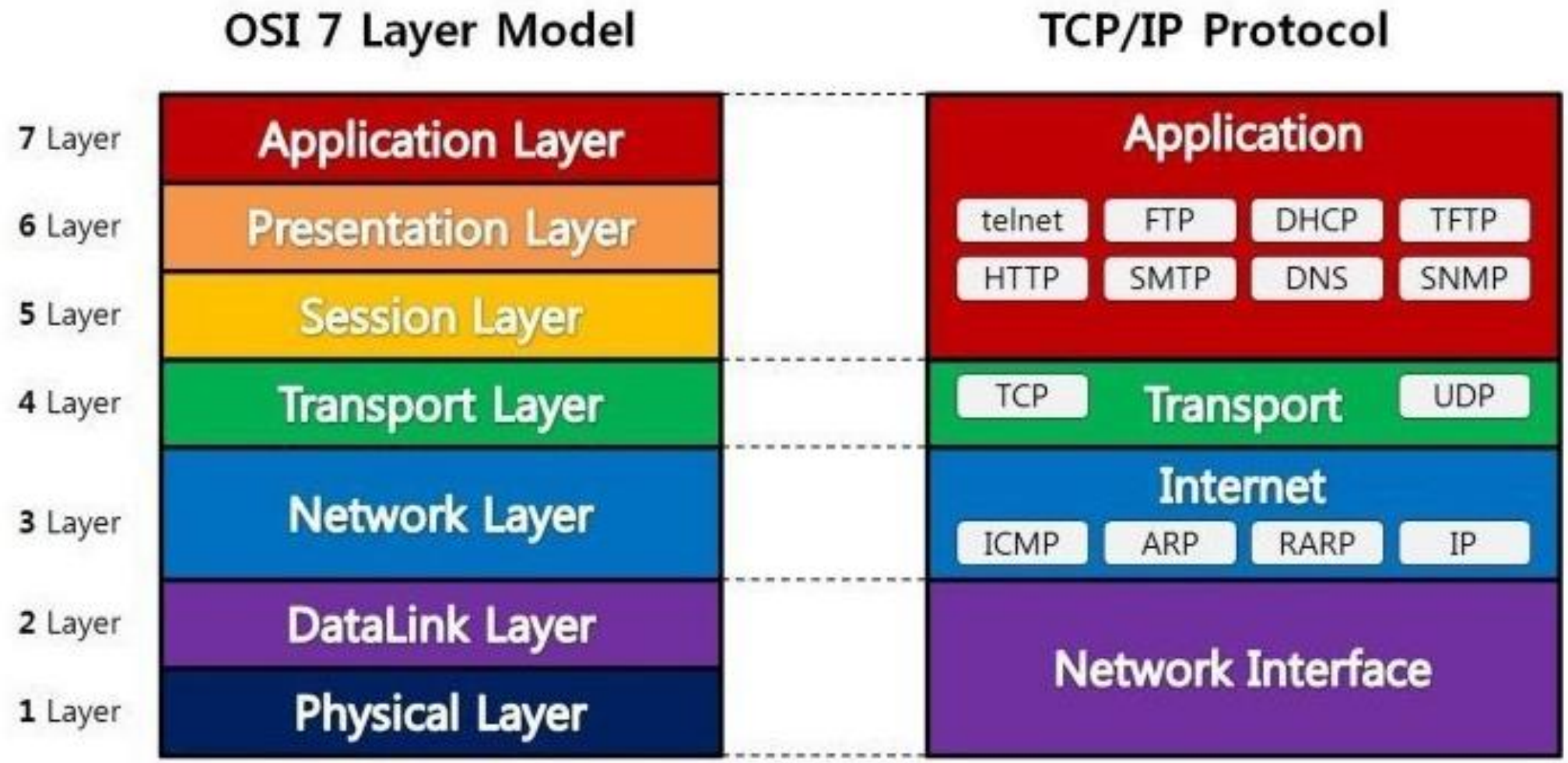
Amazon EC2 AutoScaling

박경규

OSI 7 Layer

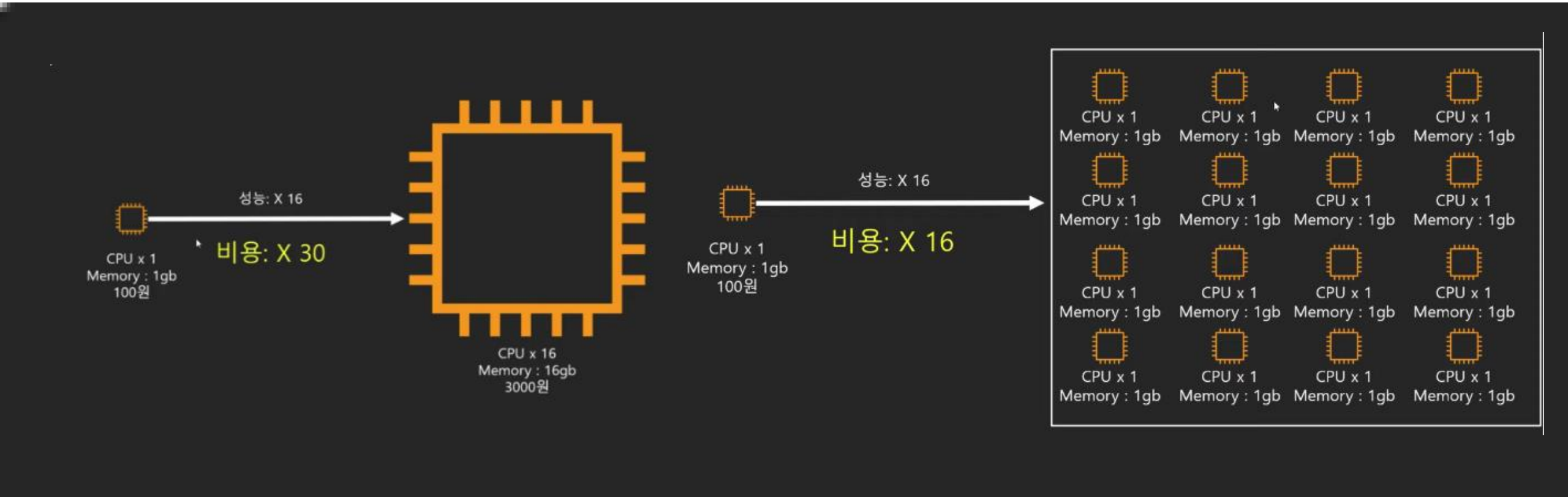

OSI 7 Layer Model
TCP/IP Protocol
7 Layer
6 Layer
5 Layer
4 Layer
3 Layer
2 Layer
1 Layer
Application Layer
Presentation Layer
Session Layer
Transport Layer
Network Layer
DataLink Layer
Physical Layer
Application
telnet
FTP
DHCP
TFTP
HTTP
SMTP
DNS
SNMP
TCP
Transport
UDP
Internet
ICMP
ARP
RARP
IP
Network Interface

확장성 (Scalability)

클라우드 환경에서는 Scale Out을 항상 염두하며 설계를 해야 합니다.
수요에 따라 인스턴스를 덜 쓸 수도, 더 쓸 수도 있음으로써 유연성을 가질 수 있습니다.

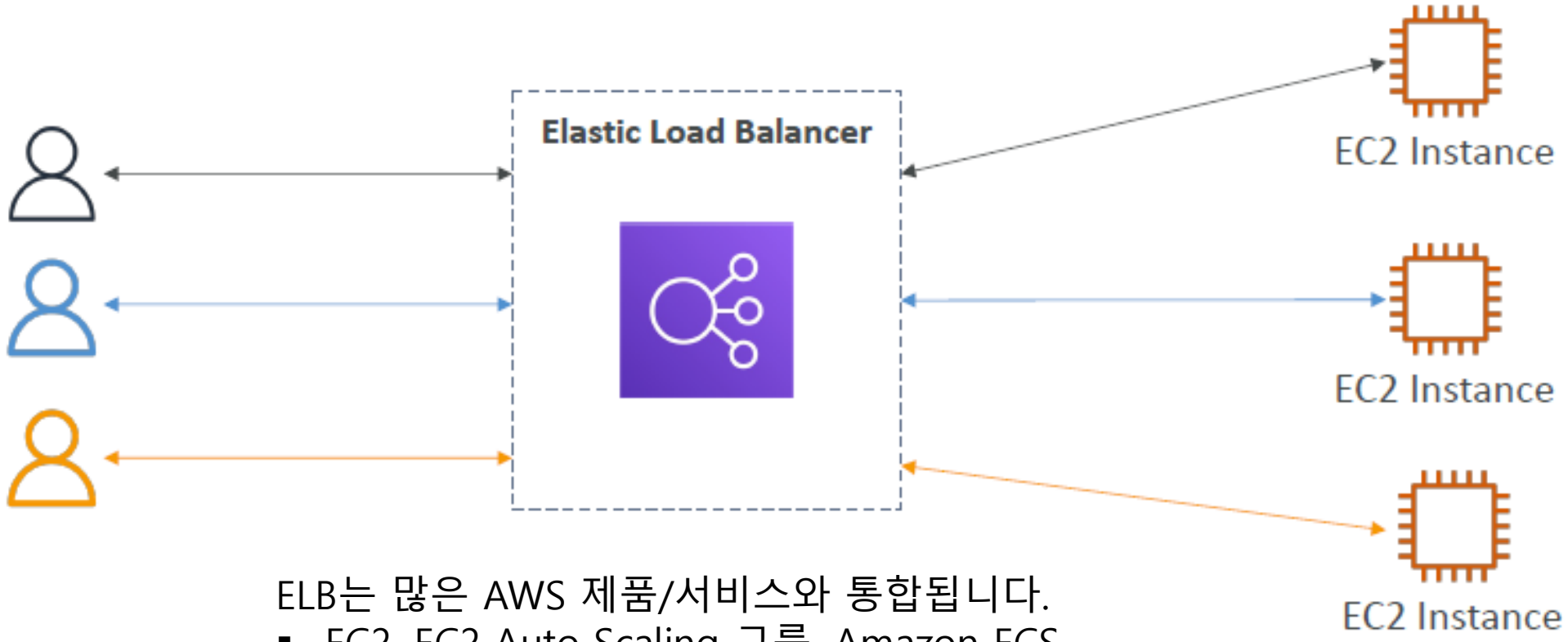
■ Vertical Scalability(스케일 업 : Scale Up)

■ Horizontal Scalability(= elasticity, 스케일 아웃 : Scale Out))



로드밸런서 (ELB : Elastic Load Balancer)

Load Balancer는 트래픽을 여러 서버 (예: EC2 인스턴스) 다운스트림으로 전달하는 서버입니다.
Elastic Load Balancer는 관리형 Load Balancer 입니다.

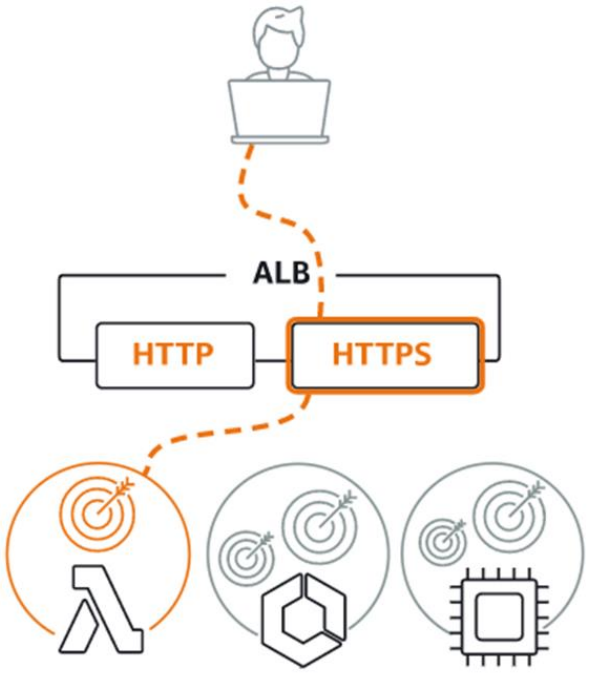


ELB는 많은 AWS 제품/서비스와 통합됩니다.

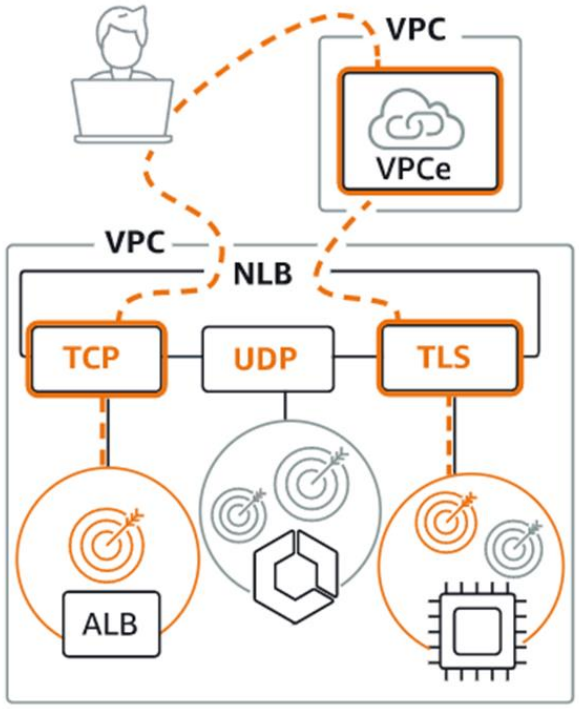
- EC2, EC2 Auto Scaling 그룹, Amazon ECS
- AWS 인증서 관리자(ACM), CloudWatch
- Route 53, AWS WAF, AWS 글로벌 액셀러레이터

로드밸런서(ELB) 종류

Application Load Balancer Info

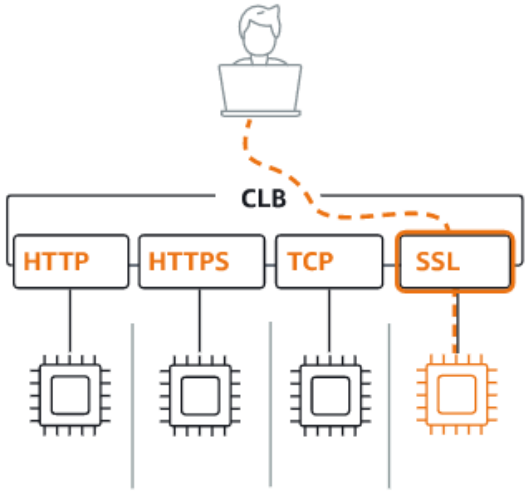


Network Load Balancer Info



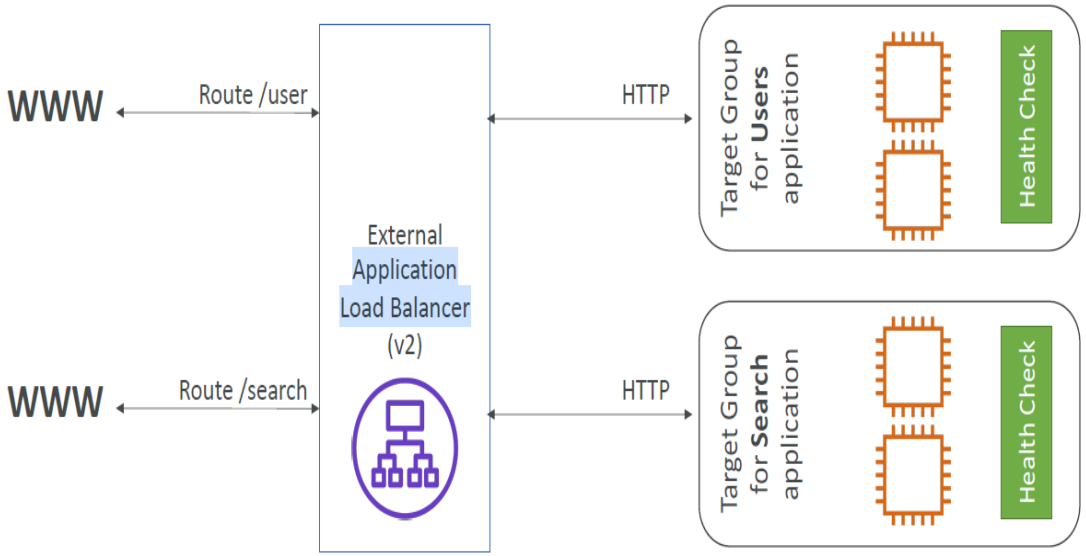
Gateway Load Balancer Info

Classic Load Balancer Info

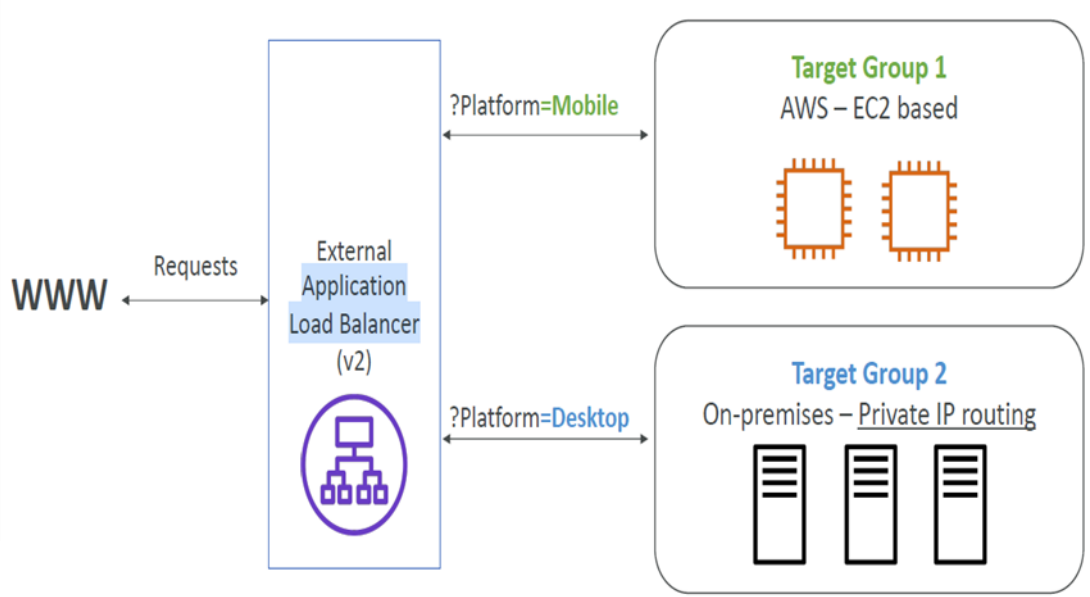


로드밸런서 – ALB(Application Load Balancer)

HTTP Based Traffic



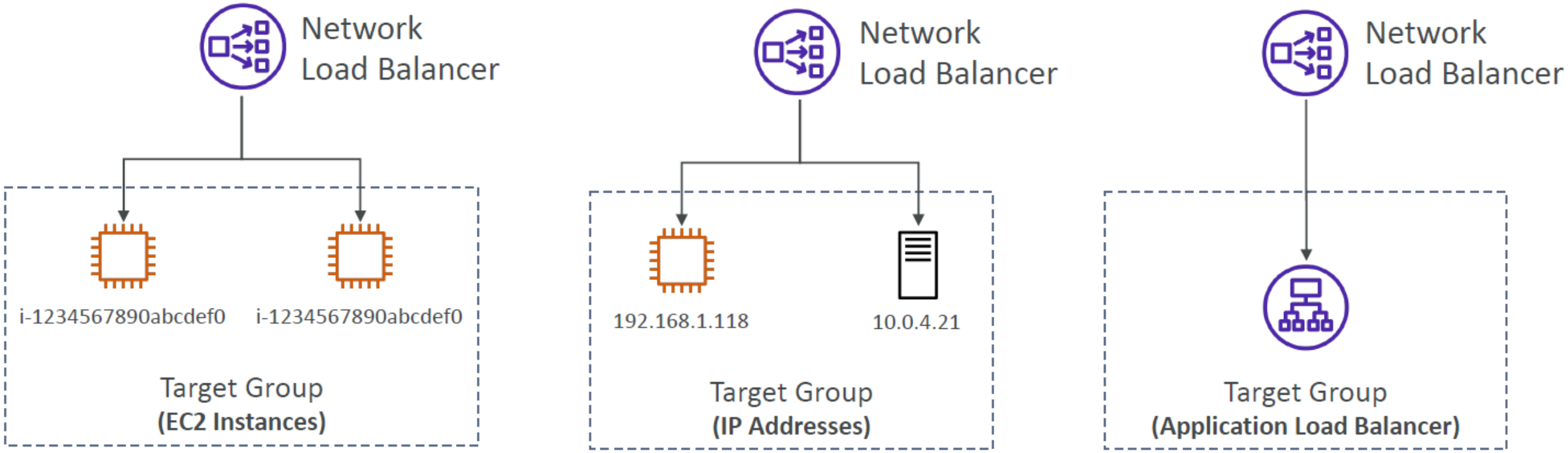
Query Strings/Parameters Routing



- ALB는 Layer 7(HTTP)에서 동작
- 여러 대상 그룹에서 여러 HTTP 애플리케이션에 대한 로드 밸런싱
- 동일한 시스템(예: 컨테이너)의 여러 애플리케이션에 대한 로드 밸런싱
- HTTP/2 및 WebSocket 지원
- ALB는 마이크로 서비스 및 컨테이너 기반 애플리케이션(예: Docker 및 Amazon ECS)에 매우 적합
- ECS의 동적 포트로 리디렉션하는 포트 매핑 기능이 있음

로드밸런서 – NLB(Network Load Balancer)

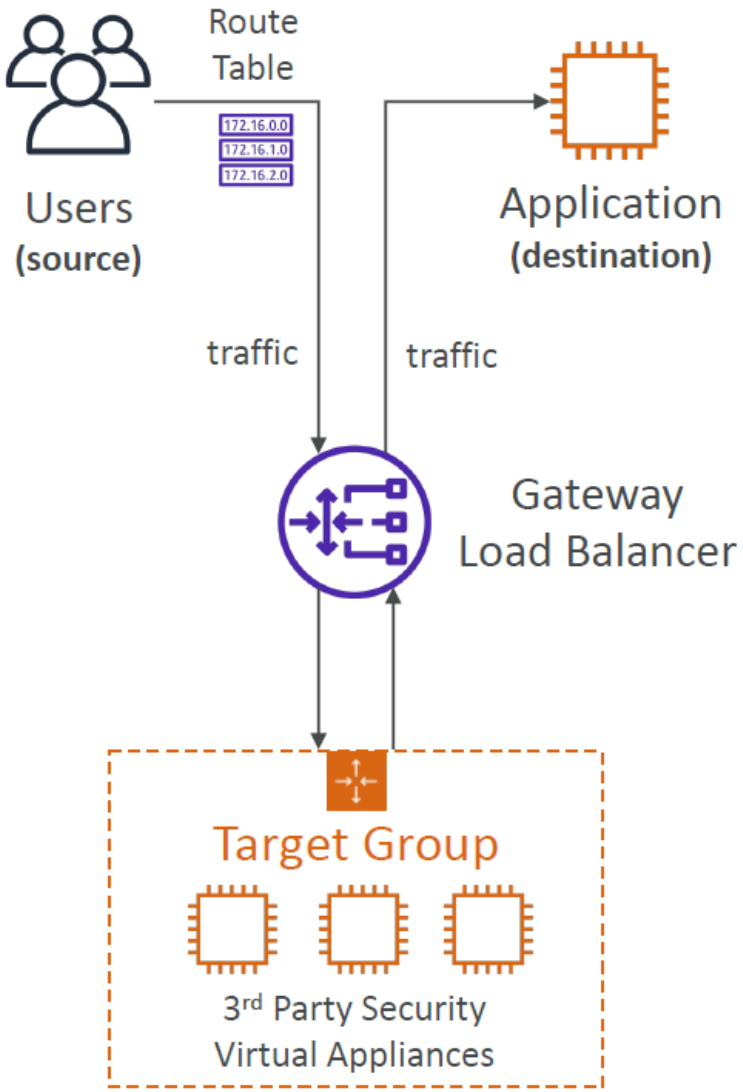
■ 타깃 그룹 : EC2, IP Address, ALB



- NLB는 Layer 4(TCP & UDP)에서 동작
- TCP 및 UDP 트래픽을 인스턴스로 전달
- 초당 수백만 건의 요청 처리
- ~100ms 미만의 지연 시간(ALB의 경우 400ms)
- NLB는 AZ당 하나의 고정 IP를 가지며 탄력적 IP 할당을 지원₩
- NLB는 고성능,TCP 또는 UDP 트래픽에 사용

로드밸런서 – GWLB(Gateway Load Balancer)

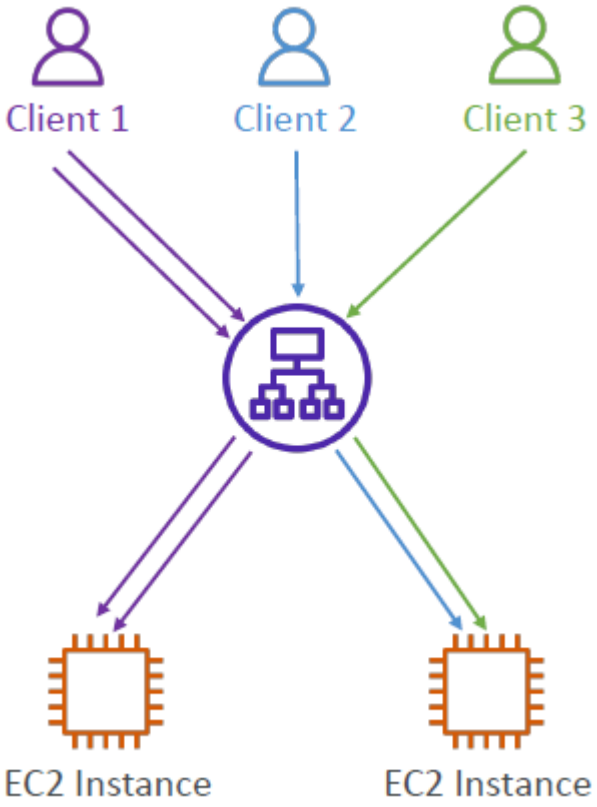
GWLB는 모든 트래픽에 대한 단일 entry/exit이며, traffic을 가상 어플라이언스로 distribute 합니다.



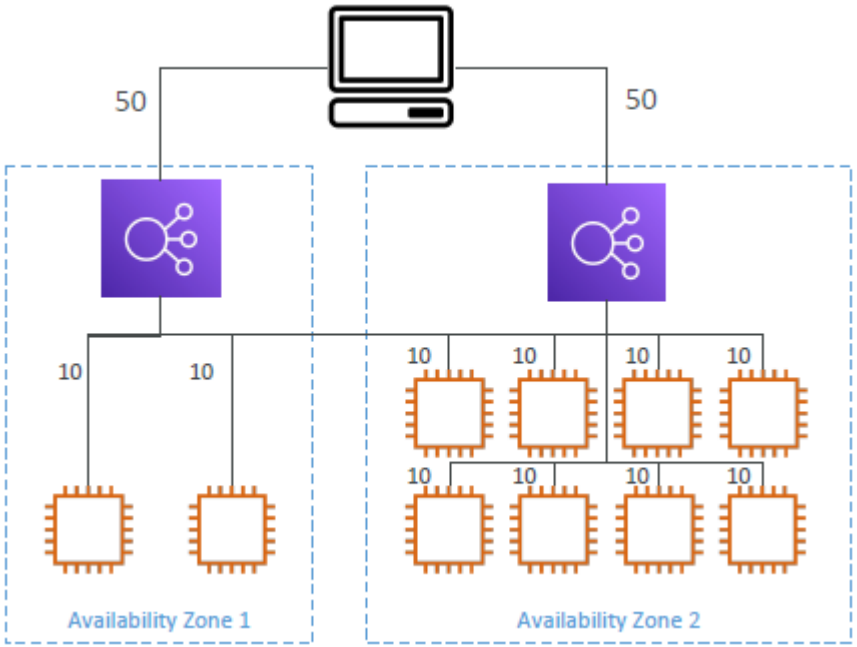
- GWLB는 Layer 3(IP 패킷)에서 동작
- Examples
  - Firewalls
  - Intrusion Detection and Prevention Systems
  - Deep Packet Inspection Systems
  - payload manipulation 등
- Transparent Network Gateway – single entry/exit for all traffic
- Load Balancer – distributes traffic to your virtual appliances

로드밸런서 – Sticky Session, Cross-Zone LB, SSL/TLS

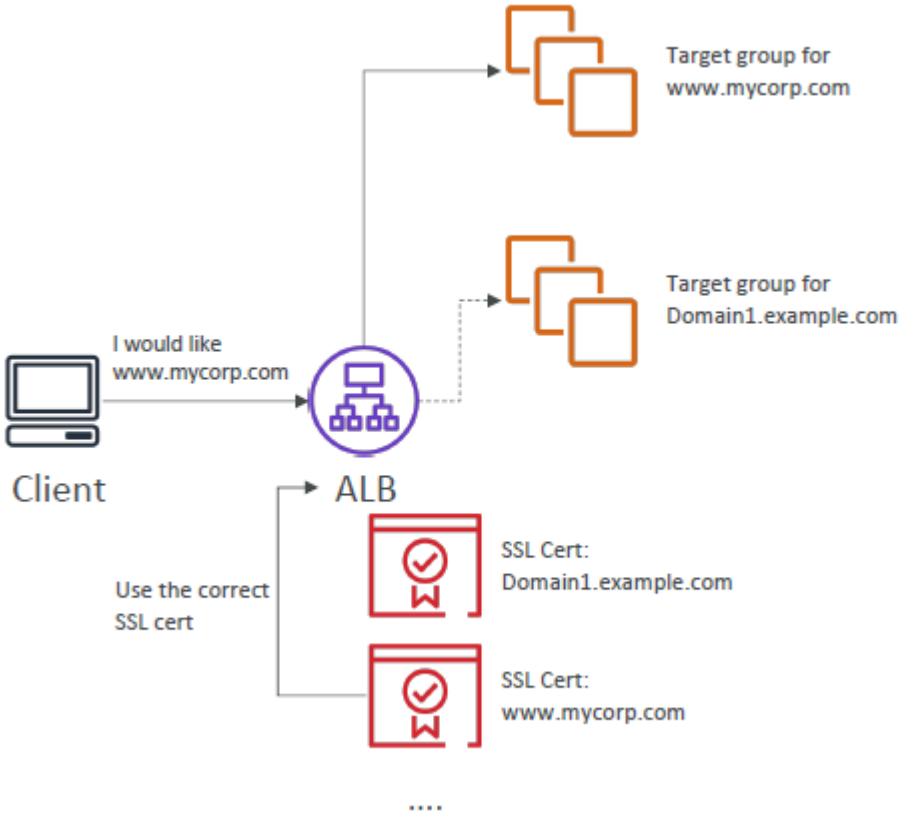
■ Sticky Sessions(Cookie Names)



■ Cross-Zone Load Balancing

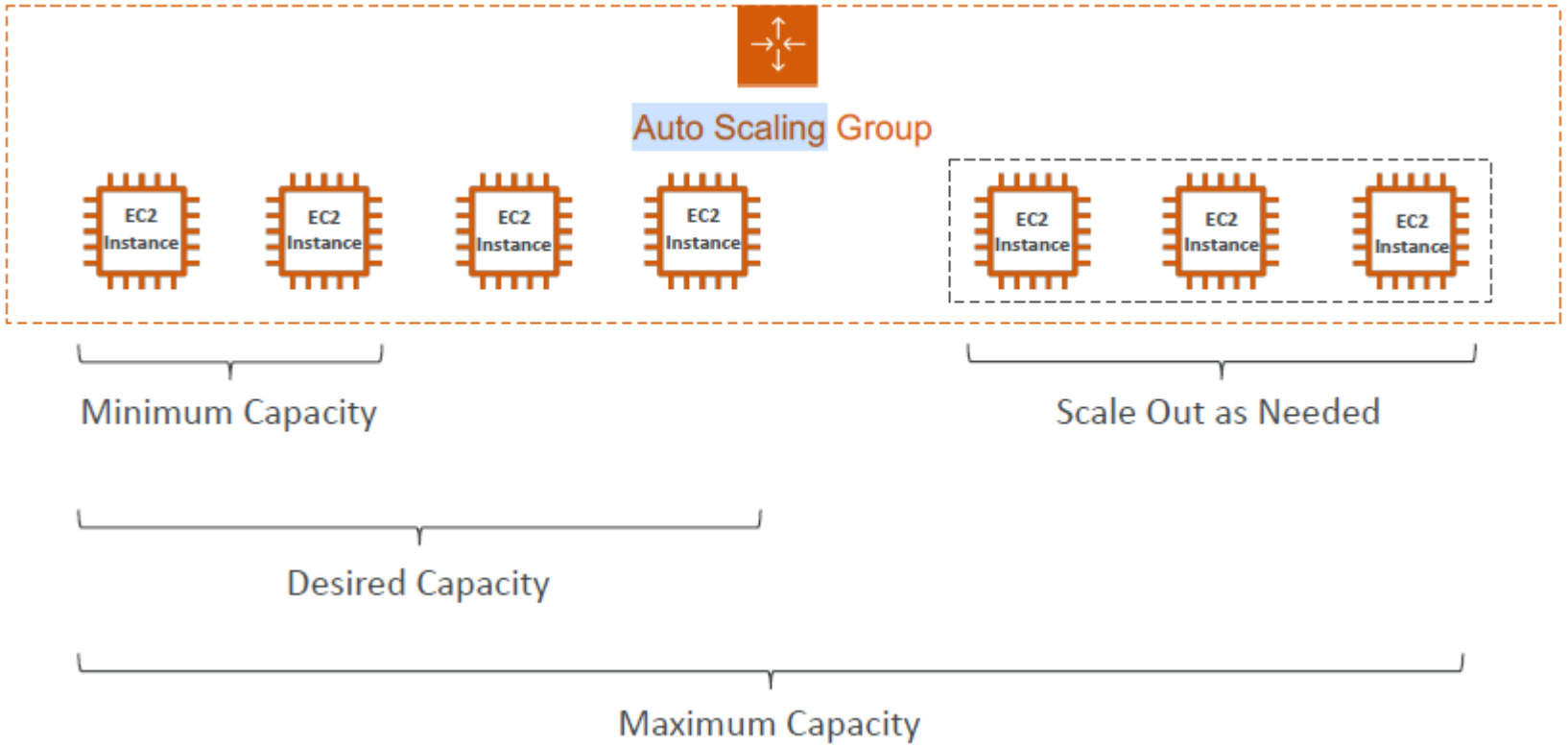


■ Server Name Indication (SNI)



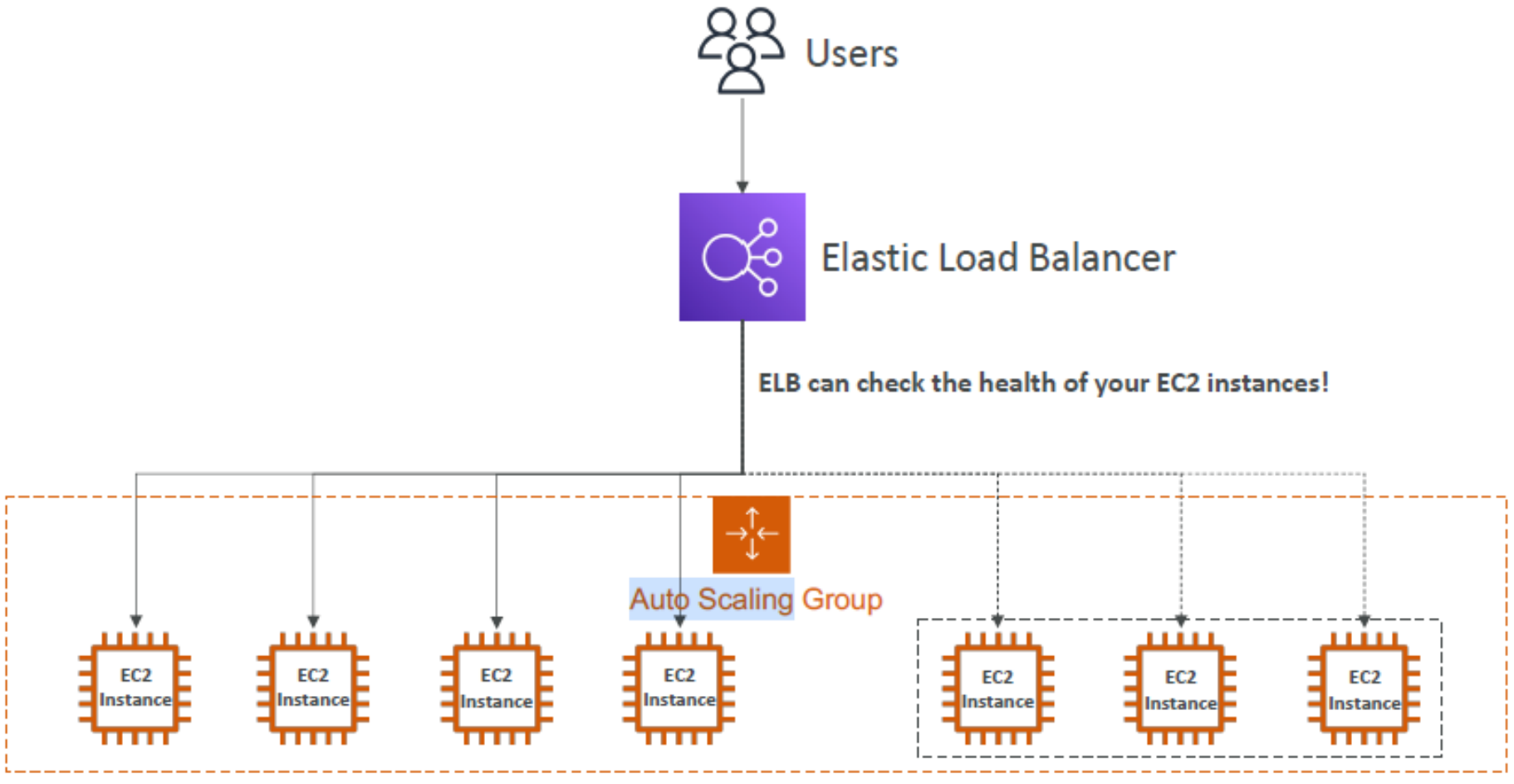
오토 스케일링 그룹(ASG : Auto Scaling Group)

ASG 를 생성할 때 EC2 인스턴스의 최소 및 최대 인스턴스 수와 원하는 인스턴스 수를 지정하고 이 범위안에서 Scale in/out 이 일어납니다.



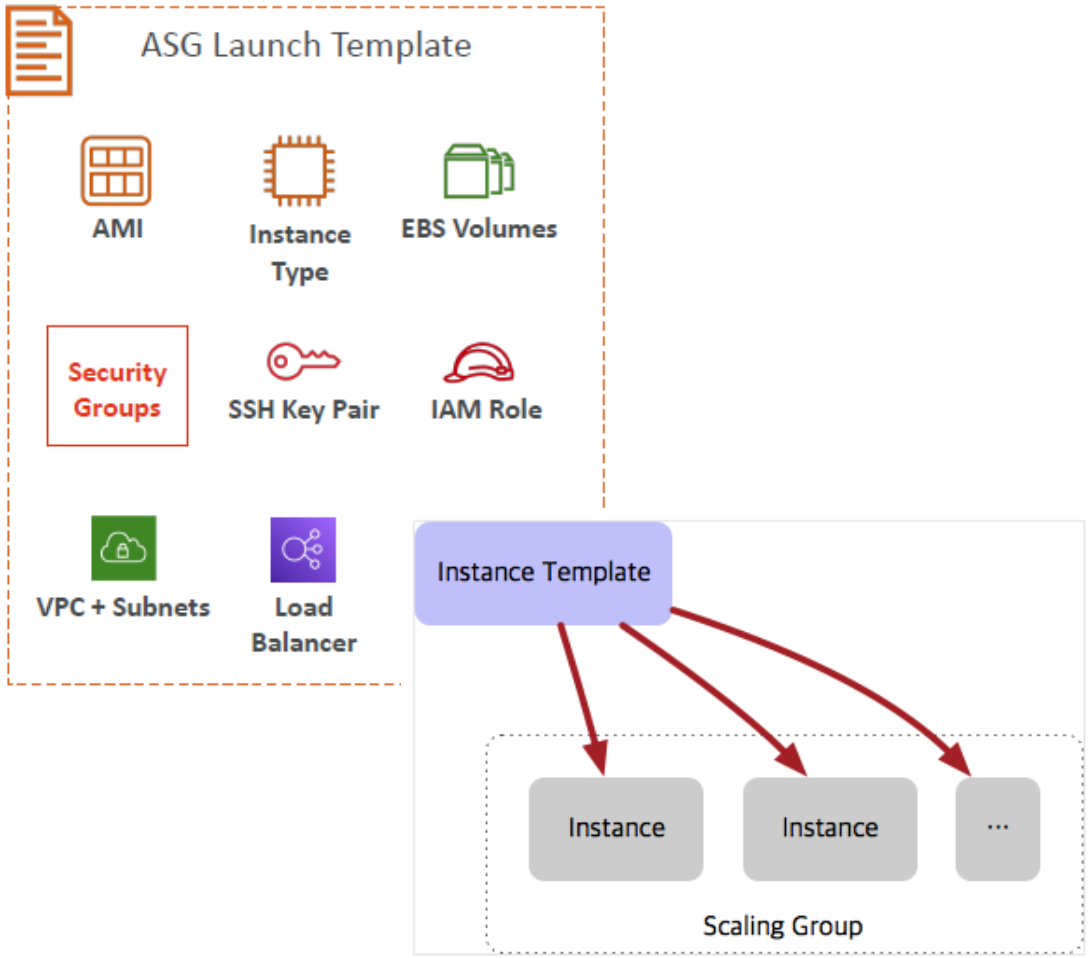
- 실행 중인 EC2 인스턴스의 최소 및 최대 수가 있는지 확인
- 인스턴스를 로드밸런서에 자동으로 등록
- 이전 인스턴스가 종료된 경우(예: 비정상인 경우) EC2 인스턴스를 다시 생성

ASG with ELB


Users
Elastic Load Balancer
ELB can check the health of your EC2 instances!
Auto Scaling Group
EC2 Instance
EC2 Instance
EC2 Instance
EC2 Instance
EC2 Instance
EC2 Instance
EC2 Instance

ASG – 시작 템플릿(Launch Template)

시작 템플릿(Launch Template)은 똑같은 환경의 인스턴스를 간편하게 복제하기 위해서 구성하는 것입니다.
인스턴스의 AMI , 인스턴스 유형, 키 페어, 보안 그룹, 블록 디바이스 매핑 등의 정보를 셋팅해서 템플릿을 구성합니다.

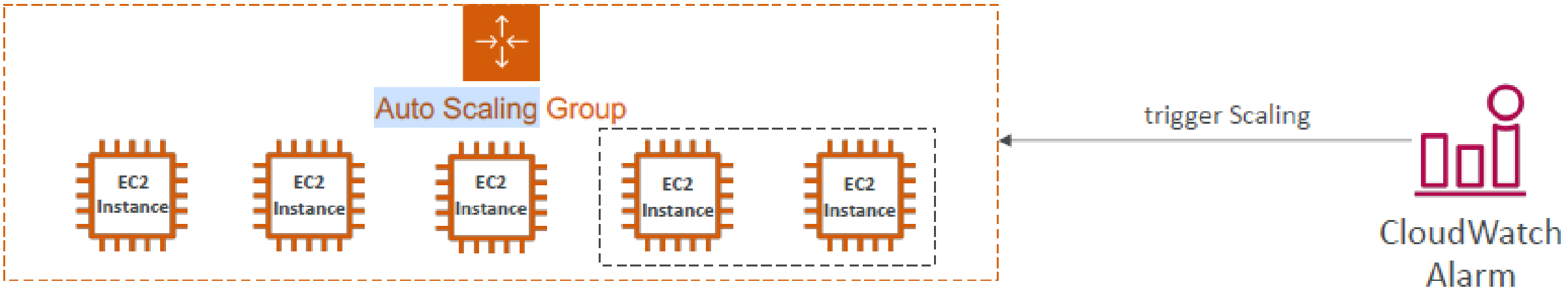


시작 템플릿(Launch Template)

- AMI + Instance Type
- EC2 User Data
- EBS Volumes
- Security Groups
- SSH Key Pair
- IAM Roles for your EC2 Instances
- Network + Subnets Information
- Load Balancer Information

ASG – CloudWatch Alarms & Scaling

CloudWatch 경보를 기반으로 ASG를 확장할 수 있습니다. 경보는 지표 (평균 CPU 또는 사용자 지정 지표)를 모니터링 합니다.



■ 스케일링 정책(Scaling Policies)

- Target Tracking Scaling : ex) I want the average ASG CPU to stay at around 40%.
- Simple / Step Scaling : ex) When a CloudWatch alarm is triggered (CPU > 70%), then add 2 units
- Scheduled Actions : ex) Increase the min capacity to 10 at 5 pm on Fridays
- Predictive Scaling: Continuously forecast load and schedule scaling ahead

애플리케이션에 대한 고가용성 구성 실습

0. 애플리케이션 개발, AMI 생성

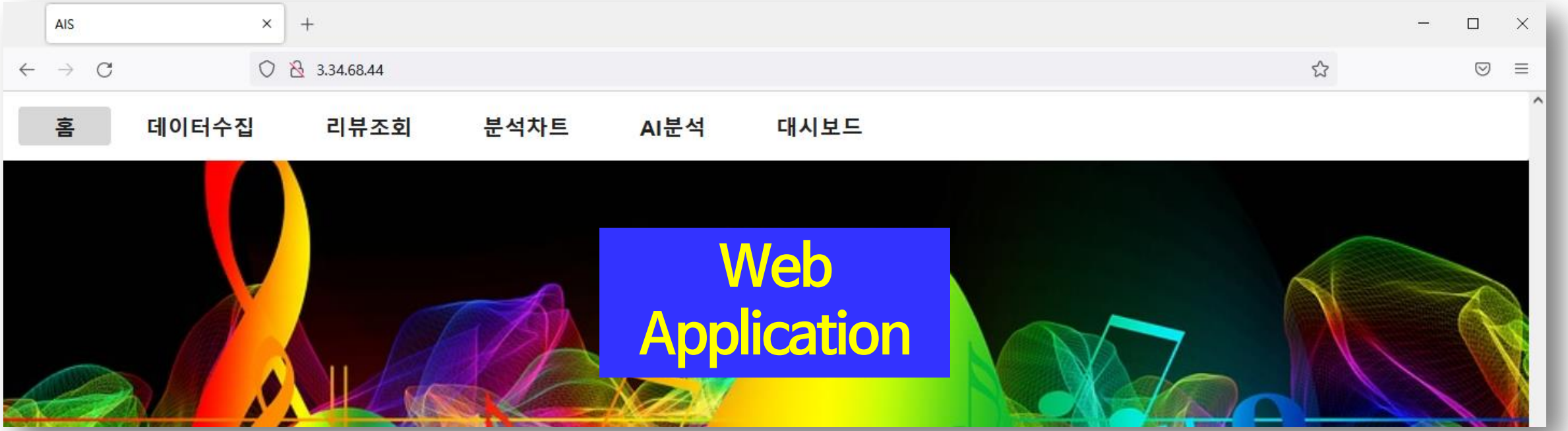
1. ELB 생성

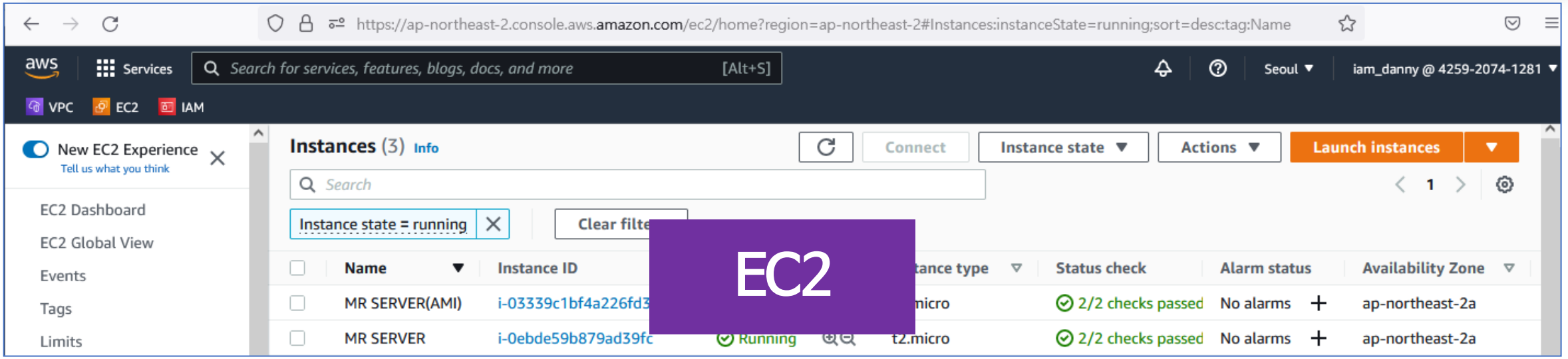
2. ASG 생성

3. Automatic Scaling

4. 4. Stress Test

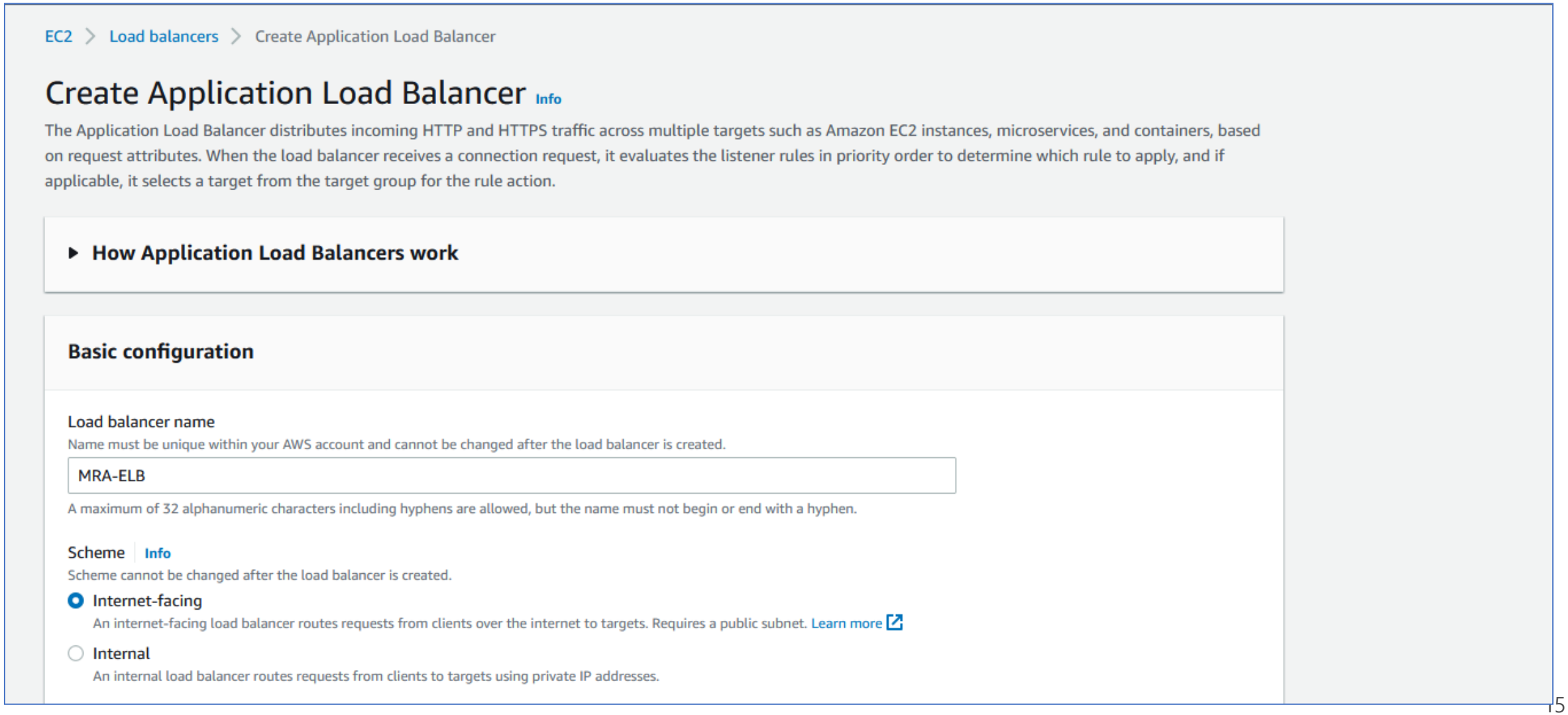
0. 애플리케이션 개발, AMI 생성



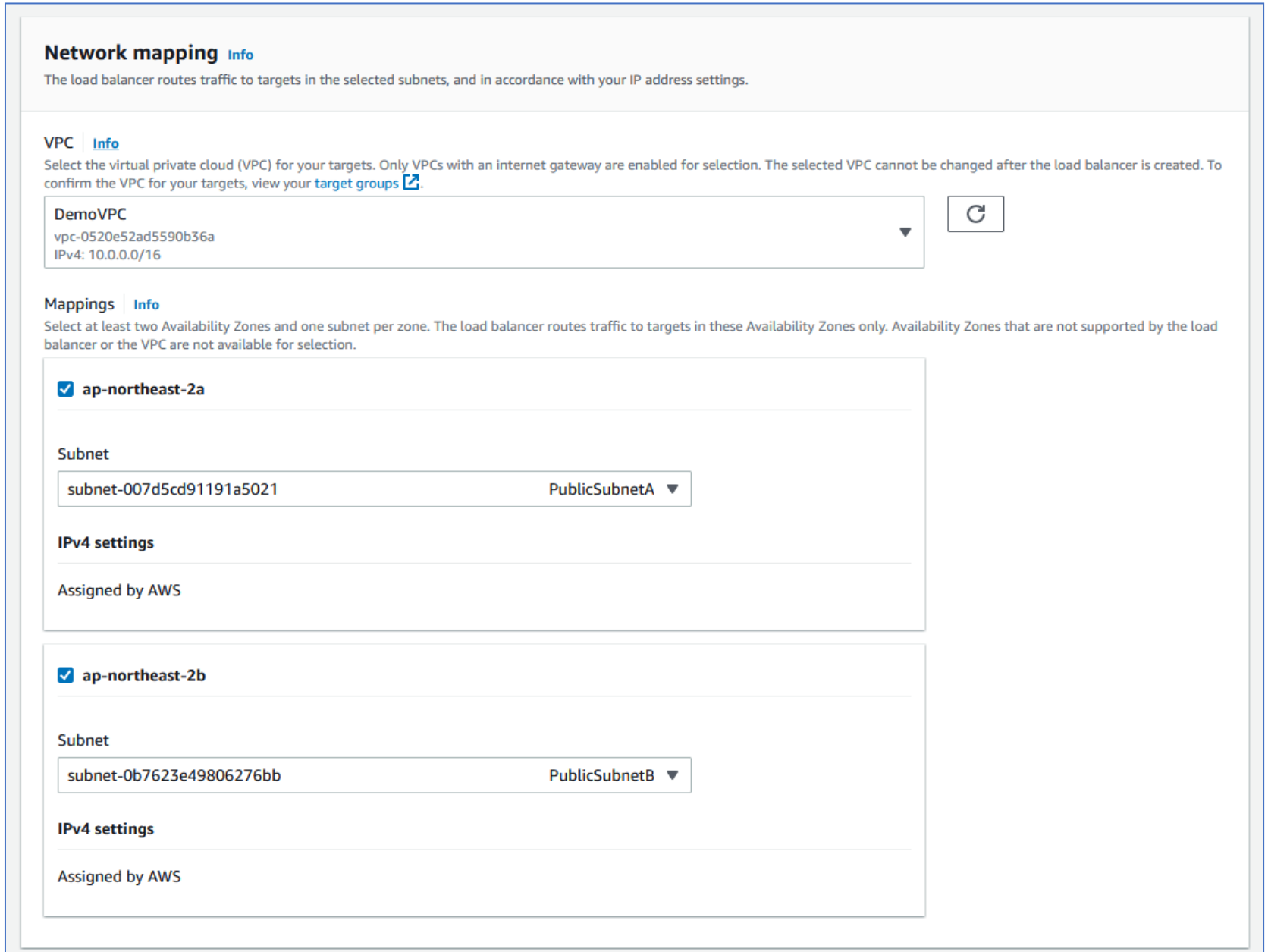


1. ELB 생성

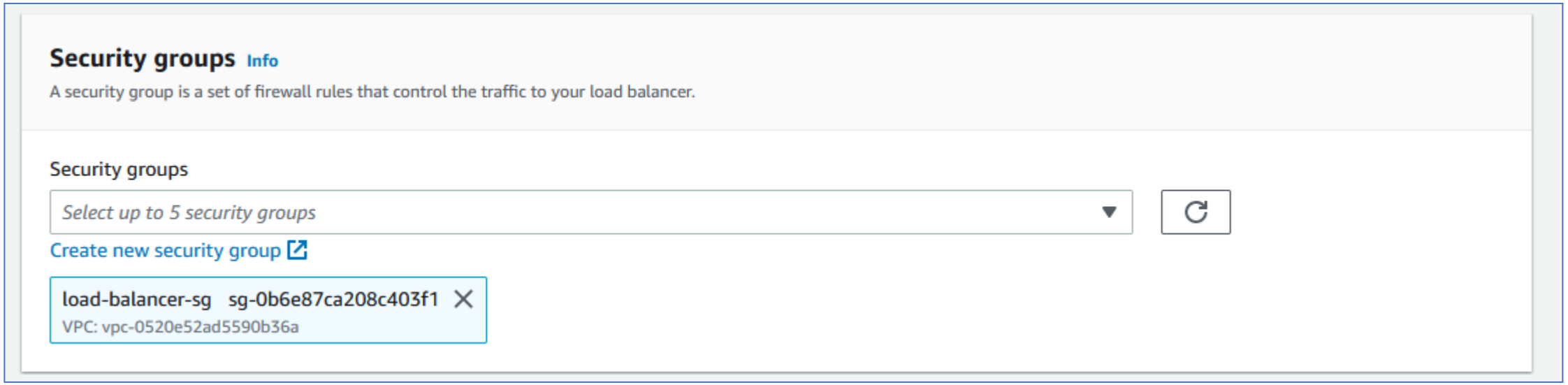
애플리케이션이 트래픽이 증가할 때 수평으로 크기 조정할 수 있도록 Application Load Balancer를 생성합니다.



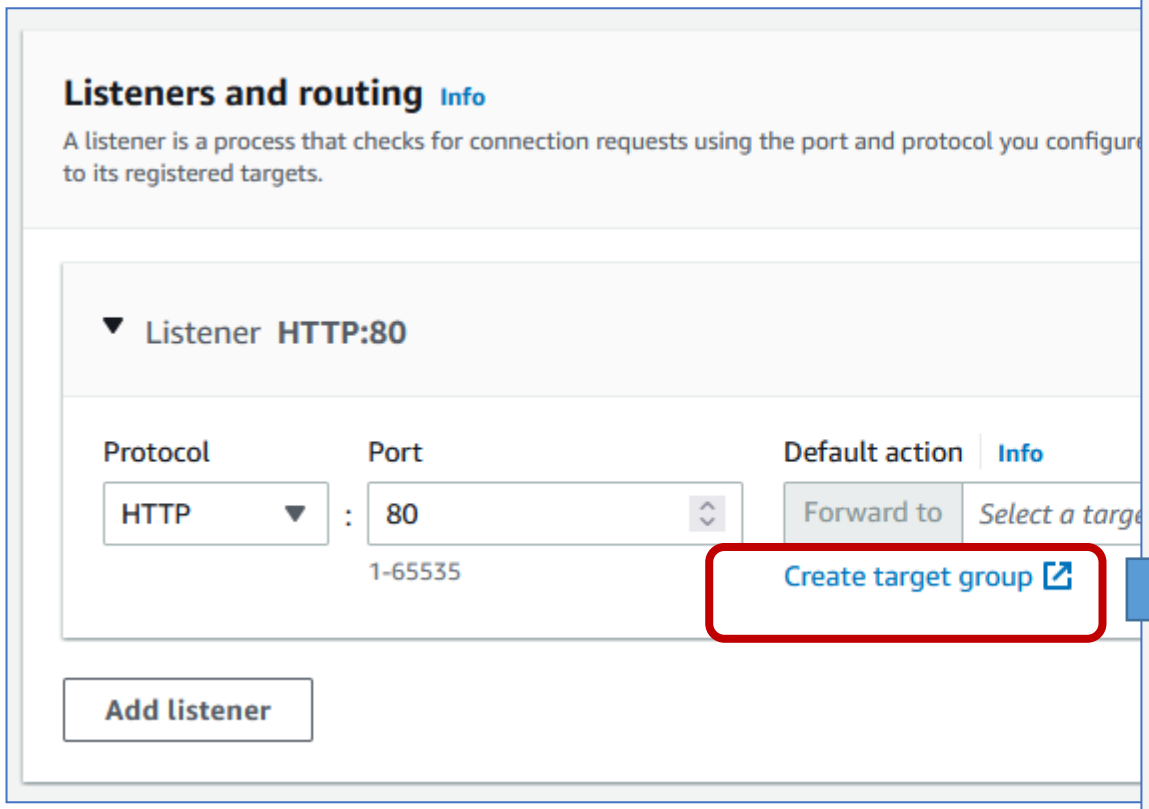
1. ELB 생성

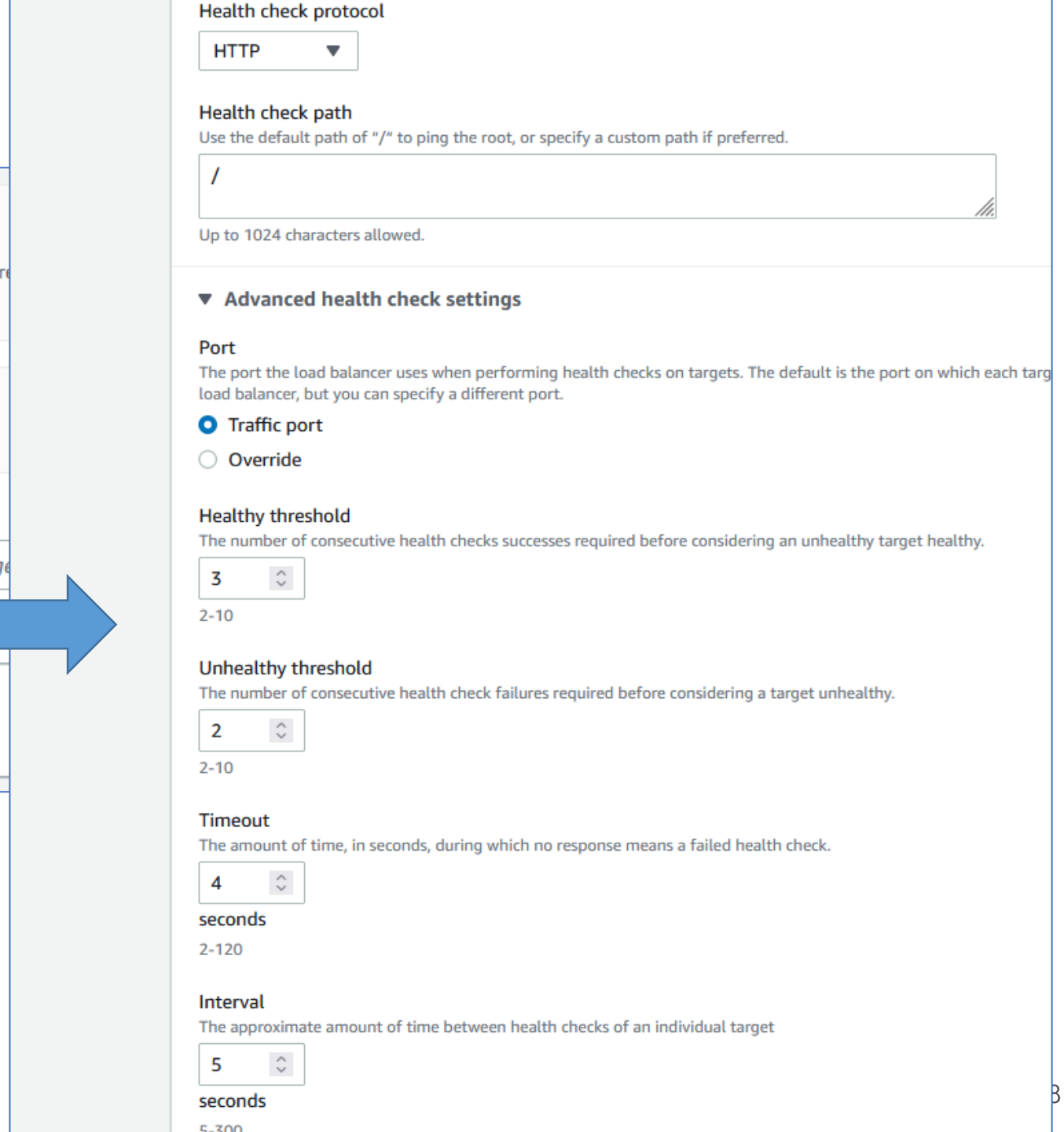

Network mapping Info
The load balancer routes traffic to targets in the selected subnets, and in accordance with your IP address settings.
VPC Info
Select the virtual private cloud (VPC) for your targets. Only VPCs with an internet gateway are enabled for selection. The selected VPC cannot be changed after the load balancer is created. To confirm the VPC for your targets, view your target groups.
DemoVPC
vpc-0520e52ad5590b36a
IPv4: 10.0.0.0/16
Mappings Info
Select at least two Availability Zones and one subnet per zone. The load balancer routes traffic to targets in these Availability Zones only. Availability Zones that are not supported by the load balancer or the VPC are not available for selection.
ap-northeast-2a
Subnet
subnet-007d5cd91191a5021 PublicSubnetA
IPv4 settings
Assigned by AWS
ap-northeast-2b
Subnet
subnet-0b7623e49806276bb PublicSubnetB
IPv4 settings
Assigned by AWS

1. ELB 생성

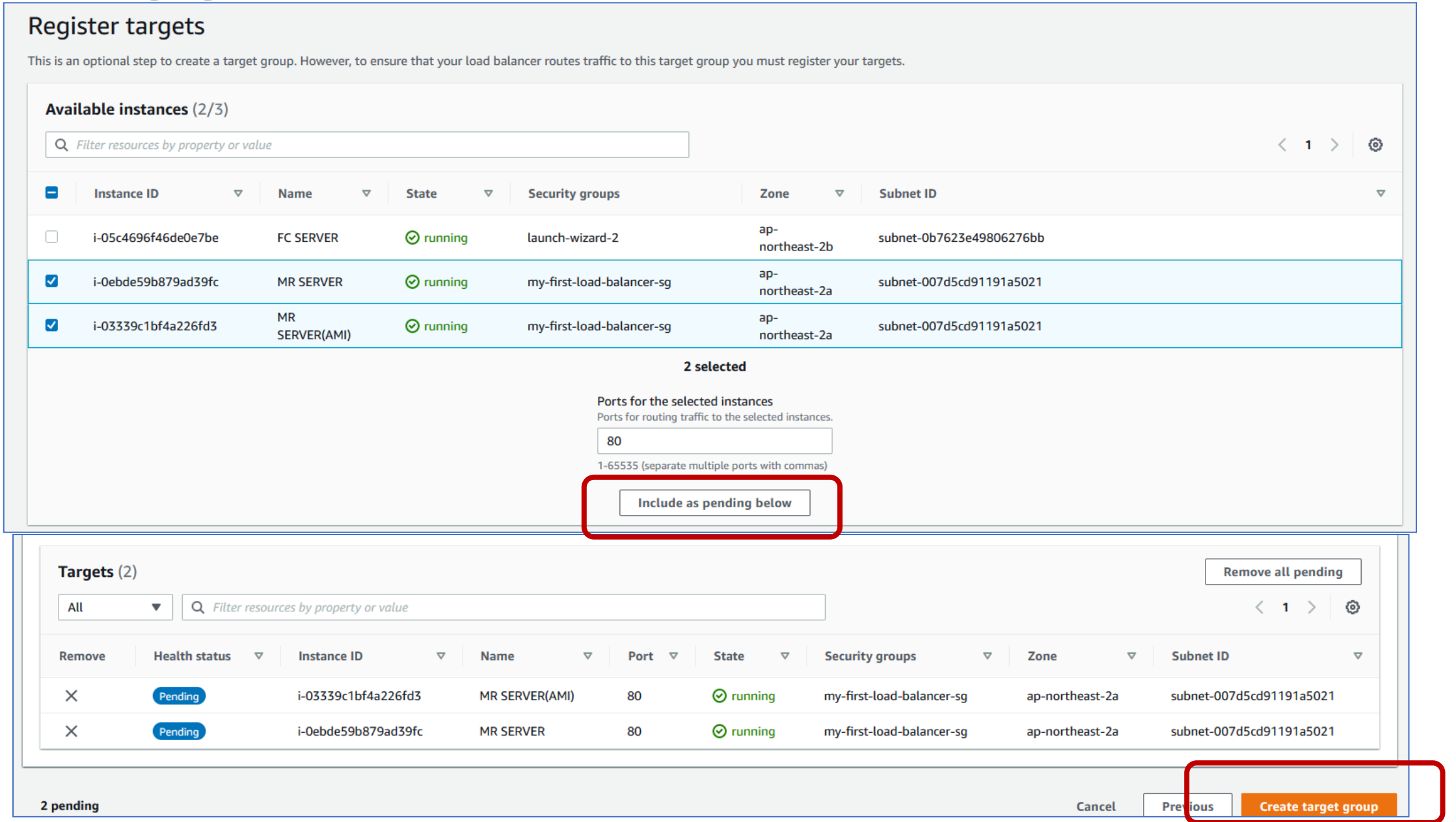

Security groups Info
A security group is a set of firewall rules that control the traffic to your load balancer.
Security groups
Select up to 5 security groups
Create new security group
load-balancer-sg sg-0b6e87ca208c403f1
VPC: vpc-0520e52ad5590b36a

1. ELB 생성

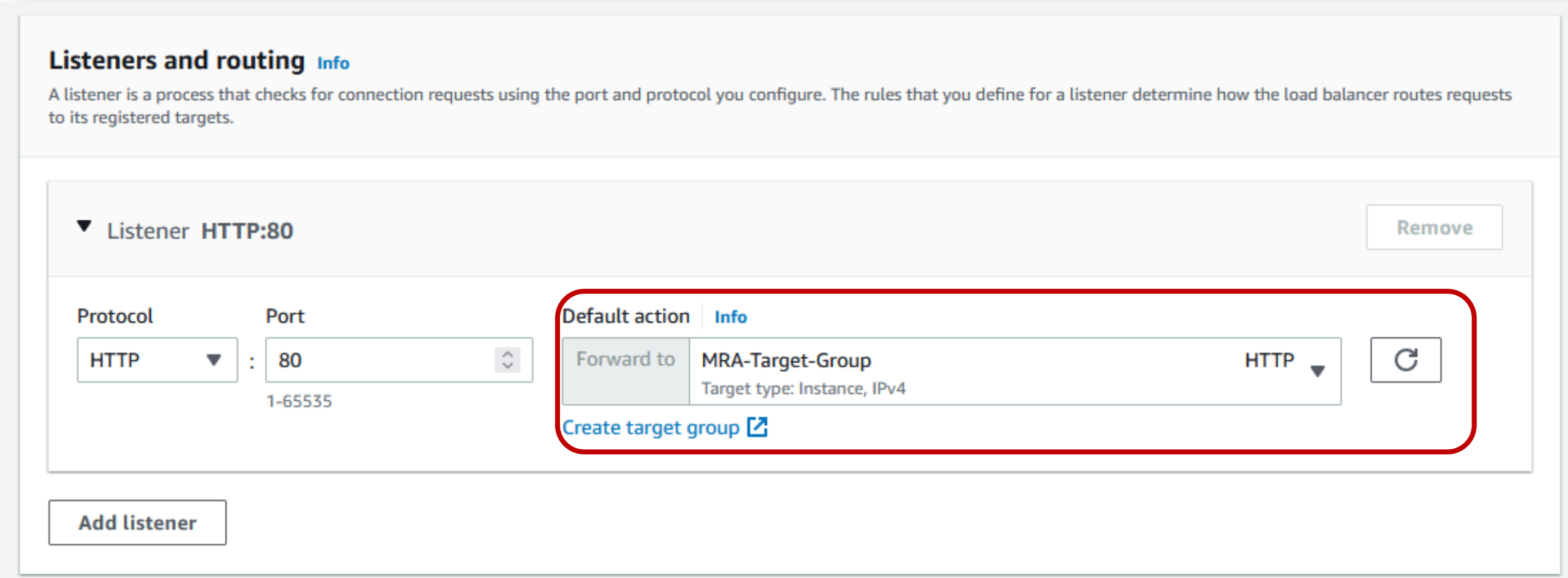

Listeners and routing Info
A listener is a process that checks for connection requests using the port and protocol you configur
to its registered targets.
Listener HTTP:80
Protocol
Port
Default action Info
HTTP
80
1-65535
Forward to
Select a targ
Create target group
Add listener


Health check protocol
HTTP
Health check path
Use the default path of "/" to ping the root, or specify a custom path if preferred.
/
Up to 1024 characters allowed.
Advanced health check settings
Port
The port the load balancer uses when performing health checks on targets. The default is the port on which each targ
load balancer, but you can specify a different port.
Traffic port
Override
Healthy threshold
The number of consecutive health checks successes required before considering an unhealthy target healthy.
3
2-10
Unhealthy threshold
The number of consecutive health check failures required before considering a target unhealthy.
2
2-10
Timeout
The amount of time, in seconds, during which no response means a failed health check.
4
seconds
2-120
Interval
The approximate amount of time between health checks of an individual target
5
seconds

1. ELB 생성


Register targets
This is an optional step to create a target group. However, to ensure that your load balancer routes traffic to this target group you must register your targets.
Available instances (2/3)
Filter resources by property or value
1
Instance ID
Name
State
Security groups
Zone
Subnet ID
i-05c4696f46de0e7be
FC SERVER
running
launch-wizard-2
ap-northeast-2b
subnet-0b7623e49806276bb
i-0ebde59b879ad39fc
MR SERVER
running
my-first-load-balancer-sg
ap-northeast-2a
subnet-007d5cd91191a5021
i-03339c1bf4a226fd3
MR SERVER(AMI)
running
my-first-load-balancer-sg
ap-northeast-2a
subnet-007d5cd91191a5021
2 selected
Ports for the selected instances
Ports for routing traffic to the selected instances.
80
1-65535 (separate multiple ports with commas)
Include as pending below
Targets (2)
Remove all pending
All
Filter resources by property or value
1
Remove
Health status
Instance ID
Name
Port
State
Security groups
Zone
Subnet ID
Pending
i-03339c1bf4a226fd3
MR SERVER(AMI)
80
running
my-first-load-balancer-sg
ap-northeast-2a
subnet-007d5cd91191a5021
Pending
i-0ebde59b879ad39fc
MR SERVER
80
running
my-first-load-balancer-sg
ap-northeast-2a
subnet-007d5cd91191a5021
2 pending
Cancel
Previous
Create target group

1. ELB 생성


Listeners and routing Info
A listener is a process that checks for connection requests using the port and protocol you configure. The rules that you define for a listener determine how the load balancer routes requests to its registered targets.
Listener HTTP:80
Remove
Protocol
Port
HTTP
:
80
1-65535
Default action Info
Forward to
MRA-Target-Group
HTTP
Target type: Instance, IPv4
Create target group
Add listener

1. ELB 생성

Summary

Review and confirm your configurations. Estimate cost

Basic configuration Edit

MRA-ELB

- Internet-facing
- IPv4

Security groups Edit

- load-balancer-sg
  sg-0b6e87ca208c403f1

Network mapping Edit

VPC vpc-0520e52ad5590b36a
DemoVPC

- ap-northeast-2a
  subnet-007d5cd91191a5021
  PublicSubnetA
- ap-northeast-2b
  subnet-0b7623e49806276bb
  PublicSubnetB

Listeners and routing Edit

- HTTP:80 defaults to
  MRA-Target-Group

Add-on services Edit

None

Tags Edit

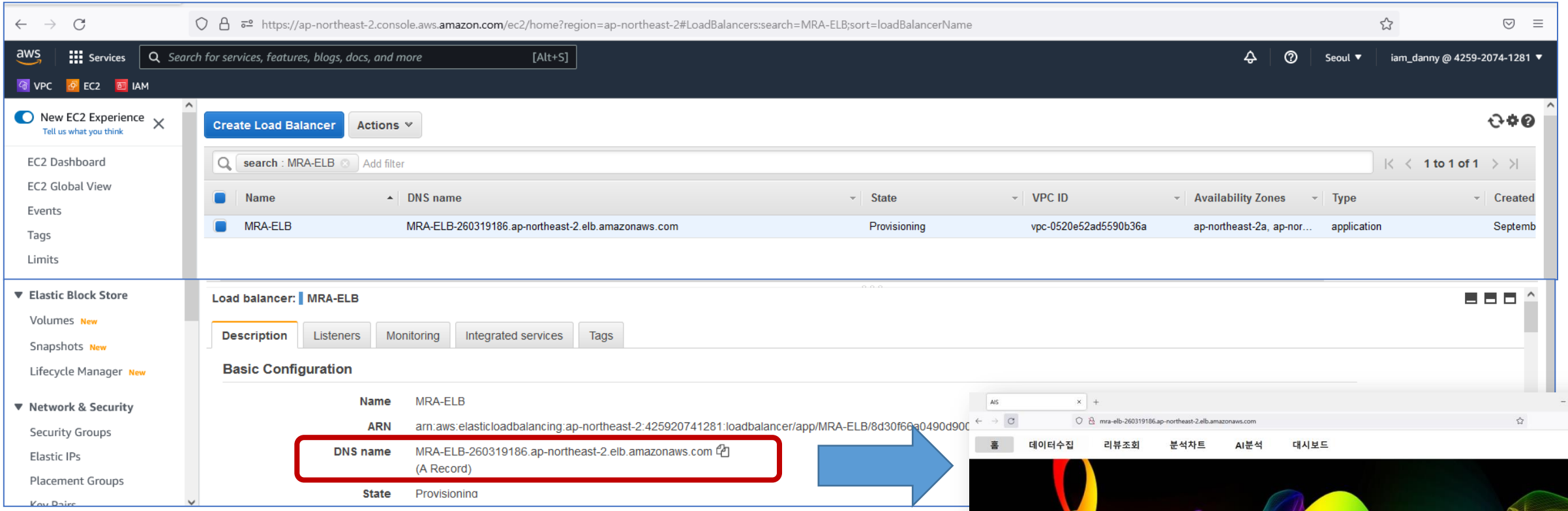
None

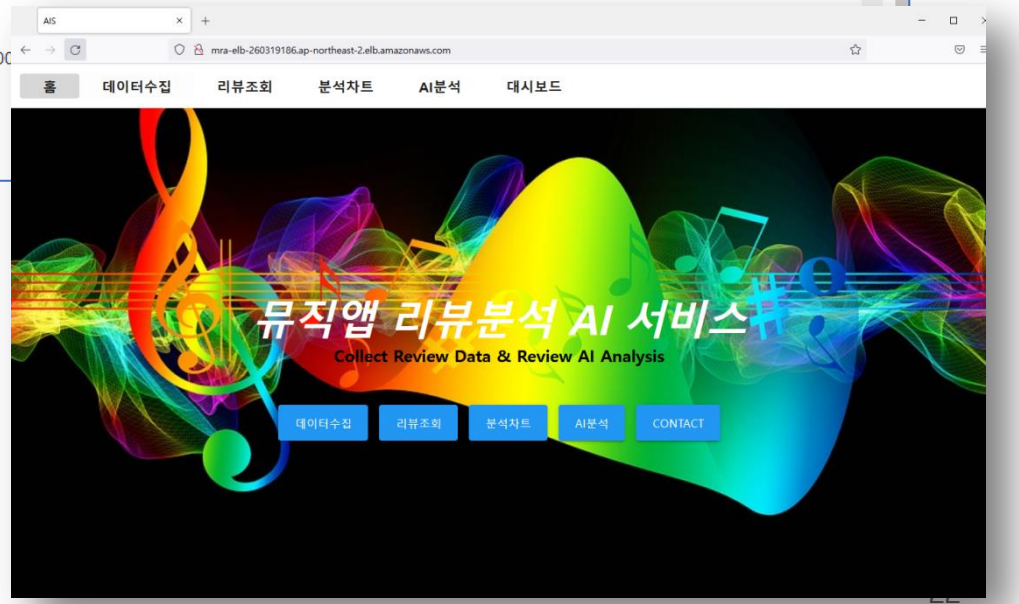
Attributes

Certain default attributes will be applied to your load balancer. You can view and edit them after creating the load balancer.

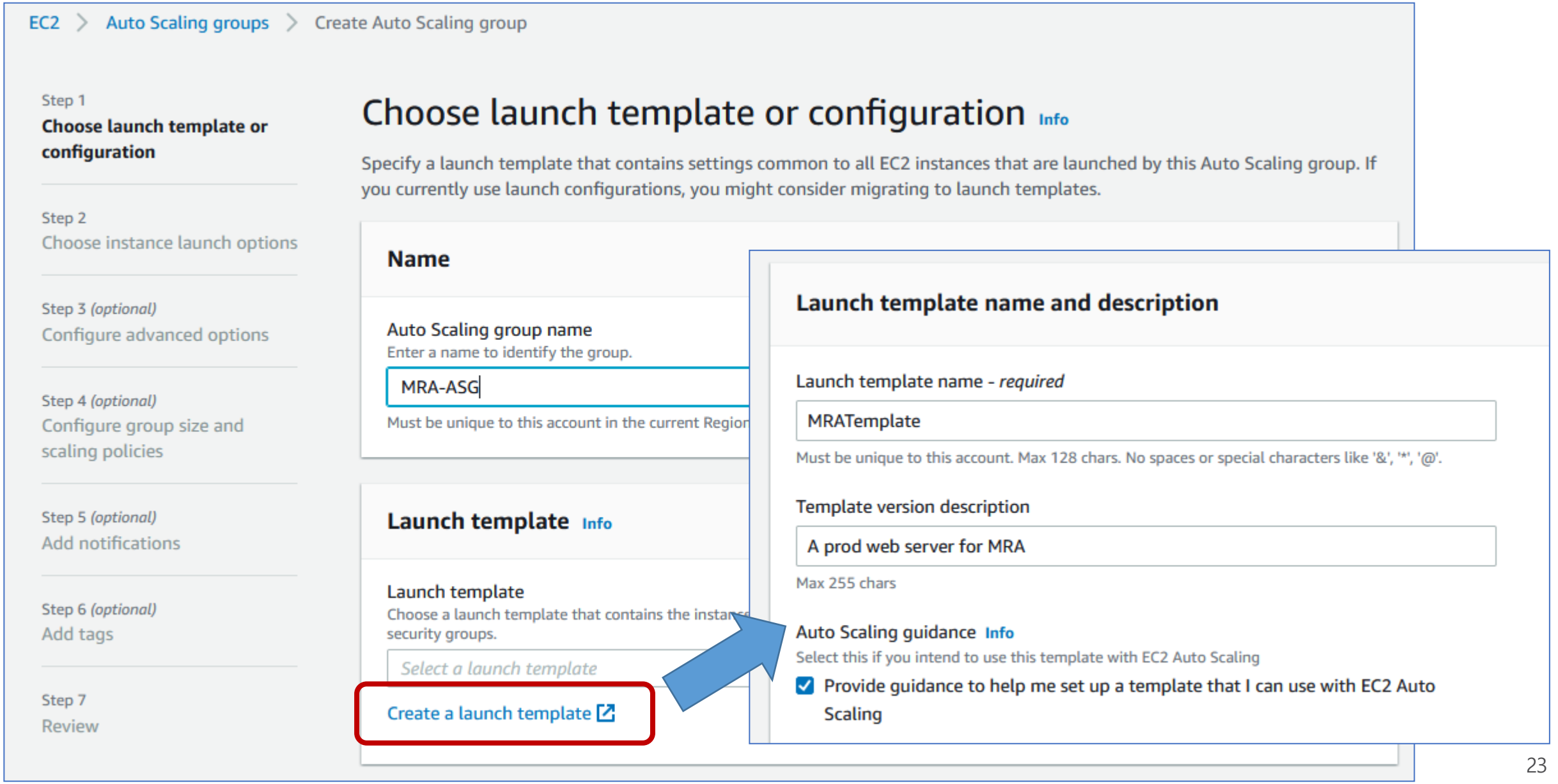
Cancel | Create load balancer

1. ELB 생성

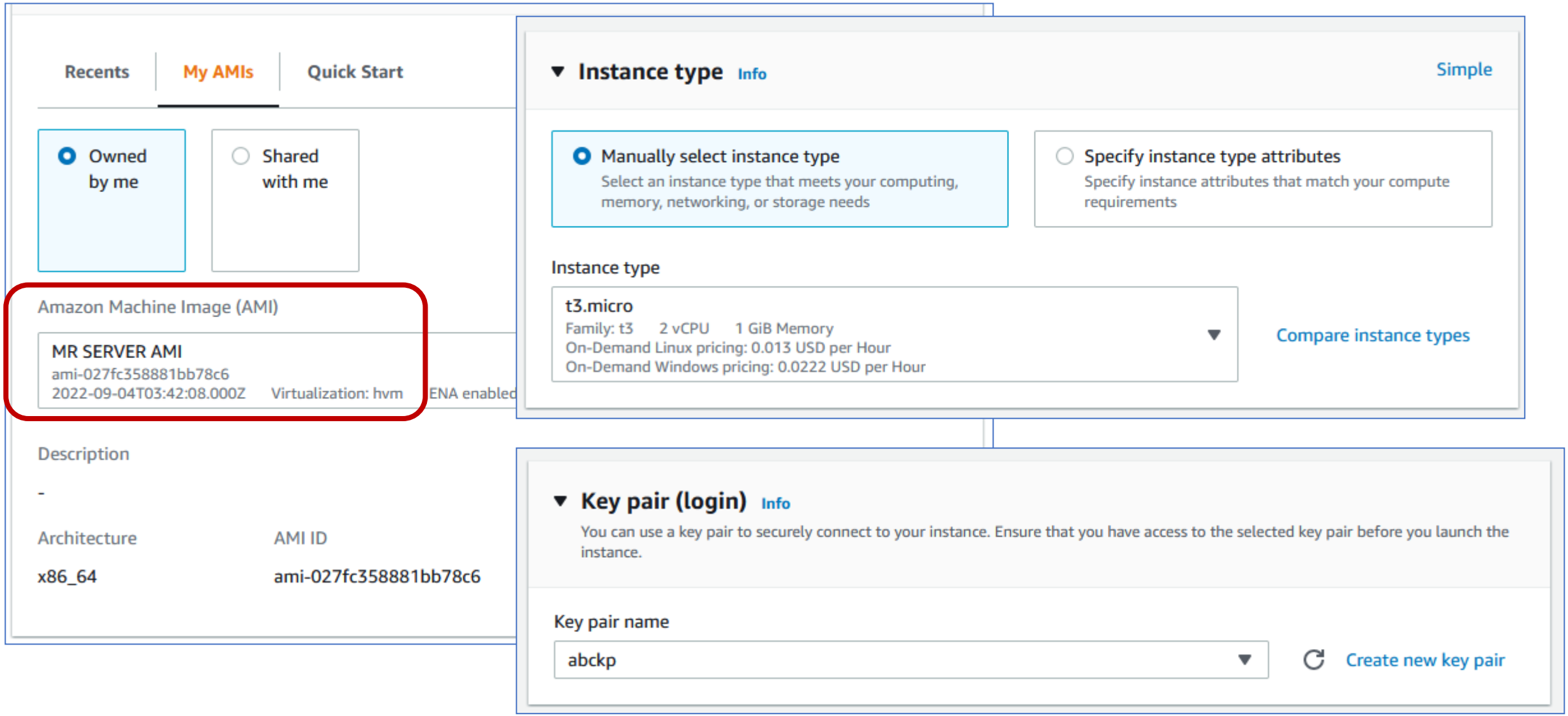

https://ap-northeast-2.console.aws.amazon.com/ec2/home?region=ap-northeast-2#LoadBalancers:search=MRA-ELB;sort=loadBalancerName
Services
Search for services, features, blogs, docs, and more [Alt+S]
Seoul
iam_danny @ 4259-2074-1281
VPC
EC2
IAM
New EC2 Experience
Tell us what you think
EC2 Dashboard
EC2 Global View
Events
Tags
Limits
Create Load Balancer
Actions
search : MRA-ELB
Add filter
1 to 1 of 1
Name
DNS name
State
VPC ID
Availability Zones
Type
Created
MRA-ELB
MRA-ELB-260319186.ap-northeast-2.elb.amazonaws.com
Provisioning
vpc-0520e52ad5590b36a
ap-northeast-2a, ap-nor...
application
Septemb
Elastic Block Store
Volumes New
Snapshots New
Lifecycle Manager New
Network & Security
Security Groups
Elastic IPs
Placement Groups
Load balancer: MRA-ELB
Description
Listeners
Monitoring
Integrated services
Tags
Basic Configuration
Name
MRA-ELB
ARN
arn:aws:elasticloadbalancing:ap-northeast-2:425920741281:loadbalancer/app/MRA-ELB/8d30f66a0490d900
DNS name
MRA-ELB-260319186.ap-northeast-2.elb.amazonaws.com
(A Record)
State
Provisioning


AIS
mra-elb-260319186.ap-northeast-2.elb.amazonaws.com
홈
데이터수집
리뷰조회
분석차트
AI분석
대시보드
뮤직앱 리뷰분석 AI 서비스
Collect Review Data & Review AI Analysis
데이터수집
리뷰조회
분석차트
AI분석
CONTACT

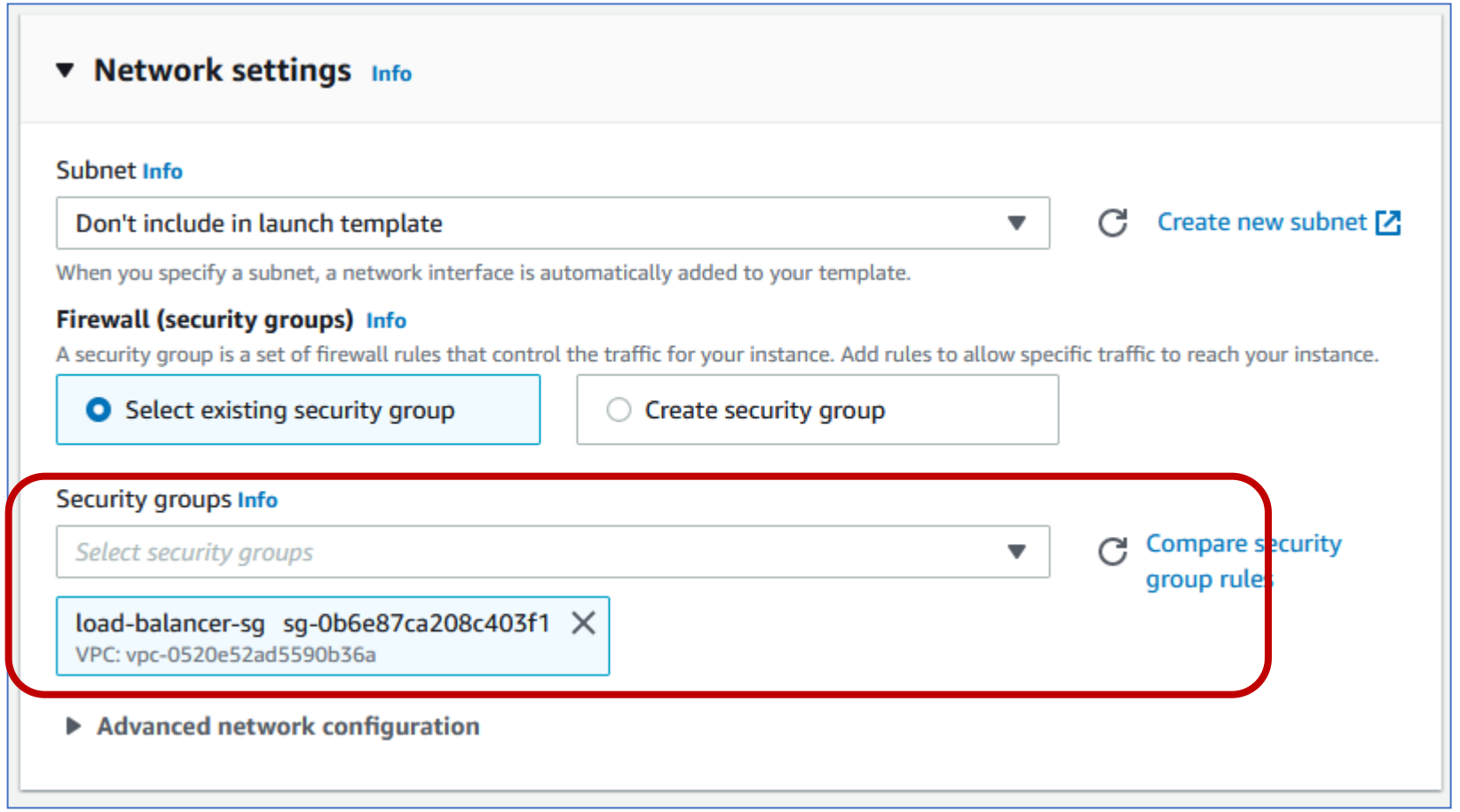
2. ASG 생성

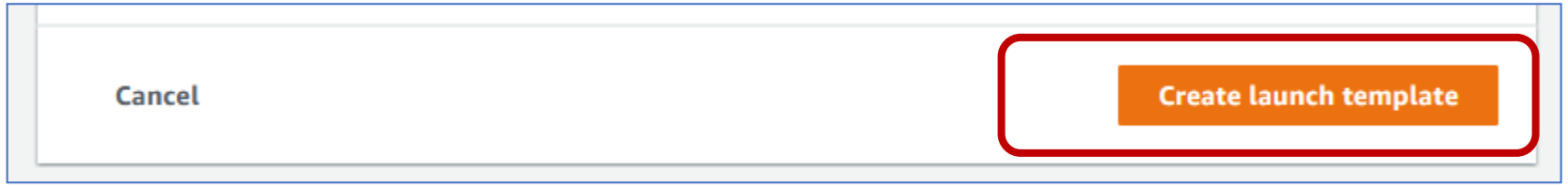

EC2 > Auto Scaling groups > Create Auto Scaling group
Step 1
Choose launch template or configuration
Step 2
Choose instance launch options
Step 3 (optional)
Configure advanced options
Step 4 (optional)
Configure group size and scaling policies
Step 5 (optional)
Add notifications
Step 6 (optional)
Add tags
Step 7
Review
Choose launch template or configuration Info
Specify a launch template that contains settings common to all EC2 instances that are launched by this Auto Scaling group. If you currently use launch configurations, you might consider migrating to launch templates.
Name
Auto Scaling group name
Enter a name to identify the group.
MRA-ASG
Must be unique to this account in the current Regior
Launch template Info
Launch template
Choose a launch template that contains the instanc
security groups.
Select a launch template
Create a launch template
Launch template name and description
Launch template name - required
MRATemplate
Must be unique to this account. Max 128 chars. No spaces or special characters like '&', '*', '@'.
Template version description
A prod web server for MRA
Max 255 chars
Auto Scaling guidance Info
Select this if you intend to use this template with EC2 Auto Scaling
Provide guidance to help me set up a template that I can use with EC2 Auto Scaling

2. ASG 생성 – 시작템플릿(Launch Template)

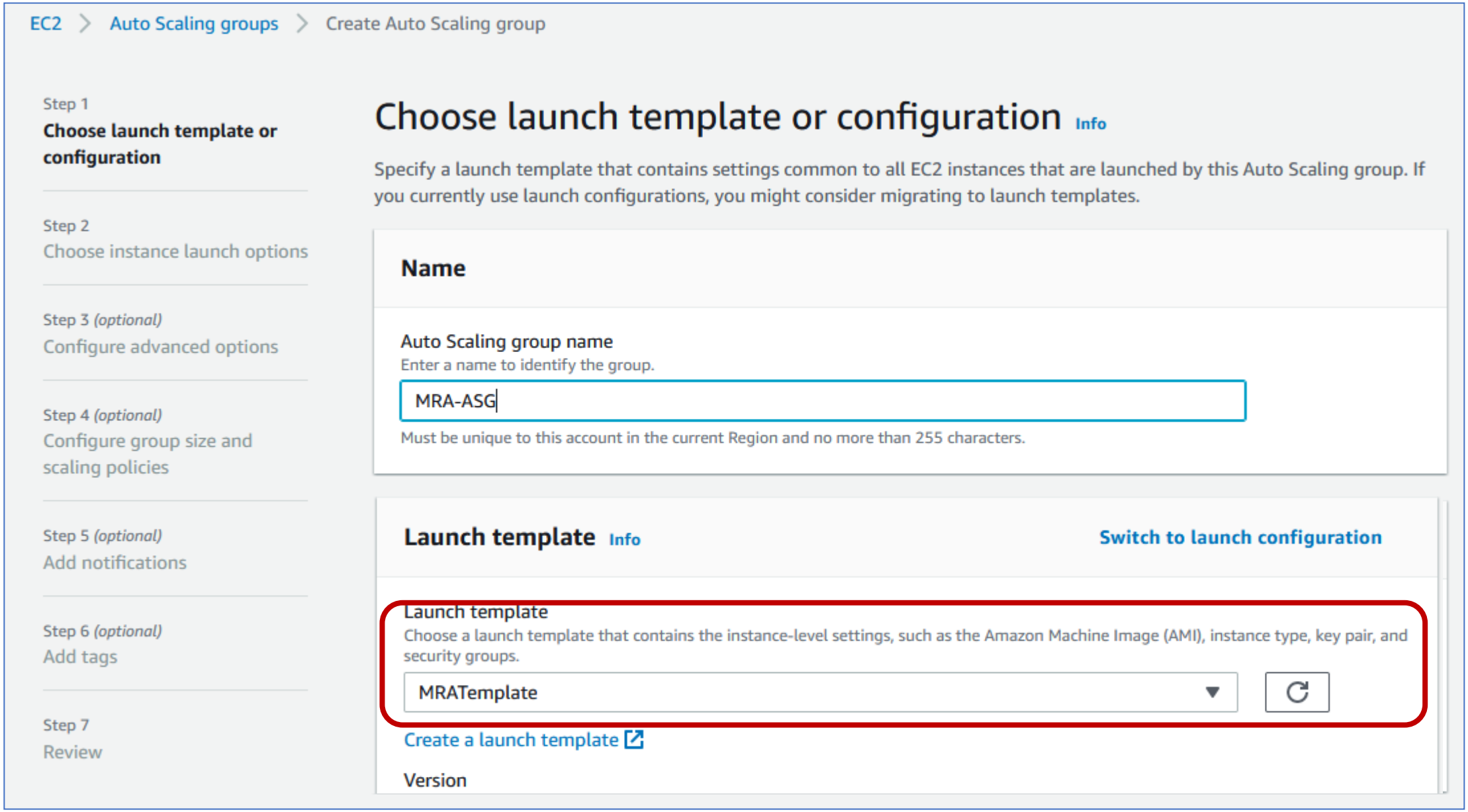


2. ASG 생성 – 시작템플릿(Launch Template)

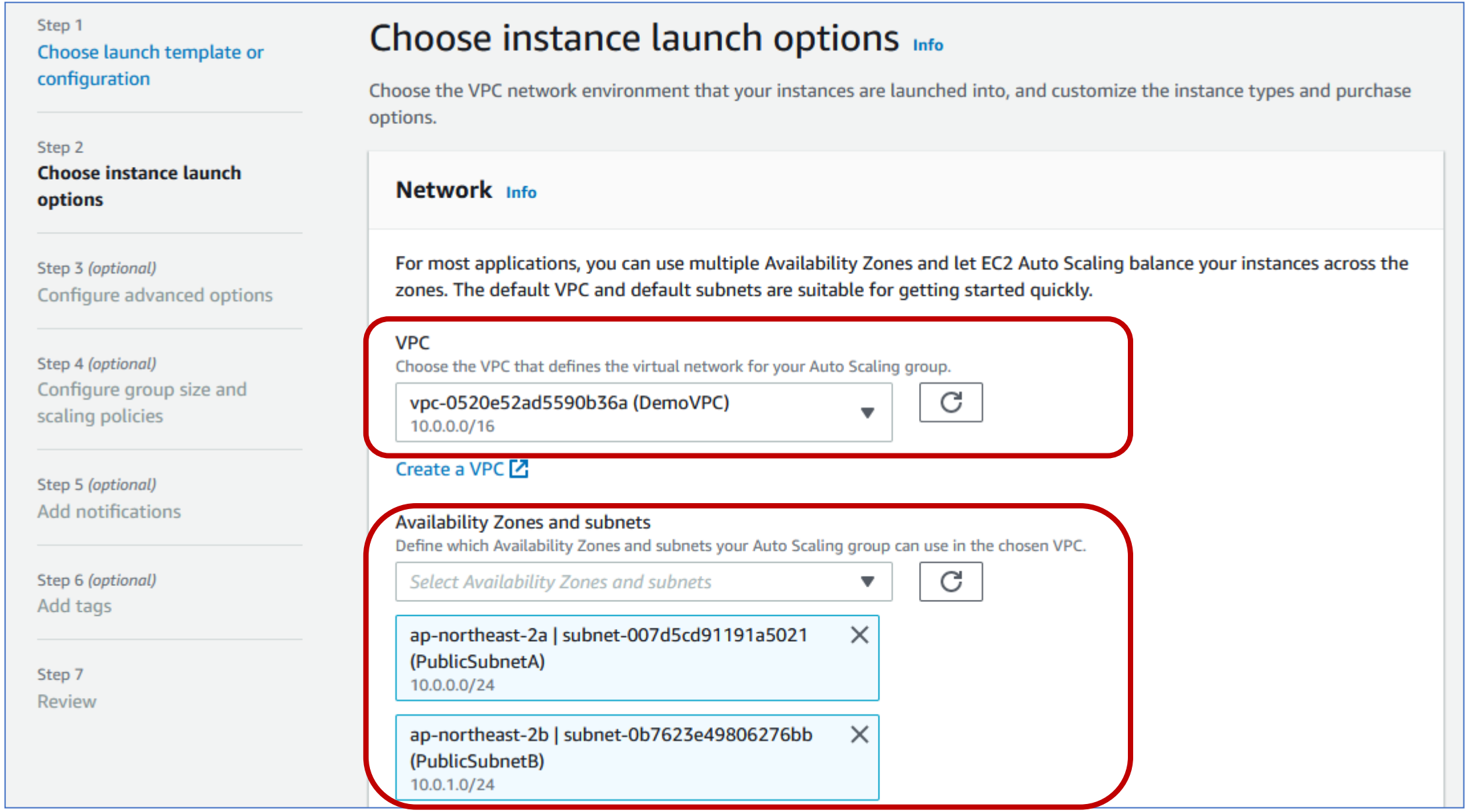




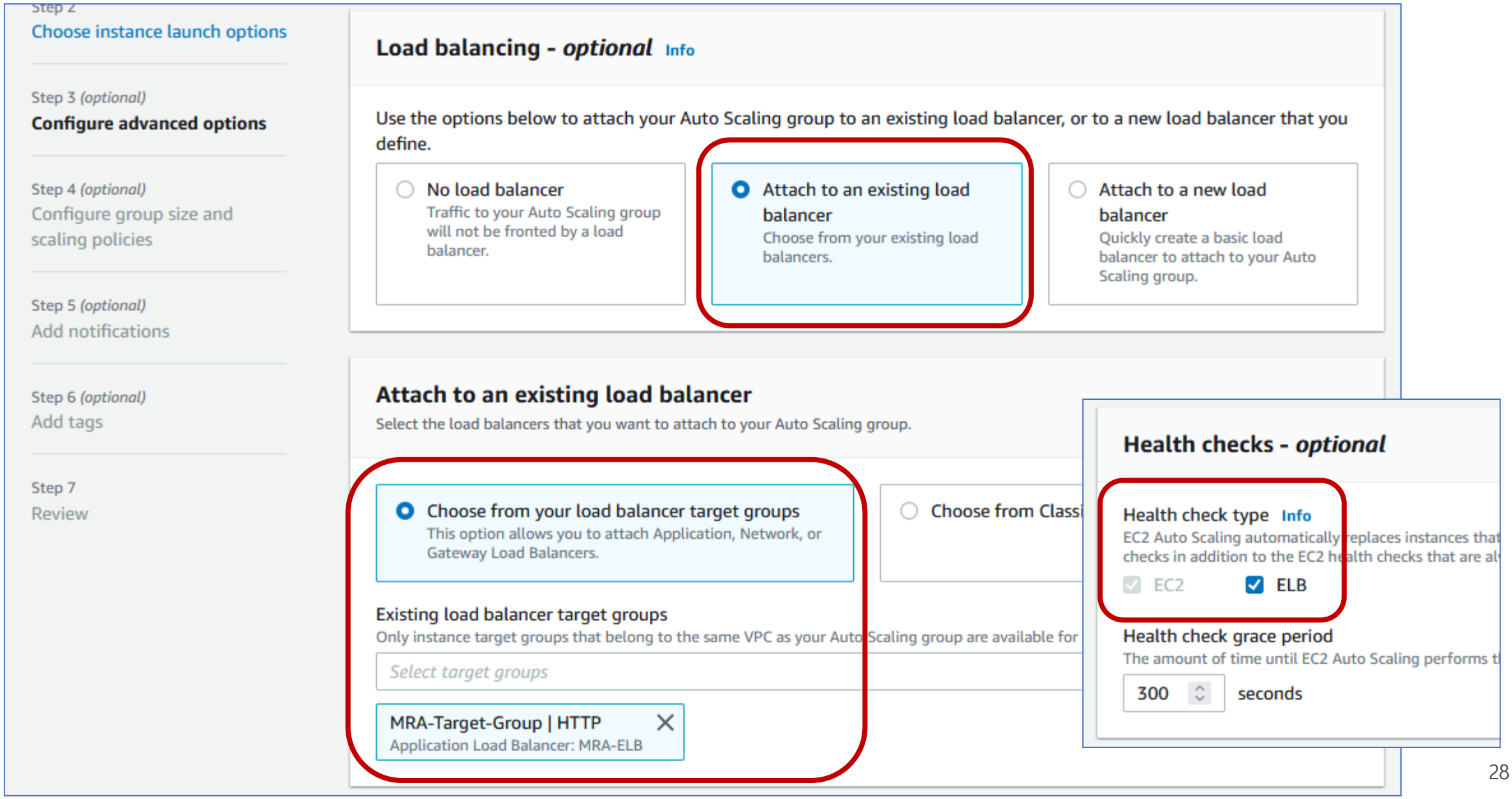
2. ASG 생성


EC2 > Auto Scaling groups > Create Auto Scaling group
Step 1
Choose launch template or configuration
Step 2
Choose instance launch options
Step 3 (optional)
Configure advanced options
Step 4 (optional)
Configure group size and scaling policies
Step 5 (optional)
Add notifications
Step 6 (optional)
Add tags
Step 7
Review
Choose launch template or configuration Info
Specify a launch template that contains settings common to all EC2 instances that are launched by this Auto Scaling group. If you currently use launch configurations, you might consider migrating to launch templates.
Name
Auto Scaling group name
Enter a name to identify the group.
MRA-ASG
Must be unique to this account in the current Region and no more than 255 characters.
Launch template Info
Switch to launch configuration
Launch template
Choose a launch template that contains the instance-level settings, such as the Amazon Machine Image (AMI), instance type, key pair, and security groups.
MRATemplate
Create a launch template
Version

2. ASG 생성


Step 1
Choose launch template or configuration
Step 2
Choose instance launch options
Step 3 (optional)
Configure advanced options
Step 4 (optional)
Configure group size and scaling policies
Step 5 (optional)
Add notifications
Step 6 (optional)
Add tags
Step 7
Review
Choose instance launch options Info
Choose the VPC network environment that your instances are launched into, and customize the instance types and purchase options.
Network Info
For most applications, you can use multiple Availability Zones and let EC2 Auto Scaling balance your instances across the zones. The default VPC and default subnets are suitable for getting started quickly.
VPC
Choose the VPC that defines the virtual network for your Auto Scaling group.
vpc-0520e52ad5590b36a (DemoVPC)
10.0.0.0/16
Create a VPC
Availability Zones and subnets
Define which Availability Zones and subnets your Auto Scaling group can use in the chosen VPC.
Select Availability Zones and subnets
ap-northeast-2a | subnet-007d5cd91191a5021 (PublicSubnetA)
10.0.0.0/24
ap-northeast-2b | subnet-0b7623e49806276bb (PublicSubnetB)
10.0.1.0/24

2. ASG 생성


Step 2
Choose instance launch options
Step 3 (optional)
Configure advanced options
Step 4 (optional)
Configure group size and scaling policies
Step 5 (optional)
Add notifications
Step 6 (optional)
Add tags
Step 7
Review
Load balancing - optional Info
Use the options below to attach your Auto Scaling group to an existing load balancer, or to a new load balancer that you define.
No load balancer
Traffic to your Auto Scaling group will not be fronted by a load balancer.
Attach to an existing load balancer
Choose from your existing load balancers.
Attach to a new load balancer
Quickly create a basic load balancer to attach to your Auto Scaling group.
Attach to an existing load balancer
Select the load balancers that you want to attach to your Auto Scaling group.
Choose from your load balancer target groups
This option allows you to attach Application, Network, or Gateway Load Balancers.
Choose from Classi
Existing load balancer target groups
Only instance target groups that belong to the same VPC as your Auto Scaling group are available for
Select target groups
MRA-Target-Group | HTTP
Application Load Balancer: MRA-ELB
Health checks - optional
Health check type Info
EC2 Auto Scaling automatically replaces instances that
checks in addition to the EC2 health checks that are al
EC2
ELB
Health check grace period
The amount of time until EC2 Auto Scaling performs t
300
seconds

2. ASG 생성

Group size - *optional* Info

Specify the size of the Auto Scaling group by changing the desired capacity. You
capacity limits. Your desired capacity must be within the limit range.

Desired capacity

2

Minimum capacity

1

Maximum capacity

4

Review Info

Step 1: Choose launch template or configuration — Edit

Group details

Auto Scaling group name
MRA-ASG

Launch template

Launch template	Version	Description
MRATemplate lt-03e52f2cb6476d958	Default	A prod web server for MRA

Step 2: Choose instance launch options — Edit

Network

Network

VPC
vpc-0520e52ad5590b36a

Availability Zone	Subnet	
ap-northeast-2a	subnet-007d5cd91191a5021	10.0.0.0/24
ap-northeast-2b	subnet-0b7623e49806276bb	10.0.1.0/24

Step 3: Configure advanced options — Edit

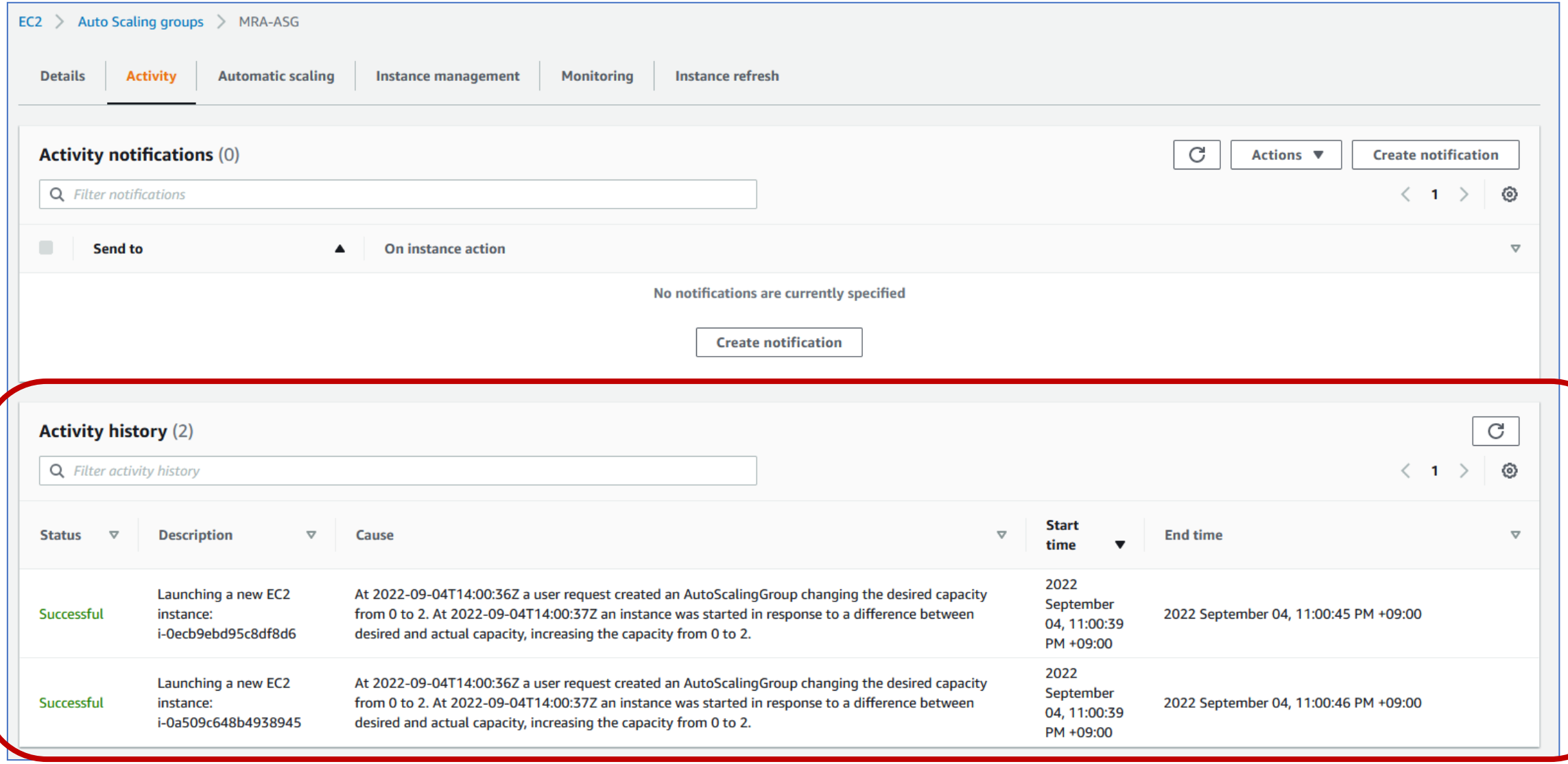
Load balancing

Load balancer 1

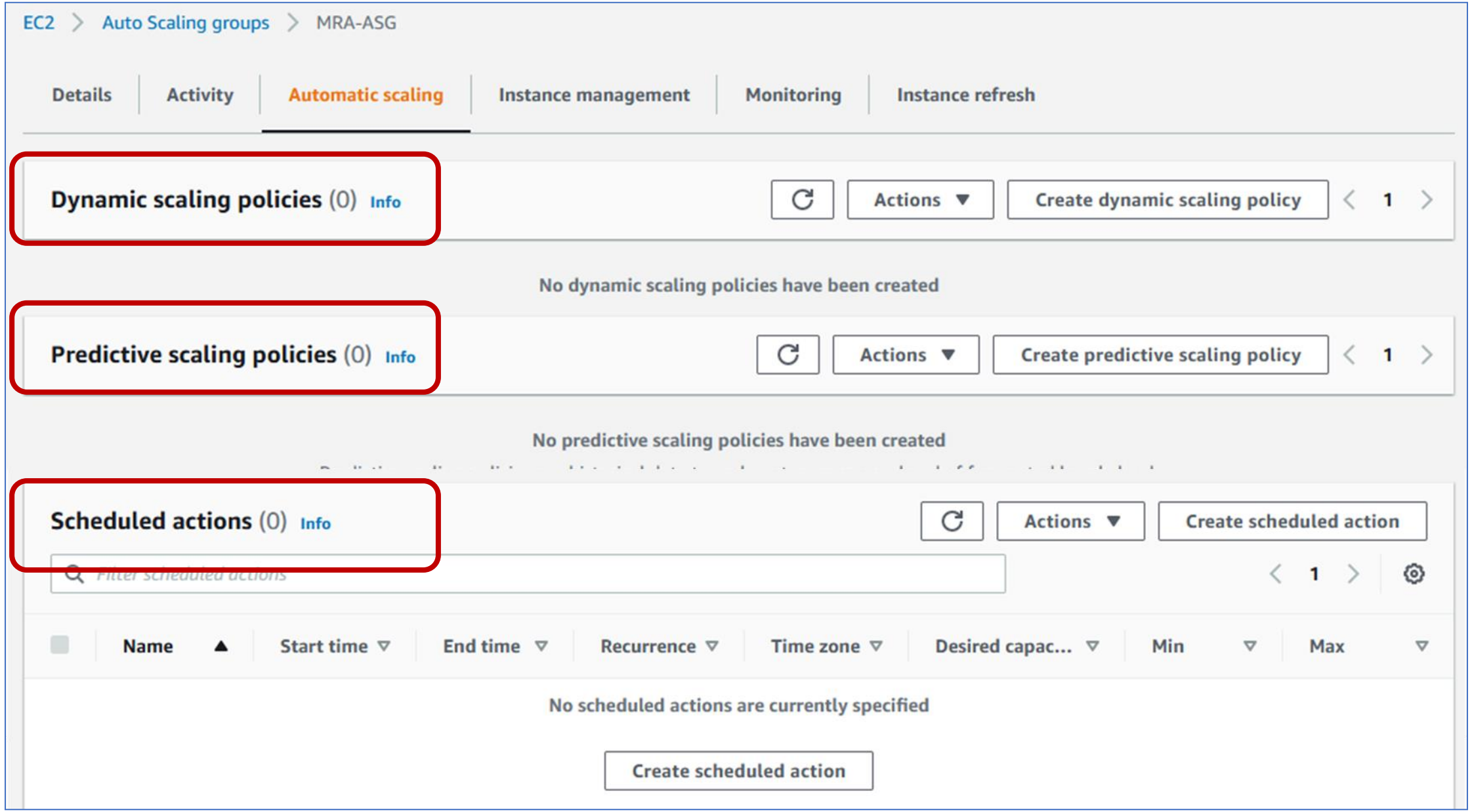
Name	Type	Target group
MRA-ELB	Application/HTTP	MRA-Target-Group

Create Auto Scaling group

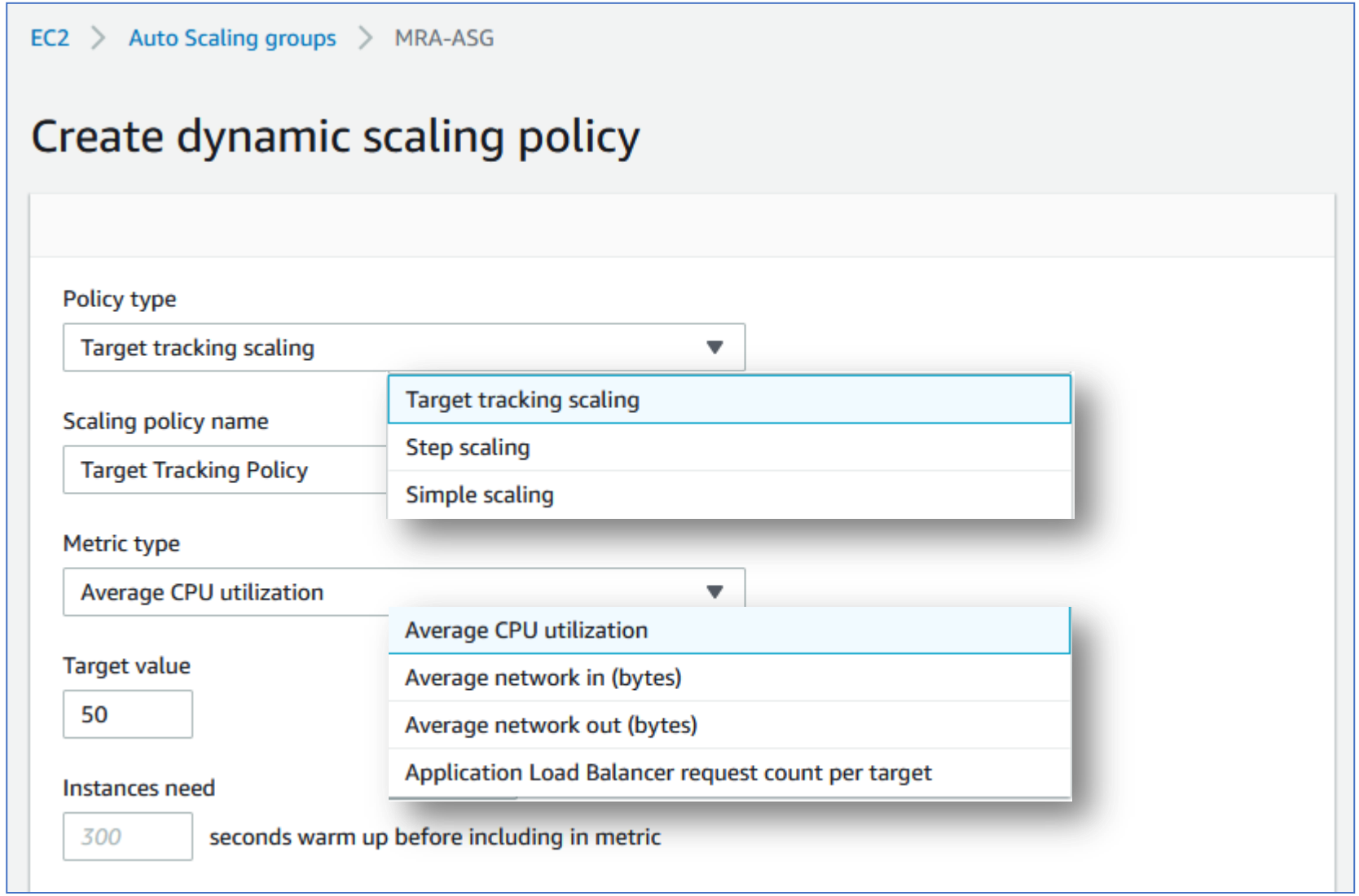
2. ASG 생성


EC2 > Auto Scaling groups > MRA-ASG
Details
Activity
Automatic scaling
Instance management
Monitoring
Instance refresh
Activity notifications (0)
Actions
Create notification
Filter notifications
1
Send to
On instance action
No notifications are currently specified
Create notification
Activity history (2)
Filter activity history
1
Status
Description
Cause
Start time
End time
Successful
Launching a new EC2 instance: i-0ecb9ebd95c8df8d6
At 2022-09-04T14:00:36Z a user request created an AutoScalingGroup changing the desired capacity from 0 to 2. At 2022-09-04T14:00:37Z an instance was started in response to a difference between desired and actual capacity, increasing the capacity from 0 to 2.
2022 September 04, 11:00:39 PM +09:00
2022 September 04, 11:00:45 PM +09:00
Successful
Launching a new EC2 instance: i-0a509c648b4938945
At 2022-09-04T14:00:36Z a user request created an AutoScalingGroup changing the desired capacity from 0 to 2. At 2022-09-04T14:00:37Z an instance was started in response to a difference between desired and actual capacity, increasing the capacity from 0 to 2.
2022 September 04, 11:00:39 PM +09:00
2022 September 04, 11:00:46 PM +09:00

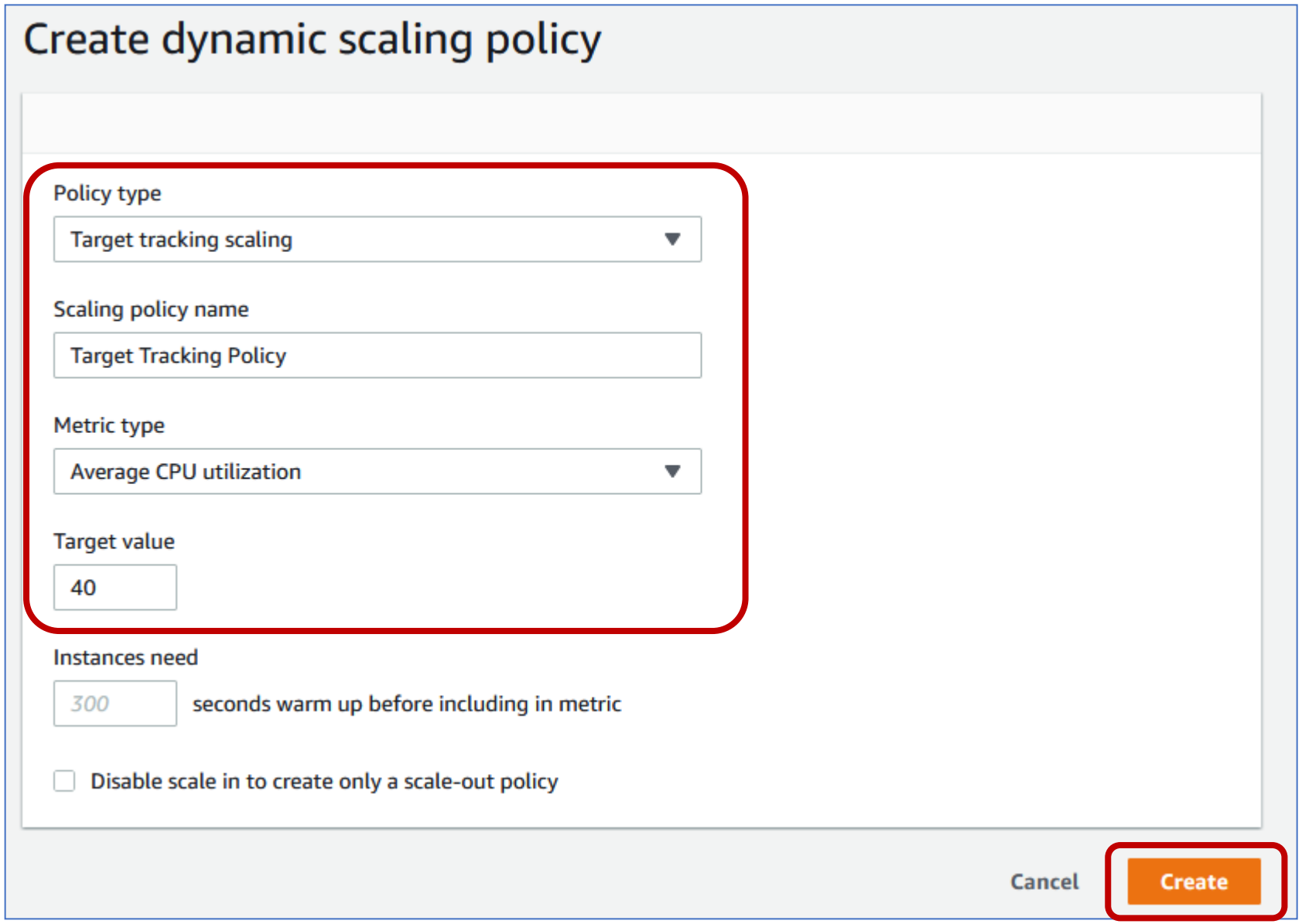
3. Automatic Scaling



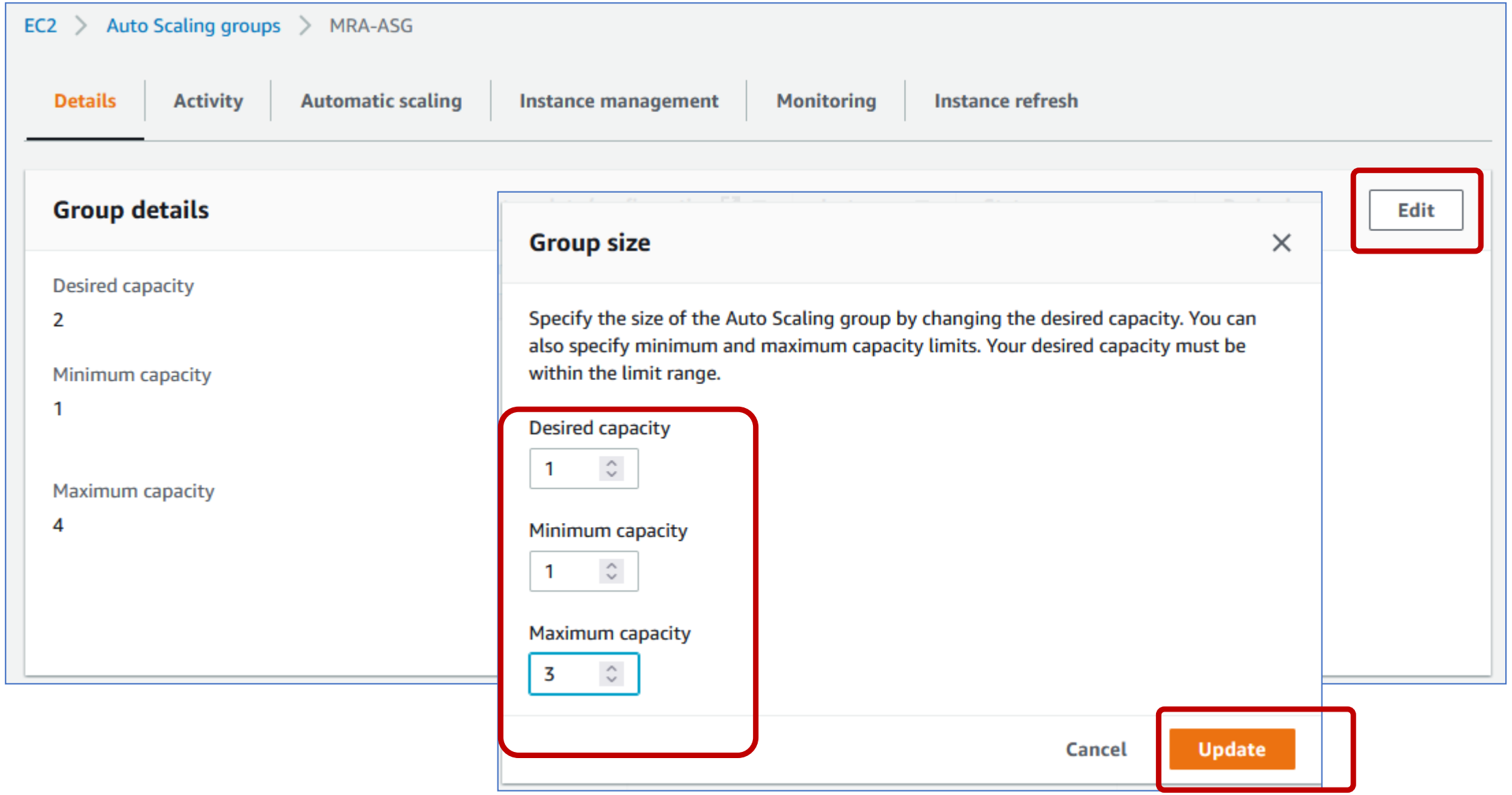
3. Automatic Scaling



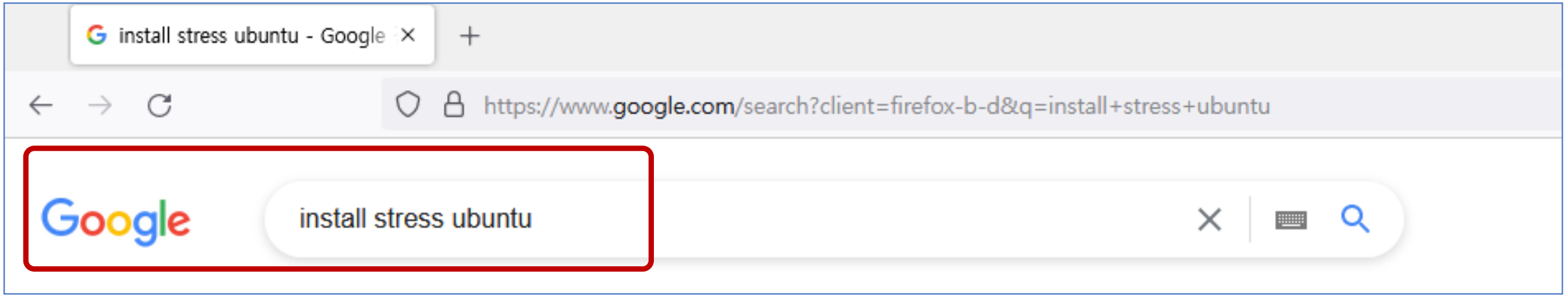
3. Automatic Scaling



3. Automatic Scaling



4. Stress Test

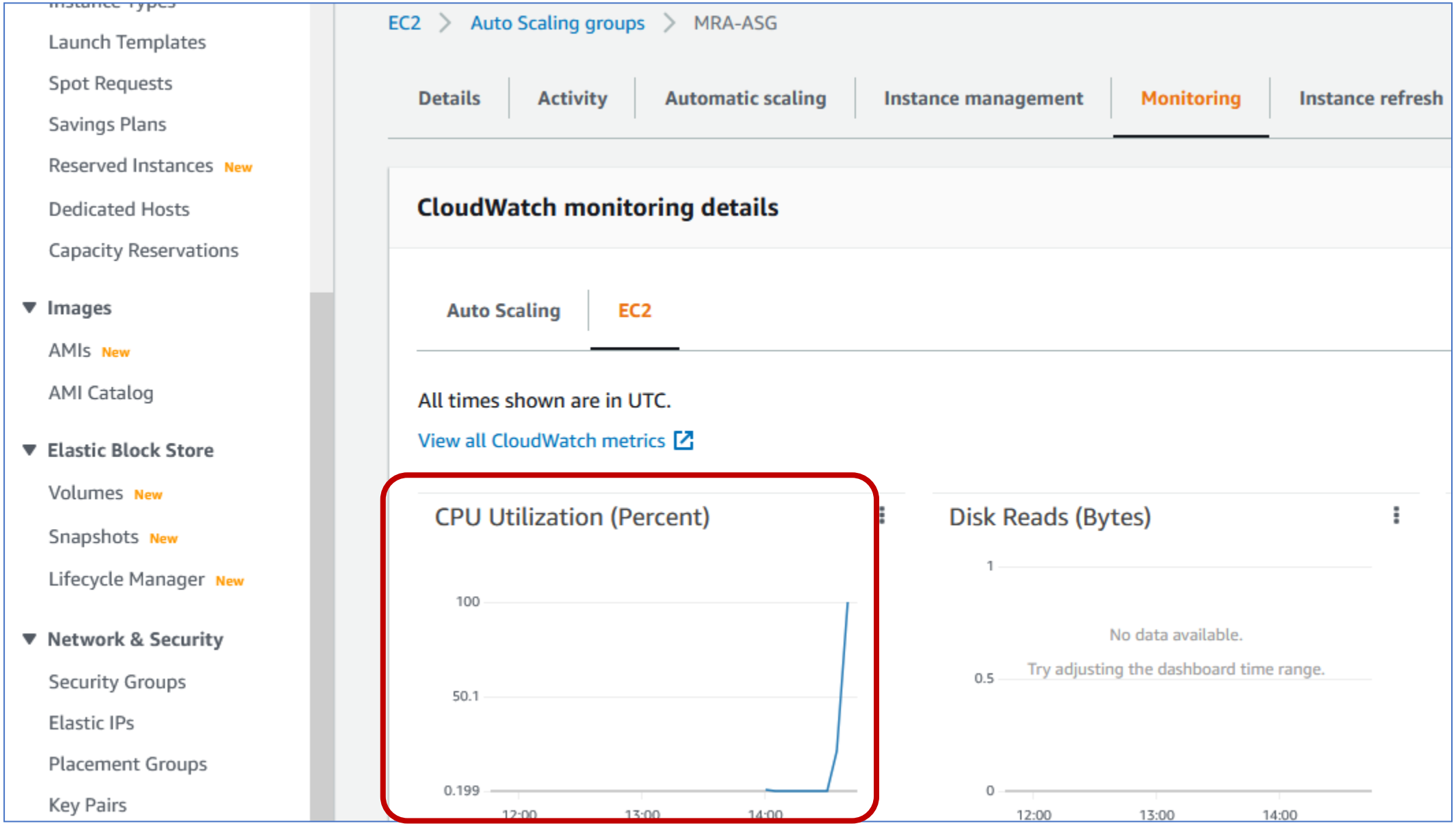



```
sudo apt -y install stress
```

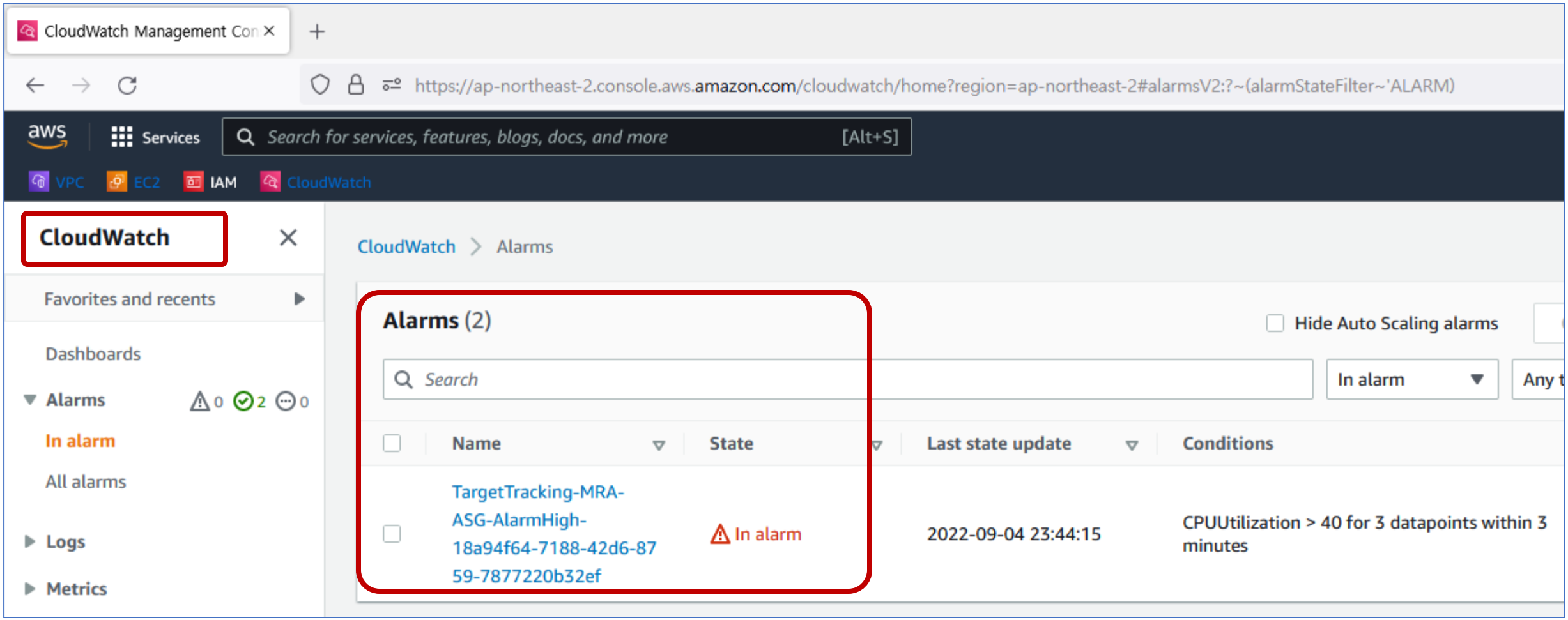
```
stress -c 4
```

```
root@ip-10-0-1-52:~# stress -c 4
stress: info: [1546] dispatching hogs: 4 cpu, 0 io, 0 vm, 0 hdd
```

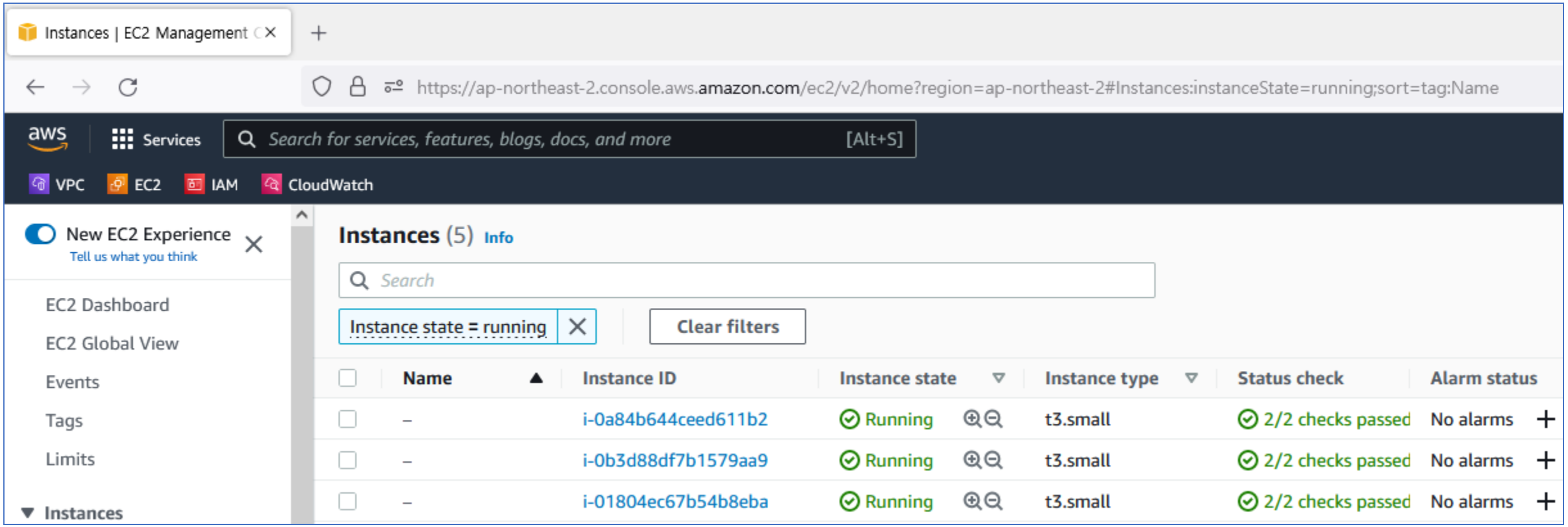
4. Stress Test



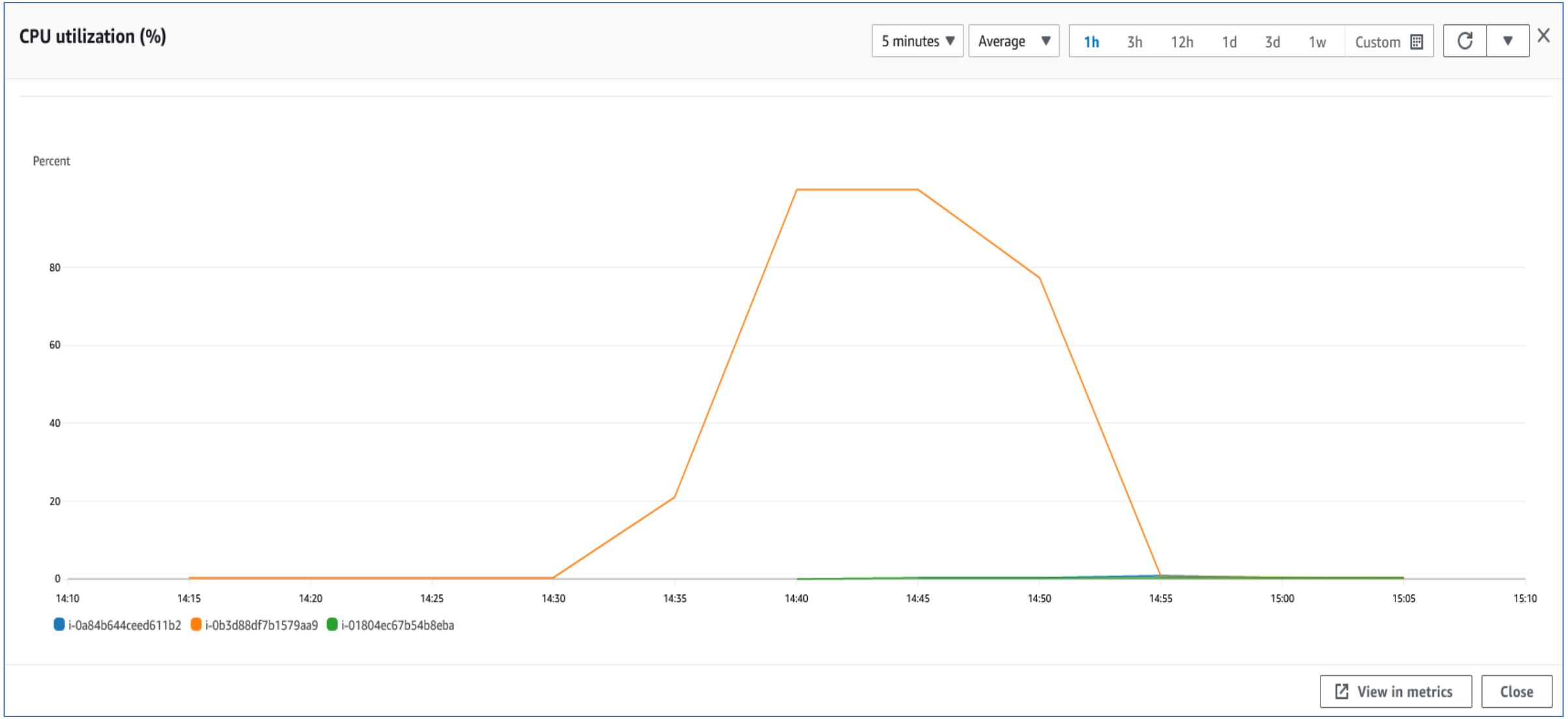
4. Stress Test



4. Stress Test



4. Stress Test



Thank you