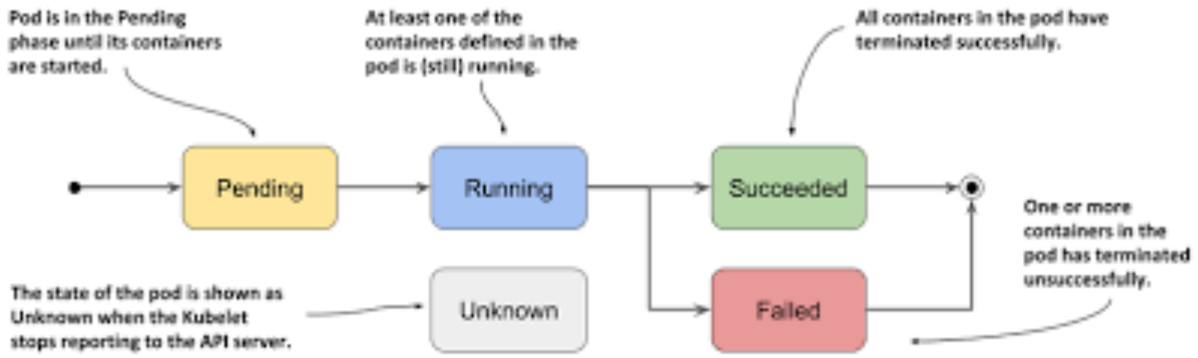
PodSupervisor
Are your pods sleeping?

Glenn West
gwest@redhat.com

Sep 30, 2022 10:31

Issue

- In a highly redundant OpenShift environment – Fail Fast is important to uptime of applications.
- Script “kills” pods on not-ready node and pod will be rescheduled as its settings dictate.



Pod is in the Pending phase until its containers are started.

At least one of the containers defined in the pod is (still) running.

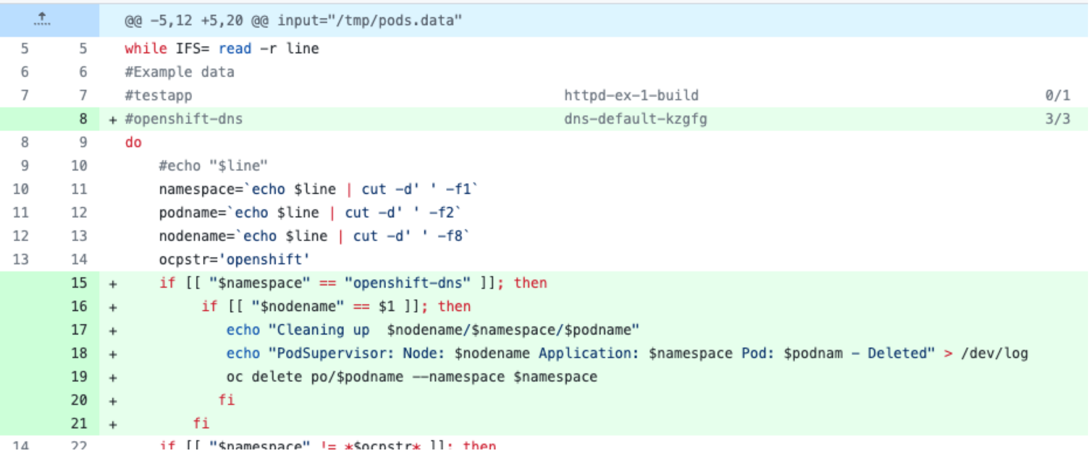
All containers in the pod have terminated successfully.

One or more containers in the pod has terminated unsuccessfully.

The state of the pod is shown as Unknown when the Kubelet stops reporting to the API server.

Solution

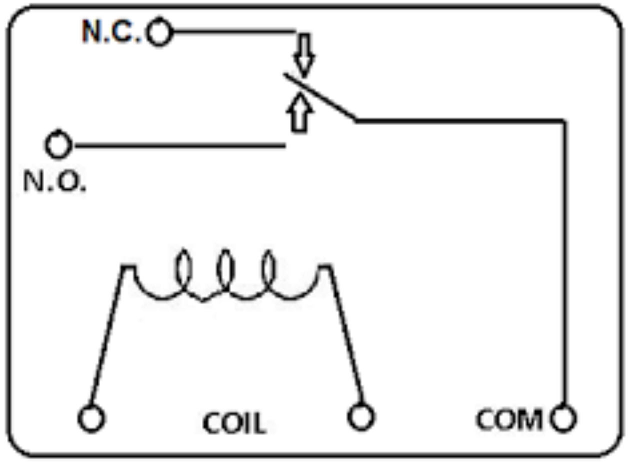
- A new monitor script that watches all nodes for not-ready then cleans up selected pods
- Pods may be selected list of names or wildcards
- Pods may be avoided by list of names or wildcard
- Script auto suspends during upgrades
- Default settings watch dns pods.



```
@@ -5,12 +5,20 @@ input="/tmp/pods.data"
  while IFS= read -r line
  #Example data
  #testapp                                                  httpd-ex-1-build                                           0/1
+ #openshift-dns                                            dns-default-kzgfg                                          3/3
  do
      #echo "$line"
      namespace=`echo $line | cut -d' ' -f1`
      podname=`echo $line | cut -d' ' -f2`
      nodename=`echo $line | cut -d' ' -f8`
      ocpstr='openshift'
+     if [[ "$namespace" == "openshift-dns" ]]; then
+          if [[ "$nodename" == $1 ]]; then
+             echo "Cleaning up  $nodename/$namespace/$podname"
+             echo "PodSupervisor: Node: $nodename Application: $namespace Pod: $podnam - Deleted" > /dev/log
+             oc delete po/$podname --namespace $namespace
+            fi
+        fi
      if [[ "$namespace" != *$ocpstr* ]]; then
```

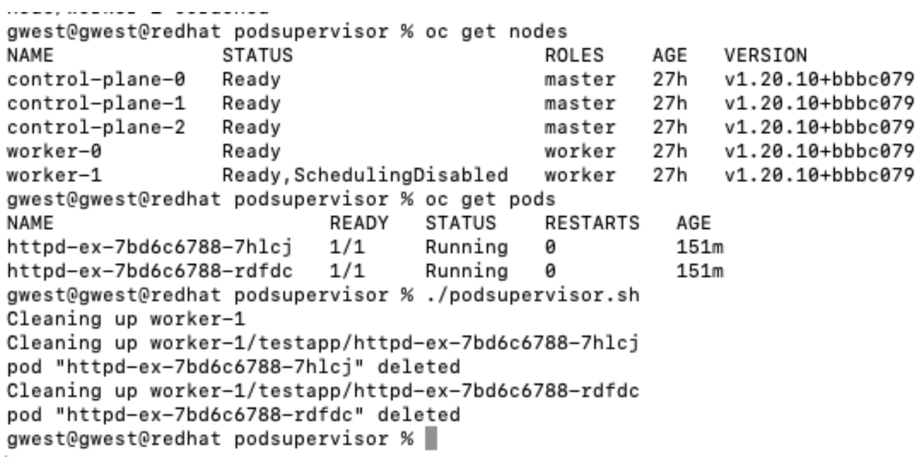
Additional Features

- Node flapping detection
  - Node must be not ready two scans
- On pod deletion, alerts are sent to OpenShift Alert Manager.



Example

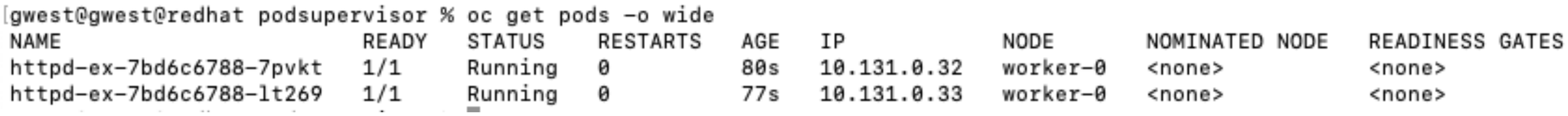
- Node is marked
- Script is run
- Two pods are deleted



```
node/worker-1 cordoned
gwest@gwest@redhat podsupervisor % oc get nodes
NAME              STATUS                     ROLES    AGE   VERSION
control-plane-0   Ready                      master   27h   v1.20.10+bbbc079
control-plane-1   Ready                      master   27h   v1.20.10+bbbc079
control-plane-2   Ready                      master   27h   v1.20.10+bbbc079
worker-0          Ready                      worker   27h   v1.20.10+bbbc079
worker-1          Ready,SchedulingDisabled   worker   27h   v1.20.10+bbbc079
gwest@gwest@redhat podsupervisor % oc get pods
NAME                        READY   STATUS    RESTARTS   AGE
httpd-ex-7bd6c6788-7hlcj    1/1     Running   0          151m
httpd-ex-7bd6c6788-rdfdc    1/1     Running   0          151m
gwest@gwest@redhat podsupervisor % ./podsupervisor.sh
Cleaning up worker-1
Cleaning up worker-1/testapp/httpd-ex-7bd6c6788-7hlcj
pod "httpd-ex-7bd6c6788-7hlcj" deleted
Cleaning up worker-1/testapp/httpd-ex-7bd6c6788-rdfdc
pod "httpd-ex-7bd6c6788-rdfdc" deleted
gwest@gwest@redhat podsupervisor % █
```

Result

- Pods have moved to worker-0.
- No system/openshift pods were touched or harmed.



```
[gwest@gwest@redhat podsupervisor % oc get pods -o wide
 NAME                         READY   STATUS    RESTARTS   AGE   IP            NODE       NOMINATED NODE   READINESS GATES
 httpd-ex-7bd6c6788-7pvkt     1/1     Running   0          80s   10.131.0.32   worker-0   <none>           <none>
 httpd-ex-7bd6c6788-lt269     1/1     Running   0          77s   10.131.0.33   worker-0   <none>           <none>
```

Source

- Upstream source:
- https://github.com/glennswest/podsupervisor