6. Kubernetes 안정성 강화 방법

02 HPA 소개 및 Metrics Server 설치

소개 및 실습 내용

순서

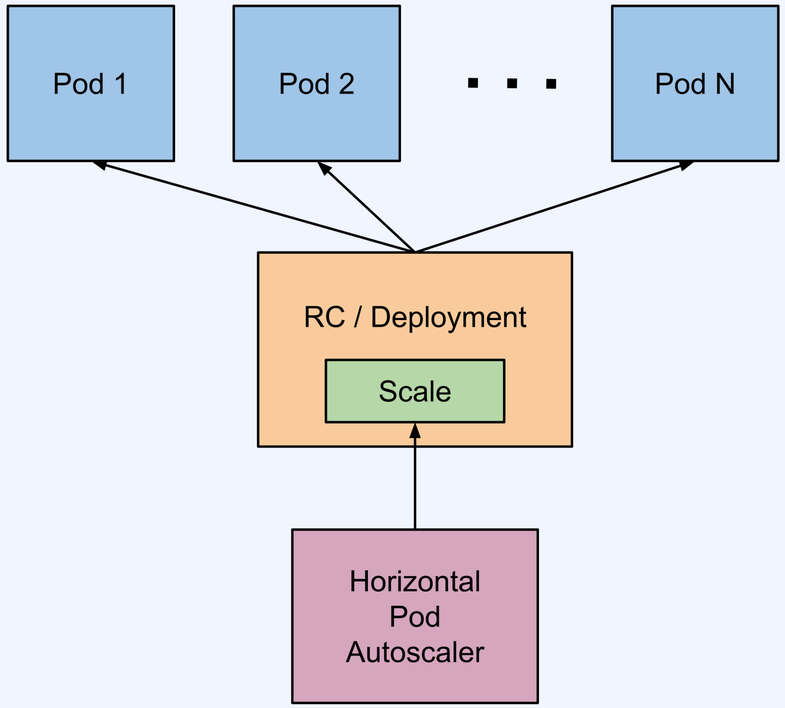
1. HPA 소개
2. Metrics Server 소개
3. Metrics Server 설치

실습 예제코드 경로

Chapter06 > Ch06_02-metrics-server

1. HPA 소개

- 워크로드 리소스(Deployment, StatefulSet)를 자동으로 업데이트
- 워크로드의 크기를 수요에 맞게 자동으로 스케일링



출처 : https://kubernetes.io/ko/docs/tasks/run-application/horizontal-pod-autoscale/

1. HPA – Scale out 수행 기준

```
$ kubectl get deploy test-deploy –o yaml
.. 중략 ..
 resources:
       limits:
        cpu: 500m
       requests:
        cpu: 200m
.. 중략 ..
```

* request > cpu 기준 50%가 넘으면 scale out 수행
200m x 0.5(50%) = 100m (1Pod)
(CPU 사용률 1분 간격 체크)

```
$ kubectl get hpa
NAME          REFERENCE                TARGETS  MINPODS  MAXPODS  REPLICAS  AGE
test-deploy   Deployment/test-deploy   0%/50%   1        10       1         5m
```

1. HPA – Replicas 산정 기준

```
$ kubectl get deploy test-deploy –o yaml
.. 중략 ..
 resources:
       limits:
        cpu: 500m
       requests:
        cpu: 200m
.. 중략 ..
```

*** request > cpu 기준 50%가 넘으면 scale out 수행**
200m x 0.5(50%) = 100m (1Pod)
200m x 2.5(250%) = 500m (5Pods - REPLICAS)
(Pod 스케쥴링 sync , 15초 간격 수행)

```
$ kubectl get hpa
NAME          REFERENCE                TARGETS     MINPODS   MAXPODS   REPLICAS   AGE
test-deploy   Deployment/test-deploy   250%/50%    1         10        5          8m
```

1. HPA – Scale in (Downscale) 산정 기준

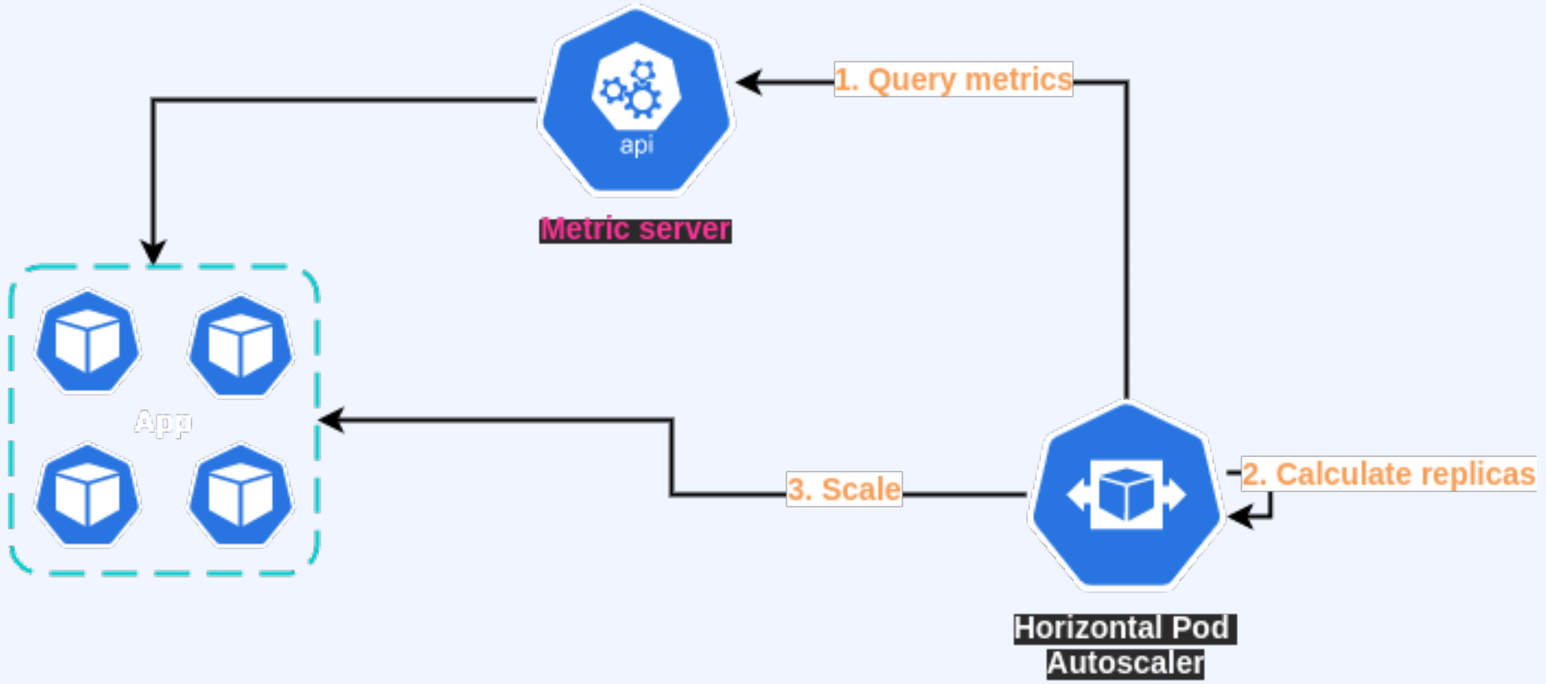
```
$ kubectl get deploy test-deploy –o yaml
.. 중략 ..
 resources:
       limits:
        cpu: 500m
       requests:
        cpu: 200m
.. 중략 ..
```

*** request > cpu 기준 50%가 이하가 되면 scale in 수행**
200m x 0.5(50%) = 100m (1Pod)
200m x 1(100%) = 200m (5에서 2로 수량감소)
(Pod downscale, 5분 간격 수행)

```
$ kubectl get hpa
NAME          REFERENCE                TARGETS     MINPODS   MAXPODS   REPLICAS   AGE
test-deploy   Deployment/test-deploy   100%/50%    1         10        2          19m
```

2. Metrics Server 소개

- Kubernetes POD Autoscale을 위한 기본 Metrics 처리 시스템
- Metrics을 Query해 생성/회수 할 Replicas를 산정한뒤 Scale



출처 : https://lzomedia.com/blog/horizontal-pod-autoscaler-on-eks-cluster/

3. Metrics Server 설치

Metrics Server 설치 Manifest 경로

- Chapter06 > Ch06_02-metrics-server

```
$ kubectl apply –f  metrics-server.yaml
```

설치 확인

```
$ kubectl get deployment metrics-server -n kube-system
```