

Automatic Scaling and Monitoring

ELASTIC LOAD BALANCING · ALB / NLB / GLB · CLOUDWATCH METRICS & ALARMS · EC2 AUTO SCALING · DYNAMIC SCALING · LAB 6

01 MUST KNOW

if you remember five things, remember these

- 01 Elastic Load Balancing** distributes incoming application or network traffic across multiple **targets** — EC2 instances, containers, IP addresses, Lambda functions — in one or **multiple Availability Zones**, and scales itself as traffic changes.
- 02 Three load balancer types, by OSI layer:** **Application** Load Balancer = layer **7** (HTTP/HTTPS, content-based routing) · **Network** Load Balancer = layer **4** (TCP/UDP, millions of requests/sec, ultra-low latency) · **Gateway** Load Balancer = layer **3** (virtual appliances).
- 03 Amazon CloudWatch** monitors AWS resources and the applications you run on AWS in **real time**: it collects **standard and custom metrics**, fires **alarms** (→ SNS notification, EC2 Auto Scaling action, or EC2 action), and **Events** rules route environment changes to targets.
- 04 An Auto Scaling group** launches instances from a **launch configuration** (the *what*) into VPC subnets (the *where*) and keeps its size between **minimum** and **maximum**, aiming at **desired capacity**. Launching = **scaling out**; terminating = **scaling in**.
- 05 The trio:** ELB + CloudWatch + EC2 Auto Scaling together deliver **dynamic scaling**. **AWS Auto Scaling** is a *separate* service that builds scaling plans for EC2/Spot Fleets, ECS tasks, DynamoDB tables/indexes, and Aurora Replicas.

02 KEY TERMS

TERM	DEFINITION
Elastic Load Balancing	Distributes app/network traffic across targets in one or more AZs; scales with your traffic
Listener	LB process that checks for connection requests — configured with protocol + port, client→LB and LB→target
Target group	Where you register targets; the load balancer routes traffic to its target groups
Health check	LB test of registered targets — traffic goes only to healthy targets, resumes when they recover
CloudWatch metric	A variable you can measure for a resource or application (e.g. CPU utilization), tracked as time-series data
CloudWatch alarm	Watches a metric or math expression; on breach → SNS notification, EC2 Auto Scaling action, or EC2 action
CloudWatch Events	Rules that match changes in your AWS environment and route them to target functions or streams
Auto Scaling group	Collection of EC2 instances treated as one logical unit for scaling and management
Launch configuration	Instance configuration template: AMI ID, instance type, IAM role, storage, security groups, EBS volumes
Desired capacity	The number of instances the Auto Scaling group tries to maintain, between min and max
Scaling out / in	Out = launching instances (add) · in = terminating instances (remove)
AWS Auto Scaling	Separate service: scaling plans across EC2 + Spot Fleets, ECS tasks, DynamoDB, Aurora Replicas

03 THREE TYPES OF LOAD BALANCERS

pick by OSI layer + traffic

TYPE	LAYER	TRAFFIC	CHOOSE IT WHEN
Application (ALB)	7 — application	HTTP / HTTPS	Content-based request routing; microservices + container apps; latest SSL/TLS ciphers; HTTPS; Lambda functions as targets

TYPE	LAYER	TRAFFIC	CHOOSE IT WHEN
Network (NLB)	4 — transport	TCP / UDP	Millions of requests/sec at ultra-low latency; sudden, volatile (spiky) traffic; static or Elastic IP; IP targets outside a VPC
Gateway (GLB)	3 — network	To virtual appliances	Deploy/scale/manage firewalls, IDS/IPS, deep-packet-inspection systems; transparent single entry-exit point

04 AMAZON CLOUDWATCH AT A GLANCE

FACET	WHAT TO KNOW
Monitors	AWS resources and applications that run on AWS, in real time
Metrics	Standard + custom ; alarm on EC2 CPU, ELB request latency, DynamoDB throughput, SQS queue length — even your AWS bill
Alarm basis	Static threshold · anomaly detection · metric math expression
Alarm anatomy	Namespace (AWS/EC2) → metric → statistic (avg, sum, min, max, sample count, percentile) → period → conditions (>, >=, <=, < + threshold) → additional config (datapoints to alarm, missing-data handling) → actions
Alarm actions	Notify an Amazon SNS topic · EC2 Auto Scaling action · EC2 action
Events	Rules match environment changes → routed to targets (EC2, Lambda, Kinesis streams, ECS tasks, Step Functions, SNS, SQS)
Pricing	No upfront commitment or minimum fee — pay at month end for what you use

05 AMAZON EC2 AUTO SCALING — COMPONENTS

what · where · when

PIECE	ROLE
Launch configuration	What you scale — template with AMI ID, instance type, IAM role, additional storage, security groups, EBS volumes
Auto Scaling group + VPC subnets	Where you scale — logical collection of EC2 instances launched into subnets
Min / max / desired	Group never goes below min or above max; Auto Scaling steers it to desired capacity
Scaling policies	When you scale — adjust instance count between min and max on the criteria you specify
ELB attachment	Attach a load balancer to the group → new instances auto-register and receive traffic
Health replacement	Detects impaired instances / unhealthy apps and replaces them without your intervention

06 FIVE WAYS TO SCALE

OPTION	HOW IT WORKS	USE WHEN
Maintain	Periodic health checks; unhealthy instance → terminated and replaced	You need a fixed number always running
Manual	You change min / max / desired yourself	Ad-hoc adjustments
Scheduled	Scaling actions run automatically by date and time	Predictable load (e.g. traffic rises every Wednesday)
Dynamic	Scaling policies respond to changing demand, often via CloudWatch alarms	You don't know when conditions will change
Predictive	AWS Auto Scaling ML forecast: needs 1 day of history, re-evaluates every 24 h, forecasts 48 h	Daily/weekly patterns; pair with dynamic to scale faster

07 DYNAMIC SCALING — THE THREE SERVICES IN A LOOP

01 CloudWatch alarm breaches (e.g. fleet CPU > 60% for 5 min) → **02 EC2 Auto Scaling** launches instance from the launch configuration → **03** New instance **registered with ELB** → **04 ELB health checks**, then routes traffic to it → **05 ELB feeds metrics back** to CloudWatch

08 EXAM TRAPS

where the knowledge check gets you

- × “The ALB is the one that handles millions of requests per second for TCP.” — That's the **Network** Load Balancer (layer 4). The ALB is layer **7**, for HTTP/HTTPS with content-based routing.
- × “A load balancer keeps sending traffic to a failed instance until you remove it.” — **Health checks** stop routing to unhealthy targets automatically, and traffic resumes when they're healthy again.
- × “A load balancer works within a single Availability Zone.” — ELB distributes across targets in a single AZ **or across multiple AZs** — that's how it delivers high availability.
- × “CloudWatch alarms send alerts themselves.” — The alarm's notification action publishes to an **Amazon SNS topic**; SNS sends the alert. (This is the module's sample exam question.)
- × “CloudWatch only tracks AWS's predefined metrics.” — It also collects **custom metrics** from your own applications and infrastructure.
- × “An alarm just needs a metric and a rough threshold.” — Conditions must use **>, >=, <=, or <** with a threshold value, plus a **statistic** (e.g. average) — “around 60%” is not valid.
- × “Scaling in means adding capacity.” — Scaling **out** = launching instances; scaling **in** = terminating instances.
- × “AWS Auto Scaling is just another name for Amazon EC2 Auto Scaling.” — They are **separate services**. EC2 Auto Scaling scales EC2 fleets; AWS Auto Scaling builds scaling plans for EC2/Spot Fleets, ECS tasks, DynamoDB, and Aurora Replicas — and powers predictive scaling.

09 SELF-QUIZ

answers are upside-down below

- Q1** Which load balancer operates at OSI layer 7 and routes based on the content of the request?
- Q2** Which load balancer handles millions of requests per second at ultra-low latency and suits spiky TCP/UDP traffic?
- Q3** Which load balancer deploys and scales virtual appliances such as firewalls and intrusion detection/prevention systems?
- Q4** What is the load balancer process that checks for connection requests, configured with a protocol and port?
- Q5** Name the three tools for monitoring a load balancer.
- Q6** On which three bases can a CloudWatch alarm be created?
- Q7** What are the three capacity settings of an Auto Scaling group?
- Q8** Which template does an Auto Scaling group use to launch instances, and name three things it specifies.
- Q9** For which four resource types can AWS Auto Scaling build scaling plans?
- Q10** Sample CCP exam: which service would you use to send alerts based on Amazon CloudWatch alarms?

ANSWER KEY (rotate the page)

Q1 Application Load Balancer **Q2** Network Load Balancer **Q3** Gateway Load Balancer **Q4** a listener **Q5** CloudWatch metrics, access logs (to S3), AWS CloudTrail logs **Q6** static threshold, anomaly detection, metric math expression **Q7** minimum, maximum, desired capacity **Q8** launch configuration — AMI ID, instance type, IAM role, storage, security groups **Q9** EC2 instances & Spot Fleets, ECS tasks, DynamoDB tables/indexes, Aurora Replicas **Q10** Amazon Simple Notification Service (SNS)