Topic Summaries

1,2, &3) IAM, S3, & AWS Organizations

4) EC2

5) Databases

6) Advanced IAM

7) Route53/DNS

8) VPC

9) HA Architecture

10) Applications

11) Security

12) Serverless

1, 2, & 3) IAM, S3, & AWS Organizations Summary

IAM Exam Tips

Identity Access Management Consists of the Following

-Users - End Users such as people, employees of an organization, etc

-Groups - A collection of users. Each user in the group will inherit the permissions of the group.

-Roles - You create roles and then assign them to AWS Resources

-Policies - Policies are made up of documents, called Policy documents. These documents are in a format called JSON and they give permissions as to what a User/Group/Role is able to do

-If you apply a policy to your group and you've got users within that group, they're going to inherit that policy automatically.

-Policies are written in JSON (Javascript Object Notation)

-IAM is universal. It does not apply to regions at this time.

-The "root account" is simply the account that's createdwhen you first set up your AWS account

and it has complete administrator access.

-New Users have **NO permissions** when first created. (Least Privilege is a common theme in AWS)

-New Users are assigned **Access Key ID & Secret Access Keys** when first created

-These are not the same as a password. You cannot use the Access key ID & Secret Access Key to Login in to the console. You can use this to access AWS via the APIs and Command Line, however.

-You only get to view your Access key ID & Secret Access Key once. If you lose them, you have to regenerate them. So, save them in a secure location.

-Always setup Multifactor Authentication on your root account.

-You can create and customise your own password rotation policies.

S3 Exam Tips

- Remember that S3 is Object-based: i.e. allows you to upload files.

-Files can be from 0 Bytes to 5 TB.

-There is unlimited storage

-Files are stored in Buckets (which is basically a folder in the cloud)

-**S3 is a universal namespace**. That is, names must be unique globally.

- example of an S3 URL https://s3-eu-west-1.amazonaws.com/acloudguru in this example eu-west-1 is the region and acloudguru is the bucket name.

-S3 is object based storage so it is **Not** suitable to install an operating system on or a database on

-Successful uploads will generate a **HTTP 200** status code.

-By default, all newly created buckets are PRIVATE. You can setup access control to your buckets using;

1) Bucket Policies

2) Access Control Lists

-S3 buckets can be configured to create access logs which log all requests made to the S3 bucket. This can be sent to another bucket and even another bucket in another account.

-The Key Fundamentals of S3 Are;

-Key (This is simply the name of the object)

-Value (This is simply the data and is made up of a sequence of bytes).

-Version ID (Important for versioning)

-Metadata (Data about data you are storing)

-Subresources;

1) Access Control Lists

2) Torrents

-Remember the consistency model for S3 going into your exam

-Read after Write consistency for PUTS of new Objects

-Eventual Consistency for overwrite PUTS and DELETES (can take some time to propagate)

(In other words, If you put an object into S3 you are going to be able to read it immediately... However, If you overwrite an object or delete an object and you read it instantly, you could potentially get the old version or you could get the new version)

S3 Storage Classes

★Very Important Exam Topic (Refer also to Most Important Charts & Diagrams Document)

1) S3 Standard

99.99% availability

99.999999999% durability (11 9's),

stored redudantly across multiple devices in multiple facilities, and is designed, and is designed to sustain the loss of 2 facilities concurrently.

2) S3 - IA

(Infrequently Accessed):

For data that is accessed less frequently, but requires rapid access when needed.

Lower fee than S3 Standard, but you are charged a retrieval fee.

3) S3 One Zone - IA

For where you want a lower-cost option for infrequently accessed data, but do not require the multiple Availability Zone data resilience.

4) S3 - Intelligent Tiering

Designed to optimize costs by automatically moving data to the most cost-effective access tier, without performance impact or operational overhead. Utilizes Machine Learning to do so.

5) S3 - Glacier

S3 Glacier is a secure, durable, and low-cost storage class for data archiving. Retrieval times configurable from minutes to hours.

6) S3 - Glacier Deep Archive

S3 Glacier Deep Archive is Amazon S3's lowest-cost storage class where a retrieval time of 12 hours is acceptable.

Understand how to get the best value out of S3

1) S3 Standard (Most Expensive)

2) S3 - IA

3) S3 - Intelligent Tiering (if you aren't accessing objects ... the objects will automatically be moved from S3 Standard to S3 - IA to save you some money)

4) S3 One Zone - IA (be advised if you lose that zone you lose that data. So it is for easily reproducible data)

5) S3 Glacier (for data archival)

6) S3 Glacier Deep Archive (Cheapest of all storage)

-Encryption in Transit is achieve by

> -SSL/TLS (if you have https in the url browser that means that traffic is going to be encrypted example https://www.google.com/ so someone in the middle won't be able to see what you are viewing on google. in terms of how this relates to S3, whenever you go to aws.amazon.com and you go into the S3 console so long as you're using HTTPS, all the traffic, all the files that you're uploading are going to be encrypted.

-Encryption at Rest (Server Side) is achieved by

-S3 Managed Keys- SSE -S3 (this is where S3 handles all of our encryption for us)

-AWS Key Management Service, Managed Keys - SSE - KMS

-Server Side Encryption with Customer Provided Keys - SSE - C (this is where you provide your keys and you manage the maintenance of those keys)

-Client Side Encryption

-This is where you encrypt the objects and then you upload them to S3

---

3 Different ways to share S3 buckets across accounts

-Using Bucket Policies & IAM (applies across the entire bucket). Programmatic Access Only.

-Using Bucket ACLs & IAM (indidvidual objects). Programmatic Access only.

-Cross-account IAM Roles. Programmatic AND Console access.

Some Best Practices with AWS Organizations;

-Always enable multi-factor authentication on root account.

-Always use a strong and complex password on root account.

-Paying account should be used for billing purposes only. Do not deploy resources into the paying account. Use the paying account to create OUs, SCPs, etc.

-Enable/Disable AWS services using Service Control Policies (SCP) either on OUs (organizational units such as the finance department or HR department) or on individual accounts

---

Advantages of Consolidated Billing

-One bill per AWS account

-Very easy to track charges and allocate costs

-Volume pricing discount

Cross Region Replication

-A way of replicating objects in S3 across regions (but you can also replicate them within the same region)

-In order for Cross Region Replication to work, Versioning must be enabled on both the source and destination buckets.

-Files in an existing bucket are **not** replicated automatically. So if you turn it on for an existing bucket those items that had already been in the bucket will not be automatically replicated to the destination bucket.

-However, all subsequent updated files will be replicated automatically.

-Delete markers are not replicated. So if you delete an object in one bucket it's not going to be deleted in the other.

-Deleting individual versions or delete markers will not be replicated.

-Going into your exam understand what Cross Region Replication is at a high level.

Lifecycle Policies

-Automates moving your objects between the different storage tiers.

-Can be used in conjunction with versioning.

-Can be applied to current versions and previous versions.

S3 Transfer Acceleration

(See Most Important Charts & Diagrams Document)

So, we have our users, they're all around the world. We have our edge locations. Our users will upload their files to the edge locations first and then those files will go over the AWS backbone network to S3.

So, if you do need to increase the performance of your, you know, of your users being able to upload files to S3, look at **S3 transfer acceleration**.

CloudFront

(See Most Important Charts & Diagrams Document)

-Edge Location -This is the location where content will be cached. This is separate to an AWS Region/AZ

-Origin - This is the origin of all the files that the CDN (Content Delivery Network) will distribute. This can be either a S3 Bucket, an EC2 instance, an Elastic Load Balancer, or Route 53.

-Distribution - This is the name given the CDN which consists of a collection of Edge Locations. There are two types of distributions:

-Web Distribution - Typically used for Websites

-RTMP - Used for Media Streaming

CloudFront (continued)

-Edge locations are not just READ only - you can write to them too. (ie put an object on to them) (We saw this when we looked at S3 Transfer Acceleration)

-Objects are cached for the life of the TTL (Time to Live) That value is always in seconds.

-You can clear cached objects by invalidating them, but you will be charged.

Snowball

-Snowball is basically a big disk that you can use to move your data in and out of the AWS cloud.

-With Snowball you can import large amounts of data into S3.

-With Snowball you can also export large amounts of data out of S3.

Storage Gateway

File Gateway

-File Gateway - For flat files, stored directly on S3. (That utilizes NFS)

Volume Gateway (utilizes iSCSI protocol)

-Stored Volumes - Entire Dataset is stored on site and is asynchronously backed up to S3.

-Cached Volumes - Entire Dataset is stored on S3 and the most frequently accessed data is cached on site.

Gateway Virtual Tape Library

-Used for backups and uses popular backup applications like NetBackup, Backup Exec, Veeam

etc. Great way to get rid of your physical tapes.

Athena Exam Tips

Remember what Athena is and what it allows you to do:

-Athena is an interactive query service

-It allows you to query data located in S3 using standard SQL

-Serverless

-Commonly used to analyse log data stored in S3 (but also works for any kinda data that you want to run SQL queries against you are going to be using Athena)

Macie Exam Tips

Remember what Macie is and what it allows you to do:

-Macie uses AI to analyze data in S3 and helps identify PII (Personally identifiable information such as social security number, address, credit card numbers, etc.)

-Can also be used to analyse CloudTrail logs for suspicious API activity

-Includes Dashboards, Reports and Alerting

-Great for PCI-DSS compliance and preventing ID theft

Recommendation Read the S3 FAQ before going into your exam

https://aws.amazon.com/s3/faqs/

S3 Lock Policies & Glacier Vault Lock Exam Tips [SAA-CO2]

-Use S3 Object Lock to store objects using a write once, read many (WORM) model.

-Object locks can be on individual objects or applied across the bucket as a whole

-Object locks come in two modes: governance mode and compliance mode

-Governance mode: users can't overwrite or delete an object version or alter its lock settings unless they have special permissions.

-Compliance mode: a protected object version can't be overwritten or deleted by any user, including the root user in your AWS account

-Use **S3 Glacier Vault Lock** to deploy and enforce compliance controls for individual s3 glacier vaults with a vault lock policy. Specify controls such as WORM in a Vault lock policy and lock the policy from future edits.

S3 Performance Exam Tips [SAA-CO2]

-The more prefixes you use the better performance you are going to get out of S3. Example: mybucketname/folder1/subfolder1/myfile.jpg > /folder1/subfolder1

-You can achieve a high number of requests: 3,500 PUT/Copy/Post/Delete and 5,500 GET/HEAD requests per second per prefix

-You can get better performance by spreading your reads across different prefixes. For example, if you are using two prefixes, you can achieve 11,000 requests per second.

-When using SSE-KMS to encrypt your objects in S3, you must keep in mind the KMS limits.

-Uploading/downloading will count toward the KMS quota

-Region-specific, however, it's either 5,500, 10,000, or 30,000 requests per second.

-You currently cannot request a quota increase for KMS.

-Use multipart uploads to increase performance when uploading files to S3.

-Should be used for any files over 100 MB and must be used for any file over 5 GB.

-Use **S3 byte-range fetches** to increase performance when downloading files to S3

S3 Select & Glacier Select Exam Tips [SAA-C02]

-S3 Select is used to retrieve only a subset of data from an object by using simple SQL expressions.

-Get data by rows or columns using simple SQL expressions.

-Save money on **data transfer** and increase speed.

AWS DataSync Exam Tips [SAA-C02]

-Used to move large amounts of data from on-premises to AWS.

-Used with NFS- and SMB- compatible file systems.

-Replication can be done hourly, daily, or weekly.

-To set it up…Install the DataSync agent to start the replication (usually done on-premise).

-Also, can be used to replicate EFS to EFS. (Install the DataSync on an EC2 instance that is connected to EFS and you can use that DataSync to replicate your EFS to another copy or another EFS in the cloud).

CloudFront Signed URLs and Cookies Exam Tips [SAA-C02]

-Use signed URLs/cookies when you want to secure content so that only the people you authorize are able to access it.

-A signed URL is for individual files. 1 file = 1 URL

-A signed cookie is for access to multiple files. 1 cookie = multiple files (an example would be a membership with A Cloud Guru which gives you access to many video files)

-If your origin is EC2 then you want to use CloudFront. If your origin is S3 and you have only got a single file in there then you are going to want to use a s3 signed URL instead of a cloudfront signed url.

---

4) EC2 Summary

Amazon Elastic Compute Cloud (Amazon EC2) is a web service that provides resizable compute capacity in the cloud. Amazon EC2 reduces the time required to obtain and boot new server instances to minutes, allowing you to quickly scale capacity, both up and down, as your computing requirements change.

There are 4 different pricing models for EC2

1) On Demand

Allows you to pay a fixed rate by the hour (or by the second) with no commitment.

2) Reserved

Provides you with a capacity reservation, and offer a significant discount on the hourly charge for an instance. Contract Terms are **1 Year or 3 Year Terms**. The more you pay up front and the longer the contract term the more discount you get.

3) Spot Instances & Spot Fleet [SAA-C02] (Pricing model has changed from before)

-Spot Instances save up to 90% of the cost of On-Demand Instances.

-Useful for any type of computing where you **don't need persistent storage** such as ephemeral computing.

-You can block Spot Instances from terminating by using Spot block.

-A Spot Fleet is a collection of Sport instances and optionally, On-Demand instances

4) Dedicated Hosts

Physical EC2 server dedicated for your use. Dedicated Hosts can help you reduce costs by allowing you to use your existing server-bound software licenses. (So it is useful where you've got existing server-bound software licenses or perhaps regulations saying that you cannot use multi-tenant virtualization)

EC2 Hibernate Exam Tips [SAA-C02]

-EC2 Hibernate preserves the in-memory RAM on persistent storage (EBS)

-Much faster to boot up because you do not need to reload the operating system

-Instance RAM must be less than 150 GB

-Instance families can include C3, C4, C5, M3, M4, M5, R3, R4 and R5

-Available for Windows, Amazon Linux 2 AMI, and Ubuntu

-Instances can't be hibernated for more than 60 days

-Available for On-Demand instances and Reserved Instances

Mneumonic for EC2 instance types (You do not need to know these for the associate exam but they required for the Professional level exam)

FIGHT DR MCPXZAU is the mnemonic

F- For FPGA

I - For IOPS

G- Graphics

H- High Disk Throughput

T- Cheap general purpose (think T2 Micro)

D- For Density

R- For Ram

M- Main choice for general purpose apps

C- For Compute

P- Graphics (think Pics)

X- Extreme Memory

Z- Extreme Memory AND CPU

A- Arm-based workloads

U- Bare Metal

EBS

-EBS is basically a virtual hard disk drive in the cloud

-Termination Protection is **turned off** by default, you must turn it on. (This will protect your EC2 instances from being accidentally deteled by your developers or system administrators)

-On an EBS-backed instance, the **default action is for the root EBS volume to be deleted** when the instance is terminated. (If you have added additional attached volumes to the EC2 instance, those additional volumes **won't** be deleted automatically when you delete your instance)

-EBS Root Volumes of your Default AMI's **CAN** be encrypted. You can also use a third party tool (such as bit locker etc) to encrypt the root volume, or this can be done when creating AMI's (remember the lab) in the AWS console or using the API.

-Additional volumes can be encrypted

Security Groups

-All Inbound traffic is blocked by default.

-All Outbound traffic is allowed.

-Changes to Security Groups take effect **immediately.**

-You can have any number of EC2 instances within a security group.

-You can have multiple security groups attached to EC2 instances. (In the EFS lab we attached our default security group as well as our Web DMZ security group)

-Security Groups are STATEFUL. (That means when you open up a port, such as port 80, it will be open for both inbound and outbound traffic) (NACLs on the other hands, are STATELESS. In that case you would have to open up Port 80 for both inbound and also do it for the outbound)

-If you create an inbound rule allowing traffic in, that traffic is automatically allowed back out again.

-You **cannot** block specific IP addresses using Security Groups, instead use Network Access Control Lists to do that.

-You can only specify **allow** rules, but not deny rules.

Compare EBS Types

★Very Important Exam Topic (Refer also to Most Important Charts & Diagrams Document)

Two different Solid-State Drives (SSD) options

-General purpose SSD (gp2 is the API name)

-Provisioned IOPS SSD (io1 is the API name) (if you want your IOPS to go above 16,000,

you want to move from general purpose over to provision IOPS)

Three different hard disk drive (HDD) types.

-Throughput Optimized Hard Disk Drives (st1 is the API name) (if you need to optimize throughput)

-Cold Hard Disk Drives (sc1 is the API name) (If you just want the lowest cost storage available,use cold hard disk drive)

-EBS magnetic (aka standard) (considered previous generation)

EBS snapshots

-Volumes exist on EBS. Think of EBS as a virtual hard disk drive in the cloud.

-Snapshots exist on S3. Think of snapshots as a photograph of the disk.

-Snapshots are point in time copies of Volumes.

-Snapshots are incremental -- this means that only the blocks that have changed since your last snapshot

are moved to S3.

-If this is your first snapshot, it may take some time to create. If you take a second snapshot it is only going to replicate the delta (or the changes so to speak)

-To create a snapshot for Amazon EBS volumes that serve as root device volumes, you should stop the instance before taking the snapshot. (That will give you a consistent snapshot)

-However you can take a snapshot while the instance is running.

-You can create AMI's from both Volumes and Snapshots

-You can change EBS volume sizes on the fly, including changing the size and storage type.

-Volumes will ALWAYS be in the same availability zone as the EC2 instance.

Migrating EBS

-To move an EC2 volume from one AZ to another, take a snapshot of it, create an AMI from the snapshot and then use the AMI to launch the EC2 instance in a new AZ.

-To move an EC2 volume from one region to another, take a snapshot of it, create an AMI from the snapshot and then copy the AMI from one region to the other. Then use the copied AMI to launch the new EC2 instance in the new region.

EBS Encryption

-Snapshots of encrypted volumes are encrypted automatically.

-Volumes restored from encrypted snapshots are encrypted automatically

-You can share snapshots, but only if they are unencrypted.

-These snapshots can be shared with other AWS accounts or made public.

Unencrypted Root Device Volumes

Root Device Volumes can now be encrypted when you provision your EC2 instance. However, if you do have an unencrypted root device volume that needs to be encrypted you need to do that following...

-Create a Snapshot of the unencrypted root device volumes

-Create a copy of the snapshot and select the encrypt option

-Create an AMI from the encrypted snapshot

-Use that AMI to launch new encrypted instances

EBS vs Instance Store

-Instance Store Volumes are sometimes called Ephemeral Storage.

-Instance store volumes cannot be stopped. If the underlying host (or hypervisor) fails, you will lose your data.

-EBS backed instances can be stopped. You will not lose the data on this instance if it is stopped.

-You can reboot both, you will not lose your data.

-By default, both ROOT volumes will be deleted on termination. However, with EBS volumes, you can tell AWS to keep the root device volume from being deleted.

Encrypting Root Device Volumes

-Create a Snapshot of the unencrypted root device volume

-Create a copy of the Snapshot and select the encrypt option

-Create an AMI from the encrypted Snapshot

-Use that AMI to launch new encrypted instances

(that's how we encrypt our root device volumes using the aws console but you can do it using software like bitlocker, etc.)

Networking with EC2

-There are three different types

In the exam you will be given different scenarios and you will be asked to choose whether you should use an ENI, EN, or EFA.

-ENI (Elastic Network Interface)(basically a virtual network card)

For basic networking. Perhaps you need a separate management network to your production network or a separate logging network and you need to do this at low cost. In this scenario use multiple ENIs for each network.

-Enhanced Network

For when you need speeds between 10Gbps and 100Gbps. Anywhere you need reliable, high throughput. (this will typically be an Enhance Network Adaptor aka ENA.) (Uses single root I/O virtualization (SR-IOV) to provide high-performance networking capabilities on supported instance types.)

-Elastic Fabric Adaptor (EFA)

For when you need to accelerate High Performance Computing (HPC) and machine learning applications or if you need to do an OS by-pass. If you see a scenario question mentioning HPC ,ML (machine learning), or OS by-pass and asking what network adaptor you want, choose EFA (Elastic Fabric Adaptor).

CloudWatch

-CloudWatch is used for monitoring performance.

-CloudWatch can monitor most of AWS as well as your applications that run on AWS.

-CloudWatch with EC2 will monitor events every 5 minutes by default.

-You can have 1 minute intervals by turning on detailed monitoring.

-You can create CloudWatch alarms which trigger notifications.

-CloudWatch is all about performance. CloudTrail is all about auditing.

What can you do with CloudWatch?

-Create **Dashboards** to see what is happening with your AWS environment

-Create **Alarms.** You can set Alarms that notify you when particular thresholds are hit. (such as 80% or 90% CPU utilization)

-Create **Events**. CloudWatch Events helps you to respond to state changes in your AWS resources.

-Create **Logs.** CloudWatch Logs help you to aggregate, monitor, and store logs.

CloudTrail vs CloudWatch

-CloudWatch monitors performance

-CloudTrail monitors API calls in the AWS platform. (So for example CloudTrail will tell you WHO

provisioned an EC2 instance, or WHO set up an S3 bucket, etc.)

The CLI (Command-line Interface)

-You can interact with AWS from anywhere in the world just by using the command line (CLI)

-You will need to set up access in IAM. (This gives you an access key id and secret access key)

-Commands themselves are not in the exam, but some basic commands are always good to know.

Roles

Roles are a more secure than storing your access key and secret access key on individual EC2 instances. (because on our EC2 instance, what happens if somebody goes into our secret directory, our .AWS directory, which is in our home directory, and then opens up our configuration details so we're able to get our access key ID and secret access key.)

-Roles are easier to manage

-Roles can be assigned to an EC2 instance after it is created using both the console & command line.

-Roles are UNIVERSAL -- you can use them in any region

BootStrap Scripts

-Bootstrap scripts run when an EC2 instance first boots.

-Can be a powerful way of automating software installs and updates. (For example, per the ACG lab, a bootstrap script that installs wordpress)

Instance Meta Data & User Data

-Used to get information about an instance (such as public ip)

-curl http://169.254.169.254/latest/meta-data/ (meta-data could be the instance type, current ip address, etc)

-curl http://169.254.169.254/lates/user-data/ (this is the bootstrap script that is passed to your instance when it first boots up)

EFS (Elastic File System)

-Supports the Network File System version 4 (NFSv4) protocol

-You only pay for the storage you use (no pre-provisioning required.)

-Can scale up to the petabytes

-Can support thousands of concurrent NFS connections

(Remember, if you need shared storage, you can't share ABS with multiple EC2 instances, doesn't work like that, but you can create an EFS mount and then you can store your files in there and multiple EC2 instances will be able to access it.)

-Data is stored across multiple AZ's within a region

-Read After Write Consistency

EFS, FSx for Windows, & FSx for Lustre

-The exam will ask what is the best storage mechanism for the scenario given.

-EFS

When you need distributed, highly resilient storage for Linux instances and Linux-based applications.

-Amazon FSx for Windows

When you need centralised storage for Windows-based applications such as Sharepoint, Microsoft SQL Server, Workspaces, IIS Web Server or any other native Microsoft Application. For SMB storage you are going to want Amazon FSx for Windows.

-Amazon FSx for Lustre

When you need high-speed, high-capacity distributed storage. This will be for applications that do High Performance Compute (HPC), financial modelling etc. Remember that FSx for Lustre can store data directly on S3.

EC2 Placement Groups

★Very Important Exam Topic (Refer also to Most Important Charts & Diagrams Document)

-Three Types of Placement Groups;

1) Clustered Placement Group ... for low network latency / high network throughput. And all of

your ec2 instances will be in the same AZ, so that they are as close together as possible.

2) Spread Placement Group ... for individual critical EC2 instances. You want to make sure they are in different AZ's and on different pieces of hardware so if a rack fails it is only going to affect the one EC2 instance, and won't take two or three out at a time.

3) Partitioned ... Multiple EC2 instances HDFS, HBase, and Cassandra. This is where you have multiple EC2 instances in a partition and each partition is always going to be on separate hardware from the other partitions.

-A clustered placement group **can't** span multiple Availability Zones.

-A spread placement and partitioned group can span Availability Zones.

-The name you specify for a placement group must be unique within your AWS account.

-Only certain types of instances can be launched in a placement group (Compute Optimized, GPU, Memory Optimized, Storage Optimized)

-AWS recommend homogenous (instances of the same type) instances within clustered placement groups.

-You can't merge placement groups.

-You can't move an existing instance into a placement group. You can create an AMI from your existing instance, and then launch a new instance from the AMI into a placement group.

HPC (high-performance computing) on AWS Exam Tips [SAA-C02]

-We can achieve HPC on AWS through **data transfer, compute and networking, storage, and orchestration & automation services**.

-1) Data Transfer if you are going to move large amounts of data into AWS you are going to be using snowball or snowmobile (terabytes/petabytes worth of data)

-Can use AWS DataSync to store our files on S3, EFS, FSx for Windows, etc.

-We can use Direct Connect (a dedicated line into the AWS data center)

-2) Compute & Networking We can use EC2 instances that are GPU or CPU optimized. We can use EC2 fleets (Spot Instances or Spot Fleets). We can use placement groups (particularly cluster placement groups within the same AZ for low latency). We can use enhanced networking single root I/O virtualization (SR-IOV). We can use Elastic Network Adapters or (Intel 82599 Virtual Function (VF) interface (which is legacy)). We can use our Elastic Fabric Adapters which have OS bypass.

-3) Storage

-instance-attached storage:

-EBS which scales up to 64,000 IOPS with Provisioned IOPS (PIOPS)

-Instance Store: Scale to millions of IOPS; low latency

-Network storage:

-Amazon S3: Distributed object-based storage; not a file system.

-Amazon EFS: Scales IOPS based on total size, or used Provisioned IOPS

-Amazon FSx for Luster: HPC-optimized distributed file system; millions of IOPs, which is backed by S3.

-4) Orchestration & Automation Services:

-**AWS Batch** enables developes to easily and efficiently run hundreds of thousands of batch computing jobs. Supports multi-node parallel jobs, which allows you to run a single job that spans multiple EC2 instances. Easily schedule jobs and launch EC2 instances for your needs.

-**AWS ParallelCluster** open-source cluster management tool that makes it easy for you to deploy and manage HPC clusters on AWS. Uses a simple text file to model and provision all the resources needed for your HPC apps. Automate creation of VPC, subnets, cluster type, and instance types.

Web Application Firewalls (WAFs)

-How do you block malicious IP addresses? You can use AWS WAF or you can Use Network ACLs

-Use AWS WAF (can also block specific countries, you can look for query string parameters, you can use WAF to block cross-site scripting, and SQL injections as well.)

-Use Network ACLs

_____

_____

5) Databases Summary

-RDS (OLTP) (is used for Online Transaction Processing)

RDS has six choices

-SQL

-MySQL

-PostgreSQL

-Oracle

-Aurora

-MariaDB

-DynamoDB (No SQL) (this is Amazon's NO SQL database service...and it is Serverless)

-Red Shift OLAP (Amazon's data warehousing solution used for Online Analytics Processing)

Elasticache

-Memcached (simple)

-Redis (used for more advanced data types and if we need multiple AZ's or if we need to be able to do backups

---

Relational Databases

Remember the following points...

-RDS runs on virtual machines

-You cannot log in to these operating systems however. so you **can't** RDP using Windows,

and you can't SSH to your Linux operating systems. (very important to know for the exam)

-Patching of the RDS Operating System and DB is Amazon's responsibility (because you do NOT get operating system level system access)

-RDS is NOT Serverless

Aurora Serverless IS Serverless (So most everything in RDS is NOT serverless with the exception of Aurora Serverless)

-There are two different types of Backups for RDS:

-Automated Backups

-Database Snapshots

Read Replicas

★Very Important Exam Topic (Refer also to Most Important Charts & Diagrams Document)

-Can be Multi-AZ

-Used to increase performance

-Must have backups turned on (in order to enable Read Replicas)

-Can be in different regions (as well as the same region)

-Can be MySQL, PostgreSQL, MariaDB, Oracle, Aurora (According to ACG, SQL server is the only one that does not support Read Replicas at this time. However, the following link says otherwise https://aws.amazon.com/rds/features/read-replicas/)

-Can be promoted to master, this will break the Read Replica

-Common exam scenario, struggling database with heavy read traffic what do you do? One way to increase performance is to add read replicas and to point your EC2 instances to those read replicas.

MultiAZ (Multi Availability Zone)

★Very Important Exam Topic (Refer also to Most Important Charts & Diagrams Document)

-Used for DR (Disaster Recovery) only.... It is NOT used for performance

-You can force a failover from one AZ to another by rebooting the RDS instance.

Encryption

-Encryption at rest is supported for MySQL, Oracle, SQL Server, PostgreSQL, MariaDB & Aurora. Encryption is done using the AWS Key Management (KMS) service. Once your RDS instance is encrypted, the data stored at rest in the underlying storage is encrypted, as are its automated backups, read replicas, and snapshots.

DynamoDB

-DynamoDB is Serverless

-Stored on SSD storage

-Spread across 3 geographically distinct data centres

-Eventual Consistent Reads (Default) (if your application doesn't need to read updated data within one

second of the update then you want eventual consistency)

-Strongly Consistent Reads (read the updated data within one second of the update)

Advanced DynamoDB Exam Tips [SAA-C02]

-DynamoDB Accelerator (DAX) is a fully managed, highly available, in-memory cache for DynamoDB which increases performance by 10 times. Reduces request time to microseconds. Both read and write performance improvement.

-On-Demand Capacity offers pay-per-request pricing. But you pay more per request than with provisioned capacity.

-On-Demand Backup and Restore: consistent within seconds and retained until deleted

-Point-in-Time Recovery (PITR): restore to any point in the last 35 days. Latest restorable time stamp is typically five minutes in the past.

-Streams are time-ordered sequence of item-level changes in a table. Like a FIFO queue of your data. Your data is stored for 24 hours. Stream records are organized into groups called shards. A shard is a container for multiple stream records.

-Global Tables is a managed multi-master, multi-region replication solutions for DynamoDB. Replication latency across regions is under one second

-Database Migration Service (DMS)

Migrate your source database to a target database (For instance, from RDS to DynamoDB). The source database remains operational during the migration. DynamoDB is not a supported source database but is a supported target.

-Security: DynamoDB is fully encrypted at rest using KMS

Redshift

-Redshift is used for business intelligence

-Available in only 1 AZ

Redshift Backups

-Enabled by default with a 1 day retention period.

-Maximum retention period is 35 days.

-Redshift always attempts to maintain at least three copies of your data (the original and replica on the compute nodes and a backup in Amazon S3).

-Redshift can also asynchronously replicate your snapshots to S3 in another region for disaster recovery.

Aurora

-2 copies of your data are contained in each availability zone, with minimum of 3 availability zones. 6 copies of your data.

-You can share Aurora Snapshots with other AWS accounts.

-3 types of replicas available. Aurora Replicas, MySQL replicas & PostgresQL replicas. Automated failover is only available with Aurora Replicas.

-Aurora has automated backups turned on by default. You can also take snapshots with Aurora. You can share these snapshots with other AWS accounts.

-Use Aurora Serverless if you want a simple, cost-effective option for infrequent, intermittent, or unpredictable workloads.

Elasticache

-Use Elasticache to increase database and web application performance (by caching).

-Redis is Multi-AZ

-You can do back ups and restores of Redis

-If you need to scale horizontally, use Memcached.

Database Migration Service (DMS) Exam Tips [SAA-C02]

-DMS allows you to migrate databases from one source to AWS.

-The source can either be on-premises, or inside AWS itself or another cloud provider such as Azure.

-You can do homogenous migrations (same DB engines) or heterogeneous migrations.

-If you do a heterogenous migration, you will ned the AWS Schema Conversion Tool (SCT)

Caching Strategies on AWS Exam Tips [SAA-C02]

Caching is a balancing act between up-to-date, accurate information and latency. We can use the following services to cache on AWS: CloudFront, API Gateway, ElastiCache – Memcached and Redis, DynamoDB Accelerator (DAX)

EMR Overview (Elastic Map Reduce Exam Tips [SAA-C02]

-EMR is used for big data processing

-Consists of a master node, a core node, and (optionally) a task node.

-By default, log data is stored on the master node.

-You can configure replication to S3 on five-minute intervals for all log data from the master node; however, this can only be configured when creating the cluster for the first time.

---

6) Advanced IAM Summary [SAA-C02]

Directory Service

-Know that Active Directory is a directory service developed by Microsoft

-Directory Service Connects AWS resources with on-premises AD

-Common use case: when you want to use SSO (single sign-on) to log into any domain-joined EC2 instance

-AWS Managed Microsoft AD = real active directory domain controllers running windows server running inside AWS

-AD trust = you can extend your existing Active Directory to on premises AD using AD trust. Use AD trust to extend existing Active Directory inside AWS to your on premises environment.

-AWS vs. customer responsibilities with regards to these services (patching, scale-out, user and group management, etc)

-Simple AD = similar to managed Microsoft AD but it does not support trusts, so you cannot join simple AD to your on premises AD. If you want to do that you will need to use AD connector.

-AD connector = a directory gateway/proxy for your on premises AD

-Cloud Directory = a service for developers looking to work with hierarchical data. It has nothing to do with Microsoft AD.

-Cognito user pools = a managed user directory that works with social media identities. Has nothing to do with Microsoft AD

-Know which services in Directory Service are AD vs. Non-AD Compatible

<u>IAM Policies</u>

-ARN Amazon Resource Name. Example: **arn:aws:s3:::bucketname**

-IAM policy structure (JSON document composed of a number of statements)

-Statements contain an Effect (such as allow or deny) an Action (API call) and a Resource (entity in AWS that the policy will affect such as an S3 bucket, DynamoDB table, etc)

-Identity vs. resource policies: Identity policies are attached to an IAM user, group, or role. Resource policies are attached to a resource.

-Policy evaluation logic (If it's not explicitly allowed then it's implicitly denied. Explicit deny overrides everything else. If two policies in effect for the same resource and one has an allow and one has a deny, the deny rule supersedes the allow.)

-AWS managed vs. customer managed (AWS managed policies can't be edited by you. Customer Managed policies are created and edited by you. You can create as many customer Managed policies as you like.)

-Permission boundaries don't allow or deny permissions on their own. They define the maximum permissions an identity can have.

<u>Resource Access Manager (RAM)</u>

-Allows resource sharing between accounts

-Works with Individual accounts and AWS Organizations accounts as well

-Types of resources you can share (app mesh, aurora, codebuild, ec2, ec2 image builder, license manager, resource groups, route53)

Single Sign-on (SSO)

-Helps us centrally manage access to AWS accounts in business applications

-Scenario: Approriate when using existing corporate identities to sign into AWS services or third party applications (Gsuite, Office 365, Salesforce accounts)

-Using existing identities to login to other contexts

-Can govern account-level permissions

-SAML2.0

---

7) DNS Summary

Route53

-ELBs do NOT have pre-defined IPv4 addresses; you resolve to them using a DNS name

-Understand the difference between an Alias Record and a CNAME. (Imagine the YELLOW PAGES and an A record is just a person's name and then their telephone number. A CNAME is something like Batman and instead of the telephone number it says SEE Bruce Wayne.)

-Exam scenario, You need to map your naked domain name or your zone apex record to an S3 bucket. What should you be using a CNAME or an alias record? You would want to use an Alias record in this case

---

Common DNS Types

- SOA Records (start of authority)

-NS Records

-A Records

-CNAMES

-MX Records

-PTR Records

The Following Routing Policies are Available with Route53:

-Simple Routing

-Weighted Routing

-Latency-based Routing

-Failover Routing

-Geolocation Routing

-Geoproximity Routing (Traffic Flow Only)

-Multivalue Answer Routing

Health Checks

-You can set health checks on individual records sets.

-If a record set fails a health check it will be removed from Route53 until it passes the health check.

-You can setup SNS notifications to alert you if a health check has failed

---

Route 53 Routing Policies

★Very Important Exam Topic (Refer also to Most Important Charts & Diagrams Document)

Simple Routing

If you choose the simple routing policy you can only have one record with multiple IP addresses. And you cannot have any Health Checks. If you specify multiple values in a record, Route53 returns all values to the user in a random order.

Weighted Routing

As an example, we've got our user, they're typing in a DNS address, and we have set weights, so that we are sending 20% of the traffic to US-EAST-1, and 80% of the traffic to US-WEST-1. So that's how weighted-routing traffic works.

Latency Routing

As an example, Latency-based routing is based on our user's location and the latency. So let's say a

South African user is going to visit our website which is behind Route 53. And let's say they are going to get there in 54 milliseconds if they are directed to US-WEST-2, and 300 milliseconds if they are directed to AP-SOUTHEAST-2. So, in this case Route 53 is going to send them to the fastest fleet of EC2 instances first, which is going to be EU-WEST-2.

Failover Routing

With failover routing, we use health checks. So essentially we've got an active/passive environment. Our active environment could be US-WEST-2, and our passive environment could be AP-SOUTHEAST-2. If for some reason our EC2 instances or a region goes down, it's going to detect this using a health check and it's going to failover to our passive environment.

Geolocation Routing Policy

Geolocation routing allows our European customers to be sent to our European servers, and it allows our US customers to be sent to our US servers. It doesn't have anything to do about latency, it actually physically knows where our customers are, and then it makes routine decisions based on that.

Geoproximity Routing (Traffic Flow Only)

Geoproximity routing lets Amazon Route 53 route traffic to you resources based on the geographic location of your users and your resources. You can also optionally choose to route more traffic or less to a given resource by specifying a value, know as a bias. A bias expands or shrinks the size of the geographic region from which traffic is routed to a resource.

To use geoproximity routing, you **must** use **Route53 traffic flow**.

Multivalue Answer Policy

Essentially the same as with simple based routing, except you get **Health Checks.** So you can have multiple values within your record sets. And if one fails a health check, it's going to remove that from the record set, or it's going to stop serving that record until it passes the health check.

---

8) VPC Summary

★Very Important Exam Topic (Refer also to Most Important Charts & Diagrams Document)

Remember the following;

-Think of a VPC as a logical datacenter in AWS.

-Consists of IGWs (OR Virtual Private Gateways), Route Tables, Network Access Control Lists, Subnets, and Security Groups

-1 Subnet = 1 Availability Zone

-Security Groups are Stateful; Network Access Control Lists are Stateless (So with a security group we just need to open up a port and the outbound is then taken care of automatically. Whereas with NACLs we have to manually make the changes to both inbound and outbound)

-No Transitive Peering (They need to be peered on a one to one basis.)

-When you create a VPC a default Route Table, Default Network Access Control List (NACL) and a default Security Group are all created automatically.

-When you create a VPC it won't create any subnets, nor will it create a default internet gateway.

-US-East-1A in your AWS account can be a completely different availability zone to US-East-1A in another AWS account. The AZ's are randomized.

-Amazon always reserves 5 IP addresses within your subnets.

-You can only have 1 Internet Gateway per VPC

-Security Groups cannot span VPCs.

NAT Instances

-When creating a NAT instance, Disable Source/Destination Checks on the instance

-NAT instances must be in a public subnet.

-There must be a route out of the private subnet to the NAT instance, in order for this to work.

-The amount of traffic that a NAT instance can support depends on the instance size. If you are bottlenecking, increase the NAT instance size.

-You can create high availability using Autoscaling Groups, multiple subnets in different AZs, and a script to automate failover.

-Behind a Security Group

How a NAT Gateway Works

★Very Important Exam Topic (Refer also to Most Important Diagrams Document)

You have got an instance inside your private subnet. It has a route in its route table, to the NAT gateway which is located in the public subnet. and the NAT gateway has a route out to the internet. And so when your instance runs a Yum update, it's going to the NAT gateway, and then traversing out to the instance.

Note that your NAT gateway is NOT behind a security group. It exists on its own. (Whereas, security groups do associate with your NAT instance and the resources behind your NAT instance to control inbound and outbound traffic...

per https://docs.aws.amazon.com/vpc/latest/userguide/vpc-nat-comparison.html)

NAT Gateways

-Redundant inside the availability zone

-Preferred by the enterprise

-Starts at 5Gbps and scales currently to 45Gbps

-No need to patch

-Not associated with security groups

-Automatically assigned a public ip address

-Remember to update your route tables.

-No need to disable Source/Destination Checks on a NAT Gateway

Nat Gateways (continued)

-If you have resources in multiple Availability Zones and they share one NAT gateway, in the event that the NAT gateway's Availability Zone is down, resources in the other Availability Zones lose internet access. So just having only a single NAT Gateway does not mean you have high availability! To create an Availability Zone-independent architecture, create a NAT gateway in each Availability Zone and configure your routing to ensure that resources use the NAT gateway in the same Availability Zone in which they are in.

Network ACLs

-Your VPC automatically comes with a default network ACL, and by default it allows all outbound and inbound traffic.

-You can create custom network ACLs. By default, each custom network ACL denies all inbound and outbound traffic until you add rules.

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

-Block IP Addresses using network ACLs not Security Groups.

-You can associate a network ACL with multiple subnets; however, a subnet can be associated with only one network ACL at a time. When you associate a network ACL with a subnet, the previous association is removed.

-Network ACLs contain a numbered list of rules that is evaluated in order, starting with the lowest numbered rule. (Remember if we have an allow and then a deny the allow is going to trump the deny, because it is evaluated first. So if you are going to deny something you must put it in front of your allow rule)

-Network ACLs have separate inbound and outbound rules, and each rule can either allow or deny traffic. (Whereas security groups just allow)_

-Network ACLs are stateless; responses to allowed inbound traffic are subject to the rules for outbound traffic (and vice versa).

ELBs And VPCs

-You need a minimum of two public subnets to deploy an internet facing loadbalancer.

VPC Flow Logs

-You cannot enable flow logs for VPCs that are peered with your VPC unless the peer VPC is in your account. You can't have flow logs across multiple AWS accounts. You can have them across multiple VPCs but they have to be within the same account.

-You can tag flow logs.

-After you've created a flow log, you cannot change its configuration; for example, you can't associate a different IAM role with the flow log.

-Not all IP Traffic is monitored;

> -Traffic generated by instances when they contact the Amazon DNS server is NOT going to be logged. If you use your own DNS server, then all traffic to that DNS server is logged.
>
> -Traffic generated by a Windows instance for Amazon Windows license activation is NOT going to be monitored.
>
> -Traffic to and from 169.254.169.254 for instance metadata and user data is not going to be monitored.
>
> -DHCP traffic is NOT going to be monitored.

-Traffic to the reserved IP addresses for the default VPC router will NOT be monitored.

Bastion Hosts

★Very Important Exam Topic (Refer also to Most Important Charts & Diagrams Document)

We've got our instance in a private subnet, if it wants to connect out to the internet it's going to do that using a NAT instance or NAT gateway. If, however, we wanna SSH in or RDP into our instances in our private subnet, we do that via a Bastion host (which is located in your public subnet), and sometimes these are called Jump Boxes as well. So, just remember that a NAT gateway or NAT instance is used to provide internet traffic to EC2 instances in private subnets. A Bastion is used to securely administer EC2 instances using SSH or RDP.

Bastions vs NAT Gateways/Instances

Remember the following;

-A Nat Gateway or NAT instance is used to provide internet traffic to EC2 instances in a private subnets.

-A Bastion is used to securely administer EC2 instances (Using SSH or RDP)(So you can SSH or RDP in to your instance located in a private subnet using a bastion which is located in a public subnet). Bastions are also called jump boxes.

-You cannot use a NAT Gateway as a a Bastion host.

Direct Connect

Remember the following;

-Direct Connect directly connects your data center to AWS

-Useful for high throughput workloads (ie lots of network traffic)

-Or if you need a stable and reliable secure connection.

Steps to Creating a Direct Connect Connection

-Create a virtual interface in the Direct Connect console. This is a PUBLIC Virtual Interface

-Go to the VPC console and then to VPN connections. Create a Customer Gateway.

-Create a Virtual Private Gateway

-Attach the Virtual Private Gateway to the desired VPC.

-Select VPN Connections and create new VPN Connection.

-Select the Virtual Private Gateway and the Customer Gateway

-Once the VPN is available, setup the VPN on the customer gateway or firewall

---

Global Accelerator

-AWS Global Accelerator is a service in which you create accelerators to improve availability and performance of your applications for local and global users.

-Your user is going to the AWS edge network and then it is traversing the AWS internal backbone network to get to your EC2, or Application Load Balancer, or various endpoints, etc.

-You are assigned two static IP addresses (or alternatively you can bring your own).

-You can control traffic using traffic dials. This is done within the endpoint group (endpoints are things like an application load balancer, a network load balancer, an EC2 instance, or an elastic IP as well)

-You can control weighting to individual end points using weights.

VPC Endpoints

★Very Important Exam Topic (Refer also to Most Important Charts & Diagrams Document)

A VPC endpoint enables you to privately connect your VPC to supported AWS services and VPC endpoint services powered by Private Link without requiring an internet gateway, NAT device, VPN connection, or AWS Direct Connect connection. Instances in your VPC do not require public IP addresses to communicate with resources in the service. Traffic between your VPC and the other service does not leave the Amazon network.

Endpoints are virtual devices. They are horizontally scaled, redundant, and highly available VPC components that allow communication between instances in your VPC and services without imposing availability risks or bandwidth constraints on your network traffic.

So our instance located within our Private subnet connects to our VPC **Gateway Endpoint** (a type of VPC Endpoint, which is also in our private subnet) which then connects to our AWS Services (either S3 or DynamoDB) (Another form of VPC endpoint is the **Interface Endpoint** which supports many AWS services unlike the Gateway Endpoint which just supports S3 and DynamoDB)

VPC Endpoints (continued)

There are two types of VPC endpoints:

-Interface Endpoints (support many AWS services)

-Gateway Endpoints (only support S3 and DynamoDB)

VPC Private Link Exam Tips (Opening Your Services in a VPC to another VPC) [SAA-C02]

-If you see a question asking about peering VPCs to tens, hundreds, or thousands of customer VPCs, think of AWS PrivateLink

-Doesn't require VPC peering; no route tables, NAT, IGW, etc.

-Requires a Network Load Balancer on the service VPC and an ENI on the customer VPC

Transit Gateway Exam Tips [SAA-C02]

-Transit Gateway allows you to have transitive peering between thousands of VPCs and on-premises data centers

-Works on a hub-and-spoke model

-Works on a regional basis, but you can have it across multiple regions

-You can use it across multiple AWS accounts using RAM (Resource Access Manager)

-You can use route tables to limt how VPCs talk to one another

-Works with Direct Connect as well as VPN connections

-Supports **IP multicast** (not supported by any other AWS services)

In order to simplify your Network topology if you have hundreds of VPN connections coming in, or direct connections coming in, and a lot of VPC peering going on, and you need to support IP multicast then Transit Gateway is what you need. It simplifies your network architecture/topology and it always uses the hub and spoke model.

AWS VPN CloudHub aka VPN Hub Exam Tips [SAA-C02]

-If you have multiple sites, each with its own VPN connection, you can use AWS VPN CloudHub to connect those sites together.

-Hub-and-spoke model

-Low cost; easy to manage

-Operates over the public internet, but all traffic between the customer gateway and the AWS VPN CloudHub is encrypted

Networking Costs on AWS Exam Tips [SAA-C02]

AWS Network Costs

-Use private IP addresses over public IP addresses to save on costs. This then utilizes the AWS backbone network.

-If you want to cut all network costs, group your EC2 instances in the same Availability Zone and use private IP addresses. This will be cost-free, but make sure to keep in mind single point of failure issues. If your AZ goes down you are going to lose your entire application.

-Scenario based questions talking about cost optimization you always want to use private IPs over public IPs and if you want to cut all your network costs then just put everything inside the same availability zone.

---

9) HA Architecture Summary

3 Different Types of Load Balancers;

-Application Load Balancers (Layer 7 Aware AKA the application layer)(Typically want you use in production because they make intelligent routing decisions)

-Network Load Balancers (For extreme performance or a static IP address) (They are NOT layer 7 aware instead they only go up to layer 4 which is the Transport Layer)

-Classic Load Balancers (do not have intelligent routing...they are for keeping costs down)

-504 Error means the gateway has timed out. This means that the application is not responding within the idle timeout period. (the problem could be the web server layer, application layer, or, the database layer)

-(With classic load balancers) If you need the IPv4 address of your end user, look for the **X-Forwarded-For** header

-Instances monitored by ELB (elastic load balancers) are reported as; InService, or OutofService

-Health Checks check the instance health by talking to it.

-(ALB & CLB) Load Balancers have their own DNS name. You are never given an IP address. (You can however, get a static IP address for a NLB)

-Read the ELB FAQ https://aws.amazon.com/elasticloadbalancing/faqs/

Advanced Load Balancer Theory

★Very Important Exam Topic (Refer also to Most Important Charts & Diagrams Document)

-Sticky Sessions enable your users to stick to the same EC2 instance. Can be useful if you are storing information locally to that instance. (Alternatively, you can disable sticky sessions)

-Cross Zone Load Balancing enables you to load balance your traffic across multiple availability zones. (In a scenario you might need to enable Cross Zone Load Balancing so that traffic is able to go to another availability zone.)

-Path patterns allow you to direct traffic to different EC2 instances based on the URL contained in the request (For example, we may want to direct myURL.com to our web servers, whereas myURL.com/images we may want to direct to our media based servers in a separate availability zone).

Auto Scaling

Auto Scaling has 3 Components

1) **Groups -** Logical component. Webserver group or Application group or Database group etc.

2) **Configuration Templates** - Groups uses a launch template or a launch configuration as a ocnfiguration 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.

3) **Scaling Options** - Scaling Options 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.

Scaling Options (there are 5 scaling options)

1-Maintain current instance levels at all times

You can configure your Auto Scaling group to maintain a specified number of running instances at all times. To maintain the current instance levels, Amazon EC2 Auto Scaling performs a periodic health check on running instances within an auto scaling group. When EC2 auto scaling finds an unhealthy instance, it terminates that instance and launches a new one.

2-Scale Manually

The most basic way to scale your resources, where you specify only the change in the maximum, minimum, or desired capacity of your auto scaling group. Amazon EC2 Auto Scaling manages the process of creating or terminating instances to maintain the updated capacity.

3-Scale based on a schedule

Scaling by schedule means that scaling actions are performed automatically as a function of time and date (i.e. Monday at 9am). This is useful when you know exactly when to increase or decrease the number of instances in your group, simply because the need arises on a predictable schedule.

4-Scale based on demand

A more advanced way to scale your resources - using scaling policies - lets you define parameters that control the scaling process. For example, let's say that you have a web application that currently runs on two instances and you want the CPU utilization of the Auto Scaling group to stay at around 50 percent when the load on the application changes. This method is useful for scaling in reponse to changing conditions, when you don't know when those conditions will change. You can set up Amazon EC2 Auto Scaling to respond for you. We will do this in the next lab.

5-Use predictive scaling

You can use Amazon EC2 Auto Scaling in combination with AWS Auto Scaling to scale resources across multiple services. AWS Auto Scaling can help you maintain optimal availability and performance by combining predictive scaling and dynamic scaling (proactive and reactive approaches, respectively) to scale your Amazon EC2 capacity faster. So Predictive scaling is basically predicting based on your previous performance when you are going to need scaling options.

<u>CloudFormation</u>

-Is a way of completely scripting your cloud environment

-Quick Start is a bunch of CloudFormation templates already built by AWS Solution Architects allowing you to create complex environments very quickly.

Elastic Beanstalk

With Elastic Beanstalk, you can quickly deploy and manage applications in the AWS Cloud without worrying about the infrastructure that runs those applications. You simply upload your application, and Elastic Beanstalk automatically handles the details of capacity provisioning, load balancing, scaling, and application health monitoring. (CloudFormation is more powerful than Elastic Beanstalk)

Highly Availability with Bastion Hosts Exam Tips [SAA-C02]

High Availability with Bastion Hosts

-Two hosts in two separate Availability Zones. Use a Network Load Balancer with static IP addresses and health checks to fail over from one host to the other.

-Can't use an Application Load Balancer (because it is layer 7) so you need to use layer 4 and that is why we use a Network Load Balancer.

-In a scenario where there is one host in one availability zone behind an auto scaling group with health checks and a fixed EIP…If the hosts fails, the health check will fail and the auto scaling group will provision a new EC2 instance in a separate availability zone. You can use a user data script to provision the same EIP to the new host. This is the cheapest option, but it is not 100% fault tolerant. There will be down time (the down time that it takes for the health check to fail and down time while provisioning the new bastion host)

On Premise Strategies Exam Tips [SAA-C02]

You need to be aware of what high-level AWS services you can use on-premise: Database Migration Service (DMS), Server Migration Service (SMS), AWS Application Discovery Service, VM Import/Export, and you can Download Amazon Linux 2 as an ISO as well.

---

10) Applications Summary

SQS (Simple Queue Service) Exam Tips

-SQS is a way to de-couple your infrastructure (for example, let's say you store your messages in SQS and your EC2 instances will go get those messages. If those EC2 instances stop working the message will

then reappear in the queue and another EC2 instance will go in and get it. In this way if something fails it doesn't break your entire application stack)

-SQS is pull based, not pushed based. (If you want push based messaging you are going to want SNS (Simple Notification Service)

-Messages are 256 KB in size. (You can have them bigger but it wont be stored in SQS it will be stored in S3)

-Messages can be kept in the queue from 1 minute to 14 days; the default retention period is 4 days.

Standard SQS and FIFO SQS

★Very Important Exam Topic (Refer also to Most Important Charts & Diagrams Document)

-Standard order is **not** guaranteed and messages can be delivered more than once.

-FIFO order is strictly maintained and messages are delivered only once.

-Visibility Time Out is the amount of time that the message is invisible in the SQS queue after a reader picks up that message. Provided the job is processed before the visibility time out expires, the message will then be deleted from the queue. If the job is not processed within that time, thte message will become visible again and another reader will process it. This could result in the same message being delivered twice. (if you are getting the same message being delivered twice and this is the root cause you may want to increase your visibility timeout, just to give your EC2 instances that little bit of extra time to process the message)

-Visibility timeout maximum is 12 hours.

-SQS guarantees that your message will be processed at least once.

-Amazon SQS long polling is a way to retrieve messages from your Amazon SQS queues. While the regular short polling returns immediately (even if the message queue being polled is empty), long polling doesn't return a response until a message arrives in the message queue, or the long poll times out. This is a way of saving money with SQS)

<u>SWF (Simple Workflow Service) vs SQS (Simple Queue Service)</u>

-SQS has a retention period of up to 14 days; with SWF, workflow executions can last up to 1 year.

-Amazon SWF presents a task-oriented API, whereas Amazon SQS offers a message-oriented API

(message based system).

-Scenario based question that includes 'human workers', the answer will be SWF

-Amazon SWF ensures that a task is assigned only once and is never duplicated. With Amazon SQS, you need to handle duplicated messages and may also need to ensure that a message is processed only once.

-Amazon SWF keeps track of all the tasks and events in an application. With Amazon SQS, you need to implement your own application-level tracking, especially if your application uses multiple queues.

SWF Actors

-Workflow Starters -- An application that can initiate (start) a workflow. Could be your ecommerce website following the placement of an order, or a mobile app searching for bus times.

-Deciders -- Control the flow of activity tasks in a workflow execution. If something has finished (or failed) in a workflow, a Decider decides what to do next.

-Activity Workers -- Carry out the activity tasks. (Amazon.com use SWF to run there warehouses. So scenarios pertaining to human workers will be referring to SWF)

SNS Benefits

-Instantaneous, push-based delivery (no polling)

-Simple APIs and easy integration with applications

-Flexible message delivery over multiple transport protocols

-Inexpensive, pay-as-you-go model with no up-front costs

-Web-based AWS Management Console offers the simplicity of a point-and-click interface

---

SNS vs SQS?

-Both Messaging Services in AWS

-SNS - Push

-SQS - Polls (pulls)

Elastic Transcoder

Just remember that Elastic Transcoder is a media transcoder in the cloud. It converts media files from their original source format in to different formats that will play on smartphones, tablets, PCs, etc.

API Gateway

-API Gateway is a gateway into your AWS resources/infrastructure

-Has caching capabilities to increase performance

-Low cost and scales automatically

-You can throttle API Gateway to prevent attacks

-You can log results to CloudWatch

-If you are using JavaScript/AJAX that uses multiple domains with API Gateway, ensure that you have enabled **CORS (Cross Origin Resource Sharing)** on API Gateway

-CORS is enforced by the client's browser

Kinesis

★Very Important Exam Topic (Refer also to Most Important Diagrams Document)

-Know the difference between Kinesis Streams and Kinesis Firehose. You will be given scenario questions and you must choose the most relevant service.

-**Kinesis Streams** has data persistence and will store your data by default for 24 hours but can go up to a longer period of time. (you have your ec2 instances going in and getting the data from the streams)

-**Kinesis Firehose** is wher you need to analyze your data in real time and then find the place to store it because there is NO persistence. (You could have lambda functions in firehose that store that data into S3 or they could put it into Elastic Search clusters, etc.)

-Shards and data persistence pertain to Kinesis Streams

-Realtime analytics, with no data persistence pertains to Kinesis Firehose

-**Kinesis Analytics** helps you analyze your data in both firehose as well as Kinesis streams.

Cognito

-Cognito allows us to to do Web Identity Federation

-Federation allows users to authenticate with a Web Identity Provider (Google, Facebook, Amazon)

-The user authenticates first with the Web ID Provider and receives an authentication token, which is exchanged for temporary AWS credentials allowing them to assume an IAM role.

-Cognito is an Identity Broker which handles interaction between your applications and the Web ID provider (You don't need to write your own code to do this.)

Cognito (continued)

-**User pool** is user based. It handles things like user registration, authentication, and account recovery.

-**Identity pools** authorise access to your AWS resources

Event Processing Patters Exam Tips [SAA-C02]

-Understand the pub/sub pattern – facilitated by SNS

-Know which AWS services support Dead Letter Queues (DLQ). They are SNS, SQS, and Lambda

-Understand the Fanout pattern and how SNS is involved in it.

-Understand how S3 event notifications work – which events trigger and which services consume.

---

11) Security [SAA-C02]

Reducing Security Threats

-To block an IP or a range of IPs use a NACL which operates on layer 4.

-In addition to a NACL you can use a host-based firewall which runs directly on your EC2 instance

-When using an ALB the incoming connection from your malicious client will terminate at the ALB so your ec2 instance will be unaware of that origin IP (so a host-based firewall would be ineffective). One potential security measure would be to allow only the ALB security group access to the EC2 security group but we still have to use a NACL.

-WAF (Web Application Firewall): for IP blocking and filtering, **blocking SQL injections & Cross-site scripting attacks**. Operates on layer 7. For a public web application a WAF is preferable.

-You can attach WAF to your CloudFront distribution. (Similar to when using an ALB) The client's connection terminates at CloudFront so the client IP won't even be visible to your NACL (only the CloudFront IP is passed along to the NACL) which is why you should use a WAF attached to your CloudFront distribution. (When using an NLB the malicious client IP is visible from end to end so you need to use a NACL in this case to block that IP)

-You can use CloudFront's Geo match feature to block a country altogether.

Key Management Service

-**Regional** secure key management and encryption and decryption

-Manages **customer master keys** (CMKs)

-Ideal for S3 objects, database passwords and API keys stored in Systems Manager Parameter Store

-Encrypt and decrypt data up to 4 KB in size

-Integrated with most AWS services

-Pay per API call

-Audit capability using CloudTrail-logs delivered to S3

-KMS has **FIPS 140-2 Level 2** security

- By the way CloudHSM has Level 3 (even higher security than Level 2)

Three Types of CMKs

AWS Managed CMK: Free; used by default if you pick encryption in most AWS services. Only that service can use them directly.

Customer Managed CMK: Allows key **rotation**; controlled via key policies and can be enabled/disabled.

AWS Owned CMK: Used by AWS on a shared basis across many accounts; you typically won't see these.

Symmetric vs. Asymmetric CMKs

-Symmetric CMKs use the same key for encryption and decryption, based on AES-256 standard algorithm, you can import you own key material.

-Asymmetric CMK's are based on RSA and elliptic-curve cryptography (ECC) algorithms. The private key never leaves AWS unencrypted. You must use the KMS APIs to use the private key. You can download the public key and use it outside of AWS. AWS service integrated with KMS do NOT support Assymetric CMKs.

CloudHSM

-CloudHSM is a Hardware Security Module (HSM) which provides secure key storage and cryptographic operations within a tamper-resistant hardware device.

-If there are regulatory compliance requirements choose CloudHSM as it is **FIPS 140-2 Level 3.**

Systems Manager Parameter Store

-Parameter Store is a component of AWS Systems Manager (SSM) which offers **serverless** storage for configuration & Secrets (Passwords, Database Connection Strings, License Codes, API Keys, etc). The data is stored **hierarchically**.

Secrets Manager

-Similar to systems manager parameter store (which is free). However, Secrets manager **charges** per secret stored & per 10,000 API calls. The difference is that Secrets manager automatically rotates secrets and generates random secrets. It applies the new key/password in RDS for you.

AWS Shield

-Protects against distributed denial of service (DDOS) attacks

AWS Shield Standard vs AWS Shield Advanced

-AWS Shield Standard: Automatically enabled for all customers at **no cost**, protects against layer 3 and 4 attacks. SYN/UDP floods, Reflection attacks.

-AWS Shield Advanced: $3,000 per month per organization. Enhanced protections from EC2, ELB, CloudFront, Global Accelerator, Route53. Business and Enterprise support customers get 24x7 access to the DDOS Response Team (DRT). Also, offers DDOS cost protection.

WAF (Web Application Firewall)

-Web Application Firewall that lets you monitor HTTP(S) requests to CloudFront, ALB, or API Gateway. WAF lets you control access to content using **filtering rules** to allow or deny traffic. You can filter by ip address, and query string parameters. WAF protects against **SQL query injections** and **Cross-site scripting**. Blocked traffic returns asHTTP 403 forbidden.

Three Behaviors of WAF

-1) Allow all requests, except the ones you specify

-2) Block all requests, except the ones you specify

-3) Count the requests that match the properties you specify

-Properties include originating IP, originating country, request size, values in headers, etc

AWS Firewall Manager: Centrally configure and manage firewall rules across an **AWS Organization**. You

can deploy WAF rules for your ALB, API Gateway, and CloudFront distributions. You can create AWS Shield advanced protection policies to automatically discover resources such as ALBs, Classic ELBs, EIPs, or CloudFront distributions and apply DDOS protection to those resources. You can use firewall manager to configure security groups across your aws organization and audit them.

---

12) Serverless Summary

Traditional vs Serverless Architecture

★Very Important Exam Topic (Refer also to Most Important Charts & Diagrams Document)

Traditional Architecture

This is where you've got your Elastic Load Balancer at the front end. You then have got your EC2 instances, so these would be your web servers, and then there's storing data in a traditional database, like RDS. (It can be made to be Highly Available, and can tolerate failure if done correctly but it is limited in scale)(Yes you do have autoscaling groups but bottlenecks will occur with your RDS instances although with Aurora it will not bottleneck as much)(The bottom line is that traditional architecture is NOT as scalable as serverless architecture)

Serverless Architecture

Typically consists of API Gateway, lambda function, and then DynamoDB on the backend.

Lambda

-Lambda scales out (not up) automatically

-example, if five people upload five different images to an S3 bucket, that is going to trigger five separate lambda functions at any given time.

-Lambda functions are independent, 1 event = 1 function

-Lambda is serverless

-RDS is NOT serverless, with the exception of Aurora Serverless

-Lambda functions can trigger other Lambda functions, 1 event can = x functions if functions trigger other functions. (can trigger 10, 20, 50 functions, etc...even a million different functions if functions are triggering other functions)

-Architectures can get extremely complicated, **AWS X-ray** allows you to debug what is happening inside your serverless application

-Lambda can do things globally (is a global service), you can use it to back up S3 buckets to other S3 buckets etc.

-You can configure your functions to run at a particular time of day, at a particular time of week.

-Know what triggers Lambda and what does NOT trigger Lambda

There is no way to directly trigger Lambda functions from RDS

https://docs.aws.amazon.com/lambda/latest/dg/lambda-invocation.html

---

Serverless Application Model (SAM) [SAA-C02]

-SAM is an open-source framework that allows you to build serverless applications really easily

-SAM is a CloudFormation extension optimized for serverless applications

-You can define functions, APIs, tables

-SAM supports anything that CloudFormation supports

-Run serverless applications locally

-Package and deploy using CodeDeploy

Elastic Container Service [SAA-C02]

-A container is a package that contains an application, libraries, runtime, and tools required to run it.

-Containers run on a container engine such as Docker

-Containers provide the isolation benefits of virtualization with less overhead and faster starts than VMs

-Containerized applications are portable and offer a consistent environment

-ECS is a FREE managed container orchestration service that lets you run and scale containerized applications. It eliminates the need for you to manage your own orchestration, tooling, or clusters.

-ECS allows you to create clusters to manage fleets of container deployments

-ECS manages EC2 or Fargate instances

-You can use docker to build and package applications to containers and then easily deploy these

applications into your AWS environment with ECS

-Schedules containers for optimal placement and monitors resource utilization

-Defines rules for CPU and memory requirements

-Allows you to deploy, update, rollback containers

-Integrates with your VPCs, security groups, and EBS volumes, Elastic Load Balancers

-ECS has Cloudtrail integration and CloudWatch integration

ECS components

-Cluster – logical collection of EC2 resources – either ECS EC2 instances or Fargate instances

-Task Definition – Defines your application. Similar to a dockerfile but for running containers in ECS. Can contain multiple containers.

-Container Definition – Inside a task definition, it defines the individual containers a task uses. Controls CPU and memory allocation and port mappings.

-Task – Single running copy of any containers defined by a task definition. One working copy of an application (DB and web containers would run as two separate tasks)

-Service – Allows task definitions to be scaled by adding tasks. Defines minimum and maximum values.

-Registry – Storage for container images (Elastic Container Register (ECR) or Docker Hub). Used to download images to create containers.

Fargate

-Serverless container engine which works with ECS and EKS (Elastic Kubernetes Service)

EKS (Elastic Kubernetes Service) aka K8S

K8s is open-source software that lets you deploy and manage containerized applications at scale

-Containers are grouped in pod

ECR

The managed docker container registry in AWS. Here you can store manage and deploy container images. It integrates with ECS and EKS. It is Highly Available and is integrated with IAM.

ECS Security

Task Roles allow us to have granular control over the permissions that our tasks have.