Autoscaling LAB (HPA)

03 January 2021   16:05

- Setup the lab environment, in my case used the minikube

```
$ minikube status
minikube
type: Control Plane
host: Running
kubelet: Running
apiserver: Running
kubeconfig: Configured
timeToStop: Nonexistent

$ minikube addons enable metrics-server
```

- Create a deployment as follow

```
#cat testdep.yml
apiVersion: apps/v1
kind: Deployment
metadata:
  name: php-apache
spec:
  selector:
    matchLabels:
      app: php-apache
  replicas: 1
  template:
    metadata:
      labels:
        app: php-apache
    spec:
      containers:
      - name: php-apache
        image: gcr.io/google_containers/hpa-example
        resources:
          requests:
            cpu: "500m"
            memory: "500M"
        ports:
        - containerPort: 80
```

- Expose the deployment

```
$ kubectl expose deployment php-apache --name=php-apache –port=80
service/php-apache exposed
```

- Test the deployment is serviing using another test pod

```
$ kubectl run -i --tty service-test --image=busybox /bin/sh
wget -q -O- http://php-apache.default.svc.cluster.local
OK!/ #

kubectl exec -i --tty service-test  /bin/sh
kubectl exec [POD] [COMMAND] is DEPRECATED and will be removed in a future version. Use kubectl exec [POD] -- [COMMAND] instead.
/ # wget -q -O- http://php-apache.default.svc.cluster.local
OK!/ #
```
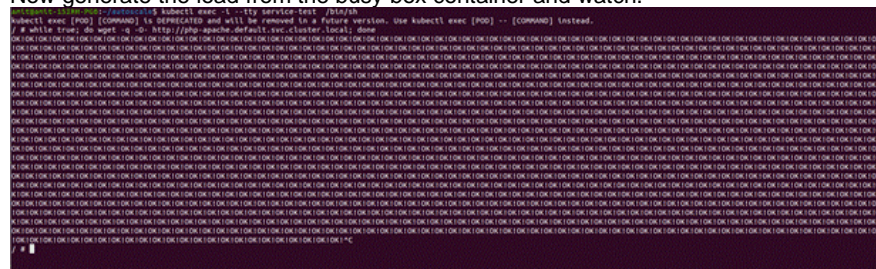
- Define the HPA (Horizontal pod autoscale)

```
$ kubectl get hpa
No resources found in default namespace.

$ kubectl autoscale deployment php-apache --cpu-percent=50 --min=1 --max=3
horizontalpodautoscaler.autoscaling/php-apache autoscaled

$ kubectl get hpa
NAME         REFERENCE               TARGETS     MINPODS  MAXPODS  REPLICAS  AGE
php-apache   Deployment/php-apache   0/50%  1       3        0         1s
```

- sdfs
- Now generate the load from the busy box container and watch.

Parallely on the second terminal observe pods and HPA output. After 2minutes you will notice more pods being launched

```
amit@amit-15IRH-PG0:~/autoscale$ kubectl top pods
NAME                          CPU(cores)   MEMORY(bytes)
php-apache-59f68b8bcc-v5dcp   735m         12Mi
service-test                  8m           1Mi
amit@amit-15IRH-PG0:~/autoscale$ kubectl top pods
NAME                          CPU(cores)   MEMORY(bytes)
php-apache-59f68b8bcc-v5dcp   735m         12Mi
service-test                  8m           1Mi
amit@amit-15IRH-PG0:~/autoscale$ kubectl get hpa
NAME         REFERENCE               TARGETS    MINPODS   MAXPODS   REPLICAS   AGE
php-apache   Deployment/php-apache   147%/50%   1         3         3          11m
amit@amit-15IRH-PG0:~/autoscale$
```

```
amit@amit-15IRH-PG0:~/autoscale$ kubectl get all
NAME                               READY   STATUS    RESTARTS   AGE
pod/php-apache-59f68b8bcc-9wzbs    1/1     Running   0          5m25s
pod/php-apache-59f68b8bcc-jm2mv    0/1     Pending   0          5m25s
pod/php-apache-59f68b8bcc-v5dcp    1/1     Running   0          29m
pod/service-test                   1/1     Running   1          27m

NAME                 TYPE        CLUSTER-IP     EXTERNAL-IP   PORT(S)   AGE
service/kubernetes   ClusterIP   10.96.0.1      <none>        443/TCP   75m
service/php-apache   ClusterIP   10.98.84.229   <none>        80/TCP    18m

NAME                          READY   UP-TO-DATE   AVAILABLE   AGE
deployment.apps/php-apache    2/3     3            2           29m

NAME                                    DESIRED   CURRENT   READY   AGE
replicaset.apps/php-apache-59f68b8bcc   3         3         2       29m

NAME                                                        REFERENCE               TARGETS   MINPODS   MAXPODS   REPLICAS   AGE
horizontalpodautoscaler.autoscaling/php-apache              Deployment/php-apache   0%/50%    1         3         3          16m
amit@amit-15IRH-PG0:~/autoscale$
```

Now press ctrl + C to stop more load generation in busy box container and after few minutes you will notice additional pods w ill disapper

```
amit@amit-15IRH-PG0:~/autoscale$ kubectl get pods
NAME                          READY   STATUS    RESTARTS   AGE
php-apache-59f68b8bcc-v5dcp   1/1     Running   0          30m
service-test                  1/1     Running   1          29m
amit@amit-15IRH-PG0:~/autoscale$ kubectl get hpa
NAME         REFERENCE               TARGETS   MINPODS   MAXPODS   REPLICAS   AGE
php-apache   Deployment/php-apache   0%/50%    1         3         3          17m
amit@amit-15IRH-PG0:~/autoscale$ kubectl get hpa
NAME         REFERENCE               TARGETS   MINPODS   MAXPODS   REPLICAS   AGE
php-apache   Deployment/php-apache   0%/50%    1         3         1          17m
amit@amit-15IRH-PG0:~/autoscale$
```