

PodSupervisor

Are your pods sleeping?

Glenn West

gwest@redhat.com

Sep 28, 2022 9:32

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 unscheduable should also be avoided for better uptime.



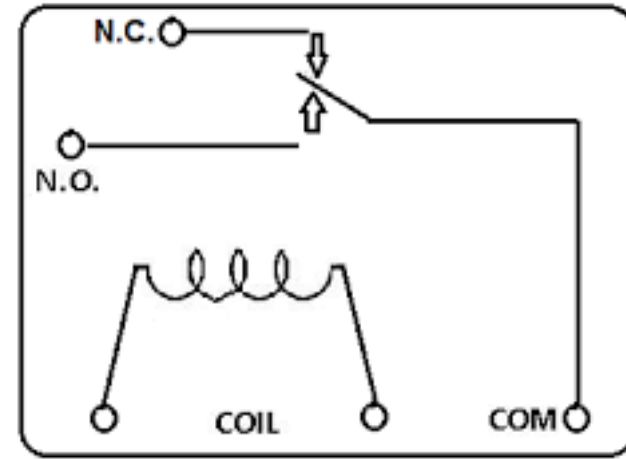
Solution

- 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      8      + #openshift-dns                        dns-default-kzgfg                                3/3
9      9      do
10     10      #echo "$line"
11     11      namespace=`echo $line | cut -d' ' -f1`
12     12      podname=`echo $line | cut -d' ' -f2`
13     13      nodename=`echo $line | cut -d' ' -f8`
14     14      ocpstr='openshift'
15     15      + if [[ "$namespace" == "openshift-dns" ]]; then
16     16      +         if [[ "$nodename" == $1 ]]; then
17     17      +             echo "Cleaning up $nodename/$namespace/$podname"
18     18      +             echo "PodSupervisor: Node: $nodename Application: $namespace Pod: $podnam - Deleted" > /dev/log
19     19      +             oc delete po/$podname --namespace $namespace
20     20      +         fi
21     21      +     fi
14     22      if [[ "$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 un-scheduable
- Script is run
- Two pods are deleted

```
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, 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>