**03. Enable HPA, Load Test, Verify and Clean-Up**

--- Reference - <https://github.com/stacksimplify/aws-eks-kubernetes-masterclass/tree/master/15-EKS-HPA-Horizontal-Pod-Autoscaler>

**Create a Horizontal Pod Autoscaler resource for the "hpa-demo-deployment"**

* This command creates an autoscaler that targets 50 percent CPU utilization for the deployment, with a minimum of one pod and a maximum of ten pods.
* When the average CPU load is below 50 percent, the autoscaler tries to reduce the number of pods in the deployment, to a minimum of one.
* When the load is greater than 50 percent, the autoscaler tries to increase the number of pods in the deployment, up to a maximum of ten

**# Template**

--- **kubectl autoscale deployment <deployment-name> --cpu-percent=50 --min=1 --max=10**

**# Replace**

--- **kubectl autoscale deployment hpa-demo-deployment --cpu-percent=50 --min=1 --max=10**

**# Describe HPA**

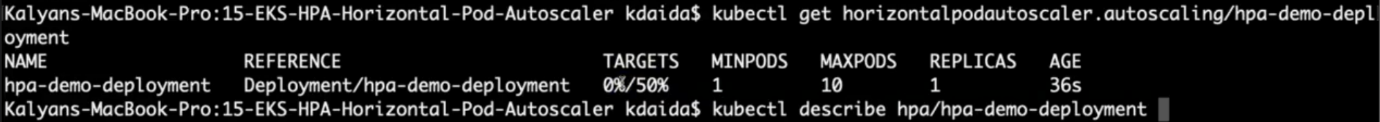
--- **kubectl describe hpa/hpa-demo-deployment**



--- **note** - **hpa-demo-deployment** is a previous deployment name.

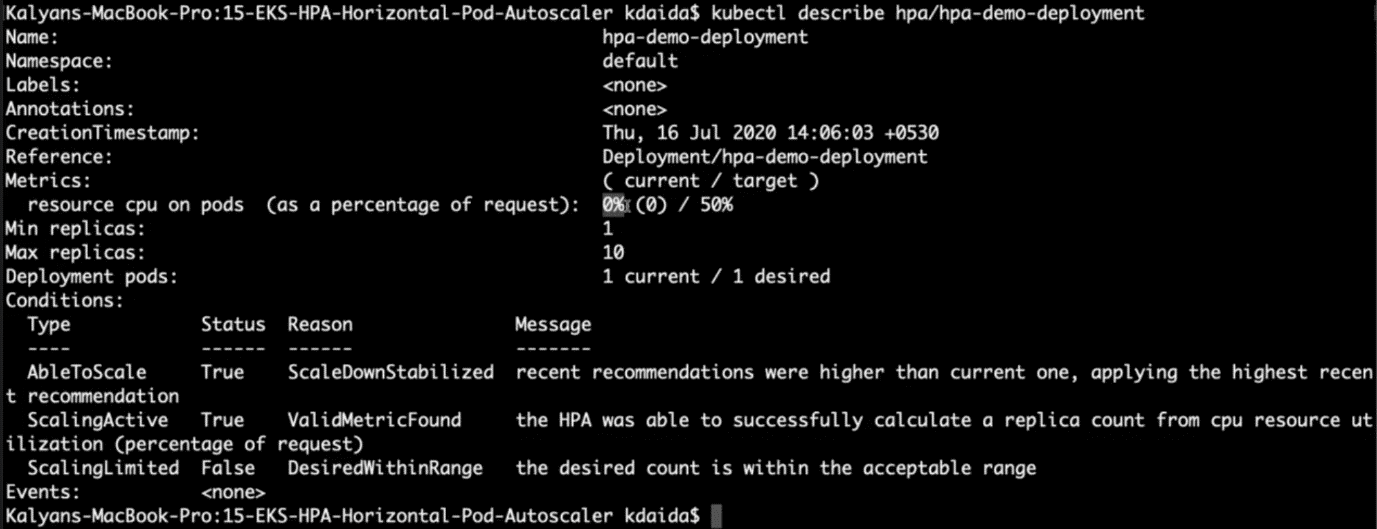
**# List HPA**

--- **kubectl get horizontalpodautoscaler.autoscaling/hpa-demo-deployment**



--- **note** – out of 50%, it is using only 0%.

--- **kubectl describe hpa/hpa-demo-deployment**



**Create the load & Verify how HPA is working**

**# Generate Load**

--- **kubectl run --generator=run-pod/v1 apache-bench -i --tty --rm --image=httpd -- ab -n 500000 -c 1000 http://hpa-demo-service-nginx.default.svc.cluster.local/**

**# List all HPA**

--- **kubectl get hpa**

**# List specific HPA**

--- **kubectl get hpa hpa-demo-deployment**

**# Describe HPA**

--- **kubectl describe hpa/hpa-demo-deployment**

**# List Pods**

--- **kubectl get pods**

**Cooldown / Scaledown**

* Default cooldown period is 5 minutes.
* Once CPU utilization of pods is less than 50%, it will start terminating pods and will reach to minimum 1 pod as configured.

**Clean-Up**

**# Delete HPA**

--- **kubectl delete hpa hpa-demo-deployment**

**# Delete Deployment & Service**

--- **kubectl delete -f kube-manifests/**

**Imperative vs Declarative for HPA**

--- From Kubernetes v1.18 onwards, we have a declarative way of defining HPA policies using behaviour object in yaml.

--- for configurable scaling behaviour

* Starting from v1.18 the v2beta2 API allows scaling behaviour to be configured through the HPA behaviour field.
* Behaviours are specified separately for scaling up and down in scaleUp or scaleDown section under the behaviour field

behavior:

  scaleDown:

    stabilizationWindowSeconds: 300

    policies:

    - type: Percent

      value: 100

      periodSeconds: 15

  scaleUp:

    stabilizationWindowSeconds: 0

    policies:

    - type: Percent

      value: 100

      periodSeconds: 15

    - type: Pods

      value: 4

      periodSeconds: 15

    selectPolicy: Max

--- Reference: Select V1.18 from top right corner on Kubernetes website for V1.18 documentation

* <https://kubernetes.io/docs/tasks/run-application/horizontal-pod-autoscale/>

**References**

--- Metrics Server Releases - <https://github.com/kubernetes-sigs/metrics-server/releases>

--- Horizontal Pod Autoscaling - Scale based on many type of metrics - <https://v1-16.docs.kubernetes.io/docs/tasks/run-application/horizontal-pod-autoscale-walkthrough/>