API Gateway

- WebSocket API @connection: send a callback message to a connected client, get connection information, or disconnect the client
- You can enable IAM authorization for HTTP API routes. When IAM authorization is enabled, clients must use Signature Version 4 (SigV4) to sign their requests with AWS credentials. API Gateway invokes your API route only if the client has **execute-api** permission for the route

REST

- Integrates directly with AWS services e.g. DynamoDB

HTTP

- Lower cost than REST
- Cannot integrate directly with services

AppFlow

- Used to securely transfer data between Software-as-a-Service (SaaS) applications like Salesforce, SAP, Zendesk, Slack, and ServiceNow, and AWS services

Application Discovery Service

- Helps enterprise customers **plan** migration projects by gathering information about their on-premise data centres
- Helpful for determining TCO
- Performs server utilization, data and dependency mapping and collects and presents configuration, usage, and behaviour data from your servers to help you better understand your workloads
- You can export this data as a CSV file and use it to estimate the TCO of running on AWS and to plan your migration to AWS
- **Agentless discovery** – Identifies VMs running on VMware
- **Agent-based discovery** – Used for physical servers and VMs running on Hyper-V

Application Migration Service (MGN)

- Primarily used for the **actual** migration of your on-premise virtual machines to AWS
- Minimizes time-intensive, error-prone manual processes by automatically converting your source servers to run natively on AWS. It also simplifies application modernization with built-in, post-launch optimization options
- Source servers are added by installing the AWS replication agent
- Automatically converts and launches your servers on AWS
- Steps:
    1. Create the Replication Settings template
    2. Add source servers to Application Migration Service by installing the AWS Replication Agent on them. The Agent can be installed on both Linux and Windows servers
    3. Launch a test instance
    4. Cutover to AWS

AppStream 2.0

- A fully managed AWS End User Computing (EUC) service designed to stream SaaS applications and convert desktop applications to SaaS without rewriting code or refactoring the application

AppSync

- **GraphQL**
- Enables developers to connect their applications and services to data and events with secure, serverless and high-performing GraphQL and Pub/Sub APIs
- Can use WebSockets for real-time communications
- Cognito groups can be used with resolvers to provide authorization based on identity
- Well suited for integration with multiple data sources

Athena

- Lets you analyse data in S3 using SQL or Apache Spark
- The Amazon Athena ODBC driver can be used to connect to Athena
- Optimizations:
    1. Store data in a columnar format such as Apache Parquet or Apache ORC
    2. Use Apache Hive partitioning in Amazon S3 using a key that includes a date

Performance tuning

- Partition your data
- Bucket your data
- Use compression
- Optimize file size
- Use columnar file formats

Query tuning

- Optimize ORDER BY
- Optimize joins
- Optimize GROUP BY
- Use approximate functions
- Only include the columns that you need

AWS backup

- Fully-managed service that makes it easy to centralize and automate data protection across AWS services, in the cloud and on-premises
- Can configure backup policies and monitor activity for your AWS resources in one place
- Supports continuous backups and point-in-time recovery (PITR) in addition to snapshot backups

Batch

- Uses ECS not lambda
- Dynamically provisions the optimal quantity and type of compute resources
    1. Job queues are mapped to one or more compute environments. Compute environments contain the Amazon ECS container instances that are used to run containerized batch jobs
    2. A specific compute environment can also be mapped to one or many job queues. Within a job queue, the associated compute environments each have an order that's used by the scheduler to determine where jobs that are ready to be run should run

- Works well with spot instances

Budgets

- Allows organizations to set custom budgets for tracking their AWS spending and usage
- Can be used to monitor resource usage in an account
- Can send notifications when thresholds are breached
- Example: to track the hours of EC2 instance operation, set monitoring interval to daily. Define a budget limit and configure alerts
- Can be used to track Reserved Instance usage and send alerts when the threshold is breached

Budget Alert

- Can trigger SNS which can invoke Lambda for manual remediations

Budgets alert actions

- Can stop EC2 & RDS instances
- Can attach an IAM policy or an SCP
- Everything else must be done via lambda

Cache

Redis

- Supports replication

Memcached

- Does not support replication
- Designed for simplicity

ElastiCache for Redis

- Security:
    1. Configure the security group for the ElastiCache cluster with the required rules to allow inbound traffic from the cluster itself as well as from the cluster's clients on port 6379
    2. Create the cluster with auth-token parameter and make sure that the parameter is included in all subsequent commands to the cluster
    3. Configure the ElastiCache cluster to have both in-transit as well as at-rest encryption

Chime

- Audio, video, chat

Certificate manager

- AWS KMS can be used to encrypt and protect ACM certificates, and the encryption context can be stored alongside the encrypted data
- Automatically renews certificates
- Certificates can only be used in one region unless attached to a CloudFront distribution

Client VPN

- Managed client-based VPN service that enables you to securely access your AWS resources and resources in your on-premises network. With Client VPN, you can access your resources from any location using an OpenVPN-based VPN client
- MFA is supported only if it's enabled in the IdP

Client VPN endpoint

- The resource that enables and manages client VPN sessions. The termination point for all client VPN sessions

Cloud Adoption Readiness Tool (CART)

- Helps organizations of all sizes develop efficient and effective plans for cloud adoption and enterprise cloud migrations
- A 16-question online survey and assessment report details your cloud migration readiness across six perspectives including business, people, process, platform, operations, and security

CloudFormation

- Use CloudFormation Resource Tags property to apply tags to certain resource types upon creation
- **AWS::IAM::InstanceProfile** – creates a new instance profile and contains an array of roles for the EC2 instance to assume

Custom resources

- Enables you to write custom provisioning logic in templates that AWS CloudFormation runs anytime you create, update (if you changed the custom resource), or delete stacks
- DeletionPolicy Options:
    1. Delete
    2. Retain
    3. Snapshot

Dynamic references

- Provide a compact, powerful way for you to specify external values that are stored and managed in other services, such as the Systems Manager Parameter Store, in your stack templates. When you use a dynamic reference, CloudFormation retrieves the value of the specified reference when necessary, during stack and change set operations

Secret rotation

- **AWS::SecretsManager::RotationSchedule** sets the rotation schedule and Lambda rotation function for a secret

CloudFront

- A path pattern (for example, images/*.jpg) specifies to which requests you want the cache behaviour to apply
- Signed cookies allow you to control who can access your content when you don't want to change your current URLs or when you want to provide access to multiple restricted files
- Origin Access Identity limits direct access to S3 to the CloudFront distribution only
- Custom headers can be configured and forwarded to the origin. This can be used to improve security. For example, WAF rule configured on an ALB can be configured to reject any traffic that does not contain the custom header

- You can use *geographic restrictions* to prevent users in specific geographic locations from accessing content that you're distributing through an Amazon CloudFront distribution
- Only accepts HTTP/S traffic
- Supports content uploads via POST, PUT, and other HTTP Methods but there is a limited connection timeout to the origin of 60 seconds
- Force HTTPS by using HTTP to HTTPS Redirect feature
- You can use your own SSL certificates with Amazon CloudFront at no additional charge with Server Name Indication (SNI) Custom SSL
- If you configure CloudFront to serve HTTPS requests using SNI, CloudFront associates your alternate domain name with a dedicated IP address for each edge location. The IP address to your domain name is determined during the SSL/TLS handshake negotiation and isn't dedicated to your distribution
- Offers basic DDOS protection

Origin failover

- Set up an origin failover by creating an origin group with two origins with one as the primary origin and the other as the second origin which CloudFront automatically switches to when the primary origin fails
- Useful for 504 errors

Field Level encryption

- Adds an additional layer of security that lets you protect specific data throughout system processing so that only certain applications can see it
- Field-level encryption allows you to enable your users to securely upload sensitive information to your web servers. The sensitive information provided by your users is encrypted at the edge, close to the user, and remains encrypted throughout your entire application stack

S3

- Cache behaviour can be specified on a partition (prefix) basis
- To ensure S3 data is only available via CloudFront
    1. Create an Origin Access Control (OAC) and associate it with your CloudFront distribution
    2. Change the permissions on your Amazon S3 bucket so that only the origin access control has read permission

HTTPS

To require HTTPS between viewers and CloudFront either:

1. Configure CloudFront to use its default SSL/TLS certificate by changing the **Viewer Protocol Policy** setting for one or more cache behaviours to require HTTPS communication
2. Set the **`Viewer Protocol Policy`** to use **`Redirect HTTP to HTTPS`** or **`HTTPS`** Only

CloudSearch

- Search large collections of data such as webpages, document files, forum posts, or product information
- Closed to new customers

OpenSearch

- Has all the search features of CloudSearch plus a vector engine supporting semantic search
- Search and analytics engine for use cases such as log analytics, real-time application monitoring, and clickstream analysis

- Suitable for dynamic schemas and semi structured JSON

OpenSearch Dashboards

- Open-source visualization tool designed to work with OpenSearch.
- Provides an installation of OpenSearch Dashboards with every OpenSearch Service domain

OpenSearch Serverless

- Allows you deploy and use OpenSearch through a REST endpoint
- You send your documents to OpenSearch Serverless, which indexes them for search using the OpenSearch REST API

OpenSearch Managed clusters deployment

- With managed clusters, you get granular control over the instances you would like to use, indexing and data-sharding strategy, and more

CloudTrail

- CloudTrail is an AWS service that helps you enable governance, compliance, and operational and risk auditing of your AWS account
- Actions taken by a user, role, or an AWS service are recorded as events in CloudTrail
- Provides a history of both API and non-API account activity made through the AWS Management Console, AWS SDKs, command line tools, and other AWS services
- CloudTrail records management events for the last 90 days free of charge, these are viewable in the Event History within the CloudTrail console. For Amazon S3 delivery of CloudTrail events, the first copy delivered is free. Additional copies of management events are charged
- Optionally, you can enable AWS CloudTrail Insights on a trail to help you identify and respond to unusual activity
- By default CloudTrail only records logs for the region it was created it. Use **`--is-multi-region-trail`** to enable for all regions

Data events

- CloudTrail data events are disabled by default
- Data events aren't viewable in CloudTrail event history and are charged for all copies at a reduced rate compared to management events

CloudWatch agent

- To integrate a server with AWS Systems Manager:
    1. Create an IAM role or user
    2. Download the agent package
    3. Modify the configuration file
    4. Install and start the agent on the server

CodeBuild

- S3 cache can store reusable pieces of the build environment and use them across multiple builds

CodePipeline

- When creating a cross-account deployment from account A to account B

1. In account B, create a service role that includes the permissions to deploy the required services
2. In account A, create a customer-managed KMS key that grants usage permissions to account A's CodePipeline service role and account B. Also, create an S3 bucket with a bucket policy that grants account B access to the bucket
3. In account B, create a cross-account IAM role. In account A, add the **AssumeRole** permission to account A's CodePipeline service role to allow it to assume the cross-account role in account B

Cognito

- User pools are for authentication. Users sign in through the user pool or federate through a third-party IdP. Identity pools are for authorization. You use identity pools to create unique identities for users and give them access to other AWS services
- Provides authentication, authorization, and user management for your web and mobile apps
- Users can sign in directly with a username and password or through a third party such as Facebook, Amazon, Google, or Apple
- You can use identity pools and user pools separately or together

User pools

- User pools are user directories that provide sign-up and sign-in options for your app users

Identity pools

- Enable you to grant your users access to other AWS services
- Provide temporary AWS credentials for users who are guests (unauthenticated) and for users who have been authenticated and have received a token

Comprehend

- Natural language processing (NLP) service that uses machine learning to uncover information in unstructured data and text within documents
- Derive and understand valuable insights from text within documents

Connect

- Managed call centre
- Custom buttons can be added to the Contact Control Panel

Config

- Allows you to assess, audit, and evaluate the configurations of your AWS resources
- Config continuously monitors and records your AWS resource configurations and allows you to automate the evaluation of recorded configurations against desired configurations
- AWS Config provides a detailed view of the configuration of AWS resources in your AWS account. This includes how the resources are related to one another and how they were configured in the past so that you can see how the configurations and relationships change over time
- Allows you to remediate noncompliant resources that are evaluated by AWS Config Rules
- You can use Config to answer questions such as - “What did my AWS resource look like at xyz point in time?”
- Does not generally auto-remediate rule violations, use Lambda to do this

Conformance pack

- A conformance pack is a collection of AWS Config rules and remediation actions that can be easily deployed as a single entity in an account and a Region or across an organization in AWS Organizations
- If a new account joins an organization, the rule or conformance pack is deployed to that account. When an account leaves an organization, the rule or conformance pack is removed

Container insights

- Gathers metrics for containerized workloads running on either ECS or EKS
- Supports Fargate
- Adds insights beyond cluster level EC2 metrics
- Container-level resource insights for nodes (EKS), tasks (ECS) and pods (EKS) for CPU, memory, storage and network utilization
- Provides metrics on **EKS** container restart failures

Control Tower

- **AWSControlTowerServiceRolePolicy** enables an administrator to manage AWS Control Tower only
- **AWSSecurityAuditors** group gives its members read-only permission

Databases

Aurora

- **Is an RDS database engine**
- Multi-master (read/write in multiple AZs) is available only for Aurora
- Modern relational database service offering performance and high availability at scale, fully open-source MySQL and PostgreSQL compatible
- Cross region MySQL read replica: replica can be promoted to stand alone cluster in the event of a disaster scenario
- MySQL error is log is generated by default
- Slow query and general logs can be generated by setting parameters in the DB parameter group
- Multi-AZ deployments follow synchronous replication
- In the event of a master failure Read Replicas are automatically promoted to be the primary instance
- Automatically scales storage

Aurora auto scaling

- Operates on the replicas
- You cannot set Auto Scaling for the master database on Amazon Aurora. You can only manually resize the instance size of the master node

Aurora global database

- Global reads, single region write
- Amazon Aurora global databases span multiple AWS Regions, enabling low latency global reads and providing fast recovery from the rare outage that might affect an entire AWS Region
- Consists of one *primary* AWS Region where your data is written, and up to five read-only *secondary* AWS Regions
- Secondary clusters can be scaled independently, by adding one or more Aurora Replicas

- An Aurora replica is a read only Aurora DB instance
- **Write forwarding** reduce the number of endpoints that applications need to manage. With write forwarding enabled, secondary clusters in an Aurora global database forward write SQL statements to the primary cluster
- To recover the database after an outage in the primary Region, use *global database failover*
- **Global database switchover** relocates the primary cluster of a healthy Aurora global database to one of its secondary Regions with no data loss

RDS

- Cross region read replicas allow replicas to be promoted to a standalone instance. In the case of failure the promoted instance can then create a read replica to ensure high availability
- Uptime SLA of 99.5%.
- Multi-AZ for high availability
- Multi-AZ cannot span regions
- Auto-scaling to handle traffic increases/decreases
- Single AZ RDS unavailable during vertical scaling, minimal downtime for multi-AZ RDS
- The only way to unencrypt an encrypted database is to export the data and import the data into another DB instance. You cannot create unencrypted snapshots of encrypted DB instances, and you cannot create unencrypted read replicas of an encrypted DB instance.
- Oracle on RDS still requires a license
- Read Replicas can be manually promoted to a standalone database instance
- Can automatically scale storage by enabling the **storage auto-scaling** option
- During maintenance window upgrades the canonical name record (CNAME) is changed from the primary database to the standby database
- Sharding can improve performance by distributing operations across different database instances
- **Source-Replica Replication for MySQL** is a replication setup where a primary database (the source) continuously replicates its data to one or more replicas

Transaction logs

- You can list the transaction log backup files for a database and copy them to a target Amazon S3 bucket
- By copying transaction log backups in an Amazon S3 bucket, you can use them in combination with full and differential database backups to perform point in time database restores
- Use RDS stored procedures to set up access to transaction log backups, list available transaction log backups, and copy them to your Amazon S3 bucket

Oracle

- RDS does not support certain features in Oracle such as Multitenant Database, Real Application Clusters (RAC), Unified Auditing, Database Vault, and many more

RDS proxy

- Fully managed, highly available database proxy for Amazon RDS that makes applications more scalable and more resilient to database failures
- Maintains a pool of established connections to your RDS database instances, reducing the stress on database compute and memory resources that typically occurs when new connections are established
- Transparently tolerates database failure. Automatically routes traffic to a new database instance while preserving application connections
- Bypasses DNS caches to reduce failover times by up to 66% for Aurora Multi-AZ databases
- Can queue write requests while a failover takes place

Data Lifecycle Manager

- Can take snapshots of EBS volumes

Database Migration Service

- Can migrate data to and from most of the widely used commercial and open-source databases
- Supports homogeneous migrations such as Oracle to Oracle, as well as heterogeneous migrations between different database infrastructures, such as Oracle to Aurora
- Replicates ongoing changes (CDC)
- Can migrate data directly to Redshift

DataSync

- AWS DataSync is a secure, online service that automates and accelerates moving data between on premises and AWS Storage services. DataSync can copy data between Network File System (NFS) shares, Server Message Block (SMB) shares, Hadoop Distributed File Systems (HDFS), self-managed object storage, AWS Snowcone, Amazon Simple Storage Service (Amazon S3) buckets, Amazon Elastic File System (Amazon EFS) file systems, Amazon FSx for Windows File Server file systems, Amazon FSx for Lustre file systems, Amazon FSz for OpenZFS file systems, and Amazon FSx for NetApp ONTAP file systems

Direct Connect

- Provides a **consistent & private** connection
- Is not encrypted, add a VPN running on DX for a secure connection
- To run an IPSec VPN over the top of a DX connection it is necessary to use a public VIF
- Requires a network device in your data centre that supports BGP and BGP MD5 authentication
- Requires a virtual interface (VIF):
    1. Private VIF is used to access an Amazon VPC using private IP addresses. To access services using a private IP an interface endpoint is required
    2. Public VIF can access all AWS public services using public IP addresses
    3. Transit VIF should be used to access one or more Amazon VPC Transit Gateways associated with Direct Connect gateways

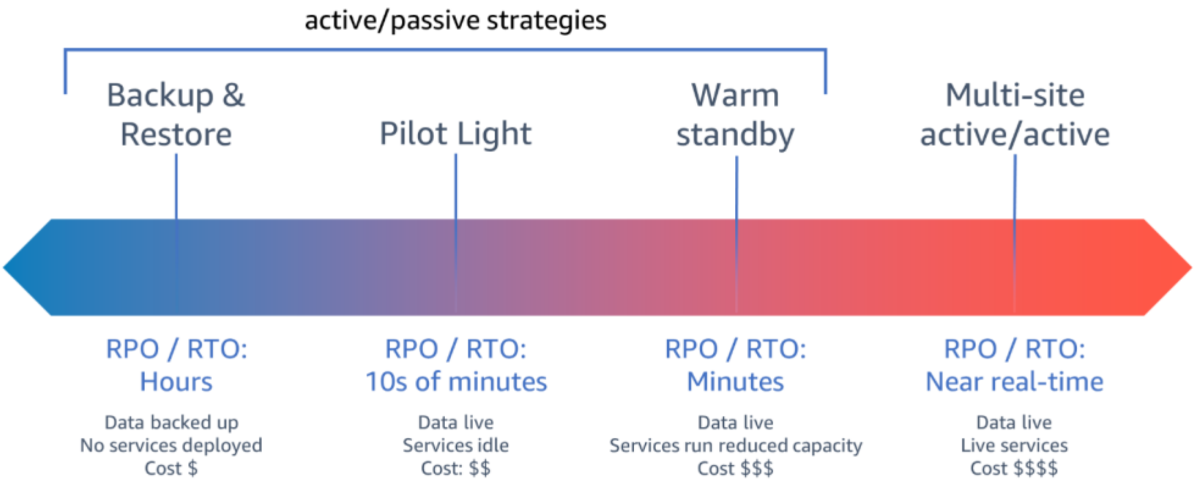
Link aggregation group

- You can use multiple connections to increase available bandwidth
- A LAG is a logical interface that uses the Link Aggregation Control Protocol (LACP) to aggregate multiple connections at a single AWS Direct Connect endpoint, allowing you to treat them as a single, managed connection
- LAGs streamline configuration because the LAG configuration applies to all connections in the group
- You can create a LAG for connections that terminate on the same AWS device and in the same location
- All connections must be dedicated connections and have a port speed of 1 Gbps, 10 Gbps, 100 Gbps, or 400 Gbps
- All connections in the LAG must use the same bandwidth
- All connections in the LAG must terminate at the same AWS Direct Connect endpoint
- LAGs are supported for all virtual interface types—public, private, and transit

Direct connect gateway

- DX gateway connects to either a transit gateway or a virtual private gateway
- Allows you to connect a DX gateway to multiple AWS Regions

Disaster recovery



DocumentDB

- Supports neither interface nor gateway VPC endpoint

DynamoDB

- Reserved capacity offers significant savings if read-and-write throughput can be predicted. Any throughput you provision in excess of your reserved capacity is billed at standard rates for provisioned throughput.
- 400kb maximum record size
- Extremely rapid spikes in load can outpace auto-scaling's ability to scale the table
- The partition key of an item is also known as its hash attribute
- Two types of primary keys:
    1. **Partition key** – A simple primary key, composed of one attribute known as the partition key. DynamoDB uses the partition key's value as input to an internal hash function. The output from the hash function determines the partition (physical storage internal to DynamoDB) in which the item will be stored
    2. **Partition key and sort key** – Referred to as a composite primary key, this type of key is composed of two attributes. The first attribute is the partition key, and the second attribute is the sort key

DynamoDB Global tables

- The only database solution that allows writes in multiple regions
- Fully managed, multi-Region, and multi-active database option that delivers fast and localized read and write performance for massively scaled global applications
- You can specify the AWS Regions where you want the tables to be available and DynamoDB will propagate ongoing data changes to all of them
- Global tables are available in all Regions
- Multi-AZ deployments follow synchronous replication

EBS

- gp2 provides 3 IOPS/GB
- Throughput Optimized HDD has maximum of 500 IOPS
- Cold HDD is the lowest cost HDD volume and is designed for less frequently accessed workloads. Has good throughput
- EBS direct API can be used to create EBS snapshots, write data directly to your snapshots, read data on your snapshots, identify the differences or changes between two snapshots and copy snapshots to S3

EC2

- Termination protection does not stop EC2 autoscaling from terminating an instance
- To prevent an ASG from terminating an instance on scale-in use instance scale-in protection
- To prevent EC2 auto scaling from terminating an unhealthy instance suspend the **terminate** process or the **ReplaceUnhealthy** process
- Cannot detach a primary network interface from an instance
- To move an instance to a different subnet: Launch a new instance in the new subnet via an AMI created from the old instance. Direct traffic to this new instance using Route 53 and then terminate the old instance
- Elastic IP address is a public IPv4 address, which is reachable from the Internet
- Use instance connect to connect to port 22 SSH via the console or CLI but only supports Linux EC2

Placement groups

- Before an instance is removed or added to a placement group it must be in the stopped state

Cluster placement group

- Recommended for applications that benefit from low network latency or high network throughput
- Also recommended when most of the network traffic is between the instances in the group
- To provide the lowest latency and the highest packet-per-second network performance, choose an instance type that supports enhanced networking

Spread placement group

- A group of instances that are each placed on distinct racks, with each rack having its own network and power source
- Recommended for applications that have a small number of critical instances that should be kept separate from each other
- Reduces the risk of simultaneous failures that might occur when instances share the same racks.

Instance role

- Retrieves credentials from the instance metadata

Traffic mirroring

- Allows you to copy traffic passing through an elastic network adaptor and send it toward another instance for further investigation
- Can be used to inspect packets

ASG

- Amazon EC2 Auto Scaling creates a new scaling activity for terminating an unhealthy instance and then terminates it. Later, another scaling activity launches a new instance to replace the terminated instance

- If an Availability Zone is unbalanced, Amazon EC2 Auto Scaling will compensate by rebalancing the Availability Zones. When rebalancing, Amazon EC2 Auto Scaling launches new instances before terminating the old ones, so that rebalancing does not compromise the performance or availability of your application
- You can only specify one launch configuration for an Auto Scaling group at a time, and you can't modify a launch configuration after you've created it
- When a launch configuration for an Auto Scaling group is changed, any new instances are launched using the new configuration options, but existing instances are not affected

Launch configuration

- Configuration template that an Auto Scaling group uses to launch EC2 instances
- The default value for the instance placement tenancy is null and the instance tenancy is controlled by the tenancy attribute of the VPC
- If you set the Launch Configuration Tenancy to default and the VPC Tenancy is set to dedicated, then the instances have dedicated tenancy
- If you set the Launch Configuration Tenancy to dedicated and the VPC Tenancy is set to default, then again, the instances have dedicated tenancy

Spot fleet

- EC2 Fleet and Spot Fleet are designed to be a useful way to launch a fleet of tens, hundreds, or thousands of EC2 instances in a single operation
- Each instance in a fleet is either configured by a launch template or a set of launch parameters that you configure manually at launch

Reserved instances

- If you purchased a zonal Reserved Instance for a specific Availability Zone, you must launch the instance into the same Availability Zone
- If you purchased a regional Reserved Instance you can launch the instance into any Availability Zone in the Region that you specified for the Reserved Instance
- Use **ModifyReservedInstances** API to change AZ, instance size & networking type
- Only RI can change instance family & operating system, tenancy

ECS

- Spot instance draining shuts down tasks upon receipt of the two-minute interruption notice
- Using an ECS cluster running behind an Application Load Balancer offers advantages such as ALBs allow containers to use dynamic host port mapping so that multiple tasks from the same service are allowed per container instance. This reduces the number of instances required for migration and therefore reduces the overall costs
- Migrating applications is not low-overhead as they will require refactoring
- Task role: permissions granted in the IAM role are assumed by the containers running in the task. This role allows your application code to use other AWS services
- Container instance role: **AmazonEC2ContainerServiceforEC2Role** contains the permissions to use the ECS feature set
- Security groups can be assigned to individual tasks

EFS

- Simple, scalable, fully managed elastic NFS file system for use with AWS Cloud services and on-premise resources
- Regional service storing data within and across multiple Availability Zones
- POSIX compliant

- You can connect to EFS from EC2 instances in other AWS Regions using an inter-Region VPC peering connection, and from on-premises servers using an AWS VPN connection
- **Max I/O mode** - optimized for high levels of throughput and I/O operations but has higher latency

Elastic Beanstalk

- Java, .NET, PHP, Node.js, Python, Ruby, Go, and Docker web applications are supported

EKS

- Using topology spread constraints based on Availability Zones is a strategic approach to enhance node resilience in an Amazon EKS cluster
- Use CloudWatch Container Insights for monitoring and collecting container logs and metrics
- Fluent Bit can be used as a log router to send logs to CloudWatch Logs, and can also send metrics to CloudWatch metrics
- A low overhead way to migrate on-premise application servers to the cloud

Elastic Disaster Recovery

- Minimizes downtime and data loss with fast, reliable recovery of on-premise and cloud-based applications using affordable storage, minimal compute, and point-in-time recovery
- Ongoing replication of data
- Can launch recovery instances on AWS within minutes, using the most up-to-date server state or a previous point in time
- After your applications are running on AWS, you can choose to keep them there, or you can initiate data replication back to your primary site when the issue is resolved
- Fail back to primary site when issue is resolved
- Must install the AWS Replication Agent on each source server that you want to add to EDR

Elastic Load balancer (ELB)

- Can only distribute requests across AZs not regions
- Sits inside a subnet
- If a static IP is required but an ALB is also required register an NLB in front of an ALB
- To route domain traffic to an ELB, use Amazon Route 53 to create an **alias record** that points to the load balancer. An **alias record** is the only valid record type
- Supports multiple TLS/SSL certificates on one ALB using Server Name Indication (SNI)
- The source port of an ALB node is a dynamically defined high number port between 1024 and 65535.

Network Load Balancer

- Can have a static IP
- Can have security group assigned, default SG allows all access
- Passes connections straight to EC2 instances with the source IP of the client preserved
- Can handle millions of requests per second
- Optimized to handle sudden and volatile traffic patterns
- Uses flow hash routing algorithm

Application Load Balancer

- Cannot have a static IP
- Can have security groups assigned

- To allow requests only from CloudFront: add a security group rule to the ALB to allow traffic from the AWS managed prefix list for CloudFront only
- Best suited for HTTP and HTTPS
- Only supports HTTP & HTTPS
- Can route traffic based on the domain name specified in the Host header
- ALB offers support for Path conditions. You can configure rules for your listener that forward requests based on the URL in the request. This enables you to structure your application as smaller services, and route requests to the correct service based on the content of the URL
- Supports hostname routing e.g. ***.ecomm.com & ecomm.com**
- Supports pathname routing e.g. **/img/*/pics**
- Routing algorithms: round robin & least outstanding requests
- Use least outstanding requests when requests vary in computation complexity

Elastic Map Reduce

- Managed Hadoop framework
- Use **S3DistCP** to combine many small files, can also be used to bulk move files between HDFS and S3

- **Master node**: A node that manages the cluster by running software components to coordinate the distribution of data and tasks among other nodes for processing. The master node tracks the status of tasks and monitors the health of the cluster. Every cluster has a master node, and it's possible to create a single-node cluster with only the master node

- **Core node**: A node with software components that run tasks and stores data in the Hadoop Distributed File System (HDFS) on your cluster. Multi-node clusters have at least one core node

- **Task node**: A node with software components that only runs tasks and does not store data in HDFS. Task nodes are optional

EMR cost optimization

- Spot node termination on the master node will terminate the entire cluster
- Avoid using Spot Instances for Core nodes if the jobs on the cluster use HDFS. That prevents a situation where Spot interruptions cause data loss for data that was written to the HDFS volumes on the instances
- Use Spot Instances for task nodes by selecting up to five instance types that match your hardware requirement. EMR fulfils the most suitable capacity by price and capacity availability

Elemental MediaConvert

- File-based video transcoding service with broadcast-grade features
- Create live stream content for broadcast and multi-screen delivery at scale
- A family of video streaming protocols including Apple's HTTP Live Streaming (HLS), Dynamic Adaptive Streaming over HTTP (DASH), Microsoft's Smooth Streaming (MSS), and Adobe's HTTP Dynamic Streaming (HDS) improves the user experience and minimises cost
- When using the adaptive streaming protocols, a manifest file is generated which should be the source of the CloudFront distribution

End User Messaging

- Empowers developers to integrate scalable and reliable messaging capabilities into their applications. Whether it's time-sensitive alerts, one-time passwords, or two-way communications
- Multiple channels like SMS, MMS, push, and text to voice

Fargate

- To enable internet access:
  1. Public subnet: set **Auto-assign public IP** to **ENABLED**
  2. Private subnet: set **Auto-assign public IP** to **DISABLED** and configure a NAT gateway

Federated web identity using IdP and SAML

- Ensure that the appropriate IAM roles are mapped to company users and groups in the IdP's SAML assertions
- Ensure that the trust policy of the IAM roles created for the federated users or groups has set the SAML provider as the principal
- Ensure that the ARN of the SAML provider, the ARN of the created IAM role, and SAML assertion from the IdP are all included when the federated identity web portal calls the AWS STS **AssumeRoleWithSAML** API

Firewall manager

- Security management service that allows you to centrally configure and manage firewall rules across your accounts and applications in AWS Organizations

FSx

- Fully managed file storage service
- For NetApp ONTAP, OpenZFS, Windows File Server, and Lustre
- You can test the failover of your multi-AZ file system by modifying its throughput capacity. When you modify your file system's throughput capacity, Amazon FSx switches out the file system's file server. Multi-AZ file systems automatically fail over to the secondary server while Amazon FSx replaces the preferred server file server first
- Monitor storage capacity and file system activity using Amazon CloudWatch, and monitor end-user actions with file access auditing using Amazon CloudWatch Logs

FSx for Lustre

- Large-scale, distributed parallel file system powering the workloads of most of the largest supercomputers
- POSIX compliant
- When linked to an Amazon S3 bucket, an FSx for Lustre file system transparently presents S3 objects as files. The file system also makes it possible for you to write file system data back to S3

FSx for Windows File Server

- Fully managed Microsoft Windows file servers, backed by a fully native Windows file system
- Supports **SMB** protocol to access file storage over a network
- Supports Windows ACLs to control access to file contents
- File system storage capacity can only be increased, not decreased
- **FreeStorageCapacity** metric is used to monitor free file space

- **update-file-system** command is used to increase file system capacity

Glue

- Compression: you can read and write bzip and gzip archives containing ORC files from S3

Global accelerator

- AWS Global Accelerator is a networking service that helps you improve the availability and performance of the applications offered to global users
- Provides you with a set of two static IPv4 addresses
- Provides static IP addresses providing a fixed entry point to your applications, eliminating the complexity of managing specific IP addresses for different AWS Regions and Availability Zones
- Provides static IP addresses as a core feature, which can be used by customers for their firewall allow lists
- Routes traffic via the private AWS global network, reducing in-game latency, jitter, and packet loss
- Standard accelerator directs traffic over the AWS global network to endpoints in the nearest Region to the client
- Can be associated with ALBs
- With AWS Global Accelerator, you can shift traffic gradually or all at once between the blue and the green environment and vice-versa without being subject to DNS caching on client devices and internet resolvers, traffic dials and endpoint weight changes are effective within seconds
- Provides automatic DDOS protection and mitigation
- A *custom routing accelerator* in AWS Global Accelerator lets you use custom application logic to direct one or more users to a specific destination among many destinations, while using the AWS global network to improve the availability and performance of your application
- Accepts non-HTTP traffic
- Well suited for cross-region failover

Custom routing accelerator

- Maps one or more users to a specific destination among many destinations
- Allows you to use your own application logic to deterministically route one or more users to a specific Amazon EC2 instance destination in a single or multiple AWS Regions. This is useful for use cases where you want to control which session on an EC2 instance your user traffic is sent to
- Supports only VPC subnet endpoints, not ALBs
- Supports only IPv4

GuardDuty

- Intelligent threat detection service that continuously monitors your AWS accounts, EC2 instances, EKS clusters, and data stored in S3 for malicious activity without the use of security software or agents
- Generates detailed security findings when suspicious/malicious behaviour is detected
- Monitors for threats, does not block them
- In an AWS organization, the management account can designate any account within this organization as the delegated GuardDuty administrator account.

HSM

- To enable Quorum Authentication Mechanism: Use the **cloudhsm_mgmt_util** command line tool, enable encrypted communication, login as a CO, and set the Quorum minimum value to two using the **setMValue** command. An asymmetric key must also be registered for signing with the **registerMofnPubKey** command

- SSL can be offloaded to HSM

IAM

- ACM is the recommended place to store certificates but in regions where it is not supported certificates should be stored in IAM

Managed policies

SecurityAudit

- Grants access to read security configuration metadata. It is useful for software that audits the configuration of an AWS account

PowerUserAccess

- Provides full access to AWS services and resources but does not allow management of Users and groups

```
-   {
-     "Version" : "2012-10-17",
-     "Statement" : [
-       {
-         "Effect" : "Allow",
-         "NotAction" : [
-           "iam:*",
-           "organizations:*",
-           "account:*"
-         ],
-         "Resource" : "*"
-       },
-       {
-         "Effect" : "Allow",
-         "Action" : [
-           "account:GetAccountInformation",
-           "account:GetPrimaryEmail",
-           "account:ListRegions",
-           "iam:CreateServiceLinkedRole",
-           "iam:DeleteServiceLinkedRole",
-           "iam:ListRoles",
-           "organizations:DescribeOrganization"
-         ],
-         "Resource" : "*"
-       }
-     ]
-   }
-
-
```

Policy

- Username policy variable can be used to restrict access to resources containing the IAM principal’s username

Statement

```
-   {
-     "Version": "2012-10-17",
-     "Statement": [
-       {
```

```
-         "Effect": "Allow",
-         "Action": "ec2:Describe*",
-         "Resource":" *"
-       },
-       {
-         "Effect": "Deny",
-         "Action": "s3:*",
-         "Resource": "*"
-       }
-     ]
-   }
```

Confused deputy

- Create an IAM role in your AWS account with a trust policy that trusts the Partner
- Take a unique external ID value from the partner and include this external ID condition in the role's trust policy

Managed policies

- **PowerUserAccess** provides full access to AWS services and resources but does not allow management of Users and groups

Permissions boundary

- Uses a managed policy to set the maximum permissions that an identity-based policy can grant to an IAM entity

IAM access analyzer

- IAM Access Analyzer analyses your AWS CloudTrail logs to identify actions and services that have been used by an IAM entity (user or role) within your specified date range. It then generates an IAM policy that is based on that access activity

IAM identity centre

- Works for any application that supports SAML 2.0
- AWS solution for connecting your workforce users to AWS managed applications such as Amazon Q Developer and Amazon QuickSight, and other AWS resources
- Can create users or pull them from service providers e.g. OKTA
- You can connect your existing identity provider and synchronize users and groups from your directory or create and manage your users directly in IAM Identity Centre. You can then use IAM Identity Centre for access to applications and accounts
- Supports only SAML 2.0–based applications
- Permission set is a template that defines a collection of one or more IAM policies
- Cannot be used for mobile applications

Inspector

- Automated security assessment service that enables you to understand and improve the security and compliance of applications deployed on AWS

Host assessment rules packages

- Checks EC2 software and configuration for vulnerabilities and deviations from best practice

Network Reachability rules package

- Analyses your VPC network configuration to determine whether your EC2 instances can be reached from external networks such as the Internet, a virtual private gateway, AWS Direct Connect, or from a peered VPC. In other words, it informs you of potential external access to your hosts

IoT core

- Managed cloud service that enables connected devices to securely interact with cloud applications and other devices
- Allows you to connect billions of IoT devices and route trillions of messages to AWS services without managing infrastructure
- Supports:
    1. MQTT (Message Queuing and Telemetry Transport)
    2. MQTT over WSS
    3. HTTPS
    4. LoRaWAN

IoT device defender

- Security management across your IoT devices and fleets
- Makes it easy to audit configurations, authenticate devices, detect anomalies, and receive alerts to help secure your IoT device fleet

IoT device management

- Helps you register, organize, monitor, and remotely manage IoT devices at scale

IoT Greengrass

- Open-source IoT edge runtime and cloud service
- Can be used to build software that enables your devices to act locally on the data that they generate, run predictions based on machine learning models, and filter and aggregate device data
- IoT Greengrass enables your devices to collect and analyse data at the edge, react autonomously to local events, and communicate securely with other devices on the local network
- Can perform ML inference on edge devices on locally generated data using cloud-trained models

Kinesis

Kinesis Client Library (KCL)

- Helps you consume and process data from a Kinesis data stream by taking care of many of the complex tasks associated with distributed computing. These include load balancing across multiple consumer application instances, responding to consumer application instance failures, checkpointing processed records, and reacting to re-sharding
- For each Amazon Kinesis Data Streams application, KCL uses a unique lease table (stored in DynamoDB) to keep track of the shards in a KDS data stream that are being leased and processed by the workers of the KCL consumer application
- Each KCL application must use its own DynamoDB table

Data Firehose

- Capture, transform, and deliver data streams into Amazon S3, Amazon Redshift, Amazon OpenSearch Service, Splunk, Snowflake, and other 3rd party analytics services
- Cannot write directly to DynamoDB
- Data can be transformed using a lambda built into the delivery stream
- Does not use kinesis clients; delivers data directly to the destination

Kinesis Data Analytics

- Collect and process large streams of data records in real time
- You cannot write directly to a KDA stream, only supports the following streaming sources:
    1. A Kinesis data stream
    2. A Kinesis Data Firehose delivery stream

Kinesis Data Streams

- Massively scalable and durable real-time data streaming service
- The data collected is available in milliseconds to enable real-time analytics use cases such as real-time dashboards, real-time anomaly detection, dynamic pricing, and more
- Near real time analytics
- Multiple applications can consume from the stream concurrently
- Highly durable, performs synchronous replication of your streaming data across three Availability zones
- Stores data for up to 365 days

Kinesis Client Library (KCL)

- One of the methods of developing custom consumer applications that can process data from KDS

Kinesis Video Streams

- Amazon Kinesis Video Streams makes it easy to securely stream video from connected devices to AWS for analytics, machine learning (ML), playback, and other processing. Kinesis Video Streams automatically provisions and elastically scales all the infrastructure needed to ingest streaming video data from millions of devices

KMS

- KMS keys are specific to an AWS Region

Lambda

- Alias traffic shifting can be used to achieve canary releases
- TooManyRequestsException (429) means the Lambda has been throttled due to too many concurrent invocations
- By default, the burst concurrency for Lambda functions is between 500-3000 requests per second
- The following example AWS CLI command points an alias to a new version, weighted at 5% (original version at 95% of traffic):

```
aws lambda update-alias --function-name myfunction --name myalias --routing-config
'{"AdditionalVersionWeights" : {"2" : 0.05} }'
```

Lambda@edge

- Allows you to run Node.js and Python Lambda functions to customize content that CloudFront delivers, executing the functions in AWS locations closer to the viewer. The functions run in response to CloudFront events
- Can serve up to 10,000 requests per second

Lex

- AI chatbot
- Voice & text

License Manager

- Not used for storing licenses
- Used to create customized licensing rules that emulate the terms of their licensing agreements and then enforce these rules

AWS Managed Services

- Managed services provided by AWS
- A request for change, or RFC, is how you make a change in your AMS-managed environment
- AMS manged-RFC mode allows customers to make changes to their infrastructure in a controlled manner, and once deployed, the application will be under prescriptive AMS governance and operations, which ensures that the application is being managed in a standardized way

Migration Hub

- Provides a single location to track the progress of application migrations across multiple AWS and partner solutions
- Allows you to choose the AWS and partner migration tools that best fits your needs while providing visibility into the status of migrations across your portfolio of applications
- Application Discovery Service can be integrated with AWS Migration Hub

MQ

- Managed message broker for Apache ActiveMQ Classic and RabbitMQ

NAT Instance vs NAT Gateway

NAT Instance

- The timeout behaviour of a NAT instance is that, when there is a connection time out, it sends a FIN packet to resources behind the NAT instance to close the connection. It does not attempt to continue the connection. For better performance, use a n instead

NAT Gateway

- When there is a connection time out, a NAT gateway returns an RST packet to any resources behind the NAT gateway that attempt to continue the connection (it does not send a FIN packet)
- NAT gateway in a private subnet will allow access to the internet
- Allows instances in a private subnet to connect to services outside your VPC but external services cannot initiate a connection with those instances
- Can have Elastic IP attached
- When you create a NAT gateway, you specify one of the following connectivity types:
    1. **Public** – (Default) Instances in private subnets can connect to the internet through a public NAT gateway but cannot receive unsolicited inbound connections from the internet. You create a public NAT gateway in a public subnet and must associate an elastic IP address with the NAT gateway at creation. You route traffic from the NAT gateway to the internet gateway for the VPC. Alternatively, you can use a public NAT gateway to connect to other VPCs or your on-premises network. In this case, you route traffic from the NAT gateway through a transit gateway or a virtual private gateway
    2. **Private** – Instances in private subnets can connect to other VPCs or your on-premises network through a private NAT gateway. You can route traffic from the NAT gateway through a transit gateway or a virtual private gateway. You cannot associate an elastic IP address with a private NAT gateway. You can attach an internet gateway to a VPC with a private NAT gateway, but if you route traffic from the private NAT gateway to the internet gateway, the internet gateway drops the traffic
- Both private and public NAT gateways map the source private IPv4 address of the instances to the private IPv4 address of the NAT gateway, but in the case of a public NAT gateway, the internet gateway then maps the private IPv4 address of the public NAT Gateway to the Elastic IP address associated with the NAT Gateway. When sending response traffic to the instances, whether it's a public or private NAT gateway, the NAT gateway translates the address back to the original source IP address
- All traffic passing through a NAT gateway will appear to originate from the same IP address
- Important:
    1. You can use either a public or private NAT gateway to route traffic to transit gateways and virtual private gateways
    2. If you use a private NAT gateway to connect to a transit gateway or virtual private gateway, traffic to the destination will come from the private IP address of the private NAT gateway
    3. If you use a public NAT gateway to connect to a transit gateway or virtual private gateway, traffic to the destination will come from the private IP address of the public NAT gateway. The public NAT gateway will only use its EIP as the source IP address when used in conjunction with an internet gateway in the same VPC

Internet Gateway

- Cannot have Elastic IP attached

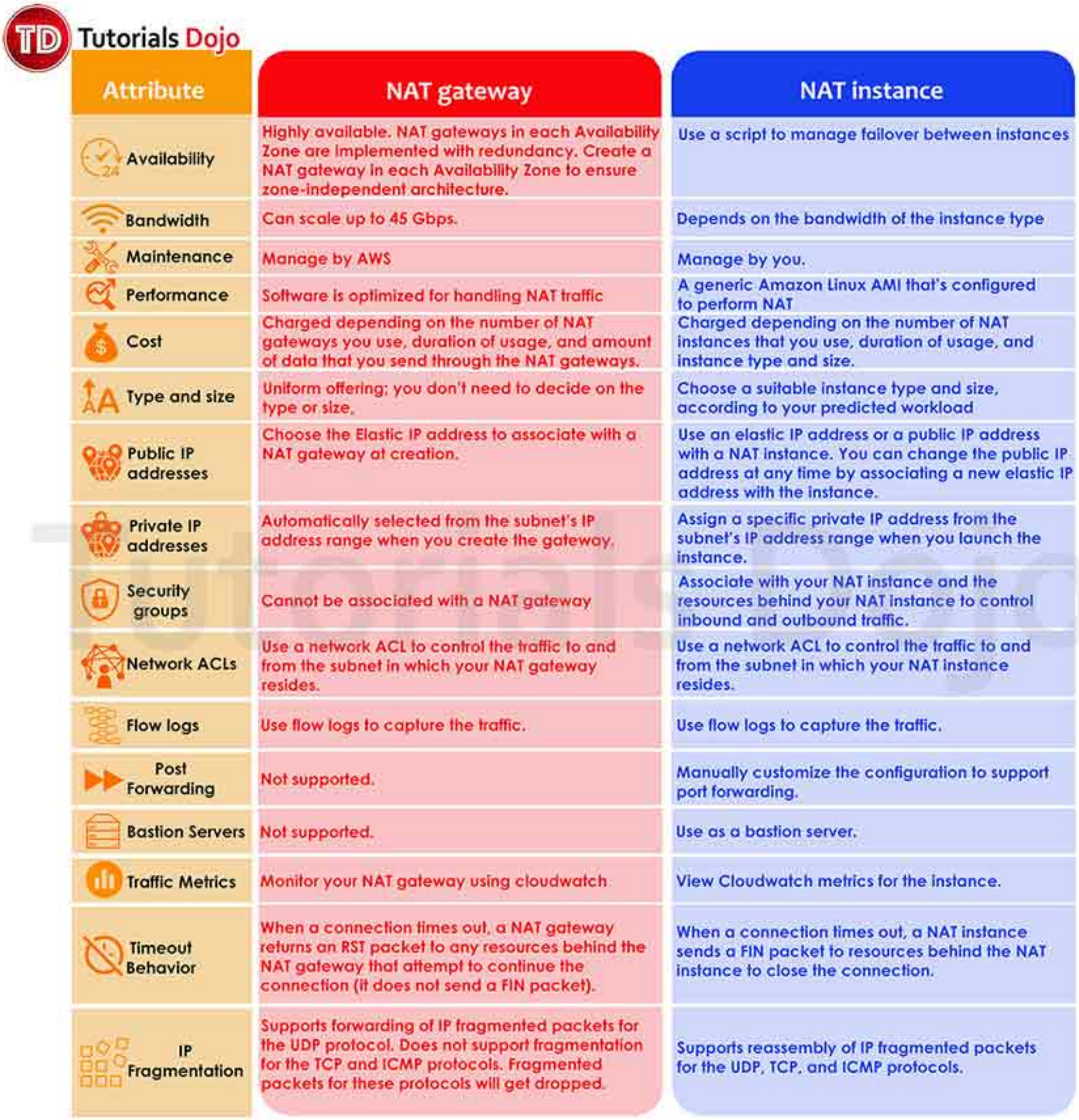
Egress only IG

- Primarily for IPv6 traffic

- Horizontally scaled, redundant, and highly available VPC component that provides a secure way for outbound-only Internet traffic from instances in a VPC to flow to the Internet
- Stateful, allows network connections initiated from instances in the VPC to automatically allow the corresponding return traffic to flow back into the VPC
- Enhances the security of instances in a VPC, as it blocks all incoming traffic from the Internet to instances in the VPC, and only allows outgoing traffic
- To use an Egress-Only Internet Gateway, you must create a VPC, configure a route table for your VPC, and associate the route table with the VPC. Then, you need to create an Egress-Only Internet Gateway and associate it with the VPC. You can also associate multiple subnets in a VPC with an Egress-Only Internet Gateway, which provides more flexibility and security

Nat Gateway vs Internet Gateway

- Internet Gateway (IGW) allows instances with public IPs to access the internet
- NAT Gateway (NGW) allows instances with no public IPs to access the internet
- NAT gateway replaces the source IP address of the instances with the IP address of the NAT gateway



Tutorials Dojo

Attribute	NAT gateway	NAT instance
Availability	Highly available. NAT gateways in each Availability Zone are implemented with redundancy. Create a NAT gateway in each Availability Zone to ensure zone-independent architecture.	Use a script to manage failover between instances
Bandwidth	Can scale up to 45 Gbps.	Depends on the bandwidth of the instance type
Maintenance	Manage by AWS	Manage by you.
Performance	Software is optimized for handling NAT traffic	A generic Amazon Linux AMI that's configured to perform NAT
Cost	Charged depending on the number of NAT gateways you use, duration of usage, and amount of data that you send through the NAT gateways.	Charged depending on the number of NAT instances that you use, duration of usage, and instance type and size.
Type and size	Uniform offering; you don't need to decide on the type or size.	Choose a suitable instance type and size, according to your predicted workload
Public IP addresses	Choose the Elastic IP address to associate with a NAT gateway at creation.	Use an elastic IP address or a public IP address with a NAT instance. You can change the public IP address at any time by associating a new elastic IP address with the instance.
Private IP addresses	Automatically selected from the subnet's IP address range when you create the gateway.	Assign a specific private IP address from the subnet's IP address range when you launch the instance.
Security groups	Cannot be associated with a NAT gateway	Associate with your NAT instance and the resources behind your NAT instance to control inbound and outbound traffic.
Network ACLs	Use a network ACL to control the traffic to and from the subnet in which your NAT gateway resides.	Use a network ACL to control the traffic to and from the subnet in which your NAT instance resides.
Flow logs	Use flow logs to capture the traffic.	Use flow logs to capture the traffic.
Post Forwarding	Not supported.	Manually customize the configuration to support port forwarding.
Bastion Servers	Not supported.	Use as a bastion server.
Traffic Metrics	Monitor your NAT gateway using cloudwatch	View Cloudwatch metrics for the instance.
Timeout Behavior	When a connection times out, a NAT gateway returns an RST packet to any resources behind the NAT gateway that attempt to continue the connection (it does not send a FIN packet).	When a connection times out, a NAT instance sends a FIN packet to resources behind the NAT instance to close the connection.
IP Fragmentation	Supports forwarding of IP fragmented packets for the UDP protocol. Does not support fragmentation for the TCP and ICMP protocols. Fragmented packets for these protocols will get dropped.	Supports reassembly of IP fragmented packets for the UDP, TCP, and ICMP protocols.

Network ACL

- Subnet level
- Stateless
- Explicit allow & deny
- Rules are numbered, and processing stops once a rule is matched Assigned to a subnet
- Can block traffic e.g UDP or IP

- Subnet level

Network ACL vs security group

- Security groups are used to control access at the EC2 instance level
- NACL controls access at the subnet level
- Neither can filter by URL, both filter via IP range

Network firewall

- Primarily used to manage multiple firewall rules across hundreds of Amazon VPCs and AWS Accounts that are usually under a single AWS Organization
- Create firewall rules that provide fine-grained control over network traffic and easily deploy firewall security across your VPCs
- Centrally manage security policies across existing accounts and VPC's and automatically enforce mandatory policies on new accounts

AWS Organizations

- Can generate Cost and Usage Reports (CUR) from the organisations management account which will allow member accounts to visualize the CUR in QuickSight
- When a new member account is invited the **OrganizationAccountAccessRole** must be created in the member account

Outposts

- Fully managed solutions delivering AWS infrastructure and services to virtually any on-premises or edge location for a truly consistent hybrid experience
- Allows you to extend and run native AWS services on premises

Pinpoint

- Offers marketers and developers one customizable tool to deliver customer communications across channels, segments, and campaigns at scale

PrivateLink

- A highly available, scalable technology that enables you to privately connect your VPC to supported AWS services, services hosted by other AWS accounts and supported AWS Marketplace partner services
- A service consumer creates a *VPC endpoint* to connect their VPC to an endpoint service. A service consumer must specify the service name of the endpoint service when creating a VPC endpoint. There are multiple types of VPC endpoints. You must create the type of VPC endpoint that's required by the endpoint service
- A PrivateLink VPC endpoint can be secured using a security group

Interface

- Create an *interface endpoint* to send traffic to endpoint services that use a Network Load Balancer to distribute traffic. Traffic destined for the endpoint service is resolved using DNS

Gateway Load Balancer

- Create a *Gateway Load Balancer endpoint* to send traffic to a fleet of virtual appliances using private IP addresses. You route traffic from your VPC to the Gateway Load Balancer endpoint using route tables. The Gateway Load Balancer distributes traffic to the virtual appliances and can scale with demand

Gateway endpoint

- Create a *gateway endpoint* to send traffic to Amazon S3 or DynamoDB using private IP addresses. You route traffic from your VPC to the gateway endpoint using route tables. Gateway endpoints do not enable AWS PrivateLink

RADIUS

- Remote Authentication Dial-In User Service
- A requirement to enable MFA for AWS services such as Amazon WorkSpaces and QuickSight

Redshift

- To copy snapshots for AWS KMS–encrypted clusters to another AWS Region, you need to create a **snapshot copy grant** for Redshift to use a KMS CMK in the destination AWS Region. Then choose that grant when you enable copying of snapshots in the source AWS Region
- When provisioned, an Amazon Redshift cluster is locked down by default, so nobody has access to it
- To grant other users inbound access to an Amazon Redshift cluster, you associate the cluster with a security group
- When audit logging is enabled, Redshift creates and uploads logs to Amazon S3 that capture data from the time audit logging is enabled. You can only use Amazon S3-managed keys (SSE-S3) encryption (AES-256) for audit logging

Rekognition

- A fully managed machine learning service that supports both real time streaming video events and stored video analysis

Replatforming vs rehosting

- Replatforming is a migration strategy where you don't change the core architecture but leverage some cloud optimizations
- Rehosting is a migration strategy where no cloud optimizations are done and the application is migrated as-is

Resource Access Manager

- AWS RAM helps you securely share your resources across AWS accounts, within your organization or organizational units (OUs), and with IAM roles and users for supported resource types. If you have multiple AWS accounts, you can create a resource once and use AWS RAM to make that resource usable by those other accounts. If your account is managed by AWS Organizations, you can share resources with all the other accounts in the organization or only those accounts contained by one or more specified OUs
- When you enable resource sharing within your organization, RAM creates a service-linked role called **AWSServiceRoleForResourceAccessManager**. This role can be assumed by only the AWS

RAM service, and grants AWS RAM permission to retrieve information about the organization it is a member of, by using the AWS managed policy AWSResourceAccessManagerServiceRolePolicy
- The API operations that principals in your account are allowed to perform vary depending on the resource type and are specified by the AWS RAM permission attached to the resource share
- Shareable resources
- Trusted access with AWS organisation is achieved by running the **enable-sharing-with-aws-organization command** in the **AWS RAM CLI**

Route53

- Designed to withstand DNS query floods
- To add failover to a latency record set the value of Evaluate Target Health to Yes on the latency alias resources for all regions
- An application cannot resolve record sets created in the private hosted zone of another AWS account. The solution to this problem is to associate the Route 53 private hosted zone in the management account with the VPC in the production account
- 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, these can be **conditionally** forwarded
- Non-authoritative queries are requests that the DNS does not have in its zone file
- Does not support static IP addresses as entry points
- To associate another account's VPC with a private hosted zone
    1. On Account A, create an authorization to associate its private hosted zone to the new VPC in Account B
    2. On Account B, associate the VPC to the private hosted zone in Account A. Delete the association authorization after the association is created

Hosted zone

- A hosted zone is a container for records, and records contain information about how you want to route traffic for a specific domain and its subdomains
- A hosted zone and the corresponding domain have the same name
- There are two types of hosted zones:
    1. ***Public hosted zones*** contain records that specify how you want to route traffic on the internet.
    2. ***Private hosted zones*** contain records that specify how you want to route traffic in an Amazon VPC

Private hosted zone

- To use private hosted zones, DNS hostnames and DNS resolution should be enabled for the VPC
- A *private hosted zone* is a container that holds information about how you want Amazon Route 53 to respond to DNS queries for a domain and its subdomains within one or more VPCs that you create with the Amazon VPC service

Health checks

- If using HTTPS health checks the endpoint must support TLS
- After a Route 53 health checker receives the HTTP status code, it must receive the response body from the endpoint within the next two seconds with the SearchString string that you specified. The string must appear entirely in the first 5,120 bytes of the response body or the endpoint fails the health check
- HTTPS health checks don't validate SSL/TLS certificates, so checks don't fail if a certificate is invalid or expired

- Route 53 must be able to establish a TCP connection with the endpoint within four seconds.
- The endpoint must respond with an HTTP status code of 2xx or 3xx within two seconds after connecting
- Route53 health checks result in a lower request count than use ALB target group health checks
- DNS records without a health check are always considered healthy. If no record is healthy, all records are deemed to be healthy
- If you're creating failover records in a private hosted zone, you must assign a public IP address to an instance in the VPC to check the health of an endpoint within a VPC by IP address

Records

- **Alias** records can be created for **root domains and subdomains**
- **CNAME** records are created only for **subdomains**
- For EC2 instances, always use a **Type A** Record without an **Alias**
- For ELB, CloudFront, and S3, always use a **Type A** Record with an **Alias**
- For RDS, always use the **CNAME** Record with **no Alias**

Routing policies

- **Simple routing policy** – Use for a single resource that performs a given function for your domain, for example, a web server that serves content for the example.com website
- **Failover routing policy** – Use when you want to configure active-passive failover
- **Geolocation routing policy** – Use when you want to route traffic based on the location of your users
- **Geoproximity routing policy** – Use when you want to route traffic based on the location of your resources and, optionally, shift traffic from resources in one location to resources in another
- **Latency routing policy** – Use when you have resources in multiple AWS Regions, and you want to route traffic to the region that provides the best latency
- **Multivalue answer routing policy** – Use when you want Route 53 to respond to DNS queries with up to eight healthy records selected at random
- **Weighted routing policy** – Use to route traffic to multiple resources in proportions that you specify

QuickSight

- To create a private connection to an RDS in a private subnet create a new private subnet in the same VPC as the Amazon RDS DB instance. Create a new security group with necessary inbound rules for QuickSight in the same VPC. Sign in to QuickSight as a QuickSight admin and create a new QuickSight VPC connection. Create a new dataset from the RDS DB instance

S3

- Cannot manage your own encryption key with SSE-S3
- KMS managed encryption keys allow the sysadmins to manage the keys e.g. rotation, revocation
- Cross region replication requires versioning to be enabled
- To detect public S3 objects, enable object-level logging for S3. Set up a EventBridge event pattern when a PutObject API call with public-read permission is detected in the AWS CloudTrail logs and set the target as an SNS topic for downstream notifications
- S3 supports both Gateway endpoints and Interface endpoints
- Can create maximum 100 buckets in account by default
- 5TB maximum object size
- **Sync** command synchronizes data between a file system and an S3 bucket, any future executions of the **sync** command will only upload the file delta

S3 access points

- S3 Access Points are unique hostnames that you can create to enforce distinct permissions and network controls for any request made through the Access Point
- By default have a specific setting to Block Public Access
- Unique to an account and Region
- Can have custom IAM permissions for a user or application
- Can have custom IAM permissions to specific objects in a bucket via a prefix to precisely control access
- Can be configured to accept requests only from a VPC to restrict Amazon S3 data access to a private network
- Can only be created from the AWS account that owns the S3 bucket

S3 Block Public Access

- Can be applied in any combination to individual access points, buckets, or entire AWS accounts
- Setting applied to an account apply to all buckets and access points that are owned by that account
- Settings applied to a bucket apply to all access points associated with that bucket

BlockPublicAcls:

- PUT Bucket acl and PUT Object acl calls fail if the specified ACL is public
- PUT Object calls fail if the request includes a public ACL
- If this setting is applied to an account, then PUT Bucket calls fail if the request includes a public ACL
- Existing policies and ACLs for buckets and objects are not modified

IgnorePublicAcls

- Setting to TRUE causes S3 to ignore all public ACLs on a bucket and any objects in that bucket
- Does not block PUT calls containing a public ACL but they have no effect
- Enabling this setting doesn't affect the persistence of any existing ACLs and doesn't prevent new public ACLs from being set

BlockPublicPolicy

- Setting to TRUE rejects PUT calls for buckets and access points if the policy allows public access
- Doesn't affect existing access point or bucket policies

RestrictPublicBuckets

- Setting to TRUE restricts access to an access point or bucket with a public policy to only AWS service principals and authorized users within the bucket owner's account and access point owner's account. Blocks all cross-account access to the access point or bucket (except by AWS service principals), while still allowing users within the account to manage the access point or bucket

S3 bucket policy

- To only allow S3 access via the VPC endpoint add an S3 bucket policy that restricts access to the VPC endpoint

Server-side encryption

- To ensure Server-Side Encryption with Customer-Provided Encryption Keys (SSE-C):
    1. For Amazon S3 REST API calls, use the following HTTP Request Headers:
        - `x-amz-server-side-encryption-customer-algorithm`
        - `x-amz-server-side-encryption-customer-key`

        - `x-amz-server-side-encryption-customer-key-MD5`
    2. For presigned URLs:
        - Specify the algorithm using the request header `x-amz-server-side-encryption-customer-algorithm`

Encryption

- To ensure in transit & at rest encryption
    1. Create a bucket policy that denies any unencrypted operations in the S3 bucket
    2. Turn on the S3 server-side encryption for the S3 bucket
    3. If using CloudFront: Configure redirection of HTTP requests to HTTPS requests in CloudFront.
- **aws:SecureTransport** enforces that the request is sent through HTTPS for encryption in transit
- If the bucket is encrypted with KMS, in addition to **Get** permissions the user/policy will also require IAM per missions to decrypt the referenced KMS key

Glacier

- Retrieval time of 3-5 hours

Glacier Flexible Retrieval

- Retrieval time is minutes to hours

Glacier Instant Retrieval

- Archive storage class that delivers the lowest-cost storage for long-lived data that is rarely accessed and requires retrieval in milliseconds
- Can save up to 68% on storage costs compared to using the S3 Standard-Infrequent Access

Glacier Select

- Cannot be used on compressed data

S3 Permissions

- If you used object ACLs for permissions management before you applied the bucket owner enforced setting and you didn't migrate these object ACL permissions to your bucket policy, after you re-enable ACLs, these permissions are restored
- You, as the bucket owner, still own any objects that were written to the bucket while the bucket owner enforced setting was applied. These objects are not owned by the object writer, even if you re-enable ACLs

S3 Replication Time Control

- Helps meet compliance or business requirements for data replication and provides visibility into Amazon S3 replication times
- Replicates most objects within seconds and 99.99 percent of those objects within 15 minutes
- Use Amazon S3 event notifications to track replication failure events

S3 Signed URLs

- Useful for distributing private content
- Can have expiration times
- Includes additional information such as an expiration date and time, that gives you more control over access to your content. This additional information appears in a policy statement, which is based on either a canned policy or a custom policy

S3 Transfer acceleration

- AWS calculates if the TA is faster than normal, if not the user is not charged for the transfer and TA may not even be used
- Must use the s3-accelerate endpoint for uploads

S3 Select

- S3 Select is an Amazon S3 feature that makes it easy to retrieve specific data from the contents of an object using simple SQL expressions without having to retrieve the entire object
- Can scan compressed files
- Can scan a subset of an object by specifying a range of bytes to query using the ScanRange parameter. This capability lets you parallelize scanning the whole object by splitting the work into separate Amazon S3 Select requests for a series of non-overlapping scan ranges

S3 website hosting

- Must have public read access
- When using S3 to host a static website, objects can't be encrypted by KMS and the account that owns the bucket must also own the object

S3 & LDAP

- To restrict S3 bucket access to a specific user's bucket:
    1. The application first authenticates against LDAP to retrieve the name of an IAM role associated with the user. It then assumes that role via a call to IAM Security Token Service (STS). Afterward, the application can now use the temporary credentials from the role to access the appropriate S3 bucket
    2. Authenticate against LDAP using an identity broker you created, and have it call IAM Security Token Service (STS) to retrieve IAM federated user credentials. The application then gets the IAM federated user credentials from the identity broker to access the appropriate S3 bucket

SAML

Schema Conversion Tool

- Convert IBM Db2 to Amazon Aurora etc

Security groups

- Instance level
- Stateful: only need rules to allow traffic in
- No explicit deny
- Implicit deny at end of rule set
- The default rules for the default security group allow inbound traffic from network interfaces that are assigned to the same security group
- The default rules for a security group that you create allow no inbound traffic

Security Hub

- Provides a comprehensive view of your security state in AWS and helps you assess your AWS environment against security industry standards and best practices
- Integrates with other AWS services such as GuardDuty, Inspector & Macie

Server Migration Service

- The SMS service may be used to implement the migrations of some servers, but it is not used for the planning phase

Service catalog

- **AWSServiceCatalogEndUserReadOnlyAccess** permissions & a launch constraint are sufficient for a user to launch products
- Launch constraint specifies IAM role that Service Catalog assumes when an end user launches, updates, or terminates a product
- An AWS Budgets can be associated with a product, which allows cost details to be viewed on the product and portfolio pages, Cloudformation - AWS::Budgets::Budget::resource -> NotificationsWithSubscribers

Session Manager

- Fully managed AWS Systems Manager capability. With Session Manager, you can manage your EC2 instances, edge devices, and on-premises servers and VMs.
- Can use either an interactive one-click browser-based shell or the AWS CLI.
- Provides secure and auditable node management without the need to open inbound ports, maintain bastion hosts, or manage SSH keys.
- To use Session Manager you must have the client installed, this is installed by default on instances launched form the Amazon Linux 2 AMI, and you must have permissions to communicate with AWS Systems Manager. The permissions should be assigned using an IAM role attached using an instance profile.
- **AmazonSSMManagedInstanceCore** - policy for Amazon EC2 Role to enable AWS Systems Manager service core functionality
- No need to open port 22 in the security group
- Supports Windows & Linux on EC2 & on-premise

SCP

- **Does not grant any access, only restricts the access that can be granted**
- Proactive: use an SCP to **prevent** something from occurring
- Default SCP allows all access
- Can ban EC2 instances from being created with public IPs
- SCPs do not affect the management account
- SCP actions affect all IAM identities including the root user of the member account
- Does not prevent a policy being added to an IAM principal but restricts the **effective permissions**

Shield

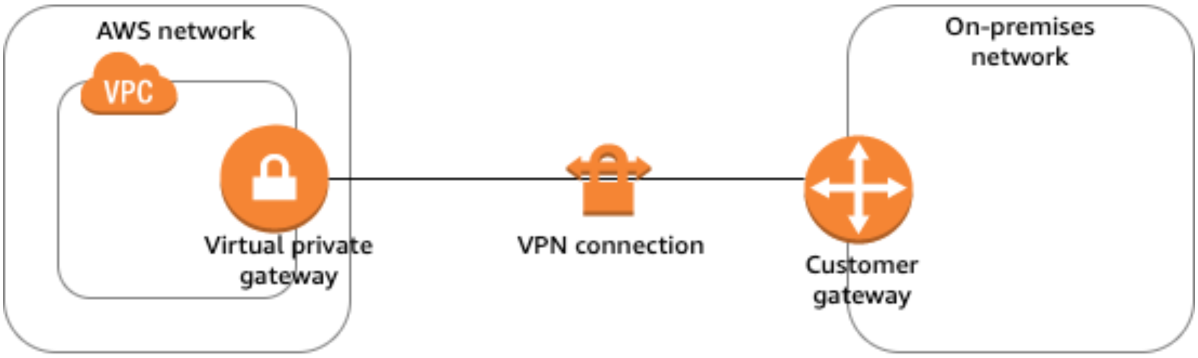
- Managed DDOS protection

Shield Advanced

- if you are subscribed to AWS Shield Advanced, you can register Elastic IPs (EIPs) as Protected Resources. DDoS attacks against EIPs that have been registered as Protected Resources are detected more quickly, which can result in a faster time to mitigate
- DDoS cost protection to safeguard against scaling charges resulting from DDoS-related usage spikes on protected EC2, ELB, CloudFront, Global Accelerator, and Route 53 resources

Site to site VPN

- Does not provide a consistent connection as uses the public internet, use Direct Connect instead
- Can be combined with an AWS Direct Connect public VIF to encrypt DX
- Connects to an on-premise customer gateway
- Customer gateway is a single point of failure, for high availability, deploy one in a second datacentre



Secure Token Service

- **AssumeRoleWithWebIdentity** is used only for Web Identity Federation (Facebook, Google, and other social logins)

Snow family

- *A 1Gbps connection at full utilization can transfer approximately 10 TB of data in a day*
- Data can be copied into an S3 bucket and later transitioned into AWS Glacier via a lifecycle policy. You can't directly copy data from Snowball devices into AWS Glacier

Snowball Edge

Storage optimized

- 100TB & 24 vCPUs

Compute optimized

- 52 vCPUs and optional GPU

Snowmobile

- Exabyte scale, up to 100PB
- Recommended for data transfers above 10PB

SNS Mobile Push Notifications

- Send push notification messages directly to apps on mobile devices

SQS

- Delay queues let you postpone the delivery of **all** new messages to consumers for up to 15 minutes
- Message timers allow you to set an initial invisibility period for a **single** message added to a queue up to a maximum of 15 minutes

- Visibility timeout is a period during which Amazon SQS prevents other consumers from receiving and processing a given message
- FIFO queues support up to 300 messages (API calls) per second (300 send, receive, or delete operations per second). When you batch 10 messages per operation (maximum), FIFO queues can support up to 3,000

Step functions

- Use SQS to start multiple workflows
- States
    1. Map - allows applying the same logic to each item in a collection and enables iterating over the collection and executing the necessary steps for each item individually
    2. Choice - designed for making decisions based on input or state conditions
    3. Parallel – branches in a Parallel state receive the same input
    4. Batch - execute a specified number of instances of a task in parallel, where each instance processes a different item from a collection

Storage Gateway

- **S3 file gateway** provides SMB or NFS based access to data in S3
- **Tape Gateway** provides virtual tapes to backup applications, data is stored in S3 or S3 Glacier
- **Volume Gateway** provides iSCSI block storage volumes to on-premises applications that you can store in S3 or migrate to EBS
- Hybrid storage solution
- Use Challenge-Handshake Authentication Protocol (CHAP) to authenticate iSCSI and initiator connections. CHAP provides protection against playback attacks by requiring authentication to access storage volume targets
- EBS snapshots can be created from and restored to storage gateway volumes

Volume gateway

- Expand block store infrastructure to the cloud
- Store your primary data locally, while asynchronously backing up that data to AWS
- Integrates with AWS backup to store volumes as EBS snapshots
- Provide your on-premises applications with low-latency access to their entire datasets
- **Cached mode:** primary data is stored in S3, frequently accessed data is stored in local cache
- **Stored mode:** primary data is stored locally, and your entire dataset is asynchronously backed up to S3

Tape gateway

- Replace physical tapes with virtual tapes backed by S3
- Integrates with existing tape-based backup infrastructure

S3 File gateway

- Hybrid file server infrastructure
- Offers a seamless way to connect to the cloud to store application data files and backup images as durable objects on Amazon S3 cloud storage
- Store and retrieve objects in Amazon S3 using industry-standard file protocols such as NFS and SMB
- Offers SMB or NFS-based access to data in Amazon S3 with local caching

Systems manager

- Amazon software that runs on EC2 instances, edge devices, on-premises servers, and VMs. SSM Agent makes it possible for Systems Manager to update, manage, and configure these resources
- AWS Systems Manager supports **AWS-RunPatchBaseline**, a SSM document for Patch Manager, a capability of AWS Systems Manager. This SSM document performs patching operations on managed nodes for both security related and other types of updates
- Use ***AWSSupport-ExecuteEC2Rescue*** automation document to recover impaired instances

Session Manager

- Session Manager is a fully managed AWS Systems Manager capability that lets you manage Amazon EC2 instances, on-premises instances, and VMs through an interactive one-click browser-based shell or through the AWS CLI

State manager

- Secure and scalable configuration management service

Patch manager

- Can patch EC2 and on-premise servers

Textract

- Automatically extracts printed text, handwriting, layout elements, and data from any document

Transit Gateway

- Allows fully transitive connections between VPCs in a Region i.e. VPC peering
- If there is only one VPC then a transit gateway will not be required
- To Peer VPCs with egress only internet access: attach each VPC to a shared transit gateway. Use an egress VPC with firewall appliances in two AZs and connect the transit gateway using IPSec VPNs with BGP

Trusted Advisor

- Inspects your AWS environment and makes recommendations for saving money, improving system performance, or closing security gaps. The Trusted Advisor notification feature helps you stay up to date with your AWS resource deployment
- Weekly email notifications

VMWare

- Application images are exported in Open Virtualization Format (OVF) format using the **VMware vSphere client** and imported using the CLI EC2 import command

VIF

- With these connections, you can create *virtual interfaces* directly to public AWS or to Amazon VPC, bypassing internet service providers in your network path

Virtual private gateway

- Only one VPG at a time can be connected to a VPC

VPC

- **enableDnsHostnames** indicates whether the instances launched in the VPC get public DNS hostnames. If this attribute is true, instances in the VPC get public DNS hostnames, but only if the **enableDnsSupport** attribute is also set to true
- **enableDnsSupport** indicates whether the DNS resolution is supported for the VPC. If this attribute is false, the Amazon-provided DNS server in the VPC that resolves public DNS hostnames to IP addresses is not enabled

Centralized VPC Endpoints (shared services VPC)

- A VPC endpoint allows you to privately connect your VPC to supported AWS services without requiring an Internet gateway, NAT device, VPN connection, or AWS Direct Connect connection
- Horizontally scaled, redundant, and highly available
- VPC endpoints enable you to reduce data transfer charges resulting from network communication between private VPC resources
- Without VPC endpoints configured, communications that originate from within a VPC destined for public AWS services must egress AWS to the public Internet to access AWS services. This network path incurs outbound data transfer charges

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
- Flow log data can be published to CloudWatch Logs, S3 or Data Firehose
- You can create a flow log for a VPC, a subnet, or a network interface
- If you create a flow log for a subnet or VPC, each network interface in that subnet or VPC is monitored
- Cannot be used to inspect packets
- Can help with several tasks, such as:
    1. Diagnosing overly restrictive security group rules
    2. Monitoring the traffic that is reaching your instance
    3. Determining the direction of the traffic to and from the network interfaces

Debugging flow logs

- If the network ACL permits the inbound and outbound traffic, the flow log displays two ACCEPT records (one for the originating ping and one for the response ping)
- If the security group denies inbound ICMP traffic, the flow log displays a single REJECT record, because the traffic was not permitted to reach your instance
- If one ACCEPT and one REJECT record is displayed, traffic is permitted entry but was denied egress

Forward web proxy

- Acts as an intermediary for requests from internal users and servers, often caching content to speed up subsequent requests
- Companies usually implement proxy solutions to provide URL and web content filtering, IDS/IPS, data loss prevention, monitoring, and advanced threat protection
- AWS customers often use a VPN or AWS Direct Connect connection to leverage existing corporate proxy server infrastructure, or build a forward proxy farm on AWS using software such as Squid proxy servers with internal Elastic Load Balancing

VPC IP Address Manager

- Amazon VPC IP Address Manager (IPAM) is a VPC feature that makes it easier for you to plan, track, and monitor IP addresses for your AWS workloads
- You can use IPAM automated workflows to more efficiently manage IP addresses
- You can use IPAM to do the following:
    1. Organize IP address space into routing and security domains
    2. Monitor IP address space that's in use and monitor resources that are using space against business rules
    3. View the history of IP address assignments in your organization
    4. Automatically allocate CIDRs to VPCs using specific business rules
    5. Troubleshoot network connectivity issues
    6. Enable cross-region and cross-account sharing of your Bring Your Own IP (BYOIP) addresses
    7. Provision Amazon-provided contiguous IPv6 CIDR blocks to pools for VPC creation

- IP address management (IPAM) is a core part of planning and managing the assignment and use of IP address space of a network used to manage available CIDR blocks
- Cannot be implemented natively, would require a custom Lambda

IPV6

- Public subnet - update the route tables to route IPv6 traffic (::/0) to an internet gateway
- Private subnets - update the route tables to route IPv6 traffic (::/0) to an egress-only internet gateway

Transit VPC

- Uses customer-managed EC2 VPN instances in a dedicated transit VPC with an Internet gateway

VPC peering

- 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. VPC peering does not facilitate centrally managed VPCs (use RAM)

VPC Sharing

- VPC sharing (part of Resource Access Manager) allows multiple AWS accounts to create their application resources such as EC2 instances, RDS databases, Redshift clusters, and Lambda functions, into shared and **centrally managed** VPCs

VPC endpoints

VPC endpoint service

- A VPC endpoint enables connections between a VPC and the supported services, without requiring that you use an internet gateway, NAT device, VPN connection, or AWS Direct Connect connection
- Can interface with an NLB but not an ALB
- **Are accessed via a private VIF**

VPC endpoint policy

- An endpoint policy is a resource-based policy that you attach to a VPC endpoint to control which AWS principals can use the endpoint to access an AWS service
- Resource based policy like an S3 bucket policy
- aws:sourceVpce condition is used to restrict access to specific VPCs

VPC Gateway Endpoint

- **S3 and DynamoDB only**
- Provide reliable connectivity without requiring an internet gateway or a NAT device for your VPC
- Does not use AWS PrivateLink, unlike other types of VPC endpoints
- Free
- A gateway that is a target for a route in your route table used for traffic destined to either S3 or DynamoDB
- Allows access from on-premise
- Does not allow access from another region
- Uses Amazon public IP addresses to access AWS

VPC Interface Endpoint

- Access AWS services using PrivateLink
- Elastic network interface with a private IP address from the IP address range of your subnet
- Serves as an entry point for traffic destined to a service that is owned by AWS or owned by an AWS customer or partner
- Does not allow access from on-premise
- Allows access from a VPC in another region using VPC peering or a transit gateway
- Billed
- Uses private IP addresses to access AWS
- Use **aws:sourceVpce** to permit access only from the VPC endpoint, for example using this in an S3 bucket policy would restrict access to the bucket to only requests coming from the VPC endpoint
- Security group on the interface endpoint must be configured to allow connectivity to the AWS services

VPN CloudHub

- AWS VPN CloudHub operates on a simple hub-and-spoke model that you can use with or without a VPC
- Use this approach if you have multiple branch offices and existing internet connections and would like to implement a convenient, potentially low-cost hub-and-spoke model for primary or backup connectivity between these remote offices

WAF

- WAF logs can be sent to CloudWatch Logs, S3 & Kinesis Data Firehose
- Cannot be directly configured in front of an ASG
- WAF web-acl with a **geo-match** rule attached to an ALB will block users based on country
- Can be associated with an API gateway, ALB, AppSync GraphQL API, Cognito user pool, App Runner, CloudFront
- Can filter based on IP addresses
- Rate based rules are evaluated over a 5-minute period

WorkDocs

- Fully managed platform for creating, sharing, and enriching digital content.