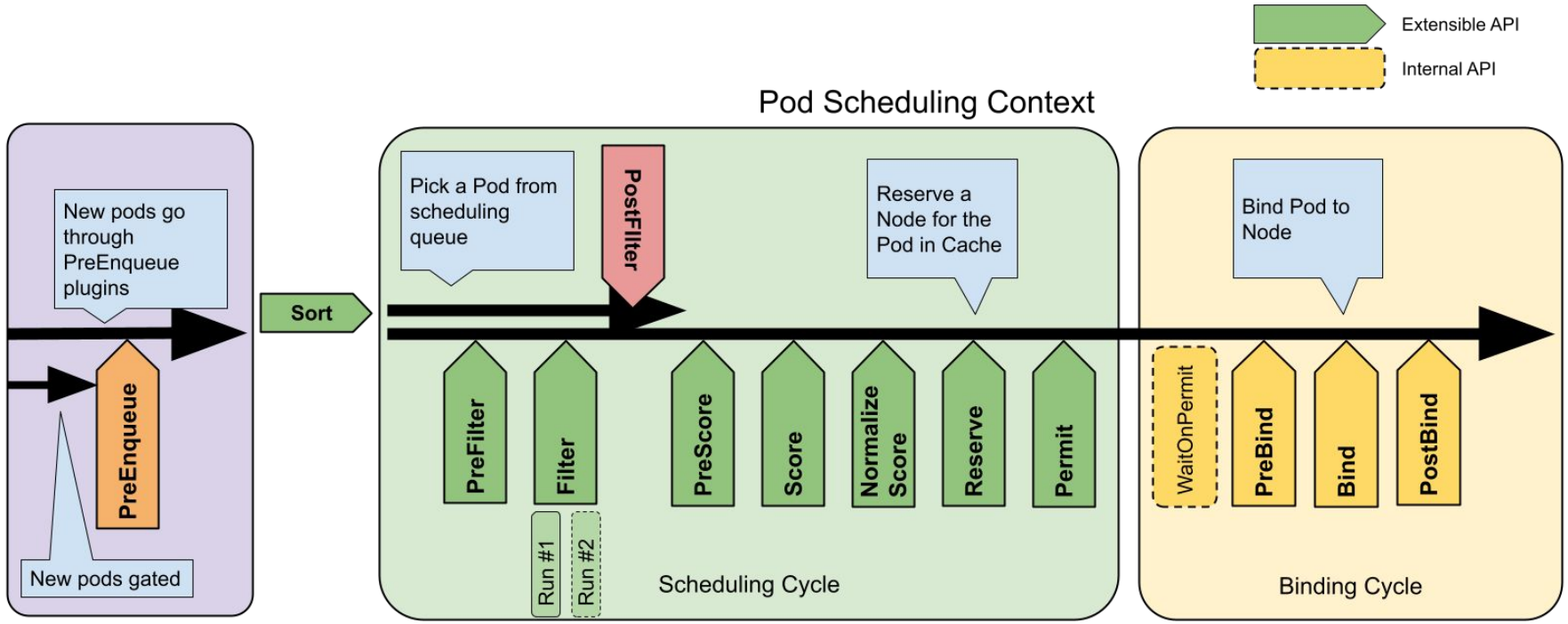
References:

- Website
- Blog

Sub-projects Updates

Scheduler-plugin

💪 **Incorporates best practices and utilities to compose a high-quality out-tree scheduler plugins based on scheduling framework**



Scheduling Framework

Scheduler-plugin

Already available, please try them

- Coscheduling/Gang Scheduling
- Capacity Scheduling Elastic resource quotas to enhance resource utilization
- Node Resource
- Node Resource Topology Scheduling based on node resource topology
- Preemption Toleration
- Trimaran Scheduling based on real node load
- Network-Aware Scheduling

Updates

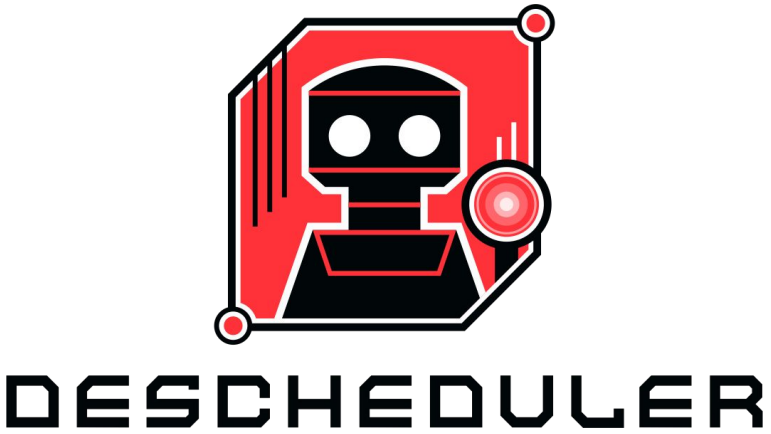
Still under discussion, please share your ideas:

- Disk IO Aware Scheduling
- Resource Policy

Main updates:

- 🔑 Kubernetes dependency bumped to v1.26.7
- 🔑 Migrated to controller-runtime
- 👉 Supported LeastNUMANodes in NodeResourceTopology plugin
- 🔔 The API Group of CRD PodGroup and ElasticQuota is migrated to scheduling.x-k8s.io
- 👉 Add a new coscheduling plugin argument podGroupBackoffSeconds to configure backoff timer for failed PodGroup

Descheduler



👉 **Rebalance clusters by evicting pods that can potentially be scheduled on better nodes**

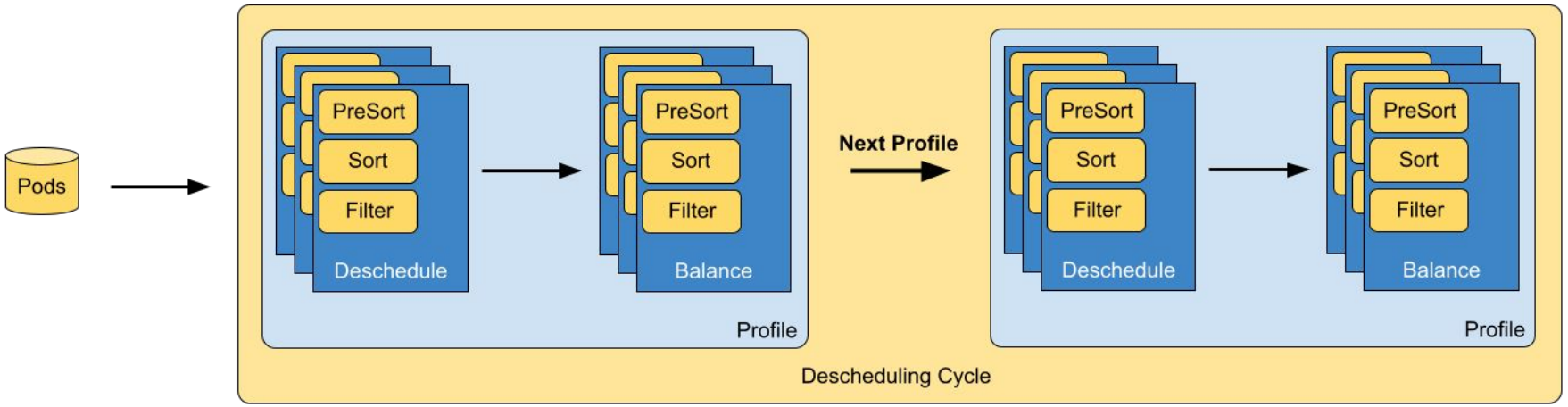
Deschedule

- RemovePodsViolatingInterPodAntiAffinity
- RemovePodsViolatingNodeAffinity
- RemovePodsViolatingNodeTaints
- RemovePodsHavingTooManyRestarts
- RemoveFailedPods

Balance

- RemoveDuplicates
- LowNodeUtilization
- HighNodeUtilization
- RemovePodsViolatingTopologySpreadConstraint

Descheduler



Multi Profile

+

Extension Plugins

- Filter
- Deschedule
- Balance
- Evict

bit.ly/descheduler-fwk

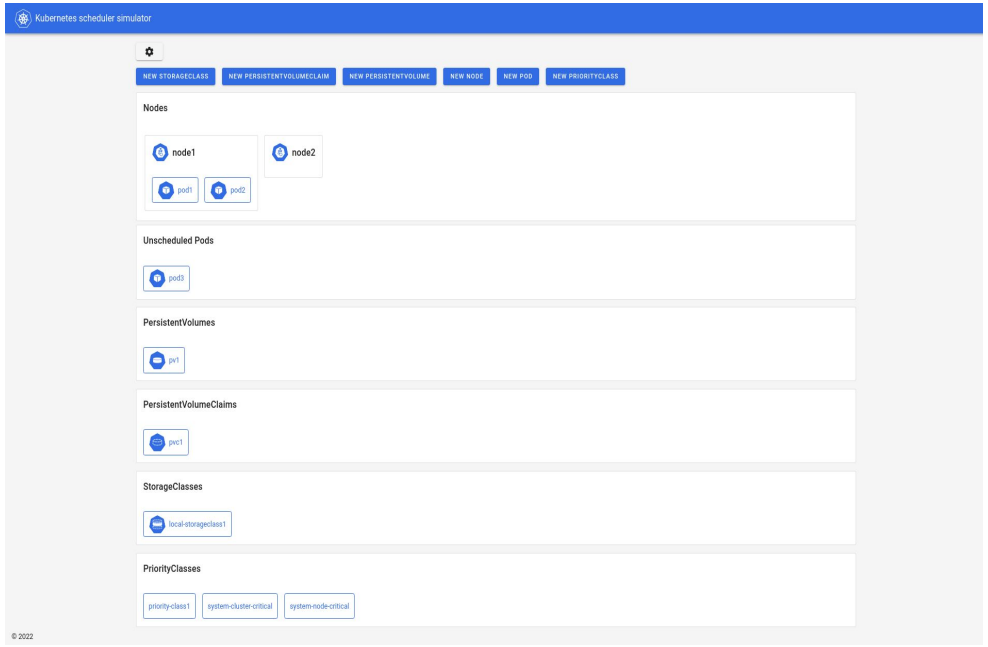
Updates

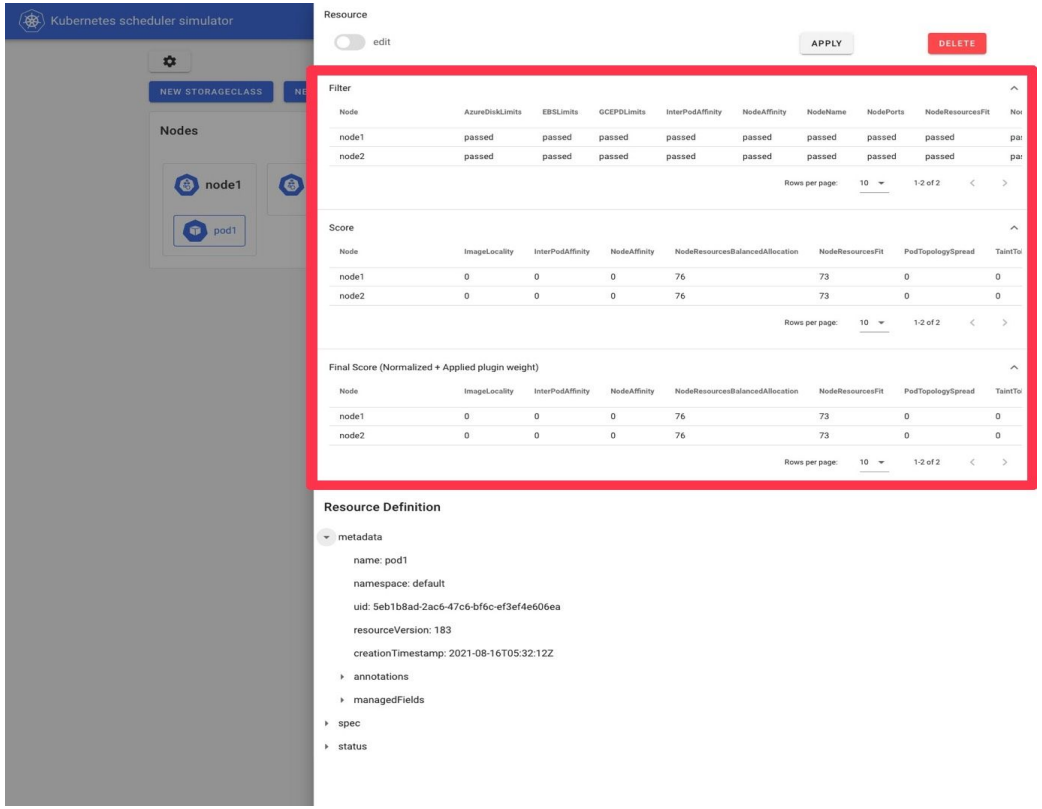
Main updates:

- 👍 Refactored all plugins based on Descheduler framework
- 💦 Supported descheduler profiles #1093
- 👋 Enable open telemetry tracing #951
- 💪 Add namespace filter to nodeutilization #967

kube-scheduler-simulator

👉 **The simulator for the Kubernetes scheduler, help you understand results of scheduling in detail easily.**





✨NEW: Kueue

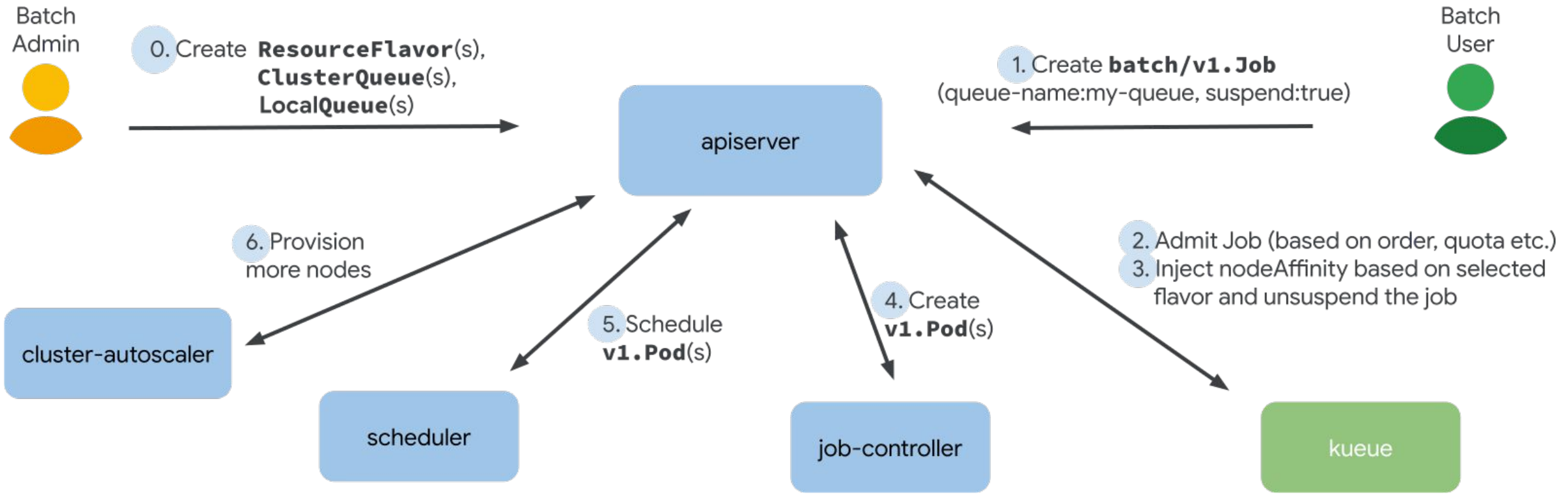
A Kubernetes-Native job queueing system, offering:

- Job management with queueing policies(FIFO, BestEffort, Preemption)
- Multi-Tenant support, but no hierarchical queue support (WIP) 🤦
- Resource quota management with fair-sharing semantics
- Resource fungibility in heterogeneous clusters
- Two-Stage admission(budget, node scaling, etc.)
- …

🎯 **Design principle**: compatibility and separation of concerns with standard k8s components: *kube-scheduler*, *kube-controller-manager*, *cluster-autoscaler*.

Opensourced at 2022.02
Latest Release: v0.4.1

How Kueue works



Integrations

Integrated
WIP

Kueue

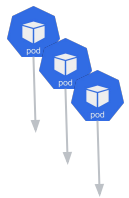
Job Framework (I/F)

BatchJob, JobSet, Pod

RayJob

Kubeflow

MPIJob, TFJob, PyTorchJob,
XGBoostJob, PaddleJob, MXJob, …



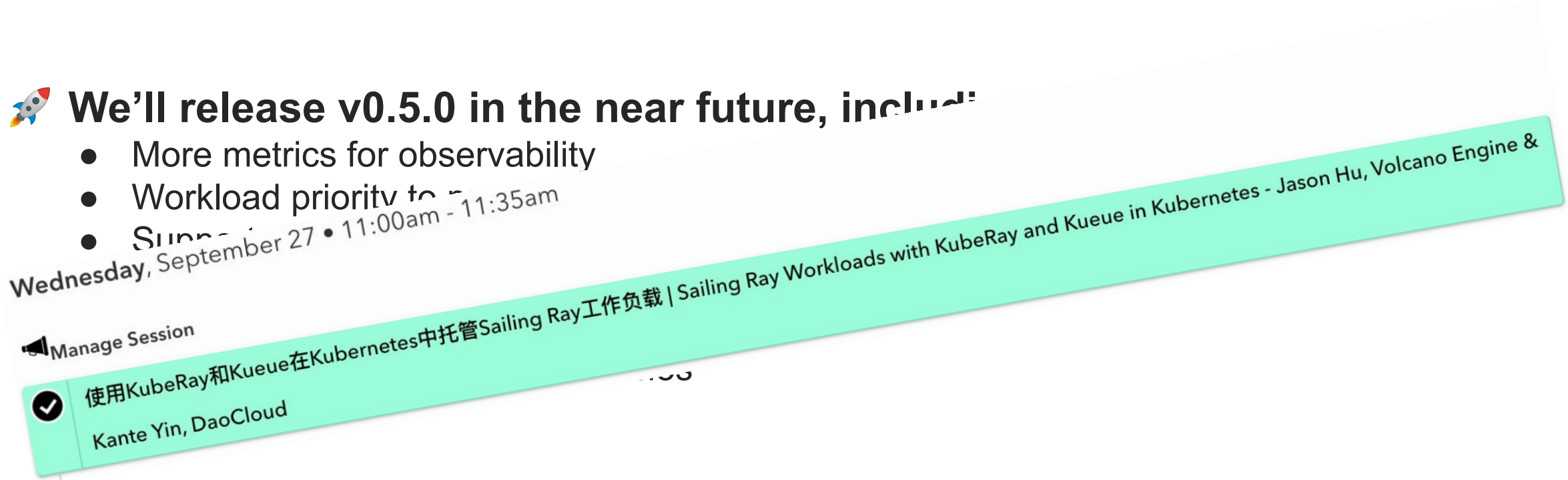
More, Flux MiniCluster etc., see *Kueue.sh*

What's Next

KubeCon CloudNativeCon OPEN SOURCE SUMMIT China 2023

🚀 **We'll release v0.5.0 in the near future, includi**

- More metrics for observability
- Workload priority to p
- Suppo



🧩 *We also have a adopter list for all the users, which helps us to better evolving Kueue project, if you're one of our users, please fill in the list. Thanks !* 😊

KWOK

KWOK is a toolkit that enables setting up a cluster of thousands of Nodes in seconds,
offering:

- **kwok**, helps to simulate the lifecycle of fake nodes, pods, and other Kubernetes API resources
- **kwokctl**, a cli tool to manage the clusters

KWOK is

- **Lightweight**: reliably to maintain 1k nodes and 100k pods easily
- **Fast**: almost 20 nodes/pods per second
- **Flexible**: you can simulate any pod or node status at your wish

Updates

M

-
-
-
- Perfor…
- A bunch of small features

Thursday, September 28 • 2:45pm - 3:20pm

深入研究：KWOK | Deep Dive: KWOK - Shiming Zhang, DaoCloud & Hao Liang, Tencent

🎮 We also see a lot of integrations with KOWK, if you're interest:

Q: How to extend the scheduler?

A: Scheduler Framework

requires recompilation

B: Multi Schedulers

resource consumption & scheduling conflict

C: Scheduler Extender

latency notable

D: Other options?

kube-scheduler-wasm-extension

Q: How to extend the scheduler?

A: Scheduler Framework

requires recompilation

B: Multi Schedulers

resource consumption & scheduling conflict

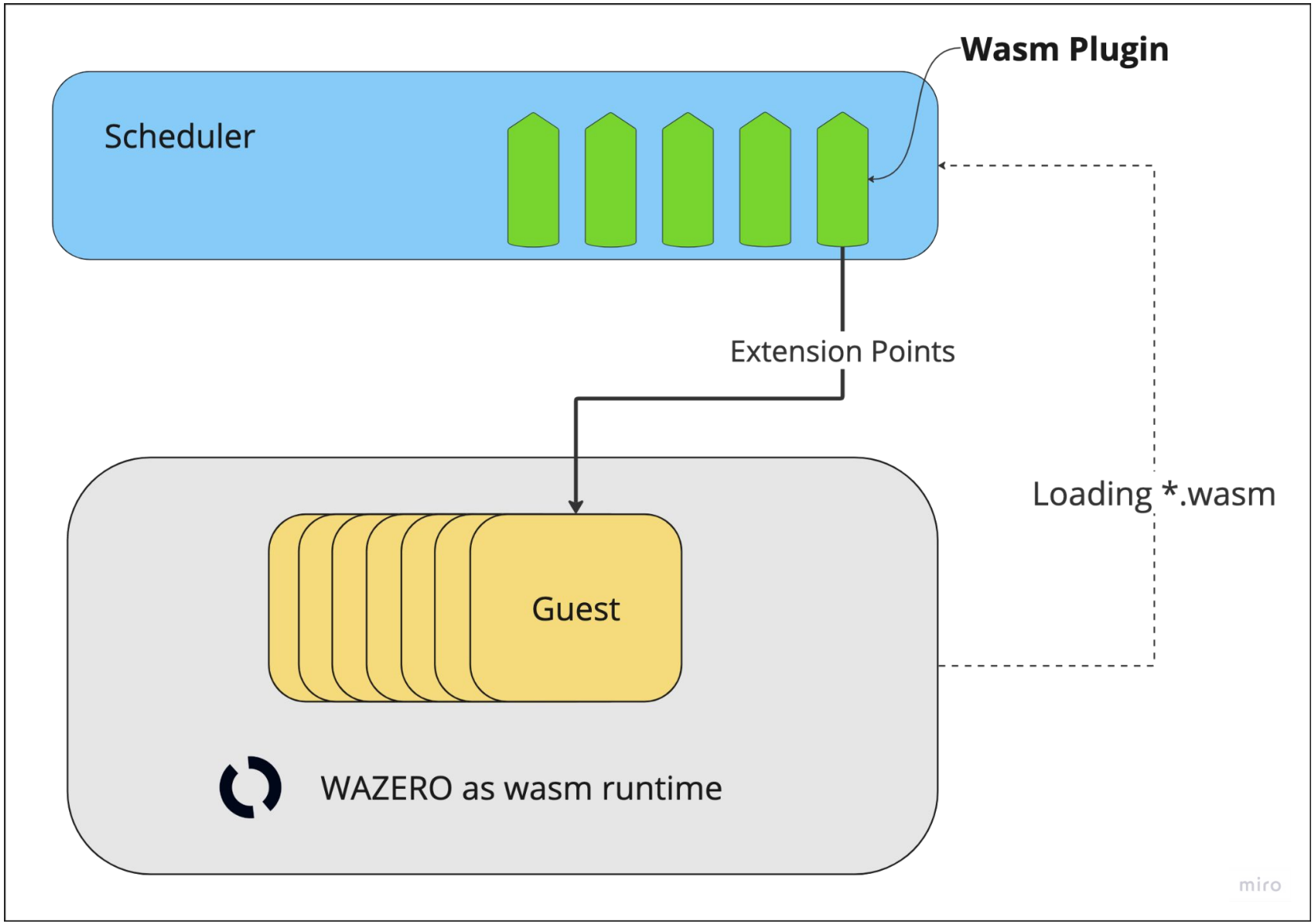
C: Scheduler Extender

latency notable

D: WebAssembly

i.e. Istio, Envoy, Dapr

How it works



Support:

- PreFilter
- Filter
- PreScore
- Score

Examples

What's Next

🥲 *Still under development, we need more volunteers!*

Roadmap:

- Support all kinds of extension potins
- Performance Improvement (2x slower than framework plugin, need more tests)
- Support other resources other than Pods, Nodes
- Other language examples

🎉 *Thanks to all these contributors which makes this happen*

Join us

- good-first-issue, help-wanted
- Slack #sig-scheduling
- Biweekly meeting (NA & Europe): Thursdays at 17:00 UTC
- Monthly meeting (APAC): First Thursday at 02:00 UTC
- KEPs, Devel Docs, Community

Come on

Q & A

Thanks!
Happy Mid-Autumn Festival!