Table of Contents

This lab does not work on OpenShift 4.0 Release 0.13.0. You can set up everything but the HPA can not get metrics. Therefore the auto scaling will not work.

LimitRanges and Horizontal Pod Autoscaler Lab

Goals

- Create a `LimitRange` object
- Create a Horizontal Pod Autoscaler (HPA)
- Test the HPA

You must have have metrics working in your cluster to create an HPA.

1. Create Limit Range

1. Create a new project:

```
oc new-project test-hpa
```

2. Deploy `hello-openshift` in the new project:

```
oc new-app openshift/hello-openshift:v3.10 -n test-hpa
oc expose svc hello-openshift
```

3. Create a `LimitRange` object with the following properties:

   Pod Limits

Min CPU	10m

Max CPU	100m
Min Memory	5Mi
Max Memory	750Mi

Container Limits

Min CPU	10m
Max CPU	100m
Min Memory	5Mi
Max Memory	750Mi
Default CPU	50m
Default Memory	100Mi

4. Confirm that your code looks similar to this:

```
echo '---
kind: LimitRange
apiVersion: v1
metadata:
  name: limits
  creationTimestamp:
spec:
  limits:
  - type: Pod
    max:
      cpu: 100m
      memory: 750Mi
    min:
      cpu: 10m
      memory: 5Mi
  - type: Container
    max:
      cpu: 100m
      memory: 750Mi
    min:
      cpu: 10m
      memory: 5Mi
    default:
      cpu: 50m
      memory: 100Mi
' | oc create -f - -n test-hpa
```

5. Create an HPA for the **`hello-openshift`** deployment to scale between one and five replicas and set it to scale up when the CPU utilization reaches 80%.

```
oc autoscale dc/hello-openshift --min 1 --max 5 --cpu-percent=80
```

6. List the status of the autoscaler:

```
oc get hpa hello-openshift -n test-hpa
```

Sample Output

```
NAME              REFERENCE                              TARGETS
MINPODS   MAXPODS   REPLICAS   AGE
hello-openshift   DeploymentConfig/hello-openshift   <unknown> / 80%   1
5         1           2h
```

It takes several minutes for the HPA to collect enough metrics to present a current status. If you see `<unknown>` in the `TARGETS` column, wait 30 seconds and repeat this step.

1. Review the autoscaler information:

```
oc describe hpa hello-openshift -n test-hpa
```

Sample Output

```
Name:                                                  hello-openshift
Namespace:                                             test-hpa
Labels:                                                <none>
Annotations:                                           <none>
CreationTimestamp:                                     Thu, 19 Apr 2018
18:12:19 +0000
Reference:
DeploymentConfig/hello-openshift
Metrics:                                               ( current / target
)
  resource cpu on pods  (as a percentage of request):  <unknown> / 80%
Min replicas:                                          1
Max replicas:                                          5
Conditions:
  Type           Status  Reason                   Message
  ----           ------  ------                   -------
  AbleToScale    True    SucceededGetScale        the HPA controller was
able to get the target's current scale
  ScalingActive  False   FailedGetResourceMetric  the HPA was unable to
compute the replica count: missing request for cpu on container hello-
openshift in pod test-hpa/hello-openshift-1-fc9l6
Events:
  Type     Reason                         Age              From
Message
  ----     ------                         ----             ----
-------
  Warning  FailedGetResourceMetric        2s (x3 over 1m)  horizontal-pod-
autoscaler  missing request for cpu on container hello-openshift in pod
test-hpa/hello-openshift-1-fc9l6
  Warning  FailedComputeMetricsReplicas  2s (x3 over 1m)  horizontal-pod-
autoscaler  failed to get cpu utilization: missing request for cpu on
container hello-openshift in pod test-hpa/hello-openshift-1-fc9l6
```

2. Fix the error displayed in the output above by rolling out the latest deployment configuration of **`hello-openshift`**.

   a. Redeploy the application to pick up the default request from the **`LimitRange`** object:

```
oc rollout latest hello-openshift -n test-hpa
```

3. In a separate window, create work for the pod and monitor the environment:

```
ROUTE=$(oc get route hello-openshift --template "{{ .spec.host }}")
for time in {1..15000}
  do
   echo time $time
   curl ${ROUTE}
  done
```

4. Run this command in a few windows concurrently to produce more work for the pods.
   - You need at least three parallel windows to generate enough load.
5. Examine the HPA information to see the effect of the workload.
6. Use the OpenShift Container Platform web console to see whether new pods are being created.

The HPA does not scale up or down instantly.

2. Clean Up Environment

1. Remove the HPA project:

```
oc delete project test-hpa
```

Build Version: 3af7325eec47b35dd9d60df5246e7b3ac0924020 : Last updated 2019-03-20 12:18:10 EDT