PodSupervisor
Are your pods sleeping?

Glenn West
gwest@redhat.com
Sep 27, 2022 17:28

Issue

- In a highly redundant OpenShift environment – Fail Fast is important to uptime of applications.
- OpenShift “kills” pods on not-ready node when node reboots. In addition, Nodes marked un-scheduable should also be avoided for better uptime.

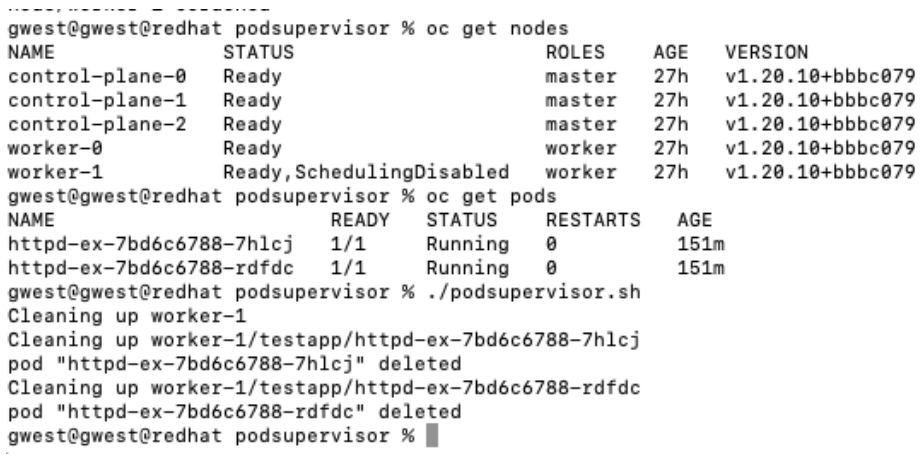
Solution

- A new monitor script that watches all nodes for not-ready or un-schedeable are cleaned up of all non-system pods
- Non-system pods do not have “openshift-” in there namespace/project name.
- DNS Pod on malfunctioning node is deleted and allowed to reschedule.

```
@@ -5,12 +5,20 @@ input="/tmp/pods.data"
5  5  while IFS= read -r line
6  6  #Example data
7  7  #testapp                                    httpd-ex-1-build                    0/1
   8  + #openshift-dns                            dns-default-kzgfg                   3/3
8  9  do
9  10     #echo "$line"
10 11     namespace=`echo $line | cut -d' ' -f1`
11 12     podname=`echo $line | cut -d' ' -f2`
12 13     nodename=`echo $line | cut -d' ' -f8`
13 14     ocpstr='openshift'
   15 +    if [[ "$namespace" == "openshift-dns" ]]; then
   16 +         if [[ "$nodename" == $1 ]]; then
   17 +            echo "Cleaning up  $nodename/$namespace/$podname"
   18 +            echo "PodSupervisor: Node: $nodename Application: $namespace Pod: $podnam - Deleted" > /dev/log
   19 +            oc delete po/$podname --namespace $namespace
   20 +           fi
   21 +       fi
14 22     if [[ "$namespace" != *$ocpstr* ]]; then
```

Example

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



Result

- Pods have moved to worker-0, as the worker-1 is un-scheduable.
- 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>
```

Things to add/Enhancements

- AvoidList for things like operators or other infra components
- State Transition Support – Wait till node is ready once before processing
- Event/Notification/Logging
- Migrate from cronjob to static pod
- Wait two cycles of not-ready to trigger

Source

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