AWS Foundation

Introduction to Load Balancing, Auto scaling
& Route 53

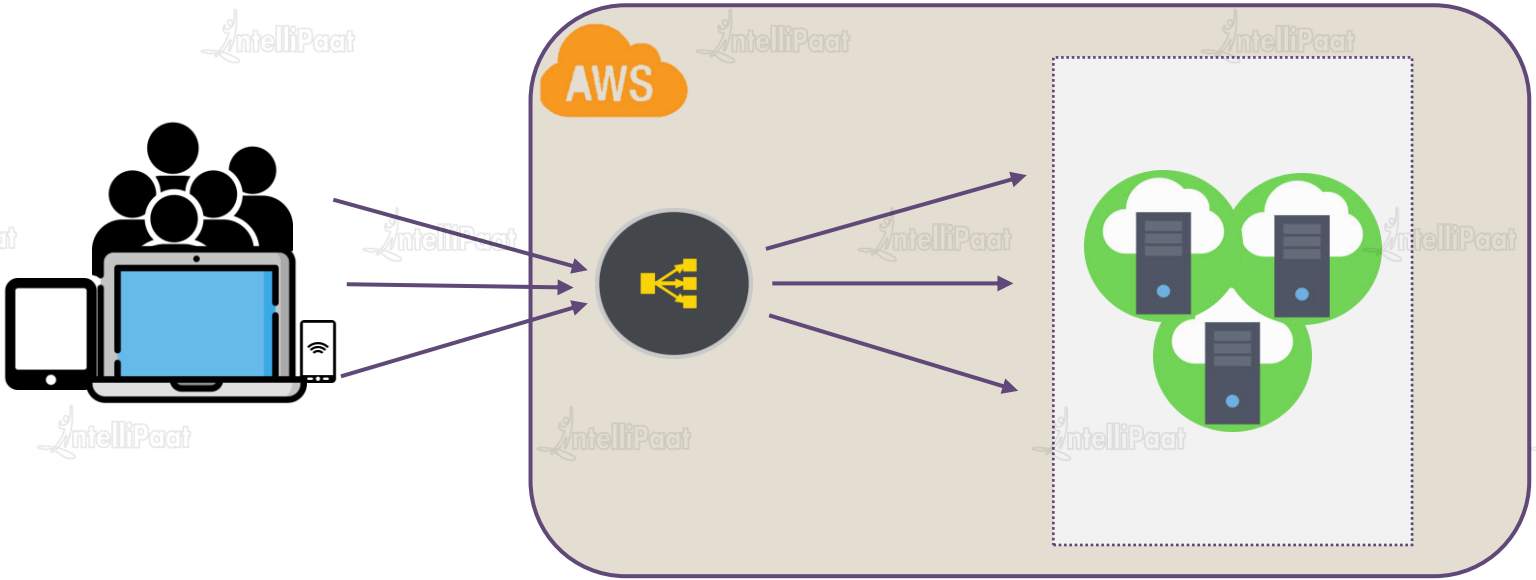
Agenda

1. Introduction to Elastic Load Balancer
2. Types of ELB - Classic, Network and
3. Load Balancer Architecture
4. Demo 1
5. Demo 2
6. Demo 3
7. Cross Zone Load Balancing
8. Introduction to Autoscaling
9. Vertical and Horizontal Scaling
10. Lifecycle of Autoscaling
11. Components of Autoscaling
9. Vertical and Horizontal Scaling
13. Instance termination
14. Using Load Balancer with Auto Scaling
15. Pre-Route53: How DNS works, A Name record, CNAME, Alias and Latency
16. Routing Policy
17. Route53 Terminologies
18. Quiz

Introduction to ELB

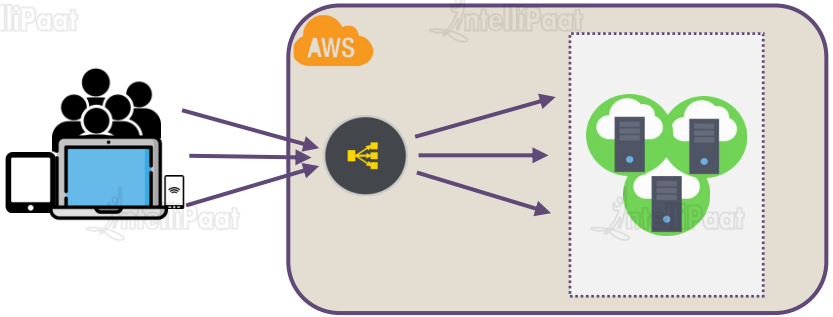
Load Balancer

Load balancer is a service which uniformly distributes network traffic and workloads across multiple servers or cluster of servers. Load balancer increases the availability and fault tolerance of an application.



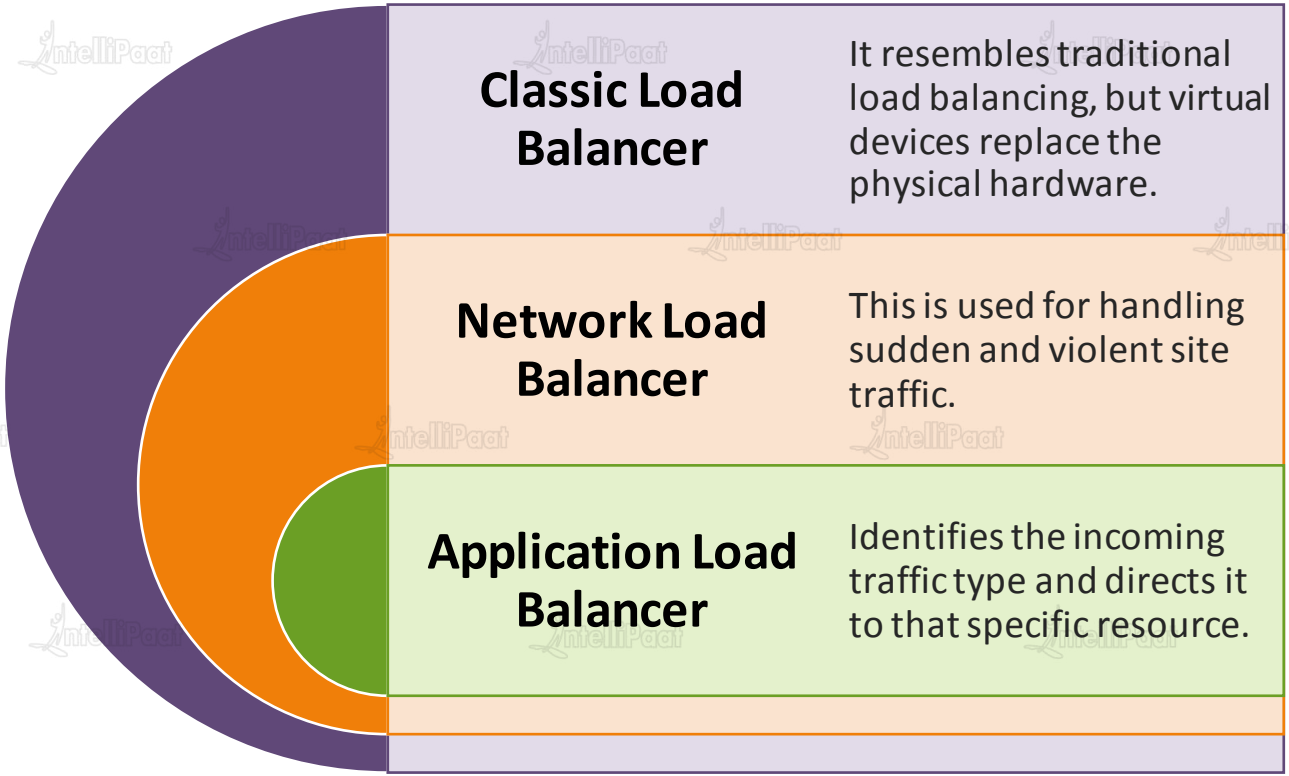
Elastic Load Balancer

- Elastic Load Balancer (ELB) is a Load balancing service for the AWS deployments.
- ELB scales the load balancer itself as necessary to handle the load.
- Incoming traffic is distributed across EC2 instances in multiple Availability Zones.
- Load balancer is the single point of contact for clients.



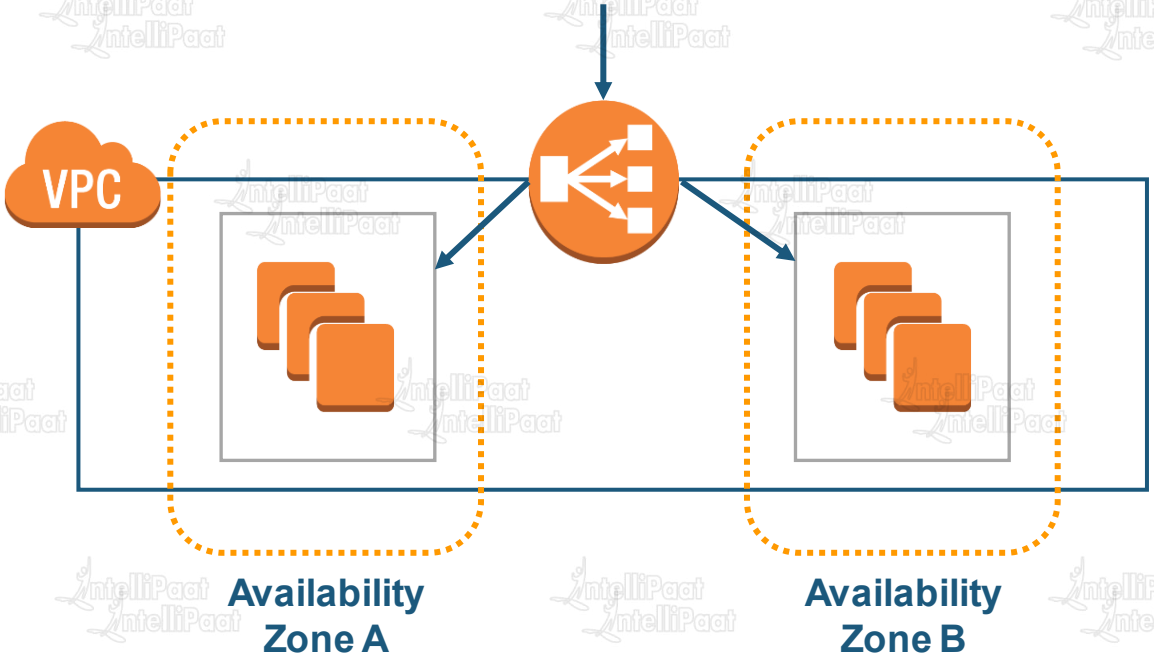
Types of Elastic Load Balancer (ELB)

Types of Elastic Load Balancer (ELB)



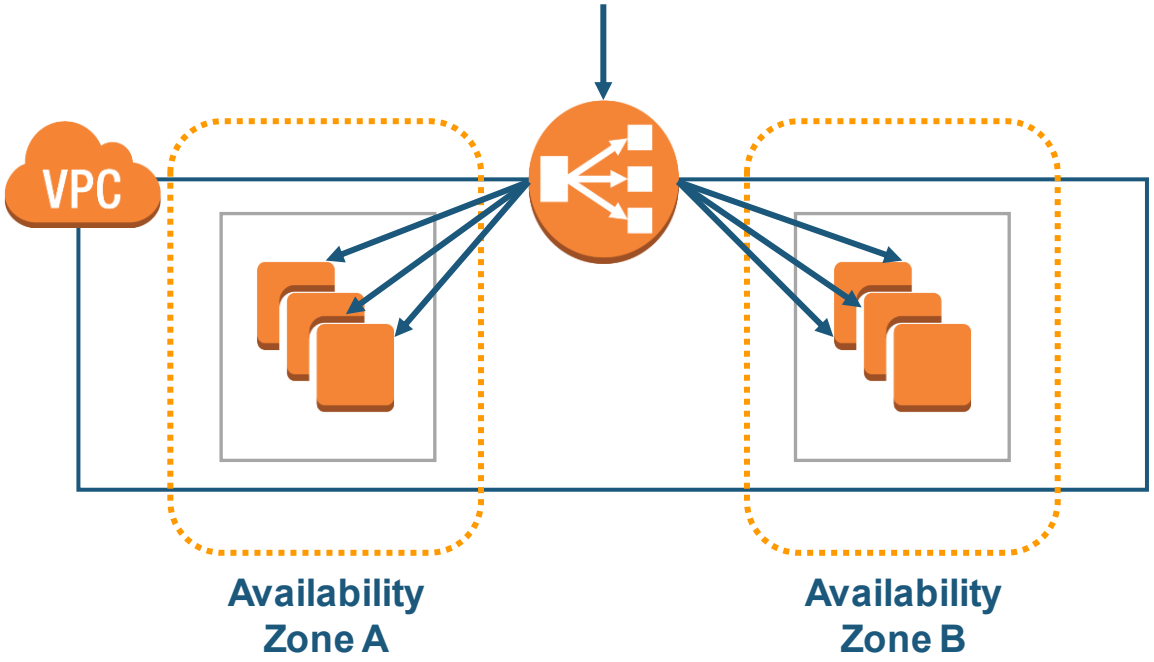
Classic Load Balancer

- Distributes incoming application traffic across ec2 instances in multiple AZs. Functions at Layer 7 (OSI Model).
- Routes traffic to healthy instances only. Evenly distributed.
- Internet and Internal Facing Load Balancer.



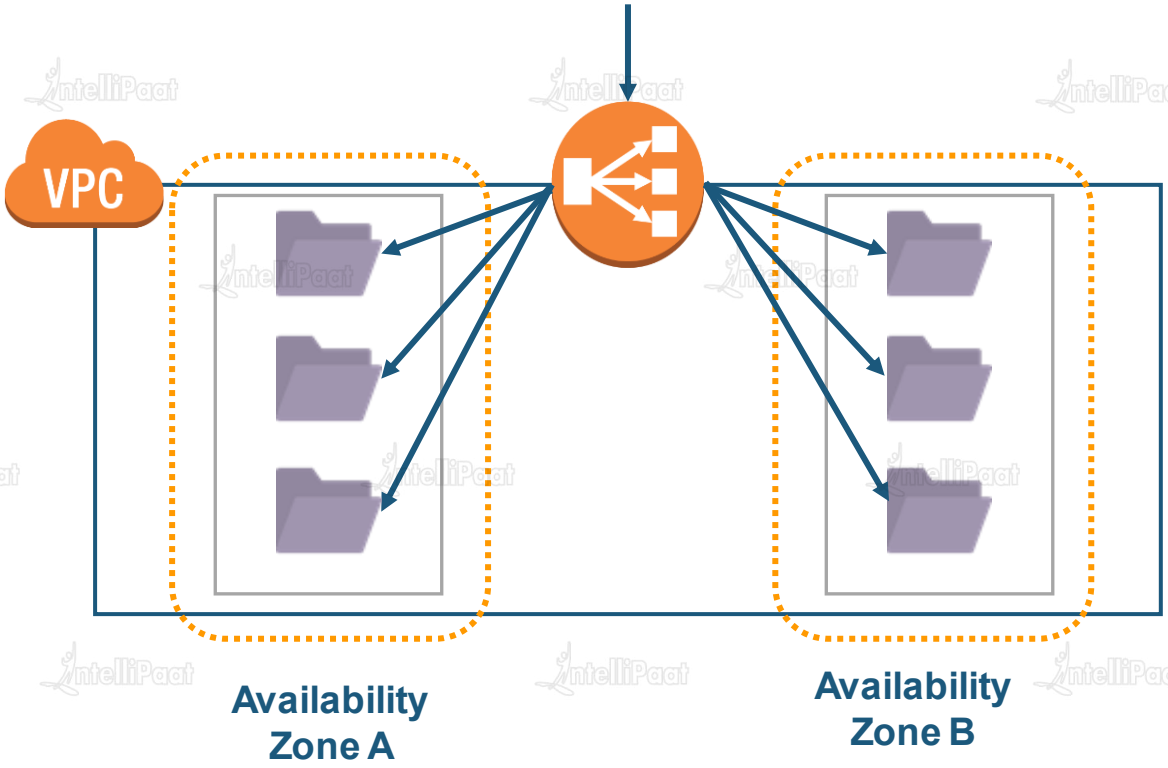
Network Load Balancer

- Network load balancer functions at the 4th layer of the OSI model
- It can handle millions of requests per second and maintain low latency
- Ideal for load balancing TCP traffic and also supports elastic or static Ip



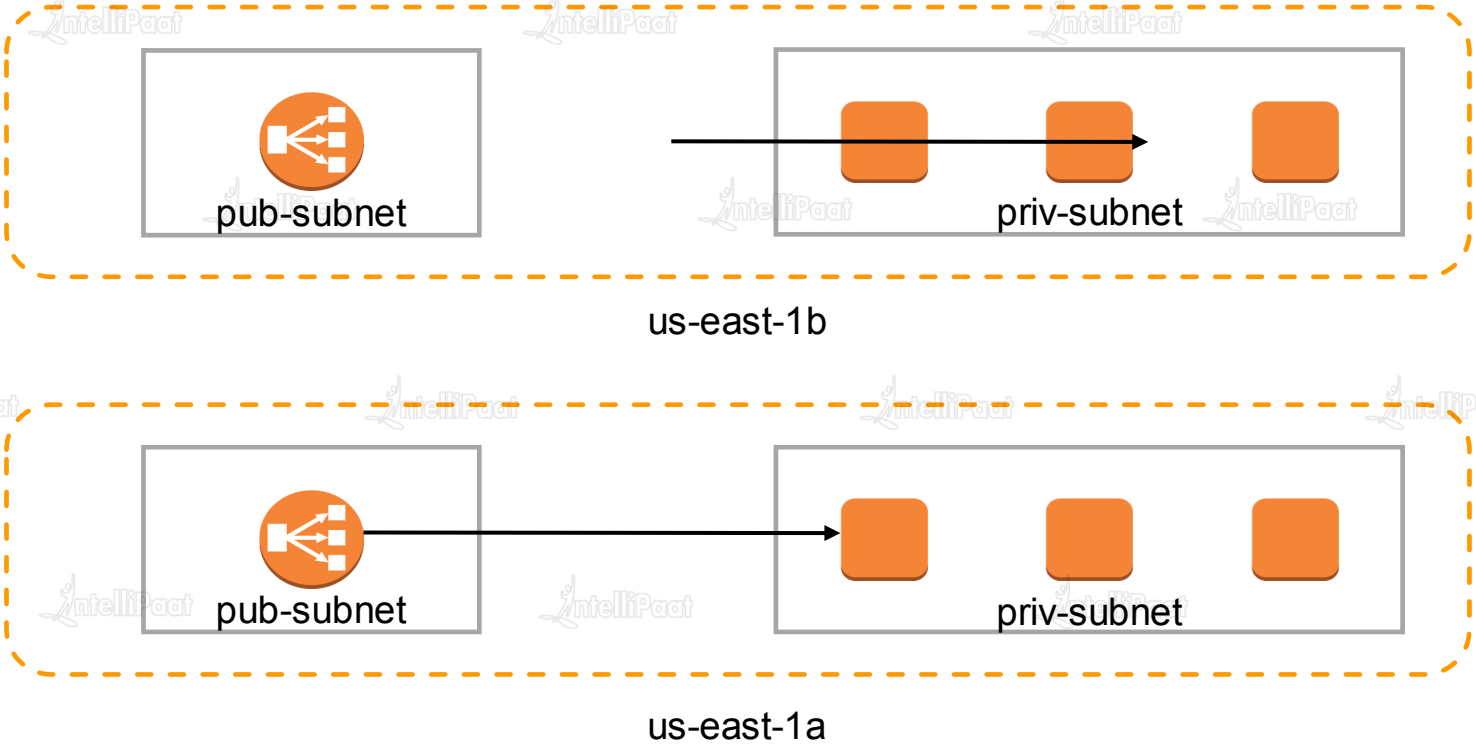
Application Load Balancer

- Application load balancer functions at the 7th layer of the OSI model
- Identifies the incoming resource type and directs it to



Load Balancer Architecture

Load Balancer Architecture



Demo 1

Demo 1: Creating Load Balancers

Classic Load Balancer Creation

1. Open AWS Management Console, click on the Services drop down box and choose EC2
2. Scroll down and choose “Load Balancers”
3. Choose create load balancer and choose Classic
4. Configure all the settings one by one – Give a name, create a new VPC, Add a tag, then choose review and launch
5. Review (optional) and choose launch
6. Classic load balancer is created

Demo 2

Demo 1: Creating Load Balancers

Network Load Balancer Creation

1. Open AWS Management Console, click on the Services drop down box and choose EC2
2. Scroll down and choose “Load Balancers”
3. Choose create load balancer and choose Network
4. Configure all the settings one by one – Give a name, create a new VPC, Add a tag, then choose review and launch
5. Review (optional) and choose launch
6. Classic load balancer is created

Demo 3

Demo 1: Creating Load Balancers

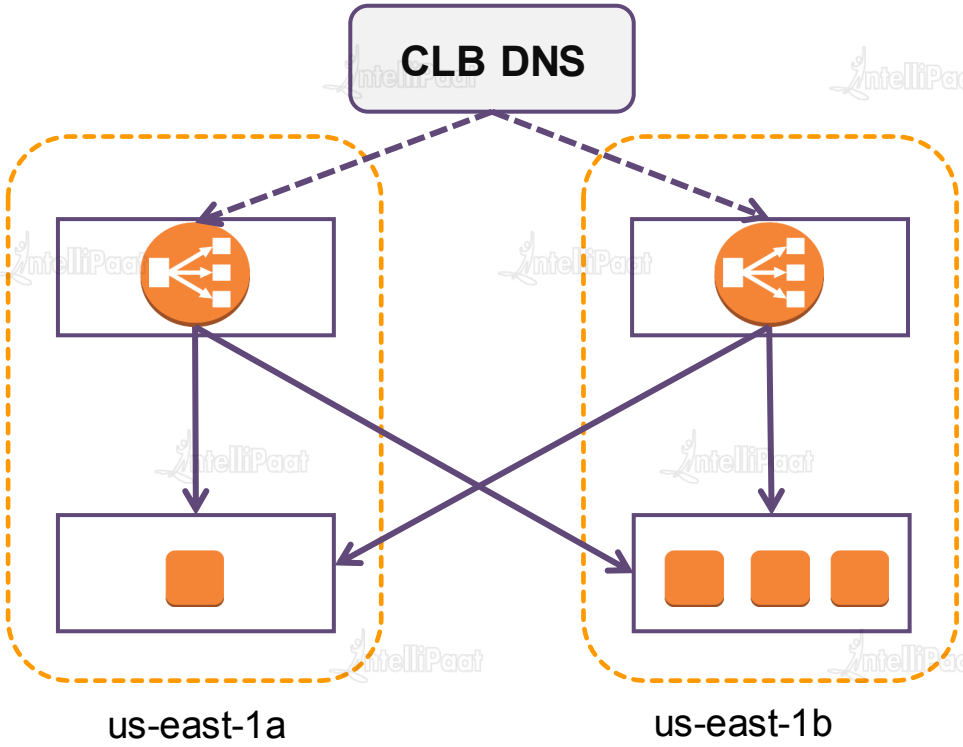
Application Load Balancer Creation

1. Open AWS Management Console, click on the Services drop down box and choose EC2
2. Scroll down and choose “Load Balancers”
3. Choose create load balancer and choose Application
4. Configure all the settings one by one – Give a name, create a new VPC, Add a tag, then choose review and launch
5. Review (optional) and choose launch
6. Classic load balancer is created

Cross Zone Load Balancing

Cross-Zone Load Balancing

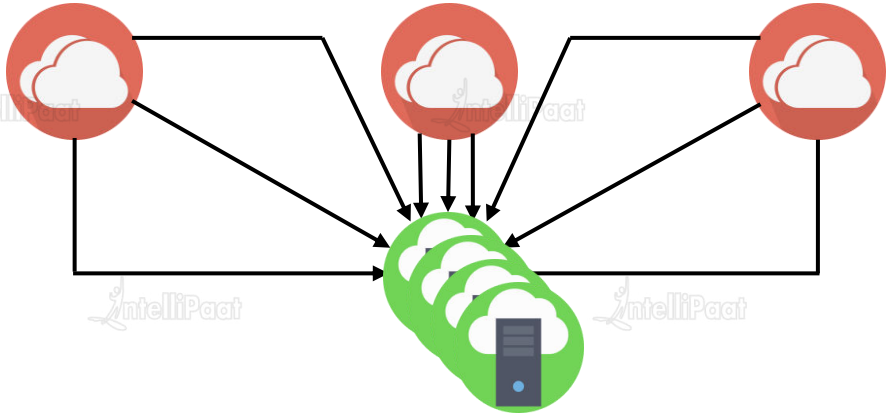
- By default CLB nodes distributes traffic to instances in its availability zone only.
- Enable cross-zone load balancing to route evenly across EC2 instances.
- Routes each request to the instance with smallest load.



Introduction to Autoscaling

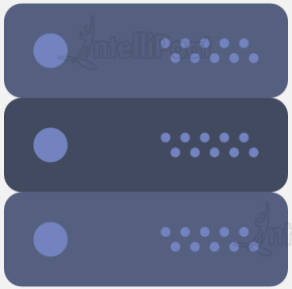
Introduction to autoscaling

- Scaling is adding or removing capacity/resource as needed.
- Scale Out is adding capacity/resources.
- Scale In is removing capacity/resources.
- Types: Vertical and Horizontal.



Introduction to autoscaling

- Scaling Types: Vertical and Horizontal
- Vertical
  - CPU: 2.0GHz to3.2 GHz
  - RAM: 1024GB to 2048GB
  - N/W Bandwidth: 4Gbps to 10Gbps



- Horizontal
  - CPU: 1 server with 1.0GHz to 3 servers with 1.0GHz
  - RAM: 1 server with 500GB to 3 servers with 500GB

Introduction to autoscaling

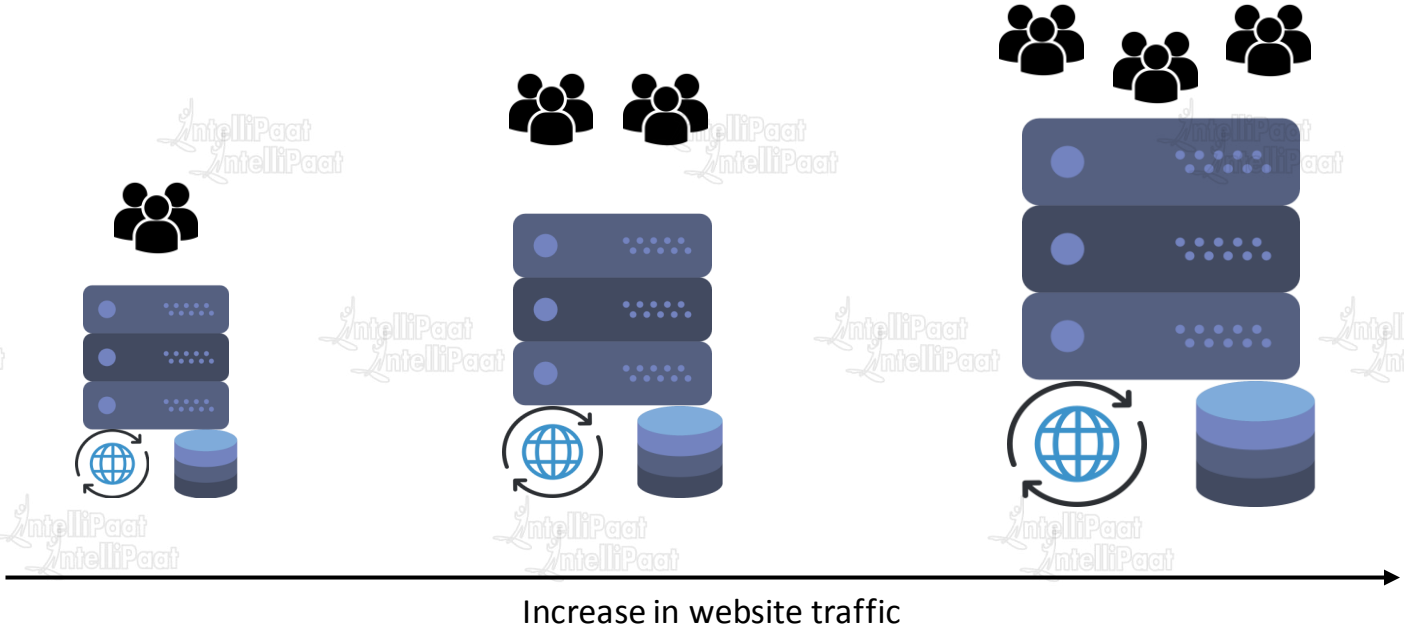
- Auto Scaling is scaling out or in automatically without any manual intervention.
- Helps to ensure correct no of ec2 instances are available to handle load.
- Multi-AZ ec2 instances provide high available solution.

- Auto Scaling can dynamically increase and decrease capacity as needed.

Vertical and Horizontal Scaling

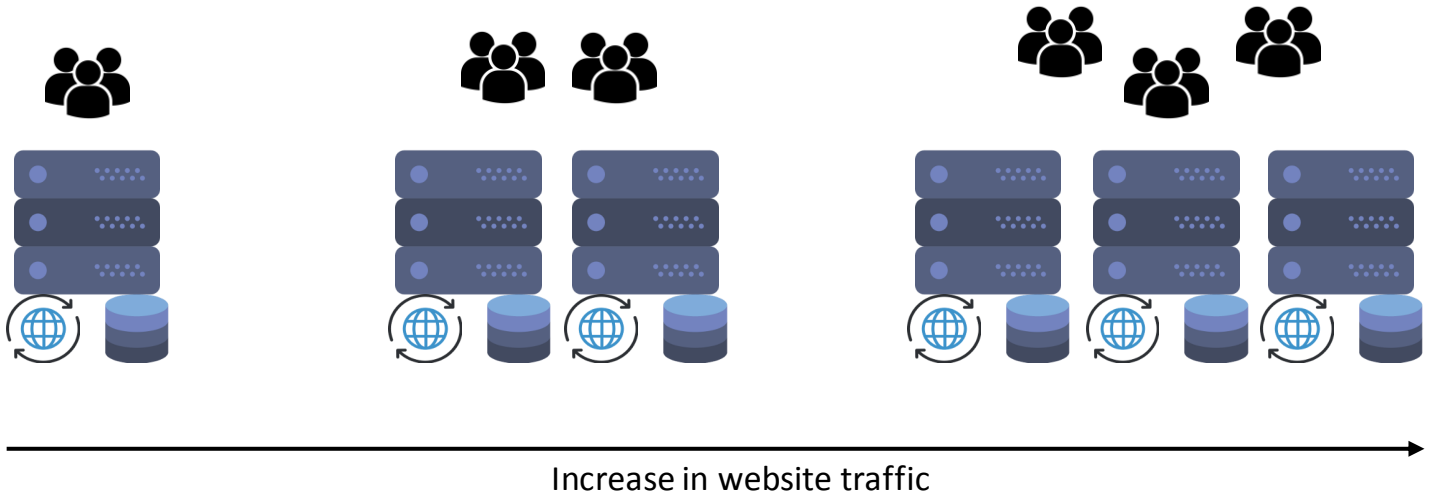
Vertical and Horizontal Scaling

Horizontal Scaling



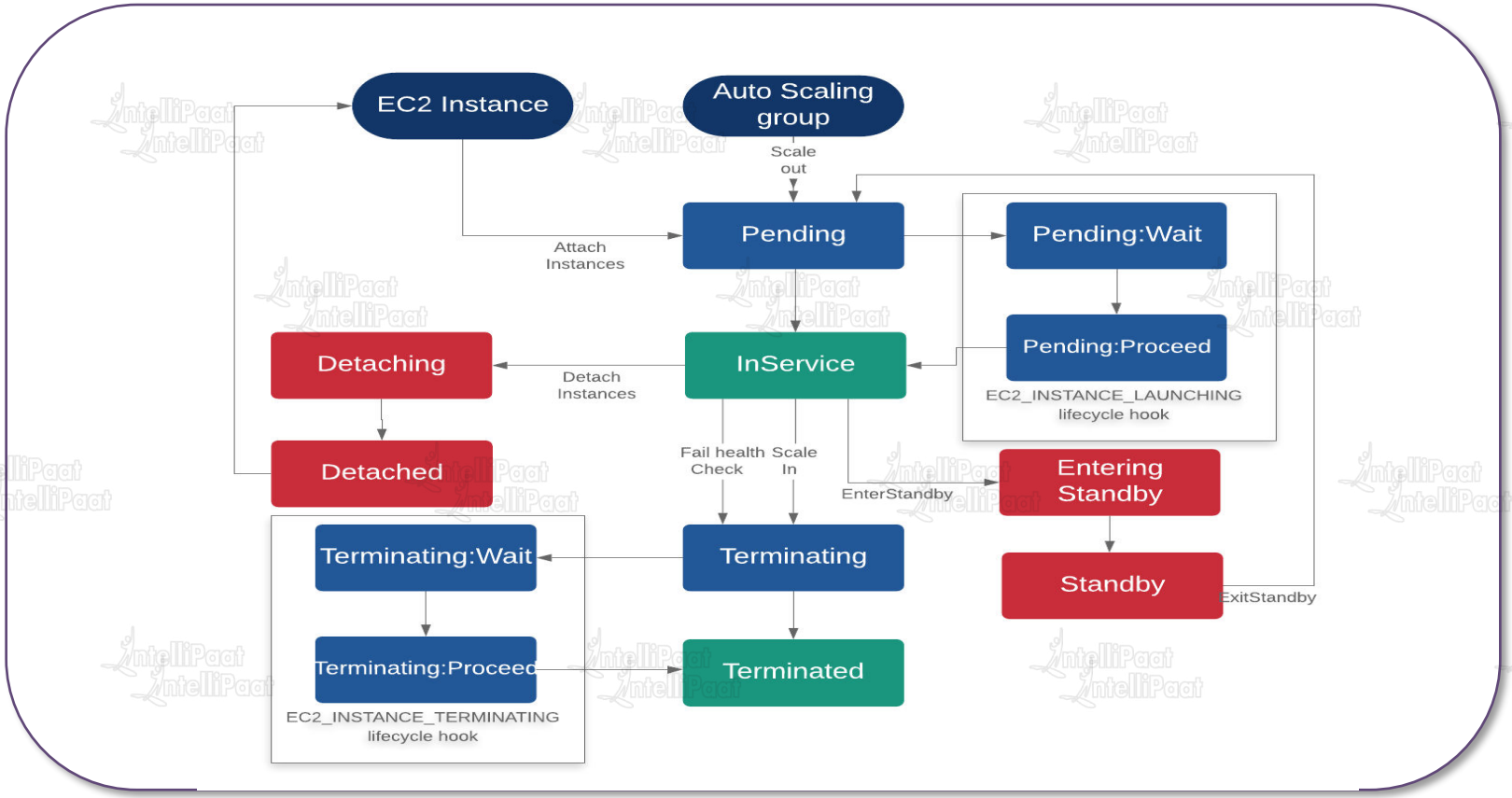
Horizontal and Vertical Scaling

Vertical Scaling



Lifecyle of Autoscaling

Lifecycle of Autoscaling



Components of Autoscaling

Autoscaling Components

Groups

- EC2 instances are in groups so that they can be considered as an logical unit (For Scaling and Management)
- When you create a group, you can mention these attributes – Max, Min and desired number of instances

Configuration Templates

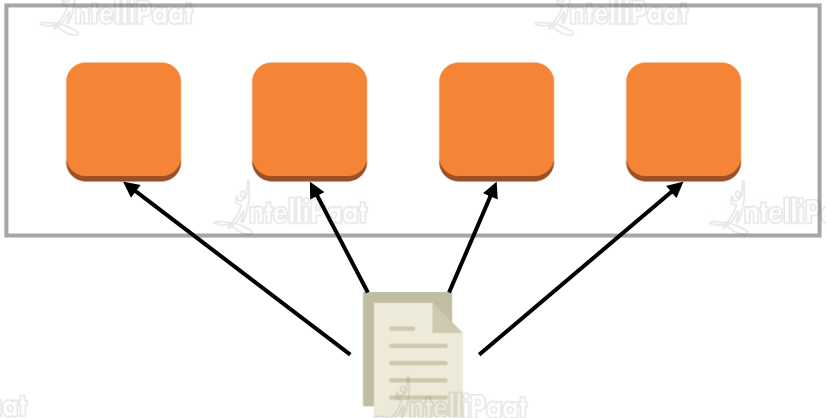
- These are used as configuration templates for the EC2 Instances.
- Launch template or Launch configuration is also used.

Scaling options

- Autoscaling provides several ways to scale the group
- Manual Scaling
- Dynamic Scaling
- Scale based on demand or schedule

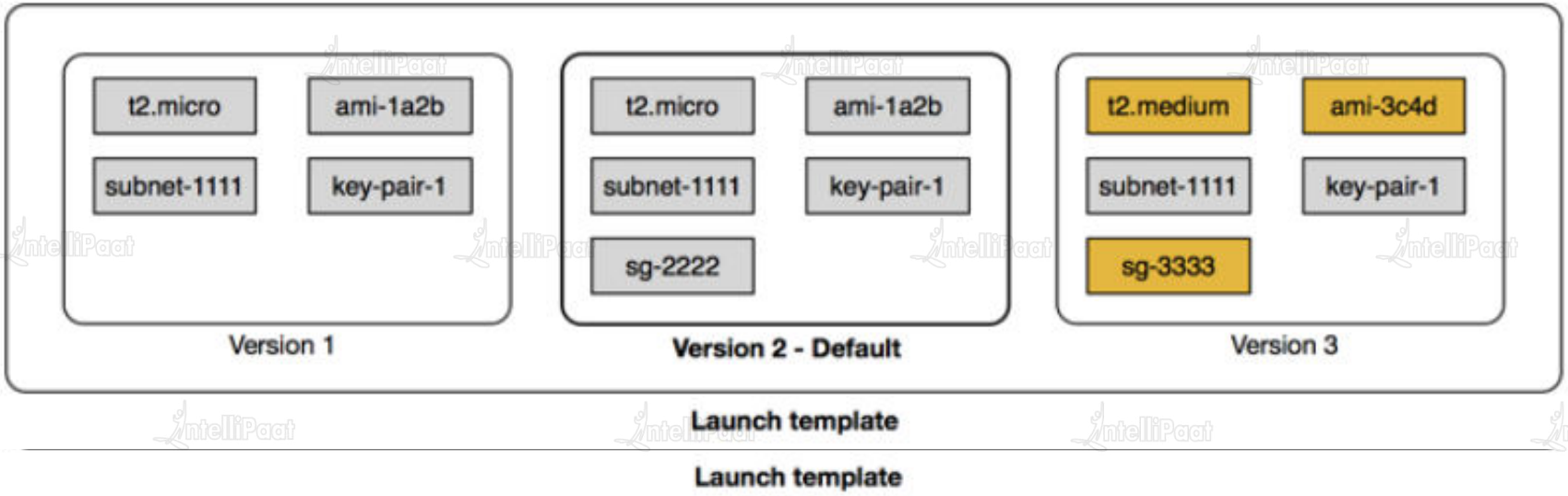
Autoscaling Groups

- Auto Scaling group contains a collection of EC2 instances that are exactly same.
- While creating an Auto Scaling group, launch configuration must be specified.
- After specifying, the launch configurations cannot be changed.
- New instances are launched using new configuration.
- EC2 instances are launched and terminated using scaling policies.



Configuration Templates

Launch template can also be used with auto scaling groups.



Configuration Templates

- **Launch configuration** is a template that is used to launch EC2 instances for Auto Scaling purpose.
- Auto scaling groups (next topic) uses launch configuration to launch instances.
- Launch Configuration cannot be modified after creation.
- Can be created in two ways
  - From scratch – Image ID, instance type, storage devices etc.
  - From an EC2 instance – Attributes from the instance are copied over. Block device mapping of the AMI is included, any additional devices which were attached after launching the instances are not considered in the launch configuration.

Scaling Options (Dynamic Scaling)

- Scaling policies and Alarms.
  - Scaling policies mention how to scale, and alarms decide when to scale.
  - CloudWatch alarms are set to monitor individual metrics, e.g. CPU Utilization etc.
  - When the threshold is breached, scaling policies are executed.
- Minimum, Maximum and Desired capacity.

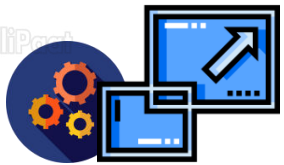
Scaling Policy:

- INCREASE 2 instances at a time
- DECREASE 1 instance at a time

Alarm:

IF CPU Utilization > 80% for more than 10 mins, ring the bell

Minimum Capacity – 2
Desired Capacity – 4
Maximum Capacity - 10



Other Scaling Options

- Scale based on a schedule – This type of a scaling method is used to scale at a given time and date.
- Scale based on a demand – Scaling occurs when the CPU utilization of the current running instances grow beyond a fixed usage limit.

Scale based on schedule:

- INCREASE the instances by 2 at 2:30 pm Today
- DECREASE the instances by 1 at 12:00 am Tomorrow

Scale based on demand:

IF CPU Utilization > 80% for more than 10 mins, INCREASE the instance by 1
IF CPU Utilization < 50% for more than 5 mins, DECREASE the instances by 2

Scaling Policy

Scaling Policy

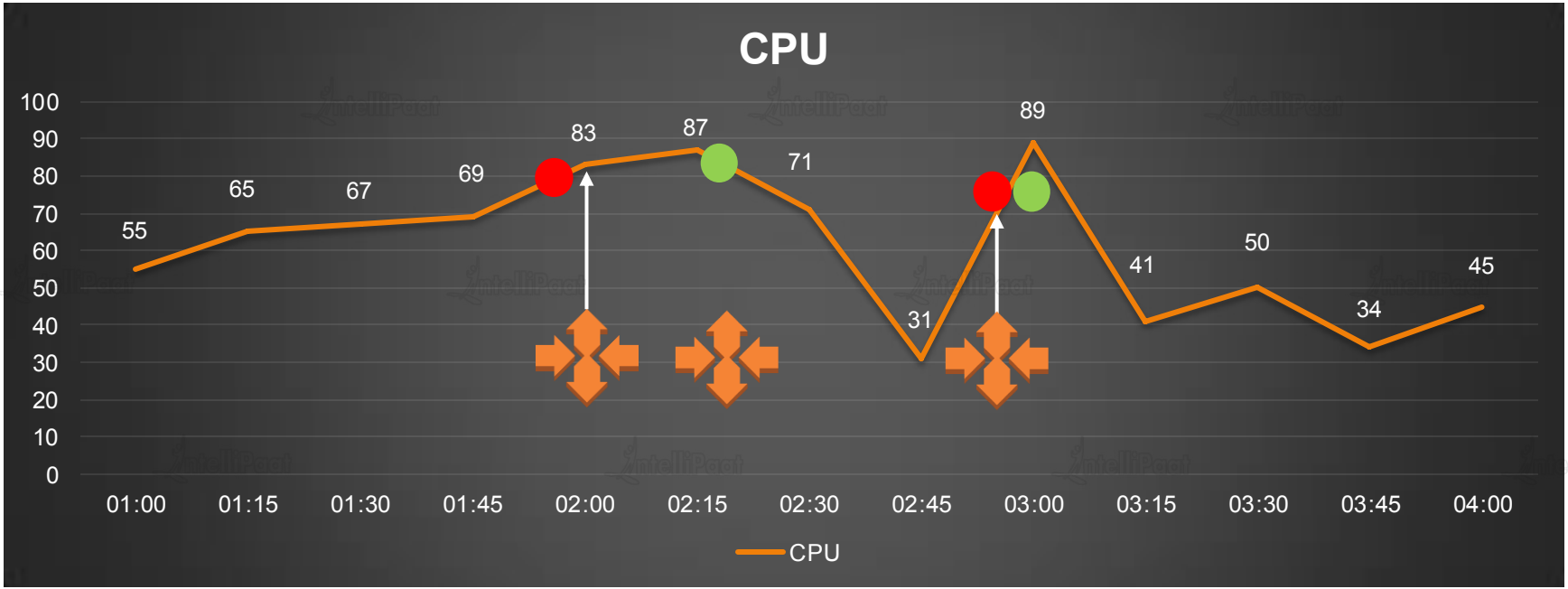
Scaling Policy:

- INCREASE 2 instances at a time
- DECREASE 1 instance at a time

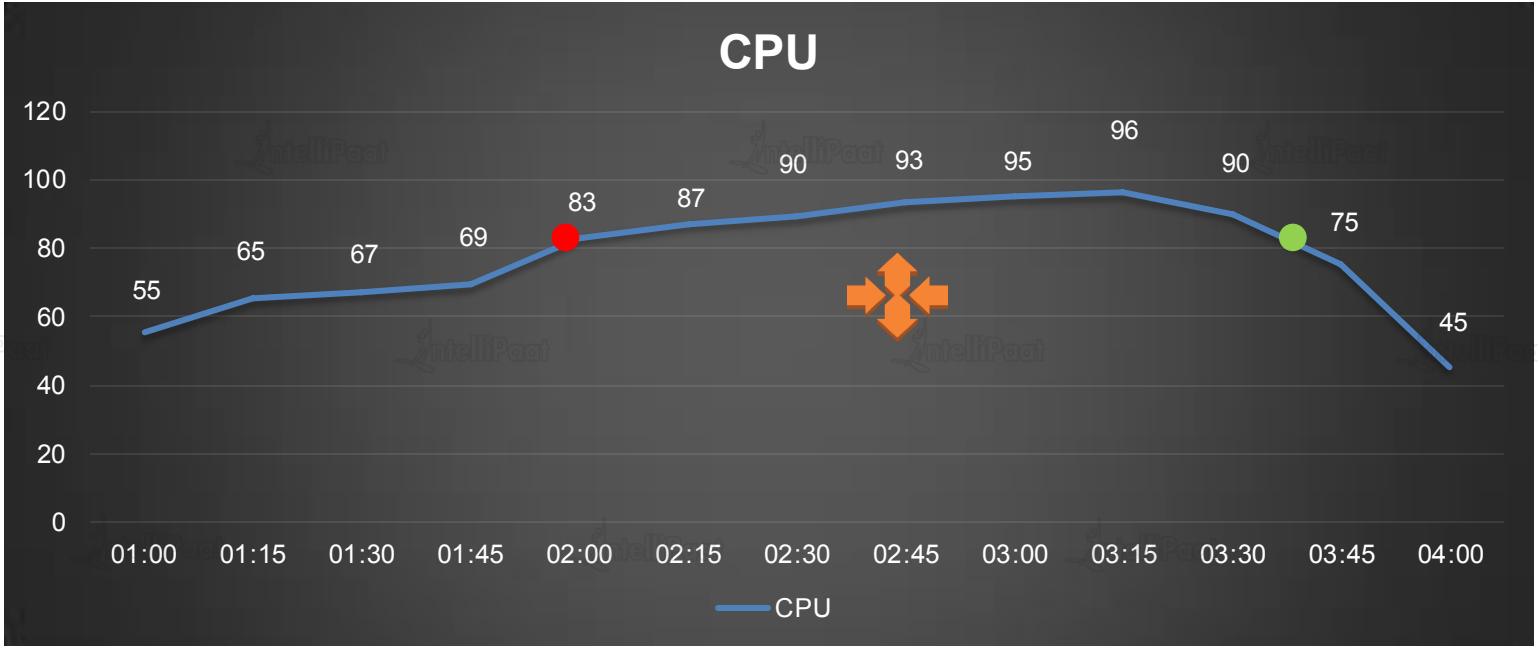
Alarm:

IF CPU Utilization > 80% for more than 10 mins, ring the bell

Minimum Capacity – 2
Desired Capacity – 4
Maximum Capacity - 10



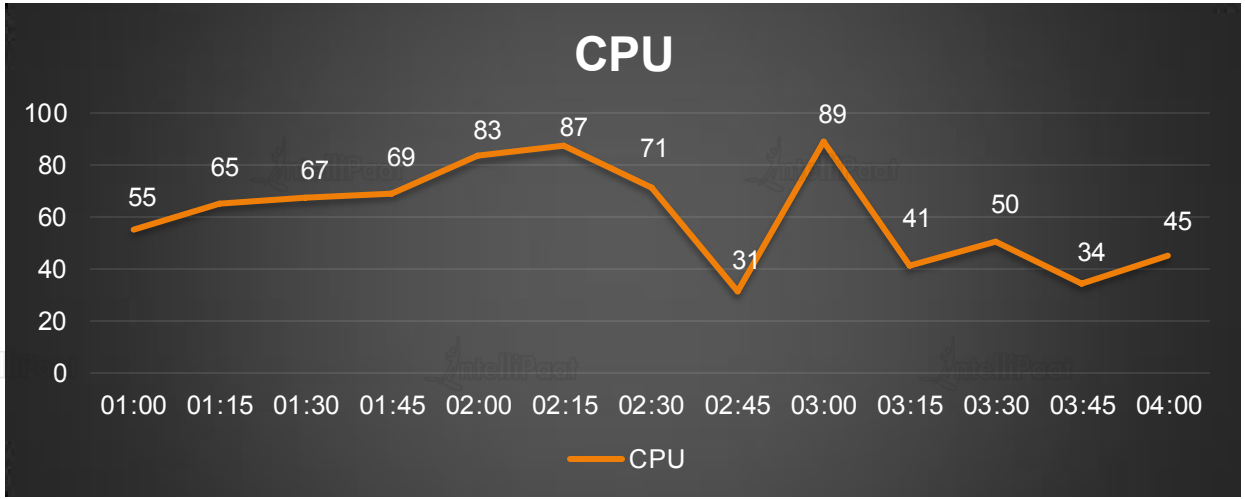
Scaling Policy



Scaling Policy

Cool-down period - Simple Scaling Policy

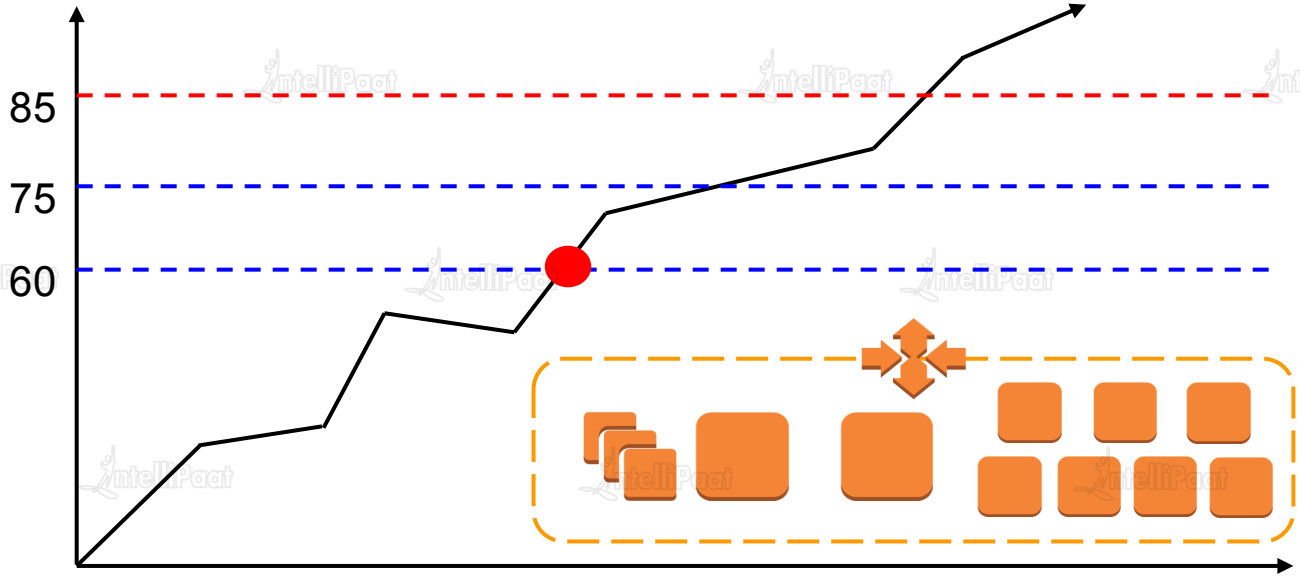
Cool-down period – Ensures that Auto Scaling does not launch or terminate any more instances until a specified time period is completed. Scaling activity is suspended until cool-down period is in effect.



Scaling Policy

Step Scaling

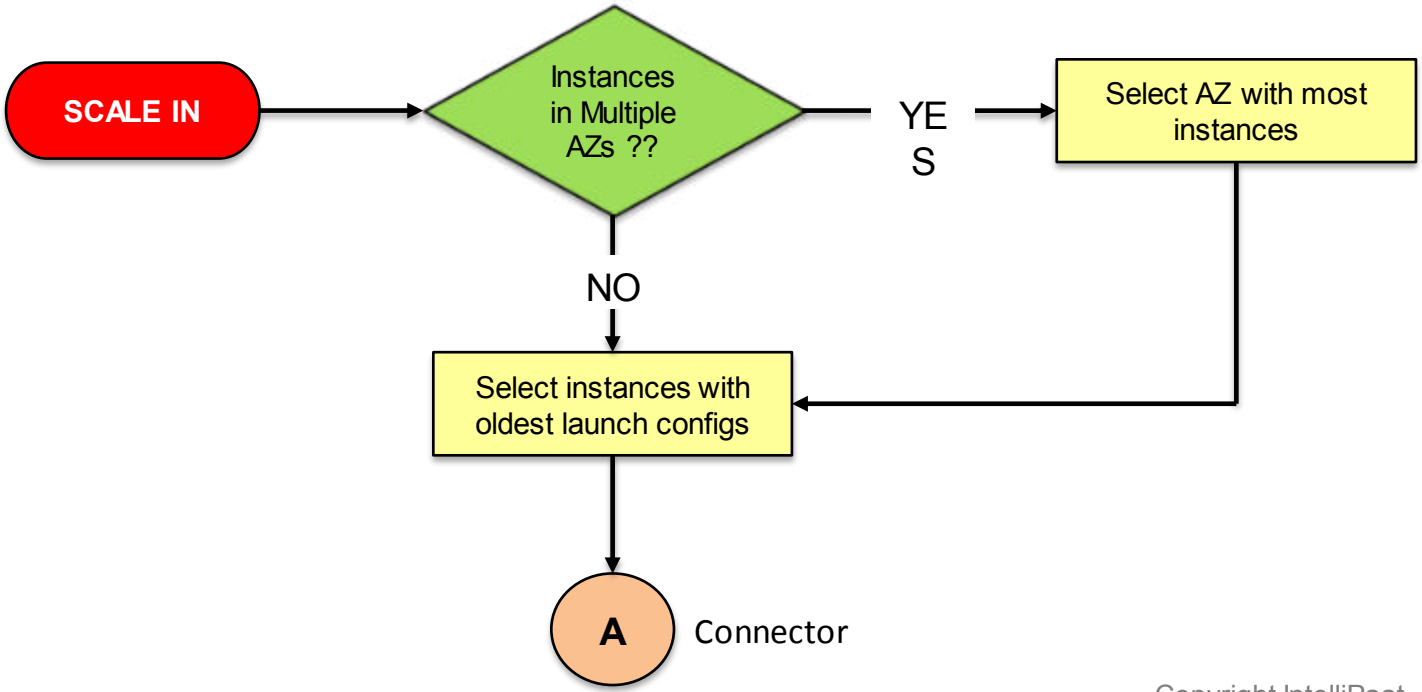
Alarm: CPU > 60%. Action: Add 2 Instances.



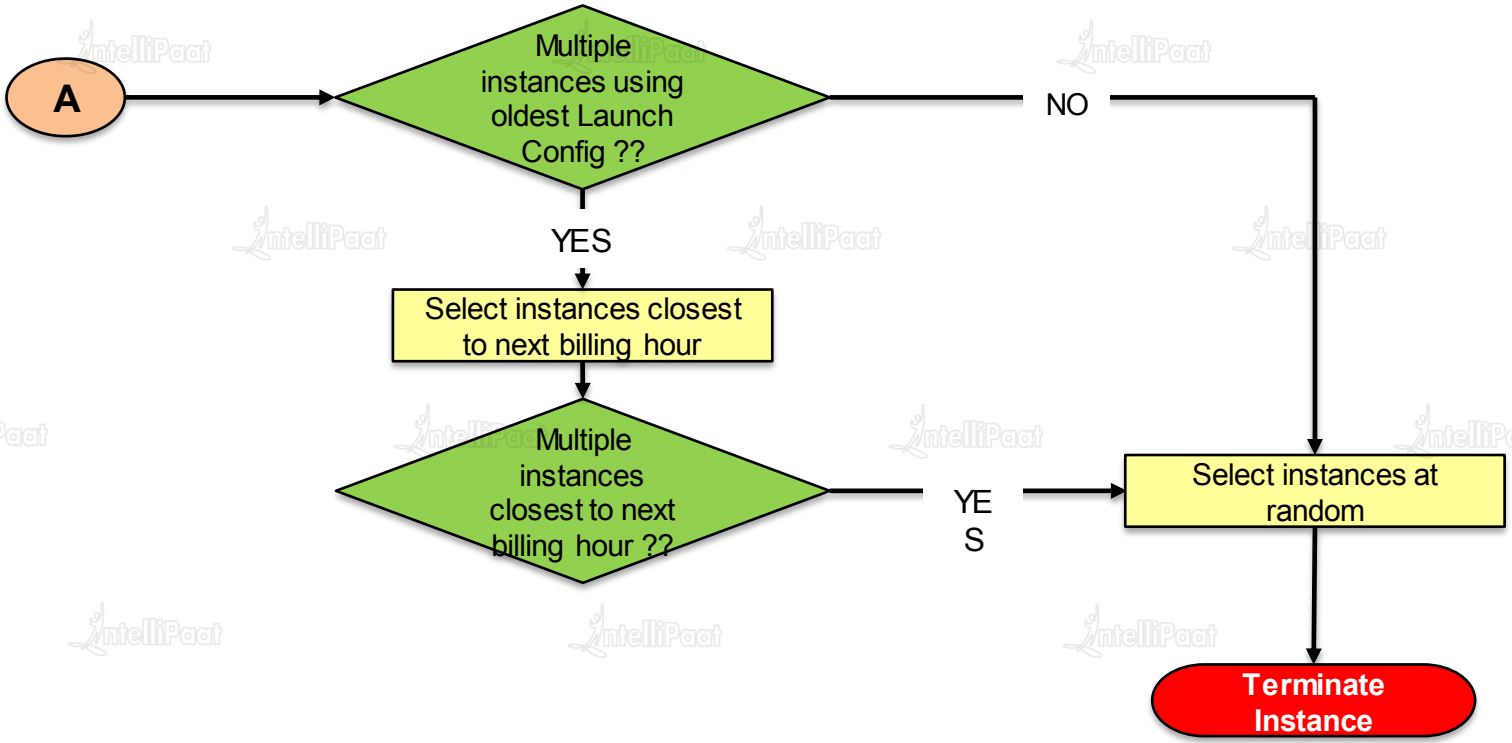
CPU	Add
> 60%	2
> 75%	3
> 85%	4

Instance Termination

Instance Termination



Instance Termination



Instance Termination

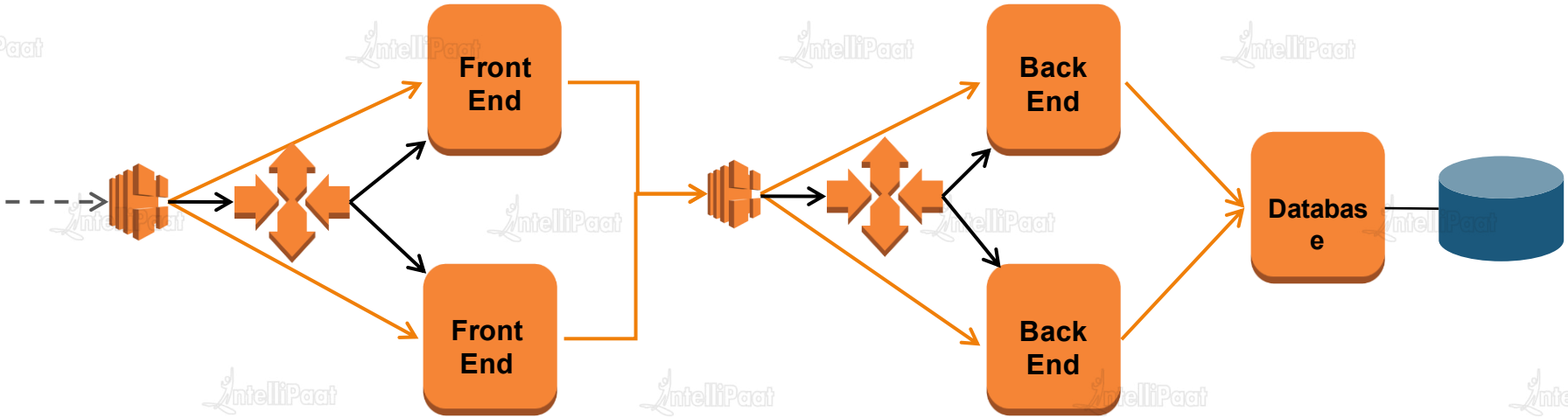
- Termination Policies (Other than DEFAULT)
  - Oldest Instance
  - Newest Instance
  - Oldest Launch Configuration
  - Closest To Next Instance Hour
- Instance protection does not terminate an instance during a scale in event. Can be enabled at Auto Scaling group or individual instance level.

Autoscaling Pricing

- No Additional fees.
- Underlying instances are charged hourly.
- Visit https://aws.amazon.com/autoscaling/pricing/ for details.

Autoscaling Design Patterns

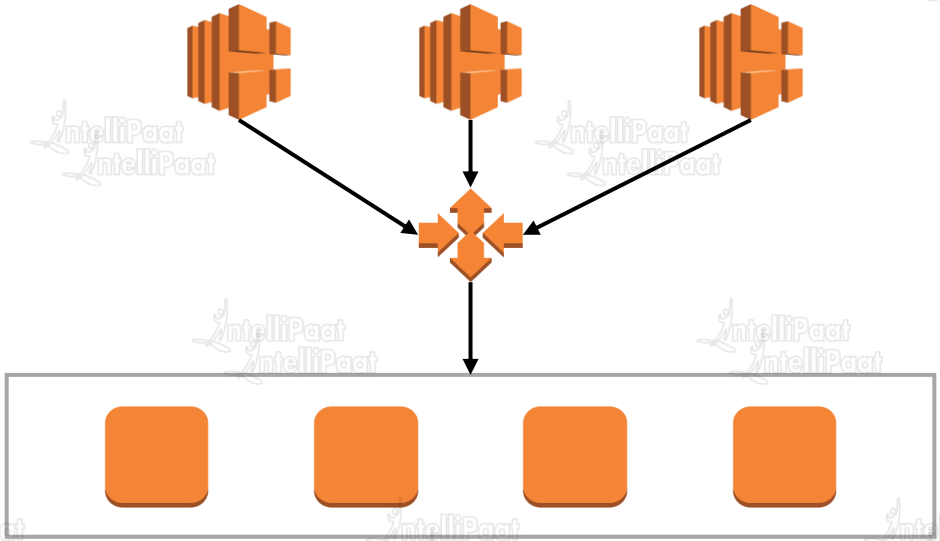
Design Patterns



ELB & AS Integration

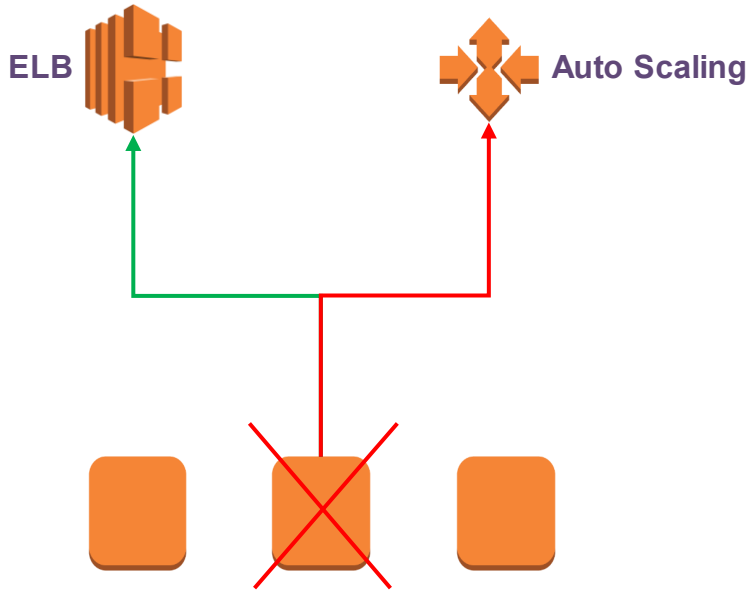
ELB & AS Integration

- **Auto Scaling:** Adds and removes capacity as per requirement.
- **Load Balancer:** Distributes incoming traffic evenly across all EC2 instances.
- Putting ELB in front of AS makes sure that all incoming traffic are distributes across dynamically changing number of EC2 instances.
- ELB is the point of contact between clients and backend ec2 instances.



ELB & AS Integration

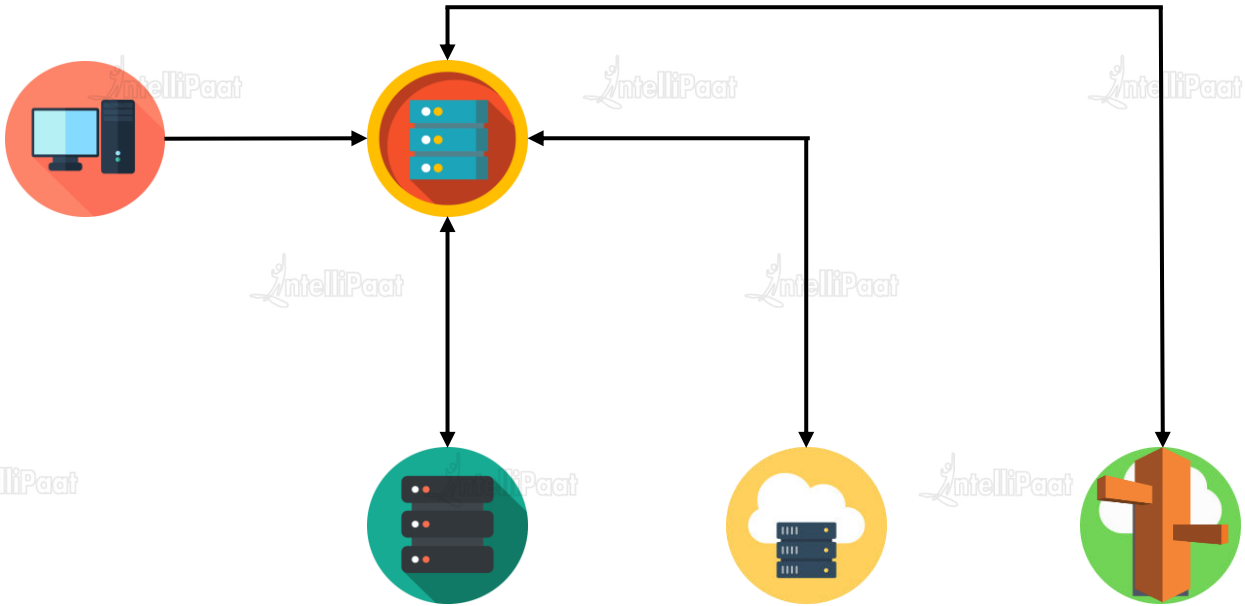
- Load balancer automatically registers instances in the group.
- Health Checks
  - EC2 instance only – EC2 status checks are considered.
  - EC2 and ELB health checks – An instance is considered unhealthy if either of the health checks fail.



Pre-Route53

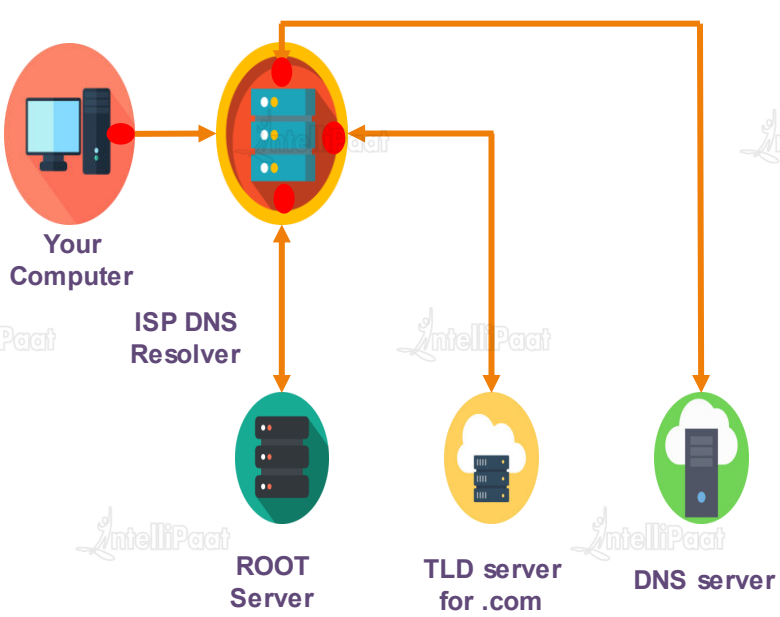
What is Route53?

Route53 is highly available and scalable Domain Name System provided by AWS.



Domain Name System

- www.amazon.com –
- “com” – Top Level Domain Name.
- “amazon” – Domain Name.
- Domain Name System is an internet service that translates Domain Names into IP Addresses.

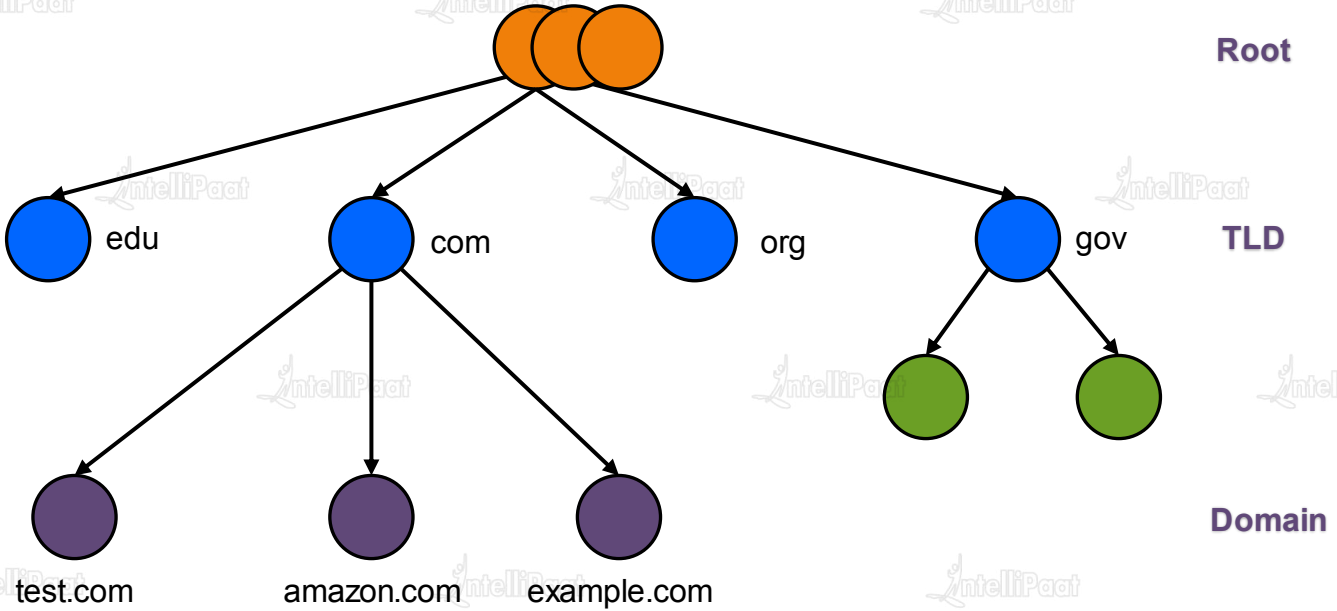


ROOT servers keep information about TLD servers.

TLD servers keep information about authoritative Name Servers.

Name Servers contain information about IP addresses for individual domains.

DNS Hierarchy

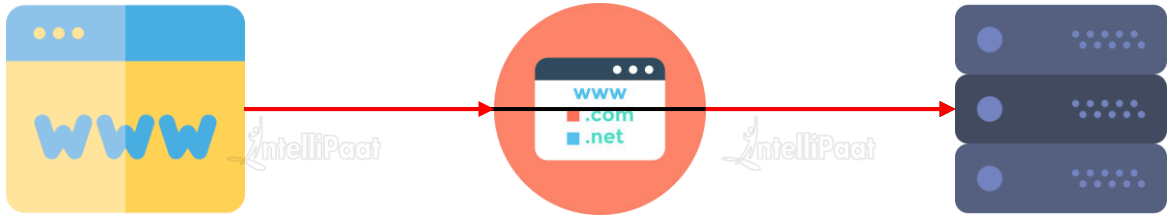


Hosting your Website

Step 1 – Start up a server/host where web service will run (Say IP Address of the server is 10.20.30.40).

Step 2 – Get a domain name from domain name providers like GoDaddy, freenom etc.

Step 3 – Link Domain name with IP address from Step 1 using Domain Name Service/System.



www.example.com	10.20.30.40

DNS Literature

- Authoritative Name Server – Server component in Domain Name System (DNS) which holds actual DNS records like A Name, CNAME, Alias etc.
- "A" NAME Record – Maps Domain Name to IP Address of the backend host. "A" is for Address. A NAME record format is mentioned below:

Type	Domain/Host Name	Address	TTL
A	www.abc.com	101.202.30.40	60
A	www.apple-orange.com	54.28.14.6	300
AAAA	www.example.com	fe80::1cb2:373a:3dd1:8f46	600

DNS Literature

- CNAME (Canonical Name) Record – Maps one name to another name instead of an IP address.

Type	Domain/Host Name	Address	TTL
CNAME	www.fruits.com	www.apple-orange.com	300
CNAME	www.vegetables.com	www.fruits.com	600
A	www.apple-orange.com	54.28.14.6	900

- Alias Name is similar to CName record with a “little” difference.

www.vegetables.com

www.fruits.com

DNS Literature

Type	Domain/Host Name	Address	TTL
CNAME	www.fruits.com	www.apple-orange.com	300
CNAME	www.vegetables.com	www.fruits.com	600
A	www.apple-orange.com	54.28.14.6	900

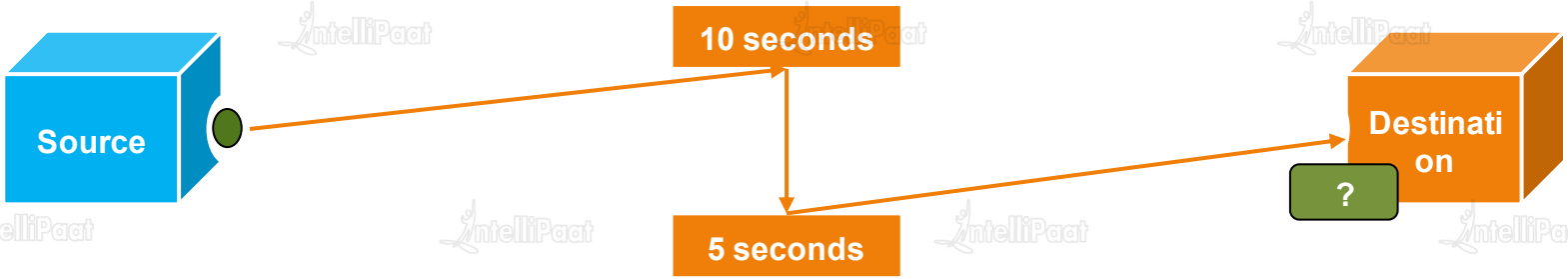
www.fruits.com

www.apple-orange.com

54.28.14.6

Network Latency and Bandwidth

Network Latency is the amount of time taken to deliver some amount of data over n/w.



Latency = 10 + 5 = 15 seconds

Routing Policy

Routing Policy

- **Public Hosted Zone** contains information about how traffic on the Internet should be routed for a Domain.
- NS record set – Authoritative Name Servers for Domain Name.
- SOA (Start of Authority) record set – Contains base DNS information about the Domain.

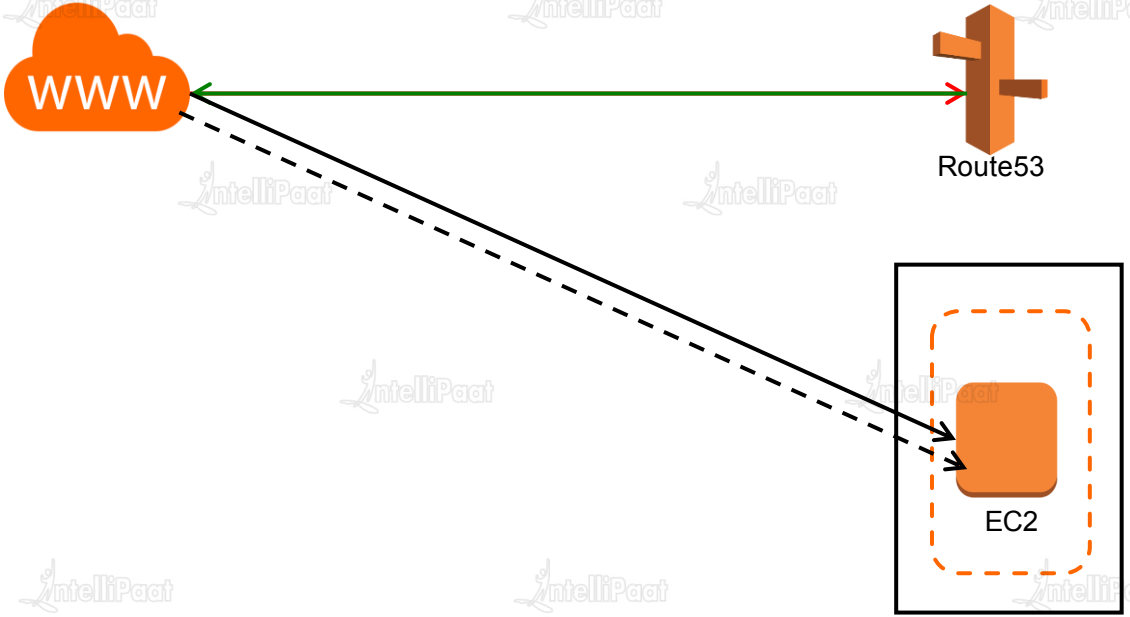
```
ns-2048.awsdns-64.net.  hostmaster.example.com.  1  7200  900  1209600  86400
```

  - ns-2048.awsdns-64.net: Host that created the SOA record.
  - hostmaster.example.com: email address of the admin with "@" being replaced by "."
  - 86400: Minimum TTL.

- **Private Hosted Zone** contains information about how to route traffic for a Domain within one or more VPCs.
  - Note: To use Private hosted zones, following VPC settings have to be set to TRUE.
    - enableDnsHostnames
    - enableDnsSupport

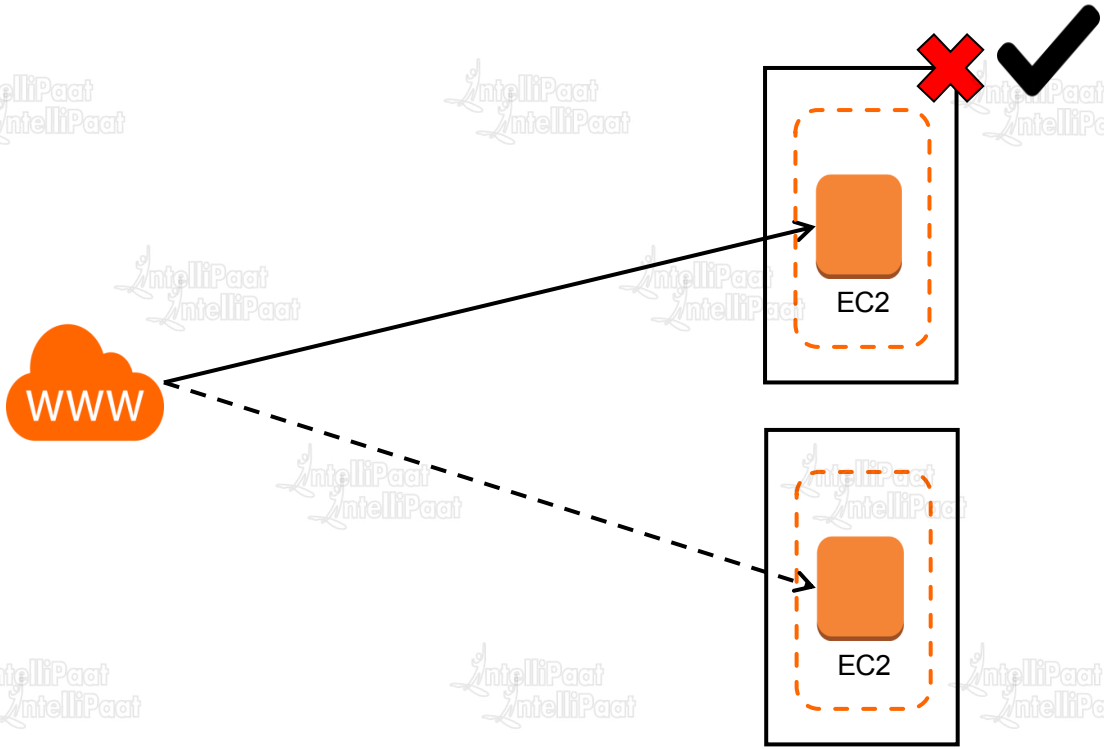
Routing Policy

Simple Routing Policy – Single server performing the desired operation.



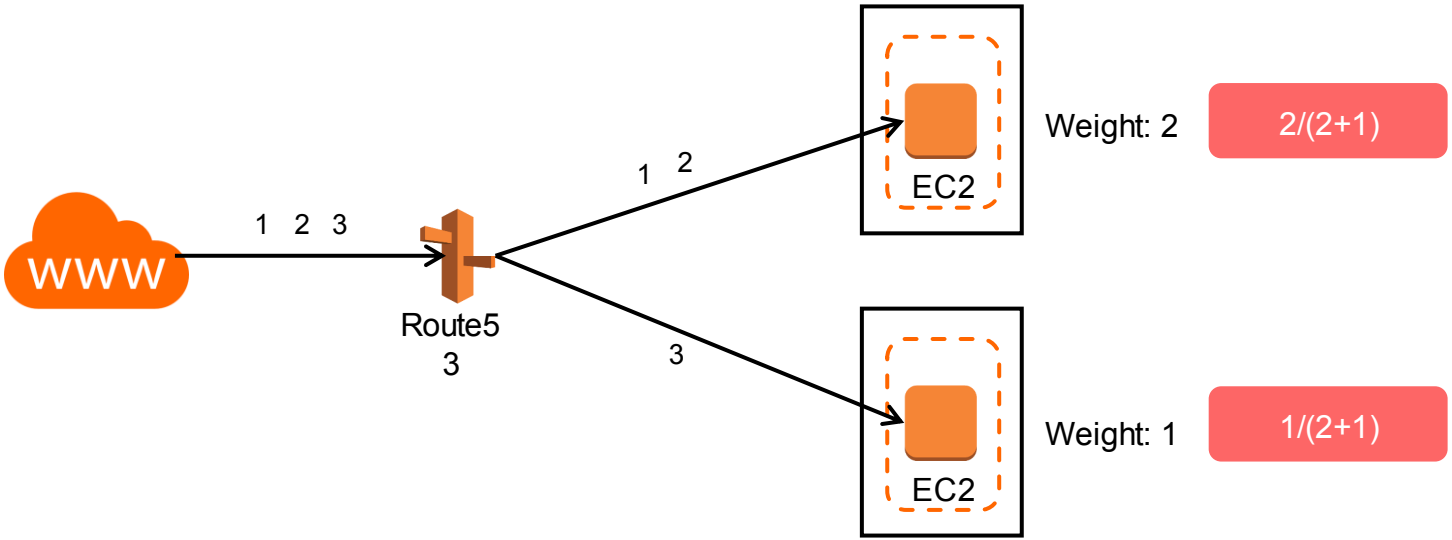
Routing Policy

Failover Routing Policy – Two servers performing the Active-Passive routing.



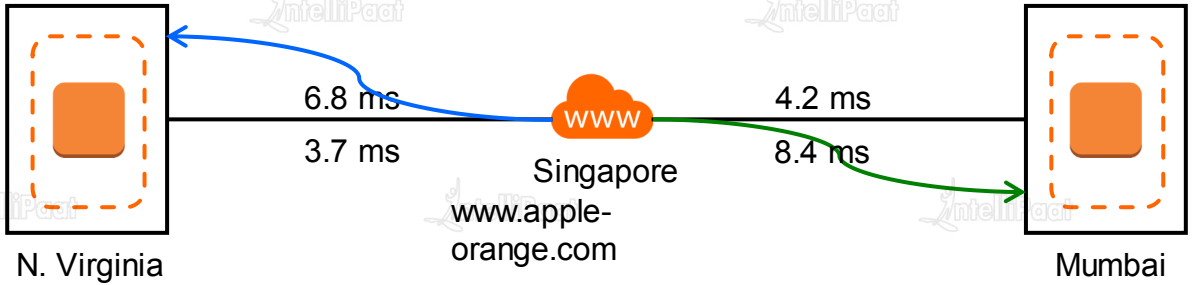
Routing Policy - Weighted

- Associate multiple resources with the same DNS name and type.
- Each Record Set is given a Weight and Set ID.



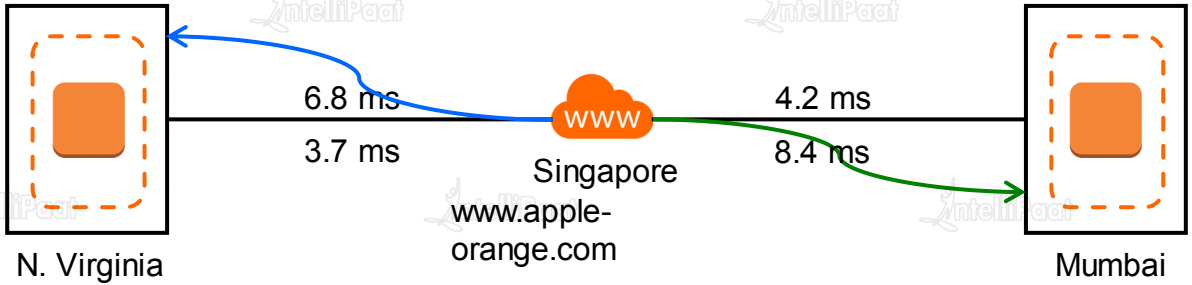
Routing Policy - Latency Based

- If an application is hosted on EC2 instances in multiple regions, user latency can be reduced by serving requests from the region where network latency is lowest.
- Create a latency resource record set for the Amazon EC2 resource in each region that hosts the application.
- Latency record sets can be created for both ELB and EC2 instances.
- Latency on the internet can change over time due to changes in routing or something else.



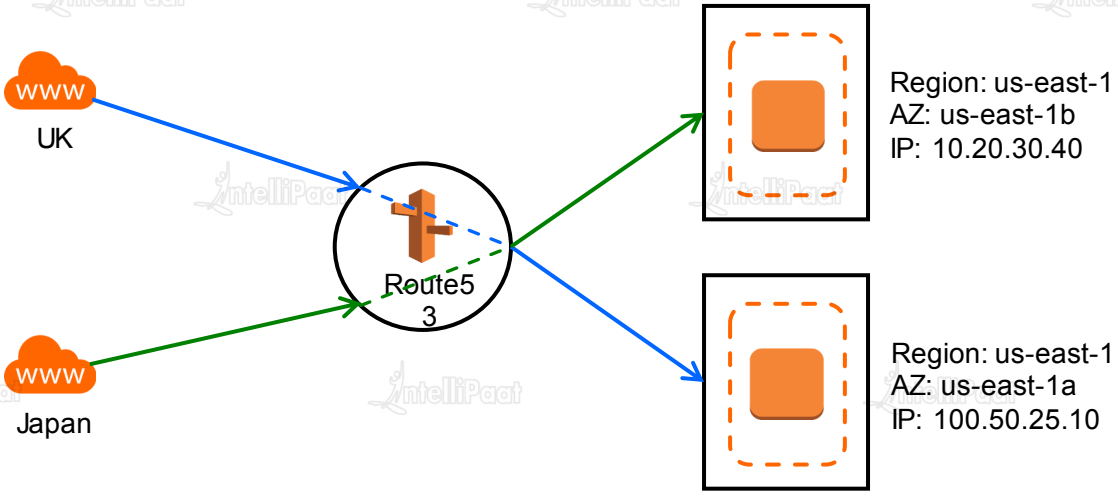
Routing Policy - Latency Based

- If an application is hosted on EC2 instances in multiple regions, user latency can be reduced by serving requests from the region where network latency is lowest.
- Create a latency resource record set for the Amazon EC2 resource in each region that hosts the application.
- Latency record sets can be created for both ELB and EC2 instances.
- Latency on the internet can change over time due to changes in routing or something else.



Routing Policy - Geolocation

- Geolocation routing can be used to send traffic to resources based on the geographical location of the users. e.g. all queries from Europe can be routed to the IP address 10.20.30.40.
- Geolocation works by mapping IP addresses, irrespective of regions, to locations.



Quiz

Quiz

1. Application Load Balancer functions at which layer of OSI Model?

A. 4

B. 7

A. 1

B. 6

Quiz

2. Network Load Balancer functions at which layer of OSI Model?

A. 4

B. 7

A. 1

B. 6

Quiz

4. In Autoscaling, Schedule scaling is based on CPU Utilization?

A. True

B. False

Quiz

5. Route53 Weighted Routing Policy is based on the latency.

A. True

B. False

Quiz

6. Route53 is allowed to create alias records.

A. True

B. False

India : +91-7847955955

US : 1-800-216-8930 (TOLL FREE)

support@intellipaat.com

24X7 Chat with our Course Advisor