

Creating the manifests

Hardcoded replicas

Lets try to create a deployment with **hardcoded** set of replicas

```
apiVersion: apps/v1
kind: Deployment
metadata:
  name: cpu-deployment
spec:
  replicas: 2
  selector:
    matchLabels:
      app: cpu-app
  template:
    metadata:
      labels:
        app: cpu-app
    spec:
      containers:
        - name: cpu-app
          image: 100xdevs/week-28:latest
          ports:
            - containerPort: 3000
```



- Create a service

```
apiVersion: v1
kind: Service
metadata:
  name: cpu-service
spec:
  selector:
```



- protocol: TCP

```
port: 80
targetPort: 3000
type: LoadBalancer
```

With a horizontal pod accelerator

- Add HPA manifest

```
apiVersion: autoscaling/v2
kind: HorizontalPodAutoscaler
metadata:
  name: cpu-hpa
spec:
  scaleTargetRef:
    apiVersion: apps/v1
    kind: Deployment
    name: cpu-deployment
  minReplicas: 2
  maxReplicas: 5
  metrics:
  - type: Resource
    resource:
      name: cpu
      target:
        type: Utilization
        averageUtilization: 50
```



- Apply all three manifests

```
kubectl apply -f service.yml
kubectl apply -f deployment.yml
kubectl apply -f hpa.yml
```



You can scale up/down based on multiple metrics.

If either of the metrics goes above the threshold, we scale up

If all the metrics go below the threshold, we scale down

