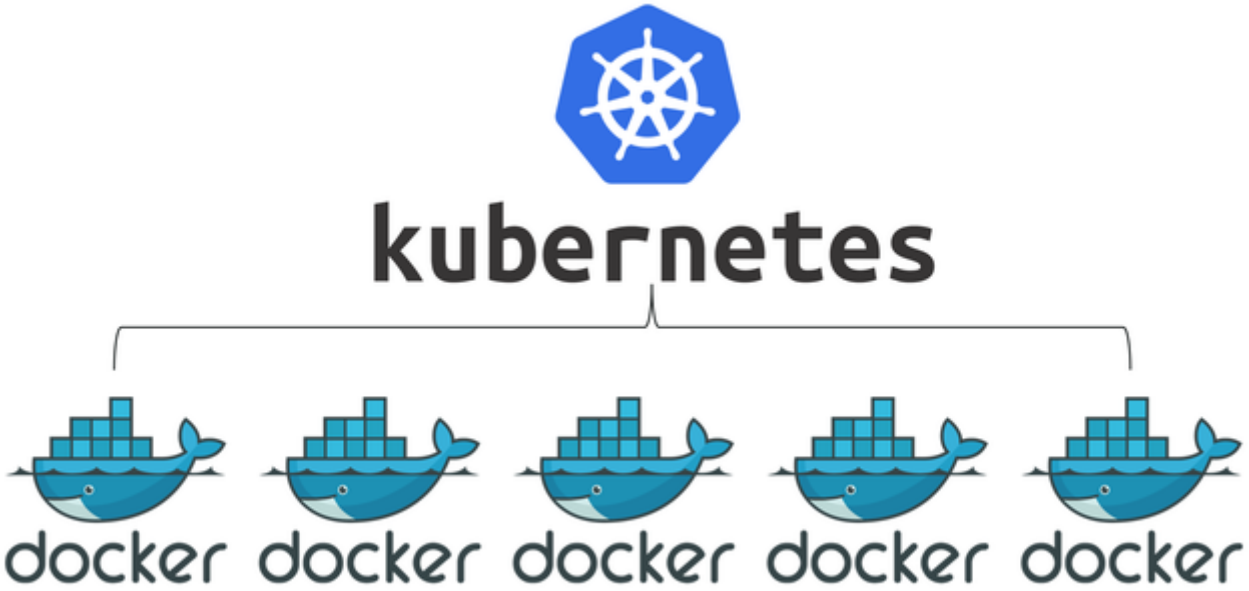
Kubernetes - Autoscaling

Autoscaling là một tính năng quan trọng trong Kubernetes cho phép tự động điều chỉnh số lượng Pod để đáp ứng với tải lưu lượng truy cập đến ứng dụng. Nó giúp tối ưu hóa sử dụng tài nguyên và đảm bảo rằng ứng dụng luôn sẵn sàng phục vụ người dùng.



Có hai loại autoscaling trong Kubernetes: Horizontal Pod Autoscaling (HPA) và Vertical Pod Autoscaling (VPA).

Horizontal Pod Autoscaling (HPA)

HPA được sử dụng để điều chỉnh số lượng Pod của một Deployment hoặc ReplicationController dựa trên tải lưu lượng truy cập đến ứng dụng. Nó hoạt động bằng cách sử dụng các metrics như CPU utilization hoặc custom metrics để quyết định số lượng Pod cần tạo ra hoặc xóa bỏ.

Để sử dụng HPA, bạn cần thực hiện các bước sau:

1. Cài đặt các metrics server để thu thập thông tin về tài nguyên và lưu lượng truy cập.
2. Định nghĩa một HPA cho Deployment hoặc ReplicationController của bạn bằng một file YAML.
3. Khi tải lưu lượng truy cập đến ứng dụng tăng, HPA sẽ tự động tạo thêm các Pod để đáp ứng với nhu cầu và giảm số lượng Pod khi tải lưu lượng truy cập giảm.

Horizontal Pod Autoscaling (HPA) là một trong những tính năng quan trọng nhất của Kubernetes. Nó cho phép tự động điều chỉnh số lượng Pod để đáp ứng với tải lượng truy cập đến ứng dụng. HPA hoạt động dựa trên các metrics như CPU utilization hoặc custom metrics, và sử dụng thông tin này

để quyết định số lượng Pod cần tạo ra hoặc xóa bỏ. Khi tải lưu lượng truy cập đến ứng dụng tăng, HPA sẽ tự động tạo thêm các Pod để đáp ứng với nhu cầu và giảm số lượng Pod khi tải lưu lượng truy cập giảm.

Ví dụ: Giả sử bạn có một Deployment trong Kubernetes chứa một ứng dụng web. Bạn muốn sử dụng HPA để tự động tạo thêm các Pod khi CPU utilization của các Pod hiện có vượt quá 80%.

Đầu tiên, bạn cần tạo một file YAML để định nghĩa HPA. Ví dụ:

```
apiVersion: autoscaling/v1
kind: HorizontalPodAutoscaler
metadata:
  name: my-app-hpa
spec:
  scaleTargetRef:
    apiVersion: apps/v1
    kind: Deployment
    name: my-app
  minReplicas: 1
  maxReplicas: 10
  targetCPUUtilizationPercentage: 80
```

Trong file YAML này, một HPA có tên là `my-app-hpa` sẽ được tạo. Nó liên kết với một Deployment có tên là `my-app`. HPA sẽ giữ số lượng Pod trong khoảng từ 1 đến 10 và sẽ tự động tạo thêm các Pod khi CPU utilization vượt quá 80%.

Sau khi triển khai file YAML này, HPA sẽ bắt đầu thu thập thông tin về CPU utilization và tự động điều chỉnh số lượng Pod để đáp ứng với tải lưu lượng truy cập đến ứng dụng.

Vertical Pod Autoscaling (VPA)

VPA được sử dụng để điều chỉnh tài nguyên của các container trong Pod. Nó hoạt động bằng cách thay đổi các giá trị tài nguyên được yêu cầu bởi container để giảm thiểu sự lãng phí và tăng khả năng sử dụng tài nguyên. VPA hỗ trợ các container chạy trên các node Kubernetes và sử dụng các metrics như CPU và memory để quyết định số lượng tài nguyên cần thiết cho mỗi container.

VPA hỗ trợ các container chạy trên các node Kubernetes và sử dụng các metrics như CPU và memory để quyết định số lượng tài nguyên cần thiết cho mỗi container. Nó có thể giúp giảm thiểu sự lãng phí và tăng khả năng sử dụng tài nguyên bằng cách thay đổi các giá trị tài nguyên được yêu cầu bởi container.

Ví dụ:

```
apiVersion: apps/v1
  kind: Deployment
  metadata:
    name: nginx-deployment
    labels:
      app: nginx
  spec:
    replicas: 2
    selector:
      matchLabels:
        app: nginx
    template:
      metadata:
        labels:
          app: nginx
      spec:
        containers:
        - name: nginx
          image: nginx:1.7.8
          ports:
          - containerPort: 80
```

- VPA resource:

```
apiVersion: autoscaling.k8s.io/v1beta1
  kind: VerticalPodAutoscaler
  metadata:
    name: nginx-deployment-vpa
  spec:
    targetRef:
      apiVersion: "apps/v1"
      kind:       Deployment
      name:       nginx-deployment
    updatePolicy:
      updateMode: "Off"
```

- Recommended resource:

```
recommendation:
  containerRecommendations:
```

```
  - containerName: nginx
    lowerBound:
      cpu: 40m
      memory: 3100k
    target:
      cpu: 60m
      memory: 3500k
    upperBound:
      cpu: 831m
      memory: 8000k
```

Tổng kết

Autoscaling là một tính năng quan trọng trong Kubernetes giúp tăng khả năng sẵn sàng và tối ưu hóa sử dụng tài nguyên. Horizontal Pod Autoscaling (HPA) và Vertical Pod Autoscaling (VPA) là hai tính năng chính trong autoscaling của Kubernetes, cung cấp các cách tiếp cận khác nhau để điều chỉnh số lượng Pod và tài nguyên của chúng.

Revision #1
Created 10 July 2023 09:43:38 by Laptrinh.vn
Updated 10 July 2023 09:51:22 by Laptrinh.vn