
HELM
TiKV
KubeCon
CloudNativeCon
Europe 2023

KubeCon | CloudNativeCon

Europe 2023

SIG Scheduling Deep Dive

Aldo Culquicondor
@alculquicondor
Google

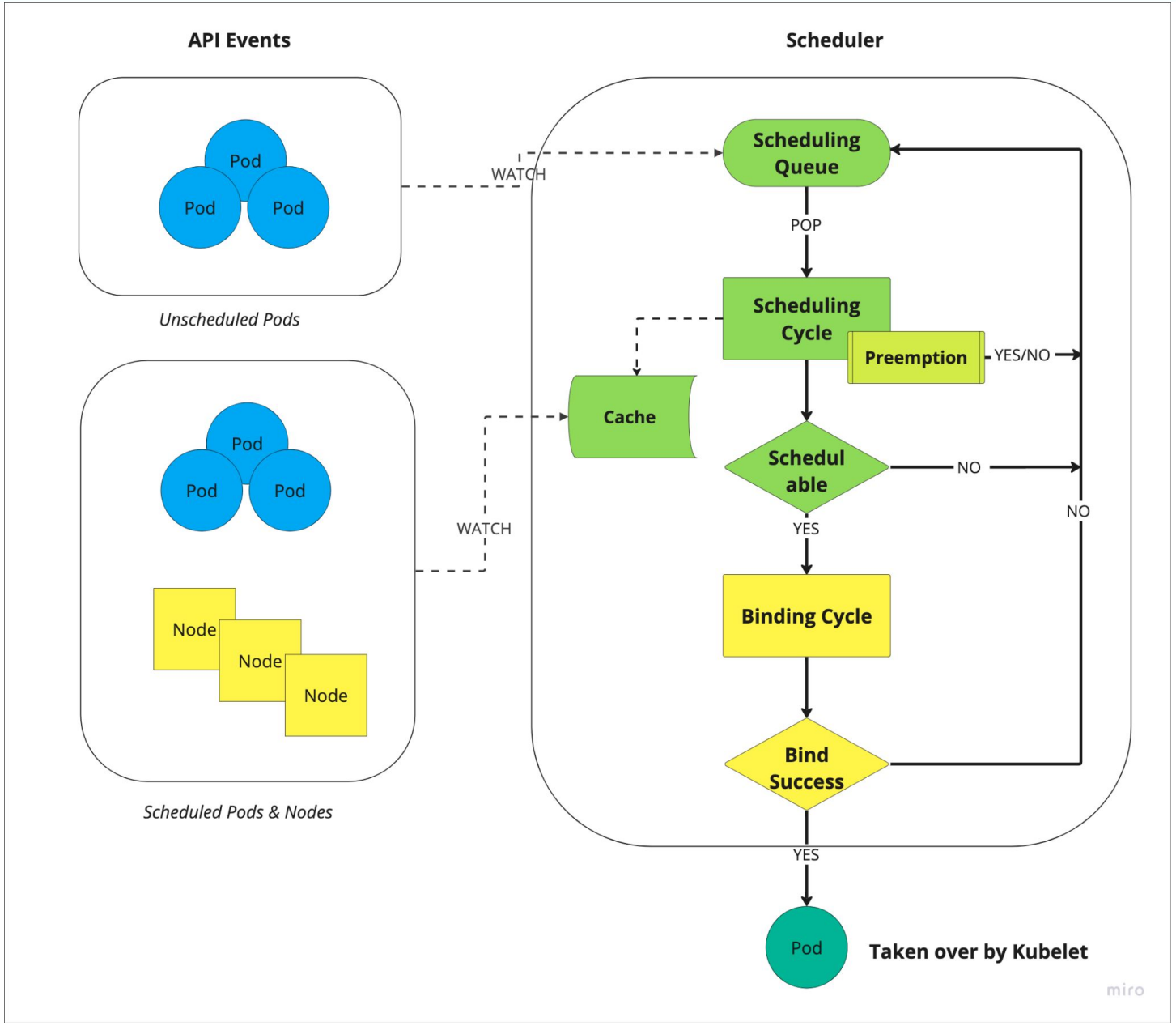
Kante Yin
@kerthcet
DaoCloud

Agenda

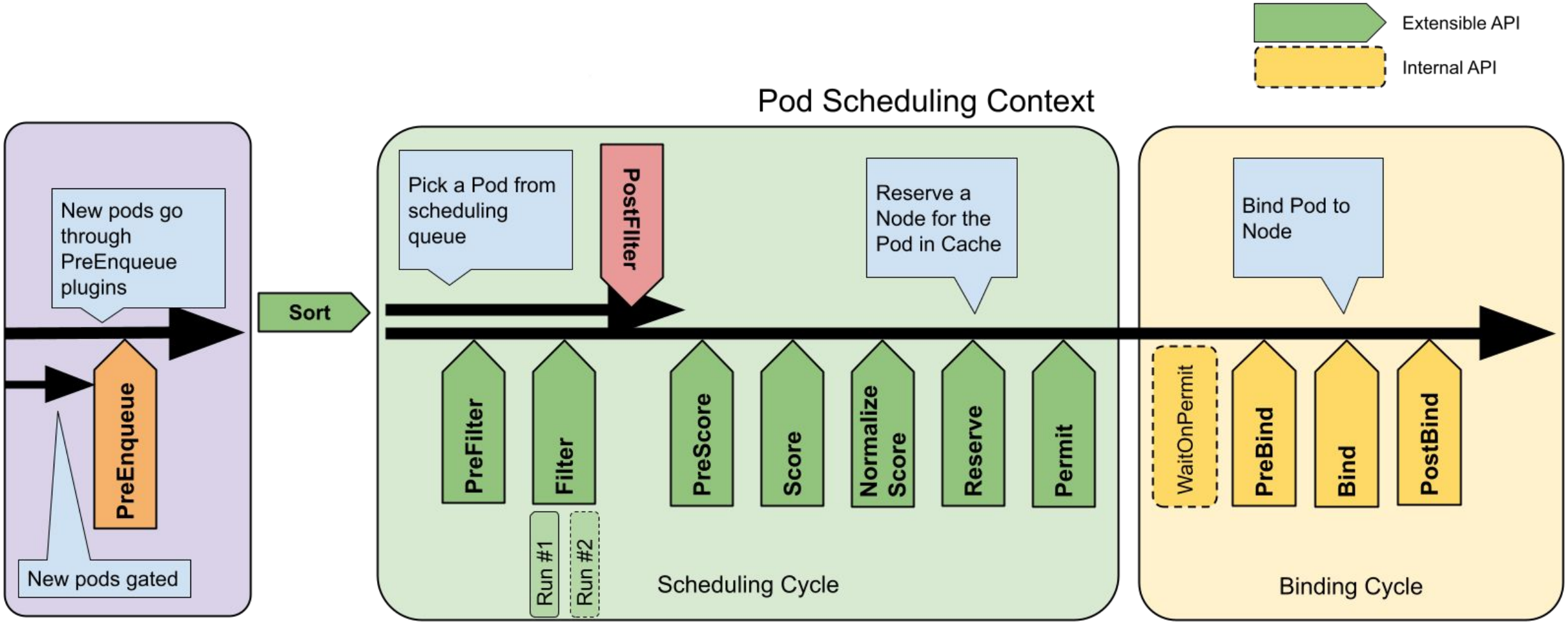
- Scheduler Overview
- Updates & Improvements
- Sub-project Updates
- Getting Involved
- Q & A

Scheduler Overview

Scheduler Flow

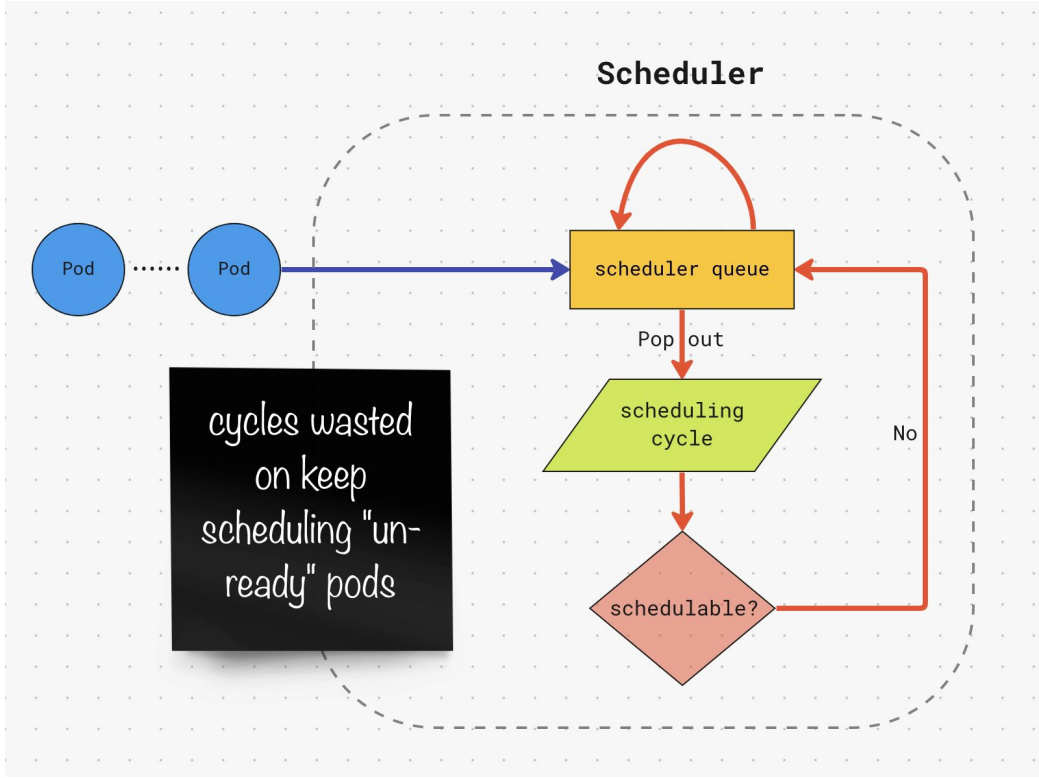


Scheduler Framework



Updates & Improvements

KEP-3521: Pod Scheduling Readiness (Beta)

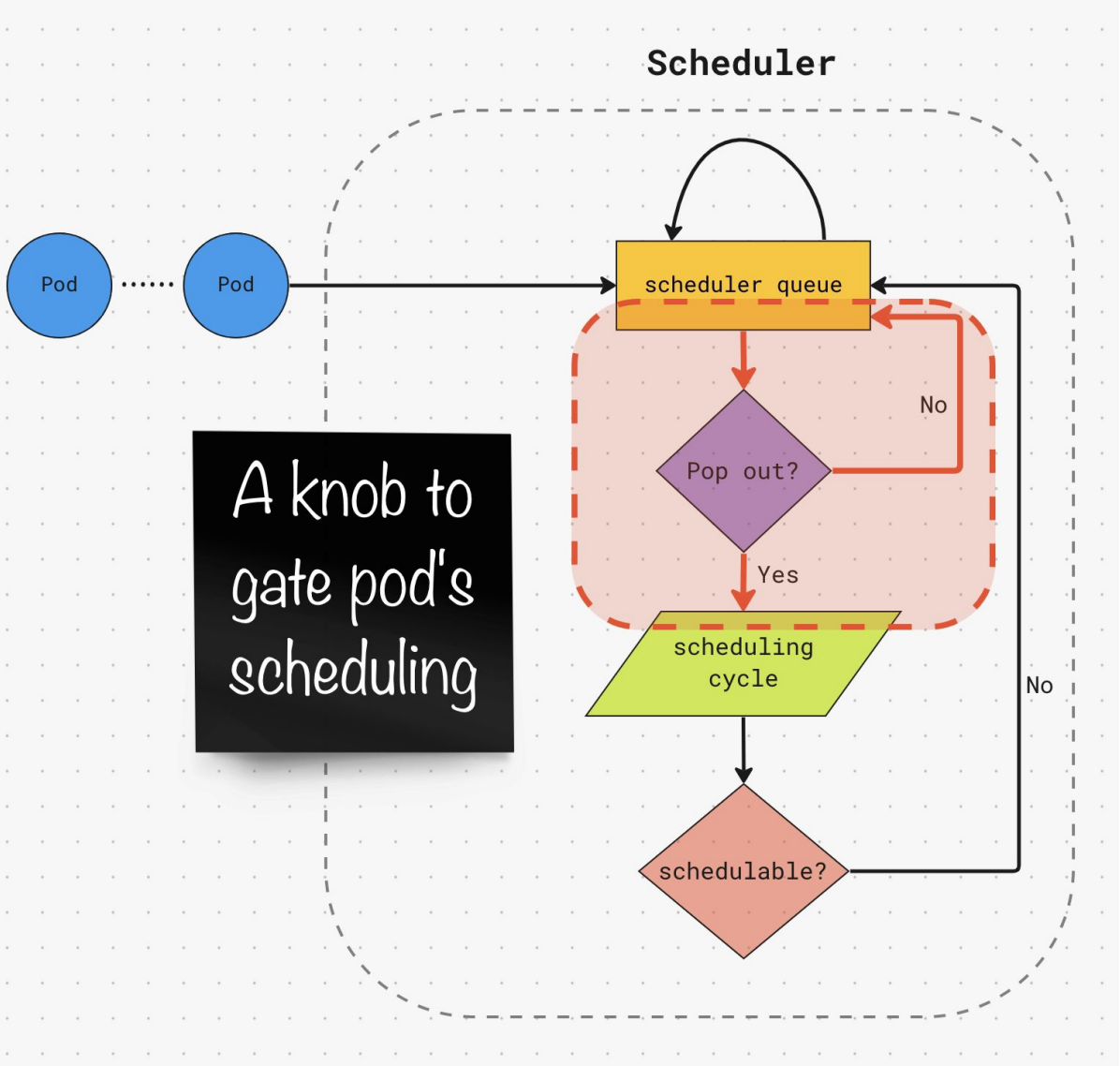


- **Problem:** Pods newly created may not be scheduling-ready:
  - Missing essential resources
  - Quota managed in queueing

  Which may lead to:
  - Degradation of scheduling throughput
  - Impact downstream components like Cluster Autoscaler

- **Demand:** Custom controllers also want to make scheduling decisions
  - No need to modify default scheduler!

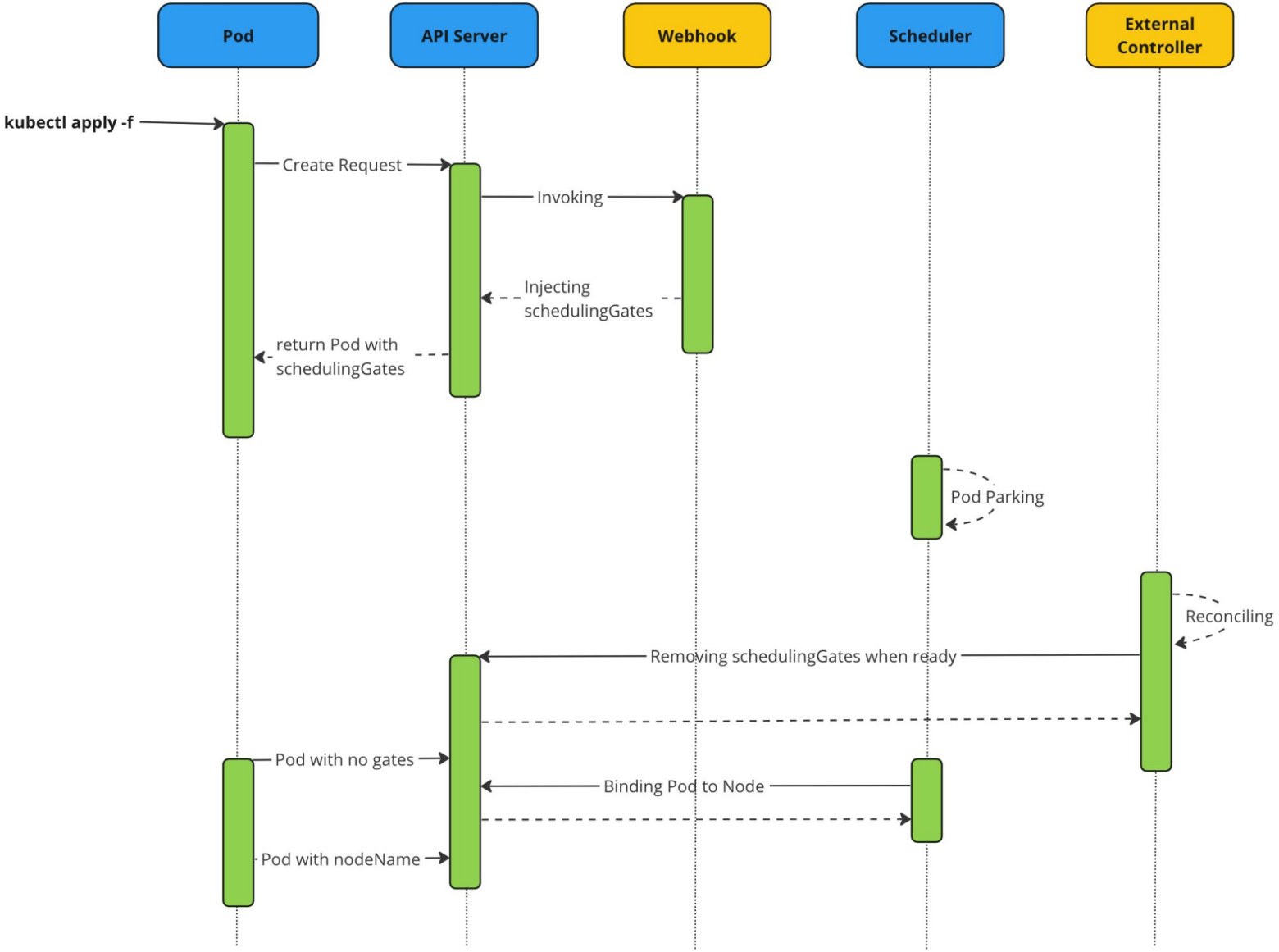
KEP-3521: Pod Scheduling Readiness (Beta)




```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
```

KEP-3521: Pod Scheduling Readiness (Beta)

Example:



KEP-2926: Mutable scheduling directives for suspended Jobs (GA)

KubeCon | CloudNativeCon Europe 2023

🤔 Motivation:

- *In batch jobs, usually pods will run with specific constraints: running in the same zone or same model of GPUs. Job hopes to be mutable when suspended.*
- *A high-level job queueing controller can take advantage of this for better pods placements.*

```yaml
apiVersion: batch/v1
kind: Job
metadata:
  name: pi
spec:
  template:
    suspend: true
    spec:
      containers:
      - name: pi
        image: perl:5.34.0
        command: ["perl",  "-Mbignum=bpi", "-wle", "print bpi(2000)"]
      restartPolicy: Never
      schedulingGates:
        - name: example.com/foo
        - name: example.com/bar
  backoffLimit: 4
```

Mutable Fields(Job's pod template):

- Node Affinity
- Node Selector
- Tolerations
- Annotations
- Labels
- 🆕 SchedulingGates

🎖️ *Enabled resource fungibility in Kueue*

NEW KEP-3838: Mutable Pod scheduling directives when gated (Beta)

KubeCon CloudNativeCon Europe 2023

🤔 **Motivation**: *influence the pod placement by external workload controllers*

```yaml
apiVersion: v1
kind: Pod
metadata:
  name: nginx
spec:
  schedulingGates:
    - name: foo
  containers:
  - name: nginx
    image: nginx:1.14.2
    ports:
    - containerPort: 80
  nodeSelector:
      …
  affinity:
    nodeAffinity:
        …
```

Requirements:

- NodeSelector
  - If empty, anything can be set
  - If not empty, only addition is allowed
- NodeAffinity
  - If empty, anything can be set
  - If not nil, only more restricted updates are allowed

🎖️ *Custom jobs with no suspend semantics can achieve similar behavior as K8s Job.*

KEPs: Fine-gained PodTopologySpread Policies

KubeCon | CloudNativeCon Europe 2023

✨ KEP-3022: Min domains in PodTopologySpread (Beta)

👉 Add MinDomains field to define the minimum number of topology domains. if not enough domains exist, getting the cluster-autoscaler on board.

✨ KEP-3094: Take taints/tolerations into consideration when calculating PodTopologySpread skew (Beta)

👉 Add nodeAffinityPolicy and nodeTaintPolicy fields for fine-gained node filters against node affinities and node taints in spreading skew calculation

```yaml
apiVersion: v1
kind: Pod
metadata:
  name: example-pod
spec:
  # Configure a topology spread constraint
  topologySpreadConstraints:
    - maxSkew: <integer>
      minDomains: <integer> # optional
      topologyKey: <string>
      whenUnsatisfiable: <string>
      labelSelector: <object>
      matchLabelKeys: <list> # optional
      nodeAffinityPolicy: [Honor|Ignore] # optional
      nodeTaintsPolicy: [Honor|Ignore] # optional
  ### other Pod fields go here
```

KEPs: Fine-gained PodTopologySpread Policies

KEP-3243: Respect PodTopologySpread after rolling upgrades(Beta)

👉 Add matchLabelKeys field to offer a solution for unbalanced scheduling in applications' rolling updates, such as Deployments

```yaml
apiVersion: apps/v1
kind: Deployment
metadata:
  name: example
spec:
  template:
    spec:
      topologySpreadConstraints:
        - maxSkew: 1
          topologyKey: kubernetes.io/hostname
          whenUnsatisfiable: DoNotSchedule
          labelSelector:
            matchLabels:
              app: foo
          matchLabelKeys:
            - pod-template-hash
```

Blog: Pod Topology Spread blog post (KEP-3022 KEP-3094 KEP-3243)

Other Notable Improvements ✨

- Won't run Filter if PreFilter returned a Skip status #114125
- Allow PreScore to return Skip status to skip running the corresponding Score extension #114827
- Add plugin_evaluation_total metric #115082

Sub-project Updates

Kueue

A Kubernetes-native job queueing system, offering:

- Resource quota management, with borrowing and preemption semantics.
- Resource fungibility in heterogeneous clusters.
- Support for k8s batch/v1.Job and kubeflow's MPIJob.
- Extension points and libraries for supporting custom job CRDs.
- More Job integrations coming soon

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

Kueue

Kueue

🎉 Release V0.3.0

(Kueue is adopting a 2-3 months release cadence from now on)

✨ Highlights:

- API is now beta, respecting k8s deprecation policy.
- Increased validation via webhooks.
- Preemption support
- Support for kubeflow MPIJob (v1beta2)
- [Optional] Sequential admission for quasi all-or-nothing
- Support for LimitRanges and Runtime Classes (pod overhead)
- Library for integrating custom job-like CRDs

Session: *Building a Batch System for the Cloud with Kueue*

Descheduler

A Post-Scheduling Eviction Component

🐝 A bunch of features, bugfixes, refactors:

- New in v0.27: **v1alpha2** + Descheduler Framework
  - Plugin-based refactor like Scheduler framework
  - New config API (v1alpha1 deprecated)
  - Descheduler Profiles
- Add namespace filter to nodeutilization #967

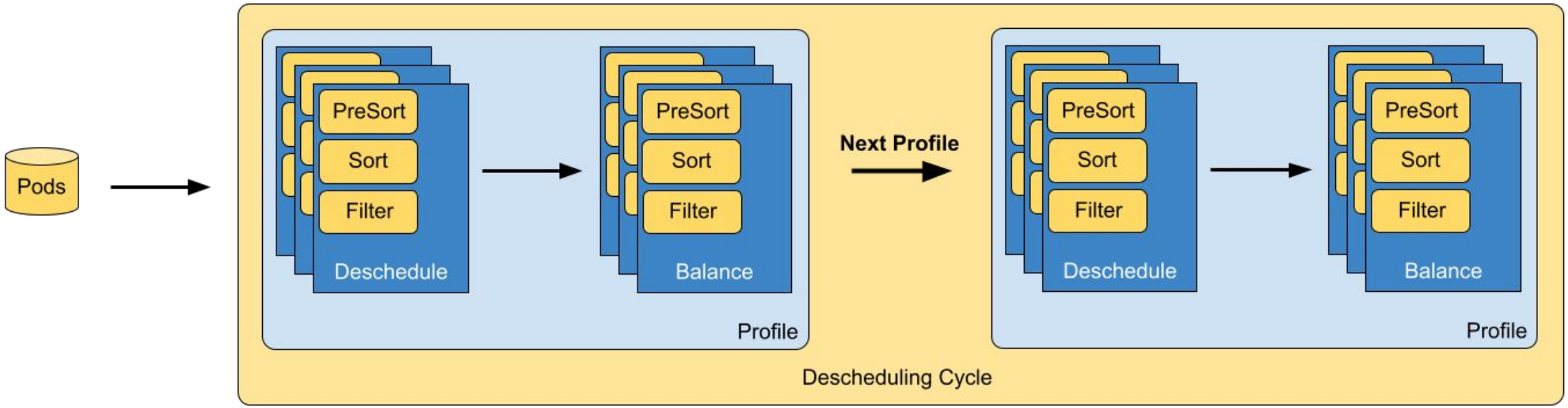
❤️ Inspiring things:

- 10+ new contributors in v0.26 🚀 🚀 🚀

DESCHEDULER

Descheduler Framework

Motivation: 💥 Multiple strategies



bit.ly/descheduler-fwk

📢 ***Calling for use-cases, please contact us!***

Scheduler Plugins

✨ **A repository hosting out-of-tree scheduler plugins:**

- Coscheduling
- Node Resource Topology
- Capacity Scheduling - elastic quota
- Preemption Toleration
- Trimaran - load-aware scheduling
- Network-Aware Scheduling

Scheduler Plugins

✨ *Major Changes:*

- Support scheduler component config v1 #466
- NodeResourceTopologyMatch
  - New ScoringStrategy LeastNUMANodes #454
  - Reservations to reduce conflicts with kubelet
- PodGroup and ElasticQuota now serve /status as a standard subresource #308

⚠️ *Breaking Changes*:

- Switch all CRDs API groups for PodGroup, ElasticQuota and Network-Aware plugins to `scheduling.x-k8s.io` #526 #528.

Kube-Scheduler-Simulator

Simulating Kubernetes scheduler without a real cluster. It displays scheduling decisions in detail.

✨ *Major Changes*:

- Display the scheduling results across all extension points (before we only support Filter/Score extension points)

🚀 *What's next?*

- KEP-140: The scenario-based simulation and benchmark

Kube-Scheduler-Simulator


```
  annotations:
    scheduler-simulator/bind-result: '{"DefaultBinder":"success"}'
    scheduler-simulator/filter-result: >-

{"node-2zc2x":{"AzureDiskLimits":"passed","EBSLimits":"passed","GCEPDLimits":"passed","InterPodAffinity":"passed","NodeAffinity":
"passed","NodeName":"passed","NodePorts":"passed","NodeResourcesFit":"passed","NodeUnschedulable":"passed","NodeVolumeLimits":"pa
ssed","PodTopologySpread":"passed","TaintToleration":"passed","VolumeBinding":"passed","VolumeRestrictions":"passed","VolumeZone"
:"passed"}}
    scheduler-simulator/finalscore-result: >-

{"node-2zc2x":{"ImageLocality":"0","InterPodAffinity":"0","NodeAffinity":"0","NodeResourcesBalancedAllocation":"76","NodeResource
sFit":"73","PodTopologySpread":"200","TaintToleration":"100"},"node-m7jqj":{"ImageLocality":"0","InterPodAffinity":"0","NodeAffin
ity":"0","NodeResourcesBalancedAllocation":"76","NodeResourcesFit":"73","PodTopologySpread":"200","TaintToleration":"100"}}
    scheduler-simulator/permit-result: '{}'
    scheduler-simulator/permit-result-timeout: '{}'
    scheduler-simulator/postfilter-result: '{}'
    scheduler-simulator/prebind-result: '{"VolumeBinding":"success"}'
    scheduler-simulator/prefilter-result: '{}'
    scheduler-simulator/prefilter-result-status: >-

{"InterPodAffinity":"success","NodeAffinity":"success","NodePorts":"success","NodeResourcesFit":"success","PodTopologySpread":"su
ccess","VolumeBinding":"success","VolumeRestrictions":"success"}
    scheduler-simulator/prescore-result: >-
      {"InterPodAffinity":"success","NodeAffinity":"success","PodTopologySpread":"success","TaintToleration":"success"}
    scheduler-simulator/reserve-result: '{"VolumeBinding":"success"}'
    scheduler-simulator/score-result: >-

{"node-2zc2x":{"ImageLocality":"0","InterPodAffinity":"0","NodeAffinity":"0","NodeResourcesBalancedAllocation":"76","NodeResource
sFit":"73","PodTopologySpread":"0","TaintToleration":"0"},"node-m7jqj":{"ImageLocality":"0","InterPodAffinity":"0","NodeAffinity"
:"0","NodeResourcesBalancedAllocation":"76","NodeResourcesFit":"73","PodTopologySpread":"0","TaintToleration":"0"}}
     scheduler-simulator/selected-node: node-2zc2x
```

NEW Kwok

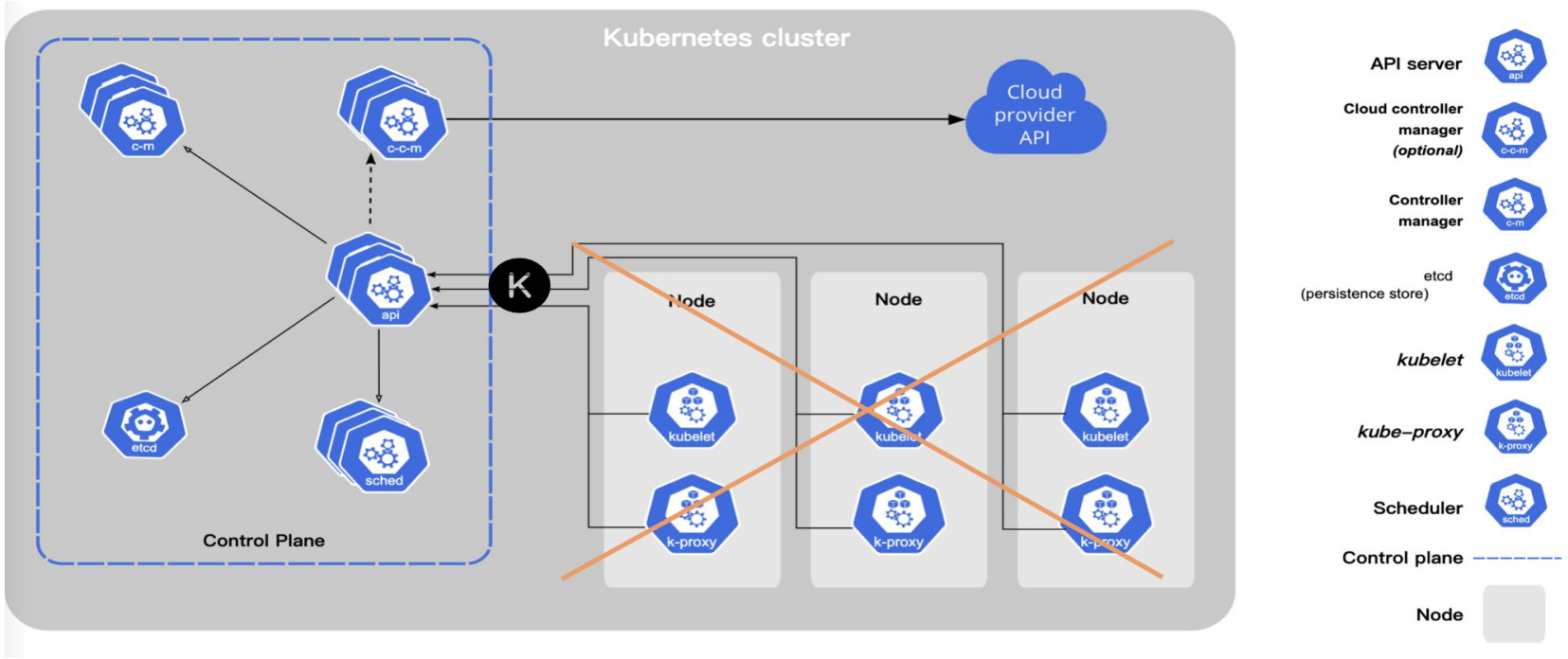
KWOK (Kubernetes without kubelet) is a toolkit that enables setting up a cluster of thousands of Nodes in seconds for control-plane scalability simulations.

It's *Lightweight*, *Fast*, *Flexible*. See blog here.

🎉 *We just released the first version v0.1.1.*

🌟 *1000 Github stars!*

Kwok - Architecture



- Kwok Controllers:
  - Node Controller - It is responsible for simulating the node's lifecycle, also updating the heartbeat of the node.
  - Pod Controller - It is responsible for simulating the pod's lifecycle.
- Kwokctl, as friendly as Kind.

- Use Cases:
  - Scheduling simulation (kube-scheduler-simulator + kwok #251)
  - Scalability tests for controller plane
  - Integration with ClusterAutoscaler, ClusterAPI…
  - Functionality tests
  - …

Getting Involved

- 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

🎉 ***Thanks for all the contributors and we're calling out reviewers!***

Q & A

Please scan the QR Code above
to leave feedback on this session

KubeCon | CloudNativeCon

Europe 2023

Thanks!