

Create folder "hpa"

Add deployment.yml & hpa.yml

Service is also include in deployment.yml

set limit of resources for pods

Add resources in containers

`requests` and `limits` are used to define the resource requirements for a container.

```
resources:
  requests:
    cpu: 200m
    memory: 100Mi
  limits:
    cpu: 400m
    memory: 200Mi
```

Create "hpa.yml"

```
apiVersion: autoscaling/v2
kind: HorizontalPodAutoscaler
metadata:
  name: my-hpa
spec:
  scaleTargetRef:      . . . . .target resource that the HPA is scaling.
    apiVersion: apps/v1
    kind: Deployment
    name: my-deployment
  minReplicas: 1
  maxReplicas: 10
  metrics:
  - type: Resource
    resource:
      name: cpu
      target:
        type: Utilization
        averageUtilization: 30
```

Apply both file

`kubectl get pod` → for check pod created by deployment, replica.

`kubectl get hpa` → to check hpa

`kubectl get deploy` → to check deployment

`kubectl apply -f https://github.com/kubernetes-sigs/metrics-server/releases/latest/download/components.yaml`
→ to install metric server

kubectl get pod -n kube-system → to check matric server created or not

kubectl exec -it my-pod -c my-container -- /bin/sh → to grt inside the container

kubectl exec -it my-deployment-7f57fccdc-p5p25 -c my-container -- /bin/sh

cat /etc/os-release → to check os of container

apt-get update

apt install stress → install stress command

stress --cpu 50 --timeout 600s & → increase stress

kubectl get pod → check pod are increses