Kubernetes Autoscaling on AWS

Kirk S. Kalvar

10/17/2016

# setup times

Provision 2 node cluster, including master: 10 minutes

Scaling Up: 20 minutes total time to scale to 10 pods and 4 nodes

Scaling Down: 10 minutes total time for 1 pod and 2 nodes

# create deployment

kubectl run php-apache --image=gcr.io/google\_containers/hpa-example

# create service

kubectl expose deployment php-apache --port=80

# show pods running

kubectl get pods --output wide

# show deployment

kubectl get deployment

# show service

kubectl get service

# autoscale deployment

kubectl autoscale deployment php-apache --cpu-percent=50 --min=1 --max=10

# show hpa

kubectl get hpa

# Modify AWS Autoscaling Group

Max: 5

Scaling Policies/Add policy

Name: Scale Up

Execute policy when: Kube Cluster >= 50 Percent

breaches the alarm threshold: CPUUtilization >= 50 for 300 seconds

Take the action: Add 1 instances when 50 <= CPUUtilization < +infinity

Instances need: 300 seconds to warm up after each step

Name: Scale Down

Execute policy when: Kube Cluster <= 10 Percent

breaches the alarm threshold: CPUUtilization <= 10 for 300 seconds

Take the action: Remove 1 instances when 10 >= CPUUtilization > -infinity

# console

https://<k8s-master public ip>/api/v1/proxy/namespaces/kube-system/services/kubernetes-dashboard/#/deployment?namespace=kube-system

userid: admin

password: <see .kube/config>

# create load generator

kubectl run -i --tty load-generator --image=busybox /bin/sh

# test load test

wget -q -O- http://php-apache.default.svc.cluster.local

# run load test

while true; do wget -q -O- http://php-apache.default.svc.cluster.local; done

# stop load test

kubectl delete deployment load-generator

# stop autoscaling application

kubectl delete hpa,service,deployment php-apache