KubeCon | CloudNativeCon

North America 2022

BUILDING FOR THE ROAD AHEAD

DETROIT 2022

How To Handle Node Shutdown in Kubernetes

KubeCon | CloudNativeCon
North America 2022

BUILDING FOR THE ROAD AHEAD

DETROIT 2022

October 24-28, 2021

Ashutosh Kumar
Engineer @VMware India

Xing Yang
Tech Lead @VMware

@sonasingh46 & @2000Xyang

Agenda

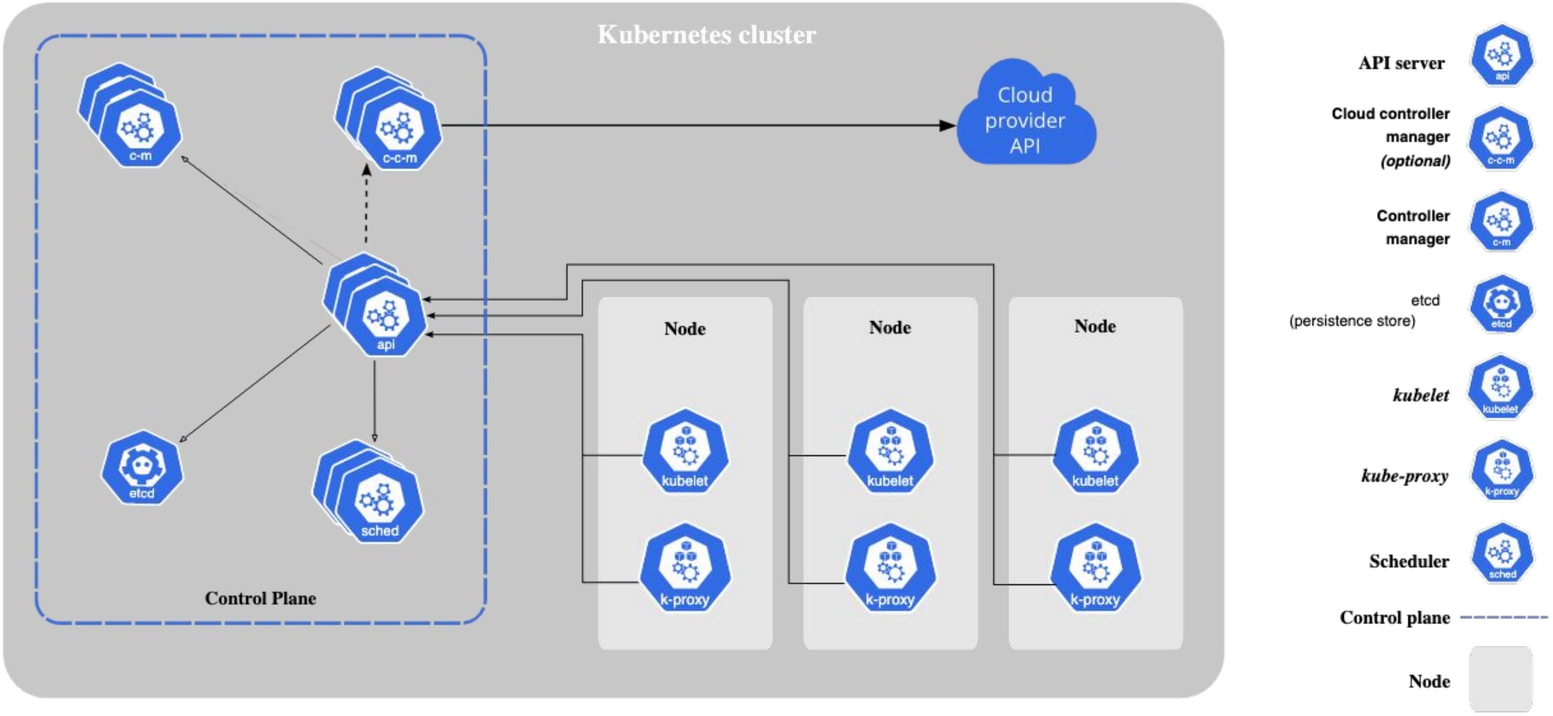
- What is a node shutdown in Kubernetes
- What is a graceful shutdown in Kubernetes
- What is a non graceful shutdown
  - Impact of a non graceful shutdown
  - How non graceful shutdown is handled in Kubernetes
- Demo: non graceful node shutdown
- Next steps

Node Shutdown: Introduction

- Node shutdowns are inevitable in K8s cluster which can result in workload failures.

- Node shutdown causes:
  - Hardware failure
  - Reboot due to a security patch
  - Preemption of short lived cloud compute instances
  - ……..

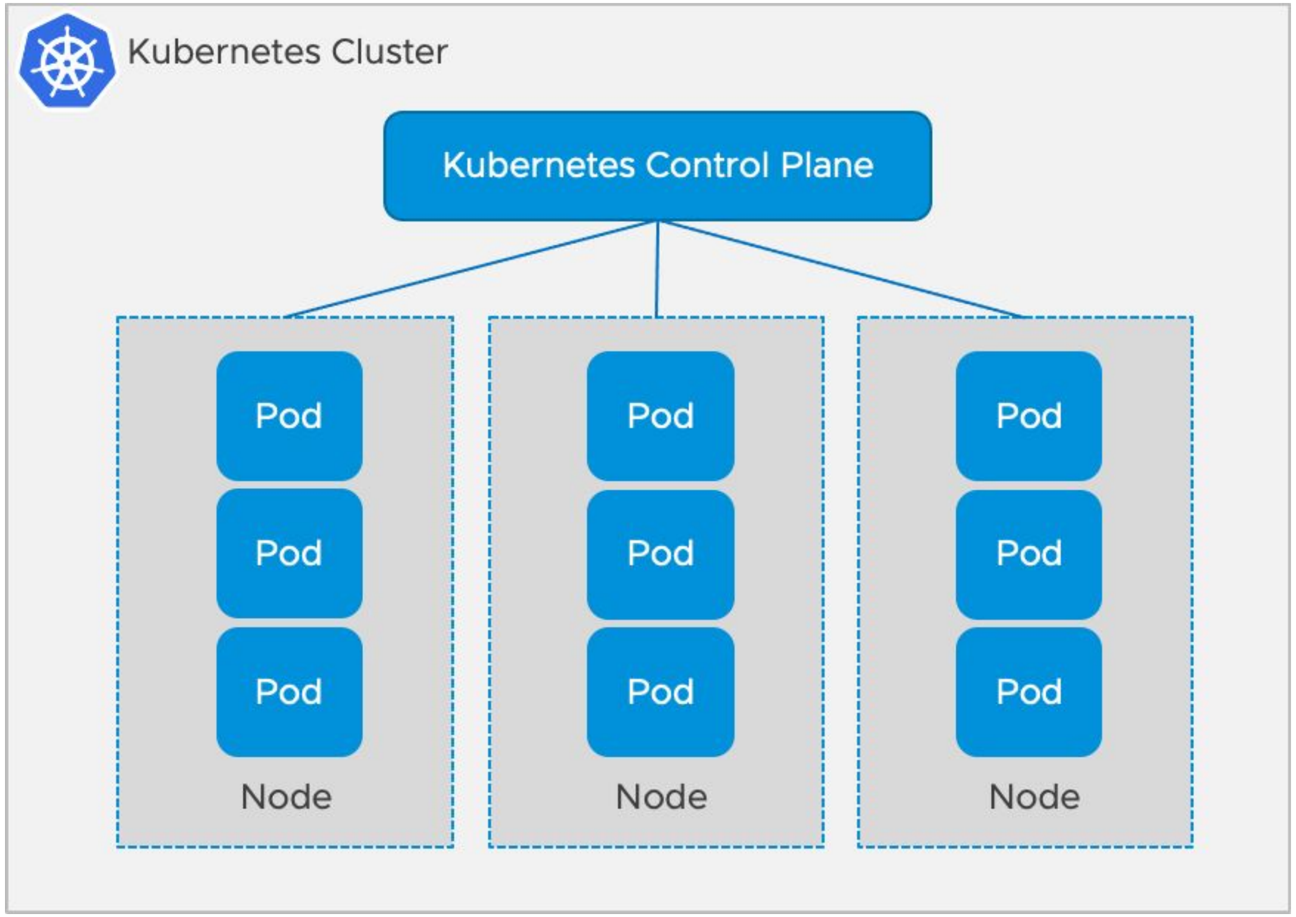
- A node shutdown could be
  - Graceful
  - Non graceful



Source: *https://kubernetes.io/docs/concepts/overview/components/*

Graceful Shutdown: The What

- Introduced in K8s v1.20 and Beta in v1.21.
- Ability of Kubelet to detect a shutdown ahead of the actual shutdown.
- Kubelet can propagate this event to pods ensuring they shut down gracefully, possibly releasing resources that are being hold.
- Pods with priority class “system-cluster-critical” or “system-node-critical” will be terminated after all other pods.



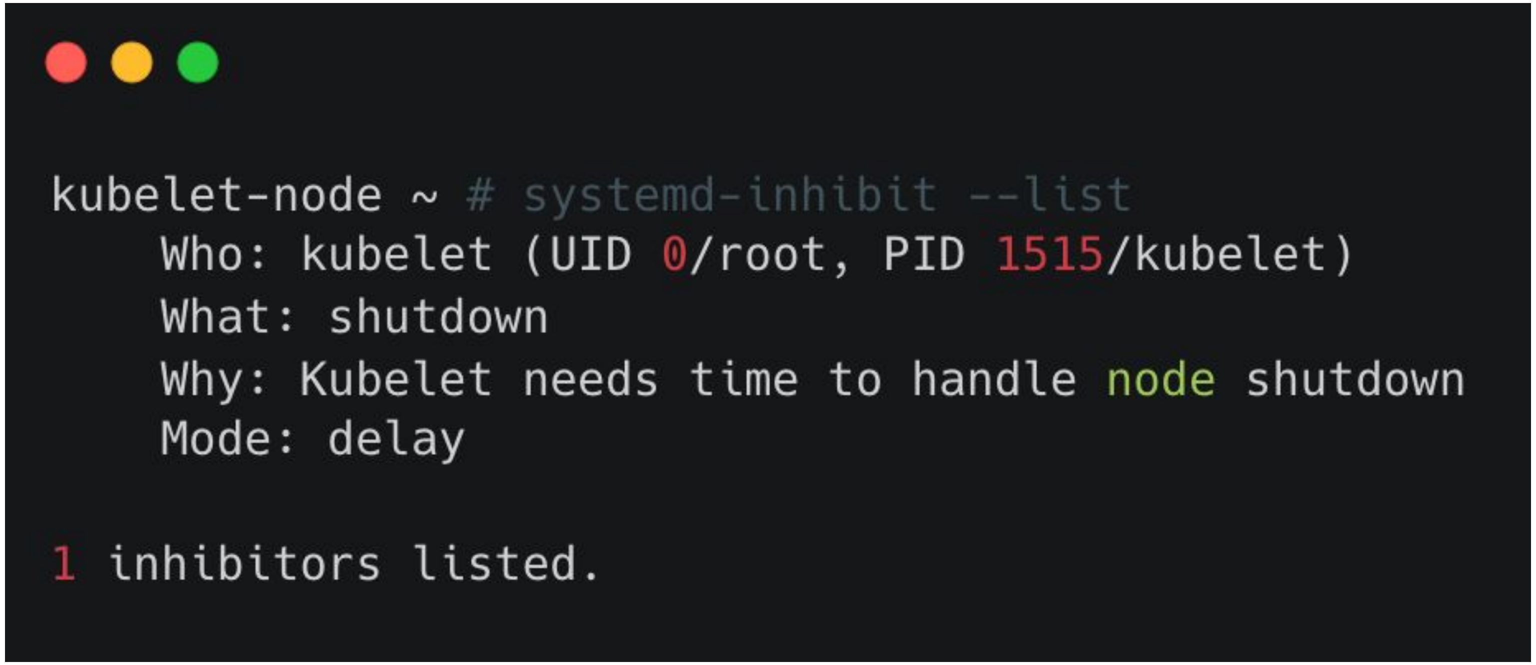
Graceful Shutdown: The Why

- Prior to this feature, safe draining of nodes required manual intervention.
- Automations that could cause a node restart, required to explicitly drain nodes for safe eviction.
- However, a node shutdown could happen unexpectedly, resulting in unsafe eviction of pods.
- Applications might see errors due to the pods exiting abruptly.

Graceful Shutdown: The How

Kubelet uses “Inhibitor Locks” to postpone shutdown for a specified duration giving a chance for the node to drain and evict pods.

- When Kubelet starts, it acquires the delay type inhibitor lock.
- At shutdown event, Kubelet delays the shutdown for a configurable period of time.



```
kubelet-node ~ # systemd-inhibit --list
    Who: kubelet (UID 0/root, PID 1515/kubelet)
    What: shutdown
    Why: Kubelet needs time to handle node shutdown
    Mode: delay

1 inhibitors listed.
```

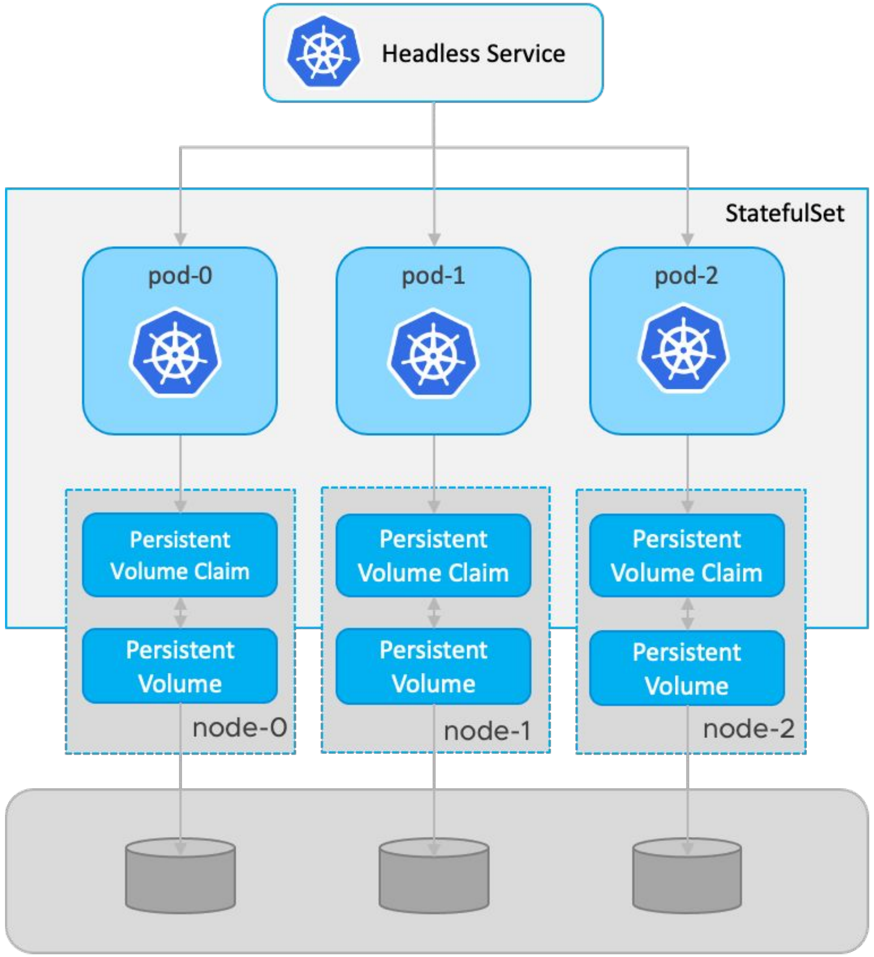
Graceful Shutdown: The How

- Configurable Kubelet parameters
  - ShutdownGracePeriod
    - Duration that the node should delay the shutdown. This is the total time available for termination of both regular and critical pods.
  - ShutdownGracePeriodCriticalPods
    - Duration used to terminate critical pods. Should be always less than ShutdownGracePeriod.
- Alpha feature Pod Priority based Graceful Node Shutdown in K8s v1.23.

Non Graceful Shutdown: The What

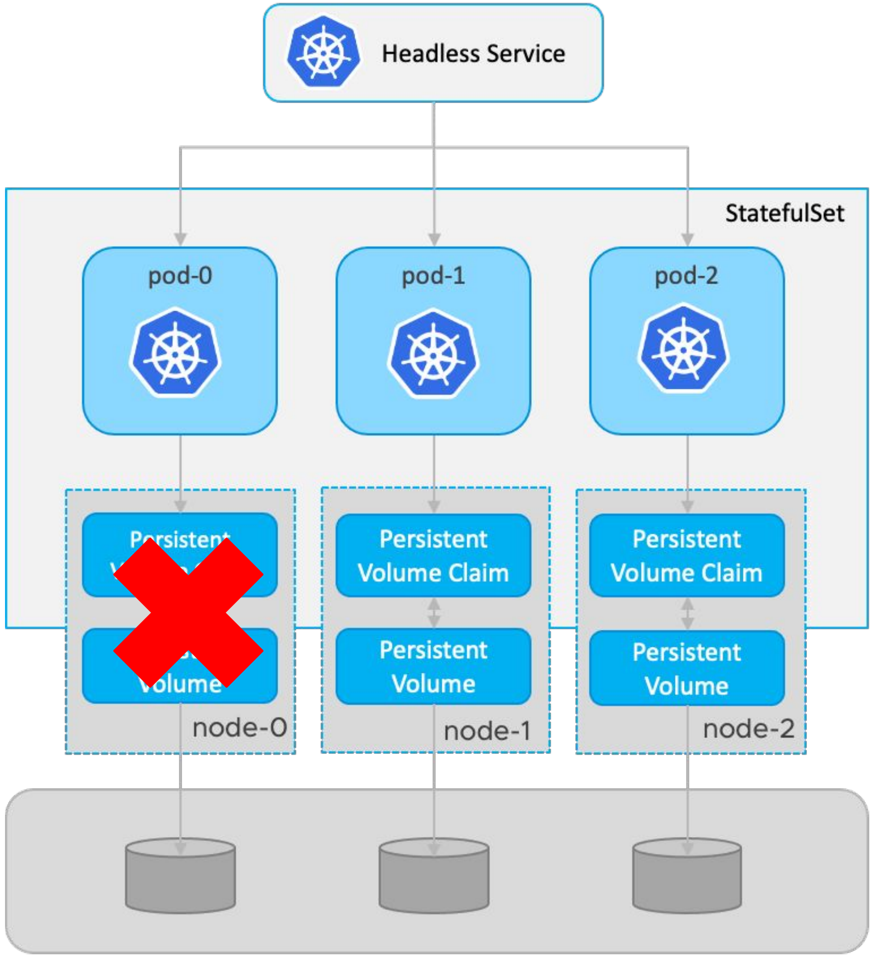
Graceful node shutdown feature in K8s

- Introduced as an Alpha feature in K8s v1.24. Disabled by default. Targeting Beta in v1.26.
- Shutdown that cannot be detected by Kubelet is a non graceful shutdown and this feature handles such shutdown.
- This is especially a problem for stateful pods.



Non Graceful Shutdown: The Why

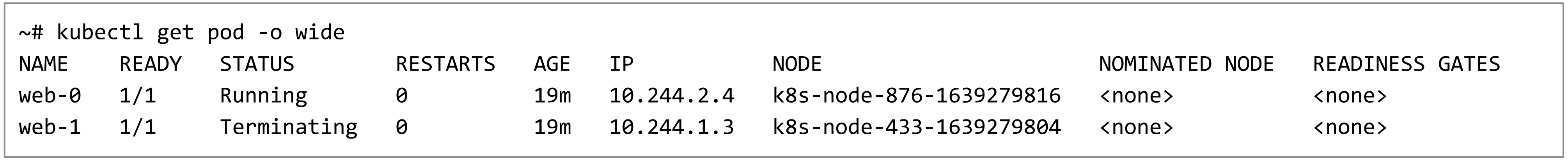
- The shutdown does not trigger the systemd inhibitor lock.
- ShutdownGracePeriod and ShutdownGracePeriodCriticalPods are not configured properly.
- In case of a non graceful shutdown, the pod moves to Terminating state.
- If the same node comes online, Kubelet detects and it works fine.
- If the original node fails to come online, the pod is stuck in Terminating state.



Non Graceful Shutdown: The Why

Feature gate is disabled

- Created a statefulset.
- Shutdown one node using the `Shut Down Guest OS` from vSphere UI.
- Observed that after 5 mins, the pod changed to `Terminating` state.
- Observed that even after 6 mins, (i.e total 6+5 = 11 mins) the pod is stuck in `Terminating` state.

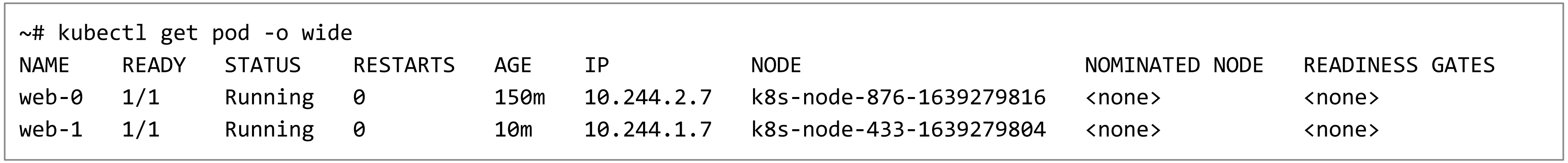


```
~# kubectl get pod -o wide
NAME     READY   STATUS        RESTARTS   AGE   IP           NODE                        NOMINATED NODE   READINESS GATES
web-0    1/1     Running       0          19m   10.244.2.4   k8s-node-876-1639279816    <none>           <none>
web-1    1/1     Terminating   0          19m   10.244.1.3   k8s-node-433-1639279804    <none>           <none>
```

Non Graceful Shutdown: The Why

Feature gate is disabled

- Created a statefulset.
- Shutdown one node using the `Shut Down Guest OS` from vSphere UI.
- Observed that after 5 mins, the pod changed to `Terminating` state.
- Manually deleted the pod using `kubectl delete pod <pod-name> --force --grace-period 0`.
- The pod immediately got scheduled to a different healthy node but was stuck in `ContainerCreating` state for 6 mins. The pod came into `Running` state after 6 mins.



```
~# kubectl get pod -o wide
NAME     READY   STATUS    RESTARTS   AGE    IP           NODE                         NOMINATED NODE   READINESS GATES
web-0    1/1     Running   0          150m   10.244.2.7   k8s-node-876-1639279816      <none>           <none>
web-1    1/1     Running   0          10m    10.244.1.7   k8s-node-433-1639279804      <none>           <none>
```

Non Graceful Shutdown: Scope

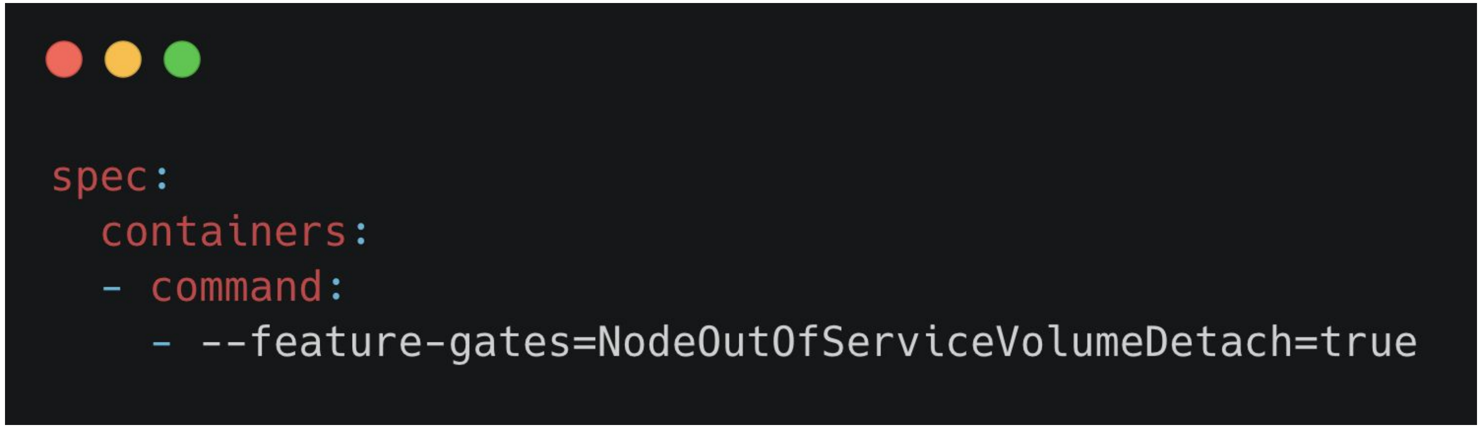
- Goals
  - Help increase availability of stateful workloads in case node goes into a non-recoverable cases e.g hardware failure or broken OS.
- Non-Goals
  - Node/control plane partitioning other than a node shutdown.
  - In-cluster logic to handle node/control plane partitioning

Non Graceful Shutdown: The How

- Uses Taint `node.kubernetes.io/out-of-service` to handle the shutdown.
- As of now, this feature requires manual intervention i.e. tainting the node that has shutdown non gracefully and may not come back.
- After the taint is applied:
  - Pod GC controller forcefully deletes the pods that do not have a matching toleration.
  - Attach-Detach controller immediately performs a force volume detach operation.
- Now, the new pod comes up successfully quickly on a different node, as the volume can be attached to this pod.

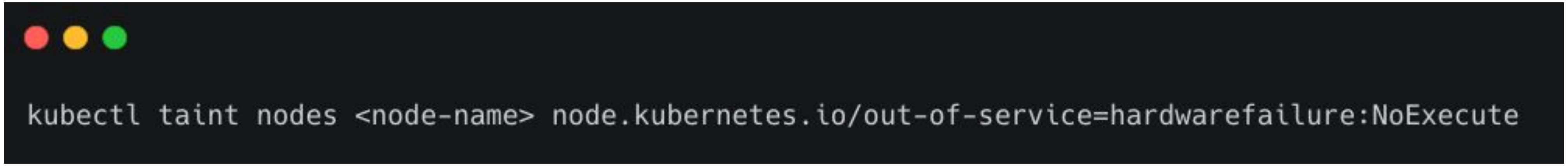
Non Graceful Shutdown: The How

- Feature gate can be enabled by setting the NodeOutOfServiceVolumeDetach flag true in the following manner.



```
spec:
  containers:
  - command:
    - --feature-gates=NodeOutOfServiceVolumeDetach=true
```

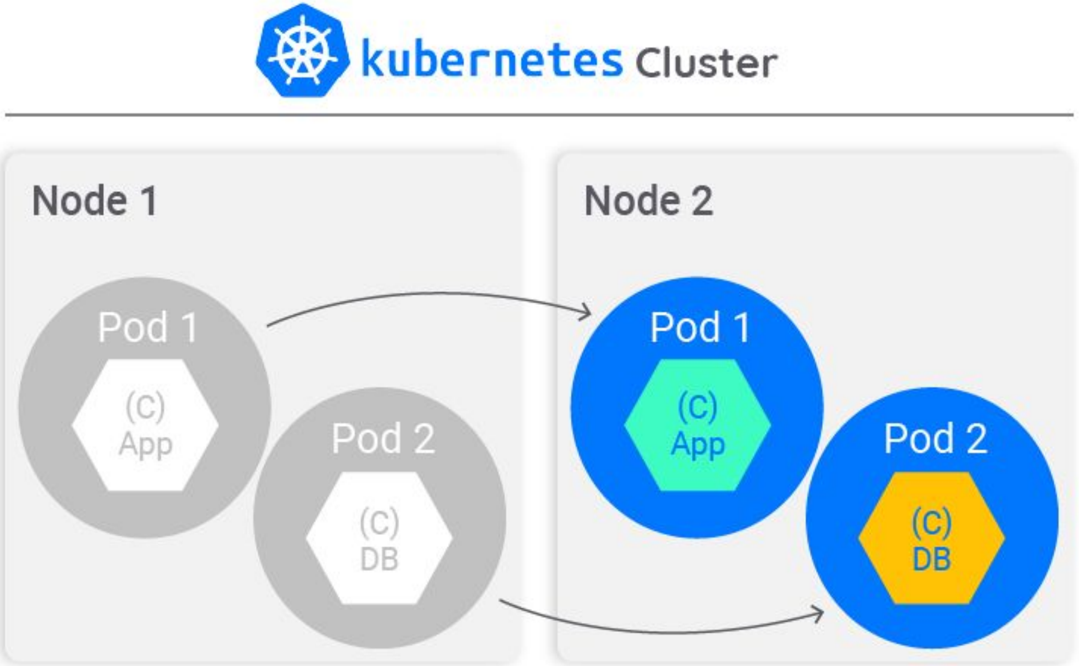
- Once a node is identified that has been shutdown non gracefully, it can be tainted using the following command.



```
kubectl taint nodes <node-name> node.kubernetes.io/out-of-service=hardwarefailure:NoExecute
```

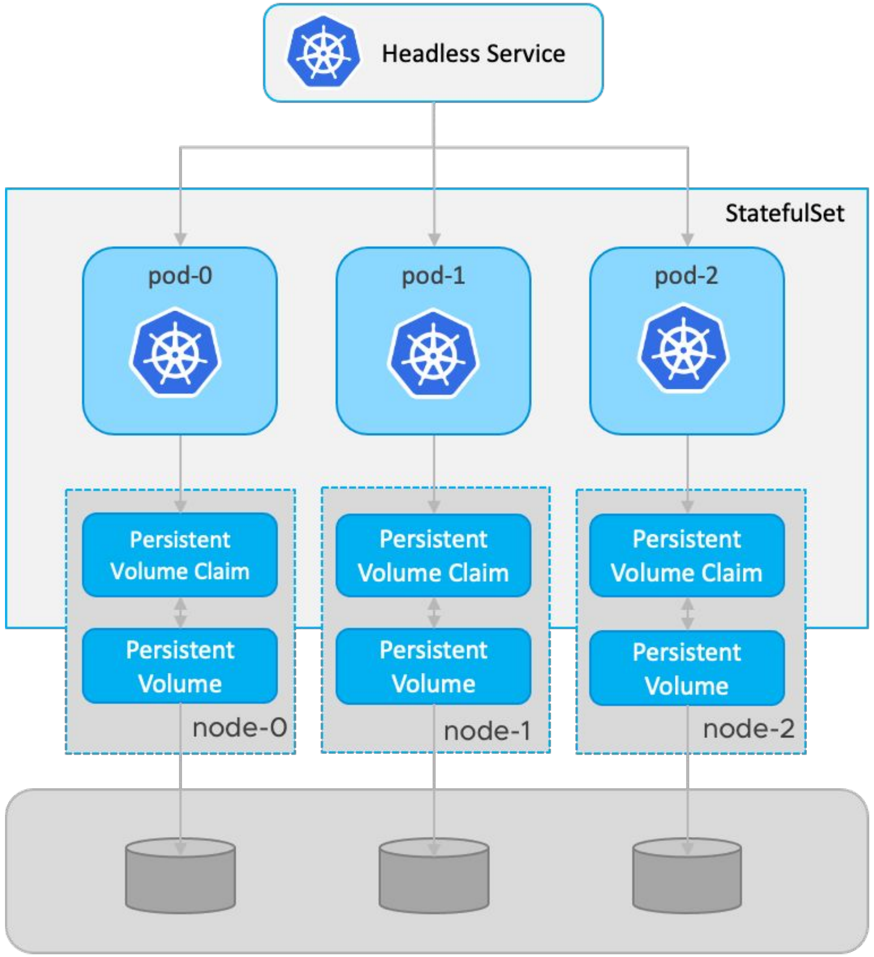
Non Graceful Shutdown: The How

- Return of a shutdown node
  - Users are required to manually remove the out of service taint after the pods have moved to a new node.
  - In case the taint is not removed from the node after it has returned, no new pods will be scheduled to the node.



Non Graceful Shutdown: The How

- Enabled the non graceful shutdown feature.
- Created a statefulset.
- Shutdown one node using `Shut Down Guest OS` from vSphere UI.
- Observed that after 5 mins, the pod changed to `Terminating` state.
- Taint the shutdown node using the command: `kubectl taint nodes <node-name> node.kubernetes.io.out-of-service=hardwarefailure:NoExecute
- The pod immediately failed over to a different healthy node without waiting for the 6 min detach timeout.



Demo

Next Steps

- Non-graceful node shutdown feature is targeting Beta in K8s v1.26.
- Alternatives
  - SafeDetach
    - Assumes CSI driver knows whether it is safe to force detach.
  - Node fencing
    - Monitors partitioned nodes and posts NodeFence CRD object.
  - CSI Force Detach
    - new CSI controller capability UNPUBLISH_FENCE and node capability FORCE_UNPUBLISH.
  - Podmon
    - Validate if host is still connected to storage and if there is IO; if not, fence and clean up.

```
// CSIDriverSpec is the specification of a CSIDriver.
type CSIDriverSpec struct {
  ...
  // +optional
  SafeDetach *bool
   }
```

```
- kind: ConfigMap
  apiVersion: v1
  metadata:
   name: fence-method-fence-rhevm-node1
   namespace: default
  data:
   method.properties: |
        agent_name=fence-rhevm
        namespace=default
        ip=ovirt.com  # address to the rhevm management
        username=admin@internal
        password-script=/usr/sbin/fetch_passwd
        ssl-insecure=true
        plug=vm-node1  # the vm name
        action=reboot
        ssl=true
        disable-http-filter=true
```

How to Get Involved

- Shoutouts
  - Ashutosh Kumar (sonasingh46), David Porter, Derek Carr (derekwaynecarr), Hemant Kumar (gnufied), Jing Xu (jingxu97), Michelle Au (msau42), Mrunal Patel, Tim Hockin (thockin), Xing Yang (xing-yang), Yassine Tijani (yastij)

- Further Readings
  - https://kubernetes.io/blog/2021/04/21/graceful-node-shutdown-beta/
  - https://kubernetes.io/blog/2022/05/20/kubernetes-1-24-non-graceful-node-shutdown-alpha/

- Get Involved
  - Kubernetes Storage SIG
  - Kubernetes Node SIG

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