Autoscaling based on CPU and Memory

In this example we will create a deployment and autoscale it based on CPU and Memory usage. The deployment will be behind a service and the exposed through an Ingress.

Explanation

The HPA will scale up whenever the CPU usage is above 50%, as specified at line 18 of `hpa.yaml`. The 50% is based on the ratio between the actual CPU utilization and the requested CPU specified at line 22 of `deployment.yaml`. When the CPU is back below 50% the HPA starts to scale down according to the settings specified at the behaviour section of the HPA configuration file. According to the `stabilizationWindowSeconds` parameter the HPA will keep considering the past statistics so greater the value the more you'll have to wait for the HPA to scale down.

Usage

Prerequisites

- Minikube
- Kubernetes CLI (kubectl)
- Enabling Ingress addon in Minikube

```
minikube addons enable ingress
```

- Enabling Metrics Server in Minikube

```
minikube addons enable metrics-server
```

wait a couple of minutes for the metrics server to be up.

Create the deployment

```
kubectl apply -f manifests
```

Then check the status of the deployment with the usual commands.

Increase the load

First start watching the deployment status:

```
kubectl get hpa php-apache --watch
```

Then using the bash script included increase the load. This script will create a temporary pod that will request the service in a loop.

```
./load.sh
```

Run it in a separate terminal, if the load is too high for the settings specificied in the configuration file then run more instances of the script. Then terminate the load script(s) and wait for the HPA to scale down.