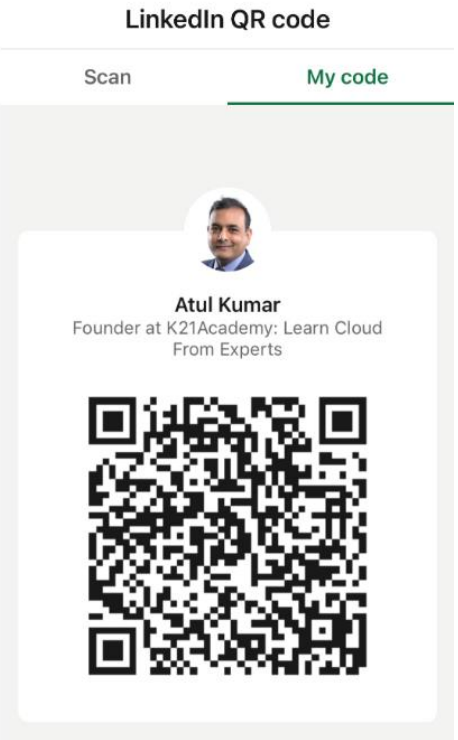
K21Academy
Learn Cloud from Experts

ORACLE Gold Partner | Silver Microsoft Partner | DevOps Institute Registered Education Partner

Load Balancing, Auto Scaling & Route 53

Atul Kumar



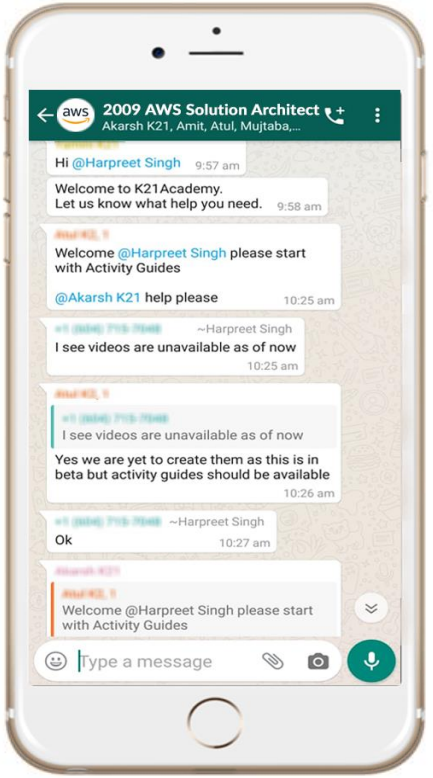
- Author & Cloud Architect
- 21+ Years working in IT & Certified Cloud Architect
- Helped **8500+ individuals** to learn Cloud & Cloud Native

WhatsApp & Ticketing System

K21Academy
Learn Cloud from Experts

support@k21academy.com

zendesk



Module Agenda

Module: Agenda

- AWS Load Balancer
- Types Of Load Balancer
- Components Of Application Load Balancer
- Load Balancer Troubleshoot
- AWS Global Accelerator
- AWS Auto - Scaling
- AWS Auto - Scaling Components
- Life Cycle of Auto Scaling
- Auto Scaling Policy
- Route 53
- Various Routing Policies

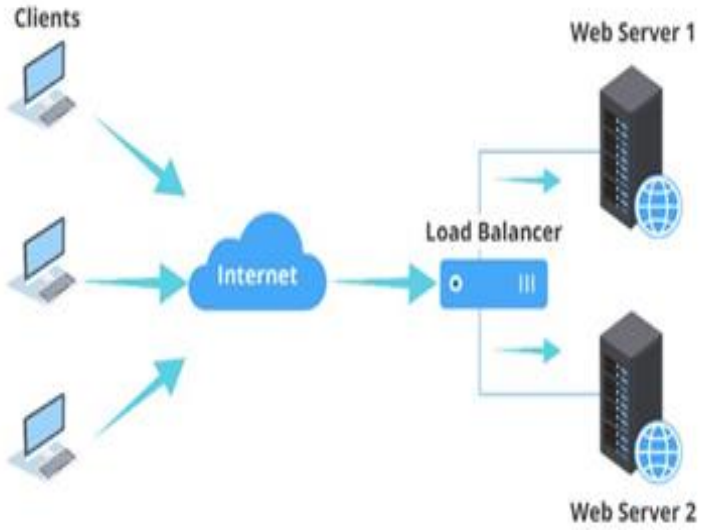
K21Academy
Learn Cloud from Experts

Load Balancer

Load Balancer

K21Academy
Learn Cloud from Experts

- Elastic Load Balancer distributes and manages the incoming traffic load among several devices to improve network performance.
- Distributes Client traffic across servers.
- Improves the performance of applications.



K21Academy
Learn Cloud from Experts

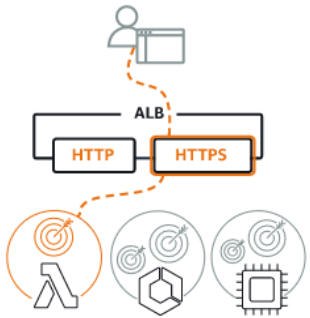
Types Of Load Balancer

Types Of Load Balancers

K21Academy
Learn Cloud from Experts

Load balancer types

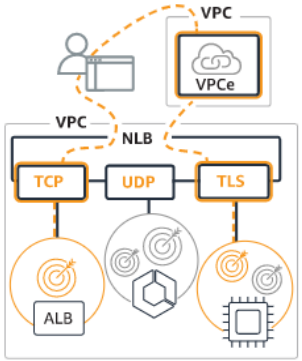
Application Load Balancer Info



Choose an Application Load Balancer when you need a flexible feature set for your applications with HTTP and HTTPS traffic. Operating at the request level, Application Load Balancers provide advanced routing and visibility features targeted at application architectures, including microservices and containers.

Create

Network Load Balancer Info



Choose a Network Load Balancer when you need ultra-high performance, TLS offloading at scale, centralized certificate deployment, support for UDP, and static IP addresses for your applications. Operating at the connection level, Network Load Balancers are capable of handling millions of requests per second securely while maintaining ultra-low latencies.

Create

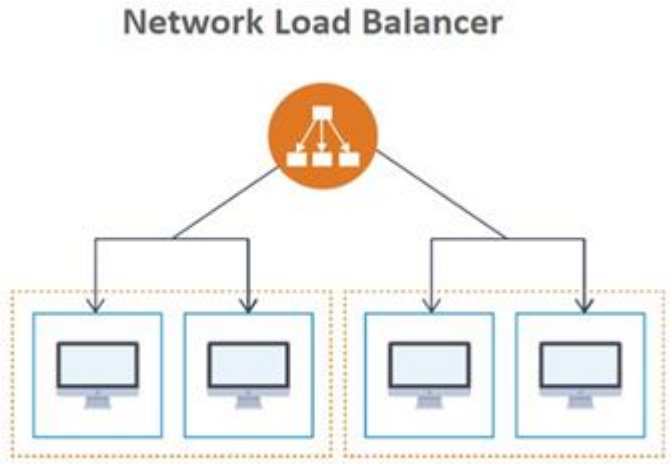
Gateway Load Balancer Info

Choose a Gateway Load Balancer when you need to deploy and manage a fleet of third-party virtual appliances that support GENEVE. These appliances enable you to improve security, compliance, and policy controls.

Create

Network Load Balancer

- **Network Load Balancer** handles sudden and violates traffic across the EC2 Instances in order to avoid any latency.
- New layer 4 load balancing platform.
- Connection base load Balancing.
- Supports TCP protocol.
- Can handle millions of requests/ seconds.



Limitation

K21Academy
Learn Cloud from Experts

Regional Limits per Region	
Number of Network LB	20
Target groups	3000

Components Limit per LB	
Listeners	50
Targets per Availability Zone With Cross-zone load balancing disabled	200
Targets With Cross-zone load balancing enabled	200
Subnets per Availability Zone	1

Application Load Balancer (ALB)

K21Academy
Learn Cloud from Experts

- The **Load Balancer** that distributes the traffic to appropriate target groups on the basis of content is called **Application Load Balancer**.
- New feature rich, layer 7 load balancing platform.
- **Reduces** hourly cost.
- Supports **web sockets, HTTP** and **HTTPS**.
- Supports micro services and container based application, including deep integration with EC2 container service.

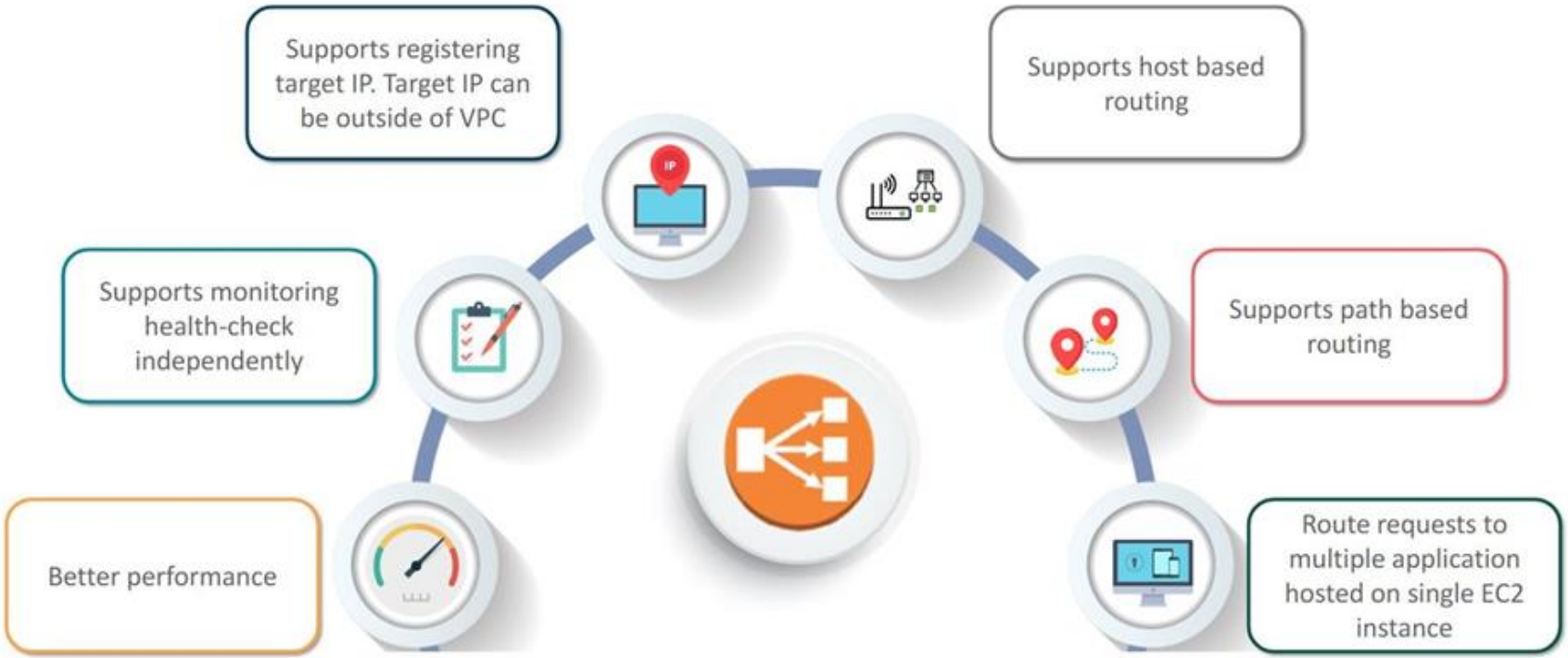
Limitations

K21Academy
Learn Cloud from Experts

Regional Limits	
LB Per Region	20
Target groups/Region	3000

Load Balancer Components Limit	
Listeners/ load balancer	50
Targets /load balancer	1000
Subnets /Availability Zone per load balancer	1
Security groups / load balancer	5
Rules(not counting default rules)	100
Certificates (not counting default certificates)	25
Number of times a target can be registered	100

Key Benefits



Gateway Load Balancer

K21Academy
Learn Cloud from Experts

- **Gateway Load Balancer** makes it easy to deploy, scale, and manage your third-party virtual appliances.
- This eliminates potential points of failure in your network and increases availability.
- Scale your virtual appliance instances automatically.
- Monitor continuous health and performance metrics.
- Runs within one Availability Zone (AZ).

Limitations Of Gateway Load Balancer

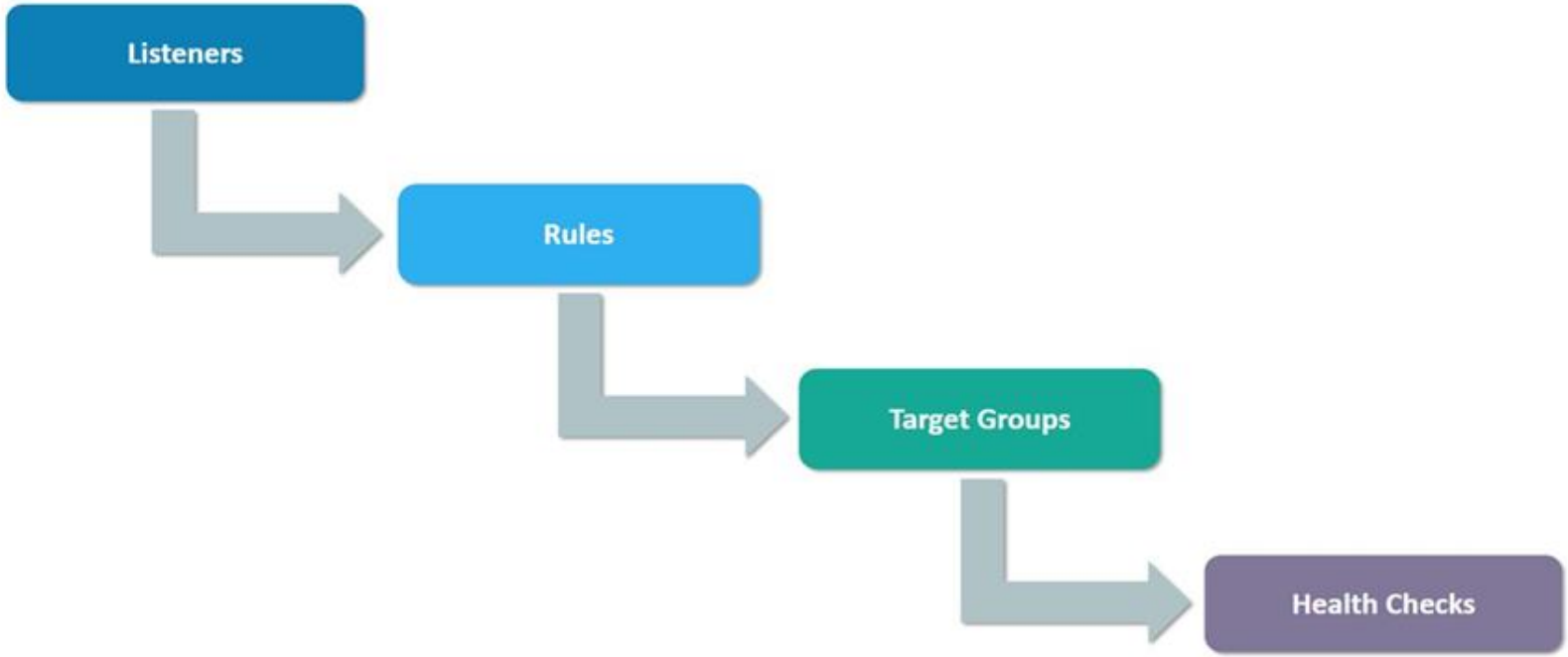
K21Academy
Learn Cloud from Experts

Regional Limit	
LB Per Region	20
Load Balancer Components Limit	
Gateway Load Balancers per VPC	10
Target groups with GENEVE protocol	100
Targets per Availability Zone per target group with GENEVE protocol	300

K21Academy
Learn Cloud from Experts

Components Of Application Load Balancer

Components of ALB



K21Academy
Learn Cloud from Experts

Load Balancer Troubleshoot

Load Balancer Troubleshoot

- Under the following circumstances you will not be able to connect to your Load Balancer.
- The registered target is not service.
- Client cannot connect to Load Balancer having Internet facing configured.
- Load Balancer sending requests to unhealthy instances.
- Load Balancer generates HTTP error.

Load Balancer Troubleshoot

K21Academy
Learn Cloud from Experts

HTTP Error Code	Reason
400	Header size exceeding the limit of 16k. Request is not following HTTP standards
403	Either Security Group or ACL is blocking your request
460	Timed out. Check ELB default timeout
463	If ELB an X-Forwarded-For request header with more than 30 IP addresses
502	Bad Gateway. E.g. SSL handshake failed or wrong port number and many more
503	Service Unavailable. If target group is missing
504	Gateway timeout. ELB failed to establish a connection with target group

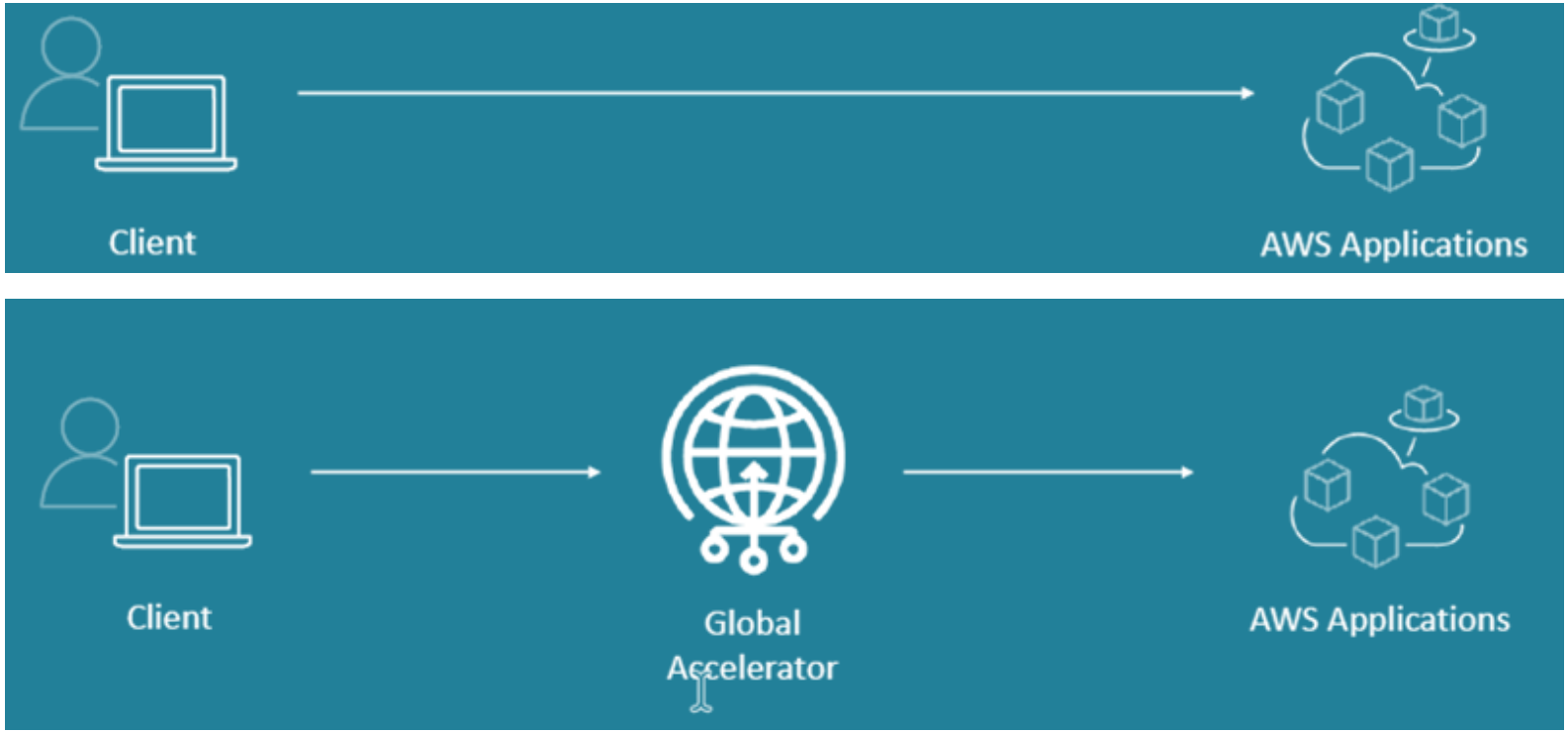
K21Academy
Learn Cloud from Experts

AWS
Global Accelerator

Global Accelerator

- A network layer service that you can deploy in front of your internet-facing applications to improve availability and performance for your globally distributed users
- A network layer service that you can deploy in front of your internet-facing applications to improve availability and performance for your globally distributed users

Global Accelerator

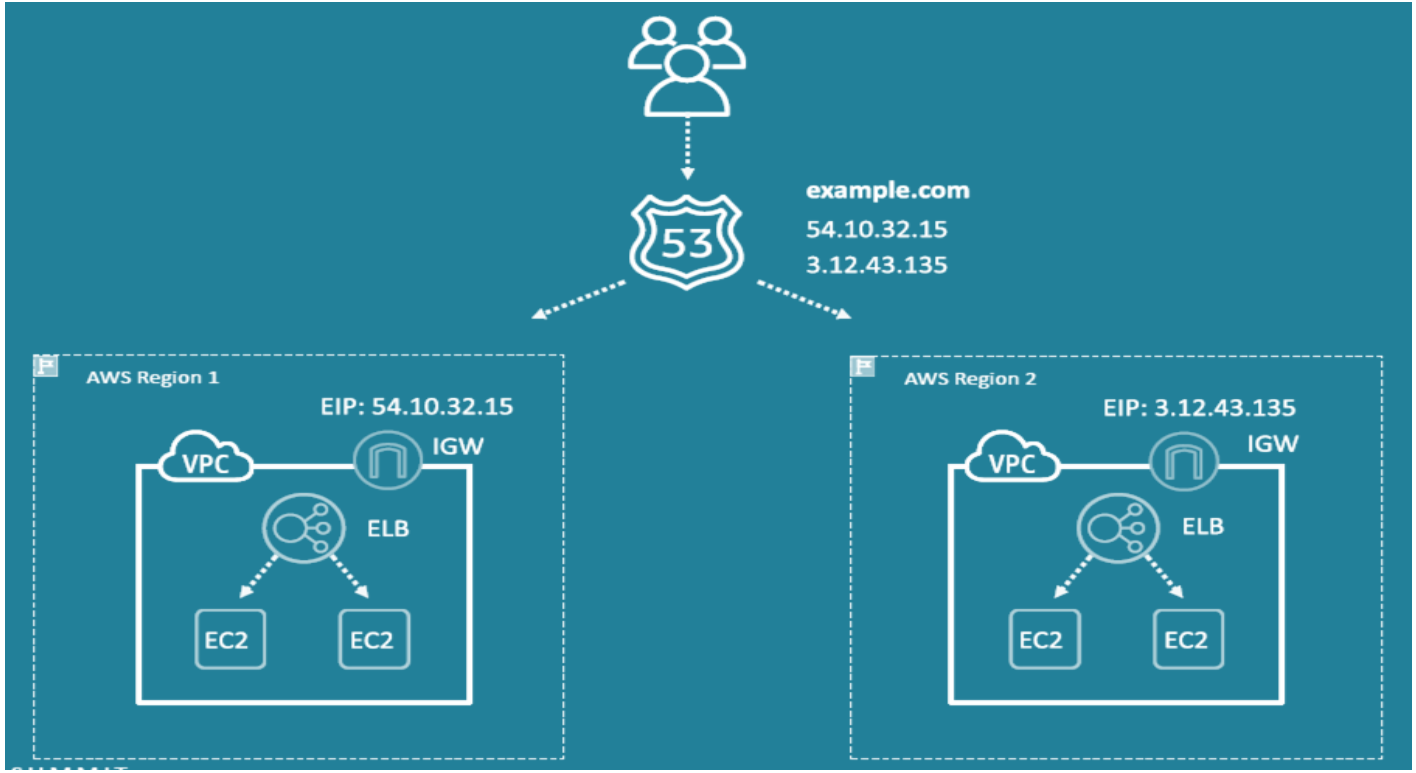


Global Accelerator – How it Work

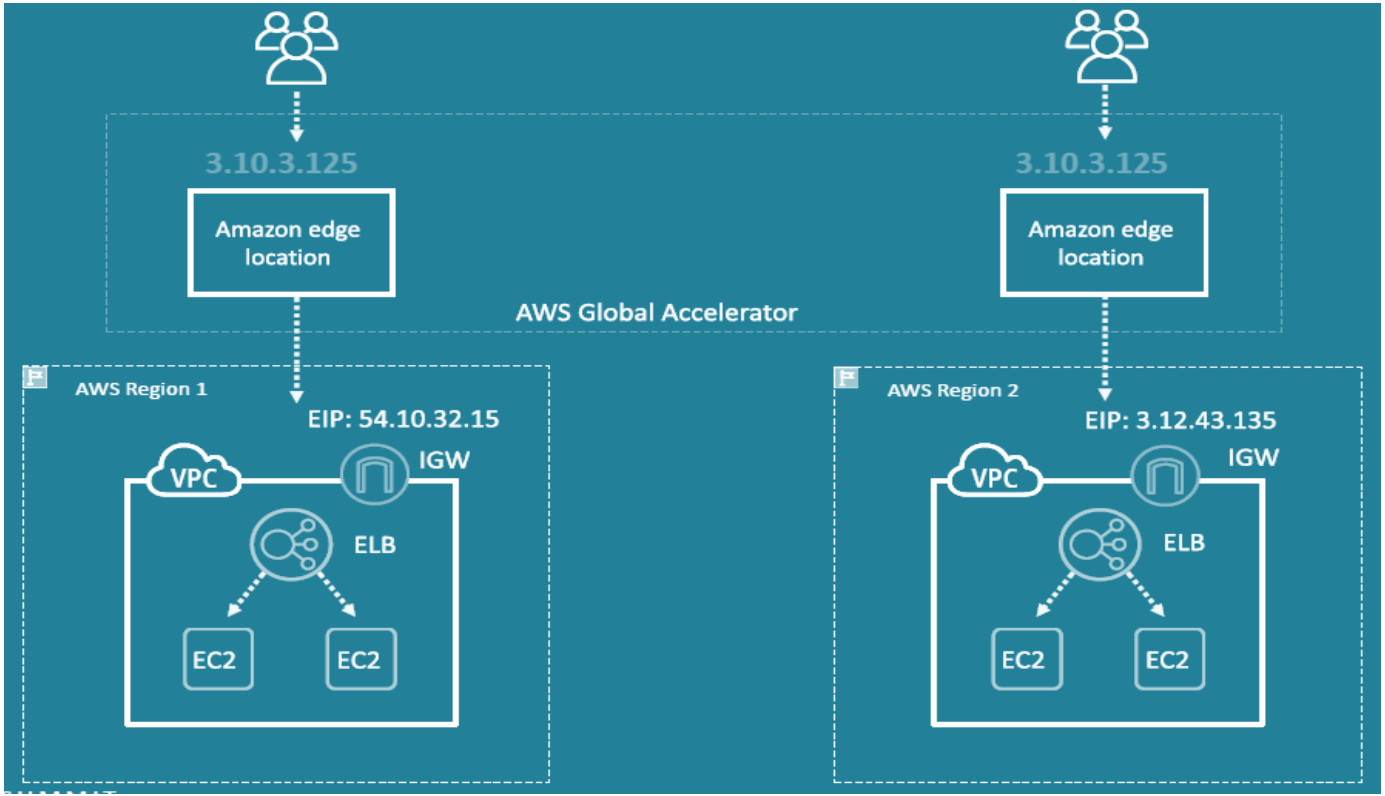
Global Accelerator – How it Work

Global Accelerator – How it Work

Taking a Closer Look..



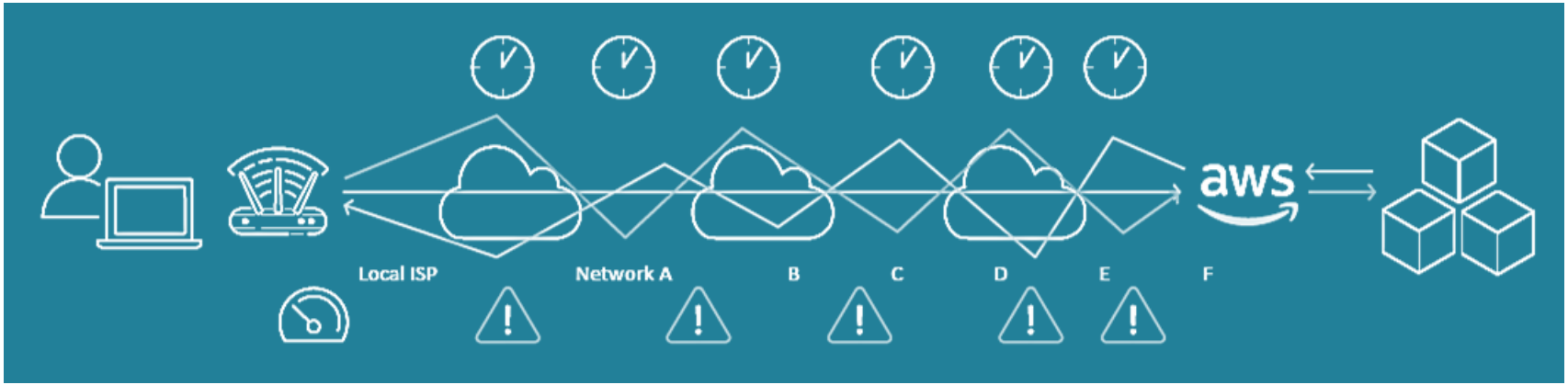
Taking a Closer Look..



Key Features

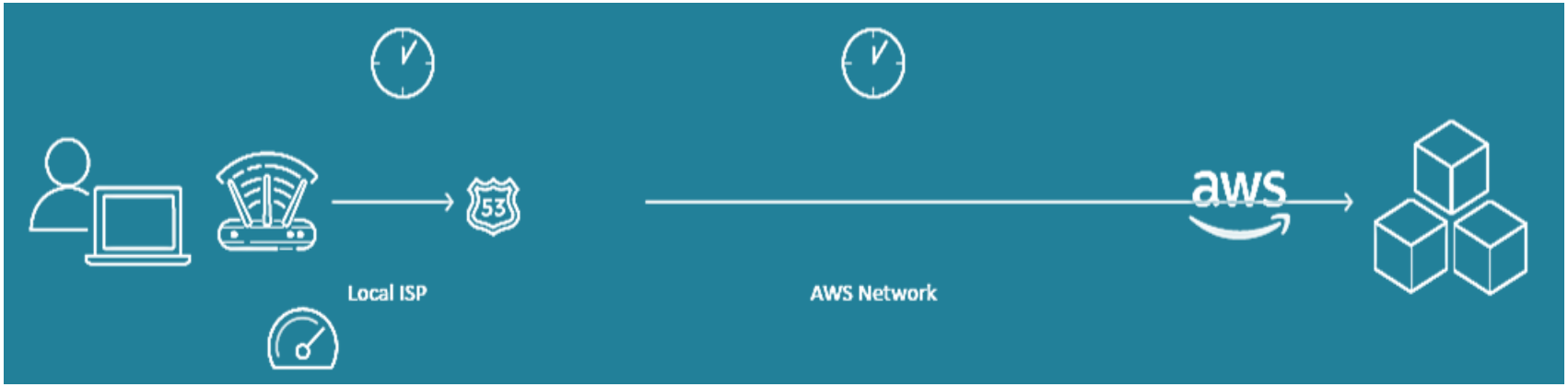
Without Global Accelerator

- Accessing your application is not this straight forward! It can take many networks to reach the application Paths to and from the application may differ Each hop impacts performance and can introduce risk



With Global Accelerator

- Traffic enters AWS global network at edge locations Leverages the Global AWS Network Resulting in improved performance.



Use Cases

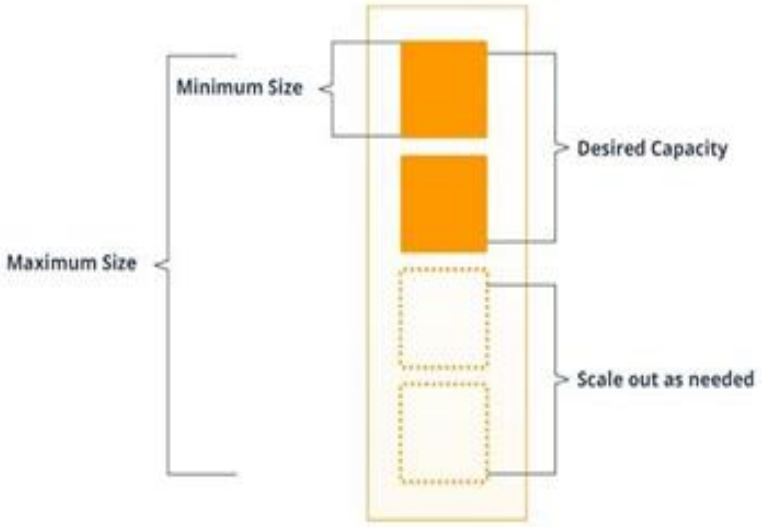
- IP whitelisting
  - Automotive / Connected vehicles
  - Retail / Payments transactions
  - Healthcare
  - IOT
- UDP Traffic
  - Gaming
  - Telco — Voice over IP (VOIP)
  - Telco DNS
- Live Video ingest
  - Media
- Multi-region applications
  - Financial Services
  - DR / Failover for applications
- API Acceleration
  - Media
  - Ad-tech

K21Academy
Learn Cloud from Experts

Auto-Scaling

Auto Scaling

- **Auto Scaling** is an AWS computing Service that automatically scales **up** or **down** compute resources as per their usage.
- Collection of EC2 instances is called **Auto Scaling Groups**.
- With Auto Scaling our application has right resource at right time.
- Auto Scaling, AWS Cloud Watch, AWS ELB work in **union**.



Auto Scaling Benefits

K21Academy
Learn Cloud from Experts

Auto-Scaling Components

Auto Scaling Components

Groups

Launch Configuration

Scaling options

EC2 Auto Scaling

K21Academy
Learn Cloud from Experts

- Amazon EC2 Auto-Scaling helps us to take care of our application availability and allows us to add or remove EC2 instances automatically according to the conditions defined.
- The EC2 Auto Scaling services offers a way to both avoid application failure and recover it when it happens.
- EC2 Auto Scaling uses either a launch configuration or a Launch template to automatically configure the instances that it launches.

Launch Configuration

K21Academy
Learn Cloud from Experts

- **Launch Configuration** is a template that defines the parameters passed to launch EC2instances in the Groups.
- It specifies AMI ID, instance type, key pair, security groups and block device for mapping the EC2 Instances.
- Launch configuration can’t be modified after creation.

Launch Template

- Launch Template identical to a launch configuration, with additional features.
- Launch Template allows multiple versions of a template to be defined.
- Launch Template allows selection of both Spot and On-Demand Instances or multiple instance types.
- Launch templates support EC2 Dedicated Hosts. Dedicated Hosts are physical servers with EC2 instance capacity that are dedicated to your use.

Auto Scaling Groups

K21Academy
Learn Cloud from Experts

- Auto Scaling group is a group of EC2 Instances that Auto Scaling Manages.
- They are the logical units for scaling and management.
- Here minimum, maximum and desired number of EC2 instances can be specified.
- It maintains the number of instances by performing the periodic health checks on the Instance in the Groups.

Auto Scaling Options

K21Academy
Learn Cloud from Experts

- **Fixed:** To maintain the same number of Instances all the time.
- **Manual:** Can specify the change in the minimum, maximum and desired capacity of an Auto Scaling Groups.
- **Scheduled:** Actions are performed automatically as a function of time and date.
- **Dynamic:** At any time you can change the size of an Auto Scaling Group by updating the desired capacity or updating the instances that are attached to the Auto Scaling Group.

Scaling Policy Types

K21Academy
Learn Cloud from Experts

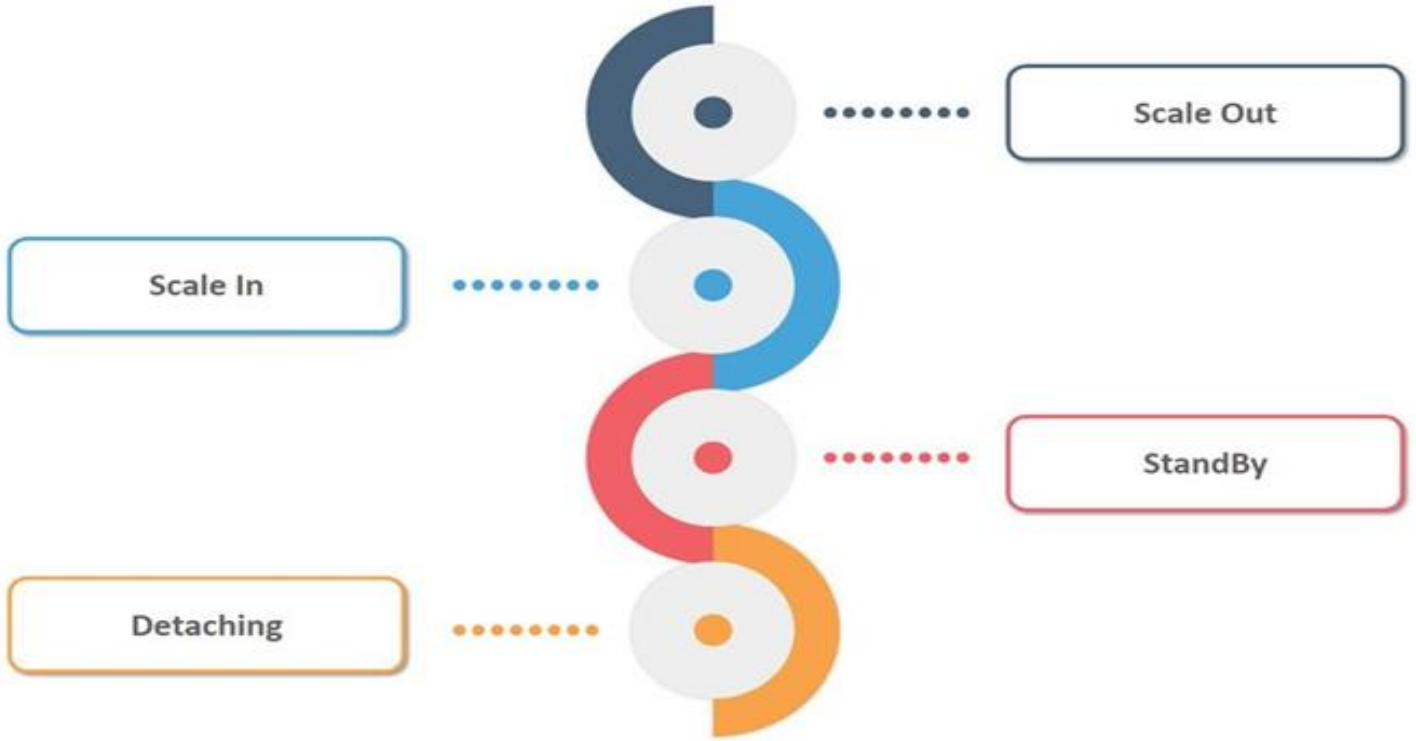
Amazon EC2 Auto Scaling supports the followings types of scaling policies:

- **Target tracking scaling:** Increases or decrease the current capacity of the group based on a target value for a specific metric.
- **Step scaling:** Increase or decrease the current capacity of the group based on a set of scaling adjustments, known as *step adjustments*, that vary based on the size of the alarm breach.
- **Simple scaling:** Increase or decrease the current capacity of the group based on a single scaling adjustment.

K21Academy
Learn Cloud from Experts

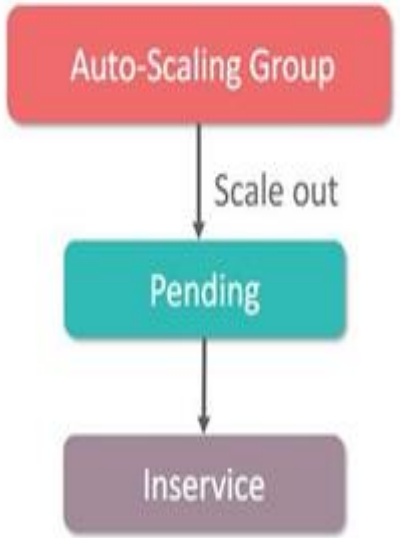
Life Cycle Of Auto-Scaling

Life Cycle of Auto Scaling



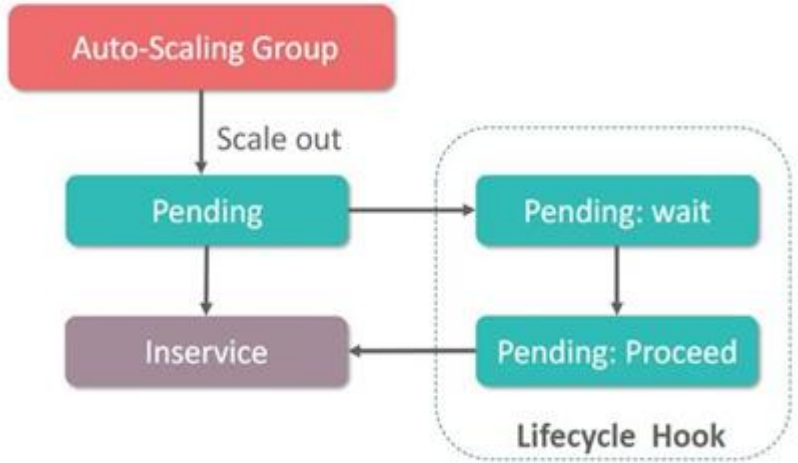
Life Cycle of Auto Scaling

- **Scale Out** is the process of launching an EC2 instance via Launch Configuration.
- These launched Instances enter the pending state.
- When all the Instances are configured completely then they enter the **In-service state.**



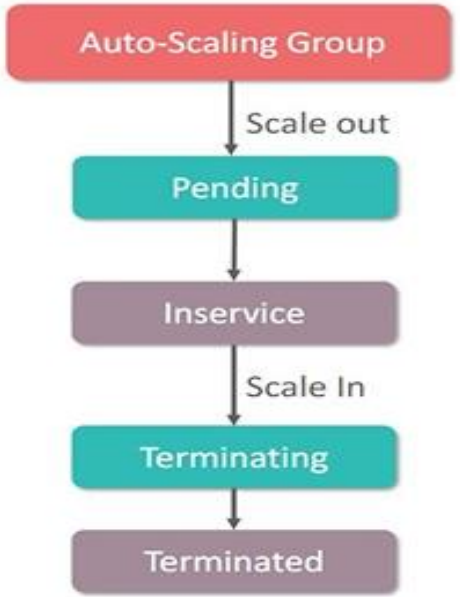
Life Cycle of Auto Scaling

- **Life Cycle Hook** can be added to Auto Scaling Group to perform the custom actions at the time of launch or termination of Instance.
- With the help of Life cycle Hook , you can install or configure software or newly launched Instances .



Life Cycle of Auto Scaling

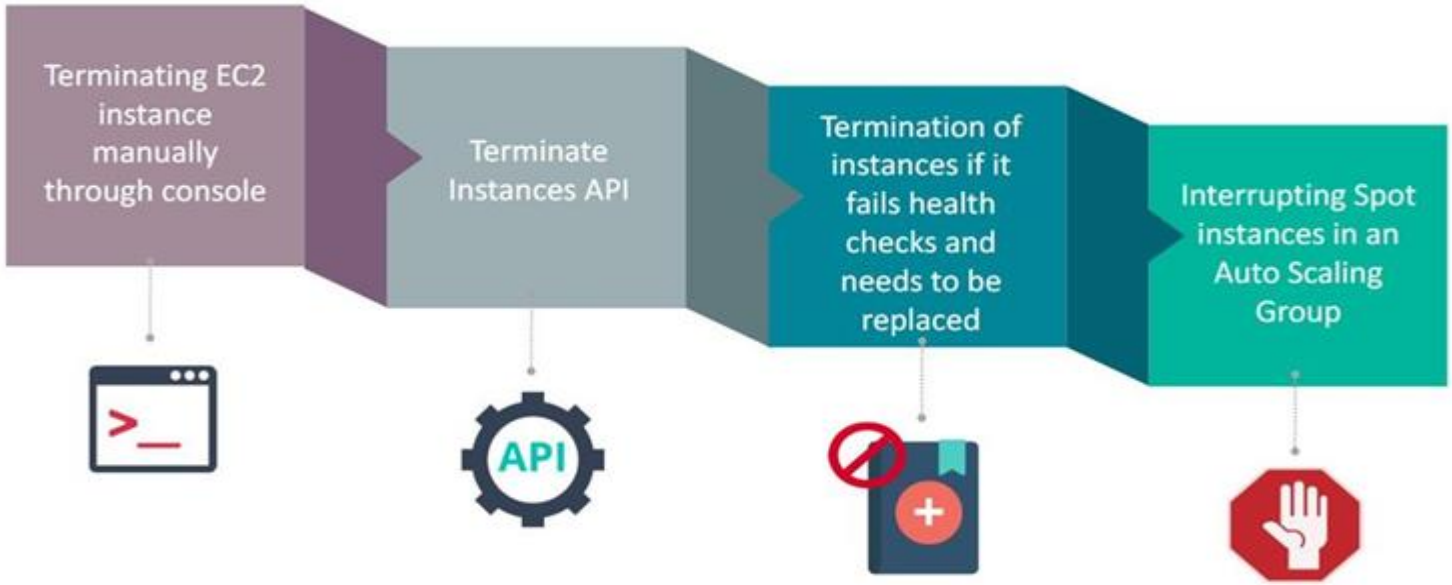
- **Scale In** is process which ensure resources attached to your application match the demands.
- It can terminate one or more instances.
- The Auto Scaling group using the termination policy determines which instances are to be terminated.



Life Cycle of Auto Scaling

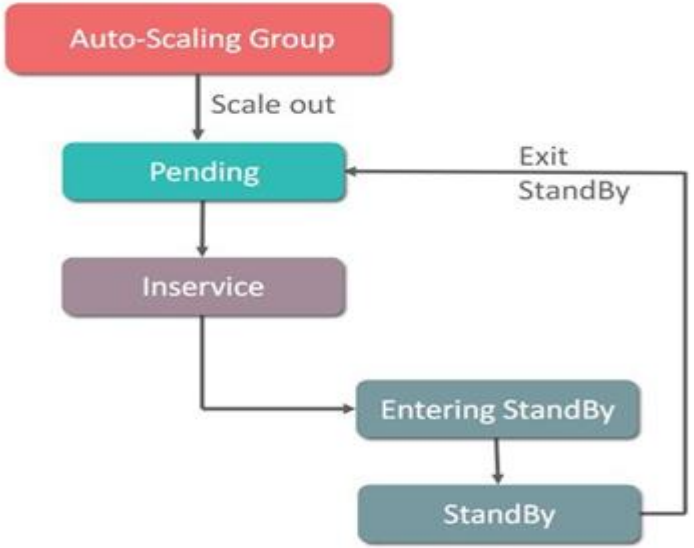
K21Academy
Learn Cloud from Experts

- Instance Protection does not work in the following state.



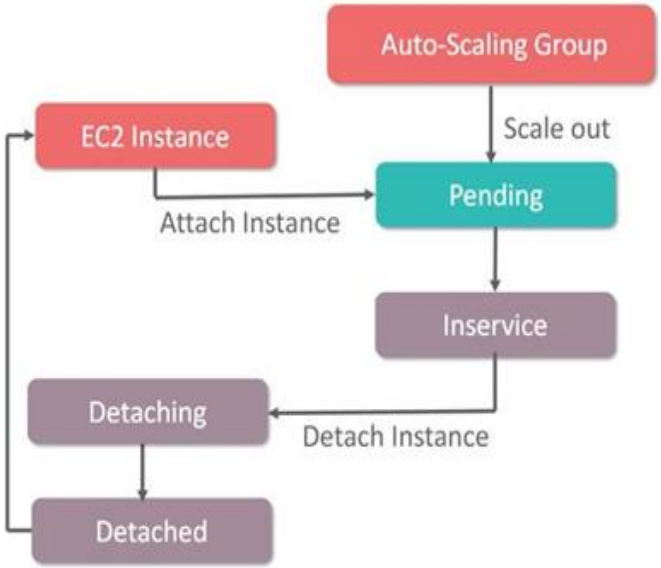
Life Cycle of Auto Scaling

- Instances in the **Stand By** state continue to be managed by the Auto Scaling Group until they are not put back into the service, they will not be active part of application.
- Stand By state can be used to **update**, **modify** or **troubleshoot** instances.

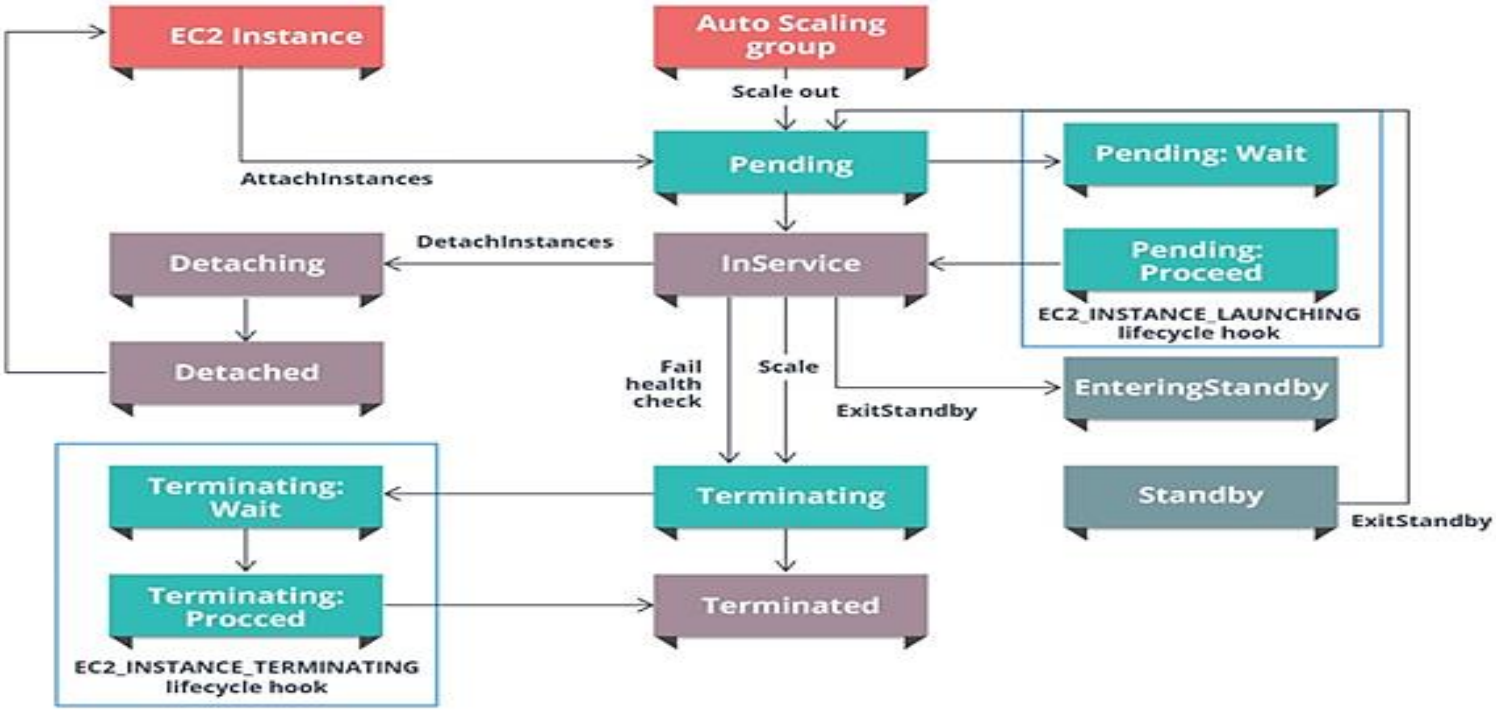


Life Cycle of Auto Scaling

- **Detaching State** is used for architectural change or update, to find the best fit for the application.
- After the instances are detached they can be moved separately from a Auto Scaling or attached to a different Auto Scaling Group.



Life Cycle of Auto Scaling



K21Academy
Learn Cloud from Experts

Auto Scaling Policy

Auto Scaling Policy

K21Academy
Learn Cloud from Experts

- Scaling Policy is a set of Instructions for Auto Scaling that tells the service how to respond to AWS CloudWatch Alarm mistakes.
- It specifies whether to scale the Auto Scaling group **up** or **down** by how much.
- It adjusts the number of instances present in the Auto Scaling Group according to the specified criteria.

Limitations of Auto Scaling

K21Academy
Learn Cloud from Experts

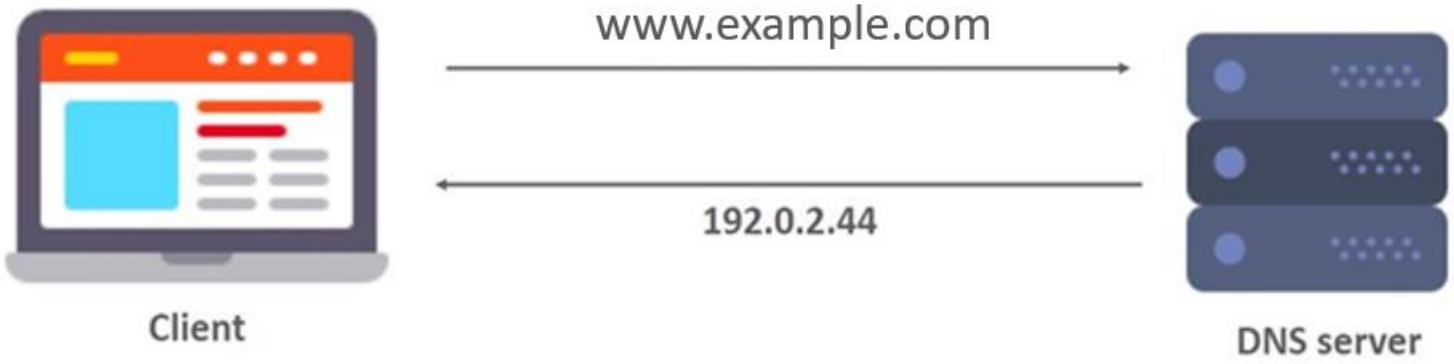
Regional Limits Per region	
Launch configurations	200
Auto Scaling groups	200
Limits Per Auto Scaling Group	
Scaling policies	50
Scheduled actions	125
Lifecycle hooks	50
SNS topics	10
Target groups	50
Scaling Policy Limits	
Step adjustments	20

K21Academy
Learn Cloud from Experts

AWS Route 53

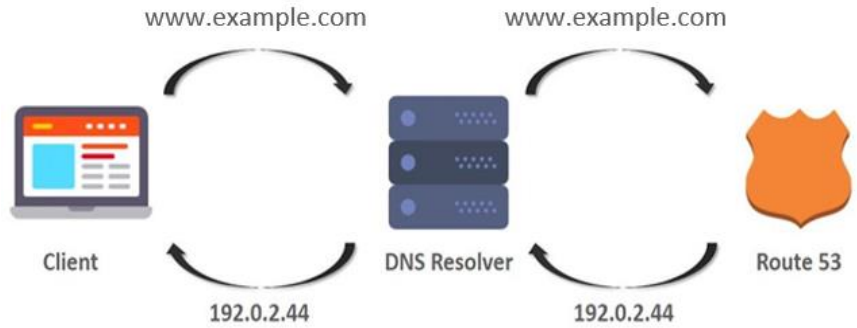
Why is DNS Used

- **DNS** Translates host name into a computer friendly IP address.
- DNS is used for managing the public names of websites and other domains.



Why Do We Need Route 53

- For hosting website we need domain name and domain name system to be accessed by any user.
- The IP address of your local DNS will be searched at your ISP.
- If the website is not listed in your local DNS, it will find on other DNS until it finds the match.
- To reduce the Hops Route 53 was introduced.

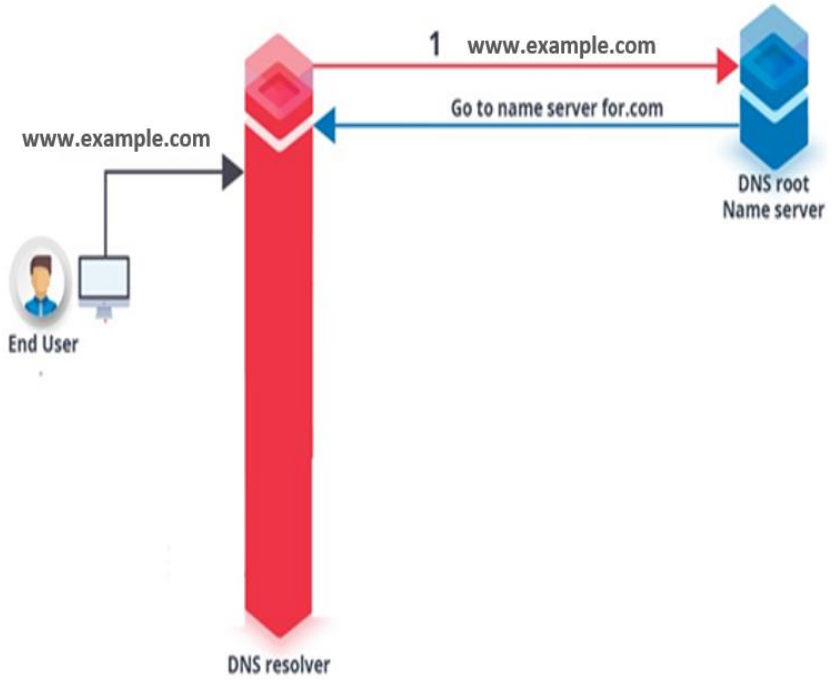


Why Do We Need Route 53

- AWS named it as **Route 53** as all the requests are handled through **Port 53.**
- **Route 53** is a reliable and cost effective way too route end users to the internet applications.
- Connects user requests to infrastructure running in and outside AWS.

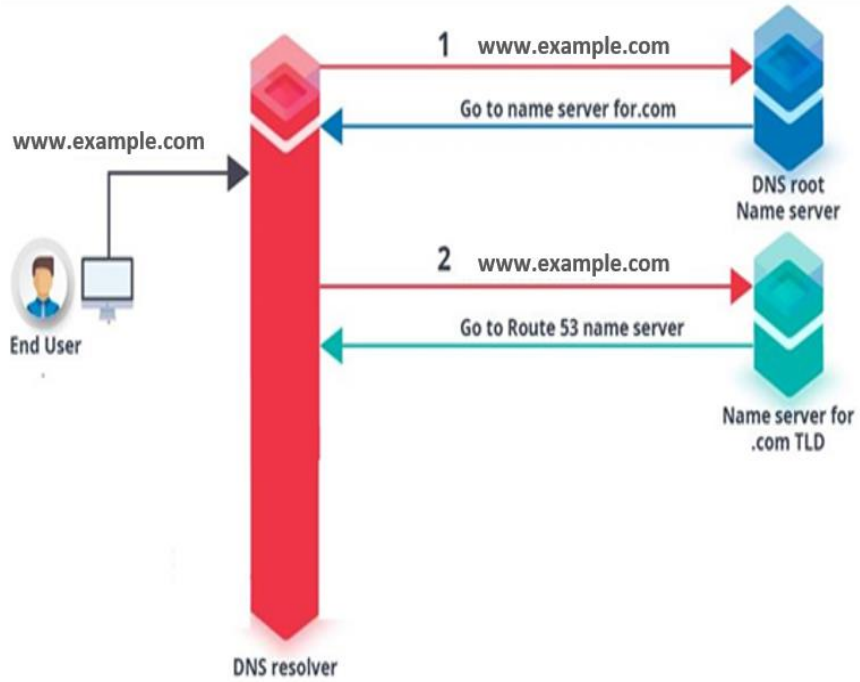
Working of Route 53

- When the request is made in the browser, it is routed to the DNS resolver.
- DNS resolver is managed by the ISP to accept the DNS name and corresponding IP of it.
- Then the DNS resolver forwards it to the root name server to find the root of DNS like .com, .net, .org etc



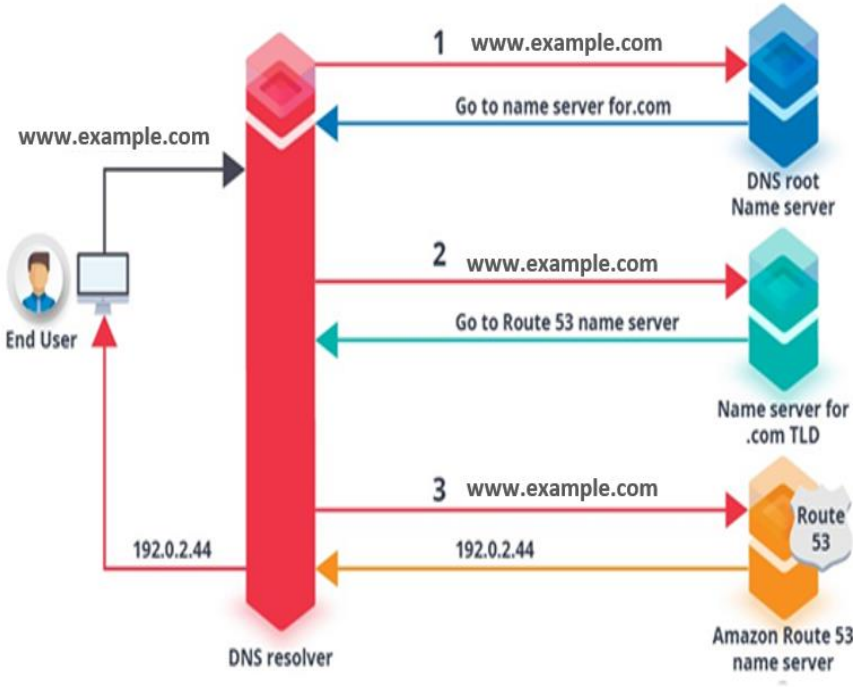
Working of Route 53

- The DNS Resolver now sends the request to one of the top level domains.
- It responds with the four Route 53 servers which are associated with it.
- DNS resolver caches the name for 2 days to reduce the latency.



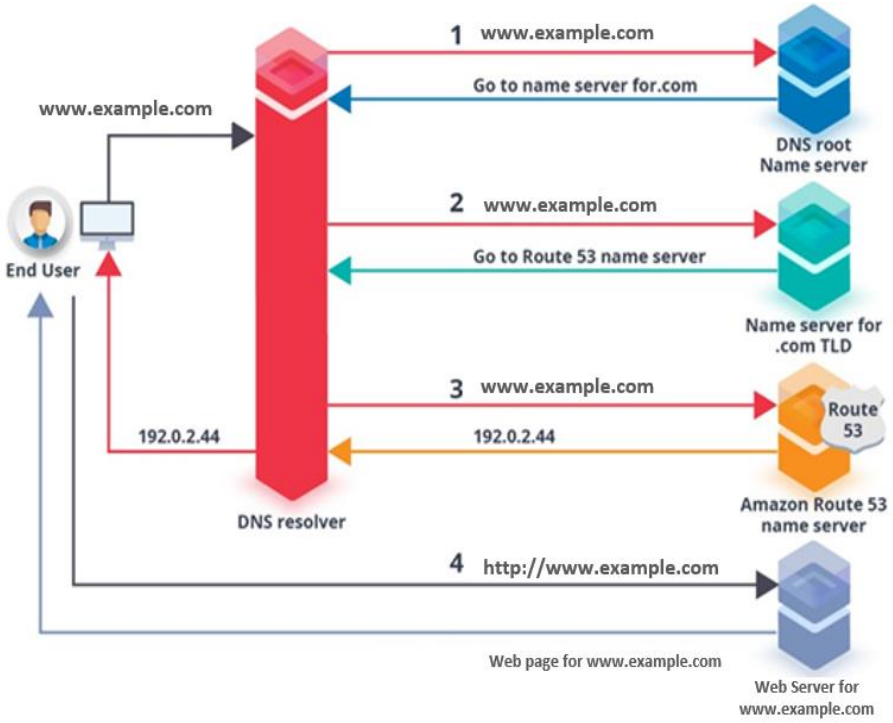
Working of Route 53

- DNS Resolver now chooses a Route 53 name server and forwards the request of the www.example.com.
- The Route 53 name server now looks into the example.com hosted zone for the www.example.com record and gets the associated IP address of it and returns it to DNS Resolver.



Working of Route 53

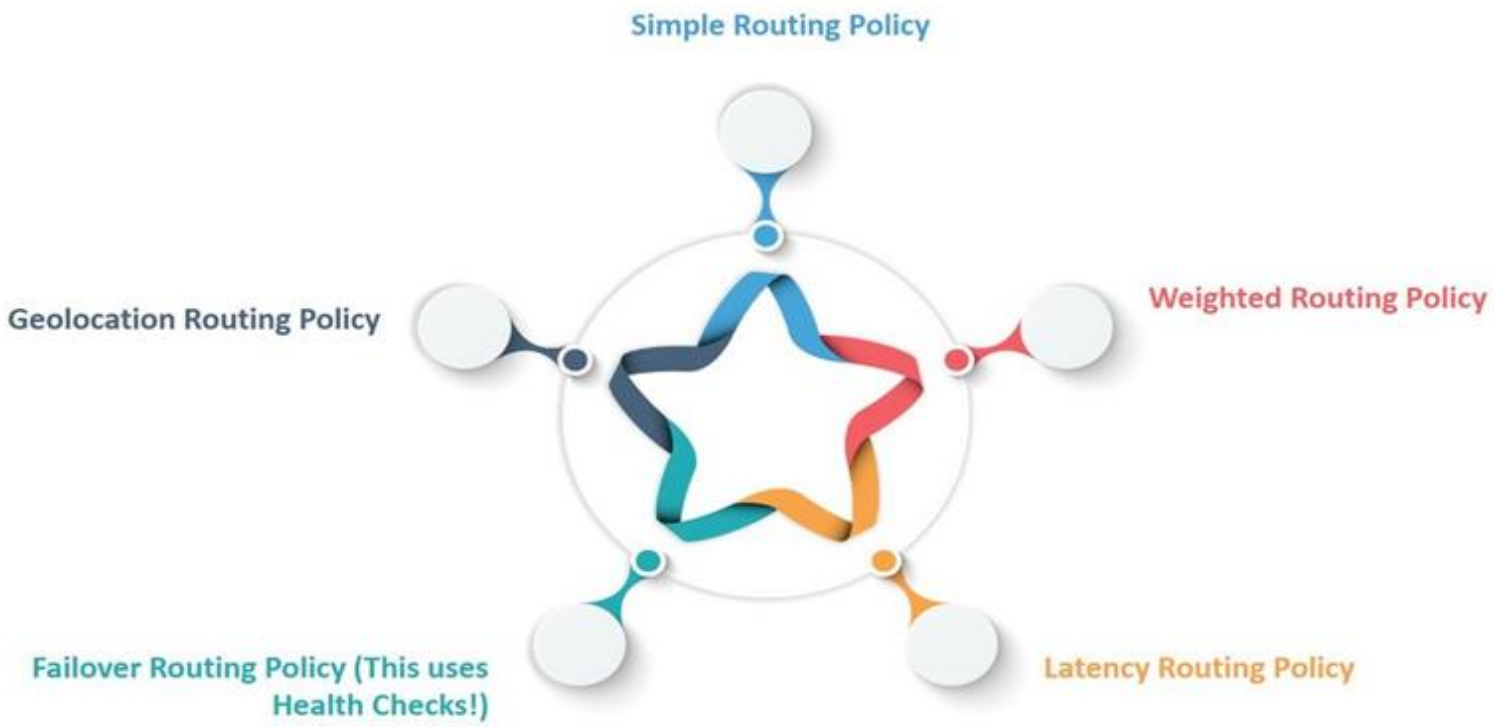
- The DNS Resolver finally sends the IP address to the browser.
- Browser now sends the request for www.example.com to the corresponding IP address that it got from the DNS Resolver.



K21Academy
Learn Cloud from Experts

Various Routing Policies

Various Routing Policies

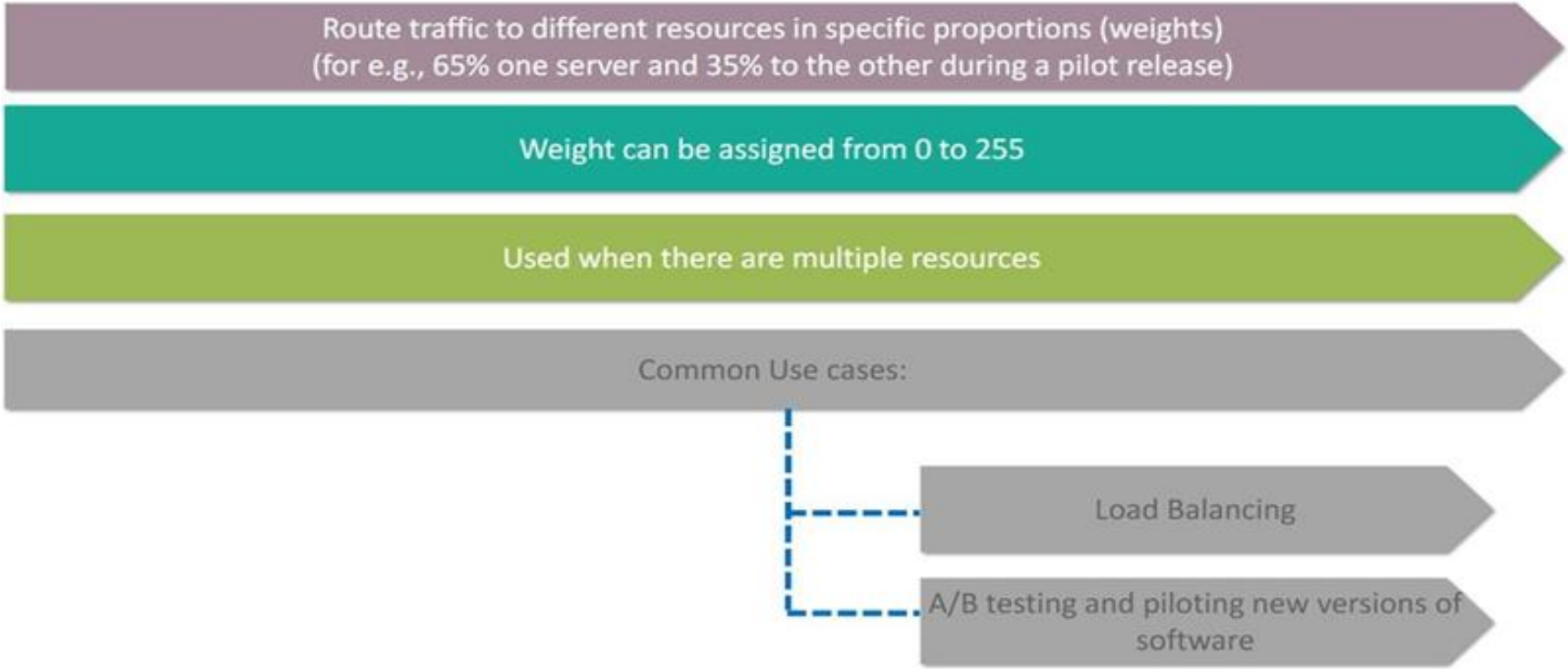


Simple Routing Policy

K21Academy
Learn Cloud from Experts

- Ideal for single resource application.
- No special routing configuration is required for simple routing policy.
- It uses Simple Round Robin policy technique.
- Responds to the DNS queries based on the values in the resource record set of Route 53.

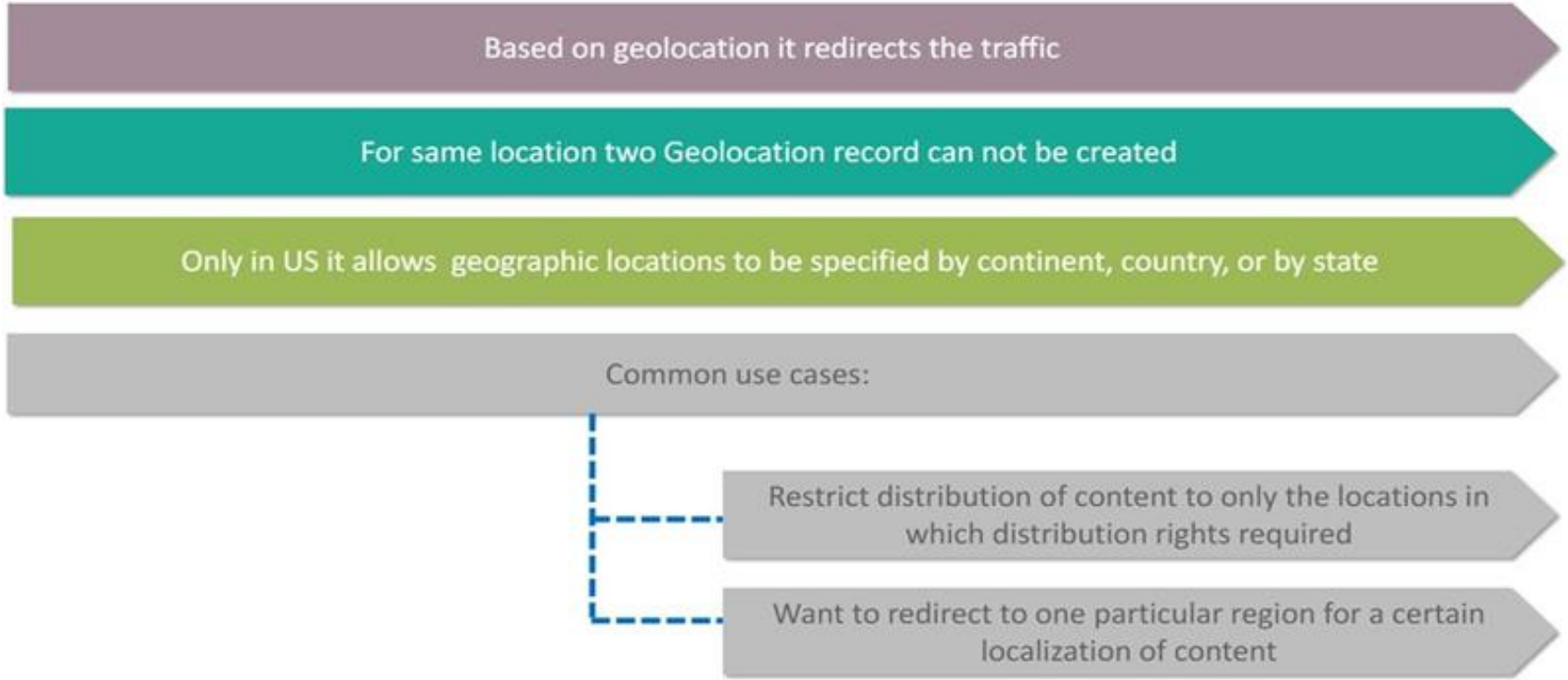
Weighted Routing Policy



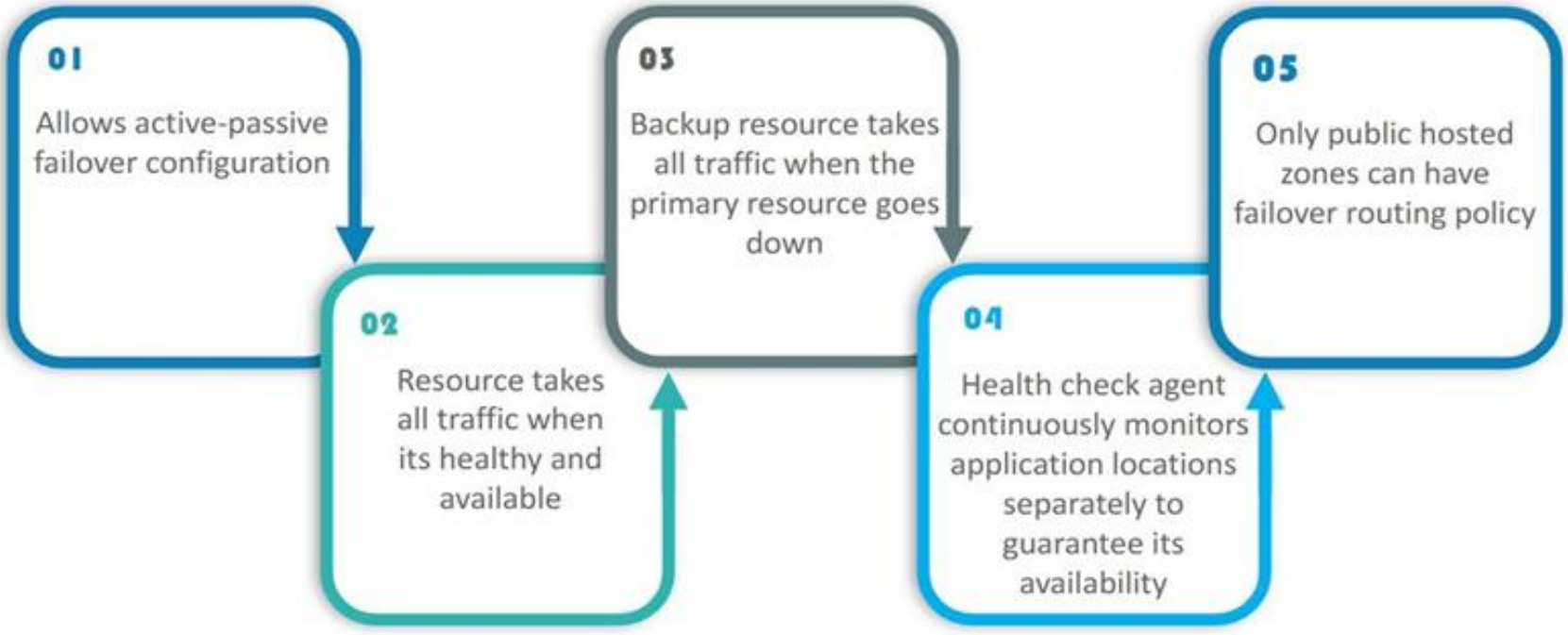
Latency Based Routing Policy

- Sends traffic to the server which has lowest network latency.
- Used when there are multiple resources performing same action.
- For the same geographical location, it does not guarantee users to be served from same location.
- Latency between servers can change over time which results change in network connectivity and routing.

Geo-location Routing Policy



Failover Routing Policy



Lab Exercises

K21Academy
Learn Cloud from Experts

Classic and Network Load Balancer

- Steps to create Classic load Balancer
- Steps to create Network Load Balancer

K21Academy
Learn Cloud from Experts

amazon web services

Contents

Lab Exercises

K21Academy
Learn Cloud from Experts

Configure a Load Balancer and Autoscaling on EC2 Instances

- Launch a Web Server in one Availability Zone
- Launch a duplicate Web Server in another Availability Zone
- Create Application Load Balancer
- Working with Autoscaling

K21Academy
Learn Cloud from Experts

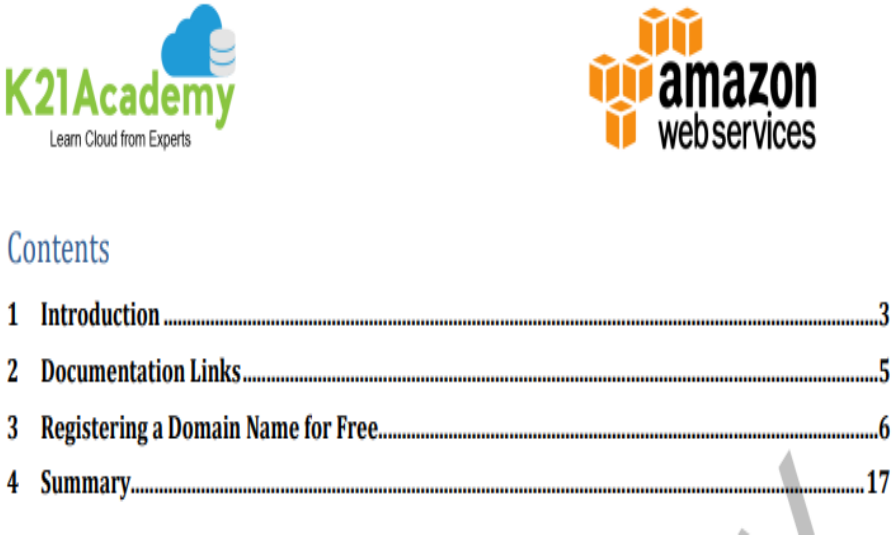
amazon web services

Contents

Lab Exercises

Register a Domain Name for Free

- Registering a Domain Name for free



Lab Exercises

Mapping DNS Using Route 53

- Mapping DNS with a Web Server Using Route 53

Contents

Quiz

K21Academy
Learn Cloud from Experts

You are helping a new DevOps Engineer to design her first architecture in AWS. She is planning to develop a highly available and fault-tolerant architecture which is composed of an Elastic Load Balancer and an Auto Scaling group of EC2 instances deployed across multiple Availability Zones. This will be used by an online accounting application which requires path-based routing, host-based routing, and bi-directional communication channels using WebSocket's.
Which is the most suitable type of Elastic Load Balancer that you should recommend for her to use?

A. Classic Load Balancer
B. Either a Classic Load Balancer or a Network Load Balancer
C. Network Load Balancer
D. Application Load Balancer

Quiz

K21Academy
Learn Cloud from Experts

Answer: D

Explanation: Application Load Balancers support path-based routing, host-based routing and support for containerized applications hence, Application Load Balancer is the correct answer.
Network Load Balancer, Classic Load Balancer, and either a Classic Load Balancer or a Network Load Balancer are all incorrect as none of these support path-based routing and host-based routing, unlike an Application Load Balancer.

Find Us


https://www.facebook.com/K21Academy


http://twitter.com/k21Academy


https://www.linkedin.com/company/k21academy

https://www.youtube.com/k21academy


https://www.instagram.com/k21academy/