

Contents

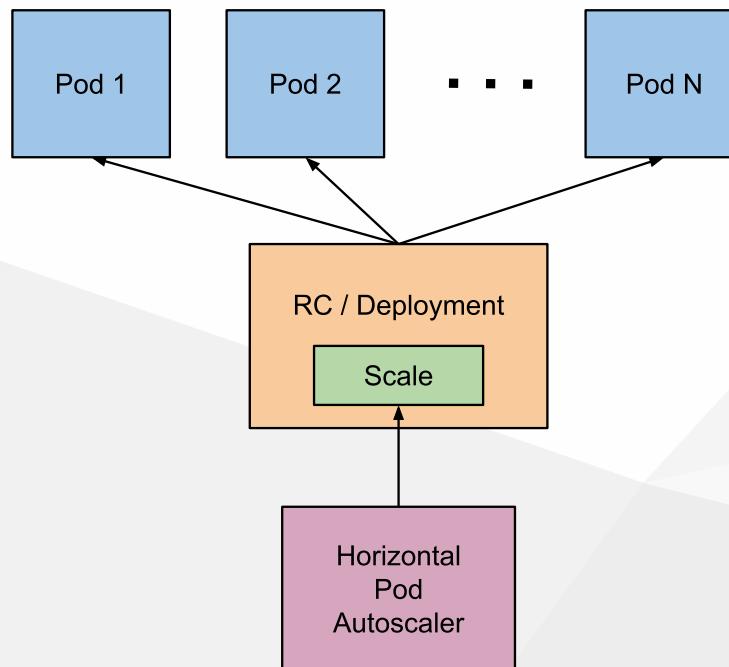
- Horizontal Pod Autoscaler
- Kubernetes Metrics Server



Horizontal Pod Autoscaler

HorizontalPodAutoscaler는 워크로드 리소스(e.g. Deployment 또는 StatefulSet)를 자동으로 업데이트하며, 워크로드의 크기(파드의 개수)를 수요에 맞게 자동으로 스케일링하는 것을 목표로 합니다.

수평 스케일링은 부하 증가에 대해 Pod의 수를 증가시키는 것을 말합니다. 이는 수직 스케일링(해당 워크로드를 위해 이미 실행 중인 파드에 더 많은 자원(예: 메모리 또는 CPU)을 할당하는 것)과는 다릅니다.



Horizontal Pod Autoscaler

Horizontal Pod Autoscaler controller는 아래와 같이 원하는(desired) 기준 값과 측정된(current) 값의 비율로 그 값을 계산합니다.

```
desiredReplicas = ceil[currentReplicas * ( currentMetricValue / desiredMetricValue )]
```

예를 들어 현재 메트릭 값이 200m이고 원하는 값이 100m인 경우 $200.0 / 100.0 == 2.0$ 이므로 복제본 수가 두 배가 됩니다.
만약 현재 값이 50m 이면, $50.0 / 100.0 == 0.5$ 이므로 복제본 수를 반으로 줄이게 됩니다.

보다 자세한 알고리즘 세부정보는 아래 링크를 참조 바랍니다.

 [Horizontal Pod Autoscaling - 알고리즘 세부정보](#)

Horizontal Pod Autoscaler

HorizontalPodAutoscaler는 다음과 같이 정의합니다.

```
apiVersion: autoscaling/v2beta2
kind: HorizontalPodAutoscaler
metadata:
  name: myphp-hpa
  namespace: default
spec:
  maxReplicas: 10
  minReplicas: 1
  scaleTargetRef:
    apiVersion: apps/v1
    kind: Deployment
    name: myphp
  metrics:
  - type: Resource
    resource:
      name: cpu
    target:
      type: Utilization
      averageUtilization: 50
```

Pod의 평균 CPU 사용율(50%)을 기준으로 1(minReplicas) 부터 10(maxReplicas)까지 Replicas의 수를 증/감 시킴.

위와같이 정의한 HorizontalPodAutoscaler는 조회 명령어에서는 다음과 같이 표시됩니다.

```
root@master:/# kubectl get hpa -o wide
NAME        REFERENCE          TARGETS      MINPODS   MAXPODS   REPLICAS     AGE
myphp-hpa   Deployment/myphp  0%/50%       1          10         1           5m48s
```

Kubernetes Metrics Server

HorizontalPodAutoscaler를 사용하려면 먼저 [Kubernetes Metrics Server](#) 를 설치해야 합니다.
사용율 측정을 위해서 필요한 요소입니다.

설치해야할 Object들에 대한 자세한 설명은 이 교재에서는 다루지 않겠습니다.
위의 링크에 자세히 설명된 내용을 참조합니다.

설치는 다음과 같이 진행하면 됩니다.

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

설치 후에는 다음과 같이 각 Node의 정보를 확인할 수 있습니다.

```
$ kubectl top node
NAME      CPU(cores)   CPU%    MEMORY(bytes)   MEMORY%
master    160m         8%     2553Mi        66%
node1     46m          2%     2289Mi        59%
node2     42m          2%     2141Mi        55%
```

Summary

- Horizontal Pod Autoscaler
- Kubernetes Metrics Server