AWS Certified Solutions Architect - Professional

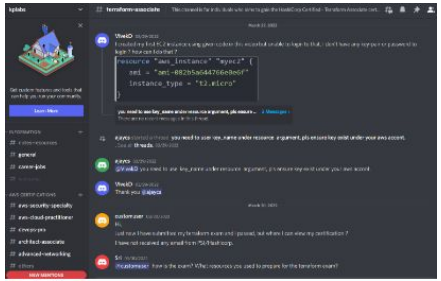
Instructed by Zeal Vora

Our Community

You can join our Discord community for any queries / discussions. You can also connect with other students going through the same course in Discord (Optional)

Discord Link: http://kplabs.in/chat

Group: #architect-professional



PPT Version

PPT Release Date = 12th Feb 2023

We regularly release new version of PPT when we update this course.

Please check regularly that you are using the latest version.

The Latest Version Details are mentioned in the PPT Lecture in Section 1.

Multi-Account Strategy

Challenges and Structure

Challenges with Multi-Account Architecture

Multiple Account provides highest amount of resources and security isolation.

When an organization has multiple AWS accounts, they need to consider following aspects:

- Identity Account Architecture
- Logging Account Architecture
- Publishing Account Structure
- Billing Structure



Identity Account Architecture

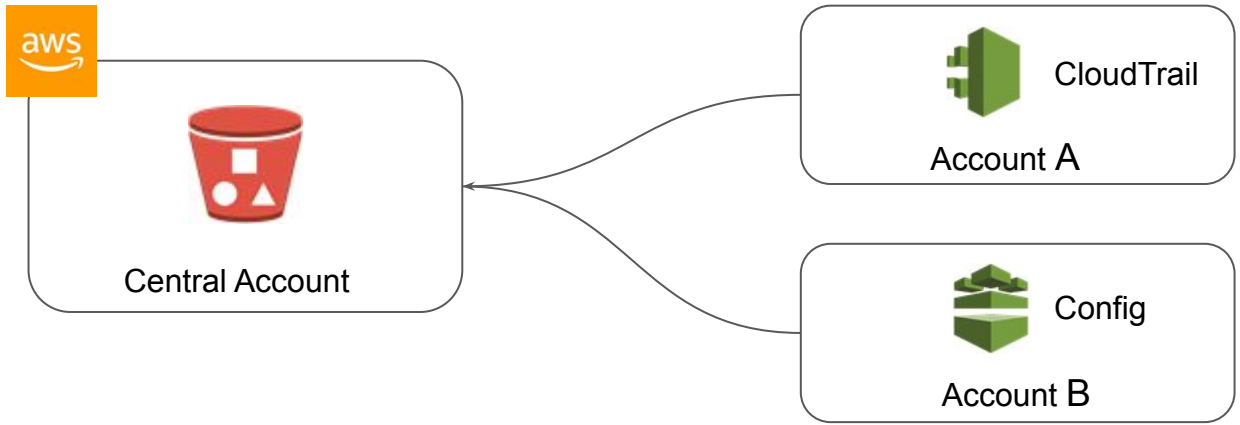
It is recommended to manage all the users at a single place and allow them to access multiple resources from other accounts

This can be achieved with the help of cross account IAM roles / Federations.



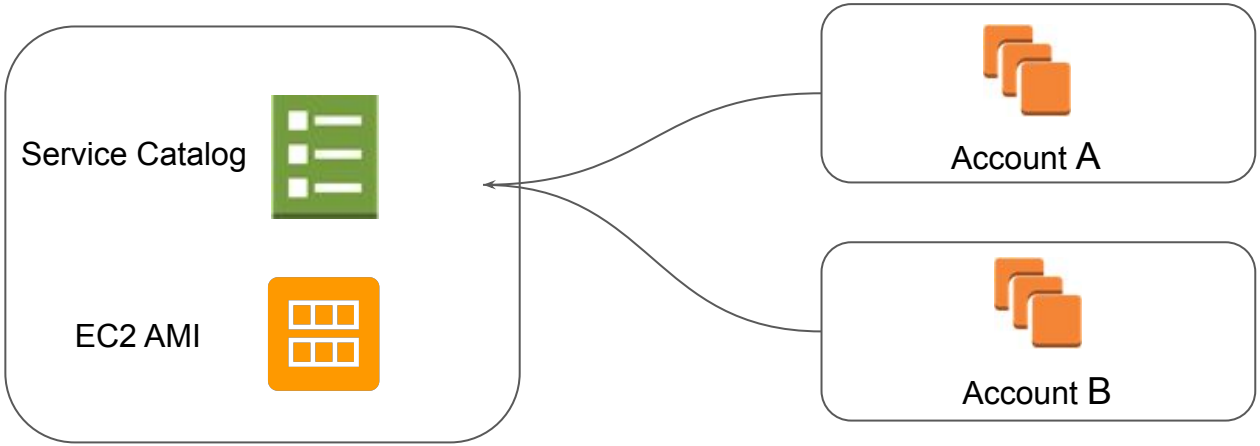
Logging Account Structure

Logs should be stored at centralized place where they can be monitored and analyzed in regular basis.



Publishing Account Architecture

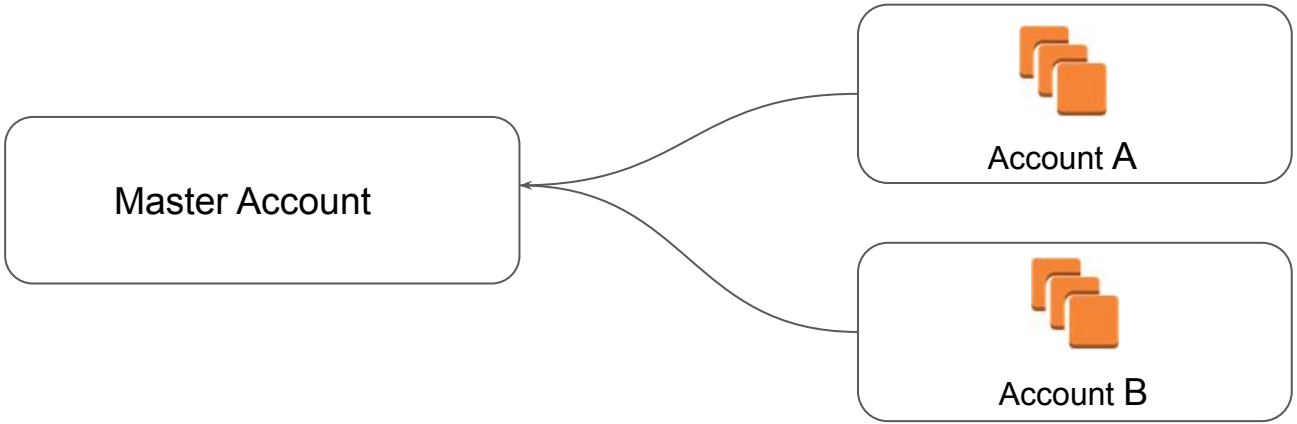
This account structure can be beneficial for customers who want to centrally manage pre approved server images and AWS CloudFormation templates across a company.



Billing Structure

You can use the consolidated billing feature in AWS Organizations to consolidate billing and payment for multiple AWS accounts accounts

One Bill + Easy Tracking + Combined Usage + No Extra Fee



Identity Account Architecture

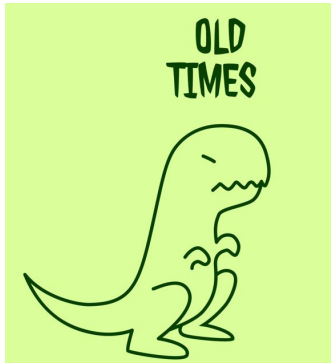
Multiple Accounts are Good

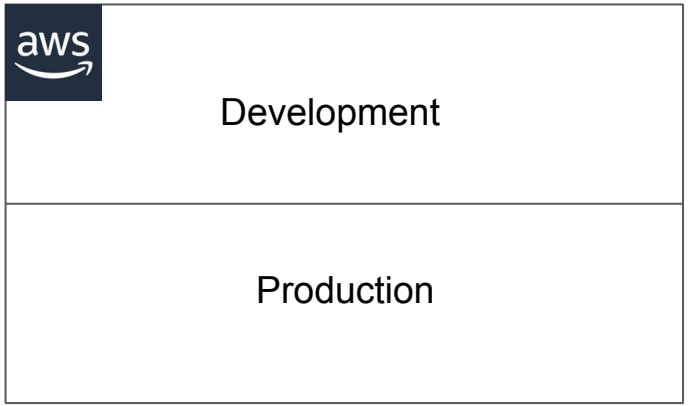
The Initial Start

During the earlier days of AWS, most of the organizations had a single AWS account.

Management was simple.

User would had a single set of username/password AND access/secret keys.





Organizations Became Big

A better architecture with multiple AWS account per function was adopted.

Each user had different username/password AND access/secret keys for each account.

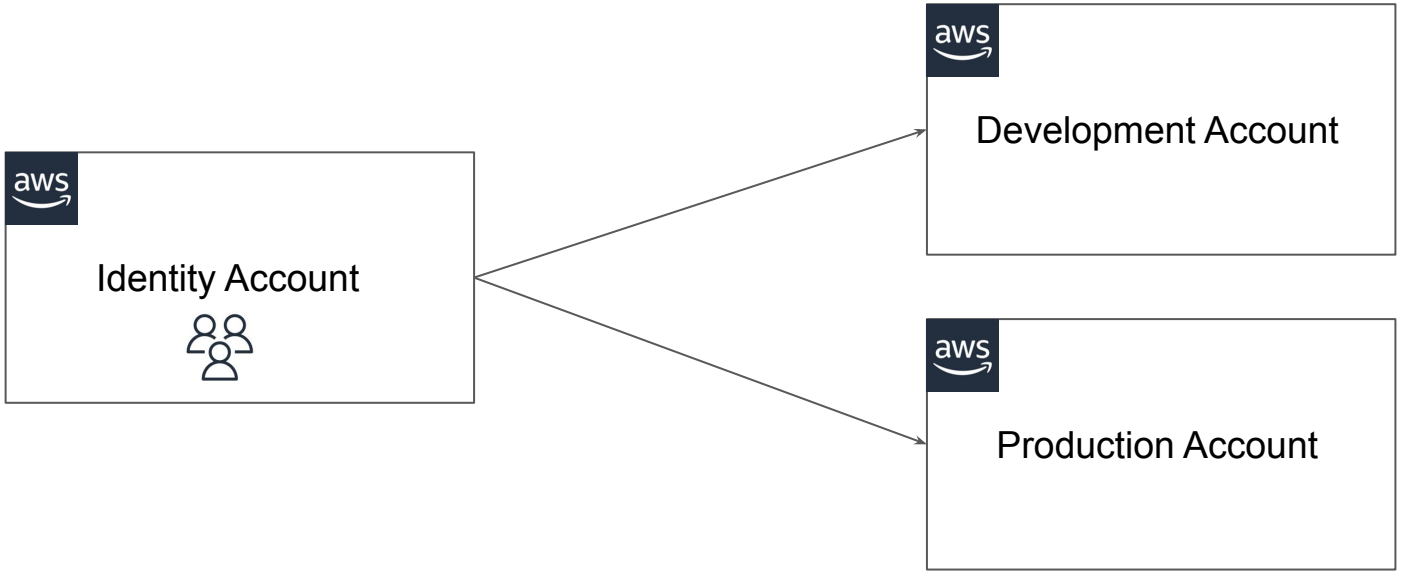
Difficult to Work with. Required a lot of Bookmarks



Rise of Identity Account

In Identity Account architecture, all the IAM Users are stored in central AWS Account.

They could easily connect to Dev/Prod accounts without separate credentials.



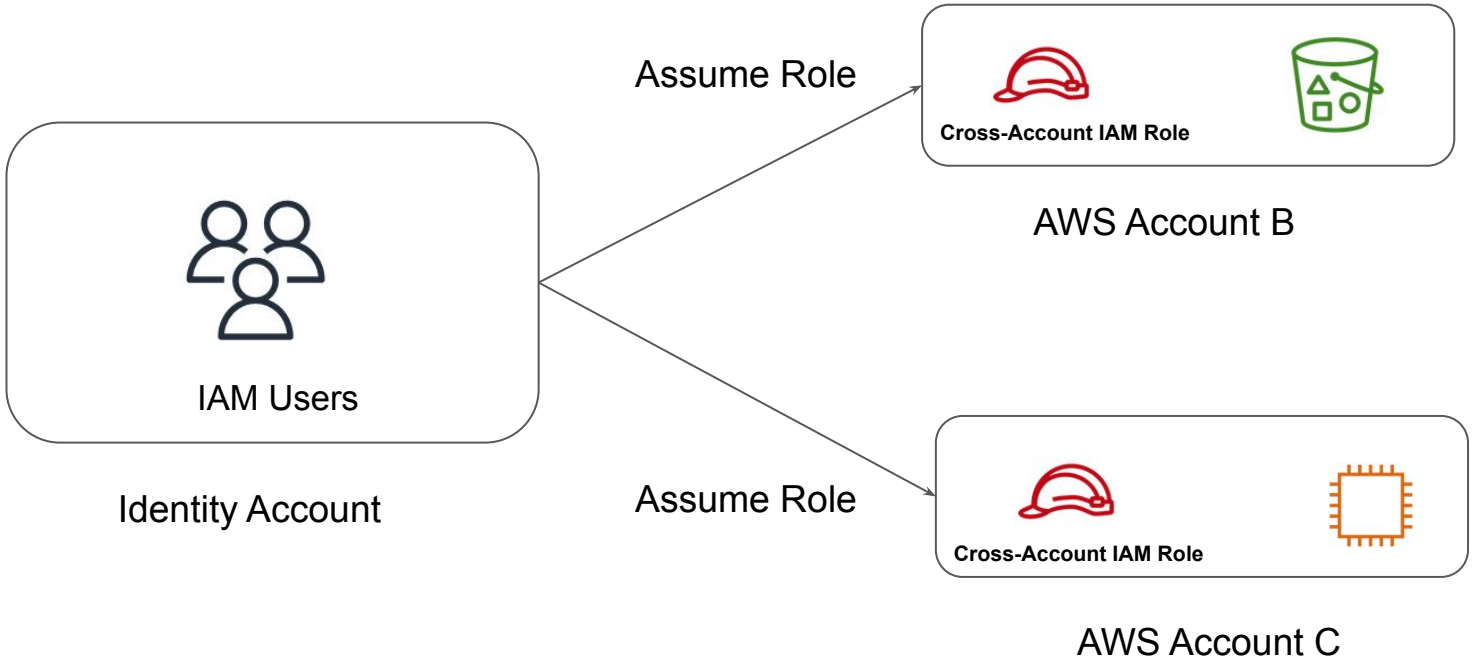
The Architecture

i) Create a user in Account A.

ii) Create a Cross-Account role in Account B.

iii) Allow User to switch to Account-B Role.

The Practical Architecture

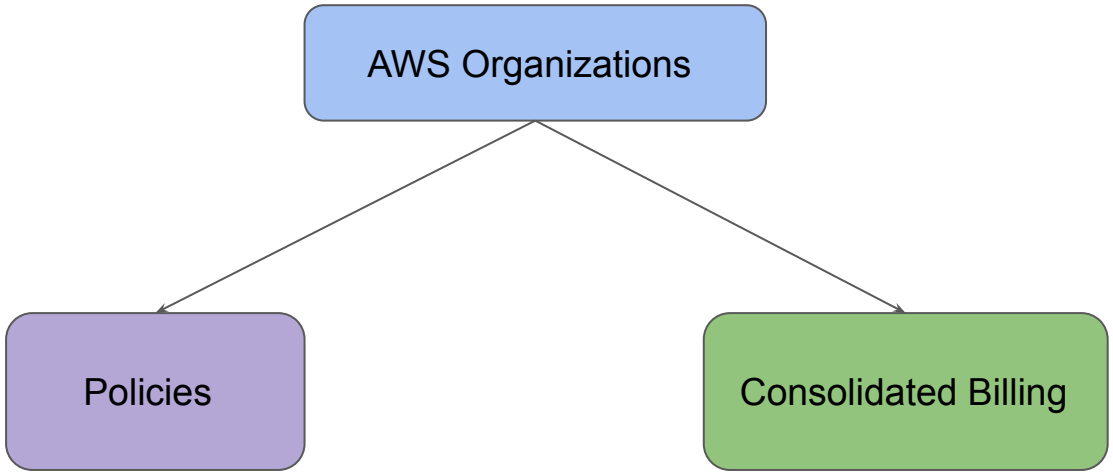


AWS Organizations

Centralized Control

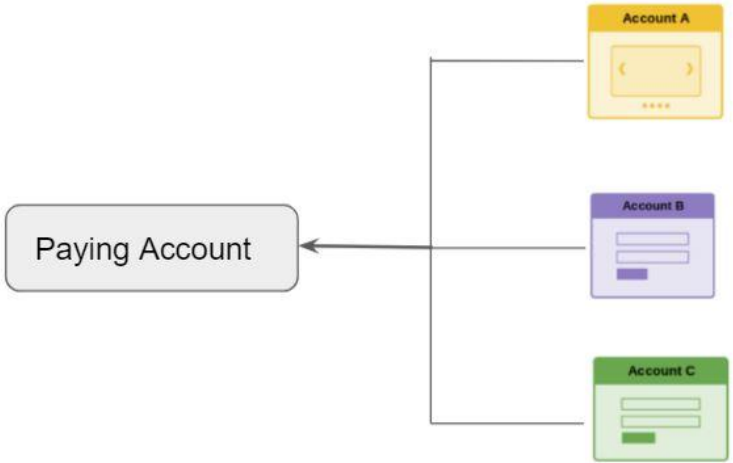
Getting the basics right

AWS offers centralized policy-based management as well as the feature of consolidated billing for multiple AWS accounts through the feature of AWS Organizations.



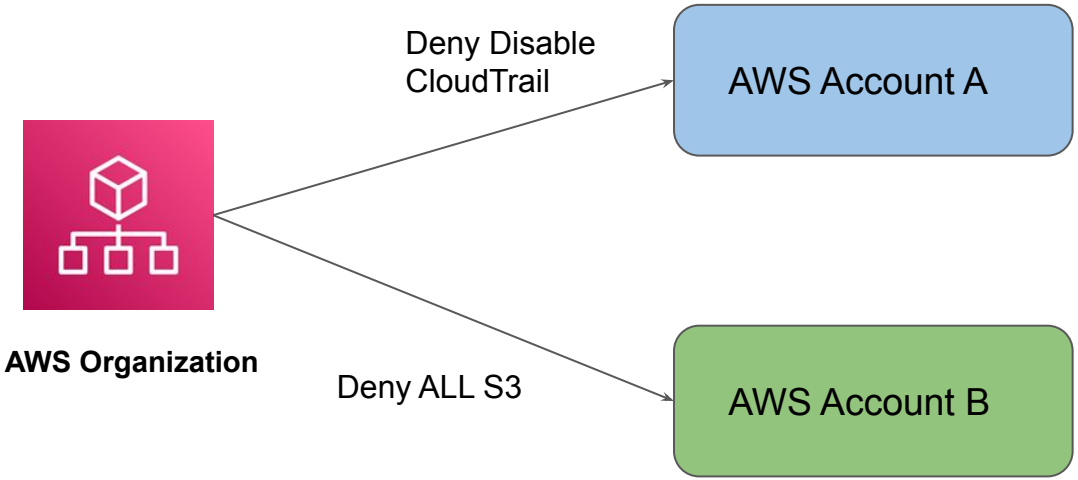
Part 1 - Consolidated Billing

In consolidated billing, management account to access the billing information and pay for all member accounts.



Part 2 - Policies

Policies in AWS Organizations enable you to apply additional types of management to the AWS accounts in your organizations.



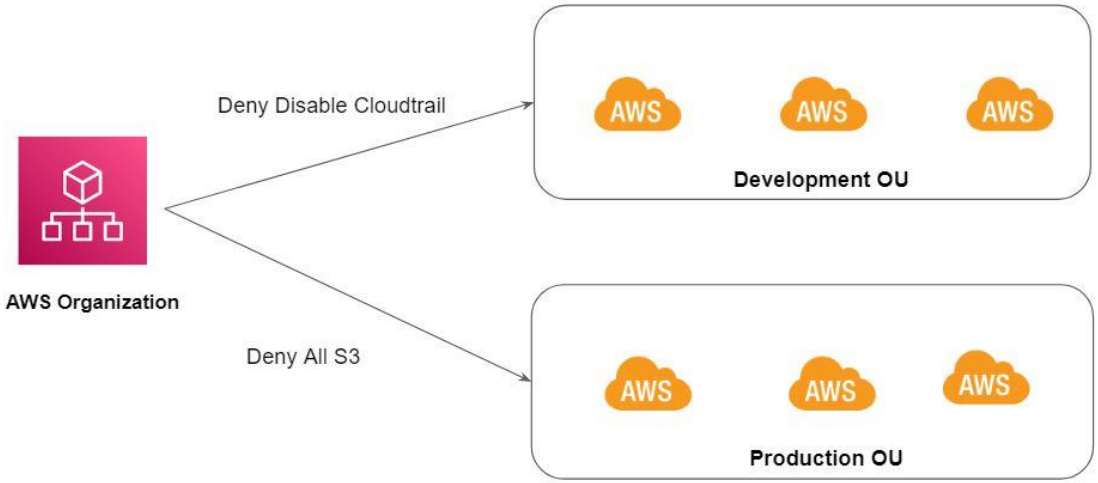
OU in AWS Organization

Were Complex becomes Easy

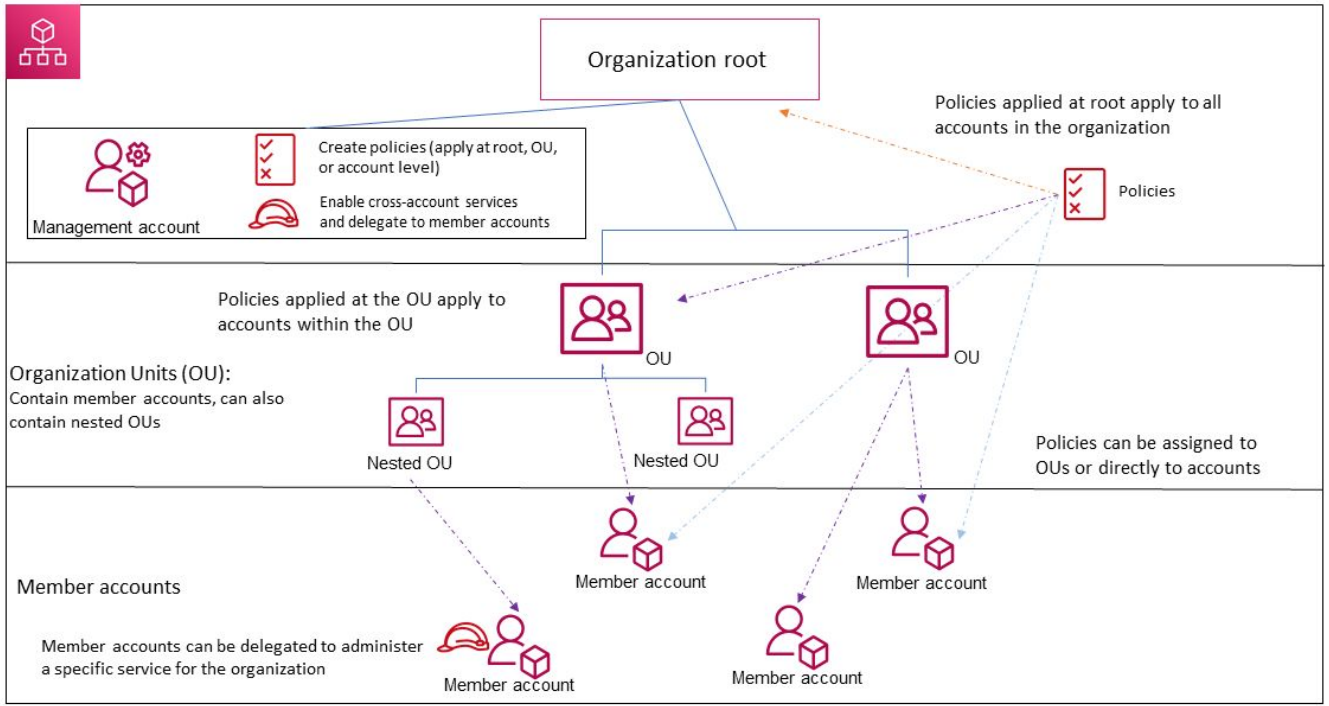
Getting the basics right

Organizational units (OUs) to group accounts together to administer as a single unit

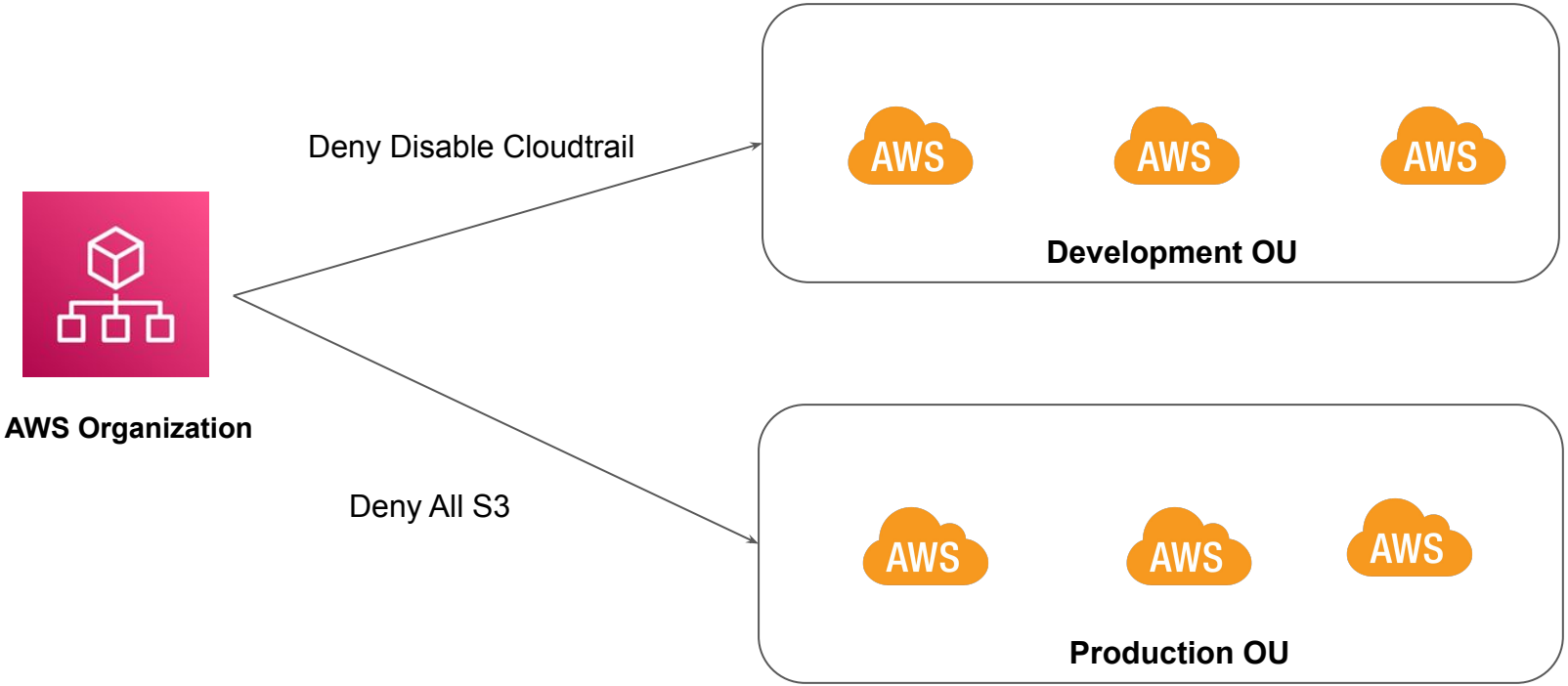
This greatly simplifies the management of your accounts. For example, you can attach a policy-based control to an OU, and all accounts within the OU automatically inherit the policy.



Important Concepts



Grouping AWS Accounts



Important Pointers

SCPs don't affect users or roles in the management account. They affect only the member accounts in your organization.

By default, AWS Organizations attaches an AWS managed policy called FullAWSAccess to all roots, OUs, and accounts. This helps ensure that, as you build your organization, nothing is blocked until you want it to be.

IAM Permissions Boundaries

Advanced IAM

Getting Started

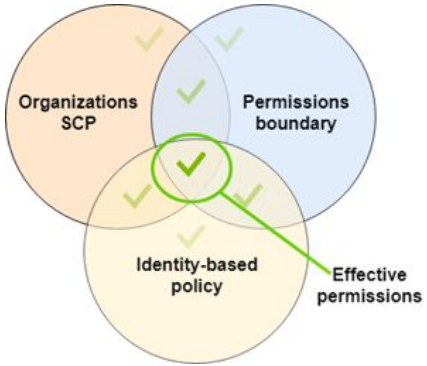
A permissions boundary is an advanced feature in which you use a managed policy to set the maximum permissions that an identity-based policy can grant to an IAM entity.

When you set a permissions boundary for an entity, the entity can perform only the actions that are allowed by both its identity-based policies and its permissions boundaries.

Evaluating Effective Permission with Boundaries

The effective permissions for an entity are the permissions that are granted by all the policies associated with the user/role/account.

Within an AWS account, the permissions for an entity can be affected by identity-based policies, resource-based policies, permissions boundaries, Organizations SCPs, or session policies.

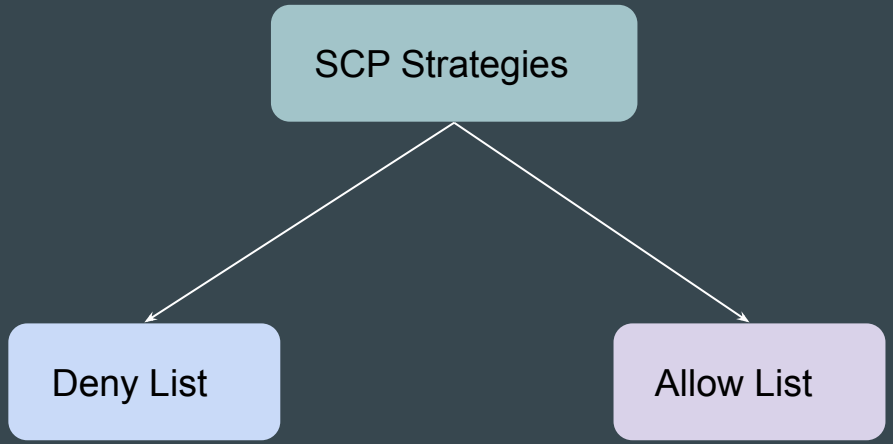


Strategies for using SCPs



Understanding the Basics

There are two strategies that you can use to configure SCPs in your account.



Strategy 1 - Deny List

In deny list, actions are allowed by default, and you specify what services and actions are prohibited

To support this, AWS Organizations attaches an AWS managed SCP named FullAWSAccess to every root and OU when it's created.

```
{
  "Version": "2012-10-17",
  "Statement": [
    {
      "Effect": "Allow",
      "Action": "*",
      "Resource": "*"
    }
  ]
}
```

Benefits of Deny List Strategy

Using a deny list strategy, account administrators can delegate all services and actions until you create and attach an SCP that denies a specific service or set of actions.

Deny statements require less maintenance, because you don't need to update them when AWS adds new services.

Deny statements usually use less space, thus making it easier to stay within the maximum size for SCPs

Sample Deny List Based Policy

```
{
    "Version": "2012-10-17",
    "Statement": [
        {
            "Sid": "AllowsAllActions",
            "Effect": "Allow",
            "Action": "*",
            "Resource": "*"
        },
        {
            "Sid": "DenyDynamoDB",
            "Effect": "Deny",
            "Action": "dynamodb:*",
            "Resource": "*"
        }
    ]
}
```

Strategy 2 - Allow List

To use SCPs as an allow list, you must replace the AWS managed FullAWSAccess SCP with an SCP that explicitly permits only those services and actions that you want to allow.

By removing the default FullAWSAccess SCP, all actions for all services are now implicitly denied.

Your custom SCP then overrides the implicit Deny with an explicit Allow for only those actions that you want to permit

Sample Allow List Based Policy



```
{
    "Version": "2012-10-17",
    "Statement": [
        {
            "Effect": "Allow",
            "Action": [
                "ec2:*",
                "cloudwatch:*"
            ],
            "Resource": "*"
        }
    ]
}
```

Points to Note

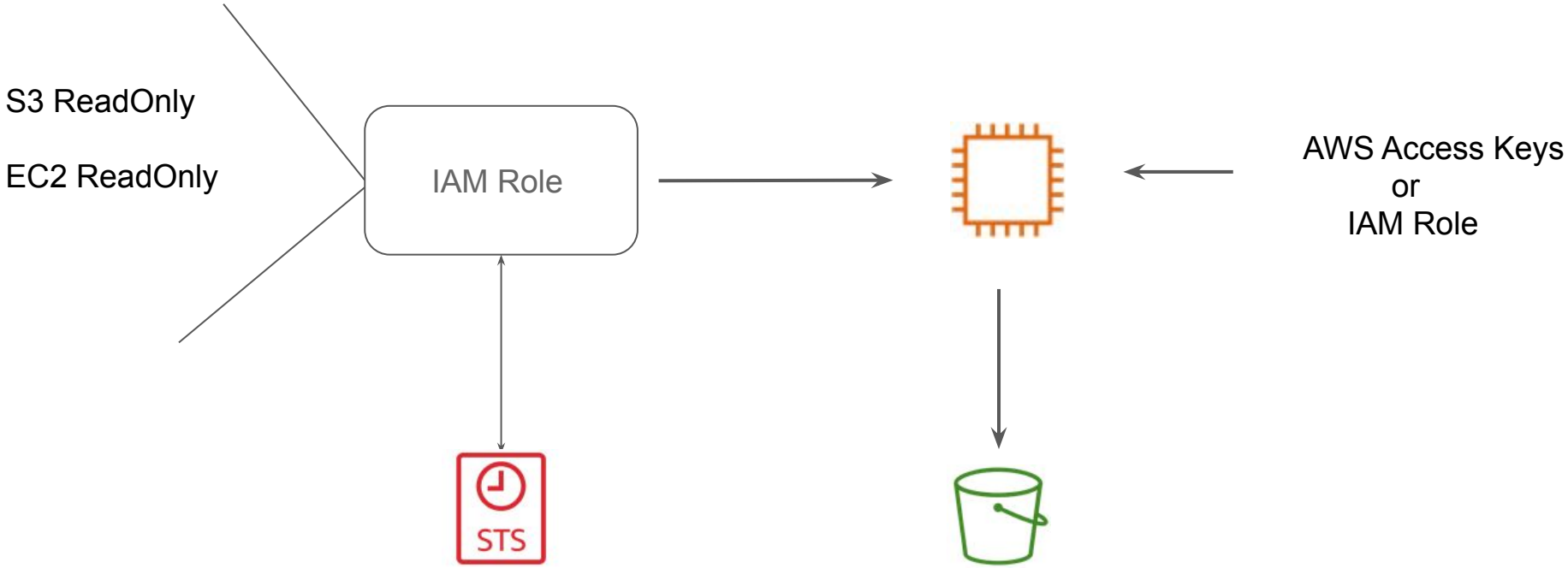
Every root, OU, and account must have at least one SCP attached.

If you want to replace the default FullAWSAccess policy with an SCP that limits the permissions that can be delegated, you must attach the replacement SCP before you can remove the default SCP.

AWS Secure Token Service (STS)

Credentials Management

How IAM Role


S3 ReadOnly
EC2 ReadOnly
IAM Role
STS
AWS Access Keys
or
IAM Role

Overview of STS

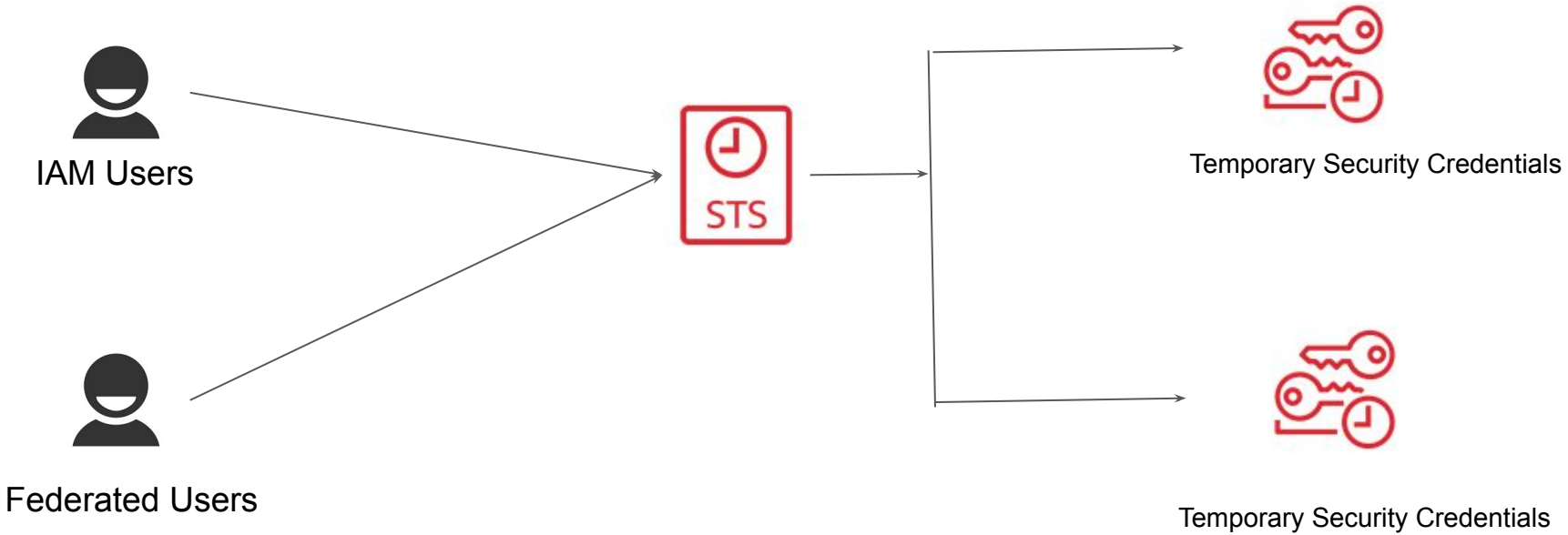
- The AWS STS is a web service that enables you to request temporary, limited-privilege credentials for AWS Identity and Access Management (IAM) users or for users that you authenticate (federated users).

- Temporary security credentials are short term and expire after a certain duration.

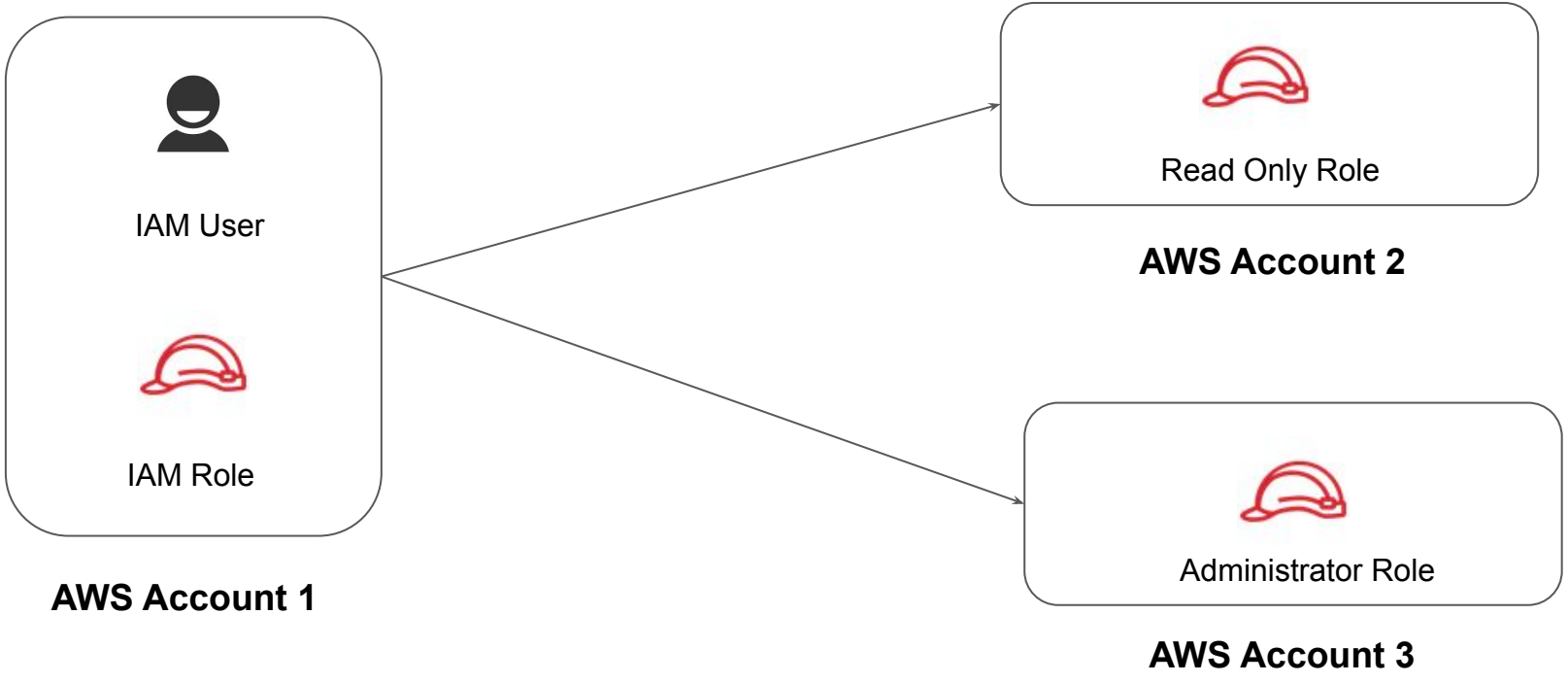
- Since they have a limited lifetime, the key rotation is no longer explicitly needed.


```
{
    "AssumedRoleUser": {
        "AssumedRoleId": "AROAJOTOADWSDZD53Z7VS:temp",
        "Arn": "arn:aws:sts::836802967410:assumed-role/CA-EC2RO/temp"
    },
    "Credentials": {
        "SecretAccessKey": "LKtyaWrhxGnBNP3tx7dMK2nv0H1VdwMP1RVP5Sob",
        "SessionToken": "FQoDYXdzEMj//////////wEaDHwScBw1Hmr5eGqKXyLHAdeXEJZ0oSuJxFd/PGtU
Z5F3XhjgIawg7ytJXXWRgpyvaq9eMKNfUqmiDca/NM+FLwqy5iek5VKPGkPut+/pAzOWOH3ddVmcuHsJowHxaDGHa
d6S2IyhyMfAF9bk9FQjMfHNtl/oDl74KvkAV6xAE4Q0cPZ4sDGes130Im4r5TulKT/I2qvg0w/LVRjraJ8UBnopMu
gU=",
        "Expiration": "2017-09-20T23:36:41Z",
        "AccessKeyId": "ASIAJWPD367QI4GHCNAQ"
    }
```

Overview Architecture of STS



STS Architecture - Cross-AWS Account Access

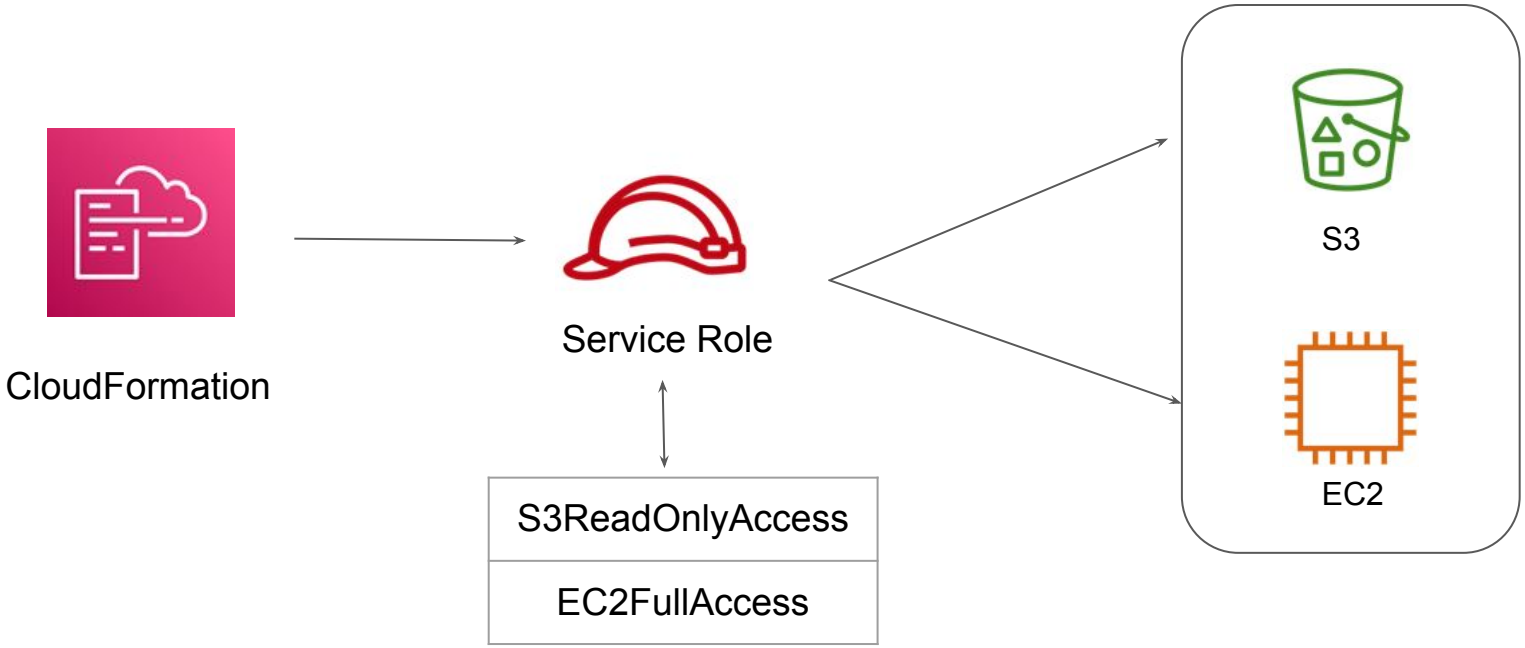


Service Role & Pass Role

Back to IAM

Overview of Service Roles

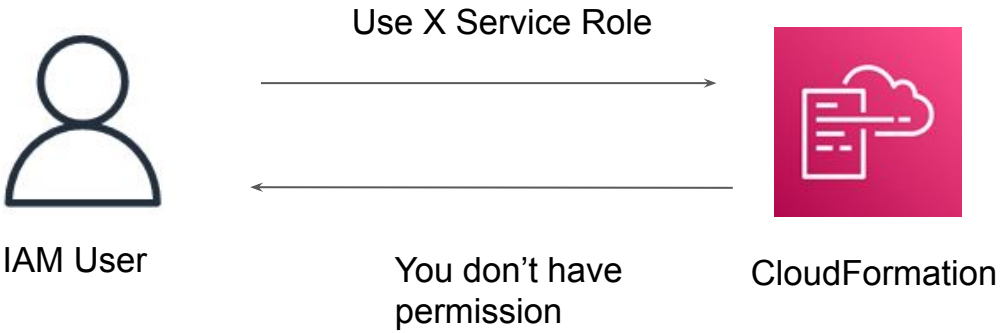
A service role is a role that an AWS service assumes to perform actions on your behalf.



Overview of PassRole

Pass Role allows the service to assume the role and perform actions on your behalf.

To pass a role (and its permissions) to an AWS service, a user must have permissions to pass the role to the service.

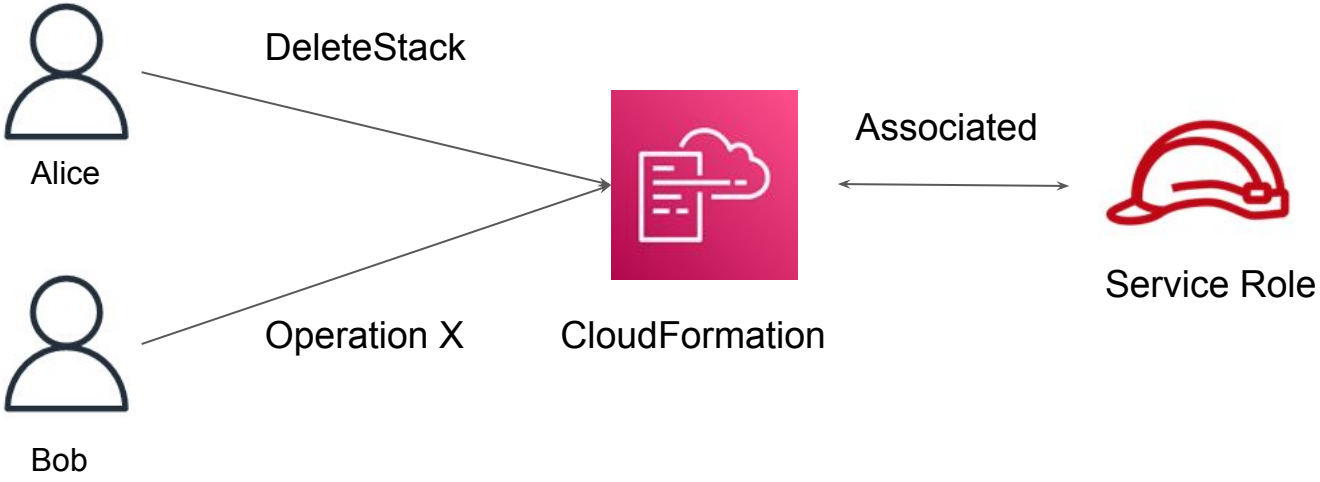


Sample PassRole Policy

```
{
    "Version": "2012-10-17",
    "Statement": [{
        "Effect": "Allow",
        "Action": [
            "iam:GetRole",
            "iam:PassRole"
        ],
        "Resource": "arn:aws:iam::<account-id>:role/EC2-roles-for-XYZ-*"
    }]
}
```

Important Pointer

Once the Role is associated with CloudFormation, other users that have permissions to operate on this stack will be able to use this role, even if they don't have permission to pass it. Ensure that this role grants least privilege.



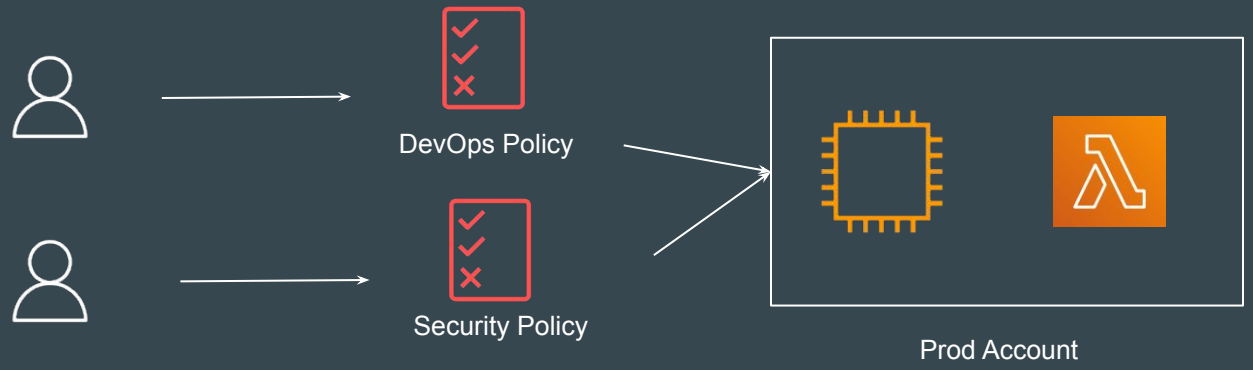
Attribute-based access control



Basics of RBAC

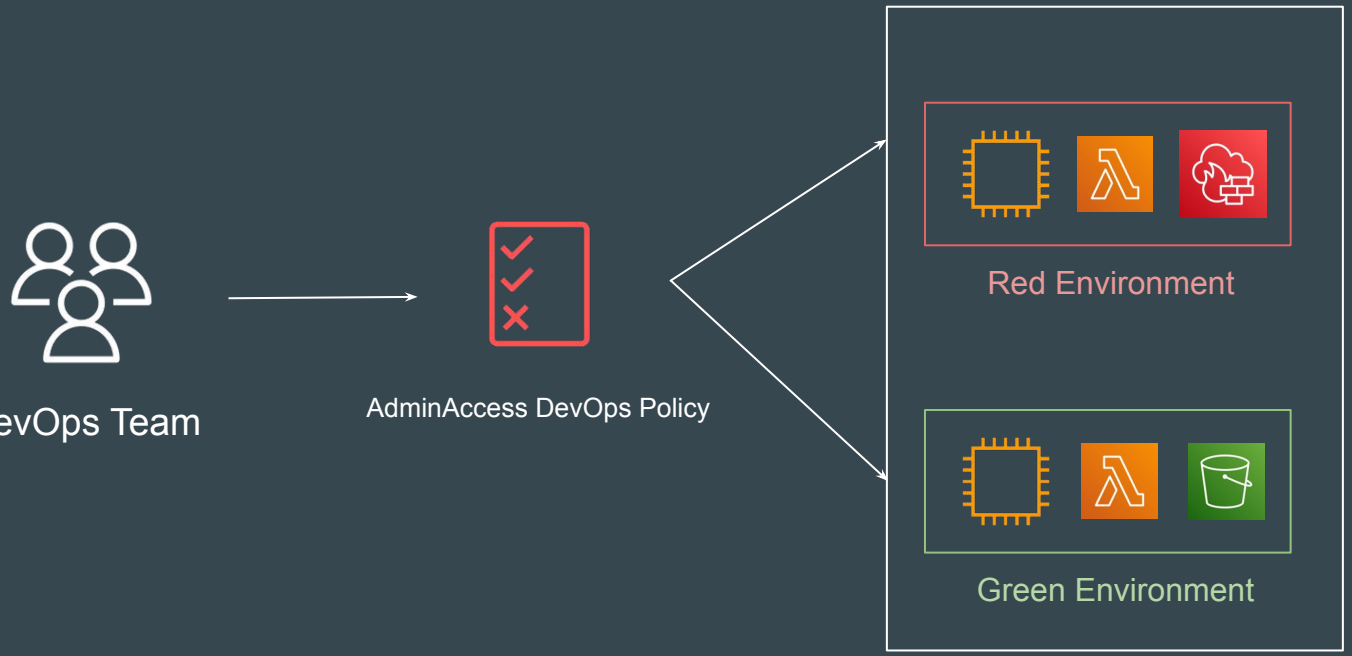
Role-based access control (RBAC) restricts access based on a person's role within an organization

In IAM, you implement RBAC by creating different policies for different job functions



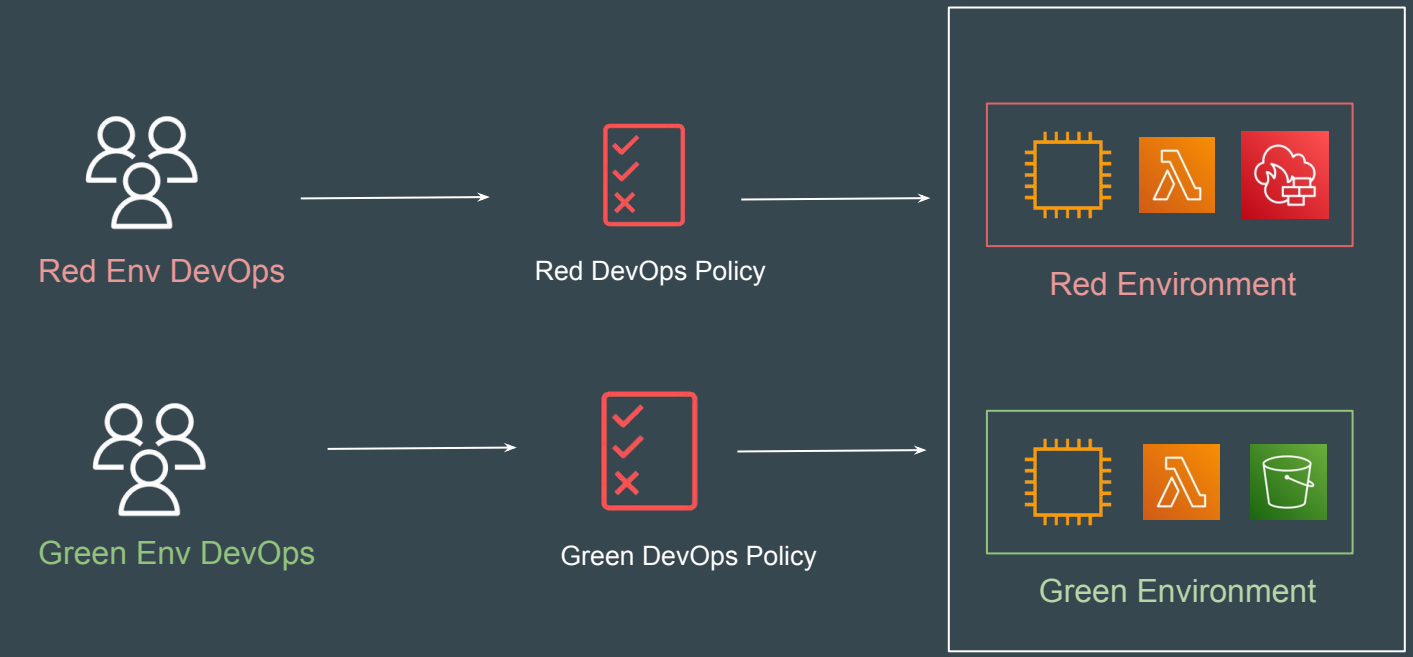
Understanding the Challenge

DevOps Team has access to Red and Green Environment



Possible Approach of Separation

Red Env DevOps only has access to Red Environment
Green Env DevOps only has access to Green Environment



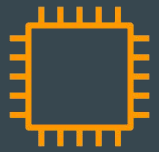
Basics of Attributes

Attributes are key-value pairs.

In AWS, these attributes are called tags.

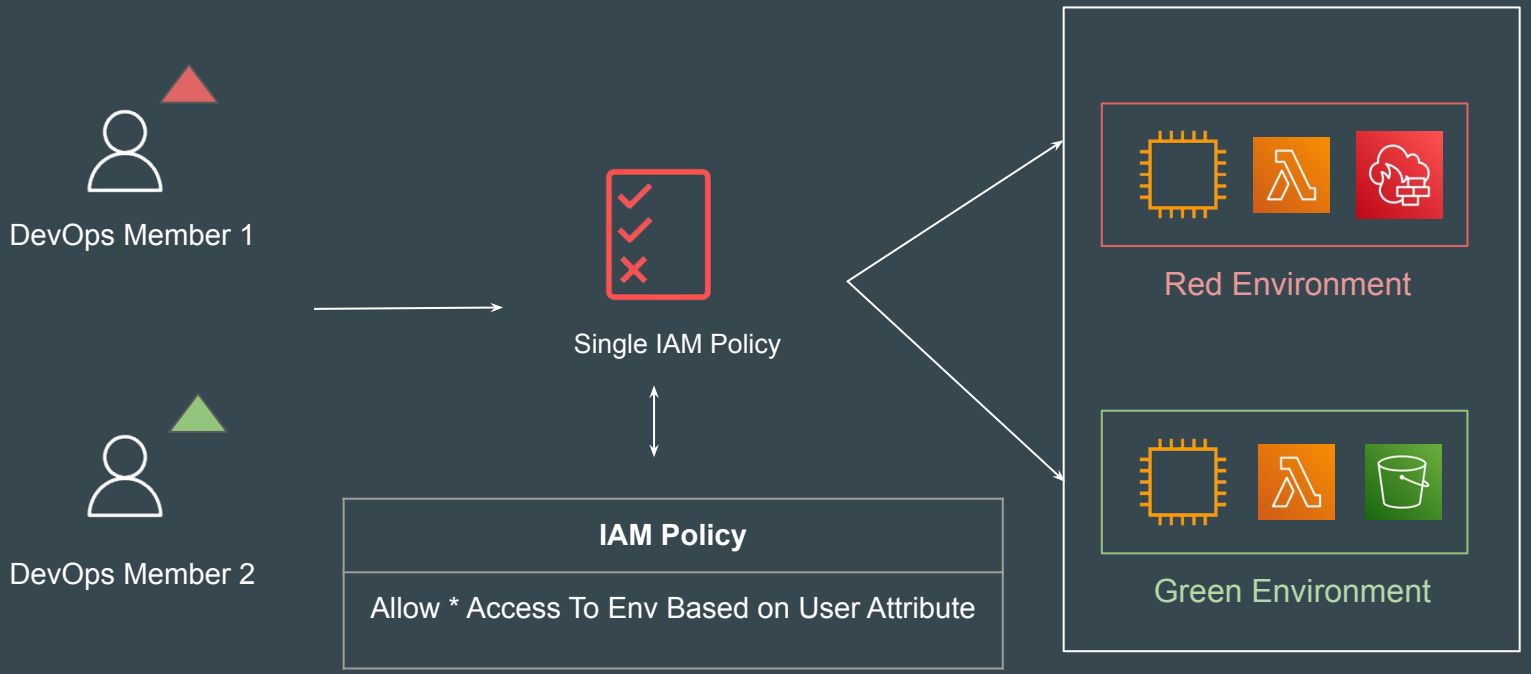


Team: DevOps
Location: India



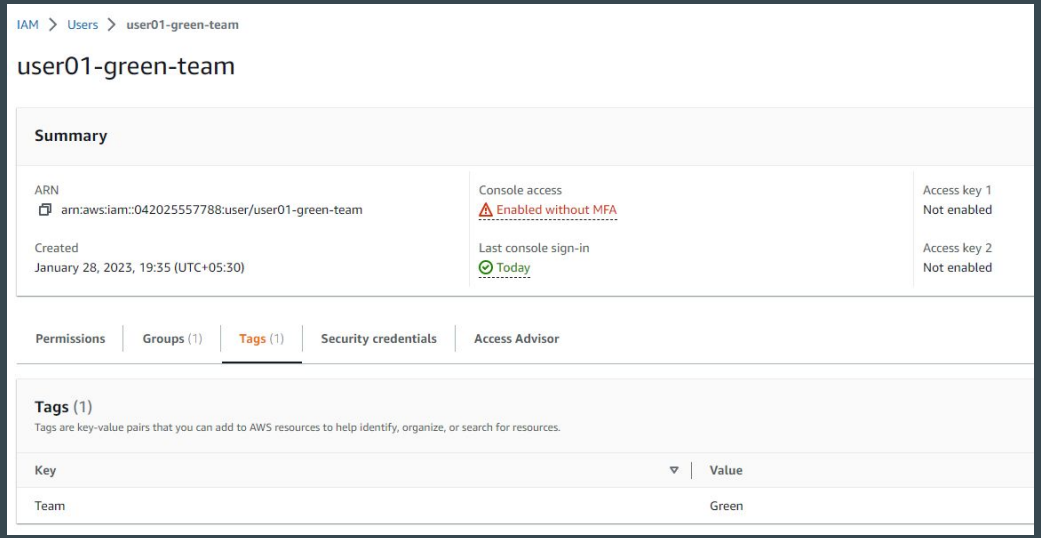
Env: Prod
Project: Green

Scalable Permission Model based on Attributes



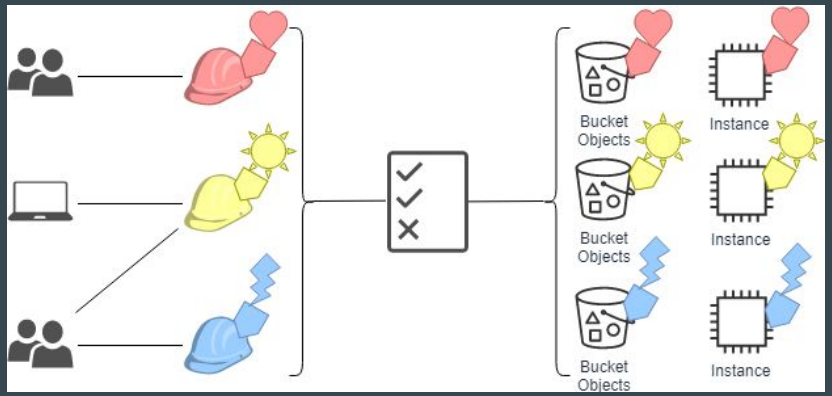
Attributes for IAM User

You can use IAM tag key-value pairs to add custom attributes to an IAM user.



Attribute-Based Access Control

Attribute-based access control (ABAC) is an authorization strategy that defines permissions based on attributes.



Permissions Based on ABAC

This example shows an IAM policy that allows a principal to start or stop an Amazon EC2 instance when the instance's resource tag and the principal's tag have the same value for the tag key Team



Key	Value
Team	Green

```
{
    "Version": "2012-10-17",
    "Statement": [
        {
            "Effect": "Allow",
            "Action": [
                "ec2:DescribeInstances"
            ],
            "Resource": "*"
        },
        {
            "Effect": "Allow",
            "Action": [
                "ec2:StartInstances",
                "ec2:StopInstances"
            ],
            "Resource": "*",
            "Condition": {
                "StringEquals": {
                    "ec2:ResourceTag/Team": "${aws:PrincipalTag/Team}"
                }
            }
        }
    ]
}
```

Benefits of ABAC

ABAC requires fewer policies. Because you don't have to create different policies for different job functions, you create fewer policies. Those policies are easier to manage.

Permissions can easily be granted and revoked based on user’s tags.

You can even use attributes of users from corporate directory to allow / deny permissions to AWS resources.

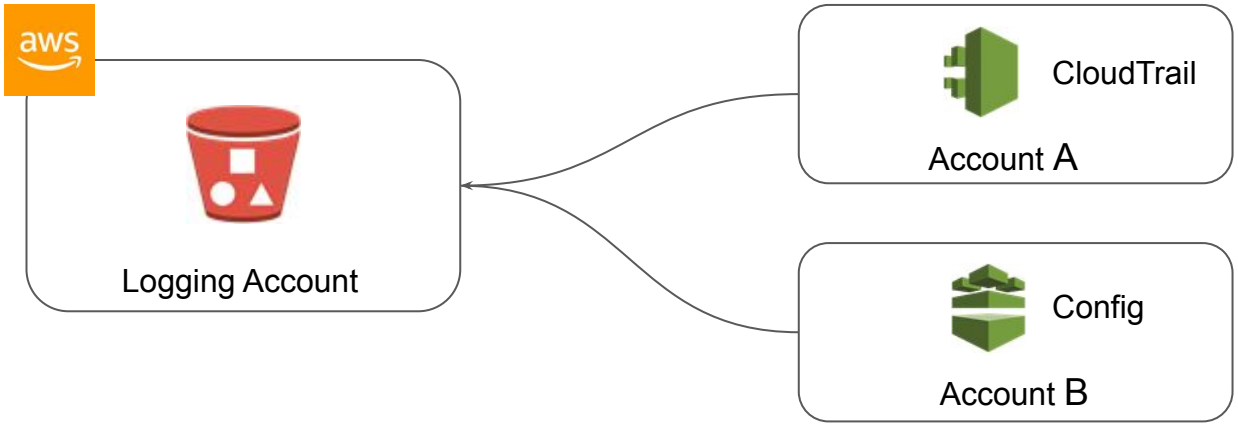
Centralized Logging

Architectural Perspective

Centralized Logging

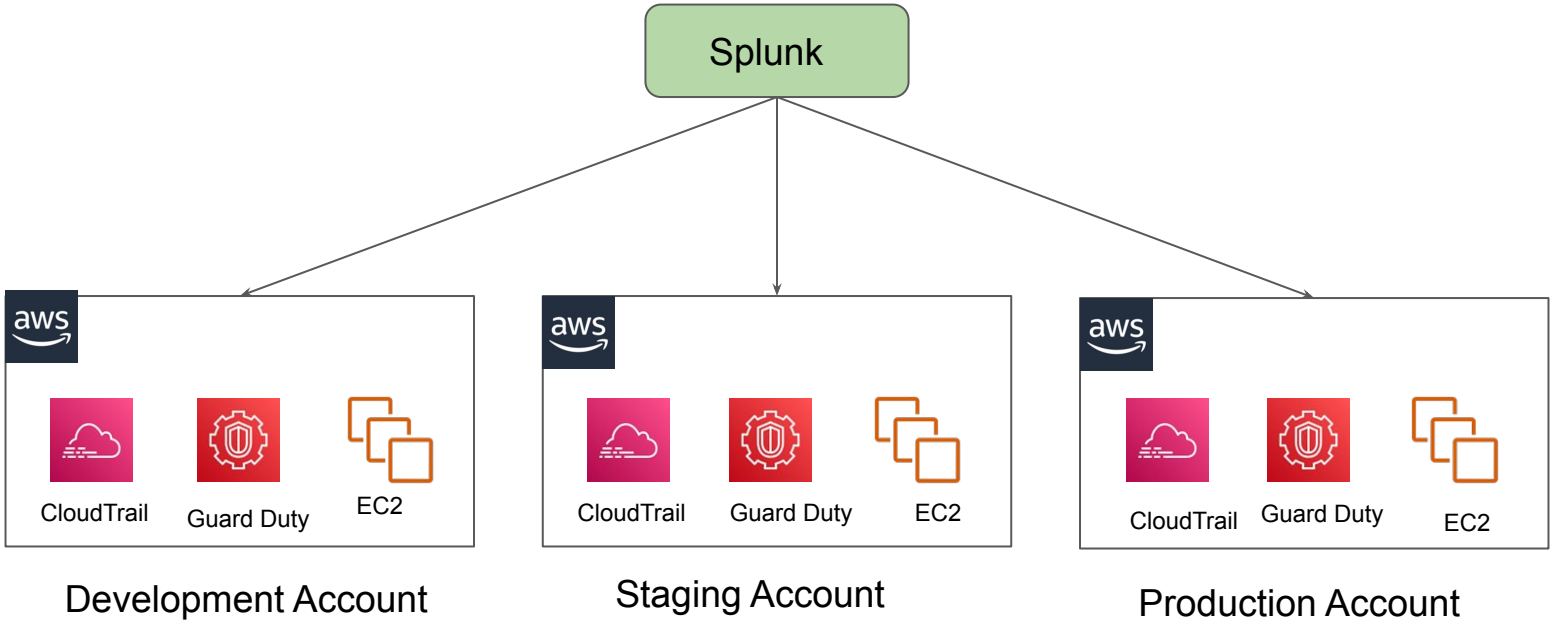
A comprehensive log management and analysis strategy is mission critical in an organization.

It enables the organizations to understand the relationship between operational, security, and change management events and maintain a comprehensive understanding of their infrastructure.



Challenges with Logging

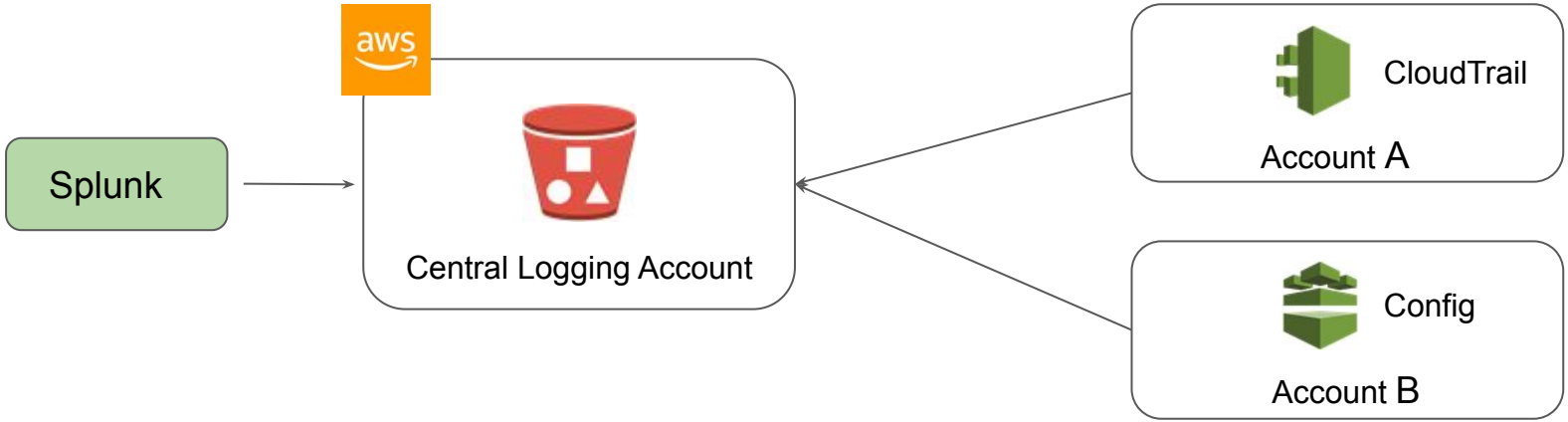
In a Multi-Account based architecture, log monitoring at an individual account level is not the best of the approaches.



Recommended Architecture for Logging

A comprehensive log management and analysis strategy is mission critical in an organization.

One of the recommended approaches is to use a Centralized Logging Account.



Considerations while implementing Logging

Define log retention requirements and lifecycle policies early on.

Incorporate tools and features to automate the lifecycle policies.

Automate the installation and configuration of log shipping agent.

Make sure the solution supports hybrid environment to support the needs.

AWS Services to Help!

We can make use of AWS Managed service to build centralized logging solutions.

Services which can help here:

- AWS ElasticSearch Service
- AWS CloudWatch Logs
- Kinesis Firehose
- AWS S3

Ways to configure centralized logging for each AWS service (CloudTrail, VPCFlow) differs.

Considerations - S3 Bucket Policy for Cross-Account

Architectural Perspective

Challenges with S3 Bucket Policy

A wildcard based S3 bucket policy allowing CloudTrail service would mean that any AWS account's CloudTrail can put its data to your S3 bucket.

```
{
  "Sid": "AWSCloudTrailWrite20131101",
  "Effect": "Allow",
  "Principal": {
    "Service": "cloudtrail.amazonaws.com"
  },
  "Action": "s3:PutObject",
  "Resource": "arn:aws:s3:::kplabs-central-log-cloudtrail/*",
  "Condition": {
    "StringEquals": {
      "s3:x-amz-acl": "bucket-owner-full-control"
    }
  }
}
```

Bucket Policy with Conditional Statement

As a security best practice, add an aws:SourceArn condition key to the Amazon S3 bucket policy. This helps prevent unauthorized access to your S3 bucket.

```
{
  "Sid": "AWSCloudTrailWrite20131101",
  "Effect": "Allow",
  "Principal": {
    "Service": "cloudtrail.amazonaws.com"
  },
  "Action": "s3:PutObject",
  "Resource": "arn:aws:s3:::kplabs-central-log-cloudtrail/*",
  "Condition": {
    "StringEquals": {
      "aws:SourceArn": "arn:aws:cloudtrail:ap-southeast-1:693331494763:trail/demo-trail",
      "s3:x-amz-acl": "bucket-owner-full-control"
    }
  }
}
```

Unified CloudWatch Agent

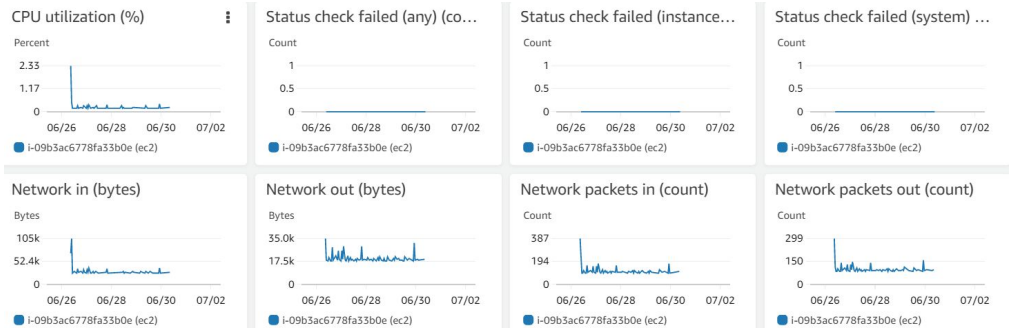
Metrics and Logs

Default CloudWatch Metrics

When we launch an EC2 instance in AWS, there are certain metrics that are captured by default.

Some of these include:

- CPU Utilization
- Network Related
- Disk Related

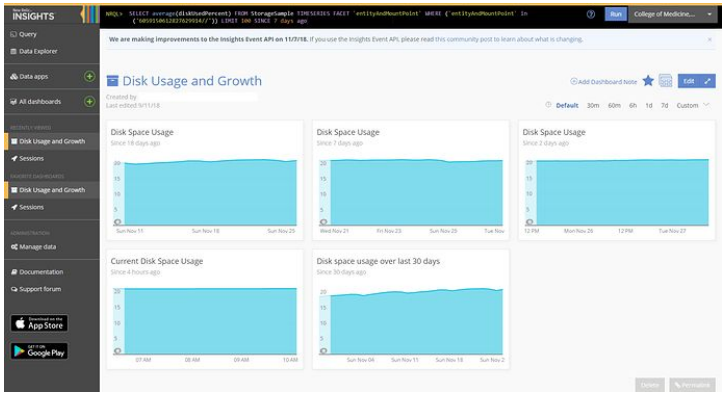


Challenge 1 -More Metrics Are Needed

There are various important metrics that needs to be collected in addition to the default ones.

Some of these include:

- Memory Metrics
- Disk Usage Metrics
- Netstat related.

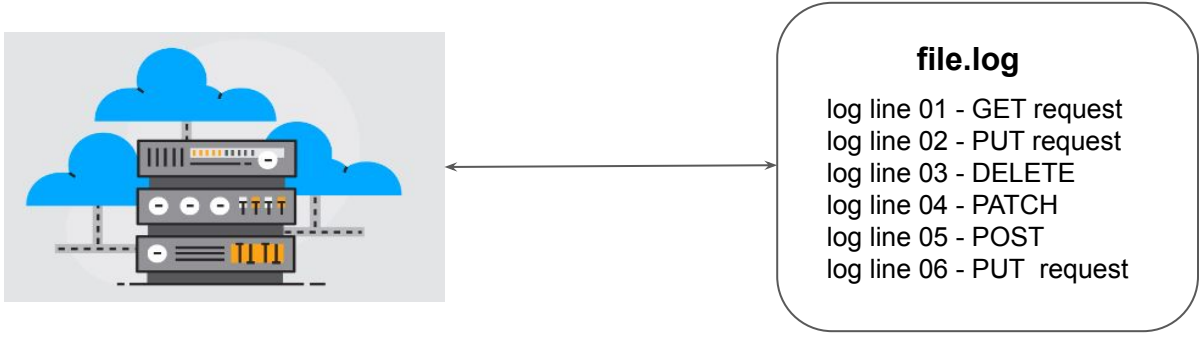


Challenge 2 - Log Monitoring

A server can contain a lot of log files, from system logs to the application logs.

During debugging, it is important to have log files at hand.

This means in default case; you need to give access to the server to an individual who wants to debug.

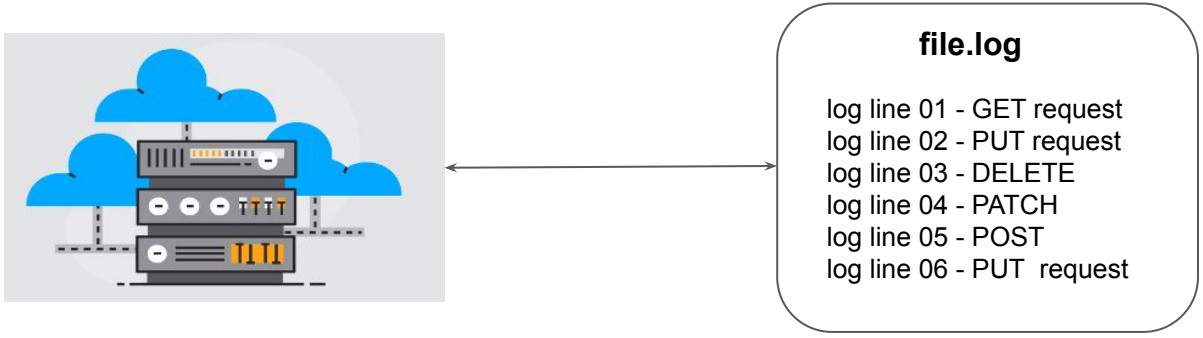


Disadvantage of the Approach

Access must be given to the server to the developers.

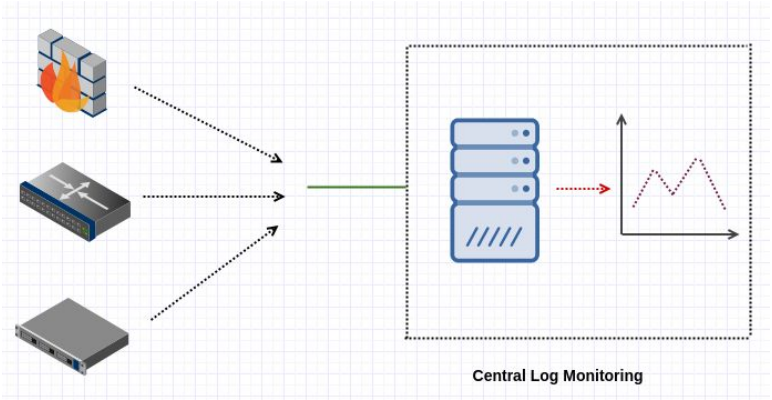
If the server gets terminated, the logs are lost.

No way to set up an alarm on certain conditions or create complex filters.



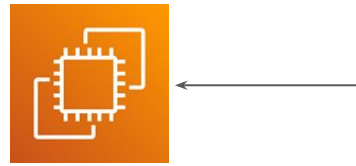
Better Way

- We create a Central Log Server.
- We push the log files from individual systems to Central Log Server.

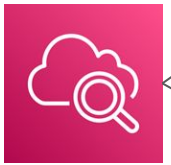


Introducing Unified CloudWatch Agent

Unified CloudWatch Agent allows customers to capture both the internal system level metrics as well as logs collection.



Amazon EC2

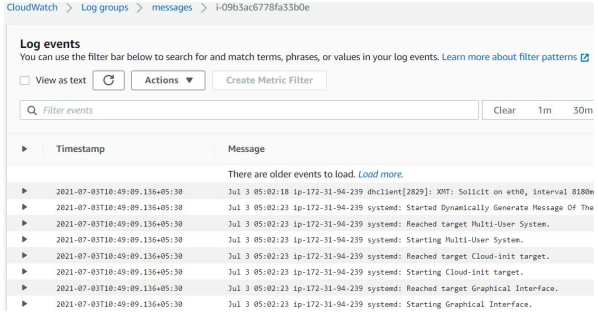


CloudWatch Unified Agent

Metrics



Logs



How-To Steps

1. Create a IAM Role with CloudWatchAgentServer policy.

2. Create EC2 using IAM Role.

3. Install CloudWatch Agent.

4. Run CloudWatch Agent Configuration Wizard

5. Start Unified CloudWatch Agent.

AWS License Manager

Let’s understand Licensing

Getting Started

In Enterprises, managing software licenses sometimes becomes quite a hassle.

Organizations uses wide variety of software licenses:

- OS Level Licenses : Windows, RedHat
- Database Licenses : Oracle DB, Microsoft SQL
- Application Licenses: SAP
- Other 3rd Party Licenses

Challenges

In Cloud, new servers can be launched in click of a button.

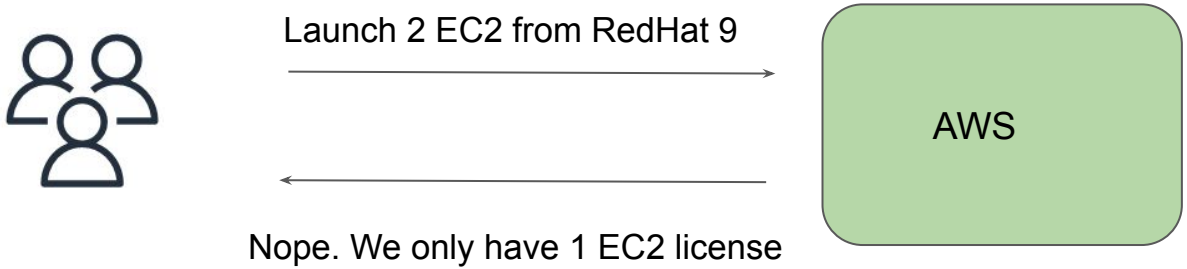
License Violation detected during audit can lead to heavy penalties.

Difficult to track licenses across multiple accounts.

Overview of AWS License Manager

AWS License Manager is a service which allows us to manage license from wide variety of software vendors across AWS and on-premise.

We can enforce policies for licenses based on various factors like CPU, sockets that will control the number of EC2 instance which can be launched.

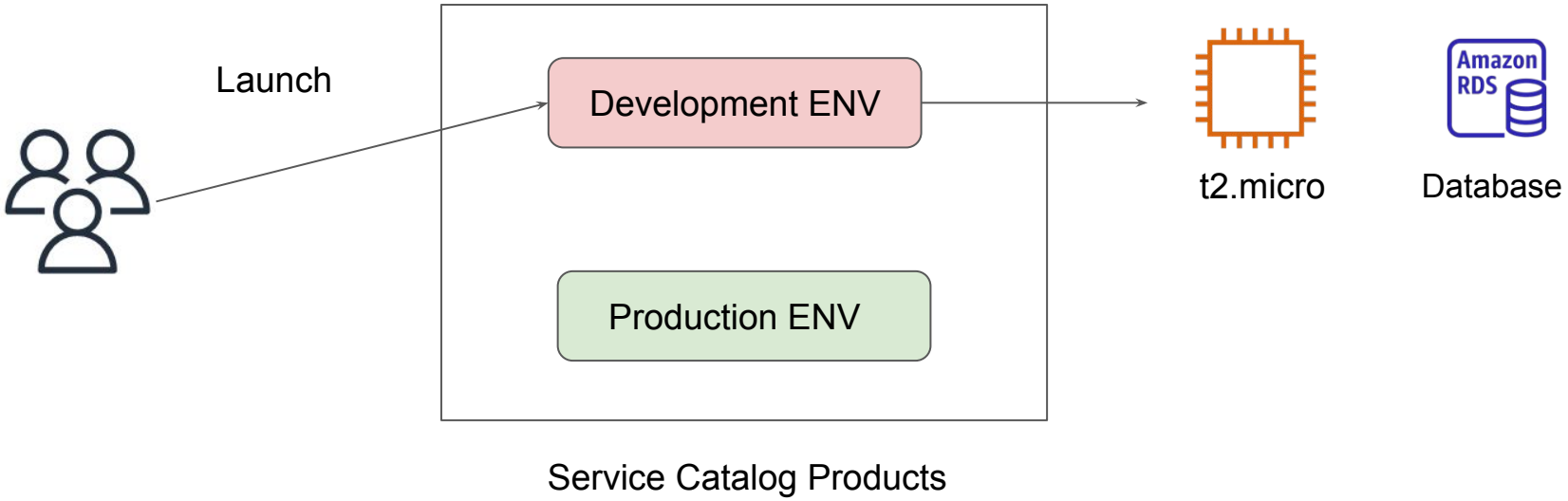


AWS Service Catalog

Standardized Stack

Understanding the Workflow

AWS Service Catalog enables organizations to create and manage catalogs of IT services that are approved for use on AWS.

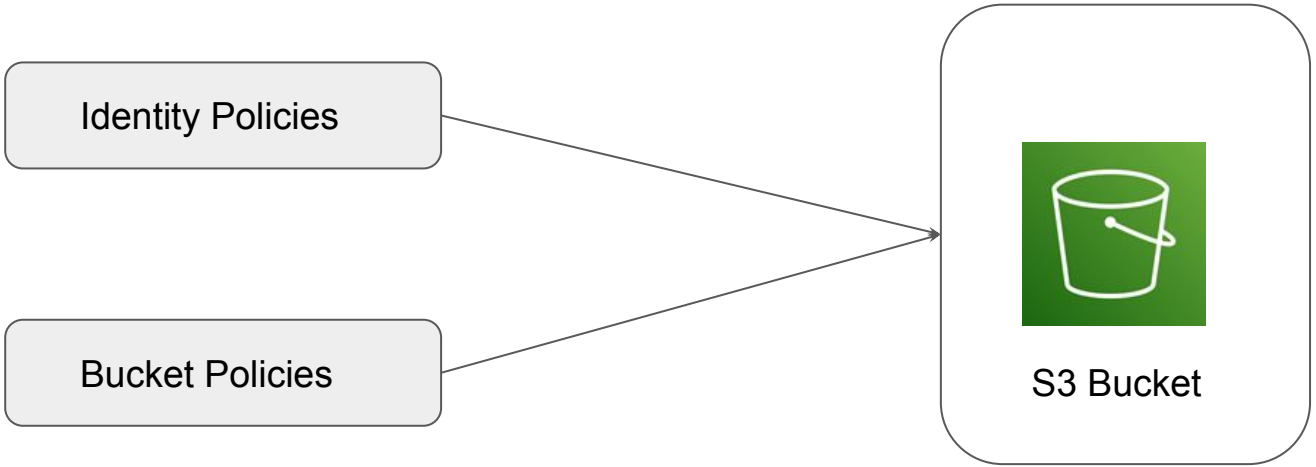


S3 Bucket Policy

Bucket Policies

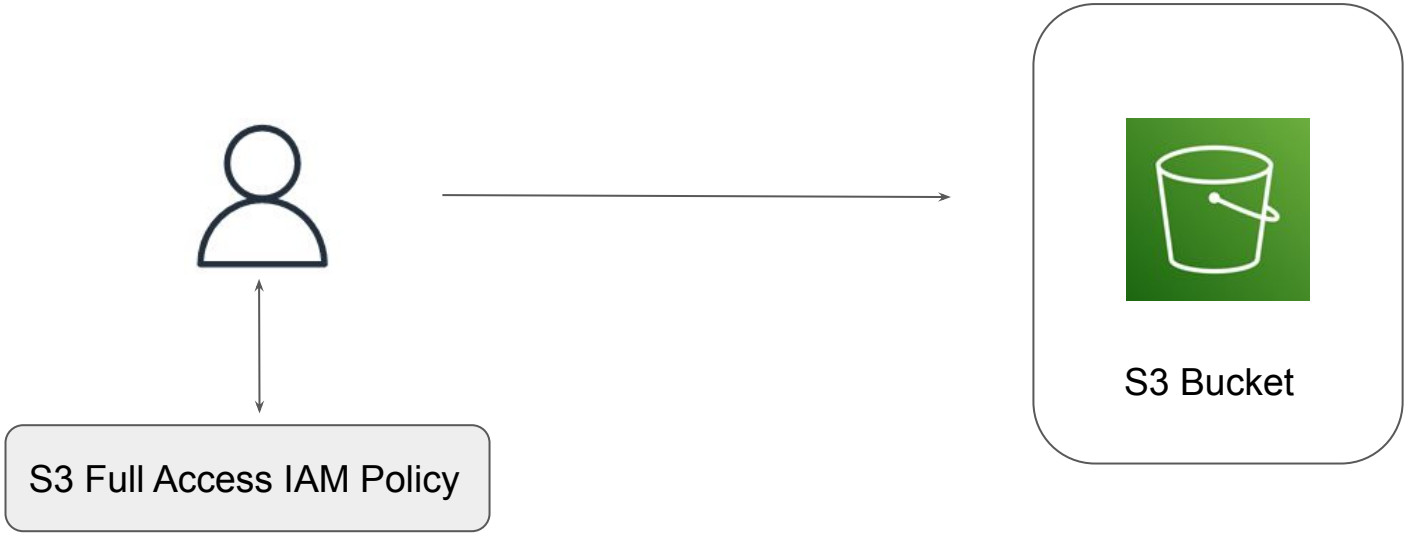
Granting Permission for S3 Resource

There are two primary ways in which a permission to a S3 resource is granted.



Use-Case 1: IAM User Needs Access to S3 Bucket

IAM User Named Bob needs Full Access to S3 Bucket.

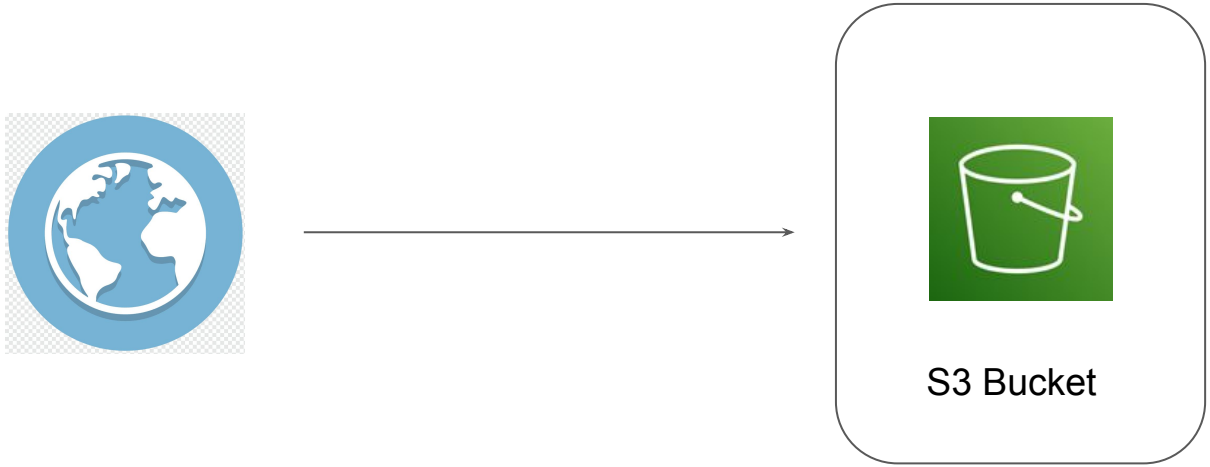


Wider Scope of S3 Bucket

Files within the S3 bucket can have scope beyond the IAM entity.

Organization can host entire websites in S3 Bucket.

S3 Buckets can even be used to host central files for download.



S3 Bucket Policy

A bucket policy is a resource-based AWS IAM policy associated with the S3 Bucket to control access permissions for the bucket and the objects in it .



S3 Bucket Policy Rules
Allow Access from Internet
Allow Access from only 10.77.3.20 IP
Only Allow Access from Specific VPC
Only allow HTTPS connections

Bucket Policy 1 - Public Access

The following example policy grants the s3:GetObject permission to any public anonymous users.

```
{
  "Version":"2012-10-17",
  "Statement":[
    {
      "Sid":"PublicRead",
      "Effect":"Allow",
      "Principal": "*",
      "Action":["s3:GetObject"],
      "Resource":["arn:aws:s3:::demo-bucket/*"]
    }
  ]
}
```

Bucket Policy 2 - Only HTTPS

Only the HTTPS requests should be allowed. All HTTP requests should be blocked.

```
{
  "Id": "ExamplePolicy",
  "Version": "2012-10-17",
  "Statement": [
    {
      "Sid": "AllowSSLRequests",
      "Action": "s3:GetObject",
      "Effect": "Allow",
      "Resource": [
        "arn:aws:s3:::demo-bucket/*"
      ],
      "Condition": {
        "Bool": {
          "aws:SecureTransport": "true"
        }
      },
      "Principal": "*"
    }
```

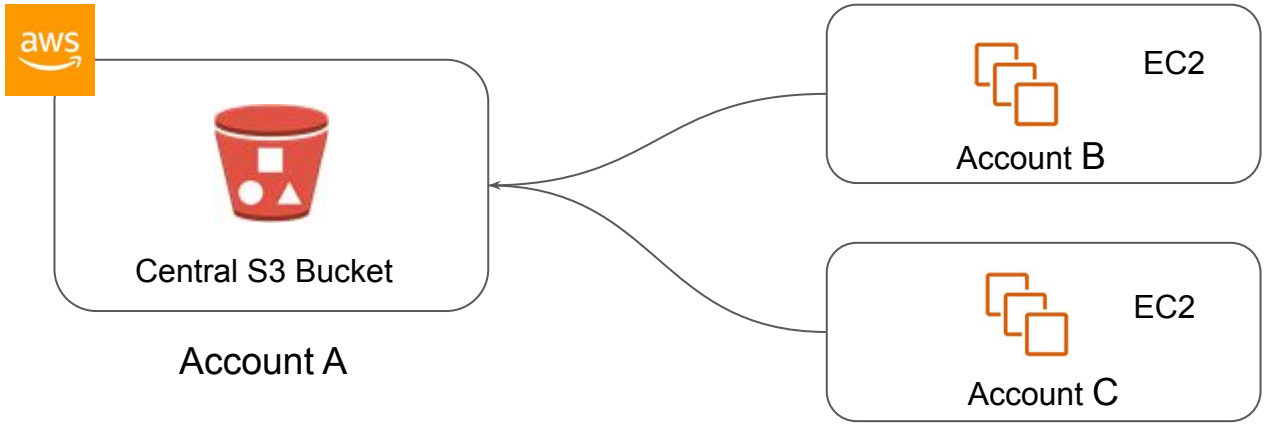
Cross Account S3 Access

Bucket Policies

Cross Account S3 Access

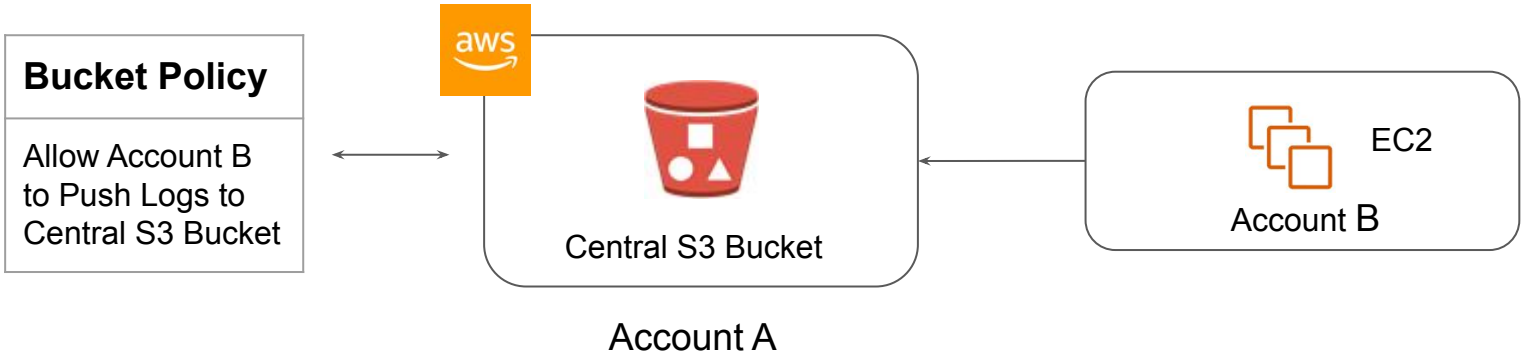
There are many requirements where logs across all AWS accounts need to be stored in a central account.

These logs can include, CloudTrail, CloudWatch, Application Logs, and others.



Creating Bucket Policy

The recommended approach is to add a Bucket Policy in the Central S3 bucket and allow the Account B to push the logs.



Bucket Policy Example - Central S3 Account

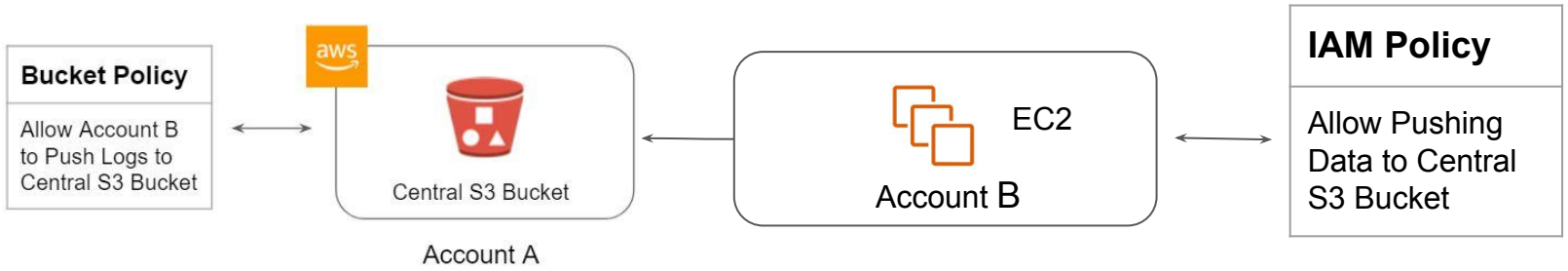
```
{
    "Version": "2012-10-17",
    "Statement": [
        {
            "Effect": "Allow",
            "Principal": {
               "AWS": "arn:aws:iam::453314488441:root"
            },
            "Action": [
                "s3:GetObject",
                "s3:PutObject",
                "s3:PutObjectAcl"
            ],
            "Resource": [
                "arn:aws:s3:::central-s3-bucket/*"
            ]
        }
    ]
}
```


Account B ARN

Central S3 Bucket

Part 2- Permission on Account B Side

The resource in the Account B also needs to have permission to push the logs to Central Account S3 Bucket.



IAM Policy - Account B

```
{
    "Version": "2012-10-17",
    "Statement": [
        {
            "Effect": "Allow",
            "Action": [
                "s3:GetObject",
                "s3:PutObject",
                "s3:PutObjectAcl"
            ],
            "Resource": "arn:aws:s3:::central-s3-bucket/*"

        }
    ]
}
```


Central S3 Bucket

Canned ACL

Setting Right Bucket Permissions

Understanding S3 Access ACL

Every bucket and it's objects have an ACL associated with them.

When a request is received, AWS S3 will check against the attached ACL to either allow or block access to that specific object.



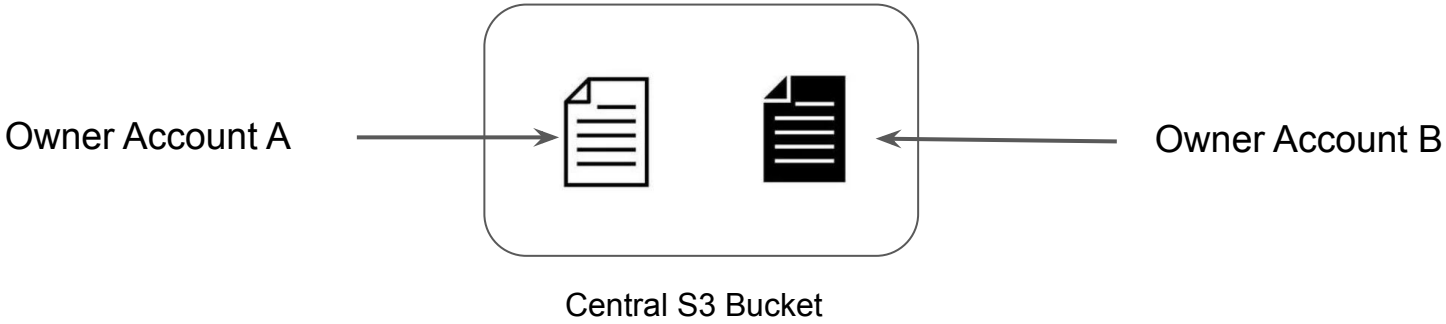
account-a.txt



account-b.txt

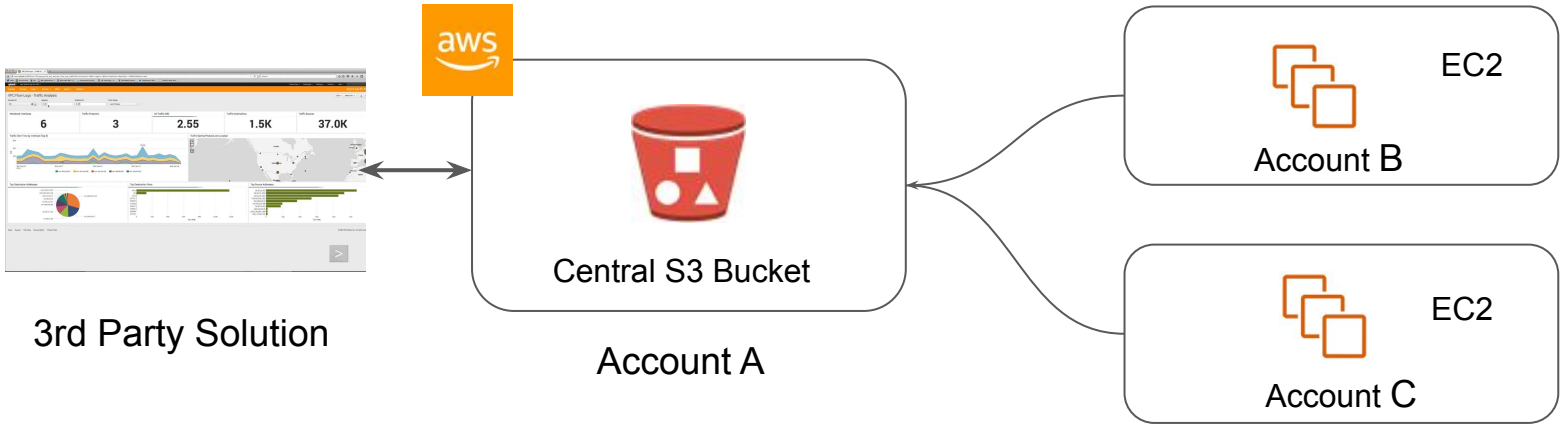
The Tricky Part

When we create a bucket or an object, AWS S3 by default will grant the resource owner full control over the resource.



Ideal Architecture

In most of the architectures, 3rd Party Log Monitoring / SIEM solutions connect to the Central S3 bucket to fetch all of the data.



Canned ACL

AWS S3 supports set of pre-defined grants, known as Canned ACL's.

Each canned ACL has predefined set of permission associated with them.

These canned ACL can be specified in the request using x-amz-acl header.

ACL Name	Description
Private	Owner gets FULL_CONTROL. No one else will have access rights (default)
Public-read	Owner has FULL_CONTROL. All others will have public read permission.
Bucket-owner-read	Owner of the object has FULL_CONTROL. Bucket owner will get read permissions.
Bucket-owner-full-control	Both the object owner and the bucket owner get FULL_CONTROL over the object.

CloudFormation - StackSets

Need to learn the backend

Getting Started

CloudFormation StackSets basically allows us to deploy stacks across multiple AWS account / AWS regions from single location.

Simple Use-Case:

- AWS Config is recommended to be enabled in all regions.
- Before we had to maintain stack across each region.
- This can now be solved easily using Stack Sets

Deployment Instruction

Two IAM Roles required:

1 for the Administrator Account of StackSets

1 for the Destination AWS Accounts.

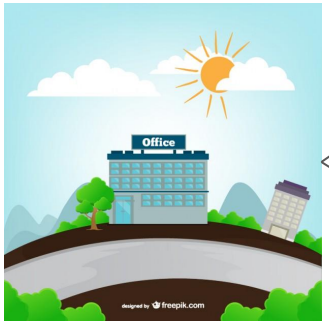
Role Name for Admin Account: AWSCloudFormationStackSetAdministrationRole

Role Name for Dest Account: AWSCloudFormationStackSetExecutionRole

Active Directory

Directory Service

Traditional Way



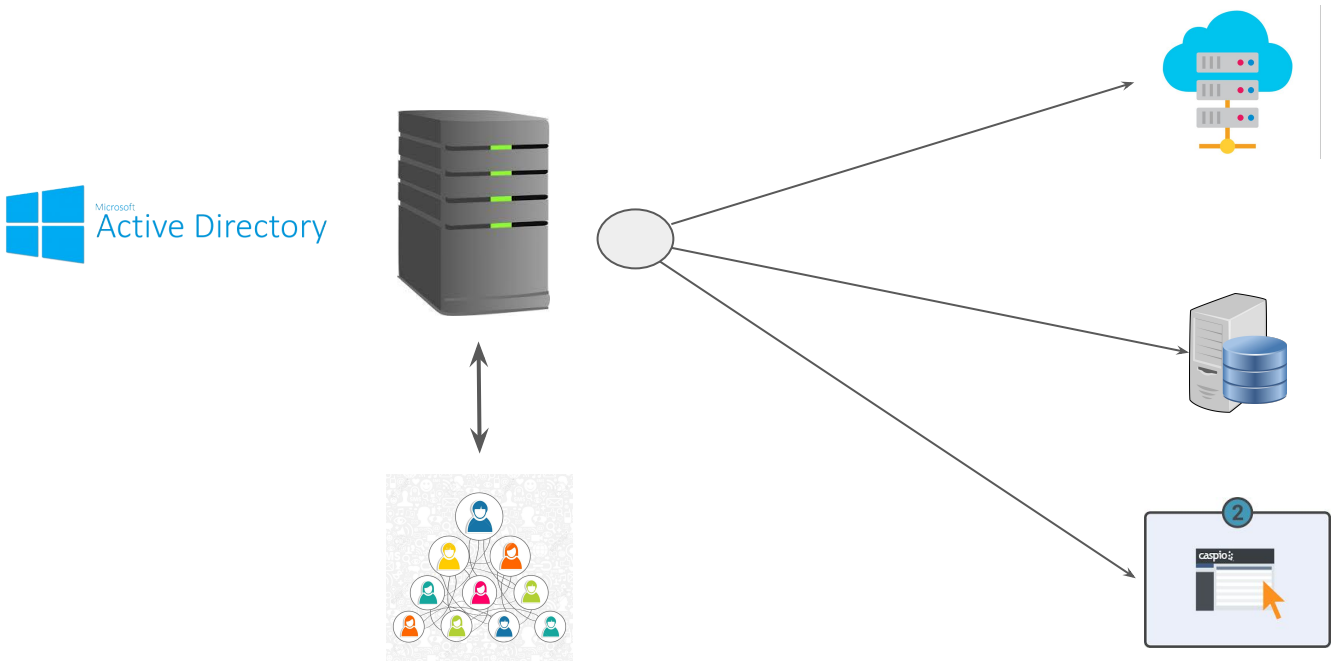








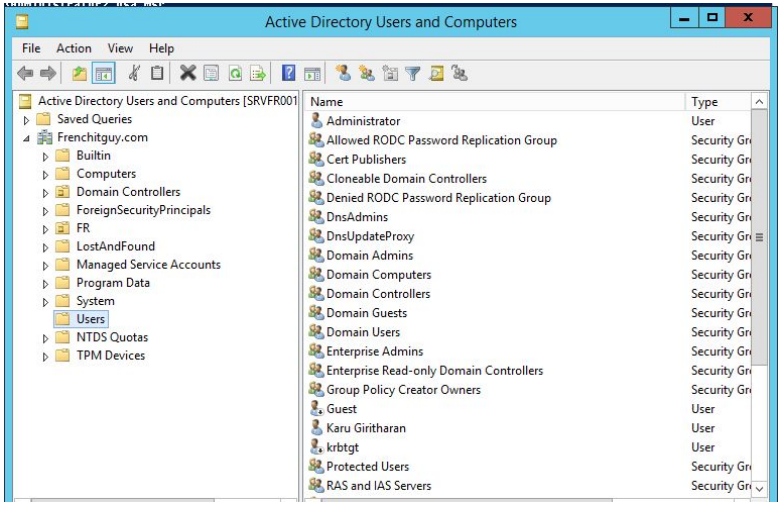
Better Way



Active Directory

Active Directory is one of the most popular directory service developed by Microsoft.

The server running the Active Directory service is called as the domain computer and it can authenticate and authorize the users and computers which are associated to it.



AWS Directory Service

Directory on the Cloud

Challenges with Active Directory

For those who have set up an AD knows, this can be a challenging and time-consuming process.

Some of the challenges involved can be:



- Provisioning the Infrastructure.
- Installing the directory software
- Getting replication setup between domain controllers for HA
- Monitoring / Patching and many more.

Directory Service in the Cloud

AWS Directory Service is a managed service based on the cloud that allows us to create directories and let AWS experts handle and manage the other parts like high availability, monitoring, backups, recovery, and others.

There are three important components :

- Active Directory Service with Microsoft Active Directory
- Simple AD
- AD Connector

Directory Service with Microsoft AD

AWS Directory Service for Microsoft Active Directory is powered by an actual Microsoft Windows Server Active Directory (AD) in the AWS Cloud.

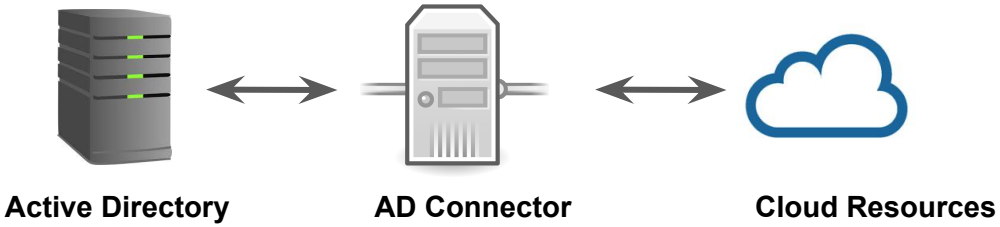
There are two types:

- Standard Edition -- For small and midsize (up to 5000 users)
- Enterprise Edition -- For larger deployments.



AD Connector

- It is a proxy service that provides easy way to connect applications in cloud to your existing on-premise Microsoft AD.
- When users log in to the applications, AD Connector forwards sign-in requests to your on-premises Active Directory domain controllers for authentication.



Simple AD

- Simple AD is a Microsoft Active Directory–compatible directory from AWS Directory Service that is powered by Samba 4.

- Simple AD supports basic Active Directory features such as user accounts, group memberships, joining a Linux domain or Windows based EC2 instances, Kerberos-based SSO, and group policies. AWS provides monitoring, daily snapshots, and recovery as part of the service.

- Simple AD does not support trust relationships, DNS dynamic update, schema extensions, multi-factor authentication, communication over LDAPS, PowerShell AD cmdlets, or FSMO role transfer.

Federation

Connecting Identities

Understanding the Challenge

Let's assume there are 500 users within an organization. Your organization are using 3 services :-

- AWS (Infrastructure)
- Jenkins (CI / CD)
- HR Activator (Payroll)

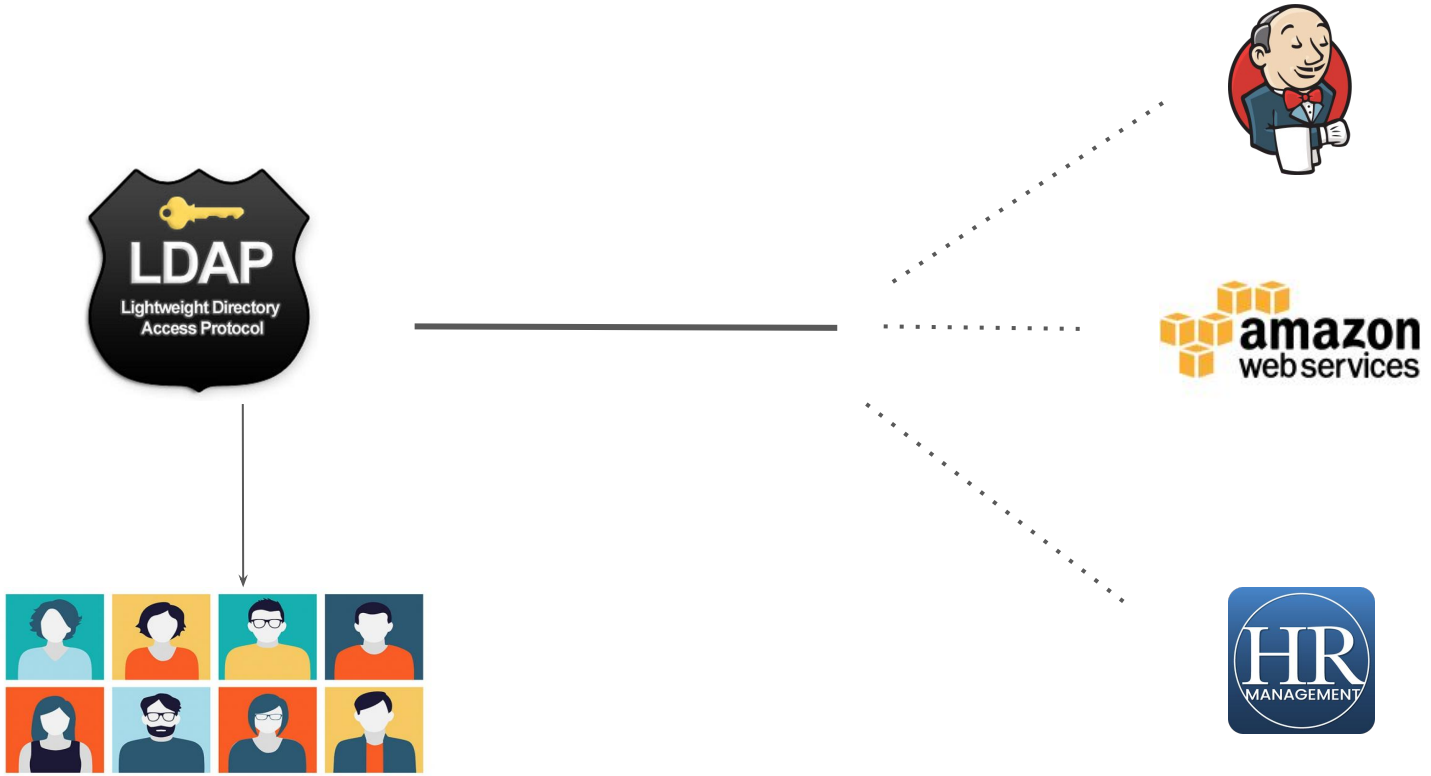


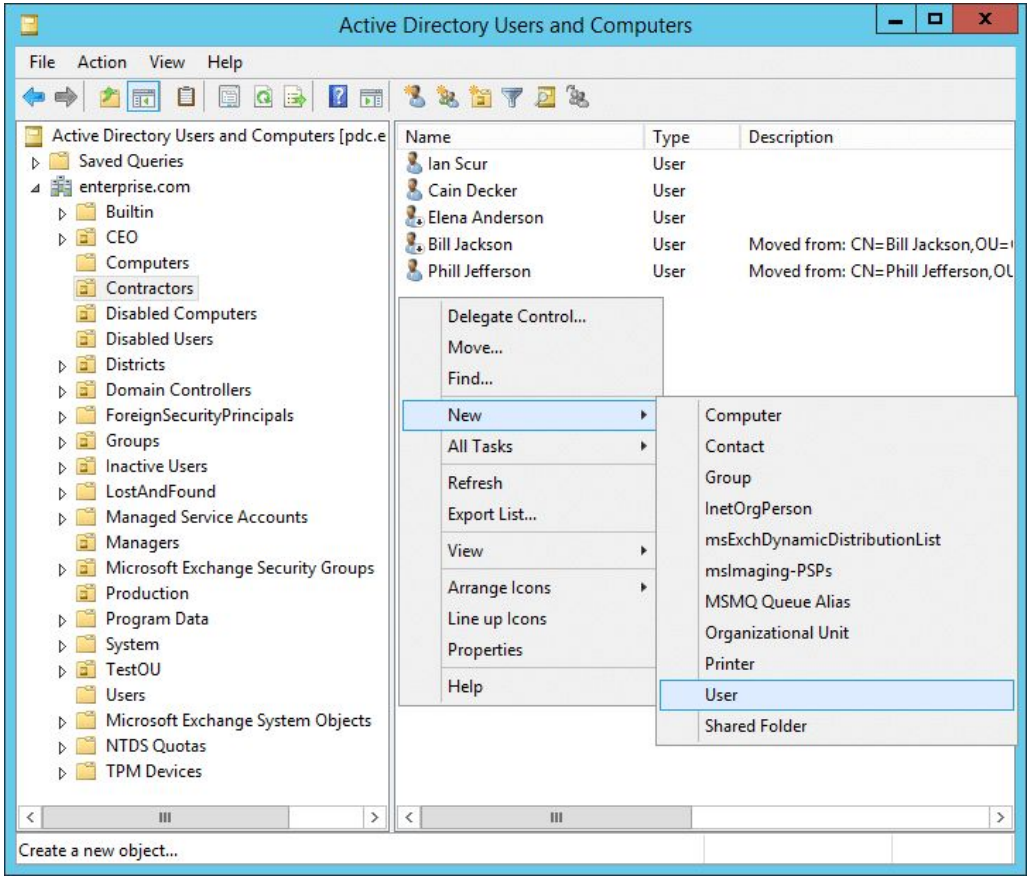




You have been assigned role to give users access to all 3 services.

Storing Users Centrally




Active Directory Users and Computers
File
Action
View
Help
Active Directory Users and Computers [pdc.e
Saved Queries
enterprise.com
Builtin
CEO
Computers
Contractors
Disabled Computers
Disabled Users
Districts
Domain Controllers
ForeignSecurityPrincipals
Groups
Inactive Users
LostAndFound
Managed Service Accounts
Managers
Microsoft Exchange Security Groups
Production
Program Data
System
TestOU
Users
Microsoft Exchange System Objects
NTDS Quotas
TPM Devices
Name
Type
Description
Ian Scur
User
Cain Decker
User
Elena Anderson
User
Bill Jackson
User
Moved from: CN=Bill Jackson,OU=
Phill Jefferson
User
Moved from: CN=Phill Jefferson,OU
Delegate Control...
Move...
Find...
New
All Tasks
Refresh
Export List...
View
Arrange Icons
Line up Icons
Properties
Help
Computer
Contact
Group
InetOrgPerson
msExchDynamicDistributionList
msImaging-PSPs
MSMQ Queue Alias
Organizational Unit
Printer
User
Shared Folder
Create a new object...

Central Users

There are various solutions available which can store users centrally :-

- Microsoft Active Directory
- RedHat Identity Management / freeIPA





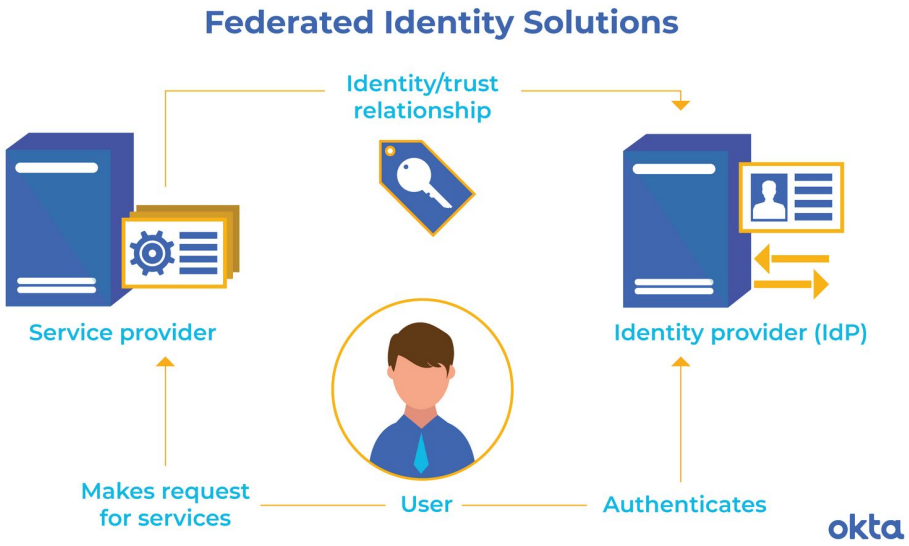
Basics of Federation - AWS Perspective

Federation allows external identities (Federated Users) to have secure access in your AWS account without having to create any IAM users.

These external identities can come from :-

- Corporate Identity Provider (AD, IPA)
- Web Identity Provider (Facebook, Google, Amazon, Cognito or OpenID)

Basic Workflow



Understanding Identity Broker

Identity Broker :-

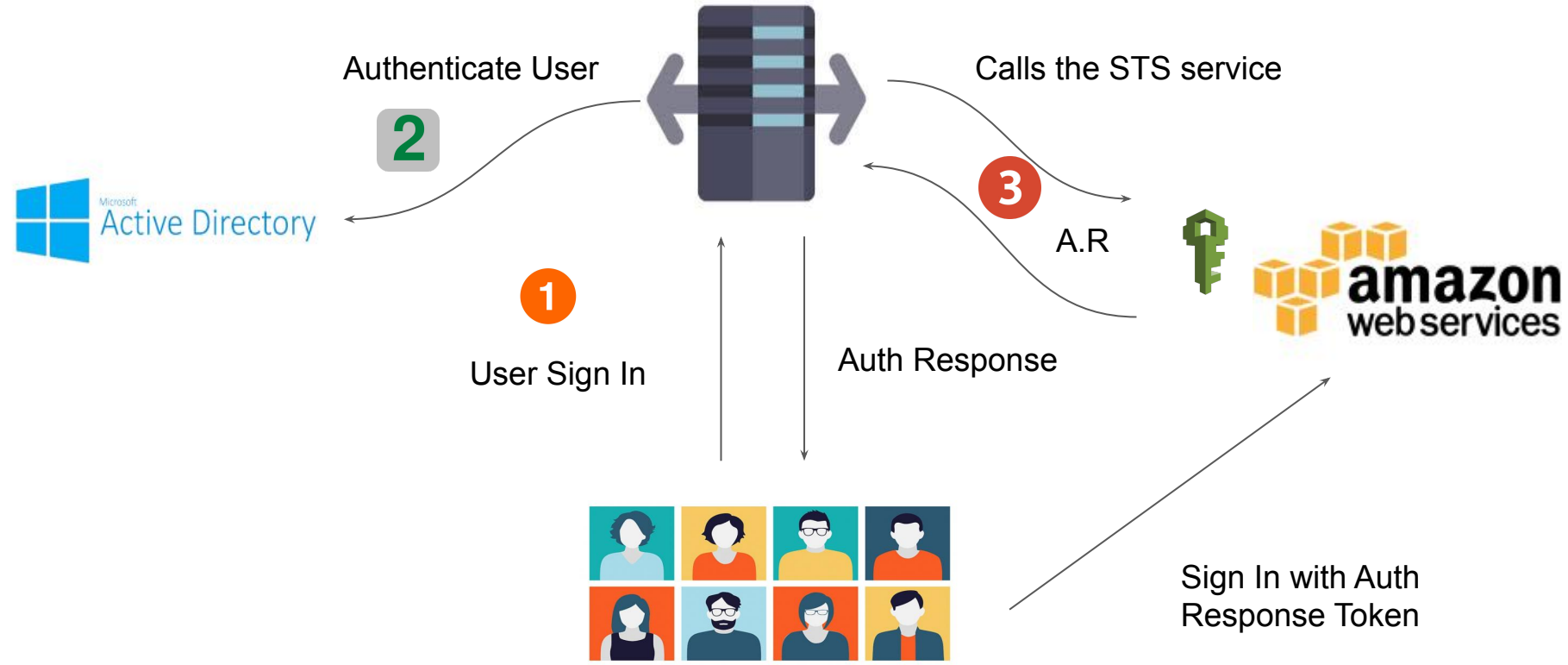
It is an intermediate service which connects multiple providers.

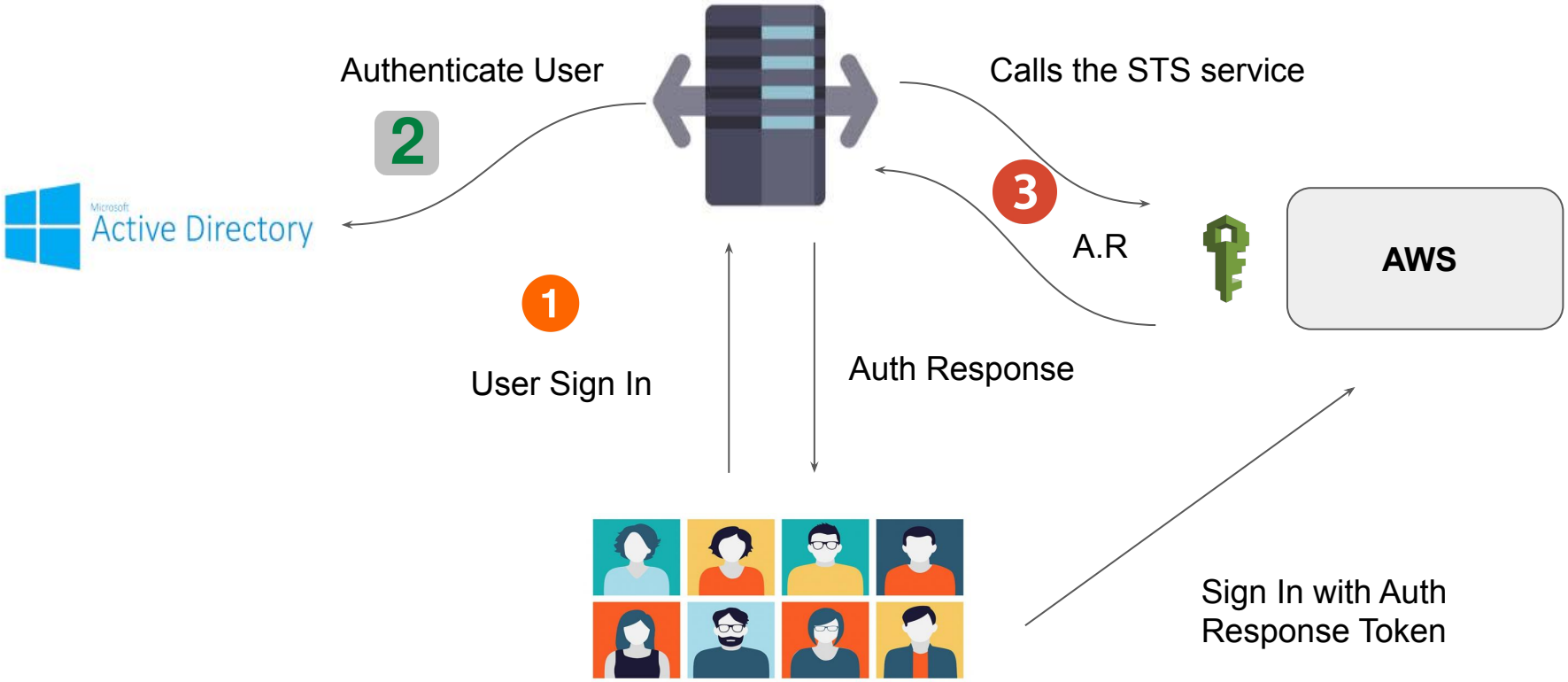


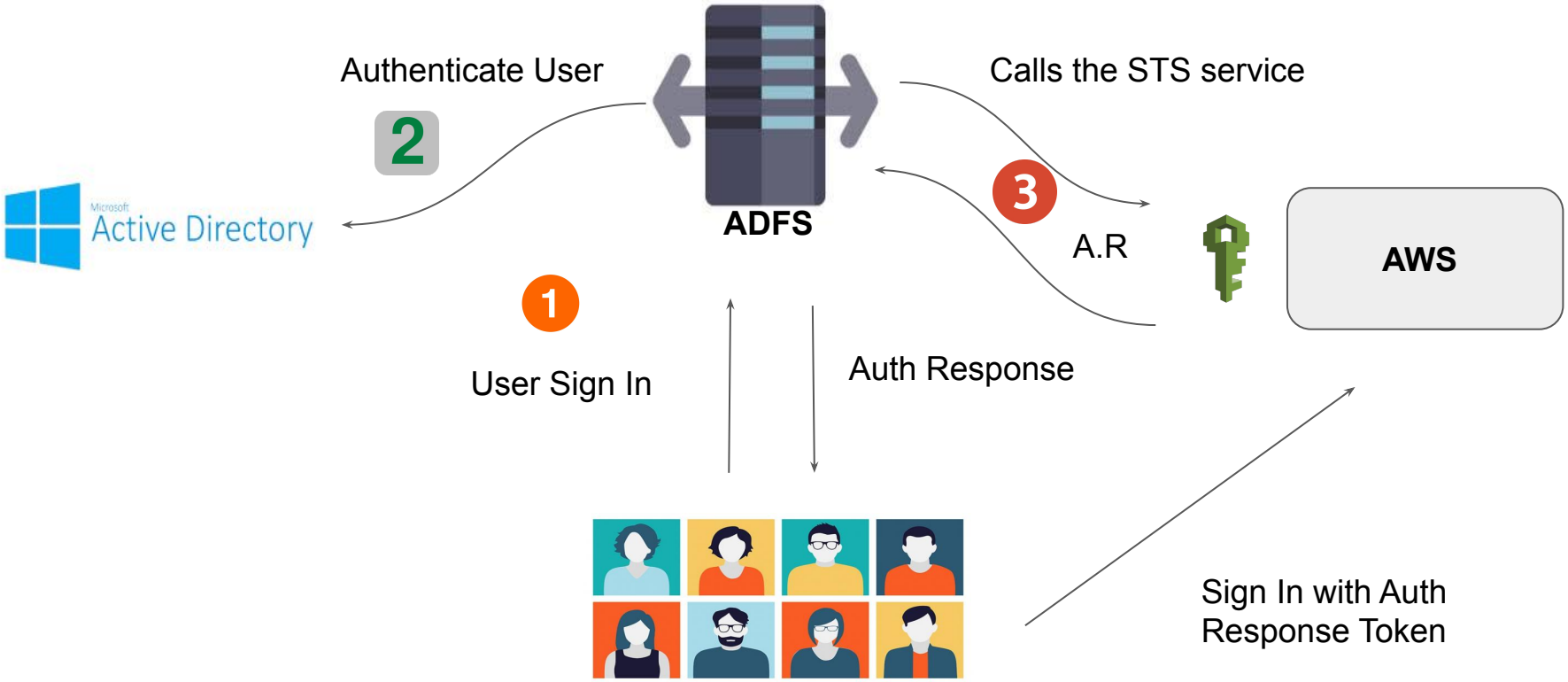


Identity Broker




Authenticate User
2
Calls the STS service
3
A.R
Microsoft
Active Directory
amazon
web services
1
User Sign In
Auth Response
Sign In with Auth
Response Token


Authenticate User
2
Microsoft
Active Directory
Calls the STS service
3
A.R
AWS
1
User Sign In
Auth Response
Sign In with Auth
Response Token


Authenticate User
2
Microsoft
Active Directory
ADFS
Calls the STS service
3
A.R
AWS
1
User Sign In
Auth Response
Sign In with Auth
Response Token

Steps to Remember

- User logs in with username & Password.
- This credentials are given to the Identity Broker.
- Identity Broker validates it against the AD.
- If credentials are valid, Broker will contact the STS token service.
- STS will share the following 4 things :-

  Access Key + Secret Key + Token + Duration

- User can now use to login to AWS Console or CLI.

Notations to Remember

Identities : Users

Identity Broker :

- It is a middlware that takes the users from point A & help connect them to point B.

Identity Store :-

- Place where users are present. Eg : AD, IPA, Facebook etc.

SAML

Single Sign On

Introduction to SAML

- SAML stands for Security Assertion Markup Language.
- It is a secure XML based communication mechanism for communicating identities across organizations.
- SAML eliminates the need to maintain multiple authentication credentials, such as passwords in multiple locations.

Classic Way













Challenges with classic way

- The administrator does not have direct visibility with the underlying database of the SAAS provider.
- If there are multiple SAAS providers, it is difficult to keep track of which user has access to which SAAS application.
- When the user leaves the organization, he needs to be removed from all the entities (Jenkins, AWS, HR app)

Different Views

Administrator's View

Have to login to different providers to manage and control the permissions of an individual user across the organization.

User forgetting username and passwords, MFA :(

User's View

I have to remember passwords of all the applications in the organization.

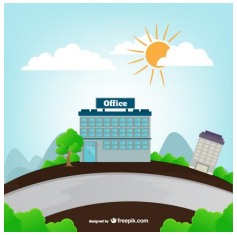
It might be possible that even userID across apps is different, so have to remember that as well.

SAAS Provider's View

Have to maintain the user and password database of customers.

This is a big security liability.

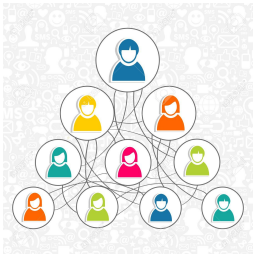
SAML



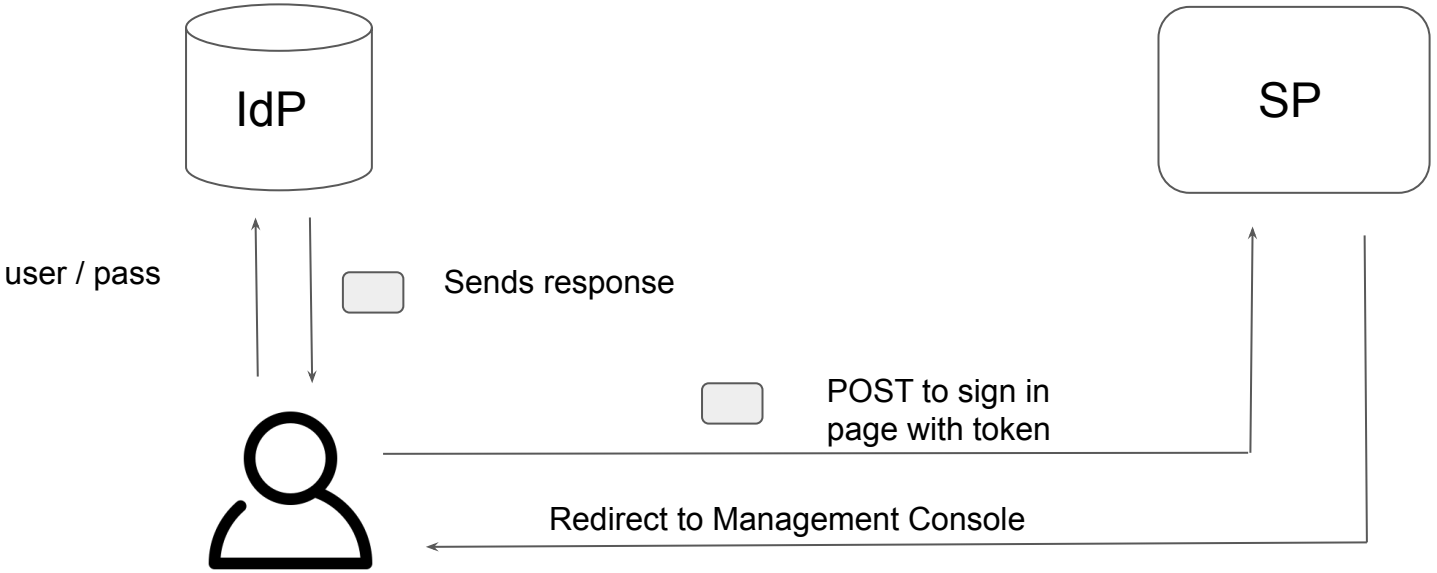
Identity Provider



Service Provider



The SAML Way


IdP
SP
user / pass
Sends response
POST to sign in
page with token
Redirect to Management Console

Introduction to SAML

- The flow gets initiated when user opens the IDP URL and enters the username and password and selects the appropriate application.

- IdP will validate the credentials and associated permissions and then user receives SAML assertion from the IdP as part of response.

- User does a POST of that SAML assertion to the SAAS sign in page and SP will validate those assertion.

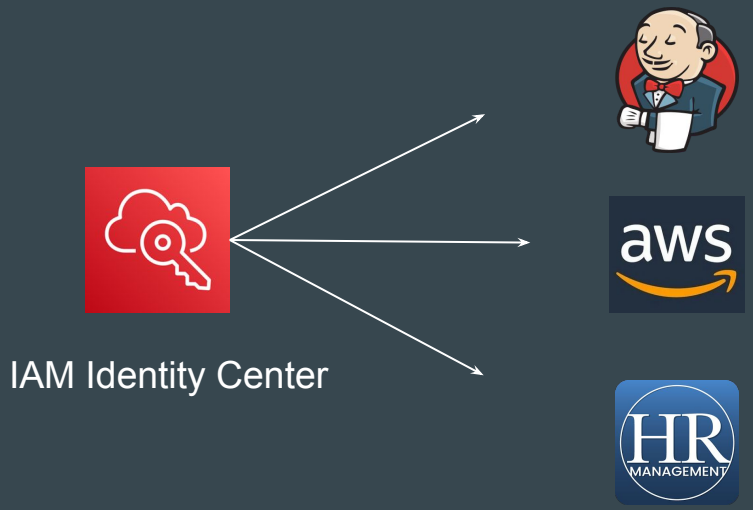
- On validation, SP will construct relevant temporary credentials, and constructs sign in URL for the console and sends to the user.

IAM Identity Center

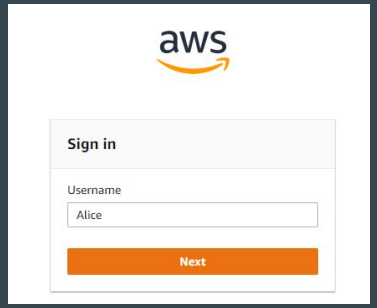


Understanding the Basics

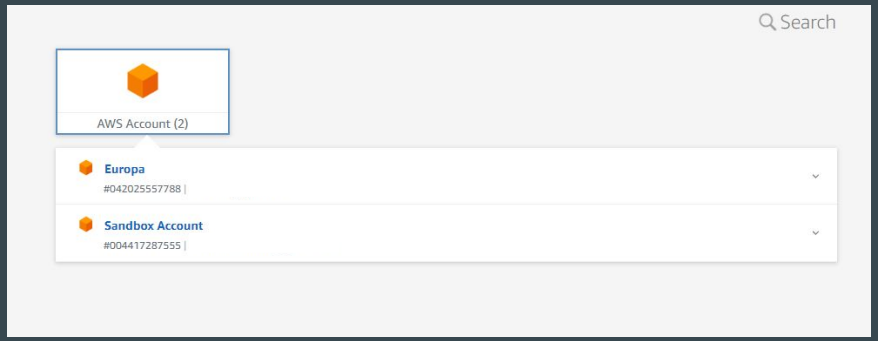
IAM Identity Center (successor to AWS Single Sign-On) allows centralized access to multiple AWS accounts and applications and provide users with single sign-on access to all their assigned accounts and applications from one place.



Basic Steps

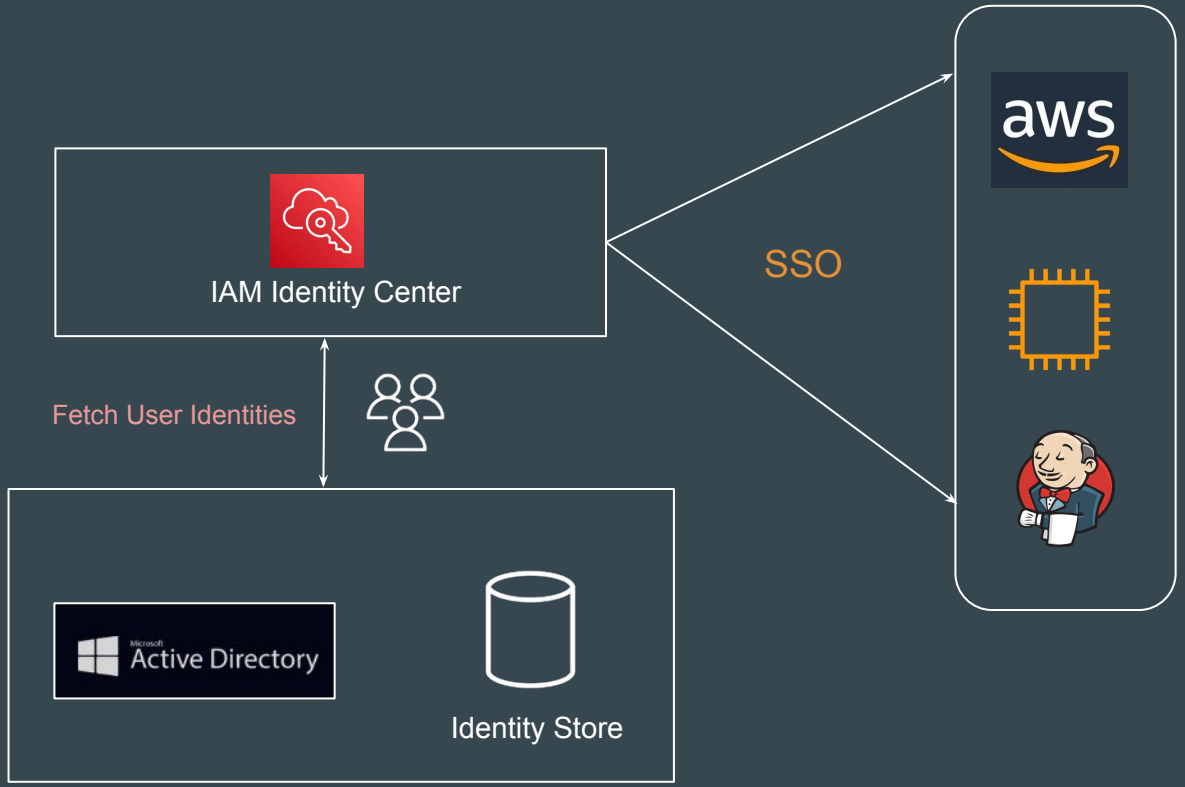


Login to Access Portal



Connect with AWS Accounts / Apps available

Understanding the Workflow



SSO with AWS CLI

AWS CLI integrates with the SSO.

SSO users can authenticate via CLI, and they will be able to perform the CLI operations without having to add keys in their ~/.aws/credentials file.

```
C:\Users\zealv>aws s3 ls --profile AdministratorAccess-004417287555

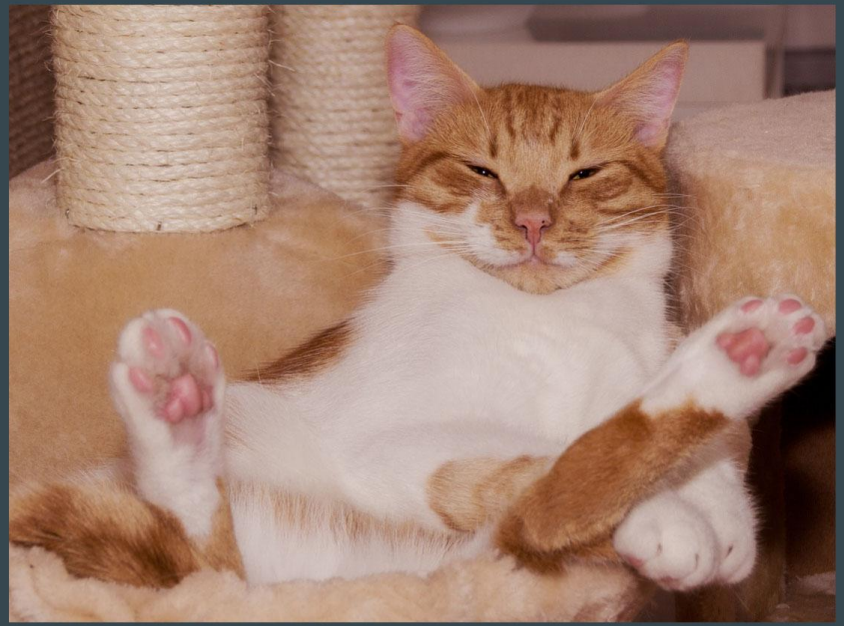
C:\Users\zealv>aws sso login --profile AdministratorAccess-004417287555
Attempting to automatically open the SSO authorization page in your default browser.
If the browser does not open or you wish to use a different device to authorize this request, open the following URL:

https://device.sso.us-east-1.amazonaws.com/

Then enter the code:

KQKZ-NRWR
Successfully logged into Start URL: https://d-9067a61937.awsapps.com/start
```

IAM Team After Implementing SSO



Benefits of IAM Identity Center

Your users can use their directory credentials for single sign-on access to multiple AWS accounts.

Enable single sign-on access to your AWS applications

Enable single sign-on access to Amazon EC2 Windows instances

Enable single sign-on access to cloud-based applications that support SAML

IAM Identity Center - Concepts & Considerations



Prerequisite for Identity Center

Your AWS account must be managed by AWS Organizations.

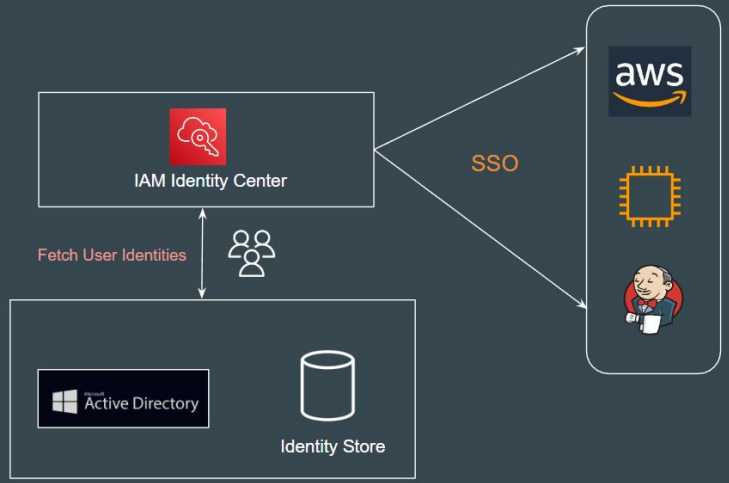
If you've already set up AWS Organizations, make sure that all features are enabled

When you enable IAM Identity Center, you will choose whether to have AWS create an organization for you.

Identity Source

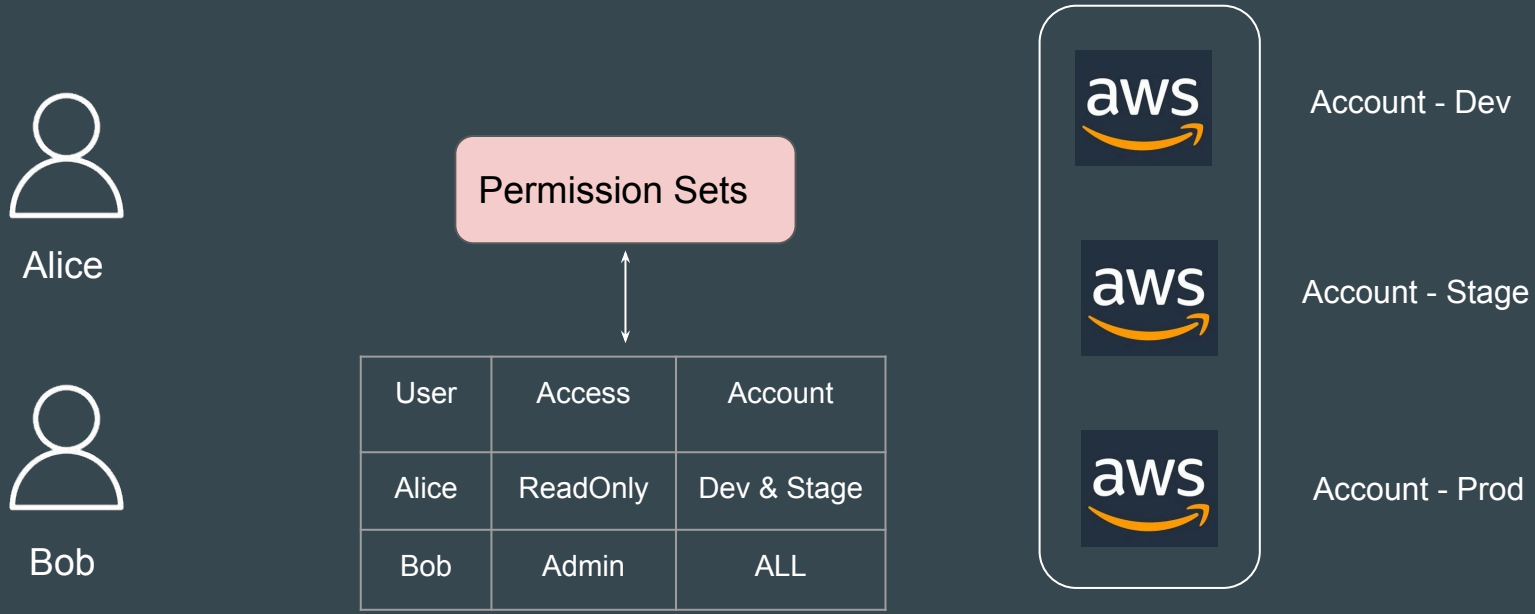
If you're already managing users and groups in Active Directory or an external IdP, it is recommended that you consider connecting this identity source when you enable IAM Identity Center and choose your identity source.

You can also create users and groups directly in IAM Identity Center.

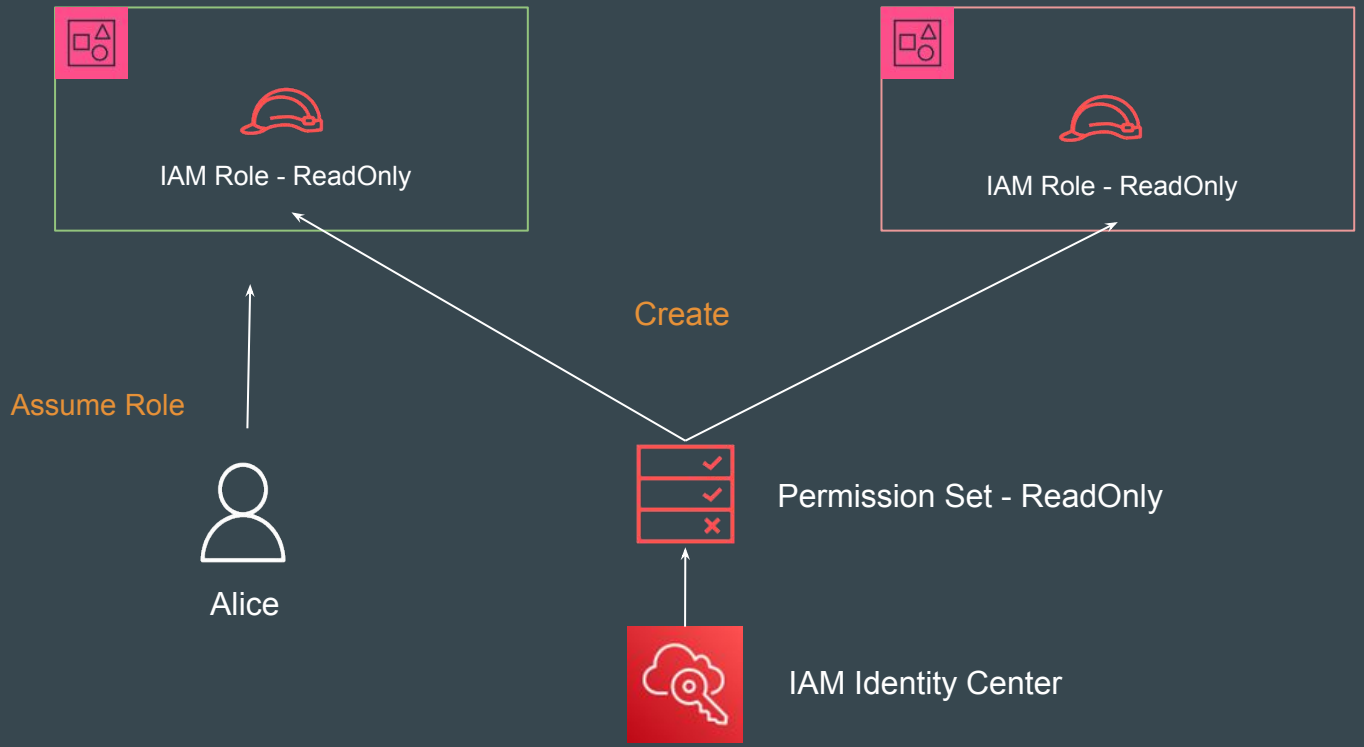


Permission Sets

Permission sets define the level of access that users in IAM Identity Center have to their assigned AWS accounts



How it Works


IAM Role - ReadOnly
IAM Role - ReadOnly
Create
Assume Role
Permission Set - ReadOnly
Alice
IAM Identity Center

Points to Note - IAM Identity Center



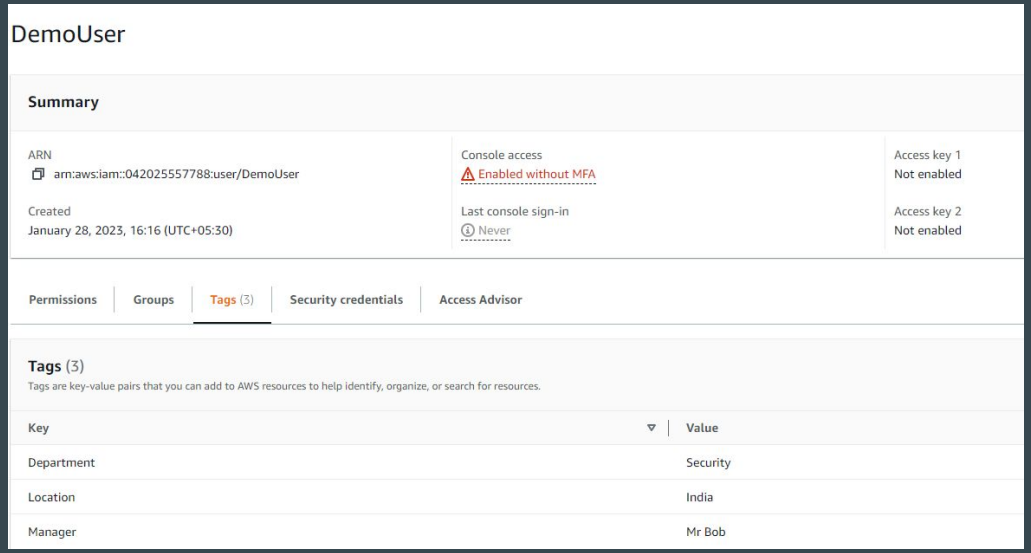
SAML Implementation

IAM Identity Center supports identity federation with SAML (Security Assertion Markup Language) 2.0.

This allows IAM Identity Center to authenticate identities from external identity providers (IdPs

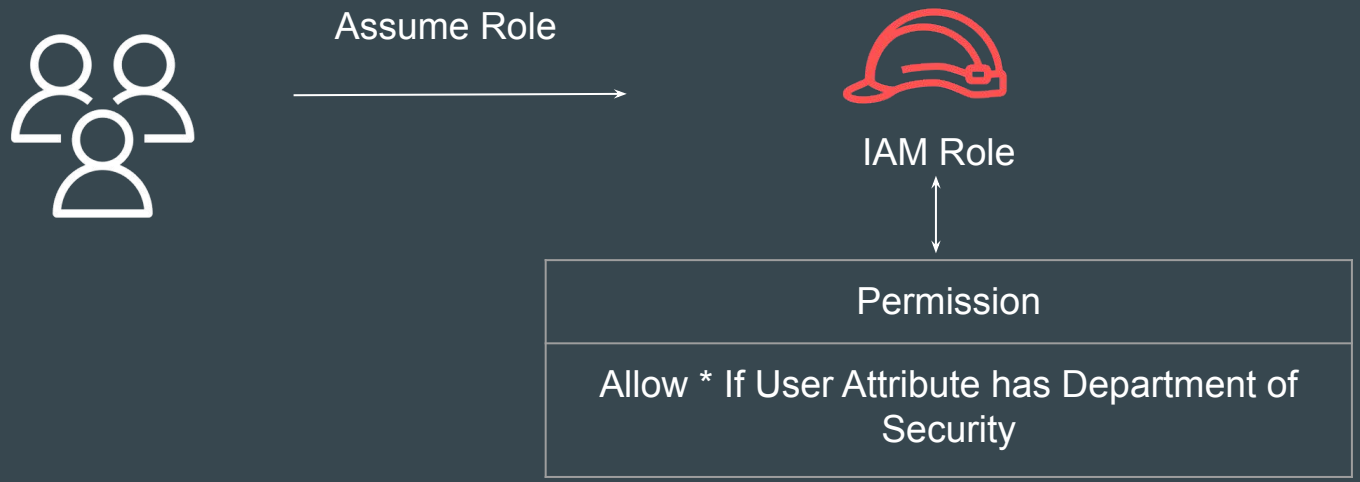
Attributes in IAM

You can use IAM tag key-value pairs to add custom attributes to an IAM user.



Attribute-Based Access Control

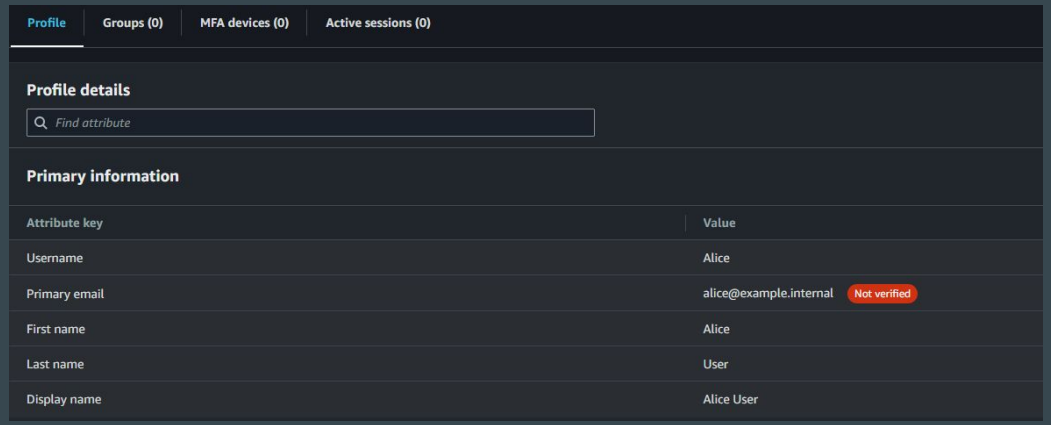
Attribute-based access control (ABAC) is an authorization strategy that defines permissions based on attributes.



How to Set Attributes?

Depending on the Identity Source, the way we set Attribute also changes.

In IAM Identity Center, we can easily set user attributes from Profile.



Permissions Based on ABAC

Depending on the Identity Source, the way we set Attribute also changes.

In IAM Identity Center, we can easily set user attributes from Profile.



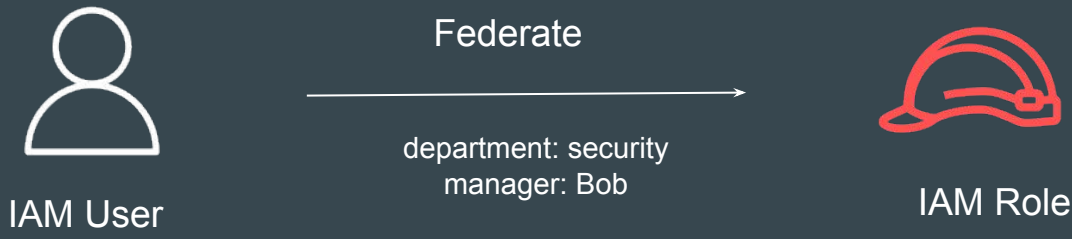
Attribute Key	Values
Department	DevOps

```
{
    "Version": "2012-10-17",
    "Statement": {
        "Effect": "Allow",
        "Action": [
            "ec2:startInstances",
            "ec2:stopInstances"
        ],
        "Resource": "*",
        "Condition": {"StringEquals":
            {"aws:ResourceTag/Department": "${aws:PrincipalTag/DevOps}"}}
    }
}
```

Importance of Session Tags

Session tags are key-value pair attributes that you pass when you assume an IAM role or federate a user in AWS STS.

Attributes are passed as session tags. They are passed as comma-separated key:value pairs

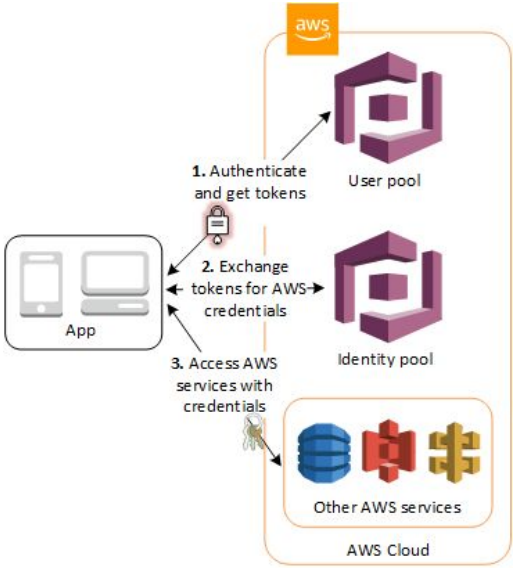


Amazon Cognito

Federation

Basics of Cognito

Amazon Cognito provides authentication, authorization, and user management service for your web and mobile apps.



Sample Use-Case

Alice is a mobile developer in a start-up organization. They have begun with mobile wallet system, and there are specific requirements as follows:

- Users should be able to sign-up with new credentials.
- User should be able to sign-in with social platforms like FB, Twitter, Google.
- There should be a post sign-up process (one-time password) for verification.
- Multi-Factor authentication should be present.
- Account recovery feature should be present.

In-Short: Build a Complete Authentication & Authorization System

Amazon Cognito

At a high level, there are two major features under Amazon Cognito

i) User Pools
ii) Identity Pools

Cognito user pool takes care of the entire authentication, authorization process .

Identity pool provides the functionality of federation for users in user pools.

Identity Pool

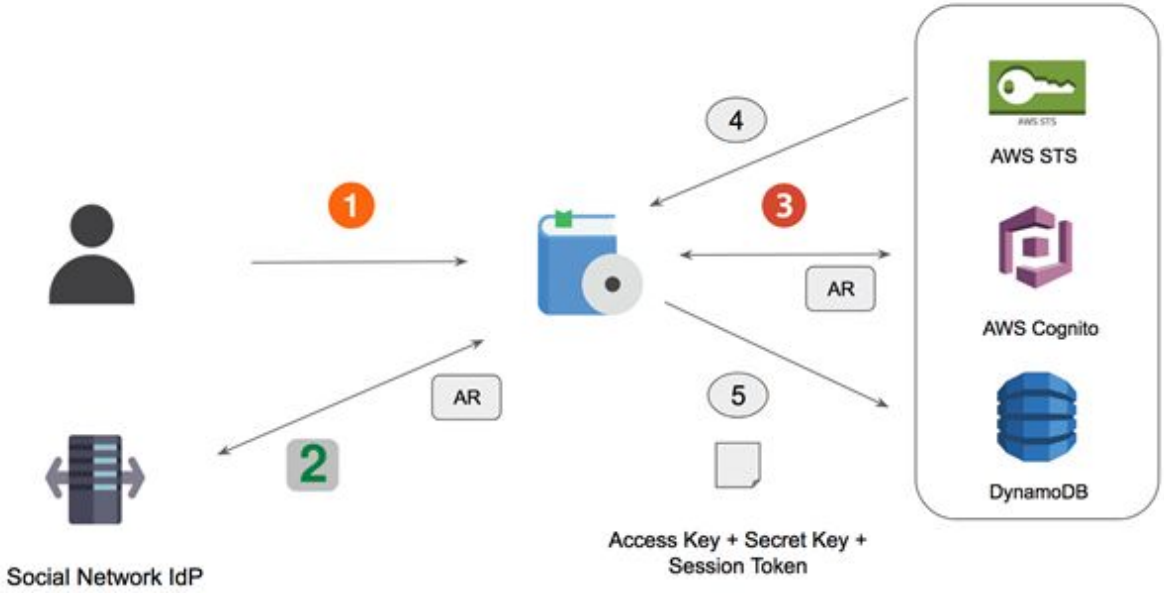
Cognito Identity pools also referred to as AWS Cognito Federated Identities allows developers to authorize the users of the application to use various AWS services.

Use-Case:

We have a quiz based mobile application. At the end of quiz, user's results should be stored in the DynamoDB table.

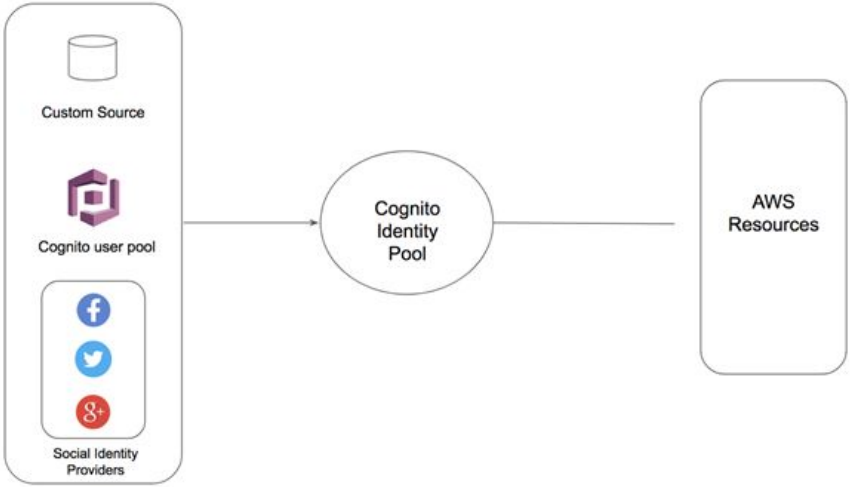
If we hard-code the access/secret keys, chances of reverse engineering are high.

Cognito Identity Pool Working - NO



User Pool vs Identity Pool - NO

The Cognito Identity pool then takes these identities and federates them and then can give secure access to the AWS services regardless of where the user comes from.

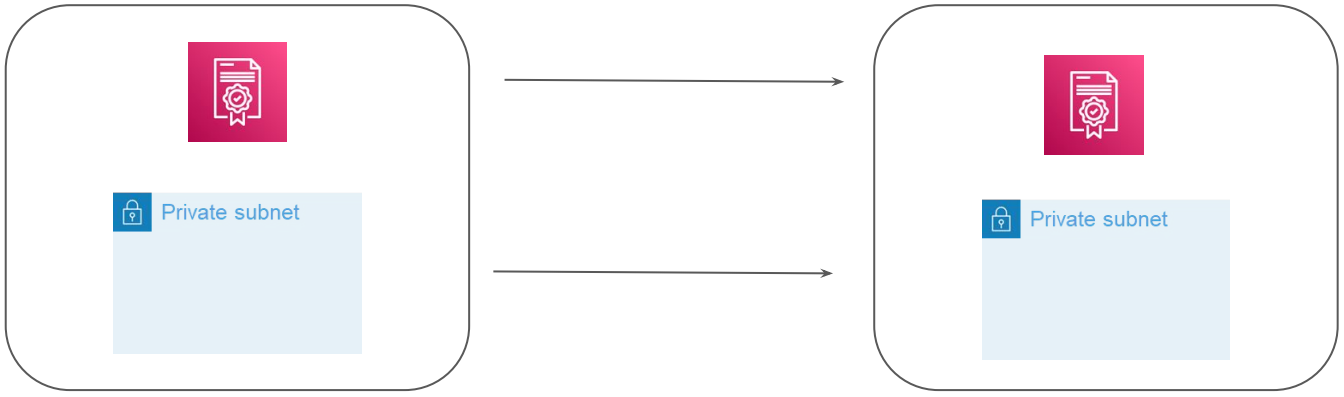


AWS Resource Access Manager

Let's Share Resources

Overview of Resource Access Manager

AWS Resource Access Manager (AWS RAM) helps you securely share the AWS resources that you create in one AWS account with other AWS accounts.



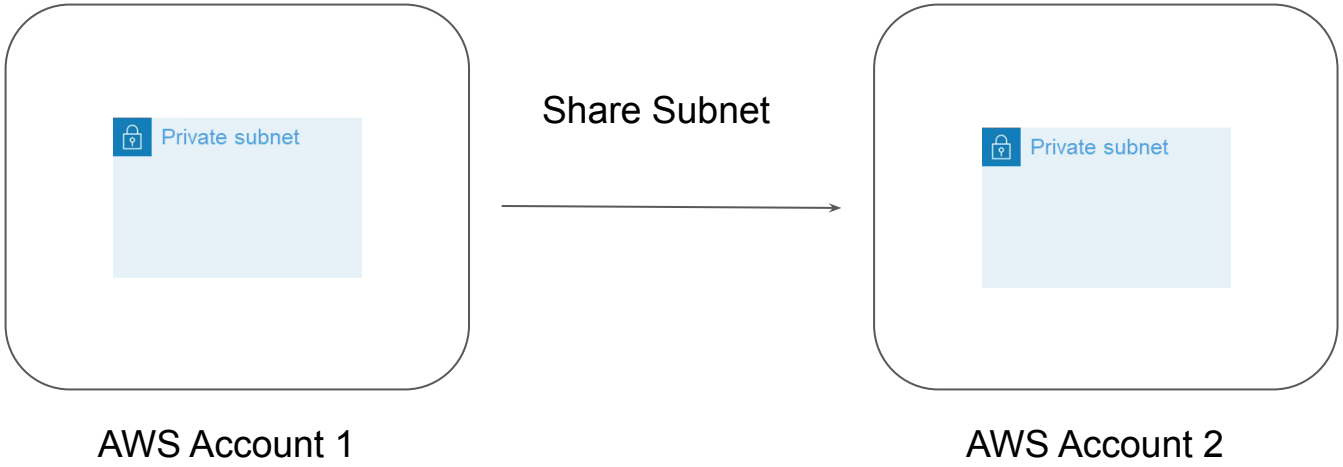
VPC Sharing in AWS

Let's Share Subnets

Understanding the Basics

VPC sharing allows multiple AWS accounts to create their application resources, such as EC2 instances, RDS, and others into shared, centrally-managed virtual private clouds (VPCs).

In this model, the account that owns the VPC (owner) shares one or more subnets with other accounts (participants) that belong to the same organization from AWS Organizations.



Important Note

VPC owners are responsible for creating, managing, and deleting the resources associated with a shared VPC. These include subnets, route tables, network ACLs and others.

VPC owners cannot modify or delete resources created by participants, such as EC2 instances and security groups

Default subnets cannot be shared.

Billing Considerations

In a shared VPC, each participant pays for their application resources including EC2 instances, RDS, Lambda functions and other resources.

Participants also pay for data transfer charges associated with inter-Availability Zone data transfer, data transfer over VPC peering connections.

VPC owners pay hourly charges across NAT gateways, virtual private gateways, transit gateways, and other VPC specific central resources.

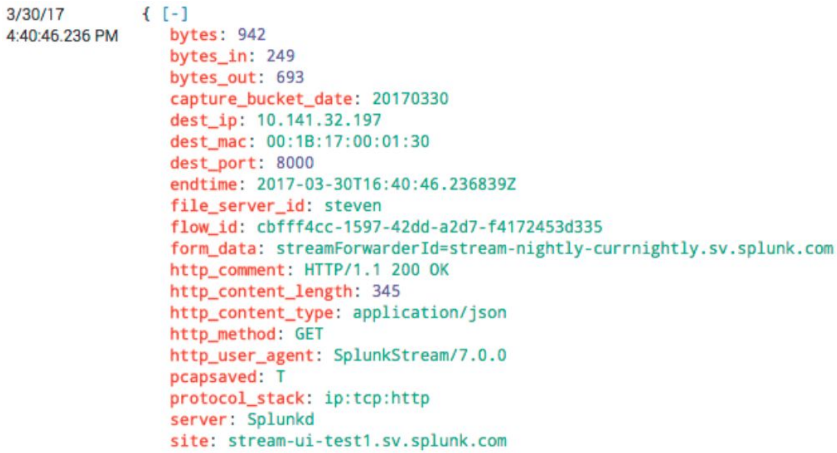
Traffic Mirroring

Capture Network Traffic

Understanding the Challenge

Many organizations use various kind of wire data collection tools like Splunk stream to capture specific type network traffic to analyze for security threats.

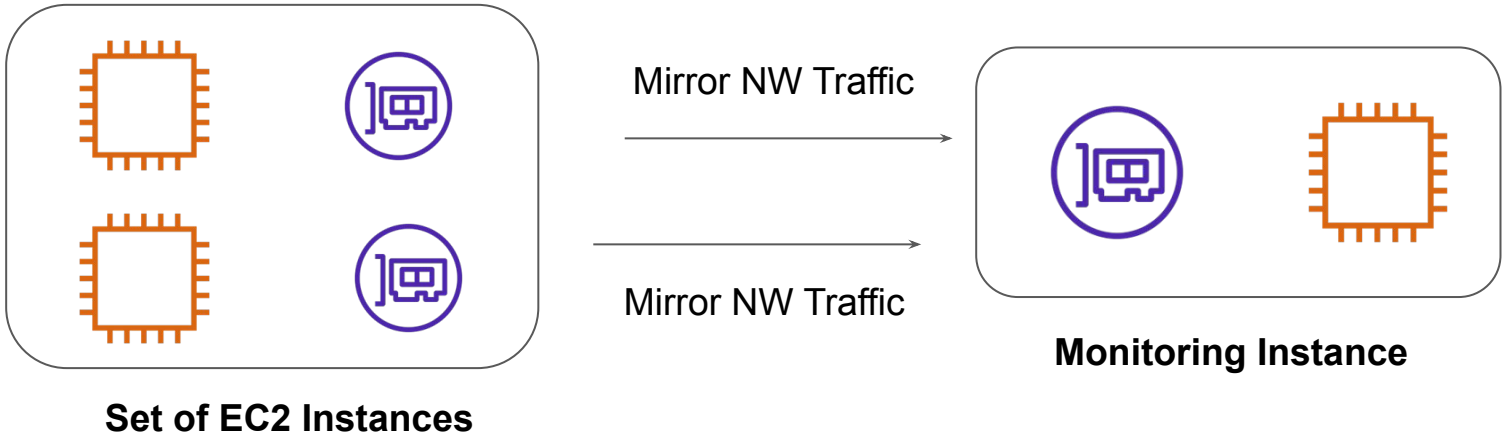
This used to impact the overall system performance.



Basics of Feature

Traffic Mirroring is an Amazon VPC feature that you can use to copy network traffic from an elastic network interface.

You can then send the traffic to out-of-band security and monitoring appliances for:



Relax and Have a Meme Before Proceeding

Do you have a special talent?

Me:

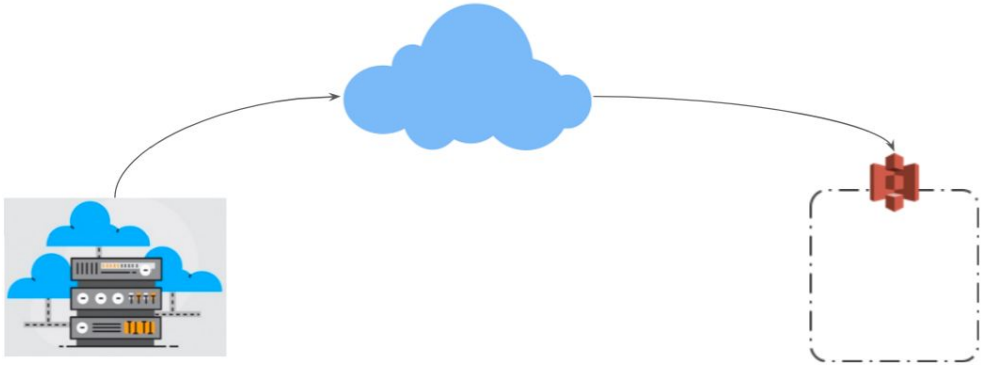


VPC Endpoints

Private Communication is Better

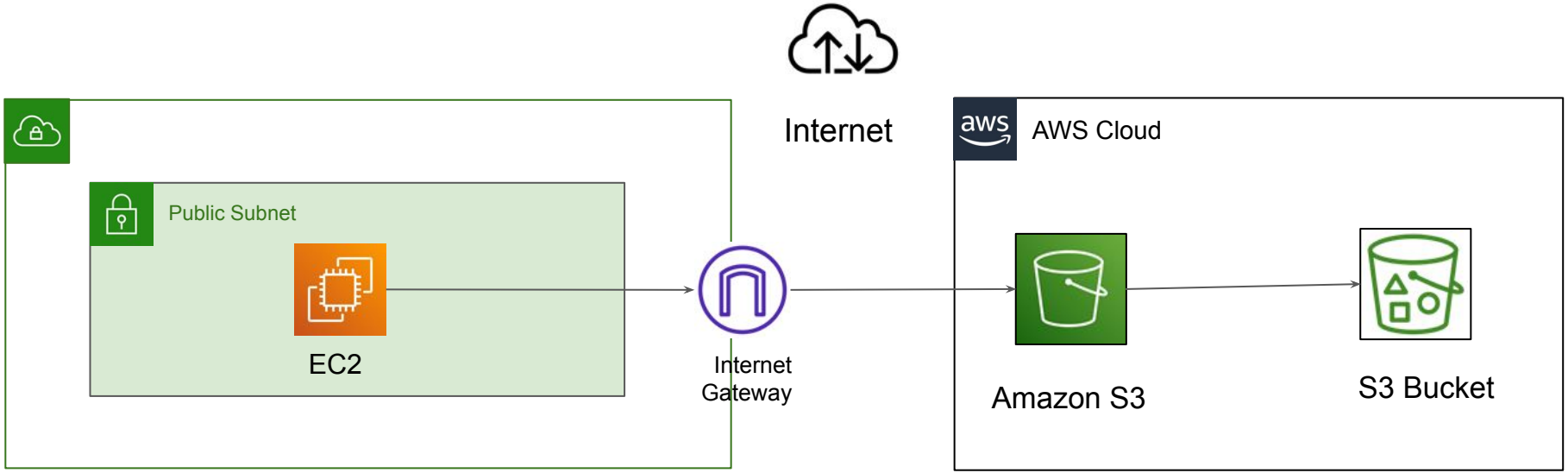
Use-Case: EC2 and S3 Communication

For EC2 instances to be able to access public resources like S3, DynamoDB and others, the traffic needed to be passed via Internet Gateway.

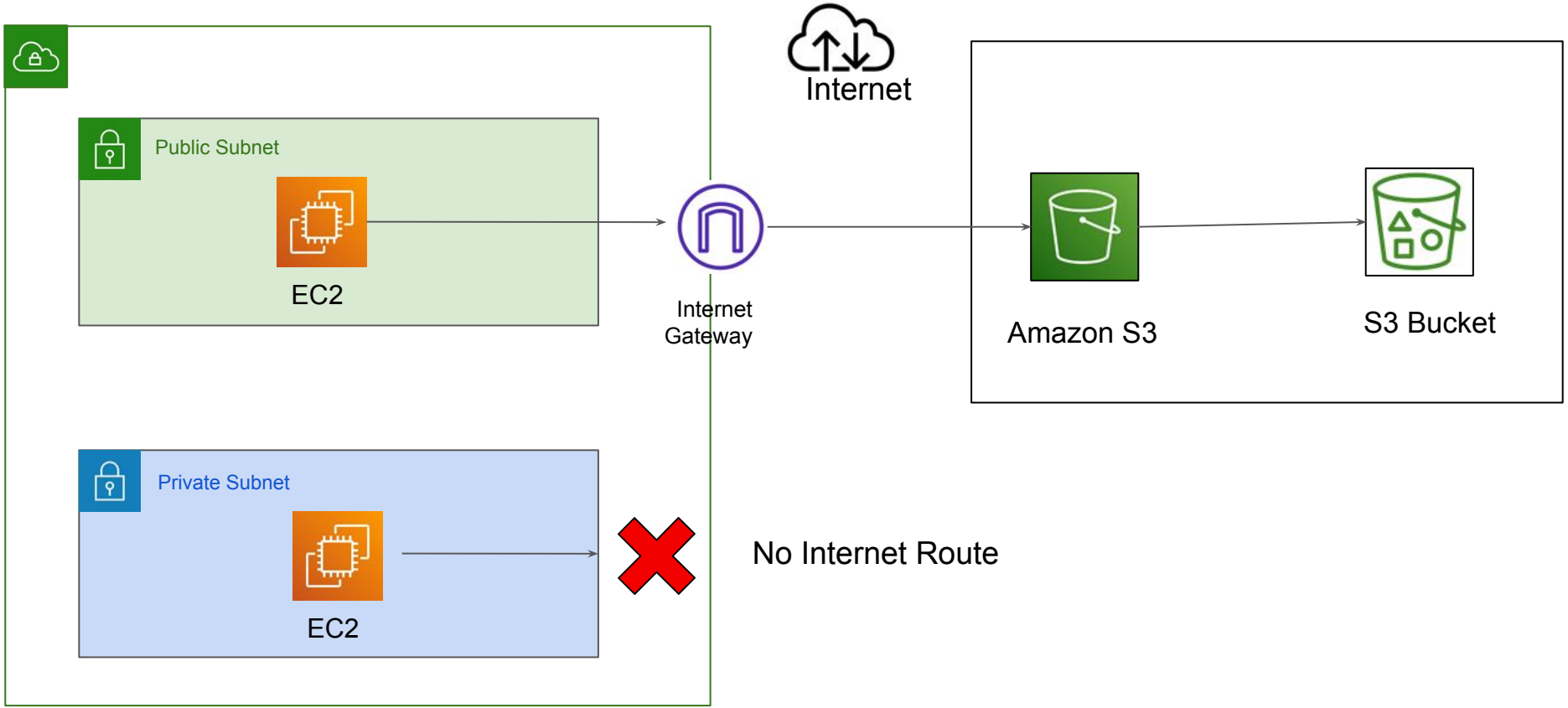


Architectural Perspective

EC2 traffic towards S3 is routed to Internet Gateway



Challenge with Private Workloads



Downsides of Using Public Internet

1. Data transfer cost of AWS

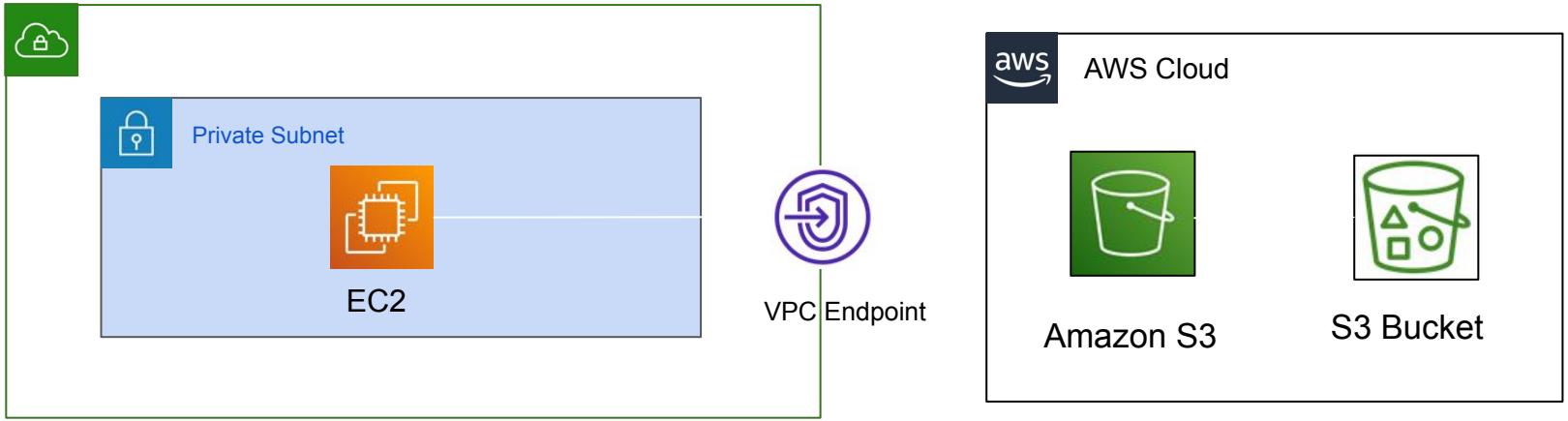
2 . Higher Latency

3. Can bottleneck your internet gateway.

4. Security

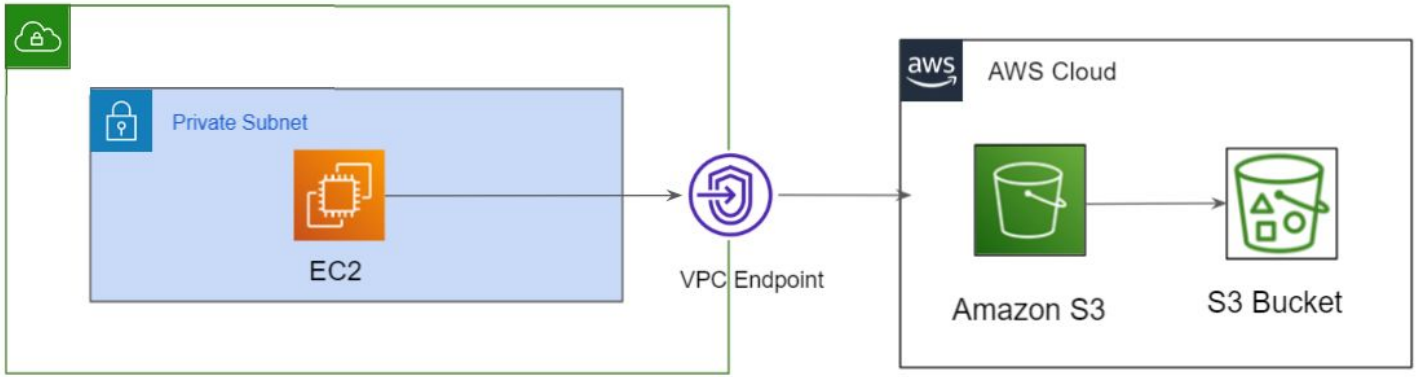
Overview of VPC Endpoints

VPC Endpoints allows us to connect VPC to another AWS services OR other supported services over AWS private network.



Overview of VPC Endpoints

VPC Endpoints allows us to connect VPC to another AWS services OR other supported services over AWS network.



Revising Important Pointers

AWS PrivateLink is a technology that enables you to privately access services by using private IP addresses.

To use AWS PrivateLink, you can create a VPC endpoint for a service in your VPC.

VPC Endpoint allows us to connect VPC to another AWS services over AWS network.

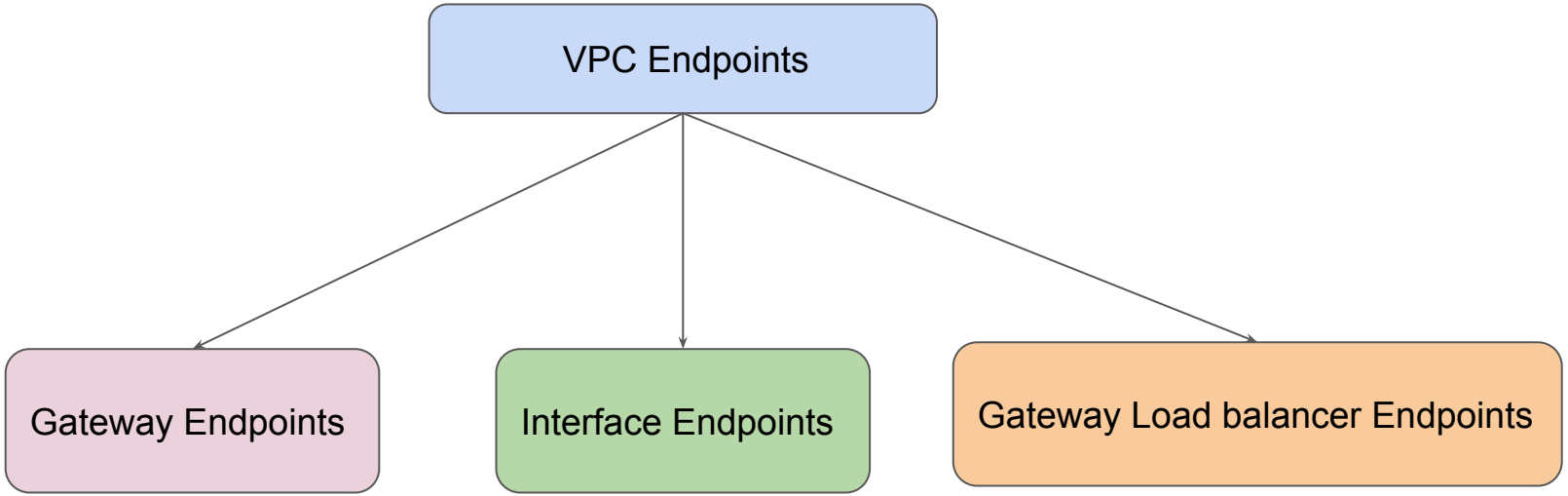
Traffic between your VPC and the other service does not leave the Amazon network.

Gateway VPC Endpoints

Understanding Types

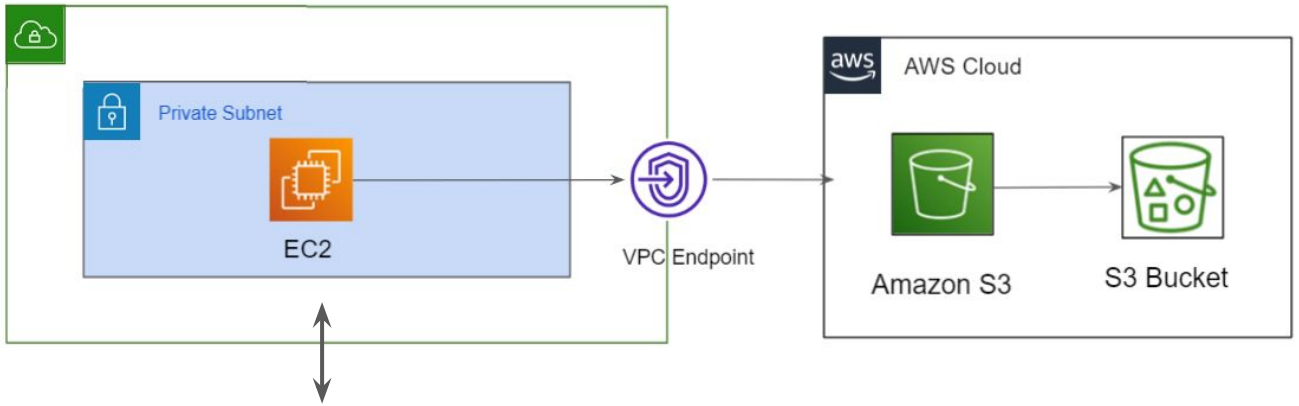
VPC Endpoints Type

There are three primary types of VPC Endpoints available.



Gateway VPC Endpoints

We specify the Gateway Endpoint as a route table target that is destined for supported AWS services.



Destination	Target
172.31.0.0/16	local
54.231.0.0/17 ⟷	pl-1a2b3c4d	vpce-11bb22cc

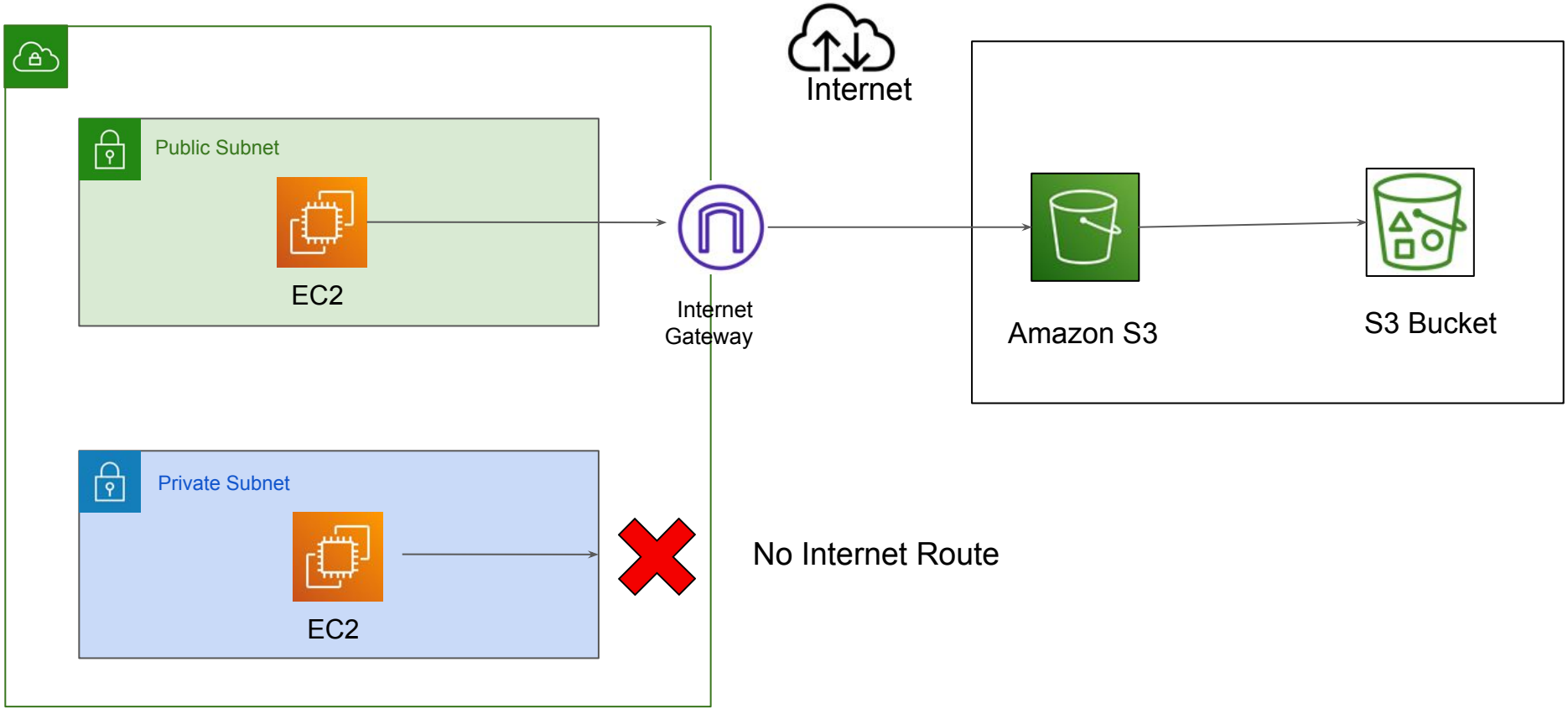
Supported Services

A gateway endpoint is for the following supported AWS services:

1. Amazon S3

2. DynamoDB

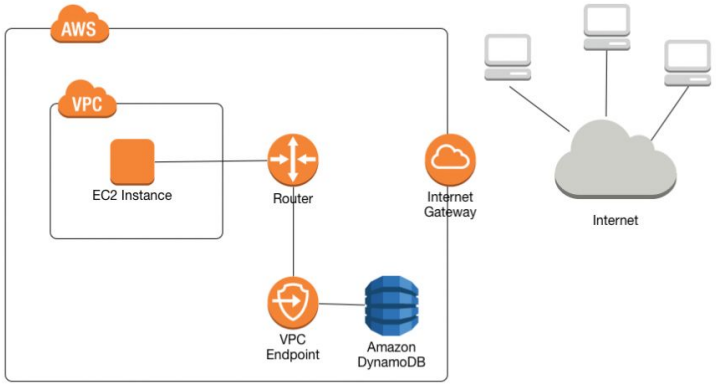
Today's Architecture



Downsides of Gateway Endpoints - 1

In Gateway endpoints approach, the VPC endpoint was created outside your VPC and traffic was routed via route table.

Thus, it is not possible to use it directly from VPN's or Direct connects and variou others.



Downsides of Gateway Endpoints - 2

Endpoints are supported within the same Region only. You cannot create an endpoint between a VPC and a service in a different Region.

Endpoints support IPv4 traffic only.

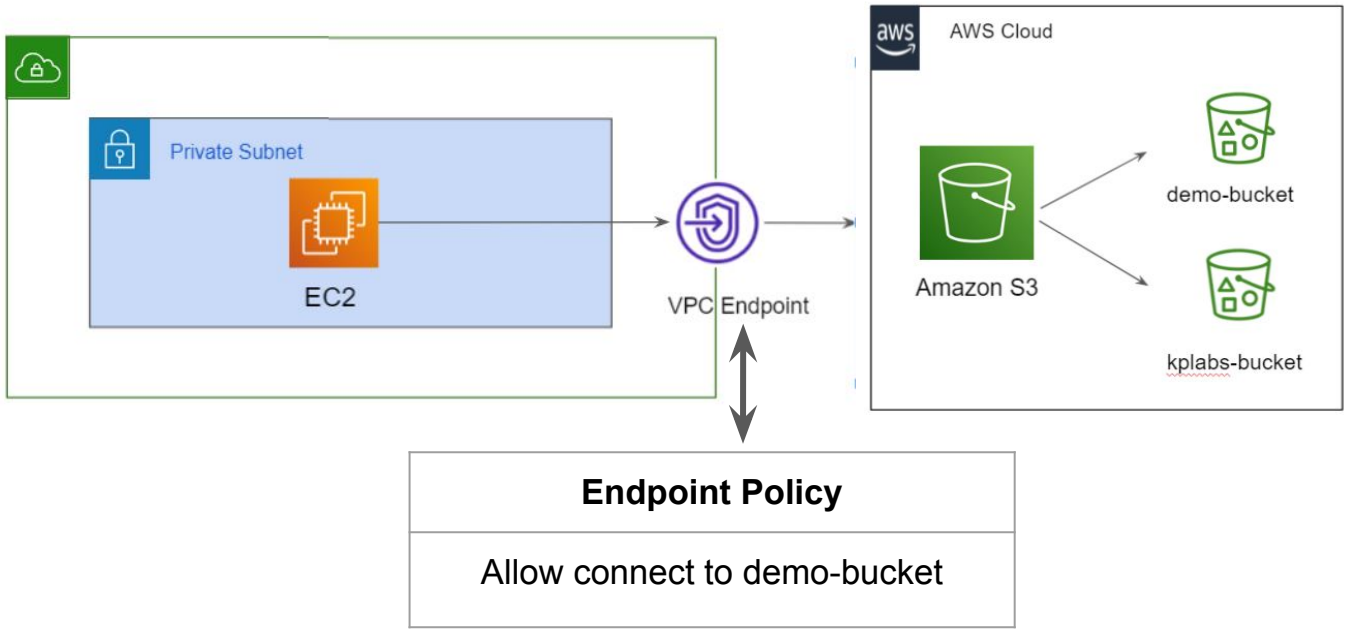
You must turn on DNS resolution in your VPC, or if you're using your own DNS server, ensure that DNS requests to the required service (such as Amazon S3) are resolved correctly to the IP addresses maintained by AWS.

VPC Endpoint Policies

Endpoint Based Access Control

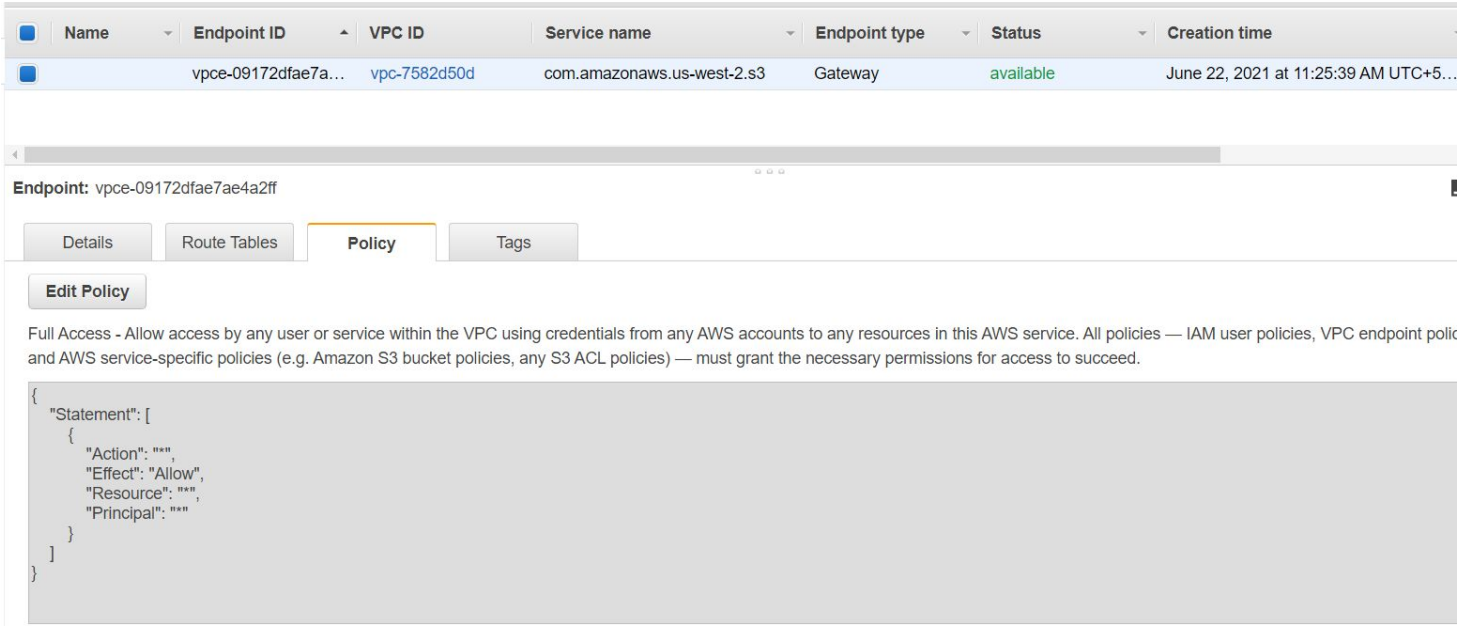
Overview of VPC Endpoint Policies

When you create a gateway endpoint, you can attach an endpoint policy to it that controls access to the service to which you are connecting.

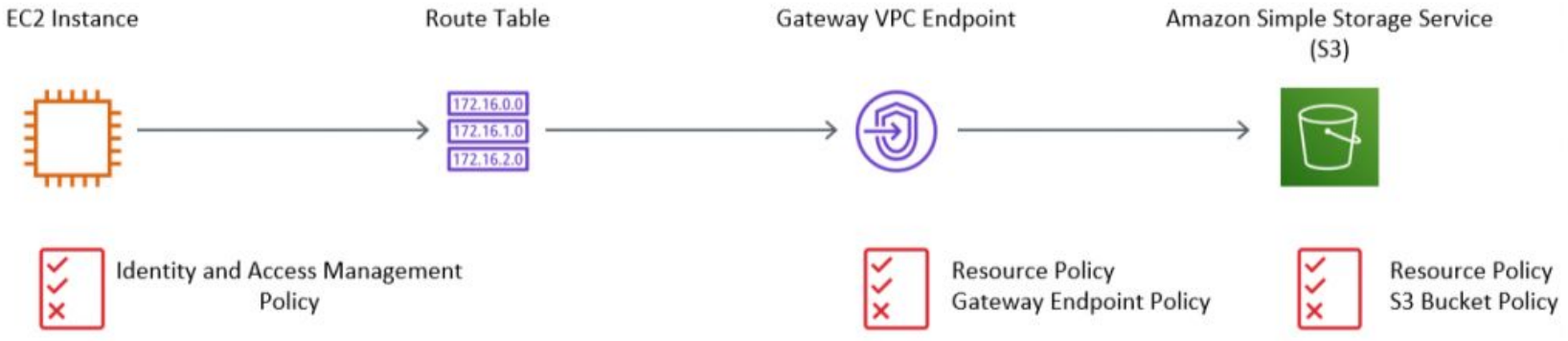


Default Policy

The default VPC Endpoint policy allows all the operations



Policy Decision



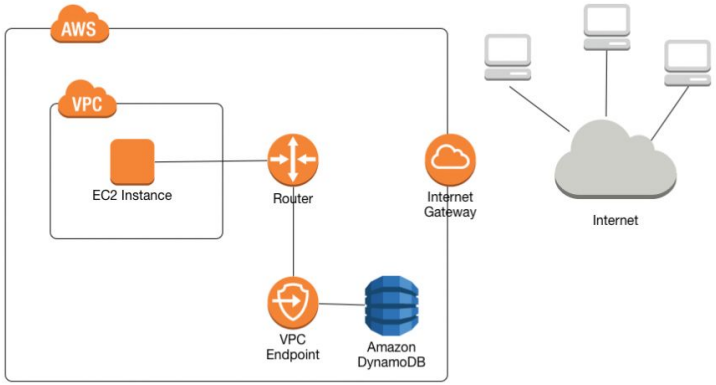
Interface Endpoints

New Generation Endpoint

Downsides of Gateway Endpoints - 1

In Gateway endpoints approach, the VPC endpoint was created outside your VPC and traffic was routed via route table.

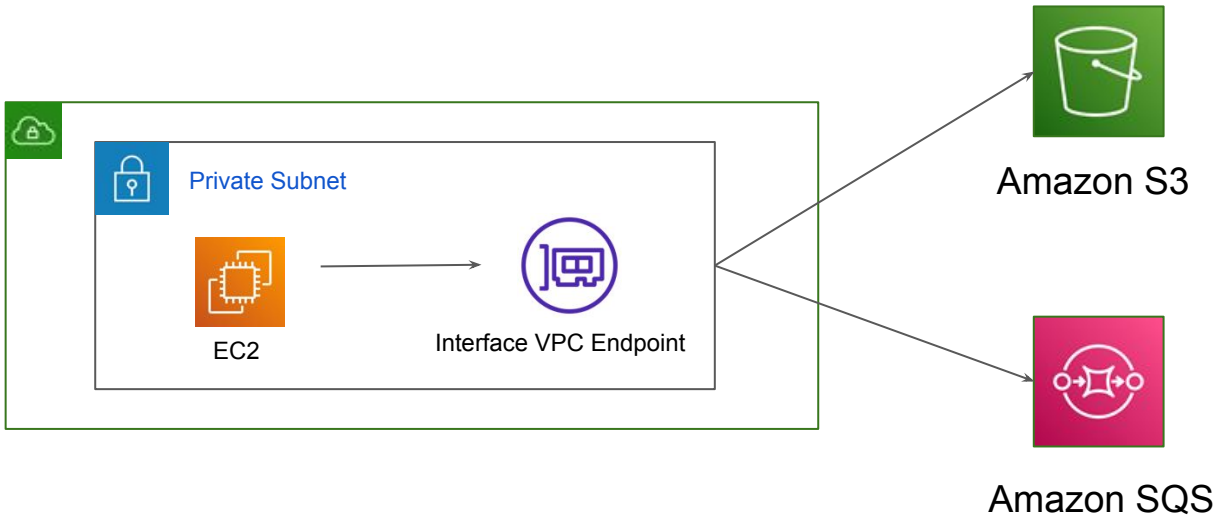
Thus, it is not possible to use it directly from VPN's or Direct connects and variou others.



Interface Endpoints

An interface endpoint is an elastic network interface with a private IP address from the IP address range of your subnet.

It serves as an entry point for traffic destined to a supported AWS service or a VPC endpoint service.



Benefits of Interface Endpoint

1. Interface endpoints enable the use of security groups to restrict access to the endpoint.

2. VPN's and Direct Connect based connections are supported.

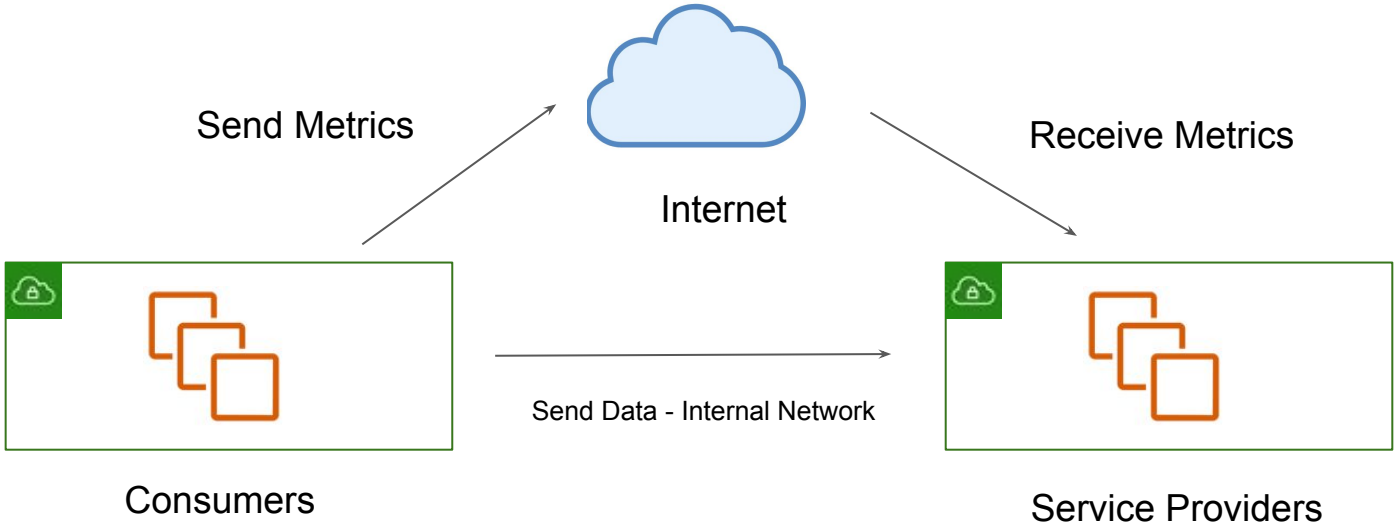
3. Interface endpoints supports lot of services unlike Gateway endpoints.

VPC Endpoint Services

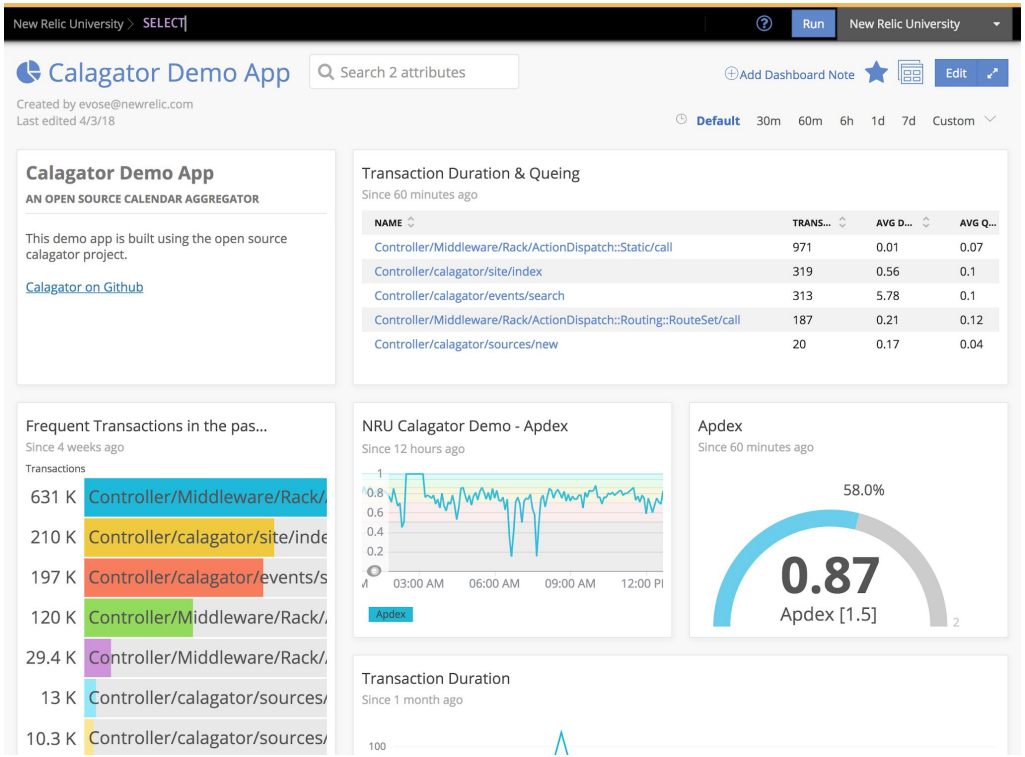
More Use-Cases Supported

Sample Use-Case

There are many service providers like DataDog, New Relic for which we need to upload our server/application metrics through Internet.

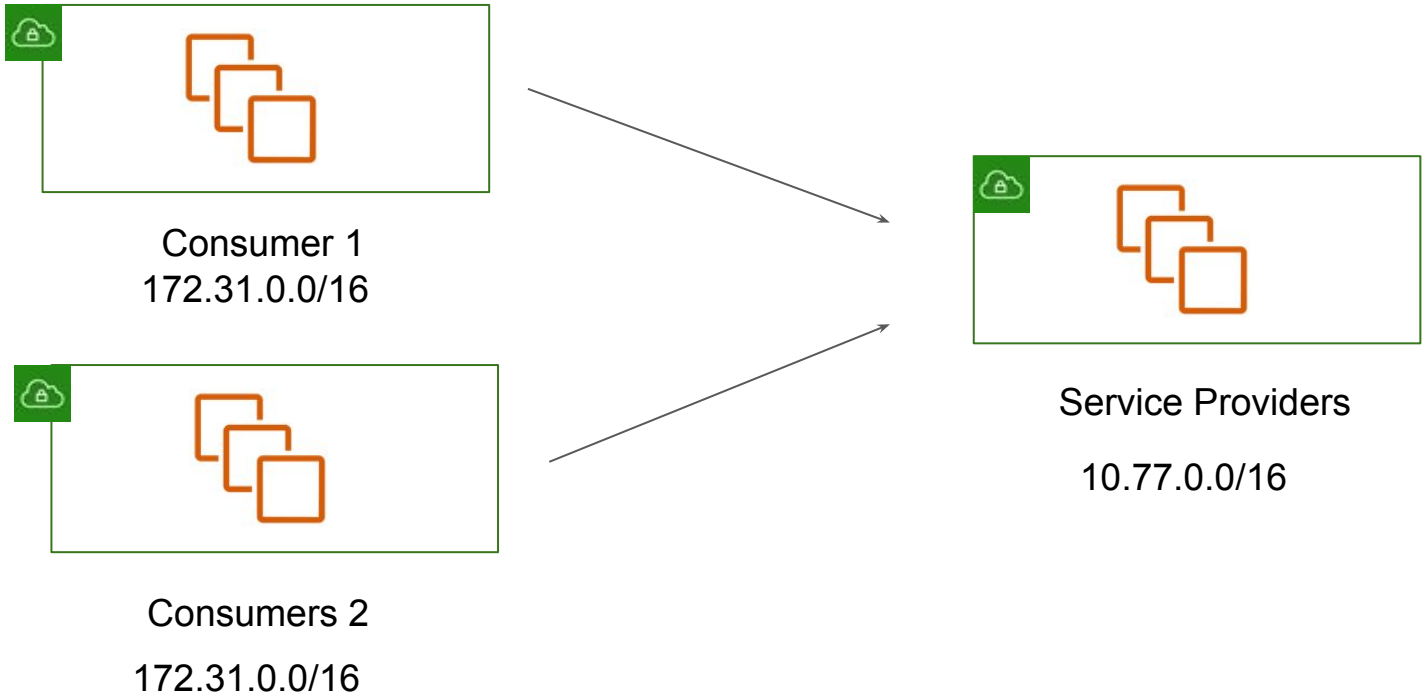


Dashboards using Metrics Collected



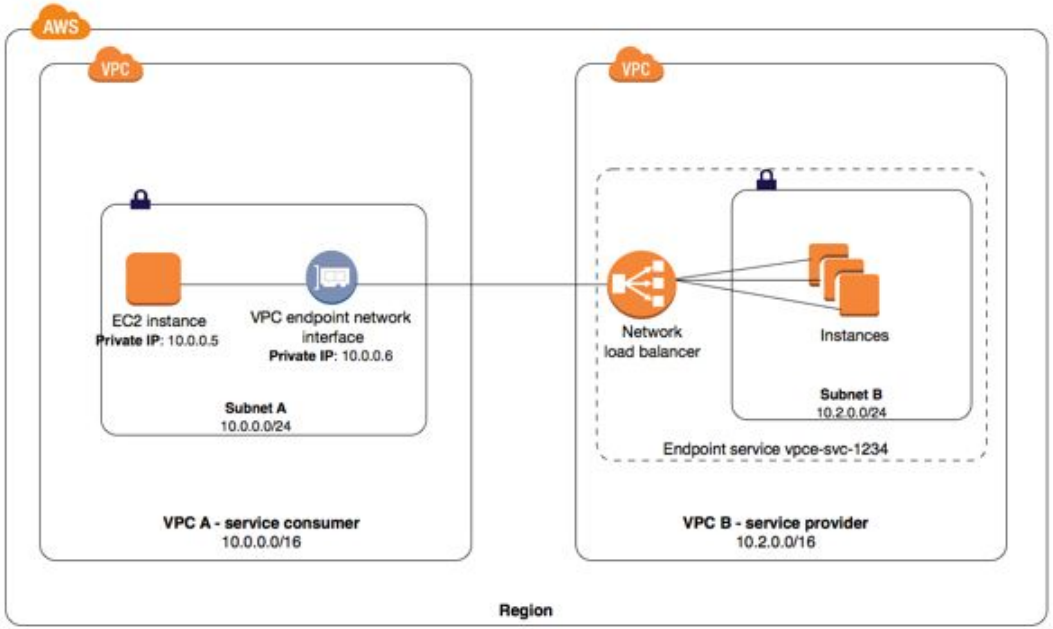
Possible Approach - VPC Peering

VPC Peering Approach will have multiple challenges related to CIDR overlap between clients.



Service VPC Endpoints

You can create your own application in your VPC and configure it as an AWS PrivateLink-powered service (referred to as an endpoint service)



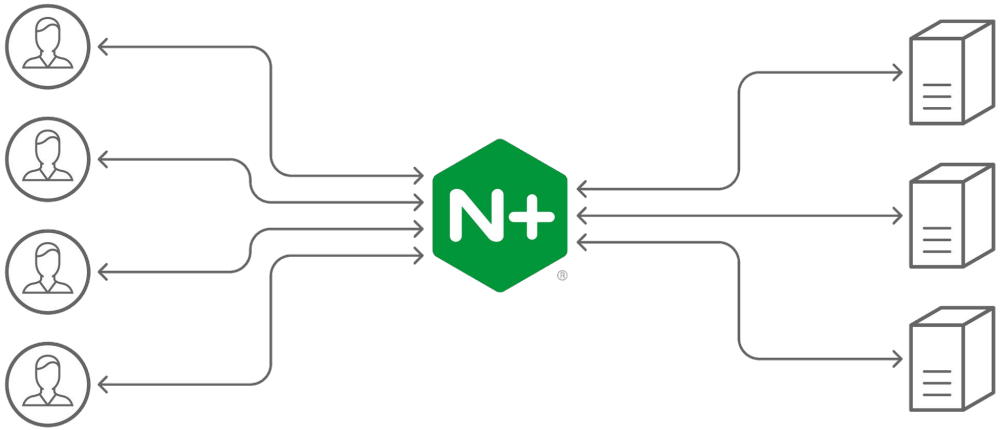
Load Balancing in AWS

Let's Load Balance Traffic in AWS

Basics of Load Balancing

There are multiple software and hardware based load balancing solutions available.

Some of the popular ones include Nginx, HA Proxy and others.



Challenges with Maintaining Load Balancing Solution

If you are using a load balancing solution, various responsibilities falls to customer.

Some of these include:

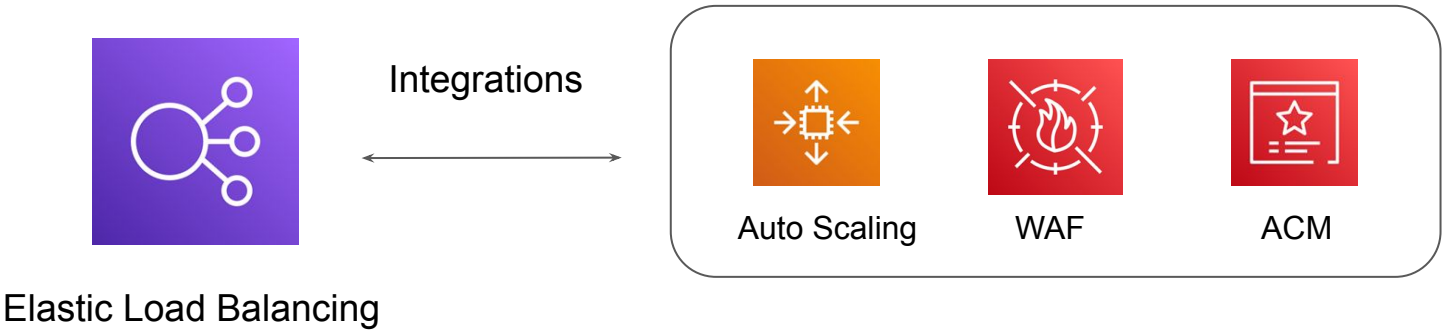
1. High-Availability of Load Balancers.
2. Security.
3. Performance.

Basics of Elastic Load Balancing Service

AWS offers managed load balancing solutions for wide variety of use-cases.

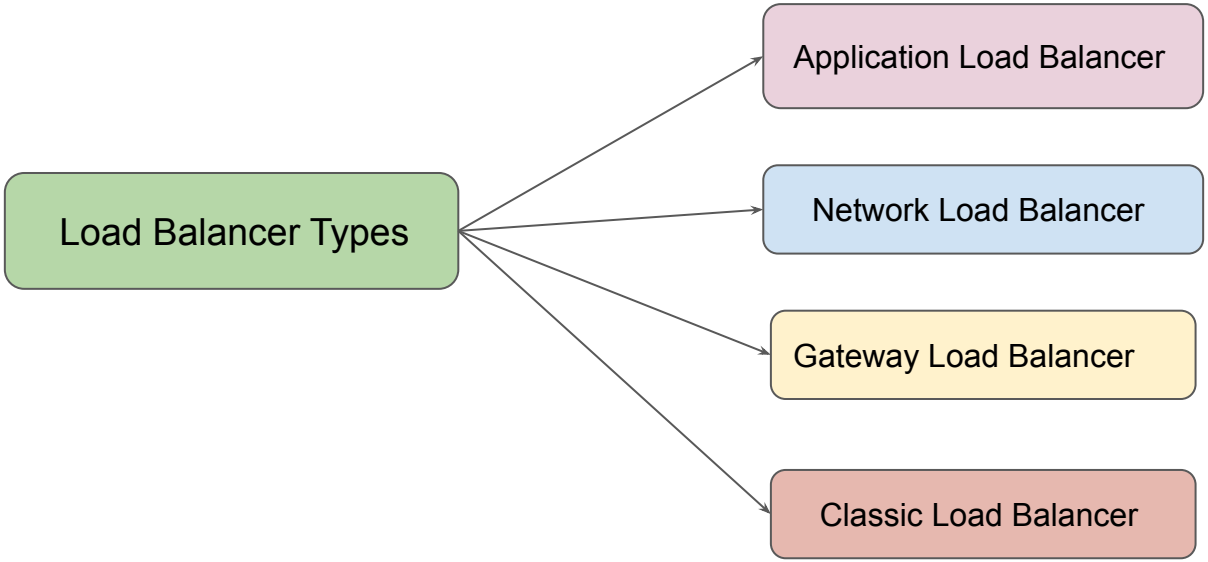
These solutions are offered under the Elastic Load Balancing feature.

Tight integration with multiple AWS Services.



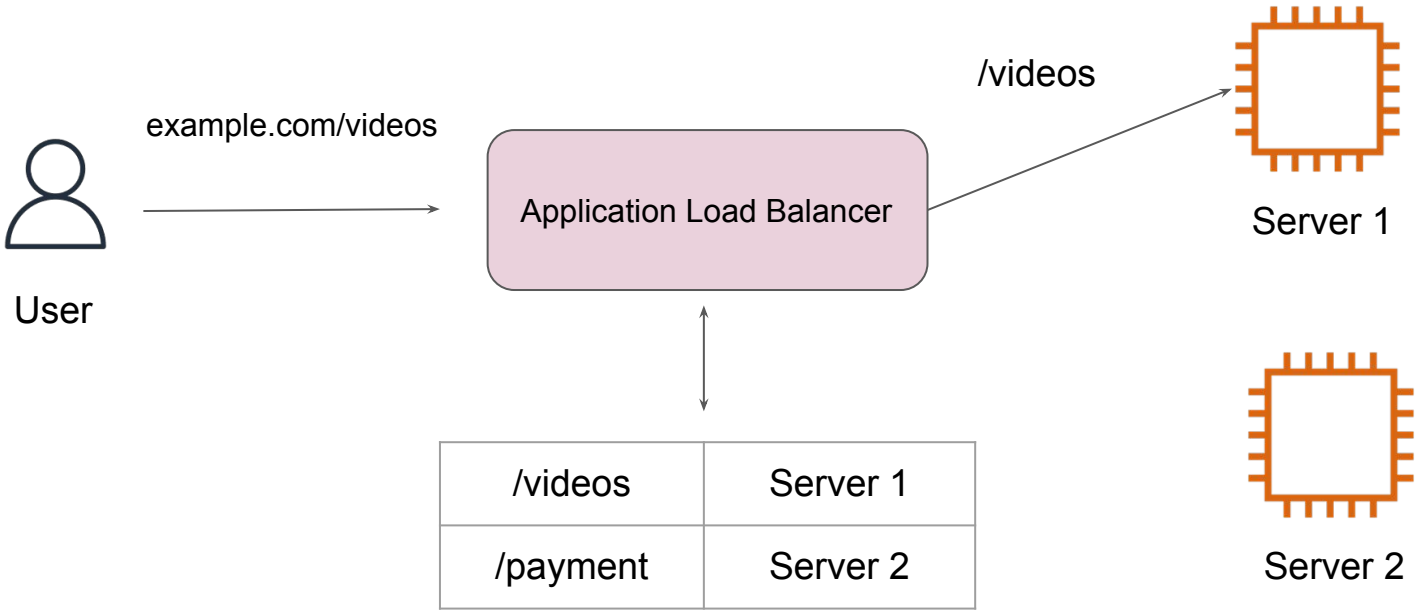
Types of Load Balancers

There are 4 primary type of Load Balancer offerings available.



Application Load Balancers

An Application Load Balancer makes routing decisions at the application layer (HTTP/HTTPS)

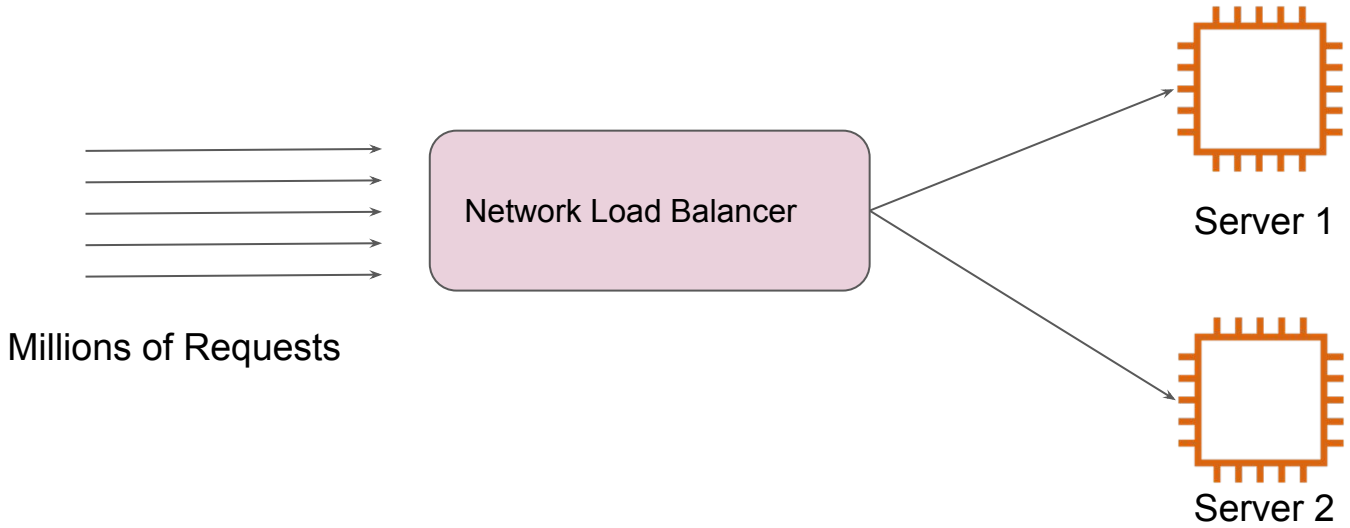


Network Load Balancers

A Network Load Balancer makes routing decisions at the transport layer (TCP/UDP/SSL).

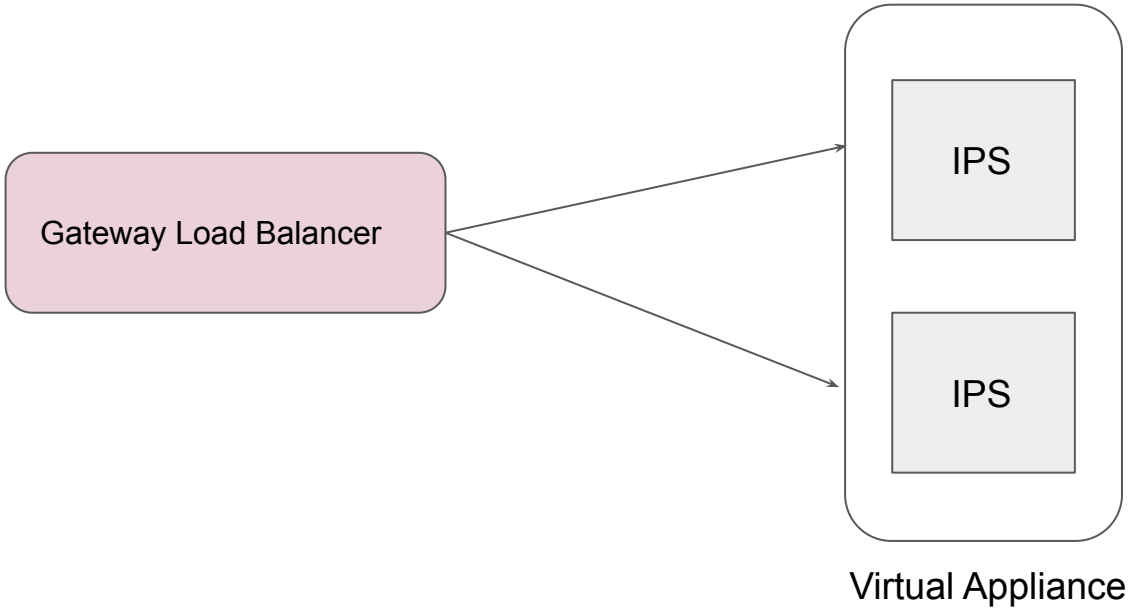
It can handle millions of requests per second.

Not all of the applications work on HTTP/HTTPS protocol.



Gateway Load Balancers

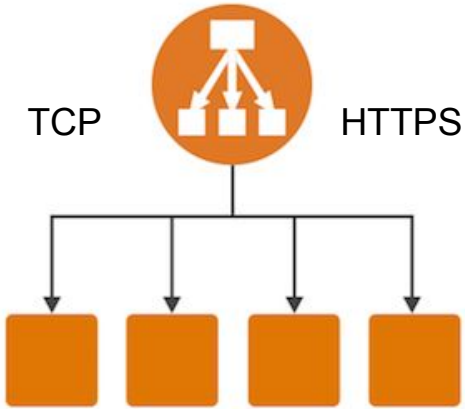
Gateway Load Balancers allow you to deploy, scale, and manage virtual appliances, such as firewalls, intrusion detection and prevention systems, and deep packet inspection systems



Classic Load Balancers

A Classic Load Balancer makes routing decisions at either the transport layer (TCP/SSL) or the application layer (HTTP/HTTPS).

Previous Generation Load Balancer and not recommended.



Summary Slide

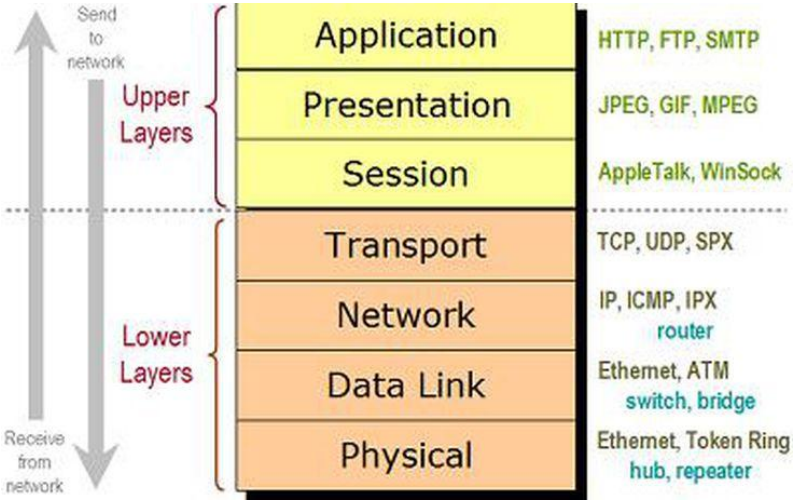
Load Balancer	Important Notes
Application Load Balancer	Use when you have websites/applications at L7 (HTTP/HTTPS)
Network Load Balancers	TCP and UDP based applications. Requirement to handle millions of requests per second. Ultra high performance.
Gateway Load Balancer	Use when you have virtual appliances: IDS/IPS Firewalls

OSI Model & Load Balancers

Revising Networking

Basics of OSI Model

The Open Systems Interconnection (OSI) model describes seven layers that computer systems use to communicate over a network. It



Load Balancer & OSI Layers

Each load balancer operates at a specific layer.

You will only be able to perform operations on requests based on Layer the ELB supports.

Feature	Application Load Balancer	Network Load Balancer	Gateway Load Balancer	Classic Load Balancer
Load Balancer type	Layer 7	Layer 4	Layer 3 Gateway + Layer 4 Load Balancing	Layer 4/7

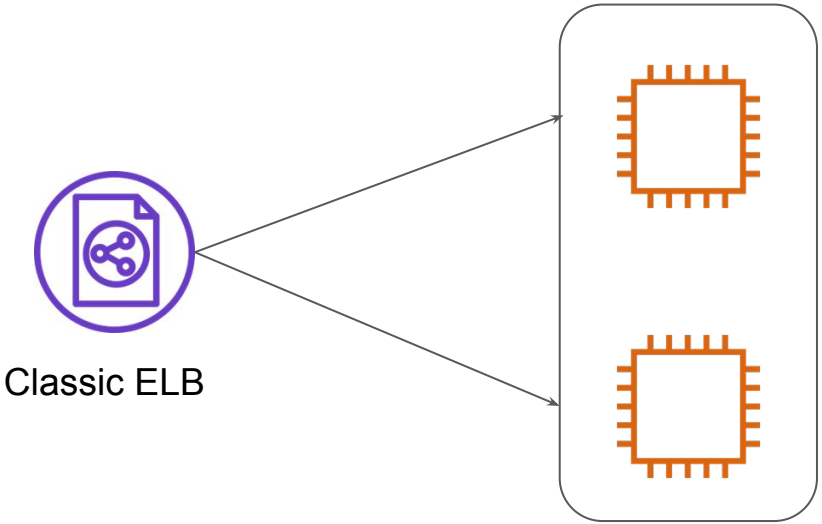
Classic Load Balancers

First generation Load Balancers

Understanding Classic Load Balancers

These are older generation of load balancers.

Provides basic set of features for HTTP, HTTPS, TCP and SSL protocols.



Limitation of Classic Load Balancers

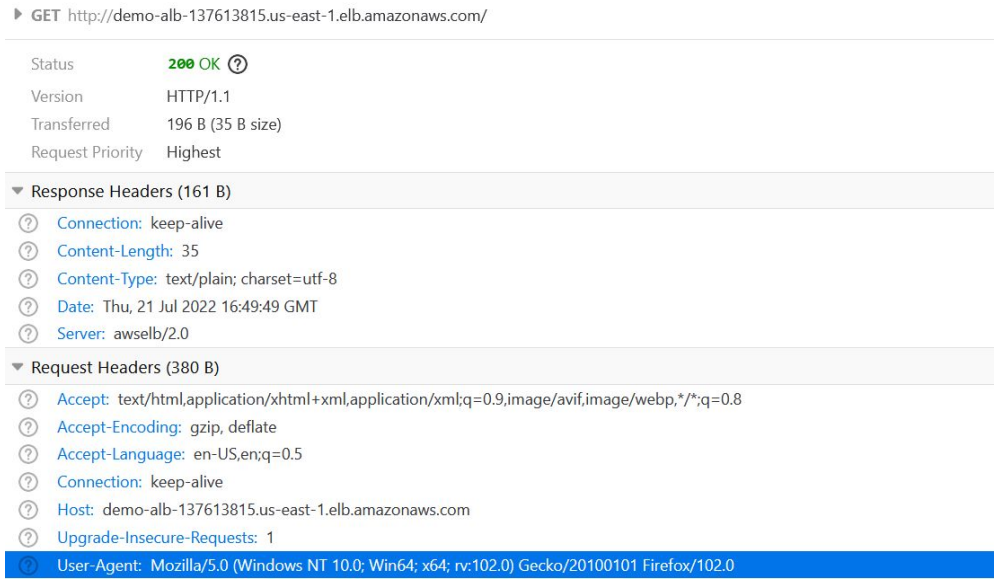
- Does not support native HTTP/2 protocol.
- IP address as targets are not supported.
- Path based routing is not supported. (eg: /images should go to server 1 & /php to server 02)
- Many Many more .....

Application Load Balancers

Next generation load balancers

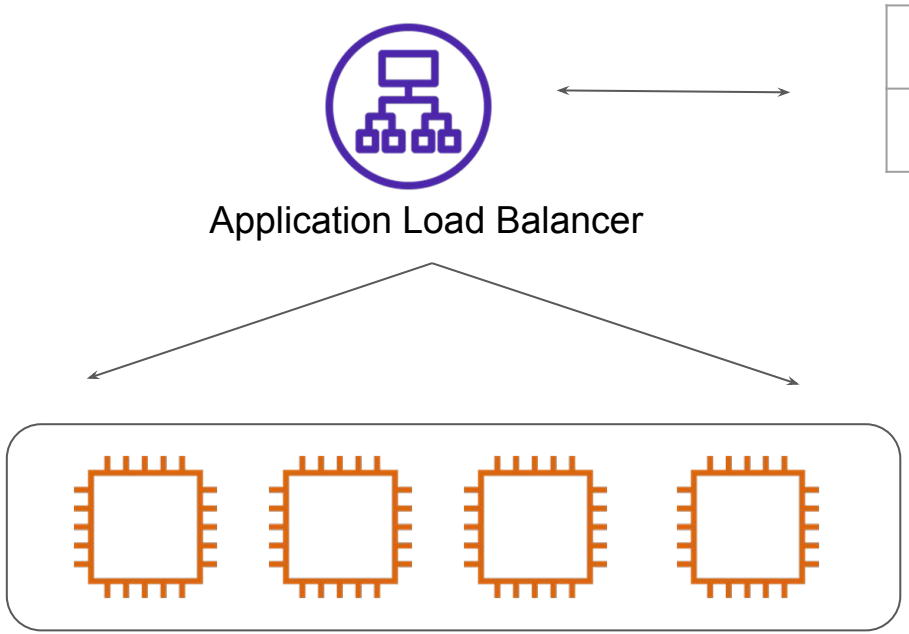
Basics of HTTP Headers

HTTP headers let the client and the server pass additional information with an HTTP request or response.



Understanding ALB

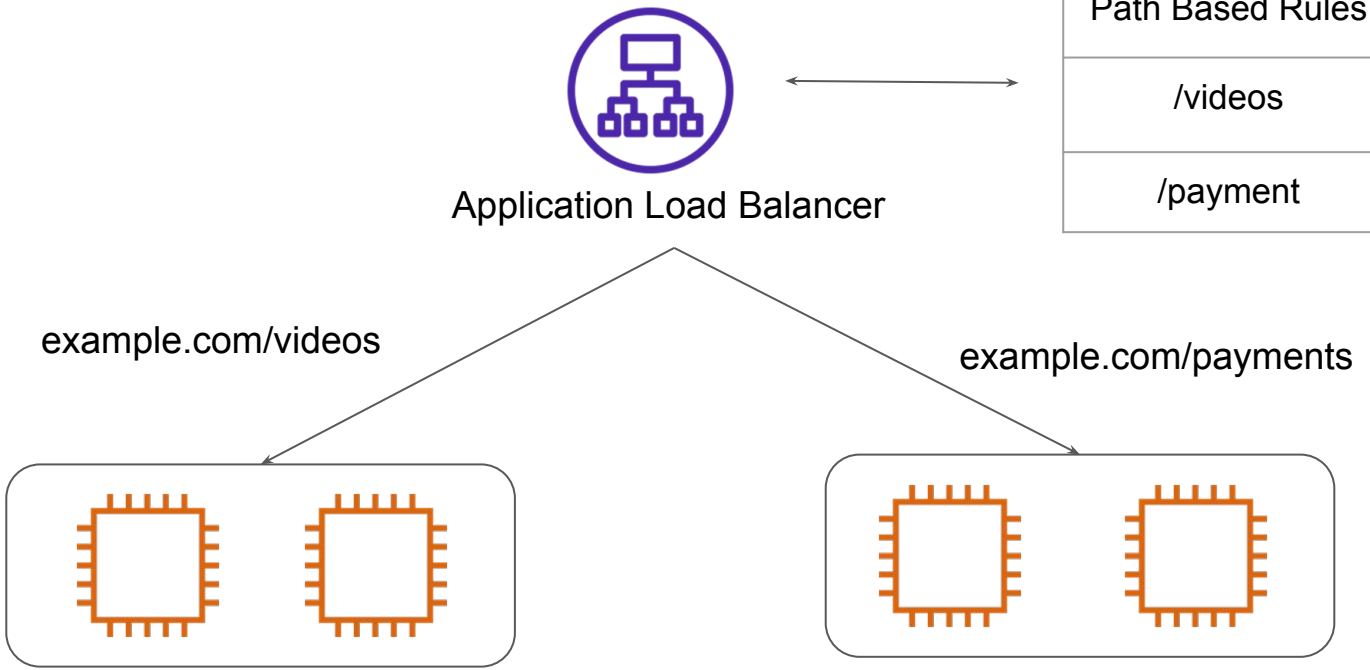
Application Load Balancer functions at Application layer and support both HTTP & HTTPS



Routing Based On
HTTP Headers

Path Based Routing

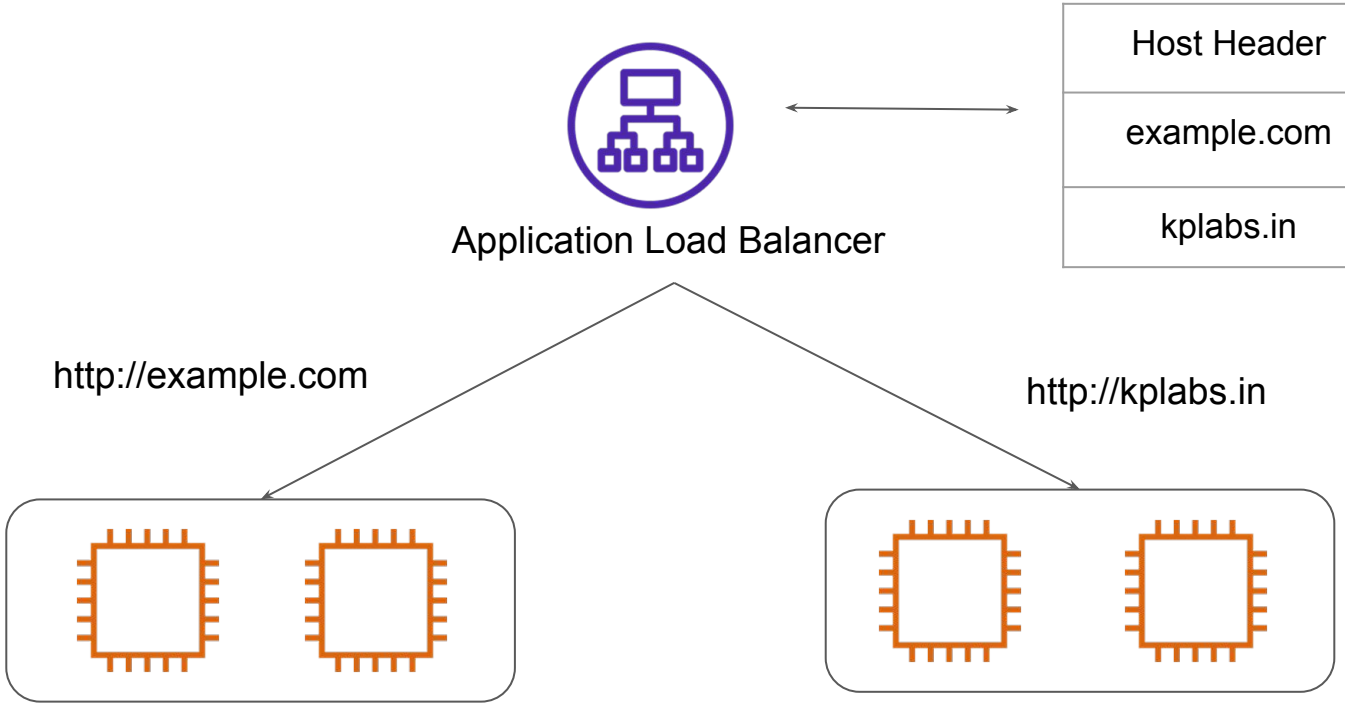
The request are routed based on the URI path.



Path Based Rules	Destination
/videos	Server 1 and 2
/payment	Server 3 and 4

Routing Using Host Headers

The request are routed based on the Host Header



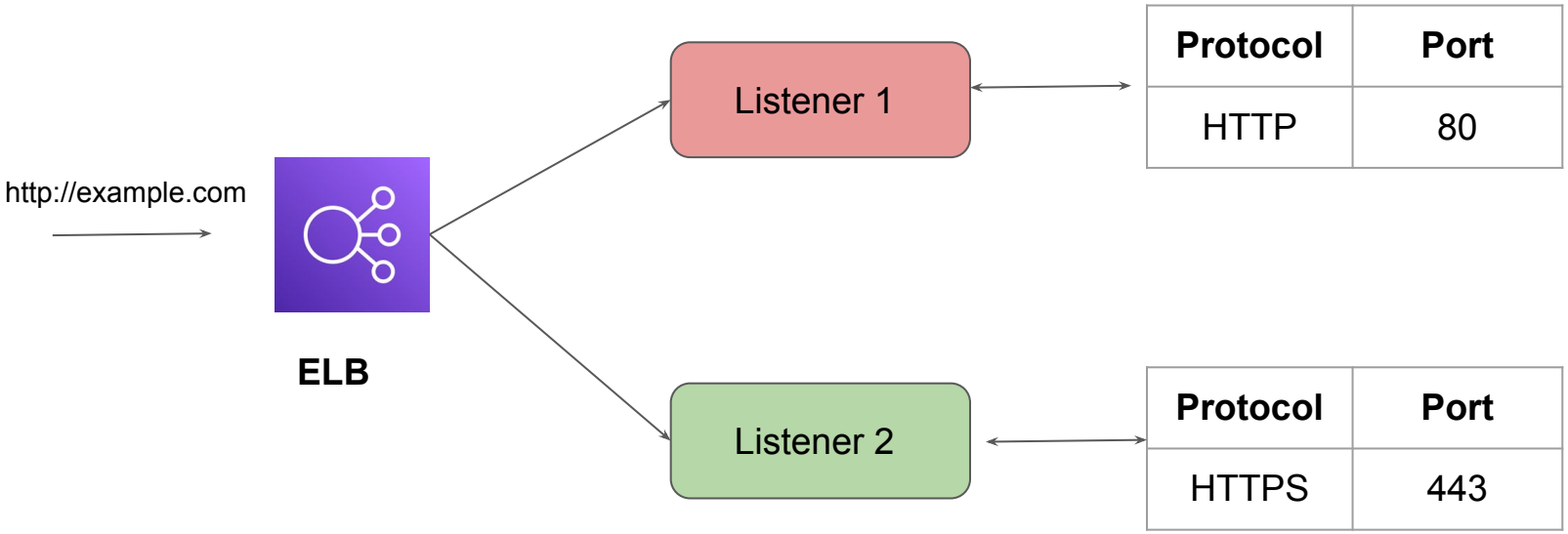
Host Header	Destination
example.com	Server 1 and 2
kplabs.in	Server 3 and 4

Listener & Target Groups

Next generation load balancers

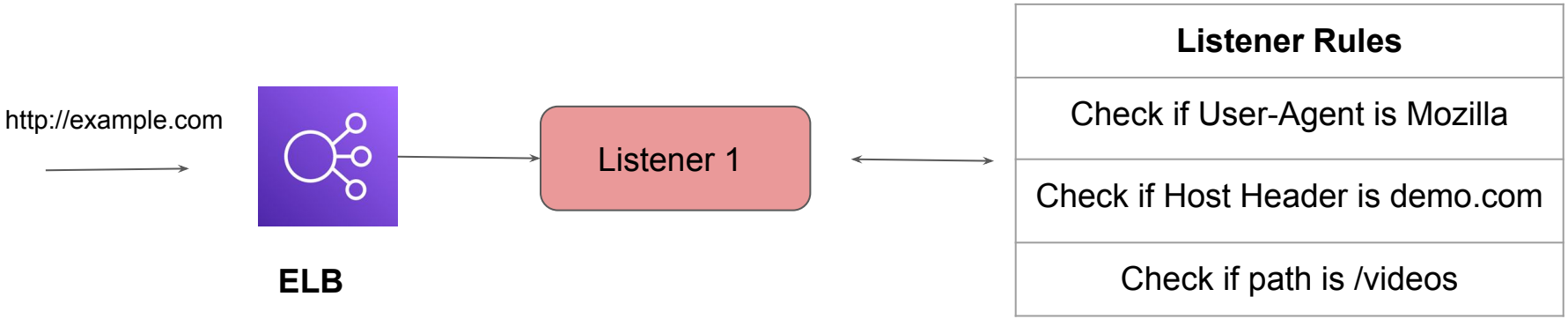
Understanding Listeners

A listener is a process that checks for connection requests, using the protocol and port that you configure.



Listener Rules

Each listener has a rule based on which an action is taken based on a request.



Listener Rule Actions

If a request matches a specific rule, what action you want to perform on that request is determined in the Rule Actions.



demo-alb | **HTTP:80** (4 rules)

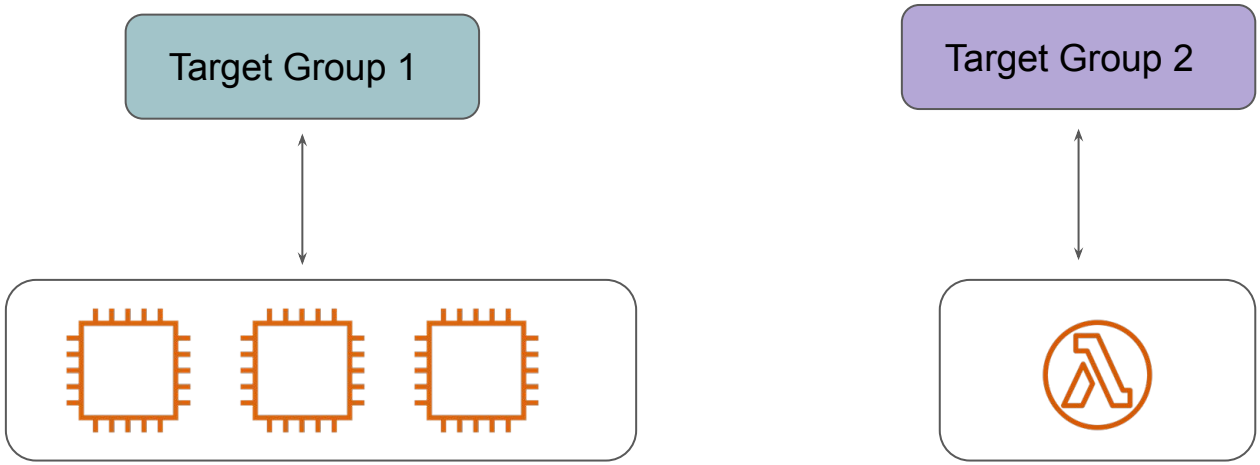
▸ Rule limits for condition values, wildcards, and total rules.

1	arn...93720	**IF** ✔ Http header User-Agent is *curl*	**THEN** **Return fixed response** 200 Content-Type: text/plain Response body: Hi curl! (less...)
2	arn...c9bc6	**IF** ✔ Http header User-Agent is *Mozilla*	**THEN** **Return fixed response** 200 Content-Type: text/plain Response body: Hey Mozilla! You have great addons! (less...)
3	arn...fdb85	**IF** ✔ Http header User-Agent is *wget*	**THEN** **Return fixed response** 200 Content-Type: text/plain Response body: Hi There wget! I detected you. (less...)

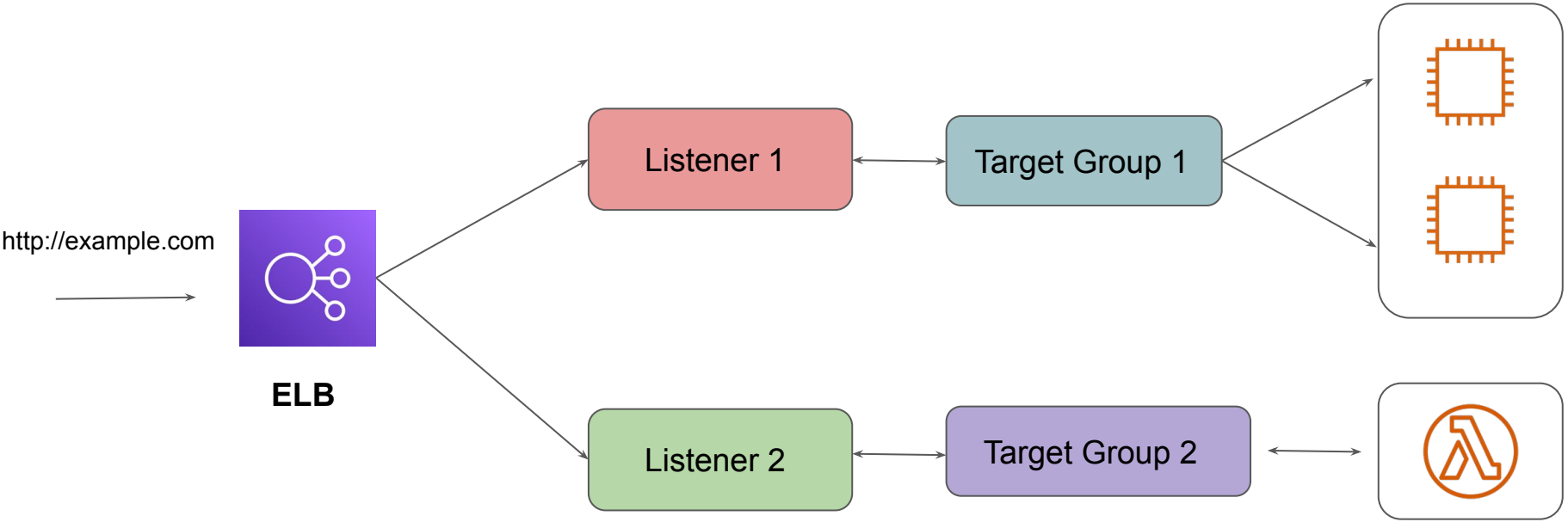
Understanding Target Groups

Target group is used to route requests to one or more registered targets.

These targets can be EC2 instances, Lambda Functions, and others.



Overall Workflow



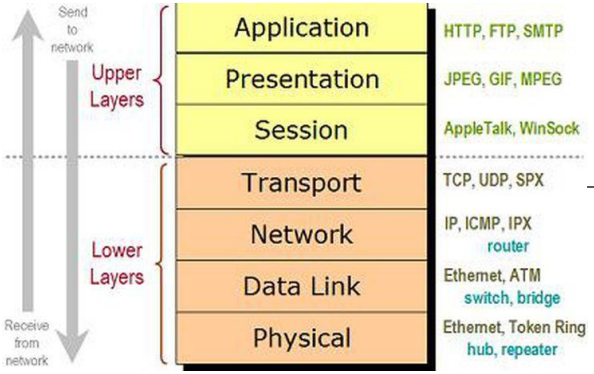
Network Load Balancers

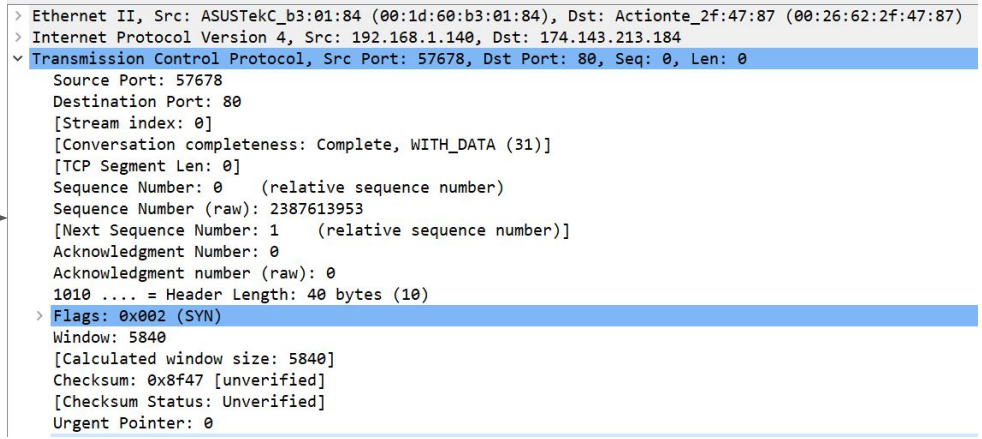
Next generation load balancers

Understanding NLB

Network Load Balancer works on the fourth layer of the OSI model.

It can handle millions of requests per second.





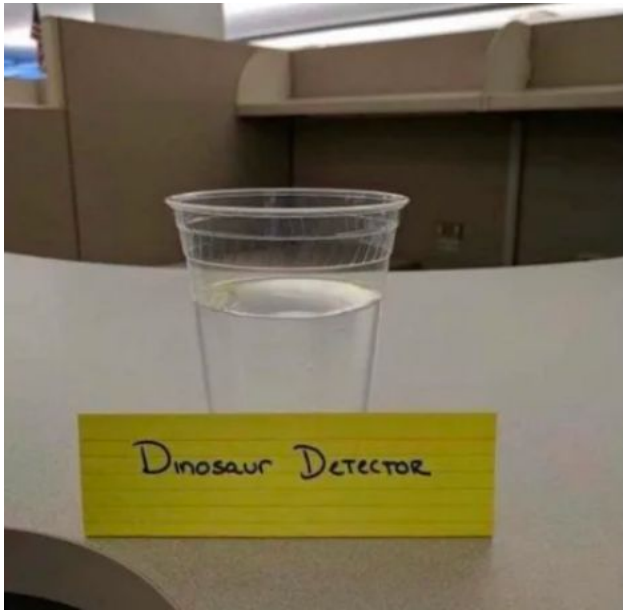
Basic Working

NLB primarily selects a target using a flow hash algorithm based on:

Protocol, Source IP address, Source port, Destination IP address, Destination port, and TCP sequence number.

Each individual TCP connection is routed to a single target for the life of the connection.

Relax and Have a Meme Before Proceeding



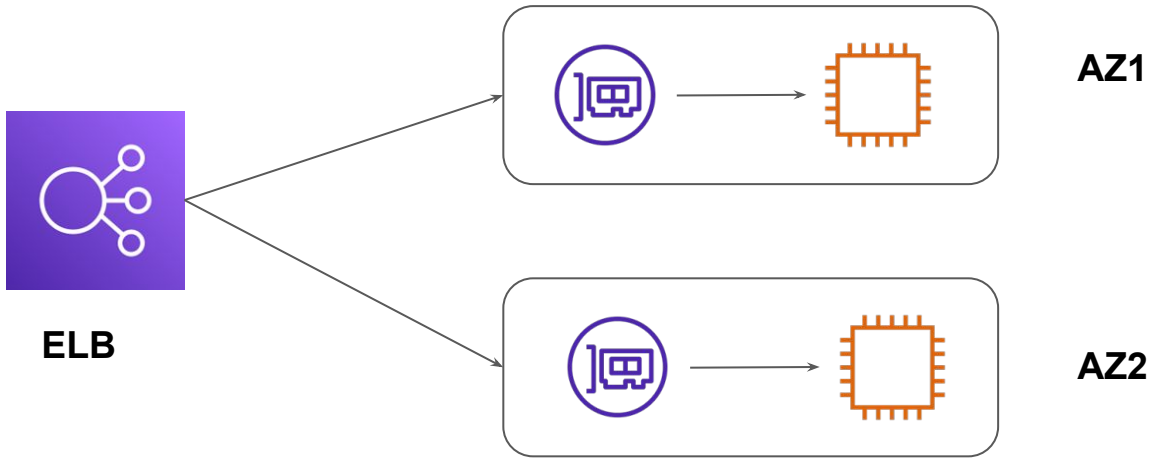
Availability Zones and ELB nodes

ELB Interfaces

Availability Zones and ELB nodes

When you enable an Availability Zone for your load balancer, Elastic Load Balancing creates a load balancer node in the Availability Zone.

If you register targets in an Availability Zone but do not enable the Availability Zone, these registered targets do not receive traffic.



Recommendations

With an Application Load Balancer, it is a requirement that you enable at least two or more Availability Zones. If one Availability Zone becomes unavailable or has no healthy targets, the load balancer can route traffic to the healthy targets in another Availability Zone.

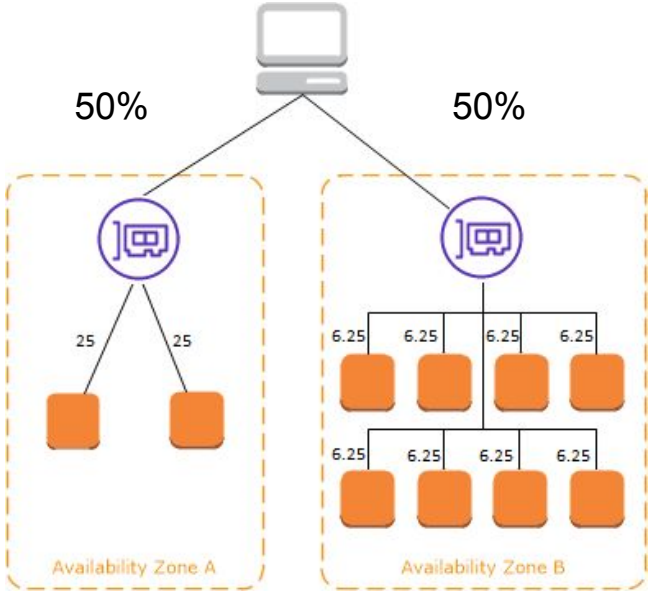
After you disable an Availability Zone, the targets in that Availability Zone remain registered with the load balancer. However, even though they remain registered, the load balancer does not route traffic to them.

Cross Zone Load Balancing

Interesting Learning

Understanding the Challenge

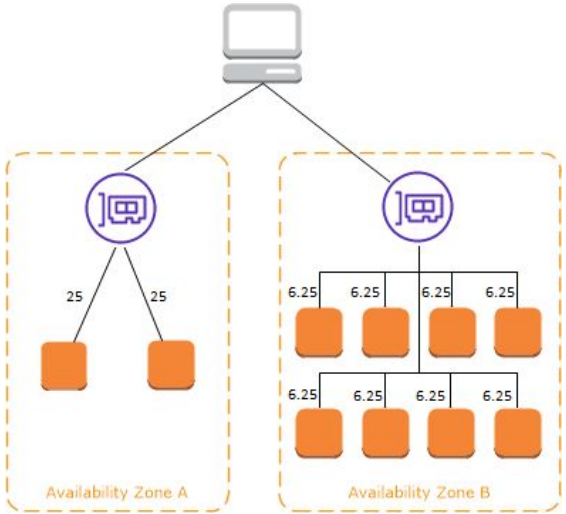
If Cross Zone Load Balancing is disabled, each load balancer node distributes traffic only across the registered targets in its Availability Zone.



Cross Zone Load Balancing Disabled

Each of the two targets in Availability Zone A receives 25% of the traffic.

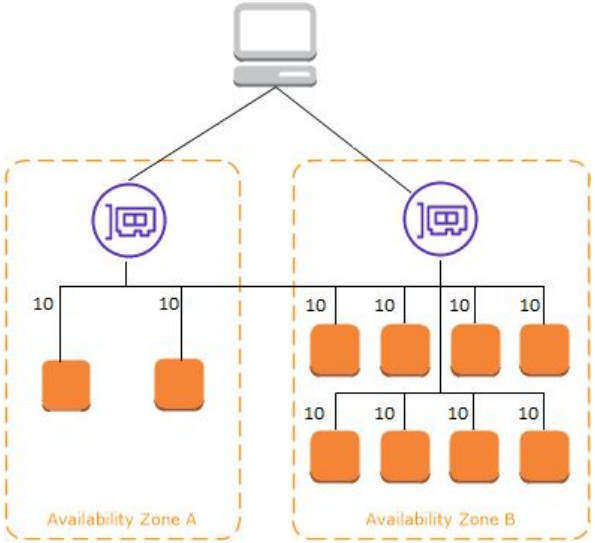
Each of the eight targets in Availability Zone B receives 6.25% of the traffic.



Cross Zone Load Balancing

When cross-zone load balancing is enabled, each load balancer node distributes traffic across the registered targets in all enabled Availability Zones.

If cross-zone load balancing is enabled, each of the 10 targets receives 10% of the traffic.



Important Pointers

With Application Load Balancers, cross-zone load balancing is always enabled.

With Network Load Balancers and Gateway Load Balancers, cross-zone load balancing is disabled by default. After you create the load balancer, you can enable or disable cross-zone load balancing at any time.

ELB Access Logs

Who is Visiting Us?

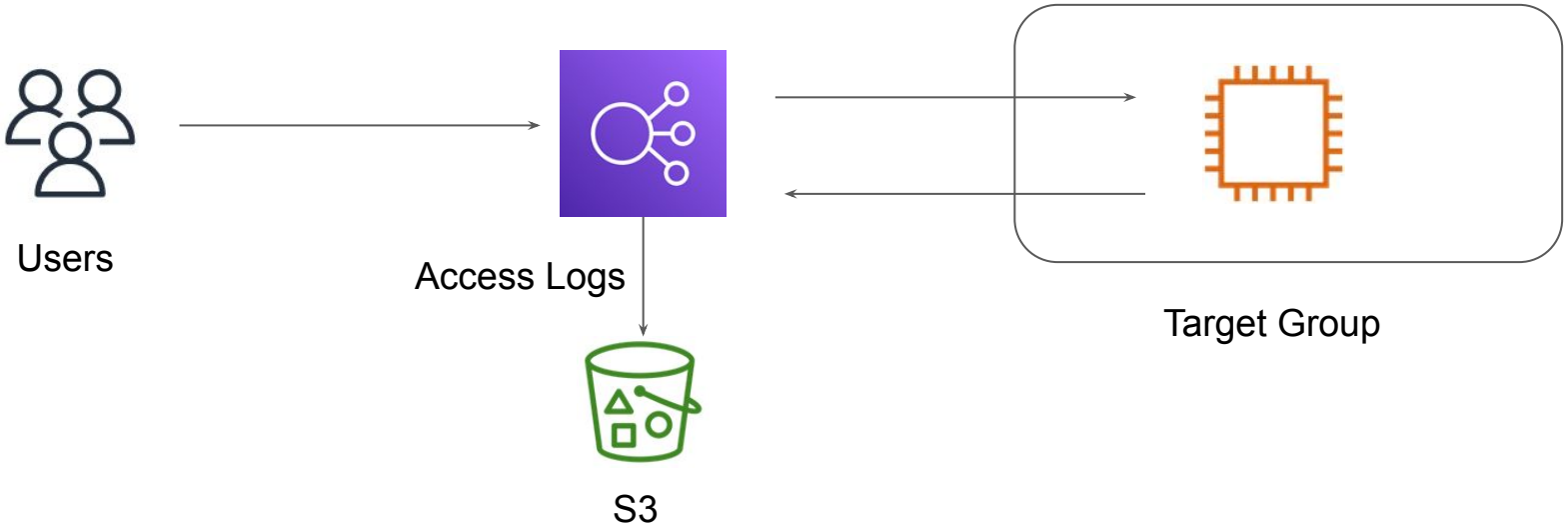
Overview of Access Logs

An access log is a list of all the requests for individual files that people have requested from a Web site



ELB Access Logs

Elastic Load Balancing provides access logs that capture detailed information about requests sent to your load balancer.



Important Pointers for Access Logs - Part 1

Access logging is an optional feature of Elastic Load Balancing that is disabled by default

Elastic Load Balancing logs requests on a best-effort basis. AWS recommend that you use access logs to understand the nature of the requests, not as a complete accounting of all requests.

Important Pointers for Access Logs - Part 2

The bucket and your load balancer must be in the same Region.

Bucket Policy should be designed so that AWS Account must be able to write to your bucket.

Elastic Load Balancing publishes a log file for each load balancer node every 5 minutes.

Relax and Have a Meme Before Proceeding



Dualstack IP Address Type for ELBs

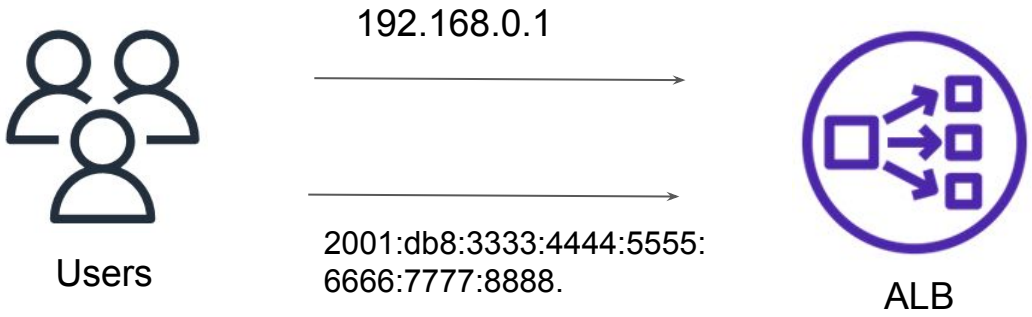
Enable IPv6 for ELBs

IP Address Type Support

ELB Supports two address types:

i) IPv4

ii) Dualstack (includes both IPv4 and IPv6 addresses)



Important Pointer

To use IPv6 addresses, the virtual private cloud (VPC) where you launch your ELB must have subnets with associated IPv6 CIDR blocks

IPv6 addresses can be associated only with internet-facing Application Load Balancers and Network Load Balancers.

Internal Application Load Balancers, Classic Load Balancers, and Network Load Balancers do not support IPv6 addresses.

Sticky Sessions

Direct Users to the Same Server

Understanding the Challenge

Generally the Load Balancers will distribute the traffic from the users to the backend servers via the round robin algorithm.

If subsequent requests are routed to different servers, your session state information is lost.



Importance of Sticky Session

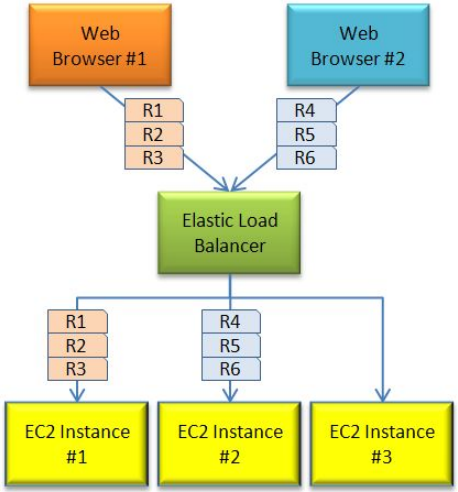
Sticky session refers to the feature of many commercial load balancing solution to route the requests for a particular session to the same physical machine that serviced the first request for that session.



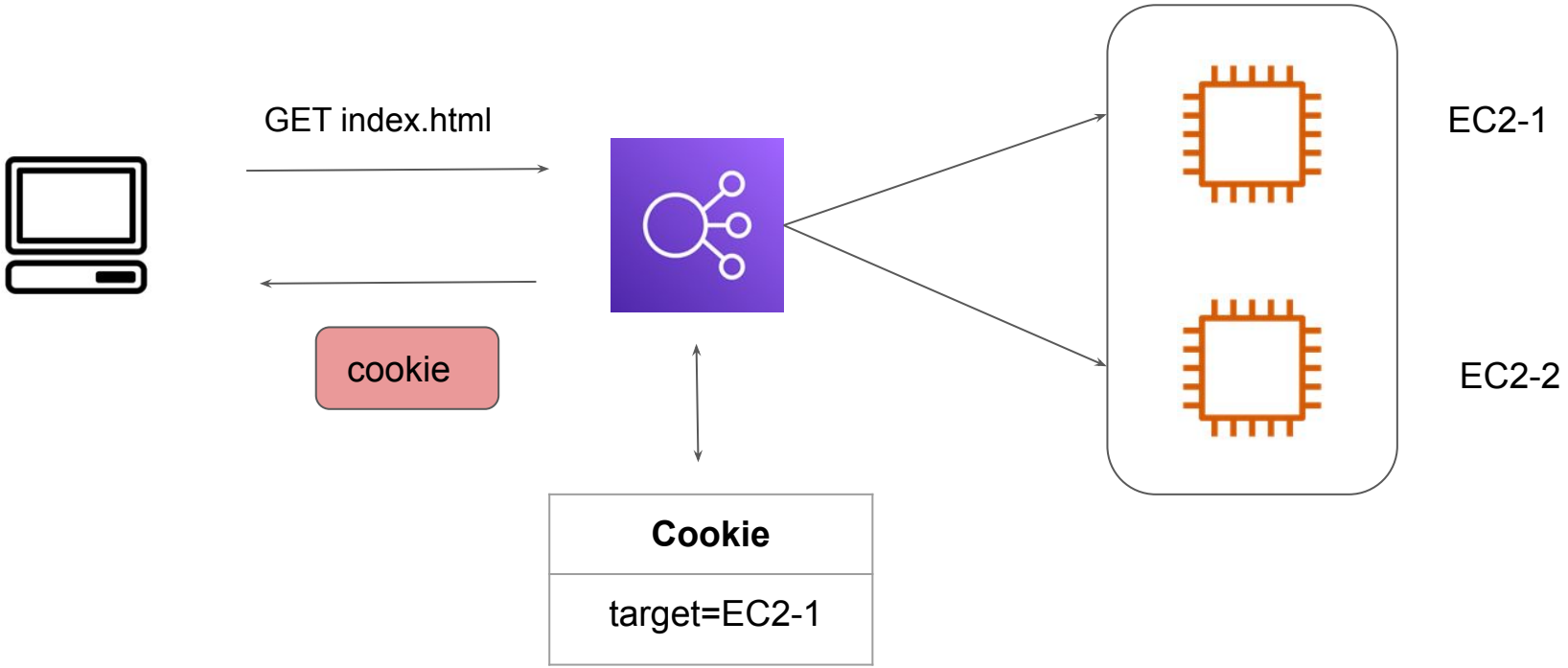
Sticky Session and AWS Load Balancers

By default, an Application Load Balancer routes each request independently to a registered target based on the chosen load-balancing algorithm.

However, you can use the sticky session feature to enable the load balancer to bind a user's session to a specific target



Overview of the Workflow



Advantages of Sticky Sessions

When sticky sessions are used, the servers do not need to exchange the sessions data which can minimize the data transfers.

Sticky Sessions can also allow better utilization of your RAM Cache that leads to better responsiveness

Disadvantage of Sticky Sessions

With sticky sessions, the overall balancing of traffic between servers can be affected.

A server can become overloaded if it accumulates too many sessions, or if specific sticky sessions require a high number of resources.

Duration Based Stickiness

When a load balancer first receives a request from a client, it routes the request to a target (based on the chosen algorithm), and generates a cookie named AWSALB.

It encodes information about the selected target, and includes the cookie in the response to the client

In subsequent requests, the client should include the AWSALB cookie. When the load balancer receives a request from a client that contains the cookie, it detects it and routes the request to the same target

Application Based Stickiness

The target is expected to set a custom application cookie.

When the ALB receives the custom application cookie from the target, it automatically generates a new encrypted application cookie to capture stickiness information.

The load balancer generated application cookie does not copy the attributes of the custom cookie set by the target. It has its own expiry of 7 days which is non-configurable

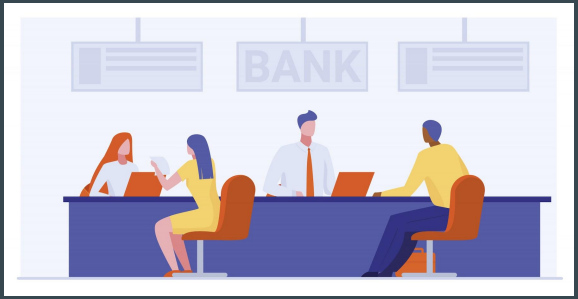
RDS Read Replicas



Use Case : Bank

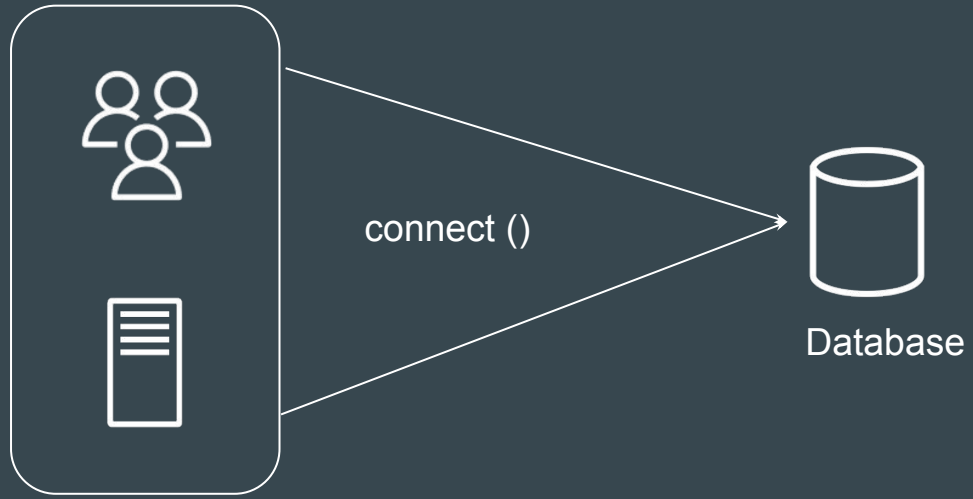
In bank, for different kind of work purpose, there are different people you might have to approach. For example :

- Cash Collector
- Cheque Counter
- Enquiry Counter



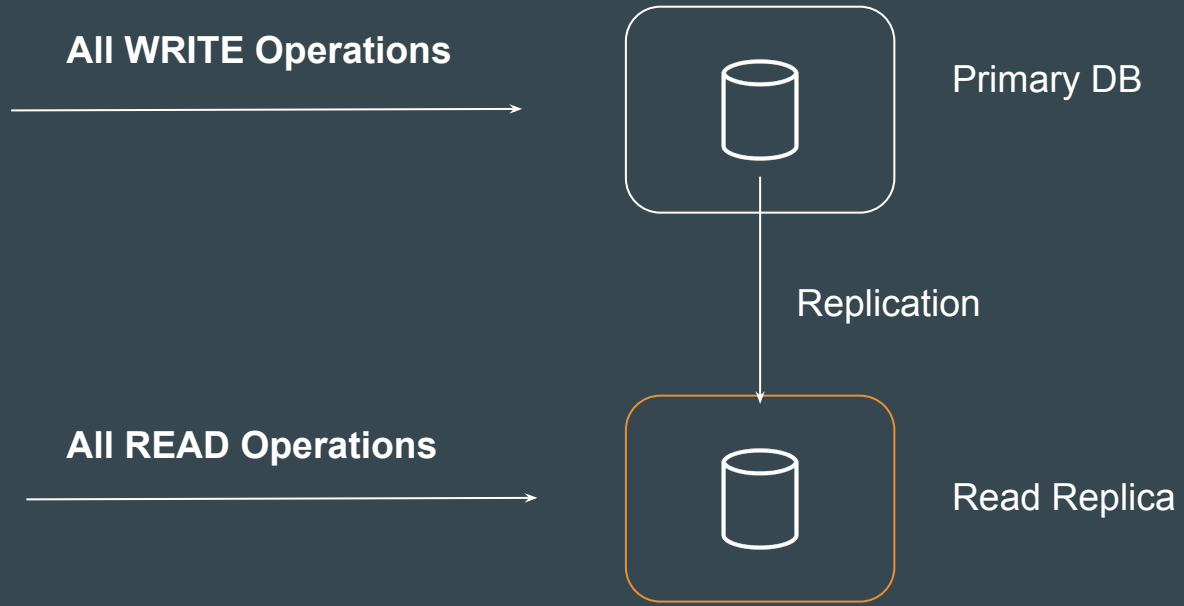
Database Way

Using a single database for all kind of activity will increase the database load and slow down the operations.



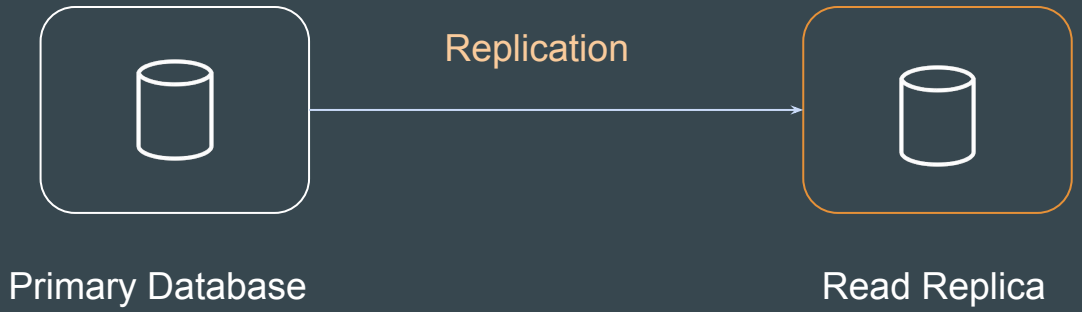
Improved Architecture - Read Replica

Read Replica allows customers to offload read requests or analytics traffic from the primary instance



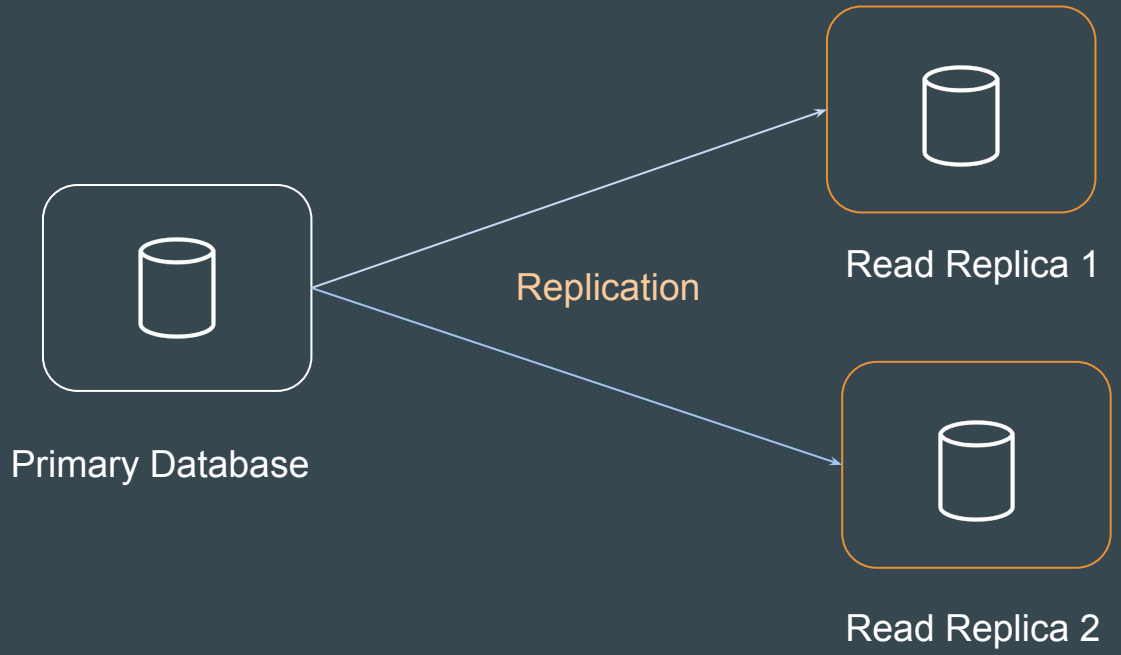
RDS Read Replica

RDS Read Replica feature allows customers to implement “Database Read Replica” functionality for RDS databases.



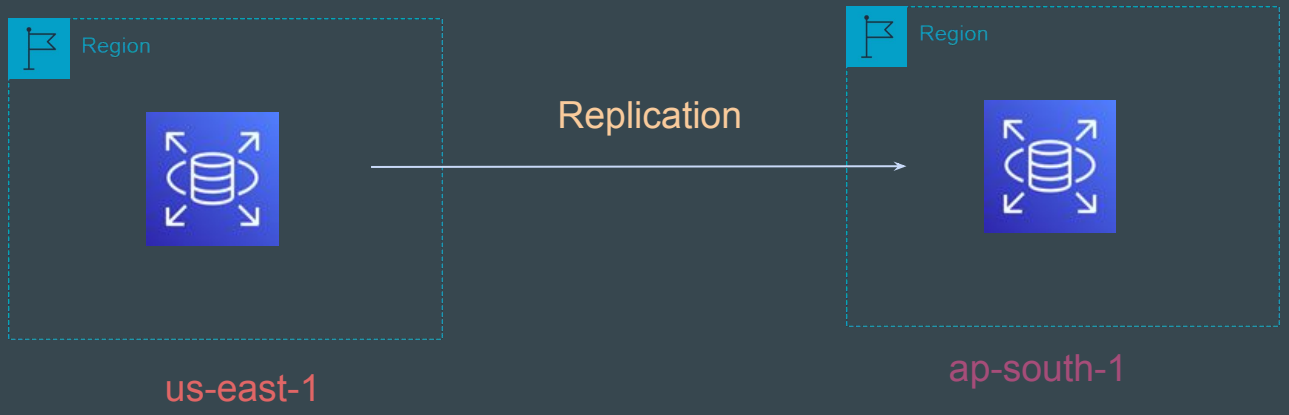
Pointers to Note - 1

You can create one or more replicas of a given source DB Instance and serve high-volume application read traffic.



Pointers to Note - 2

With Amazon RDS, you can create a read replica in a different AWS Region from the source DB instance.

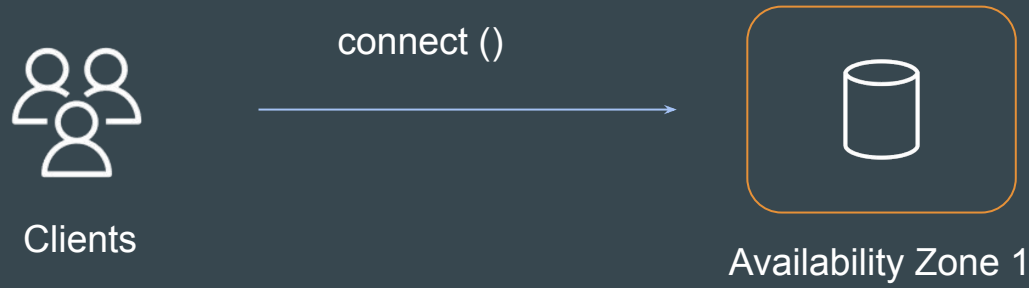


Amazon RDS Multi AZ Deployments



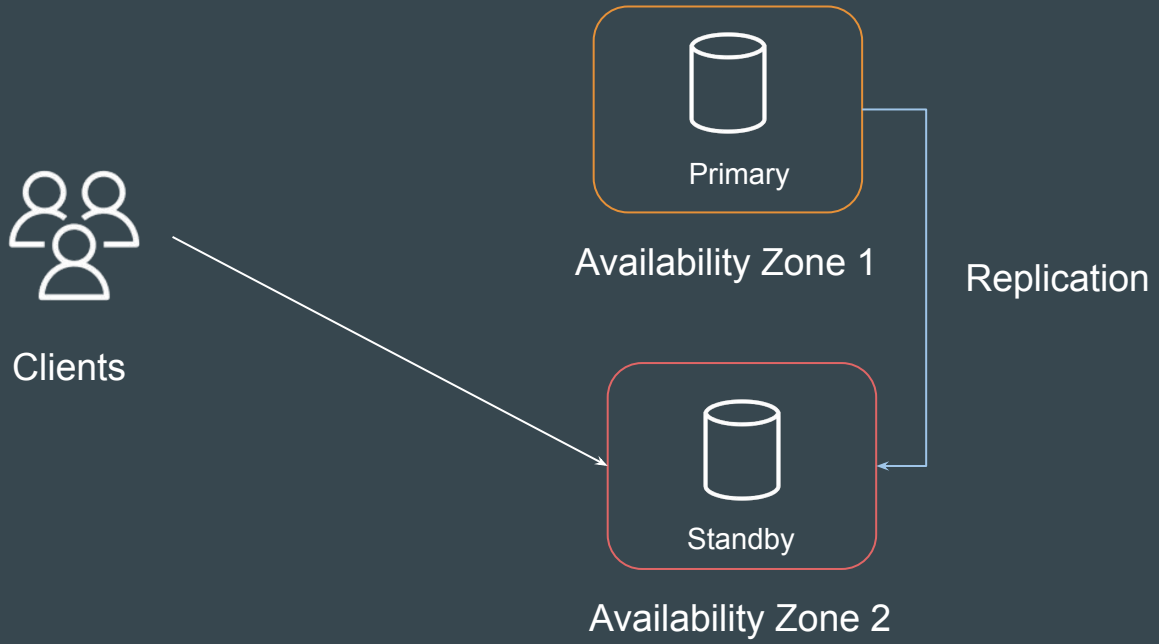
Understanding the Challenge

If database is running in a specific availability zone and if the AZ is down or unreachable then your entire application can be impacted.



Multi-AZ Architecture

In this approach, Amazon creates a standby DB instance and synchronously replicates data from the primary DB instance in a different availability zone.



Failover Condition

If a planned or unplanned outage of your DB instance results from an infrastructure defect, Amazon RDS automatically switches to a standby replica in another Availability Zone if you have turned on Multi-AZ.

- Loss of availability in primary Availability Zone
- Loss of network connectivity to primary
- Compute unit failure on primary
- Storage failure on primary

Failover times are typically 60–120 seconds.

Multi AZ Deployment Types



Types of Multi-AZ Deployments

There are two primary deployment types available in Multi-AZ based approach.



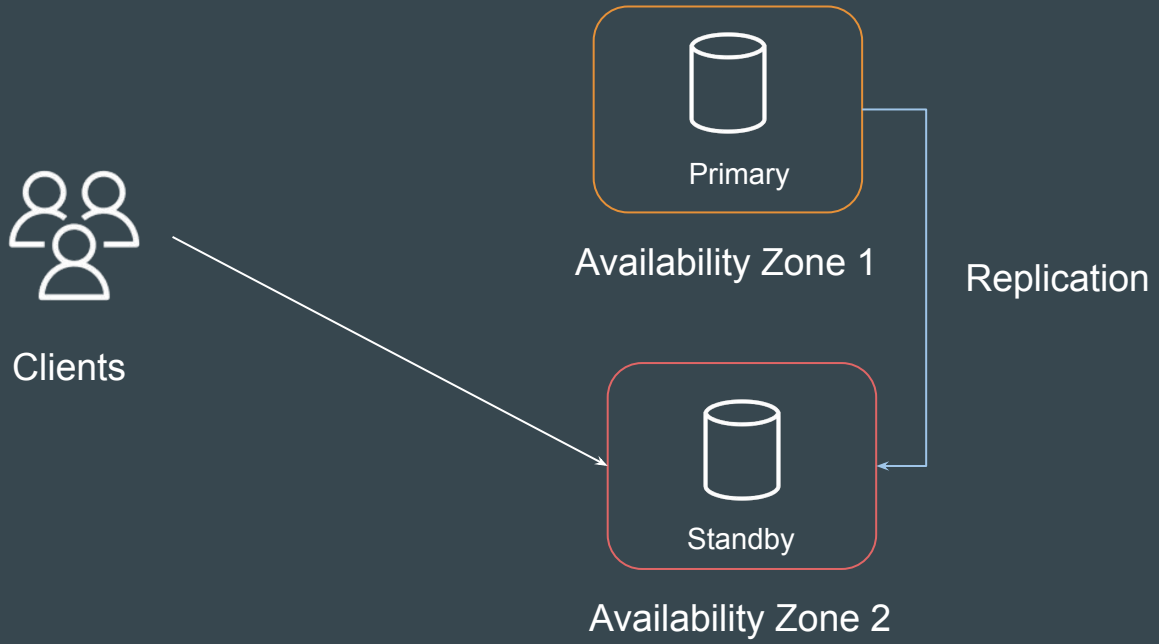
Multi-AZ Instance Deployment



Multi-AZ Cluster Deployment

Approach 1 - Multi-AZ Instance Deployment

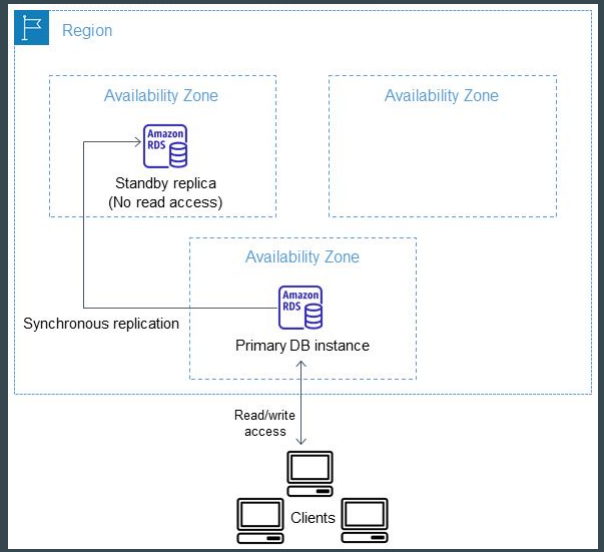
Referred to as RDS Multi-AZ with one standby



Approach 1 - Multi-AZ Instance Deployment

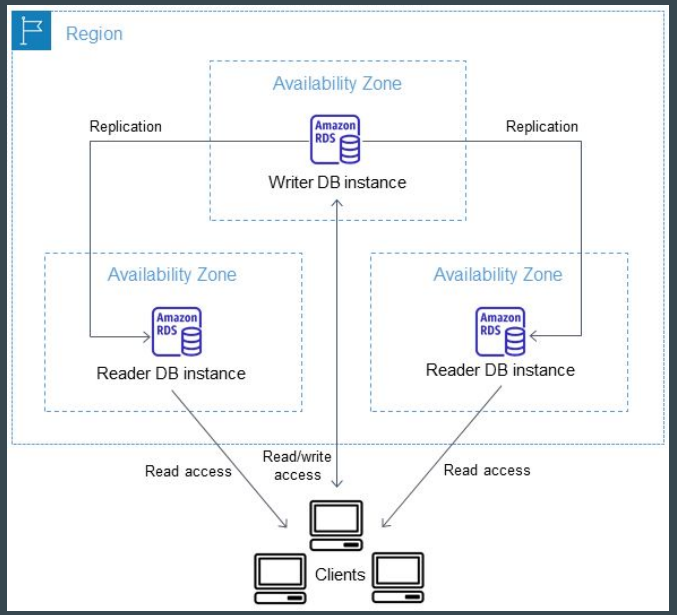
Referred to as RDS Multi-AZ with one standby

Cannot perform any operation on standby replica (including read)



Approach 2 - Multi-AZ Cluster Deployment

A Multi-AZ DB cluster has a writer DB instance and two reader DB instances in three separate Availability Zones in the same AWS Region.



Different Deployment Types

Feature	Single-AZ	Multi-AZ with 1 Standby	Multi-AZ with 2 readable Standby
Additional Read Capacity	None (only primary)	None (only primary)	2 standby DB instance
Automatic Failover Detection	None	Yes	Yes
Failover Duration	NA	New primary is available to serve workload in as quickly as 60 seconds	New primary is available to serve workload in typically under 35 seconds

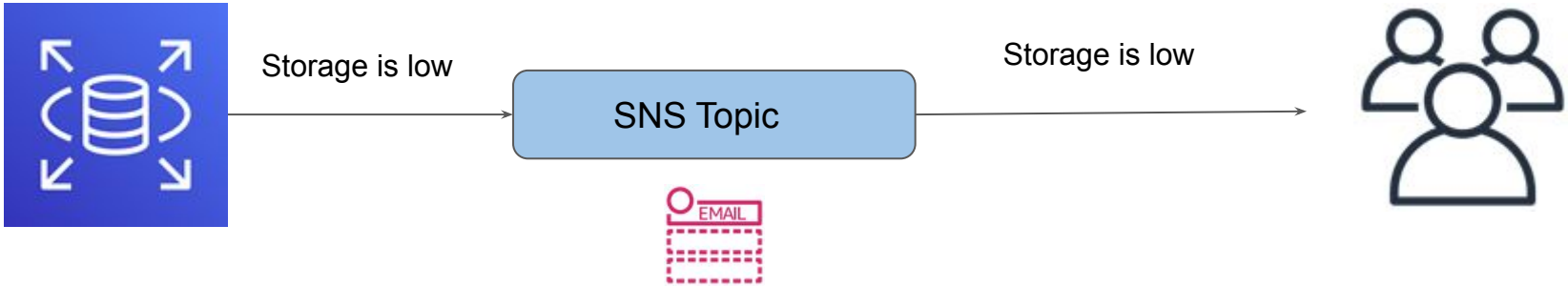
RDS Event Notification

Back to Notifications!

RDS Event Notification

RDS Event Notification provides notification when a specific type of RDS event occurs.

These events are categorized into multiple categories like Availability, Configuration Change, Failure, Deletion, Low Storage and others.



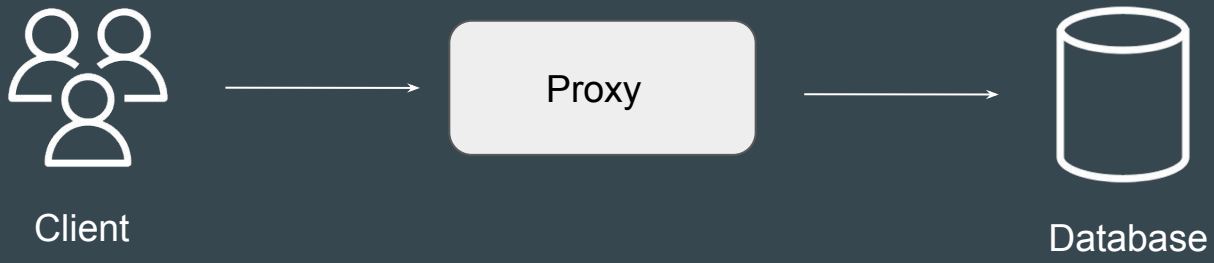
RDS Proxy



Basics of Database Proxy

Database Proxy is a intermediary between a user and database

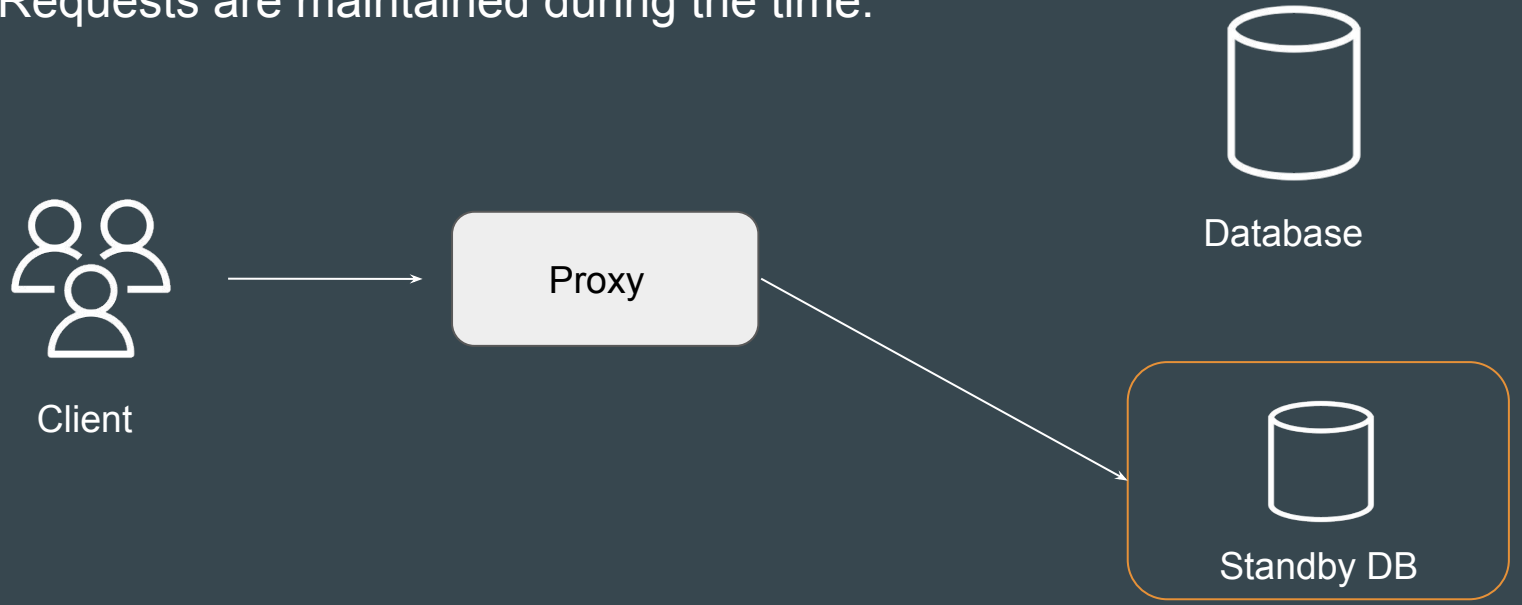
Since the queries goes through Proxy and then to database, there are lot of controls and optimizations that can be applied.



Use-Case - FailOver Scenarios

Proxies can be used in the scenario of master failover

Requests are maintained during the time.



Basics of Database Connections

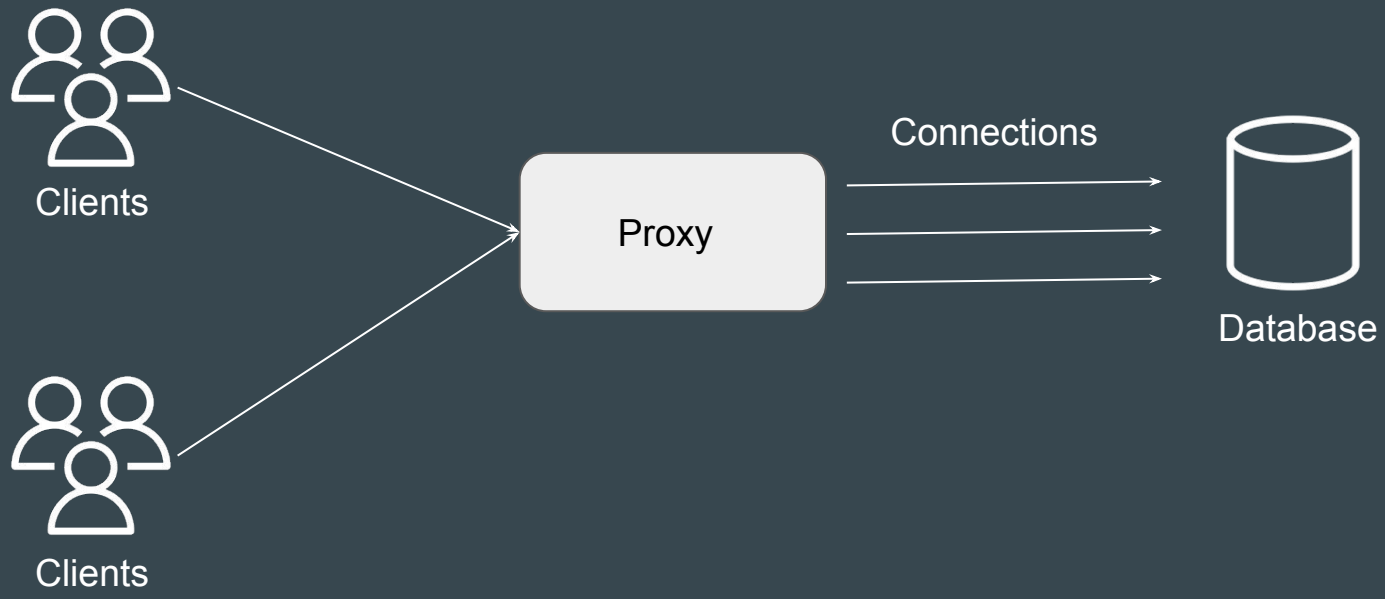
Following are high-level steps that take place when app connects to database.

1. Application makes use of Database driver to open connection to DB.
2. Network Socket is opened in OS to connect application to DB.
3. Authentication takes place.
4. Operation Complete and Connection can be closed.
5. Network Socket is closed.



Connection Pooling

Database connection pooling is a way to reduce the cost of opening and closing connections by maintaining a “pool” of open connections.



RDS Proxy

AWS RDS proxy is a fully-managed database proxy for Amazon RDS.

Benefits	Description
Connection Pooling	Improves application performance by reducing the number of open database connections
Availability	Makes applications more resilient to database failures by automatically connecting to a standby DB instance while preserving application connections.
Authentication	Can also enforce AWS Identity and Access Management (IAM) authentication for databases,

Amazon Aurora

Closed Source Database

Overview of Database Offerings

Databases are generally divided into two types:

- Open Source Databases
- Commercial Databases

Commercial Offering does come with various aspects that are not found in the open source databases.

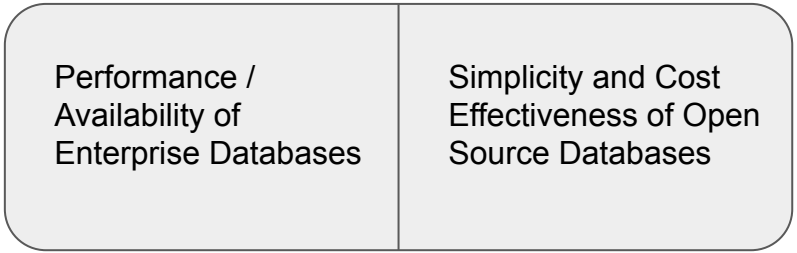


Introducing Aurora

Amazon Aurora is a MySQL and PostgreSQL-compatible relational database built for the cloud, that combines the performance and availability of traditional enterprise databases with the simplicity and cost-effectiveness of open source databases.

Amazon Aurora is up to five times faster than standard MySQL databases and three times faster than standard PostgreSQL databases.

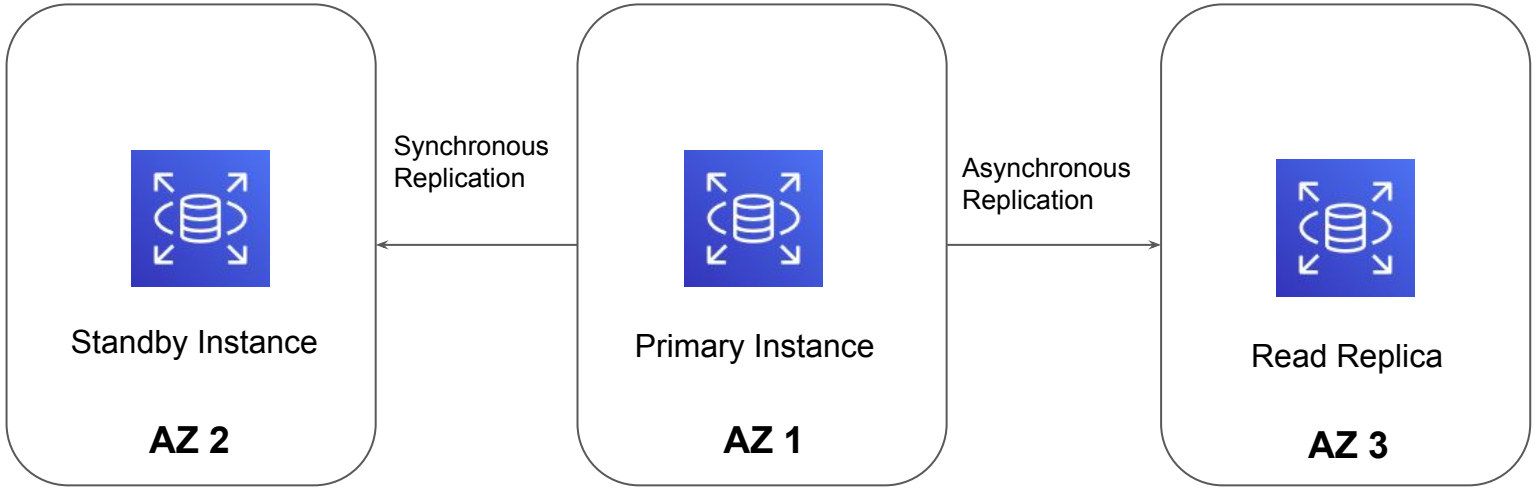
It provides the security, availability, and reliability of commercial databases at 1/10th the cost



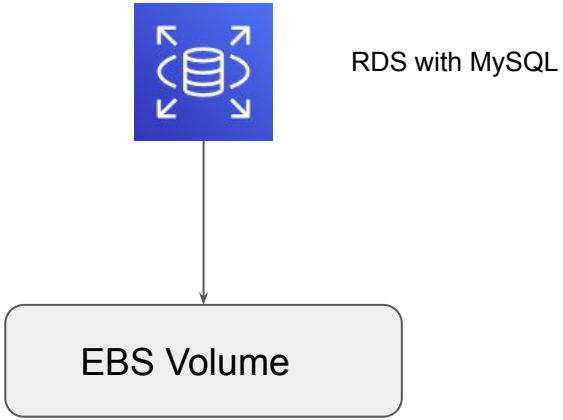
Amazon Aurora

RDS - Multi-AZ & Read Replica Architecture

In a typical setup, primary, standby and read replicas are three different instances in multiple availability zones. The underlying storage is EBS volume.



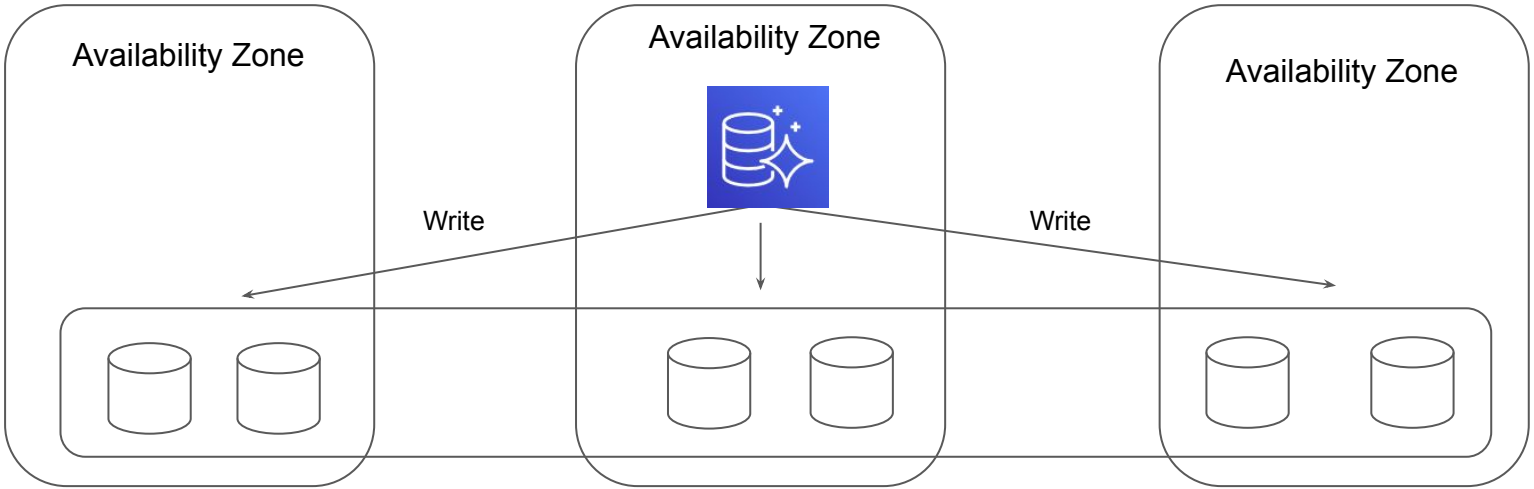
RDS with EBS Storage



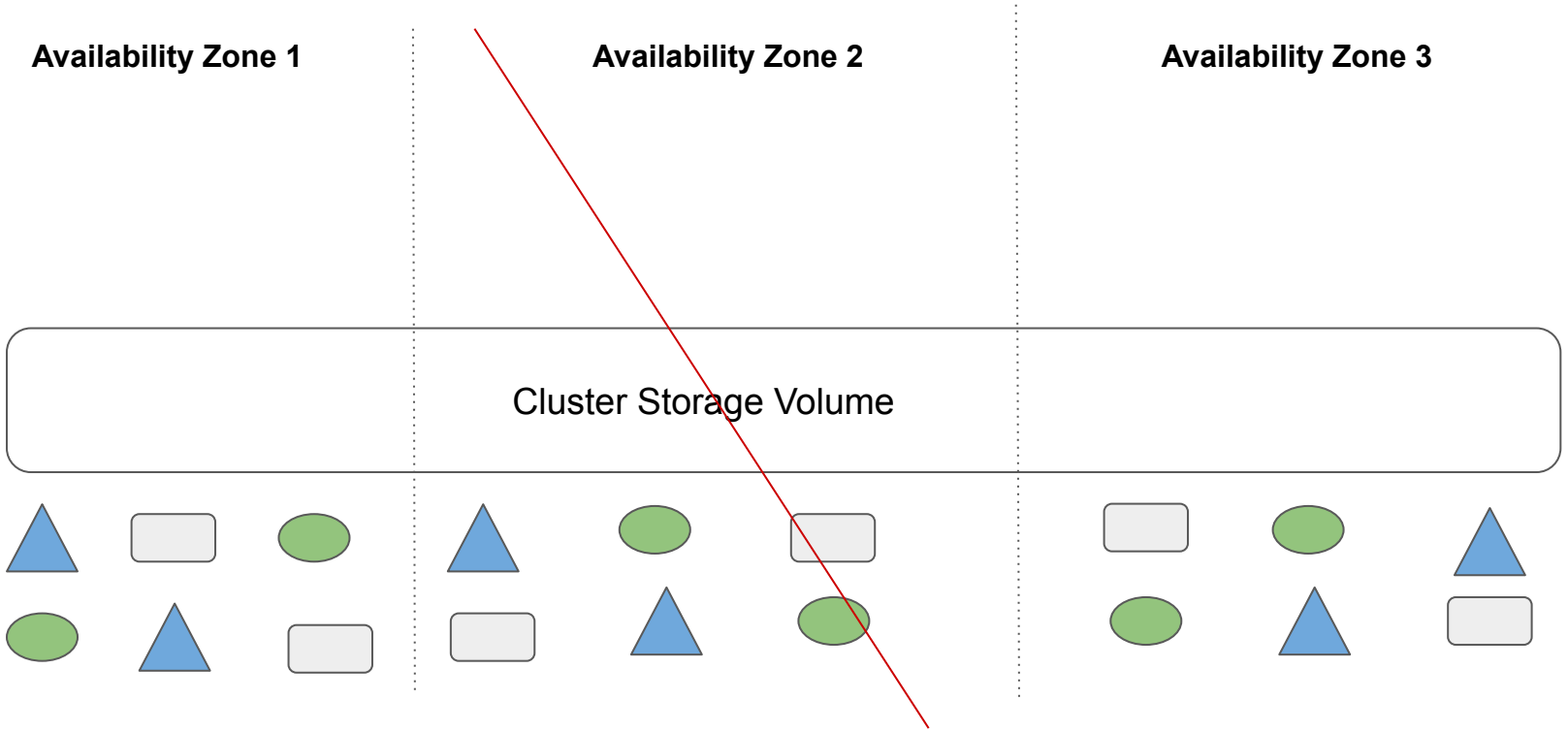
Aurora Architecture

Two Primary Components: DB Instances + Storage Cluster Volume

Since Aurora and Storage Layer are independent, we can scale the storage easily.

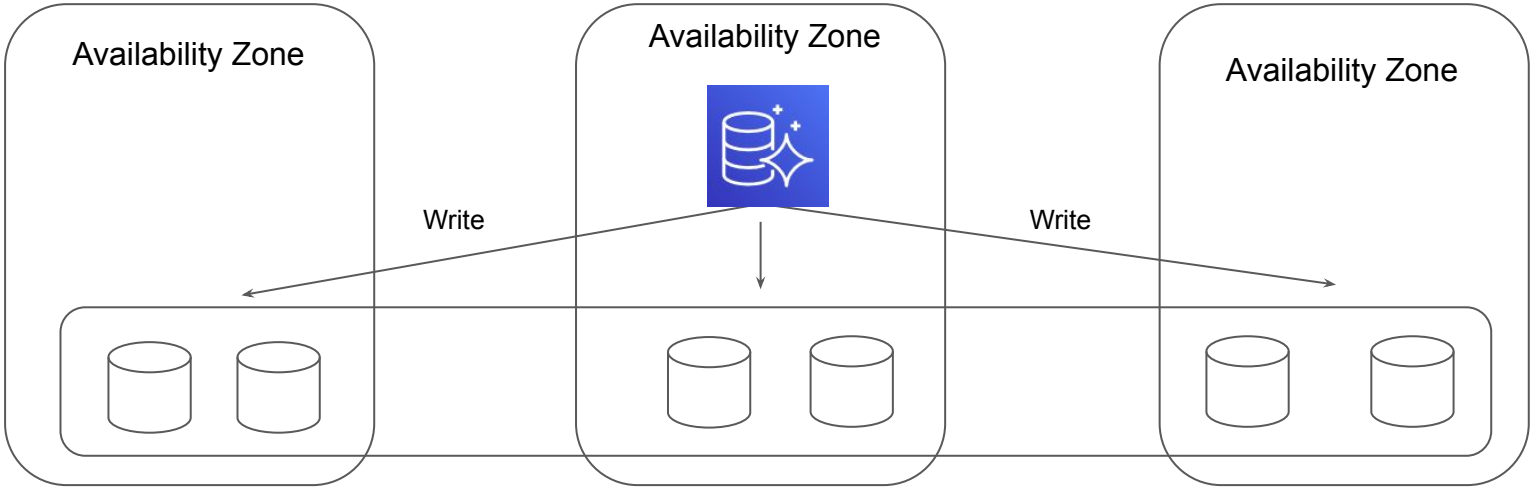


Overview of Storage Volume

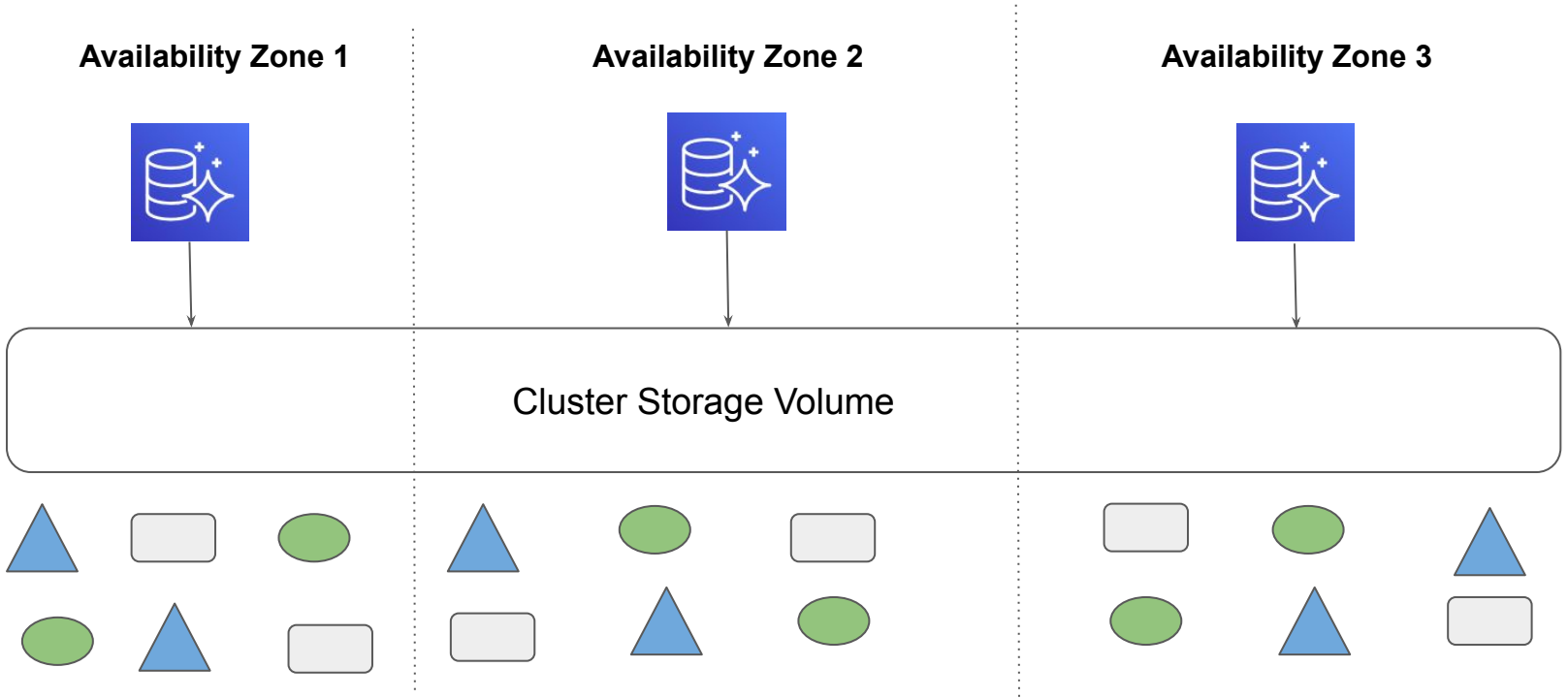


Scalability Aspect in Aurora

With this architecture, you can add a DB instance quickly because Aurora doesn't make a new copy of the table data. Instead, the DB instance connects to the shared volume that already contains all your data.



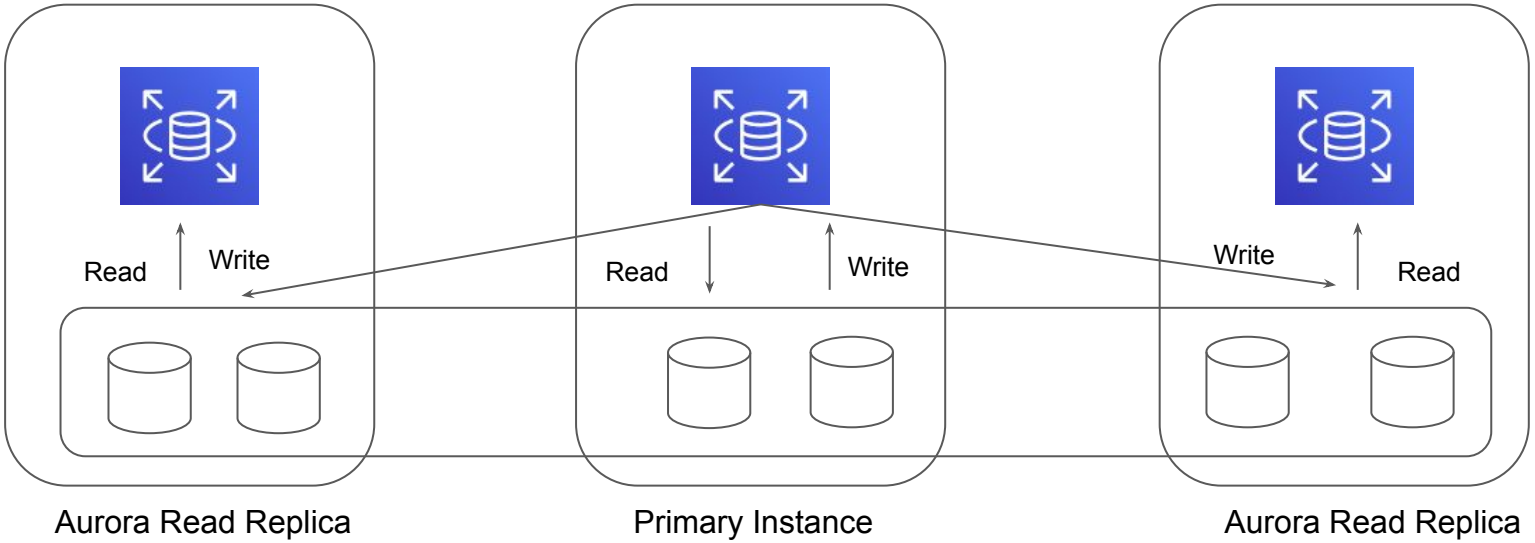
Scale at a Faster Pace



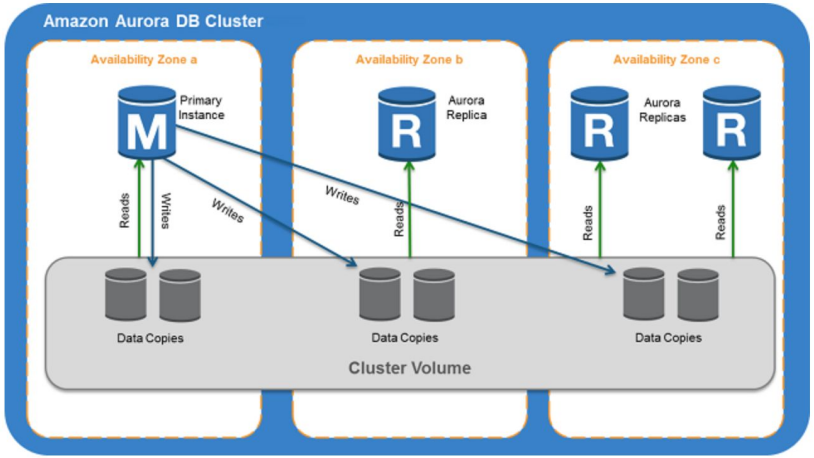
Aurora Architecture

Two Primary Components: DB Instances + Storage Cluster Volume

Since Aurora and Storage Layer are independent, we can scale the storage easily.



Aurora Architecture



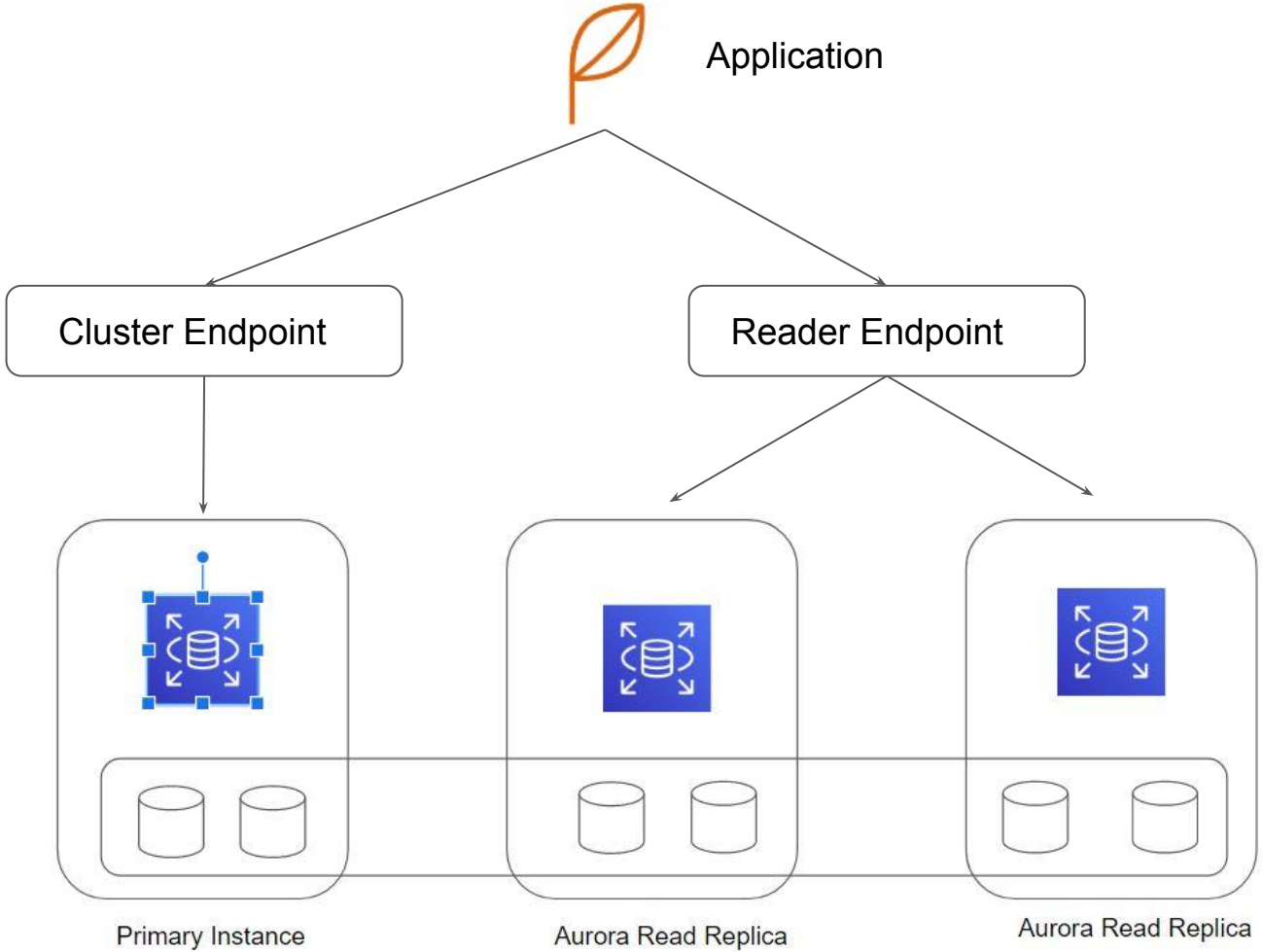
Aurora Endpoints

You can connect to Aurora Cluster through endpoints.

Endpoints is Aurora Specific URL consisting of host and port.

There are three primary types of endpoints available:

- Cluster Level Endpoints
- Reader Level Endpoints
- Instance Level Endpoints


Application
Cluster Endpoint
Reader Endpoint
Primary Instance
Aurora Read Replica
Aurora Read Replica

Aurora Endpoints

Endpoint Types	Description
Cluster Level Endpoints	Connects to current primary DB instance in the cluster. Used for performing write operations.
Reader Level Endpoints	Built-In endpoints for Read Replicas. For Multiple Read Replicas, this endpoint will balance load among all read replicas.
Instance Endpoints	Allows connection directly to the instance.
Custom Endpoints	Ability to create custom endpoints for our own requirements.

Aurora Features

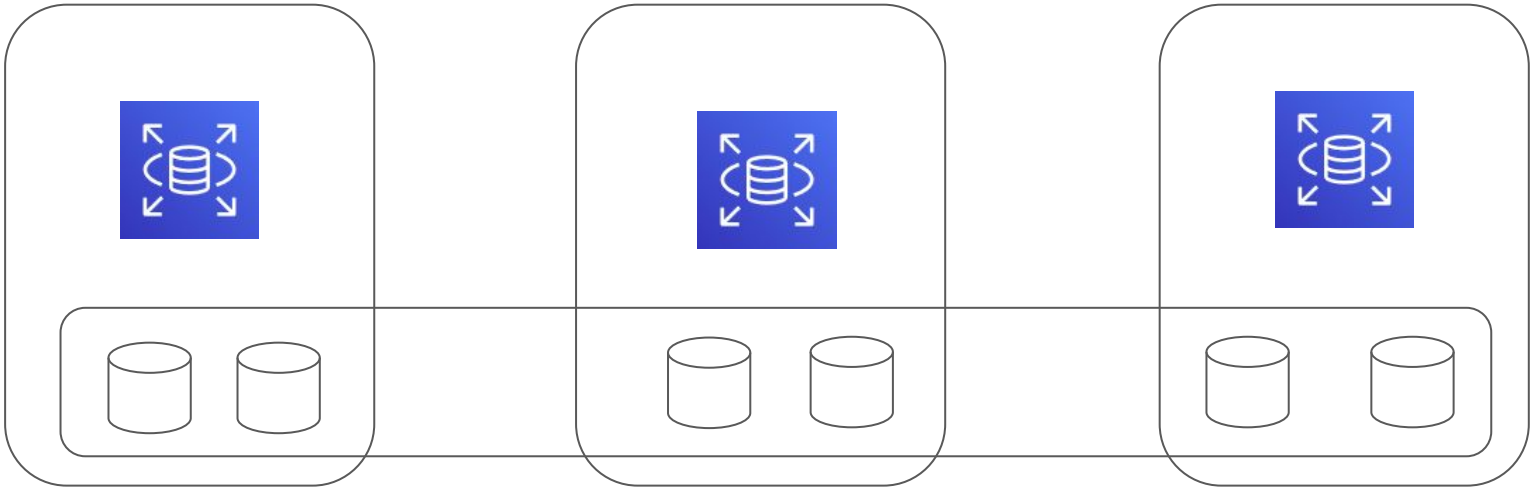
Aurora provides wide variety of interesting features. Some of these includes:

Global Databases	Serverless
Cross Region Replication	Auto-Scaling
BackTrack	IAM DB Authentication
Sharing DB Clusters	RDS Proxy

Aurora Architecture

Two Primary Components: DB Instances + Cluster Volume

Since Aurora and Storage Layer are independent, we can scale the storage easily.



Aurora Serverless

Auto-Scaling Database

Understanding the Typical Setup

In a typical DB setup, one of the primary configuration during database setup is DB instance size.

Example: t2.small

If your workload changes, you can modify the DB instance class size



Challenge with the Approach

In some environments, workloads can be intermittent and unpredictable.

There can be periods of heavy workloads that might last only a few minutes or hours, and also long periods of light activity, or even no activity.

In these cases, it can be difficult to configure the correct capacity at the right times. It can also result in higher costs when you pay for capacity that isn't used.



Choices for Customers

Provision For Peak → Expensive

Provision Less Than Peak → End-user (business) impact

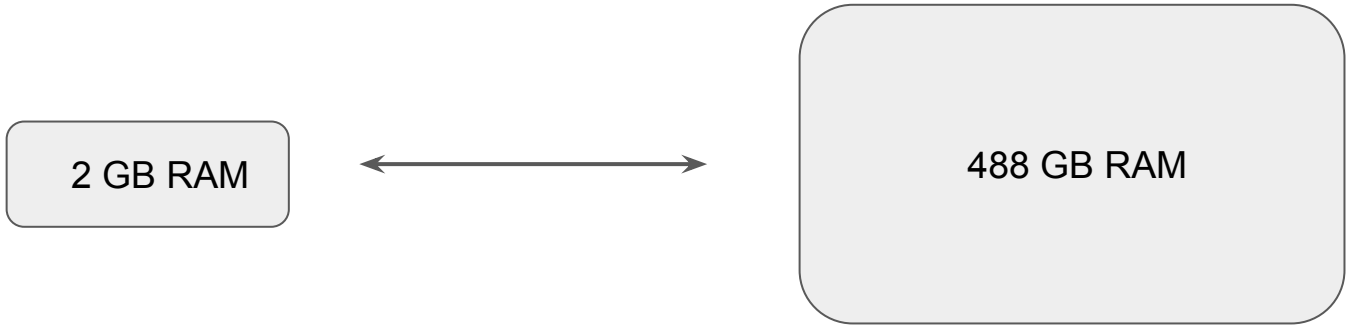
Continually Monitor and Adjust capacity manually → Require Experts & Risk Outages

Aurora Serverless

Aurora Serverless automatically scales up and down based on the capacity your workload consumes.

When DB is idle, it will automatically be shut down, and when workload resumes, it will automatically spin it back up.

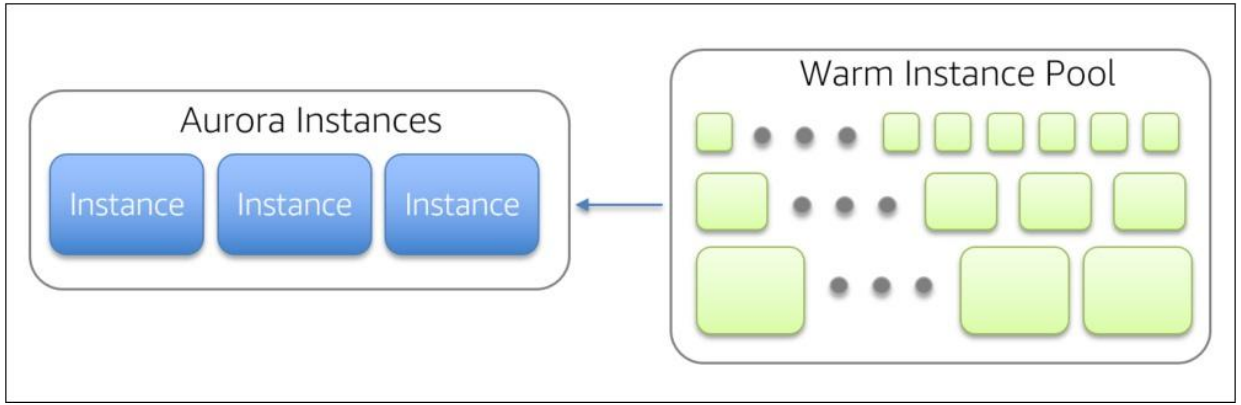
You set the minimum and maximum capacity.



Pooled Aurora Resources

Warm Instance Pool represents a warm fleet of instances that can be easily swapped in to add capacity to your environment.

These instances are allocated in a range of sizes, providing Aurora Serverless with a more granular approach to how it responds to variations in load.



Aurora Global Database

Scalability Aspect

Overview of Global Database

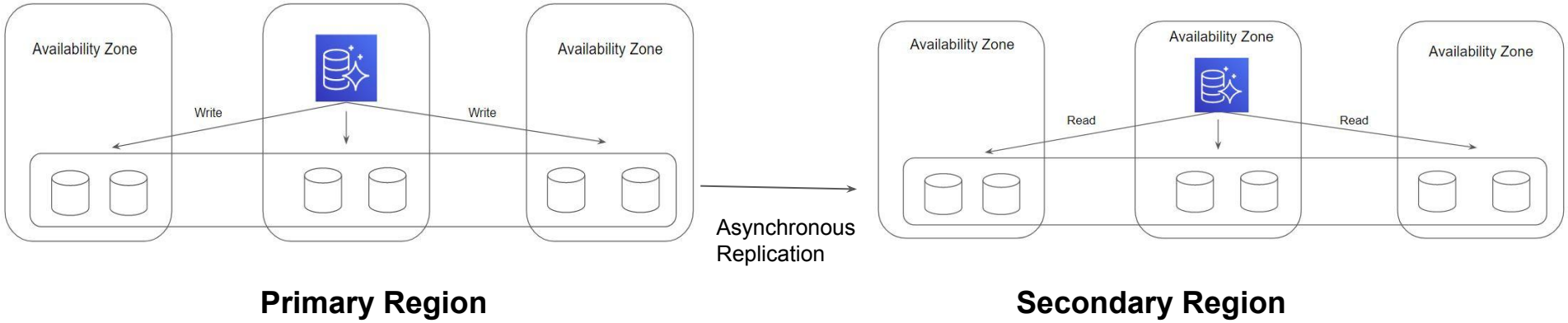
Aurora Global Database allows a single Amazon Aurora database to span multiple AWS regions.

It replicates your data with no impact on database performance, enables fast local reads with low latency in each region, and provides disaster recovery from region-wide outages.



Replication Approach

Data is replicated based on asynchronous replication between the storage layer of the two regions.



Important Pointers

Global Database does not support automated failover to the secondary region. This step is manual.

Not all instance types are supported. You can't use db.t2 or db.t3 instance classes.

Certain features like Backtrack are not supported.

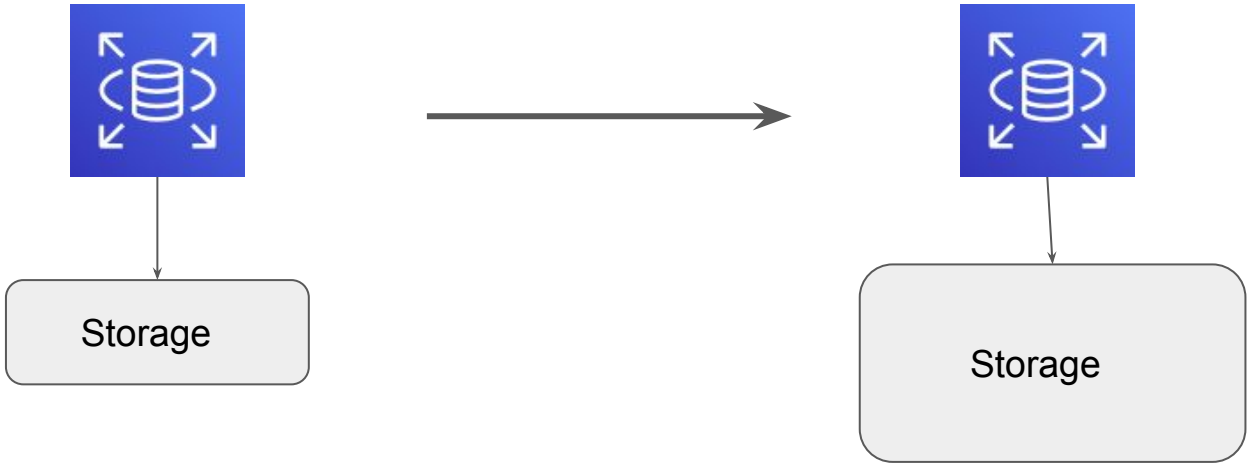
Stopping and starting the DB clusters within the global database is not supported.

RDS Storage Auto-Scaling

Auto-Scaling Storage

Overview of Storage Auto Scaling

RDS Storage Auto Scaling automatically scales storage capacity in response to growing database workloads, with zero downtime



Important Pointers to Remember

Amazon RDS starts a storage modification for an auto scaling-enabled DB instance when these factors apply:

1. Free available space is less than 10 percent of the allocated storage.
2. The low-storage condition lasts at least five minutes.
3. At least six hours have passed since the last storage modification.

Important Pointers to Remember - 2

Autoscaling can't completely prevent storage-full situations for large data loads, because further storage modifications can't be made until six hours after storage optimization has completed on the instance.

The additional storage is in increments of whichever of the following is greater:

- 5 GiB
- 10 percent of currently allocated storage

Autoscaling can't be used with the following previous-generation instance classes that have less than 6 TiB of orderable storage: db.m3.large, db.m3.xlarge, and db.m3.2xlarge.

Relax and Have a Meme Before Proceeding



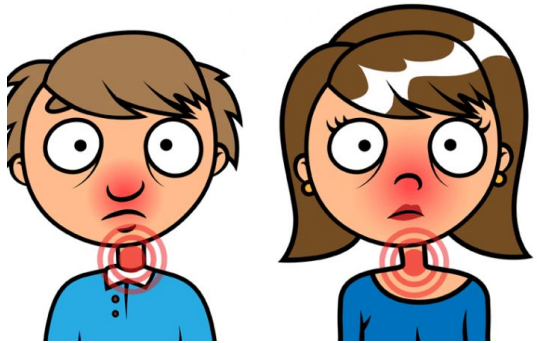
ElastiCache

Let's Cache

Simple Analogy

- There is a new vegetable shop in the locality which has become very popular.
- Every day 300-500 people visit and buy vegetables.
- Each visitor asks the price of at-least two-three veggies before making a purchase.

Imaging the condition of the employee inside that shop after a few days.



Simple Analogy - Smart Approach

Vegetable Shop Owner decided to create a dashboard that has a list of all the common vegetable prices which are requested by the buyers.



Price List

Veggie A - $1

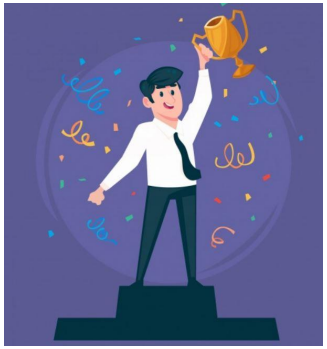
Veggie B - $2

Veggic C - $C

Simple Analogy - Learning

1. Since the price list of common items was listed, users no longer need to ask the employees about it. This reduces the overall load on the employee.

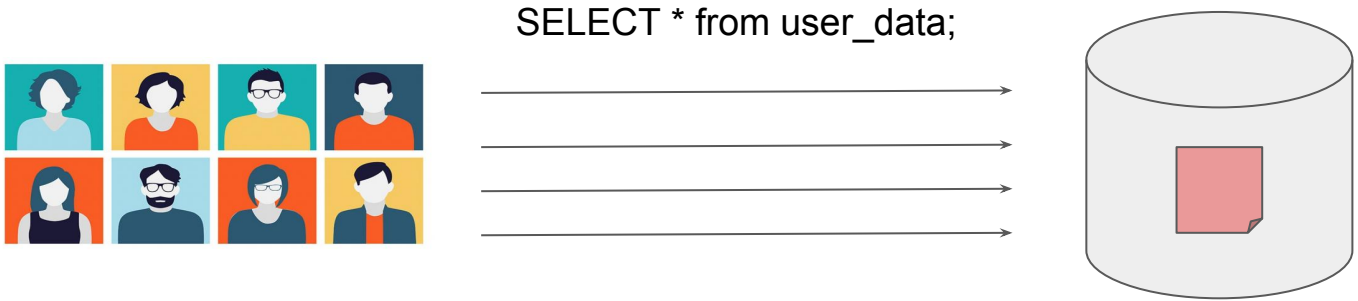
2. Visitors can quickly get to go through the price list - Better Efficiency.



Challenges with Database Workloads

There can be certain common queries within the database that hundreds of users might request.

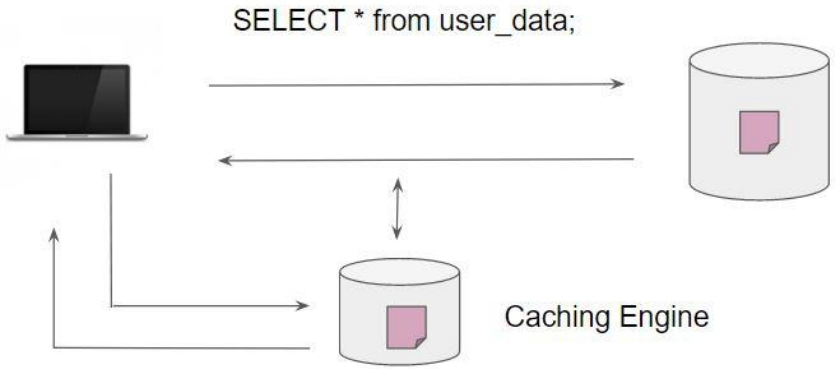
This would increase the load on the database and can lead to performance degradation.



Caching Solutions

With caching solutions, you can cache the response associated with frequent queries.

This allows better response time and decreases the load on the database servers.



Popular Caching Solutions

Two of the most popular caching solutions used for databases are:

1. Memcached
2. Redis

To use them, you will have to install, configure, optimize and secure the EC2 instances where these engines would be running.





Introducing AWS ElastiCache

ElastiCache is a fully managed AWS service that makes it easier to deploy, operate and scale an in-memory data-store or cache in the cloud.

It is like a managed service and within a few clicks, we can have an in-memory layer in our infra.

ElastiCache can also detect and replace failed nodes thus reducing the overhead.



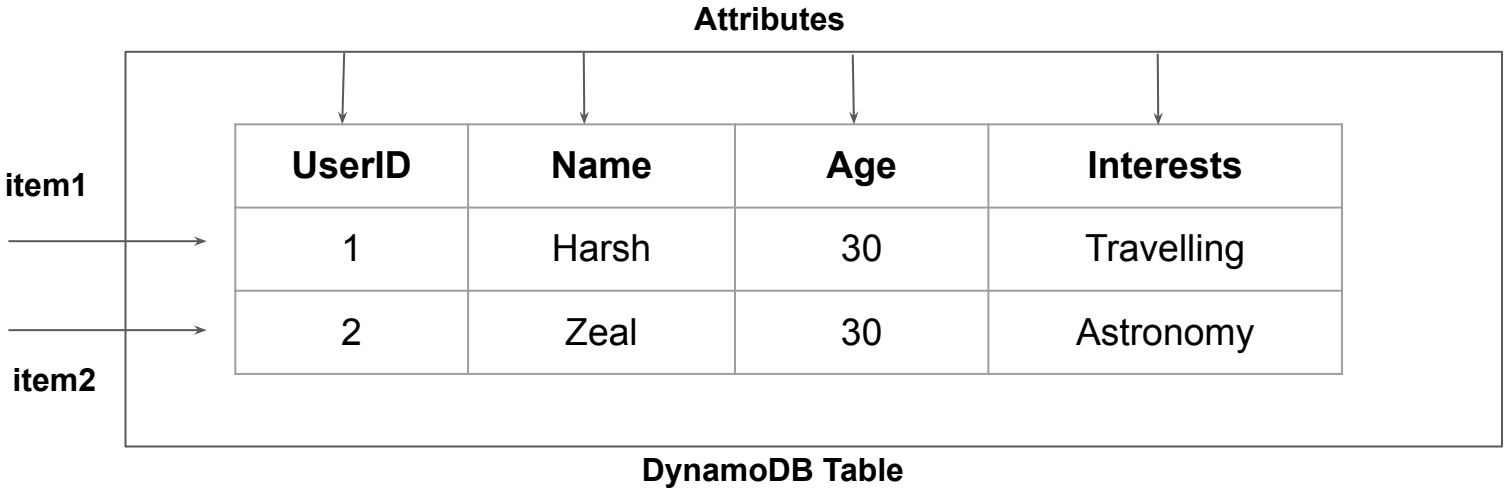
Core Components - DynamoDB

DynamoDB Basics

Understanding the Basics

In DynamoDB, tables, items, and attributes are the core components that you work with.

Table is a collection of items, and each item is a collection of attributes.



Importance of Primary Key

Each item in the table has a unique identifier, or primary key, that distinguishes the item from all of the others in the table

Other than the primary key, the table is schemaless, which means that neither the attributes nor their data types need to be defined beforehand.

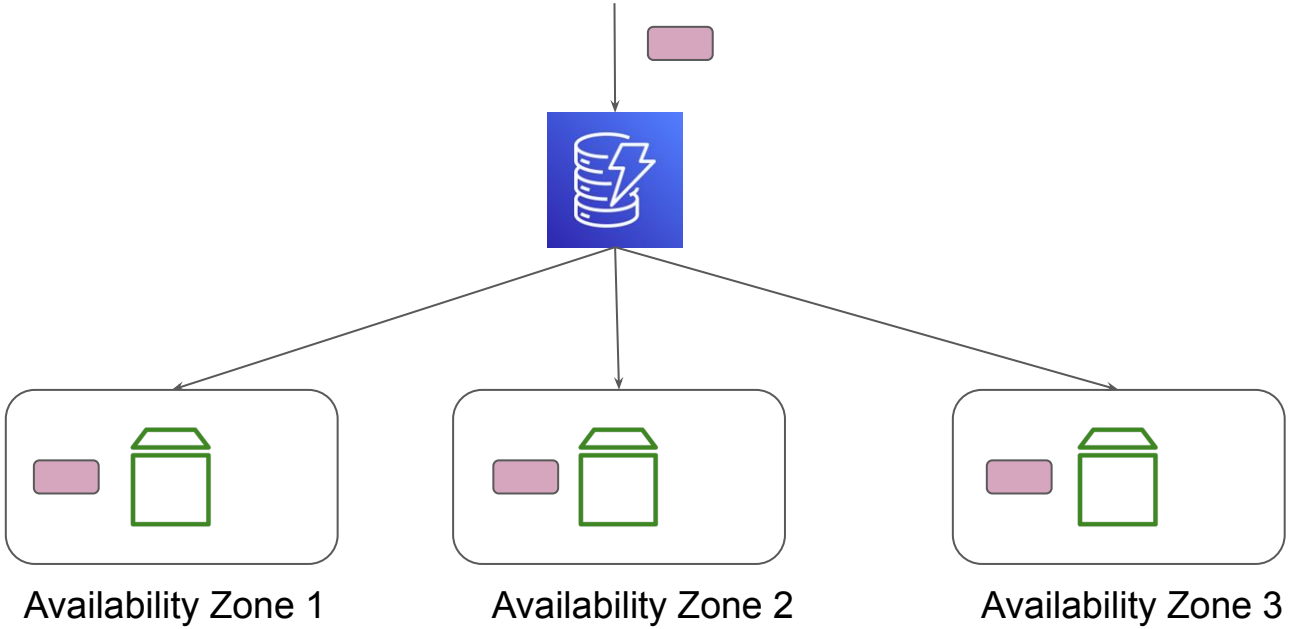


Consistency Model

Important Storage Concept

Understanding Consistency Model

In DynamoDB, all of your data is stored on SSDs and is automatically replicated across multiple Availability Zones in an Amazon Region, providing built-in high availability and data durability.



Consistency Timeline

When your application writes data to a DynamoDB table and receives an HTTP 200 response (OK), the write has occurred and is durable.

The data is eventually consistent across all storage locations, usually within one second or less.

Eventual Consistency Reads

When you read data from a DynamoDB table, the response might not reflect the results of a recently completed write operation.

The response might include some stale data.

If you repeat your read request after a short time, the response should return the latest data.

Strong Consistency Reads

When you request a strongly consistent read, DynamoDB returns a response with the most up-to-date data, reflecting the updates from all prior write operations that were successful.

1. A strongly consistent read might not be available if there is a network delay or outage. In this case, DynamoDB may return a server error (HTTP 500).

2. Strongly consistent reads may have higher latency than eventually consistent reads.

3. Strongly consistent reads use more throughput capacity than eventually consistent reads.

Important Note

DynamoDB uses eventually consistent reads, unless you specify otherwise.

Read operations (such as GetItem, Query, and Scan) provide a ConsistentRead parameter. If you set this parameter to true, DynamoDB uses strongly consistent reads during the operation.

Example Command:

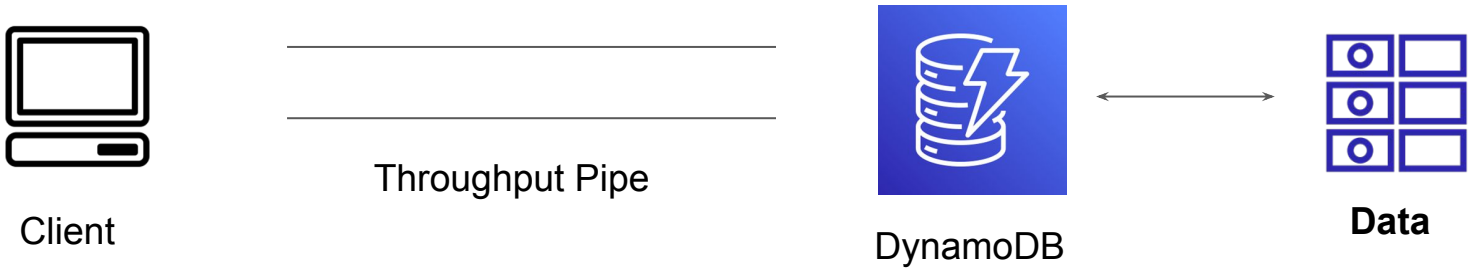
aws dynamodb get-item --table-name MusicCollection --key file://key.json
--consistent-read

Read/Write Capacity Units

Managing Throughput

Throughput in DynamoDB

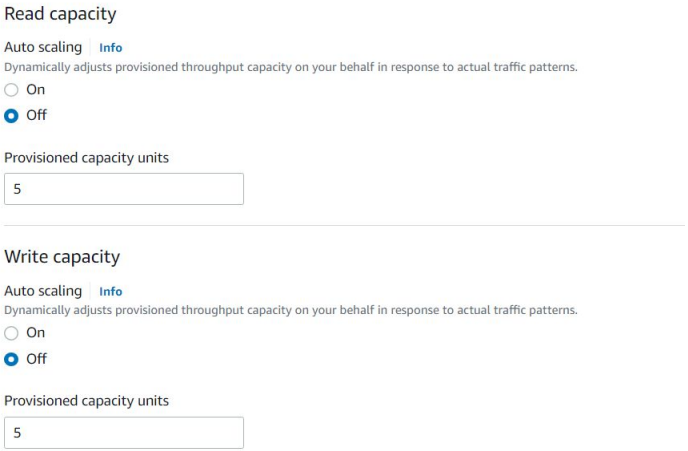
Throughput is the maximum amount of capacity that an application can consume from a table or index



Setting the Read & Write Capacity

We can specify throughput capacity in terms of read capacity units (RCUs) and write capacity units.

.



Read Request Unit

One read request unit represents one strongly consistent read request, or two eventually consistent read requests, for an item up to 4 KB in size.

If you need to read an item that is larger than 4 KB, DynamoDB needs additional read request units.

Item Size	Read Capacity Unit (Strong)	Read Capacity Unit (Eventual)
4 KB	1	1
8 KB	2	1
10 KB	3	2

Write Request Unit

One write request unit represents one write for an item up to 1 KB in size.

If you need to write an item that is larger than 1 KB, DynamoDB needs to consume additional write request units.

.

Item Size	Write Capacity Unit
1 KB	1
4 KB	4
10 KB	10

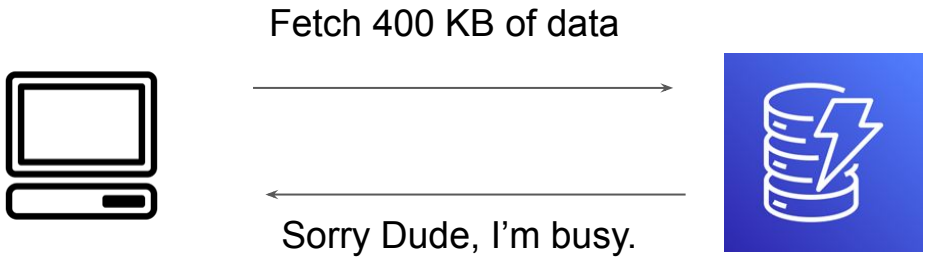
Capacity Modes in DynamoDB

Adjust Throughput Automatically

Understanding the Challenge

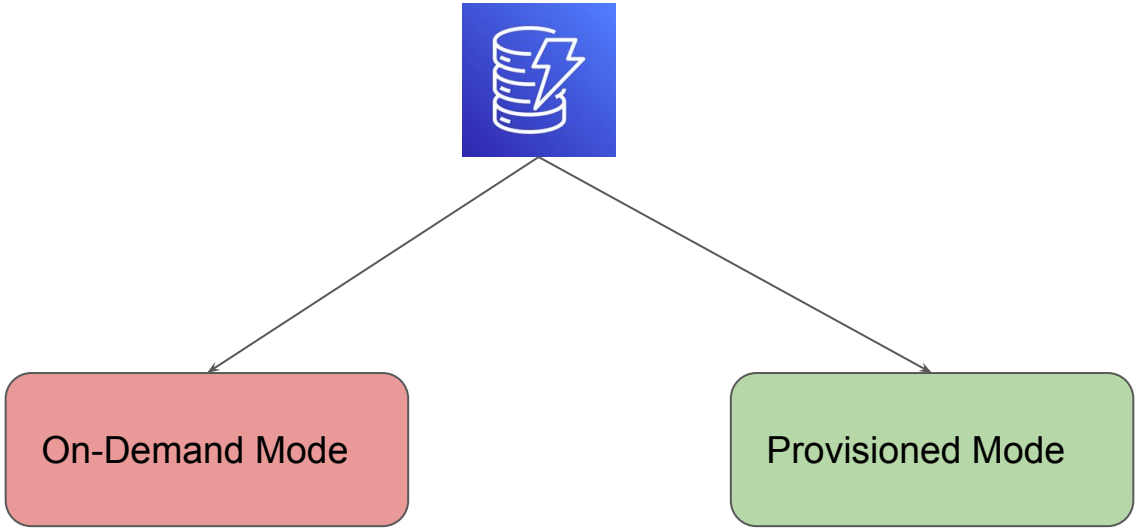
We can specify throughput capacity in terms of read capacity units (RCUs) and write capacity units

If your application exceeds your provisioned throughput capacity on a table or index, it is subject to request throttling.



Types of Capacity Modes

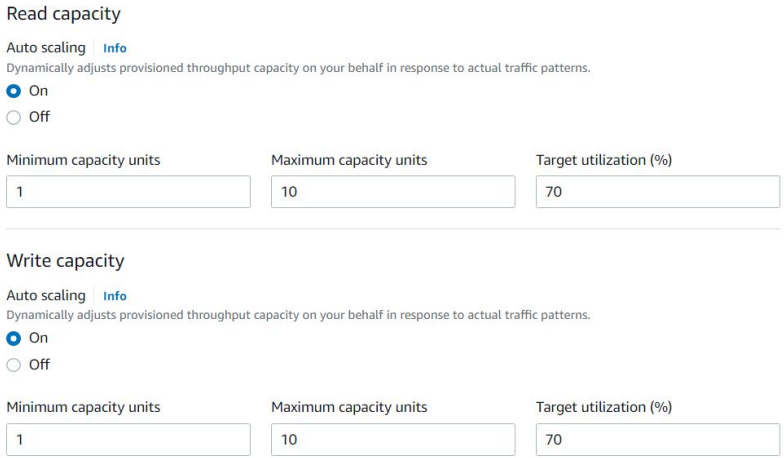
There are two primary capacity modes available in DynamoDB.



Provisioned Mode

If you choose provisioned mode, you specify the number of reads and writes per second that you require for your application.

You can use auto scaling to adjust your table's provisioned capacity automatically in response to traffic changes.



Recommended Traffic Patterns for Provisioned Mode

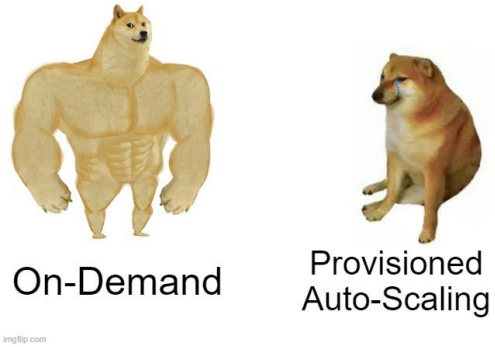
Provisioned mode is a good option if any of the following are true:

- You have predictable application traffic.
- You run applications whose traffic is consistent or ramps gradually.
- You can forecast capacity requirements to control costs.

On-Demand Mode

Amazon DynamoDB on-demand is capable of serving thousands of requests per second without capacity planning.

DynamoDB on-demand offers pay-per-request pricing for read and write requests so that you pay only for what you use.



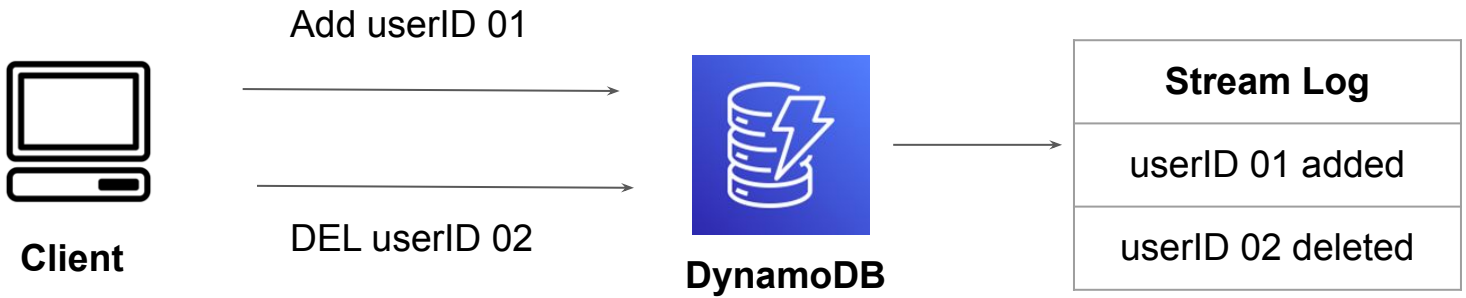
DynamoDB Streams

Stream Records Real-Time

Understanding the Basics

DynamoDB Streams captures a time-ordered sequence of item-level modifications in any DynamoDB table and stores this information in a log for up to 24 hours.

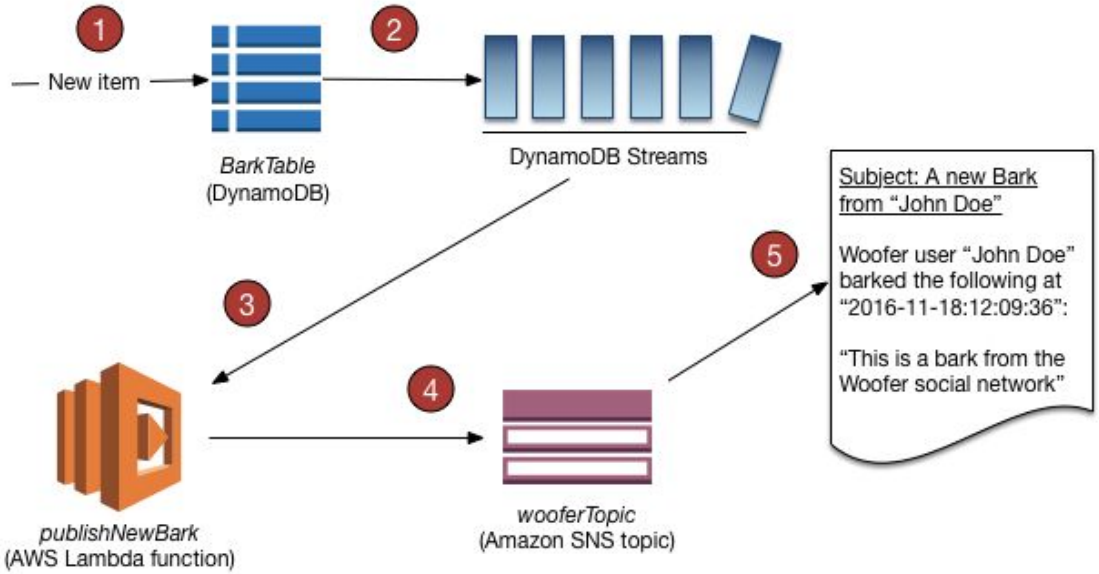
Applications can access this log and view the data items as they appeared before and after they were modified, in near-real time.



Sample Record Log in CloudWatch

▶	2022-07-15T21:06:51.711+05:30	REMOVE
▶	2022-07-15T21:06:51.711+05:30	DynamoDB Record: {
▶	2022-07-15T21:06:51.711+05:30	"ApproximateCreationDateTime": 1657899411.0,
▶	2022-07-15T21:06:51.711+05:30	"Keys": {
▶	2022-07-15T21:06:51.711+05:30	"userID": {
▶	2022-07-15T21:06:51.711+05:30	"S": "01"
▶	2022-07-15T21:06:51.711+05:30	}
▶	2022-07-15T21:06:51.711+05:30	},
▶	2022-07-15T21:06:51.711+05:30	"OldImage": {
▶	2022-07-15T21:06:51.711+05:30	"courseName": {
▶	2022-07-15T21:06:51.711+05:30	"S": "AWS Certification Course"
▶	2022-07-15T21:06:51.711+05:30	},
▶	2022-07-15T21:06:51.711+05:30	"userID": {
▶	2022-07-15T21:06:51.711+05:30	"S": "01"

A Sample Use-Case



Use-Cases

1. Allows setting up a relationship across multiple tables in which, based on the value of an item from one table, you update the item in a second table

2. Triggering an event based on a particular item change

3. Audit or Archive Data

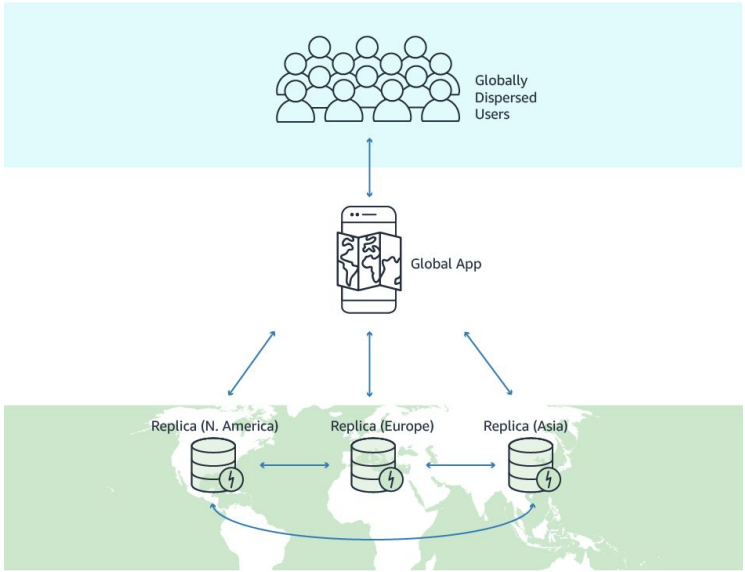
4. Replicating Data Across Multiple Tables

DynamoDB - Global Table

Let's Replicate

Basics of Global Tables

Global tables feature provides us with a fully managed, multi-Region, and multi-active database that delivers fast, local, read and write performance for massively scaled, global applications.



Basic Terminology

A global table is a collection of one or more replica tables, all owned by a single AWS account.

A replica table is a single DynamoDB table that functions as a part of a global table. Each replica stores the same set of data items.

When an application writes data to a replica table in one Region, DynamoDB propagates the write to the other replica tables in the other AWS Regions automatically.

Important Pointers

In a global table, a newly-written item is usually propagated to all replica tables within seconds.

With a global table, each replica table stores the same set of data items. DynamoDB does not support partial replication of only some of the items.

Conflicts can arise if applications update the same item in different regions at about the same time. To ensure eventual consistency, DynamoDB global tables use a “last writer wins”

DynamoDB Accelerator (DAX)

Let's Accelerate Read Requests

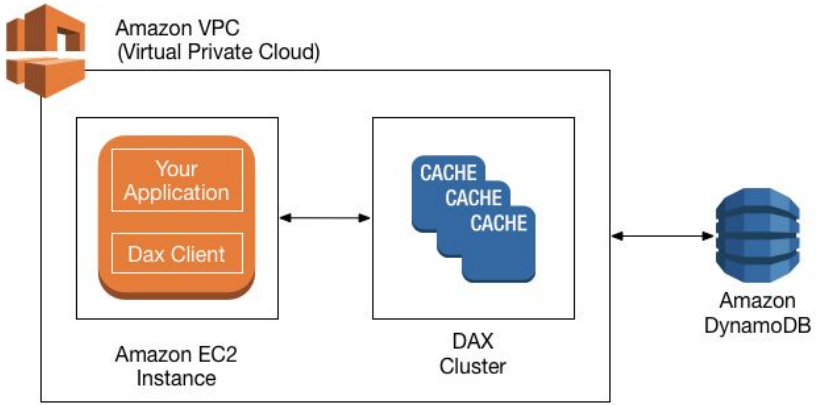
Understanding the Need

In most cases, the DynamoDB response times can be measured in single-digit milliseconds. However, there are certain use cases that require response times in microseconds.

For these use cases, DynamoDB Accelerator (DAX) delivers fast response times for accessing eventually consistent data.

Overview of the Feature

Amazon DynamoDB Accelerator (DAX) is a fully managed, highly available, in-memory cache for Amazon DynamoDB that delivers up to a 10x performance improvement.



Use-Case for DAX

Applications that require the fastest possible response time for reads. Some examples include real-time bidding, social gaming, and trading applications.

Applications that read a small number of items more frequently than others

Applications that are read-intensive, but are also cost-sensitive.

Where it is not suitable for ?

DAX is not ideal for the following types of applications:

Applications that require strongly consistent reads (or that cannot tolerate eventually consistent reads).

Applications that do not require microsecond response times for reads, or that do not need to offload repeated read activity from underlying tables.

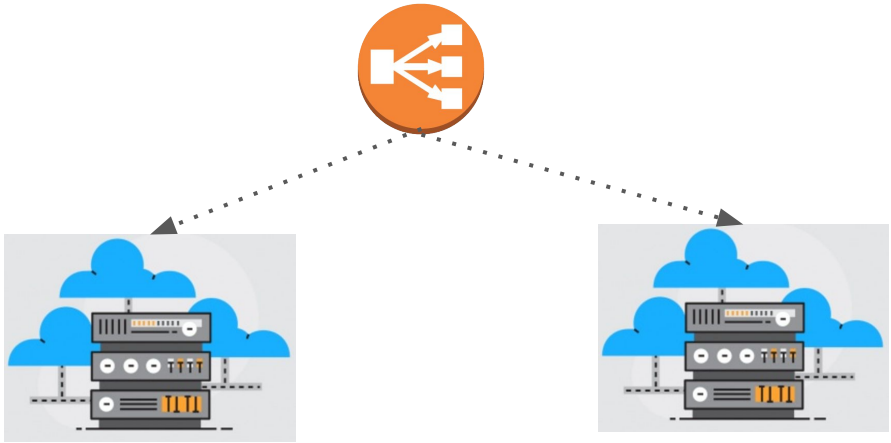
Applications that are write-intensive, or that do not perform much read activity.

RTO & RPO

Health should always be good

Everything comes at price

- High Availability Architecture is driven by your requirements.
- An highly available, multi-AZ, fault tolerant infrastructure is certainly possible, however there is cost associated with it.



Recovery Time Objective

- Recovery Time Objective (RTO) is the amount of time frame it takes for you to recover your infrastructure and business operations after disaster has struck.

Sample Example:

- If RTO is 3 hours, then one needs to invest quiet good amount of money to make sure DR region is always ready in-case main region goes down due to disaster.

Recovery Point Objective

- Recovery Point Objective (RPO) is concerned with data and maximum tolerance period to which data can be lost.

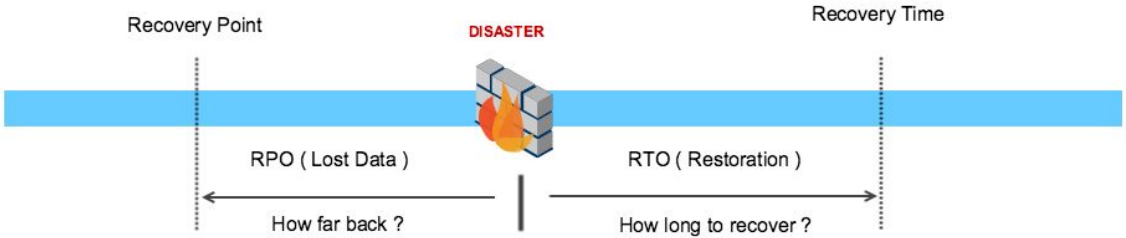
- It helps in determining how well we should be designing the infrastructure.

Sample Example:

- If RPO is 5 hours for database, then we should be taking backup of database every five hours .

RTO vs RPO

- RTO is more broader scope and covers whole business and systems involved while RPO is more directly related to interval of backup to take to avoid data loss.



IAM DB Authentication

Challenges and Structure

Getting Started

We can authenticate to your DB instance using AWS Identity and Access Management (IAM) database authentication.

IAM database authentication works with MySQL and PostgreSQL.

In this method, we don't need to have a password, instead we can make use of authentication token.

Things to Remember

Network traffic to and from the database is encrypted using Secure Sockets Layer (SSL).

You can use IAM to centrally manage access to your database resources, instead of managing access individually on each DB instance.

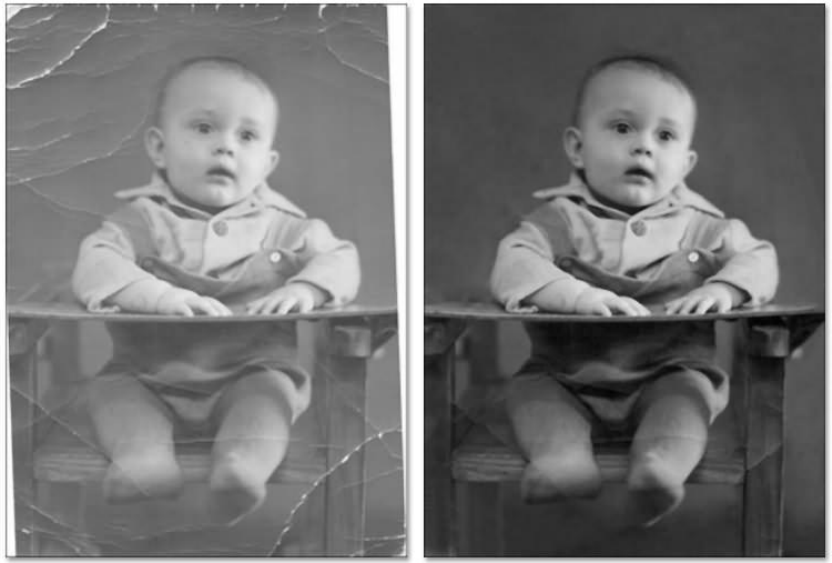
When using IAM database authentication with MySQL, you are limited to a maximum of 20 new connections per second.

Intro to Amazon SQS

Message Queuing Service

Use-Case: Restoring Image Application

Medium Corp is designing an application that will enhance and restore the images that users submit through the online portal.



Current Architecture

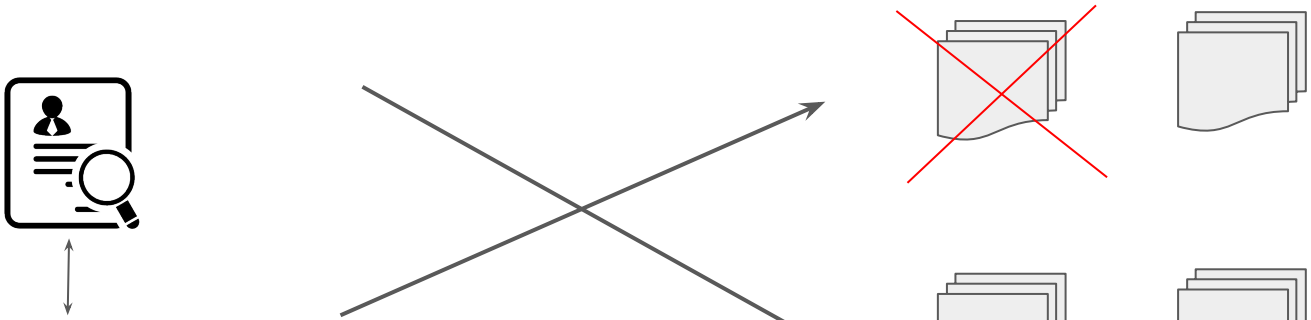
The overall architecture involves two components:

1. Image Gatherer - Takes the Images from the user via Upload button.
2. Imager Enhancer - Receives the Image from Image Gatherer.



Challenges

Due to popularity of the application and huge traffic spike, Medium Corp has decided to add more image enhancer servers.

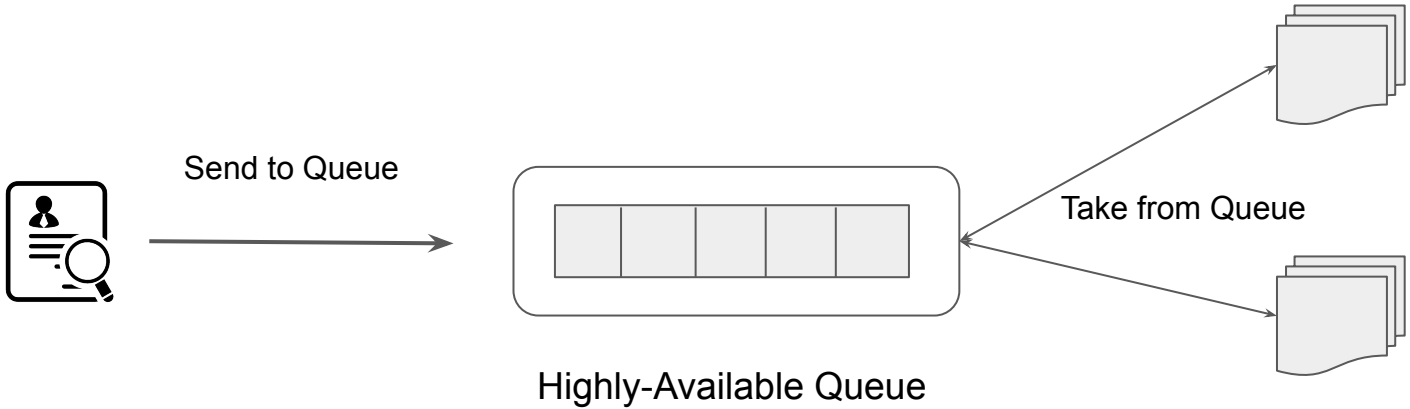


server-a	10.77.2.5
server-b	10.66.0.10
server-c	192.168.0.2
server-d	10.66.10.10

Better Architecture

One of the main function of message queue service is to take message from a Publisher and forward that to a consumer.

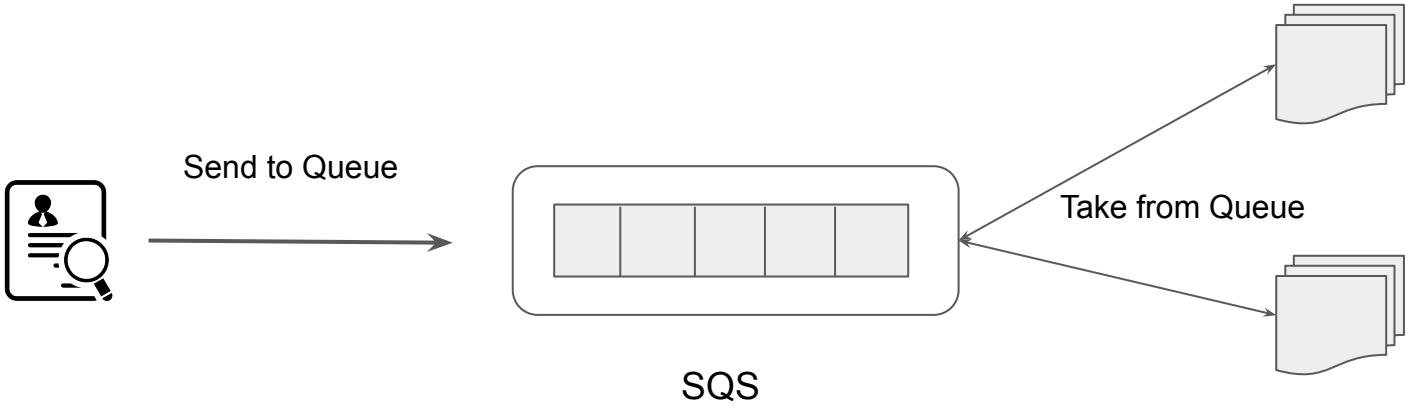
The queue stores these messages internally.



Introduction to SQS

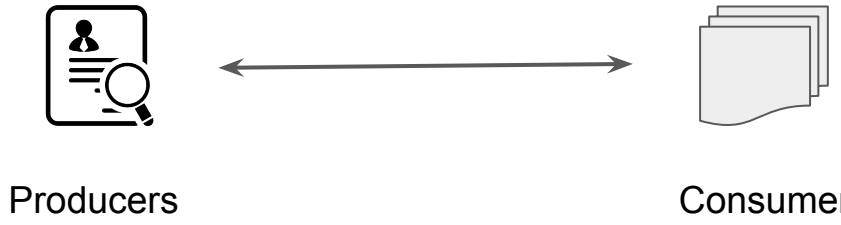
Amazon SQS is a fast reliable, scalable, and fully managed message queuing service.

Amazon SQS makes it simple and quiet cost effective to decouple the components of a specific application.



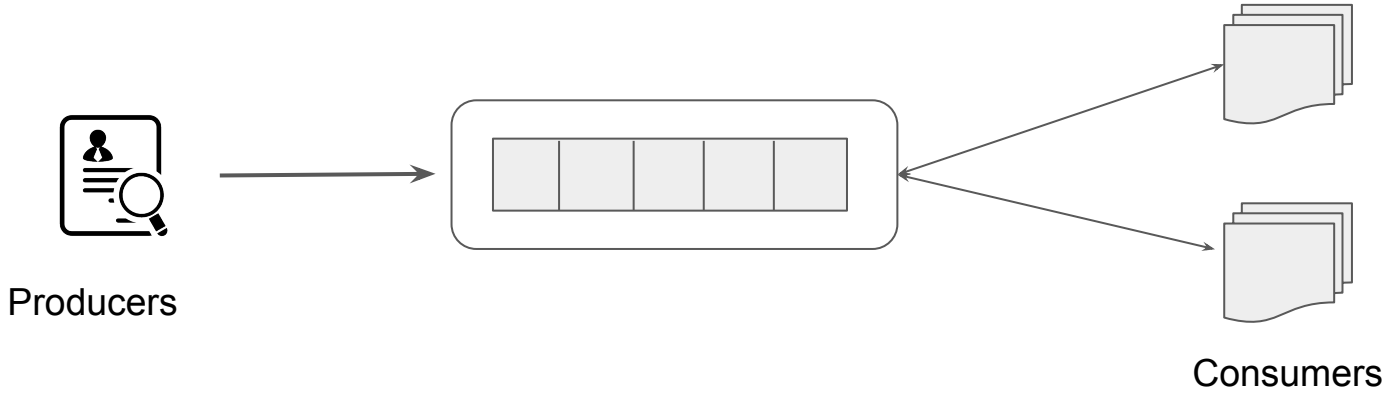
Tightly Coupled Systems

Components of system architecture directly communicate with each other and have hard-dependency on each other.



Loosely Coupled System

Components of system architecture that can process the information without being directly connected.

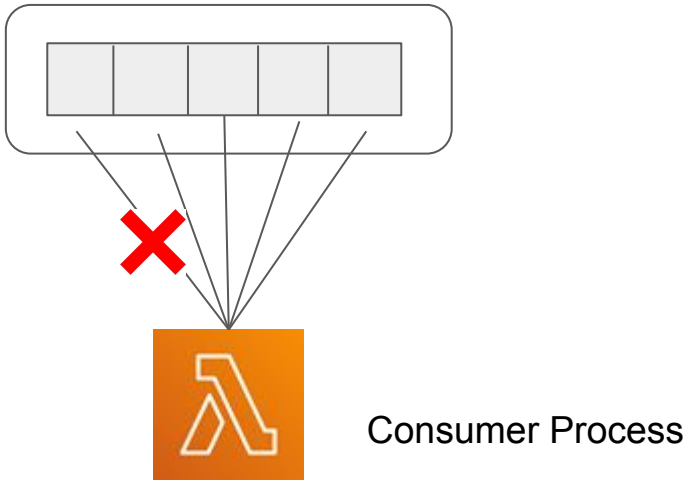


SQS Dead-Letter Queues

Troubleshooting Problematic messages

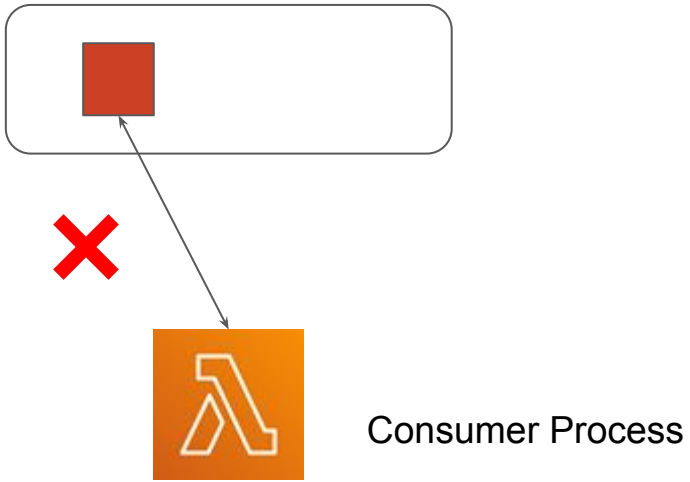
Understanding the Challenge

Amazon SQS supports dead-letter queues, which other queues (source queues) can target for messages that can't be processed (consumed) successfully.



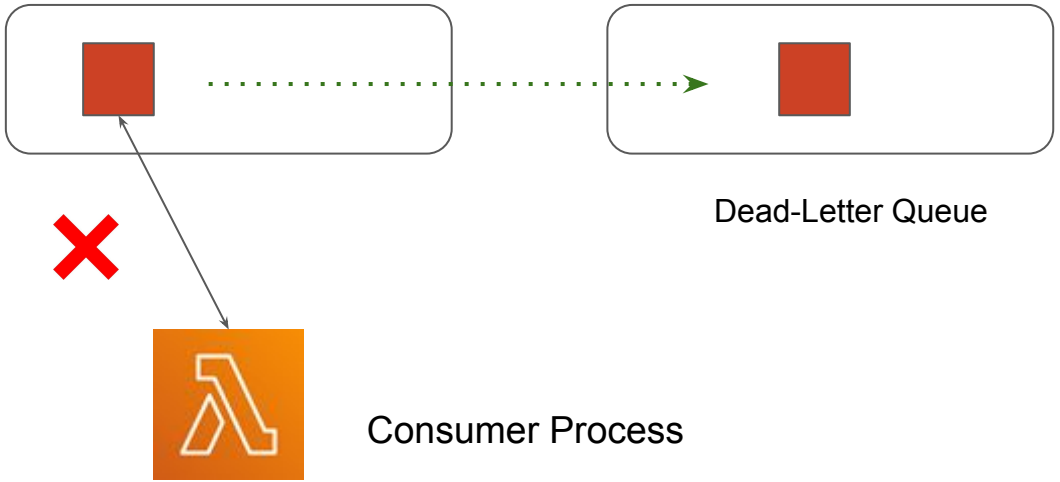
Understanding the Challenge

Amazon SQS supports dead-letter queues, which other queues (source queues) can target for messages that can't be processed (consumed) successfully.



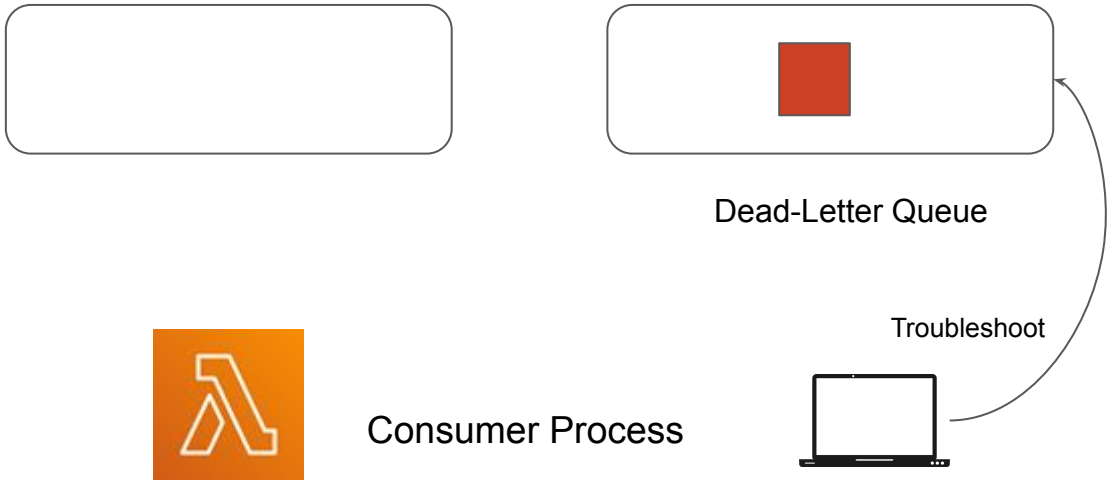
Moving to Dead-Letter Queue

Move the message that cannot be processed to dead letter queue.



Moving to Dead-Letter Queue

Move the the message that cannot be processed to dead letter queue.



Overview of Dead Letter Queue

Amazon SQS supports dead-letter queues, which other queues (source queues) can target for messages that can't be processed (consumed) successfully

Dead-letter queues are useful for debugging your application or messaging system because they let you isolate problematic messages to determine why their processing doesn't succeed.

The messages are sent to the dead letter queue after exceeding maximum receives.

Important Pointers to Remember

When a message moves to a dead-letter queue, the timestamp remains unchanged.

Let's understand this with an example:

- Message has been in the source queue for 1 day and moved to dead-letter queue.
- Message Retention Period in Dead Letter Queue is 4 days.
- Message will be deleted from the Dead Letter queue after 3 days.

Best practice is to have higher retention period for dead-letter queues then the source queue.

Relax and Have a Meme Before Proceeding

me: i'll do it at 6

time: 6:05

me: wow looks like i gotta wait til 7 now

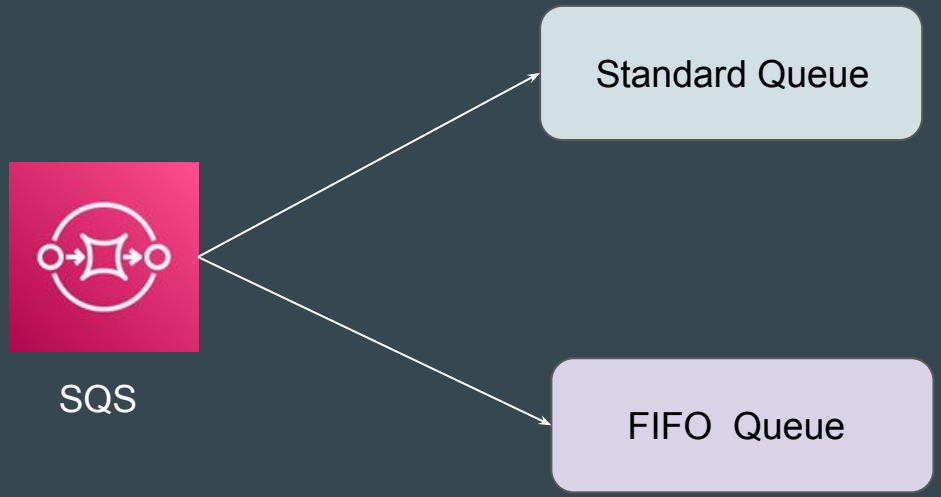


Amazon SQS queue types



Types of SQS Queue

There are two primary types of SQS queues.



Message Ordering

Standard Queue	Occasionally, messages are delivered in an order different from which they were sent.
FIFO Queue	The order in which messages are sent and received is strictly preserved



Standard Queue



FIFO Queue

Difference

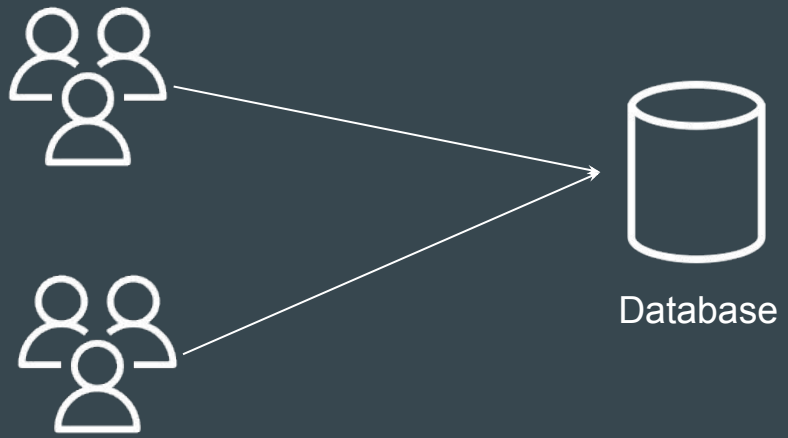
Characteristic	Standard Queue	FIFO Queue
Throughput	Standard queues support a nearly unlimited number of API calls per second, per API action (SendMessage, ReceiveMessage, or DeleteMessage).	FIFO queues support up to 300 API calls per second, per API method (SendMessage, ReceiveMessage, or DeleteMessage).
Delivery	A message is delivered at least once, but occasionally more than one copy of a message is delivered.	A message is delivered once and remains available until a consumer processes and deletes it. Duplicates aren't introduced into the queue.
Ordering	Messages can be out of order.	Messages are in Order.

Message Queues in Database Transactions



Understanding with a Use-Case

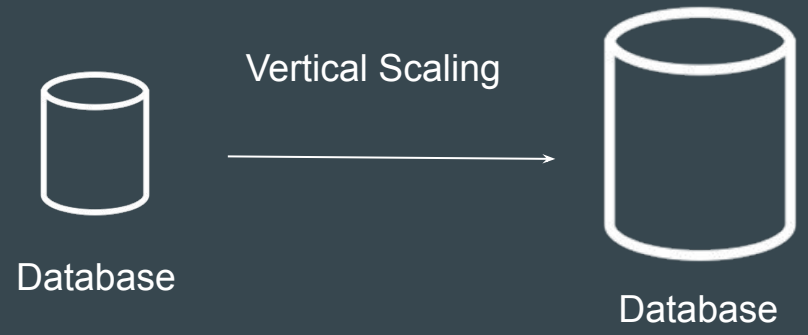
- Let’s assume you have a single database hosted on RDS.
- Due to sales, the number of write transactions has reached 20x normal load.
- Many requests are failing regularly.
- New sale promotions are scheduled every alternate month.



Possible Solution - Vertical Scaling

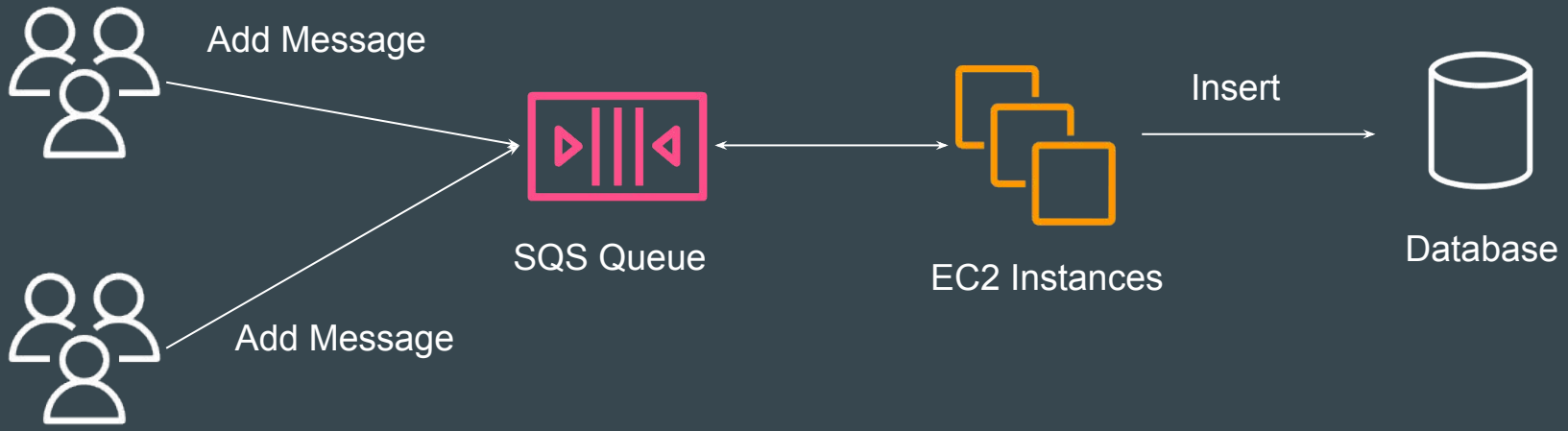
Increase the DB Instance Size of RDS and Provisioned IOPS to handle 20x capacity

Challenge: Downtime + Increased Cost



Better Approach - Add a Queue

In this approach, the messages are temporarily stored in SQS queue which can handle nearly infinite messages.



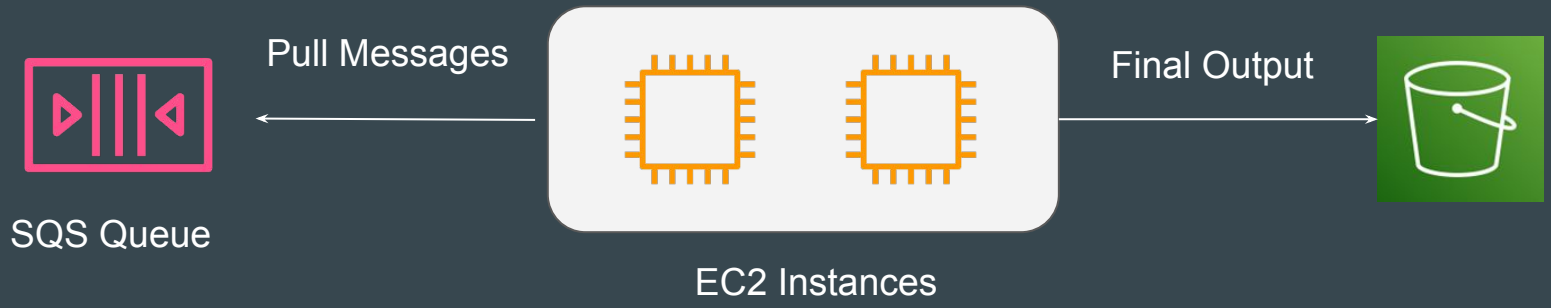
Scaling based on Amazon SQS



Setting the Base

In many scenarios, the number of EC2 instances that are required directly correlates with number of messages in SQS queue.

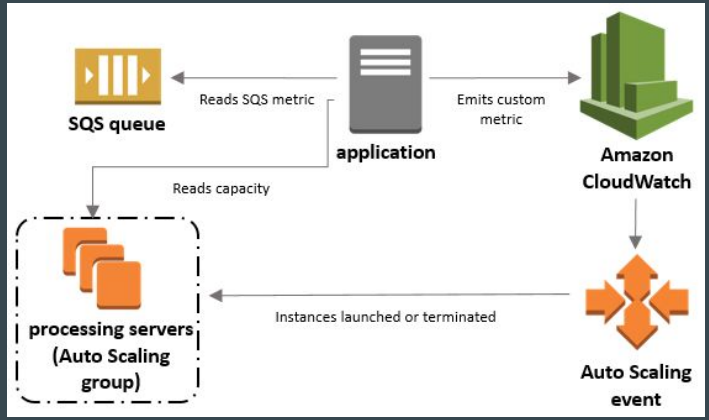
If number of messages increases in SQS, there would be a need to increase EC2 instances.



Scaling based on Amazon SQS

We can configure Auto-Scaling group to launch or terminate EC2 instances based on the number of messages in the SQS queue.

SQS Attribute to check: ApproximateNumberOfMessages



```
PS C:\Users\zealv> aws sqs get-queue-attributes --queue-url https://sqs.ap-southeast-1.amazonaws.com/693331494763/queue-1  --attribute-names ApproximateNumberOfMessages --region ap-southeast-1
{
    "Attributes": {
        "ApproximateNumberOfMessages": "2"
    }
}
```

Amazon MQ



Current Stage

SQS is a simple queuing service.

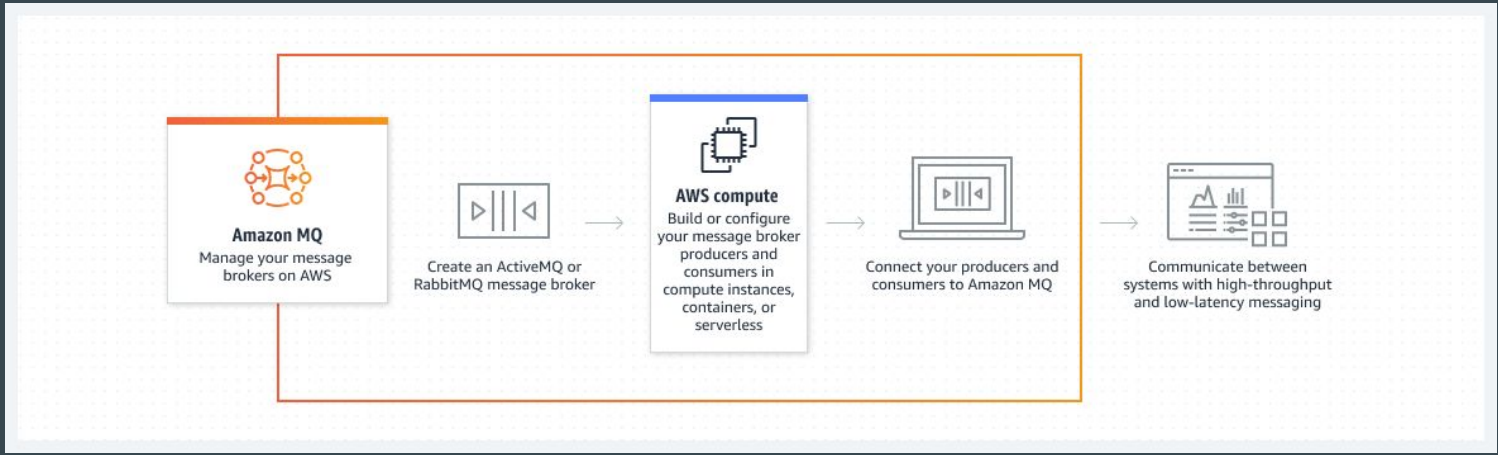
Limited set of functionalities.

There are various Message Broker services like Apache Active MQ that provides many set of features that are extensively used in on-premise organizations.



Moving to MQ

If you're using messaging with existing applications, and want to move your messaging to the cloud quickly and easily, we recommend you consider Amazon MQ.



Pointers to Note

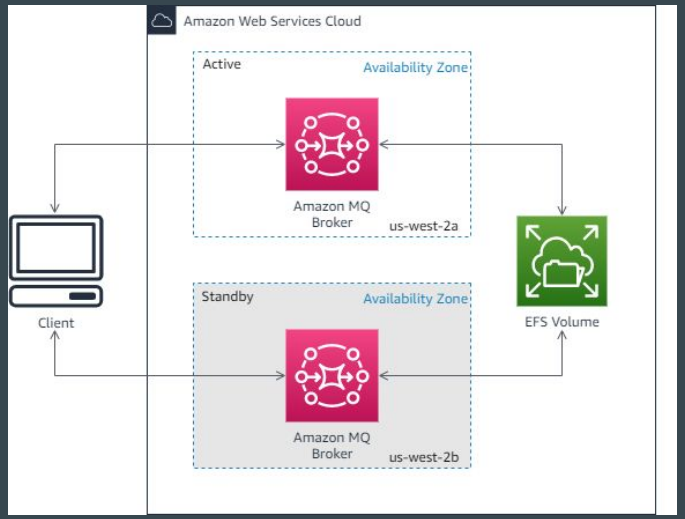
If you are building brand new applications in the cloud, AWS recommends you consider Amazon SQS and Amazon SNS.

Amazon SQS and SNS are lightweight, fully managed message queue and topic services that scale almost infinitely and provide simple, easy-to-use APIs

Active/standby broker for high availability

An active/standby broker is comprised of two brokers in two different Availability Zones, configured in a redundant pair.

Data is written on the shared EFS Volume.



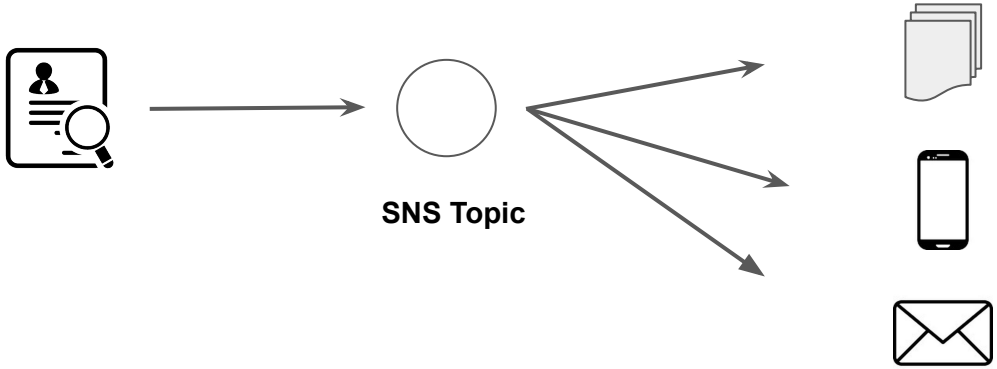
Simple Notification Service

Notification Service

Let's Message

SNS stands for simple notification service.

SNS is a fully managed messaging and mobile notification service for delivering messages to the subscribed endpoints.



Use-Cases for SNS

AWS CloudWatch integrates well with SNS.

Whenever a disk usage of a server exceeds 95%, send an EMAIL and SMS notification to the NOC team.

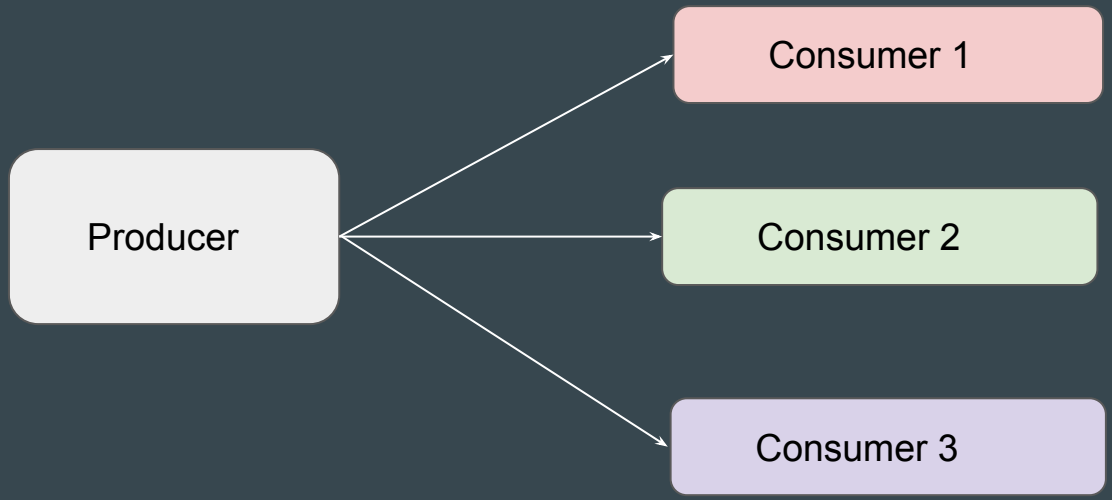
Whenever a server load in production is more than 90%, send and email and SMS notification.

SNS Fanout



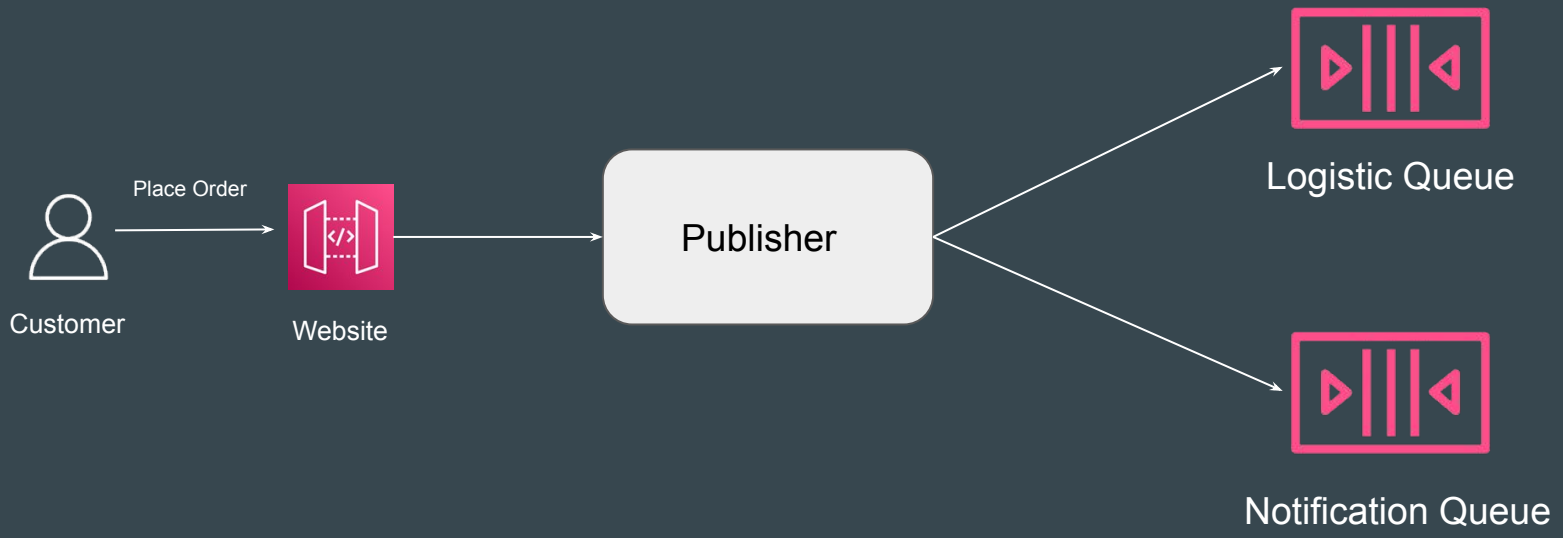
Basics of Fanout Pattern

Fanout is a pattern in which message is delivered to multiple destinations.



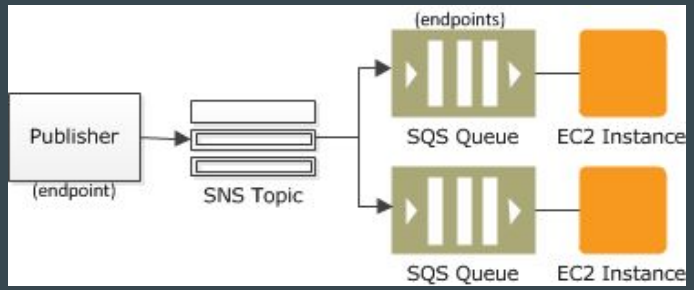
Simple Use-Case: Ordering a Product

Fanout is a pattern in which message is delivered to multiple destinations.



SNS Fanout Pattern

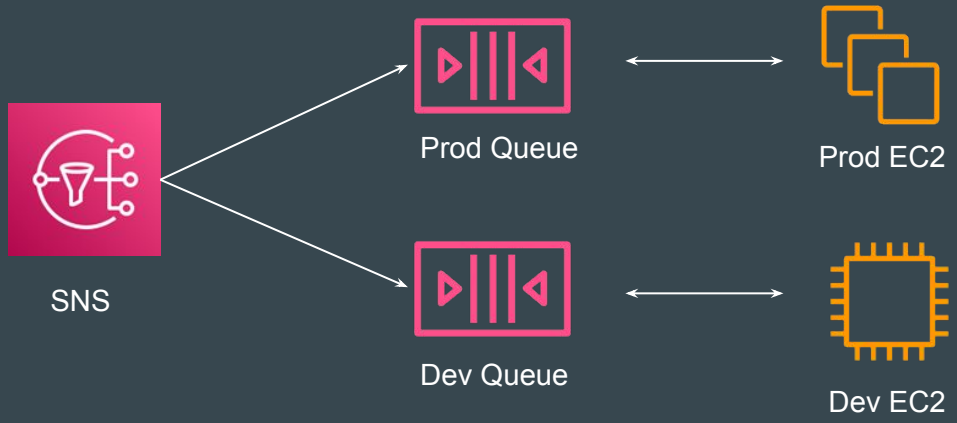
The Fanout scenario is when a message published to an SNS topic is replicated and pushed to multiple endpoints, such as Kinesis Data Firehose delivery streams, Amazon SQS queues, HTTP(S) endpoints, and Lambda functions.



Another Use-Case

You can also use fanout to replicate data sent to your production environment with your test environment

In production, you can attach a new SQS queue for test environment and can continue to improve and test your application using data received from your production environment.

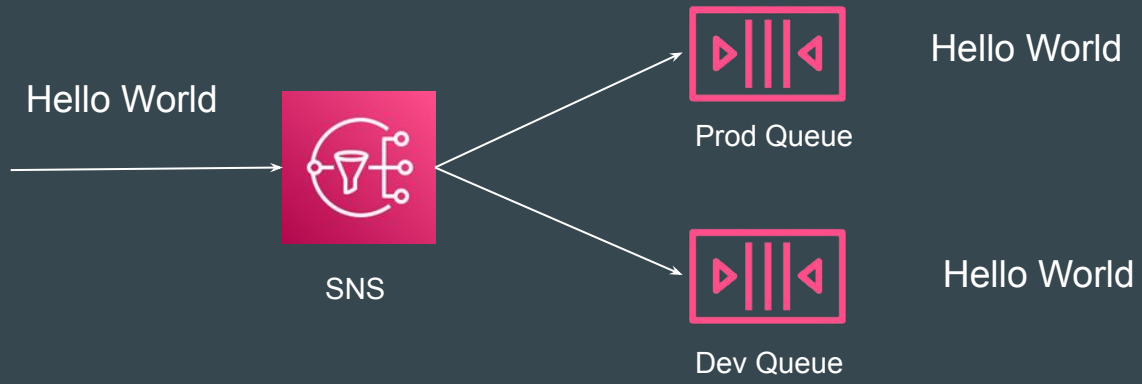


SNS Message Filtering



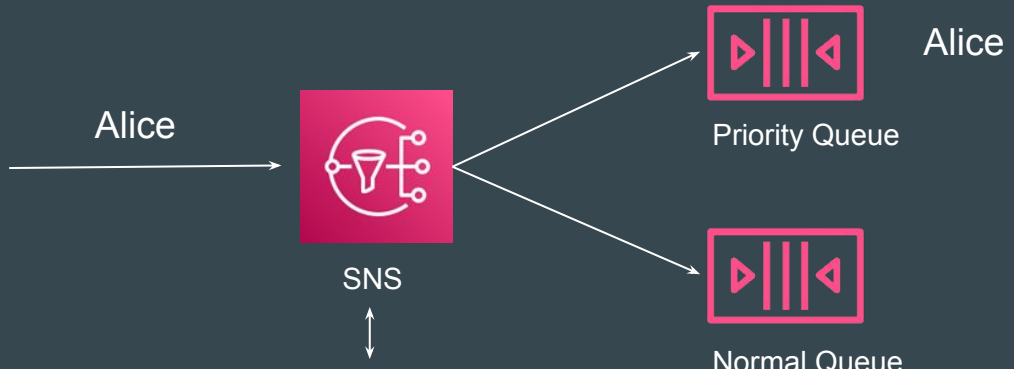
Revising the Basics

By default, an Amazon SNS topic subscriber receives every message that's published to the topic.



Basics of SNS Filtering

A filter policy is a JSON object containing properties that define which messages the subscriber receives



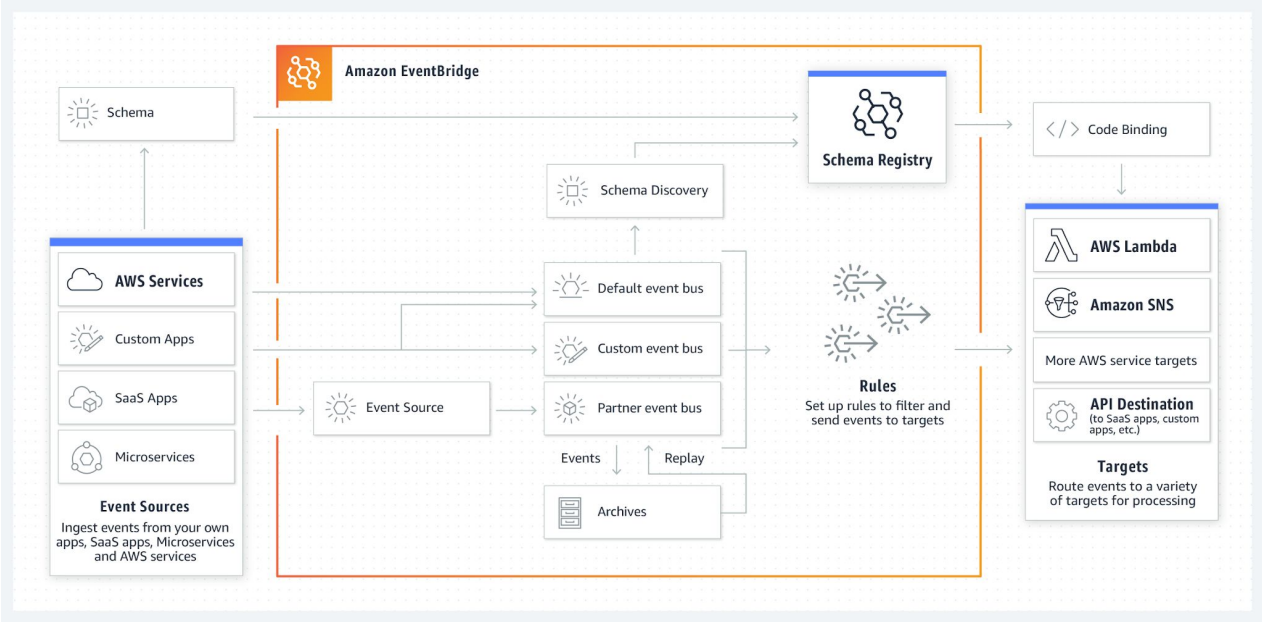
Filter	Queue
Alice	Priority Queue
Bob	Normal Queue

Amazon EventBridge

Connecting Services

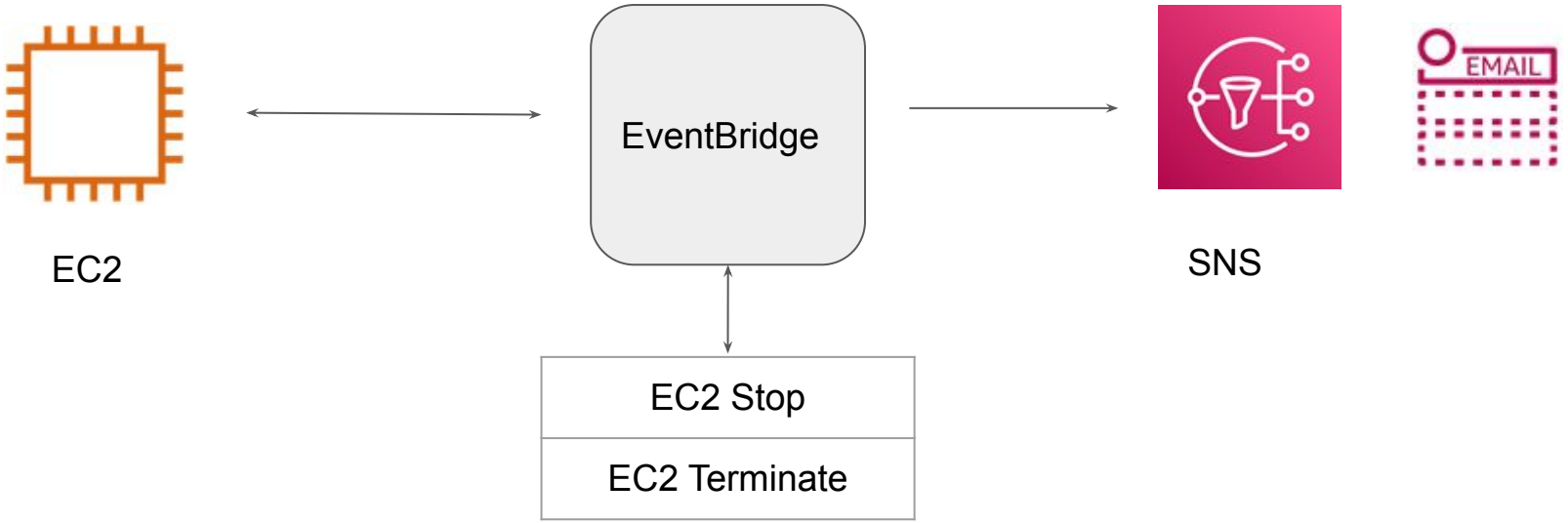
Overview of Amazon Event Bridge

EventBridge delivers a stream of real-time data from event sources to targets.



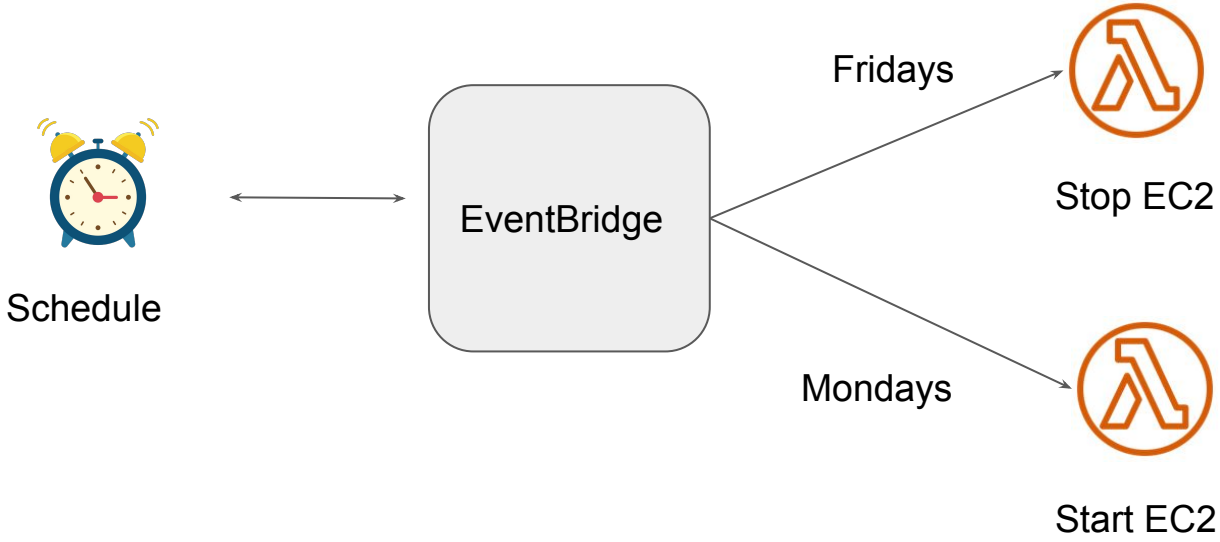
Use-Case 1: EC2 to SNS

Whenever a EC2 instance is stopped, Administrator should be notified.



Use-Case 2: Stop Dev EC2 Instances

Stop all DEV instances at 8PM on Fridays and Start at 9 AM on Mondays.



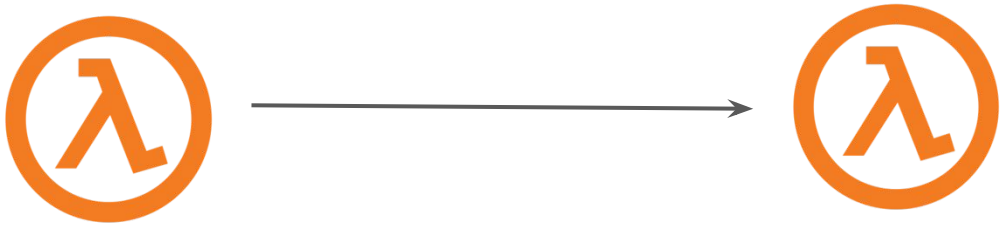
Step Functions

Coordinating across distributed components

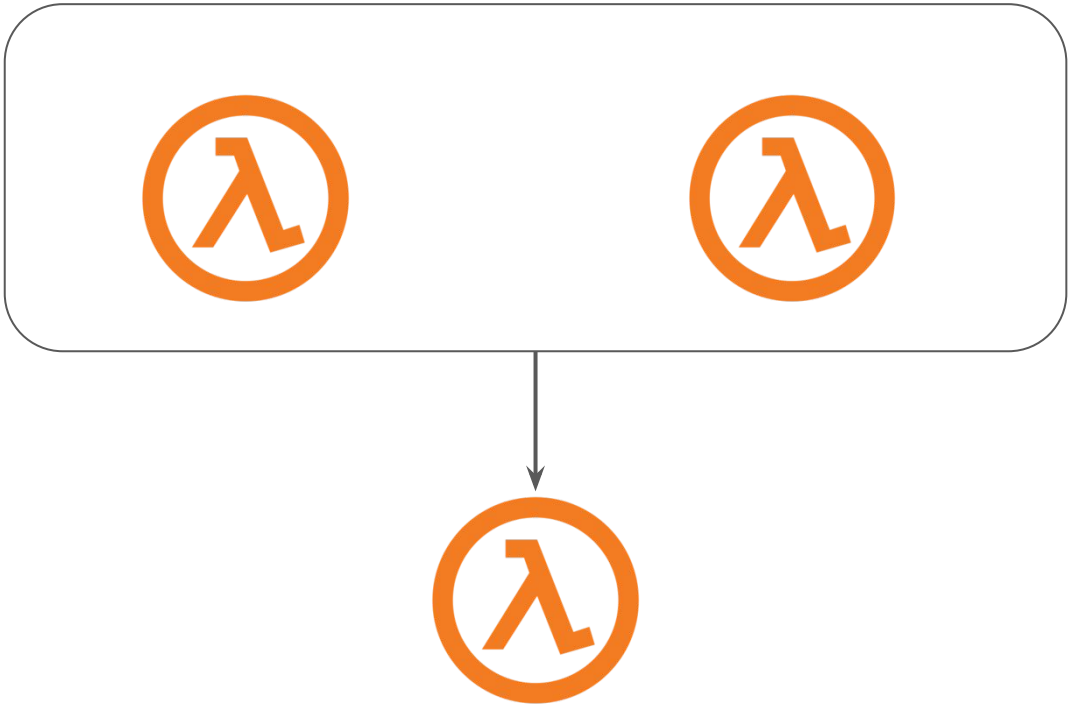
Overview of Step Functions

Step Functions are generally used as an orchestration for serverless functions.

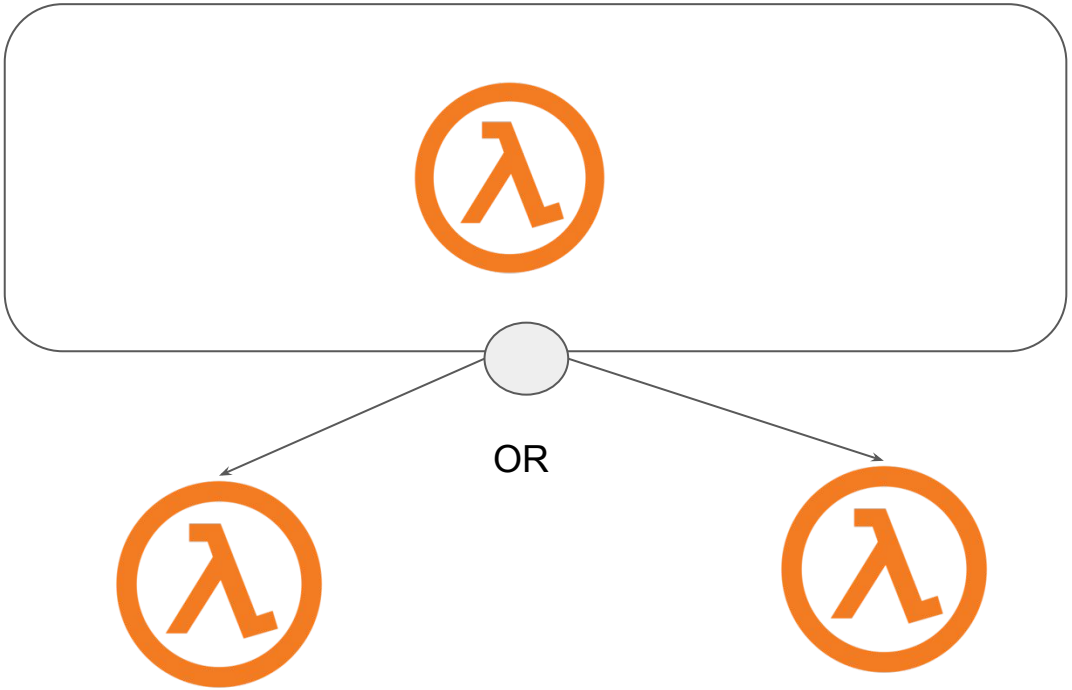
One of the question that comes when you use serverless is, how can we turn serverless into apps ?



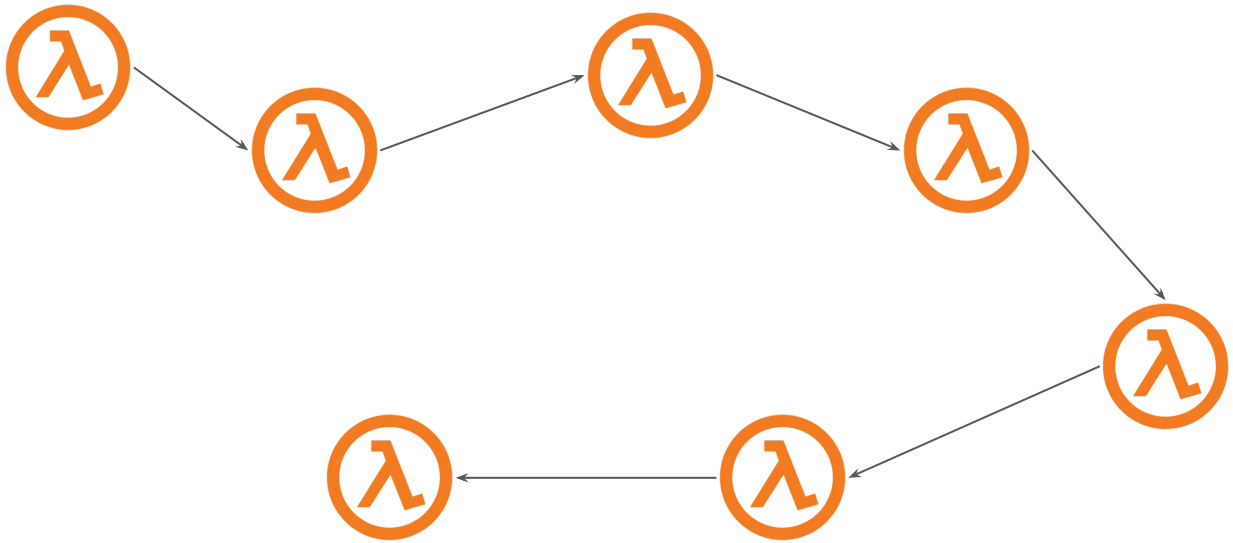
Running Functions in Parallel



Selecting Function Based on Data

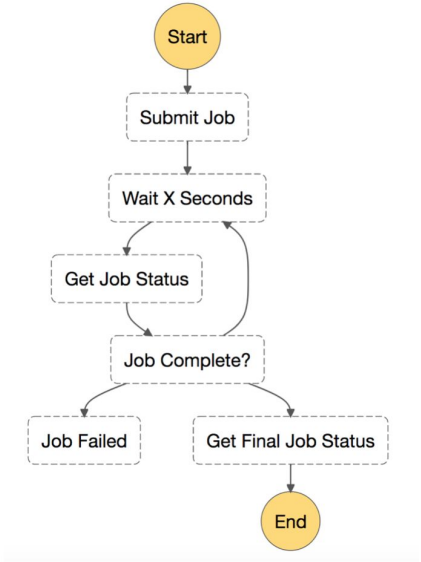


Coordinating Lambda Function



Overview of Step Functions

Step Functions makes it easy to coordinate the components of distributed application using visual workflow.



Simple Email Service (SES)



Understanding the Basics

Amazon SES is an email platform that provides an easy, cost-effective way for you to send and receive email using your own email addresses and domains.

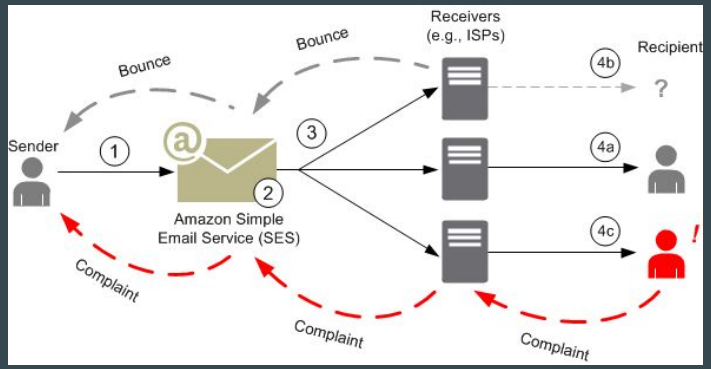
Many organization has generic emails like noreply@example.com which is used to send emails to users upon registration or other use-cases.



How email sending works in Amazon SES

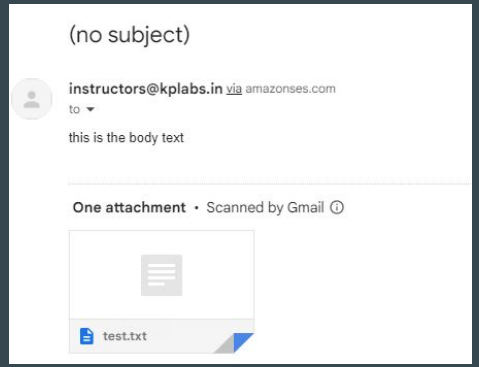
- Email sender makes a request to SES to send email to recipients.
- If the request is valid, SES accepts the email.
- SES sends the message over the Internet to the recipient's receiver.

Bounce Notifications (email not exist) & Complaints (feedback) are sent back to SES which then forwards it to the sender.



Email format in Amazon SES

Email Format	Description
Formatted	Construct simple test message using the form provided.
Raw	For more complex use-cases like using HTML or attachments.



Raw Mail Example

Relax and Have a Meme Before Proceeding



Types of Amazon SES credentials



Understanding the Basics

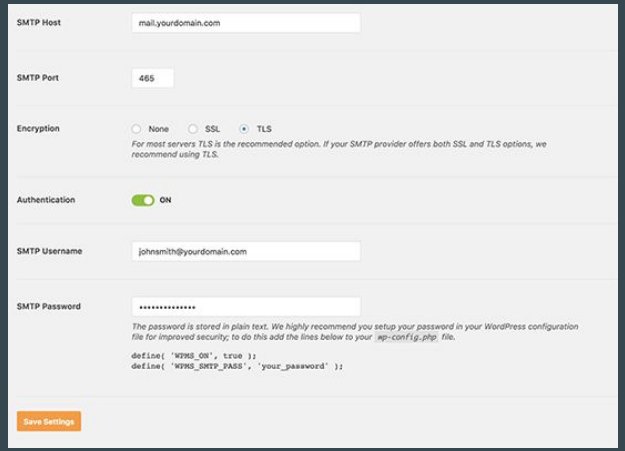
To interact with Amazon SES you use security credentials to verify who you are and whether you have permission to interact with Amazon SES

Access Type	Credentials to Use
Amazon SES API	AWS Access Keys
SES SMTP Interface	Username and Password
SES Console	IAM User and Password

Use-Case: SMTP Interface

There are a number of commercial and open source software packages that support sending email through SMTP

You can configure any such SMTP-enabled software to send email through the Amazon SES SMTP interface.



SMTP Endpoint

Region Name	Region	Endpoint	Protocol
US East (Ohio)	us-east-2	email-smtp.us-east-2.amazonaws.com	SMTP
US East (N. Virginia)	us-east-1	email-smtp.us-east-1.amazonaws.com email-smtp-fips.us-east-1.amazonaws.com	SMTP
US West (N. California)	us-west-1	email-smtp.us-west-1.amazonaws.com	SMTP
US West (Oregon)	us-west-2	email-smtp.us-west-2.amazonaws.com email-smtp-fips.us-west-2.amazonaws.com	SMTP
Asia Pacific (Mumbai)	ap-south-1	email-smtp.ap-south-1.amazonaws.com	SMTP
Asia Pacific (Osaka)	ap-northeast-3	email-smtp.ap-northeast-3.amazonaws.com	SMTP

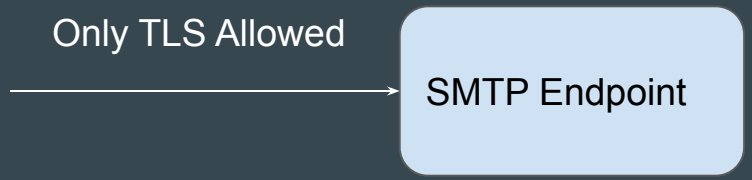
Connecting to an Amazon SES SMTP endpoint



Understanding the Basics

To send email using the Amazon SES SMTP interface, you connect to an SMTP endpoint.

The Amazon SES SMTP endpoint requires that all connections be encrypted using Transport Layer Security (TLS).



Mechanism for TLS

Amazon SES supports two mechanisms for establishing a TLS-encrypted connection

1. STARTTLS

2. TLS Wrapper

Approach 1 - STARTTLS

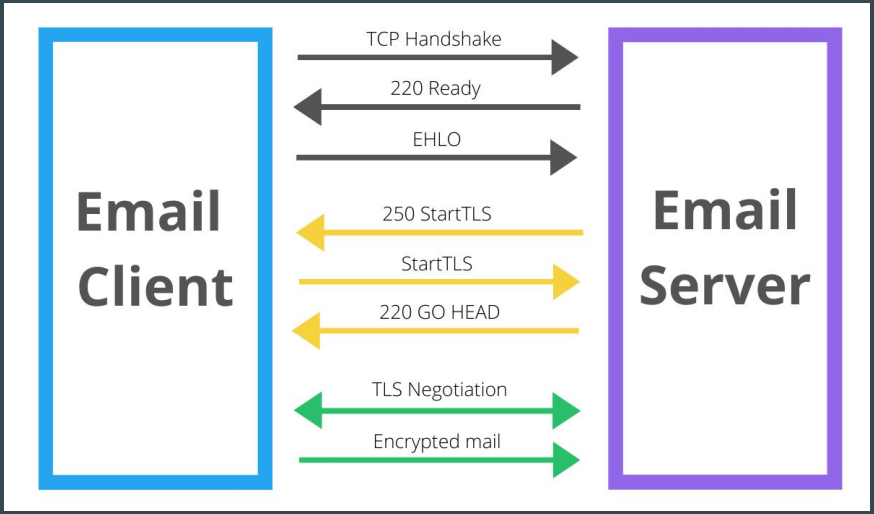
STARTTLS is a means of upgrading an unencrypted connection to an encrypted connection

To set up a STARTTLS connection, the SMTP client connects to the SES SMTP endpoint on port 25, 587, or 2587, issues an EHLO command, and waits for the server to announce that it supports the STARTTLS SMTP extension.

The client then issues the STARTTLS command, initiating TLS negotiation.

When negotiation is complete, the client issues an EHLO command over the new encrypted connection, and the SMTP session proceeds normally.

Overall Flow



Approach 2 - TLS Wrapper

TLS Wrapper is a means of initiating an encrypted connection without first establishing an unencrypted connection.

With TLS Wrapper, the Amazon SES SMTP endpoint doesn't perform TLS negotiation: it's the client's responsibility to connect to the endpoint using TLS, and to continue using TLS for the entire conversation.

To set up a TLS Wrapper connection, the SMTP client connects to the Amazon SES SMTP endpoint on port 465 or 2465.

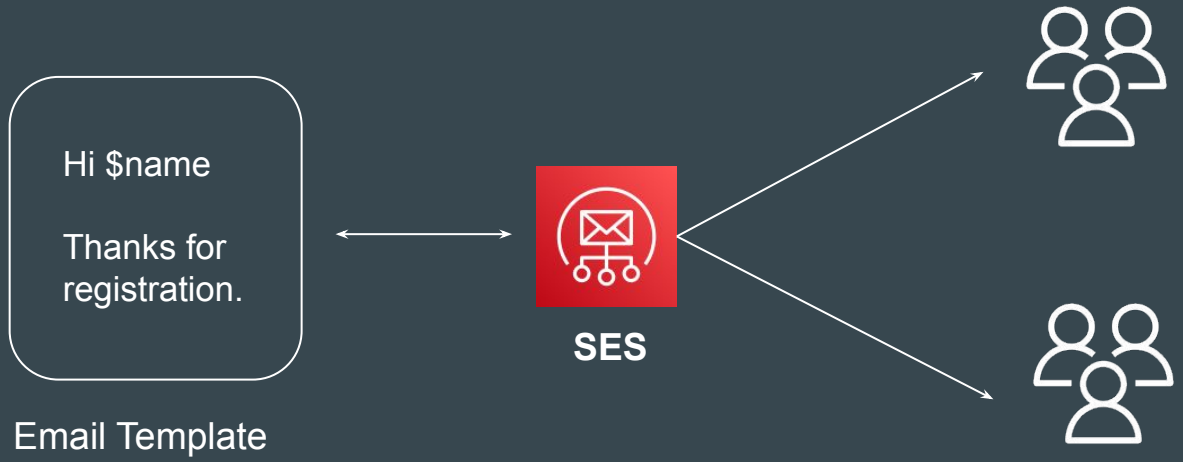
SES Templates



Basics of Email Template

An email template is a pre-defined email layout.

Rather than create a new email from scratch each time, you can use a template as a base.



Using templates in SES

You can use the CreateTemplate API operation to create email templates.

These templates include a subject line, and the text and HTML parts of the email body.

```
C:\Users\zealv\Desktop\tmp\1>aws ses get-template --template-name MyTemplate
{
    "Template": {
        "TemplateName": "MyTemplate",
        "SubjectPart": "Greetings, {{name}}!",
        "TextPart": "Dear {{name}},\r\nYour favorite animal is {{favoriteanimal}}.",
        "HtmlPart": "<h1>Hello {{name}},</h1><p>Your favorite animal is {{favoriteanimal}}.</p>"
    }
}
```

Sending Personalized Email

You can use the SendTemplatedEmail operation to send an email to a single destination.

You can include the values associated with the variables in TemplateData

```
{
    "Source":"Zeal Vora <instructors@kplabs.in>",
    "Template": "MyTemplate",
    "ConfigurationSetName": "ConfigSet",
    "Destination": {
      "ToAddresses": [ "instructors@kplabs.in"
      ]
    },
    "TemplateData": "{ \"name\":\"Zeal\", \"favoriteanimal\": \"Elephant\" }"
  }
```

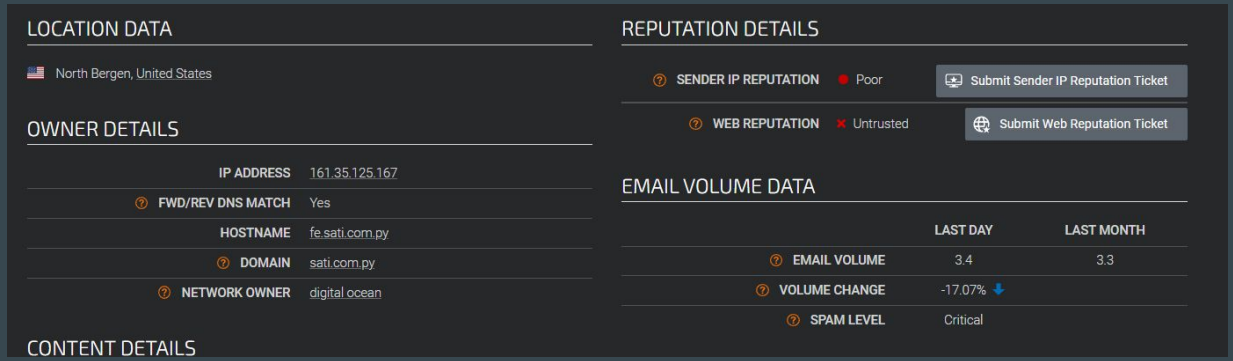
Bring your own IP addresses



Basics of IP Reputation

IP reputation is a measure that helps evaluate the quality of an IP address and determine how legitimate its requests are

Bad IP Reputation generally corresponds to activities like sending spam emails, viruses etc that originate from the IP.



Use-Case: Organization Migrating to Cloud

Organization’s infrastructure is hosted in the on-premise datacenter.

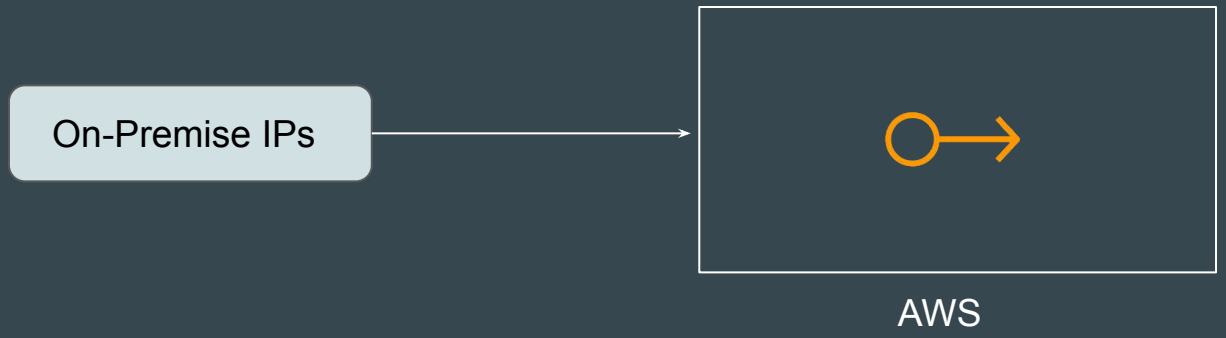
They have certain Public IPs from years with very good reputation.

They decide to migrate to Cloud and server receive IP with NOT as good reputation as their previous IPs.



Introducing Bring Your Own IP

You can bring part or all of your publicly routable IPv4 or IPv6 address range from your on-premises network to your AWS account.



Benefits of Bring Your Own IP

Benefits	Description
IP Reputation	Many customers consider the reputation of their IP addresses to be a strategic asset and want to use those IPs on AWS with their resources.
Customer whitelisting	BYOIP also enables customers to move workloads that rely on IP address whitelisting to AWS without the need to re-establish the whitelists with new IP addresses
Regulation and compliance	Many customers are required to use certain IPs because of regulation and compliance reasons. They too are unlocked by BYOIP.

Important Requirements - Part 1

The address range must be registered with your regional internet registry (RIR) such as ARIN, RIPE, APNIC.

It must be registered to a business or institutional entity and cannot be registered to an individual person.

The most specific IPv4 address range that you can bring is /24.

The most specific IPv6 address range that you can bring is /48 for CIDRs that are publicly advertised, and /56 for CIDRs that are not publicly advertised.

Important Requirements - Part 2

The addresses in the IP address range must have a clean history. AWS might investigate the reputation of the IP address and reserve the right to reject an IP address range an IP has a poor reputation or is associated with malicious behavior.

Points to Note

Customers can create Elastic IPs from the IPv4 space they bring to AWS and use them with EC2 instances, NAT Gateways, and Network Load Balancers.

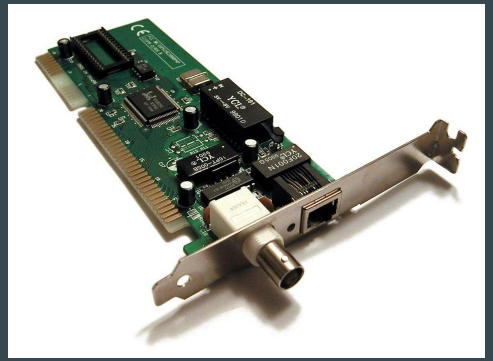
Elastic Network Interface (ENI)



Revising Basics of Network Interface

Network interface is a hardware component that connects a computer to a computer network

A virtual network interface (VIF) is an abstract virtualized representation of a computer network interface.



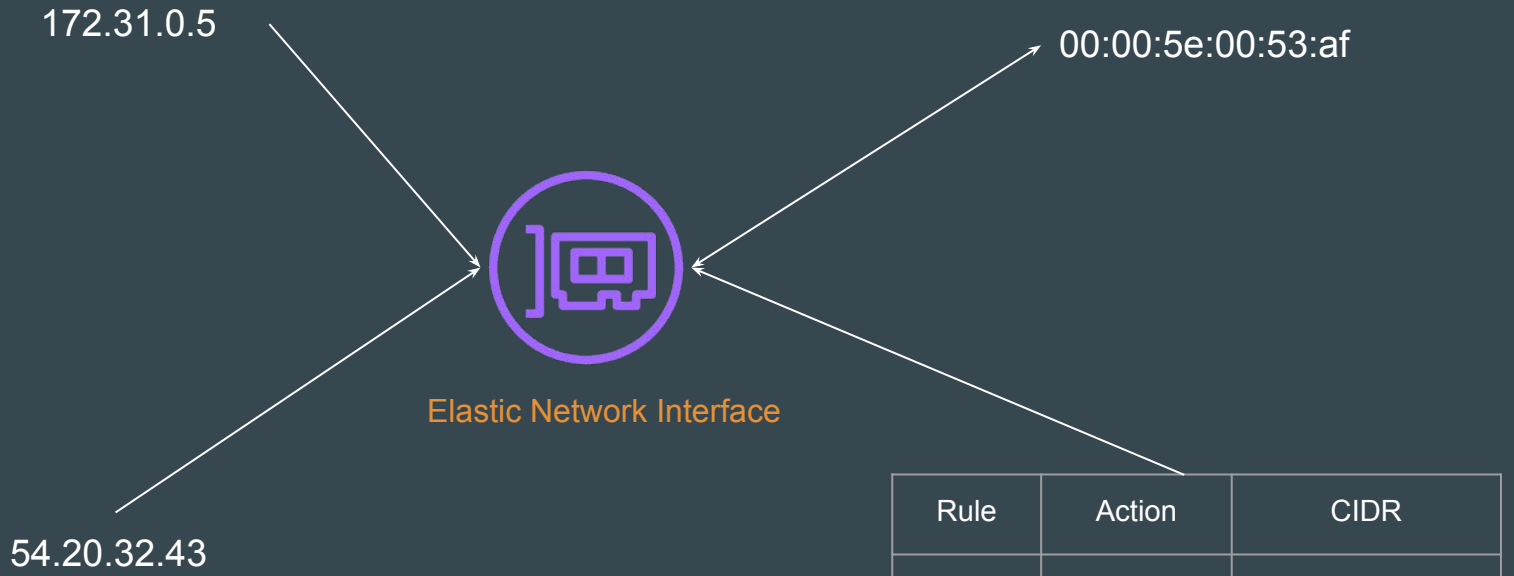
Elastic network interfaces

An elastic network interface is a logical networking component in a VPC that represents a virtual network card.

Some of the following attributes include:

- A primary private IPv4 address
- One or more secondary private IPv4 addresses
- One Elastic IP address (IPv4) per private IPv4 address
- One or more security groups
- A MAC address
- A source/destination check flag

Sample Attributes of ENI

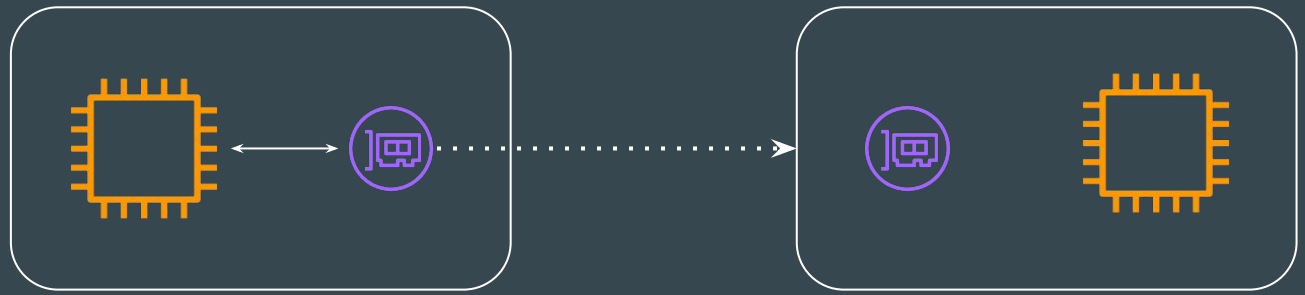


Rule	Action	CIDR
Port 22	ALLOWED	192.168.0.5/32

Portable NICs

You can create a network interface, attach it to an instance, detach it from an instance, and attach it to another instance.

The attributes of a network interface follow it as it's attached or detached from an instance and reattached to another instance.



Importance of Default NICs

Each instance has a default network interface, called the primary network interface. You cannot detach a primary network interface from an instance.

You can create and attach additional network interfaces.

The maximum number of network interfaces that you can use varies by instance type.

NICs are availability zone specific.

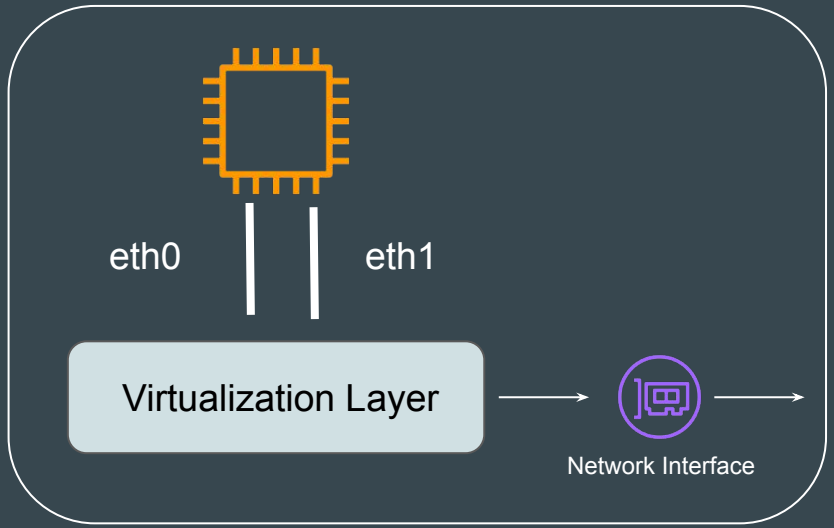
Enhanced Networking



Understanding the Basics

Every network interface card has a specific bandwidth.

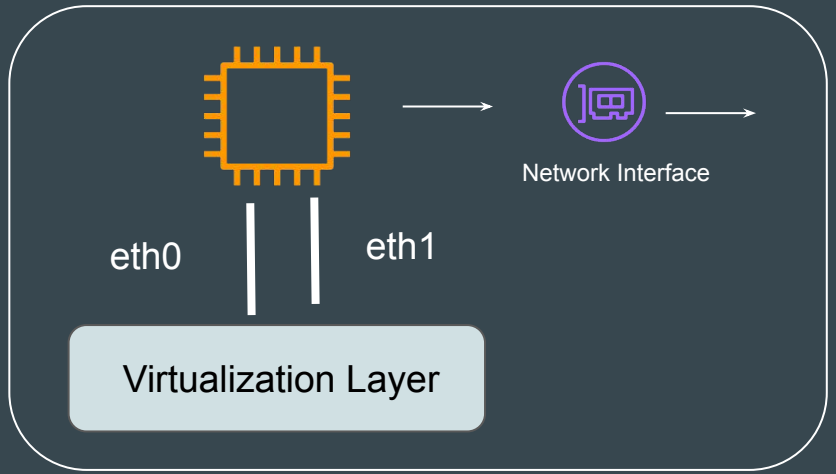
The networking bandwidth in-turn gets affected when we virtualization layer comes into picture.



Basics of Enhanced Networking

Enhanced Networking uses single root I/O virtualization technique (SR-IOV) to provide high performance networking capabilities on supported instance types.

SR-IOV is a method of device virtualization that provides higher I/O performance and lower CPU utilization when compared to traditional virtualized network interfaces.



Mechanism to Enable Enhanced Networking

All current generation instance types support enhanced networking, except for T2 instances.

You can enable enhanced networking using one of the following mechanisms:

Approach	Description
Elastic Network Adapter (ENA)	Supports network speeds of up to 100 Gbps for supported instance types.
Intel 82599 Virtual Function (VF) interface (ixgbevf driver)	Supports network speeds of up to 10 Gbps for supported instance types.

Instance and Supported Mechanism

Depending on the instance type, the supported mechanism to enable Enhanced Networking changes.

Instance type	EBS only	NVMe EBS	Instance store	Placement group	Enhanced networking
C4	Yes	No	No	Yes	Intel 82599 VF
C5	Yes	Yes	No	Yes	ENA
C5a	Yes	Yes	No	Yes	ENA
C5ad	No	Yes	NVMe *	Yes	ENA
C5d	No	Yes	NVMe *	Yes	ENA
C5n	Yes	Yes	No	Yes	ENA
C6a	Yes	Yes	No	Yes	ENA
C6g	Yes	Yes	No	Yes	ENA
C6gd	No	Yes	NVMe *	Yes	ENA
C6gn	Yes	Yes	No	Yes	ENA
C6i	Yes	Yes	No	Yes	ENA

Verify if Module is Used in a Interface

ethtool -i eth0

```
[ec2-user@ip-172-31-19-108 ~]$ ethtool -i eth0
driver: ixgbevf
version: 5.10.157-139.675.amzn2.x86_64
firmware-version:
expansion-rom-version:
bus-info: 0000:00:03.0
supports-statistics: yes
supports-test: yes
supports-eeprom-access: no
supports-register-dump: yes
supports-priv-flags: yes
```

Intel VF

```
[ec2-user@ip-172-31-40-246 ~]$ ethtool -i eth0
driver: ena
version: 2.8.0g
firmware-version:
expansion-rom-version:
bus-info: 0000:00:05.0
supports-statistics: yes
supports-test: no
supports-eeprom-access: no
supports-register-dump: no
supports-priv-flags: yes
```

ENA

Placement Groups

Time to go fast

Placement Groups

- Placement group are recommended for applications that require low latency, high network throughput.

- Placement groups can also be used to influence placement of a group of EC2 instances.

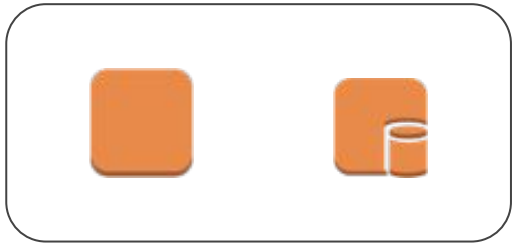


Small Road vs Highway





Let's understand GUI way



Placement Group

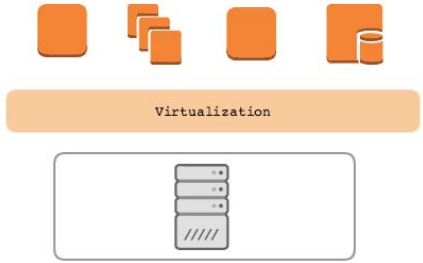


Point 2 - Influencing Placement of EC2

- A single server can run multiple virtual machines.
- This can lead to issues if you are running a cluster of servers.

Example:

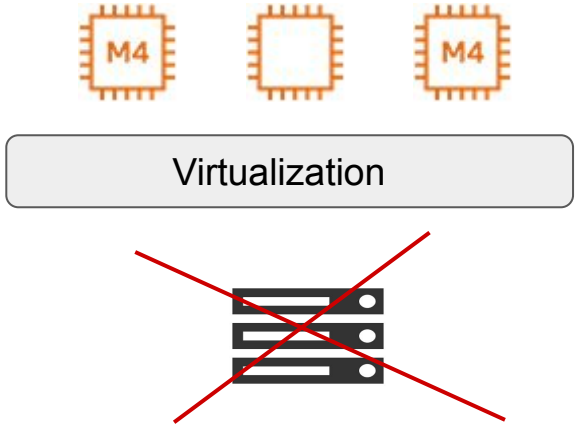
- Medium Corp is running a MySQL cluster consisting of two servers in single AZ. In the background, both the EC2 are part of the same underlying host.



Example Use-Case

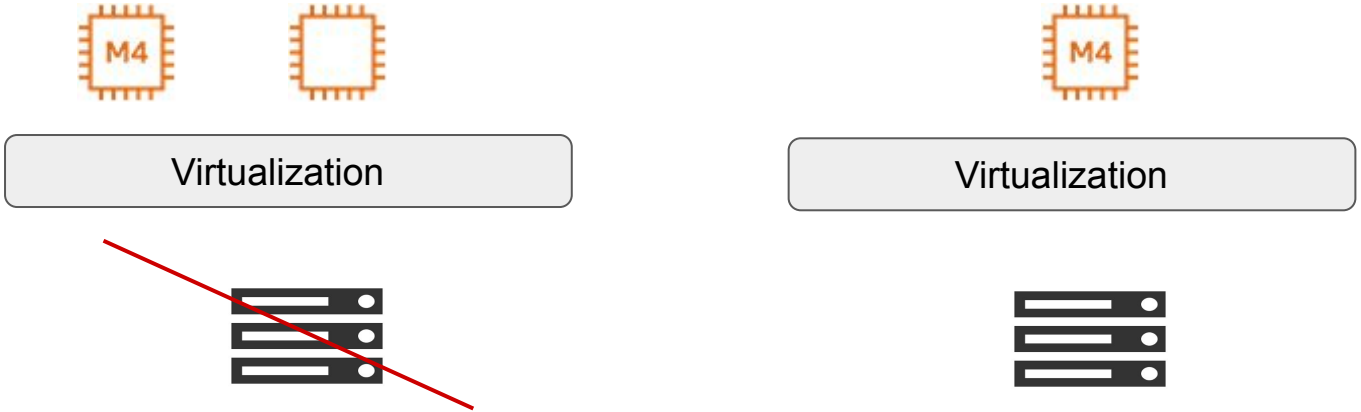
Medium Corp is running a MySQL cluster consisting of two servers in single AZ. The server are of type m4.large.

In the background, both the EC2 are part of the same underlying host.



Solution - Placement Group

With placement group, we can explicitly specify that two EC2 instance should not be part of the same server (same rack of servers)



Racks in Data Center



Types of Placement Groups

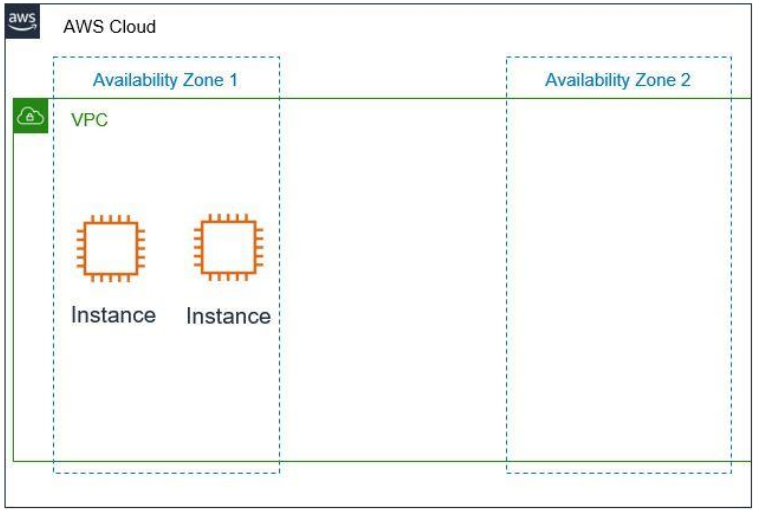
There are three types of placement groups available:

Sr No	Type	Description
1	Cluster	Packs instances close to each other in an Availability Zone.
2	Partition	Spreads instances in logical partition such that group of instances in one partition do not share underlying hardware.
3	Spread	Strictly places group of instances across distinct hardware to reduce failures.

Cluster Placement Groups

Logical grouping of instances within a single Availability Zone.

Intended for applications that require low network latency and high network throughput.

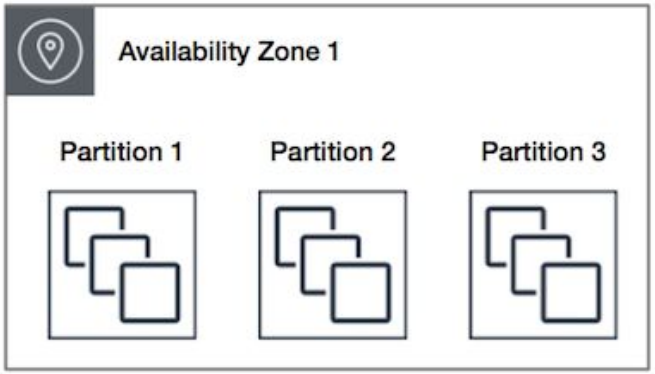


Partition Placement Groups

AWS ensures that each partition within a placement group has its own set of racks.

In the below diagram, there are 3 partition and each partition has multiple EC2 instances.

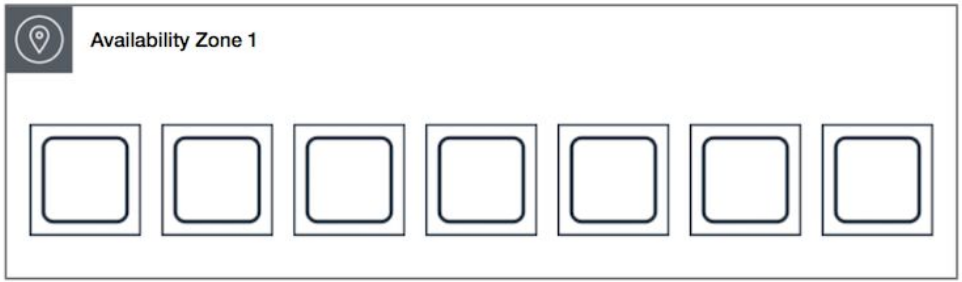
Each of these partition resides in a different rack inside the Data center.



Spread Placement Group

A spread placement group is a group of instances that are each placed on distinct racks, with each rack having its own network and power source.

In the following diagram, there are 7 EC2 instances and each instance is in a separate rack.



Important Points - Cluster Placement Groups

- A cluster placement group can't span multiple Availability Zones.
- Only specific types of EC2 instances can be launched.
- Maximum network throughput traffic between two instance in placement group is limited by the slower of the two instance.
- Recommended to launch all instance together. Launching instance later can lead to capacity errors. In such-case, stop and start all instances in the placement group.

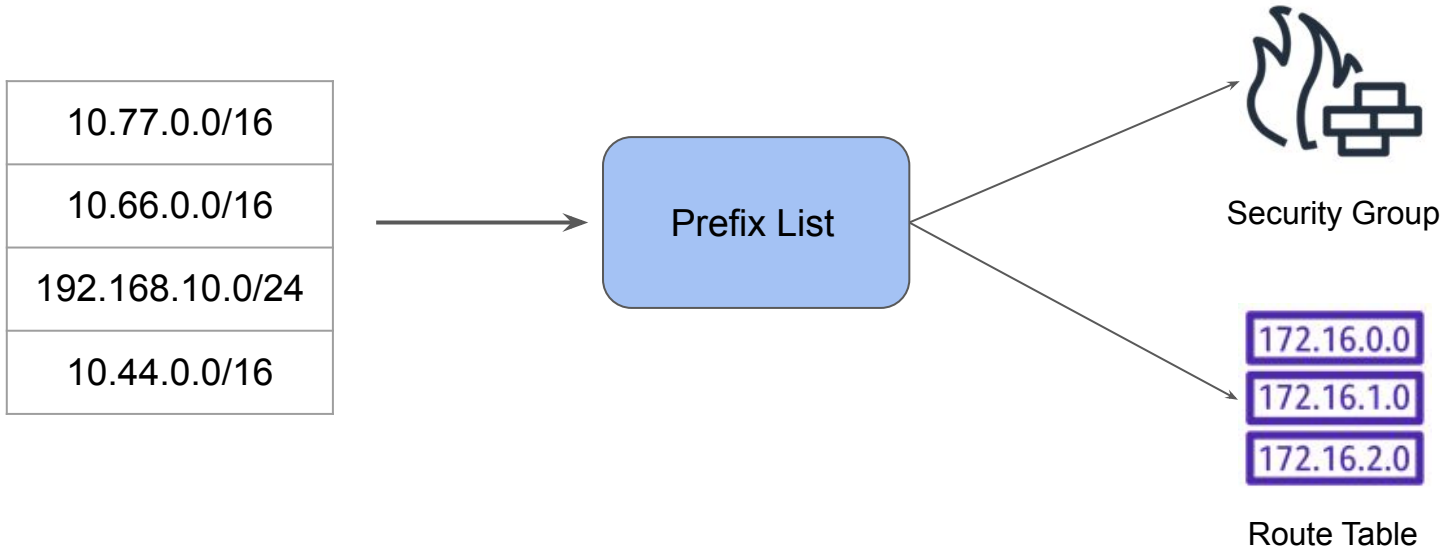
Prefix Lists

Centralizing IP Address Data

Overview of Prefix Lists

A prefix list is a set of one or more CIDR blocks.

You can create a prefix list from the IP addresses that you frequently use, and reference them as a set in security group rules and routes instead of referencing them individually.



Types of Prefix List

There are two types of prefix lists:

Types of Prefix List	Description
Customer-managed prefix lists	Sets of IP address ranges that you define and manage.
AWS-managed prefix lists	Sets of IP address ranges for AWS services.

Important Pointers

A prefix list supports a single type of IP addressing only (IPv4 or IPv6). You cannot combine IPv4 and IPv6 CIDR blocks in a single prefix list.

A prefix list applies only to the Region where you created it.

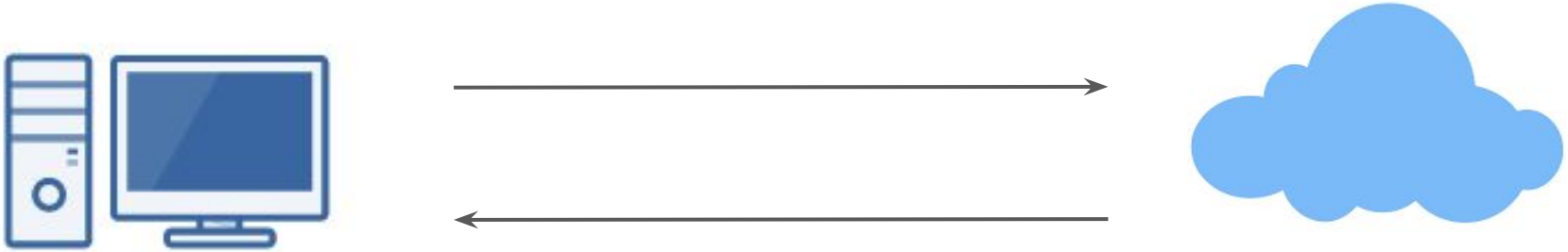
When you reference a prefix list in a resource, the maximum number of entries for the prefix lists counts against the quota for the number of entries for the resource. For example, if you create a prefix list with 20 maximum entries and you reference that prefix list in a security group rule, this counts as 20 security group rules.

Virtual Private Network

Let's Route

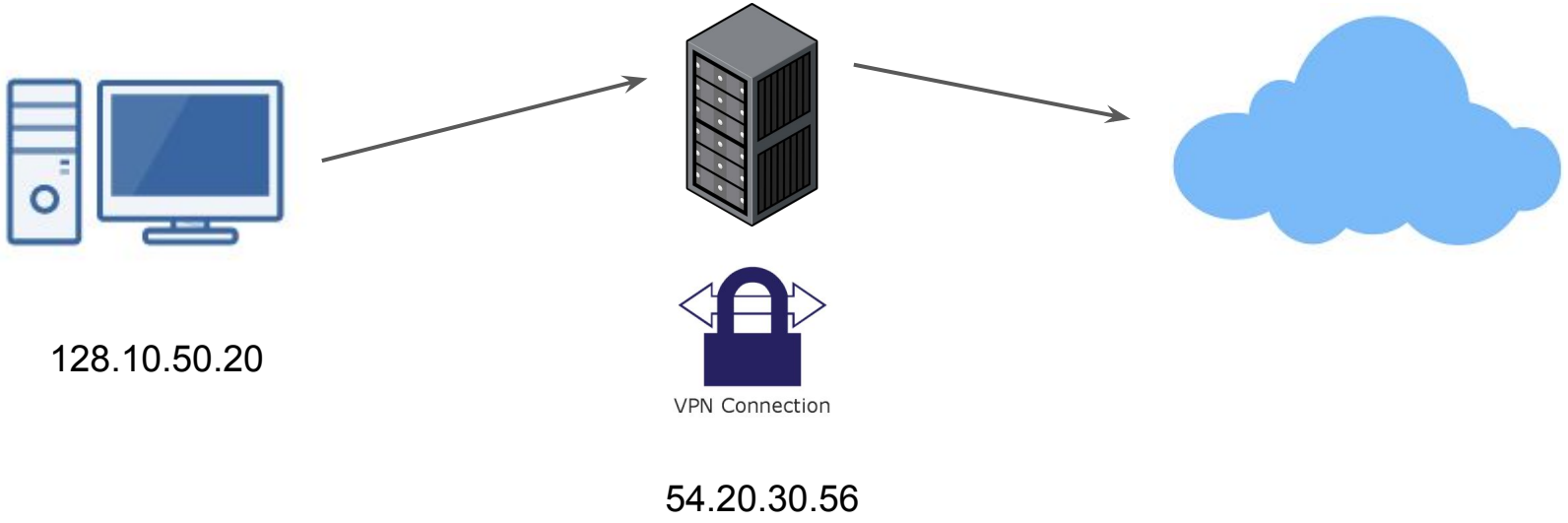
VPN

- VPN enables you to route traffic from yourself towards destination through itself.
- Something similar to Proxy.



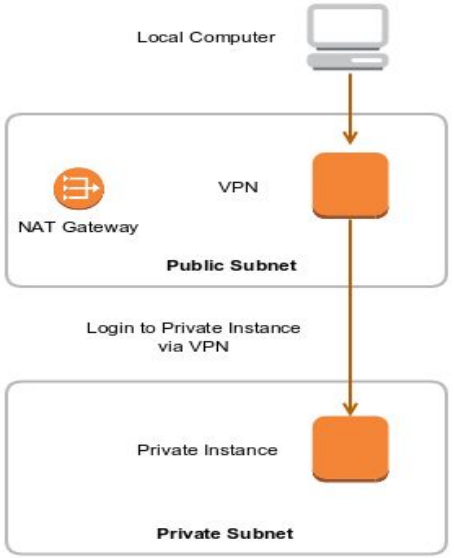
128.10.50.20

Routing via VPN Server



VPN use in Corporate Network

- In Corporate environments, VPN is used to connect to instances in Private Subnet.
- VPN Server resides in the Public Subnet and you route your traffic via VPN server to instances in Public Subnet.

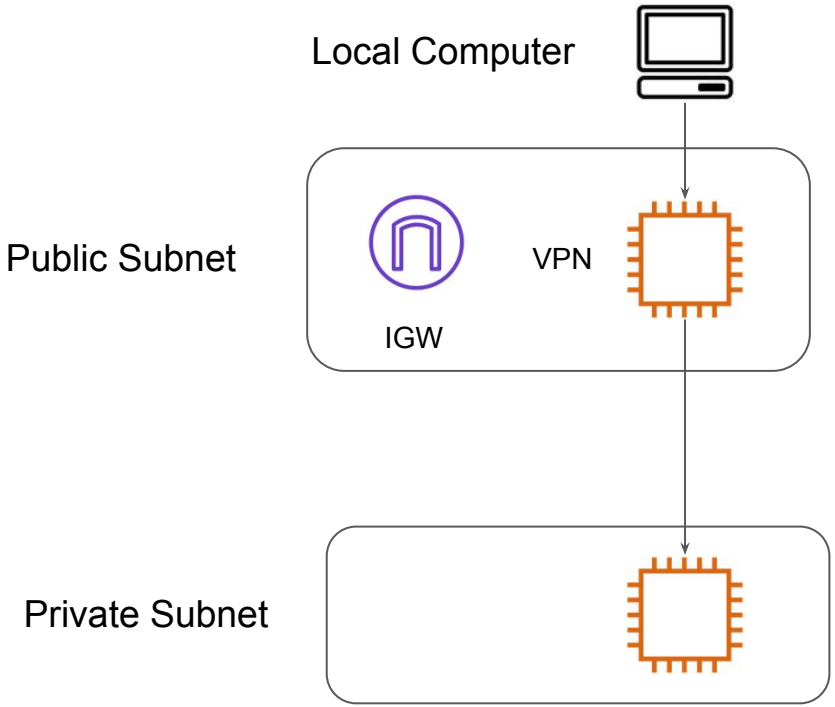


AWS Client VPN

Creating our First VPN in AWS

EC2 Based VPN Architecture

In this approach, you install VPN softwares like OpenVPN in the EC2 instance and use it to route traffic to private subnets.

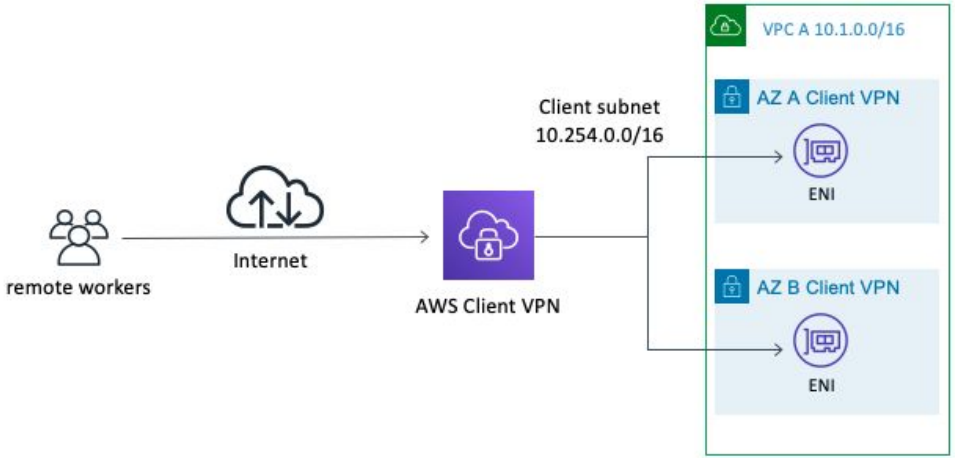


Challenges with EC2 VPN Based Architectures

1. High-Availability (What if VPN EC2 goes down)
2. Patch Management.
3. Upgrade of VPN Software
4. Performance Optimization
5. VPN Server Configuration

AWS Client VPN

AWS Client VPN is a managed client-based VPN service that enables you to securely access your AWS resources and resources in your on-premises network.



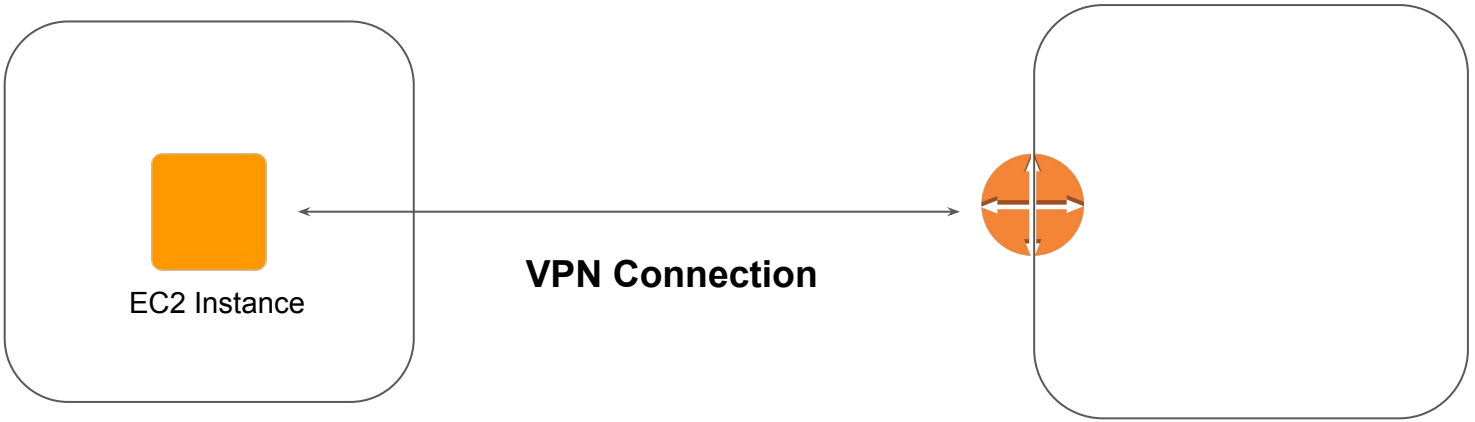
Site to Site Tunnel

Let's Route

Site to Site VPN

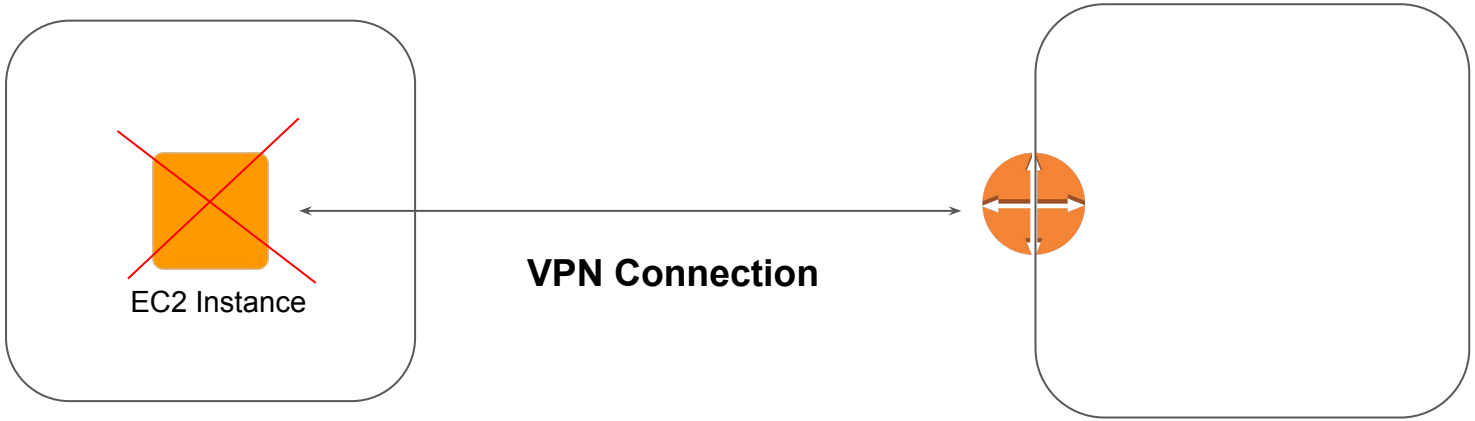
A Site to Site (S2S) VPN allows two networking domains to communicate securely between each other over an untrusted network like Internet.

The two sites can be AWS and on-premise data-center or even two different VPC's.

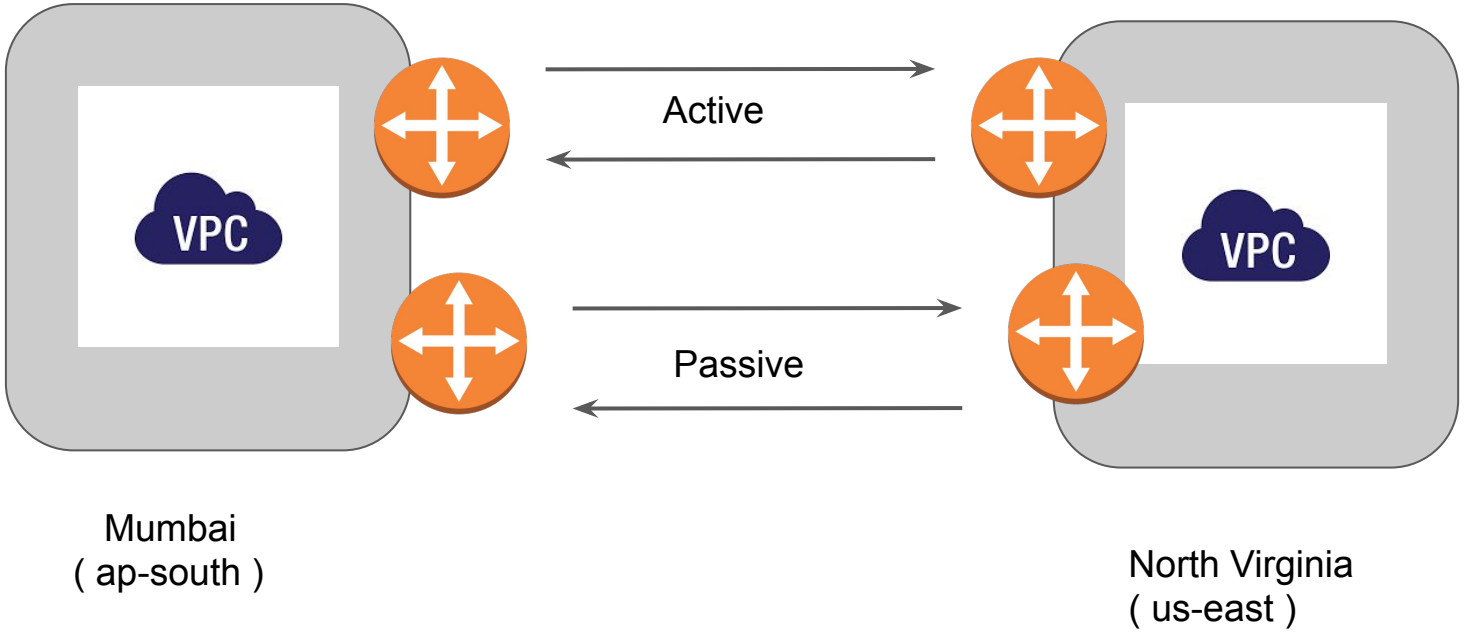


Availability Challenges in S2S VPN

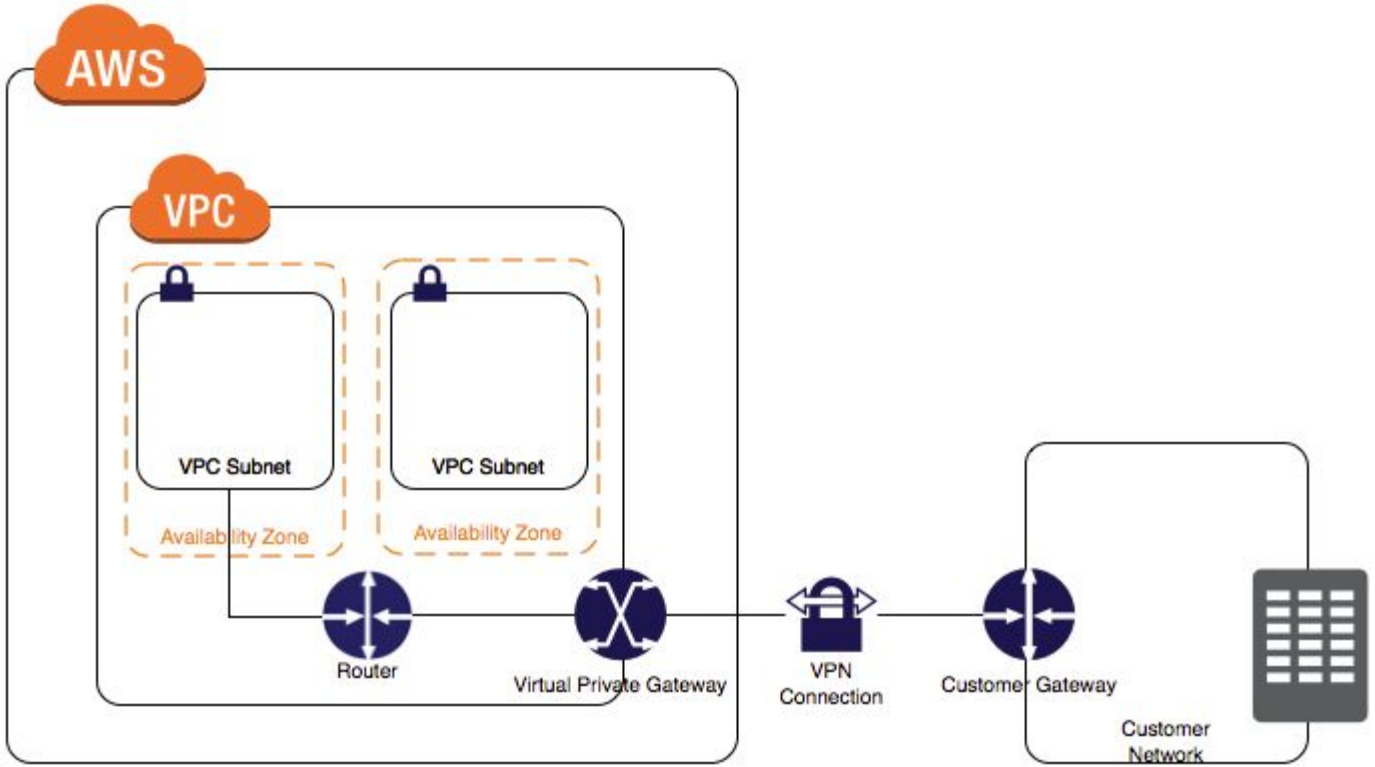
If you have a single tunnel endpoint and if one of the side goes down, then the entire tunnel breaks.



High Availability in S2S VPN

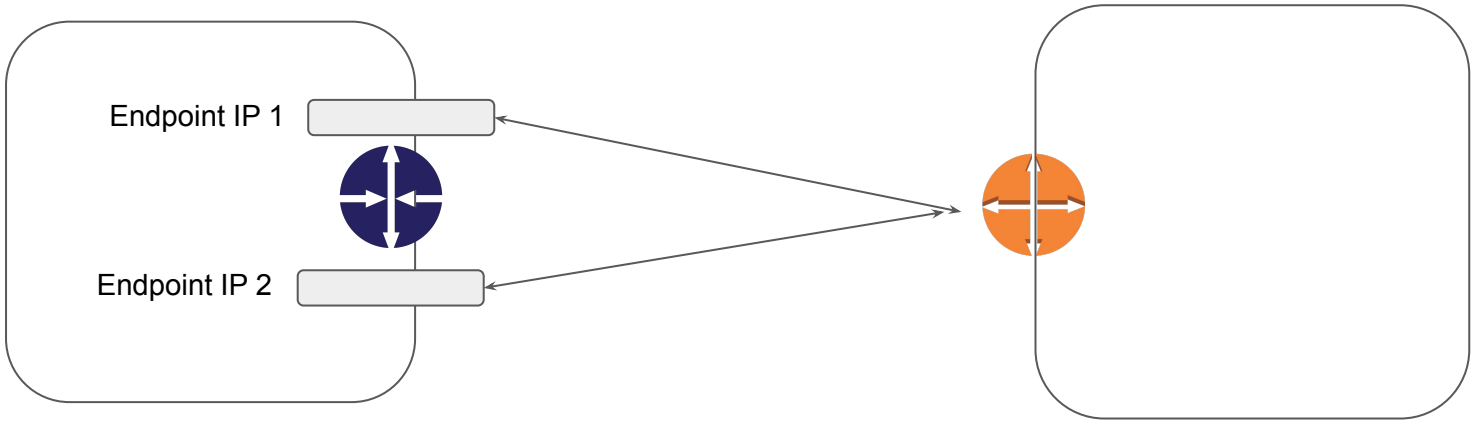


Site to Site VPN



Importance of VGW

- A Virtual Private Gateway (VGW) has built-in high-availability for VPN connection.
- AWS automatically creates 2 HA endpoints, each in a different AZ.



VPN Connection = 2 VPN Tunnels

Importance of VGW

Name	VPN ID	State	Virtual Private Gateway	Customer Gateway
ohio-mumbai	vpn-5cdf0a6b	available	vgw-7072fd40 | ohio-mumbai	cgw-27058b17 | ohio-mumbai

VPN Connection: vpn-5cdf0a6b

Details | **Tunnel Details** | Static Routes | Tags

Outside IP Address	Inside IP CIDR	Status	Status Last Changed	Details
18.216.150.193	169.254.59.32/30	UP	December 24, 2017 at 7:42:19 PM U...	-
18.220.211.76	169.254.57.128/30	DOWN	December 24, 2017 at 7:36:56 PM U...	-

Direct Connect

Let's Route Centrally

Customer to VPC



Packets travels via Hops

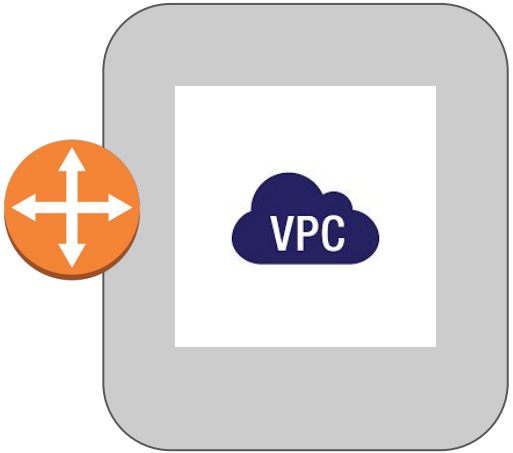


Challenges

- Internet is a good option if amount of traffic is within a certain limit.
- There are always latencies which can also be involved.
- Many of the organization have hybrid architecture : DataCenter + AWS
- In such cases, latency can cause major challenges for the application



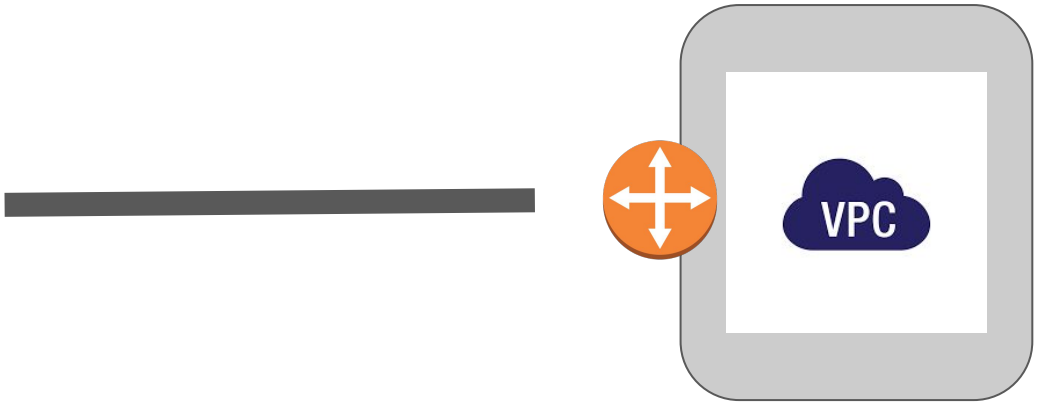




Introducing DX

- In order to solve this challenge, AWS introduced Direct Connect.
- AWS Direct connect let's customer establish a dedicated direct network connection between the client's network and one of the direct connect locations.





Benefits of DX

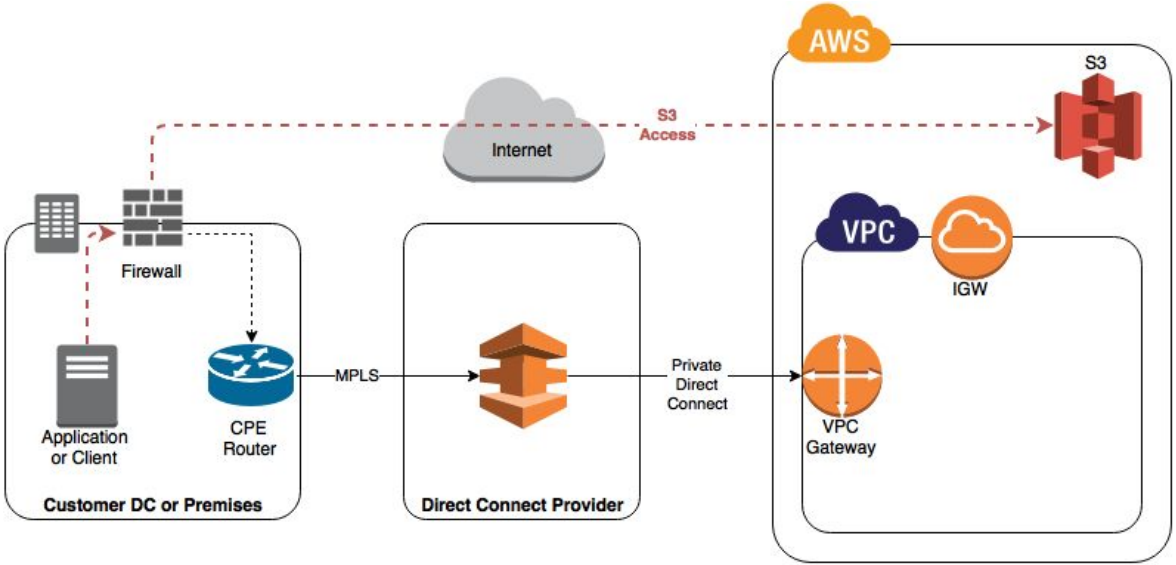
Having direct connection between customer's datacenter to AWS, brings tremendous amount of benefits, some of them includes:

i) Consistent Network Performance:

ii) Reduces our bandwidth costs

iii) Private connectivity to our AWS VPC

Architecture of DX

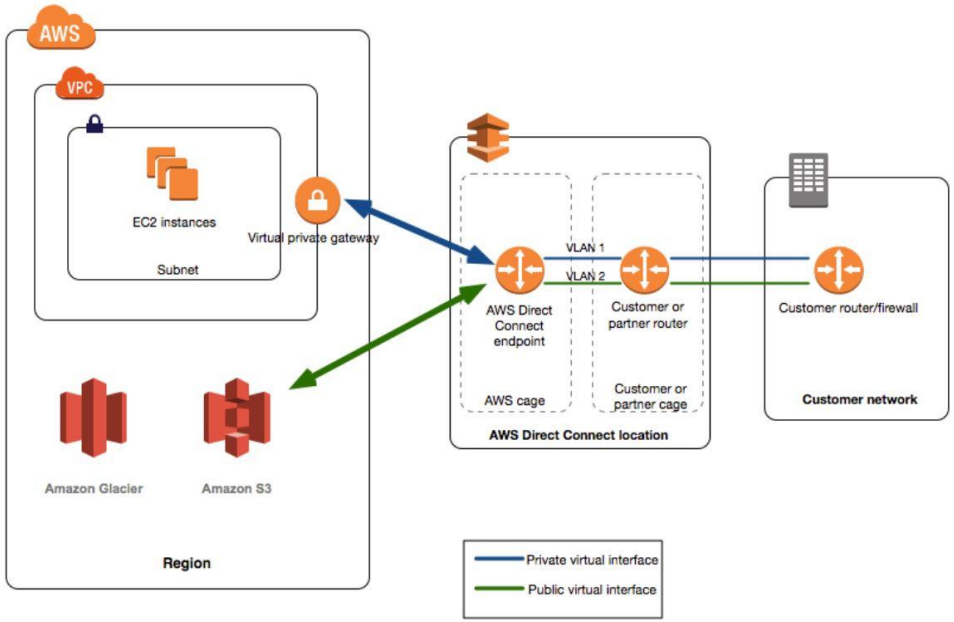


Virtual Interfaces

Connecting to AWS via DX

What after DX Line is Established?

The next primary step after Direct Connect is established is to configure the Virtual Interfaces.



Virtual Interface Types

Depending on your requirement, appropriate Virtual Interface (VIF) can be created.

Virtual Interface Type	Description
Public Virtual Interface	Enables access to public AWS services like AWS S3, and others that are not in the VPC.
Private Virtual Interface	Enables access to your VPC
Transit Virtual Interface	Access one or more Amazon VPC Transit Gateways associated with Direct Connect gateways

Steps 1 : Establish DX connection request

- Here we specify the connection name, location and the port speed.
- After we click on create, it will go for review to AWS and if approved, we get a LOA which we can download and give it to provider who will be establishing DX connection on your behalf.
- It takes upto 3 working days for the LOA to be approved.

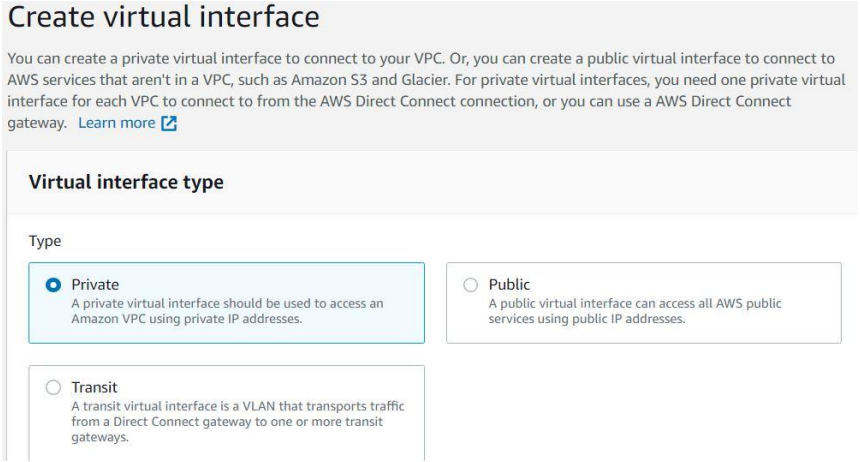




Steps 2 : Create Virtual Interface

Create Virtual Interface based on your requirement.

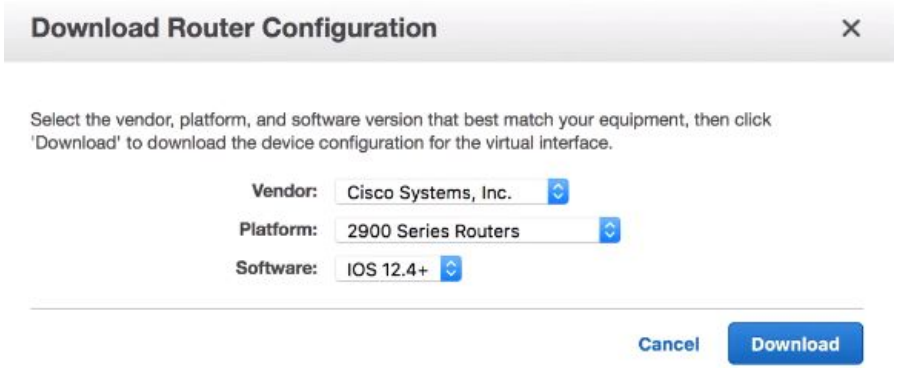
Can be associated with Direct Connect Gateway or Virtual private Gateways.



Step 3 - Download Router Configuration

After you have created the virtual interface for your AWS Direct Connect connection, you can download the router configuration file.

The file contains the necessary commands to configure your router for use with your private or public virtual interface



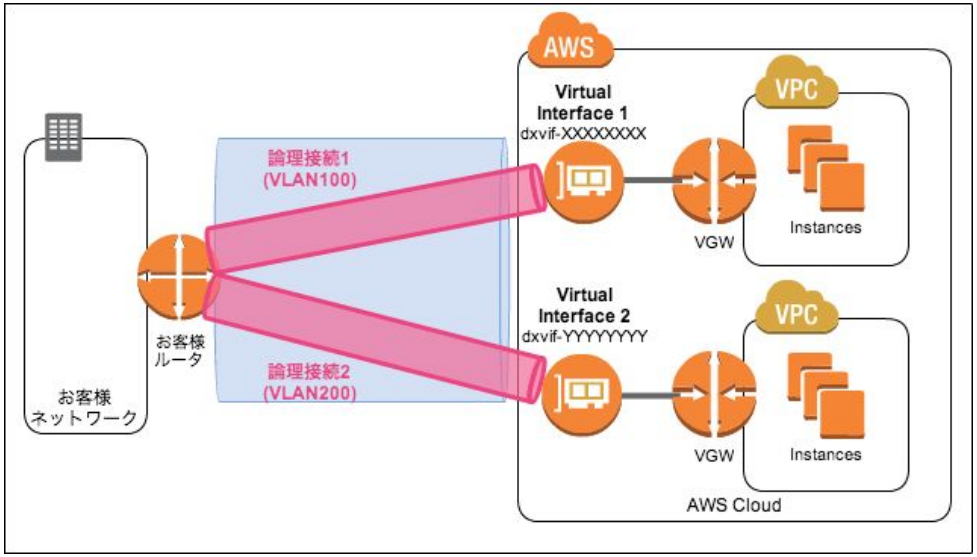
Important Pointers

- By default 1Gbps and 10 Gbps connections are available, we can also have sub-1GB connection from direct connect partners which includes 50 mbps, 100 mbps, 200 mbps, 400 mbps, 500 mbps.

- Direct connect is not fault tolerant, so we need to either have secondary Direct Connect or use VPN as backup. Use BGP to automatic failover to backup connection.

- In US, direct connect will grant you access in all the US related region.

Virtual Interfaces



Direct Connect Gateway

Direct connect all the way

Gateway Types for Virtual Interface

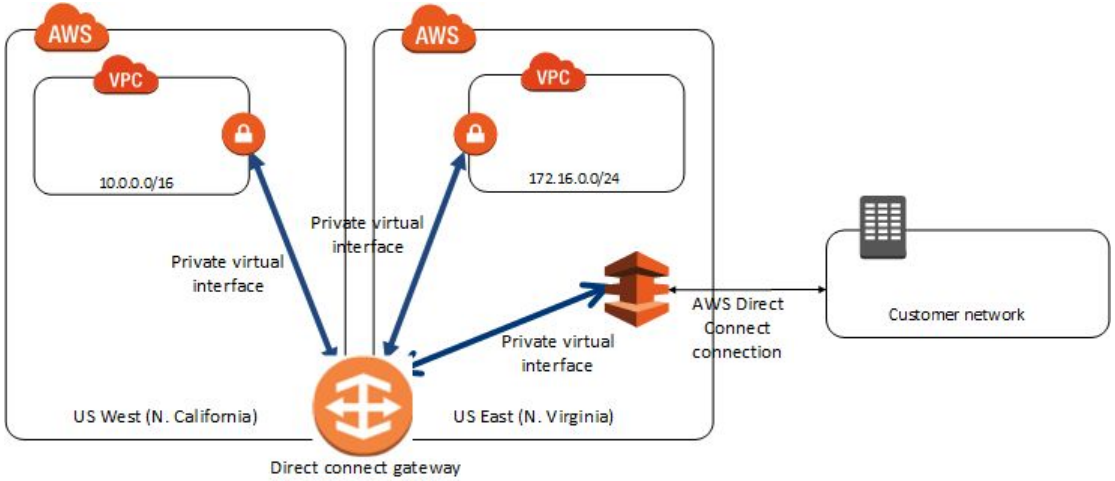
Following tables illustrates the supported gateway types for the Virtual Interfaces

Gateway Types	Description
Virtual Private Gateway	Allows connections to a single VPC in the same region
Direct Connect Gateway	Allows connections to multiple VPCs and regions

Overview of DX Gateway

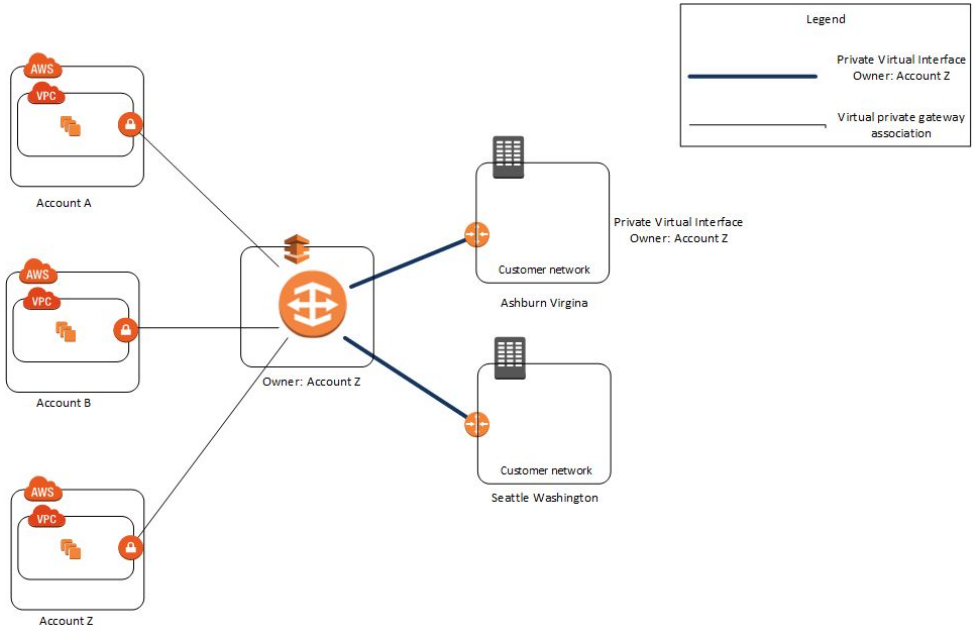
Direct Connect Gateway can be used to connect your direct connect connection over private VIF to one or more VPC's within the account that are located in same or multiple regions.

It allows us to combine private VIF's with multiple VGW's in local or in remote region.



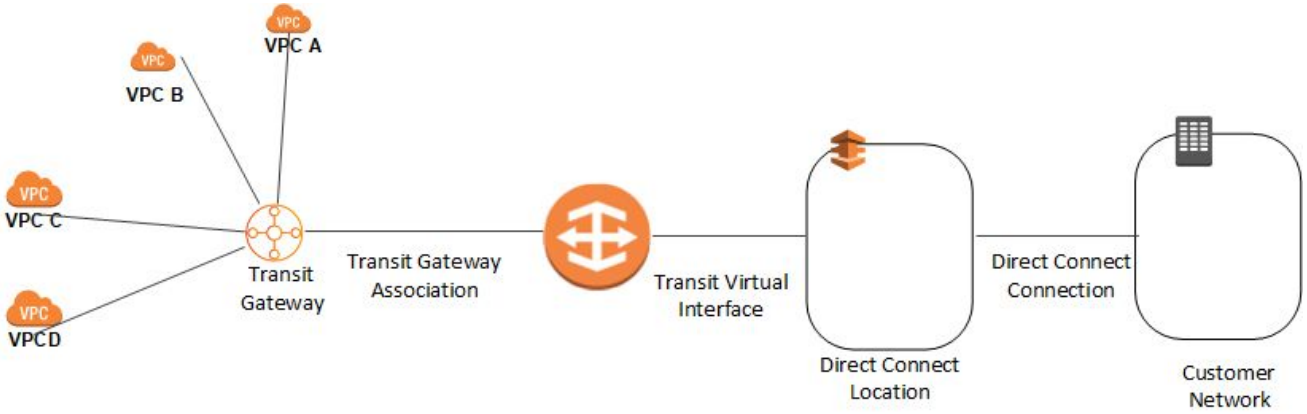
Multiple Accounts

Multiple AWS accounts can also be integrated with DX Gateways.



Transit Gateway Association

The following diagram illustrates how the Direct Connect gateway enables you to create a single connection to your Direct Connect connection that all of your VPCs can use.



Direct Connect - High Availability

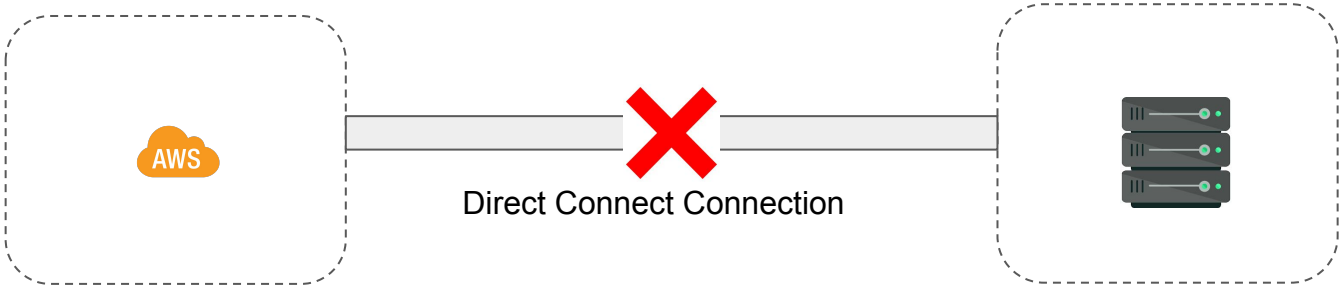
Allows us to have a better sleep.

Getting Started

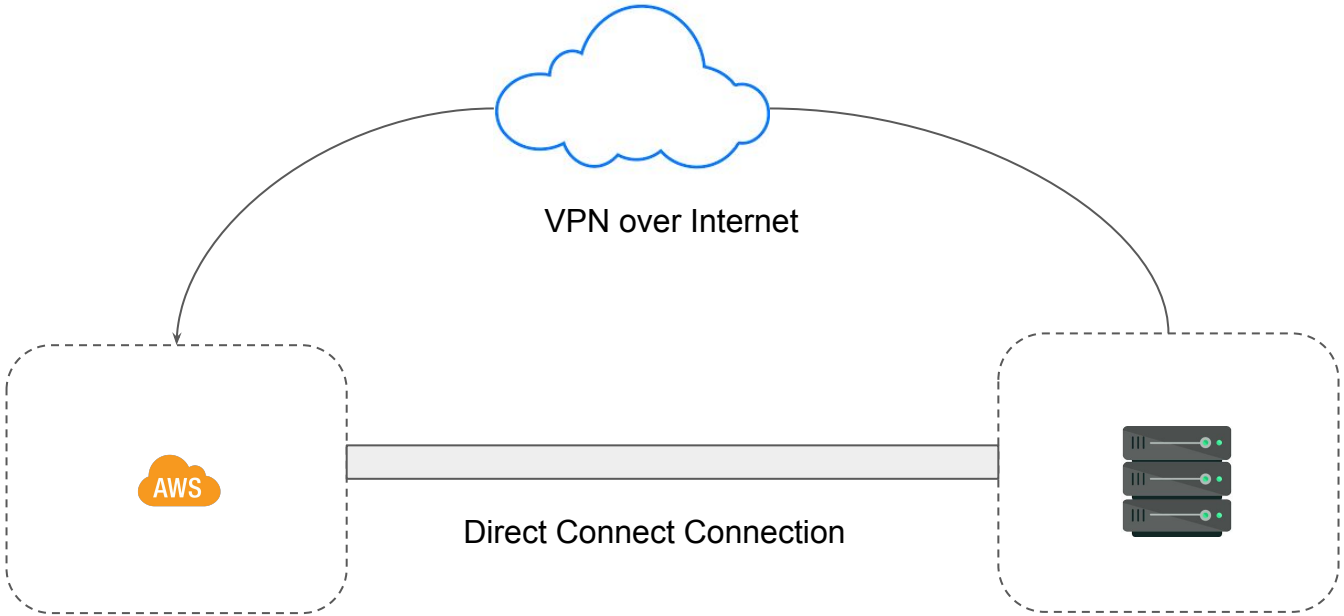
If you have a single DX connection, it is subjected to scenario of failover.

In-case if your DX connection breaks, your link will break completely.

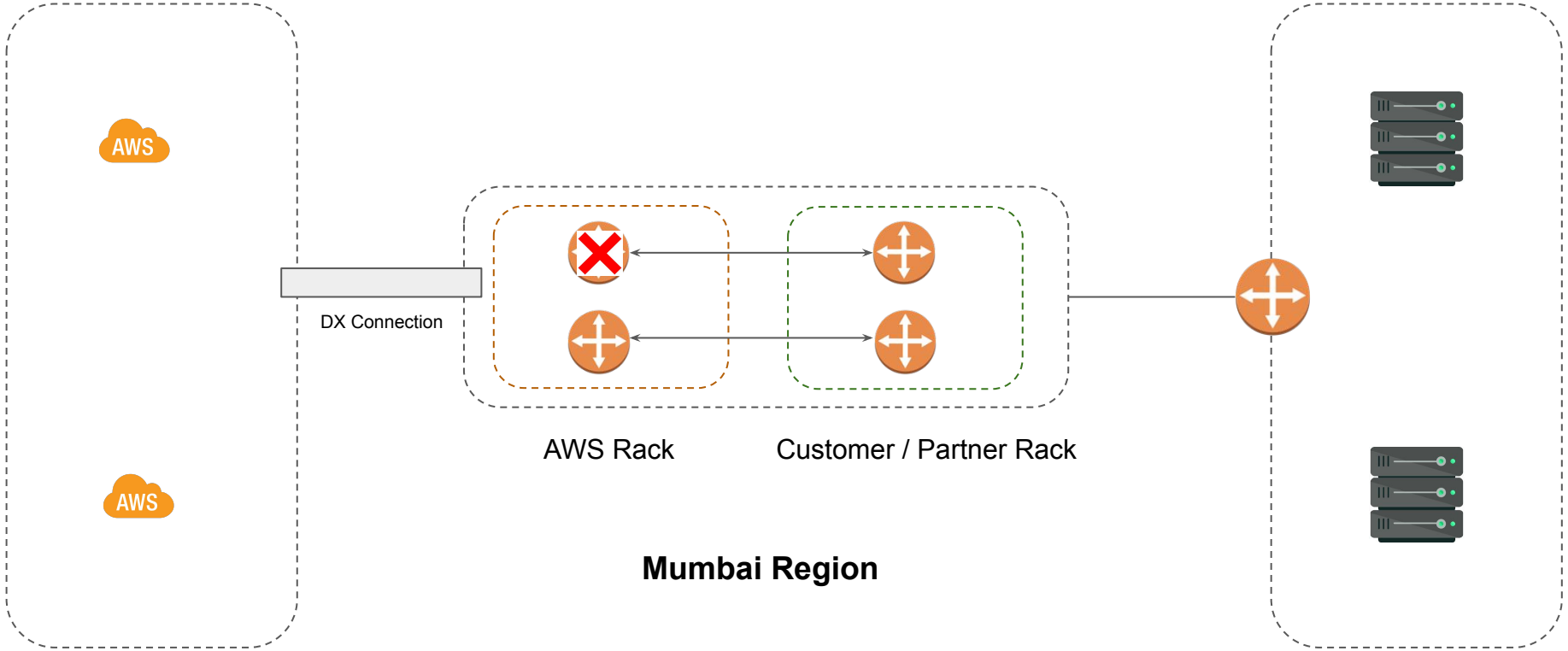
Hence it is recommended to have a backup connection of VPN over the Internet.



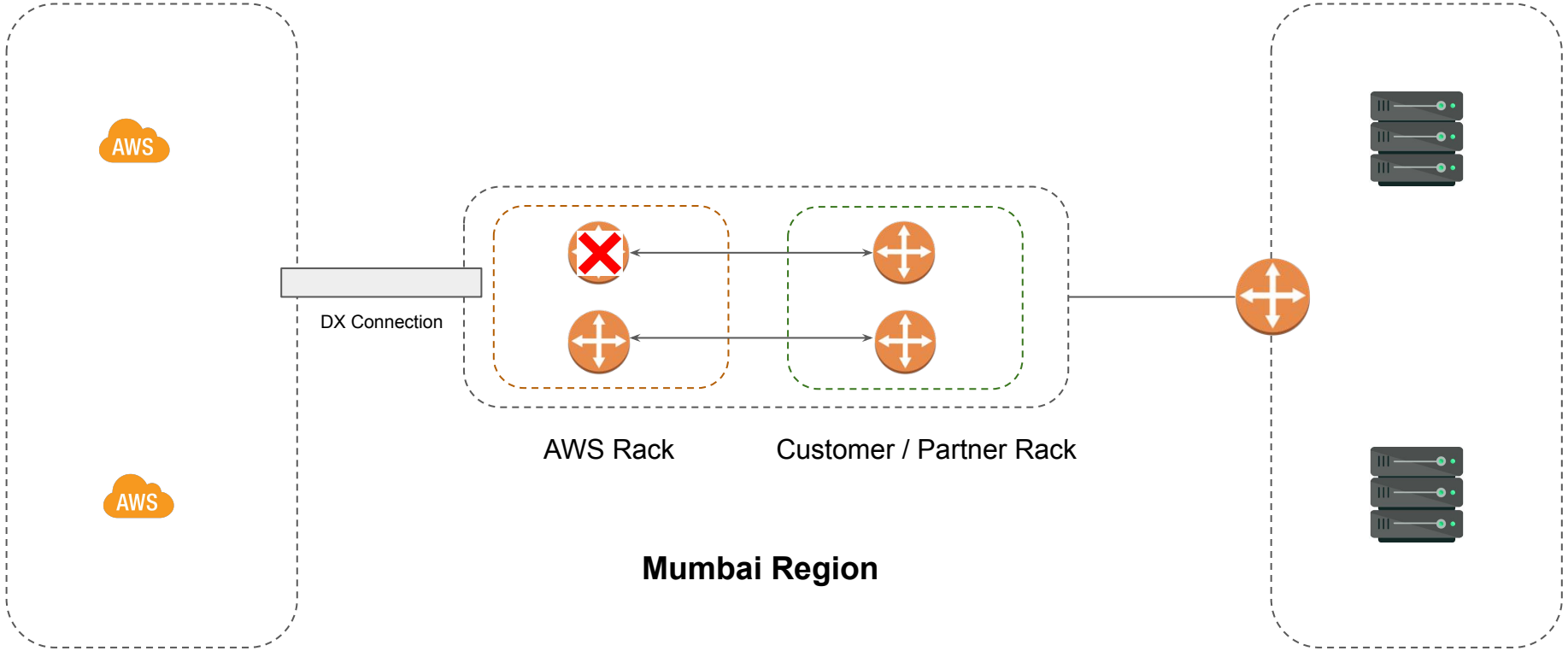
Secondary Backup Connection



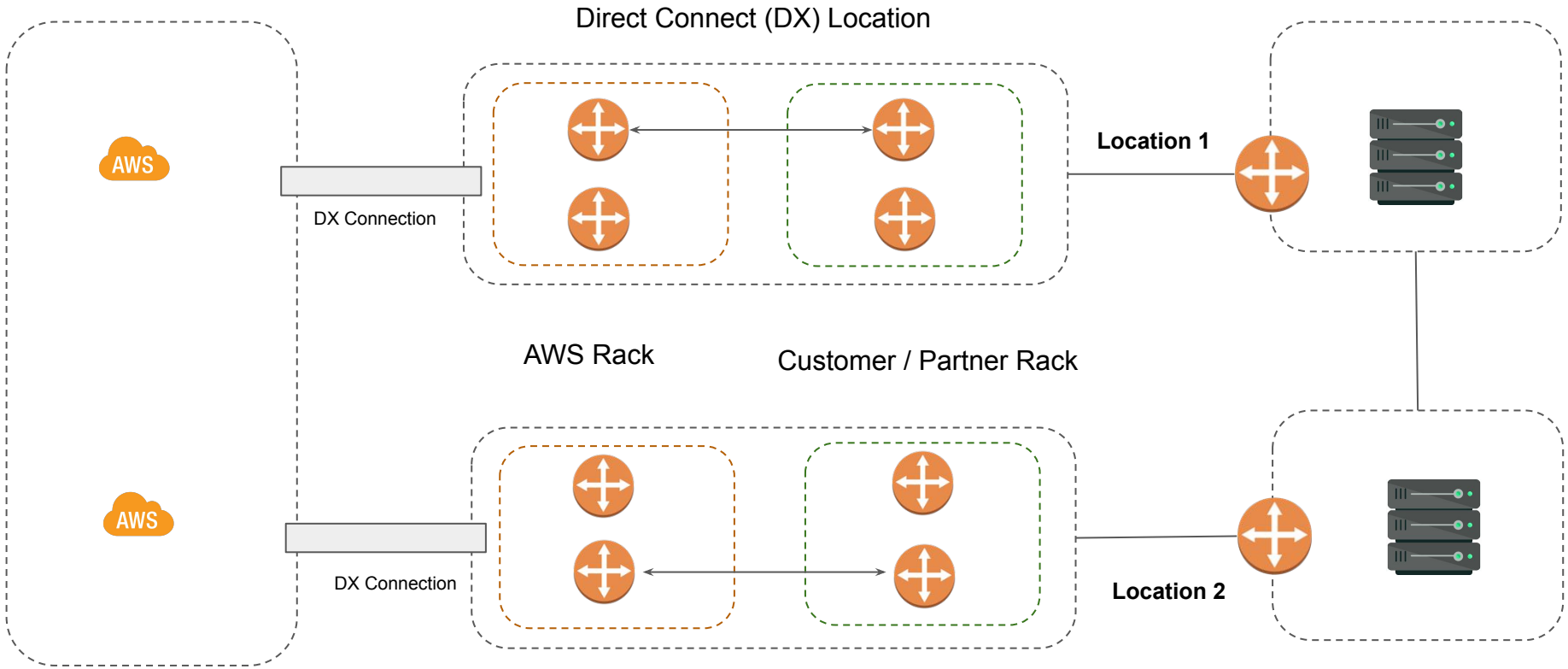
Dual Connection - Single Location



Dual Connection - Single Location



Dual Connection - Dual Location

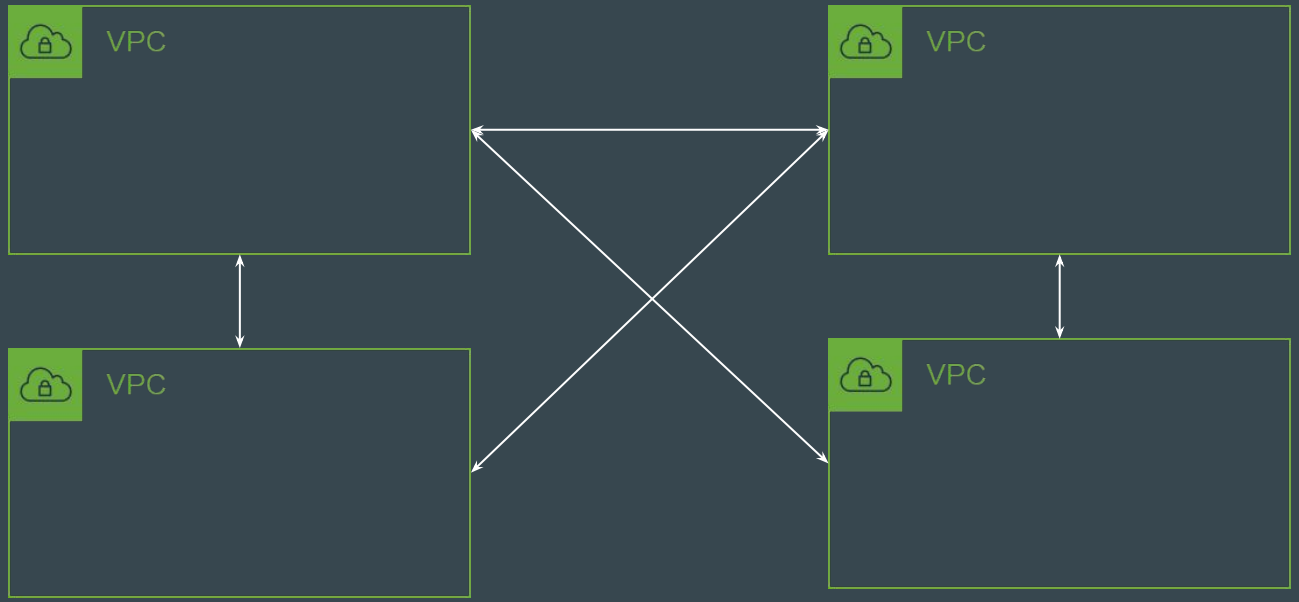


Transit Gateways



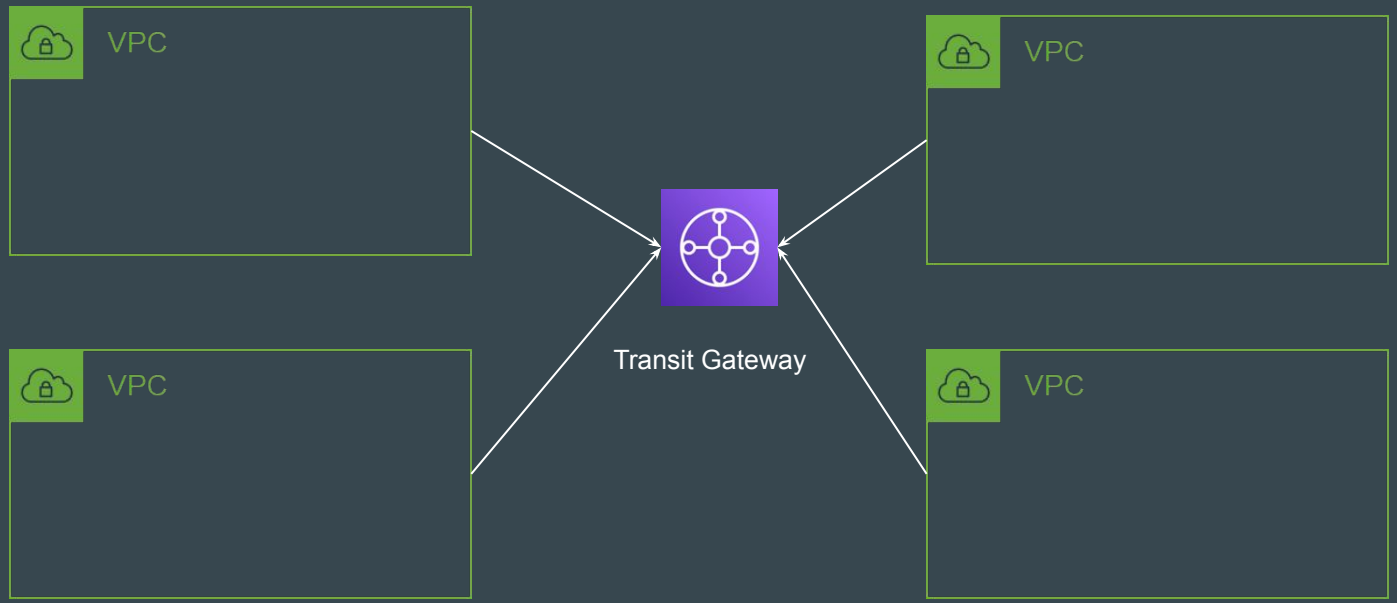
Use-Case: Connecting 4 VPCs

More the Number of VPCs, more the number of peering connection you have to establish for inter-connectivity related use-case.

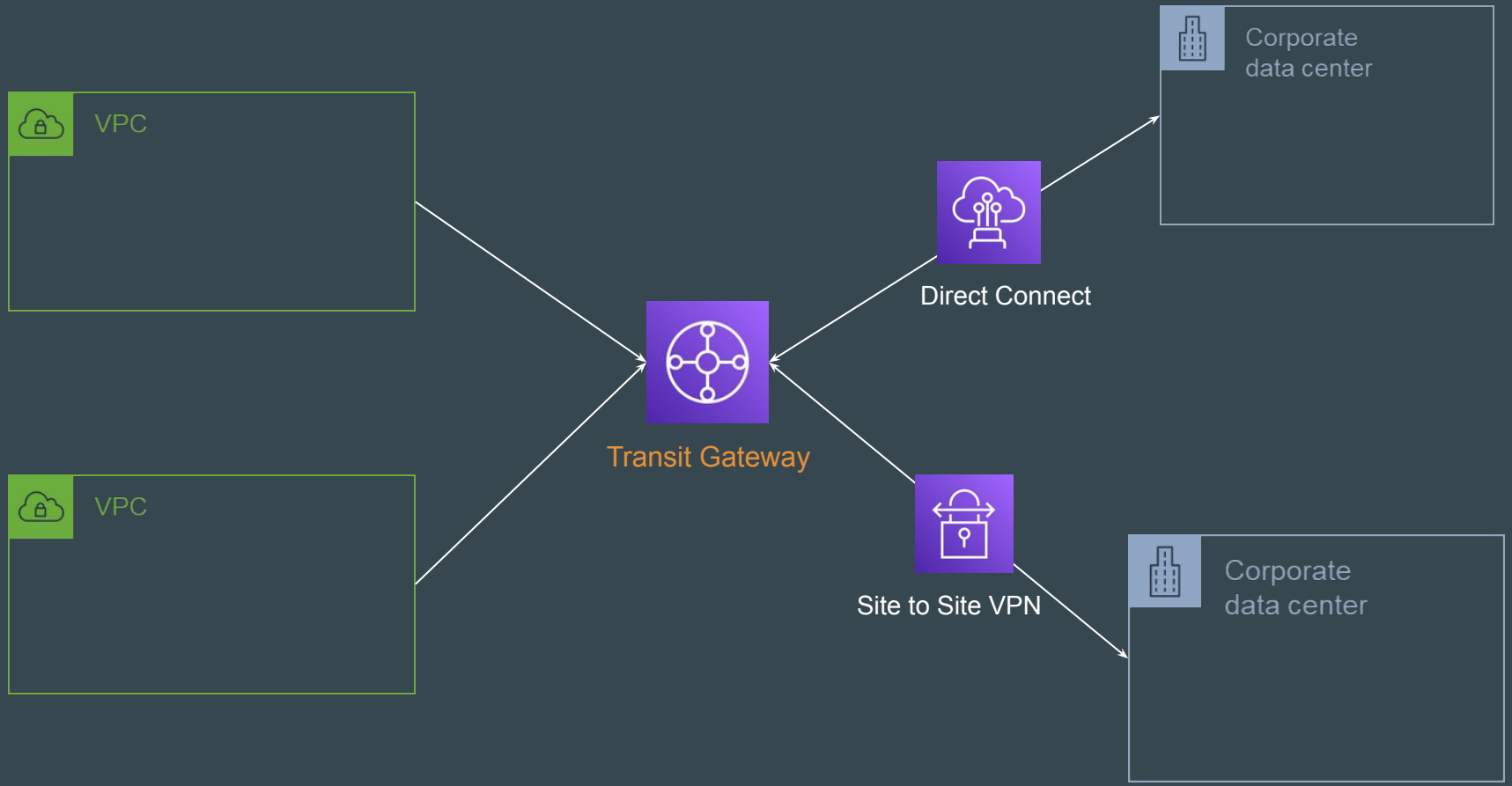


Introducing Transit Gateway

AWS Transit Gateway connects your Amazon Virtual Private Clouds (VPCs) and on-premises networks through a central hub



Larger Setup



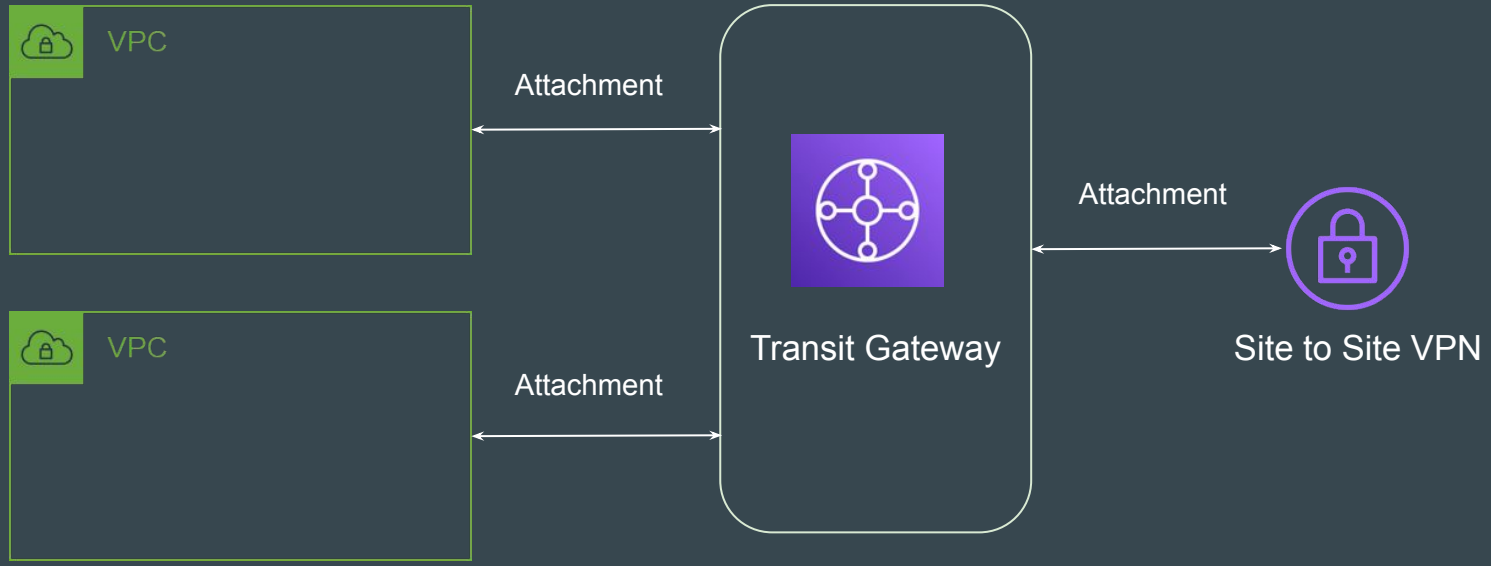
Transit gateway concepts



Concept 1 - Attachments

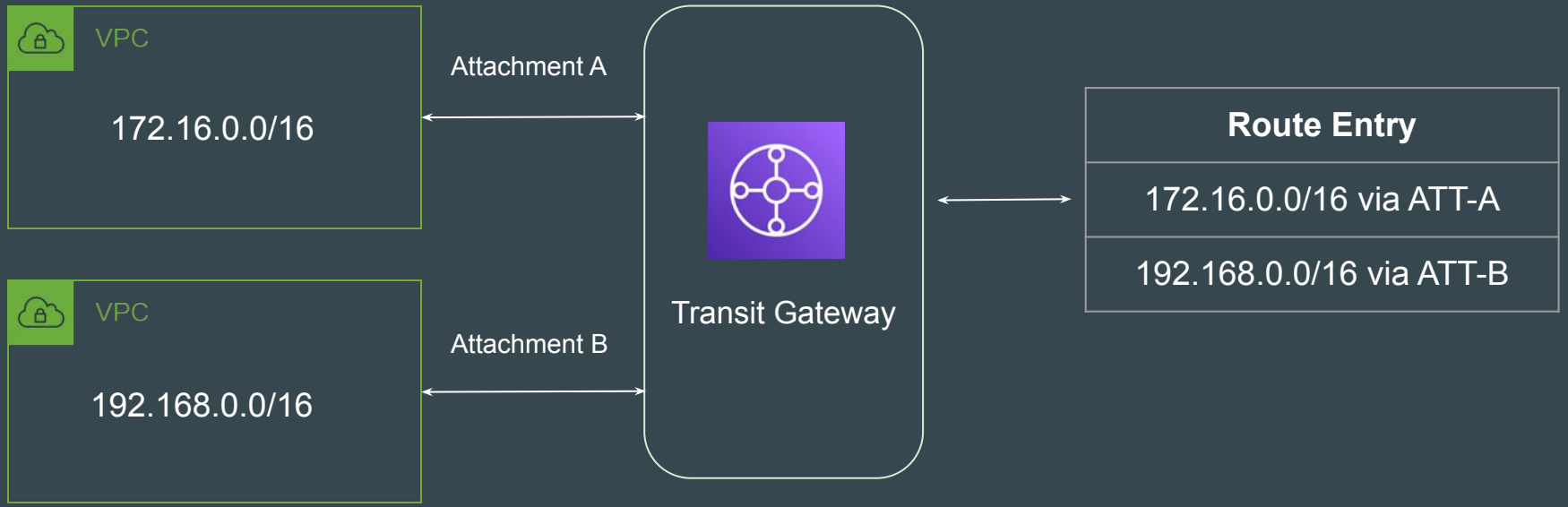
Multiple resources can be attached to Transit Gateway.

Some of the supported entities: VPCs, Direct Connect Gateway, VPN, SD-WAN



Concept 2 - Route Table

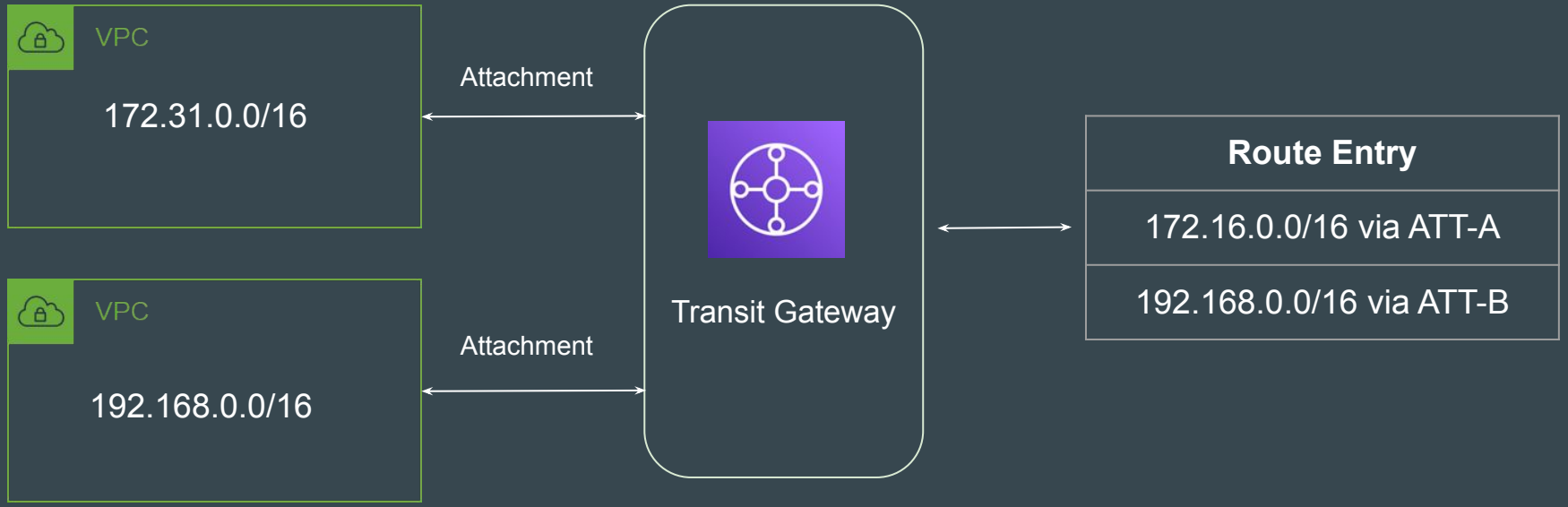
Defines how the traffic is routed between the connected resources.



Transit gateway Practical



Our Practical Setup



Success Criteria

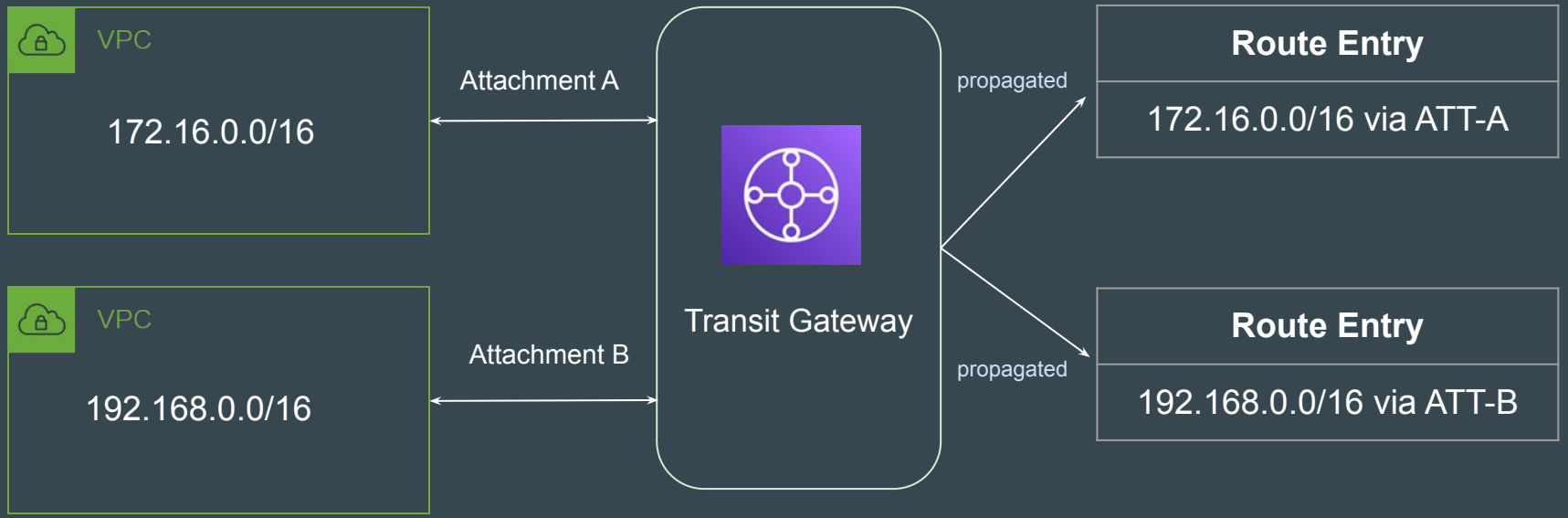
EC2 Instance from VPC-1 should be able to communicate to E2 Instance from VPC-2 through Transit Gateway.

Routes in Transit Gateways



Route Propagation

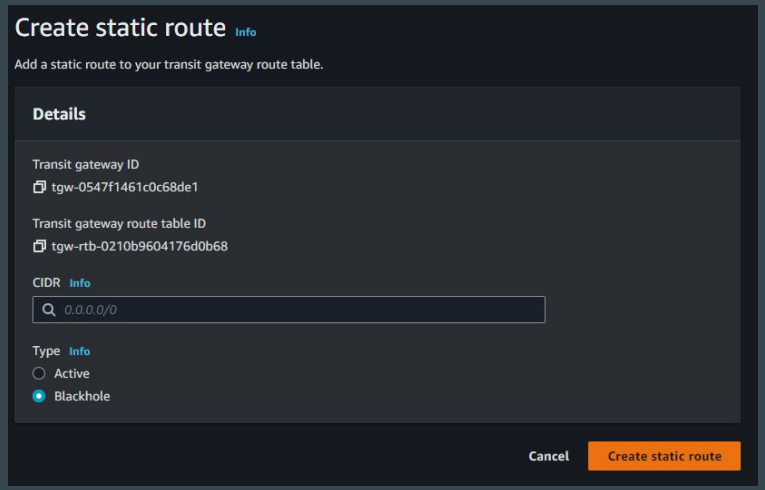
For a VPC attachment, the CIDR blocks of the VPC are propagated to the transit gateway route table.



Static Routes

In addition to propagated routes, you can also add static routes.

Static routes allows more flexible routing policy with option to even DROP traffic.



Route Evaluation Order

The most specific route for the destination address is given higher priority.

For routes with the same destination IP address but different targets, the route priority is as follows:

1. Static routes (for example, Site-to-Site VPN static routes)
2. Prefix list referenced routes
3. VPC propagated routes
4. Direct Connect gateway propagated routes
5. Transit Gateway Connect propagated routes
6. Site-to-Site VPN propagated routes

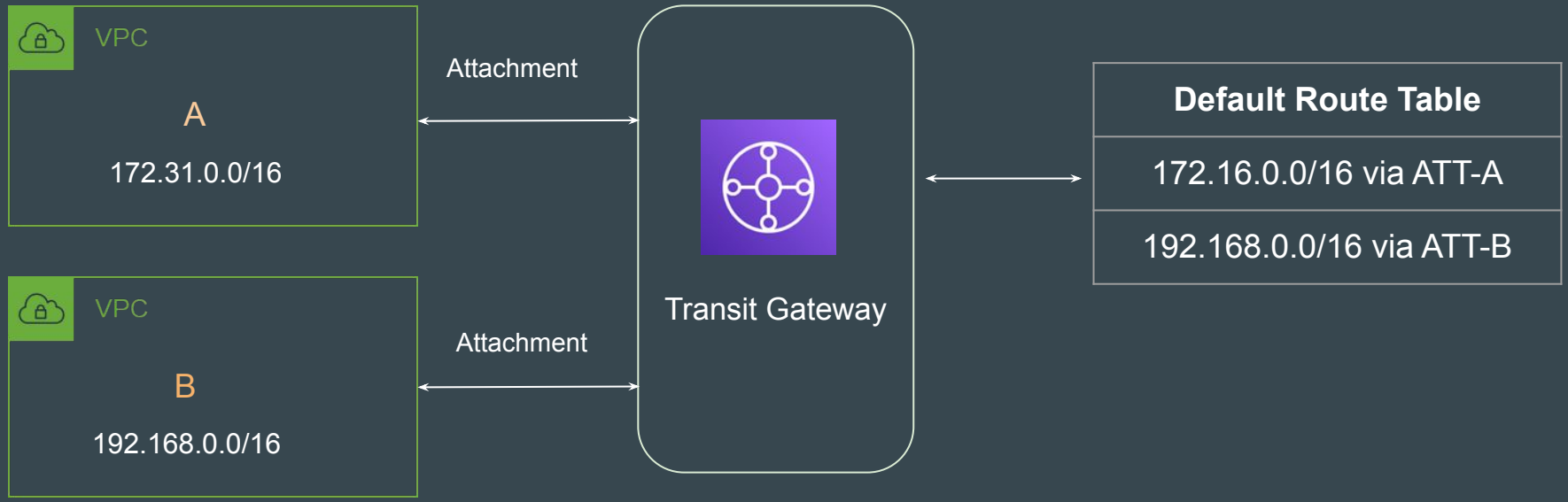
Attachment Specific Routing



Default Route Table

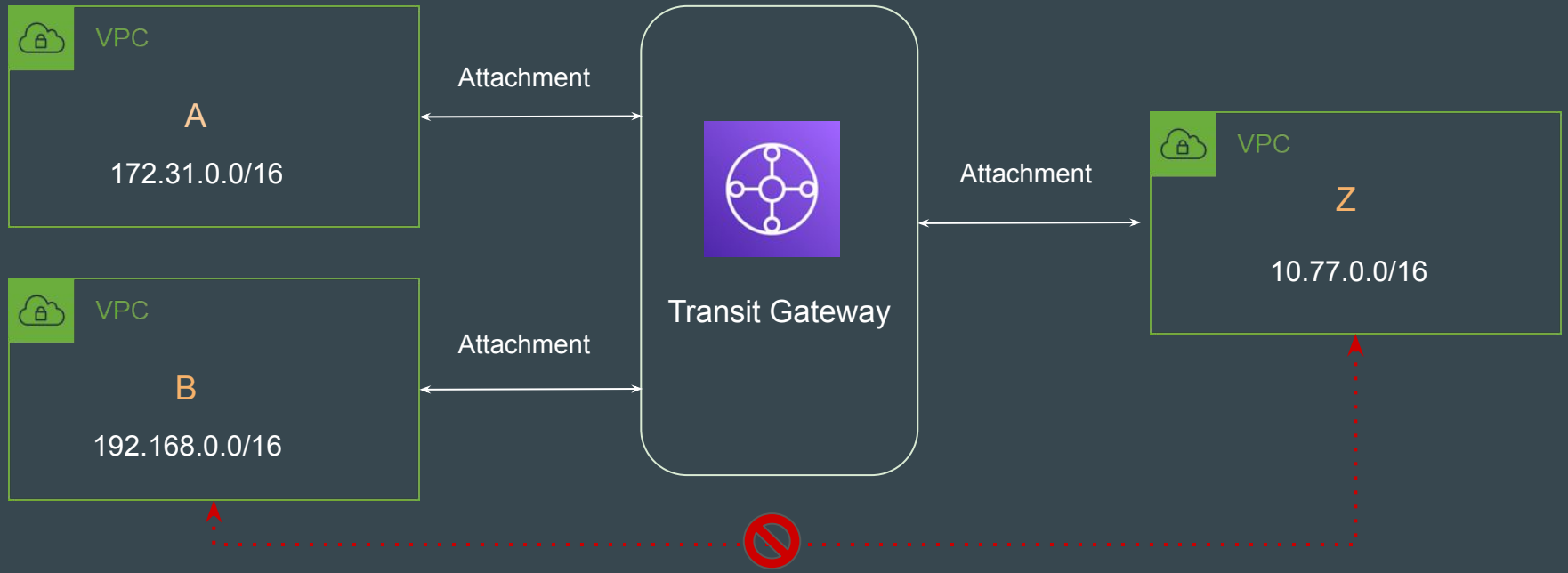
Your transit gateway automatically comes with a default route table.

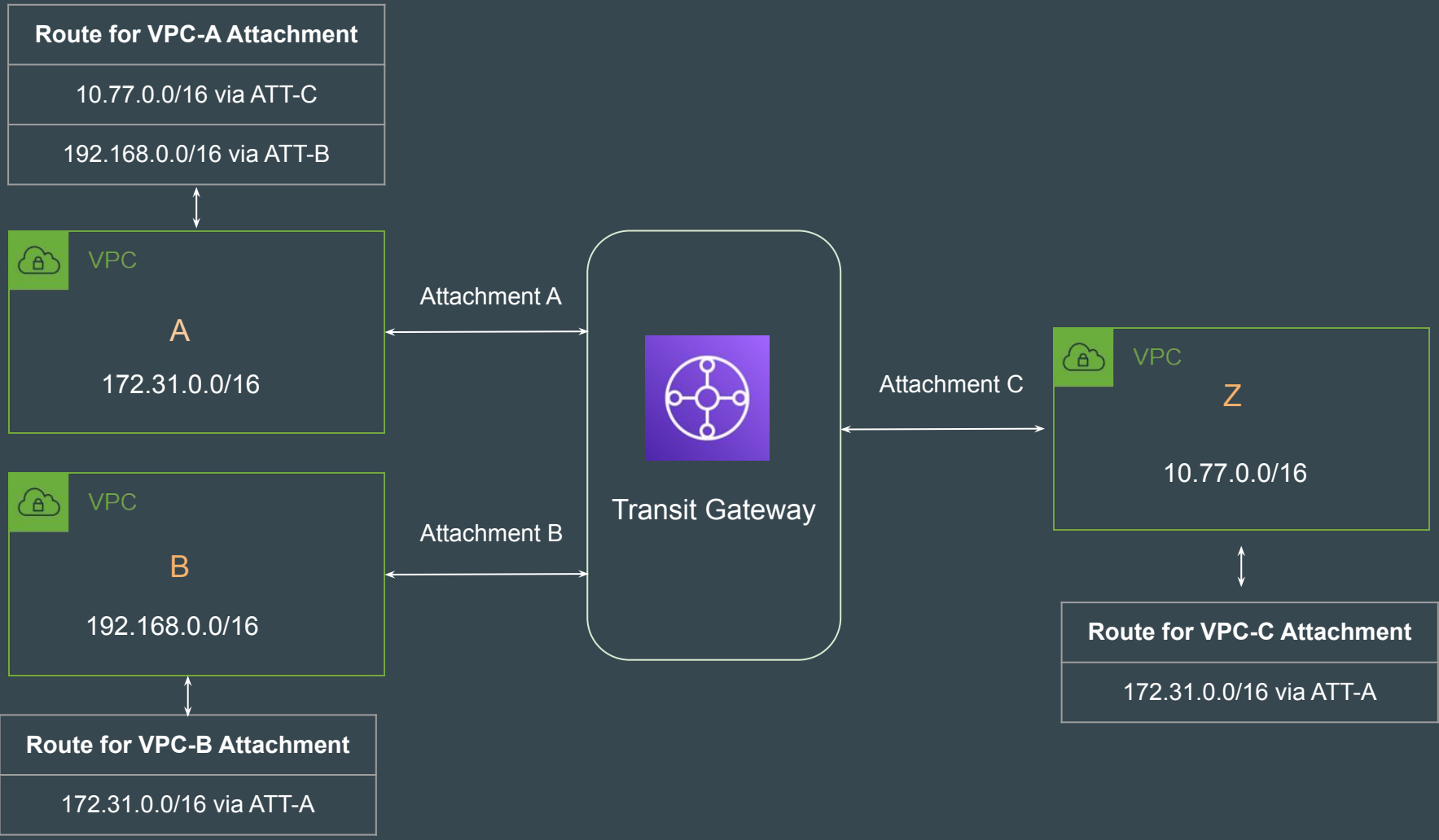
By default, this route table is the default association route table and the default propagation route table.



Understanding the Use-Case

All communication should be allowed EXCEPT VPC-B to VPC-Z



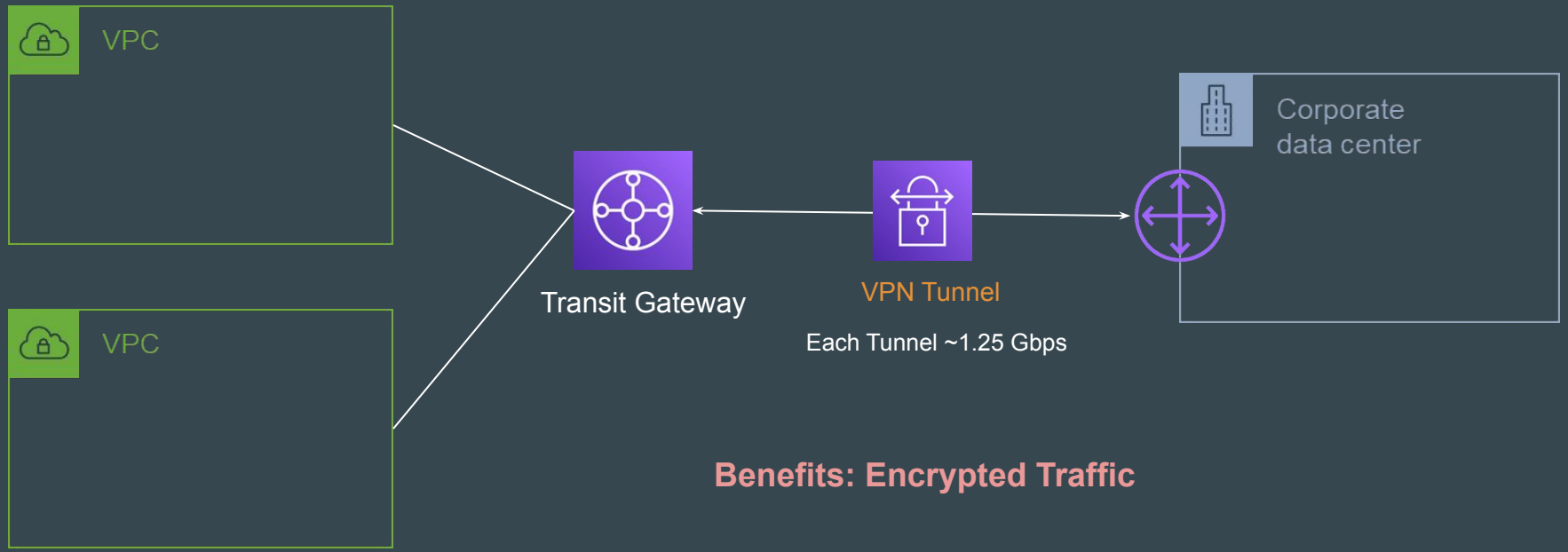

Route for VPC-A Attachment
10.77.0.0/16 via ATT-C
192.168.0.0/16 via ATT-B
VPC
A
172.31.0.0/16
Attachment A
Transit Gateway
Attachment C
VPC
Z
10.77.0.0/16
VPC
B
192.168.0.0/16
Attachment B
Route for VPC-C Attachment
172.31.0.0/16 via ATT-A
Route for VPC-B Attachment
172.31.0.0/16 via ATT-A

Transit Gateway - VPN Attachments



AWS Transit Gateway + VPN

AWS Transit Gateway + VPN, using the Transit Gateway VPN attachment, provides the option of creating an IPsec VPN connection between your remote network and the Transit Gateway over the internet



Benefits: Encrypted Traffic

Points to Note

A single VPN tunnel has a maximum throughput of 1.25 Gbps

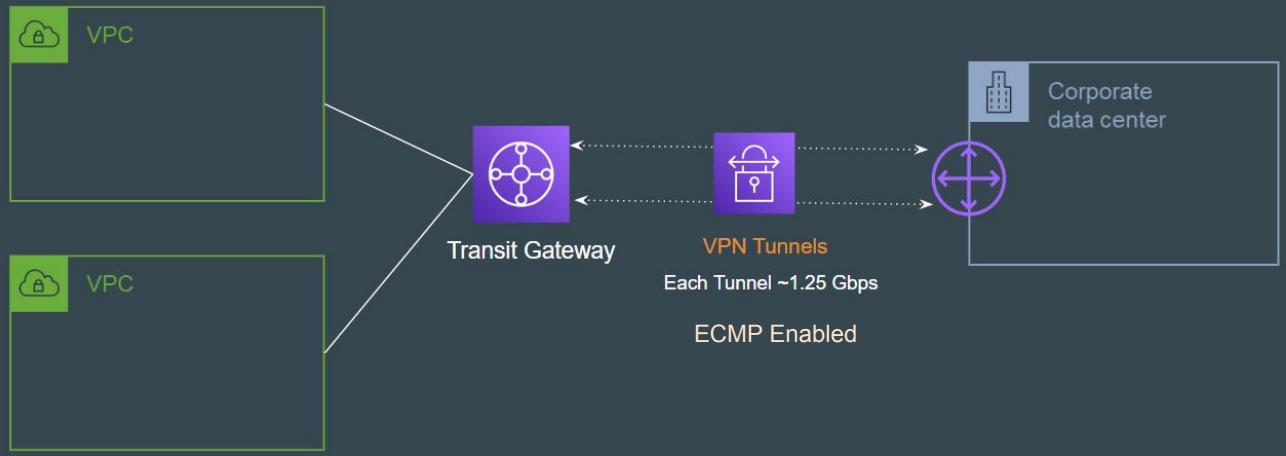
If you establish multiple VPN tunnels to an ECMP-enabled transit gateway, it can scale beyond the default maximum limit of 1.25 Gbps.

Equal-cost multi-path routing (ECMP) is a routing strategy where packet forwarding to a single destination can occur over multiple best paths with equal routing priority.

Transit Gateway with Multiple VPN Connection

When you create your VPN, you must choose Dynamic for Routing options. Static routing does not support ECMP.

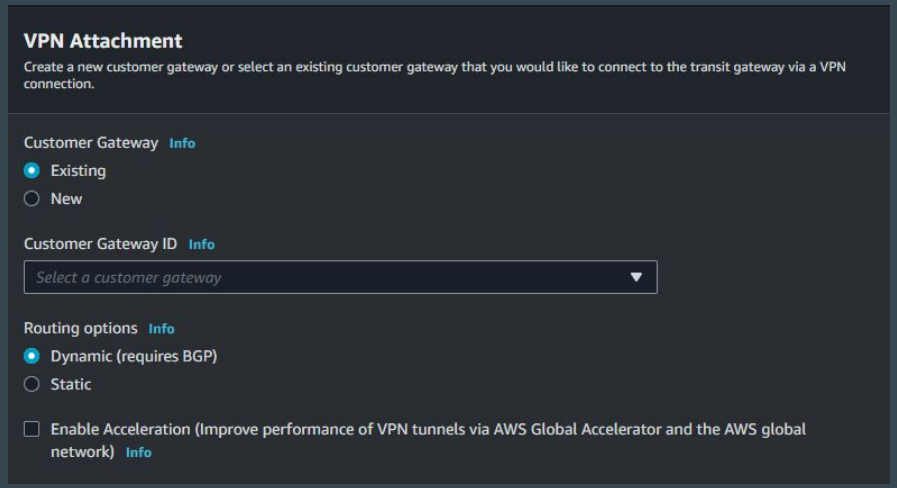
When you create your transit gateway, you must enable VPN ECMP support.



VPN Attachment

To attach a VPN connection to your transit gateway, you must specify the customer gateway.

For static VPNs, we have to add the static routes to the transit gateway route table.



ECMP with multiple VPN tunnels with a transit gateway

Ensure that your customer gateway is configured to perform ECMP for traffic going out to AWS for all VPN tunnels.

Confirm that your customer gateway is advertising the on-premises prefix to AWS with the same BGP AS PATH attribute.

For AWS to choose all the available ECMP paths, the AS Path and AS Number must match.

Example Configuration

You plan to use ECMP with two VPN connections. The AS Number of your customer gateway is 65270. In this scenario, you configure your VPNs as follows:

VPN-A

- Tunnel 1 – AS PATH: 65270 (while advertising the prefix)
- Tunnel 2 – AS PATH: 65270 (while advertising the prefix)

VPN-B

- Tunnel 1 – AS PATH: 65270 (while advertising the prefix)
- Tunnel 2 – AS PATH: 65270 (while advertising the prefix)

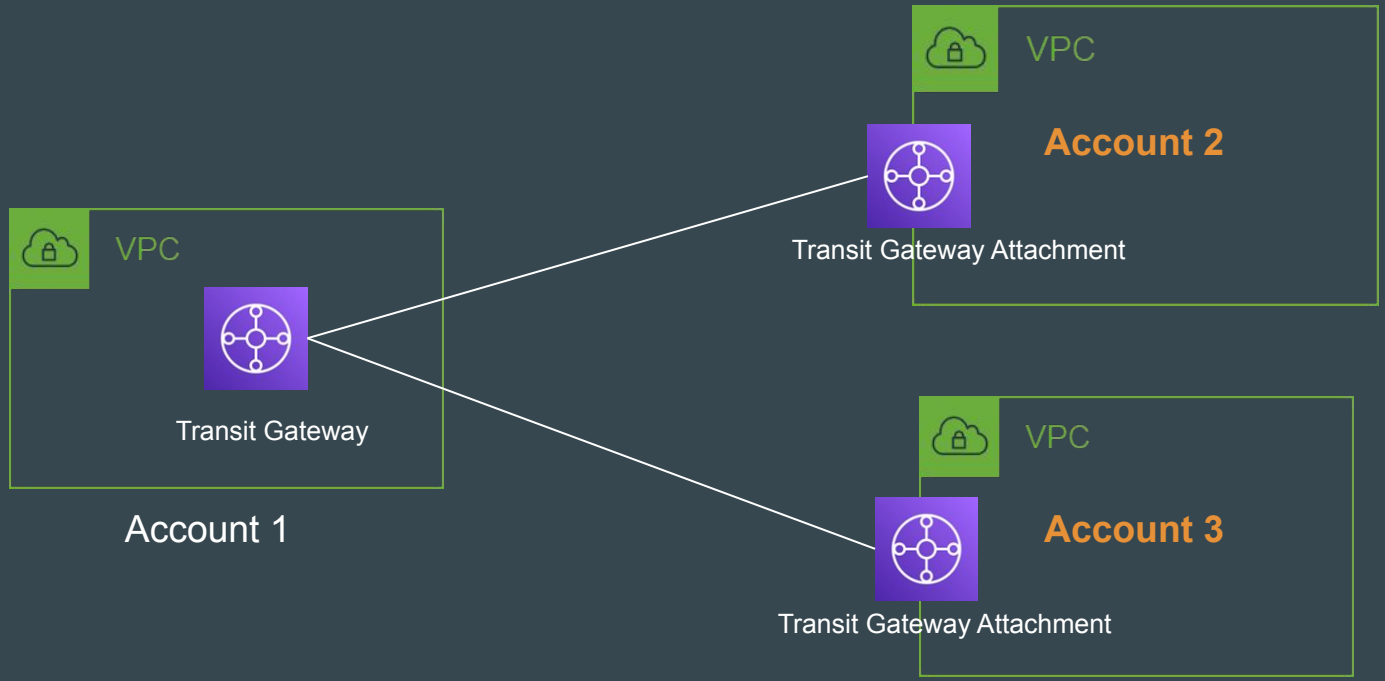
With a configuration similar to this information, AWS sends out traffic with ECMP on all four VPN tunnels.

Transit Gateway Sharing



Base Architecture

Transit Gateway sharing allows VPCs across multiple accounts to use Transit gateway for inter-connectivity.



Points to Note

An AWS Site-to-Site VPN attachment must be created in the same AWS account that owns the transit gateway.

When a transit gateway is shared with you, you cannot create, modify, or delete its transit gateway route tables, or its transit gateway route table propagations and associations.

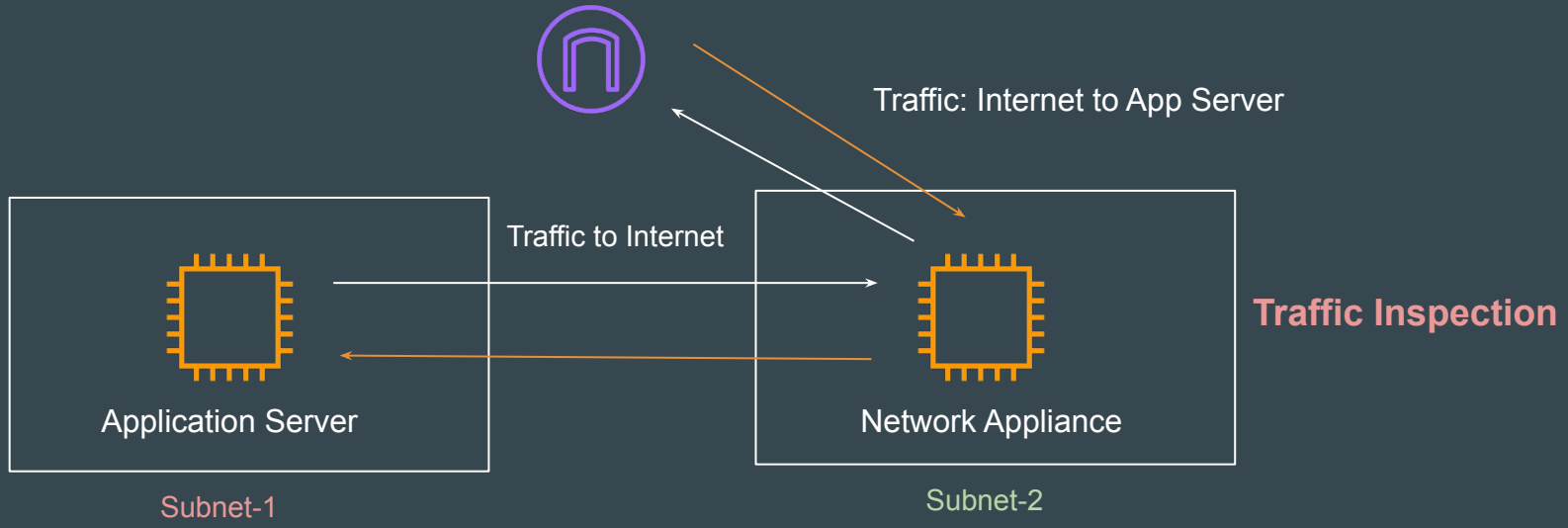
Gateway Load Balancer



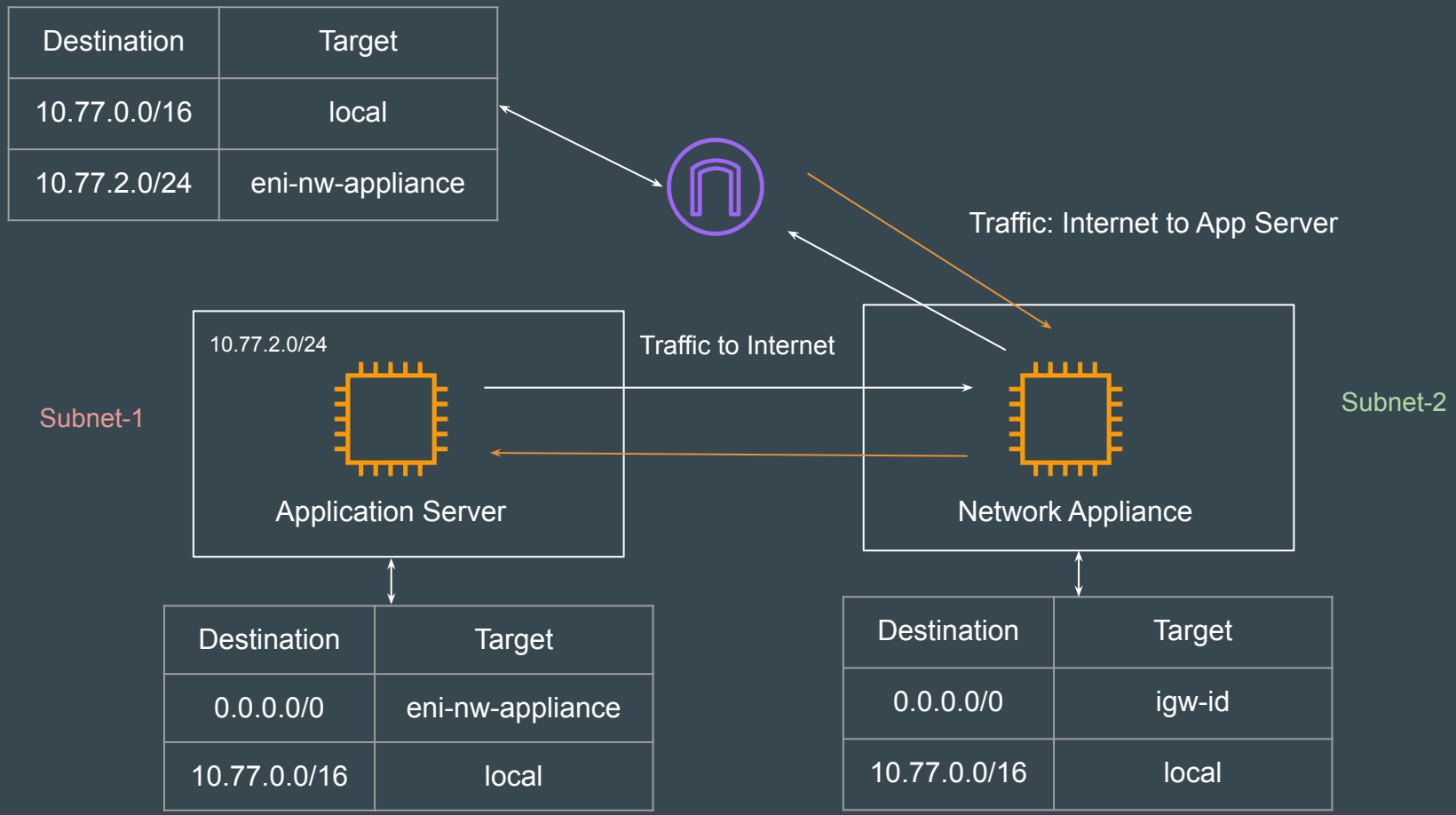
Problem Statement

Traffic Inspection is one of the common use-cases in Enterprises.

Many providers offers virtual appliance related to IDS/IPS, Firewalls etc.



The Architecture



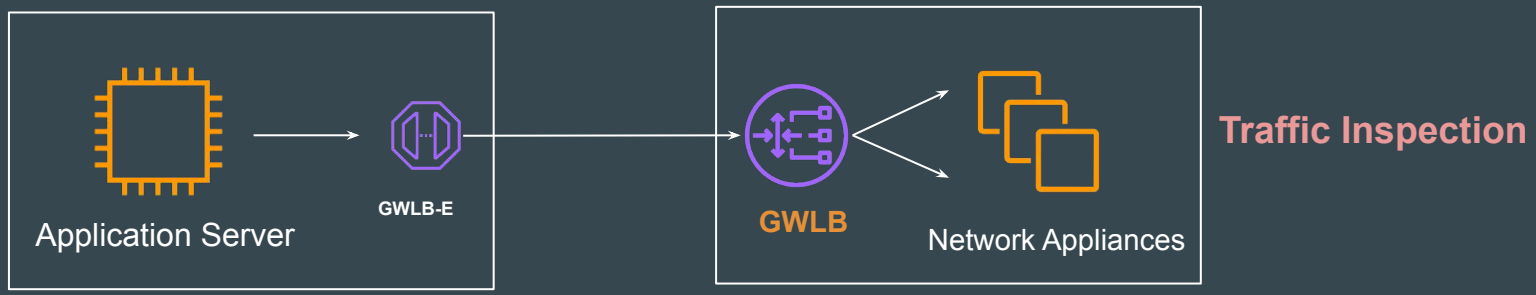
Challenges with the Architecture

The routing is done at a ENI level of the Network Appliance.

Issues: High-Availability, Scaling

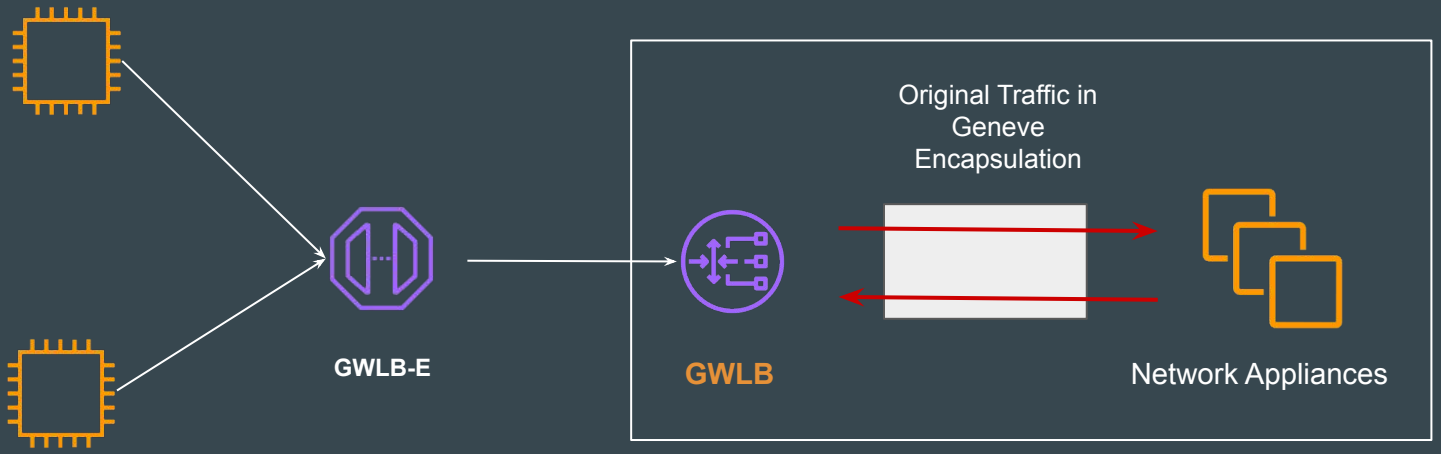
Introducing GWLB

Gateway Load Balancers enable you to deploy, scale, and manage virtual appliances, such as firewalls, IDS/IPS , and deep packet inspection systems



Points to Keep In Mind

In order to work with GWLB, appliances need to support Geneve protocol to exchange traffic with GWLB.



NAT Gateway Performance

Multiple is better

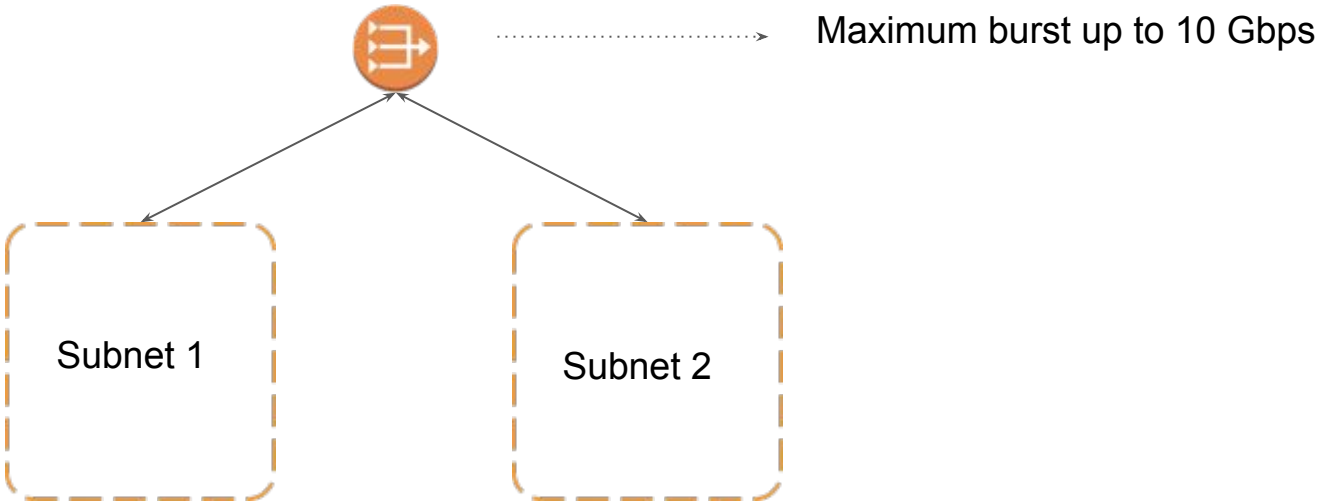
Performance Aspect

NAT Gateway supports a burst of up to 10 Gbps of bandwidth.

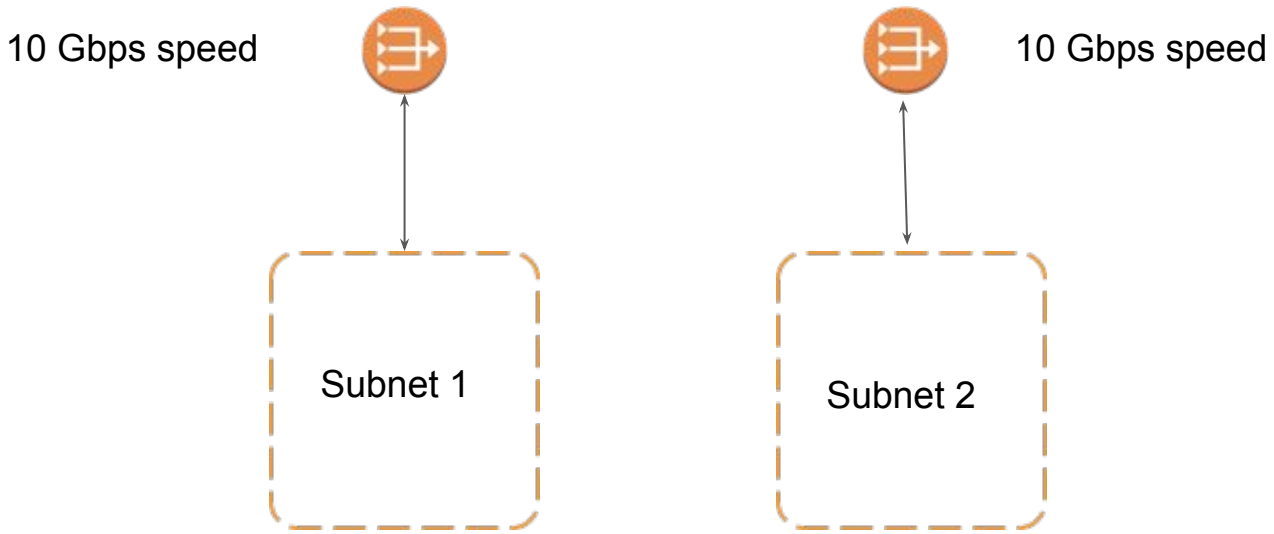
Thus all the instances within the private subnet need to have traffic less than that of 10 Gbps. If more than 10 Gbps, then the network will be the bottleneck.

Thus when we need more bandwidth, than the recommended design is to split the instance across multiple subnets and attach different NAT gateway to each of those subnets.

Normal NAT Gateway based Architecture



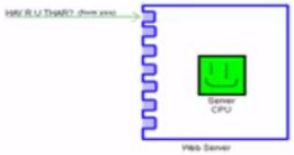
Multiple NAT Gateway Approach



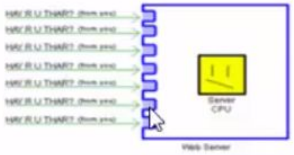
Denial of Service

Attack difficult to mitigate

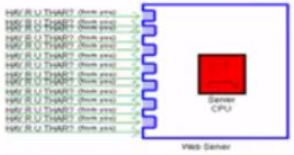
normal service →



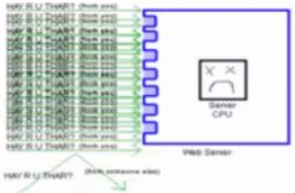
high traffic →



single DOS →



LOL DDOS'D →



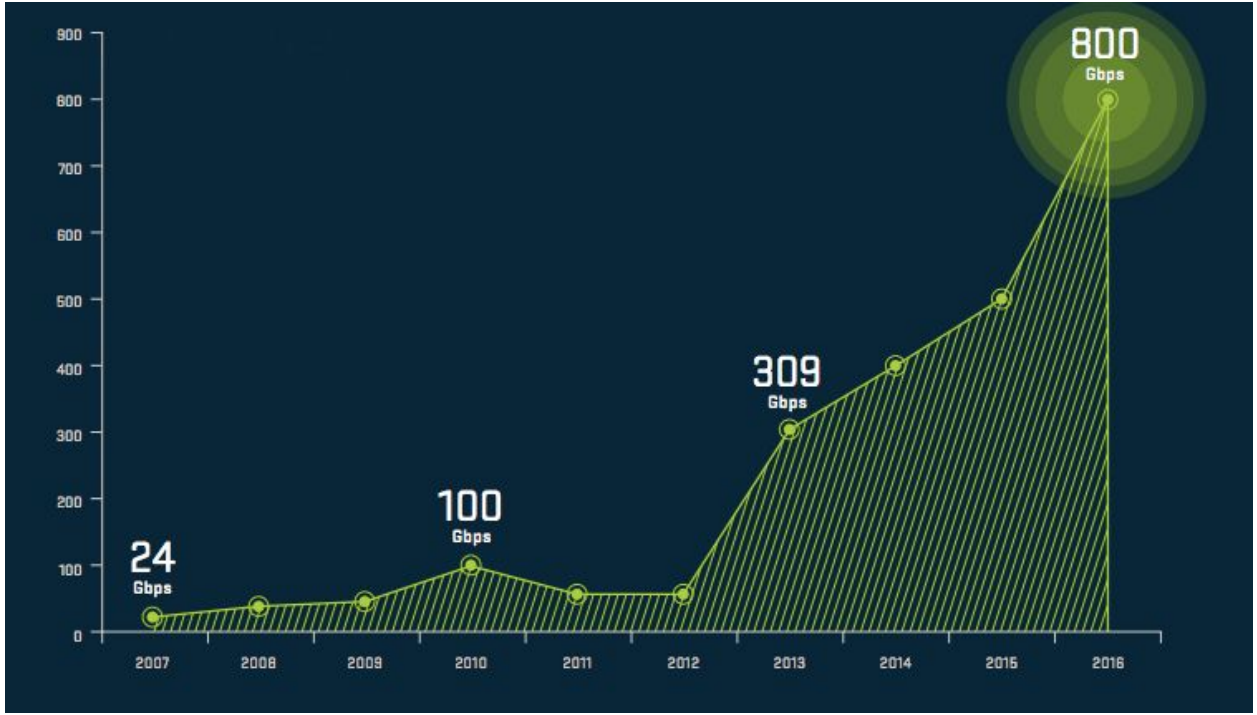
DOS and DDoS are part and parcel of servers life

DOS and DDoS attacks are very common attack vectors used nowadays to bring down the servers or flood the network.

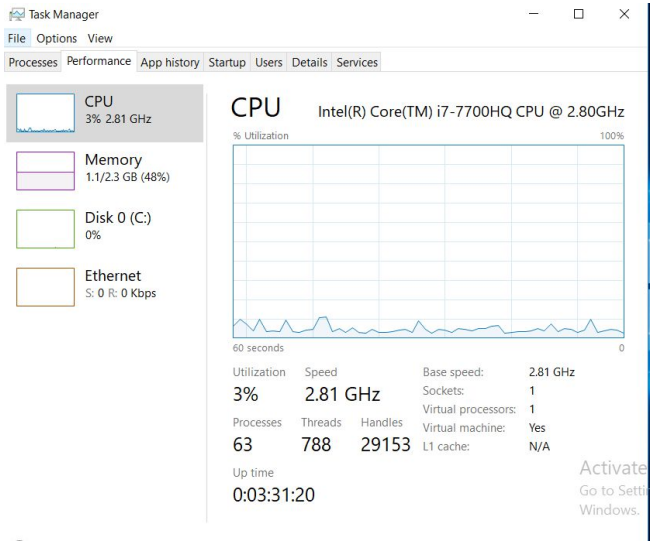
The reason why they are so successful is because of ease of ability to launch the attack and most of the protection mechanisms are based on expensive hardware.



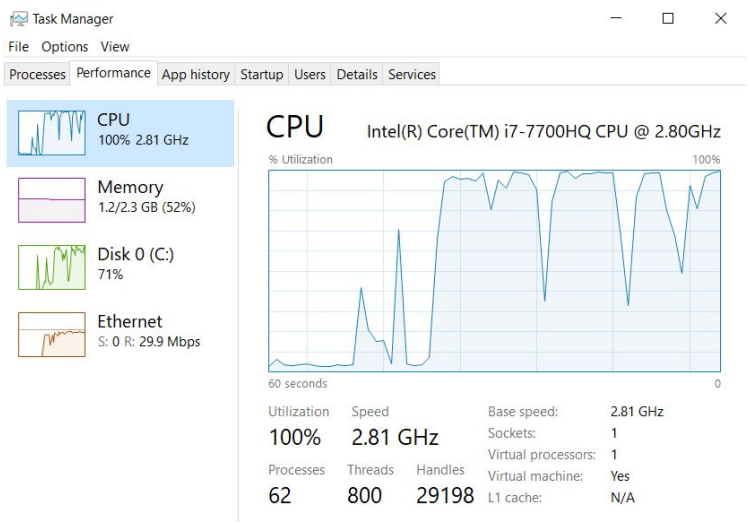
DDOS attacks are going really big!



Before vs After (DOS Attack)

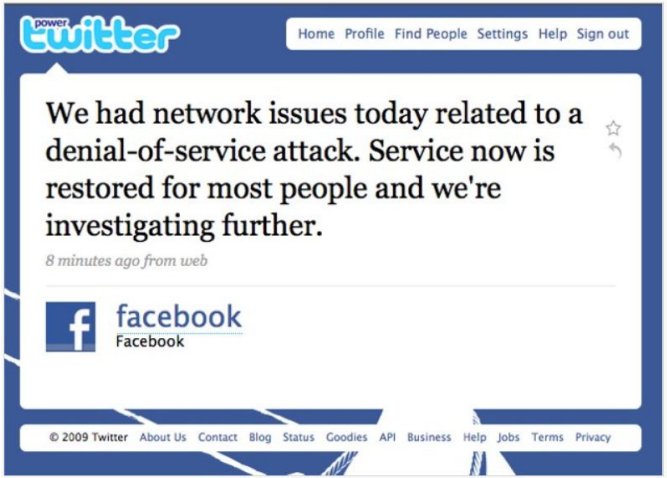






DDOS Attacks Crush Twitter, Hobble Facebook

Posted Aug 6, 2009 by Michael Arrington (@arrington)



Crunchbase

Facebook –

FOUNDED
2004

OVERVIEW
Facebook is an online social networking service that allows its users to connect with friends and family as well as make new connections. It provides its users with the ability to create a profile, update information, add images, send friend requests, and accept requests from other users. Its features include status update, photo tagging and sharing, and more. Facebook's profile structure includes ...

LOCATION
Menlo Park, California

CATEGORIES
Social Media, Social Network, Social

WEBSITE
http://www.facebook.com

Egress-Only Internet Gateway

IPv6

Understanding Egress-Only IGW

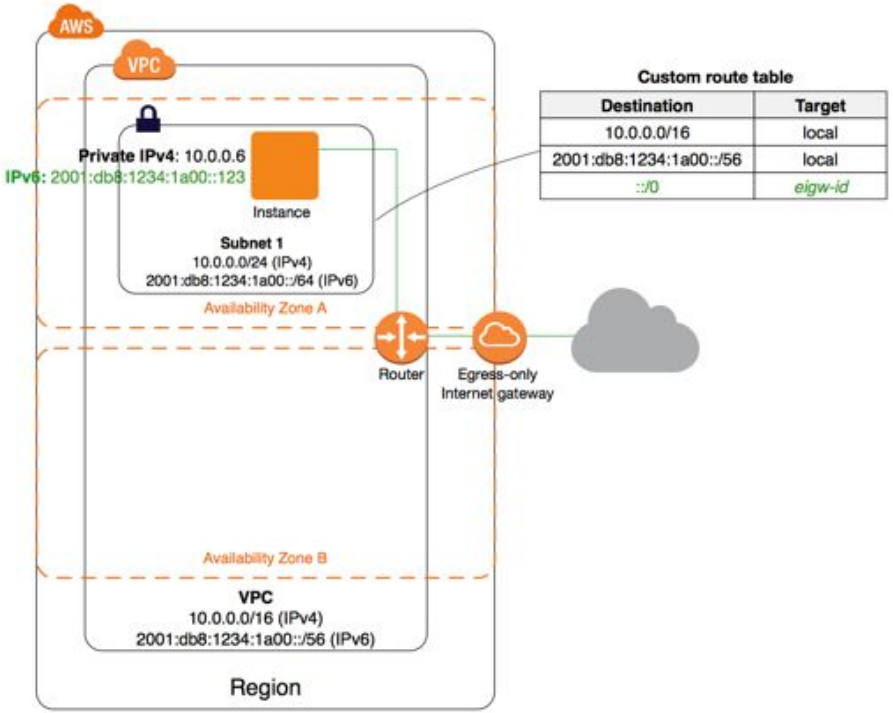
The IPv6 ddresses which are assigned from AWS are public routable addresses.

Thus, instance in the public subnet can initiate connection to Internet via the Internet Gateway. Similarly resources from Internet can also initiate connection to the EC2 instance via it's public IPv4 or IPv6 addresses.

IPv6 addresses are globally unique, and are therefore public by default.

Egress-Only Gateway allows EC2 instance with IPv6 address to access internet directly but prevent resource from internet to directly initiate connection with the EC2 instance.

Architecture for Egress-Only IGW

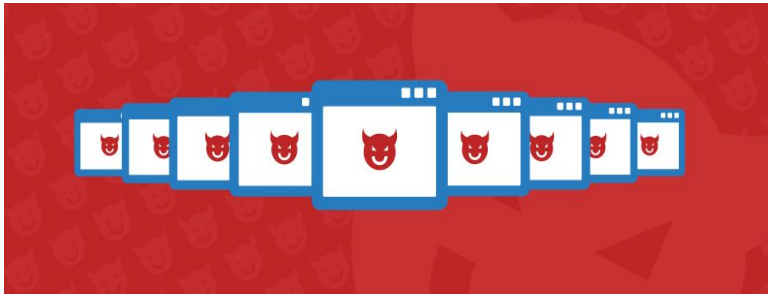


Mitigating DDOS

The stronghold for Fort

Mitigating DDOS

- Be ready to scale as traffic surges.
- Minimize the attack surface area.
- Know what is normal and abnormal.
- Create a Plan for Attacks.



Be Ready to Scale

1. Be Ready to Scale

- Your infrastructure should be designed to scale when the traffic increases.
- It not only helps in Business but also during DDOS Attacks.

Example :

Whenever CPU load is more than 70% in Application servers, automatically add one more Application server to meet the needs.

AWS Services : ELB, Auto Scaling

Let's Minimizing is the Key

2. Minimize the attack surface area.

Decouple your infrastructure.

Example :

Application and Database should not be on the same server.

AWS Services : SQS, Elastic BeanStalk

Normal and Abnormal

3. Know what is normal and abnormal

- Key metrics need to be defined to understand the behavior.

Example :

Website getting a huge surge in traffic in the middle of the night at 3 AM

AWS Services :- CloudWatch, SNS.

Create a Plan

4. Create a Plan for Attacks.



For example :

- Check whether the Source IP Address is the same.
- Check from which country the increased traffic is coming from.
- Nature of the attack (SYN Flood, Application Level)
- Can it be blocked with NACL or Security Group level.

It is recommended to have AWS Support. At-least Business Support.

AWS Services for DDoS Attack Mitigation

Following are some of the key AWS services involved in DDoS attack mitigation

- **AWS Shield**
- **Amazon CloudFront**
- **Amazon Route53**
- AWS WAF
- Elastic Load Balancing
- VPC & Security Groups

AWS Shield

DDoS Protection

Understanding AWS Shield

AWS Shield is a managed Distributed Denial of Service (DDoS) service that safeguards the workloads running on AWS against DDoS attacks.

There are two tiers of AWS Shield:

- Shield Standard
- Shield Advanced

Understanding AWS Shield

AWS Shield standard provides basic level protection against most common network and transport layer DDoS attacks.

For a higher level of protection, we can subscribe to the Shield Advanced. Shield Advanced protects against large and sophisticated DDoS attacks with near-real-time visibility into the attacks that might be occurring.

AWS Shield Advanced also gives customers 24x7 access to the AWS DDoS Response Team (DRT) during ongoing attacks.

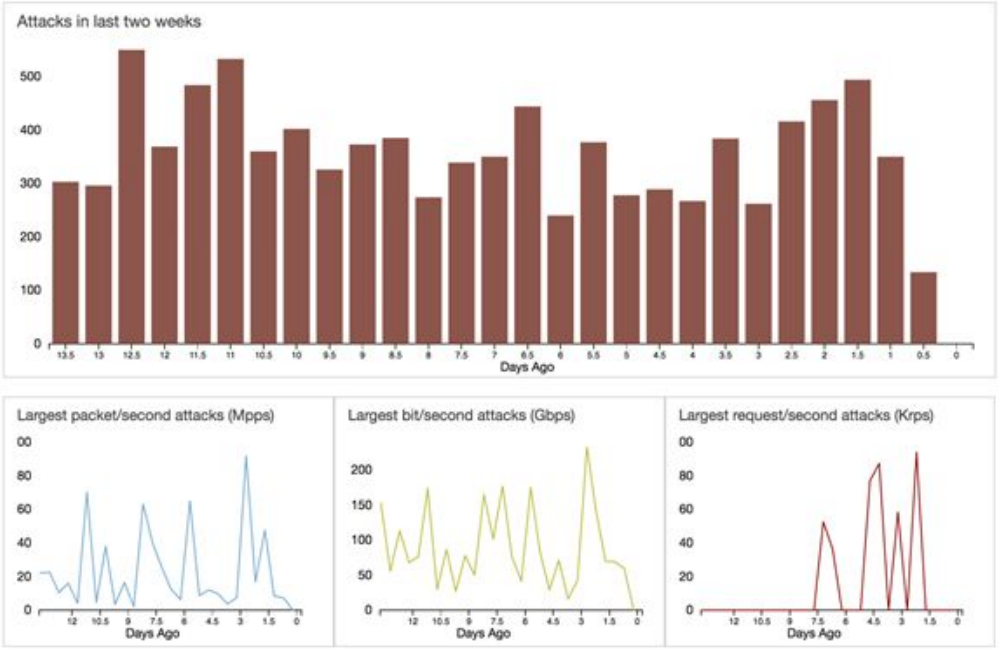
AWS Shield Costs and Credits

AWS Shield Advanced costs 3000$ per organization and requires Business or Enterprise Support.

One interesting part about AWS Shield Advanced is that during the attack, if your infrastructure has scaled, AWS will return you the amount occurred during scaling in the form of credits. This is also referred to as Cost protection.



AWS Shield Dashboard

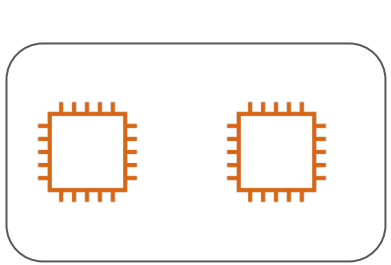


AWS Network Firewall

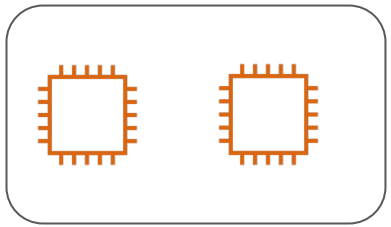
Yet Another Firewall

Basics of Network Firewall

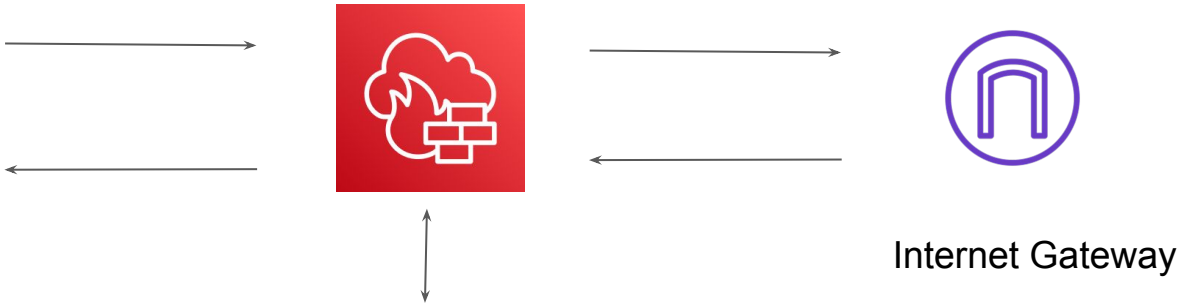
AWS Network Firewall is a stateful, managed, network firewall and intrusion detection and prevention service for your virtual private cloud (VPC)



Public Subnet



Private Subnet



Internet Gateway

IP / Domain	Action
facebook.com	BLOCK
twitter.com	ALLOW
10.77.0.0/16	ALLOW

Benefits of Network Firewall

You can use Network Firewall to monitor and protect your Amazon VPC traffic in a number of ways, including the following:

1. Pass traffic through only from known AWS service domains or IP address endpoints, such as Amazon S3.

2. Use custom lists of known bad domains to limit the types of domain names that your applications can access

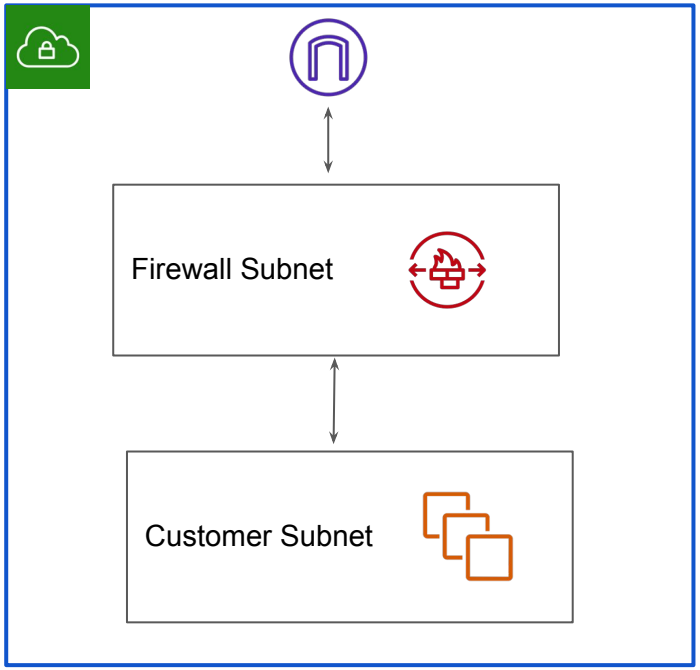
3. Perform deep packet inspection on traffic entering or leaving your VPC

Deploying Network Firewall

Let's Deploy Network Firewall

Basic Deployment Architecture

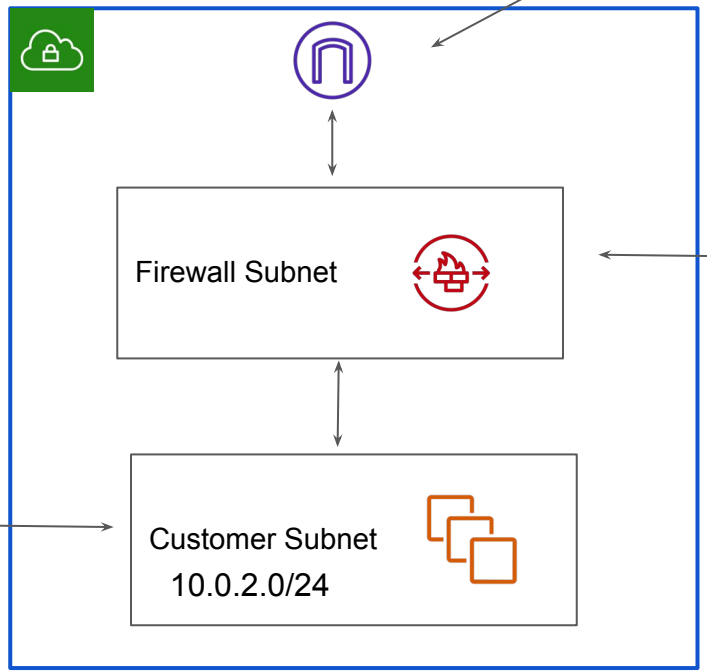
The Network firewall protects the subnets within your VPC by filtering traffic going between the subnets and locations outside of your VPC



Route Table Entries

Destination	Target
10.0.2.0/24	vpce-1234

IGW



Destination	Target
10.0.0.0/16	local
0.0.0.0/0	igw-1234

Firewall Subnet

Destination	Target
10.0.0.0/16	local
0.0.0.0/0	vpce-1234

Customer Subnet

Configuration Steps

Following are the 3 resource types that Network Firewall Manages.

Resource Type	Description
RuleGroup	Defines a set of rules to match against VPC traffic, and the actions to take when Network Firewall finds a match.
FirewallPolicy	Allows adding multiple rule groups and configure other settings.
Firewall	Provides traffic filtering logic for the subnets in a VPC.

CloudHSM

Secure Storage

Storing Expensive House Hold Items

You have an expensive jewellery in your house and you are planning to go on a long vacation.

Where will you prefer to store the jewellery?



Cupboard



Bank Locker

Storing Sensitive Digital Keys

You have sensitive encryption keys that needs to be stored

Where will you prefer to store the keys?



Notepad



Special Security Devices

Special Security Device - HSM

A hardware security module (HSM) is a physical device that provides extra security for sensitive data

This type of device is used to provision cryptographic keys for critical functions such as encryption, decryption and authentication for the use of applications, identities and databases.



Tamper Resistant

- These devices are tamper resistant , that means if anyone tries to tamper, they will automatically delete the keys stored.

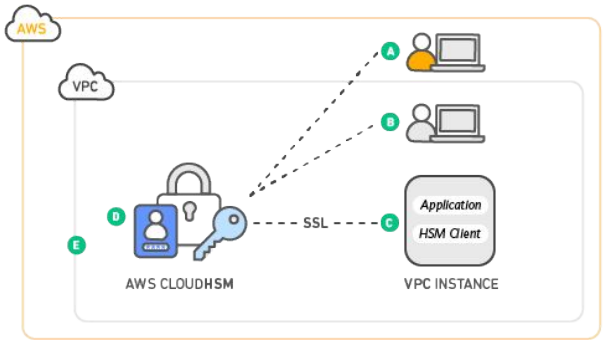


CloudHSM

AWS CloudHSM is a cloud-based hardware security module (HSM).

With CloudHSM, you can manage your own encryption keys using FIPS 140-2 Level 3 validated HSMs.

Prior to this, company's had to store HSM on-premise and if infrastructure was on AWS, there were lot of latency involved.



Important Points for Exams

- Cloud HSM is Single Tenanted (Single Physical Device only for you)
- It must be used within a VPC.
- We can integrate Cloud HSM with RedShift & RDS for Oracle.
- For fault tolerance, we will need to build cluster of 2 Cloud HSM.
- AWS uses Safenet Luna SA HSM appliance for Cloud HSM.
- They are FIPS validated.

- It generally has 2 partitions, one for AWS to monitor and second is cryptographic partition which you have access to and has stored keys.

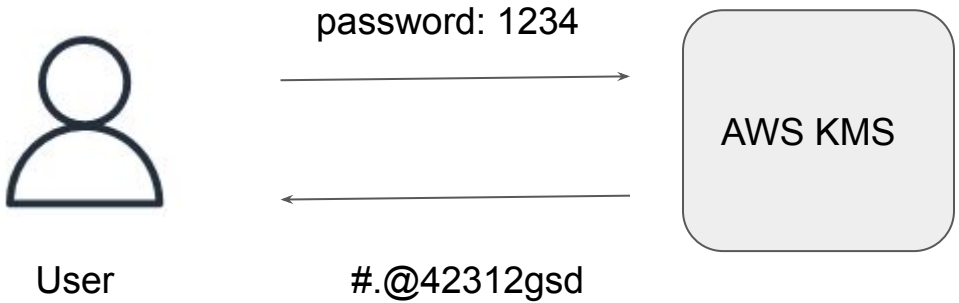
AWS KMS

Do things the right way

Basics of KMS

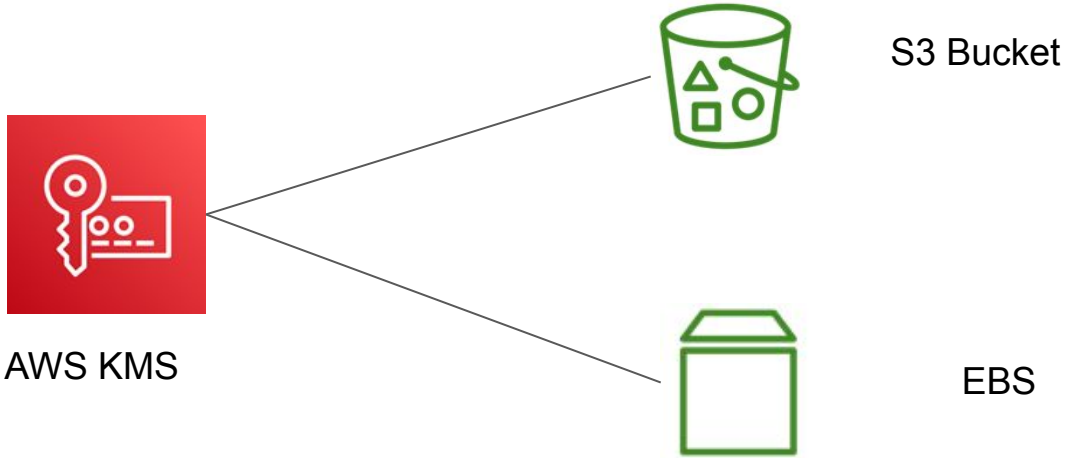
AWS KMS stands for AWS Key Management Service.

This service provides capability to encrypt and decrypt the data.



Integration of KMS

AWS KMS also integrates with various AWS services like S3, DynamoDB, EBS and others.



KMS Practical

Time to Defend Easily

Revising Cryptography Concepts

Plaintext can refer to anything which humans can understand and/or relate to. This may be as simple as English sentences or even Python code.

Ciphertext, or encrypted text, is a series of randomized letters and numbers which humans cannot make any sense of.

An encryption algorithm is step by step approach that tells on how the PT will be converted to the CipherText.

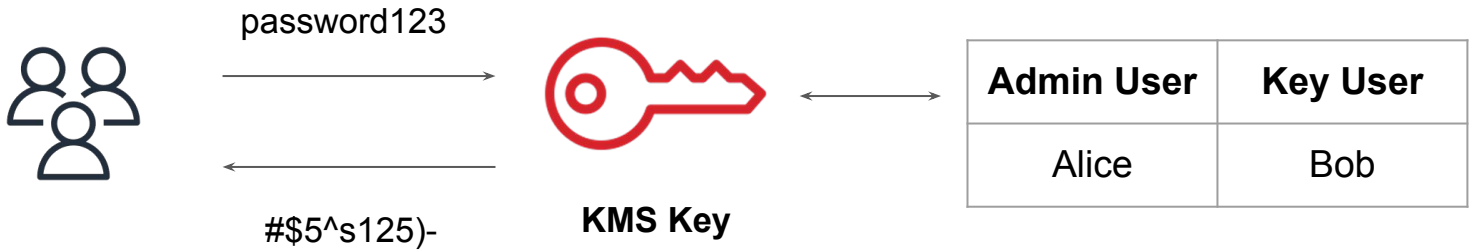


KMS Practical Workflow

1. Create a Customer Managed Key (CMK)

2. Define the Administrative User & Key User.

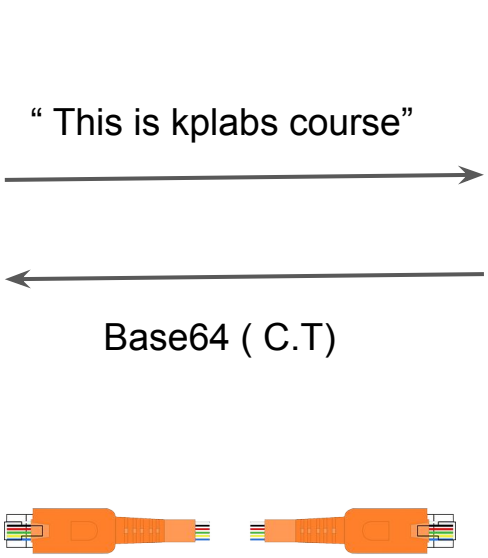
3. Encrypt and Decrypt data with the CMK.

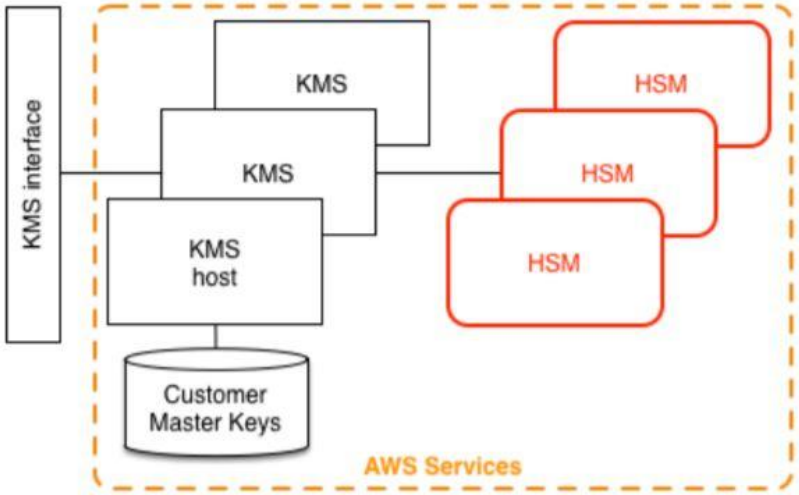


KMS Architecture

Let's Scramble

AWS KMS Architecture





Some of Caveats

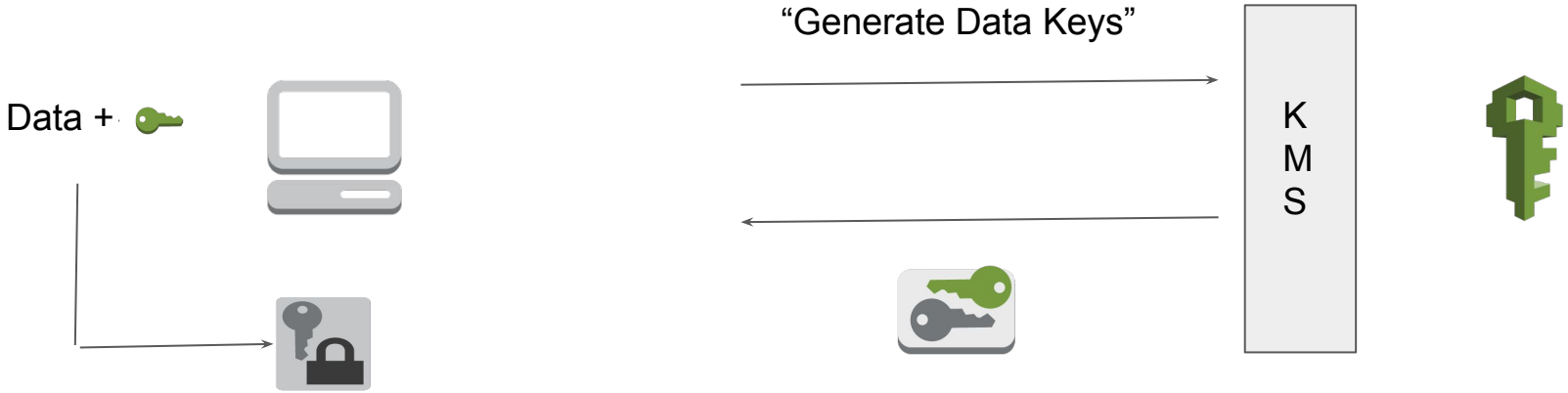
- We can encrypt of maximum 4 KB of data with CMK.
- Since data travels over network, there can be latency issue.

- AWS suggested the Customer Master Key + Data Key based approach.



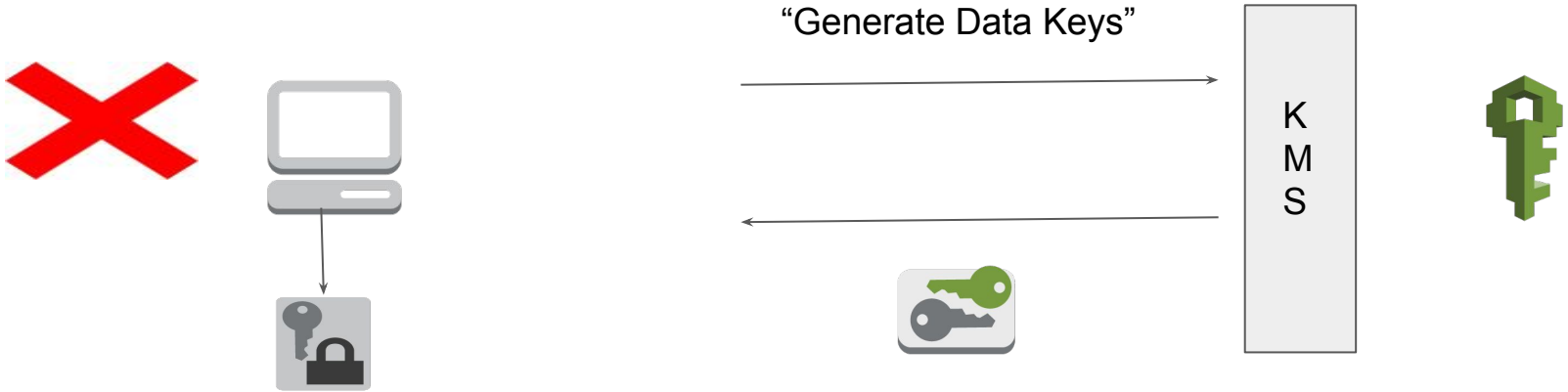
Envelope Encryption

- We generate 1 CMK.
- We then generate the Data Key. AWS returns PT & CT version of it.
- We use the PlainText data key to encrypt the files in server.
- We then store CipherText Data Key along with Encrypted file.



Envelope Encryption

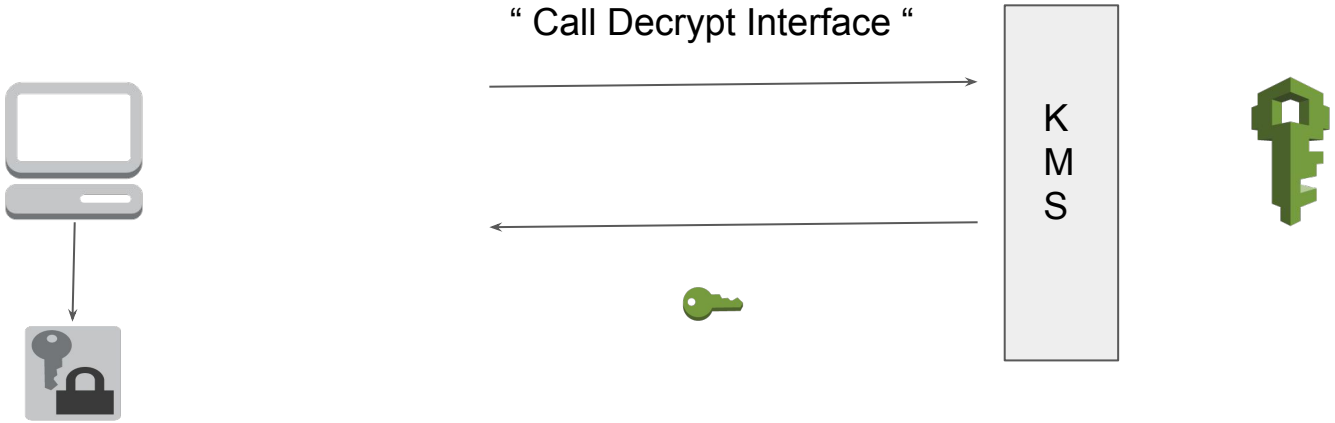
- We generate 1 CMK.
- We then generate the Data Key. AWS returns PT & CT version of it.
- We use the PlainText data key to encrypt the files in server.
- We then store CipherText Data Key along with Encrypted file.



Decryption Steps

- Use the decrypt operation to decrypt the encrypted data key into a plaintext copy of the data key.

- Use the plaintext data key to decrypt data locally.



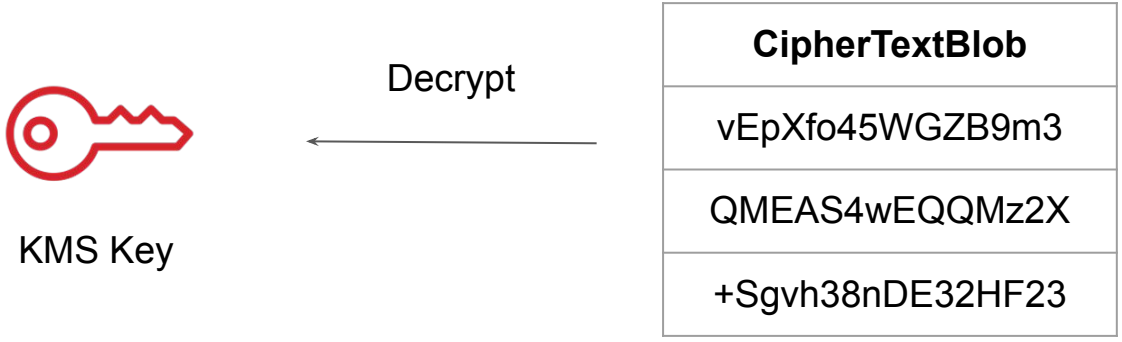
Schedule Key Deletion

Delete the KMS Key

Deleting Key in KMS

Deleting KMS key is destructive and potentially dangerous and an irreversible process.

After a KMS key is deleted, you can no longer decrypt the data that was encrypted under that KMS key, which means that data becomes unrecoverable.



Important Note

You should delete a KMS key only when you are sure that you don't need to use it anymore.

If you are not sure, consider disabling the KMS key instead of deleting it.

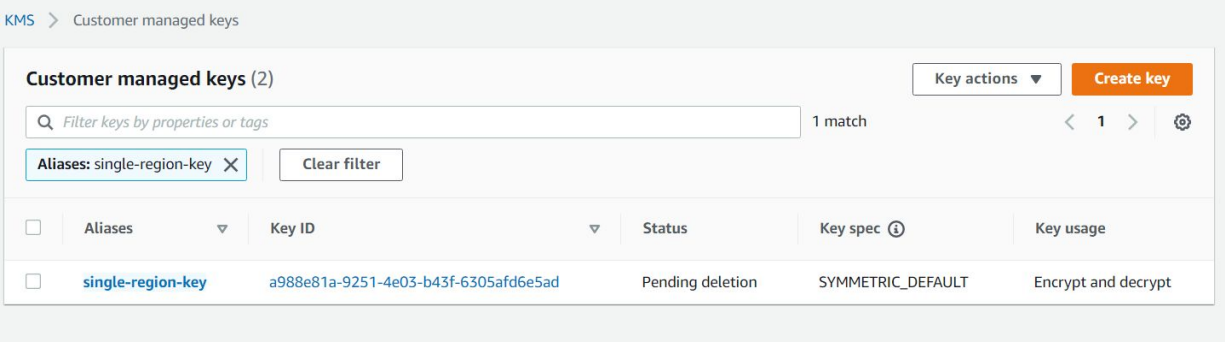
If you disable a KMS key, it cannot be used to encrypt or decrypt data until you re-enable it.

You can re-enable a disabled KMS key if you need to use it again later

Waiting Period for Key Deletion

Because it is destructive and potentially dangerous to delete a KMS key, AWS KMS requires you to set a waiting period of 7 – 30 days. The default waiting period is 30 days.

During the waiting period, A KMS key pending deletion cannot be used in any cryptographic operations.



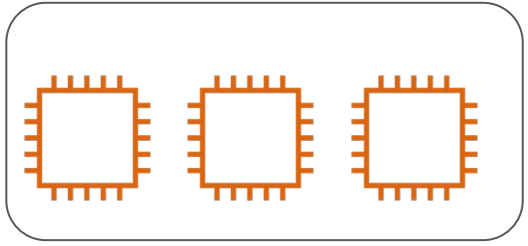
Network ACL

Multiple Layers for Defense

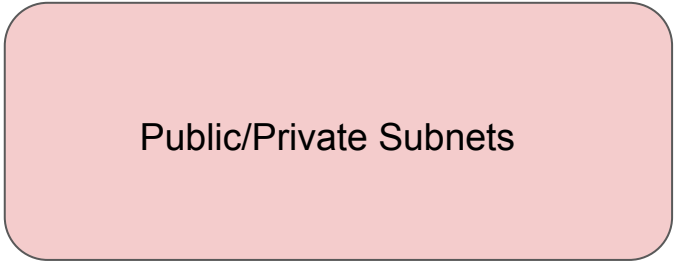
Understanding the Basics

A network access control list (ACL) is an optional layer of security for your VPC that acts as a firewall for controlling traffic in and out of one or more subnets.

- Security Group works at an EC2 instance level.
- Network ACL works at a Subnet Level.



Security Group

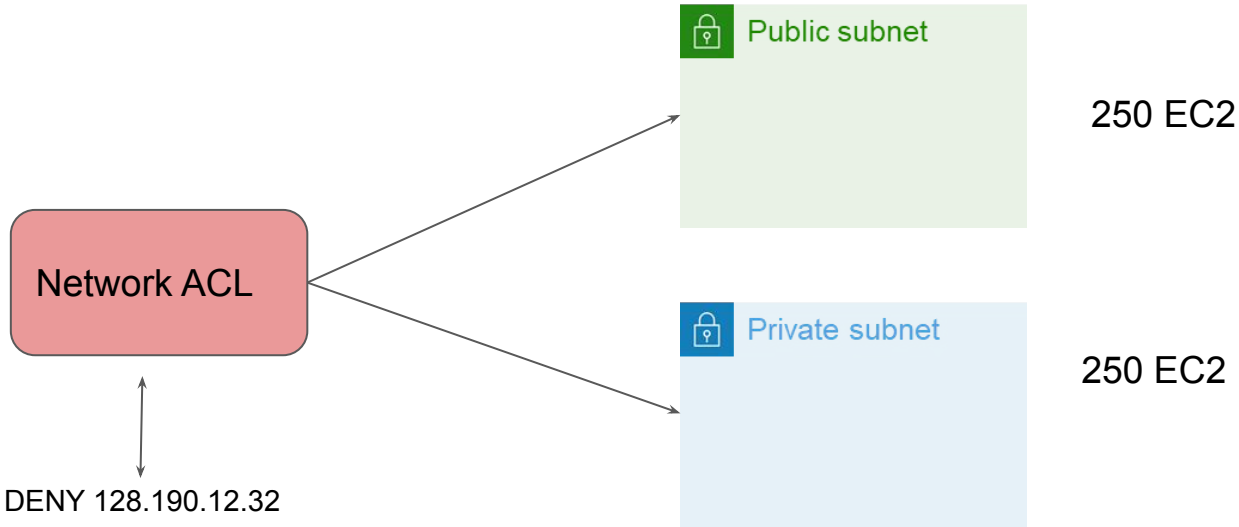


Network ACL

Understanding with Use-Case

Company XYZ is getting lot of attacks from a random IP 128.190.12.32. The company has more than 500 servers and Security team decided to block that IP in firewall for all the servers.

How to go ahead and achieve that goal ?



Important Pointers

Each subnet in your VPC must be associated with a network ACL. If you don't explicitly associate a subnet with a network ACL, the subnet is automatically associated with the default network ACL.

Default NACL allows all inbound and outbound IPv4 traffic and, if applicable, IPv6 traffic.

You can associate a network ACL with multiple subnets. However, a subnet can be associated with only one network ACL at a time.

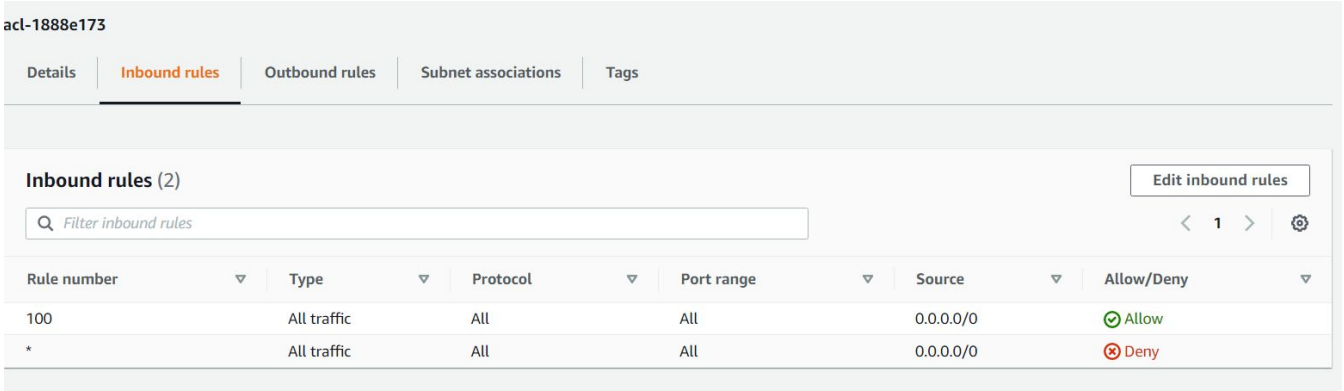
Network ACL - Rule Ordering

Setting Right Set of NACL Rules

Basics of Rules

You can add or remove rules from the default network ACL

When you add or remove rules from a network ACL, the changes are automatically applied to the subnets that it's associated with.



Rule Ordering

Rules are evaluated starting with the lowest numbered rule.

As soon as a rule matches traffic, it's applied regardless of any higher-numbered rule that might contradict it.

Rule Number	Rule Contents
99	ALLOW from 10.77.0.5
100	DENY from ALL

Important Pointers - Deciding Ports

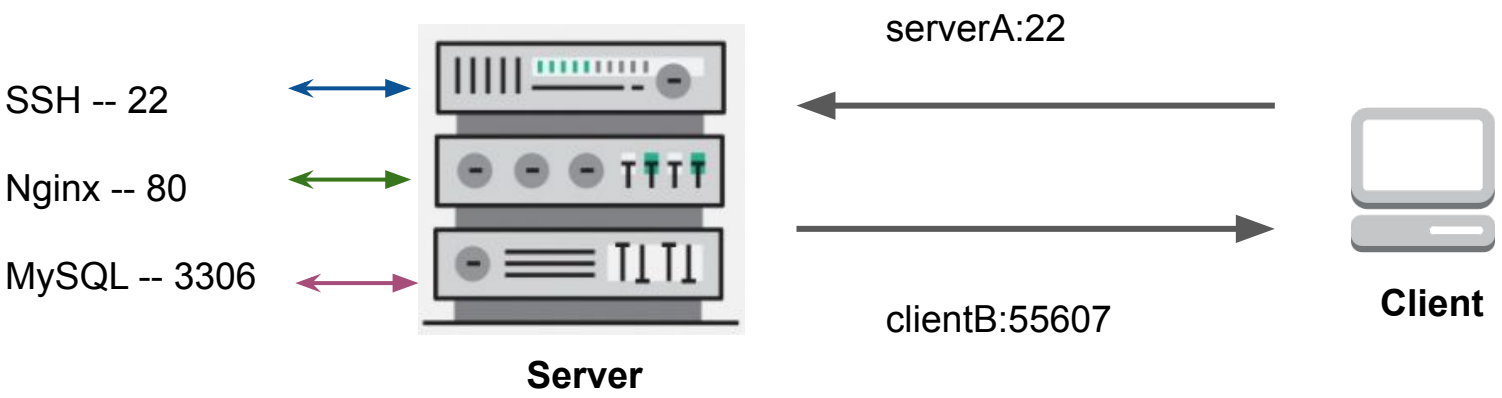
- Clients that initiates the request chooses ephemeral port range.
- Port 0 to 1023 are well known or reserved ports.
- This range varies depending on the Operating System.

Example :-

Many Linux kernels uses ports 32768-61000.

Request originating from the ELB uses 1024-65535

Windows XP uses 1025-5000 port range.



- Clients opens an port 55607 from which it sends data to serverA port 22

- serverA has to respond back to the same IP (clientB) & port (55607).

TCP/IP Communication

```
fczv329@fcblr-l003:~/Documents$ cat handshake
19:46:27.378297 3c:a9:f4:a0:fa:e0 > 08:5b:0e:47:be:1e, ethertype IPv4 (0x0800), length 74: 172.20.1.55.55427 > 128.199.106.4.80: Flags [S], seq
 3002048179, win 29200, options [mss 1460,sackOK,TS val 202385607 ecr 0,nop,wscale 7], length 0


19:46:27.798037 08:5b:0e:47:be:1e > 3c:a9:f4:a0:fa:e0, ethertype IPv4 (0x0800), length 74: 128.199.106.4.80 > 172.20.1.55.55427: Flags [S.], se
q 2402250441, ack 3002048180, win 14480, options [mss 1460,sackOK,TS val 2028995051 ecr 202385607,nop,wscale 8], length 0


19:46:27.798119 3c:a9:f4:a0:fa:e0 > 08:5b:0e:47:be:1e, ethertype IPv4 (0x0800), length 66: 172.20.1.55.55427 > 128.199.106.4.80: Flags [.], ack
 1, win 229, options [nop,nop,TS val 202385712 ecr 2028995051], length 0
```

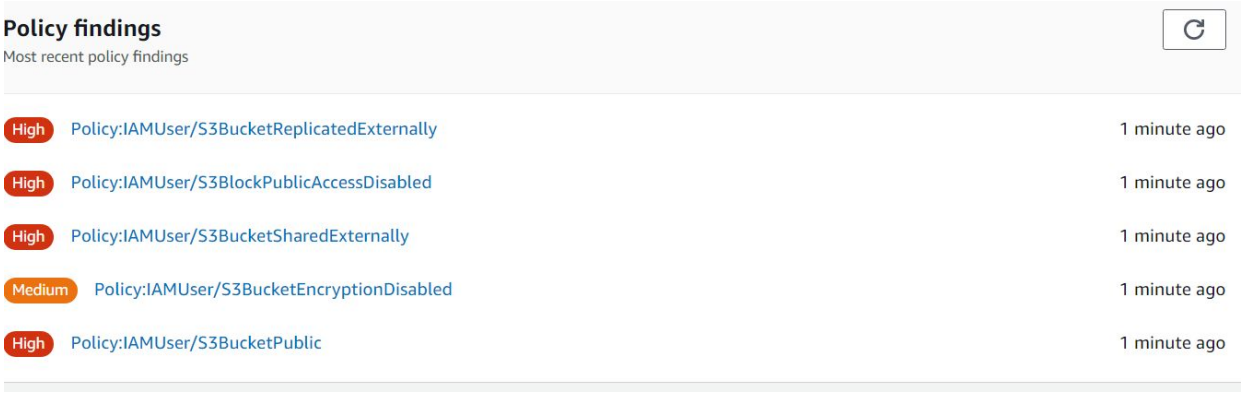
Amazon Macie

Machine Learning based Security

Core Feature of Macie

S3 might contain sensitive information like PII data, database backups, SSL private keys and various others.

Amazon Macie makes use of machine learning to identify sensitive data stored in AWS.

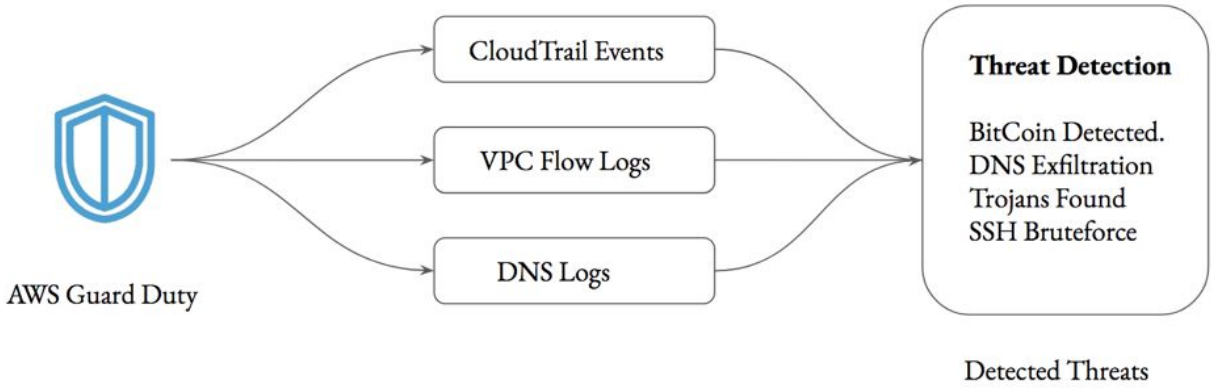


AWS GuardDuty

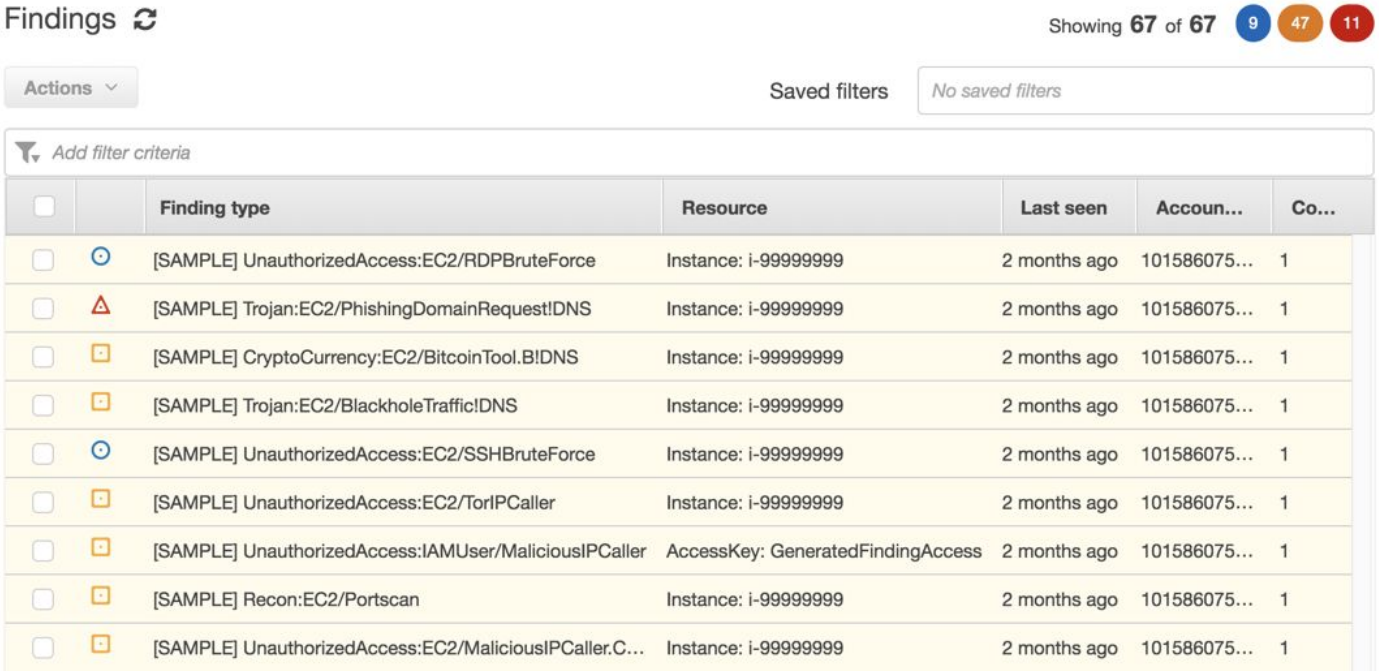
Let's start Rolling !

Understanding GuardDuty

AWS Guard Duty is a threat intelligence service by AWS which monitors for malicious behavior to help customers protect their AWS workloads.



GuardDuty Findings



Important Pointers

Guard Duty will only monitor the Route53 for DNS Logs.

Lot of organizations makes use of Active Directory DNS. The logs from these servers will not be monitored.

Relax and Have a Meme Before Proceeding



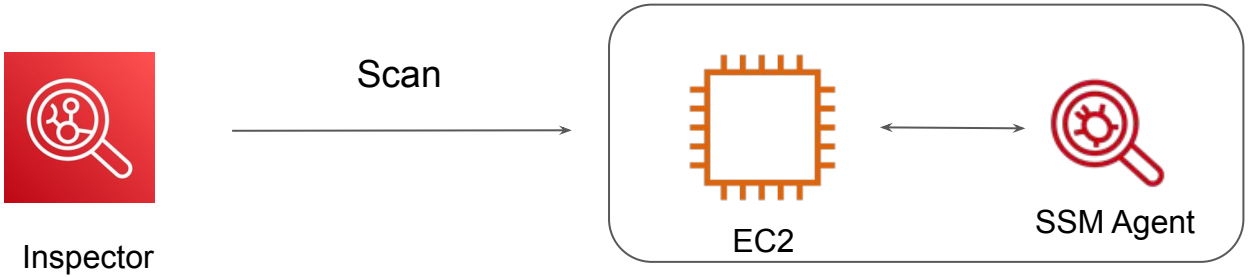
AWS Inspector

Vulnerability Scanner

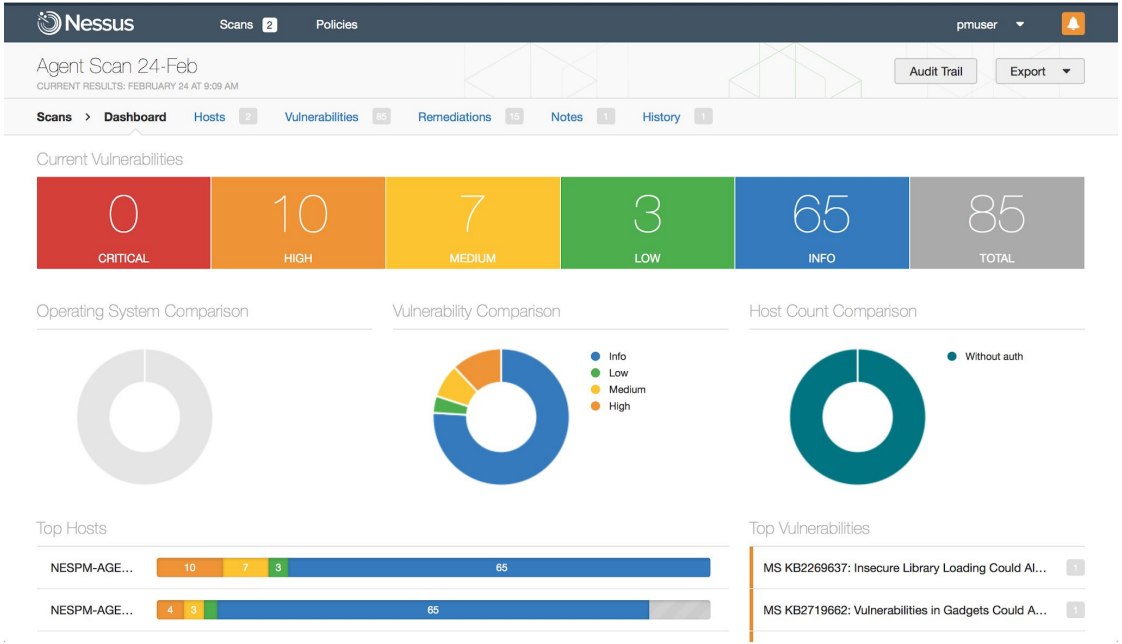
Basics of AWS Inspector

AWS Inspector is similar to a vulnerability scanner which will scan the system for specific vulnerabilities and provide the results.

It relies on the agent installed on the server to scan the server.

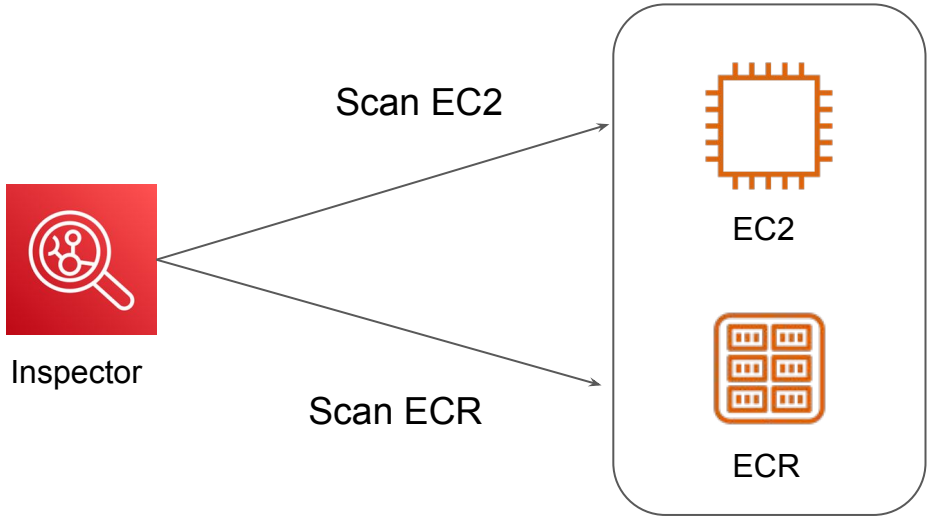


Similar to Nessus



Supported Scans

Amazon Inspector gives you the flexibility to enable either EC2 scanning or ECR container image scanning, or both.

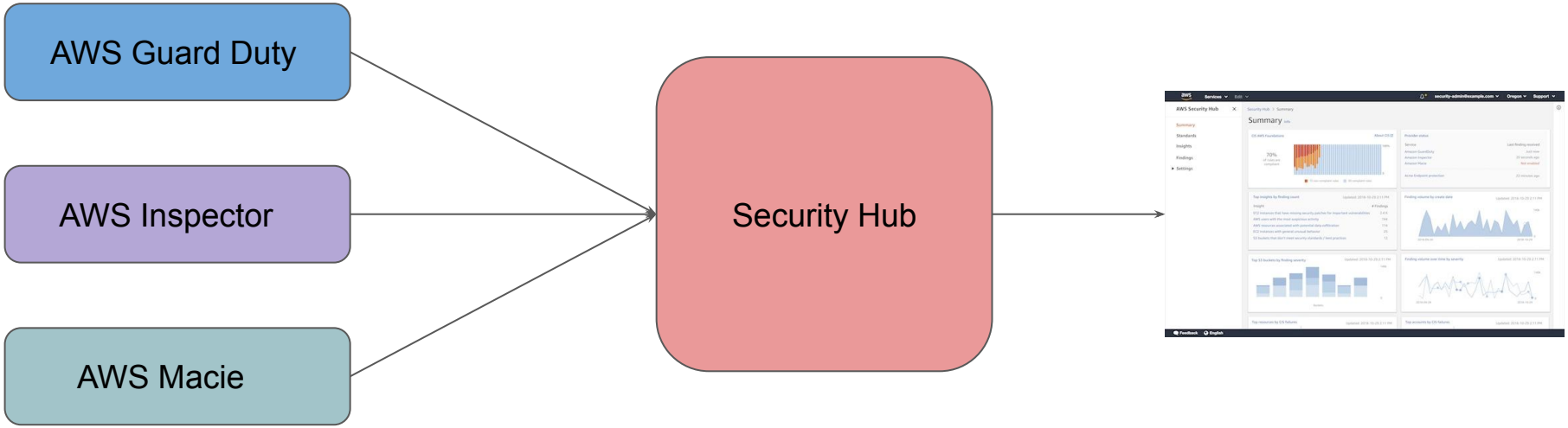


Security Hub

Centralized Security Hub

Overview of Security Hub

AWS Security Hub gives you a comprehensive view of your high-priority security alerts and compliance status across AWS accounts.



Supported Compliance Standard

AWS Security Hub also has ability to generate its own findings by running automated and continuous checks against the rules in a set of supported security standards.

Following Standards are supported:

- CIS AWS Foundation
- PCI DSS



Standard	Passed	Failed	Score ▲
CIS AWS Foundations Benchmark v1.2.0	12	30	29%
PCI DSS v3.2.1	22	9	69%

Web Application Firewall

Next generation firewalls

Getting started

We all know about Firewalls and in some way might have worked as well.

Firewall works on the Layer 3 and Layer 4 of the OSI model.

Main aim of firewall: Block malicious and unauthorized traffic.

However what about malicious traffic like SQL Injection attacks, XSS and many more ?

Introducing WAF

A Web Application Firewall is an application level firewall for HTTP applications.

It applies set of rules for the HTTP based conversations.

WAF generally are deployed to protect against attacks targeted towards application, specifically the ones defined in the OWASP Top 10 metrics.



WAF Vendors

There are lot of ways in which you can implement WAF and various vendors as well.

Naxsi and Modsecurity are some of the famous open sources ones.

Signal Sciences, Akamai, AWS WAF are some of the commercial vendors that offer WAF related functionalities.





AWS WAF

Protection against Layer 7 Attacks

Understanding AWS WAF Concepts

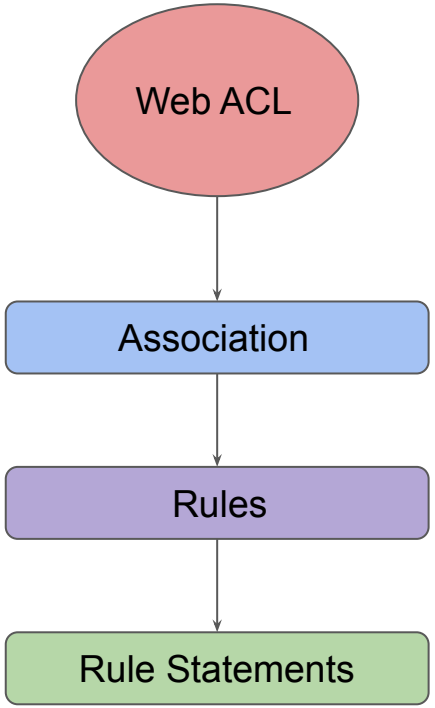
I live in a place A in Bangalore and want to meet my friend living in place B in Bangalore.

Rule Statement: If traffic is less on the roads?
Are there any Uber / OLA available?

Rules: If traffic is less AND uber ola available then yes or no

WebACL: Container for all the things + default action.

Association: Zeal



Rule Statements

Rule Statements define basic characteristics that would be analyzed within a web request.

These can be custom-defined or you can use ready-made ones from AWS and marketplace.

1. Block all the requests which are coming from out of India.
2. Block request which has a URI Path of /admin

You can even build custom condition based on:

Headers, HTTP Method, Query Strings, URI Path, Geo-Location, Body.

Rules in WAF

We can combine multiple statements into rules to precisely target requests.

WAF provides two primary rule types: Regular Rule & Rate-Based rule

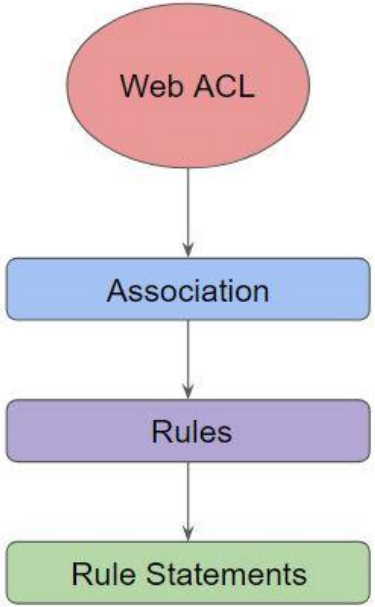
Let's look into sample regular rule:

1. If a request comes from 172.30.0.50
2. Request has SQL-like code

Rules with multiple statements can be AND, OR, NOT

Rate-Based rule = Regular Rule + Rate limiting feature

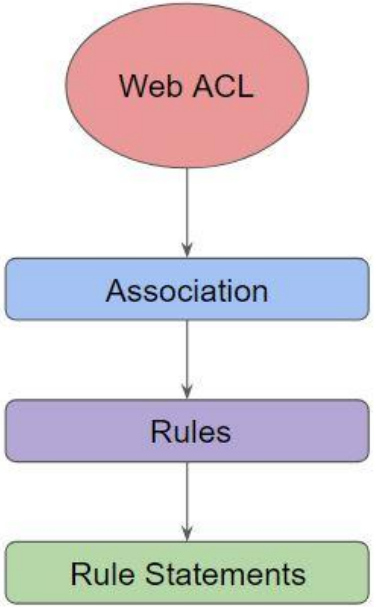
1 If a request comes from 172.30.0.50
2. If requests exceeds 1000 request in 10 minutes



Web ACL in WAF

Web ACL is a centralized place that contains the rules, rule statements and associated configuration.

It is used to define the protection strategy.

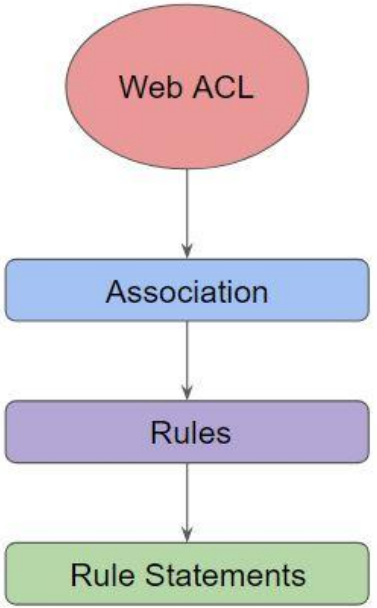


Association in WAF

Association defines to which entity WAF is associated to.

WAF cannot be associated with EC2 instances directly.

Support Association: ALB and CloudFront, API Gateway



Important Pointers

Rule Groups can be configured which has multiple rules that can be used across multiple Web ACLs.

Customers can decide to use ready-made AWS-Managed rules or even rules from AWS Marketplace.

Every Rule has a priority. If a request matches Priority 0 rule, none of the other rules will inspect the request

Pricing Aspect:

Web-ACL ($5 per month), Rule ($1 per month), Requests ($0.60 / 1 million)

Relax and Have a Meme Before Proceeding



HTTPS

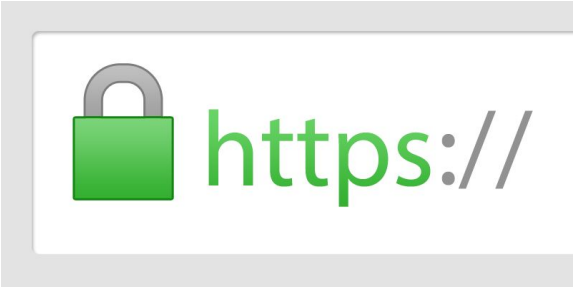
Secure Communication

Overview of HTTPS

HTTPS is an extension of HTTP.

In HTTPS, the communication is encrypted using Transport Layer Security (TLS)

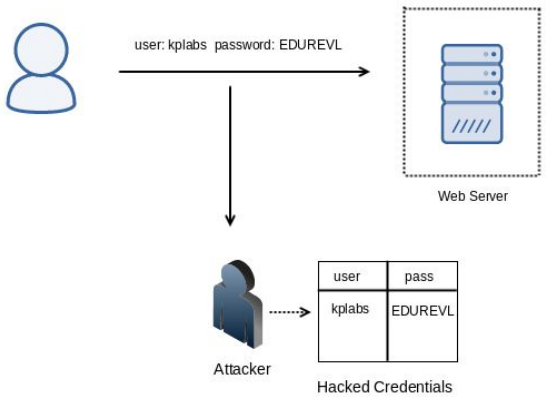
The protocol is therefore also often referred to as HTTP over TLS or HTTP over SSL.



Scenario 1: MITM Attacks

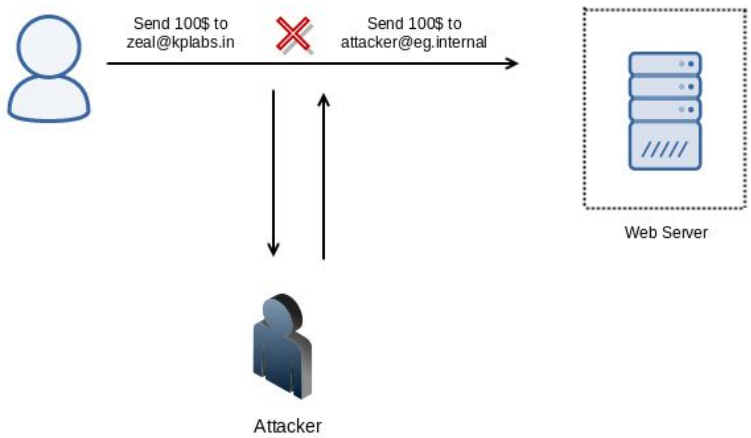
- User is sending their username and password in plaintext to a Web Server for authentication over a network.

- There is an Attacker sitting between them doing a MITM attack and storing all the credentials he finds over the network to a file:



Scenario 2: MITM & Integrity Attacks

- Attacker changing the payment details while packets are in transit.



Introduction to SSL/TLS

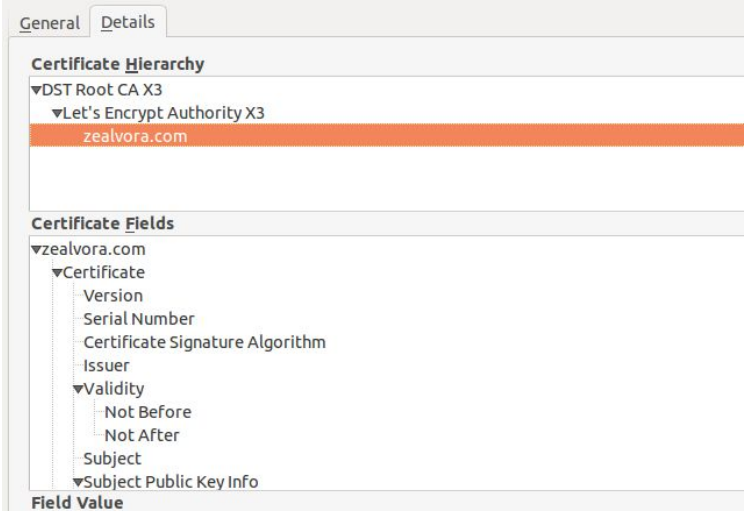
To avoid the previous two scenarios (and many more), various cryptographic standards were clubbed together to establish a secure communication over an untrusted network and they were known as SSL/TLS.

Protocol	Year
SSL 2.0	1995
SSL 3.0	1996
TLS 1.0	1999
TLS 1.1	2006
TLS 1.2	2008
TLS 1.3	2018

Understanding it in easy way

Every website has a certificate (like a passport which is issued by a trusted entity).

Certificate has lot of details like domain name it is valid for, the public key, validity and others.



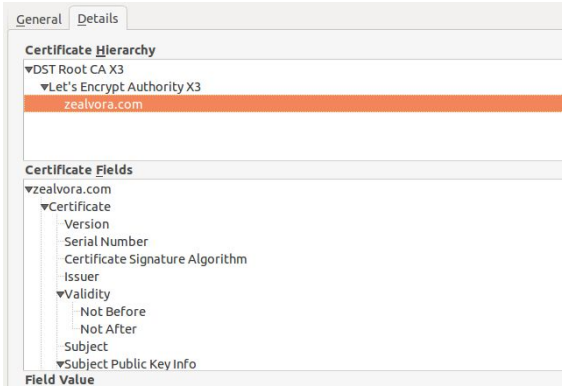
Understanding it in easy way

Browser (clients) verifies if it trusts the certificate issuer.

It will verify all the details of the certificate.

It will take the public key and initiate a negotiation.

Asymmetric key encryption is used to generate a new temporary symmetric key which will be used for secure communication.



Web Server Configuration

```
server {
       listen         80;
       server_name    zealvora.com;
       return         301 https://$server_name$request_uri;
}


server {
  server_name zealvora.com;
  listen 443 default ssl;
  server_name zealvora.com;
  ssl_certificate /etc/letsencrypt/archive/zealvora.com/fullchain1.pem;
  ssl_certificate_key /etc/letsencrypt/archive/zealvora.com/privkey1.pem;

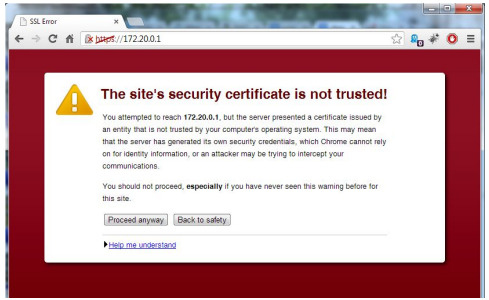
  location / {
    root /websites/zealvora/;
    include location-php;
    index index.php;
 }
 location ~ /.well-known {
        allow all;
 }
}
```

AWS Certificate Manager

Certificates Again :)

Earlier Approach

I have a website and I need to use HTTPS. There are two ways, self-signed certificate and the CA signed certificate.



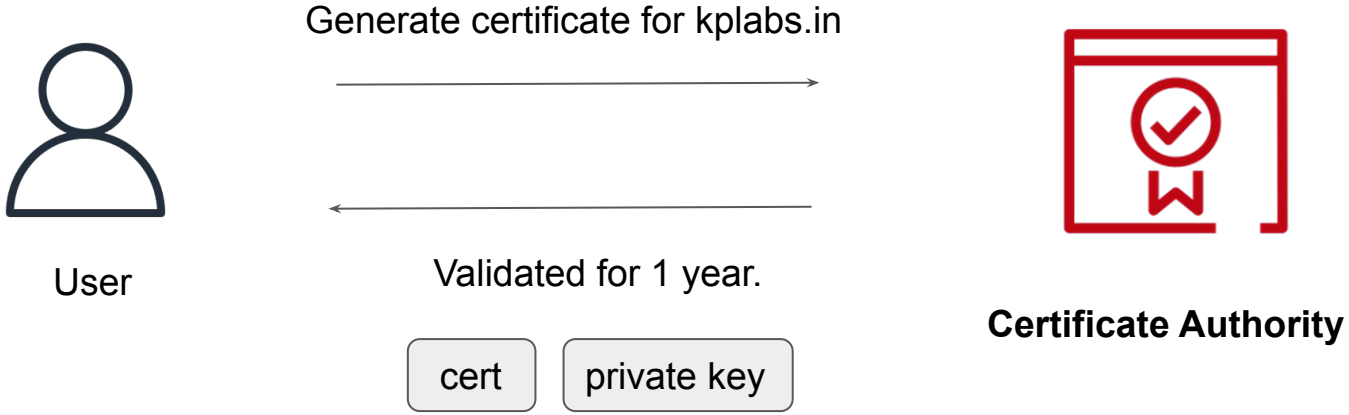
Self Signed Certificate



CA Signed Certificate

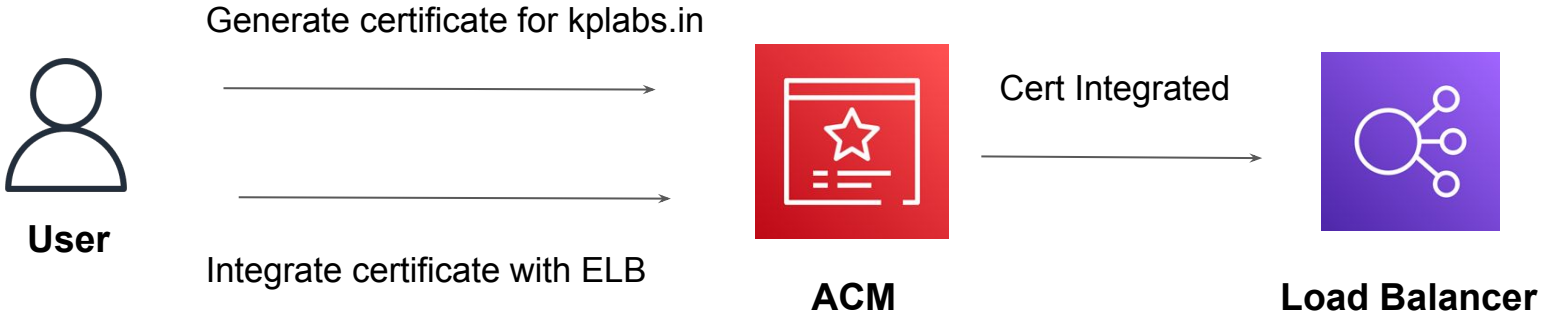
Generating Certificates

To generate a certificate for your domain, you will have to go to a Certificate Authority and after required level of validation, you would be issued a certificate.



AWS Certificate Manager

AWS Certificate Manager (ACM) handles the complexity of creating, storing, and renewing public and private SSL/TLS X.509 certificates and keys that protect your AWS websites and applications.



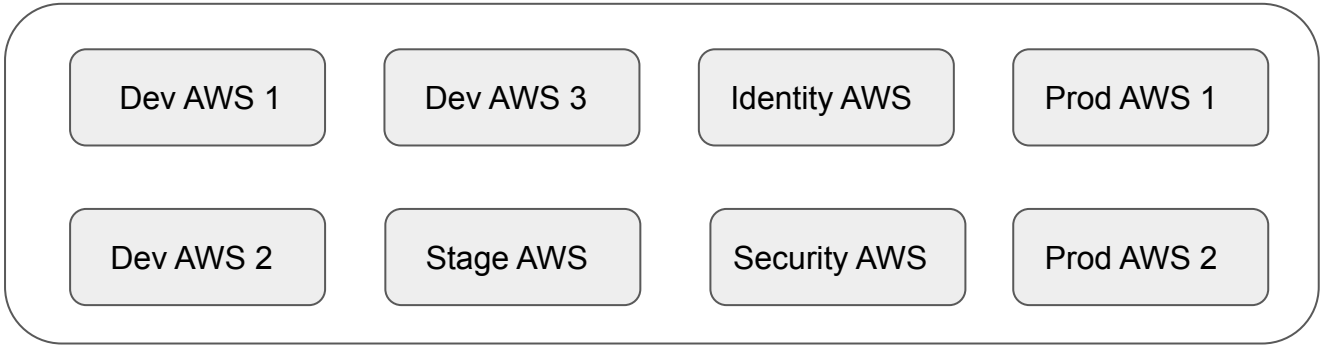
AWS Control Tower

Agility and Governance

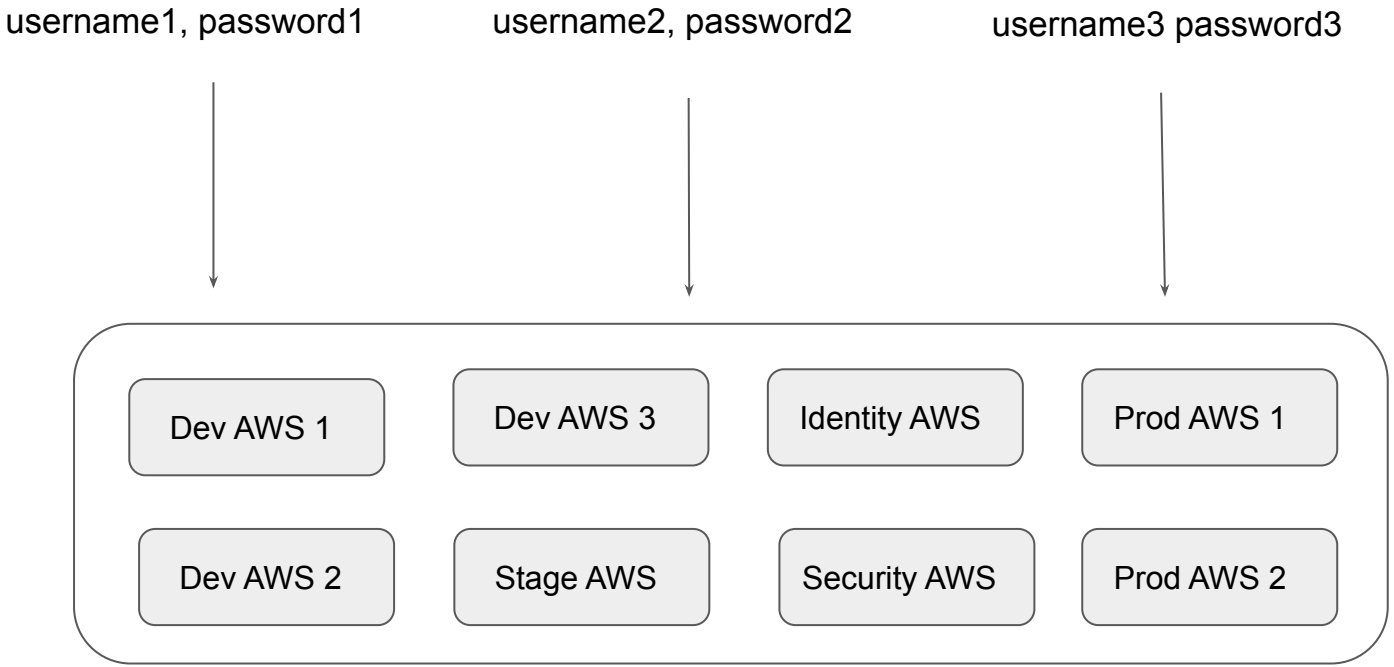
Challenges with Multi-Account Environments

Most of the organizations follow a multi-account based architecture.

When the amount of AWS account increases, it leads to own set of challenges.

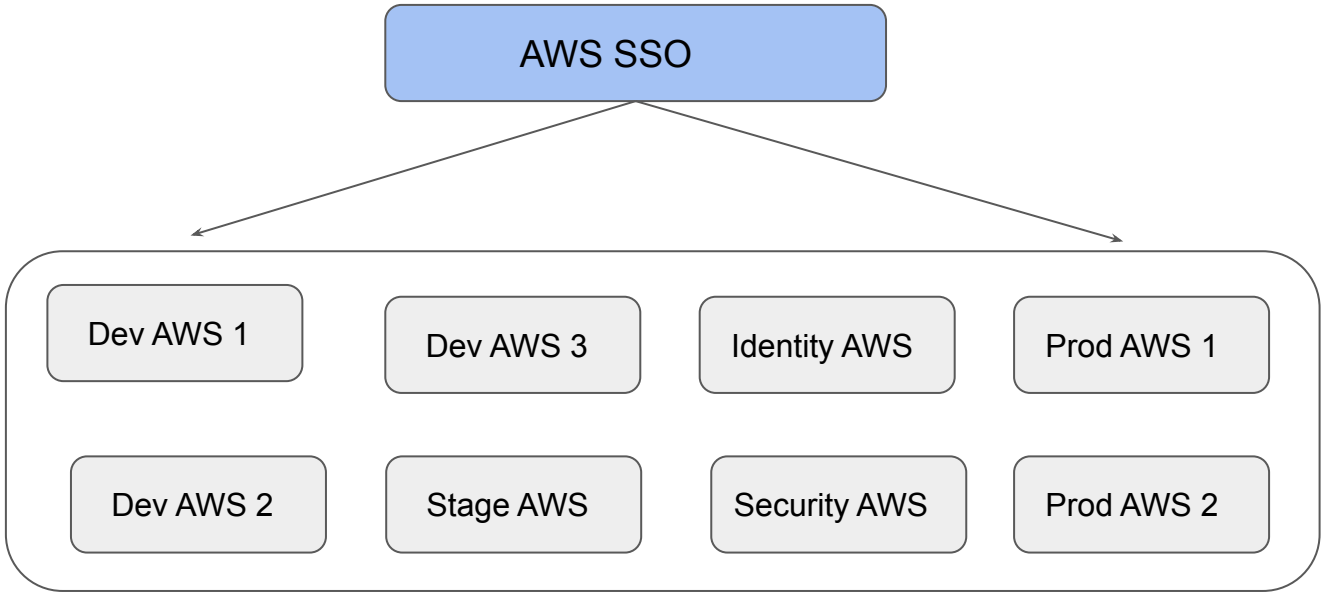


Challenge 1 - Identity Management

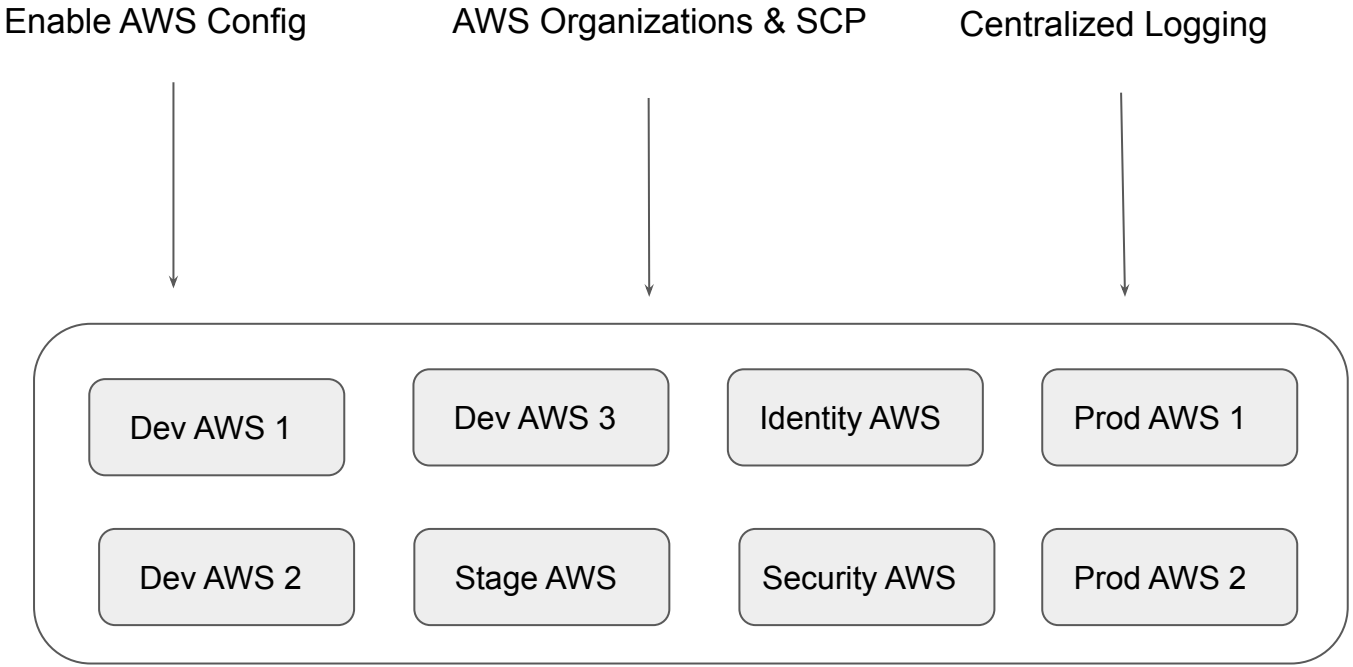


Solution 1 - Single Sign On

Single sign-on (SSO) is an authentication method that enables users to securely authenticate with multiple applications and websites by using just one set of credentials.

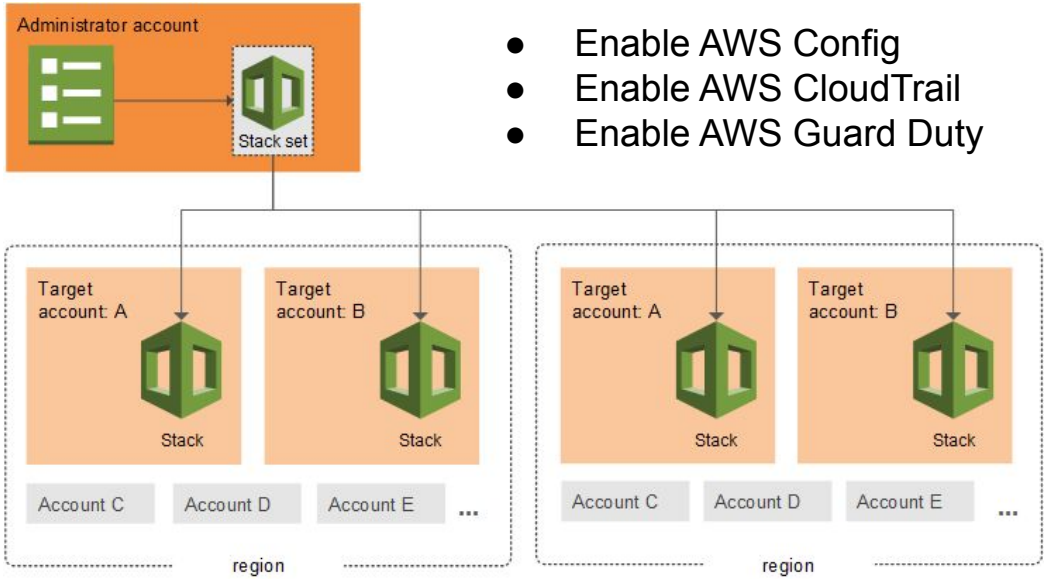


Challenge 2 - Security Hardening



Solution 2 - Security Automation

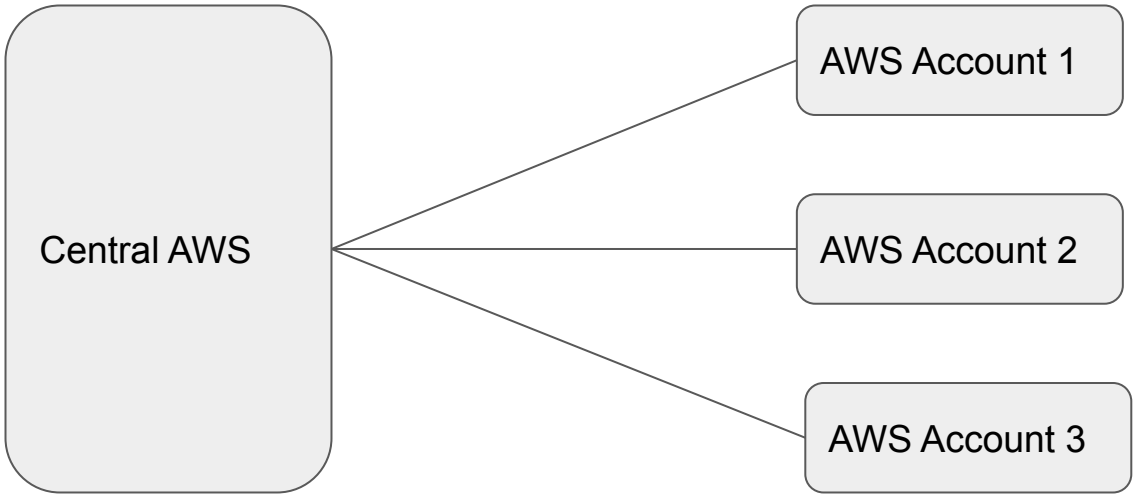
AWS CloudFormation StackSets allows you to create, update, or delete stacks across multiple accounts and Regions with a single operation



- Enable AWS Config
- Enable AWS CloudTrail
- Enable AWS Guard Duty

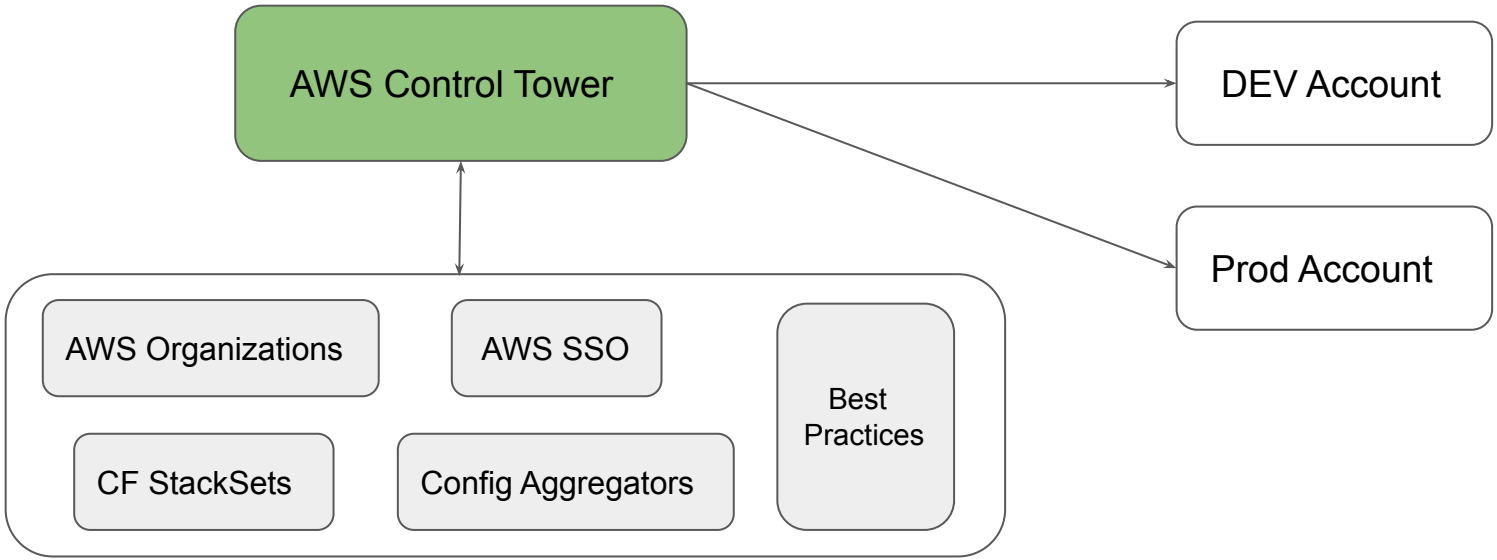
Challenge 3 - Centralized Console

We need to have a centralized console that shows details of all AWS accounts, their security compliance level, and other information



AWS Control Tower

AWS Control Tower offers a straightforward way to set up and govern an AWS multi-account environment, following the best practices.



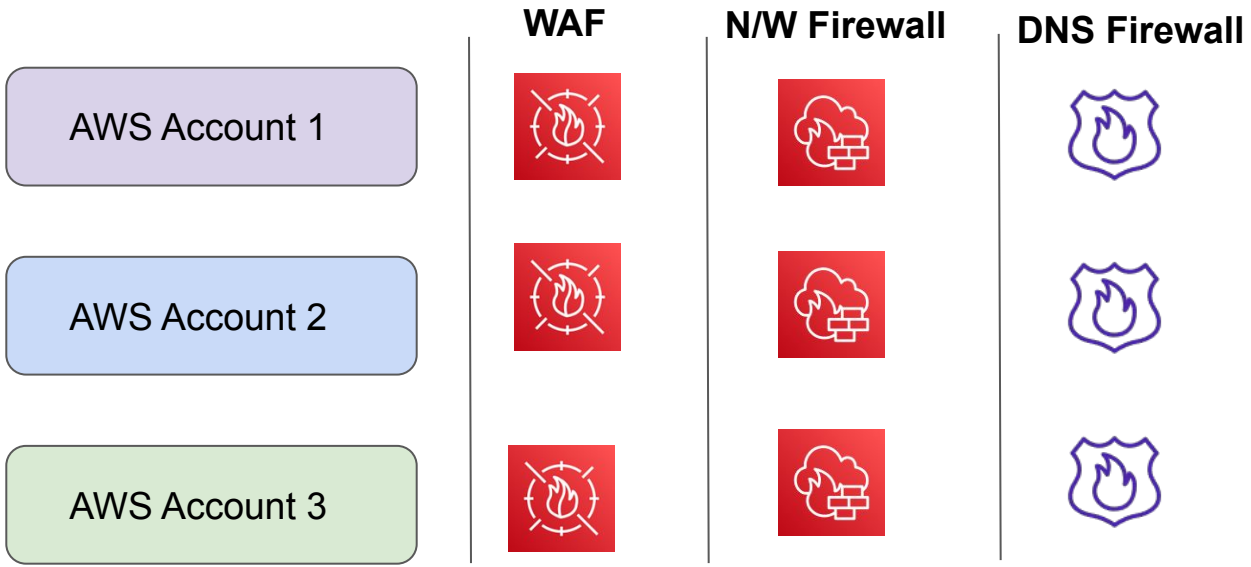
Firewall Manager

Centrally Manage Rules

Understanding the Challenge

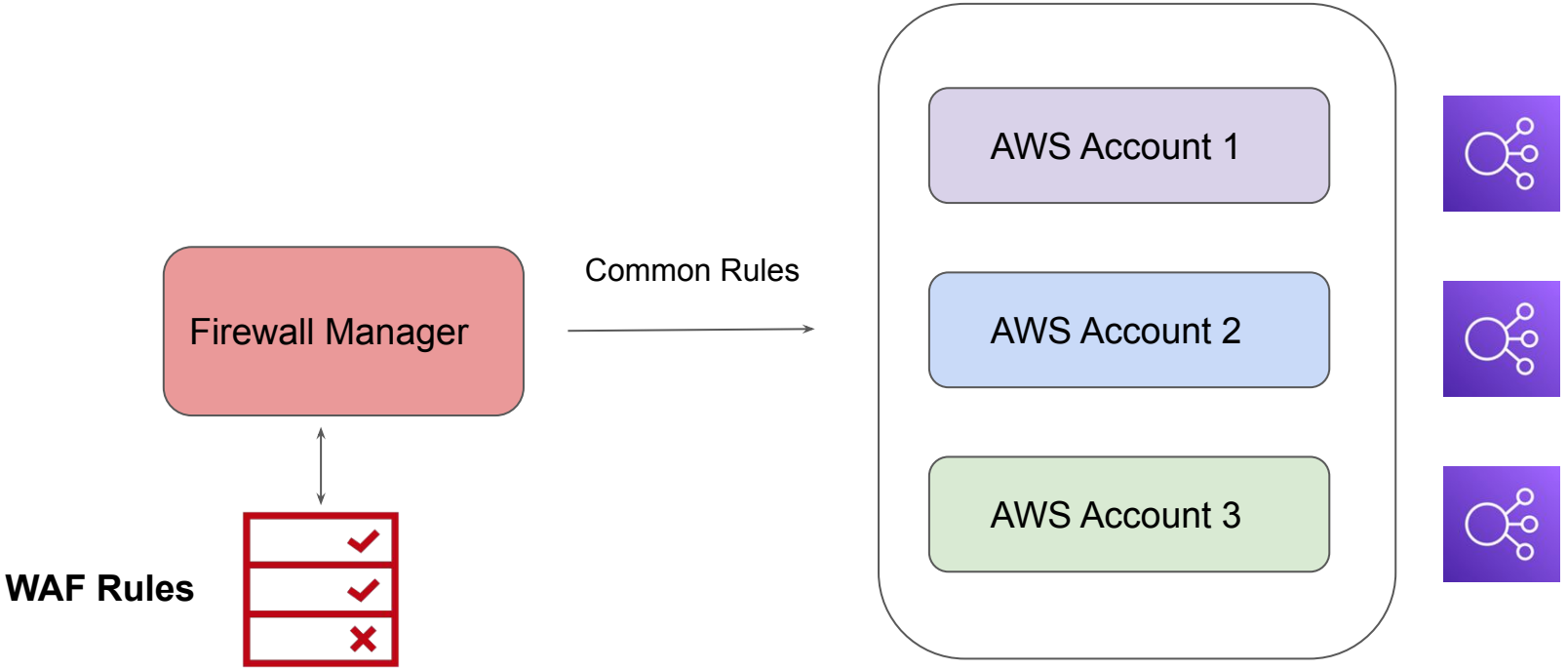
Most of the organizations are opting for Multi-Account based strategy for separation of environments (dev, stage, prod)

Security Team needs to create, maintain and update security services across all of the accounts.



Understanding the Basics

AWS Firewall Manager is a security management service which allows you to centrally configure and manage firewall rules across your accounts and applications in AWS Organizations



Supported Service

Firewall Manager supports wide variety of services, including:

- AWS WAF
- VPC Security Groups
- AWS Network Firewall
- Route53 DNS Firewall
- AWS Shield Advanced
- Palo Alto Cloud Next-generation firewalls

Important Prerequisite: AWS Organizations + AWS Config.

Benefits of Firewall Manager

1. Simplify management of firewall rules across your accounts

2. Ensure compliance of existing and new applications

3. Easily deploy managed rules across accounts

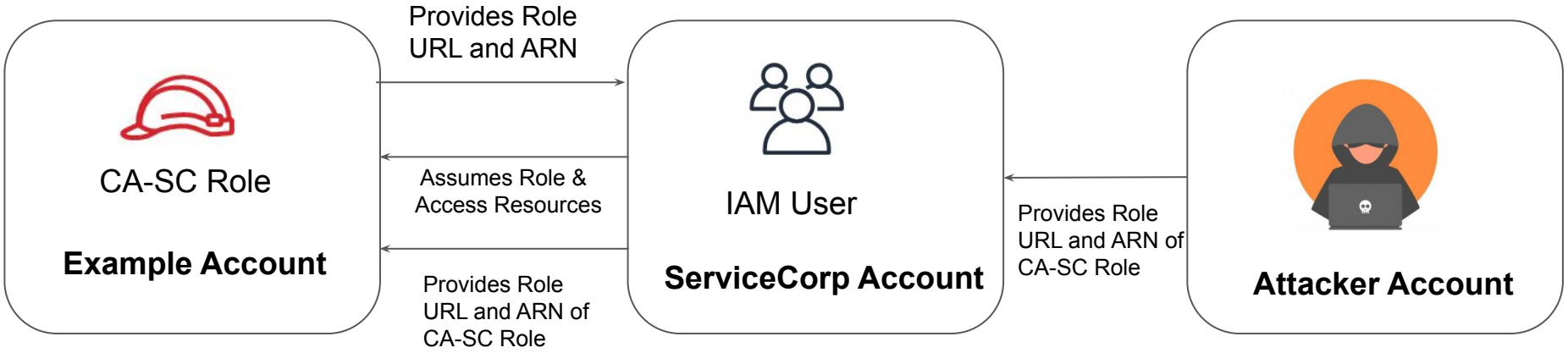
4. Centrally deploy protections for your VPCs

External ID

Secure Delegation

Understanding the Challenge

External ID is a piece of data that can be passed to the AssumeRole API of the Security Token Service (STS).



Important Points to Remember

Once the External ID is set, when someone uses the AWS CLI or AWS API to assume that role, they must provide the external ID.

You cannot switch roles at the AWS console to a role that has an external ID condition in its trust policy.

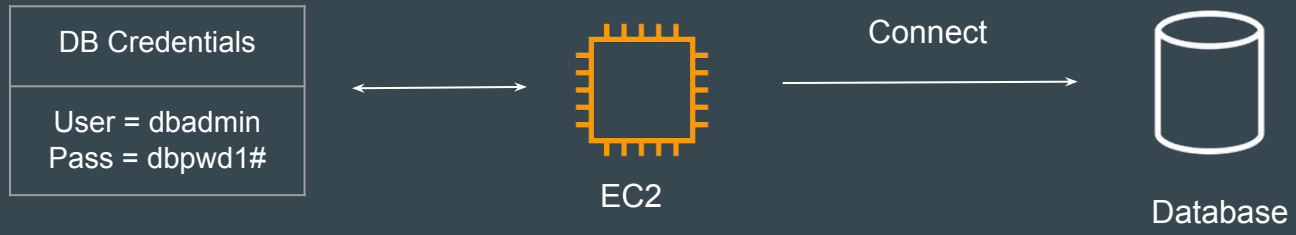
AWS Secrets Manager



Understanding the Challenge

In many organizations, secrets are hard coded directly as part of the application.

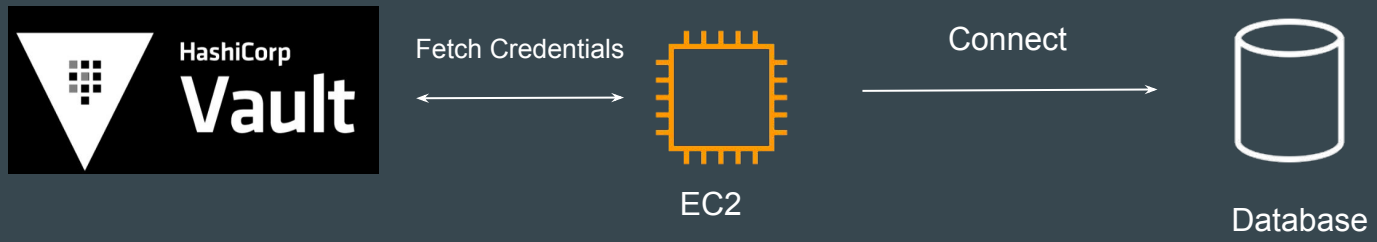
If you want to rotate the secret credential, all the application server needs to be updated. If you miss one, the production can go down.



Introducing Secrets Management

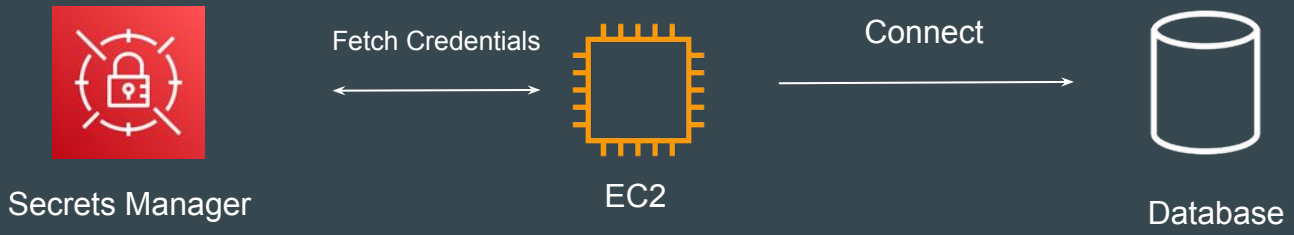
Secret management is a practice that allows developers to securely store sensitive data, such as passwords, keys, and tokens, in a secure environment with strict access controls.

Popular Tools: HashiCorp Vault, AWS Secrets Manager

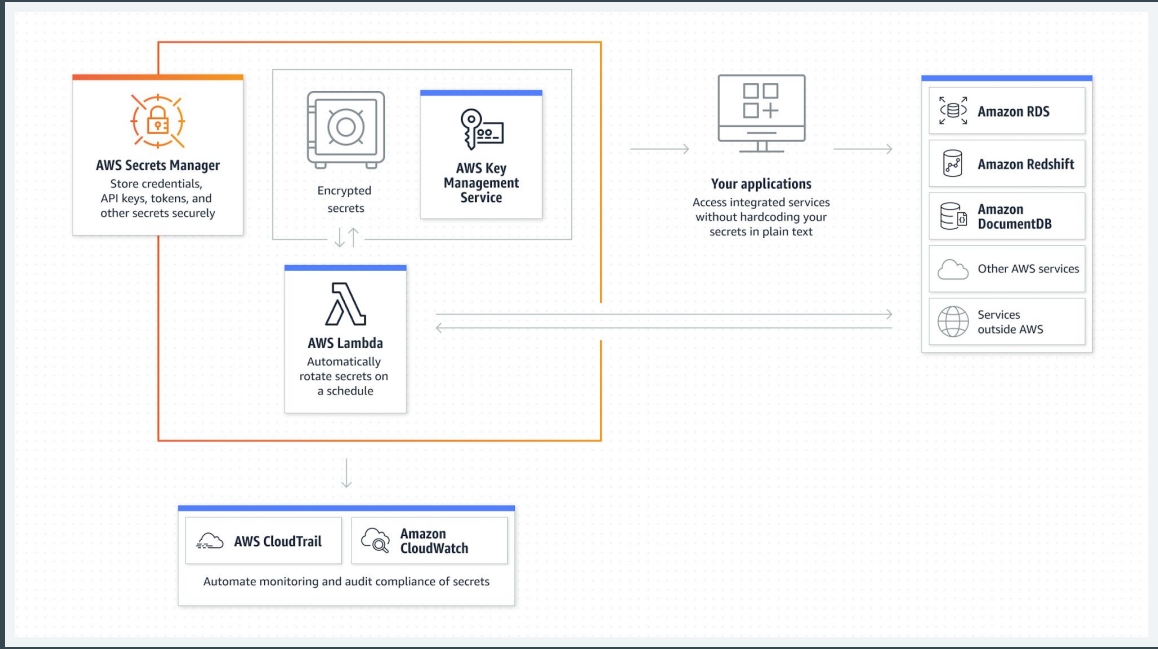


Introduction to Topic

AWS Secrets Manager helps you manage, retrieve, and rotate database credentials, API keys, and other secrets throughout their lifecycles.



Referenced from Docs



Rotate AWS Secrets Manager secrets

Rotation is the process of periodically updating a secret.

Secrets Manager rotation uses an AWS Lambda function to update the secret and the database.

To rotate a secret, Secrets Manager calls a Lambda function according to the schedule you set up. You can set a schedule to rotate after a period of time, for example, every 30 days.

Relax and Have a Meme Before Proceeding



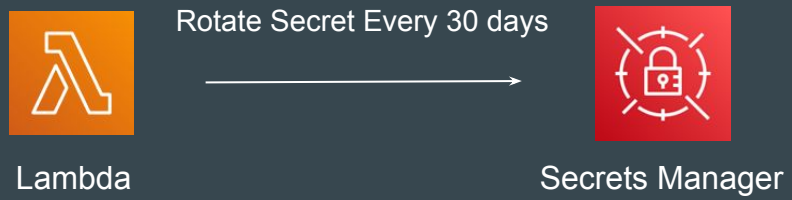
Rotating Secrets



Basics of Rotation

Rotation is the process of periodically updating a secret.

Secrets Manager rotation uses an AWS Lambda function to update the secret and the database.



Points to Note

To rotate a secret, Secrets Manager calls a Lambda function according to the schedule you set up. You can set a schedule to rotate after a period of time, for example, every 30 days.

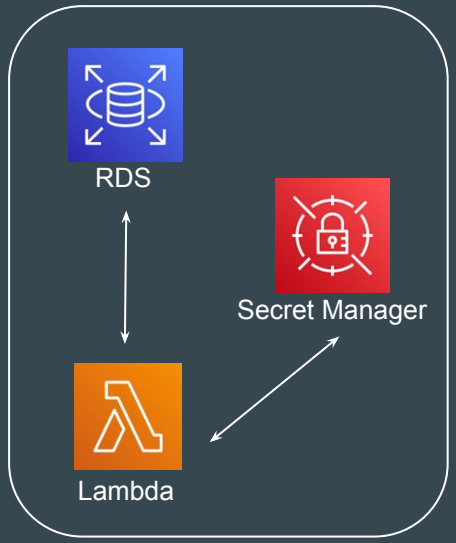
Secrets Manager provides rotation function templates for various use-cases related to RDS, DocumentDB, RedShift etc.

Replicate AWS Secrets Manager secrets



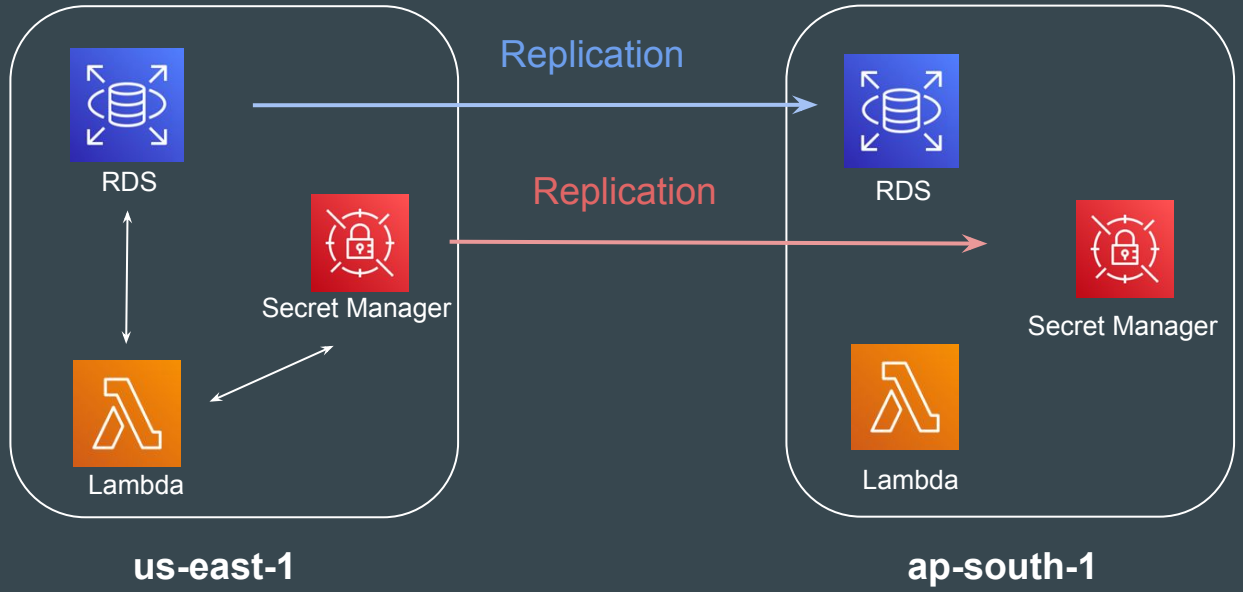
Understanding with Use-Case

In a Disaster Recovery based architecture, it is necessary to setup necessary level of replication across regions for failover.



Replicating Data Across Regions

In this architecture, the data and secrets are replicated across regions.



Points to Note

You can replicate your secrets in multiple AWS Regions to support applications spread across those Regions to meet Regional access and low latency requirements.

If you later need to, you can promote a replica secret to a standalone and then set it up for replication independently.

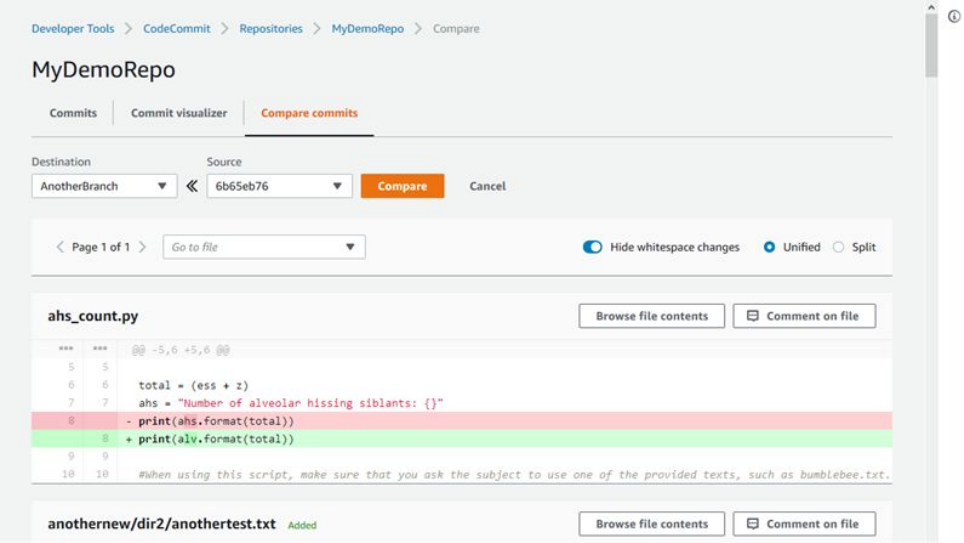
If you turn on rotation for your primary secret, Secrets Manager rotates the secret in the primary Region, and the new secret value propagates to all of the associated replica secrets.

AWS CodeCommit

Git Repository as Service

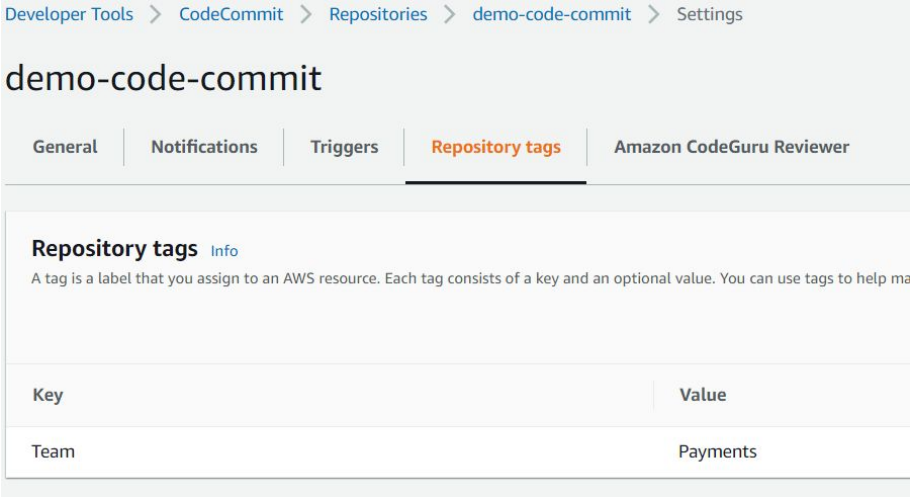
Overview of AWS CodeCommit

AWS CodeCommit is a managed source control service provided by AWS for hosting GIT repos.



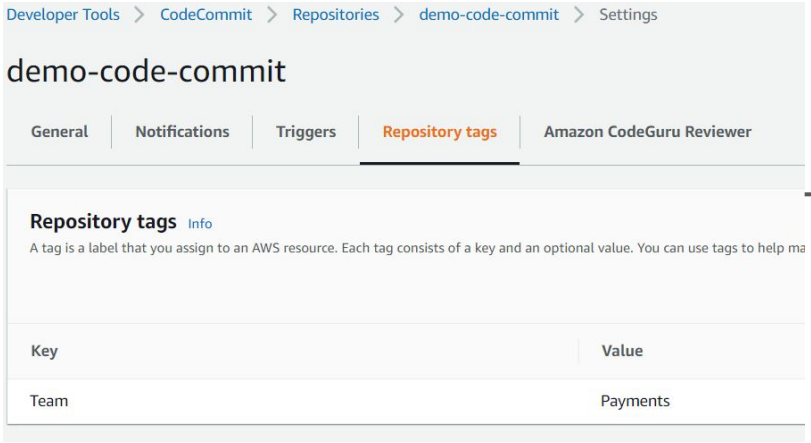
Repository Tags

Repositories can be tagged in AWS CodeCommit which further helps to identify and organize your AWS resources.



Identity Policies Based On Tags

You can create a policy that allows or denies actions on repositories based on the AWS tags associated with those repositories.




```
{
  "Version": "2012-10-17",
  "Statement" : [
    {
      "Effect" : "Allow",
      "Action" : "codecommit:*"
      "Resource" : "*",
      "Condition" : {
         "StringEquals" : "aws:ResourceTag/Team": "Payments"
        }
    }
  ]
}
```

Identity Policies For AWS CodeCommit

IAM Policy allows administrators to put a granular level of control over AWS CodeCommit service.

Sample Use-Cases:

- Deny all access to a repository with tag Status as Secret.
- Allow access to repository only in Mumbai region.
- Allow user only connecting from 10.77.2.50 to connect to repository.
- Deny Push actions to Master branch.

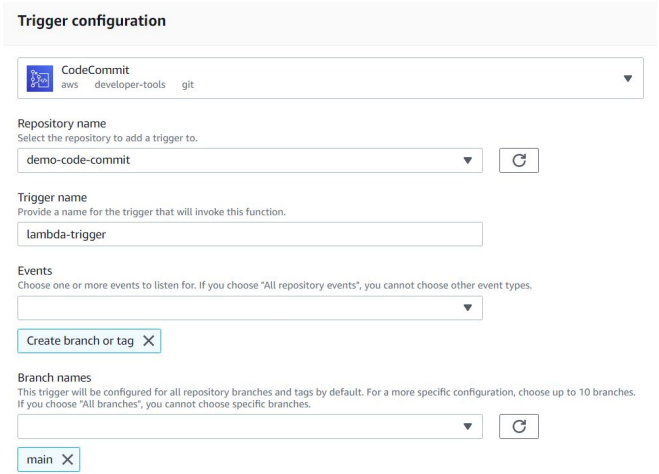
Integration with AWS Services

AWS CodeCommit integrates with various other services like Lambda, EventsBridge, CloudTrail, CodeBuild, AWS KMS and others.



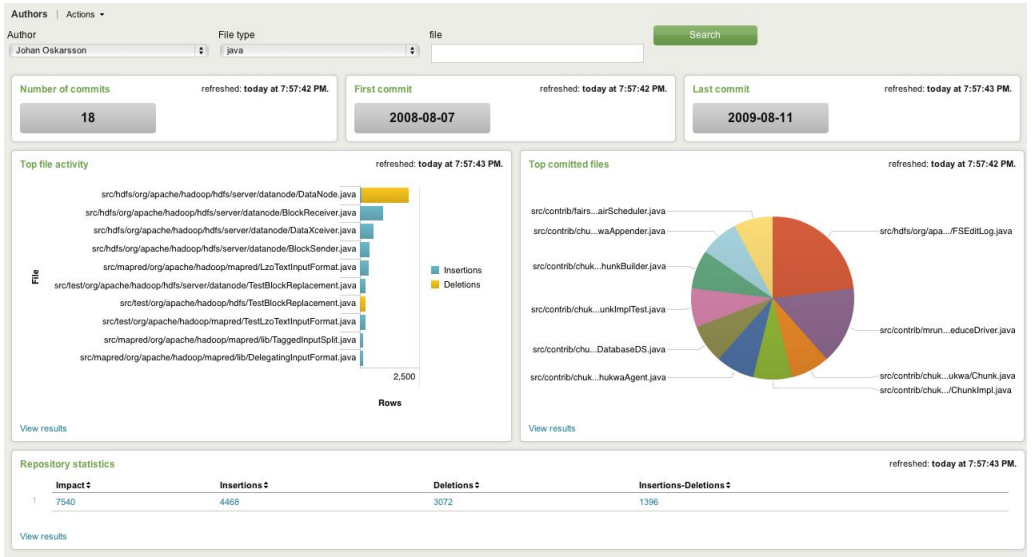
AWS CodeCommit

Trigger Lambda



Logging CodeCommit API Calls

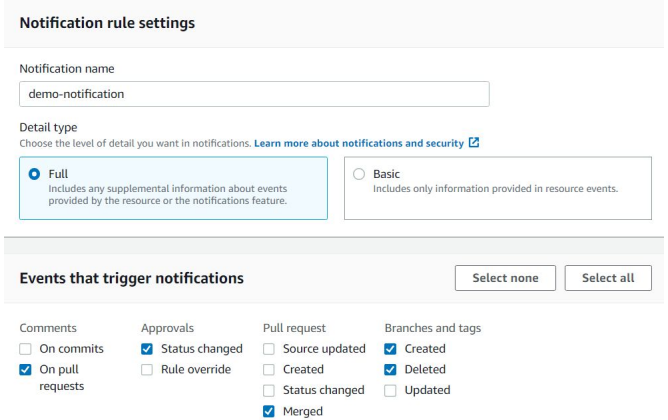
AWS CodeCommit is integrated with CloudTrail that will allow administrators to capture insights into activities by users.



Notification Rules

You can set up notification rules for a repository so that repository users receive emails about the repository event types you specify.

Notifications are sent when events match the notification rule settings



Data Protection

Data in CodeCommit repositories is encrypted in transit and at rest. When data is pushed into a CodeCommit repository (for example, by calling git push), CodeCommit encrypts the received data as it is stored in the repository.

When data is pulled from a CodeCommit repository (for example, by calling git pull), CodeCommit decrypts the data and then sends it to the caller.

Data sent or received is transmitted using the HTTPS or SSH encrypted network protocols.

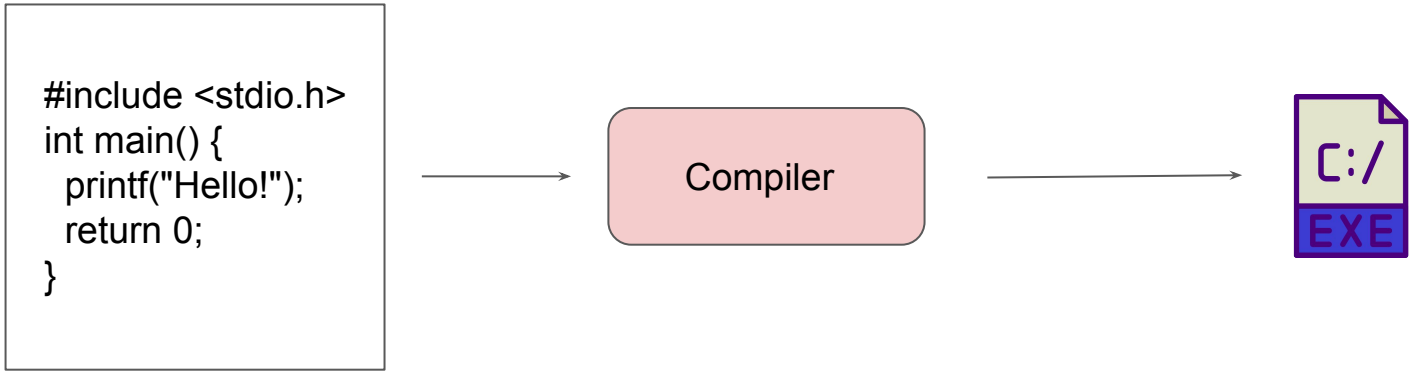
AWS Code Build

Let's Build Software

Software Build Process

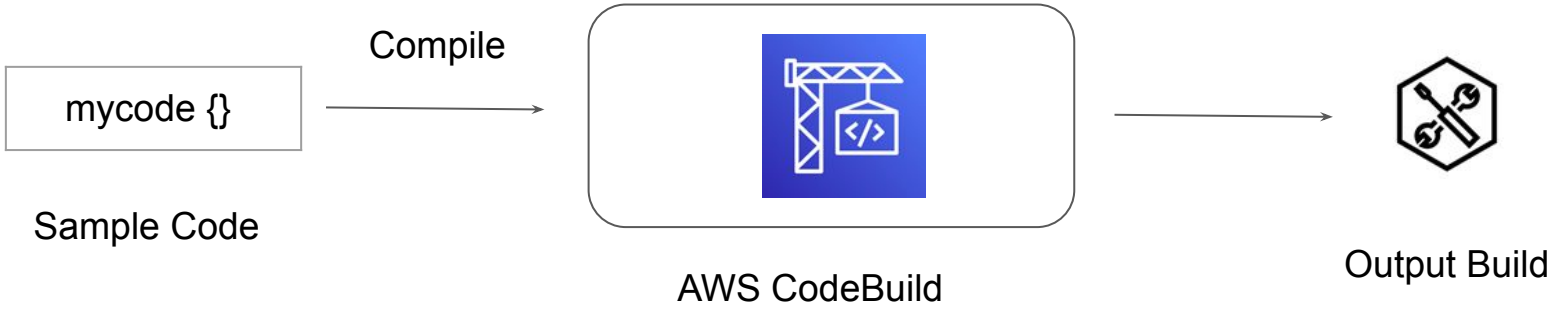
When we write code on various languages like Java, C and others, we need to compile to code.

The output of the compilation is the binary that is executable.



Overview of AWS CodeBuild

AWS CodeBuild is a fully managed continuous integration service that compiles source code, runs tests, and produces software packages that are ready to deploy.

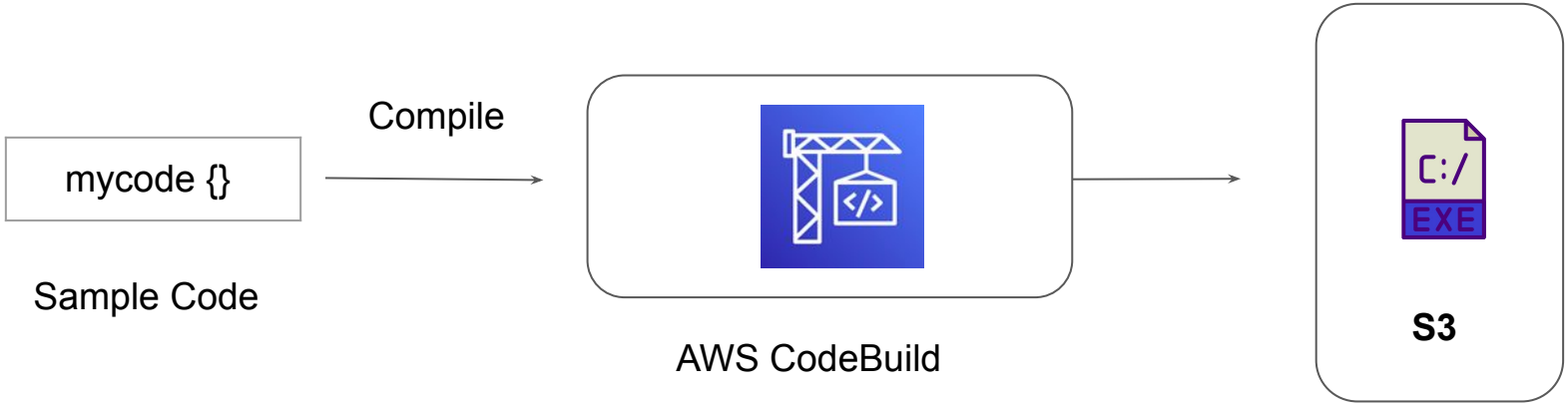


AWS CodeDeploy

Let's Deploy Software

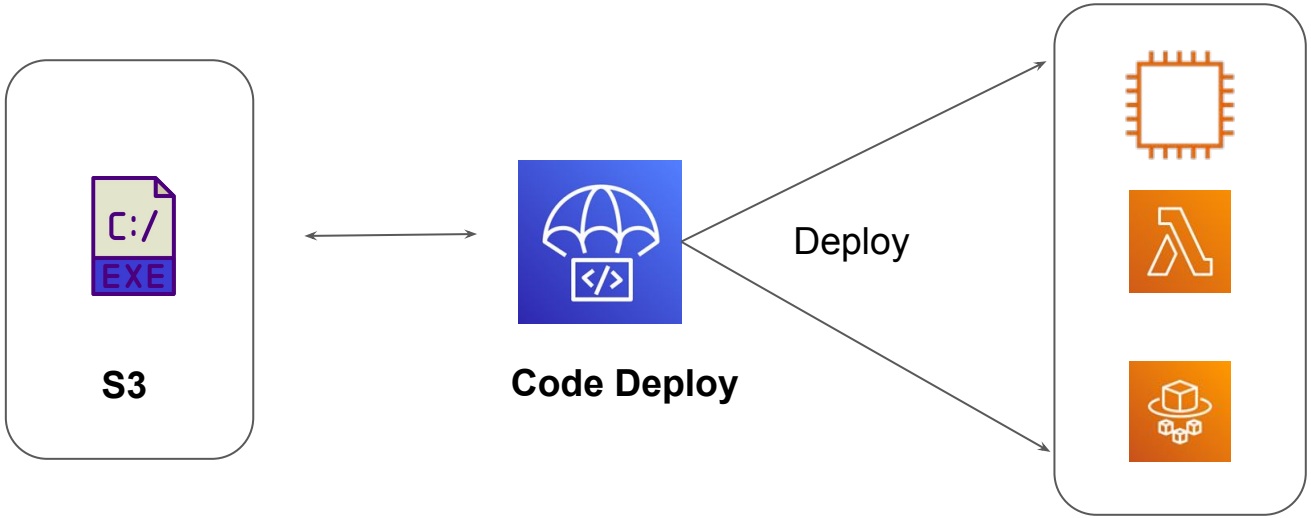
CodeBuild Workflow

AWS CodeBuild compiles the application and uploads the artifacts to S3 bucket.



Overview of CodeDeploy

AWS CodeDeploy is a managed deployment service that automates software deployments to a variety of compute services such as Amazon EC2, Fargate, Lambda and others.



Practical Steps

1. Create IAM Role for CodeDeploy with S3ReadOnly Access.

2. Create IAM Role for EC2 with S3ReadOnlyAccess

3. Launch EC2 Instance with Appropriate Role.

4. Install CodeDeploy Agent in EC2

5. Configure CodeDeploy Service.

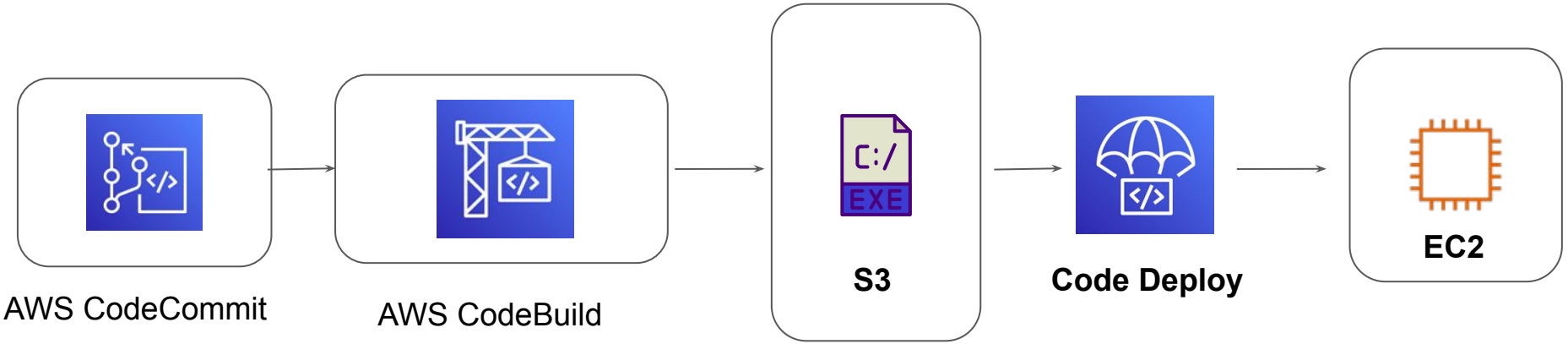
Code Pipeline

Automating Deployments

Current Setup

At this stage, we have the pipeline setup using Code Commit, Code Build and CodeDeploy

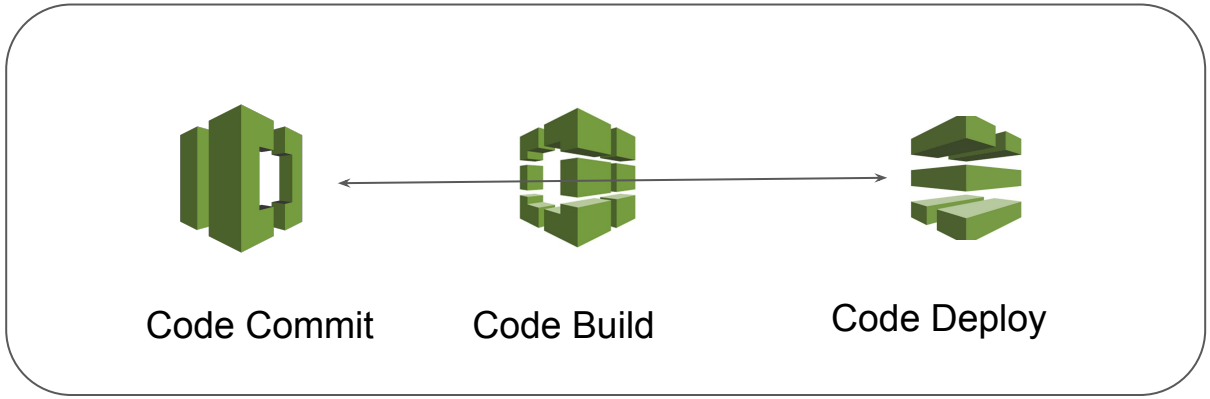
Challenge: The Entire Process is Manual.



Overview of Code Pipeline

AWS Codepipeline is a continuous delivery service to automate steps required to release the software.

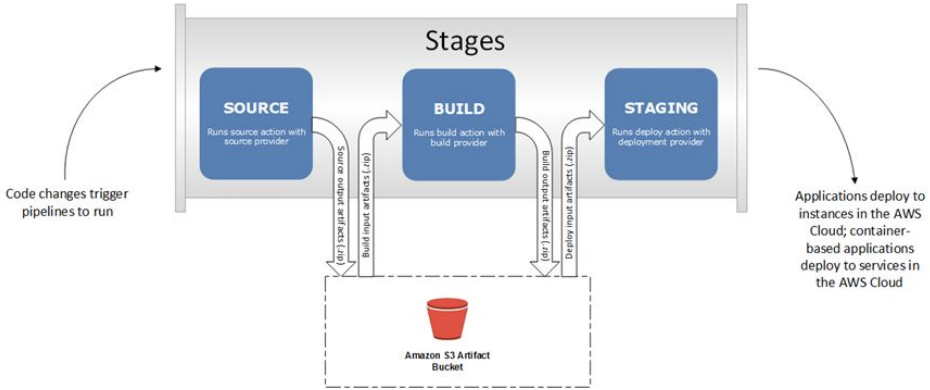
It allows us to launch the entire continuous delivery toolchain in minutes, allowing releasing code faster.



Important Pointer - Code Pipeline

Codepipeline automatically triggers your pipeline whenever there is a commit in the source repository.

- Output artifact is ingested into input artifact to the Build stage.
- Output artifact from build stage (build) acts as input to the deploy stage.



Launch Templates

Launching EC2 The Easy Way

Understanding the Challenge

When you launch an EC2 instance, there are various configurations that needs to be set.

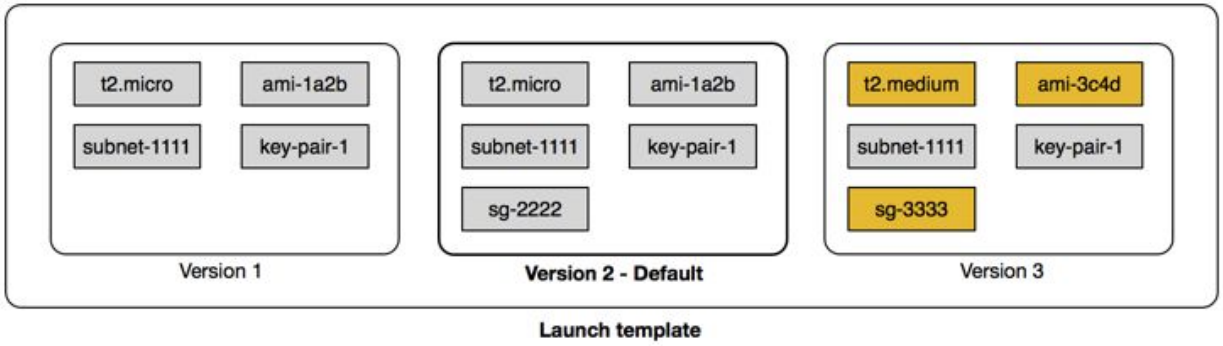
Some of the common configuration includes:

- AMI ID
- Instance Type
- Security Group
- Key Pair
- Storage
- IAM Role
- VPC

Everytime when you intend to launch instance, going through process is time consuming,

Introduction to Launch Templates

Launch templates enable you to store launch parameters so that you do not have to specify them every time you launch an instance.

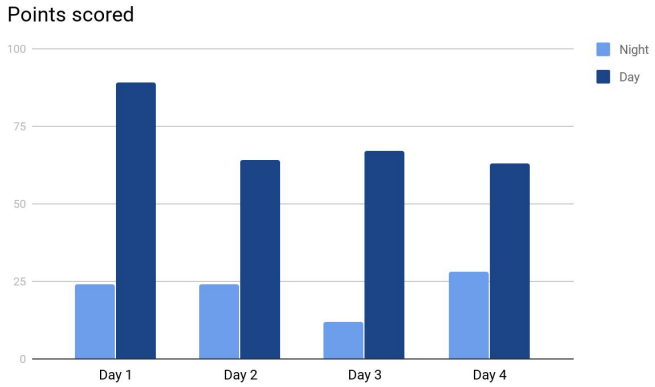


EC2 Auto Scaling

Up and Down, Round and Round

Understanding Scalability

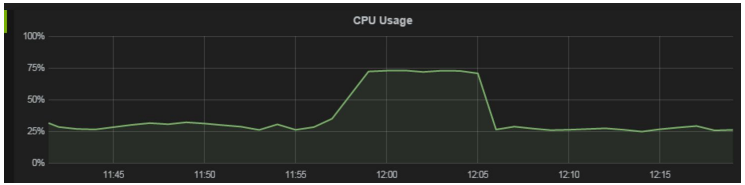
- Scalability is the ability of a system to change in size depending on the needs.
- Infrastructure should scale to support changing in traffic patterns.



Launch and Remove Servers Based on Load

What if new servers automatically get launched on high load?

Simple Scaling Policy:



Base: 2 servers

Scalable :

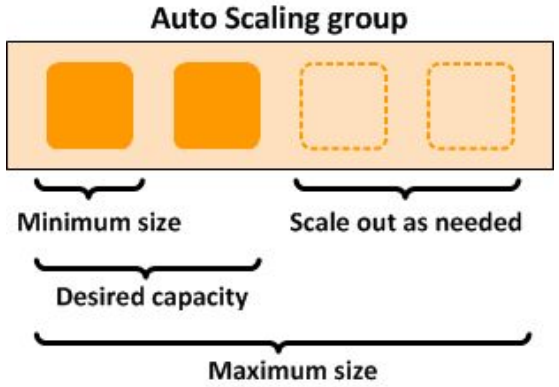
If average CPU utilization > 60% ; add two more instance
If average CPU utilization < 30% ; remove two instance

Overview of EC2 Auto-Scaling

Amazon EC2 Auto Scaling helps you maintain application availability and allows you to automatically add or remove EC2 instances according to conditions you define.

Example Scenario:

- Minimum: 2 EC2 instance
- Maximum: 10 EC2 instance
- Threshold: 50% of CPU



Multiple Types of Scaling

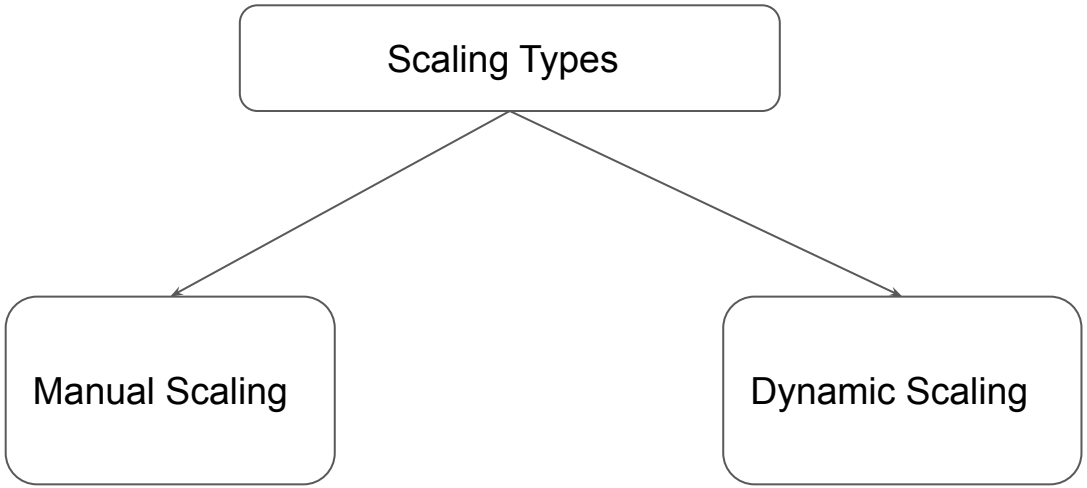
Type of Scaling	Description
Scheduled Scaling	Servers are scaled based on a specific schedule. For example, every week the traffic to your web application starts to increase on Wednesday, remains high on Thursday, and starts to decrease on Friday
Dynamic Scaling	Follow the demand curve for scaling activities. CPU Utilization higher then 90%
Predictive Scaling	Predictive Scaling has machine learning algorithms that detect changes in daily and weekly patterns, automatically adjusting their forecasts.

Dynamic Scaling

Overview

2 Types of Scaling

There are two primary types of scaling approaches that are available:



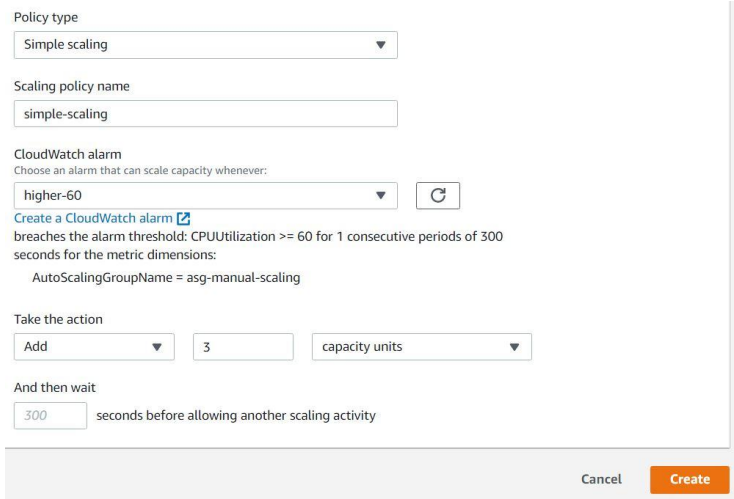
Dynamic Scaling

When you configure dynamic scaling, you define how to scale the capacity of your Auto Scaling group in response to changing demand.

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

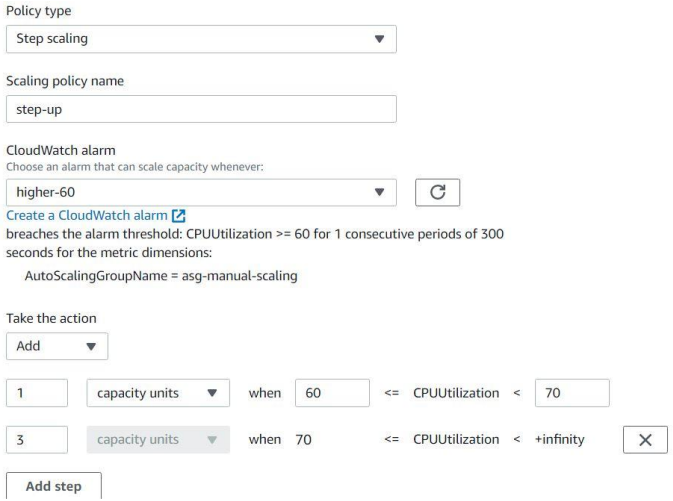
Simple Scaling Policy

With simple scaling policy, you can configure a specific number of instances to be added when a threshold reaches certain value.



Step Scaling Policy

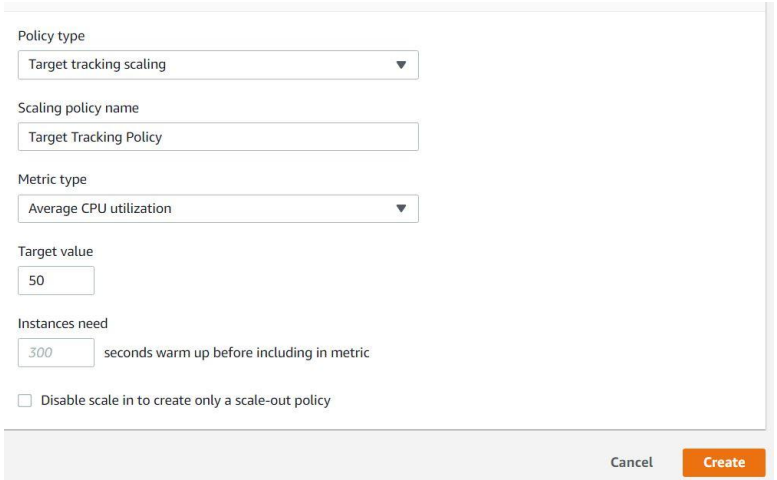
In step scaling, the adjustment of the current capacity of instances vary based on the size of the alarm breach.



Target Tracking Policy

With target tracking scaling policies, you select a scaling metric and set a target value.

The scaling policy adds or removes capacity as required to keep the metric at, or close to, the specified target value.



Use Case of Thermostat

A thermostat is a component which senses the temperature of a physical system and performs actions so that the system's temperature is maintained near a desired setpoint.

Example:

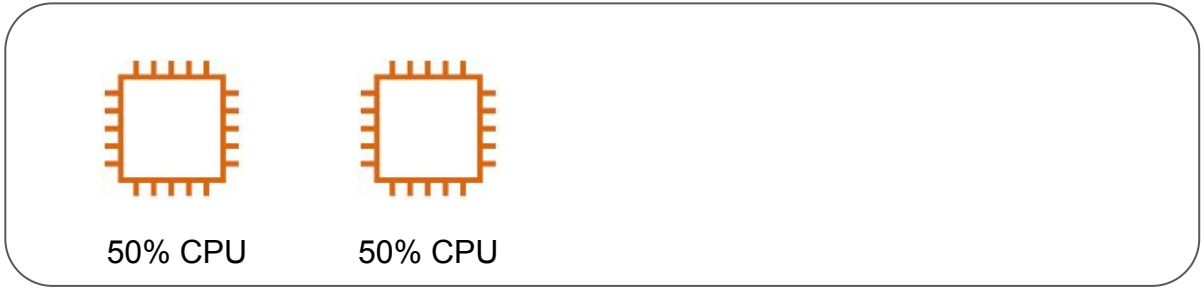
- Desired = 24
- Current = 18



Example Target Tracking Policy

Metric Type = CPU Utilization

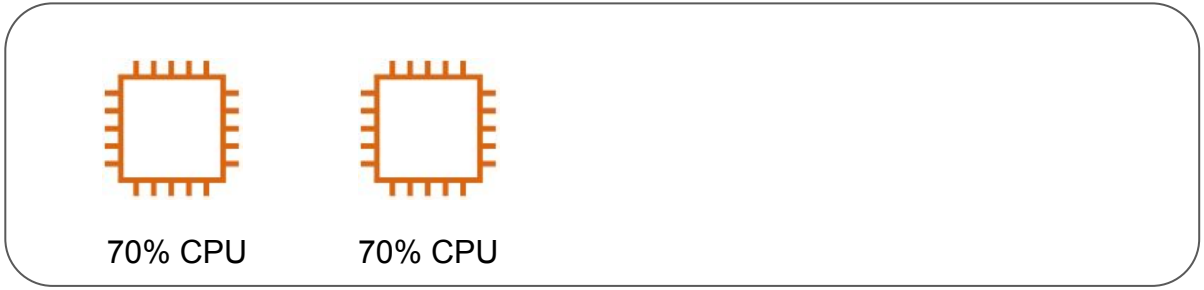
Target Value = 50%



Example Target Tracking Policy

Metric Type = CPU Utilization

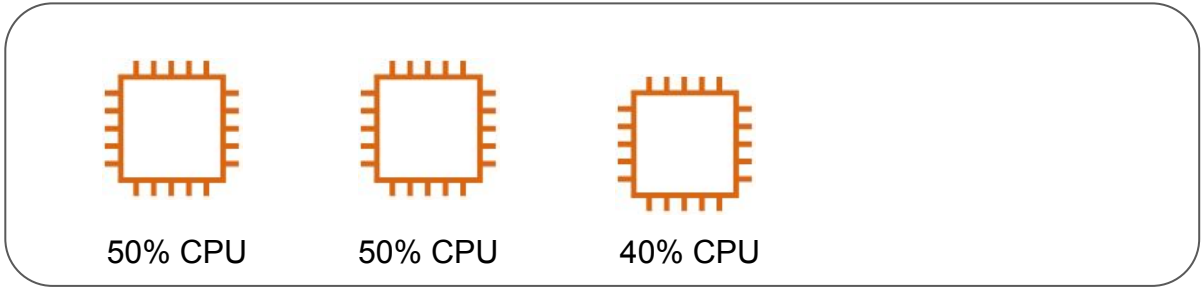
Target Value = 50%
Actual Value = 70%



Example Target Tracking Policy

Metric Type = CPU Utilization

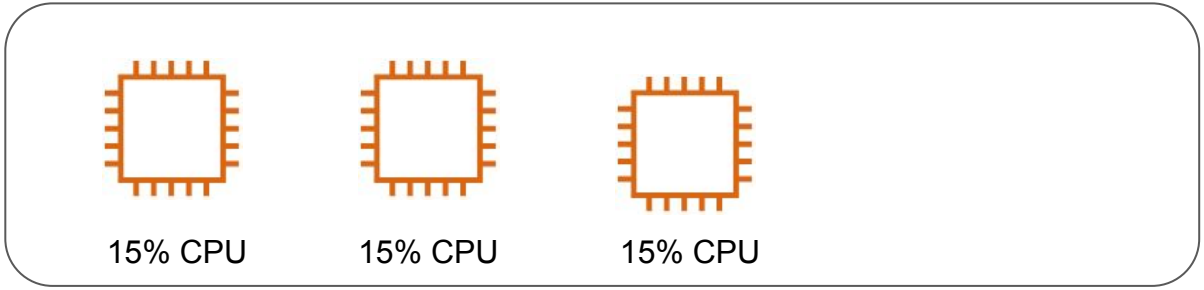
Target Value = 50%



Example Target Tracking Policy

Metric Type = CPU Utilization

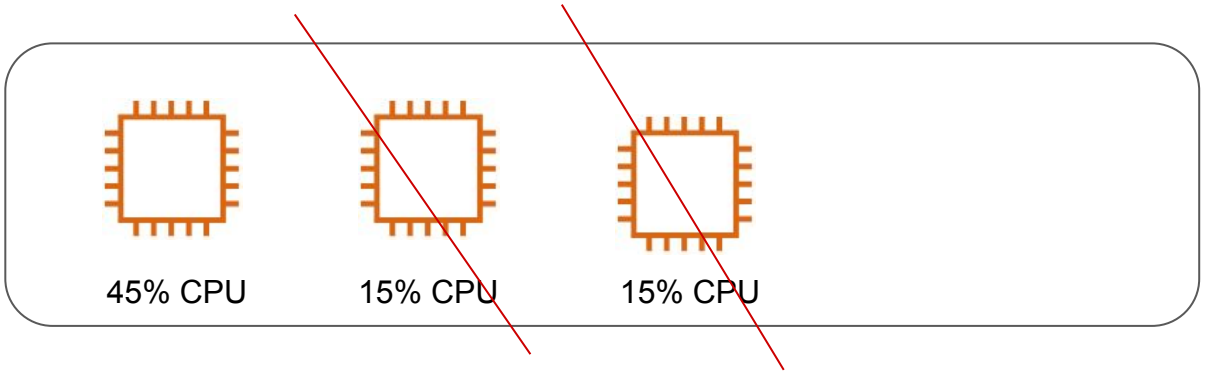
Target Value = 50%
Actual Value = 15%



Example Target Tracking Policy

Metric Type = CPU Utilization

Target Value = 50%
Actual Value = 45% (average)



Scheduled Scaling

Overview and Practical

Overview of Scheduled Scaling

Scheduled scaling allows you to set your own scaling schedule.

For example, let's say that every week the traffic to your web application starts to increase on Wednesday, remains high on Thursday, and starts to decrease on Friday.

Scaling actions are performed automatically as a function of time and date.

Relax and Have a Meme Before Proceeding



Auto-Scaling LifeCycle Hooks

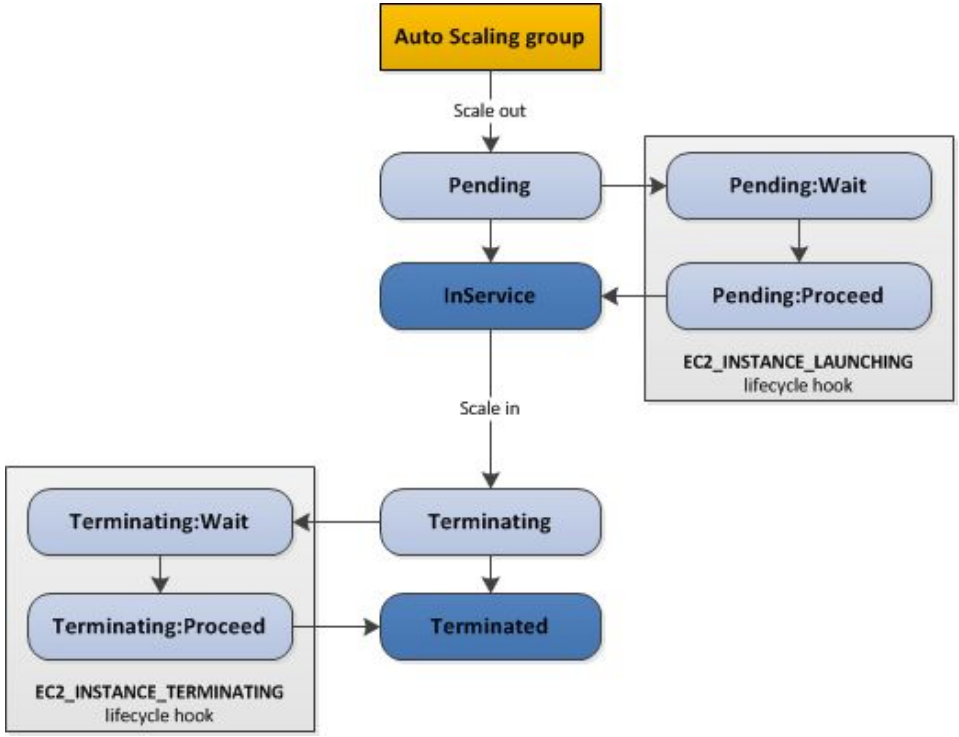
Challenges and Structure

Overview of LifeCycle Hooks

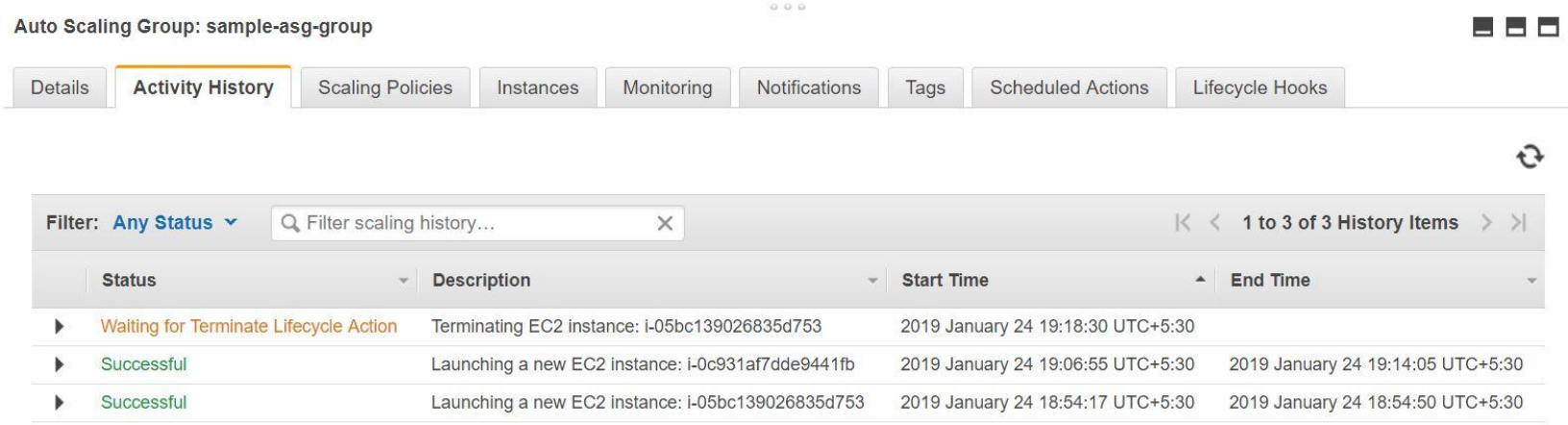
Auto-Scaling Lifecycle hooks allows us to have control over instance launch and termination state within auto-scaling group.

Sample Use-Case:

- You have EC2 instance which is scheduled to be terminated.
- You want to backup all it's logs to S3 and run some deregistration scripts.
- Terminate instance once 2nd steps is completed.


Auto Scaling group
Scale out
Pending
Pending:Wait
InService
Pending:Proceed
EC2_INSTANCE_LAUNCHING
lifecycle hook
Scale in
Terminating:Wait
Terminating
Terminating:Proceed
Terminated
EC2_INSTANCE_TERMINATING
lifecycle hook

Step 1: Instance Awaiting Termination

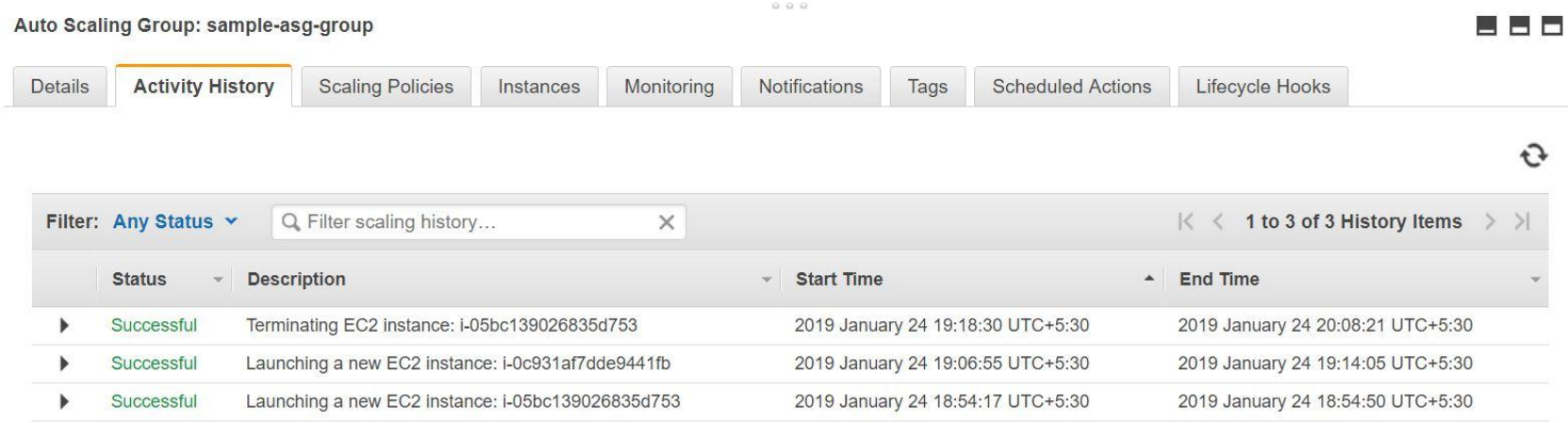


Step 2: Confirmation from Automation

```
[root@ip-172-31-34-10 ~]# aws autoscaling complete-lifecycle-action --lifecycle-hook-name SampleTer
minateHook --auto-scaling-group-name sample-asg-group --lifecycle-action-result CONTINUE --instance
-id i-05bc139026835d753 --region us-east-1
[root@ip-172-31-34-10 ~]#
Broadcast message from root@ip-172-31-34-10
        (unknown) at 14:36 ...

The system is going down for power off NOW!
Connection to 54.160.201.201 closed by remote host.
Connection to 54.160.201.201 closed.
```

Step 3: Go Ahead!



OpsWorks

Automation

Introduction

AWS OpsWorks is a configuration management service that provides managed instances of Chef and Puppet.

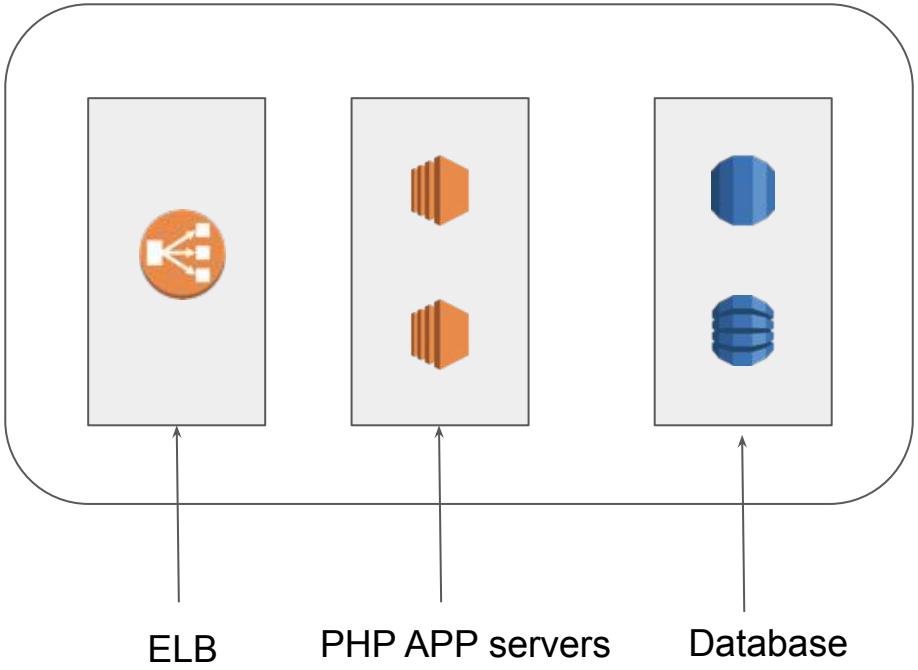
Integration of EC2 with the configuration management tool brings up great possibilities on how servers are configured, deployed and managed.

Example Use Case:

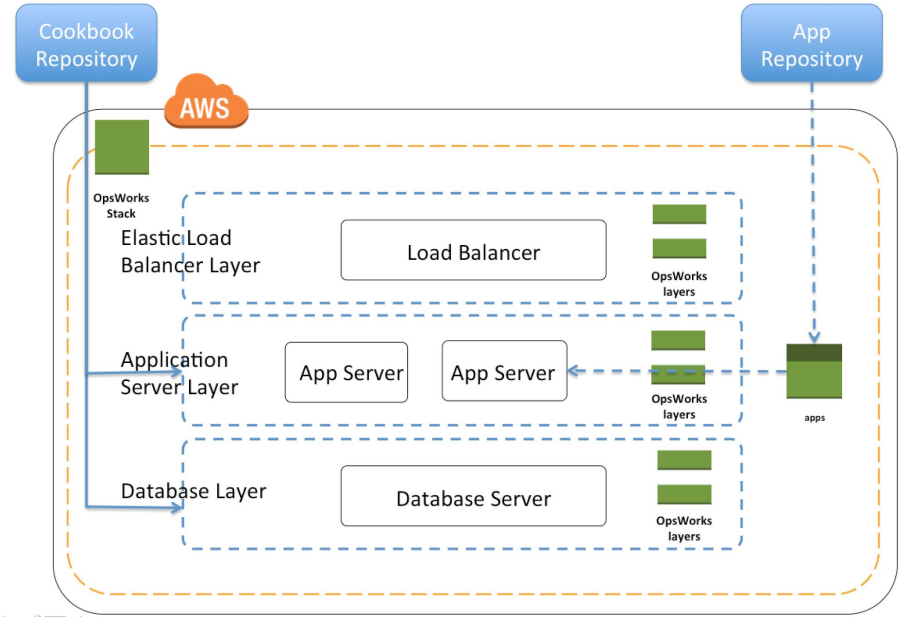
While an EC2 instance is getting launched, we want to install certain packages like Nginx, PHP-FPM and MySQL along with having custom SSH configuration file followed by restart of SSH server.

OpsWorks Concepts

OpsWorks Stack



Stacks & Layers



OpsWorks - Lifecycle Events

Deployments, yet again!

Getting Started

OpsWorks has “five” events that occur during the event lifecycle.

These events are :

1. Set up
2. Configure
3. Deploy
4. Undeploy
5. Shutdown

When an event occurs, it runs a set of chef recipes assigned to that event.

Getting Started

i) Setup Event:

- Event occurs when the started instance has finished booting up.
- Used for initial installation of software packages.

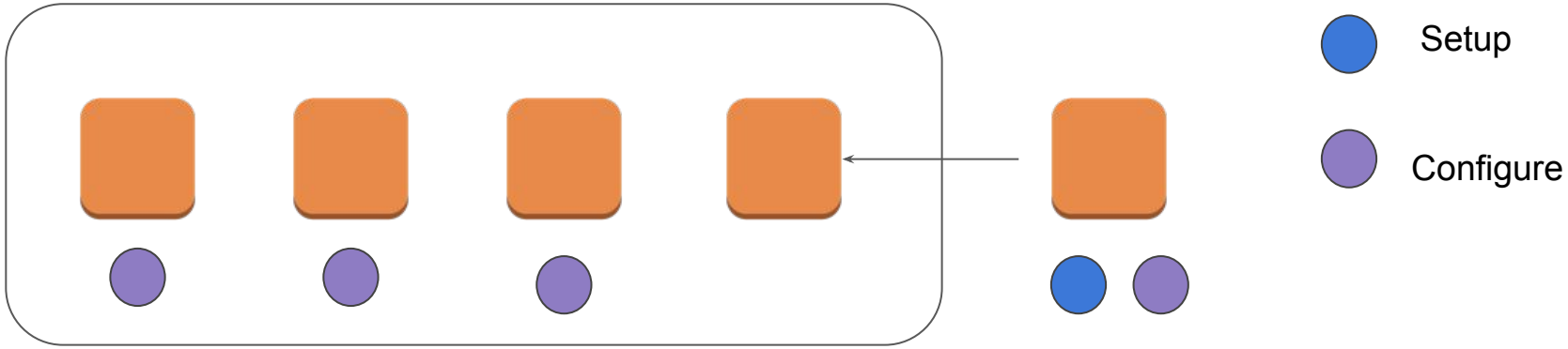
Example:

Installing PHP and Apache through layer recipes.

Configure Event

ii) Configure [Event occurs on all the stack's instances when]

- Events in configure are executed whenever instance enters or leaves online state.
- You associate or disassociate EIP of the instance.
- You attach or detach an ELB from the layer.



Deploy and Undeploy events

iii) Deploy

Deploy event allows us to manually define when we want to deploy a new version of app.

iv) Undeploy:

This event occurs when we delete the app or run the undeploy command to remove the app from the set of application servers.

ShutDown events

v) ShutDown

The event in this stage are executed when we inform OpsWorks to shut the instance down before the EC2 instances are terminated.

The default shutdown timeout if 120 seconds.

AWS Config

Overview of Infrastructure Changes

AWS Config - High Level Overview

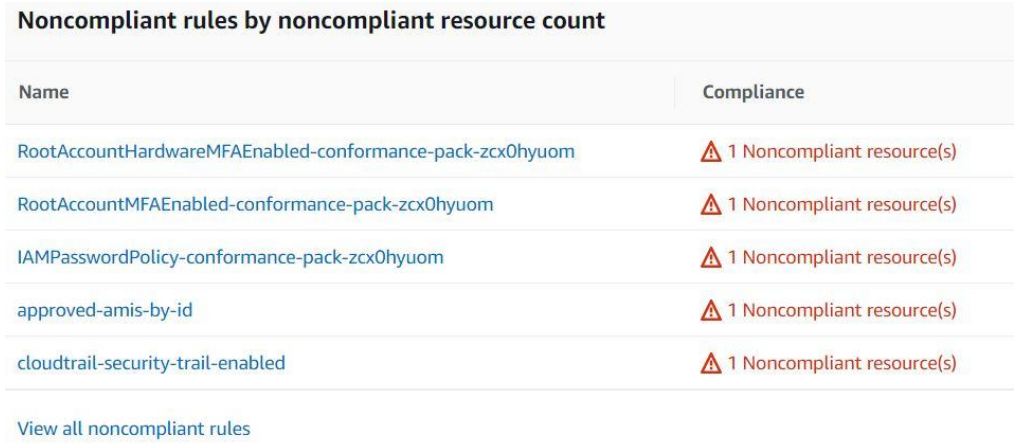
AWS Config is primarily used to record the resource configuration changes over time.

An EC2 instance was hosting website from past 90 days. Suddenly in last one week, there have been a lot of issues with the requests. What was changed?



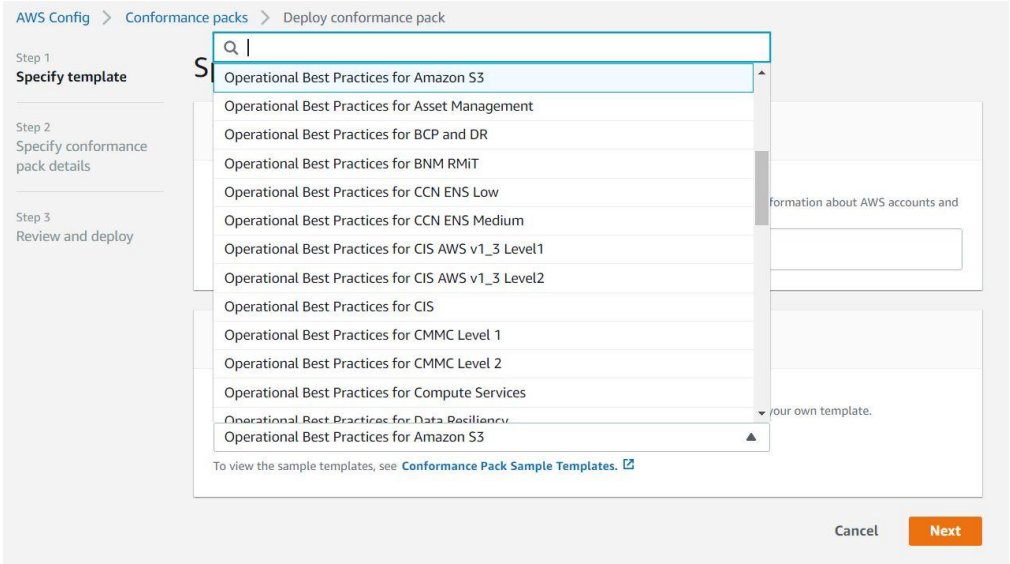
Audit and Compliance

AWS Config comes with large set of rules that can continuously monitor your AWS environment and report the findings.



Conformance Packs

A conformance pack is a collection of AWS Config rules and remediation actions that can be easily deployed



Pricing of AWS Config

You pay $0.003 per configuration item recorded in your AWS account per AWS Region. A configuration item is recorded whenever a resource undergoes a configuration change or a relationship change.

Based on rule evaluation. A rule evaluation is recorded every time a resource is evaluated for compliance against an AWS Config rule.

You are charged per conformance pack evaluation in your AWS account per AWS Region based on the tier below.

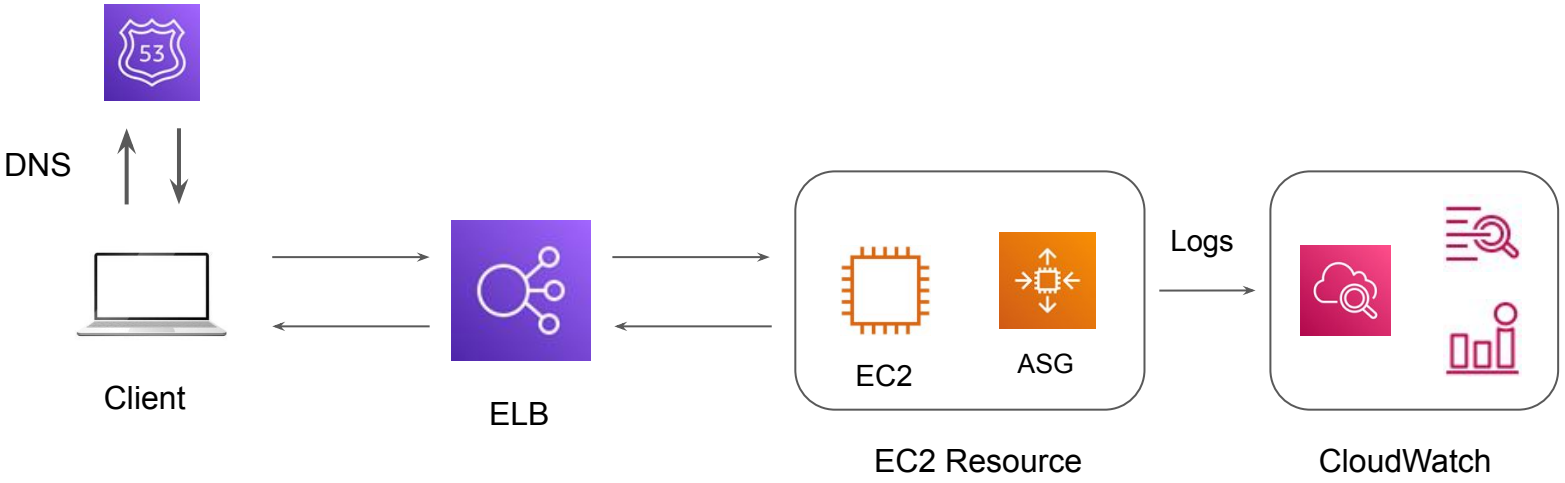
Elastic Beanstalk

Orchestration

Traditional Deployment Approach

Use-Case: Deploy a simple Hello World application for production.

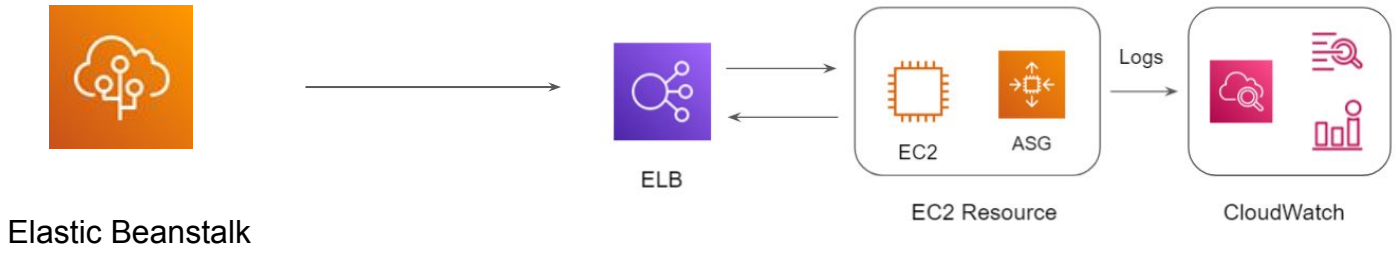
Resources to be created: AWS EC2, ELB, Auto-Scaling, Web-Server Configuration, and others.



Elastic Beanstalk Deployment Approach

Use-Case: Deploy a simple Hello World application for production.

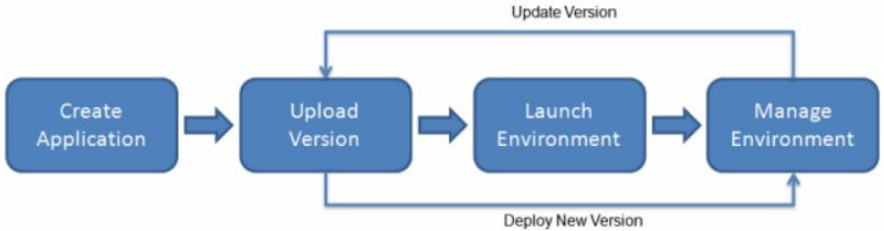
Create Elastic Beanstalk Environment



Overview of Elastic Beanstalk

AWS Elastic Beanstalk is an easy-to-use service for deploying and scaling web applications and services.

You can simply upload your code and Elastic Beanstalk automatically handles the deployment, from capacity provisioning, load balancing, auto-scaling to application health monitoring.



EB Deployment Policy

Deploying Application Updates

Deployment Policy Options

Deployment Policy specifies how new updates for the applications are pushed into the EB environment.

Elastic Beanstalk supports several options for deployments

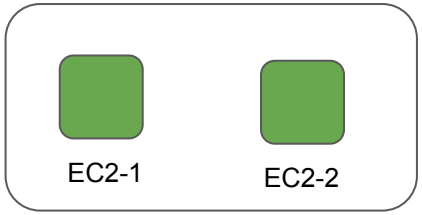
- All at Once
- Rolling
- Rolling with Additional Batch
- Immutable
- Traffic Splitting
- Blue/Green

All at Once

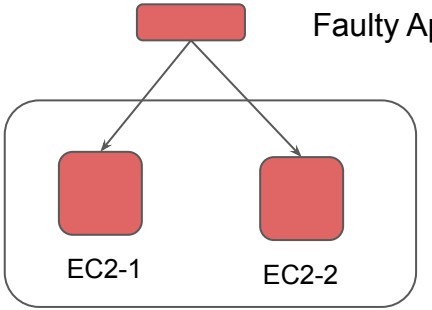
Deploys new version to all the instances simultaneously.

Applications might be unavailable for the users for short-period of time.

If updates fail then you will need to roll back changes by re-deploying the previous working version.



Working Environment

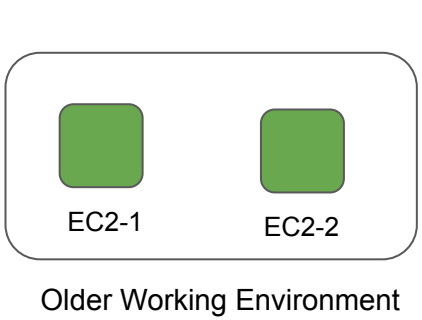


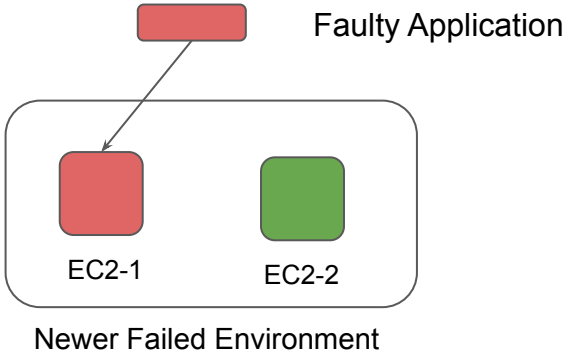
Newer Failed Environment

Rolling Deployment Policy

Your application is deployed to your environment one batch of instances at a time.

Each batch of instances is taken out of service while deployment is taking place.





Important Pointer - Rolling Deployment Policy

The overall capacity (in terms of servers) will be reduced while the deployment is happening.

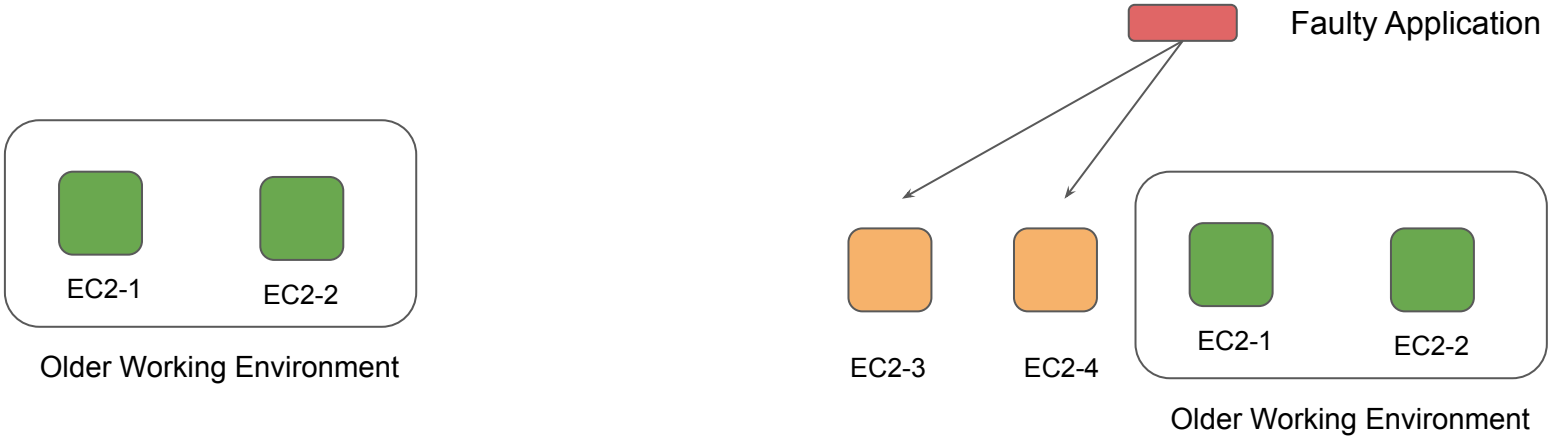
Not recommended for performance-critical applications.

The overall deployment time is much longer then All at Once approach.

Rolling with additional batch

With this method, Elastic Beanstalk launches an extra batch of instances, then performs a rolling deployment.

This maintains full capacity during deployments.



Important Pointer - Rolling with Additional Batch

The deployment is first made to the additional batch instances.

If the deployment fails, these additional batch instances are terminated.

If the deployment succeeds, these additional batch instances are registered under load balancer and the older instances are terminated.

Immutable Deployment Policy

Deploys newer version of the application in a completely new servers under new auto-scaling group.

When new instances passes their health-check, they are moved to older auto-scaling group and older ones are terminated.

Impact of failed update is less as all we need to do is delete the new auto-scaling group and EC2.

Preferred option for mission critical production systems.

Blue Green Deployments

Deployments, yet again!

Getting Started

Blue environment is an existing environment in production receiving live traffic.

Green environment is parallel environment running different version of the application.

Deployment = Routing production traffic from blue to green environment.



Various ways to achieve Blue Green

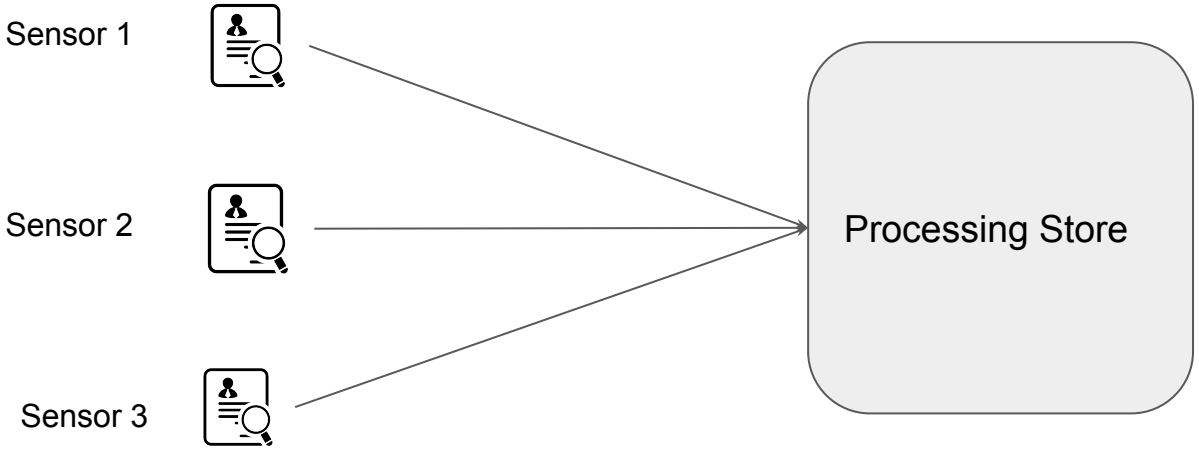
1. Updating DNS Routing via Route53
2. Swap the Auto-Scaling Groups behind your ELB
3. Swap Launch Configurations
4. Swap Beanstalk Environments (applies only to EB environments)
5. Clone Stack with AWS OpsWorks and updating DNS

Amazon Kinesis

Streaming Data

Basics of Streaming Data.

Streaming data is the continuous flow of data generated by various sources



Examples of Streaming Data

A financial institution tracks changes in the stock market in real time and adjust it's portfolio accordingly.

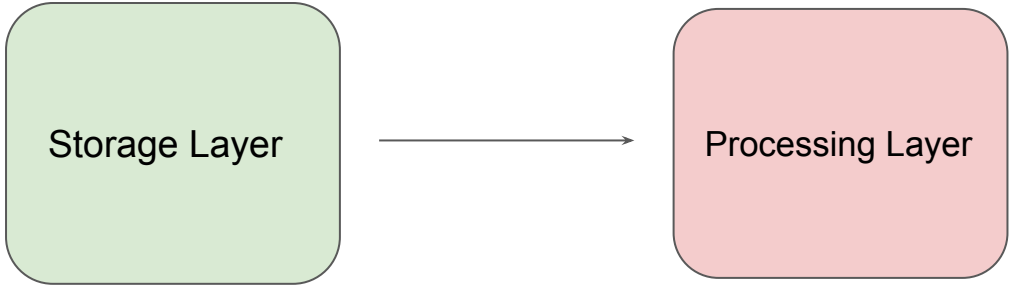
A media publisher streams billions of clickstream records from its online properties



Challenges with Working of Streaming Data

Streaming data processing requires two layers: a storage layer and a processing layer.

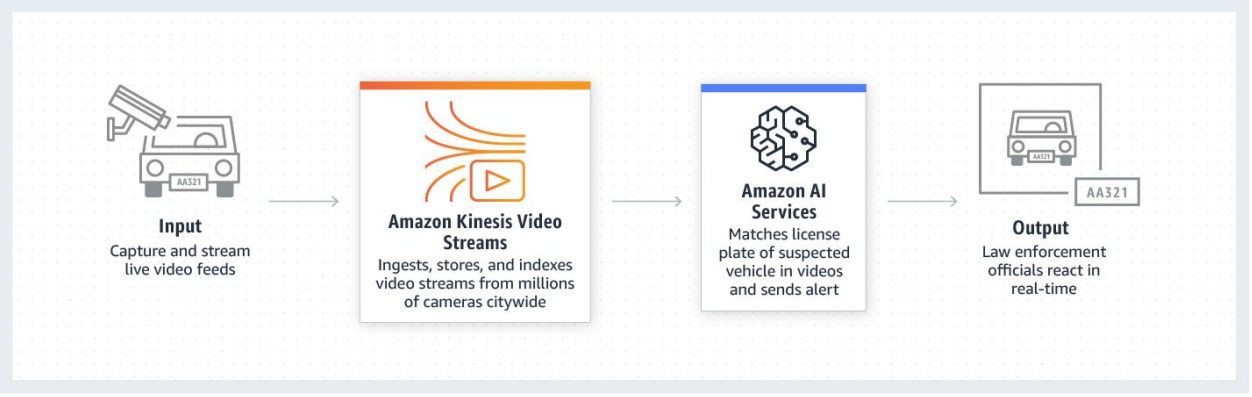
The storage layer needs to support record ordering and strong consistency, replayable reads and the processing layer is responsible for consuming data from the storage layer, running computation on that data and many other tasks.



Basics of Amazon Kinesis

Amazon Kinesis makes it easy to collect, process, and analyze real-time, streaming data so you can get timely insights and react quickly to new information.

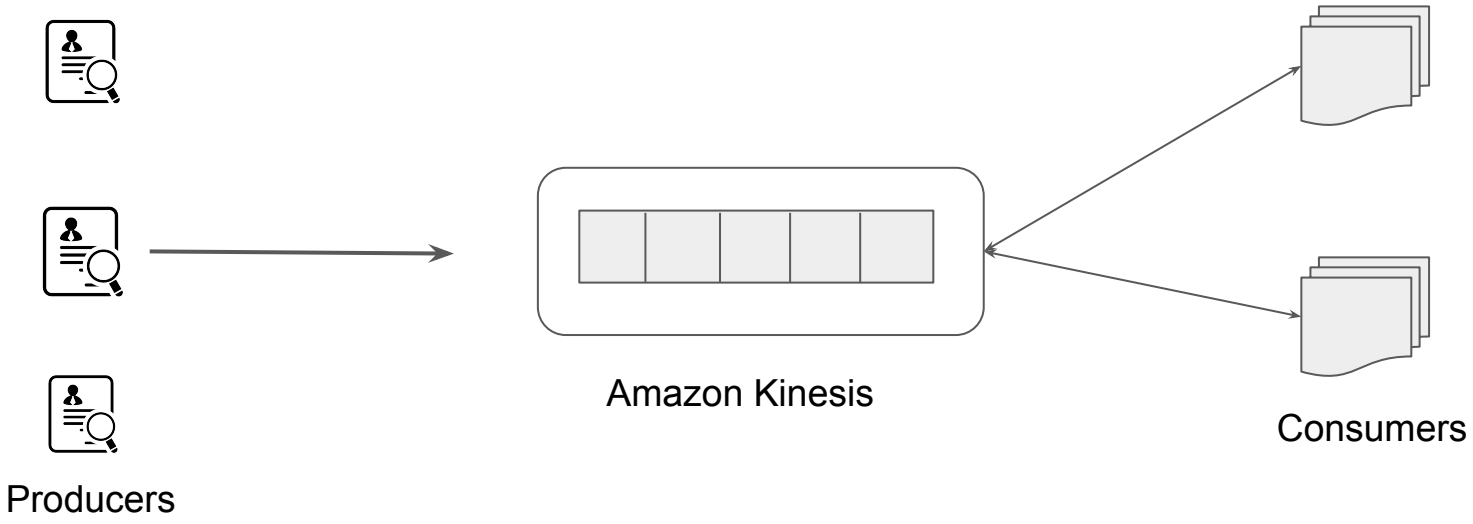
Amazon Kinesis offers key capabilities to cost-effectively process streaming data at any scale



3 entities

There are 3 entities in this kind of use case:

Producer, Stream Store, Consumer



Amazon Kinesis Services

Capabilities of Kinesis Set of Services

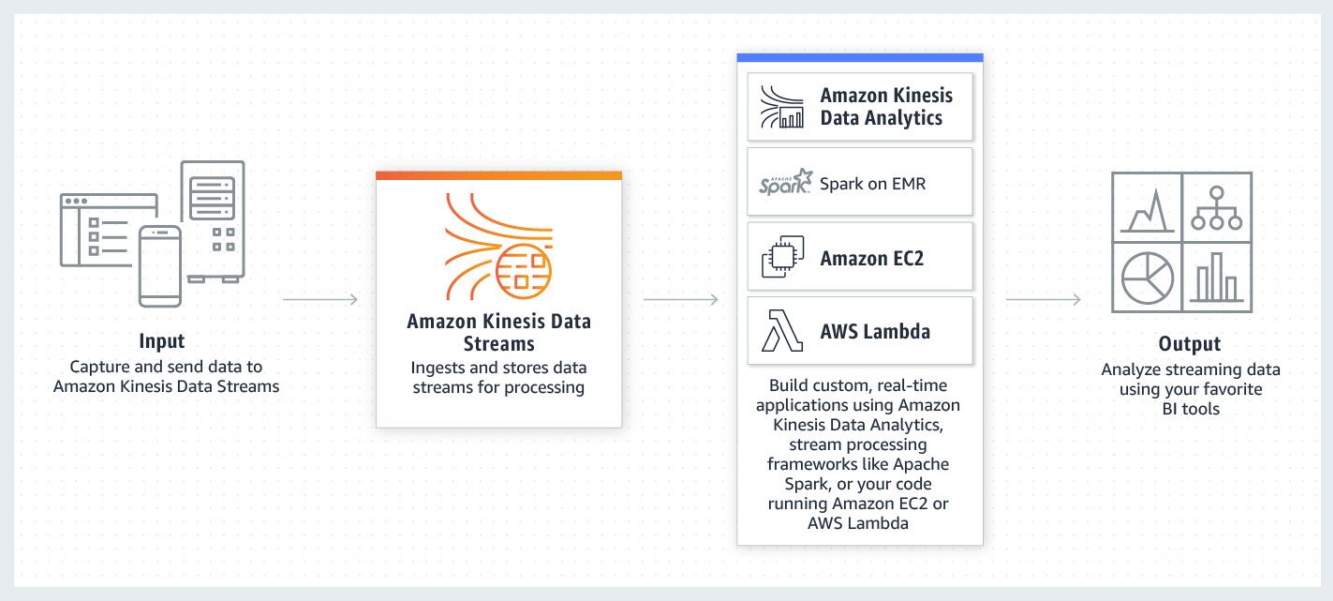
Kinesis Offerings

Amazon Kinesis is a set of services which makes it easy to work with set of streaming data on AWS.

Sr No	Kinesis Services	Description
1	Kinesis Data Stream	Captures, processes and stores data streams in real-time
2	Kinesis Data Firehose	Primary to move data from point A to point B.
3	Kinesis Data Analytics	Analyze streaming data in real-time with SQL / Java code.
4	Kinesis Video Stream	Capture, processes and stores video streams.

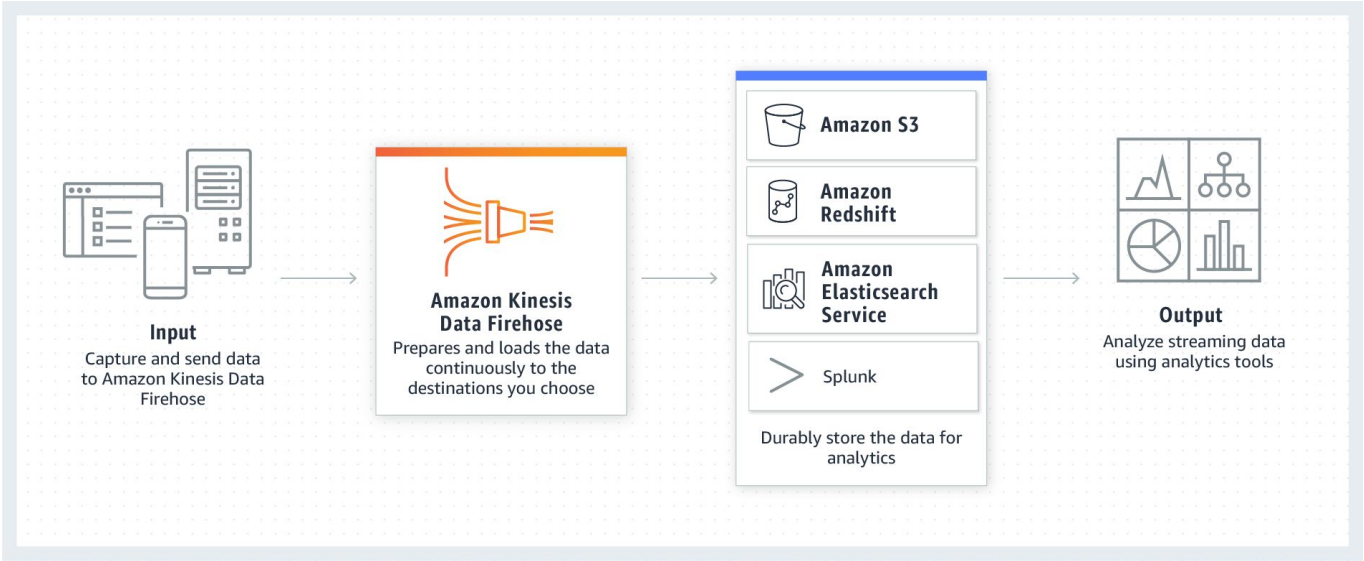
Kinesis Data Stream

It allows us to capture, process and store data streams.



Kinesis Firehose

Kinesis firehose delivers data from point A to point B.



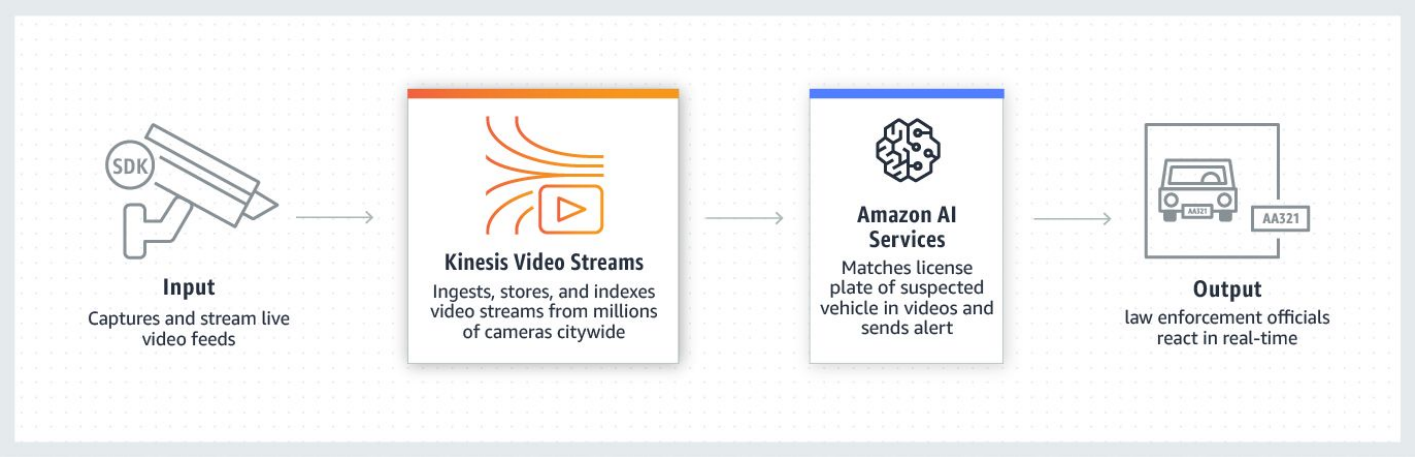
Kinesis Data Analytics

Kinesis Data Analytics has ability to analyze data streams in real time.



Kinesis Video Stream

Amazon Kinesis Video Streams makes it easy to securely stream video from connected devices to AWS



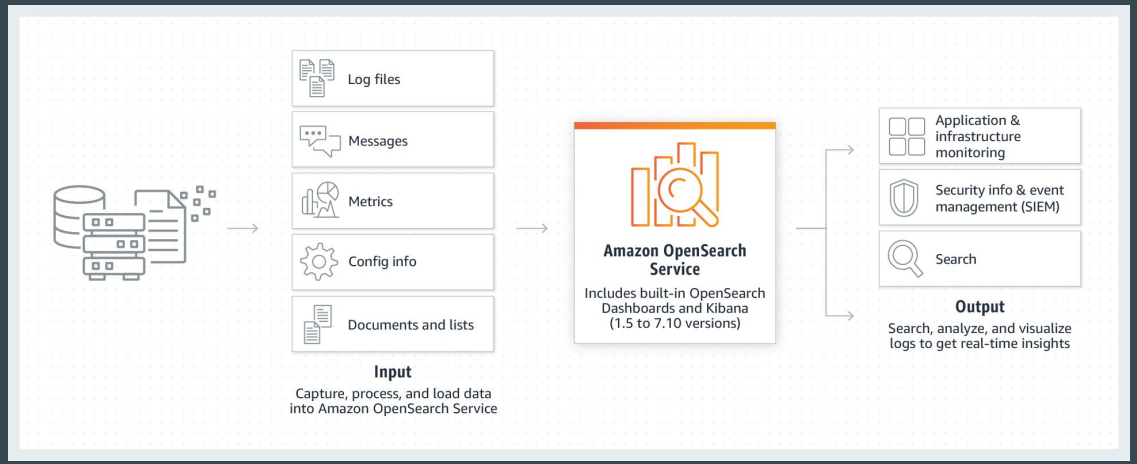
Amazon OpenSearch



Understanding the Basics

Amazon OpenSearch is initially based on the forked version of ElasticSearch

Allow ingesting, searching and visualization of data.

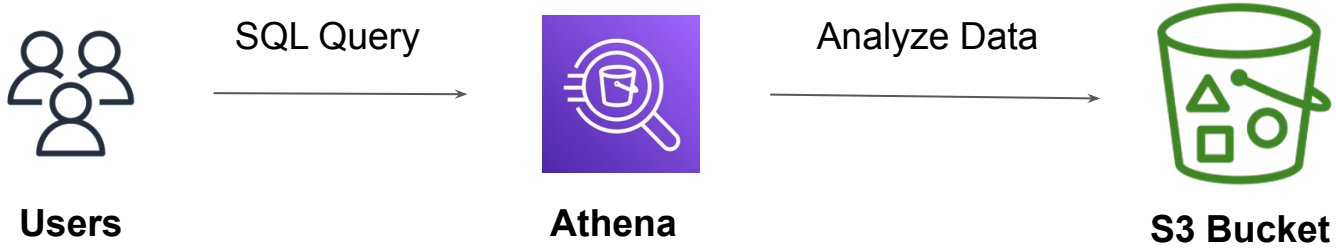


Amazon Athena

Query Logs from S3

Getting the basics right

Amazon Athena is service that allows us to analyze various log files from a data source using standard SQL



Approach Before Athena

You have CloudTrail logs in S3 and you want to see who has logged in, in the past 10 days.

- Create EC2 instances.
- Deploy monitoring stack like Splunk, ELK or others.
- Add the data source from S3 to import CloudTrail logs.
- Begin Analyzing.

VPC Flow Logs

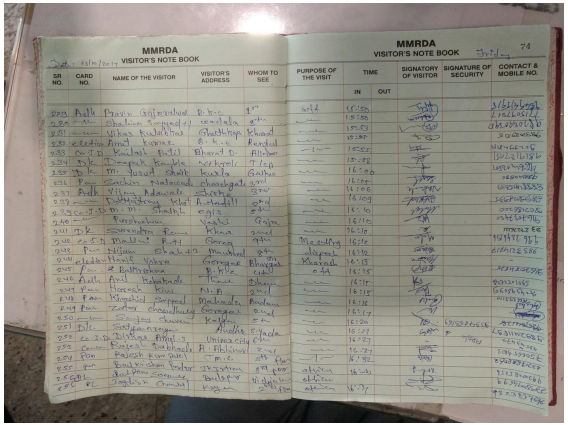
Logs are Awesome

Simple Analogy - Visitor Register

In many of the societies across India, whenever a visitor visits, they first have to fill in their information in the visitor register.

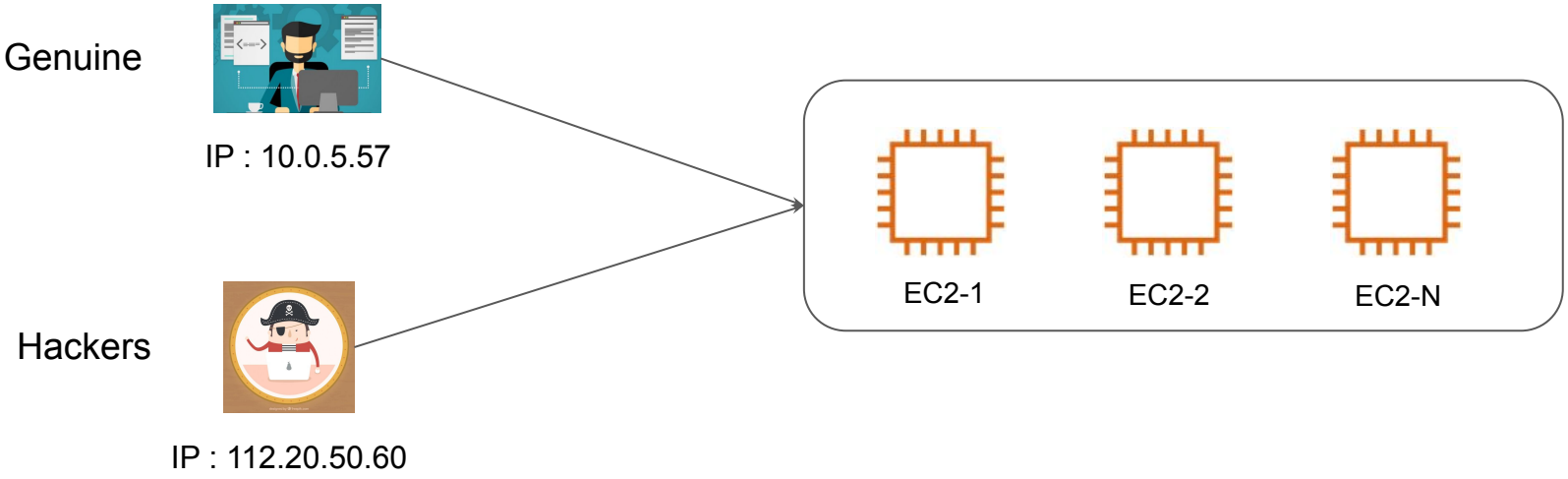
Some of the information includes:

- Name
- Source Place.
- Destination Place.
- Entry and Exit Date/Time
- Purpose of Work



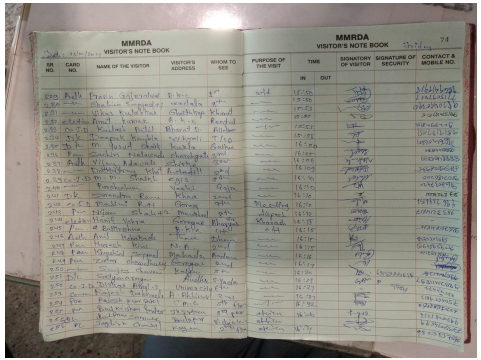
Comparing Analogy with AWS Environment

Even in AWS, there can be thousands of users across the world who might be visiting your environment.

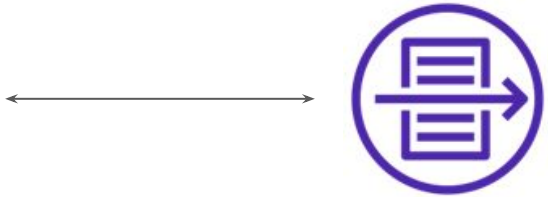


VPC Flow Logs

VPC Flow Logs is a feature that enables you to capture information about the IP traffic going to and from network interfaces in your VPC.



Visitor Register

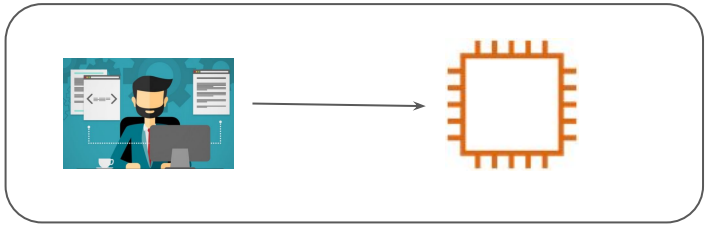


VPC Flow Log

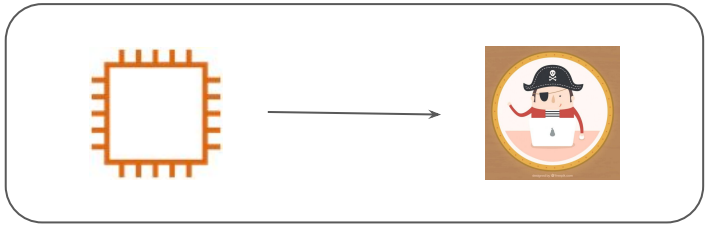
Capture Information Scope

The scope of the VPC Flow logs:

1. Record the traffic information that is visiting the resource (eg EC2)
2. Record data about resource connecting to which outbound endpoint.



10.77.2.50 → EC2 Instance



EC2 Instance → 192.168.0.5

Log events

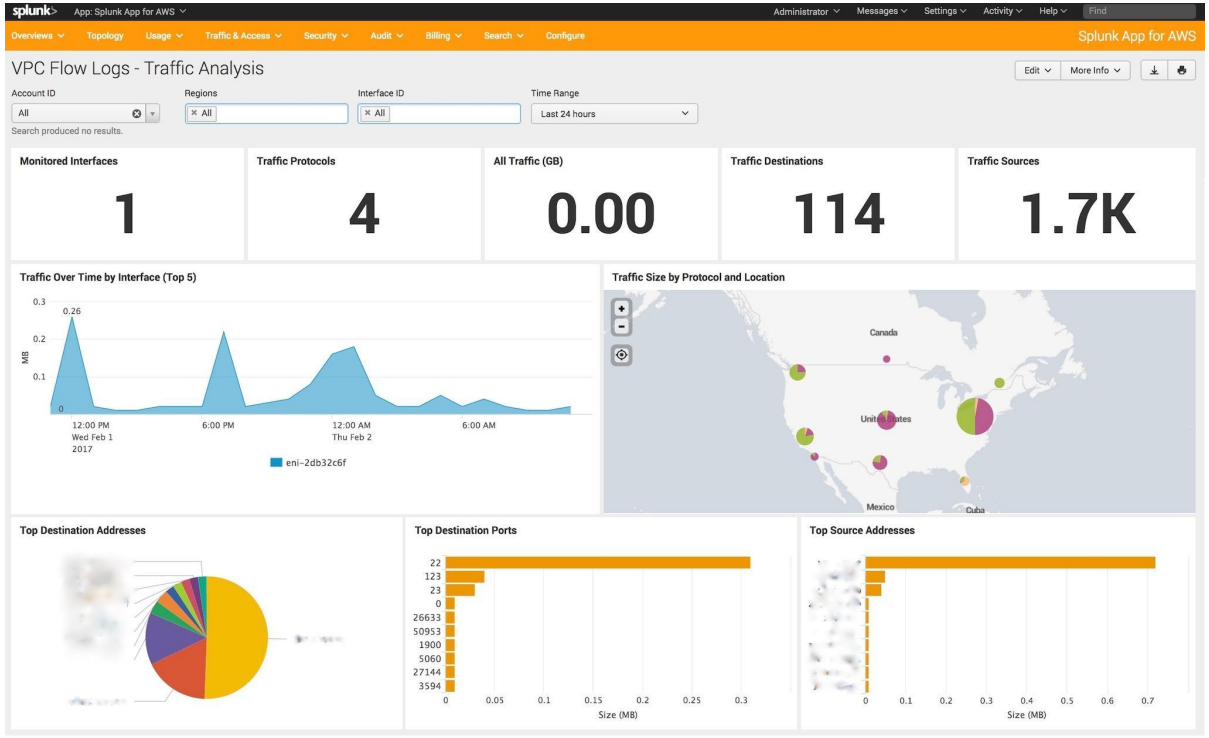
You can use the filter bar below to search for and match terms, phrases, or values in your log events. Learn more about filter patterns

View as text | Actions | Create Metric Filter

Filter events | Clear | 1m | 30m | 1h | 12h | Custom

Timestamp	Message
	No older events at this moment. *Retry*
2021-06-08T21:40:11.000+05:30	2 693331494763 eni-025fffb751de82493 50.205.244.36 172.31.94.239 123 34874 17 1 76 1623168611 1623168640…
2021-06-08T21:40:11.000+05:30	2 693331494763 eni-025fffb751de82493 172.31.94.239 1.116.229.53 80 59807 6 1 40 1623168611 1623168640 AC…
2021-06-08T21:40:11.000+05:30	2 693331494763 eni-025fffb751de82493 8.129.43.176 172.31.94.239 48507 2376 6 1 40 1623168611 1623168640 …
2021-06-08T21:40:11.000+05:30	2 693331494763 eni-025fffb751de82493 172.31.94.239 204.11.201.12 39609 123 17 1 76 1623168611 1623168640…
2021-06-08T21:40:11.000+05:30	2 693331494763 eni-025fffb751de82493 172.31.94.239 138.68.201.49 55618 123 17 1 76 1623168611 1623168640…
2021-06-08T21:40:11.000+05:30	2 693331494763 eni-025fffb751de82493 204.11.201.12 172.31.94.239 123 39609 17 1 76 1623168611 1623168640…
2021-06-08T21:40:11.000+05:30	2 693331494763 eni-025fffb751de82493 138.68.201.49 172.31.94.239 123 55618 17 1 76 1623168611 1623168640…
2021-06-08T21:40:11.000+05:30	2 693331494763 eni-025fffb751de82493 172.31.94.239 69.89.207.199 53680 123 17 1 76 1623168611 1623168640…
2021-06-08T21:40:11.000+05:30	2 693331494763 eni-025fffb751de82493 1.116.229.53 172.31.94.239 59807 80 6 1 40 1623168611 1623168640 AC…
2021-06-08T21:40:11.000+05:30	2 693331494763 eni-025fffb751de82493 69.89.207.199 172.31.94.239 123 53680 17 1 76 1623168611 1623168640…
2021-06-08T21:40:11.000+05:30	2 693331494763 eni-025fffb751de82493 162.142.125.150 172.31.94.239 62446 9143 6 1 44 1623168611 16231686…
2021-06-08T21:40:11.000+05:30	2 693331494763 eni-025fffb751de82493 172.31.94.239 50.205.244.36 34874 123 17 1 76 1623168611 1623168640…
2021-06-08T21:40:46.000+05:30	2 693331494763 eni-025fffb751de82493 107.173.140.175 172.31.94.239 49640 8088 6 1 44 1623168646 16231687…
2021-06-08T21:40:46.000+05:30	2 693331494763 eni-025fffb751de82493 50.205.244.36 172.31.94.239 123 35182 17 1 76 1623168646 1623168700…
2021-06-08T21:40:46.000+05:30	2 693331494763 eni-025fffb751de82493 172.31.94.239 50.205.244.36 35182 123 17 1 76 1623168646 1623168700…

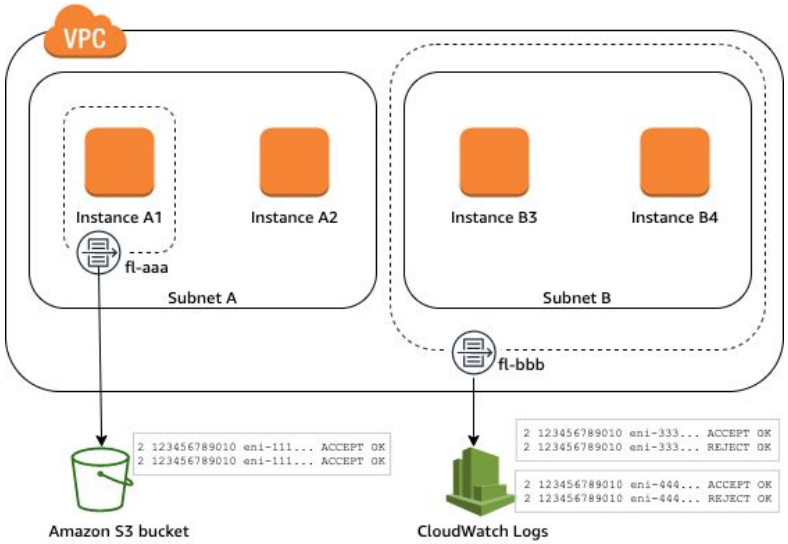
Dashboards Built using VPC Flow Logs Data



Interface Level Flow Logs

VPC Flow Logs captures traffic at an interface level.

Flow logs do not capture real-time log streams for your network interfaces.



High-Level Flow Logs Format

version	-	The VPC Flow Logs Version
account-id	-	AWS Account ID
interface-id	-	The network interface id
srcaddr	-	The source address
destaddr	-	Destination Address
src port	-	Source Port
dest port	-	Destination Port
protocol	-	The protocol number
packets	-	Number of packets transferred
bytes	-	Number of bytes transferred
start	-	Start time in unix seconds
end	-	End time in unix seconds
action	-	ACCEPT or REJECT
log status	-	Logging status of flow log

2 7742829482 eni-4d788e3d 115.73.149.218 10.0.5.157 12053 23 6 2 88 1485439809 1485440090 REJECT OK

Type of Traffic Not Logged

Flow logs do not capture all IP traffic. Some of these include:

- Traffic generated by instances when they contact the Amazon DNS server. If you use your own DNS server, then all traffic to that DNS server is logged.
- Traffic generated by a Windows instance for Amazon Windows license activation.
- Traffic to and from 169.254.169.254 for instance metadata.
- DHCP traffic.

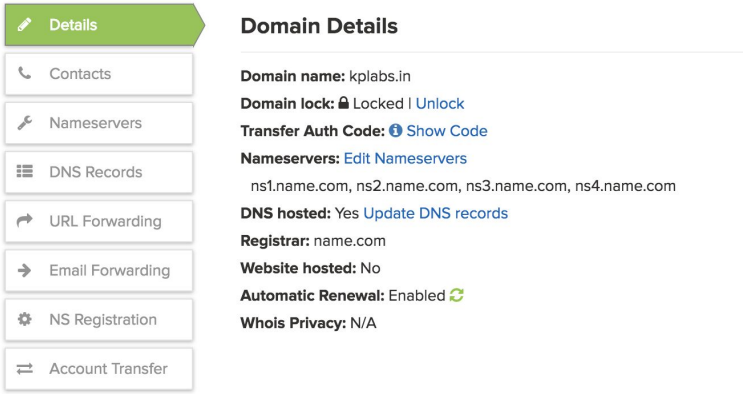
Advance Route53 Features

Interesting Features of Route53

Managed DNS Providers

Generally a managed DNS server supports basic functionality like :

- Domain Registration
- GUI for putting DNS records
- Mapping & Resolving various DNS Records.
- WHOIS Management



Route53 does a lot more

- Support of Public and Private Hosted Zones.
- Routing - Weighted, Latency, Geolocation, Round Robin
- Health Checks & Monitoring
- Route53 Endpoints
- DNS Firewall



Traffic Flow



Hosted Zone



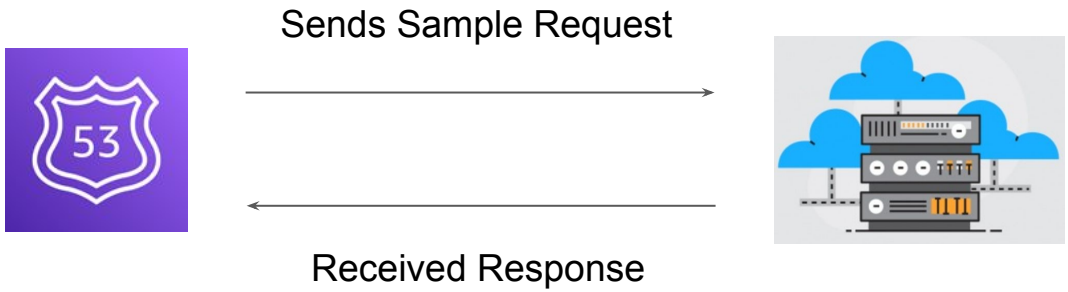
Health Checks

Route53 Health Checks

Back to Monitoring!

Overview of Health Checks

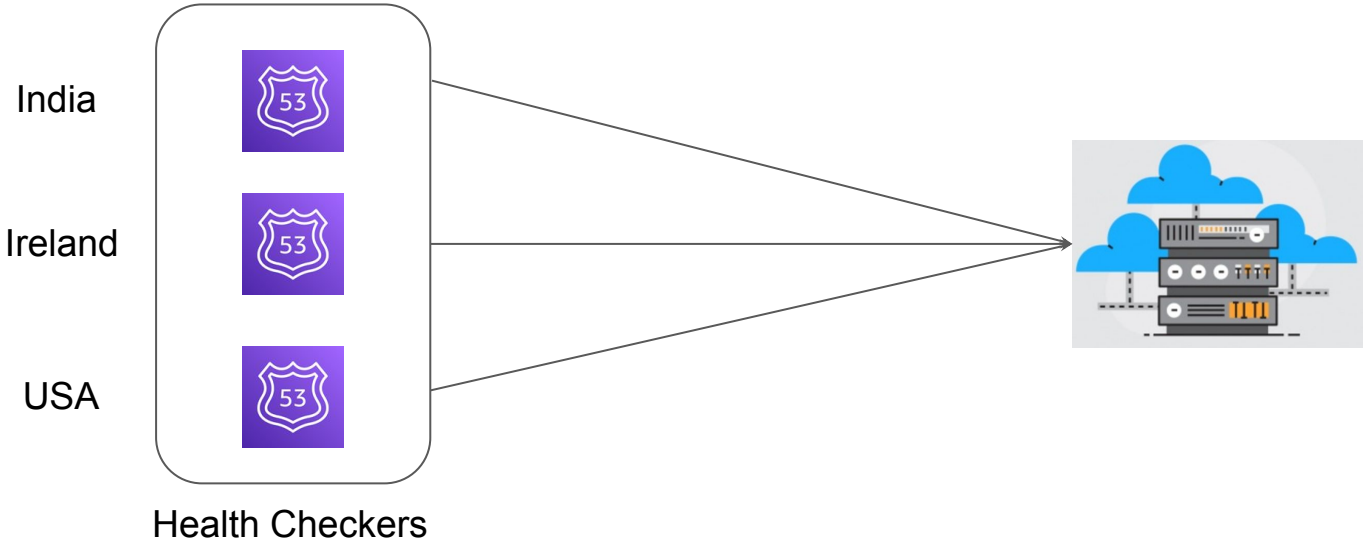
Amazon Route 53 health checks monitor the health and performance of your web applications, web servers, and other resources



Route53 Health Checkers

Route 53 has health checkers in locations around the world.

When you create a health check that monitors an endpoint, health checkers start to send requests to the endpoint that you specify to determine whether the endpoint is healthy.



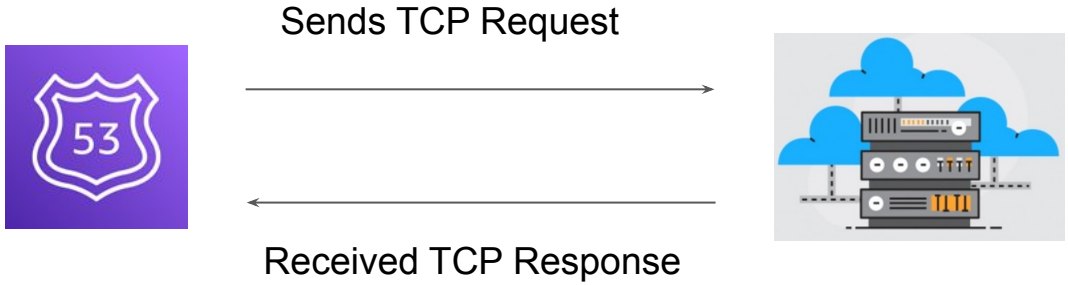
Types of Health Checks

Back to Monitoring!

Type of Health Checks

There are three primary type of Health Checks supported by Route53

1. HTTP and HTTPS health checks
2. TCP health checks
3. HTTP and HTTPS health checks with string matching

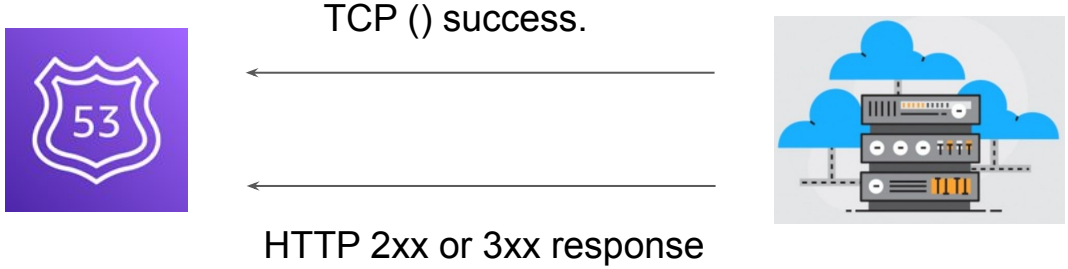


Type 1 - HTTP/HTTPS

Two important factors as part of this health check:

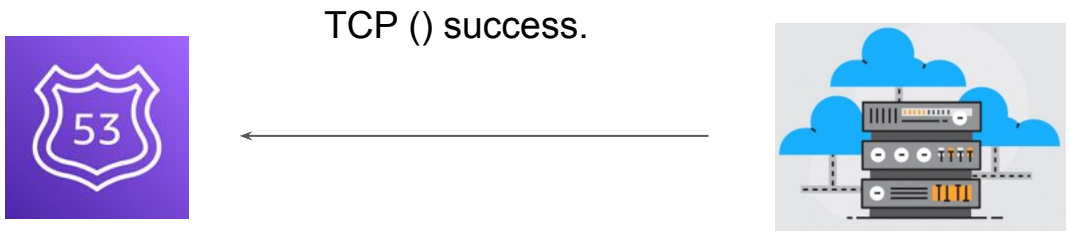
1. Route 53 must be able to establish a TCP connection with the endpoint within four seconds.

2. In addition, the endpoint must respond with an HTTP status code of 2xx or 3xx within two seconds after connecting.



Type 2 - TCP

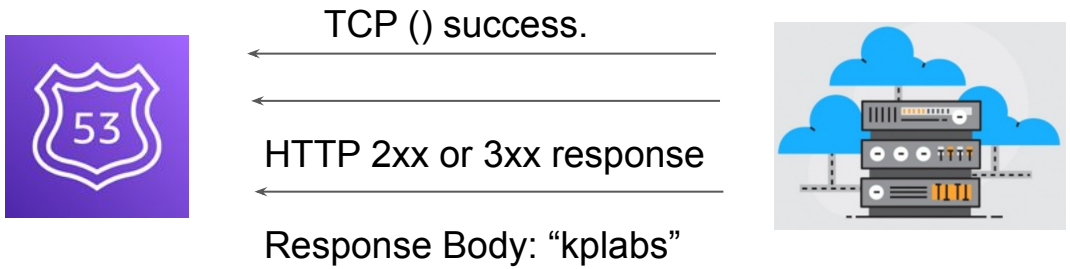
Route 53 must be able to establish a TCP connection with the endpoint within ten seconds.



Type 3 - HTTP/HTTPS with string matching

1. Route 53 must be able to establish a TCP connection with the endpoint within four seconds.
2. In addition, the endpoint must respond with an HTTP status code of 2xx or 3xx within two seconds after connecting.
3. Must receive response body within next two seconds containing a specific string.

4. The string must appear entirely in the first 5,120 bytes of the response body or the endpoint fails the health check.



Routing Policies

Great DNS Provider

Routing Policies

Routing Policies determine how Amazon Route53 responds to the queries.

There are various supported routing policies available in Route53.

Each policy supports a specific use-case.

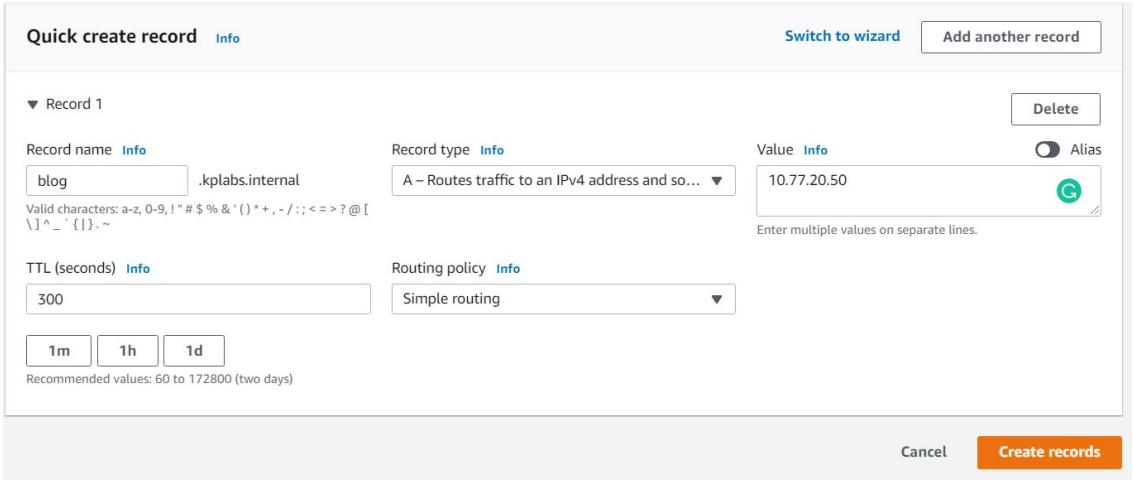


- Simple
- Weighted
- Latency
- Failover
- Geolocation
- Multi-value answer

Simple Routing Policy

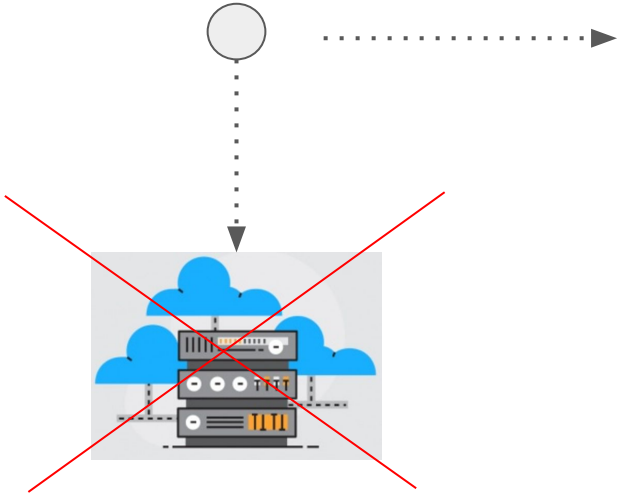
In simple routing, there is a plain one to one mapping between domain and host.

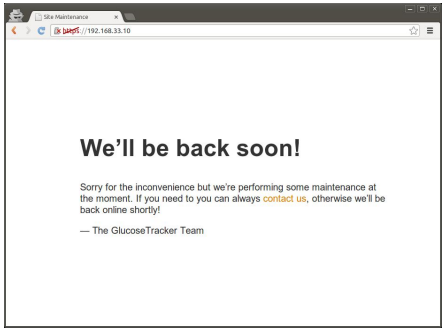
Example: blog.kplabs.internal A 128.199.241.125



Failover Routing

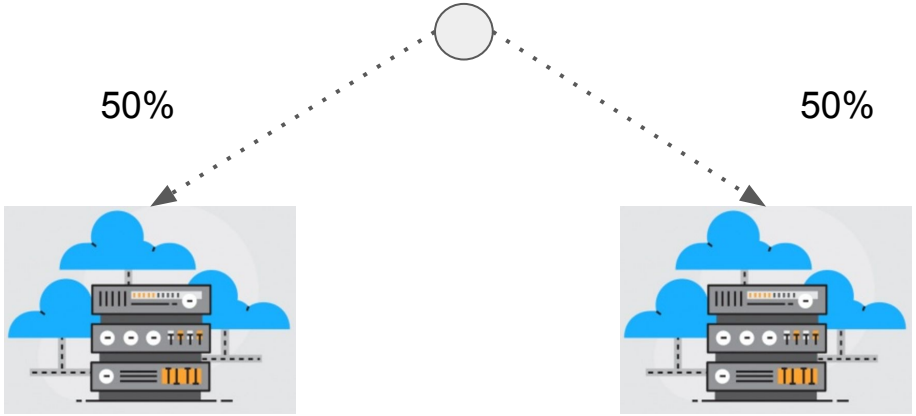
Failover routing lets you route traffic to a resource when the resource is healthy or to a different resource when the first resource is unhealthy.





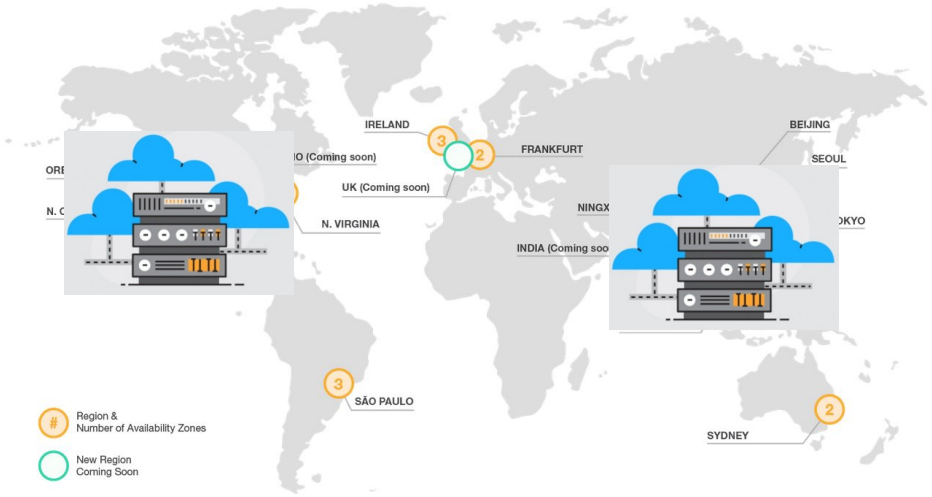
Weighted Routing

Weighted routing helps us to route the traffic to multiple resources in a proportion that we specify from our end.



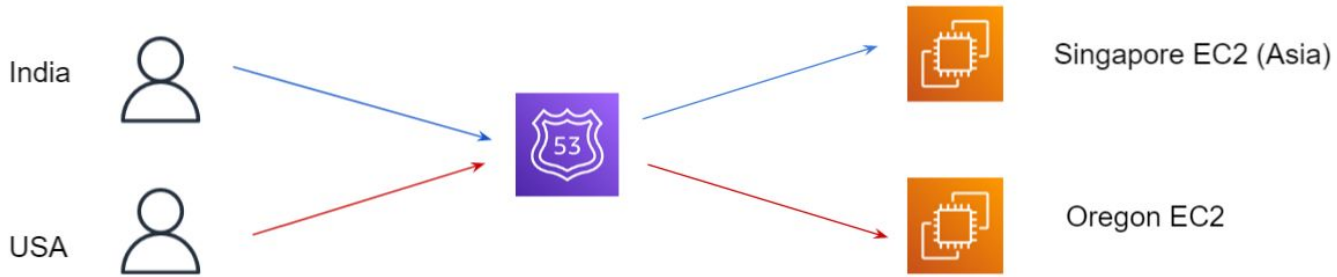
Latency Based Routing

If your application is hosted in multiple AWS regions, we can improve the performance for the users by serving their request from AWS region that provides lowest latency.



GeoLocation Routing

Geolocation routing allows us to choose different resources for different users based on different countries / continents.



Join us in our Adventure

Be Awesome



kplabs.in/twitter



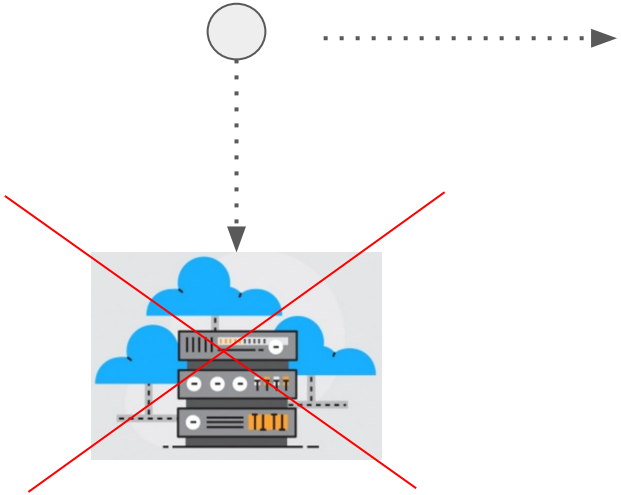
kplabs.in/linkedin

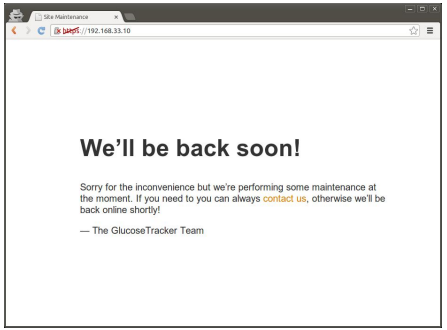
Failover Routing Policy

Back to Monitoring!

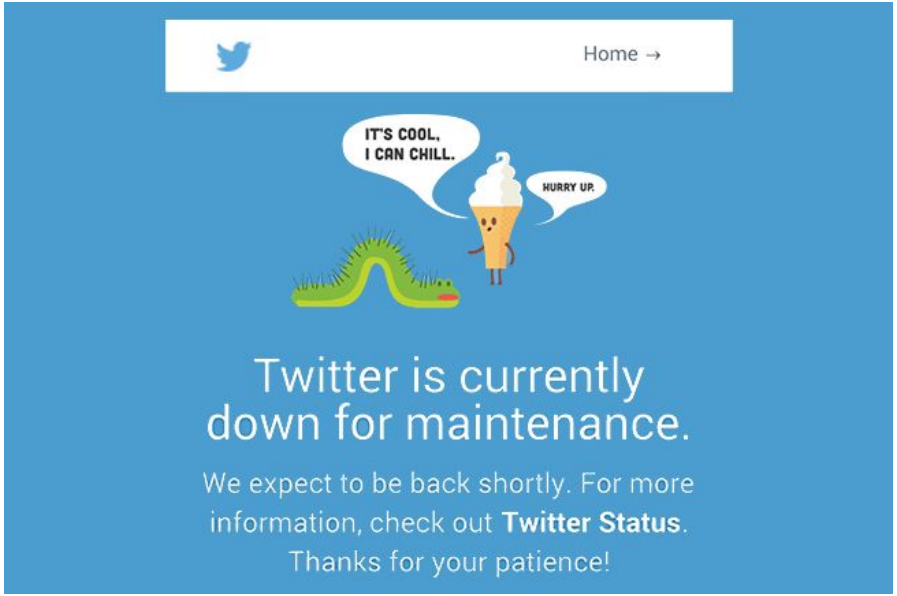
Failover Routing

Failover routing lets you route traffic to a resource when the resource is healthy or to a different resource when the first resource is unhealthy.





Maintenance Page



Relax and Have a Meme Before Proceeding



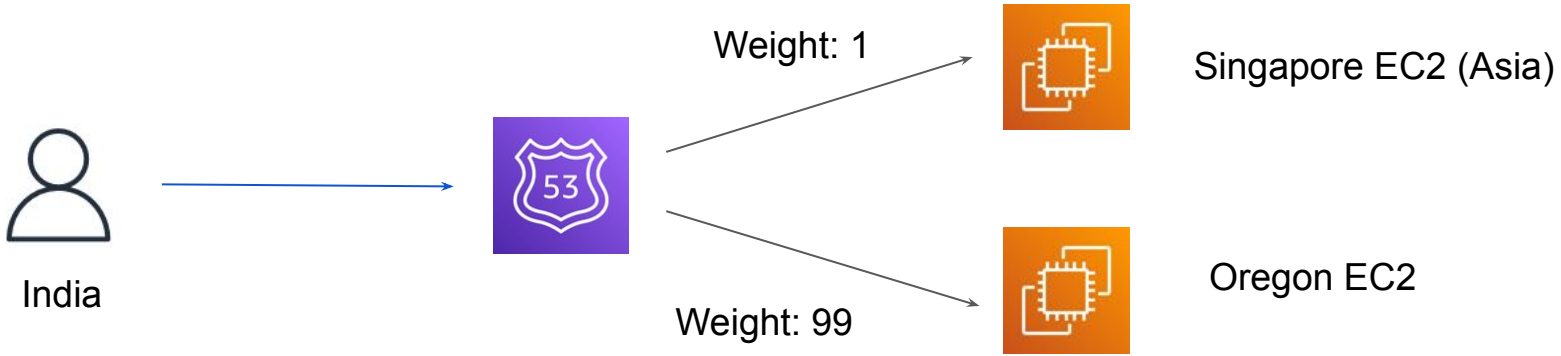
Weighted Routing

Route53 Routing Policy

Overview of Weighted Routing Policy

Weighted Routing allows us to specify the proportion in which traffic should be routed to the underlying servers.

If we want to send small portion of traffic to a new website theme, you can specify the weight of 1 and 99. The resource with 1 gets 1% of the traffic and other gets 99% of traffic.



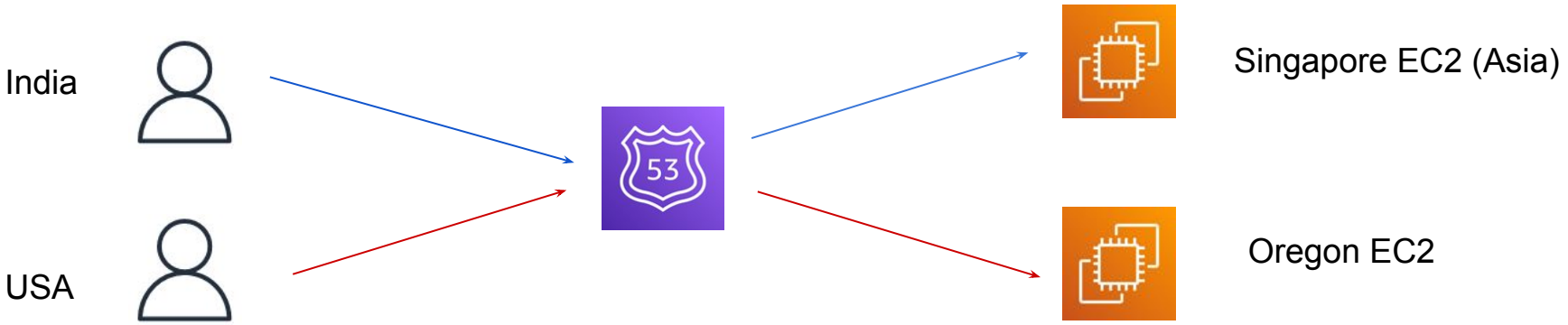
Geolocation Routing

Route53 Routing Policy

Overview of Geolocation Routing

Geolocation Routing allows us to choose resources based on the geographic location of the users

For example, you might want all queries from Asia to be routed to an ELB load balancer in the Singapore region.



Important Caution

Geolocation Routing works by mapping database to IP address.

The results are not always accurate as some ISP might not have any geolocation data associated with them, and some ISP might move the IP block to different country without notification.

For such cases, Route 53 allows us to have a default resource block associated with the routing policy.

Multivalue Answer Routing

Route53 Routing Policy

Overview of Multivalue Answer Routing

Multivalue Answer Routing allows us to return multiple values (such as IP address) in response to a DNS query.

Multivalue Answer Routing also allows us to check the health of the resource so that Route53 responds with details of only healthy resources.

Route53 can respond to up to DNS query with up to 8 healthy records.

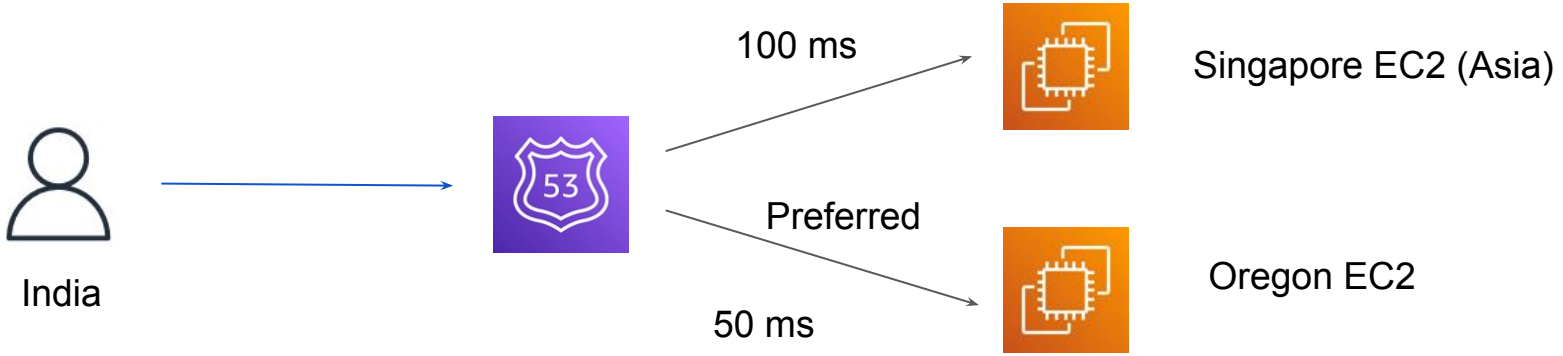
Latency Routing

Route53 Routing Policy

Overview of Latency Based Routing

If your application is hosted in multiple AWS regions, we can improve the performance for the users by serving their request from AWS region that provides lowest latency.

A request that is routed to Singapore today might be routed to India tomorrow.

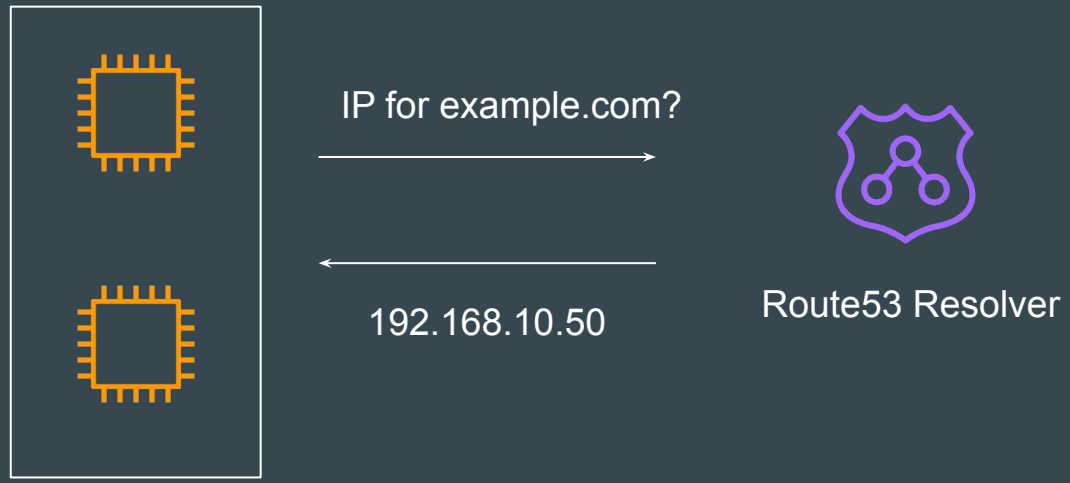


Route53 Resolver



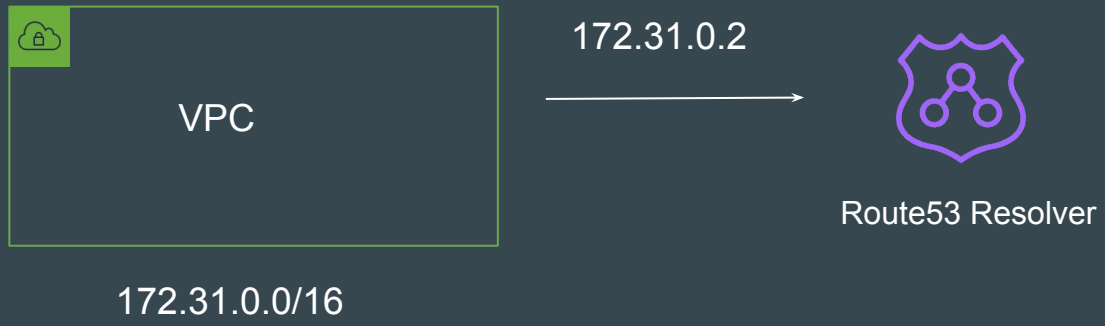
Understanding the Basics

Amazon Route 53 Resolver responds to DNS queries from AWS resources for public records, Amazon VPC-specific DNS names, and Amazon Route 53 private hosted zones, and is available by default in all VPCs.



Address of Route53 Resolver

An Amazon VPC connects to a Route 53 Resolver at a VPC+2 IP address.



Contents of /etc/resolv.conf file of EC2 instance.

```
[ec2-user@ip-172-31-86-117 ~]$ cat /etc/resolv.conf
; generated by /usr/sbin/dhclient-script
search ec2.internal
options timeout:2 attempts:5
nameserver 172.31.0.2
```

Query Resolution

A Route 53 Resolver automatically answers DNS queries for:

1. Local VPC domain names for EC2 instances (for example, ec2-192-0-2-44.compute-1.amazonaws.com).

2. Records in private hosted zones (for example, acme.example.com).

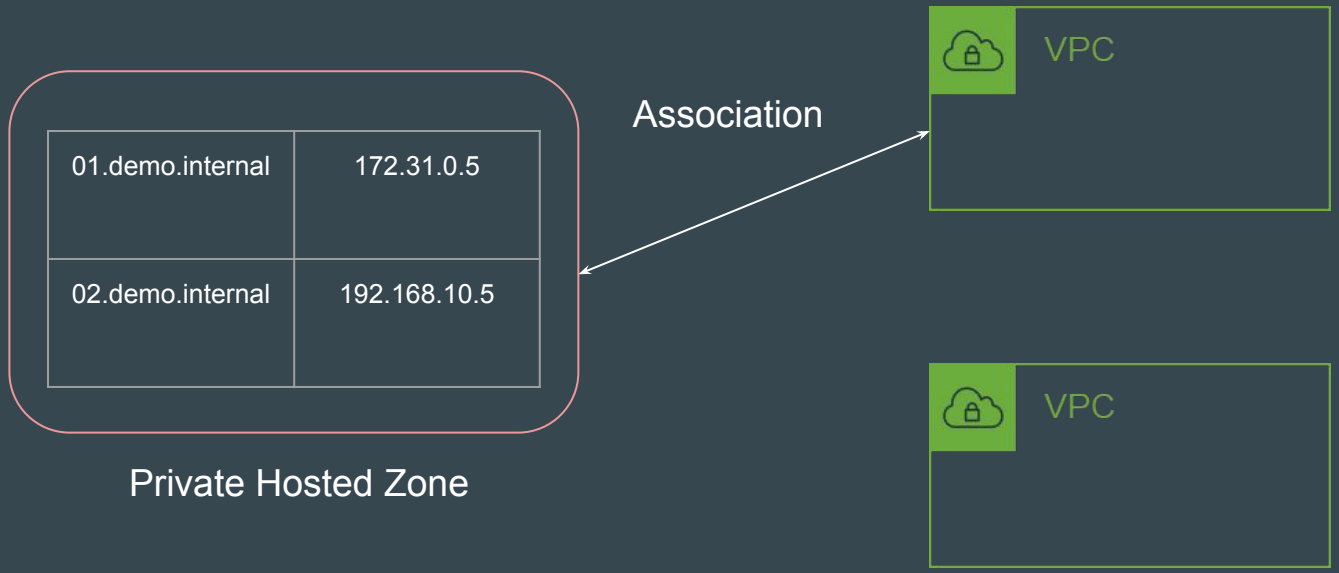
3. For public domain names, Route 53 Resolver performs recursive lookups against public name servers on the internet.

Hybrid DNS



Revising the Basics

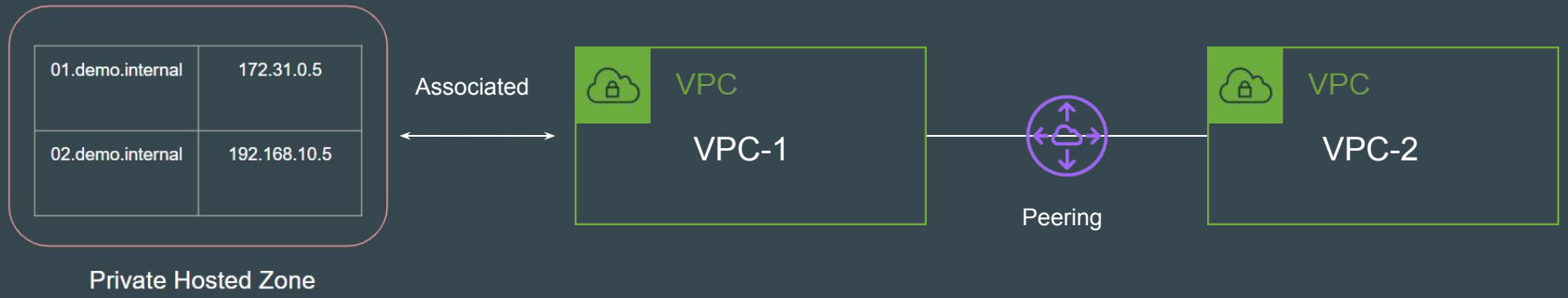
Private hosted zones contain records that specify how you want to route traffic in an Amazon VPC.



Problem Statement

Route53 does not respond to queries which are not originating from the VPC.

If EC2 from VPC-2 sends query for 01.demo.internal to VPC-1 + 2 Address, it will not be resolvable.



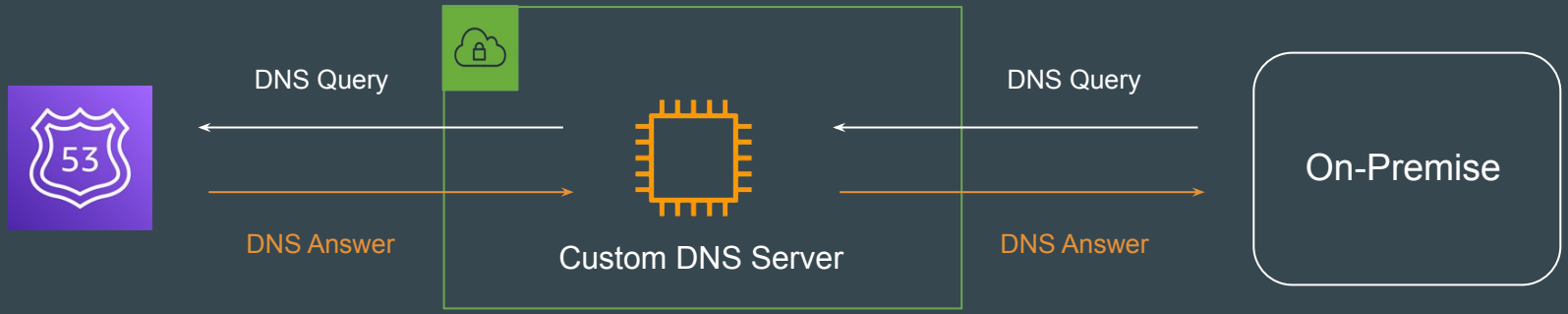
Hybrid Architecture

On-Premise servers will not be able to resolve the domain of the Private Hosted Zone through VPC-1 + 2 address.



Probable Architecture

- Host Custom DNS Server in VPC.
- On-Premise sends queries to Custom DNS Server.
- Custom DNS Server forwards it to Route53 and responds back.



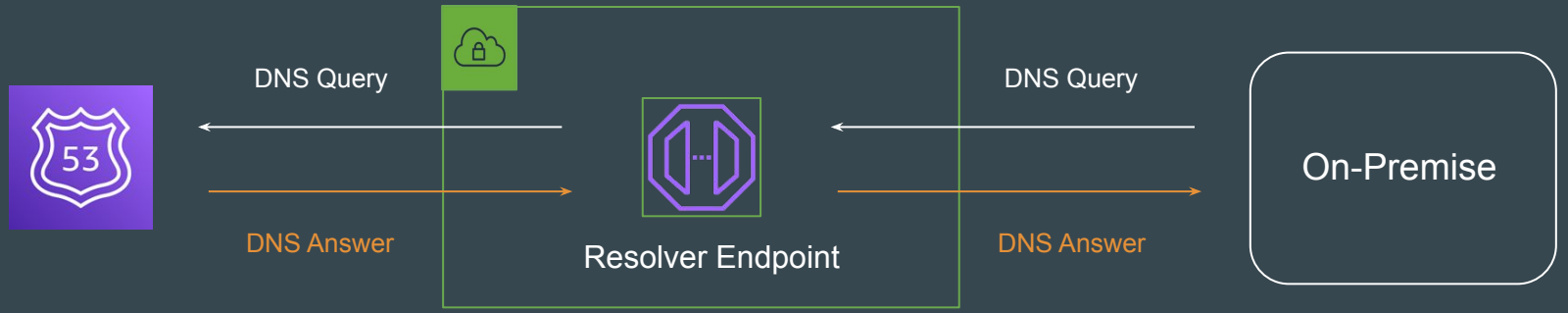
Challenge with Custom DNS Architecture

Customer has to manage Custom DNS Server.

High-Availability, Security, Scalability, Optimization = Customer Responsibility

Better Solution - Resolver Endpoints

AWS has released Route53 Resolver Endpoints that can allow resolution of private hosted zone domains from outside of the VPC.

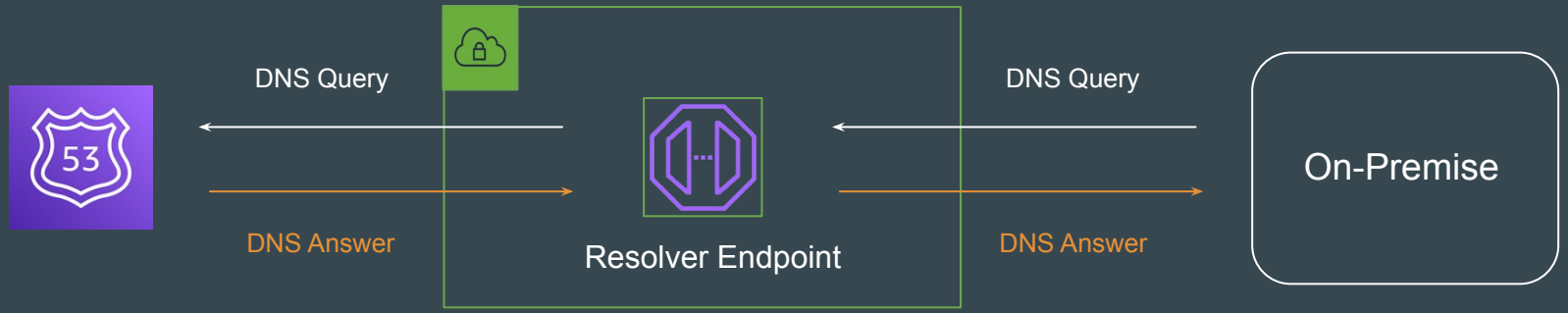


Route53 Resolver Endpoints



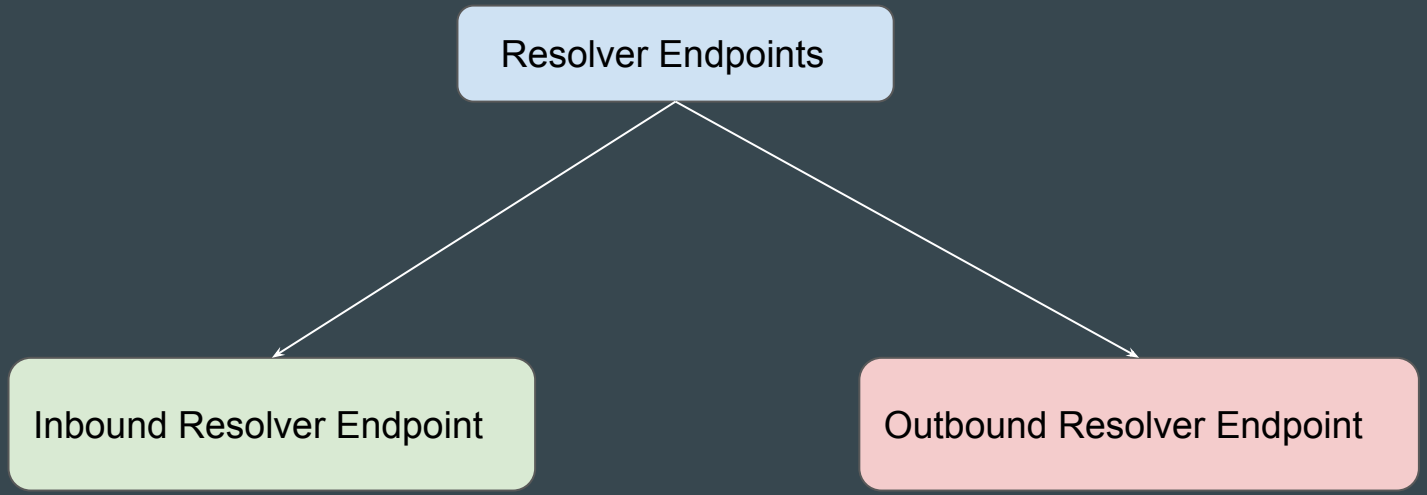
Basics of Route53 Resolver Endpoints

AWS has released Route53 Resolver Endpoints that can allow resolution of private hosted zone domains from outside of the VPC.



Types of Resolver Endpoints

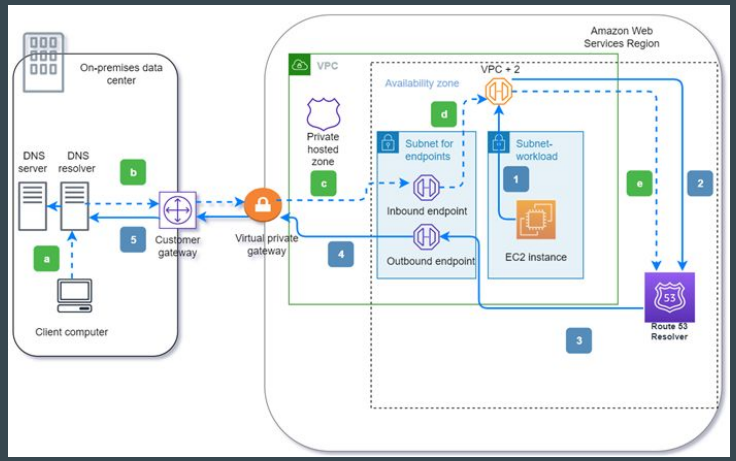
There are 2 primary types of Resolver Endpoints



Understanding Resolver Endpoints

Inbound Resolver endpoints allow DNS queries to your VPC from your on-premises network or another VPC.

Outbound Resolver endpoints allow DNS queries from your VPC to your on-premises network or another VPC.

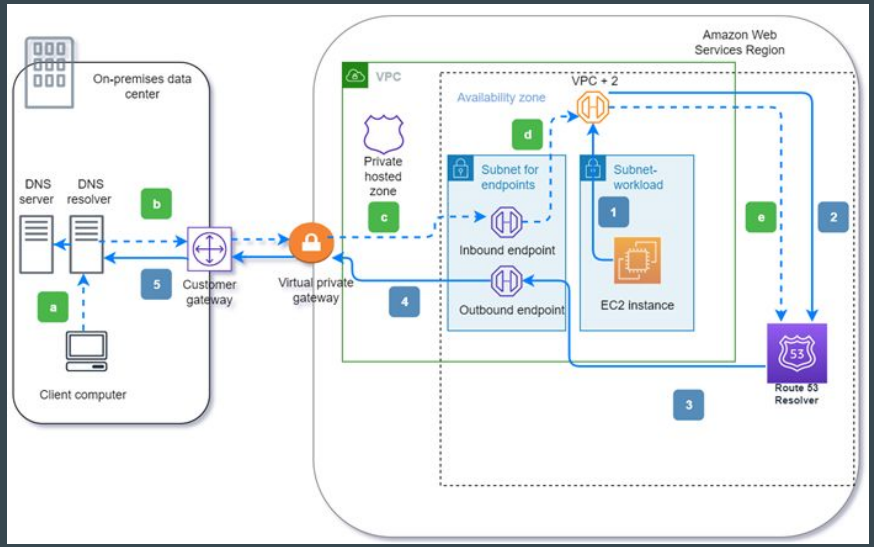


Route53 Resolver - Inbound Endpoints



Understanding Resolver Endpoints

Inbound Resolver endpoints allow DNS queries to your VPC from your on-premises network or another VPC.



Inbound Endpoint Workflow - 1

1. A client in the on-premises data center needs to resolve a DNS query to an AWS resource for the domain dev.example.com. It sends the query to the on-premises DNS resolver.

2. The on-premises DNS resolver has a forwarding rule that points queries to dev.example.com to an inbound endpoint.

3. The query arrives at the inbound endpoint through a private connection, such as AWS Direct Connect or AWS Site-to-Site VPN, depicted as a virtual gateway.

Inbound Endpoint Workflow - 2

4. The inbound endpoint sends the query to Route 53 Resolver at the VPC +2.

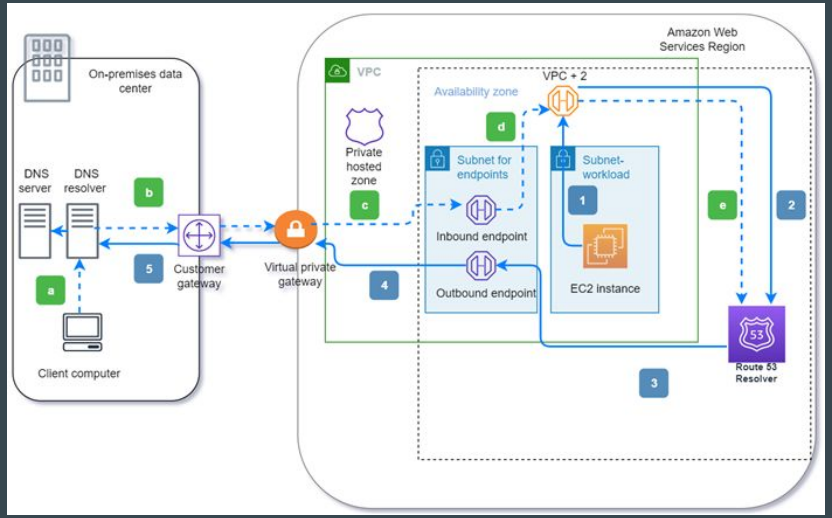
5. Route 53 Resolver resolves the DNS query for dev.example.com and returns the answer to the client via the same path in reverse.

Route53 Resolver - Outbound Endpoints



Understanding Outbound Endpoints

Outbound Resolver endpoints allow DNS queries from your VPC to your on-premises network or another VPC.



Outbound Endpoint Workflow - 1

1. An Amazon EC2 instance needs to resolve a DNS query to the domain internal.example.com. The authoritative DNS server is in the on-premises data center. This DNS query is sent to the VPC+2 in the VPC that connects to Route 53 Resolver.

2. A Route 53 Resolver forwarding rule is configured to forward queries to internal.example.com in the on-premises data center.

3. The query is forwarded to an outbound endpoint.

Outbound Endpoint Workflow - 2

4. The outbound endpoint forwards the query to the on-premises DNS resolver through a private connection between AWS and the data center. The connection can be either AWS Direct Connect or AWS Site-to-Site VPN, depicted as a virtual private gateway.

5. The on-premises DNS resolver resolves the DNS query for internal.example.com and returns the answer to the Amazon EC2 instance via the same path in reverse.

CNAME vs ALIAS

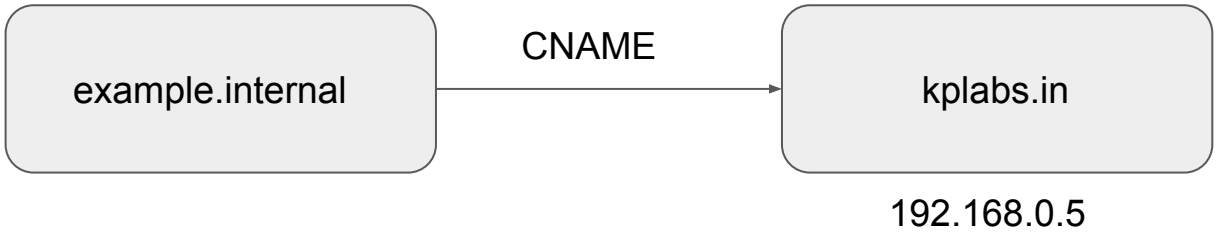
Understand the Use-Case

CNAME Records

CNAME records are used for pointing a domain name to another hostname.

In below diagram, example.internal has CNAME to kplabs.in

When we resolve example.internal, the response will be 192.168.0.5

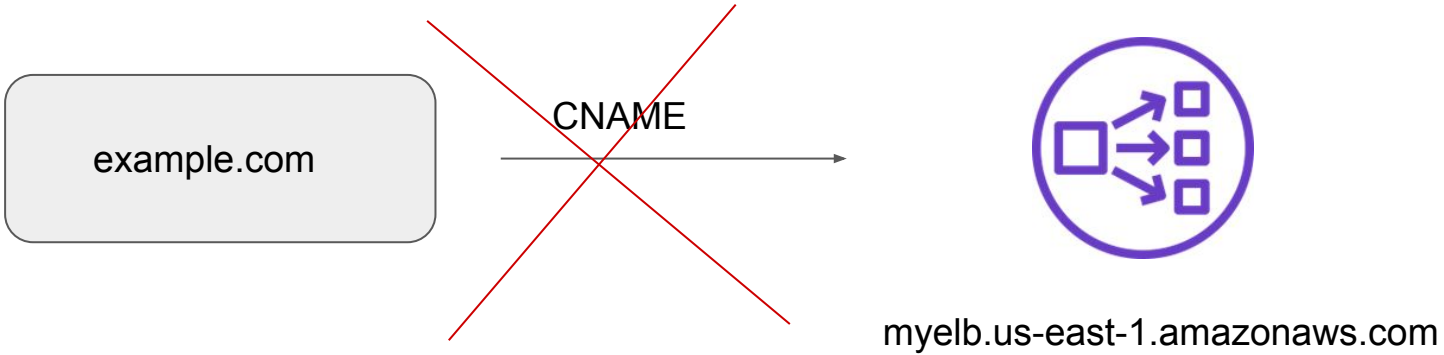


Challenge with CNAME Record

There is one important drawback of CNAME Record.

It cannot be used with the ROOT domain.

example.com CNAME acme.com << CANNOT WORK



Use ALIAS Record

If we want ROOT domain to point to ELB, S3 Bucket, CloudFront distribution, it will not work with CNAME Records.

To resolve this drawback of CNAME record, we make use of ALIAS record which allows us to even point ROOT domain to DNS of AWS Services.

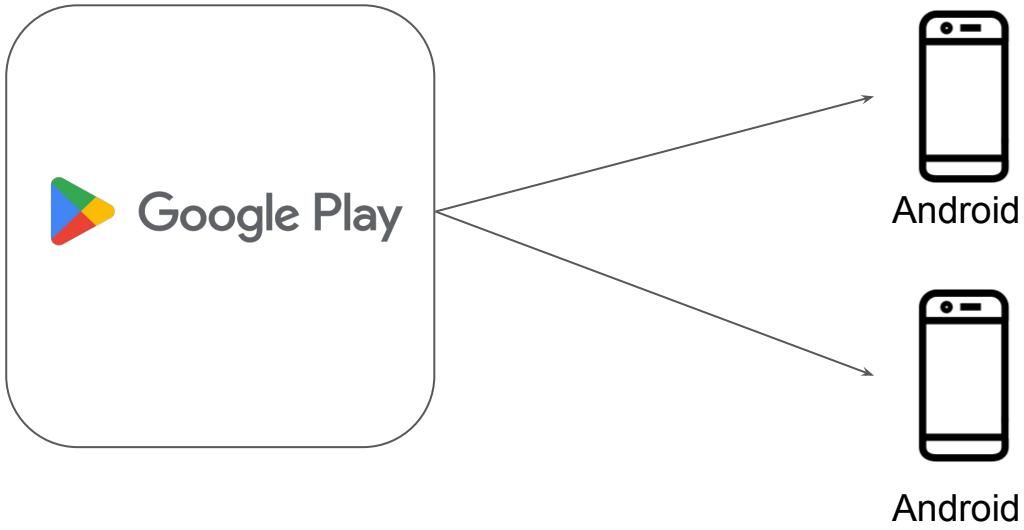


Elastic Container Registry (ECR)

Storing Container Images

Understanding with Analogy

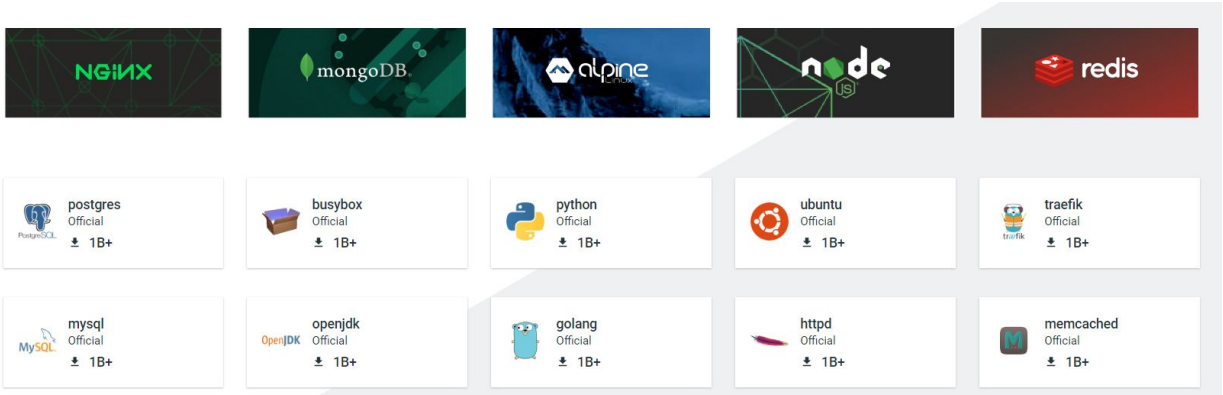
Google Play is an online store where people go to find their favorite apps, games, movies, TV shows, books, and mor



Importance of Container Registry

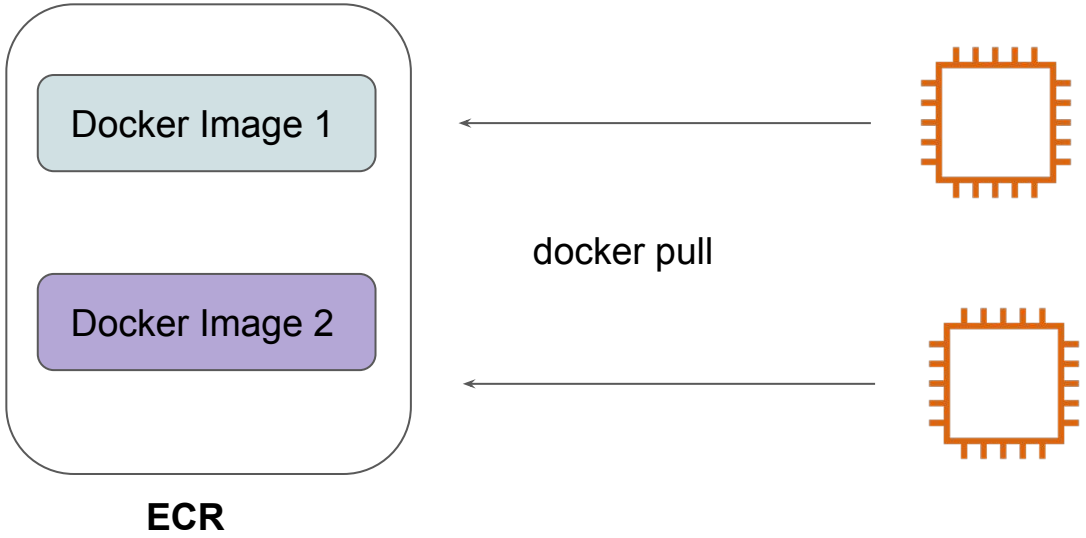
Container Registry is a single place for your team to manage Docker images.

Whenever you launch a Docker Container, the associated image is pulled from Registry.



Basics of ECR

Amazon ECR is a fully managed container registry for storing Docker Images.

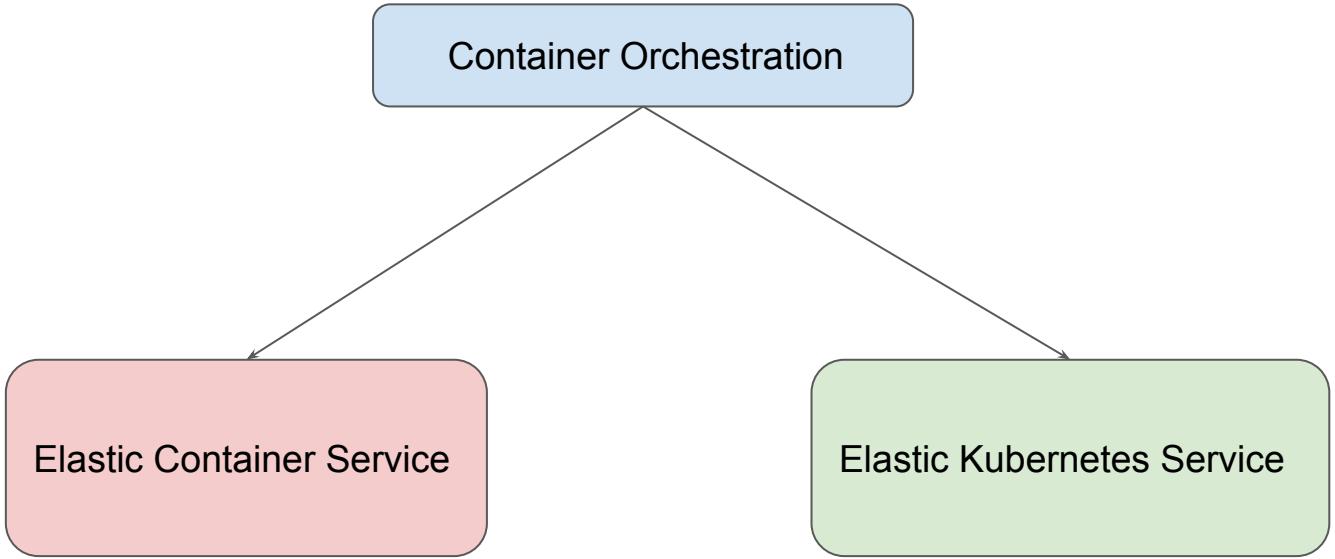


Container Orchestration in AWS

Choosing Right Orchestrator

Container Orchestration in AWS

There are two primary services that are extensively used for container orchestration use-cases.



Important Difference

Pointers	AWS EKS	AWS ECS
Open-Source	Yes	No
Complexity	More Complex	Less Complex
Community Support	More	Less

Choosing Right Orchestrator

If you plan to work exclusively on AWS, you should choose ECS as it offers more in-depth AWS integration than Amazon EKS.

Organizations with limited expertise and insufficient resources to invest in learning Kubernetes can go with ECS.

If you plan to deploy containers across multiple platforms, you can choose EKS.

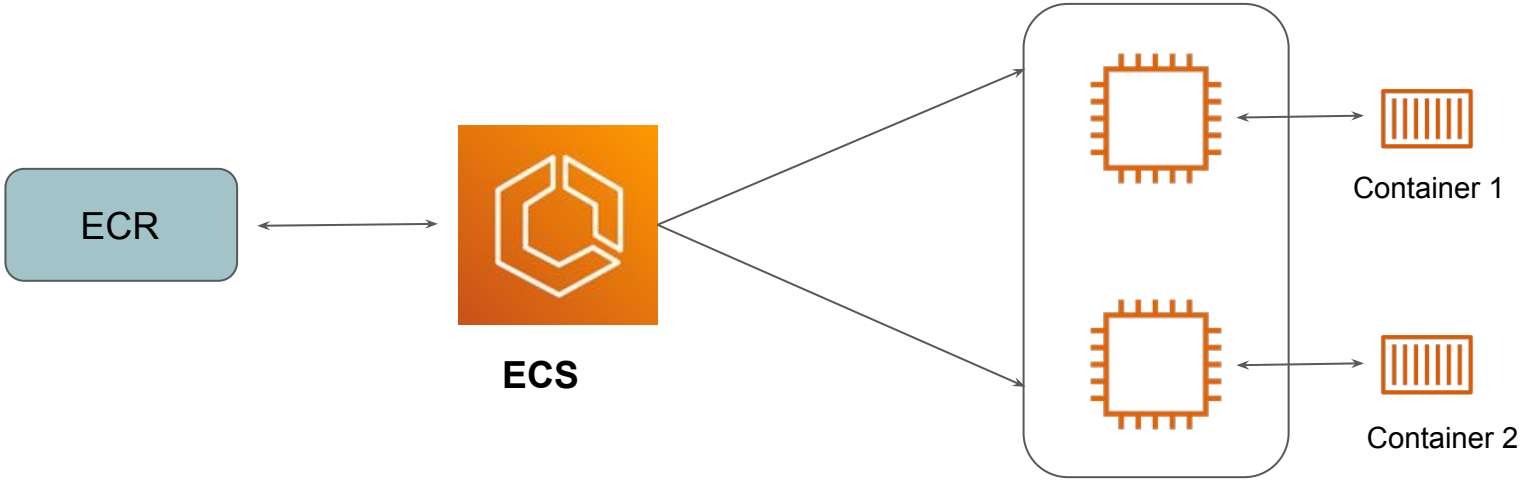
Elastic Container Service (ECS)

Container Management

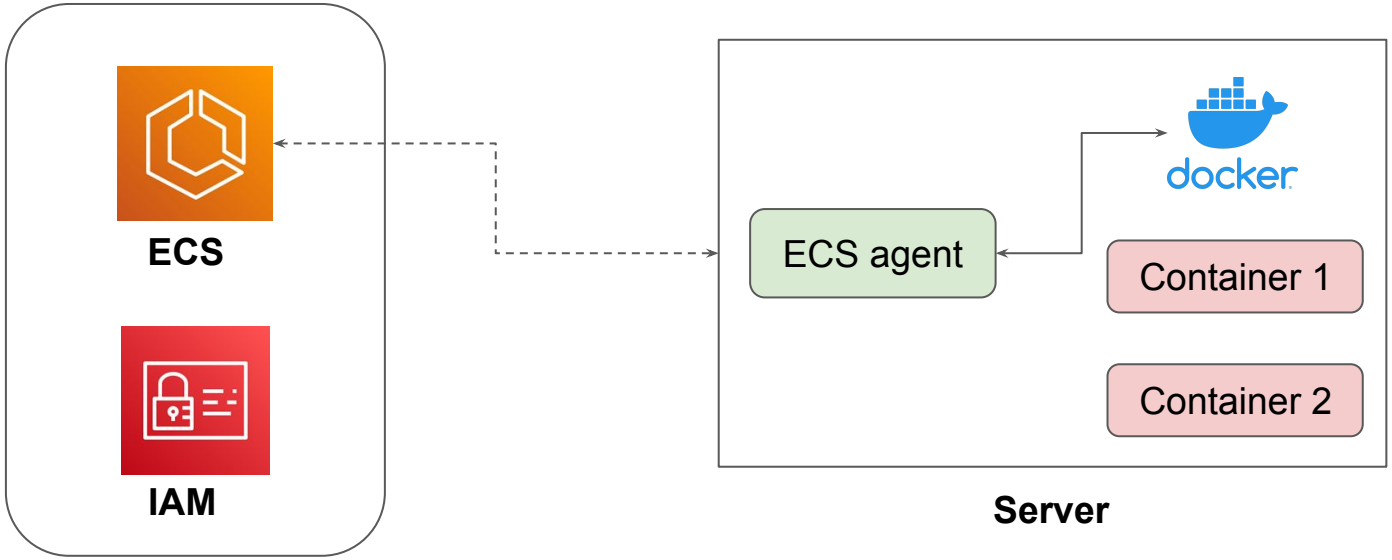
Basics of Service

Amazon Elastic Container Service (Amazon ECS) is a highly scalable and fast container management service.

You can use it to run, stop, and manage containers on a cluster.



High-Level Workflow



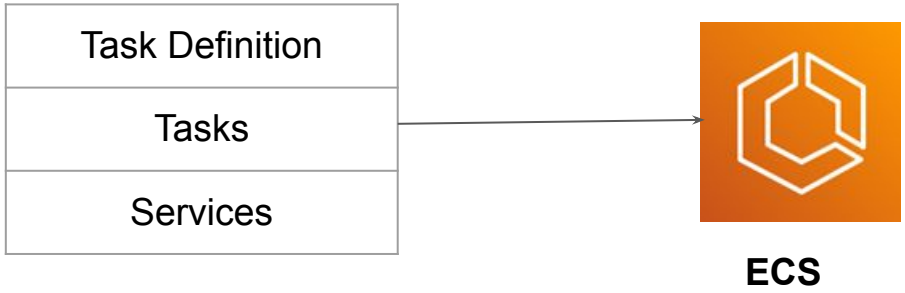
Components of ECS

Container Management

Basic Components

There are three primary components of ECS Cluster:

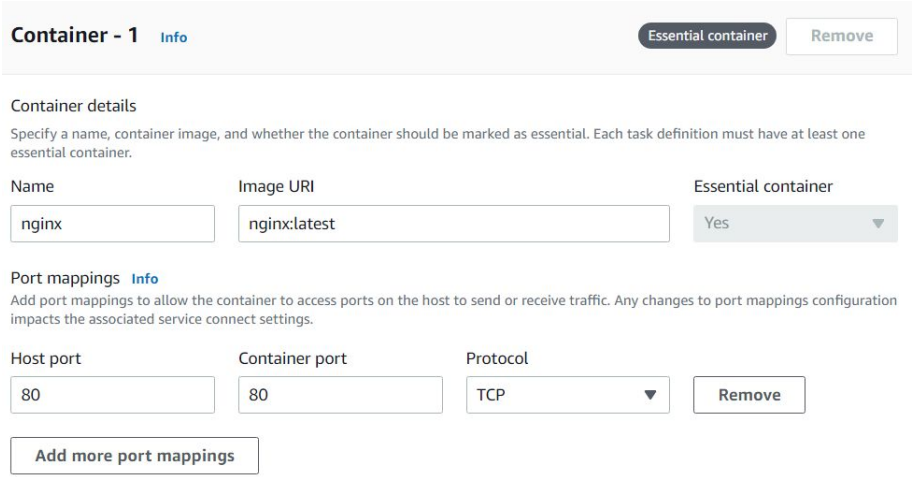
Task Definition, Tasks and Service



Component - Task Definition

A task definition is a text file that describes one or more containers that form your application.

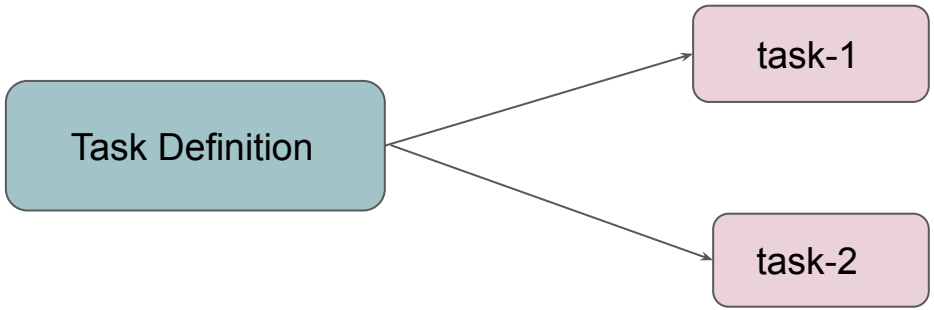
It contains information like operating system, containers to use, ports to open, storage



Component - Task

A task is the instantiation of a task definition within a cluster.

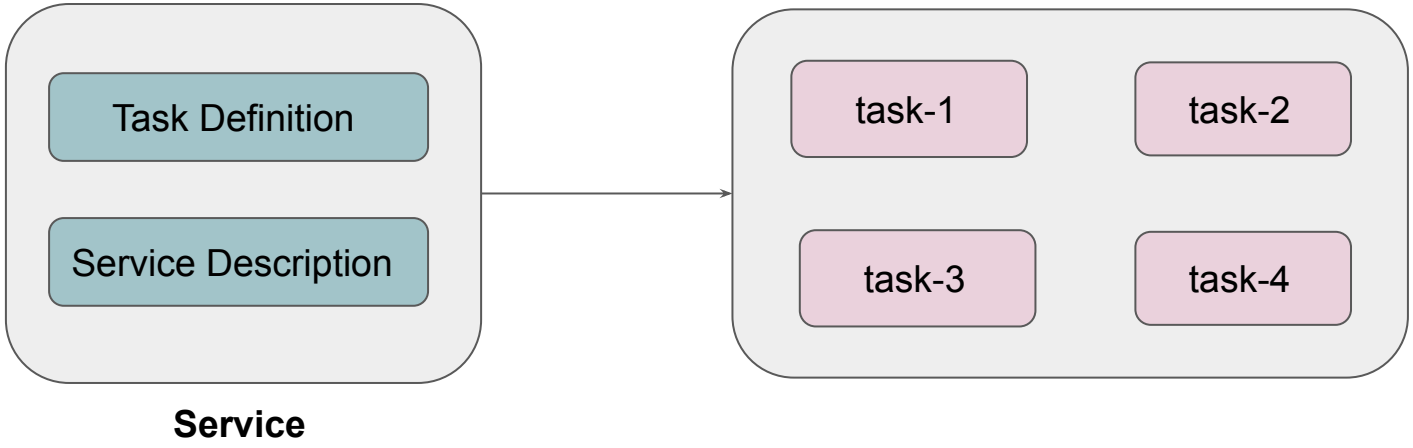
After you create a task definition for your application within Amazon ECS, you can specify the number of tasks to run on your cluster.



Component - Service

Service to run and maintain your desired number of tasks simultaneously in an Amazon ECS cluster.

If any of your tasks fail or stop for any reason, the Amazon ECS service scheduler launches another instance based on your task definition



ECS Networking

Container Management

Let's Network

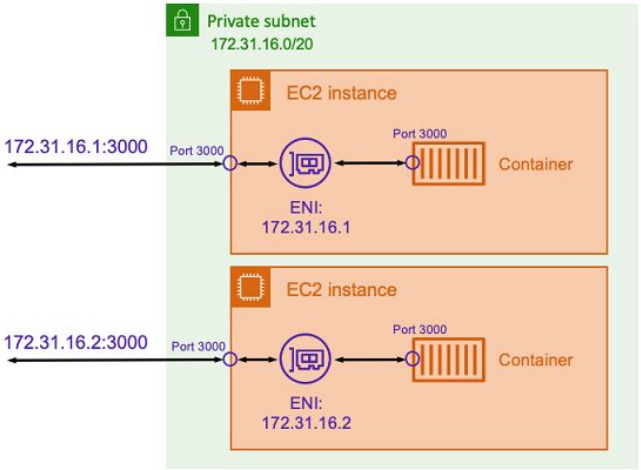
There are 3 primary networking mode that you can use in ECS

Network Mode	Description
Host Mode	The networking of the container is tied directly to the underlying host that's running the container.
Bridge Mode	The bridge network mode allows you to use a virtual network bridge to create a layer between the host and the networking of the container.
AWS VPC Mode	With the awsvpc network mode, Amazon ECS creates and manages an Elastic Network Interface (ENI) for each task and each task receives its own private IP address within the VPC.

Host Mode (Not Recommended)

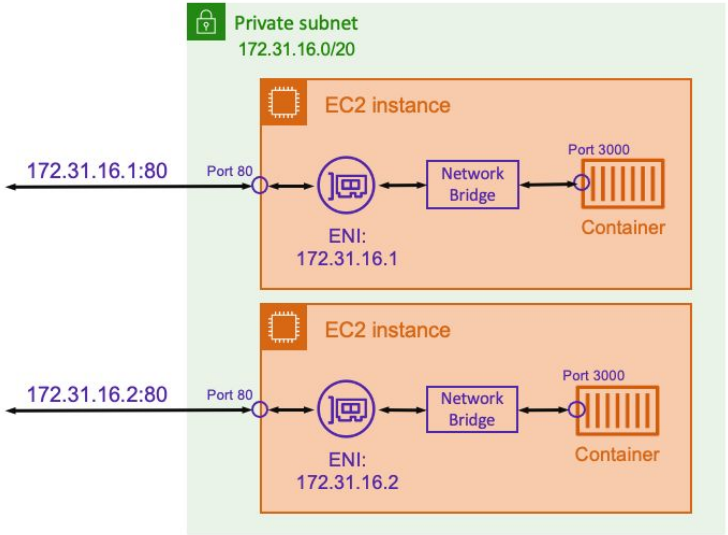
The networking of the container is tied directly to the underlying host that's running the container.

To connect to container: HOST IP + Port



Bridge Mode

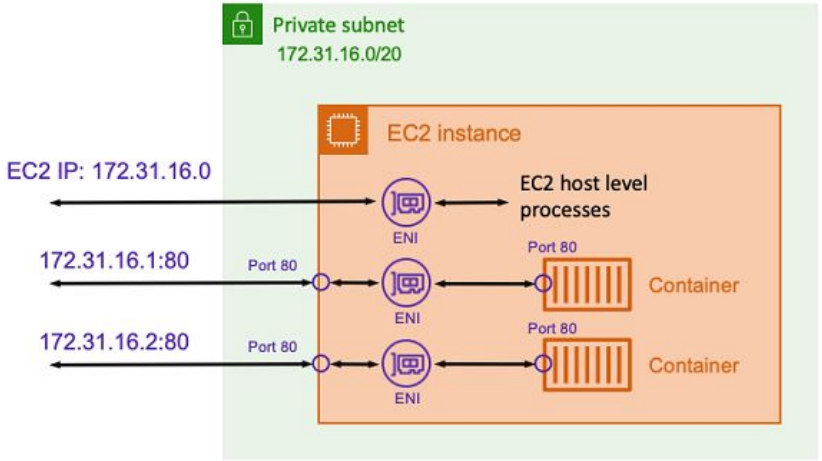
With bridge mode, you're using a virtual network bridge to create a layer between the host and the networking of the container.



AWS VPC Mode

Amazon ECS creates and manages an Elastic Network Interface (ENI) for each task and each task receives its own private IP address within the VPC.

This ENI is separate from the underlying hosts ENI.



Introduction to Kubernetes

Orchestrator Engine

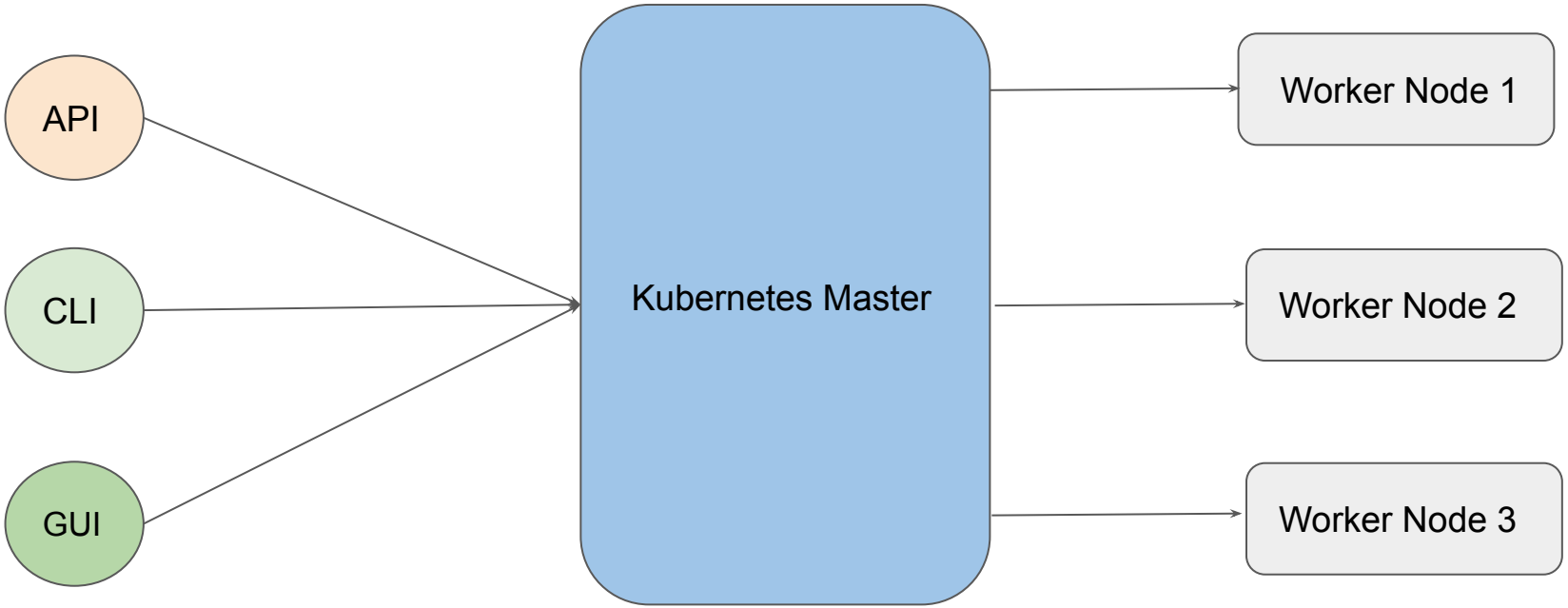
Introduction to Kubernetes

Kubernetes (K8s) is an open-source container orchestration engine developed by Google.

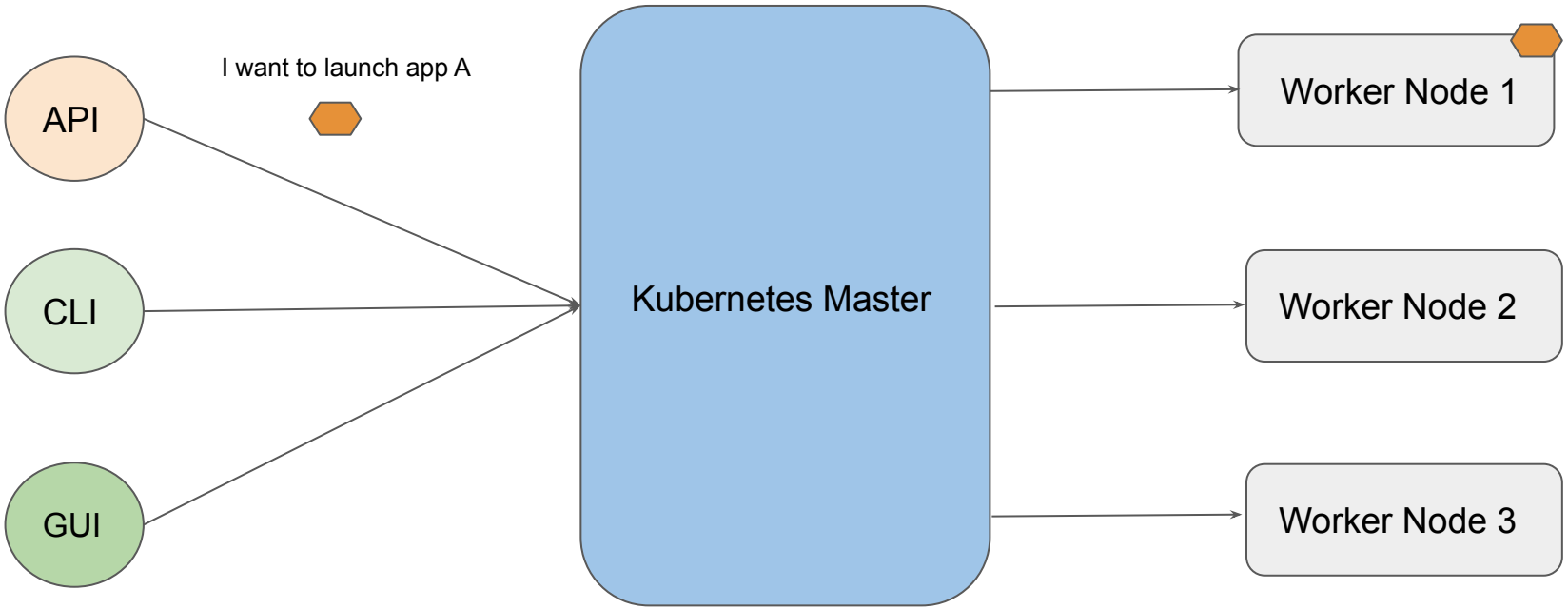
It was originally designed by Google, and is now maintained by the Cloud Native Computing Foundation.



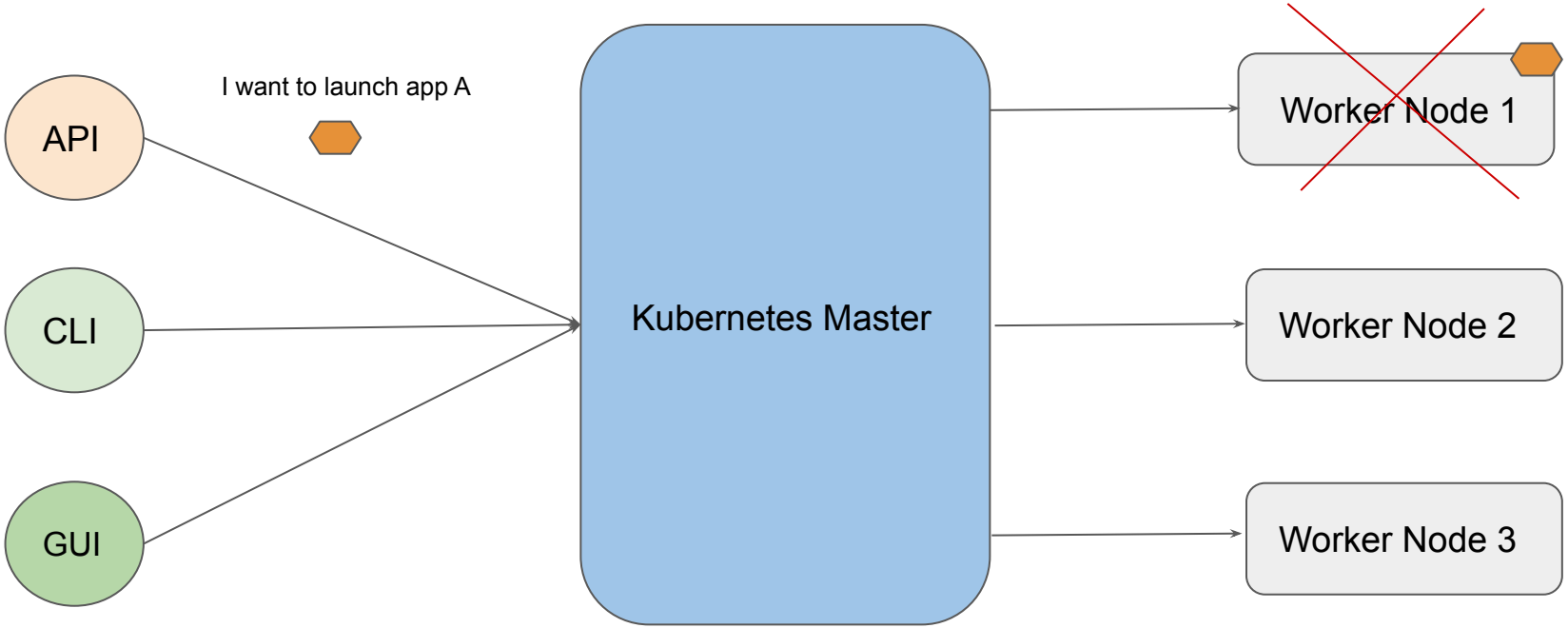
Architecture of Kubernetes



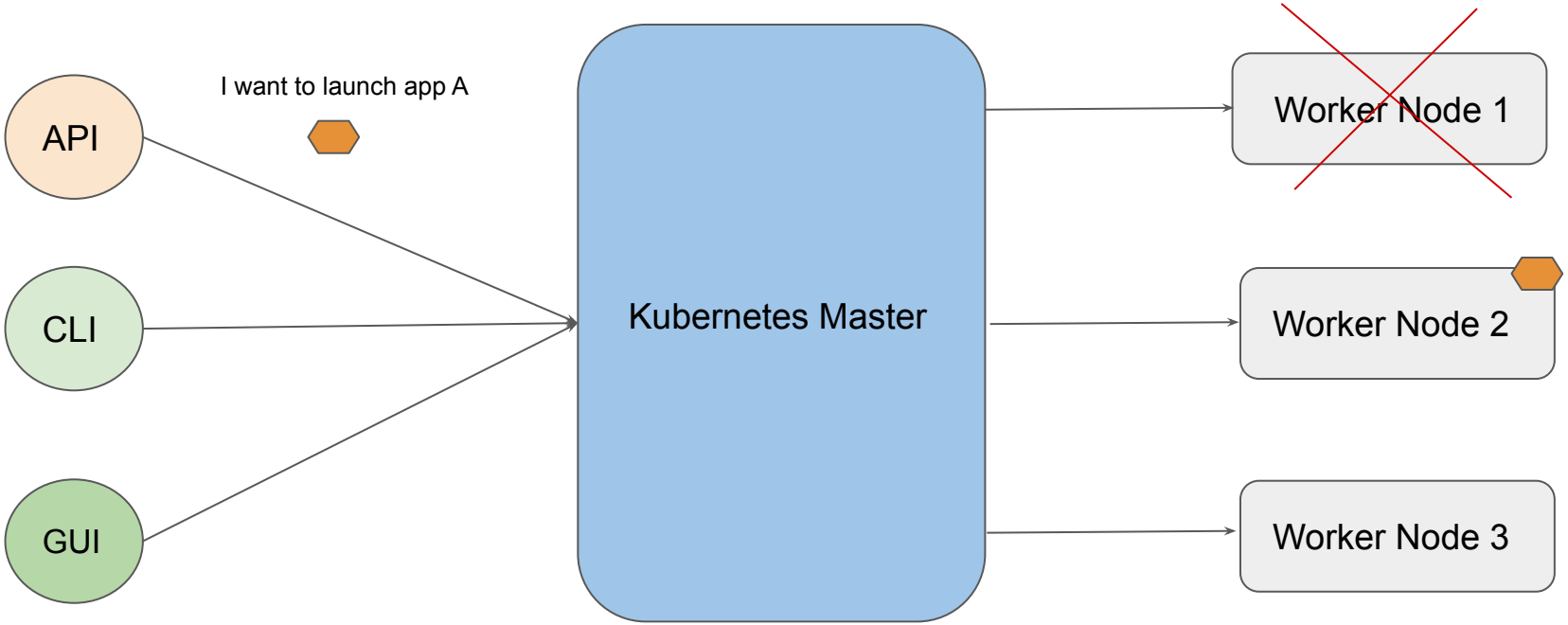
Architecture of Kubernetes



Architecture of Kubernetes



Architecture of Kubernetes

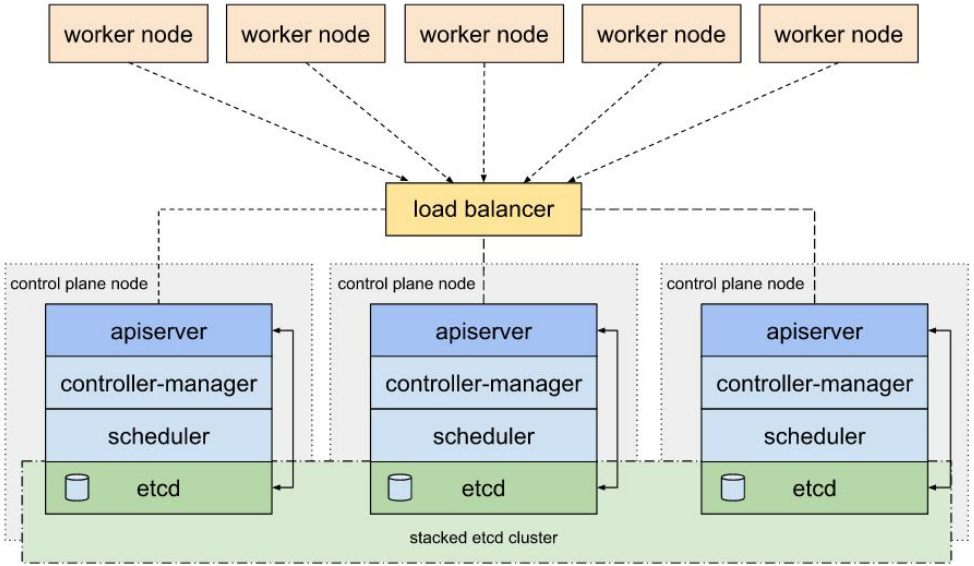


Elastic Kubernetes Service

Managed Kubernetes in AWS

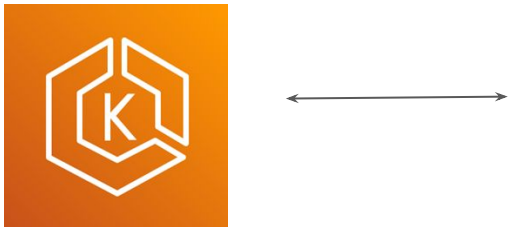
Operating Kubernetes is Hard

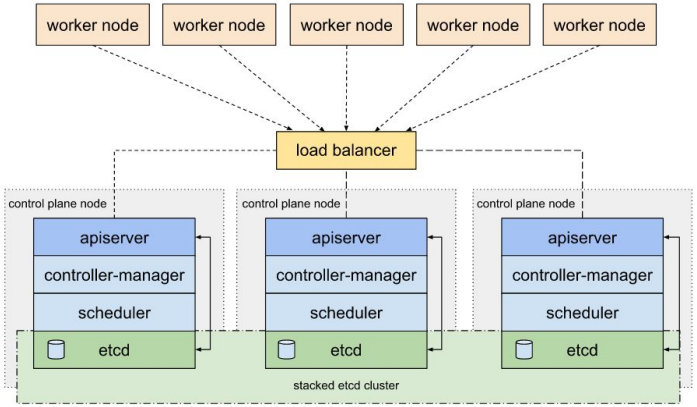
Building and Maintaining entire Kubernetes cluster takes lot of time and resources.



Understanding the Basics

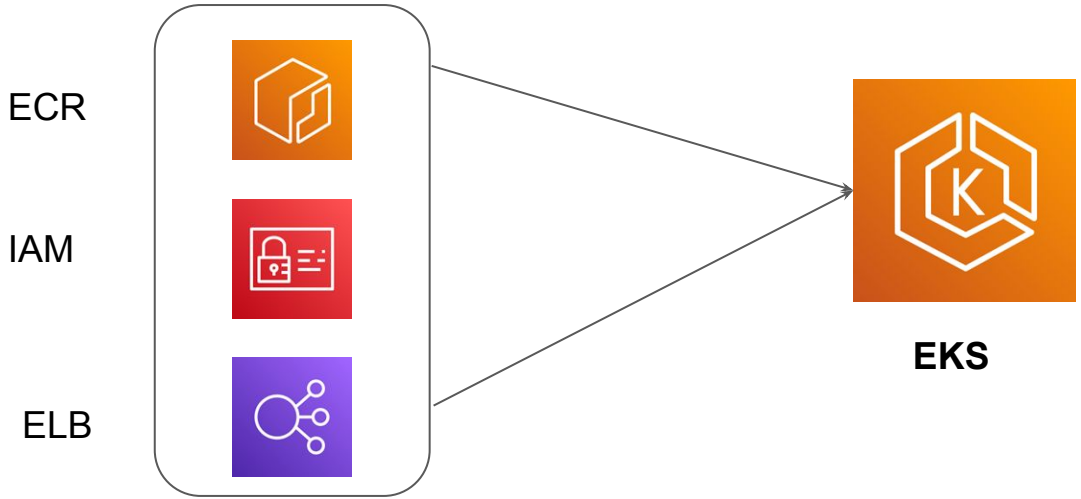
Amazon Elastic Kubernetes Service (Amazon EKS) is a managed service that you can use to run Kubernetes on AWS.





Benefits of EKS

EKS provides tight integration with various other AWS services like ECR, IAM, ELB to provide end to end features for application deployments.

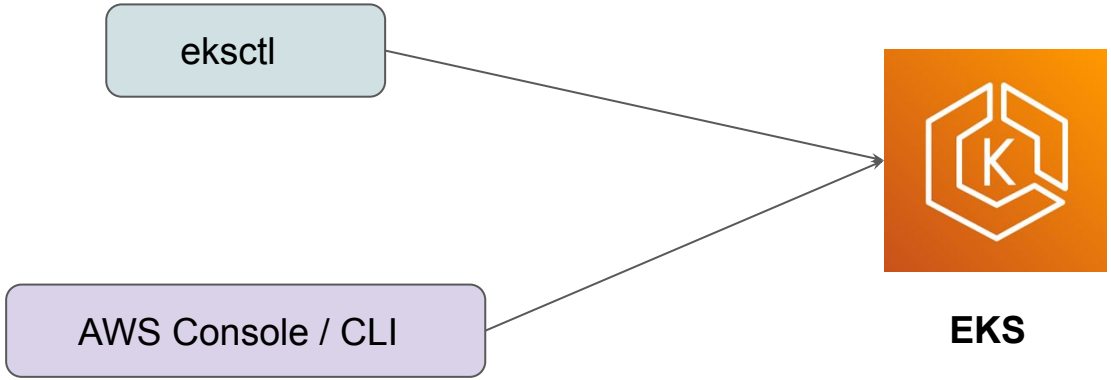


EKS Practical Steps

Let's Create EKS Cluster

Approaches to Create EKS Cluster

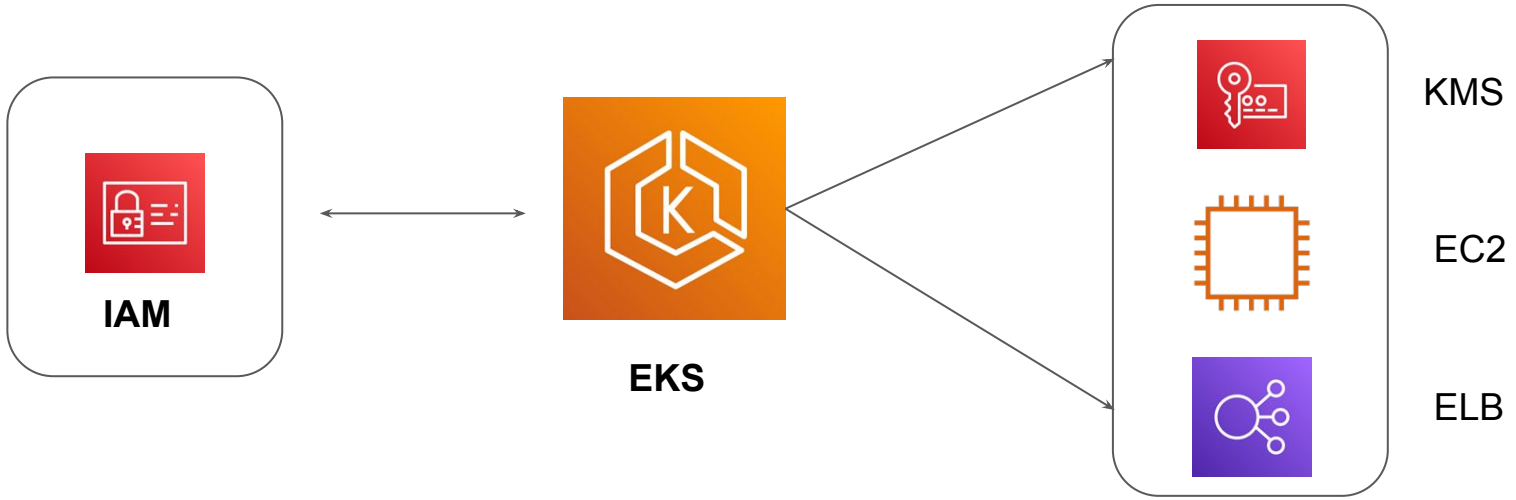
There are two primary ways to create EKS Cluster



Step 1 - Build EKS Cluster

In this step, we build the base EKS Cluster.

Appropriate IAM Role needs to be associated so EKS can manage resource on your behalf.



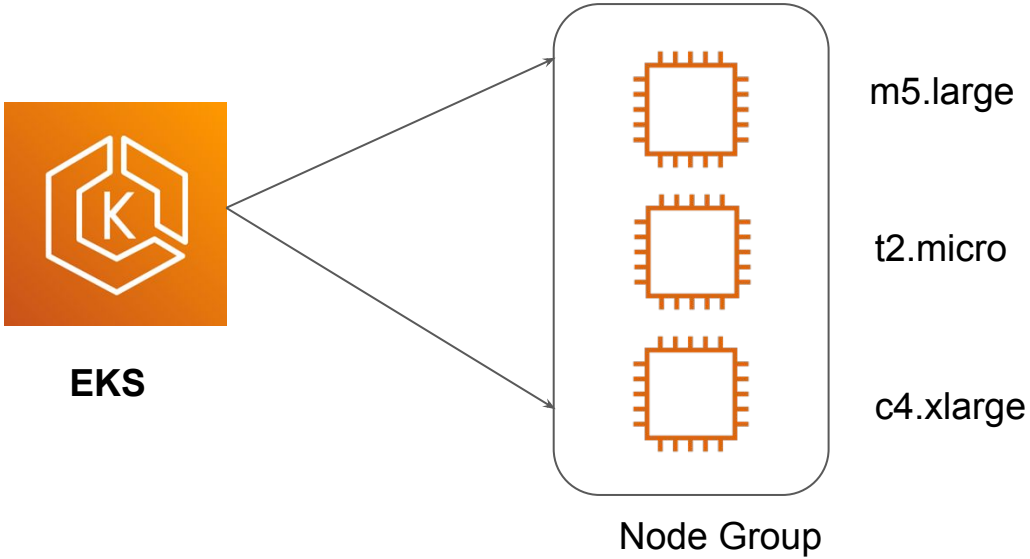
Important Configuration - Building Cluster

Configs	Description
Kubernetes Version	Sets the K8s Version for your cluster.
Cluster Service Role	Allows EKS to manage resources.
VPC	VPC for cluster resources.
Cluster Endpoint Access	Public / Private Access to EKS Cluster.
Networking Add-Ons	To Configure appropriate networking in cluster.
Logging	Enable Logging for K8s Components.

Step 2 - Create Node Group

A node group is a group of EC2 instances that supply compute capacity to your Amazon EKS cluster.

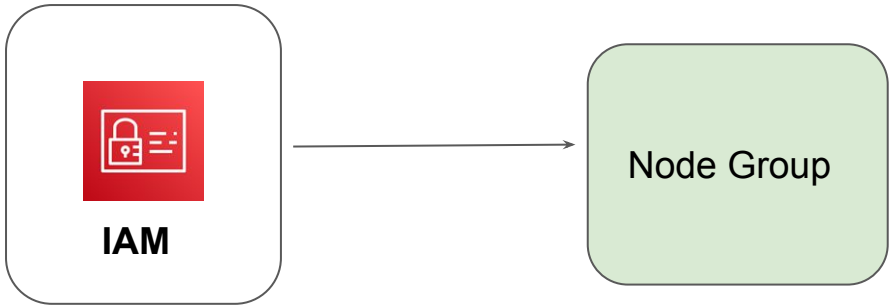
Configuration: AMI ID, Instance Type, Auto-Scaling Configuration, Disk Size.



IAM Role for NodeGroup

An IAM Role needs to be associated with NodeGroup to ensure EC2 instance can perform following operations:

Fetch Images from ECR, Manage Network Interfaces, and others.



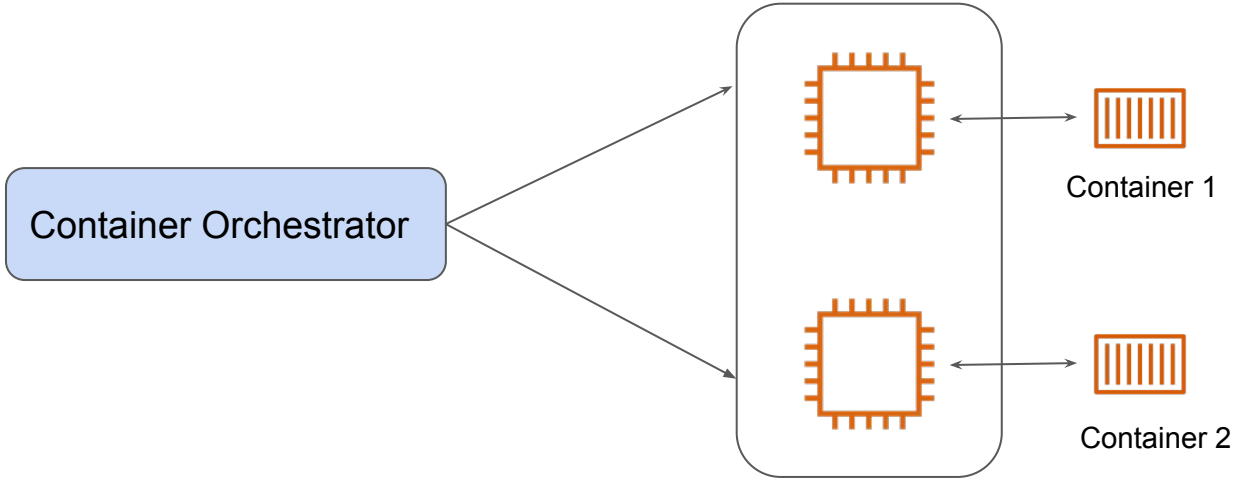
AWS Fargate

Serverless Compute

Basic Approach

In traditional approach, there is a need to create set of EC2 instances where containers can run.

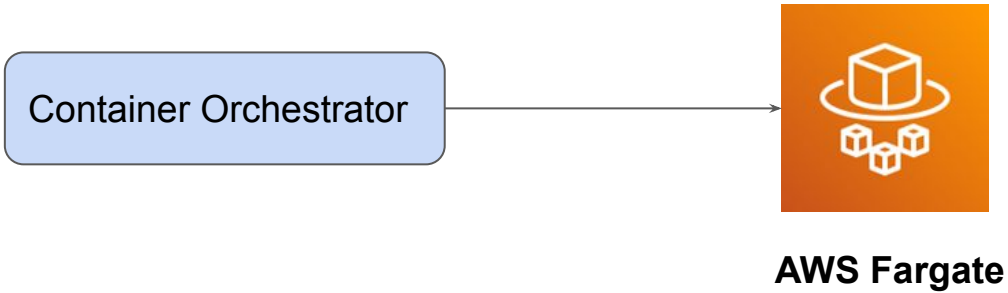
Challenges: Define and Deploy EC2, Security of EC2, Manage EC2



Serverless Approach

In the serverless approach, we do not have to worry about provisioning and managing EC2

AWS Fargate is a serverless, pay-as-you-go compute engine that lets you focus on building applications without managing servers



Migration Strategies

Challenges and Structure

Migration Strategy	Description	Use-Case / Example
Re-Host -	Referred to as a “lift and shift. Move application without any changes.	MySQL to EC2 MySQL
Re-Platform	Referred to as “lift, tinker, and shift.” Make a few cloud optimizations to achieve a tangible benefit.	Move on-premise MySQL to RDS.
Re-Factor/Re-Architect	Re-imagine how the application is architected and developed using cloud-native features.	Migrating to serverless.
Re-Purchase	Move from perpetual licenses to a software-as-a-service model.	Nessus to AWS Inspector
Retire	Remove applications which are no longer needed.	On-Premise FTP server
Retain	Keep applications that are critical for the business but that require major refactoring before they can be migrated.	Old good servers still running.

Migration - Tools and Services

Migration is not always related to servers.

Organizations might still continue to use on-premise and might want to migrate data from old magnetic tapes.

AWS services that helps in migrations:

- AWS Snowball
- Server Migration Service
- Database Migration Service
- Application Discovery Service
- AWS Snowmobile

AWS Snowball Family



Understanding with Use-Case

Organization A has hosted all of it's storage infrastructure in data-center.

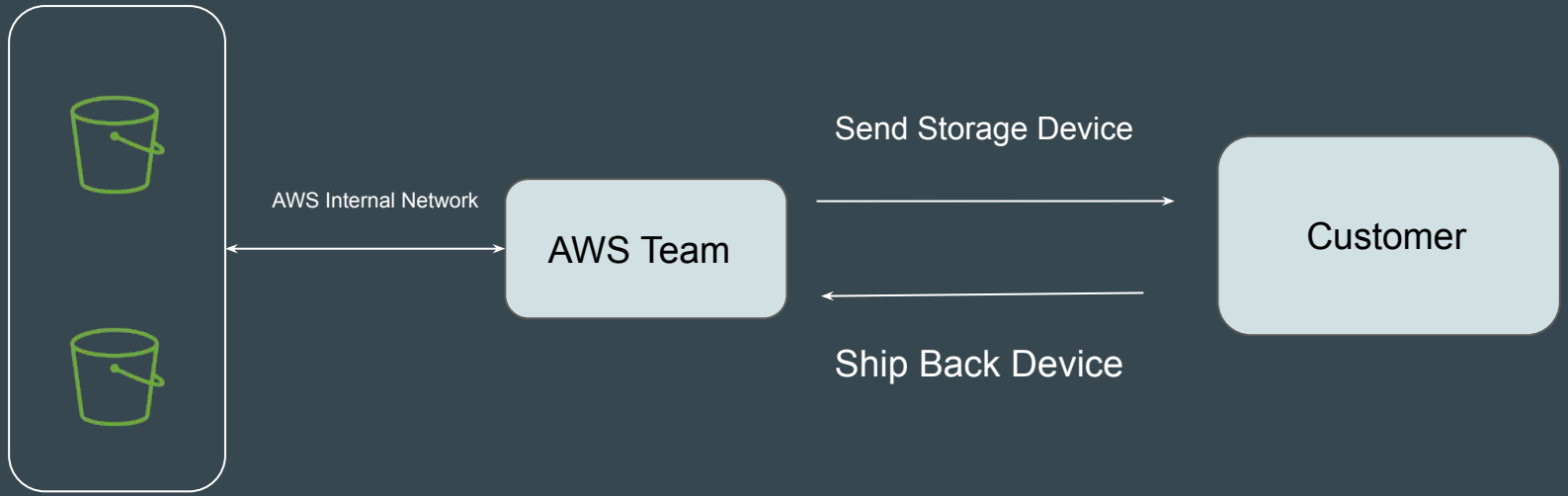
Total Storage: 500 TB.

They have now decided to use S3 due to the benefits that it provides.

Bandwidth	Transfer Time
100 Mbps	510 days.
500 Mbps	101 days
1 Gbps	50 days

AWS Snowball Family

Allows customers to Accelerate moving offline data or remote storage to the cloud



How the Storage Devices Look



Snowball Edge



Snowmobile

Edge Computing Functionality

These devices can also come with edge computing capabilities.

This means, you can run your applications in EC2 instances in the devices so that you can work in edge environments with limited connectivity.

Process data locally (Image/ Video Processing, Machine Learning etc)





Snowcone

AWS Snowcone is a small, rugged, and secure device offering edge computing, data storage, and data transfer on-the-go, in severe environment with little or no connectivity.

Can carry in backpack, drones and others.

8 TB of usable storage



Snowball Edge

AWS Snowball Edge is a type of Snowball device with on-board storage and compute power for select AWS capabilities

Available in Multiple Storage Capacity like 100 TB, 40 TB and others.



Snowball Edge

Device	Description
Snowball Edge Storage Optimized (for data transfer)	This option has a 100 TB (80 TB usable) storage capacity.
Snowball Edge Storage Optimized (with EC2 compute functionality)	This option has up to 80 TB of usable storage space, 24 vCPUs, and 80 GB of memory for compute functionality
Snowball Edge Compute Optimized	Most compute functionality, with 104 vCPUs, 416 GB of memory, and 28 TB of dedicated NVMe SSD for compute instances.
Snowball Edge Compute Optimized with GPU	Identical to the Compute Optimized option, except for an installed GPU

AWS Snowmobile

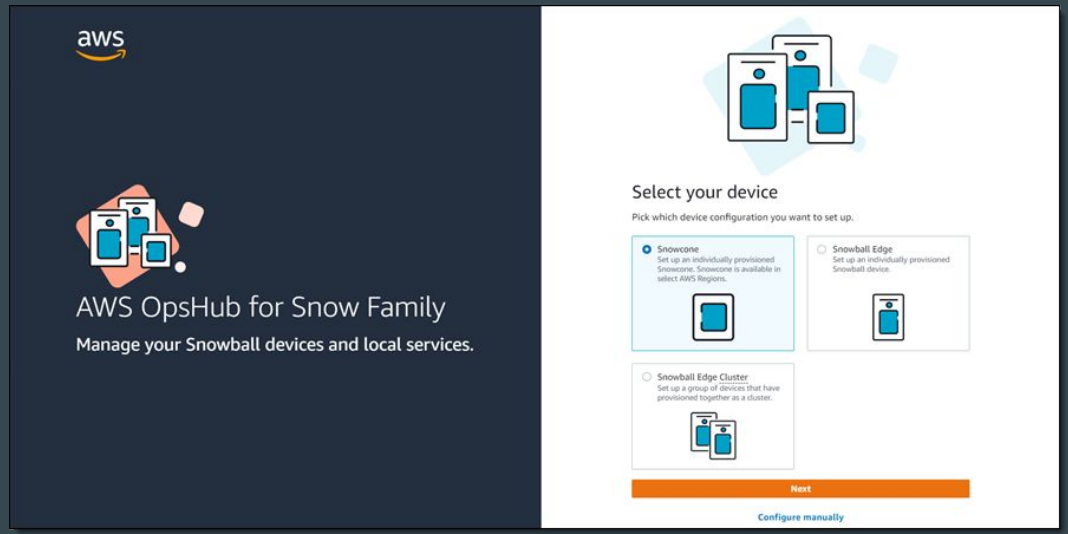
AWS Snowmobile moves extremely large amounts of data to AWS.

Transfer up to 100 PB per Snowmobile, a 45-foot-long ruggedized shipping container pulled by a semi-trailer truck.



AWS OpsHub

AWS OpsHub is a graphical user interface you can use to manage your AWS Snowball devices.

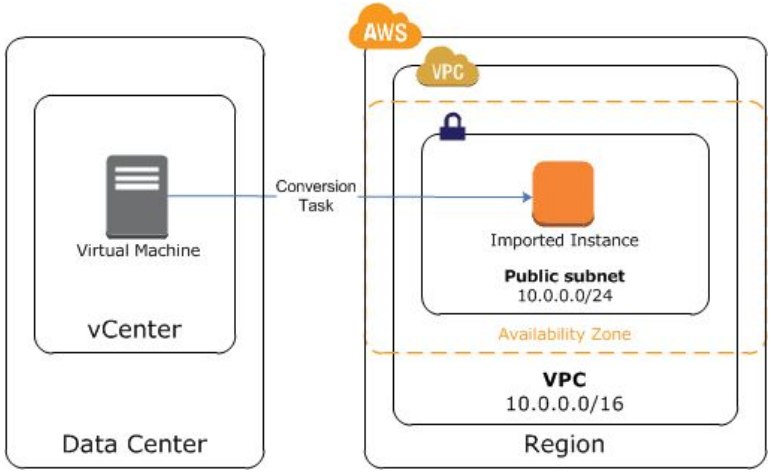


VMWare Migrations

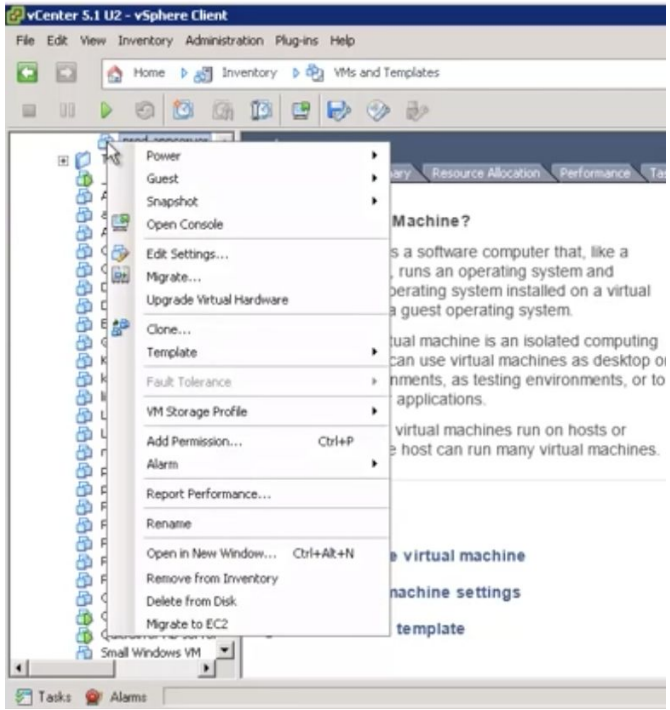
Virtualization & Cloud

Many organizations are migrating

- Many organizations are migrating from their on-premise environments to EC2.
- VMWare was one of the most used solutions for virtualization and private clouds.



Migrate VM to EC2



Choose the appropriate AWS region & Subnets

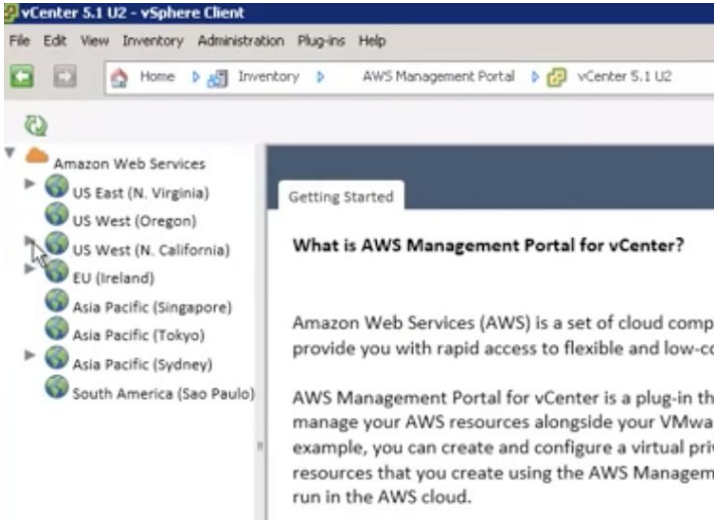


Migration Completed



Important Pointer

AWS Management Portal for vCenter allows customer to manage the AWS resources from their VMWare vCenter dashboard itself.



AWS SMS

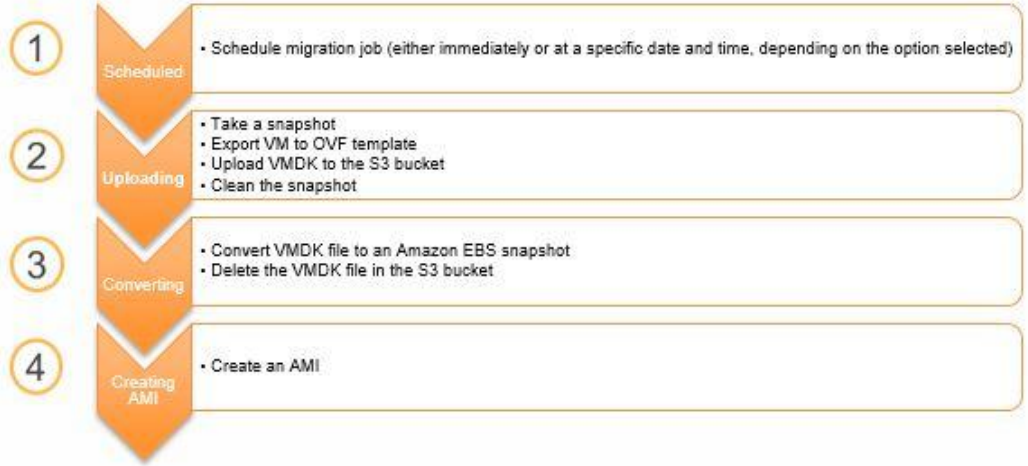
Server Migration Service

Getting Started

AWS Server Migration Service (SMS) is an agentless service which makes it easier and faster for you to migrate thousands of on-premises workloads to AWS.

Supportd Platforms: vSphere and Hyper-V

-



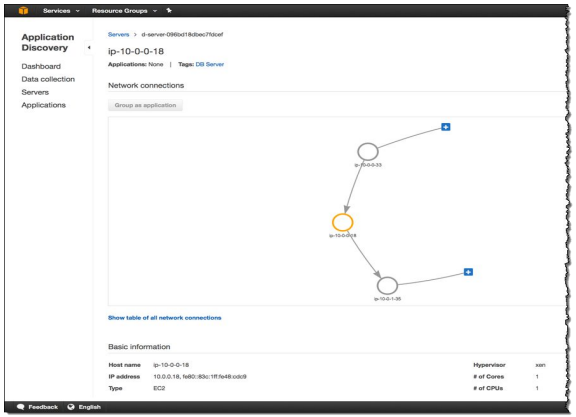
Application Discovery Service

Overview Video

Overview of AWS Application Discovery Service

AWS Application Discovery Service helps enterprise customers plan migration projects by gathering information about their on-premises data centers.

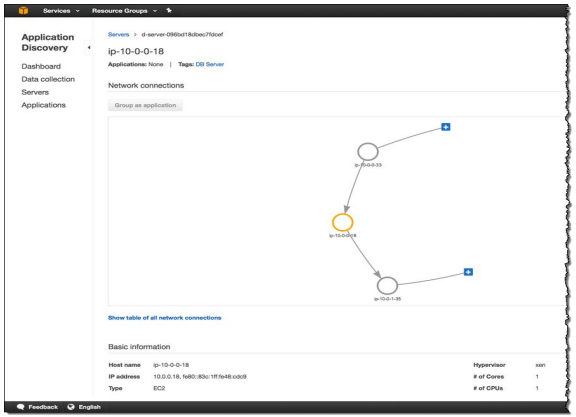
For enterprises which has hundreds to thousands of servers on-premise, getting to know network dependency, right instance types for them can be challenging.



Important Pointers

It supports both agentless discovery and agent-based discovery

- Agentless service needs installation of the discovery connector in VMware vCenter .
- Agent-based service requires agent to be installed in OS.



Database Migration Service

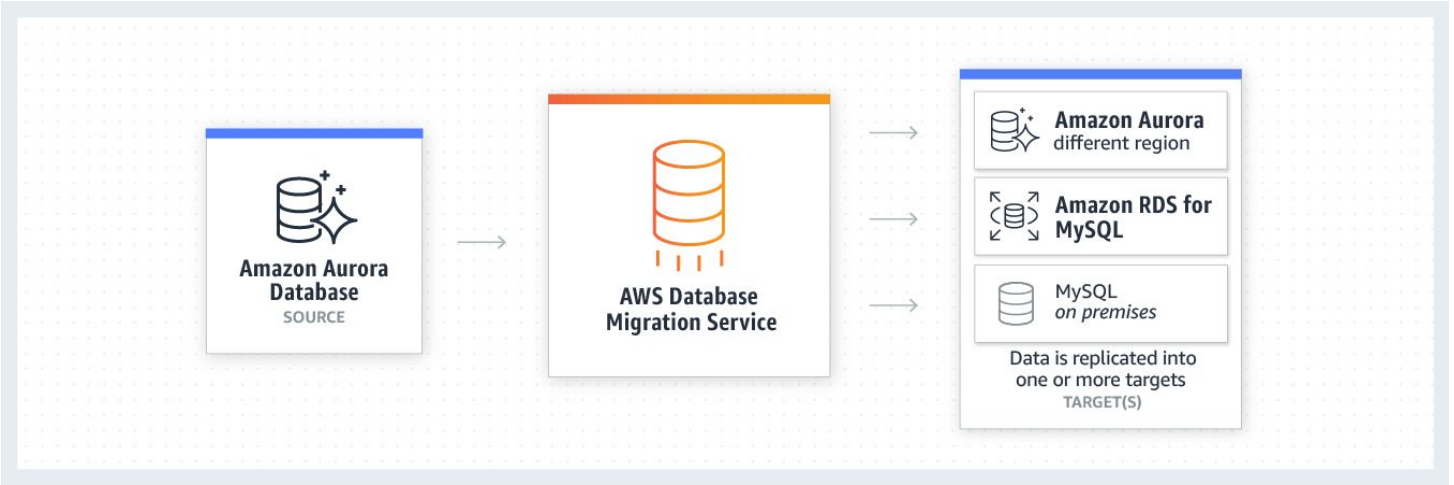
Challenges and Structure

Overview of DMS

AWS Database Migration Service helps you migrate databases to AWS quickly and securely.

The source database remains fully operational during the migration, minimizing downtime to applications that rely on the database.

AWS Database Migration Service supports homogeneous migrations such as Oracle to Oracle, as well as heterogeneous migrations between different database platforms, such as Oracle or Microsoft SQL Server to Amazon Aurora.


Amazon Aurora
Database
SOURCE
AWS Database
Migration Service
Amazon Aurora
different region
Amazon RDS for
MySQL
MySQL
on premises
Data is replicated into
one or more targets
TARGET(S)

AWS Outposts

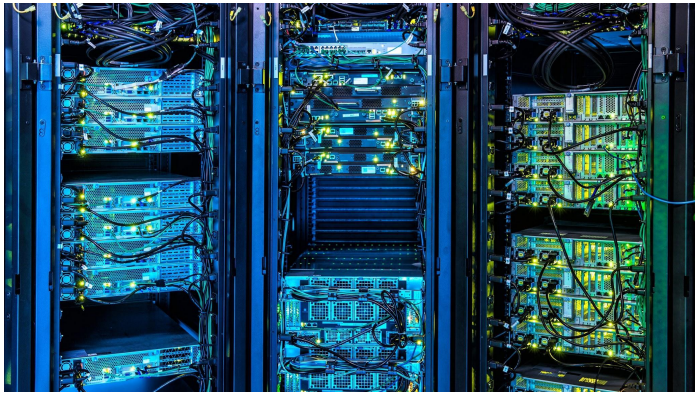
AWS in On-Premise

Cloud is Servers Behind the Scenes

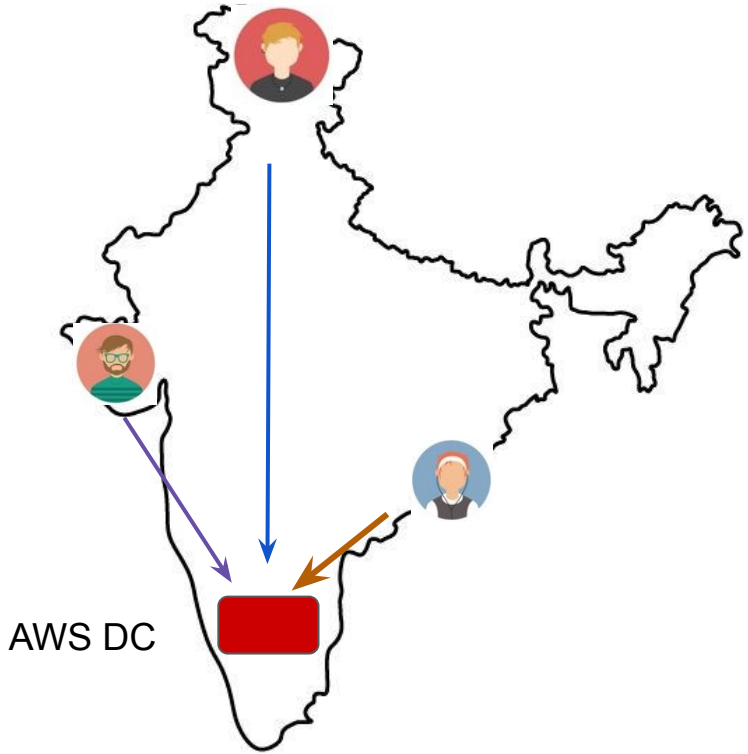
Cloud is basically set of Servers behind the scenes.

These servers reside in the AWS Datacenter.

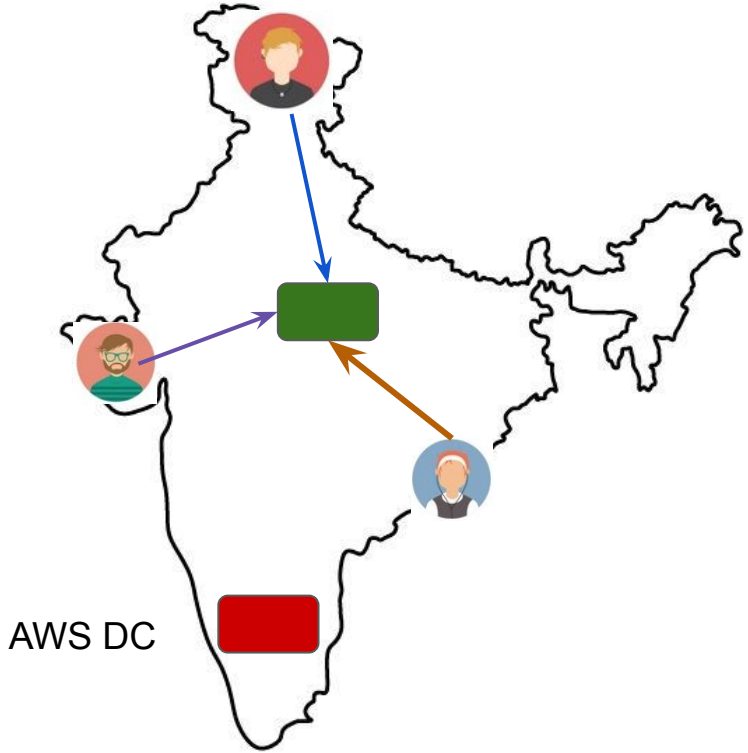
Everything from Power, Cooling, Internet Connectivity, Physical Security is managed by AWS.



Challenge 1: Latency



Possible Solution 1 - On-Premise Servers



Challenges with Hybrid Architecture

1. Different set of API's to manage servers and services.

2. Automation is difficult.

3. Additional learning required.

AWS Outposts

AWS Outposts is a fully managed service that offers the same AWS infrastructure, AWS services, APIs, and tools to virtually any datacenter, co-location space, or on-premises facility.



AWS Side



Customer Side

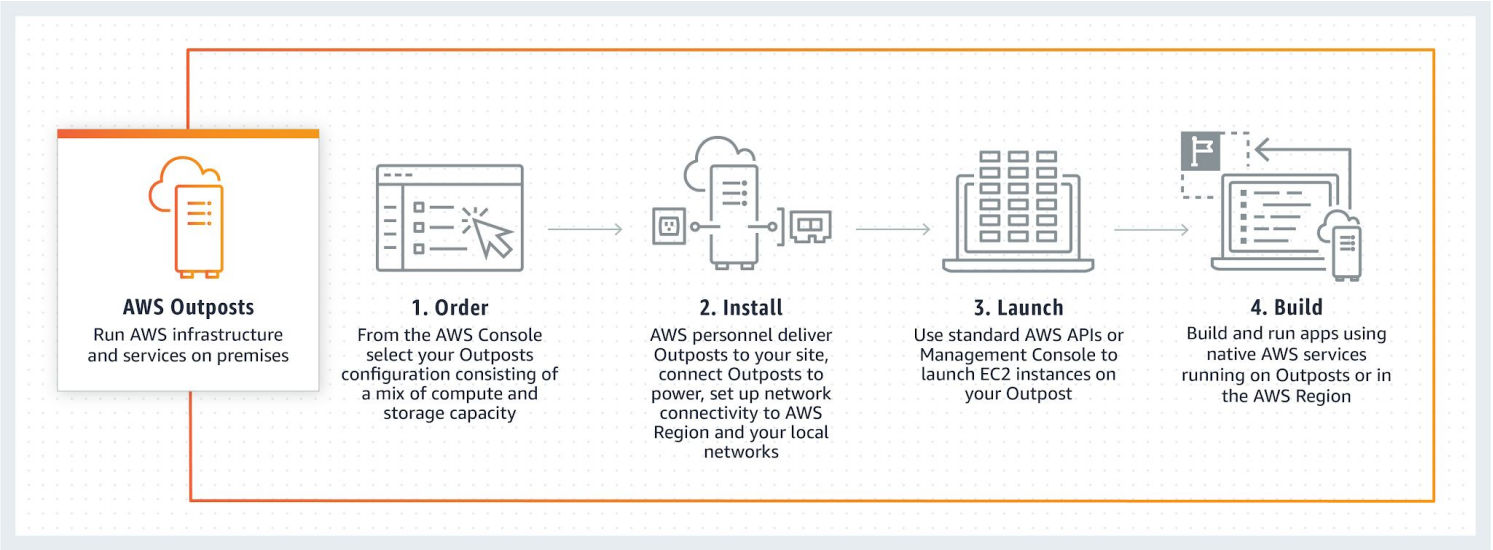
Services that you can Run

With AWS Outposts, we can run wide variety of AWS services locally.

Some of these include:

- Amazon EC2
- Amazon EBS
- S3
- RDS
- EKS
- EMR

Installation Step



Use-Case for AWS Outposts

AWS Outposts can be used for wide-variety of use-cases.

Some of these include:

- Low-Latency Requirements
- Data Residency
- Local Data Processing

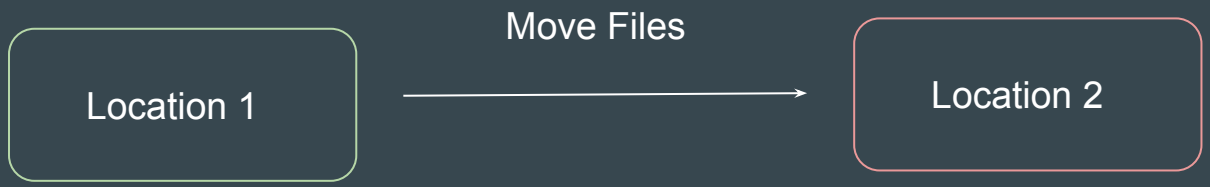
AWS DataSync



Understanding the Basics

AWS DataSync is an online data transfer service that simplifies, automates, and accelerates moving data between storage systems and services.

DataSync provides end-to-end security, including encryption and integrity validation, to help ensure that your data arrives securely, intact, and ready to use.

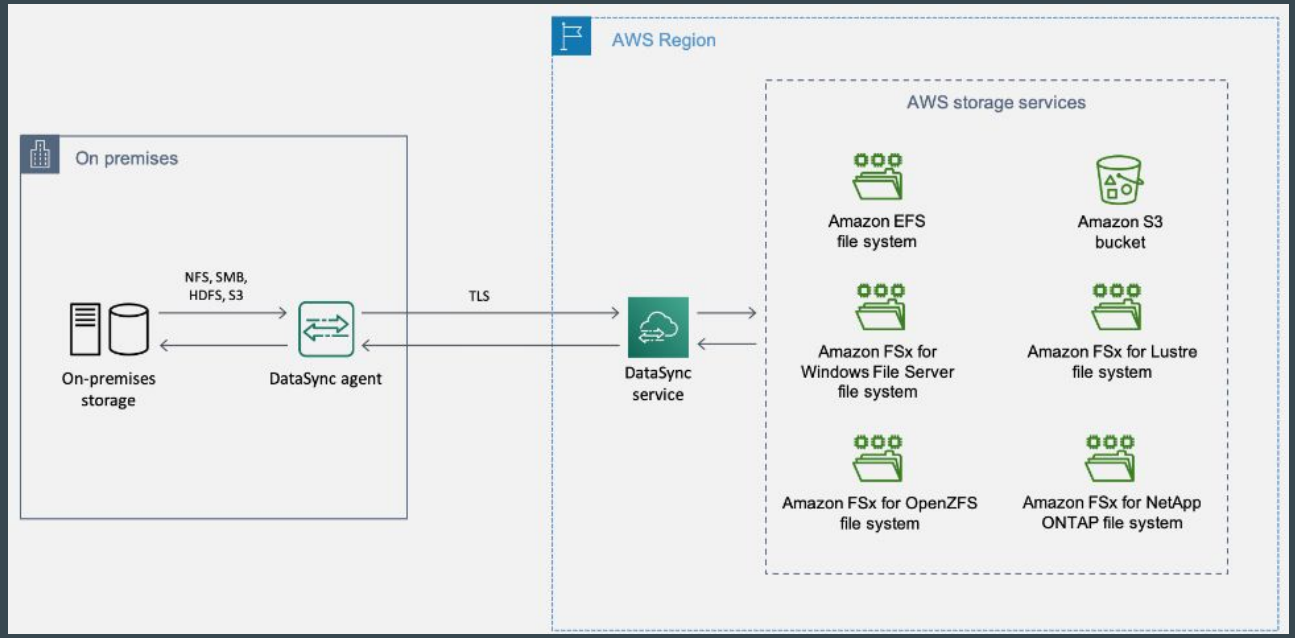


Supported Endpoints

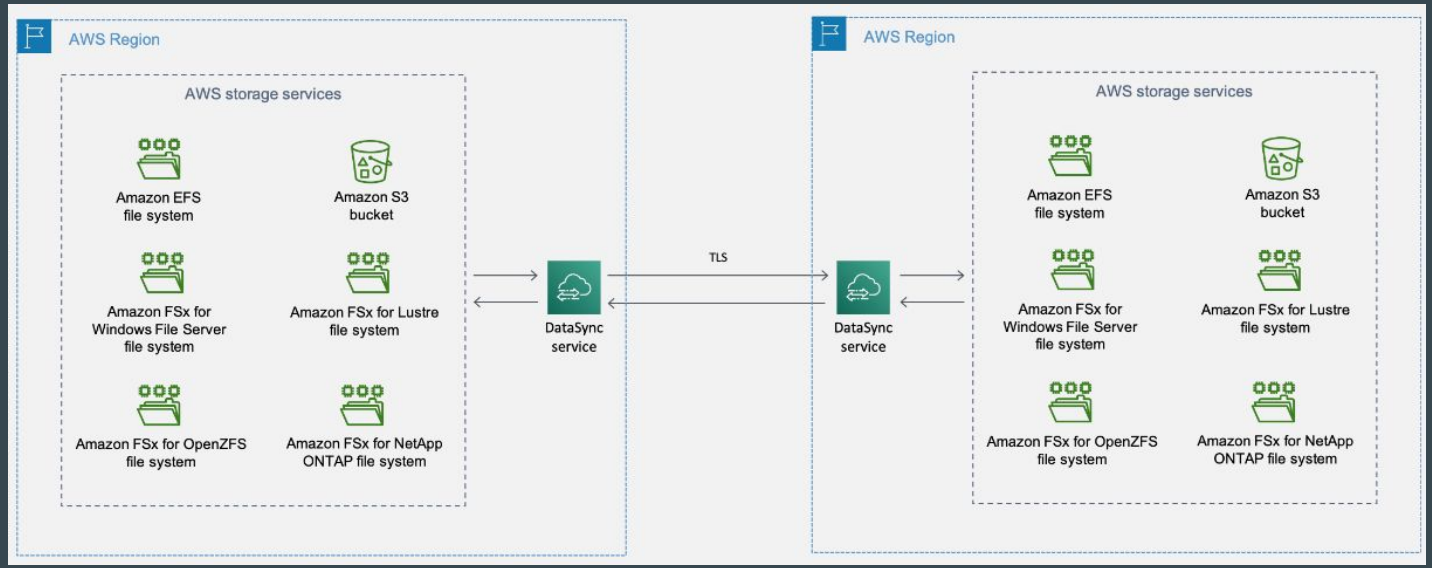
DataSync can copy data to and from:

- Network File System (NFS) file servers
- Server Message Block (SMB) file servers
- Hadoop Distributed File System (HDFS)
- Object storage systems
- Amazon Simple Storage Service (Amazon S3) buckets
- Amazon EFS file systems
- Amazon FSx for Windows File Server file systems
- Amazon FSx for Lustre file systems
- Amazon FSx for OpenZFS file systems
- Amazon FSx for NetApp ONTAP file systems
- AWS Snowcone devices
- Google Cloud Storage buckets
- Azure Files

Transferring between on-premises storage and AWS



Transferring between AWS storage services

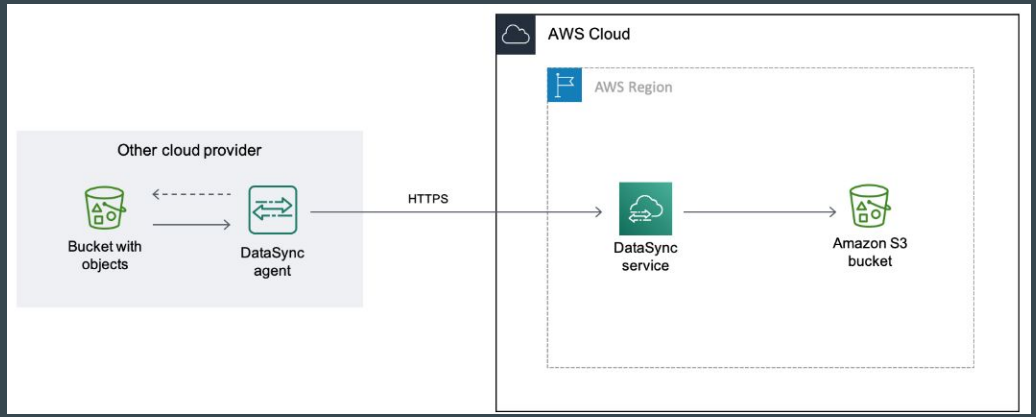


Transferring data from Google Cloud Storage to Amazon S3

1. You deploy a DataSync agent in your Google Cloud environment.
2. The agent reads your Google Cloud Storage bucket

3. Objects from your Google Cloud Storage bucket move securely through TLS 1.2 into the AWS Cloud by using a public endpoint.

4. The DataSync service writes the data to your S3 bucket.



Trusted Advisor

Recommendations are always good

What is Trusted Advisor ?

AWS Trusted Advisor analyzes your AWS environment and provides best practice recommendations in five major categories:



Trusted Advisor Check Categories

Categories	Description
Cost optimization	Recommendations that can potentially save you money.
Performance	Recommendations that can improve the speed and responsiveness of your applications.
Security	Recommendations for security settings that can make your AWS solution more secure.
Fault tolerance	Recommendations that help increase the resiliency of your AWS solution.
Service limits	Checks the usage for your account and whether your account approaches or exceeds the limit for AWS services and resources.

AWS Tags

Meta-Data to AWS Resources

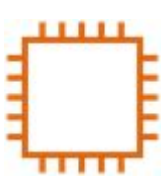
Understanding the Basics

A tag is a label that you assign to an AWS resource.

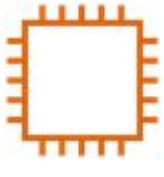
Each tag consists of a key and an optional value, both of which you define.

Let's understand this with a simple use-case:

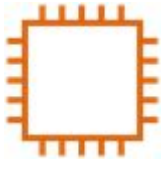
There are three EC2 instances running. How will you identify them?



email-server



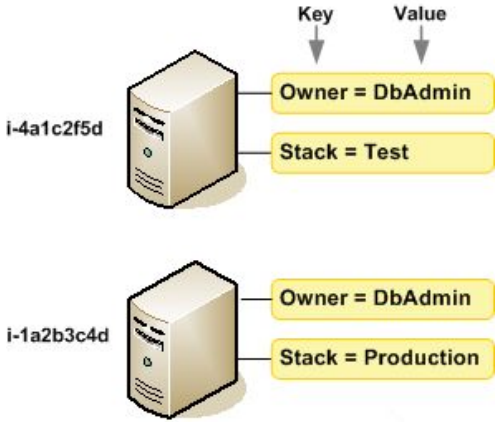
payment-app



compliance-system

Architecture Perspective

A resource can have multiple tags assigned to it.



Important Restrictions

AWS has hundreds of services. Not all of them support tagging.

Maximum number of tags per resource – 50

Tag keys and values are case-sensitive.

.

Important Use-Cases for Tags - Billing

Tags can be integrated in Billing and that allows customers to understand their AWS bill in a granular way.

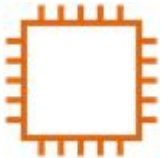
Total Cost	user:Owner	user:Stack	user:Cost Center	user:Application
0.95	DbAdmin	Test	80432	Widget2
0.01	DbAdmin	Test	80432	Widget2
3.84	DbAdmin	Prod	80432	Widget2
6.00	DbAdmin	Test	78925	Widget1
234.63	SysEng	Prod	78925	Widget1
0.73	DbAdmin	Test	78925	Widget1
0.00	DbAdmin	Prod	80432	Portal
2.47	DbAdmin	Prod	78925	Portal

Important Use-Cases for Tags - IAM

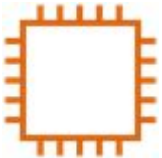
Tags can be used with IAM to control the access to resources.

Example Use-case:

Allow Alice to only delete resources which has a tag of env as development.



env: production



env: development

Follow the Tagging Strategy

It is important to have a consistent and effective tagging strategy.

Example

Alice creates a resource with the tag of key of env and value of production

Matthew creates a resource with the tag of Env and value of Production

Follow the AWS Tagging Strategies document.

Consolidated Billing

Money Optimization

Many AWS Accounts

- Lots of organizations are following approach of having multiple AWS accounts.
- For example, separate AWS account for DEV, Staging, Production.

- It brings quiet lot of benefits which includes **security** as well as **costing**.



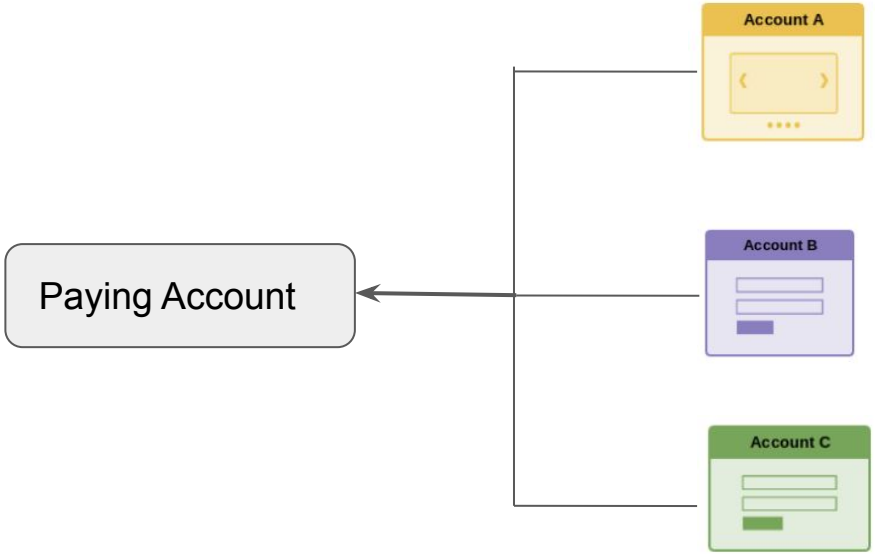
What many startups do ?

A startup has 3 AWS accounts which they use to separate their environment as part of good security best practice. At end of the month, they get 3 separate bills and they pay for each and every account individually.



Consolidated Billing

- We need to link all our AWS accounts with the central paying account through AWS Organizations.



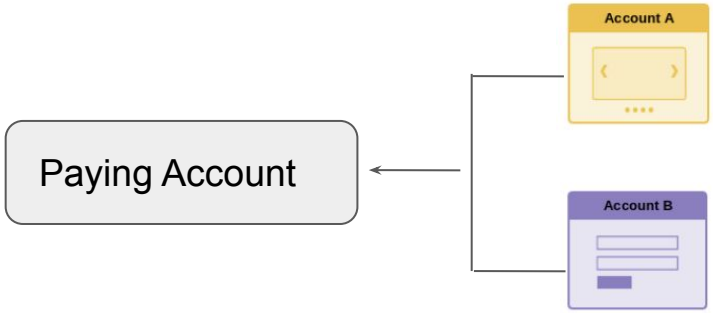
Benefits

There are a lot of benefits of consolidated billing :

- One bill per linked AWS account.
- It becomes quite easy to allocate costing and budgets.
- Benefits of volume pricing discounts.
- Benefits when using reserved instances as well.

	Standard Storage
First 50 TB / month	$0.023 per GB
Next 450 TB / month	$0.022 per GB
Over 500 TB / month	$0.021 per GB

Reserved Instance Benefits



DEV Account: Running 6 on-demand instances

Prod Account: Has 6 RI and running 3 instances

Total Bill : **We will be billed for 6 RI and only 3 on-demand instances.**

Important Pointers & Best Practices

- Paying account should be used for billing purpose only. Do not deploy resource there.
- By default, we can have maximum of 20 linked accounts.

- Even when linked, paying account is individual and does not have access on resources of the other accounts.



TEST

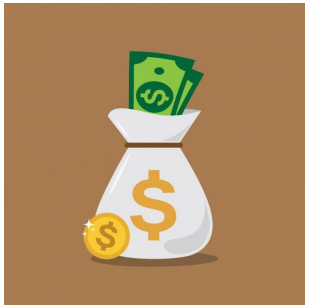
EC2 Pricing

The backbone of internet

Types of Instances

There are 4 ways to pay for an EC2 instance :

- On-demand instance
- Reserved instance
- Spot instances
- Dedicated hosts



On-Demand instance

With On-demand instances, we pay for compute capacity per hour or per second depending on the instances which is being run.

No upfronts payments are needed and we can increase or decrease the capacity whenever it is needed.



Challenges with On-Demand

Monday: 500 customers using 16GB RAM on-demand servers individually.

Wednesday: 10 customers using 16GB RAM on-demand servers individually.

A “Cloud Service Provider” will not have a clear picture on how many servers should the provision. Too high → resources might unused and too low → money loss



Reserved Instance

Reserved Instance provides us with significant discount (upto 75%) compared to on-demand instance pricing.

Reserved instance are assigned to a specific availability zone and provides capacity reservation for AWS EC2 instances.

Example :

You know you will always be running 20 servers of m4.2xlarge type, then buy reserved instances for them.

Reserved Instance - Part 2

Example: m5.4xlarge instance type

Pricing Option	Monthly Cost	Total 3 year cost
On-Demand Instance	$0.096	$20,484
3 year all up-front - Reserved	-	$7589
Savings		~37%

Spot Instance

Spot instances allows us to bind on spare Amazon EC2 computing capacity for up to 90% of the on-demand cost.

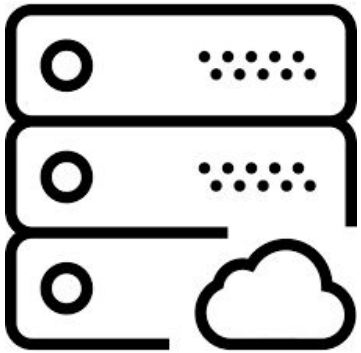
Such instances are recommended for applications which can have flexible start and end times



Dedicated Host

A dedicated host is a physical EC2 server dedicated for your use.

It can be purchased on-demand as well as reserved instance.



Reserved Instances

Money Optimization

Types of Reserved Instances

Type of RIs	Description
Standard RIs	These provide the most significant discount (up to 72% off On-Demand) and are best suited for steady-state usage.
Convertible RIs	These provide a discount (up to 54% off On-Demand) and the capability to change the attributes of the RI
Scheduled RIs	These are available to launch within the time windows you reserve.

RI Types

With convertible RI, we can :

- Convert to new instance family eg R3 to M4 to C4 to T2
- Convert to new operating system eg Windows to Linux
- Convert to new instance price [eg if AWS reduces the public rate for our instance]
- Convert to new instance size [eg : from m4.xlarge to m4.2xlarge]
- Convert tenancy [eg dedicated instance to default]
- Convert to different payment option [no upfront to partial upfront]

Reservation Term

Reservation Term	Description
No Upfront	No upfront required. Lower discount rate compared to others.
Partial Upfront	You make a low upfront payment and are then charged a discounted hourly rate for the instance for the duration of the Reserved Instance term.
All Upfront	You pay for the entire Reserved Instance term with one upfront payment. Provides the largest discount.

Regional vs Zonal RIs

	Regional RI	Zonal RI
Ability to Reserve Capacity	No Reservation in Capacity.	Capacity Reserved in the specific Availability Zone.
Availability Zone Flexibility	The Reserved Instance discount applies to instance usage in any Availability Zone in the specified Region.	Reserved Instance discount applies to instance usage in the specified Availability Zone only.
Instance size flexibility	The Reserved Instance discount applies to instance usage within the instance family, regardless of size.	No instance size flexibility—the Reserved Instance discount applies to instance usage for the specified instance type and size only.

Scenario

Scenario 1 :

Customer has following instances running:

- 2 x m4.large instance running in us-east-1a and us-east-1b region.
- 2 x t2.large instance running across us-east-1b and us-east-1c region

Customer has following RI:

- 2 x m4.large, default tenancy, us-east-1b region (zonal RI)
- 2 x t2.large, default tenancy, us-east-1 regional RI

Additional pay : 1 x m4.large instance will be charged at the on-demand rate.

On-Demand Capacity Reservations

Reserving EC2 Capacity

Understanding the Challenge

With On-Demand Instance, there are chances that new instance might not launch due to insufficient capacity error.

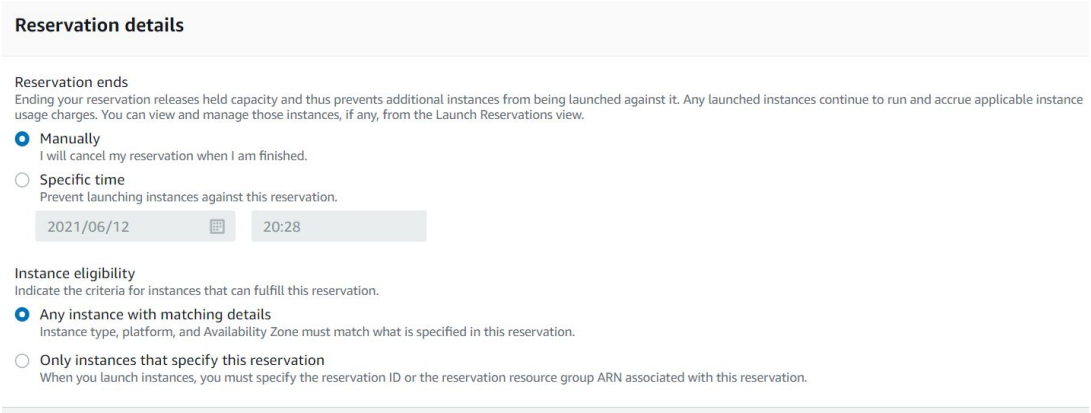
Typical Solution → Go with Reserved Instances (1 year, 3 year term)



Benefits of On-Demand Capacity Reservation

On-Demand Capacity Reservations enable you to reserve compute capacity for your Amazon EC2 instances in a specific Availability Zone for any duration

You can create Capacity Reservations at any time, without entering into a one-year or three-year term commitment, and the capacity is available immediately.



On-Demand Capacity Reservation with RI

You can combine Capacity Reservations with Regional Reserved Instances to receive a discount.

	On-Demand Capacity Reservation	Regional RI
Term	No commitment required. Can be created and canceled as needed.	Requires a fixed one-year or three-year commitment
Capacity benefit	Capacity reserved in a specific Availability Zone.	No capacity reserved.
Billing discount	No billing discount.	Provides a billing discount.

Pricing

When the Capacity Reservation enters the active state, you are charged the equivalent On-Demand rate whether you run instances in the reserved capacity or not.

If you do not use the reservation, this shows up as unused reservation on your EC2 bill.

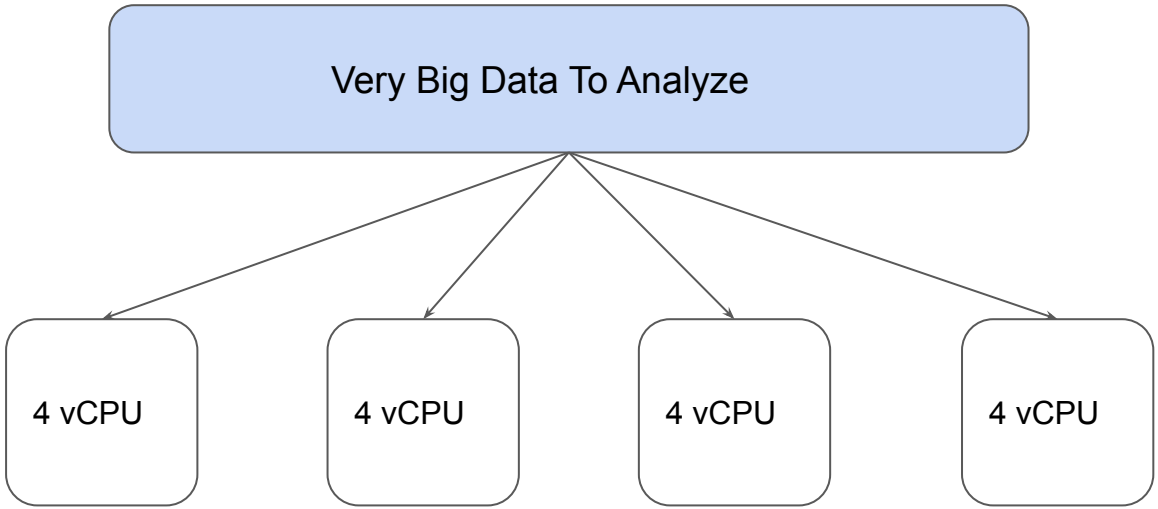
For example, if you create a Capacity Reservation for 20 m4.large Linux instances and run 15 m4.large Linux instances in the same Availability Zone, you will be charged for 15 active instances and for 5 unused instances in the reservation.

EC2 Fleet

Launching EC2 based on Requirements

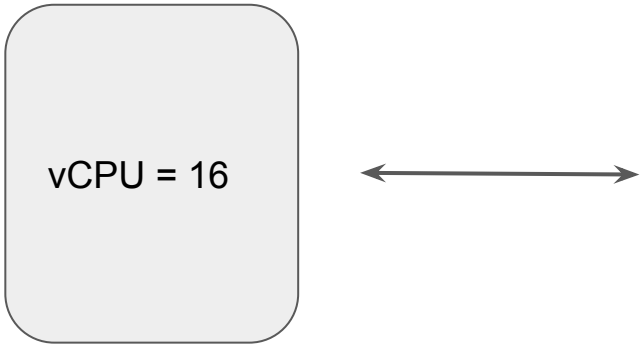
Original Feature

EC2 Fleet allows users to launch a fleet of Spot Instances that spans EC2 instance types and Availability Zones without having to write custom code to discover capacity or monitor prices.



Understanding With A Use-Case

Let us assume, to perform the analysis, you require in total of 16 vCPU.



Instances Types for 16 vCPU
1 x c5.4xlarge
2 x c5.2xlarge
4 x c5.xlarge

Setting Weight

You can assign a set of weighted capacity to a set of EC2 instance Types.

Instances Types for 16 vCPU
1 x c5.4xlarge
2 x c5.2xlarge
4 x c5.xlarge



Instance Type	Weight
c5.4xlarge	16
c5.2xlarge	8
c5.xlarge	4

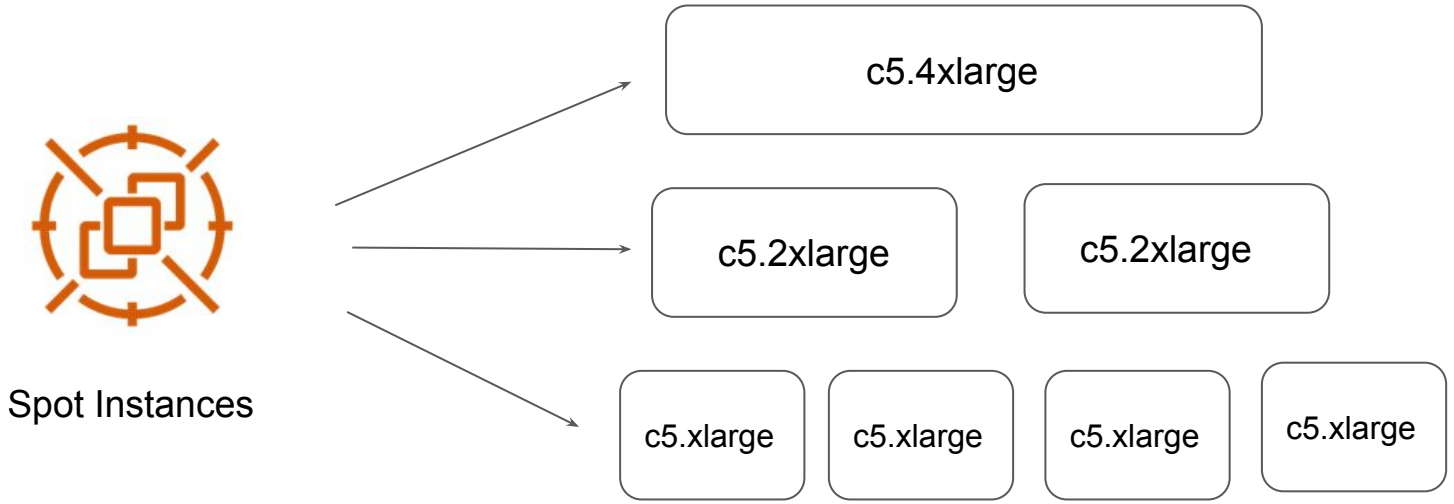
Setting Weight

```
"Overrides": [
{
  "InstanceType": "c5.4xlarge",
  "WeightedCapacity": 16,
},
{
  "InstanceType": "c5.2xlarge",
  "WeightedCapacity": 8,
},
{
  "InstanceType": "c5.xlarge",
  "WeightedCapacity": 4,
},
]
```

Target Capacity = 16

Setting Target Capacity

EC2 Fleet will select the most cost effective combination of instance types and Availability Zones (both specified in the template) using the current prices for the Spot Instances and public prices for the On-Demand Instances



Overview of EC2 Fleet - New

The EC2 Fleet attempts to launch the number of instances that are required to meet the target capacity that you specify in the fleet request.

The fleet can comprise only On-Demand Instances, only Spot Instances, or a combination of both On-Demand Instances and Spot Instances.

```
"TargetCapacitySpecification": {
"TotalTargetCapacity": 16,
"OnDemandTargetCapacity": 4,
"SpotTargetCapacity": 8,
"DefaultTargetCapacityType": "Spot"
}
```

Interpret the Requirements

I want total of 16 vCPUs.
8 vCPU should be fulfilled using On-Demand instance type.
8 vCPU should be fulfilled using Spot instance type.

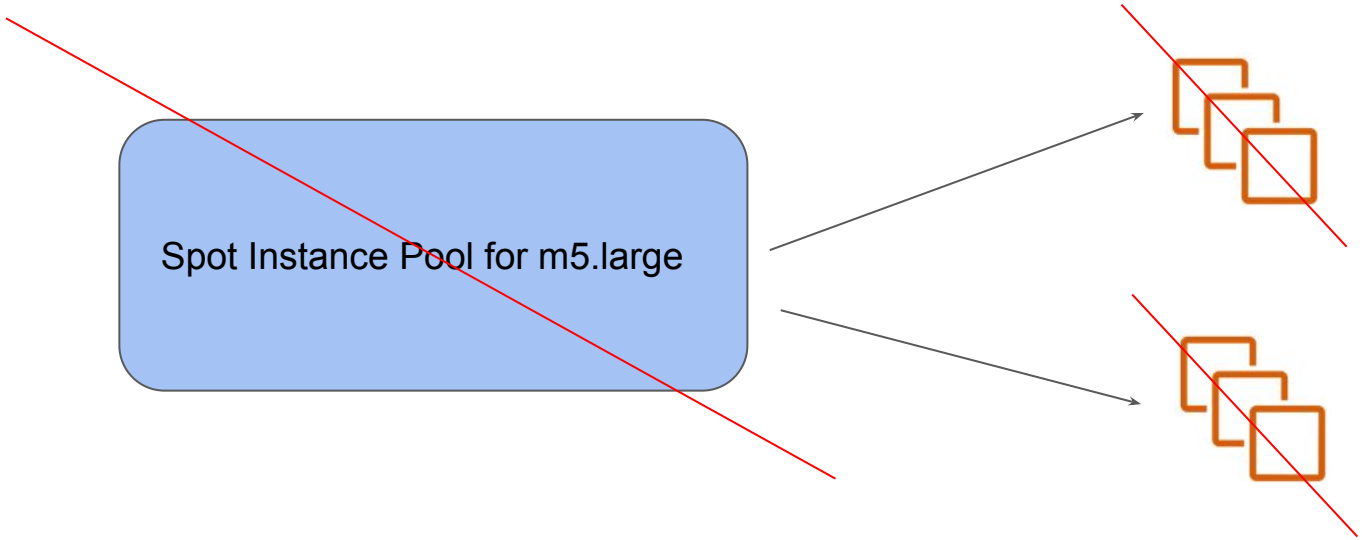
```
"TargetCapacitySpecification": {
"TotalTargetCapacity": 16,
"OnDemandTargetCapacity": 8,
"SpotTargetCapacity": 8,
"DefaultTargetCapacityType": "Spot"
}
```

Allocation Strategy for Spot Instances

Optimizing Costs

Understanding Spot Instance Pool

A Spot instance pool is a set of unused EC2 instances with the same instance type and size (for example, m5.large), availability zone (AZ), in the same region



Allocation Strategy

Depending on your requirements and use-case, there are three primary allocation strategy for Spot instances.

Allocation Strategy	Description
lowest-price	The Spot Instances come from the Spot capacity pool with the lowest price. This is the default strategy.
diversified	The Spot Instances are distributed across all Spot capacity pools.
capacity-optimized	Provisions Spot Instances from the most-available Spot Instance pools by analyzing capacity metrics.

Preference - lowest-price

Choose the lowest-price allocation strategy if:

If your fleet is small or runs for a short time, the probability that your Spot Instances will be interrupted is low, even with all of the instances in a single Spot capacity pool.

Therefore, the lowest-price strategy is likely to meet your needs while providing the lowest cost.

Since the price constantly changes, the existing instances in ASG can be terminated and be replaced by new, cheaper once thus potentially disrupting your service at a higher rate.

Preference - diversified

If your fleet is large or runs for a long time, you can improve the availability of your fleet by distributing the Spot Instances across multiple pools using the diversified strategy.

For example, if your EC2 Fleet specifies 10 pools and a target capacity of 100 instances, the fleet launches 10 Spot Instances in each pool.

If the Spot price for one pool exceeds your maximum price for this pool, only 10% of your fleet is affected.

Preference - capacity-optimized

If your fleet runs workloads that may have a higher cost of interruption associated with restarting work and checkpointing, then use the capacity-optimized strategy

This strategy does not look at the prices of the instance types in each pool configure but instead looks for the optimal capacity volume and chooses those instances to run your service on.

While the overall hourly cost of capacity-optimized allocation strategy might be slightly higher, the possibility of having fewer interruptions can lower the overall cost of your workload.

EC2 Tenancy

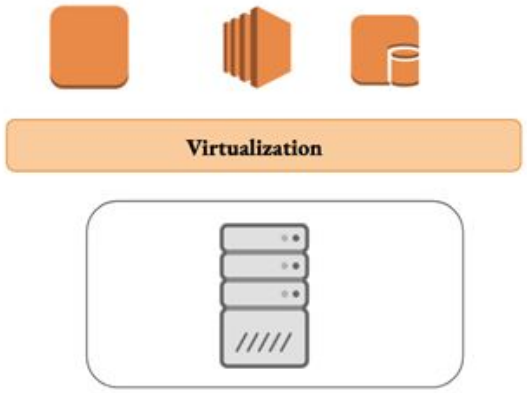
Understanding the EC2 Tenancy

Every EC2 instance that we launch in the VPC has a specific tenancy attribute associated with it. There are three tenancy attributes which are available:

Tenancy Attribute	Description
Shared	The EC2 instance runs on shared hardware.
Dedicated	EC2 instance runs on hardware which will only be shared between same account AWS instances.
Hosts	Instance runs on dedicated hosts with very granular level of hardware access.

Shared Tenancy

In this approach, your EC2 instance is launched on the shared hardware where EC2 instances of other customers also run.

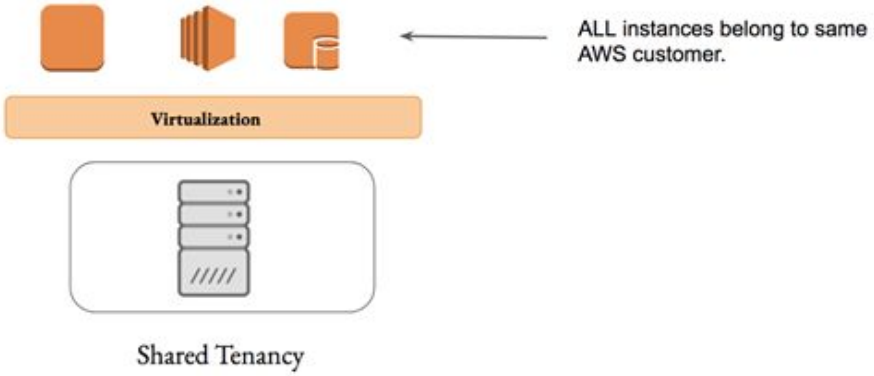


Shared Tenancy

Dedicated Instance

Dedicated Instances are EC2 instances that run on the hardware which is dedicated to a single customer.

Dedicated instances may share the hardware with other EC2 instances that belongs to the same AWS accounts.



Dedicated Hosts

Dedicated Host is a physical server that allows us to use our existing per-socket, per-core or even per-VM based software licenses which includes Windows Server, SUSE, and various others.

With dedicated hosts, we can use the same physical server over the time, even if the instance is stopped and started.

EBS Volume Types



Performance Metrics in Storage Device

Storage Device is a piece of equipment on which information can be stored

Common disk performance metrics are :

- Input / Output operations per second (IOPS)
- Throughput (MB/s or MiB/s)





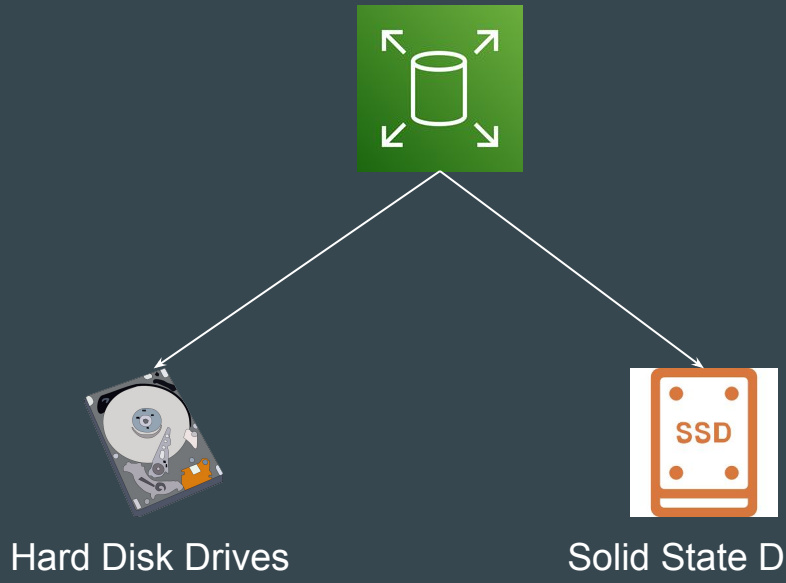
Basic Metrics Information

IOPS is a count of the read/write operations per second

Throughput is the actual measurement of read/write bits per second that are transferred over a network

EBS Volume Types

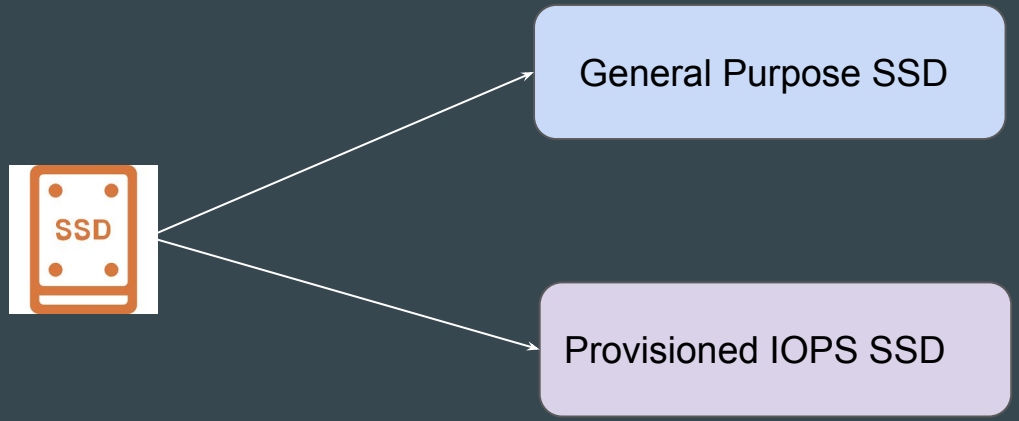
EBS provides different volume types which differs in performance and price.



Solid State Drives (SSD)

Optimized for transactional workloads involving frequent read/write operations with small I/O size, where the dominant performance attribute is IOPS.

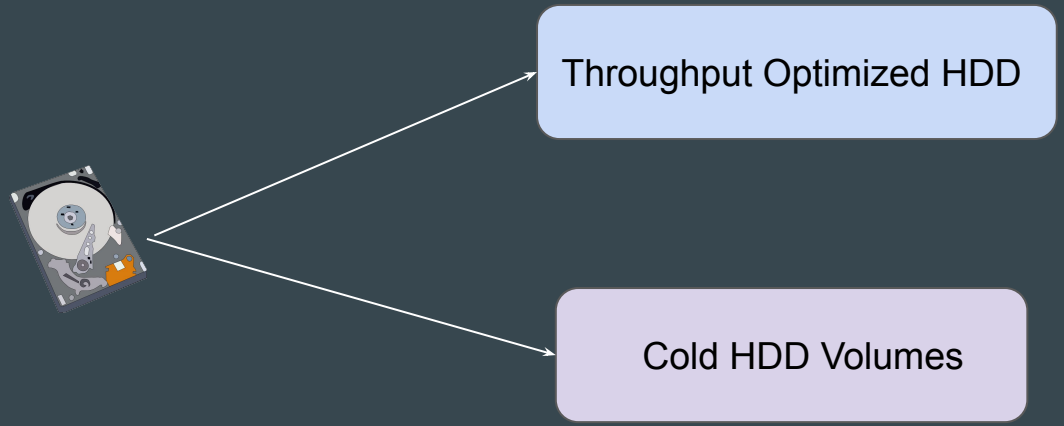
SSD-backed volume types include:



Hard Disk Drives (HDD)

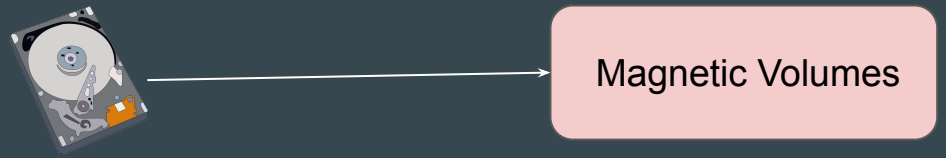
Optimized for large streaming workloads where the dominant performance attribute is throughput.

HDD-backed volume types include:



Previous Generation

Hard disk drives that you can use for workloads with small datasets where data is accessed infrequently and performance is not of primary importance.



General Purpose SSD

Characteristic	gp3	gp2
Baseline IOPS	Baseline of 3000 IOPS	3 IOPS/GiB (minimum 100 IOPS) to a maximum of 16,000 IOPS
Baseline Throughput	125 MiB/s Maximum throughput = 1000 MiB/s	Throughput between 128 MiB/s and 250 MiB/s, depending on the volume size.
Generation	Newer	Older

gp3 offers SSD-performance at a 20% lower cost per GB than gp2 volumes.

Provisioned IOPS

Highest performance SSD volume designed for mission critical application workloads.

Characteristic	io2 Block Express	io2	io1
Durability	99.999% durability (0.001% annual failure rate)	99.999% durability (0.001% annual failure rate)	99.8% - 99.9% durability (0.1% - 0.2% annual failure rate)
Use cases	Workloads that require: Sub-millisecond latency More than 64,000 IOPS or 1,000 MiB/s of throughput	Workloads that require sustained IOPS performance or more than 16,000 IOPS	Workloads that require sustained IOPS performance or more than 16,000 IOPS
Max IOPS per volume	256,000	64,000 †	64,000 †
Max throughput per volume	4,000 MiB/s	1,000 MiB/s †	1,000 MiB/s †

Hard Disk Drives

Characteristic	Throughput Optimized HDD	Cold HDD
Volume type	st1	sc1
Max IOPS per volume	500	250
Max throughput per volume	500 MiB/s	250 MiB/s
Use cases	Big data, Data Warehouse	Throughput-oriented storage for data that is infrequently accessed Scenarios where the lowest storage cost is important

Sample Question

Medium Corp is an E-Commerce organization and you have been assigned responsibility related to performance optimization of servers. They have a critical database server which receives lot of connections and they tried increasing RAM and CPU but still it is slow. What type of EBS volume type will you suggest ?

- General Purpose SSD
- Throughput Optimized
- Provisioned IOPS
- Cold HDD

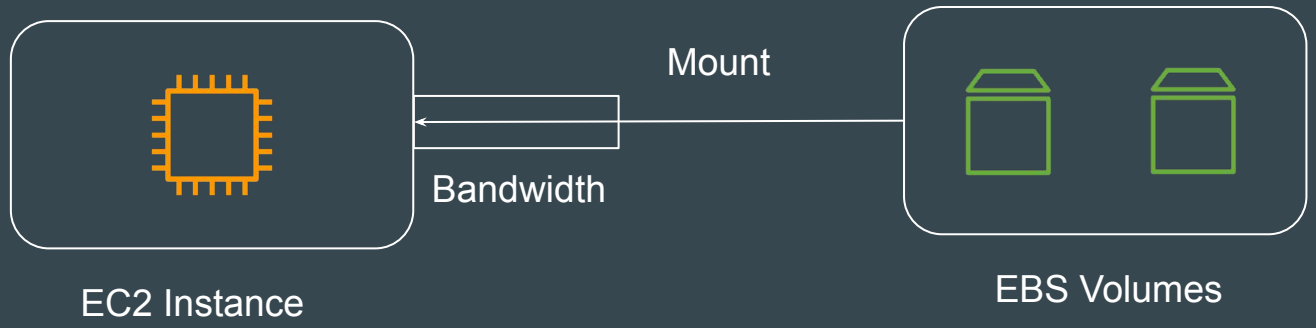
EBS Optimized Instances



Understanding the Basics

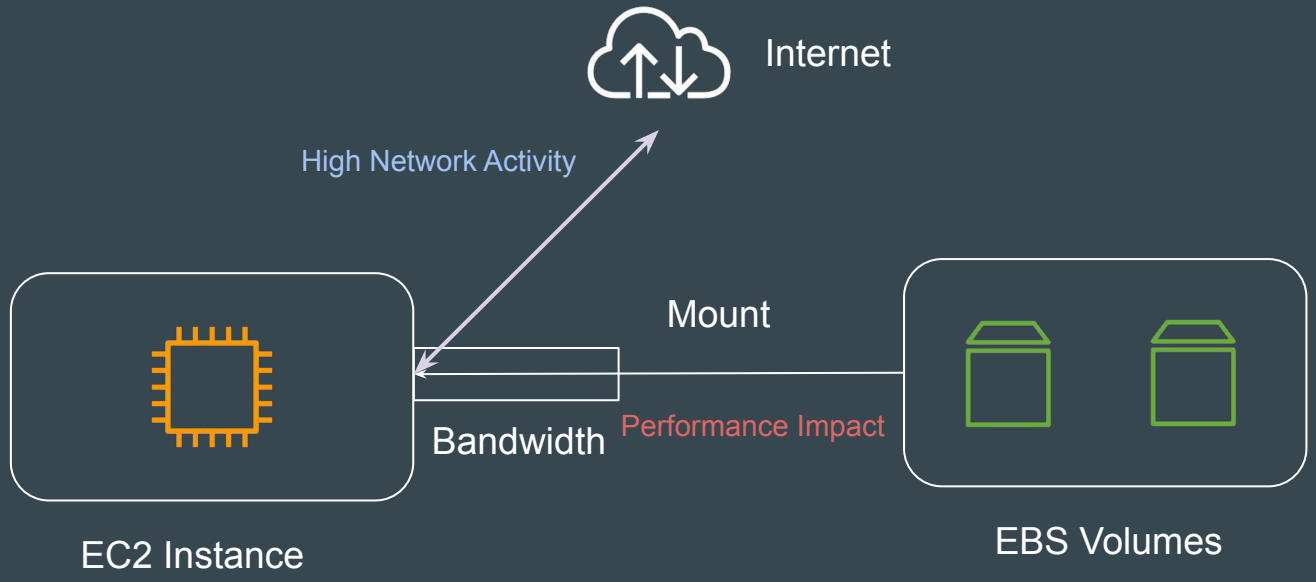
The available network bandwidth of an instance depends on the number of vCPUs that it has.

EBS volumes are mounted to the E2 instance via Network.



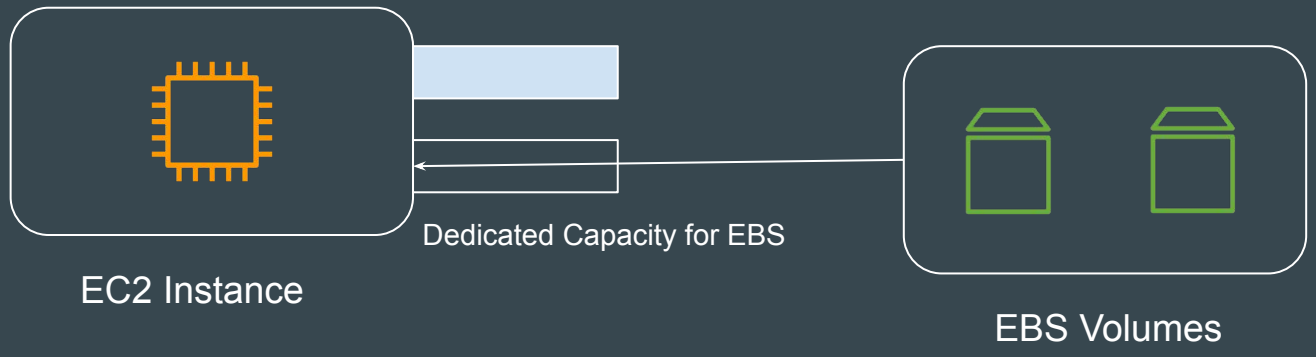
Understanding the Challenge

If EC2 instance is using the available bandwidth and has high Network I/O, it can impact the overall performance at EBS level.



EBS Optimized EC2 Instances

An Amazon EBS–optimized instance uses an optimized configuration stack and provides additional, dedicated capacity for Amazon EBS I/O.



Supported Instance Types

Not ALL instance types support EBS Optimization.

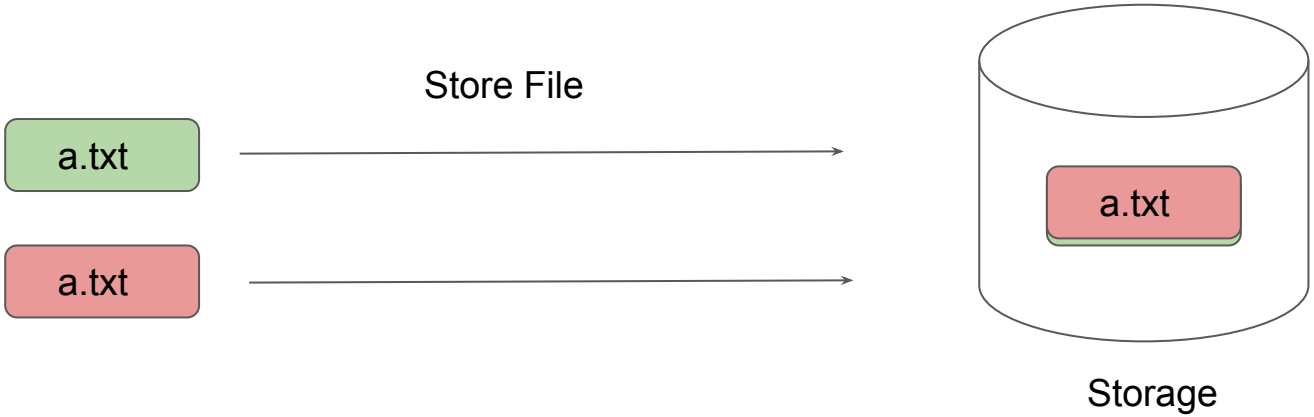
Some instances types have EBS Optimization enabled by default.

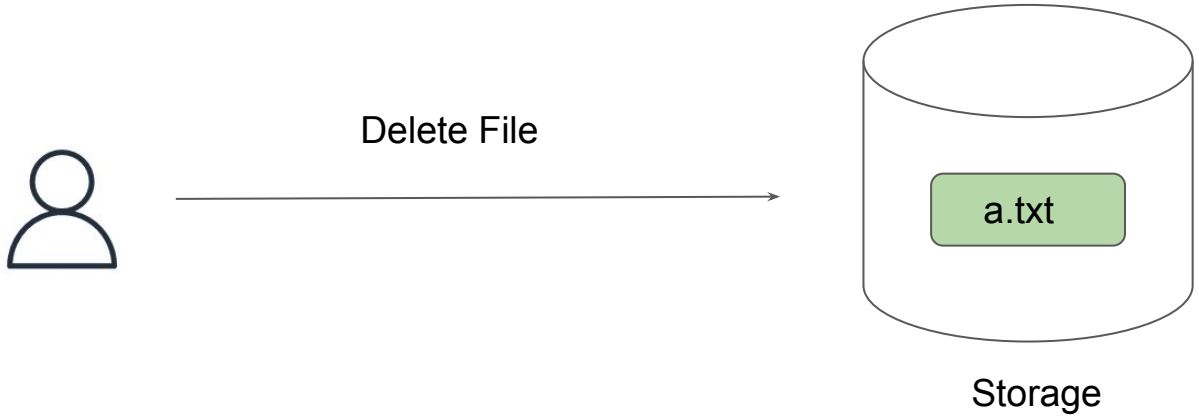
For certain instance types, you have to explicitly enable EBS Optimization.



S3 Versioning

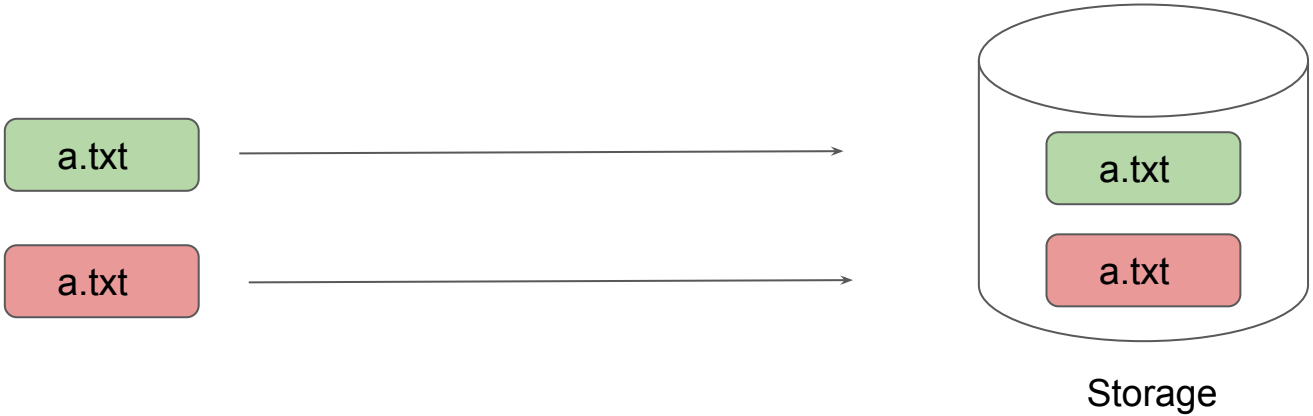
Versioning in Object Storage


Store File
a.txt
a.txt
a.txt
Storage


Delete File
a.txt
Storage

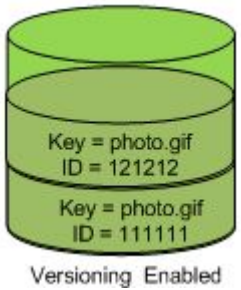
Versioning allows users to keep multiple variants of an object in the same S3 bucket.

You can use versioning to preserve, retrieve, and restore every version of every object stored in your Amazon S3 bucket.



Once you version enable a bucket, it can never return to an unversioned state. You can, however, suspend versioning on that bucket.

The versioning state applies to all (never some) of the objects in that bucket.



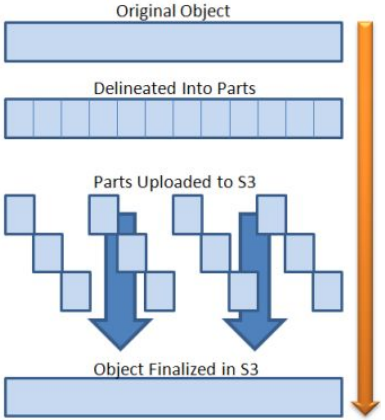
Multi-Part Upload

Saves our computer lives

Understanding Multi-Part

- Multi-Part upload is a way in which we upload an entire file in the form of small individual chunks to the storage device.

- While uploading data via multi-part, we need to specify the part number and its position in the uploaded object. This will help AWS reconstruct data.

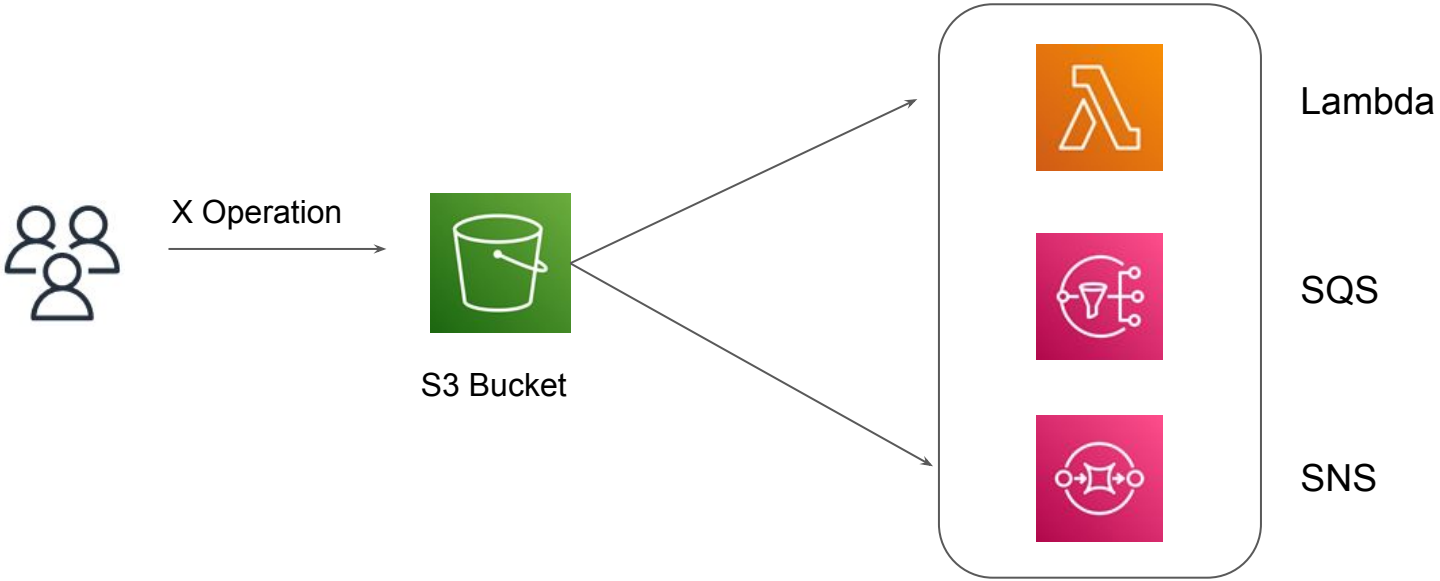


S3 Event Notification

S3 is more than just storage

Overview of S3 Event Notification

The Amazon S3 notification feature enables you to receive notifications when certain events happen in your bucket.



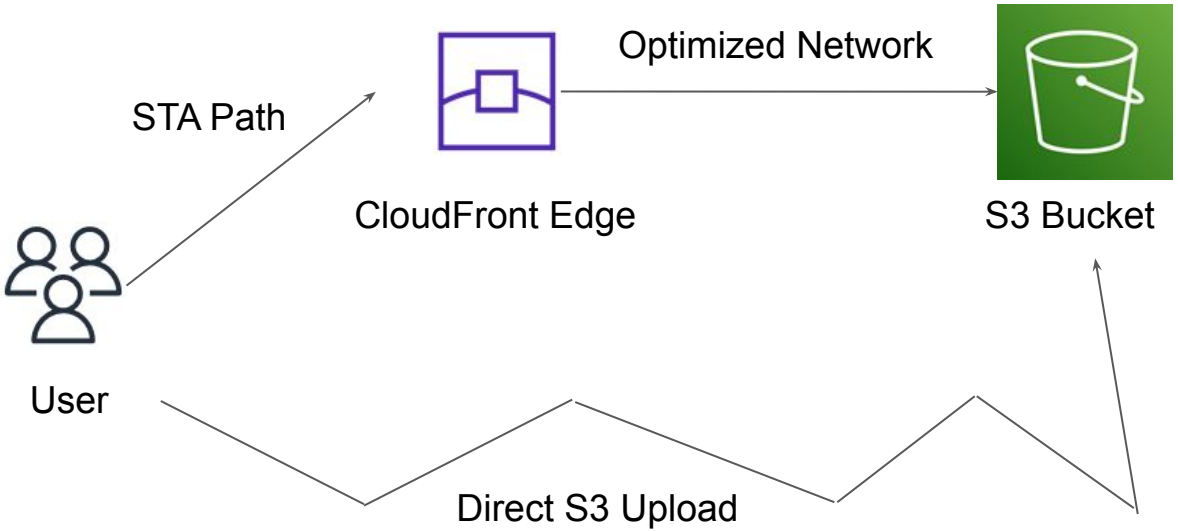
S3 Transfer Acceleration

Least Latencies

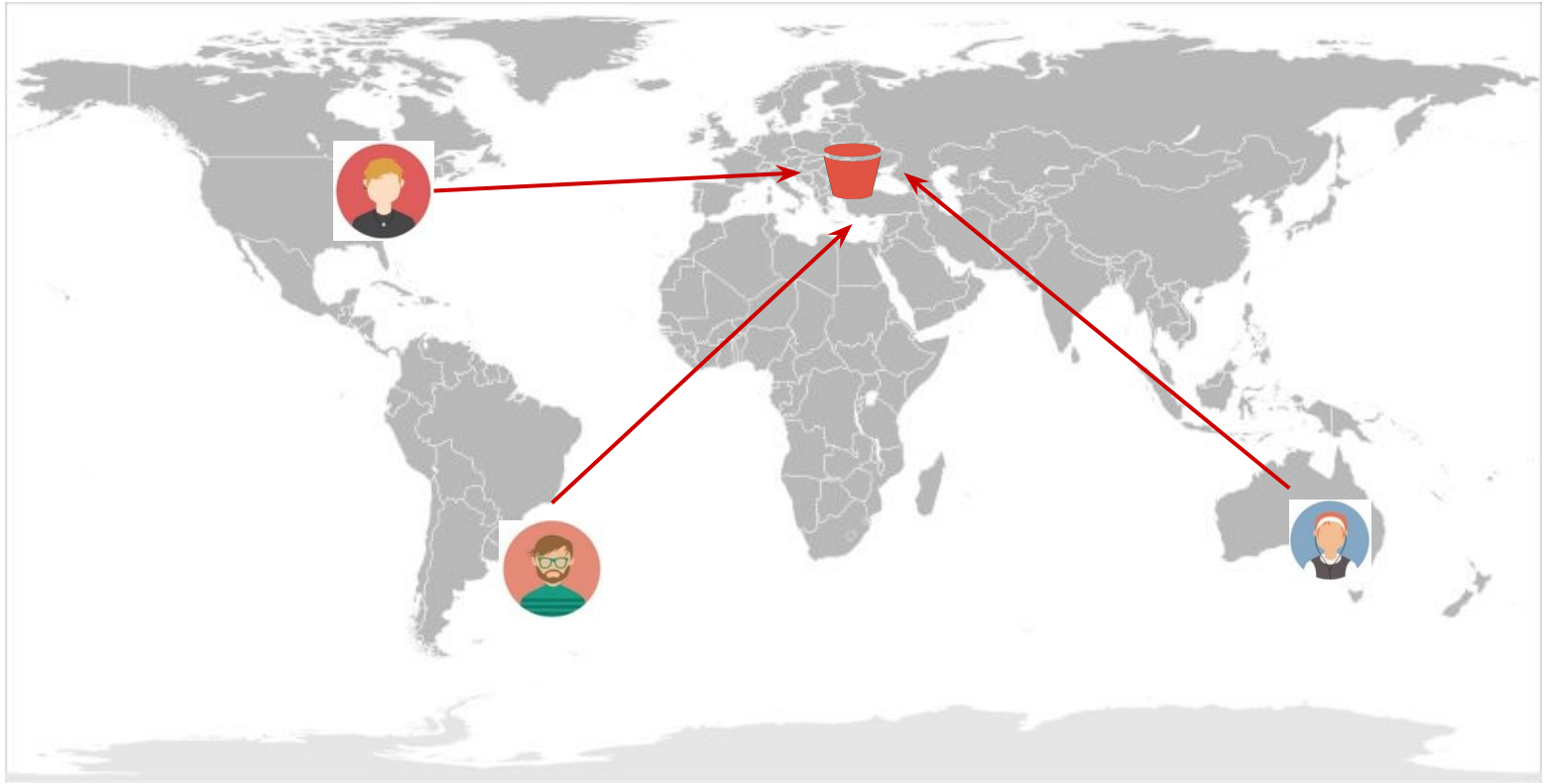
Overview of S3 Transfer Acceleration

S3 Transfer Acceleration allows users to accelerate data uploads from all over the world to centralized S3 bucket.

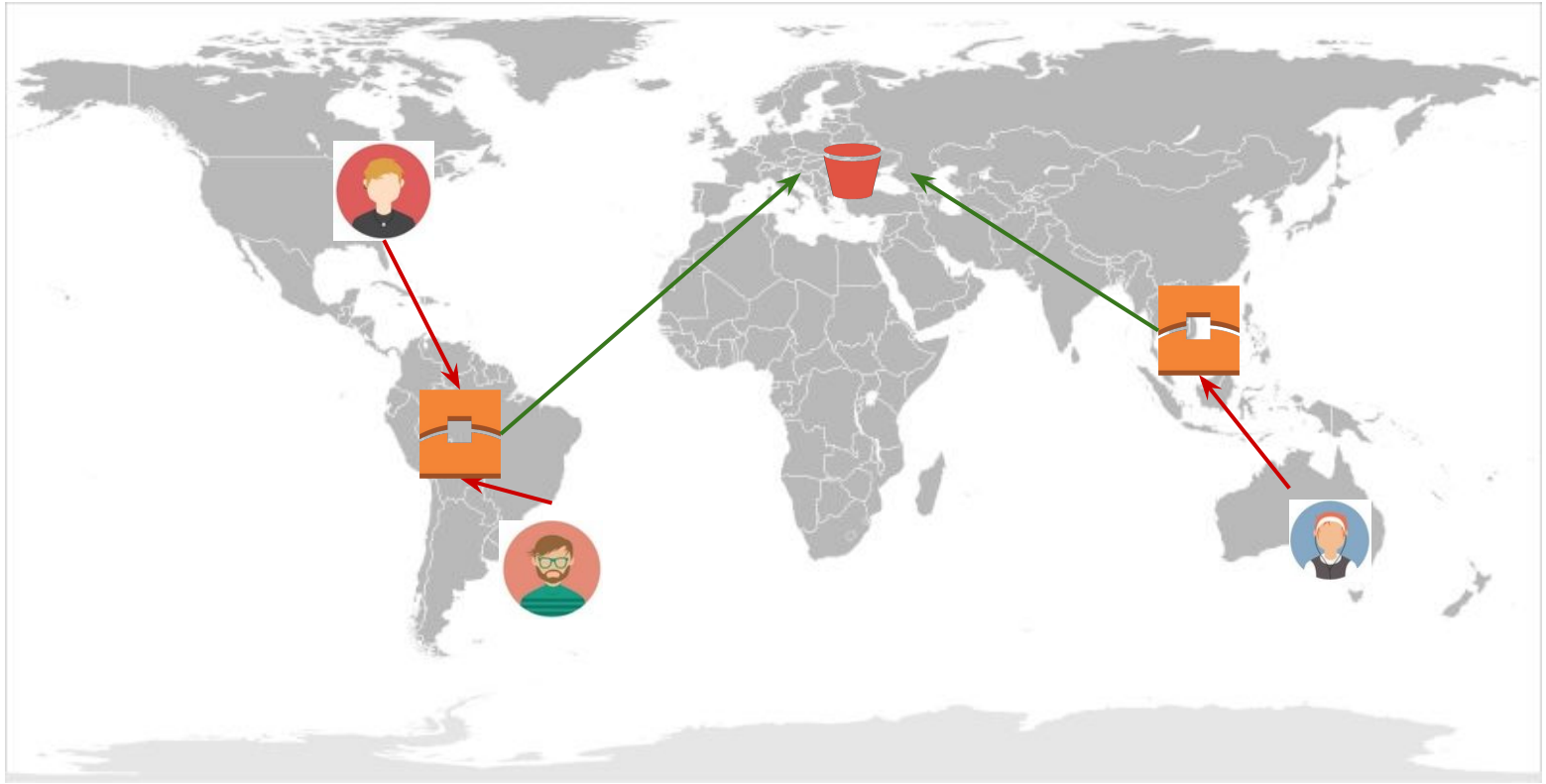
The transfers are accelerated by routing data to the closest edge location.



Edge Locations



S3 Transfer Acceleration



Range GET in AWS S3

Back to Logging!

Data Retrieval Options

There are two options while retrieving data from AWS S3.

Retrieval Configuration	Description
Retrieve an entire object	A single GET operation can return you the entire object stored in Amazon S3.
Retrieve object in parts	Using the Range HTTP header in a GET request, you can retrieve a specific range of bytes in an object stored in Amazon S3.

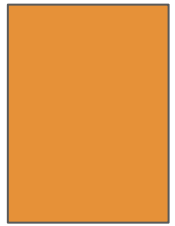
How it Works



Give me only first 10 MB of the file.







Alright, here you go.



Benefits

There are two important benefits of using this type of operation:

1. This resumable download is useful when you need only portions of your object data.

2. It is also useful where network connectivity is poor and you need to react to failures.

S3 Storage Classes

Cloud Storage is Saviour

Use-Case - Netflix

Netflix offers various different subscription plans for various category of requirements.

Main Aim: Watch the Entertainment Content



	Basic	Standard	Premium
Monthly cost* (Arab Emirates Dirham)	**29 AED**	**39 AED**	**56 AED**
Number of screens you can watch on at the same time	1	2	4
Number of phones or tablets you can have downloads on	1	2	4
Unlimited movies and TV shows	✓	✓	✓
Watch on your laptop, TV, phone or tablet	✓	✓	✓
HD available		✓	✓
Ultra HD available			✓

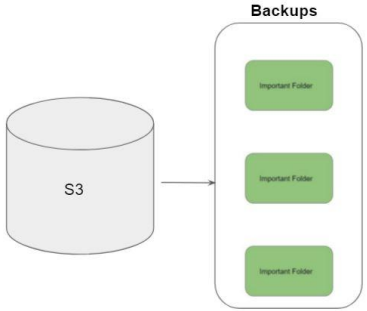
Initial Challenge - S3

AWS has millions of active customers.

Each customer might have different requirements for data storage.

Main Aim: Store Data.





S3 Storage Classes

Amazon S3 offers a range of storage classes designed for different use cases.

Storage Classes	Description
S3 Standard	Offers high durability, availability, and performance object storage for frequently accessed data
S3 Standard-Infrequent Access	For data that is accessed less frequently, but requires rapid access when needed.
Amazon S3 Glacier	Low-cost storage class for data archiving

AWS S3 Standard

S3 Standard offers high durability, availability, and performance object storage for frequently accessed data.

Designed for durability of 99.999999999% of objects (eleven nines)

Example :-

If we have 10,000 files stored in S3 (11 nines durability) then you can expect to lose one file every ten million years.

AWS S3 Standard IA

S3 Standard-IA is for data that is accessed less frequently, but requires rapid access when needed.

Comparing storage cost of 1TB data stored in S3 based on accessibility patterns.

Criteria	Amazon S3	Amazon S3 IA
Storage of 1TB Data	$23.44	$23.50
50% storage accessed in last 30 days	-	$18.18
0% storage accessed in last 30 days	-	$12.80

Amazon S3 Glacier

Glacier is meant to be for archiving and for storing long-term backups.

Ideally meant for data that needs to be archived for years without much requirement of access.

Criteria	Amazon S3	Glacier
Storage of 1TB Data	$23.44	$4.10

Multiple S3 Storage Classes

Performance across the S3 Storage Classes

	S3 Standard	S3 Intelligent-Tiering*	S3 Standard-IA	S3 One Zone-IA†	S3 Glacier	S3 Glacier Deep Archive
Designed for durability	99.999999999% (11 9's)	99.999999999% (11 9's)	99.999999999% (11 9's)	99.999999999% (11 9's)	99.999999999% (11 9's)	99.999999999% (11 9's)
Designed for availability	99.99%	99.9%	99.9%	99.5%	99.99%	99.99%
Availability SLA	99.9%	99%	99%	99%	99.9%	99.9%
Availability Zones	≥3	≥3	≥3	1	≥3	≥3
Minimum capacity charge per object	N/A	N/A	128KB	128KB	40KB	40KB
Minimum storage duration charge	N/A	30 days	30 days	30 days	90 days	180 days

Durability vs Availability

- Durability is percent (%) over one year period of time that the file which is stored in S3 will not be lost.

- Availability is percent (%) over one year period of time that the file stored in S3 will not be available.

Example :-

For Servers, Availability is one of the key metric and any minute of downtime is a loss. However what happens if component of server itself fails and server goes down ?

S3 Intelligent-Tiering

Smart Automated System

Overview of S3 Intelligent Tiering

The S3 Intelligent Tiering is primarily designed to optimize cost by automatically moving data to most cost-effective tier.

- 1TB of data stored in Standard S3 = $23.44
- 1TB of data stored in Standard IA = $12.80

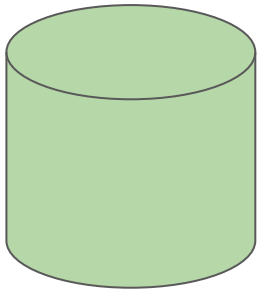
Organization stores terabytes of data in S3.

It will be great if a solution automatically moves infrequent data to Standard IA.

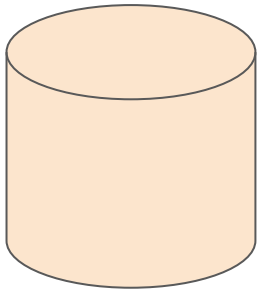
Overview of S3 Intelligent Tiering

The S3 Intelligent Tiering works by storing data in one of the two access tiers:

- Frequent Access Tier (Costly)
- Infrequent Access Tier (Much cheaper)



Frequent Access Tier

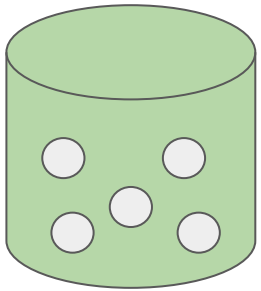


Infrequent Access Tier

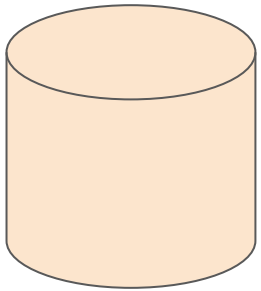
Overview of S3 Intelligent Tiering

The S3 Intelligent Tiering works by storing data in one of the two access tiers:

- Frequent Access Tier (Costly)
- Infrequent Access Tier (Much cheaper)

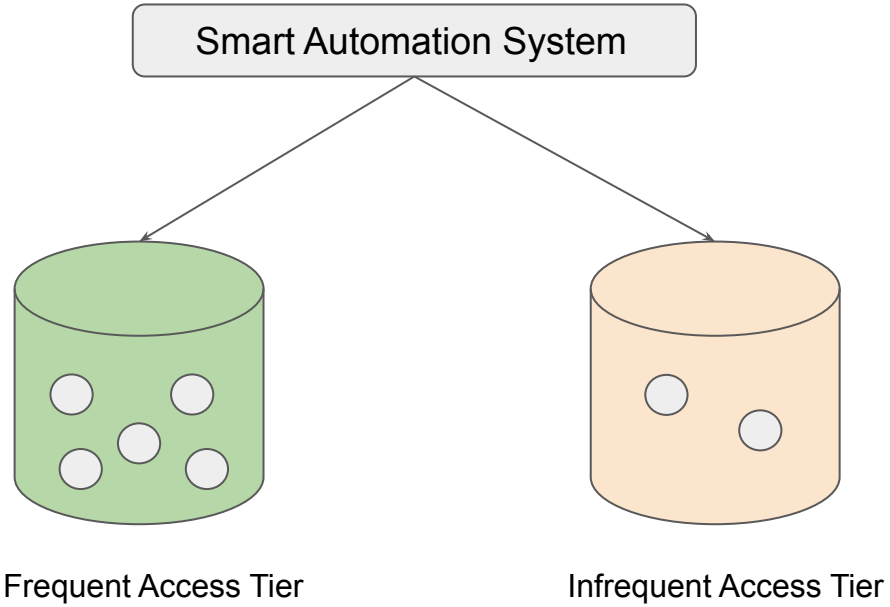


Frequent Access Tier



Infrequent Access Tier

Overview of S3 Intelligent Tiering

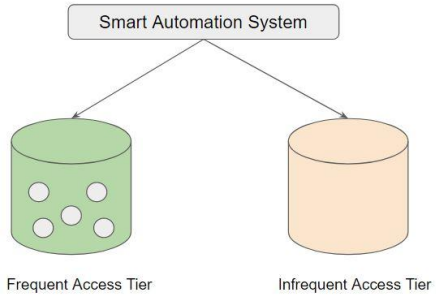


Revising S3 Intelligent Tiering

Amazon S3 monitors access patterns of the objects in S3 Intelligent-Tiering, and moves the ones that have not been accessed for 30 consecutive days to the infrequent access tier.

If an object in the infrequent access tier is accessed, it is automatically moved back to the frequent access tier.

A monthly monitoring and automation fee is charged at a per object level.



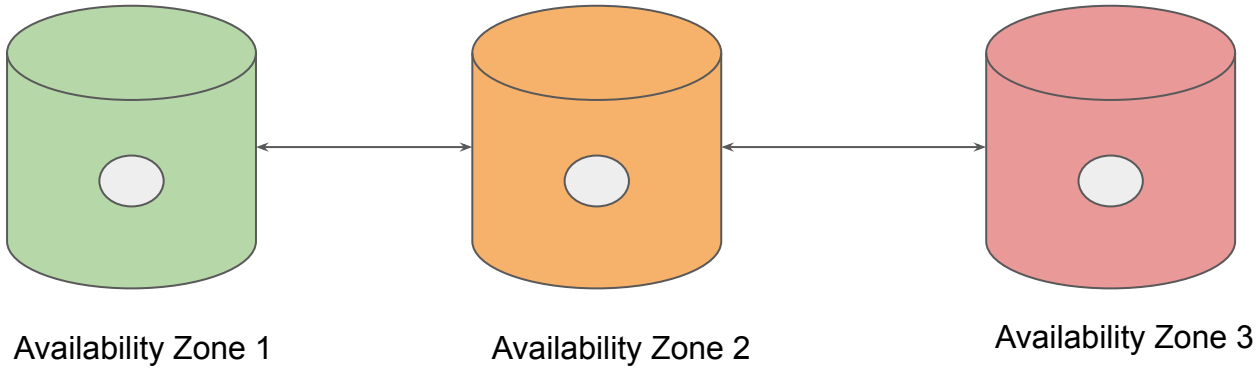
S3 Storage Class - One Zone IA

Back again!

Understanding the Basics

Storage classes like Standard S3, Standard IA stores the data in minimum 3 availability zones.

Due to this, the overall cost per of storage is increased with such architecture.

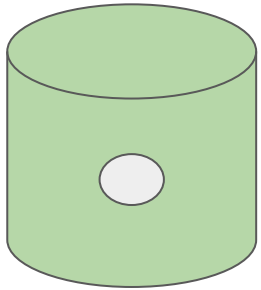


Overview of One Zone IA

S3 One Zone-IA stores data in a single AZ and costs 20% less than S3 Standard-IA.

It's a good choice for storing secondary backup copies of on-premises data or easily recreatable data.

Data will be lost in-case of availability zone destruction.



Availability Zone 1

Pricing Comparison

Overview of Pricing comparison between storage classes:

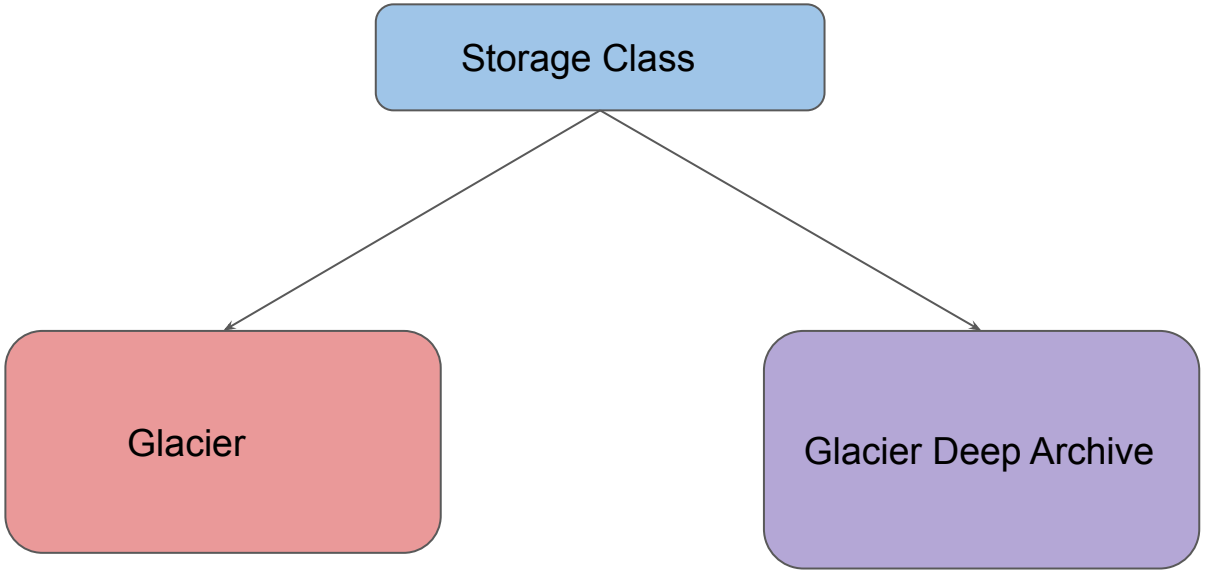
- 1TB of data stored in Standard S3 = $23.44
- 1TB of data stored in Standard IA = $12.80
- 1 TB of data stored in One Zone IA = $10.24

S3 Storage Class - Glacier

Back again!

Overview of Glacier

Amazon S3 Glacier is a low-cost, cloud-archive storage service that provides secure and durable storage for data archiving and online backup.



Pricing Comparison

Storage Class & Data Storage	Pricing
1 TB Data stored in S3 Standard	$23.44
1 TB Data stored in One Zone IA	$12.80
1 TB Data stored in Glacier	$4.10
1 TB Data stored in Glacier Archive	$1.02

Glacier vs Glacier Deep Archive

To keep costs low, Amazon S3 Glacier provides three options for access to archives, from a few minutes to several hours.

Glacier Deep Archive provides two access options, which range from 12 to 48 hours

Storage Class	Expedited	Standard	Bulk
Amazon S3 Glacier	1–5 minutes	3–5 hours	5–12 hours
S3 Glacier Deep Archive	Not available	Within 12 hours	Within 48 hours

Important Note

Amazon S3 Glacier for archiving data that might infrequently need to be restored from minutes to few hours.

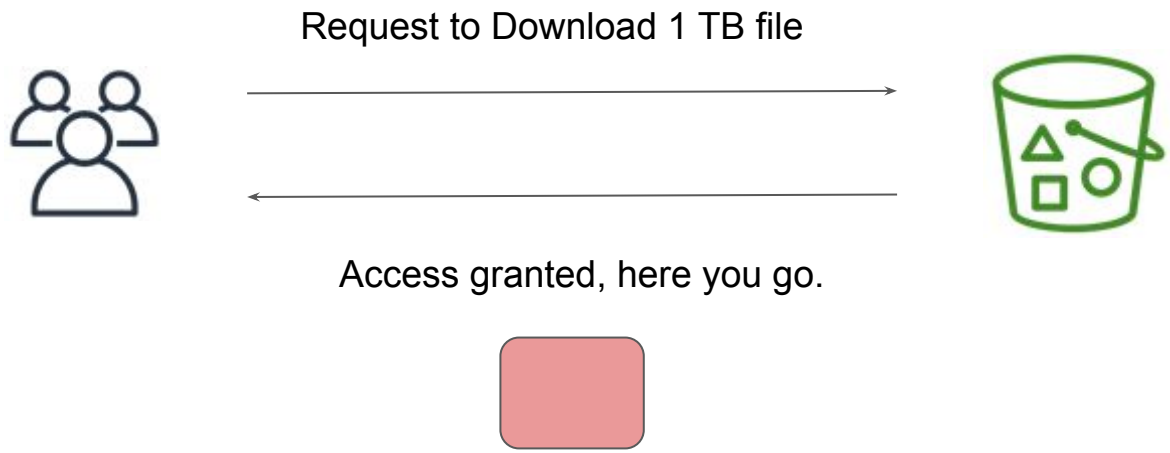
S3 Glacier Deep Archive for archiving long-term backup cycle data that might infrequently need to be restored within 12 hours

S3 Requester Pays

Back to Billing!

Understanding the Challenge

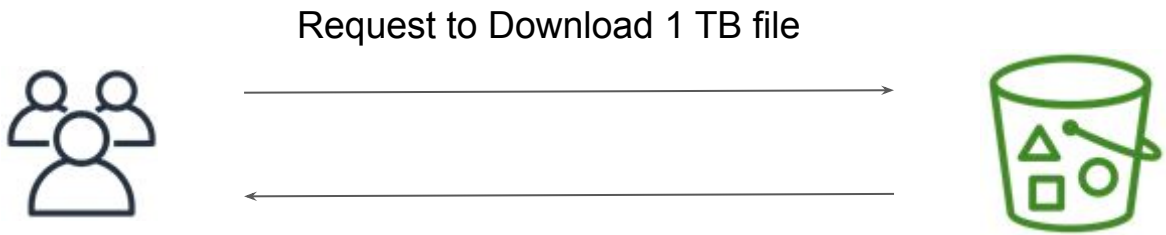
In general, bucket owners pay for all Amazon S3 storage and data transfer costs associated with their bucket.



After Requester Pays

With Requester Pays buckets, the requester instead of the bucket owner pays the cost of the request and the data download from the bucket.

The bucket owner always pays the cost of storing data.



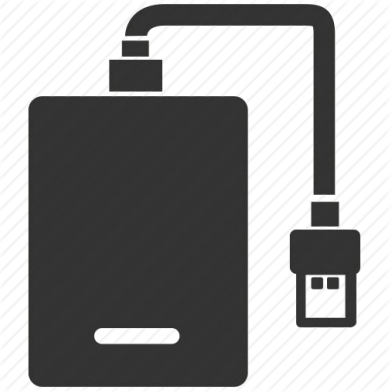
S3 Encryption

S3 is Back

What's the Need ?







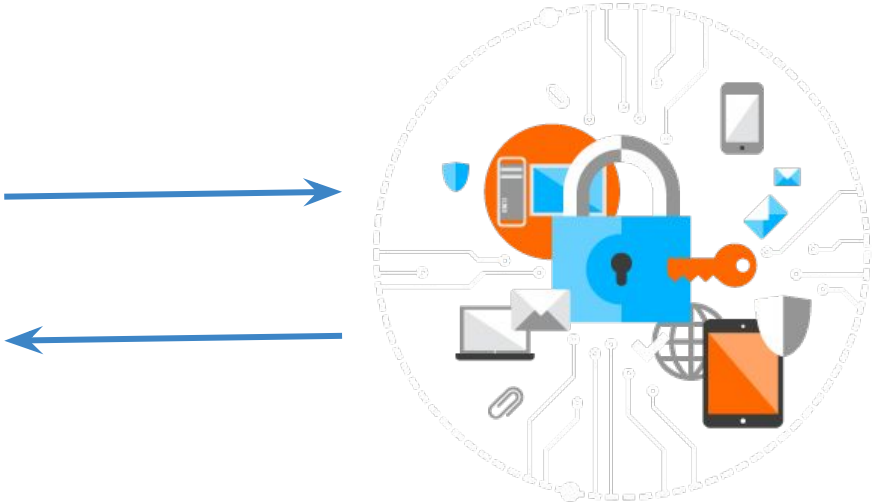
Let's be Proactive

Western Digital external HDs with hardware-based encryption

Aspiring to be a CISSP in 2017? Download the free planning kit!

WD introduced its new My Book Essential and My Book for Mac desktop external hard drives equipped with the new WD SmartWare software and hardware-based encryption.





S3 also needs Encryption

AWS S3 offers multiple approaches to encrypt the data being stored in S3.

i) Server Side Encryption

- Request Amazon S3 to encrypt your object before saving it on disks in its data centers and then decrypt it when you download the objects.

ii) Client Side Encryption

- Encrypt data client-side and upload the encrypted data to Amazon S3. In this case, you manage the encryption process, the encryption keys, and related tools.

Server Side Encryption

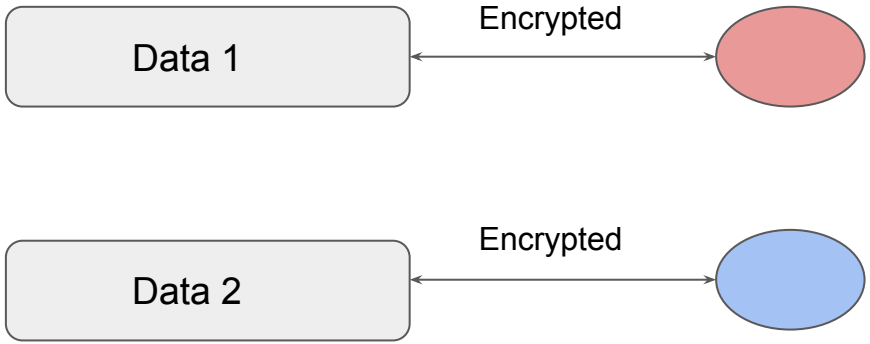
Within Server-Side encryption, there are three options that can be used depending on the use-case.

- Server-Side Encryption with Amazon S3-Managed Keys (SSE-S3)
- Server-Side Encryption with Customer Master Keys (CMKs) Stored in AWS Key Management Service (SSE-KMS)
- Server-Side Encryption with Customer-Provided Keys (SSE-C)

SSE with Amazon S3-Managed Keys (SSE-S3)

i) Server-Side Encryption with Amazon S3-Managed Keys (SSE-S3)

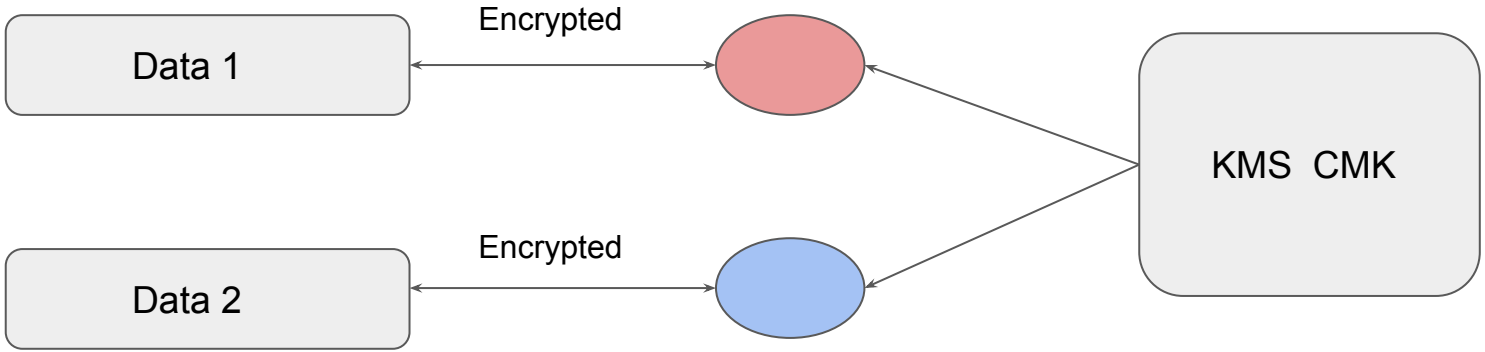
- In this approach, each object is encrypted with a unique key.
- Uses one of the strongest block ciphers to encrypt the data, AES 256.



SSE with CMK (SSE-KMS)

ii) Server-Side Encryption with CMKs Stored in AWS Key Management Service (SSE-KMS)

Encrypting data with own CMK allows customers to create, rotate, disable customer managed CMK's. We can also define access controls and enable auditing.



SSE with Customer-Provided Keys (SSE-C)

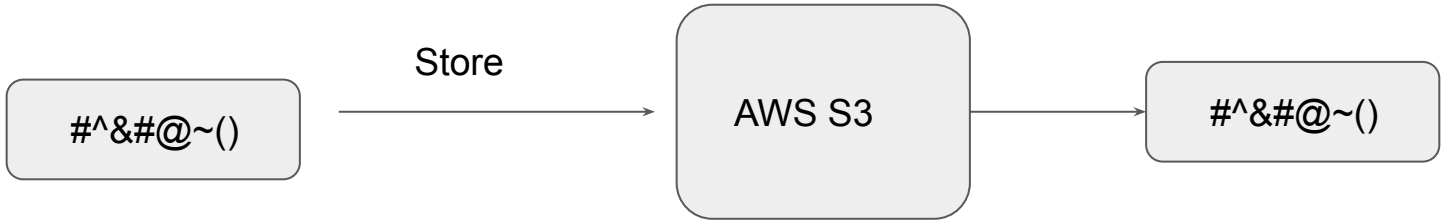
Allows customers to set their own encryption keys.

Encryption key needs to be provided as part of the request and S3 will manage both the encryption as well as the decryption options.



Client Side Encryption

Client-side encryption is the act of encrypting data before sending it to Amazon S3.



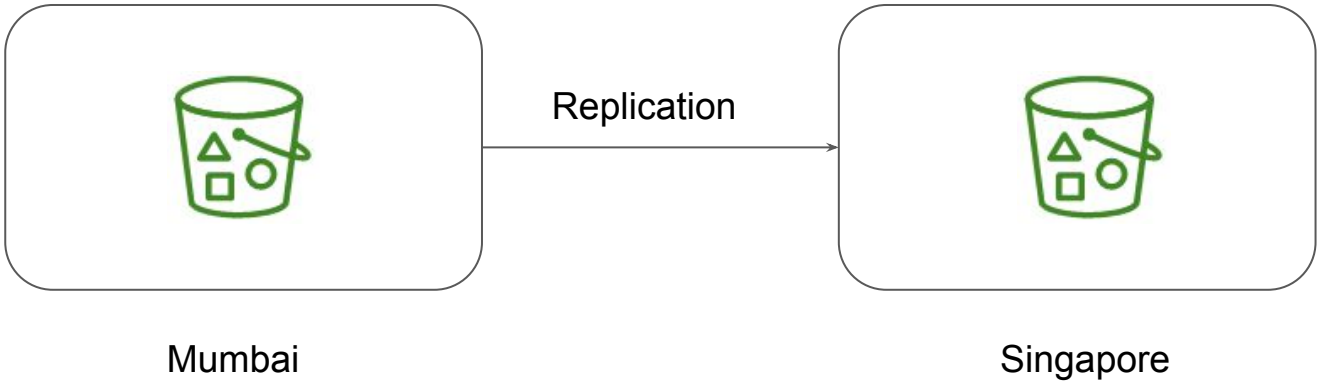
S3 - Cross Region Replication

Storage Service

Understanding the Use-Case

Many compliance has a requirement that the data must be replicated across greater distances.

Cross-Region Replication allows data from S3 buckets to be replicated across regions.



Important Pointers

Both source and destination buckets must have versioning enabled.

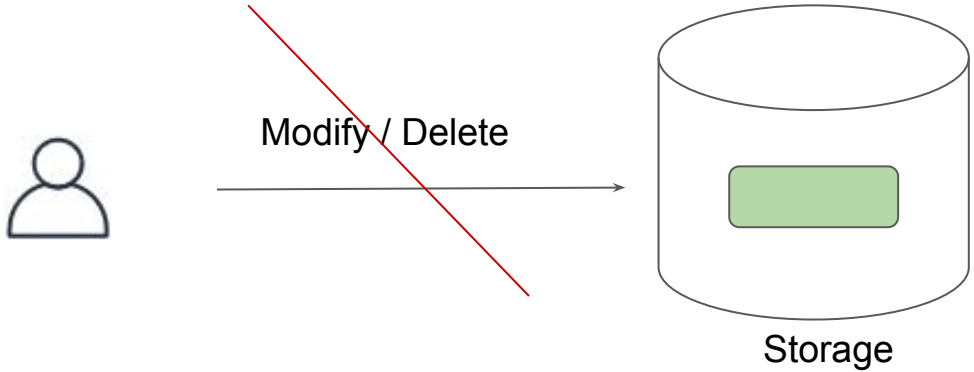
S3 Object Lock

Mastering S3

Overview of WORM

Write once read many (WORM) describes a data storage device in which information, once written, cannot be modified.

This write protection affords the assurance that the data cannot be tampered with once it is written to the device.



Use-Case - Ransomware

Ransomware also blackmail trojans , blackmail software are malicious programs with the help of which an intruder can prevent the computer owner from accessing data, its use or the entire computer system.

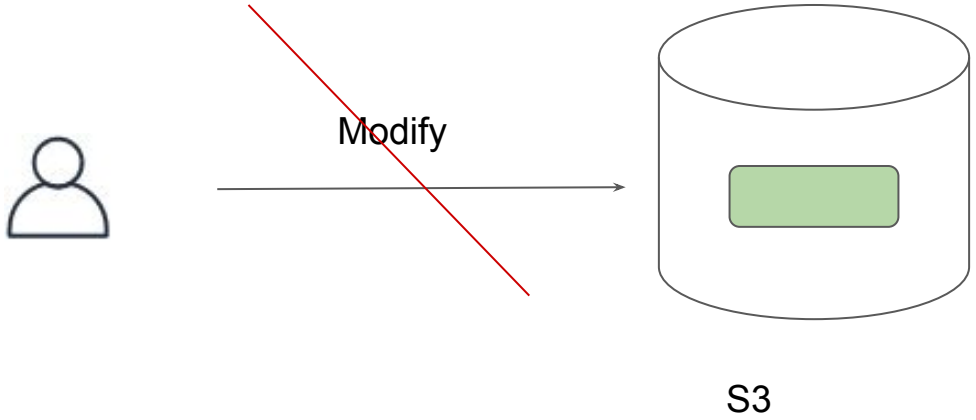
Private data on the foreign computer is encrypted or access to it is prevented in order to demand a ransom for decryption or release.



S3 Object Lock

With S3 Object Lock, you can store objects using a write-once-read-many (WORM) model.

You can use it to prevent an object from being deleted or overwritten for a fixed amount of time or indefinitely.



Retention Modes

Retention Mode	Description
Governance Mode	When deployed in Governance mode, AWS accounts with specific IAM permissions are able to remove object locks from objects.
Compliance Mode	In Compliance Mode, the protection cannot be removed by any user, including the root account.

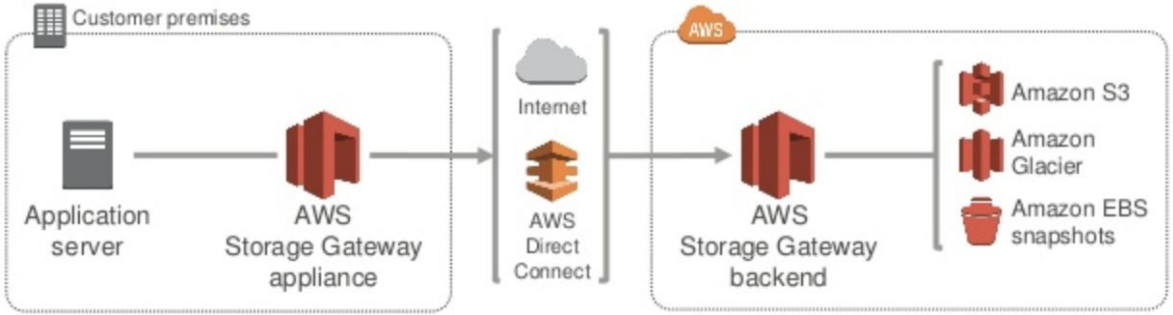
Storage Gateway

Hybrid Storage

Introduction

AWS Storage Gateway is a hybrid storage service that allows the on-premise application to easily use the cloud storage

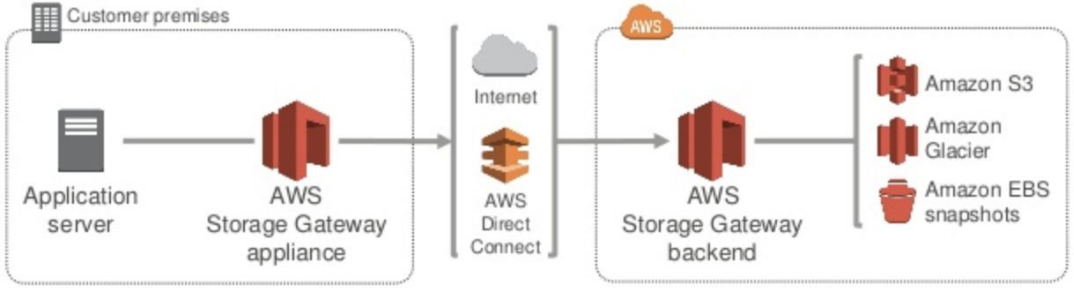
-



Storage Gateway

Storage Gateway appliance uses standard storage protocols like NFS, iSCSI which the application connects to and stores the data.

The other end of storage gateway connects to AWS storage services like S3, Glacier and EBS Snapshots



Storage Gateway Configuration

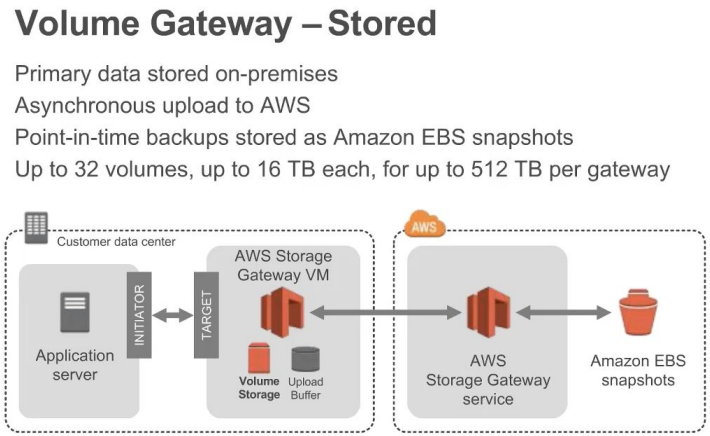
There are three different configuration available :

- Gateway Stored Volume
- Gateway Cached Volume
- Gateway-virtual tape library

Gateway Stored Volumes

Gateway Stored Volume :

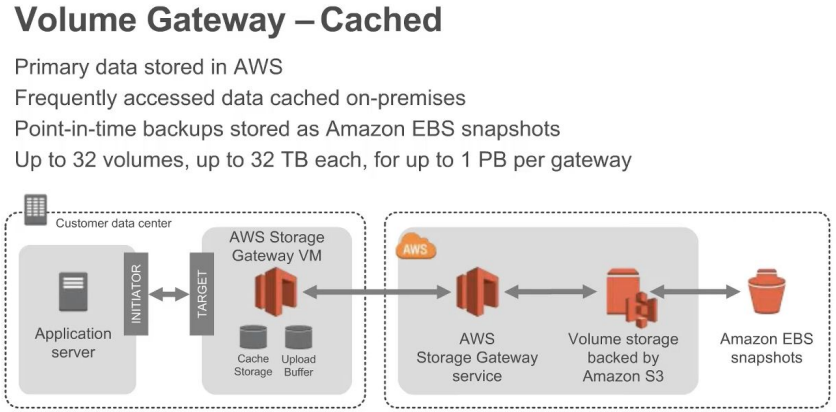
Stores primary data locally while asynchronously backing up data to AWS.



Gateway Cached Volumes

Gateway Stored Volume :

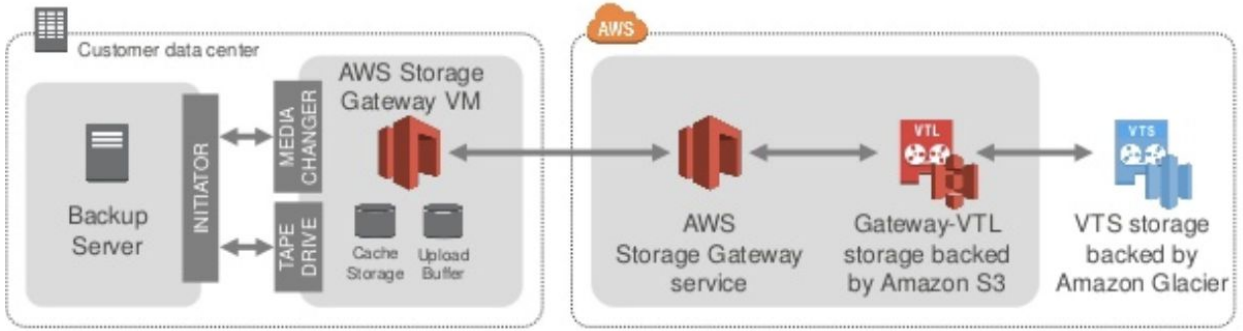
Data is stored primarily on AWS S3 with cache of recently read or written data stored locally in the on-premise server.



Gateway-virtual tape library

Gateway-virtual tape library

Virtual tapes stored in S3 with frequently accessed data stored on-premise.



Tape based storage

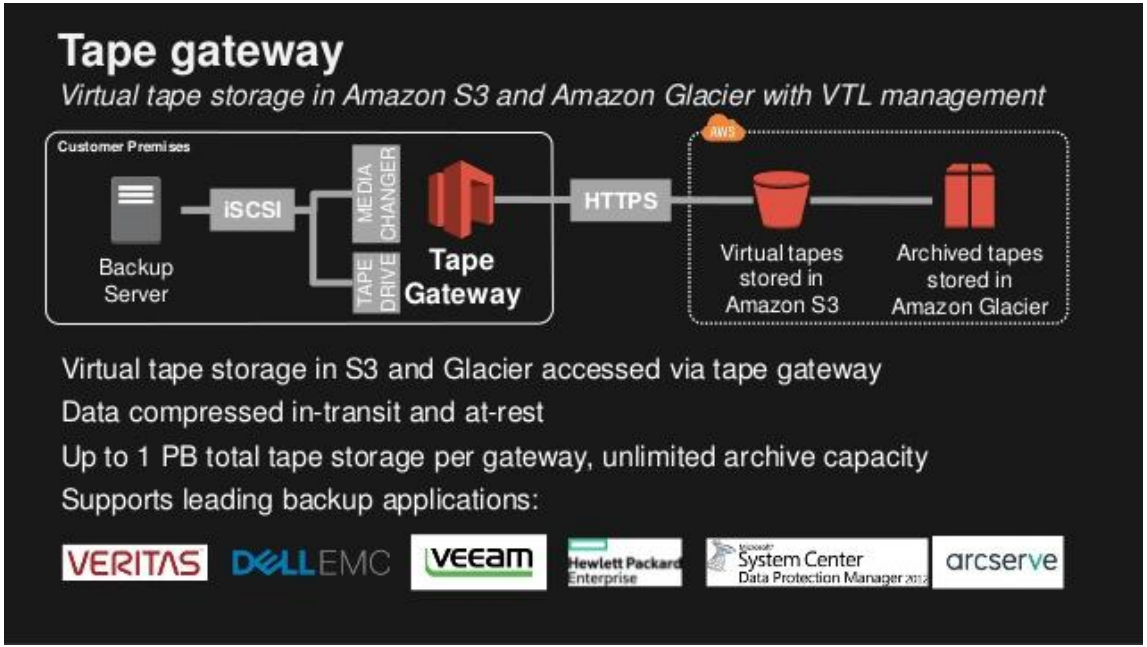
Tape backup is practice of periodically copying data from primary storage device to a tape cartridge so data can be recovered if there is any data crash or failure on primary device.

Tape solutions remains most cost effective solution till date.





VTS and VTL


Tape gateway
Virtual tape storage in Amazon S3 and Amazon Glacier with VTL management
Customer Premises
Backup Server
iSCSI
MEDIA CHANGER
TAPE DRIVE
Tape Gateway
HTTPS
AWS
Virtual tapes stored in Amazon S3
Archived tapes stored in Amazon Glacier
Virtual tape storage in S3 and Glacier accessed via tape gateway
Data compressed in-transit and at-rest
Up to 1 PB total tape storage per gateway, unlimited archive capacity
Supports leading backup applications:
VERITAS
DELL EMC
veeam
Hewlett Packard Enterprise
System Center Data Protection Manager
arcserve

File Gateway

Storage Gateway Type

Overview of File Gateway

AWS Storage Gateway's file interface, or file gateway provides an interface via NFS, SMB for synchronization of data from on-premise to S3 bucket.



Amazon FSx



Basics of Filesystems

There are multiple popular set of file systems / storage platforms available in the industry that are used extensively based on specific use-cases

File Systems	Description
Lustre	Parallel distributed file system, generally used for large-scale cluster computing (HPC)
Open ZFS	Encompasses the functionality of traditional file systems and logical volume manager. Benefits: protection against data corruption, efficient data compression etc

Understanding the Challenges

Many organizations have use-case to leverage the rich feature sets and fast performance of widely-used open source and commercially-licensed file systems.

This would lead to lot of time-consuming administrative tasks like hardware provisioning, software configuration, patching, and backups.

Introduction to FSx

Amazon FSx makes it easy and cost effective to launch and run popular file systems.

It provides cost-efficient capacity and high levels of reliability, and integrates with other AWS services so that you can manage and use the file systems in cloud-native ways.



Benefits of FSx

Benefits	Description
Simple and fully managed	In minutes and with a few clicks, you can launch a fully managed file system. No need to worry about configuring, patching, backups etc.
Secure and compliant	Amazon FSx automatically encrypts your data at-rest and in-transit. Complies with PCI-DSS, ISO, SOC certifications
Integration with AWS services	Integrate with AWS services, including Amazon S3, AWS KMS, Amazon SageMaker, Amazon WorkSpaces, AWS ParallelCluster.

Amazon FSx for Lustre

Provides cost-effective, high-performance, scalable file storage for compute workloads such as machine learning, high performance computing (HPC), video processing, and financial modeling.

Integrates seamlessly with Amazon S3, SageMaker, EKS etc

Amazon FSx for Windows File Server

Provides simple, fully managed, highly reliable file storage that's accessible over the industry-standard Server Message Block (SMB) protocol.

Built on Windows Server, providing full SMB support and a wide range of administrative features like user quotas, data deduplication, and end-user file restore. Accessible from Windows, Linux, and macOS.

Integrates with Microsoft Active Directory (AD) to support Windows-based environments and enterprises.

FSx for OpenZFS

Provides simple, cost-effective, high-performance file storage built on the OpenZFS file system accessible over the industry-standard NFS protocol.

Provides powerful OpenZFS data management capabilities including Z-Standard/LZ4 compression, instant point-in-time snapshots, and data cloning, thin provisioning, and user/group quotas.

Amazon FSx for NetApp ONTAP

Provides feature-rich, high-performance, and highly-reliable storage built on NetApp's popular ONTAP file system and fully managed by AWS.

Accessible via industry-standard NFS, SMB, and iSCSI protocols.

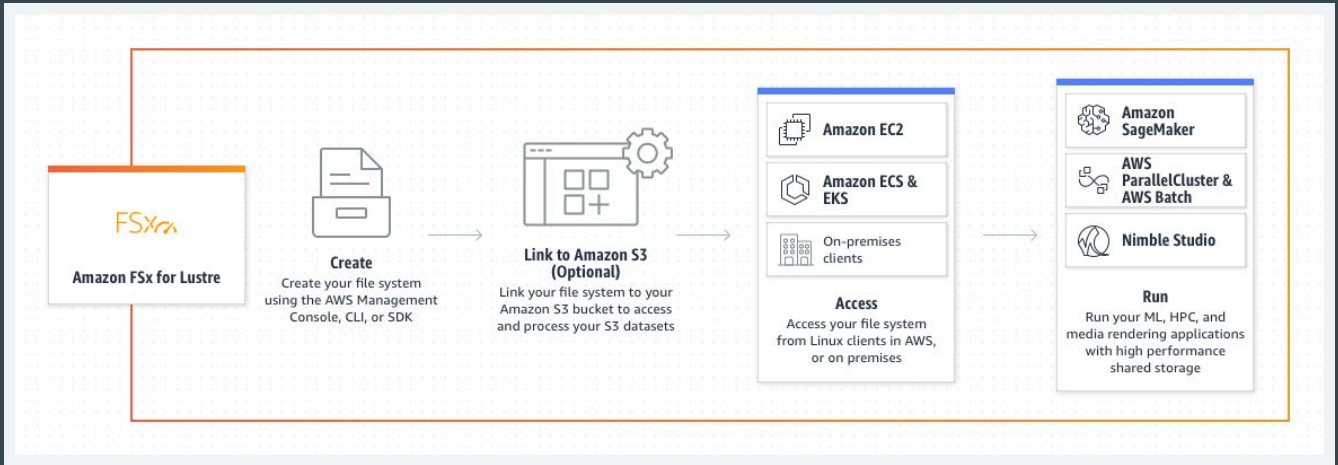
Integrates with Microsoft Active Directory (AD) to support Windows-based environments and enterprises.

Amazon FsX for Lustre



Understanding the Basics

Amazon FSx for Lustre provides fully managed shared storage with the scalability and performance of the popular Lustre file system.



Use-Cases

Accelerate machine learning (ML)

Enable high performance computing (HPC)

Unlock big data analytics

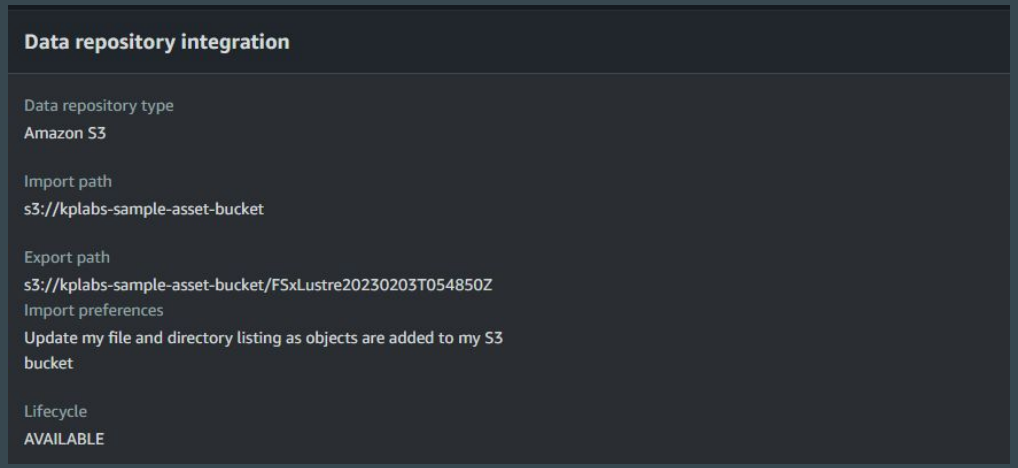
Increase media workload agility (VFX)

Use-Case Workflow



Data Repository

FSx for Lustre is natively integrated with data repositories such as Amazon S3, making it easier to process datasets with the Lustre file system.



Lazy Loading of Data

Lazy loading is the practice of delaying load or initialization of resources or objects until they're actually needed to improve performance and save system resources



S3 - Petabytes of Data

File system deployment options

Amazon FSx for Lustre provides two file system deployment options: scratch and persistent.

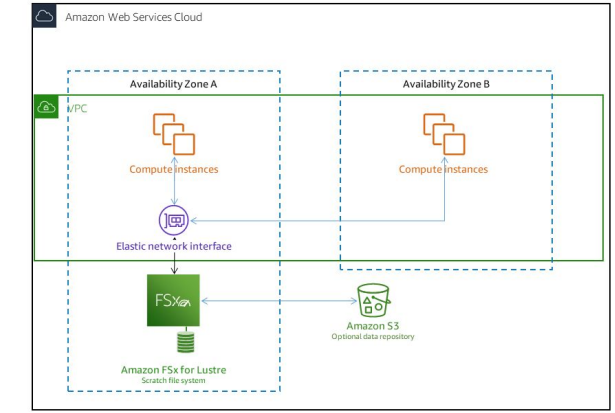
Depending on the option, the overall storage capacity and throughput changes.



Deployment Option 1 - Scratch

Scratch file systems are designed for temporary storage and shorter-term processing of data. Provides high burst throughput.

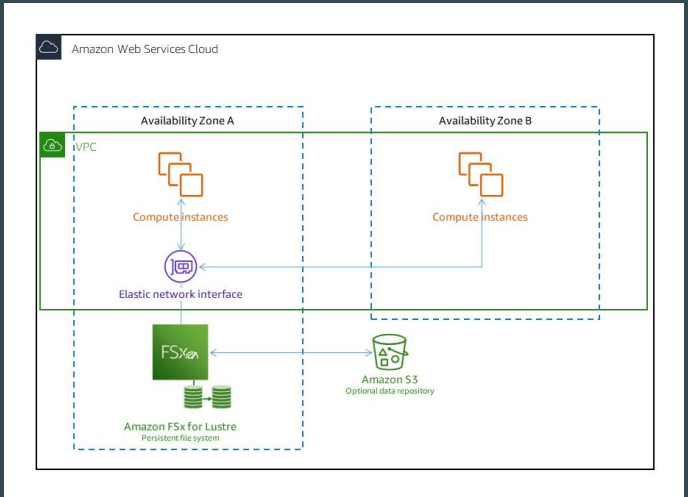
Data isn't replicated and doesn't persist if a file server fails



Deployment Option 2 - Persistent

Persistent file systems are designed for longer-term storage and workloads.

The file servers are highly available, and data is automatically replicated within the same Availability Zone in which the file system is located.



Points to Note

You can configure FSx for Lustre to keep content synchronized in both directions between the file system and the linked S3 buckets

Alternatively, you also have the option to import and export batches of new and changed data between the file system and S3 for fine-grained control over data synchronization.

Increasing FSx Storage Capacity



Understanding the Basics

When you increase the storage capacity of your Amazon FSx file system, behind the scenes, Amazon FSx adds a new, larger set of disks to your file system.

Amazon FSx then runs a storage optimization process in the background to transparently migrate data from the old disks to the new disks.

Understanding the Use-Case

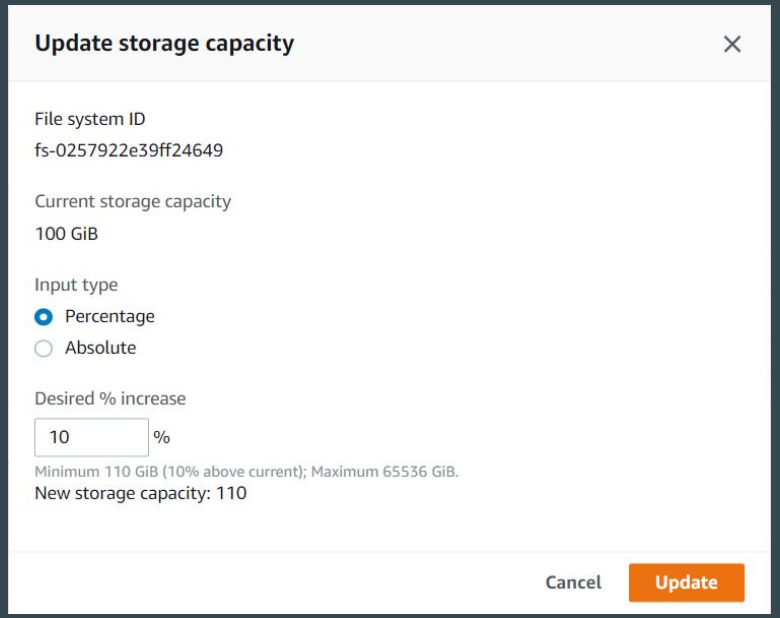


Points to Note

Storage optimization can take between a few hours and a few days, with minimal noticeable impact on the workload performance.

You can only increase the amount of storage capacity for a file system; you cannot decrease storage capacity.

To increase the storage capacity for an FSx for Windows File Server file system, use the AWS CLI command update-file-system.

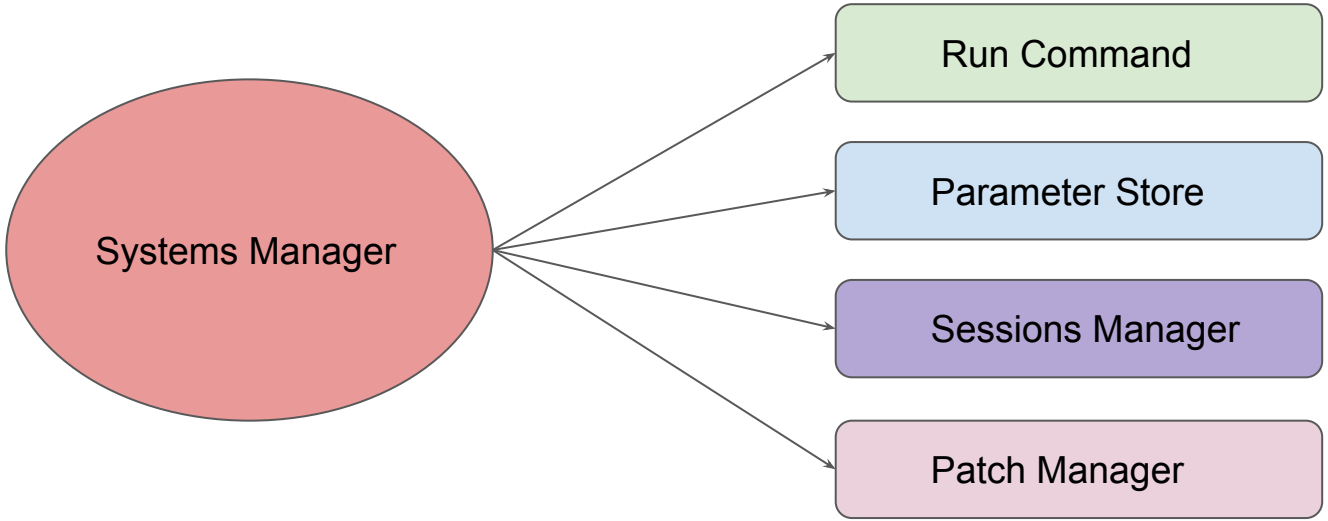

Update storage capacity
File system ID
fs-0257922e39ff24649
Current storage capacity
100 GiB
Input type
Percentage
Absolute
Desired % increase
10
%
Minimum 110 GiB (10% above current); Maximum 65536 GiB.
New storage capacity: 110
Cancel
Update

AWS Systems Manager

Interesting Set of Services

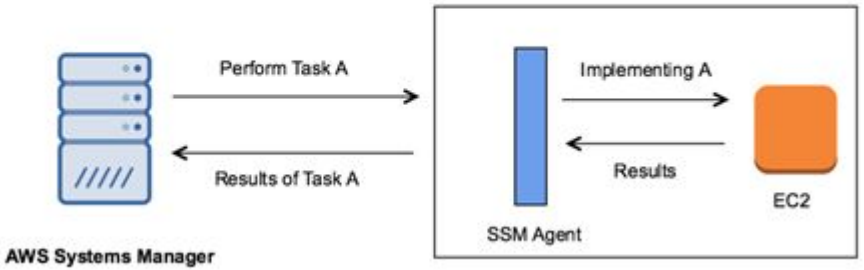
Overview of Systems Manager

AWS Systems Manager is a group of services which allows customers to have a better visibility and control of the infrastructure.



High Level Overview

The basic idea behind the " Systems Manager" is that there will be an SSM agent installed in the EC2 instances, and the customer can provide specific tasks to the installed agent from the systems manager console.

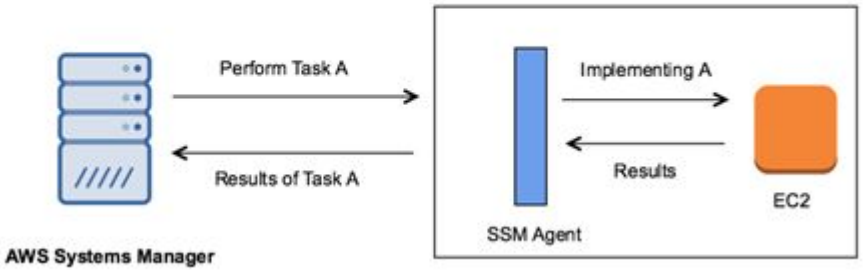


Configuring SSM Agent

Systems Manager Agent

High Level Overview

The basic idea behind the " Systems Manager" is that there will be an SSM agent installed in the EC2 instances, and the customer can provide specific tasks to the installed agent from the systems manager console.



Overview of the SSM Agent

AWS Systems Manager Agent (SSM Agent) is Amazon software that can be installed and configured on an Amazon EC2 instance, an on-premises server, or a virtual machine (VM).

SSM Agent is preinstalled, by default, on the following Amazon Machine Images (AMIs):

- Amazon Linux
- Amazon Linux 2
- Ubuntu Server 16.04, 18.04, and 20.04
- Amazon Linux 2 ECS-Optimized Base AMIs

Required Permissions

By default, AWS Systems Manager doesn't have permission to perform actions on your instances

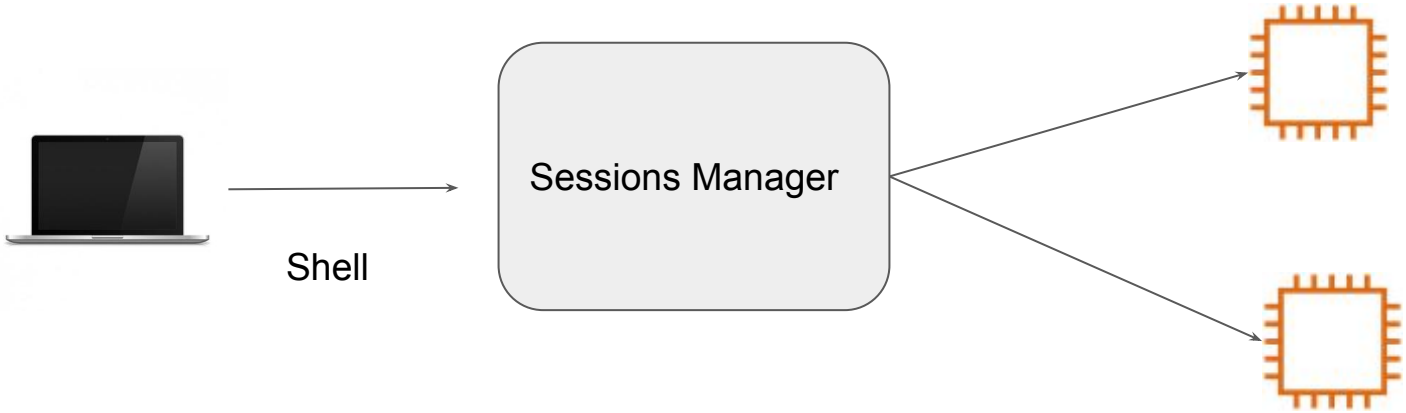
You need to attach IAM role with AmazonSSMManagedInstanceCore policy to allow an instance to use Systems Manager service core functionality.

Systems Manager - Sessions Manager

Interesting Set of Services

Overview of Sessions Manager

Sessions Manager allows customers to connect to the instances through an interactive one-click browser-based shell or through the AWS CLI.



Difference Between EC2 Connect & Sessions Manager

	EC2 Connect	Sessions Manager
IAM Role for EC2	Not Required	Required
Security Group (22)	Required	Not Required
Public IP	Required	Not Required

Benefits of Sessions Manager

Some of the notable benefits of Sessions Manager are as follows:

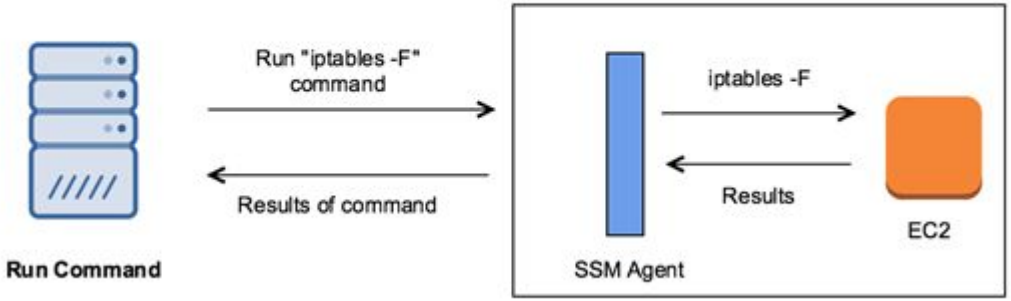
- Centralized Access Control using IAM Policies
- No Inbound Ports Needs to be Open
- Logging and auditing session activity
- One-click access to instances from the console and CLI
- No need of VPN to connect to instances.

Systems Manager - Run Command

Running Commands Remotely

Overview of Run Command

Run Command, as the name suggests allows us to run specific commands in the instances where SSM agent is installed.



Document Feature

Run Command provides much more granular features because of its “command document” feature.

There are various command document available that can perform certain ready-made actions.

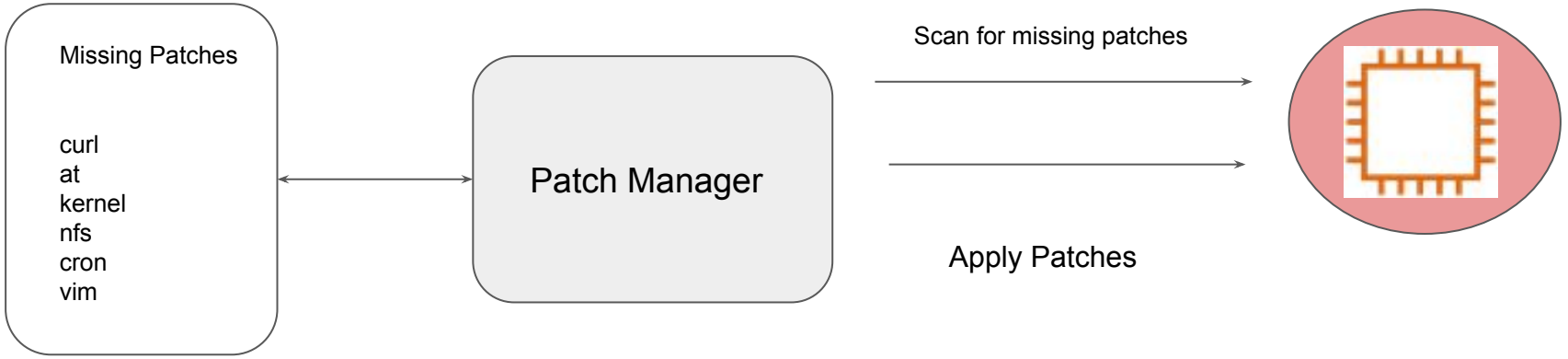
- AWS-RunAnsiblePlaybook
- AWS-ConfigureDocker
- AWS-InstallMissingWindowsUpdates
- AWS-RunShellScript

Systems Manager - Patch Manager

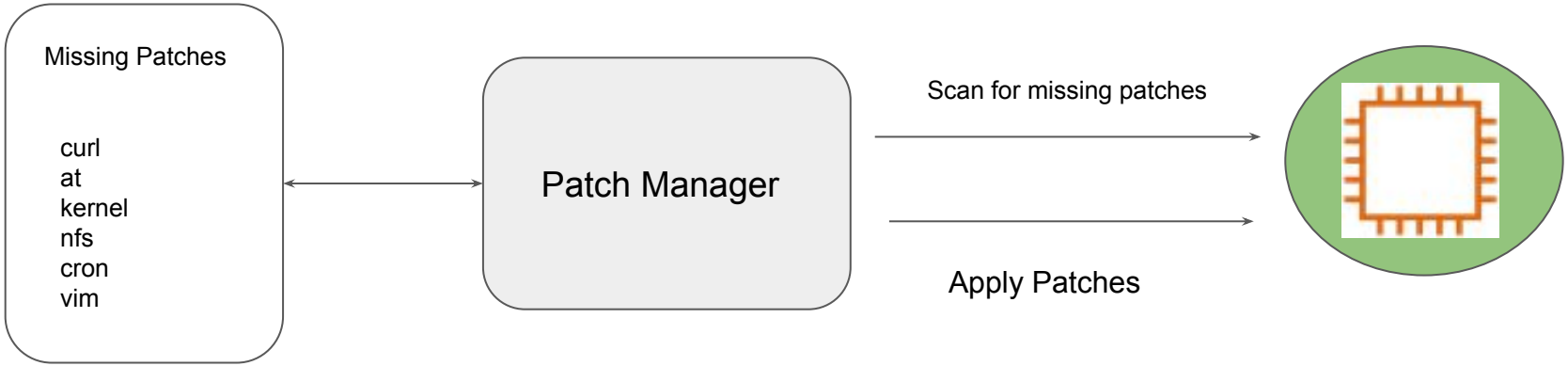
Interesting Set of Services

Overview of Patch Manager

Patch Manager automates the process of patching managed instances with both security related and other types of updates.



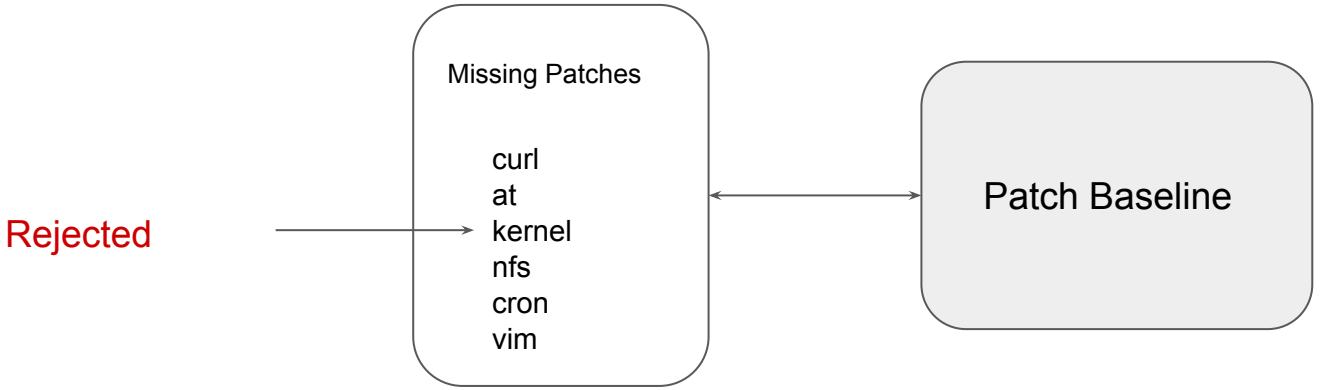
After Patching



Patch Baseline

Patch Baseline service determines the list of missing patches that need to be installed in the EC2 instance.

A patch baseline defines which patches are approved for installation on your instances. You can specify approved or rejected patches one by one.



Maintenance Window

Maintenance Window provides a mechanism for scheduling a particular activity on the specific target.

Example: Perform Patching activity at 2 AM in the morning.



SSM - Automation

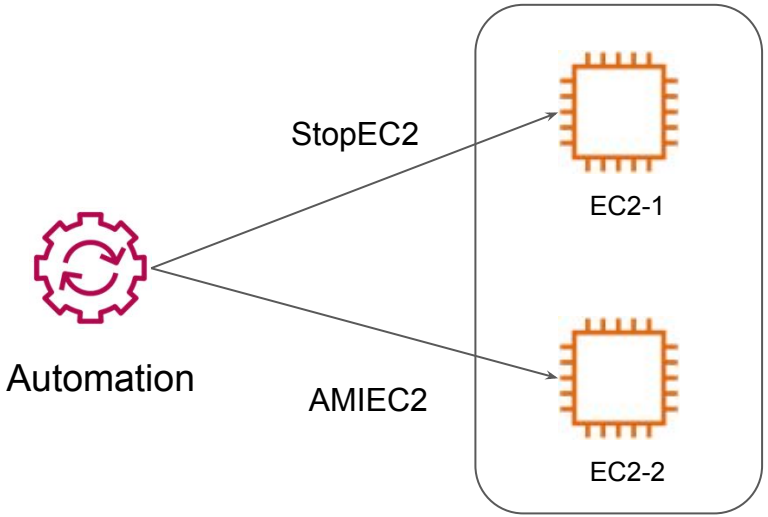
Automate Everything

Overview of SSM Automation

Automation, a capability of AWS Systems Manager, simplifies common maintenance and deployment tasks Amazon EC2 instances and other AWS resources.

Example Automation Tasks:

- Attach IAM to EC2 Instances
- Create AMI of Instances
- Perform Patching Activities



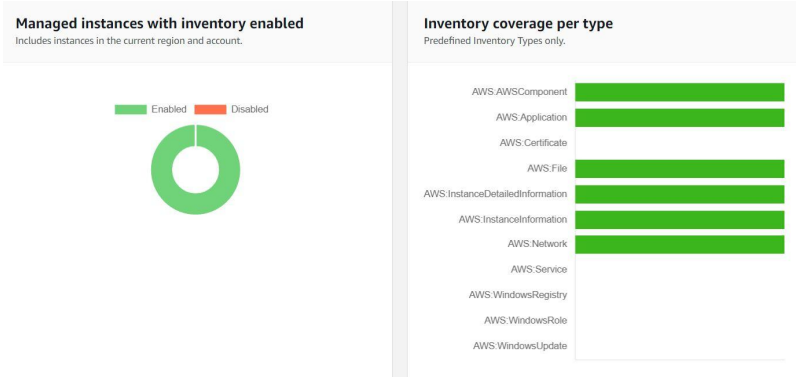
SSM - Inventory

Automate Everything

Overview of SSM Inventory

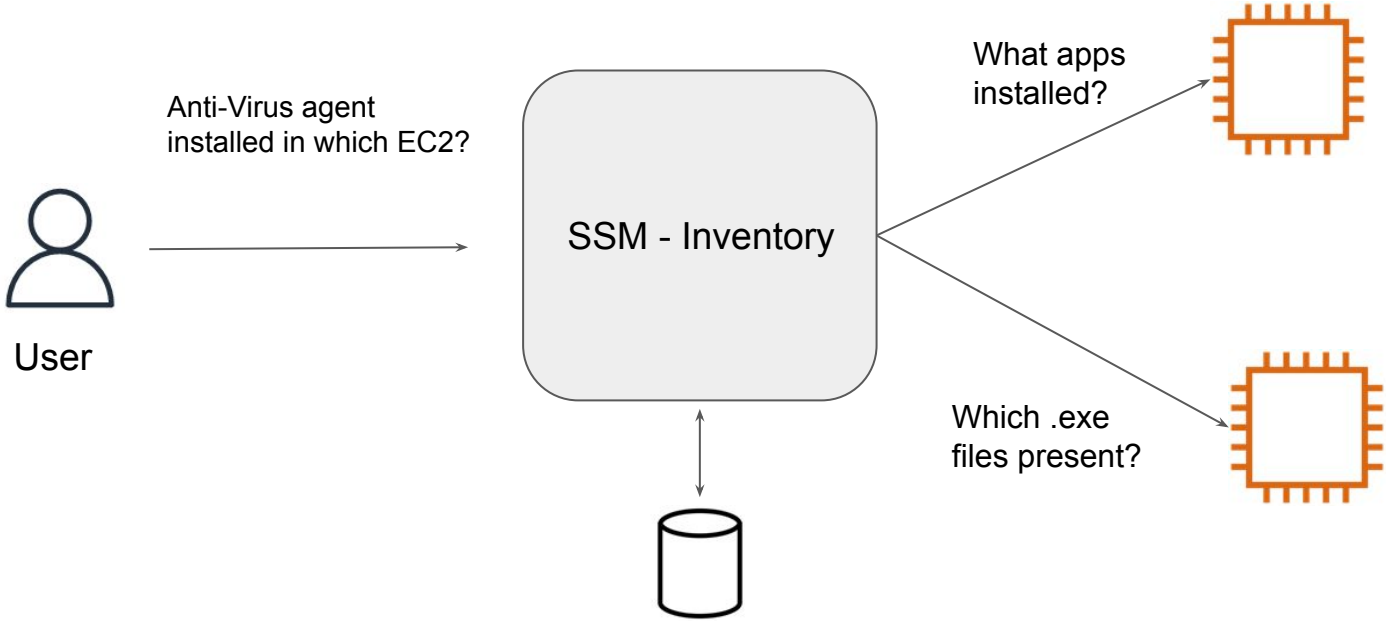
AWS Systems Manager Inventory provides visibility into your Amazon EC2 and on-premises computing environment.

It can capture various informations like Application Names, Files, Network Configuration, Instance Details, Windows Registry and others.

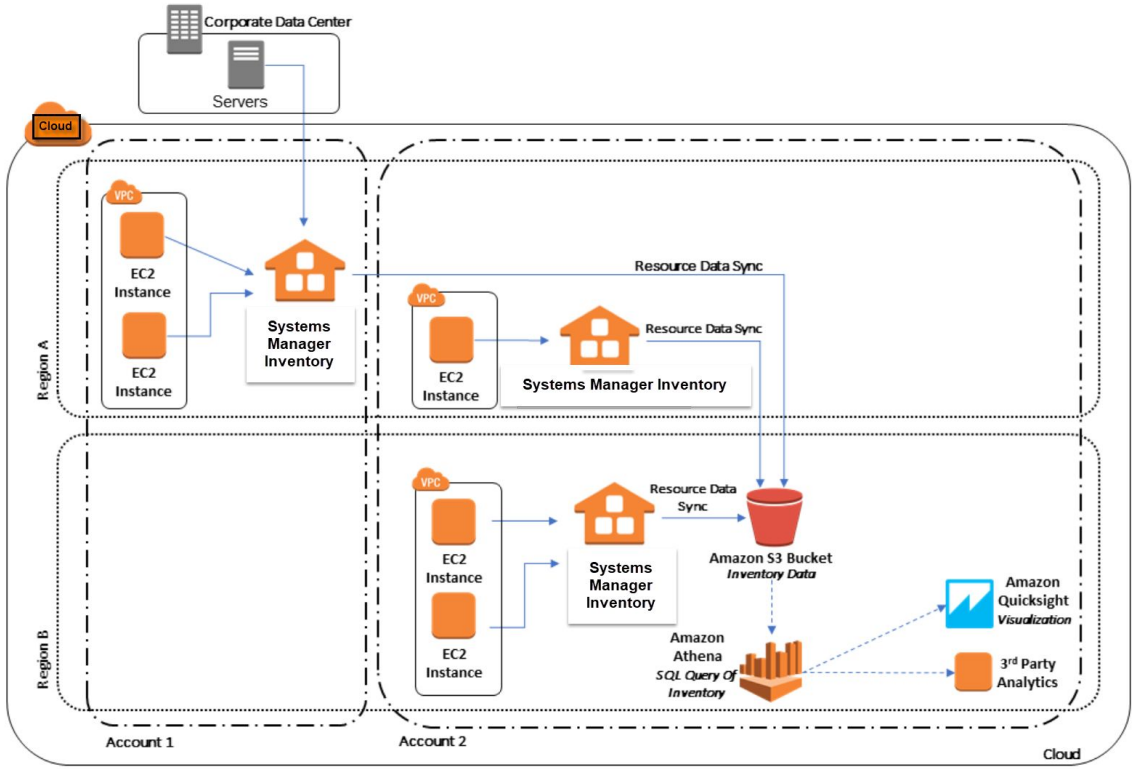


Overview of SSM Inventory

Administrator can run various queries to search for specific data based on the use-case.



Centralized Architecture



AWS Batch

Running Batch Jobs

Getting Started

A batch job is a collection, or list, of commands that are processed in sequence often without requiring user input or intervention.

Generally batch jobs accumulate during working hours, and are then executed during the evening or another time the computer is idle.

Batch jobs wait in a job queue for processing when the system has the available resources.

Overview of AWS Batch

AWS Batch enables developers, scientists, and engineers to easily and efficiently run hundreds of thousands of batch computing jobs on AWS.

AWS Batch dynamically provisions the optimal quantity and type of compute resources (e.g., CPU or memory optimized instances) based on the volume and specific resource requirements of the batch jobs submitted.

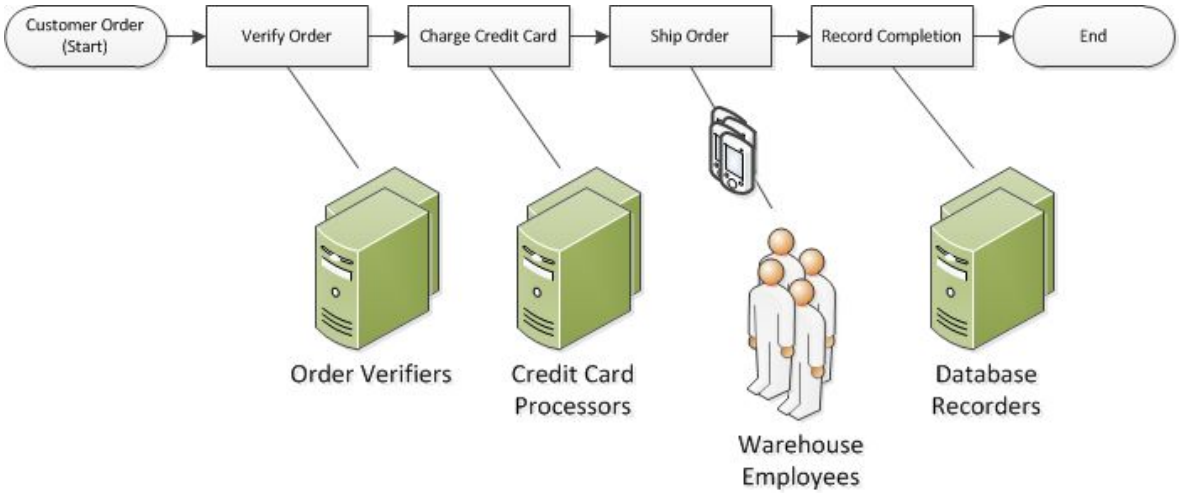
Simple Workflow Service

Workflow execution

What is a Workflow?

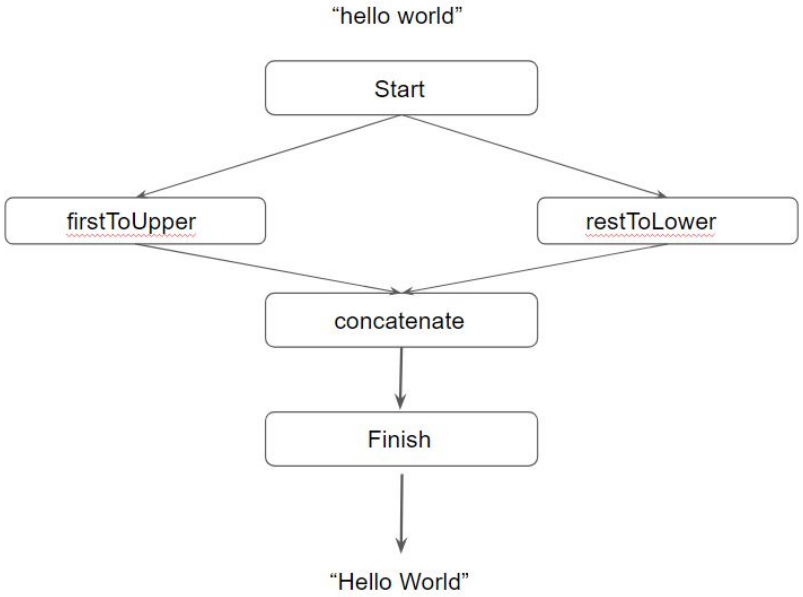
A Workflow is a sequence of tasks that processes a set of data.

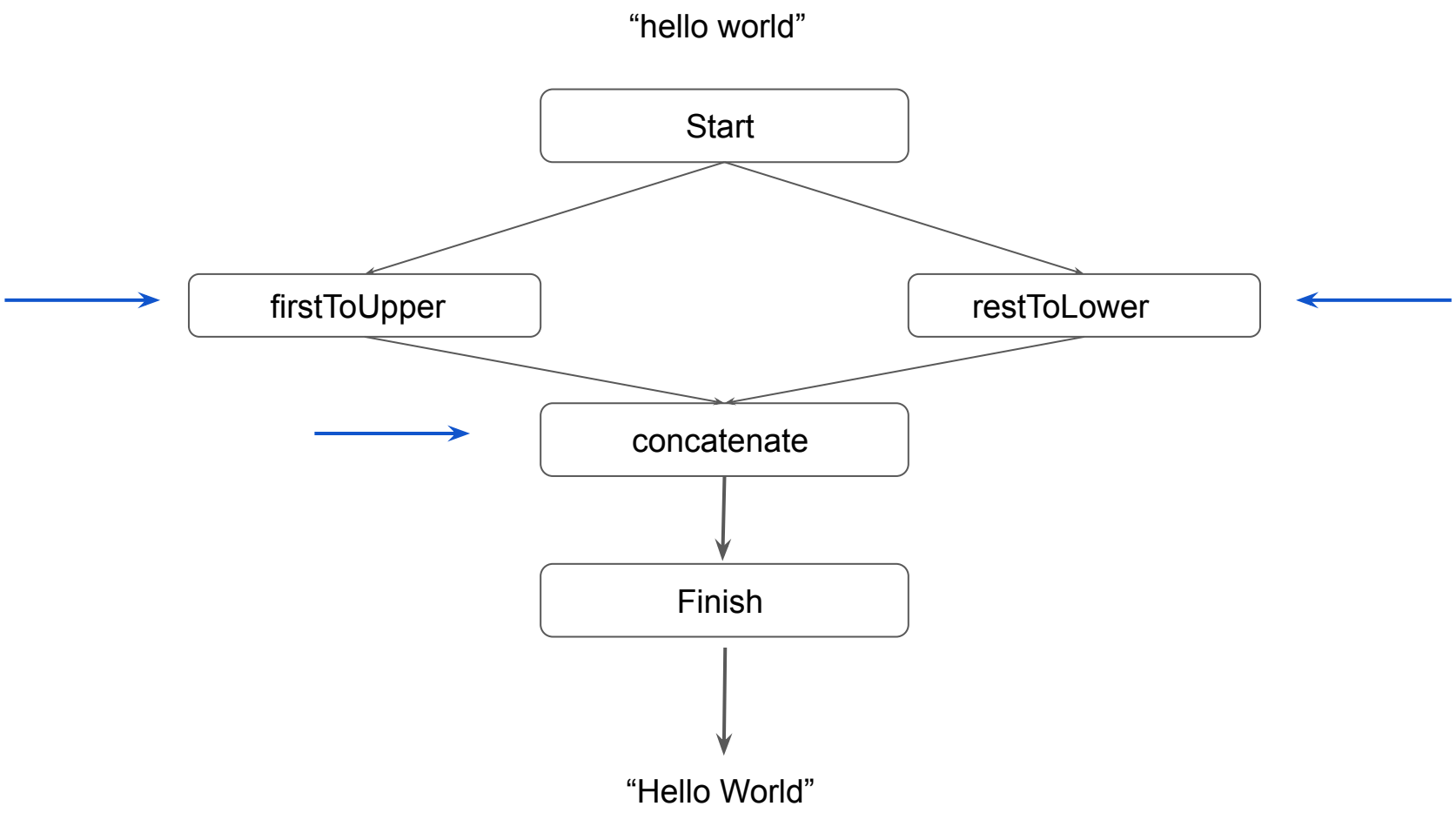
A workflow is a set of activities that carry out some objective



Overview of SWF

The Amazon Simple Workflow Service (Amazon SWF) makes it easy to build applications that coordinate work across distributed components.




“hello world”
Start
firstToUpper
restToLower
concatenate
Finish
“Hello World”

Overview of Activities

Activities are were the actual processing takes place.

```
function firstToUpper (input: String) {
    return input[0].toUpperCase();
  }
```

Overview of Decider

Defines execution order of the processes.

Flow of input/output between multiple processes.

Conditionals and Concurrency.

One decider per workflow.

AppStream 2.0

Interesting Service

Getting Started

AppStream 2.0 allows us to centrally manage our desktop application and securely deliver them to any computer.

Sample Use-Case:

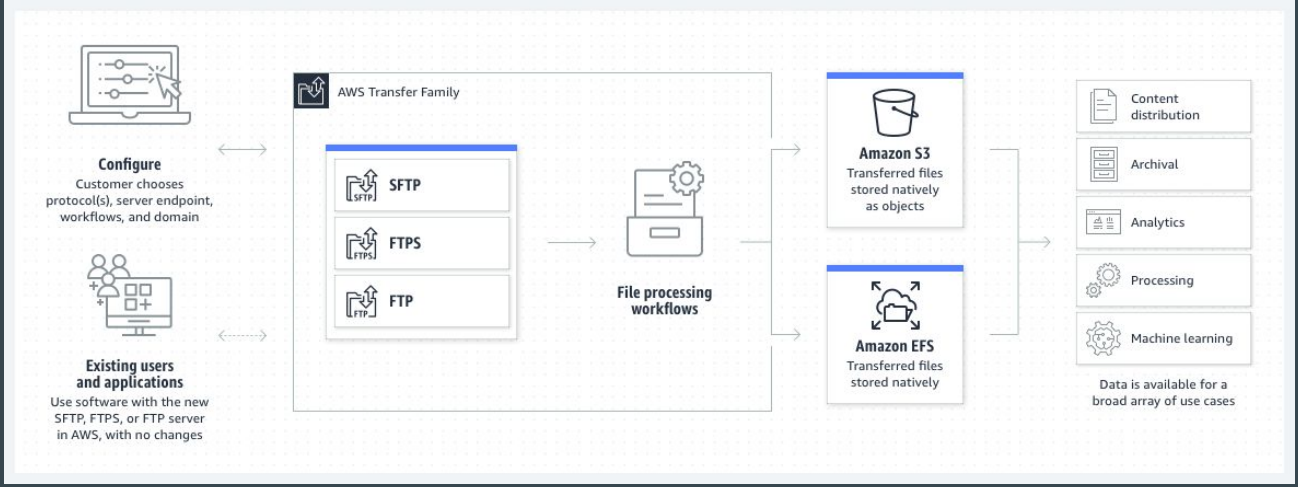
Software vendors can use AppStream 2.0 to deliver trials, demos, and training for their applications with no downloads or installations.

AWS Transfer Family



Understanding the Basics

AWS Transfer Family securely scales your recurring business-to-business file transfers to AWS Storage services using SFTP, FTPS, FTP, and AS2 protocols.

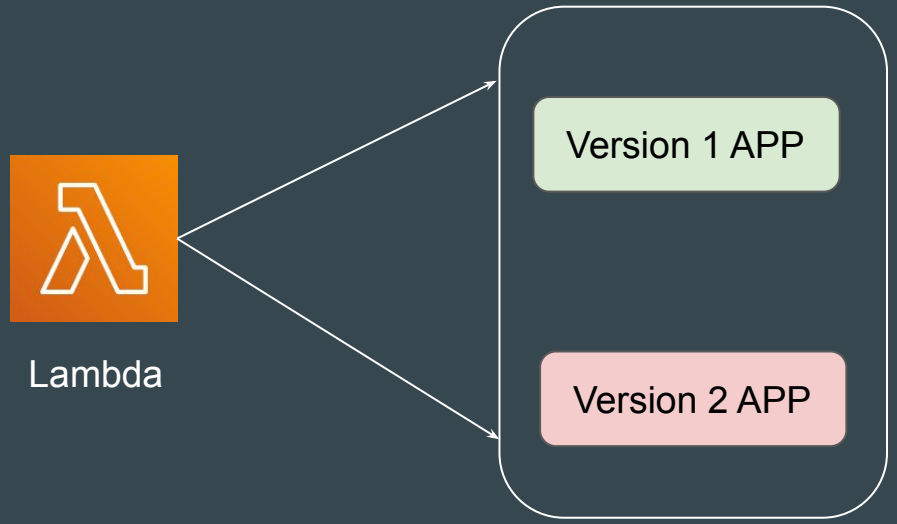


Lambda Versioning



Understanding the Basics

Lambda allows developers to host multiple versions of their code.



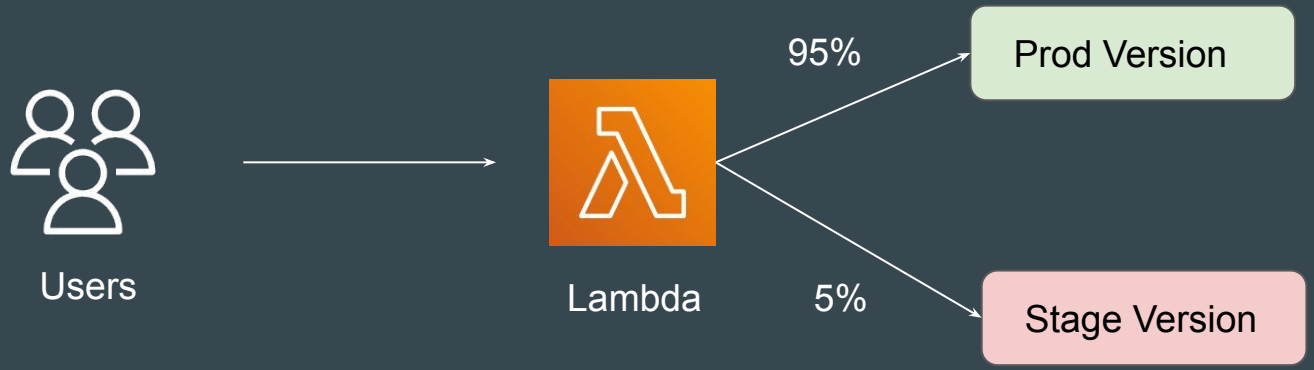
Function Version

A function version includes the following information:

1. The function code and all associated dependencies.
2. The Lambda runtime identifier and runtime version used by the function.
3. All the function settings, including the environment variables.
4. A unique Amazon Resource Name (ARN) to identify the specific version of the function.

Testing Before Prod

You can allow certain amount of traffic to a specific version of function to test before rolling out changes.



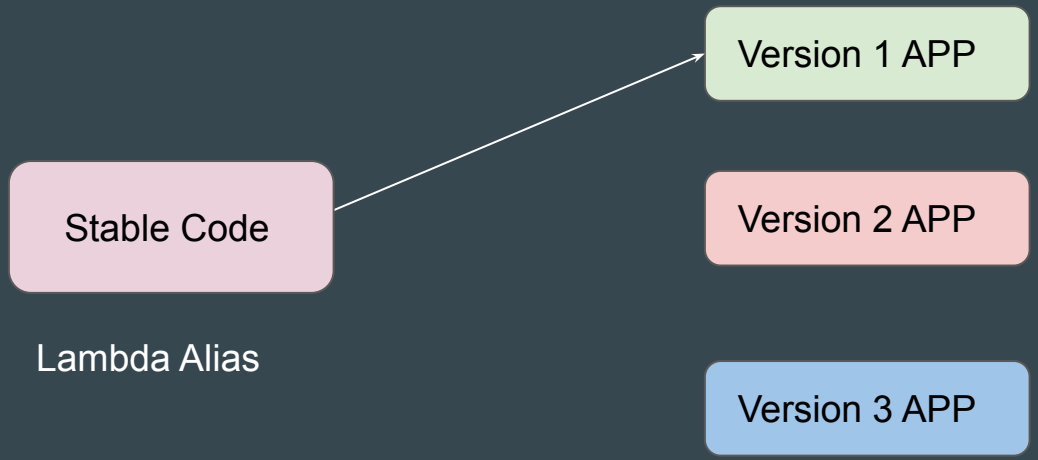
Lambda Alias



Understanding the Basics

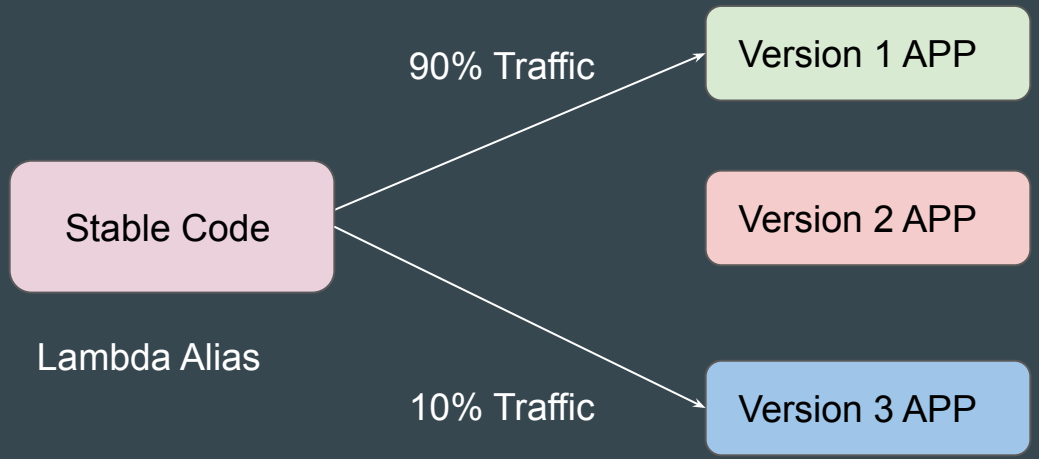
A Lambda alias is like a pointer to a specific function version.

Users can access the function version using the alias Amazon Resource Name (ARN).



Alias routing configuration

We can use routing configuration on an alias to send a portion of traffic to a second function version



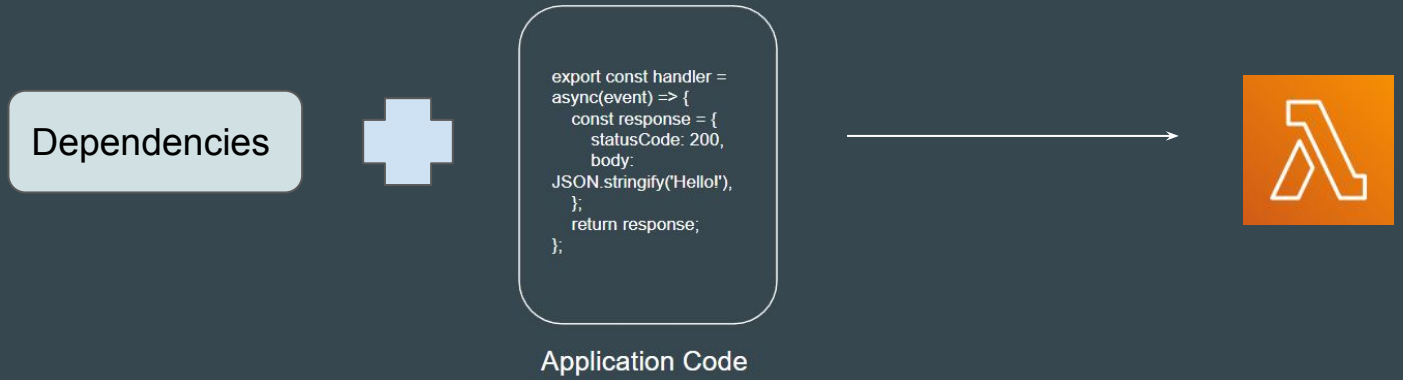
Lambda Deployment Package



Understanding the Basics

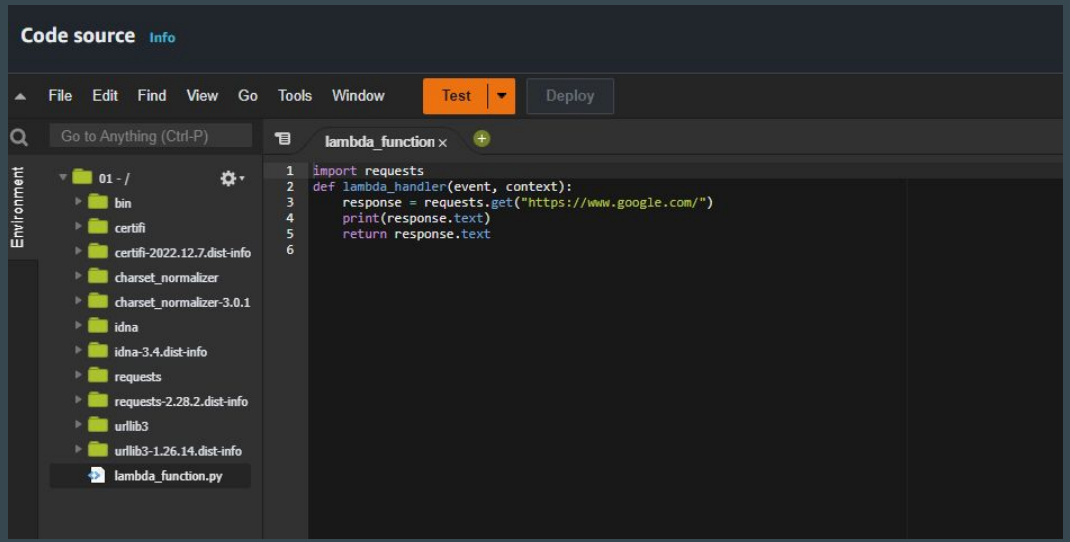
Your AWS Lambda function's code can consists of scripts or compiled programs and their dependencies

A dependency can be any package, module or other assembly dependency that is not included with the Lambda runtime environment for your function's code.



Overview of Deployment Package

The deployment package acts as the source bundle to run your function's code and dependencies (if applicable) on Lambda.



Points to Note

Lambda supports two types of deployment packages:

Container images and .zip file archives.

Containers

Zip Archives

Migrating Lambda Function



Understanding the Use-Case

You have a Lambda function that is running in AWS Account A.

You now want the exact same function in AWS Account B.

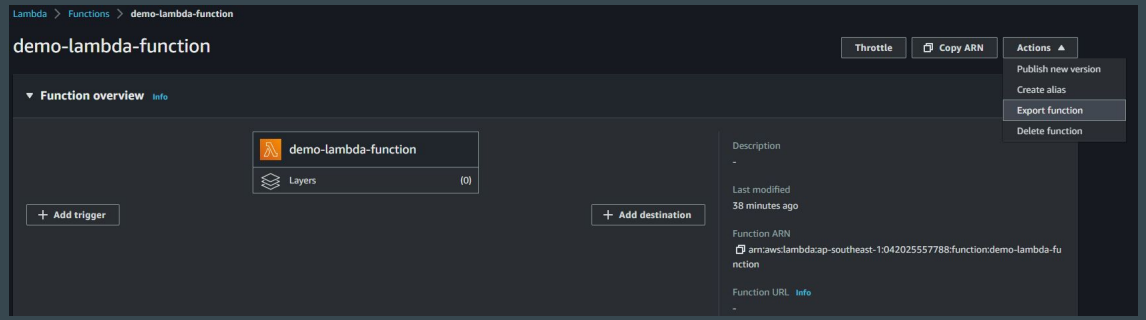


Step 1 - Exporting Function

To migrate a Lambda function to another AWS account or AWS Region using the Lambda console, do the following:

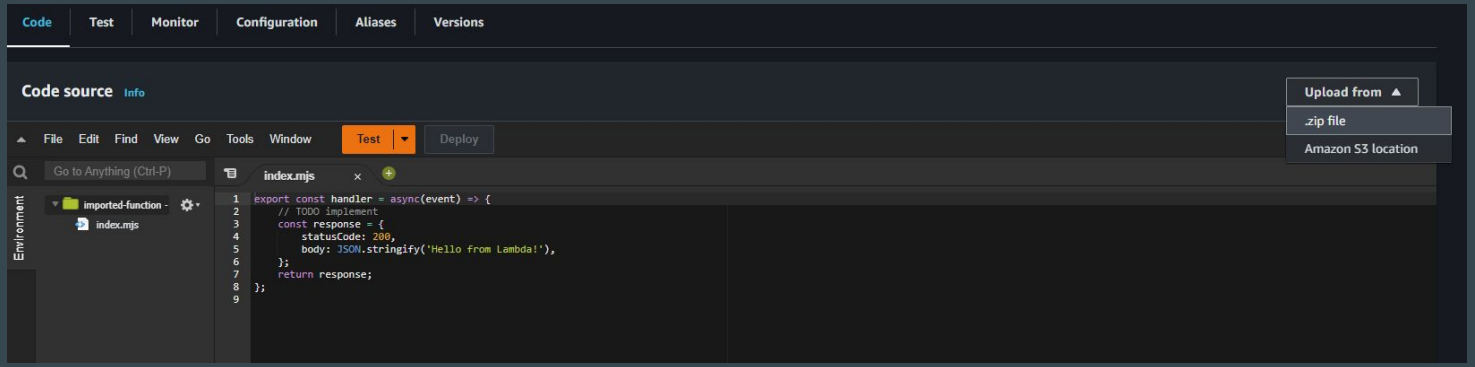
1. Download the Lambda function's deployment package.

2. Use the Lambda function's deployment package to create a new Lambda function in another AWS account or Region.



Step 2 - Importing Function

Use the “Upload From” option in the new account’s Lambda function and upload the zip file.



Points to Note

The deployment package contains only the Lambda function's code.

The rest of your function's configurations, such as timeout and memory size, must be entered manually in the console when you create the new function.

To migrate all your function's code and configurations automatically, you can use an AWS SAM file.

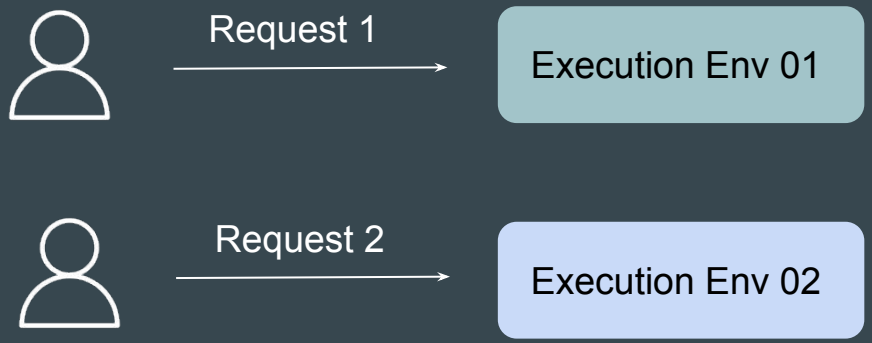
Basics of Lambda Concurrency



Understanding the Basics

Concurrency is the number of in-flight requests your AWS Lambda function is handling at the same time

For each concurrent request, Lambda provisions a separate instance of your execution environment.



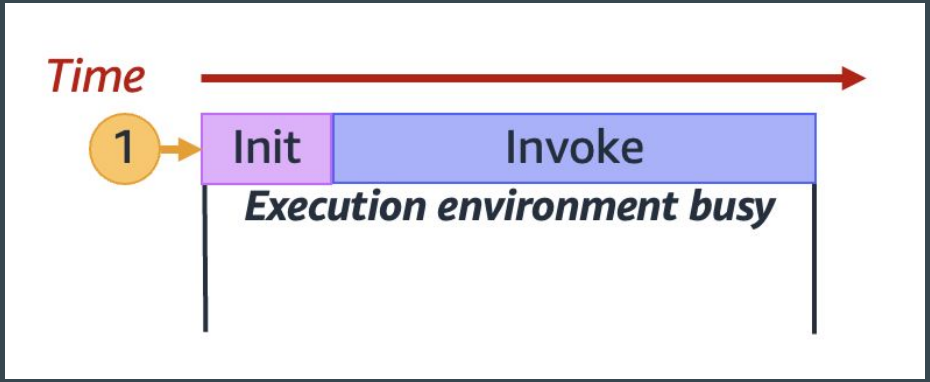
Point to Note

As your functions receive more requests, Lambda automatically handles scaling the number of execution environments until you reach your account's concurrency limit.

Understanding and visualizing concurrency

To handle a request, Lambda must first initialize an execution environment (the Init phase), before using it to invoke your function (the Invoke phase)

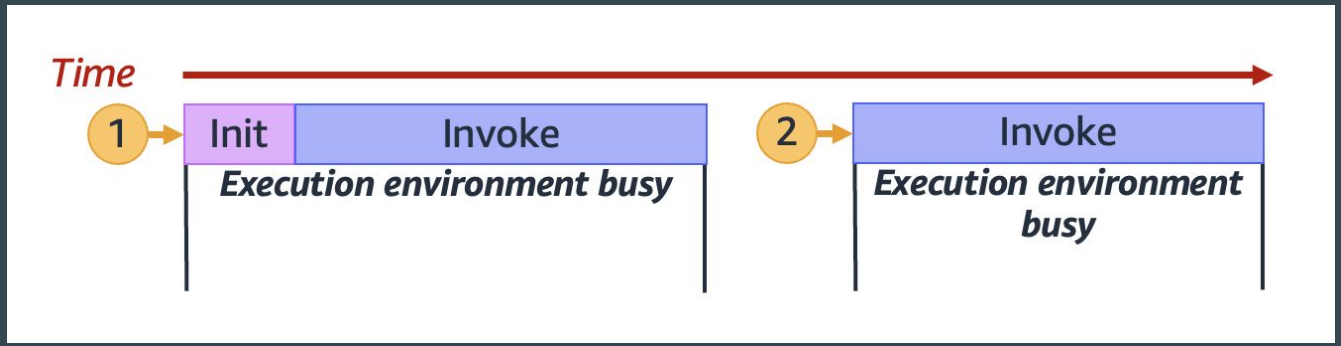
When Lambda finishes processing the first request, this execution environment can then process additional requests for the same function.



Point to Note

Lambda can reuse the same execution environment to handle the second request.

Single instance of your execution environment = Concurrency of 1



Creating Base Workflow

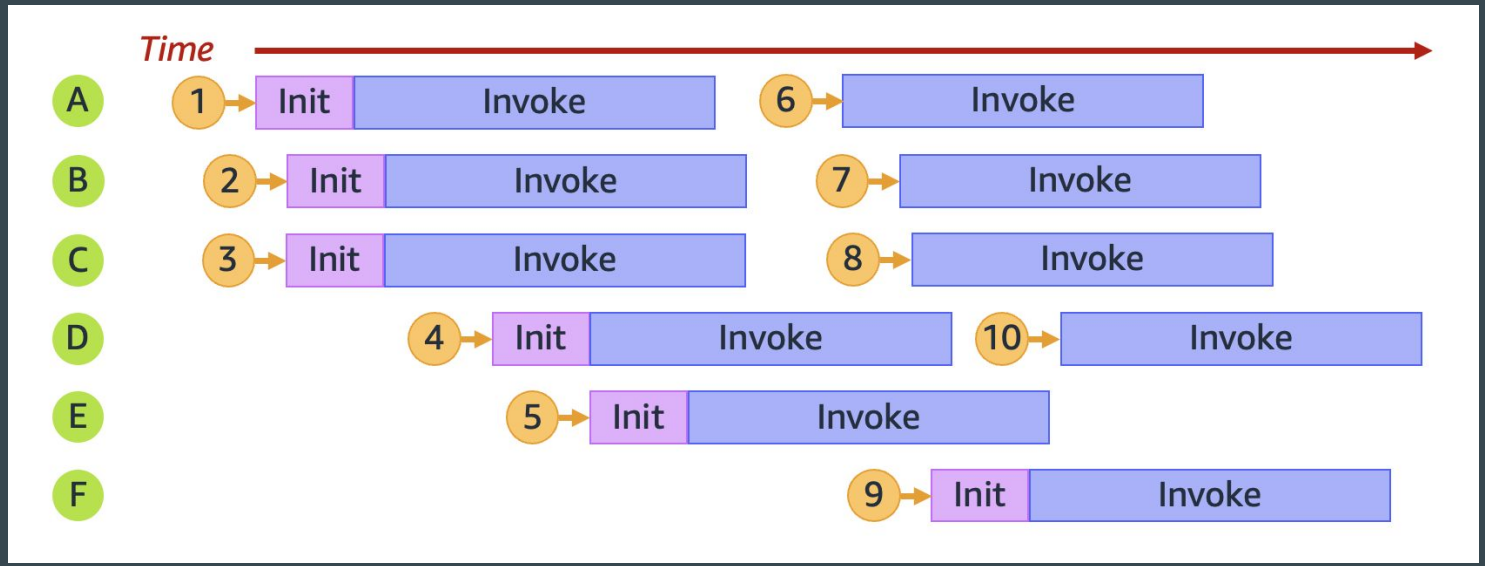
In real world scenario, Lambda may need to provision multiple execution environment instances in parallel to handle all incoming requests.

When your function receives a new request, one of two things can happen:

1. If a pre-initialized execution environment instance is available, Lambda uses it to process the request.

2. Otherwise, Lambda creates a new execution environment instance to process the request.

Sample Workflow - 10 Requests



Reserved and Provisioned Concurrency



Understanding the Basics

The default concurrency limit per AWS Region is 1,000 invocations at any given time

Your functions share this pool of 1,000 concurrency on an on-demand basis.

Your function experiences throttling (i.e. it starts to drop requests) if you run out of available concurrency.

Practical Point of View

Some of your functions might be more critical than others.

As a result, you might want to configure concurrency settings to ensure that critical functions get the concurrency they need.



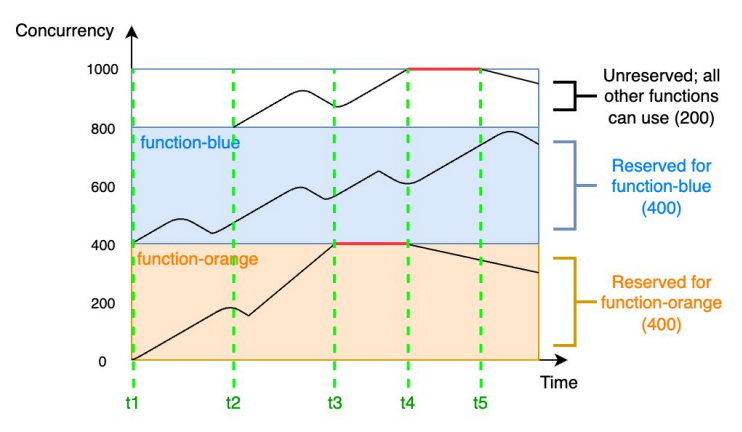
Concurrency Controls

Concurrency Control	Description
Reserved concurrency	Reserve a portion of your account's concurrency for a function. Useful if you don't want other functions taking up all the available unreserved concurrency.
Provisioned concurrency	Pre-initialize a number of environment instances for a function. Useful for reducing cold start latencies.

Reserved concurrency

If you want to guarantee that a certain amount of concurrency is available for your function at any time, use reserved concurrency.

When you dedicate reserved concurrency to a function, no other function can use that concurrency.



Challenge with Reserved Concurrency

You use reserved concurrency to define the maximum number of execution environments reserved for a Lambda function.

However, none of these environments come pre-initialized

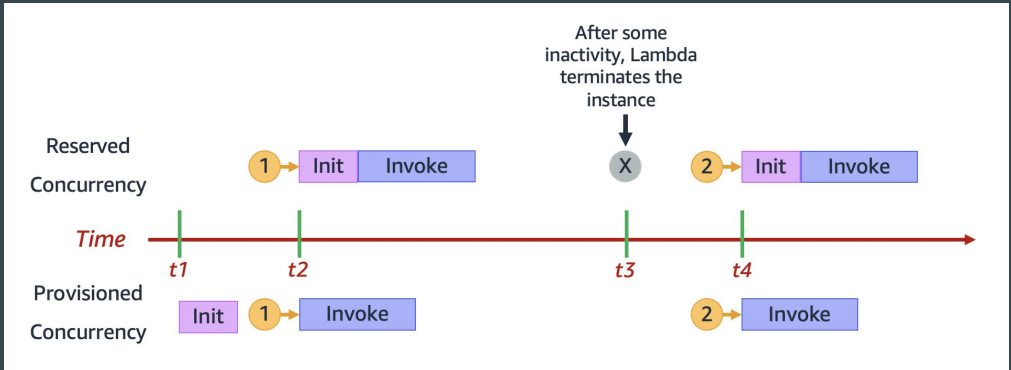
As a result, your function invocations may take longer because Lambda must first initialize the new environment before being able to use it to invoke your function

When initialization takes longer than expected, this is known as a cold start. To mitigate cold starts, you can use provisioned concurrency.

Provisioned concurrency

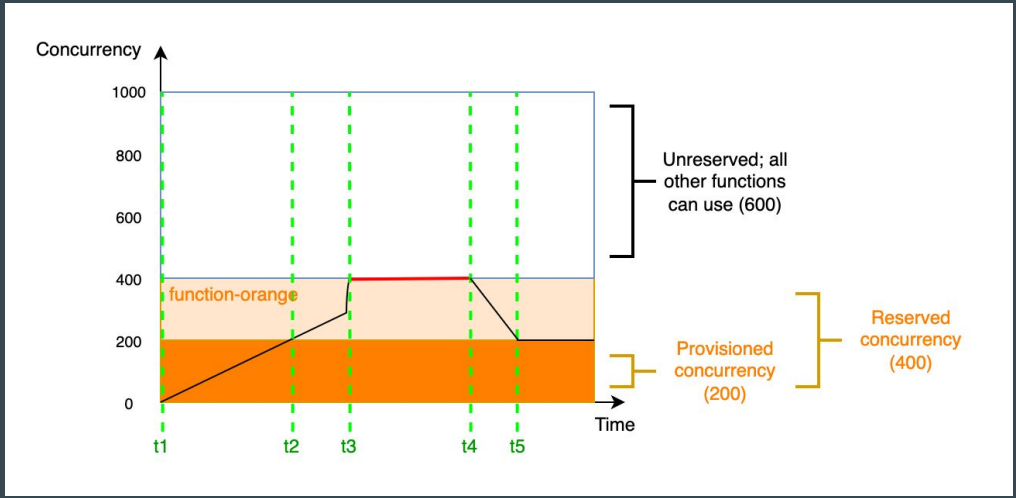
Provisioned concurrency is the number of pre-initialized execution environments you want to allocate to your function

If you set provisioned concurrency on a function, Lambda initializes that number of execution environments so that they are prepared to respond immediately to function requests.



Practical Point of View

You could use provisioned concurrency to set a baseline amount of environments to handle request during weekdays, and use reserved concurrency to handle the weekend spikes.



Comparing Reserved and Provisioned concurrency

Topic	Reserved concurrency	Provisioned concurrency
Definition	Maximum number of execution environment instances for your function.	Set number of pre-provisioned execution environment instances for your function.
Provisioning behavior	Lambda provisions new instances on an on-demand basis.	Lambda pre-provisions instances (i.e. before your function starts receiving requests).
Cold start behavior	Cold start latency possible, since Lambda must create new instances on-demand.	Cold start latency eliminated, since Lambda doesn't have to create instances on-demand.
Throttling behavior	Function throttled when reserved concurrency limit reached.	If reserved concurrency not set: function uses unreserved concurrency when provisioned concurrency limit reached. If reserved concurrency set: function throttled when reserved concurrency limit reached.
Default behavior if not set	Function uses unreserved concurrency available in your account.	Lambda doesn't pre-provision any instances. Instead, if reserved concurrency not set: function uses unreserved concurrency available in your account. If reserved concurrency set: function uses reserved concurrency.
Pricing	No additional charge.	Incurs additional charges.

Lambda Layers

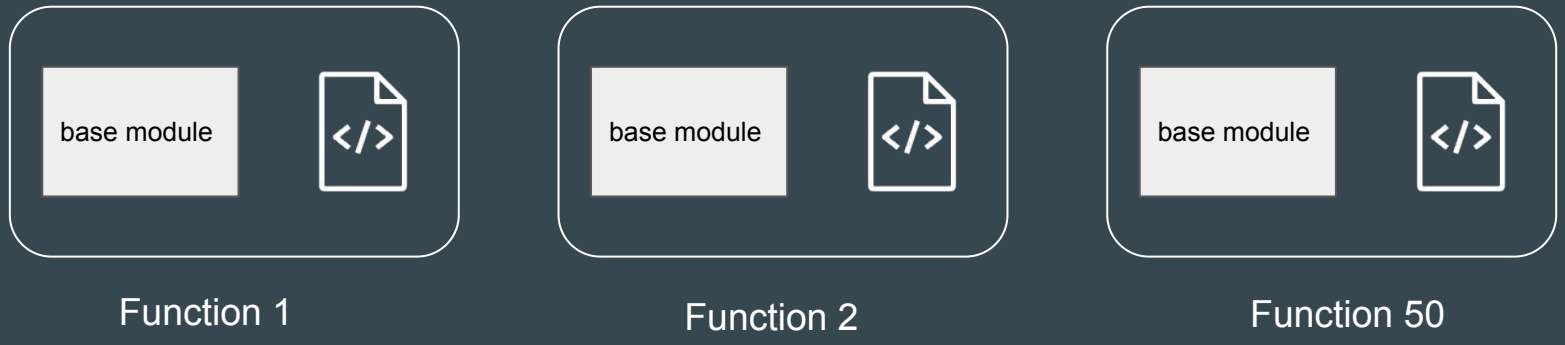


Understanding the Challenge

Let us assume there are 50 Lambda functions created based on Python.

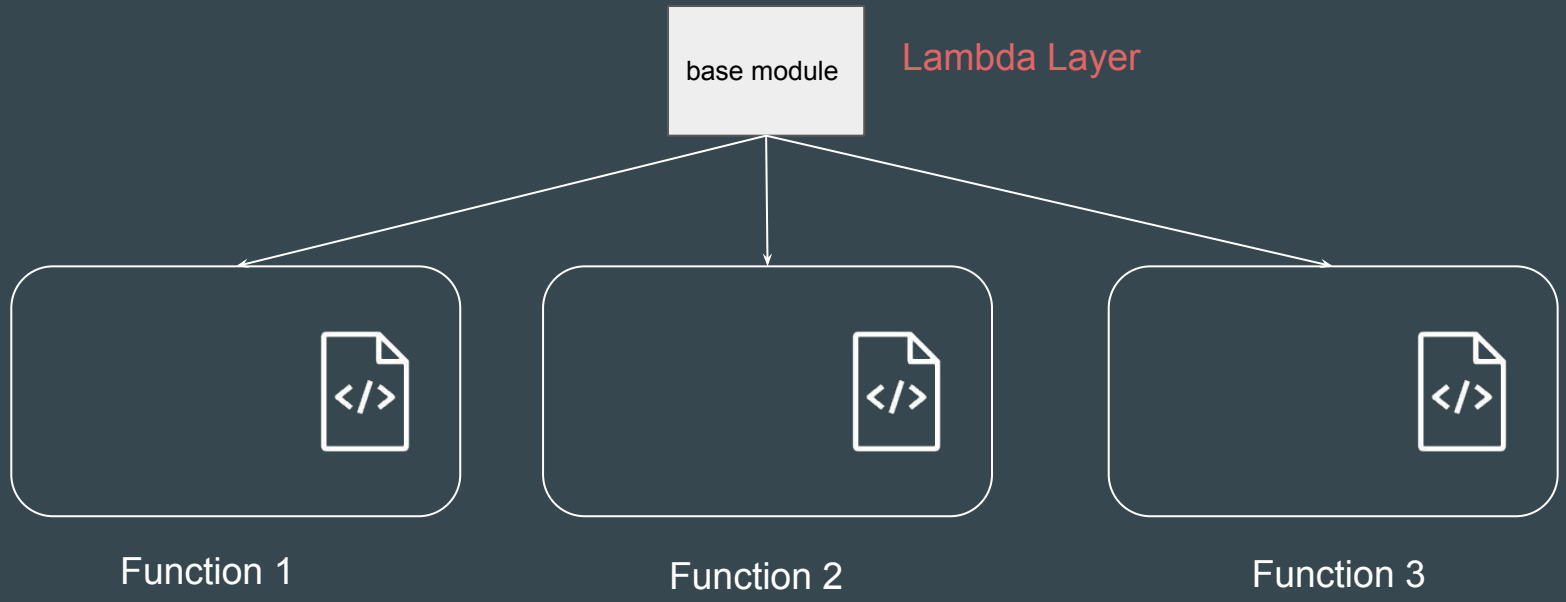
All 50 function uses same set of base libraries.

Challenge: Large Size Deployment Package, Update Difficult, Deployment Time



Understanding the Basics

Lambda layers provide a convenient way to package libraries and other dependencies that you can use with your Lambda functions.



Layer Path for Lambda Runtime

Layer paths for each Lambda runtime	
Runtime	**Path**
Node.js	`nodejs/node_modules`
	`nodejs/node14/node_modules` (`NODE_PATH`)
	`nodejs/node16/node_modules` (`NODE_PATH`)
	`nodejs/node18/node_modules` (`NODE_PATH`)
Python	`python`
	`python/lib/python3.9/site-packages` (site directories)
Java	`java/lib` (`CLASSPATH`)
Ruby	`ruby/gems/2.7.0` (`GEM_PATH`)
	`ruby/lib` (`RUBYLIB`)
All runtimes	`bin` (`PATH`)
	`lib` (`LD_LIBRARY_PATH`)

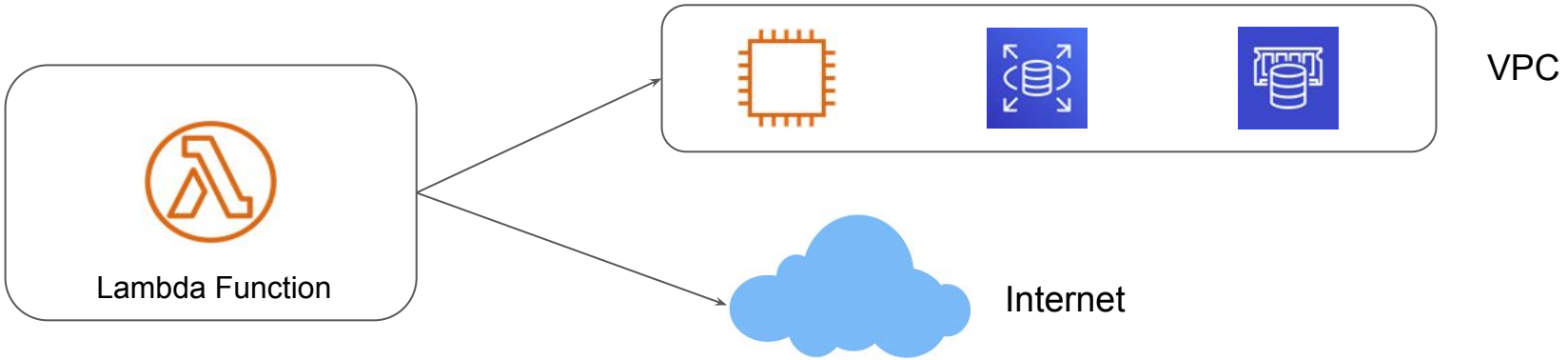
Connectivity Features of Lambda

Networking Again!

Connectivity Features

There can be scenarios where Lambda function needs to connect to EC2 instances, RDS databases which are running inside the VPC (private subnet)

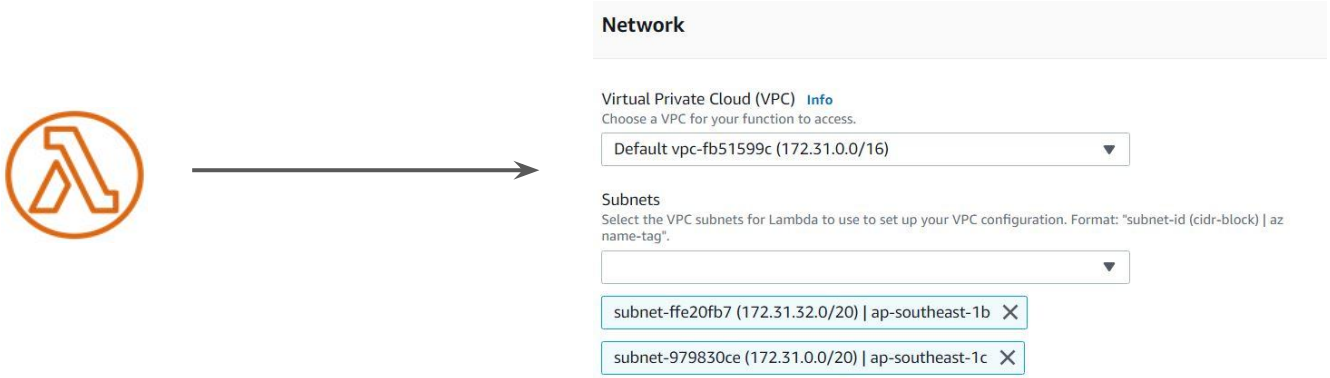
Lambda can connect to both the resources inside the VPC and Public Resources.



Lambda and VPC

By default, when your Lambda function is not configured to connect to your own VPCs, the function can access anything available on the public internet

However we can configure Lambda to connect to VPC by associate VPC and subnets with the function.



Important Point to Note

When launching in a specific subnet in VPC, make sure that NAT gateway is attached in-case if you need internet access to the Lambda function.

Lambda can also connect to AWS services like SQS via Private Link (VPC Endpoints)

If Lambda wants to connect to SQS to perform certain operations, appropriate IAM role will be needed.

If launched in VPC, you need to assign appropriate IAM role so that Lambda can perform certain operations like creating/deleting network interfaces.

Content Delivery Network

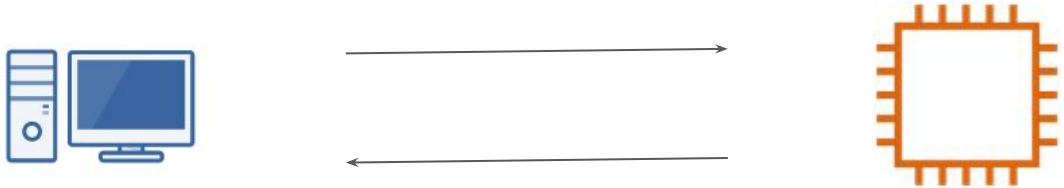
CDN is Awesome

Generic Scenario

Let's consider a typical scenario where everything is hosted in a single server.

On a smaller scale this seems to be an ideal approach.

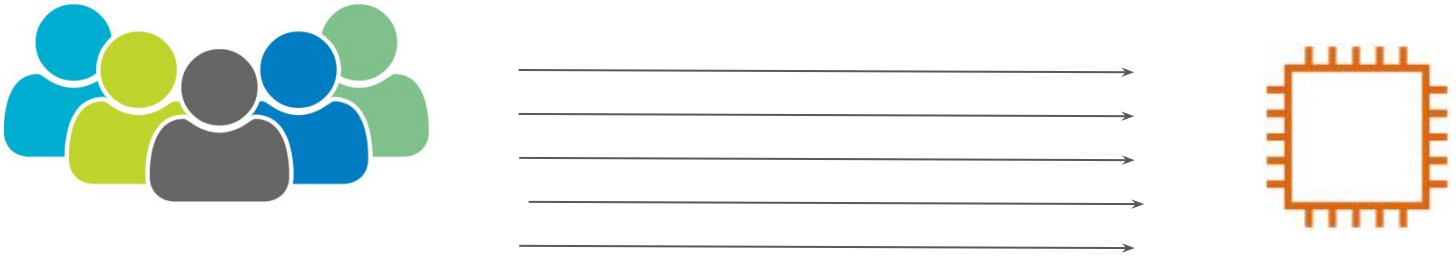
But when the traffic and popularity grows, there are a lot of challenges.



Challenge 1 - Performance

With an increase in number of visitors, the performance can go down.

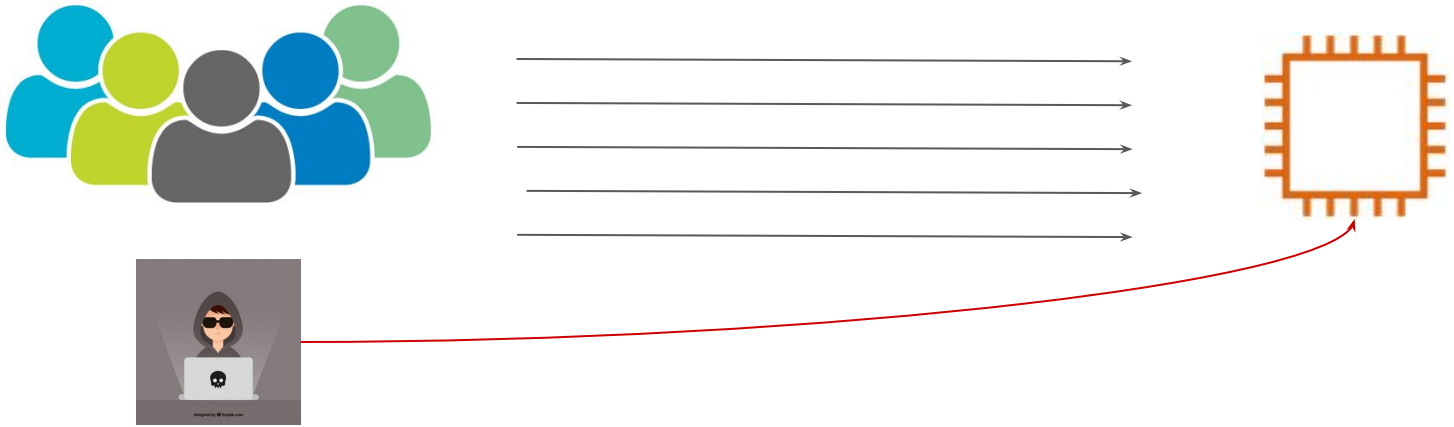
If website has 1 image and 1000 users are visiting it, then the same single image will need to be sent to 1000 users.



Challenge 2 - Security

Attackers love the Internet.

A typical website and web-application face various type of attacks ranging from DOS, Web-Application attacks and so on.



Typical Solution

Some of the approaches to the challenges that we discussed:

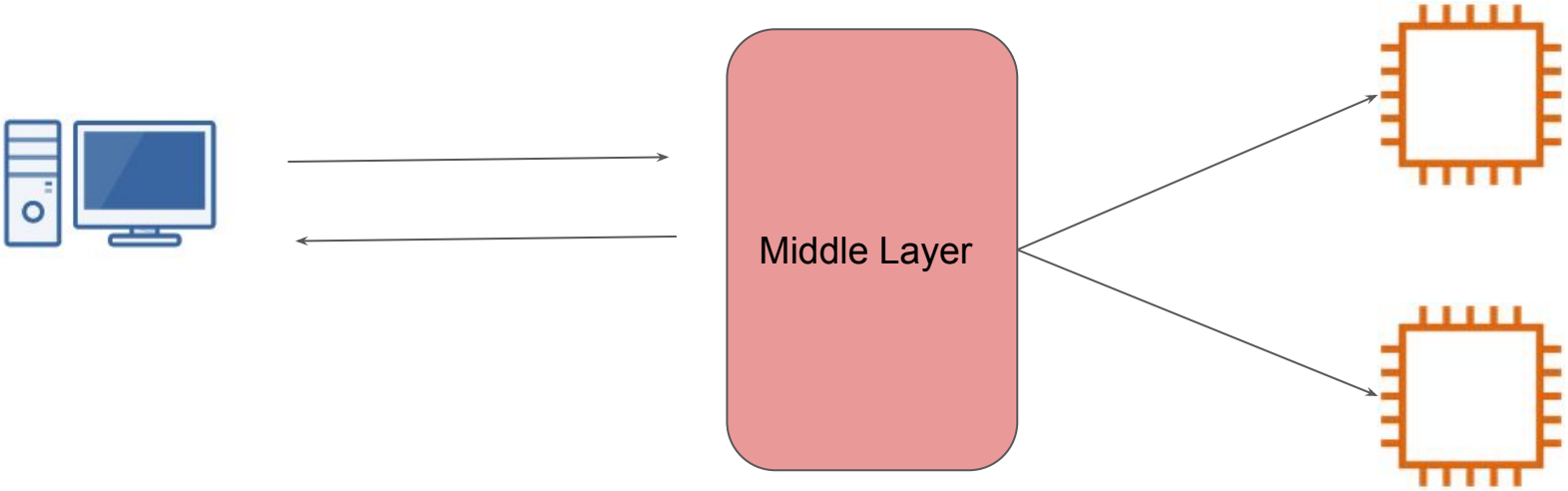
1. Increase the size of server / increase number of servers for better performance.

2. Configure DDoS protection, Web-Application Firewall etc at the server level.

Doing these things on each server is a tedious task and it cannot scale very well.

Better Architecture

Better architecture would be to introduce a middle layer that has all functionalities related to protecting against attacks, caching of commonly requested objects for better performance.



Content Delivery Network

A CDN acts as a proxy that receives the request and then forwards it to the backend systems.

Various CDN's also comes with features like DDoS Protection, WAF, Cachig and others.

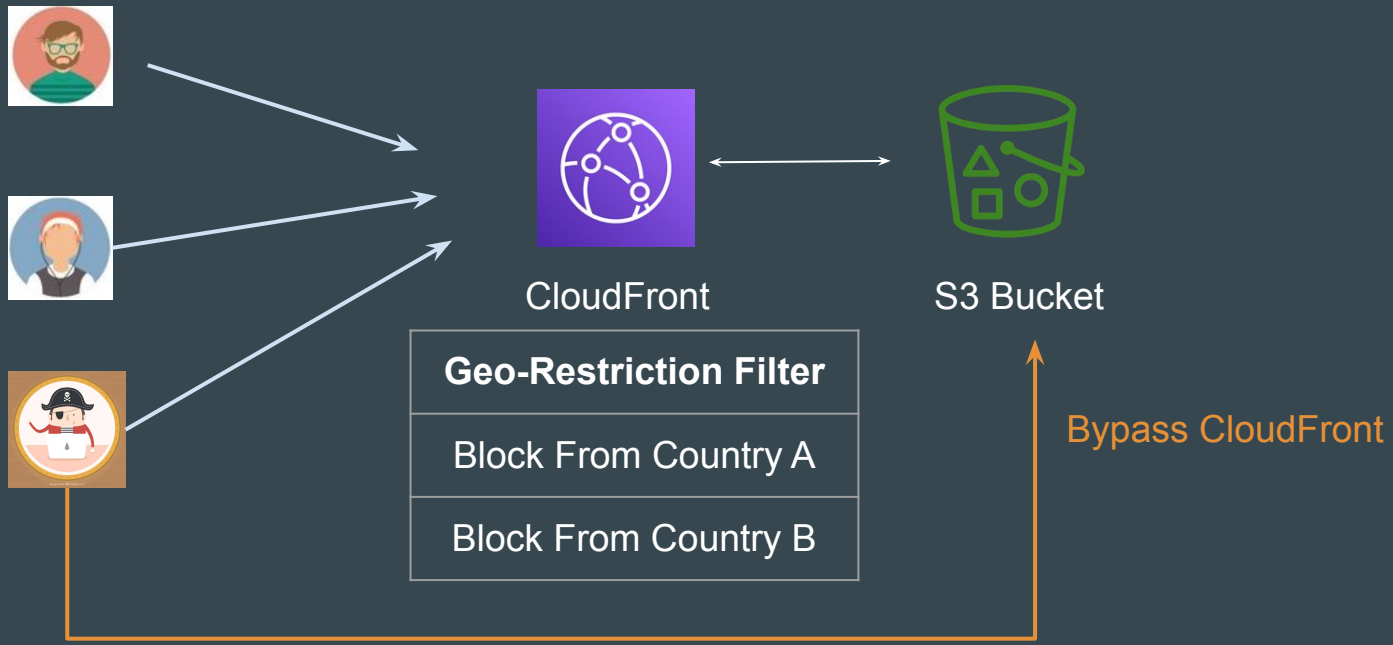


Origin Access Identity



Understanding the Challenge

Security measures applied at Cloudfront can easily be bypassed if attacker sends queries directly to the origin.



CloudFront Origin Access Identity

CloudFront Origin Access Identity ensures that only users coming through CloudFront distribution are able to access the contents of your S3 Buckets.



Join us in our Adventure

Be Awesome



kplabs.in/chat



kplabs.in/linkedin

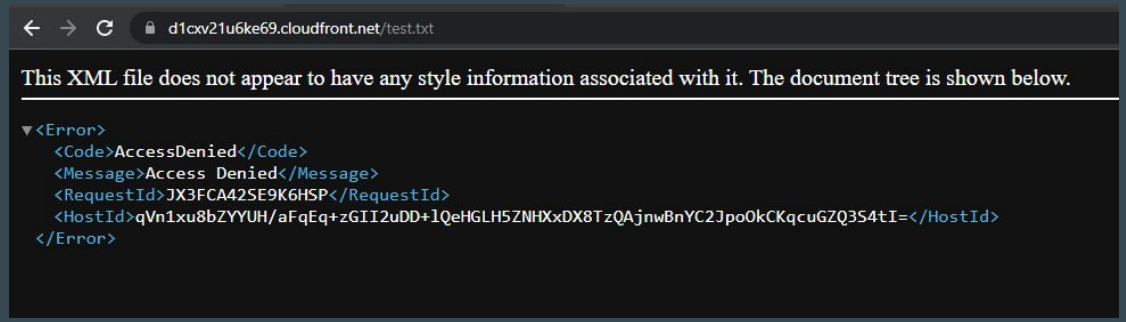
Custom Error Pages - CloudFront



Understanding the Challenge

The status code and the error page that CloudFront receives from the origin is also returned to the viewer.

This will not create a good experience for the user accessing the website.



Generating Custom Error Response

You can configure CloudFront to return a custom error response to the viewer that is different from the origin response.



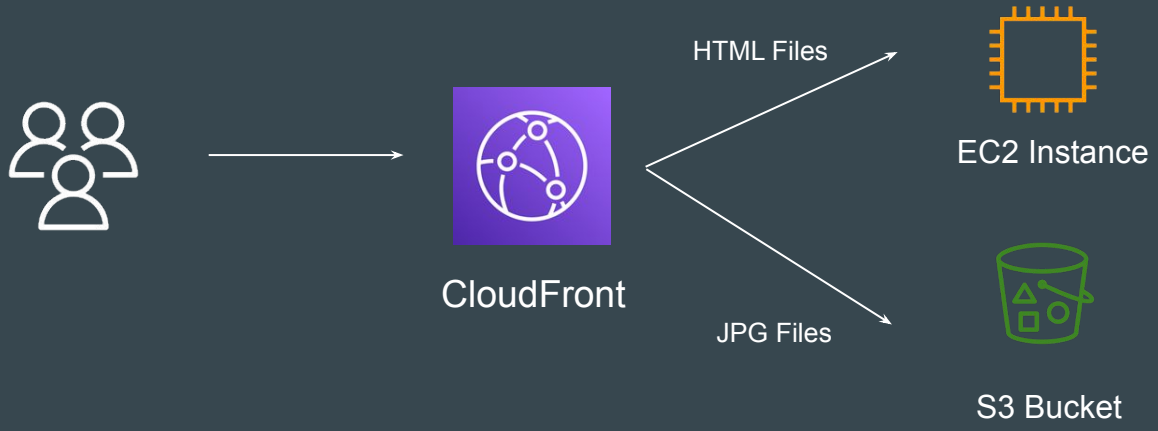
Multiple Origins Configuration - CloudFront



Understanding the Basics

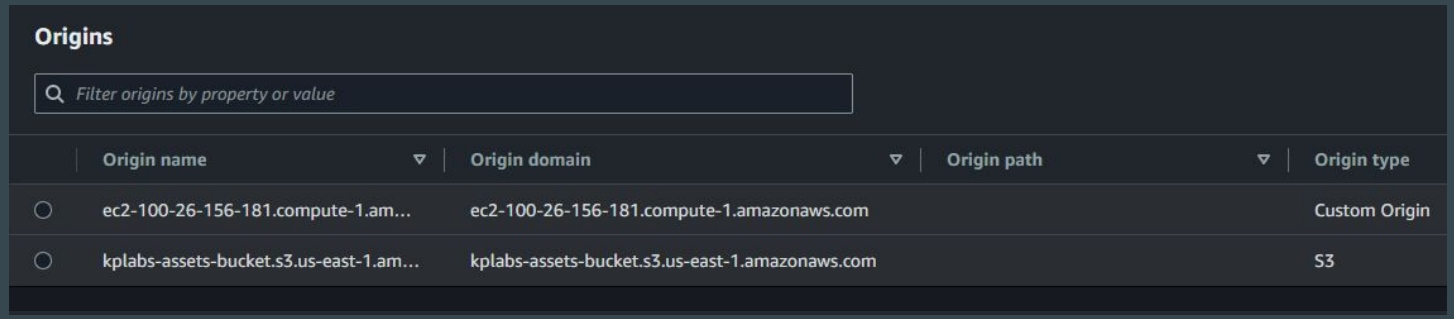
You can configure a single CloudFront web distribution to serve different types of requests from multiple origins.

For example, your website might serve static content from an Amazon Simple Storage Service (Amazon S3) bucket and dynamic content from EC2.



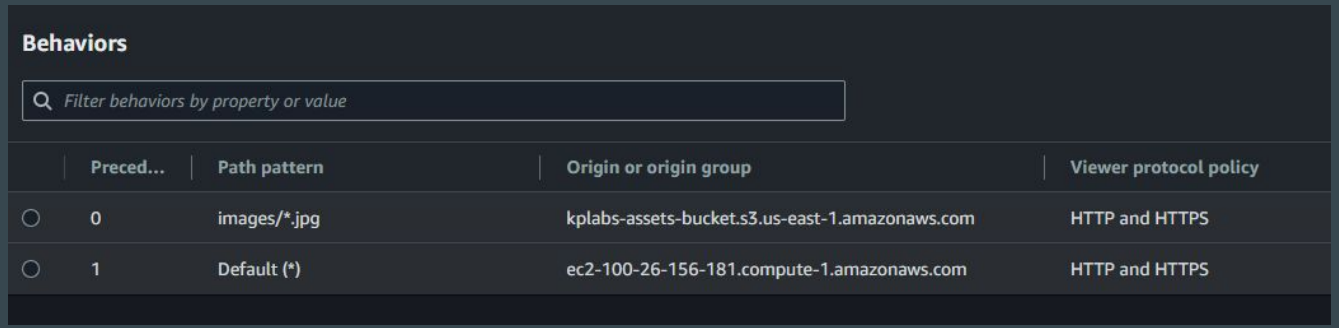
Practical Approach - Step 1

1. You need to have two origins (EC2 and S3 Bucket)



Practical Approach - Step 2

2. Create a behavior that specifies a path pattern to route all static content requests to the S3 bucket

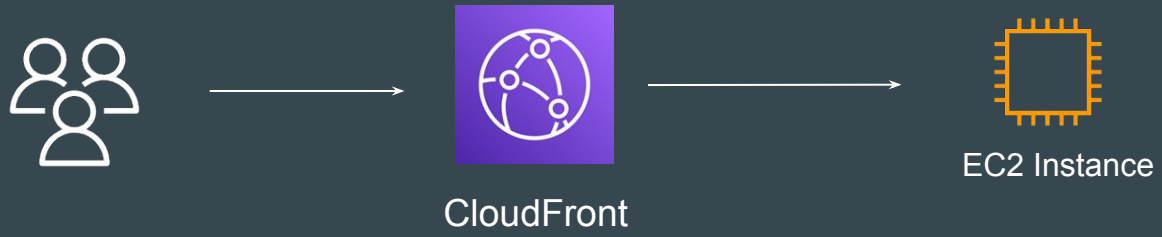


High-Availability with CloudFront Origin FailOver



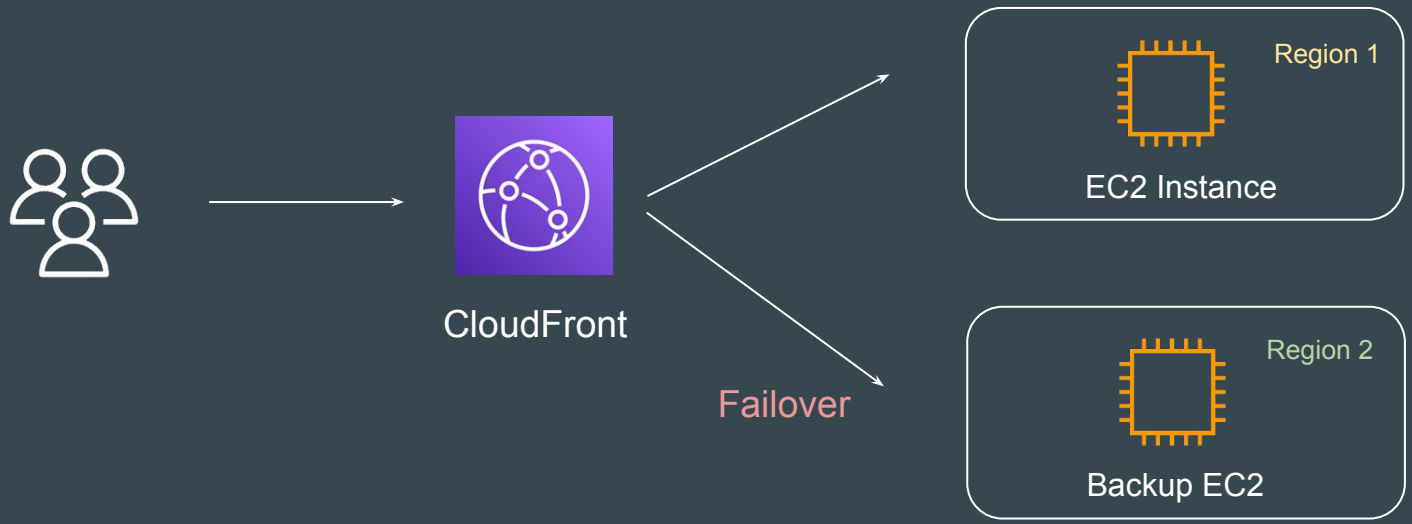
Understanding the Challenge

If the CloudFront origin is down, then it can impact the production environment.



Achieving High-Availability

You can set up CloudFront with origin failover for scenarios that require high availability.



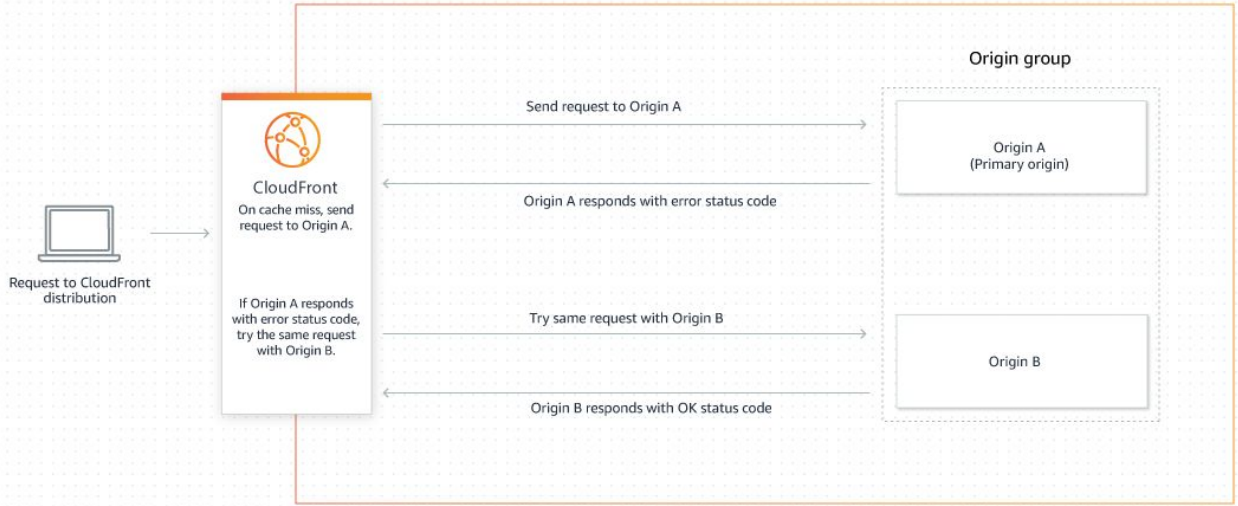
Points to Note

When there’s a cache hit, CloudFront returns the requested object.

When there’s a cache miss, CloudFront routes the request to the primary origin in the origin group.

When the primary origin returns a status code that is not configured for failover, such as an HTTP 5xx or CloudFront fails to connect to the origin, then CloudFront routes the request to the secondary origin in the origin group.

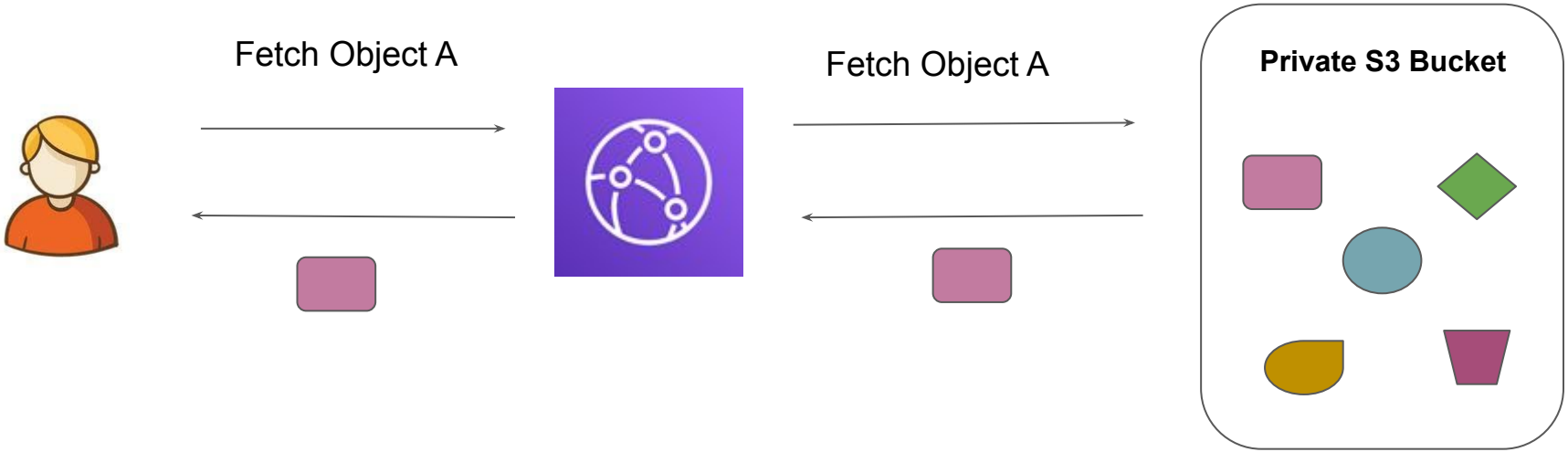
Figure 1 : Origin failover on cache miss.



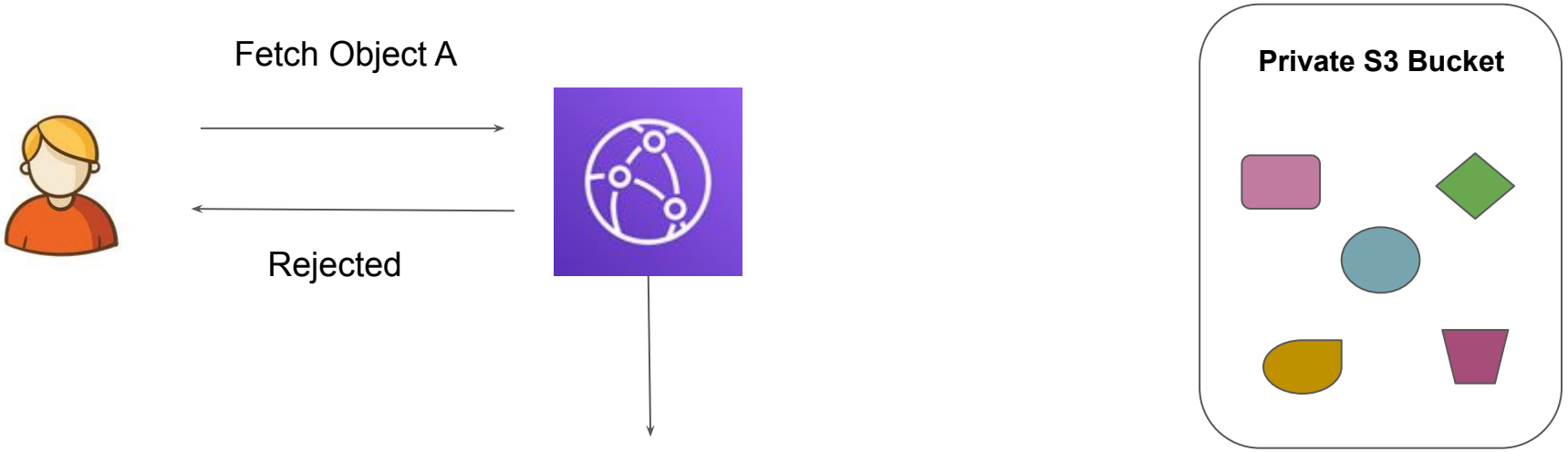
CloudFront Signed URLs

CDN is Awesome

Generic Approach

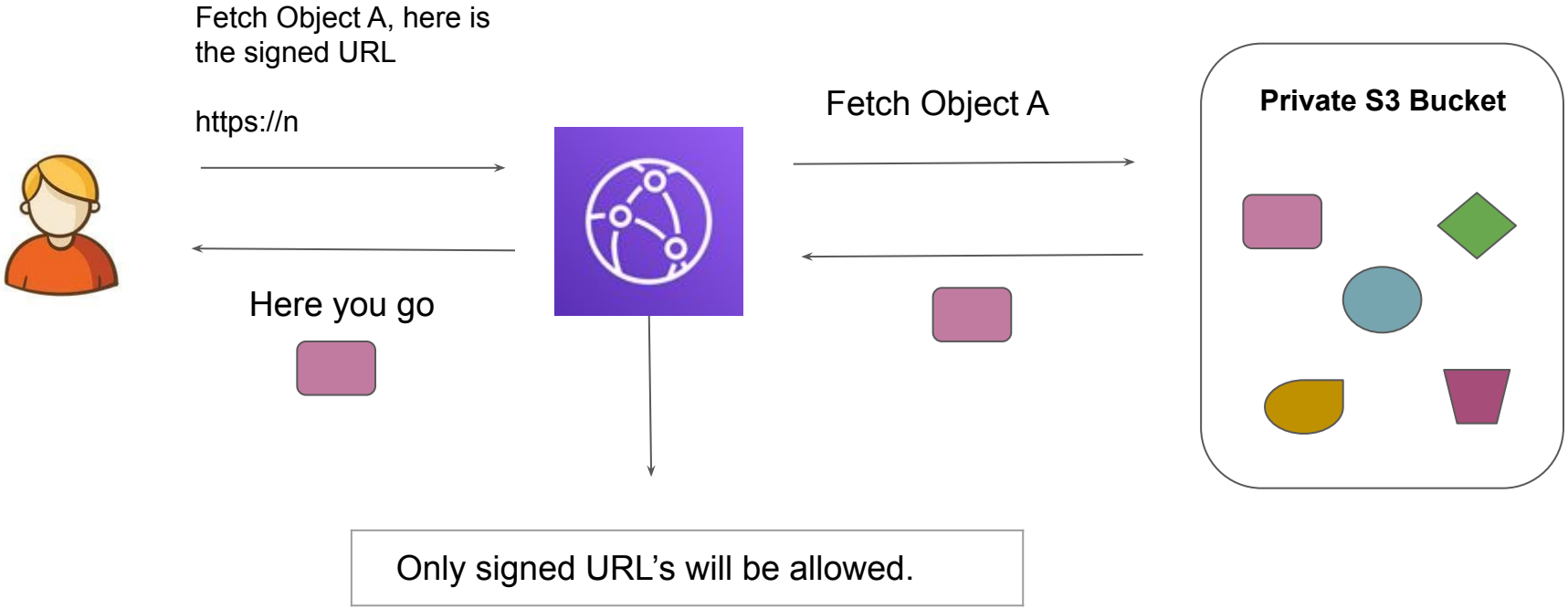


Allow only special URLs



Only special URL's will be allowed.

Architecture Overview of Signed URLs



CloudFront Signed URLs

CloudFront Signed URLs mandates users to provide signed URLs or signed cookies to access the private content.

CloudFront signed URLs can be generated by the trusted signers assigned in your AWS account.

Lambda@Edge

Running Serverless at the Edge

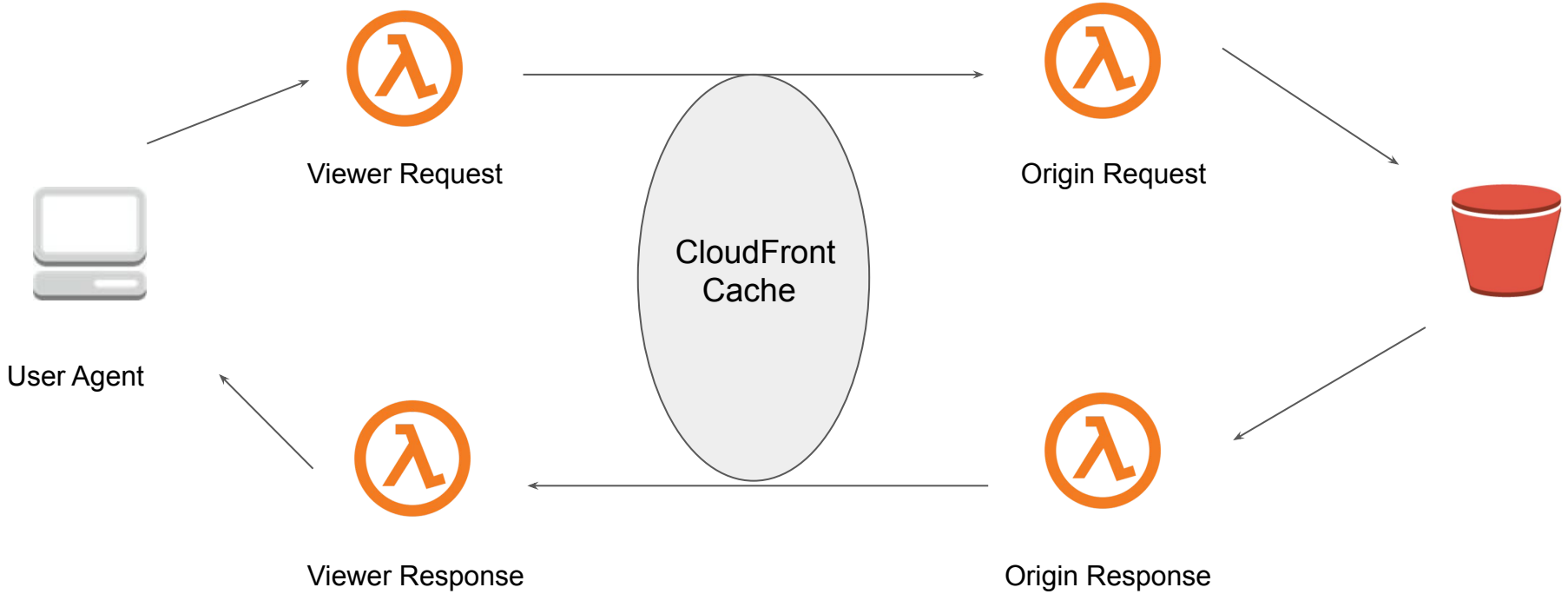
Getting started

Lambda@Edge lets you run Lambda functions to customize content that CloudFront delivers.

You can use Lambda functions to change CloudFront requests and responses at the following points:

1. After CloudFront receives a request from a viewer (viewer request)
2. Before CloudFront forwards the request to the origin (origin request)
3. After CloudFront receives the response from the origin (origin response)
4. Before CloudFront forwards the response to the viewer (viewer response)

Diagrammatic Representation



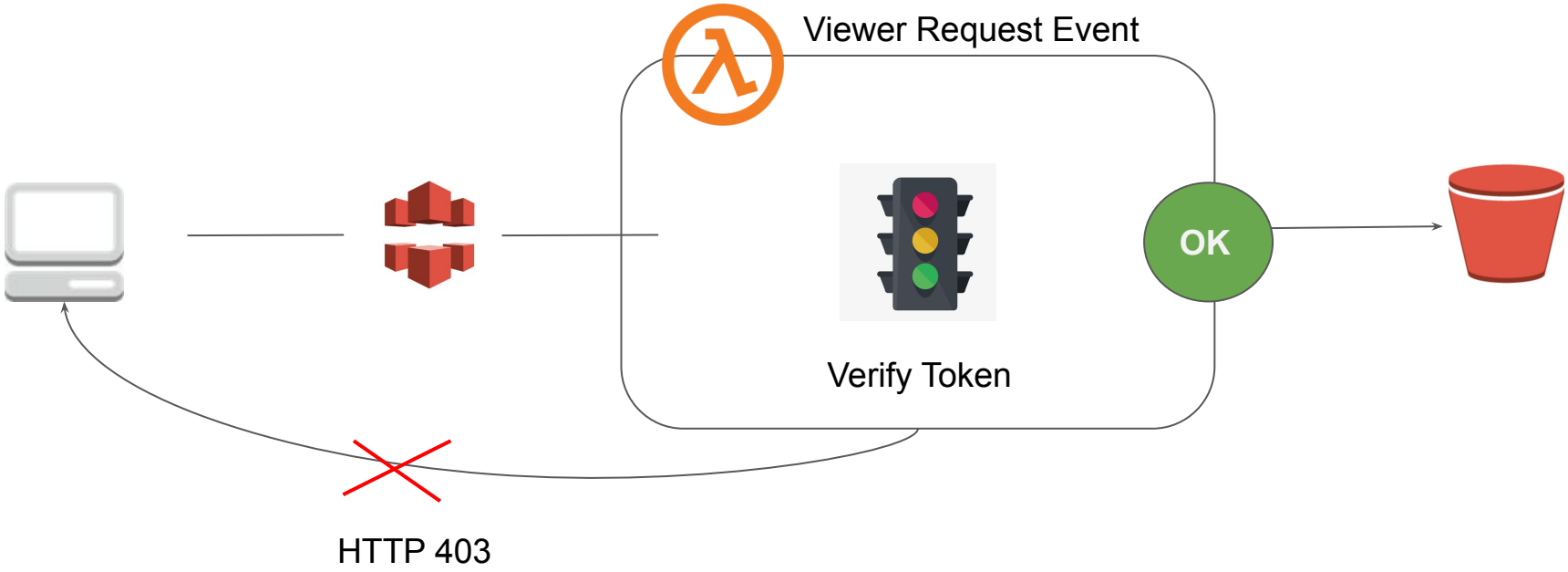
Viewer Request

Viewer Request is executed on every request before CloudFront cache is checked.

There are various things that we can do at this stage, like:

- Modify URLs, cookies query strings etc.
- Perform Authentication and Authorization Checks.

Viewer Request



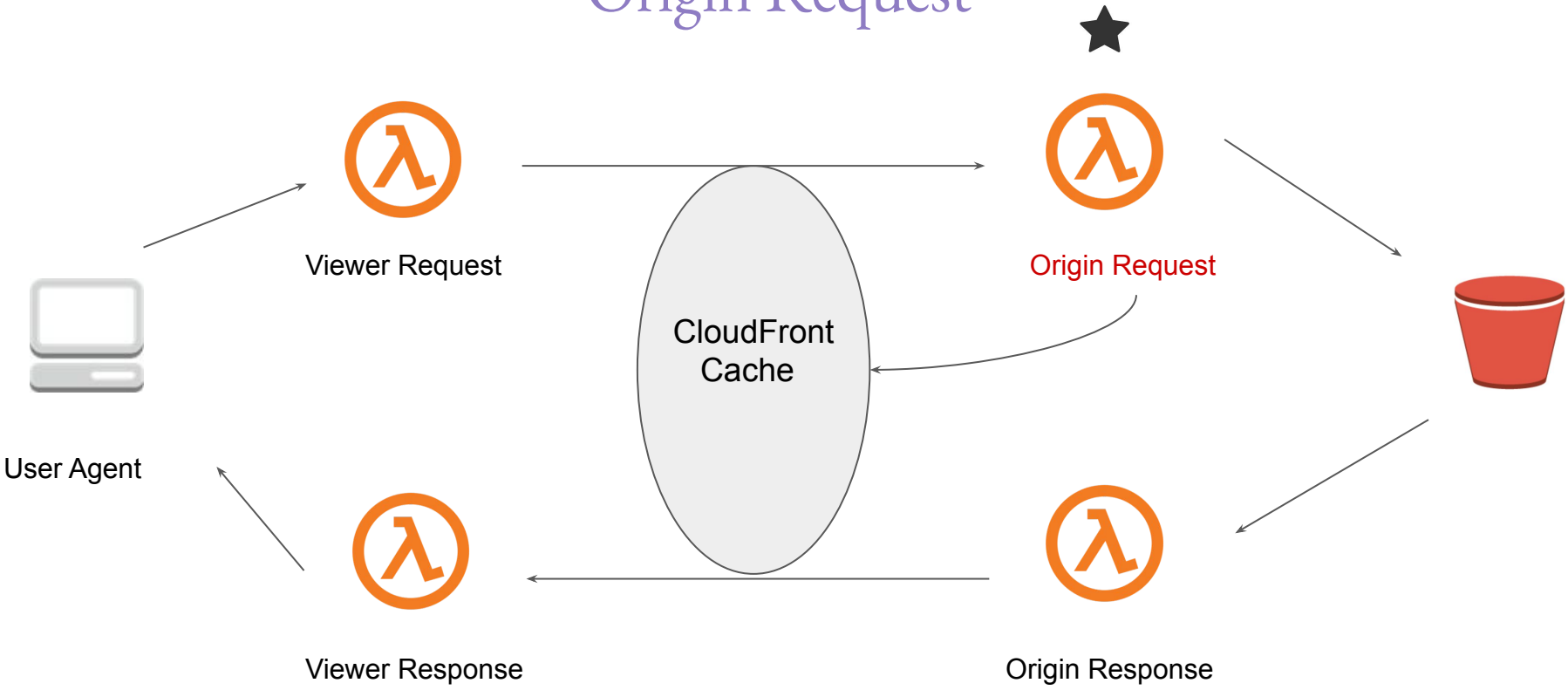
Origin Request

Executed on cache miss, before a request is forwarded to the origin.

There are various things that we can do at this stage, like:

- Dynamically select origin based on the request headers

Origin Request



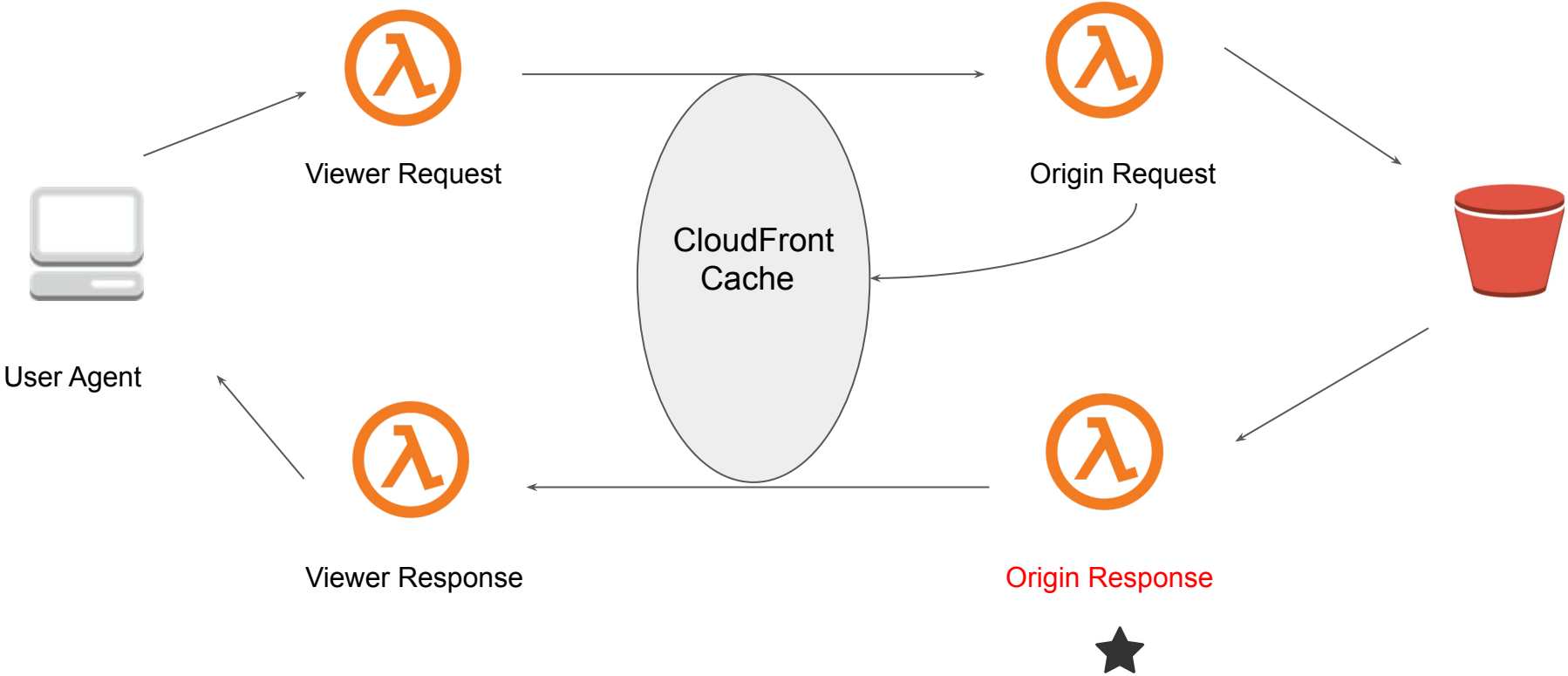
Origin Response

Executed on a cache miss, after a response is received from the origin.

There are various things that we can do at this stage, like:

- Modify the response headers.
- Intercept and replace various 4XX and 5XX errors from the origin.

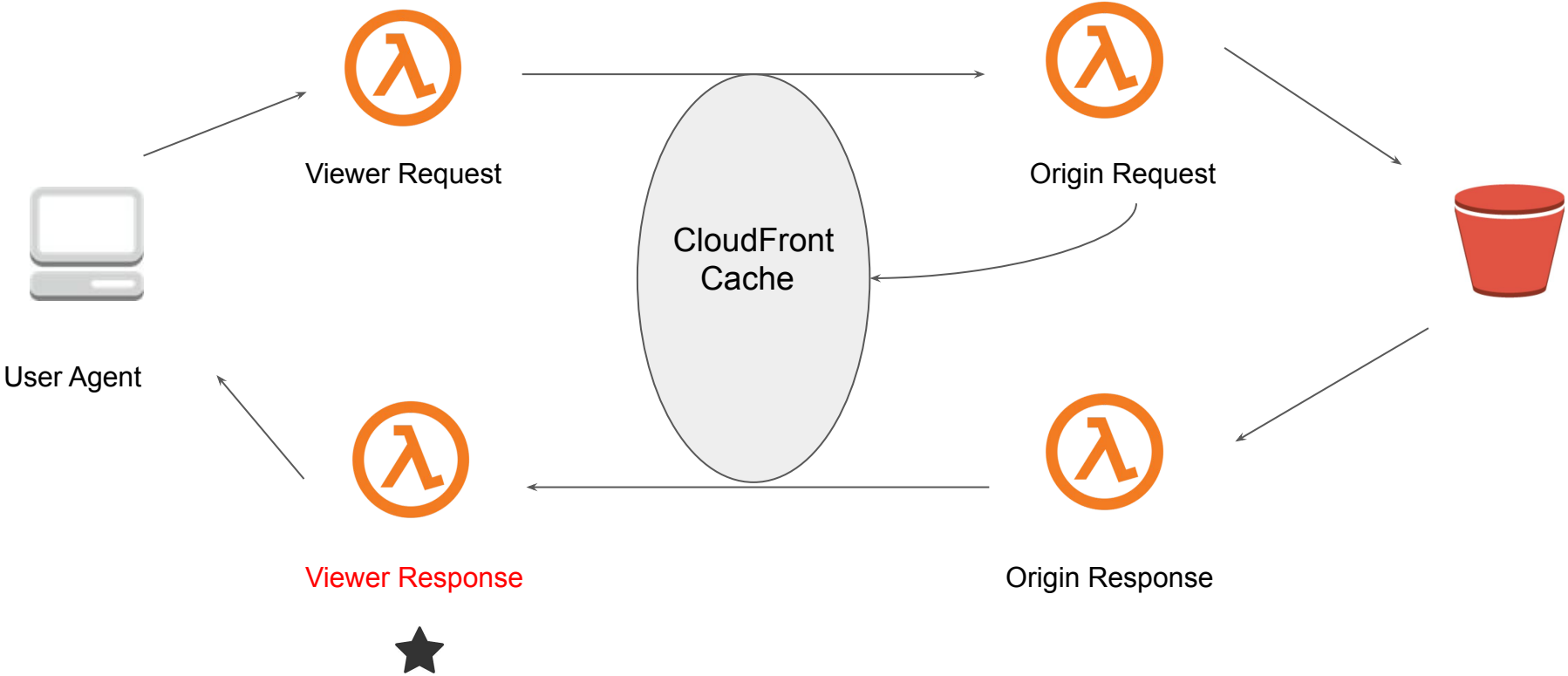
Origin Response



Viewer Response

Executed on all the responses received either from the origin or the cache.

Modifies the response headers before caching the response.


Viewer Request
Origin Request
CloudFront
Cache
User Agent
Viewer Response
Origin Response

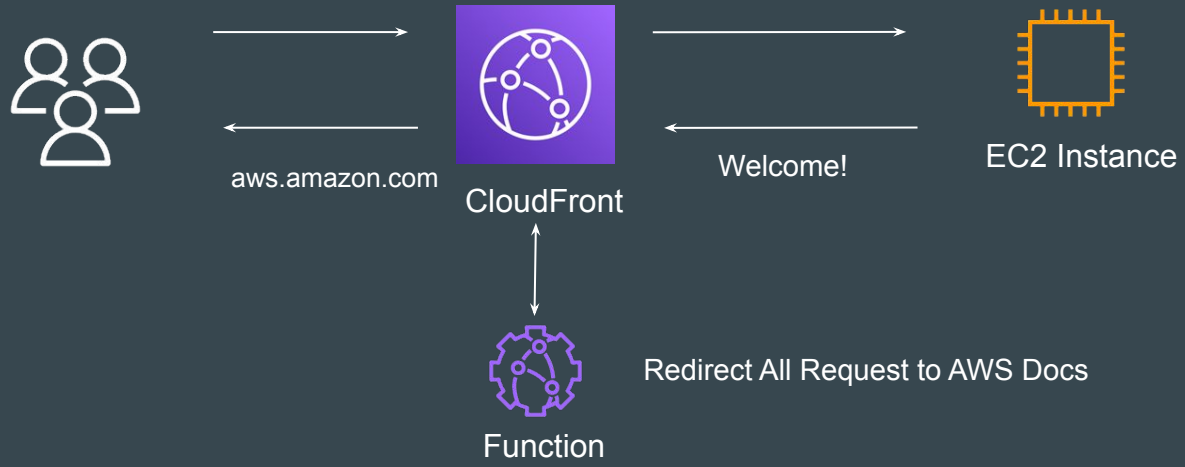
CloudFront Functions



Understanding the Basics

CloudFront Functions allows users to write lightweight functions in JavaScript for high-scale, latency-sensitive CDN customizations.

Your functions can manipulate the requests and responses that flow through CloudFront



Basics of API

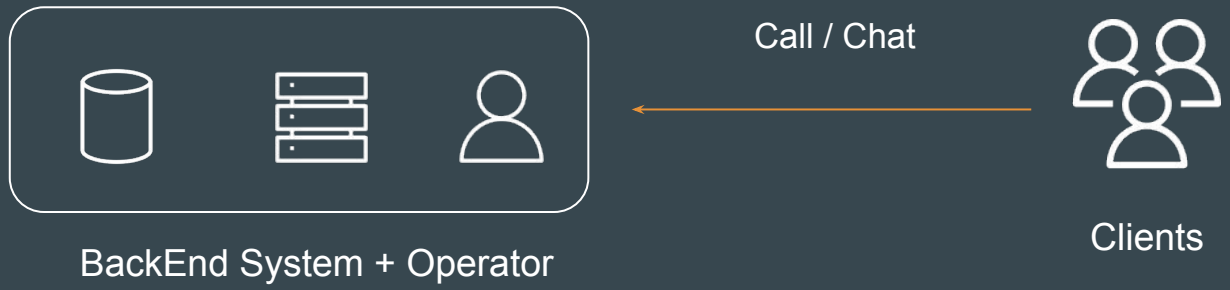


Understanding the Challenge

Book Distributor maintains the list of available books in it's backend systems.

Operator has access to Backend system to check the availability.

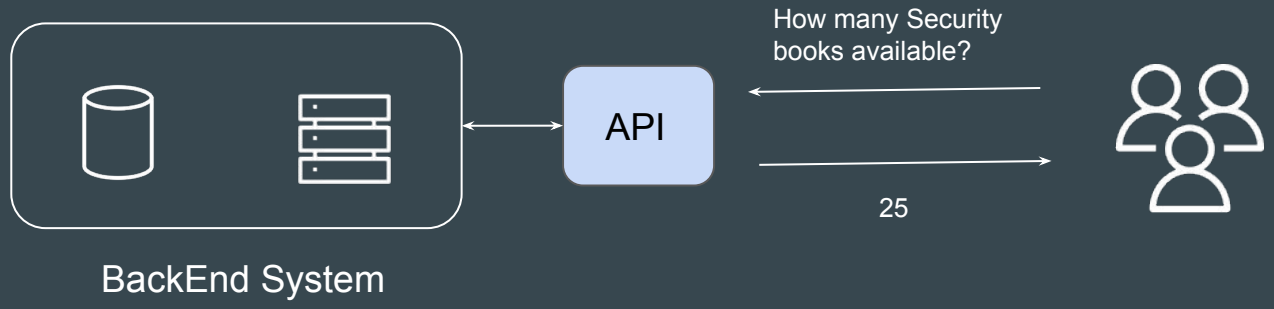
Clients they connect to Operator via Phone call / Chat option



API Based Approach

The book distributor could provide an API to check stock availability.

APIs let you open up access to your resources while maintaining security and control.



Simple Use-Case

James wants to build a weather report application.

OpenWeatherMap is an online service that provides global weather data via API.

He decided to connect his application to OpenWeatherMap API to fetch the latest reports and populate it in application.



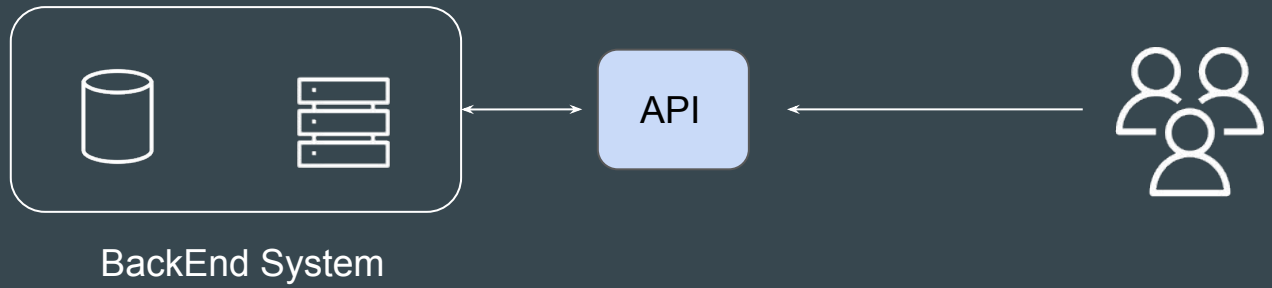
API Gateway



Introduction to Topic

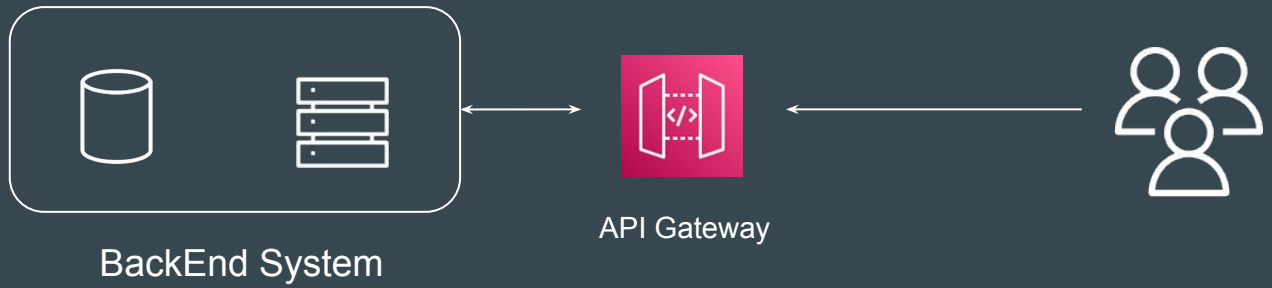
APIs act as the "front door" for applications to access data, business logic, or functionality from your backend services.

Hence API should be able to be highly available and handle thousands of requests.



Understanding the Basics

Amazon API Gateway is a fully managed service that makes it easy for developers to create, publish, maintain, monitor, and secure APIs at any scale.



REST APIs vs HTTP APIs



Understanding the Basics

REST APIs and HTTP APIs are both RESTful API products.

REST APIs support more features than HTTP APIs, while HTTP APIs are designed with minimal features so that they can be offered at a lower price.

Which to Choose?

Choose REST APIs if you need features such as API keys, per-client throttling, request validation, AWS WAF integration, or private API endpoints.

Choose HTTP APIs if you don't need the features included with REST APIs.

Core Differences - Security

API Gateway provides a number of ways to protect your API from certain threats, like malicious actors or spikes in traffic.

Security features	REST API	HTTP API
Mutual TLS authentication	✓	✓
Certificates for backend authentication	✓	
AWS WAF	✓	

Core Differences - API Management

Choose REST APIs if you need API management capabilities such as API keys and per-client rate limiting

Features	REST API	HTTP API
Custom domains	✓	✓
API keys	✓	
Per-client rate limiting	✓	
Per-client usage throttling	✓	

Core Differences - Monitoring

API Gateway supports several options to log API requests and monitor your APIs

Feature	REST API	HTTP API
Amazon CloudWatch metrics	✓	✓
Access logs to CloudWatch Logs	✓	✓
Access logs to Amazon Kinesis Data Firehose	✓	
Execution logs	✓	
AWS X-Ray tracing	✓	

Core Differences - Endpoint Type

The endpoint type refers to the endpoint that API Gateway creates for your API

Endpoint types	REST API	HTTP API
Edge-optimized	✓	
Regional	✓	✓
Private	✓	

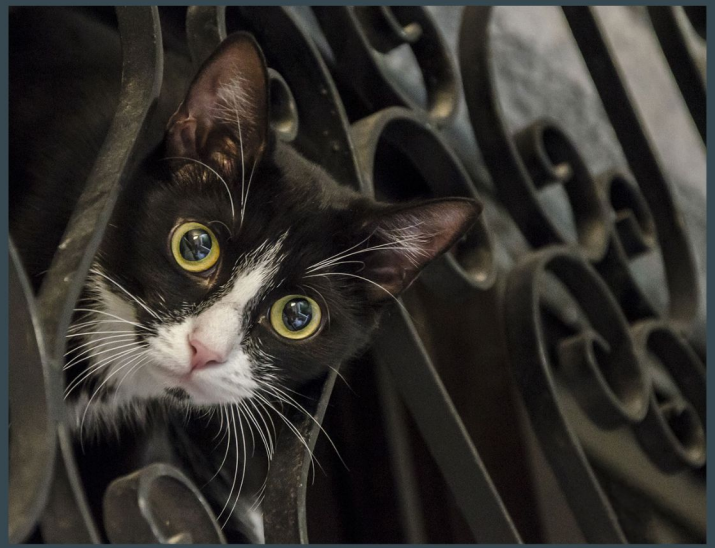
Core Differences - Development

As you're developing your API Gateway API, you decide on a number of characteristics of your API.

These characteristics depend on the use case of your API.

Features	REST API	HTTP API
CORS configuration	✓	✓
Test invocations	✓	
Caching	✓	
User-controlled deployments	✓	✓
Automatic deployments		✓
Custom gateway responses	✓	
Canary release deployments	✓	
Request validation	✓	
Request parameter transformation	✓	✓
Request body transformation	✓	

When someone deployed HTTP API for prod environment



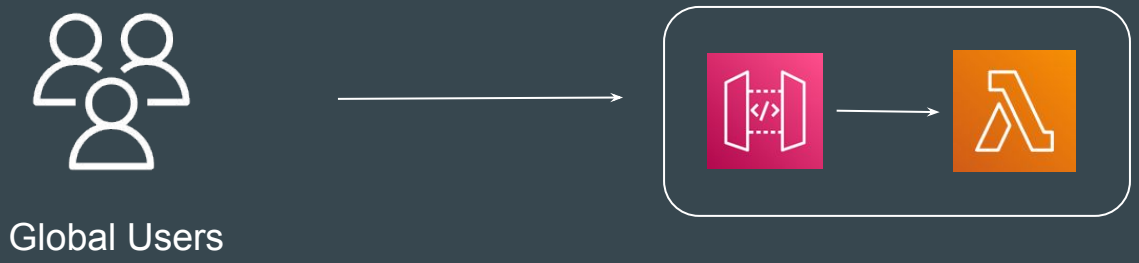
API Gateway Practical



Overall Implementation Architecture

1. Create HTTP API

2. API will invoke a backend Lambda function.



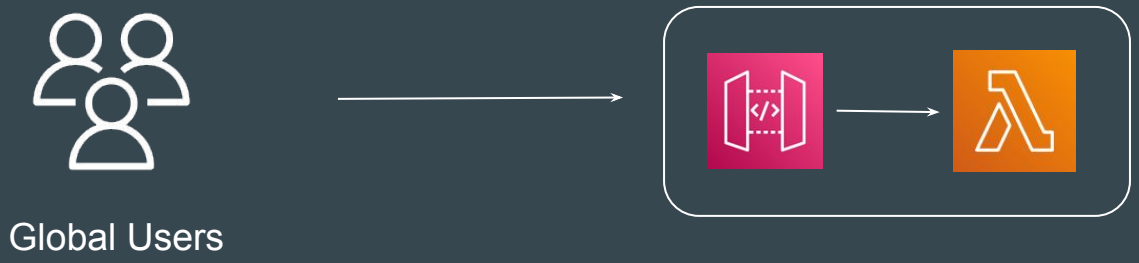
Creating REST API



Overall Implementation Architecture

1. Create REST API

2. API will invoke a backend Lambda function.

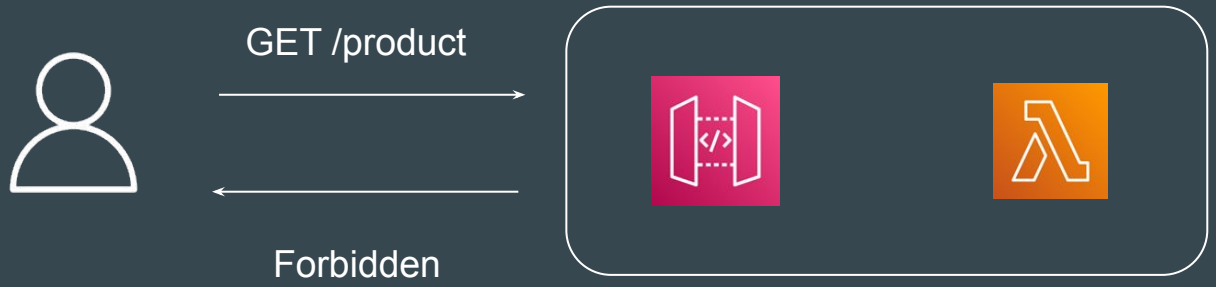


API Keys and Usage Plans



Basics of API Keys

API keys are alphanumeric string values that you distribute to application developer customers to grant access to your API.



Connecting Through API Key

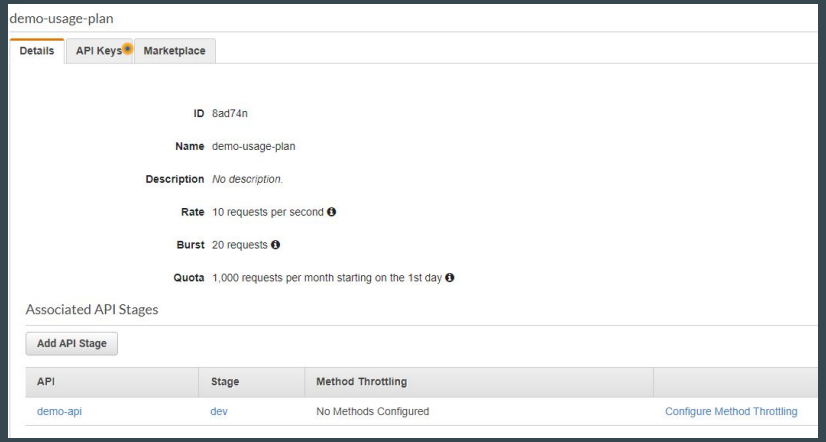
You can use the X-API-KEY header while connecting to the API Endpoint.

```
C:\Users\zealv>curl --header "x-api-key: bDa2vO89lF9TBgusPLptR253M4QzpVVlrTzKPPg3" https://9jxbr4wdac.execute-api.us-east-1.amazonaws.com/dev
{"statusCode":200,"body":"\"Hello from Lambda!\""}
```

Usage Plan

A usage plan specifies who can access one or more deployed API stages and methods—and optionally sets the target request rate to start throttling requests.

The plan uses API keys to identify API clients and who can access the associated API stages for each key.



Points to Note

After you create, test, and deploy your APIs, you can use API Gateway usage plans to make them available as product offerings for your customers.

You can configure usage plans and API keys to allow customers to access selected APIs, and begin throttling requests to those APIs based on defined limits and quotas.

These can be set at the API, or API method level.

Points to Note

API Gateway throttles requests to your API using the token bucket algorithm, where a token counts for a request

When request submissions exceed the steady-state request rate and burst limits, API Gateway begins to throttle requests. Clients may receive 429 Too Many Requests

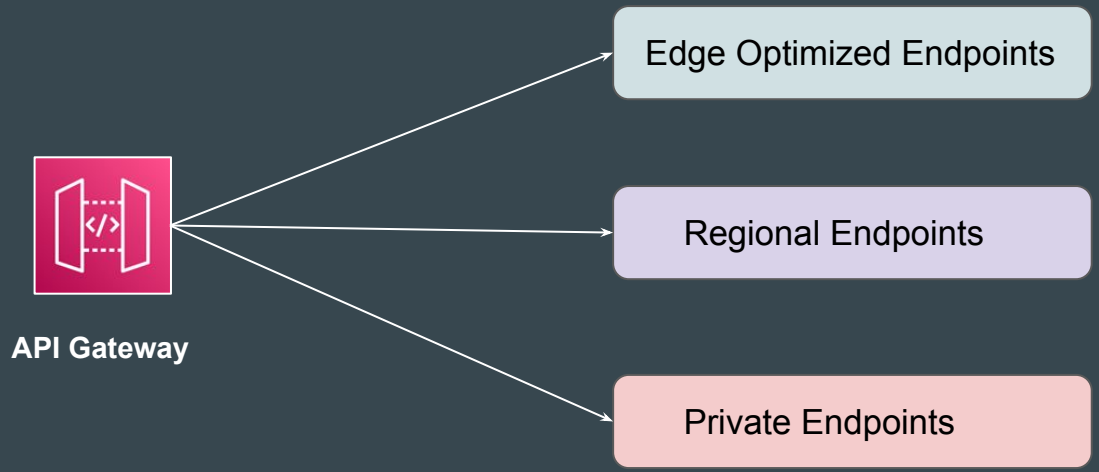
There is a default quota of 10,000 requests per second (RPS) applicable at per account per region.

API Gateway Endpoint Types



API Endpoints

Depending on where the majority of your API traffic originates from, you can create a appropriate API Gateway endpoint type.

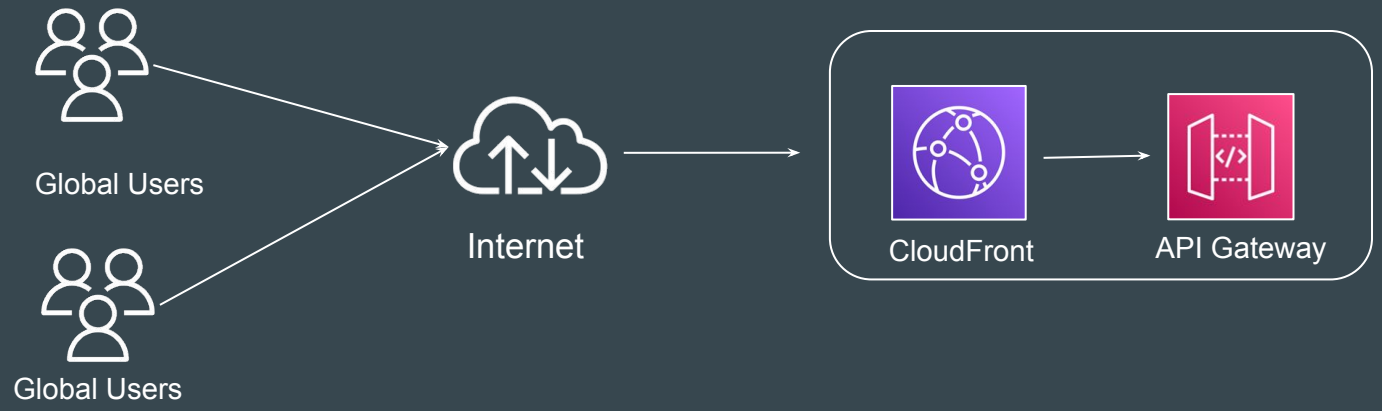


Edge-optimized API endpoints

An edge-optimized API endpoint is best for geographically distributed clients.

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

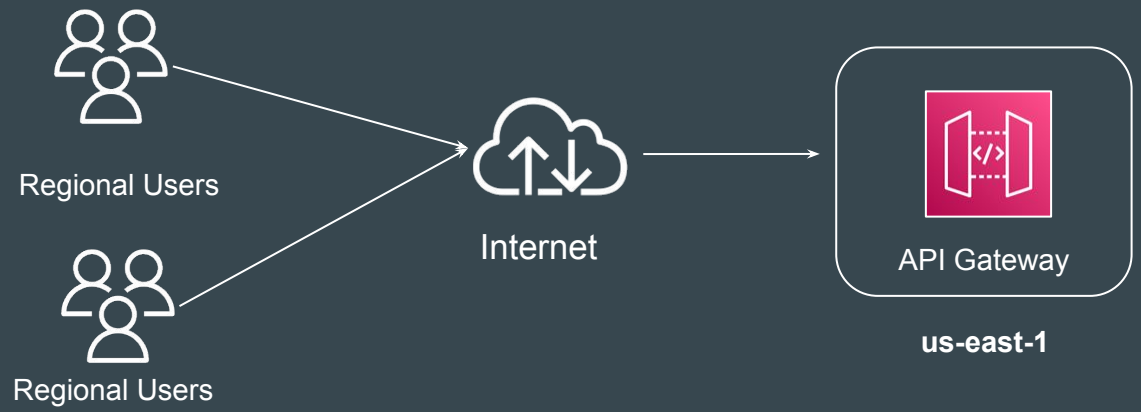
This is the default endpoint type for API Gateway REST APIs.



Regional API endpoints

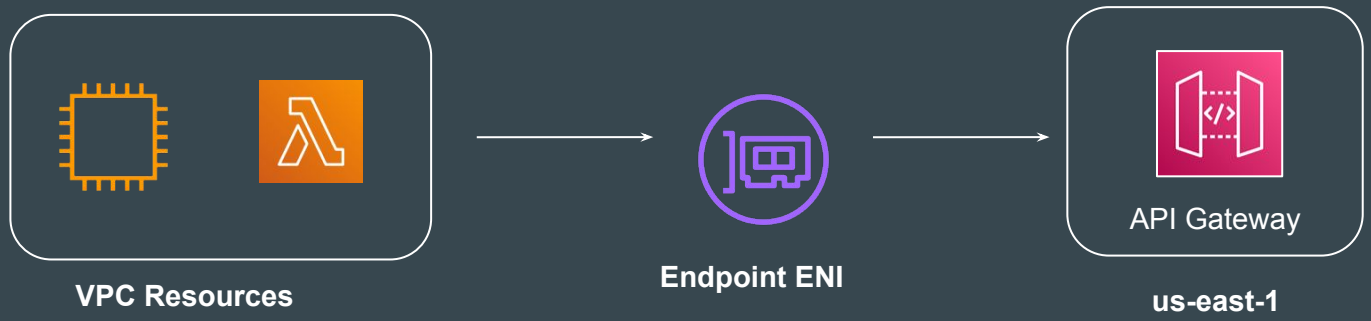
A regional API endpoint is intended for clients in the same region.

When a client running on an EC2 instance calls an API in the same region, or when an API is intended to serve a small number of clients with high demands, a regional API reduces connection overhead.



Private API endpoints

A private API endpoint is an API endpoint that can only be accessed from your Amazon Virtual Private Cloud (VPC) using an interface VPC endpoint

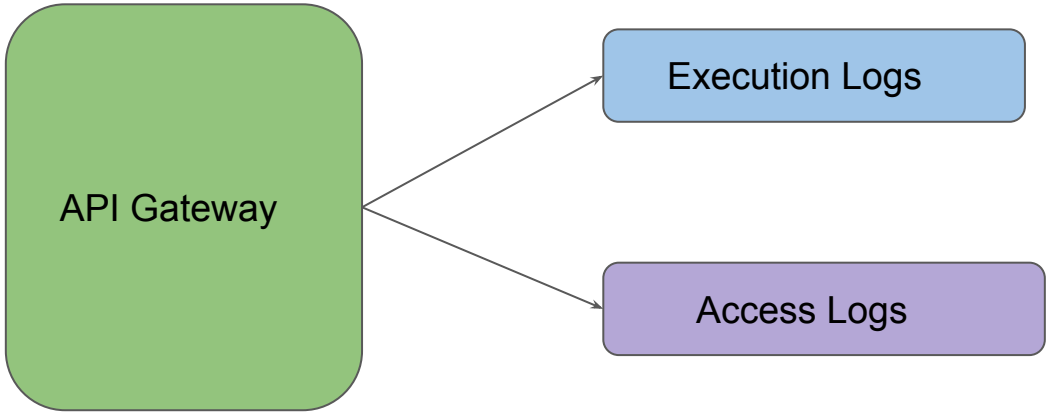


API Gateway Logging

Back to Logging!

Logging at API Gateway Level

Logging at API Gateway allows customers to log calls that are made to the API Gateway along with detailed information as API Gateway goes through each step of processing the request.

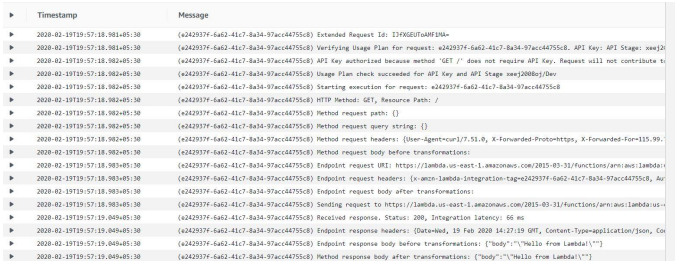


1. Execution Logs

Records the API Gateway internal information as the request is processed.
These are fully managed by the API Gateway.

Contains information like:

- The request URL
- The request data received by API Gateway
- The request data sent to the Lambda function
- The response received from the Lambda function
- The response data sent by API Gateway



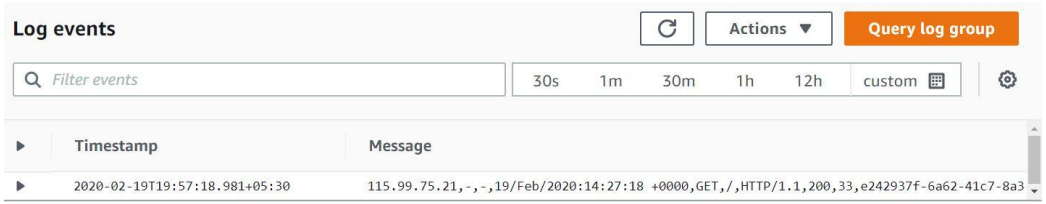
Useful when a specific request needs troubleshooting.

2. Access Logs

Logs related to who has accessed the API.
Very similar to the Apache / Nginx Logs.

Contains information like:

- The caller's IP address
- The request time
- The request HTTP method
- The request URL
- The response HTTP status code, etc.



CloudWatch Metrics for API Gateway

There are certain metrics that are made available in CloudWatch for the API Gateway resource.

Some of these metrics include:

- 5XX Error
- Latency
- Count
- 4XX Error



Amazon Comprehend

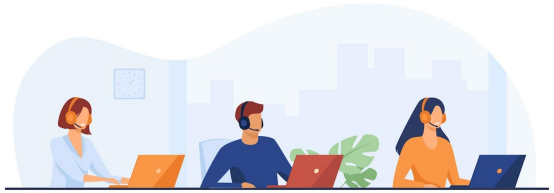
ML to Analyze Text

Simple Use-Case

There are 100 customer representatives working in a call center.

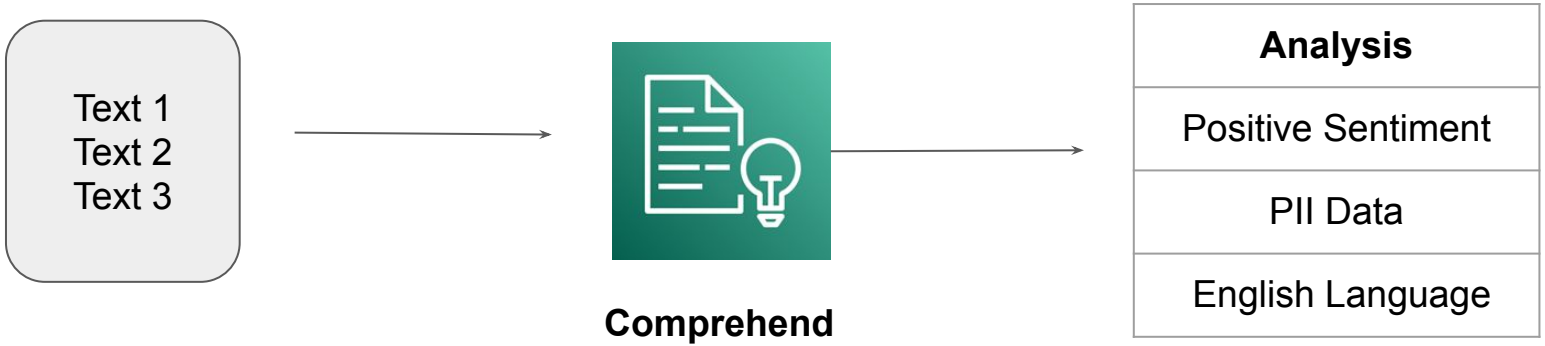
All the conversation is recorded into text (speech to text converter)

Management wants to know the overall sentiment of conversation (positive/negative).



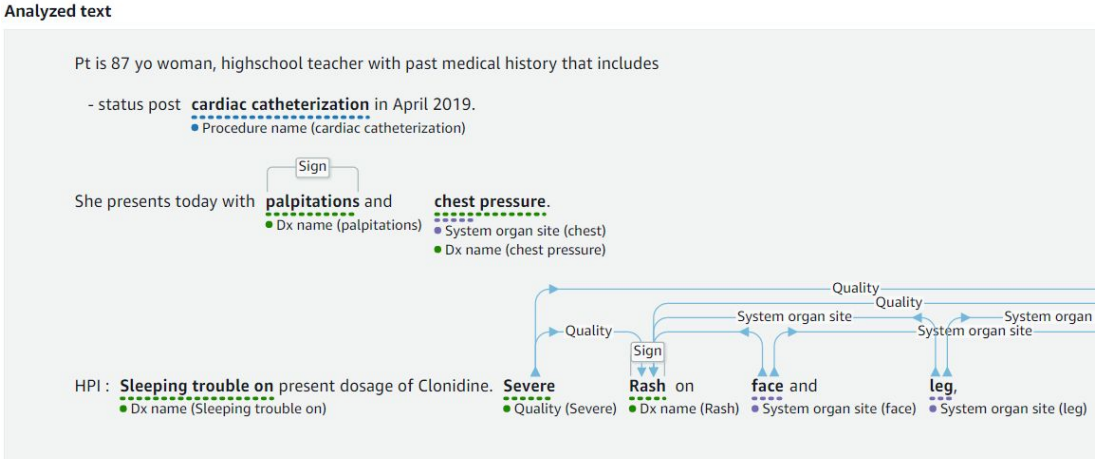
Amazon Comprehend

Amazon Comprehend is a natural-language processing (NLP) service that uses machine learning to uncover valuable insights and connections in text.



Amazon Comprehend - Medical

Amazon Comprehend Medical is a HIPAA-eligible NLP service that uses machine learning to understand and extract health data from medical text, such as prescriptions, procedures, or diagnoses.

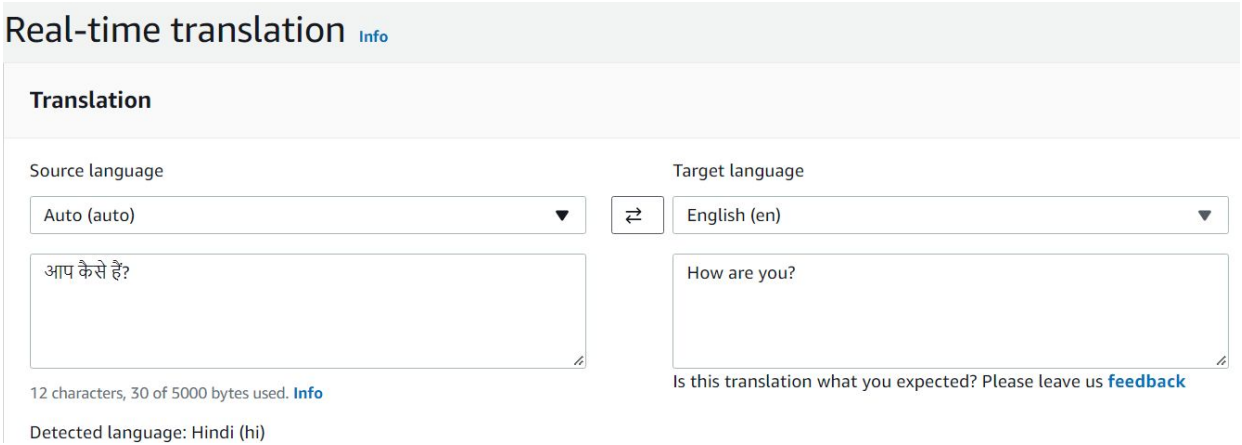


Amazon Translate

Translate Languages

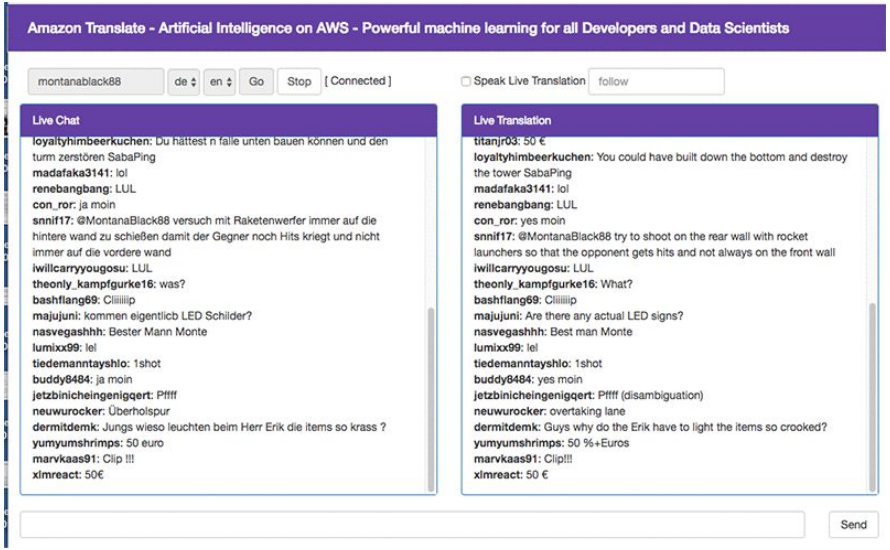
Understanding the Basics

Amazon Translate is a neural machine translation service that delivers fast, high-quality, affordable, and customizable language translation.



Sample Use-Case - Chat Application

You can translate messages in real-time based received on services like Twitch.

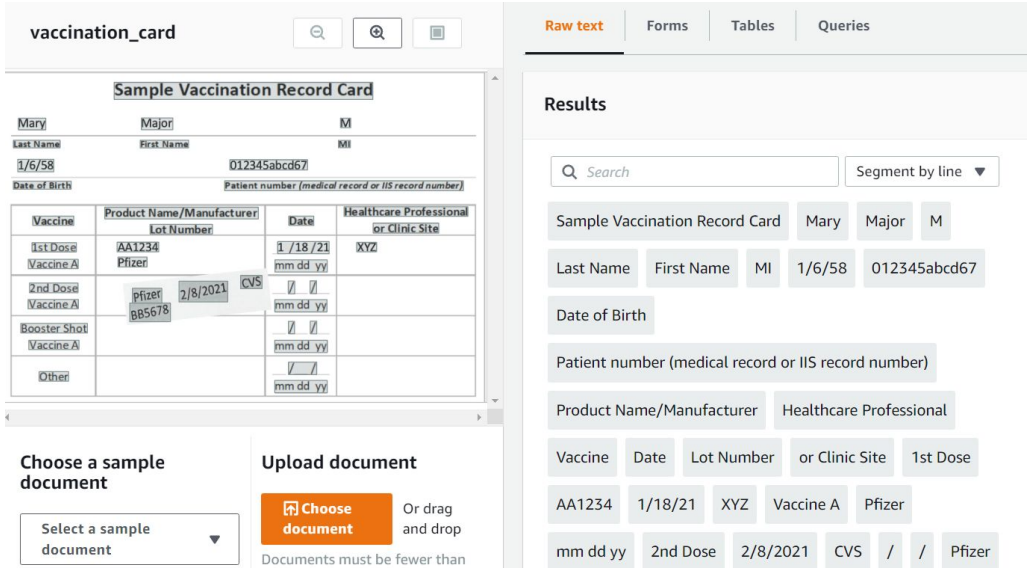


Amazon Textract

Handwriting to Text!

Understanding the Basics

Amazon Textract is a machine learning (ML) service that automatically extracts text, handwriting, and data from scanned documents.

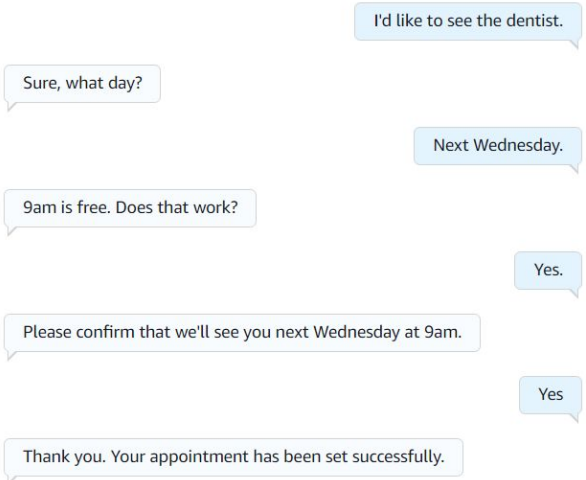


Amazon Lex

Automate Conversations

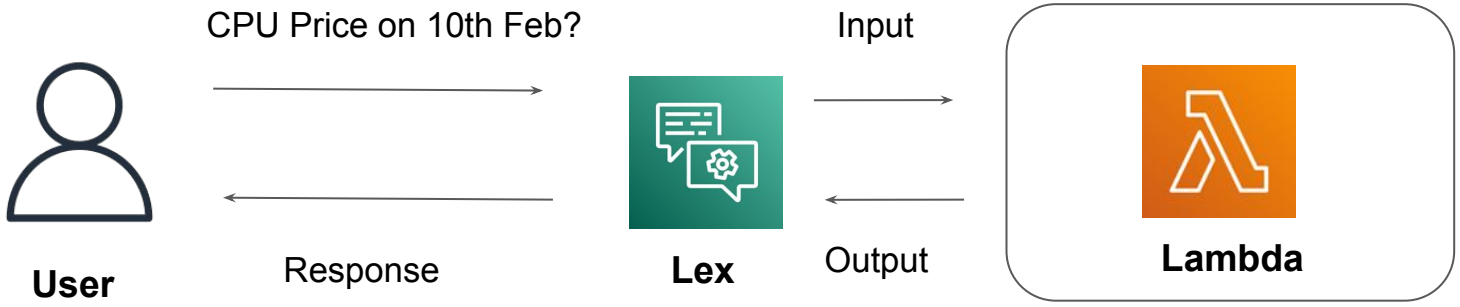
Basic of Amazon Lex

Amazon Lex is a fully managed AI service with advanced natural language models to design, build, test, and deploy conversational interfaces in applications.



Integration with Lambda

Amazon Lex can also be integrated with Lambda Function to achieve a specific use-case.

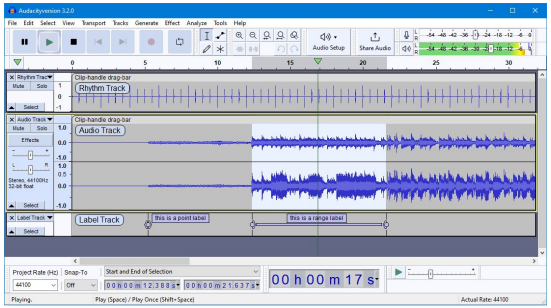


Amazon Transcribe

Speech to Text Converter

Understanding the Basics

Amazon Transcribe is an automatic speech recognition service that uses machine learning models to convert audio to text.



Speech

Hi Everyone

Welcome Back

This is Demo

Text

Call Analytics

Amazon Transcribe Call Analytics allows organizations to gain insight into customer-agent interactions. Call Analytics provides you with:

1. Call characteristics, including talk time, non-talk time, speaker loudness, interruptions, and talk speed

2. Speaker sentiment for each caller at various points in a call

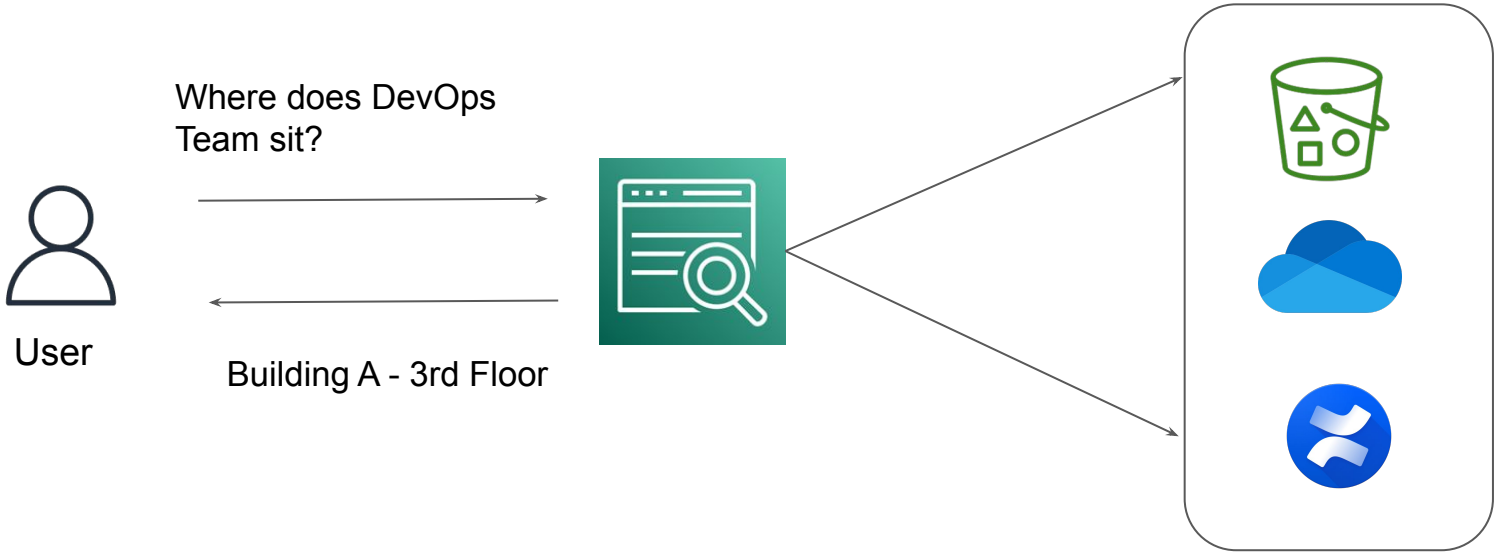
3. Call summarization, which detects issues, action items, and outcomes

Amazon Kendra

Enterprise Searching

Understanding the Basics

Amazon Kendra is a highly accurate and intelligent search service that enables your users to search unstructured and structured data using natural language processing and advanced search algorithms.



AWS Rekognition

Deep Learning

Overview of AWS Rekognition

AWS AWS Rekognition is a deep learning based virtual analysis service.

It allows us to easily integrate powerful visual analysis into our application

-

Best of Luck for the Exams

