O'REILLY

AWS Certified Solutions Architect Associate Exam Prep

Domains of Knowledge for SSA-C03 Exam

- Design Secure Architectures 30 %
- Design Resilient Architectures 26 %
- Design High-Performing Architecture 24 %
- Design Cost-Optimized Architecture 20%

Required Knowledge

- AWS Management Console and the AWS Command Line Interface (CLI).
- Understanding of the AWS Well-Architected Framework.
- AWS networking, security services, and AWS global services.
- Ability to identify which AWS services meet a given technical requirement and to define technical requirements for an AWS-based application.

Required Knowledge

- Compute, networking, storage, and database services.
- AWS deployment and management services.
- Knowledge and skills in deploying, managing, and operating workloads on AWS.
- Implementing security controls and compliance requirements.

Steps for AWS Certification Success

- The Solutions Associated Architect exam is based on common sense.
- Every question is a “situation” - current or select a proposed solution.
- Architects propose solutions based on existing building blocks.

AWS Documentation

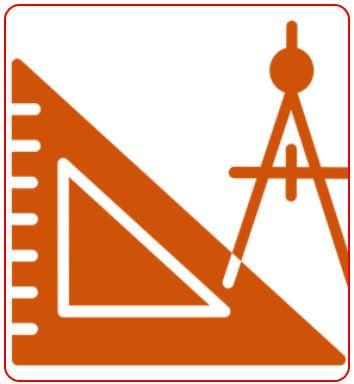
- Well-Architected Framework Labs.
- Well-Architected Framework PDFs.
- AWS FAQ’s.
- AWS free tier.

How the Exam is Graded

- 65 questions - 130 minutes to complete.
- Only 50 count!
- 72% is a pass.
- During the exam, mark questions for future review.
- Answer all questions.
- Questions are multiple-choice.
- Single-selection and multiple-selection.
- No penalty for guessing.
- Certification is valid for three years.

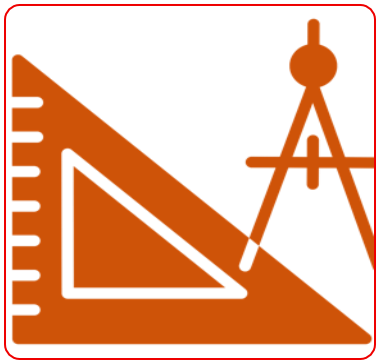
Well-Architected Framework

- The exam is based on the Well-Architected Framework.
- Compare your workload against Well-Architected Framework guidance to produce resilient, stable, and efficient workloads.
- AWS Well-Architected Tool.



Well-Architected Framework

- Security - Protect data, systems and assets.
- Reliability - The application stack performs its intended function correctly and consistently.
- Performance Efficiency - Use compute resources efficiently to meet system requirements and maintain efficiency as demand changes.
- Cost Optimization - Business value at the lowest price point.



Well-Architected Framework Pillars

- Security
- Reliability
- Performance Efficiency
- Cost Optimization

Designing for Reliability and Resiliency

Resilient Architecture Concepts

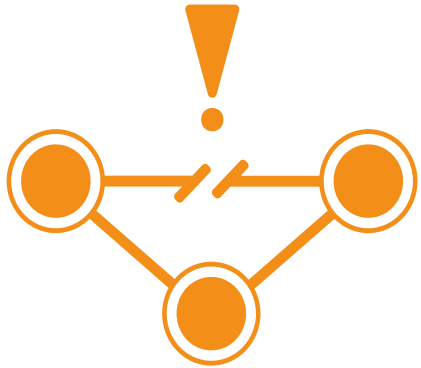
- Workload architecture - high availability (Multi-AZ, multi-region)
- Service quotas of deployed resources.
- Constraints to design - Instance size (vCPU, memory, network, and storage speeds).
- Failure management - backups, fault isolation, managing component failures, disaster recovery.

High Availability (HA)

- RTO – How long until recovery?
- RPO – How much data was lost?
- High availability designs lower the Recovery Time Objective and the Restore Point Objective.
- Automate the recovery of system components that are part of the application stack.

Fault Tolerance

- Fault-tolerant application stacks withstand subsystem failure and maintain availability.
- Workloads use spare (or redundant) subsystems when necessary.

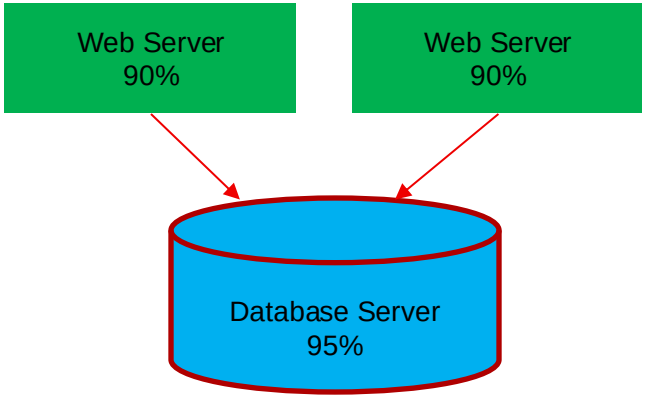


Redundant Component Design

- Use multiple independent availability zones.
- A single AZ has an SLA = 99.9 %.
- Designing with two AZs = 99.9999 %

SLA Example: Desired 99.5%

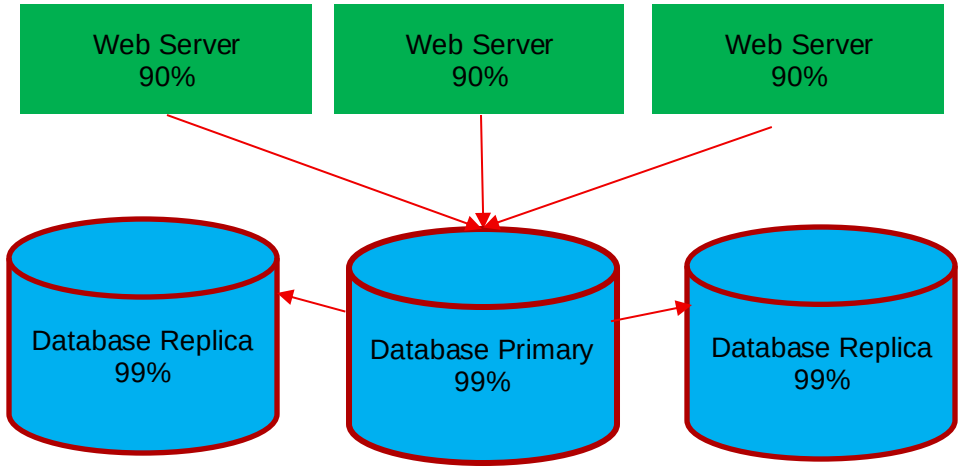
- **Weekly:** 25m 30s
- **Monthly:** 1h 49m 58s
- **Quarterly:** 5h 29m 54s
- **Yearly:** 21h 59m 36s



Total Availability: 94.%

- **Weekly:** 5h 6m 0s
- **Monthly:** 21h 59m 36s
- **Quarterly:** 2d 17h 58m 48s
- **Yearly:** 10d 23h 55m 10s

SLA Example: Desired 99.5%



Total Availability: 99.8.%

- **Weekly:** 10m 12s
- **Monthly:** 43m 59s
- **Quarterly:** 2h 11m 58s
- **Yearly:** 8h 47m 50s

AWS Global Infrastructure

AWS Global Infrastructure

Online Regions

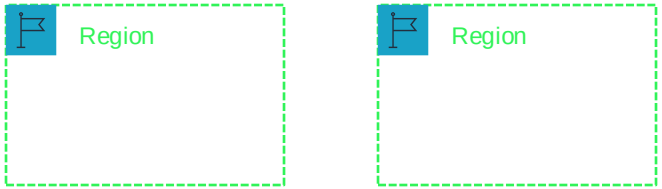
Availability Zones

Local Zones

Edge locations / Points of Presence

AWS Regions

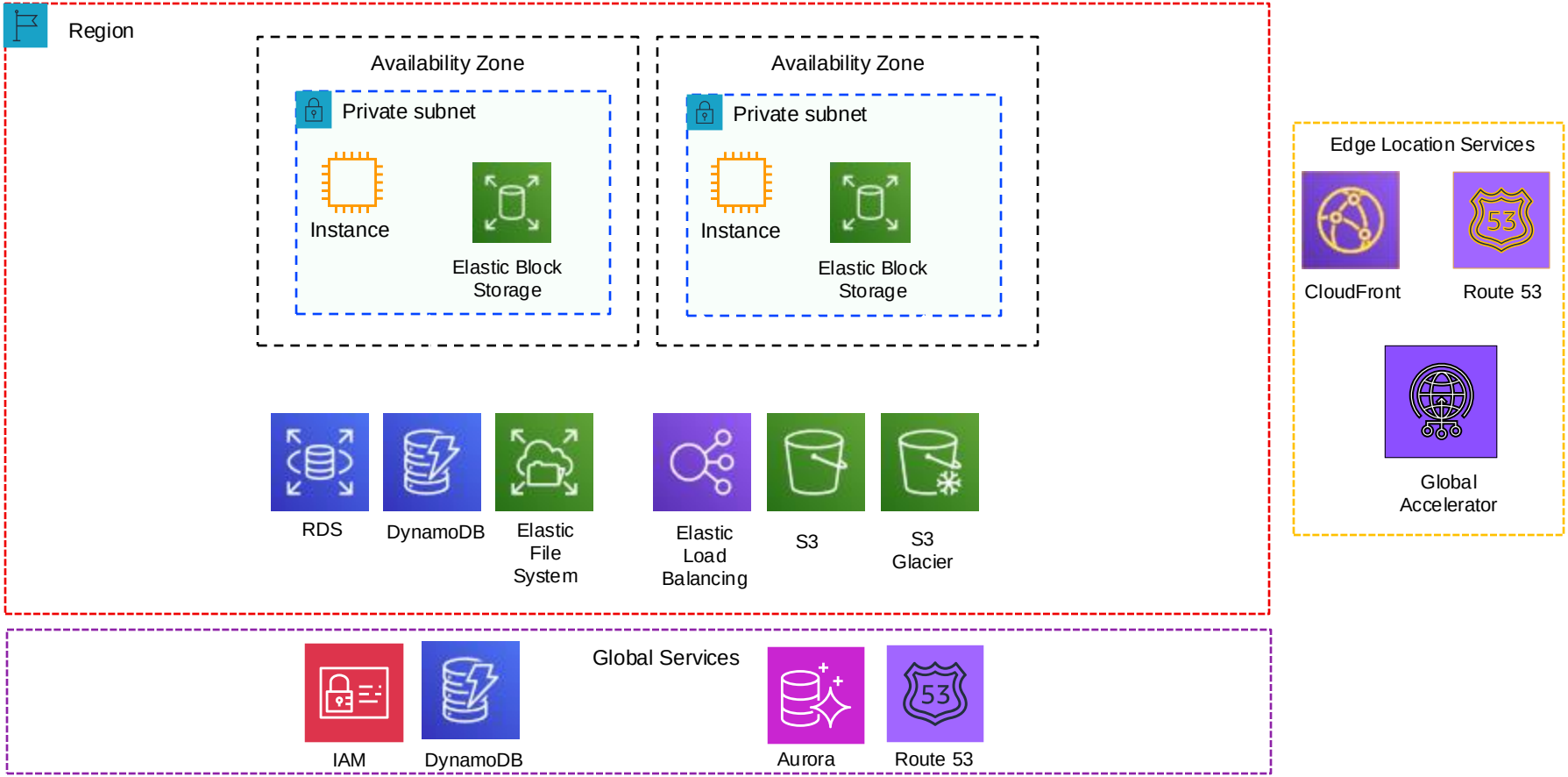
- Areas of the world where Amazon offers AWS cloud services.
- Each region is a geographical location.
  - Where do you operate?
  - Where are your customers?
  - Where are you allowed to operate?



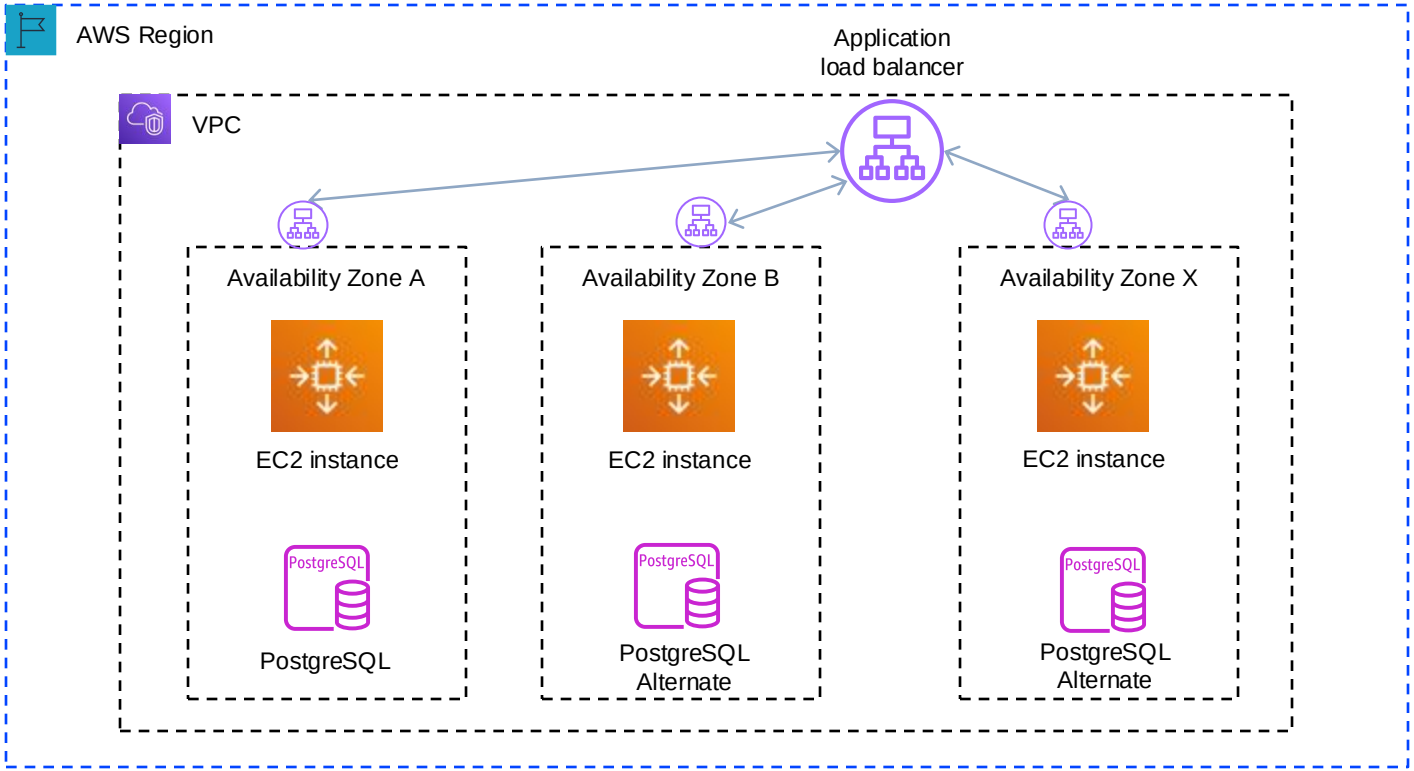
- Each region is completely independent and isolated.
- Pricing differences depend on geographical location.
- Traffic sent across AWS regions faces additional charges for egress traffic flow.

AWS Regions

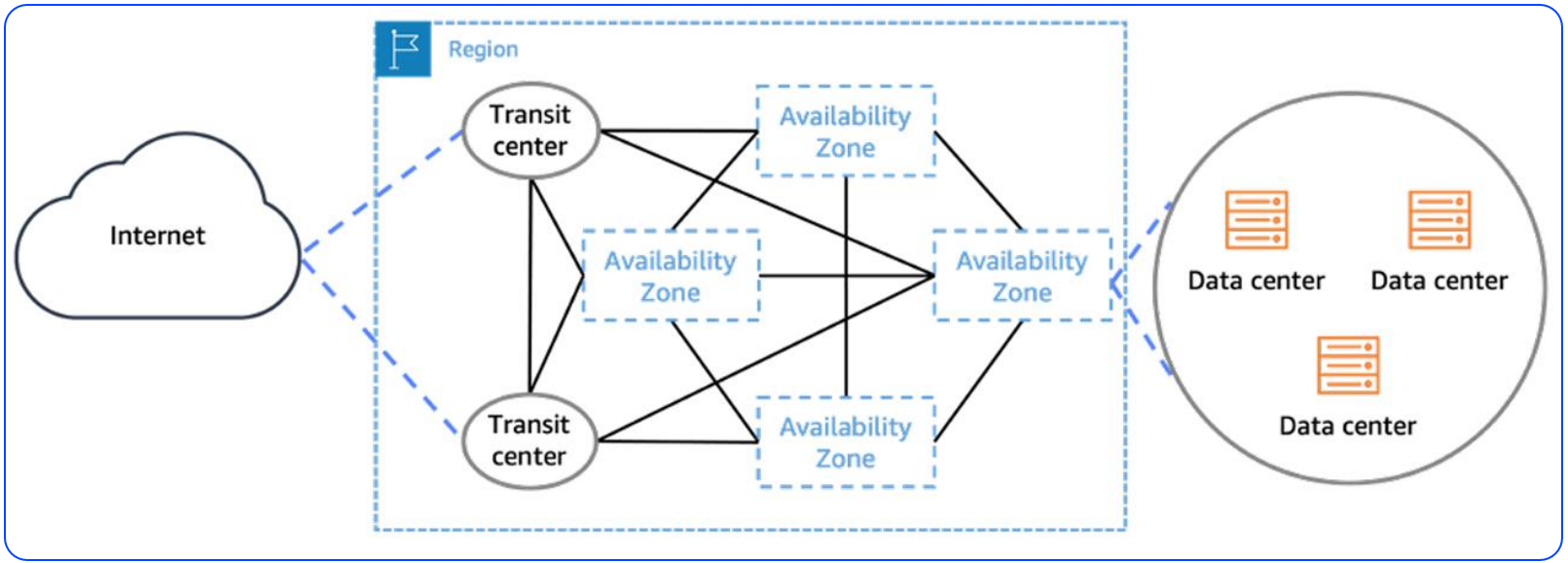
AWS Cloud Services



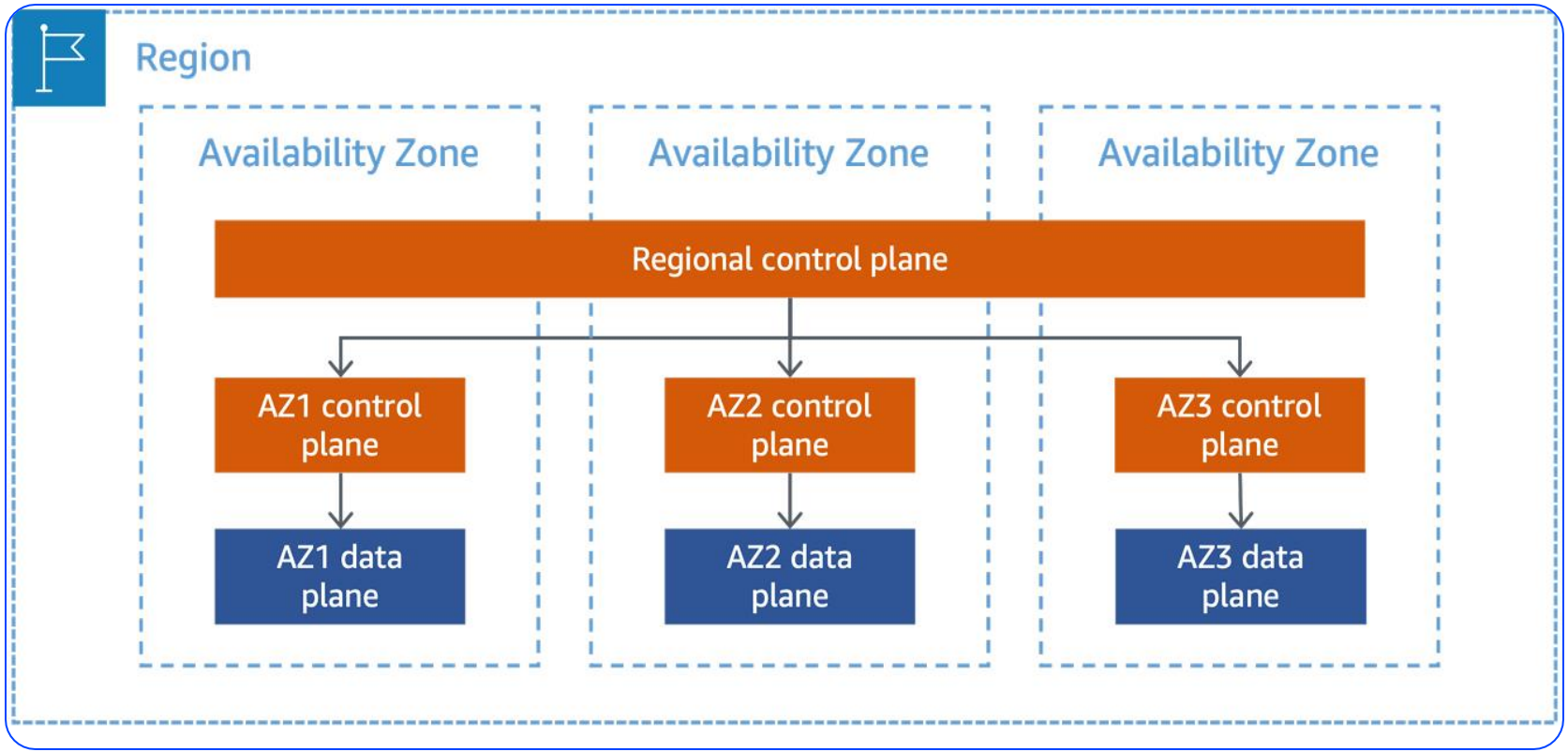
Multiple-Availability Zone Design



Availability Zone Design



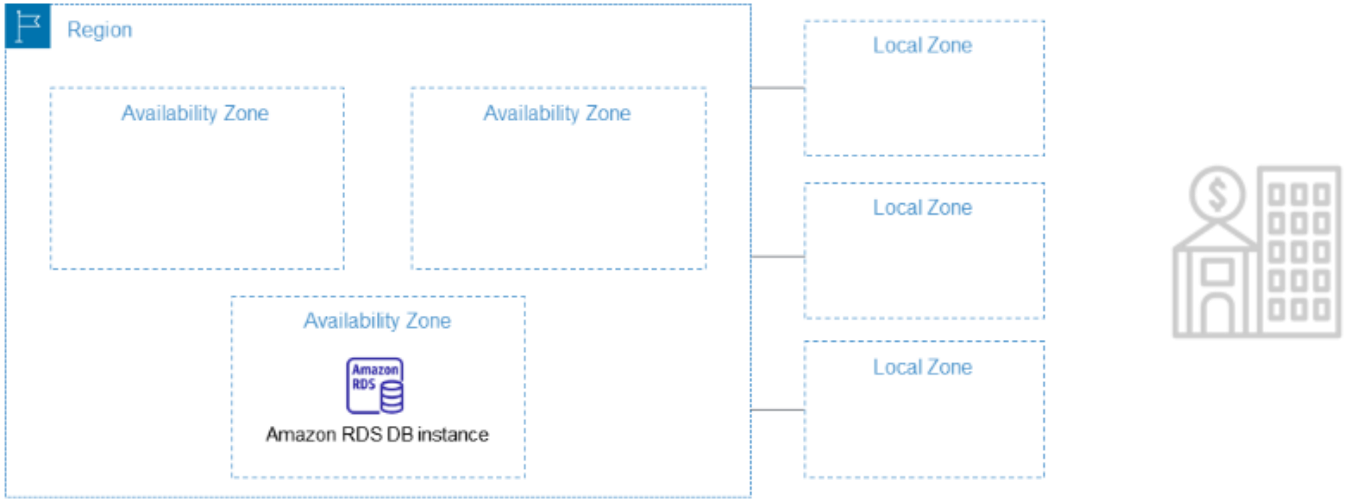
Availability Zone Independence



Availability Zones in Operation

- EC2 instances can launch across multiple subnets hosted in multiple availability zones.
- ELB can target EC2 instances across multiple availability zones.
- EC2 auto-scaling scales EC2 instances across multiple availability zones.
- RDS solutions are replicated across multiple availability zones.
- RDS Aurora uses three AZs (Can deploy as a multi-region DB).
- DynamoDB uses three AZs (Can deploy as a multi-region global table).

Local Zone Architecture



- Access compute and storage services with single-digit millisecond latency.
- VPCs can include local zone subnets.

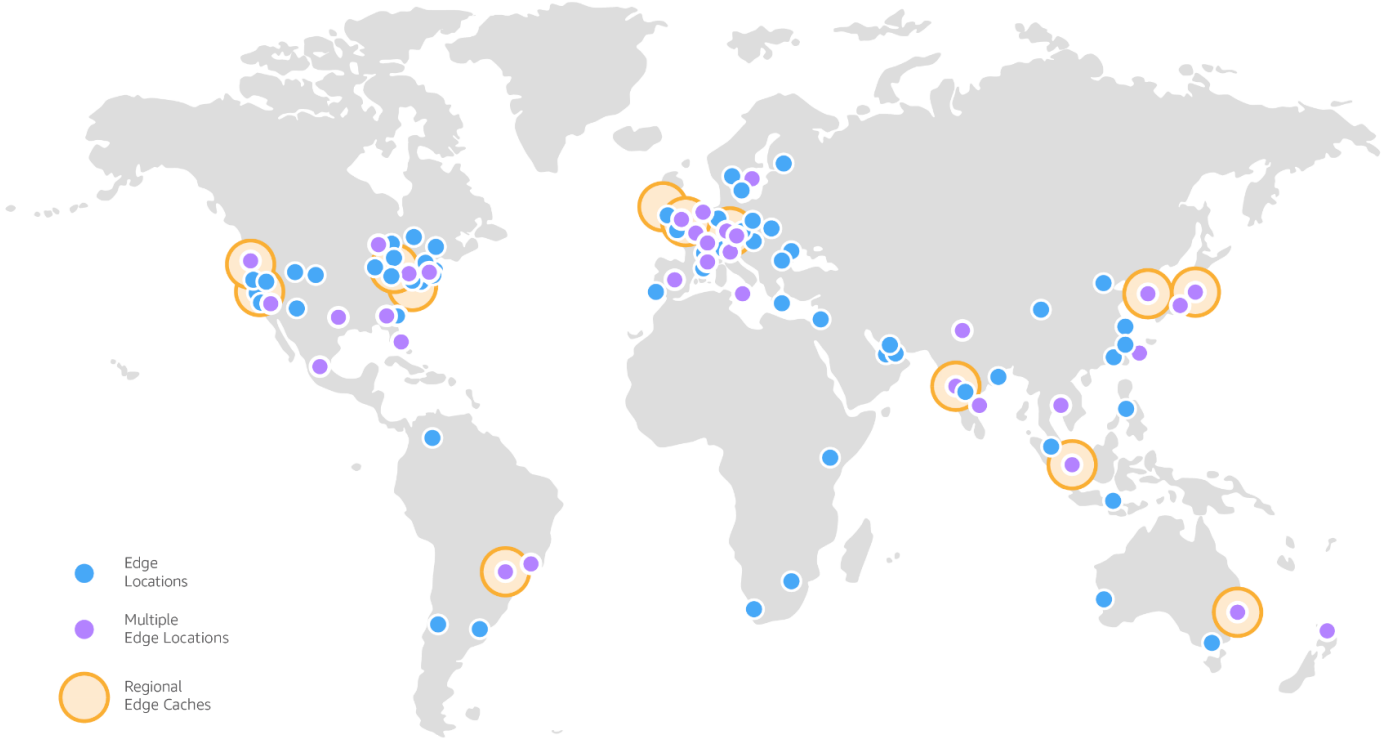
Low Latency Solutions at AWS

- **AWS Local Zones** provide single-digit millisecond latency for video rendering and CAD services.
- **AWS Outposts** allow you to run AWS computing and storage services on-premises.
- **AWS Wavelength Zones** extend AWS infrastructure into 3rd party telco 5G data centers.

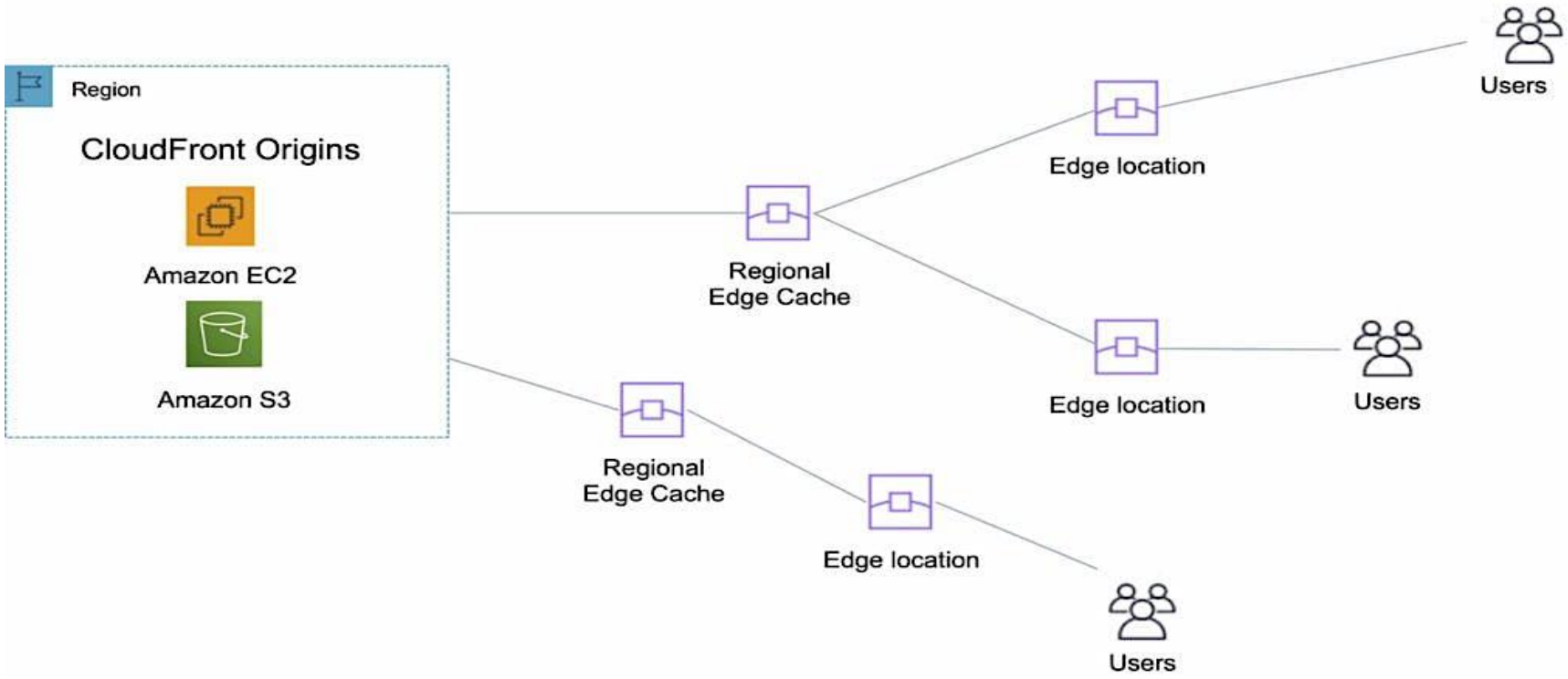
Edge Locations

Each edge location has caching data centers and edge services directly connected to the AWS cloud using high-speed private network links.

AWS Edge Locations



Regional Edge Cache



In Class Project

- Choose the AWS region for deployment and availability zones for several scenarios.

Edge Services

CloudFront

Caches your static and dynamic content.

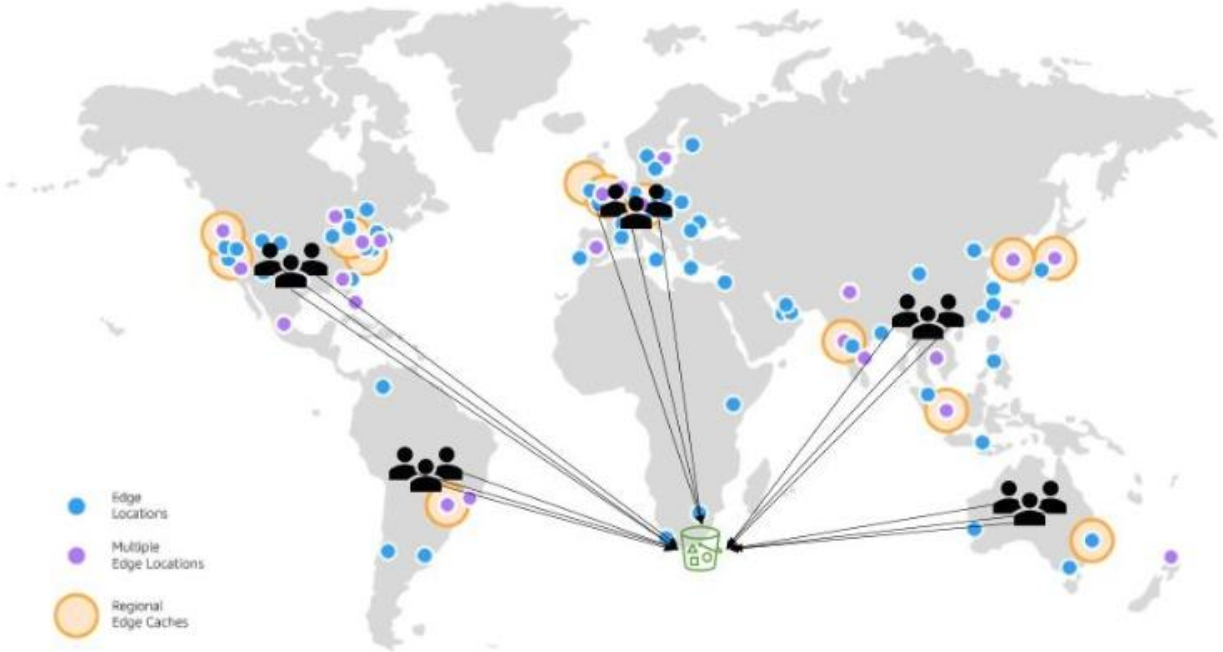
Route 53

Delivers your request from closest edge location.

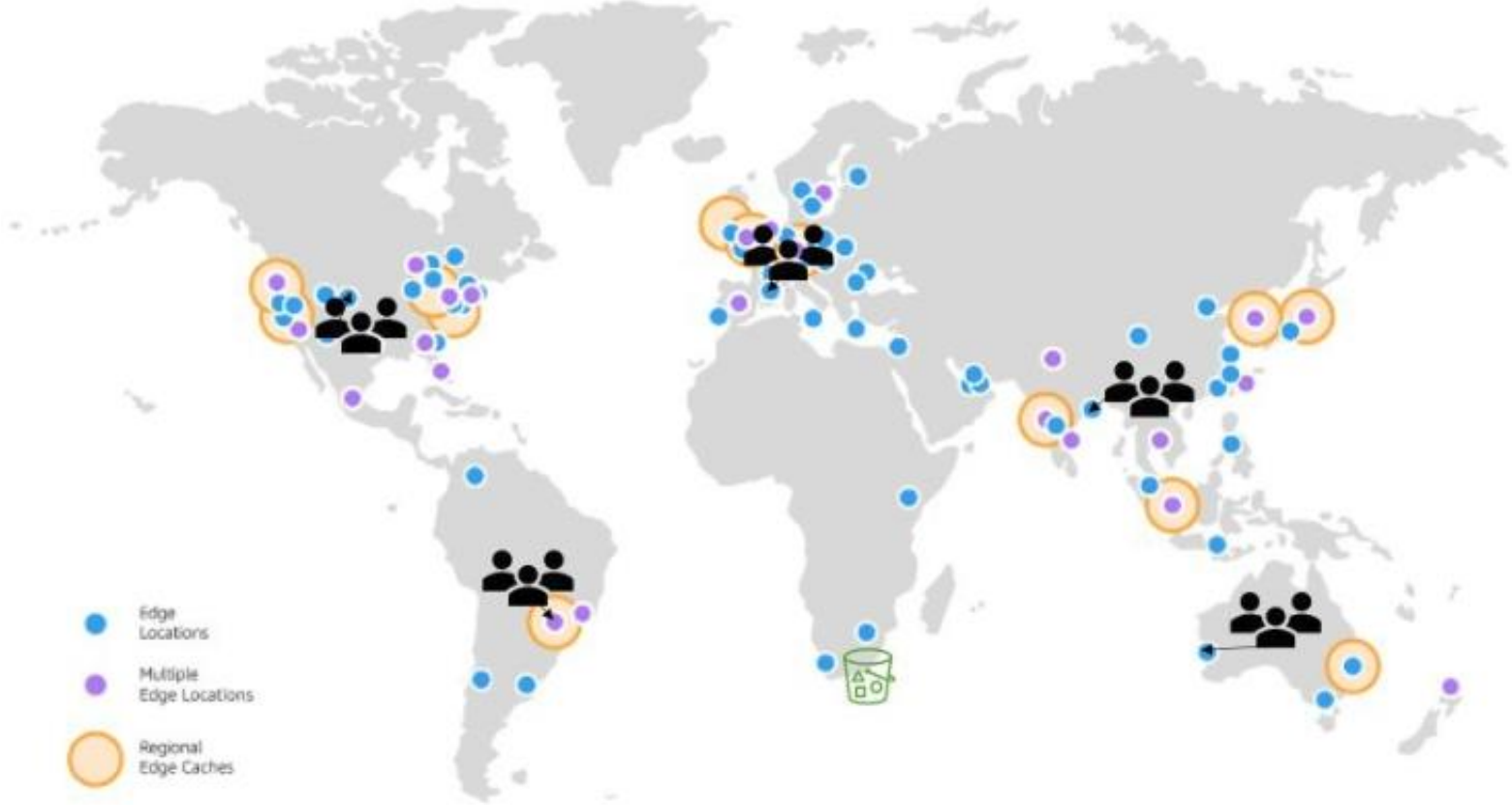
WAF

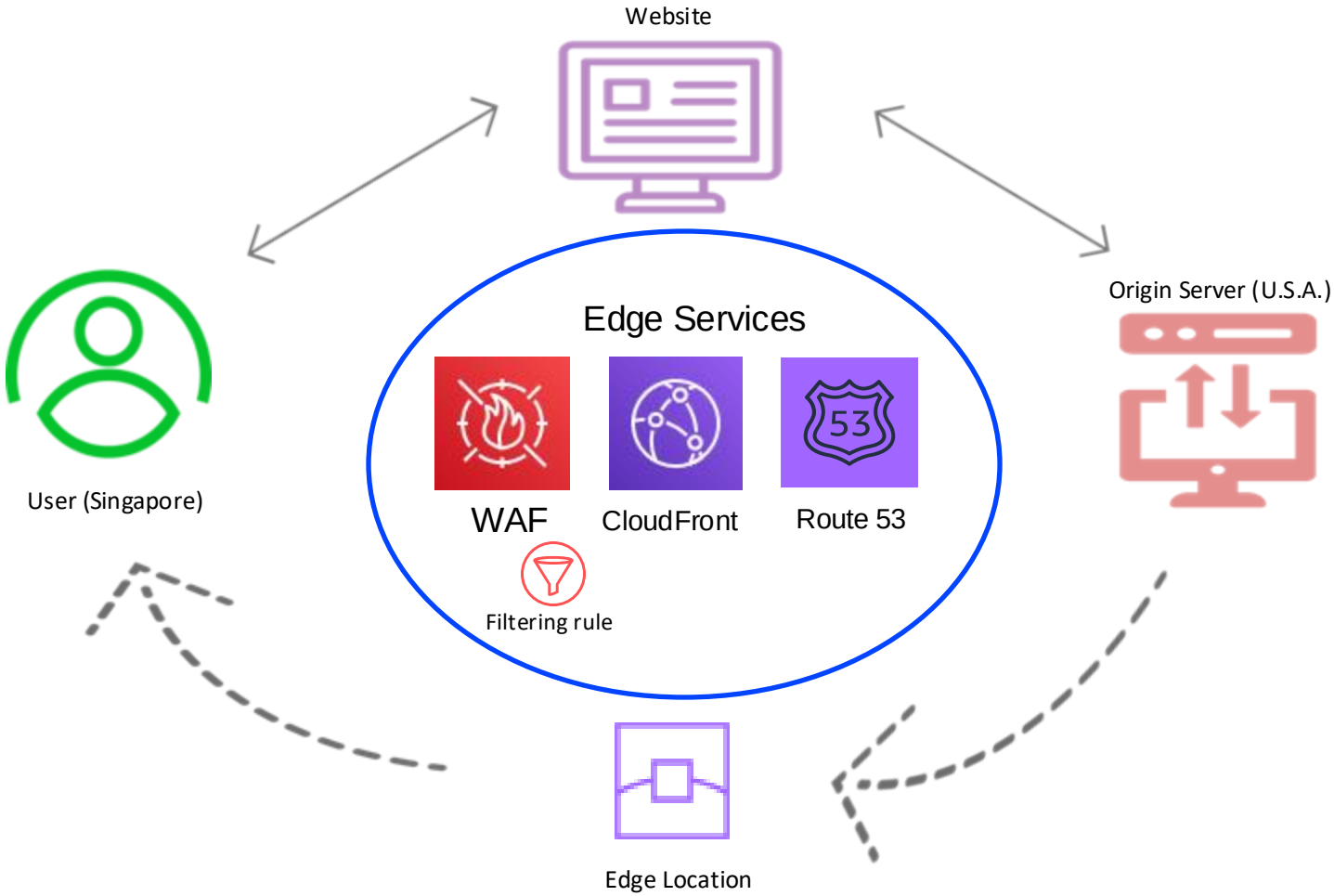
Filters incoming public traffic.

Without CloudFront



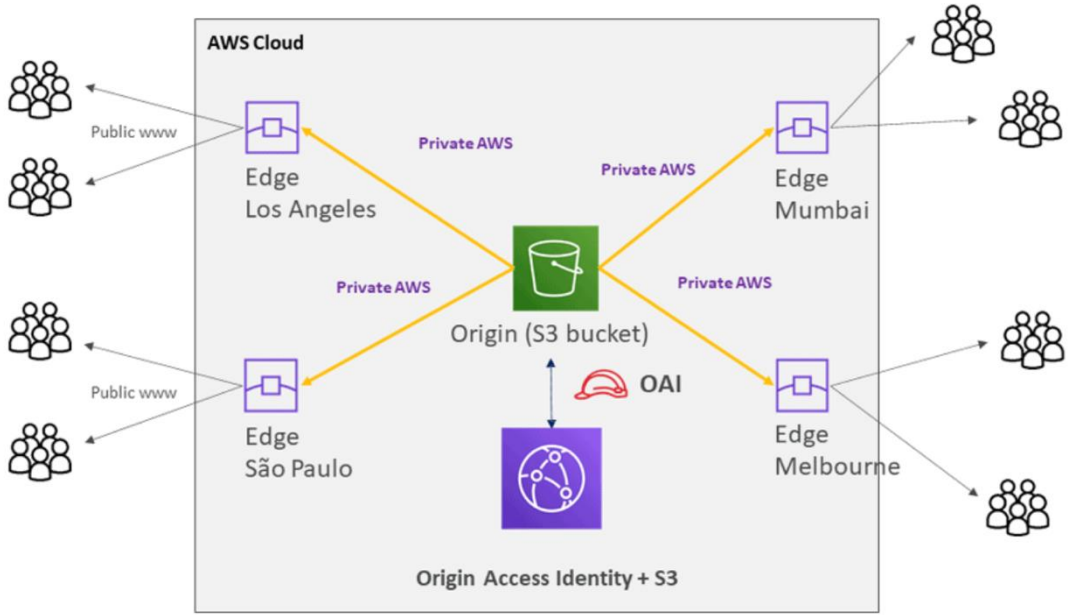
With CloudFront




Website
Origin Server (U.S.A.)
User (Singapore)
Edge Services
WAF
CloudFront
Route 53
Filtering rule
Edge Location

CloudFront Operation

- Content is delivered to a worldwide distribution of edge locations.
- Requests for content served with CloudFront are routed to the closest edge location with the lowest latency.



Delivering Content to End Users

1. A user accesses a website and requests files (image and HTML files).
2. Route 53 routes the request to the lowest latency edge location.
3. CloudFront checks the edge cache for the requested file; if the files are present, they are delivered to the end-user.
4. If files are not present, CloudFront compares the request with the distribution and forwards requests for the files to your origin.
5. Requested files are delivered to the end-user and are added to the cache.

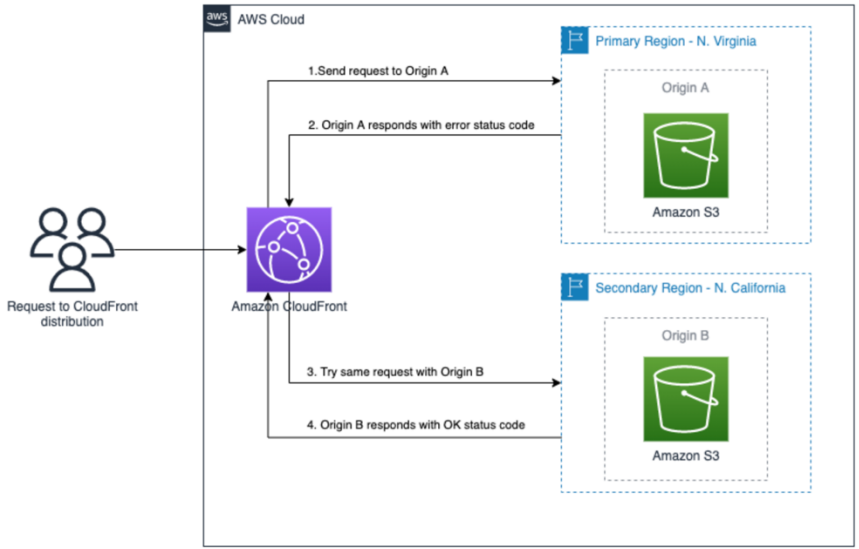
Serving Private Content

- Configure CloudFront for users to access content using *signed URLs* or *signed cookies.*
  - Signed URLs restrict access to individual files.
  - Signed cookies provide access to multiple restricted files.

- Secure access only through CloudFront with Origin Access Identity (OAI).

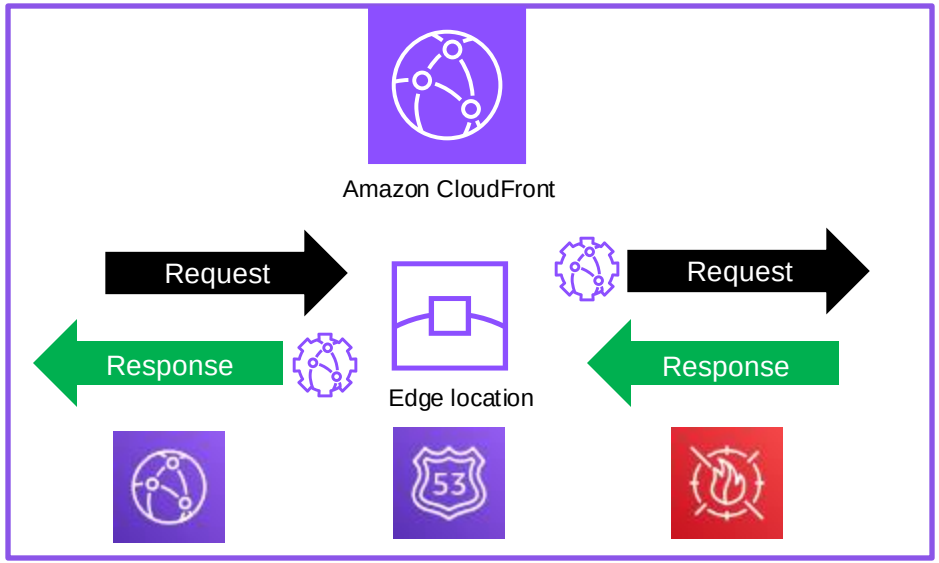
CloudFront Origin Failover

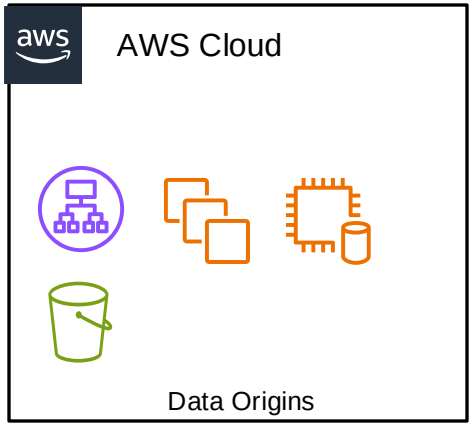
- Set up CloudFront with origin failover for scenarios that require high availability.
- Create an *origin group* with two origins: a primary and a secondary.
- Specify the connection timeout: 1 - 10 seconds.
- CloudFront automatically switches to the secondary origin when the primary origin is unavailable.



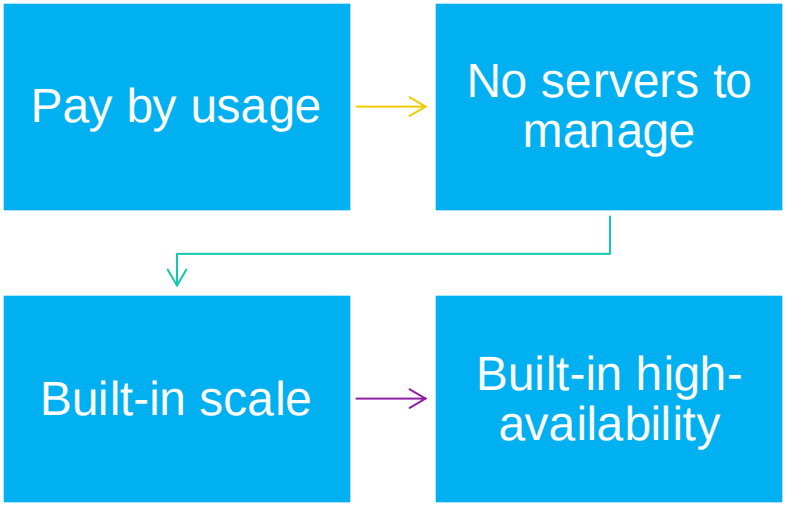
Lambda@EDGE

.


Amazon CloudFront
Request
Response
Edge location
Request
Response
53


aws
AWS Cloud
Data Origins

Serverless Computing

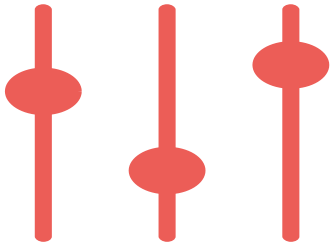


AWS Lambda

- No EC2 instances to manage and scale.
- Background scaling is handled by AWS.
- Sub-second metering – pay when each function executes.
- Bring your own code- Node.js, Java, Ruby, Python, C#, Go.
- Integrates with other AWS services.
- Select power rating from 128 MB to 1.5 GB: CPU and network will be proportionally allocated

Customizing with Lambda@Edge

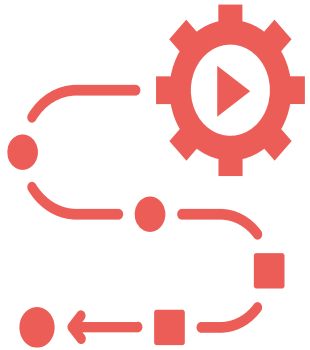
- Lambda@Edge allows you to execute functions at the edge location.
- Lambda functions can be executed for the following CloudFront events:
  - CloudFront receives a request from a viewer.
  - CloudFront forwards a request to the origin server.
  - CloudFront receives a response from the origin.
  - CloudFront returns the response to the viewer.



Lambda Edge Processing

Lambda@Edge processing could include:

- Inspecting cookies.
- Rewriting URLs.
- Return different objects to viewers based on the device they're using.
- Make network calls to confirm user credentials.



WAF and Shield

Protecting the Application Perimeter

AWS Shield Standard

Protects against DDOS attacks.

AWS Shield Advanced

Enhanced protection.

AWS WAF

Protection with rules.

Filtering rule

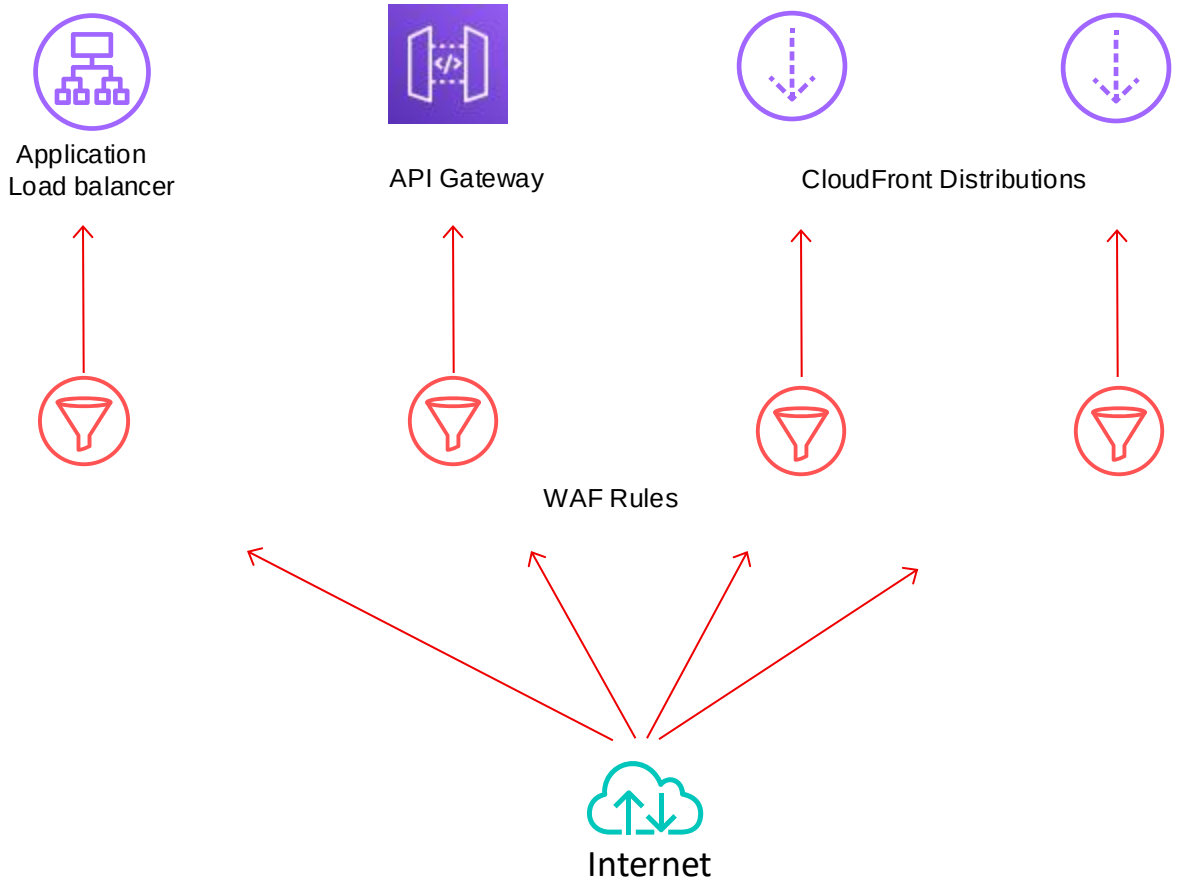
WAF Traffic Filtering

Traffic is allowed except for specific requests.

Traffic is denied except for specific requests.

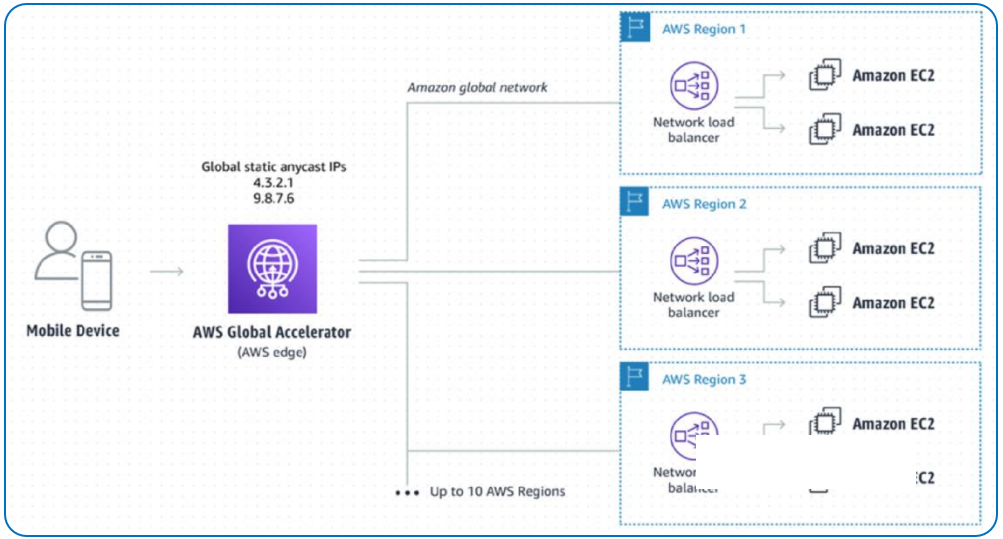
Count incoming requests and decide.

WAF Design


Application
Load balancer
API Gateway
CloudFront Distributions
WAF Rules
Internet

AWS Global Accelerator

- User app requests are directed across the AWS private network using edge locations.



Route 53

Route 53 Features

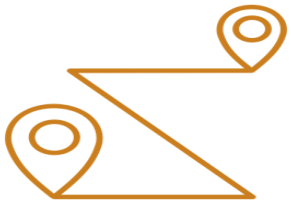
- Global traffic management; direct requests to the correct endpoint.
- Health checks and monitoring of AWS resources.
- Zone-apex support for CloudFront distribution and S3-hosted websites.
- Resolve with recursive DNS for both VPC and on-premise networks.

Hosted zone

Resolver

Routing Policy Types

- Simple
- Weighted
- Latency
- Failover
- Geolocation



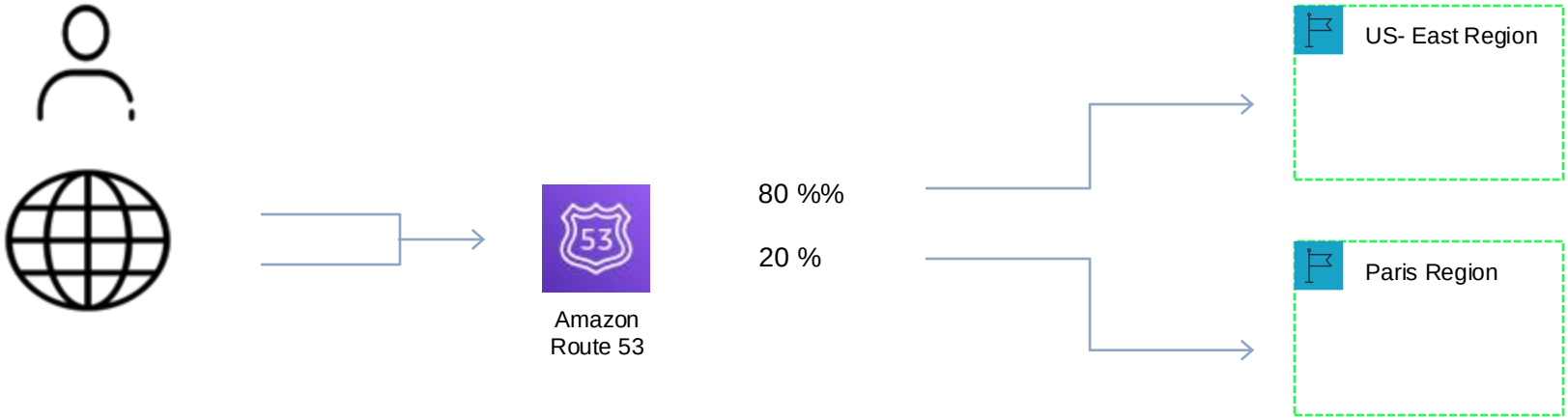
Simple Routing Policy

- The default routing policy when a new record is created in Route 53.

Used when you have a single resource.

Weighted Routing Policy

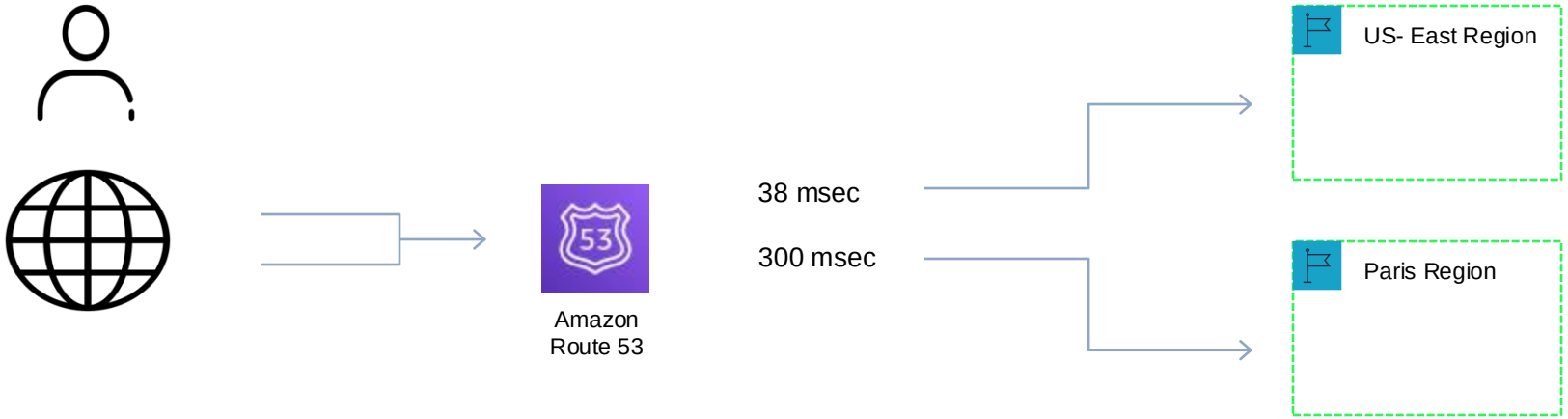
- Split traffic based on different weights assigned to resources.



Assign 20% of your traffic to one region and 80% to the other region.

Latency Based Routing

- Route traffic based on the lowest network latency for your end user.
- Create a latency resource record set in each region that hosts your resource.



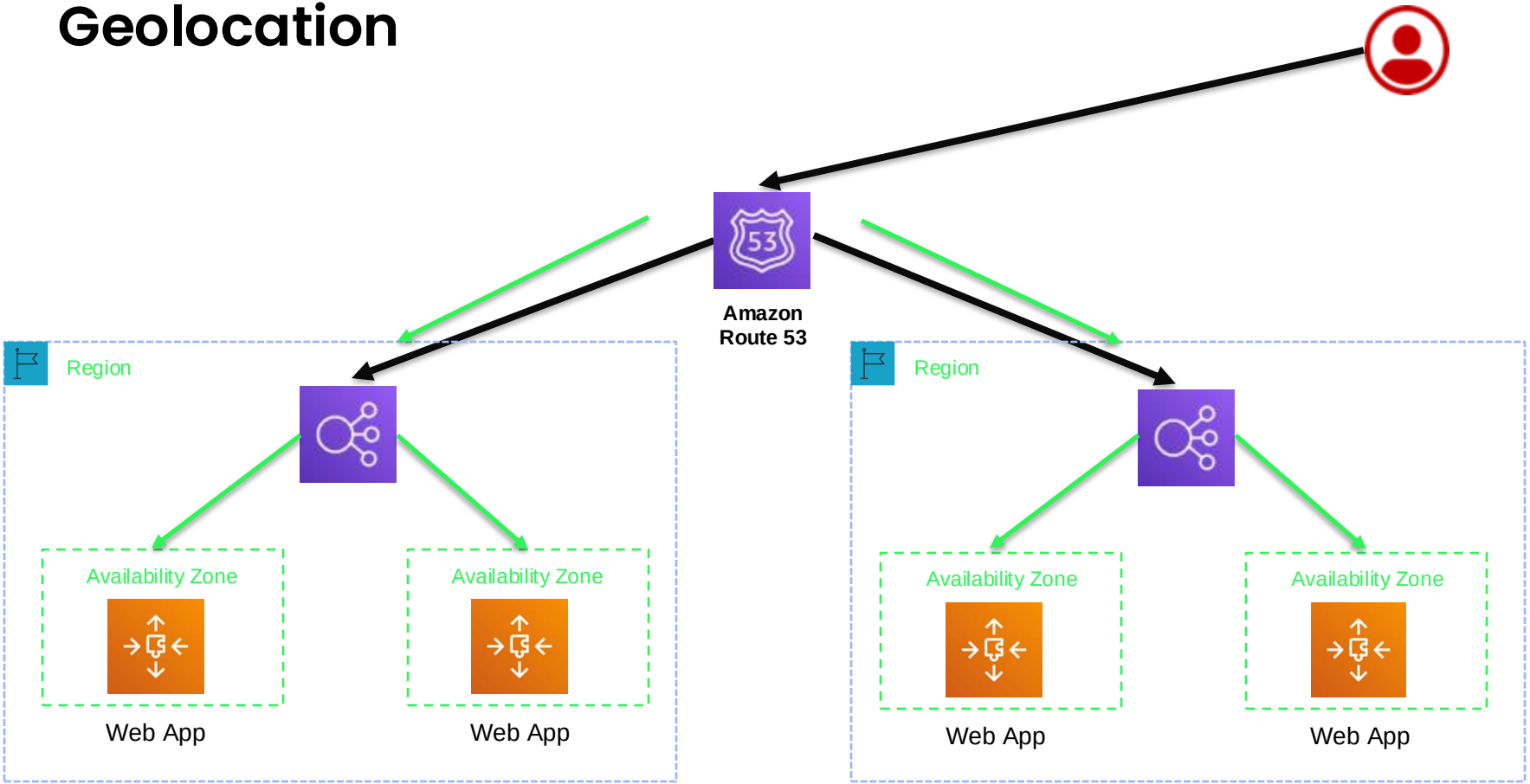
Route 53 selects the latency resource record for the region with the lowest latency.

Failover Routing Policy

- Highly available ELB load balancers in separate regions.

Create health check status; health of ELB, health of site / region.

Geolocation



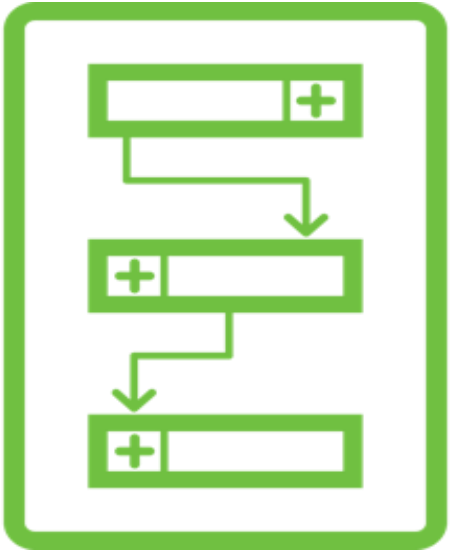
Networking Services

What’s a VPC?

- Networking layer at AWS Cloud.
- Logically isolated to your AWS account.
- EC2 instances are hosted on virtual private cloud subnets.

Multi-VPC Based on Requirements

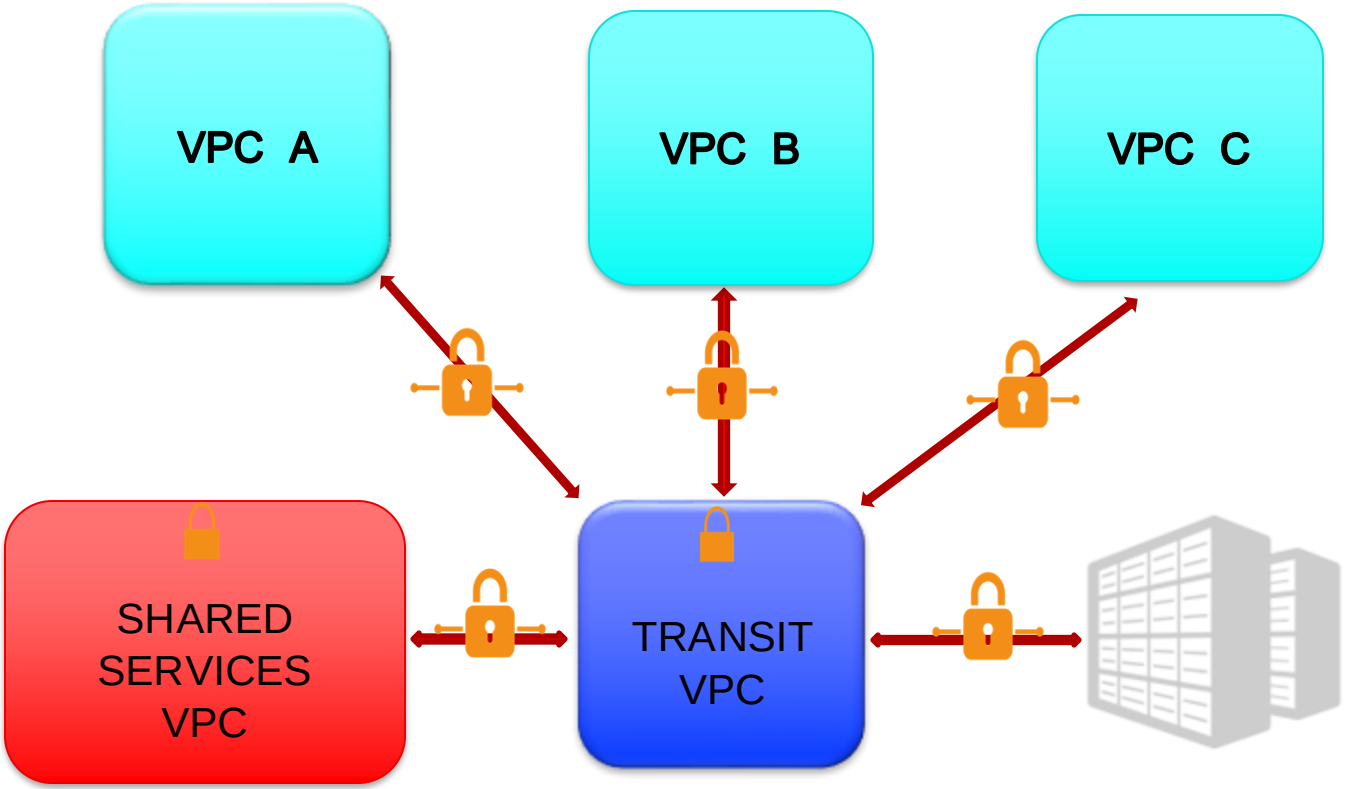
- Application stack isolation.
- Production from non-production.
- Multi-tenant / single-tenant isolation.
- Business unit design.
- Shared corporate services.



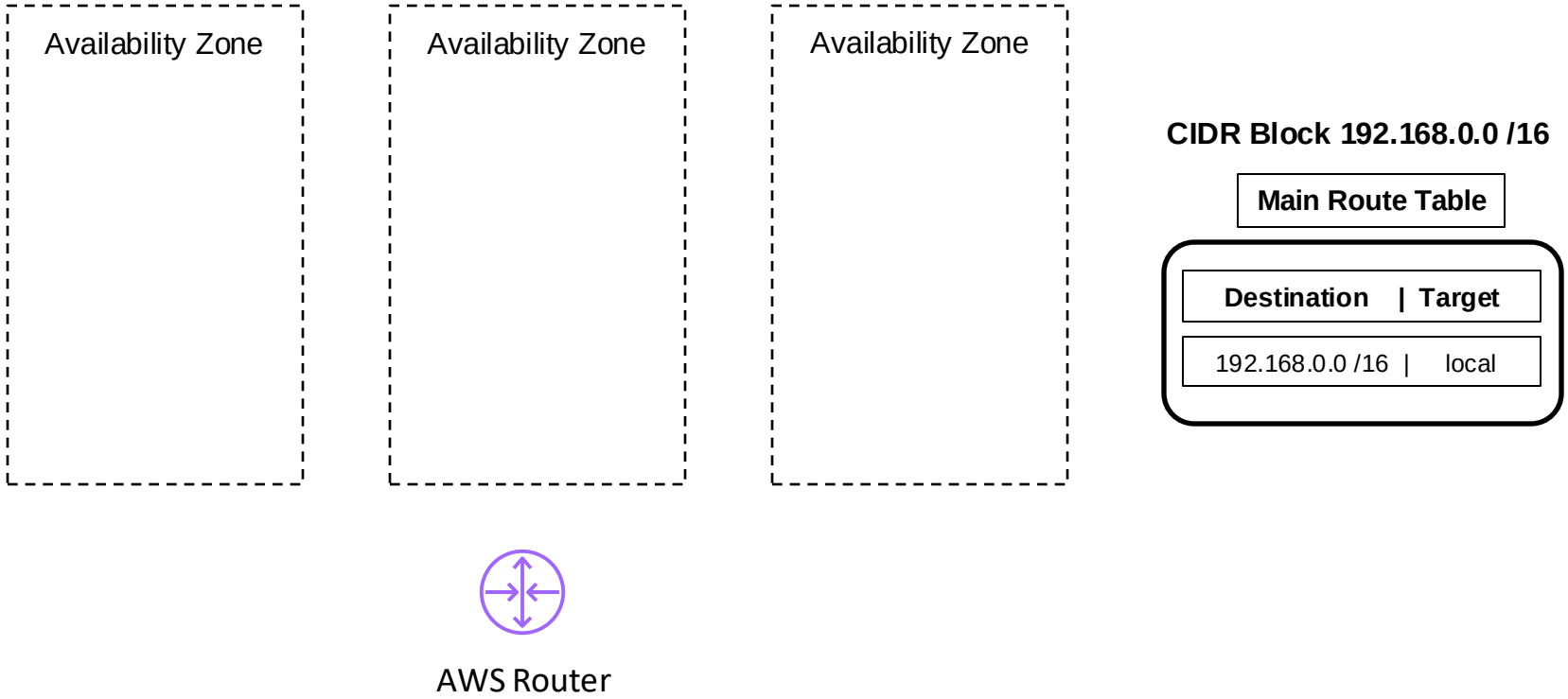
VPC Functional Design

- VPCs can span all availability zones of the AWS region.
- Subnets are hosted within the selected availability zone.
- Route table entries control subnet traffic.
- NACLs allow or deny subnet traffic.

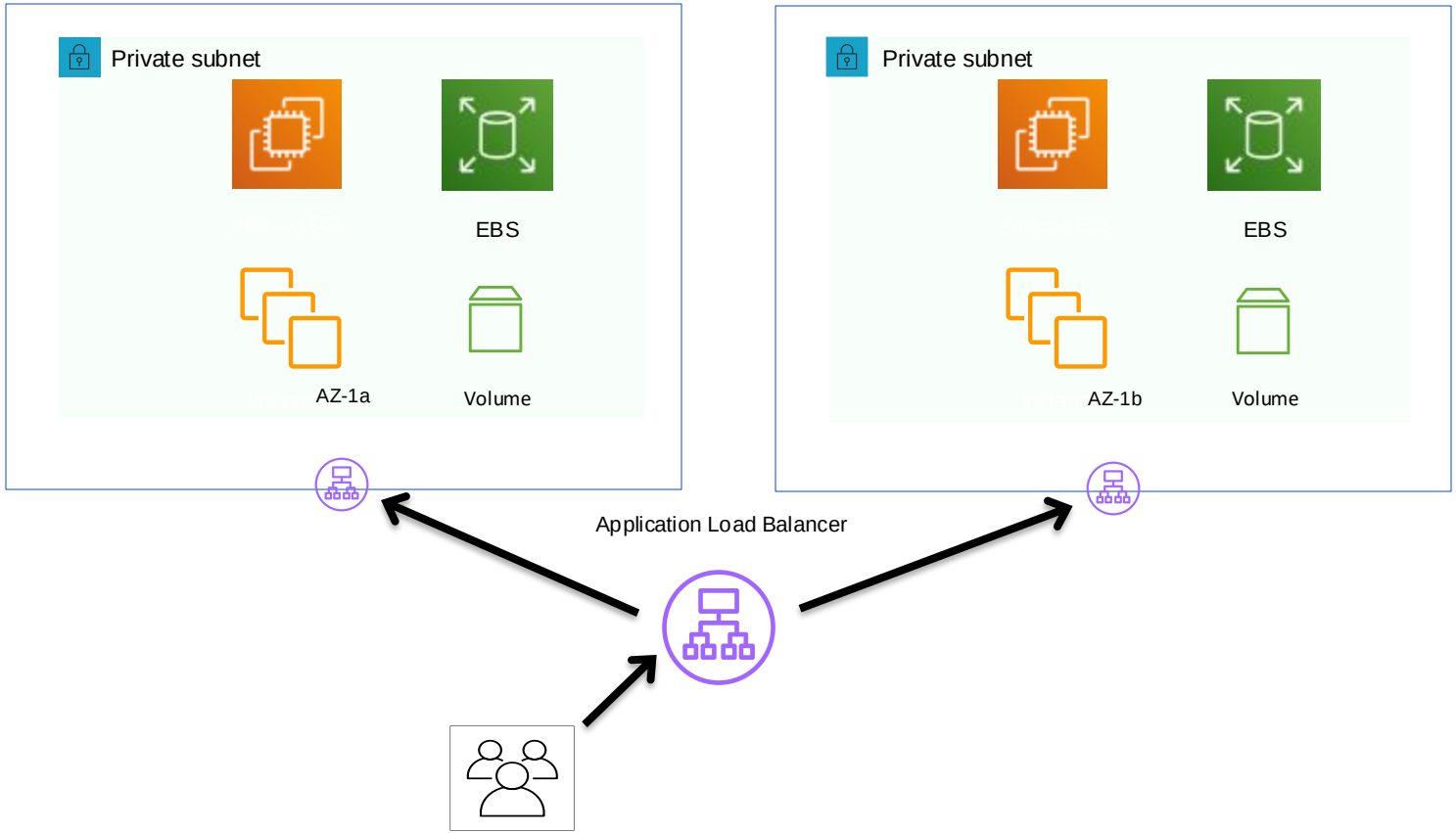
Multiple VPCs



VPC’s Span AZs



Failover Possibilities



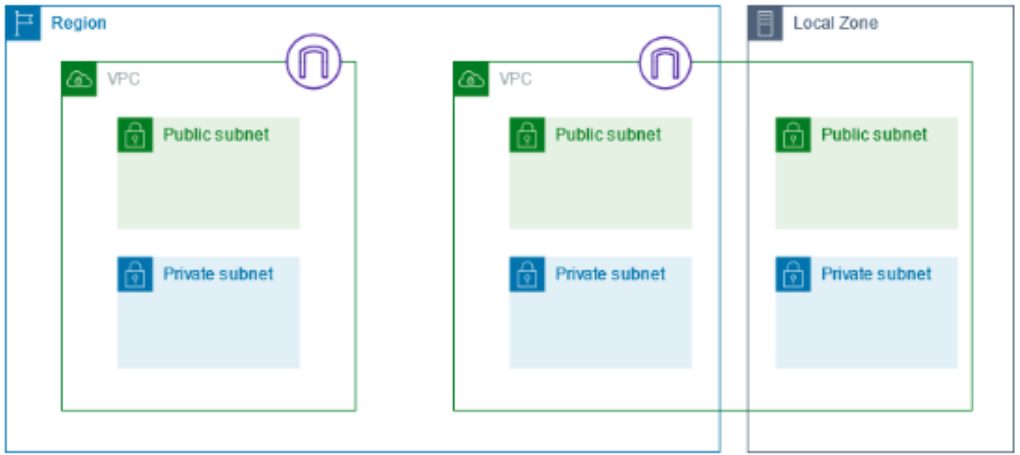
VPC Core Components

- Subnets
- Route tables
- Security groups (SG)
- Network ACLs

- Gateway devices
- Endpoints
- External connections
- Peering connections

Subnets

- EC2 instances are hosted on subnets.
- Public subnets can be used for resources /services that need direct Internet access (IGW, ELB, NAT Services).
- Private subnets host resources that don't directly connect to the Internet (Web and Application servers, RDS instances).



Subnet Rules

- Public or private subnets can be created in availability zones.
- Subnets cannot span across multiple availability zones.
- A subnet that doesn’t directly route to the Internet gateway is private.
- A subnet that directly routes to an Internet gateway is public.

IP Addresses

- Each EC2 instance is assigned a private DNS hostname associated with its IP address.
- IPv4 is default and required; IPv6 is optional.

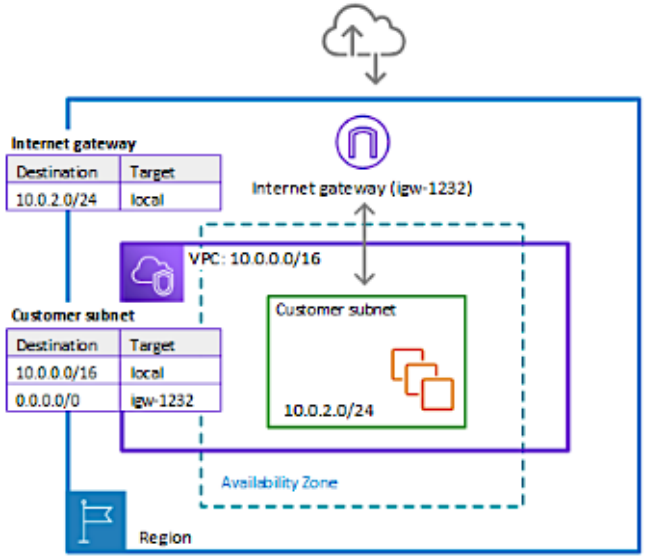
- EC2 instance address types:
  - Private / Public IP v4
  - Static public IPv4 - elastic IP address, BYOIP
  - Public IP v6

AWS Gateways

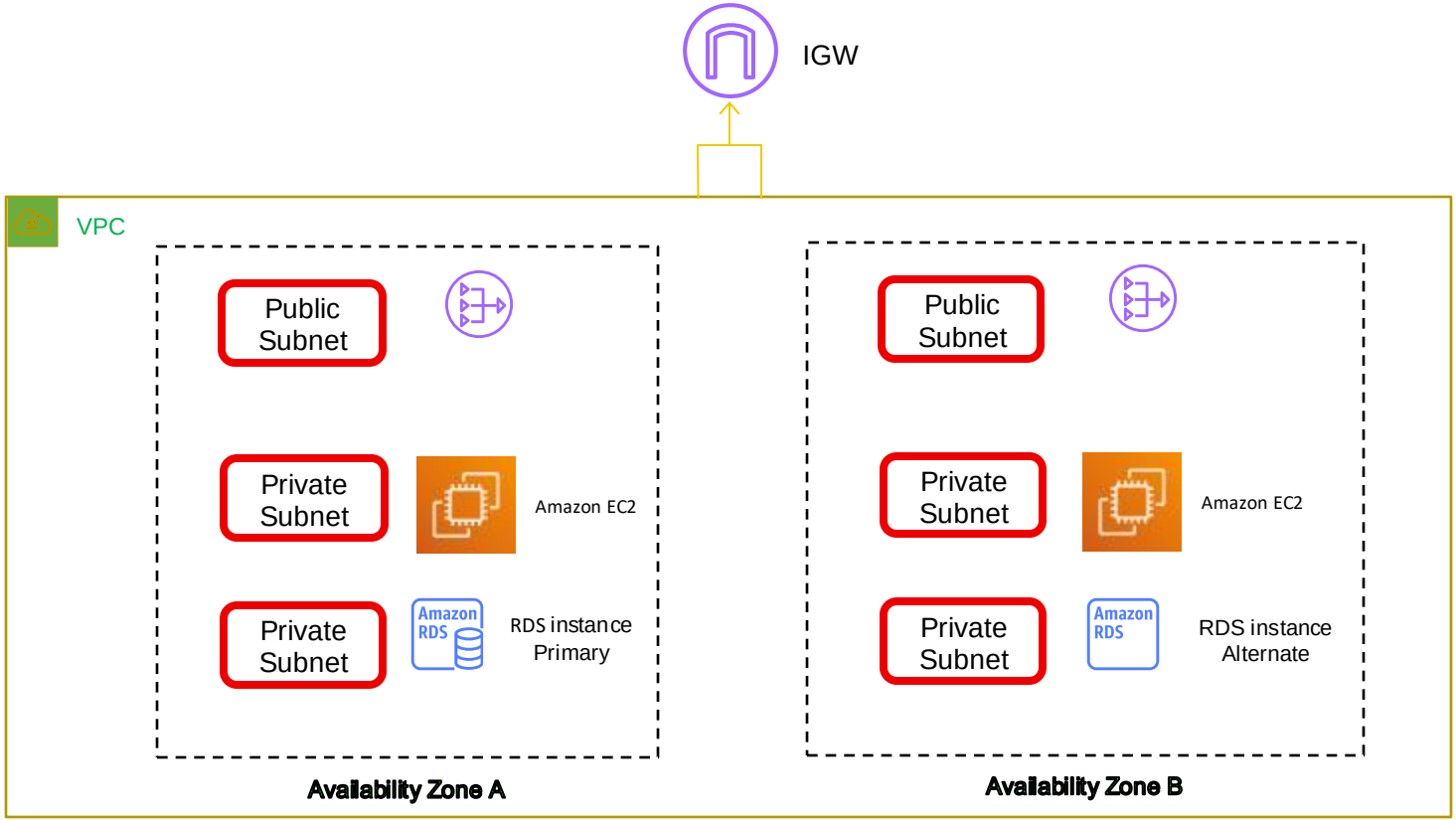
- Internet Gateway (IGW) – VPC side of the public Internet.
- VGW - VPN endpoint on AWS side of private connection.
- NAT Gateway - Private IPv4 instance indirect Internet access.
- Egress-only Internet Gateway – Public IPv6 protection from Internet access.
- Transit Gateway – Custom customer private routing paths.

Internet Gateway

- Allows direct communication between public subnet resources and the Internet.
- The Internet gateway is attached to the VPC.
- To enable Internet access, order and attach Internet Gateway to VPC.
- Add route table Internet gateway entry for the public subnet route table.

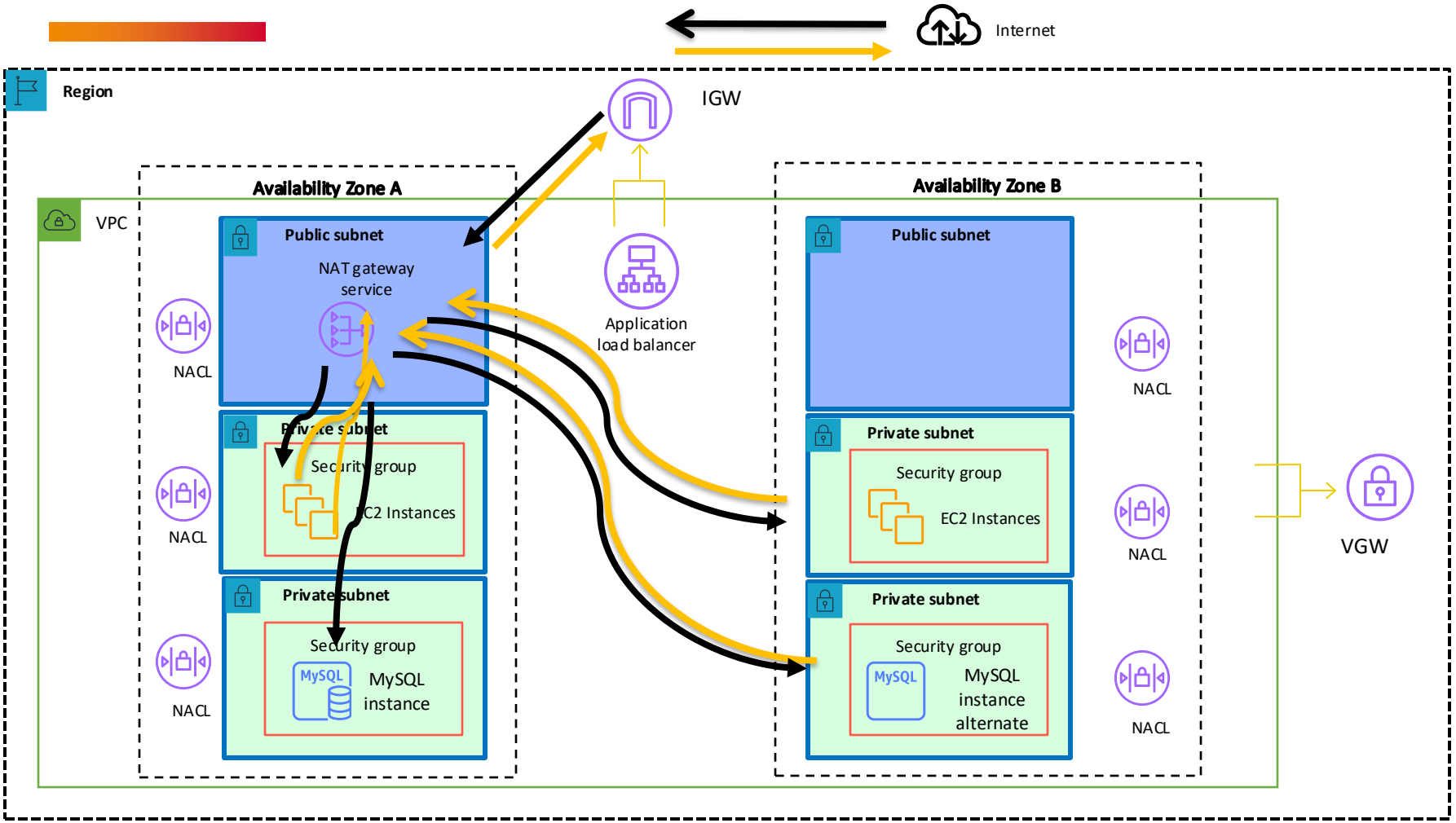


Two-tier Workload

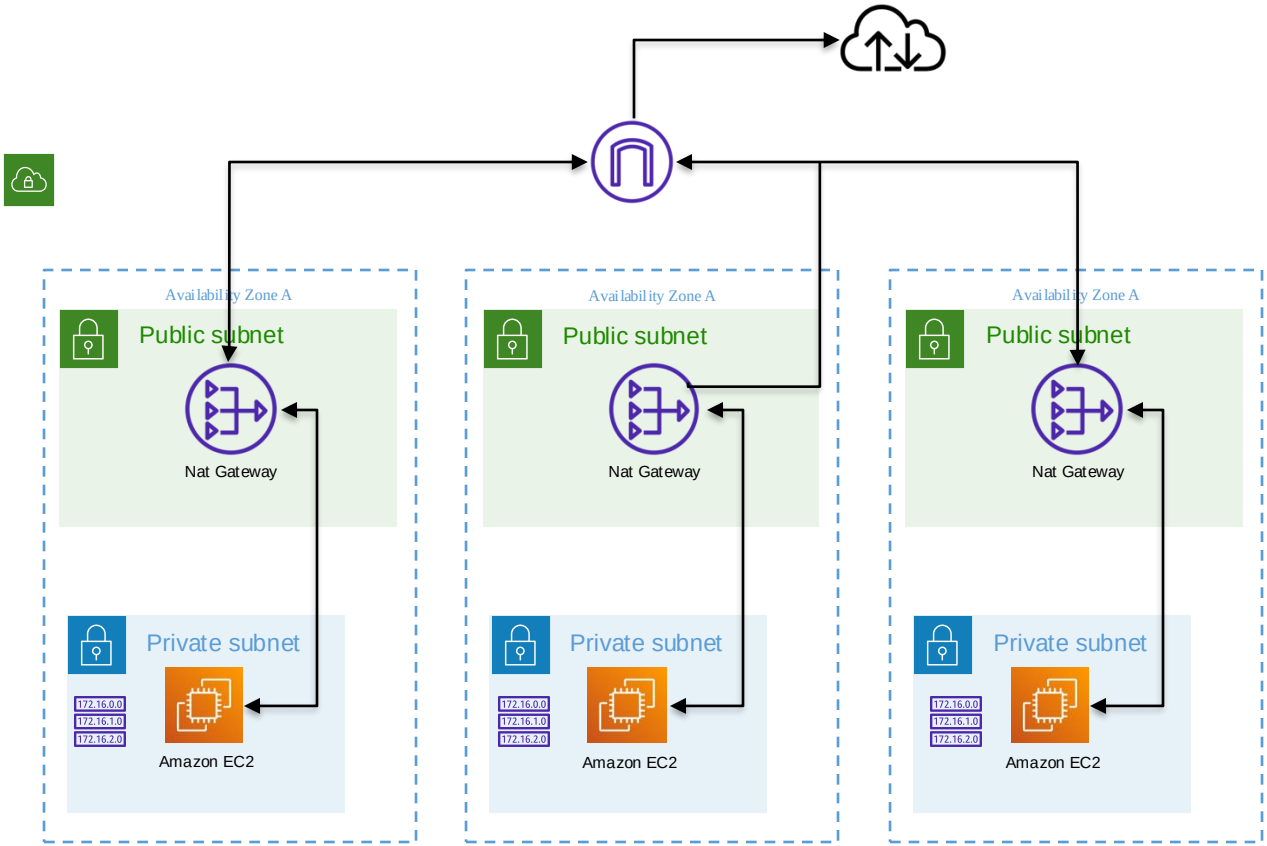


NAT Gateway Services

- NAT service enables instances in private subnets to connect to the Internet to get updates.
- Traffic requests are forwarded to the NAT service hosted in the public subnet.
- Internet response is sent back to the instance that made the request.
- NAT Options:
  - NAT gateway service provided by AWS
  - NAT instance – compute instance created with a NAT AMI

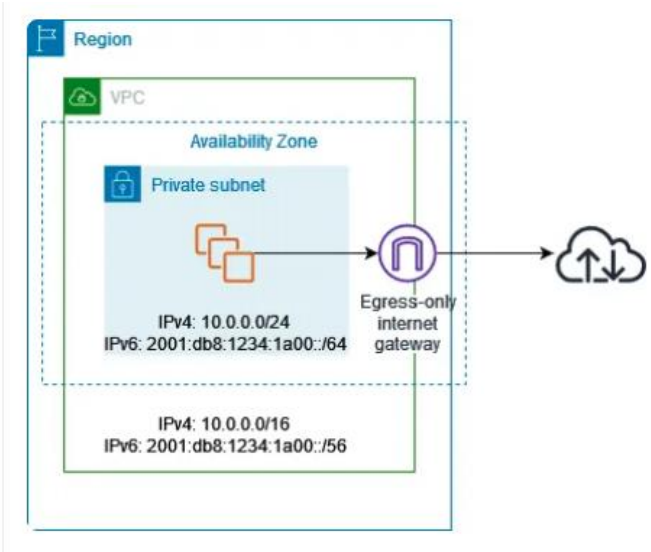

Internet
Region
IGW
Availability Zone A
Availability Zone B
VPC
Public subnet
NAT gateway service
Application load balancer
NACL
Private subnet
Security group
EC2 Instances
Private subnet
Security group
MySQL
MySQL instance
Public subnet
Private subnet
Security group
EC2 Instances
Private subnet
Security group
MySQL
MySQL instance alternate
NACL
VGW

NAT Gateway Service HA Deployment



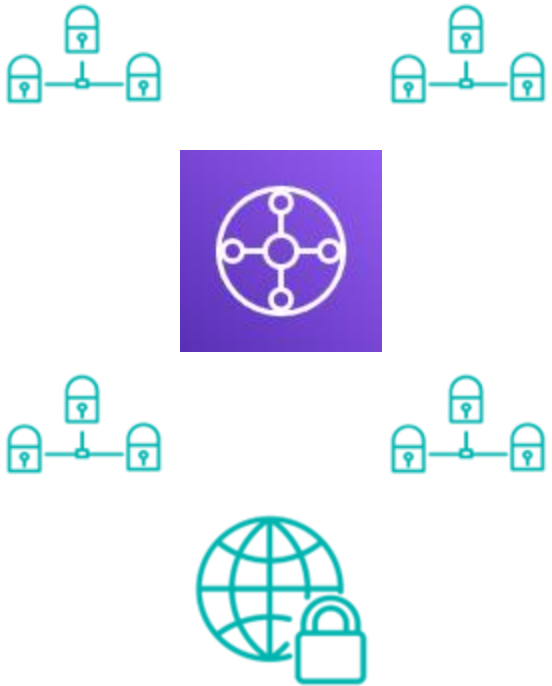
Egress-only Internet Gateway

- Control outbound Internet access for IPv6 EC2 Instances.
- Stops direct inbound access.
- IPv6 addresses are public.
- Forwards request to the Internet and send back the response.
- Egress-Only Internet Gateway **instead** of NAT gateway for IPv6 instances.

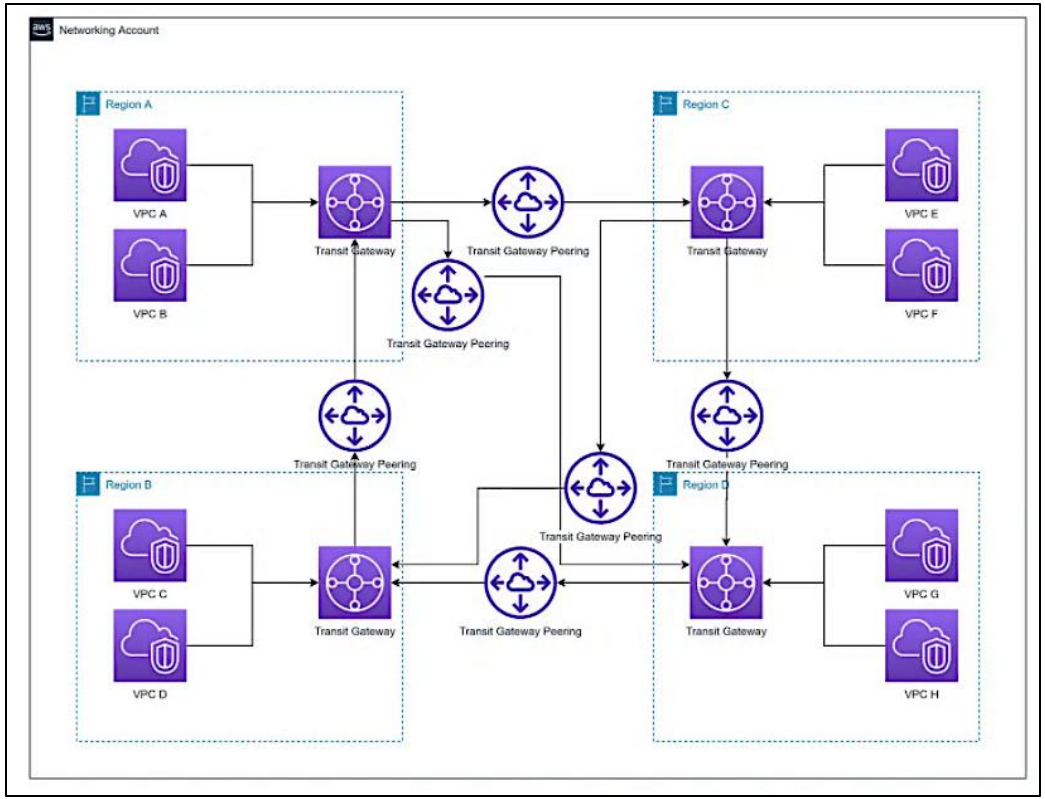


Transit Gateway

- Segment and isolate networks.
- Each transit gateway can connect to 5000 VPCs.
- VPCs connect to the Transit gateway using a hub and spoke model.
- Transit gateway traffic remains on AWS's private network.
- Custom transit gateway route tables allow you to segment and isolate your networks.
- Connect AWS regions using Transit Gateway Peering connections.

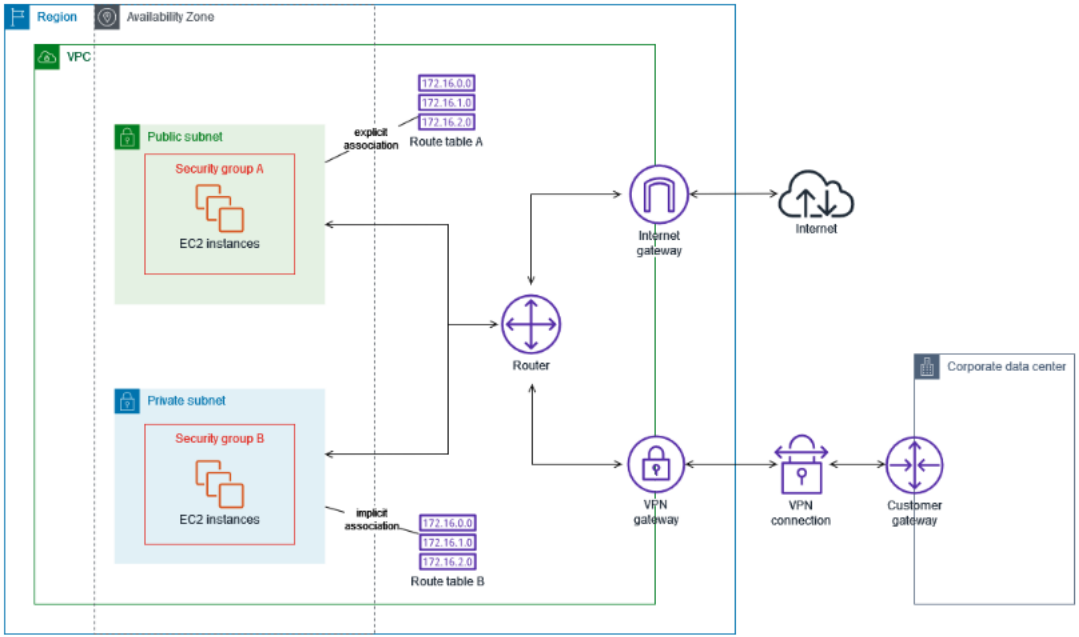


Transit Gateway Architecture



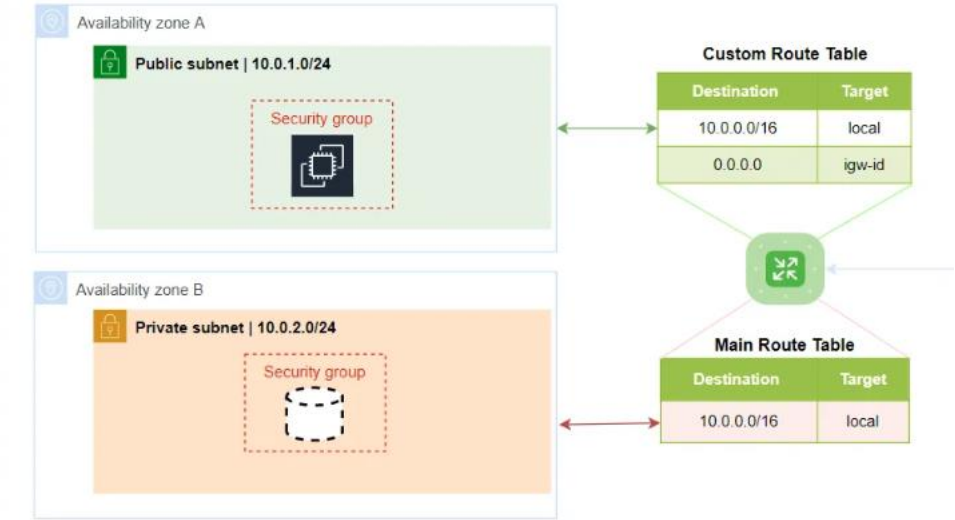
Route Tables

- Route table entries define network traffic patterns.
- Route tables connect subnets to an Internet gateway (IGW), a VPN gateway (VGW), or a NAT gateway.

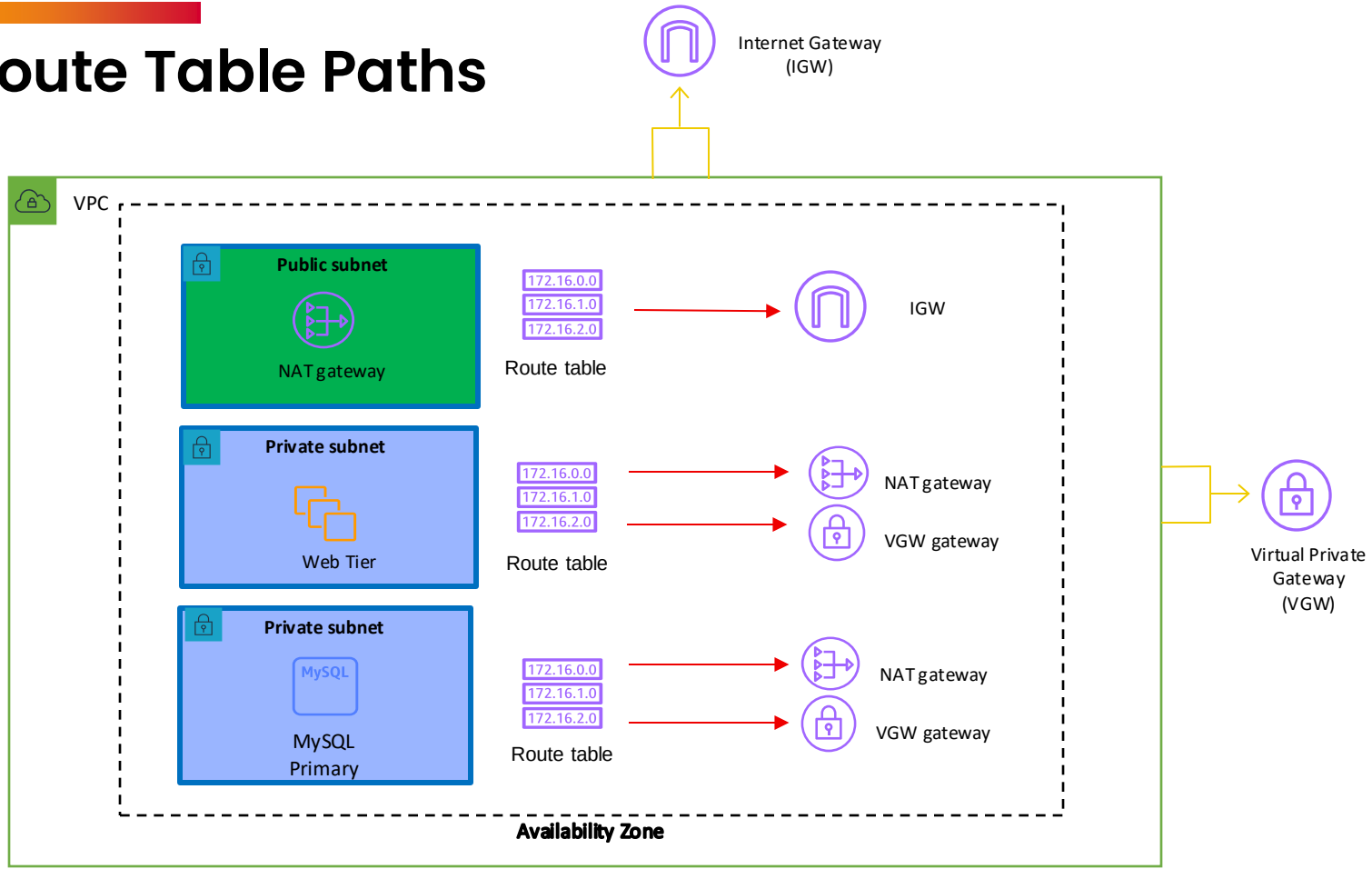


The Main Route Table

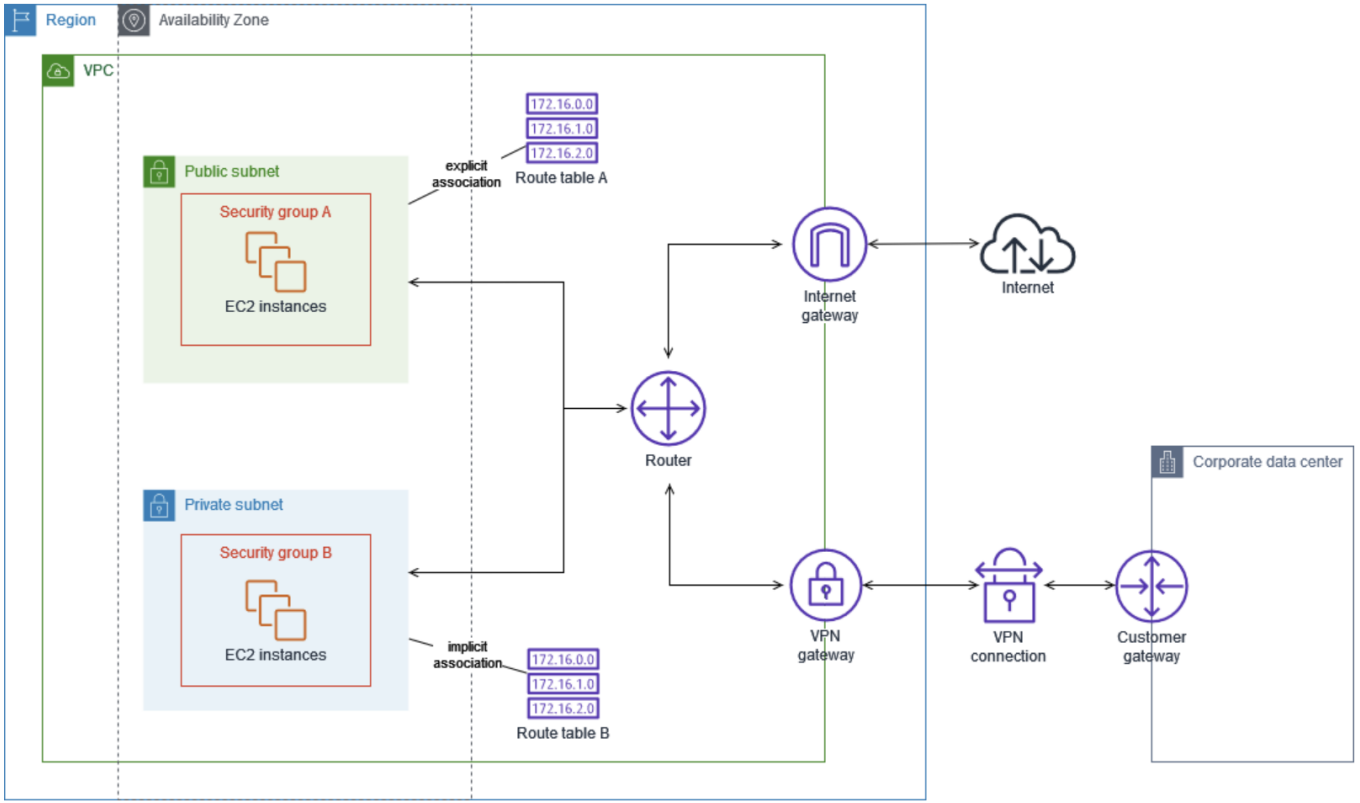
- The main route table is implicitly attached to subnets that are not explicitly associated with a custom route table.
- A custom route table is assigned to a custom subnet.
- Each custom route table has a local route entry.



Route Table Paths

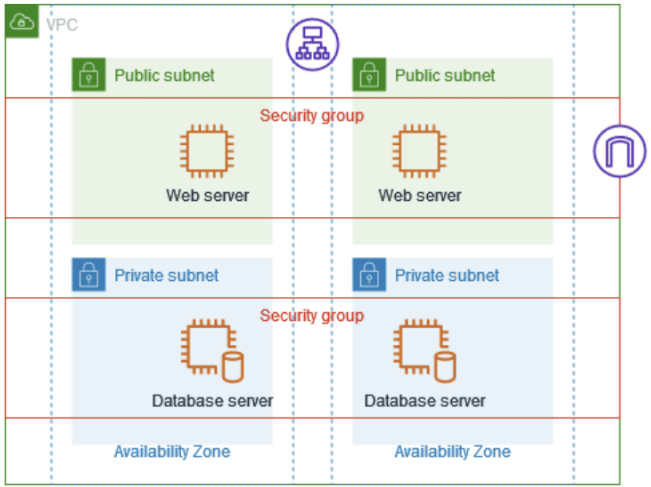


Route Tables

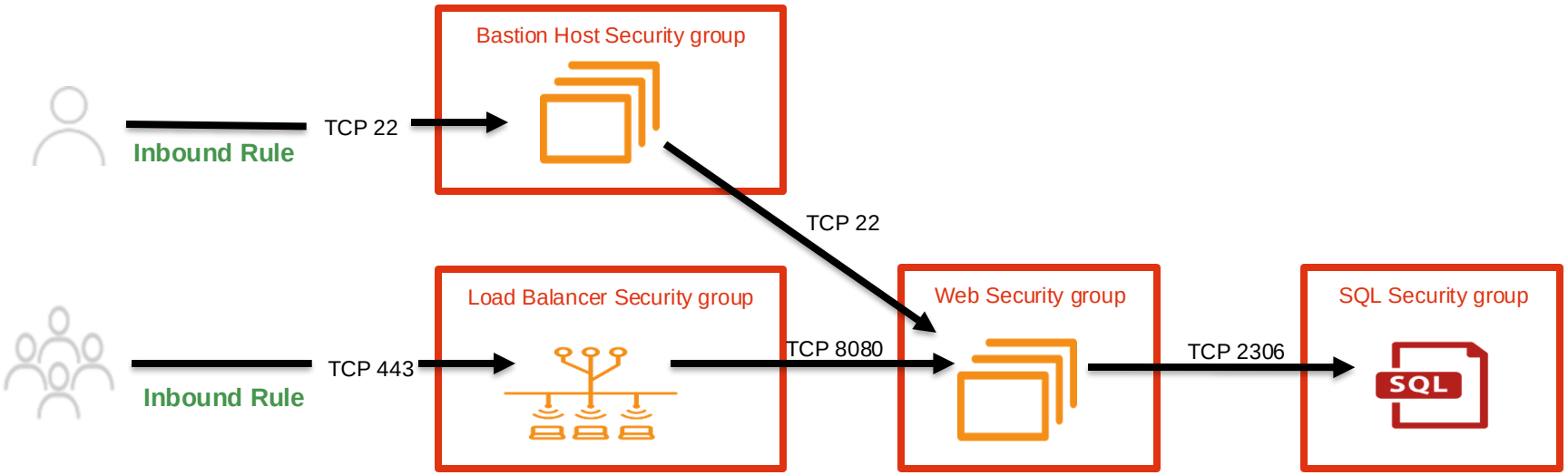


Security Groups

- Required EC2 instance firewall allows rules.
- Explicit deny rules can’t be specified.
- Inbound rules define the source of the traffic and the destination port, port range, or security group.
- A security group can be a source or destination for other security groups.

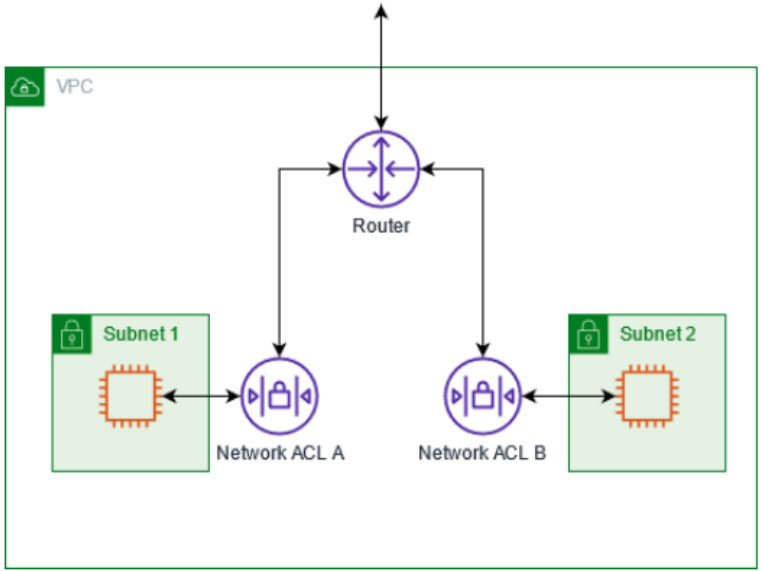


Security Group Design



Network Access Control List

- Nacls control traffic in and out of the subnet.
- Each subnet must be associated with a Nacl.
- Nacl rules are defined as stateless.
- Rules are evaluated in order.
- Use Nacls for blocking IP address ranges.



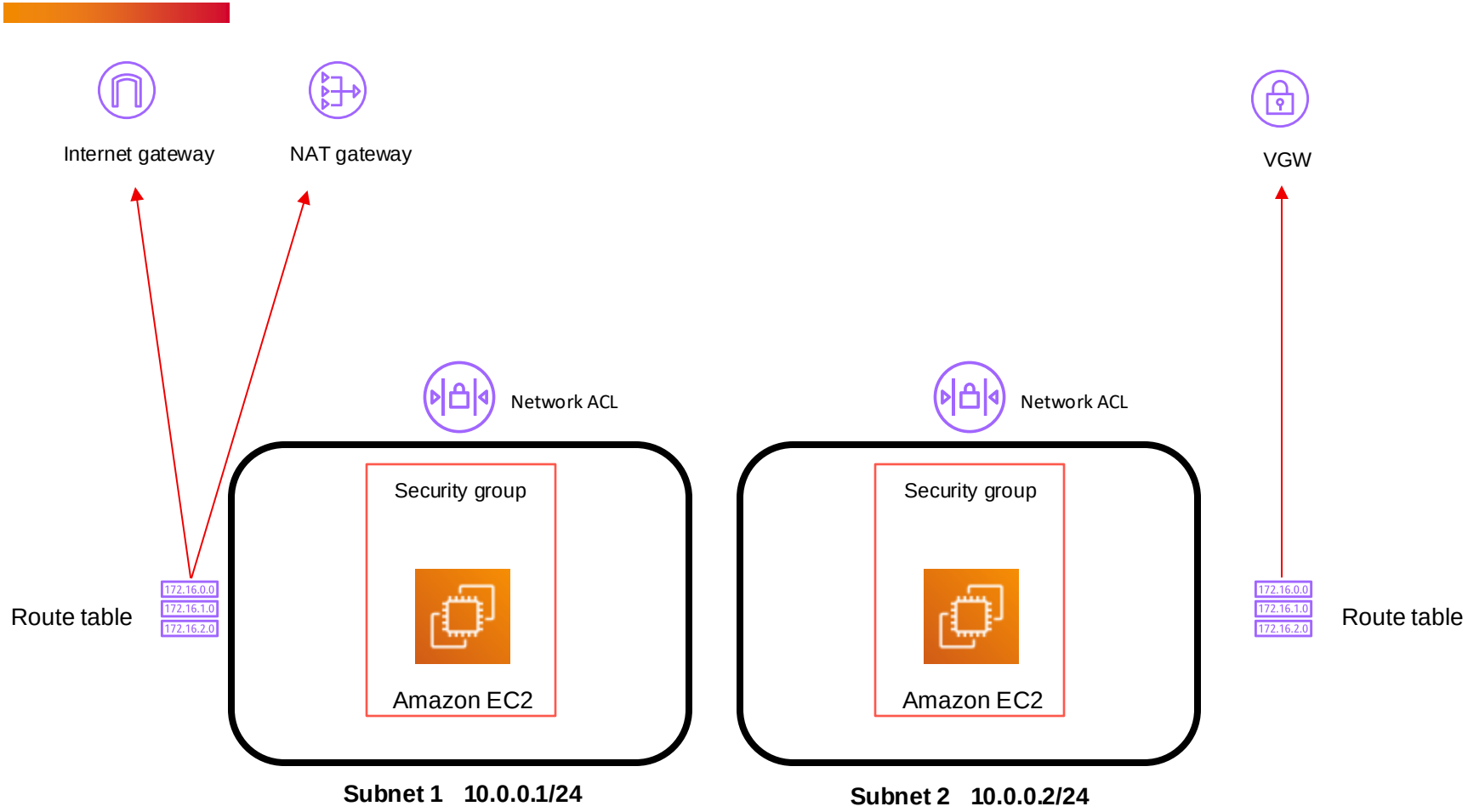
Network ACL Rules

- Inbound Rule

Rule #	Type	Protocol	Port range	Source	Allow/Deny	Comments
100	HTTP	TCP	80	0.0.0.0/0	ALLOW	Allows inbound HTTP traffic from any IPv4 address.
110	HTTPS	TCP	443	0.0.0.0/0	ALLOW	Allows inbound HTTPS traffic from any IPv4 address.
120	SSH	TCP	22	192.0.2.0/24	ALLOW	Allows inbound SSH traffic from your home network's public IPv4 address range (over the internet gateway).
130	RDP	TCP	3389	192.0.2.0/24	ALLOW	Allows inbound RDP traffic to the web servers from your home network's public IPv4 address range (over the internet gateway).
140	Custom TCP	TCP	32768-65535	0.0.0.0/0	ALLOW	Allows inbound return IPv4 traffic from the internet (that is, for requests that originate in the subnet). This range is an example only.
*	All traffic	All	All	0.0.0.0/0	DENY	Denies all inbound IPv4 traffic not already handled by a preceding rule (not modifiable).

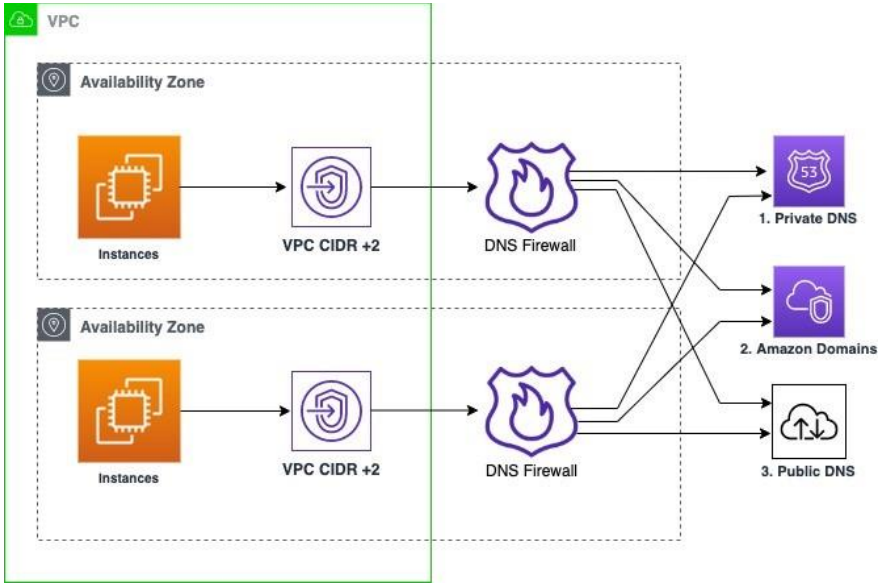
- Outbound Rule - Allow or deny the specified traffic pattern.

Rule #	Type	Protocol	Port range	Destination	Allow/Deny	Comments
100	HTTP	TCP	80	0.0.0.0/0	ALLOW	Allows outbound IPv4 HTTP traffic from the subnet to the internet.
110	HTTPS	TCP	443	0.0.0.0/0	ALLOW	Allows outbound IPv4 HTTPS traffic from the subnet to the internet.
120	SSH	TCP	1024-65535	192.0.2.0/24	ALLOW	Allows outbound return SSH traffic to your home network's public IPv4 address range (over the internet gateway).
140	Custom TCP	TCP	32768-65535	0.0.0.0/0	ALLOW	Allows outbound IPv4 responses to clients on the internet (for example, serving webpages to people visiting the web servers in the subnet). This range is an example only.
*	All traffic	All	All	0.0.0.0/0	DENY	Denies all outbound IPv4 traffic not already handled by a preceding rule (not modifiable).


Internet gateway
NAT gateway
VGW
Network ACL
Network ACL
Security group
Security group
Amazon EC2
Amazon EC2
Route table
172.16.0.0
172.16.1.0
172.16.2.0
172.16.0.0
172.16.1.0
172.16.2.0
Route table
Subnet 1 10.0.0.1/24
Subnet 2 10.0.0.2/24

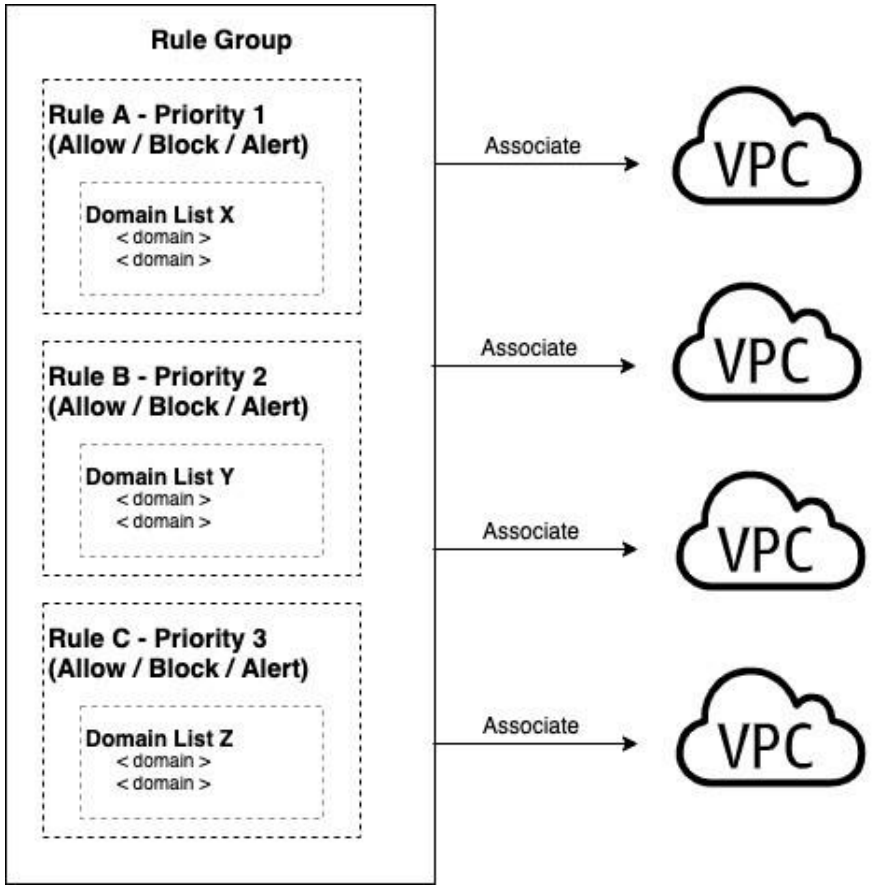
DNS Firewall

- Filter and regulate outbound DNS traffic for your virtual private cloud (VPC).
- Block DNS queries that are made for known malicious domains while allowing DNS queries to trusted domains.
- DNS Firewall consists of domain lists and rule groups associated with a VPC.



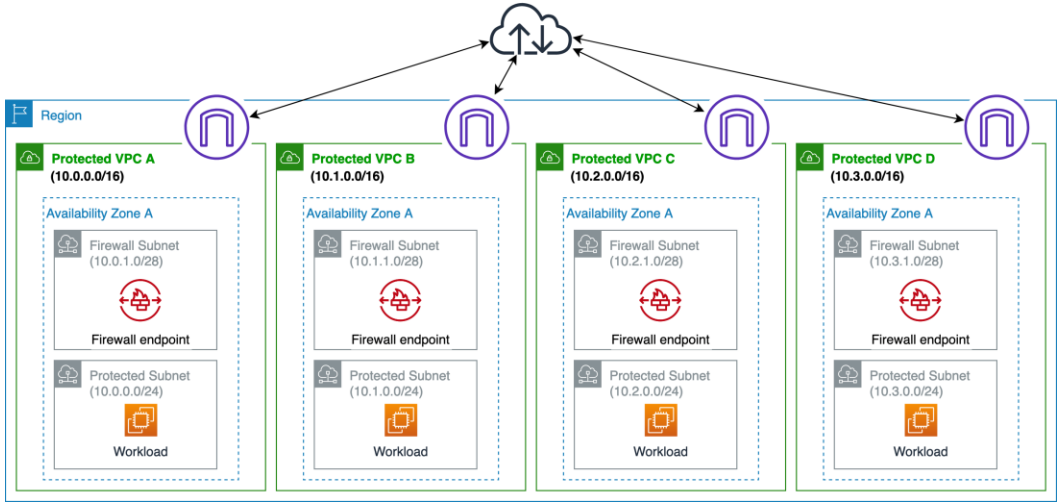
Rules and Rule Groups

- Rules - Single domain and action to take(Allow, Block, Alert).
- Domain List - Specify domains and actions to take(Allow, Block, Alert).
- Rule Groups - Collection of rules. Can be associated with multiple VPCs.
- Managed using AWS Firewall Manager (WAF, DNS Firewall)



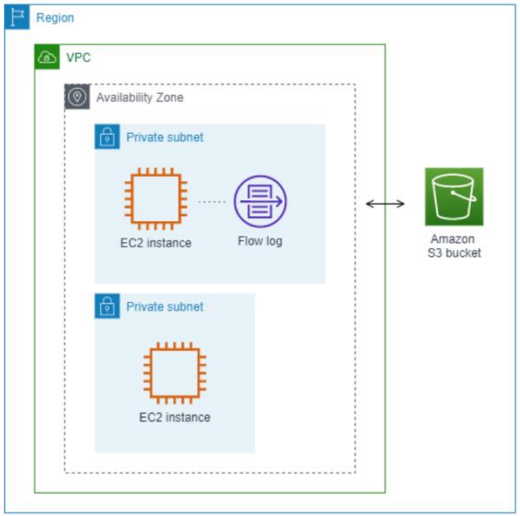
AWS Network Firewall

- Network Firewall provides filtering for both network and application layer traffic.
- For each Availability Zone, choose a subnet to host the firewall endpoint that filters your traffic.
- Deploy and manage stateful inspection, intrusion prevention and detection, and web filtering.



VPC Flow Logs

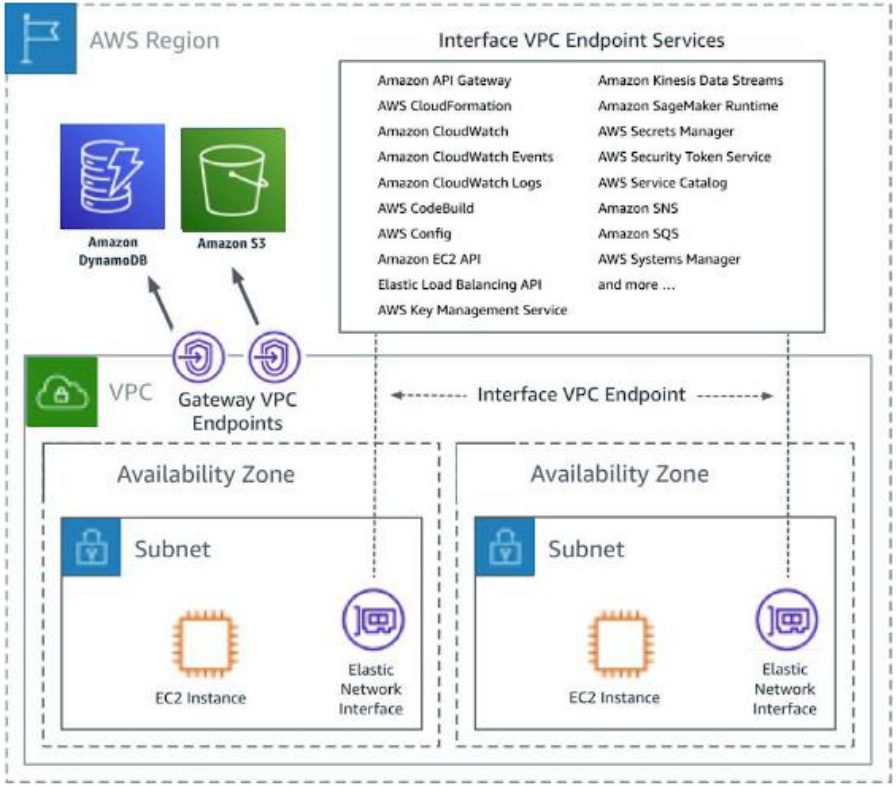
- Flow logs can be created for a VPC, a subnet, or a network interface.
- Logs IP traffic to and from network interfaces in a VPC (accepted/rejected).
- Each NIC has a unique log stream.
- Flow log data is published to a log group stored as a CloudWatch log group or S3 Bucket.



Troubleshooting Network Connectivity

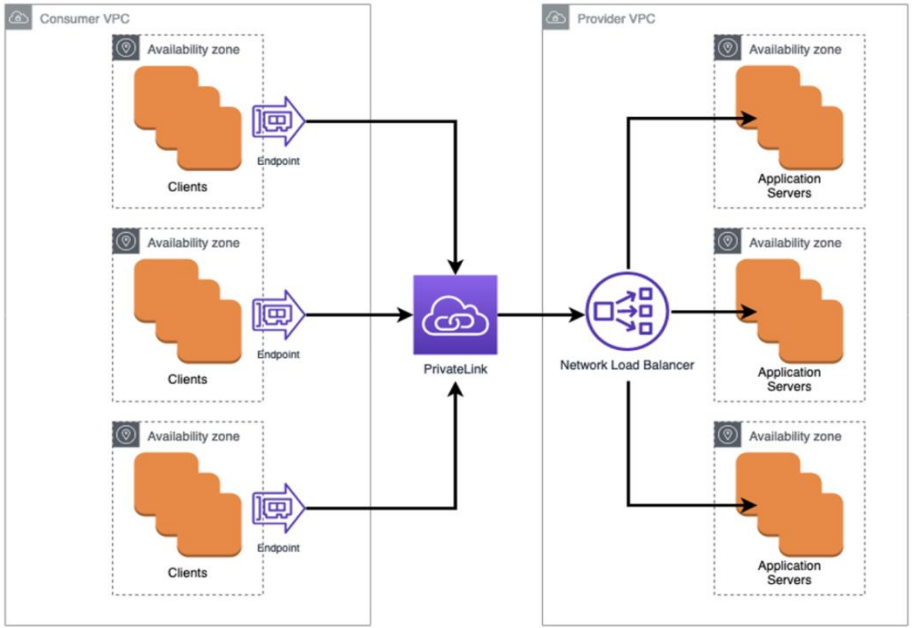
- VPC Flow Logs.
- VPC Reachability Analyzer (Traffic source, destination, protocol, and optionally a destination port).
- EC2 instance health - status checks.
- Subnet and route table configuration.
- Security group or network ACL configuration.
- Internet access or NAT gateway configuration.

VPC Endpoints



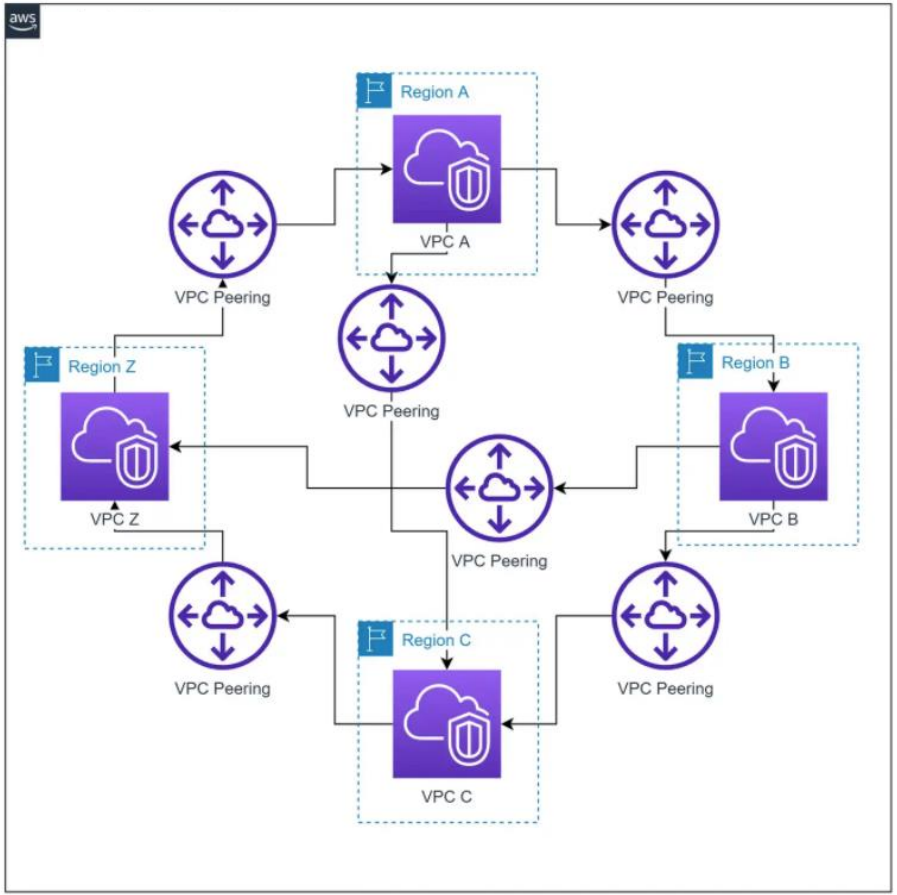
Endpoint Services / PrivateLink

- Provides private connectivity between VPC, and third-party services.



Peering VPC’s

- Networking connection between two VPCs.
- Peer your VPCs or between other account holders' VPCs using a private IP address.
- Peering is a one-to-one relationship.
- Peering connections are not transitive.
- CIDR blocks can’t overlap in a peering relationship.
- Peering connections can be created between VPCs in the same or different regions.

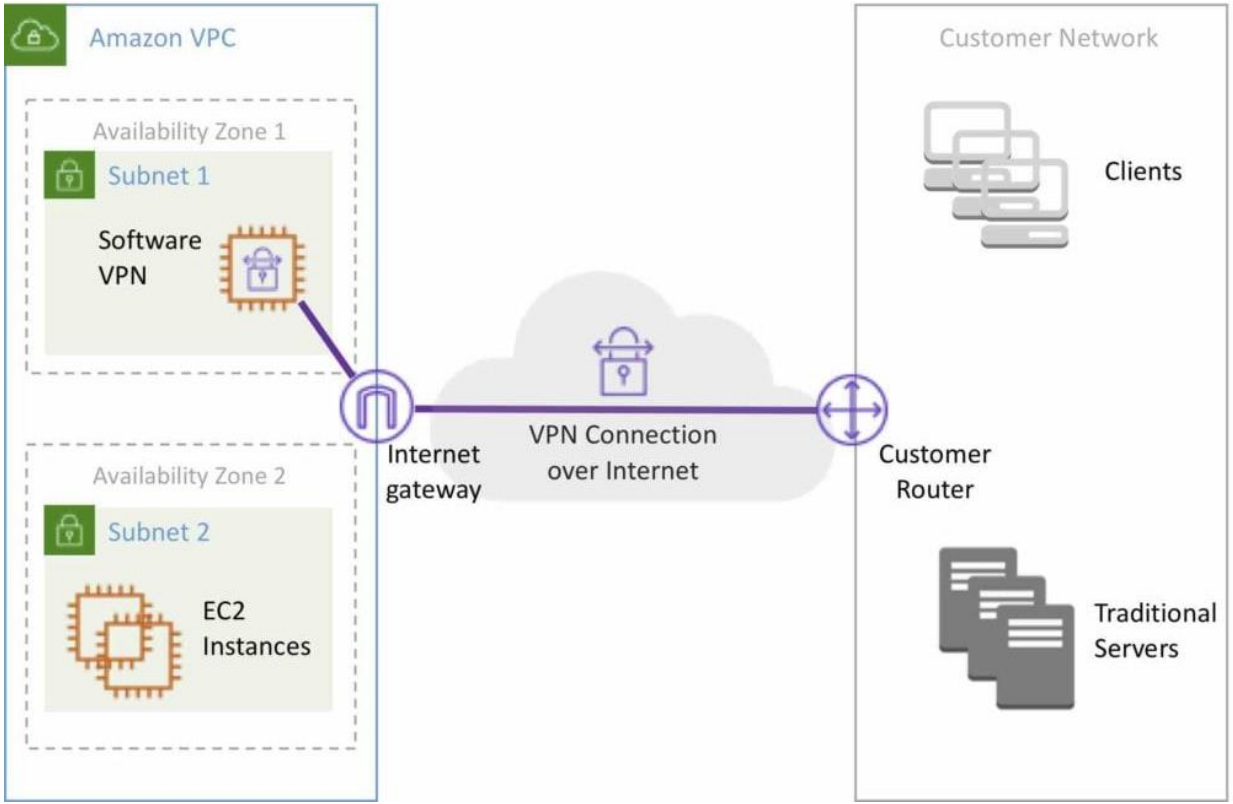

aws
Region A
VPC A
VPC Peering
VPC Peering
VPC Peering
Region Z
VPC Z
Region B
VPC B
VPC Peering
VPC Peering
VPC Peering
Region C
VPC C

External Connectivity

External Terminology

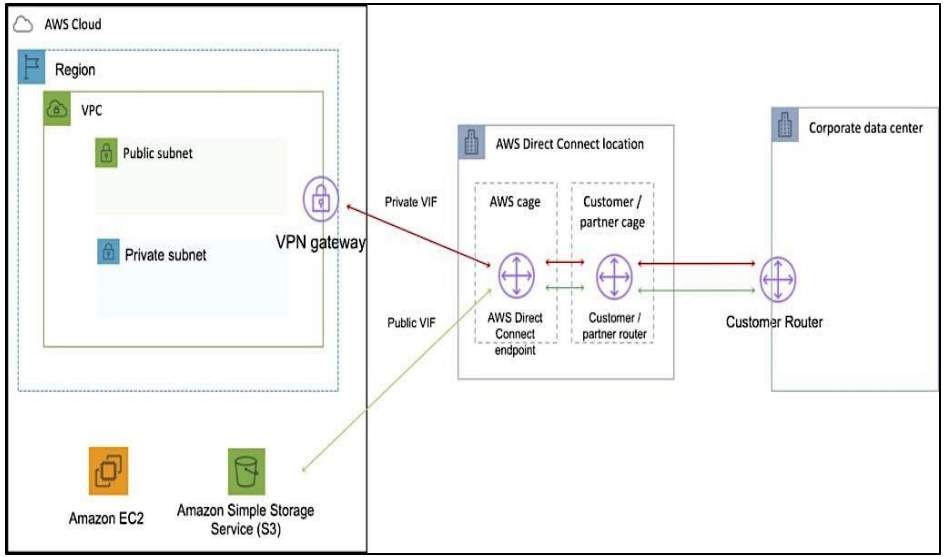
- The AWS side of the VPN tunnel is the Virtual Private Gateway (VGW).
- The customer gateway (CGW) is a hardware device or software application on the customer’s side of the VPN tunnel.
- The VPN tunnel is initiated from the CGW to the VPC.
- Static and dynamic routing (BGP) is supported.

Software VPN

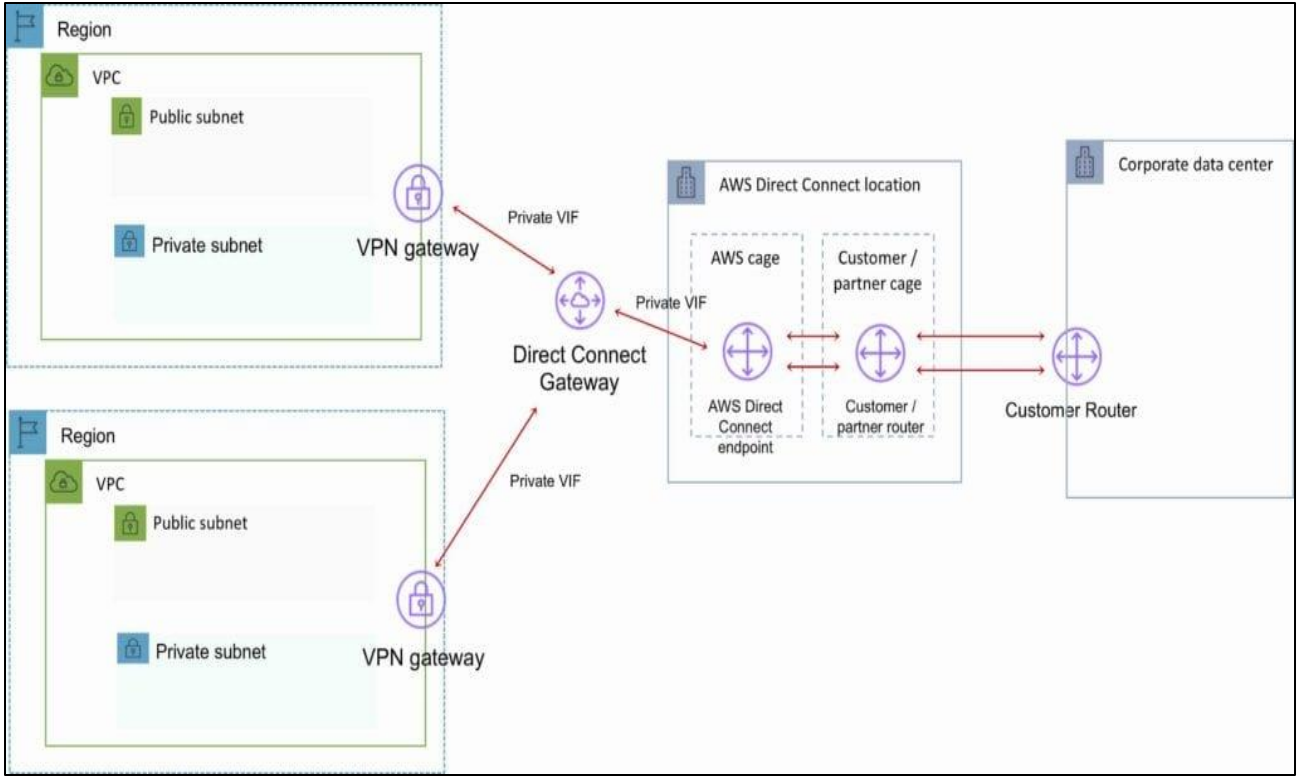


AWS Direct Connect

- Dedicated private fiber connection to AWS Cloud.
- Connection from company router to a Direct Connect router.
- Public virtual connections to AWS service.
- Private virtual interfaces connect to VPC.
- Direct Connect Gateways connect multiple VPCs to Direct Connect connection.



AWS Direct Connect Gateway



In Class Project

- Review the proposed VPC design and suggest improvements.

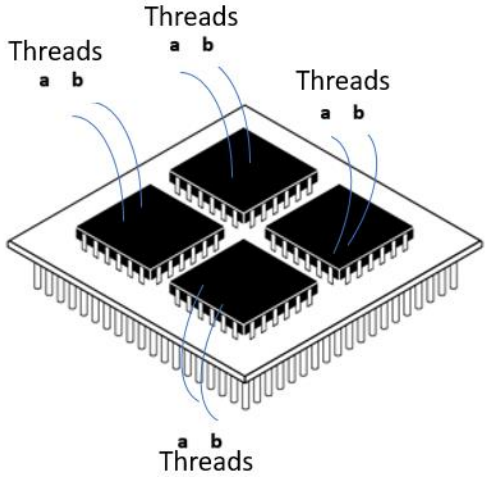
Compute Services

EC2 Instance FYI

- EC2 instances are members of compute families.
- For each instance's name, the first letter is the instance family and describes the resources allocated to the instance.
- EC2 resources (vCPUs, memory, storage, network bandwidth) are assigned to your account.

What’s a vCPU?

- AWS defines the amount of CPU power assigned to each instance as a virtual CPU (vCPU).
- vCPUs are software cores that “thread” to the physical CPU core on the host bare metal server.
- Hyper-threading associates two virtual threads to each physical core, working in a multitasking operating mode.



4 Core CPU / 8 Threads

EBS-optimized instances

- EBS–optimize instances have an optimized configuration stack with dedicated capacity for Amazon EBS.
- EBS–optimize instances utilize 500 Mbps and 14,000 Mbps dedicated bandwidth to Amazon EBS.
    - General purpose SSD volumes: within 10% of baseline and 99% of their burst performance over a given year.
    - Provisioned IOPS SSD volumes: within 10% of baseline and 99.9% of their provisioned performance over a given year.

Non-Volatile Memory Express

- SSD instance flash storage volumes.
- EBS volumes are exposed as NVMe block devices on Nitro-based instances.

Instance Name	vCPUs	RAM	Local Storage	EBS-Optimized Bandwidth	Network Bandwidth
m5d.large	2	8 GiB	1 x 75 GB NVMe SSD	Up to 2.120 Gbps	Up to 10 Gbps
m5d.xlarge	4	16 GiB	1 x 150 GB NVMe SSD	Up to 2.120 Gbps	Up to 10 Gbps
m5d.2xlarge	8	32 GiB	1 x 300 GB NVMe SSD	Up to 2.120 Gbps	Up to 10 Gbps
m5d.4xlarge	16	64 GiB	2 x 300 GB NVMe SSD	2.210 Gbps	Up to 10 Gbps
m5d.12xlarge	48	192 GiB	2 x 900 GB NVMe SSD	5.0 Gbps	10 Gbps
m5d.24xlarge	96	384 GiB	4 x 900 GB NVMe SSD	10.0 Gbps	25 Gbps

Dedicated Hosts

- Physical servers with EC2 instance capacity fully dedicated for customer use.
- Allows using existing per-socket, per-core, or per-VM software licenses (Windows Server, SQL Server, and SUSE Linux Enterprise Server) to meet compliance requirements.
- Visibility into the number of sockets and physical cores on a host, aiding in software licensing and compliance.
- AWS License Manager integration helps in managing software licenses and license usage.

Dedicated Instance

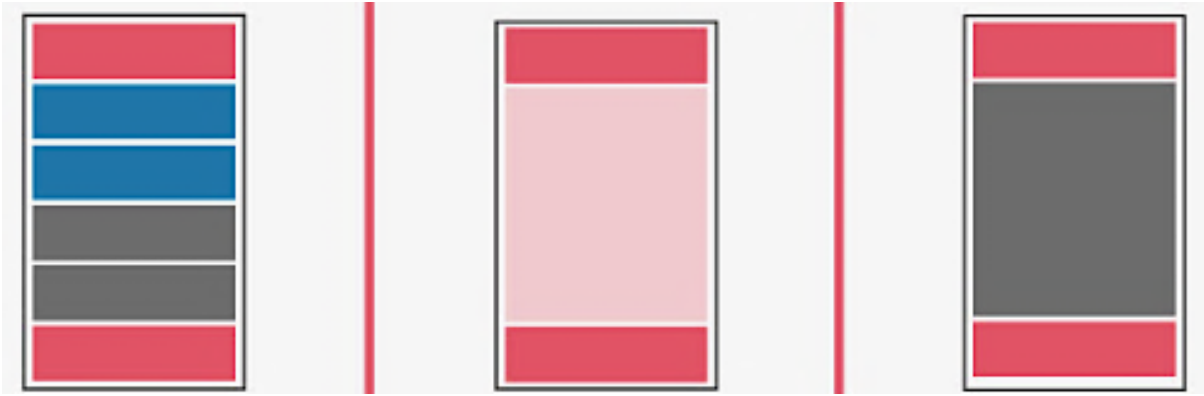
- Dedicated Instances are **physically isolated at the host hardware level** from other dedicated instances and instances in other AWS accounts.
- Compliance with specific regulatory requirements that don’t permit multi-tenant virtualization.
- When you launch a Dedicated Instance backed by EBS volumes, the EBS volume itself doesn't run on single-tenant hardware.

EC2 Purchase Options

No host exposure.

Pay for the host; there are no instance charges.

Extra charges for instances on dedicated hardware.



EC2 Host
Default / Shared

Dedicated Host

Dedicated Instance

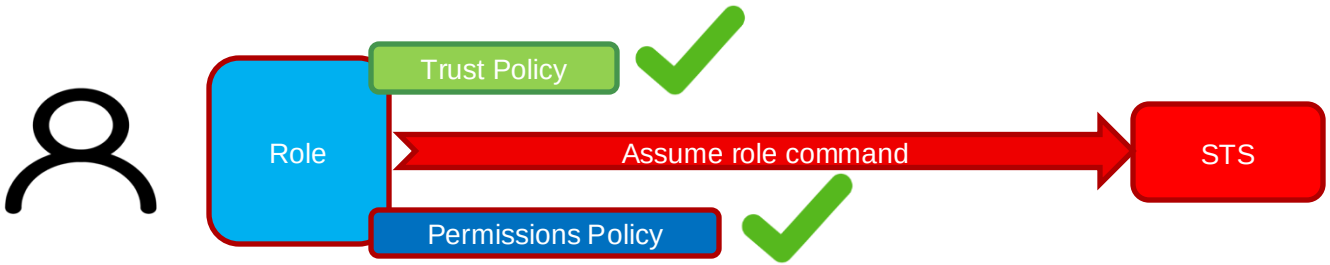
IAM Roles with EC2 Instances

IAM Roles with EC2 Instances

- Simplify management and deployment of access keys to applications hosted on EC2 instances.
- EC2 instances use EC2 profiles to store IAM Roles.
- EC2 instance-secured memory locations hold the temporary security credentials used by the application.
- Applications that run on the EC2 instance retrieve temporary security credentials from the EC2 instance metadata.

Role Basics

- The user/application must also have been assigned a trust policy with the permission to run the Assume Role command.
- The Assume Role command is executed at the Secure Token Service.
- The trust policy is attached to the role.



STS Operation

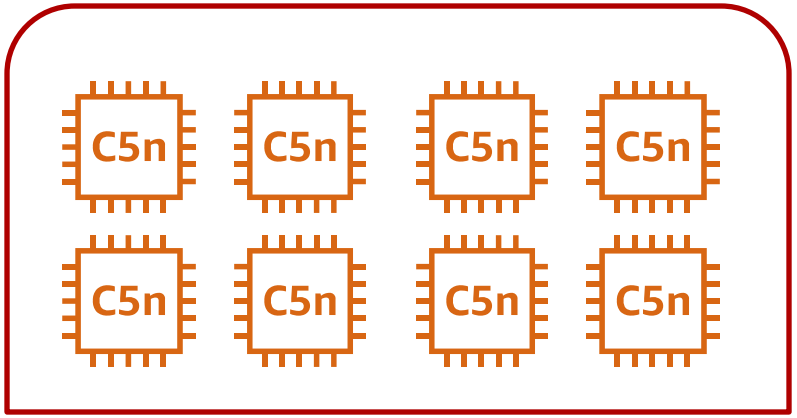
- By default, the AWS Security Token Service (AWS STS) is available as a global service; STS requests go to a single endpoint at https://sts.amazonaws.com.
- Regional AWS STS endpoints can also be chosen for faster access speeds.

Security token lifetime can be set from 1 to 36 hours

Placement Groups

Cluster Placement Group

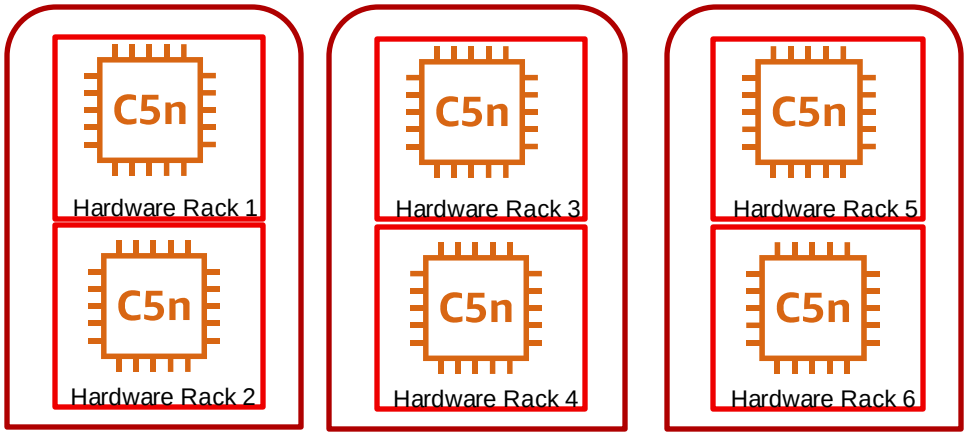
- A cluster design has the lowest latency possible within a single AZ.
- EC2 instances are placed on the same underlying hardware, which minimizes the network latency between them.
- Use case: Big Data or HPC.



Same Hardware Rack

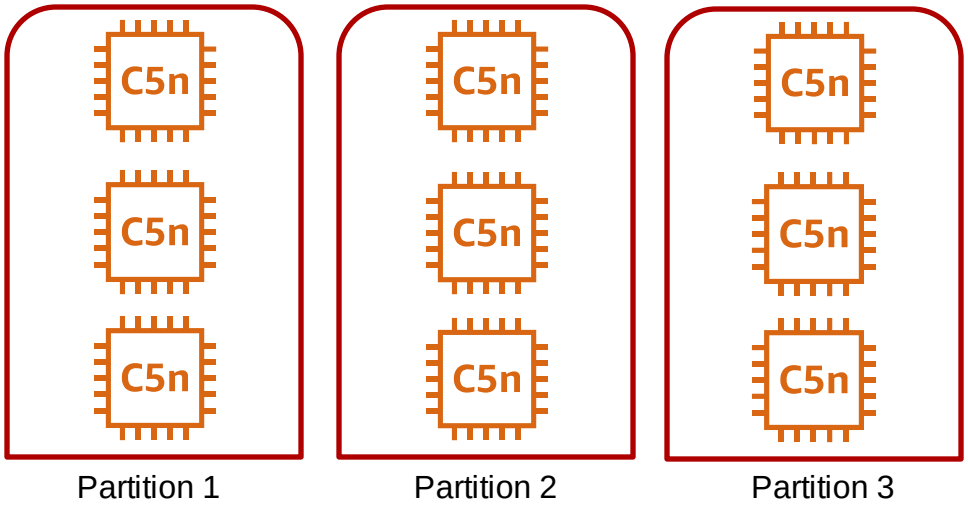
Spread Placement Group

- Spread your instances across distinct underlying hardware to reduce the risk of simultaneous hardware failures.
- Racks have separate power sources.



Partition Placement Group

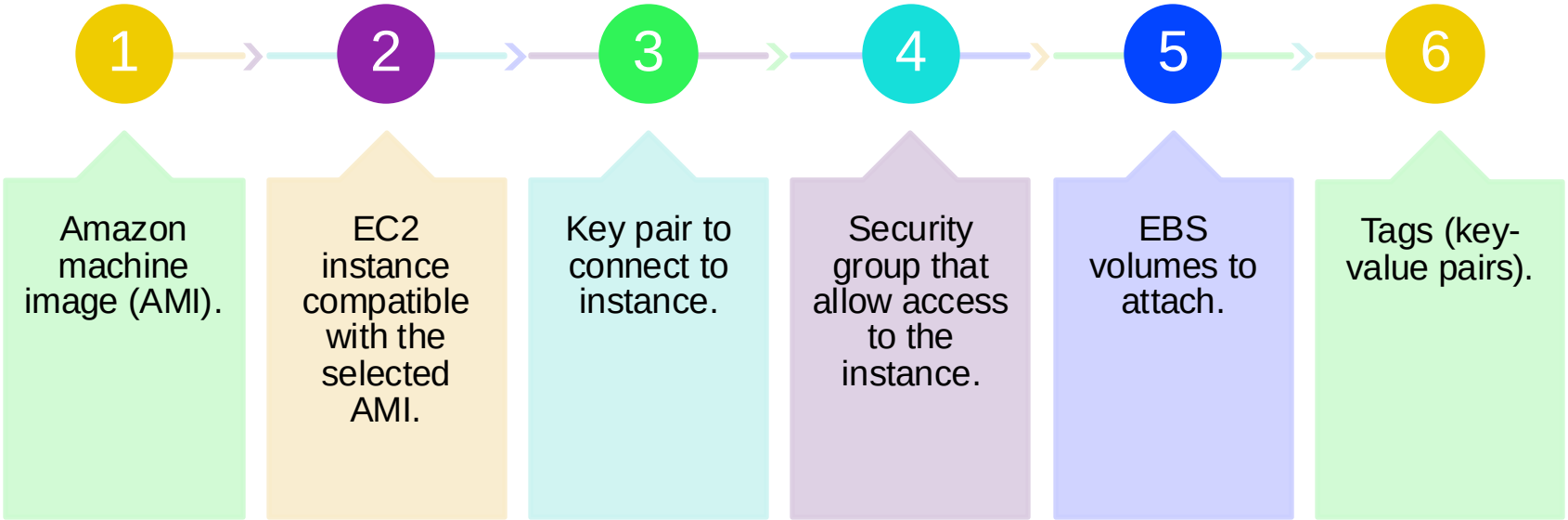
- Create multiple “partitions” of nodes with separate racks.
- No two partitions within a placement group share the same rack.
- Distributed, fault-tolerant, and low-latency workloads.



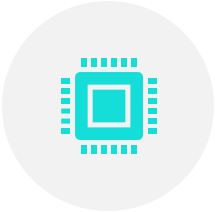
AMI components

- Volume: Describes the instance boot and data volumes.
- Launch permissions: Define the AWS accounts permitted to use the AMI to launch the instances.
- Encryption - Custom AMIs can be encrypted using KMS keys.
- Volumes to attach: Volumes to attach are contained in a block device mapping document.
- Default region: AMIs are region-specific but can be manually copied.

Launch Templates



EC2 Image Builder



Automate the creation, maintenance, and deployment of Linux or Windows images.

Images can be generated based on source AMIs.

Define collections of security settings that you can edit, update, and use to harden your images.

AWS-provided tests and customer tests can be run before image finalization.

EC2 Auto Recovery

- A CloudWatch alarm can monitor and automatically recover the instance if it becomes impaired due to an underlying hardware failure or a problem.
    - StatusCheckFailed
    - Network connectivity (System power loss, software or hardware issues on the physical host.)
- The recovered instance is identical to the original instance, including the instance ID, private and public IP addresses, and instance metadata.

EC2 Instance Pricing

EC2 Instances: Pricing Options

On-Demand

Pay by per hour/ second

Short-term, unpredictable workloads

Reserved Instances

Discount for 1 - 3-year commitment

Applications with consistent usage

Spot Requests

Spare AWS capacity > 90% discount

Applications with flexible start and end times

On-demand EC2 Instances

There are over 200 EC2 instance types.

On-demand instances each AWS account can are defined by a default vCPU quota.

Service Quotas can be increased upon request.

Reserved EC2 Instance Pricing (RI)

RI is a billing discount for on-demand EC2 instances.

When a running EC2 instance matches a reserved instance pricing agreement, the RI price is charged.

The number of regional reserved instances that you can purchase depends on your current quota limit(s).

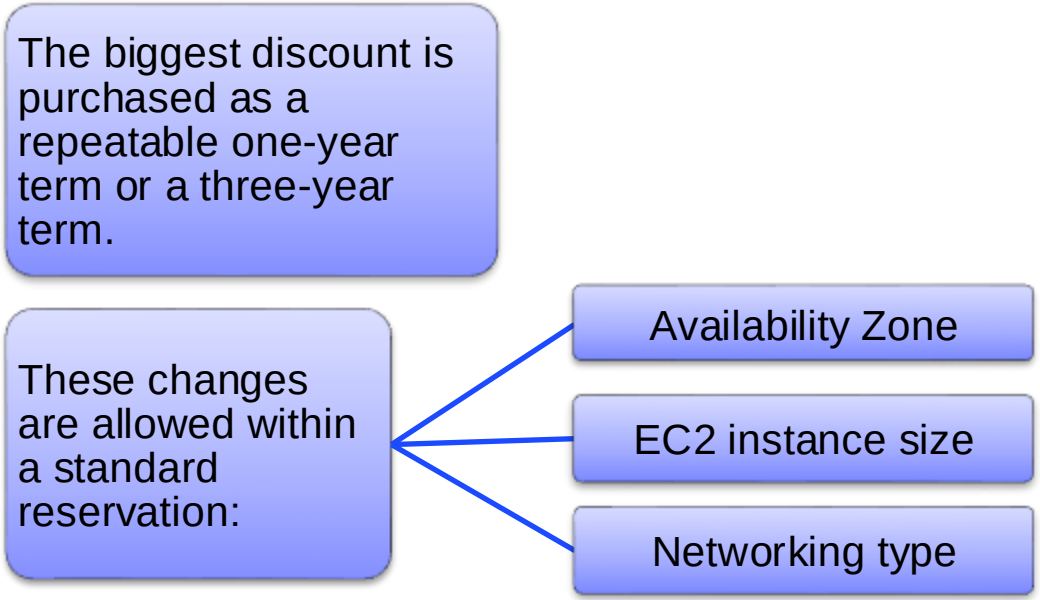
Reserved instances quota limits are based on the region and the number of availability zones.

Reserved instances are defined as regional or zonal.

Regional and Zonal Limits

	Regional Reserved	Zonal Reserved
Reserved capacity	Does not reserve capacity.	Reserves capacity in the specified AZ.
Flexibility	Instance usage in any AZ in the specified Region.	Instance usage in the specified AZ only.
Instance size	Applies to instance usage within the instance family regardless of size.	Instance discount applies to instance usage for the specified instance type and size only.
Tenancy	Default	Default

Standard Reservation



Convertible Reservation

Change EC2 instance types and operating systems, or switch from multi-tenancy to single-tenancy compute operation.

The convertible reserved discount could be over 50%; and the term can be a one, or three-year term.

Spot Instance

A spot instance is spare compute capacity that AWS is not currently using.

Save up to 90% of the purchase price; when AWS wants your instance back, a two-minute warning is provided (CloudWatch alert).

Spot instance pricing is based on supply and demand; the instance will run until the EC2 service needs the capacity back.

To counteract this possibility, you can define a maximum spot price that you're willing to pay.

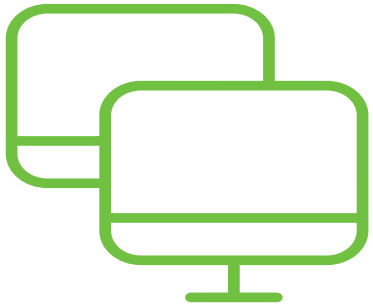
On-Demand Capacity Reservation

Guarantee that on-demand instances are always available for use in a specific availability zone.

After you've created a capacity reservation, you will be charged for the capacity reservation whether you run the instances or not.

Spot Fleets

- Spot Fleet is a set of Spot instances and optionally On-Demand instances.
- Launches the number of Spot and On-Demand Instances to meet the target capacity specified in the Spot Fleet request.
- Spot Fleet can launch a replacement Spot instance.
- EC2 emits a rebalance recommendation that a Spot Instance is at an elevated risk of interruption.
- Spot Fleet choices: request and maintain.
    - A one-time request for your desired capacity.
    - Maintain a target capacity over time.



Target capacity

Total target capacity

Set your total target capacity (number of instances or vCPUs) to launch. If you specified a launch template, you can allocate part of the target capacity as On-Demand. The number of On-Demand Instances always persists, while Spot Instances can be scaled.

50 | instances

☑ Include On-Demand base capacity

Allocate part of target capacity as On-Demand instances

25 instances

☑ Maintain target capacity

Automatically replace interrupted Spot Instances

Interruption behavior

Terminate

☑ Capacity rebalance

When a rebalance notification is sent to a Spot Instance, Spot Fleet automatically attempts to replace the instance before it is interrupted. Learn More

Instance replacement strategy

Launch only

Fleet only launches a replacement instance and will not terminate the instance that receives the rebalance recommendation. You can t…

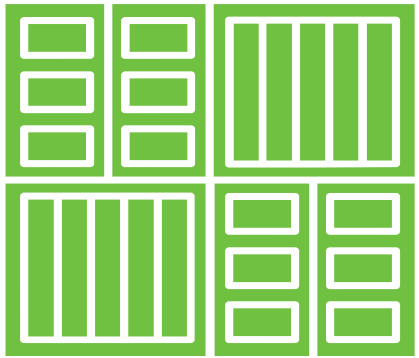
Savings Plans

- Savings Plans provide up to 72% savings on your AWS compute usage.
  - Compute: Applies to all Amazon EC2 instances (OS, tenancy or Region), Fargate and Lambda usage.
  - EC2 Instances.
  - Amazon Sage Maker.
- Commit to using a specific amount of compute power (measured in $/hour) for one or three years.

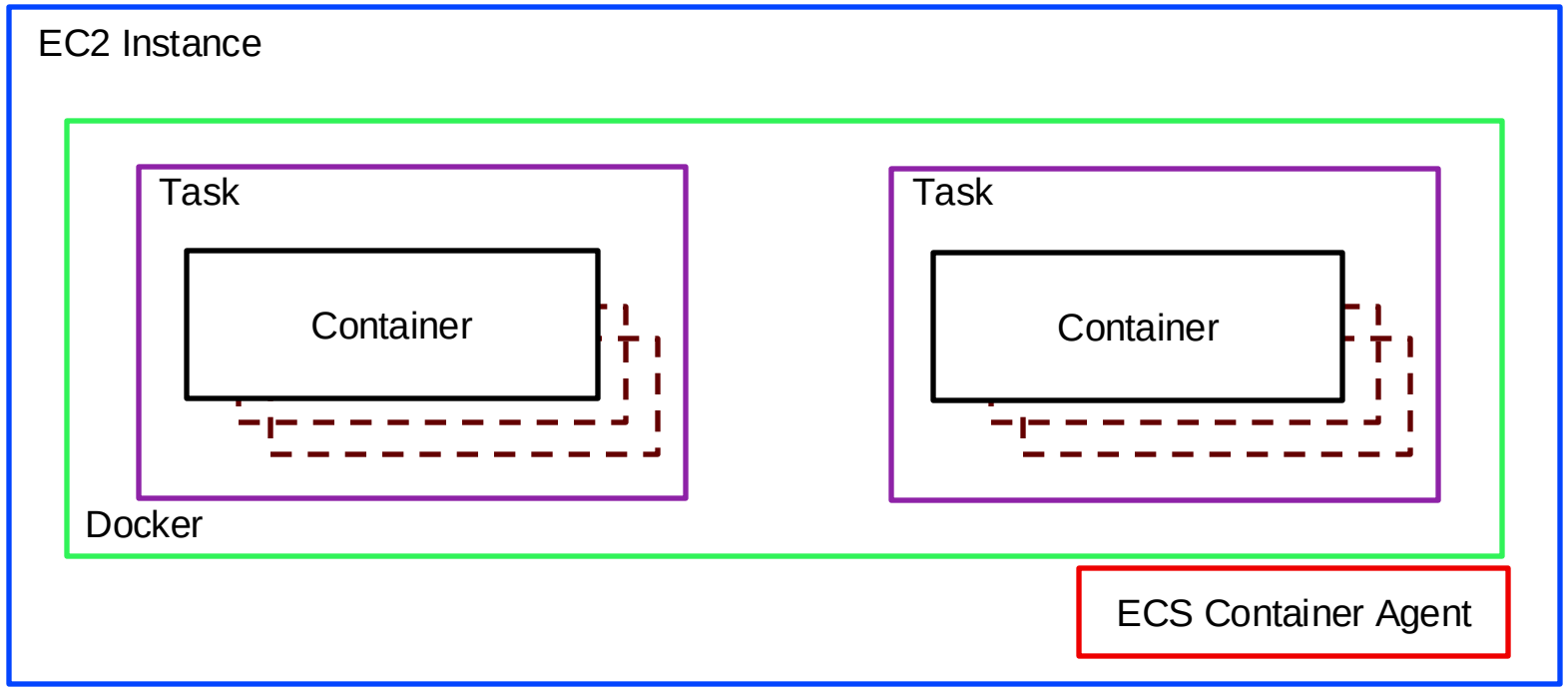
Containers

Container Options at AWS

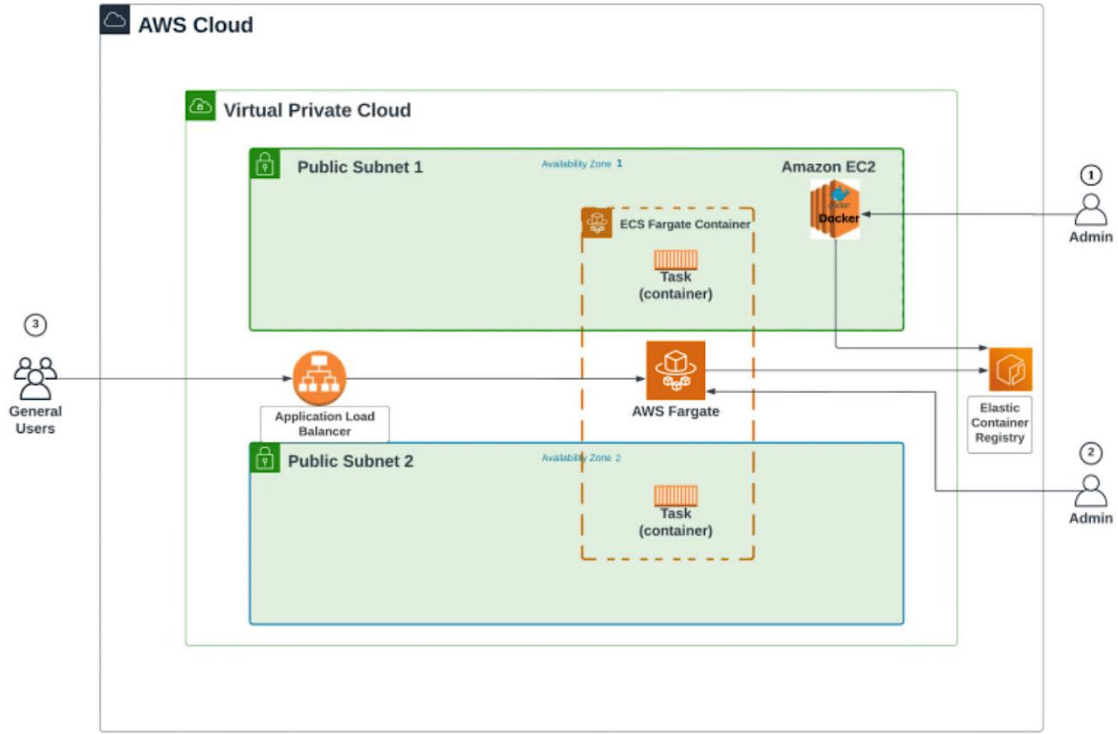
- Elastic Container Service: Run Docker containers: applications or micro-services
- Amazon Elastic Kubernetes Service: Manage containers with Kubernetes.
- AWS Fargate: Orchestration service for Docker or Kubernetes container deployments.



Task Definition Logic

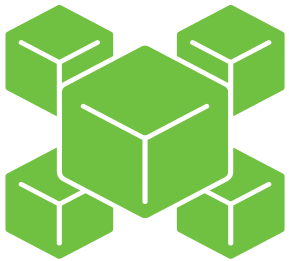


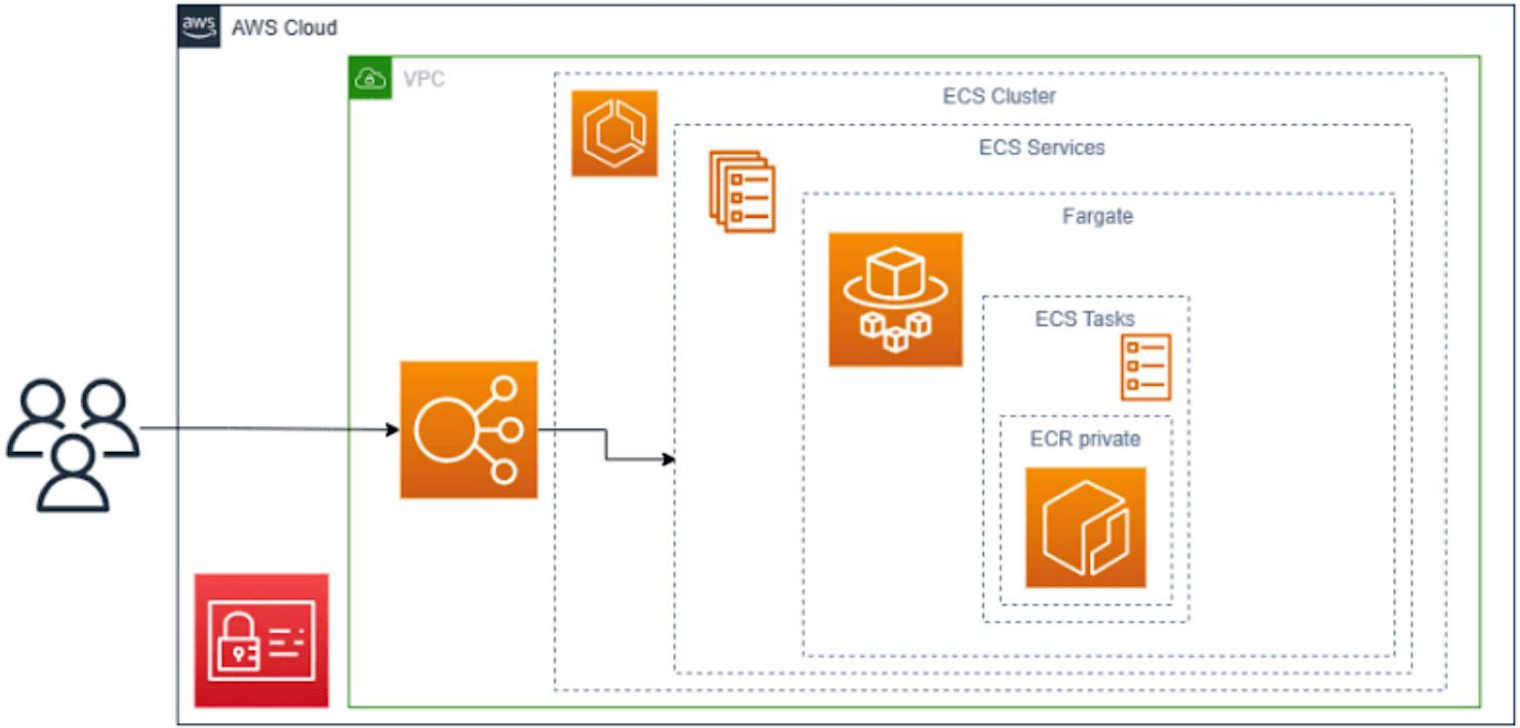
Amazon ECS (Elastic Container Service)



AWS Fargate

- Orchestration container management service managing the underlying instances and clusters.
- Specify the container image, memory and CPU requirements using an ECS task definition.
- Fargate schedules the containers' orchestration, management and placement throughout the cluster.




AWS Cloud
VPC
ECS Cluster
ECS Services
Fargate
ECS Tasks
ECR private

AWS Storage Services

Data Storage Options at AWS

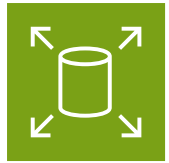
STORAGE TYPE	DETAILS	AWS SERVICES
Persistent data Storage	Data is persistent and durable.	S3, S3 Glacier, EBS, EFS, FSx for Windows File Server.
Transient data Storage	Data is temporarily stored before being consumed.	SQS, SNS
Ephemeral data Storage	Data records are lost when system(s) are stopped.	EC2 Instance Storage, ElastiCache for Redis.

EBS Features

- **Scalability**: Scale up or down typically without detaching the volume or stopping your instance.
- **Security**: EBS volumes offer encryption capabilities to secure your data at rest.
- **Availability and Reliability**: EBS volumes automatically replicate within the Availability Zone to protect from component failure.
- **Snapshots**: Backup EBS volumes to S3 using point-in-time snapshots.

Amazon Elastic Block Store

- High-performance block storage service designed for use with the Amazon Elastic Compute Cloud (EC2).
- Virtual hard drives attached to EC2 instances.
- Data volumes are persistent; no data loss when the associated instance is stopped or terminated.



Volume

EBS Performance

General Purpose SSD Volumes

gp2 SSD Drives (1G to 16 TB)

- Performance scales linearly between a minimum of 100 and a maximum of 16,000 at a rate of 3 IOPS per GiB of volume size.
- Volumes burst based on I/O credits.
- More expensive than gp3.

gp3 SSD Drives (1G to 16 TB)

- Single-digit latency (millisecond).
- Consistent IOPS 3,000 / volume.
- Provision up to 16,000 IOPS.
- Scale volume performance independently of volume size.
- Lowest cost than gp2 drives.

EBS Performance

Provisioned IOPS SSD Volumes

Provisioned IOPS SSD (io2)
(Block Express)

- Sub millisecond latency.
- Built on the Nitro System.
- Uses the SRD protocol.
- Up to 64 TB volume.
- Provisioned up to 256,000 IOPS.

Provisioned IOPS SSD (io1)

- Volumes use a defined IOPS rate
- 4G to 16TB.
- 100 - 64,000 IOPS per volume.

EBS Performance

HDD Volumes 0.125 TB to 16 TB

Throughput Optimized HDD (st1)

- Low-cost magnetic storage
- Throughput not IOPS
- Not bootable
- Baseline of 5 MB/sec. up to a cap of 500 MB/sec.

Cold HDD (sc1)

- Low-cost magnetic storage
- Infrequent data access
- Not bootable
- Baseline of 1.5 MiB/s to a maximum of 192 MiB/s.

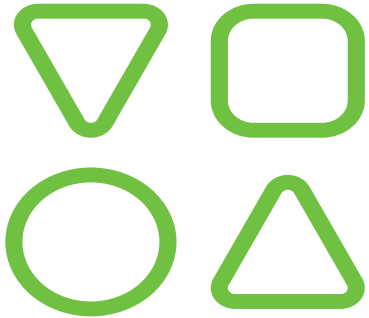
Amazon S3 Object Storage

- Secure, durable and scalable.
  - Stored in three availability zones per region minimum.
- Each object contains:
  - Key (Name), Value (Data), Metadata (System, User), Access control
- Each bucket can hold an unlimited number of objects.
  - Maximum object size is 5 TB.
- Objects are stored in buckets or vaults. (Amazon S3 Glacier)

Object

Amazon S3 Objects

- S3 can store any data type.
  - Up to 5 TB max for a single object.
- Each object has a unique key:
  - Key = filename
  - Must be unique within each bucket.
  - Metadata describes the data.



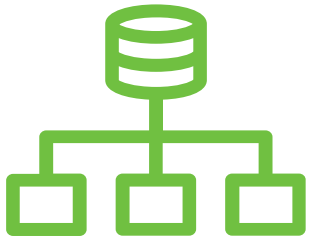
- System metadata - AWS date, size, storage class.
- User metadata - tags specified at the time the object is created.

S3 Architecture

- S3 buckets are stored in a single region.
- An S3 bucket name is a DNS namespace (Route 53); names must be globally unique.

  https://s3-***eu-east-1***.amazonaws.com/***<bucketname***>

- S3 automatically scales to high request rates:
  - 3,500 PUT/POST/DELETE and 5,500 GET requests per second.
  - For faster read requests, use CloudFront edge locations.



S3 Durability

Standard Storage Class

- 11 9’s durability
- 4 9’s availability
- No minimum storage time.

Standard IA Storage Class

- 11 9’s durability
- 4 9’s durability
- Minimum storage time of 30 days

S3 Storage Classes

S3 Standard - No minimum storage time.

S3 Intelligent-tiering - Move to the most cost-effective tier after 30 days.

S3 Standard-IA - Min 30 days.

S3 One Zone-IA - Store in one AZ, less resilience - min 30 days.

S3 Glacier Deep Archive - Long-term retention min 180 days.

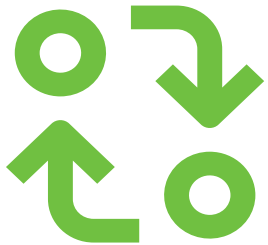
S3 on Outposts.

Cross-Region Replication (CCR)

- Asynchronous replication from source bucket in AWS region to destination bucket in another AWS region.
- Versioning must be enabled for both the source and destination buckets.
- Separate S3 lifecycle rules can be configured on both the source and destination buckets.
- Helps move data closer to end-users.
- Compliance / additional durability.

Same Region Replication (SSR)

- Replicate objects to a destination bucket within the same AWS region as the source bucket.
- Replication triggers include uploading objects, deleting objects, or changes to the object.



S3 Encryption

- Amazon S3 and S3 Glacier automatically encrypt all object uploads to all buckets.
- Amazon S3 supports three additional encryption options.

SSE – C

The client supplies the encryption keys.

SSE – KMS

Key Management Service supplies the encryption keys.

DSSE + KMS

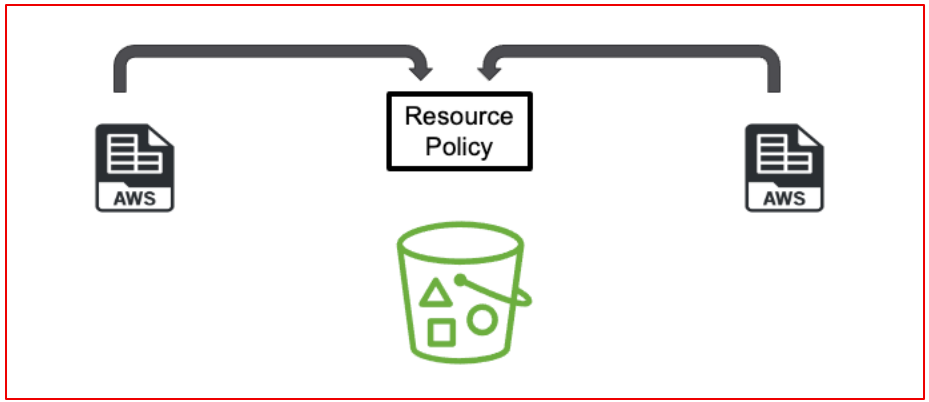
Enable dual-layer - SSE (Default SSE +KMS encryption)

Access Control

- Only the owner of the S3 bucket has access by default.
- IAM policies.
- Bucket policies – Resource-based policies.
- Query string authentication (Time-limited URL).
- Access auditing can be configured by configuring access log records for all requests made.

Bucket Policies

- Bucket policies can grant permission to access buckets and objects for other AWS accounts or IAM users.
- S3 buckets can be accessed by different AWS accounts using a bucket policy.



S3 Glacier Storage

- Archives from 100MB up to 40TB can be uploaded to S3 Glacier.
- Archives are assigned a unique ID and are stored in vaults.
- Each AWS account can have up to 1,000 vaults.
- Enable compliance controls per vault using a vault lock policy (WORM).
- Retrieval policies control the frequency of access.

S3 Glacier Storage Classes and Retrieval Options

Instant Retrieval

- Long live archive data accessed once a year.
- Retrieval in milliseconds.

Flexible Retrieval

- Long-lived archive data can be accessed once a year.
- Retrieval time of minutes to hours.

Deep Archive

- Long-lived archive data can be accessed once a year.
- Retrieval time of hours.

Amazon Simple Storage Service Glacier

- Data archiving and long-term backup.
- Stored in archives up to 40 TB in size.
- Archives are stored in vaults.
- Retrieval times from seconds to hours.

Archive

Vault

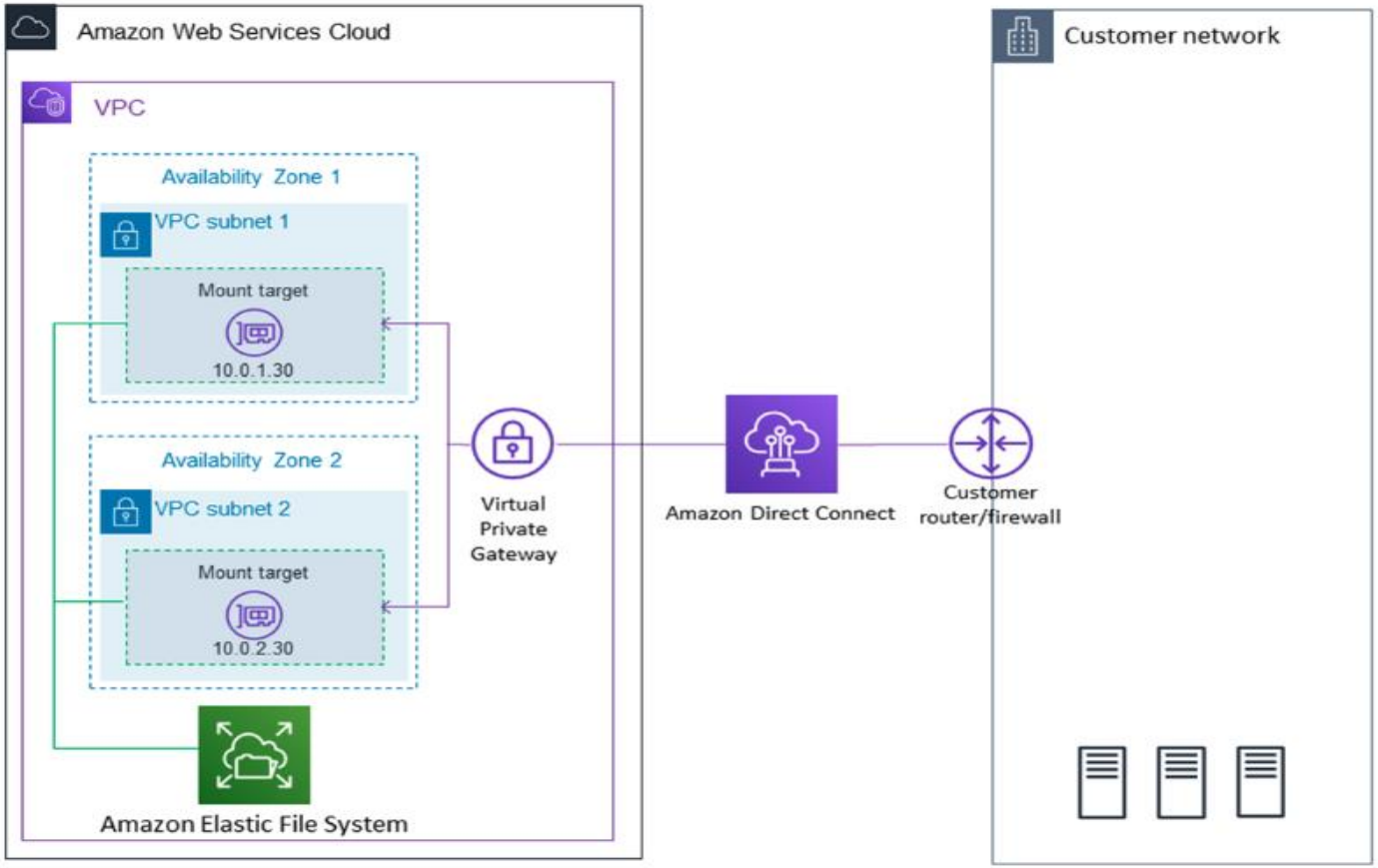
Amazon FSX for Windows File Server

- Fully managed native Microsoft Windows file system.
- Automatically grows and shrinks as files are added and removed; no provisioning storage in advance.
- Supports SMB protocol and Windows NTFS.
- Integrates with Microsoft Active Directory deployments.

File system

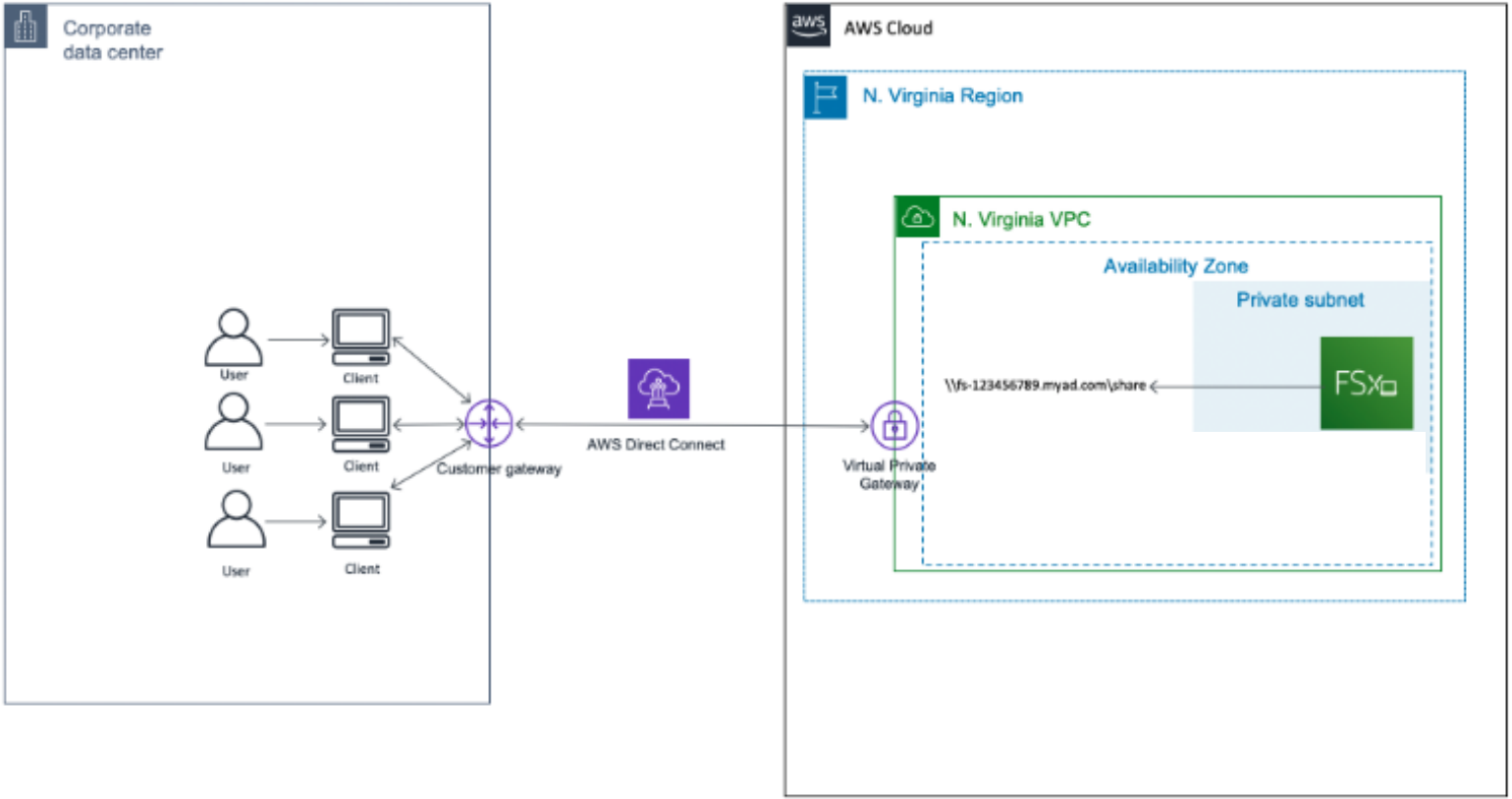
File system On-premises with Direct Connect or VPN.

- Create an Amazon EFS file system and a mount target in your Amazon VPC.
- Download and install the amazon-efs-utils tools.
- Access the file system from your on-premises client.


Amazon Web Services Cloud
VPC
Availability Zone 1
VPC subnet 1
Mount target
10.0.1.30
Availability Zone 2
VPC subnet 2
Mount target
10.0.2.30
Amazon Elastic File System
Virtual Private Gateway
Amazon Direct Connect
Customer router/firewall
Customer network

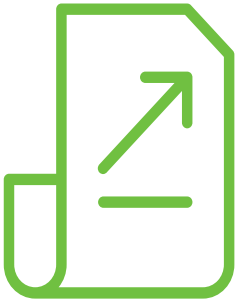
Windows Home Directories with FSx

- FSx for Windows File Server fully managed native Windows file systems can be used to host user home directories and user file shares.
- Map a user home directory from the FSx for Windows File Server file system.


Corporate
data center
User
Client
User
Client
User
Client
Customer gateway
AWS Direct Connect
Virtual Private
Gateway
AWS Cloud
N. Virginia Region
N. Virginia VPC
Availability Zone
Private subnet
\\fs-123456789.myad.com\share
FSx

FSx for Windows File Server Features

- **Native Windows File System**: Supports the SMB protocol and Windows NTFS and is integrated with Microsoft Active Directory.
- **Data Deduplication**: Optimize storage utilization by removing duplicate data blocks.
- **Highly Available and Durable**: Data is automatically replicated within multiple Availability Zones. A single AZ can also be chosen.
- **Backup and Restore**: Supports automatic daily backups and user-initiated backups.
- **Windows Authentication:** Choose a self-managed or managed Active Directory deployment for user authentication and file system access control.



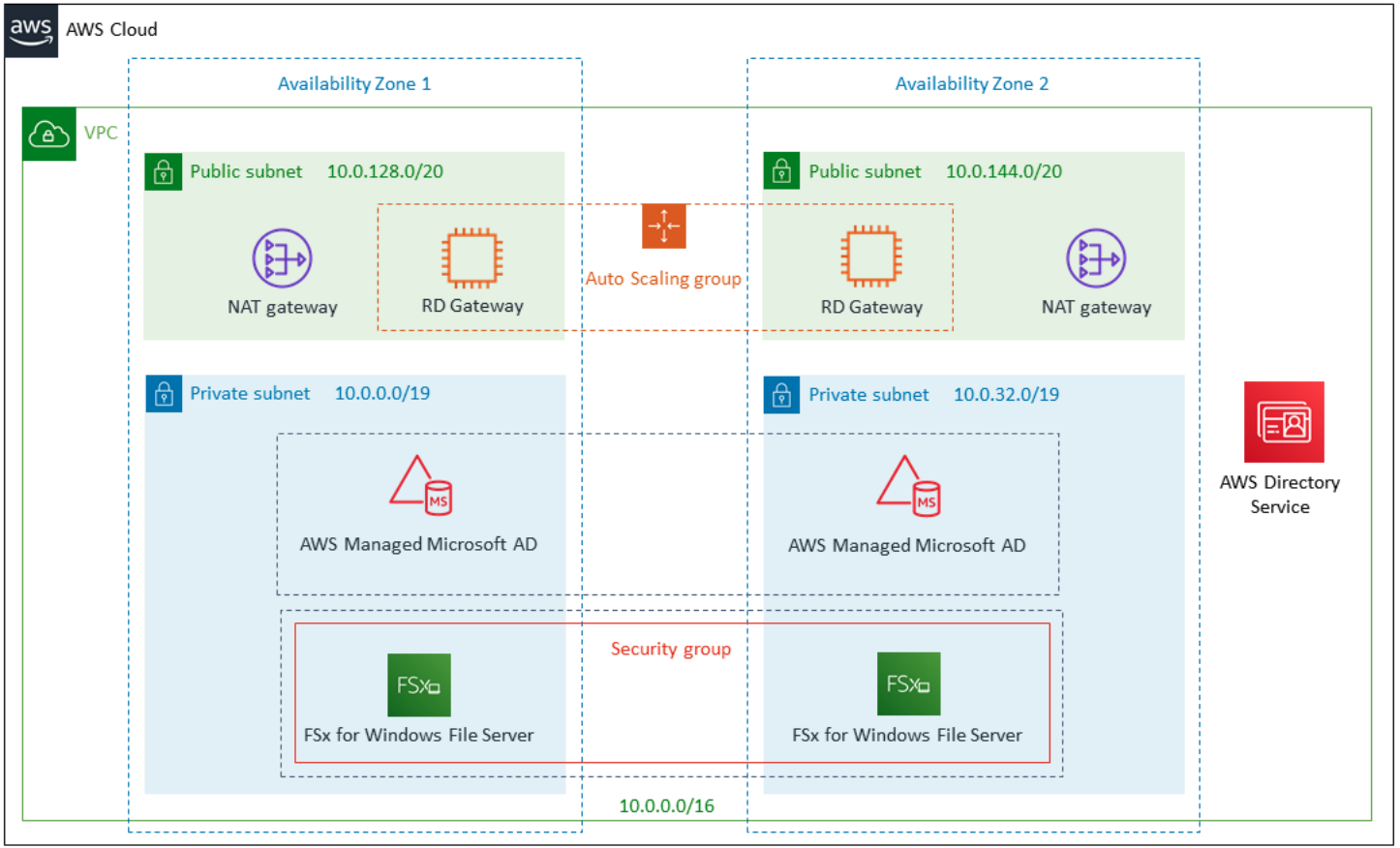
FSx for Windows File Server Performance Options

Storage Types

- **SSD Storage**: Offers high IOPS and throughput, suitable for performance-sensitive databases and media processing.
- **HDD Storage**: More cost-effective, suitable for applications such as file sharing

Throughput Capacity

- **Configurable Throughput**: Select the throughput capacity of the file system in megabytes per second (MB/s).
- **Automatic Throughput Scaling**: Some configurations support automatic scaling of throughput capacity based on your workload’s needs.


AWS Cloud
Availability Zone 1
Availability Zone 2
VPC
Public subnet 10.0.128.0/20
Public subnet 10.0.144.0/20
NAT gateway
RD Gateway
Auto Scaling group
RD Gateway
NAT gateway
Private subnet 10.0.0.0/19
Private subnet 10.0.32.0/19
AWS Managed Microsoft AD
AWS Managed Microsoft AD
AWS Directory Service
Security group
FSx for Windows File Server
FSx for Windows File Server
10.0.0.0/16

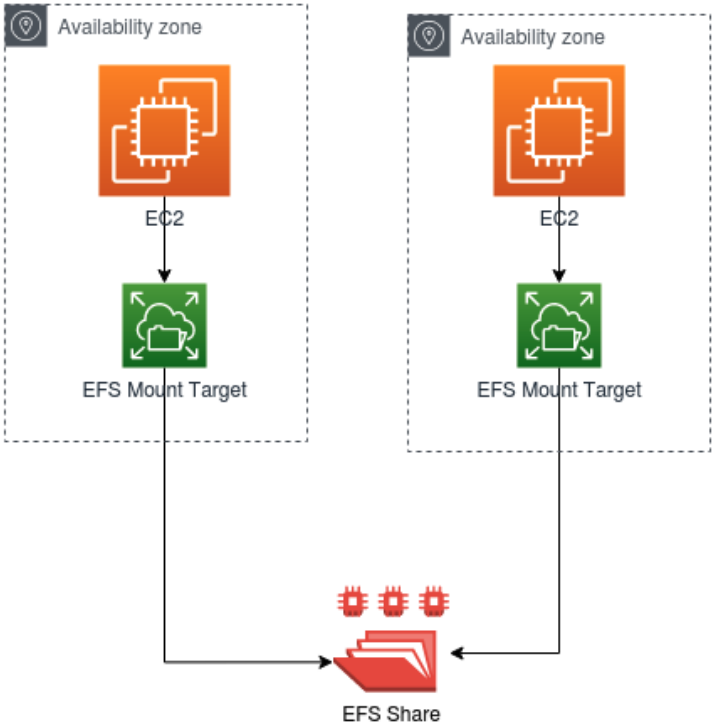
Amazon Elastic File System

- Amazon Elastic File System (Amazon EFS) is a fully managed elastic NFS file system service provided by AWS.
- Automatically grows and shrinks as files are added and removed; no provisioning storage in advance.
- Multiple EC2 instances can access an EFS file system at the same time.

File system

Elastic File System

- Fully managed scalable and shareable storage service (Up to petabytes).
- NFS file share access using the NFS protocol and mount points in one or many AZs.
- Elastic storage capacity: pay for what you use.
- It can be mounted from on-prem systems using Direct Connect or a VPN connection.
- AWS Backup solution to backup file system.
- General-purpose or Max I/O operation.
- EFS can burst to high throughput levels.


Availability zone
EC2
EFS Mount Target
Availability zone
EC2
EFS Mount Target
EFS Share

EFS Performance Modes

- **Automatic Scaling**: Storage capacity scales up or down automatically as files are added or removed, without manual intervention or provisioning.
- **Performance Modes**
  - **General Purpose**: Recommended starting mode of operation.
  - **Max I/O Performance**: Optimized for high levels of aggregate throughput and operations, suitable for large-scale data processing and media processing workloads.

AWS Storage Gateway

- Volume Gateway - local volume synched with S3 storage.
- File Gateway - local file share synched with S3 storage.
- Tape Gateway - The virtual tape library is stored in S3 storage.

- Amazon S3 File Gateway - Applications directly access objects stored in S3.
- Amazon FSx File Gateway - Applications directly access data stored in FSx Windows File Server storage.

Amazon S3
File Gateway

Amazon FSx
File Gateway

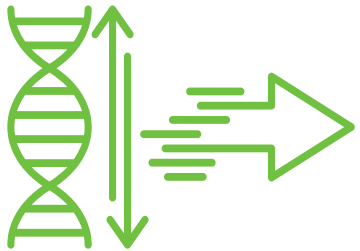
Tape
Gateway

File
gateway

Volume
Gateway

EFS Throughput Modes

- **Bursting Throughput**: Automatically scales with the size of the file system. Suitable for workloads with sporadic read/write operations.
- **Provisioned Throughput**: Allows you to specify the throughput independent of the amount of data stored. It is beneficial for applications that require a higher sustained level of throughput.
- Elastic Throughput: If the application has low baseline activity but has unexpected spikes, then select elastic throughput mode.

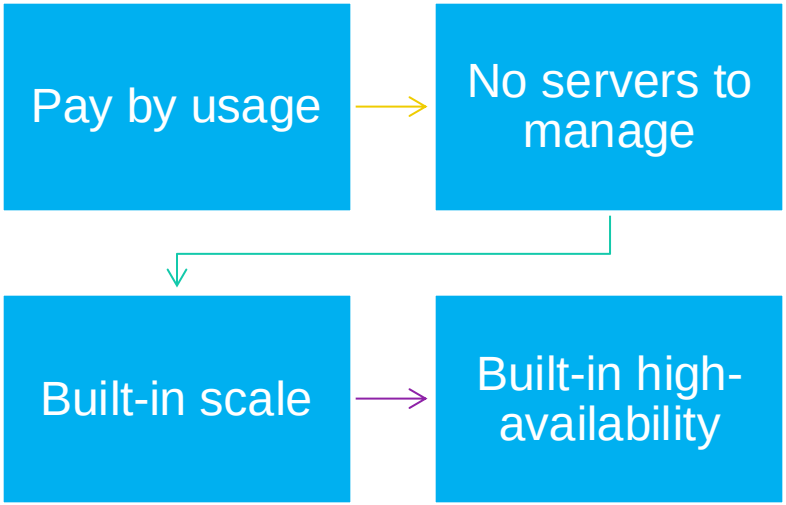


EFS Storage Classes

- **Standard**: For frequently accessed files.
- **Infrequent Access (IA)**: Lower-cost option for files accessed less frequently. EFS automatically moves files to and from the IA storage class based on access patterns.
- **One Zone Infrequent Access:** The same operation as the Infrequent Access storage class but stored in a single availability zone.

Serverless Computing

Serverless Computing



AWS Lambda

- No EC2 instances to manage and scale.
- Background scaling is handled by AWS.
- Sub-second metering – pay when each function executes.
- Bring your own code- Node.js, Java, Ruby, Python, C#, Go.
- Integrates with other AWS services.
- Select power rating from 128 MB to 1.5 GB: CPU and network will be proportionally allocated

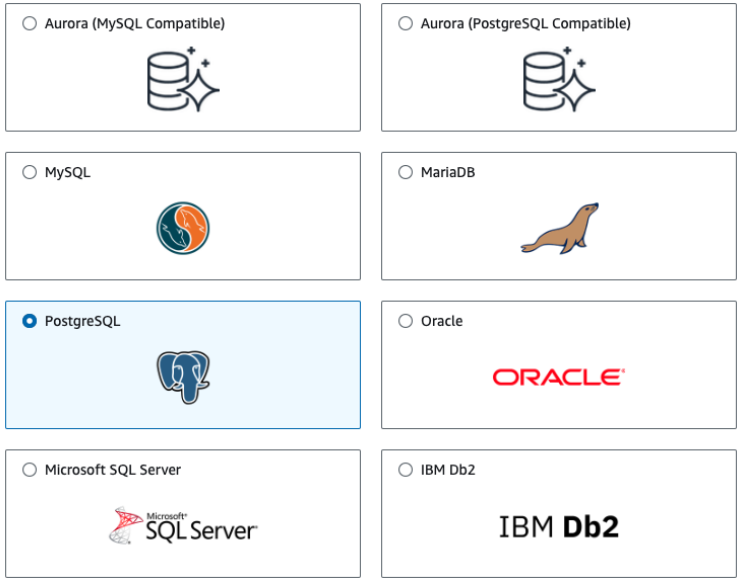
AWS Lambda Triggers

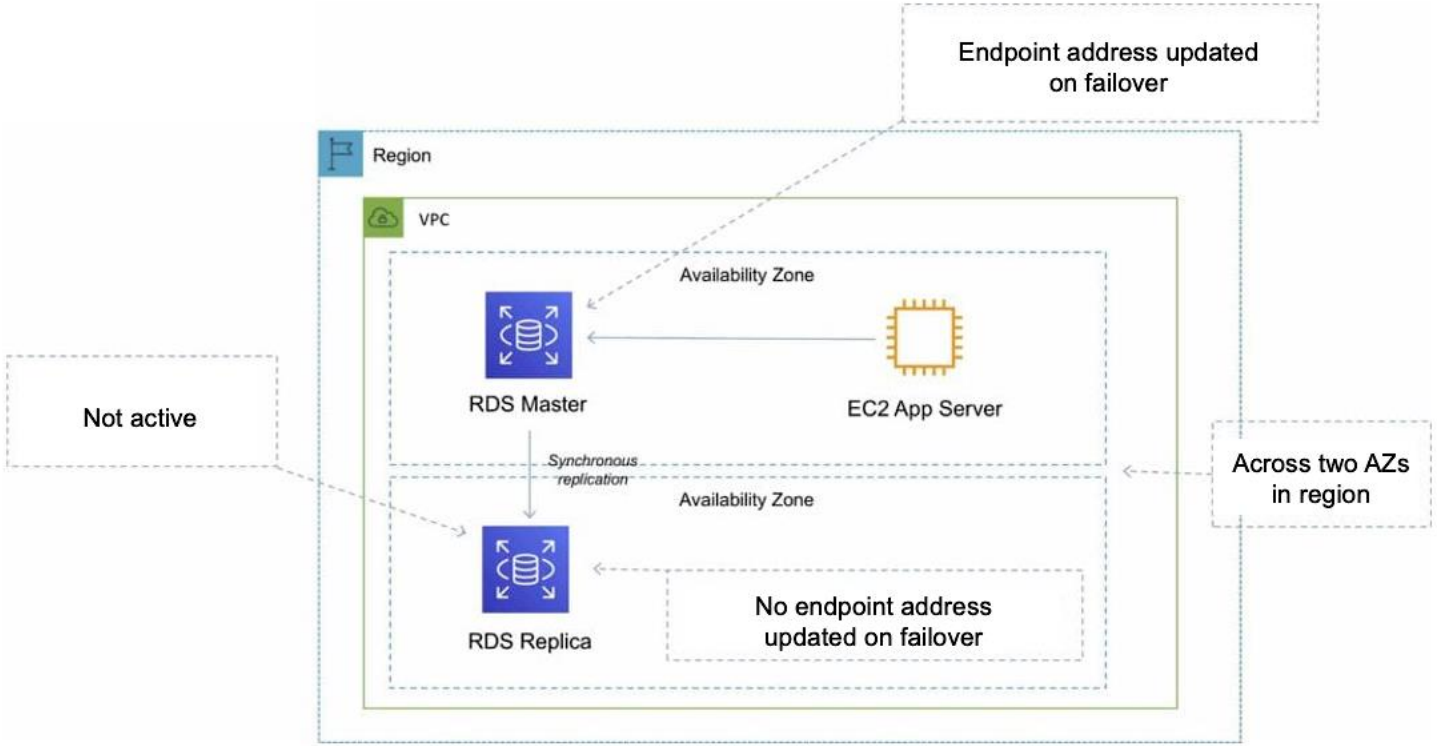
- **S3 bucket:** An object is uploaded to a bucket.
- **DynamoDB table:** An entry is made in a DynamoDB table.
- **Amazon Kinesis:** Data is added to an Amazon Kinesis stream.
- **Amazon SNS:** A message is published to an SNS topic.
- **Amazon API:** RESTful APIs call a Lambda function to access backend AWS services.
- **Application Load Balancer (ALB):** Requests can be directed to an AWS Lambda function hosted by a target group.

Databases

RDS Database Instances

- RDS uses AWS-managed database instances (Primary, Standbys)
- Single, Multi-AZ deployments.
- Up to 15 read replicas.

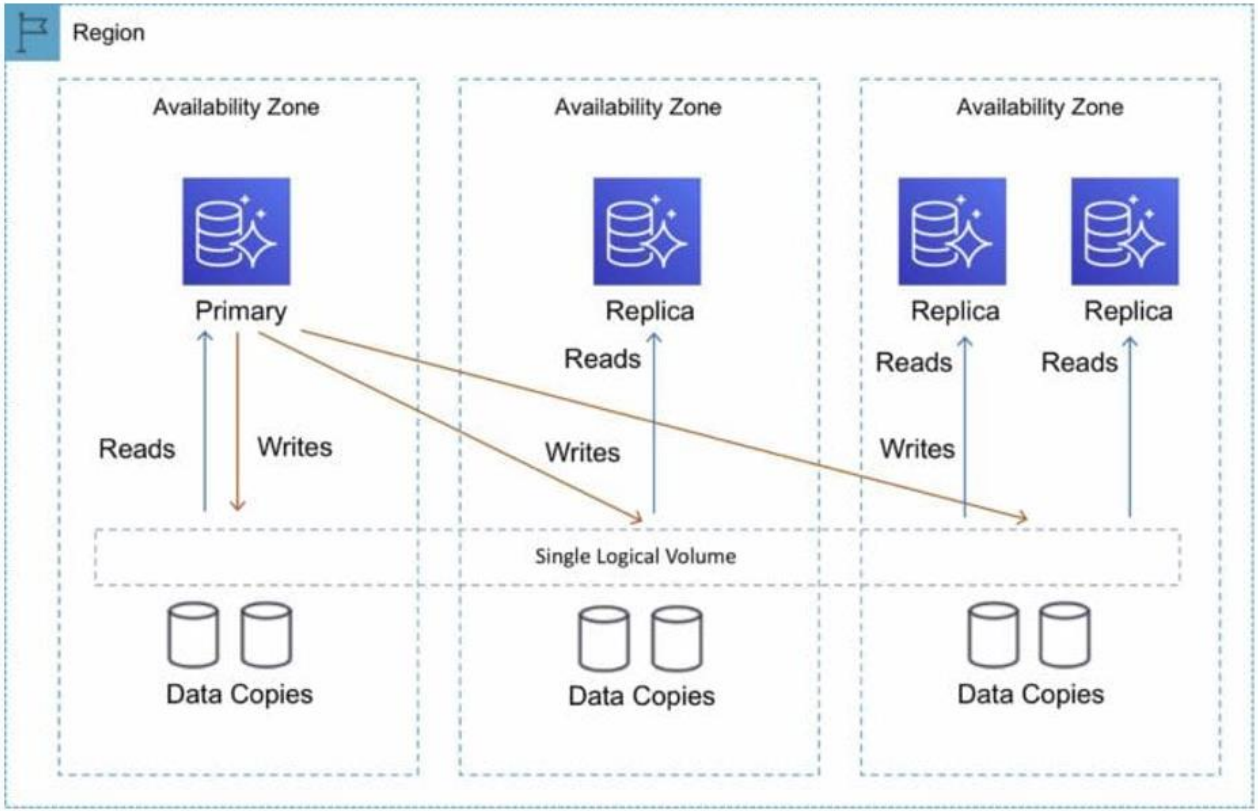



Endpoint address updated on failover
Region
VPC
Availability Zone
RDS Master
EC2 App Server
Not active
Synchronous replication
Availability Zone
Across two AZs in region
RDS Replica
No endpoint address updated on failover

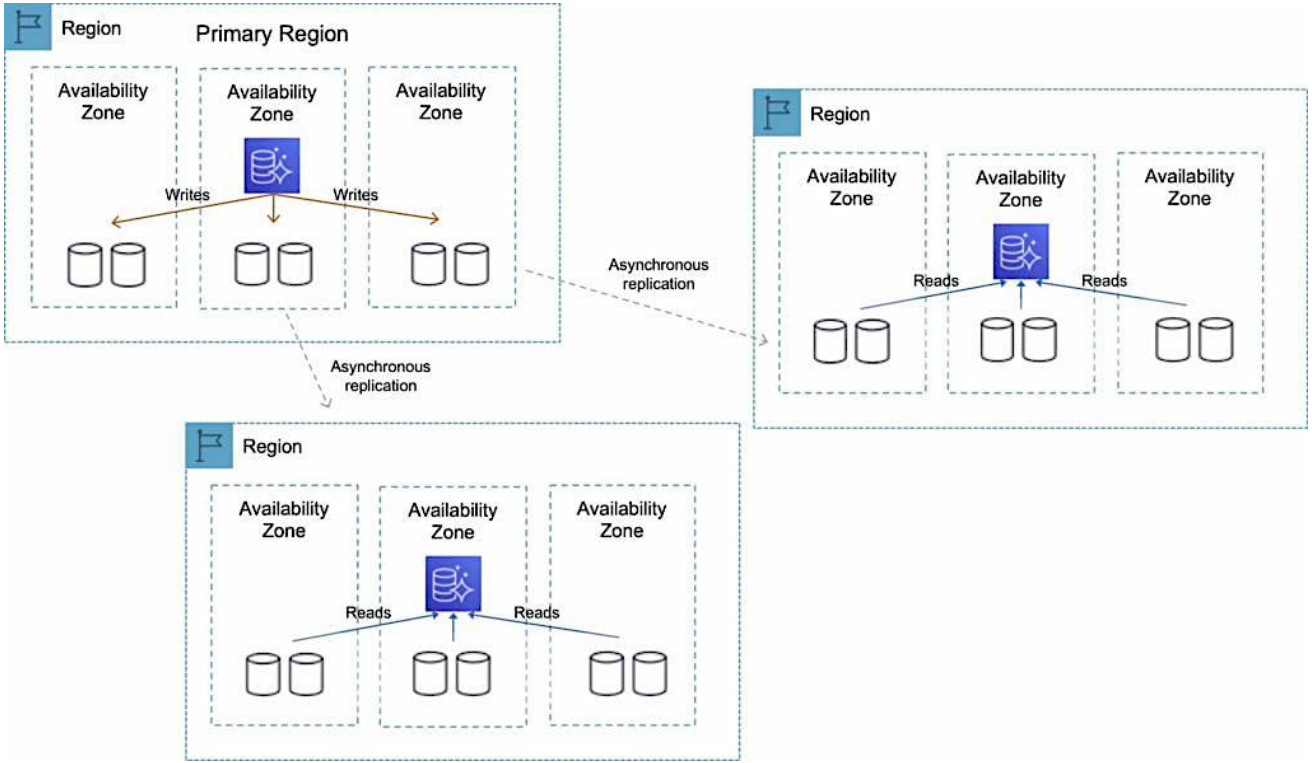
Amazon Aurora

- An Aurora cluster storage volume consists of six storage nodes in three Availability Zones, providing high availability and durability of data.
- Fully managed 6-way replication across three availability zones with SSD Virtual SAN storage.
- Two copies of data are kept in each AZ.
- Can be deployed as a global database with cross-region replication and read replicas.
- Aurora read replicas can be created in up to five AWS regions.

Amazon Aurora

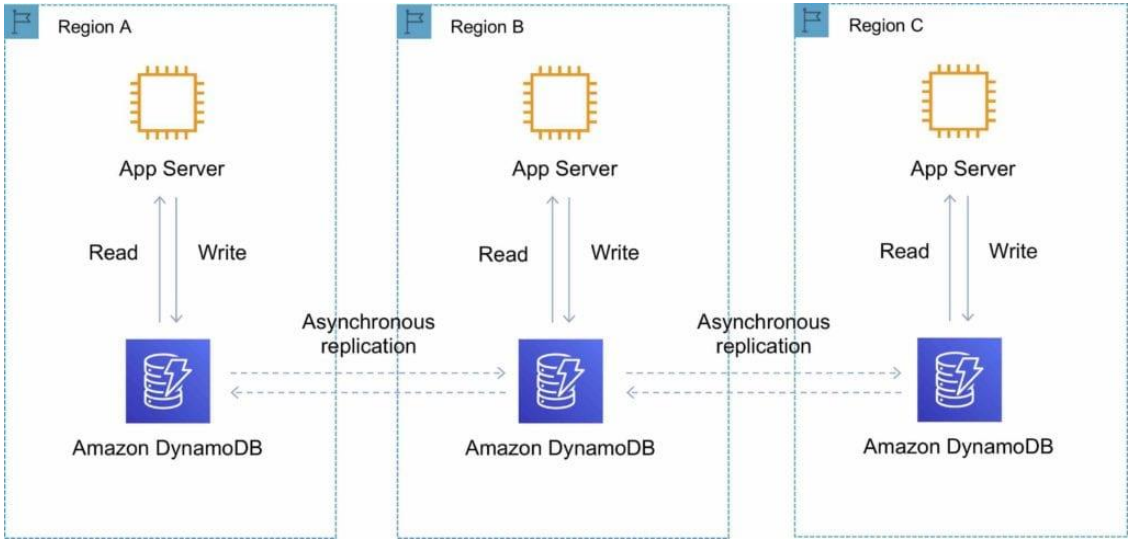


Amazon Aurora: Multi-region



DynamoDB

- Managed NoSQL database with fast performance and seamless scalability with no downtime.
- DynamoDB is designed with automatic synchronous data replication across three AZs.
- DynamoDB supports cross-region replication across regions with Global tables.

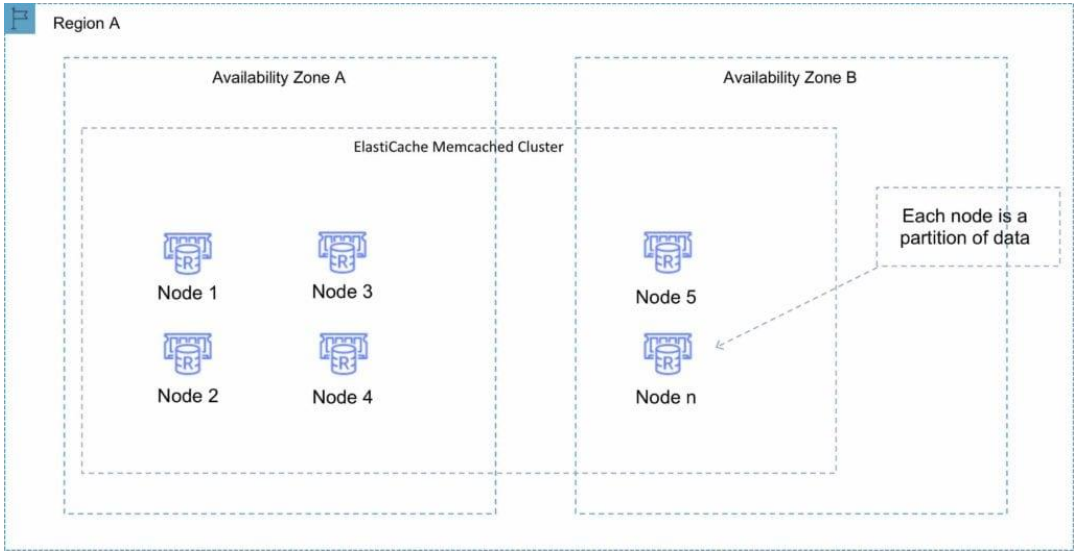


ElastiCache

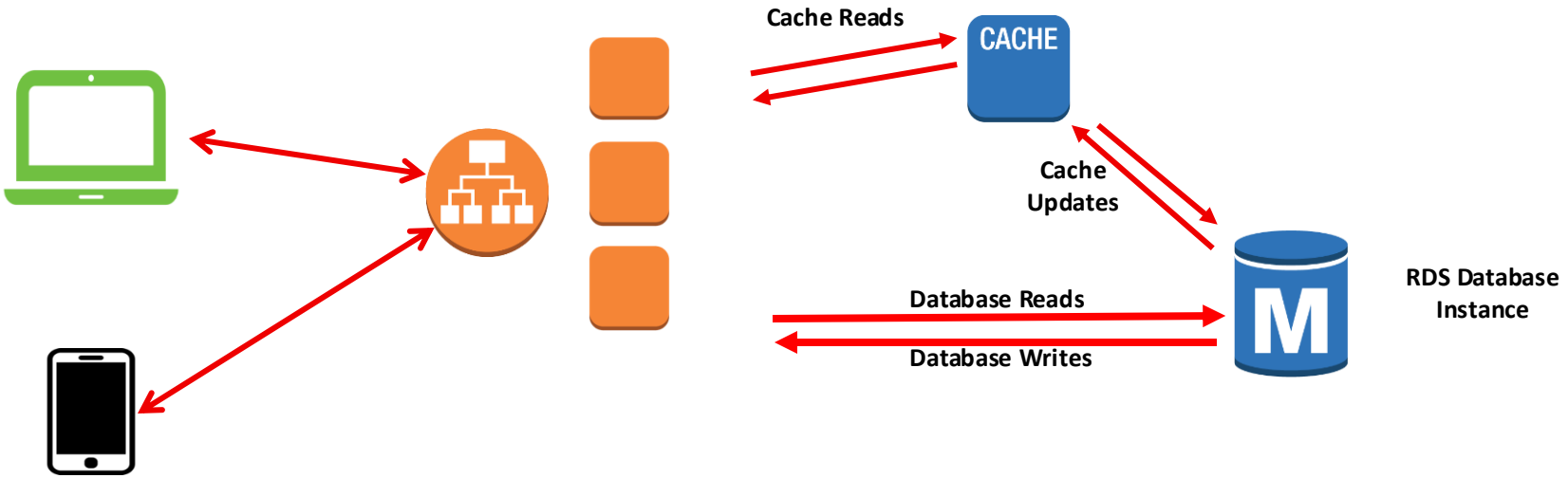
- Managed implementation of Redis and Memcached in-memory data stores.
- EC2 nodes cannot be accessed from the Internet.
- A node is a fixed size of network-secured RAM.
- Nodes are deployed in clusters utilizing the selected caching engine.

Memcached

- Can run large nodes with multiple cores and threads.
- Memcached is not persistent; node failure results in data loss.
- User session storage.
- Database caching (RDS).

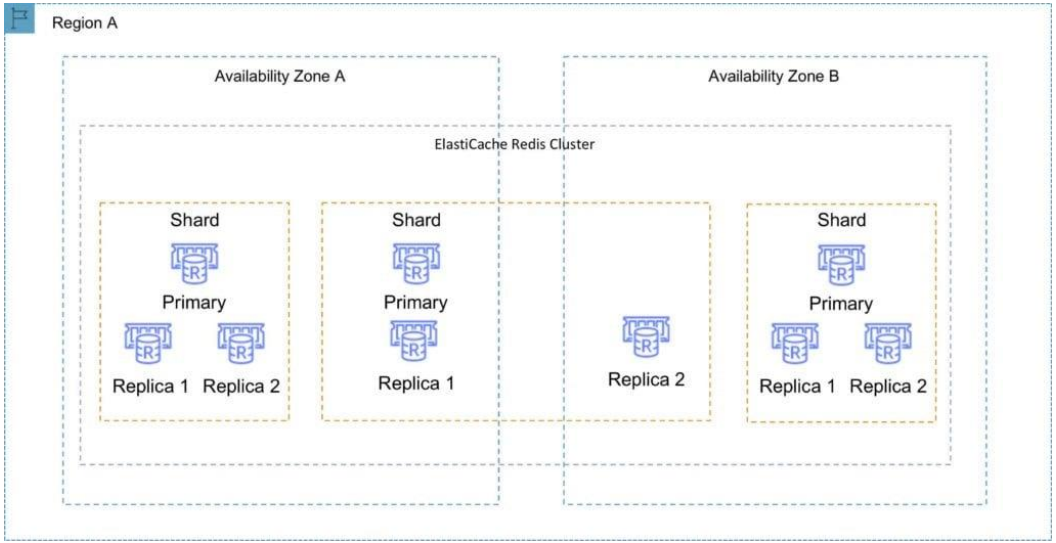


ElastiCache Operation



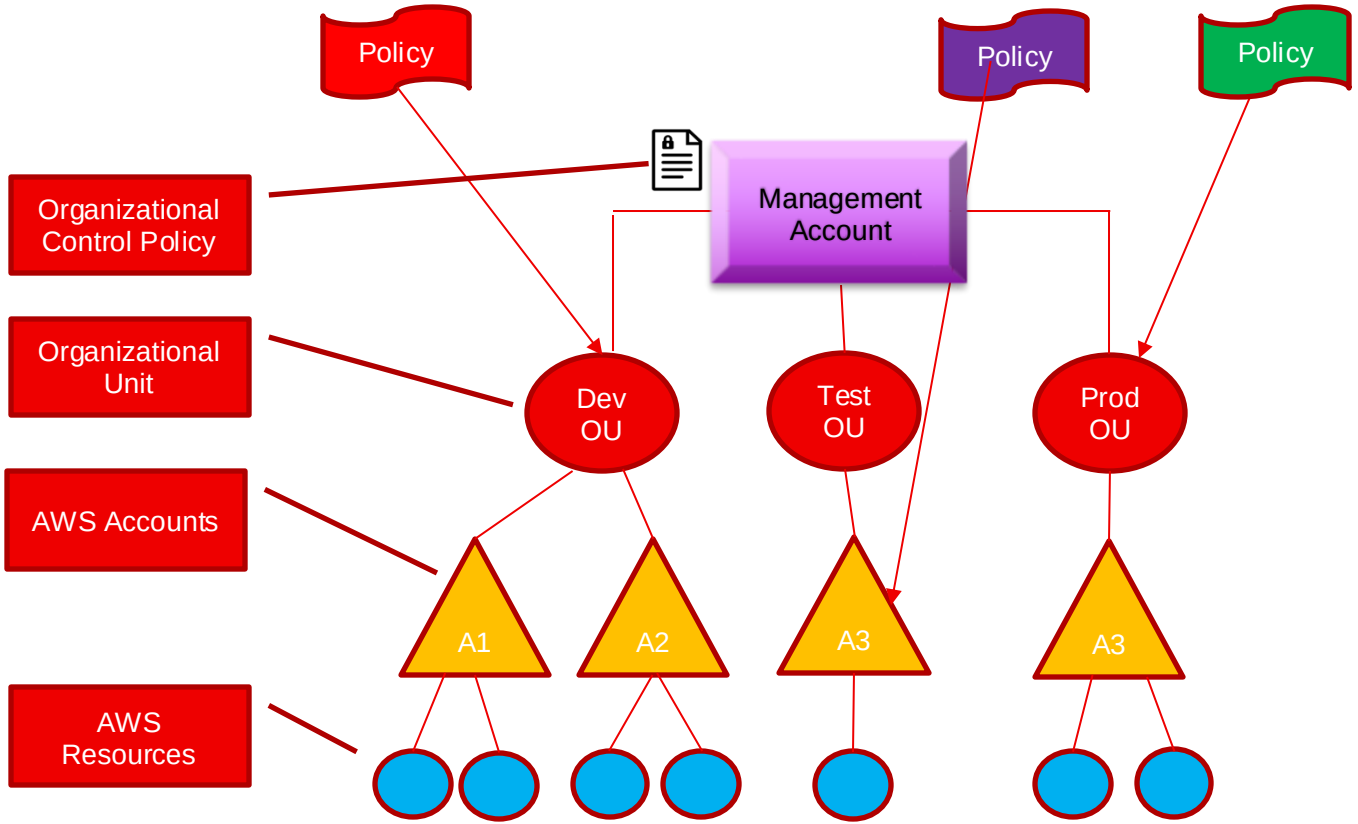
Redis

- Used as a persistent data store.
- Shards include a primary node and optionally read replicas.
- Supports automatic snapshots and backups.
- Multi-AZ support for read replicas.



AWS Organizations

AWS Organizations



AWS Organizations

Consolidate multiple AWS accounts into a single organization.

- All accounts are centrally managed.
- Consolidated billing using the root management account.
- Group accounts into organizational units (OUs).

Integration with AWS Services.	Standardize consistent tags across resources.
Granular control over users and roles.	Configure automatic backups for resources.

Control Tower

- Manage and govern AWS Organization deployments.
- Integrates with AWS Organizations, AWS Service Catalog, and AWS IAM Identity Center.
- Automate AWS account deployment and configuration in an AWS Organization.
- Maintain compliance and governance across an AWS Organization.

Landing Zone

Enterprise-wide container with all OUs, accounts, users, and resources.

Account Factory

Automates account deployment and enrolment in the AWS organization.

Controls

Provide ongoing governance with guard rails.

IAM Identity Center

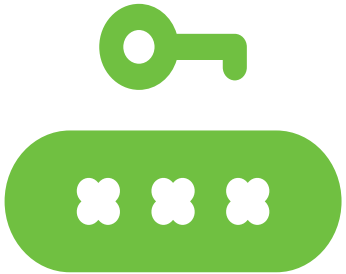
- Manage sign-in security for user identities.
- Connect and centrally manage access across AWS Organization accounts.
- Connect and manage AWS accounts, individual IAM users and Roles, and Identity Centre-enabled applications.

Resource Access Manager

Share resources
Across AWS Organization or multi-account environments.

Central Management
Manage Cloud Services centrally automating account deployment and enrolment in the AWS organization.

Least privilege controls
Grant required permissions for essential tasks.

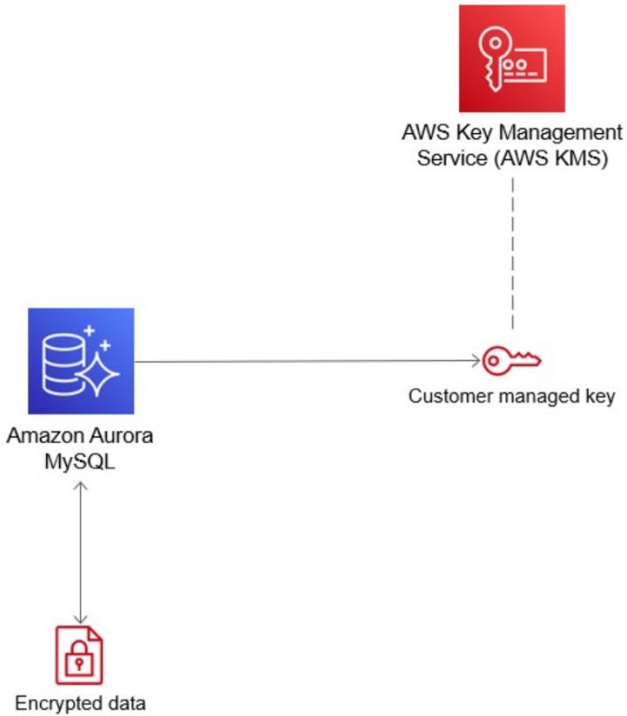


AWS Security Services

AWS Key Management Service

- Manage encryption/decryption of AWS data services.
- Centrally store your customer master key (CMK).
- CMK can be generated using AWS KMS or AWS CloudHSM.
- Unique data keys are used for each encryption request.
- KMS stores multiple copies of encrypted keys with 99.999999999% durability.

KMS Encryption



Data Keys

- Data keys are encrypted by your Customer Master Key (CMK).
- Data keys are not reused.
- Encrypted data is stored along with an encrypted copy of the data key.
- When an AWS service needs to decrypt your data, KMS decrypts the data key using your CMK.
- All requests to use your CMKs are logged in AWS CloudTrail.

AWS Secrets Manager

- Store, rotate, and manage organizational secrets used to access your applications, services, and IT resources.
- Store and manage database credentials and API keys.
- Secure secrets for SaaS applications, SSH keys, RDS databases, third-party services, and on-premises resources.

AWS GuardDuty

- Reconnaissance by attacks: Port scanning, API attacks, unusual logins.
- EC2 instance compromise: Backdoor control, Temp EC2 credentials.
- Account compromise: Disabled CloudTrail logging.
- S3 Bucket compromise: Data access patterns, Unusual S3 API activity.

GuardDuty Analysis

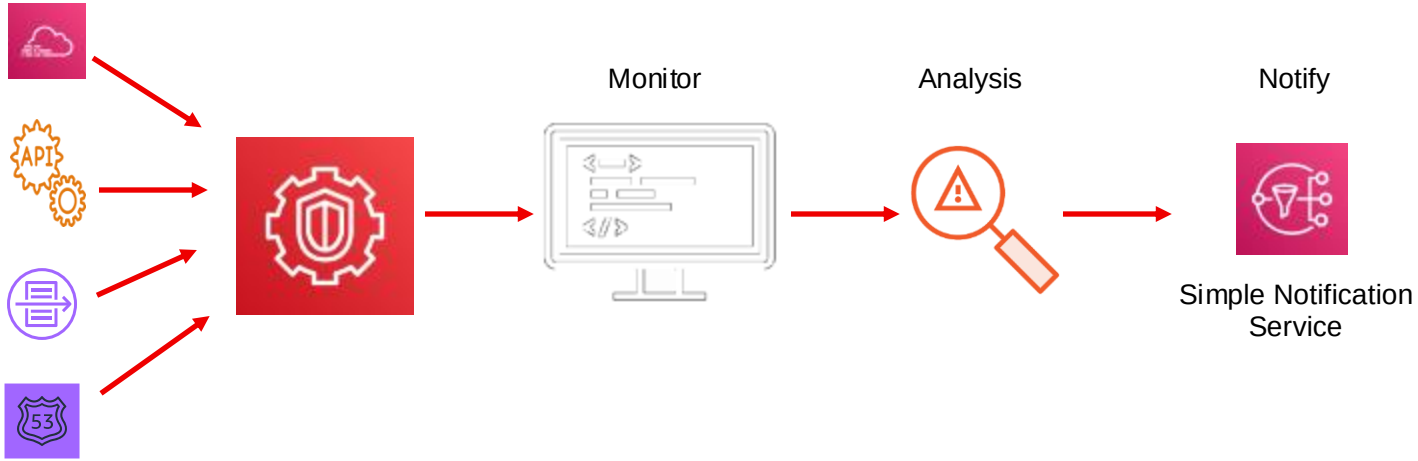
CloudTrail Authentications

CloudTrail APIs

DNS Query Logs

VPC Flow Logs

GuardDuty Integration



AWS Config

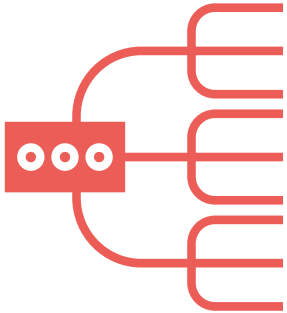
- AWS resource inventory
- Configuration history
- Configuration change notifications
    - “What did my resources look like on this date?”
    - “Who made an API call to modify this resource?”

AWS Config Rules

- Config rules are evaluated against configuration changes on the resource
- If violations are found, the resource is flagged as non-compliant
- A “Configuration item” is the configuration of a resource at a given time

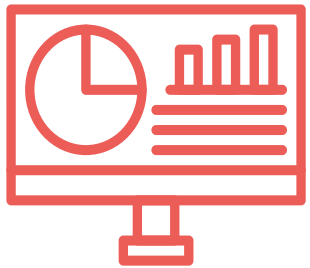
Amazon Macie

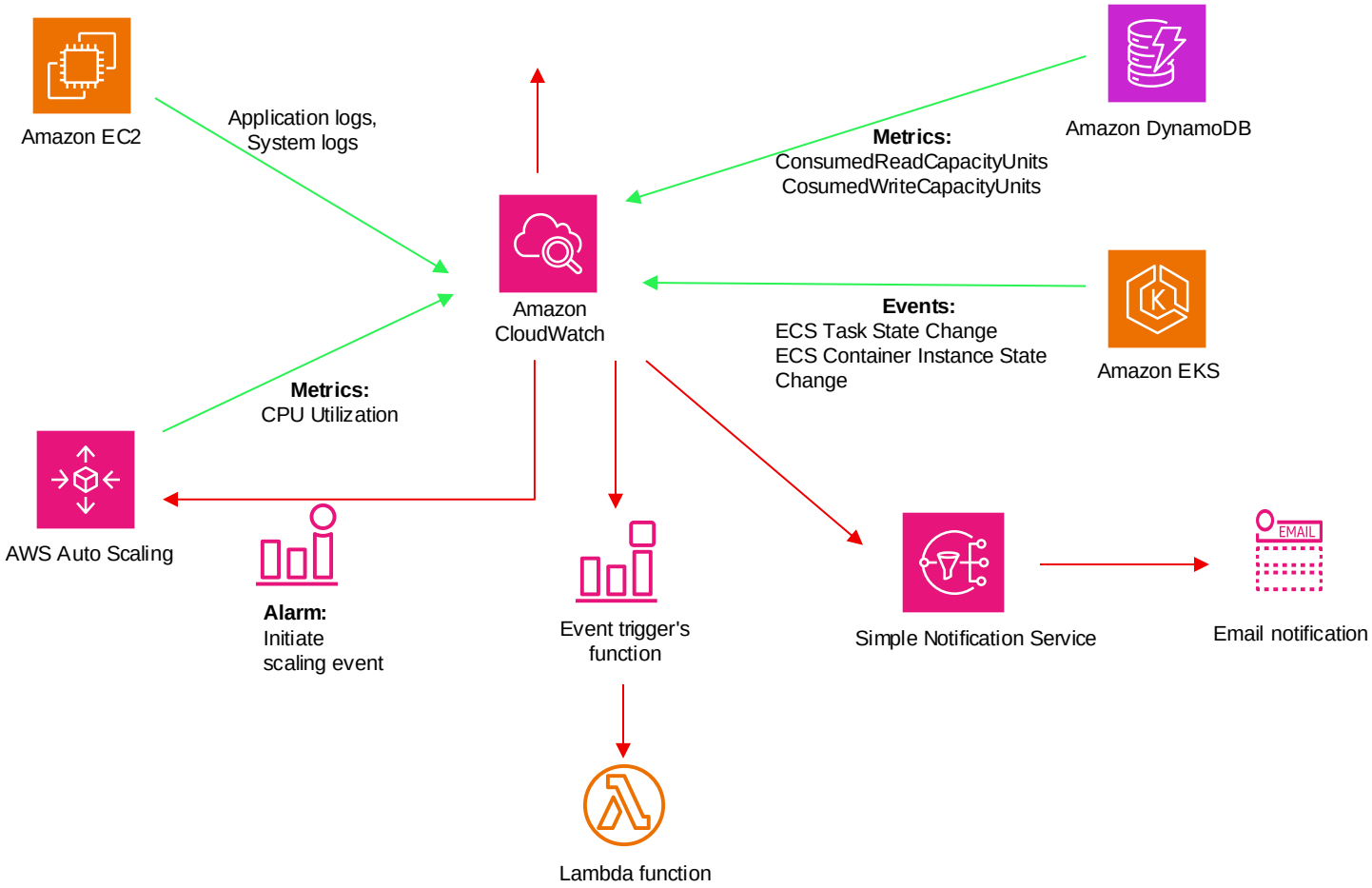
- Automated security classification of data access patterns.
- Monitor data usage for anomalies.
- Proactive data loss through data visibility.
- Custom report and alert management.



AWS CloudWatch

- Regional monitoring services for over 70 AWS cloud services.
- Metrics are added to the dashboard when AWS service is ordered.
- Metrics events trigger alarms.
- Metric filters analyze captured CloudWatch logs.
- Insights provide detailed information and preset search filters.
- Dashboards customize monitoring information.

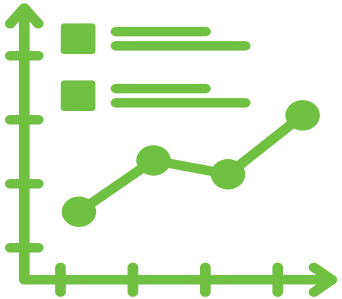



Amazon EC2
Application logs,
System logs
Amazon DynamoDB
Metrics:
ConsumedReadCapacityUnits
CosumedWriteCapacityUnits
Amazon
CloudWatch
Events:
ECS Task State Change
ECS Container Instance State
Change
Amazon EKS
Metrics:
CPU Utilization
AWS Auto Scaling
Alarm:
Initiate
scaling event
Event trigger's
function
Simple Notification Service
EMAIL
Email notification
Lambda function

Metrics

- Each metric collects data published to CloudWatch based on standard or high-resolution.
    - Standard resolution data displays changes every minute.
    - High-resolution data displays changes every second.

- Statistics are calculated based on metric data points. (Minimum, Maximum, Sum, Average)



Alarms

- Each CloudWatch alarm is associated with a single metric.
- Alarms are triggered by state changes to the defined threshold over the defined time period.
- Alarms invoke actions for sustained state changes, sending notifications to an Amazon SNS topic.

CloudWatch Logs

- CloudWatch Logs store logs from Amazon EC2 instances, AWS CloudTrail, VPC Flow Logs, and AWS Lambda.
- Application and system logs can trigger notifications.
- Alarms can be created in CloudWatch for real-time monitoring based on API activity captured by CloudTrail.

AWS CloudTrail

- AWS account audit trail enabled by default.
- Stores API calls and authentications as events (for 90 days).
- Create custom trails for storing events permanently (S3 bucket, CloudWatch log).
- Management, Data, and Insights data types.
- Integrates with AWS Organizations.

CloudWatch Logs

- CloudWatch logs store log files from Amazon EC2 instances, AWS CloudTrail, Route 53, and more.
- For near real-time log monitoring, use Kinesis Firehose.

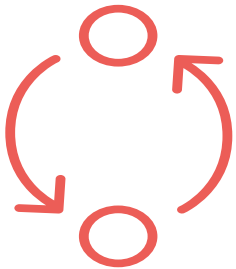
Amazon CloudWatch vs AWS CloudTrail

Performance Monitoring

- Log events across AWS services.
- High-level monitoring event service.
- Log from multiple accounts.
- Logs stored indefinitely.

Auditing

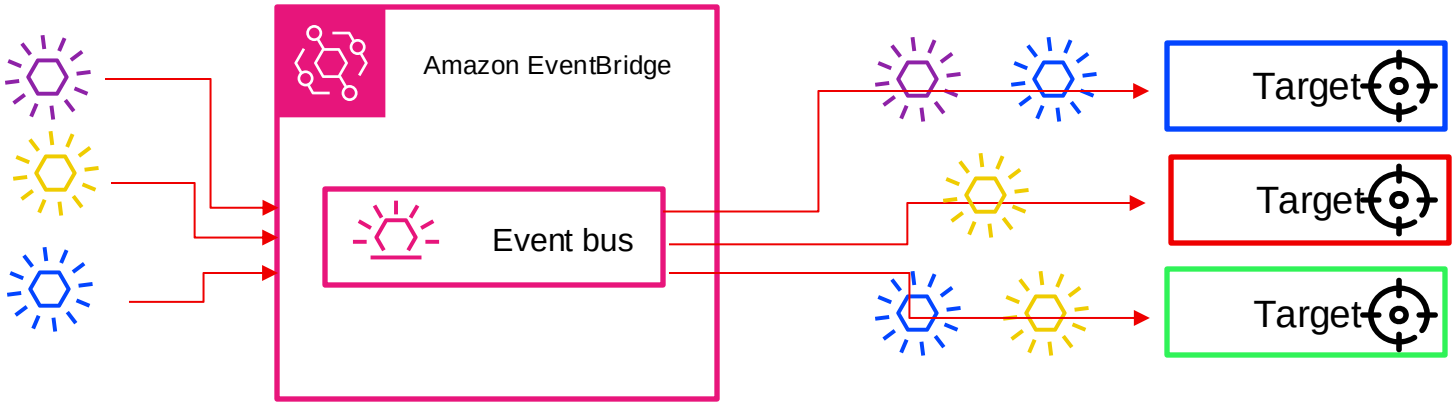
- Log API activity across AWS services.
- Low-level granular events.
- Log from multiple AWS accounts.
- Logs are stored in S3 buckets or CloudWatch logs indefinitely.



Amazon EventBridge

Amazon EventBridge

- Events connect application components together.
- Route events from applications, AWS services, and 3rd party SaaS applications.
- Works across regions.



Amazon EventBridge

Custom
event bus

SaaS
event bus

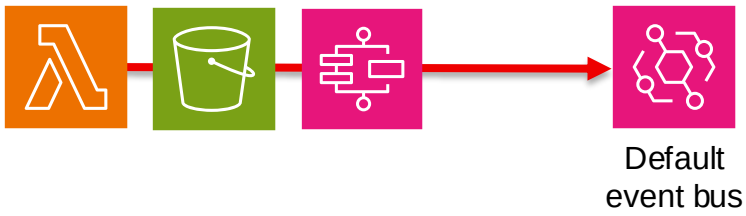
Default
event bus

- Event buses - Routers receive events and deliver them to targets.
- Pipes - Point-to-point integration (Single source to single target)

Pipes

Amazon EventBridge

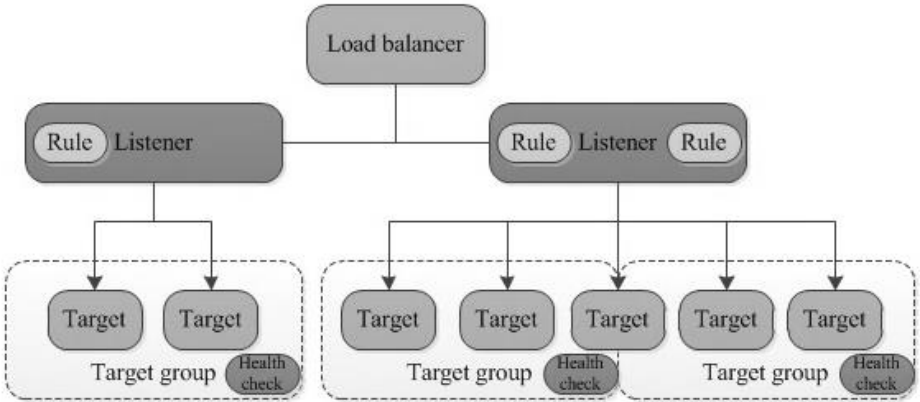
- AWS native events are sent to the default bus, without any prior configuration.
- EventBridge integrates with many AWS services (API Gateway, EC2, Glue, Kinesis Firehose, Lambda, Step Functions, SQS and SNS).

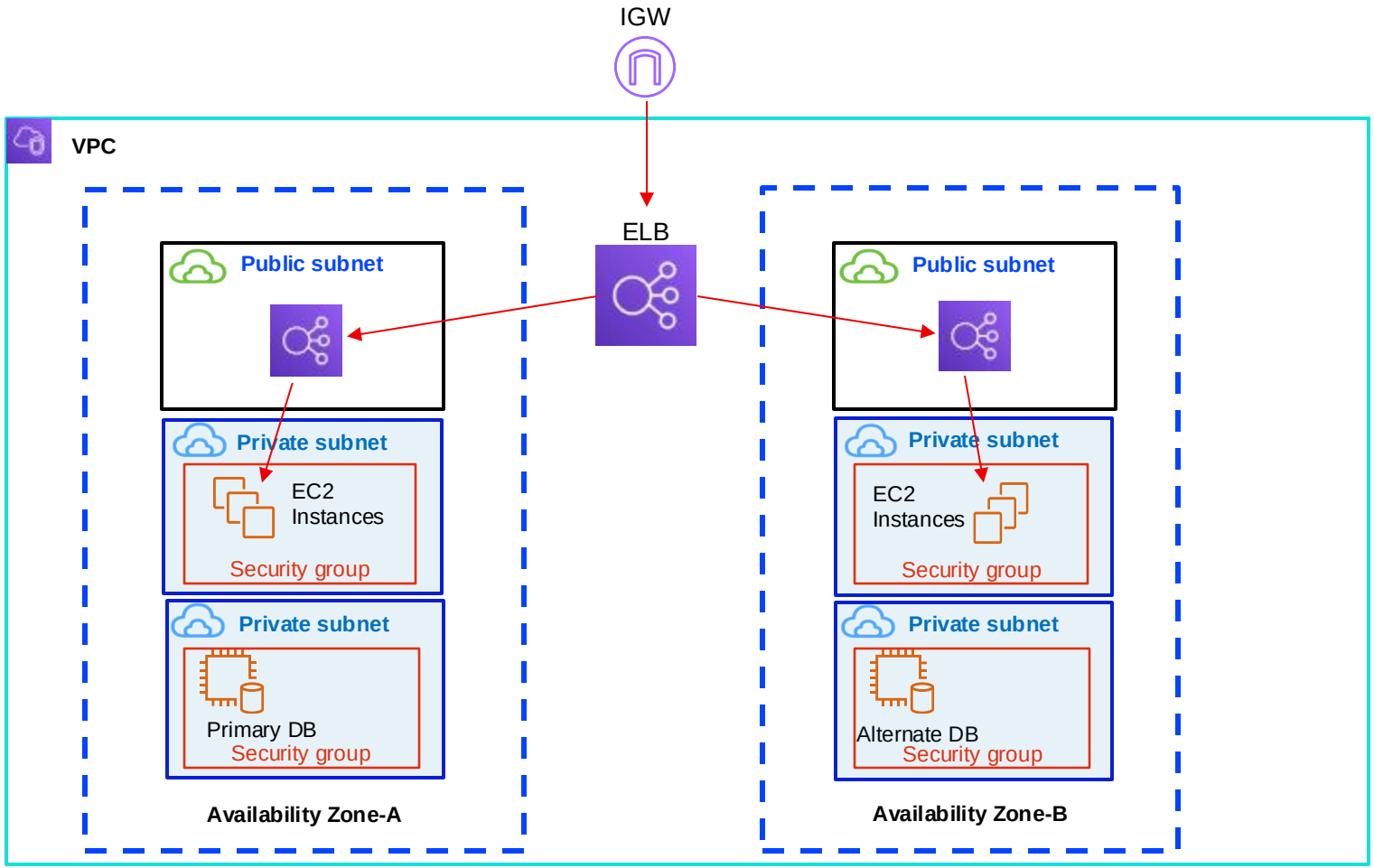


Load Balancing

Elastic Load Balancing Service

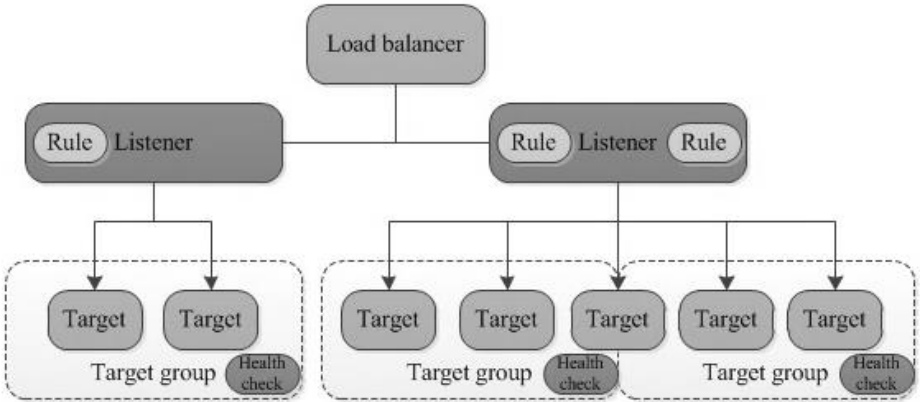
- Distribute incoming traffic across multiple EC2 instances or containers across one or multiple availability zones.
- Network load balancer: Target groups support the TCP, UDP, TCP_UDP, and TLS protocols.
- Application load balancer: Target groups support the HTTP and HTTPS protocols.




IGW
VPC
ELB
Public subnet
Public subnet
Private subnet
EC2
Instances
Security group
Private subnet
EC2
Instances
Security group
Private subnet
Primary DB
Security group
Private subnet
Alternate DB
Security group
Availability Zone-A
Availability Zone-B

Target Groups

- Target groups route requests to individual registered targets, such as EC2 instances.
- When cross-zone load balancing is on, each load balancer node distributes traffic across the registered targets in all registered Availability Zones.
- When cross-zone load balancing is off, each load balancer node distributes traffic only across the registered targets in its Availability Zone.

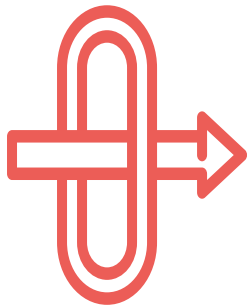


Health Checks

- Health checks ensure the resource is available and ready to accept traffic.
- A registered target is defined as *healthy* or *unhealthy.*
  - A newly added target to the target group is marked as *initial*.
  - Once a health check is successful, the target is marked as *healthy.*
  - When a health check is unsuccessful, the target is marked as unhealthy.
  - When connection draining is underway, the target is marked as *draining.*

Gateway Load Balancer

- Third-party solution - virtual appliance.
- Gateway Load Balancer (GWLB): Deploy and manage a fleet of virtual appliances.
- Gateway Load Balancer Endpoint (GWLBE): Link GWLB with VPC using VPC endpoints.

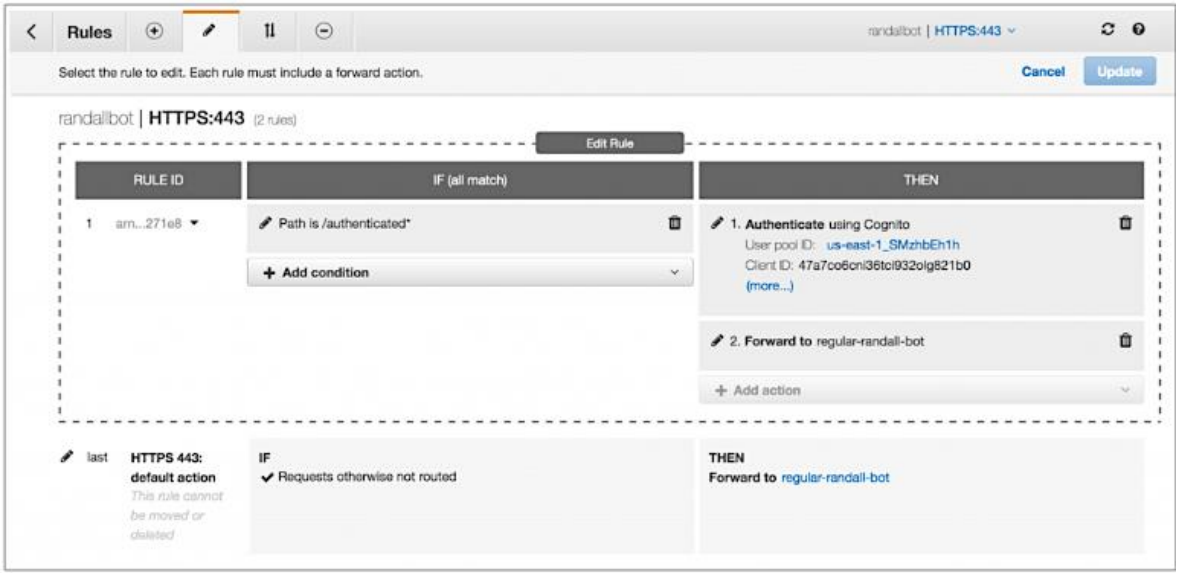


Application Load Balancer Features

- TLS termination.
- Authentication (Cognito).
- Web application firewall support (WAF).

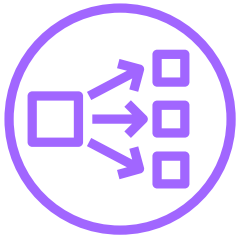
User Authentication

- The ALB can authenticate users when they request access to cloud applications.
- The ALB integrates with AWS Cognito, allowing Web-based and enterprise identity providers to authenticate.



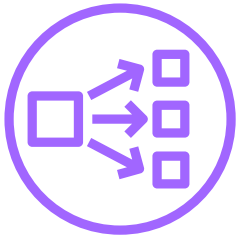
NLB Features

- Designed for supporting resources using the TCP and UDP protocol.
- Handle workloads that scale to millions of requests per second.
- Support static IP addresses for the load balancer.
- NLBs preserve the source IP and port of the ingress traffic.

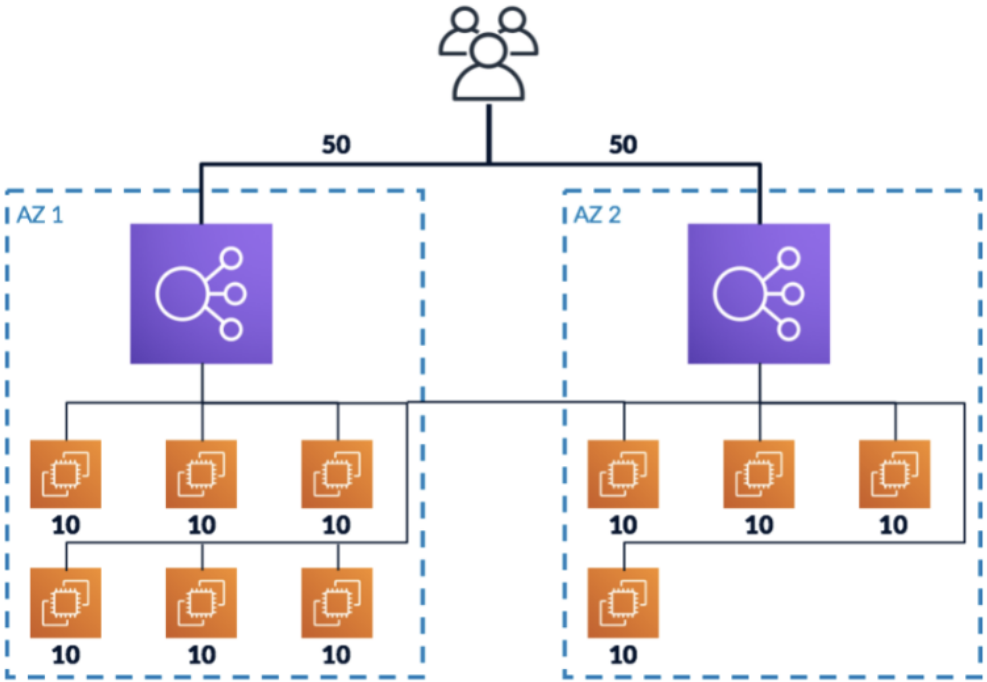


Network Load Balancer

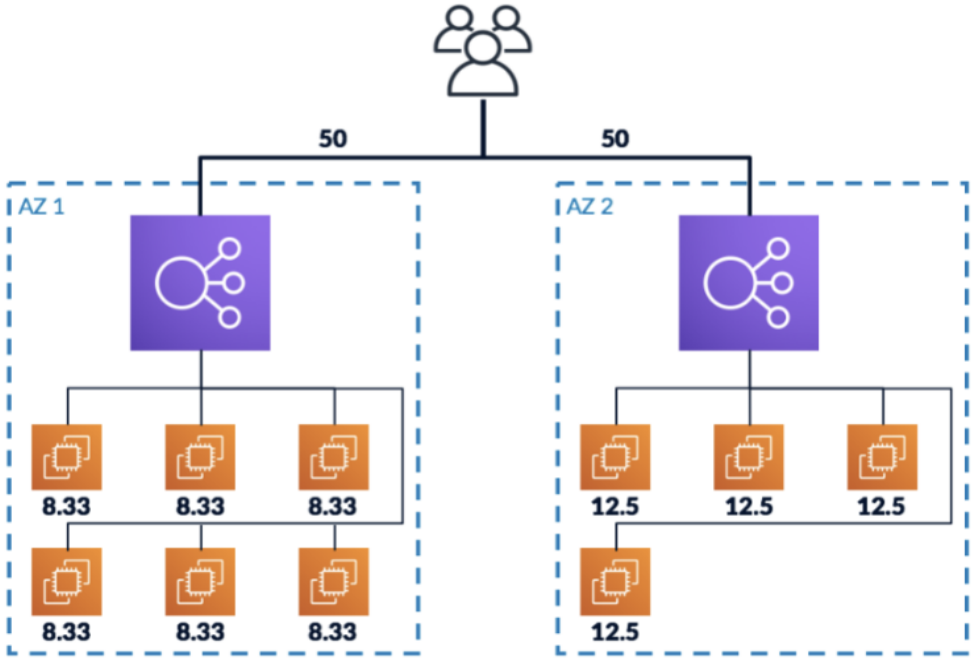
- The NLB doesn’t support HTTPS or the Web Application Firewall.
- End-to-end encryption uses TLS 1.3.
- Support connections from clients across VPC peering connections, Direct Connect and VPN connections.



Cross-Zone Load Balancing Enabled



Cross-Zone Load Balancing Disabled



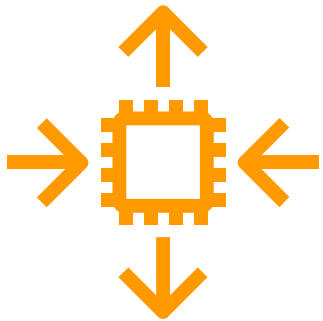
Auto Scale

Auto Scale Concepts

- Create collections of EC2 instances with Auto Scaling groups.
- Provide automatic horizontal scaling (scale-out) for your EC2 instances.
- Auto Scaling can span multiple AZs within the same AWS region.
- Auto Scaling integrates with ELB and CloudWatch.

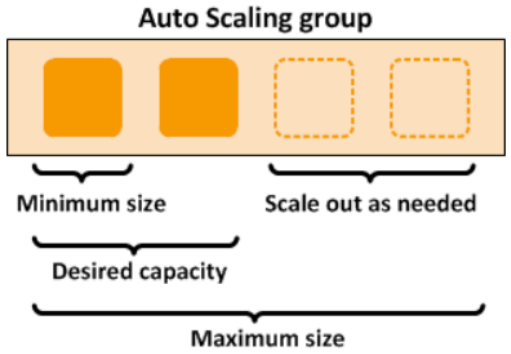
Auto Scaling Components

- Auto Scaling Groups – a managed group of EC2 instances.
- Launch Template – Configuration template for adding EC2 instances.
- Scaling Options – Auto scale groups scaling options.
- CloudWatch – Metrics and alarms that initiate scaling.



Setting Capacity Limits

- The minimum, maximum, and desired capacity are initially set to one.
- The ASG uses the maximum value to define the maximum number of instances that can be launched.
- The minimum size defines the number of instances deployed.

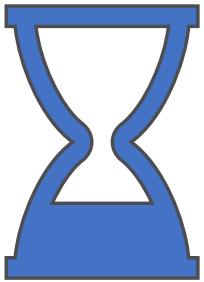


Scaling Policy Options

Scaling Policy	Details
Target tracking scaling	Increase or decrease the capacity of the ASG based on a target value for specific cloud watch metric.
Step scaling	Increase or decrease the capacity based on a set of scaling percentages.
Simple scaling	Increase or decrease based on a single scaling adjustment.

Predictive Scaling

- Create a load forecast – 14 days of history are analyzed for a specific load metric, and future demand is forecasted.
- Schedule scaling actions – Schedule the scaling actions that proactively add and remove resource capacity to match the load forecast.
- Maximum capacity – Automatically add additional resources when the forecast capacity exceeds the maximum capacity resources allocated for scaling.

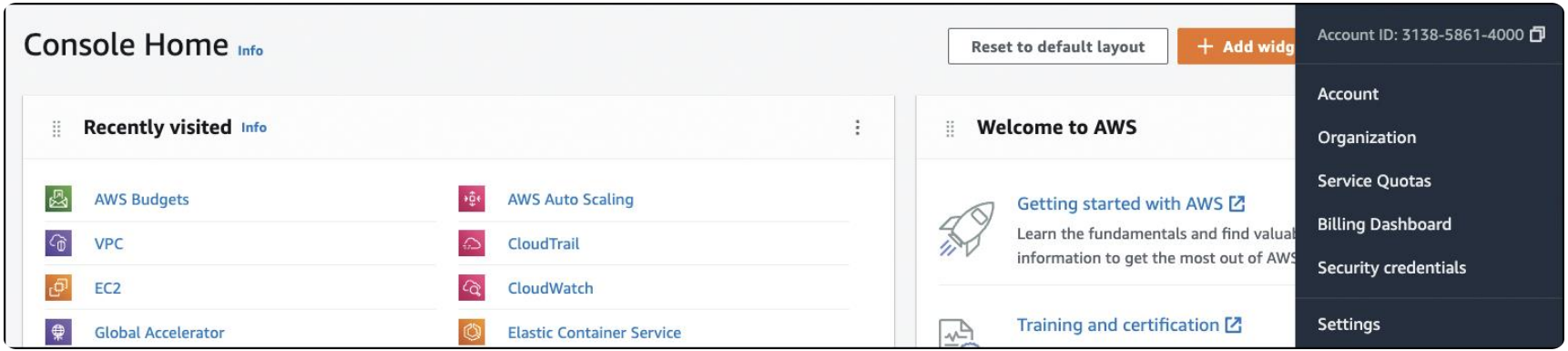


AWS Auto Scaling

- Configure and manage scaling for your stacked resources using a scaling plan.
- Dynamic and predictive scaling automatically scales AWS resources.
    - EC2 Auto scaling groups.
    - EC2 Spot fleet requests.
    - Amazon Elastic Container Service.
    - Amazon DynamoDB.
    - Amazon Aurora.

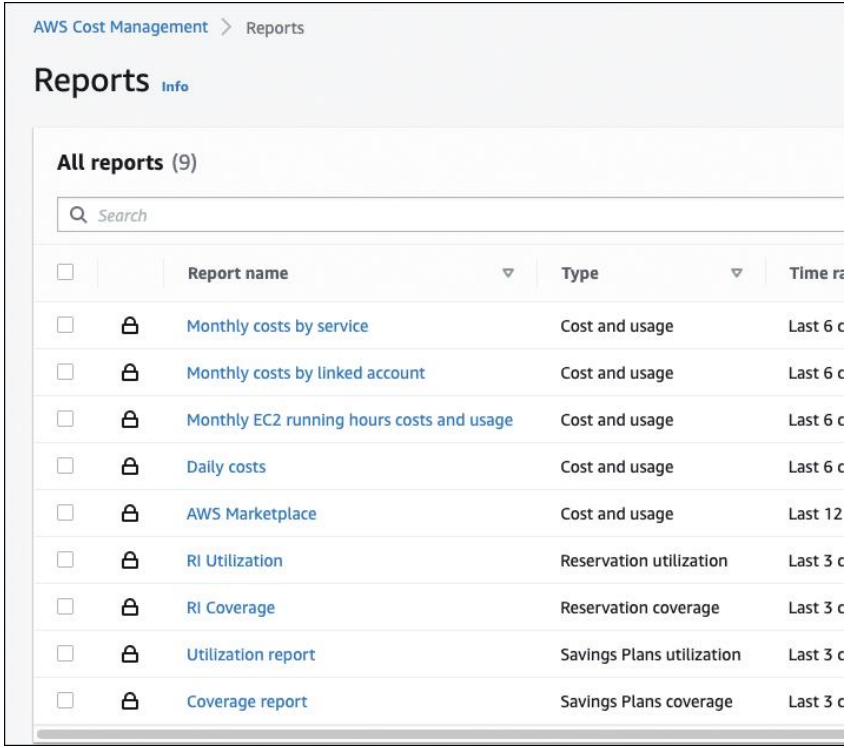
Pricing Tools

Billing Dashboard



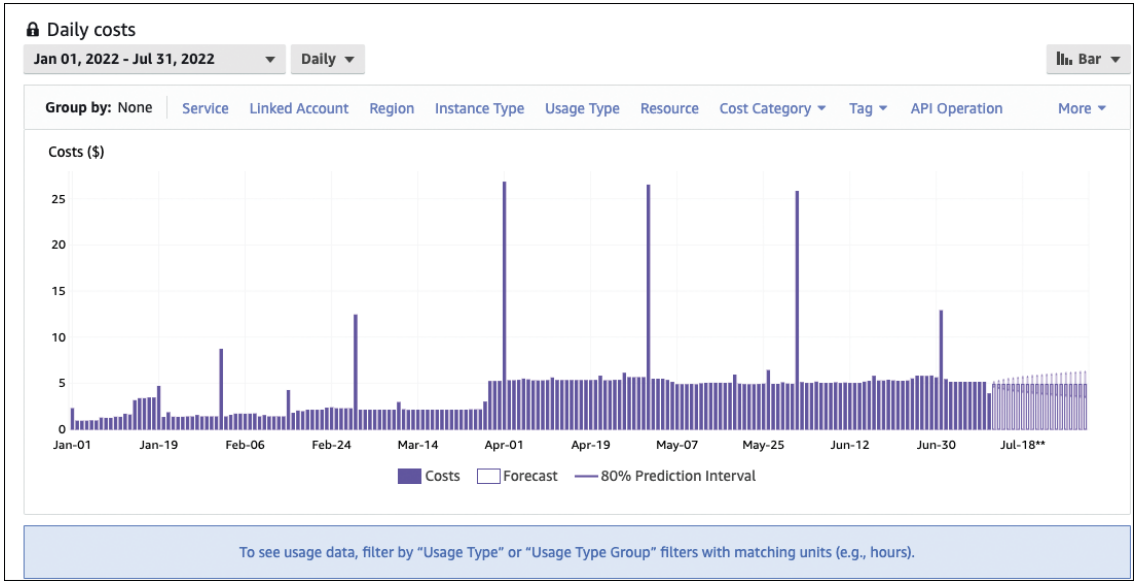
Cost Explorer

- Cost Explorer helps customers analyze AWS account costs.
- Reports breakdown AWS services that are incurring the most costs.
- Filter with numerous dimensions by AWS service and by region.
- AWS Organization can take advantage of consolidated billing and review the charges by member accounts.



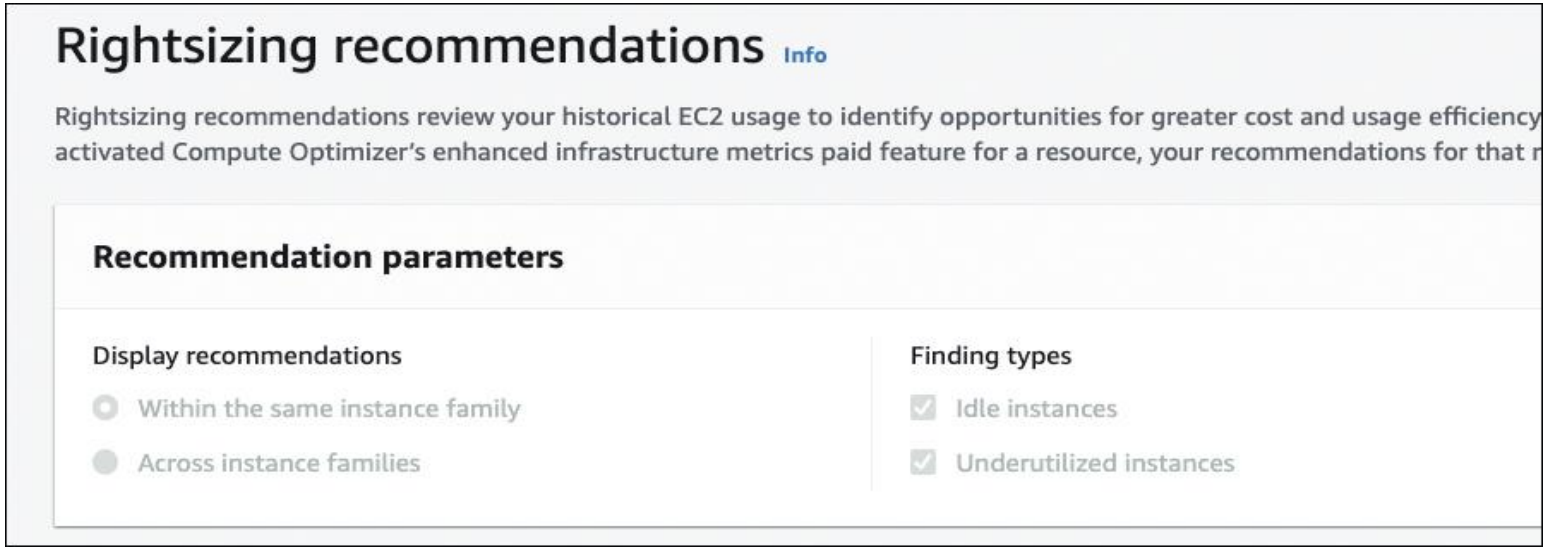
AWS Cost and Usage Reports

- Cost and Usage Reports (CUR) provide a comprehensive overview of the monthly costs and usage of AWS services per AWS account or AWS organization.
- Show hourly, daily, or monthly expenses based on resources used or defined resource tags.



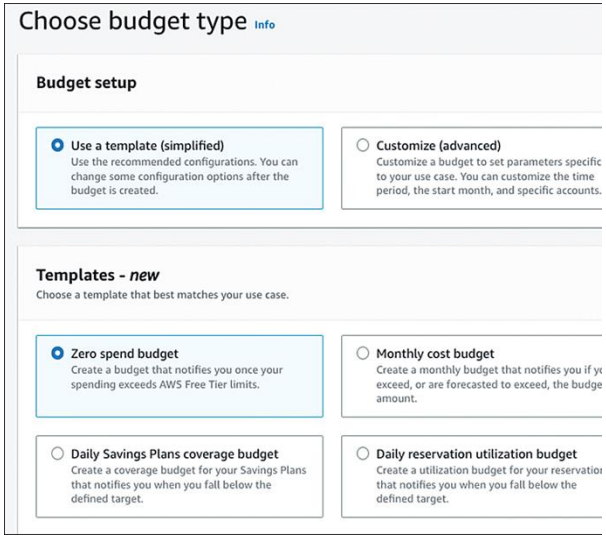
Rightsizing Recommendations

- Cost Explorer recommendations for improving the use of reserved instances and AWS resources.



AWS Budgets

- Budgets can monitor the overall utilization and coverage of your existing reserved instances or savings plans.
- Billing alerts to alert when costs are outside defined budget guidelines.
- Alert notifications can be sent to Amazon SNS topic / email addresses.

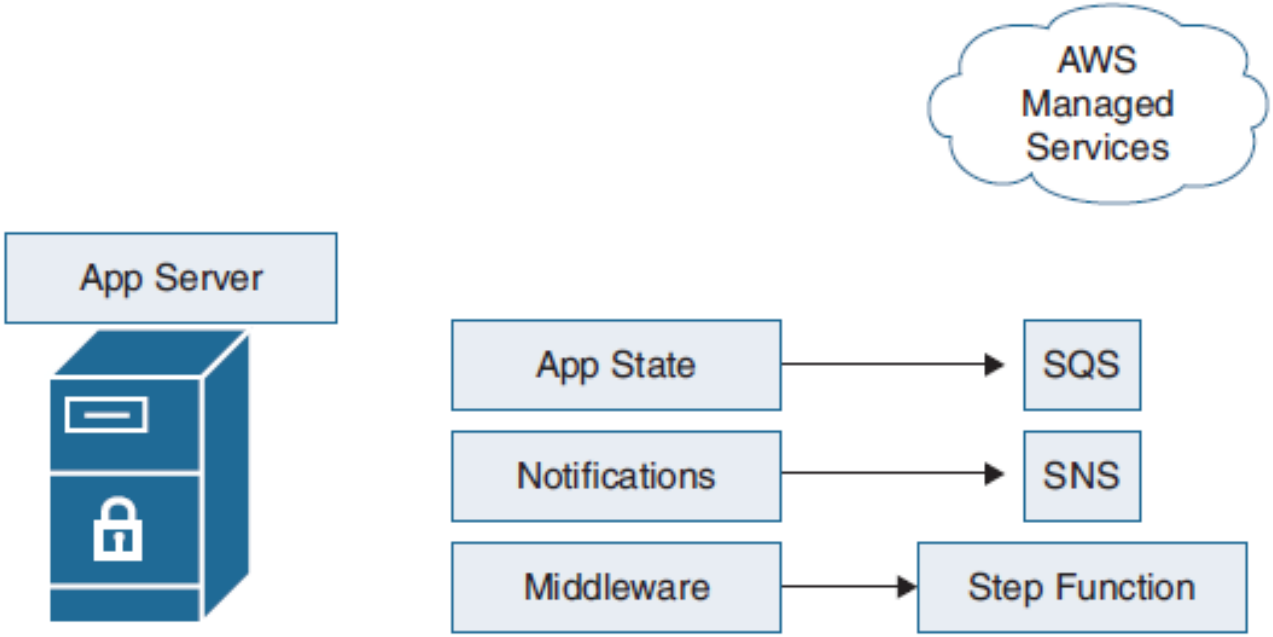


Application Integration Services

Application Integration Services

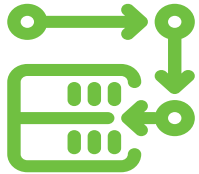
AWS Service	Purpose	Use case
Simple Notification Service (SNS)	Send notifications from AWS	Email, text, or Lambda
Step Functions	Visual workflow of AWS services	Order processing workflow
Simple Queue Service (SQS)	Messaging queue (FIFO, Standard)	Distributed applications
Amazon MQ	Message broker based on Apache MQ	Helps with migration from on premise message brokers

Integration Services



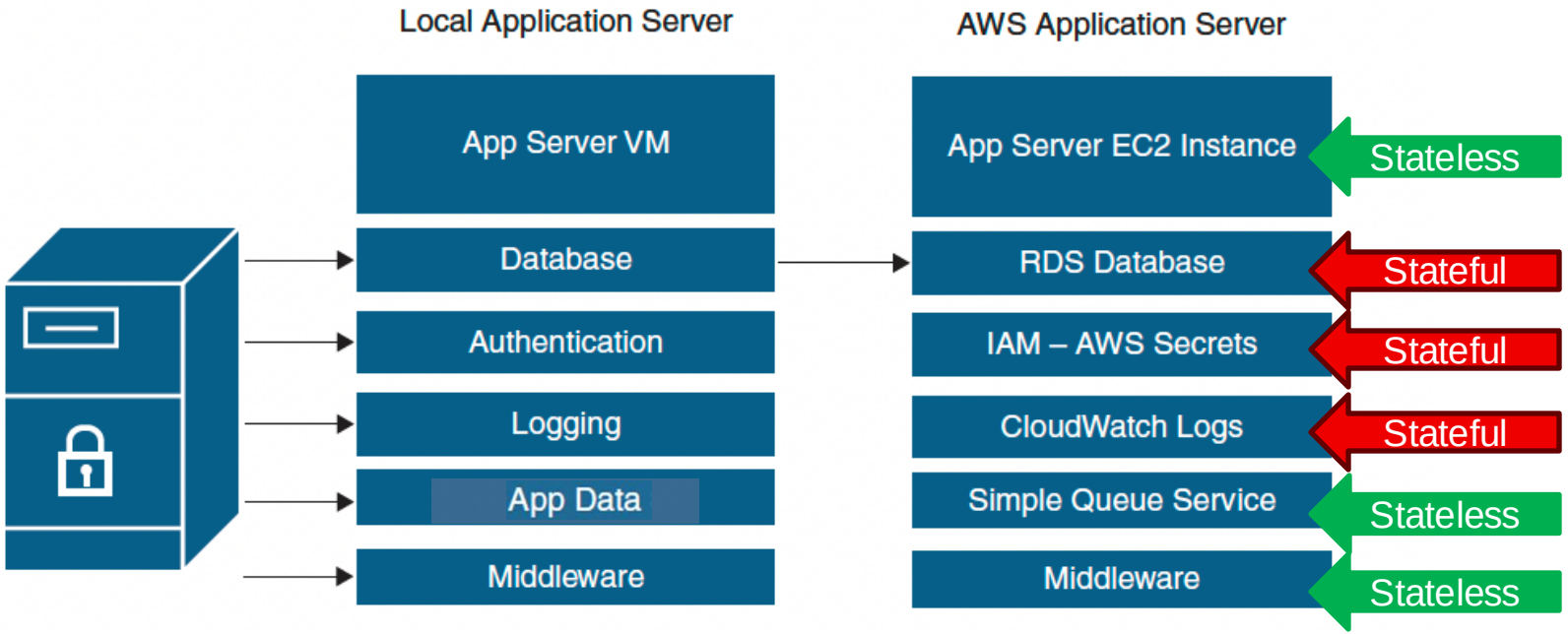
Stateless vs Stateful Services

- Stateful services retain information, or “state,” regarding previous interactions.
  - Online shopping carts that keep track of purchases.
  - Banking systems that track account information.



- Stateless services operate without knowledge of previous interactions.
  - RESTful API requests contain all required information for processing.
  - Stateless applications perform specified activities without storing state information.

On-Premises vs. AWS Hosted



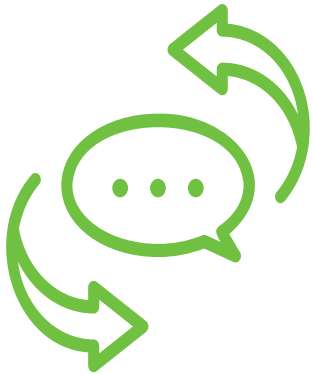
Amazon SQS

Amazon SQS

- Amazon Simple Queue Service (SQS) is a fully managed message queue.
- Send, store, and receive messages between applications, services, and servers.
- Decouple communications between distributed workload components such as applications and microservices.
- Use cases include system-to-system messaging and request offloading.

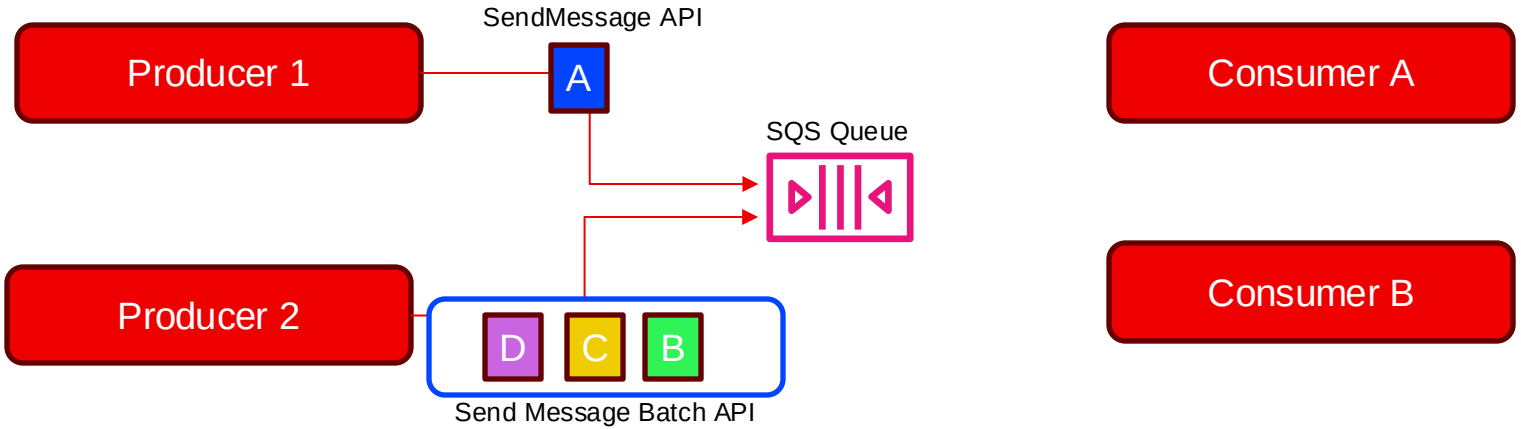
Amazon SQS

- Each message is a discrete data unit sent from one application or service to another using the queue.
- Messages are stored in SQS queues and can be sent and received by applications, services, and servers.
- Messages can be stored in SQS queues for up to 14 days; the default is 4 days.



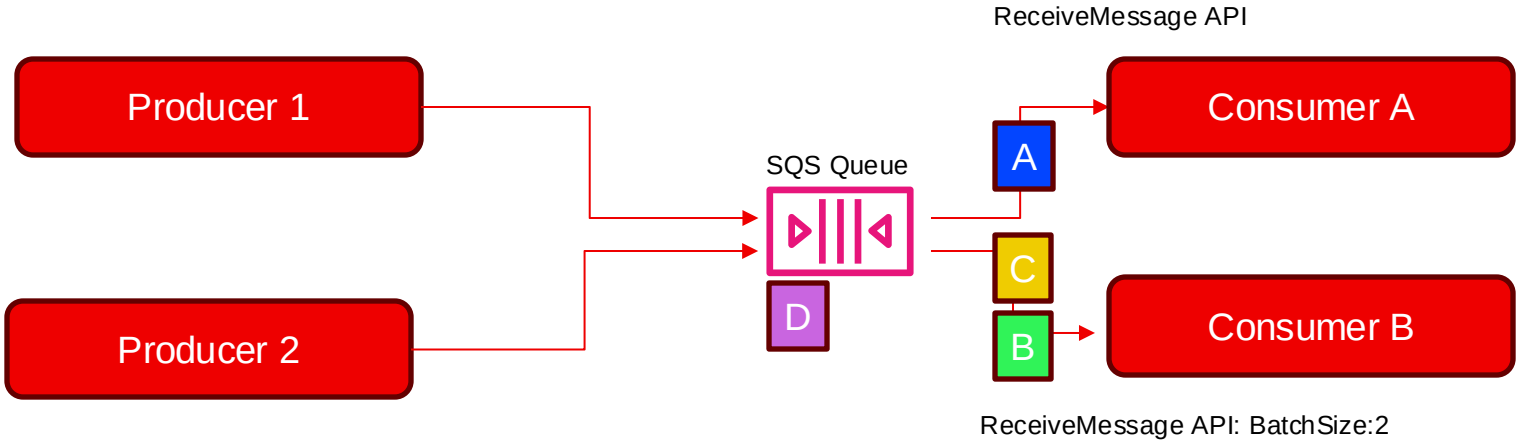
Sending Messages

- SendMessage API or SendMessage Batch API is called by the message producer.



Receiving Messages

- SendMessage or SendMessage API is called by the message producer.



SQS Components

Standard Queues

Best-effort message ordering with at least once delivery.

Use case: Decoupling application processing, sending notifications.

FIFO Queues

Preserves the order in which messages are sent and received. Each message is delivered exactly once.
Use case: Financial,microservices.

Message Polling

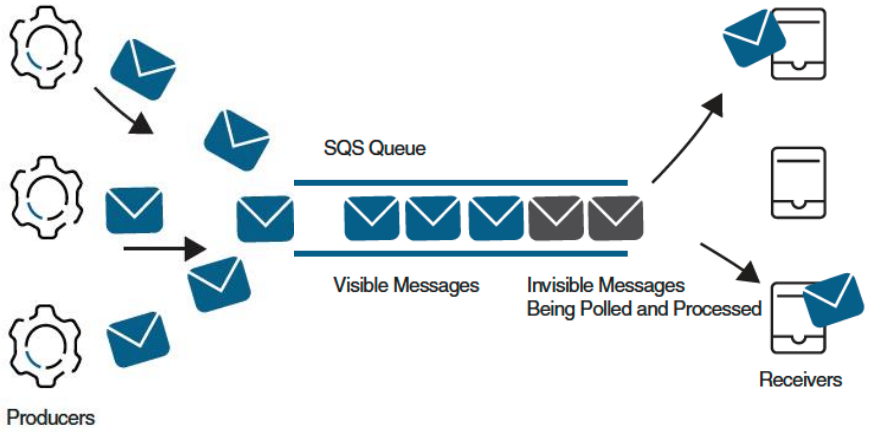
Default short polling.

Optional long polling.

Long polling can help reduce the number of empty received message responses returned.

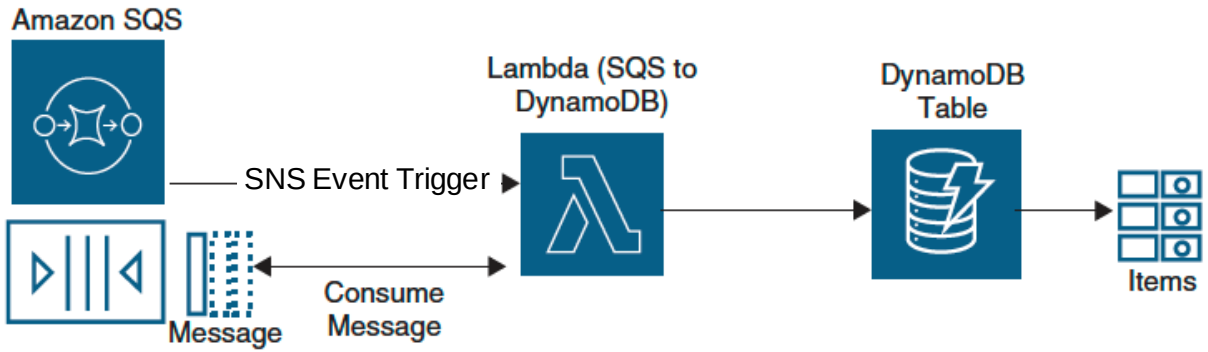
Visibility Timeouts

- During message processing of a message, the message is not visible.
- After processing, messages are deleted from the queue.



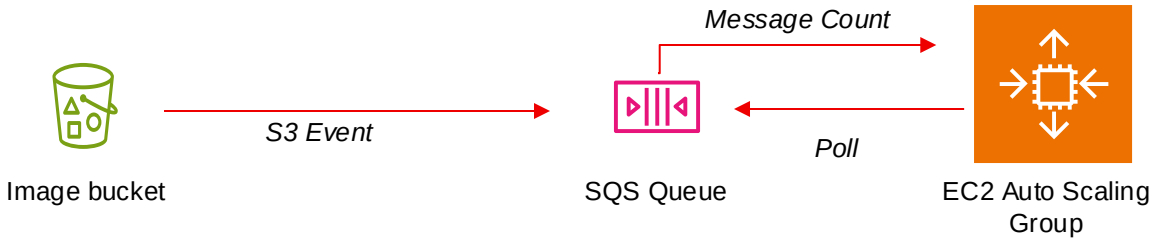
SQS Message Processing

- AWS Lambda receives notifications from SNS when there are messages to be processed.



SQS Message Processing

- EC2 Auto Scale: Scale up when messages in the SQS queue increase.
- The number of instances/containers in the worker pool can be scaled according to the number of messages in the queue.



SQS Components

- **Standard queues**: This default queue type has the best-effort ordering of messages, which can be delivered at least once.
- **FIFO (first-in, first-out) queues:** This queue type preserves the order in which messages are sent and received, and each message is delivered exactly once. FIFO queues should be used when the order of operations and events is essential.
- **Polling:** Applications can receive messages from an SQS queue using either the default short polling method or the long polling method.
- **Dead-letter queue (DLQ):** Messages are stored in the DLQ after the maximum number of processing attempts have not been completed.
- **Visibility timeout:** During the processing of a message, there is a period where each message being processed is not visible.

Amazon SNS

Amazon SNS

- Enables applications and devices to send and receive notifications from applications and services.
- Publishers, applications or AWS services send messages to topics.
- Topics are access points to which publishers send messages asynchronously.
- Clients subscribe to SNS topics to receive published messages.

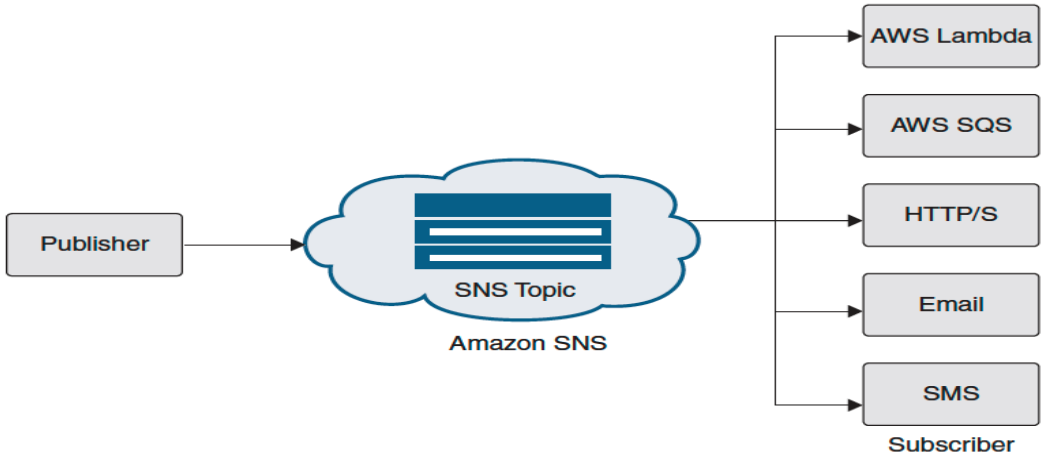
Amazon SQS

- Messages can be application-to-application messaging:
  - Kinesis Data Firehose delivery streams, Lambda functions, SQS queues, and HTTP/S endpoints.

- Messages can also be application-to-person:
  - Push notifications for mobile applications, phone numbers, and email addresses.

SNS Integration

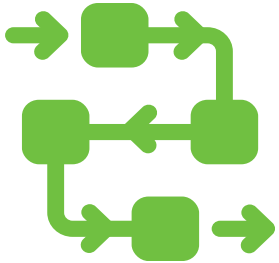
- Amazon SNS is integrated with Amazon CloudWatch.
- Alerts, alarms, and SNS notifications should be part of workload design.
- More than 70 AWS infrastructure services have CloudWatch metrics, allowing detailed monitoring and alerting.



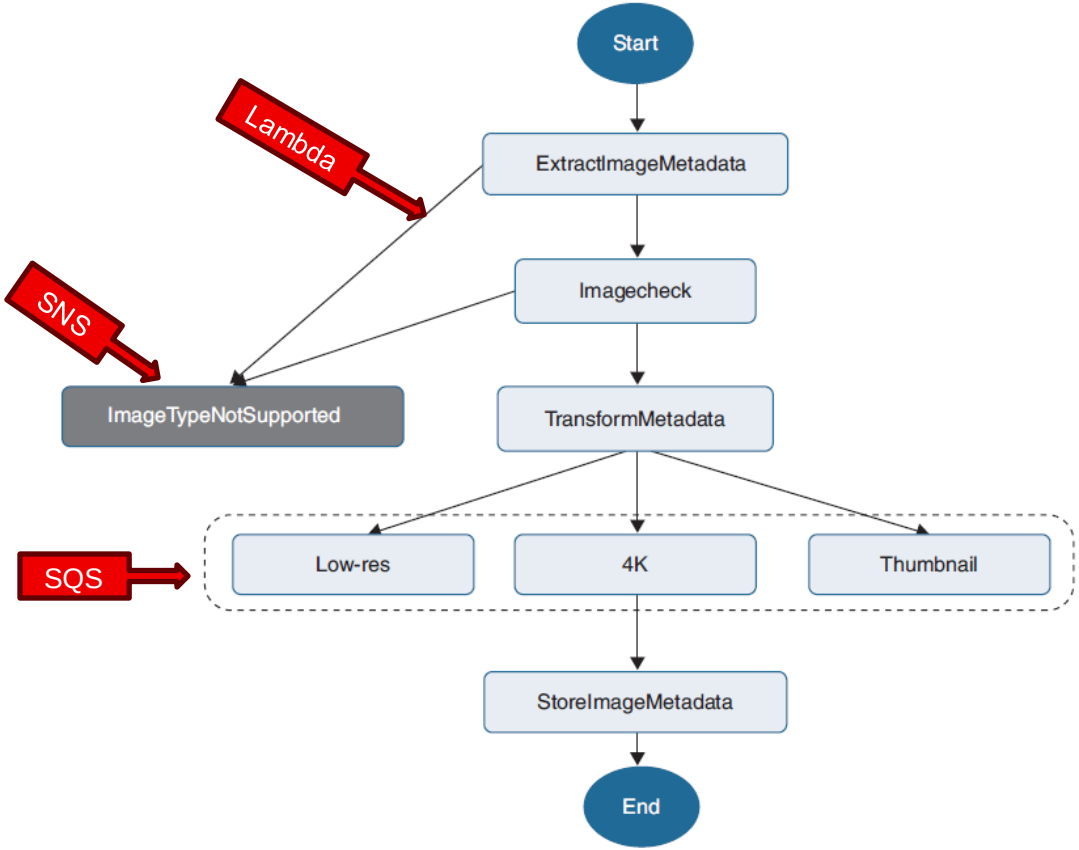
Amazon Step Functions

AWS Step Functions

- Developers can coordinate the execution of multiple AWS services and components using visual workflows.
- Integrates with Amazon EC2, Amazon ECS, AWS Lambda, Amazon SQS, and Amazon SNS.
- Each workflow has checkpoints that maintain workflow and its state throughout each defined stage.
- The output of each stage in the workflow is the starting point of the next stage.



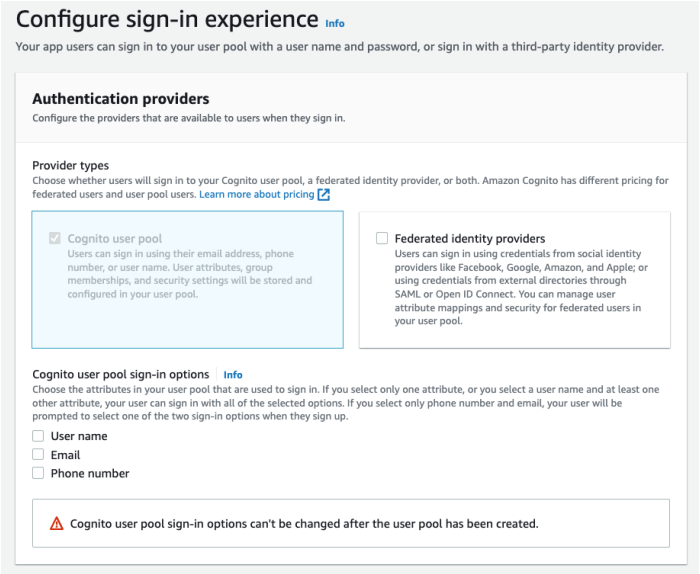
Step Function Workflow



Amazon Cognito

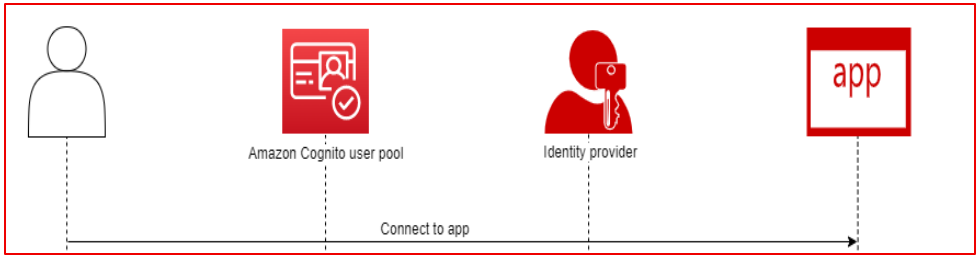
Amazon Cognito

- Amazon Cognito provides authentication, authorization, and user management for web and mobile applications hosted at AWS.
- End users sign in using either a user pool or a federated identity provider.



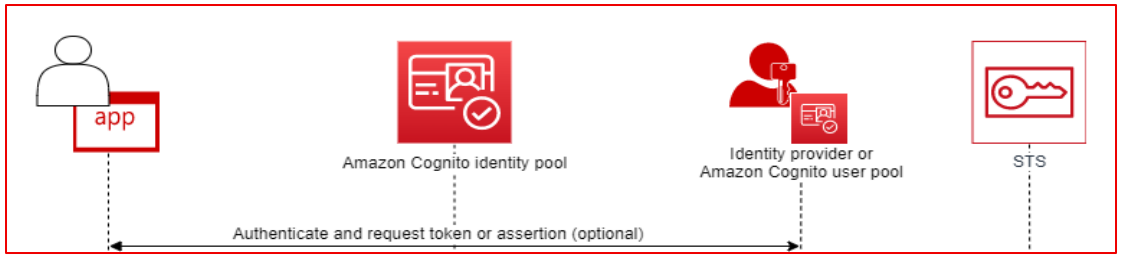
User Pool

- Used to authenticate and authorize users to an application or API.
- Independent user directory, OIDC identity provider, service provider of 3rd party SAML 2.0 and consumer identities (Apple, Google or Facebook).

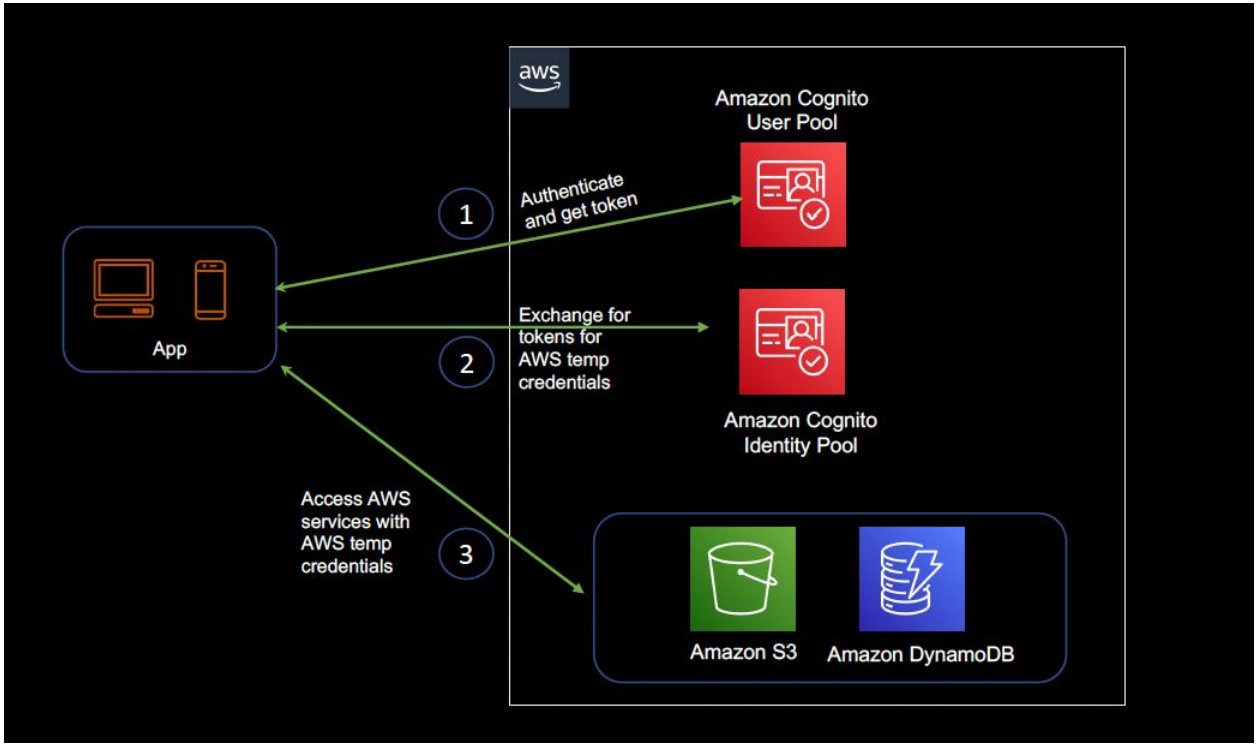


Identity Pool

- Used to authorize authenticated users access to AWS resources.
- AWS credentials are issued from the Cognito identity pool.

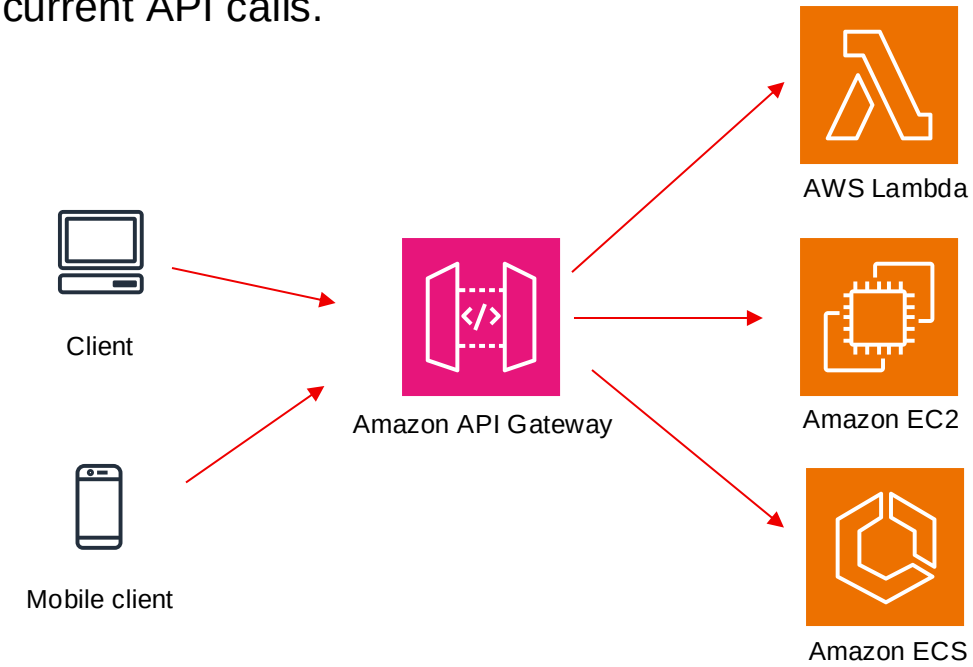


User Pool and Identity Pool Together



API Gateway

- API Gateway acts as a front door for applications to access data, business logic, or functionality from back-end AWS services.
- API Gateway handles all the tasks of accepting and processing up to hundreds of thousands of concurrent API calls.



API Types

REST APIs	WebSocket APIs	HTTP APIs
• Resources and methods integrated with backend HTTP endpoints, Lambda functions, and AWS services. • Clients send requests to an AWS service through the API gateway; the service responds synchronously.	• APIs are accessed using WebSocket protocol. • Full-duplex communication allows real-time data exchange. • Chat, gaming, and financial trading applications.	• REST APIs and HTTP APIs are both RESTful API products; HTTP APIs are cheaper. • Integrate with Lambda functions or to a routable HTTP endpoint. • Works with private VPC endpoints (ALB/ NLB)

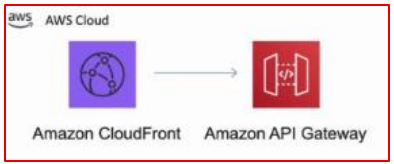
API Gateway in Operation

- API Gateway supports RESTful, WebSocket, and HTTP APIs.
- Amazon API Gateway exposes HTTPS endpoints.
- API calls can also include traffic management, authorization, access control, monitoring, and API version management.
- Traffic spikes can be managed with throttling; backend operations can withstand traffic spikes.

Endpoints

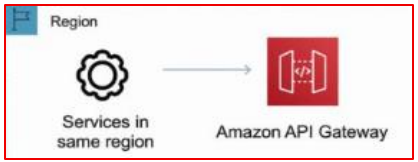
Edge - Optimized

API requests are routed to the nearest CloudFront Point of Presence (POP).



Regional

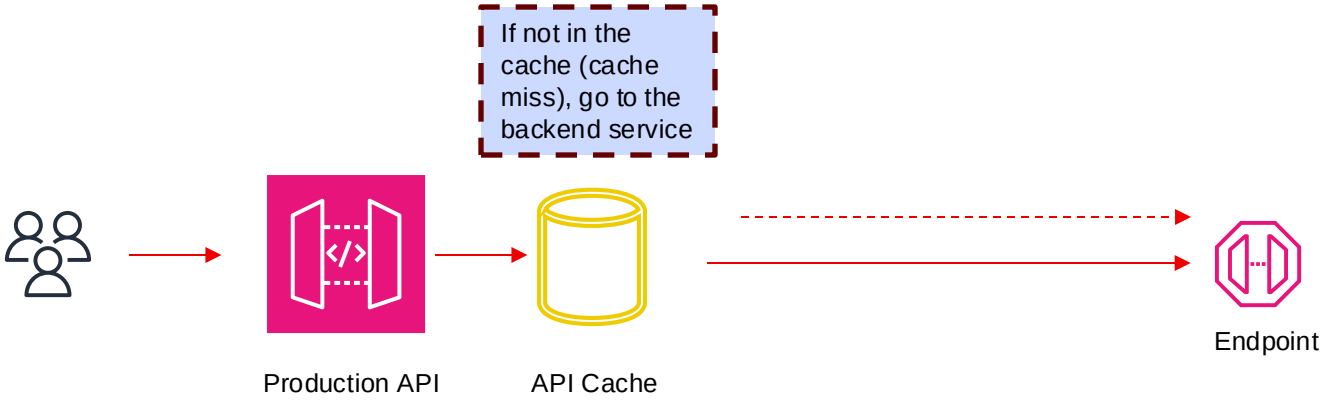
A regional API endpoint is accessed by clients in the same AWS region.



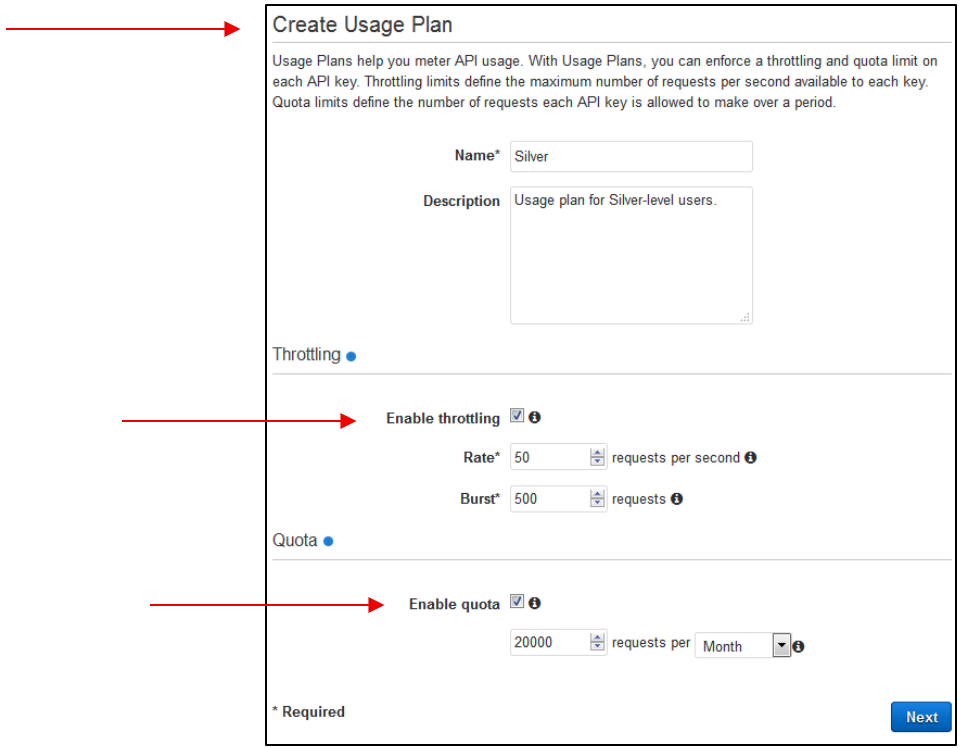
Private

A private API endpoint is an API endpoint accessed from a VPC using an interface VPC endpoint.

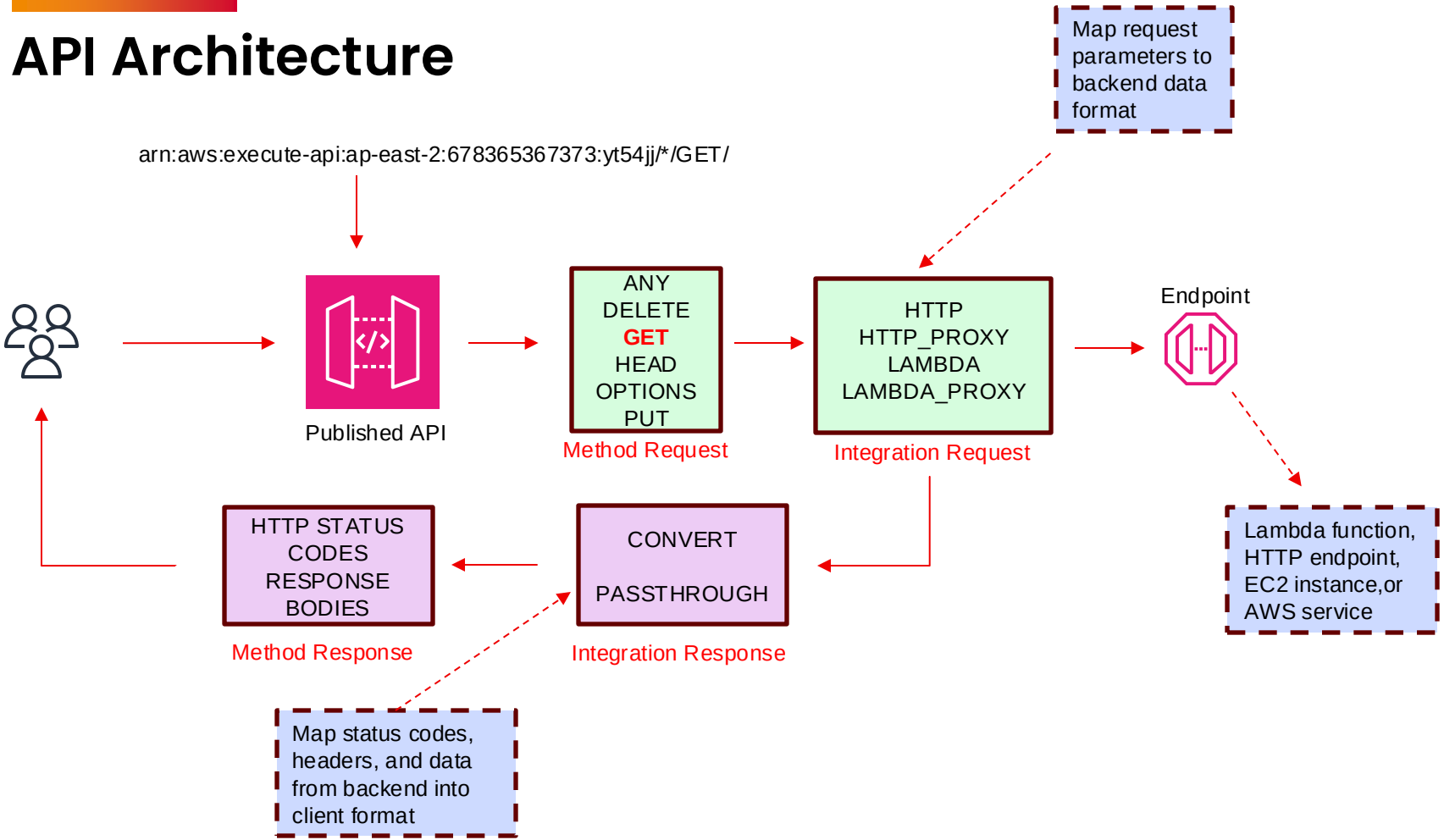
API Gateway Caching



API Throttling with Usage Plans



API Architecture



Next Steps

- Read Well-Architected documentation (Security, Performance, Reliability, Cost)
- Answer supplied test questions.
- Sign up for AWS Skill Builder.
- Take the exam!