75Kubernetes 系列（六十八）Scheduler 调度器深度挖掘

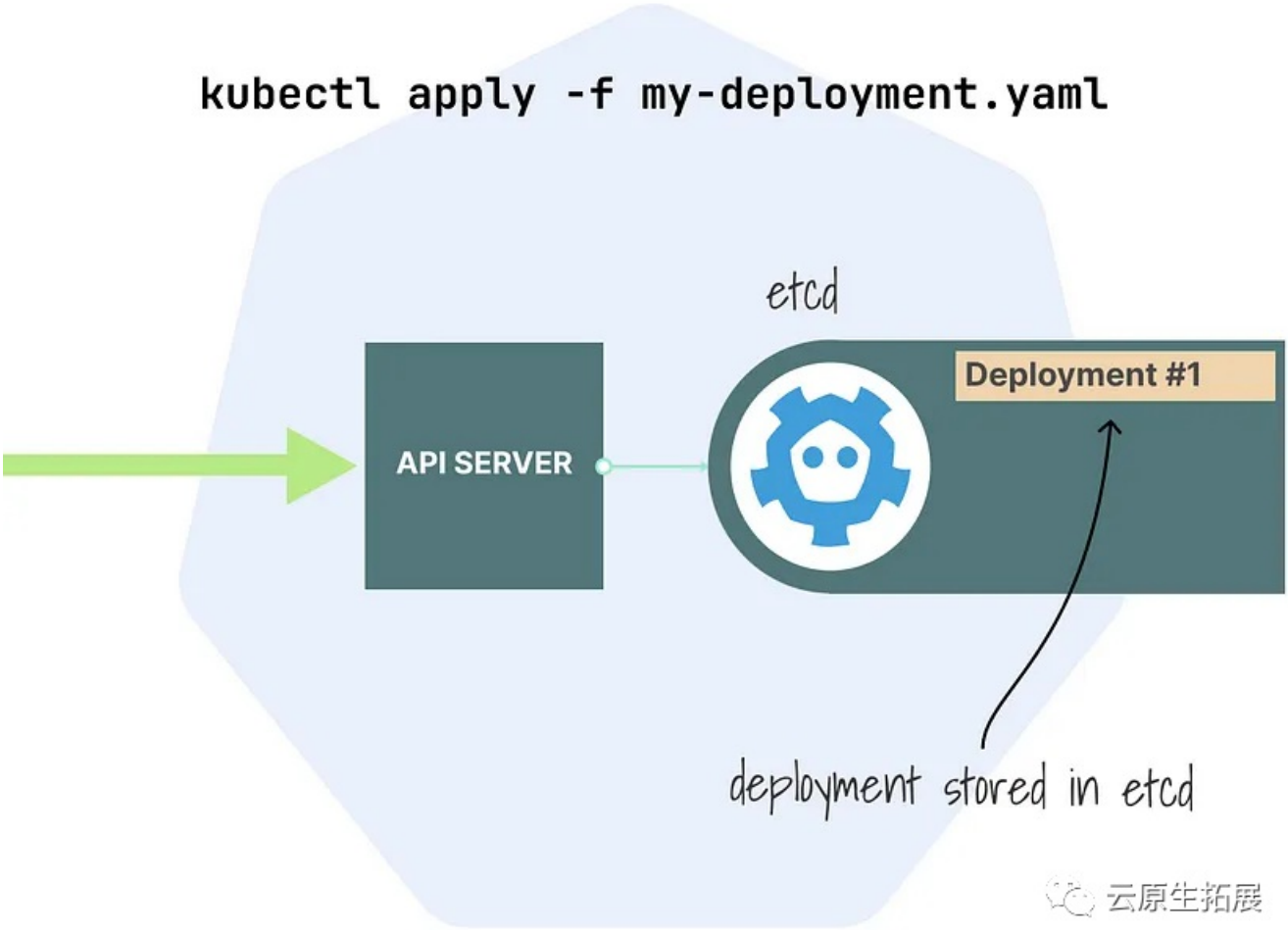
调度程序(scheduler)负责决定 Pod 在集群中的部署位置。

这听起来像是一件容易的事，但它相当复杂！

让我们从基础开始。

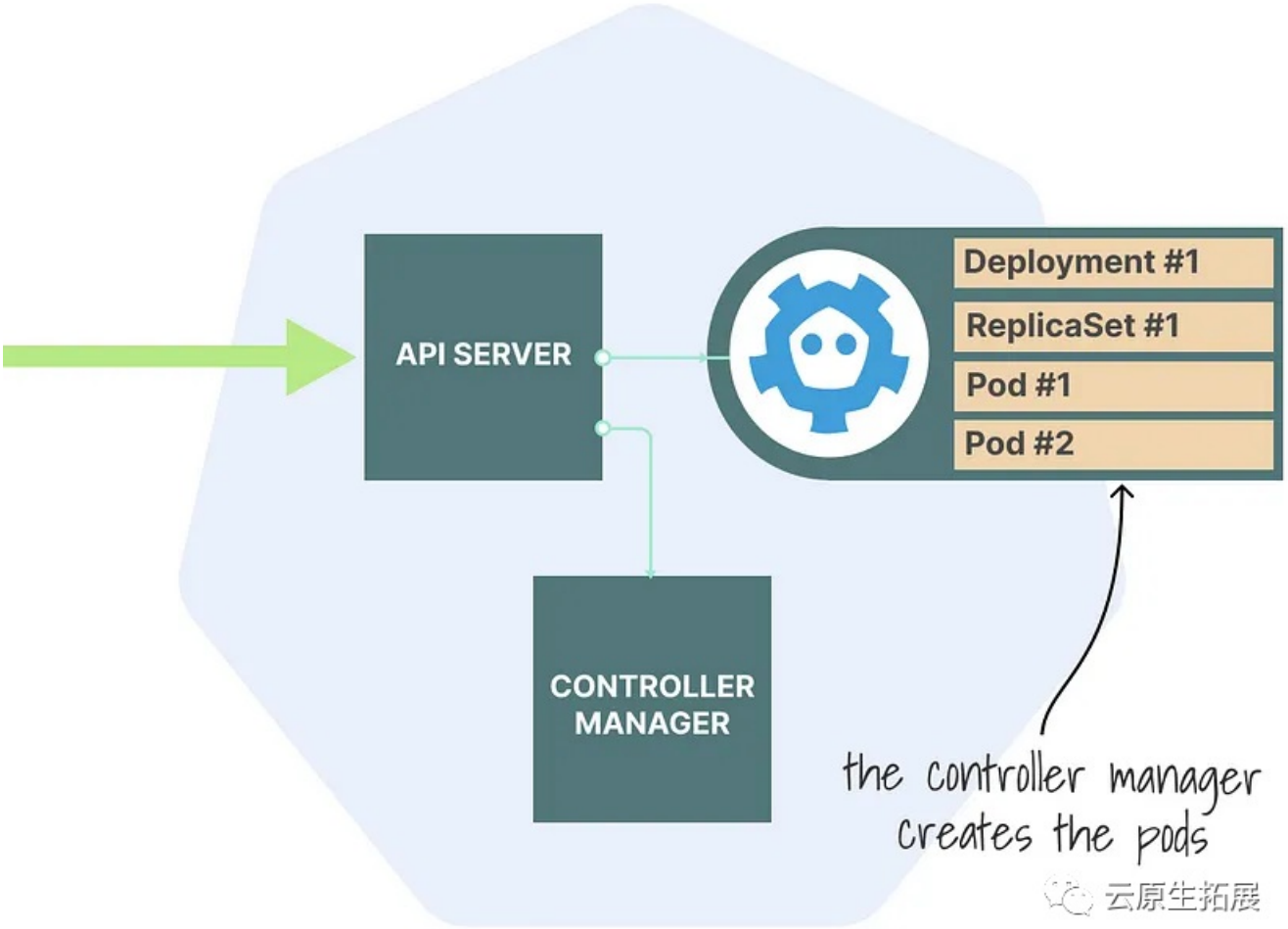
当你使用 kubectl 提交部署时，API Server 会收到请求，资源存储在 etcd 中。

谁创建 POD？



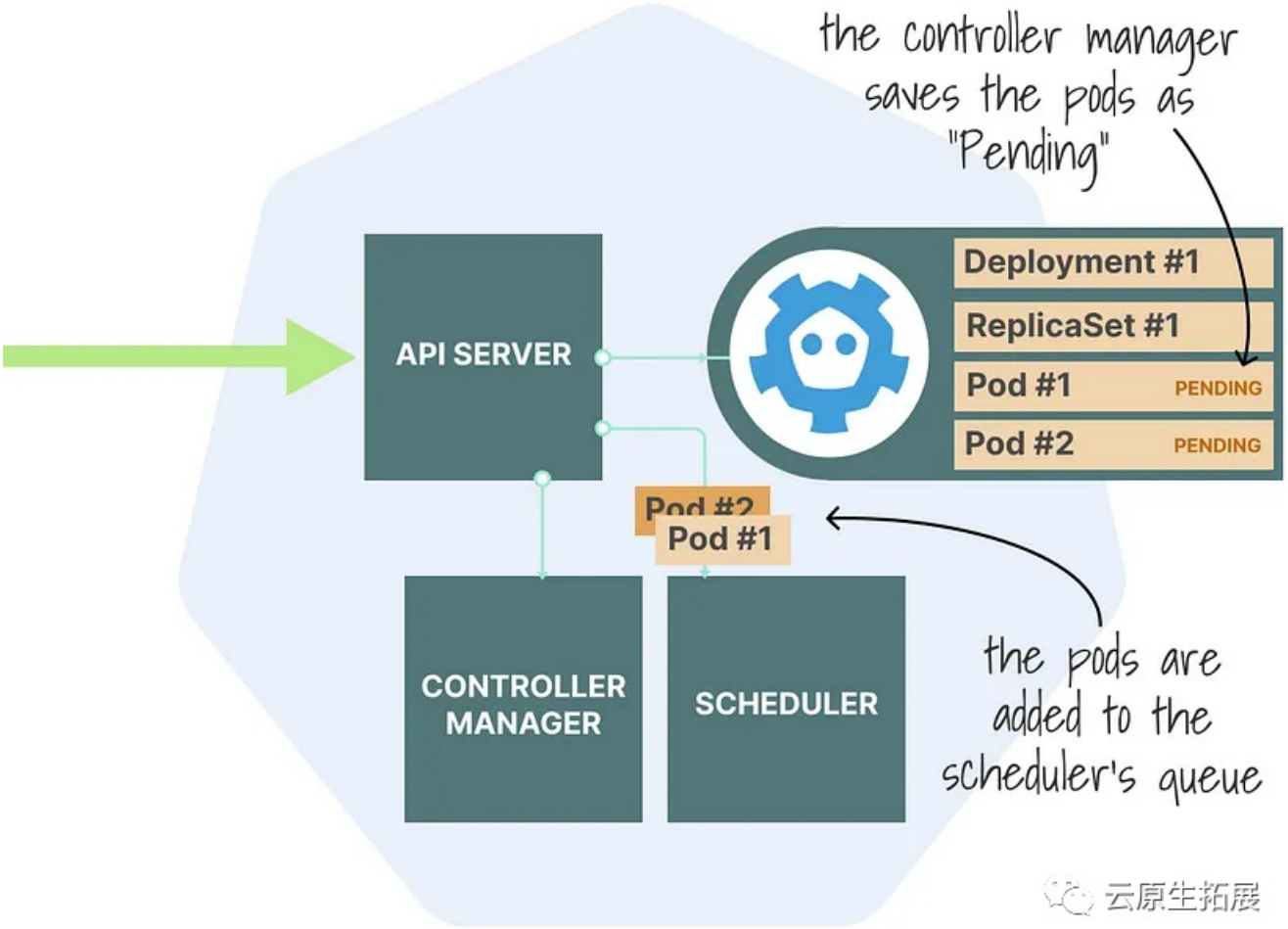
一个常见的误解是，创建 Pod 是调度程序的工作。

相反，控制器管理器会创建它们（以及关联的副本集- ReplicaSet）。



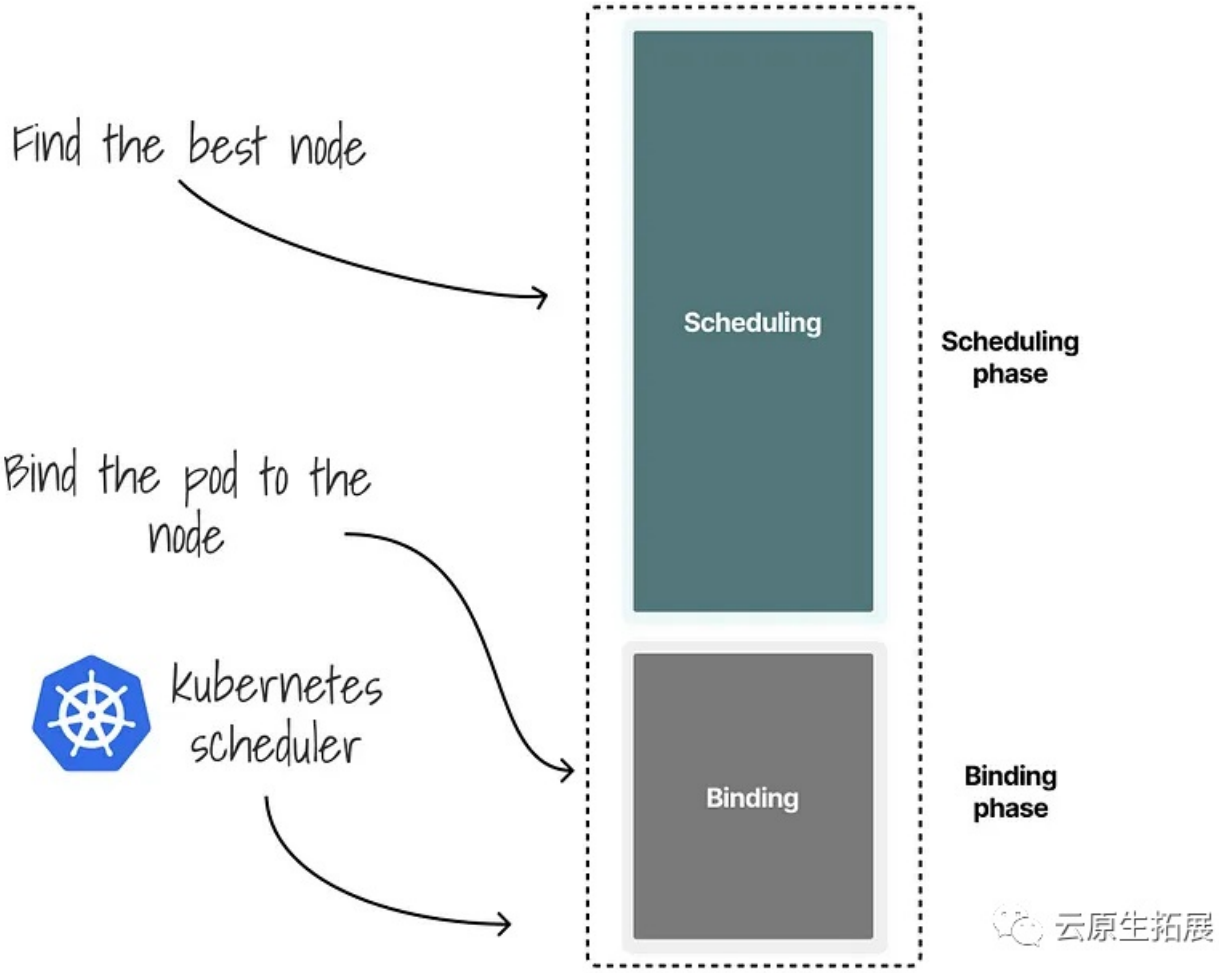
此时，Pod 在 etcd 中存储为 “Pending”，并且不会分配给任何节点。

它们也会添加到调度程序的队列中，准备分配。



调度程序通过两个阶段逐一处理 Pod：

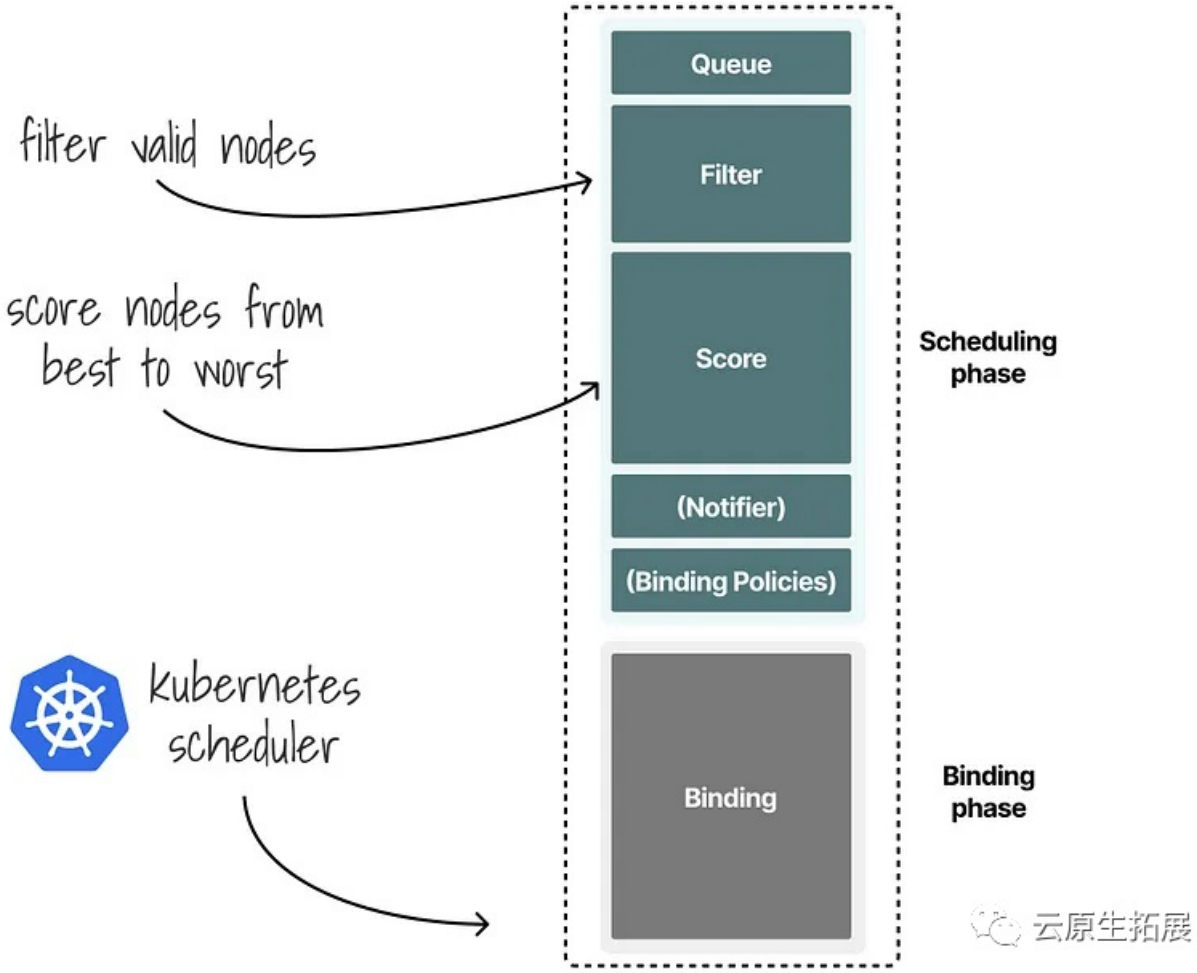
1. 调度阶段（我应该选择哪个节点？）
2. 绑定阶段（让我们将此 Pod 所属节点写入数据库）。



调度阶段分为两部分。调度程序：

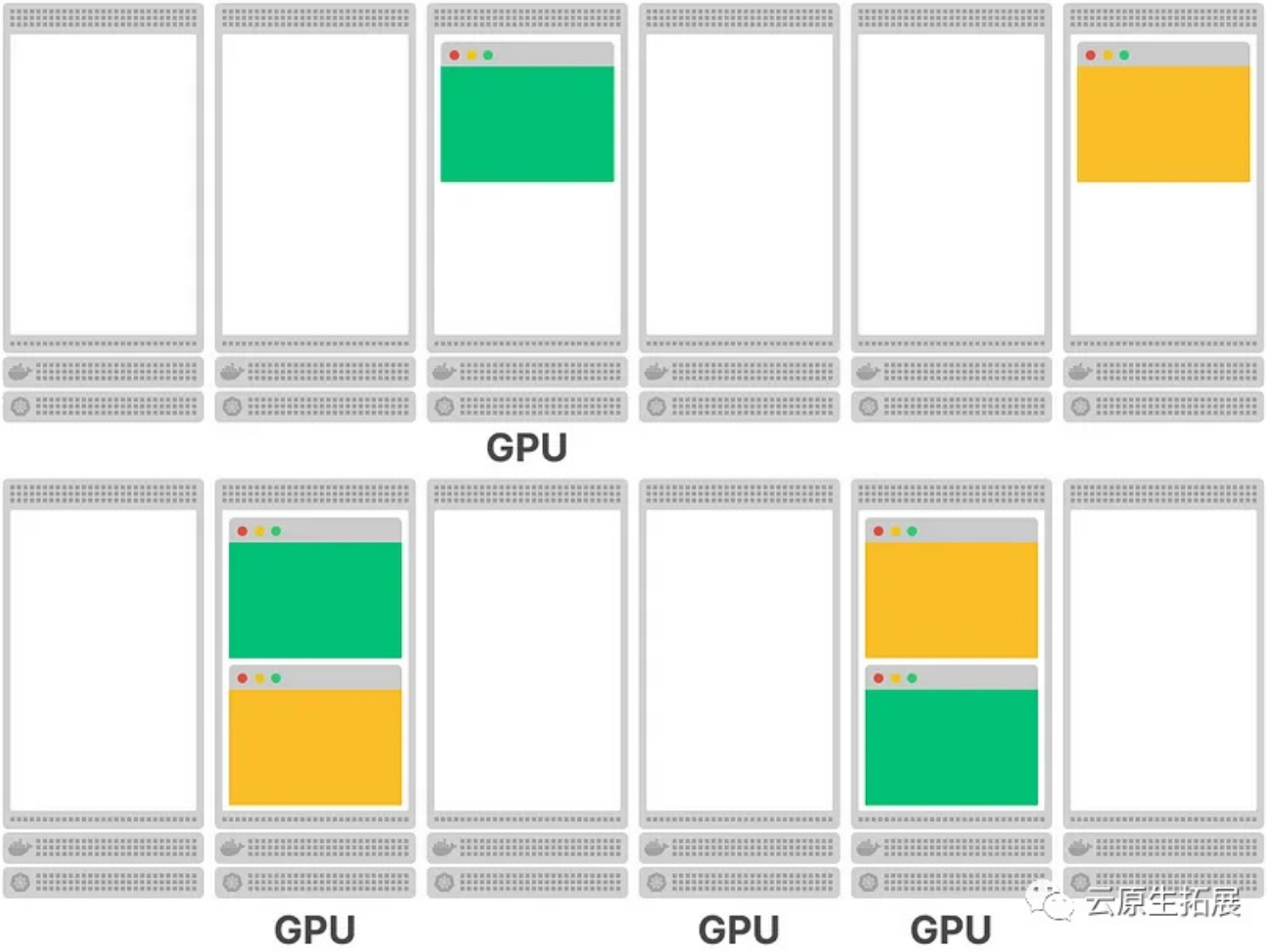
1. 筛选相关节点（使用称为 predicates 的函数列表）
2. 对其余节点进行排名（使用称为 priorities 的函数列表）

让我们看一个例子。



请考虑以下具有带和不带 GPU 的节点的群集。

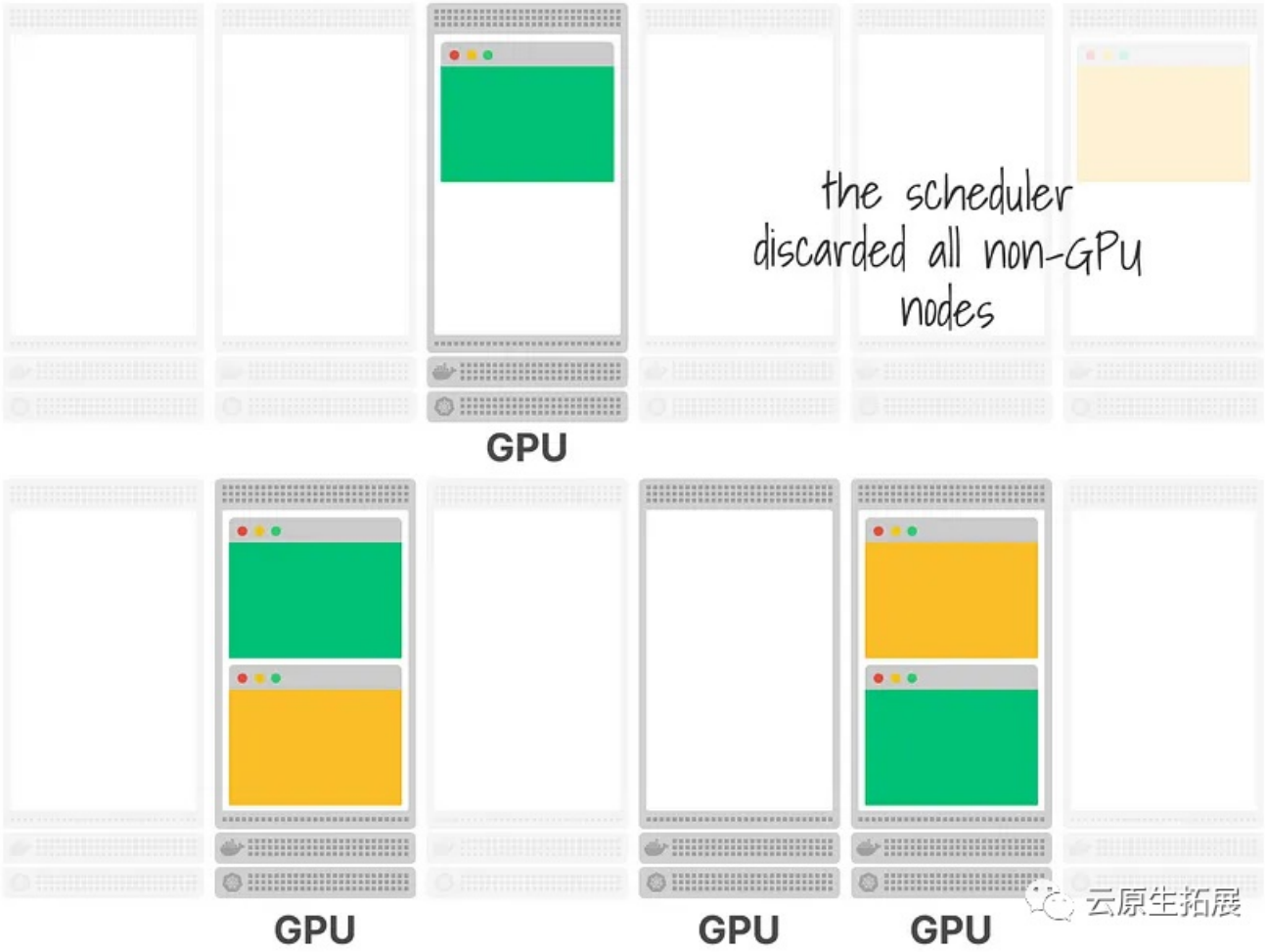
此外，一些节点已经在以总容量运行。



你想要部署一个需要一些 GPU 的 Pod。

将 Pod 提交到集群，然后将其添加到调度程序队列中。

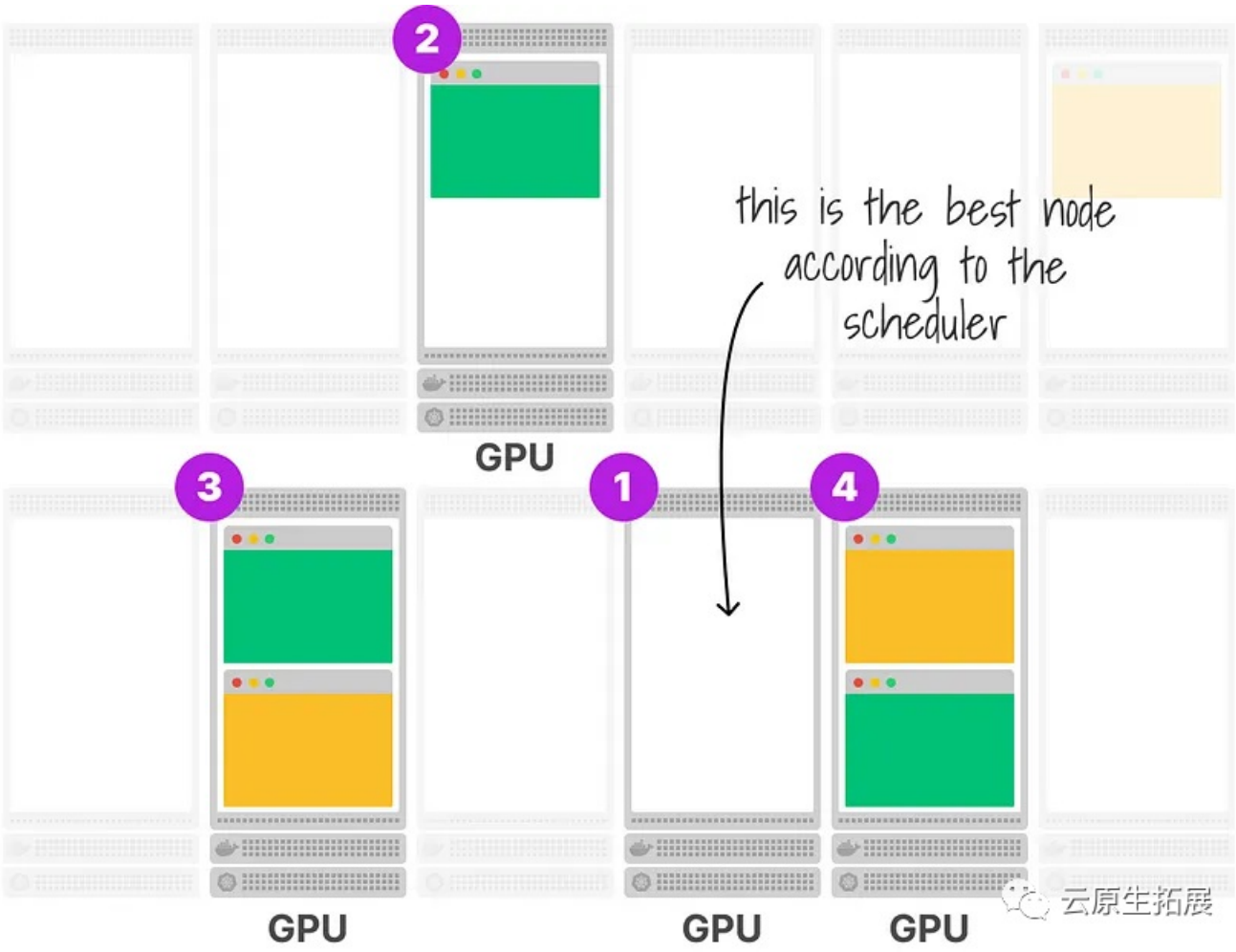
调度程序会丢弃所有没有 GPU 的节点（筛选阶段）。



接下来，调度程序对剩余节点进行评分。

在此示例中，充分利用的节点得分较低。

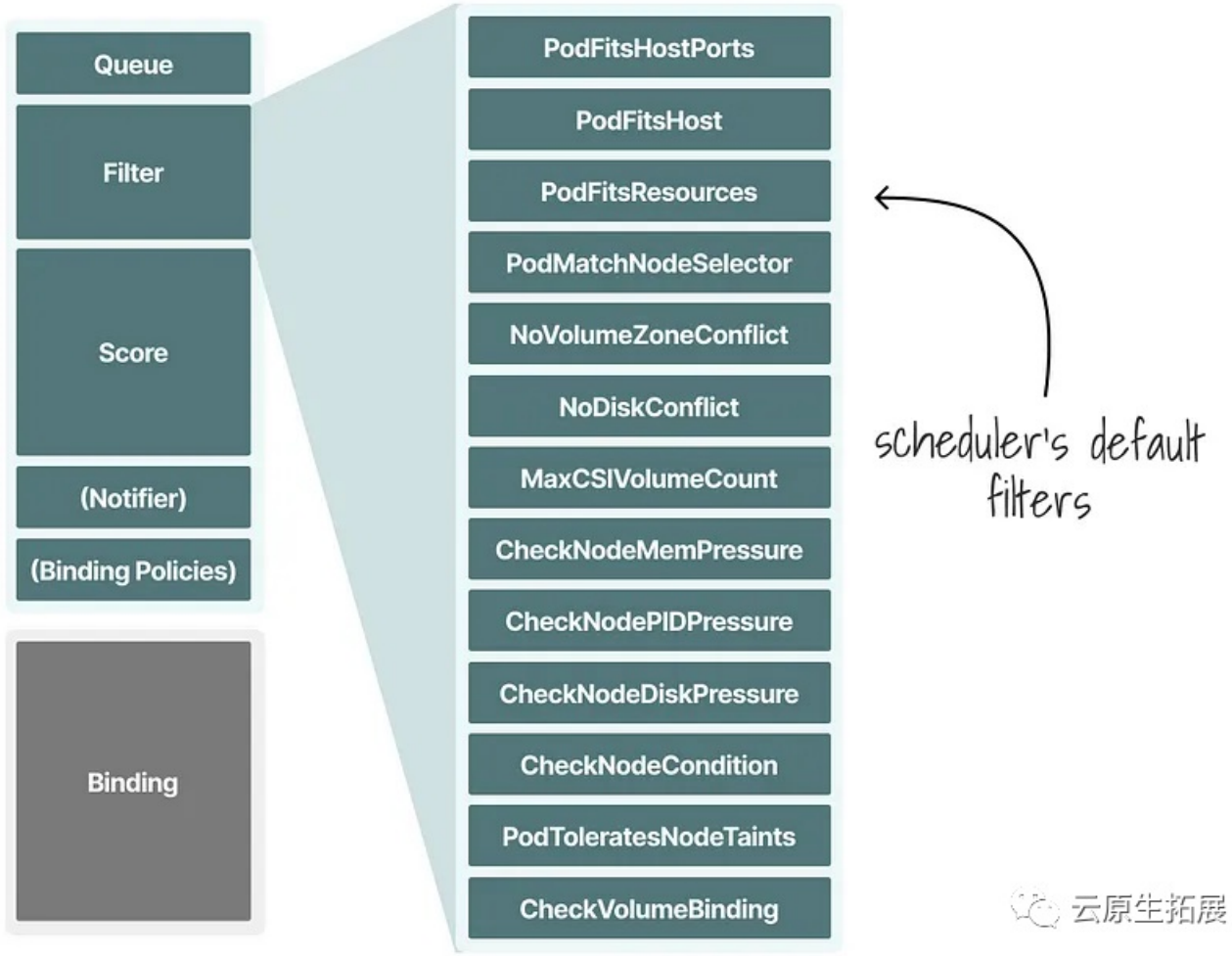
最后，选择空节点。



过滤器有哪些示例？

- `NodeUnschedulable` 防止 Pod 标记为不可调度的节点。
- `VolumeBinding` 检查节点是否可以绑定请求的卷。

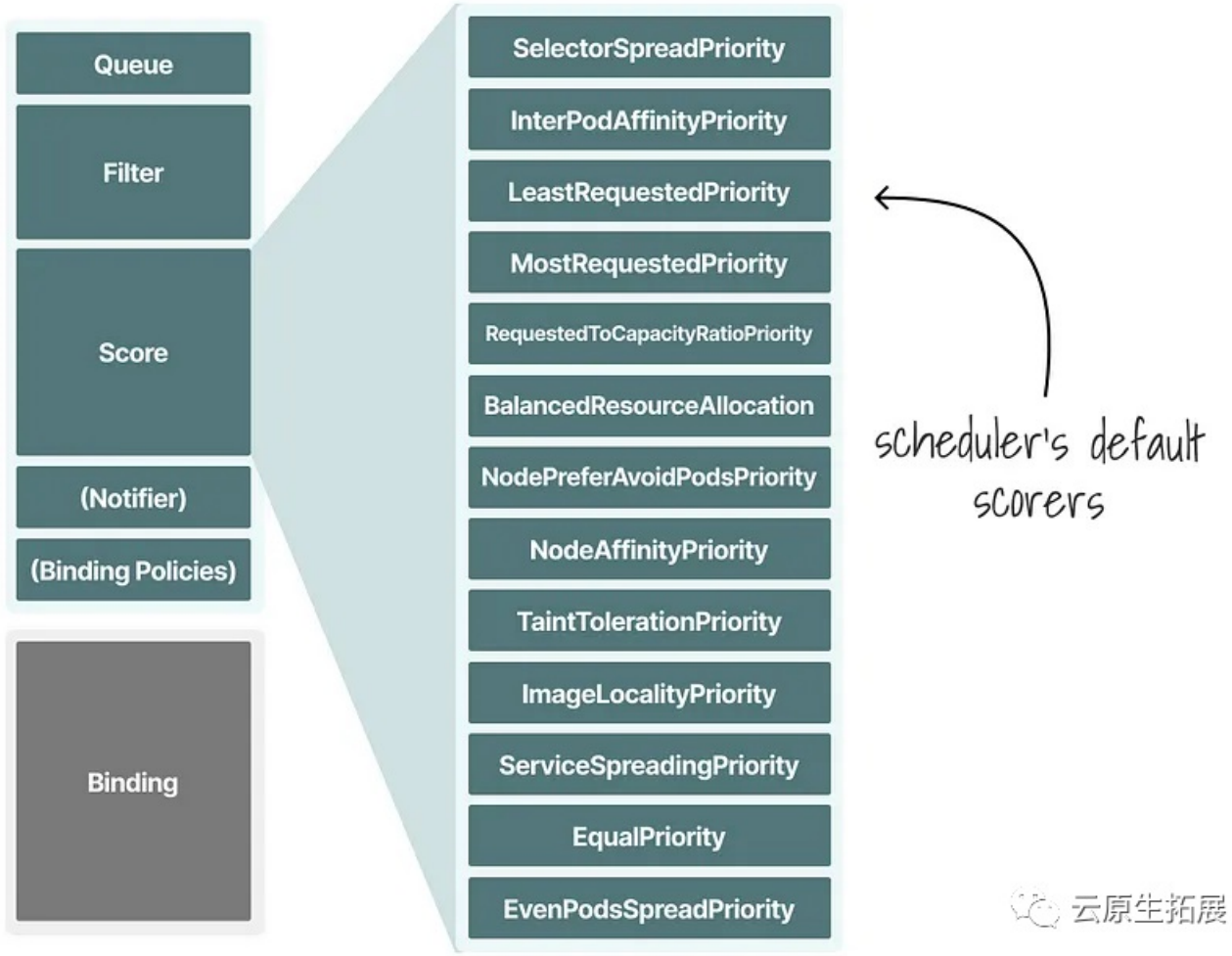
默认筛选阶段有 13 个 predicates。



以下是一些评分示例：

- `ImageLocality` 首选已在本地下载容器镜像的节点。
- `NodeResourcesBalancedAllocation` 首选未充分利用的节点。

有 13 个函数来决定如何对节点进行评分和排名。



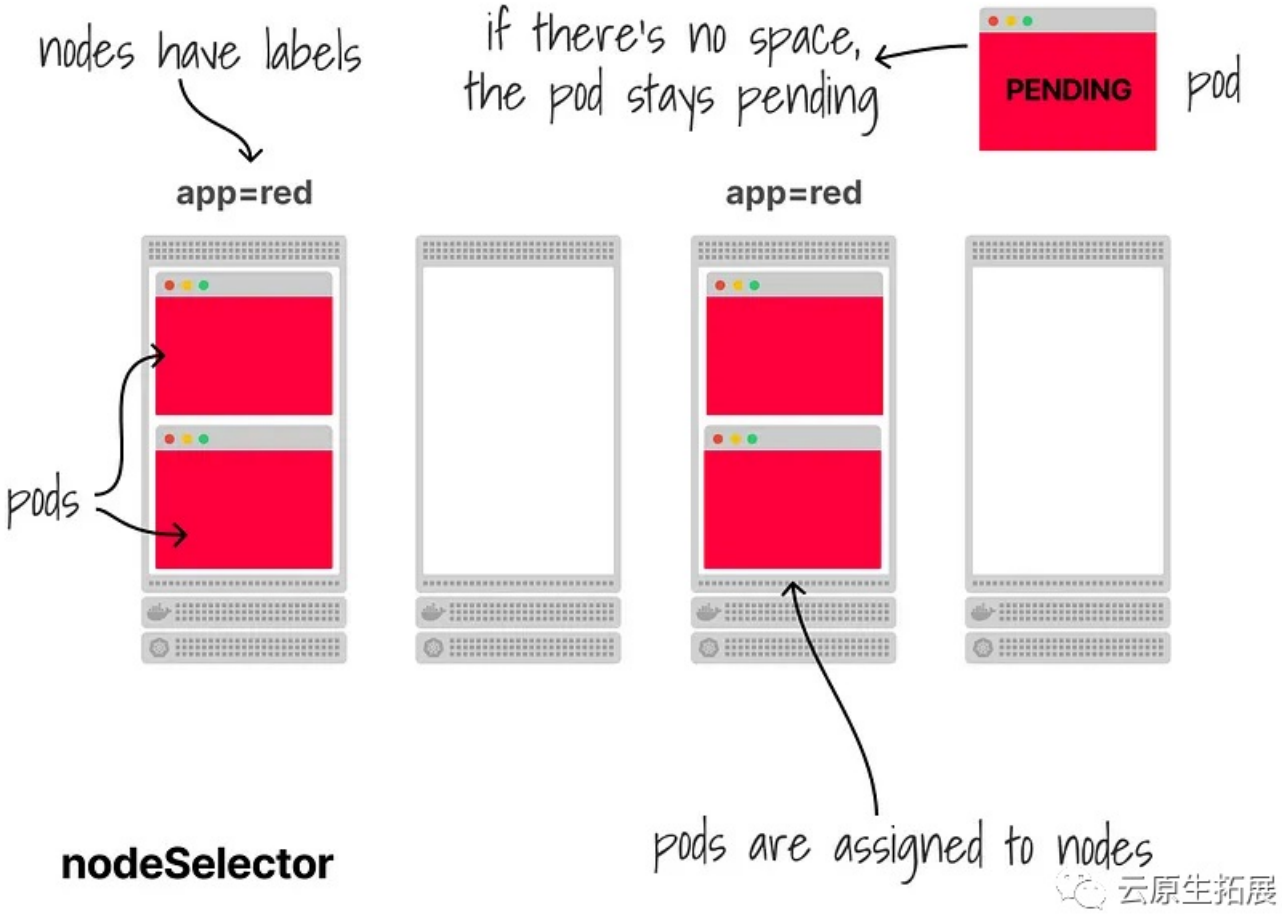
您如何影响调度程序的决策？

- nodeSelector
- node affinity
- Pod affinity / anti-affinity
- Taints 、tolerations
- Topology constraints
- Scheduler profiles

`nodeSelector` 是最直接的机制。

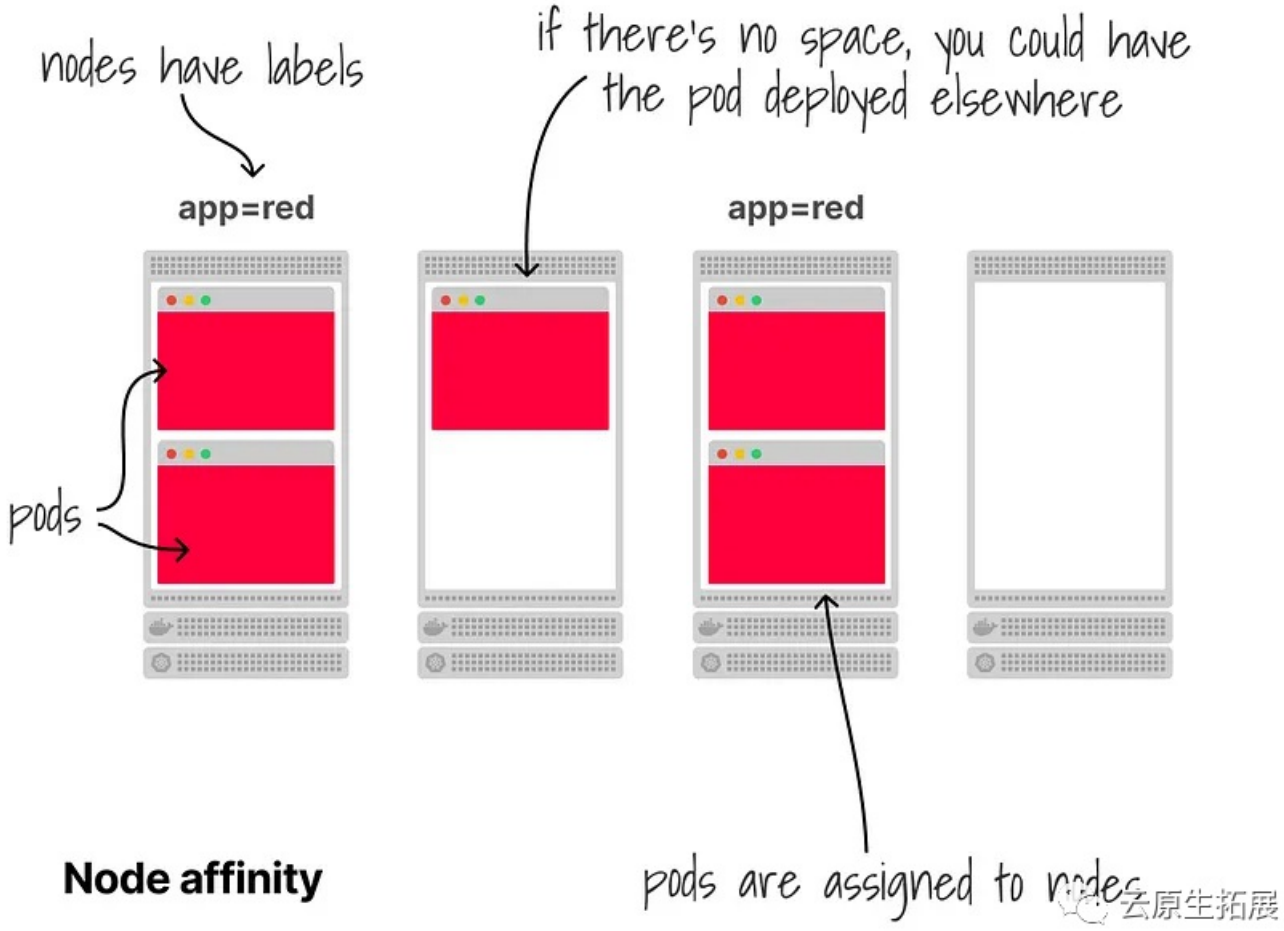
您可以为节点分配标签，并将该标签添加到 Pod 中。

Pod 只能部署在具有该标签的节点上。



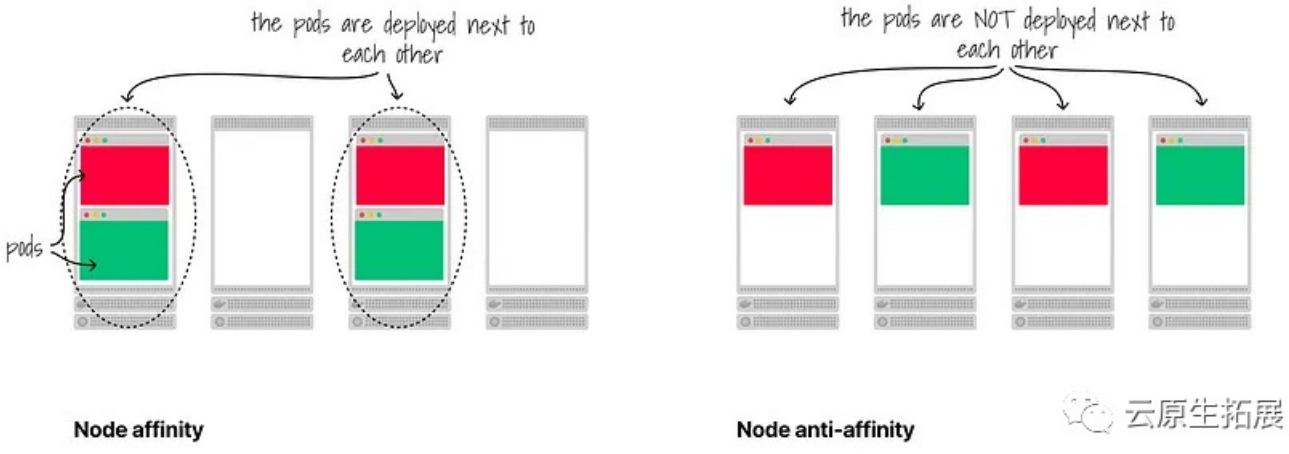
Node affinity 通过更灵活的接口扩展了 Node Selector。

你仍然可以告诉调度程序应该将 Pod 部署在哪里，但你也可以有软约束和硬约束。



使用 Pod 亲和/反亲和，您可以要求调度程序在特定 Pod 旁边放置一个 Pod(或者不)。

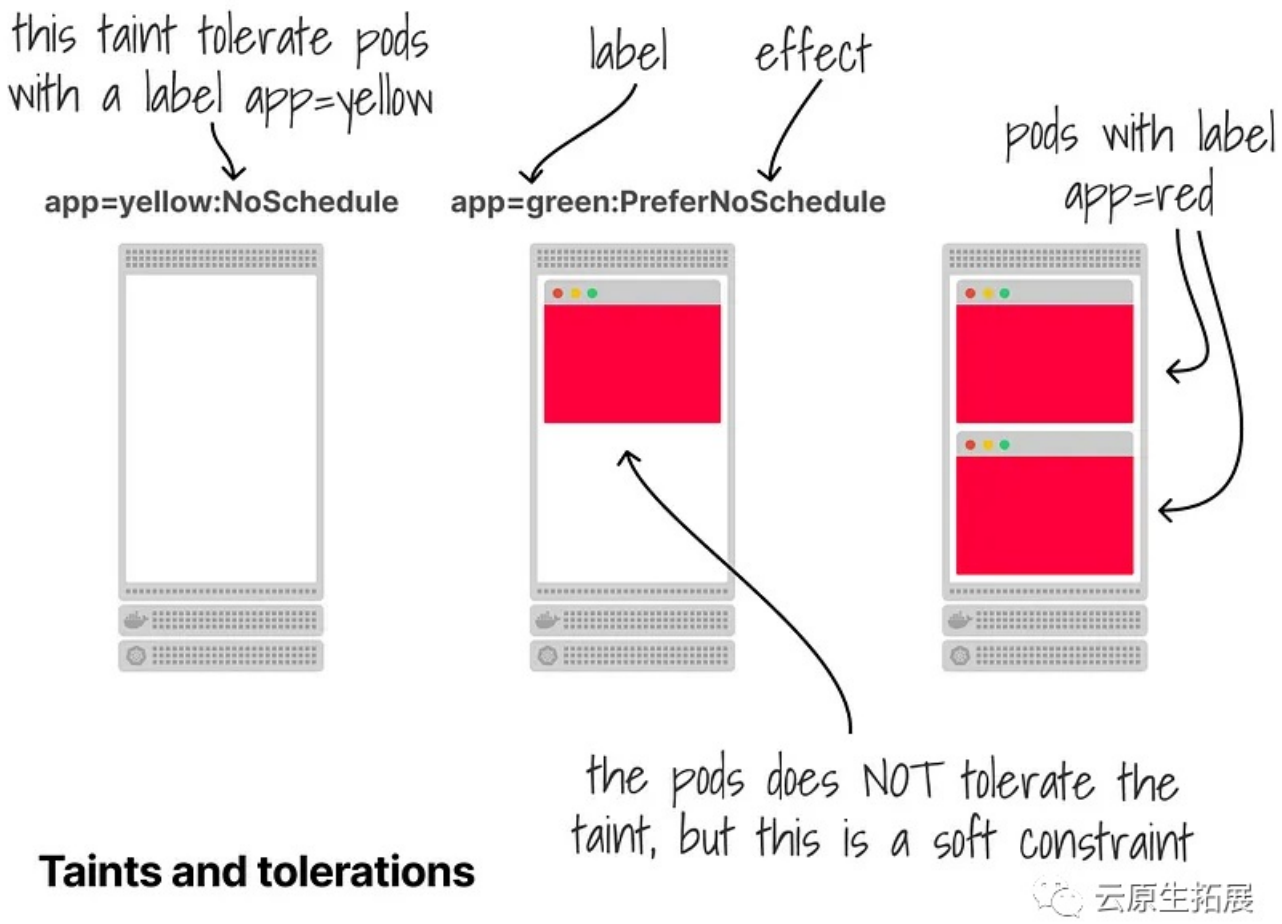
例如，您可以有一个具有反关联性的部署，以强制扩展 Pod。



通过污点（taints）和容忍度(tolerations)，Pod 被污染(tained)，节点排斥（或容忍）Pods。

这类似于节点亲和力，但有一个显着的区别：通过节点亲和力，Pod 会被节点吸引。

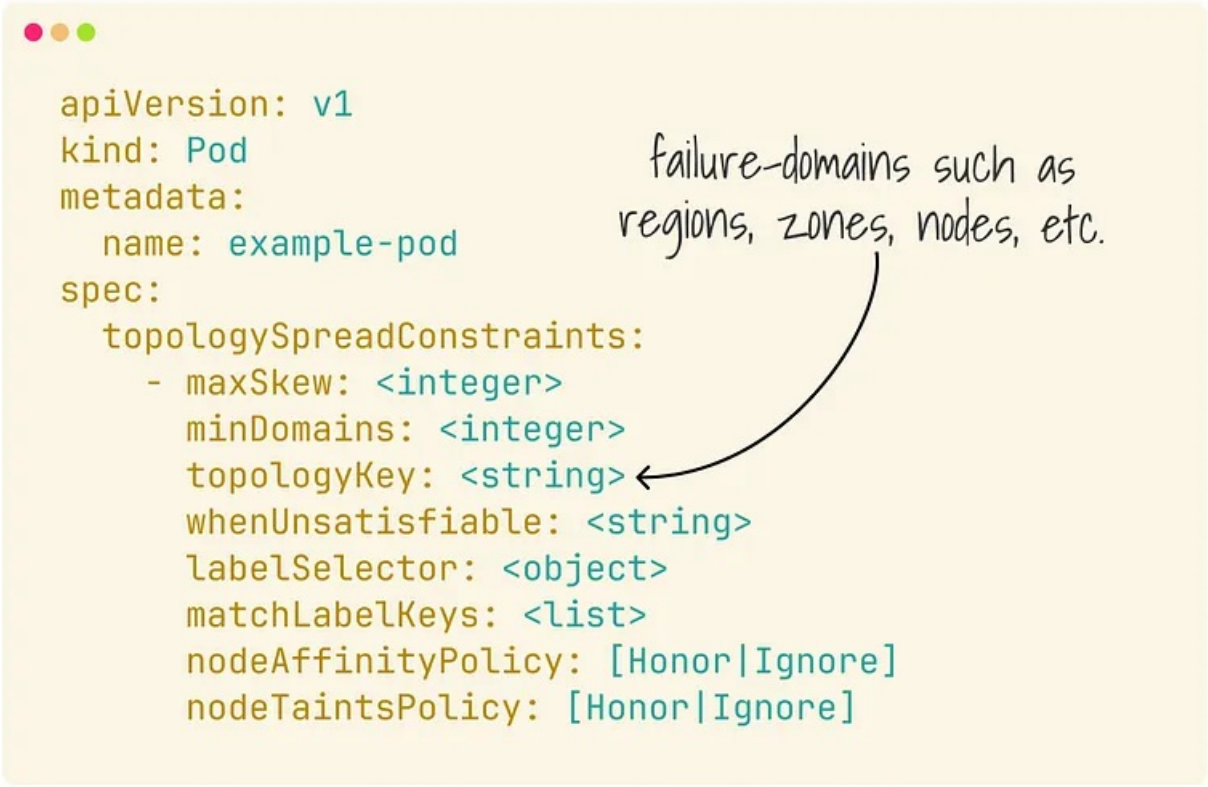
污点则相反——它们允许节点排斥 Pod。



此外，容忍可以通过三种效果排斥 Pod：驱逐(evict)、不调度（dont scheduler） 和 宁愿不调度。

您可以使用拓扑分布约束来控制 Pod 在集群中的分布方式。

当您想要确保所有 Pod 不会登陆同一节点时，这很方便。



最后，您可以使用调度程序策略自定义调度程序如何使用 filters 和 predicates 将节点分配给 Pod。

这个相对较新的功能 （>1.25） 允许您关闭调度程序或向调度程序添加新逻辑。

```yaml
apiVersion: kubescheduler.config.k8s.io/v1
kind: KubeSchedulerConfiguration
profiles:
  - plugins:
      score:
        disabled:
        - name: PodTopologySpread
        enabled:
        - name: MyCustomPluginA
          weight: 2
        - name: MyCustomPluginB
          weight: 1
```