



AWS Certified Solutions
Architect Associate Exam Crash
Course SAA-C02
-Chad Smith



Exam Guide and General Strategies

Exam Guide As A Job Description

Validating an examinee's ability to:

Define a solution using architectural design principles based on customer requirements.

Multiple services can be present in the same question!

Exam Guide As A Job Description

Validating an examinee's ability to:

Provide implementation guidance based on best practices to the organization throughout the lifecycle of the project.

You must understand modular design as well as operational ramifications of architectural decisions

Exam Guide As A Job Description

Recommended AWS Knowledge:

One year of hands-on experience designing available, cost-efficient, fault-tolerant, and scalable distributed systems on AWS

Hands-on means “get those hands dirty”. That said, hands-on is important, but not as important as understanding the architectural design principles and basic deployment patterns to meet requirements.

Exam Guide As A Job Description

Recommended AWS Knowledge:

Hands-on experience using compute, networking, storage, and database AWS services

Hints on which services are in scope for the exam - these are the foundational areas of Infrastructure and Platform services in the AWS ecosystem. Use that personal account to create resources in these categories!

Exam Guide As A Job Description

Recommended AWS Knowledge:

Hands-on experience with AWS deployment and management services

More hints on exam service scope. This points to IAC, monitoring, and compliance-based services.

Exam Guide As A Job Description

Recommended AWS Knowledge:

Ability to identify and define technical requirements for an AWS-based application

This is a hint that you should be a technology professional with some experience, understanding commonly-used technologies outside of AWS.

Exam Guide As A Job Description

Recommended AWS Knowledge:

Ability to identify which AWS services meet a given technical requirement

Back to the Well-Architected Framework. This sort of exercise is common in the whitepapers.

Exam Guide As A Job Description

Recommended AWS Knowledge:

Knowledge of recommended best practices for building secure and reliable applications on the AWS platform

This focuses on two of the Well-Architected pillars, a hint that they will show up in a significant number of questions.

Exam Guide As A Job Description

Recommended AWS Knowledge:

An understanding of the basic architectural principles of building on the AWS cloud

This is a hint to learn the buzzwords and recognize potential answer choices based on the use of one or more principle in a question.

Exam Guide As A Job Description

Recommended AWS Knowledge:

An understanding of the AWS global infrastructure

This should be basic knowledge for anyone who uses AWS, and can be readily learned through documentation.

Exam Guide As A Job Description

Recommended AWS Knowledge:

An understanding of network technologies as they relate to AWS

Networking, networking, networking! Learn the basics of ipv4 and the 7 layer model. Then apply this knowledge to networking services in AWS

Exam Guide As A Job Description

Recommended AWS Knowledge:

An understanding of security features and tools that AWS provides and how they relate to traditional services

Another emphasis on security. Learn those security services, and the security features within each service, as well as security best practices!

Question Domain Focus

Domain 1	Design Resilient Architectures	30%
Domain 2	Design High-Performing Architectures	28%
Domain 3	Design Secure Applications and Architectures	24%
Domain 4	Design Cost-Optimized Architectures	18%



Focus Here!



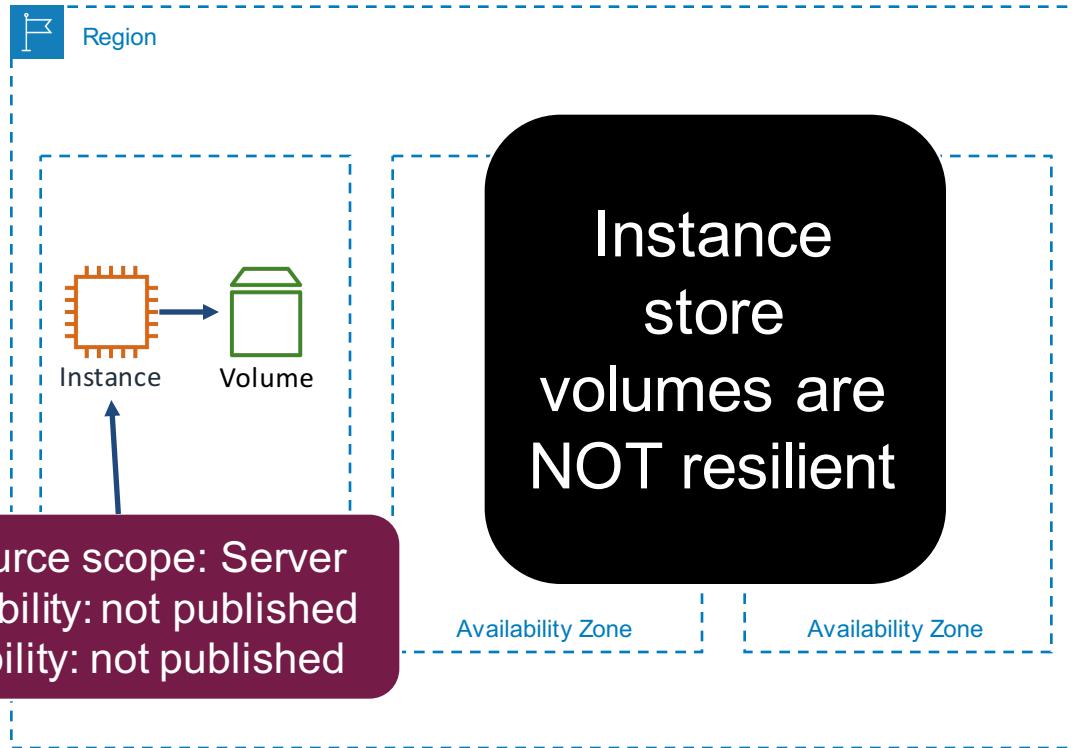
Design Resilient Architectures
30%



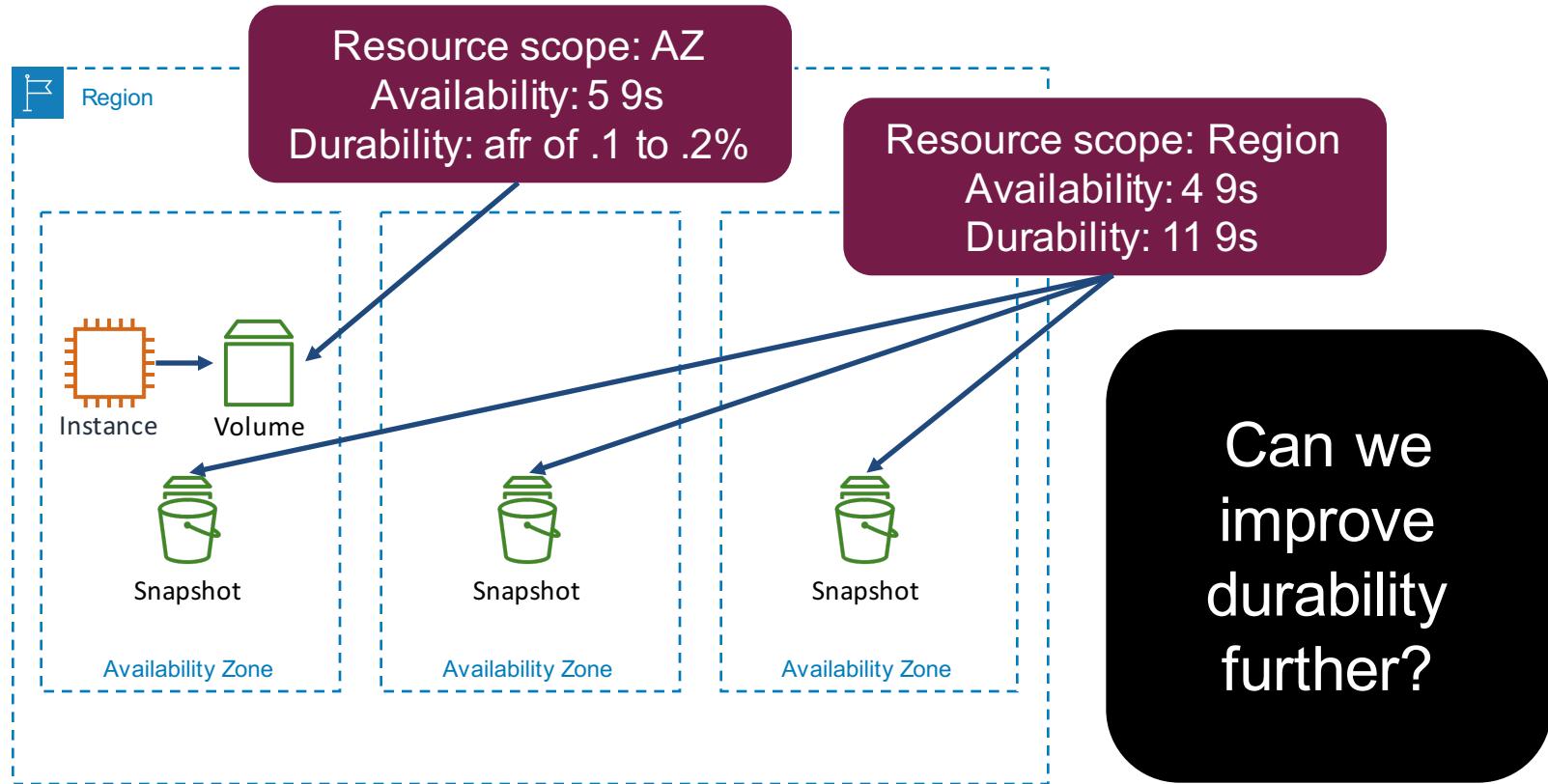
Design Resilient Architectures

Storage Principles

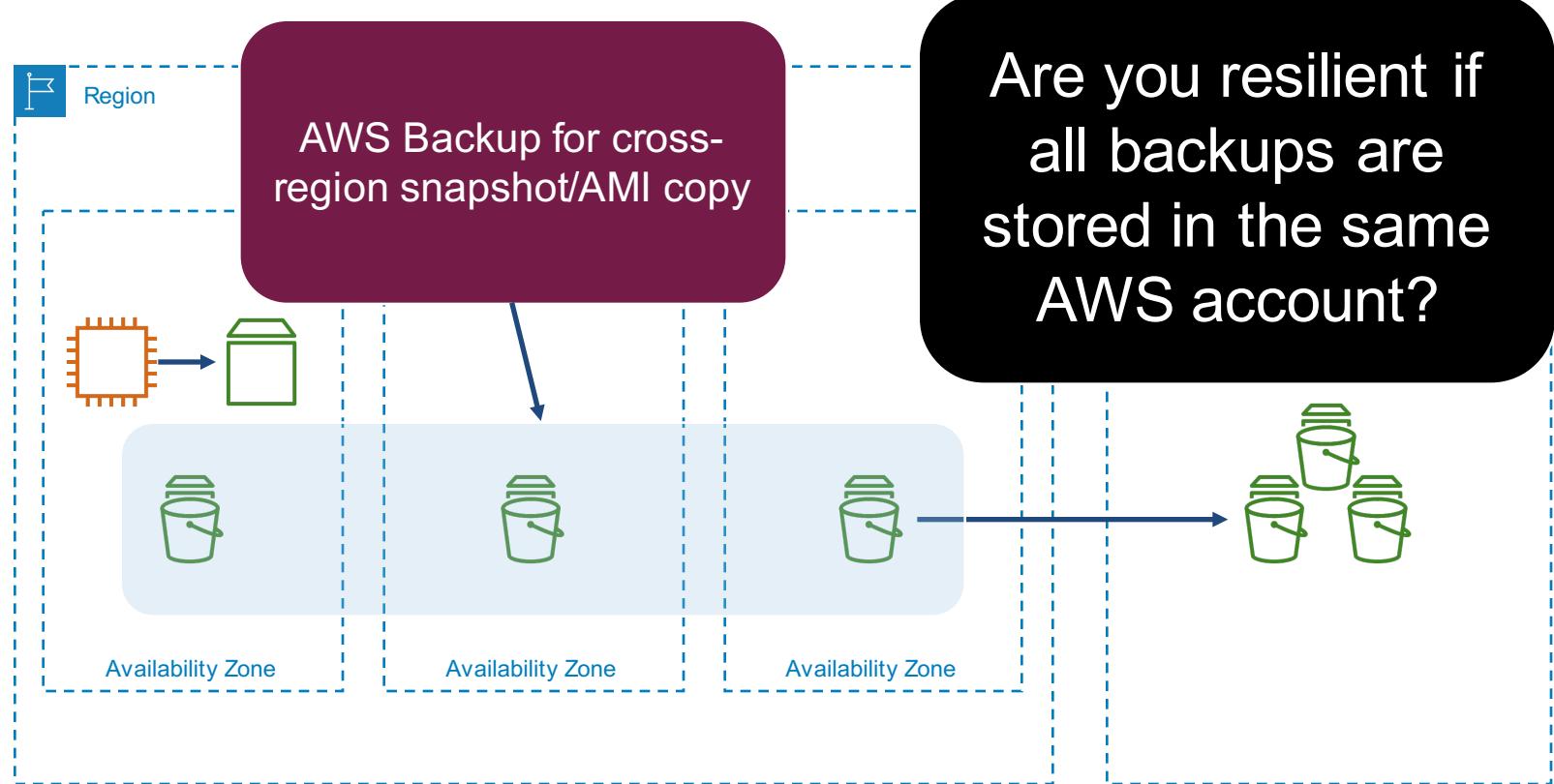
Resilient File Storage - Instance Store



Resilient File Storage - EBS



Resilient File Storage - EBS



Storage Principles

EASY Question Breakdown

Question Breakdown

What is the availability SLA for a single EBS volume?

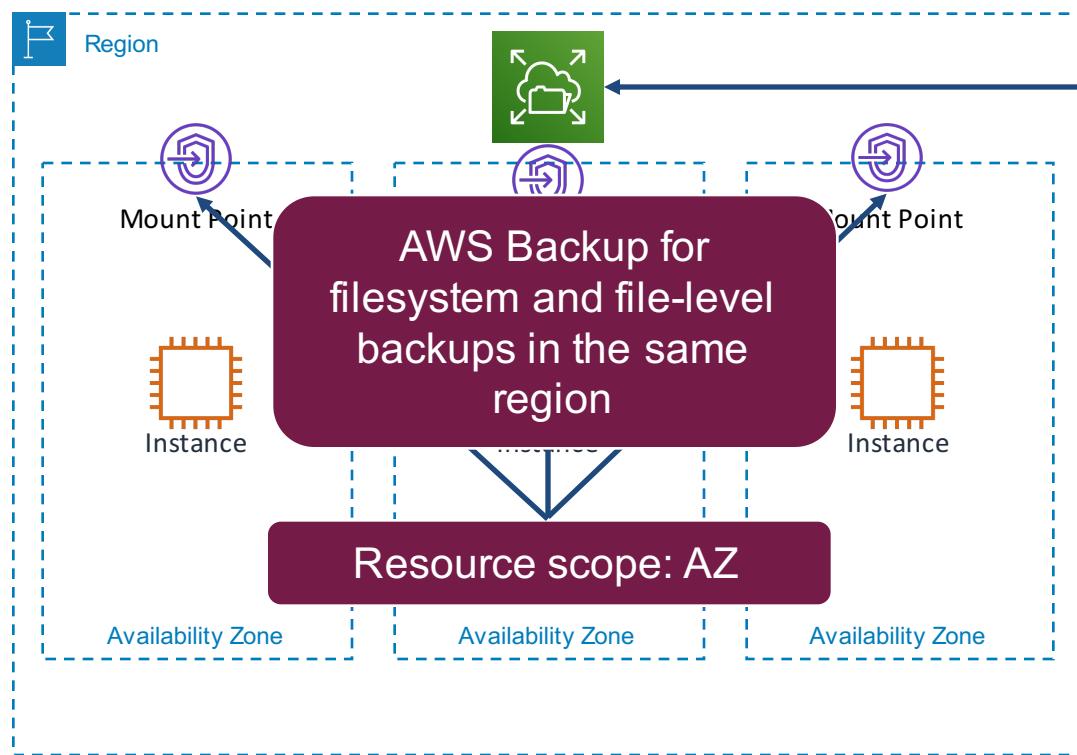
- A. 3 9s
- B. 4 9s
- C. 5 9s
- D. 6 9s

Question Breakdown - Correct Answer

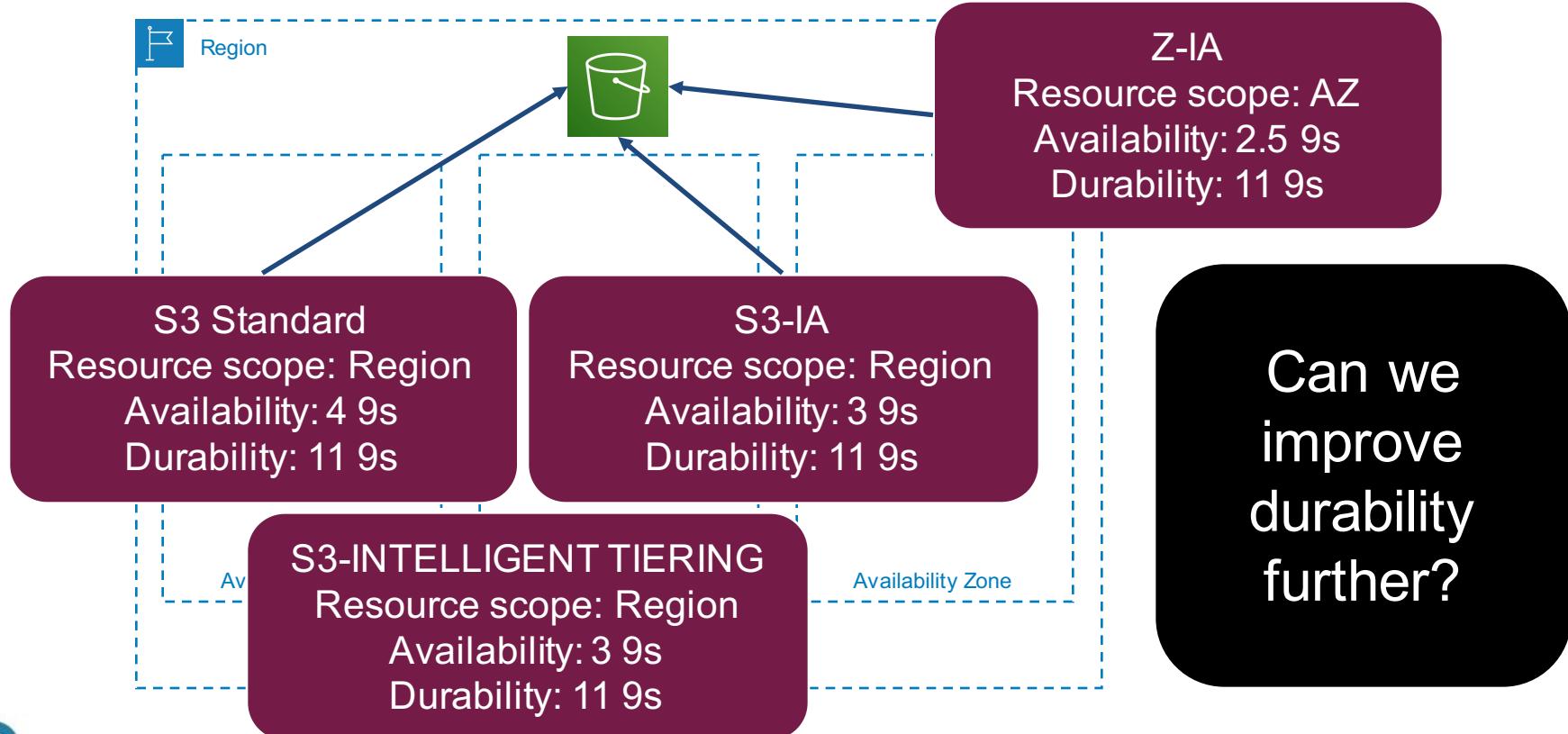
Correct Answer: C

- A. 3 9s
- B. 4 9s
- C. 5 9s
- D. 6 9s

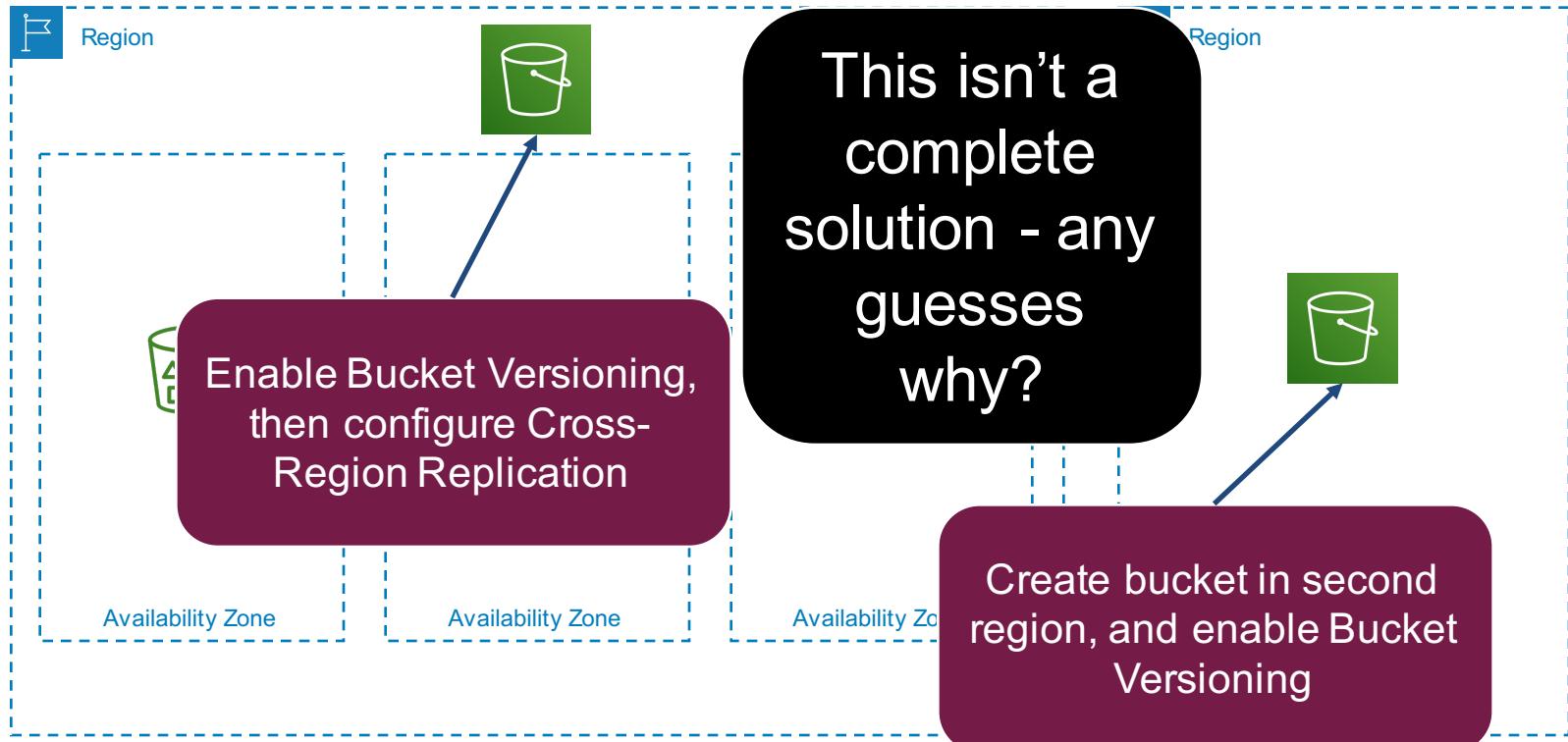
Resilient File Storage - EFS



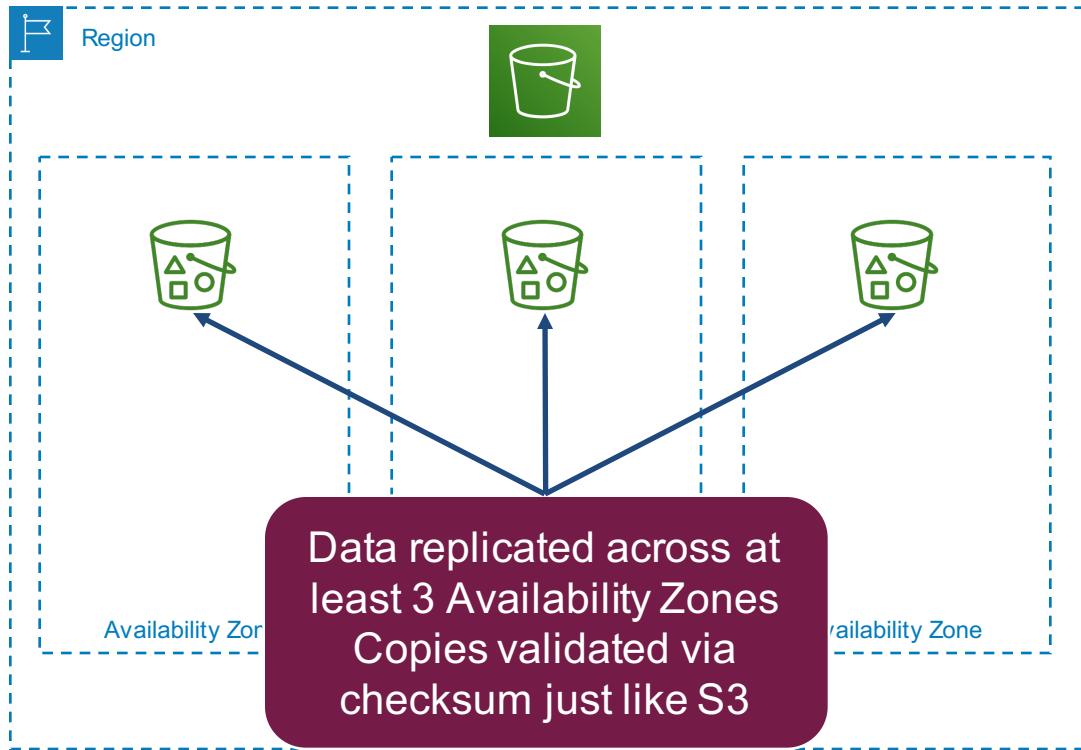
Resilient Object Storage - S3



Resilient Object Storage - S3

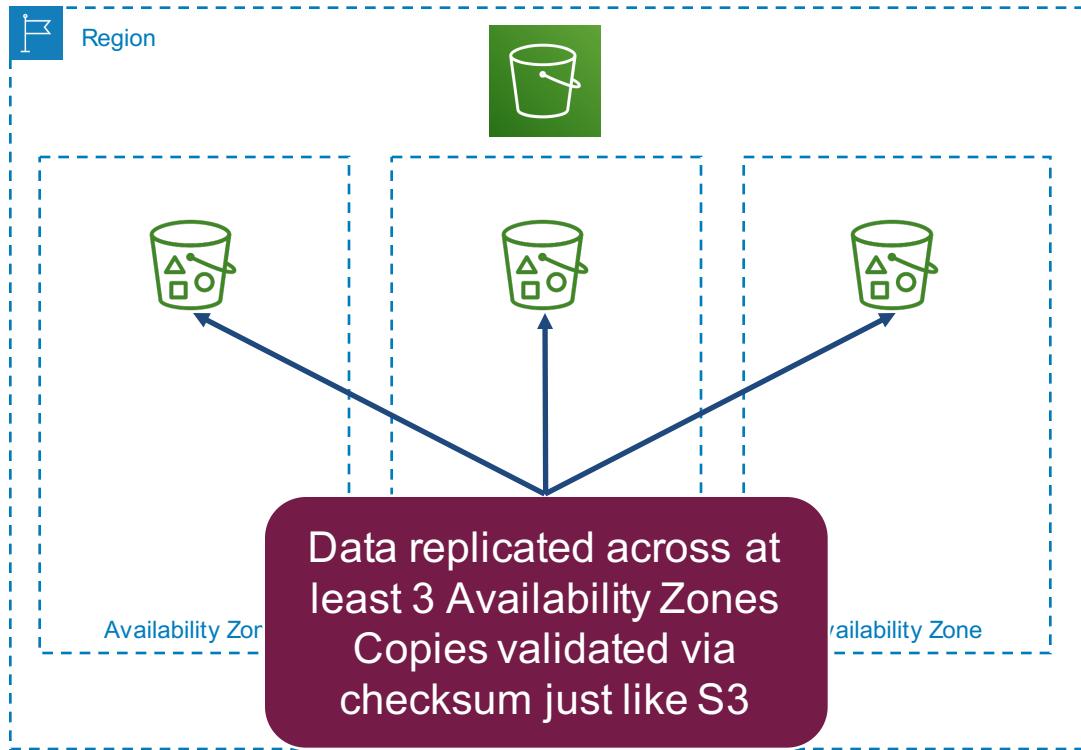


Resilient Object Storage - Glacier



S3 GLACIER
Resource scope: Region
Availability: 4 9s
Durability: 11 9s
Latency: a few minutes to several hours; configurable

Resilient Object Storage - Glacier



S3 GLACIER DEEP ARCHIVE
Resource scope: Region
Availability: 4 9s
Durability: 11 9s
Latency: within 12 hours

Storage Principles

Question Breakdown

Question Breakdown - Key Terms

Your team has been asked to implement an AWS **storage** infrastructure that can support **multiple AZs** within a region. **Multiple EC2 instances** will require access to the data. **High availability** is more important than performance. Which of the following solutions meet the requirements with the **least operational overhead**?

- A. AWS Storage Gateway in volume cache mode. Data stored in S3.
- B. Individual EBS volumes attached to instances. Data downloaded from S3.
- C. GlusterFS installed on all instances with multiple partitions and replicas of data.
- D. EFS volume deployed in the region. Each EC2 instance mounts the volume via NFS.

Question Breakdown - Answers

The Storage Gateway is a single point of failure, thus reducing availability of the infrastructure.

- A. AWS Storage Gateway in volume cache mode. Data stored in S3.
- B. Individual EBS volumes attached to instances. Data downloaded from S3.
- C. GlusterFS installed on all instances with multiple partitions and replicas of data.
- D. EFS volume deployed in the region. Each EC2 instance mounts the volume via NFS.

Question Breakdown - Answers

There is a lot of operational overhead associated with downloading the data from S3 and updating the data on EBS periodically.

- A. AWS Storage Gateway in volume cache mode. Data stored in S3.
- B. Individual EBS volumes attached to instances. Data downloaded from S3.
- C. GlusterFS installed on all instances with multiple partitions and replicas of data.
- D. EFS volume deployed in the region. Each EC2 instance mounts the volume via NFS.

Question Breakdown - Answers

This solution is highly available, but requires the extra overhead of maintaining GlusterFS.

- A. AWS Storage Gateway in volume cache mode. Data stored in S3.
- B. Individual EBS volumes attached to instances. Data downloaded from S3.
- C. GlusterFS installed on all instances with multiple partitions and replicas of data.
- D. EFS volume deployed in the region. Each EC2 instance mounts the volume via NFS.

Question Breakdown - Answers

EFS volumes mount points can be deployed to multiple AZs within a region, and the operational overhead is minimal, as the volume is elastic according to the amount of data present.

- A. AWS Storage Gateway in volume cache mode. Data stored in S3.
- B. Individual EBS volumes attached to instances. Data downloaded from S3.
- C. GlusterFS installed on all instances with multiple partitions and replicas of data.
- D. EFS volume deployed in the region. Each EC2 instance mounts the volume via NFS.

Question Breakdown - Correct Answer

Correct Answer: D

- A. AWS Storage Gateway in volume cache mode. Data stored in S3.
- B. Individual EBS volumes attached to instances. Data downloaded from S3.
- C. GlusterFS installed on all instances with multiple partitions and replicas of data.
- D. EFS volume deployed in the region. Each EC2 instance mounts the volume via NFS.



Design Resilient Architectures

High Availability and
Fault Tolerance

Definitions

- **High Availability** - The system will continue to function despite the complete failure of any component of the architecture
- **Fault Tolerance** - The system will continue to function **without degradation in performance** despite the complete failure of any component of the architecture.
- **Availability** - determined by percentage uptime, in 9s
- **Redundant** - multiple resources dedicated to performing the same task

Availability Documentation

- aws.amazon.com/<service>/sla
- Well-Architected Reliability Pillar Whitepaper, pages 54-58
- Many services have 4 9s availability, but don't assume
- Route 53 health checks can help with arbitrary endpoints

EASY Question Breakdown

Question Breakdown

Which of the following services maintains higher than 5 9s availability SLA?

- A. Route 53
- B. S3
- C. EBS
- D. API Gateway

Question Breakdown - Correct Answer

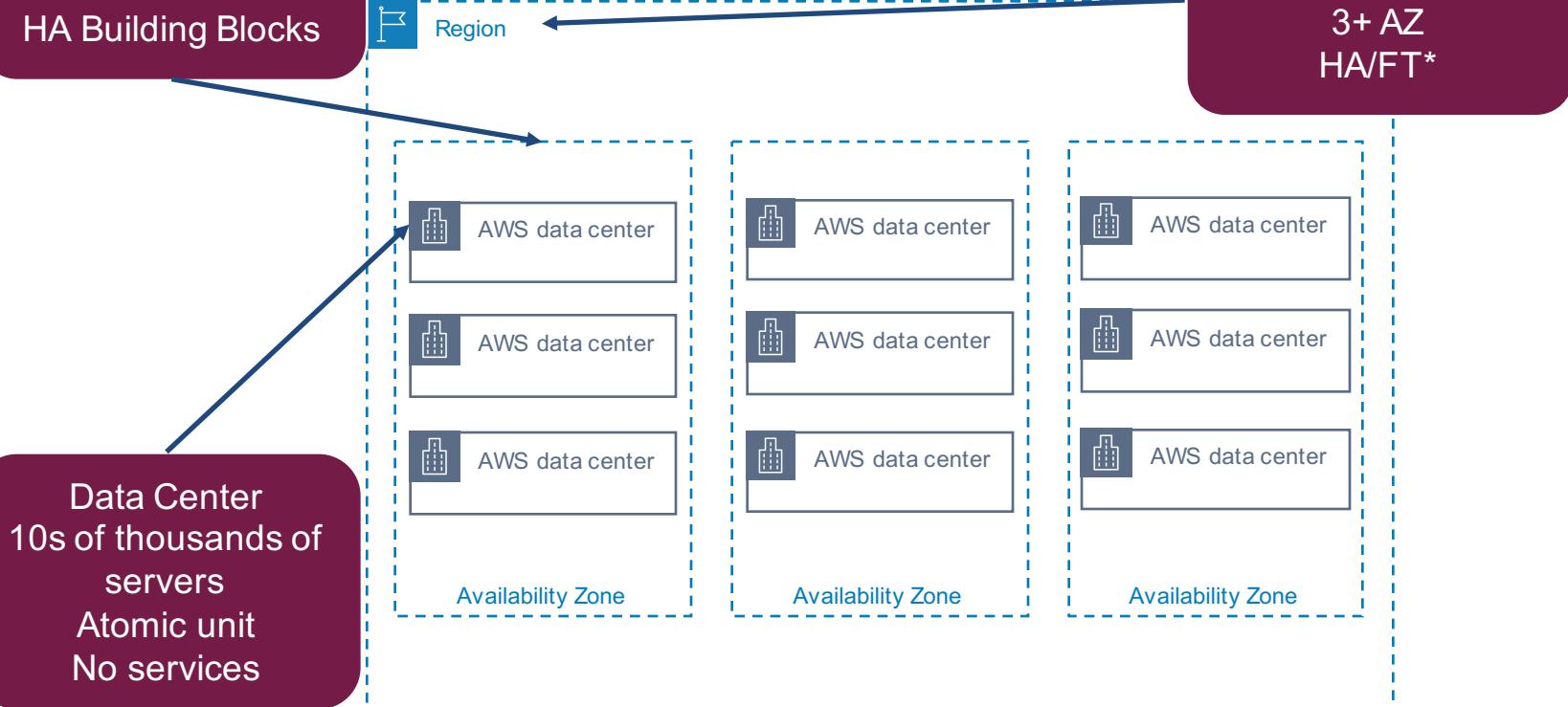
Correct Answer: A

- A. Route 53
- B. S3
- C. EBS
- D. API Gateway

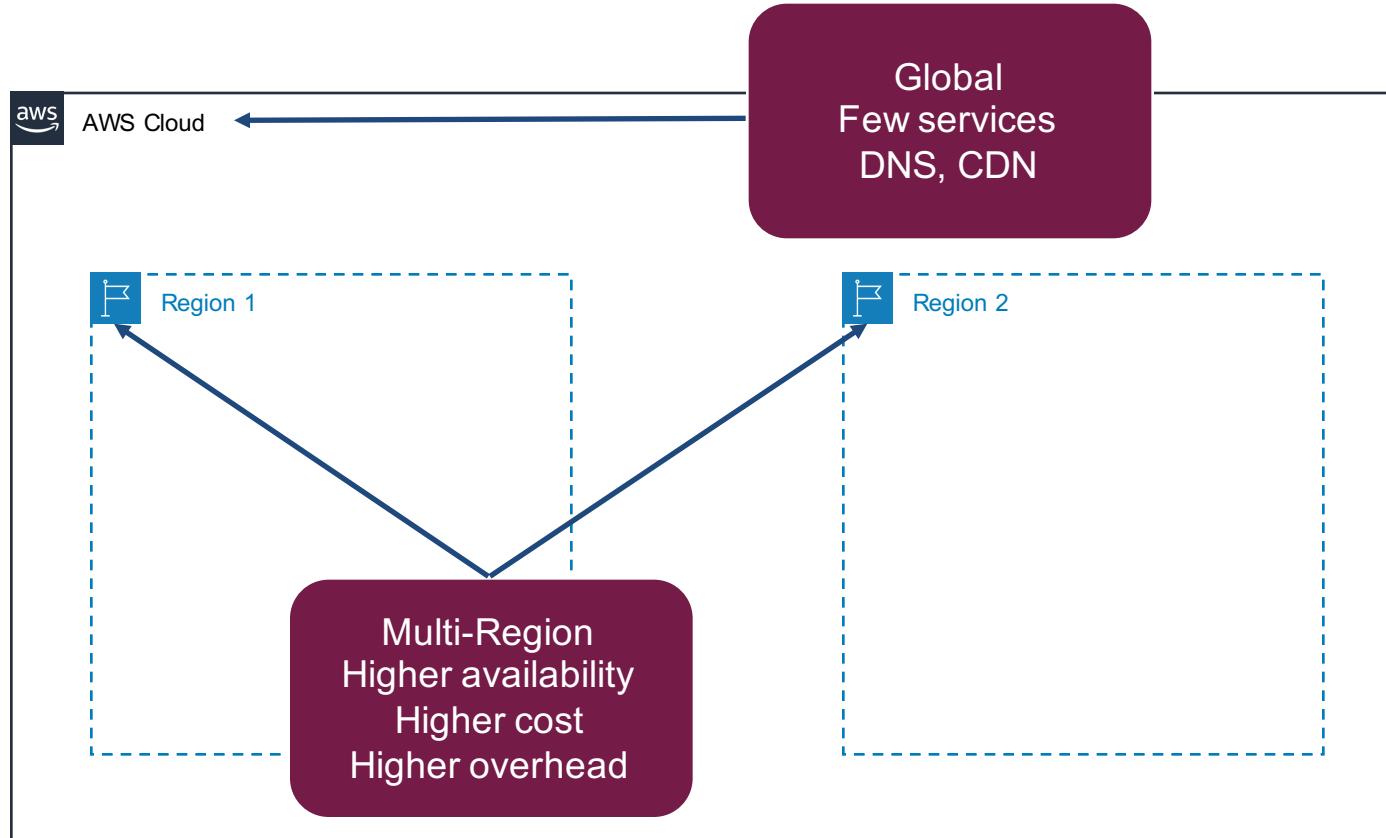
HA/FT - Global Infrastructure

Availability Zone
Failure zone
3+ Data centers
HA Building Blocks

Region
Fully Independent
3+ AZ
HA/FT*



HA/FT - Global Infrastructure



High Availability and Fault Tolerance

Question Breakdown

Question Breakdown - Key Terms

Assuming your application infrastructure has an **availability requirement of 99.99%**, which of the following resilience strategies **would likely NOT achieve** the required uptime?

- A. Deploying the database back end via RDS with Multi-AZ enabled and practicing restores daily.
- B. Deploying infrastructure via CloudFormation templates with disaster plans to re-deploy.
- C. Monitoring on all application layer KPIs with sensitive alarms and early notification, automated mitigation wherever possible.
- D. All web services are hosted behind ALB and use Auto Scaling, both in multiple availability zones.

Question Breakdown - Answers

These tasks will optimize database restore time to a minimum value, usually less than 60 minutes, depending on the size of the database.

- A. Deploying the database back end via RDS with Multi-AZ enabled and practicing restores daily.
- B. Deploying infrastructure via CloudFormation templates with disaster plans to re-deploy.
- C. Monitoring on all application layer KPIs with sensitive alarms and early notification, automated mitigation wherever possible.
- D. All web services are hosted behind ALB and use Auto Scaling, both in multiple availability zones.

Question Breakdown - Answers

CloudFormation is great for automation, and a complete re-deploy of a stack might maintain 4 9s of availability. However, as infrastructure complexity increases, with many dependencies, recreating from scratch will quickly become a problem.

- A. Deploying the database back end via RDS with Multi-AZ enabled and practicing restores daily.
- B. Deploying infrastructure via CloudFormation templates with disaster plans to re-deploy.
- C. Monitoring on all application layer KPIs with sensitive alarms and early notification, automated mitigation wherever possible.
- D. All web services are hosted behind ALB and use Auto Scaling, both in multiple availability zones.

Question Breakdown - Answers

Proper monitoring with fast responses will ensure a high level of availability. Automated mitigation, often in the form of Auto Scaling, will reduce dependency on human actors that can be delayed by many factors and increase risk.

- A. Deploying the database back end via RDS with Multi-AZ enabled and practicing restores daily.
- B. Deploying infrastructure via CloudFormation templates with disaster plans to re-deploy.
- C. Monitoring on all application layer KPIs with sensitive alarms and early notification, automated mitigation wherever possible.
- D. All web services are hosted behind ALB and use Auto Scaling, both in multiple availability zones.

Question Breakdown - Answers

By utilizing a combination of the Application Load Balancer service and Auto Scaling, the infrastructure becomes elastic, making it able to expand and contract according to the traffic requests. This reduces the chances of catastrophic failure due to capacity failure, thus increasing availability.

- A. Deploying the database back end via RDS with Multi-AZ enabled and practicing restores daily.
- B. Deploying infrastructure via CloudFormation templates with disaster plans to re-deploy.
- C. Monitoring on all application layer KPIs with sensitive alarms and early notification, automated mitigation wherever possible.
- D. All web services are hosted behind ALB and use Auto Scaling, both in multiple availability zones.

Question Breakdown - Correct Answer

Correct Answer: B

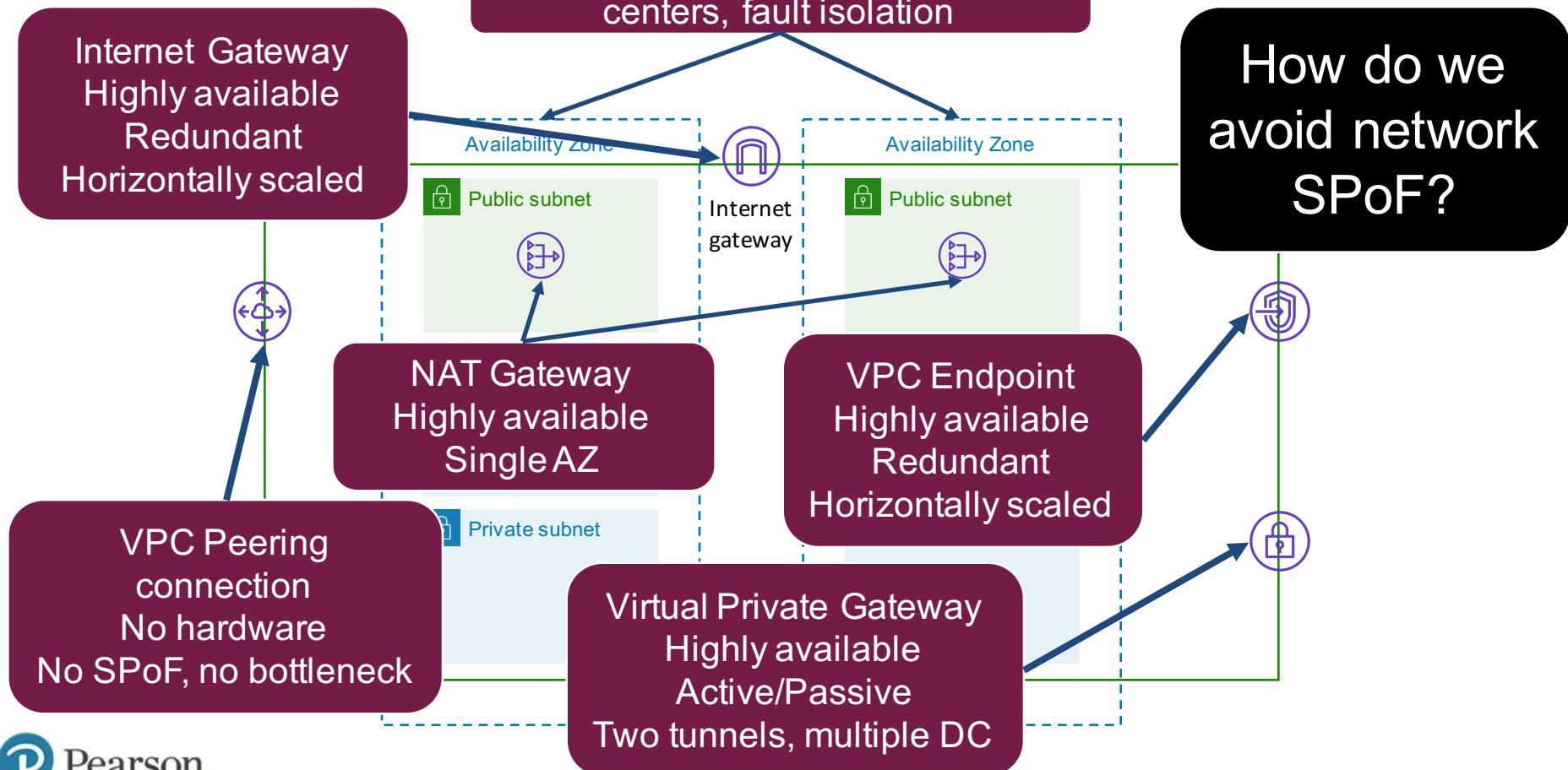
- A. Deploying the database back end via RDS with Multi-AZ enabled and practicing restores daily.
- B. Deploying infrastructure via CloudFormation templates with disaster plans to re-deploy.
- C. Monitoring on all application layer KPIs with sensitive alarms and early notification, automated mitigation wherever possible.
- D. All web services are hosted behind ALB and use Auto Scaling, both in multiple availability zones.



Design Resilient Architectures

Multi-tier Architectures

Multi-Tier Traditional Web/App/DB



EASY Question Breakdown

Question Breakdown

Which of the following resilience strategies could result in increased network throughput charges?

- A. Utilize managed services instead of un-managed
- B. Deploy independent, regional copies of application tiers
- C. Migrate data to region-scoped services such as S3 and DynamoDB
- D. Deploy infrastructure into multiple AZ within a VPC

Question Breakdown - Correct Answer

Correct Answer:

- A. Utilize managed services instead of un-managed
- B. Deploy independent, regional copies of application tiers
- C. Migrate data to region-scoped services such as S3 and DynamoDB
- D. Deploy infrastructure into multiple AZ within a VPC

Multi-Tier Traditional Web/App/DB

ELB
Multi-AZ
Redundant
Availability 4 9s

Auto Scaling
Multi-AZ
Redundant (EC2)
Availability 1 9/EC2

Entire
infrastructure:
Highly available
Self-healing
Automatic Scaling**

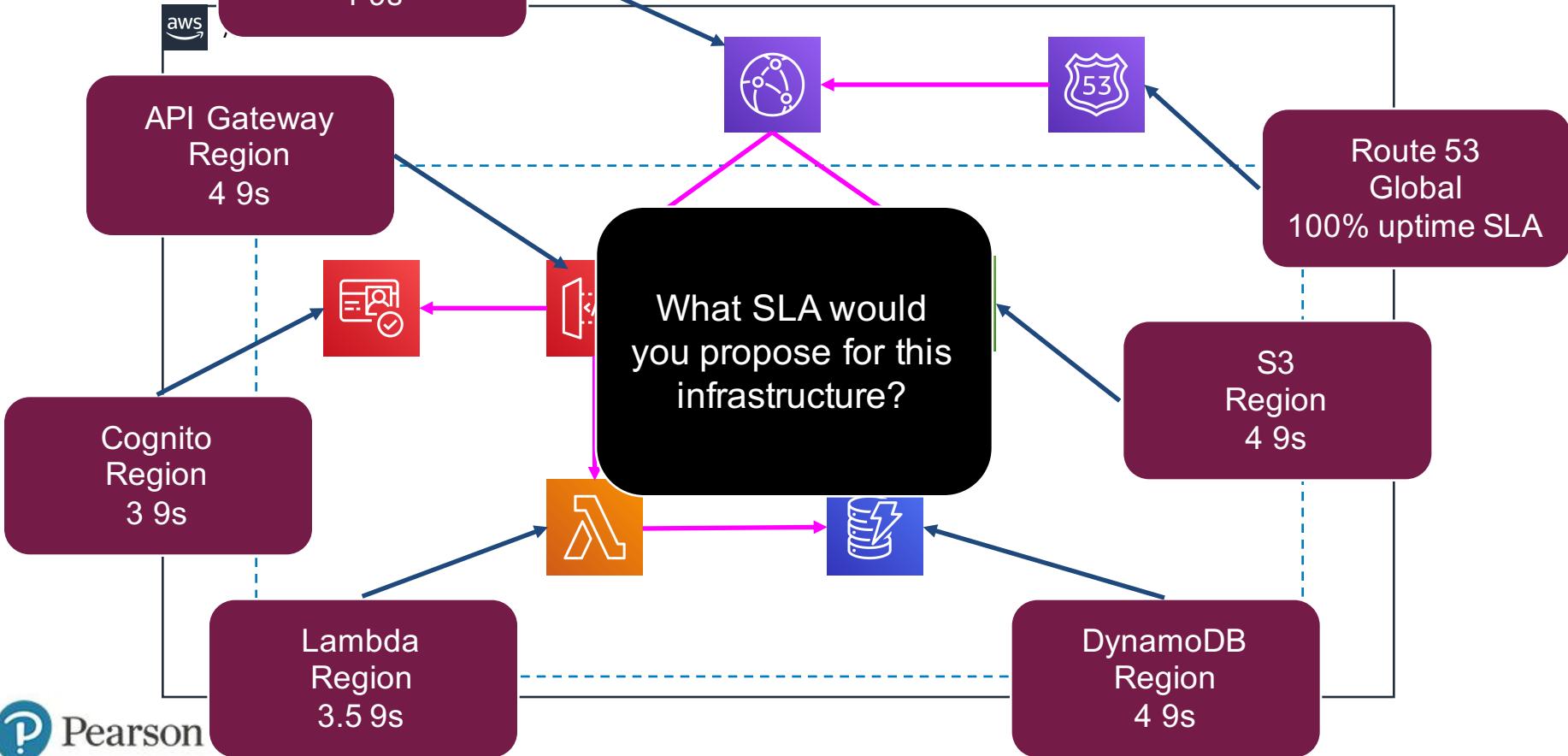
Can we do better?

How do we
avoid
application
SPoF?

RDS/Aurora
Multi-AZ
HA - active/passive writes
Availability 3.5 9s/Node OR

Aurora
Multi-Master
Active/active writes
Availability 4 9s

Serverless



Multi-tier Architectures

Question Breakdown

Question Breakdown - Key Terms

An application is currently hosted on an **EC2 instance** and consists of **static images**, **JavaScript**, **Java code**, and a **MySQL database**. What steps could be performed to improve the **resilience**? (pick two)

- A. Move the database to RDS and enable Multi-AZ.
- B. Resize the EC2 instance to increase memory and CPU.
- C. Move the static images and JavaScript to an EFS volume
- D. Move the static images and JavaScript to an S3 bucket.
- E. Move the static images/JavaScript/Java to one EBS volume, and the database to a second volume

Question Breakdown - Answers

Relocating the database to a managed service like RDS with Multi-AZ enabled will improve the resilience of the infrastructure by splitting architectural tiers.

- A. Move the database to RDS and enable Multi-AZ.
- B. Resize the EC2 instance to increase memory and CPU.
- C. Move the static images and JavaScript to an EFS volume
- D. Move the static images and JavaScript to an S3 bucket.
- E. Move the static images/JavaScript/Java to one EBS volume, and the database to a second volume

Question Breakdown - Answers

Resizing the EC2 instance may improve the overall performance of the application but does not do anything to improve resilience.

- A. Move the database to RDS and enable Multi-AZ.
- B. Resize the EC2 instance to increase memory and CPU.
- C. Move the static images and JavaScript to an EFS volume
- D. Move the static images and JavaScript to an S3 bucket.
- E. Move the static images/JavaScript/Java to one EBS volume, and the database to a second volume

Question Breakdown - Answers

As a standalone task, migrating the static content to an EFS volume does nothing to improve the overall resilience of the application, as there is still a single entry point for all assets.

- A. Move the database to RDS and enable Multi-AZ.
- B. Resize the EC2 instance to increase memory and CPU.
- C. Move the static images and JavaScript to an EFS volume
- D. Move the static images and JavaScript to an S3 bucket.
- E. Move the static images/JavaScript/Java to one EBS volume, and the database to a second volume

Question Breakdown - Answers

Migrating the static assets to an S3 bucket will improve resilience by hosting them on a managed service that is highly available. There is a secondary effect of potential improved performance of the application because the EC2 resources are no longer required for hosting those static assets.

- A. Move the database to RDS and enable Multi-AZ.
- B. Resize the EC2 instance to increase memory and CPU.
- C. Move the static images and JavaScript to an EFS volume
- D. Move the static images and JavaScript to an S3 bucket.
- E. Move the static images/JavaScript/Java to one EBS volume, and the database to a second volume

Question Breakdown - Answers

Splitting the static assets and application code from the database on different EBS volumes will not measurably improve resilience, as EBS is already rated for 5 nines of availability.

- A. Move the database to RDS and enable Multi-AZ.
- B. Resize the EC2 instance to increase memory and CPU.
- C. Move the static images and JavaScript to an EFS volume
- D. Move the static images and JavaScript to an S3 bucket.
- E. Move the static images/JavaScript/Java to one EBS volume, and the database to a second volume

Question Breakdown - Correct Answer

Correct Answers: A and D

- A. Move the database to RDS and enable Multi-AZ.
- B. Resize the EC2 instance to increase memory and CPU.
- C. Move the static images and JavaScript to an EFS volume
- D. Move the static images and JavaScript to an S3 bucket.
- E. Move the static images/JavaScript/Java to one EBS volume, and the database to a second volume

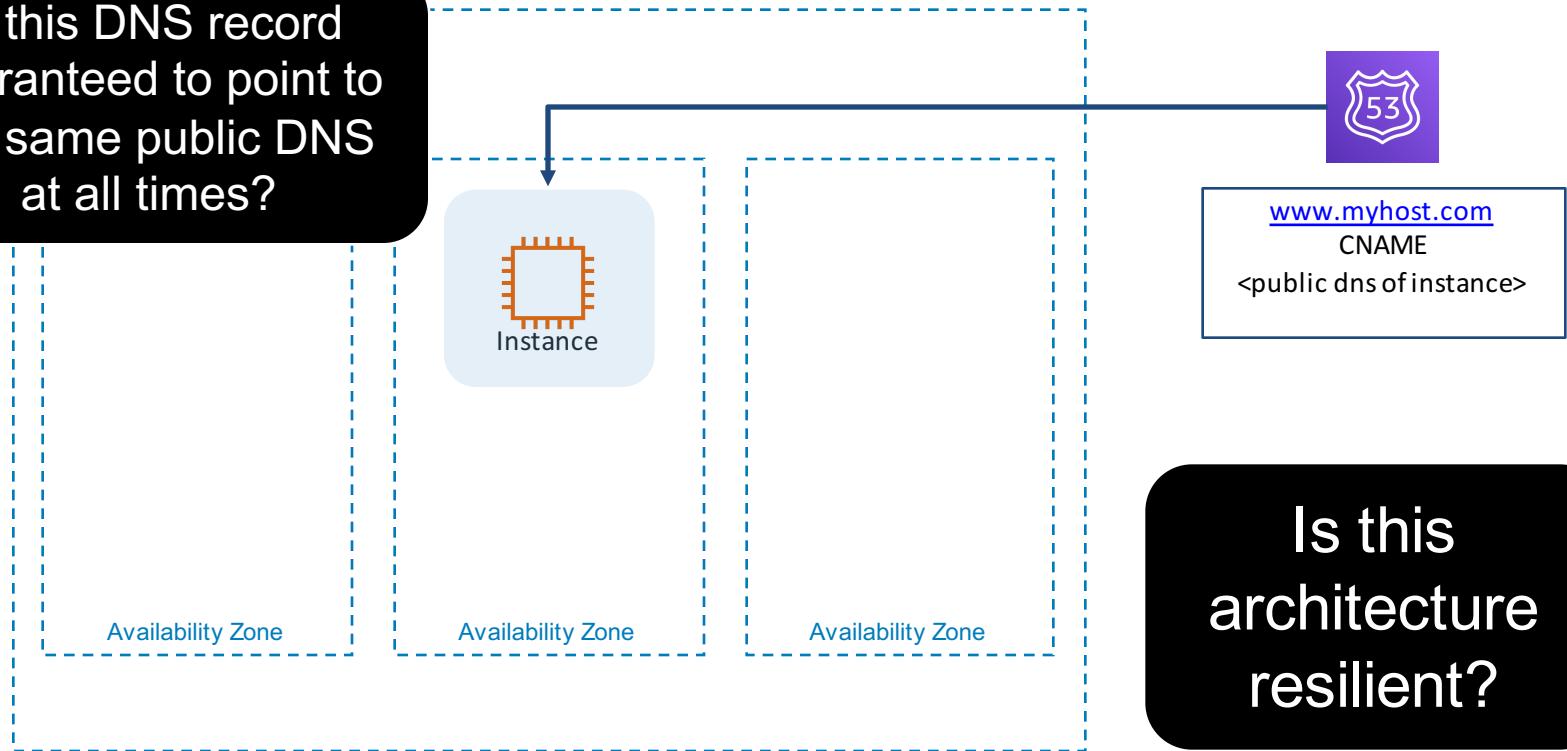


Design Resilient Architectures

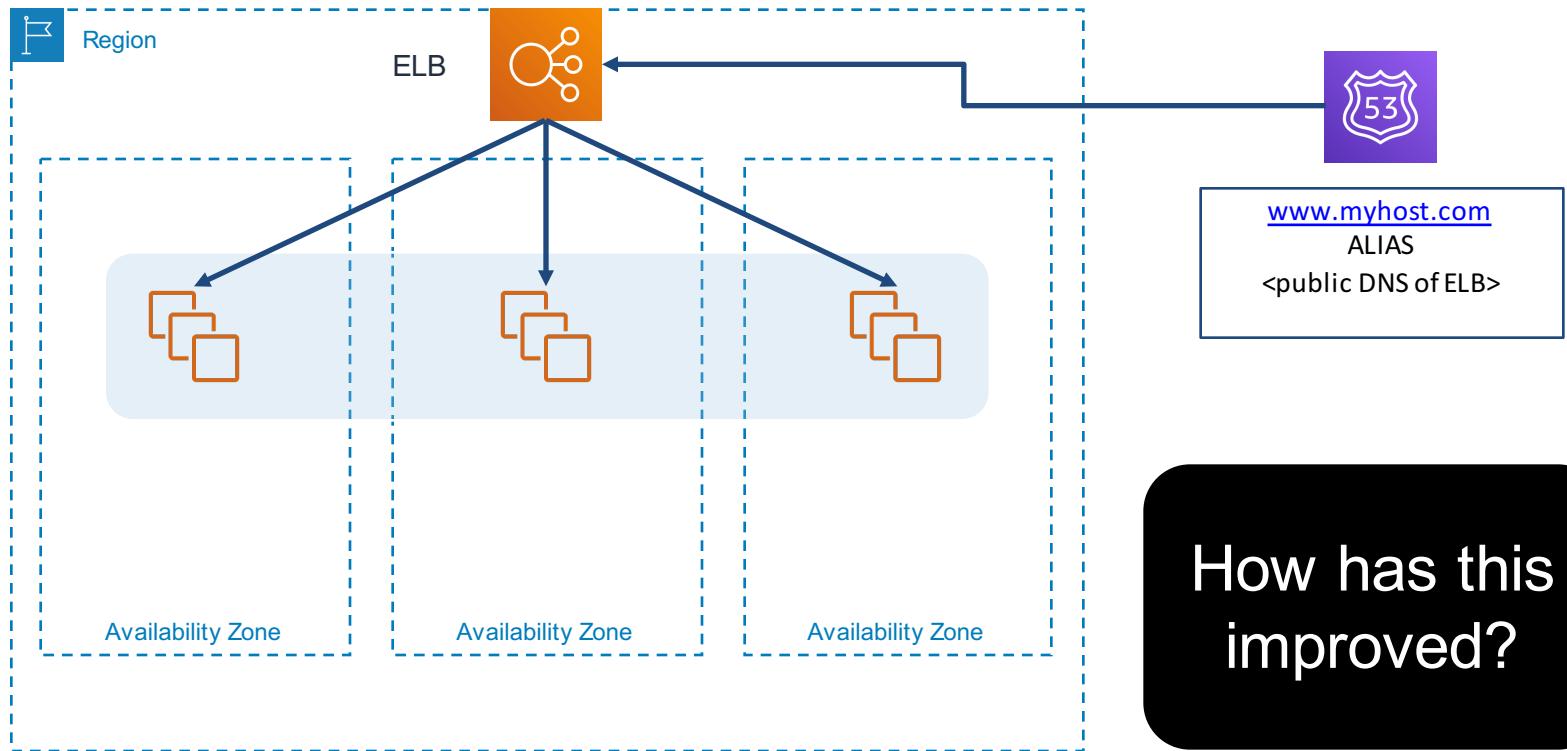
Decoupling Mechanisms

Decoupling - Single Region

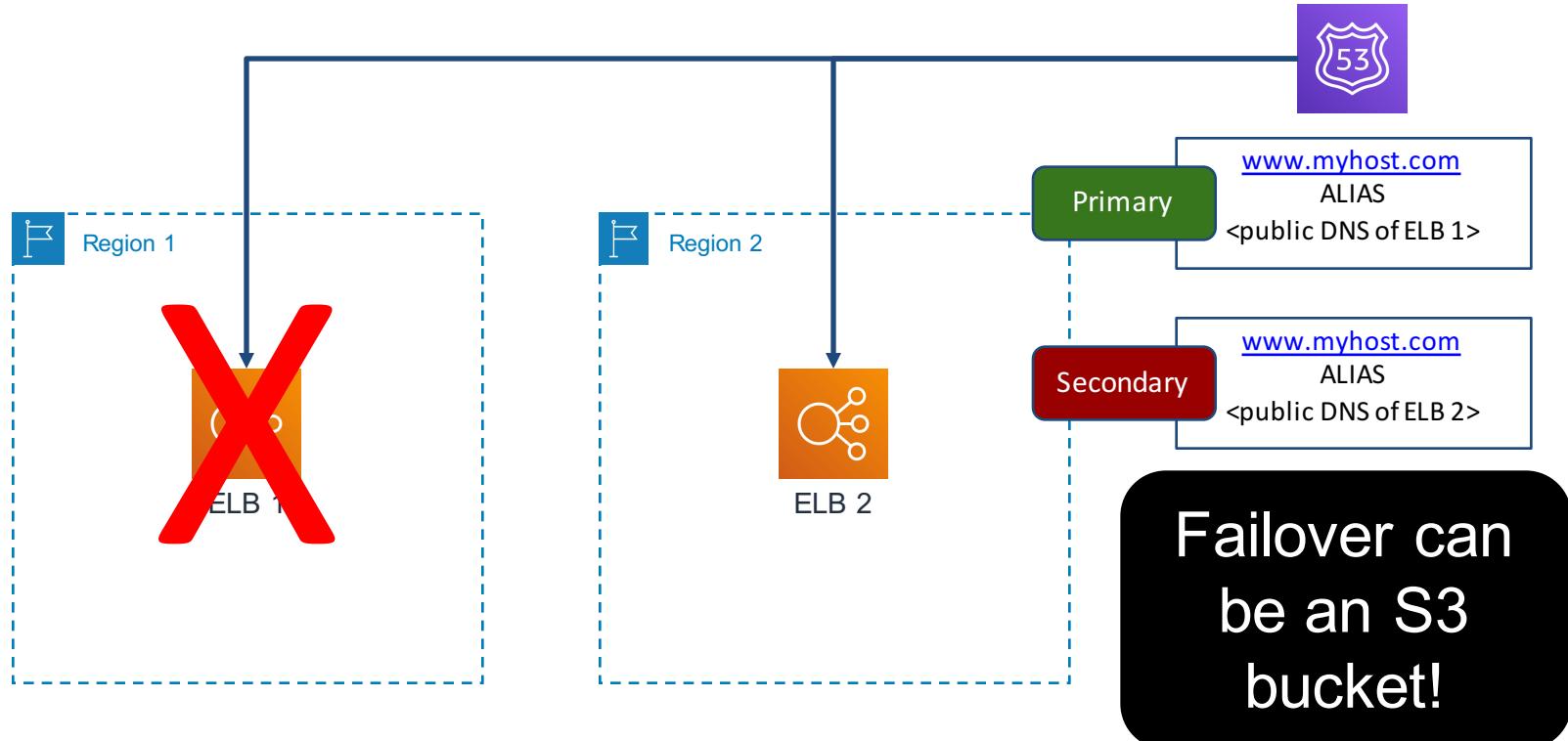
Is this DNS record guaranteed to point to the same public DNS at all times?



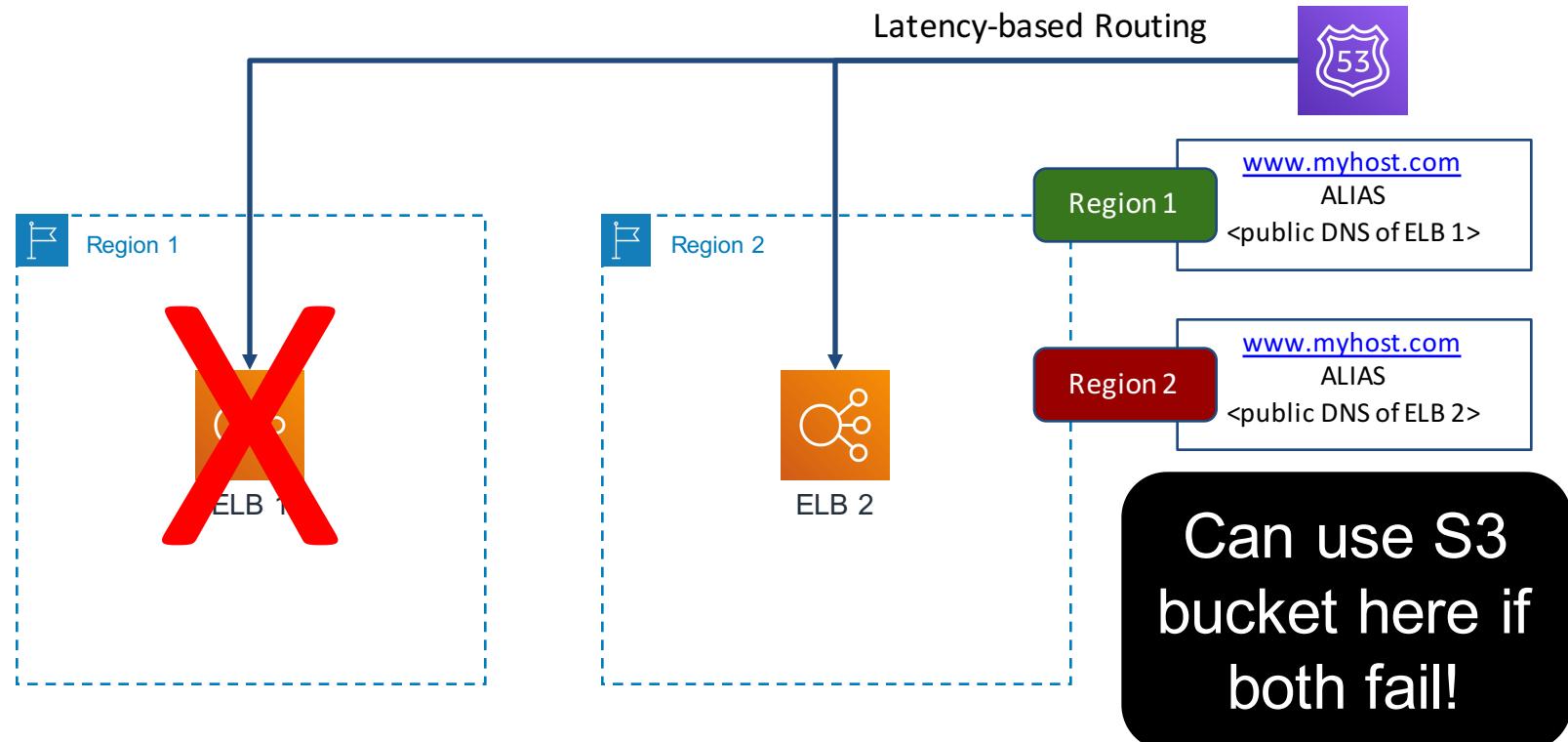
Decoupling - Single Region



Decoupling - Multi Region Failover



Decoupling - Multi Region Concurrent



Decoupling

EASY Question Breakdown

Question Breakdown

What is one potential side benefit to application decoupling?

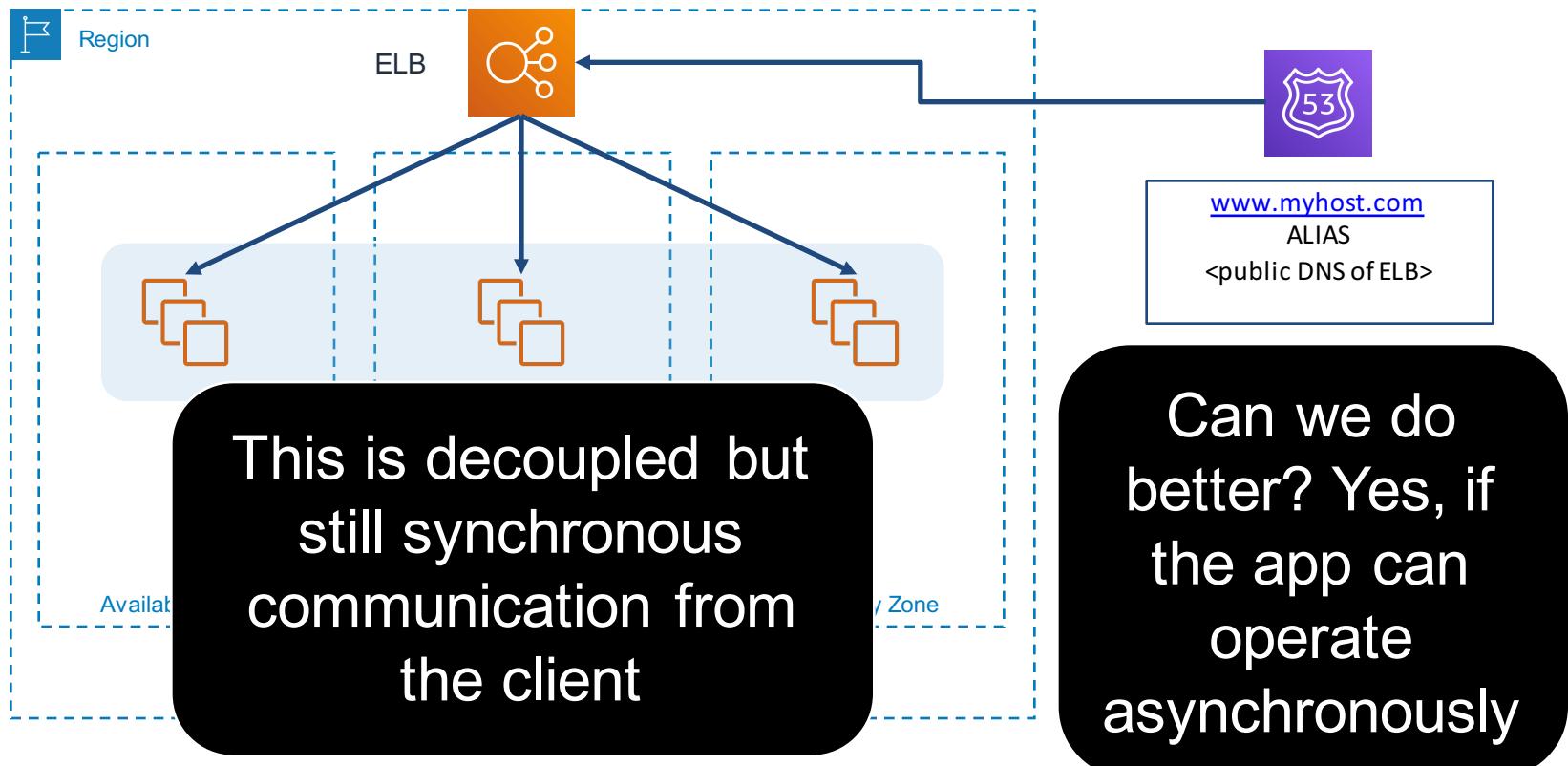
- A. Decreased operational overhead
- B. Easier future service adoption
- C. Lower service latencies
- D. Improved vertical scaling

Question Breakdown - Correct Answer

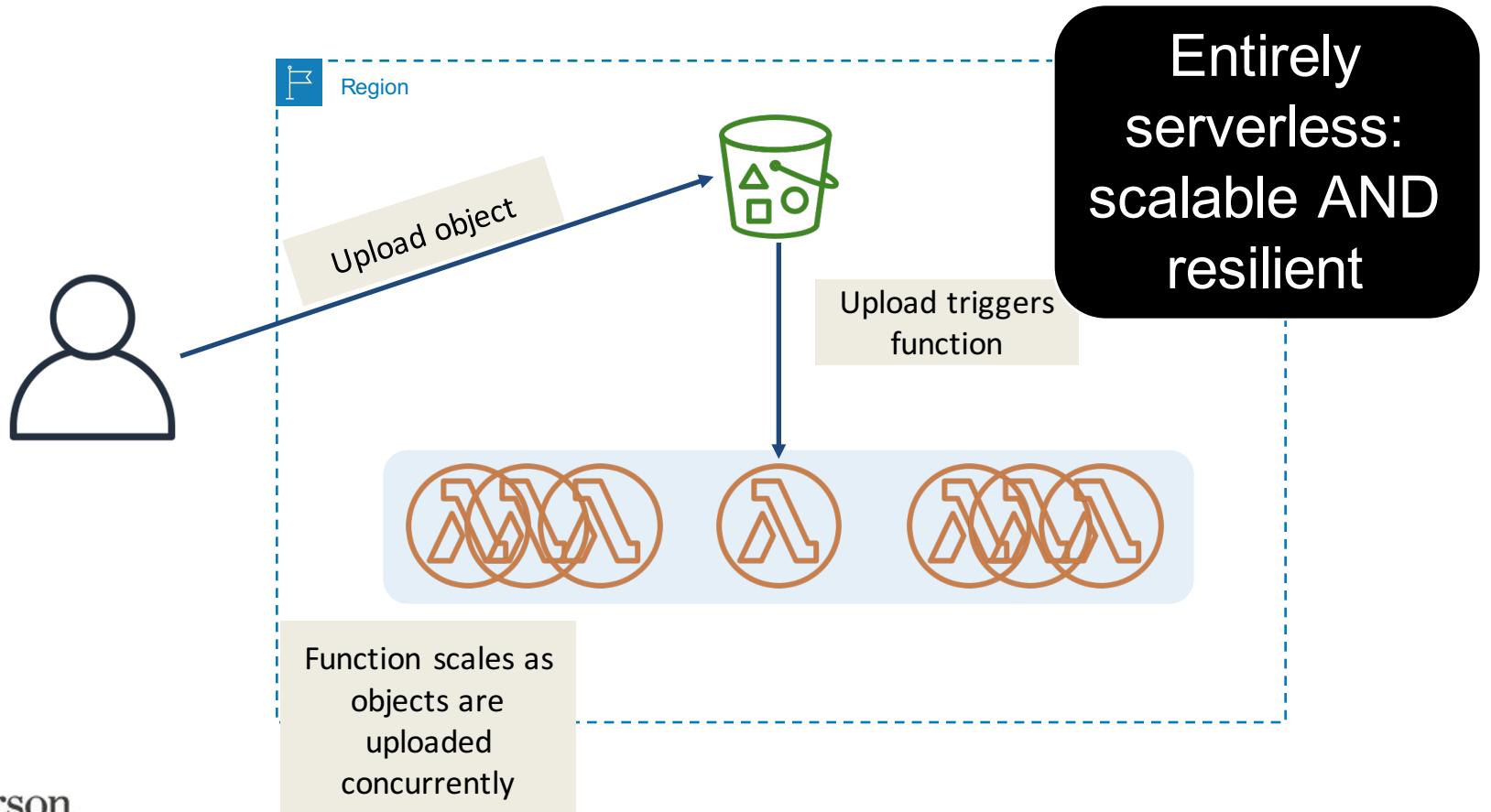
Correct Answer: B

- A. Decreased operational overhead
- B. Easier future service adoption
- C. Lower service latencies
- D. Improved vertical scaling

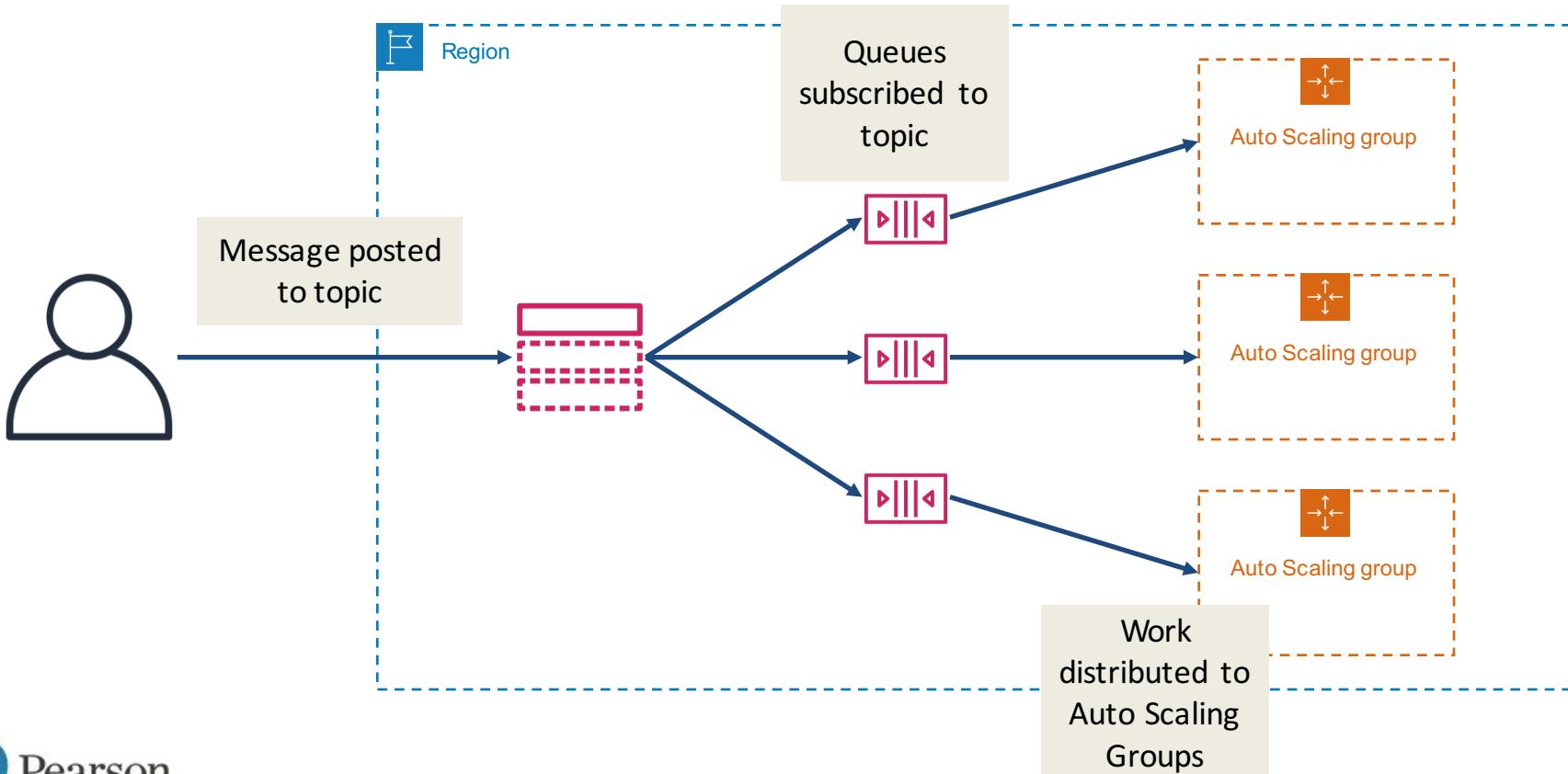
Decoupling - Lambda + S3



Decoupling - Lambda + S3



Decoupling - SNS + SQS + Auto Scaling



Decoupling Mechanisms

Question Breakdown

Question Breakdown - Key Terms

An application team supports a service that runs on a **single EC2 instance** with an **EIP** attached. The service accepts **HTTP requests** and performs **asynchronous work** before placing results in an **S3 bucket**. There is a new requirement to **improve the overall resilience** of the application. Which of the following **decoupling solutions** will best improve the resilience of the infrastructure?

- A. Create an AMI of the instance. Launch two instances from the AMI and place them behind an Application Load Balancer.
- B. Create an AMI of the instance. Create an Auto Scaling group using the AMI in a Launch Template, and associate the ASG with an Application Load Balancer.
- C. Create an SQS Queue. Place requests in the queue, and migrate the app code to a Lambda function that is triggered by messages in the queue.
- D. Create an SQS Queue. Place requests in the queue and poll the queue from the EC2 instance.

Question Breakdown - Answers

This will improve resilience, but will rely on synchronous HTTP communication through the ALB to accept tasks, even though the work is performed asynchronously afterward.

- A. Create an AMI of the instance. Launch two instances from the AMI and place them behind an Application Load Balancer.
- B. Create an AMI of the instance. Create an Auto Scaling group using the AMI in a Launch Template, and associate the ASG with an Application Load Balancer.
- C. Create an SQS Queue. Place requests in the queue, and migrate the app code to a Lambda function that is triggered by messages in the queue.
- D. Create an SQS Queue. Place requests in the queue and poll the queue from the EC2 instance.

Question Breakdown - Answers

This will improve resilience (more so than A), but will also rely on synchronous HTTP communication through the ALB.

- A. Create an AMI of the instance. Launch two instances from the AMI and place them behind an Application Load Balancer.
- B. Create an AMI of the instance. Create an Auto Scaling group using the AMI in a Launch Template, and associate the ASG with an Application Load Balancer.
- C. Create an SQS Queue. Place requests in the queue, and migrate the app code to a Lambda function that is triggered by messages in the queue.
- D. Create an SQS Queue. Place requests in the queue and poll the queue from the EC2 instance.

Question Breakdown - Answers

This will improve resilience, and also completely decouple the service from upstream applications. The message queue is also highly available, making this a good candidate for the correct answer choice.

- A. Create an AMI of the instance. Launch two instances from the AMI and place them behind an Application Load Balancer.
- B. Create an AMI of the instance. Create an Auto Scaling group using the AMI in a Launch Template, and associate the ASG with an Application Load Balancer.
- C. Create an SQS Queue. Place requests in the queue, and migrate the app code to a Lambda function that is triggered by messages in the queue.
- D. Create an SQS Queue. Place requests in the queue and poll the queue from the EC2 instance.

Question Breakdown - Answers

The message queue will remove the synchronous communication from requesting applications, but the EC2 instance is still a single point of failure.

- A. Create an AMI of the instance. Launch two instances from the AMI and place them behind an Application Load Balancer.
- B. Create an AMI of the instance. Create an Auto Scaling group using the AMI in a Launch Template, and associate the ASG with an Application Load Balancer.
- C. Create an SQS Queue. Place requests in the queue, and migrate the app code to a Lambda function that is triggered by messages in the queue.
- D. Create an SQS Queue. Place requests in the queue and poll the queue from the EC2 instance.

Question Breakdown - Correct Answer

Correct Answer: C

- A. Create an AMI of the instance. Launch two instances from the AMI and place them behind an Application Load Balancer.
- B. Create an AMI of the instance. Create an Auto Scaling group using the AMI in a Launch Template, and associate the ASG with an Application Load Balancer.
- C. Create an SQS Queue. Place requests in the queue, and migrate the app code to a Lambda function that is triggered by messages in the queue.
- D. Create an SQS Queue. Place requests in the queue and poll the queue from the EC2 instance.



Design Performant Architectures, Part 1 of 2

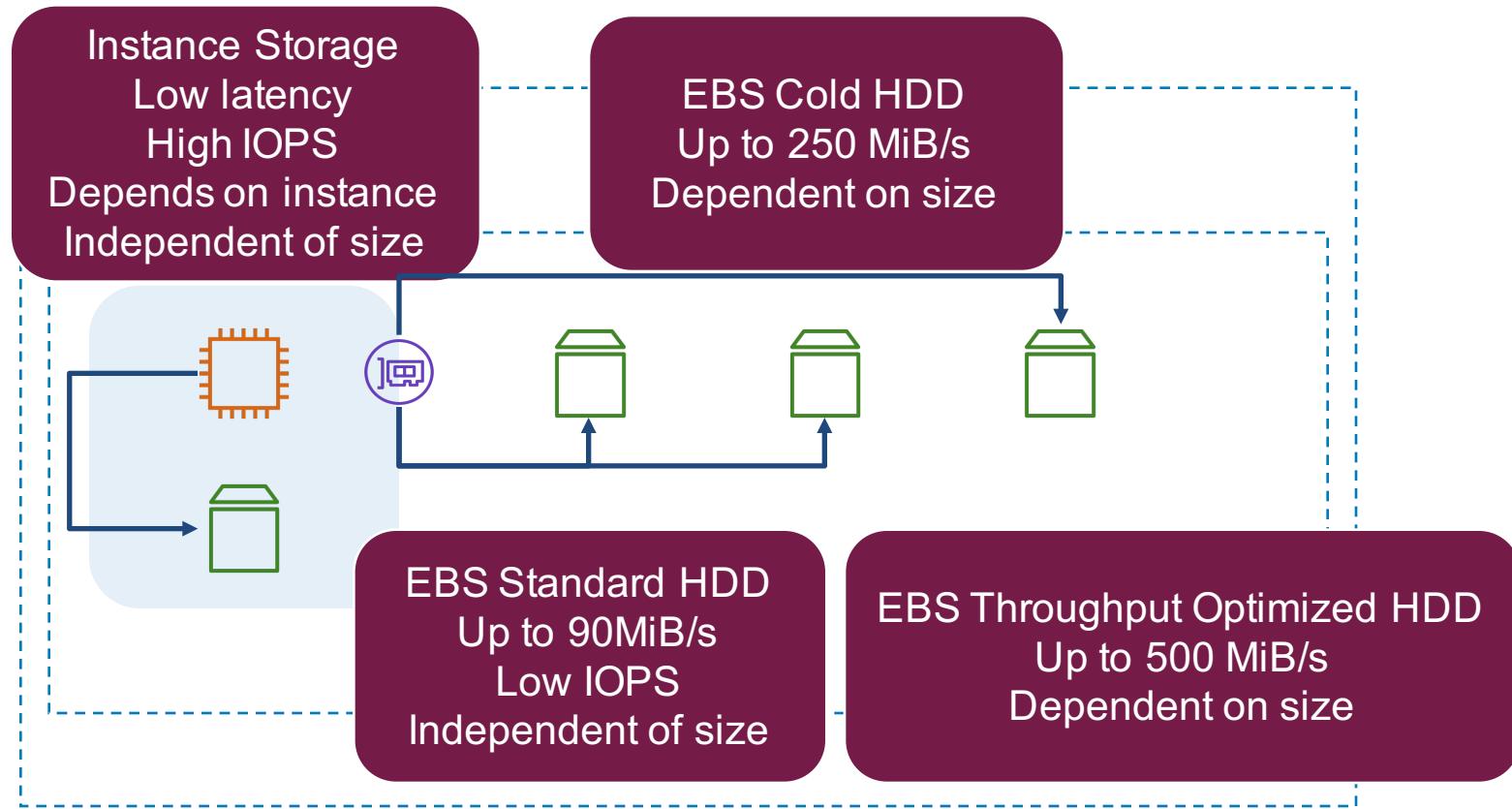
28%



Design Performant
Architectures, Part 1 of 2

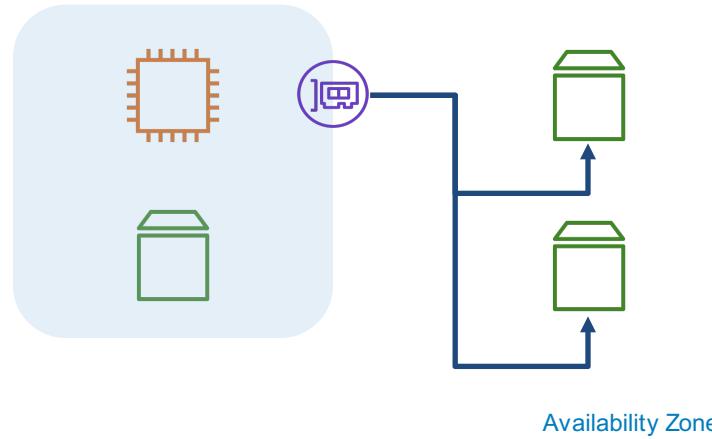
Performant Storage

Performant Storage - Block



Performant Storage - Block

Instance maximums
Up to 160,000 IOPS across multiple volumes
3500 Mbps across multiple volumes



EBS General Purpose SSD
Up to 16000 IOPS
128-250MiB/s
Dependent on size

EBS Provisioned IOPS SSD
Up to 64000* IOPS
up to 500-1000* MiB/s
Dependent on size

Performant Storage

EASY Question Breakdown

Question Breakdown

Which EBS volume type would be the most cost-effective for 7 TB of infrequently-accessed data?

- A. gp2
- B. piops
- C. sc1
- D. st1
- E. standard

Question Breakdown - Correct Answer

Correct Answer: C

- A. gp2
- B. piops
- C. sc1
- D. st1
- E. standard

Performant Storage - File



Region



General Purpose performance mode
35000 READ, 7000 WRITE IOPS
Lowest metadata latency

EFS File system resource
Up to 3Gb/s throughput
Depends on region
250MB/s per client
Latency: unpublished but
can be 1ms



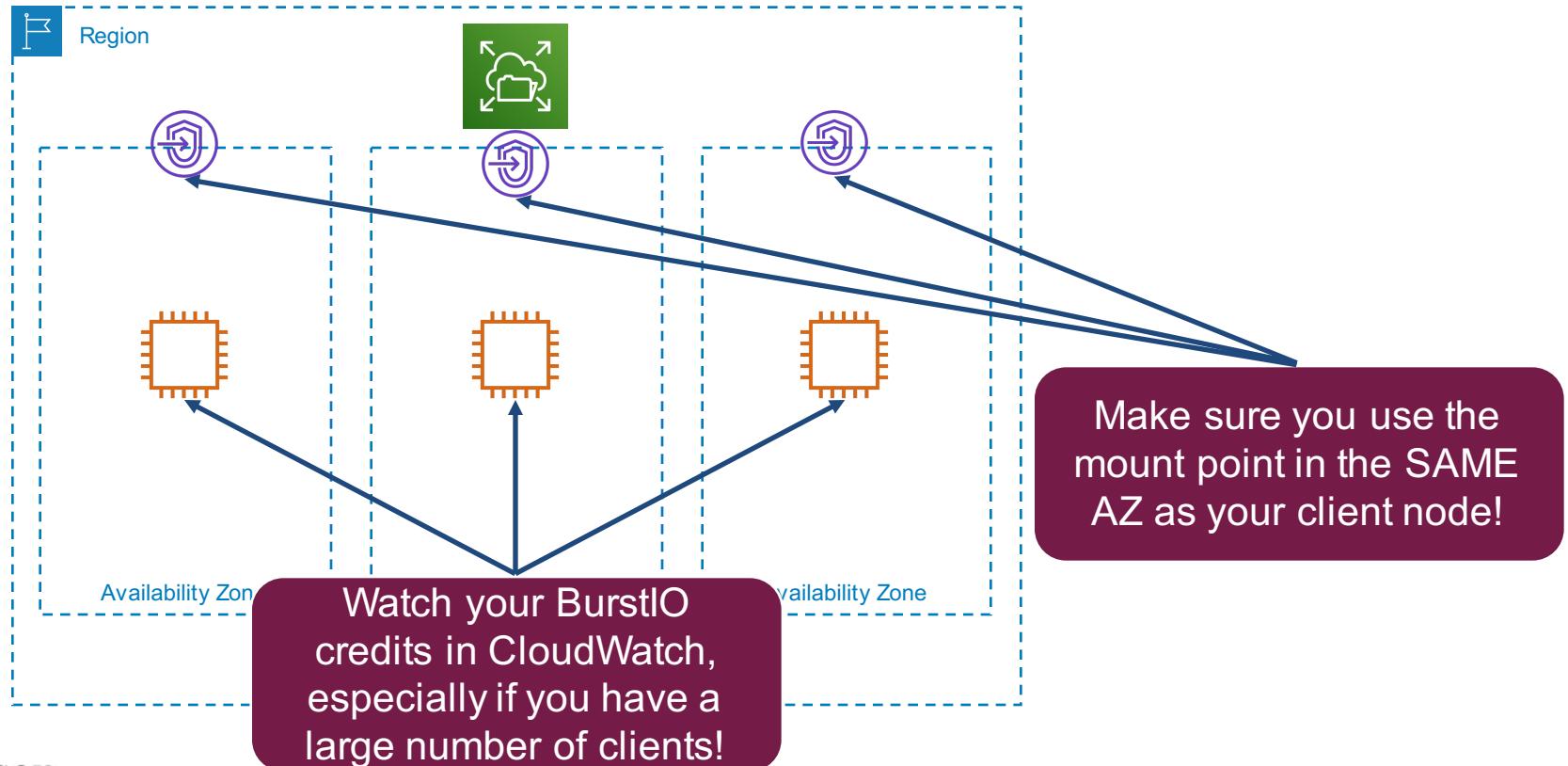
MaxIO performance mode
500k+ IOPS
Highest metadata latency

Availability Zone

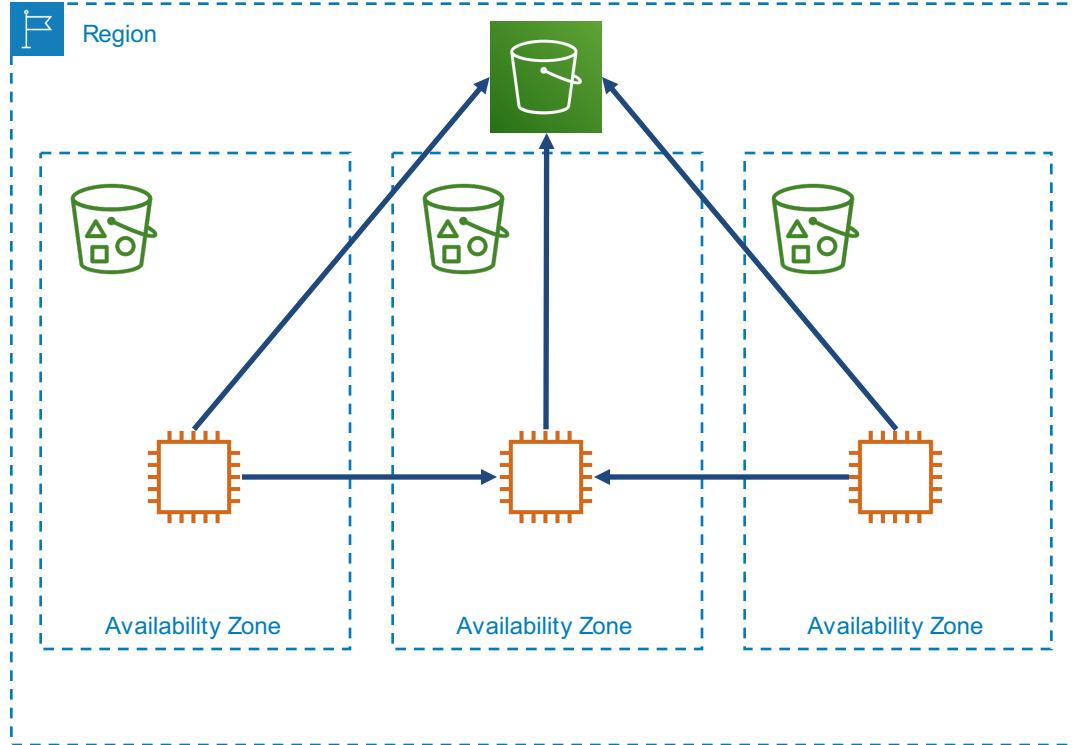
Availability Zone

Availability Zone

Performant Storage - File



Performant Storage - Object



Parallelize requests
to improve S3
performance!

Anti-pattern!
Don't aggregate S3
requests through a
single node!

Performant Storage - Object

Per-prefix performance

PUT/POST/COPY/DELETE 3500/sec

GET/HEAD 5500/sec

unlimited prefixes allowed

NO RANDOM PREFIXES REQUIRED

Latency performance

Small objects 100-200ms

Large objects 100-200ms for
first-byte-out

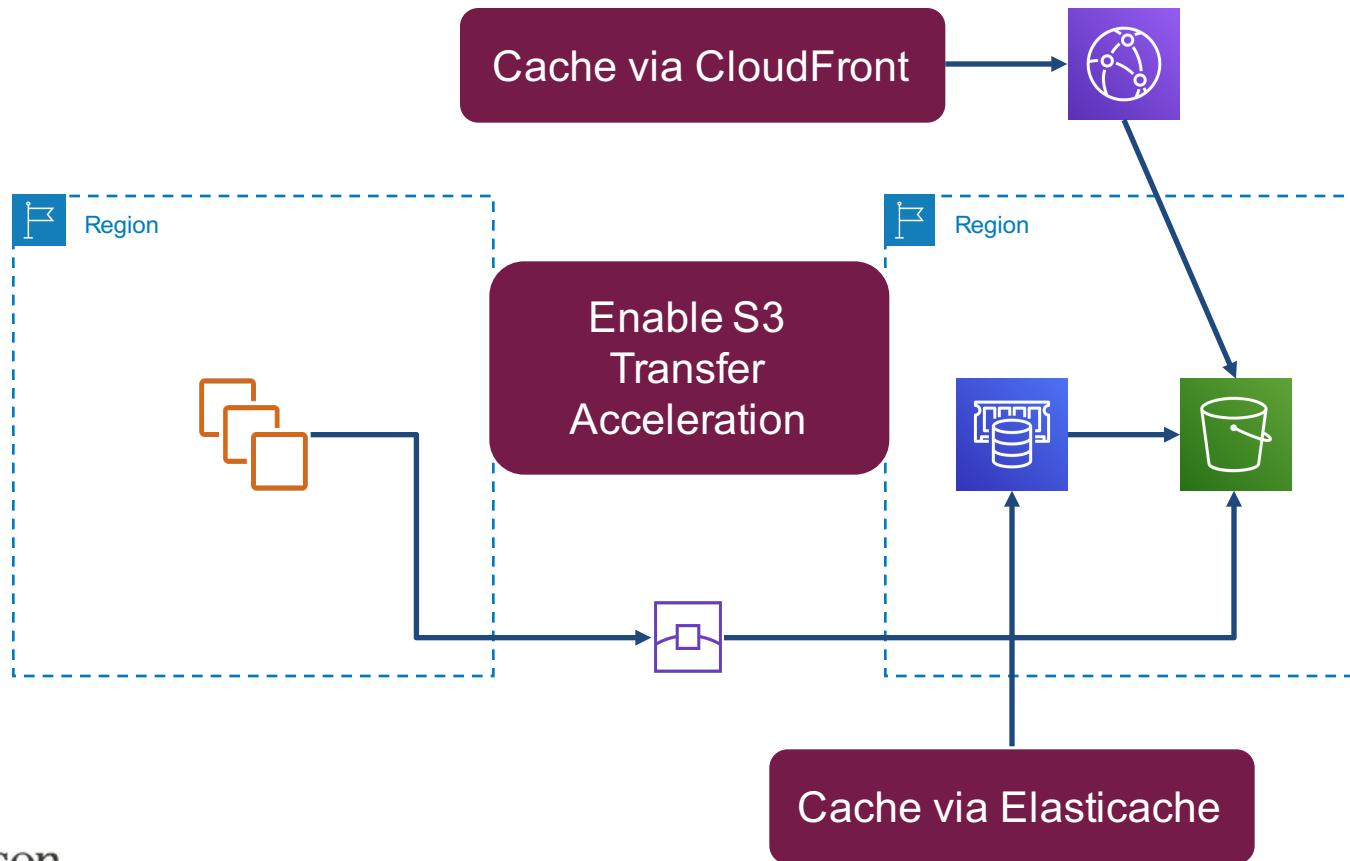
What other S3
performance options
are available?



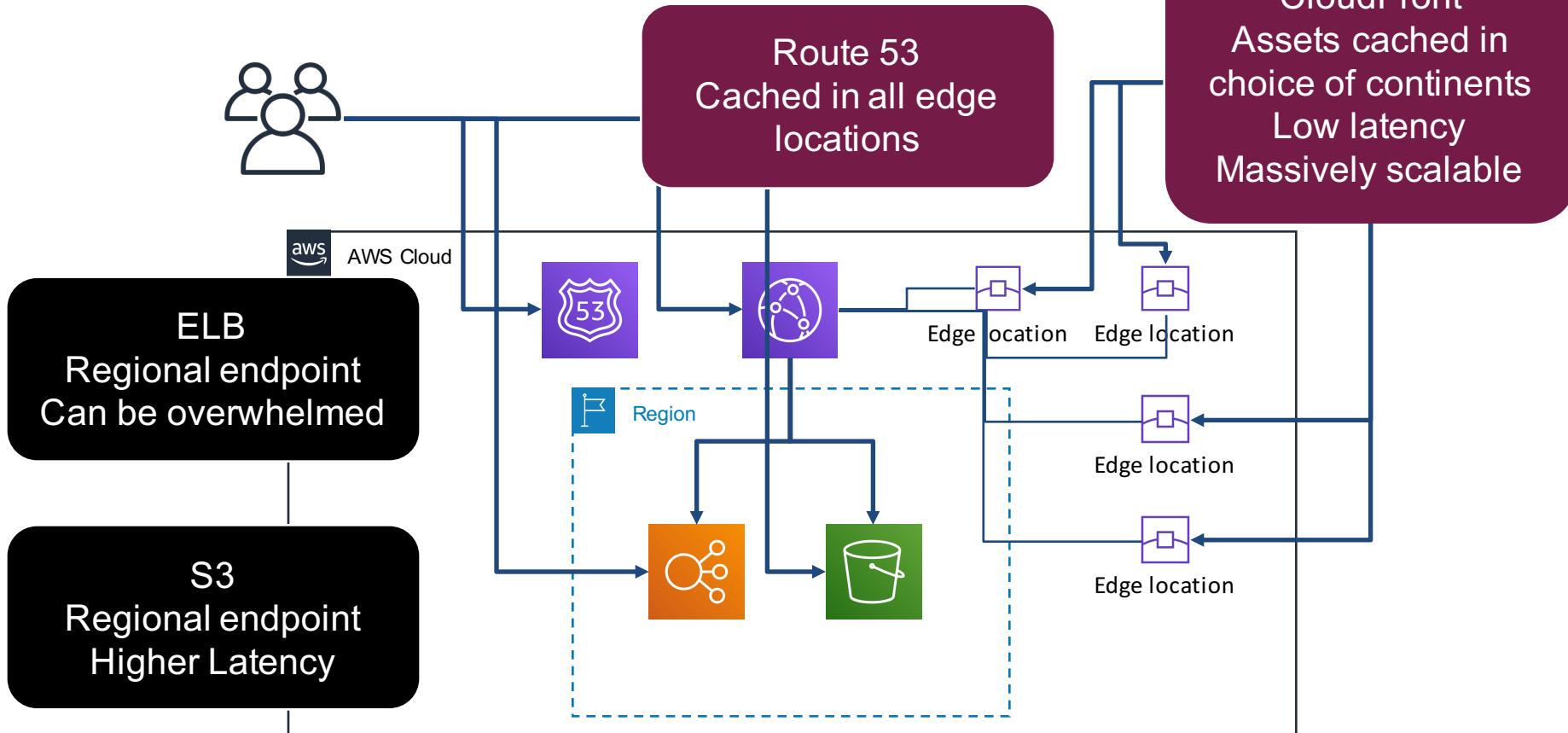
Region



Performant Storage - Object



Caching Solutions - CloudFront



Caching Solutions

EASY Question Breakdown

Question Breakdown

AWS has caching offerings that can be implemented for _____ operations at various layers in an application architecture.

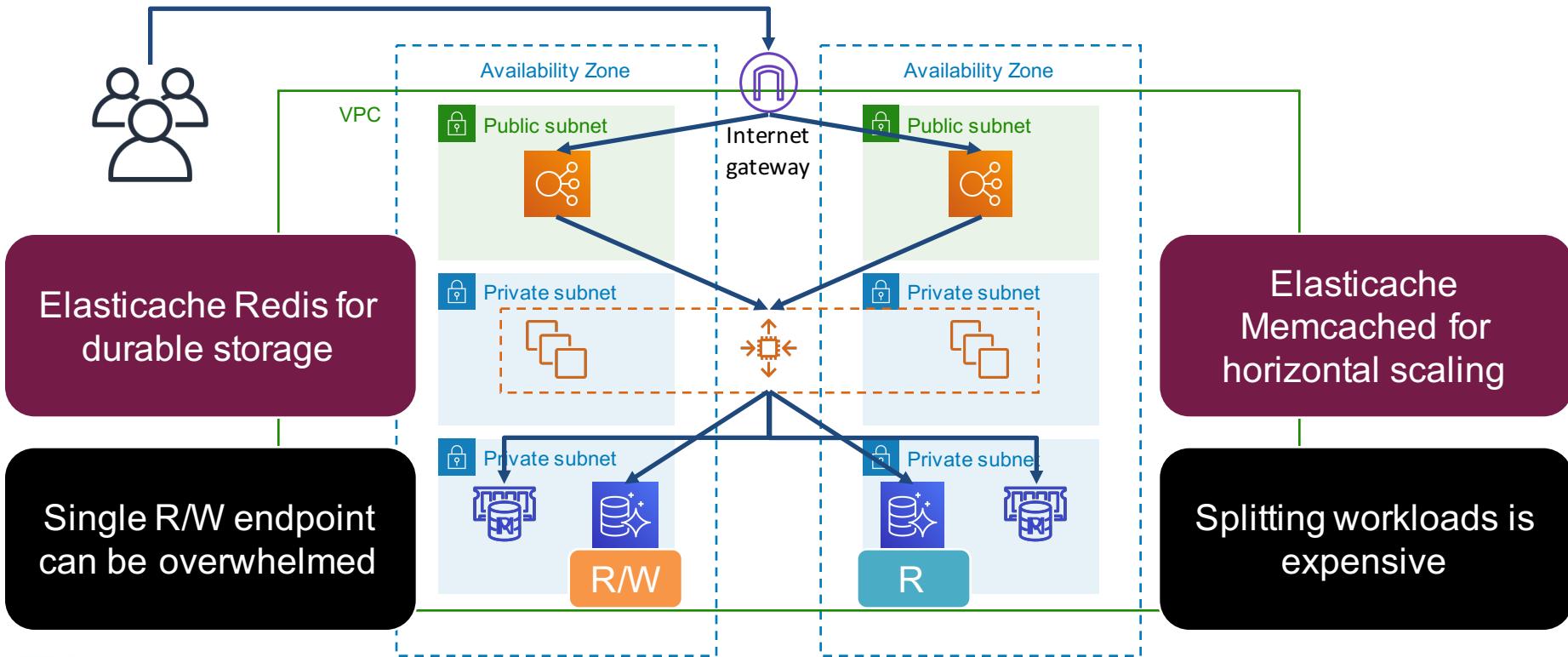
- A. read and write
- B. read
- C. write
- D. no

Question Breakdown - Correct Answer

Correct Answer: A

- A. **read and write**
- B. read
- C. write
- D. no

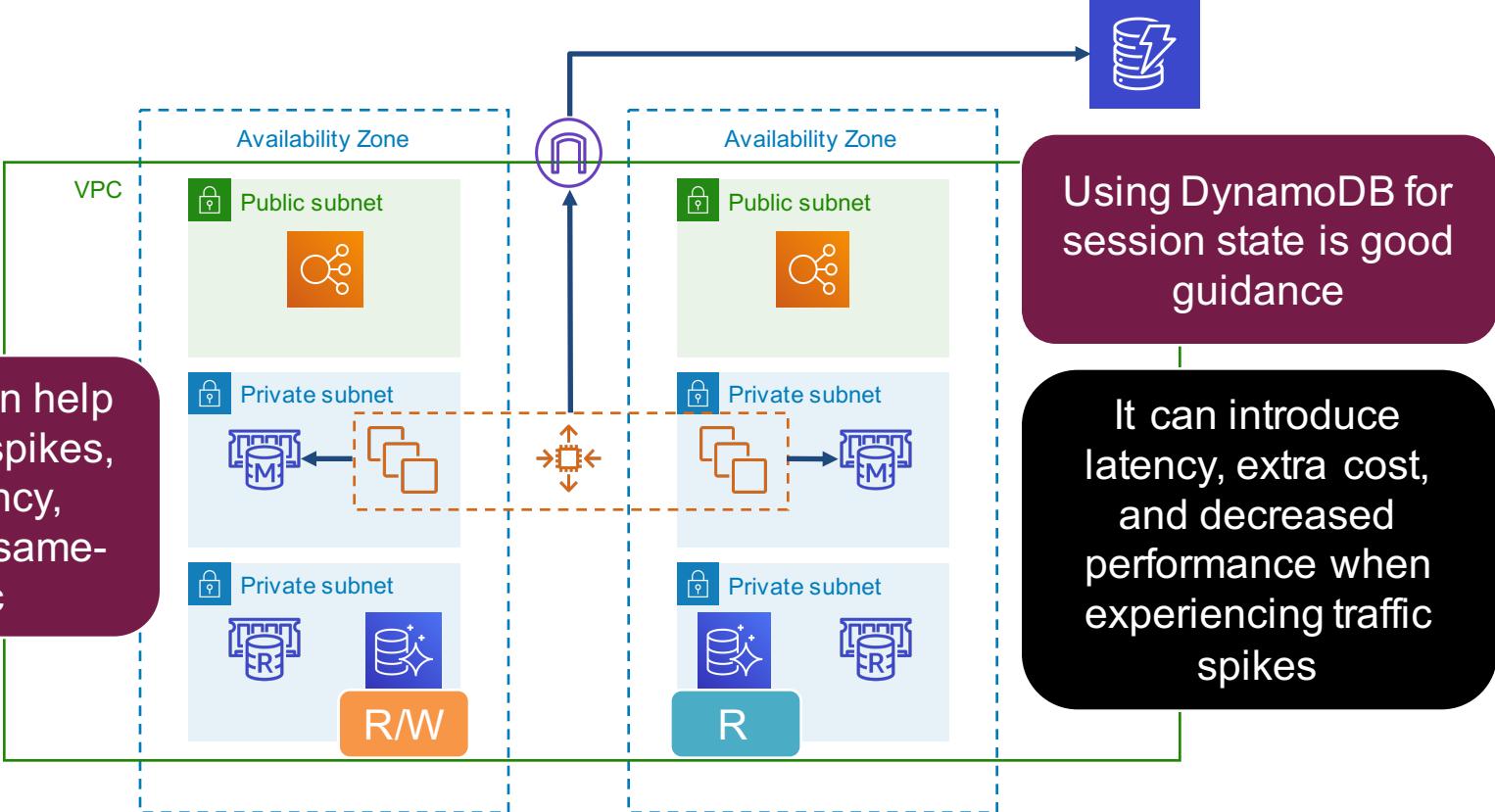
Caching Solutions - ElastiCache



Caching Solutions - ElastiCache



ElastiCache can help absorb traffic spikes, reduce latency, especially for same-AZ traffic



Performant Storage

Question Breakdown

Question Breakdown - Key Terms

During the peak load every weekday, an **MSSQL RDS** database becomes **overloaded** due to heavy **read traffic**, impacting user request latencies. You've been asked to recommend a solution that **improves the user experience** and **enables easier scaling** during future anticipated **increased load**. Which of the following will best meet the requirements?

- A. Configure an Elasticache cluster to cache database reads. Query the cache from the application before issuing reads to the database.
- B. Increase either the RDS storage size or PIOPS to maximum value to improve database performance.
- C. Upsize the RDS database instance to improve database performance.
- D. Scale the application tier horizontally to accommodate more concurrent requests.

Question Breakdown - Answers

This strategy will allow reads to scale without requiring vertical scaling of the database instance.

- A. Configure an Elasticache cluster to cache database reads. Query the cache from the application before issuing reads to the database.
- B. Increase either the RDS storage size or PIOPS to maximum value to improve database performance.
- C. Upsize the RDS database instance to improve database performance.
- D. Scale the application tier horizontally to accommodate more concurrent requests.

Question Breakdown - Answers

Improving storage performance will directly impact the user experience, but is a costly solution, and one that has a vertical ceiling (the limits of EBS storage on RDS).

- A. Configure an ElastiCache cluster to cache database reads. Query the cache from the application before issuing reads to the database.
- B. Increase either the RDS storage size or PIOPS to maximum value to improve database performance.
- C. Upsize the RDS database instance to improve database performance.
- D. Scale the application tier horizontally to accommodate more concurrent requests.

Question Breakdown - Answers

Increasing the CPU and/or memory available to the database instance can improve the user experience, but is also a costly and limited approach similar to B.

- A. Configure an ElastiCache cluster to cache database reads. Query the cache from the application before issuing reads to the database.
- B. Increase either the RDS storage size or PIOPS to maximum value to improve database performance.
- C. Upsize the RDS database instance to improve database performance.
- D. Scale the application tier horizontally to accommodate more concurrent requests.

Question Breakdown - Answers

Scaling the application tier does not really address the issue of database reads becoming a problem. In fact, this might make the problem worse by increasing the number of connections/queries to the database tier.

- A. Configure an ElastiCache cluster to cache database reads. Query the cache from the application before issuing reads to the database.
- B. Increase either the RDS storage size or PIOPS to maximum value to improve database performance.
- C. Upsize the RDS database instance to improve database performance.
- D. Scale the application tier horizontally to accommodate more concurrent requests.

Question Breakdown - Correct Answer

Correct Answer: A

- A. Configure an ElastiCache cluster to cache database reads. Query the cache from the application before issuing reads to the database.
- B. Increase either the RDS storage size or PIOPS to maximum value to improve database performance.
- C. Upsize the RDS database instance to improve database performance.
- D. Scale the application tier horizontally to accommodate more concurrent requests.



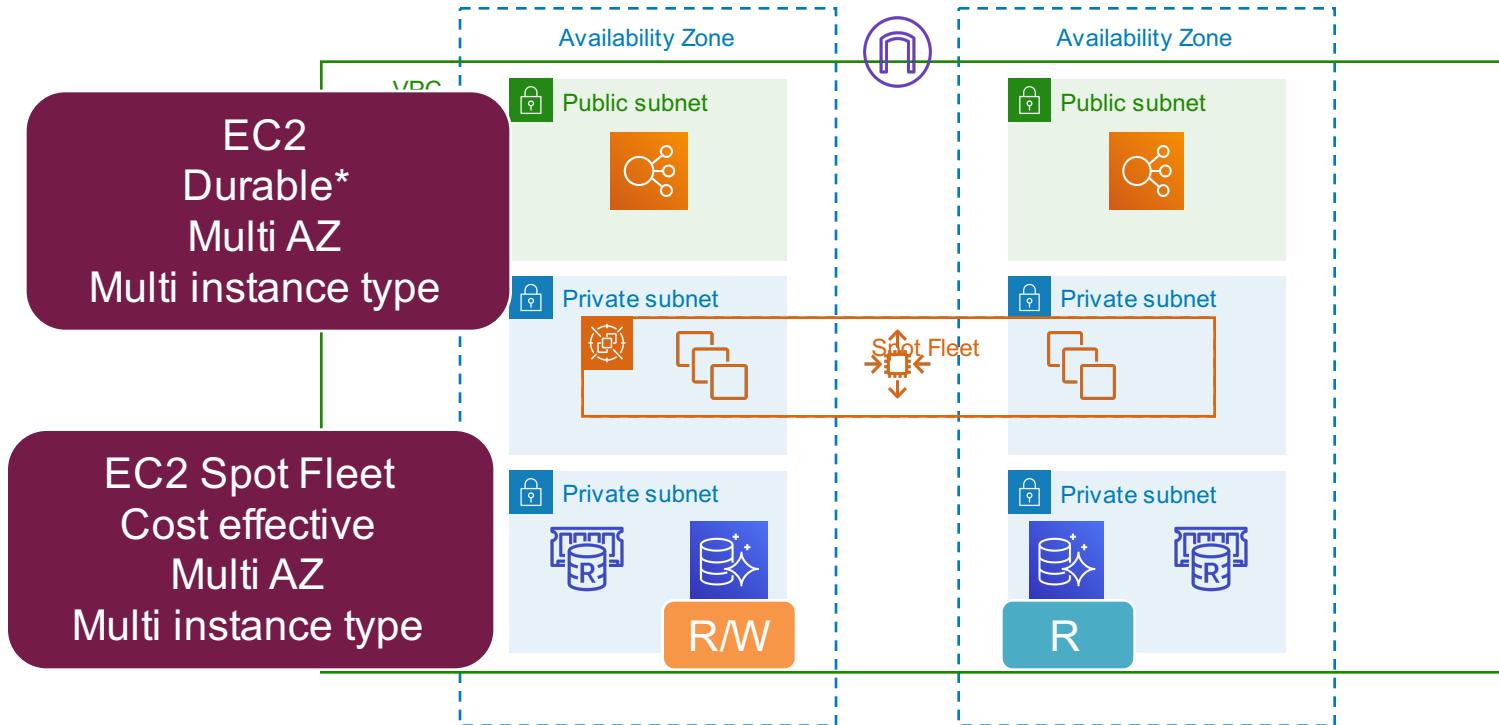
Design Performant
Architectures, Part 1 of 2

Performant Compute
Solutions

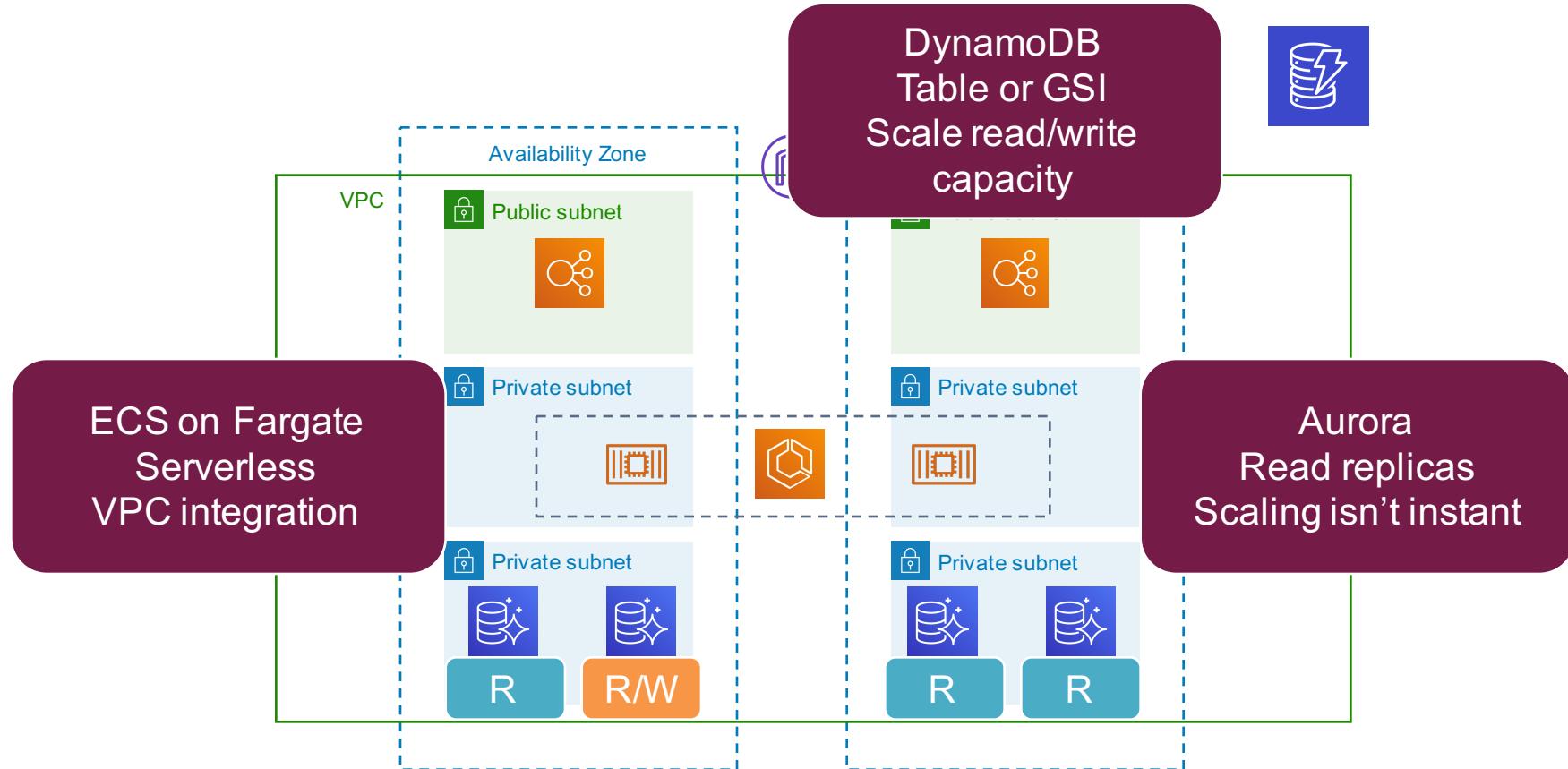
Definitions

1. **Scalability** - the ability of a system to **increase** resources to accommodate increased demand. This can be done vertically or horizontally, and is not necessarily automated.
2. **Elasticity** - the ability of a system to **increase** and **decrease** resources allocated (usually horizontally) to match demand, and implies automation.
3. In general, an **elastic** resource is also **scalable**, but the reverse isn't always true.
4. Ironically, EBS volume size is an example of a *scalable* but NOT *elastic* property, while EBS volume IOPS/throughput can be both.

Elastic/Scalable - Unified Auto Scaling



Elastic/Scalable - Unified Auto Scaling



Elasticity and Scalability

EASY Question Breakdown

Question Breakdown

Which of the following elastic services have the most operational overhead for scaling activities?

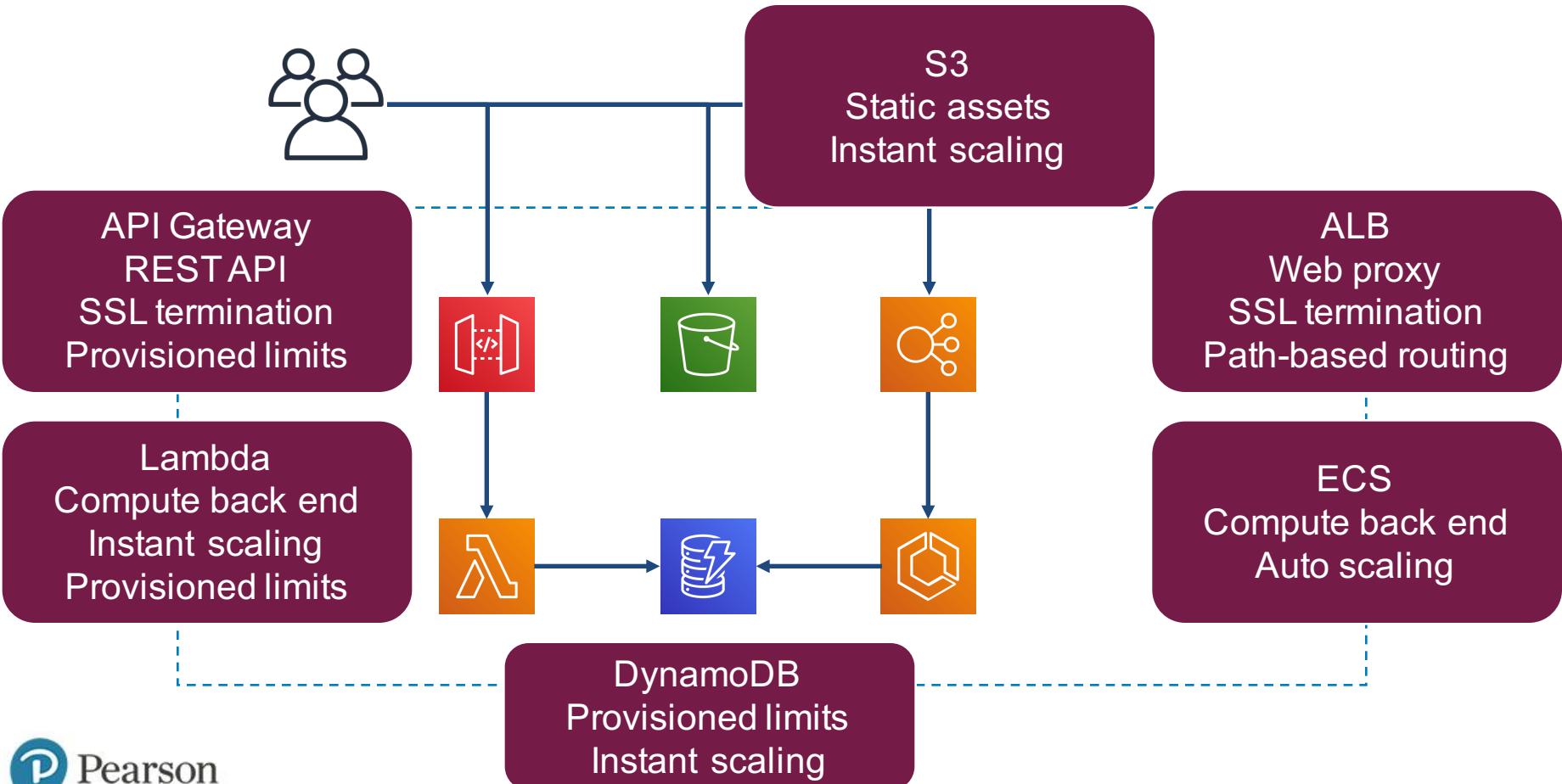
- A. EC2 Auto Scaling
- B. ECS on Fargate
- C. Aurora Read Replica Auto Scaling
- D. DynamoDB table read/write capacity Auto Scaling
- E. None of these

Question Breakdown - Correct Answer

Correct Answer: E

- A. EC2 Auto Scaling
- B. ECS on Fargate
- C. Aurora Read Replica Auto Scaling
- D. DynamoDB table read/write capacity Auto Scaling
- E. None of these

Elastic/Scalable - Managed Services



Performant Compute Resources

Question Breakdown

Question Breakdown - Key Terms

An application is currently deployed using AWS Auto Scaling on EC2. The application experiences a steep traffic spike twice per week, but not always at the same time. The spike does usually start within the same 60 minute window, and can occur anywhere within that window. What minimal-overhead strategy could be employed to ensure a good user experience every time, as the current Auto Scaling configuration is not able to scale fast enough at the start of the traffic spike?

- A. Configure Scheduled scale-out at the beginning of the hour window on the spike days.
- B. Increase the minimum instance number to more effectively handle the spikes.
- C. Write a shell script to execute manual scaling out before the hour window on spike days.
- D. Configure Predictive Scaling on the Auto Scaling group.

Question Breakdown - Answers

This will accommodate the traffic spikes as long as they dont ever change the hour window, and might not be an appropriate long-term solution.

- A. Configure Scheduled scale-out at the beginning of the hour window on the spike days.
- B. Increase the minimum instance number to more effectively handle the spikes.
- C. Write a shell script to execute manual scaling out before the hour window on spike days.
- D. Configure Predictive Scaling on the Auto Scaling group.

Question Breakdown - Answers

While this is a guaranteed fix for the spikes, it goes against the best practice of elasticity and renders the Auto Scaling service less useful overall.

- A. Configure Scheduled scale-out at the beginning of the hour window on the spike days.
- B. Increase the minimum instance number to more effectively handle the spikes.
- C. Write a shell script to execute manual scaling out before the hour window on spike days.
- D. Configure Predictive Scaling on the Auto Scaling group.

Question Breakdown - Answers

A scheduled script will be similar to solution A, with the added overhead of maintaining the compute resource that runs the script.

- A. Configure Scheduled scale-out at the beginning of the hour window on the spike days.
- B. Increase the minimum instance number to more effectively handle the spikes.
- C. Write a shell script to execute manual scaling out before the hour window on spike days.
- D. Configure Predictive Scaling on the Auto Scaling group.

Question Breakdown - Answers

Within a few weeks, Predictive Scaling will have created a model that accounts for the spikes by scaling in advance. It has the further advantage of catching any other regular traffic changes, and so this is a zero-overhead solution.

- A. Configure Scheduled scale-out at the beginning of the hour window on the spike days.
- B. Increase the minimum instance number to more effectively handle the spikes.
- C. Write a shell script to execute manual scaling out before the hour window on spike days.
- D. **Configure Predictive Scaling on the Auto Scaling group.**

Question Breakdown - Correct Answer

Correct Answer: D

- A. Configure Scheduled scale-out at the beginning of the hour window on the spike days.
- B. Increase the minimum instance number to more effectively handle the spikes.
- C. Write a shell script to execute manual scaling out before the hour window on spike days.
- D. Configure Predictive Scaling on the Auto Scaling group.



Design Performant Architectures, Part 2 of 2

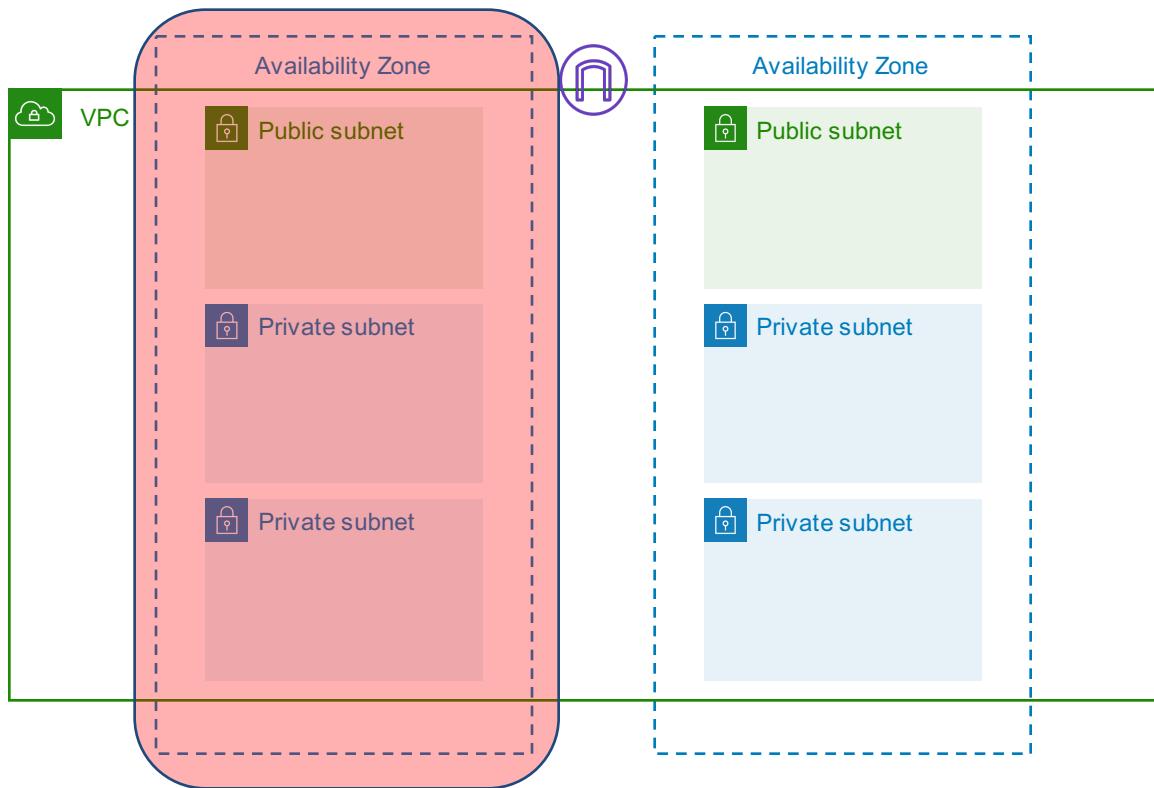
28%



Design Performant
Architectures, Part 2 of 2

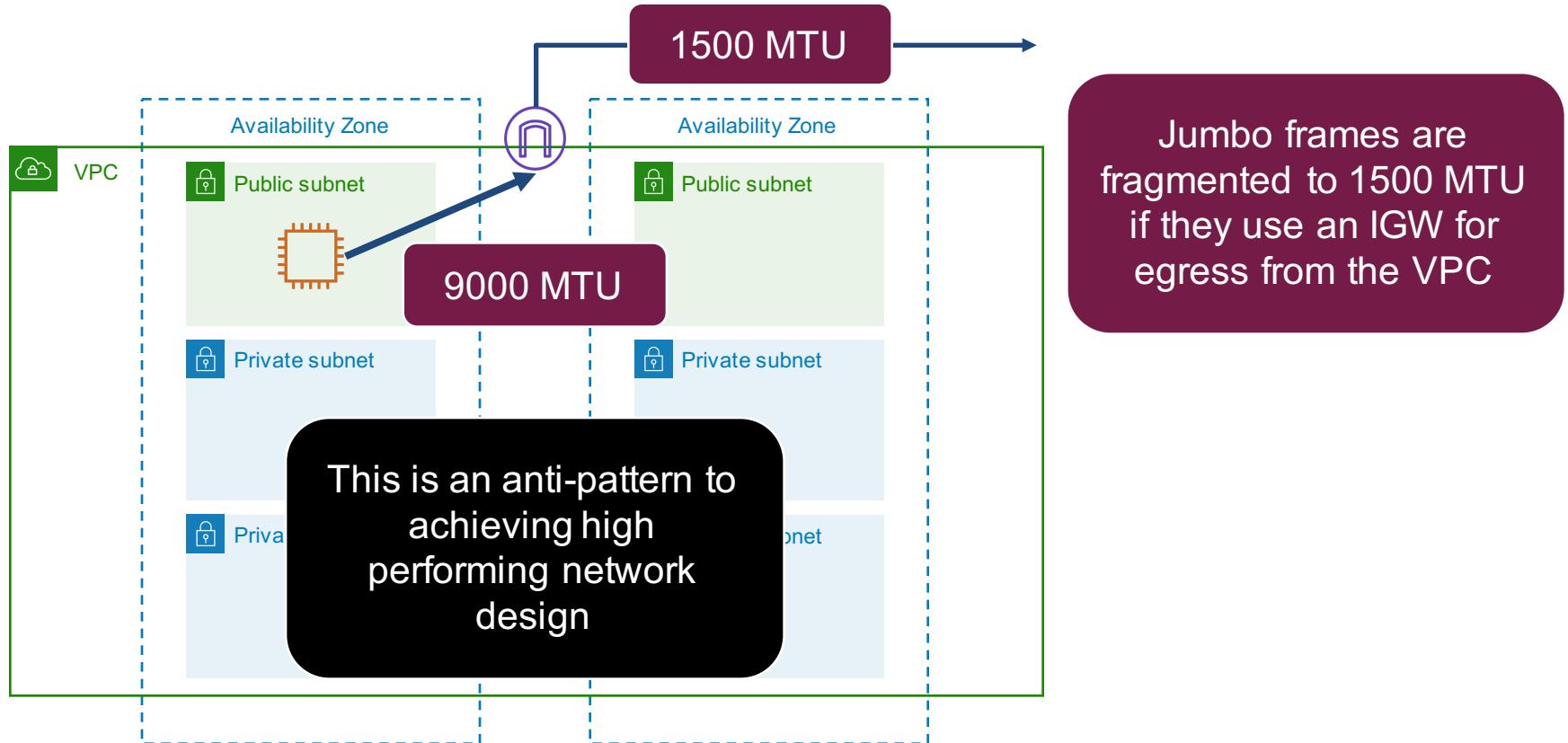
Performant Network Solutions

Single-AZ Design



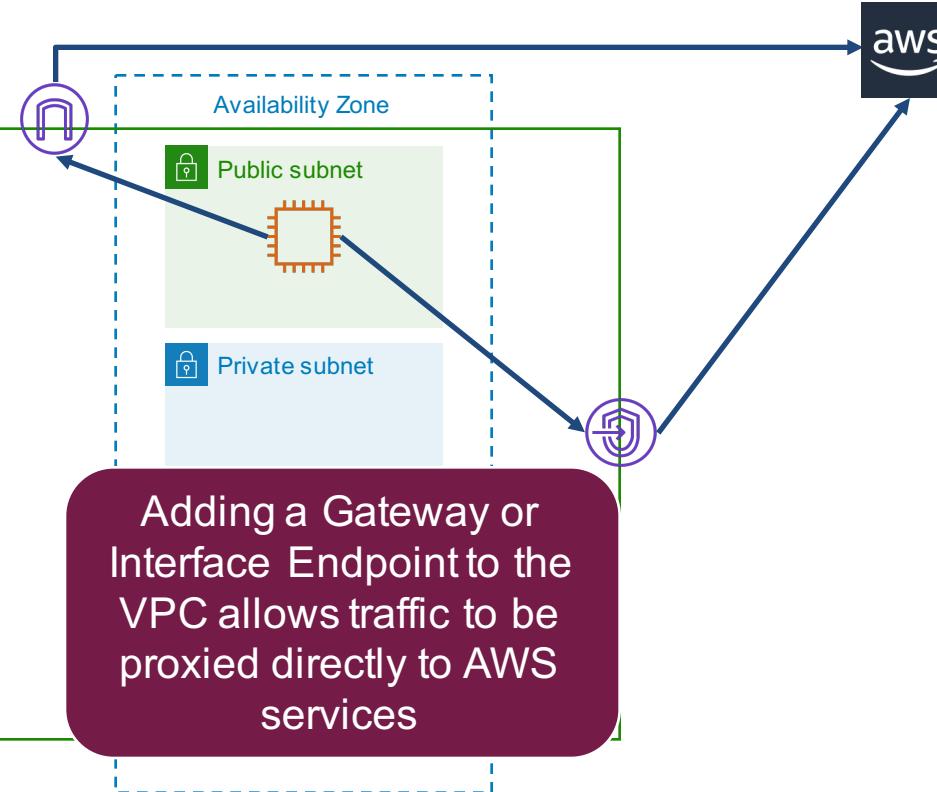
Consolidate resources
within a single AZ to
minimize latency (<1ms)

Traffic Egress Fragmentation

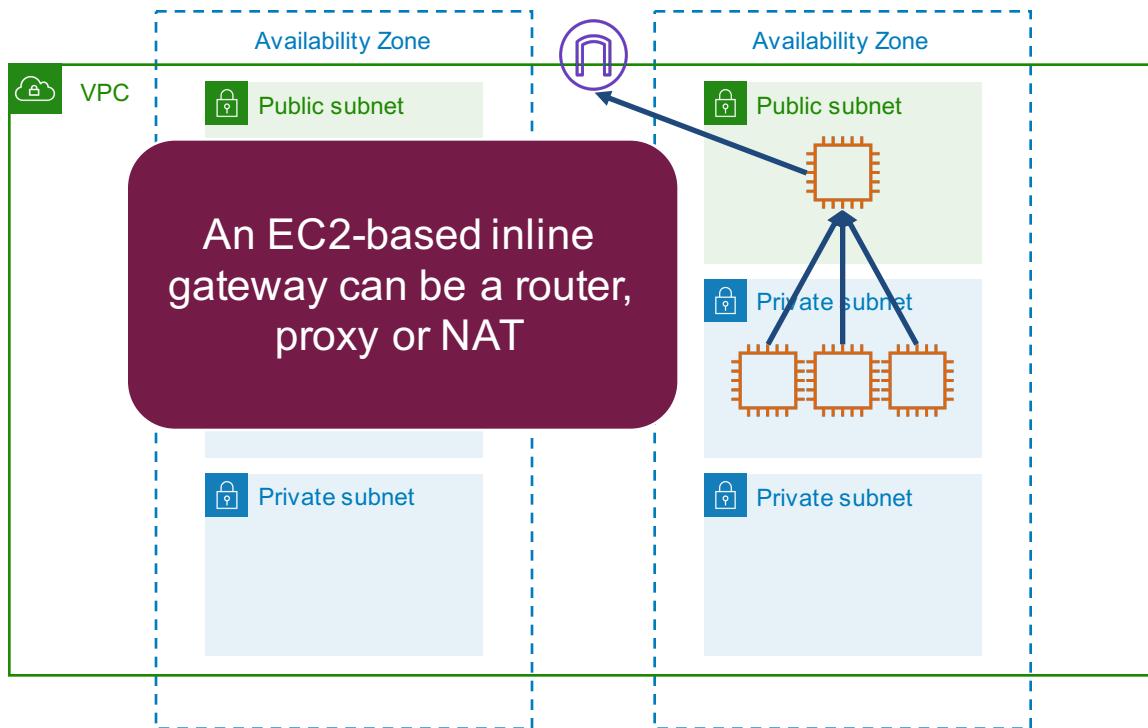


Avoid Public AWS Network

Using an IGW to reach AWS services requires traffic crossing public AWS networks



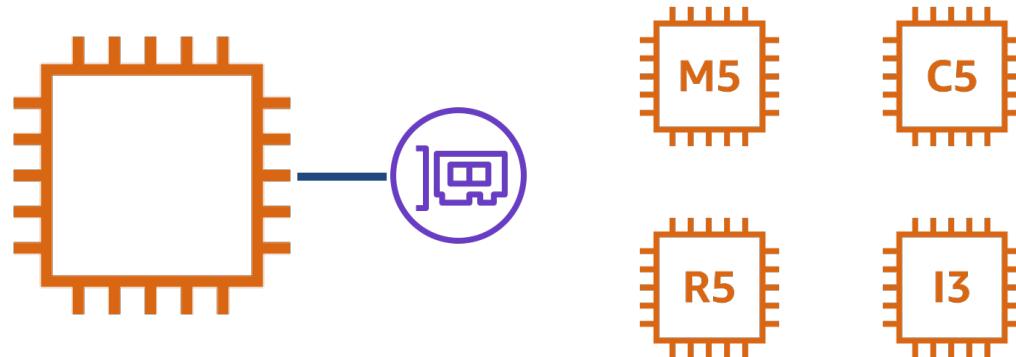
Avoid Inline Gateways



This is an anti-pattern because it will be a bottleneck and a single point of failure

EC2 Instance Type with ENI

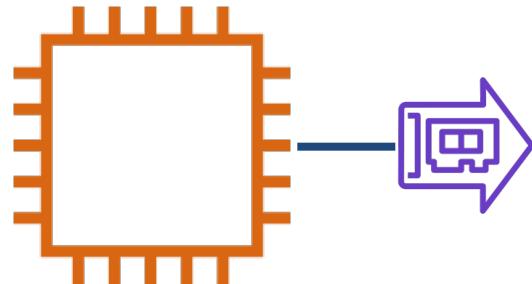
Maximum 25Gbps



Many instance types are limited to much lower throughput. Always test your infrastructure!

EC2 Instance Type with ENA

Maximum? ~100Gbps

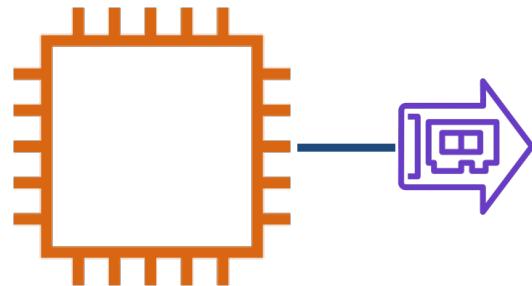


Instance types that end with “n”
Some are bare metal
Some use Nitro virtualization

Elastic Network Adapters require Enhanced Networking to be enabled on the instance

EC2 Instance Type with EFA

Maximum? ~100Gbps

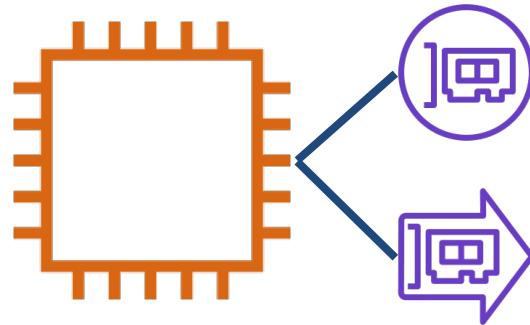


2 largest instance types of each family that end with “n”
MPI one-way latency as low as 15.5 microseconds

Elastic Fabric Adapter is an ENA with added capabilities, and benefits from Cluster Placement Groups

EC2 Enhanced Networking

Uses single root I/O virtualization



For ENI, just switch the kernel driver

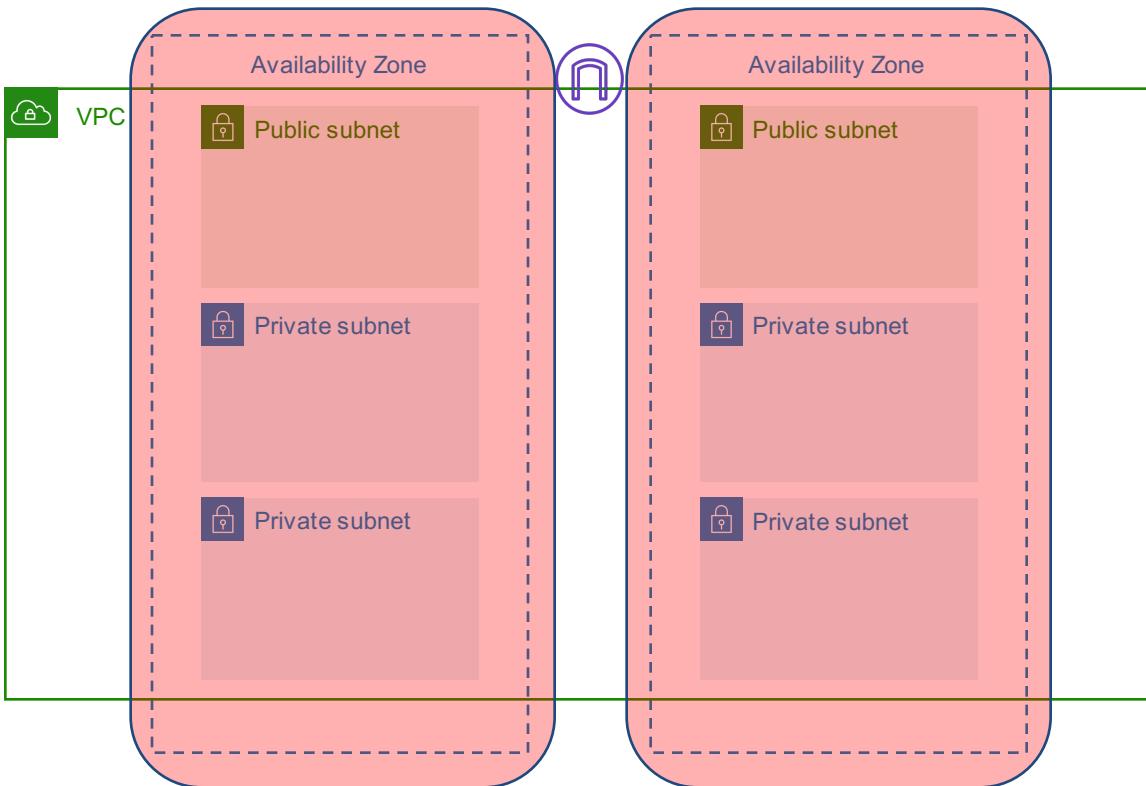
For ENA/EFA, switch the kernel driver then attach the resource

Higher bandwidth

Higher PPS

Lower latency

EC2 Placement Groups



Designed to influence
EC2 instance placement
within the region for
performance

Placement Groups can be
used within a single AZ or
across multiple AZ

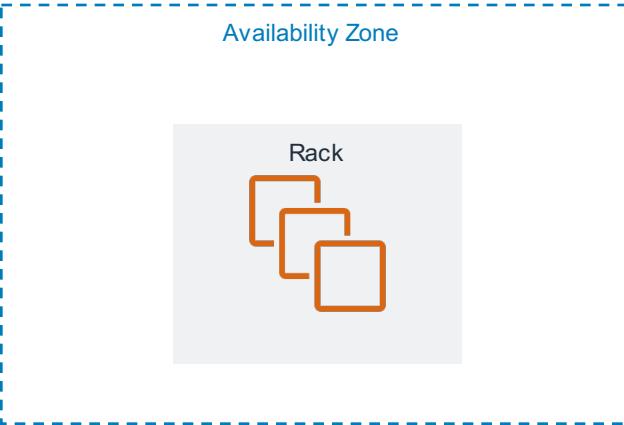
EC2 Cluster Placement Group

No blocking
No oversubscription

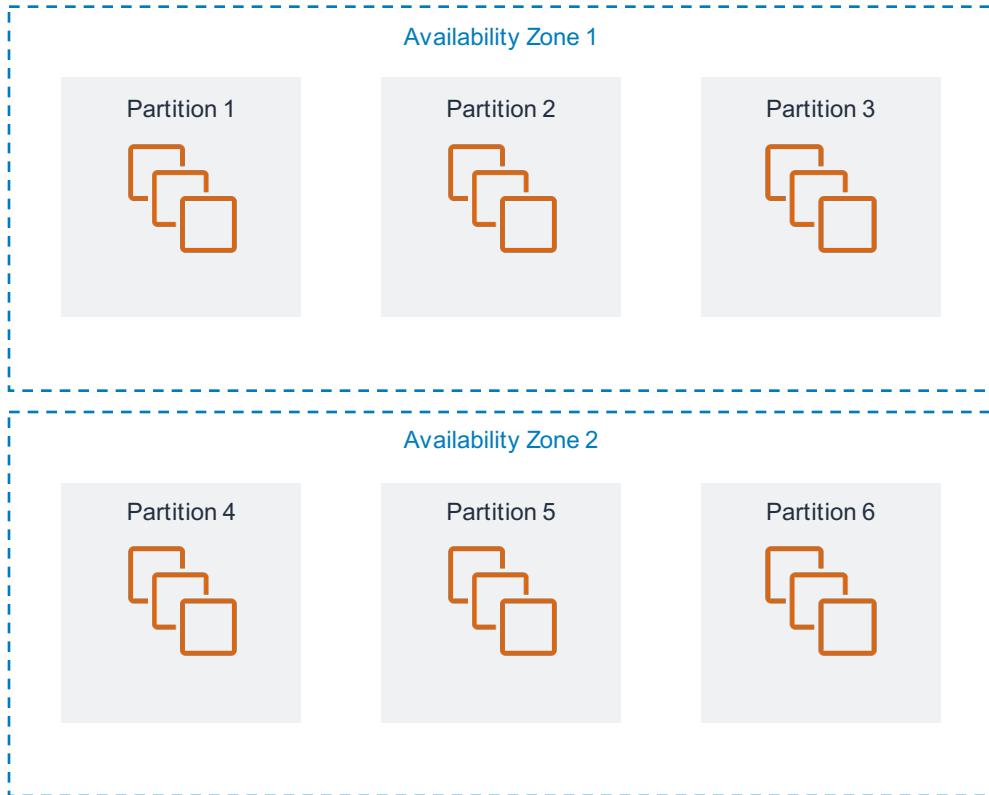
High throughput
Low latency

Majority of traffic
between group nodes

Not designed for
resilience



EC2 Partition Placement Group



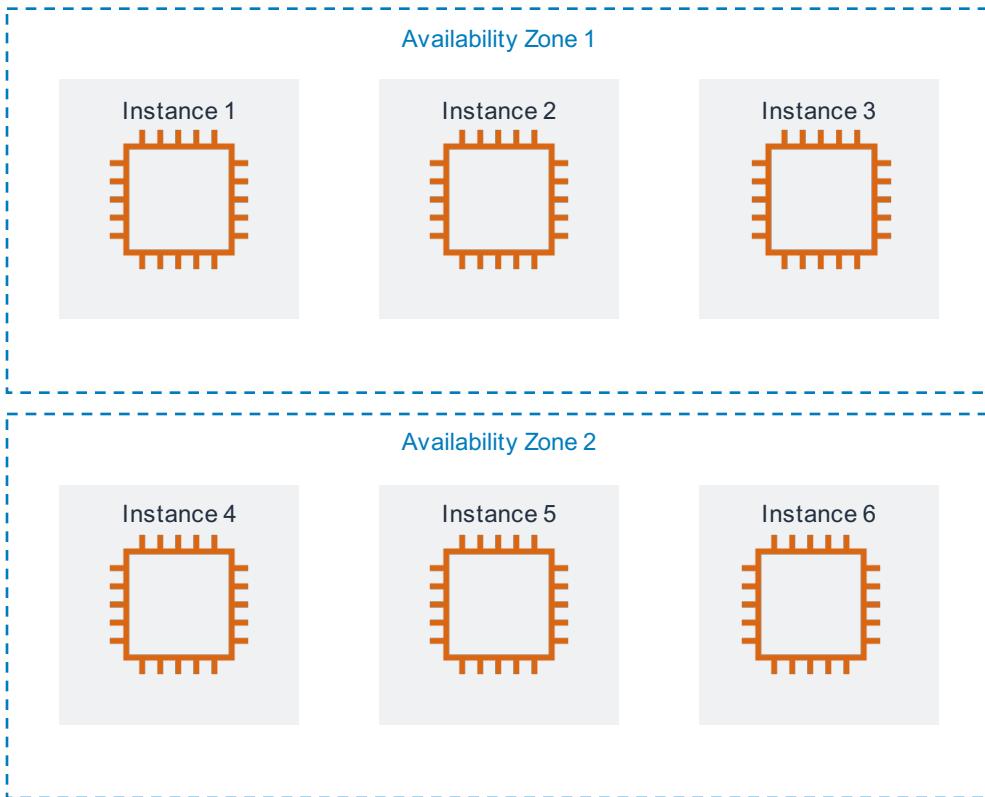
Partition = set of racks

Distributed + Replicated workloads

Can span multiple AZ

Designed for performance and resilience

EC2 Spread Placement Group



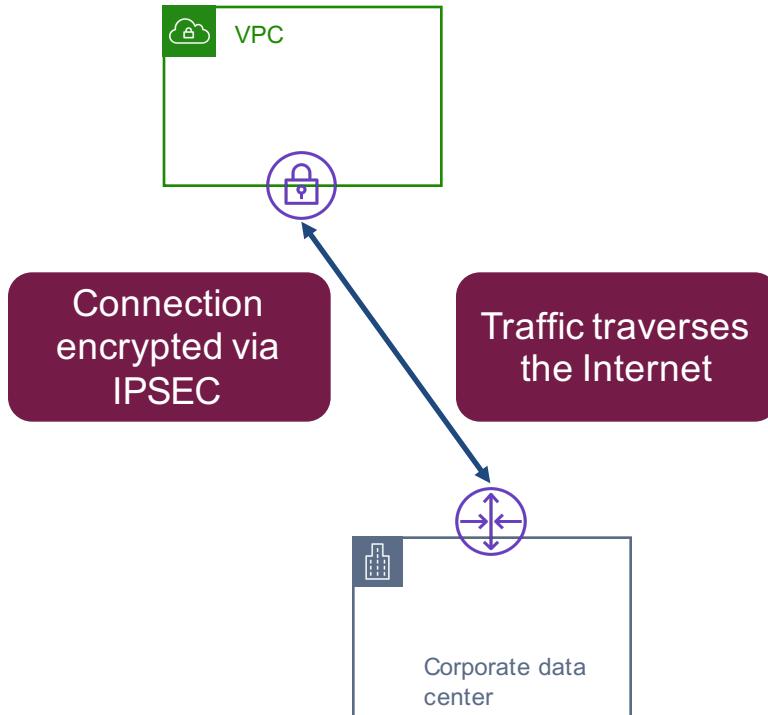
Each instance =
separate rack

Critical instances that
must be kept separate

Designed for guaranteed
resilience

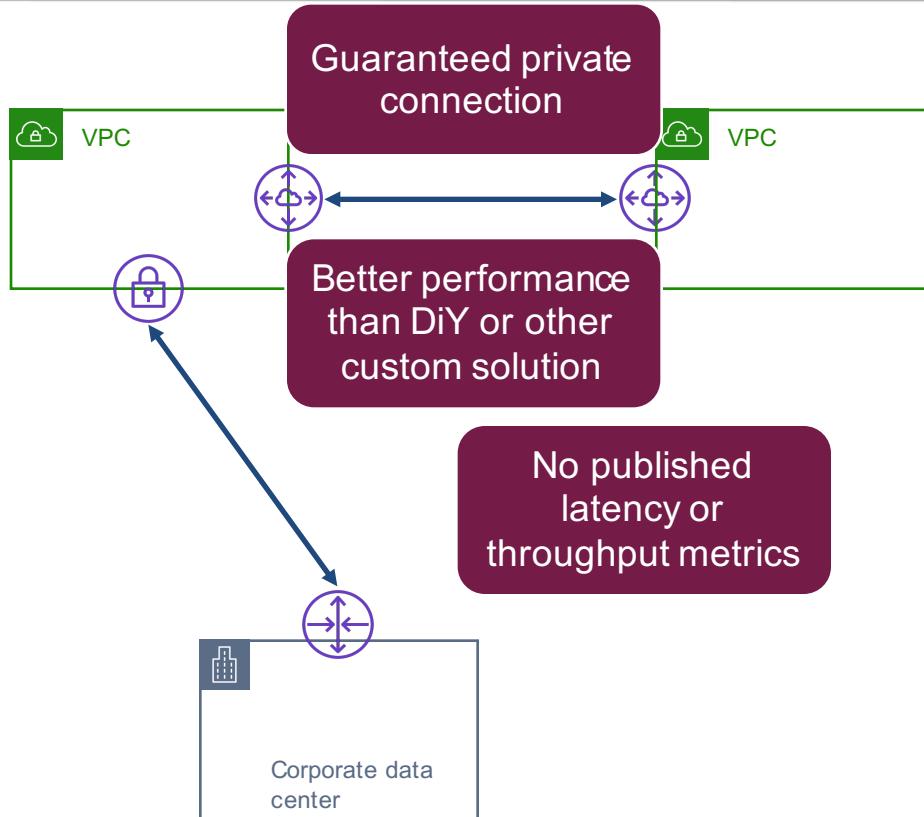
Can span multiple AZ

VPN Gateway (VPG)



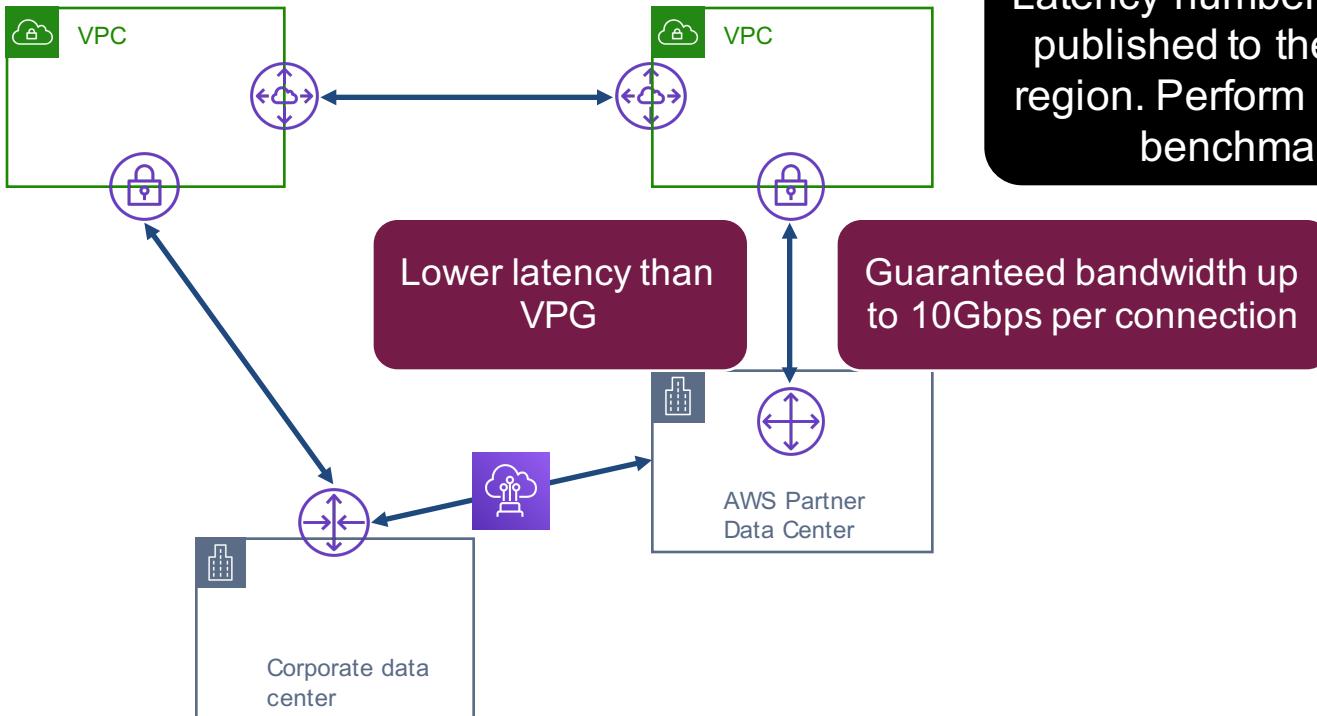
There is no way to guarantee specific throughput or latency through this solution

VPC Peering

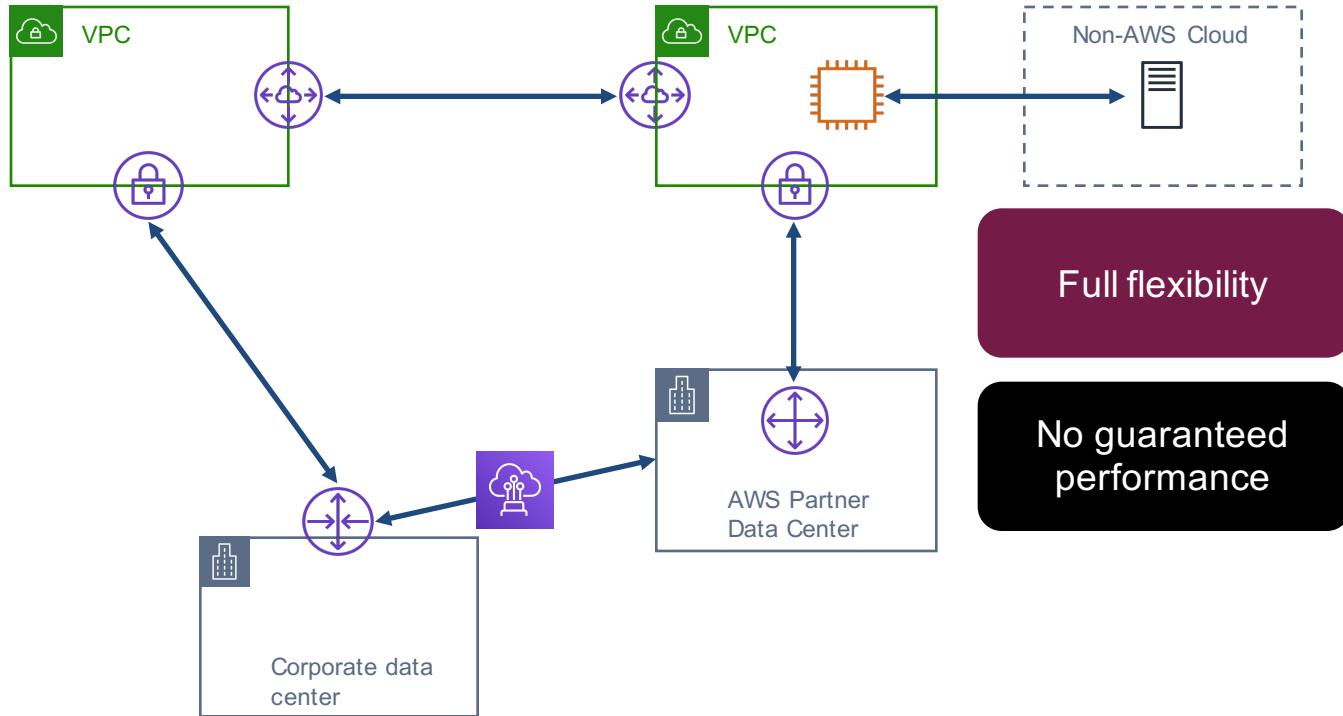


Cross-region bandwidth varies according to the regions involved. Test your connection to create a baseline!

Direct Connect Gateway



Do it Yourself



Question Breakdown

Question Breakdown - Key Terms

An application is deployed with Apache Kafka onto a fleet of EC2 instances. There are multiple Kafka topics and multiple partitions per topic. The application requires high performance and low latency. Which of the following recommendations would achieve this? (Choose two.)

- A. Use an EC2 Spread Placement Group during instance launch.
- B. Use an EC2 Cluster Placement Group during instance launch.
- C. Use an EC2 Partition Placement Group during instance launch.
- D. Configure jumbo frames on the EC2 instances.
- E. Use EC2 instance types that support Enhanced Networking.

Question Breakdown - Answers

An EC2 Spread Placement Group is only for a small group of instances, spreading them across distinct hardware to reduce the chance of correlated failures.

- A. Use an EC2 Spread Placement Group during instance launch.
- B. Use an EC2 Cluster Placement Group during instance launch.
- C. Use an EC2 Partition Placement Group during instance launch.
- D. Configure jumbo frames on the EC2 instances.
- E. Use EC2 instance types that support Enhanced Networking.

Question Breakdown - Answers

An EC2 Cluster Placement Group will pack instances close together within an Availability Zone for high performing node-to-node connections but not necessarily optimizing for outside communication.

- A. Use an EC2 Spread Placement Group during instance launch.
- B. Use an EC2 Cluster Placement Group during instance launch.
- C. Use an EC2 Partition Placement Group during instance launch.
- D. Configure jumbo frames on the EC2 instances.
- E. Use EC2 instance types that support Enhanced Networking.

Question Breakdown - Answers

An EC2 Partition Placement Group can be used to spread instances across hardware such that instances in one partition do not share underlying hardware with instances in other partitions, and is ideal for distributed and replicated workloads.

- A. Use an EC2 Spread Placement Group during instance launch.
- B. Use an EC2 Cluster Placement Group during instance launch.
- C. Use an EC2 Partition Placement Group during instance launch.
- D. Configure jumbo frames on the EC2 instances.
- E. Use EC2 instance types that support Enhanced Networking.

Question Breakdown - Answers

Jumbo frames allow for more data to be transferred with each packet, but will not impact latency of node-to-node communication. Furthermore, the larger frame size may end up fragmented if it leaves the VPC, resulting in zero net improvement.

- A. Use an EC2 Spread Placement Group during instance launch.
- B. Use an EC2 Cluster Placement Group during instance launch.
- C. Use an EC2 Partition Placement Group during instance launch.
- D. Configure jumbo frames on the EC2 instances.
- E. Use EC2 instance types that support Enhanced Networking.

Question Breakdown - Answers

Enhanced Networking uses a different network driver to achieve higher bandwidth, higher packet-per-second performance and lower inter-instance latencies.

- A. Use an EC2 Spread Placement Group during instance launch.
- B. Use an EC2 Cluster Placement Group during instance launch.
- C. Use an EC2 Partition Placement Group during instance launch.
- D. Configure jumbo frames on the EC2 instances.
- E. Use EC2 instance types that support Enhanced Networking.

Question Breakdown - Correct Answer

Correct Answers: C and E

- A. Use an EC2 Spread Placement Group during instance launch.
- B. Use an EC2 Cluster Placement Group during instance launch.
- C. Use an EC2 Partition Placement Group during instance launch.
- D. Configure jumbo frames on the EC2 instances.
- E. Use EC2 instance types that support Enhanced Networking.



Design Performant
Architectures, Part 2 of 2

Performant Databases

Performance

Unmanaged DB

Relational

EC2 unmanaged DB

Instance Storage

Lowest latency

High/Very High throughput

Fixed performance

EC2 unmanaged DB

EBS Storage, 1 volume

Low latency

Medium/High throughput

Some elasticity

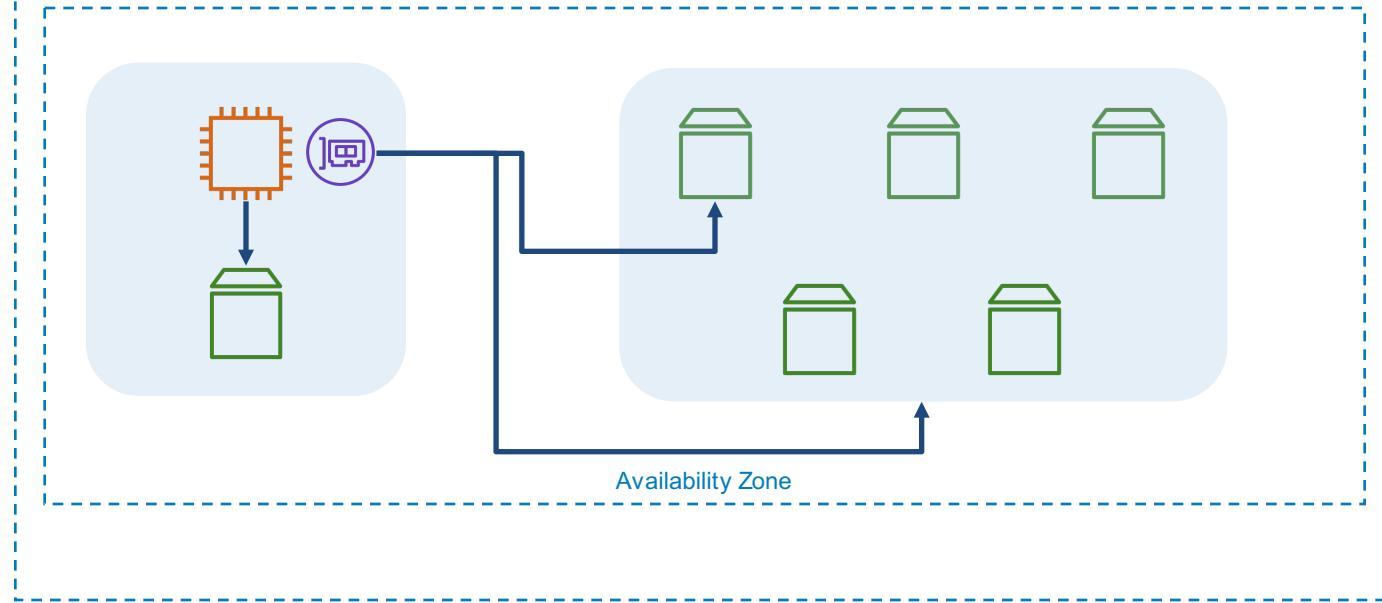
EC2 unmanaged DB

EBS Storage, 2+ volume

Low latency

Very High throughput

Significant elasticity

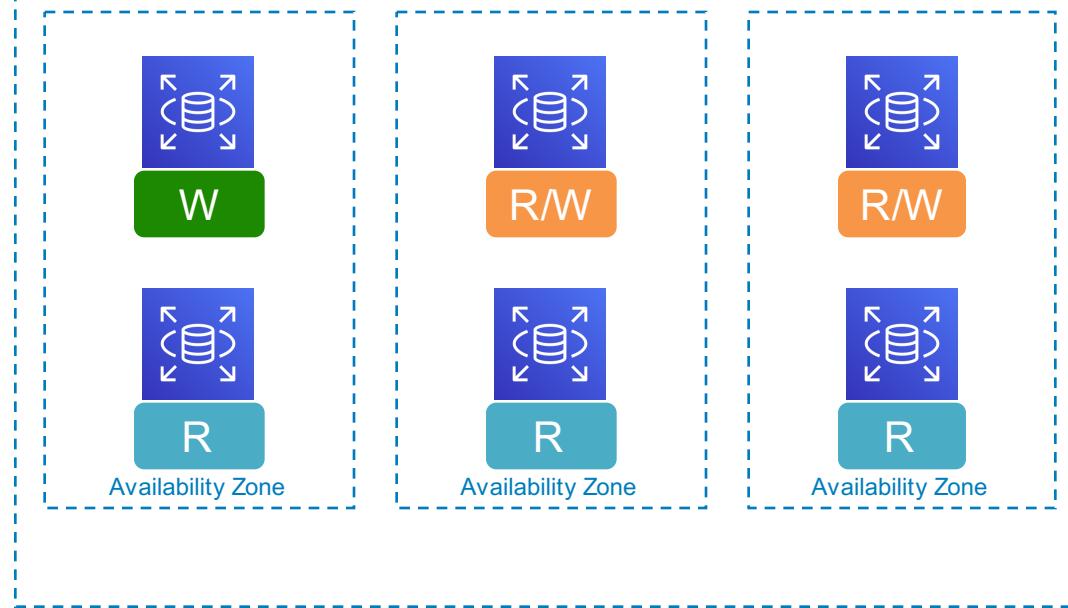


Performant Database - RDS

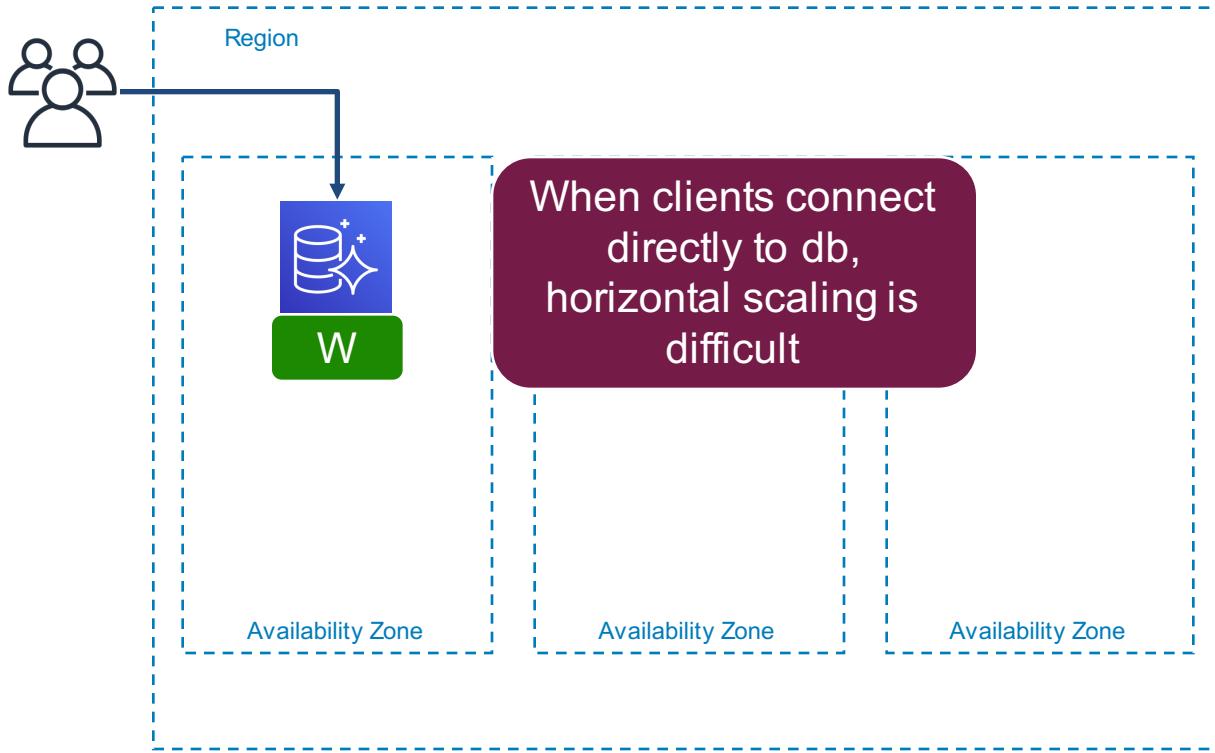
Upsize instance type
Upsize storage

Deploy read replicas

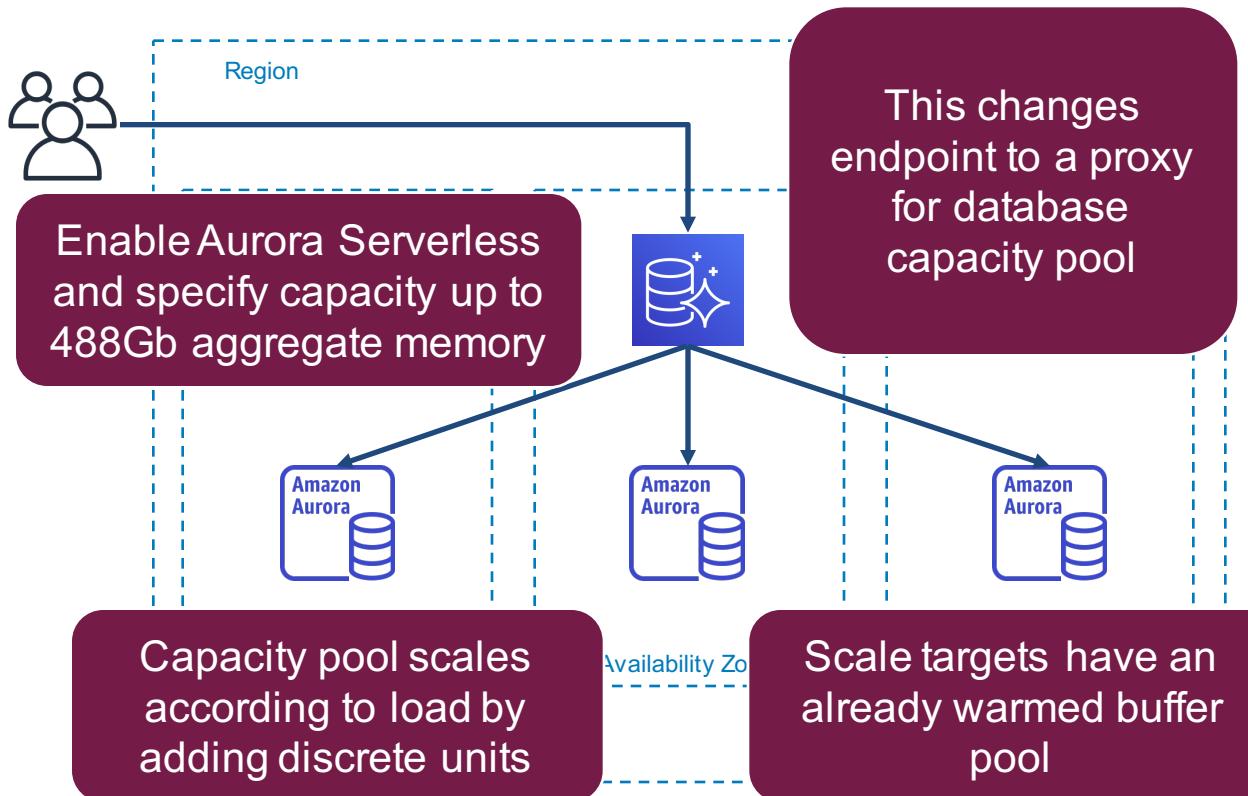
Deploy multi-master +
Read replicas
(Aurora only)



Aurora Database Performance



Aurora Database Performance



Performant Database - DynamoDB

Can we scale performance further?

Static Provisioning
Read Ops
Write Ops

Auto Scaling
Minimums
Scaling Thresholds
Maximums



Region

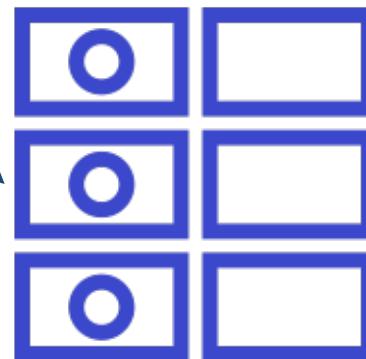
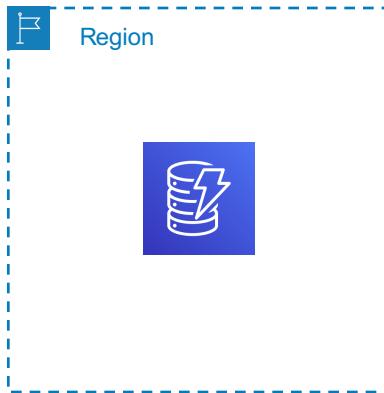


On demand
No capacity planning

Account service limits
Region service limits

Performant Database - DynamoDB

Partition key
Single most important
decision for
performance



Global tables
In accordance with
multiple best
practices!



Performant Storage and Databases

Question Breakdown

Question Breakdown - Key Terms

When migrating an on-premises legacy database, an AWS architect must recommend an **Oracle database** hosting option that supports **32Tb** of database storage and handles a **sustained load higher than 60,000 IOPS**. Which of the following choices should the architect recommend? (pick two)

- A. r5.12xlarge EC2 instance with multiple PIOPS EBS volumes configured as a striped RAID.
- B. r4.16xlarge EC2 instance with multiple PIOPS EBS volumes configured as a striped RAID.
- C. r4.16xlarge EC2 instance with a single GP2 EBS volume.
- D. db.r5.24xlarge RDS instance with PIOPS storage.
- E. db.r5.24xlarge RDS instance with GP2 storage.

Question Breakdown - Answers

The r5.12xlarge instance only supports 40,000 total IOPS across all EBS volumes and will not meet the requirement.

- A. r5.12xlarge EC2 instance with multiple PIOPS EBS volumes configured as a striped RAID.
- B. r4.16xlarge EC2 instance with multiple PIOPS EBS volumes configured as a striped RAID.
- C. r4.16xlarge EC2 instance with a single GP2 EBS volume.
- D. db.r5.24xlarge RDS instance with PIOPS storage.
- E. db.r5.24xlarge RDS instance with GP2 storage.

Question Breakdown - Answers

The r4.16xlarge instance type supports a total of 75,000 IOPS across all EBS volumes, and will require multiple volumes to achieve this.

- A. r5.12xlarge EC2 instance with multiple PIOPS EBS volumes configured as a striped RAID.
- B. r4.16xlarge EC2 instance with multiple PIOPS EBS volumes configured as a striped RAID.
- C. r4.16xlarge EC2 instance with a single GP2 EBS volume.
- D. db.r5.24xlarge RDS instance with PIOPS storage.
- E. db.r5.24xlarge RDS instance with GP2 storage.

Question Breakdown - Answers

A single General Purpose SSD volume only supports a maximum of 16,000 IOPS and thus will not meet the requirement.

- A. r5.12xlarge EC2 instance with multiple PIOPS EBS volumes configured as a striped RAID.
- B. r4.16xlarge EC2 instance with multiple PIOPS EBS volumes configured as a striped RAID.
- C. r4.16xlarge EC2 instance with a single GP2 EBS volume.
- D. db.r5.24xlarge RDS instance with PIOPS storage.
- E. db.r5.24xlarge RDS instance with GP2 storage.

Question Breakdown - Answers

The db.r5.24xlarge instance type supports up to 80,000 IOPS assuming the storage is PIOPS, and does meet the requirement.

- A. r5.12xlarge EC2 instance with multiple PIOPS EBS volumes configured as a striped RAID.
- B. r4.16xlarge EC2 instance with multiple PIOPS EBS volumes configured as a striped RAID.
- C. r4.16xlarge EC2 instance with a single GP2 EBS volume.
- D. db.r5.24xlarge RDS instance with PIOPS storage.
- E. db.r5.24xlarge RDS instance with GP2 storage.

Question Breakdown - Answers

While the db instance type does support up to 80,000 IOPS, the General Purpose SSD storage only supports up to 16,000 IOPS.

- A. r5.12xlarge EC2 instance with multiple PIOPS EBS volumes configured as a striped RAID.
- B. r4.16xlarge EC2 instance with multiple PIOPS EBS volumes configured as a striped RAID.
- C. r4.16xlarge EC2 instance with a single GP2 EBS volume.
- D. db.r5.24xlarge RDS instance with PIOPS storage.
- E. db.r5.24xlarge RDS instance with GP2 storage.

Question Breakdown - Correct Answer

Correct Answers: B and D

- A. r5.12xlarge EC2 instance with multiple PIOPS EBS volumes configured as a striped RAID.
- B. r4.16xlarge EC2 instance with multiple PIOPS EBS volumes configured as a striped RAID.
- C. r4.16xlarge EC2 instance with a single GP2 EBS volume.
- D. db.r5.24xlarge RDS instance with PIOPS storage.
- E. db.r5.24xlarge RDS instance with GP2 storage.



**Specify Secure Applications and
Architectures**
24%



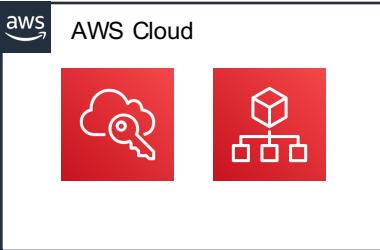
Specify Secure Applications and
Architectures

Securing AWS Account
Resources

Securing Account Resources

AWS SSO
account
access
federation

Organizations
Manage multiple AWS
accounts



Manage account
access via SCPs

IAM policies and
permission
boundaries

IAM Roles for
temporary privileges

S3 Bucket Policy



Only applies to the associated bucket

Required for some functions

Watch for S3 Block Public Access override

Glacier Vault Access Policy



Only applies to the associated vault

Similar to S3 Bucket Policy

Vault Lock Policy is separate

KMS CMK Key Policy



Applies to specific CMK

Required for all CMKs

Key Admin and Key User
permissions

Lambda Function Access Policy

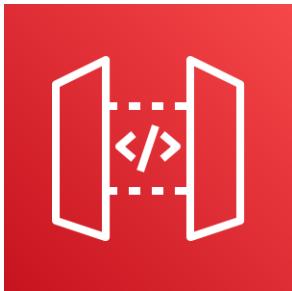


Only applies to the associated function

Can be used to share layers

Be careful of IAM vs. Function Policy interaction

API Gateway Resource Policy



Only applies to the associated API

Usable for all endpoint types

Can affect authorization workflows

SNS Access Policy



Only applies to the associated topic

Required for some use cases (Budgets)

Overrides IAM in some cases

Securing Application Tiers

EASY Question Breakdown

Question Breakdown

Which of these would be an appropriate least-privilege policy addition for an SCP to be applied to all accounts in an AWS Organization? (pick two)

- A. Grant application specific permissions for one of the AWS accounts
- B. Deny deletes on CloudTrail log objects in S3
- C. Grant administrative permissions to all IT admin staff
- D. Deny IAM Policy changes that grant administrative permissions

Question Breakdown - Correct Answer

Correct Answers: B and D

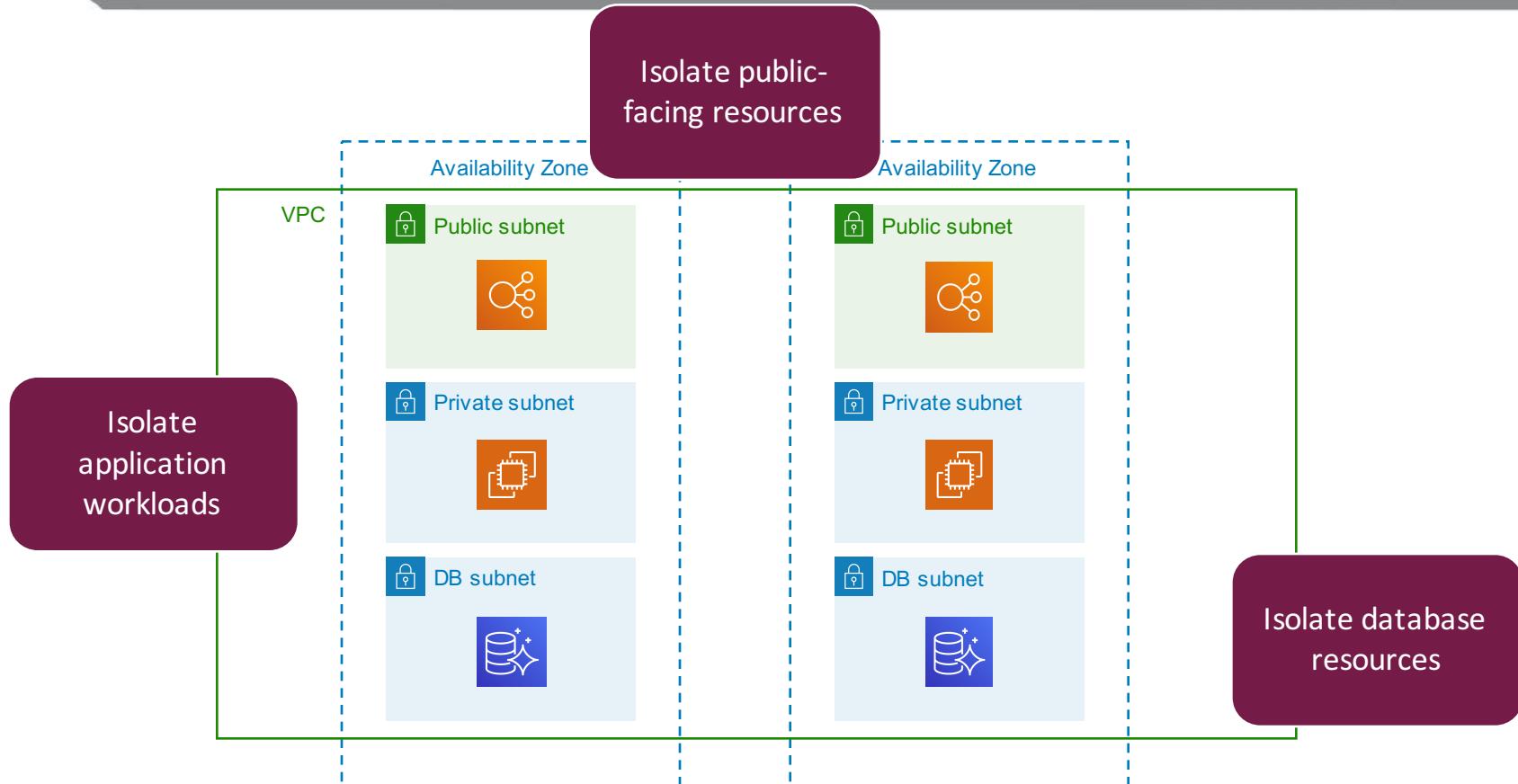
- A. Grant application specific permissions for one of the AWS accounts
- B. Deny deletes on CloudTrail log objects in S3
- C. Grant administrative permissions to all IT admin staff
- D. Deny IAM Policy changes that grant administrative permissions



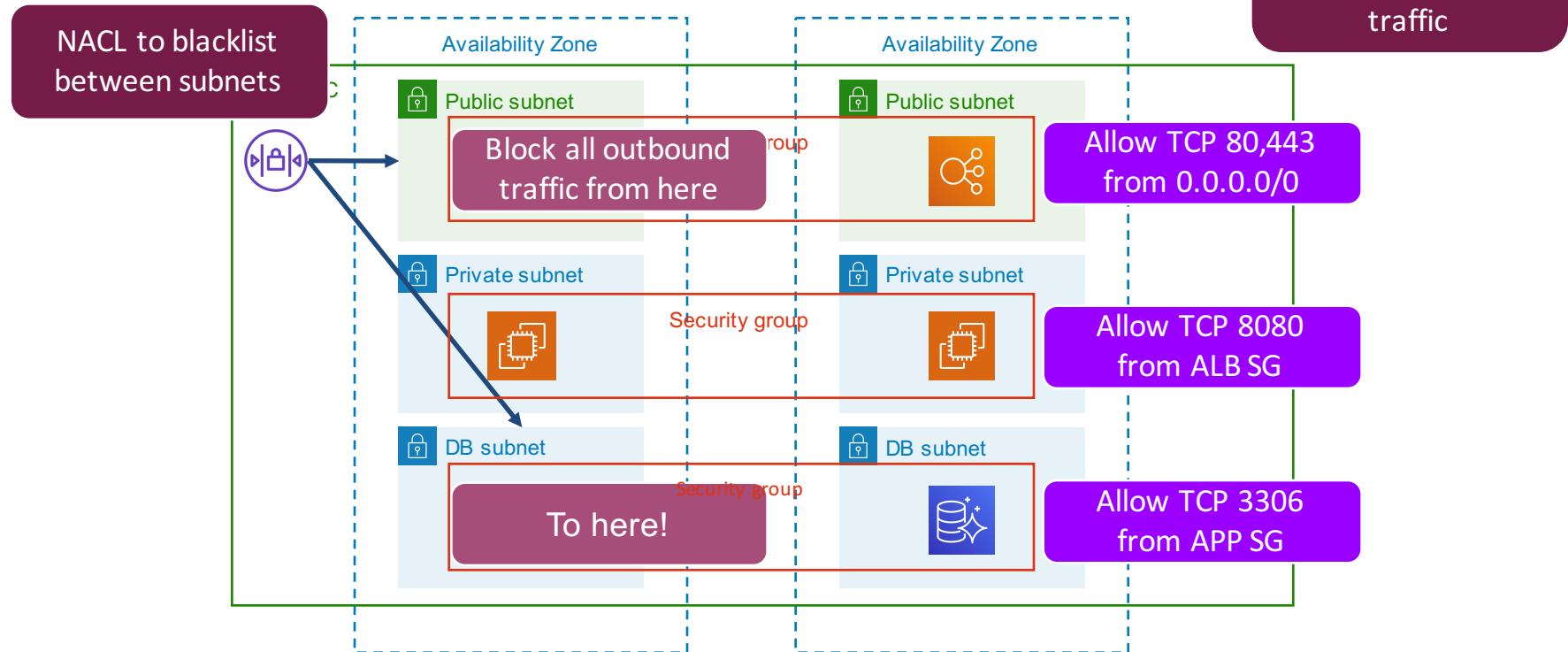
Specify Secure Applications and
Architectures

Securing Networks

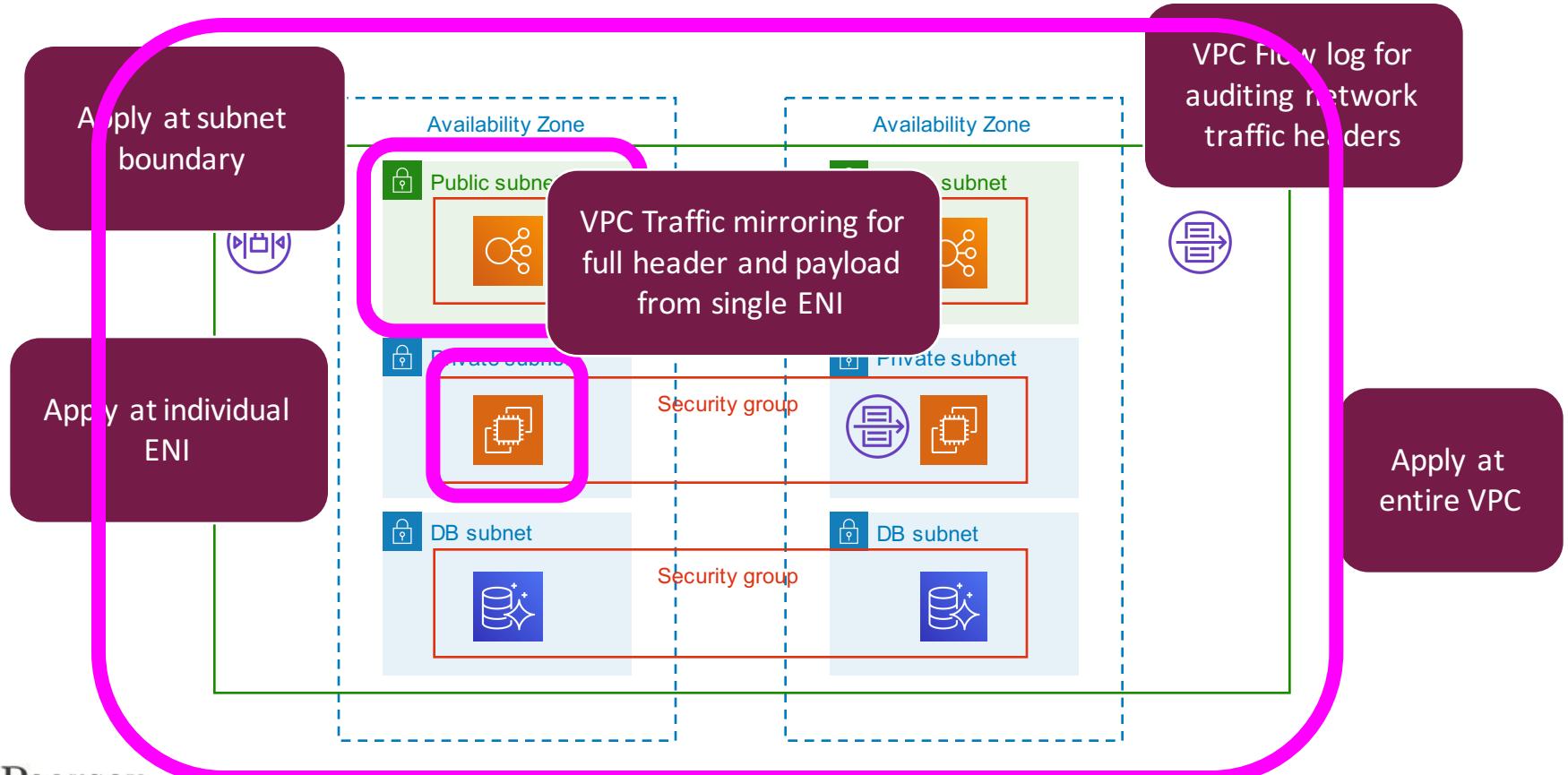
Secure VPC Internal Network



Secure VPC Internal Network



Secure VPC Internal Network



Single VPC Network Security

EASY Question Breakdown

Question Breakdown

Which of the following acts as a stateless firewall function in a VPC?

- A. Route Table
- B. Network Access Control List
- C. Security Group
- D. NAT Gateway

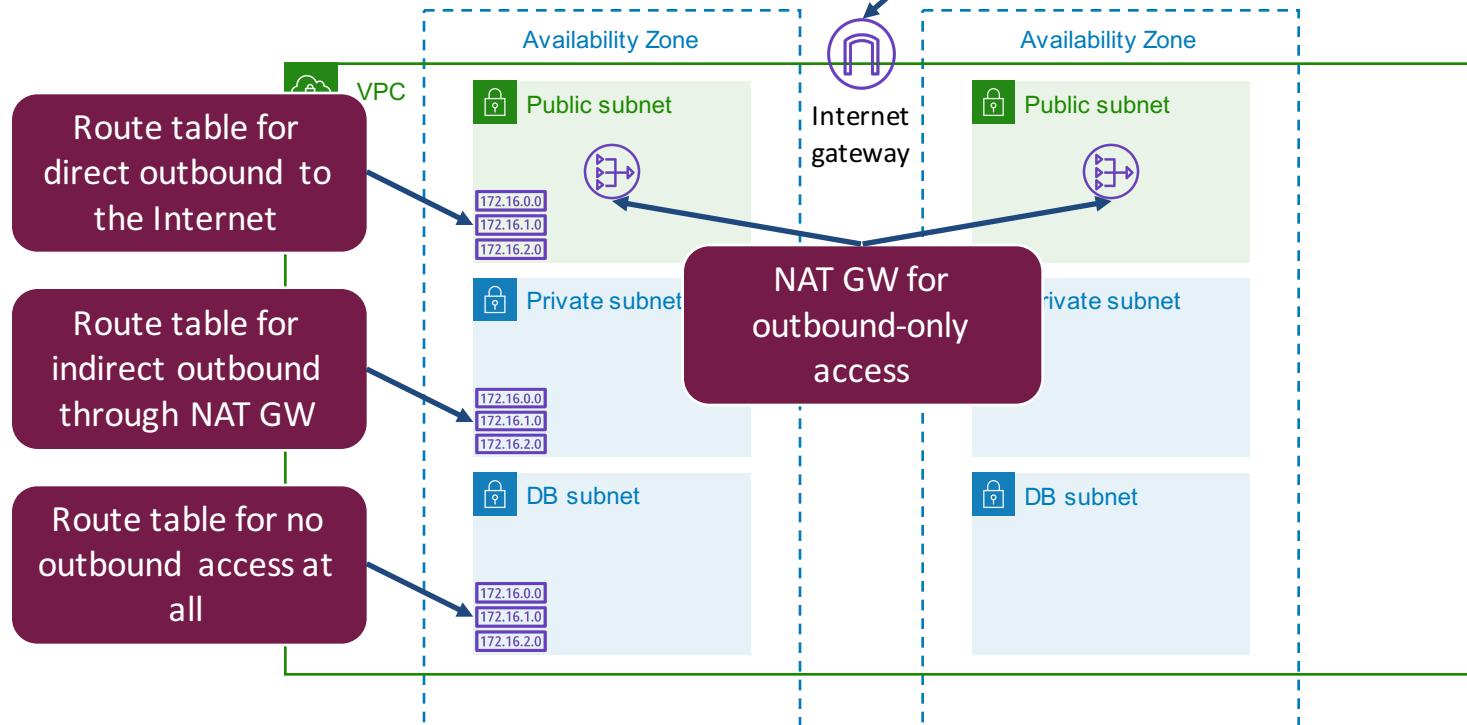
Question Breakdown - Correct Answer

Correct Answer: B

- A. Route Table
- B. Network Access Control List
- C. Security Group
- D. NAT Gateway

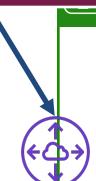
Secure VPC Egress

Internet Gateway
can now attach
route tables!
(12/2019)

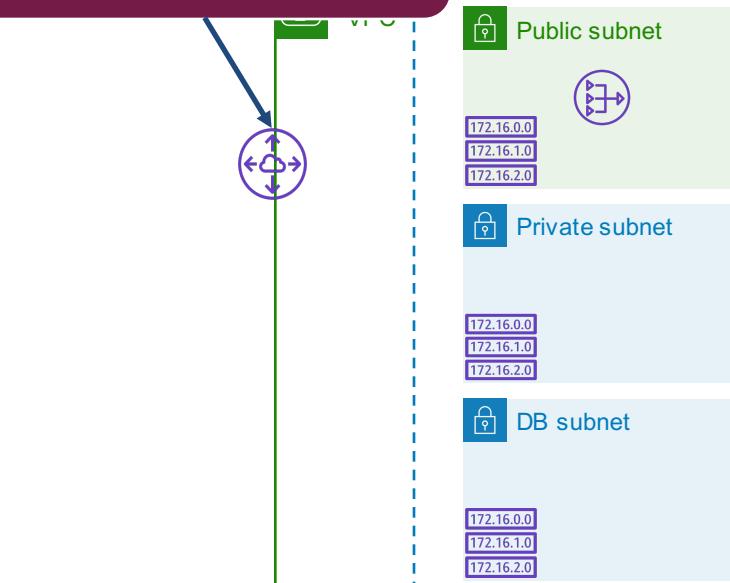


Secure VPC Egress

VPC Peering
Private network access to another VPC



Availability Zone



Gateway Endpoint
Private network access to S3 and DynamoDB



Interface Endpoint
Private network access to other services and PrivateLink



Virtual Private Gateway
VPN to outside network
Can attach route tables
(12/2019)



Single-VPC Network Security

Question Breakdown

Question Breakdown - Key Terms

As an AWS network architect, you've been asked to design a **VPC** that must host the following: 1) **ALB** front end, 2) Docker containers managed by **ECS**, and 3) **RDS Aurora** database. Which of the following VPC security strategies would ensure the **greatest security control** over each of the application tiers?

- A. All applications in the same public subnets. Isolate workloads via Security Groups.
- B. Each application in dedicated subnets (ALB - public, ECS - private, RDS - private). Isolate workloads via Security Groups and NACLs
- C. ALB and ECS containers in the same public subnets, RDS in dedicated private subnets. Isolate workloads via Security Groups and NACLs.
- D. ALB in dedicated public subnets, ECS and RDS colocated in the same private subnets. Isolate workloads via Security Groups and NACLs.

Question Breakdown - Answers

While putting all app tiers in the same subnet might simplify operations, it does not provide granular control over security, especially if they are all public subnets.

- A. All applications in the same public subnets. Isolate workloads via Security Groups.
- B. Each application in dedicated subnets (ALB - public, ECS - private, RDS - private). Isolate workloads via Security Groups and NACLs
- C. ALB and ECS containers in the same public subnets, RDS in dedicated private subnets. Isolate workloads via Security Groups and NACLs.
- D. ALB in dedicated public subnets, ECS and RDS colocated in the same private subnets. Isolate workloads via Security Groups and NACLs.

Question Breakdown - Answers

Separating each tier into their own subnet will allow for individual controls to be applied. Placing ECS and RDS in private subnets will improve security by denying all direct inbound traffic from the Internet.

- A. All applications in the same public subnets. Isolate workloads via Security Groups.
- B. Each application in dedicated subnets (ALB - public, ECS - private, RDS - private). Isolate workloads via Security Groups and NACLs
- C. ALB and ECS containers in the same public subnets, RDS in dedicated private subnets. Isolate workloads via Security Groups and NACLs.
- D. ALB in dedicated public subnets, ECS and RDS colocated in the same private subnets. Isolate workloads via Security Groups and NACLs.

Question Breakdown - Answers

By combining ALB and ECS into the same subnets, it is more difficult to apply network security to either.

- A. All applications in the same public subnets. Isolate workloads via Security Groups.
- B. Each application in dedicated subnets (ALB - public, ECS - private, RDS - private). Isolate workloads via Security Groups and NACLs
- C. ALB and ECS containers in the same public subnets, RDS in dedicated private subnets. Isolate workloads via Security Groups and NACLs.
- D. ALB in dedicated public subnets, ECS and RDS colocated in the same private subnets. Isolate workloads via Security Groups and NACLs.

Question Breakdown - Answers

Separating the ALB from the ECS and RDS tiers is good guidance, by keeping the public-facing resources separate from the important application and database resources. Combining the ECS and RDS tiers into the same subnet will increase the difficulty of applying specific security controls to either.

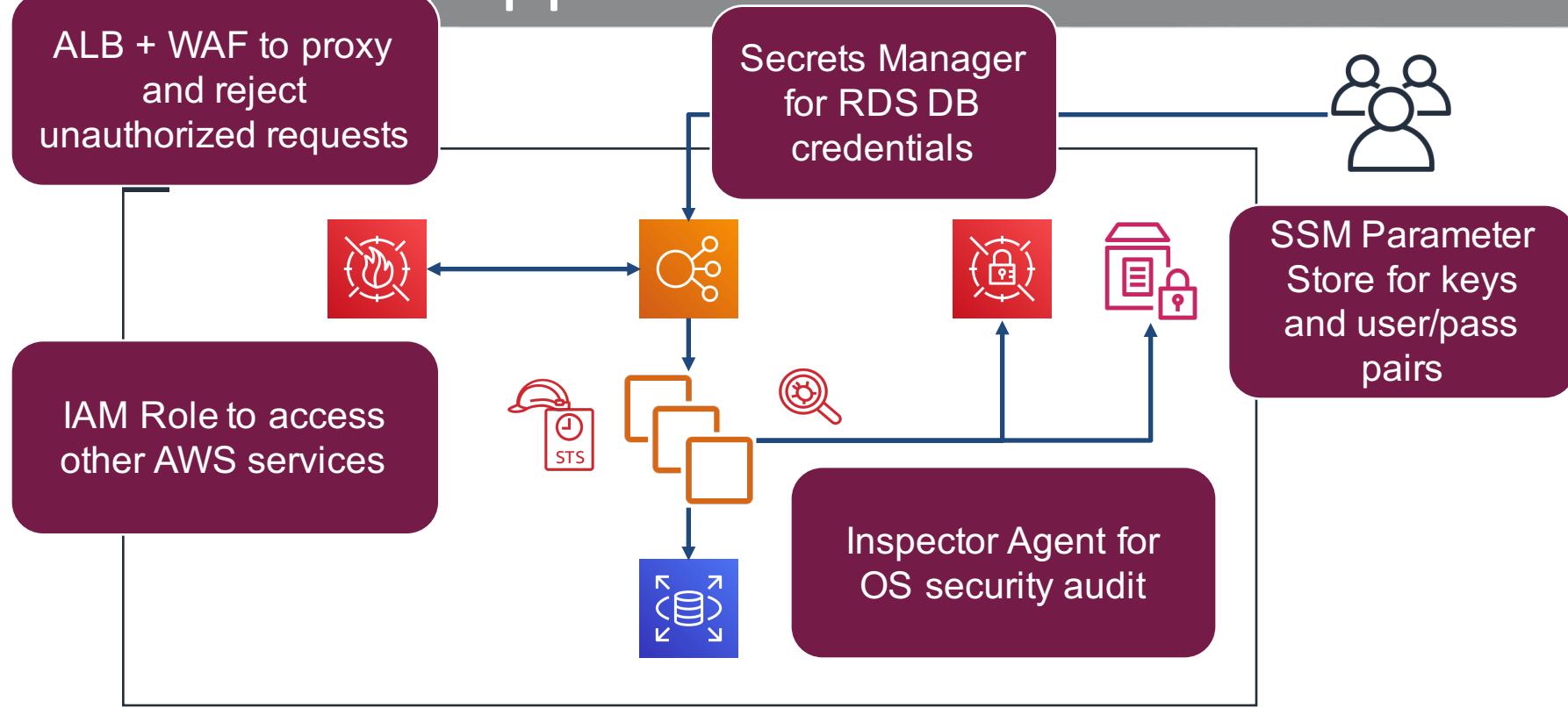
- A. All applications in the same public subnets. Isolate workloads via Security Groups.
- B. Each application in dedicated subnets (ALB - public, ECS - private, RDS - private). Isolate workloads via Security Groups and NACLs
- C. ALB and ECS containers in the same public subnets, RDS in dedicated private subnets. Isolate workloads via Security Groups and NACLs.
- D. ALB in dedicated public subnets, ECS and RDS colocated in the same private subnets. Isolate workloads via Security Groups and NACLs.

Question Breakdown - Correct Answer

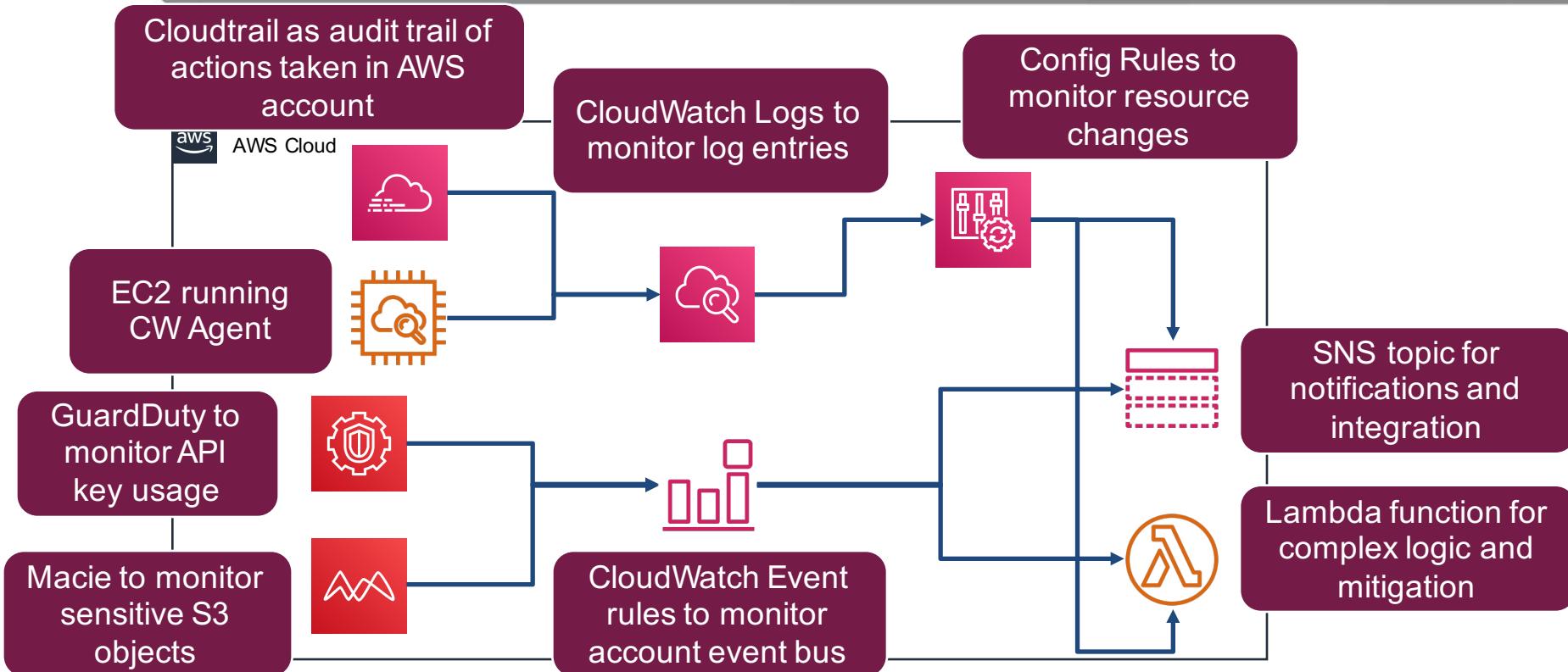
Correct Answer: B

- A. All applications in the same public subnets. Isolate workloads via Security Groups.
- B. Each application in dedicated subnets (ALB - public, ECS - private, RDS - private). Isolate workloads via Security Groups and NACLs
- C. ALB and ECS containers in the same public subnets, RDS in dedicated private subnets. Isolate workloads via Security Groups and NACLs.
- D. ALB in dedicated public subnets, ECS and RDS colocated in the same private subnets. Isolate workloads via Security Groups and NACLs.

Securing Application Access



Monitoring Application Activity



Securing Application Tiers

Question Breakdown

Question Breakdown

Your team supports a Java-based application that uses a JDBC connection to an RDS database running MySQL. The connection string contains hard-coded credentials. You've been asked to improve the security of the database credentials, and must account for a new 30-day password rotation policy on RDS. Which of the following meet the requirements with the least ongoing overhead?

- A. Move the database credentials to a text file on each instance. Read the text file upon application start. Update the text file on each instance when password is rotated.
- B. Move the database credentials to SSM Parameter Store. Read the Parameter upon application start. Update the Parameter when password is rotated.
- C. Move the database credentials to AWS Secrets Manager. Read the Secret upon application start. Configure the Secret to rotate automatically.
- D. Move the database credentials to S3. Download the object upon application start. Update the S3 object when password is rotated.

Question Breakdown - Key Terms

Your team supports a Java-based application that uses a **JDBC connection** to an **RDS database** running **MySQL**. The connection string contains **hard-coded credentials**. You've been asked to **improve the security** of the database credentials, and must account for a new **30-day password rotation** policy on RDS. Which of the following meet the requirements with the **least ongoing overhead**?

- A. Move the database credentials to a text file on each instance. Read the text file upon application start. Update the text file on each instance when password is rotated.
- B. Move the database credentials to SSM Parameter Store. Read the Parameter upon application start. Update the Parameter when password is rotated.
- C. Move the database credentials to AWS Secrets Manager. Read the Secret upon application start. Configure the Secret to rotate automatically.
- D. Move the database credentials to S3. Download the object upon application start. Update the S3 object when password is rotated.

Question Breakdown - Answers

This will improve security by removing the text file from the source code repository, but the improvement is only marginal because it is still stored in plaintext.

- A. Move the database credentials to a text file on each instance. Read the text file upon application start. Update the text file on each instance when password is rotated.
- B. Move the database credentials to SSM Parameter Store. Read the Parameter upon application start. Update the Parameter when password is rotated.
- C. Move the database credentials to AWS Secrets Manager. Read the Secret upon application start. Configure the Secret to rotate automatically.
- D. Move the database credentials to S3. Download the object upon application start. Update the S3 object when password is rotated.

Question Breakdown - Answers

Similar to A, this improves security, with the added benefit of removing the credentials from the individual instances. The parameter can be updated in a single location, which makes operations easier.

- A. Move the database credentials to a text file on each instance. Read the text file upon application start. Update the text file on each instance when password is rotated.
- B. Move the database credentials to SSM Parameter Store. Read the Parameter upon application start. Update the Parameter when password is rotated.
- C. Move the database credentials to AWS Secrets Manager. Read the Secret upon application start. Configure the Secret to rotate automatically.
- D. Move the database credentials to S3. Download the object upon application start. Update the S3 object when password is rotated.

Question Breakdown - Answers

This will improve security by removing the credentials from the application code AND from the instance, and reduces operational overhead by enabling automated password rotation.

- A. Move the database credentials to a text file on each instance. Read the text file upon application start. Update the text file on each instance when password is rotated.
- B. Move the database credentials to SSM Parameter Store. Read the Parameter upon application start. Update the Parameter when password is rotated.
- C. Move the database credentials to AWS Secrets Manager. Read the Secret upon application start. Configure the Secret to rotate automatically.
- D. Move the database credentials to S3. Download the object upon application start. Update the S3 object when password is rotated.

Question Breakdown - Answers

This improves security by removing credentials from the source code and instance. S3 has a number of mechanisms for protecting objects which give flexibility for access control.

- A. Move the database credentials to a text file on each instance. Read the text file upon application start. Update the text file on each instance when password is rotated.
- B. Move the database credentials to SSM Parameter Store. Read the Parameter upon application start. Update the Parameter when password is rotated.
- C. Move the database credentials to AWS Secrets Manager. Read the Secret upon application start. Configure the Secret to rotate automatically.
- D. Move the database credentials to S3. Download the object upon application start. Update the S3 object when password is rotated.

Question Breakdown - Correct Answer

Correct Answer: C

- A. Move the database credentials to a text file on each instance. Read the text file upon application start. Update the text file on each instance when password is rotated.
- B. Move the database credentials to SSM Parameter Store. Read the Parameter upon application start. Update the Parameter when password is rotated.
- C. Move the database credentials to AWS Secrets Manager. Read the Secret upon application start. Configure the Secret to rotate automatically.
- D. Move the database credentials to S3. Download the object upon application start. Update the S3 object when password is rotated.



Specify Secure Applications and
Architectures

Securing Data

Securing Data At-Rest

KMS for encryption key management



VPC

EFS volume encryption

RDS storage encryption

EBS volume encryption

What do all of these encryption features have in common?

RDS database encryption MSSQL and Oracle

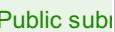
S3 SSE and bucket policy

Glacier SSE, vault access policy and vault lock policy

RedShift database encryption



Availability Zone



Availability Zone



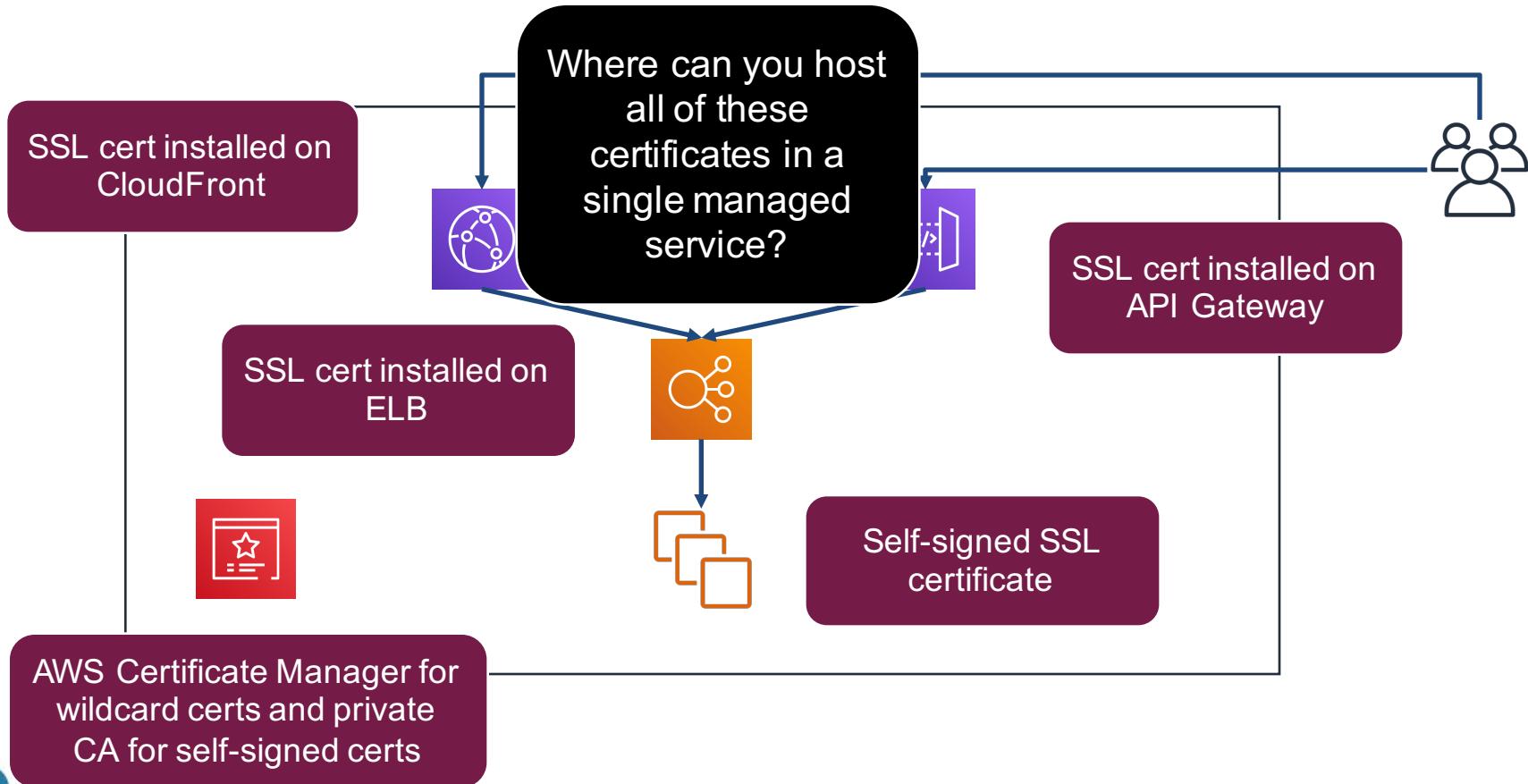
Private subnet



Private subnet



Securing Data In Transit - SSL



Securing Data

EASY Question Breakdown

Question Breakdown

Which of these services/features has no capability for enforcing encryption for data at rest?

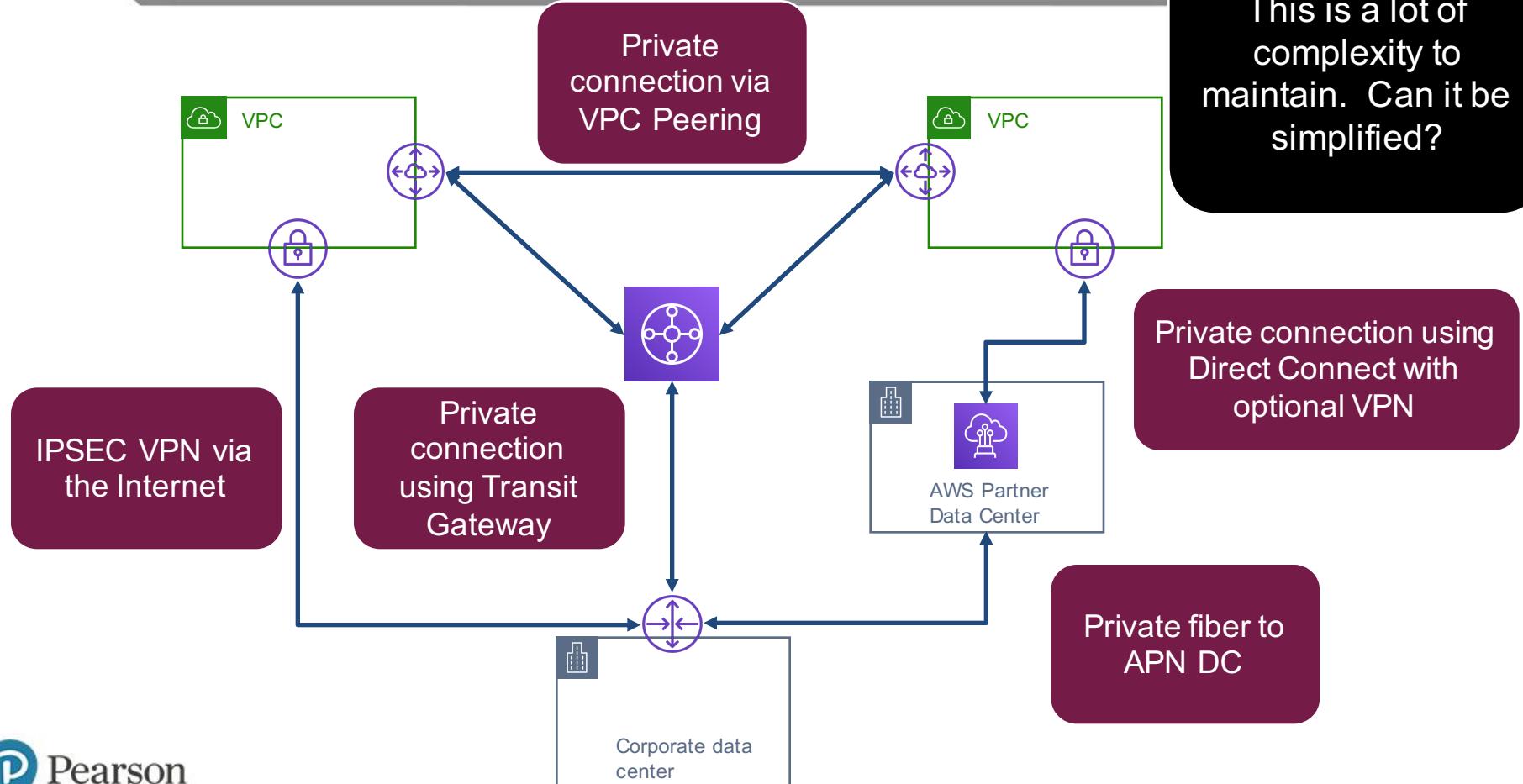
- A. S3
- B. DynamoDB
- C. Storage Gateway
- D. EBS
- E. Instance-store

Question Breakdown - Correct Answer

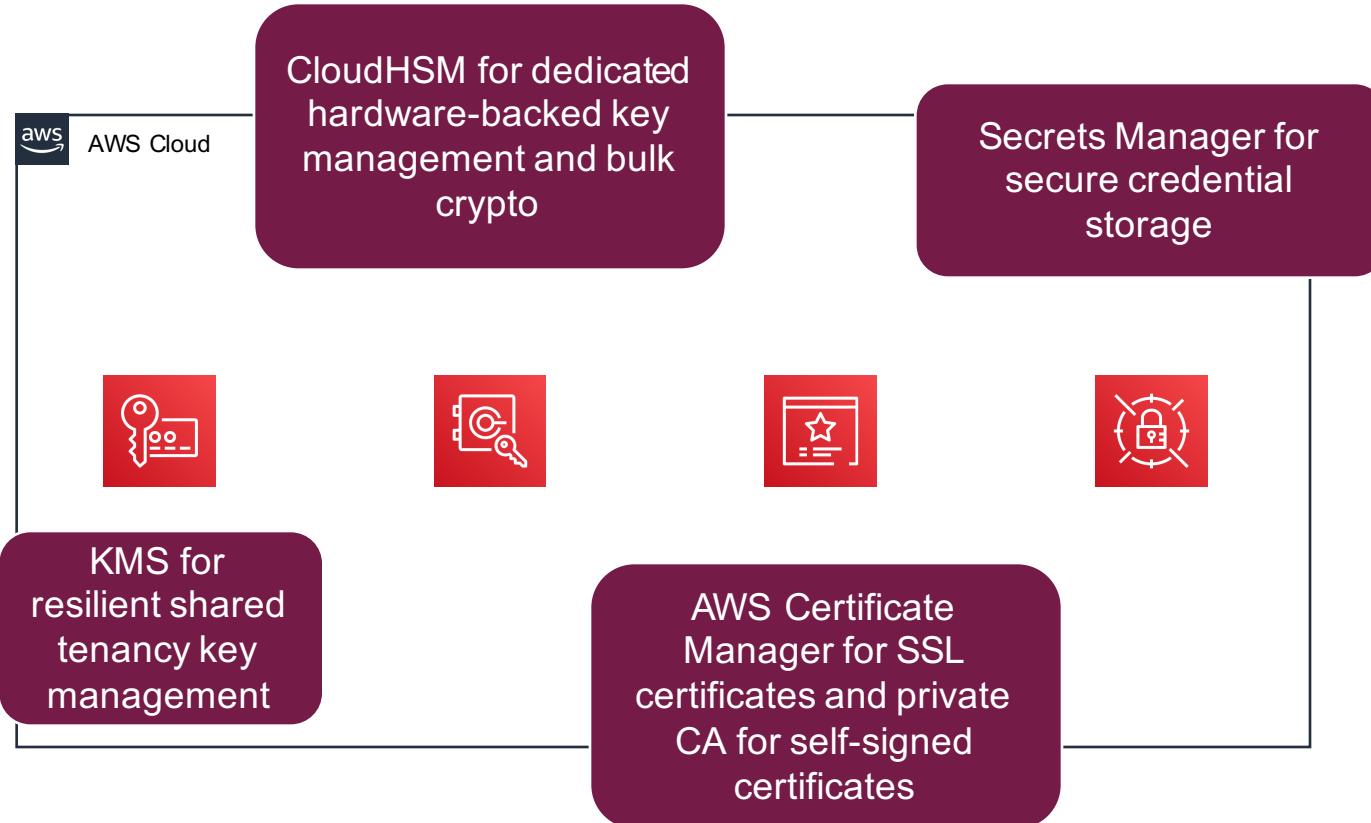
Correct Answer: E

- A. S3
- B. DynamoDB
- C. Storage Gateway
- D. EBS
- E. Instance-store

Securing Data in Transit - Network



Key Management Services



Securing Data

Question Breakdown

Question Breakdown

A company is building a data lake containing healthcare data that must be properly secured. The data will be stored in S3 and accessed by users who will be separated into two groups: 1) those that can view PHI (protected health information), and 2) those that cannot. Which of the following strategies will meet the requirements using least privilege techniques and low operational overhead? (pick two)

- A. Tag all S3 buckets and objects to indicate the presence of PHI. Create IAM Policies and S3 bucket policies using conditions based on the tags.
- B. Create an S3 full-access IAM policy and associate with users requiring PHI access. Create a more restrictive IAM policy for the non-PHI users.
- C. Tag all IAM users based on PHI access. Test for those tags using IAM Policy and S3 bucket policy conditions.
- D. Write an application to interface with S3 and implement access using custom code. Create IAM Policies and S3 bucket policies to allow access only through the application.

Question Breakdown - Key Terms

A company is building a **data lake** containing **healthcare data** that must be properly secured. The data will be stored in **S3** and accessed by users who will be separated into **two groups**: 1) those that **can view PHI** (protected health information), and 2) those that **cannot**. Which of the following strategies will meet the requirements using **least privilege techniques** and **low operational overhead**? (pick two)

- A. Tag all S3 buckets and objects to indicate the presence of PHI. Create IAM Policies and S3 bucket policies using conditions based on the tags.
- B. Create an S3 full-access IAM policy and associate with users requiring PHI access. Create a more restrictive IAM policy for the non-PHI users.
- C. Tag all IAM users based on PHI access. Test for those tags using IAM Policy and S3 bucket policy conditions.
- D. Write an application to interface with S3 and implement access using custom code. Create IAM Policies and S3 bucket policies to allow access only through the application.

Question Breakdown - Answers

Tagging alone is not a security control, but it can be used as a building block toward least privilege. Tags can be tested using policy conditions and are an appropriate way to implement access control with some flexibility.

- A. Tag all S3 buckets and objects to indicate the presence of PHI. Create IAM Policies and S3 bucket policies using conditions based on the tags.
- B. Create an S3 full-access IAM policy and associate with users requiring PHI access. Create a more restrictive IAM policy for the non-PHI users.
- C. Tag all IAM users based on PHI access. Test for those tags using IAM Policy and S3 bucket policy conditions.
- D. Write an application to interface with S3 and implement access using custom code. Create IAM Policies and S3 bucket policies to allow access only through the application.

Question Breakdown - Answers

This solution will meet the functional requirements by splitting the two groups into different layers of access. Any strategy that involves “full” access to any service will struggle to meet a least privilege requirement.

- A. Tag all S3 buckets and objects to indicate the presence of PHI. Create IAM Policies and S3 bucket policies using conditions based on the tags.
- B. Create an S3 full-access IAM policy and associate with users requiring PHI access. Create a more restrictive IAM policy for the non-PHI users.
- C. Tag all IAM users based on PHI access. Test for those tags using IAM Policy and S3 bucket policy conditions.
- D. Write an application to interface with S3 and implement access using custom code. Create IAM Policies and S3 bucket policies to allow access only through the application.

Question Breakdown - Answers

As mentioned in answer A, tagging is a building block that enables least privilege with flexibility, and this covers the other side of that solution by providing a mechanism for testing users for access rights.

- A. Tag all S3 buckets and objects to indicate the presence of PHI. Create IAM Policies and S3 bucket policies using conditions based on the tags.
- B. Create an S3 full-access IAM policy and associate with users requiring PHI access. Create a more restrictive IAM policy for the non-PHI users.
- C. Tag all IAM users according on PHI access. Test for those tags using IAM Policy and S3 bucket policy conditions.
- D. Write an application to interface with S3 and implement access using custom code. Create IAM Policies and S3 bucket policies to allow access only through the application.

Question Breakdown - Answers

This solution meets the functional requirements and can possibly implement least privilege, but will be more operational overhead to maintain. It can possibly become an unnecessary single point of failure or bottleneck.

- A. Tag all S3 buckets and objects to indicate the presence of PHI. Create IAM Policies and S3 bucket policies using conditions based on the tags.
- B. Create an S3 full-access IAM policy and associate with users requiring PHI access. Create a more restrictive IAM policy for the non-PHI users.
- C. Tag all IAM users based on PHI access. Test for those tags using IAM Policy and S3 bucket policy conditions.
- D. Write an application to interface with S3 and implement access using custom code. Create IAM Policies and S3 bucket policies to allow access only through the application.

Question Breakdown - Correct Answer

Correct Answers: A and C

- A. Tag all S3 buckets and objects to indicate the presence of PHI. Create IAM Policies and S3 bucket policies using conditions based on the tags.
- B. Create an S3 full-access IAM policy and associate with users requiring PHI access. Create a more restrictive IAM policy for the non-PHI users.
- C. Tag all IAM users based on PHI access. Test for those tags using IAM Policy and S3 bucket policy conditions.
- D. Write an application to interface with S3 and implement access using custom code. Create IAM Policies and S3 bucket policies to allow access only through the application.



Design Cost-Optimized Architectures

18%



Design Cost-Optimized
Architectures

Cost-Optimized Storage

EBS Standard



Cheapest storage pricing

Also charged for consumed IOPS

Lower limit on size than other volume types

Lower IOPS capacity than other volume types

IOPS not dependent on volume size

EBS SC1



More expensive than Standard but less than ST1

Appropriate for cold storage data sets

Easy to upsize as data increases

Throughput dependent on volume size

EBS ST1



More expensive than SC1 but less than GP2

Appropriate for high throughput data sets

Easy to upsize as data increases

Throughput dependent on volume size

EBS GP2



More expensive than ST1 but less than PIOPS

Appropriate for medium-high IOPS bound workloads

Easy to upsize as data increases

IOPS dependent on volume size

EBS PIOPS



Most expensive EBS storage

Also charged for provisioned IOPS

Appropriate for high IOPS and throughput bound workloads

Easy to upsize as data increases

IOPS dependent on volume size

EFS



More expensive than any EBS storage

Charged for data used

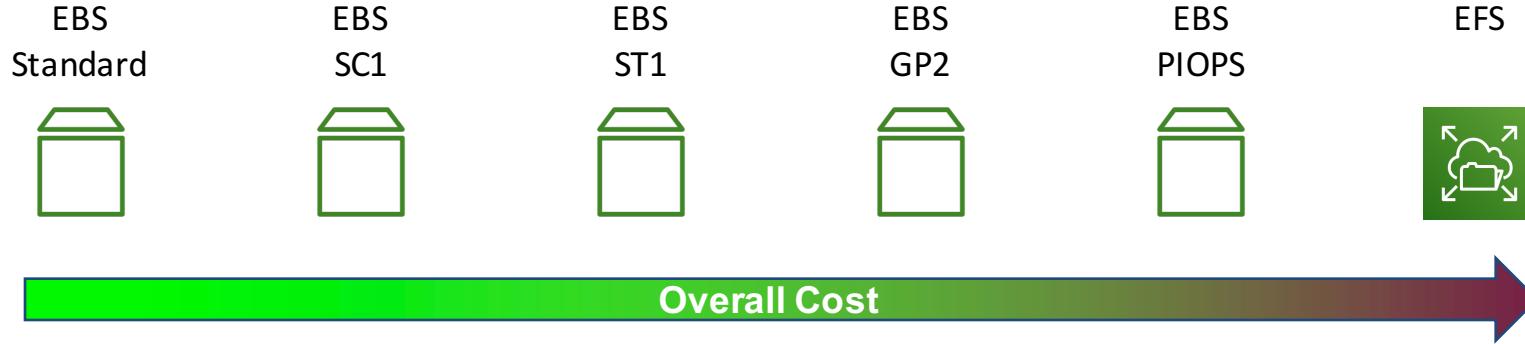
Appropriate for large data sets of larger file sizes

Filesystem is elastic, no need to provision size

IOPS/throughput dependent on amount of data

Storage Cost - Block and File

Storage prices drop over time. It is better to understand current
RELATIVE pricing!



Cost-Optimized Storage

EASY Question Breakdown

Question Breakdown

Which EBS volume type is charged by both the volume size and the number of consumed IOPS?

- A. Standard
- B. GP2
- C. PIOPS
- D. SC1
- E. ST1

Question Breakdown - Correct Answer

Correct Answer: A

- A. Standard
- B. GP2
- C. PIOPS
- D. SC1
- E. ST1

S3 Standard



Highest storage cost

Lowest cost for data access

Highest availability of S3 storage classes

Appropriate for static website objects

S3 Infrequent-Access (S3-IA)



Storage cost less than S3 Standard

Access cost more than S3 Standard

Lower availability than Standard but more than Z-IA

Appropriate for backups requiring low latency access

S3 Intelligent-Tiering



Storage cost according to current storage class

Monitoring and automation charges

Availability according to current storage class

Appropriate for objects with changing access patterns

S3 Onezone Infrequent-Access (Z-IA)



Storage cost less than S3-IA

Access cost identical to S3-IA

Lowest availability of S3 storage classes

Appropriate for backups with lower availability needs

Glacier



Storage cost less than Z-IA

Access cost significantly higher than Z-IA

Designed for 4 9s availability

Appropriate for archives with min-hours latency needs

Glacier Deep Archive



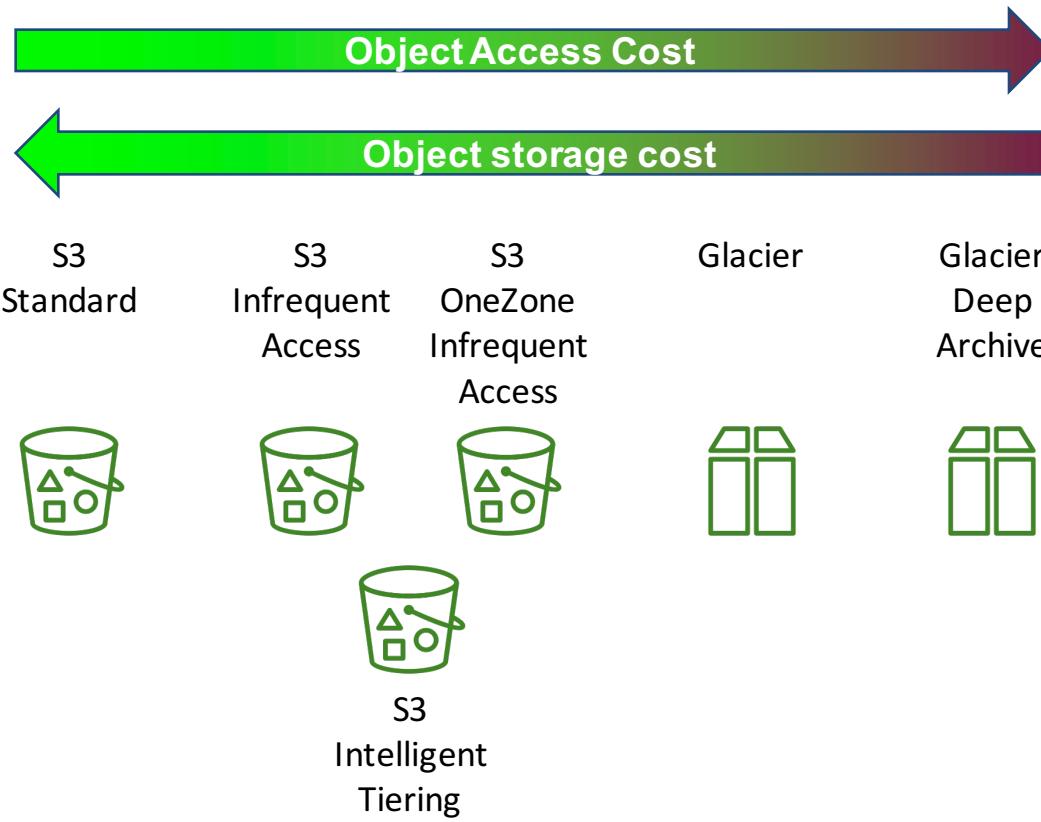
Lowest storage cost

Also charged for consumed
IOPS

Designed for 4 9s availability

Appropriate for archives with
up to 12 hours latency needs

Storage Cost - Object



Cost-Optimized Storage

Question Breakdown

Question Breakdown

After an audit of your company's AWS bill, there is an initiative to reduce costs, and you've been asked to focus on S3 usage. There are tens of millions of large objects spread across many buckets. The usage patterns are varied by bucket and prefix, and are not always predictable. Which of the following cost optimization strategies would be the most appropriate?

- A. Provision CloudFront distributions using the S3 buckets as origins to reduce the cost of accessing the objects by caching.
- B. Manually migrate all objects to S3 Infrequent Access to reduce storage costs.
- C. Create lifecycle policies on the S3 buckets that migrate objects to cheaper storage classes as they age, regardless of usage patterns.
- D. Migrate objects to the S3 Intelligent-Tiering storage class to automate the optimization of storage costs based on access frequency.

Question Breakdown - Key Terms

After an audit of your company's AWS bill, there is an initiative to **reduce costs**, and you've been asked to focus on **S3 usage**. There are **tens of millions of large objects spread across many buckets**. The **usage patterns are varied by bucket and prefix**, and are **not always predictable**. Which of the following cost optimization strategies would be the most appropriate?

- A. Provision CloudFront distributions using the S3 buckets as origins to reduce the cost of accessing the objects by caching.
- B. Manually migrate all objects to S3 Infrequent Access to reduce storage costs.
- C. Create lifecycle policies on the S3 buckets that migrate objects to cheaper storage classes as they age, regardless of usage patterns.
- D. Migrate objects to the S3 Intelligent-Tiering storage class to automate the optimization of storage costs based on access frequency.

Question Breakdown - Answers

CloudFront can be used to reduce the cost of data transfer between regions or from S3 to the Internet, but won't impact actual S3 storage costs.

- A. Provision CloudFront distributions using the S3 buckets as origins to reduce the cost of accessing the objects by caching.
- B. Manually migrate all objects to S3 Infrequent Access to reduce storage costs.
- C. Create lifecycle policies on the S3 buckets that migrate objects to cheaper storage classes as they age, regardless of usage patterns.
- D. Migrate objects to the S3 Intelligent-Tiering storage class to automate the optimization of storage costs based on access frequency.

Question Breakdown - Answers

This might reduce the storage cost for objects, but may increase overall cost if a large enough number of objects are accessed on a very frequent basis, as S3-IA has higher access costs than S3 Standard.

- A. Provision CloudFront distributions using the S3 buckets as origins to reduce the cost of accessing the objects by caching.
- B. Manually migrate all objects to S3 Infrequent Access to reduce storage costs.
- C. Create lifecycle policies on the S3 buckets that migrate objects to cheaper storage classes as they age, regardless of usage patterns.
- D. Migrate objects to the S3 Intelligent-Tiering storage class to automate the optimization of storage costs based on access frequency.

Question Breakdown - Answers

This solution is similar to B in migration of objects to less expensive storage, but if this is done regardless of frequency of access, the overall S3 charges could actually increase.

- A. Provision CloudFront distributions using the S3 buckets as origins to reduce the cost of accessing the objects by caching.
- B. Manually migrate all objects to S3 Infrequent Access to reduce storage costs.
- C. Create lifecycle policies on the S3 buckets that migrate objects to cheaper storage classes as they age, regardless of usage patterns.
- D. Migrate objects to the S3 Intelligent-Tiering storage class to automate the optimization of storage costs based on access frequency.

Question Breakdown - Answers

The S3 Intelligent-Tiering storage class is designed to monitor object access frequency, and migrate objects between S3 Standard and S3-Infrequent Access as soon as it is determined that the move will reduce overall cost.

- A. Provision CloudFront distributions using the S3 buckets as origins to reduce the cost of accessing the objects by caching.
- B. Manually migrate all objects to S3 Infrequent Access to reduce storage costs.
- C. Create lifecycle policies on the S3 buckets that migrate objects to cheaper storage classes as they age, regardless of usage patterns.
- D. Migrate objects to the S3 Intelligent-Tiering storage class to automate the optimization of storage costs based on access frequency.

Question Breakdown - Correct Answer

Correct Answer: D

- A. Provision CloudFront distributions using the S3 buckets as origins to reduce the cost of accessing the objects by caching.
- B. Manually migrate all objects to S3 Infrequent Access to reduce storage costs.
- C. Create lifecycle policies on the S3 buckets that migrate objects to cheaper storage classes as they age, regardless of usage patterns.
- D. Migrate objects to the S3 Intelligent-Tiering storage class to automate the optimization of storage costs based on access frequency.



Design Cost-Optimized
Architectures

Cost-Optimized Compute
Resources

Compute Cost - EC2 Pricing

Spot Instances

No guaranteed pricing
Pay for unused capacity
Volatile
Specify maximum bid
+Specific duration
+Multiple instance types
+Multiple AZ

Reserved Instances

Guaranteed pricing for up to 3 years
+Capacity guarantee
Variable up-front for more discount

On Demand Instances

Pay as you go
No discount
No capacity guarantee

Dedicated Instances

Dedicated hardware
Can share with non-dedicated VMs
Per-region fee
+Spot
+Reservations
+On Demand

Dedicated Hosts

Dedicated hardware
Single instance type
Pay for host capacity, not instance
+Reservations
+On Demand

Overall Cost



Cost-Optimized Compute Resources

EASY Question Breakdown

Question Breakdown

Which EC2 pricing model would be the most appropriate for a massively-parallel, short term job?

- A. Dedicated Instances
- B. On-Demand
- C. Spot
- D. Reserved

Question Breakdown - Correct Answer

Correct Answer: C

- A. Dedicated Instances
- B. On-Demand
- C. Spot
- D. Reserved

Compute Cost - EC2 Strategy

Switching from On Demand shared tenancy to Dedicated Instances

Increased cost?

Decreased cost?

Compute Cost - EC2 Strategy

Switching from On Demand to a mix of Reserved and Spot

Increased cost?

Decreased cost?

Compute Cost - EC2 Strategy

Switching to
Instances from
Dedicated H

It depends! Were
the Dedicated Hosts
fully utilized?

reased cost?

reased cost?

Compute Cost - Overhead

What managed services and features reduce operational overhead (and thus TCO)?

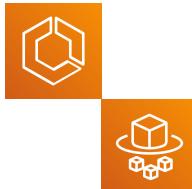
Auto Scaling for minimizing unused resources



Elastic Beanstalk for automating app env deployments



ECS on Fargate for eliminating infrastructure overhead



Lambda for eliminating infrastructure and platform overhead



ECS on EC2



Priced according to EC2 instance resources

No charge for ECS management layer

More resilient than DiY container management

Performance dependent on EC2 resources

ECS on Fargate



On-demand and Spot pricing available

Cheaper than EC2 if tasks are temporary

Charged for vCPU and memory separately

Pricier than EC2 for persistent tasks

More resilient than EC2

Auto scaled for concurrency

Performance depends on task resources

Lambda



Pay for provisioned
CPU and memory

Cheaper than EC2 if
tasks are temporary

Pay for execution time

Pricier than EC2 for
long-running code

More resilient than
EC2

Auto scaled for
concurrency

Performance depends
on task resources

Pre-warmed functions
for lower latency

Cost-Optimized Compute Resources

Question Breakdown

Question Breakdown

A new application is being deployed onto EC2 instances, with a requirement for horizontal scaling. The EC2 instance type doesn't need to be static, as long as the instances meet minimum CPU and memory requirements. What would be the most cost-effective deployment strategy for the application as well as lowest operational overhead?

- A. Deploy one Auto Scaling group using a single launch template with multiple instance types defined. Specify an appropriate percentage of On Demand instances to maintain resilience.
- B. Deploy two Auto Scaling groups for On Demand and Spot pricing. Specify baseline maximum instances for On Demand and everything else will be Spot instances, with multiple instance types defined.
- C. Deploy one steady-state Auto Scaling group with reserved instances for baseline traffic. Deploy a second Auto Scaling group with On Demand instances for variable traffic.
- D. Deploy one Auto Scaling group using only Spot instances in two AZ to minimize chances of spot price spikes having a cost impact.

Question Breakdown - Key Terms

A new application is being deployed onto EC2 instances, with a requirement for horizontal scaling. The EC2 instance type doesn't need to be static, as long as the instances meet minimum CPU and memory requirements. What would be the most cost-effective deployment strategy for the application as well as lowest operational overhead?

- A. Deploy one Auto Scaling group using a single launch template with multiple instance types defined. Specify an appropriate percentage of On Demand instances to maintain resilience.
- B. Deploy two Auto Scaling groups for On Demand and Spot pricing. Specify baseline maximum instances for On Demand and everything else will be Spot instances, with multiple instance types defined.
- C. Deploy one steady-state Auto Scaling group with reserved instances for baseline traffic. Deploy a second Auto Scaling group with On Demand instances for variable traffic.
- D. Deploy one Auto Scaling group using only Spot instances in two AZ to minimize chances of spot price spikes having a cost impact.

Question Breakdown - Answers

A single Auto Scaling group means fewer moving parts, and fewer parameters that must be monitored. The ASG configuration includes multiple instance types which can help with cost control.

- A. Deploy one Auto Scaling group using a single launch template with multiple instance types defined. Specify an appropriate percentage of On Demand instances to maintain resilience.
- B. Deploy two Auto Scaling groups for On Demand and Spot pricing. Specify baseline maximum instances for On Demand and everything else will be Spot instances, with multiple instance types defined.
- C. Deploy one steady-state Auto Scaling group with reserved instances for baseline traffic. Deploy a second Auto Scaling group with On Demand instances for variable traffic.
- D. Deploy one Auto Scaling group using only Spot instances in two AZ to minimize chances of spot price spikes having a cost impact.

Question Breakdown - Answers

Splitting the Auto Scaling groups into On Demand and Spot pricing can potentially make cost optimization easier, but will increase operational overhead (by needing to manage both).

- A. Deploy one Auto Scaling group using a single launch template with multiple instance types defined. Specify an appropriate percentage of On Demand instances to maintain resilience.
- B. Deploy two Auto Scaling groups for On Demand and Spot pricing. Specify baseline maximum instances for On Demand and everything else will be Spot instances, with multiple instance types defined.
- C. Deploy one steady-state Auto Scaling group with reserved instances for baseline traffic. Deploy a second Auto Scaling group with On Demand instances for variable traffic.
- D. Deploy one Auto Scaling group using only Spot instances in two AZ to minimize chances of spot price spikes having a cost impact.

Question Breakdown - Answers

Reserved instances must be a single instance type to be valid, which will lock you into that choice whether it remains appropriate or not. The steady-state group should require near-zero operations, which helps, but doesn't outweigh the lack of flexibility.

- A. Deploy one Auto Scaling group using a single launch template with multiple instance types defined. Specify an appropriate percentage of On Demand instances to maintain resilience.
- B. Deploy two Auto Scaling groups for On Demand and Spot pricing. Specify baseline maximum instances for On Demand and everything else will be Spot instances, with multiple instance types defined.
- C. Deploy one steady-state Auto Scaling group with reserved instances for baseline traffic.
Deploy a second Auto Scaling group with On Demand instances for variable traffic.
- D. Deploy one Auto Scaling group using only Spot instances in two AZ to minimize chances of spot price spikes having a cost impact.

Question Breakdown - Answers

This is an interesting choice, and definitely minimizes operational overhead. There is increased risk to availability if AWS experiences an event or lack of capacity, you could lose all your instances and experience your own outage.

- A. Deploy one Auto Scaling group using a single launch template with multiple instance types defined. Specify an appropriate percentage of On Demand instances to maintain resilience.
- B. Deploy two Auto Scaling groups for On Demand and Spot pricing. Specify baseline maximum instances for On Demand and everything else will be Spot instances, with multiple instance types defined.
- C. Deploy one steady-state Auto Scaling group with reserved instances for baseline traffic. Deploy a second Auto Scaling group with On Demand instances for variable traffic.
- D. Deploy one Auto Scaling group using only Spot instances in two AZ to minimize chances of spot price spikes having a cost impact.

Question Breakdown - Correct Answer

Correct Answer: A

- A. Deploy one Auto Scaling group using a single launch template with multiple instance types defined. Specify an appropriate percentage of On Demand instances to maintain resilience.
- B. Deploy two Auto Scaling groups for On Demand and Spot pricing. Specify baseline maximum instances for On Demand and everything else will be Spot instances, with multiple instance types defined.
- C. Deploy one steady-state Auto Scaling group with reserved instances for baseline traffic. Deploy a second Auto Scaling group with On Demand instances for variable traffic.
- D. Deploy one Auto Scaling group using only Spot instances in two AZ to minimize chances of spot price spikes having a cost impact.



Design Cost-Optimized
Architectures

Cost-Optimized Database
Resources

RDS



Pay for provisioned compute resources

Pay for provisioned storage resources

Resilience dependent on single node limits

Performance dependent on single node limits

Aurora



Pay for provisioned OR actual compute resources

Pay for actual storage

Resilience higher than RDS

Serverless capability enables horizontal scaling

RedShift



Pay for provisioned compute resources

Storage is charged according to compute resources

Much higher storage capacity than RDS/Aurora

Performance scales according to number of cluster nodes

DynamoDB



Pay for provisioned OR actual
read/write ops

Pay for actual storage

Higher resilience than RDS,
Aurora, RedShift

Performance only limited by
account quotas

Elasticache



Pay for provisioned compute resources

Storage is charged according to compute resources

Memcached - no SPoF
Redis - depends on 1 node

Performance depends on number of nodes



Design Cost-Optimized
Architectures

Cost-Optimized Networks

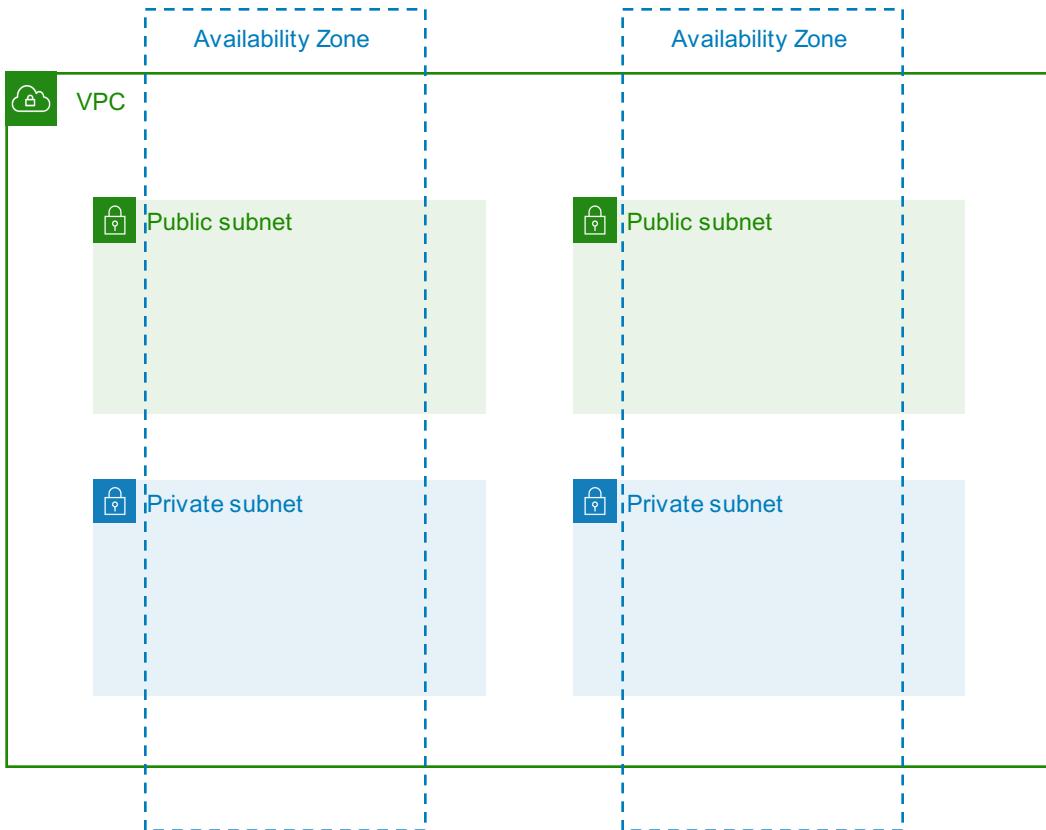
Zero-cost VPC Network Resources



VPC

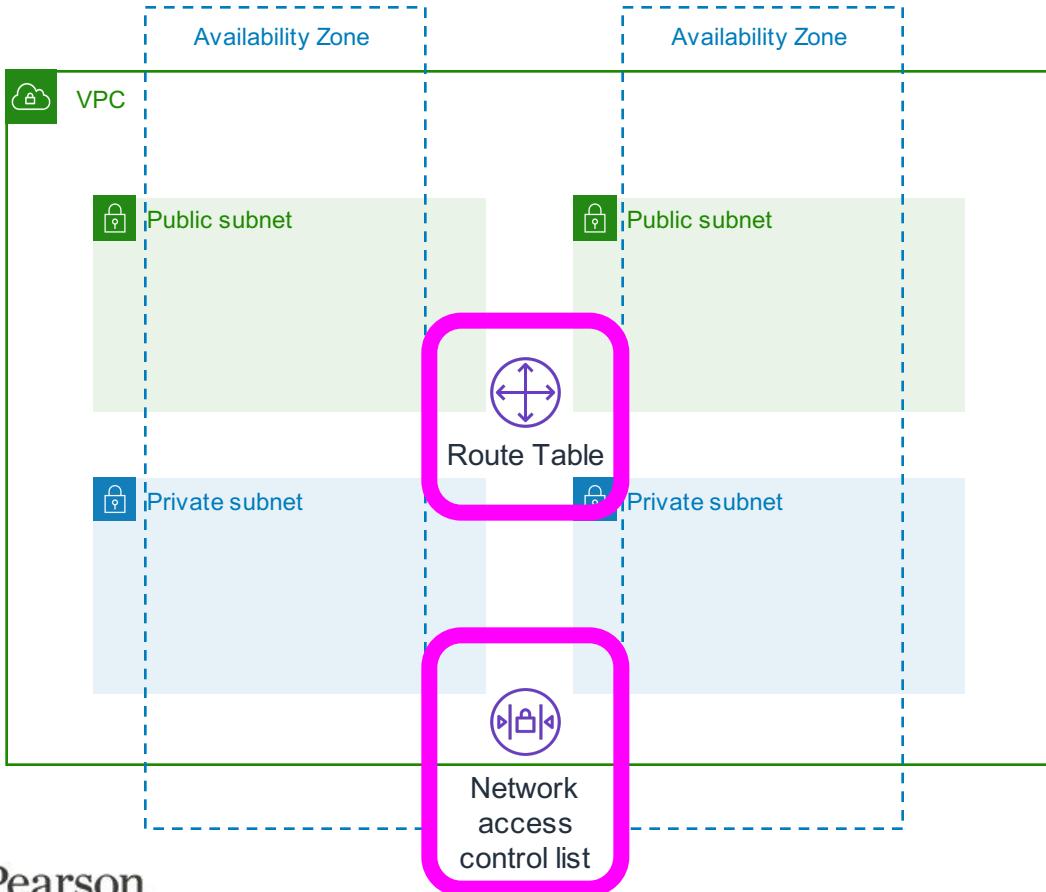
The VPC network may be
free, but it is useless
without other features!

Zero-cost VPC Network Resources



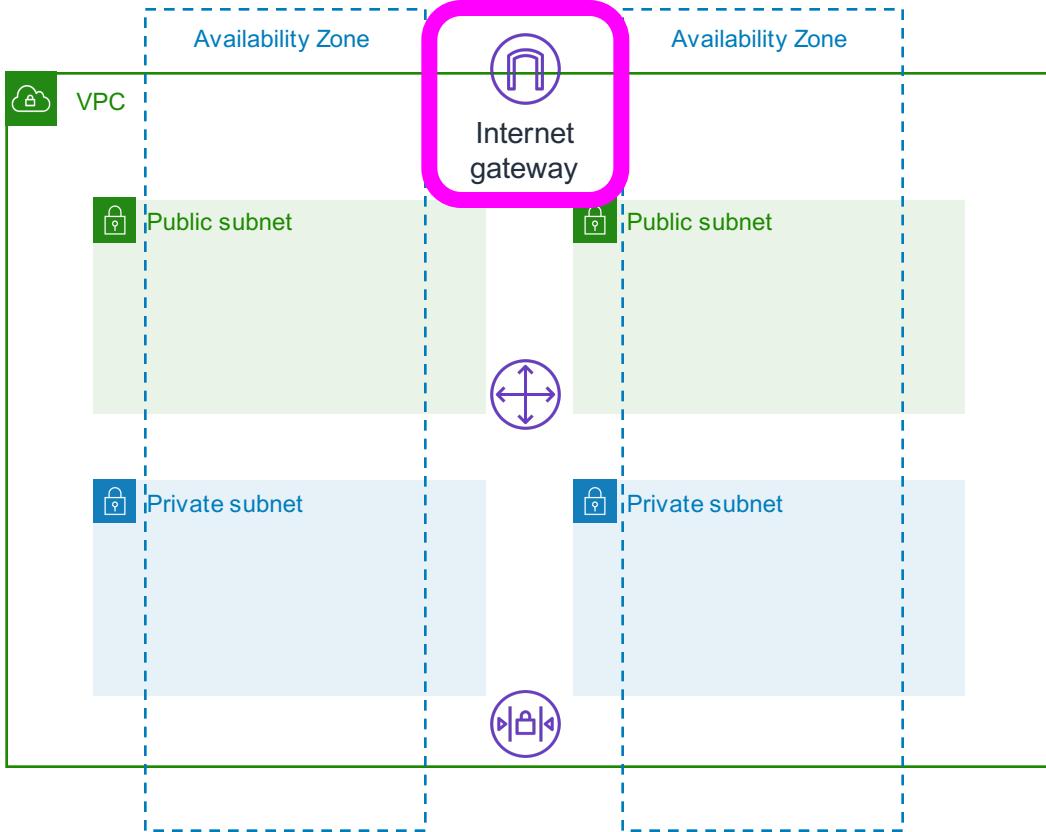
Subnets are free,
regardless of how many
AZ are used in the region

Zero-cost VPC Network Resources



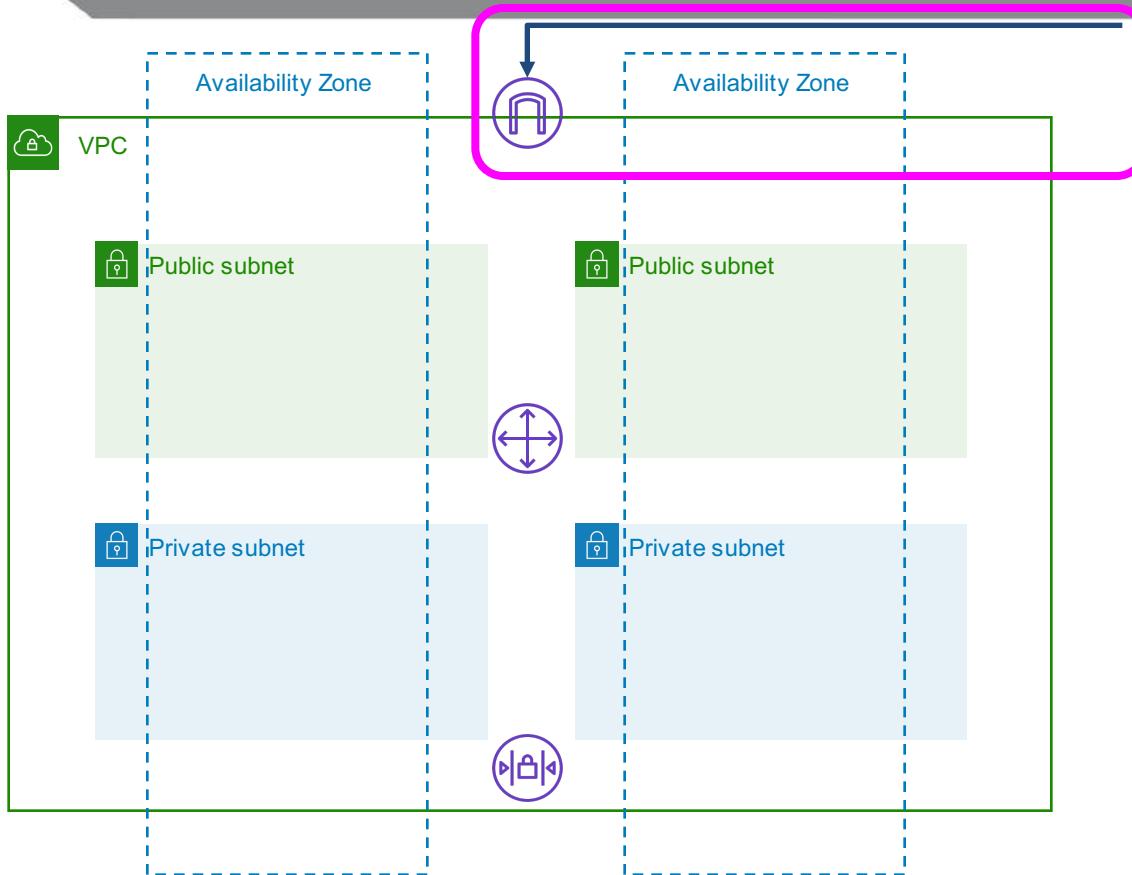
Route tables and NACLs
are free, and only limited
by account quotas

Zero-cost VPC Network Resources



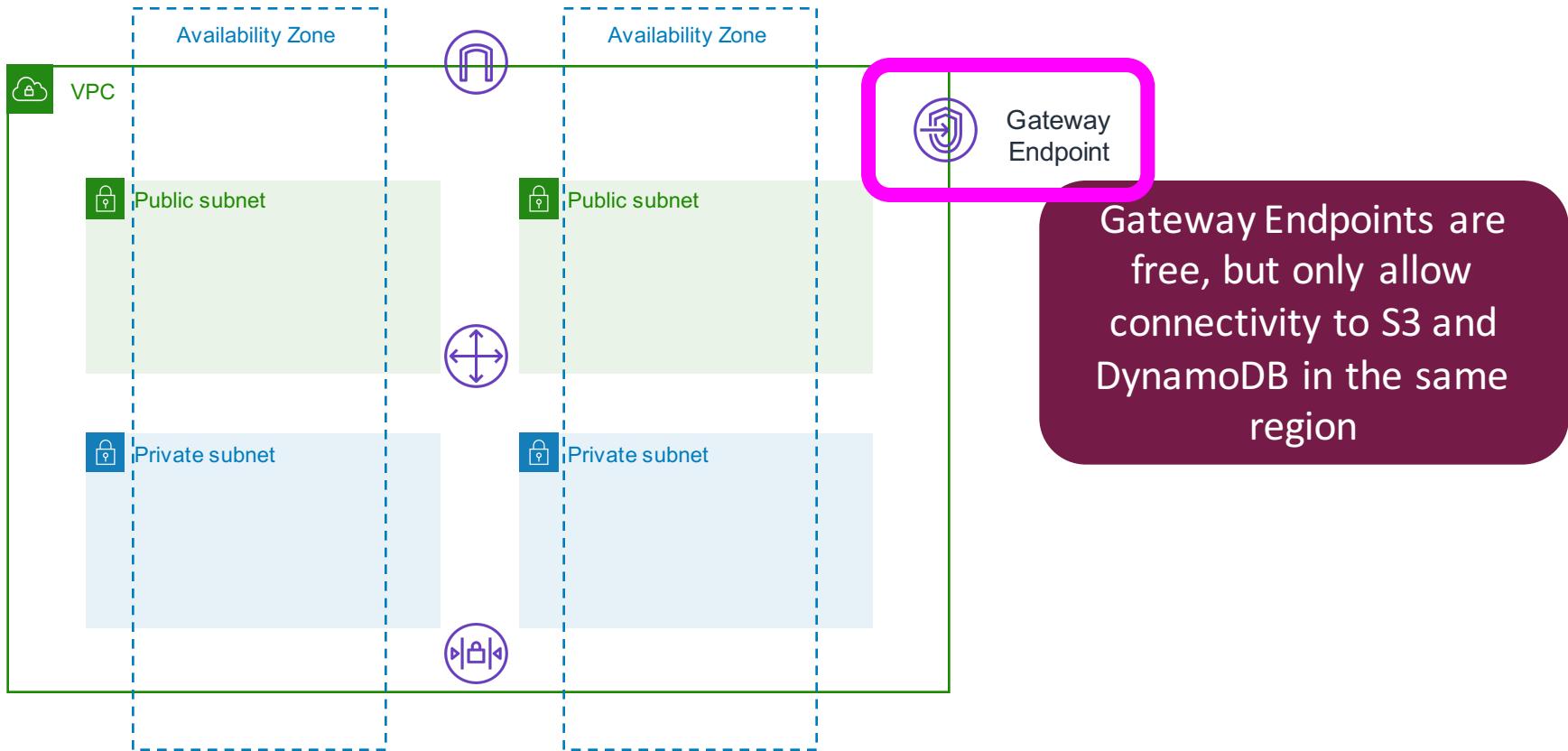
Internet Gateway
resources are free,
regardless of traffic
throughput

Zero-cost VPC Network Resources

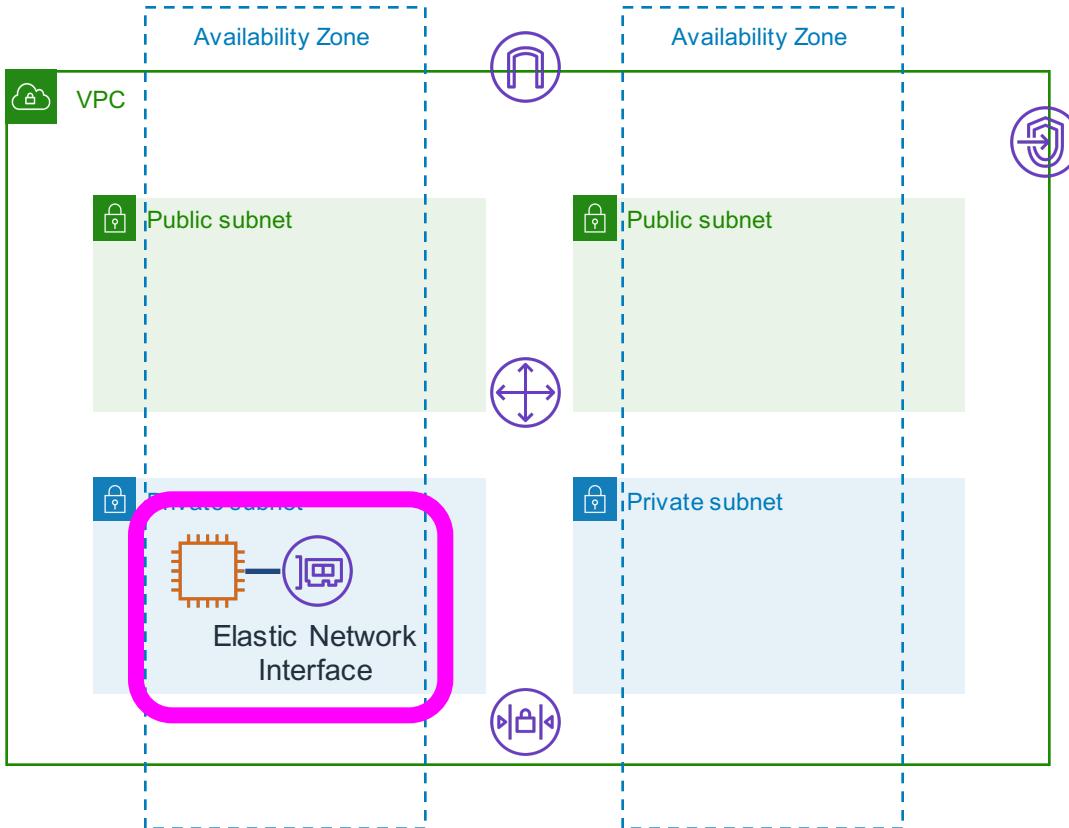


Inbound traffic from the Internet is free, regardless of source

Zero-cost VPC Network Resources

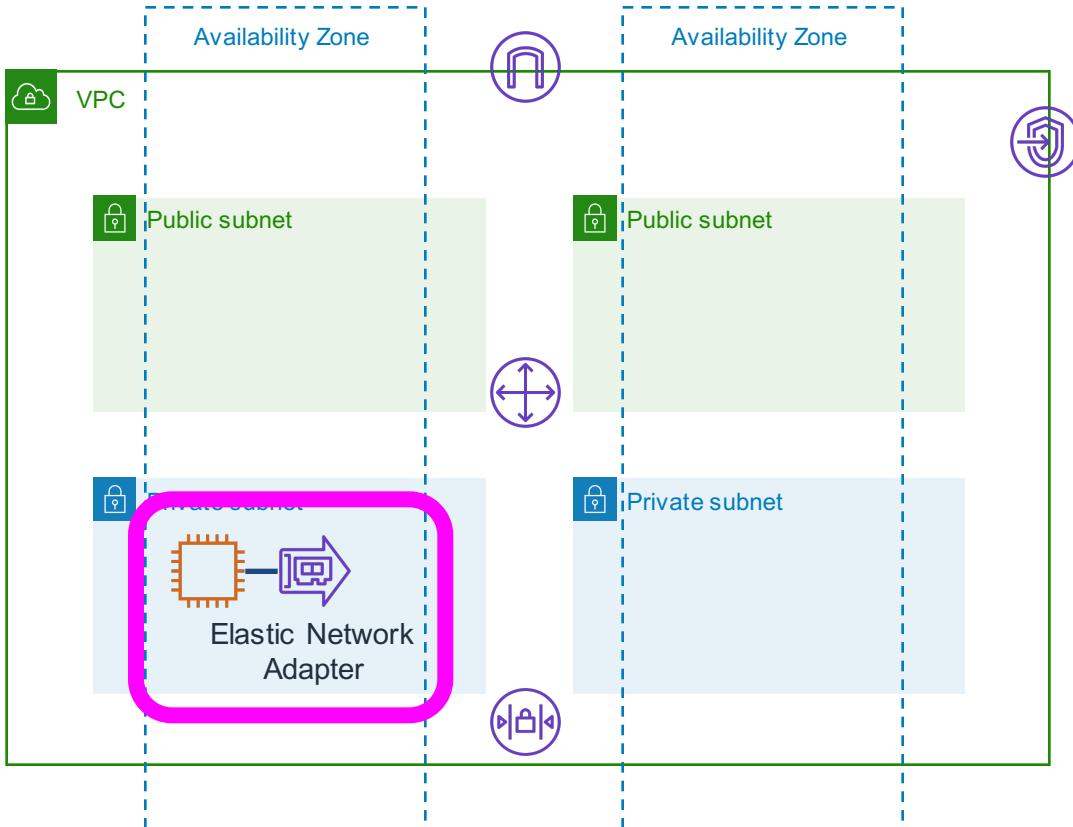


Zero-cost VPC Compute Resources



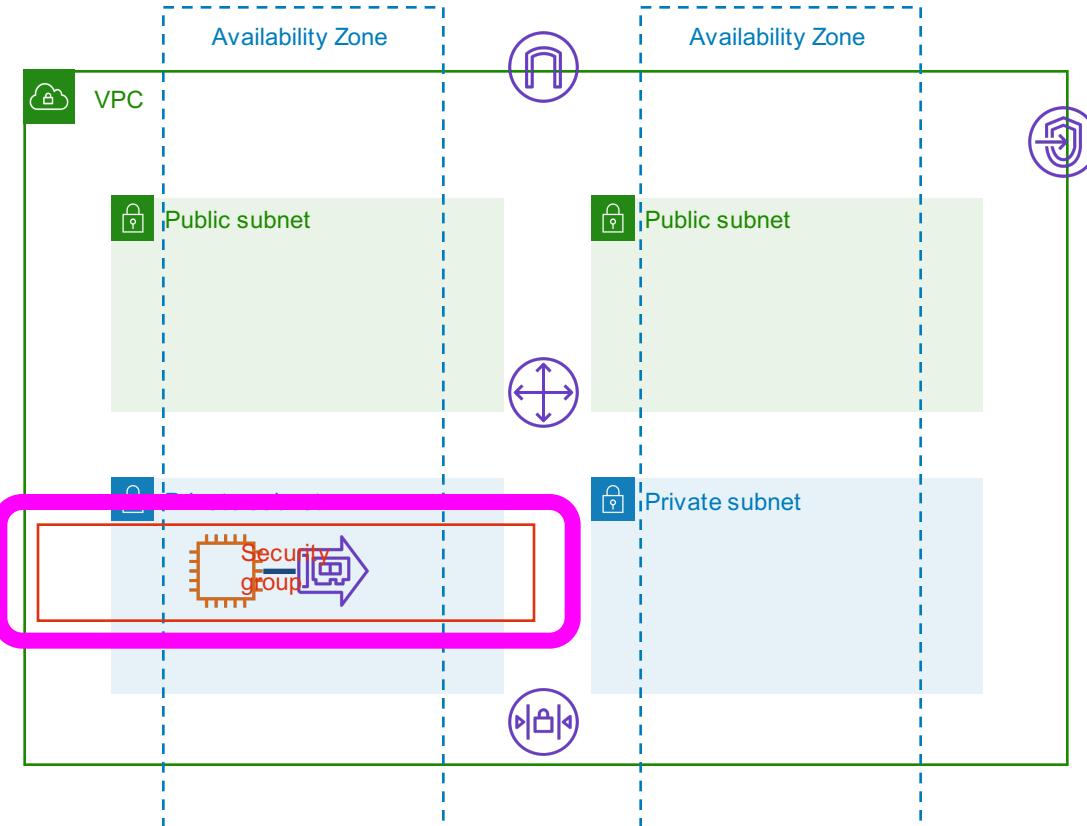
The ENI resource is free, but you may be charged for traffic depending on destination

Zero-cost VPC Compute Resources



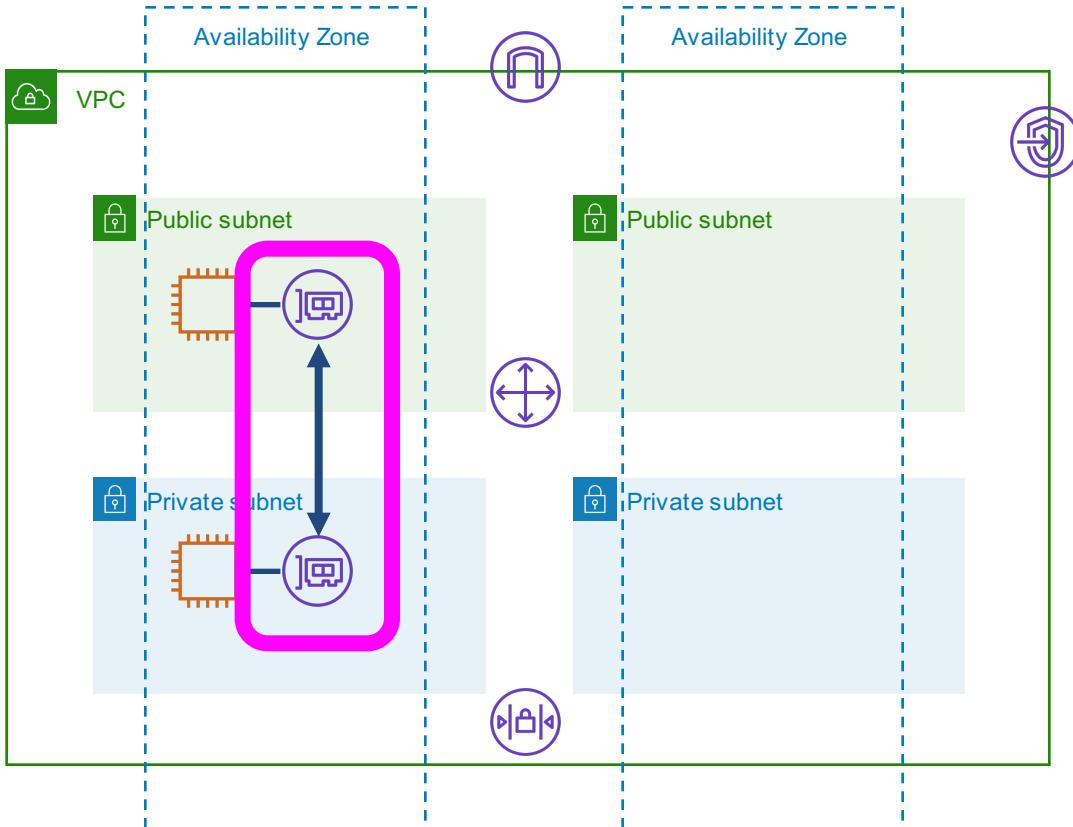
The ENA and Elastic Fabric Adapters are similar to ENI - free but possible charges for network activity

Zero-cost VPC Compute Resources



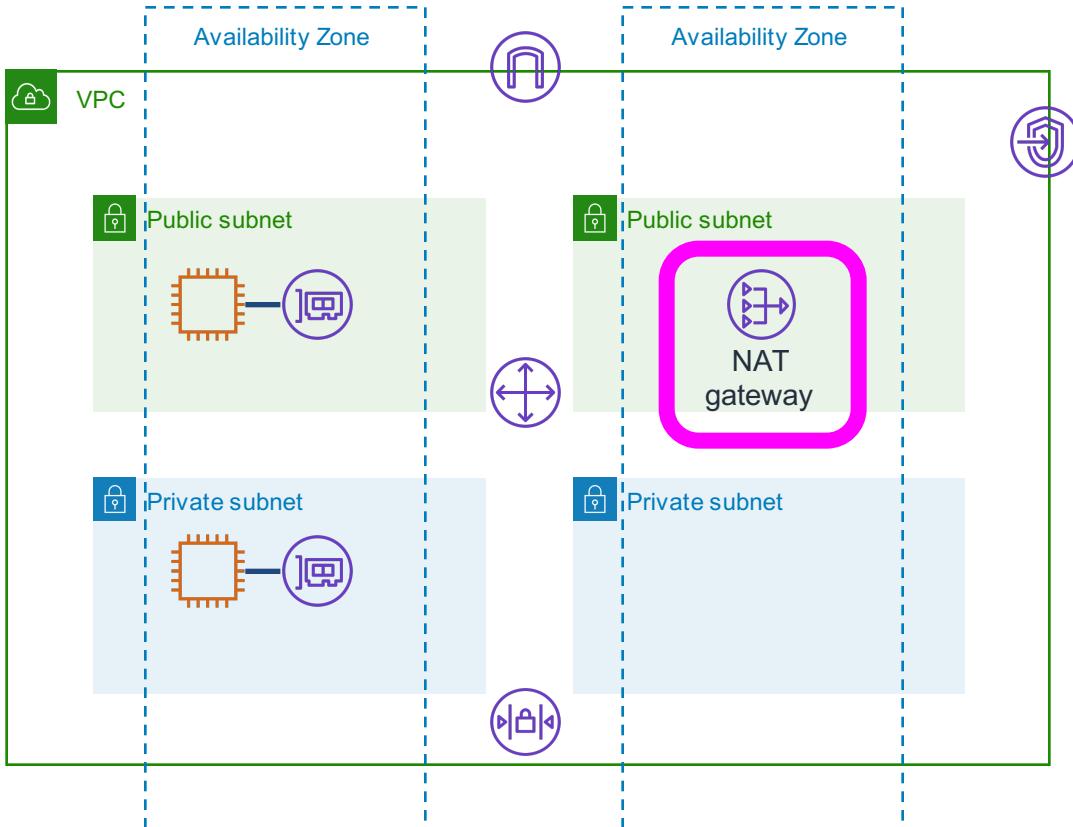
Security groups and rules
are free, and only limited
by account quotas

Zero-cost VPC Compute Resources



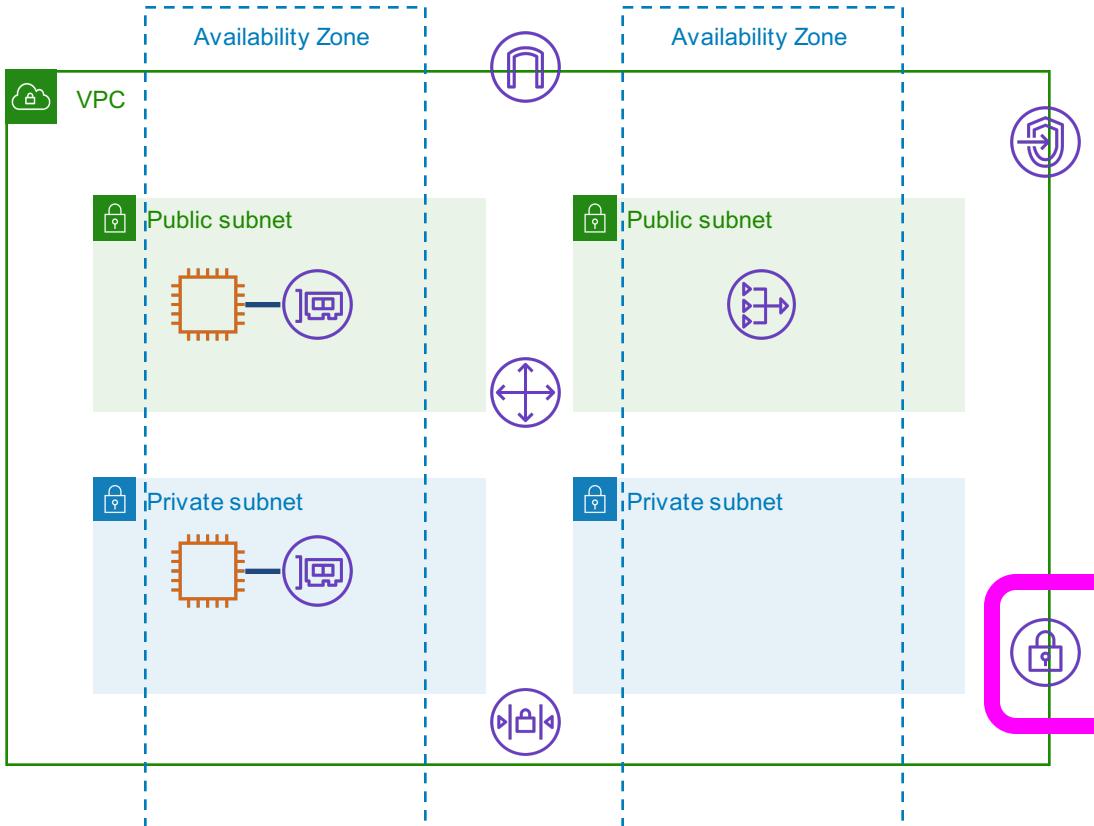
Same-AZ network traffic is free EXCEPT if a public IP is the destination

Charged VPC Network Resources



NAT Gateways are charged by the hour and based on throughput

Charged VPC Network Resources



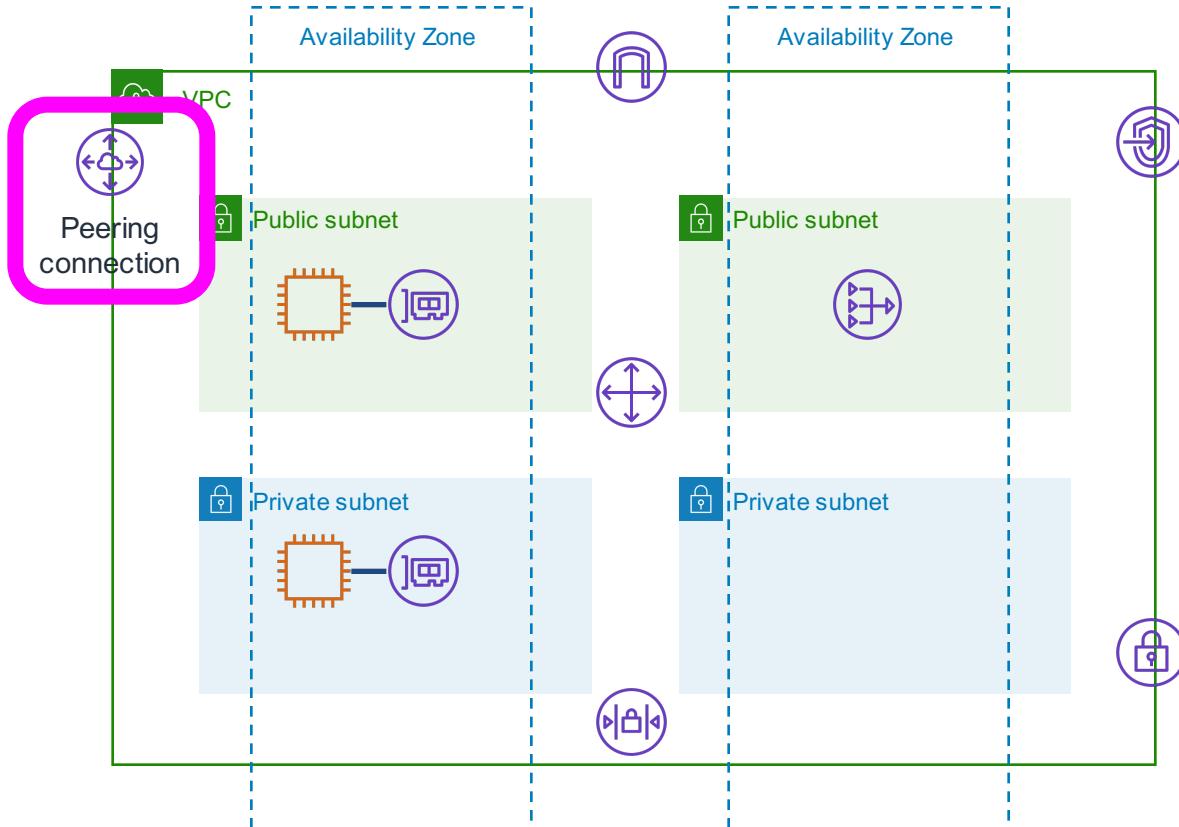
VPGs are charged by the hour and for VPN throughput



VPN gateway

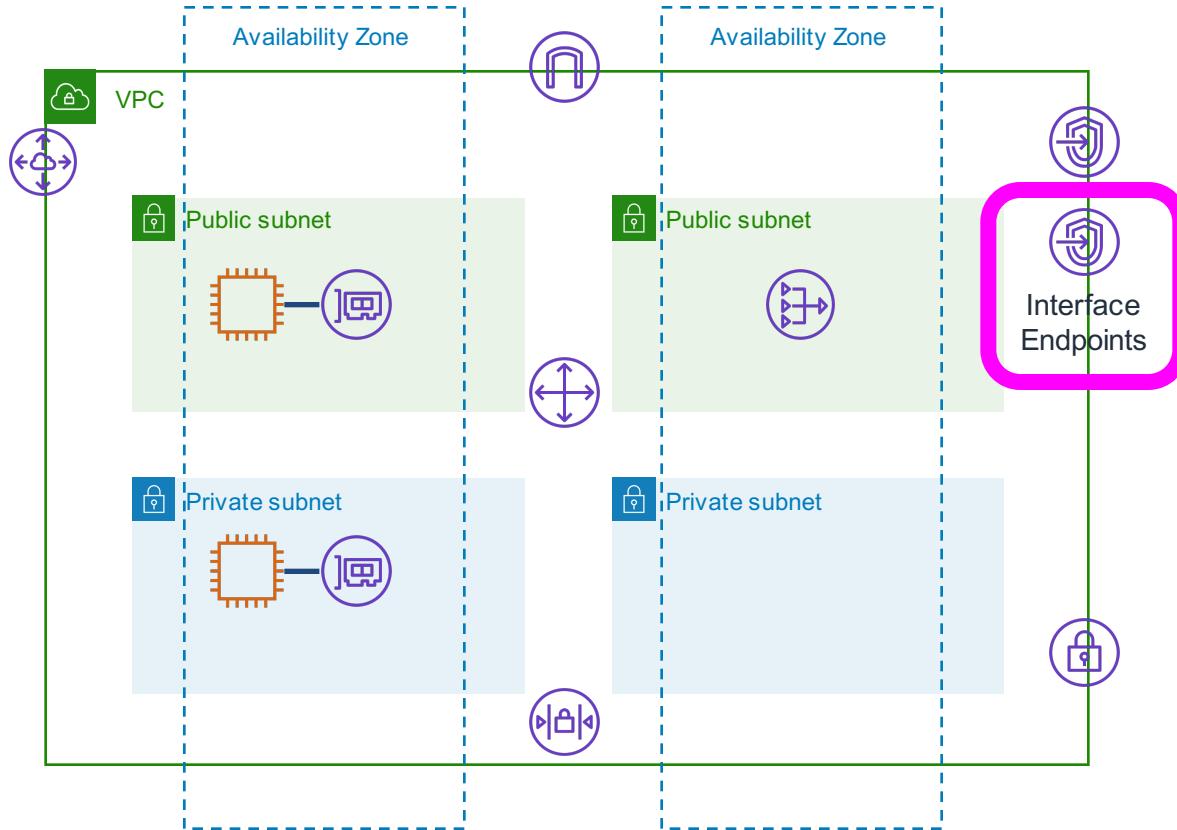


Charged VPC Network Resources



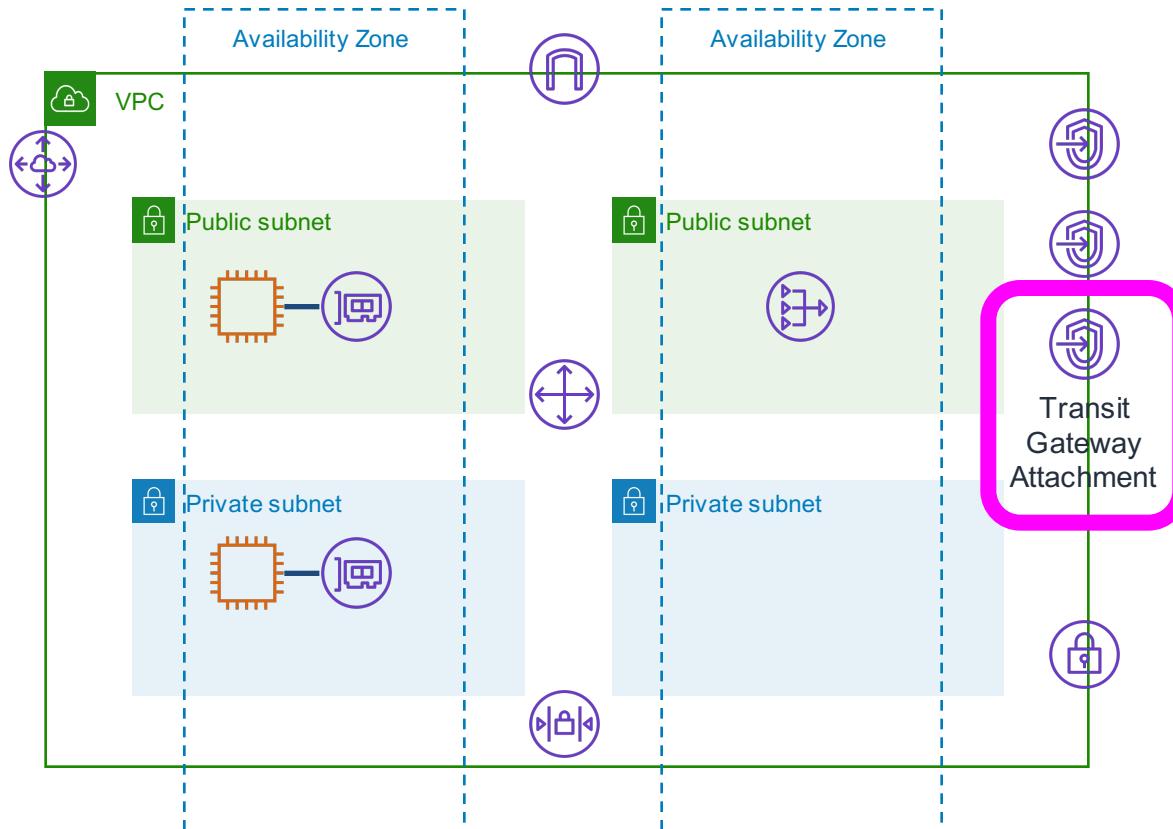
VPC Peering connections are charged by the hour and for traffic throughput, even to the same AZ

Charged VPC Network Resources



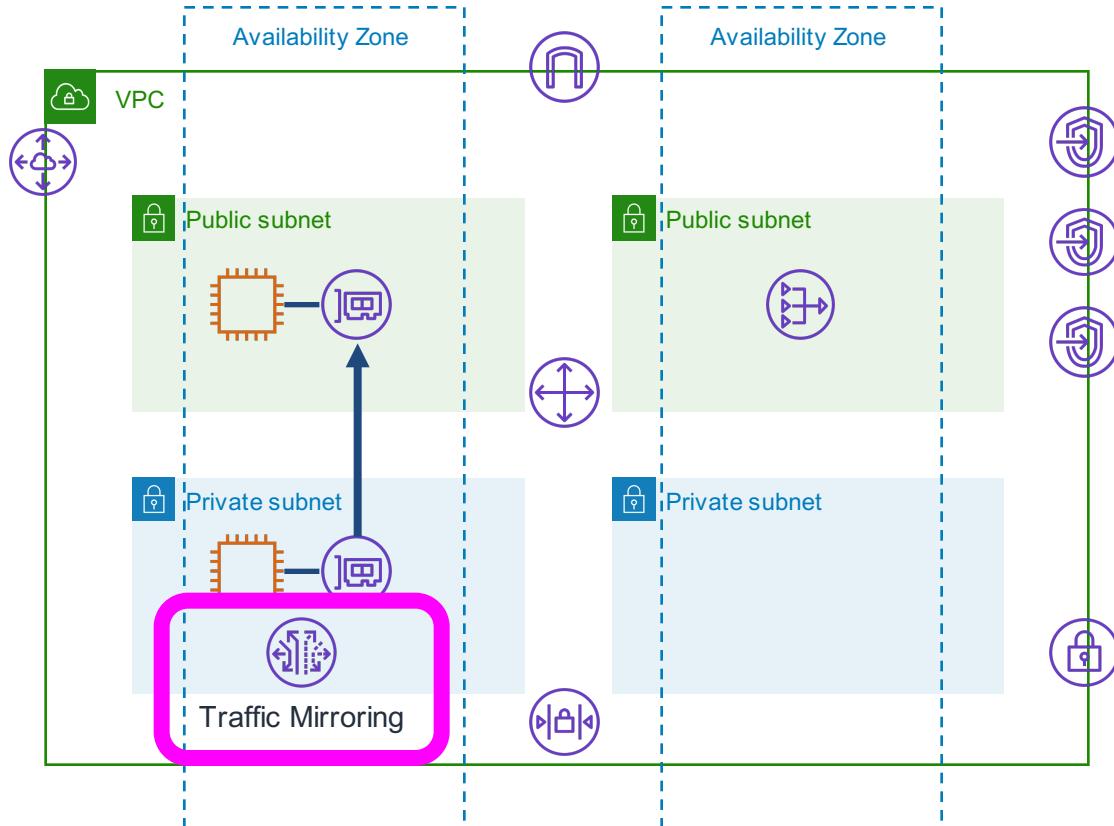
Interface Endpoints and
Privatelink are charged by
the hour and for traffic
throughput

Charged VPC Network Resources



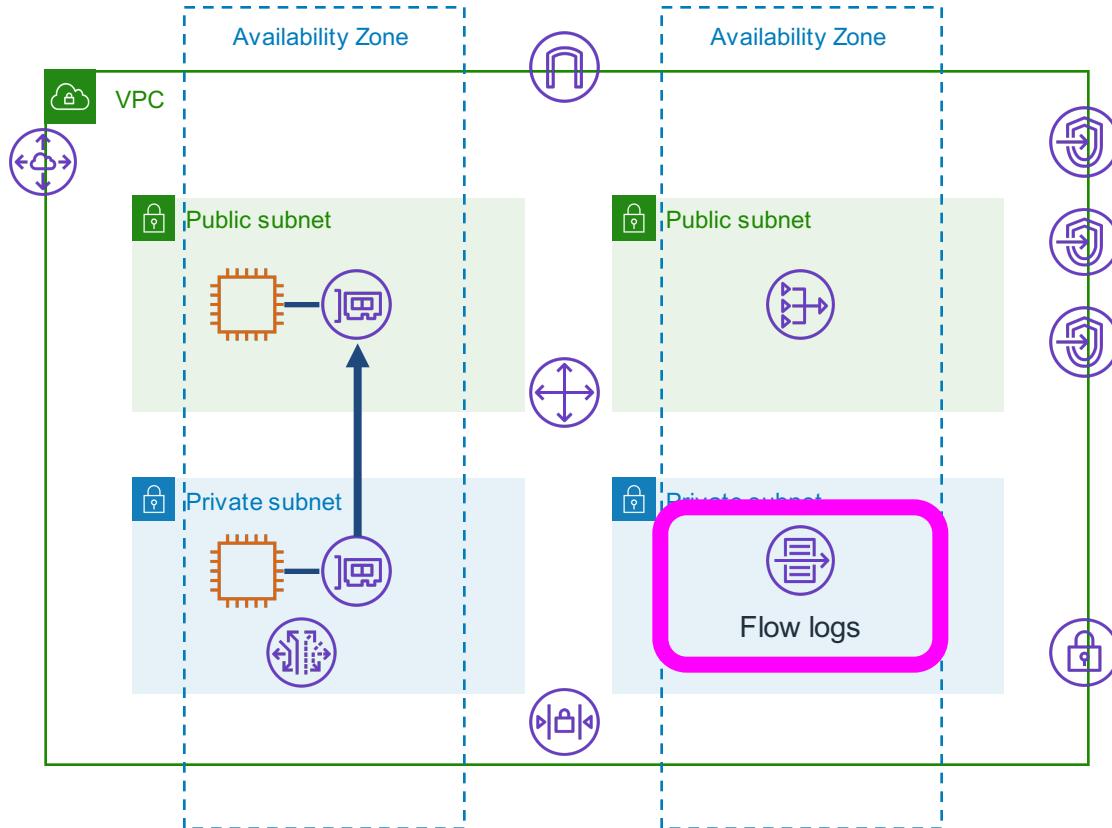
Transit Gateway Attachments are charged hourly and for traffic throughput, and can be more expensive than other options

Charged VPC Network Resources



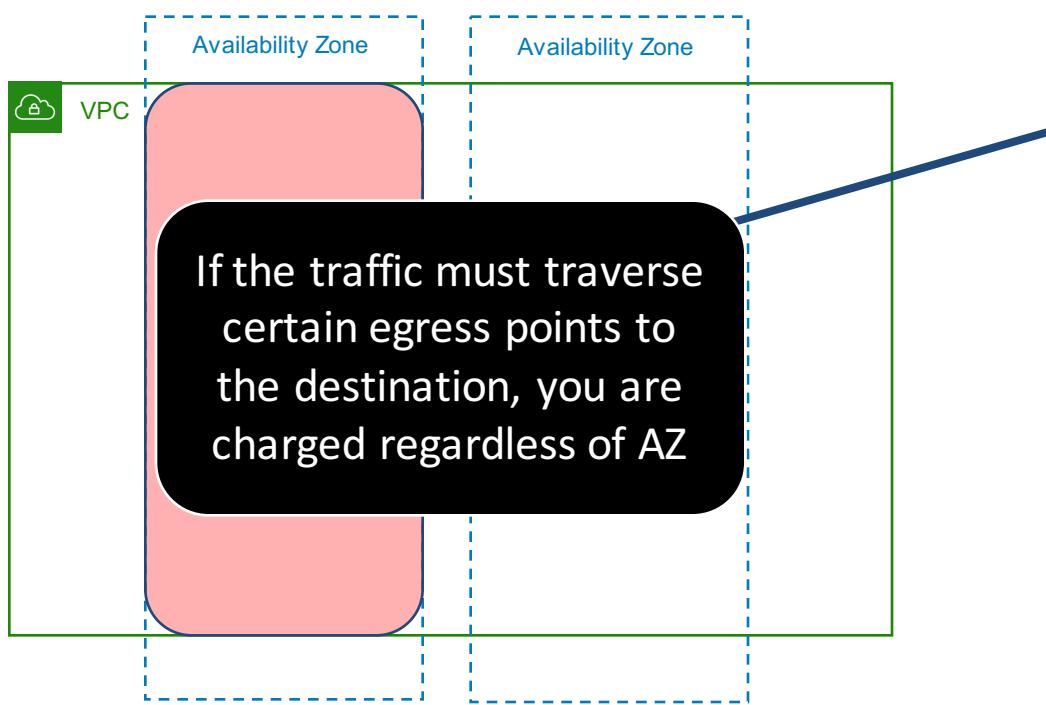
Traffic Mirroring is charged hourly per ENI that has mirroring enabled

Charged VPC Network Resources



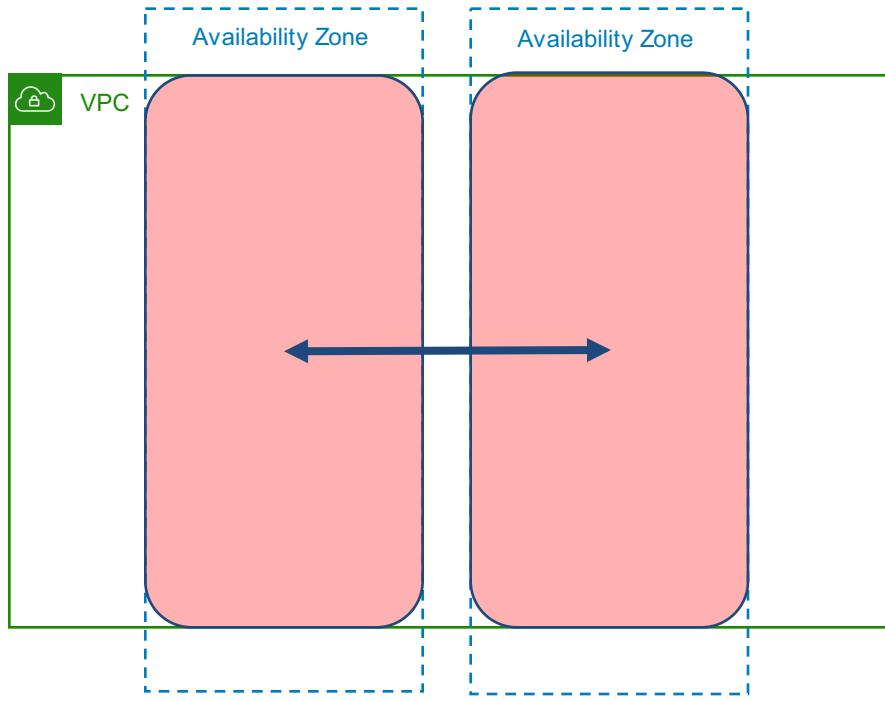
VPC Flow logs are charged according to the amount of traffic processed (and for log storage in the destination service)

Same-Region Traffic Charges



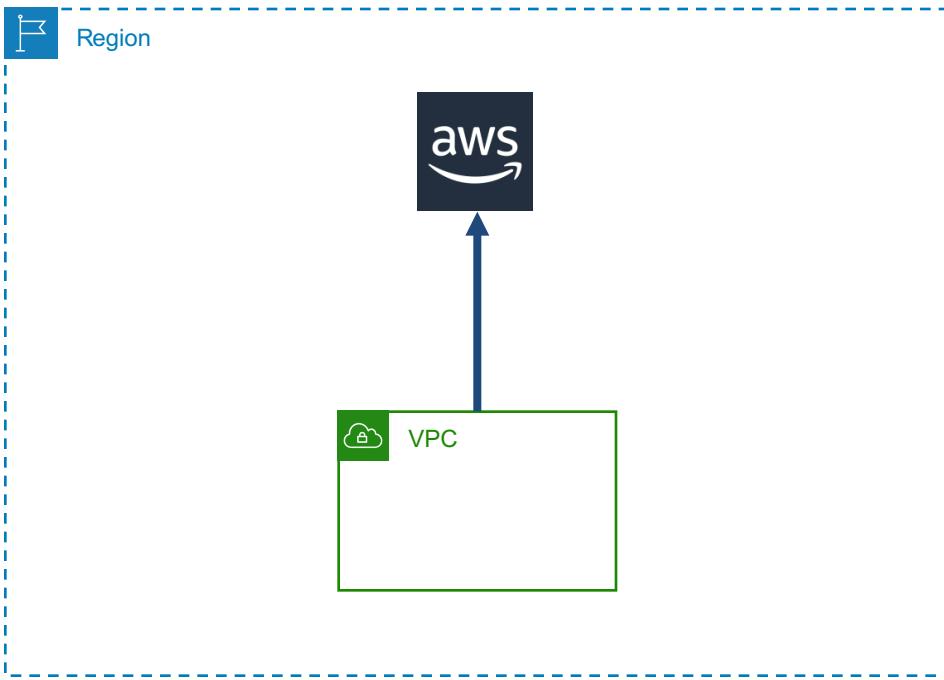
Any traffic with source and destination in here is free unless the destination uses the public IP

Same-Region Traffic Charges



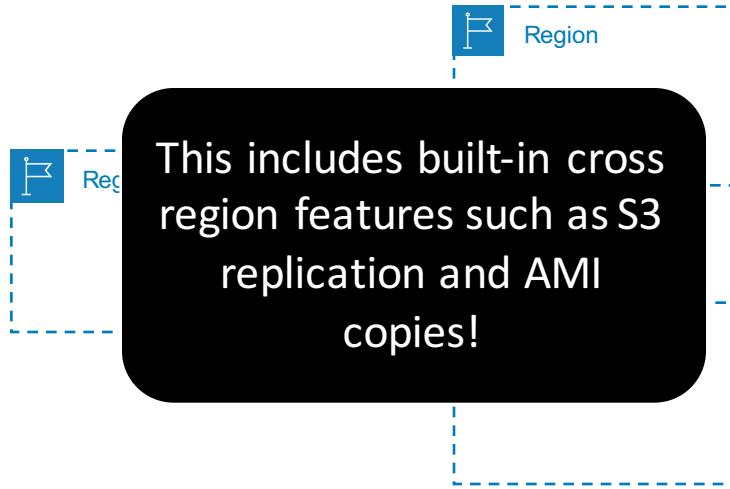
Any traffic with source and destination in different AZ is charged if the resource is AZ scoped

Same-Region Traffic Charges



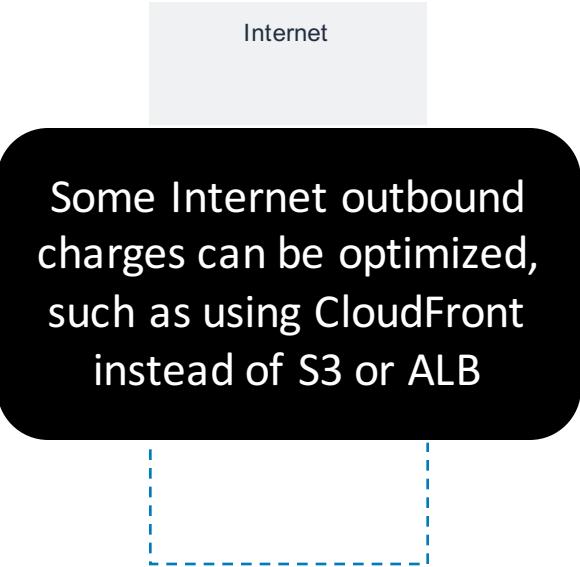
Most same-region traffic from VPC to AWS services will be free, such as S3 bucket access, unless otherwise noted

Cross-Region Traffic Charges



All outbound cross-region traffic is charged, and there can be additional fees based on the gateway used

Cross-Region Traffic Charges



All outbound Internet traffic is charged, and there can be additional fees based on the gateway used

Question Breakdown - Key Terms

Your company has deployed a **high-bandwidth website** that is entirely **static content** and served directly from **S3**. The monthly charges are significant and you've been asked to **reduce cost** if possible. Which of the following strategies would result in lower charges for the site?

- A. Deploy an ALB with EC2 instances and migrate the content to an EFS volume shared to EC2.
- B. Deploy a CloudFront distribution which uses the S3 bucket as an origin and migrate DNS to the CloudFront distribution endpoint.
- C. Replicate the S3 content to multiple regions and configure Route 53 latency-based routing entries to direct traffic to the appropriate region.
- D. Write a script to migrate all of the static S3 objects to S3-IA storage class.

Question Breakdown - Answers

The ALB will incur charges for simply existing, as well as for processing data. The EC2 instances will incur charges, and you will still have to pay for all of the outbound traffic to the clients, at a rate equal to or above that of S3.

- A. Deploy an ALB with EC2 instances and migrate the content to an EFS volume shared to EC2.
- B. Deploy a CloudFront distribution which uses the S3 bucket as an origin and migrate DNS to the CloudFront distribution endpoint.
- C. Replicate the S3 content to multiple regions and configure Route 53 latency-based routing entries to direct traffic to the appropriate region.
- D. Write a script to migrate all of the static S3 objects to S3-IA storage class.

Question Breakdown - Answers

CloudFront outbound charges are lower than that of S3, and there is no charge at all for transfer between CloudFront and S3.

- A. Deploy an ALB with EC2 instances and migrate the content to an EFS volume shared to EC2.
- B. Deploy a CloudFront distribution which uses the S3 bucket as an origin and migrate DNS to the CloudFront distribution endpoint.
- C. Replicate the S3 content to multiple regions and configure Route 53 latency-based routing entries to direct traffic to the appropriate region.
- D. Write a script to migrate all of the static S3 objects to S3-IA storage class.

Question Breakdown - Answers

Storing the content in multiple regions may have a possible benefit of decreased latency, but serving content from those multiple regions will not lower overall cost of the solution.

- A. Deploy an ALB with EC2 instances and migrate the content to an EFS volume shared to EC2.
- B. Deploy a CloudFront distribution which uses the S3 bucket as an origin and migrate DNS to the CloudFront distribution endpoint.
- C. Replicate the S3 content to multiple regions and configure Route 53 latency-based routing entries to direct traffic to the appropriate region.
- D. Write a script to migrate all of the static S3 objects to S3-IA storage class.

Question Breakdown - Answers

Assuming the objects are currently in S3 Standard storage class, transitioning them to S3-IA will result in increased overall transfer charges.

- A. Deploy an ALB with EC2 instances and migrate the content to an EFS volume shared to EC2.
- B. Deploy a CloudFront distribution which uses the S3 bucket as an origin and migrate DNS to the CloudFront distribution endpoint.
- C. Replicate the S3 content to multiple regions and configure Route 53 latency-based routing entries to direct traffic to the appropriate region.
- D. Write a script to migrate all of the static S3 objects to S3-IA storage class.

Question Breakdown - Correct Answer

Correct Answer: B

- A. Deploy an ALB with EC2 instances and migrate the content to an EFS volume shared to EC2.
- B. Deploy a CloudFront distribution which uses the S3 bucket as an origin and migrate DNS to the CloudFront distribution endpoint.
- C. Replicate the S3 content to multiple regions and configure Route 53 latency-based routing entries to direct traffic to the appropriate region.
- D. Write a script to migrate all of the static S3 objects to S3-IA storage class.



QA and Wrap-up!