Autoscaling EC2 with ALB

> Create an autoscaling mechanism which will automatically increase the number of instances by one if CPU utilization is more than 40% again by one if cpu utilization is more than 80%. Use below services –
> VPC, SG, EC2, Target groups, Hosted zones, ALB, Route 53, SNS, Cloudwatch, Autoscaling group, Dynamic scaling policies.

Here's a step-by-step guide to setting up an autoscaling mechanism that increases the number of EC2 instances when CPU utilization crosses 40% and again when it exceeds 80%.

Step 1: Set Up a VPC and Security Group (SG)

1. **Create a VPC**
   - Go to **VPC Console** → Create VPC
   - Name: `` `MyVPC` ``
   - Choose **IPv4 CIDR block** (e.g., `` `10.0.0.0/16` ``)
2. **Create a Security Group (SG)**
   - Go to **EC2 Console** → Security Groups
   - Create a new security group, name it `` `AutoScalingSG` ``
   - Allow inbound traffic for **SSH (22), HTTP (80), and HTTPS (443)**
   - Associate this SG with your VPC

Step 2: Launch EC2 Instances and Configure Target Groups

1. **Create a Target Group (TG)**
   - Go to **EC2 Console** → Load Balancing → Target Groups
   - Choose **Instance-based** target group
   - Register at least one instance
2. **Launch EC2 Instances**
   - Choose an AMI (Amazon Linux 2)
   - Select an appropriate instance type (e.g., `` `t3.micro` ``)
   - Attach the previously created **SG and VPC**
   - In **Advanced details**, add a startup script to install a web server (if needed).

Step 3: Set Up an Application Load Balancer (ALB)

1. **Create an ALB**
   - Go to **EC2 Console** → Load Balancers → Create Load Balancer
   - Choose **Application Load Balancer**

  - Attach it to the **VPC and Subnets**
  - Attach the **Security Group** (AutoScalingSG)
  - Register the **Target Group**

2. **Update Route 53 for Domain Name**
   - Go to **Route 53 Console**
   - Create a **Hosted Zone** for your domain
   - Add an **Alias Record** pointing to the ALB

Step 4: Create an Auto Scaling Group (ASG)

1. **Go to EC2 Console → Auto Scaling Groups → Create ASG**
2. **Attach the ALB Target Group**
3. **Set the minimum and maximum number of instances (e.g., min=1, max=5)**
4. **Use Launch Template** with AMI, Instance Type, SG, and VPC settings
5. **Configure Health Check to use the ALB target group**

Step 5: Set Up CloudWatch Alarms and Scaling Policies

1. **Create a CloudWatch Alarm for 40% CPU Utilization**
   - Go to **CloudWatch Console** → Alarms → Create Alarm
   - Choose **EC2: CPUUtilization**
   - Set condition: **Greater than 40%** for 2 consecutive periods
   - Set actions → **Trigger Auto Scaling Policy to add 1 instance**
2. **Create a CloudWatch Alarm for 80% CPU Utilization**
   - Repeat the above steps, but change condition: **Greater than 80%**
   - Set actions → **Trigger Auto Scaling Policy to add 1 instance**
3. **Create a Scale-In Policy**
   - Create an alarm when **CPU utilization is below 20%**
   - Set action → **Remove 1 instance**

Step 6: Configure SNS for Notifications

1. **Go to SNS Console → Create Topic** (`` `AutoScalingAlerts` ``)
2. **Subscribe your email/SMS** to receive alerts
3. **Attach SNS to CloudWatch alarms** to notify on scaling events

---

Step 7: Testing

- Simulate high CPU usage (`` `stress` `` command in EC2)
- Monitor EC2 instances scaling up/down based on CPU utilization

---

Summary of Key Components

Service	Purpose
VPC	Network isolation
Security Group (SG)	Controls traffic
EC2	Compute instances
Target Groups (TG)	Routing for ALB
ALB	Distributes traffic
Route 53	DNS resolution
CloudWatch	Monitoring and alarms
Auto Scaling Group (ASG)	Manages instance scaling
Scaling Policies	Defines when to scale
SNS	Sends notifications

This setup ensures that EC2 instances automatically scale up when CPU usage exceeds thresholds and scale down when demand decreases. 🚀