Cluster Auto-Scaler

ow

Overview

- Autoscaling is a function that automatically scales resources up or down to meet inconsistent demands. This is a major Kubernetes function that would otherwise require extensive human resources to perform manually.
- If we deploy our cluster with a “managed” node group, AWS will create an auto-scaling group and manage it as a part of an EKS cluster deployment
- We can also define the ***maximum, minimum,*** *and* ***desired capacity*** for the auto-scaling group (ASG).

..

- **When to use Cluster Autoscaler?**

ASG will scale when **CPU** and **Memory** usage goes high. But in our current situation, CPU or Memory usage isn’t really high; what’s stopping the pod to be created is the resource requests exceeding the available resource. ASG does not have information about the allocated resources for the pods. This is where **Cluster Autoscaler** comes into the image.

- **What will Cluster Autoscaler accomplish?**

The Cluster Autoscaler automatically adds or removes nodes in a cluster based on resource requests from pods. The Cluster Autoscaler doesn’t directly measure CPU and memory usage values to make a scaling decision. Instead, it checks every 10 seconds to detect any pods in a pending state, suggesting that the scheduler could not assign them to a node due to insufficient cluster capacity.

How Cluster Autoscaler works on AWS EKS?

- Cluster Autoscaler needs to be deployed as a deployment on the EKS cluster. The autoscaler pods will detect that any pods are in a pending state due to the lack of resources. If there is any pod is in the pending state, the cluster autoscaler will call the AWS API to update the auto-scaling group automatically.
- Since Custer autoscaler requires the ability to examine and modify EC2 Auto Scaling Groups. By using IAM roles for Service Accounts to associate the Service Account that the Cluster Autoscaler Deployment runs as with an IAM role, the autoscaler pod can manipulate ASG to scale in or out the EC2 instances of a node group.


EC2 Dashboard
EC2 Global View
Events
Console-to-Code Preview
Instances
Instances
Instance Types
Launch Templates
Spot Requests
Savings Plans
Reserved Instances
Resources
EC2 Global view
You are using the following Amazon EC2 resources in the US East (N. Virginia) Region:
Instances (running) 2
Auto Scaling Groups 1
Dedicated Hosts 0
Elastic IPs 2
Instances 2
Key pairs 2
Load balancers 2
Placement groups 0
Security groups 8
Snapshots 0
Volumes 2
EC2 Free
Offers for all
2 EC2 fre
End of mor
0 offers
er limit.
Exceeds fre
0 offers
s-you-go pr
View Globa