Designing Computing Environment

[Edition 02]

[Last Update 210508]

K21Academy
Learn Cloud from Experts

amazon web services

Contents

1 DOCUMENTATION LINK

1. Connect to your Linux instance using SSH

https://docs.aws.amazon.com/AWSEC2/latest/UserGuide/AccessingInstancesLinux.html

2. Amazon EC2

https://aws.amazon.com/ec2/?ec2-whats-new.sortby=item.additionalFields.postDateTime&ec2-whats-new.sort-order=desc

3. Amazon EC2 FAQs

https://aws.amazon.com/ec2/faqs/

4. Amazon EC2 pricing

https://aws.amazon.com/ec2/pricing/

5. Amazon FSx

https://aws.amazon.com/fsx/

6. Amazon FSx for Windows File Server

https://docs.aws.amazon.com/fsx/latest/WindowsGuide/getting-started-step1.html

7. Amazon FSx for Linux

https://aws.amazon.com/fsx/lustre/?nc=sn&loc=3

8. Amazon FSx for Windows File Server FAQ's

https://aws.amazon.com/fsx/windows/faqs/?nc=sn&loc=8

9. Amazon FSx for Windows File Server, reference Architecture

https://aws.amazon.com/quickstart/architecture/amazon-fsx-windows-file-server/

10. Getting Started with Amazon FSx for Lustre

https://docs.aws.amazon.com/fsx/latest/LustreGuide/getting-started.html

11. Amazon EBS
https://aws.amazon.com/ebs/

12. Amazon EBS snapshots
https://docs.aws.amazon.com/AWSEC2/latest/UserGuide/EBSSnapshots.html

2 AMAZON EC2

Amazon Elastic Compute Cloud (Amazon EC2) is a web service that provides resizable compute capacity in the cloud. It is designed to make web-scale computing easier for developers.

With Amazon EC2 you launch virtual server instances on the AWS cloud. Each virtual server is known as an “instance”.

You are limited to running up to a total of 20 On-Demand instances across the instance family, purchasing 20 Reserved Instances, and requesting Spot Instances per your dynamic spot limit per region (by default).

AWS are transitioning to a vCPU based, rather than instance based, limit. This is currently being rolled out and may not feature on the exam yet.

Amazon EC2 currently supports a variety of operating systems including: Amazon Linux, Ubuntu, Windows Server, Red Hat Enterprise Linux, SUSE Linux Enterprise Server, Fedora, Debian, CentOS, Gentoo Linux, Oracle Linux, and FreeBSD.

EC2 compute units (ECU) provide the relative measure of the integer processing power of an Amazon EC2 instance.

With EC2 you have full control at the operating system layer.

Key pairs are used to securely connect to EC2 instances:

- A key pair consists of a public key that AWS stores, and a private key file that you store.
- For Windows AMIs, the private key file is required to obtain the password used to log into your instance.
- For Linux AMIs, the private key file allows you to securely SSH (secure shell) into your instance.

Metadata and User Data:

- User data is data that is supplied by the user at instance launch in the form of a script.
- Instance metadata is data about your instance that you can use to configure or manage the running instance.
- User data is limited to 16KB.
- User data and metadata are not encrypted.
- Instance metadata is available at http://169.254.169.254/latest/meta-data/ (the trailing “/” is required).
- Instance user data is available at: http://169.254.169.254/latest/user-data.

- The IP address 169.254.169.254 is a link-local address and is valid only from the instance.
- On Linux you can use the curl command to view metadata and userdata, e.g., "curl http://169.254.169.254/latest/meta-data/".
- The Instance Metadata Query tool allows you to query the instance metadata without having sto type out the full URI or category names.

2.1 Billing & Provisioning

2.1.1 ON DEMAND:

• Pay for hours used with no commitment.

• Low cost and flexibility with no upfront cost.

• Ideal for auto scaling groups and unpredictable workloads.

• Good for dev/test.

2.1.2 SPOT:

• Amazon EC2 Spot Instances let you take advantage of unused EC2 capacity in the AWS cloud.

• Spot Instances are available at up to a 90% discount compared to On-Demand prices.

• You can use Spot Instances for various stateless, fault-tolerant, or flexible applications such as big data, containerized workloads, CI/CD, web servers, high-performance computing (HPC), and other test & development workloads.

• You can request Spot Instances by using the Spot management console, CLI, API or the same interface that is used for launching On-Demand instances by indicating the option to use Spot.

• You can also select a Launch Template or a pre-configured or custom Amazon Machine Image (AMI), configure security and network access to your Spot instance, choose from multiple instance types and locations, use static IP endpoints, and attach persistent block storage to your Spot instances.

• New pricing model: The Spot price is determined by long term trends in supply and demand for EC2 spare capacity.

• You don't have to bid for Spot Instances in the new pricing model, and you just pay the Spot price that's in effect for the current hour for the instances that you launch.

• Spot Instances receive a two-minute interruption notice when these instances are about to be reclaimed by EC2, because EC2 needs the capacity back.

• Instances are not interrupted because of higher competing bids.

• To reduce the impact of interruptions and optimize Spot Instances, diversify and run your application across multiple capacity pools.

• Each instance family, each instance size, in each Availability Zone, in every Region is a separate Spot pool.

• You can use the RequestSpotFleet API operation to launch thousands of Spot Instances and diversify resources automatically.

• To further reduce the impact of interruptions, you can also set up Spot Instances and Spot Fleets to respond to an interruption notice by stopping or hibernating rather than terminating instances when capacity is no longer available.

2.1.3 RESERVED:

• Purchase (or agree to purchase) usage of EC2 instances in advance for significant discounts over On-Demand pricing.

• Provides a capacity reservation when used in a specific AZ.

• AWS Billing automatically applies discounted rates when you launch an instance that matches your purchased RI.

• Capacity is reserved for a term of 1 or 3 years.

• EC2 has three RI types: Standard, Convertible, and Scheduled.

• Standard = commitment of 1 or 3 years, charged whether it's on or off.

• Scheduled = reserved for specific periods of time, accrue charges hourly, billed in monthly increments over the term (1 year).

• Scheduled RIs match your capacity reservation to a predictable recurring schedule.

• For the differences between standard and convertible RIs, see the table below.

• RIs are used for steady state workloads and predictable usage.

• Ideal for applications that need reserved capacity.

• Upfront payments can reduce the hourly rate.

• Can switch AZ within the same region.

• Can change the instance size within the same instance type.

• Instance type modifications are supported for Linux only.

• Cannot change the instance size of Windows RIs.

• Billed whether running or not.

• Can sell reservations on the AWS marketplace.

• Can be used in Auto Scaling Groups.

• Can be used in Placement Groups.

• Can be shared across multiple accounts within Consolidated Billing.

• If you don't need your RI's, you can try to sell them on the Reserved Instance Marketplace.

	Standard	Convertible
Terms	1 year, 3 year	1 year, 3 year
Average discount off On-Demand price	40% - 60%	31% - 54%
Change AZ, instance size, networking type	Yes via ModifyReservedInstance API or console	Yes via ExchangeReservedInstance API or console
Change instance family, OS, tenancy, payment options	No	Yes
Benefit from price reductions	No	Yes

RI Attributes:

• Instance type – designates CPU, memory, networking capability.

• Platform – Linux, SUSE Linux, RHEL, Microsoft Windows, Microsoft SQL Server.

• Tenancy – Default (shared) tenancy, or Dedicated tenancy.

• Availability Zone (optional) – if AZ is selected, RI is reserved, and discount applies to that AZ (Zonal RI). If no AZ is specified, no reservation is created but the discount is applied to any instance in the family in any AZ in the region (Regional RI).

2.2 Comparing Amazon EC2 Pricing Models

The following table provides a brief comparison of On-demand, Reserved and Spot pricing models:

On-Demand	Reserved	Spot
No upfront fee	Options: No upfront, partial upfront or all upfront	No upfront fee
Charged by hour or second	Charged by hour or second	Charged by hour or second
No commitment	1-year or 3-year commitment	No commitment
Ideal for short term needs or unpredictable workloads	Ideal for steady-state workloads and predictable usage	Ideal for cost-sensitive, compute intensive use cases that can withstand interruption

Dedicated hosts:

- Physical servers dedicated just for your use.
- You then have control over which instances are deployed on that host.
- Available as On-Demand or with Dedicated Host Reservation.
- Useful if you have server-bound software licenses that use metrics like per-core, per-socket, or per-VM.
- Each dedicated host can only run one EC2 instance size and type.
- Good for regulatory compliance or licensing requirements.
- Predictable performance.
- Complete isolation.
- Most expensive option.
- Billing is per host.

Dedicated instances:

- Virtualized instances on hardware just for you.
- Also uses physically dedicated EC2 servers.
- Does not provide the additional visibility and controls of dedicated hosts (e.g., how instance is placed on a server).
- Billing is per instance.
- May share hardware with other non-dedicated instances in the same account.
- Available as On-Demand, Reserved Instances, and Spot Instances.

- Cost additional $2 per hour per region.

The following table describes some of the differences between dedicated instances and dedicated hosts:

Characteristic	Dedicated Instances	Dedicated Hosts
Enables the use of dedicated physical servers	X	X
Per instance billing (subject to a $2 per region fee)	X	
Per host billing		X
Visibility of sockets, cores, host ID		X
Affinity between a host and instance		X
Targeted instance placement		X
Automatic instance placement	X	X
Add capacity using an allocation request		X

- Partial instance-hours consumed are billed based on instance usage.
- Instances are billed when they're in a running state – need to stop or terminate to avoid paying.
- Charging by the hour or second (by the second with Linux instances only).
- Data between instances in different regions is charged (in and out).
- Regional Data Transfer rates apply if at least one of the following is true, but are only charged once for a given instance even if both are true:
- The other instance is in a different Availability Zone, regardless of which type of address is used.
- Public or Elastic IP addresses are used, regardless of which Availability Zone the other instance is in.

2.3 Instance Types

Amazon EC2 provides a wide selection of instance types optimized to fit different use cases.

Instance types comprise varying combinations of CPU, memory, storage, and networking capacity and give you the flexibility to choose the appropriate mix of resources for your applications.

Each instance type includes one or more instance sizes, allowing you to scale your resources to the requirements of your target workload.

Category	Families	Purpose/Design
General Purpose	A1, T3, T3a, T2, M5, M5a, M4	General purpose instances provide a balance of compute, memory and networking resources, and can be used for a variety of diverse workloads
Compute Optimized	C5, C5n, C4	Compute Optimized instances are ideal for compute bound applications that benefit from high performance processors
Memory Optimized	R5, R5a, R4, X1e, X1, High Memory, z1d	Memory optimized instances are designed to deliver fast performance for workloads that process large data sets in memory
Accelerated Computing	P3, P2, G4, G3, F1	Accelerated computing instances use hardware accelerators, or co-processors, to perform functions, such as floating-point number calculations, graphics processing, or data pattern matching
Storage Optimized	I3, I3en, D2, H1	This instance family provides Non-Volatile Memory Express (NVMe) SSD-backed instance storage optimized for low latency, very high random I/O performance, high sequential read throughput and provide high IOPS at a low cost

- Options when launching Instances Choose whether to auto-assign a public IP – default is to use the subnet setting. Can add an instance to a placement group.
- Instances can be assigned to IAM roles which configures them with credentials to access AWS resources.
- Termination protection can be enabled and prevents you from terminating an instance.

- Basic monitoring is enabled by default (5-minute periods), detailed monitoring can be enabled (1-minute periods, chargeable).
- Can define shared or dedicated tenancy.
- T2 unlimited allows applications to burst past CPU performance baselines as required (chargeable).
- Can add a script to run on startup (user data).
- Can join to a directory (Windows instances only).
- There is an option to enable an Elastic GPU (Windows instances only).
- Storage options include adding additional volumes and choosing the volume type.
- Non-root volumes can be encrypted.
- Root volumes can be encrypted if the instance is launched from an encrypted AMI.
- There is an option to create tags (or can be done later).
- You can select an existing security group or create a new one.
- You must create or use an existing key pair – this is required.

2.4 Amazon Machine Images

An Amazon Machine Image (AMI) provides the information required to launch an instance.

An AMI includes the following:

- A template for the root volume for the instance (for example, an operating system, an application server, and applications).
- Launch permissions that control which AWS accounts can use the AMI to launch instances.
- A block device mapping that specifies the volumes to attach to the instance when it's launched.
- AMIs are regional. You can only launch an AMI from the region in which it is stored. However, you can copy AMI's to other regions using the console, command line, or the API.
- Volumes attached to the instance are either EBS or Instance store:
- Amazon Elastic Block Store (EBS) provides persistent storage. EBS snapshots, which reside on Amazon S3, are used to create the volume.
- Instance store volumes are ephemeral (non-persistent). That means data is lost if the instance is shut down. A template stored on Amazon S3 is used to create the volume.

2.5 Networking

Networking Limits (per Region or as Specified):

Name	Default Limit
EC2-Classic Elastic IPs	5
EC2-VPC Elastic IPs	5
VPCs	5
Subnets per VPC	200
Security groups per VPC	500
Rules per VPC security group	50
VPC security groups per elastic network interface	5
Network interfaces	350
Network ACLs per VPC	200
Rules per network ACL	20
Route tables per VPC	200
Entries per route table	50
Active VPC peering connections	50
Outstanding VPC peering connection requests	25
Expiry time for an unaccepted VPC peering connection	168

2.6 IP Addresses

There are three types of IP address that can be assigned to an Amazon EC2 instance:

1. Public – public address that is assigned automatically to instances in public subnets and reassigned if instance is stopped/started.
2. Private – private address assigned automatically to all instances.
3. Elastic IP – public address that is static.

- Public IPv4 addresses are lost when the instance is stopped but private addresses (IPv4 and IPv6) are retained.
- Public IPv4 addresses are retained if you restart the instance.
- Elastic IPs are retained when the instance is stopped.
- Elastic IP addresses are static public IP addresses that can be remapped (moved) between instances.
- All accounts are limited to 5 elastic IP's per region by default.
- AWS charge for elastic IP's when they're not being used.
- An Elastic IP address is for use in a specific region only.
- You can assign custom tags to your Elastic IP addresses to categorize them.

- By default, EC2 instances come with a private IP assigned to the primary network interface (eth0).
- Public IP addresses are assigned for instances in public subnets (VPC).
- Public IP addresses are always assigned for instances in EC2-Classic.
- DNS records for elastic IP's can be configured by filling out a form.
- Secondary IP addresses can be useful for hosting multiple websites on a server or redirecting traffic to a standby EC2 instance for HA.
- You can choose whether secondary IP addresses can be reassigned.
- You can associate a single private IPv4 address with a single Elastic IP address and vice versa.
- When reassigned the IPv4 to Elastic IP association is maintained.
- When a secondary private address is unassigned from an interface, the associated Elastic IP address is disassociated.
- You can assign or remove IP addresses from EC2 instances while they are running or stopped.
- All IP addresses (IPv4 and IPv6) remain attached to the network interface when detached or reassigned to another instance.
- You can attach a network interface to an instance in a different subnet as long as it's within the same AZ.

The following table compares the different types of IP address available in Amazon EC2:

Name	Description
Public IP address	Lost when the instance is stopped Used in Public Subnets No charge Associated with a private IP address on the instance Cannot be moved between instances
Private IP address	Retained when the instance is stopped Used in Public and Private Subnets
Elastic IP address	Static Public IP address You are charged if not used Associated with a private IP address on the instance Can be moved between instances and Elastic Network Adapters

2.7 Elastic Network Interfaces

An elastic network interface (referred to as a network interface in this documentation) is a logical networking component in a VPC that represents a virtual network card.

A network interface can include the following attributes:

- A primary private IPv4 address from the IPv4 address range of your VPC
- One or more secondary private IPv4 addresses from the IPv4 address range of your VPC
- One Elastic IP address (IPv4) per private IPv4 address
- One public IPv4 address
- One or more IPv6 addresses
- One or more security groups
- A MAC address
- A source/destination check flag
- A description

- You can create and configure network interfaces in your account and attach them to instances in your VPC.
- You cannot team by adding ENIs to an instance.
- eth0 is the primary network interface and cannot be moved or detached.
- By default, eth0 is the only Elastic Network Interface (ENI) created with an EC2 instance when launched.
- You can add additional interfaces to EC2 instances (number dependent on instances family/type).
- An ENI is bound to an AZ and you can specify which subnet/AZ you want the ENI to be added in.
- You can specify which IP address within the subnet to configure or leave it be auto-assigned.
- You can only add one extra ENI when launching but more can be attached later.
- ENIs can be “hot attached” to running instances.
- ENIs can be “warm-attached” when the instance is stopped.
- ENIs can be “cold-attached” when the instance is launched.
- If you add a second interface AWS will not assign a public IP address to eth0 (you would need to add an Elastic IP).
- Default interfaces are terminated with instance termination.
- Manually added interfaces are not terminated by default.
- You can change the termination behavior.

2.8 Enhanced Networking

2.8.1 ELASTIC NETWORK ADAPTAR (ENA)

- Enhanced networking provides higher bandwidth, higher packet-per-second (PPS) performance, and consistently lower inter-instance latencies.
- Enhanced networking is enabled using an Elastic Network Adapter (ENA).
- If your packets-per-second rate appears to have reached its ceiling, you should consider moving to enhanced networking because you have likely reached the upper thresholds of the VIF driver.
- AWS currently supports enhanced networking capabilities using SR-IOV.
- SR-IOV provides direct access to network adapters, provides higher performance (packets-per-second) and lower latency.
- Must launch an HVM AMI with the appropriate drivers.

- Only available for certain instance types.
- Only supported in VPC.

2.8.2 ELASTIC FABRIC ADAPTER (EFA)

- An Elastic Fabric Adapter is an AWS Elastic Network Adapter (ENA) with added capabilities.
- An EFA can still handle IP traffic, but also supports an important access model commonly called OS bypass.
- This model allows the application (most commonly through some user-space middleware) access the network interface without having to get the operating system involved with each message.
- Elastic Fabric Adapter (EFA) is a network interface for Amazon EC2 instances that enables customers to run applications requiring high levels of inter-node communications at scale on AWS.
- Its custom-built operating system (OS) bypass hardware interface enhances the performance of inter-instance communications, which is critical to scaling these applications.
- With EFA, High Performance Computing (HPC) applications using the Message Passing Interface (MPI) and Machine Learning (ML) applications using NVIDIA Collective Communications Library (NCCL) can scale to thousands of CPUs or GPUs.
- As a result, you get the application performance of on-premises HPC clusters with the on-demand elasticity and flexibility of the AWS cloud.
- EFA is available as an optional EC2 networking feature that you can enable on any supported EC2 instance at no additional cost.

2.8.3 ENI VS ENA VS EFA

When to use ENI:

- This is the basic adapter type for when you don't have any high-performance requirements.
- Can use with all instance types.

When to use ENA:

- Good for use cases that require higher bandwidth and lower inter-instance latency.
- Supported for limited instance types (HVM only).

When to use EFA:

- High Performance Computing.
- MPI and ML use cases.

- Tightly coupled applications.
- Can use with all instance types.

2.9 Placement Groups

Placement groups are a logical grouping of instances in one of the following configurations.

Cluster – clusters instances into a low-latency group in a single AZ: © 2021 Digital Cloud Training 27

1. Cluster

• A cluster placement group is a logical grouping of instances within a single Availability Zone.

• Cluster placement groups are recommended for applications that benefit from low network latency, high network throughput, or both, and if the majority of the network traffic is between the instances in the group.

2. Spread

• Spreads instances across underlying hardware (can span AZs):

• A spread placement group is a group of instances that are each placed on distinct underlying hardware.

• Spread placement groups are recommended for applications that have a small number of critical instances that should be kept separate from each other.

3. Partition

• Partition — divides each group into logical segments called partitions:

• Amazon EC2 ensures that each partition within a placement group has its own set of racks.

• Each rack has its own network and power source. No two partitions within a placement group share the same racks, allowing you to isolate the impact of hardware failure within your application.

• Partition placement groups can be used to deploy large distributed and replicated workloads, such as HDFS, HBase, and Cassandra, across distinct racks.

The table below describes some key differences between clustered and spread placement groups:

	Clustered	Spread
What	Instances are placed into a low-latency group within a single AZ	Instances are spread across underlying hardware
When	Need low network latency and/or high network throughput	Reduce the risk of simultaneous instance failure if underlying hardware fails
Pros	Get the most out of enhanced networking Instances	Can span multiple AZs
Cons	Finite capacity: recommend launching all you might need up front	Maximum of 7 instances running per group, per AZ

- Launching instances in a spread placement group reduces the risk of simultaneous failures that might occur when instances share the same underlying hardware.
- Recommended for applications that benefit from low latency and high bandwidth.
- Recommended to use an instance type that supports enhanced networking.
- Instances within a placement group can communicate with each other using private or public IP addresses.
- Best performance is achieved when using private IP addresses.
- Using public IP addresses, the performance is limited to 5Gbps or less.
- Low-latency 10 Gbps or 25 Gbps network.
- Recommended to keep instance types homogenous within a placement group.
- Can use reserved instances at an instance level but cannot reserve capacity for the placement group.
- The name you specify for a placement group must be unique within your AWS account for the Region.
- You can't merge placement groups.
- An instance can be launched in one placement group at a time; it cannot span multiple placement groups.

- On-Demand Capacity Reservation and zonal Reserved Instances provide a capacity reservation for EC2 instances in a specific Availability Zone. The capacity reservation can be used by instances in a placement group. However, it is not possible to explicitly reserve capacity for a placement group.
- Instances with a tenancy of host cannot be launched in placement groups.

2.10 IAM Roles

- IAM roles are more secure than storing access keys and secret access keys on EC2 instances.
- IAM roles are easier to manage.
- You can attach an IAM role to an instance at launch time or at any time after by using the AWS CLI, SDK, or the EC2 console.
- IAM roles can be attached, modified, or replaced at any time.
- Only one IAM role can be attached to an EC2 instance at a time.
- IAM roles are universal and can be used in any region.

2.11 Bastion/Jump Hosts

- You can configure EC2 instances as bastion hosts (aka jump boxes) in order to access your VPC instances for management.
- Can use the SSH or RDP protocols to connect to your bastion host.
- Need to configure a security group with the relevant permissions.
- Can use auto-assigned public IPs or Elastic IPs.
- Can use security groups to restrict the IP addresses/CIDRs that can access the bastion host.
- Use auto-scaling groups for HA (set to 1 instance to just replace if it fails).
- Best practice is to deploy Linux bastion hosts in two AZs, use auto-scaling and Elastic IP addresses.

2.12 EC2 Migration

VM Import/Export is a tool for migrating VMware, Microsoft, XEN VMs to the Cloud.

Can also be used to convert EC2 instances to VMware, Microsoft or XEN VMs.

Supported for:

- Windows and Linux.
- VMware ESX VMDKs and (OVA images for export only).
- Citrix XEN VHD.
- Microsoft Hyper-V VHD.

Can only be used via the API or CLI (not the console).

Stop the VM before generating VMDK or VHD images.

AWS has a VM connector plugin for vCenter:

- Allows migration of VMs to S3.
- Then converts into a EC2 AMI.
- Progress can be tracked in vCenter.

2.13 Monitoring

EC2 status checks are performed every minute and each returns a pass or a fail status.

- If all checks pass, the overall status of the instance is OK.
- If one or more checks fail, the overall status is impaired.
- System status checks detect (StatusCheckFailed_System) problems with your instance that require AWS involvement to repair.
- Instance status checks (StatusCheckFailed_Instance) detect problems that require your involvement to repair.
- Status checks are built into Amazon EC2, so they cannot be disabled or deleted.
- You can, however, create or delete alarms that are triggered based on the result of the status checks.
- You can create Amazon CloudWatch alarms that monitor Amazon EC2 instances and automatically perform an action if the status check fails.

Actions can include:

- Recover the instance (only supported on specific instance types and can be used only with StatusCheckFailed_System).
- Stop the instance (only applicable to EBS-backed volumes).
- Terminate the instance (cannot terminate if termination protection is enabled).

- Reboot the instance.
- It is a best practice to use EC2 to reboot instance rather than the OS (create a CloudWatch record).

CloudWatch Monitoring frequency:

- Standard monitoring = 5 mins
- Detailed monitoring = 1 min (chargeable)

2.14 Tags

- A tag is a label that you assign to an AWS resource.
- Used to manage AWS assets.
- Tags are just arbitrary name/value pairs that you can assign to virtually all AWS assets to serve as metadata.
- Each tag consists of a key and an optional value, both of which you define.
- Tagging strategies can be used for cost allocation, security, automation, and many other uses. For example, you can use a tag in an IAM policy to implement access control.
- Enforcing standardized tagging can be done via AWS Config rules or custom scripts. For example, EC2 instances not properly tagged are stopped or terminated daily.
- Most resources can have up to 50 tags.

2.15 Resource Groups

- Resource groups are mappings of AWS assets defined by tags.
- Create custom consoles to consolidate metrics, alarms and config details around given tags.

2.16 High Availability Approaches For Compute

- Up-to-date AMIs are critical for rapid fail-over.
- AMIs can be copied to other regions for safety or DR staging.
- Horizontally scalable architectures are preferred because risk can be spread across multiple smaller machines versus one large machine.
- Reserved instances are the only way to guarantee that resources will be available when needed.
- Auto Scaling and Elastic Load Balancing work together to provide automated recovery by maintaining minimum instances.

- Route 53 health checks also provide “self-healing” redirection of traffic.

2.17 Migration

- AWS Server Migration Service (SMS) is an agent-less service which makes it easier and faster for you to migrate thousands of on-premises workloads to AWS.
- AWS SMS allows you to automate, schedule, and track incremental replications of live server volumes, making it easier for you to coordinate large-scale server migrations.
- Automates migration of on-premises VMware vSphere or Microsoft Hyper-V/SCVMM virtual machines to AWS.
- Replicates VMs to AWS, syncing volumes and creating periodic AMIs.
- Minimizes cutover downtime by syncing VMs incrementally.
- Supports Windows and Linux VMs only (just like AWS).
- The Server Migration Connector is downloaded as a virtual appliance into your on-premises vSphere or Hyper-V environments.

2.18 Sample Questions

Q1 : You terminate an EC2 instance and find that the EBS root volume that was attached to the instance was also deleted. How can you correct this?

A. You can't, a root volume is always deleted when the EC2 instance attached to that volume is deleted.
B. Take a snapshot of the EBS volume while the EC2 instance is running. Then, when the EC2 instance is terminated, you can restore the EBS volume from the snapshot.
C. Remove termination protection from the EC2 instance.
D. Use the AWS CLI to change the DeleteOnTermination attribute for the EBS volume to "false."

Answer: D
Explanation: By default, EBS root volumes are terminated when the associated instance is terminated. However, this is only the default value; therefore, A is not correct. Option B is not directly addressing the question; the EBS volume would still be deleted even if you take a snapshot. Option C is not relevant, but option D is: You can use the AWS CLI (or the console) to set the root volume to persist after instance termination.

Q2 : Can you attach an EBS volume to more than one EC2 instance at the same time?

A. Yes, as long as the volume is not the root volume.
B. No, EBS volumes cannot be attached to more than one instance at the same time.
C. Yes, as long as the volume is one of the SSD classes and not magnetic storage.
D. Yes, as long as at least one of the instances uses the volume as its root volume.

Answer: B
Explanation: EBS volumes can only attach to a single instance at one time. The other options are all simply to distract.

For more Questions Please check Certification Sample Quiz under each module
Link: https://k21academy.com/awssaquizm05

3 AMAZON EBS

- EBS is the Elastic Block Store.
- EBS volumes are network attached storage that can be attached to EC2 instances.
- EBS volume data persists independently of the life of the instance.
- EBS volumes do not need to be attached to an instance.
- You can attach multiple EBS volumes to an instance.
- You cannot attach an EBS volume to multiple instances (use Elastic File Store instead).
- EBS volume data is replicated across multiple servers in an AZ.
- EBS volumes must be in the same AZ as the instances they are attached to.
- EBS is designed for an annual failure rate of 0.1%-0.2% & an SLA of 99.95%.
- Termination protection is turned off by default and must be manually enabled (keeps the volume/data when the instance is terminated).
- Root EBS volumes are deleted on termination by default.
- Extra non-boot volumes are not deleted on termination by default.
- The behavior can be changed by altering the “DeleteOnTermination” attribute.
- You can now create AMIs with encrypted root/boot volumes as well as data volumes (you can also use separate CMKs per volume).
- Volume sizes and types can be upgraded without downtime (except for magnetic standard).
- Elastic Volumes allow you to increase volume size, adjust performance, or change the volume type while the volume is in use.
- To migrate volumes between AZ’s, create a snapshot then create a volume in another AZ from the snapshot (possible to change size and type).
- Auto-enable IO setting prevents the stopping of IO to a disk when AWS detects inconsistencies.
- The root device is created under /dev/sda1 or /dev/xvda.
- Magnetic EBS is for workloads that need throughput rather than IOPS.
- Throughput optimized EBS volumes cannot be a boot volume.
- Each instance that you launch has an associated root device volume, either an Amazon EBS volume or an instance store volume.

- You can use block device mapping to specify additional EBS volumes or instance store volumes to attach to an instance when it's launched. © 2021 Digital Cloud Training 32

- You can also attach additional EBS volumes to a running instance.
- You cannot decrease an EBS volume size.
- When changing volumes, the new volume must be at least the size of the current volume's snapshot.
- Images can be made public but not if they're encrypted.
- AMIs can be shared with other accounts.
- You can have up to 5,000 EBS volumes by default.
- You can have up to 10,000 snapshots by default.

3.1 Instance Store

- An instance store provides temporary (non-persistent) block-level storage for your instance.
- This is different to EBS which provides persistent storage but is also a block storage service that can be a root or additional volume.
- Instance store storage is located on disks that are physically attached to the host computer.
- Instance store is ideal for temporary storage of information that changes frequently, such as buffers, caches, scratch data, and other temporary content, or for data that is replicated across a fleet of instances, such as a load-balanced pool of web servers.
- You can specify instance store volumes for an instance only when you launch it.
- You can't detach an instance store volume from one instance and attach it to a different instance.
- The instance type determines the size of the instance store available and the type of hardware used for the instance store volumes.
- Instance store volumes are included as part of the instance's usage cost.
- Some instance types use NVMe or SATA-based solid-state drives (SSD) to deliver high random I/O performance.
- This is a good option when you need storage with very low latency, but you don't need the data to persist when the instance terminates, or you can take advantage of fault-tolerant architectures.

EXAM TIP: Instance stores offer very high performance and low latency. As long as you can afford to lose an instance, i.e. you are replicating your data, these can be a good solution for high performance/low latency requirements. Look out for questions that mention distributed or replicated databases that need high I/O. Also, remember that the cost of instance stores is included in the instance charges so it can also be more cost-effective than EBS Provisioned IOPS.

3.2 EBS VS Instance Store

- EBS-backed means the root volume is an EBS volume and storage is persistent.
- Instance store-backed means the root volume is an instance store volume and storage are not persistent.
- On an EBS-backed instance, the default action is for the root EBS volume to be deleted upon termination.
- Instance store volumes are sometimes called Ephemeral storage (non-persistent).
- Instance store backed instances cannot be stopped. If the underlying host fails, the data will be lost.
- Instance store volume root devices are created from AMI templates stored on S3.
- EBS backed instances can be stopped. You will not lose the data on this instance if it is stopped (persistent).
- EBS volumes can be detached and reattached to other EC2 instances.
- EBS volume root devices are launched from AMI's that are backed by EBS snapshots.
- Instance store volumes cannot be detached/reattached.
- When rebooting the instances for both types data will not be lost.
- By default, both root volumes will be deleted on termination unless you configured otherwise.

3.3 EBS Volume Types

3.3.1 SSD, GENERAL PURPOSE – GP2

• Baseline of 3 IOPS per GiB with a minimum of 100 IOPS.

• Burst up to 3000 IOPS (for volumes >= 334GB).

• Up to 16,000 IOPS per volume.

• AWS designs gp2 volumes to deliver 90% of the provisioned performance 99% of the time. A gp2 volume can range in size from 1 GiB to 16 TiB.

3.3.2 SSD, PROVISIONED IOPS – I01

• More than 16,000 IOPS.

• Up to 64,000 IOPS per volume.

• Up to 50 IOPS per GiB.

• Amazon EBS delivers the provisioned IOPS performance 99.9 percent of the time.

3.3.3 HDD, THROUGHPUT OPTIMIZED – (ST1):

• Frequently accessed, throughput intensive workloads with large datasets and large I/O sizes, such as MapReduce, Kafka, log processing, data warehouse, and ETL workloads.

• Throughput measured in MB/s, and includes the ability to burst up to 250 MB/s per TB, with a baseline throughput of 40 MB/s per TB and a maximum throughput of 500 MB/s per volume.

• Cannot be a boot volume.

3.3.4 HDD, COLD – (SC1):

• Lowest cost storage – cannot be a boot volume.

• Less frequently accessed workloads with large, cold datasets.

• These volumes can burst up to 80 MB/s per TB, with a baseline throughput of 12 MB/s per TB and a maximum throughput of 250 MB/s per volume.

• HDD, Magnetic – Standard – cheap, infrequently accessed storage – lowest cost storage that cannot be a boot volume.

3.3.5 EBS OPTIMIZED INSTANCES:

• Dedicated capacity for Amazon EBS I/O.

• EBS-optimized instances are designed for use with all EBS volume types.

• Max bandwidth: 400 Mbps – 12000 Mbps.

• IOPS: 3000 – 65000.

• GP-SSD within 10% of baseline and burst performance 99.9% of the time.

• PIOPS within 10% of baseline and burst performance 99.9% of the time.

• Additional hourly fee.

• Available for select instance types.

• Some instance types have EBS-optimized enabled by default.

	Solid State Drives (SSD)		Hard Disk Drives (HDD)	
Volume Type	EBS Provisioned IOPS SSD (io1)	EBS General Purpose SSD (gp2)	Throughput Optimized HDD (st1)	Cold HDD (sc1)
Short Description	Highest performance SSD volume designed for latency-sensitive transactional workloads	General Purpose SSD volume that balances price performance for a wide variety of transactional workloads	Low cost HDD volume designed for frequently accessed, throughput intensive workloads	Lowest cost HDD volume designed for less frequently accessed workloads
Use Cases	I/O-Intensive NoSQL and relational databases	Boot volumes, low-latency interactive apps, dev & test	Big data, data warehouses, log processing	Colder data requiring fewer scans per day
API Name	io1	gp2	st1	sc1
Volume Size	4GB - 16TB	1 GB - 16 TB	500 GB - 16 TB	500 GB - 16 TB
Max IOPS/Volume	64,000	16,000	500	250
Max Throughput/Volume	1,000 MB/s	250 MB/s	500 MB/s	250 MB/s
Max IOPS/Instance	80,000	80,000	80,000	80,000
Max Throughput/Instance	1,750 MB/s	1,750 MB/s	1,750 MB/s	1,750 MB/s

3.4 Snapshots

- Snapshots capture a point-in-time state of an instance.
- Cost-effective and easy backup strategy.
- Share data sets with other users or accounts.
- Can be used to migrate a system to a new AZ or region.
- Can be used to convert an unencrypted volume to an encrypted volume.
- Snapshots are stored on Amazon S3.
- Does not provide granular backup (not a replacement for backup software).
- If you make periodic snapshots of a volume, the snapshots are incremental, which means that only the blocks on the device that have changed after your last snapshot are saved in the new snapshot.
- Even though snapshots are saved incrementally, the snapshot deletion process is designed so that you need to retain only the most recent snapshot in order to restore the volume.
- Snapshots can only be accessed through the EC2 APIs.
- EBS volumes are AZ specific but snapshots are region specific.

- Volumes can be created from EBS snapshots that are the same size or larger.
- Snapshots can be taken of non-root EBS volumes while running.
- To take consistent snapshots writes must be stopped (paused) until the snapshot is complete – if not possible the volume needs to be detached, or if it's an EBS root volume the instance must be stopped.
- To lower storage costs on S3 a full snapshot and subsequent incremental updates can be created.
- You are charged for data traffic to S3 and storage costs on S3.
- You are billed only for the changed blocks.
- Deleting a snapshot removes only the data not needed by any other snapshot.
- You can resize volumes through restoring snapshots with different sizes (configured when taking the snapshot).
- Snapshots can be copied between regions (and be encrypted). Images are then created from the snapshot in the other region which creates an AMI that can be used to boot an instance.
- You can create volumes from snapshots and choose the availability zone within the region.

3.5 Encryption

- You can encrypt both the boot and data volumes of an EC2 instance. When you create an encrypted EBS volume and attach it to a supported instance type, the following types of data are encrypted:
- Data at rest inside the volume.
- All data moving between the volume and the instance.
- All snapshots created from the volume.
- All volumes created from those snapshots.
- Encryption is supported by all EBS volume types.
- Expect the same IOPS performance on encrypted volumes as on unencrypted volumes.
- All instance families support encryption.
- Amazon EBS encryption is available on the instance types listed below:
- General purpose: A1, M3, M4, M5, M5a, M5ad, M5d, T2, T3, and T3a.
- Compute optimized: C3, C4, C5, C5d, and C5n.
- Memory optimized: cr1.8xlarge, R3, R4, R5, R5a, R5ad, R5d, u-6tb1.metal, u-9tb1.metal, u-12tb1.metal, X1, X1e, and z1d.

- Storage optimized: D2, h1.2xlarge, h1.4xlarge, I2, and I3.
- Accelerated computing: F1, G2, G3, G4, P2, and P3.
-
- EBS encrypts your volume with a data key using the industry-standard AES-256 algorithm.
- Your data key is stored on-disk with your encrypted data, but not before EBS encrypts it with your CMK. Your data key never appears on disk in plaintext.
- The same data key is shared by snapshots of the volume and any subsequent volumes created from those snapshots.
- Snapshots of encrypted volumes are encrypted automatically.
- EBS volumes restored from encrypted snapshots are encrypted automatically.
- EBS volumes created from encrypted snapshots are also encrypted.
- You can share snapshots, but if they're encrypted it must be with a custom CMK key.
- There is no direct way to change the encryption state of a volume.
- Either create an encrypted volume and copy data to it or take a snapshot, encrypt it, and create a new encrypted volume from the snapshot.
- To encrypt a volume or snapshot you need an encryption key, these are customer managed keys (CMK) and they are managed by the AWS Key Management Service (KMS).
- A default CMK key is generated for the first encrypted volumes.
- Subsequent encrypted volumes will use their own unique key (AES 256 bit).
- The CMK used to encrypt a volume is used by any snapshots and volumes created from snapshots.
- You cannot share encrypted volumes created using a default CMK key.
- You cannot change the CMK key that is used to encrypt a volume.
- You must create a copy of the snapshot and change encryption keys as part of the copy.
- This is required in order to be able to share the encrypted volume.
- By default, only the account owner can create volumes from snapshots.
- You can share unencrypted snapshots with the AWS community by making them public.
- You can also share unencrypted snapshots with other AWS accounts by making them private and selecting the accounts to share them with.
- You cannot make encrypted snapshots public.

- You can share encrypted snapshots with other AWS accounts using a non-default CMK key and configuring cross-account permissions to give the account access to the key, mark as private and configure the account to share with.
- The receiving account must copy the snapshot before they can then create volumes from the snapshot.
- It is recommended that the receiving account re-encrypt the shared and encrypted snapshot using their own CMK key.

The following information applies to snapshots:

- Snapshots are created asynchronously and are incremental.
- You can copy unencrypted snapshots (optionally encrypt).
- You can copy an encrypted snapshot (optionally re-encrypt with a different key).
- Snapshot copies receive a new unique ID.
- You can copy within or between regions.
- You cannot move snapshots, only copy them.
- You cannot take a copy of a snapshot when it is in a “pending” state, it must be “complete”.
- S3 Server-Side Encryption (SSE) protects data in transit while copying.
- User defined tags are not copied.
- You can have up to 5 snapshot copy requests running in a single destination per account.
- You can copy Import/Export service, AWS Marketplace, and AWS Storage Gateway snapshots.
- If you try to copy an encrypted snapshot without having access to the encryption keys it will fail silently (cross-account permissions are required).

Copying snapshots may be required for:

- Creating services in other regions.
- DR – the ability to restore from snapshot in another region.
- Migration to another region.
- Applying encryption.
- Data retention.

To take application-consistent snapshots of RAID arrays:

- Stop the application from writing to disk.
- Flush all caches to the disk.
- Freeze the filesystem.
- Unmount the RAID array.
- Shut down the associated EC2 instance.

3.6 AMI's

An Amazon Machine Image (AMI) is a special type of virtual appliance that is used to create a virtual machine within the Amazon Elastic Compute Cloud ("EC2").

An AMI includes the following:

- A template for the root volume for the instance (for example, an operating system, an application server, and applications).
- Launch permissions that control which AWS accounts can use the AMI to launch instances.
- A block device mapping that specifies the volumes to attach to the instance when it's launched.
- AMIs are either instance store-backed or EBS-backed.

Instance store-backed:

- Launch an EC2 instance from an AWS instance store-backed AMI.
- Update the root volume as required.
- Create the AMI which will upload to a user-specified S3 bucket (user bucket).
- Register the AMI with EC2 (creates another EC2 controlled S3 image).
- To make changes update the source then deregister and reregister.
- Upon launch the image is copied to the EC2 host.
- Deregister an image when the AMI is not needed anymore (does not affect existing instances created from the AMI).
- Instance store-backed volumes can only be created at launch time.

EBS-backed:

- Must stop the instance to create a consistent image and then create the AMI.
- AWS registers the AMIs manually.

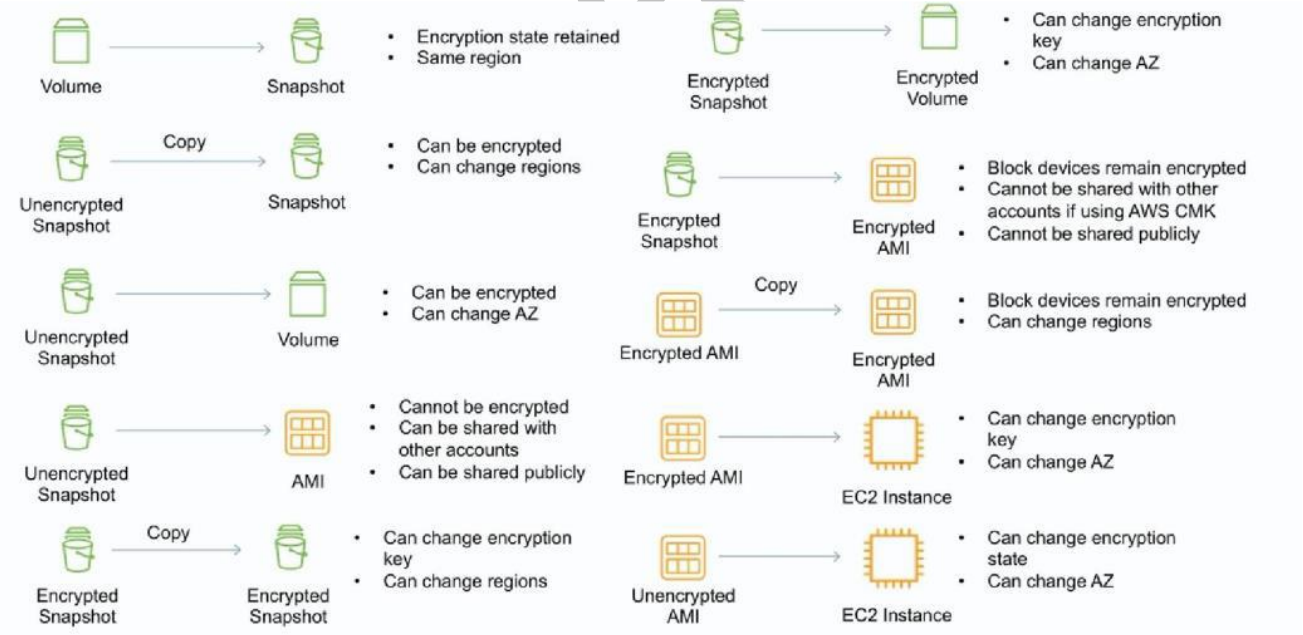
- During creation AWS creates snapshots of all attached volumes – there is no need to specify a bucket but you will be charged for storage on S3.
- You cannot delete the snapshot of the root volume as long as the AMI is registered (deregister and delete).
- You can now create AMIs with encrypted root/boot volumes as well as data volumes (can also use separate CMKs per volume).

Copying AMIs:

- You can copy an Amazon Machine Image (AMI) within or across an AWS region using the AWS Management Console, the AWS Server-Side Command Line Interface or SDKs, or the Amazon EC2 API, all of which support theCopyImage action.
- You can copy both Amazon EBS-backed AMIs and instance store-backed AMIs.
- You can copy encrypted AMIs and AMIs with encrypted snapshots.

3.7 EBS Copying, Sharing & Encryption Methods

The following diagram aims to articulate the various possible options for copying EBS volumes, sharing AMIs and snapshots and applying encryption:



3.8 RAID

- RAID can be used to increase IOPS.
- RAID 0 = 0 striping – data is written across multiple disks and increases performance but no redundancy.
- RAID 1 = 1 mirroring – creates 2 copies of the data but does not increase performance, only redundancy.
- RAID 10 = 10 combination of RAID 1 and 2 resulting in increased performance and redundancy (at the cost of additional disks).
- You can configure multiple striped gp2 or standard volumes (typically RAID 0).
- You can configure multiple striped PIOPS volumes (typically RAID 0).
- RAID is configured through the guest OS.
- EBS optimized EC2 instances are another way of increasing performance.
- Ensure the EC2 instance can handle the bandwidth required for the increased performance.
- Use EBS optimized instances or instances with a 10 Gbps network interface.
- Not recommended to use RAID for root/boot volumes.

3.9 EBS Limits (Per Region)

Name	Default Limit
Provisioned IOPS	300,000
Provisioned IOPS (SSD) volume storage (TiB)	300
General Purpose (SSD) volume storage (TiB)	300
Magnetic volume storage (TiB)	300
Max Cold HDD (SC1) Storage in (TiB)	300
Max Throughput Optimized HDD (ST1) Storage (TiB)	300

3.10 Sample Questions

Q1: What of the following types of data is not encrypted automatically when an encrypted EBS volume is attached to an EC2 instance?

A. Data in transit to the volume
B. Data at rest on the volume
C. Data in transit from the volume
D. All of these are encrypted.

Answer: D

Expalaination: All of these are encrypted. Data moving to and from the volume as well as data at rest on the volume are all encrypted.

Q2: Which of the following will allow you to bring a non-encrypted EBS volume into compliance with an “all data must be encrypted at rest” policy?

A. Create a new volume, attach the new volume to an EC2 instance, copy the data from the non-encrypted volume to the new volume, and then encrypt the new volume.
B. Create a new volume with encryption turned on, attach the new volume to an EC2 instance, and copy the data from the non-encrypted volume to the new volume.
C. Create a new volume, attach the new volume to an EC2 instance, and use the encrypted-copy command to copy the data from the non-encrypted volume to the new volume.
D. None of these will encrypt all data on the volume.

Answer: B

Expalaination: The only way to encrypt an EBS volume is to encrypt it at creation time. Remembering this one detail will help on lots of questions in this vein.

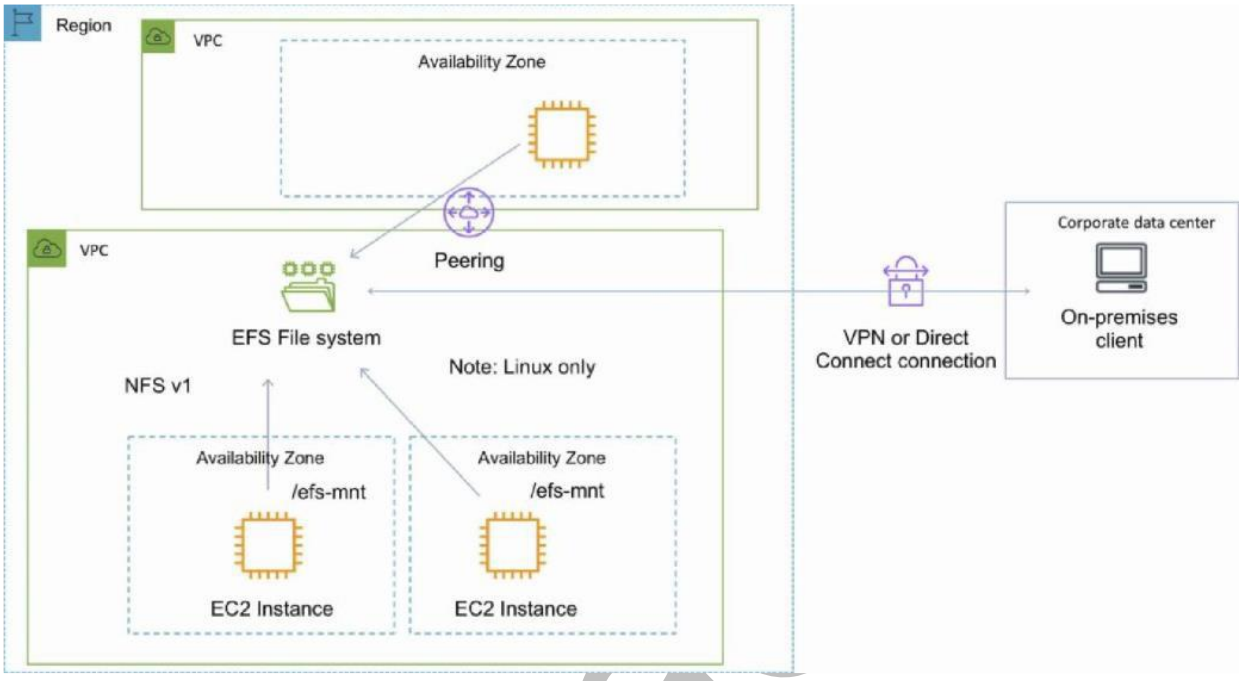
For more Questions Please check Certification Sample Quiz under each module

Link: https://k21academy.com/awssaquizm05

4 AMAZON EFS

- EFS is a fully-managed service that makes it easy to set up and scale file storage in the Amazon Cloud.
- Implementation of an NFS file share and is accessed using the NFSv4.1 protocol.
- Elastic storage capacity and pay for what you use (in contrast to EBS with which you pay for what you provision).
- Multi-AZ metadata and data storage.
- Can configure mount-points in one, or many, AZs.
- Can be mounted from on-premises systems ONLY if using Direct Connect or a VPN connection.
- Alternatively, use the EFS File Sync agent.
- Good for big data and analytics, media processing workflows, content management, web serving, home directories etc.
- Pay for what you use (no pre-provisioning required).
- Can scale up to petabytes.
- EFS is elastic and grows and shrinks as you add and remove data.
- Can concurrently connect 1 to 1000s of EC2 instances, from multiple AZs.
- A file system can be accessed concurrently from all AZs in the region where it is located.

The following diagram depicts the various options for mounting an EFS filesystem:



- By default, you can create up to 10 file systems per account.
- Access to EFS file systems from on-premises servers can be enabled via Direct Connect or AWS VPN.
- You mount an EFS file system on your on-premises Linux server using the standard Linux mount command for mounting a file system via the NFSv4.1 protocol.
- Can choose General Purpose or Max I/O (both SSD).
- The VPC of the connecting instance must have DNS hostnames enabled.
- EFS provides a file system interface, file system access semantics (such as strong consistency and file locking).
- Data is stored across multiple AZ's within a region.
- Read after write consistency.
- Need to create mount targets and choose AZ's to include (recommended to include all AZ's).
- Instances can be behind an ELB.
- EC2 Classic instances must mount via ClassicLink.
- EFS is compatible with all Linux-based AMIs for Amazon EC2.

- Using the EFS-to-EFS Backup solution, you can schedule automatic incremental backups of your Amazon EFS file system.

The following table provides a comparison of the storage characteristics of EFS vs EBS:

	Amazon EFS	Amazon EBS Provisioned IOPS
Availability and durability	Data is stored redundantly across multiple AZs	Data is stored redundantly in a single AZ
Access	Up to thousands of Amazon EC2 instances, from multiple AZs, can connect concurrently to a file system	A single Amazon EC2 instance in a single AZ can connect to a file system
Use cases	Big data and analytics, media processing and workflows, content management, web serving and home directories	Boot volumes, transactional and NoSQL databases, data warehousing and ETL

4.1 Performance

There are two performance modes:

1. "General Purpose" performance mode is appropriate for most file systems.
2. "Max I/O" performance mode is optimized for applications where tens, hundreds, or thousands of EC2 instances are accessing the file system.

- Amazon EFS is designed to burst to allow high throughput levels for periods of time.
- Amazon EFS file systems are distributed across an unconstrained number of storage servers, enabling file systems to grow elastically to petabyte scale and allowing massively parallel access from Amazon EC2 instances to your data.
- This distributed data storage design means that multithreaded applications and applications that concurrently access data from multiple Amazon EC2 instances can drive substantial levels of aggregate throughput and IOPS.

The table below compares high-level performance and storage characteristics for AWS's file (EFS) and block (EBS) cloud storage offerings:

	Amazon EFS	Amazon EBS Provisioned IOPS
Per-operation latency	Low, consistent latency	Lowest, consistent latency
Throughput scale	10+ GB per second	Up to 2 GB per second

4.2 Access Control

- When you create a file system, you create endpoints in your VPC called "mount targets".
- When mounting from an EC2 instance, your file system's DNS name, which you provide in your mount command, resolves to a mount target's IP address.
- You can control who can administer your file system using IAM.
- You can control access to files and directories with POSIX-compliant user and group-level permissions.
- POSIX permissions allow you to restrict access from hosts by user and group.
- EFS Security Groups act as a firewall, and the rules you add define the traffic flow.

4.3 EFS Encryption

- EFS offers the ability to encrypt data at rest and in transit.
- Encryption keys are managed by the AWS Key Management Service (KMS).
- Data encryption in transit uses industry standard Transport Layer Security (TLS) 1.2 to encrypt data sent between your clients and EFS file systems.
- Data encrypted at rest is transparently encrypted while being written, and transparently decrypted while being read.
- Enable encryption at rest in the EFS console or by using the AWS CLI or SDKs.
- Encryption of data at rest and of data in transit can be configured together or separately to help meet your unique security requirements.

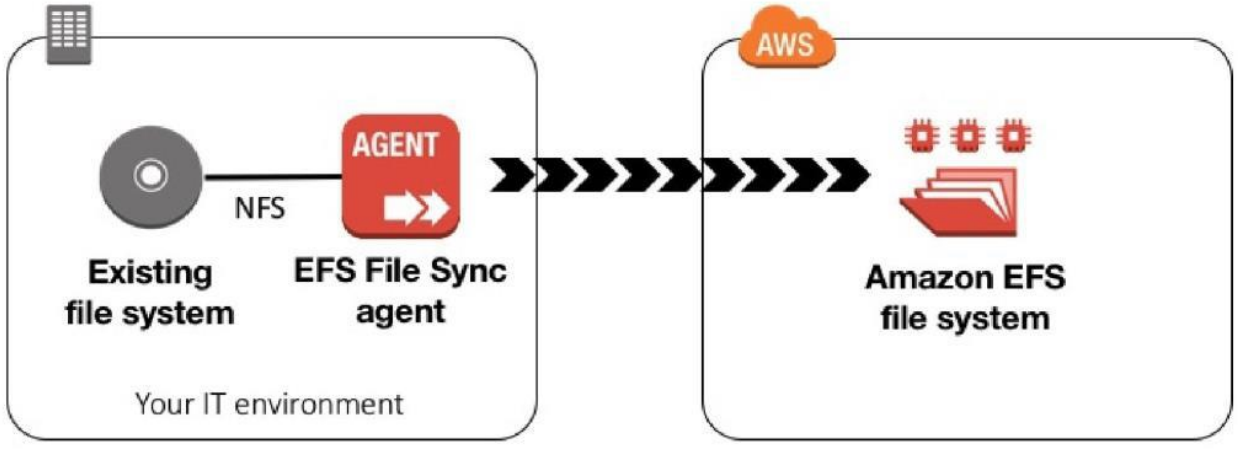
4.4 EFS File Sync

- EFS File Sync provides a fast and simple way to securely sync existing file systems into Amazon EFS.
- EFS File Sync copies files and directories into Amazon EFS at speeds up to 5x faster than standard Linux copy tools, with simple setup and management in the AWS Console.
- EFS File Sync securely and efficiently copies files over the internet or an AWS Direct Connect connection.
- Copies file data and file system metadata such as ownership, timestamps, and access permissions.

EFS File Sync provides the following benefits:

- Efficient high-performance parallel data transfer that tolerates unreliable and high-latency networks.
- Encryption of data transferred from your IT environment to AWS.
- Data transfer rate up to five times faster than standard Linux copy tools.
- Full and incremental syncs for repetitive transfers.

The following diagram shows a high-level view of the EFS File Sync architecture:



Note: EFS File Sync currently doesn't support syncing from an Amazon EFS source to an NFS destination.

- When deploying Amazon EFS File Sync on EC2, the instance size must be at least xlarge for your EFS File Sync to function.
- Recommended to use one of the Memory optimized r4.xlarge instance types.

- Can choose to run EFS File Sync either on-premises as a virtual machine (VM), or in AWS as an EC2 instance.
- Supports VMware ESXi.

4.5 Compatibility

- EFS is integrated with a number of other AWS services, including CloudWatch, CloudFormation, CloudTrail, IAM, and Tagging services.
- CloudWatch allows you to monitor file system activity using metrics.
- CloudFormation allows you to create and manage file systems using templates.
- CloudTrail allows you to record all Amazon EFS API calls in log files.
- IAM allows you to control who can administer your file system.
- Tagging services allows you to label your file systems with metadata that you define.
- PRICING AND BILLING
- You pay only for the amount of file system storage you use per month.
- When using the Provisioned Throughput mode, you pay for the throughput you provision per month.
- There is no minimum fee and there are no set-up charges.
- With EFS File Sync, you pay per-GB for data copied to EFS.

4.6 Sample Questions

Q1: What type of services are associated with EFS?

A. Storage services
B. Networking services
C. Compute services
D. All of the above

Answer: A

Explaination : EFS is the Elastic File System, a scalable file system concerned with storage.

Q2: Which AWS service functions like a NAS in the cloud?

A. EBS
B. Tape gateway
C. EFS
D. DynamoDB

Answer: C

Explanation : EFS, Elastic File System, provides scalable storage accessible from multiple compute instances. EBS is Elastic Block Storage and is tied to one instance at a time and therefore not like a NAS (network attached storage). DynamoDB is a NoSQL database, and tape gateway is a client device for interacting with S3, but locally rather than in the cloud.

For more Questions Please check Certification Sample Quiz under each module

Link: https://k21academy.com/awssaquizm05

5 AMAZON FSX

Amazon FSx provides fully managed third-party file systems.

Amazon FSx provides you with the native compatibility of third-party file systems with feature sets for workloads such as Windows-based storage, high-performance computing (HPC), machine learning, and electronic design automation (EDA).

You don't have to worry about managing file servers and storage, as Amazon FSx automates the time-consuming administration tasks such as hardware provisioning, software configuration, patching, and backups.

Amazon FSx integrates the file systems with cloud-native AWS services, making them even more useful for a broader set of workloads.

Amazon FSx provides you with two file systems to choose from:

1. Amazon FSx for Windows File Server for Windows-based applications
2. Amazon FSx for Lustre for compute-intensive workloads.

5.1 Amazon FSX For Windows File Server

- Amazon FSx for Windows File Server provides a fully managed native Microsoft Windows file system so you can easily move your Windows-based applications that require shared file storage to AWS.
- Built on Windows Server, Amazon FSx provides the compatibility and features that your Microsoft applications rely on, including full support for the SMB protocol, Windows NTFS, and Microsoft Active Directory (AD) integration.
- Amazon FSx uses SSD storage to provide fast performance with low latency.
- This compatibility, performance, and scalability enables business-critical workloads such as home directories, media workflows, and business applications.
- Amazon FSx helps you optimize TCO with Data Deduplication, reducing costs by 50-60% for general-purpose file shares.
- User quotas give you the option to better monitor and control costs. You pay for only the resources used, with no upfront costs, or licensing fees.

Details and Benefits

- High availability: Amazon FSx automatically replicates your data within an Availability Zone (AZ) it resides in (which you specify during creation) to protect it from component failure, continuously monitors for hardware failures, and automatically replaces infrastructure components in the event of a failure.
- Multi-AZ: Amazon FSx offers a multiple availability (AZ) deployment option, designed to provide continuous availability to data, even in the event that an AZ is unavailable. Multi-AZ file systems include an active and standby file server in separate AZs, and any changes written to disk in your file system are synchronously replicated across AZs to the standby.

Supports Windows-native file system features:

- Access Control Lists (ACLs), shadow copies, and user quotas.
- NTFS file systems that can be accessed from up to thousands of compute instances using the SMB protocol.
- Works with Microsoft Active Directory (AD) to easily integrate file systems with Windows environments.
- Built on SSD-storage, Amazon FSx provides fast performance with up to 2 GB/second throughput per file system, hundreds of thousands of IOPS, and consistent sub-millisecond latencies.
- Can choose a throughput level that is independent of your file system size.
- Using DFS Namespaces, you can scale performance up to tens of gigabytes per second of throughput, with millions of IOPS, across hundreds of petabytes of data.
- Amazon FSx can connect file systems to Amazon EC2, VMware Cloud on AWS, Amazon WorkSpaces, and Amazon AppStream 2.0 instances.
- Amazon FSx also supports on-premises access via AWS Direct Connect or AWS VPN, and access from multiple VPCs, accounts, and regions using VPC Peering or AWS Transit Gateway.
- Amazon FSx automatically encrypts your data at-rest and in-transit.
- Assessed to comply with ISO, PCI-DSS, and SOC certifications, and is HIPAA eligible.
- Integration with AWS CloudTrail monitors and logs your API calls letting you see actions taken by users on Amazon FSx resources.
- Pay only for the resources you use, with no minimum commitments or up-front fees.
- Can optimize costs by removing redundant data with Data Deduplication.
- User quotas provide tracking, monitoring, and enforcing of storage consumption to help reduce costs.

5.2 Amazon FSX For Lustre

- Amazon FSx for Lustre provides a high-performance file system optimized for fast processing of workloads such as machine learning, high performance computing (HPC), video processing, financial modeling, and electronic design automation (EDA).
- These workloads commonly require data to be presented via a fast and scalable file system interface, and typically have data sets stored on long-term data stores like Amazon S3.
- Amazon FSx for Lustre provides a fully managed high-performance Lustre file system that allows file-based applications to access data with hundreds of gigabytes per second of data, millions of IOPS, and sub millisecond latencies.
- Amazon FSx works natively with Amazon S3, letting you transparently access your S3 objects as files on Amazon FSx to run analyses for hours to months.
- You can then write results back to S3, and simply delete your file system. FSx for Lustre also enables you to burst your data processing workloads from on-premises to AWS, by allowing you to access your FSx file system over Amazon Direct Connect or VPN.
- You can also use FSx for Lustre as a standalone high-performance file system to burst your workloads from on-premises to the cloud.
- By copying on-premises data to an FSx for Lustre file system, you can make that data available for fast processing by compute instances running on AWS.
- With Amazon FSx, you pay for only the resources you use. There are no minimum commitments, upfront hardware or software costs, or additional fees.

Details and Benefits

- Lustre is a popular open-source parallel file system that is designed for high-performance workloads. These workloads include HPC, machine learning, analytics, and media processing.
- A parallel file system provides high throughput for processing large amounts of data and performs operations with consistently low latencies.
- It does so by storing data across multiple networked servers that thousands of compute instances can interact with concurrently.
- The Lustre file system provides a POSIX-compliant file system interface.
- Amazon FSx can scale up to hundreds of gigabytes per second of throughput, and millions of IOPS.
- Amazon FSx provides high throughput for processing large amounts of data and performs operations with consistent, sub-millisecond latencies.

- Amazon FSx for Lustre supports file access to thousands of EC2 instances, enabling you to provide file storage for your high-performance workloads, like genomics, seismic exploration, and video rendering.

Amazon S3:

- Amazon FSx works natively with Amazon S3, making it easy to access your S3 data to run data processing workloads.
- Your S3 objects are presented as files in your file system, and you can write your results back to S3.
- This lets you run data processing workloads on FSx for Lustre and store your long-term data on S3 or on-premises data stores.

On-premises:

- You can use Amazon FSx for Lustre for on-premises workloads that need to burst to the cloud due to peak demands or capacity limits.
- To move your existing on-premises data into Amazon FSx, you can mount your Amazon FSx for Lustre file system from an on-premises client over AWS Direct Connect or VPN, and then use parallel copy tools to import your data to your Amazon FSx for Lustre file system.
- At any time, you can write your results back to be durably stored in your data lake.

Security:

- All Amazon FSx file system data is encrypted at rest.
- You can access your file system from your compute instances using the open-source Lustre client.
- Once mounted, you can work with the files and directories in your file system just like you would with a local file system.
- FSx for Lustre is compatible with the most popular Linux-based AMIs, including Amazon Linux, Red Hat Enterprise Linux (RHEL), CentOS, Ubuntu, and SUSE Linux.
- You access your Amazon FSx file system from endpoints in your Amazon VPC, which enables you to isolate your file system in your own virtual network.
- You can configure security group rules and control network access to your Amazon FSx file systems.
- Amazon FSx is integrated with AWS Identity and Access Management (IAM).

- This integration means that you can control the actions your AWS IAM users and groups can take to manage your file systems (such as creating and deleting file systems).
- You can also tag your Amazon FSx resources and control the actions that your IAM users and groups can take based on those tags.

5.3 Sample Questions

Q1 : A large quantity of data is stored on a NAS device on-premises and accessed using the SMB protocol. The company require a managed service for hosting the filesystem and a tool to automate the migration.

Which actions should a Solutions Architect take?

A. Migrate the data to Amazon FSx for Windows File Server using AWS DataSync
B. Migrate the data to Amazon S3 using and AWS Snowball Edge device
C. Migrate the data to Amazon FSx for Lustre using AWS DataSync
D. Migrate the data to Amazon EFS using the AWS Server Migration Service (SMS)

Answer: A

Explanation : Amazon FSx for Windows File Server provides fully managed, highly reliable, and scalable file storage that is accessible over the industry-standard Server Message Block (SMB) protocol. This is the most suitable destination for this use case.

Q2 : A company is migrating from an on-premises infrastructure to the AWS Cloud. One of the company's applications stores files on a Windows file server farm that uses Distributed File System Replication (DFSR) to keep data in sync. A solutions architect needs to replace the file server farm.

Which service should the solutions architect use?

A. AWS Storage Gateway
B. Amazon S3
C. Amazon FSx
D. Amazon EFS

Answer : C

Explanation: Amazon FSx for Windows File Server provides fully managed, highly reliable file storage that is accessible over the industry-standard Server Message Block (SMB) protocol.

Amazon FSx is built on Windows Server and provides a rich set of administrative features that include end-user file restore, user quotas, and Access Control Lists (ACLs).

For more Questions Please check Certification Sample Quiz under each module

Link: https://k21academy.com/awssaquizm05