O'REILLY

Amazon Certified Solutions Associate Prep

Steps for AWS Certification Success

- Think like a Cloud Architect.
- Architects “build” (i.e., design) “construct.
- Architects propose solutions based on existing building blocks.
- The Associated Architect is based on common sense.
- Every question is a “situation” ; current or proposed.
- The correct answer is the best answer based on suggested answers to the multiple-choice question.

AWS Documentation

- AWS FAQ’s.
- AWS Quick Starts – Automated application stacks.
- Well-Architected Framework PDFs.
- Well-Architected Framework labs.

AWS Well-Architected Framework

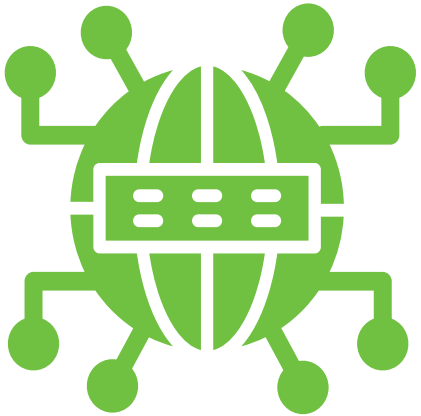
- The exam is based on the AWS Well-Architected Framework.
- Deploy and manage your workload following best practices.
- Guidance for resilient, stable, and efficient workloads.
- Each pillar has hands-on labs.
- AWS Well-Architected Tool.

Re-Certification

- Every three years you must recertify.

Domains of Knowledge（SSA-C03）

- Designing Resilient Architectures　30 %
- Design High-Performance Architectures　26 %
- Secure Applications and Architectures　24 %
- Design Cost-Optimized Architecture　20 %

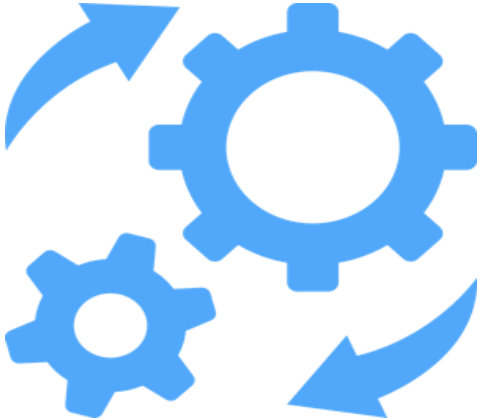


How the Exam is Graded

- Questions are multiple-choice.
- Both single selection and multiple selection.
- No penalty for guessing.
- 65 questions on the exam.
- 15 questions are beta questions.
- Mark questions for future review.
- Answer all questions.
- Sketch out your answers using the supplied plastic sheet and marker. (Test center)

Design and Deployment

- Understand use cases.
- Know how each AWS service works.
- Be able to sketch use.
- Hands-on experience.
- Don’t memorize syntax.
- Read the FAQs.



Domain 1

Design Secure Architectures

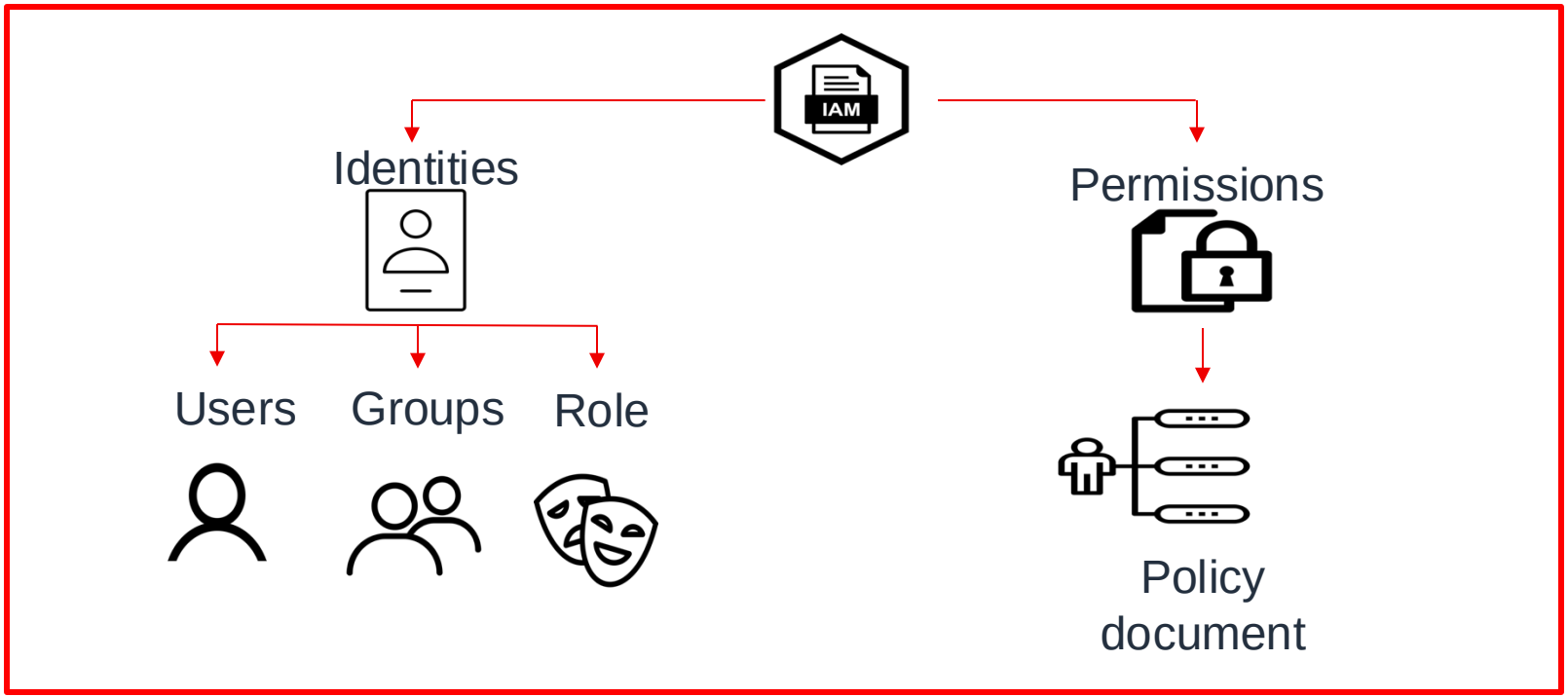
Task Statement 1.1

Design secure access to AWS resources.

Design secure access to AWS resources

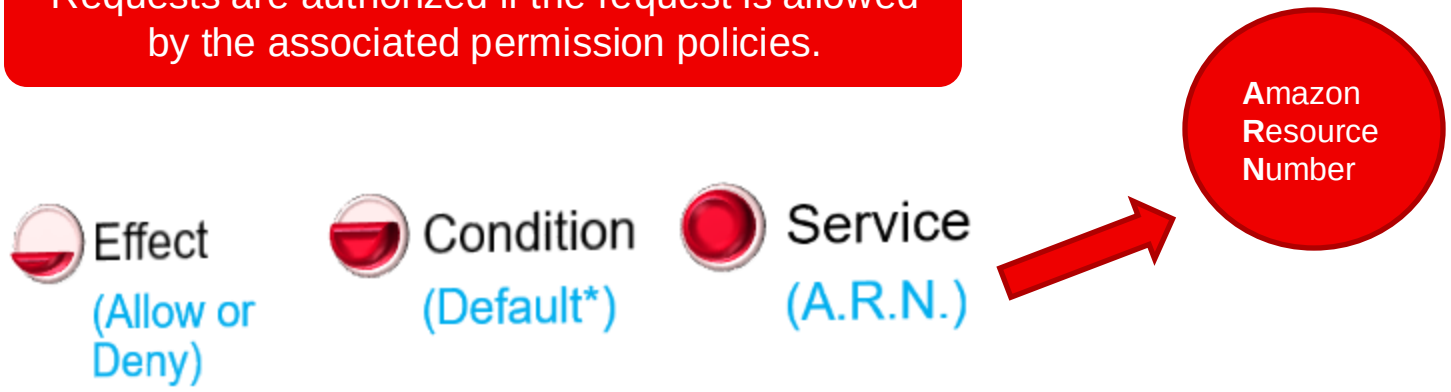
- AWS IAM (Users, Groups, Roles, Policies)
- AWS Organizations (Control Tower, AWS IAM Identity Center)
- The AWS shared responsibility model.

Identity and Access Management



IAM Authorization

Requests are authorized if the request is allowed by the associated permission policies.



Conditions Define limits, rather than *

Multi-factor authentication

Google

YubiKey

AWS

Authy

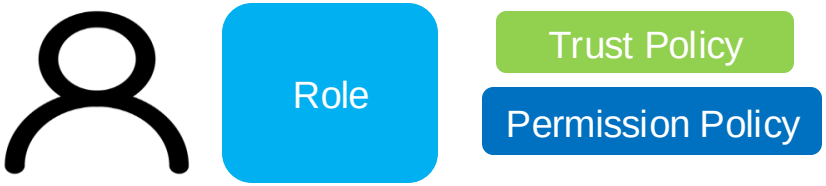
MFA adds a second layer of authentication using an OTP. (a one-time-password)

Username + Password + One-time-use code

1. Something you know.
2. Something you have.

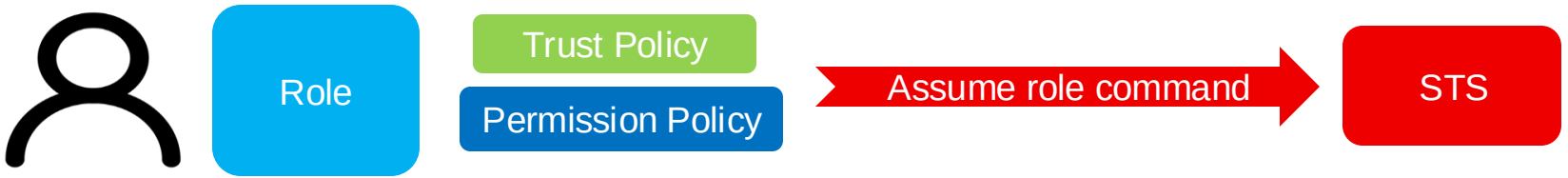
IAM Roles

- Roles are used to provide temporary delegation of permissions.
- Externally authenticated users, or IAM users in another AWS account.
- Role delegation allows each identity to perform the approved actions listed in the permissions policy.



Role Execution

- The user / application is been assigned a trust policy.
- The trust policy is attached to the role.
- The IAM user or IAM Role has permission to run the Assume role command at STS.



Cross AWS Account Access

Production AWS account has live application.

Development AWS account has test application.

- **Problem:** Developers need to update live applications in the production account.
- **Solution:** Use roles to delegate access to production resources for developers.

Production AWS account is the Trusting account.

Development AWS account is the Trusted account.

IAM roles allow developers to access resources in production account

STS Operation

- The AWS Security Token Service (AWS STS) is available as a global service.
- All STS requests go to a single endpoint at https://sts.amazonaws.com.
- Regional AWS STS endpoints can also be chosen for faster access speeds.
- AWS STS endpoints can be activated in the console for AWS regions that are not activated by default.
- AWS STS can also be disabled in an AWS region.

Security token lifetime can be set from 1 to 36 hours

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).

Central control over AWS services and API requests.

Granular control over users and roles.

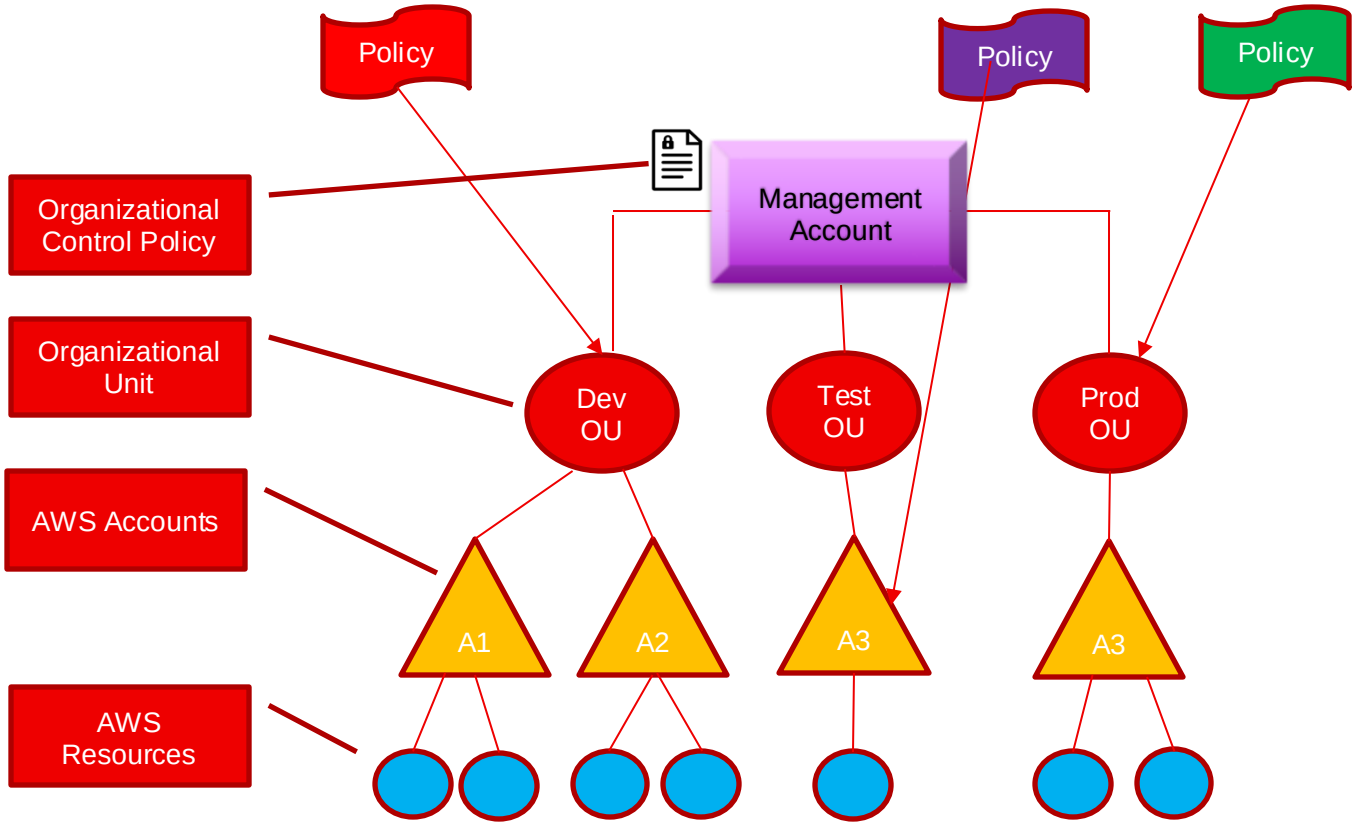
Standardize with consistent tags across resources.

Integration with AWS Services.

Configure automatic backups for resources.

Global access from all AWS regions.

AWS Organizations



Control Tower

- Manage and govern a secure AWS Organization deployment.
- Integrated with AWS Organizations, AWS Service Catalog, and AWS IAM Identity Center.
- Automate AWS account deployment and configuration in an AWS Organization.
- Provides organizations with the ability to 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

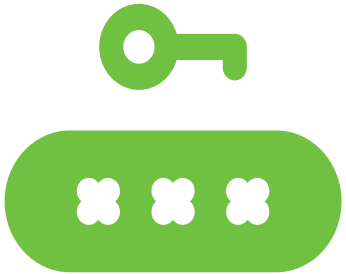
- Manage sign-in security for user identities.
- Connect and centrally manage access across AWS Organization accounts.
- Connect and manage AWS accounts, individual IAM users and Roles, and Identity Centre-enabled applications.

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.



Task Statement 1.2

Determine Secure Workloads and Applications

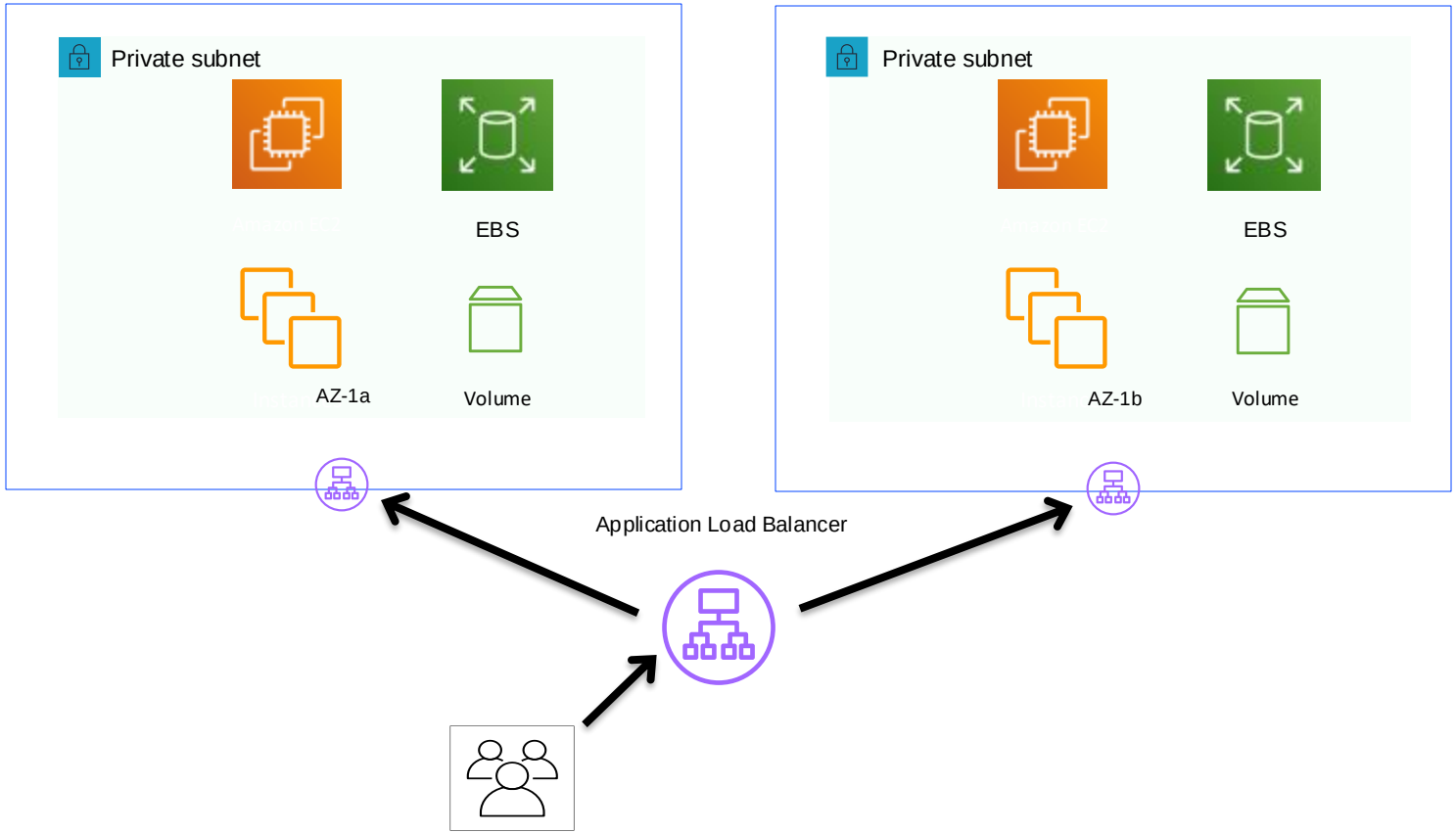
Determine Secure Workloads and Applications

- Security groups, Route tables, Network ACLs, NAT gateways
- VPN, Direct Connect
- Secure application access (AWS Shield, AWS Secrets Manager)
- Amazon Cognito, Amazon GuardDuty, Amazon Macie)

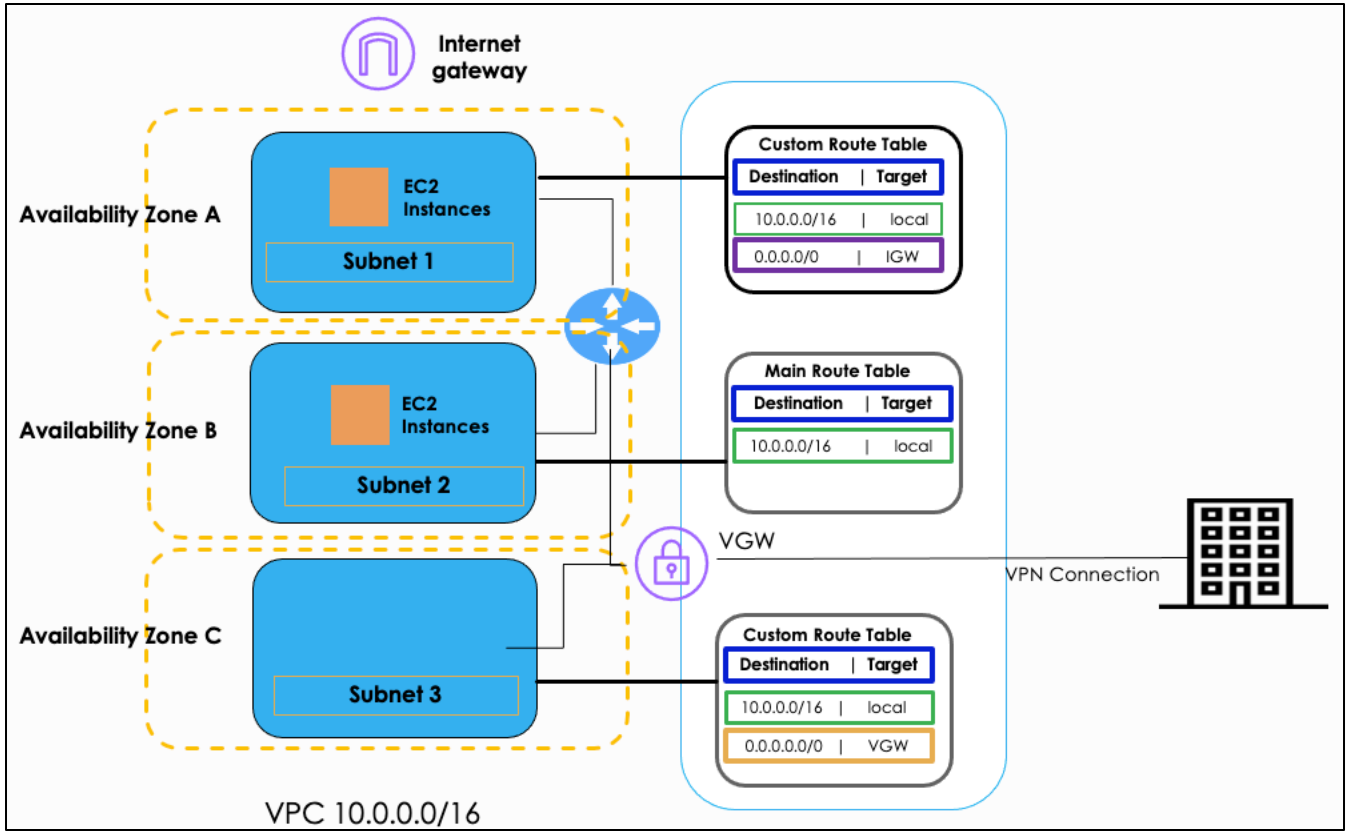
Virtual Private Cloud

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

Failover Possibilities



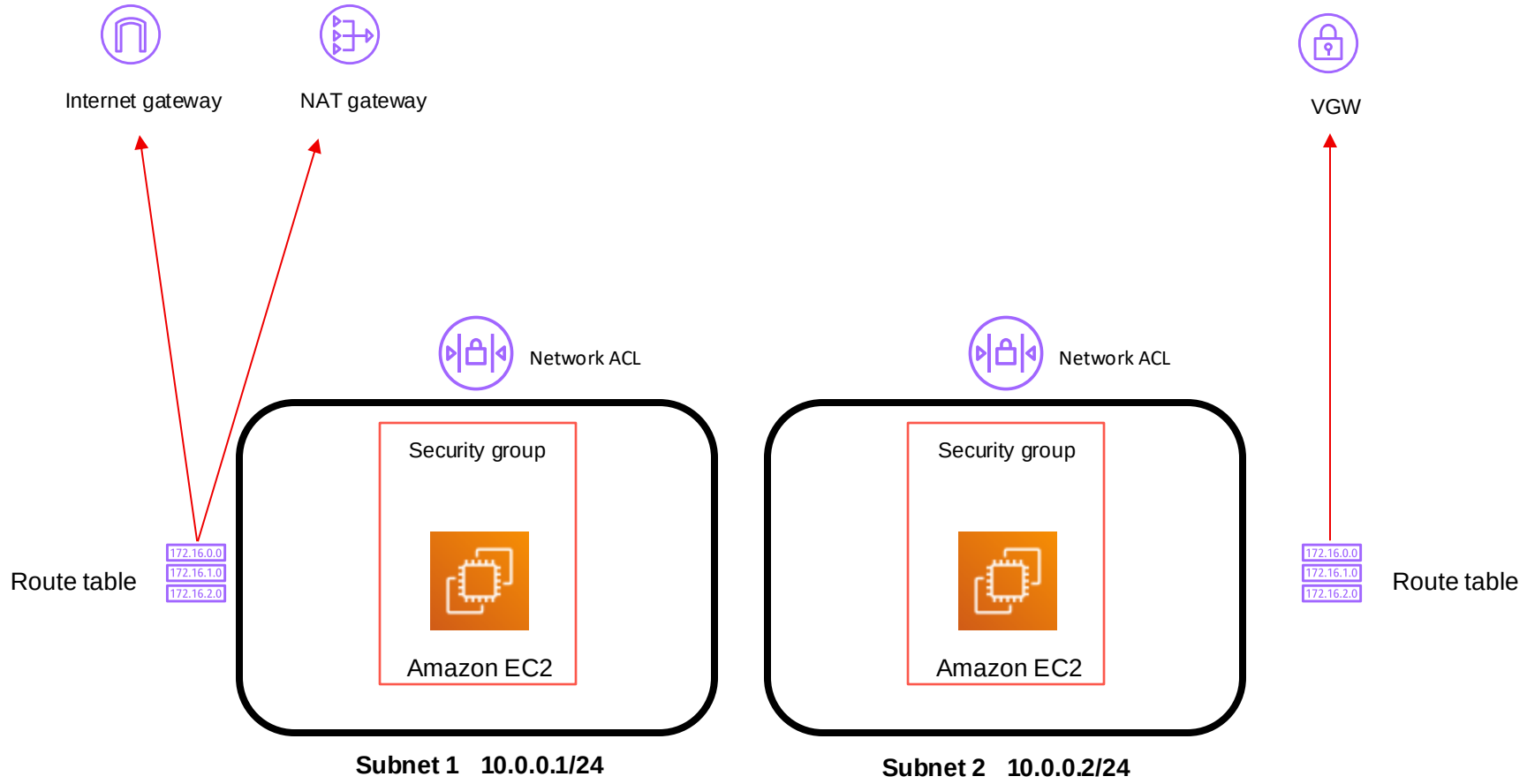
Route Tables



Network Access Control List

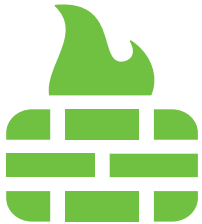
- NACLs control traffic in and out of the subnet.
- Each subnet must be associated with a NACL.
- NACL rules are defined as stateless.
- Rules are evaluated in order.
    - Inbound Rule - Allow or deny the specified traffic pattern
    - Outbound Rule - Allow or deny the specified traffic pattern

- Use NACLs for blocking IP address ranges.

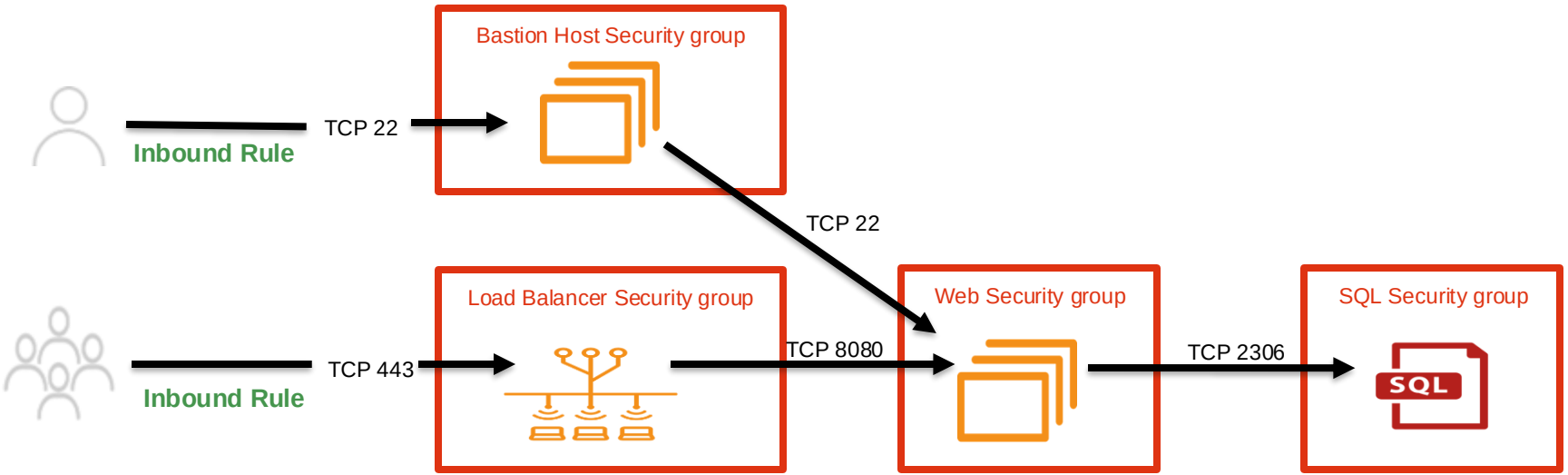

Internet gateway
NAT gateway
VGW
Network ACL
Network ACL
Security group
Security group
Amazon EC2
Amazon EC2
Route table
172.16.0.0
172.16.1.0
172.16.2.0
172.16.0.0
172.16.1.0
172.16.2.0
Route table
Subnet 1 10.0.0.1/24
Subnet 2 10.0.0.2/24

Security Groups

- EC2 instance firewall uses allow 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.
- Any TCP protocol that is defined with a standard port number.
- Requests made inbound are allowed outbound.
- A security group can be a source or destination for other security groups.

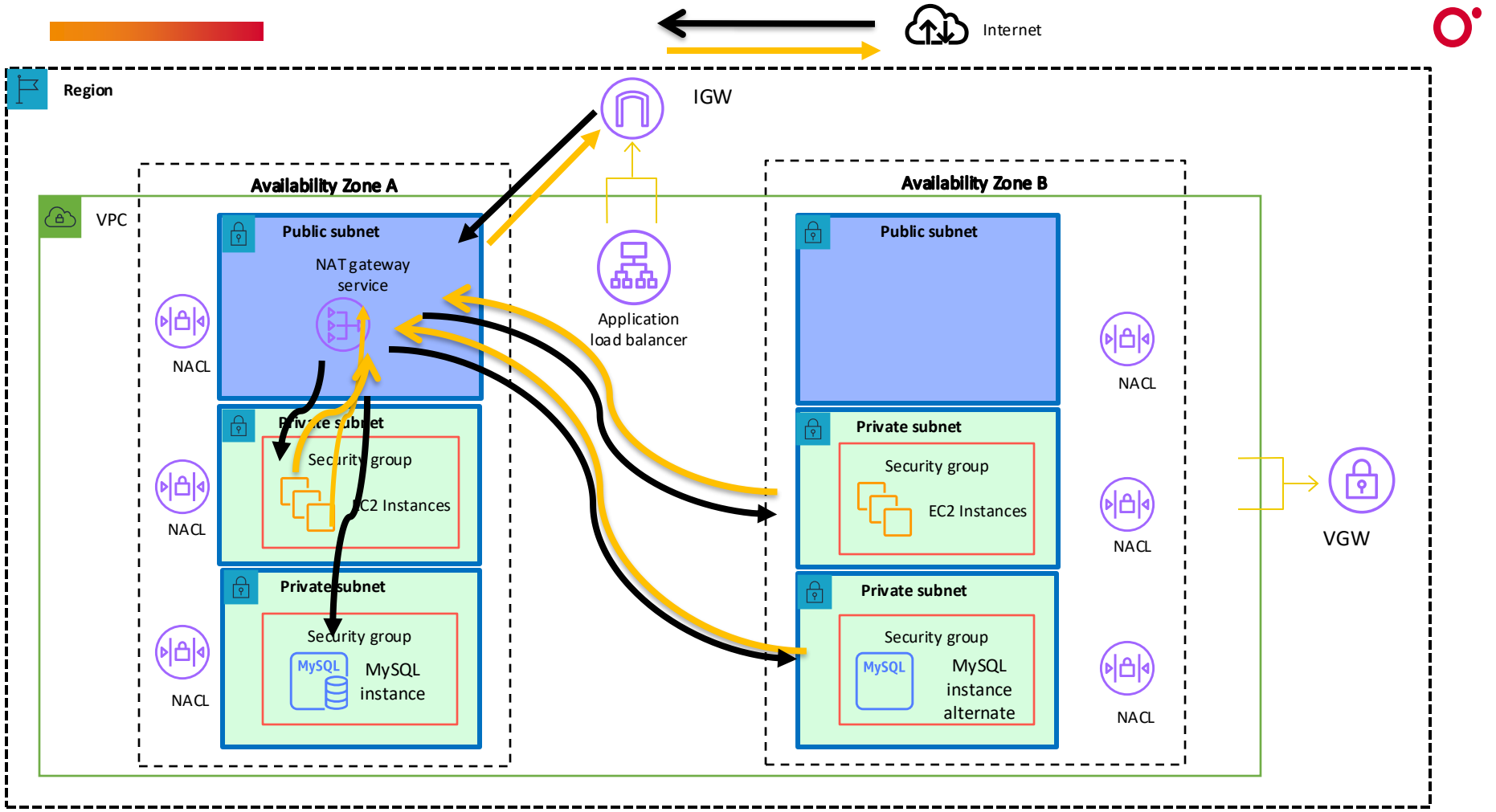


Security Group Design



NAT Gateway Services

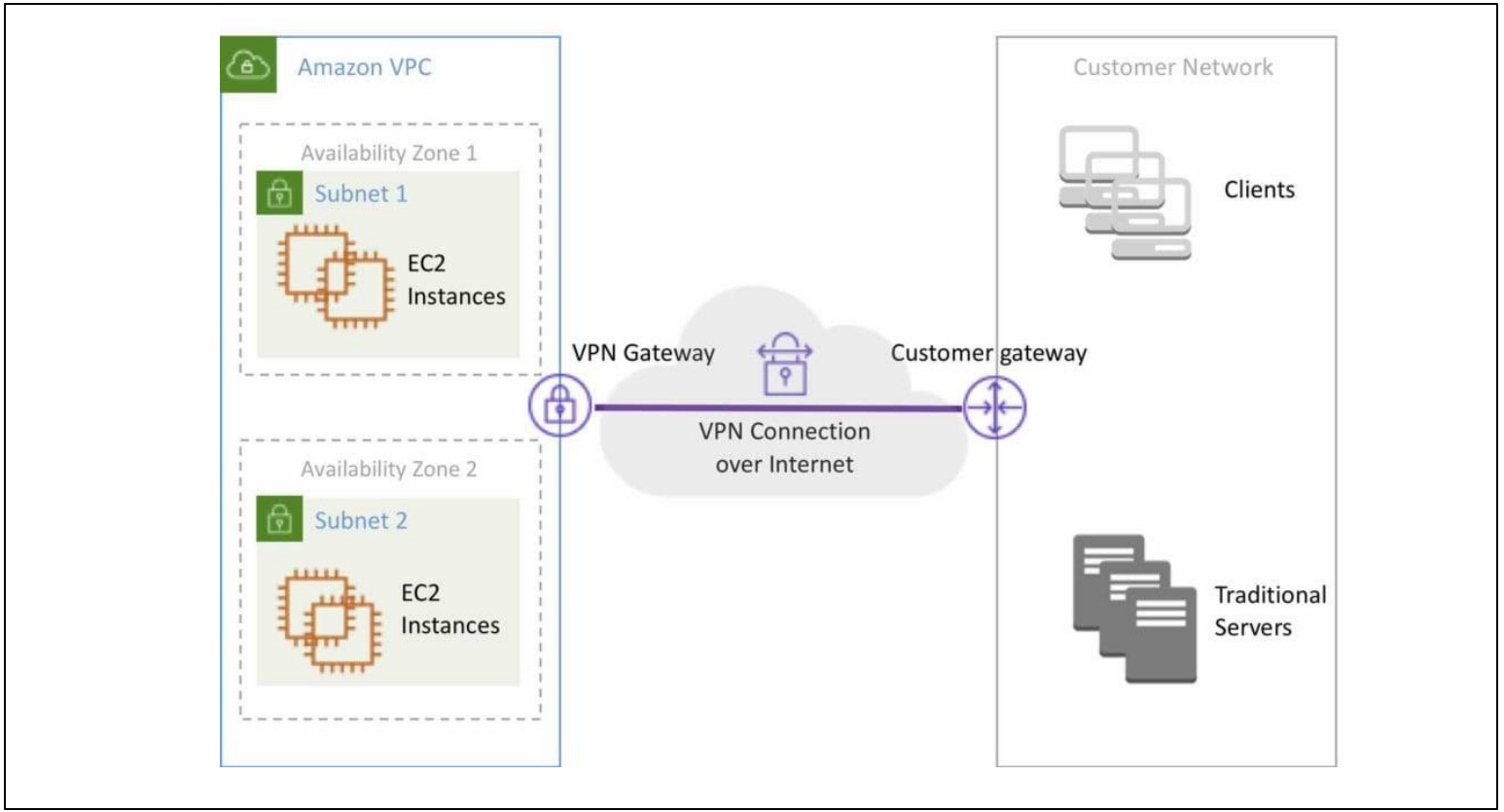
- NAT service enables instances in private subnets to connect to the Internet to get updates.
- Traffic requests are forwarded to the NAT service hosted in the public subnet.
- Internet response is sent back to the instance that made the request
- NAT Options:
    - NAT gateway service provided by AWS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

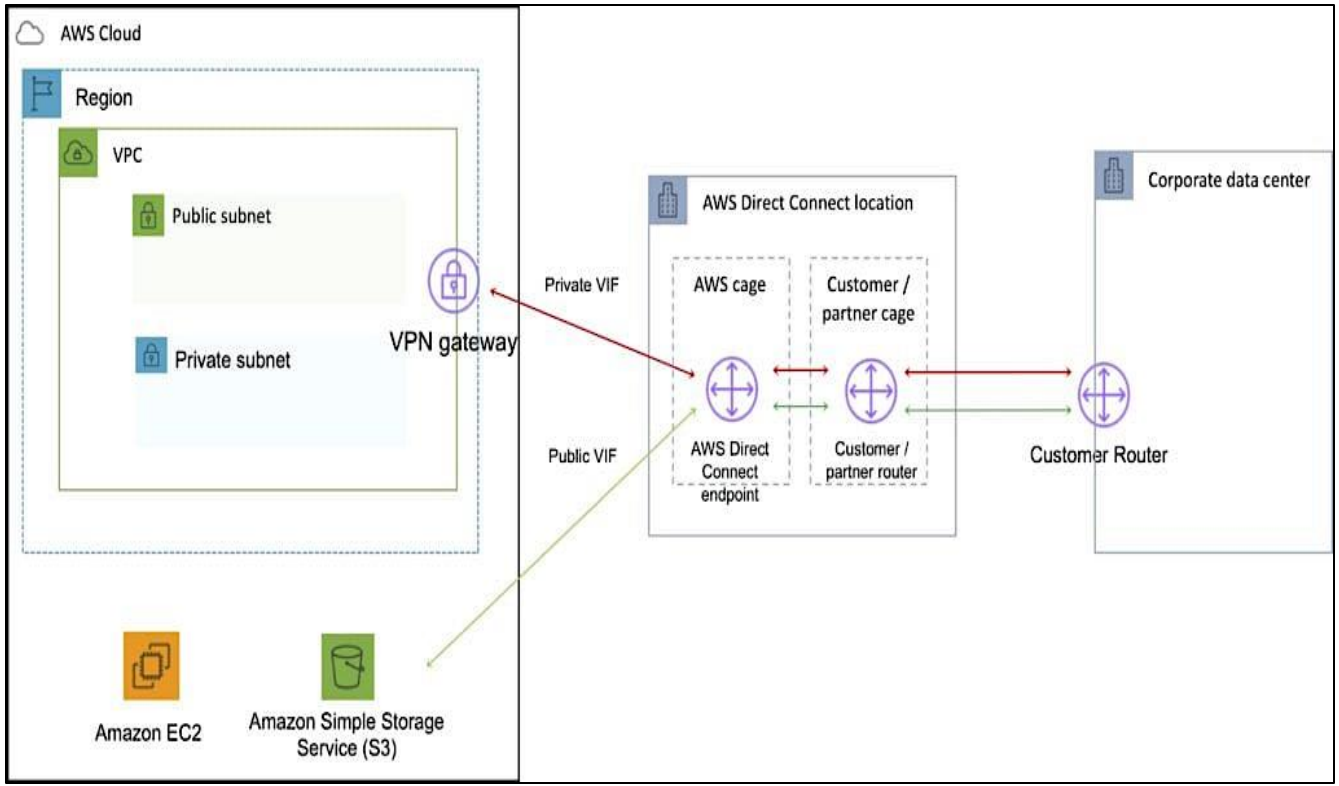
Egress-only Internet Gateway

- 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.

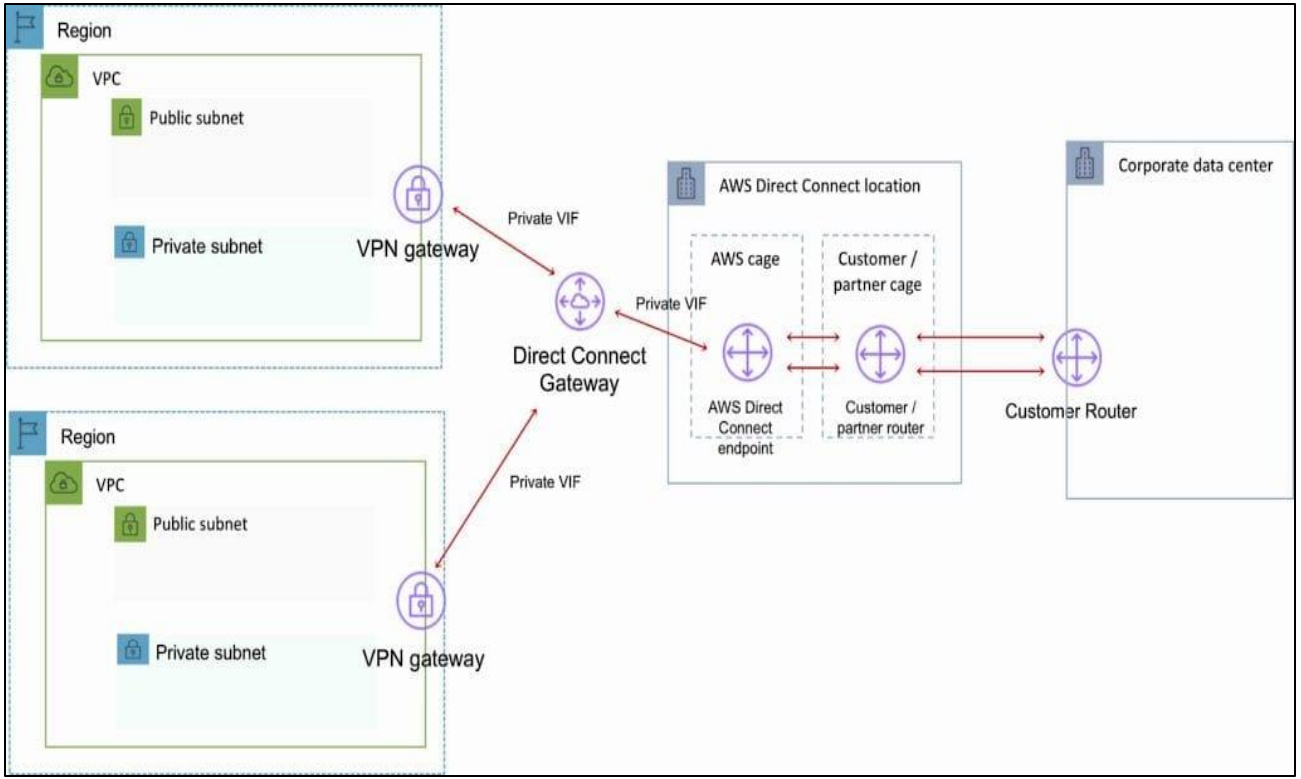
Managed VPN


Amazon VPC
Availability Zone 1
Subnet 1
EC2
Instances
Availability Zone 2
Subnet 2
EC2
Instances
VPN Gateway
VPN Connection
over Internet
Customer gateway
Customer Network
Clients
Traditional
Servers

AWS Direct Connect

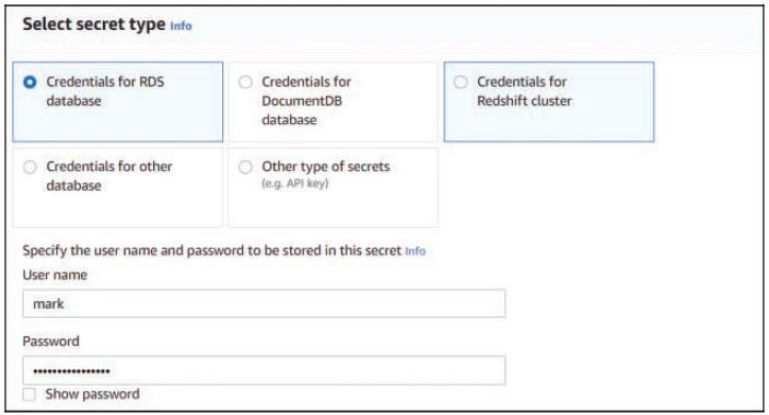


AWS Direct Connect Gateway



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.

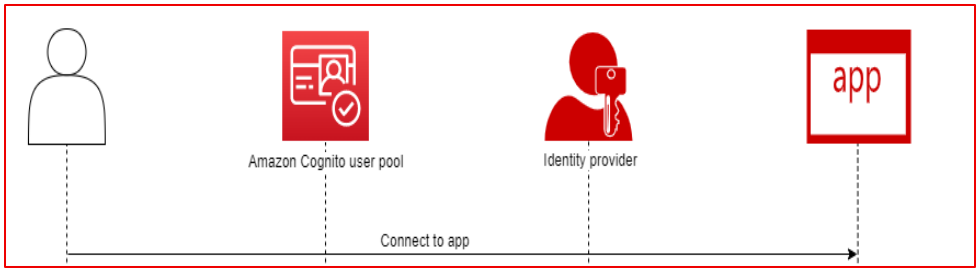


What is Amazon Cognito?

- Identity platform for web and mobile applications.
- User directory (pool) to authenticate and authorize users to an application.
- Identity pools provide AWS credentials for access to AWS Services.

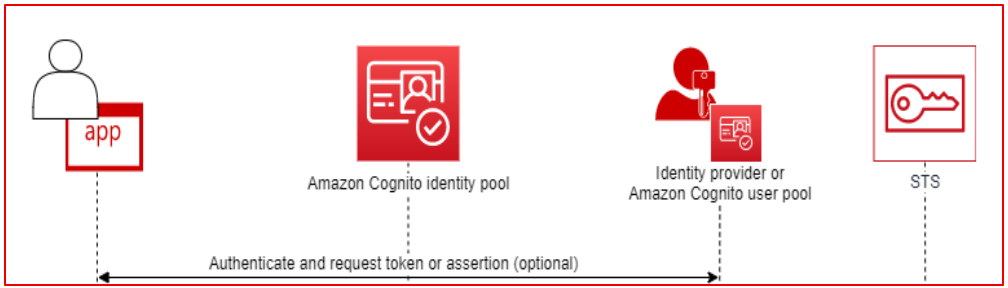
User Pool

Used to authenticate and authorize users to an application or API.



Identity Pool

Used to authorize authenticated users access to AWS resources.



AWS GuardDuty

CloudTrail

CloudTrail APIs

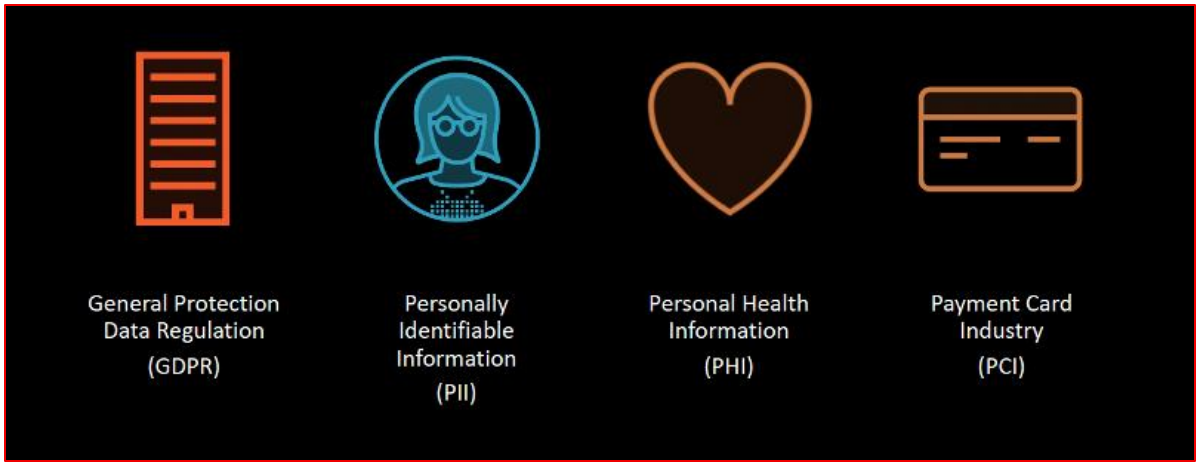
VPC Flow Logs

Route 53 logs and requests

S3 bucket access

Amazon Macie

- Automated security classification of S3 objects and access patterns.
- Monitor S3 for anomalies.
- Proactive data loss through continuous data visibility.
- Custom reports and alert management.



Task Statement 1.3

Determine Appropriate Data Security Controls

Determine Secure Workloads and Applications

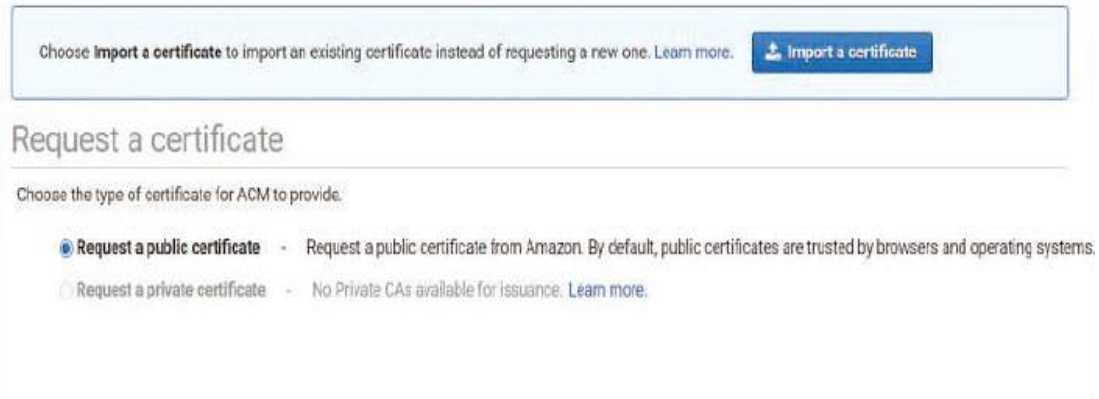
- AWS Key Management Service
- AWS Certificate Manager
- AWS Backup

AWS Key Management Service

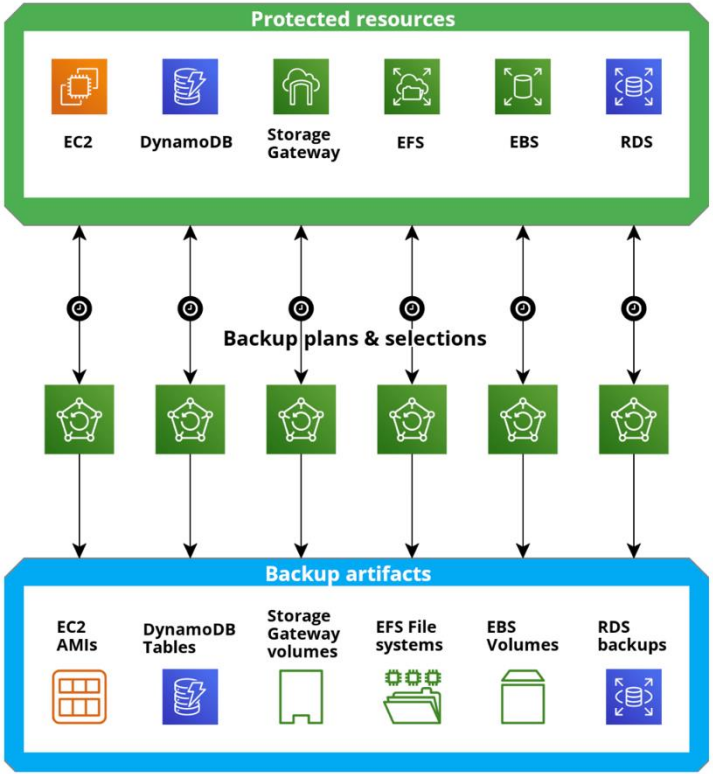
- Manage encryption/decryption of AWS data services.
- Centrally store your customer master key (CMK).
- CMK can be generated using KMS, AWS CloudHSM cluster, or imported.
- Unique data keys are used for each encryption request.
- KMS stores multiple copies of encrypted keys with 99.999999999% durability.

AWS Certificate Manager

- Provision, manage, and deploy public and private SSL / TLS certificates used with AWS services and AWS-hosted websites and applications.
- Certificates can be deployed on ELB load balancers, Amazon CloudFront distributions, AWS Elastic Beanstalk, and APIs hosted on Amazon API Gateway.



AWS Backup


Protected resources
EC2
DynamoDB
Storage Gateway
EFS
EBS
RDS
Backup plans & selections
Backup artifacts
EC2 AMIs
DynamoDB Tables
Storage Gateway volumes
EFS File systems
EBS Volumes
RDS backups

Domain 2

Design Resilient Architecture

Task Statement 2.1

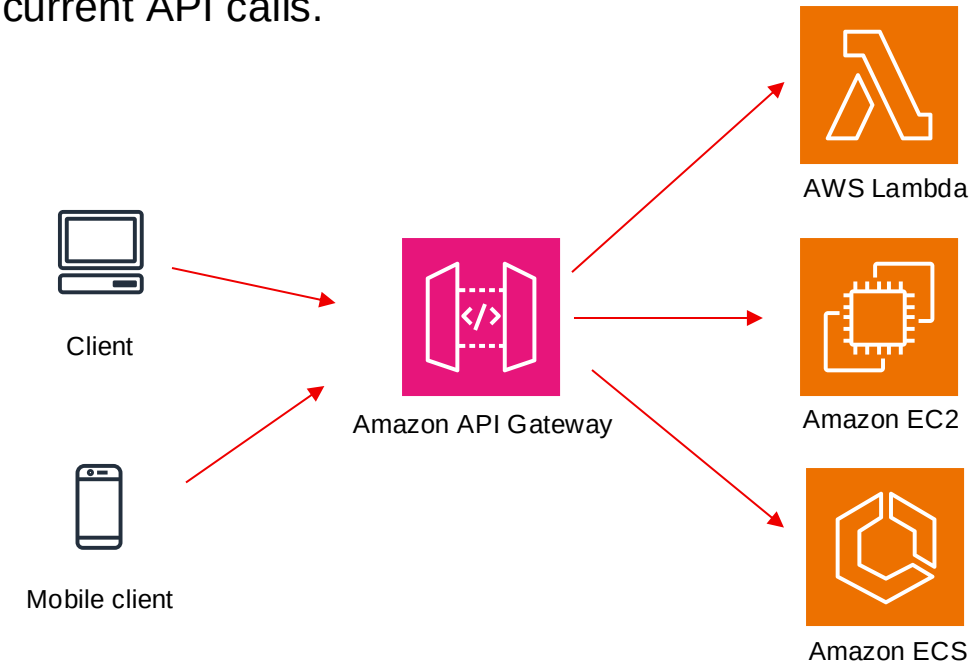
Design scalable and loosely coupled architectures.

Design scalable and loosely coupled architectures

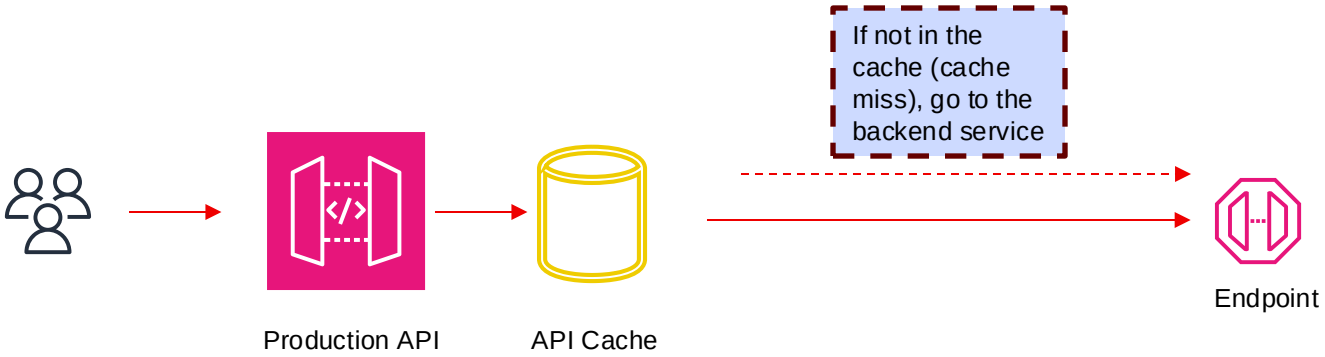
- API Gateway
- Amazon SQS, AWS SNS
- Stateless vs Stateful Applications.
- Amazon CloudWatch
- AWS CloudFront
- Storage Types (Object, File, Block)
- Container Services

API Gateway

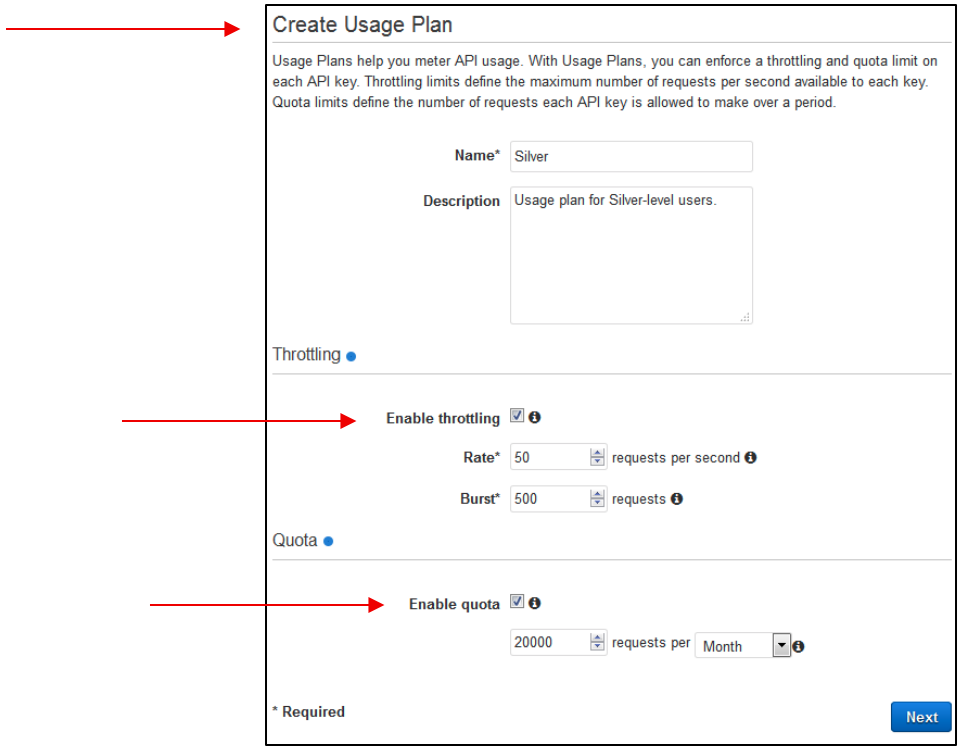
- API Gateway acts as a front door for applications to access data, business logic, or functionality from back-end AWS services.
- API Gateway handles all the tasks of accepting and processing up to hundreds of thousands of concurrent API calls.



API Gateway Caching

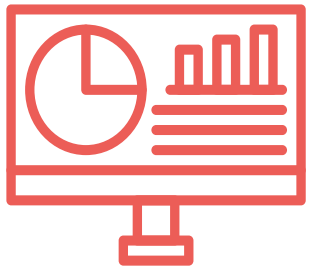


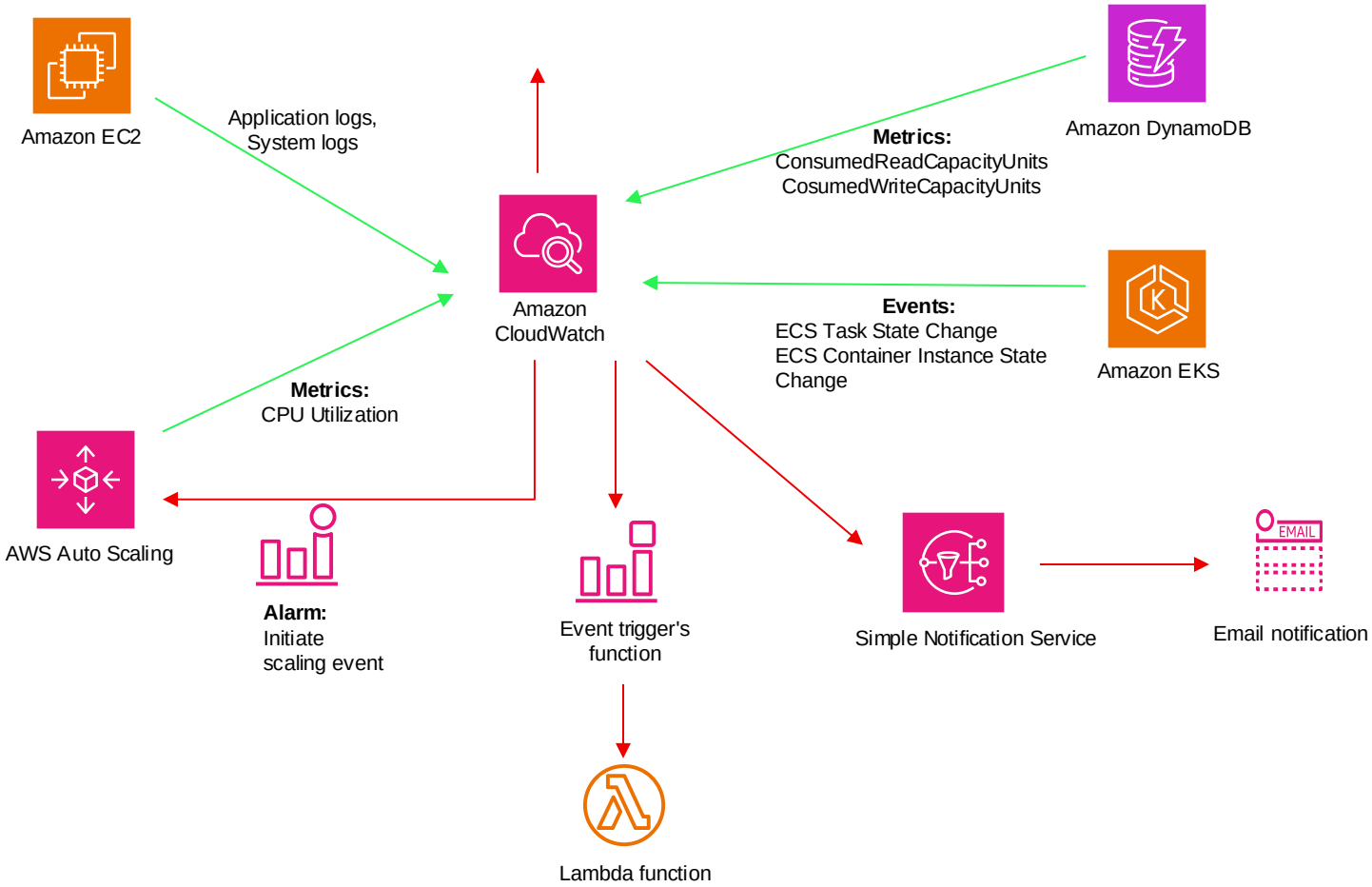
API Throttling with Usage Plans



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.



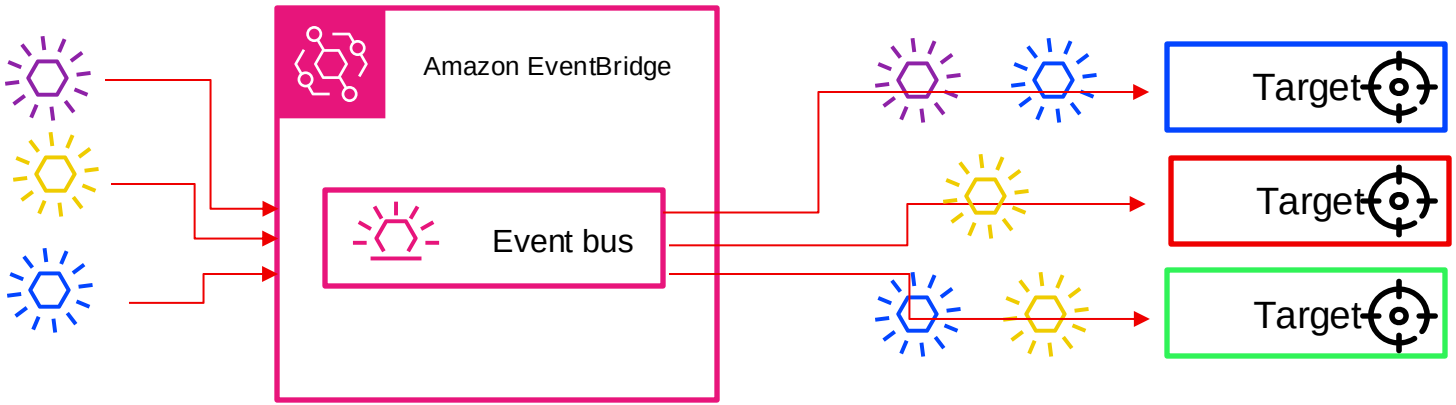

Amazon EC2
Application logs,
System logs
Amazon DynamoDB
Metrics:
ConsumedReadCapacityUnits
CosumedWriteCapacityUnits
Amazon
CloudWatch
Events:
ECS Task State Change
ECS Container Instance State
Change
Amazon EKS
Metrics:
CPU Utilization
AWS Auto Scaling
Alarm:
Initiate
scaling event
Event trigger's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

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.

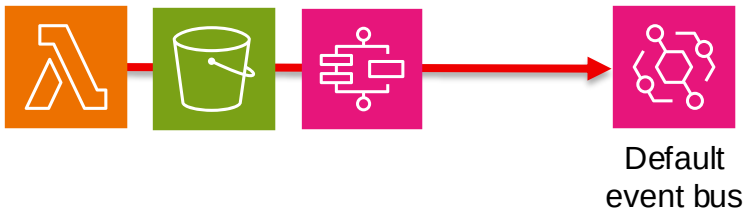
Amazon EventBridge

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



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).



Containers at AWS

Elastic Container Service

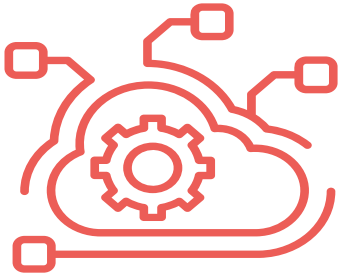
Run Docker containers: Applications, Micro-services.

Amazon Elastic Kubernetes Service

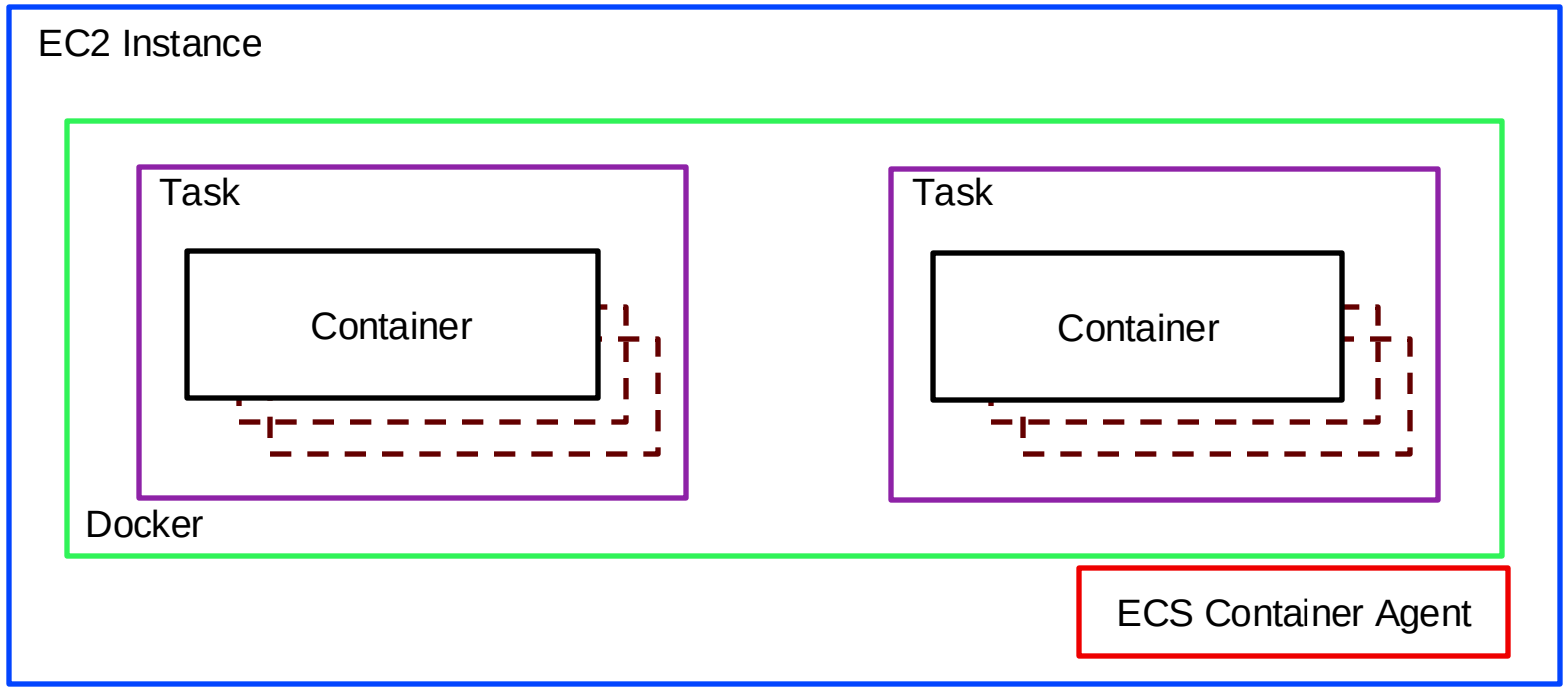
Manage containers with Kubernetes.

AWS Fargate

Management service for Docker or Kubernetes containers.



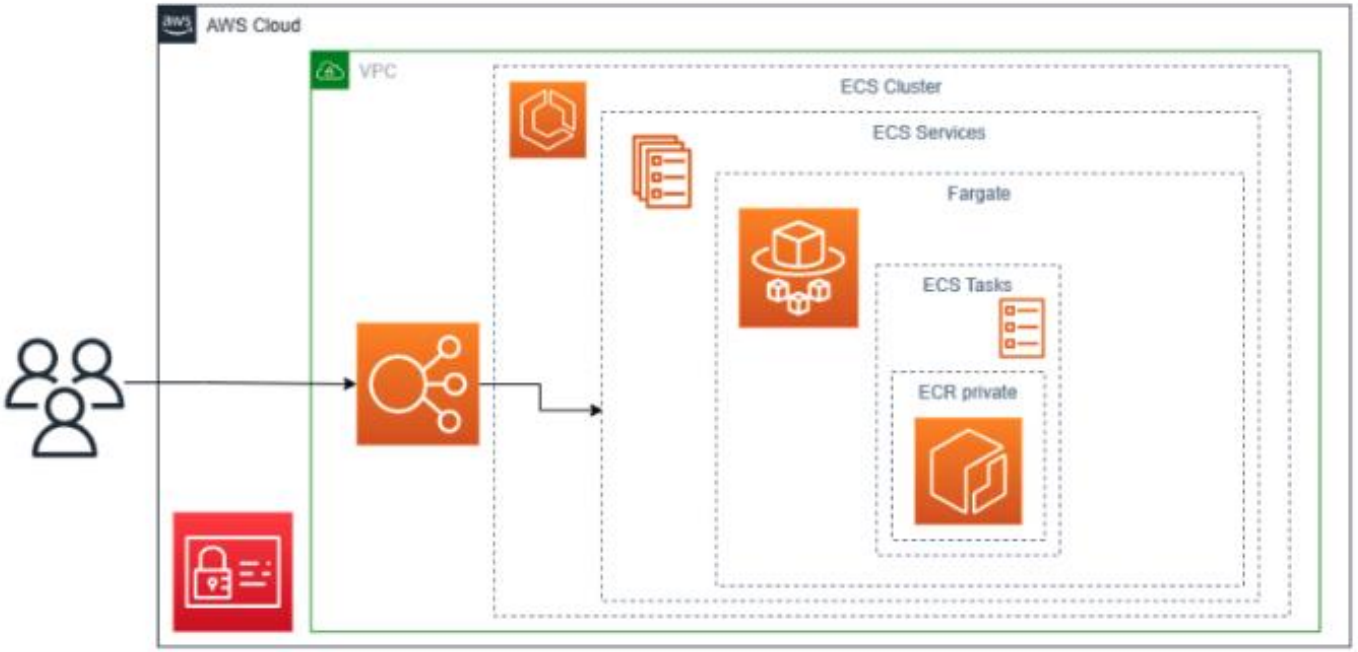
Task Definition Logic



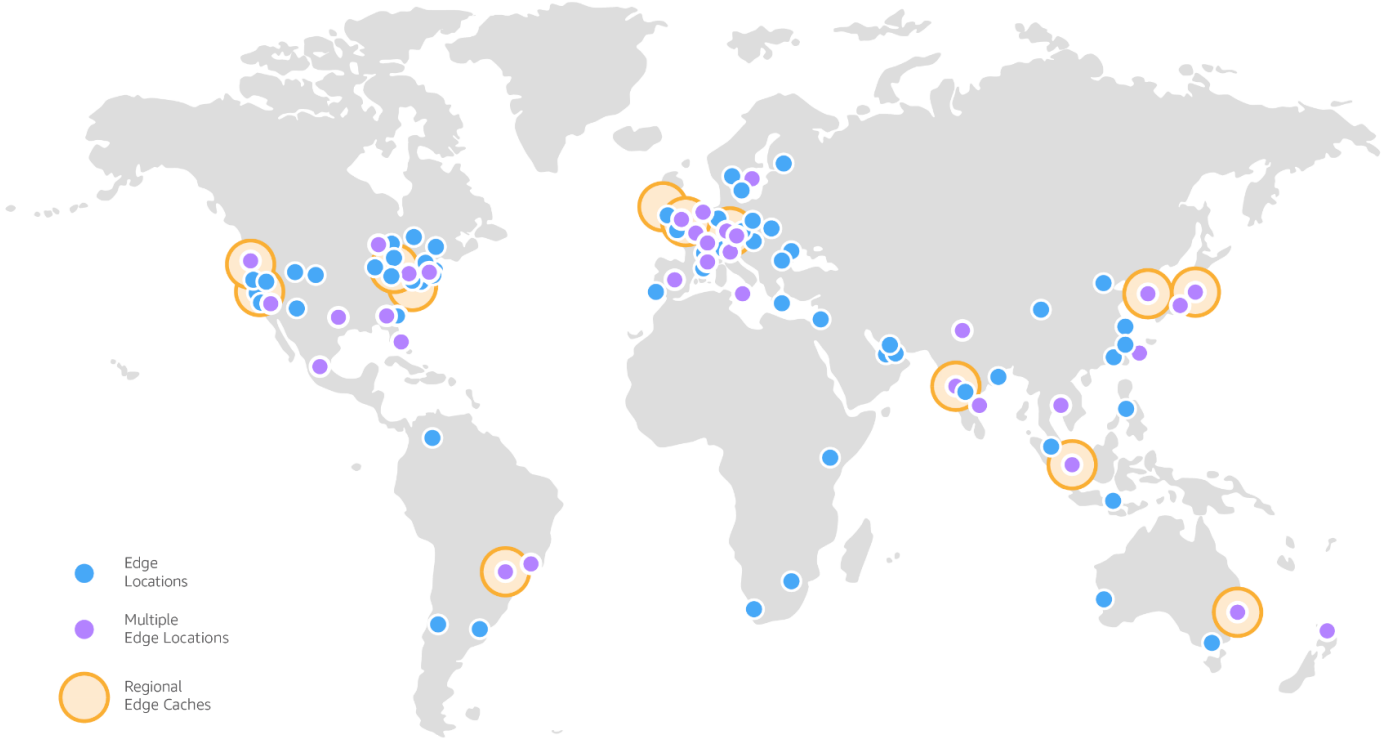
Container Orchestration with Fargate

- Fargate is a container orchestration service.
- Run your applications in containers and manage the underlying EC2 instances and clusters with Fargate.
- Specify the container image, memory, and CPU requirements using an ECS task definition.
- Fargate schedules the containers' orchestration, management and placement throughout the cluster, providing a highly available operation environment.

AWS Fargate


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

AWS Edge Locations



Edge Location Services

CloudFront

Caches your static and dynamic content

Route 53

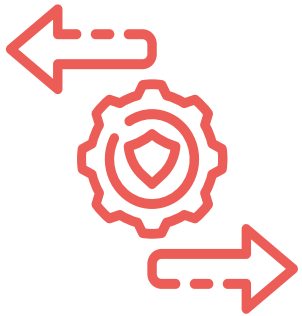
Delivers your request from closest edge location

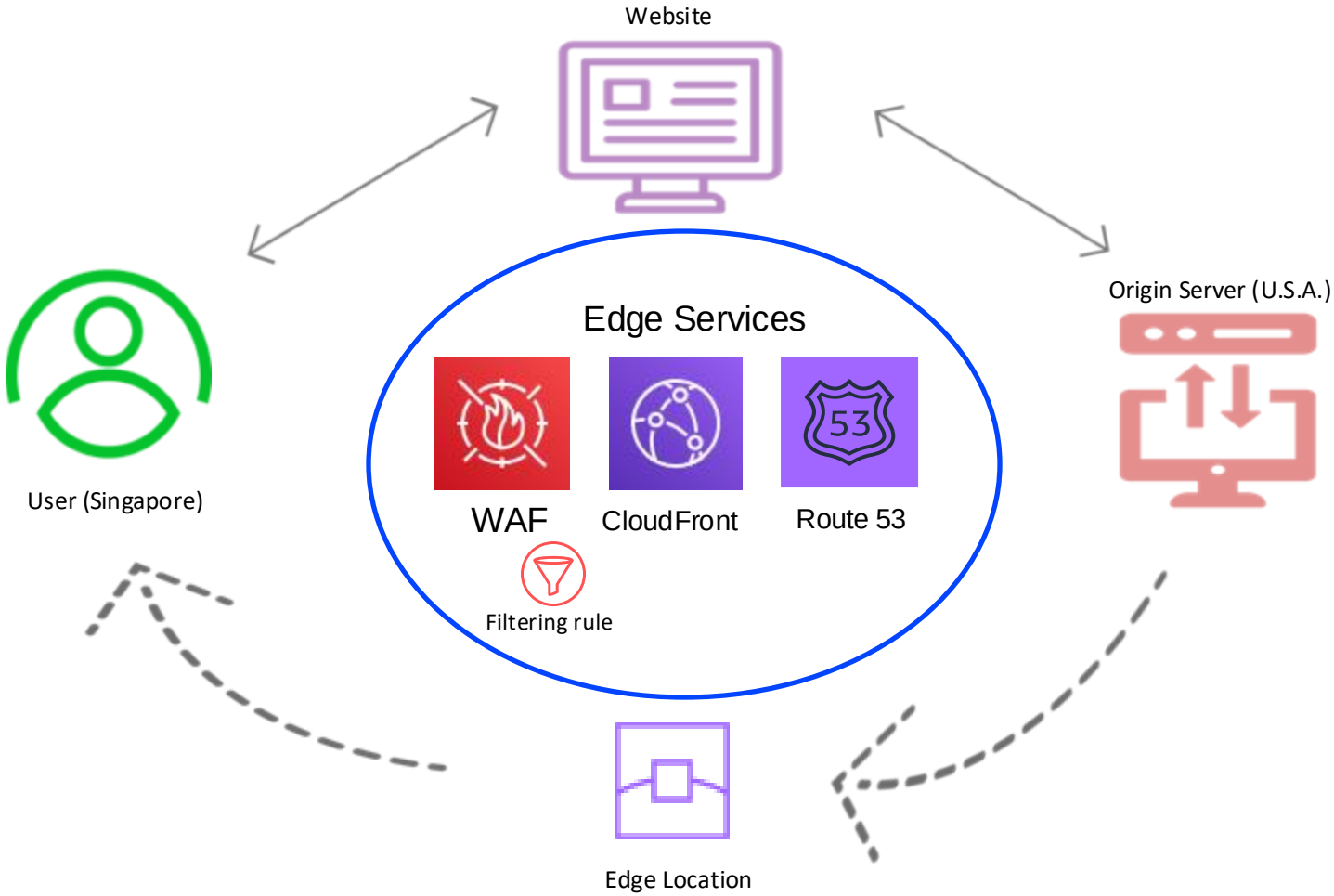
WAF

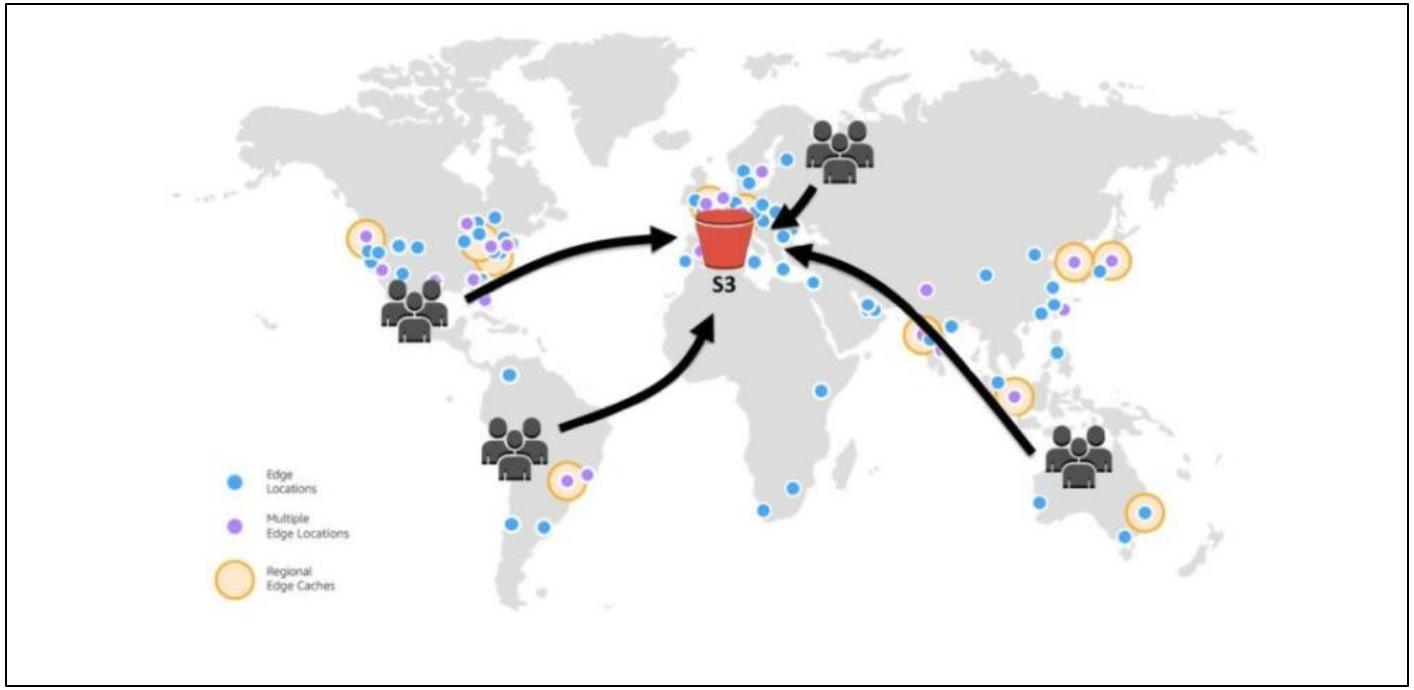
Filters incoming public traffic at the edge

CloudFront in Operation

- Content is delivered through a worldwide distribution of edge locations.
- Requests for content served with CloudFront are routed by Route 53 to the closest edge location with the lowest latency.
- Origin fetches from Amazon S3, Amazon EC2, or Elastic Load Balancing are free from data origin to CloudFront Edge locations.
- GET and DELETE requests are charged.



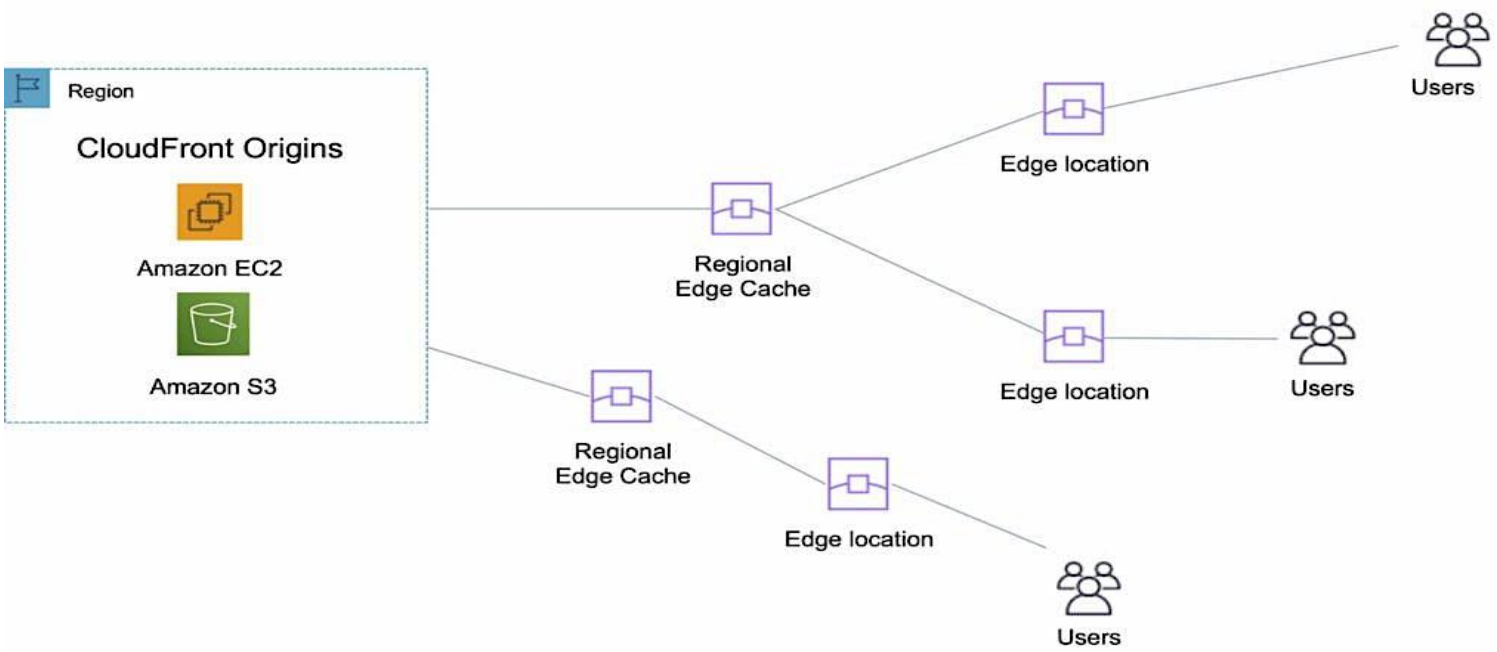

Website
Origin Server (U.S.A.)
User (Singapore)
Edge Services
WAF
CloudFront
Route 53
Filtering rule
Edge Location



Without CloudFront

With CloudFront

Regional Edge Cache



Protecting the Application Perimeter

AWS Shield Standard

Protects against DDOS attacks

AWS Shield Advanced

Enhanced visibility and economic protection

AWS WAF

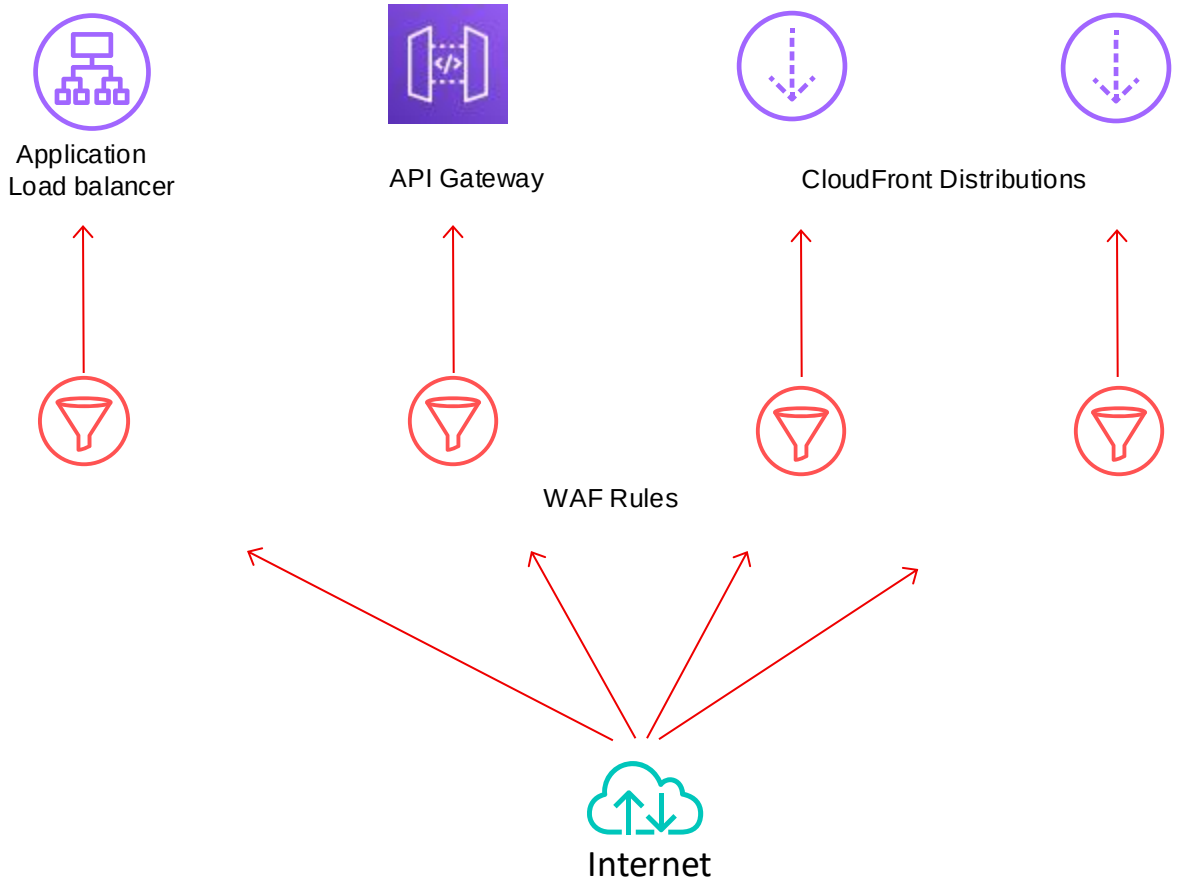
Filtering rule

Web app protection with custom or managed rules

WAF Protection

- Pre-configured rules.
- Common attack vectors and threats.
- Regular or rate-based rules.
- Actions to take (block, allow, count).

WAF Design


Application
Load balancer
API Gateway
CloudFront Distributions
WAF Rules
Internet

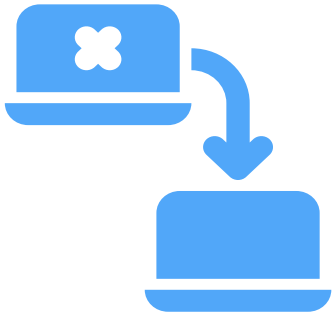
Serving Private Content

- 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).

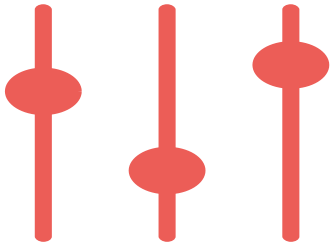
CloudFront Origin Failover

- 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.

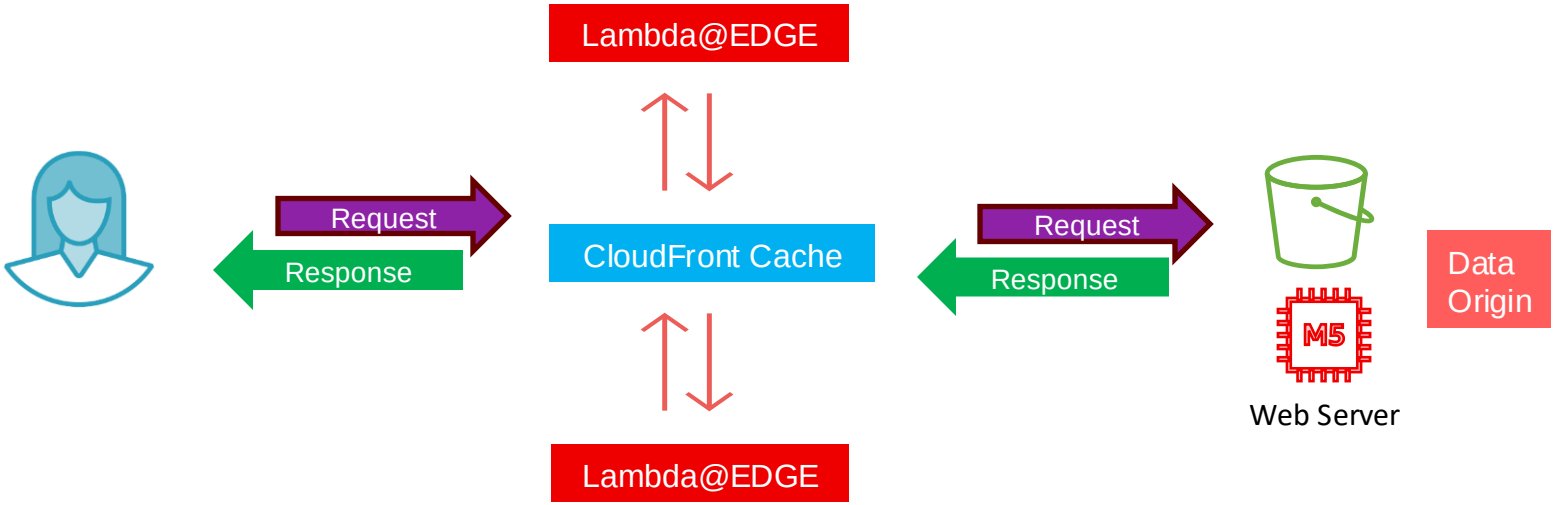


Customizing with Lambda@Edge

- 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.

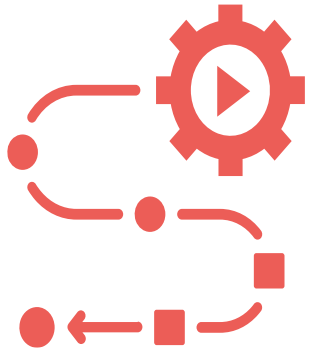


Lambda@EDGE

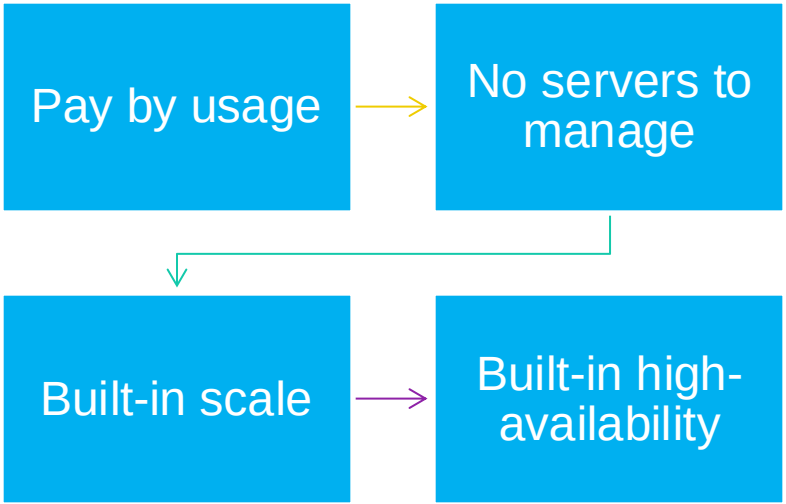

Lambda@EDGE
Request
Response
CloudFront Cache
Request
Response
Data Origin
M5
Web Server
Lambda@EDGE

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.



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

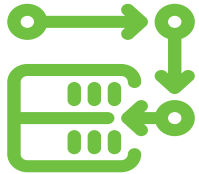
- **S3 bucket:** An object is uploaded to a bucket.
- **DynamoDB table:** An entry is made in a DynamoDB table.
- **Amazon Kinesis:** Data is added to an Amazon Kinesis stream.
- **Amazon SNS:** A message is published to an SNS topic.
- **Amazon API:** RESTful APIs call a Lambda function to access backend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

Stateful Applications

- Databases store and manage data records.
  - Amazon RDS, Amazon DynamoDB.
- E-commerce sites store account information.
- Gaming sites store user account information and gaming state.

Stateless vs Stateful Applications

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



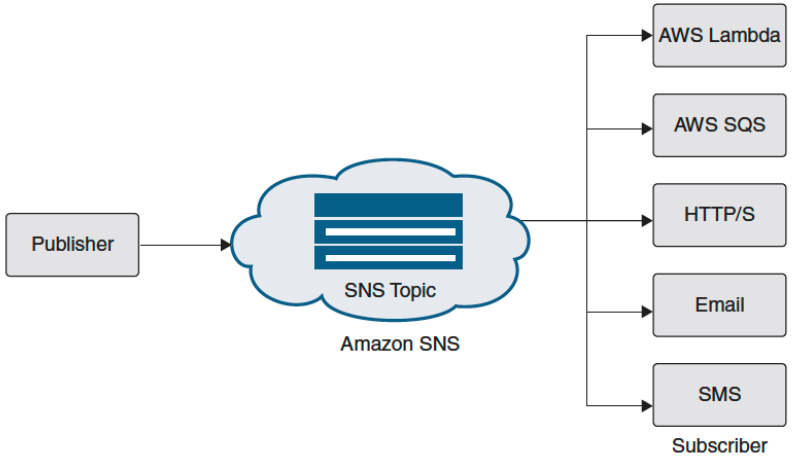
- Stateless applications operate without knowledge of previous interactions.
  - RESTful API requests contain all required information for processing.
  - Stateless applications perform specified activities without storing state information.

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.

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.



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

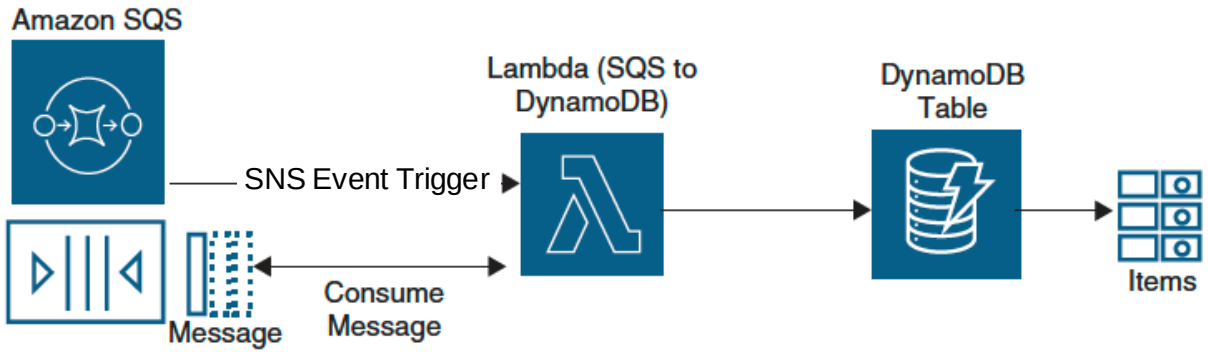
Default short polling.

Optional long polling.

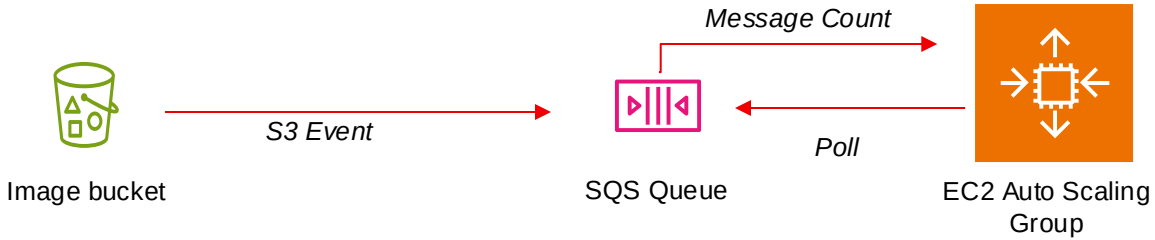
Long polling can help reduce the number of empty received message responses returned.

SQS Message Processing

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



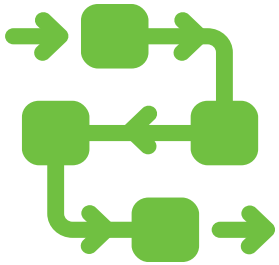
SQS Message Processing



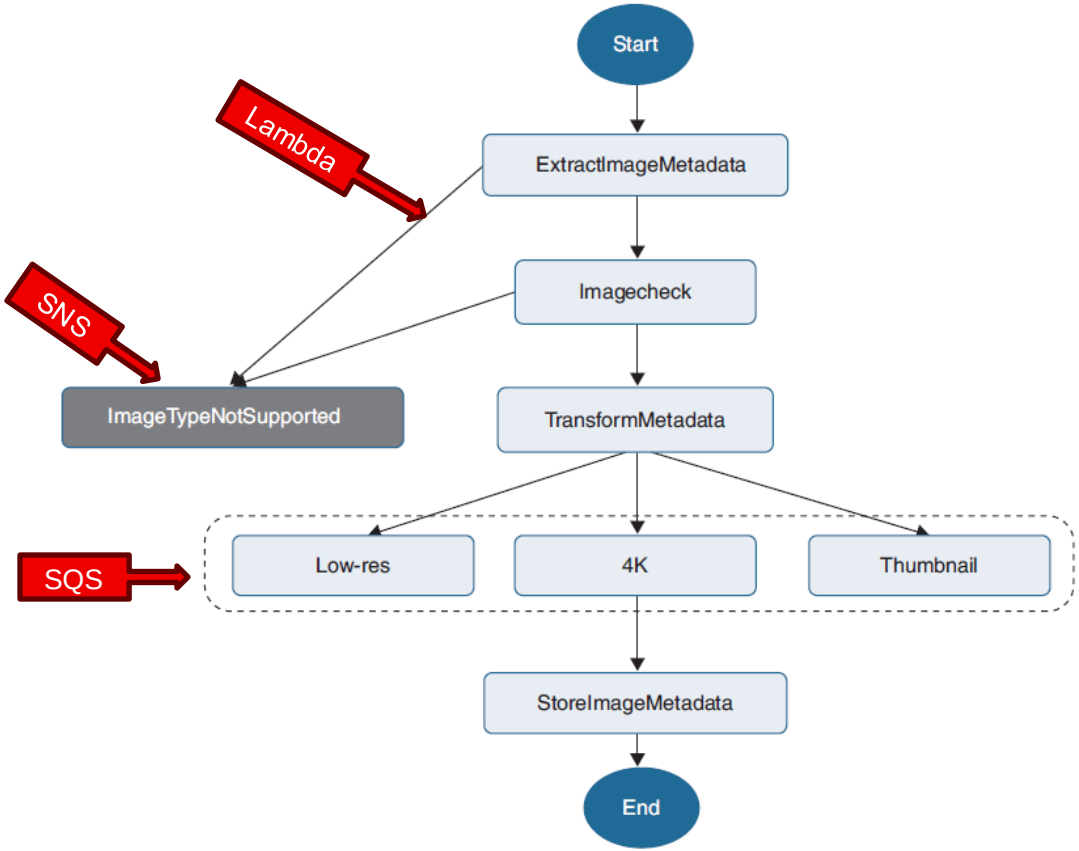
- EC2 Auto Scale: Scale up when messages in the SQS queue increase.
- The number of instances/containers in the worker pool can be scaled according to the number of messages in the queue.

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



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.

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

Task Statement 2.2

Design highly available and fault tolerant architectures.

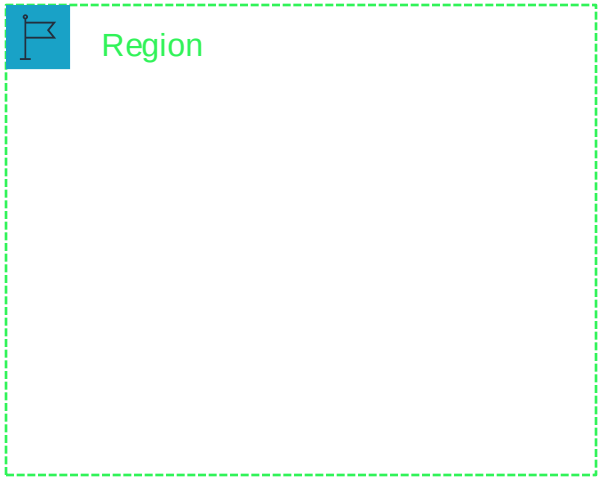
Design Highly Available and Fault Tolerant Architectures

- AWS Regions and Availability Zones.
- Amazon Route 53 Traffic Patterns
- Disaster Recovery Strategies
- Immutable infrastructure
- AWS X-Ray

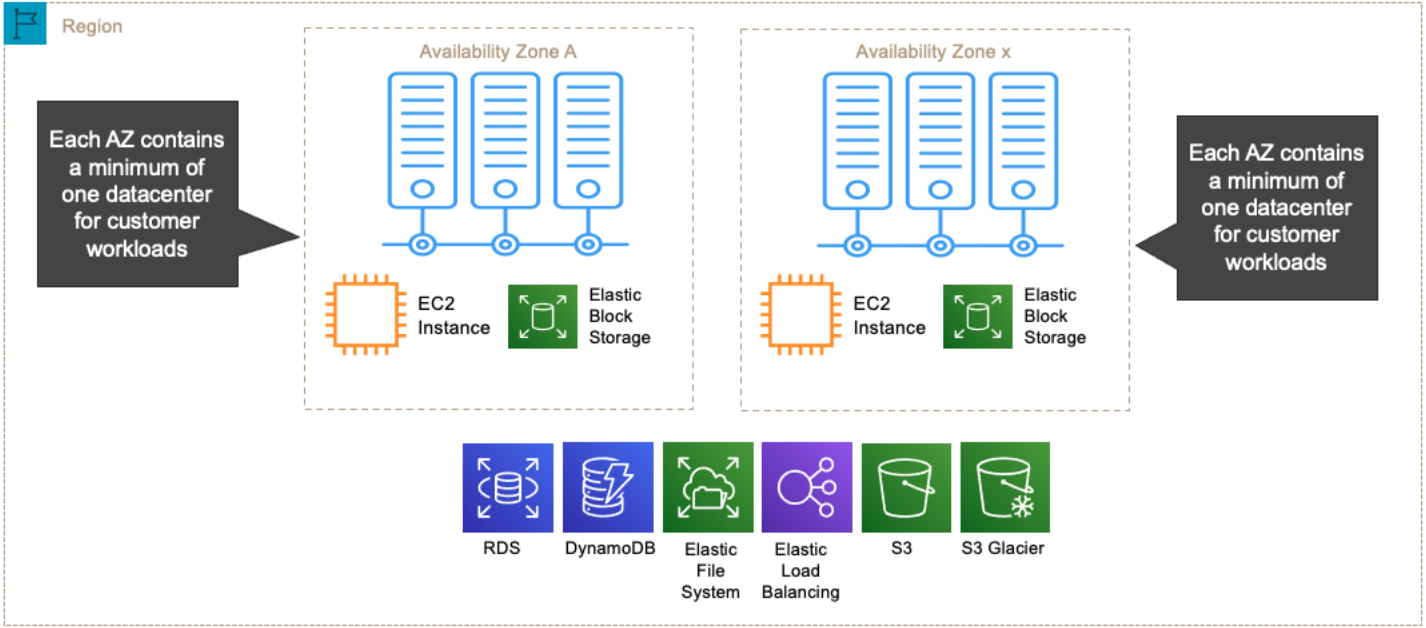
AWS Regions

AWS Regions

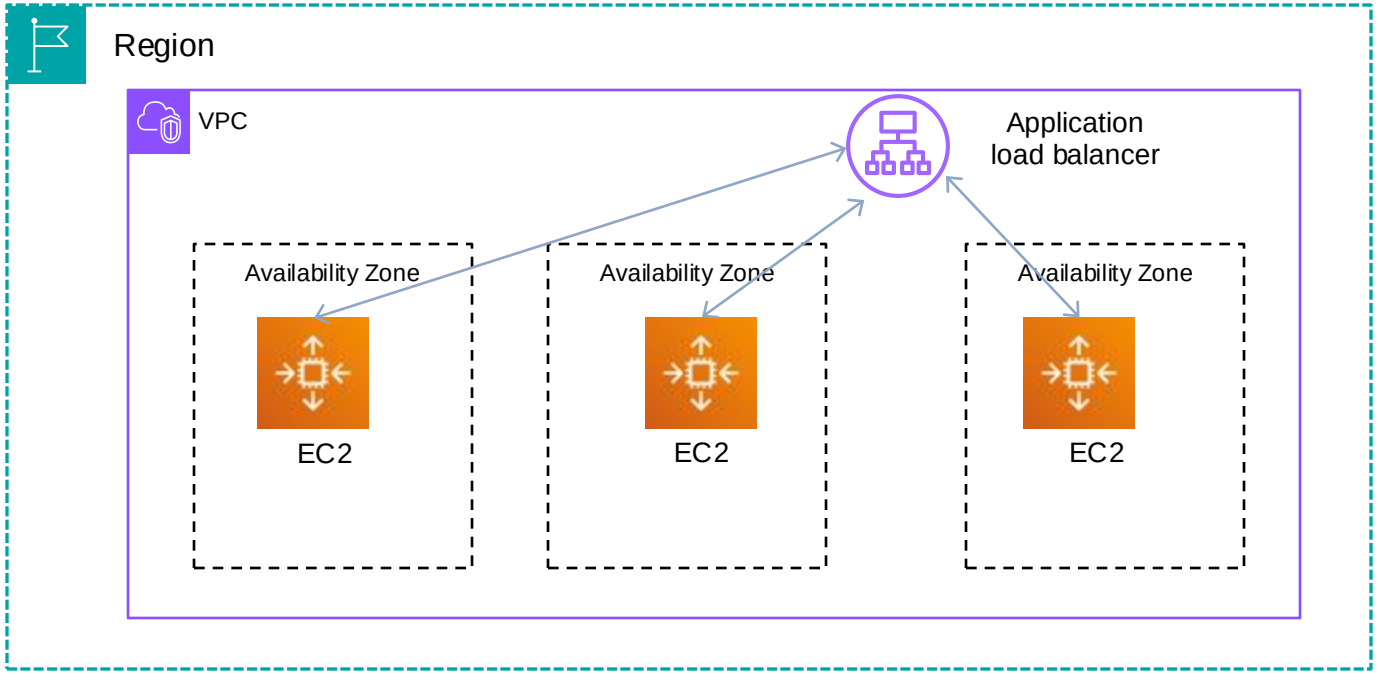
- Areas of the world where Amazon offers AWS cloud services.
- Each region is a geographical location.
- Each region is mostly independent and isolated.
- AWS regional services are not replicated to other regions by default.



Regional Services

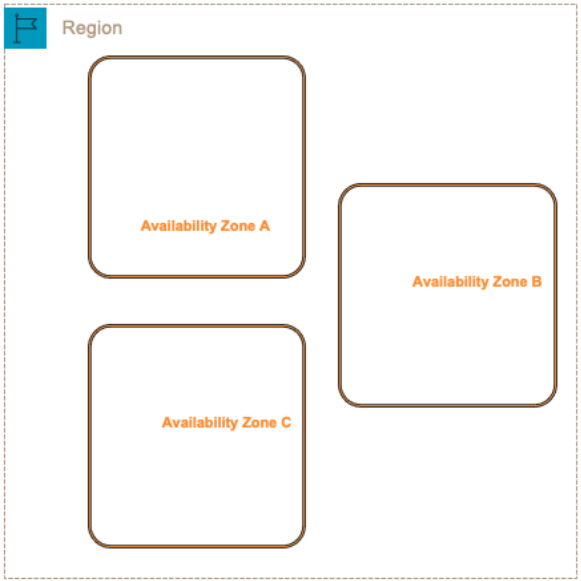


Multi-AZ Design



Availability Zones Summary

- Each availability zone contain at least one data center for customer workloads.
- Many availability zones contain multiple data centers for customer workloads.
- Each availability zone has inexpensive low latency network connectivity to the other availability zones in the same region.
- Designing with two AZ’s is a best practice to consider.



Availability Zones In Operation

- EC2 instances are deployed in multiple subnets in multiple availability zones.
- ELB target EC2 instances in multiple availability zones.
- EC2 Auto Scaling can scale instances in multiple availability zones.
- RDS solutions are deployed in multiple availability zones.
- Amazon Aurora cluster-shared storage is replicated across multiple availability zones.
- DynamoDB tables are replicated across multiple availability zones.

Route 53 Features

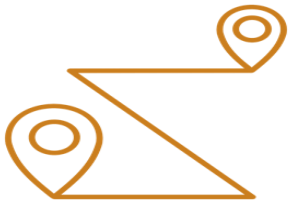
- Global traffic management; direct requests to the correct endpoint.
- Health checks and monitoring of AWS resources.
- Zone-apex support for CloudFront distribution and S3-hosted websites.
- Resolve with recursive DNS for both VPC and on-premise networks.

Hosted zone

Resolver

Routing Policy Types

- Simple
- Weighted
- Latency
- Failover
- Geolocation



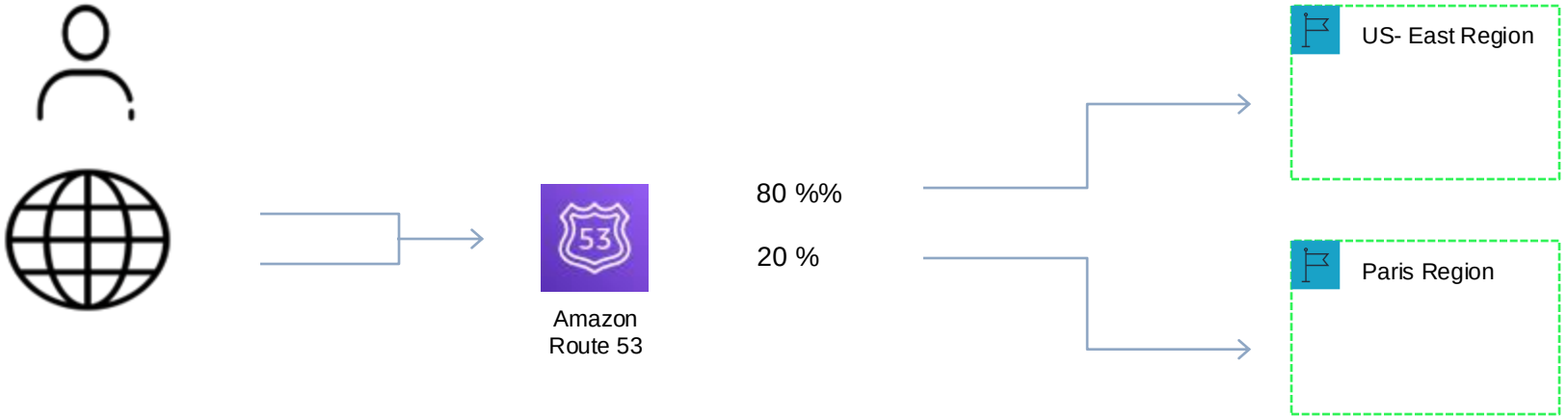
Simple Routing Policy

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

Used when you have a single resource.

Weighted Routing Policy

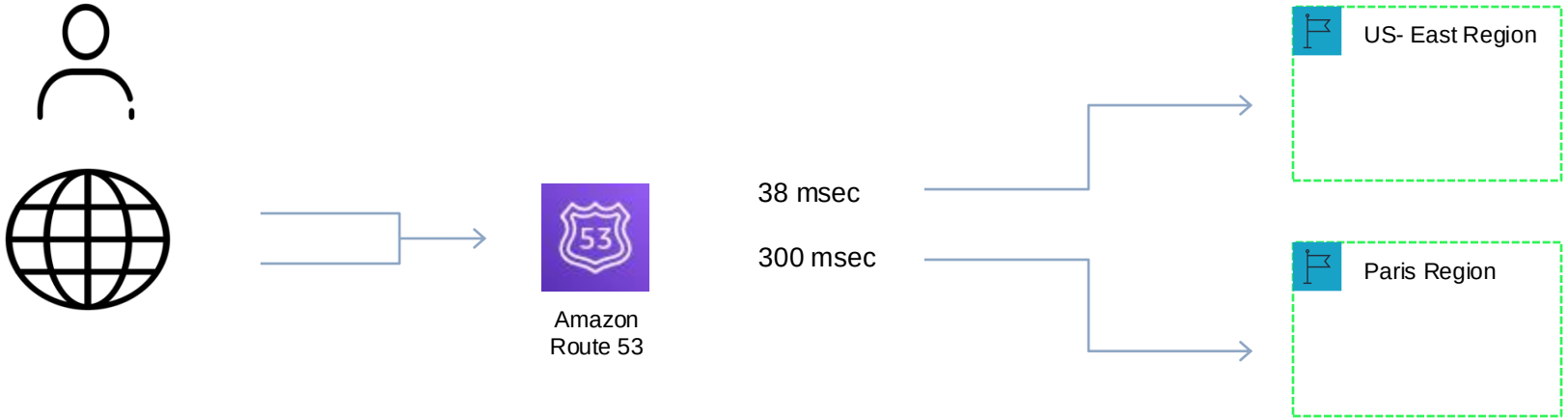
- Split traffic based on different weights assigned to resources.



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 latency resource record set in each region that hosts your resource.



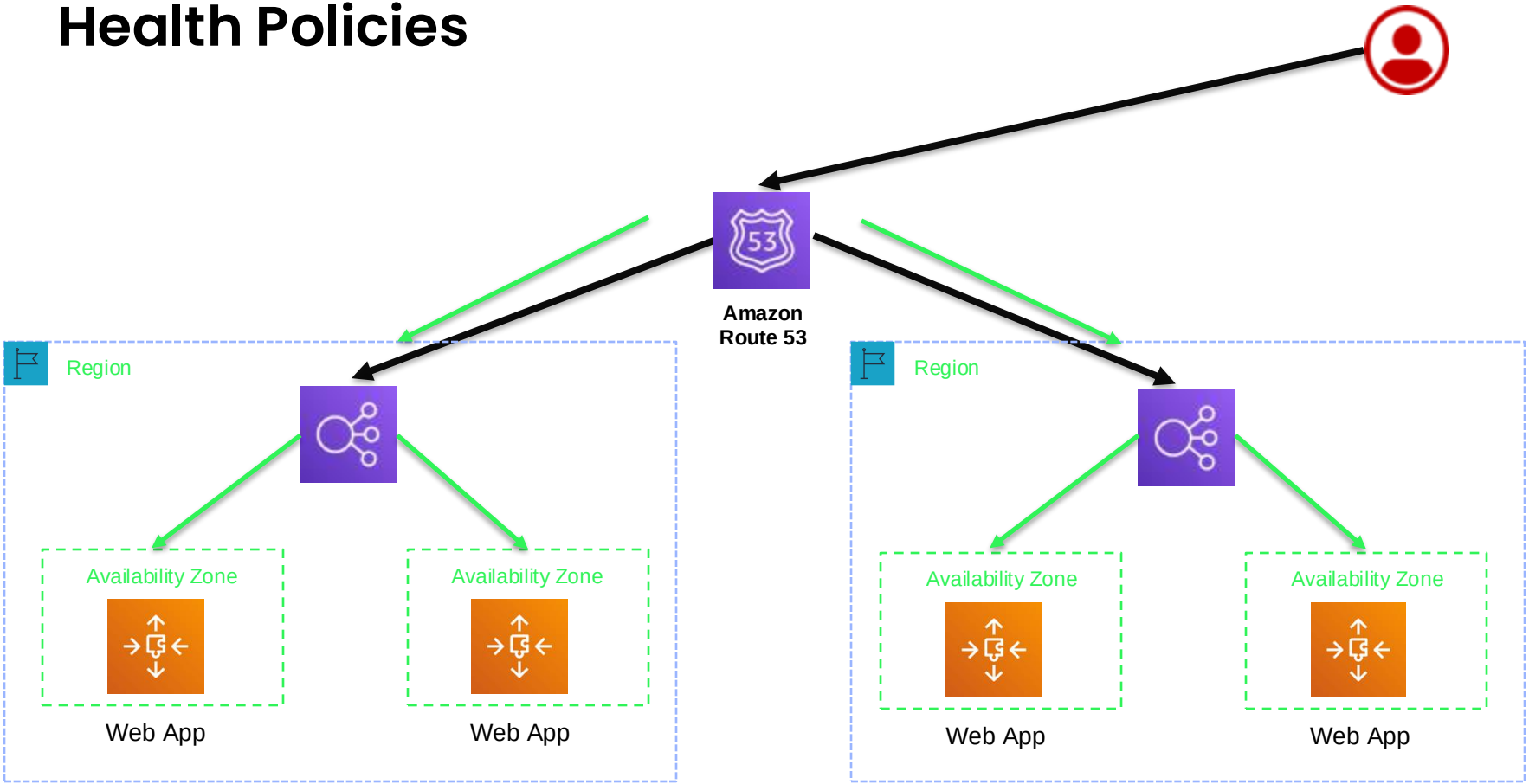
Route 53 selects the latency resource record for the region with the lowest latency.

Failover Routing Policy

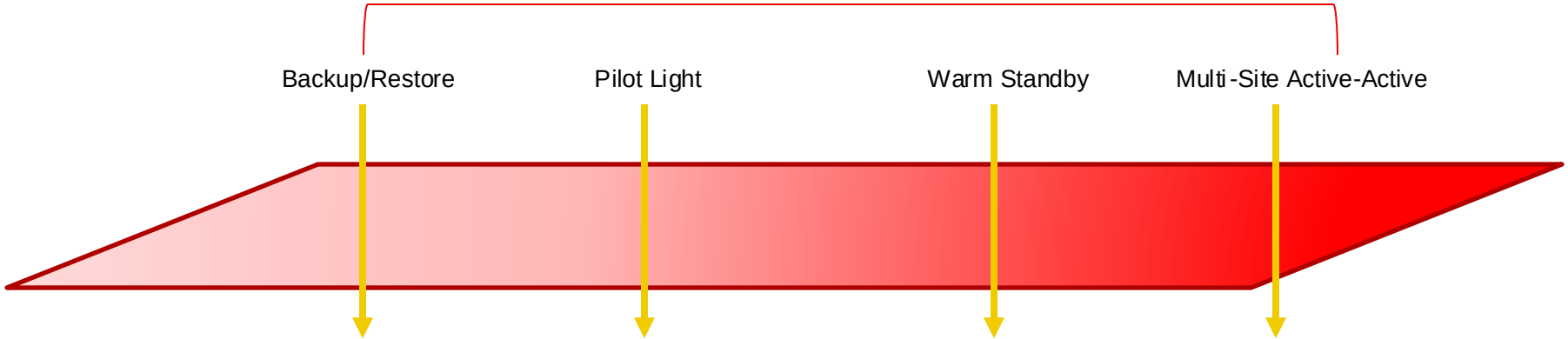
- Highly available ELB load balancers in separate regions.

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

Health Policies

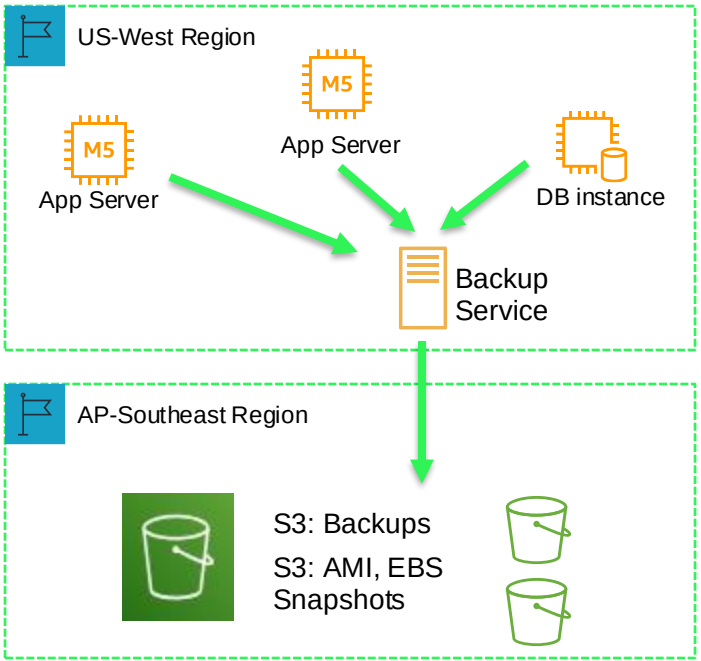


Disaster Strategies

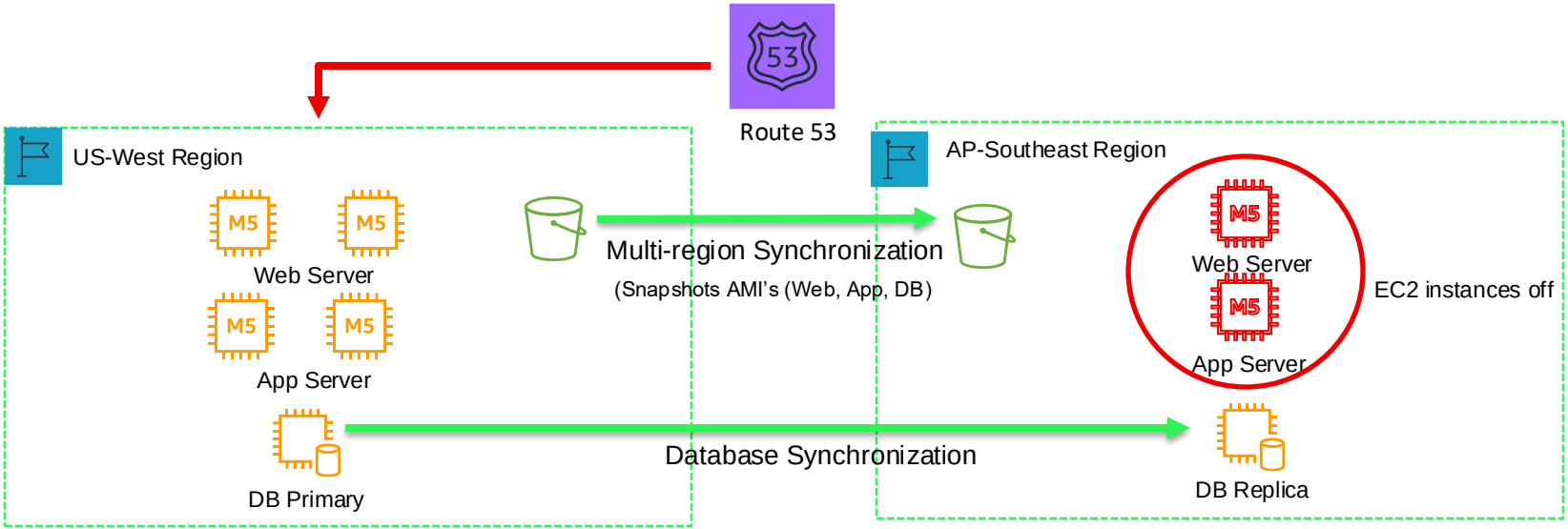


RPO / RTO	RPO / RTO	RPO / RTO	RPO / RTO
Hours	< 1hour	Minutes	Real-time
Low priority Provision resources after failure $	Data live; Services idle Provision and scale after failure $$	Services smaller Scale after failure Business critical $$$	No downtime Near-zero data loss Mission critical $$$$

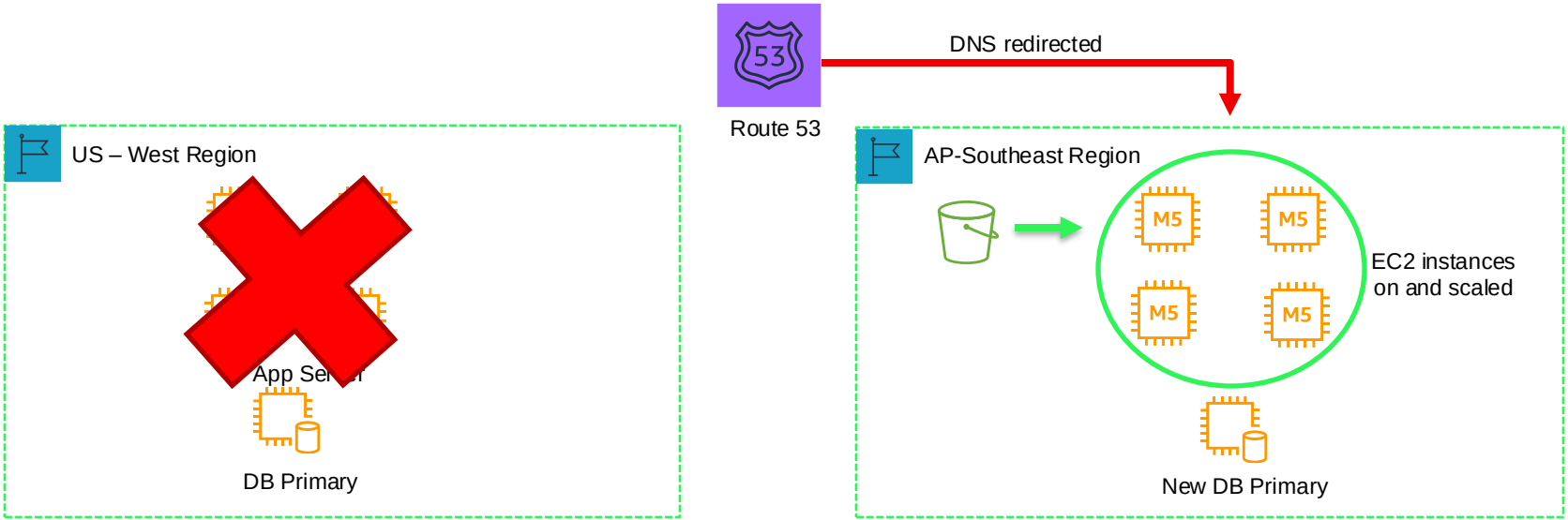
Multi-Region Backup



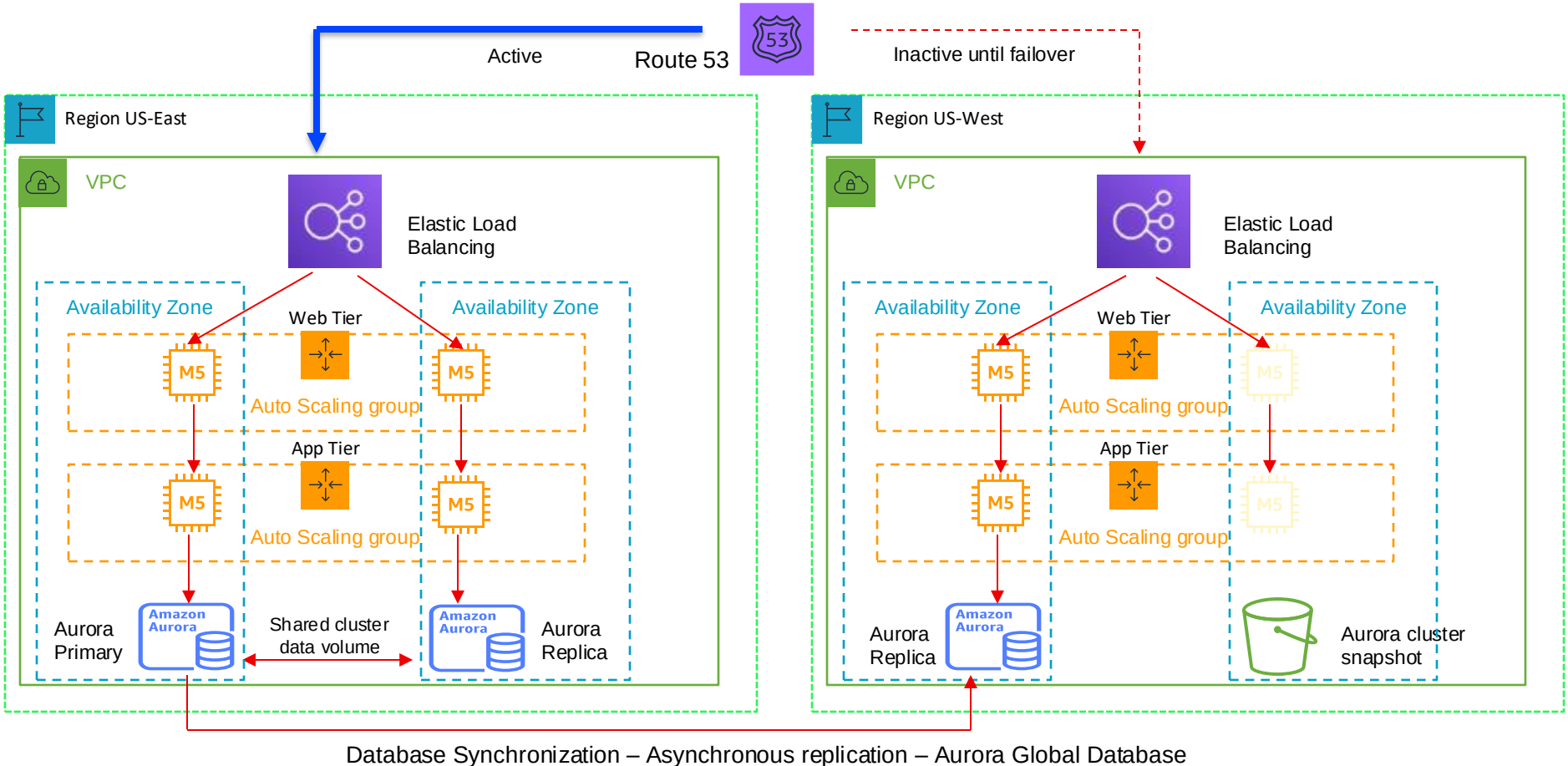
Multi-Region Pilot Light Setup



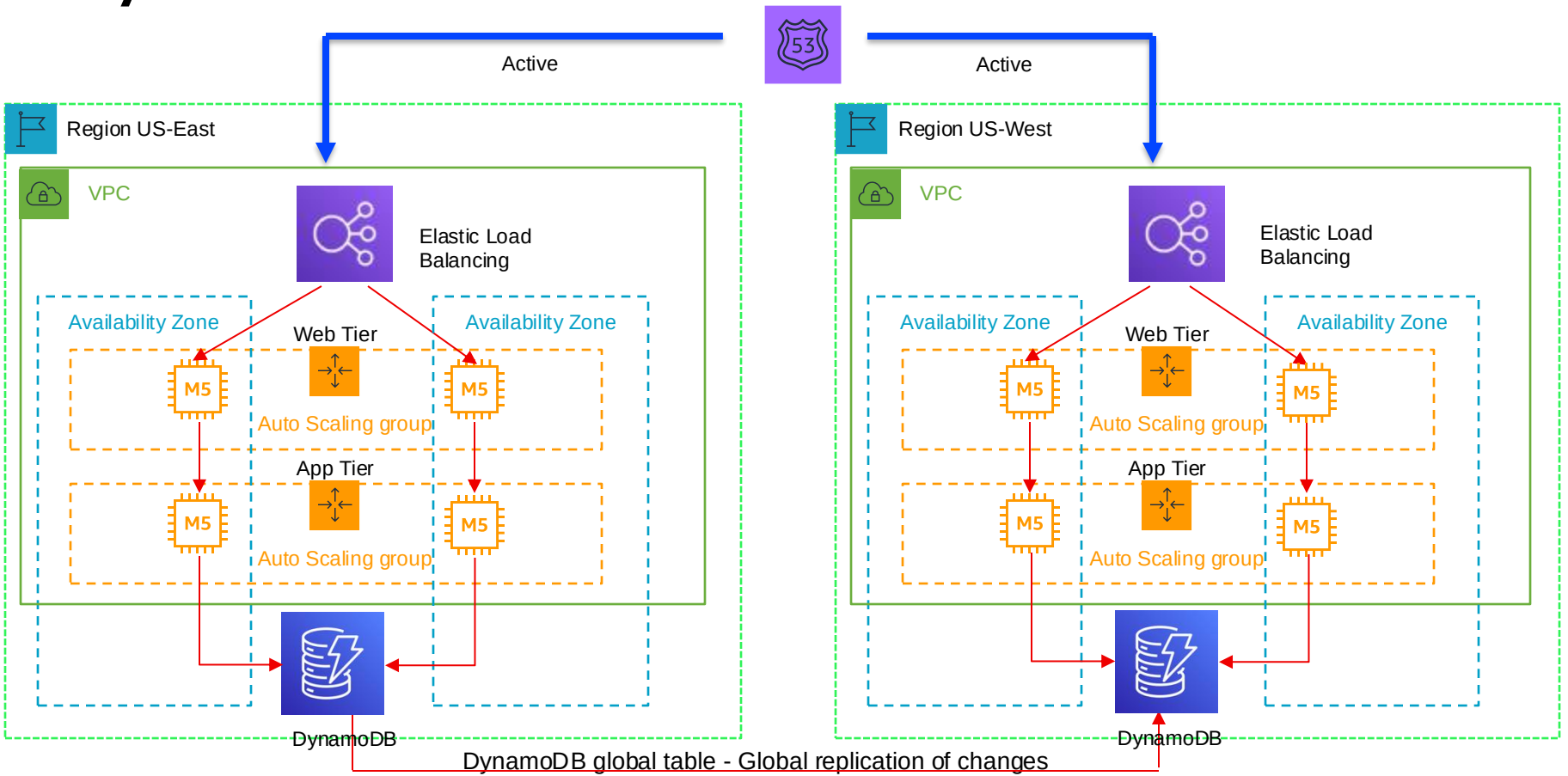
Multi-Region Pilot Light Failover



Warm Standby with Aurora Global Database

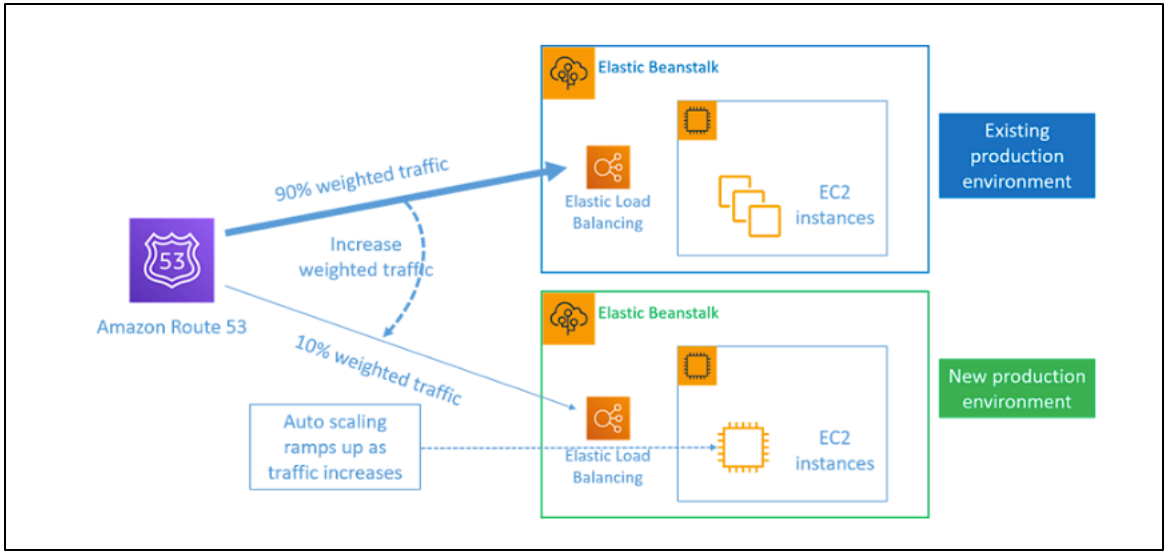


DynamoDB Global Tables - Active-Active

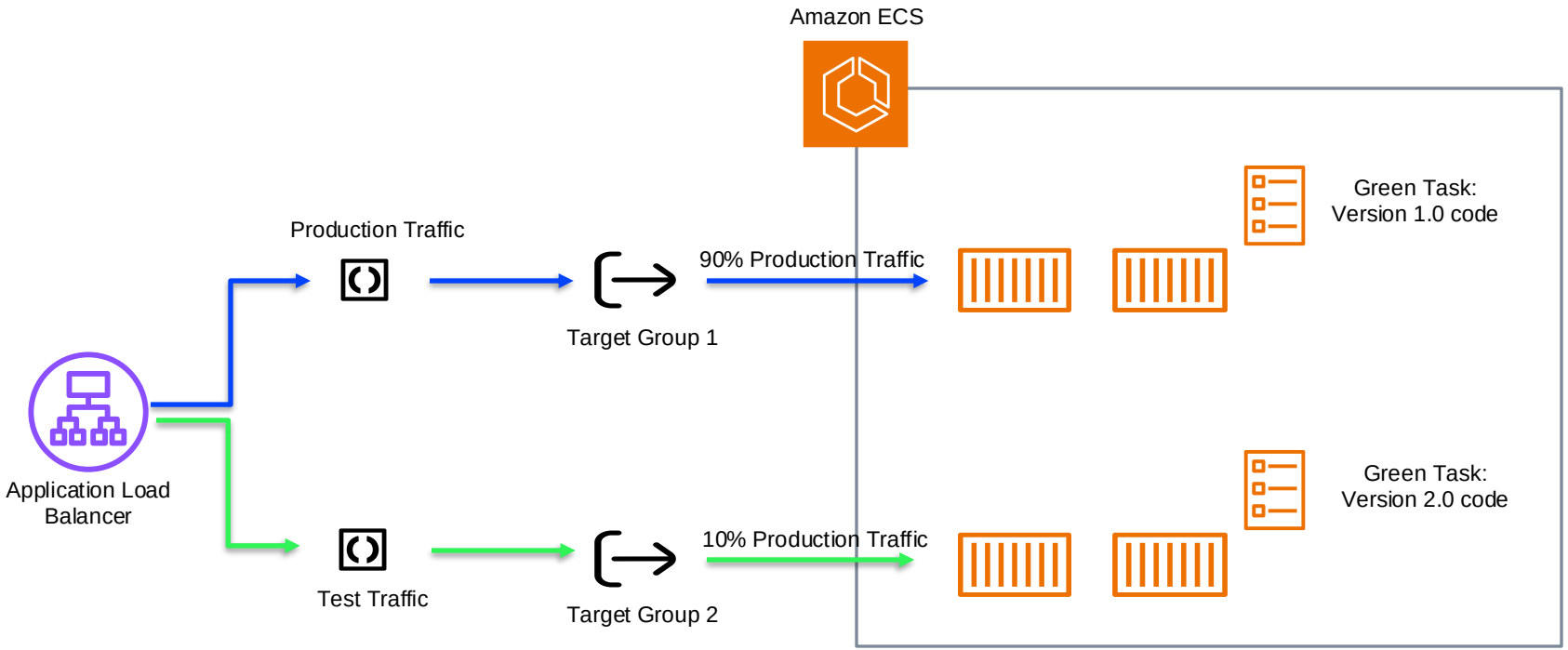


AWS Elastic Beanstalk

- AWS Elastic Beanstalk supports blue/green deployments.



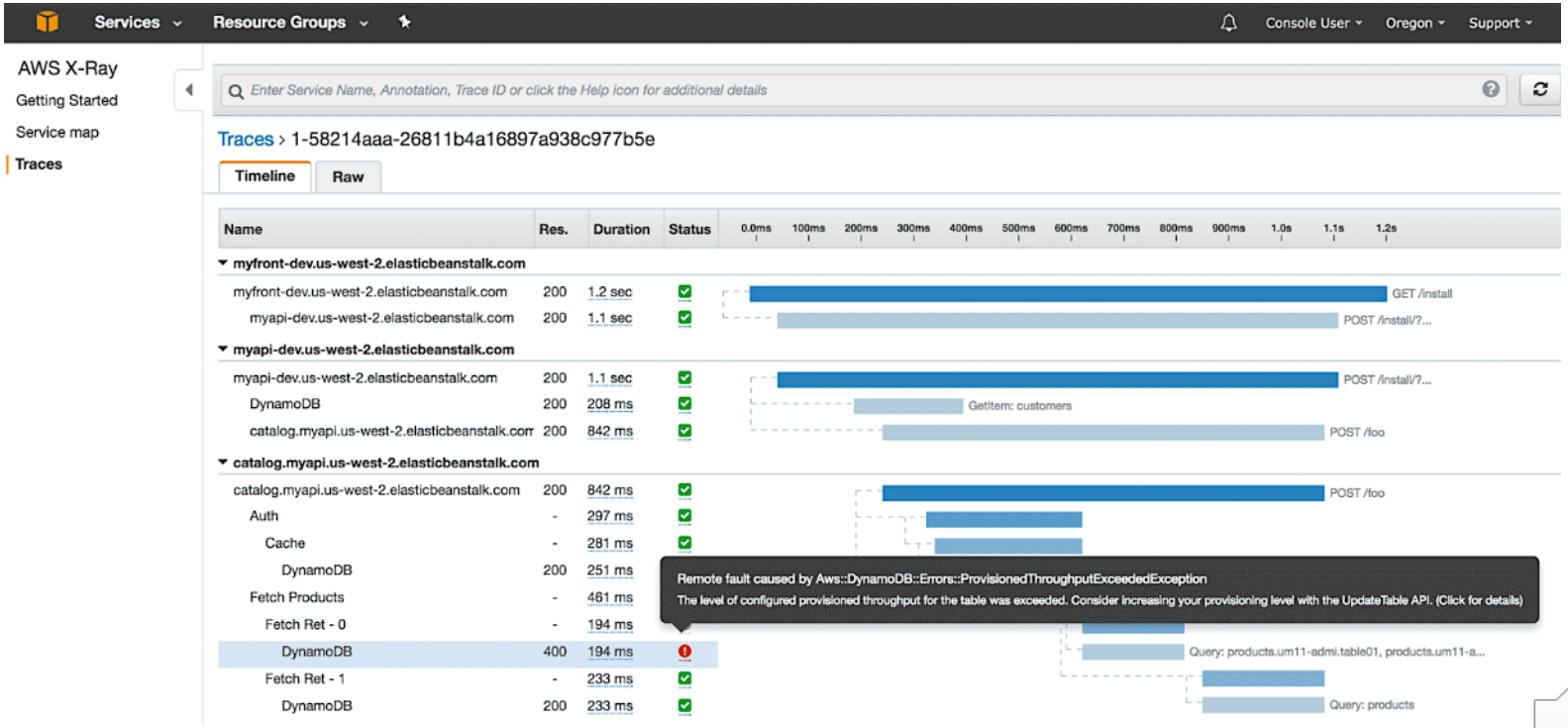
ECS Blue / Green Deployment



AWS X-Ray

- Analyze the behaviour of applications with end-to-end tracing.
- Use with applications running on EC2, ECS, AWS Lambda, and AWS Elastic Beanstalk.
- Applications written with the X-Ray SDK and X-Ray agent installed on EC2 instance.

X-Ray Tracing



Domain 3

Design High-Performing Architectures

Task Statement 3.1

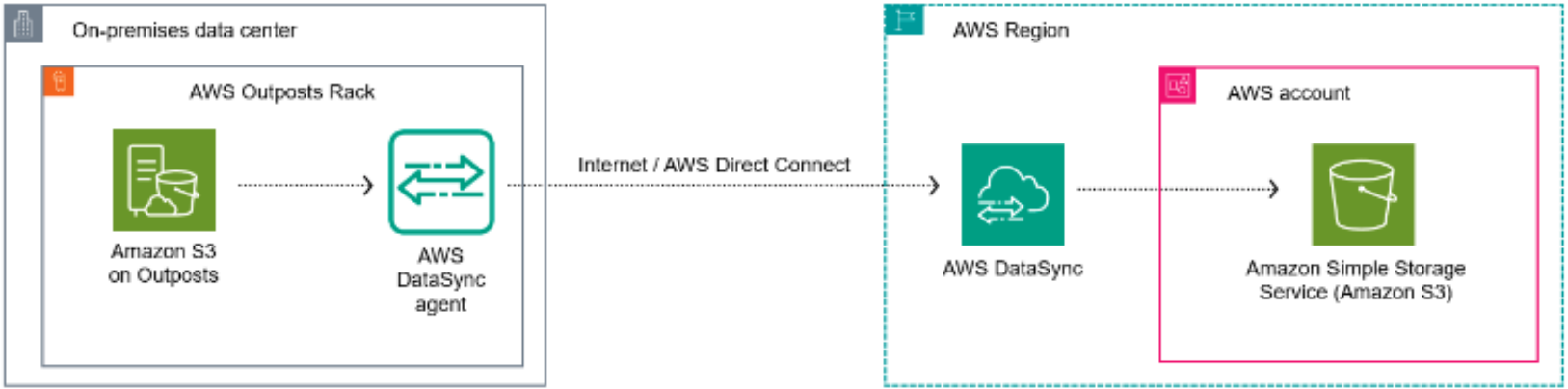
Determine high-performing and scalable storage solutions.

Determine high-performing and scalable storage solutions

- Hybrid storage (EFS, FSx for Windows File Server)
- Use cases for storage services (EBS, S3, EFS, FSx for Windows File Server)

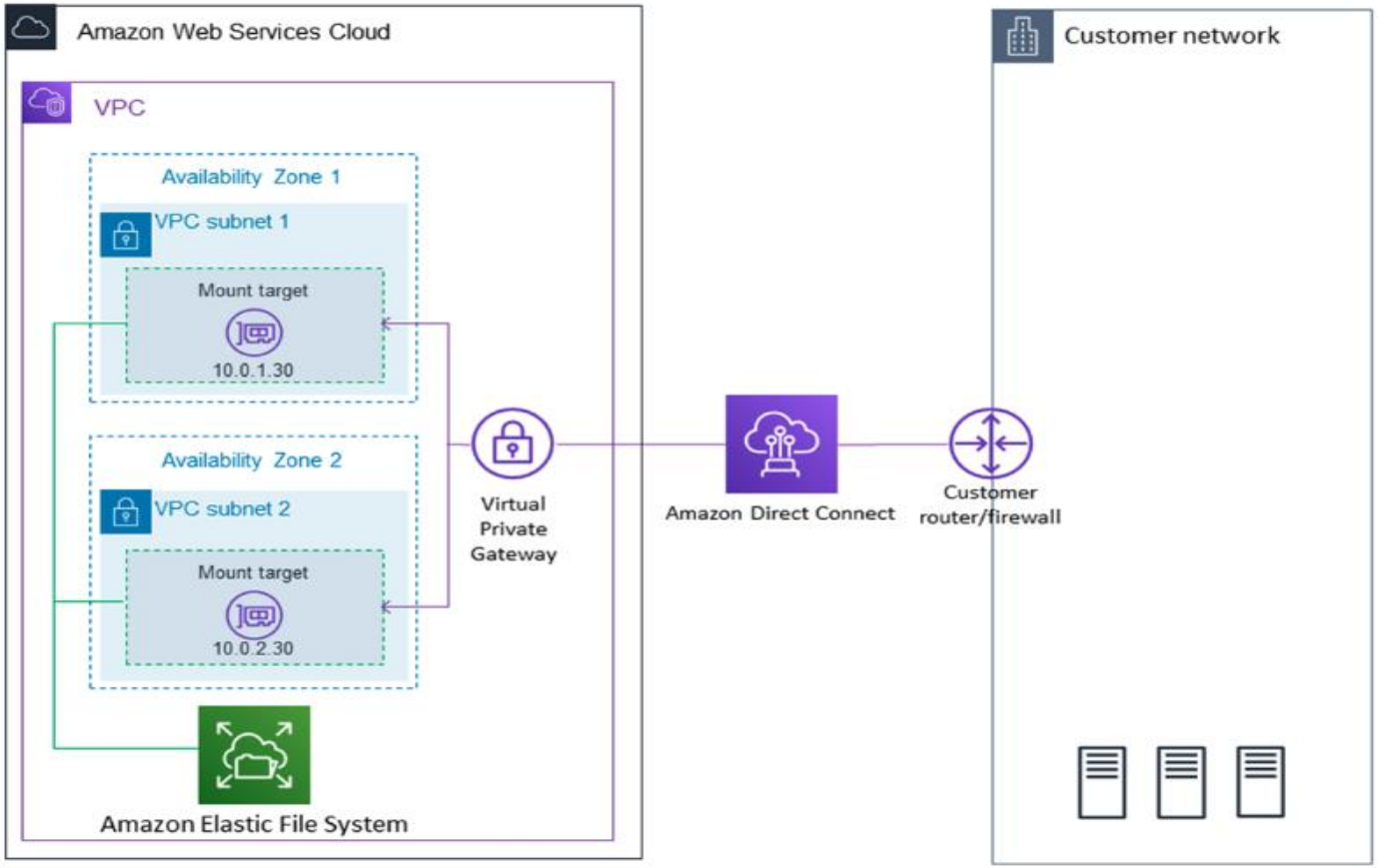
AWS S3 on Outposts

- S3 on Outposts allows deployment of object storage on-premises.
- Capacity can be selected in 26TB, 48TB, 96 TB, 240TB, or 380TB.
- Process and securely store data locally in your on-premises environment.
- Transfer data between Amazon S3 on Outposts and AWS Regions by using AWS DataSync.


On-premises data center
AWS Outposts Rack
Amazon S3
on Outposts
AWS
DataSync
agent
Internet / AWS Direct Connect
AWS Region
AWS DataSync
AWS account
Amazon Simple Storage
Service (Amazon S3)

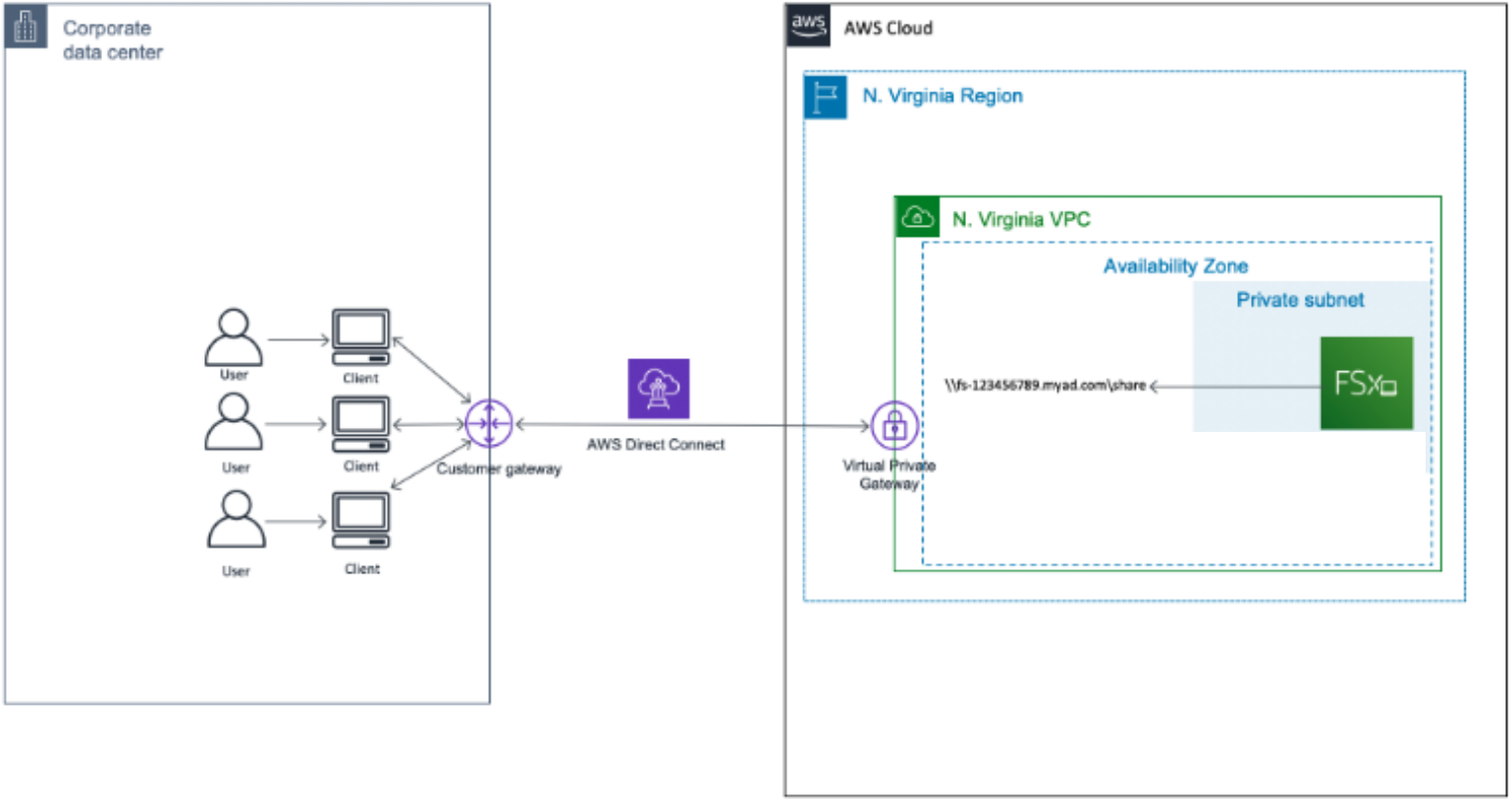
File system On-premises with Direct Connect or VPN.

- Create an Amazon EFS file system and a mount target in your Amazon VPC.
- Download and install the amazon-efs-utils tools.
- Test 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

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

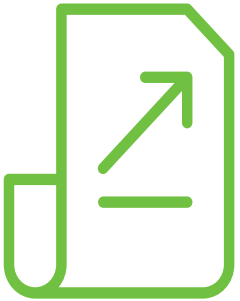
- Long-lived archive data can be accessed once a year.
- Retrieval time of minutes to hours.

Deep Archive

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

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.



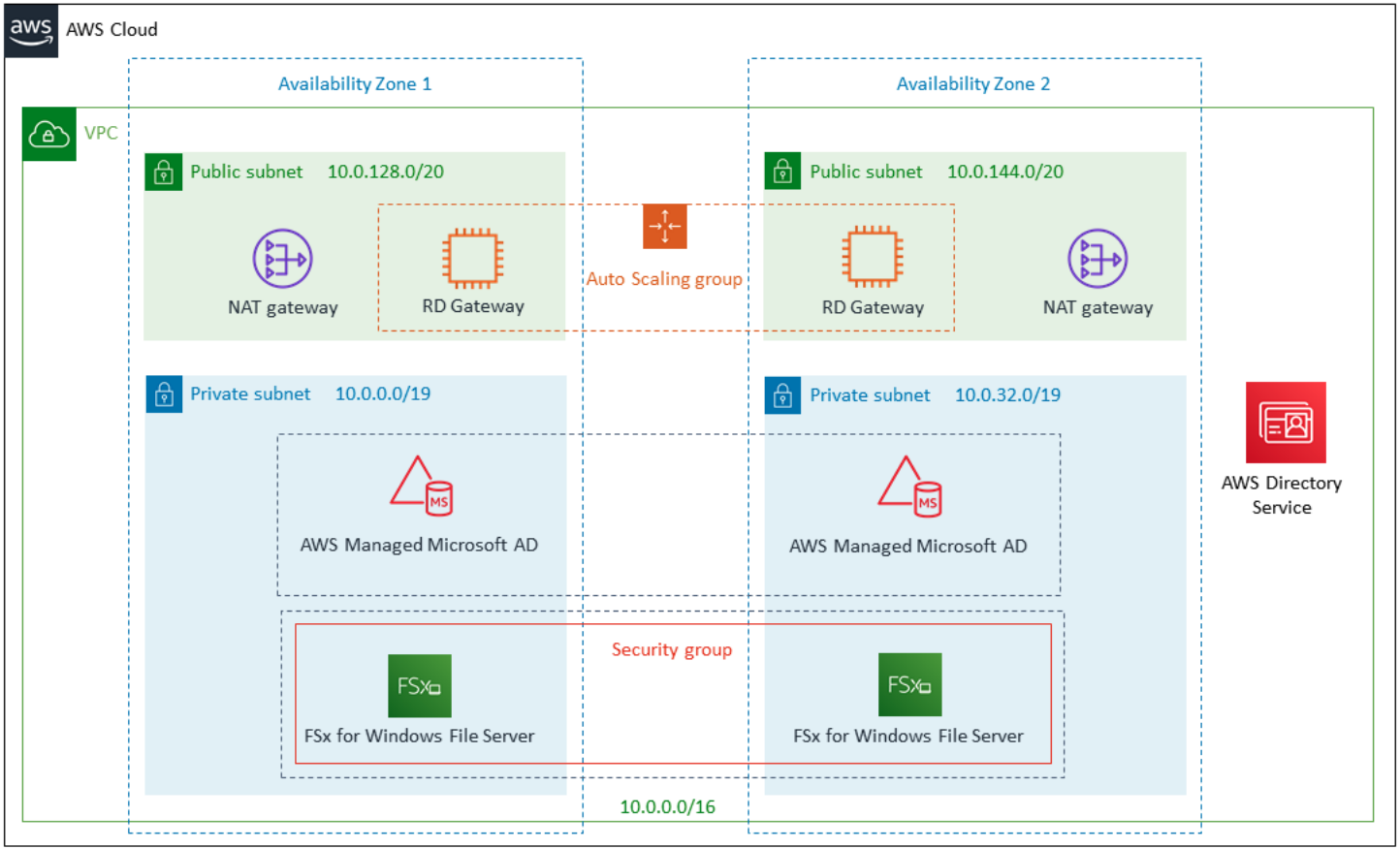
FSx for Windows File Server Performance Options

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

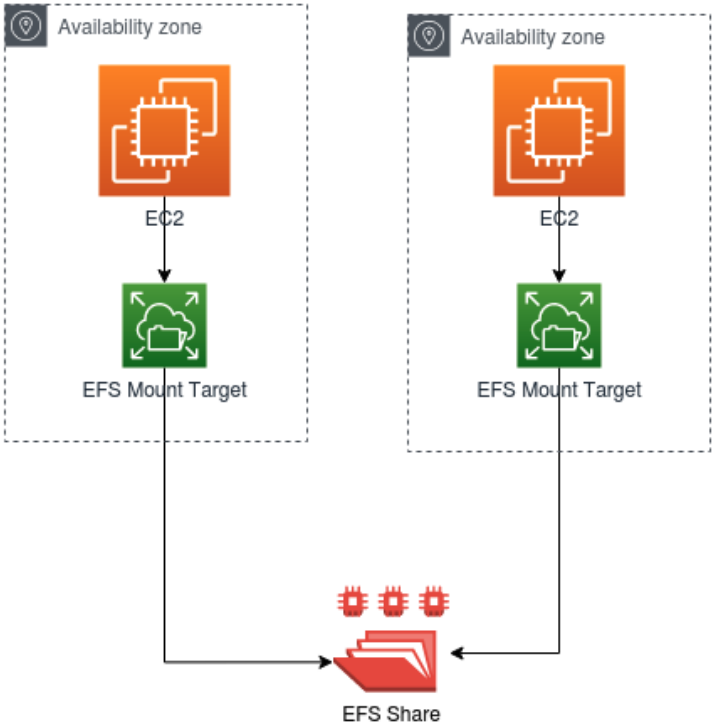
Throughput Capacity

- **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

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.

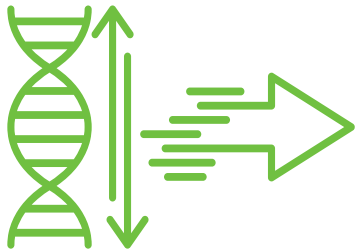

Availability zone
EC2
EFS Mount Target
Availability zone
EC2
EFS Mount Target
EFS Share

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.

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.

Amazon EBS

- High-performance block storage service designed for use with Amazon Elastic Compute Cloud (EC2)
- Throughput and transaction-intensive workloads

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)

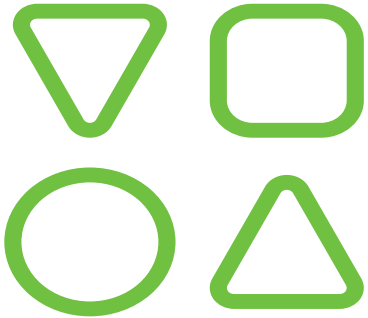
- Low-cost magnetic storage
- Throughput not IOPS
- Not bootable
- Baseline of 5 MB/sec. up to a cap of 500 MB/sec.

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 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.



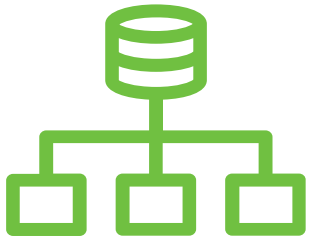
- System metadata - AWS date, size, storage class.
- User metadata - tags specified at the time the object is created.

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.

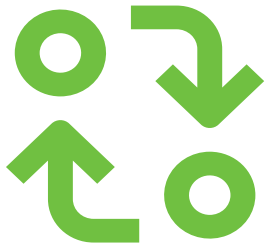
S3 on Outposts.

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

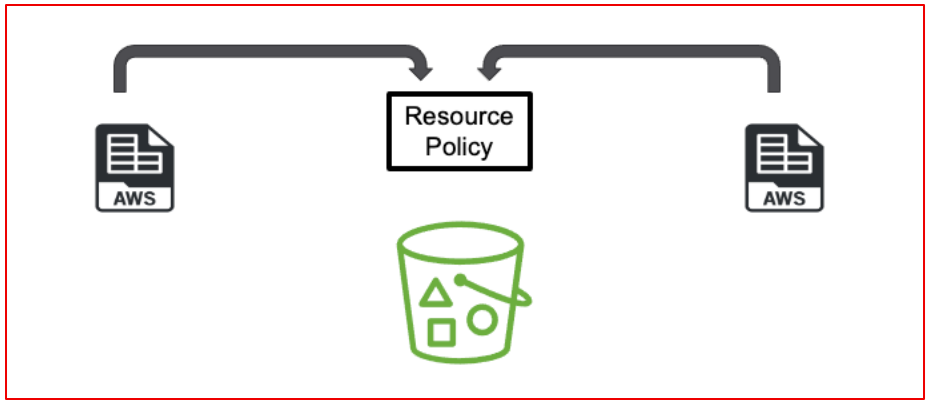
Enable dual-layer - SSE (Default SSE +KMS encryption)

S3 Security

- 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.

S3 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.



Task Statement 3.2

Design high-performing and elastic compute solutions.

Design high-performing and elastic compute solutions

- AWS EC2
- EC2 Auto Scaling
- AWS Auto Scaling

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 and are never shared with other AWS customers.

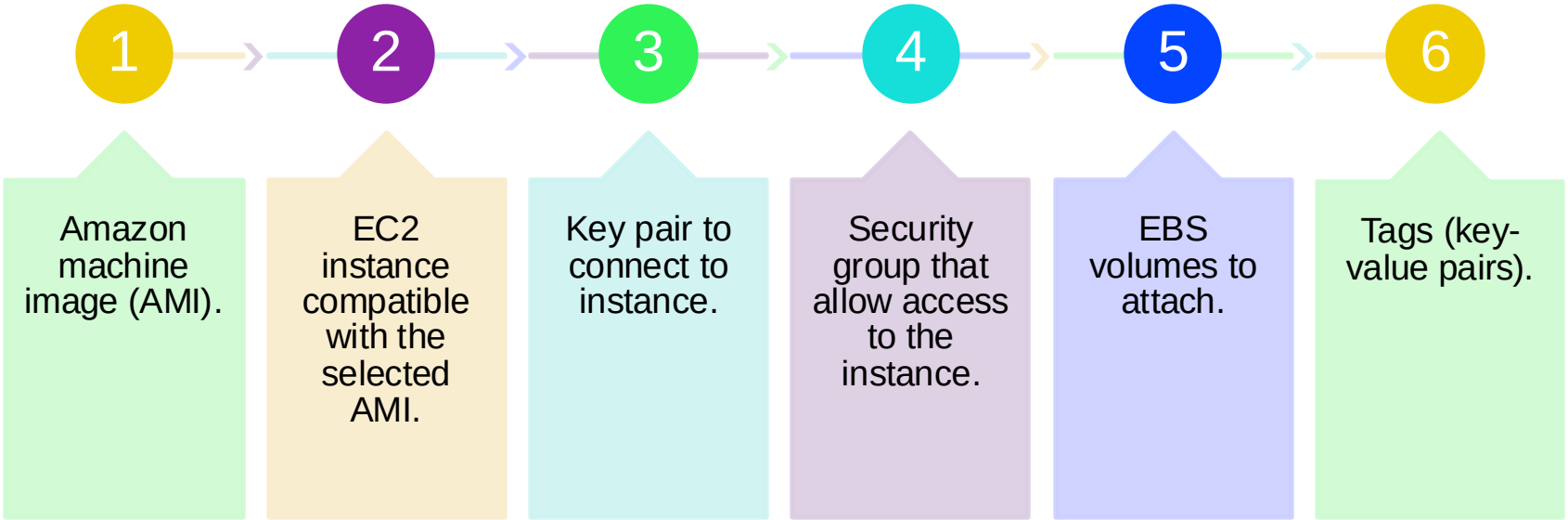
Dedicated Hosts

- 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.

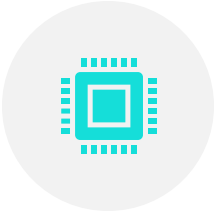
Dedicated Instance

- Physically isolated at the host hardware level from other dedicated instances and instances in different AWS accounts.
- Compliance with specific regulatory requirements that don’t permit multi-tenant virtualization.
- The attached EBS volumes don’t run on single-tenant hardware.

Launch Templates



EC2 Image Builder



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

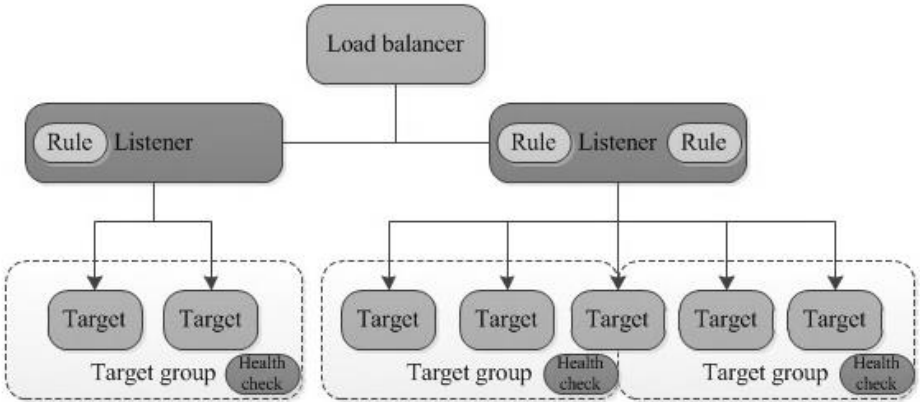
Images can be generated based on source AMIs.

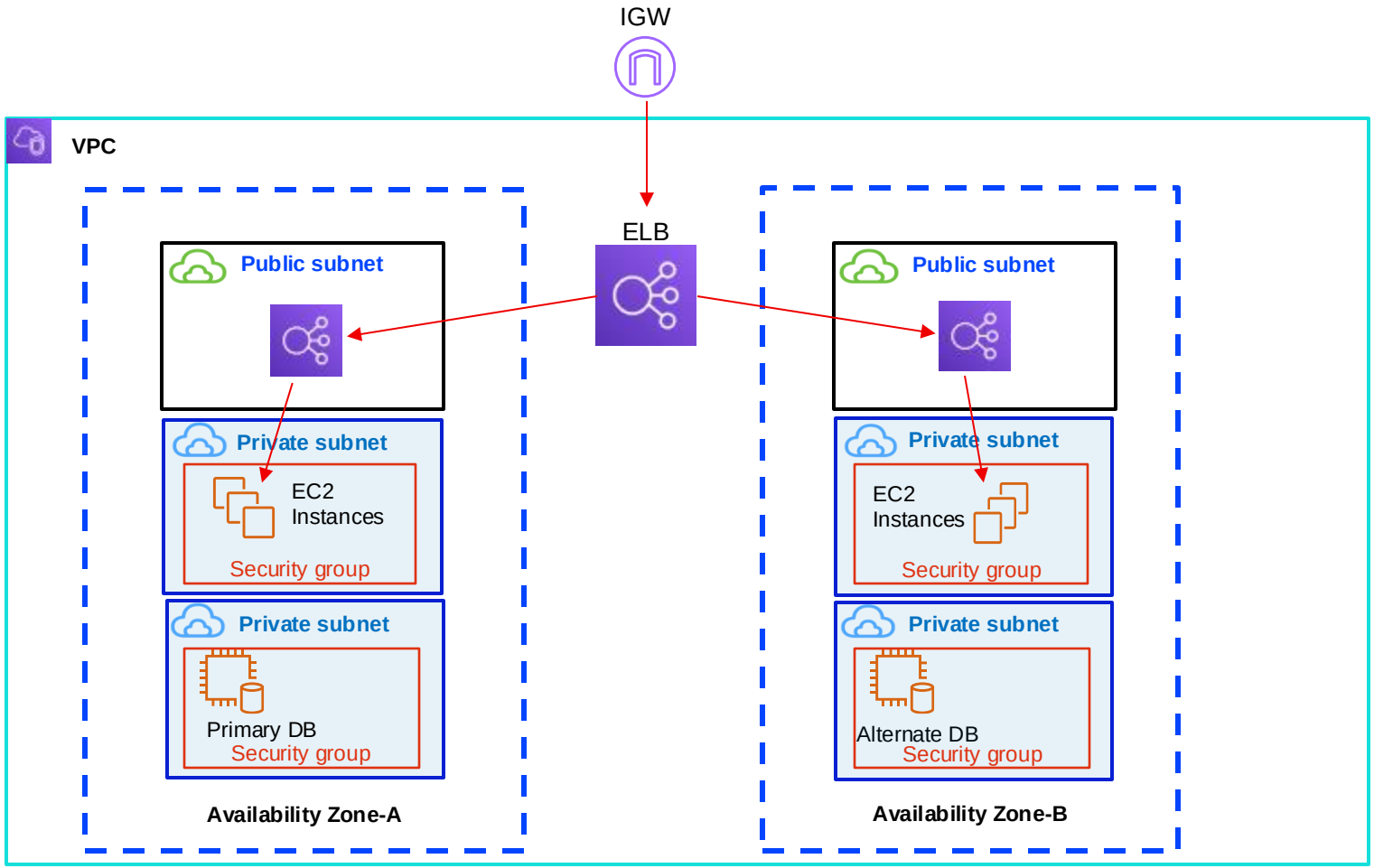
Define collections of security settings that you can edit, update, and use to harden your images.

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

Elastic Load Balancing Service

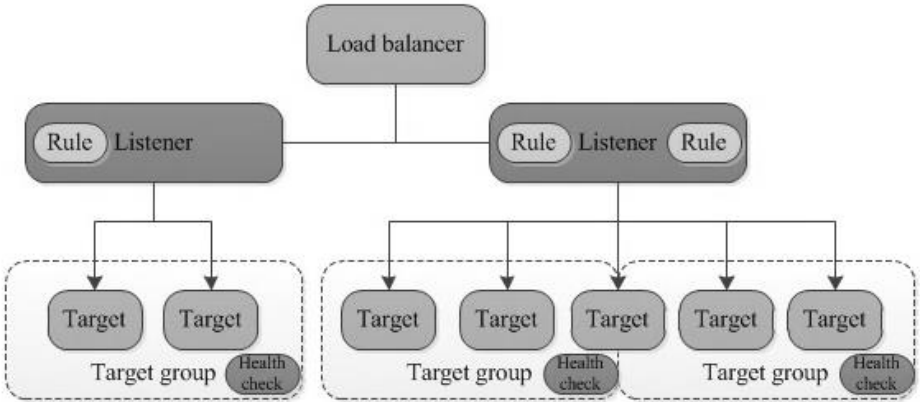
- Distribute incoming traffic across multiple EC2 instances or containers across one or multiple availability zones.
- Network load balancer: Target groups support the TCP, UDP, TCP_UDP, and TLS protocols.
- Application load balancer: Target groups support the HTTP and HTTPS protocols.




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.

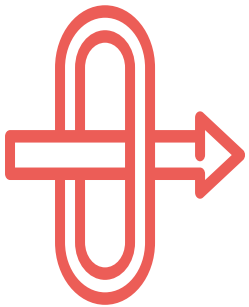


Health Checks

- 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

- Gateway Load Balancer (GWLB): Deploy and manage a fleet of 3rd party virtual appliances.

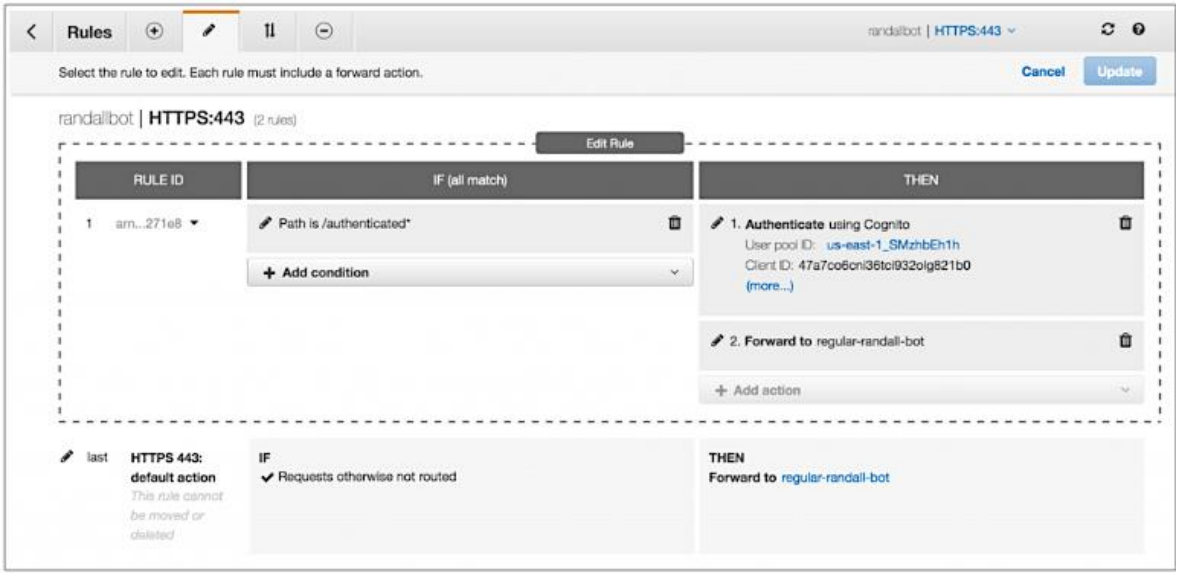


Application Load Balancer Features

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

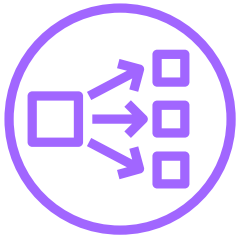
User Authentication

- 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.

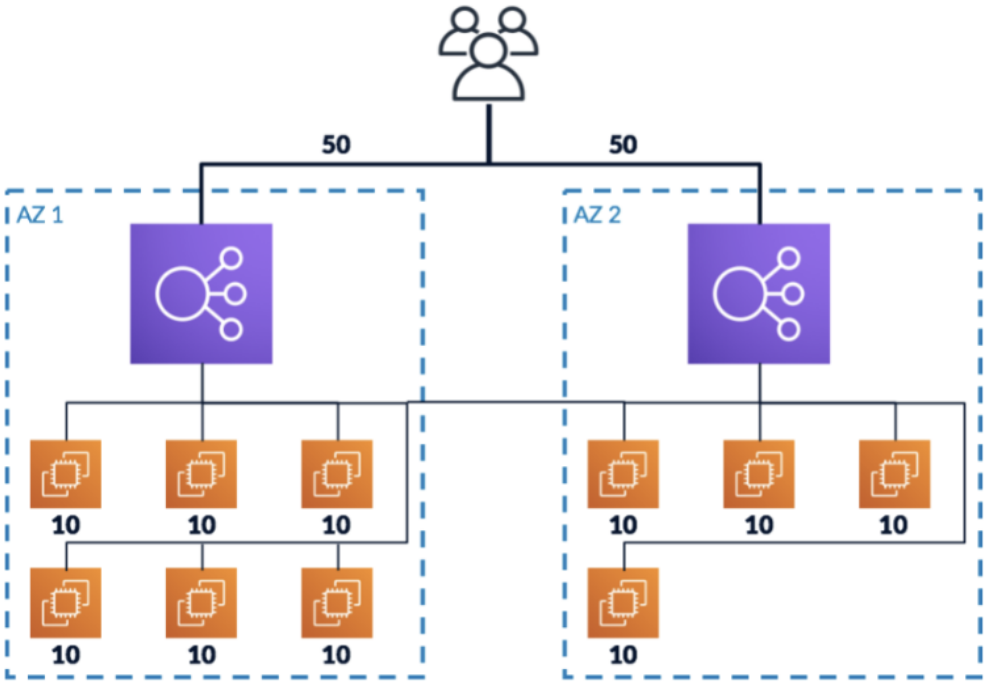


Network Load Balancer

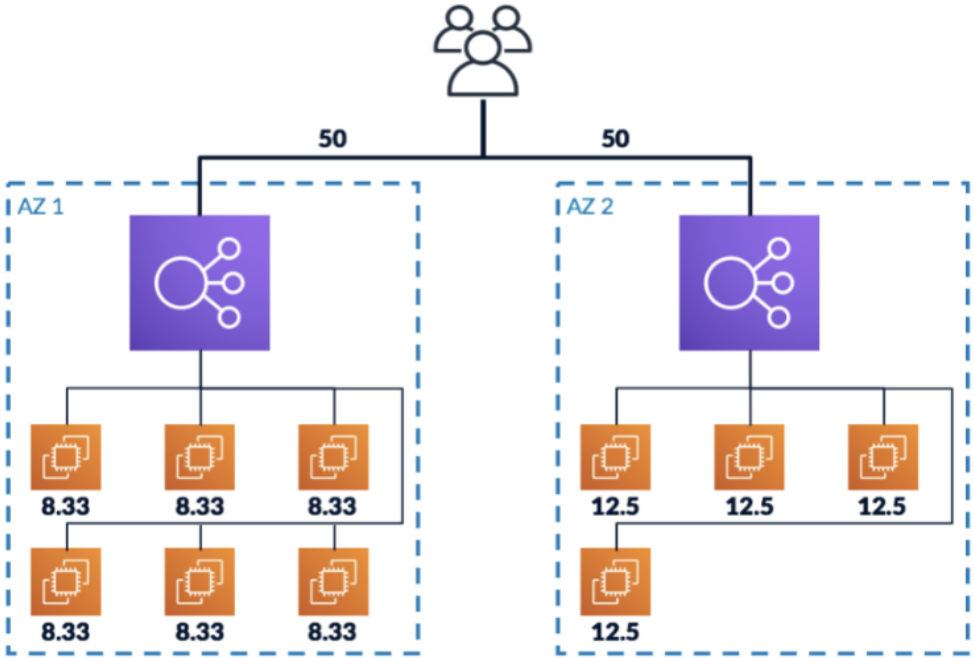
- End-to-end encryption uses TLS 1.3.
- Support connections from clients across VPC peering connections, Direct Connect and VPN connections.



Cross-Zone Load Balancing Enabled



Cross-Zone Load Balancing Disabled

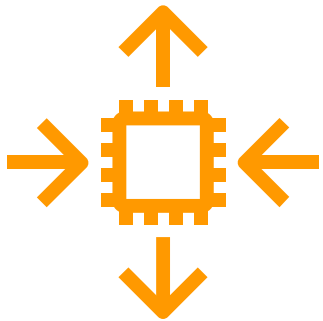


EC2 Auto Scale Concepts

- 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.

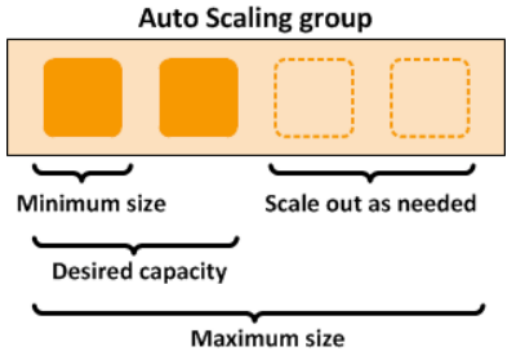
EC2 Auto Scale Components

- Auto Scaling Group – a managed group of EC2 instances.
- Launch template – Configuration template for adding EC2 instances.
- Scaling options – Auto scale groups scaling options.
- CloudWatch – Metrics linked to 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.

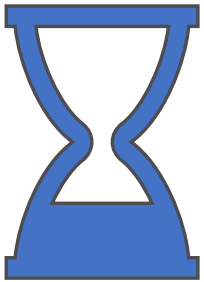


Scaling Policy Options

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 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 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.

Task Statement 3.3

Determine High-Performing Database Solutions

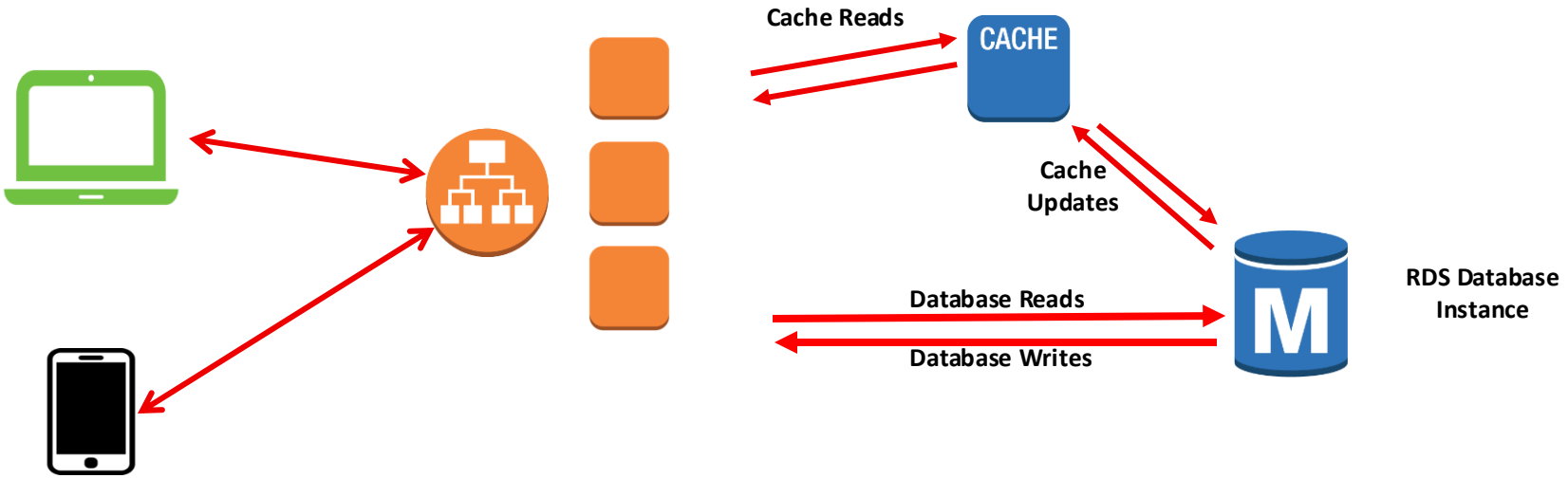
Determine High-Performing Database Solutions

- Amazon ElastiCache.
- Database engines with appropriate use cases.
- Database capacity planning (for example, capacity units, instance types, Provisioned IOPS).
- Determining an appropriate database type (Amazon Aurora, Amazon DynamoDB).
- Configuring read replicas.

Amazon ElastiCache

- Managed implementation of Redis and Memcached in-memory data stores (node).
- A node is a fixed size of network-secured RAM.
- Nodes are deployed in clusters utilizing the selected caching engine.

ElastiCache Operation

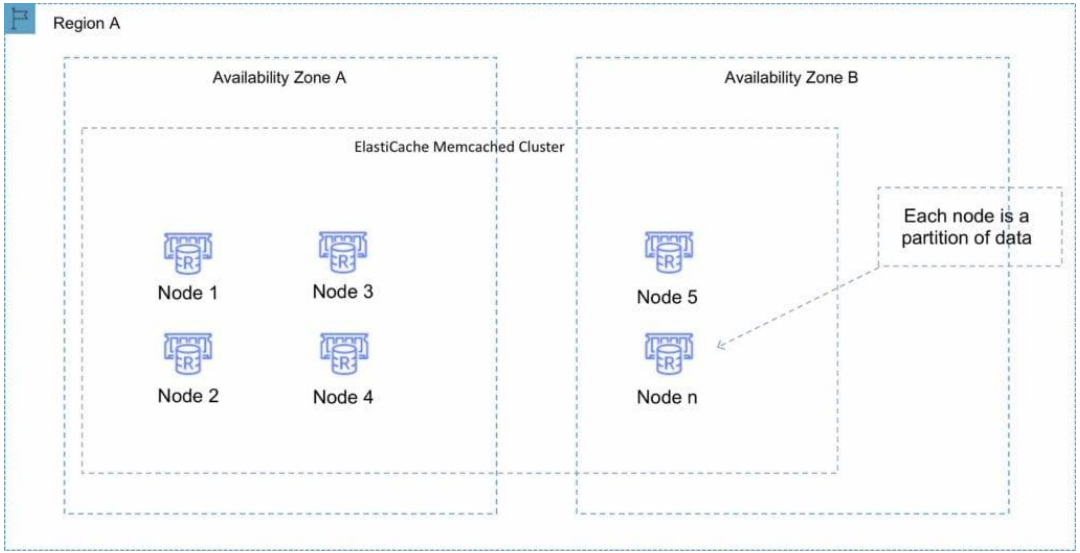


ElastiCache Use Cases

- User session storage.
- Database caching (RDS).
- Leaderboards for mobile applications.

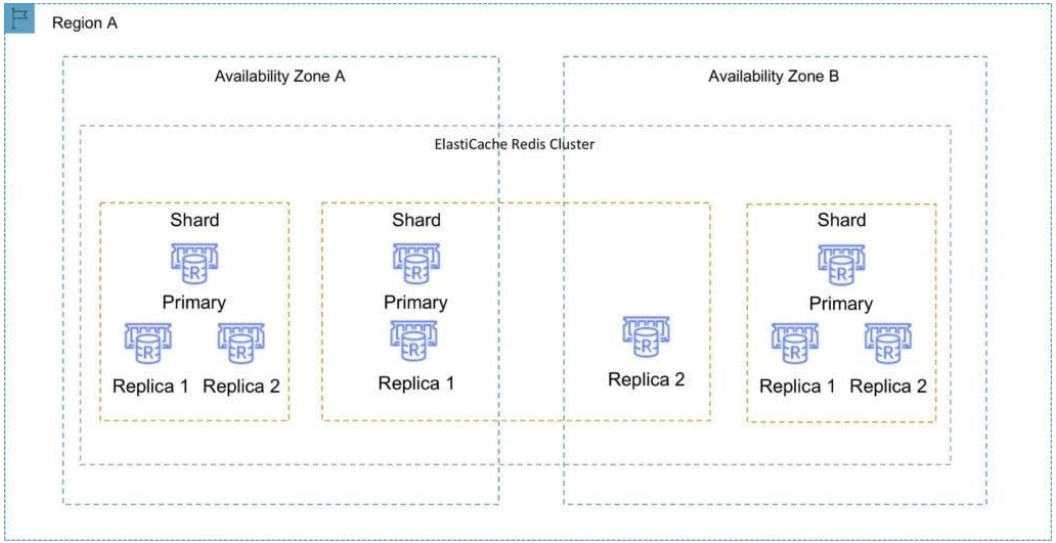
Memcached

- Can run large nodes with multiple cores and threads.
- Can be scaled in and out.
- Memcached is not persistent; node failure results in data loss.



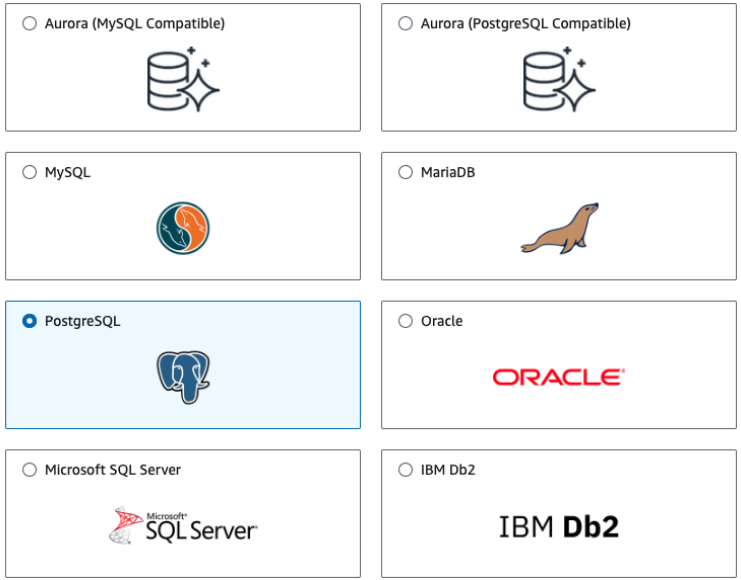
Redis

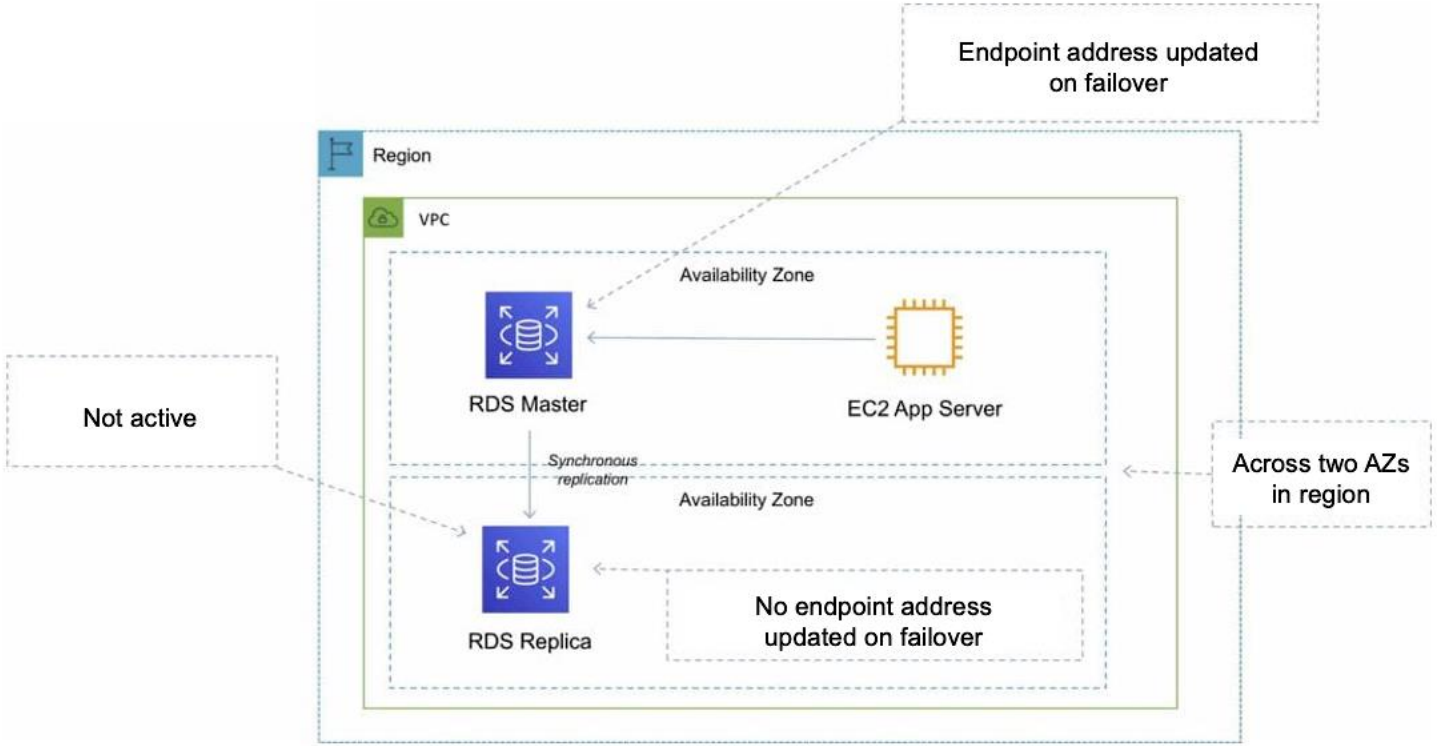
- Can be used as a data store.
- Data is persistent.
- Scales by adding shards that can include a primary node and optionally read replicas.
- Supports automatic snapshots and backups.
- Multi-AZ support for read replicas.



RDS Database Instances

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

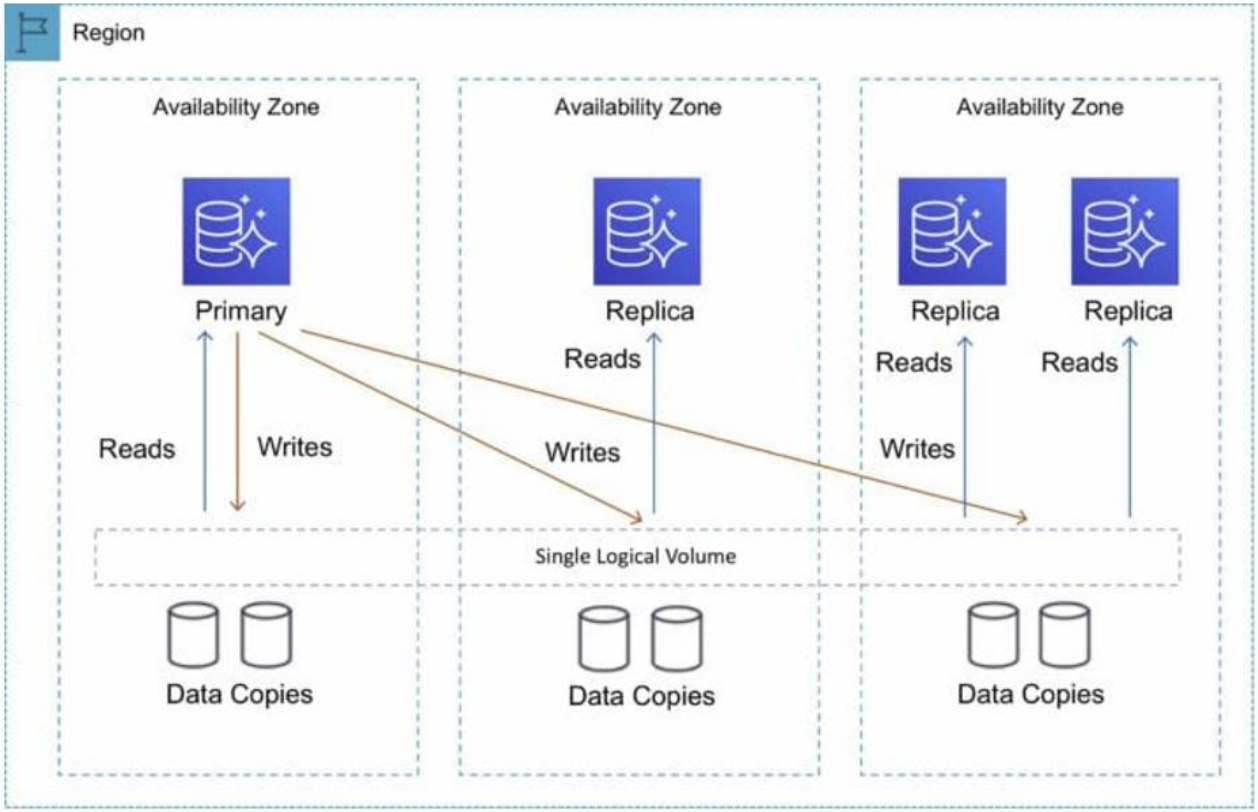



Endpoint address updated on failover
Region
VPC
Availability Zone
RDS Master
EC2 App Server
Not active
Synchronous replication
Availability Zone
Across two AZs in region
RDS Replica
No endpoint address updated on failover

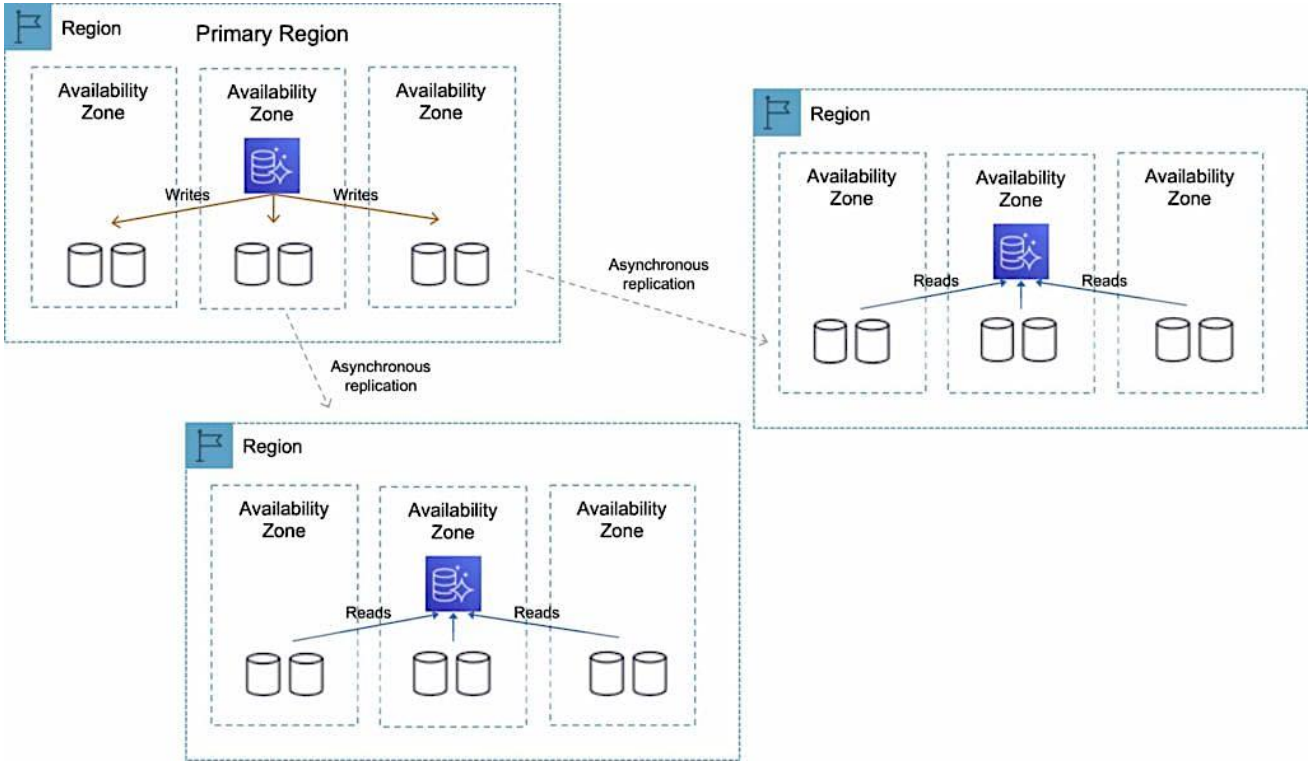
Amazon Aurora

- Fully managed 6-way replication across three availability zones with SSD Virtual SAN storage.
- Aurora can also 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

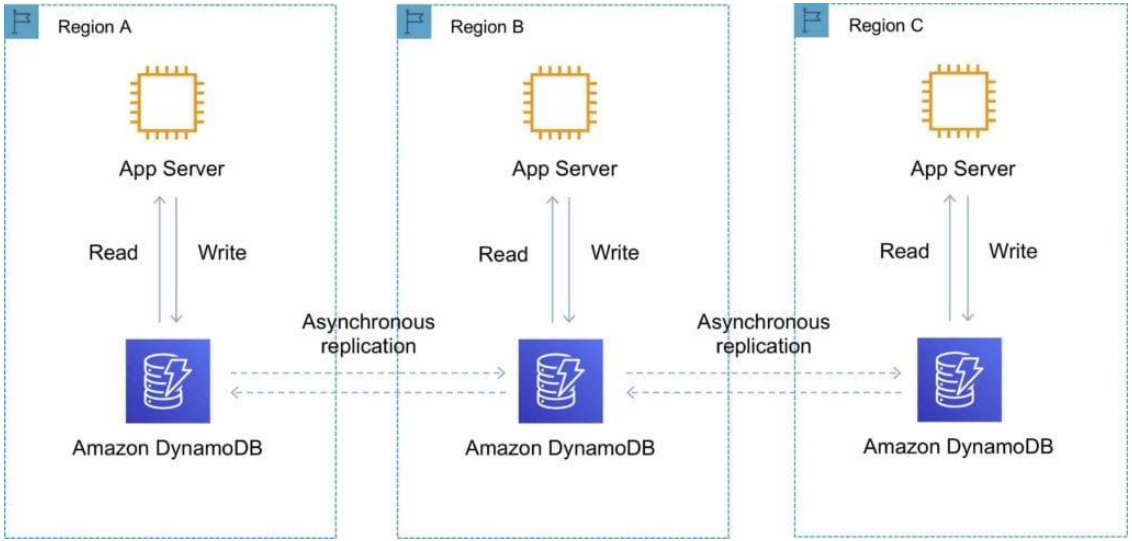


Amazon Aurora: Multi-region



Amazon 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.



Task Statement 3.4

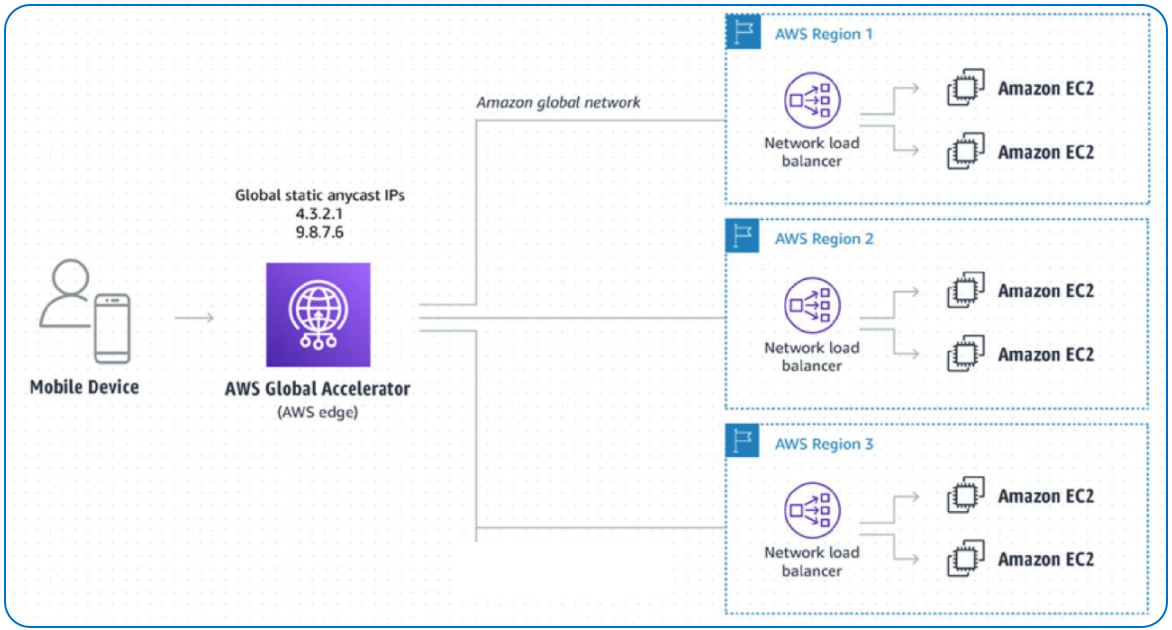
Determine High-Performing and scalable network architectures

Determine High-Performing and scalable network architectures

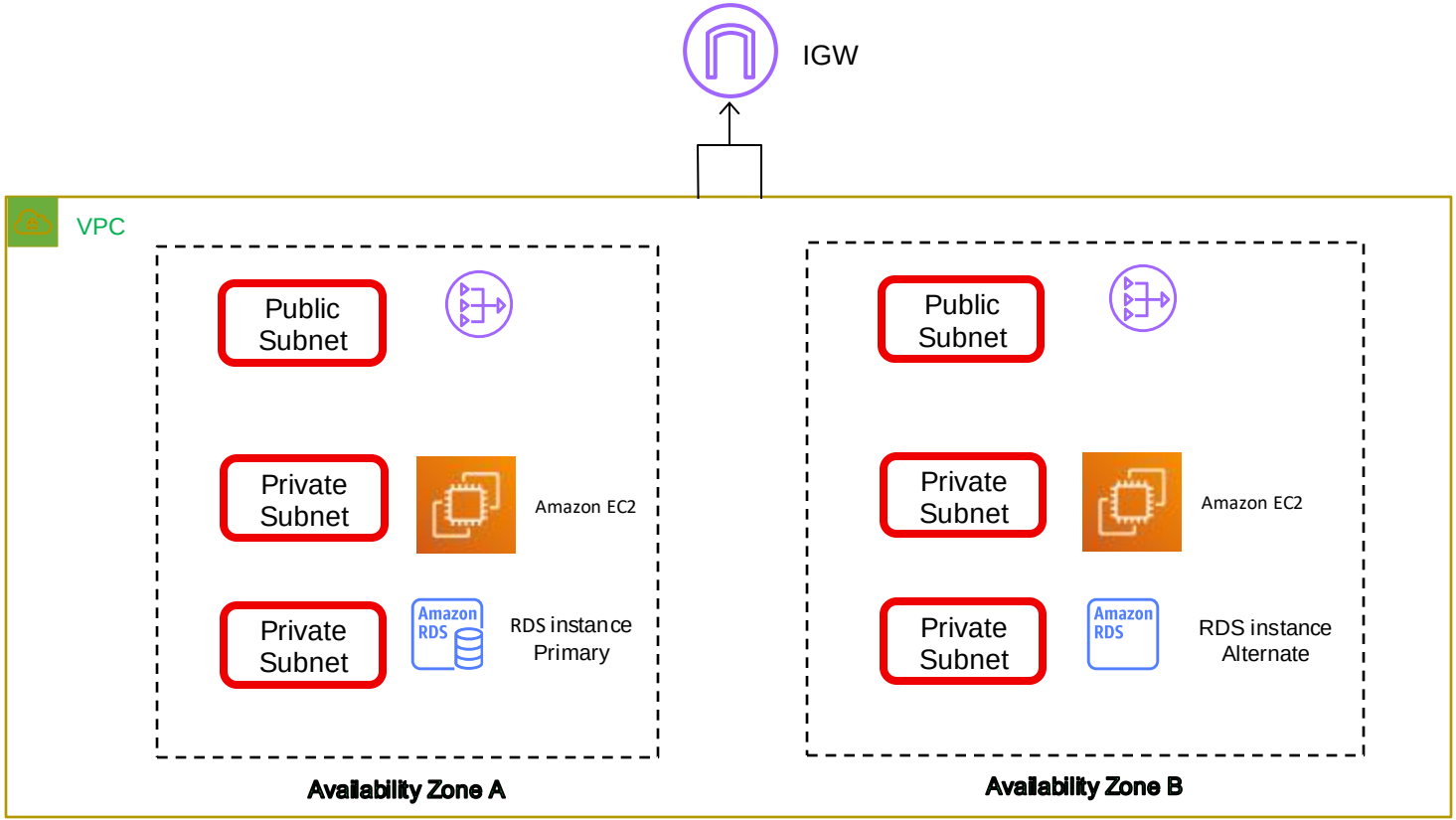
- AWS Global Accelerator
- How to design network architecture (Subnet IP addressing)

AWS Global Accelerator

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



Public -Private Subnets

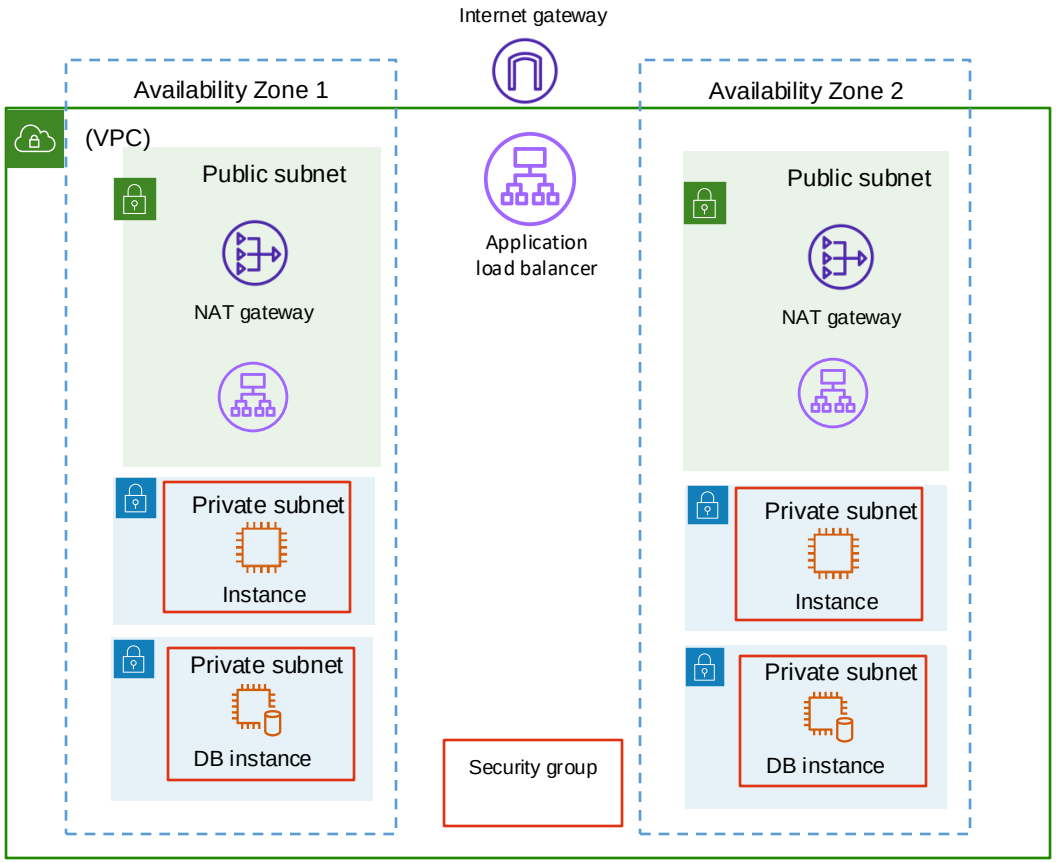


Subnets

- 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.


Internet gateway
Availability Zone 1
Availability Zone 2
(VPC)
Public subnet
NAT gateway
Application
load balancer
Public subnet
NAT gateway
Private subnet
Instance
Private subnet
Instance
Private subnet
DB instance
Private subnet
DB instance
Security group

IP Addresses

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

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

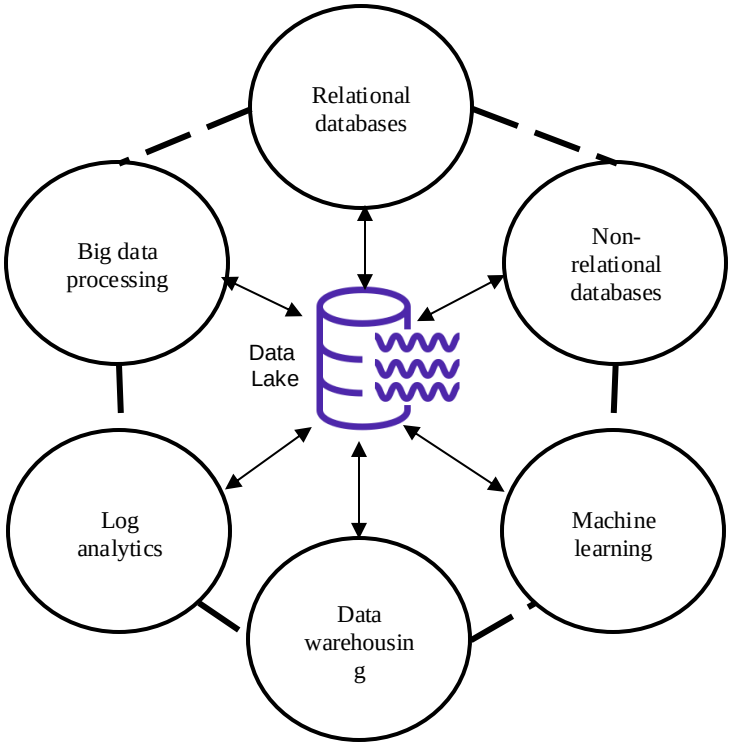
Task Statement 3.5

Determine High-Performing data Ingestation and transformation solutions.

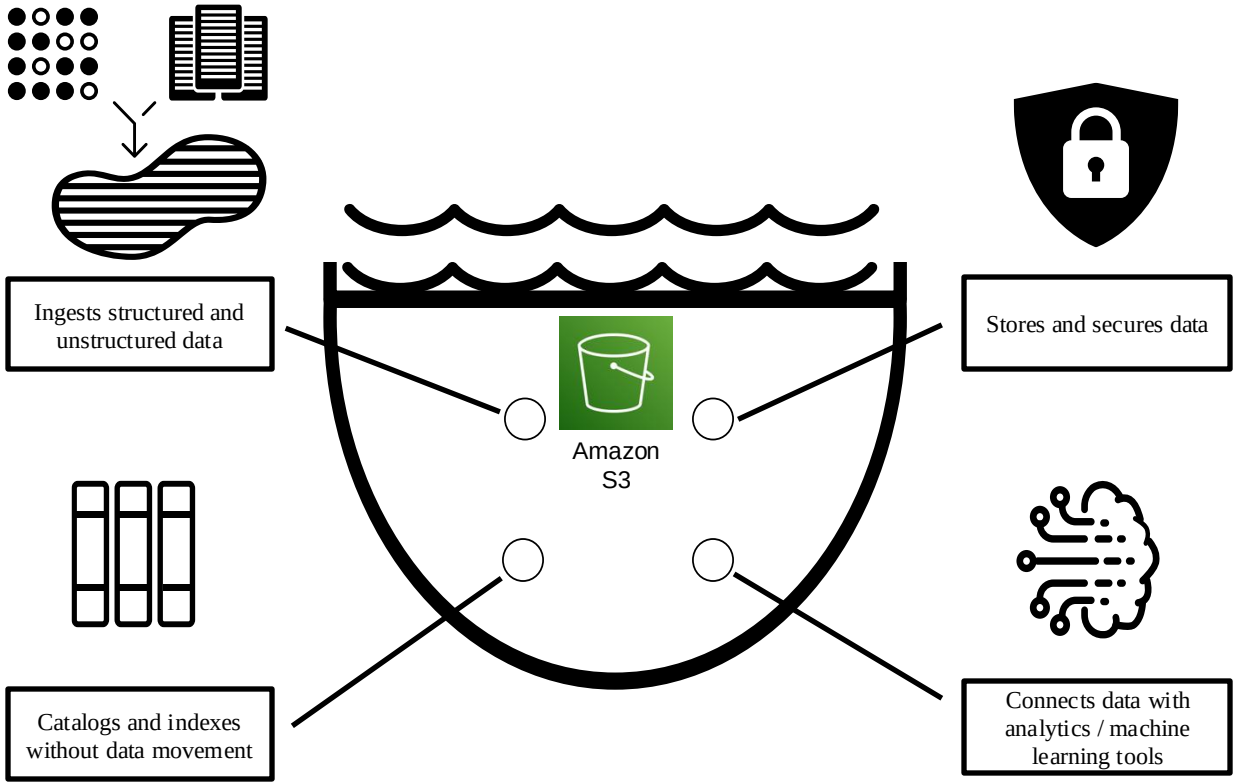
Determine High-Performing data Ingestation and transformation solutions

- Amazon Athena, AWS Lake Formation, Amazon QuickSight
- AWS Glue
- Amazon Kinesis
- AWS DataSync
- AWS Storage Gateway

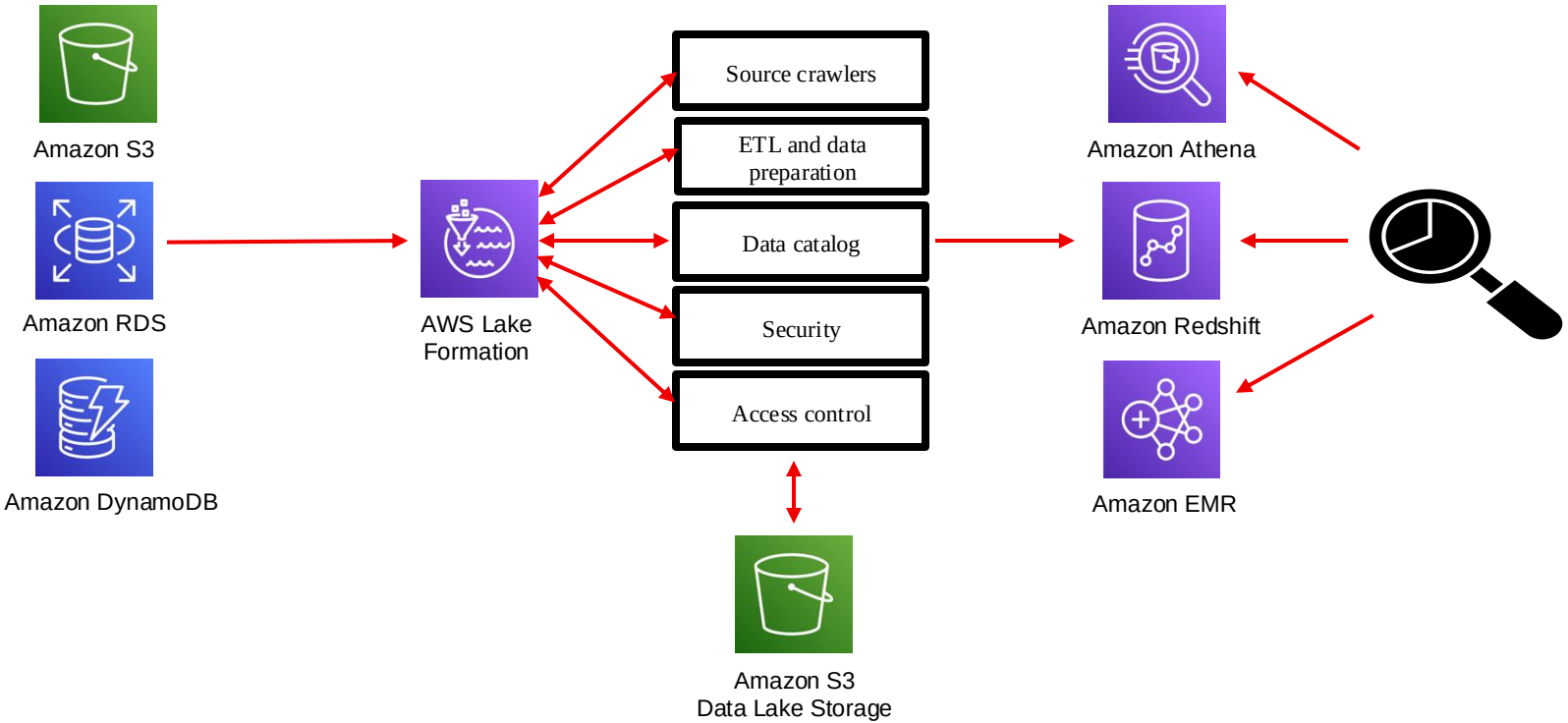
Corporate Data

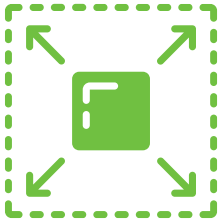


AWS Lake Formation



Data Lake Storage





Data Lake

Amazon S3

AWS Lake Formation

Analytics

Amazon Athena

Amazon EMR

Amazon Kinesis

Amazon Redshift

Data Movement

AWS Glue

AWS Glue Elastic Views

Governance

AWS Lake Formation Row-level security

Perfo rmance / Cost

Amazon EC2

Savings Plans

Amazon Kinesis

- Amazon Kinesis Data Streams capture, process, and store data streams at any scale.
    - Application monitoring, fraud detection, and live leaderboards.
    - Amazon Managed Streaming for Apache Kafka
- Amazon Kinesis Data Firehose - Capture, transform (extract, transform, and load ETL) data to data lakes, data stores, and analytics services.
    - Stream data from IoT devices, video streams.
- Amazon Kinesis Video Streams - Securely stream video from connected devices to AWS for analytics, ML, playback, and other processing.

Domain 4

Design Cost-Optimized Architectures

Task Statement 4.1

Design Cost-Optimized Storage Solutions.

Design Cost-Optimized Storage Solutions

- Cost Allocation Tags
- Cost Management Tools (Cost Explorer, AWS Budgets, AWS Cost and Usage Report)
- Data Transfer Family
- S3 Object Lifecycles

EBS Pricing Considerations

Volumes

- Charged by gigabytes provisioned per month
- Varies by volume type selected

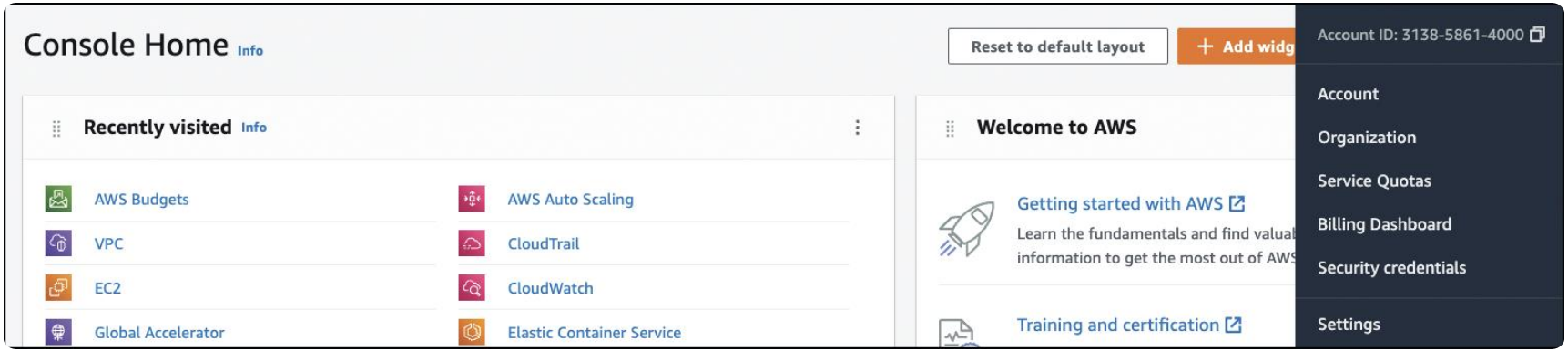
Snapshots

- Charged by storage space used in S3
- Charged for snapshots copied across regions

Data Transfer

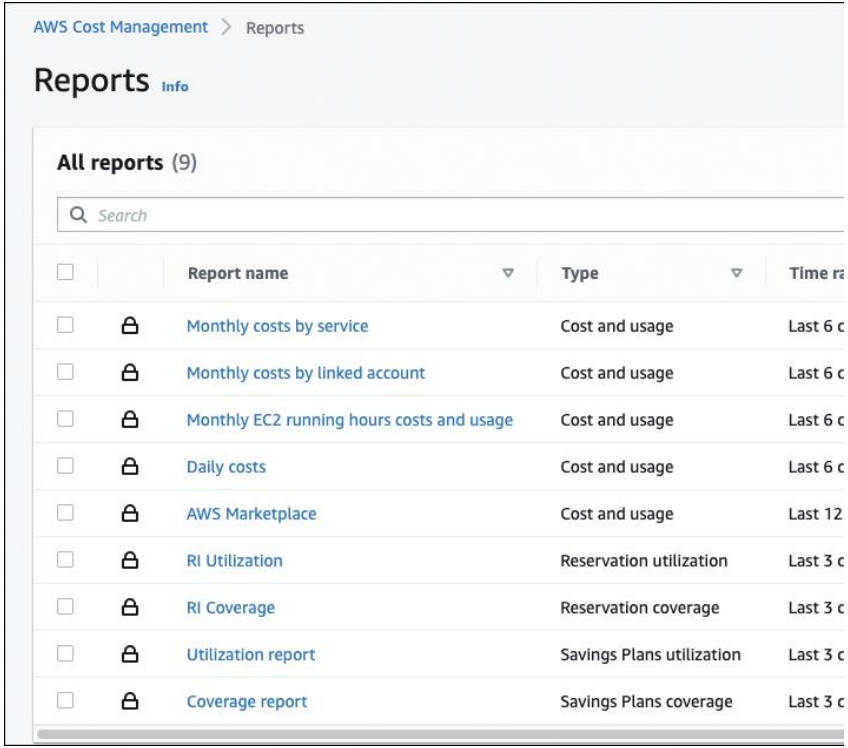
- Inbound data transfer is free
- Outbound data transfer charges

Billing Dashboard



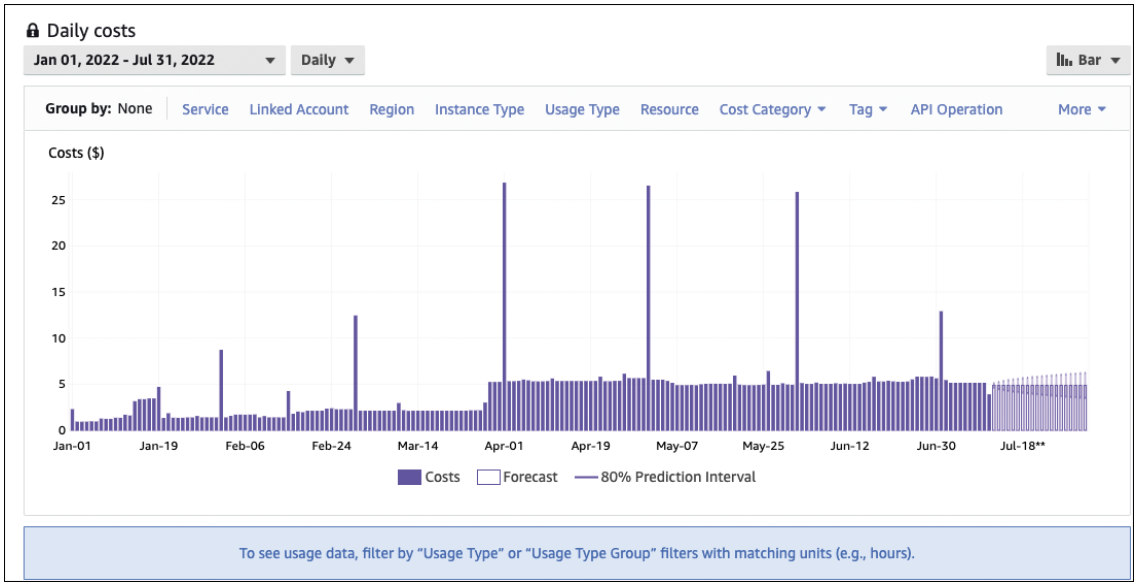
Cost Explorer

- 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.



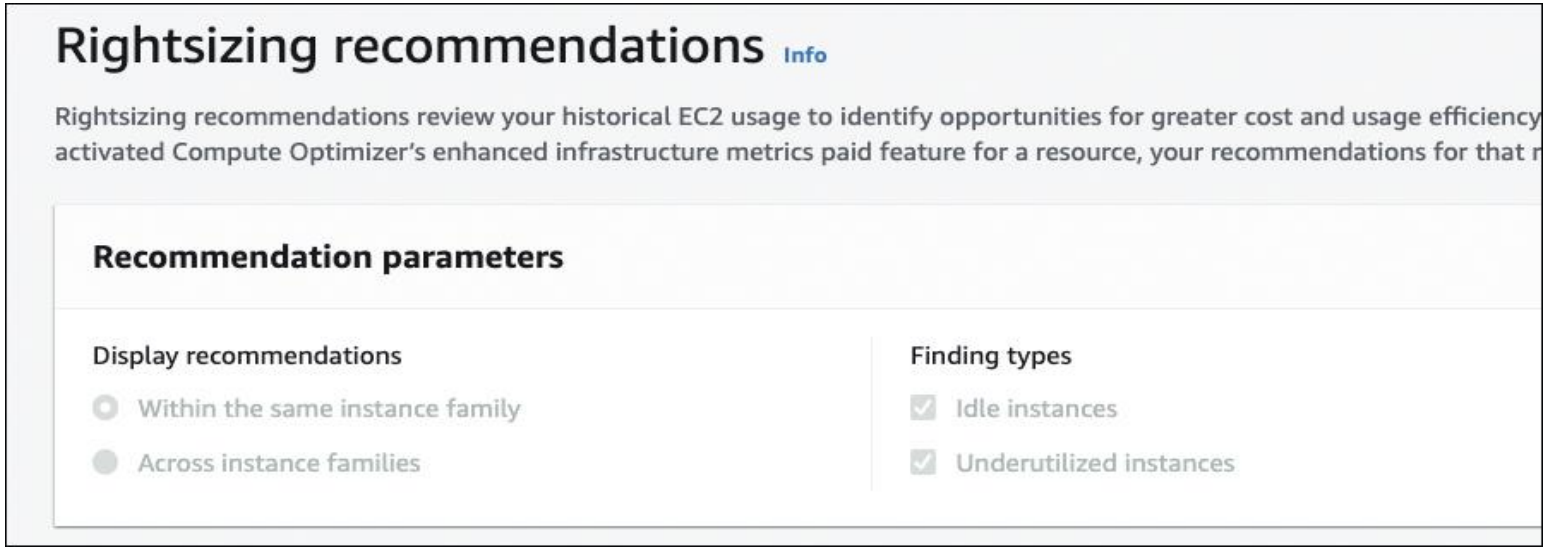
AWS Cost and Usage Reports

- 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.



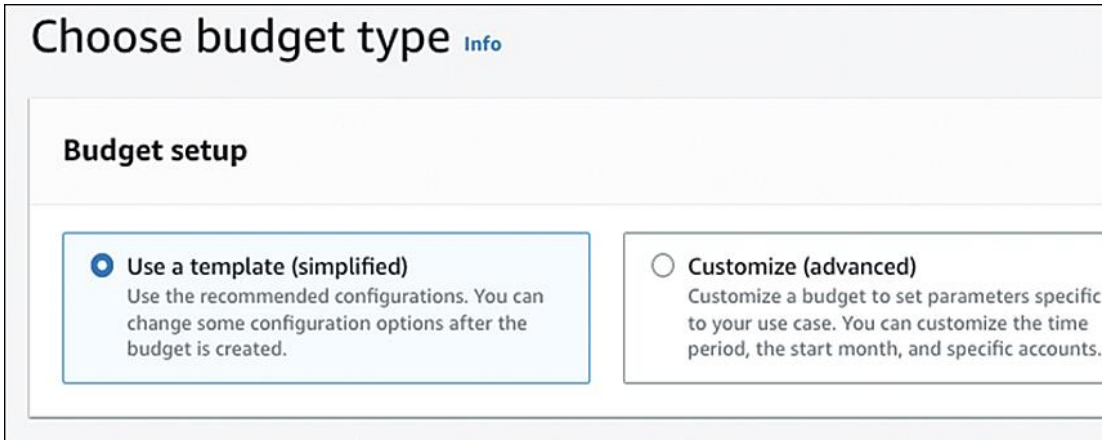
Rightsizing Recommendations

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



AWS Budgets

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



Task Statement 4.2

Design Cost Optimized Compute Solutions.

Design Cost Optimized Compute Solutions.

- Spot Instances, Reserved Instances, Savings Plans
- Cost-effective AWS compute services (EC2, Lambda, Fargate)
- Availability of workload types (Production, Non-production)

EC2 Pricing Considerations

Server runtime	Instance type	Pricing model
Number of instances	Load Balancing	Auto Scaling
Detailed monitoring enabled	Elastic IP addresses	Operating system

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 are defined by a 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).

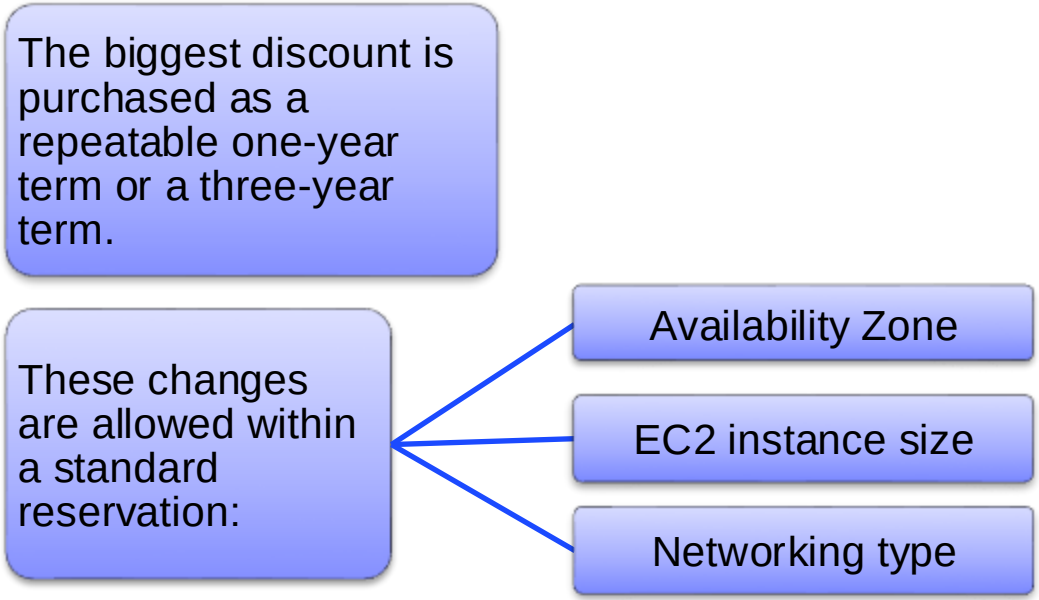
Reserved instances quota limits are based on the region and the number of availability zones.

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 request is spare EC2 compute capacity that AWS is not currently using. Save up to 90% of the purchase price.

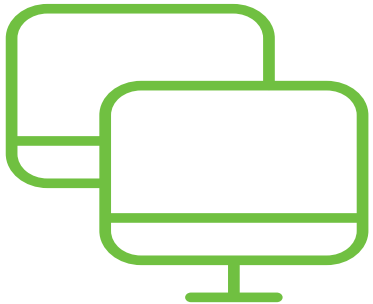
Spot instance pricing is based on supply and demand; the instance will run until the EC2 service needs the capacity back.

A two-minute warning is provided when AWS needs the instance back (CloudWatch alert).

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

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.
- Two types of Spot Fleet requests: 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 savings of up to 72% 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.

Task Statement 4.3

Design Cost-optimized Database Solutions.

Design Cost-optimized Database Solutions.

- Database replication
- Database capacity planning (capacity units)
- Backup and retention policies
- Determining appropriate database engines (Amazon RDS, Aurora Serverless, time series format, columnar format.

Database concepts

- Amazon RDS - choose EC2 size and EBS storage.
- Amazon Aurora - choose EC2 size and EBS storage.
- Amazon DynamoDB - Set target utilization levels, and DynamoDB will adjust the read (RCU) and write capacity (WCU) to match the desired performance.
- Amazon Aurora Serverless v2
  - Choose minimum and maximum capacity ranges.
  - Storage automatically scales (10GB to 64 TB)

- “Continuous data backup” is stored in Amazon S3 storage.

Task Statement 4.4

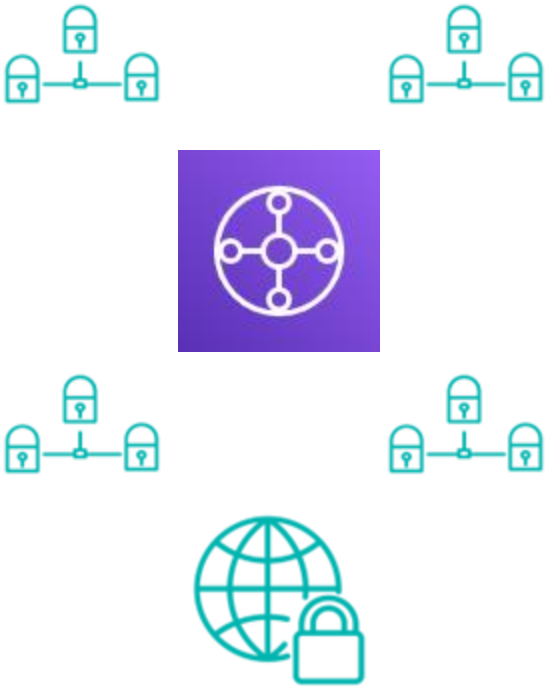
Design Cost-Optimized Network Architectures.

Design Cost-Optimized Network Architectures.

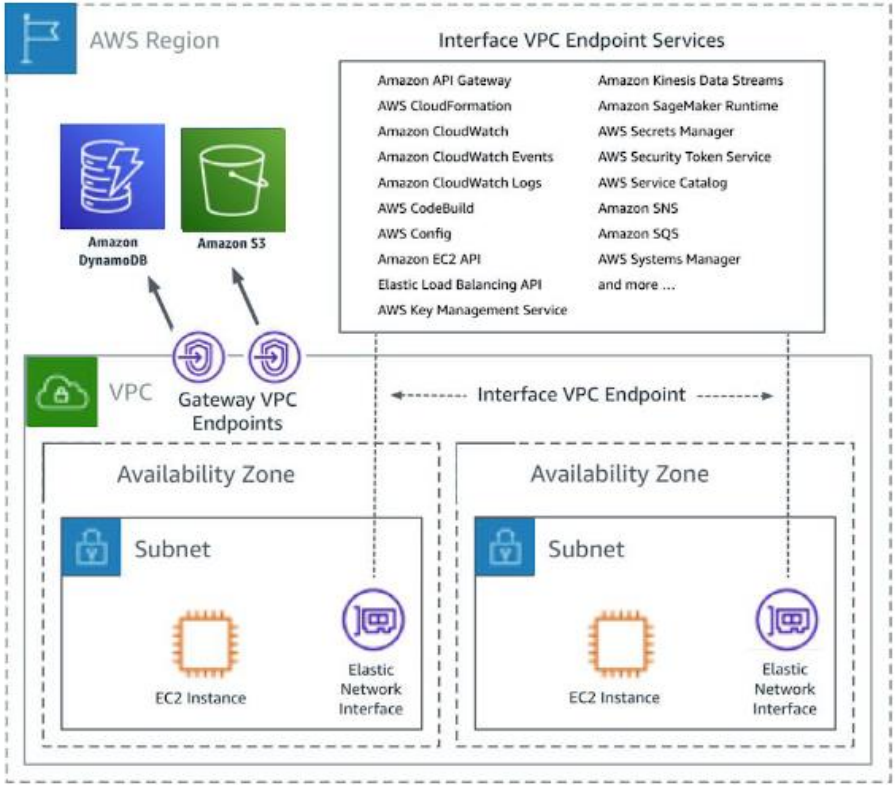
- Load balancing concepts.
- NAT Gateways (Single- multi-AZ).
- Network connectivity.
- Routing and peering (AWS Transit Gateway, VPC Peering).
- Minimize data transfer costs (Region to Region, Availability Zone to Availability. Zone, private to public, Global Accelerator, VPC endpoints).
- CDN edge caching.

Transit Gateway

- VPCs connect to the Transit gateway using a hub and spoke model.
- Transit gateway traffic remains on AWS's private network.
- Supports dynamic and static routing between VPCs and VPN, Direct Connect connections.
- Custom transit gateway route tables allow you to segment and isolate your networks.
- Connect AWS regions using Transit Gateway Peering connections.



VPC Endpoints

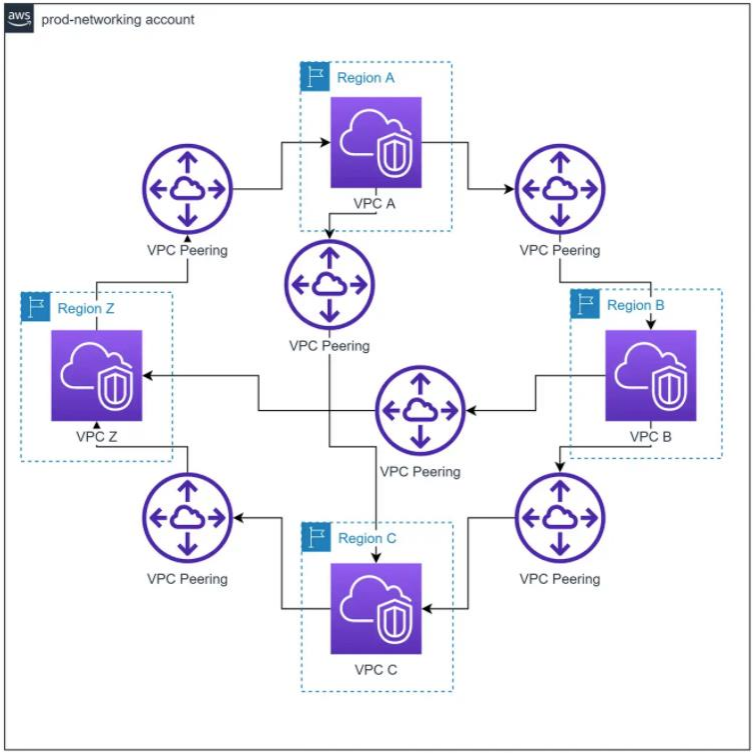


Endpoint Services

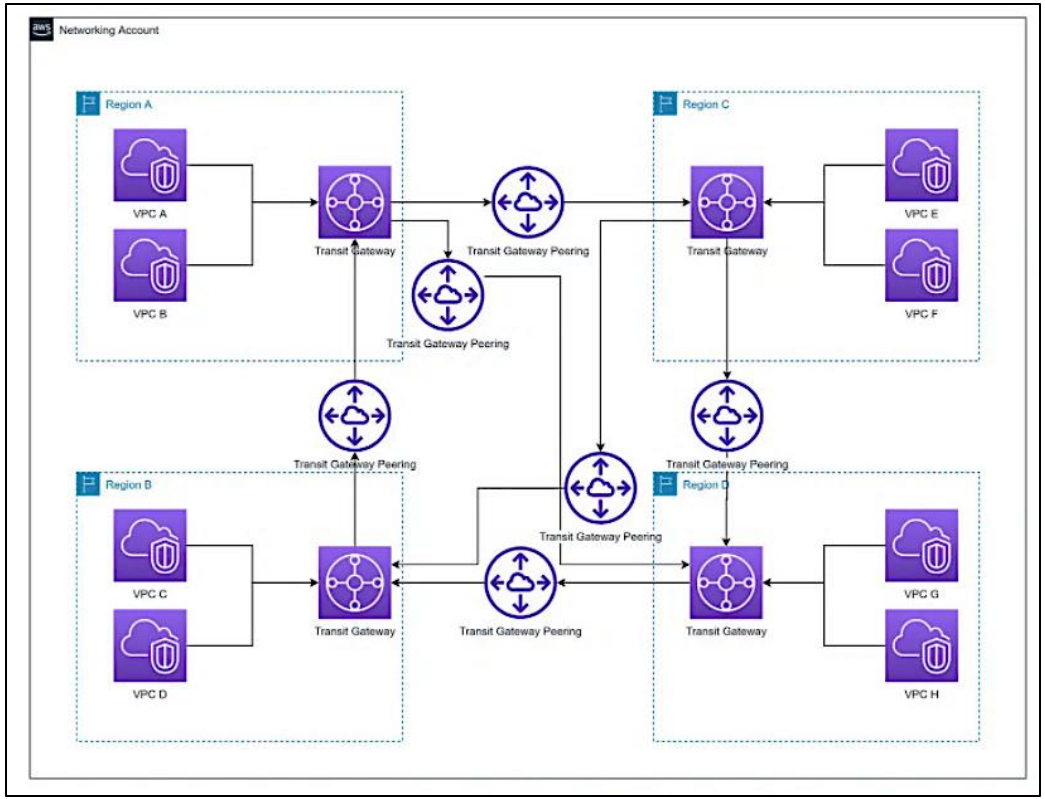
- Provides private connectivity between VPCs, AWS services, and on-premises applications securely on the Amazon network
- Connect services across different accounts and VPCs to simplify the network architecture significantly.
- Create endpoints to AWS or Marketplace service in subnets.
- Applications in a VPC can securely access PrivateLink endpoints across regions using Inter-Region VPC Peering.

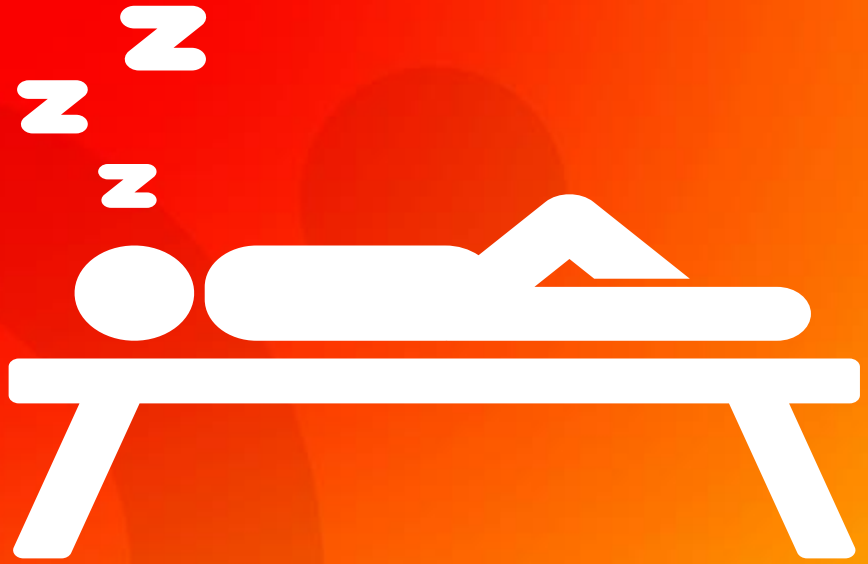
Peering VPC’s

- 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.


prod-networking account
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
VPC Peering

Transit Gateway Architecture






O'REILLY®