HTTPS request

AWS Cloud

Region

VPC ⬡ SSL/TLS cert 172.31.0.0/16

CloudWatch metrics

Availability zone 1a | Availability zone 1b | Availability zone 1c

Subnet 172.31.16.0/20 | Subnet 172.31.32.0/20 | Subnet 172.16.48.0/20

Auto Scaling group

Target group #1 | **Target group #2** | **Target group #3**

Security group

web-srv-1 web-srv-2 | web-srv-1 web-srv-2 | web-srv-1 web-srv-2
SSL/TLS cert | SSL/TLS cert | SSL/TLS cert

Horizontal scaling

Auto scaling groups: Elasticity
scaling in & out (horizontal scaling)

Targets

Load balancers

create launch template

ami, instance type, ram, storage, security group, key pair, user data, az

create target group

Network Load balancer + listener

Application Load balancers + listener

create auto scaling group

parameters, scaling policies

Subnet | Subnet

AZs → each get a node with EIP (static) | AZs → dynamic

generate servers

generated based on min cap and launch template, replaced if down to meet the set max cap

Listener forward to target group | Listener forward to target group

Scaling policies
CloudWatch gather metrics from the auto scaling groups

Dynamic scaling
Simple scaling: scale out whenever a metric is met (e.g: cpu goes over a specific percentage)
Step scaling: add additional alarm breach and launch additional instances if met
Scheduled scaling: scale out at a desired time