**AWS Intro**

**Cloud Computing**  
               Cloud computing is the on-demand delivery of IT resources over the Internet with pay-as-you-go pricing. Instead of buying, owning, and maintaining physical data centers and servers, you can access technology services, such as computing power, storage, and databases, on an as-needed basis from a cloud provider.

**Top Cloud Providers**  
               Amazon Web Services (AWS)  
               Google Cloud Platform (GCP)  
               Microsoft Azure

**On Prem**  
               An on-premises data center is a group of servers that you privately own and control.  
               **"On-prem"** refers to private data centers that companies house in their own facilities and maintain themselves.

**On Prem vs Cloud Models (IaaS, PaaS, SaaS)**

**Infrastructure as a Service** provides you with the highest level of flexibility and management control over your IT resources and is most similar to existing IT resources that many IT departments and developers are familiar with today.

**Platforms as a Service** remove the need for organizations to manage the underlying infrastructure (usually hardware and operating systems) and allow you to focus on the deployment and management of your applications

**Software as a Service** provides you with a completed product that is run and managed by the service provider.

**Amazon Web Services –**[**What is AWS ?**](https://aws.amazon.com/what-is-aws/?pg=TOCC)

Amazon Web Services (AWS) is the world’s most comprehensive and broadly adopted cloud, offering **over 200** fully featured services from data centers globally. Millions of customers—including the fastest-growing startups, largest enterprises, and leading government agencies—are using AWS to lower costs, become more agile, and innovate faster.

**AWS Service:** is a collection of resources and tools that AWS provides to users to accomplish specific tasks or functions in the cloud.

**AWS Resources:** It refers to any computing or storage component that is used to build and deploy applications in the cloud. This can include virtual machines, databases, load balancers, network interfaces, and many other components that are managed by AWS.

**Example:** **EC2** is an AWS service that provides virtual machines as resources, and **S3** is an AWS service that provides object storage as resources.

We can Access AWS cloud using   
a) AWS Console  
b) AWS CLI  
c) AWS SDKs

**AWS Region and Availability Zone**

**AWS Region** is a separate geographic area where we cluster data centers. Each Region is designed to be isolated from the other Regions. This achieves the greatest possible fault tolerance and stability.

When you view your resources, you see only the resources that are tied to the Region that you specified.

**Availability Zones** are multiple, isolated locations within each Region. The code for Availability Zone is its Region code followed by a letter identifier. For example, us-east-1a.

Each Region will have a minimum of 3 separate AZ.

**Amazon Machine Image (AMI):**

* An Amazon Machine Image (AMI) provides the information required to launch an instance.
* You can launch multiple instances from a single AMI when you need multiple instances with the same configuration.
* You must specify an AMI whenever you launch an instance.

**Custom AMI**

* You can launch an instance from an existing AMI, customize the instance by installing/uninstalling packages on the instance, and then save this updated configuration as a custom AMI. Instances launched from this new custom AMI include the customizations that you made when you created the AMI.
* AMI can be shared across different accounts without making it public, but in the same region.  If the Custom AMI is required in another region, it needs to be copied and then shared again. ([Reference](https://docs.aws.amazon.com/AWSEC2/latest/UserGuide/sharing-amis.html))

**Difference between EBS Snapshot and AMI**

An **AMI** image is a backup of an entire EC2 instance. Associated with an AMI image are EBS snapshots. Those EBS snapshots are the backups of the individual EBS volumes attached to the EC2 instance at the time the AMI image was created.

An **EBS snapshot** is a backup of a single EBS volume. The EBS snapshot contains all the data stored on the EBS volume at the time the EBS snapshot was created.

**User Data**script in EC2 instances**:**  
We can run commands on the instance at launch by passing user data to the instance.  It can be used to perform some common configuration tasks or even run scripts after an instance starts.

**Security Groups**  
It is a stateful service which acts like a firewall at the instance to route inbound and outbound traffic by defining inbound and outbound rules. Security groups are stateful — if you send a request from your instance, the response traffic for that request is allowed to flow in regardless of inbound security group rules.

A security group acts as a virtual firewall for your instance to control inbound and outbound traffic

#### Assignment - 2023-08-29

1. Create EC2 instance with custom user-data.
2. Create an AMI and launch EC2 instance with custom AMI
3. How can I connect to an EC2 instance if I lost my SSH key pair after its launch?
4. **User Data in EC2:**
5. When you launch an instance in Amazon EC2, you have the option of passing user data to the instance that can be used to perform common automated configuration tasks and even run scripts after the instance starts.  
   There are two types of user data in linux ([reference](https://docs.aws.amazon.com/AWSEC2/latest/UserGuide/user-data.html)) - shell scripts and cloud-init directives. For Windows instances ([reference](https://docs.aws.amazon.com/AWSEC2/latest/WindowsGuide/ec2-windows-user-data.html)), you can use batch script and Powershell script.
6. Note: The user data run only during the first launch of instance. In order to run the user data scripts on every restart of the instance, we have to update some configuration ([Reference](https://aws.amazon.com/premiumsupport/knowledge-center/execute-user-data-ec2/)).
7. **EC2 Purchasing Options and Cost Saving: (**[**Reference**](https://docs.aws.amazon.com/AWSEC2/latest/UserGuide/instance-purchasing-options.html)**)**

### **Amazon IAM (Identity and Access Management)**

IAM provides access to AWS accounts and services where we can manage Users, Groups, Roles and Policies.

It is a global service. Anything you create/change applies **globally** to all AWS regions.

**IAM Users** - We create users and assign necessary permissions to them in the form of policies.

**IAM Groups** - We can create groups for ex. Developers, Administrators, Testers/QA etc and attach policies at the group level.

**IAM Roles** - These are entities used to create and assign permissions to allow users/resources to perform actions

**IAM Policy** defines permissions for users/roles to perform an action/operation on AWS cloud

**Types of IAM Policy: (Reference:**[**link**](https://docs.aws.amazon.com/IAM/latest/UserGuide/access_policies.html)**)**

1. **Identity-based Policies:** Applicable on users, groups of users, and roles

* AWS Managed policy
* Custom Managed Policy
* Inline Policy

1. **Resource-based policies:** Attach to a resource such as an Amazon S3 bucket
2. **Session Policies:** create a temporary session for a role or federated user
3. There are also **Permissions Boundaries**, **Organizations SCPs**and **Access control lists (ACLs)**

**Note:** More than one policy can be attached to a user or a group at the same time.  
Policies can’t be attached directly to resources like EC2 instance, S3 bucket etc.

**IAM Roles** - These are entities used to create and assign permissions to allow users/resources to perform actions on AWS account.

**Cross Account Role (**[Reference](https://docs.aws.amazon.com/IAM/latest/UserGuide/id_roles_common-scenarios_aws-accounts.html))- Cross-account IAM roles allow users from one account to securely connect and access AWS resources in another account. There must be relevant permissions provided in the policies attached to the user (account1) and the role (account2) to assume the role.

#### ****AWS CLI****

The AWS Command Line Interface (AWS CLI) is an open source tool that enables you to interact with AWS services using commands in your command-line shell. With minimal configuration, the AWS CLI enables you to start running commands that implement functionality equivalent to that provided by the browser-based AWS Management Console from the command prompt in your terminal program:

Linux shells – Use common shell programs such as bash, zsh, and tcsh to run commands in Linux or macOS.  
Windows command line – On Windows, run commands at the Windows command prompt or in PowerShell.  
Remotely – Run commands on Amazon Elastic Compute Cloud (Amazon EC2) instances through a remote terminal program such as PuTTY or SSH, or with AWS Systems Manager.

**Install** or **Update** AWS CLI by following the link [here](https://docs.aws.amazon.com/cli/latest/userguide/getting-started-install.html).

**STS – Getting Session Token for temporary access to AWS accounts**

The AWS Security Token Service (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).

Run the following commands to configure AWS CLI and get session token

*aws configure*  
               <Enter Access Key Id and Secret Access Key, Region and Output>

*aws sts get-session-token*  
example output:  
{  
   "Credentials": {  
        "AccessKeyId": "ASIATFOKVBZRFMBZYVGJ",  
        "SecretAccessKey": "mN7FUNC4RyGWs9sqYR2HBJ189VRgIU4D1lO6vSh8",  
        "SessionToken": "",  
        "Expiration": “2023-07-03T23:06:22+00:00”  
   }  
}

**Assume Role: (**[**Reference**](https://repost.aws/knowledge-center/iam-assume-role-cli)**)**

*aws sts assume-role*

Returns a set of temporary security credentials that you can use to access AWS resources that you might not normally have access to. These temporary credentials consist of an access key ID, a secret access key, and a security token.

**Example using STS and Assuming Role to query AWS resources:**

1. Create a IAM user   
2. Add ec2 full access policy   
3. Try to list the bucket (*aws s3 list-buckets*) - you can't list the bucket because the user is not having the permission to list bucket (access denied)  
4. Grant the user to assume a role   
> Create a role   
> Attach a policy s3 full access policy   
> Edit the role under **Trust Relationship**, modify the policy to user instead of ec2  
"Principal": {  
                 **"Service": “ec2.amazonaws.com”**  
}

to

"Principal": {  
               **"AWS": “arn:aws:iam::217857724000:user/test-user-cli”**  
}

5. *aws sts assume-role --role-arn <enter\_role\_arn> --role-session-name s3-access --duration-seconds 3600*  
6. copy the Accesskey, Secretkey and Session token (Reference: [Linux or Windows](https://docs.aws.amazon.com/IAM/latest/UserGuide/id_credentials_temp_use-resources.html#using-temp-creds-sdk-cli))  
7.  *export AWS\_ACCESS\_KEY\_ID=<enter the copied access key>*  
*export AWS\_SECRET\_ACCESS\_KEY=<enter the copied session key>*  
*export AWS\_SESSION\_TOKEN=<sessiontoken>*  
8. *aws s3 ls***OR***aws s3api list-objects --bucket <bucket\_name>*

### **Amazon Elastic Block Store (Amazon EBS)**

* Amazon Elastic Block Store (Amazon EBS) provides block level storage volumes for use with EC2 instances. EBS volumes behave like raw, unformatted block devices.
* EBS volumes are persistent. They can live beyond the life of the ec2 instance that they are attached to. They can be attached/detached to and from different ec2 instances.
* An EBS volume can only be attached to 1 ec2 instance at a time (Multi Attach EBS Volume can be attached to multiple instances). They are mounted to the file system.
* You can back up the EBS volume into a snapshot, which can later be used/restored into a new EBS volume.

**Types of EBS Volumes**

Reference: [EBS volume types](https://docs.aws.amazon.com/AWSEC2/latest/UserGuide/ebs-volume-types.html)

* Solid State Drives (SSD)
  1. General Purpose SSD
  2. Provisioned IOPS SSD
* Hard Disk Drives (HDD)
  1. Throughput Optimized HDD
  2. Cold HDD
* Previous Generation
  1. Magnetic

**EBS Snapshots**

* You can back up the data on your Amazon EBS volumes to Amazon S3 by taking point-in-time snapshots.
* Snapshots are incremental backups.
* Each snapshot contains all of the information that is needed to restore your data (from the moment when the snapshot was taken) to a new EBS volume.
* You can use Amazon Data Lifecycle Manager to automate the creation, retention, and deletion of snapshots taken to back up your Amazon EBS volumes

### **Amazon EFS (Elastic File System):**

* Amazon Elastic file systemis a regional service storing data within and across multiple Availability Zones (AZs) for high availability and durability
* Amazon EFS is an NFS file system service offered by AWS.
* An Amazon EFS file system is excellent as a managed network file system that can be shared across different Amazon EC2 instances.
* Amazon EFS works like NAS devices and performs well for big data analytics, media processing workflows, and content management.
* EFS can be accessed by multiple instances at a time through NFS protocol. It is used as a clustered database and document sharing.

**Benefits of EFS:**

* With EFS you need not worry about managing file servers or storage, updating hardware, configuring software, or performing backups as EFS is a fully managed service
* The distributed architecture of Amazon EFS provides data protection from an AZ outage, system and component failures, and network connection errors.
* Network Access to the files can be controlled using Virtual Private Cloud security group rules and with Identity Access Management policies and EFS access points you can control the access to your files
* Amazon EFS is designed to provide the throughput, IOPS, and low latency needed for a broad range of workloads
* With Amazon EFS, storage capacity is elastic, growing and shrinking automatically as you add and remove files, dynamically providing the storage capacity to applications as needed.
* AWS EFS provides encryption of data both at rest and in transit so that your data is secure.

**EFS Access Points:**(Reference: [Access Points](https://docs.aws.amazon.com/efs/latest/ug/efs-access-points.html))

* Access Points are application-specific entry points into an EFS file system that make it easier to manage application access to shared datasets.
* Access points can enforce a user identity, including the user's POSIX groups
* Access points can also enforce a different root directory for the file system so that clients can only access data in the specified directory or its subdirectories.

**NOTE:**We can mount and access EFS volumes across VPCs, i.e., EFS in VPC1 can be accessed from an ec2 instance running in VPC2, provided **both the VPCs are connected** using VPC peering or Transit Gateway.  (Reference: [Mount a File System from a Different VPC](https://itdefined.org/course/training-material/2/4/248/Mount%20a%20File%20System%20from%20a%20Different%20VPC))

### **Amazon S3 (Simple storage service)**

* Amazon S3 is an object storage service. It is a scalable, high-speed, web-based cloud storage service.
* It has a simple web services interface that you can use to store and retrieve any amount of data, at any time, from anywhere on the web.
* S3 provides 99.999999999% durability for objects stored in the service and supports multiple security and compliance certifications.

**Single operation upload:**

* It’s a traditional upload where you will upload the object in one part
* A single operation upload can upload the file up to 5GB in size.

**Upload object in parts:**

* Using multipart upload, you can upload the large objects up to 5TB.
* You can use multipart upload for the objects from 5MB to 5TB in size.

**Rules for bucket naming:**

* Bucket names must be between 3 and 63 characters long.
* Bucket names can consist only of lowercase letters, numbers, dots (.) and hyphens (-)
* Bucket names must begin and end with a letter or number.
* Bucket names must not be formatted as an IP address (for example, 192.168.5.4).
* Bucket names can't begin with xn-- (for buckets created after February 2020).

**Limitation of S3 bucket:**

* Only 100 buckets can be created per account.
* Can hold unlimited objects

**S3 Storage classes: (Reference:**[**https://aws.amazon.com/s3/storage-classes/**](https://aws.amazon.com/s3/storage-classes/)**)**

* Standard
* S3 Intelligent-Tiering
* S3 Standard-IA: Infrequent Access
* S3 One Zone-IA
* Glacier
  + Glacier Instant Retrieval
  + Glacier Flexible Retrieval
  + Glacier Deep Archive
* S3 Outposts

**S3 Lifecycle policy:**

* An object lifecycle policy is a set of rules that automate the migration of the object storage class to different storage class
* By default, lifecycle policies are disabled for a bucket
* Lifecycle Management can be applied to both current and previous versions.
* It can be used either in conjunction with the versioning or without versioning.
* There are 2 types of actions:

**Transition actions:** Moving objects from one storage class to another storage class. Each storage class has a different cost associated with it.

**Expiration actions:** When objects expire after a span of time (say 30 days,60 days, etc). Amazon S3 deletes expired objects on your behalf.

**S3 Encryption:**

**Two ways of protecting information while transferring data with S3**

1. **Server side/ At rest:**
2. **In-transit/Client-side encryption:**
3. Using KMS or (b) Using Client side Master key

**S3 Bucket Versioning**

Buckets can be in one of three states:

1. Un-versioned (the default)
2. Versioning-enabled
3. Versioning-suspended

* You can enable and suspend versioning at the bucket level. After you enable versioning on a bucket, it can never return to an unversioned state. But you can *suspend* versioning on that bucket.
* Objects that are stored in your bucket before you set the versioning state have a version ID of**null**.   
  When you enable versioning, existing objects in your bucket do not change. What changes is how Amazon S3 handles the objects in future requests. (Reference: [Working with objects in a versioning-enabled bucket](https://docs.aws.amazon.com/AmazonS3/latest/userguide/manage-objects-versioned-bucket.html))
* The bucket owner (or any user with appropriate permissions) can suspend versioning to stop accruing object versions.   
  When you suspend versioning, existing objects in your bucket do not change. What changes is how Amazon S3 handles objects in future requests. (Reference: [Working with objects in a versioning-suspended bucket](https://docs.aws.amazon.com/AmazonS3/latest/userguide/VersionSuspendedBehavior.html))

**S3 Bucket Policy (**Reference: [Bucket Policy](https://docs.aws.amazon.com/AmazonS3/latest/userguide/bucket-policies.html))

* A bucket policy is a resource-based policy that you can use to grant access permissions to your Amazon S3 bucket and the objects in it.
* Only the bucket owner can associate a policy with a bucket.
* The permissions attached to the bucket apply to all of the objects in the bucket that are owned by the bucket owner. These permissions do not apply to objects that are owned by other AWS accounts.
* Bucket policies use JSON-based IAM policy language.
* You can use bucket policies to add or deny permissions for the objects in a bucket. You can allow or deny requests based on the elements in the policy.

**Difference between EBS v/s EFS v/s S3**

|  |  |  |
| --- | --- | --- |
| AMAZON S3 | AMAZON EBS | AMAZON EFS |
| Can be publicly accessible Web interface Object Storage Scalable Slower than EBS and EFS | Accessible only via the given EC2 Machine File System interface Block Storage Hardly scalable Faster than S3 and EFS | Accessible via several EC2 machines and AWS services Web and file system interface Object storage Scalable Faster than S3, slower than EBS |
| Good for storing backups and other static data | Is meant to be EC2 drive | Good for applications and shareable workloads |

#### Assignment - 2023-09-09

Create a s3 bucket and create 2 folders in it. Create 2 IAM users and provide access to specific folders to these users using bucket policy.  
Ex: folder1 should be accessible to user1 to read and write and similarly, folder2 to user2.

Interview Questions:

1. If you delete a s3 bucket, can you create new bucket with same name?
2. Can you recover a deleted object in s3?

### **CloudWatch**

Amazon CloudWatch monitors your Amazon Web Services (AWS) resources and the applications you run on AWS in real time.

Default metrics of EC2 instance: Network usage, CPU Usage, Storage

We can install CloudWatch agent ([installation guide](https://docs.aws.amazon.com/AmazonCloudWatch/latest/monitoring/install-CloudWatch-Agent-on-EC2-Instance.html)) on the servers in case we need system level data like Storage, Memory, Application data etc.

**CW Metrics:**

Metrics are data about the performance of your systems/resources.

Metric data is kept for **15 months**

Basic monitoring: polls for every 5 minutes

Detailed monitoring: polls for every 1 minute.

**CW Alarm:**

CloudWatch Alarms feature allows you to watch CloudWatch metrics and to receive notifications when the metrics fall outside of the levels (high or low thresholds) that you configure and take action accordingly.

**Example:** If CPU utilization goes beyond the static threshold alarm goes to alarm state and triggers **Alarm Notification**. We can also configure actions **Autoscaling action, EC2 action or Systems Manager action**.

**Steps to create the alarm**

* Go to create alarm
* Select the metric
* Set the condition
* Select the type of alarm
* **In alarm** - The metric or expression is outside of the defined threshold.
* **Ok** - The metric or expression is within the defined threshold.
* **Insufficient data** - The alarm has just started or not enough data is available.
* Select SNS topic if already created. Or create a new topic.
* If you have chosen, EC2 metric, then you can select Autoscaling action or EC2 action
* Then give a name to the alarm and create alarm.

**CW Event:**

A CloudWatch Event indicates a change in AWS environment or A change in the state of AWS resource.

**Steps to create an event**

* Create rule – Give a name to the rule and choose event pattern
* Choose Event source (ex: AWS service)
* Choose event to watch for that service (ex: EC2 instance state-change notification)
* Add target – SNS topic or Lambda function or CodePipeline
* Add relevant tags (optional) and Create

**CW Logs:**

Amazon CloudWatch Logs service allows you to collect and store logs from your resources, applications, and services in near real-time.

**Steps to create logs**

* Give a log group name
* Choose retention setting

### **CloudTrail**

AWS CloudTrail tracks the activities of users and APIs. It continuously logs your AWS account activity

* Auditing tool that records all account activity.
* Any action taken by users, roles and AWS services are recorded to cloud trial.
* Cloud trial events are kept for **90 days** in event history.
* You can create a trail of your own to store the event history in s3 bucket.

**Working of Cloud Trail**

* **Capture** - Record activity in AWS services as AWS CloudTrail events
* **Store** - AWS CloudTrail delivers events to the AWS CloudTrail console, Amazon S3 buckets, and optionally Amazon CloudWatch Logs
* **Act** - Use Amazon CloudWatch Alarms and Events to take action when important events are detected
* **Review** - View recent events in the AWS CloudTrail console, or analyze log files with Amazon Athena

**Steps to create trails**

* Create trial
* Specify the trial name
* Choose storage location - s3 bucket (either create new bucket or use existing)
* Choose the encryption method - Server side or custom managed encryption
* If you want to enable the cloudwatch alarm, you can select that.
* Choose the types of events.
* Review and create.

There are three types of events –

**1. Data events** - Data events provide visibility into the resource operations performed on or within a resource.  
Example – s3, lambda, dynamodb.

* Amazon S3 object-level API activity (for example, GetObject, DeleteObject, and PutObject API operations)
* AWS Lambda function execution activity (the Invoke API)
* Amazon DynamoDB object-level API activity on tables (for example, PutItem, DeleteItem, and UpdateItem API operations).

**2. Management events** - Management events provide visibility into management operations that are performed on resources in your AWS account.

* Configuring security (for example, IAM AttachRolePolicy API operations)
* Registering devices (for example, Amazon EC2 CreateDefaultVpc API operations)
* Configuring rules for routing data (for example, Amazon EC2 CreateSubnet API operations)
* Setting up logging (for example, AWS CloudTrail CreateTrail API operations)

**3. Insight events** - AWS users identify and respond to unusual activity associated with write API calls.

CloudTrail Insights measures your normal patterns of API call volume, also called the baseline, and generates Insights events when the volume is outside normal patterns. Insights events are generated for write management APIs.

#### Simple Notification service (SNS)

Amazon Simple Notification Service is a notification service provided as part of Amazon Web Service.  
It provides a low-cost infrastructure for the mass delivery of messages, predominantly to mobile users

**Topic:**  
An Amazon SNS topic is a logical access point that acts as a communication channel.  
A topic is a message channel. When you publish a message to a topic, it fans out the message to all subscribed endpoints.

**Virtual Private Cloud**

Amazon Virtual Private Cloud is a commercial cloud computing service that provides users a virtual private cloud, by provision[ing] a logically isolated section of Amazon Web Services Cloud.

**Amazon VPC**(Reference: [VPC](https://docs.aws.amazon.com/vpc/latest/userguide/what-is-amazon-vpc.html)) lets you provision a logically isolated section of the AWS Cloud where you can launch AWS resources in a virtual network that you define like EC2 instance, Databases, lambda function.

**CIDR**  
Classless Inter-Domain Routing - When you create a VPC, you must specify a range of IPv4 addresses for the VPC in the form of CIDR. For example, 10.0.10.0/22 (consists of 1024 ips). **This is the primary CIDR block for your VPC**.  
Classless Inter-Domain Routing is a method for allocating IP addresses and for IP routing.

**Subnet**  
Subnet is a logical subdivision of an IP network. It is a range of IP-address.

* A subnet is a range of IP addresses in your VPC. You can launch AWS resources into a specified subnet.
* Some IP addresses are reserved for a subnet with CIDR block 10.0.10.0/24 are
  + 10.0.10.0 Network address
  + 10.0.10.1 VPC Router
  + 10.0.10.2 DNS server (DNS. (Domain Name System) The Internet's system for converting alphabetic names into numeric IP addresses)
  + 10.0.10.3 Future use
  + 10.0.10.255 N/W Broadcast address
* VPC spans multiple Availability zones.
* Subnets must be associated with route table
  + A public subnet has a route to internet
  + A private subnet doesn’t have route to internet. It creates higher level of security.
* You can use a network address translation (NAT) gateway to enable instances in a private subnet to connect to the internet or other AWS services, but prevent the internet from initiating a connection with those instances.

**Route table**  
It is a interface which connects IGW, NAT, subnets in the VPC. It helps to configure route by specifying gateway and associate specific subnets to the table.

* VPC Quota or VPC limitations
* 5 VPC per region
* 5 IGW per region
* Subnet per VPC 200
* IPv4 CIDR blocks per VPC 4
* Elastic IP addresses per Region 5
* Internet gateways per Region 5
* NAT gateways per Availability Zone 5
* Network ACLs per VPC 200
* Rules per network ACL 200

**Internet Gateway** – It is a gateway which is connected to the VPC to route traffic to and from the internet.

**NAT Gateway** – It is a gateway through which resources in the private subnet can get access to the internet. A NAT Gateway is a device used to enable instances in a private subnet to connect to the internet or other AWS services.

**Egress only Internet Gateway –**It allows outbound communication over IPv6 from instances in your VPC to the internet, and prevents the internet from initiating an IPv6 connection with your instances.

**VPC Peering**

A VPC peering ([reference](https://docs.aws.amazon.com/vpc/latest/peering/vpc-peering-basics.html)) connection is a networking connection between two VPCs that enables you to route traffic between them using private IPv4 addresses or IPv6 addresses.   
Instances in either VPC can communicate with each other as if they are within the same network.

1. You can create a VPC peering connection between your own VPCs, or with a VPC in another AWS account.
2. The VPCs can be in different regions (also known as an inter-region VPC peering connection).

Inter-Region and Intra-region VPC Peering

**Conditions:**

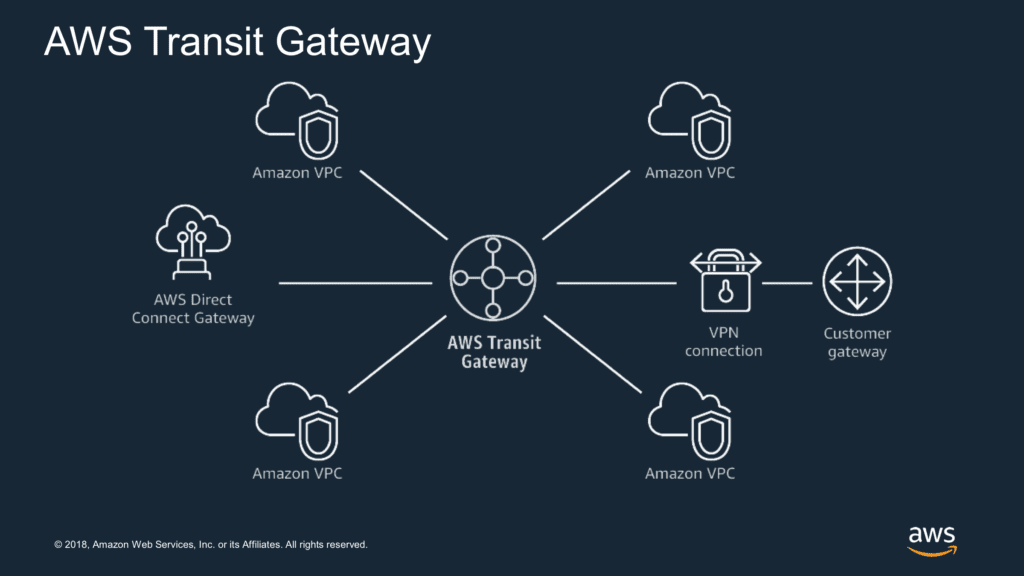
* CIDR block shouldn’t overlap
* Transitive peering relationships are not supported. i.e here VPC B cannot connect with VPC C.
* If the VPCs are in different regions, inter-region data transfer costs apply.
* You cannot have more than one VPC peering connection between the same two VPCs at the same time.

**Assignment - 2023-09-19**

Create a new VPC  
Establish VPC peering between default VPC and new VPC  
Create ec2 in both VPC and test the connection by ssh/copying a sample file

**Transit Gateway**  
The Transit Gateway ([reference](https://docs.aws.amazon.com/vpc/latest/tgw/how-transit-gateways-work.html)) is a managed service from AWS that acts as a hub interconnecting VPCs and VPN connections within a single region.

Transit Gateway is Highly Scalable. It can support bandwidths up to 50 Gbps between it and each VPC attachment. And, each Transit Gateway supports up to 5,000 VPCs and 10,000 routes



**Advantages of TGW**

* Simplified management of VPC connections. Each spoke VPC only needs to connect to the TGW to gain access to other connected VPCs.
* Supports more VPCs compared to VPC peering.
* TGW Route Tables per attachment allow for fine-grained routing.

**Disadvantages**

* Additional hop introduces some latency.
* Extra cost of hourly charge per attachment in addition to data fees.

**Limitations**

* You can connect to a maximum of three Transit Gateways over a single Direct Connect Connection for hybrid connectivity.
* Transit Gateway doesn’t support routing between VPCs with overlapping CIDRs.

**Some important points to note on TGW:**

* It can be connected across different regions and accounts
* Static routes are required
* Data is Encrypted
* Choose unique ASN (The autonomous system number) for BGP (Border Gateway Protocol)
* VPCs with Overlapping CIDR Ranges cannot be Attached to Same Transit Gateway
* TGW Attachments can only be associated with 1 Route Table
* 5000 attachments per transit gateway and 10000 static routes per transit gateway
* Up to 50 peering attachments (inter-region) per TGW

**Elastic IP - (Permanent IP address)**  
An Elastic IP address is a static, public IPv4 address designed for dynamic cloud computing. You can associate an Elastic IP address with any instance or network interface in any VPC in your account. With an Elastic IP address, you can mask the failure of an instance by rapidly remapping the address to another instance in your VPC.

**Bastion Host or Jump Server**  
BASTION host are EC2 instances present in the public subnet and connected to the private instance in the private subnet in order to route traffic from external network to the private instance. It is used to communicate instances in the private subnet.

**VPN – Virtual Private Network**  
It is a tunnel established to the private network to access resources using private IP address ([video reference](https://www.youtube.com/watch?v=3j1MLlgc5Eg)).

There are two types of VPN  
1. **Site-to-site VPN** - A secure connection between your on-premises equipment and your VPCs.  
2. **Client VPN** - is a managed client-based VPN service that enables you to securely access AWS resources and resources in your on-premises network.

**Firewall**  
A firewall is a network security system that monitors and controls incoming and outgoing network traffic based on predetermined security rules.

**Security Groups**  
It is a stateful service which acts like a firewall at the instance to route inbound and outbound traffic by defining inbound and outbound rules. Security groups are stateful — if you send a request from your instance, the response traffic for that request is allowed to flow in regardless of inbound security group rules.

A security group acts as a virtual firewall for your instance to control inbound and outbound traffic

**Network ACL**  
Network ACL is a stateless service work as firewall to the subnet which route inbound and outbound traffic to the subnet by configuring inbound and outbound rules. Network ACLs are stateless, which means that responses to allowed inbound traffic are subject to the rules for outbound traffic (and vice versa).

It 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. (Firewall at subnet level)

• Inbound means – incoming (ingress)  
• Outbound means – outgoing (egress)  
• Always explicit deny take precedence over allow

**VPC Endpoints**  
A VPC endpoint enables customers to privately connect to supported AWS services and VPC endpoint services powered by AWS PrivateLink. ([Reference](https://docs.aws.amazon.com/whitepapers/latest/aws-privatelink/what-are-vpc-endpoints.html))

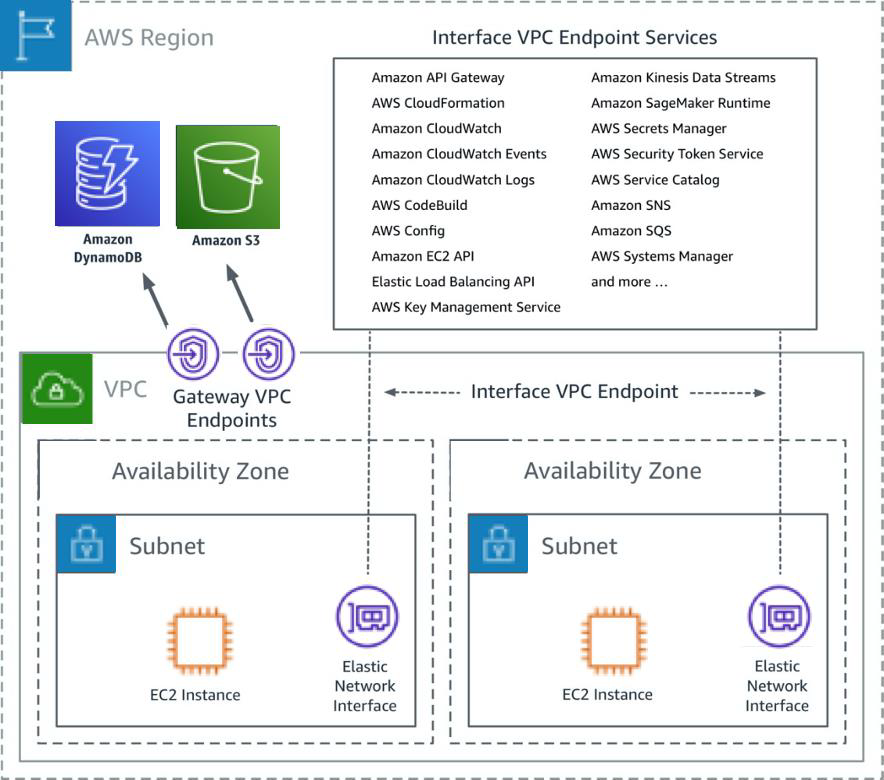
Amazon VPC instances do not require public IP addresses to communicate with resources of the service. Traffic between an Amazon VPC and a service does not leave the Amazon network.

2 types of endpoints  
Gateway endpoints – for Amazon S3 and Dynamo DB  
Interface endpoints – Many services listed below

**VPC Endpoints**  
A VPC endpoint enables customers to privately connect to supported AWS services and VPC endpoint services powered by AWS PrivateLink. ([Reference](https://docs.aws.amazon.com/whitepapers/latest/aws-privatelink/what-are-vpc-endpoints.html))

Amazon VPC instances do not require public IP addresses to communicate with resources of the service. Traffic between an Amazon VPC and a service does not leave the Amazon network.

2 types of endpoints  
Gateway endpoints – for Amazon S3 and Dynamo DB  
Interface endpoints – Many services listed below



**Auto Scaling Groups**

Amazon EC2 Auto Scaling ([Reference](https://docs.aws.amazon.com/autoscaling/ec2/userguide/what-is-amazon-ec2-auto-scaling.html)) helps you ensure that you have the correct number of Amazon EC2 instances available to handle the load for your application. You create collections of EC2 instances, called Auto Scaling groups.

* You can specify the minimum number of instances in each Auto Scaling group, and Amazon EC2 Auto Scaling ensures that your group never goes below this size.
* You can specify the maximum number of instances in each Auto Scaling group, and Amazon EC2 Auto Scaling ensures that your group never goes above this size.
* If you specify the desired capacity, either when you create the group or at any time thereafter, Amazon EC2 Auto Scaling ensures that your group has this many instances.
* If you specify scaling policies, then Amazon EC2 Auto Scaling can launch or terminate instances as demand on your application increases or decreases.

**What is Autoscaling?**  
AWS Auto Scaling lets you build scaling plans that automate how groups of different resources respond to changes in demand. AWS Auto Scaling monitors your application and automatically adds or removes capacity from your resource groups in real-time as demands change.

For example, the following Auto Scaling group has a minimum size of one instance, a desired capacity of two instances, and a maximum size of four instances. The scaling policies that you define adjust the number of instances, within your minimum and maximum number of instances, based on the criteria that you specify.

Auto Scaling components: The following describes the key components of Amazon EC2 Auto Scaling.

**Groups**  
Your EC2 instances are organized into groups so that they can be treated as a logical unit for the purposes of scaling and management. When you create a group, you can specify its minimum, maximum, and desired number of EC2 instances.

**Configuration templates**  
Your group uses a **launch template**, or a launch configuration (not recommended, offers fewer features), as a configuration template for its EC2 instances. You can specify information such as the AMI id, instance type, key pair, security groups, and block device mapping for your instances.

**Scaling options**  
Amazon EC2 Auto Scaling provides several ways for you to scale your Auto Scaling groups. For example, you can configure a group to scale based on the occurrence of specified conditions (dynamic scaling) or on a schedule.

**Launch template**  
Launch Templates is a new capability that enables a new way to templatize your launch requests. Launch Templates streamline and simplify the launch process for Auto Scaling, Spot Fleet, Spot, and On-Demand instances.

**EC2 Auto Scaling lifecycle hooks (**[reference](https://docs.aws.amazon.com/autoscaling/ec2/userguide/lifecycle-hooks-overview.html)**)**  
These hooks can be used to perform actions on the instances based on lifecycle events configured in the Auto Scaling instance lifecycle. A lifecycle hook provides a specified amount of time (defaults to 1hour) to wait for the action to complete before the instance transitions to the next state.

#### Elastic Load Balancing

Elastic Load Balancing ([Reference](https://docs.aws.amazon.com/elasticloadbalancing/latest/userguide/what-is-load-balancing.html)) automatically distributes your incoming traffic across multiple targets, such as EC2 instances, containers, and IP addresses, in one or more Availability Zones.

It monitors the health of its registered targets, and routes traffic only to the healthy targets. Elastic Load Balancing scales your load balancer as your incoming traffic changes over time. It can automatically scale to the vast majority of workloads.

It will assign the requests to the multiple target or server in a round-robin manner by evenly distributing the requests.

**What is Application Load Balancer?**

The Application Load Balancer is a feature of Elastic Load Balancing that allows a developer to configure and route incoming end-user traffic to applications based in the AWS public cloud. Security group rules are used to define the inbound traffic to the load balancer.

**What is network load balancer?**

A Network Load Balancer functions at the fourth layer (transport layer) of the Open Systems Interconnection (OSI) model. It can handle millions of requests per second. It manages TCP/UDP connection requests.

**What is classic load balancer?**

Classic Load Balancer provides basic load balancing across multiple Amazon EC2 instances and operates at both the request level and connection level. Classic Load Balancer is intended for applications that are built within the EC2-Classic network.

**Gateway Load Balancers**

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

Auto Scaling components: The following describes the key components of Amazon EC2 Auto Scaling.

**Groups**  
Your EC2 instances are organized into groups so that they can be treated as a logical unit for the purposes of scaling and management. When you create a group, you can specify its minimum, maximum, and desired number of EC2 instances.

**Configuration templates**  
Your group uses a **launch template**, or a launch configuration (not recommended, offers fewer features), as a configuration template for its EC2 instances. You can specify information such as the AMI id, instance type, key pair, security groups, and block device mapping for your instances.

**Scaling options**  
Amazon EC2 Auto Scaling provides several ways for you to scale your Auto Scaling groups. For example, you can configure a group to scale based on the occurrence of specified conditions (dynamic scaling) or on a schedule.

**Launch template**  
Launch Templates is a new capability that enables a new way to templatize your launch requests. Launch Templates streamline and simplify the launch process for Auto Scaling, Spot Fleet, Spot, and On-Demand instances.

**EC2 Auto Scaling lifecycle hooks (**[reference](https://docs.aws.amazon.com/autoscaling/ec2/userguide/lifecycle-hooks-overview.html)**)**  
These hooks can be used to perform actions on the instances based on lifecycle events configured in the Auto Scaling instance lifecycle. A lifecycle hook provides a specified amount of time (defaults to 1hour) to wait for the action to complete before the instance transitions to the next state.

#### Elastic Load Balancing

Elastic Load Balancing ([Reference](https://docs.aws.amazon.com/elasticloadbalancing/latest/userguide/what-is-load-balancing.html)) automatically distributes your incoming traffic across multiple targets, such as EC2 instances, containers, and IP addresses, in one or more Availability Zones.

It monitors the health of its registered targets, and routes traffic only to the healthy targets. Elastic Load Balancing scales your load balancer as your incoming traffic changes over time. It can automatically scale to the vast majority of workloads.

It will assign the requests to the multiple target or server in a round-robin manner by evenly distributing the requests.

**What is Application Load Balancer?**

The Application Load Balancer is a feature of Elastic Load Balancing that allows a developer to configure and route incoming end-user traffic to applications based in the AWS public cloud. Security group rules are used to define the inbound traffic to the load balancer.

**What is network load balancer?**

A Network Load Balancer functions at the fourth layer (transport layer) of the Open Systems Interconnection (OSI) model. It can handle millions of requests per second. It manages TCP/UDP connection requests.

**What is classic load balancer?**

Classic Load Balancer provides basic load balancing across multiple Amazon EC2 instances and operates at both the request level and connection level. Classic Load Balancer is intended for applications that are built within the EC2-Classic network.

**Gateway Load Balancers**

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

uto Scaling components: The following describes the key components of Amazon EC2 Auto Scaling.

**Groups**  
Your EC2 instances are organized into groups so that they can be treated as a logical unit for the purposes of scaling and management. When you create a group, you can specify its minimum, maximum, and desired number of EC2 instances.

**Configuration templates**  
Your group uses a **launch template**, or a launch configuration (not recommended, offers fewer features), as a configuration template for its EC2 instances. You can specify information such as the AMI id, instance type, key pair, security groups, and block device mapping for your instances.

**Scaling options**  
Amazon EC2 Auto Scaling provides several ways for you to scale your Auto Scaling groups. For example, you can configure a group to scale based on the occurrence of specified conditions (dynamic scaling) or on a schedule.

**Launch template**  
Launch Templates is a new capability that enables a new way to templatize your launch requests. Launch Templates streamline and simplify the launch process for Auto Scaling, Spot Fleet, Spot, and On-Demand instances.

**EC2 Auto Scaling lifecycle hooks (**[reference](https://docs.aws.amazon.com/autoscaling/ec2/userguide/lifecycle-hooks-overview.html)**)**  
These hooks can be used to perform actions on the instances based on lifecycle events configured in the Auto Scaling instance lifecycle. A lifecycle hook provides a specified amount of time (defaults to 1hour) to wait for the action to complete before the instance transitions to the next state.

#### Elastic Load Balancing

Elastic Load Balancing ([Reference](https://docs.aws.amazon.com/elasticloadbalancing/latest/userguide/what-is-load-balancing.html)) automatically distributes your incoming traffic across multiple targets, such as EC2 instances, containers, and IP addresses, in one or more Availability Zones.

It monitors the health of its registered targets, and routes traffic only to the healthy targets. Elastic Load Balancing scales your load balancer as your incoming traffic changes over time. It can automatically scale to the vast majority of workloads.

It will assign the requests to the multiple target or server in a round-robin manner by evenly distributing the requests.

**What is Application Load Balancer?**

The Application Load Balancer is a feature of Elastic Load Balancing that allows a developer to configure and route incoming end-user traffic to applications based in the AWS public cloud. Security group rules are used to define the inbound traffic to the load balancer.

**What is network load balancer?**

A Network Load Balancer functions at the fourth layer (transport layer) of the Open Systems Interconnection (OSI) model. It can handle millions of requests per second. It manages TCP/UDP connection requests.

**What is classic load balancer?**

Classic Load Balancer provides basic load balancing across multiple Amazon EC2 instances and operates at both the request level and connection level. Classic Load Balancer is intended for applications that are built within the EC2-Classic network.

**Gateway Load Balancers**

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

