Scaling Applications

» Autoscaling Deployments

Overview

Verify resource usage of all pods in the environment with the `kubectl top` command.

```
kubectl top pods --all-namespaces
```

Create a basic nginx deployment with one replica. The container in this deployment requires 100m (.1 vcpu).

```
cat > autoscale-nginx.yaml <<EOF
apiVersion: extensions/v1beta1
kind: Deployment
metadata:
  name: nginx
spec:
  replicas: 1
  template:
    metadata:
      labels:
        app: nginx
    spec:
      containers:
      - name: nginx
        image: nginx:1.13
        imagePullPolicy: IfNotPresent
        ports:
        - containerPort: 80
        resources:
          requests:
            cpu: "100m"
  EOF
```

Create the deployment.

```
kubectl create -f autoscale-nginx.yaml
```

Autoscale

Apply a Horizonal Pod Autoscaler to the nginx deployment.

```
kubectl autoscale deployment nginx --min 1 --max 10 --cpu-percent 70
```

Get the list of pods created by the deployment.

```
kubectl get pods -l app=nginx -o wide
```

Save the name of the single pod in an environment variable.

```
export POD=$(kubectl get pods -l app=nginx -o jsonpath={.items[0].metadata.name})
echo $POD
```

Execute a command in the pod to produce cpu load.

```
kubectl exec -it $POD -- /bin/dd if=/dev/zero of=/dev/null
```

In a separate terminal window: Watch the HPA in real-time.

```
kubectl get hpa -w
```

In a separate terminal window: Watch the deployment in real-time.

```
kubectl get deployment nginx -w
```

Type `Ctrl+c` in the first terminal window to cancel the `exec` command that was used to produce cpu load.

However, the command is still running on the pod! Keep watching the HPA window to verify.

Remove the pod priducing load and watch the deployment and HPA update accordingly.

```
kubectl delete pod $POD
```

Get the list of pods currently in the deployment.

```
kubectl get pods -l app=nginx -o wide
```

Wait for a while and the deployment should scale back down to a single pod.

Clean Up

You can close the additional terminal windows and delete the deployment and HPA when finished.

```
kubectl delete deployment,hpa nginx
```

View any remaining resources.

```
kubectl get all
```