### 2.2.11 HPA

2.2.11.1 什么是HPA

HPA（Horizontal Pod Autoscaler，水平Pod自动伸缩器）可根据观察到的CPU、内存使用率或自定义度量标准来自动扩展或缩容Pod的数量。HPA不适用于无法缩放的对象，比如DaemonSet。

HPA控制器会定期调整RC或Deployment的副本数，以使观察到的平均CPU利用率与用户指定的目标相匹配。

HPA需要metrics-server（项目地址：https://github.com/kubernetes-incubator/metrics-server）获取度量指标，由于在高可用集群安装中已经安装了metrics-server，所以本节的实践部分无须再次安装。

2.2.11.2 HPA实践

在生产环境中，总会有一些意想不到的事情发生，比如公司网站流量突然升高，此时之前创建的Pod已不足以撑住所有的访问，而运维人员也不可能24小时守着业务服务，这时就可以通过配置HPA，实现负载过高的情况下自动扩容Pod副本数以分摊高并发的流量，当流量恢复正常后，HPA会自动缩减Pod的数量。

本节将测试实现一个Web服务器的自动伸缩特性，具体步骤如下：

首先启动一个Nginx服务：

[root@k8s-master01 ~]#kubectl run nginx-server --requests=cpu=10m --image=nginx --port=80

service/php-apache created

deployment.apps/php-apache created

临时开启nginx-server的端口，实际使用时需要定义service：

kubectl expose deployment nginx-server --port=80

使用kubectl autoscale创建HPA：

[root@k8s-master01 ~]# kubectl autoscale deployment nginx-server --cpu-percent=10 --min=1 --max=10

此HPA将根据CPU的使用率自动增加和减少副本数量，上述设置的是CPU使用率超过10%（--cpu-percent参数指定）即会增加Pod的数量，以保持所有Pod的平均CPU利用率为10%，允许最大的Pod数量为10（--max），最少的Pod数为1（--min）。

查看当前HPA状态，因为未对其发送任何请求，所以当前CPU使用率为0%：

[root@k8s-master01 metric-server]# kubectl get hpa

NAME REFERENCE TARGETS MINPODS MAXPODS REPLICAS AGE

nginx-server Deployment/nginx-server 0%/10% 1 10 1 5m

查看当前Nginx的Service地址：

[root@k8s-master01 ~]# kubectl get service -n default

NAME TYPE CLUSTER-IP EXTERNAL-IP PORT(S) AGE

kubernetes ClusterIP 10.96.0.1 <none> 443/TCP 1d

nginx-server ClusterIP 10.108.160.23 <none> 80/TCP 5m

增加负载：

[root@k8s-master01 ~]# while true; do wget -q -O- http://10.108.160.23 > /dev/null; done

1分钟左右再次查看HPA：

[root@k8s-master01 metric-server]# kubectl get hpa

NAME REFERENCE TARGETS MINPODS MAXPODS REPLICAS AGE

nginx-server Deployment/nginx-server 540%/10% 1 10 1 15m

再次查看Pod，可以看到nginx-server的Pod已经在扩容阶段：

[root@k8s-master01 metric-server]# kubectl get po

NAME READY STATUS RESTARTS AGE

nginx-server-589c8db585-5cbxl 0/1 ContainerCreating 0 <invalid>

nginx-server-589c8db585-7whl8 1/1 Running 0 <invalid>

nginx-server-589c8db585-cv4hs 1/1 Running 0 <invalid>

nginx-server-589c8db585-m5dn6 0/1 ContainerCreating 0 <invalid>

nginx-server-589c8db585-sxbfm 1/1 Running 0 19m

nginx-server-589c8db585-xbctd 0/1 ContainerCreating 0 <invalid>

nginx-server-589c8db585-xffs9 1/1 Running 0 <invalid>

nginx-server-589c8db585-xlb8s 0/1 ContainerCreating 0 <invalid>

在增加负荷的终端，按Ctrl+C键终止访问。

停止1分钟后再次查看HPA和deployment，此时副本已经恢复为1：

[root@k8s-master01 metric-server]# kubectl get hpa

NAME REFERENCE TARGETS MINPODS MAXPODS REPLICAS AGE

nginx-server Deployment/nginx-server 0%/10% 1 10 10 20m