**Text

Description automatically generated with medium confidence**

**Purpose**

Services fluctuate in resource usage. At peak times, services and EC2 instances may be running at full utilization levels, leading to a sluggish service for incoming users. With the implementation of AWS load balancers, incoming traffic can be directed and handled by multiple EC2 instances. The implementation of auto scaling allows EC2 instances to be either started or terminated depending on usage levels, making sure there are enough EC2 instances to handle the given traffic. These two services combine to provide strong redundancy with multiple functional concurrent EC2 instances that both reduce impact of overload and single individually faulty virtual machines.

**Lab Summary**

A target group and a load balancer were created in order to designate the group of EC2 instances that traffic should be directed towards. An Amazon machine image (AMI) was created alongside a launch configuration to designate a template for which new EC2 instances can be launched using. An autoscaling group was then created using the load balancer and launch configuration in order to launch and new template EC2 instances and share load across them. Cloudwatch was used to set launch conditions, including one that launches a new EC2 instance when CPU utilization goes over 60%. The setup was tested using a predefined tool that creates large amounts of traffic.

**Creating a Target Group:**

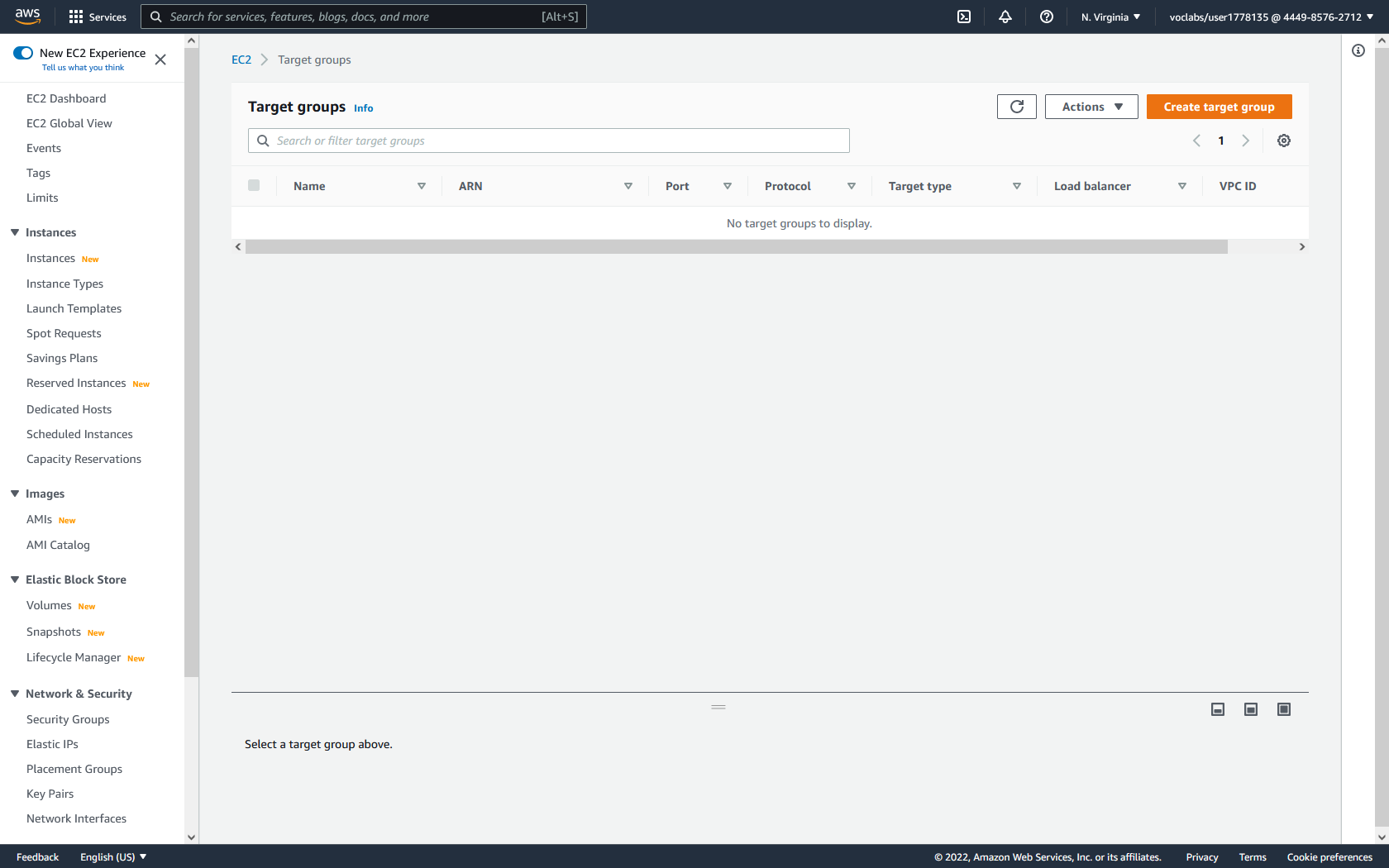
1. In the left menu, click Target groups to reach the Target groups dashboard. Then click Create target group.

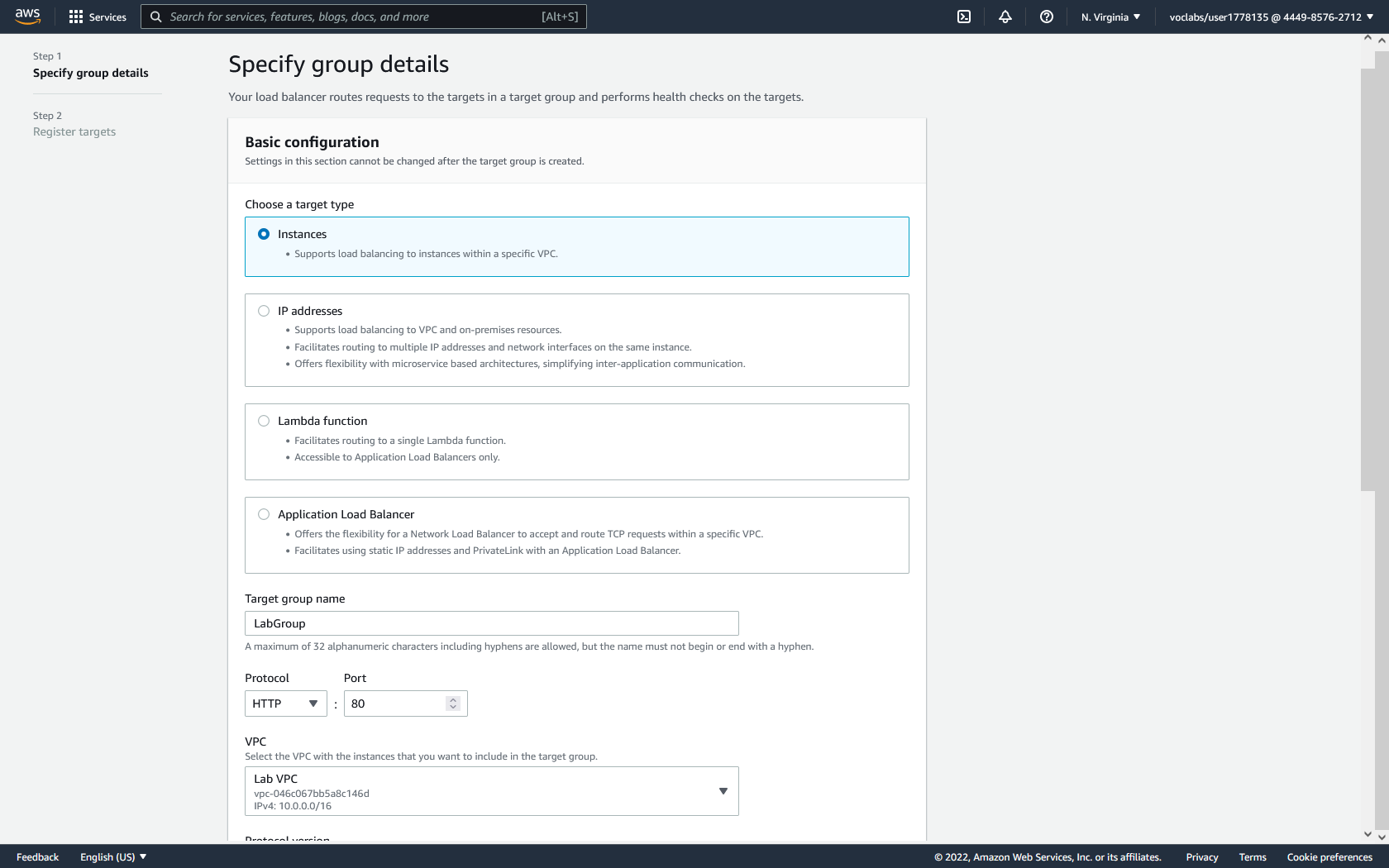
2. Choose the target type to be instances.

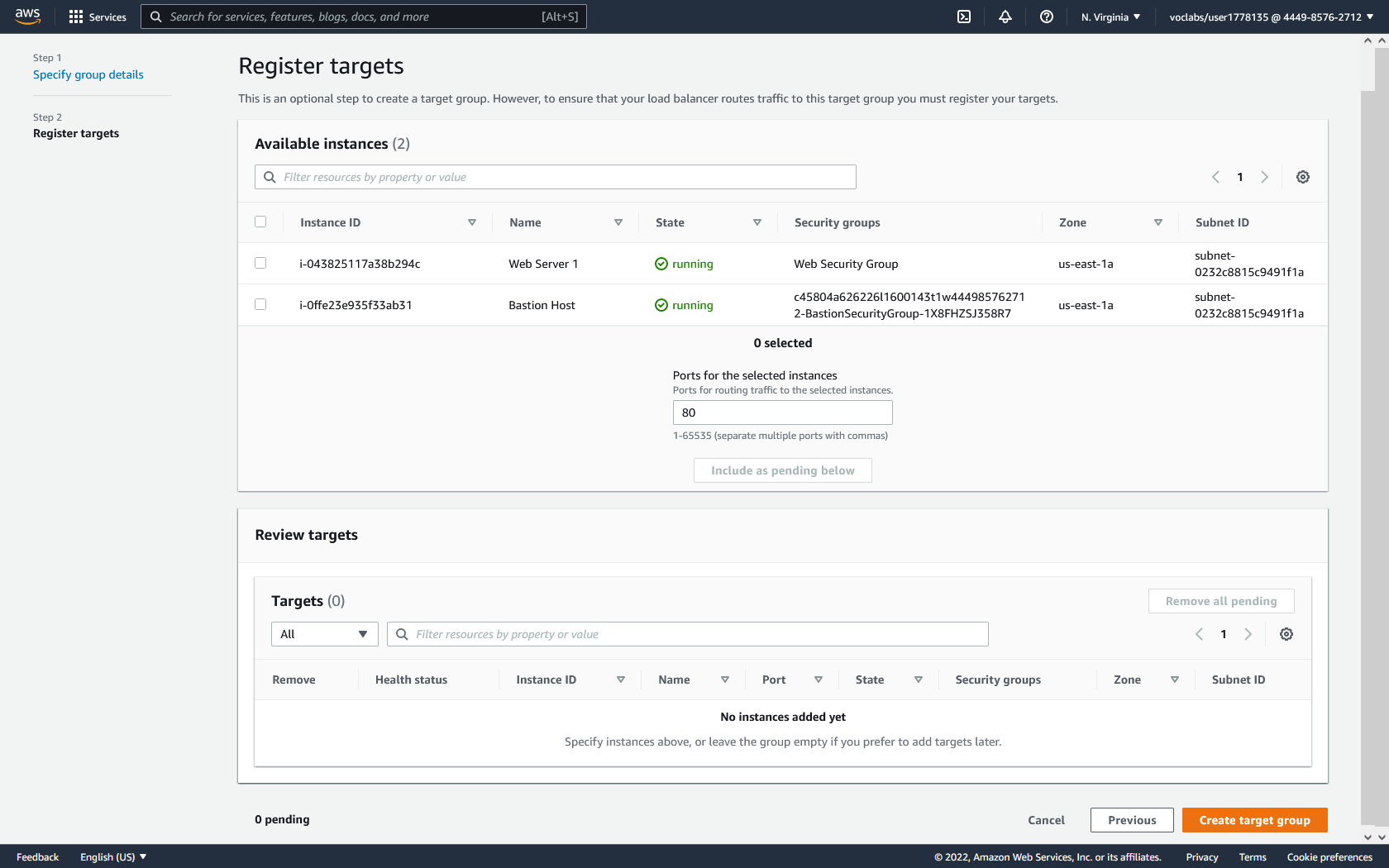
3. Give the target group the name of LabGroup.

4. Select the VPC to use Lab VPC. Then, click next.

5. For the register target steps, nothing needs to be done as the instances will be automatically created in a later step. Click Create target group.







**Creating a Load Balancer:**

1. In the left menu, click Load Balancers to reach the Load Balancers dashboard. Then click Create Load Balancer.

2. Click create under Application Load Balancer.

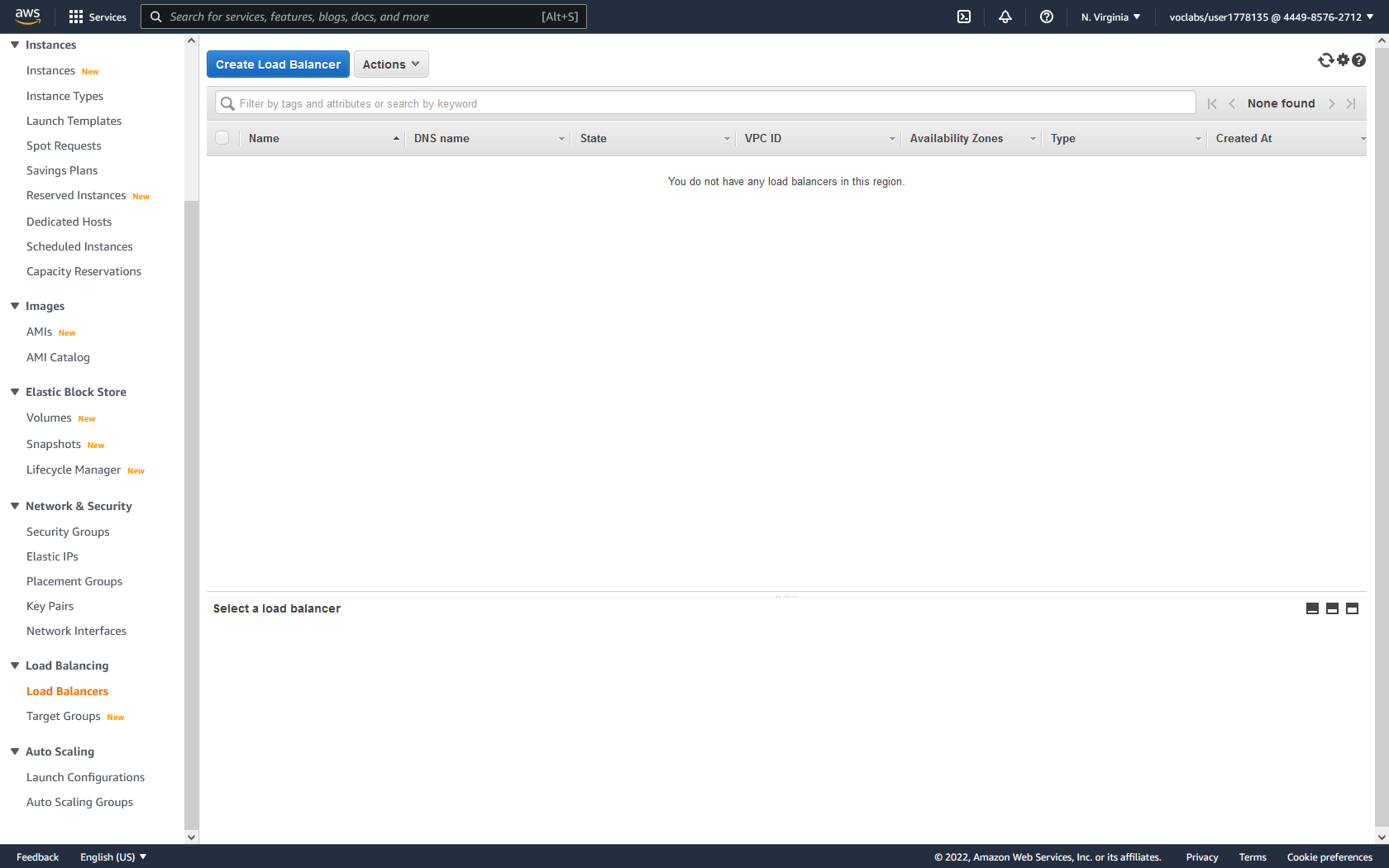
3. Give the load balancer the name of LabELB.

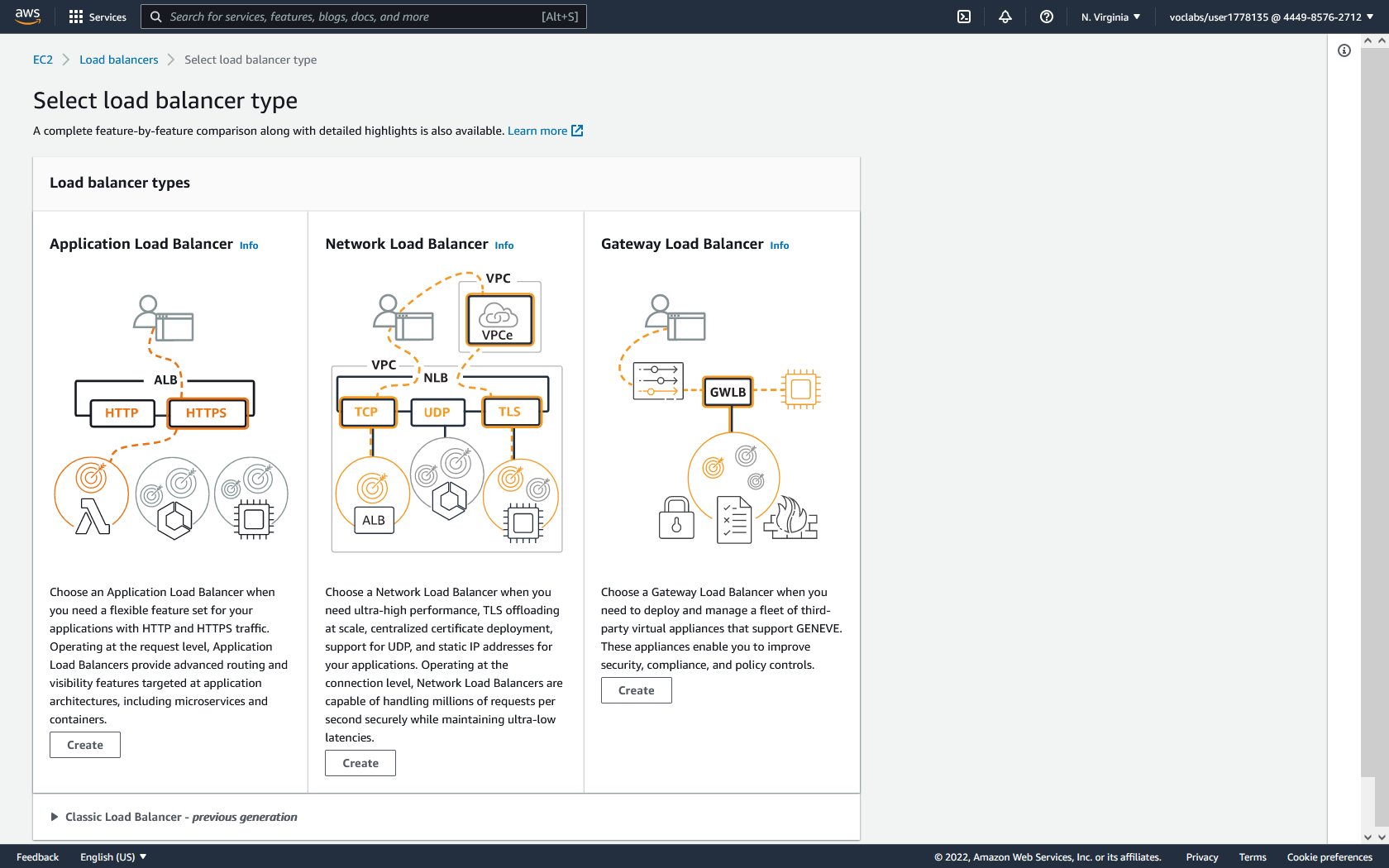
4. Under Network Mapping, select Lab VPC. For both the us-east1a and us-east1b subnets, select the public subnets.

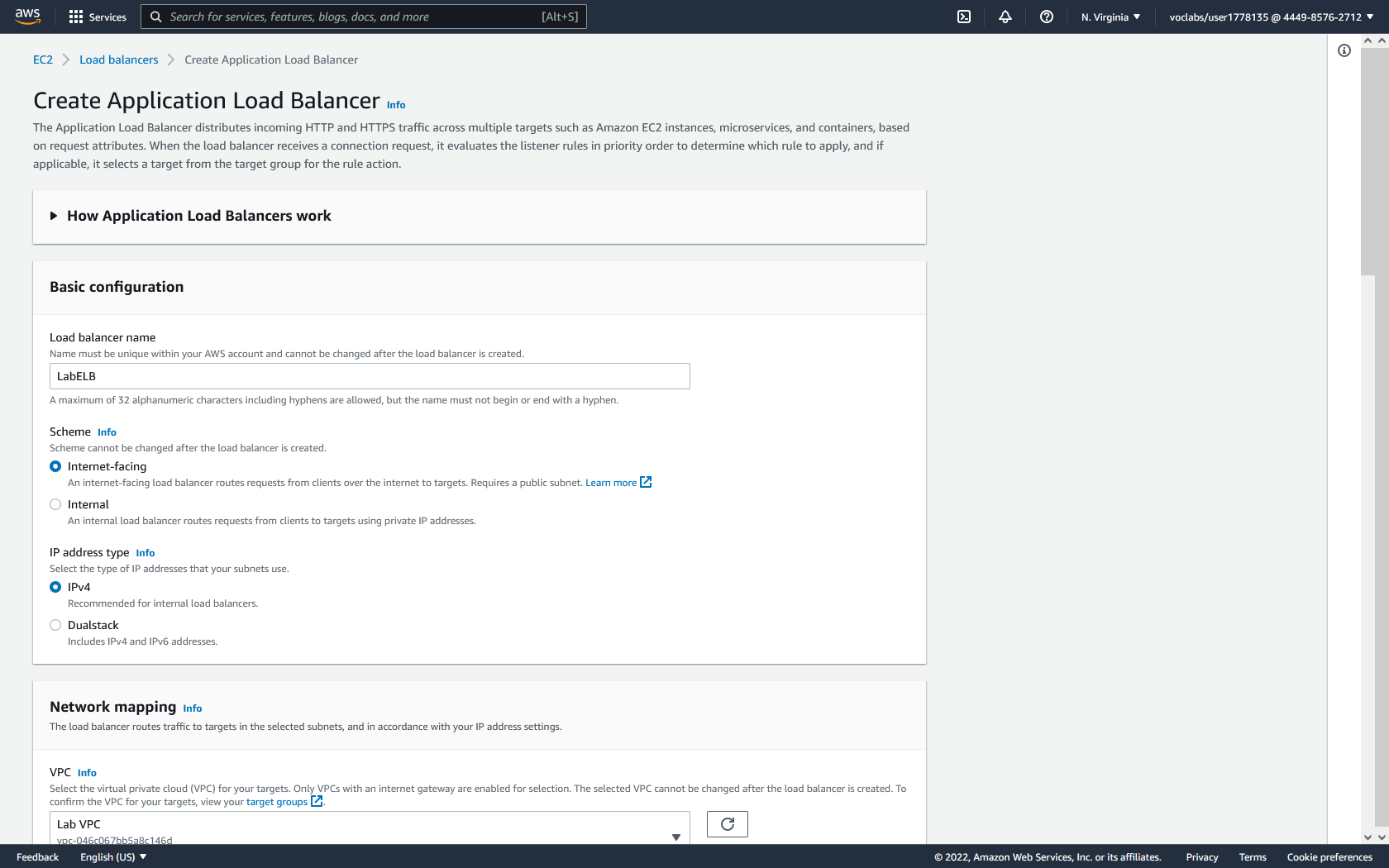
5. Under Security Groups, click the x next to the default security group to remove the default security group. Then select the Web Security Group.

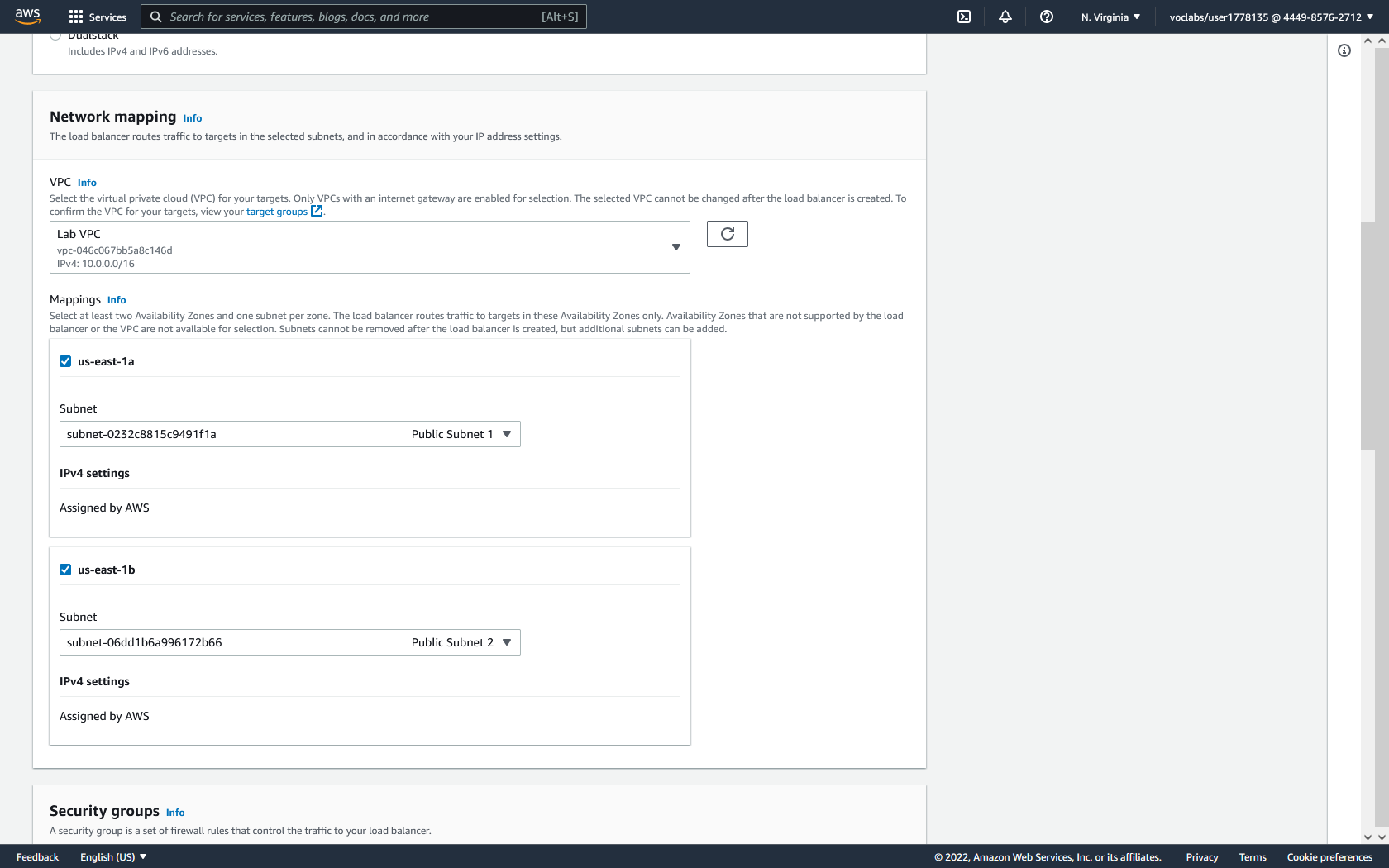
6. Under Listeners and routing, set the HTTP listener’s default action to forward to the previously created LabGroup target group.

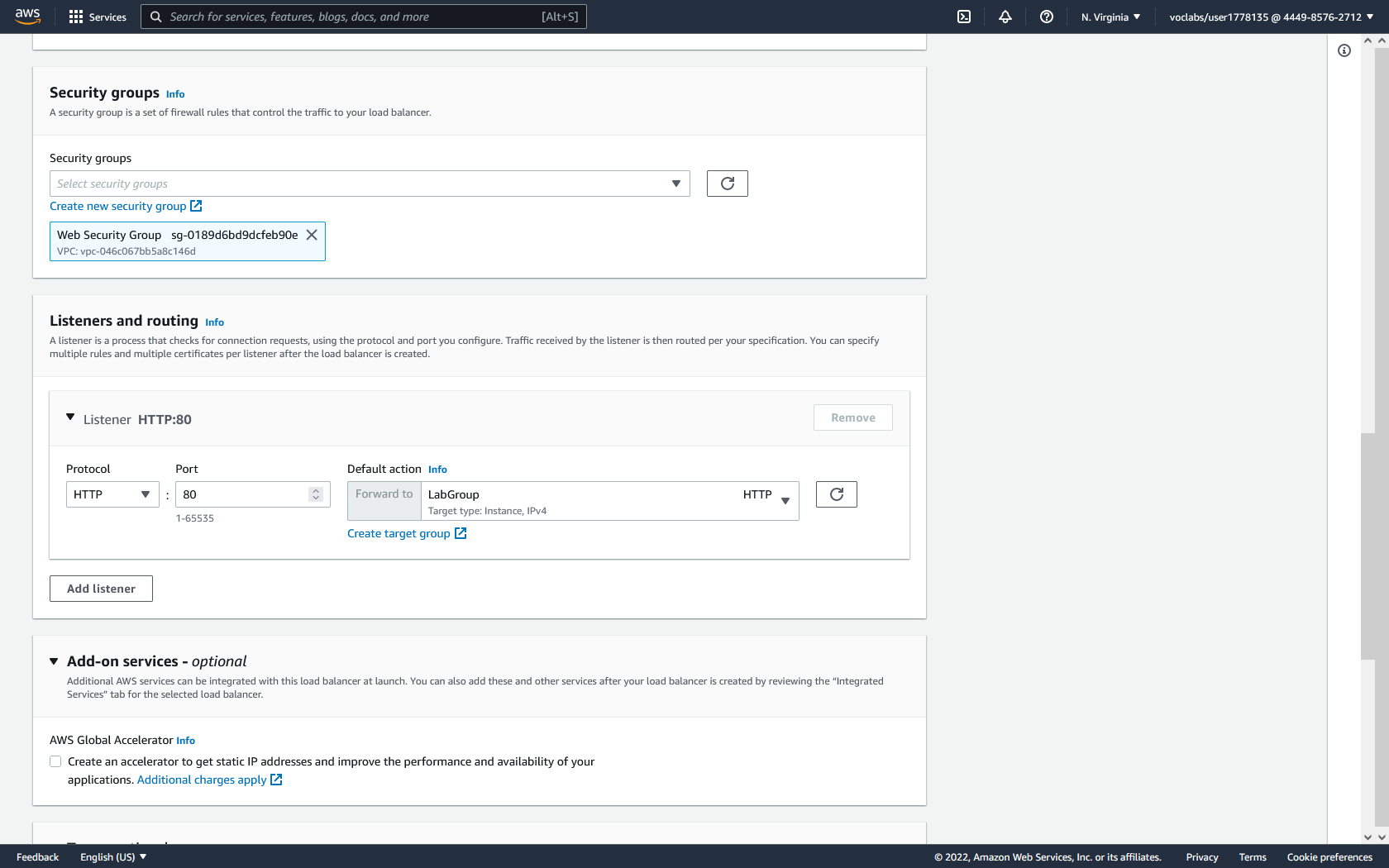
7. Click Create load balancer.

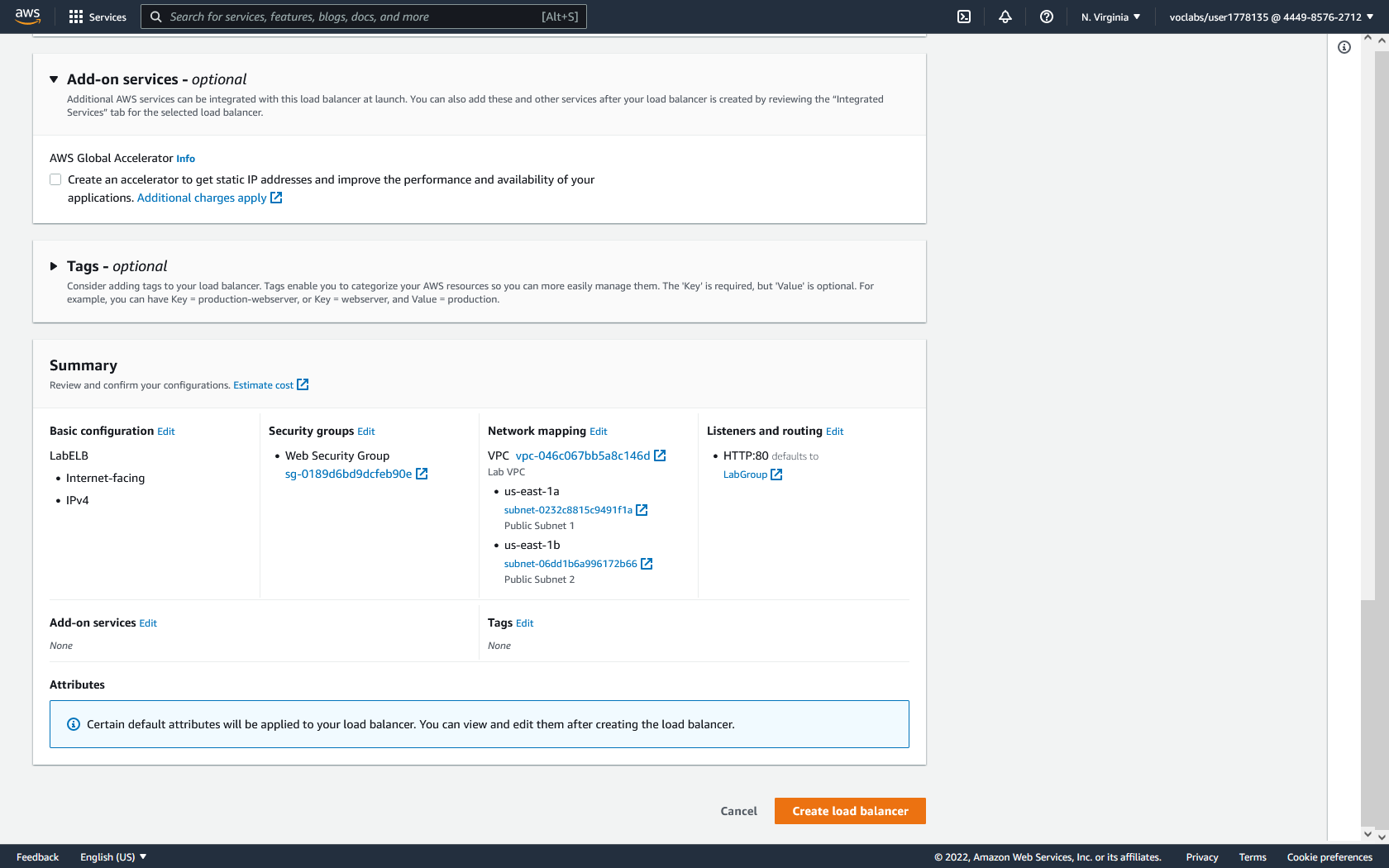












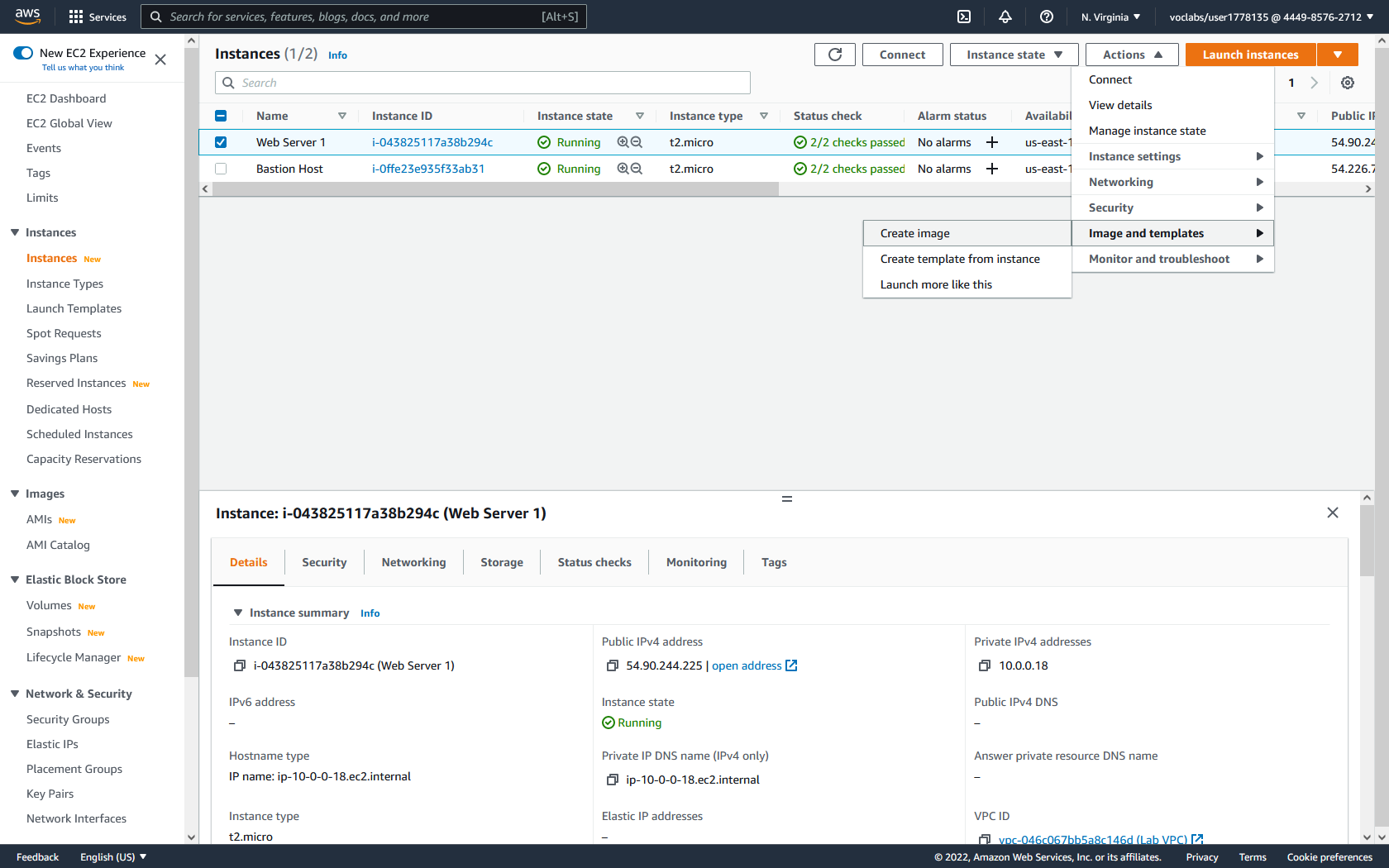


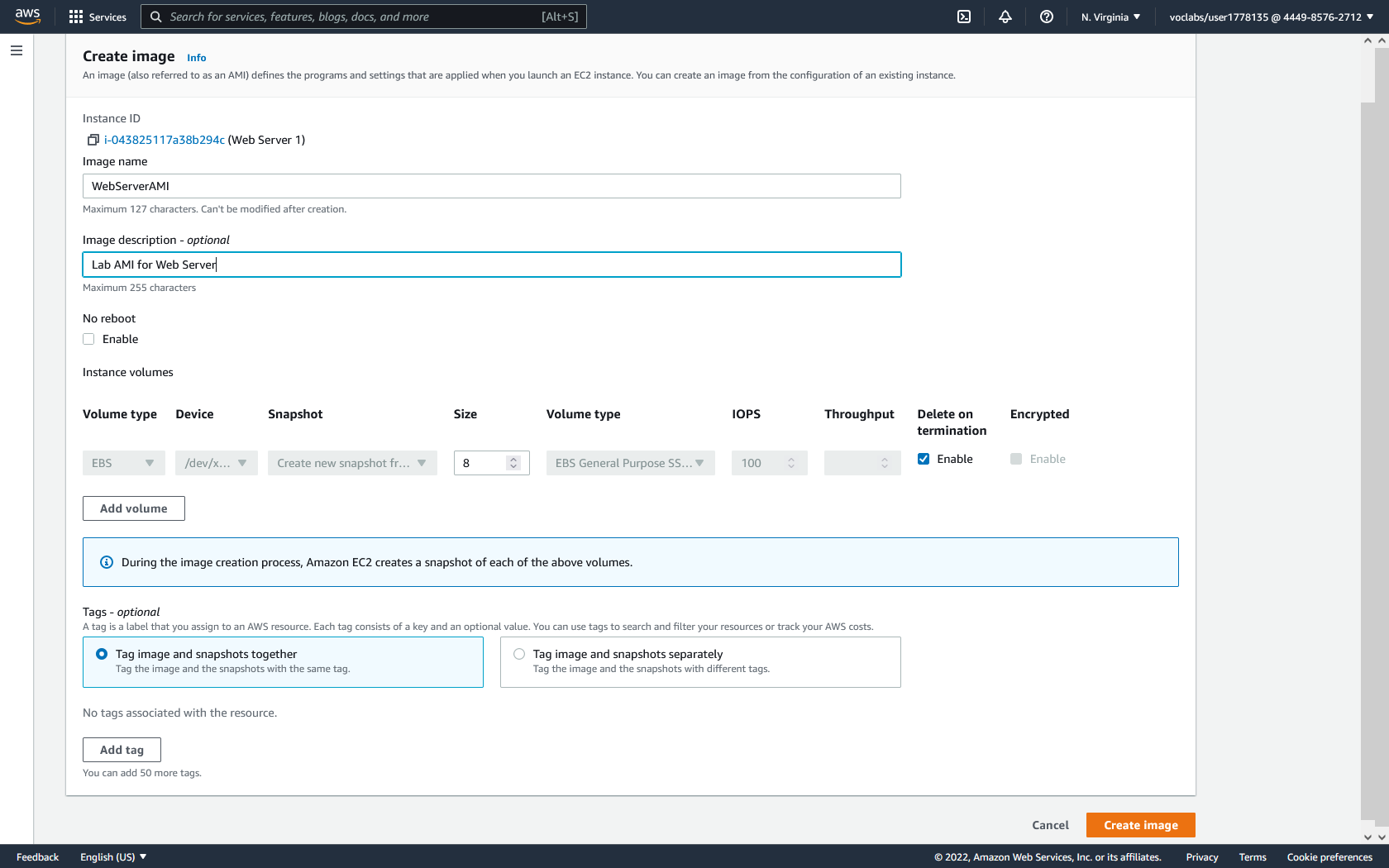
**Creating an Image of an Instance:**

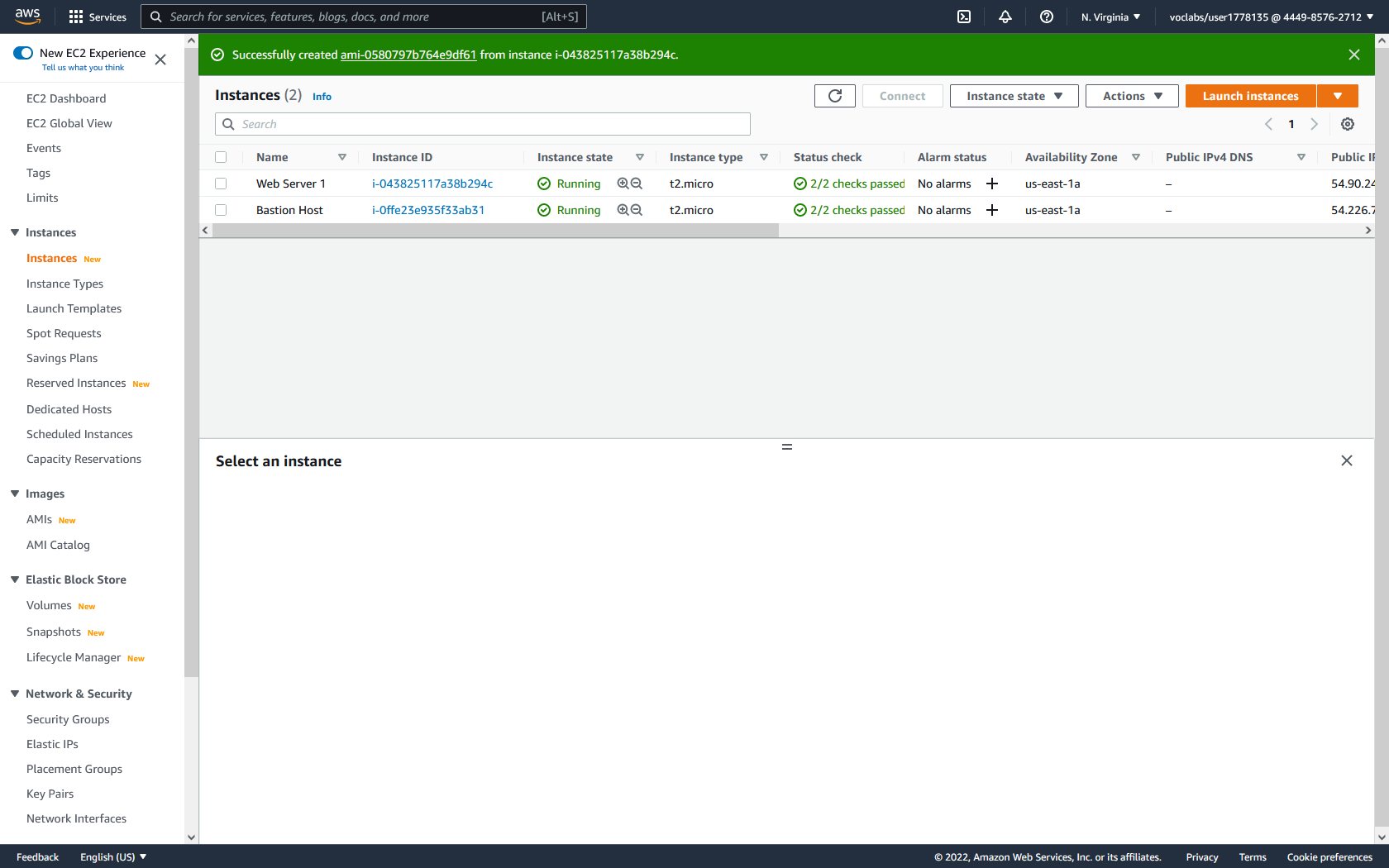
1. In the EC2 instances dashboard, select Web Server 1, then click actions, Image and templates, and click create image.

2. Give the image the name of WebServerAMI and the description of Lab AMI for Web Server.

3. Click Create Image.







**Creating a Launch Configuration:**

1. In the left menu, click Launch configurations to reach the Launch configurations dashboard. Then click create launch configuration.

2. Give the launch configuration the name of LabConfig.

3. For the Amazon Machine Image, use the previously created WebServer AMI.

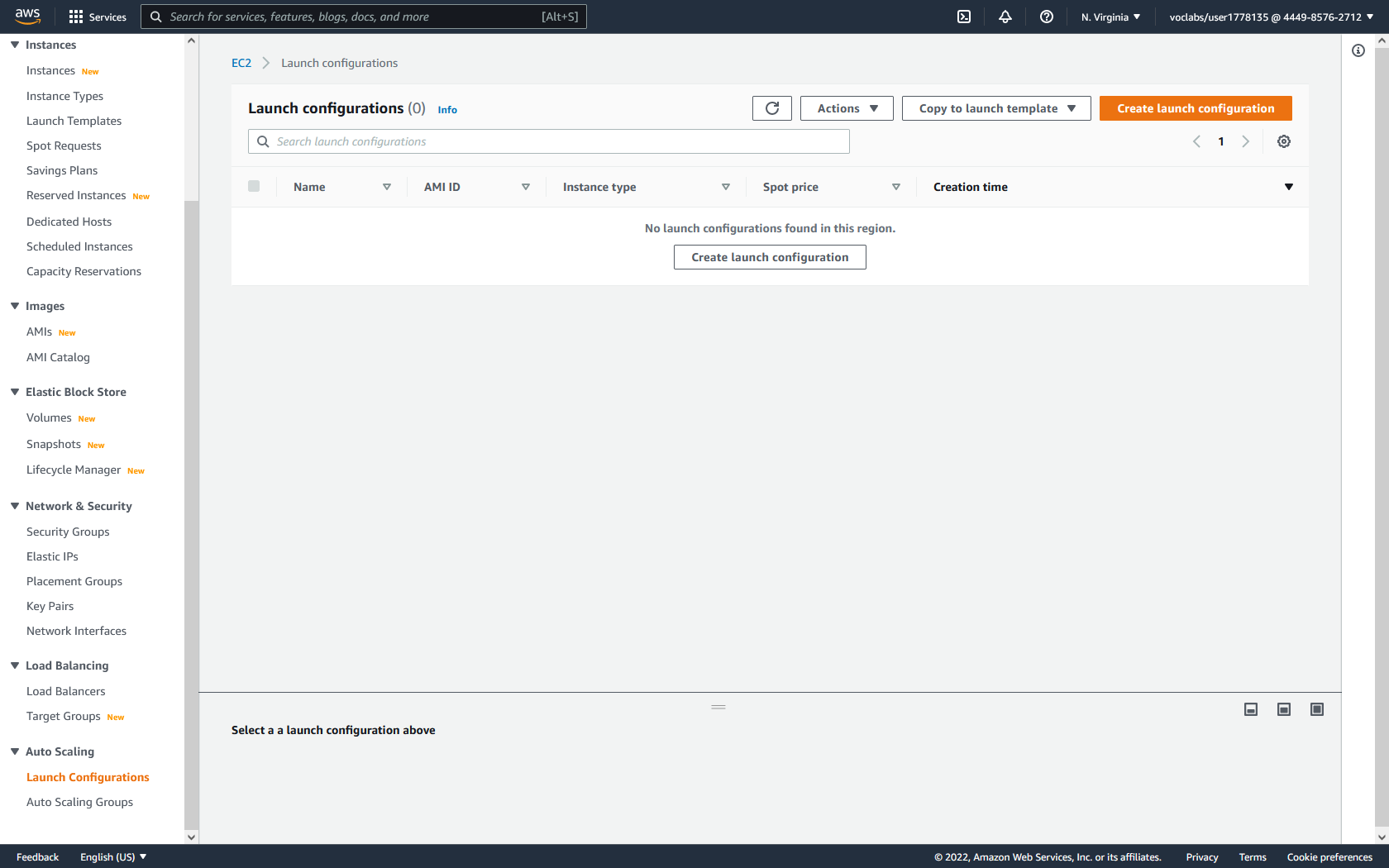
4. For Instance Type, choose the t3.micro instance.

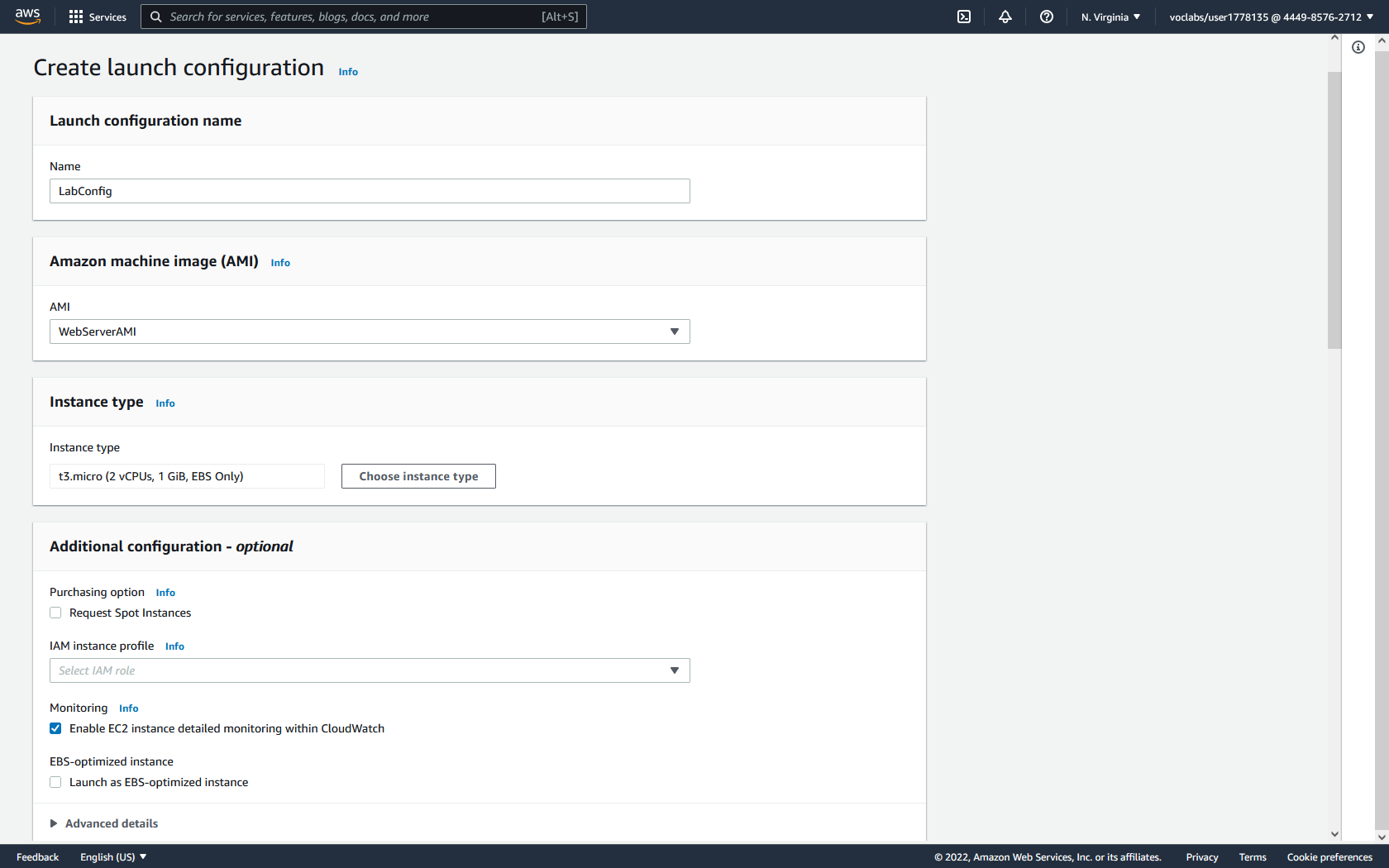
5. For Additional Configuration, check Enable EC2 instance detailed monitoring within CloudWatch.

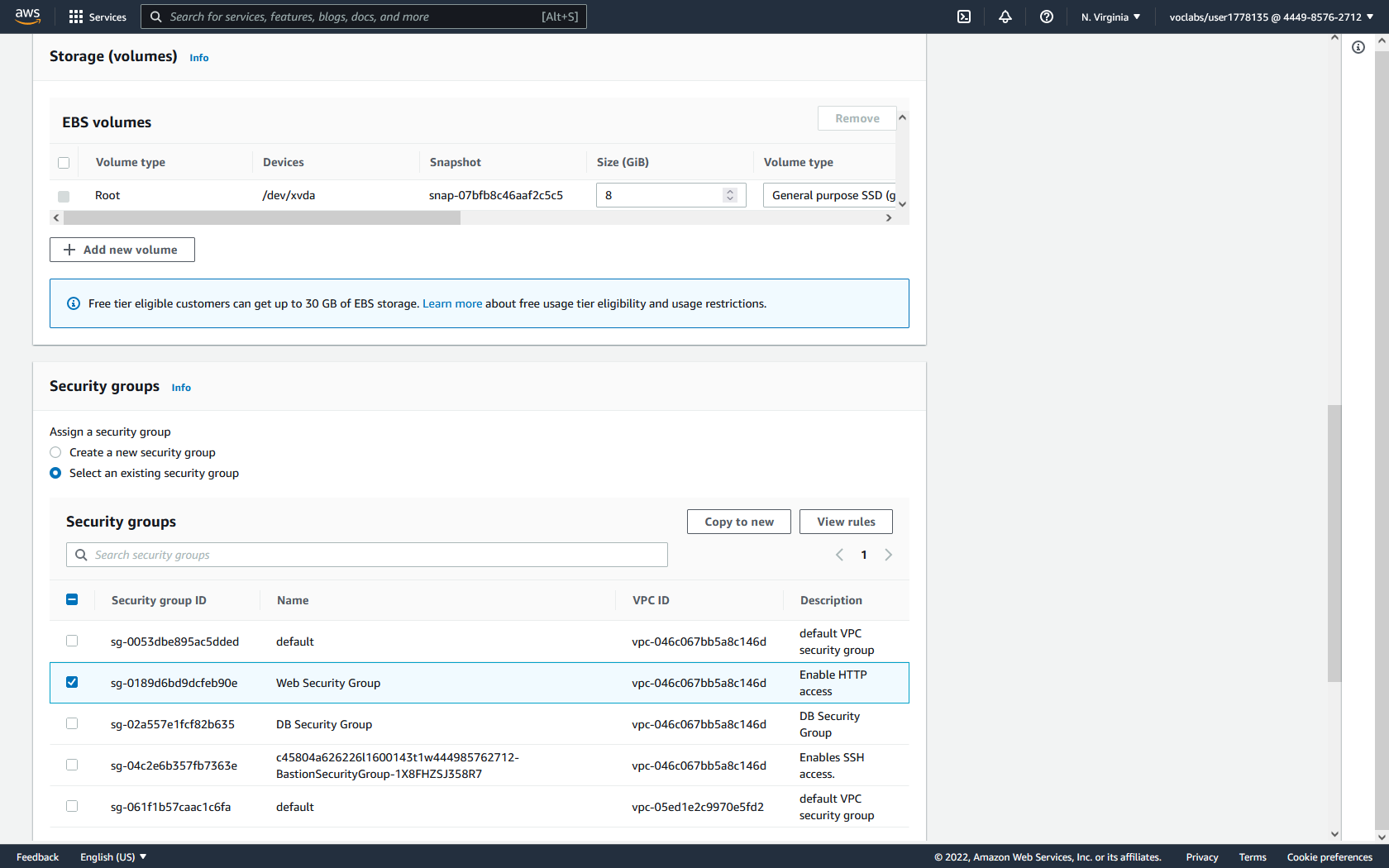
6. For Security Groups select the existing security group of Web Security Group.

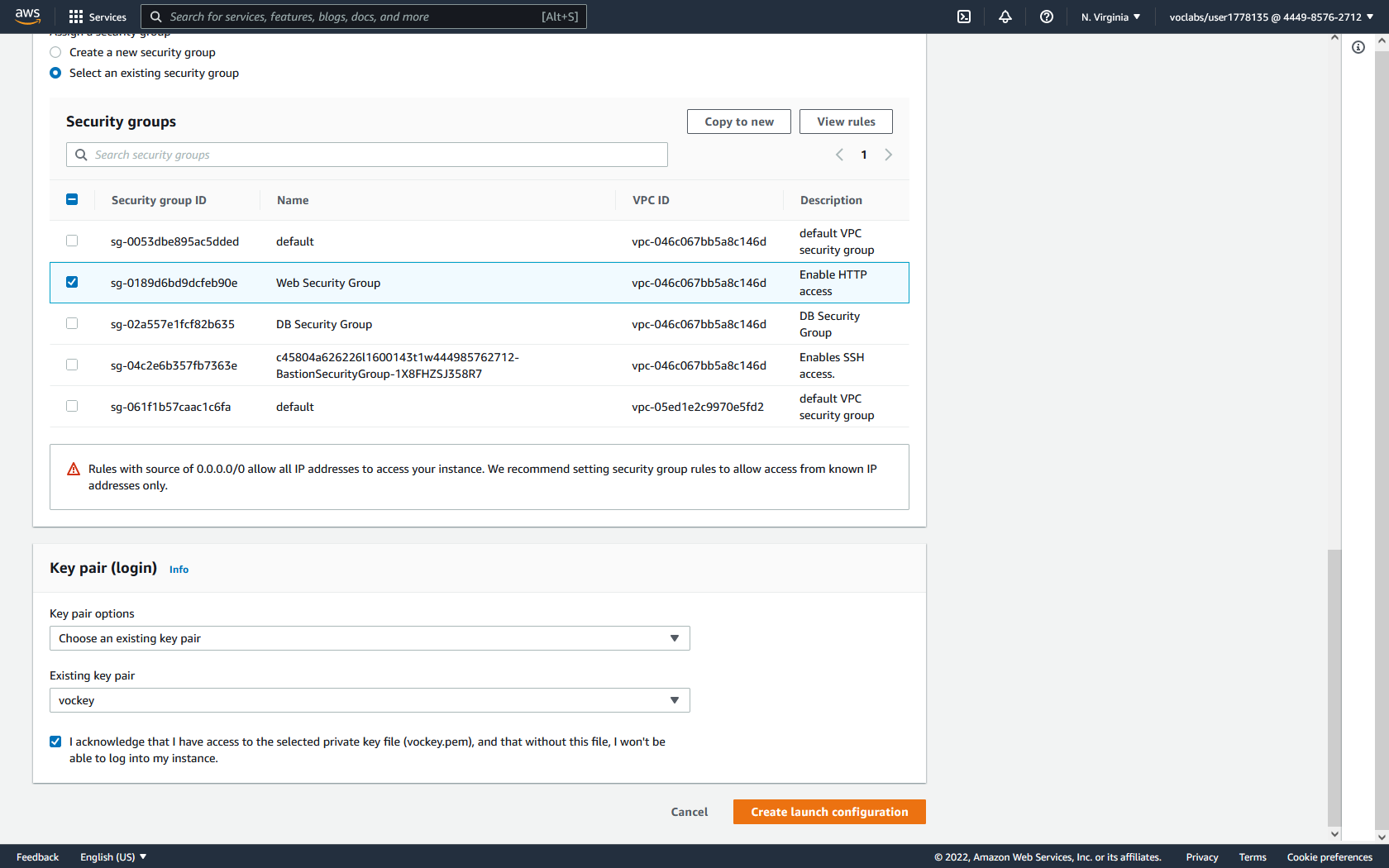
7. For Key Pair, choose the existing key pair of vockey.

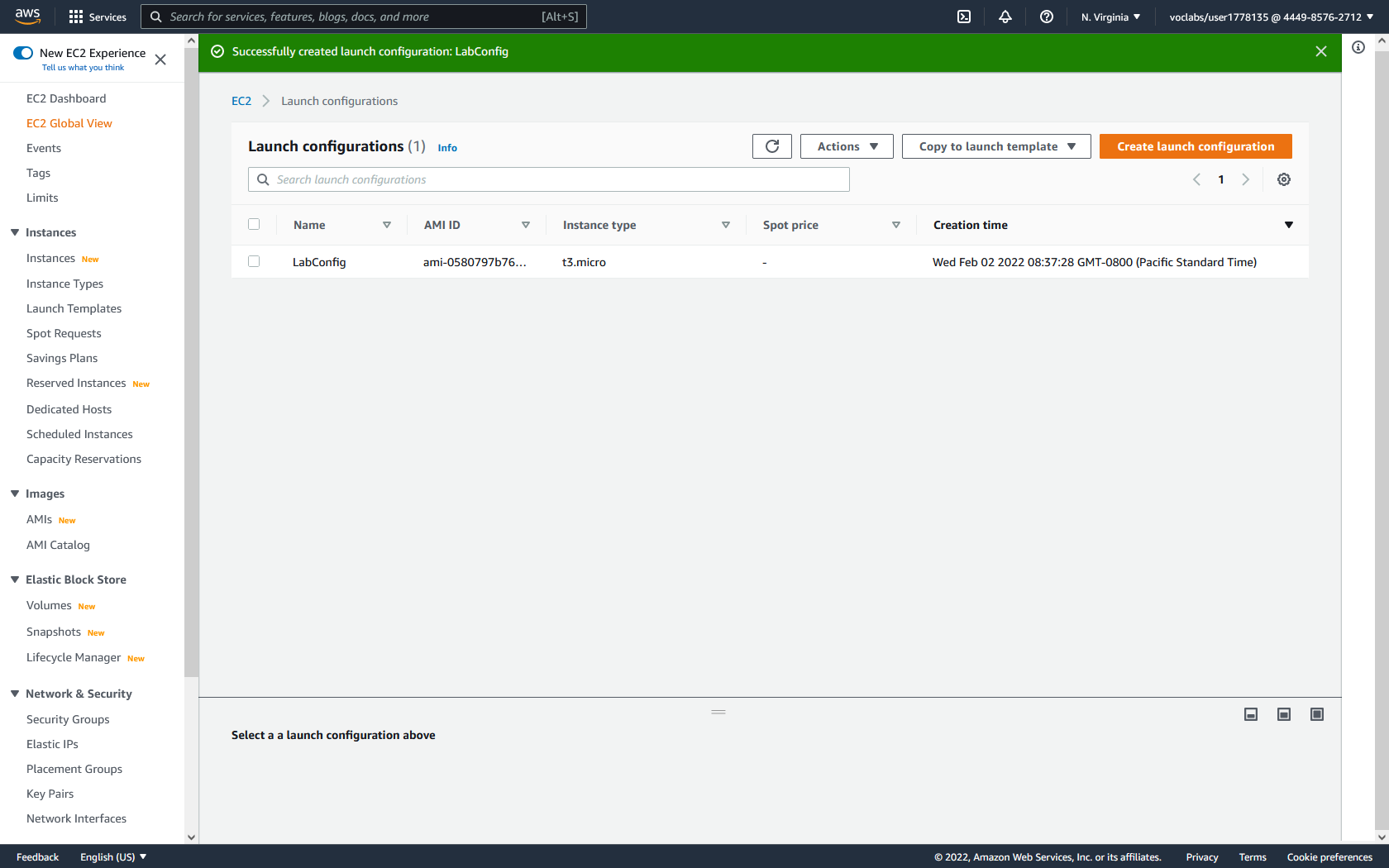
8. Click Create launch configuration.











**Creating an Auto Scaling Group:**

1. Select the configured LabConfig launch configurations. Then click actions and Create Auto Scaling group.

2. For the Launch Template or Configuration step, configure the Auto Scaling Group Name as Lab Auto Scaling Group. Also verify that it is using the LabConfig launch configuration. Then click next.

3. For the Instance Launch Options, choose the Lab VPC and both the private subnets. This is because privately launched instances will be encapsulated and appear as a single instance publicly while being load balanced.

4. For Advanced Options, attach the auto scaling group to the previously created LabELB load balancer. Then, check Enable group metrics collection with CloudWatch. Click next.

5. For Group Size and Scaling Policies, configure the following:

Desired Capacity: 2

Minimum Capacity: 2

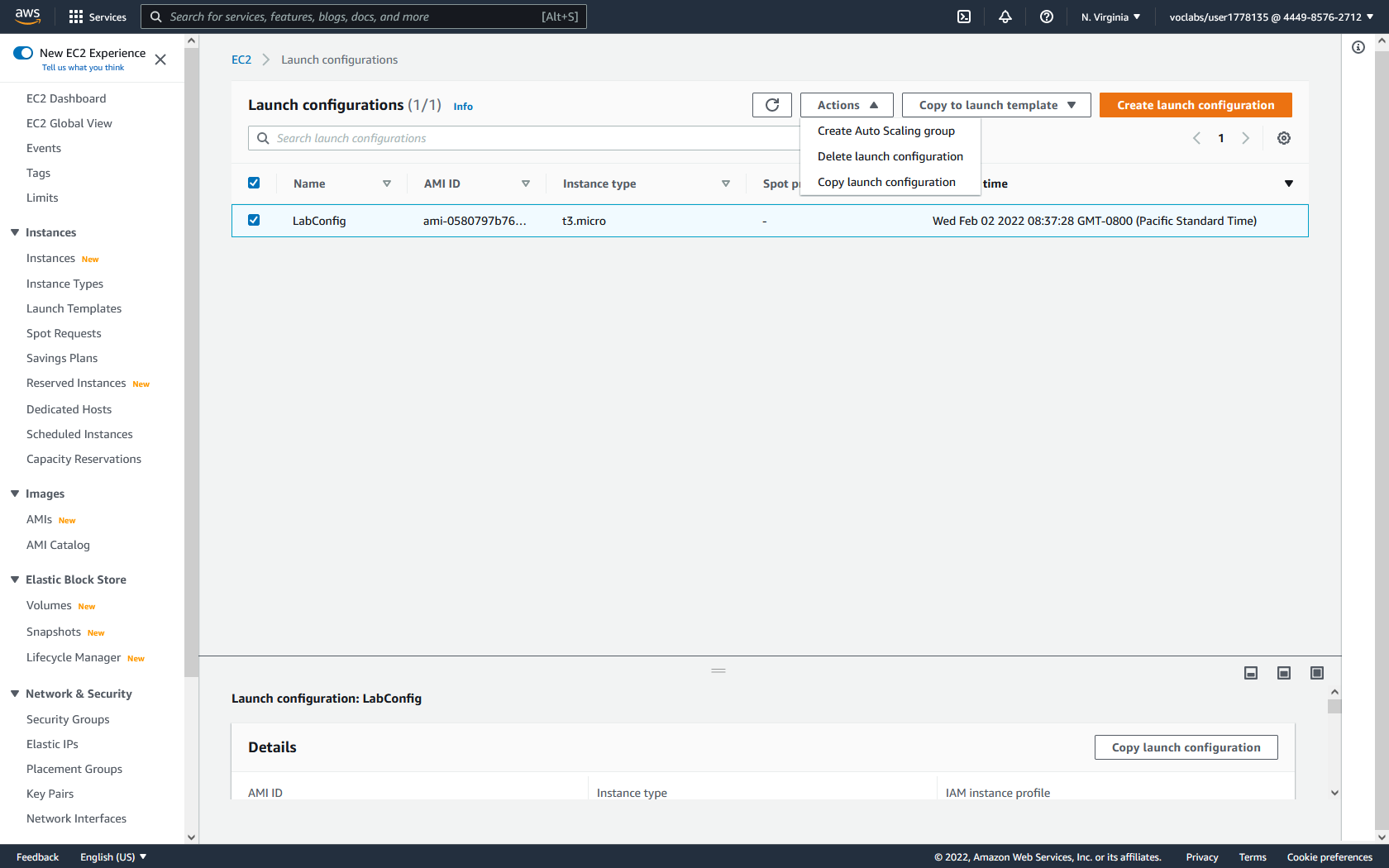
Maximum Capacity: 6

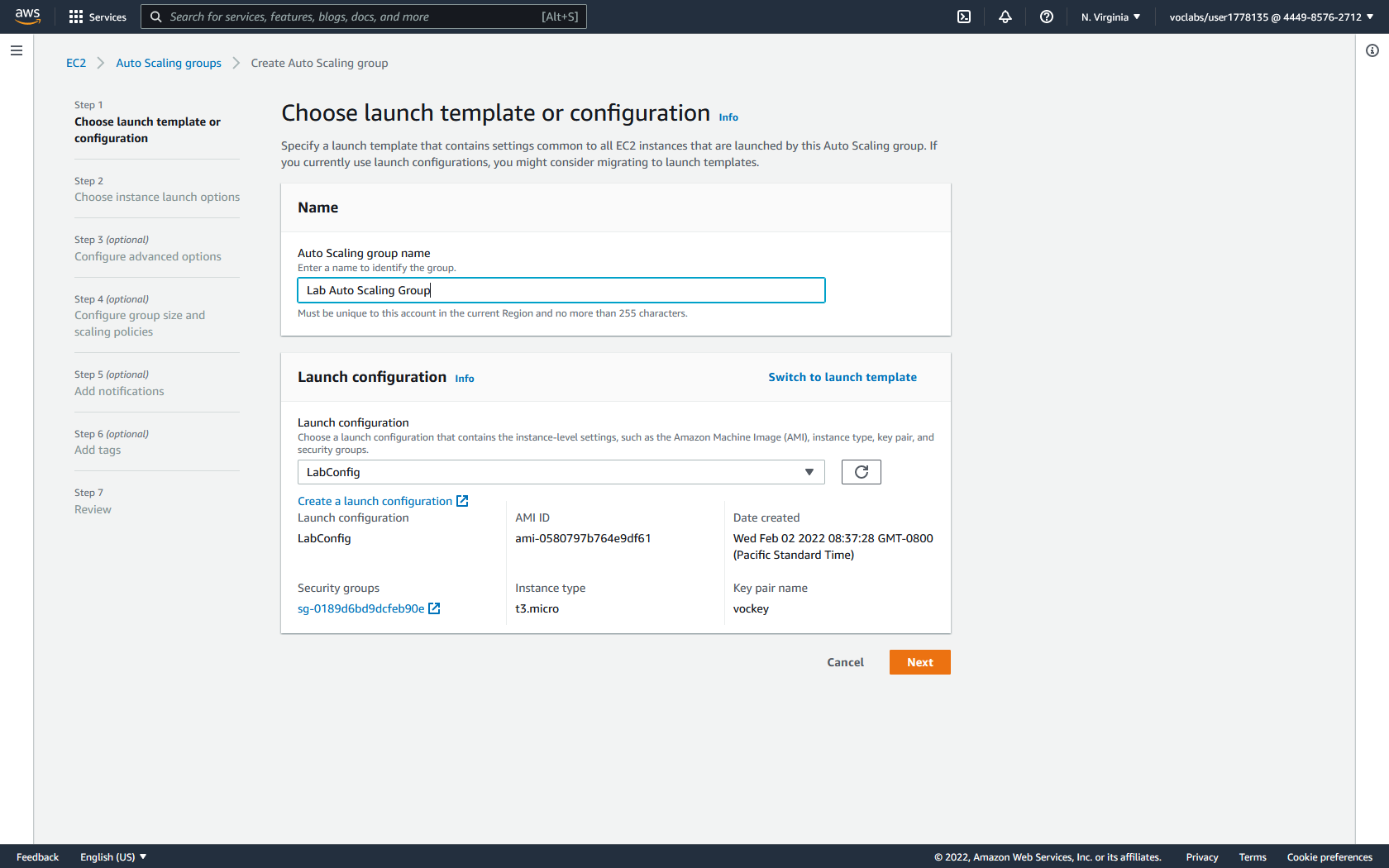
6. Under Scaling policies, create a target tracking scaling policy called LabScalingPolicy that uses Average CPU utilization of 60. Then, click next.

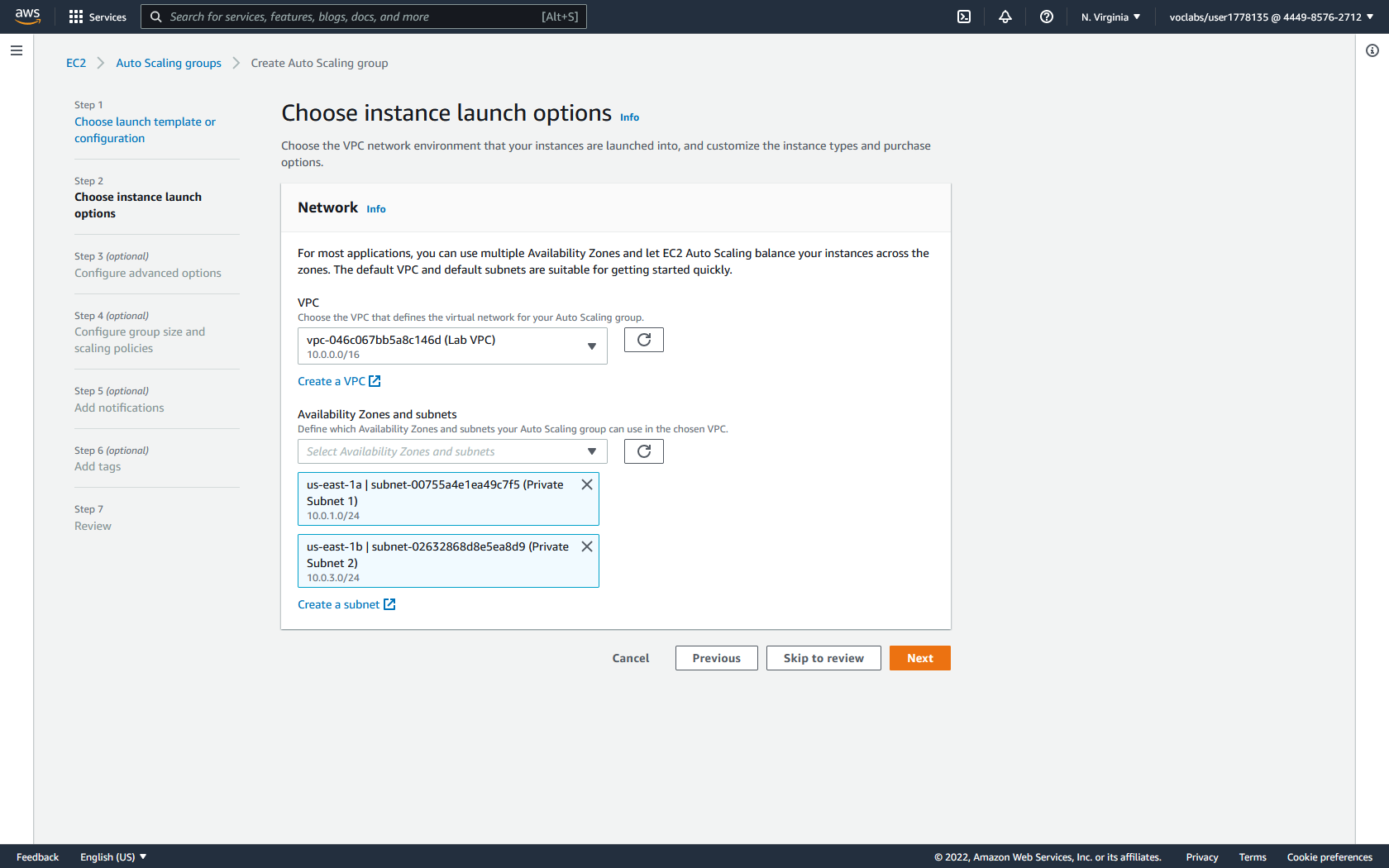
7. Notifications give optional information. This step can be skipped. Click next.

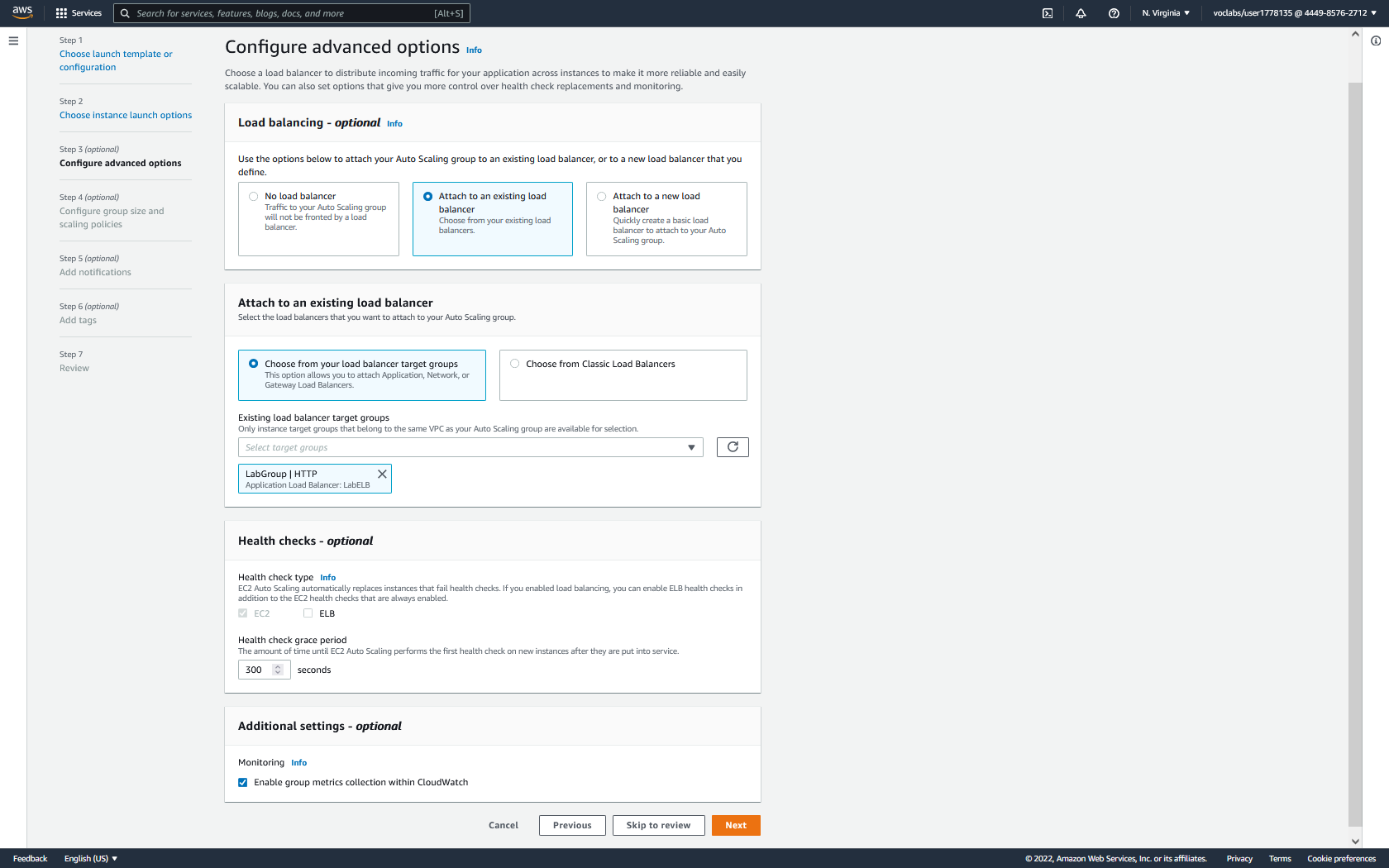
8. Tags give optional organizational marking for easier documentation and further configuration. Create the Key / Value pair of Name / Lab Instance and check tag new instances. Then, click next.

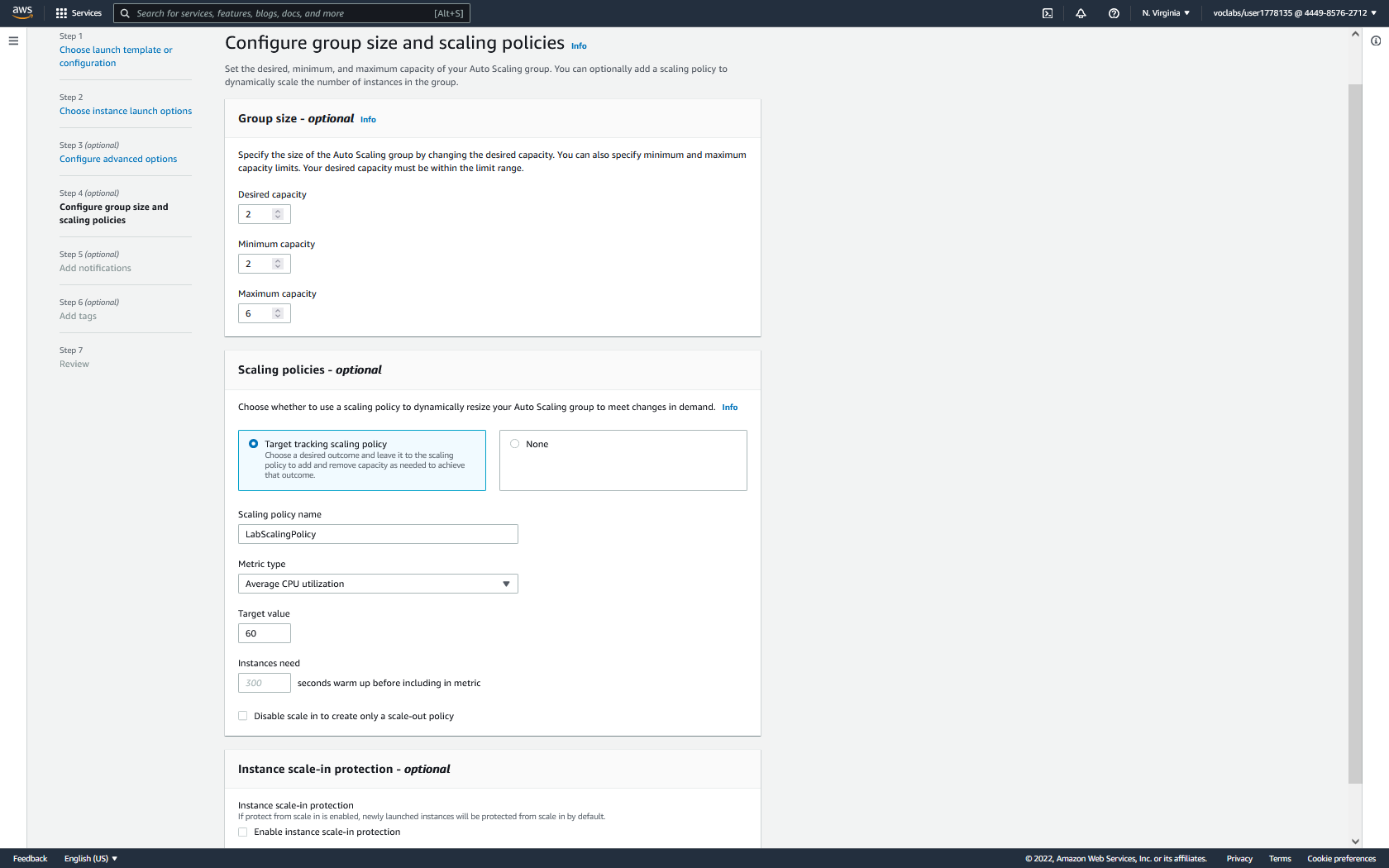
9. Review the configuration and click Create Auto Scaling Group.

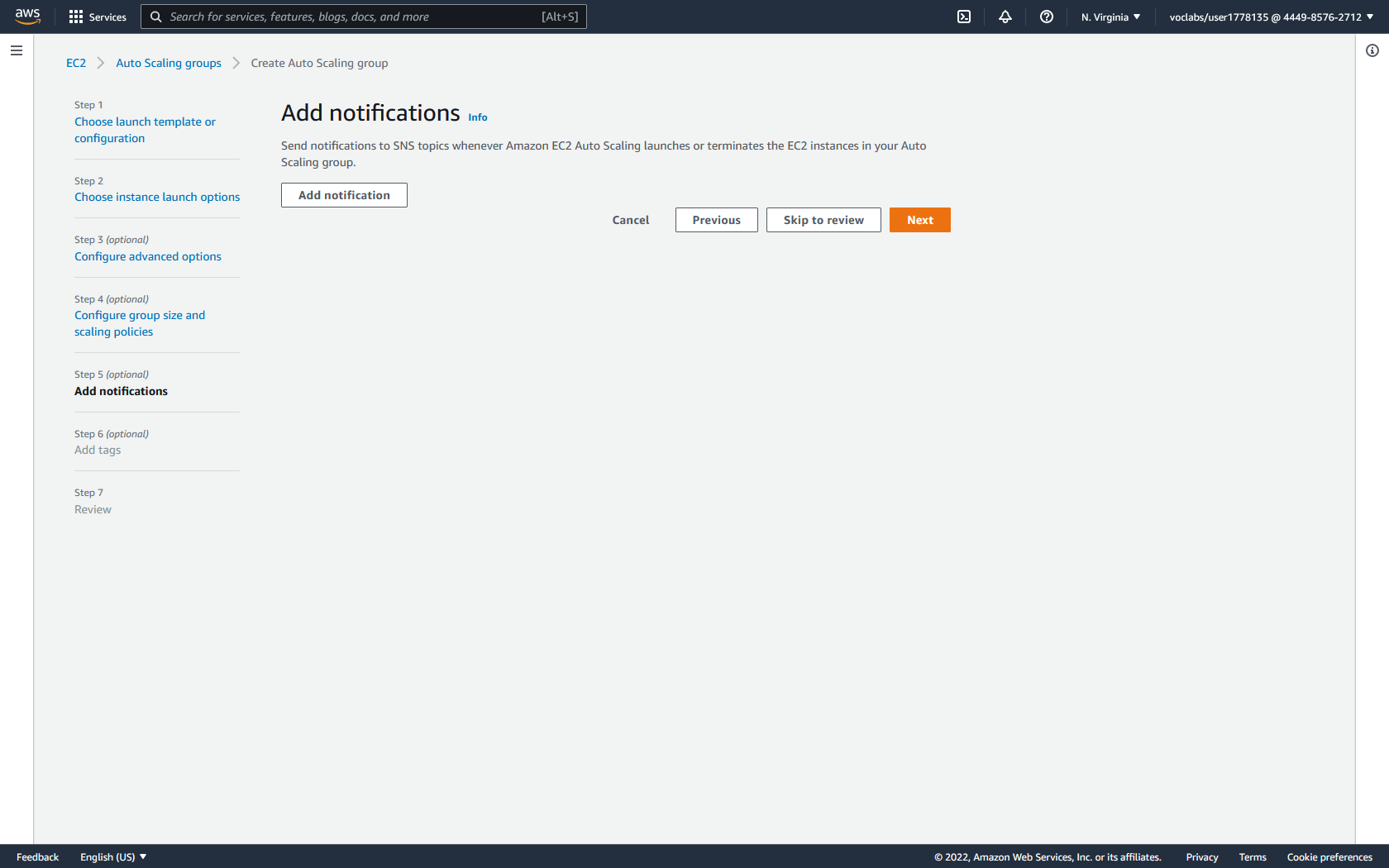


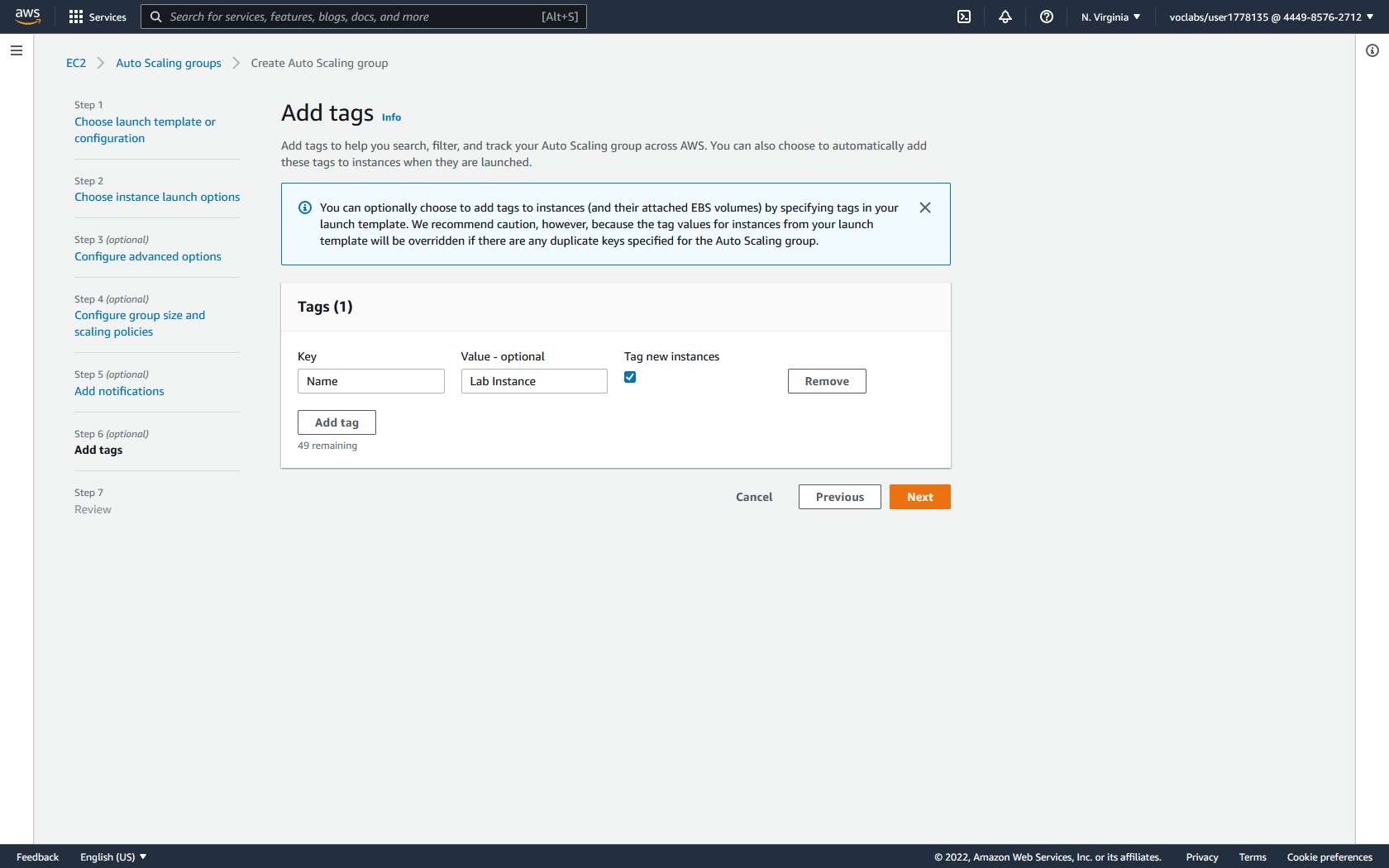


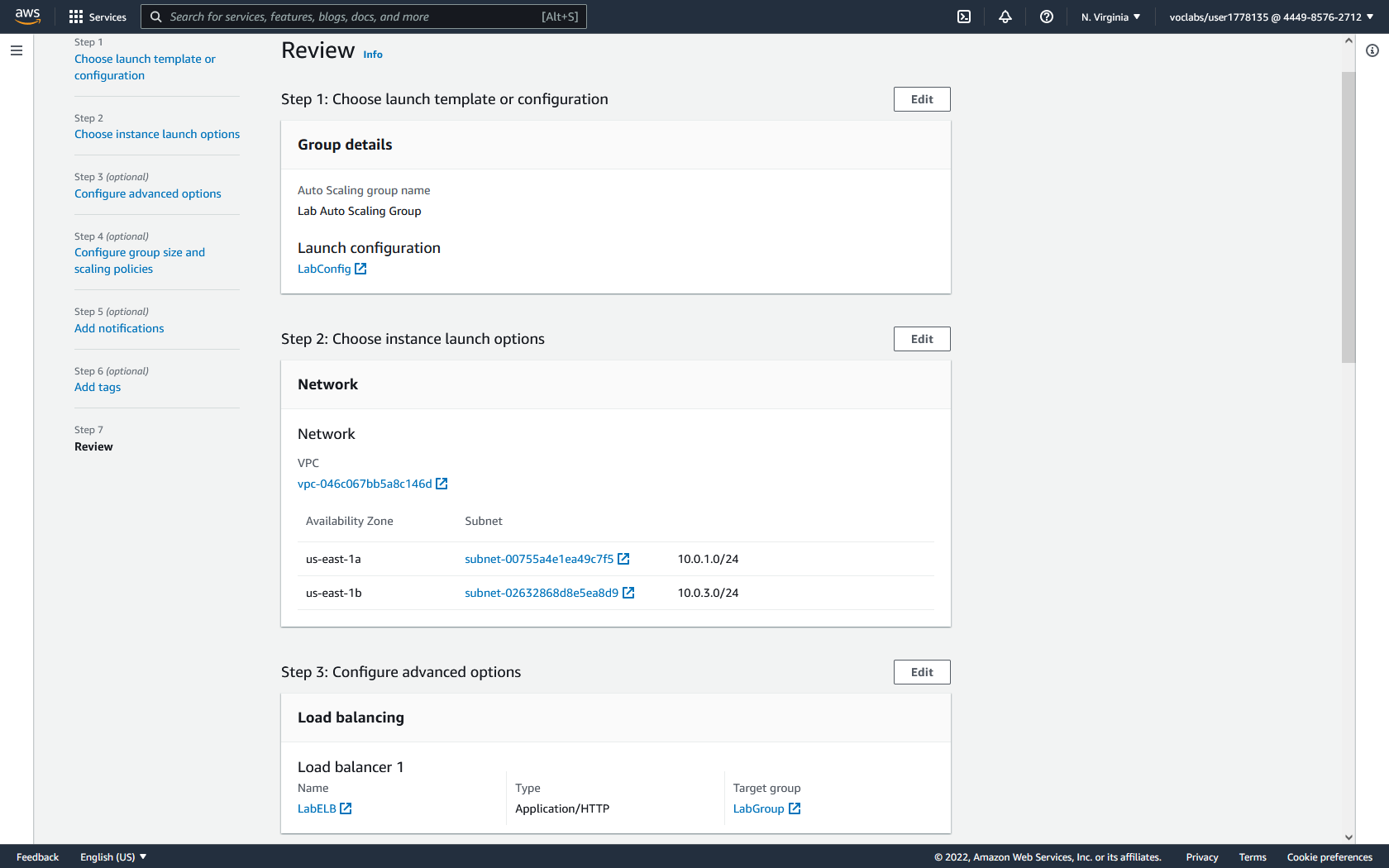


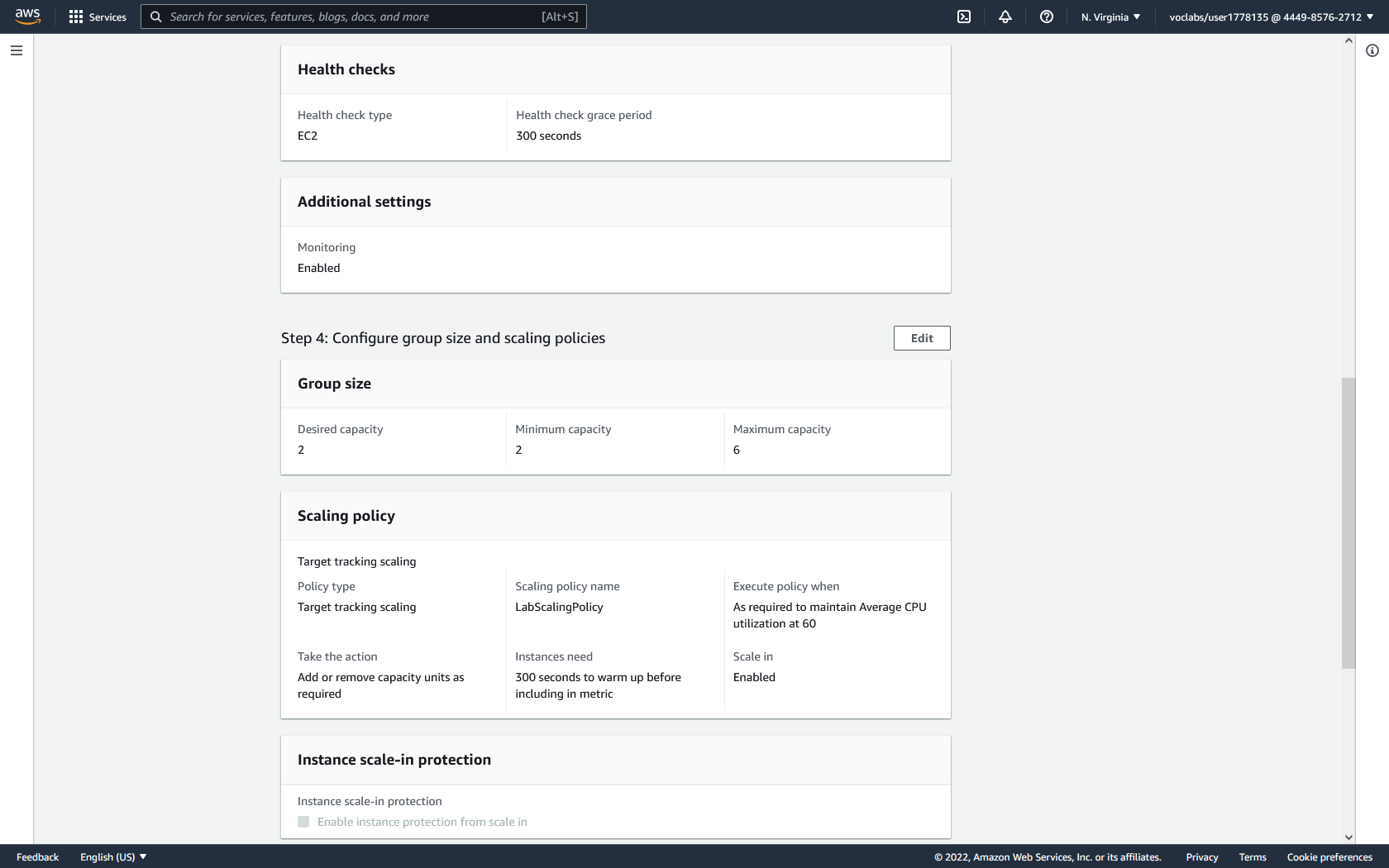


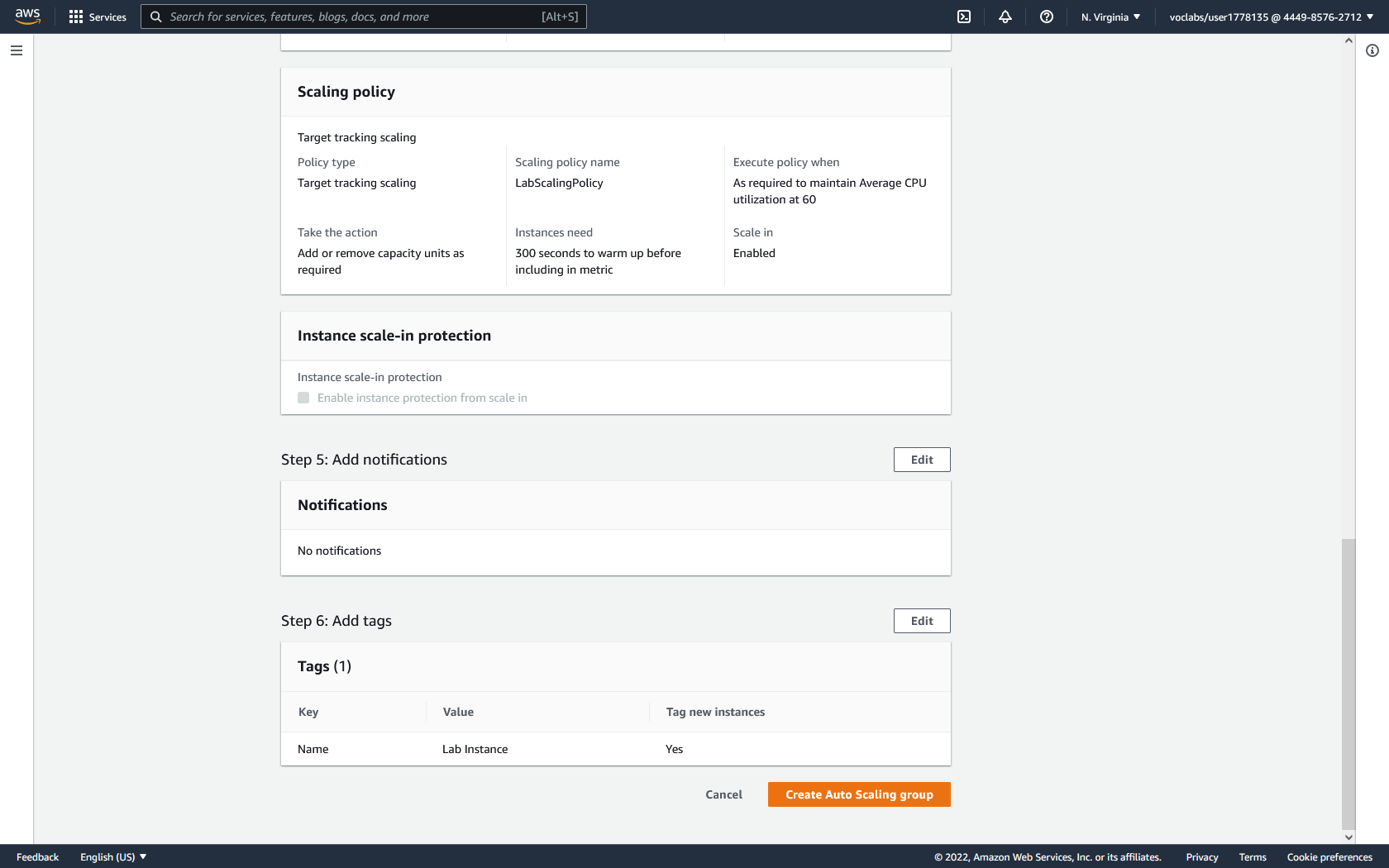


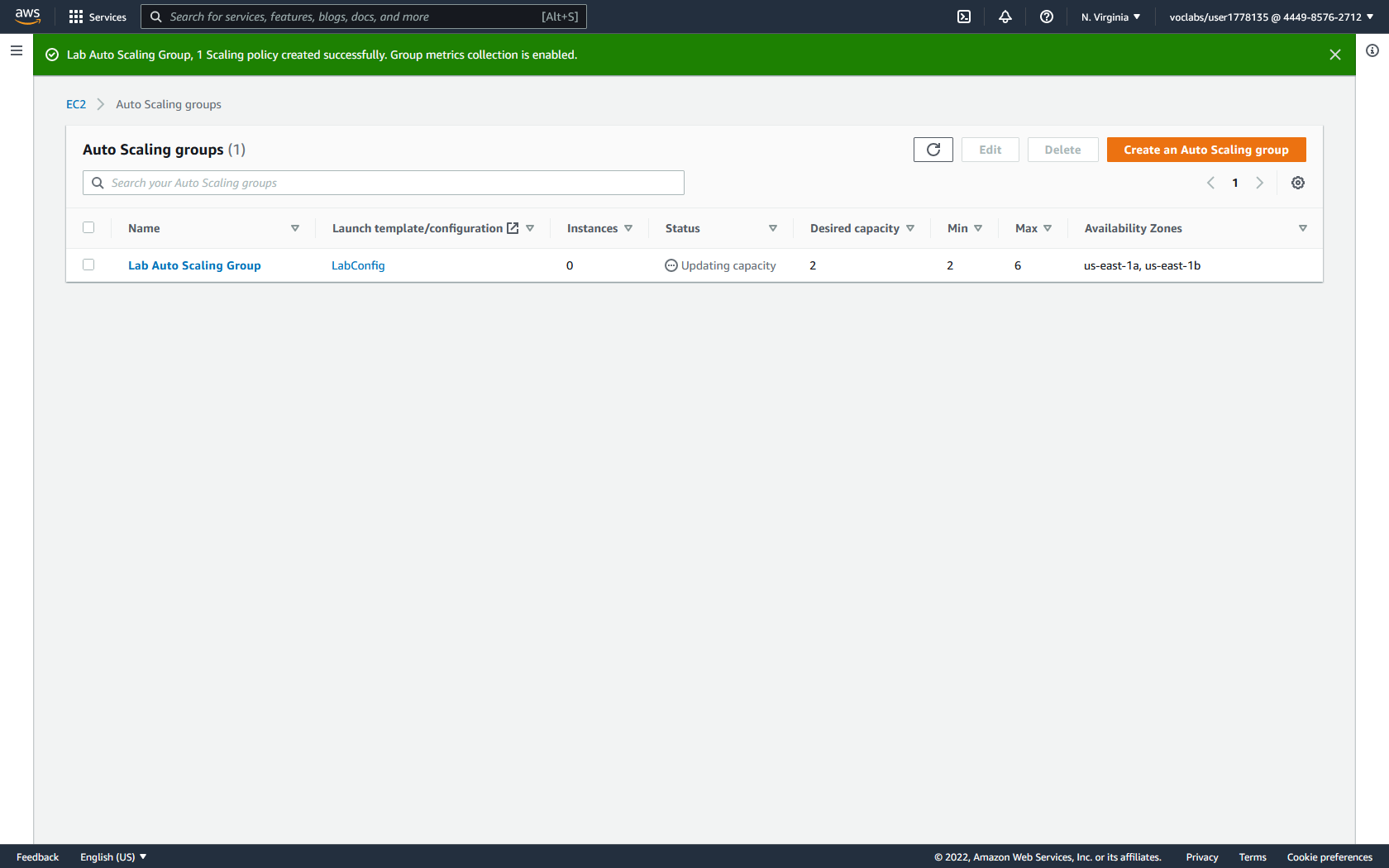


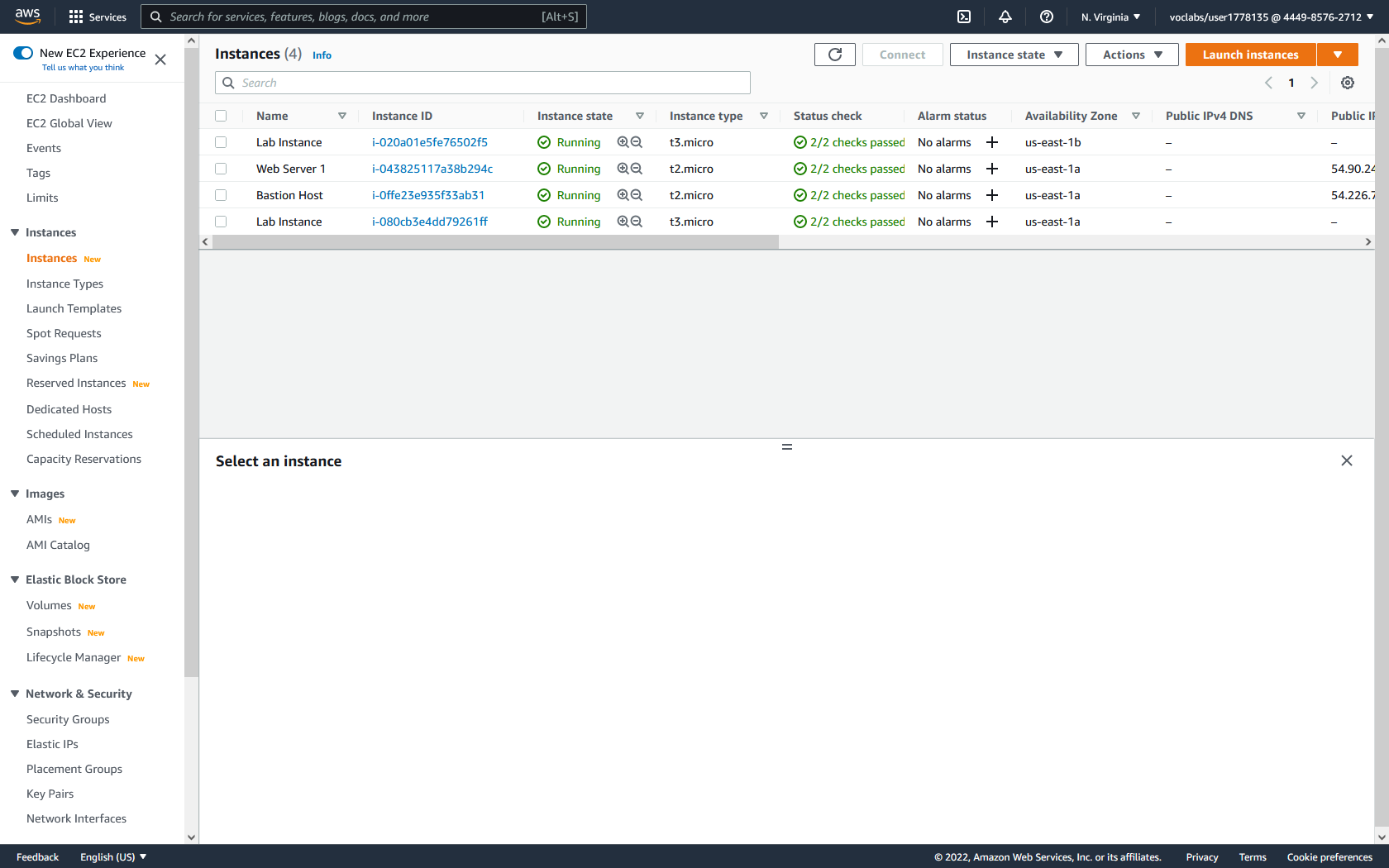










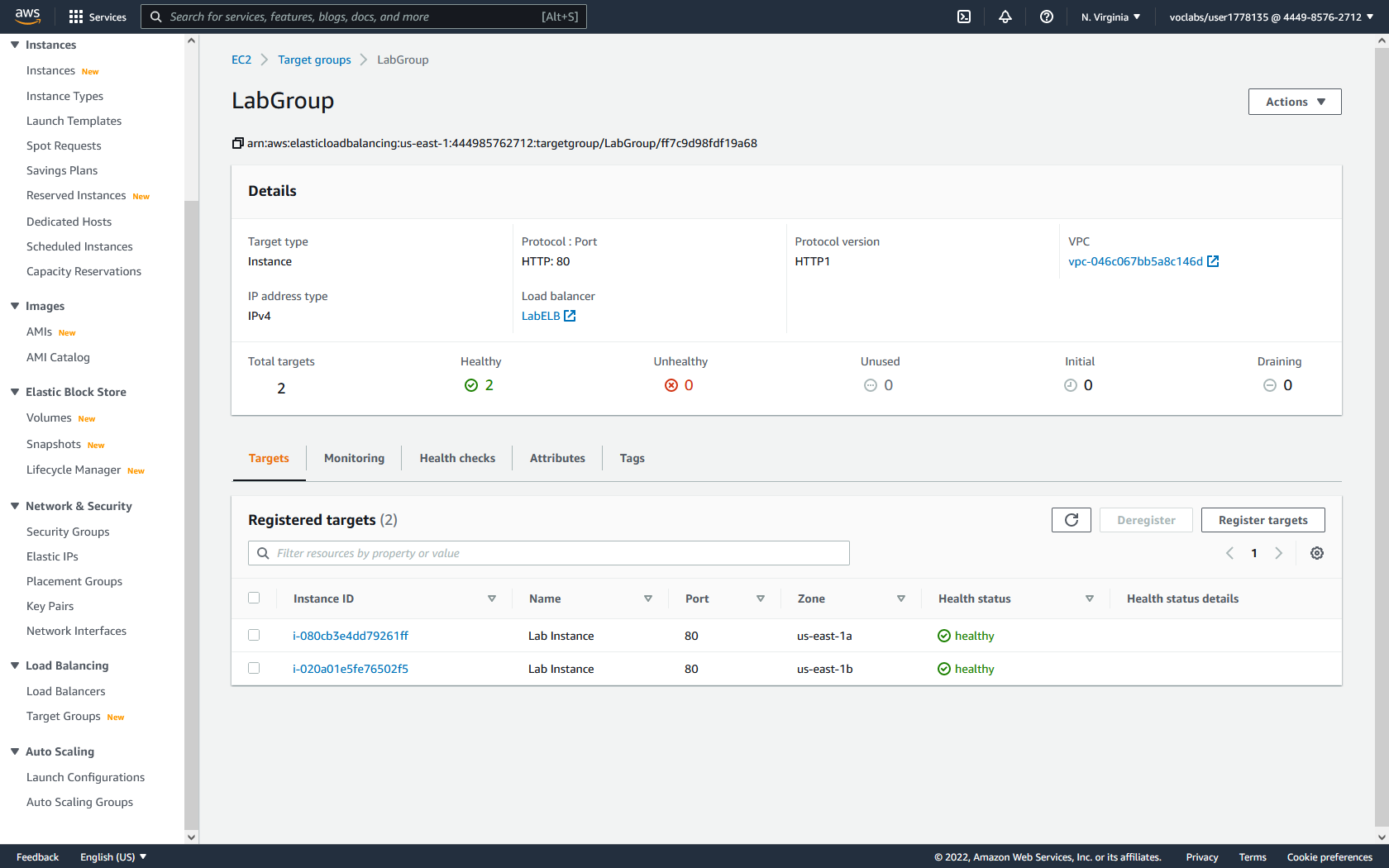


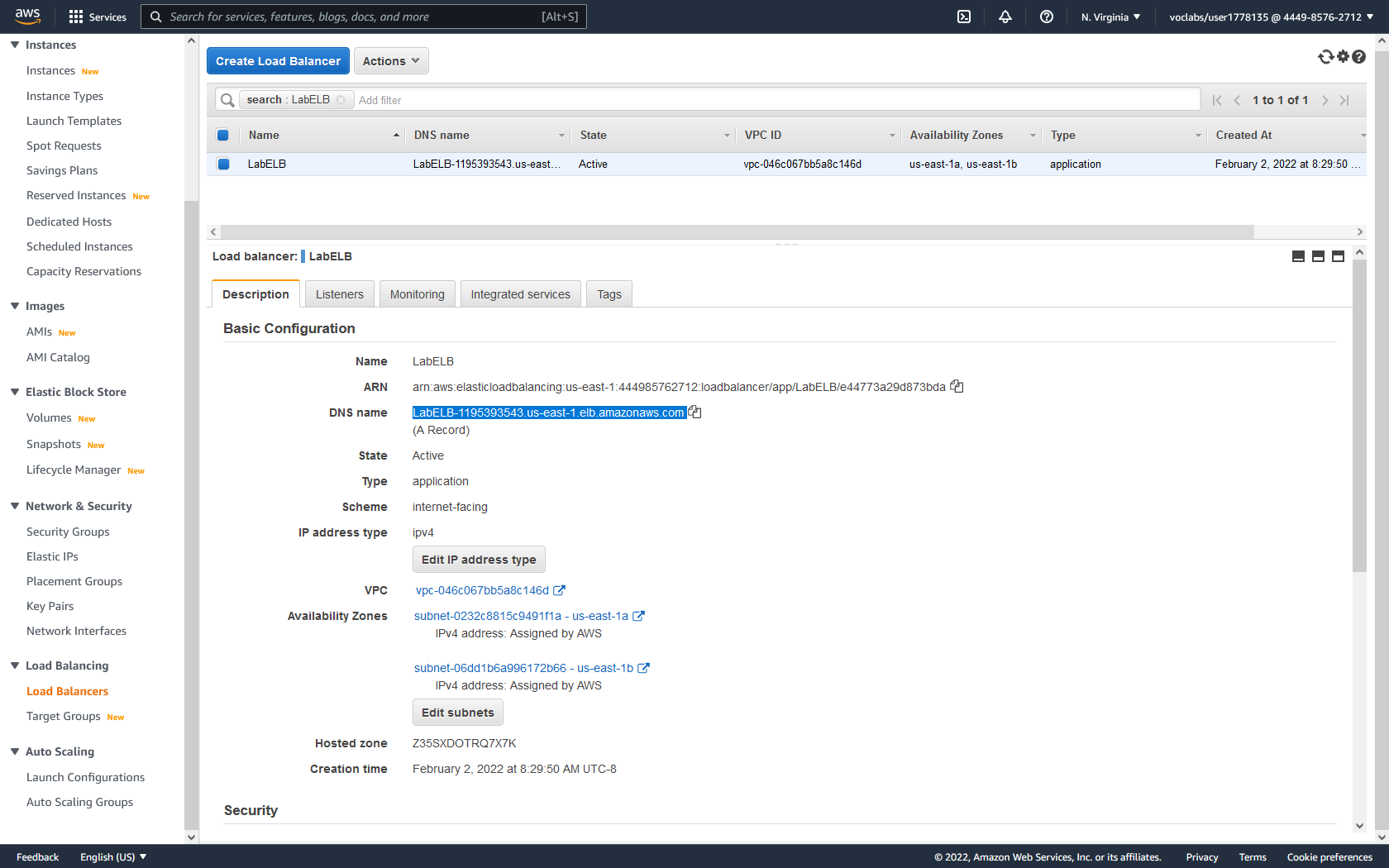
**Verify Load Balancing:**

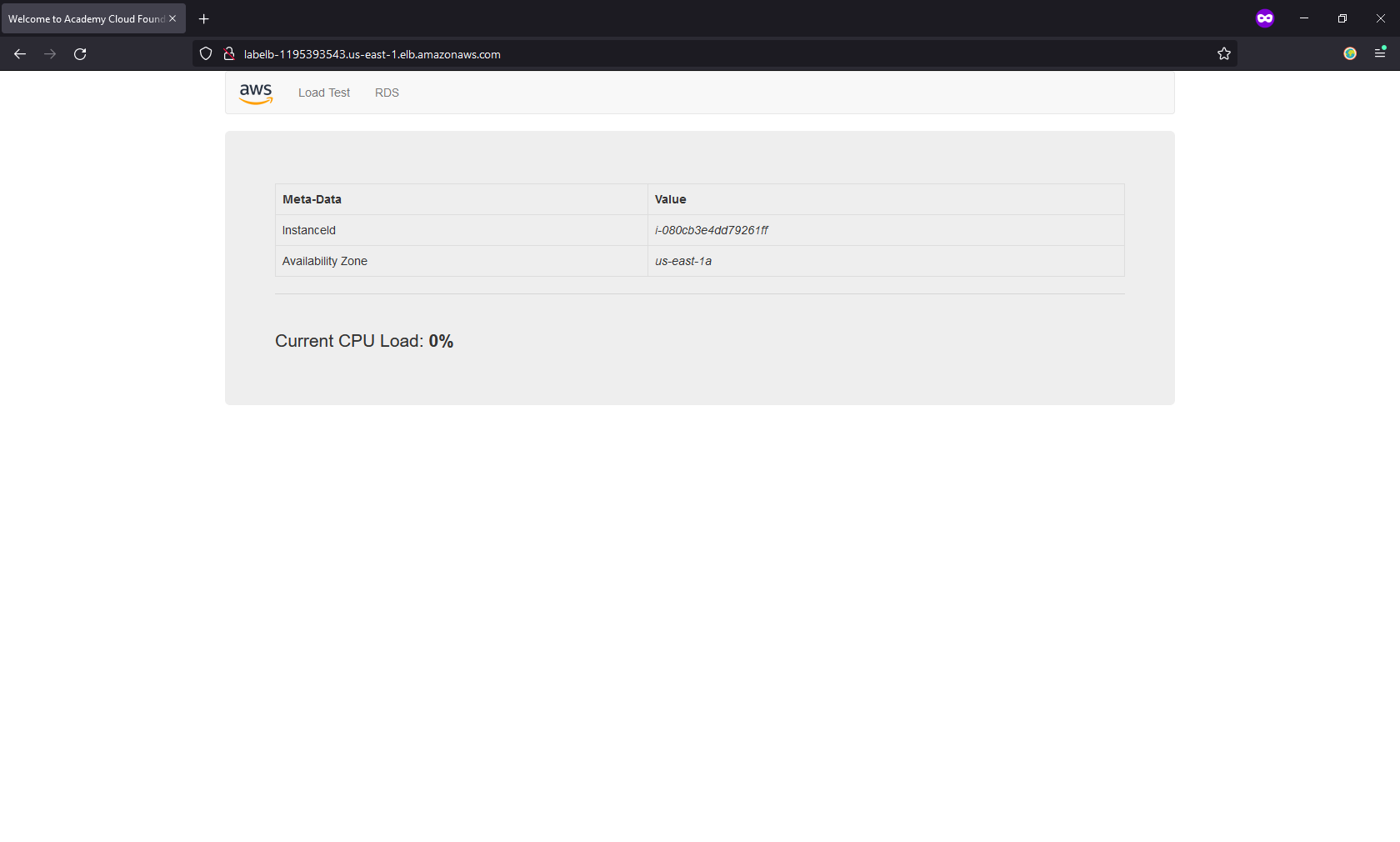
1. In the left menu, click target groups. Click into the previously created LabGroup target group to view further details. There should be two instances under registered targets.

2. In the left menu, click load balancers. Select the previously created LabELB load balancer to view further details. Under the description tab, copy the DNS name.

3. Open a web browser and paste the DNS name. A web application should open. Notice that the web application displays the instance ID and availability zone of one of the previously seen lab instances.







**Verify Auto Scaling:**

1. Locate CloudWatch by typing CloudWatch in the services search bar at the top of the page. Then click CloudWatch.

2. In the left menu, click all alarms to reach the Alarms dashboard. Click into the alarm with AlarmHigh in its name to view further details. This is the alarm that signals when CPU utilization goes above 60%.

3. Return to the Web Server application and click Load Test in the top banner. This should generate CPU load greater than 60%.

4. Return to the Alarms dashboard. Wait and refresh the dashboard until the alarm with AlarmHigh signals that it is in alarm.

5. In the EC2 service, reach the instance dashboard. There should be a new Lab Instance initializing automatically.

