Scaling & Monitoring

Cloud Computing

Brenden west

Contents

Learning Outcomes

- Elastic Load Balancing
- Amazon CloudWatch
- Amazon EC2 Auto Scaling

Reading

- AWS Cloud Foundations - Module 10

Elastic Load Balancing (ELB)

- Distributes incoming traffic across multiple targets
- Targets can be in one or more availability zones
- Targets can be EC2 instances, IP addresses, containers, or Lambda functions
-

Types of Amazon Load Balancers

Application

- Operates on application level (OSI layer 7)
- handles HTTP/HTTPs traffic
- Routes traffic based on content

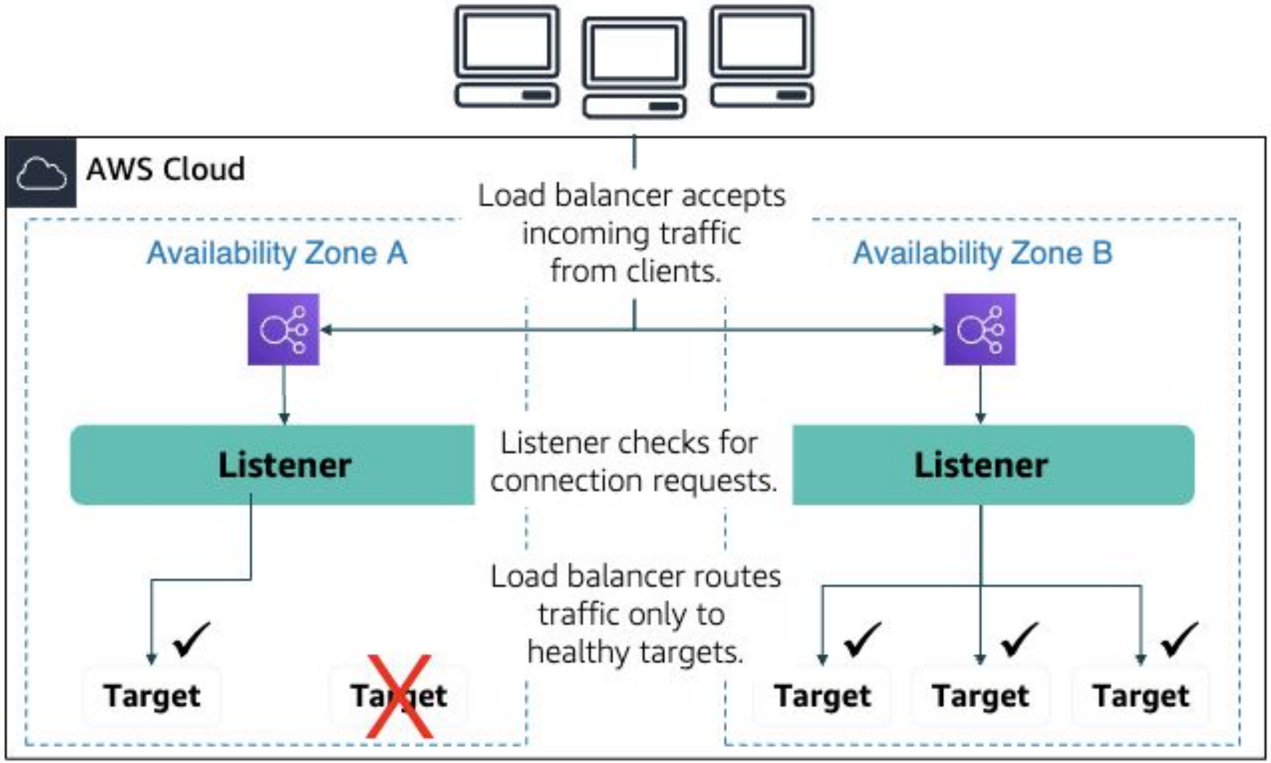
Network

- Operates at network transport level (OSI layer 4)
- Works well for TCP & UDP traffic
- Can handle millions of requests per second

Classic - operates on application & transport layers. Not recommended

How Elastic Load Balancing works

- With Application Load Balancers and Network Load Balancers, you register targets in target groups, and route traffic to the target groups.
- With Classic Load Balancers, you register instances with the load balancer.



Load balancer accepts incoming traffic from clients.

Listener checks for connection requests.

Load balancer routes traffic only to healthy targets.

Load balancer performs health checks to monitor health of registered targets.

How Amazon Load Balancers work

- Configured to accept traffic with **listeners** that check for connection requests
- Registers **target groups** that can receive traffic
- Can be configured to perform health checks on registered targets.
- Only sends traffic to healthy targets
- Automatically balances load across targets
- Can register new EC2 instances provisioned by EC2 Auto Scaling
- Can be configured with security groups to control access for allowed sources
- Can invoke Lambda functions over HTTP(S)
- Integrates with CloudWatch and CloudTrail for monitoring activity

Amazon CloudWatch

- Monitors AWS resources & applications
- Collects standard & custom metrics
- Can send notifications via SNS (Simple Notification Service) or trigger EC2 Auto Scaling based on alarms
- Alarms can be based on a single metric or the result of a math expression
- Can trigger target functions based on defined events

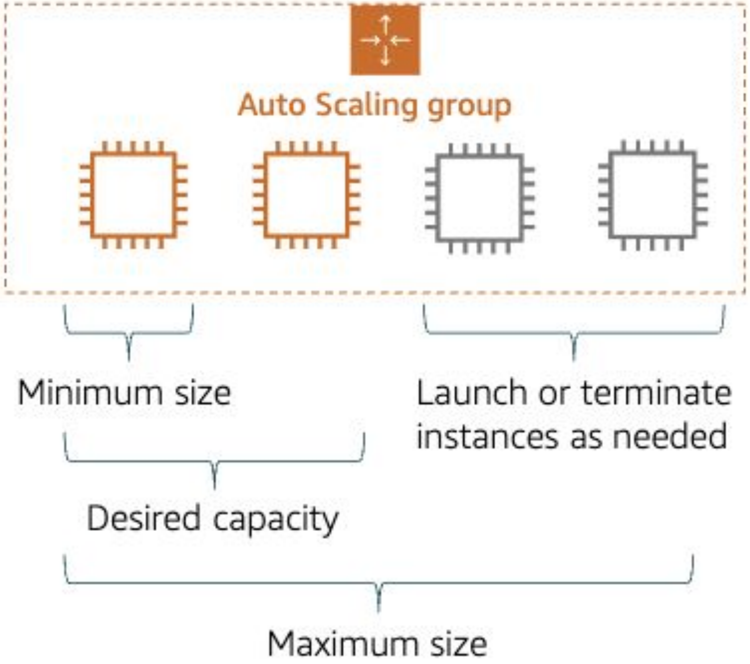
Amazon EC2 Auto Scaling

- Automatically adds or removes EC2 instances to handle changes in demand
- Detects & replaces impaired EC2 instances & unhealthy applications
- Provides a choice of scaling options - manual, scheduled, dynamic (on-demand), & predictive
- Scaling is based on an **Auto Scaling group** with a defined policy for min, max and desired capacity
- Uses a **launch configuration** that specifies the type of EC2 instances
- ELB load balancer can be attached to the Auto Scaling group to be notified about changes to instances

Types of Scaling

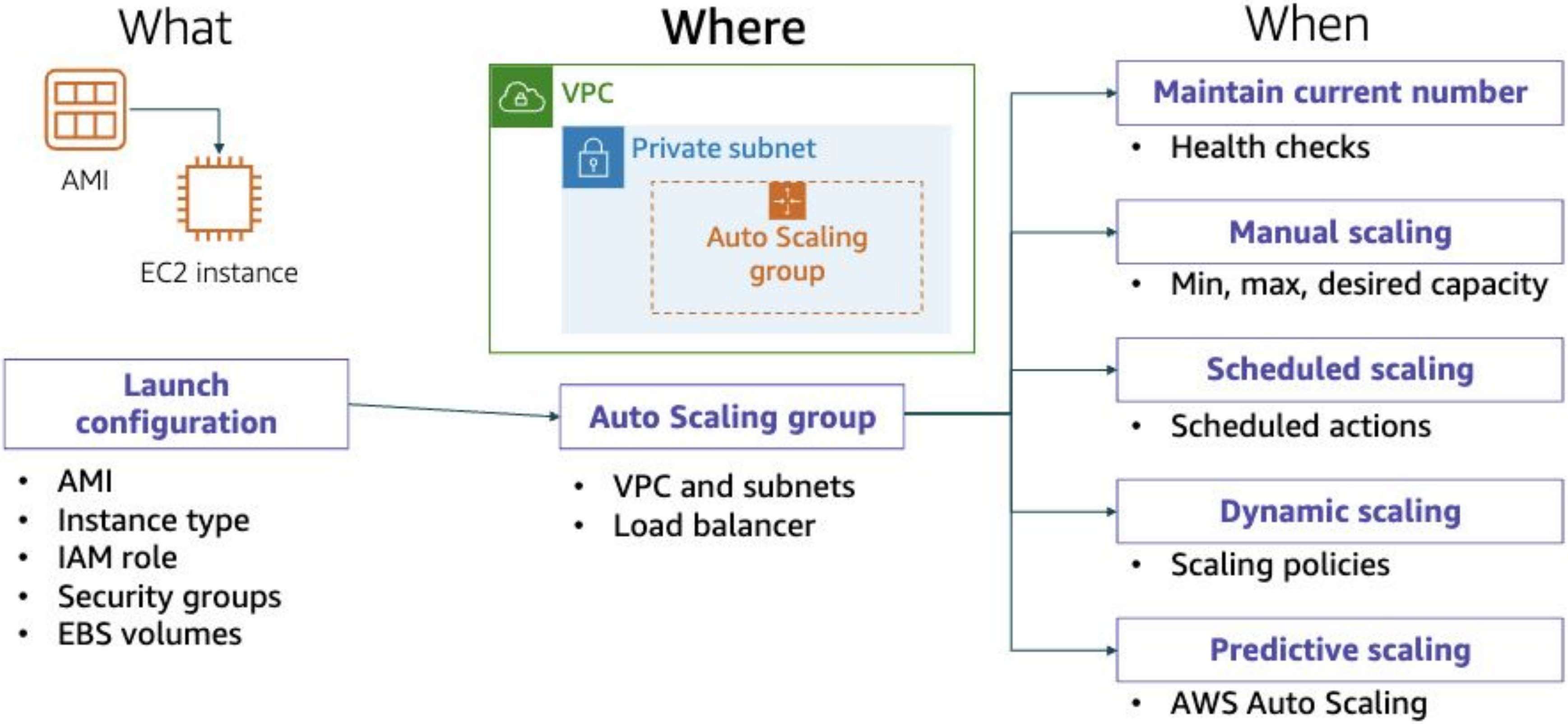
- Manual -
- Scheduled - based on date & time settings
- Dynamic - based on pre-defined load parameters
- Predictive - based on predicted demand

Auto Scaling groups



An **Auto Scaling group** is a collection of EC2 instances that are treated as a logical grouping for the purposes of automatic scaling and management.

How Amazon EC2 Auto Scaling works



What

Launch configuration

- AMI
- Instance type
- IAM role
- Security groups
- EBS volumes

Where

Auto Scaling group

- VPC and subnets
- Load balancer

When

Maintain current number

- Health checks

Manual scaling

- Min, max, desired capacity

Scheduled scaling

- Scheduled actions

Dynamic scaling

- Scaling policies

Predictive scaling

- AWS Auto Scaling