Cloud Computing Exercise #18

Auto-Scaling

A. Preparation

1. Sign in to your AWS account as the non-root admin user.

B. Create and test a launch template

1. Go to EC2 Security groups (EC2/Network & security/Security groups) and create a security group for your instances called “MySecurityGroup” that allows only SSH inbound traffic (port 22) from anywhere and allows all outbound traffic.
2. Go to EC2 launch templates (EC2/Instances/Launch templates) and select “Create launch template”. Give it a name (e.g. “MyTemplate”), select the “Amazon Linux 3 2023 AMI” as the AMI for the template and t2.micro as the instance type. Choose your existing key pair (e.g. “mykey”) for instance access, and set the template’s security group to the security group you created in step 2 (MySecurityGroup). Open the “Advanced details” sub-menu, go to “User data”, and add the following commands to the text window:

#! /bin/bash

amazon-linux-extras install -y epel

yum install -y stress

This Bash script will run each time a new EC2 instance is launched and will install the “stress” utility on each EC2 instance automatically. (This is an alternative to creating an EC2 instance, manually installing SW on it and then creating a custom AMI from the EC2 instance.) Scroll down to the bottom and create the launch template. In the Launch Templates sub-menu, your new launch template should show up.

1. Select your launch template and click on “Actions” and select “Launch instance from template”. Scroll all the way down and launch the instance.
2. If you go to the list of EC2 instances (EC2/Instances/Instances), you will see your EC2 instance initializing and then running. Log into the EC2 instance using your SSH client.
3. At the OS prompt, type: stress. You should see help/usage information printed for the “stress” utility. Thus, you could launch an instance from a launch template (without manually specifying its details) and provide a script that automatically installed SW on this EC2 instance at launch time.
4. Exit from the SSH session and terminate your EC2 instance.

C. Create an auto-scaling group

1. Go to Auto Scaling Groups (EC2/Auto scaling/Auto scaling groups), and click on “Create Auto Scaling group”. Provide an auto scaling group name (e.g. “MyASG”) and select the launch template you created (“MyTemplate”). Go to the next page – you can leave most of the settings on the “Configure settings page” with their default values, except for the “Subnets” setting; you can pick the subnet in the “us-east-1a” AZ. Thus, in this lab, we will use auto-scaling within a single AZ. (If you want your auto-scaling group to operate across multiple AZs for high reliability/availability, you need to specify multiple subnets in multiple AZs here.) On the “Configure advanced options” page in the “Additional settings” section, enable group metrics collection and go to the next page. On the “Configure group size and scaling policies” page, configure the auto-scaling group size by setting the desired capacity to 2 instances, the minimum capacity to 2 instances, and the maximum capacity to 4 instances. Thus, you will run at least to EC2 instances, and at most four EC2 instances. In the “Scaling policies” section, select target tracking scaling policy, set the metric type to “Average CPU utilization” and set the target value to 60%.
2. On the “Add notifications” page, configure email notifications so that you would get an email when the auto-scaling group launches new EC2 instances or terminates running EC2 instances. Click on the “Add notification” button, and then on the “Create a topic” button. Give the topic a name (e.g. “AutoScalingTest”), and provide your email address where you want to receive the notifications. Go to the next page, skip adding tags and create the auto-scaling group. If you look at the list of your auto-scaling groups (EC2/Auto Scaling Groups), you should see the ASG you just created.
3. Check your email. You will receive an email with the subject “AWS Notification - Subscription Confirmation”. Click on the “Confirm subscription” link to confirm that you intend to receive email messages sent to the topic “AutoScalingTest”.

D. Test your auto-scaling group

1. Go to the list of your auto-scaling groups ((EC2/Auto Scaling Groups), and select your ASG. On the “Details” tab, you will see all the configured ASG details. On the “Activity” tab, you will see a list of auto-scaling events with two successful EC2 instance launches. On the “Instance management” tab, you will see your two EC2 instances with “InService” status. (The same instances can also be found with “Running” state on the list of your EC2 instances at EC2/Instances/Instances).
2. The “Monitoring” tab shows the CloudWatch metric details for your ASG. The “Auto Scaling” tab shows the group metrics; within this tab, the “In Service Instances” graph shows how the number of running instances changes as a function of time. The “EC2” tab shows instance-specific metrics; within this tab, the “CPU Utilization” graph shows the average CPU utilization for the ASG.
3. Log into both of your running instances and open two SSH sessions per instance. In the first window, enter the following command: stress --cpu 2 --timeout 600. This will create two processes that will increase the CPU load for 600 seconds. In the second window, run the top utility and check how your CPU utilization will go up to 100% (and the CPU load average should go up to 2.0). Do this for both EC2 instances (run stress in both the ssh windows)
4. Monitor the CPU utilization CloudWatch metric in the AWS management console. You may have to wail a few minutes, but after a while the metric should go up to close to 100%. When the metric breaches the 50% mark, the auto-scaling group will launch a third, and then a fourth EC2 instance. You will see the new instances in your auto-scaling group (“Instance management” tab), and you will also receive emails with the subject “Auto Scaling: launch for group MyASG" describing the EC2 launch event. This event will also show up in the list of ASG events under the “Activity” tab.
5. After a few minutes, the stress utility will time out and you will see that the CPU utilization values in the “top” utility’s SSH windows will go back to near zero. Monitor the CloudWatch metrics in the “Monitoring” tab; the CPU utilization metric will also go back to zero, even though after a few minutes of delay. When this happens, one of the four instances will be terminated, and then another one, ending up with two running EC2 instances in the ASG. Note: the instances to be terminated could be the ones you are currently logged into, so do not be surprised if you lose the SSH session with some of your EC2 instances. Check the “Instance management” tab and verify that you only have two remaining instances. The “Activity” tab will show the termination events. Also, check your email to see the notifications about the EC2 instance termination.

F. Clean up after yourself

1. Go to the list of your auto-scaling groups (EC2/Auto Scaling Groups), select your ASG and delete it (“Delete” button). Note that when you delete your ASG, the two remaining EC2 instances will also be terminated. Make sure that you are not leaving any running EC2 instances.
2. Log out of AWS.