K8s 污点和容忍度

AUTHOR: 彭玲 TIME: 2022/5/12

污点和容忍度

节点亲和性 是 Pod 的一种属性，它使 Pod 被吸引到一类特定的 节点 （这可能出于一种偏好，也可能是硬性要求）。 污点 （Taint）则相反——它使节点能够排斥一类特定的 Pod。

给 node 加污点

使用命令 kubectl taint 给节点增加一个污点。比如，给节点 `node1` 增加一个污点，它的键名是 `key1`，键值是 `value1`，效果是 `NoSchedule`。 这表示只有拥有和这个污点相匹配的容忍度的 Pod 才能够被分配到 `node1` 这个节点。

```
kubectl taint nodes node1 key1=value1:NoSchedule # NoSchedule 为 effect(效果)
```

定义 Pod 的容忍度

您可以在 PodSpec 中定义 Pod 的容忍度。Pod 的 容忍度 (`tolerations`) 配置中， `operator` 的默认值是 `Equal`。

一个容忍度和一个污点相“匹配”是指它们有一样的键名和效果，并且：

- 如果 `operator` 是 `Exists` （此时容忍度不能指定 `value`），或者
- 如果 `operator` 是 `Equal`，则它们的 `value` 应该相等。

存在两种特殊情况：

> - 如果一个容忍度的 `key` 为空且 operator 为 `Exists`， 表示这个容忍度与任意的 key 、value 和 effect 都匹配，即这个容忍度能容忍任意 taint。
> - 如果 `effect` 为空，则可以与所有键名 `key1` 的效果相匹配。

下面两个容忍度均与上面例子中使用 `kubectl taint` 命令创建的污点相匹配，因此如果一个 Pod 拥有其中的任何一个容忍度都能够被分配到 `node1`：

```yaml
tolerations:
- key: "key1"
  operator: "Equal"
  value: "value1"
  effect: "NoSchedule"

# 或者
tolerations:
- key: "key1"
  operator: "Exists"
  effect: "NoSchedule"
```

效果（`effect`）

K8s 处理多个`污点`和`容忍度`的过程就像一个过滤器：

从一个节点的所有污点开始遍历， 过滤掉那些 Pod 中存在与之相匹配的容忍度的污点。余下未被过滤的污点的`effect`值决定了 Pod 是否会被分配到该节点，特别是以下情况：

- 如果未被过滤的污点中存在至少一个 effect 值为 `NoSchedule` 的污点， 则 Kubernetes 不会将 Pod 分配到该节点。
- 如果未被过滤的污点中不存在 effect 值为 `NoSchedule` 的污点， 但是存在 effect 值为 `PreferNoSchedule` 的污点， 则 Kubernetes 会 尝试 不将 Pod 分配到该节点。
- 如果未被过滤的污点中存在至少一个 effect 值为 `NoExecute` 的污点， 则 Kubernetes 不会将 Pod 分配到该节点（如果 Pod 还未在节点上运行）， 或者将 Pod 从该节点驱逐（如果 Pod 已经在节点上运行）。

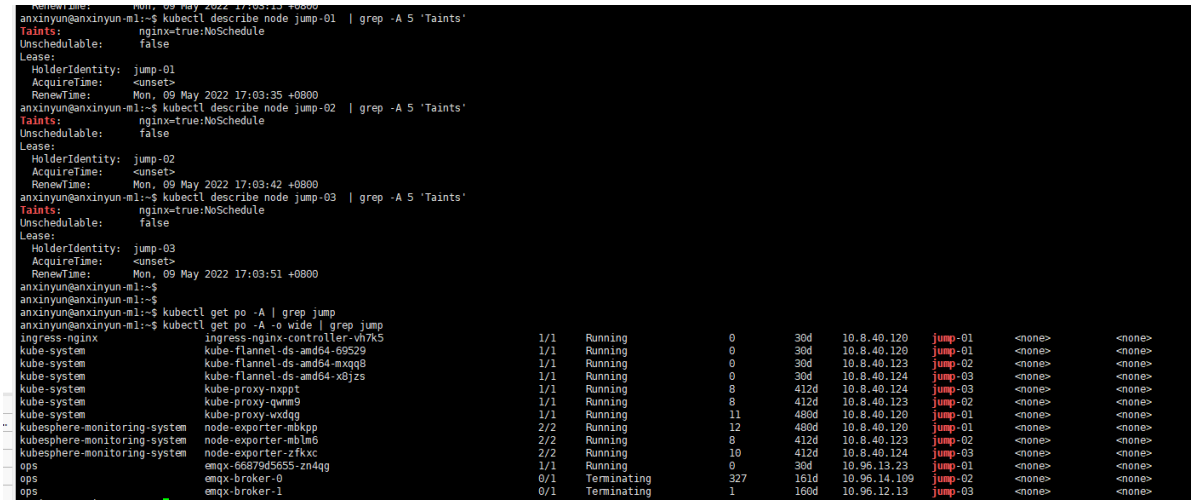
Pod 示例

下面是一个使用了容忍度的 Pod：

```yaml
# pods/pod-with-toleration.yaml
apiVersion: v1
kind: Pod
metadata:
  name: nginx
  labels:
    env: test
spec:
  containers:
  - name: nginx
    image: nginx
    imagePullPolicy: IfNotPresent
  tolerations: # 容忍度
  - key: "example-key"
    operator: "Exists"
    effect: "NoSchedule"
```

应用案例

污点 (Taints)



```
anxinyun@anxinyun-m1:~$ kubectl describe node jump-01  | grep -A 5 'Taints'
Taints:             nginx=true:NoSchedule
Unschedulable:      false
Lease:
  HolderIdentity:  jump-01
  AcquireTime:     <unset>
  RenewTime:       Mon, 09 May 2022 17:03:35 +0800
anxinyun@anxinyun-m1:~$ kubectl describe node jump-02  | grep -A 5 'Taints'
Taints:             nginx=true:NoSchedule
Unschedulable:      false
Lease:
  HolderIdentity:  jump-02
  AcquireTime:     <unset>
  RenewTime:       Mon, 09 May 2022 17:03:42 +0800
anxinyun@anxinyun-m1:~$ kubectl describe node jump-03  | grep -A 5 'Taints'
Taints:             nginx=true:NoSchedule
Unschedulable:      false
Lease:
  HolderIdentity:  jump-03
  AcquireTime:     <unset>
  RenewTime:       Mon, 09 May 2022 17:03:51 +0800
anxinyun@anxinyun-m1:~$
anxinyun@anxinyun-m1:~$
anxinyun@anxinyun-m1:~$ kubectl get po -A | grep jump
anxinyun@anxinyun-m1:~$ kubectl get po -A -o wide | grep jump
ingress-nginx               ingress-nginx-controller-vh7k5               1/1     Running             0          30d     10.8.40.120   jump-01   <none>           <none>
kube-system                 kube-flannel-ds-amd64-69529                  1/1     Running             0          30d     10.8.40.120   jump-01   <none>           <none>
kube-system                 kube-flannel-ds-amd64-mxqq8                  1/1     Running             0          30d     10.8.40.123   jump-02   <none>           <none>
kube-system                 kube-flannel-ds-amd64-x8jzs                  1/1     Running             0          30d     10.8.40.124   jump-03   <none>           <none>
kube-system                 kube-proxy-nxppt                             1/1     Running             8          412d    10.8.40.124   jump-03   <none>           <none>
kube-system                 kube-proxy-qwnm9                             1/1     Running             8          412d    10.8.40.123   jump-02   <none>           <none>
kube-system                 kube-proxy-wxdqg                             1/1     Running             11         480d    10.8.40.120   jump-01   <none>           <none>
kubesphere-monitoring-system   node-exporter-mbkpp                       2/2     Running             12         480d    10.8.40.120   jump-01   <none>           <none>
kubesphere-monitoring-system   node-exporter-mblm6                       2/2     Running             8          412d    10.8.40.123   jump-02   <none>           <none>
kubesphere-monitoring-system   node-exporter-zfkxc                       2/2     Running             10         412d    10.8.40.124   jump-03   <none>           <none>
ops                         emqx-66879d5655-zn4qg                        1/1     Running             0          30d     10.96.13.23   jump-01   <none>           <none>
ops                         emqx-broker-0                                0/1     Terminating         327        161d    10.96.14.109  jump-02   <none>           <none>
ops                         emqx-broker-1                                0/1     Terminating         1          160d    10.96.12.13   jump-03   <none>           <none>
```

容忍度 (Tolerations) 配置文件

```yaml
kind: DaemonSet
apiVersion: apps/v1
metadata:
  name: ingress-nginx-controller
  namespace: ingress-nginx
  labels:
    app.kubernetes.io/component: controller
    app.kubernetes.io/instance: ingress-nginx
    app.kubernetes.io/name: ingress-nginx
spec:
  selector:
    matchLabels:
      app.kubernetes.io/component: controller
      app.kubernetes.io/instance: ingress-nginx
      app.kubernetes.io/name: ingress-nginx
  template:
    metadata:
      creationTimestamp: null
      labels:
        app.kubernetes.io/component: controller
        app.kubernetes.io/instance: ingress-nginx
        app.kubernetes.io/name: ingress-nginx
    spec:
      containers:
        - name: controller
          image: 'k8s.gcr.io/ingress-nginx/controller:v0.40.2'
          args:
            - /nginx-ingress-controller
            - '--election-id=ingress-controller-leader'
            - '--ingress-class=nginx'
            - '--configmap=$(POD_NAMESPACE)/ingress-nginx-controller'
            - '--tcp-services-configmap=$(POD_NAMESPACE)/tcp-services'
            - '--udp-services-configmap=$(POD_NAMESPACE)/udp-services'
            - '--validating-webhook=:8443'
            - '--validating-webhook-
certificate=/usr/local/certificates/cert'
            - '--validating-webhook-key=/usr/local/certificates/key'
          ports:
```

```
            - name: http
              hostPort: 80
              containerPort: 80
              protocol: TCP
            - name: https
              hostPort: 443
              containerPort: 443
              protocol: TCP
            - name: webhook
              hostPort: 8443
              containerPort: 8443
              protocol: TCP
          env:
            - name: POD_NAME
              valueFrom:
                fieldRef:
                  apiVersion: v1
                  fieldPath: metadata.name
            - name: POD_NAMESPACE
              valueFrom:
                fieldRef:
                  apiVersion: v1
                  fieldPath: metadata.namespace
            - name: LD_PRELOAD
              value: /usr/local/lib/libmimalloc.so
          livenessProbe:
            httpGet:
              path: /healthz
              port: 10254
              scheme: HTTP
            initialDelaySeconds: 10
            timeoutSeconds: 1
            periodSeconds: 10
            successThreshold: 1
            failureThreshold: 5
          readinessProbe:
            httpGet:
              path: /healthz
              port: 10254
              scheme: HTTP
            initialDelaySeconds: 10
            timeoutSeconds: 1
            periodSeconds: 10
            successThreshold: 1
            failureThreshold: 3
          lifecycle:
            preStop:
              exec:
                command:
                  - /wait-shutdown
          securityContext:
            capabilities:
              add:
                - NET_BIND_SERVICE
              drop:
                - ALL
            runAsUser: 101
            allowPrivilegeEscalation: true
```

```
      nodeSelector:
        ingress: nginx
      hostNetwork: true
      tolerations: # 容忍度
        - key: nginx
          operator: Exists
```