



# Amazon

## SAA-C03

**AWS Certified Solutions Architect - Associate**  
**QUESTION & ANSWERS**

## QUESTION 1

A Solutions Architect identified a series of DDoS attacks while monitoring the VPC. The Architect needs to fortify the current cloud infrastructure to protect the data of the clients.

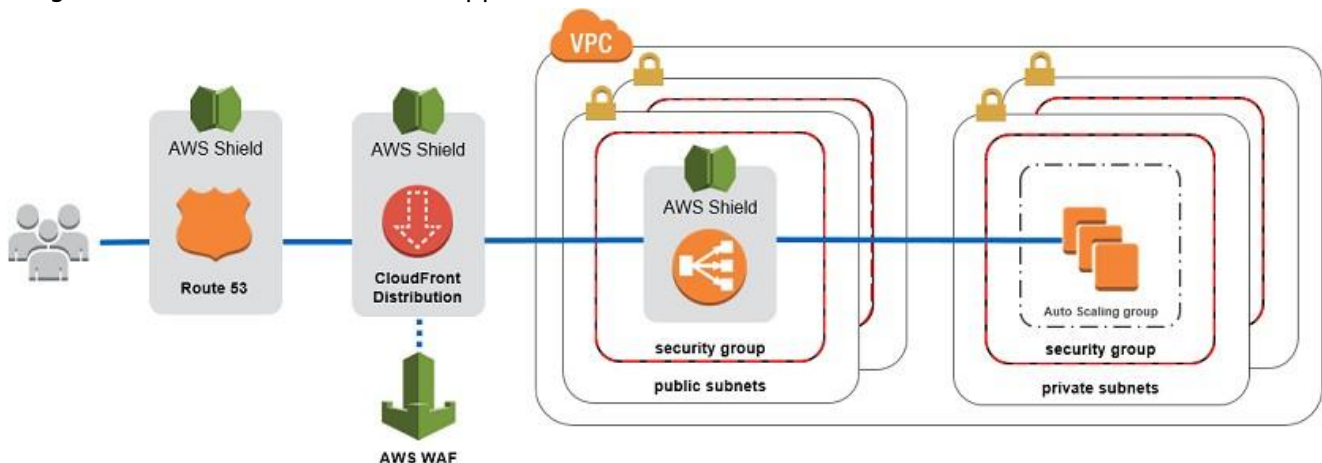
Which of the following is the most suitable solution to mitigate these kinds of attacks?

- A. Using the AWS Firewall Manager, set up a security layer that will prevent SYN floods, UDP reflection attacks, and other DDoS attacks.
- B. Set up a web application firewall using AWS WAF to filter, monitor, and block HTTP traffic.
- C. A combination of Security Groups and Network Access Control Lists to only allow authorized traffic to access your VPC.
- D. Use AWS Shield Advanced to detect and mitigate DDoS attacks.

**Correct Answer: D**

### Explanation/Reference:

For higher levels of protection against attacks targeting your applications running on Amazon Elastic Compute Cloud (EC2), Elastic Load Balancing(ELB), Amazon CloudFront, and Amazon Route 53 resources, you can subscribe to AWS Shield Advanced. In addition to the network and transport layer protections that come with Standard, AWS Shield Advanced provides additional detection and mitigation against large and sophisticated DDoS attacks, near real-time visibility into attacks, and integration with AWS WAF, a web application firewall.



AWS Shield Advanced also gives you 24x7 access to the AWS DDoS Response Team (DRT) and protection against DDoS related spikes in your Amazon Elastic Compute Cloud (EC2), Elastic Load Balancing(ELB), Amazon CloudFront, and Amazon Route 53 charges.

Hence, the correct answer is: Use AWS Shield Advanced to detect and mitigate DDoS attacks.

The option that says: Using the AWS Firewall Manager, set up a security layer that will prevent SYN floods, UDP reflection attacks and other DDoS attacks is incorrect because AWS Firewall Manager is mainly used to simplify your AWS WAF administration and maintenance tasks across multiple accounts and resources. It does not protect your VPC against DDoS attacks.

The option that says: Set up a web application firewall using AWS WAF to filter, monitor, and block HTTP traffic is incorrect. Even though AWS WAF can help you block common attack patterns to your VPC such as SQL injection or cross-site scripting, this is still not enough to withstand DDoS attacks. It is better to use AWS Shield in this scenario.

The option that says: A combination of Security Groups and Network Access Control Lists to only allow authorized traffic to access your VPC is incorrect. Although using a combination of Security Groups and NACLs are valid to provide security to your VPC, this is not enough to mitigate a DDoS attack. You should use AWS Shield for better security protection.

References:

[https://d1.awsstatic.com/whitepapers/Security/DDoS\\_White\\_Paper.pdf](https://d1.awsstatic.com/whitepapers/Security/DDoS_White_Paper.pdf)

<https://aws.amazon.com/shield/>

Check out this AWS Shield Cheat Sheet:

<https://tutorialsdojo.com/aws-shield/>

AWS Security Services Overview - WAF, Shield, CloudHSM, KMS:

<https://youtu.be/-1S-RdeAmMo>

## QUESTION 2

An online cryptocurrency exchange platform is hosted in AWS which uses ECS Cluster and RDS in Multi-AZ Deployments configuration. The application is heavily using the RDS instance to process complex read and write database operations. To maintain the reliability, availability, and performance of your systems, you have to closely monitor how the different processes or threads on a DB instance use the CPU, including the percentage of the CPU bandwidth and total memory consumed by each process.

Which of the following is the most suitable solution to properly monitor your database?

- A. Create a script that collects and publishes custom metrics to CloudWatch, which tracks the real-time CPU Utilization of the RDS instance, and then set up a custom CloudWatch dashboard to view the metrics.
- B. Check the CPU% and MEM% metrics which are readily available in the Amazon RDS console that shows the percentage of the CPU bandwidth and total memory consumed by each database process of your RDS instance.
- C. Use Amazon CloudWatch to monitor the CPU Utilization of your database.
- D. Enable Enhanced Monitoring in RDS.

## Correct Answer: D


### Explanation/Reference:

Amazon RDS provides metrics in real time for the operating system (OS) that your DB instance runs on. You can view the metrics for your DB instance using the console, or consume the Enhanced Monitoring JSON output from CloudWatch Logs in a monitoring system of your choice. By default, Enhanced Monitoring metrics are stored in the CloudWatch Logs for 30 days. To modify the amount of time the metrics are stored in the CloudWatch Logs, change the retention for the RDSOSMetrics log group in the CloudWatch console.

Take note that there are certain differences between CloudWatch and Enhanced Monitoring Metrics. CloudWatch gathers metrics about CPU utilization from the hypervisor for a DB instance, and Enhanced Monitoring gathers its metrics from an agent on the instance. As a result, you might find differences between the measurements, because the hypervisor layer performs a small amount of work. Hence, enabling Enhanced Monitoring in RDS is the correct answer in this specific scenario.

The differences can be greater if your DB instances use smaller instance classes, because then there are likely more virtual machines (VMs) that are managed by the hypervisor layer on a single physical

instance. Enhanced Monitoring metrics are useful when you want to see how different processes or threads on a DB instance use the CPU.

Process List					
<input type="text" value="Filter process list"/>					
<div>&lt; 1 2 &gt; </div>					
NAME ▼	VIRT ▼	RES ▼	CPU% ▼	MEM% ▼	VMLIMIT ▼
▼ postgres [3181]†	283.55 MB	17.11 MB	0.02	1.72	
postgres: rdsadmin rdsadmin localhost(40156) idle [2953]†	384.7 MB	9.51 MB	0.02	0.95	

Using Amazon CloudWatch to monitor the CPU Utilization of your database is incorrect. Although you can use this to monitor the CPU Utilization of your database instance, it does not provide the percentage of the CPU bandwidth and total memory consumed by each database process in your RDS instance. Take note that CloudWatch gathers metrics about CPU utilization from the hypervisor for a DB instance while RDS Enhanced Monitoring gathers its metrics from an agent on the instance. The option that says: Create a script that collects and publishes custom metrics to CloudWatch, which tracks the real-time CPU Utilization of the RDS instance and then set up a custom CloudWatch dashboard to view the metrics is incorrect. Although you can use Amazon CloudWatch Logs and CloudWatch dashboard to monitor the CPU Utilization of the database instance, using CloudWatch alone is still not enough to get the specific percentage of the CPU bandwidth and total memory consumed by each database processes. The data provided by CloudWatch is not as detailed as compared with the Enhanced Monitoring feature in RDS. Take note as well that you do not have direct access to the instances/servers of your RDS database instance, unlike with your EC2 instances where you can install a CloudWatch agent or a custom script to get CPU and memory utilization of your instance.

The option that says: Check the CPU% and MEM% metrics which are readily available in the Amazon RDS console that shows the percentage of the CPU bandwidth and total memory consumed by each database process of your RDS instance is incorrect because the CPU% and MEM% metrics are not readily available in the Amazon RDS console, which is contrary to what is being stated in this option.

References:

[https://docs.aws.amazon.com/AmazonRDS/latest/UserGuide/USER\\_Monitoring.OS.html#USER\\_Monitoring.OS.CloudWatchLogs](https://docs.aws.amazon.com/AmazonRDS/latest/UserGuide/USER_Monitoring.OS.html#USER_Monitoring.OS.CloudWatchLogs)

<https://docs.aws.amazon.com/AmazonRDS/latest/UserGuide/MonitoringOverview.html#monitoring-cloudwatch>

Check out this Amazon CloudWatch Cheat Sheet:

<https://tutorialsdojo.com/amazon-cloudwatch/>

Check out this Amazon RDS Cheat Sheet:

<https://tutorialsdojo.com/amazon-relational-database-service-amazon-rds/>

### QUESTION 3

A media company has two VPCs: VPC-1 and VPC-2 with peering connection between each other. VPC-1 only contains private subnets while VPC-2 only contains public subnets. The company uses a single AWS Direct Connect connection and a virtual interface to connect their on-premises network with VPC-1.

Which of the following options increase the fault tolerance of the connection to VPC-1? (Select TWO.)

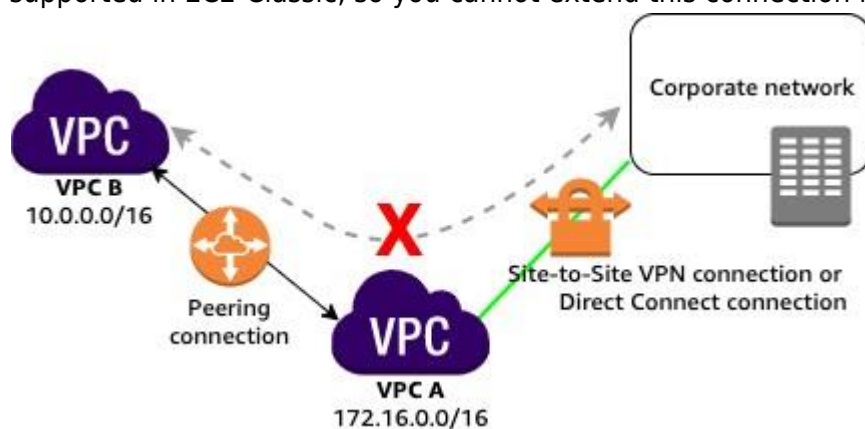
- A. Establish a new AWS Direct Connect connection and private virtual interface in the same region as VPC-2.
- B. Establish another AWS Direct Connect connection and private virtual interface in the same AWS region as VPC-1.
- C. Establish a hardware VPN over the Internet between VPC-2 and the on-premises network.
- D. Use the AWS VPN CloudHub to create a new AWS Direct Connect connection and private virtual interface in the same region as VPC-2.
- E. Establish a hardware VPN over the Internet between VPC-1 and the on-premises network.

**Correct Answer: B,E**

#### Explanation/Reference:

In this scenario, you have two VPCs which have peering connections with each other. Note that a VPC peering connection does not support edge to edge routing. This means that if either VPC in a peering relationship has one of the following connections, you cannot extend the peering relationship to that connection:

- A VPN connection or an AWS Direct Connect connection to a corporate network
- An Internet connection through an Internet gateway
- An Internet connection in a private subnet through a NAT device
- A gateway VPC endpoint to an AWS service; for example, an endpoint to Amazon S3.
- (IPv6) A ClassicLink connection. You can enable IPv4 communication between a linked EC2-Classical instance and instances in a VPC on the other side of a VPC peering connection. However, IPv6 is not supported in EC2-Classical, so you cannot extend this connection for IPv6 communication.



For example, if VPC A and VPC B are peered, and VPC A has any of these connections, then instances in VPC B cannot use the connection to access resources on the other side of the connection. Similarly, resources on the other side of a connection cannot use the connection to access VPC B.

Hence, this means that you cannot use VPC-2 to extend the peering relationship that exists between

VPC-1 and the on-premises network. For example, traffic from the corporate network can't directly access VPC-1 by using the VPN connection or the AWS Direct Connect connection to VPC-2, which is why the following options are incorrect:

- Use the AWS VPN CloudHub to create a new AWS Direct Connect connection and private virtual interface in the same region as VPC-2.
- Establish a hardware VPN over the Internet between VPC-2 and the on-premises network.
- Establish a new AWS Direct Connect connection and private virtual interface in the same region as VPC-2.

You can do the following to provide a highly available, fault-tolerant network connection:

- Establish a hardware VPN over the Internet between the VPC and the on-premises network.
- Establish another AWS Direct Connect connection and private virtual interface in the same AWS region.

References:

<https://docs.aws.amazon.com/vpc/latest/peering/invalid-peering-configurations.html#edge-to-edge-vgw>

<https://aws.amazon.com/premiumsupport/knowledge-center/configure-vpn-backup-dx/>

<https://aws.amazon.com/answers/networking/aws-multiple-data-center-ha-network-connectivity/>

Check out this Amazon VPC Cheat Sheet:

<https://tutorialsdojo.com/amazon-vpc/>

## QUESTION 4

A company is hosting EC2 instances that are on non-production environment and processing non-priority batch loads, which can be interrupted at any time.

What is the best instance purchasing option which can be applied to your EC2 instances in this case?

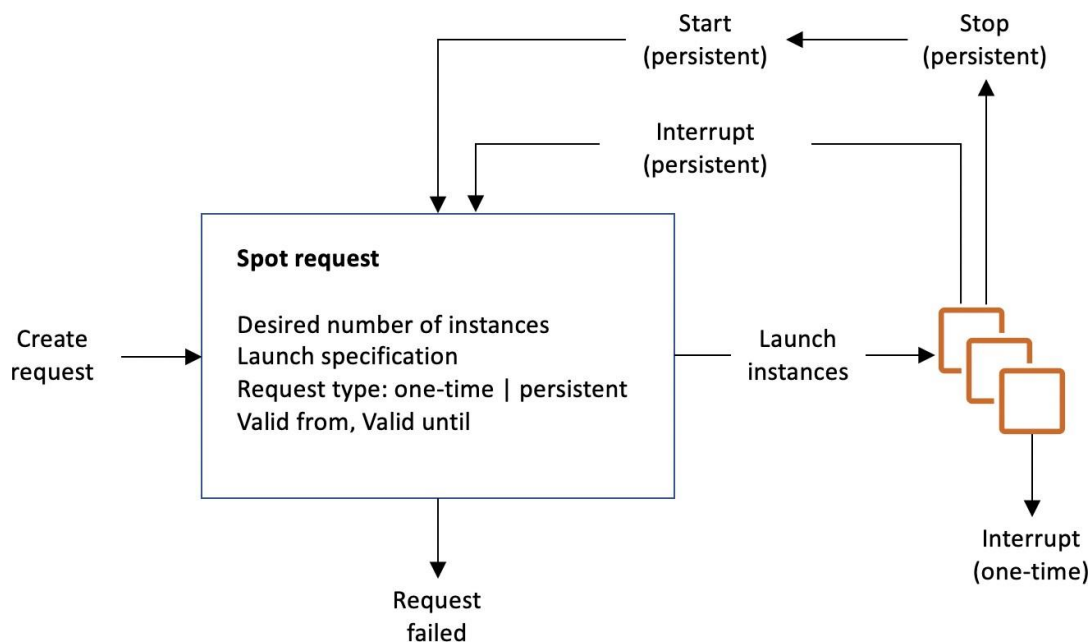
- A. Spot Instances
- B. Reserved Instances
- C. On-Demand Instances
- D. On-Demand Capacity Reservations

## Correct Answer: A

### Explanation/Reference:

Amazon EC2 Spot instances are spare compute capacity in the AWS cloud available to you at steep discounts compared to On-Demand prices. It can be interrupted by AWS EC2 with two minutes of notification when the EC2 needs the capacity back.

To use Spot Instances, you create a Spot Instance request that includes the number of instances, the instance type, the Availability Zone, and the maximum price that you are willing to pay per instance hour. If your maximum price exceeds the current Spot price, Amazon EC2 fulfills your request immediately if capacity is available. Otherwise, Amazon EC2 waits until your request can be fulfilled or until you cancel the request.



#### References:

<http://docs.aws.amazon.com/AWSEC2/latest/UserGuide/using-spot-instances.html>

<https://aws.amazon.com/ec2/spot/>

Amazon EC2 Overview:

[https://youtu.be/7VsGIHT\\_jQE](https://youtu.be/7VsGIHT_jQE)

Check out this Amazon EC2 Cheat Sheet:

<https://tutorialsdojo.com/amazon-elastic-compute-cloud-amazon-ec2/>

## QUESTION 5

Due to the large volume of query requests, the database performance of an online reporting application significantly slowed down. The Solutions Architect is trying to convince her client to use Amazon RDS Read Replica for their application instead of setting up a Multi-AZ Deployments configuration.

What are two benefits of using Read Replicas over Multi-AZ that the Architect should point out? (Select TWO.)

- A. Provides asynchronous replication and improves the performance of the primary database by taking read-heavy database workloads from it.
- B. It enhances the read performance of your primary database by increasing its IOPS and accelerates its query processing via AWS Global Accelerator.
- C. It elastically scales out beyond the capacity constraints of a single DB instance for read-heavy database workloads.
- D. Provides synchronous replication and automatic failover in the case of Availability Zone service failures.
- E. Allows both read and write operations on the read replica to complement the primary database.

**Correct Answer: A,C**



## Explanation/Reference:

Amazon RDS Read Replicas provide enhanced performance and durability for database (DB) instances. This feature makes it easy to elastically scale out beyond the capacity constraints of a single DB instance for read-heavy database workloads.

You can create one or more replicas of a given source DB Instance and serve high-volume application read traffic from multiple copies of your data, thereby increasing aggregate read throughput. Read replicas can also be promoted when needed to become standalone DB instances.

For the MySQL, MariaDB, PostgreSQL, and Oracle database engines, Amazon RDS creates a second DB instance using a snapshot of the source DB instance. It then uses the engines' native asynchronous replication to update the read replica whenever there is a change to the source DB instance. The read replica operates as a DB instance that allows only read-only connections; applications can connect to a read replica just as they would to any DB instance. Amazon RDS replicates all databases in the source DB instance.

Multi-AZ deployments	Multi-Region deployments	Read replicas
Main purpose is high availability	Main purpose is disaster recovery and local performance	Main purpose is scalability
Non-Aurora: synchronous replication; Aurora: asynchronous replication	Asynchronous replication	Asynchronous replication
Non-Aurora: only the primary instance is active; Aurora: all instances are active	All regions are accessible and can be used for reads	All read replicas are accessible and can be used for readscaling
Non-Aurora: automated backups are taken from standby; Aurora: automated backups are taken from shared storage layer	Automated backups can be taken in each region	No backups configured by default
Always span at least two Availability Zones within a single region	Each region can have a Multi-AZ deployment	Can be within an Availability Zone, Cross-AZ, or Cross-Region
Non-Aurora: database engine version upgrades happen on primary; Aurora: all instances are updated together	Non-Aurora: database engine version upgrade is independent in each region; Aurora: all instances are updated together	Non-Aurora: database engine version upgrade is independent from source instance; Aurora: all instances are updated together
Automatic failover to standby (non-Aurora) or read replica (Aurora) when a problem is detected	Aurora allows promotion of a secondary region to be the master	Can be manually promoted to a standalone database instance (non-Aurora) or to be the primary instance (Aurora)

When you create a read replica for Amazon RDS for MySQL, MariaDB, PostgreSQL, and Oracle, Amazon RDS sets up a secure communications channel using public-key encryption between the source DB instance and the read replica, even when replicating across regions. Amazon RDS establishes any AWS security configurations such as adding security group entries needed to enable the secure channel.

You can also create read replicas within a Region or between Regions for your Amazon RDS for MySQL, MariaDB, PostgreSQL, and Oracle database instances encrypted at rest with AWS Key Management Service (KMS).

Hence, the correct answers are:

- It elastically scales out beyond the capacity constraints of a single DB instance for read-heavy database workloads.
- Provides asynchronous replication and improves the performance of the primary database by taking read-heavy database workloads from it.

The option that says: Allows both read and write operations on the read replica to complement the primary database is incorrect as Read Replicas are primarily used to offload read-only operations from the primary database instance. By default, you can't do a write operation to your Read Replica.



The option that says: Provides synchronous replication and automatic failover in the case of Availability Zone service failures is incorrect as this is a benefit of Multi-AZ and not of a Read Replica. Moreover, Read Replicas provide an asynchronous type of replication and not synchronous replication.

The option that says: It enhances the read performance of your primary database by increasing its IOPS and accelerates its query processing via AWS Global Accelerator is incorrect because Read Replicas do not do anything to upgrade or increase the read throughput on the primary DB instance per se, but it provides a way for your application to fetch data from replicas. In this way, it improves the overall performance of your entire database-tier (and not just the primary DB instance). It doesn't increase the IOPS nor use AWS Global Accelerator to accelerate the compute capacity of your primary database. AWS Global Accelerator is a networking service, not related to RDS, that directs user traffic to the nearest application endpoint to the client, thus reducing internet latency and jitter. It simply routes the traffic to the closest edge location via Anycast.

References:

<https://aws.amazon.com/rds/details/read-replicas/>

<https://aws.amazon.com/rds/features/multi-az/>

Check out this Amazon RDS Cheat Sheet:

<https://tutorialsdojo.com/amazon-relational-database-service-amazon-rds/>

Additional tutorial - How do I make my RDS MySQL read replica writable?

<https://youtu.be/j5da6d2TIPc>

## QUESTION 6

A media company needs to configure an Amazon S3 bucket to serve static assets for the public-facing web application. Which methods ensure that all of the objects uploaded to the S3 bucket can be read publicly all over the Internet? (Select TWO.)

- A. Configure the cross-origin resource sharing (CORS) of the S3 bucket to allow objects to be publicly accessible from all domains.
- B. Grant public read access to the object when uploading it using the S3 Console.
- C. Create an IAM role to set the objects inside the S3 bucket to public read.
- D. Configure the S3 bucket policy to set all objects to public read.
- E. Do nothing. Amazon S3 objects are already public by default.

**Correct Answer: B,D**

## Explanation/Reference:

By default, all Amazon S3 resources such as buckets, objects, and related subresources are private which means that only the AWS account holder (resource owner) that created it has access to the resource. The resource owner can optionally grant access permissions to others by writing an access policy. In S3, you also set the permissions of the object during upload to make it public.

Amazon S3 offers access policy options broadly categorized as resource-based policies and user policies. Access policies you attach to your resources (buckets and objects) are referred to as resource-based policies.

For example, bucket policies and access control lists (ACLs) are resource-based policies. You can also attach access policies to users in your account. These are called user policies. You may choose to use resource-based policies, user policies, or some combination of these to manage permissions to your

Amazon S3 resources.

You can also manage the public permissions of your objects during upload. Under Manage public permissions, you can grant read access to your objects to the general public (everyone in the world), for all of the files that you're uploading. Granting public read access is applicable to a small subset of use cases such as when buckets are used for websites.

The screenshot shows the Amazon S3 Upload console interface. At the top, there's a blue header with the word 'Upload' and a close button. Below the header, a progress bar shows four steps: 1. Select files, 2. Set permissions (current step), 3. Set properties, and 4. Review. Below the progress bar, a summary bar shows '1 Files', 'Size: 263.3 KB', and 'Target path: 1-baket'. The main content area is divided into sections: 'Manage users' with a table of users and permissions, 'Access for other AWS account' with an 'Add account' button, and 'Manage public permissions'. The 'Manage public permissions' section is highlighted with a green border and contains a dropdown menu with three options: 'Do not grant public read access to this object(s) (Recommended)' (selected), 'Do not grant public read access to this object(s) (Recommended)', and 'Grant public read access to this object(s)'. At the bottom, there are 'Upload', 'Previous', and 'Next' buttons.

Hence, the correct answers are:

- Grant public read access to the object when uploading it using the S3 Console.
- Configure the S3 bucket policy to set all objects to public read.

The option that says: Configure the cross-origin resource sharing (CORS) of the S3 bucket to allow objects to be publicly accessible from all domains is incorrect. CORS will only allow objects from one domain (travel.cebu.com) to be loaded and accessible to a different domain (palawan.com). It won't necessarily expose objects for public access all over the internet.

The option that says: Creating an IAM role to set the objects inside the S3 bucket to public read is incorrect. You can create an IAM role and attach it to an EC2 instance in order to retrieve objects from the S3 bucket or add new ones. An IAM Role, in itself, cannot directly make the S3 objects public or change the permissions of each individual object.

The option that says: Do nothing. Amazon S3 objects are already public by default is incorrect because, by default, all the S3 resources are private, so only the AWS account that created the resources can access them.

References:

<http://docs.aws.amazon.com/AmazonS3/latest/dev/s3-access-control.html>

<https://docs.aws.amazon.com/AmazonS3/latest/dev/BucketRestrictions.html>

Check out this Amazon S3 Cheat Sheet:

<https://tutorialsdojo.com/amazon-s3/>

## QUESTION 7

A company plans to deploy an application in an Amazon EC2 instance. The application will perform the following tasks:

- Read large datasets from an Amazon S3 bucket.
- Execute multi-stage analysis on the datasets.
- Save the results to Amazon RDS.

During multi-stage analysis, the application will store a large number of temporary files in the instance storage. As the Solutions Architect, you need to recommend the fastest storage option with high I/O performance for the temporary files.

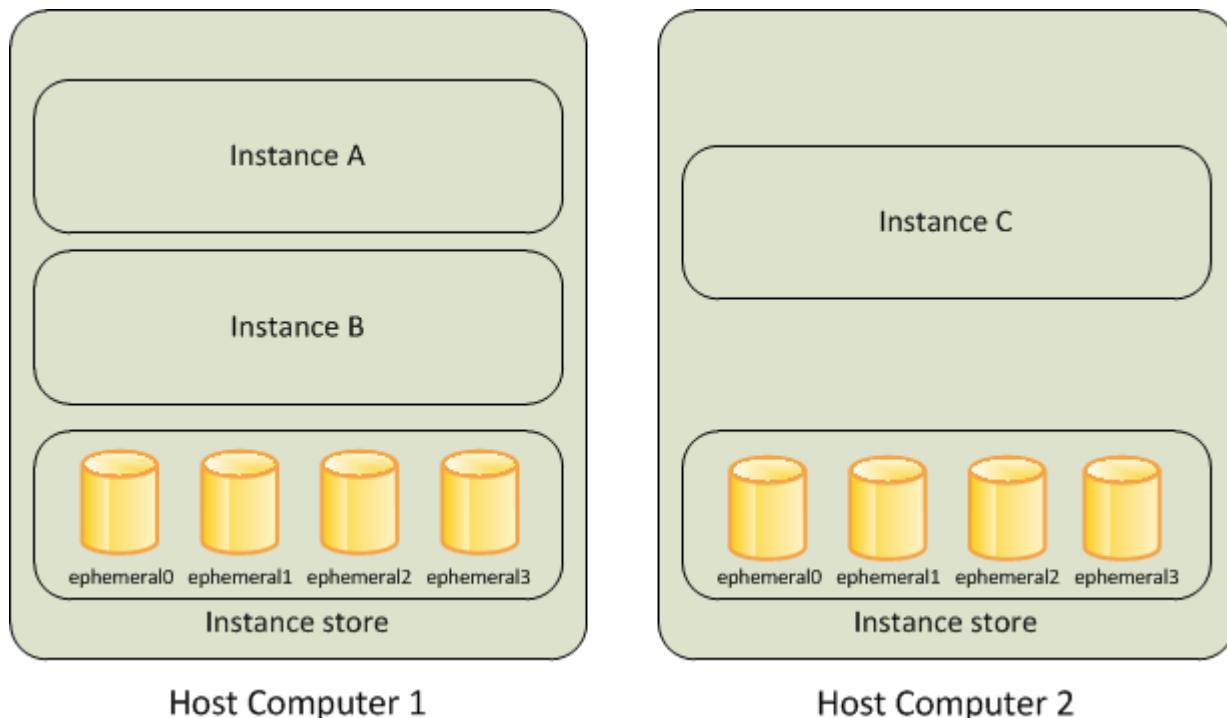
Which of the following options fulfills this requirement?

- A. Configure RAID 1 in multiple instance store volumes.
- B. Attach multiple Provisioned IOPS SSD volumes in the instance.
- C. Configure RAID 0 in multiple instance store volumes.
- D. Enable Transfer Acceleration in Amazon S3.

**Correct Answer: C**

### Explanation/Reference:

Amazon Elastic Compute Cloud (Amazon EC2) provides scalable computing capacity in the Amazon Web Services (AWS) Cloud. You can use Amazon EC2 to launch as many or as few virtual servers as you need, configure security and networking, and manage storage. Amazon EC2 enables you to scale up or down to handle changes in requirements or spikes in popularity, reducing your need to forecast traffic.



RAID 0 configuration enables you to improve your storage volumes' performance by distributing the I/O across the volumes in a stripe. Therefore, if you add a storage volume, you get the straight addition of throughput and IOPS. This configuration can be implemented on both EBS or instance

store volumes. Since the main requirement in the scenario is storage performance, you need to use an instance store volume. It uses NVMe or SATA-based SSD to deliver high random I/O performance. This type of storage is a good option when you need storage with very low latency, and you don't need the data to persist when the instance terminates.

Hence, the correct answer is: Configure RAID 0 in multiple instance store volumes.

The option that says: Enable Transfer Acceleration in Amazon S3 is incorrect because S3 Transfer Acceleration is mainly used to speed up the transfer of gigabytes or terabytes of data between clients and an S3 bucket.

The option that says: Configure RAID 1 in multiple instance volumes is incorrect because RAID 1 configuration is used for data mirroring. You need to configure RAID 0 to improve the performance of your storage volumes.

The option that says: Attach multiple Provisioned IOPS SSD volumes in the instance is incorrect because persistent storage is not needed in the scenario. Also, instance store volumes have greater I/O performance than EBS volumes.

References:

<https://docs.aws.amazon.com/AWSEC2/latest/UserGuide/InstanceStorage.html>

<https://docs.aws.amazon.com/AWSEC2/latest/UserGuide/raid-config.html>

Check out this Amazon EC2 Cheat Sheet:

<https://tutorialsdojo.com/amazon-elastic-compute-cloud-amazon-ec2/>

## QUESTION 8

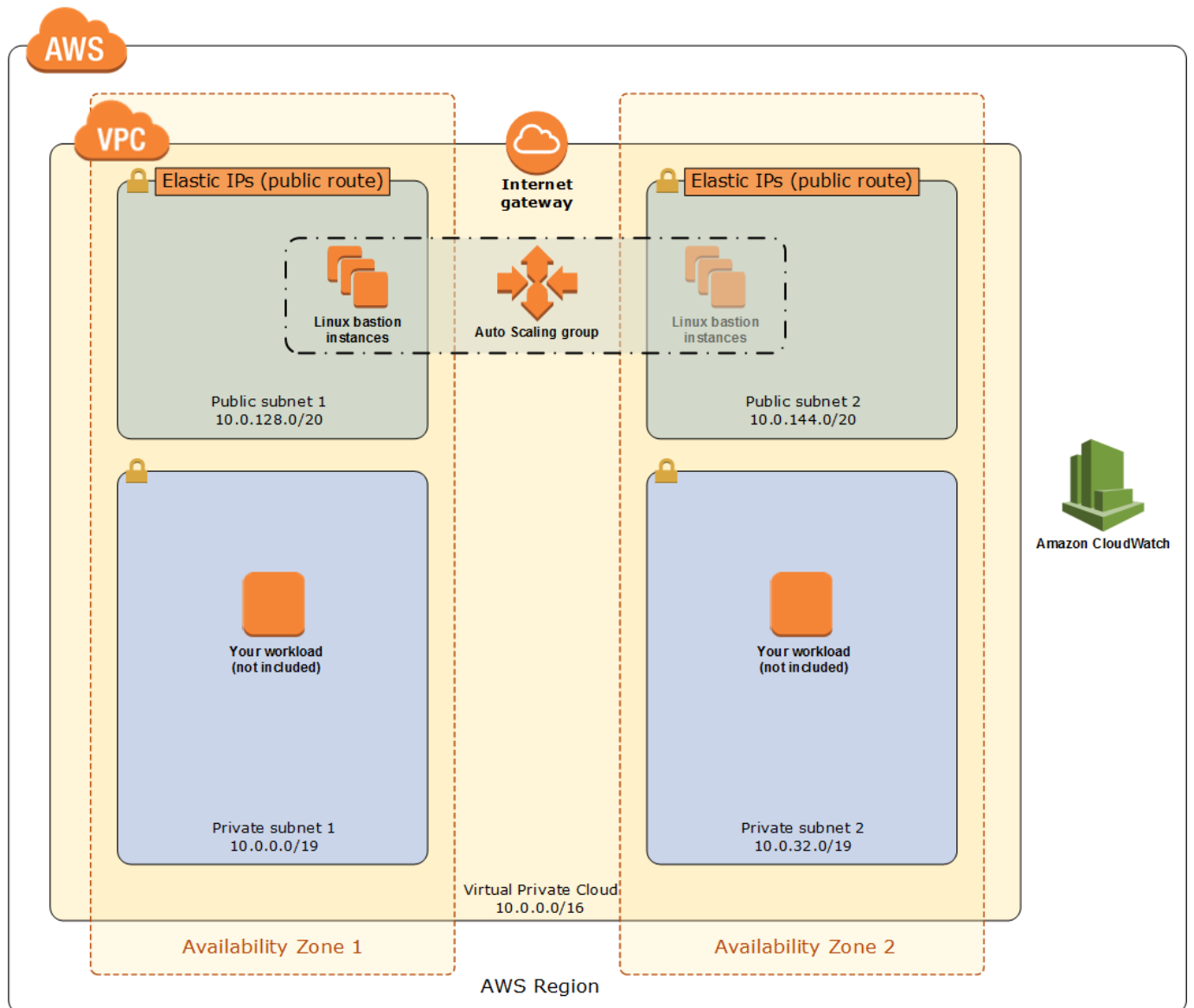
A Solutions Architect needs to set up a bastion host in Amazon VPC. It should only be accessed from the corporate data center via SSH. What is the best way to achieve this?

- A. Create a large EC2 instance with a security group which only allows access on port 22 via the IP address of the corporate data center. Use a private key (.pem) file to connect to the bastion host.
- B. Create a small EC2 instance with a security group which only allows access on port 22 using your own pre-configured password.
- C. Create a small EC2 instance with a security group which only allows access on port 22 via the IP address of the corporate data center. Use a private key (.pem) file to connect to the bastion host.
- D. Create a large EC2 instance with a security group which only allows access on port 22 using your own pre-configured password.

## Correct Answer: C

### Explanation/Reference:

The best way to implement a bastion host is to create a small EC2 instance which should only have a security group from a particular IP address for maximum security. This will block any SSH Brute Force attacks on your bastion host. It is also recommended to use a small instance rather than a large one because this host will only act as a jump server to connect to other instances in your VPC and nothing else.



Therefore, there is no point of allocating a large instance simply because it doesn't need that much computing power to process SSH (port 22) or RDP (port 3389) connections. It is possible to use SSH with an ordinary user ID and a pre-configured password as credentials but it is more secure to use public key pairs for SSH authentication for better security.

Hence, the right answer for this scenario is the option that says: Create a small EC2 instance with a security group which only allows access on port 22 via the IP address of the corporate data center.

Use a private key (.pem) file to connect to the bastion host.

Creating a large EC2 instance with a security group which only allows access on port 22 using your own pre-configured password and creating a small EC2 instance with a security group which only allows access on port 22 using your own pre-configured password are incorrect. Even though you have your own pre-configured password, the SSH connection can still be accessed by anyone over the Internet, which poses as a security vulnerability.

The option that says: Create a large EC2 instance with a security group which only allows access on port 22 via the IP address of the corporate data center. Use a private key (.pem) file to connect to the bastion host is incorrect because you don't need a large instance for a bastion host as it does not require much CPU resources.

References:

<https://docs.aws.amazon.com/quickstart/latest/linux-bastion/architecture.html>

<https://aws.amazon.com/blogs/security/how-to-record-ssh-sessions-established-through-a-bastion-host/>

Check out this Amazon VPC Cheat Sheet:  
<https://tutorialsdojo.com/amazon-vpc/>

## QUESTION 9

A Solutions Architect joined a large tech company with an existing Amazon VPC. When reviewing the Auto Scaling events, the Architect noticed that their web application is scaling up and down multiple times within the hour.

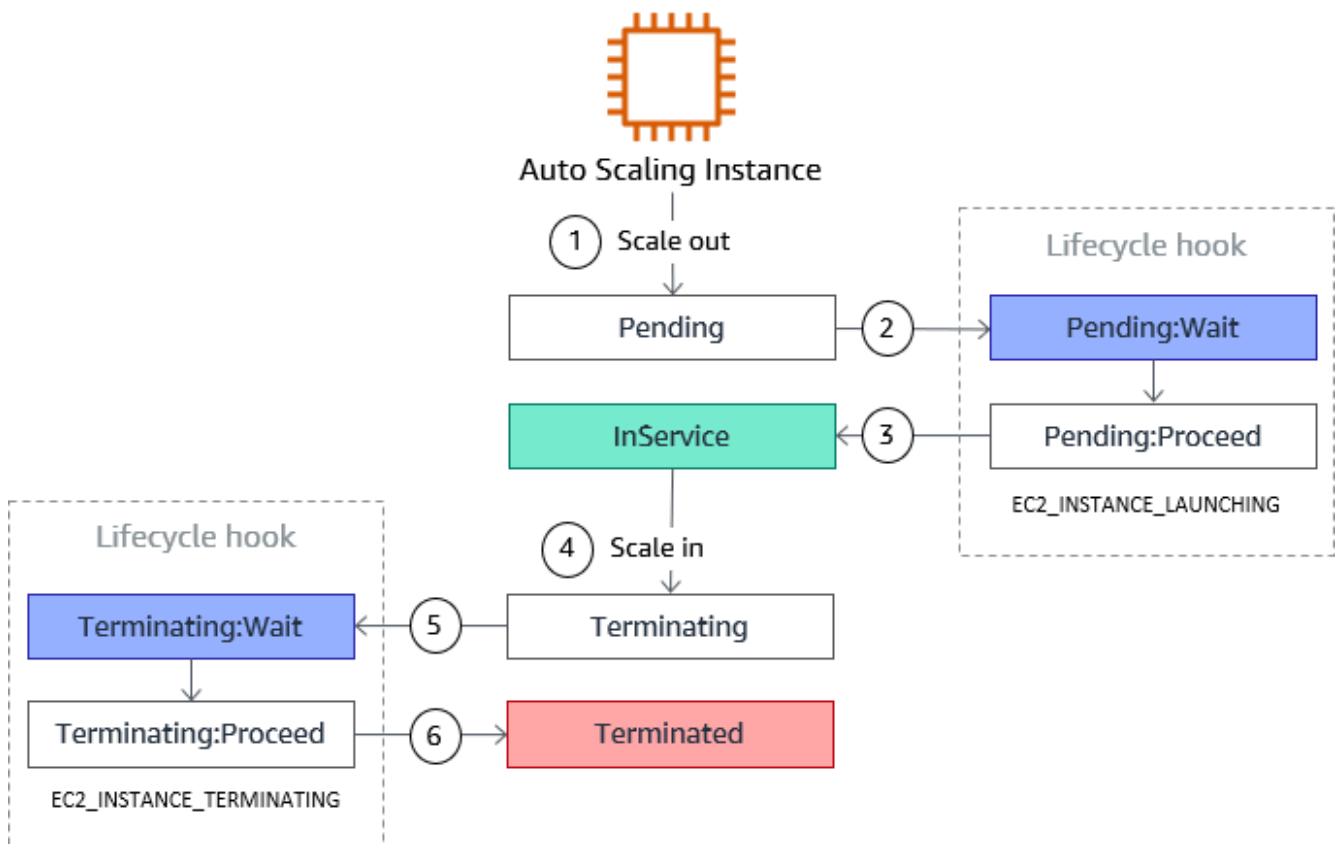
What design change could the Architect make to optimize cost while preserving elasticity?

- A. Increase the base number of Auto Scaling instances for the Auto Scaling group
- B. Change the cooldown period of the Auto Scaling group and set the CloudWatch metric to a higher threshold
- C. Add provisioned IOPS to the instances
- D. Increase the instance type in the launch configuration

## Correct Answer: B

### Explanation/Reference:

Since the application is scaling up and down multiple times within the hour, the issue lies on the cooldown period of the Auto Scaling group.



The cooldown period is a configurable setting for your Auto Scaling group that helps to ensure that it doesn't launch or terminate additional instances before the previous scaling activity takes effect. After the Auto Scaling group dynamically scales using a simple scaling policy, it waits for the

cooldown period to complete before resuming scaling activities.

When you manually scale your Auto Scaling group, the default is not to wait for the cooldown period, but you can override the default and honor the cooldown period. If an instance becomes unhealthy, the Auto Scaling group does not wait for the cooldown period to complete before replacing the unhealthy instance.

Reference:

<http://docs.aws.amazon.com/autoscaling/latest/userguide/as-scale-based-on-demand.html>

Check out this Amazon EC2 Cheat Sheet:

<https://tutorialsdojo.com/amazon-elastic-compute-cloud-amazon-ec2/>

## QUESTION 10

A Solutions Architect is working for a fast-growing startup that just started operations during the past 3 months. They currently have an on-premises Active Directory and 10 computers. To save costs in procuring physical workstations, they decided to deploy virtual desktops for their new employees in a virtual private cloud in AWS. The new cloud infrastructure should leverage the existing security controls in AWS but can still communicate with their on-premises network.

Which set of AWS services will the Architect use to meet these requirements?

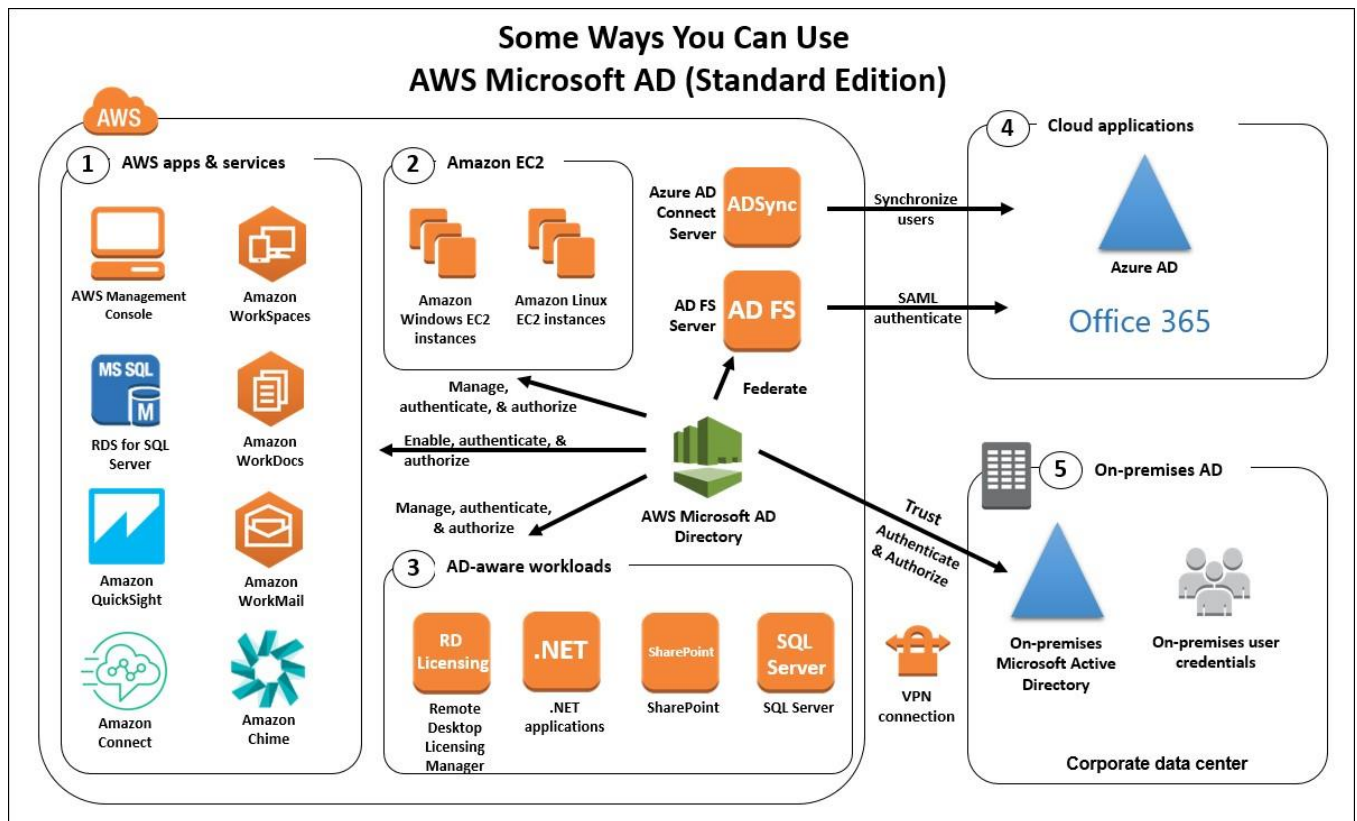
- A. AWS Directory Services, VPN connection, and Amazon Workspaces
- B. AWS Directory Services, VPN connection, and Amazon S3
- C. AWS Directory Services, VPN connection, and ClassicLink
- D. AWS Directory Services, VPN connection, and AWS Identity and Access Management

## Correct Answer: A

### Explanation/Reference:

For this scenario, the best answer is: AWS Directory Services, VPN connection, and Amazon Workspaces.





First, you need a VPN connection to connect the VPC and your on-premises network. Second, you need AWS Directory Services to integrate with your on-premises Active Directory and lastly, you need to use Amazon WorkSpace to create the needed virtual desktops in your VPC.

References:

<https://aws.amazon.com/directoryservice/>

<https://docs.aws.amazon.com/AmazonVPC/latest/UserGuide/vpn-connections.html>

<https://aws.amazon.com/workspaces/>

AWS Identity Services Overview:

<https://www.youtube.com/watch?v=AldUw0i8rr0>

Check out these cheat sheets on AWS Directory Service, Amazon VPC, and Amazon WorkSpaces:

<https://tutorialsdojo.com/aws-directory-service/>

<https://tutorialsdojo.com/amazon-vpc/>

## QUESTION 11

A company needs to set up a cost-effective architecture for a log processing application that has frequently accessed, throughput-intensive workloads with large, sequential I/O operations. The application should be hosted in an already existing On-Demand EC2 instance in the VPC. You have to attach a new EBS volume that will be used by the application.

Which of the following is the most suitable EBS volume type that you should use in this scenario?



- A. EBS Throughput Optimized HDD (st1)
- B. EBS Cold HDD (sc1)
- C. EBS General Purpose SSD (gp2)
- D. EBS Provisioned IOPS SSD (io1)

## Correct Answer: A

### Explanation/Reference:

In the exam, always consider the difference between SSD and HDD as shown on the table below. This will allow you to easily eliminate specific EBS-types in the options which are not SSD or not HDD, depending on whether the question asks for a storage type which has small, random I/O operations or large, sequential I/O operations.

Since the scenario has workloads with large, sequential I/O operations, we can narrow down our options by selecting HDD volumes, instead of SDD volumes which are more suitable for small, random I/O operations.

FEATURES	SSD Solid State Drive	HDD Hard Disk Drive
Best for workloads with:	<i>small, random</i> I/O operations	<i>large, sequential</i> I/O operations
Can be used as a bootable volume?	Yes	No
Suitable Use Cases	<ul style="list-style-type: none"><li>- Best for <b>transactional workloads</b></li><li>- Critical business applications that require sustained IOPS performance</li><li>- Large database workloads such as MongoDB, Oracle, Microsoft SQL Server and many others...</li></ul>	<ul style="list-style-type: none"><li>- Best for <i>large streaming workloads</i> requiring consistent, fast throughput at a low price</li><li>- Big data, Data warehouses, Log processing</li><li>- Throughput-oriented storage for large volumes of data that is <i>infrequently</i> accessed</li></ul>
Cost	moderate / high 	low 
Dominant Performance Attribute	IOPS	Throughput (MiB/s)

Throughput Optimized HDD (st1) volumes provide low-cost magnetic storage that defines performance in terms of throughput rather than IOPS. This volume type is a good fit for large, sequential workloads such as Amazon EMR, ETL, data warehouses, and log processing. Bootable st1 volumes are not supported.

Throughput Optimized HDD (st1) volumes, though similar to Cold HDD (sc1) volumes, are designed to support frequently accessed data.

EBS Provisioned IOPS SSD (io1) is incorrect because Amazon EBS Provisioned IOPS SSD is not the most cost-effective EBS type and is primarily used for critical business applications that require sustained IOPS performance.

EBS General Purpose SSD (gp2) is incorrect. Although an Amazon EBS General Purpose SSD volume balances price and performance for a wide variety of workloads, it is not suitable for frequently accessed, throughput-intensive workloads. Throughput Optimized HDD is a more suitable option to

use than General Purpose SSD.

EBS Cold HDD (sc1) is incorrect. Although this provides lower cost HDD volume compared to General Purpose SSD, it is much suitable for less frequently accessed workloads.

Reference:

[https://docs.aws.amazon.com/AWSEC2/latest/UserGuide/EBSVolumeTypes.html#EBSVolumeTypes\\_st1](https://docs.aws.amazon.com/AWSEC2/latest/UserGuide/EBSVolumeTypes.html#EBSVolumeTypes_st1)

Amazon EBS Overview - SSD vs HDD:

<https://www.youtube.com/watch?v=LW7x8wyLFvw&t=8s>

Check out this Amazon EBS Cheat Sheet:

<https://tutorialsdodojo.com/amazon-ebs/>

## QUESTION 12

A startup is building a microservices architecture in which the software is composed of small independent services that communicate over well-defined APIs. In building large-scale systems, fine-grained decoupling of microservices is a recommended practice to implement. The decoupled services should scale horizontally from each other to improve scalability.

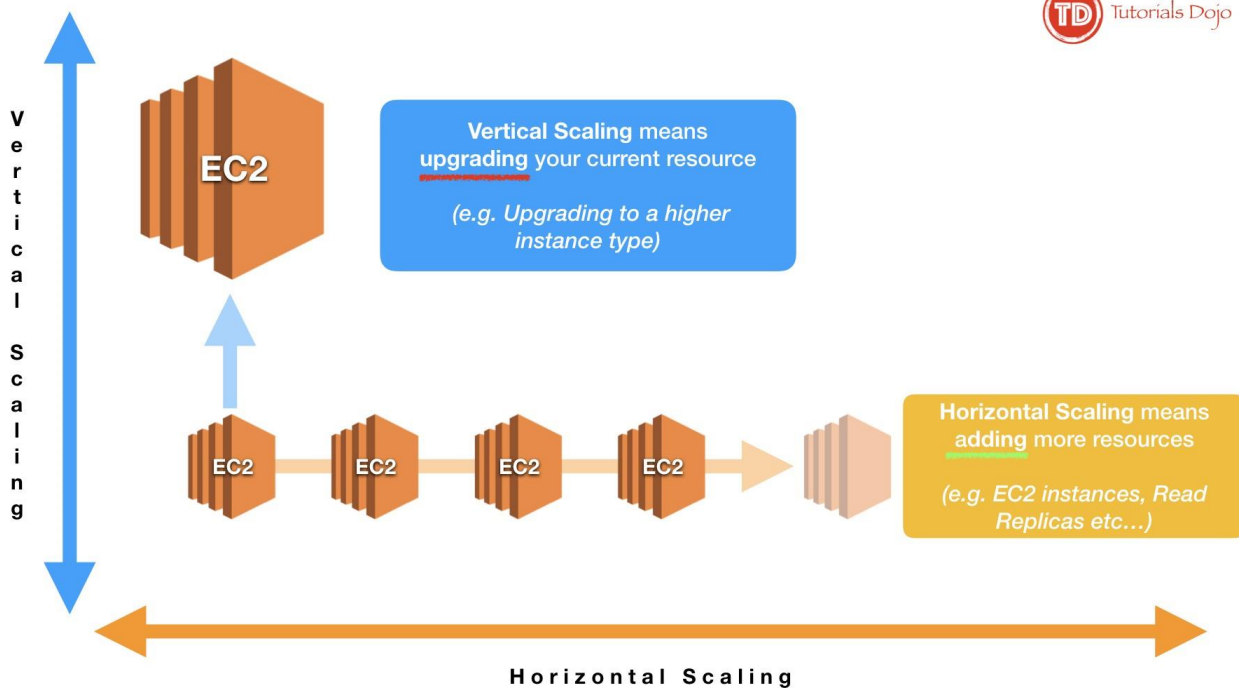
What is the difference between Horizontal scaling and Vertical scaling?

- A. Vertical scaling means running the same software on a fully serverless architecture using Lambda. Horizontal scaling means adding more servers to the existing pool and it doesn't run into limitations of individual servers.
- B. Horizontal scaling means running the same software on smaller containers such as Docker and Kubernetes using ECS or EKS. Vertical scaling is adding more servers to the existing pool and doesn't run into limitations of individual servers.
- C. Vertical scaling means running the same software on bigger machines which is limited by the capacity of the individual server. Horizontal scaling is adding more servers to the existing pool and doesn't run into limitations of individual servers.
- D. Horizontal scaling means running the same software on bigger machines which is limited by the capacity of individual servers. Vertical scaling is adding more servers to the existing pool and doesn't run into limitations of individual servers.

## Correct Answer: C

### Explanation/Reference:

Vertical scaling means running the same software on bigger machines which is limited by the capacity of the individual server. Horizontal scaling is adding more servers to the existing pool and doesn't run into limitations of individual servers.



Fine-grained decoupling of microservices is a best practice for building large-scale systems. It's a prerequisite for performance optimization since it allows choosing the appropriate and optimal technologies for a specific service. Each service can be implemented with the appropriate programming languages and frameworks, leverage the optimal data persistence solution, and be fine-tuned with the best performing service configurations.

Properly decoupled services can be scaled horizontally and independently from each other. Vertical scaling, which is running the same software on bigger machines, is limited by the capacity of individual servers and can incur downtime during the scaling process. Horizontal scaling, which is adding more servers to the existing pool, is highly dynamic and doesn't run into limitations of individual servers. The scaling process can be completely automated.

Furthermore, the resiliency of the application can be improved because failing components can be easily and automatically replaced. Hence, the correct answer is the option that says: Vertical scaling means running the same software on bigger machines which is limited by the capacity of the individual server. Horizontal scaling is adding more servers to the existing pool and doesn't run into limitations of individual servers.

The option that says: Vertical scaling means running the same software on a fully serverless architecture using Lambda. Horizontal scaling means adding more servers to the existing pool and it doesn't run into limitations of individual servers is incorrect because Vertical scaling is not about running the same software on a fully serverless architecture. AWS Lambda is not required for scaling. The option that says: Horizontal scaling means running the same software on bigger machines which is limited by the capacity of individual servers. Vertical scaling is adding more servers to the existing pool and doesn't run into limitations of individual servers is incorrect because the definitions for the two concepts were switched. Vertical scaling means running the same software on bigger machines which is limited by the capacity of the individual server. Horizontal scaling is adding more servers to the existing pool and doesn't run into limitations of individual servers.

The option that says: Horizontal scaling means running the same software on smaller containers such as Docker and Kubernetes using ECS or EKS. Vertical scaling is adding more servers to the existing pool and doesn't run into limitations of individual servers is incorrect because Horizontal scaling is not related to using ECS or EKS containers on a smaller instance.

Reference:

<https://docs.aws.amazon.com/aws-technical-content/latest/microservices-on-aws/microservices-on-aws.html>

## QUESTION 13

A company has a top priority requirement to monitor a few database metrics and then afterward, send email notifications to the Operations team in case there is an issue. Which AWS services can accomplish this requirement? (Select TWO.)

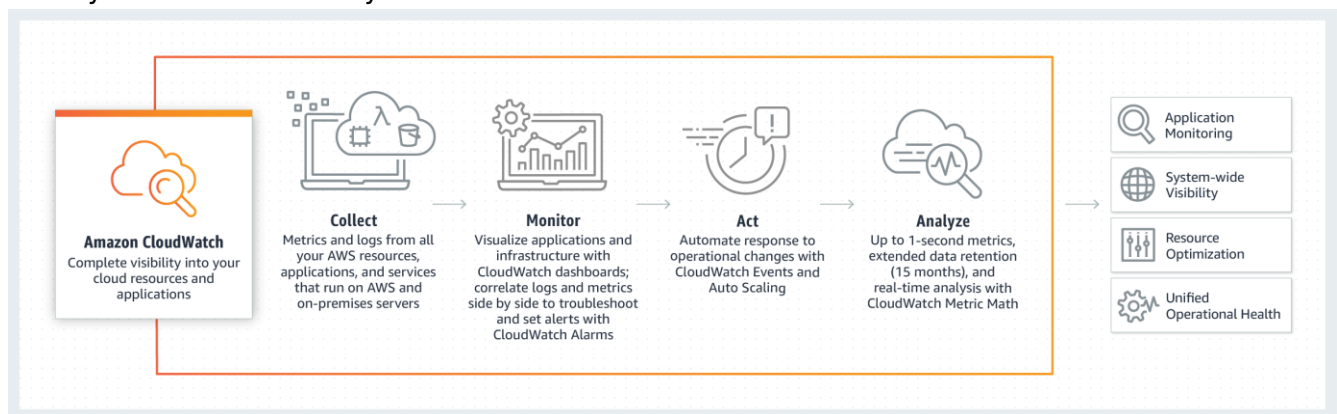
- A. Amazon Simple Notification Service (SNS)
- B. Amazon Simple Email Service
- C. Amazon Simple Queue Service (SQS)
- D. Amazon CloudWatch
- E. Amazon EC2 Instance with a running Berkeley Internet Name Domain (BIND) Server.

**Correct Answer: A,D**

### Explanation/Reference:

Amazon EC2 Instance with a running Berkeley Internet Name Domain (BIND) Server.

Amazon CloudWatch and Amazon Simple Notification Service (SNS) are correct. In this requirement, you can use Amazon CloudWatch to monitor the database and then Amazon SNS to send the emails to the Operations team. Take note that you should use SNS instead of SES (Simple Email Service) when you want to monitor your EC2 instances.



CloudWatch collects monitoring and operational data in the form of logs, metrics, and events, providing you with a unified view of AWS resources, applications, and services that run on AWS, and on-premises servers.

SNS is a highly available, durable, secure, fully managed pub/sub messaging service that enables you to decouple microservices, distributed systems, and serverless applications.

Amazon Simple Email Service is incorrect. SES is a cloud-based email sending service designed to send notification and transactional emails.

Amazon Simple Queue Service (SQS) is incorrect. SQS is a fully-managed message queuing service. It does not monitor applications nor send email notifications unlike SES.

Amazon EC2 Instance with a running Berkeley Internet Name Domain (BIND) Server is incorrect because BIND is primarily used as a Domain Name System (DNS) web service. This is only applicable if you have a private hosted zone in your AWS account. It does not monitor applications nor send



email notifications.

References:

<https://aws.amazon.com/cloudwatch/>

<https://aws.amazon.com/sns/>

Check out this Amazon CloudWatch Cheat Sheet:

<https://tutorialsdojo.com/amazon-cloudwatch/>

## QUESTION 14

A company is using an Amazon RDS for MySQL 5.6 with Multi-AZ deployment enabled and several web servers across two AWS Regions. The database is currently experiencing highly dynamic reads due to the growth of the company's website. The Solutions Architect tried to test the read performance from the secondary AWS Region and noticed a notable slowdown on the SQL queries.

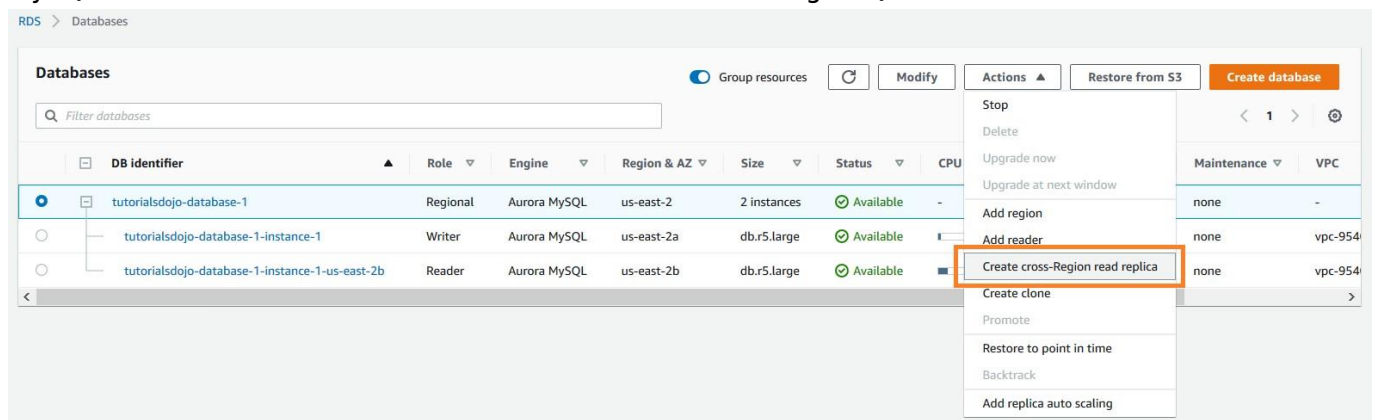
Which of the following options would provide a read replication latency of less than 1 second?

- A. Upgrade the MySQL database engine.
- B. Migrate the existing database to Amazon Aurora and create a cross-region read replica.
- C. Create an Amazon RDS for MySQL read replica in the secondary AWS Region.
- D. Use Amazon ElastiCache to improve database performance.

## Correct Answer: A

### Explanation/Reference:

Amazon Aurora is a MySQL and PostgreSQL-compatible relational database built for the cloud, that combines the performance and availability of traditional enterprise databases with the simplicity and cost-effectiveness of open source databases. Amazon Aurora is up to five times faster than standard MySQL databases and three times faster than standard PostgreSQL databases.



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

Amazon Aurora is fully managed by Amazon RDS, which automates time-consuming administration tasks like hardware provisioning, database setup, patching, and backups.

Based on the given scenario, there is a significant slowdown after testing the read performance from the secondary AWS Region. Since the existing setup is an Amazon RDS for MySQL, you should migrate the database to Amazon Aurora and create a cross-region read replica.

Feature	Amazon Aurora Replicas	MySQL Replicas
Number of replicas	Up to 15	Up to 5
Replication type	Asynchronous (milliseconds)	Asynchronous (seconds)
Performance impact on primary	Low	High
Replica location	In-region	Cross-region
Act as failover target	Yes (no data loss)	Yes (potentially minutes of data loss)
Automated failover	Yes	No
Support for user-defined replication delay	No	Yes
Support for different data or schema vs. primary	No	Yes

Less than 1 second!

The read replication latency of less than 1 second is only possible if you would use Amazon Aurora replicas. Aurora replicas are independent endpoints in an Aurora DB cluster, best used for scaling read operations and increasing availability. You can create up to 15 replicas within an AWS Region. Hence, the correct answer is: Migrate the existing database to Amazon Aurora and create a cross-region read replica.

The option that says: Upgrade the MySQL database engine is incorrect because upgrading the database engine wouldn't improve the read replication latency to milliseconds. To achieve the read replication latency of less than 1-second requirement, you need to use Amazon Aurora replicas.

The option that says: Use Amazon ElastiCache to improve database performance is incorrect. Amazon ElastiCache won't be able to improve the database performance because it is experiencing highly dynamic reads. This option would be helpful if the database frequently receives the same queries.

The option that says: Create an Amazon RDS for MySQL read replica in the secondary AWS Region is incorrect because MySQL replicas won't provide you a read replication latency of less than 1 second. RDS Read Replicas can only provide asynchronous replication in seconds and not in milliseconds. You have to use Amazon Aurora replicas in this scenario.

References:

<https://aws.amazon.com/rds/aurora/faqs/>

<https://docs.aws.amazon.com/AmazonRDS/latest/AuroraUserGuide/AuroraMySQL.Replication.CrossRegion.html>

Amazon Aurora Overview:

<https://youtu.be/iwS1h7rLNBQ>

Check out this Amazon Aurora Cheat Sheet:

<https://tutorialsdojo.com/amazon-aurora/>

## QUESTION 15

To save costs, your manager instructed you to analyze and review the setup of your AWS cloud infrastructure. You should also provide an estimate of how much your company will pay for all of the AWS resources that they are using. In this scenario, which of the following will incur costs? (Select TWO.)

- A. Using an Amazon VPC
- B. Public Data Set
- C. EBS Volumes attached to stopped EC2 Instances
- D. A stopped On-Demand EC2 Instance
- E. A running EC2 Instance



## Correct Answer: C,E

### Explanation/Reference:

Billing commences when Amazon EC2 initiates the boot sequence of an AMI instance. Billing ends when the instance terminates, which could occur through a web services command, by running "shutdown -h", or through instance failure. When you stop an instance, AWS shuts it down but doesn't charge hourly usage for a stopped instance or data transfer fees. However, AWS does charge for the storage of any Amazon EBS volumes.

Hence, a running EC2 Instance and EBS Volumes attached to stopped EC2 Instances are the right answers and conversely, a stopped On-Demand EC2 Instance is incorrect as there is no charge for a stopped EC2 instance that you have shut down.

Using Amazon VPC is incorrect because there are no additional charges for creating and using the VPC itself. Usage charges for other Amazon Web Services, including Amazon EC2, still apply at published rates for those resources, including data transfer charges.

Public Data Set is incorrect due to the fact that Amazon stores the data sets at no charge to the community and, as with all AWS services, you pay only for the compute and storage you use for your own applications.

References:

<https://aws.amazon.com/cloudtrail/>

<https://aws.amazon.com/vpc/faqs>

<https://docs.aws.amazon.com/AWSEC2/latest/UserGuide/using-public-data-sets.html>

Check out this Amazon EC2 Cheat Sheet:

<https://tutorialsdojo.com/amazon-elastic-compute-cloud-amazon-ec2/>

## QUESTION 16

A company launched a website that accepts high-quality photos and turns them into a downloadable video montage. The website offers a free and a premium account that guarantees faster processing. All requests by both free and premium members go through a single SQS queue and then processed by a group of EC2 instances that generate the videos. The company needs to ensure that the premium users who paid for the service have higher priority than the free members.

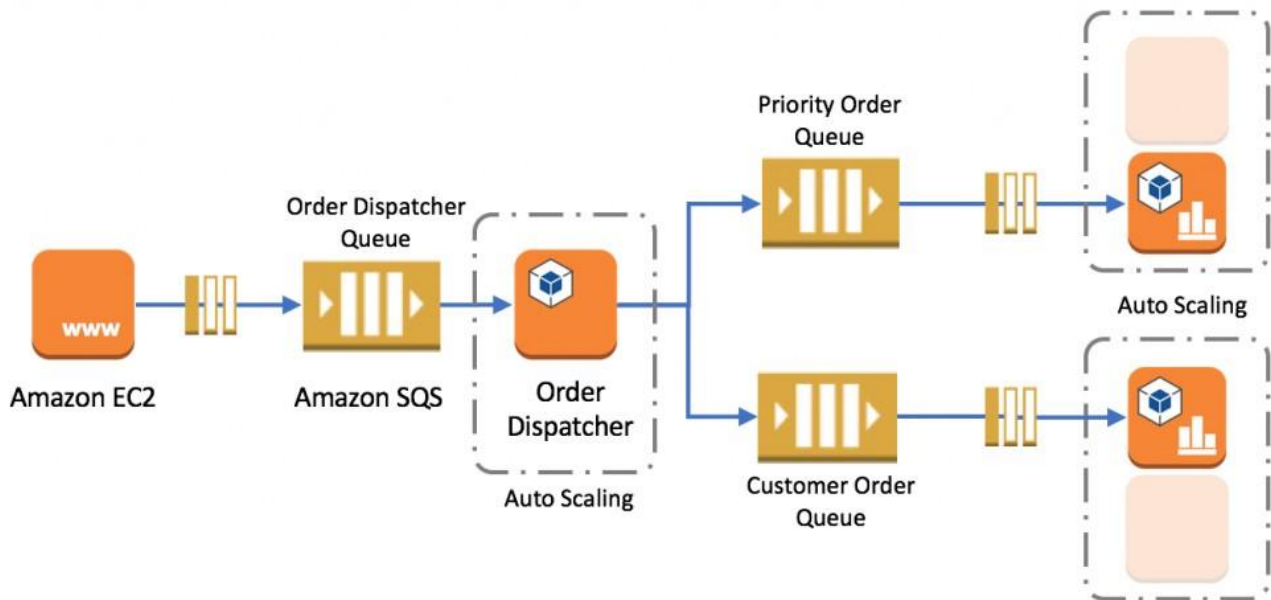
How should the company re-design its architecture to address this requirement?

- A. Use Amazon S3 to store and process the photos and then generate the video montage afterward.
- B. Create an SQS queue for free members and another one for premium members. Configure your EC2 instances to consume messages from the premium queue first and if it is empty, poll from the free members' SQS queue.
- C. For the requests made by premium members, set a higher priority in the SQS queue so it will be processed first compared to the requests made by free members.
- D. Use Amazon Kinesis to process the photos and generate the video montage in real-time.

## Correct Answer: B

## Explanation/Reference:

Amazon Simple Queue Service (SQS) is a fully managed message queuing service that enables you to decouple and scale microservices, distributed systems, and serverless applications. SQS eliminates the complexity and overhead associated with managing and operating message oriented middleware, and empowers developers to focus on differentiating work. Using SQS, you can send, store, and receive messages between software components at any volume, without losing messages or requiring other services to be available.



In this scenario, it is best to create 2 separate SQS queues for each type of members. The SQS queues for the premium members can be polled first by the EC2 Instances and once completed, the messages from the free members can be processed next.

Hence, the correct answer is: Create an SQS queue for free members and another one for premium members. Configure your EC2 instances to consume messages from the premium queue first and if it is empty, poll from the free members' SQS queue.

The option that says: For the requests made by premium members, set a higher priority in the SQS queue so it will be processed first compared to the requests made by free members is incorrect as you cannot set a priority to individual items in the SQS queue.

The option that says: Using Amazon Kinesis to process the photos and generate the video montage in real time is incorrect as Amazon Kinesis is used to process streaming data and it is not applicable in this scenario.

The option that says: Using Amazon S3 to store and process the photos and then generating the video montage afterwards is incorrect as Amazon S3 is used for durable storage and not for processing data.

Reference:

<https://docs.aws.amazon.com/AWSSimpleQueueService/latest/SQSDeveloperGuide/sqs-best-practices.html>

Check out this Amazon SQS Cheat Sheet:

<https://tutorialsdojo.com/amazon-sqs/>

## QUESTION 17

In Amazon EC2, you can manage your instances from the moment you launch them up to their

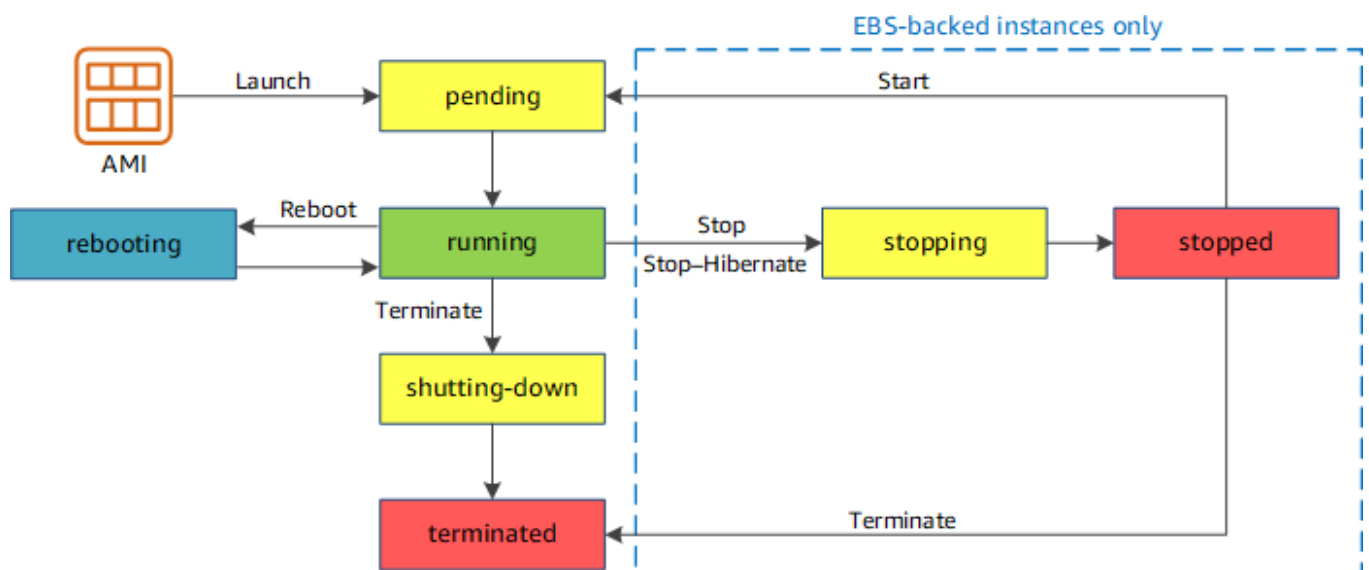
termination. You can flexibly control your computing costs by changing the EC2 instance state. Which of the following statements is true regarding EC2 billing? (Select TWO.)

- A. You will be billed when your Spot instance is preparing to stop with a stopping state.
- B. You will be billed when your On-Demand instance is preparing to hibernate with a stopping state.
- C. You will not be billed for any instance usage while an instance is not in the running state.
- D. You will be billed when your On-Demand instance is in pending state.
- E. You will be billed when your Reserved instance is in terminated state.

**Correct Answer: B,E**

### Explanation/Reference:

By working with Amazon EC2 to manage your instances from the moment you launch them through their termination, you ensure that your customers have the best possible experience with the applications or sites that you host on your instances. The following illustration represents the transitions between instance states. Notice that you can't stop and start an instance store-backed instance:



Below are the valid EC2 lifecycle instance states:

**pending** - The instance is preparing to enter the running state. An instance enters the pending state when it launches for the first time, or when it is restarted after being in the stopped state.

**running** - The instance is running and ready for use.

**stopping** - The instance is preparing to be stopped. Take note that you will not be billed if it is preparing to stop however, you will still be billed if it is just preparing to hibernate.

**stopped** - The instance is shut down and cannot be used. The instance can be restarted at any time.

**shutting-down** - The instance is preparing to be terminated.

**terminated** - The instance has been permanently deleted and cannot be restarted. Take note that Reserved Instances that applied to terminated instances are still billed until the end of their term according to their payment option.

The option that says: You will be billed when your On-Demand instance is preparing to hibernate with a stopping state is correct because when the instance state is stopping, you will not be billed if it is preparing to stop however, you will still be billed if it is just preparing to hibernate.

The option that says: You will be billed when your Reserved instance is in terminated state is correct

because Reserved Instances that applied to terminated instances are still billed until the end of their term according to their payment option. I actually raised a pull-request to Amazon team about the billing conditions for Reserved Instances, which has been approved and reflected on your official AWS Documentation: <https://github.com/awsdocs/amazon-ec2-user-guide/pull/45>

The option that says: You will be billed when your On-Demand instance is in pending state is incorrect because you will not be billed if your instance is in pending state.

The option that says: You will be billed when your Spot instance is preparing to stop with a stopping state is incorrect because you will not be billed if your instance is preparing to stop with a stopping state.

The option that says: You will not be billed for any instance usage while an instance is not in the running state is incorrect because the statement is not entirely true. You can still be billed if your instance is preparing to hibernate with a stopping state.

References:

<https://github.com/awsdocs/amazon-ec2-user-guide/pull/45>

<http://docs.aws.amazon.com/AWSEC2/latest/UserGuide/ec2-instance-lifecycle.html>

Check out this Amazon EC2 Cheat Sheet:

<https://tutorialsdojo.com/amazon-elastic-compute-cloud-amazon-ec2/>

## QUESTION 18

You have built a web application that checks for new items in an S3 bucket once every hour. If new items exist, a message is added to an SQS queue. You have a fleet of EC2 instances which retrieve messages from the SQS queue, process the file, and finally, send you and the user an email confirmation that the item has been successfully processed. Your officemate uploaded one test file to the S3 bucket and after a couple of hours, you noticed that you and your officemate have 50 emails from your application with the same message.

Which of the following is most likely the root cause why the application has sent you and the user multiple emails?

- A. There is a bug in the application.
- B. The `sqsSendMessage` attribute of the SQS queue is configured to 50.
- C. By default, SQS automatically deletes the messages that were processed by the consumers. It might be possible that your officemate has submitted the request 50 times which is why you received a lot of emails.
- D. Your application does not issue a delete command to the SQS queue after processing the message, which is why this message went back to the queue and was processed multiple times.

## Correct Answer: D

### Explanation/Reference:

In this scenario, the main culprit is that your application does not issue a delete command to the SQS queue after processing the message, which is why this message went back to the queue and was processed multiple times.

The option that says: The `sqsSendMessage` attribute of the SQS queue is configured to 50 is incorrect as there is no `sqsSendMessage` attribute in SQS.

The option that says: There is a bug in the application is a valid answer but since the scenario did not mention that the EC2 instances deleted the processed messages, the most likely cause of the

problem is that the application does not issue a delete command to the SQS queue as mentioned above.

The option that says: By default, SQS automatically deletes the messages that were processed by the consumers. It might be possible that your officemate has submitted the request 50 times which is why you received a lot of emails is incorrect as SQS does not automatically delete the messages.

Reference:

<https://aws.amazon.com/sqs/faqs/>

Check out this Amazon SQS Cheat Sheet:

<https://tutorialsdojo.com/amazon-sqs/>

## QUESTION 19

A company developed a financial analytics web application hosted in a Docker container using MEAN (MongoDB, Express.js, AngularJS, and Node.js) stack. You want to easily port that web application to AWS Cloud which can automatically handle all the tasks such as balancing load, auto-scaling, monitoring, and placing your containers across your cluster.

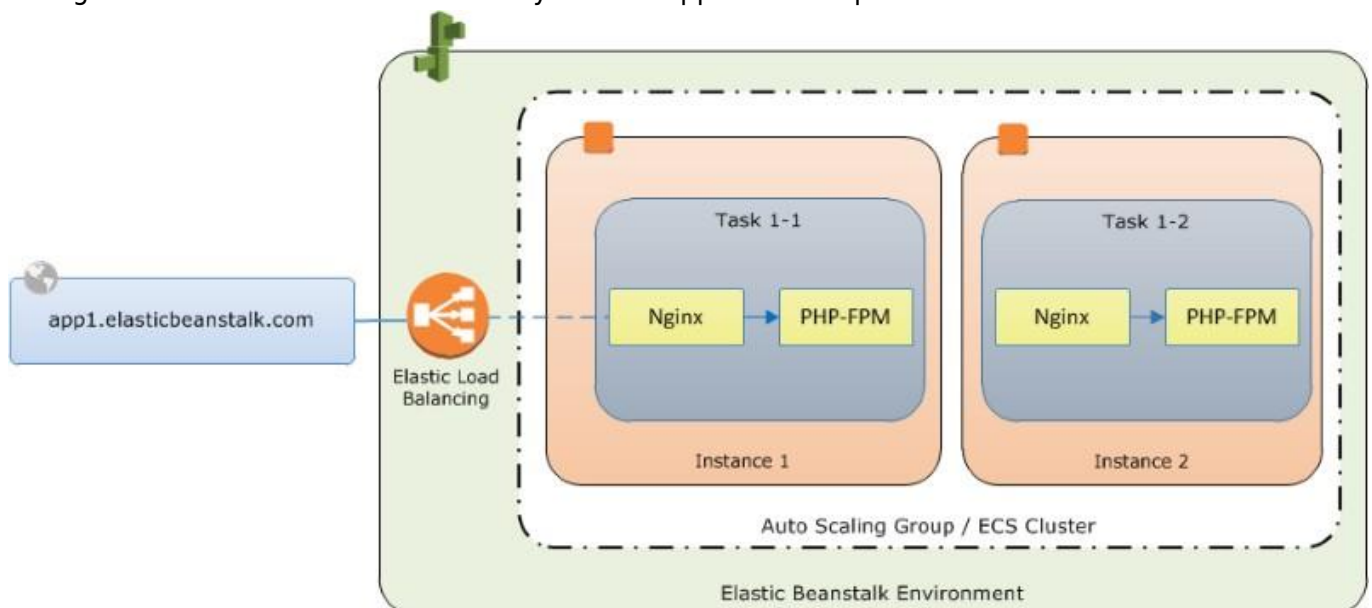
Which of the following services can be used to fulfill this requirement?

- A. AWS Elastic Beanstalk
- B. AWS Compute Optimizer
- C. Amazon Elastic Container Service (Amazon ECS)
- D. AWS CloudFormation

**Correct Answer: A**

### Explanation/Reference:

AWS Elastic Beanstalk supports the deployment of web applications from Docker containers. With Docker containers, you can define your own runtime environment. You can choose your own platform, programming language, and any application dependencies (such as package managers or tools), that aren't supported by other platforms. Docker containers are self-contained and include all the configuration information and software your web application requires to run.



By using Docker with Elastic Beanstalk, you have an infrastructure that automatically handles the details of capacity provisioning, load balancing, scaling, and application health monitoring. You can manage your web application in an environment that supports the range of services that are integrated with Elastic Beanstalk, including but not limited to VPC, RDS, and IAM.

Hence, the correct answer is: AWS Elastic Beanstalk.

Amazon Elastic Container Service (Amazon ECS) is incorrect. Although it also provides Service Auto Scaling, Service Load Balancing, and Monitoring with CloudWatch, these features are not automatically enabled by default unlike with Elastic Beanstalk. Take note that the scenario requires a service that will automatically handle all the tasks such as balancing load, auto-scaling, monitoring, and placing your containers across your cluster. You will have to manually configure these things if you wish to use ECS. With Elastic Beanstalk, you can manage your web application in an environment that supports the range of services easier.

AWS CloudFormation is incorrect. While you can deploy the infrastructure for your application thru CloudFormation templates, you will be the one responsible for connecting the AWS resources needed to build your application environment. With ElasticBeanstalk, all you have to do is upload your code; ElasticBeanstalk will automatically set up the environment for your application.

AWS Compute Optimizer is incorrect. Compute Optimizer simply analyzes your workload and recommends the optimal AWS resources needed to improve performance and reduce costs.

Reference:

[https://docs.aws.amazon.com/elasticbeanstalk/latest/dg/create\\_deploy\\_docker.html](https://docs.aws.amazon.com/elasticbeanstalk/latest/dg/create_deploy_docker.html)

Check out this AWS Elastic Beanstalk Cheat Sheet:

<https://tutorialsdojo.com/aws-elastic-beanstalk/>

AWS Elastic Beanstalk Overview:

<https://youtu.be/rx7e7Fej1Oo>

Elastic Beanstalk vs CloudFormation vs OpsWorks vs CodeDeploy:

<https://tutorialsdojo.com/elastic-beanstalk-vs-cloudformation-vs-opsworks-vs-codedeploy/>

Comparison of AWS Services Cheat Sheets:

<https://tutorialsdojo.com/comparison-of-aws-services/>

## QUESTION 20

An application consists of multiple EC2 instances in private subnets in different availability zones. The application uses a single NAT Gateway for downloading software patches from the Internet to the instances. There is a requirement to protect the application from a single point of failure when the NAT Gateway encounters a failure or if its availability zone goes down.

How should the Solutions Architect redesign the architecture to be more highly available and cost-effective

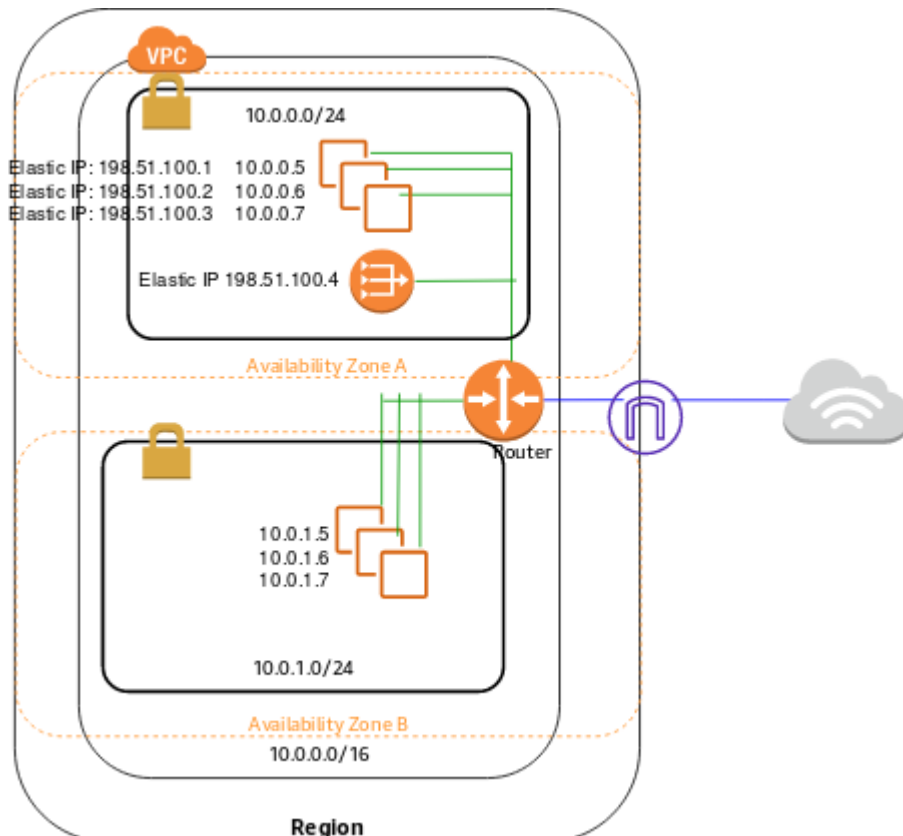
- A. Create a NAT Gateway in each availability zone. Configure the route table in each public subnet to ensure that instances use the NAT Gateway in the same availability zone.
- B. Create a NAT Gateway in each availability zone. Configure the route table in each private subnet to ensure that instances use the NAT Gateway in the same availability zone
- C. Create three NAT Gateways in each availability zone. Configure the route table in each private subnet to ensure that instances use the NAT Gateway in the same availability zone.
- D. Create two NAT Gateways in each availability zone. Configure the route table in each public subnet to ensure that instances use the NAT Gateway in the same availability zone.

## Correct Answer: B

### Explanation/Reference:

A NAT Gateway is a highly available, managed Network Address Translation (NAT) service for your resources in a private subnet to access the Internet. NAT gateway is created in a specific Availability Zone and implemented with redundancy in that zone.

You must create a NAT gateway on a public subnet to enable instances in a private subnet to connect to the Internet or other AWS services, but prevent the Internet from initiating a connection with those instances.



If you have resources in multiple Availability Zones and they share one NAT gateway, and if the NAT gateway's Availability Zone is down, resources in the other Availability Zones lose Internet access. 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.

Hence, the correct answer is: Create a NAT Gateway in each availability zone. Configure the route table in each private subnet to ensure that instances use the NAT Gateway in the same availability zone.

The option that says: Create a NAT Gateway in each availability zone. Configure the route table in each public subnet to ensure that instances use the NAT Gateway in the same availability zone is incorrect because you should configure the route table in the private subnet and not the public subnet to associate the right instances in the private subnet.

The options that say: Create two NAT Gateways in each availability zone. Configure the route table in each public subnet to ensure that instances use the NAT Gateway in the same availability zone and Create three NAT Gateways in each availability zone. Configure the route table in each private subnet to ensure that instances use the NAT Gateway in the same availability zone are both incorrect because a single NAT Gateway in each availability zone is enough. NAT Gateway is already redundant in nature, meaning, AWS already handles any failures that occur in your NAT Gateway in an



availability zone.

References:

<https://docs.aws.amazon.com/vpc/latest/userguide/vpc-nat-gateway.html>

<https://docs.aws.amazon.com/vpc/latest/userguide/vpc-nat-comparison.html>

Check out this Amazon VPC Cheat Sheet:

<https://tutorialsdojo.com/amazon-vpc/>

## QUESTION 21

A company has an On-Demand EC2 instance with an attached EBS volume. There is a scheduled job that creates a snapshot of this EBS volume every midnight at 12 AM when the instance is not used. One night, there has been a production incident where you need to perform a change on both the instance and on the EBS volume at the same time when the snapshot is currently taking place. Which of the following scenario is true when it comes to the usage of an EBS volume while the snapshot is in progress?

- A. The EBS volume cannot be detached or attached to an EC2 instance until the snapshot completes
- B. The EBS volume can be used while the snapshot is in progress.
- C. The EBS volume can be used in read-only mode while the snapshot is in progress.
- D. The EBS volume cannot be used until the snapshot completes.

## Correct Answer: B

### Explanation/Reference:

Snapshots occur asynchronously; the point-in-time snapshot is created immediately, but the status of the snapshot is pending until the snapshot is complete (when all of the modified blocks have been transferred to Amazon S3), which can take several hours for large initial snapshots or subsequent snapshots where many blocks have changed.

## Create Snapshot

Select resource type ☐ Volume  
☒ Instance

Instance ID\* i-11111111

Description Multiple volume snapshot

Exclude root volume ☐

1 to 4 of 4		
Volume ID	Volume Type	Encryption
vol-11111111	Root	Encrypted
vol-22222222	EBS	Not Encrypted
vol-33333333	EBS	Not Encrypted
vol-44444444	EBS	Not Encrypted

Copy tags from volume ☒

Key	Value
-----	-------

*This resource currently has no tags*

*Choose the Add tag button or [click to add a Name tag](#)*

Add Tag 50 remaining (Up to 50 tags maximum)

\* Required

Cancel

Create Snapshot

While it is completing, an in-progress snapshot is not affected by ongoing reads and writes to the volume hence, you can still use the EBS volume normally.

When you create an EBS volume based on a snapshot, the new volume begins as an exact replica of the original volume that was used to create the snapshot. The replicated volume loads data lazily in the background so that you can begin using it immediately. If you access data that hasn't been loaded yet, the volume immediately downloads the requested data from Amazon S3, and then continues loading the rest of the volume's data in the background.

References:

<https://docs.aws.amazon.com/AWSEC2/latest/UserGuide/ebs-creating-snapshot.html>

<https://docs.aws.amazon.com/AWSEC2/latest/UserGuide/EBSSnapshots.html>

Check out this Amazon EBS Cheat Sheet:

<https://tutorialsdojo.com/amazon-ebs/>

## QUESTION 22

A Solutions Architect designed a real-time data analytics system based on Kinesis Data Stream and Lambda. A week after the system has been deployed, the users noticed that it performed slowly as the data rate increases. The Architect identified that the performance of the Kinesis Data Streams is causing this problem.

Which of the following should the Architect do to improve performance?

- A. Implement Step Scaling to the Kinesis Data Stream.
- B. Increase the number of shards of the Kinesis stream by using the UpdateShardCount command.
- C. Improve the performance of the stream by decreasing the number of its shards using the

MergeShard command.

D. Replace the data stream with Amazon Kinesis Data Firehose instead.

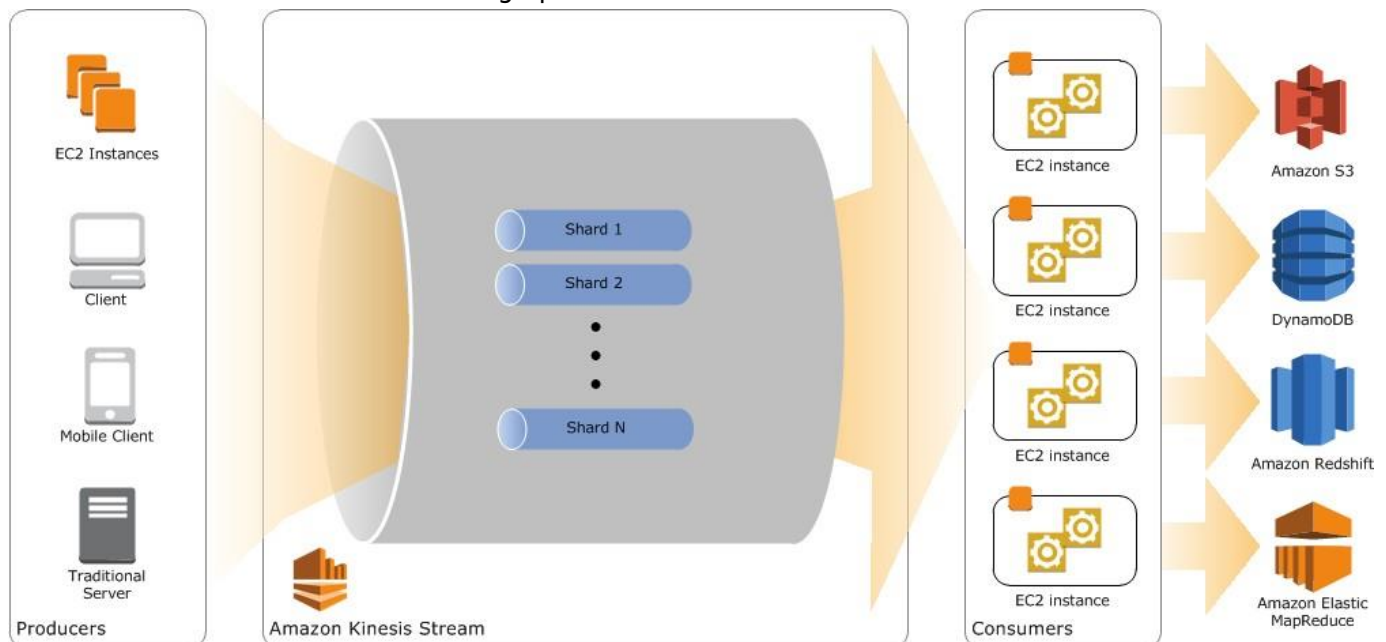
## Correct Answer: B

### Explanation/Reference:

Amazon Kinesis Data Streams supports resharding, which lets you adjust the number of shards in your stream to adapt to changes in the rate of data flow through the stream. Resharding is considered an advanced operation.

There are two types of resharding operations: shard split and shard merge. In a shard split, you divide a single shard into two shards. In a shard merge, you combine two shards into a single shard.

Resharding is always pairwise in the sense that you cannot split into more than two shards in a single operation, and you cannot merge more than two shards in a single operation. The shard or pair of shards that the resharding operation acts on are referred to as parent shards. The shard or pair of shards that result from the resharding operation are referred to as child shards.



Splitting increases the number of shards in your stream and therefore increases the data capacity of the stream. Because you are charged on a per-shard basis, splitting increases the cost of your stream. Similarly, merging reduces the number of shards in your stream and therefore decreases the data capacity—and cost—of the stream.

If your data rate increases, you can also increase the number of shards allocated to your stream to maintain the application performance. You can reshard your stream using the `UpdateShardCount` API. The throughput of an Amazon Kinesis data stream is designed to scale without limits via increasing the number of shards within a data stream. Hence, the correct answer is to increase the number of shards of the Kinesis stream by using the `UpdateShardCount` command.

Replacing the data stream with Amazon Kinesis Data Firehose instead is incorrect because the throughput of Kinesis Firehose is not exceptionally higher than Kinesis Data Streams. In fact, the throughput of an Amazon Kinesis data stream is designed to scale without limits via increasing the number of shards within a data stream.

Improving the performance of the stream by decreasing the number of its shards using the `MergeShard` command is incorrect because merging the shards will effectively decrease the

performance of the stream rather than improve it.

Implementing Step Scaling to the Kinesis Data Stream is incorrect because there is no Step Scaling feature for Kinesis Data Streams. This is only applicable for EC2.

References:

<https://aws.amazon.com/blogs/big-data/scale-your-amazon-kinesis-stream-capacity-with-updateshardcount/>

<https://aws.amazon.com/kinesis/data-streams/faqs/>

<https://docs.aws.amazon.com/streams/latest/dev/kinesis-using-sdk-java-resharding.html>

Check out this Amazon Kinesis Cheat Sheet:

<https://tutorialsdojo.com/amazon-kinesis/>

## QUESTION 23

A startup is using Amazon RDS to store data from a web application. Most of the time, the application has low user activity but it receives bursts of traffic within seconds whenever there is a new product announcement. The Solutions Architect needs to create a solution that will allow users around the globe to access the data using an API.

What should the Solutions Architect do meet the above requirement?

- A. Create an API using Amazon API Gateway and use Amazon Elastic Beanstalk with Auto Scaling to handle the bursts of traffic in seconds.
- B. Create an API using Amazon API Gateway and use an Auto Scaling group of Amazon EC2 instances to handle the bursts of traffic in seconds.
- C. Create an API using Amazon API Gateway and use the Amazon ECS cluster with Service Auto Scaling to handle the bursts of traffic in seconds.
- D. Create an API using Amazon API Gateway and use AWS Lambda to handle the bursts of traffic in seconds.

## Correct Answer: D

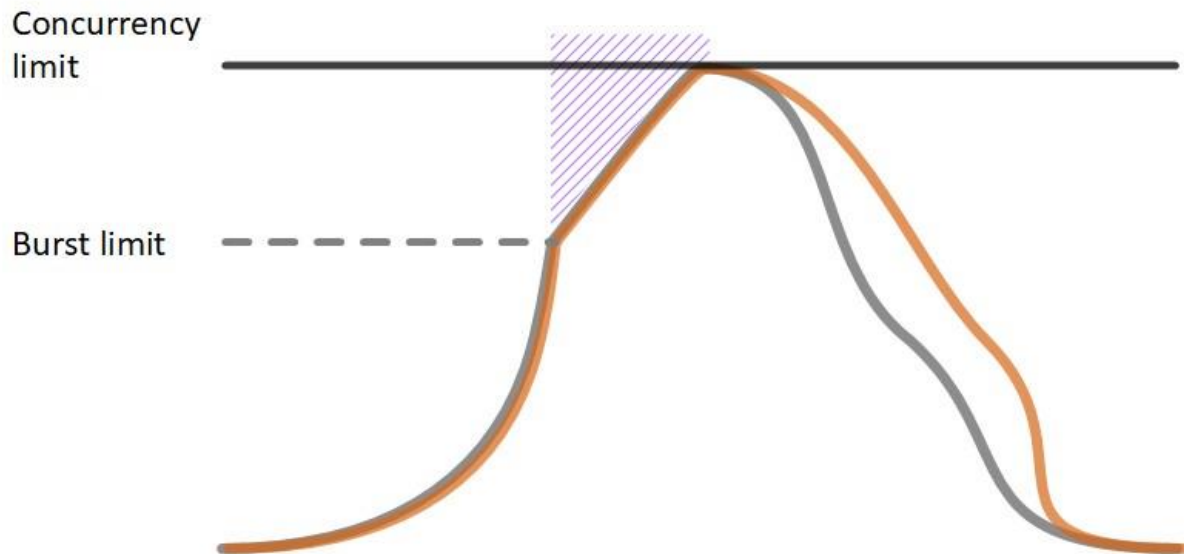
### Explanation/Reference:

AWS Lambda lets you run code without provisioning or managing servers. You pay only for the compute time you consume. With Lambda, you can run code for virtually any type of application or backend service - all with zero administration. Just upload your code, and Lambda takes care of everything required to run and scale your code with high availability. You can set up your code to automatically trigger from other AWS services or call it directly from any web or mobile app.

The first time you invoke your function, AWS Lambda creates an instance of the function and runs its handler method to process the event. When the function returns a response, it stays active and waits to process additional events. If you invoke the function again while the first event is being processed, Lambda initializes another instance, and the function processes the two events concurrently. As more events come in, Lambda routes them to available instances and creates new instances as needed.

When the number of requests decreases, Lambda stops unused instances to free up the scaling capacity for other functions.

## Function Scaling with Concurrency Limit



Your functions' concurrency is the number of instances that serve requests at a given time. For an initial burst of traffic, your functions' cumulative concurrency in a Region can reach an initial level of between 500 and 3000, which varies per Region.

Based on the given scenario, you need to create a solution that will satisfy the two requirements. The first requirement is to create a solution that will allow the users to access the data using an API. To implement this solution, you can use Amazon API Gateway. The second requirement is to handle the burst of traffic within seconds. You should use AWS Lambda in this scenario because Lambda functions can absorb reasonable bursts of traffic for approximately 15-30 minutes.

Lambda can scale faster than the regular Auto Scaling feature of Amazon EC2, Amazon Elastic Beanstalk, or Amazon ECS. This is because AWS Lambda is more lightweight than other computing services. Under the hood, Lambda can run your code to thousands of available AWS-managed EC2 instances (that could already be running) within seconds to accommodate traffic. This is faster than the Auto Scaling process of launching new EC2 instances that could take a few minutes or so. An alternative is to overprovision your compute capacity but that will incur significant costs. The best option to implement given the requirements is a combination of AWS Lambda and Amazon API Gateway.

Hence, the correct answer is: Create an API using Amazon API Gateway and use AWS Lambda to handle the bursts of traffic.

The option that says: Create an API using Amazon API Gateway and use the Amazon ECS cluster with Service Auto Scaling to handle the bursts of traffic in seconds is incorrect. AWS Lambda is a better option than Amazon ECS since it can handle a sudden burst of traffic within seconds and not minutes. The option that says: Create an API using Amazon API Gateway and use Amazon Elastic Beanstalk with Auto Scaling to handle the bursts of traffic in seconds is incorrect because just like the previous option, the use of Auto Scaling has a delay of a few minutes as it launches new EC2 instances that will be used by Amazon Elastic Beanstalk.

The option that says: Create an API using Amazon API Gateway and use an Auto Scaling group of Amazon EC2 instances to handle the bursts of traffic in seconds is incorrect because the processing time of Amazon EC2 Auto Scaling to provision new resources takes minutes. Take note that in the scenario, a burst of traffic within seconds is expected to happen.

References:

<https://aws.amazon.com/blogs/startups/from-0-to-100-k-in-seconds-instant-scale-with-aws-lambda/>

<https://docs.aws.amazon.com/lambda/latest/dg/invoke-scaling.html>

Check out this AWS Lambda Cheat Sheet:

<https://tutorialsdojo.com/aws-lambda/>

## QUESTION 24

A company recently launched an e-commerce application that is running in eu-east-2 region, which strictly requires six EC2 instances running at all times. In that region, there are 3 Availability Zones (AZ) that you can use - eu-east-2a, eu-east-2b, and eu-east-2c.

Which of the following deployments provide 100% fault tolerance if any single AZ in the region becomes unavailable? (Select TWO.)

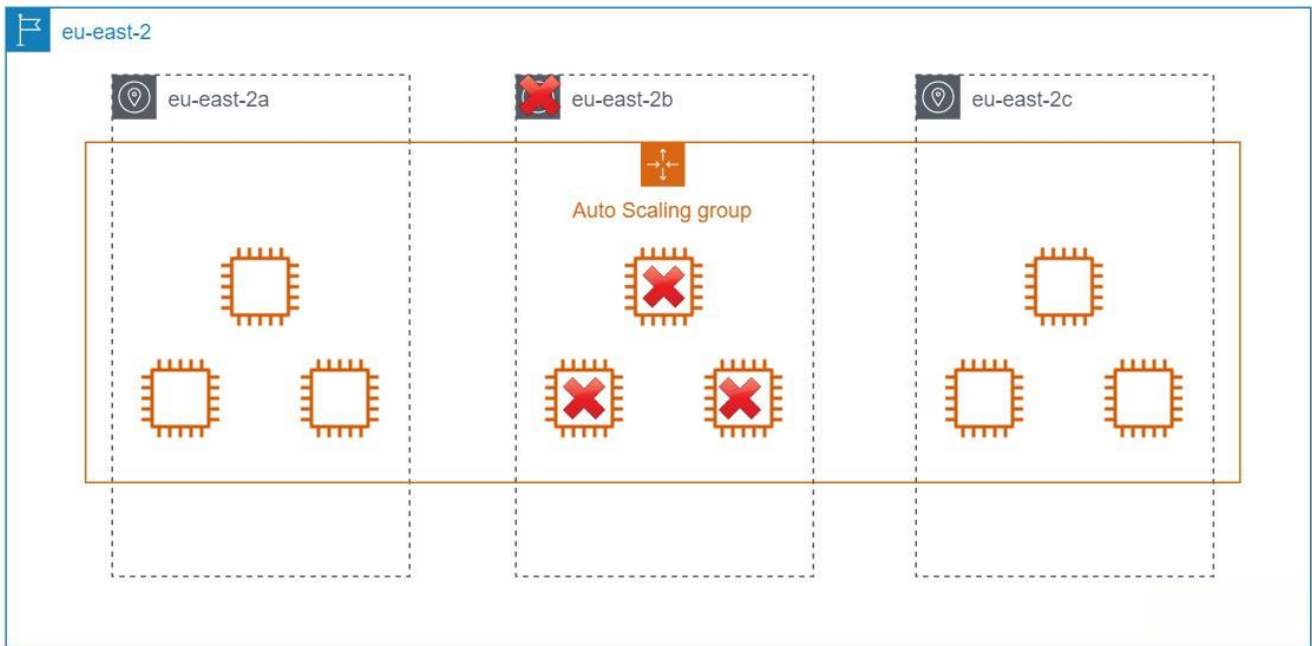
- A. eu-east-2a with four EC2 instances, eu-east-2b with two EC2 instances, and eu-east-2c with two EC2 instances
- B. eu-east-2a with two EC2 instances, eu-east-2b with four EC2 instances, and eu-east-2c with two EC2 instances
- C. eu-east-2a with six EC2 instances, eu-east-2b with six EC2 instances, and eu-east-2c with no EC2 instances
- D. eu-east-2a with three EC2 instances, eu-east-2b with three EC2 instances, and eu-east-2c with three EC2 instances
- E. eu-east-2a with two EC2 instances, eu-east-2b with two EC2 instances, and eu-east-2c with two EC2 instances

**Correct Answer: C,D**

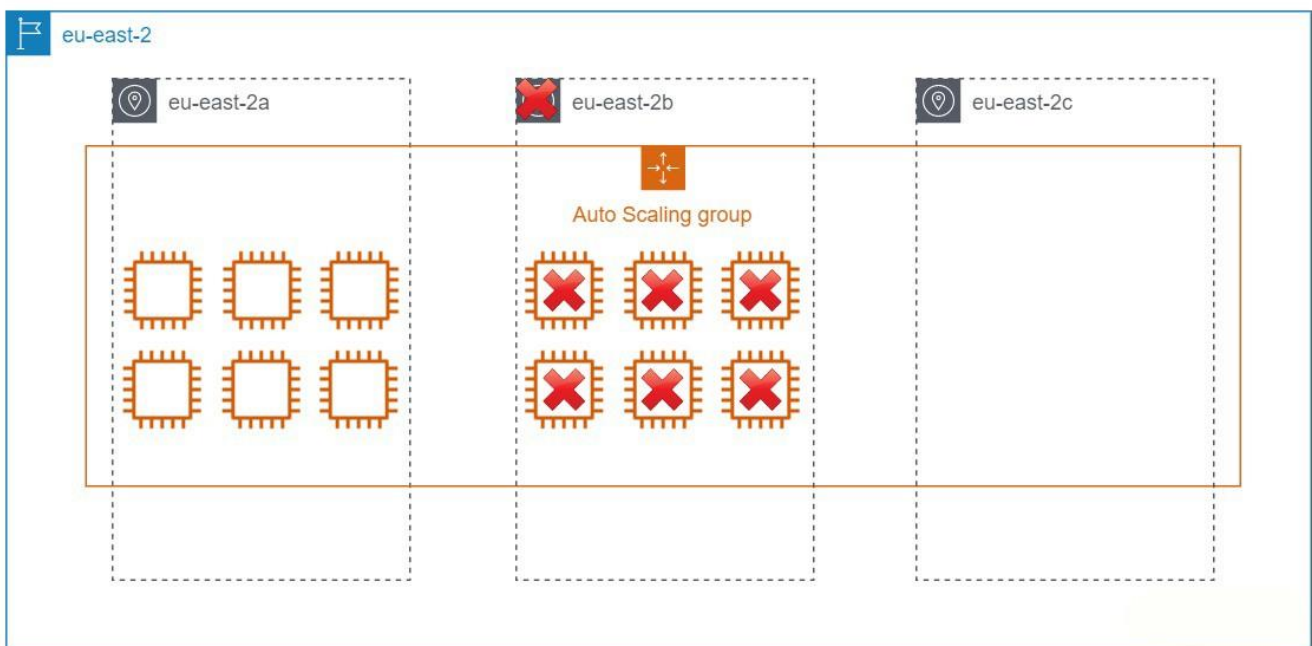
## Explanation/Reference:

Fault Tolerance is the ability of a system to remain in operation even if some of the components used to build the system fail. In AWS, this means that in the event of server fault or system failures, the number of running EC2 instances should not fall below the minimum number of instances required by the system for it to work properly. So if the application requires a minimum of 6 instances, there should be at least 6 instances running in case there is an outage in one of the Availability Zones or if there are server issues.

## Case 1



## Case 2



In this scenario, you have to simulate a situation where one Availability Zone became unavailable for each option and check whether it still has 6 running instances.

Hence, the correct answers are: eu-east-2a with six EC2 instances, eu-east-2b with six EC2 instances, and eu-east-2c with no EC2 instances and eu-east-2a with three EC2 instances, eu-east-2b with three EC2 instances, and eu-east-2c with three EC2 instances because even if one of the availability zones were to go down, there would still be 6 active instances.

Reference:

[https://media.amazonwebservices.com/AWS\\_Building\\_Fault\\_Tolerant\\_Applications.pdf](https://media.amazonwebservices.com/AWS_Building_Fault_Tolerant_Applications.pdf)



## QUESTION 25

A company conducted a surprise IT audit on all of the AWS resources being used in the production environment. During the audit activities, it was noted that you are using a combination of Standard and Convertible Reserved EC2 instances in your applications.

Which of the following are the characteristics and benefits of using these two types of Reserved EC2 instances? (Select TWO.)

- A. Convertible Reserved Instances allow you to exchange for another convertible reserved instance of a different instance family.
- B. It can enable you to reserve capacity for your Amazon EC2 instances in multiple Availability Zones and multiple AWS Regions for any duration.
- C. It runs in a VPC on hardware that's dedicated to a single customer.
- D. Unused Standard Reserved Instances can later be sold at the Reserved Instance Marketplace.
- E. Unused Convertible Reserved Instances can later be sold at the Reserved Instance Marketplace.

**Correct Answer: A,D**

### Explanation/Reference:

Reserved Instances (RIs) provide you with a significant discount (up to 75%) compared to On-Demand instance pricing. You have the flexibility to change families, OS types, and tenancies while benefiting from RI pricing when you use Convertible RIs. One important thing to remember here is that Reserved Instances are not physical instances, but rather a billing discount applied to the use of On-Demand Instances in your account.

The offering class of a Reserved Instance is either Standard or Convertible. A Standard Reserved Instance provides a more significant discount than a Convertible Reserved Instance, but you can't exchange a Standard Reserved Instance unlike Convertible Reserved Instances. You can modify Standard and Convertible Reserved Instances. Take note that in Convertible Reserved Instances, you are allowed to exchange another Convertible Reserved instance with a different instance type and tenancy.

The configuration of a Reserved Instance comprises a single instance type, platform, scope, and tenancy over a term. If your computing needs change, you might be able to modify or exchange your Reserved Instance.

When your computing needs change, you can modify your Standard or Convertible Reserved Instances and continue to take advantage of the billing benefit. You can modify the Availability Zone, scope, network platform, or instance size (within the same instance type) of your Reserved Instance. You can also sell your unused instance for Standard RIs but not Convertible RIs on the Reserved Instance Marketplace.

Hence, the correct options are:

- Unused Standard Reserved Instances can later be sold at the Reserved Instance Marketplace.
- Convertible Reserved Instances allow you to exchange for another convertible reserved instance of a different instance family.

The option that says: Unused Convertible Reserved Instances can later be sold at the Reserved Instance Marketplace is incorrect. This is not possible. Only Standard RIs can be sold at the Reserved Instance Marketplace.

The option that says: It can enable you to reserve capacity for your Amazon EC2 instances in multiple Availability Zones and multiple AWS Regions for any duration is incorrect because you can reserve

capacity to a specific AWS Region (regional Reserved Instance) or specific Availability Zone (zonal Reserved Instance) only. You cannot reserve capacity to multiple AWS Regions in a single RI purchase.

The option that says: It runs in a VPC on hardware that's dedicated to a single customer is incorrect because that is the description of a Dedicated instance and not a Reserved Instance. A Dedicated instance runs in a VPC on hardware that's dedicated to a single customer.

References:

<https://docs.aws.amazon.com/AWSEC2/latest/UserGuide/ri-modifying.html>

<https://aws.amazon.com/ec2/pricing/reserved-instances/>

<https://docs.aws.amazon.com/AWSEC2/latest/UserGuide/ec2-reserved-instances.html>

<https://docs.aws.amazon.com/AWSEC2/latest/UserGuide/reserved-instances-types.html>

Amazon EC2 Overview:

[https://youtu.be/7VsGIHT\\_jQE](https://youtu.be/7VsGIHT_jQE)

Check out this Amazon EC2 Cheat Sheet:

<https://tutorialsdojo.com/amazon-elastic-compute-cloud-amazon-ec2/>

## QUESTION 26

A Solutions Architect is building a cloud infrastructure where EC2 instances require access to various AWS services such as S3 and Redshift. The Architect will also need to provide access to system administrators so they can deploy and test their changes.

Which configuration should be used to ensure that the access to the resources is secured and not compromised? (Select TWO.)

- A. Store the AWS Access Keys in the EC2 instance.
- B. Assign an IAM user for each Amazon EC2 Instance.
- C. Store the AWS Access Keys in ACM.
- D. Assign an IAM role to the Amazon EC2 instance.
- E. Enable Multi-Factor Authentication.

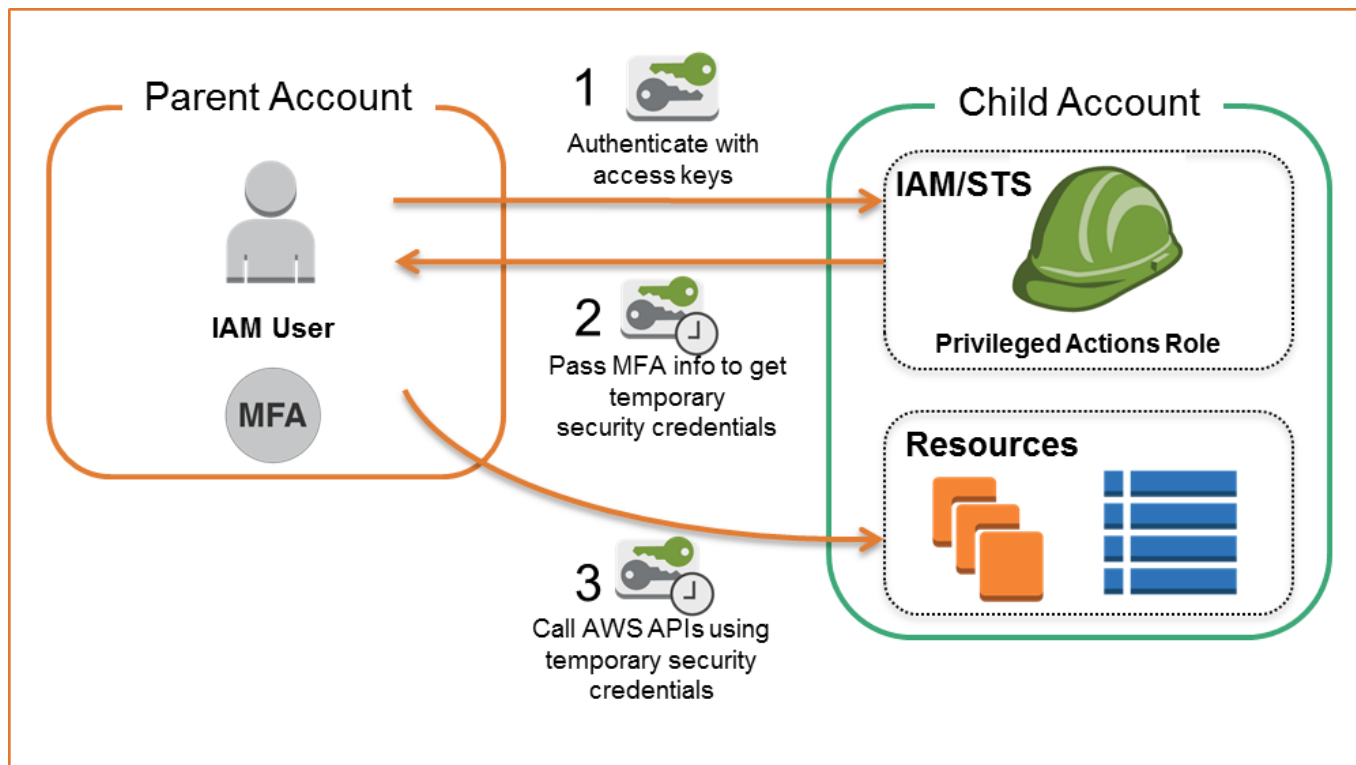
**Correct Answer: D,E**

## Explanation/Reference:

In this scenario, the correct answers are:

- Enable Multi-Factor Authentication
- Assign an IAM role to the Amazon EC2 instance

Always remember that you should associate IAM roles to EC2 instances and not an IAM user, for the purpose of accessing other AWS services. IAM roles are designed so that your applications can securely make API requests from your instances, without requiring you to manage the security credentials that the applications use. Instead of creating and distributing your AWS credentials, you can delegate permission to make API requests using IAM roles.



AWS Multi-Factor Authentication (MFA) is a simple best practice that adds an extra layer of protection on top of your user name and password. With MFA enabled, when a user signs in to an AWS website, they will be prompted for their user name and password (the first factor—what they know), as well as for an authentication code from their AWS MFA device (the second factor—what they have). Taken together, these multiple factors provide increased security for your AWS account settings and resources. You can enable MFA for your AWS account and for individual IAM users you have created under your account. MFA can also be used to control access to AWS service APIs.

Storing the AWS Access Keys in the EC2 instance is incorrect. This is not recommended by AWS as it can be compromised. Instead of storing access keys on an EC2 instance for use by applications that run on the instance and make AWS API requests, you can use an IAM role to provide temporary access keys for these applications.

Assigning an IAM user for each Amazon EC2 Instance is incorrect because there is no need to create an IAM user for this scenario since IAM roles already provide greater flexibility and easier management.

Storing the AWS Access Keys in ACM is incorrect because ACM is just a service that lets you easily provision, manage, and deploy public and private SSL/TLS certificates for use with AWS services and your internal connected resources. It is not used as a secure storage for your access keys.

References:

<https://aws.amazon.com/iam/details/mfa/>

<https://docs.aws.amazon.com/AWSEC2/latest/UserGuide/iam-roles-for-amazon-ec2.html>

Check out this AWS IAM Cheat Sheet:

<https://tutorialsdojo.com/aws-identity-and-access-management-iam/>

## QUESTION 27

A company has a web application that is relying entirely on slower disk-based databases, causing it to perform slowly. To improve its performance, the Solutions Architect integrated an in-memory data store to the web application using ElastiCache.

How does Amazon ElastiCache improve database performance?

A It securely delivers data to customers globally with low latency and high transfer speeds.

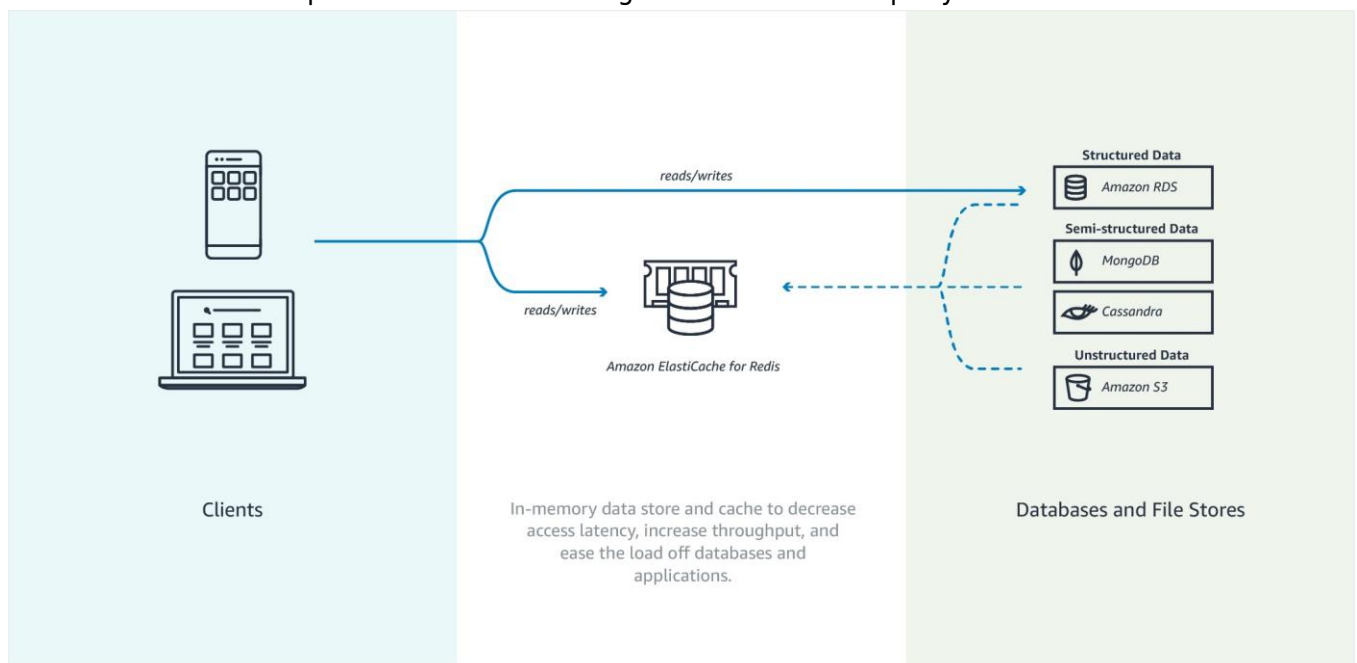
- B. It provides an in-memory cache that delivers up to 10x performance improvement from milliseconds to microseconds or even at millions of requests per second.
- C. By caching database query results.
- D. It reduces the load on your database by routing read queries from your applications to the Read Replica.

**Correct Answer: C**

### Explanation/Reference:

ElastiCache improves the performance of your database through caching query results. The primary purpose of an in-memory key-value store is to provide ultra-fast (submillisecond latency) and inexpensive access to copies of data. Most data stores have areas of data that are frequently accessed but seldom updated. Additionally, querying a database is always slower and more expensive than locating a key in a key-value pair cache. Some database queries are especially expensive to perform, for example, queries that involve joins across multiple tables or queries with intensive calculations.

By caching such query results, you pay the price of the query once and then are able to quickly retrieve the data multiple times without having to re-execute the query.



The option that says: It securely delivers data to customers globally with low latency and high transfer speeds is incorrect because this option describes what CloudFront does and not ElastiCache.

The option that says: It provides an in-memory cache that delivers up to 10x performance improvement from milliseconds to microseconds or even at millions of requests per second is incorrect because this option describes what Amazon DynamoDB Accelerator (DAX) does and not ElastiCache. Amazon DynamoDB Accelerator (DAX) is a fully managed, highly available, in-memory cache for DynamoDB. Amazon ElastiCache cannot provide a performance improvement from milliseconds to microseconds, let alone millions of requests per second like DAX can.

The option that says: It reduces the load on your database by routing read queries from your applications to the Read Replica is incorrect because this option describes what an RDS Read Replica does and not ElastiCache. Amazon RDS Read Replicas enable you to create one or more read-only

copies of your database instance within the same AWS Region or in a different AWS Region.

References:

<https://aws.amazon.com/elasticache/>

<https://docs.aws.amazon.com/AmazonElastiCache/latest/red-ug/elasticache-use-cases.html>

Check out this Amazon Elasticache Cheat Sheet:

<https://tutorialsdojo.com/amazon-elasticache/>

## QUESTION 28

A company has several microservices that send messages to an Amazon SQS queue and a backend application that poll the queue to process the messages. The company also has a Service Level Agreement (SLA) which defines the acceptable amount of time that can elapse from the point when the messages are received until a response is sent. The backend operations are I/O-intensive as the number of messages is constantly growing, causing the company to miss its SLA. The Solutions Architect must implement a new architecture that improves the application's processing time and load management.

Which of the following is the MOST effective solution that can satisfy the given requirement?

- A. Create an AMI of the backend application's EC2 instance and replace it with a larger instance size.
- B. Create an AMI of the backend application's EC2 instance. Use the image to set up an Auto Scaling group and configure a target tracking scaling policy based on the CPUUtilization metric with a target value of 80%.
- C. Create an AMI of the backend application's EC2 instance and launch it to a cluster placement group.
- D. Create an AMI of the backend application's EC2 instance. Use the image to set up an Auto Scaling group and configure a target tracking scaling policy based on the ApproximateAgeOfOldestMessage metric.

## Correct Answer: D

### Explanation/Reference:

Amazon Simple Queue Service (SQS) is a fully managed message queuing service that enables you to decouple and scale microservices, distributed systems, and serverless applications. SQS eliminates the complexity and overhead associated with managing and operating message-oriented middleware and empowers developers to focus on differentiating work. Using SQS, you can send, store, and receive messages between software components at any volume, without losing messages or requiring other services to be available.



The `ApproximateAgeOfOldestMessage` metric is useful when applications have time-sensitive messages and you need to ensure that messages are processed within a specific time period. You can use this metric to set Amazon CloudWatch alarms that issue alerts when messages remain in the queue for extended periods of time. You can also use alerts to take action, such as increasing the number of consumers to process messages more quickly.

With a target tracking scaling policy, you can scale (increase or decrease capacity) a resource based on a target value for a specific CloudWatch metric. To create a custom metric for this policy, you need to use AWS CLI or AWS SDKs. Take note that you need to create an AMI from the instance first before you can create an Auto Scaling group to scale the instances based on the `ApproximateAgeOfOldestMessage` metric.

Hence, the correct answer is: Create an AMI of the backend application's EC2 instance. Use the image to set up an Auto Scaling Group and configure a target tracking scaling policy based on the `ApproximateAgeOfOldestMessage` metric.

The option that says: Create an AMI of the backend application's EC2 instance. Use the image to set up an Auto Scaling Group and configure a target tracking scaling policy based on the `CPUUtilization` metric with a target value of 80% is incorrect. Although this will improve the backend processing, the scaling policy based on the `CPUUtilization` metric is not meant for time-sensitive messages where you need to ensure that the messages are processed within a specific time period. It will only trigger the scale-out activities based on the CPU Utilization of the current instances, and not based on the age of the message, which is a crucial factor in meeting the SLA. To satisfy the requirement in the scenario, you should use the `ApproximateAgeOfOldestMessage` metric.

The option that says: Create an AMI of the backend application's EC2 instance and replace it with a larger instance size is incorrect because replacing the instance with a large size won't be enough to dynamically handle workloads at any level. You need to implement an Auto Scaling group to automatically adjust the capacity of your computing resources.

The option that says: Create an AMI of the backend application's EC2 instance and launch it to a cluster placement group is incorrect because a cluster placement group is just a logical grouping of EC2 instances. Instead of launching the instance in a placement group, you must set up an Auto Scaling group for your EC2 instances and configure a target tracking scaling policy based on the `ApproximateAgeOfOldestMessage` metric.

References:

<https://aws.amazon.com/about-aws/whats-new/2016/08/new-amazon-cloudwatch-metric-for-amazon-sqs-monitors-the-age-of-the-oldest-message/>

<https://docs.aws.amazon.com/AWSSimpleQueueService/latest/SQSDeveloperGuide/sqs-available-cloudwatch-metrics.html>

<https://docs.aws.amazon.com/autoscaling/ec2/userguide/as-using-sqs-queue.html>

Check out this Amazon SQS Cheat Sheet:

<https://tutorialsdojo.com/amazon-sqs/>

## QUESTION 29

A company has multiple AWS Site-to-Site VPN connections placed between their VPCs and their remote network. During peak hours, many employees are experiencing slow connectivity issues, which limits their productivity. The company has asked a solutions architect to scale the throughput of the VPN connections.

Which solution should the architect carry out?

- A. Associate the VPCs to an Equal Cost Multipath Routing (ECMR)-enabled transit gateway and attach additional VPN tunnels.
- B. Re-route some of the VPN connections to a secondary customer gateway device on the remote network's end.

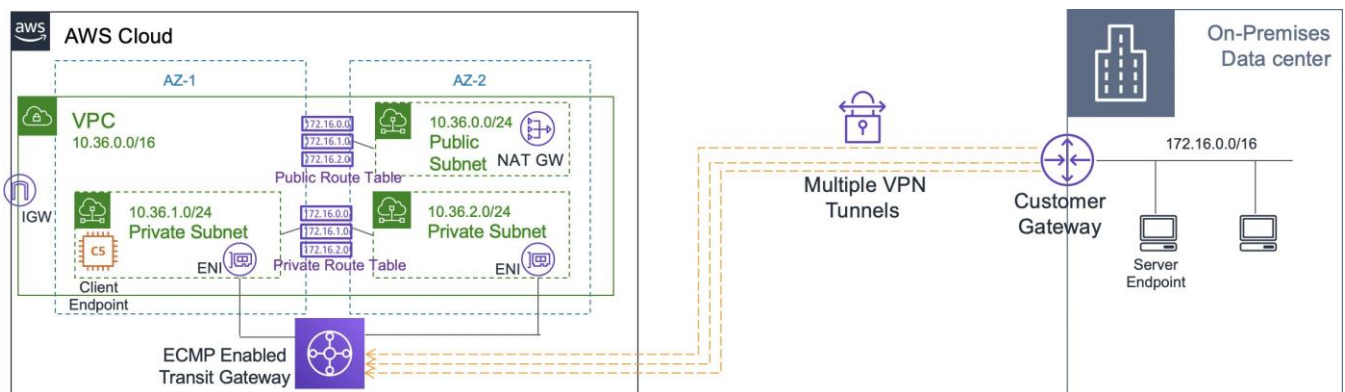


- C. Modify the VPN configuration by increasing the number of tunnels to scale the throughput.
- D. Add more virtual private gateways to a VPC and enable Equal Cost Multipath Routing (ECMR) to get higher VPN bandwidth.

**Correct Answer: A**

### Explanation/Reference:

With AWS Transit Gateway, you can simplify the connectivity between multiple VPCs and also connect to any VPC attached to AWS Transit Gateway with a single VPN connection.



AWS Transit Gateway also enables you to scale the IPsec VPN throughput with equal-cost multi-path (ECMP) routing support over multiple VPN tunnels. A single VPN tunnel still has a maximum throughput of 1.25 Gbps. If you establish multiple VPN tunnels to an ECMP-enabled transit gateway, it can scale beyond the default limit of 1.25 Gbps.

Hence, the correct answer is: Associate the VPCs to an Equal Cost Multipath Routing (ECMR)-enabled transit gateway and attach additional VPN tunnels.

The option that says: Add more virtual private gateways to a VPC and enable Equal Cost Multipath Routing (ECMR) to get higher VPN bandwidth is incorrect because a VPC can only have a single virtual private gateway attached to it one at a time. Also, there is no option to enable ECMP in a virtual private gateway.

The option that says: Modify the VPN configuration by increasing the number of tunnels to scale the throughput is incorrect. The maximum tunnel for a VPN connection is two. You cannot increase this beyond its limit.

The option that says: Re-route some of the VPN connections to a secondary customer gateway device on the remote network's end is incorrect. This would only increase connection redundancy and won't increase throughput. For example, connections can failover to the secondary customer gateway device in case the primary customer gateway device becomes unavailable.

References:

<https://aws.amazon.com/premiumsupport/knowledge-center/transit-gateway-ecmp-multiple-tunnels/>  
<https://aws.amazon.com/blogs/networking-and-content-delivery/scaling-vpn-throughput-using-aws-transit-gateway/>

Check out this AWS Transit Gateway Cheat Sheet:

<https://tutorialsdojo.com/aws-transit-gateway/>



## QUESTION 30

A company is planning to launch a High Performance Computing (HPC) cluster in AWS that does Computational Fluid Dynamics (CFD) simulations. The solution should scale-out their simulation jobs to experiment with more tunable parameters for faster and more accurate results. The cluster is composed of Windows servers hosted on t3a.medium EC2 instances. As the Solutions Architect, you should ensure that the architecture provides higher bandwidth, higher packet per second (PPS) performance, and consistently lower inter-instance latencies.

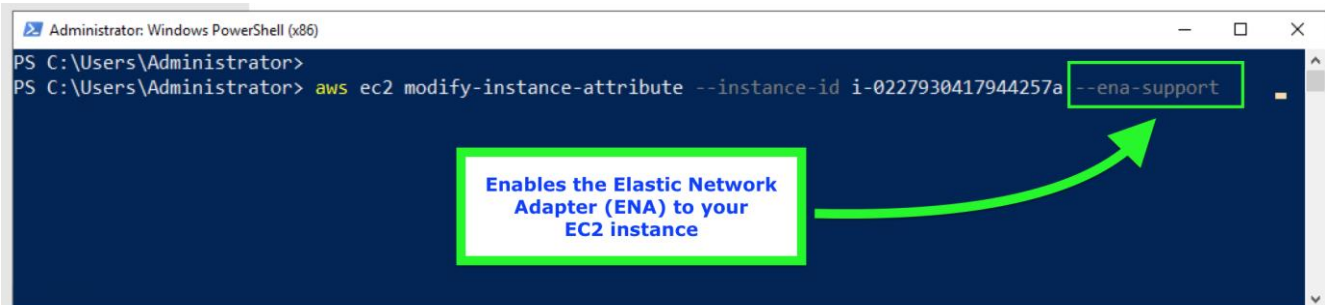
Which is the MOST suitable and cost-effective solution that the Architect should implement to achieve the above requirements?

- A. Enable Enhanced Networking with Elastic Fabric Adapter (EFA) on the Windows EC2 Instances.
- B. Use AWS ParallelCluster to deploy and manage the HPC cluster to provide higher bandwidth, higher packet per second (PPS) performance, and lower inter-instance latencies.
- C. Enable Enhanced Networking with Elastic Network Adapter (ENA) on the Windows EC2 Instances.
- D. Enable Enhanced Networking with Intel 82599 Virtual Function (VF) interface on the Windows EC2 Instances.

**Correct Answer: D**

### Explanation/Reference:

Enhanced networking uses single root I/O virtualization (SR-IOV) to provide high-performance networking capabilities on supported instance types. SR-IOV is a method of device virtualization that provides higher I/O performance and lower CPU utilization when compared to traditional virtualized network interfaces. Enhanced networking provides higher bandwidth, higher packet per second (PPS) performance, and consistently lower inter-instance latencies. There is no additional charge for using enhanced networking.



Amazon EC2 provides enhanced networking capabilities through the Elastic Network Adapter (ENA). It supports network speeds of up to 100 Gbps for supported instance types. Elastic Network Adapters (ENAs) provide traditional IP networking features that are required to support VPC networking.

An Elastic Fabric Adapter (EFA) is simply an Elastic Network Adapter (ENA) with added capabilities. It provides all of the functionality of an ENA, with additional OS-bypass functionality. OS-bypass is an access model that allows HPC and machine learning applications to communicate directly with the network interface hardware to provide low-latency, reliable transport functionality.

The OS-bypass capabilities of EFAs are not supported on Windows instances. If you attach an EFA to a Windows instance, the instance functions as an Elastic Network Adapter, without the added EFA capabilities.

Hence, the correct answer is to enable Enhanced Networking with Elastic Network Adapter (ENA) on the Windows EC2 Instances.

Enabling Enhanced Networking with Elastic Fabric Adapter (EFA) on the Windows EC2 Instances is incorrect because the OS-bypass capabilities of the Elastic Fabric Adapter (EFA) are not supported on Windows instances. Although you can attach EFA to your Windows instances, this will just act as a regular Elastic Network Adapter, without the added EFA capabilities. Moreover, it doesn't support the t3a.medium instance type that is being used in the HPC cluster.

Enabling Enhanced Networking with Intel 82599 Virtual Function (VF) interface on the Windows EC2 Instances is incorrect because although you can attach an Intel 82599 Virtual Function (VF) interface on your Windows EC2 Instances to improve its networking capabilities, it doesn't support the t3a.medium instance type that is being used in the HPC cluster.

Using AWS ParallelCluster to deploy and manage the HPC cluster to provide higher bandwidth, higher packet per second (PPS) performance, and lower inter-instance latencies is incorrect because an AWS ParallelCluster is just an AWS-supported open-source cluster management tool that makes it easy for you to deploy and manage High Performance Computing (HPC) clusters on AWS. It does not provide higher bandwidth, higher packet per second (PPS) performance, and lower inter-instance latencies, unlike ENA or EFA.

References:

<https://docs.aws.amazon.com/AWSEC2/latest/UserGuide/enhanced-networking.html>

<https://docs.aws.amazon.com/AWSEC2/latest/UserGuide/efa.html>

### QUESTION 31

A travel photo sharing website is using Amazon S3 to serve high-quality photos to visitors of your website. After a few days, you found out that there are other travel websites linking and using your photos. This resulted in financial losses for your business.

What is the MOST effective method to mitigate this issue?

- A. Configure your S3 bucket to remove public read access and use pre-signed URLs with expiry dates.
- B. Block the IP addresses of the offending websites using NACL.
- C. Use CloudFront distributions for your photos.
- D. Store and privately serve the high-quality photos on Amazon WorkDocs instead.

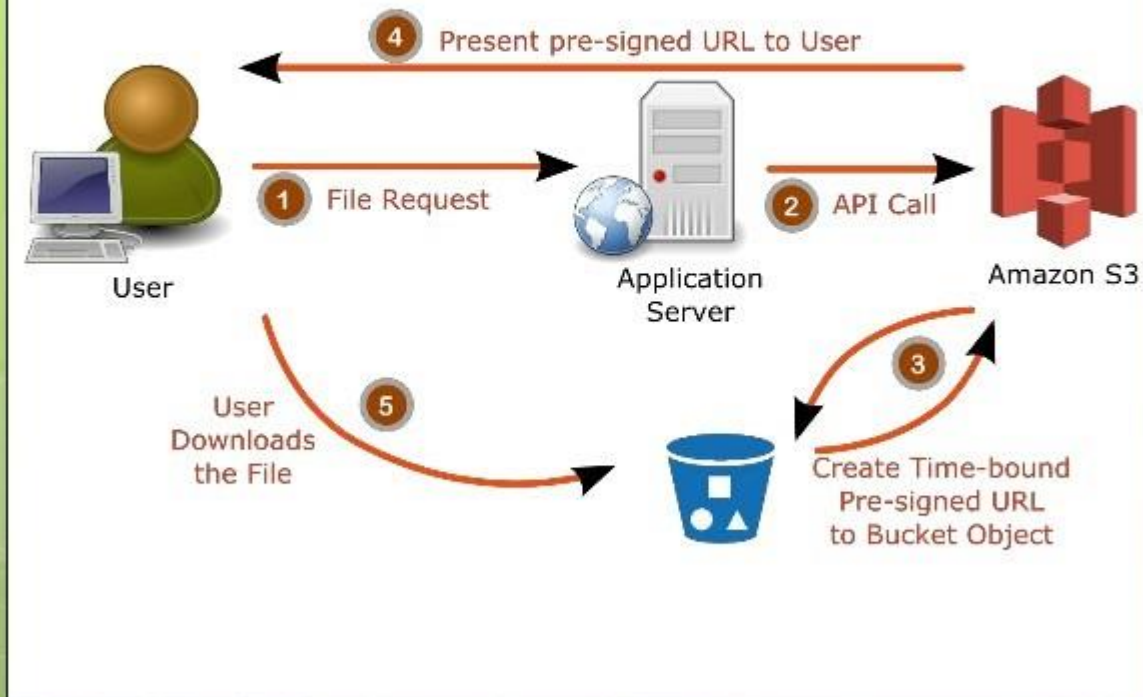
### Correct Answer: A

### Explanation/Reference:

In Amazon S3, all objects are private by default. Only the object owner has permission to access these objects. However, the object owner can optionally share objects with others by creating a pre-signed URL, using their own security credentials, to grant time-limited permission to download the objects. When you create a pre-signed URL for your object, you must provide your security credentials, specify a bucket name, an object key, specify the HTTP method (GET to download the object) and expiration date and time. The pre-signed URLs are valid only for the specified duration.

Anyone who receives the pre-signed URL can then access the object. For example, if you have a video in your bucket and both the bucket and the object are private, you can share the video with others by generating a pre-signed URL.

## Complete Flow



Using CloudFront distributions for your photos is incorrect. CloudFront is a content delivery network service that speeds up delivery of content to your customers.

Blocking the IP addresses of the offending websites using NACL is also incorrect. Blocking IP address using NACLs is not a very efficient method because a quick change in IP address would easily bypass this configuration.

Storing and privately serving the high-quality photos on Amazon WorkDocs instead is incorrect as WorkDocs is simply a fully managed, secure content creation, storage, and collaboration service. It is not a suitable service for storing static content. Amazon WorkDocs is more often used to easily create, edit, and share documents for collaboration and not for serving object data like Amazon S3.

References:

<https://docs.aws.amazon.com/AmazonS3/latest/dev/ShareObjectPreSignedURL.html>

<https://docs.aws.amazon.com/AmazonS3/latest/dev/ObjectOperations.html>

Check out this Amazon CloudFront Cheat Sheet:

<https://tutorialsdojo.com/amazon-cloudfront/>

S3 Pre-signed URLs vs CloudFront Signed URLs vs Origin Access Identity (OAI)

<https://tutorialsdojo.com/s3-pre-signed-urls-vs-cloudfront-signed-urls-vs-origin-access-identity-oai/>

Comparison of AWS Services Cheat Sheets:

<https://tutorialsdojo.com/comparison-of-aws-services/>

## QUESTION 32

A company plans to use a cloud storage service to temporarily store its log files. The number of files to be stored is still unknown, but it only needs to be kept for 12 hours.

Which of the following is the most cost-effective storage class to use in this scenario?

A Amazon S3 Glacier Deep Archive

- B. Amazon S3 Standard
- C. Amazon S3 One Zone-IA
- D. Amazon S3 Standard-IA

**Correct Answer: B**

### Explanation/Reference:

Amazon Simple Storage Service (Amazon S3) is an object storage service that offers industry-leading scalability, data availability, security, and performance. Amazon S3 also offers a range of storage classes for the objects that you store. You choose a class depending on your use case scenario and performance access requirements. All of these storage classes offer high durability.

	Storage class	Designed for	Availability Zones	Min storage duration
<input checked="" type="radio"/>	Standard	Frequently accessed data (more than once a month) with milliseconds access	$\geq 3$	-
<input type="radio"/>	Intelligent-Tiering	Data with changing or unknown access patterns	$\geq 3$	-
<input type="radio"/>	Standard-IA	Infrequently accessed data (once a month) with milliseconds access	$\geq 3$	30 days
<input type="radio"/>	One Zone-IA	Recreatable, infrequently accessed data (once a month) stored in a single Availability Zone with milliseconds access	1	30 days
<input type="radio"/>	Glacier Instant Retrieval	Long-lived archive data accessed once a quarter with instant retrieval in milliseconds	$\geq 3$	90 days
<input type="radio"/>	Glacier Flexible Retrieval (formerly Glacier)	Long-lived archive data accessed once a year with retrieval of minutes to hours	$\geq 3$	90 days
<input type="radio"/>	Glacier Deep Archive	Long-lived archive data accessed less than once a year with retrieval of hours	$\geq 3$	180 days
<input type="radio"/>	Reduced redundancy	Noncritical, frequently accessed data with milliseconds access (not recommended as S3 Standard is more cost effective)	$\geq 3$	-

The scenario requires you to select a cost-effective service that does not have a minimum storage duration since the data will only last for 12 hours. Among the options given, only Amazon S3 Standard has the feature of no minimum storage duration. It is also the most cost-effective storage service because you will only be charged for the last 12 hours, unlike in other storage classes where you will still be charged based on its respective storage duration (e.g. 30 days, 90 days, 180 days). S3

Intelligent-Tiering also has no minimum storage duration and this is designed for data with changing or unknown access patterns.

S3 Standard-IA is designed for long-lived but infrequently accessed data that is retained for months or years. Data that is deleted from S3 Standard-IA within 30 days will still be charged for a full 30 days.

S3 Glacier Deep Archive is designed for long-lived but rarely accessed data that is retained for 7-10 years or more. Objects that are archived to S3 Glacier Deep Archive have a minimum of 180 days of storage, and objects deleted before 180 days incur a pro-rated charge equal to the storage charge for the remaining days.

Hence, the correct answer is: Amazon S3 Standard.

Amazon S3 Standard-IA is incorrect because this storage class has a minimum storage duration of at least 30 days. Remember that the scenario requires the data to be kept for 12 hours only.

Amazon S3 One Zone-IA is incorrect. Just like S3 Standard-IA, this storage class has a minimum storage duration of at least 30 days.

Amazon S3 Glacier Deep Archive is incorrect. Although it is the most cost-effective storage class among all other options, it has a minimum storage duration of at least 180 days which is only suitable for backup and data archival. If you store your data in Glacier Deep Archive for only 12 hours, you will still be charged for the full 180 days.

References:

<https://docs.aws.amazon.com/AmazonS3/latest/dev/storage-class-intro.html>

<https://aws.amazon.com/s3/storage-classes/>

Check out this Amazon S3 Cheat Sheet:

<https://tutorialsdojo.com/amazon-s3/>

S3 Standard vs S3 Standard-IA vs S3 One Zone-IA Cheat Sheet:

<https://tutorialsdojo.com/s3-standard-vs-s3-standard-ia-vs-s3-one-zone-ia/>

### QUESTION 33

A solutions architect is managing an application that runs on a Windows EC2 instance with an attached Amazon FSx for Windows File Server. To save cost, management has decided to stop the instance during off-hours and restart it only when needed. It has been observed that the application takes several minutes to become fully operational which impacts productivity.

How can the solutions architect speed up the instance's loading time without driving the cost up?

- A. Migrate the application to a Linux-based EC2 instance.
- B. Migrate the application to an EC2 instance with hibernation enabled.
- C. Disable the Instance Metadata Service to reduce the things that need to be loaded at startup.
- D. Enable the hibernation mode on the EC2 instance.

### Correct Answer: B

### Explanation/Reference:

Hibernation provides the convenience of pausing and resuming the instances, saves time by reducing the startup time taken by applications, and saves effort in setting up the environment or applications all over again. Instead of having to rebuild the memory footprint, hibernation allows applications to pick up exactly where they left off.



The screenshot displays the AWS Management Console for configuring a new EC2 instance. On the left, the 'Stop - Hibernate behavior' dropdown menu is highlighted with a yellow border and set to 'Enable'. Below it, a note states: 'To enable hibernation, space is allocated on the root volume to store the instance memory (RAM). Make sure that the root volume is large enough to store the RAM contents and accommodate your expected usage, e.g. OS, applications. To use hibernation, the root volume must be an encrypted EBS volume. [Learn more](#)'. Other configuration options include IAM instance profile, Hostname type (IP name), DNS Hostname (with checkboxes for IPv4 and IPv6), Instance auto-recovery, Shutdown behavior, and Termination protection. On the right, the 'Summary' sidebar shows the instance configuration: 1 instance, Amazon Linux 2 Kernel 5.10 AMI, t2.micro instance type, new security group, and 1 volume of 8 GiB. A 'Free tier' notification indicates that the first year includes 750 hours of t2.micro or t3.micro usage in certain regions. At the bottom right, there are 'Cancel' and 'Launch instance' buttons.

While the instance is in hibernation, you pay only for the EBS volumes and Elastic IP Addresses attached to it; there are no other hourly charges (just like any other stopped instance). Therefore, the correct answer is: Migrate the application to an EC2 instance with hibernation enabled. The option that says: Migrate the application to a Linux-based EC2 instance is incorrect. This does not guarantee a faster load time. Moreover, it is a risky thing to do as the application might have dependencies tied to the previous operating system that won't work on a different OS. The option that says: Enable the hibernation mode on the EC2 instance is incorrect. It is not possible to enable or disable hibernation for an instance after it has been launched. The option that says: Disable the instance metadata service to reduce the things that need to be loaded at startup is incorrect. This won't affect the startup load time at all. The Instance Metadata Service is just a service that you can access over the network from within an EC2 instance.

References:

<https://aws.amazon.com/about-aws/whats-new/2019/10/amazon-ec2-hibernation-now-available-on-windows/>

<https://docs.aws.amazon.com/AWSEC2/latest/UserGuide/enabling-hibernation.html>

<https://aws.amazon.com/blogs/aws/new-hibernate-your-ec2-instances/>

Check out this Amazon EC2 Cheat sheet:

<https://tutorialsdojo.com/amazon-elastic-compute-cloud-amazon-ec2/>

## QUESTION 34

A company needs to implement a solution that will process real-time streaming data of its users across the globe. This will enable them to track and analyze globally-distributed user activity on their website and mobile applications, including clickstream analysis. The solution should process the data in close geographical proximity to their users and respond to user requests at low latencies. Which of the following is the most suitable solution for this scenario?

- A. Use a CloudFront web distribution and Route 53 with a Geoproximity routing policy in order to process the data in close geographical proximity to users and respond to user requests at low latencies. Process real-time streaming data using Kinesis and durably store the results to an Amazon S3 bucket.
- B. Use a CloudFront web distribution and Route 53 with a latency-based routing policy, in order to

process the data in close geographical proximity to users and respond to user requests at low latencies. Process real-time streaming data using Kinesis and durably store the results to an Amazon S3 bucket.

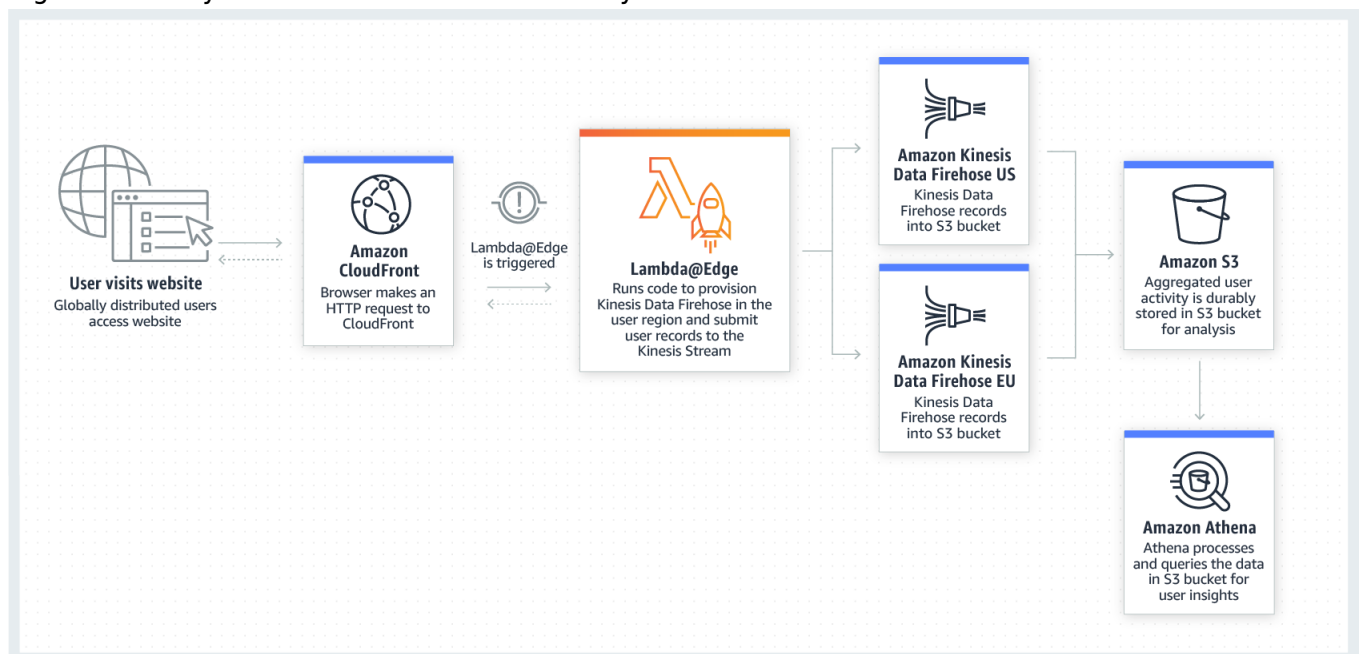
- C. Integrate CloudFront with Lambda@Edge in order to process the data in close geographical proximity to users and respond to user requests at low latencies. Process real-time streaming data using Amazon Athena and durably store the results to an Amazon S3 bucket.
- D. Integrate CloudFront with Lambda@Edge in order to process the data in close geographical proximity to users and respond to user requests at low latencies. Process real-time streaming data using Kinesis and durably store the results to an Amazon S3 bucket.

## Correct Answer: D

### Explanation/Reference:

Lambda@Edge is a feature of Amazon CloudFront that lets you run code closer to users of your application, which improves performance and reduces latency. With Lambda@Edge, you don't have to provision or manage infrastructure in multiple locations around the world. You pay only for the compute time you consume - there is no charge when your code is not running.

With Lambda@Edge, you can enrich your web applications by making them globally distributed and improving their performance — all with zero server administration. Lambda@Edge runs your code in response to events generated by the Amazon CloudFront content delivery network (CDN). Just upload your code to AWS Lambda, which takes care of everything required to run and scale your code with high availability at an AWS location closest to your end user.



By using Lambda@Edge and Kinesis together, you can process real-time streaming data so that you can track and analyze globally-distributed user activity on your website and mobile applications, including clickstream analysis. Hence, the correct answer in this scenario is the option that says: Integrate CloudFront with Lambda@Edge in order to process the data in close geographical proximity to users and respond to user requests at low latencies. Process real-time streaming data using Kinesis and durably store the results to an Amazon S3 bucket.

The options that say: Use a CloudFront web distribution and Route 53 with a latency-based routing policy, in order to process the data in close geographical proximity to users and respond to user



requests at low latencies. Process real-time streaming data using Kinesis and durably store the results to an Amazon S3 bucket and Use a CloudFront web distribution and Route 53 with a Geoproximity routing policy in order to process the data in close geographical proximity to users and respond to user requests at low latencies. Process real-time streaming data using Kinesis and durably store the results to an Amazon S3 bucket are both incorrect because you can only route traffic using Route 53 since it does not have any computing capability. This solution would not be able to process and return the data in close geographical proximity to your users since it is not using Lambda@Edge.

The option that says: Integrate CloudFront with Lambda@Edge in order to process the data in close geographical proximity to users and respond to user requests at low latencies. Process real-time streaming data using Amazon Athena and durably store the results to an Amazon S3 bucket is incorrect because although using Lambda@Edge is correct, Amazon Athena is just an interactive query service that enables you to easily analyze data in Amazon S3 using standard SQL. Kinesis should be used to process the streaming data in real-time.

References:

<https://aws.amazon.com/lambda/edge/>

<https://aws.amazon.com/blogs/networking-and-content-delivery/global-data-ingestion-with-amazon-cloudfront-and-lambdaedge/>

## QUESTION 35

An online shopping platform is hosted on an Auto Scaling group of On-Demand EC2 instances with a default Auto Scaling termination policy and no instance protection configured. The system is deployed across three Availability Zones in the US West region (us-west-1) with an Application Load Balancer in front to provide high availability and fault tolerance for the shopping platform. The us-west-1a, us-west-1b, and us-west-1c Availability Zones have 10, 8 and 7 running instances respectively. Due to the low number of incoming traffic, the scale-in operation has been triggered.

Which of the following will the Auto Scaling group do to determine which instance to terminate first in this scenario? (Select THREE.)

- A. Choose the Availability Zone with the least number of instances, which is the us-west-1c Availability Zone in this scenario.
- B. Choose the Availability Zone with the most number of instances, which is the us-west-1a Availability Zone in this scenario.
- C. Select the instances with the oldest launch configuration.
- D. Select the instances with the most recent launch configuration.
- E. Select the instance that is closest to the next billing hour.
- F. Select the instance that is farthest to the next billing hour.

**Correct Answer: B,C,E**

## Explanation/Reference:

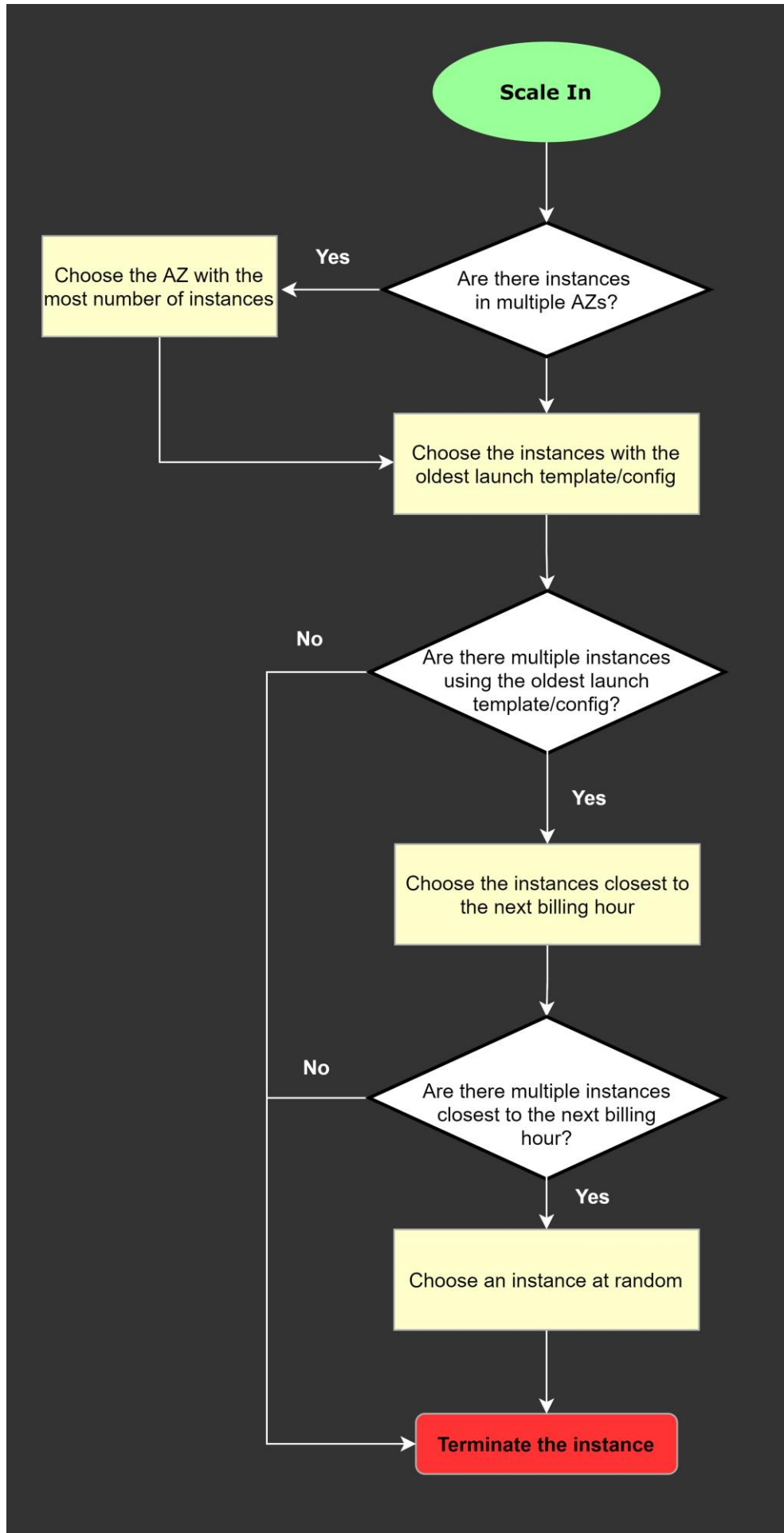
The default termination policy is designed to help ensure that your network architecture spans Availability Zones evenly. With the default termination policy, the behavior of the Auto Scaling group is as follows:

1. If there are instances in multiple Availability Zones, choose the Availability Zone with the most instances and at least one instance that is not protected from scale in. If there is more than one Availability Zone with this number of instances, choose the Availability Zone with the instances that

use the oldest launch configuration.

2. Determine which unprotected instances in the selected Availability Zone use the oldest launch configuration. If there is one such instance, terminate it.
3. If there are multiple instances to terminate based on the above criteria, determine which unprotected instances are closest to the next billing hour. (This helps you maximize the use of your EC2 instances and manage your Amazon EC2 usage costs.) If there is one such instance, terminate it.
4. If there is more than one unprotected instance closest to the next billing hour, choose one of these instances at random.

The following flow diagram illustrates how the default termination policy works:



Reference:

<https://docs.aws.amazon.com/autoscaling/ec2/userguide/as-instance-termination.html#default-termination-policy>

Check out this AWS Auto Scaling Cheat Sheet:

<https://tutorialsdojo.com/aws-auto-scaling/>

### QUESTION 36

A startup launched a new FTP server using an On-Demand EC2 instance in a newly created VPC with default settings. The server should not be accessible publicly but only through the IP address 175.45.116.100 and nowhere else.

Which of the following is the most suitable way to implement this requirement?

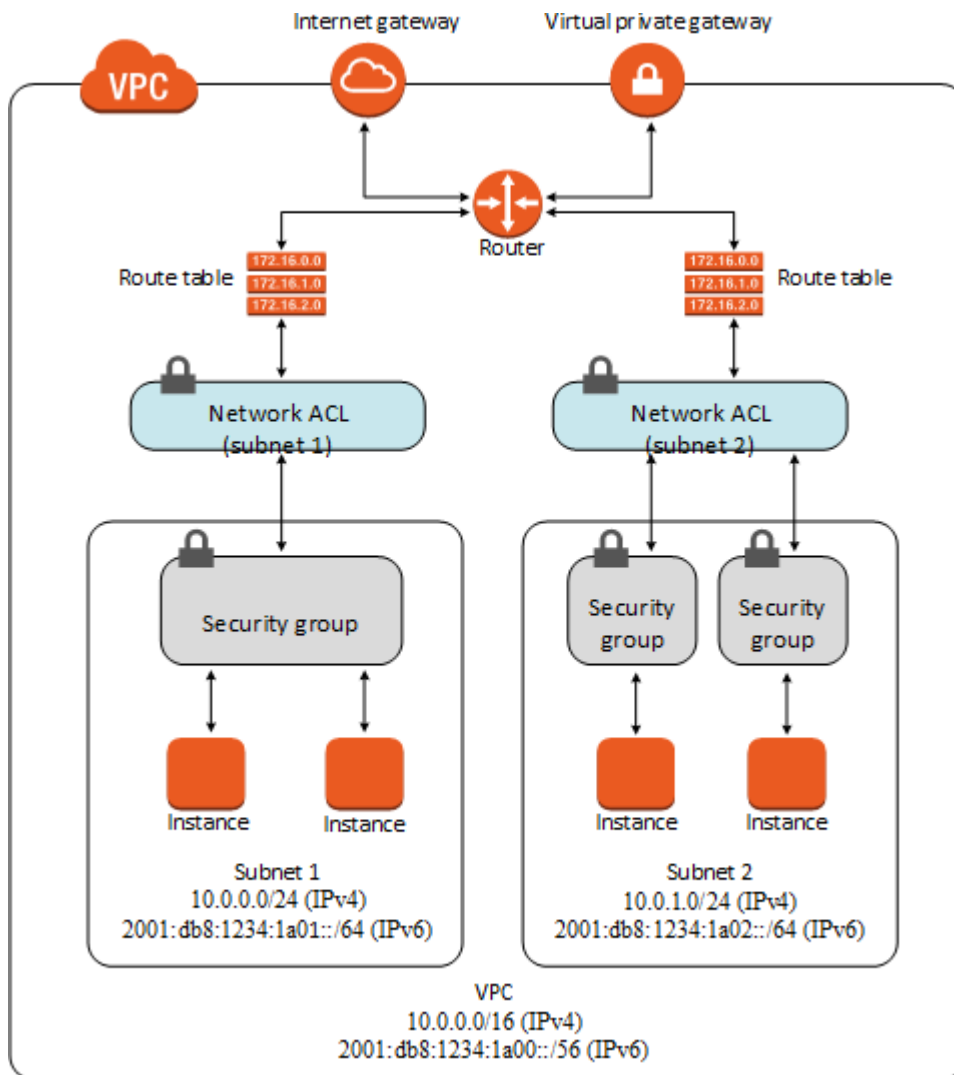
- A. Create a new Network ACL inbound rule in the subnet of the EC2 instance with the following details: Protocol: UDP Port Range: 20 - 21 Source: 175.45.116.100/0 Allow/Deny: ALLOW
- B. Create a new inbound rule in the security group of the EC2 instance with the following details: Protocol: TCP Port Range: 20 - 21 Source: 175.45.116.100/32
- C. Create a new Network ACL inbound rule in the subnet of the EC2 instance with the following details: Protocol: TCP Port Range: 20 - 21 Source: 175.45.116.100/0 Allow/Deny: ALLOW
- D. Create a new inbound rule in the security group of the EC2 instance with the following details: Protocol: UDP Port Range: 20 - 21 Source: 175.45.116.100/32

### Correct Answer: A

### Explanation/Reference:

The FTP protocol uses TCP via ports 20 and 21. This should be configured in your security groups or in your Network ACL inbound rules. As required by the scenario, you should only allow the individual IP of the client and not the entire network. Therefore, in the Source, the proper CIDR notation should be used. The /32 denotes one IP address and the /0 refers to the entire network.

It is stated in the scenario that you launched the EC2 instances in a newly created VPC with default settings. Your VPC automatically comes with a modifiable default network ACL. By default, it allows all inbound and outbound IPv4 traffic and, if applicable, IPv6 traffic. Hence, you actually don't need to explicitly add inbound rules to your Network ACL to allow inbound traffic, if your VPC has a default setting.



The below option is incorrect:

Create a new inbound rule in the security group of the EC2 instance with the following details:

Protocol: UDP

Port Range: 20 - 21

Source: 175.45.116.100/32

Although the configuration of the Security Group is valid, the provided Protocol is incorrect. Take note that FTP uses TCP and not UDP.

The below option is also incorrect:

Create a new Network ACL inbound rule in the subnet of the EC2 instance with the following details:

Protocol: TCP

Port Range: 20 - 21

Source: 175.45.116.100/0

Allow/Deny: ALLOW

Although setting up an inbound Network ACL is valid, the source is invalid since it must be an IPv4 or IPv6 CIDR block. In the provided IP address, the /0 refers to the entire network and not a specific IP address. In addition, it is stated in the scenario that the newly created VPC has default settings and by default, the Network ACL allows all traffic. This means that there is actually no need to configure your Network ACL.

Likewise, the below option is also incorrect:

Create a new Network ACL inbound rule in the subnet of the EC2 instance with the following details:

Protocol: UDP

Port Range: 20 - 21

Source: 175.45.116.100/0

Allow/Deny: ALLOW

Just like in the above, the source is also invalid. Take note that FTP uses TCP and not UDP, which is one of the reasons why this option is wrong. In addition, it is stated in the scenario that the newly created VPC has default settings and by default, the Network ACL allows all traffic. This means that there is actually no need to configure your Network ACL.

References:

[https://docs.aws.amazon.com/vpc/latest/userguide/VPC\\_SecurityGroups.html](https://docs.aws.amazon.com/vpc/latest/userguide/VPC_SecurityGroups.html)

<https://docs.aws.amazon.com/vpc/latest/userguide/vpc-network-acls.html>

Check out this Amazon VPC Cheat Sheet:

<https://tutorialsdojo.com/amazon-vpc/>

## QUESTION 37

A company has a web application hosted in their on-premises infrastructure that they want to migrate to AWS cloud. Your manager has instructed you to ensure that there is no downtime while the migration process is on-going. In order to achieve this, your team decided to divert 50% of the traffic to the new application in AWS and the other 50% to the application hosted in their on-premises infrastructure. Once the migration is over and the application works with no issues, a full diversion to AWS will be implemented. The company's VPC is connected to its on-premises network via an AWS Direct Connect connection.

Which of the following are the possible solutions that you can implement to satisfy the above requirement? (Select TWO.)

- A. Use AWS Global Accelerator to divert and proportion the HTTP and HTTPS traffic between the on-premises and AWS-hosted application. Ensure that the on-premises network has an AnyCast static IP address and is connected to your VPC via a Direct Connect Gateway.
- B. Use Route 53 with Weighted routing policy to divert the traffic between the on-premises and AWS-hosted application. Divert 50% of the traffic to the new application in AWS and the other 50% to the application hosted in their on-premises infrastructure.
- C. Use an Application Elastic Load balancer with Weighted Target Groups to divert and proportion the traffic between the on-premises and AWS-hosted application. Divert 50% of the traffic to the new application in AWS and the other 50% to the application hosted in their on-premises infrastructure.
- D. Use a Network Load balancer with Weighted Target Groups to divert the traffic between the on-premises and AWS-hosted application. Divert 50% of the traffic to the new application in AWS and the other 50% to the application hosted in their on-premises infrastructure.
- E. Use Route 53 with Failover routing policy to divert and proportion the traffic between the on-premises and AWS-hosted application. Divert 50% of the traffic to the new application in AWS and the other 50% to the application hosted in their on-premises infrastructure.

**Correct Answer: B,C**

## Explanation/Reference:

Application Load Balancers support Weighted Target Groups routing. With this feature, you will be able to do weighted routing of the traffic forwarded by a rule to multiple target groups. This enables various use cases like blue-green, canary and hybrid deployments without the need for multiple load balancers. It even enables zero-downtime migration between on-premises and cloud or between different compute types like EC2 and Lambda.

Select the rule to edit. Each rule must include one action of type forward, redirect, fixed response.

AlbWt-LB A2 | HTTP:80 (1 rules)

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

RULE ID	IF (all match)	THEN
last arn...3de0d	✓ Requests otherwise not routed	<p>1. Forward to...</p> <p>Target group : Weight (0-999)</p> <p>blue 50 <input type="checkbox"/></p> <p>Traffic distribution 50%</p> <p>green 50 <input type="checkbox"/></p> <p>Traffic distribution 50%</p> <p>Select a target group 0 <input type="checkbox"/></p> <p>► Group-level stickiness <input checked="" type="checkbox"/></p> <p>+ Add action</p>

To divert 50% of the traffic to the new application in AWS and the other 50% to the application, you can also use Route 53 with Weighted routing policy. This will divert the traffic between the on-premises and AWS-hosted application accordingly.

Weighted routing lets you associate multiple resources with a single domain name (tutorialsdodo.com) or subdomain name (portal.tutorialsdodo.com) and choose how much traffic is routed to each resource. This can be useful for a variety of purposes, including load balancing and testing new versions of software. You can set a specific percentage of how much traffic will be allocated to the resource by specifying the weights.

For example, if you want to send a tiny portion of your traffic to one resource and the rest to another resource, you might specify weights of 1 and 255. The resource with a weight of 1 gets 1/256th of the traffic (1/1+255), and the other resource gets 255/256ths (255/1+255).

You can gradually change the balance by changing the weights. If you want to stop sending traffic to a resource, you can change the weight for that record to 0.

When you create a target group in your Application Load Balancer, you specify its target type. This determines the type of target you specify when registering with this target group. You can select the following target types:

1. instance - The targets are specified by instance ID.
2. ip - The targets are IP addresses.
3. Lambda - The target is a Lambda function.



1. Configure Load Balancer
2. Configure Security Settings
3. Configure Security Groups
4. Configure Routing
5. Register Targets
6. Review

## Step 5: Register Targets

Register targets with your target group. If you register a target in an enabled Availability Zone, the load balancer starts routing requests to the targets as soon as the registration process completes and the target passes the initial health checks.

**ip-target-1** (target group)

Specify one or more IP addresses to register as targets

Network ⓘ  
Other private IP address

Availability Zone ⓘ  
all

IP (allowed ranges)

Port ⓘ  
80

Add to list

To be registered

3 total IP addresses.

Clear all ✕

10.1.200.1	: 80	all	private network resource	✕
10.0.100.2	: 80	us-east-1b	private network resource	✕
10.0.100.1	: 80	us-east-1a	private network resource	✕

When the target type is ip, you can specify IP addresses from one of the following CIDR blocks:

- 10.0.0.0/8 (RFC 1918)
- 100.64.0.0/10 (RFC 6598)
- 172.16.0.0/12 (RFC 1918)
- 192.168.0.0/16 (RFC 1918)
- The subnets of the VPC for the target group

These supported CIDR blocks enable you to register the following with a target group: ClassicLink instances, instances in a VPC that is peered to the load balancer VPC, AWS resources that are addressable by IP address and port (for example, databases), and on-premises resources linked to AWS through AWS Direct Connect or a VPN connection.

Take note that you can not specify publicly routable IP addresses. If you specify targets using an instance ID, traffic is routed to instances using the primary private IP address specified in the primary network interface for the instance. If you specify targets using IP addresses, you can route traffic to an instance using any private IP address from one or more network interfaces. This enables multiple applications on an instance to use the same port. Each network interface can have its own security group.

Hence, the correct answers are the following options:

- Use an Application Elastic Load balancer with Weighted Target Groups to divert and proportion the traffic between the on-premises and AWS-hosted application. Divert 50% of the traffic to the new application in AWS and the other 50% to the application hosted in their on-premises infrastructure.
- Use Route 53 with Weighted routing policy to divert the traffic between the on-premises and AWS-hosted application. Divert 50% of the traffic to the new application in AWS and the other 50% to the application hosted in their on-premises infrastructure.

The option that says: Use a Network Load balancer with Weighted Target Groups to divert the traffic between the on-premises and AWS-hosted application. Divert 50% of the traffic to the new application in AWS and the other 50% to the application hosted in their on-premises infrastructure is incorrect because a Network Load balancer doesn't have Weighted Target Groups to divert the traffic between the on-premises and AWS-hosted application.

The option that says: Use Route 53 with Failover routing policy to divert and proportion the traffic between the on-premises and AWS-hosted application. Divert 50% of the traffic to the new application in AWS and the other 50% to the application hosted in their on-premises infrastructure is incorrect because you cannot divert and proportion the traffic between the on-premises and AWS-hosted application using Route 53 with Failover routing policy. This is primarily used if you want to configure active-passive failover to your application architecture.

The option that says: Use AWS Global Accelerator to divert and proportion the HTTP and HTTPS traffic between the on-premises and AWS-hosted application. Ensure that the on-premises network has an AnyCast static IP address and is connected to your VPC via a Direct Connect Gateway is incorrect because although you can control the proportion of traffic directed to each endpoint using AWS Global Accelerator by assigning weights across the endpoints, it is still wrong to use a Direct Connect Gateway and an AnyCast IP address since these are not required at all. You can only associate static IP addresses provided by AWS Global Accelerator to regional AWS resources or endpoints, such as Network Load Balancers, Application Load Balancers, EC2 Instances, and Elastic IP addresses. Take note that a Direct Connect Gateway, per se, doesn't establish a connection from your on-premises network to your Amazon VPCs. It simply enables you to use your AWS Direct Connect connection to connect to two or more VPCs that are located in different AWS Regions.

References:

<http://docs.aws.amazon.com/Route53/latest/DeveloperGuide/routing-policy.html>

<https://aws.amazon.com/blogs/aws/new-application-load-balancer-simplifies-deployment-with-weighted-target-groups/>

<https://docs.aws.amazon.com/elasticloadbalancing/latest/application/load-balancer-target-groups.html>

Check out this Amazon Route 53 Cheat Sheet:

<https://tutorialsdojo.com/amazon-route-53/>

## QUESTION 38

A company has a web-based ticketing service that utilizes Amazon SQS and a fleet of EC2 instances. The EC2 instances that consume messages from the SQS queue are configured to poll the queue as often as possible to keep end-to-end throughput as high as possible. The Solutions Architect noticed that polling the queue in tight loops is using unnecessary CPU cycles, resulting in increased operational costs due to empty responses.

In this scenario, what should the Solutions Architect do to make the system more cost-effective?

- A. Configure Amazon SQS to use short polling by setting the `ReceiveMessageWaitTimeSeconds` to a number greater than zero.
- B. Configure Amazon SQS to use long polling by setting the `ReceiveMessageWaitTimeSeconds` to zero.
- C. Configure Amazon SQS to use short polling by setting the `ReceiveMessageWaitTimeSeconds` to zero.
- D. Configure Amazon SQS to use long polling by setting the `ReceiveMessageWaitTimeSeconds` to a number greater than zero.

## Correct Answer: D

### Explanation/Reference:

In this scenario, the application is deployed in a fleet of EC2 instances that are polling messages from a single SQS queue. Amazon SQS uses short polling by default, querying only a subset of the servers (based on a weighted random distribution) to determine whether any messages are available for inclusion in the response. Short polling works for scenarios that require higher throughput. However, you can also configure the queue to use Long polling instead, to reduce cost.

The `ReceiveMessageWaitTimeSeconds` is the queue attribute that determines whether you are using Short or Long polling. By default, its value is zero which means it is using Short polling. If it is set to a

value greater than zero, then it is Long polling.

Hence, configuring Amazon SQS to use long polling by setting the `ReceiveMessageWaitTimeSeconds` to a number greater than zero is the correct answer.

Quick facts about SQS Long Polling:

- Long polling helps reduce your cost of using Amazon SQS by reducing the number of empty responses when there are no messages available to return in reply to a `ReceiveMessage` request sent to an Amazon SQS queue and eliminating false empty responses when messages are available in the queue but aren't included in the response.
- Long polling reduces the number of empty responses by allowing Amazon SQS to wait until a message is available in the queue before sending a response. Unless the connection times out, the response to the `ReceiveMessage` request contains at least one of the available messages, up to the maximum number of messages specified in the `ReceiveMessage` action.
- Long polling eliminates false empty responses by querying all (rather than a limited number) of the servers. Long polling returns messages as soon any message becomes available.

Reference:

<https://docs.aws.amazon.com/AWSSimpleQueueService/latest/SQSDeveloperGuide/sqs-long-polling.html>

Check out this Amazon SQS Cheat Sheet:

<https://tutorialsdojo.com/amazon-sqs/>

## QUESTION 39

There are a lot of outages in the Availability Zone of your RDS database instance to the point that you have lost access to the database. What could you do to prevent losing access to your database in case that this event happens again?

- A. Make a snapshot of the database
- B. Increase the database instance size
- C. Enabled Multi-AZ failover
- D. Create a read replica

## Correct Answer: C

### Explanation/Reference:

Amazon RDS Multi-AZ deployments provide enhanced availability and durability for Database (DB) Instances, making them a natural fit for production database workloads. For this scenario, enabling Multi-AZ failover is the correct answer. When you provision a Multi-AZ DB Instance, Amazon RDS automatically creates a primary DB Instance and synchronously replicates the data to a standby instance in a different Availability Zone (AZ). Each AZ runs on its own physically distinct, independent infrastructure, and is engineered to be highly reliable.

In case of an infrastructure failure, Amazon RDS performs an automatic failover to the standby (or to a read replica in the case of Amazon Aurora), so that you can resume database operations as soon as the failover is complete.

Making a snapshot of the database allows you to have a backup of your database, but it does not provide immediate availability in case of AZ failure. So this is incorrect.

Increasing the database instance size is not a solution for this problem. Doing this action addresses the need to upgrade your compute capacity but does not solve the requirement of providing access to

your database even in the event of a loss of one of the Availability Zones.

Creating a read replica is incorrect because this simply provides enhanced performance for read-heavy database workloads. Although you can promote a read replica, its asynchronous replication might not provide you the latest version of your database.

Reference:

<https://aws.amazon.com/rds/details/multi-az/>

Check out this Amazon RDS Cheat Sheet:

<https://tutorialsdojo.com/amazon-relational-database-service-amazon-rds/>

Tutorials Dojo's AWS Certified Solutions Architect Associate Exam Study Guide:

<https://tutorialsdojo.com/aws-certified-solutions-architect-associate-saa-c02/>

## QUESTION 40

A startup is building an AI-based face recognition application in AWS, where they store millions of images in an S3 bucket. As the Solutions Architect, you have to ensure that each and every image uploaded to their system is stored without any issues.

What is the correct indication that an object was successfully stored when you put objects in Amazon S3?

- A. You will receive an email from Amazon SNS informing you that the object is successfully stored.
- B. HTTP 200 result code and MD5 checksum.
- C. You will receive an SMS from Amazon SNS informing you that the object is successfully stored.
- D. Amazon S3 has 99.999999999% durability hence, there is no need to confirm that data was inserted.

## Correct Answer: B

### Explanation/Reference:

If you triggered an S3 API call and got HTTP 200 result code and MD5 checksum, then it is considered as a successful upload. The S3 API will return an error code in case the upload is unsuccessful.

The option that says: Amazon S3 has 99.999999999% durability hence, there is no need to confirm that data was inserted is incorrect because although S3 is durable, it is not an assurance that all objects uploaded using S3 API calls will be successful.

The options that say: You will receive an SMS from Amazon SNS informing you that the object is successfully stored and You will receive an email from Amazon SNS informing you that the object is successfully stored are both incorrect because you don't receive an SMS nor an email notification by default, unless you added an event notification.

Reference:

<https://docs.aws.amazon.com/AmazonS3/latest/API/RESTObjectPOST.html>

Check out this Amazon S3 Cheat Sheet:

<https://tutorialsdojo.com/amazon-s3/>