

Computing (Scaling and Name Resolution)

ELASTIC LOAD BALANCING · ALB / NLB / CLB · EC2 AUTO SCALING · SCALING POLICIES · ROUTE 53 · DNS
ROUTING & FAILOVER

STUDY & REVIEW

01 MUST KNOW

if you remember five things, remember these

- 01 The scaling trio:** a scalable, on-demand architecture wires three services together — **Amazon Route 53** resolves the domain name and points to an **ELB load balancer**, which distributes traffic across the **EC2 instances in an Auto Scaling group**. You pay for the capacity you need, when you need it, instead of buying idle hardware for peak periods.
- 02 Three ELB types:** the **Application Load Balancer (ALB)** works at **Layer 7** — HTTP/HTTPS, content-based routing, containers and microservices; the **Network Load Balancer (NLB)** works at **Layer 4** — TCP/UDP, extreme performance, one static IP per Availability Zone; the **Classic Load Balancer (CLB)** is previous-generation, for the EC2-Classic network.
- 03 Auto Scaling has three parts:** the **launch configuration/template** defines *what* to launch (AMI, instance type, security group, key pair, storage, IAM role, user data); the **Auto Scaling group** defines *where* and the min/desired/max size across AZs; the **scaling policy** defines *when* and by how much. Scale out to meet demand, scale in to cut cost.
- 04 Cooldown vs. warmup:** **simple scaling** uses a **cooldown period** (pause further scaling after an action); **step scaling** uses an **instance warmup period** and does *not* support cooldown. Dynamic policy types are target tracking, step, and simple; **scheduled** and **predictive** scaling handle known, recurring patterns.
- 05 Route 53 routing:** seven routing policies — simple, weighted, latency, failover, geolocation, geoproximity, and multivalued answer. **Latency-based routing** sends users to the fastest Region; **failover** gives active-passive; **weighted** shifts traffic gradually for blue/green deployments.

02 KEY TERMS

TERM	DEFINITION
Elastic Load Balancing (ELB)	Managed service that distributes incoming traffic across multiple targets — EC2 instances, containers, IP addresses, and Lambda functions — in one or more Availability Zones
Listener	Process that checks for connection requests on a configured protocol and port; each load balancer needs at least one and can have up to 50
Target group	Group of registered targets (EC2 instances or ECS containers) that a rule routes requests to; health checks are set per target group
Listener rule	Specifies a target group, a condition, and a priority; the matched condition forwards traffic — the basis of content-based routing
Cross-zone load balancing	Distributes traffic evenly across targets in all enabled AZs; on by default for ALBs, with no extra bandwidth charge
Launch configuration / template	The instance blueprint (AMI, type, security group, key pair, storage, IAM role, user data) that Auto Scaling uses to launch instances
Auto Scaling group (ASG)	Logical group of EC2 instances that scales between a minimum, desired, and maximum size and spreads instances across AZs
Scaling policy	Parameters that define when an Auto Scaling action fires (CloudWatch, health, schedule, manual) and by how much capacity changes
Cooldown period	Pause after a scaling action before another can occur; used with simple scaling policies to prevent thrashing
Instance warmup period	Seconds a newly launched instance takes to warm up; used with step scaling — until it elapses the instance is not counted in group metrics

TERM	DEFINITION
Alias record	Route 53 record that maps a name to selected AWS resources (ELB, CloudFront, S3) or another record in the same hosted zone
Latency-based routing (LBR)	Uses DNS to route each user to the AWS Region that gives the fastest response time

03 THE THREE ELB LOAD BALANCER TYPES

match the layer and protocol to the workload

TYPE	OSI LAYER	TRAFFIC	BEST FOR
Application (ALB)	Layer 7 (application)	HTTP, HTTPS, WebSocket, HTTP/2	Content-based routing, microservices, containers; supports static IP; recommended for most VPC use cases
Network (NLB)	Layer 4 (connection)	TCP, UDP	Extreme performance, ultra-low latency, millions of requests/sec; one static IP per AZ
Classic (CLB)	Request + connection level	HTTP, HTTPS, TCP	Previous generation; applications on the EC2-Classic network only

04 EC2 AUTO SCALING POLICY TYPES

dynamic = target tracking · step · simple

POLICY	HOW IT SCALES	REMEMBER
Target tracking	Holds a metric at a target value, like a thermostat	Best when the metric moves proportionally to instance count
Step scaling	Applies step adjustments sized to the alarm breach	Uses an instance warmup period; does not support cooldown
Simple scaling	Makes a single adjustment per alarm	Uses a cooldown period between actions
Scheduled	Sets a new min/desired/max at specified times	For predictable, recurring load changes
Predictive	ML forecasts load and schedules minimum capacity	Needs ≥1 day of history; forecasts 48h, re-runs every 24h; cyclic spikes only

05 AMAZON ROUTE 53 ROUTING POLICIES

seven policies – know each by its trigger

ROUTING POLICY	USE IT TO...
Simple	Route to a single resource that serves a function for your domain (one web server)
Weighted	Route to multiple resources in proportions you specify — powers blue/green deployments
Latency	Route to the AWS Region that gives the lowest latency when you have resources in several Regions
Failover	Configure active-passive failover to a standby resource
Geolocation	Route based on the location of your users (where the DNS query originates)
Geoproximity	Route based on the location of your resources; optionally shift (bias) traffic between them
Multivalue answer	Return up to eight healthy records selected at random

06 ON-DEMAND SCALABLE ARCHITECTURE

how the pieces connect end to end

01 Client looks up the domain name → **02** **Route 53** resolves it and points to an **ELB load balancer** → **03** **ELB** distributes the request across healthy targets → **04** The **Auto Scaling group** keeps EC2 capacity balanced across AZs → **05** **CloudWatch** alarms fire scaling policies to scale out or in

- × “Use a Network Load Balancer to route on URL path.” — Content-based routing on HTTP/HTTPS is an **ALB (Layer 7)** feature. The **NLB is Layer 4** (TCP/UDP) for extreme performance; it can't see URL paths.
- × “Geolocation and geoproximity are the same.” — **Geolocation** routes on the location of your **users** (the DNS query origin); **geoproximity** routes on the location of your **resources**, and can bias traffic between them.
- × “A launch configuration sets the VPC and subnet.” — It does **not**. The **Auto Scaling group** specifies the VPC/subnet; the launch configuration's only network option is whether to auto-assign a public IP. Only one launch configuration is active at a time.
- × “Cooldown periods apply to every scaling policy.” — Cooldown is for **simple scaling** (and manual, if chosen). **Step scaling** uses an **instance warmup period** and does not support cooldown; neither do target tracking or scheduled scaling.
- × “Predictive scaling handles any traffic spike.” — It uses ML on **daily/weekly patterns** and needs at least a day of history — it's for **cyclic, predictable** spikes only. For random spikes, rely on dynamic scaling.
- × “A steady-state group (min = max = desired = 1) removes the single point of failure.” — It recreates a failed instance automatically, but the instance still recycles with **downtime**, so a single NAT instance built this way is still a SPOF.
- × “An alias record can point anywhere, like a CNAME.” — An **alias** points only to selected **AWS resources** (ELB, CloudFront, S3) or another record in the same hosted zone. A **CNAME** can redirect to any DNS name.
- × “Cross-zone load balancing costs extra and must be turned on for ALBs.” — It is **on by default** for Application Load Balancers, with **no additional bandwidth charge** for cross-zone traffic.

08 SELF-QUIZ

answers are upside-down below

- Q1** At which OSI layer does a Network Load Balancer operate, and which protocols does it handle?
- Q2** Which load balancer type supports content-based routing of HTTP/HTTPS traffic at Layer 7?
- Q3** Name the three parts of an EC2 Auto Scaling configuration.
- Q4** A launch configuration does not specify the VPC or subnet — which component does?
- Q5** Which dynamic scaling policy uses step adjustments sized to the alarm breach, and which timing setting does it rely on?
- Q6** Which scaling policy type keeps a chosen CloudWatch metric at a target value?
- Q7** How much history does predictive scaling need before forecasting, and how far ahead does it forecast?
- Q8** Which Route 53 routing policy configures active-passive failover?
- Q9** Which Route 53 routing policy shifts traffic gradually and enables a blue/green deployment?
- Q10** Which Route 53 record type can point only to selected AWS resources such as an ELB, CloudFront, or S3?

ANSWER KEY (rotate the page)

Q1 Layer 4 (connection level); TCP and UDP **Q2** the Application Load Balancer (ALB) **Q3** launch configuration/launch template, Auto Scaling group, and scaling policy **Q4** the Auto Scaling group **Q5** step scaling; an instance warmup period (no cooldown) **Q6** target tracking scaling **Q7** at least 1 day of data; it forecasts the next 48 hours, re-evaluated every 24 hours **Q8** the failover routing policy **Q9** the weighted routing policy **Q10** an alias record