阿里云 | 奥运会全球指定云服务商

© Copyright by Alibaba Cloud All rights reserved

阿里巴巴如何扩展Kubernetes调度器支持 AI 和大数据作业？

背景

阿里云 | 奥运会全球指定云服务商

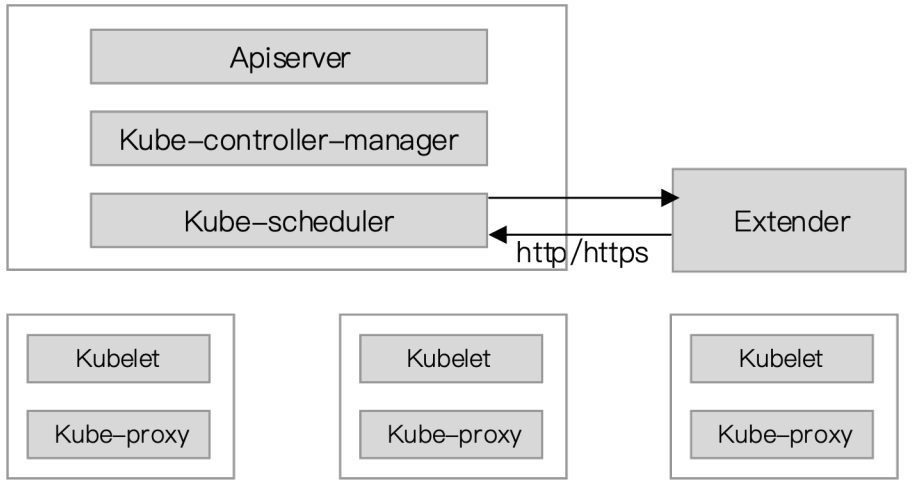
Kubernetes目前已经广泛的应用于在线服务编排，为了提升集群的的利用率和运行效率，我们希望将Kubernetes作为一个统一的管理平台来管理在线服务和离线作业。但是目前Kubernetes原生的调度器缺少例如Coscheduling、Capacity Scheduling等对于离线作业非常重要的功能，所以无法有效支持离线作业迁移到Kubernetes平台。

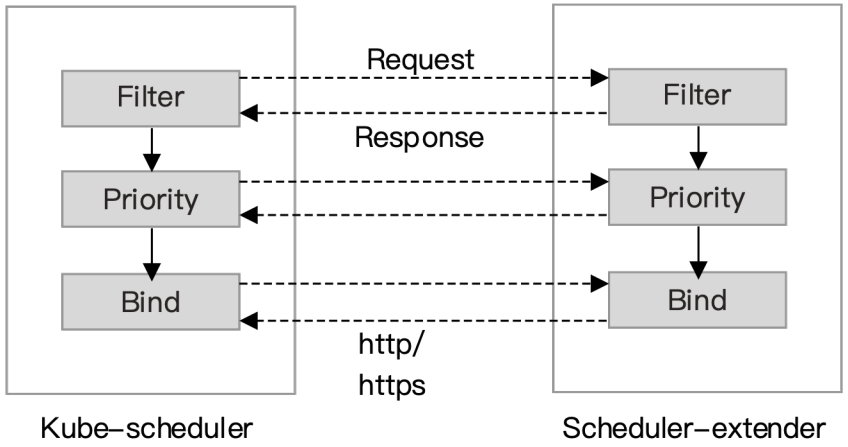
由于Scheduler Framework v2的出现，Kube-scheduler开发出多个扩展点方便开发者定制自己的调度器插件，给我们带来了将离线作业所需的重要功能集成到Kube-scheduler的可能。

早期Scheduler扩展方案

阿里云 | 奥运会全球指定云服务商

■Scheduler Extender





◆缺点：

- ✓ 调用为http请求，受到网络环境的影响，性能远低于本地的函数调用。
- ✓ 可以扩展的点比较有限，位置比较固定，无法支持灵活的扩展

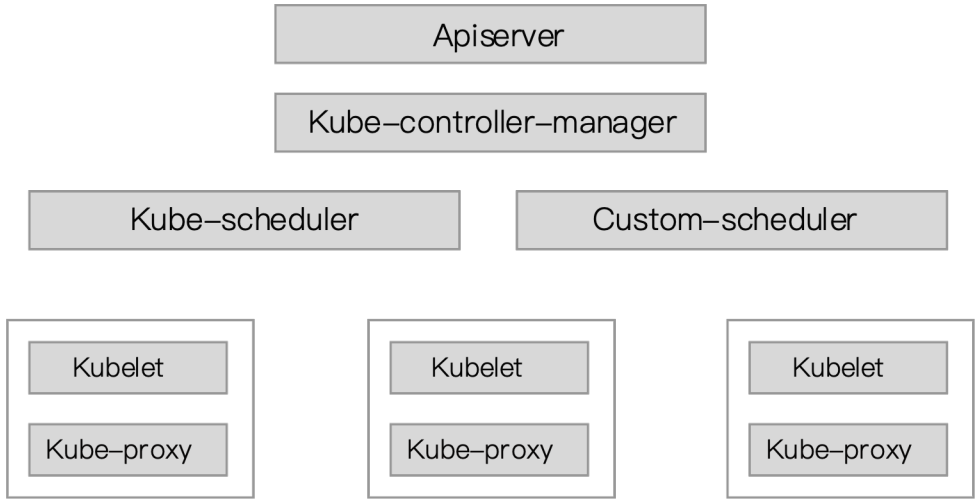
◆优点：

- ✓ 对Scheduler代码无侵入性，维护成本较低
- ✓ 灵活升级部署

早期Scheduler扩展方案

阿里云 | 奥运会全球指定云服务商

■Multiple schedulers



◆缺点：

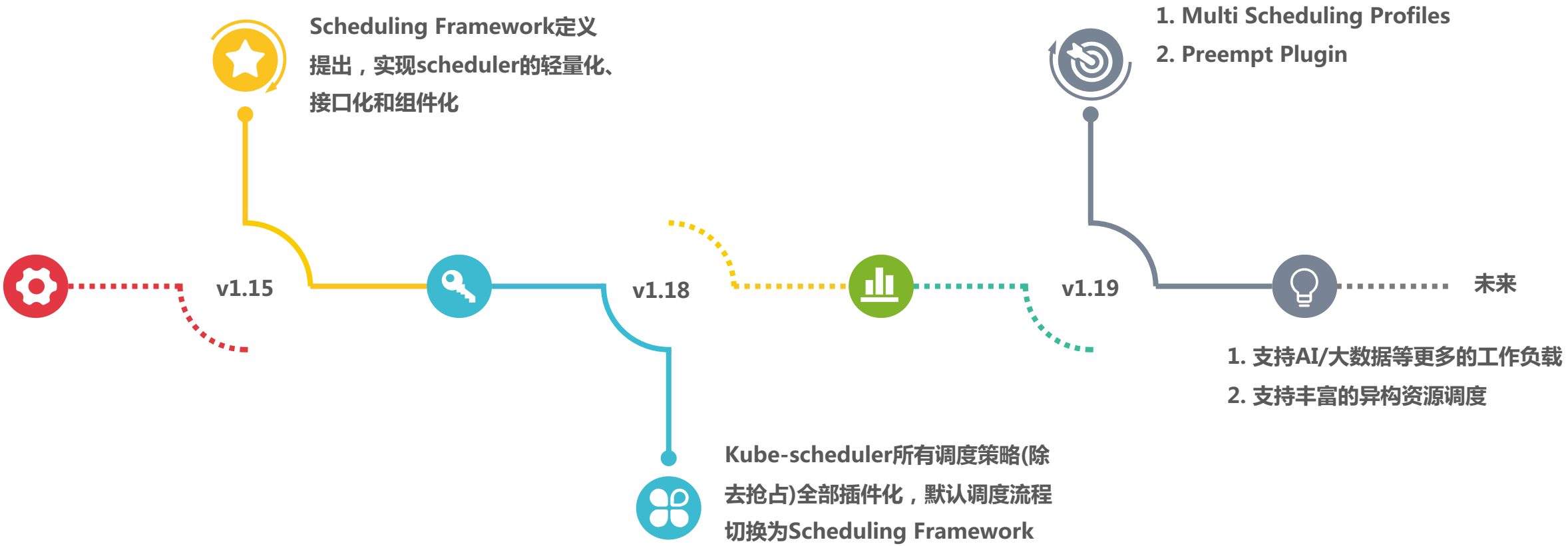
- ✓ 与default scheduler同时部署时会导致资源冲突(通过label划分资源池又会导致资源利用率下降)
- ✓ 研发和维护成本较高，版本兼容性风险

◆优点：

- ✓ 调度性能相比Extender更好
- ✓ Scheduler可扩展性强

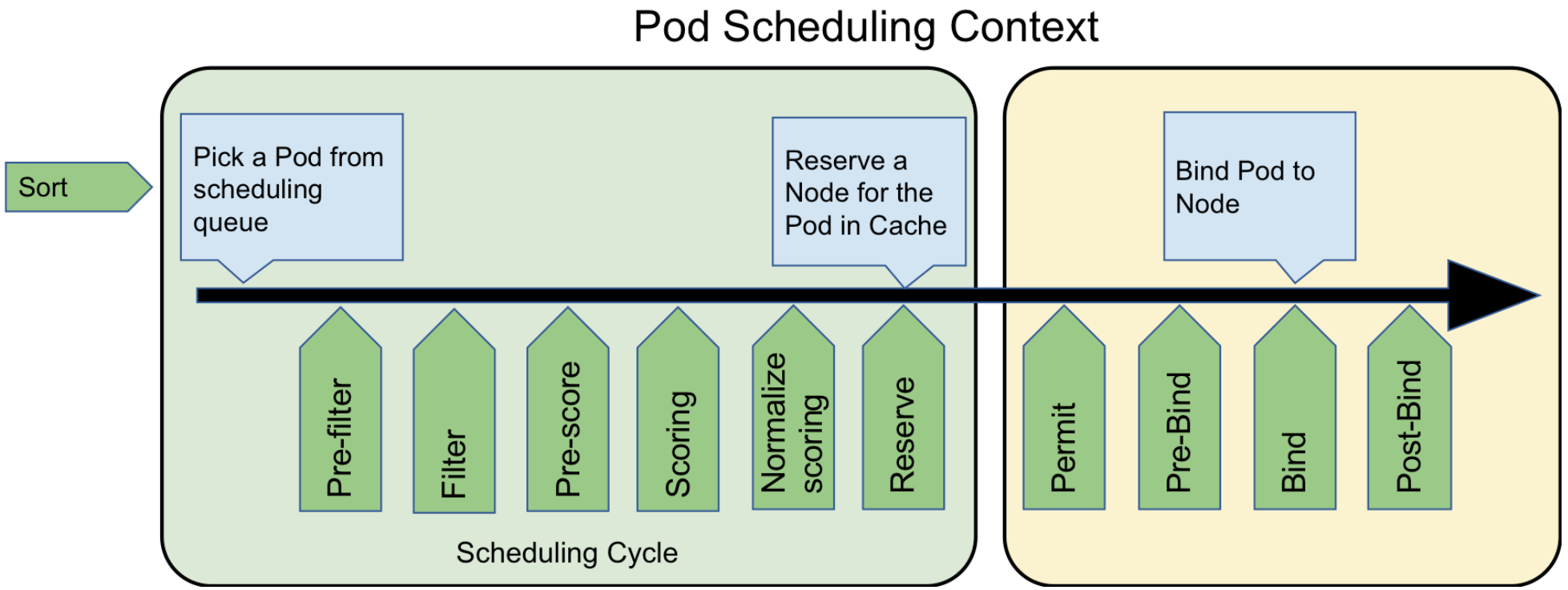
Scheduling Framework

阿里云 | 奥运会全球指定云服务商



Scheduling Framework

阿里云 | 奥运会全球指定云服务商



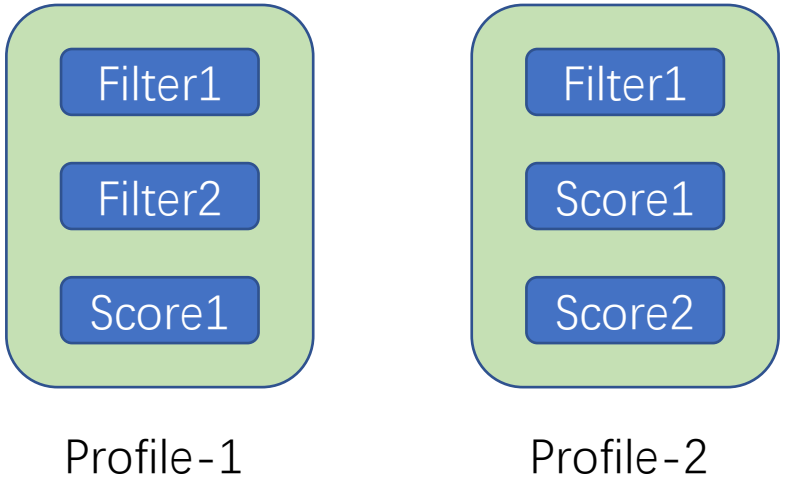
Multi Scheduling Profiles

阿里云 | 奥运会全球指定云服务商

■Goals：

✓ 使得kube-scheduler可以满足不同Workloads对于调度策略差异化的支持

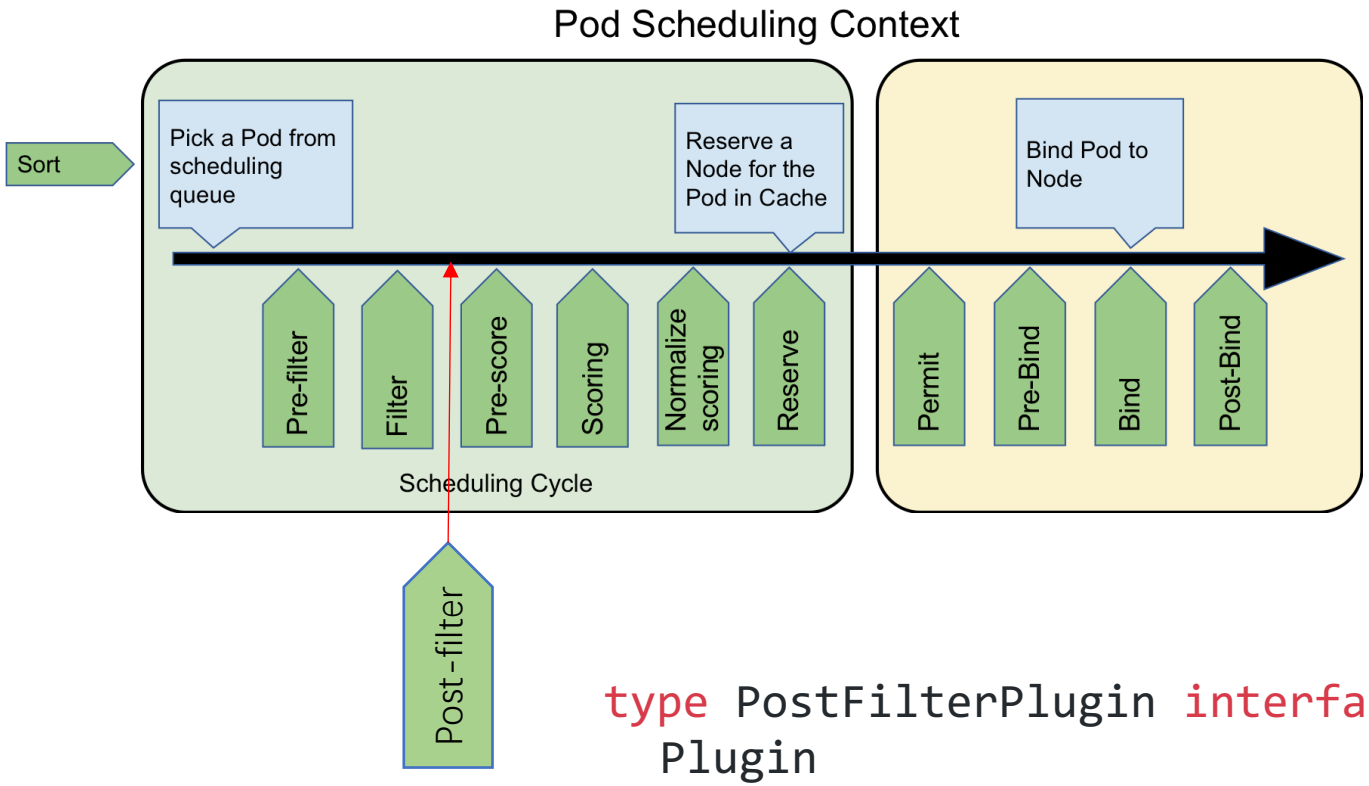
```
type KubeSchedulerProfile struct {
    SchedulerName string
    Plugins *Plugins
    PluginConfig []PluginConfig
}
```



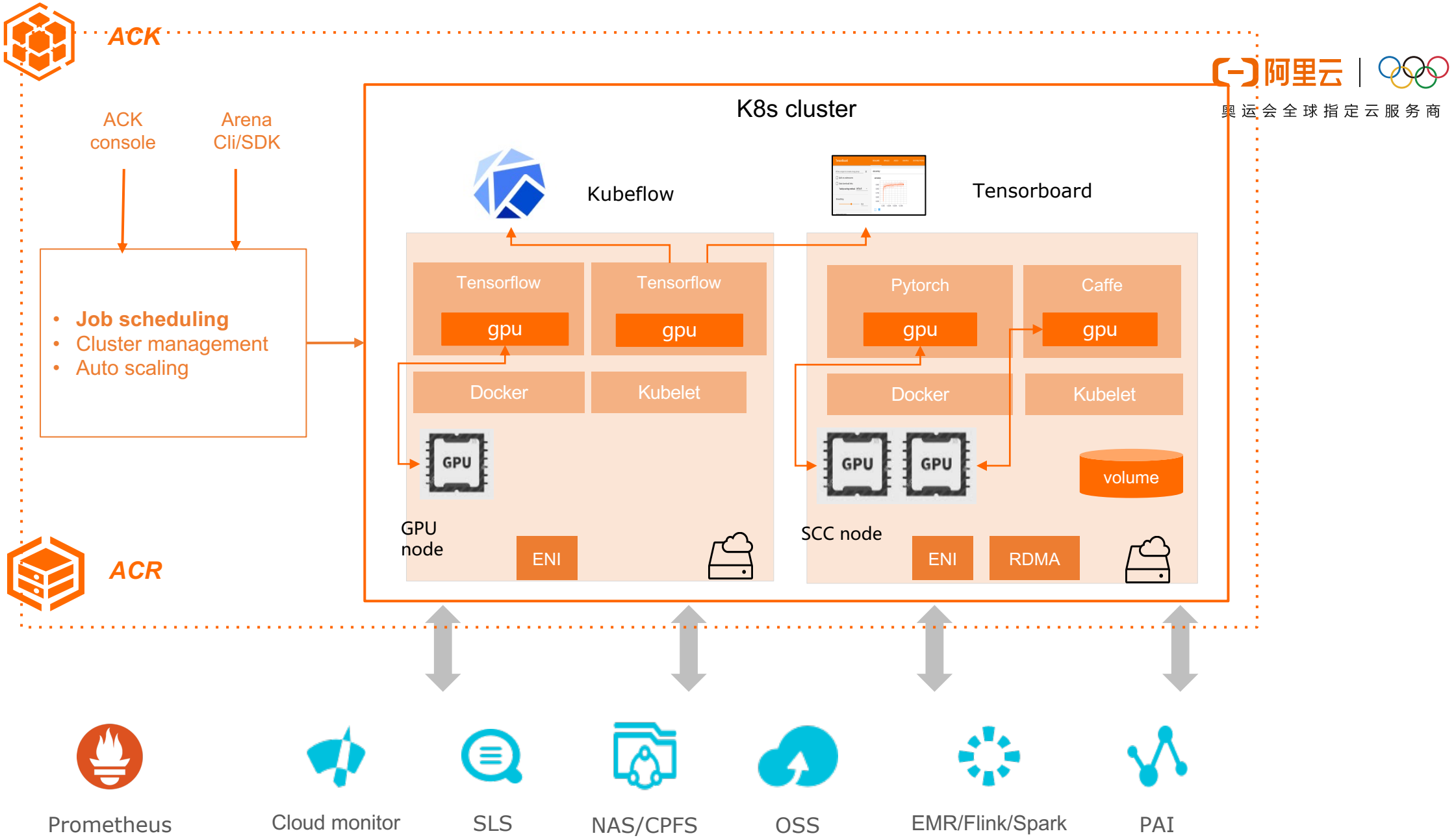

```
type PodSpec struct {
    SchedulerName string
}
```

Preempt Plugin

阿里云 | 奥运会全球指定云服务商




```
type PostFilterPlugin interface {
  Plugin
  PostFilter(ctx context.Context, state *CycleState, pod *v1.Pod,
  filteredNodeStatusMap NodeToStatusMap) (*PostFilterResult, *Status)
}
```


ACK
阿里云
奥运会全球指定云服务商
ACK console
Arena Cli/SDK
Job scheduling
Cluster management
Auto scaling
K8s cluster
Kubeflow
Tensorboard
Tensorflow
gpu
Tensorflow
gpu
Docker
Kubelet
GPU
GPU node
ENI
Pytorch
gpu
Caffe
gpu
Docker
Kubelet
GPU
GPU
volume
SCC node
ENI
RDMA
ACR
Prometheus
Cloud monitor
SLS
NAS/CPFS
OSS
EMR/Flink/Spark
PAI

AI/Batch作业级调度 Coscheduling

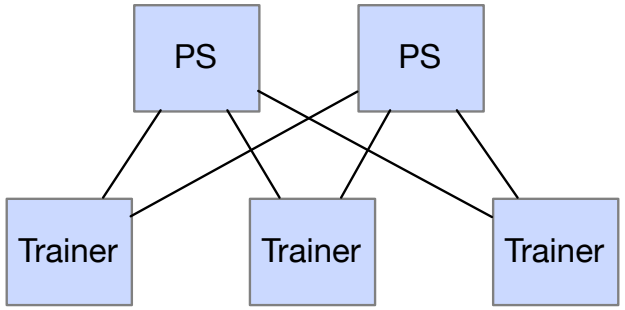
阿里云 | 奥运会全球指定云服务商

■需求：

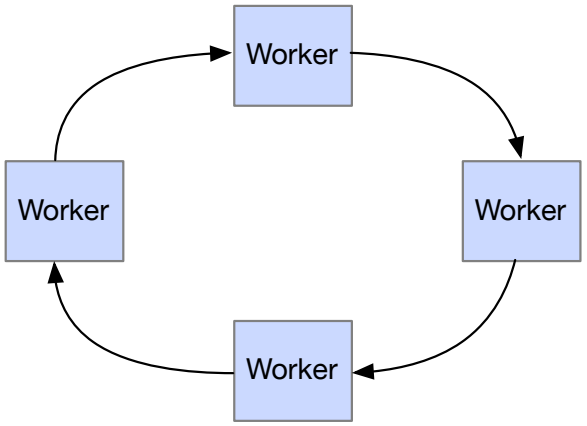
- ✓ 对于Tensorflow和MPI的作业，同一个Job下要求所有Pod同时启动才能运行
- ✓ 对于Spark的作业，需要至少保证启动Driver和Excuter作业满足最小数目才能运行

■痛点：

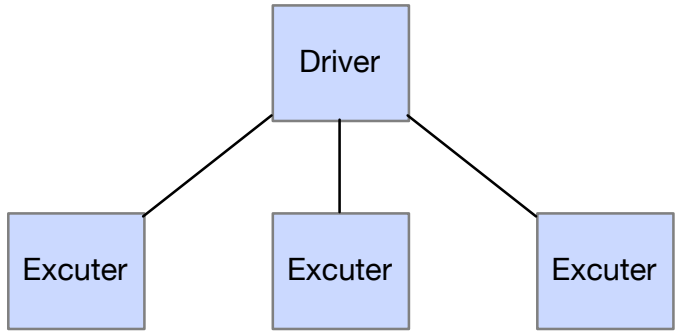
- ✓ Kubernetes原生的调度器是以Pod为单位依次调度的，不会在调度过程中考虑Pod之间的关系
- ✓ 当集群资源无法满足Job所有资源请求时，部分Pod无法启动，已经创建的Pod又无法运行，死锁导致资源的浪费



Tensorflow



MPI



Spark

AI/Batch作业级调度 Coscheduling

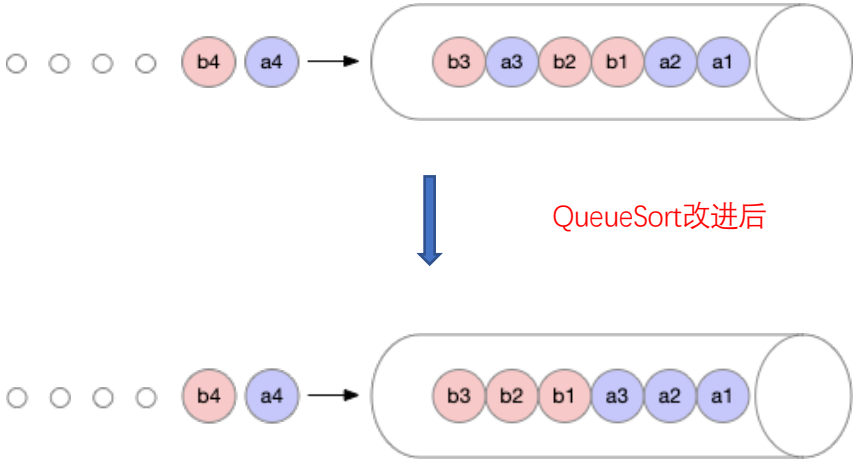
阿里云 | 奥运会全球指定云服务商

■解决方案：

- ✓ 实现QueueSort插件, 保证在队列中属于同一个PodGroup的Pod能够排列在一起
- ✓ Permit的延迟绑定的功能，对于不满足PodGruop Min限制的Pod进行等待，等待积累的Pod数目满足足够的数目时，再将同一个PodGruop的所有Pod创建

```
labels:
    pod-group.scheduling.sigs.k8s.io/name: nginx
    pod-group.scheduling.sigs.k8s.io/min-available: "2"
```

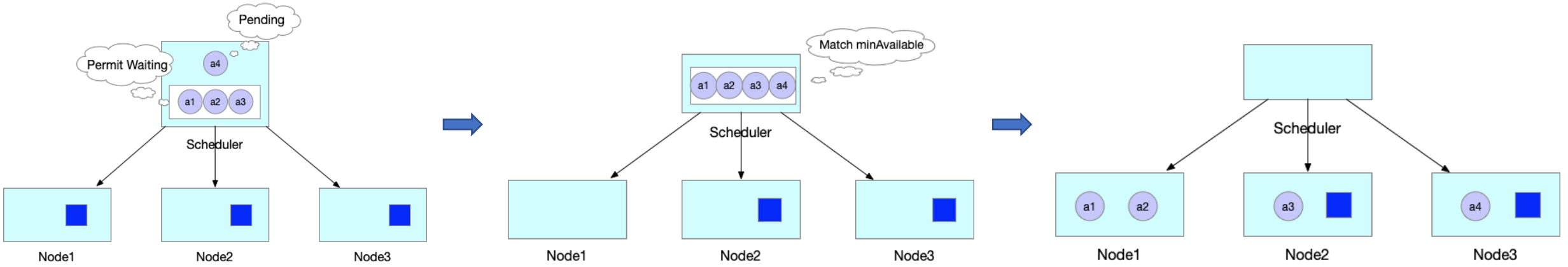
- name标识podGroup的name
- min-available是用来标识该PodGroup的作业能够正式运行时所需要的最小副本数



AI/Batch作业级调度 Coscheduling

阿里云 | 奥运会全球指定云服务商

■解决方案：



资源不足，无法满足min则等待

资源足够时，判断是否满足min

满足min，pod被成功创建

多用户下AI/Batch作业级调度 Capacity Scheduling

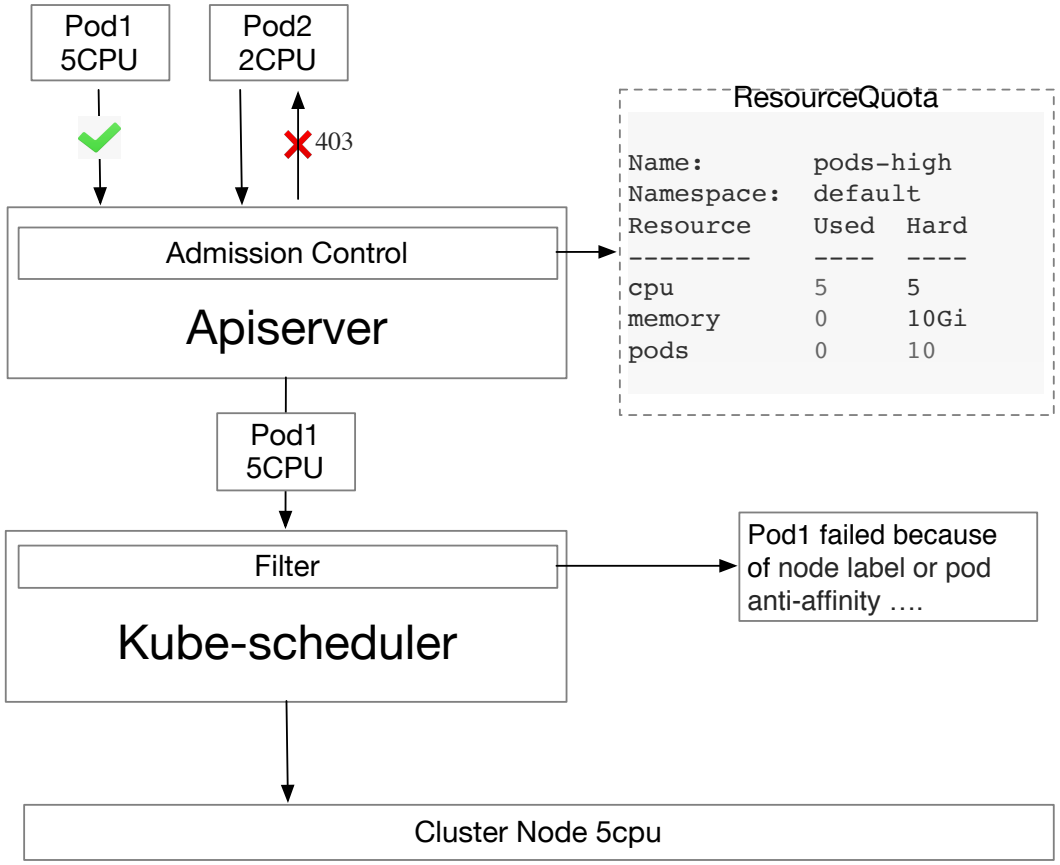
阿里云 | 奥运会全球指定云服务商

■需求：

- ✓ 多个用户同时使用同一个Kubernetes集群时，需要配额管理，保障每个用户的资源请求
- ✓ 为了提升集群的资源利用率，当某用户有空闲资源时可被其他用户借用

■痛点：

- ✓ Kubernetes支持多租户资源管理是通过ResourceQuota + Namespace 实现的，通过Admission Control来进行Quota限制，导致集群的资源利用率降低
- ✓ ResourceQuota为硬隔离，无法支持资源的共享



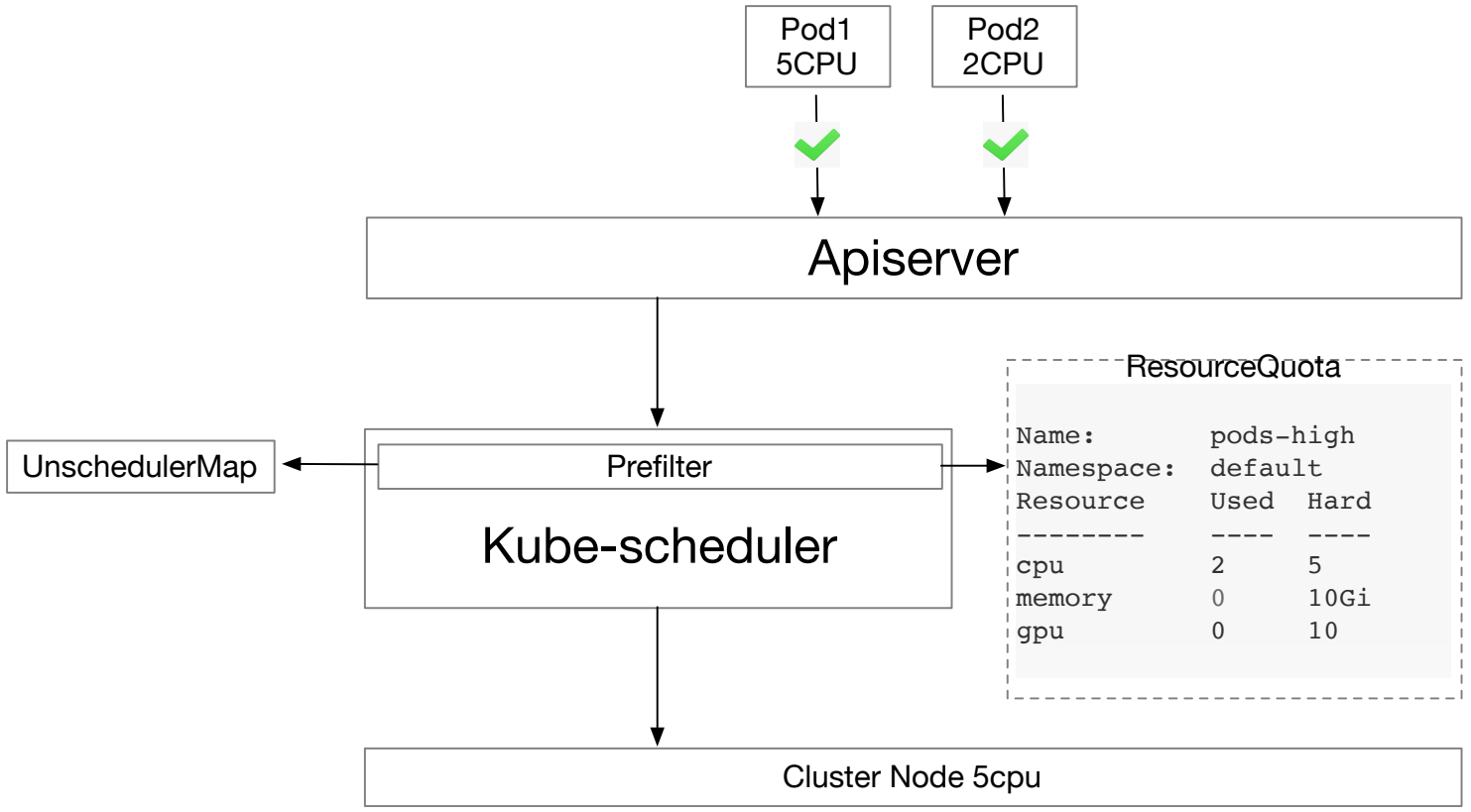
多用户下AI/Batch作业级调度 Capacity Scheduling

阿里云 | 奥运会全球指定云服务商

■解决方案：

1. 将Quota的限制判断从Admission Control迁移到Kube-scheduler中，

在调度过程中进行Quota判断



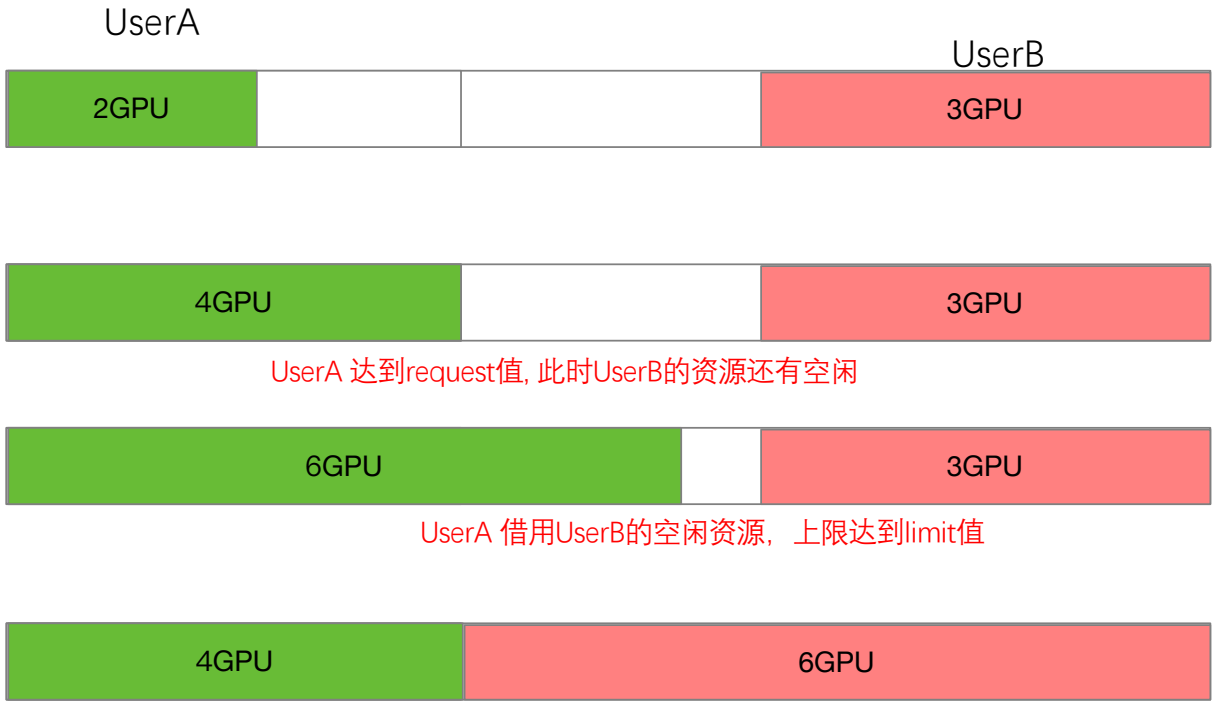
多用户下AI/Batch作业级调度 Capacity Scheduling

阿里云 | 奥运会全球指定云服务商

■解决方案：

2. 实现弹性的资源配额管理

- Request：资源紧张时该用户保证分配的资源量
- Limit：该用户所能使用资源的最大值




```yaml
apiVersion: queue.alibabacloud.com/v1alpha1
kind: Queue
metadata:
  name: user1
  namespace: user1
spec:
  resources:
    requesets:
      gpu: 4
    limits:
      gpu: 6
```

UserA

```yaml
apiVersion: queue.alibabacloud.com/v1alpha1
kind: Queue
metadata:
  name: user2
  namespace: user2
spec:
  resources:
    requesets:
      gpu: 6
    limits:
      gpu: 8
```

UserB

AI/Batch作业级调度 Binpack

阿里云 | 奥运会全球指定云服务商

■问题：

- ✓ Kubernetes默认开启的资源调度策略是Spread的策略，资源尽量打散，但是会导致较多的资源碎片，使得整体资源利用率下降。
- ✓ Kubernetes提供的MostRequested策略，仅支持CPU+Mem的形式进行Binpack，无法支持Extend Resource, 例如GPU

■解决方案：

- ✓ 通过RequestedToCapacityRatioPriority配置支持GPU的权重
- ✓ 在打分阶段计算对应资源的利用率，通过利用率进行排序，优先打满一个节点后再向后调度

如何构建自己的Scheduling Framework 插件

阿里云 | 奥运会全球指定云服务商

Kubernetes 负责 Kube-scheduler 的小组 sig-scheduling 为了更好的管理调度相关的 Plugin，新建了项目 scheduler-plugins 来方便用户管理不同的插件，用户可以直接基于这个项目来定义自己的插件。

https://github.com/kubernetes-sigs/scheduler-plugins

如何构建自己的Scheduling Framework 插件

阿里云 | 奥运会全球指定云服务商

进击的 Kubernetes 调度系统系列文章

(一) Scheduling Framework
https://www.infoq.cn/article/lYUw79lJH9bZv7HrgGH5

(二) 支持批任务的 Coscheduling/Gang scheduling
https://www.infoq.cn/article/Q1l845yOl2GAF8WIVtCW

其他内容敬请期待

阿里云 | 奥运会全球指定云服务商

Q&A