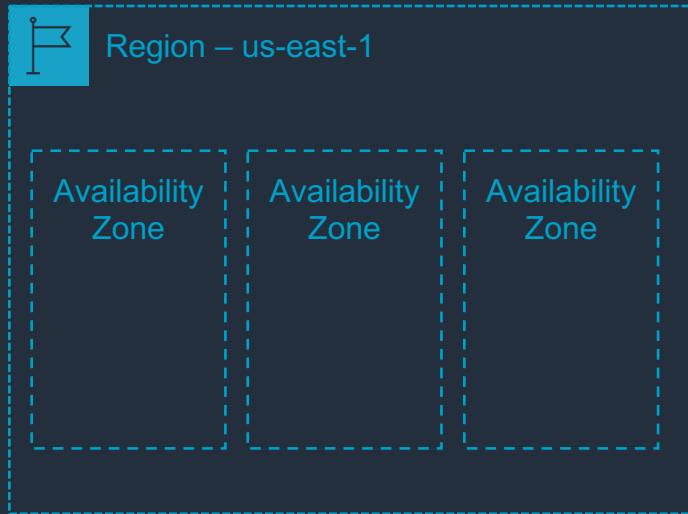


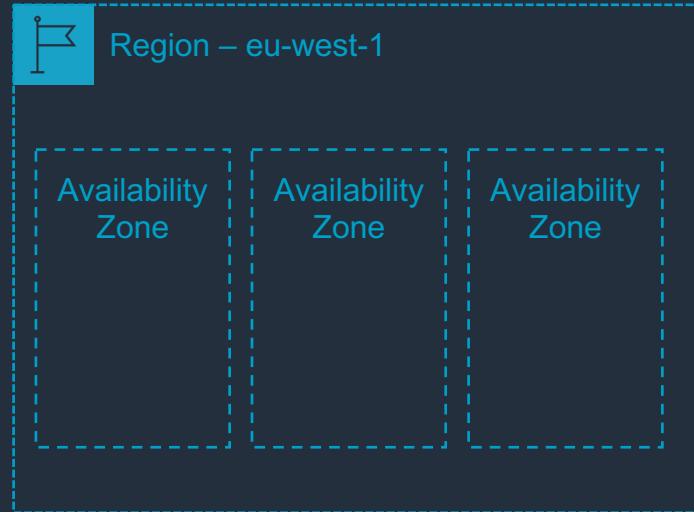
Section 2: AWS Global Infrastructure

Name	Description
Region	A geographical area with 2 or more AZs, isolated from other AWS regions
Availability Zone (AZ)	One or more data centers that are physically separate and isolated from other AZs
Edge Location	A location with a cache of content that can be delivered at low latency to users – used by CloudFront
Regional Edge Cache	Also part of the CloudFront network. These are larger caches that sit between AWS services and Edge Locations
Global Network	Highly available, low-latency private global network interconnecting every data center, AZ, and AWS region

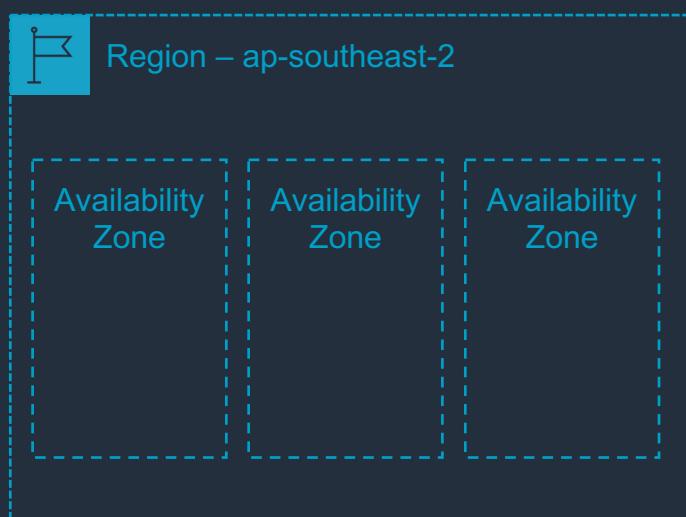
Section 2: AWS Global Infrastructure



Every region is connected via a high bandwidth, full redundant network



There are 23 regions around the world

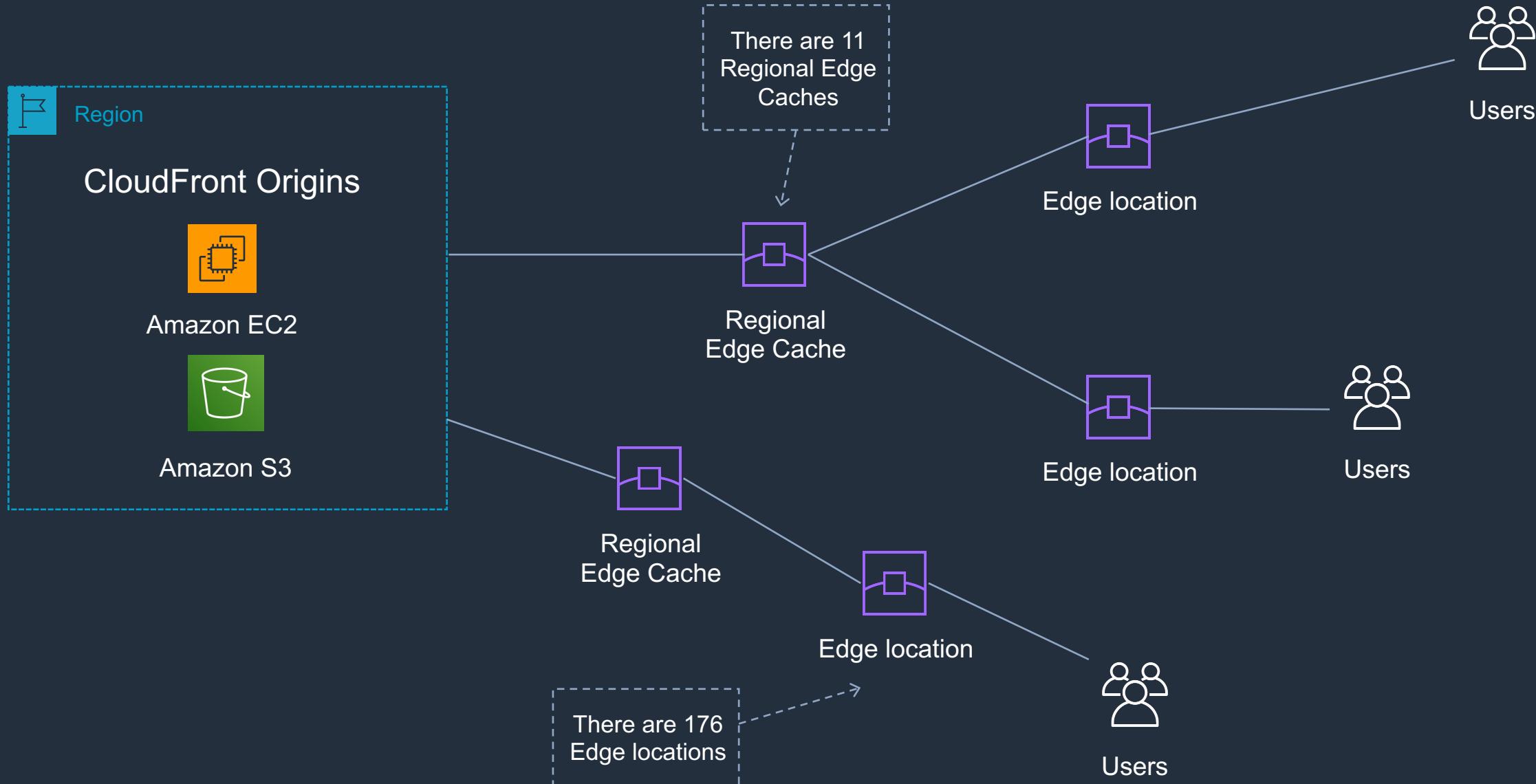


Each region is completely independent

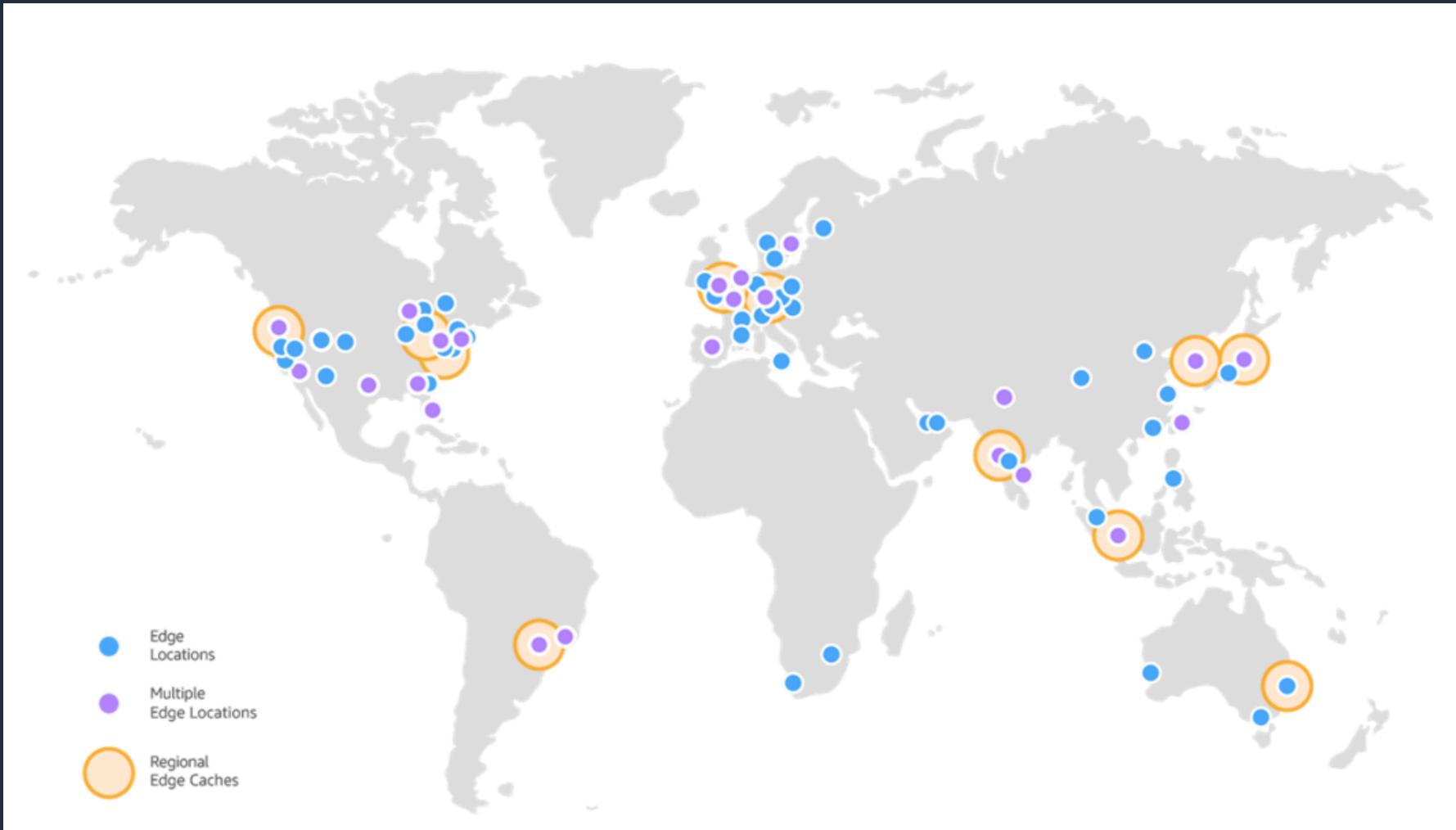
Section 2: AWS Global Infrastructure



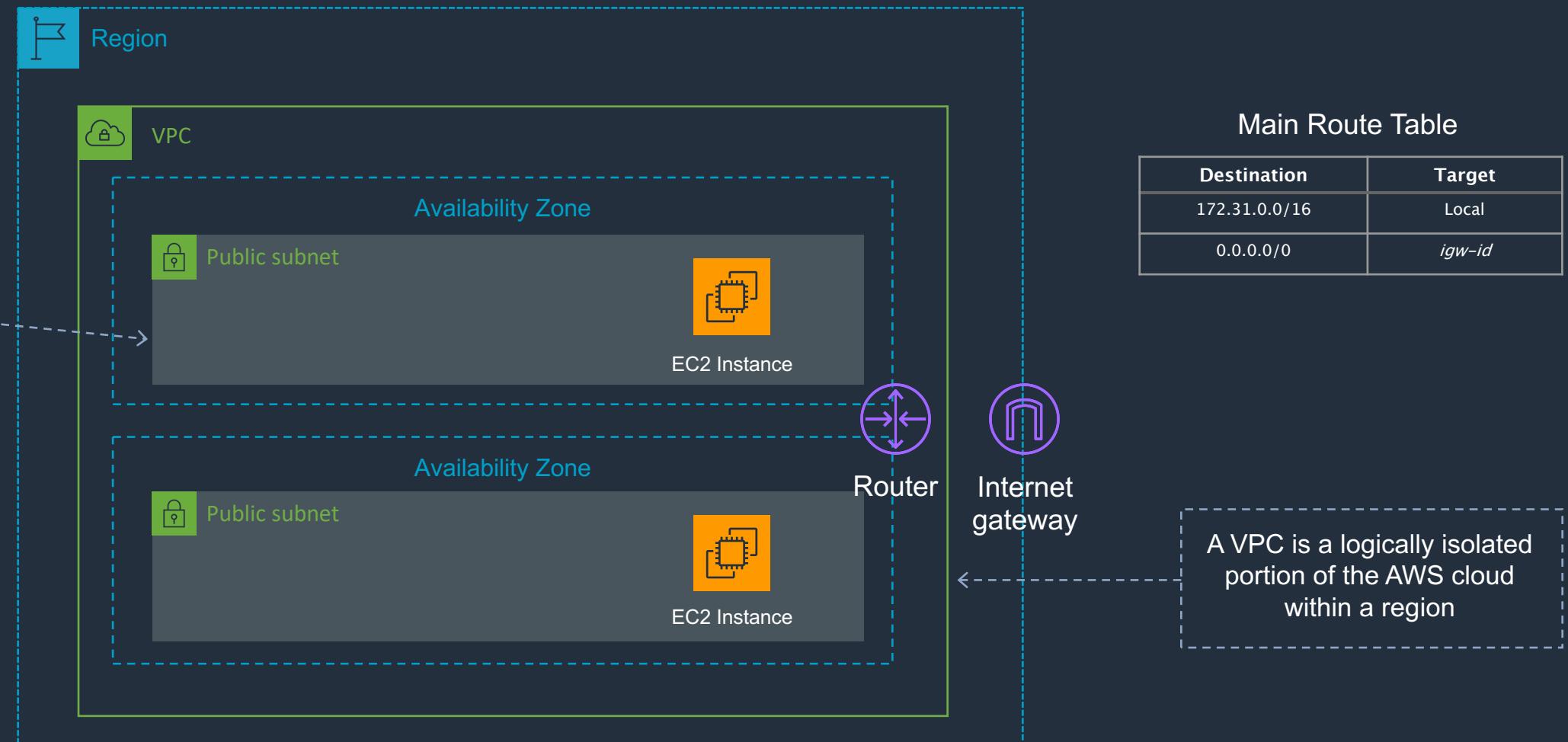
Section 2: CloudFront Edge Locations



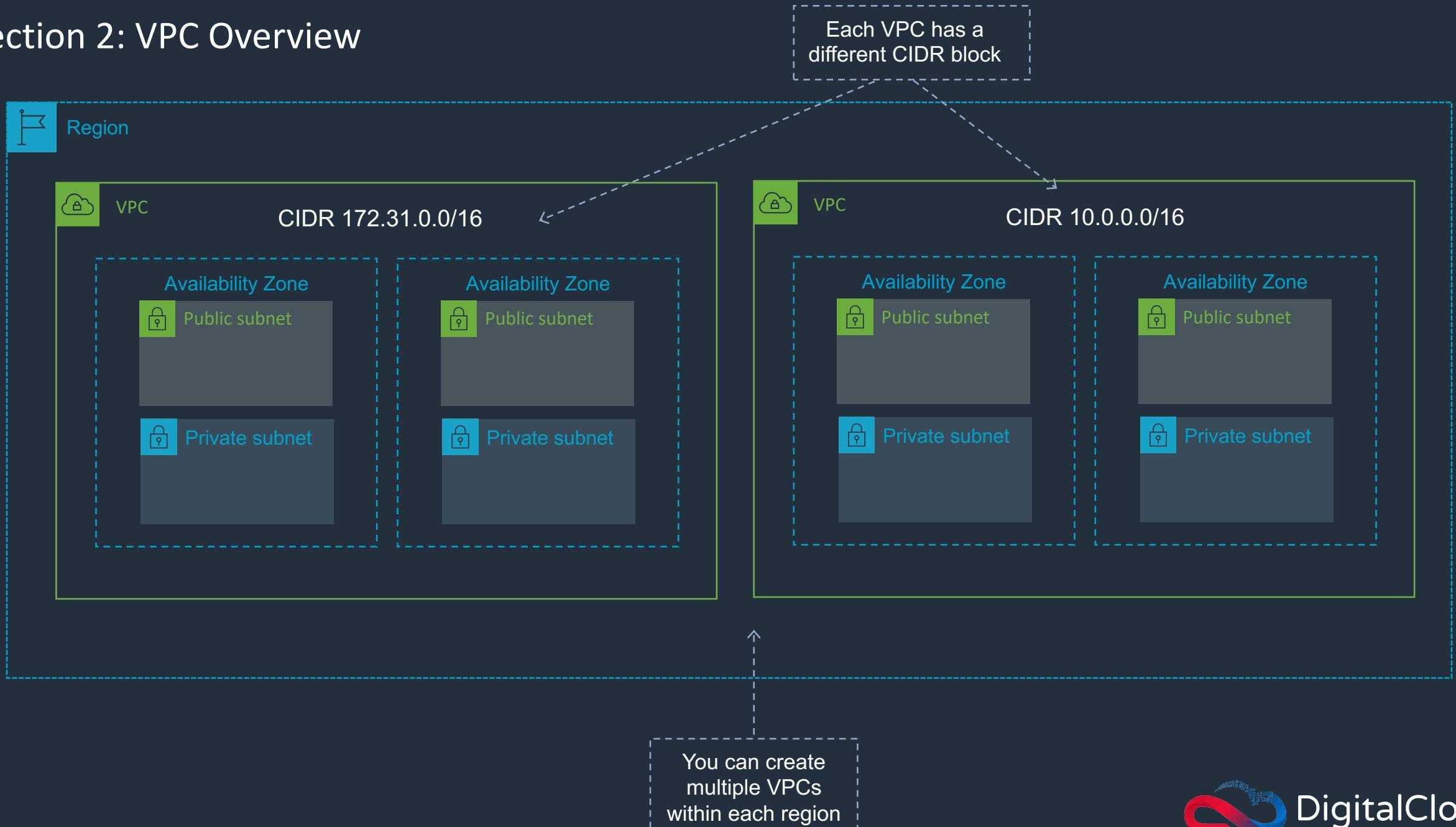
Section 2: CloudFront Edge Locations



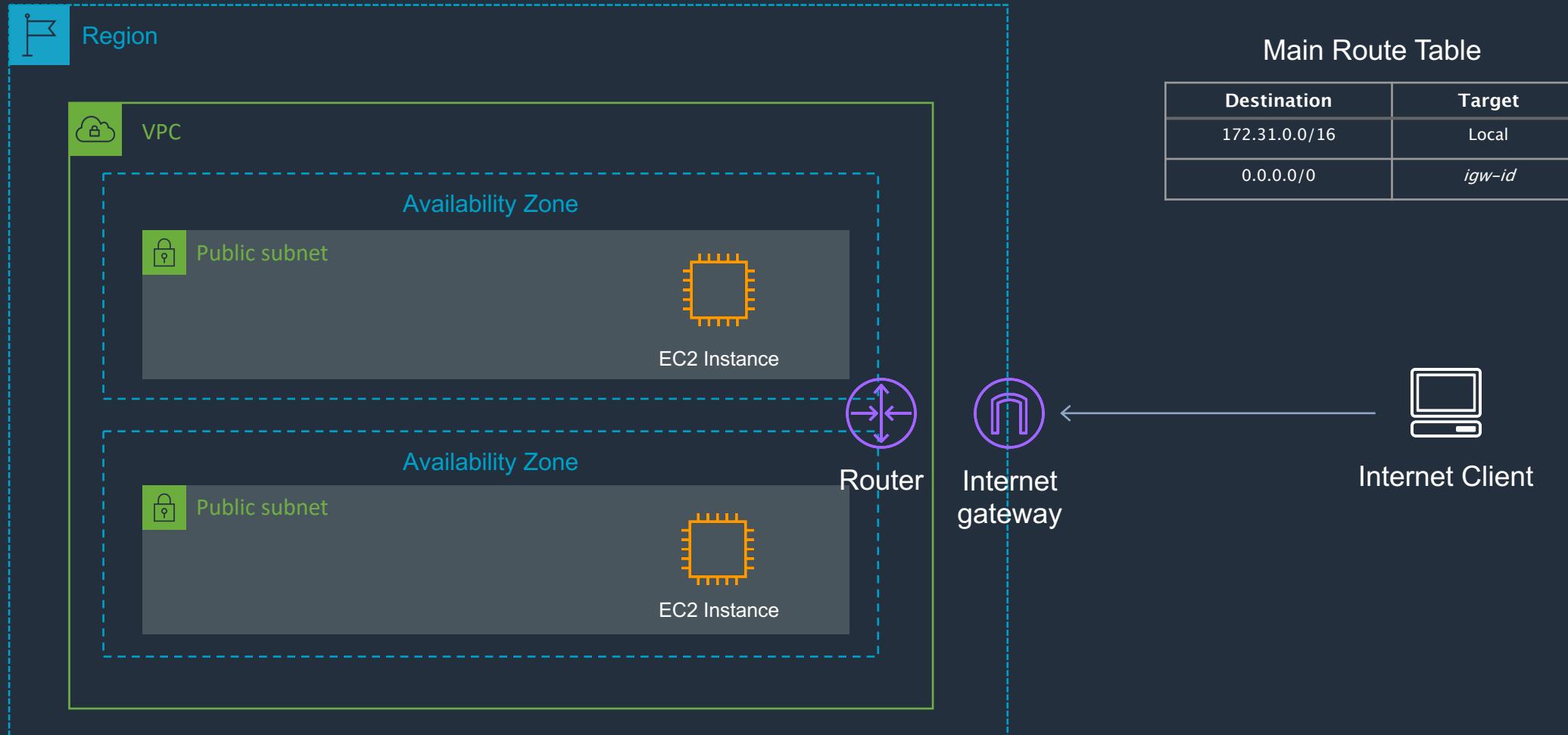
Section 2: VPC Overview



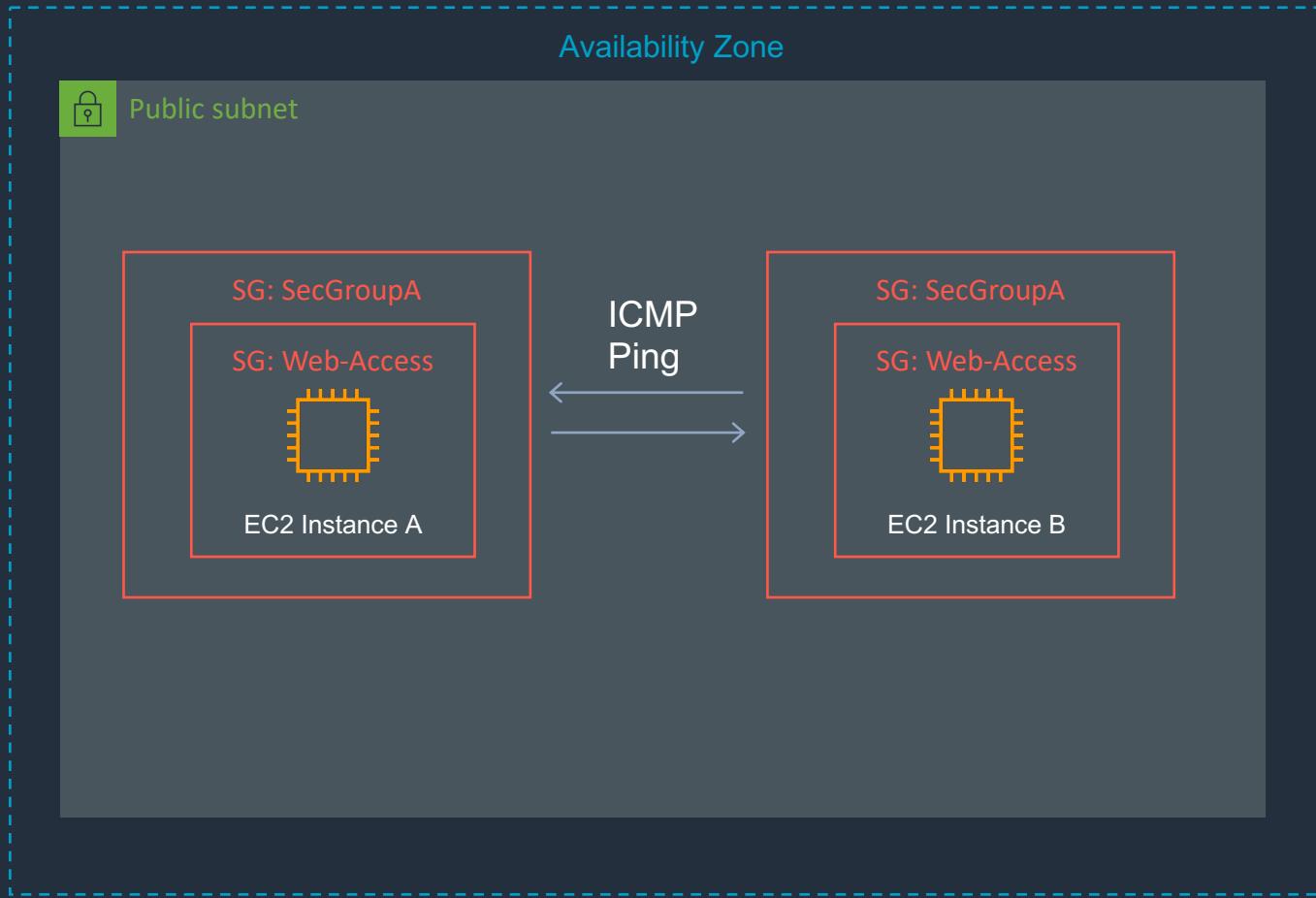
Section 2: VPC Overview



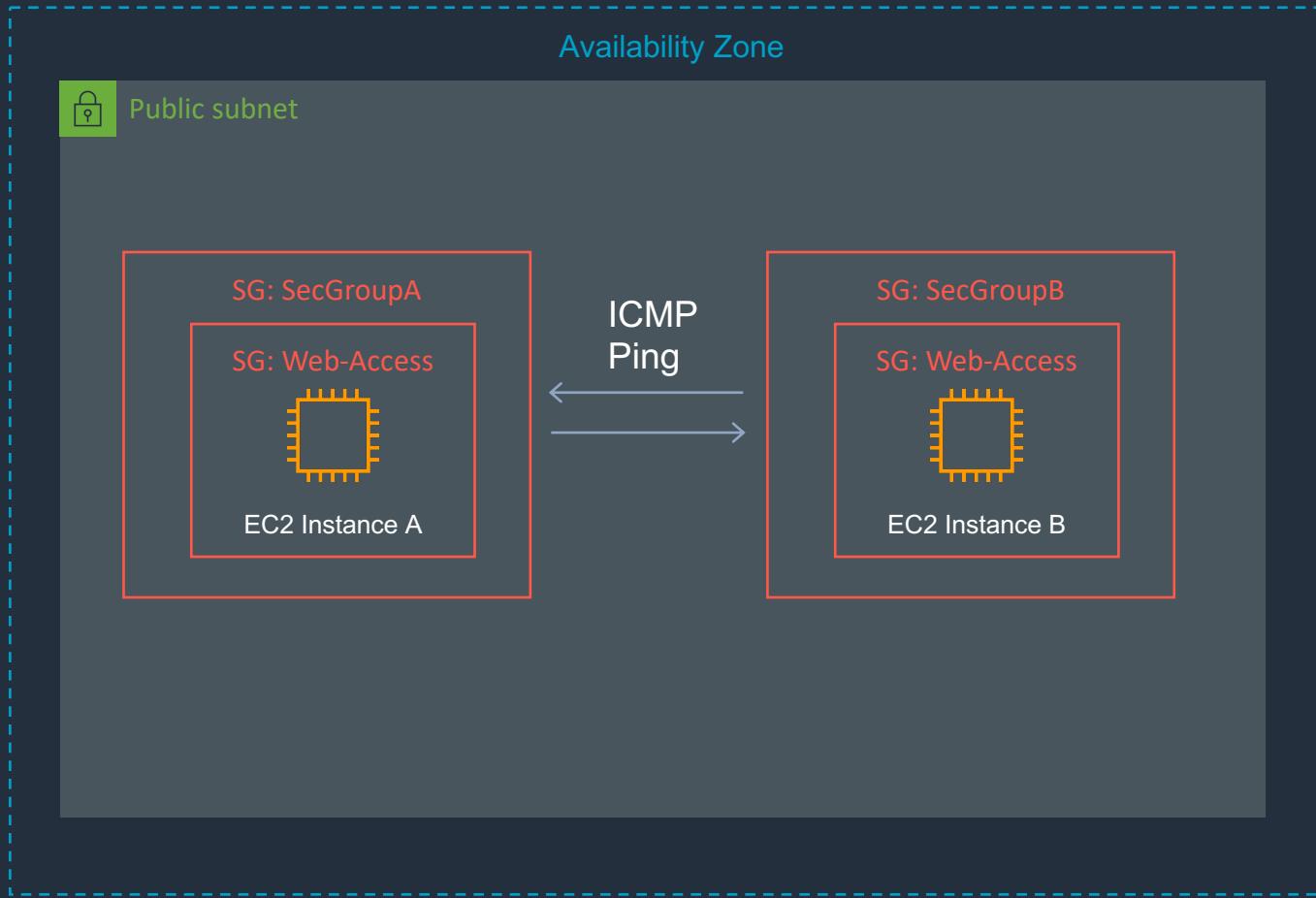
Section 3: Launch an EC2 Instance



Section 3: Security Group Slide 1



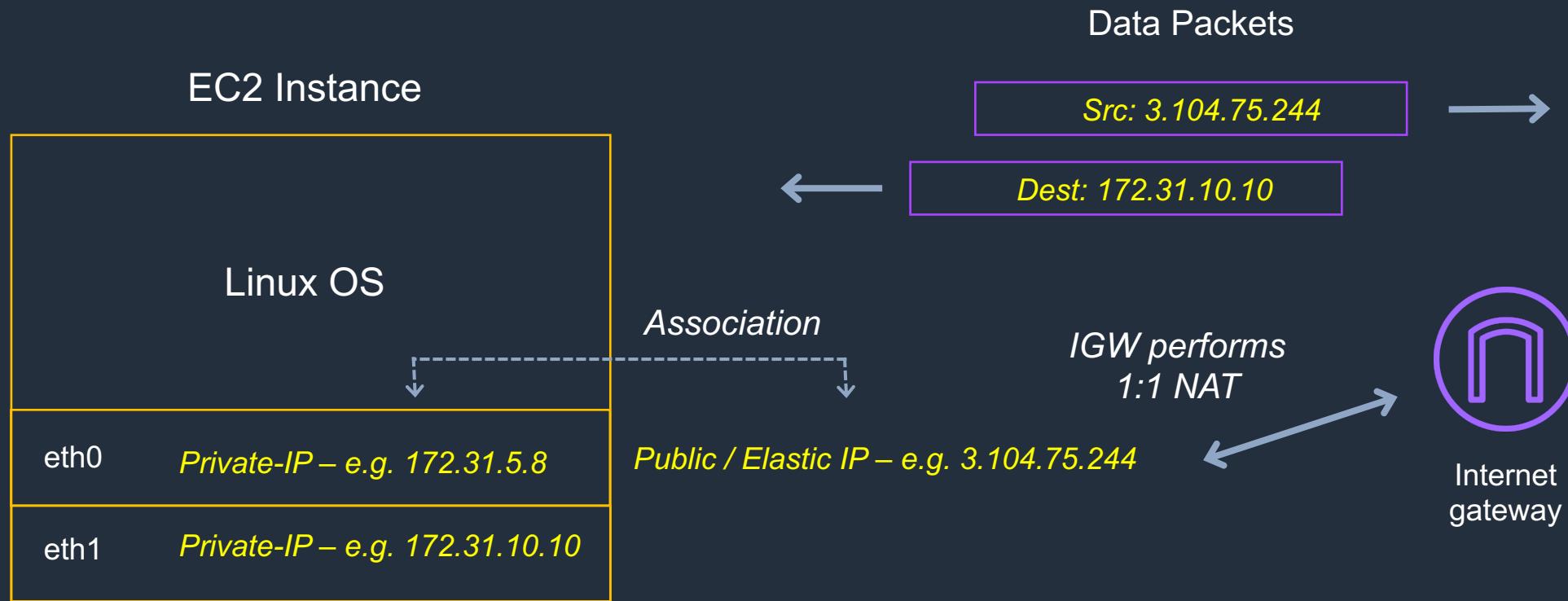
Section 3: Security Group Slide 2



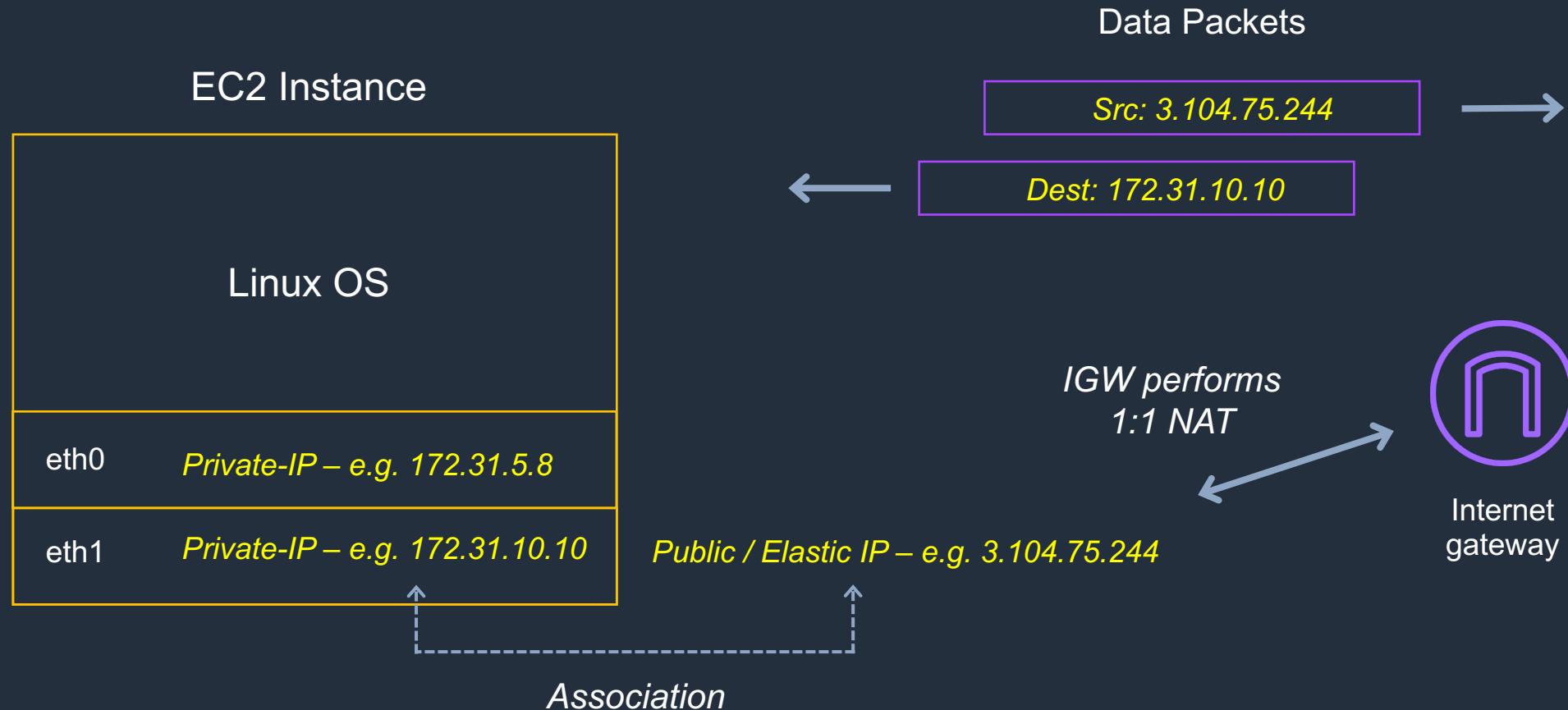
Section 3: Public, Private, and Elastic IP addresses

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

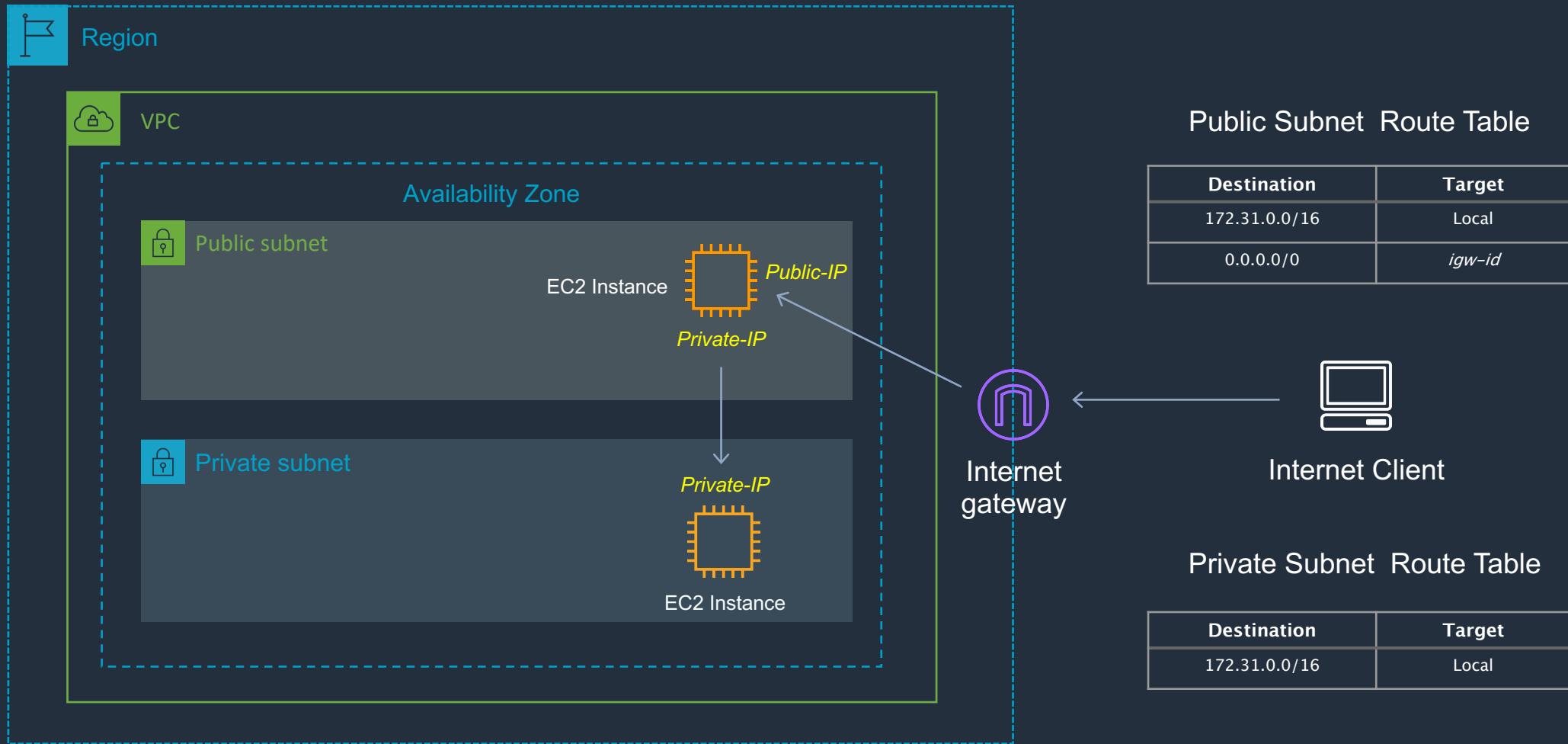
Section 3: Public, Private and Elastic IPs - Slide 1



Section 3: Public, Private and Elastic IPs - Slide 2



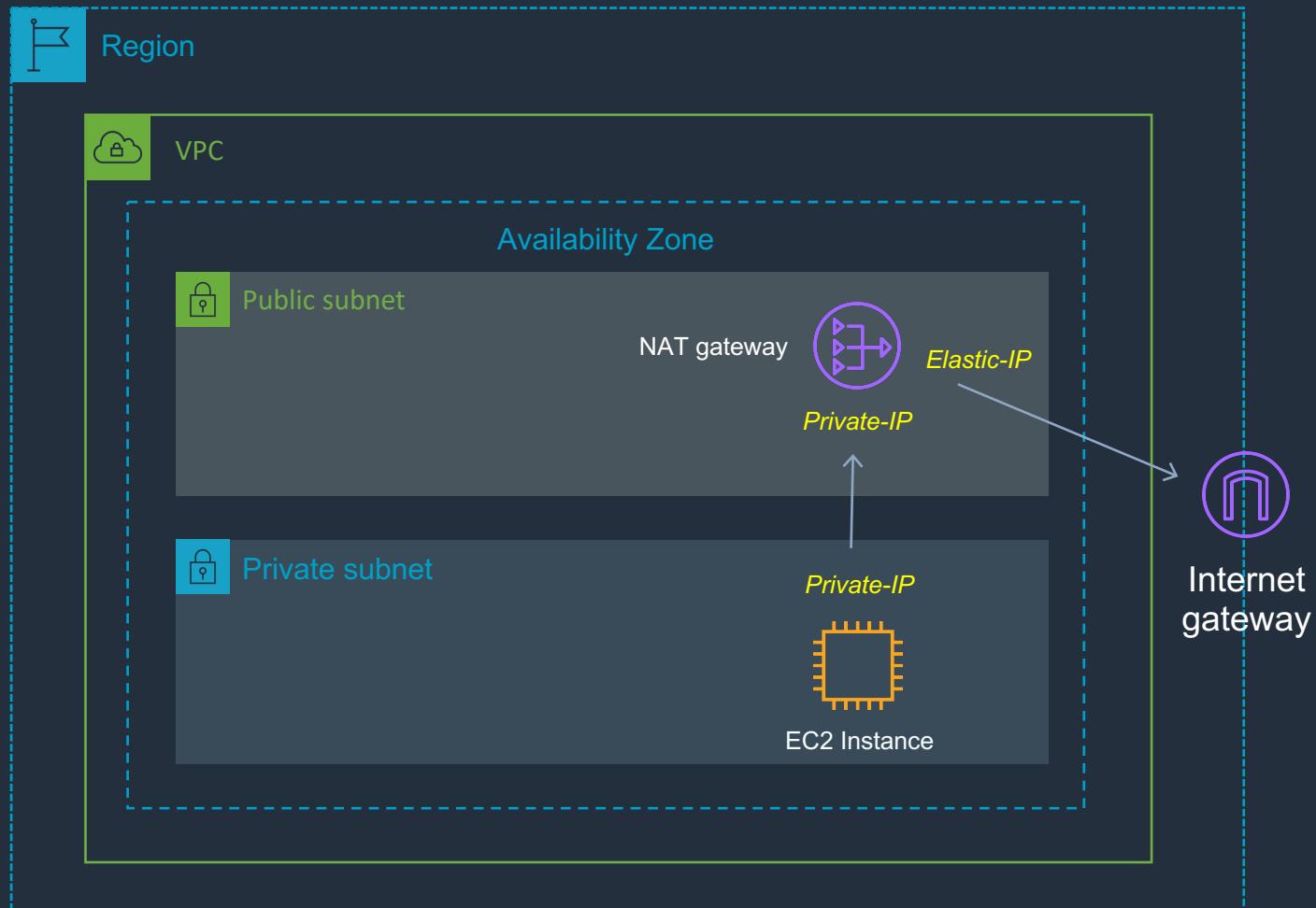
Section 3: Private Subnets and Bastion Hosts



Section 3: NAT Instance vs NAT Gateway

NAT Instance	NAT Gateway
Managed by you (e.g. software updates)	Managed by AWS
Scale up (instance type) manually and use enhanced networking	Elastic scalability up to 45 Gbps
No high availability – scripted/auto-scaled HA possible using multiple NATs in multiple subnets	Provides automatic high availability within an AZ and can be placed in multiple AZs
Need to assign Security Group	No Security Groups
Can use as a bastion host	Cannot access through SSH
Use an Elastic IP address or a public IP address with a NAT instance	Choose the Elastic IP address to associate with a NAT gateway at creation
Can implement port forwarding through manual customisation	Does not support port forwarding

Section 3: Private Subnet with NAT Gateway



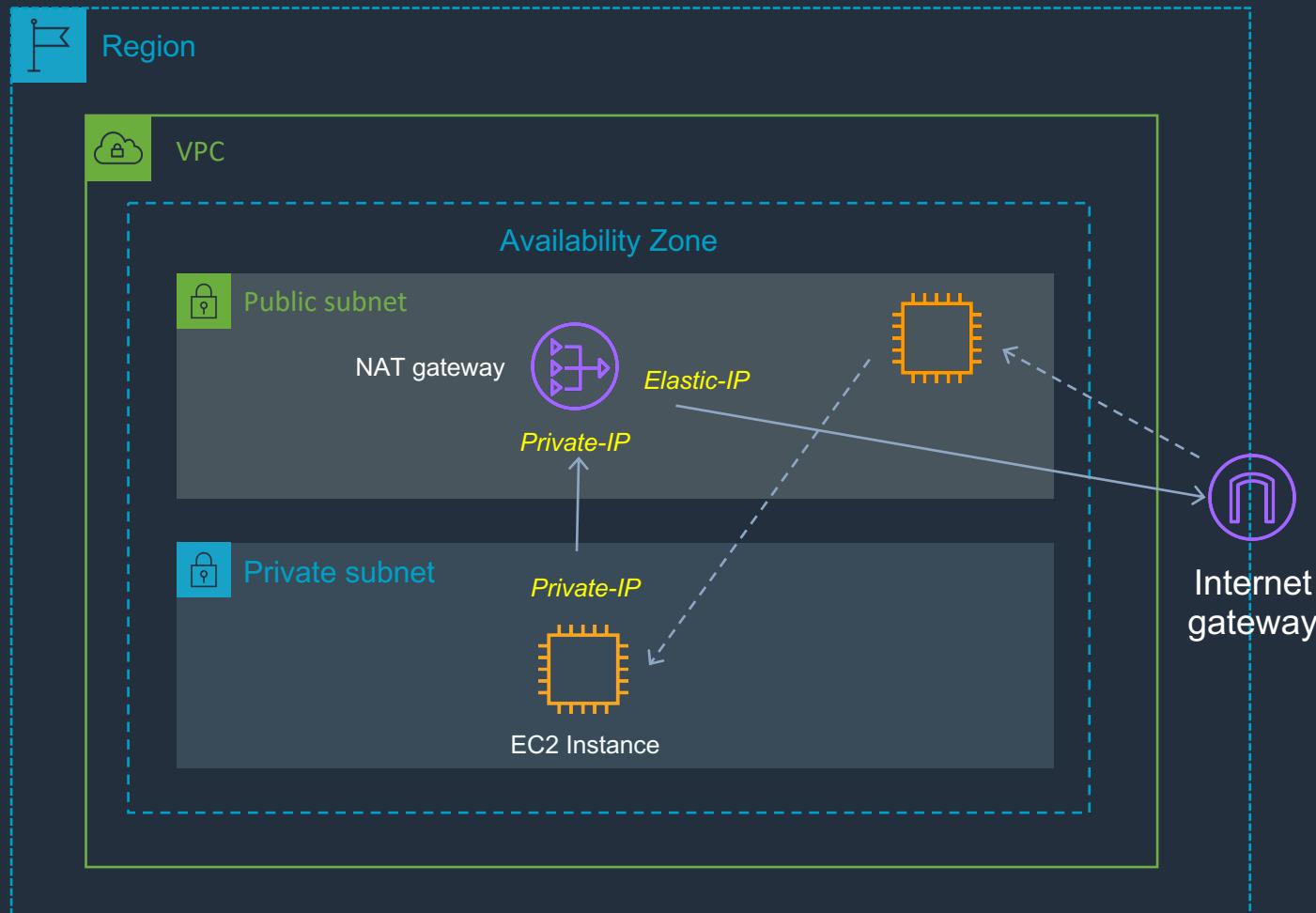
Public Subnet Route Table

Destination	Target
172.31.0.0/16	Local
0.0.0.0/0	<i>igw-id</i>

Private Subnet Route Table

Destination	Target
172.31.0.0/16	Local
0.0.0.0/0	<i>nat-gateway-id</i>

Section 3: Agent Forwarding with Putty



Public Subnet Route Table

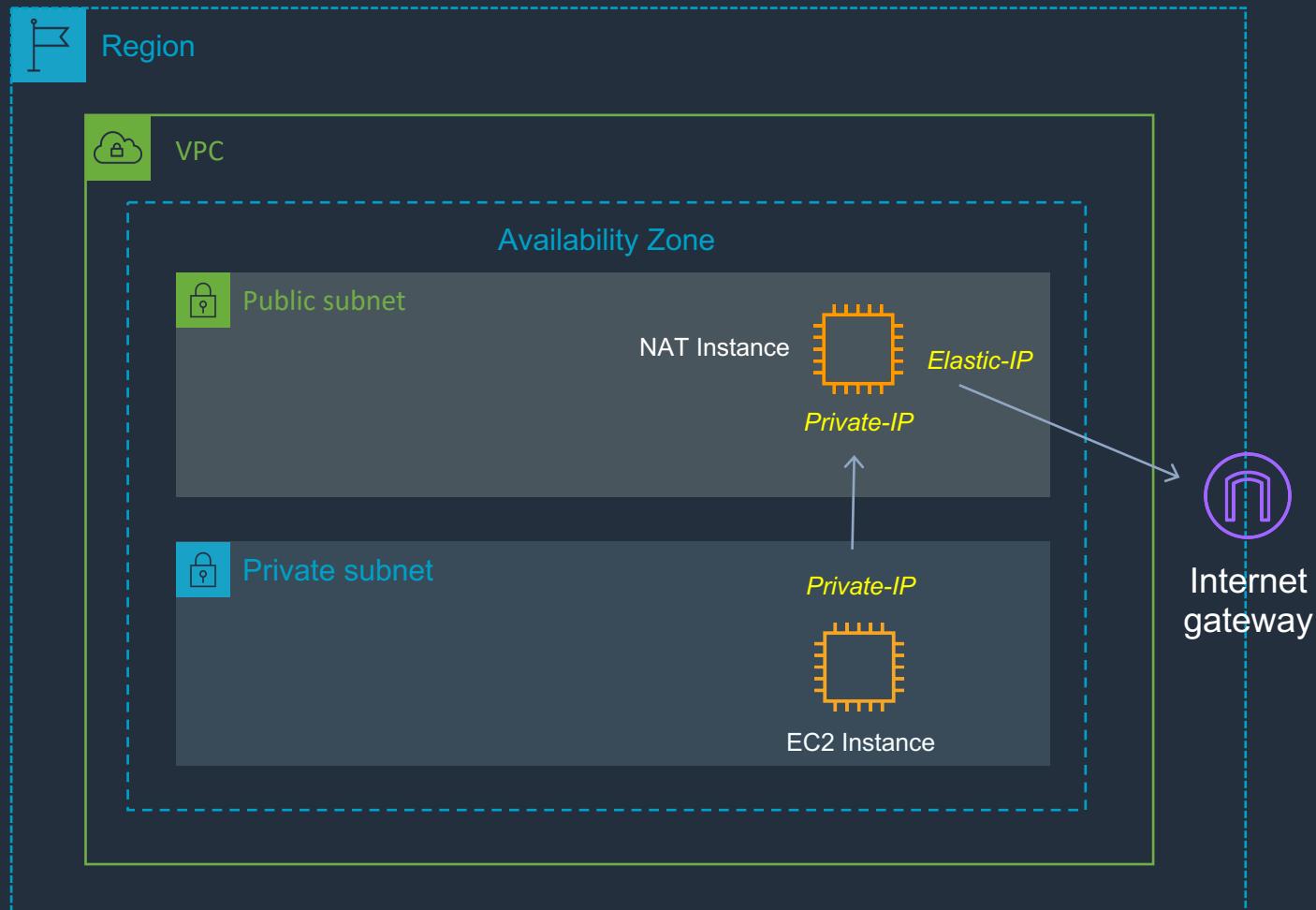
Destination	Target
172.31.0.0/16	Local
0.0.0.0/0	<i>igw-id</i>



Private Subnet Route Table

Destination	Target
172.31.0.0/16	Local
0.0.0.0/0	<i>nat-gateway-id</i>

Section 3: Private Subnet with NAT Instance



Public Subnet Route Table

Destination	Target
172.31.0.0/16	Local
0.0.0.0/0	<i>igw-id</i>

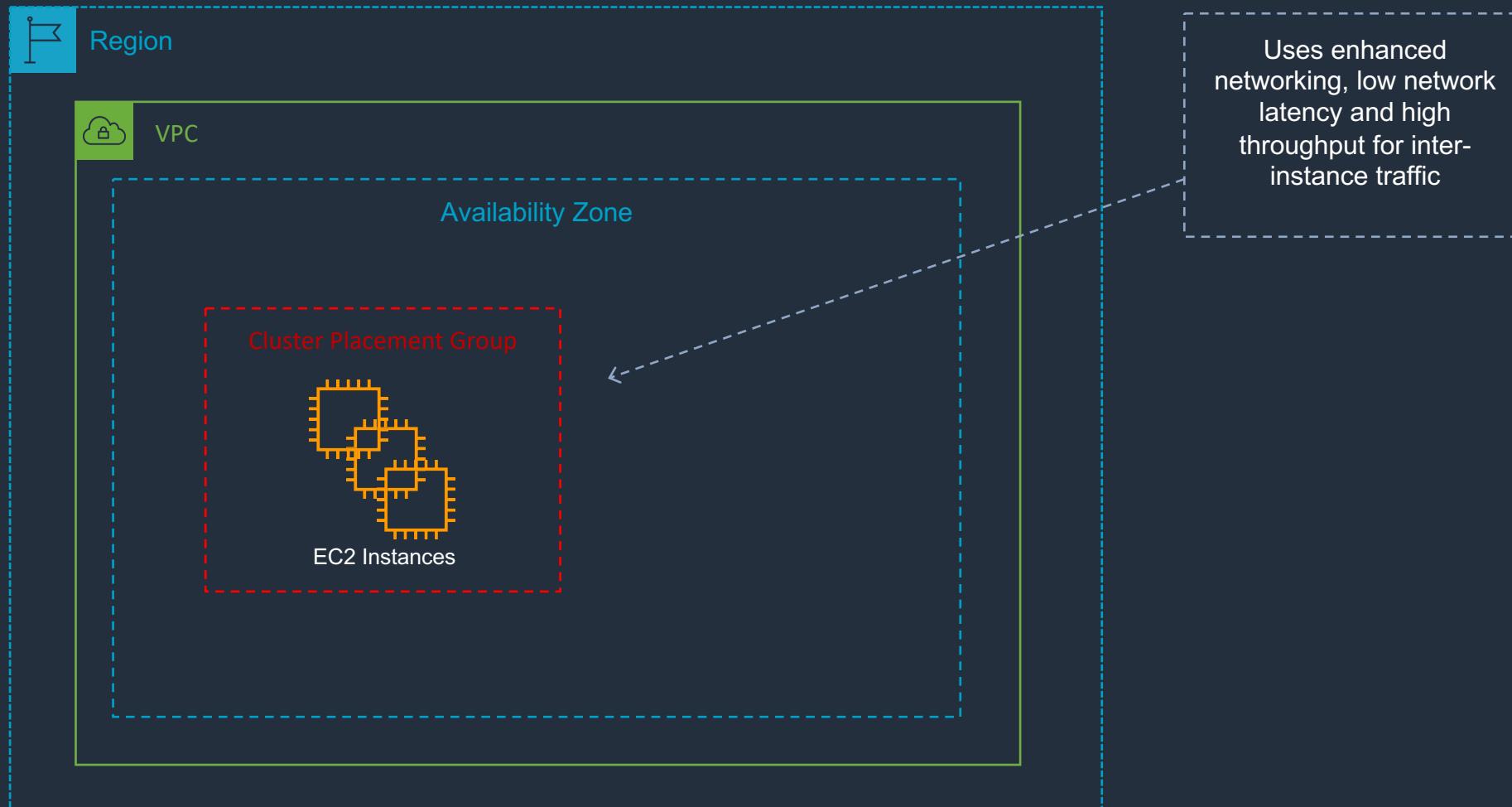
Private Subnet Route Table

Destination	Target
172.31.0.0/16	Local
0.0.0.0/0	<i>nat-instance-id</i>

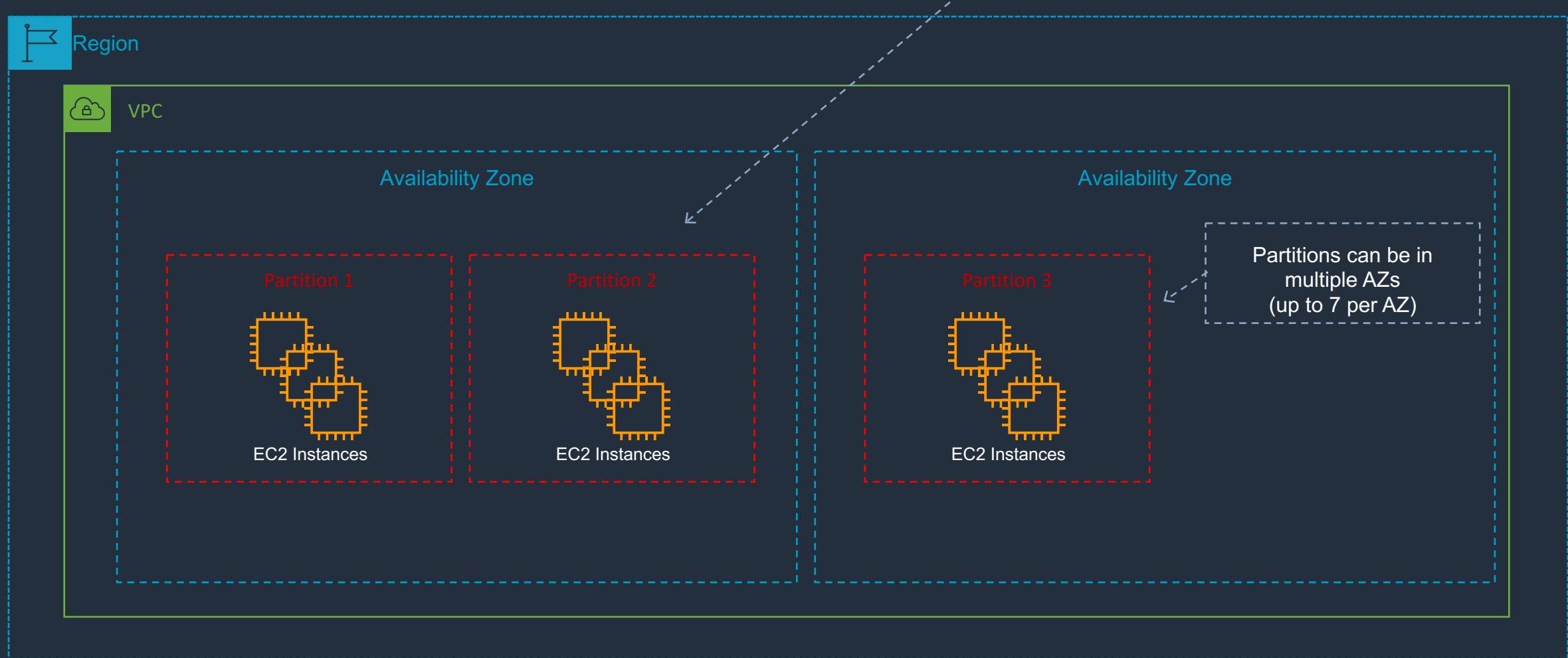
Section 3: Placement Groups

- **Cluster** – packs instances close together inside an Availability Zone. This strategy enables workloads to achieve the low-latency network performance necessary for tightly-coupled node-to-node communication that is typical of HPC applications.
- **Partition** – spreads your instances across logical partitions such that groups of instances in one partition do not share the underlying hardware with groups of instances in different partitions. This strategy is typically used by large distributed and replicated workloads, such as Hadoop, Cassandra, and Kafka.
- **Spread** – strictly places a small group of instances across distinct underlying hardware to reduce correlated failures.

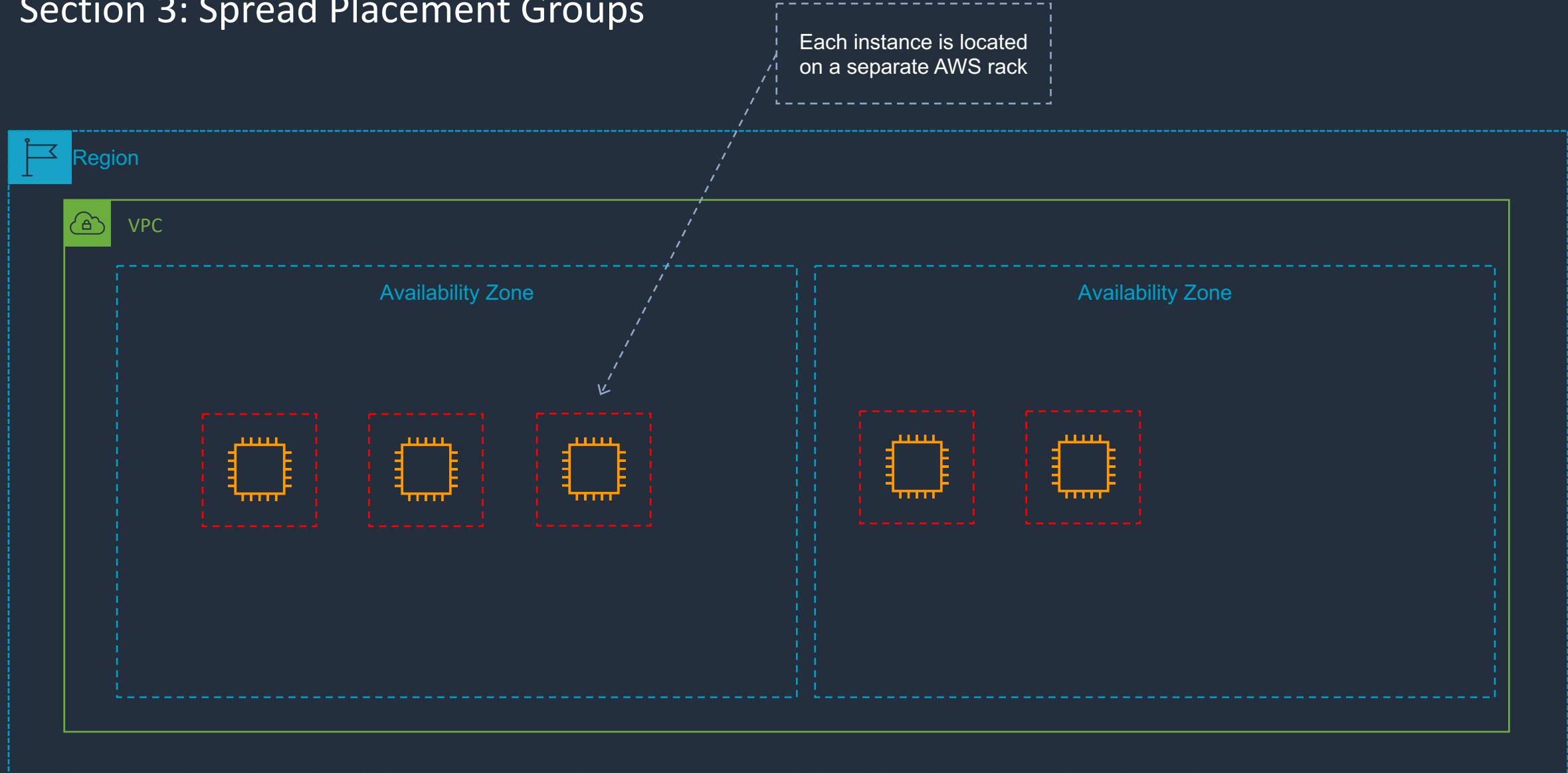
Section 3: Cluster Placement Groups



Section 3: Partition Placement Groups



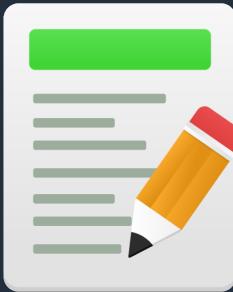
Section 3: Spread Placement Groups



Section 3: Placement Groups

	Clustered	Spread	Partition
What	Instances are placed into a low-latency group within a single AZ	Instances are spread across underlying hardware	Instances are grouped into logical segments called partitions which use distinct hardware
When	Need low network latency and/or high network throughput	Reduce the risk of simultaneous instance failure if underlying hardware fails	Need control and visibility into instance placement
Pros	Get the most out of enhanced networking Instances	Can span multiple AZs	Reduces likelihood of correlated failures for large workloads.
Cons	Finite capacity: recommend launching all you might need up front	Maximum of 7 instances running per group, per AZ	Partition placement groups are not supported for Dedicated Hosts

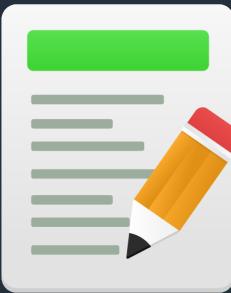
Section 3: Exam Cram



Exam cram is:

- A summary of important facts you need to know for the exam
- Relevant to the section - more detail on topics within the section
- Fast-paced run through of facts to supplement the hands-on
- No need to take notes unless you want to – all of these facts and more
are documented in the Training Notes on our website
- Additionally, we provide the Training Notes in a PDF for offline study
- After exam cram, you can complete a short quiz with unique questions
from the topics covered

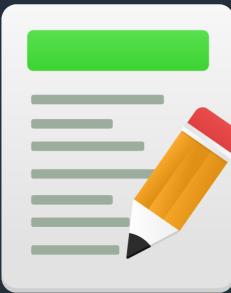
Section 3: Exam Cram



Amazon EC2

- Amazon Elastic Compute Cloud (Amazon EC2) is a web service that provides resizable compute capacity in the cloud.
- With EC2 you have full control at the operating system layer.
- Key pairs are used to securely connect to EC2 instances:
 - A key pair consists of a public key that AWS stores, and a private key file that you store.
 - For Windows AMIs, the private key file is required to obtain the password used to log into your instance.
 - For Linux AMIs, the private key file allows you to securely SSH (secure shell) into your instance.

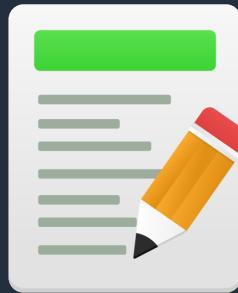
Section 3: Exam Cram



Amazon EC2

- User data is data that is supplied by the user at instance launch in the form of a script.
- Instance metadata is data about your instance that you can use to configure or manage the running instance.
- Instance metadata is available at <http://169.254.169.254/latest/meta-data/> (the trailing "/" is required).
- Instance user data is available at: <http://169.254.169.254/latest/user-data>.

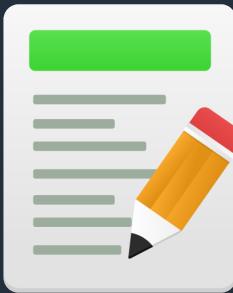
Section 3: Exam Cram



Amazon EC2 Pricing Models

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

Section 3: Exam Cram

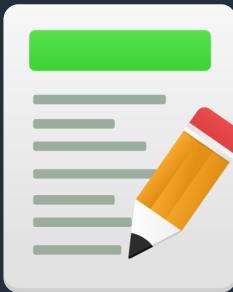


Amazon EC2 Dedicated Hosts and Instances

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

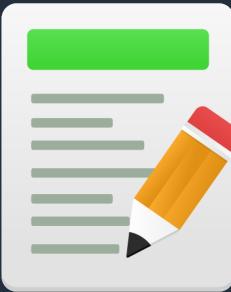
Section 3: Exam Cram

Amazon EC2 Instance Types



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

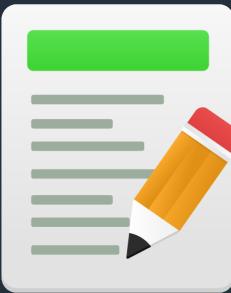
Section 3: Exam Cram



Amazon EC2 AMIs

- An Amazon Machine Image (AMI) provides the information required to launch an instance.
- An AMI includes the following:
 - A template for the root volume for the instance (for example, an operating system, an application server, and applications).
 - Launch permissions that control which AWS accounts can use the AMI to launch instances.
 - A block device mapping that specifies the volumes to attach to the instance when it's launched.

Section 3: Exam Cram



Amazon EC2 AMIs

- Volumes attached to the instance are either EBS or Instance store:
 - Amazon Elastic Block Store (EBS) provides persistent storage. EBS snapshots, which reside on Amazon S3, are used to create the volume.
 - Instance store volumes are ephemeral (non-persistent). That means data is lost if the instance is shut down. A template stored on Amazon S3 is used to create the volume.
- AMIs are regional. You can only launch an AMI from the region in which it is stored. However, you can copy AMI's to other regions using the console, command line, or the API.

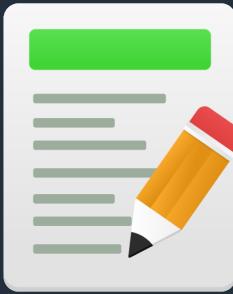
Section 3: Exam Cram

Amazon EC2 IP Addresses



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

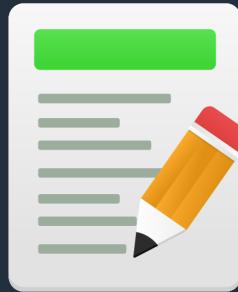
Section 3: Exam Cram



Elastic Network Interface (ENI).

- A logical networking component in a VPC that represents a virtual network card.
- Can include attributes such as IP addresses, security groups, MAC address, source/destination check flag, description.
- You can create and configure network interfaces in your account and attach them to instances in your VPC.
- eth0 is the primary network interface and cannot be moved or detached.

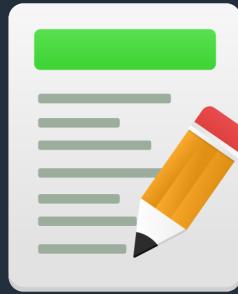
Section 3: Exam Cram



Elastic Network Interface (ENI):

- A logical networking component in a VPC that represents a virtual network card.
- Can include attributes such as IP addresses, security groups, MAC address, source/destination check flag, description.
- You can create and configure network interfaces in your account and attach them to instances in your VPC.
- eth0 is the primary network interface and cannot be moved or detached.
- An ENI is bound to an AZ and you can specify which subnet/AZ you want the ENI to be added in.

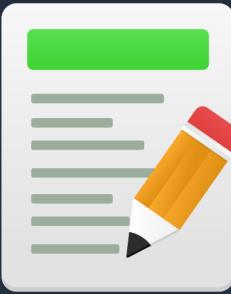
Section 3: Exam Cram



Elastic Network Adapter (ENA):

- Used for Enhanced Networking.
- Provides higher bandwidth, higher packet-per-second (PPS) performance and lower latency.
- Must launch an HVM AMI.
- Available for certain instance types within a VPC.

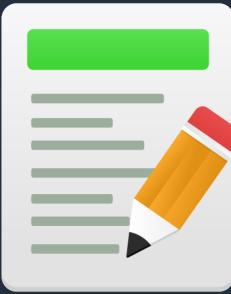
Section 3: Exam Cram



Elastic Fabric Adapter (EFA):

- An AWS Elastic Network Adapter (ENA) with added capabilities.
- Enables customers to run applications requiring high levels of inter-node communications at scale on AWS.
- With EFA, High Performance Computing (HPC) applications using the Message Passing Interface (MPI) and Machine Learning (ML) applications using NVIDIA Collective Communications Library (NCCL) can scale to thousands of CPUs or GPUs.

Section 3: Exam Cram

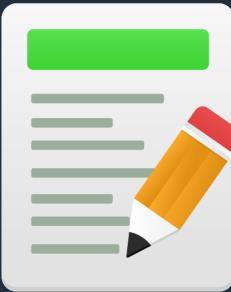


ENI vs ENA vs EFA:

- **When to use ENI:** This is the basic adapter type for when you don't have any high-performance requirements. Can use with all instance types.
- **When to use ENA:** Good for use cases that require higher bandwidth and lower inter-instance latency. Supported for limited instance types (HVM only).
- **When to use EFA:** High Performance Computing. MPI and ML use cases. Tightly coupled applications. Can use with all instance types.

Section 3: Exam Cram

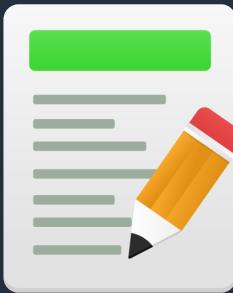
Placement Groups



	Clustered	Spread	Partition
What	Instances are placed into a low-latency group within a single AZ	Instances are spread across underlying hardware	Instances are grouped into logical segments called partitions which use distinct hardware
When	Need low network latency and/or high network throughput	Reduce the risk of simultaneous instance failure if underlying hardware fails	Need control and visibility into instance placement
Pros	Get the most out of enhanced networking Instances	Can span multiple AZs	Reduces likelihood of correlated failures for large workloads.
Cons	Finite capacity: recommend launching all you might need up front	Maximum of 7 instances running per group, per AZ	Partition placement groups are not supported for Dedicated Hosts



Section 3: Exam Cram



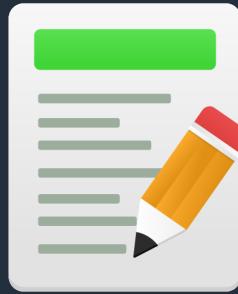
IAM Roles:

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

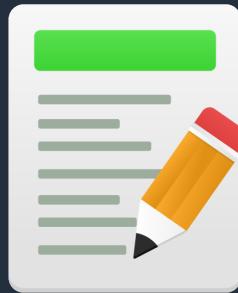
Section 3: Exam Cram

Monitoring

- EC2 status checks are performed every minute and each returns a pass or a fail status.
- If all checks pass, the overall status of the instance is OK. If one or more checks fail, the overall status is impaired.
- System status checks detect (`StatusCheckFailed_System`) problems with your instance that require [AWS](#) involvement to repair.
- Instance status checks (`StatusCheckFailed_Instance`) detect problems that require [your](#) involvement to repair.
- You can create Amazon CloudWatch alarms that monitor Amazon EC2 instances and automatically perform an action if the status check fails.



Section 3: Exam Cram



NAT Instance	NAT Gateway
Managed by you (e.g. software updates)	Managed by AWS
Scale up (instance type) manually and use enhanced networking	Elastic scalability up to 45 Gbps
No high availability – scripted/auto-scaled HA possible using multiple NATs in multiple subnets	Provides automatic high availability within an AZ and can be placed in multiple AZs
Need to assign Security Group	No Security Groups
Can use as a bastion host	Cannot access through SSH
Use an Elastic IP address or a public IP address with a NAT instance	Choose the Elastic IP address to associate with a NAT gateway at creation
Can implement port forwarding through manual customisation	Does not support port forwarding

Section 4: Amazon S3 Overview



S3 Bucket

[http://*bucket*.s3.*aws-region*.amazonaws.com](http://bucket.s3.amazonaws.com)
[http://s3.*aws-region*.amazonaws.com/*bucket*](http://s3.amazonaws.com/bucket)



Object

- Key
- Version ID
- Value
- Metadata
- Subresources
- Access control information



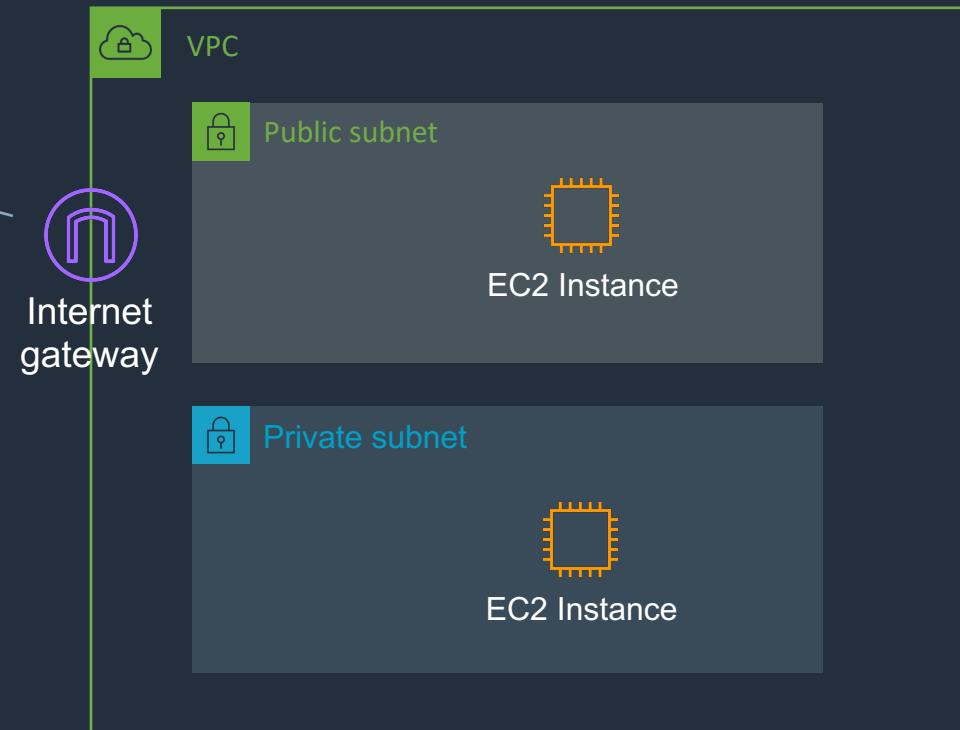
Public Internet



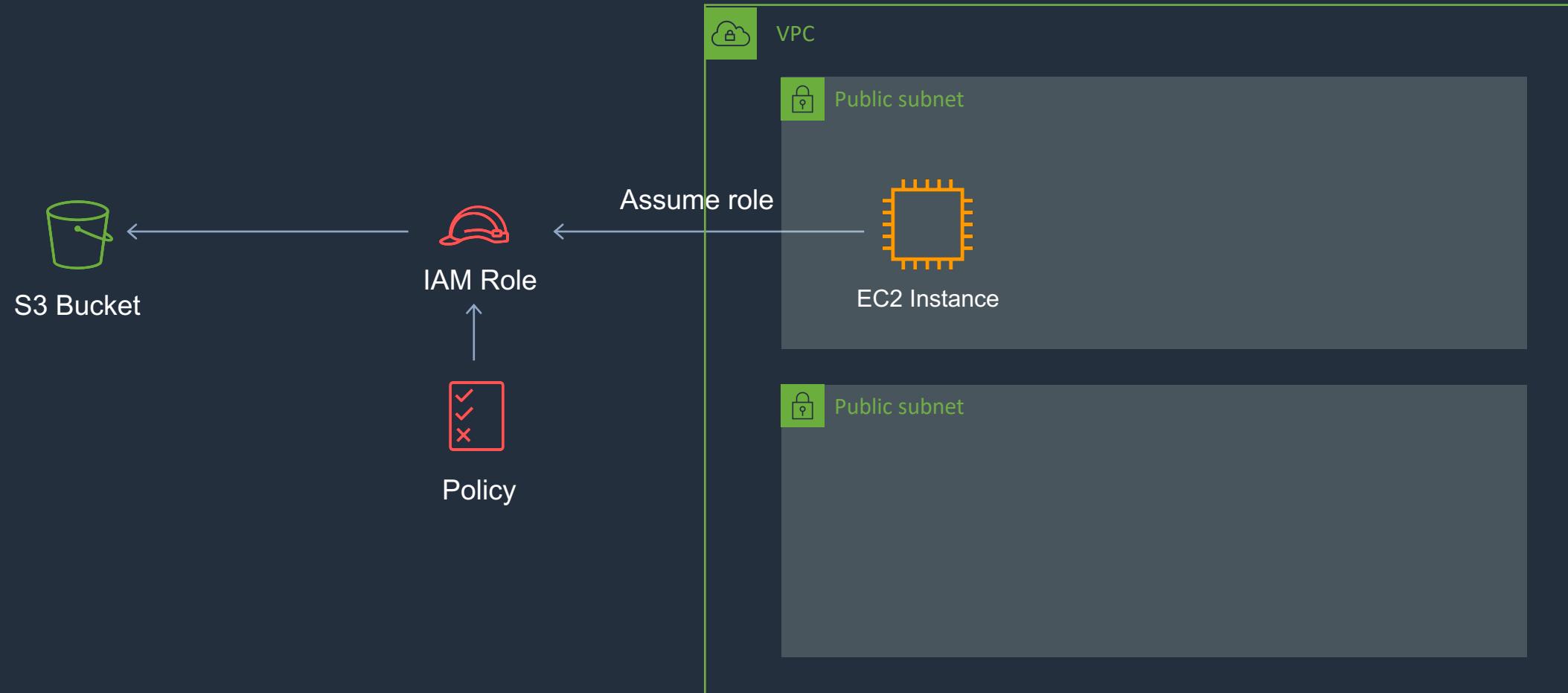
Internet Client



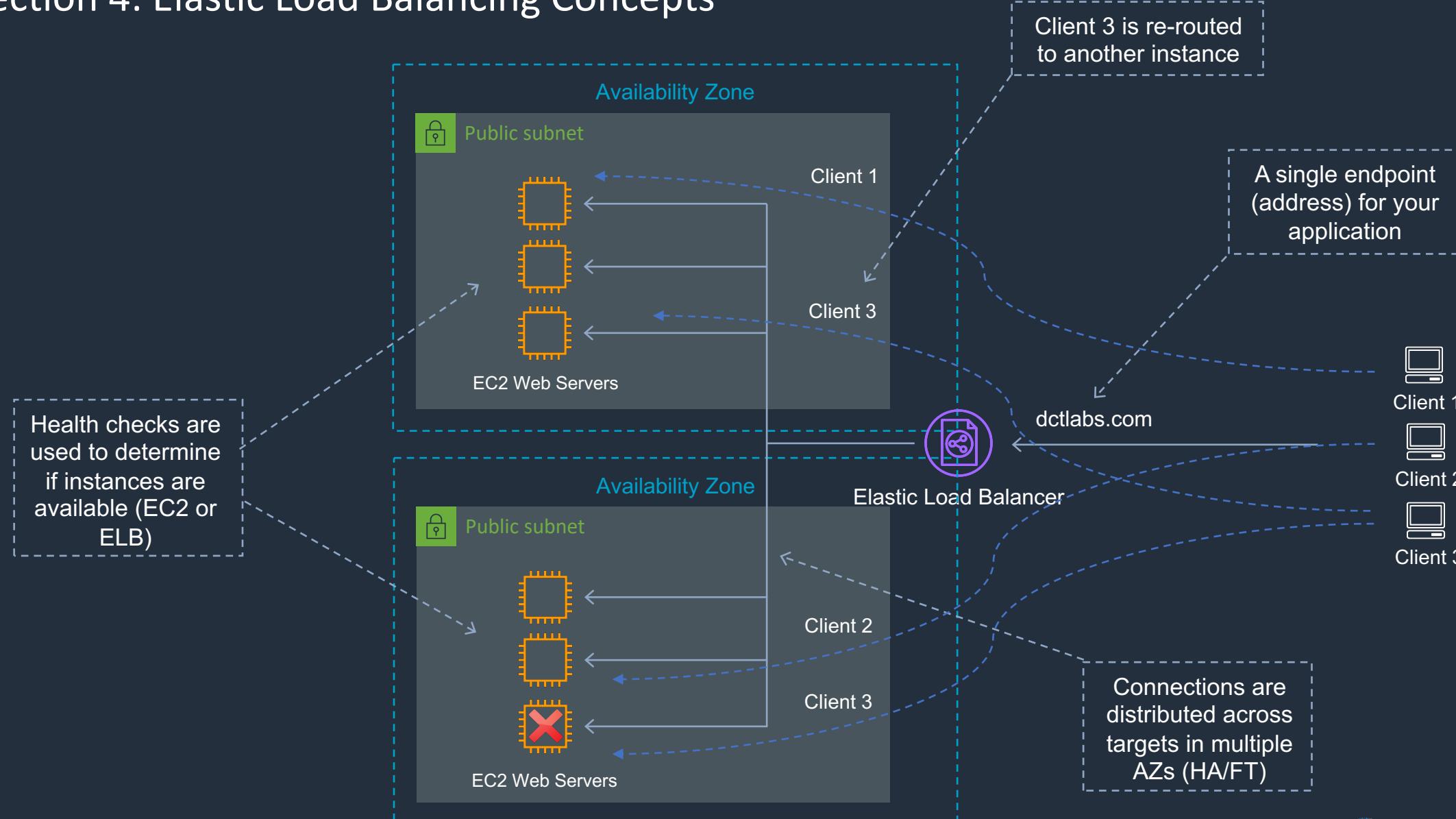
Amazon S3



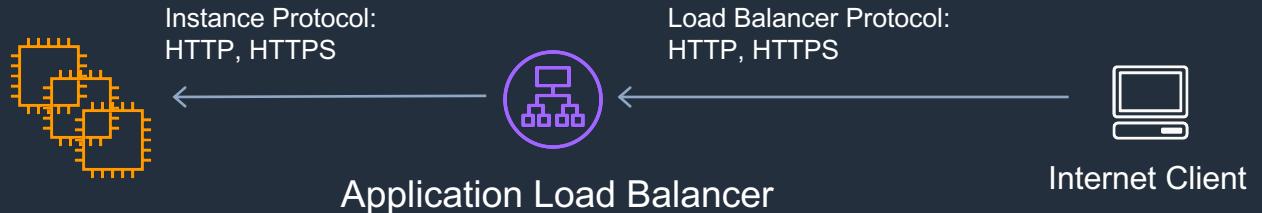
Section 4: IAM Roles



Section 4: Elastic Load Balancing Concepts

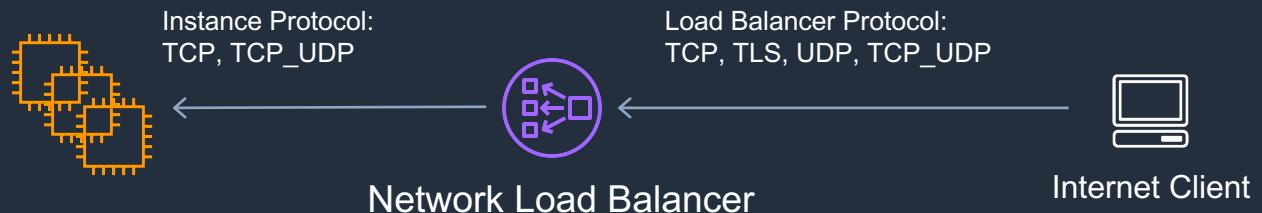


Section 4: Elastic Load Balancing (ELB) Types



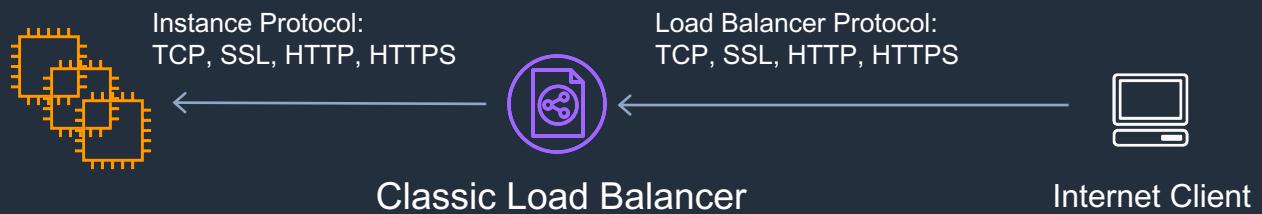
Application Load Balancer

- Operates at the request level
- Routes based on the content of the request (layer 7)
- Supports path-based routing, host-based routing, query string parameter-based routing, and source IP address-based routing
- Supports IP addresses, Lambda Functions and containers as targets



Network Load Balancer

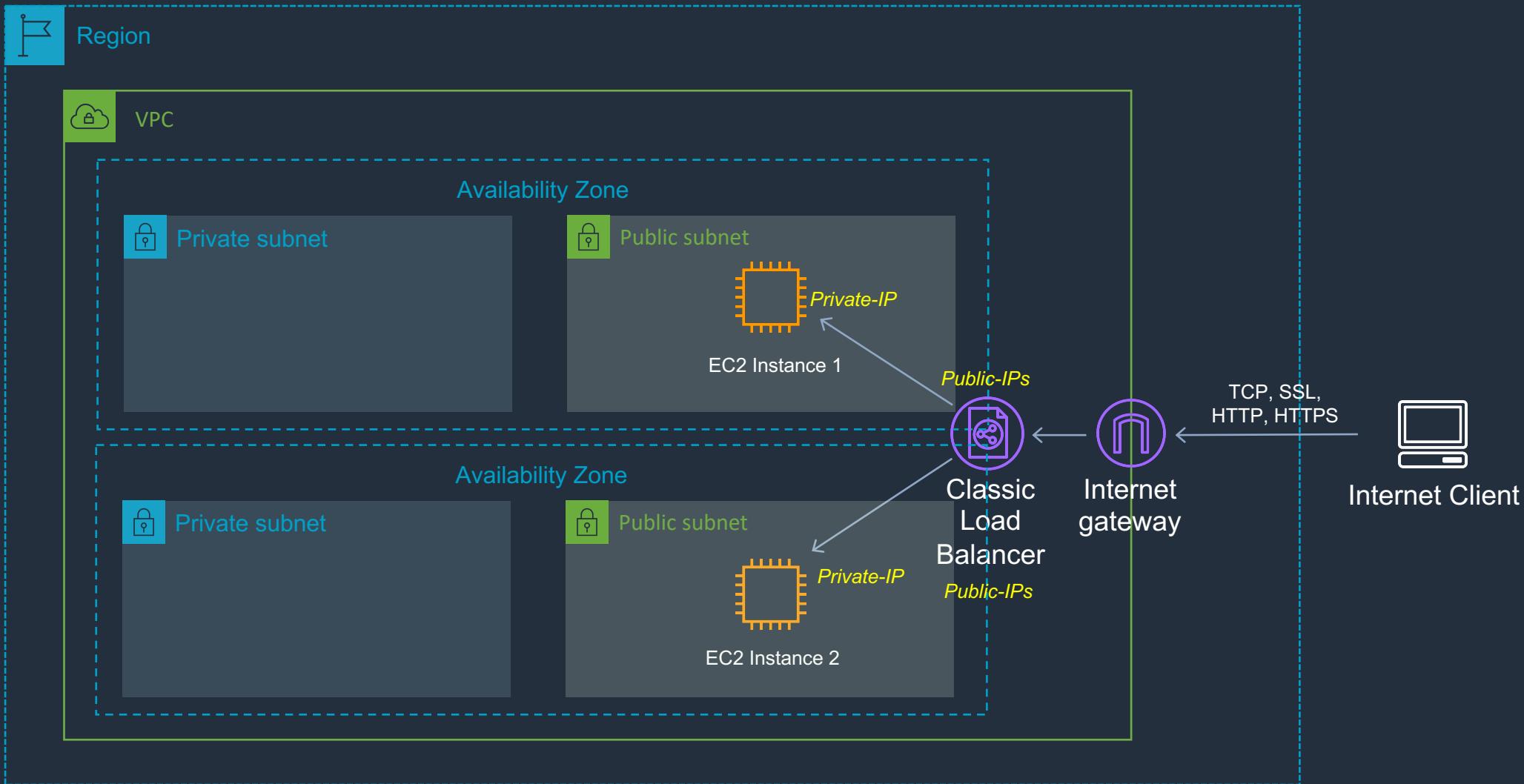
- Operates at the connection level
- Routes connections based on IP protocol data (layer 4)
- Offers ultra high performance, low latency and TLS offloading at scale
- Can have static IP / Elastic IP
- Supports UDP and static IP addresses as targets



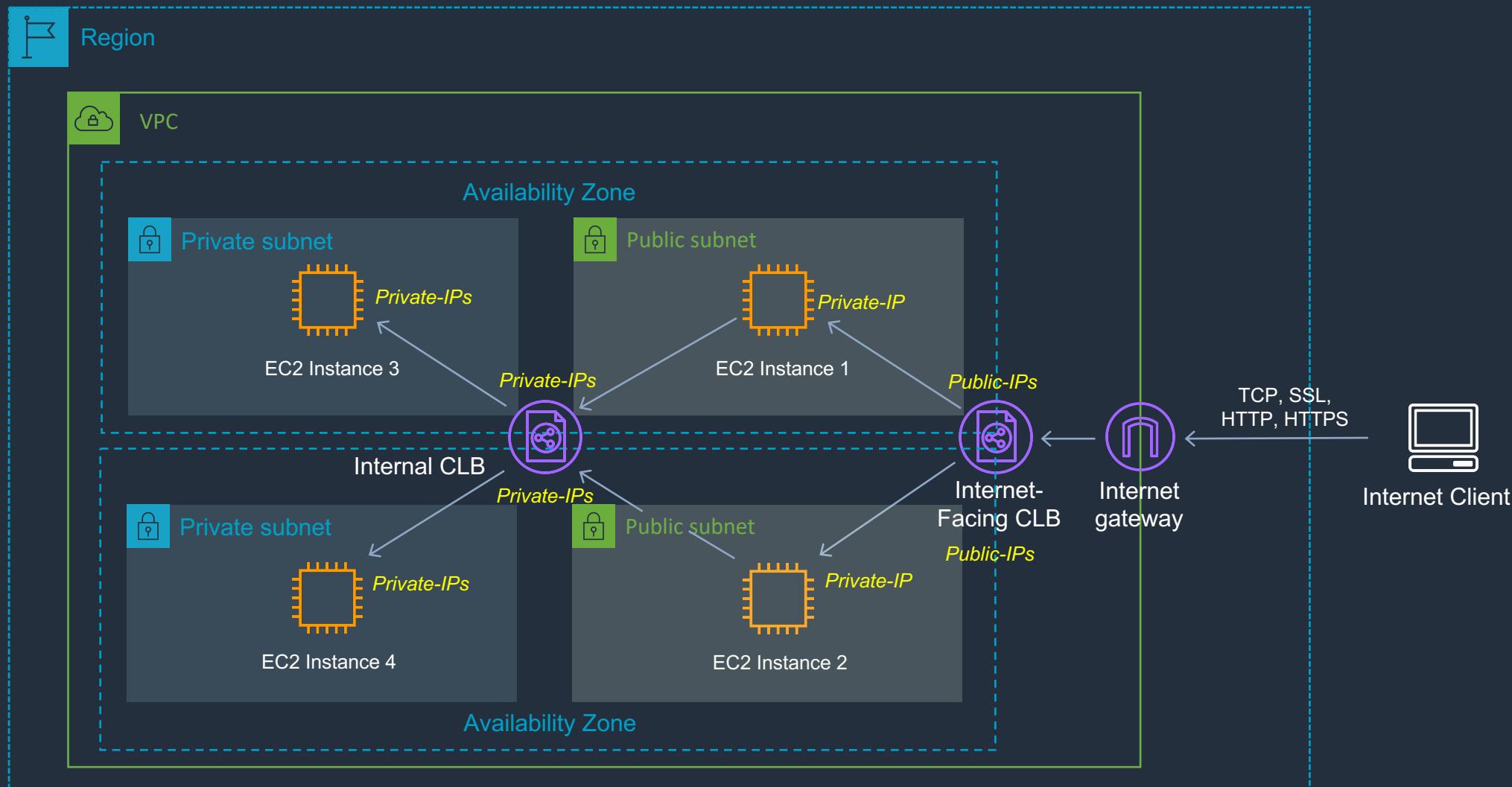
Classic Load Balancer

- Old generation; not recommended for new applications
- Performs routing at Layer 4 and Layer 7
- Use for existing applications running in EC2-Classic

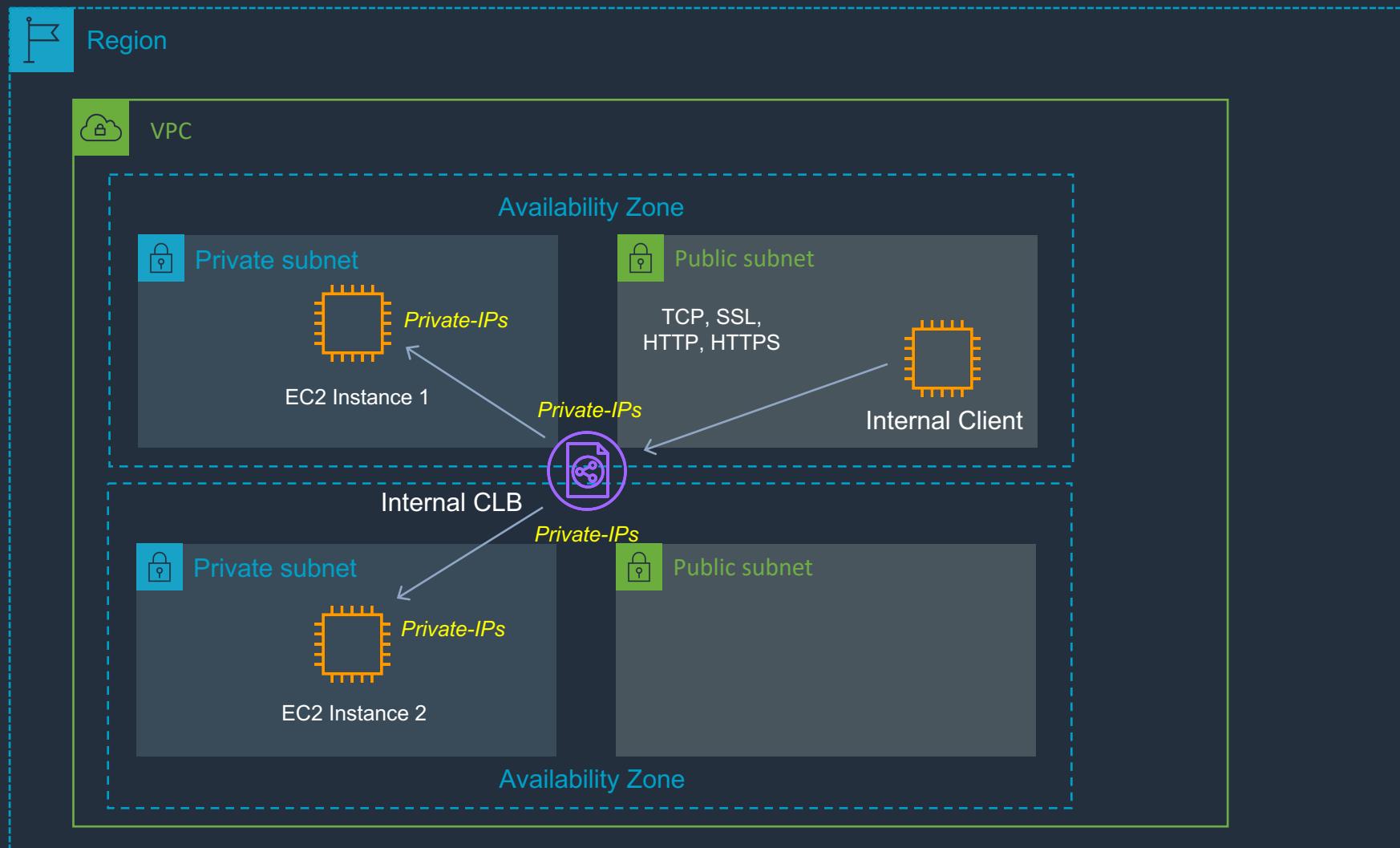
Section 4: Classic Load Balancer (Internet-Facing)



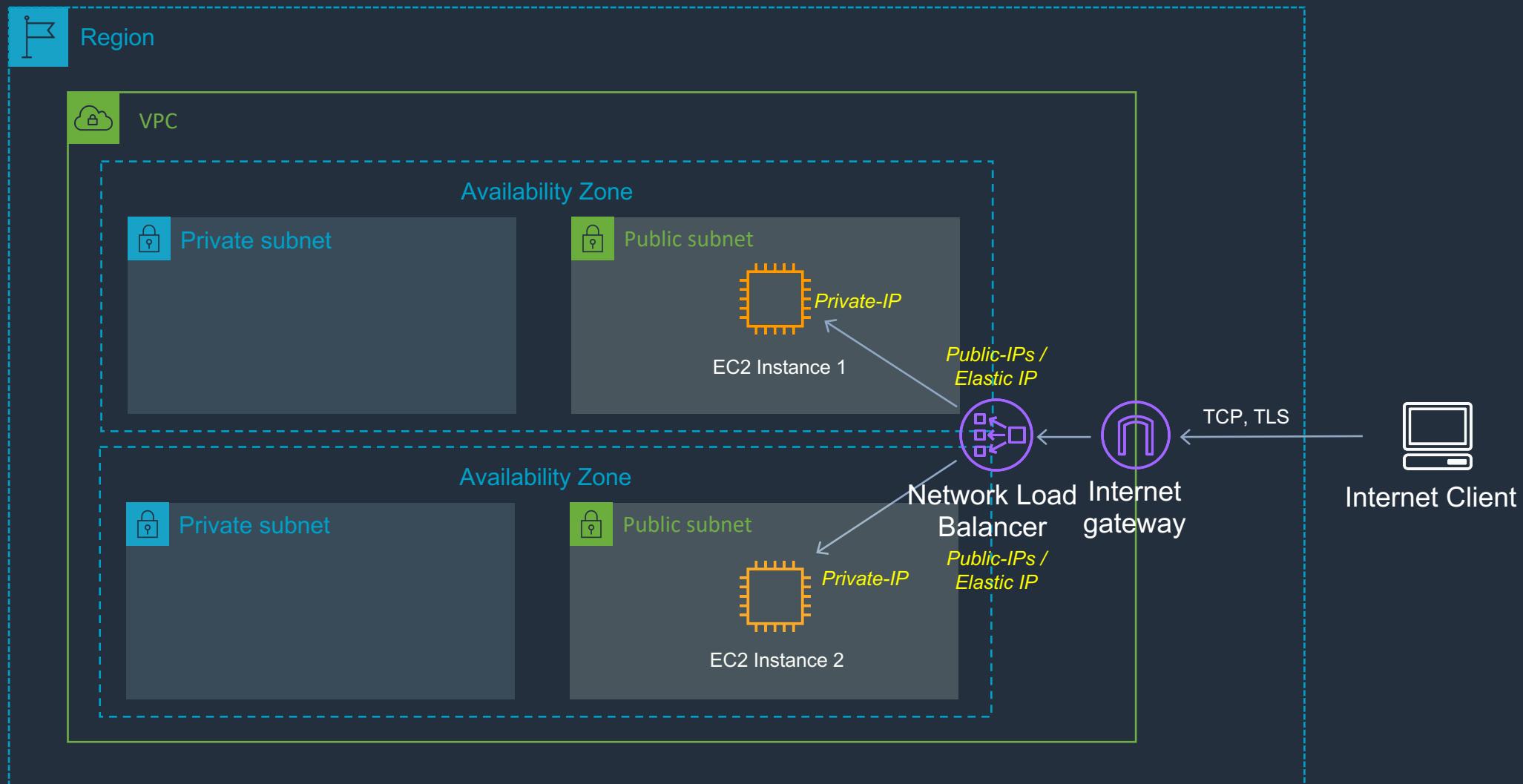
Section 4: Classic Load Balancer - Multi-tier



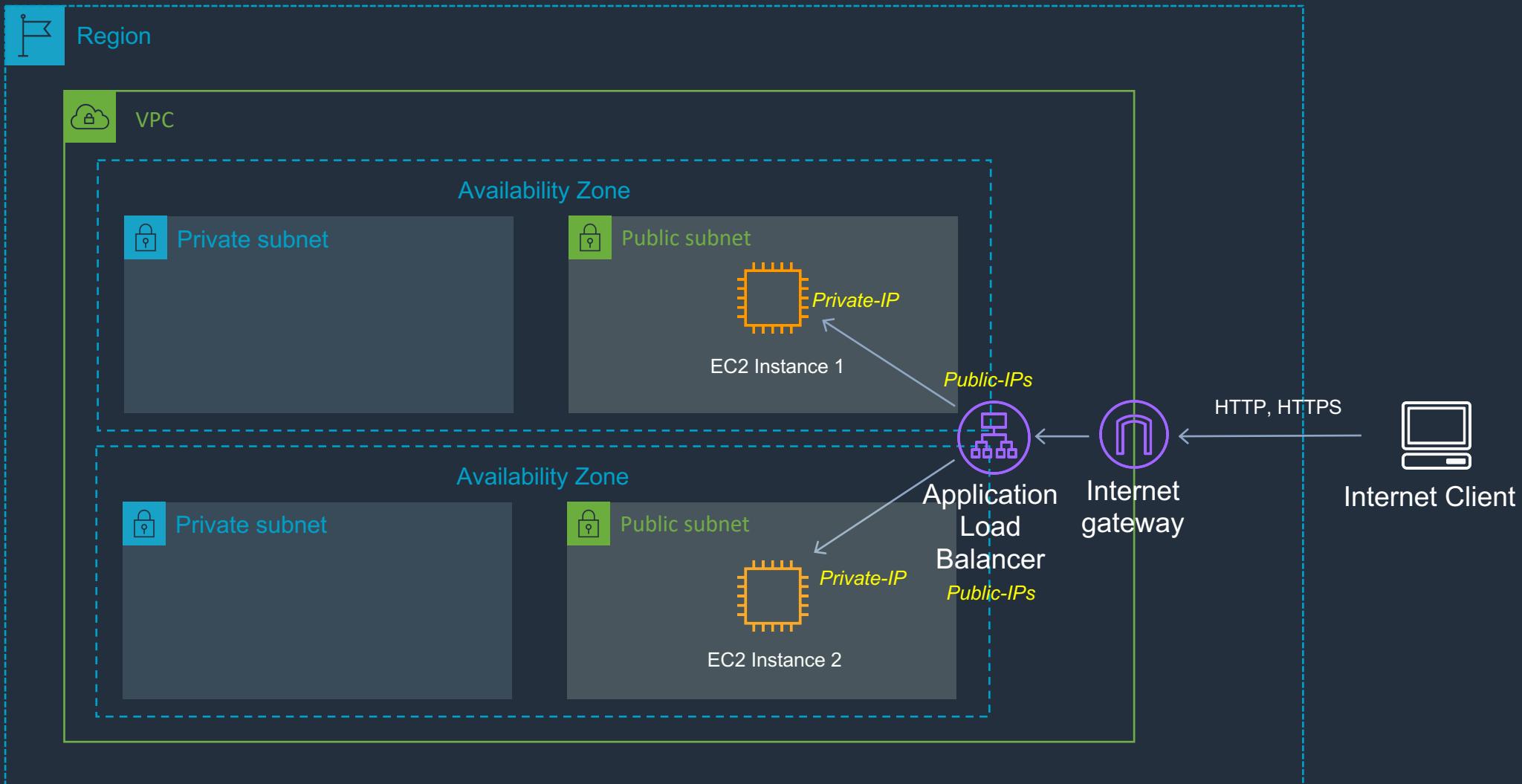
Section 4: Classic Load Balancer (Internal)



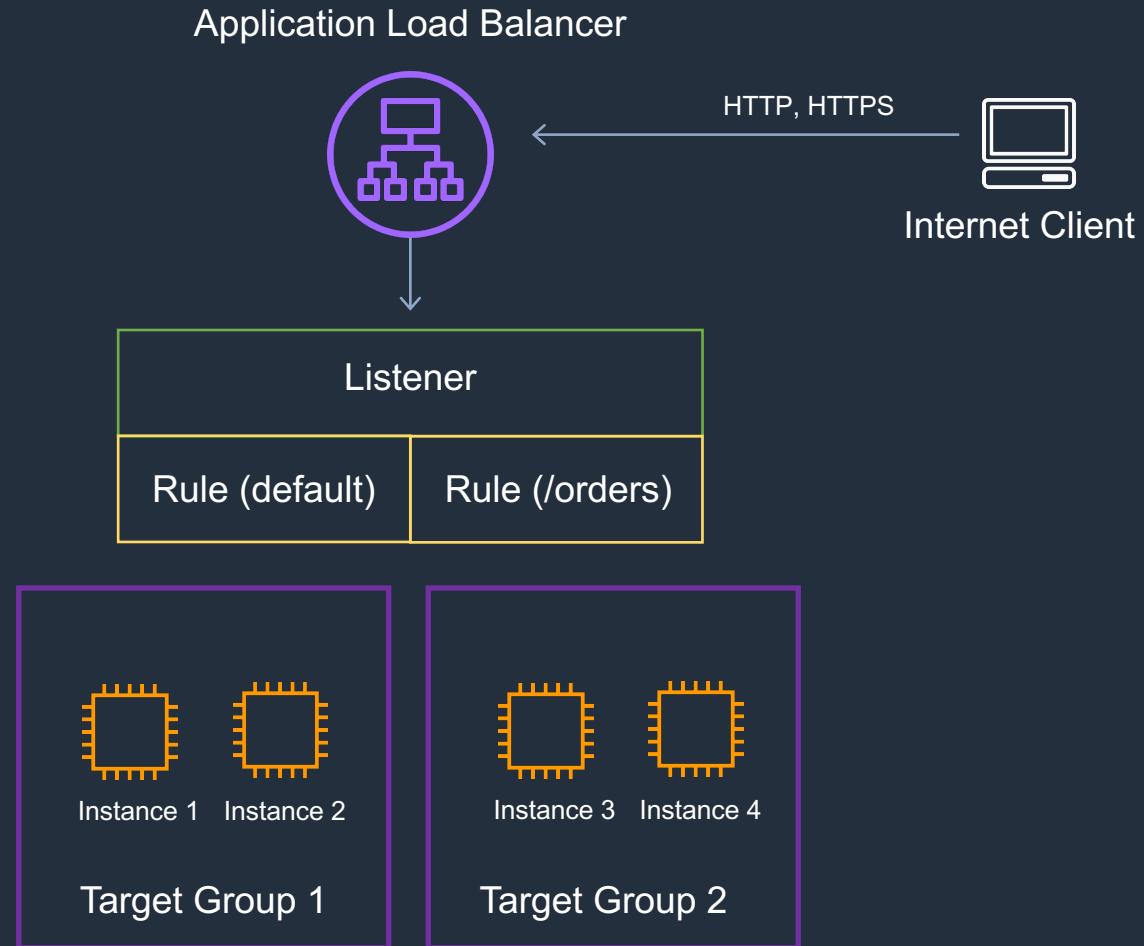
Section 4: Network Load Balancer (Internet-Facing)



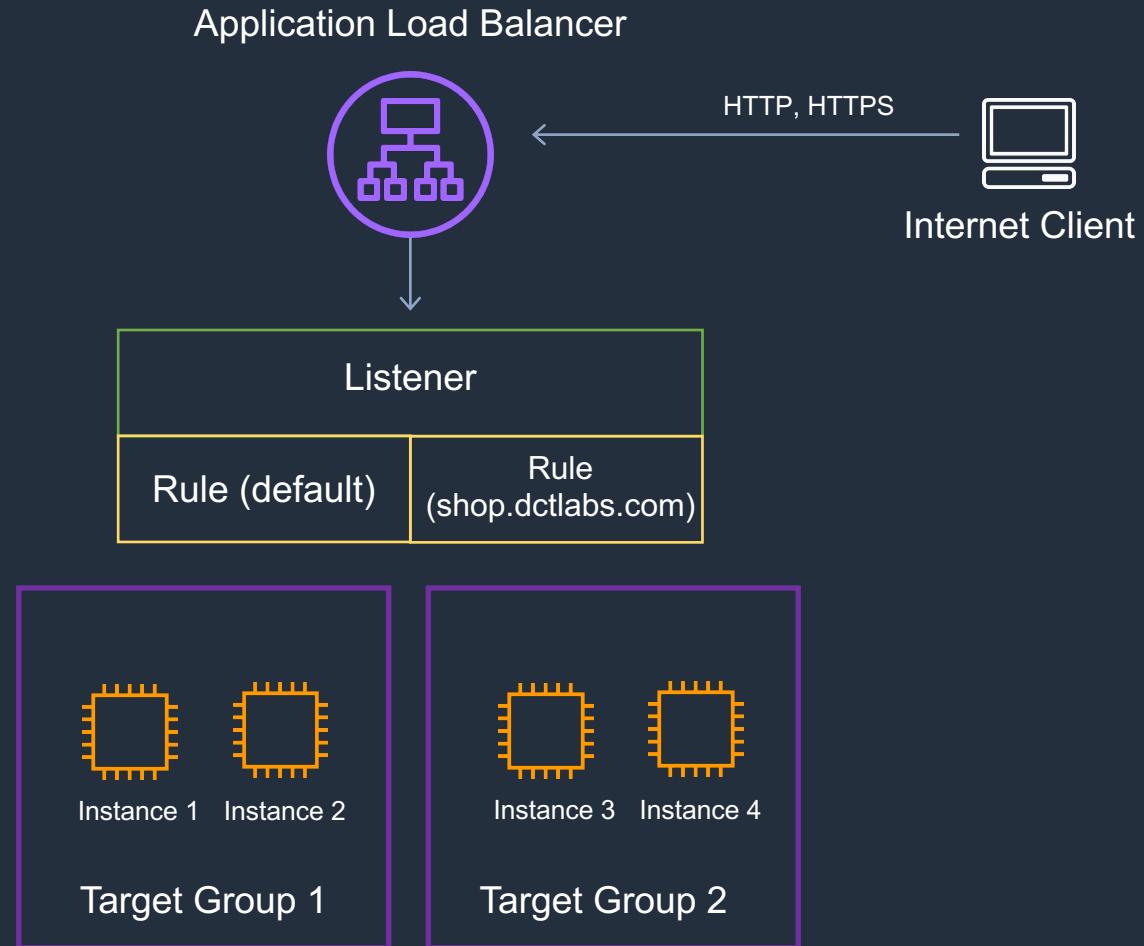
Section 4: Application Load Balancer (Internet-Facing)



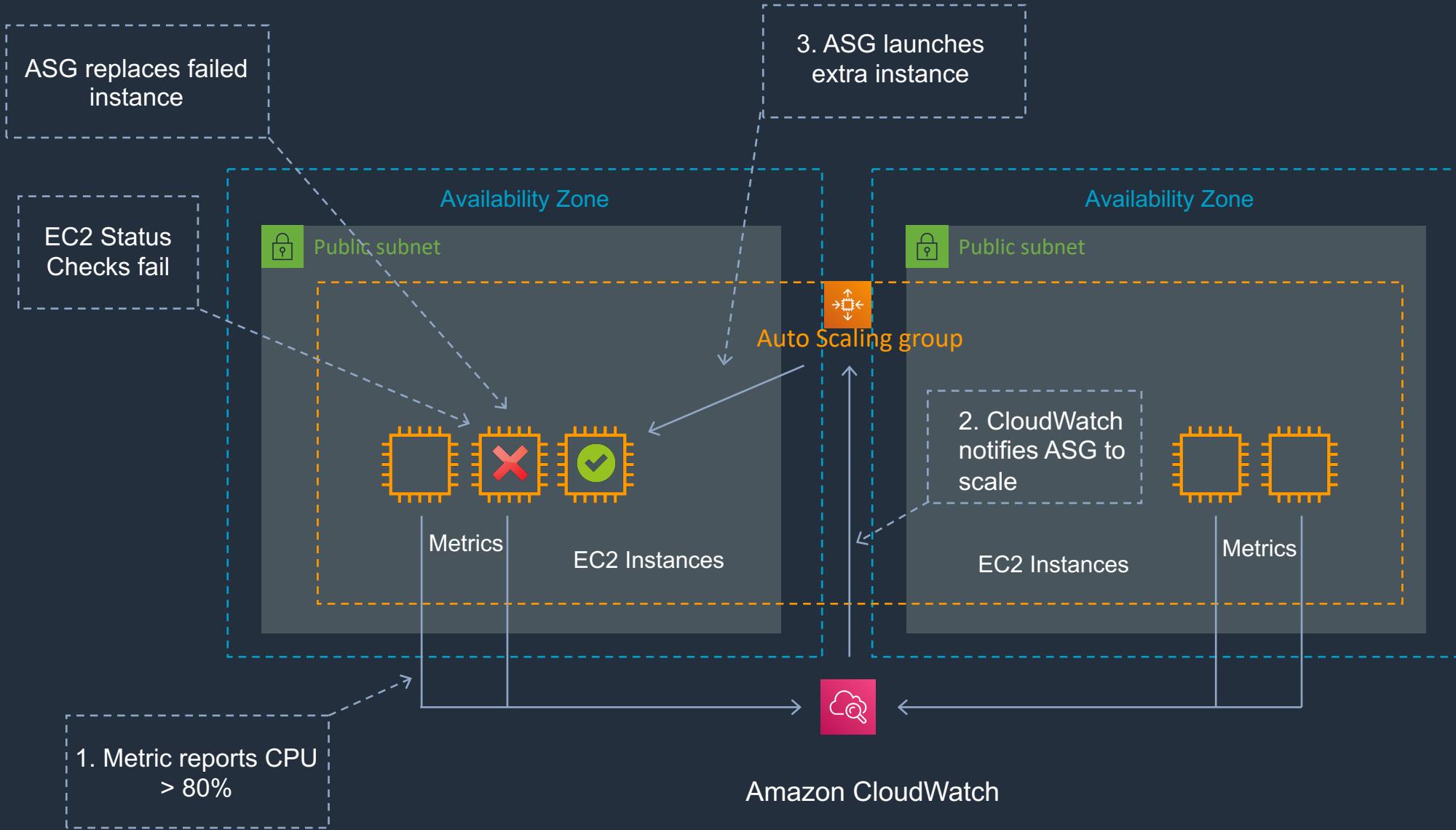
Section 4: Application Load Balancer – Path-based Routing



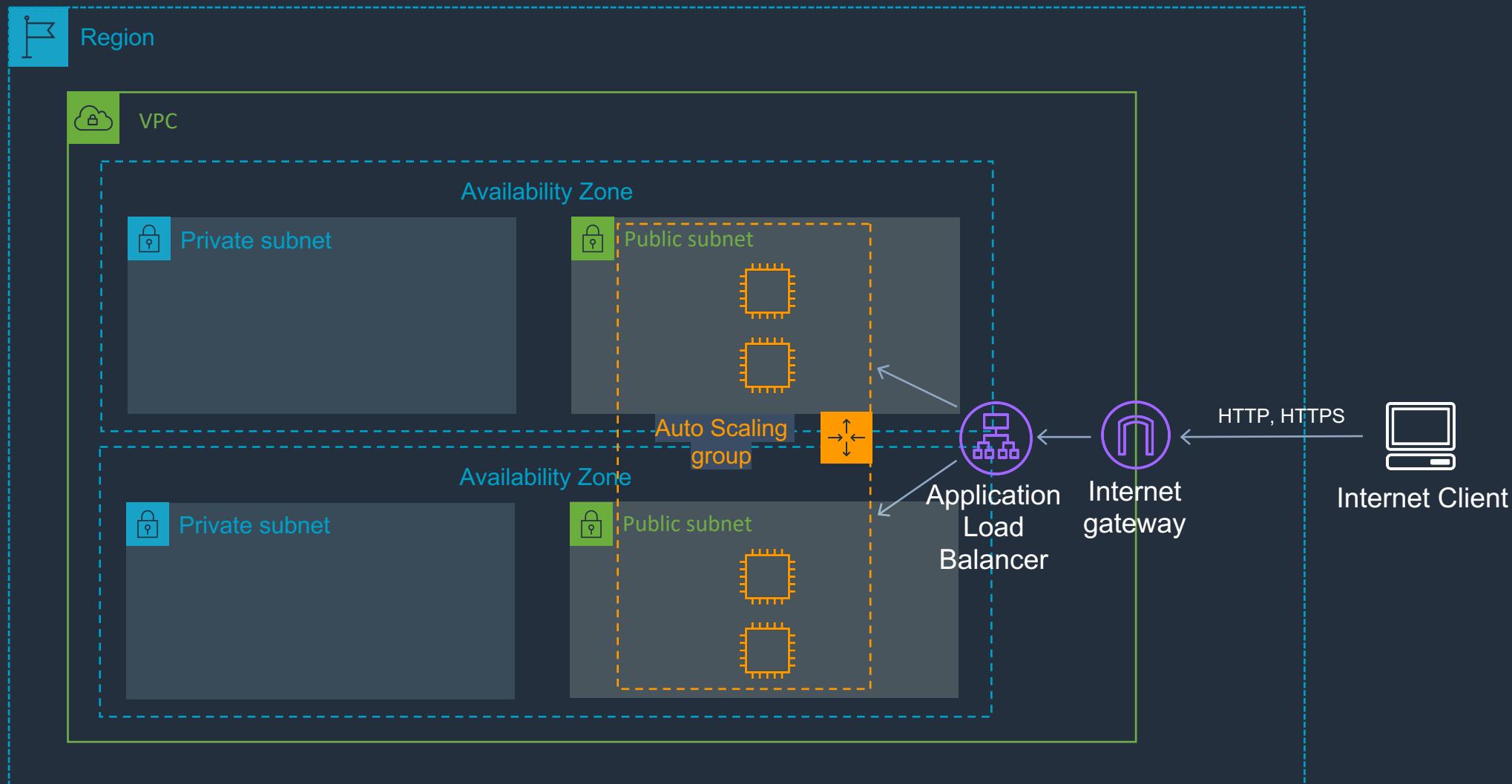
Section 4: Application Load Balancer – Host-based Routing



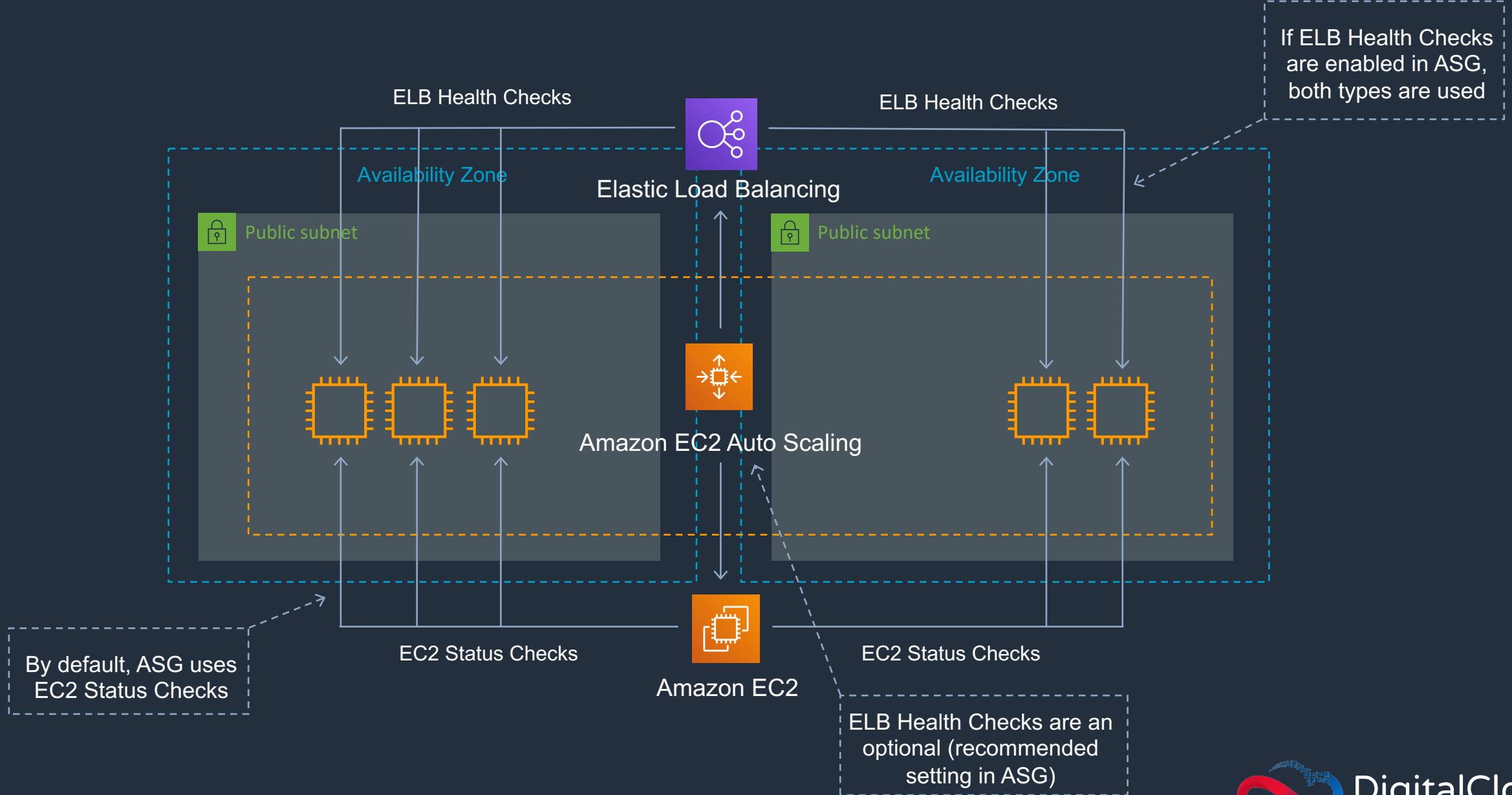
Section 4: Auto Scaling Overview



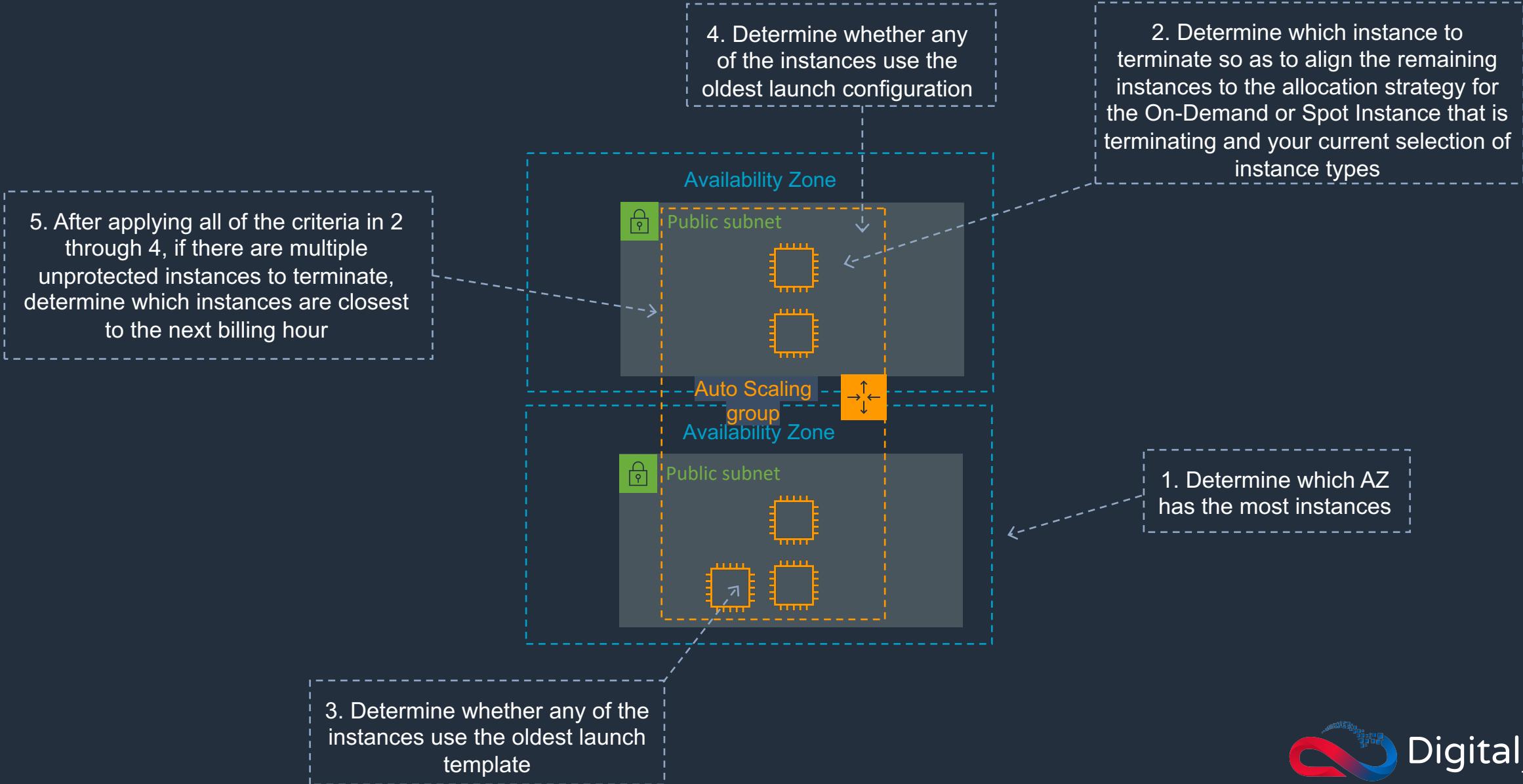
Section 4: Auto Scaling Group with ALB



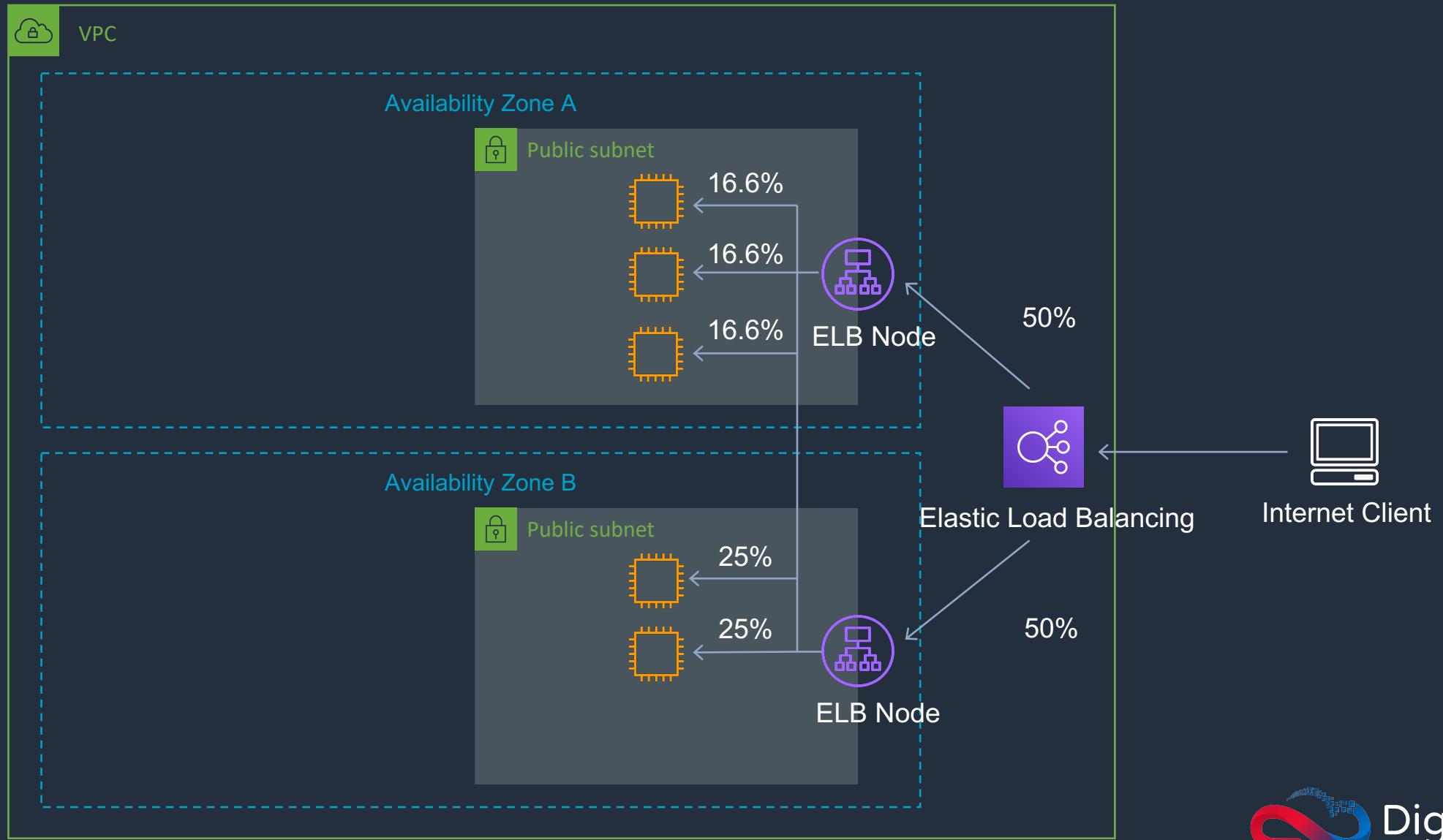
Section 4: EC2 and ELB Health Checks



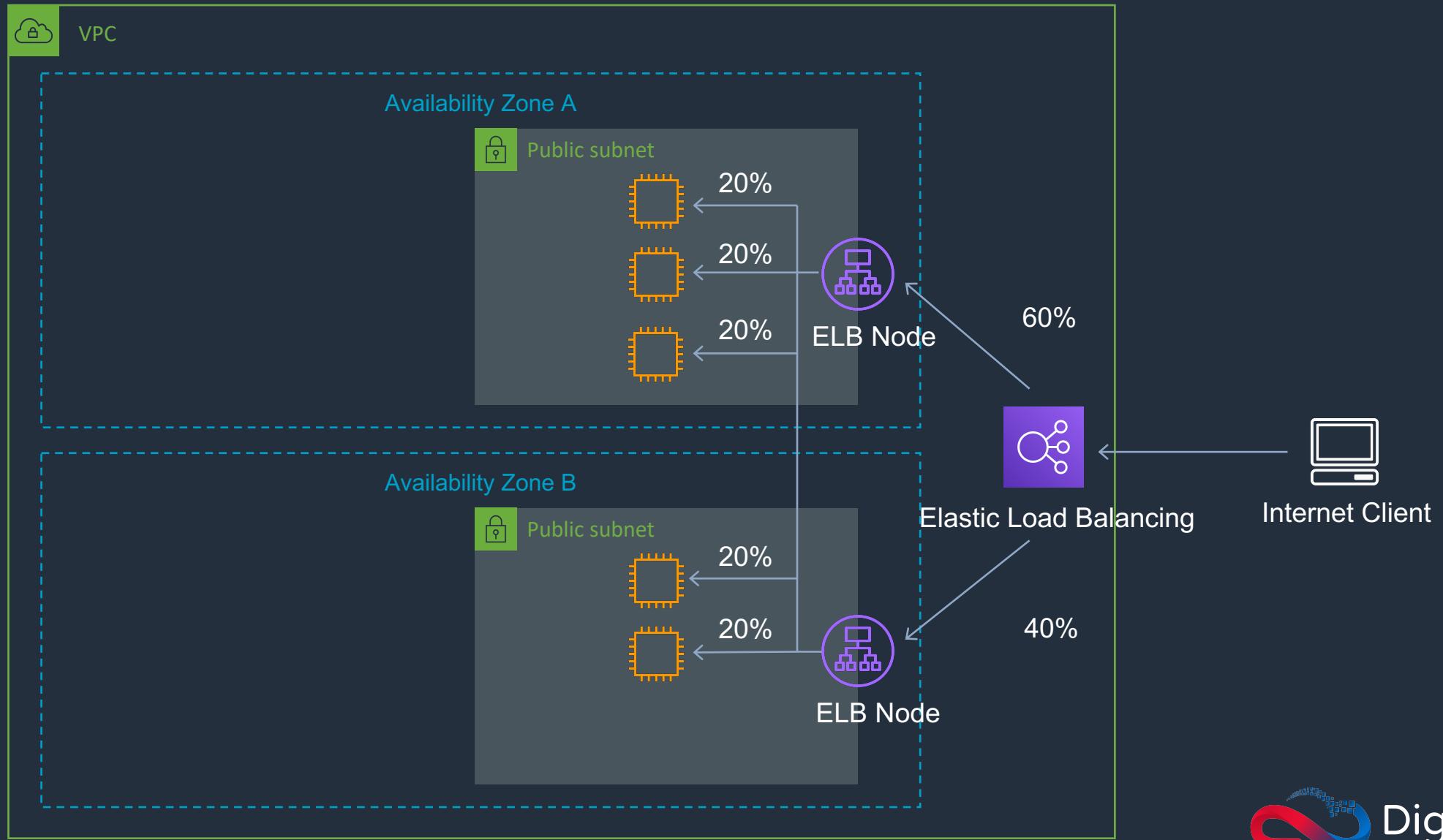
Section 4: Auto Scaling Termination Policies – Default Policies



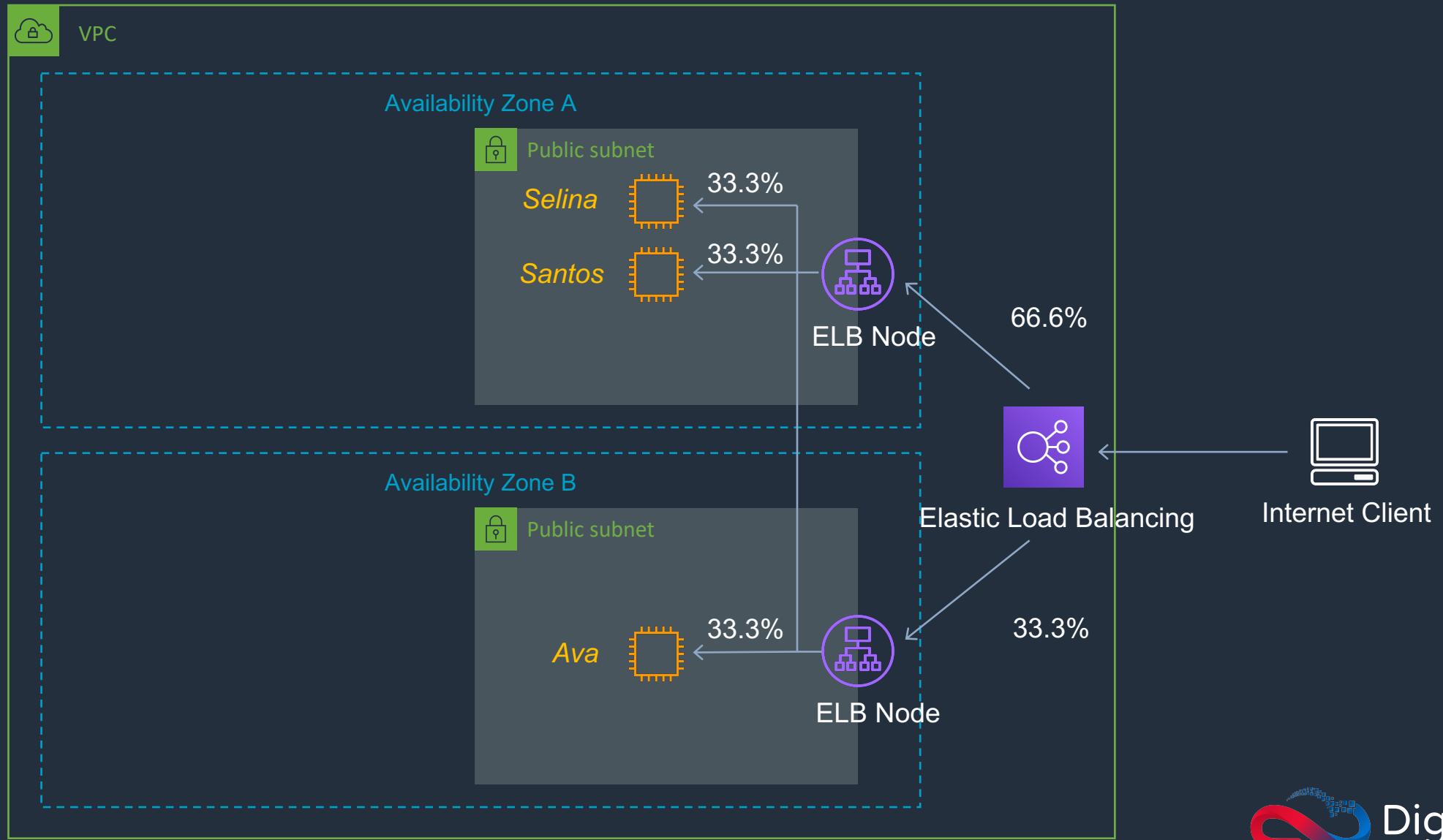
Section 4: Cross-Zone Load Balancing - Disabled



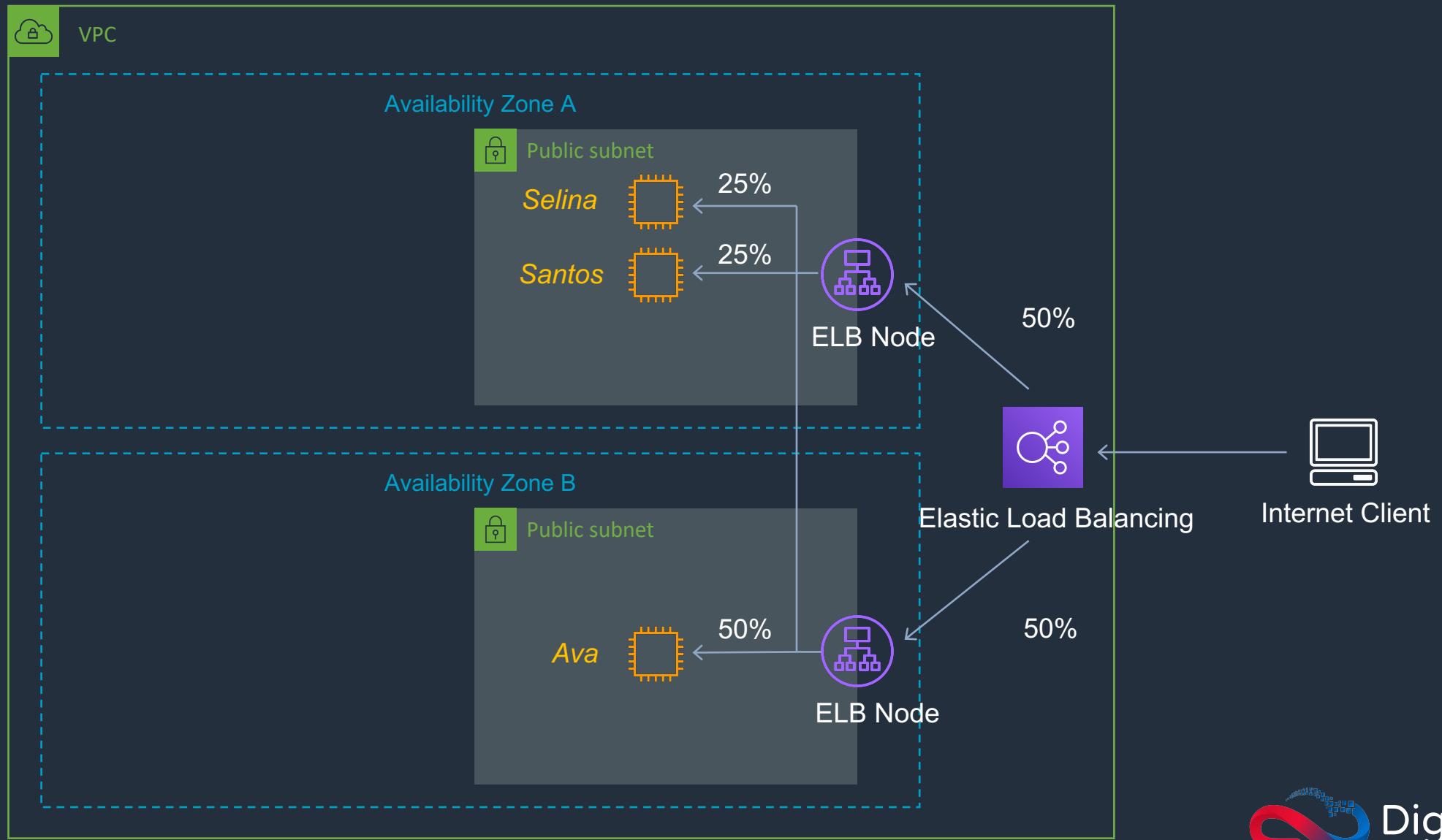
Section 4: Cross-Zone Load Balancing - Enabled



Section 4: Cross-Zone Load Balancing - Enabled



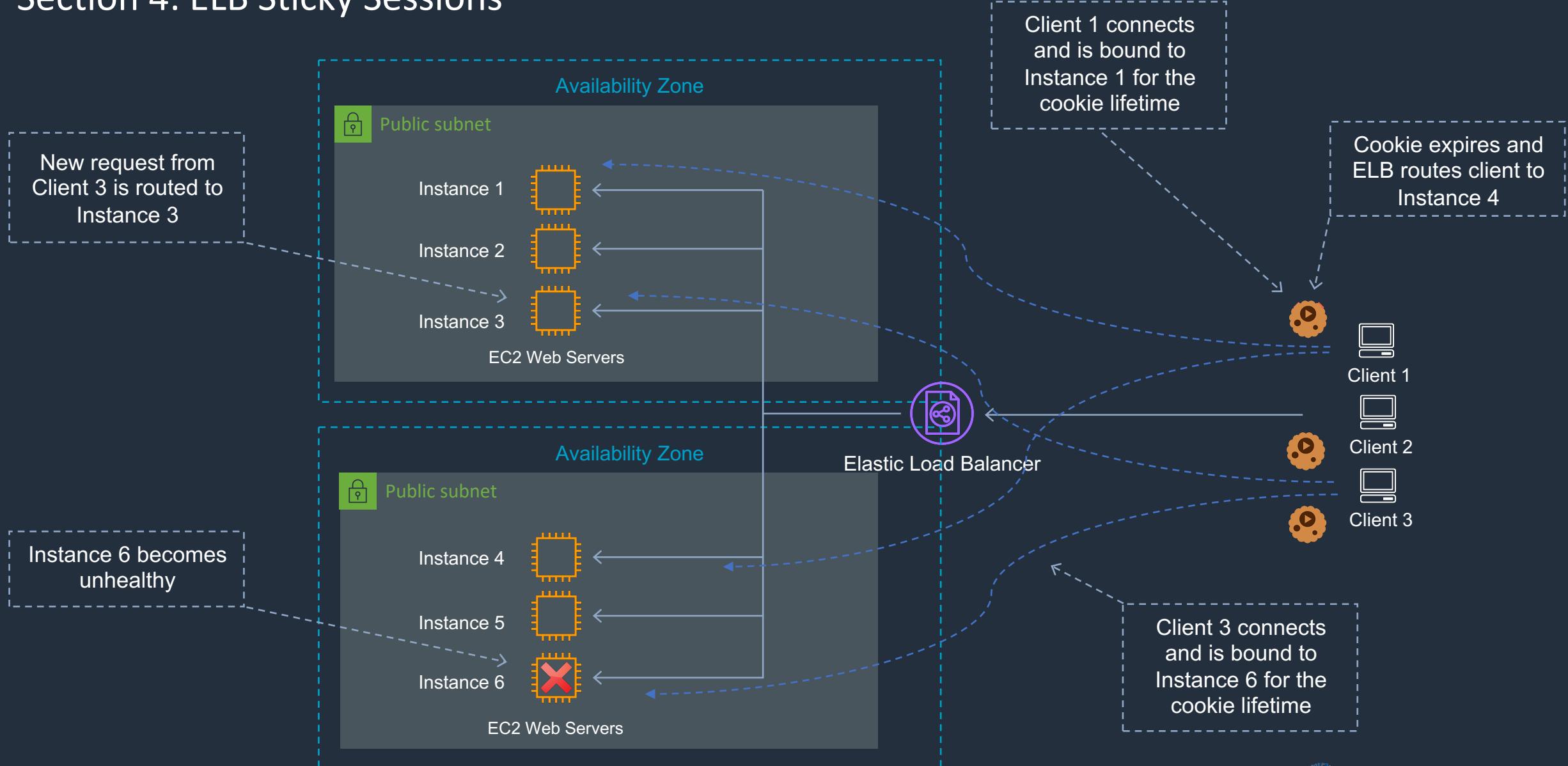
Section 4: Cross-Zone Load Balancing - Disabled



Section 4: Cross-Zone Load Balancing

Name	Created through Console	Created through CLI/API	Can be enabled/disabled?
ALB	Enabled	Enabled	No
NLB	Disabled	Disabled	Yes
CLB	Enabled	Disabled	Yes

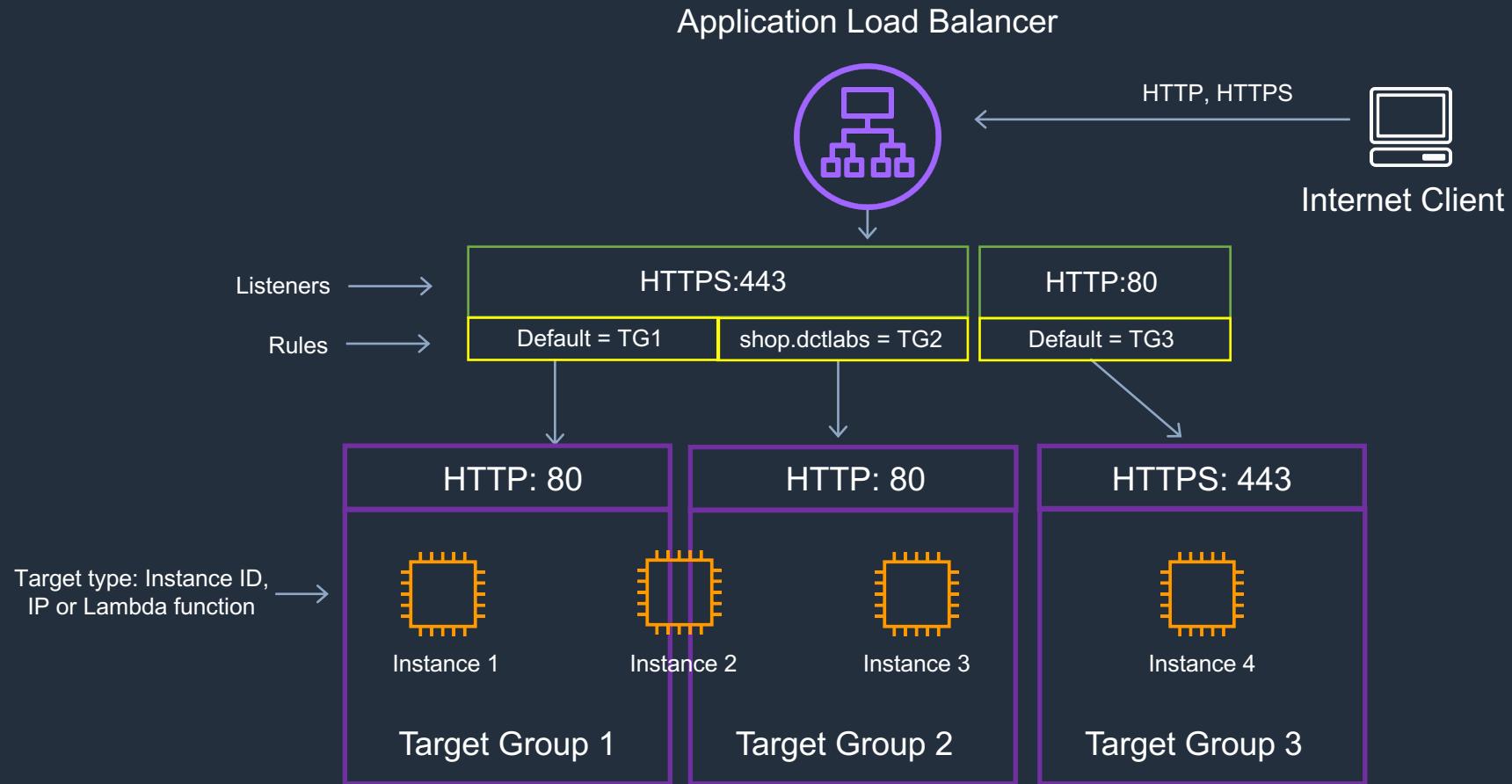
Section 4: ELB Sticky Sessions



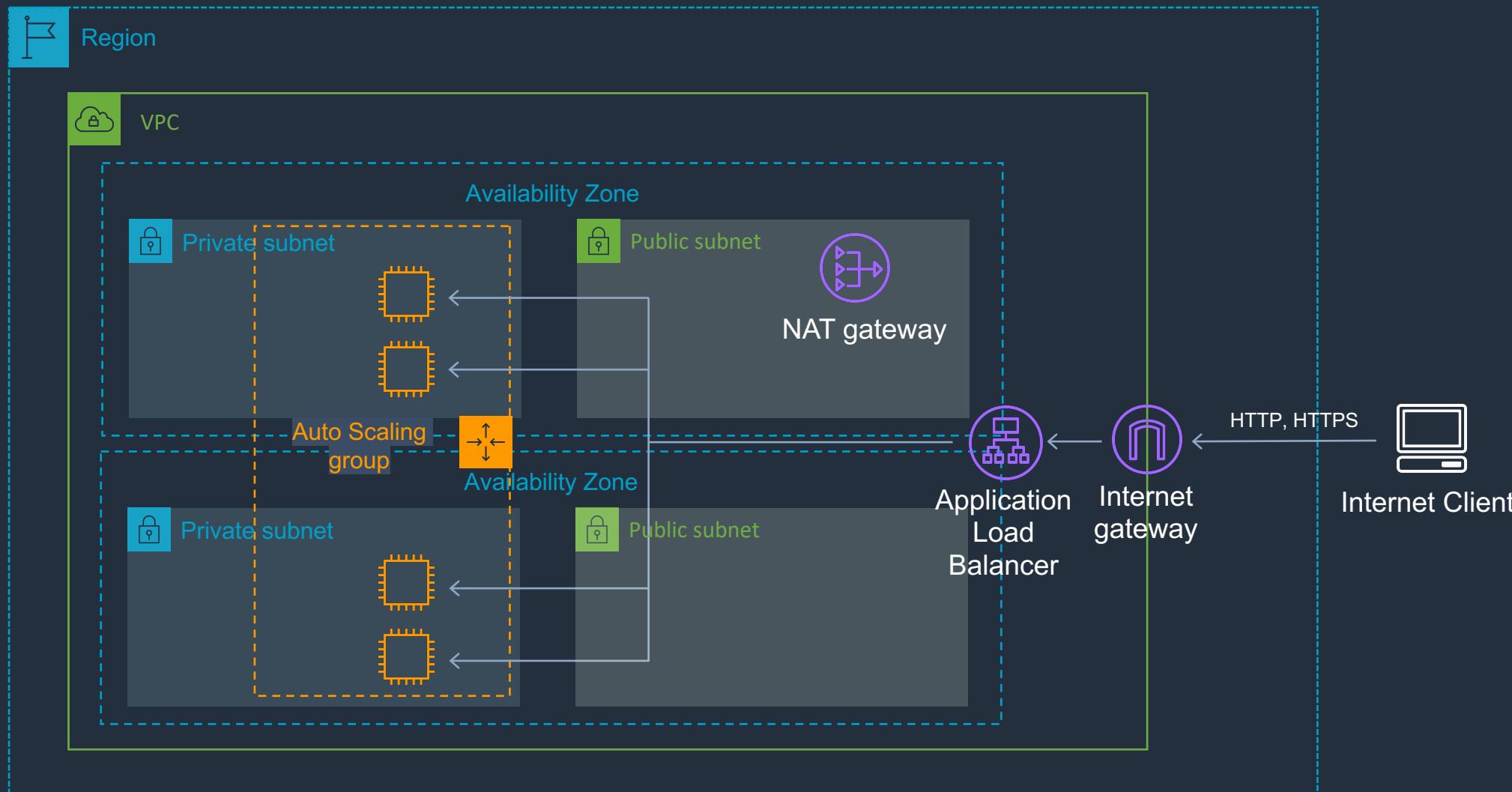
Section 4: Sticky Sessions

Name	Supported?	Load Balancer Generated Cookie	Application Generated Cookie
ALB	Yes	Yes, "AWSALB"	Not supported
NLB	No	N/A	N/A
CLB	Yes	Yes	Yes

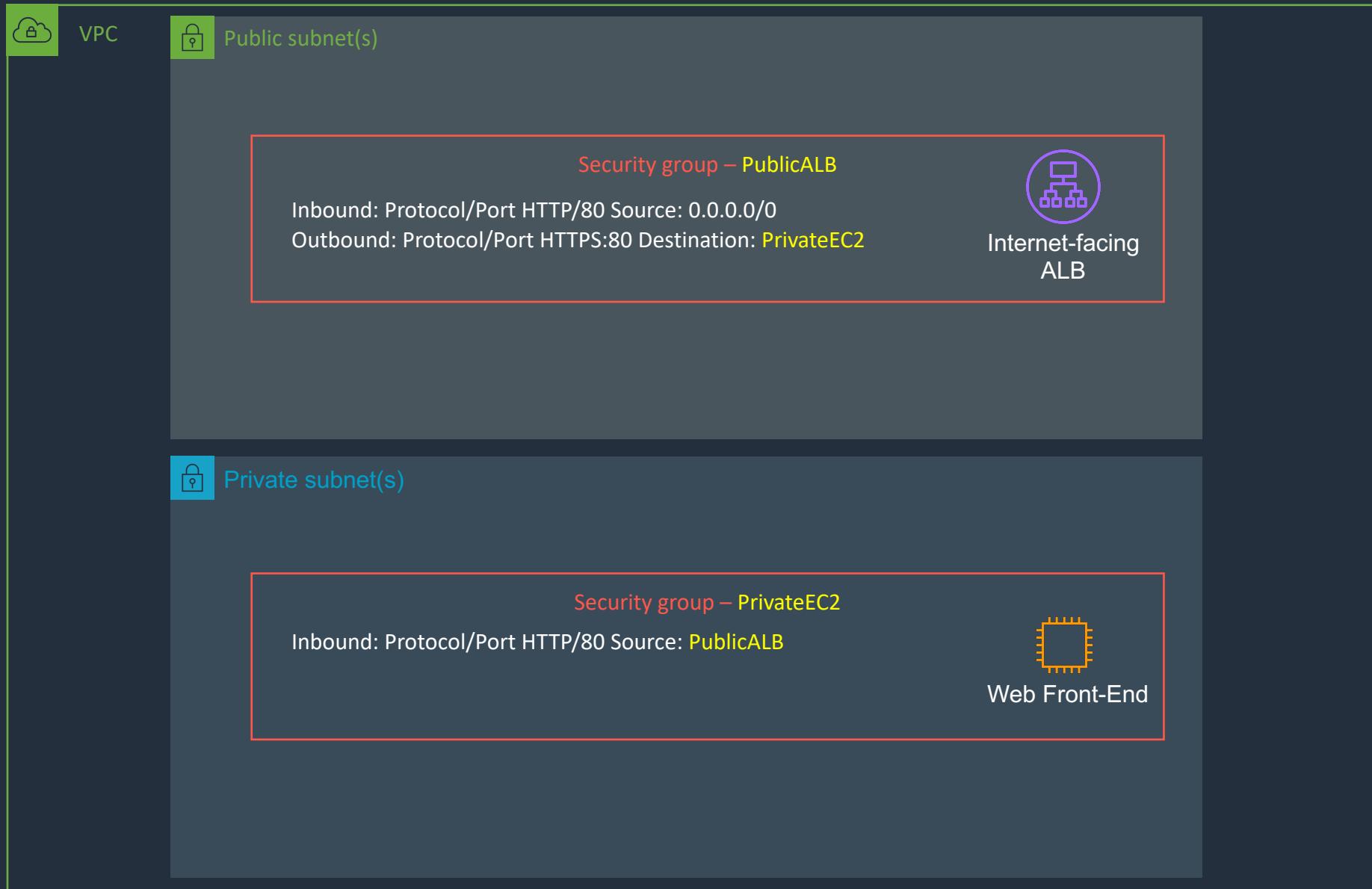
Section 4: Application Load Balancer – Listeners and SSL/TLS



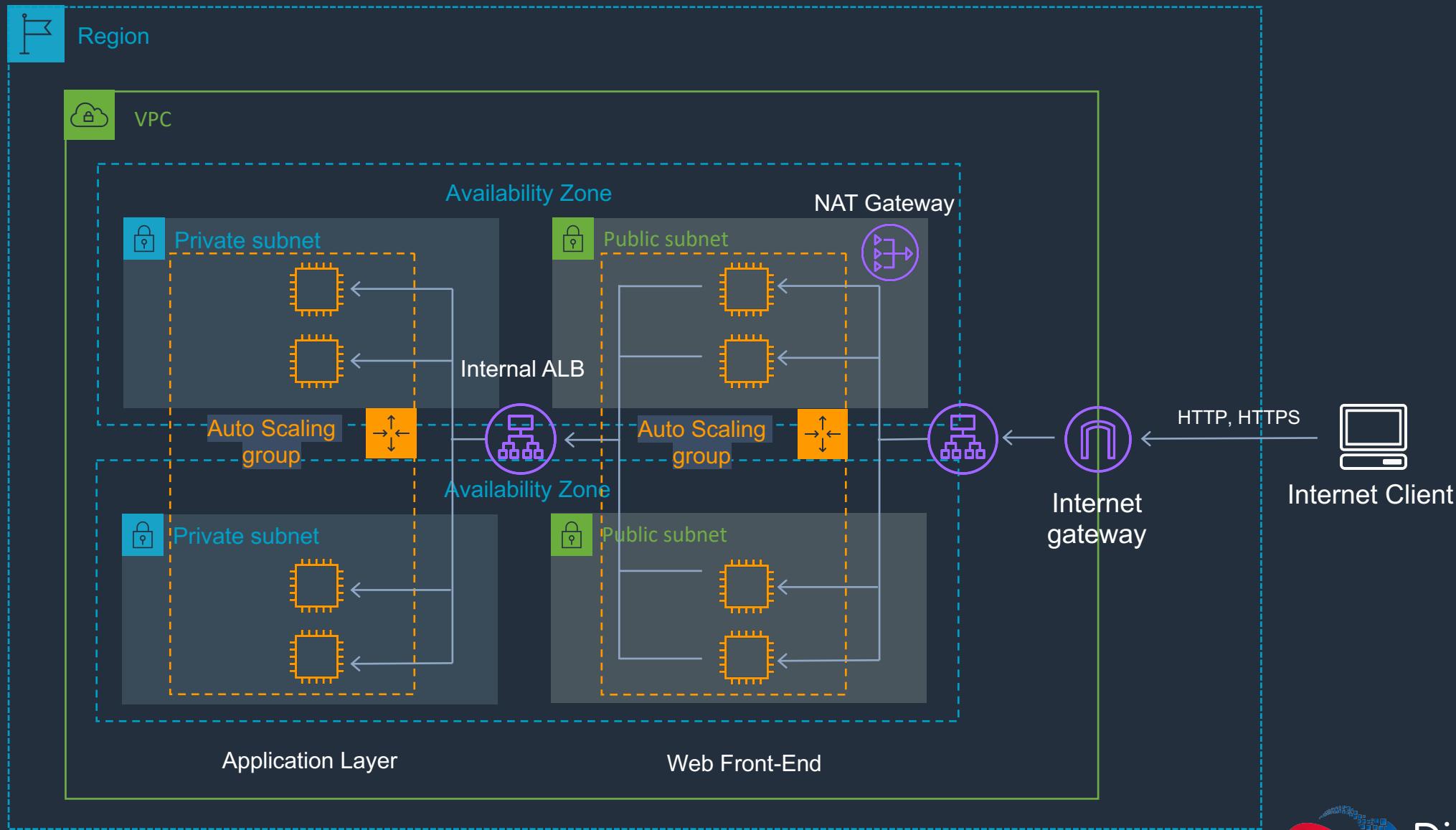
Section 4: Public ALB with Private Instances



Section 4: Public ALB with Private Instances– Security Groups



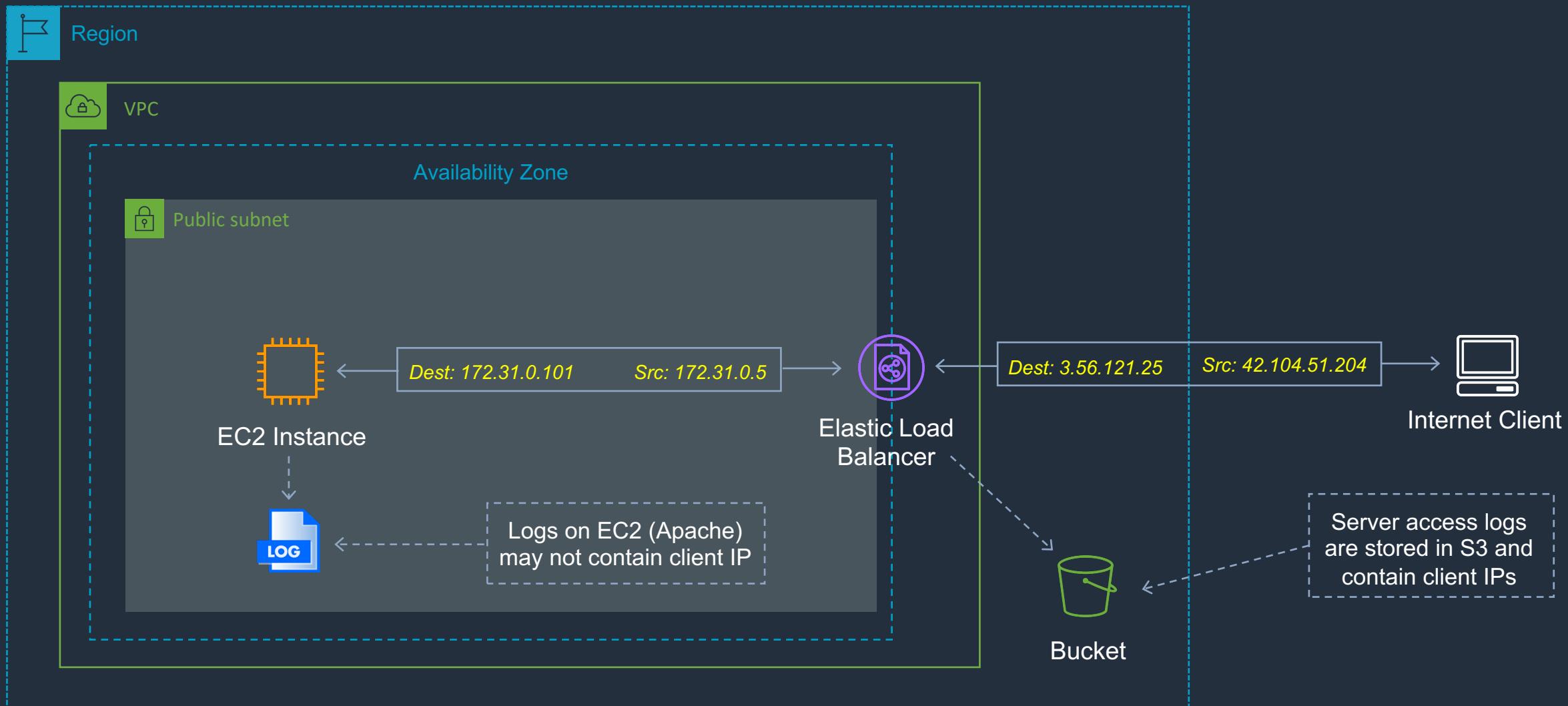
Section 4: Multi-Tier Web Architecture



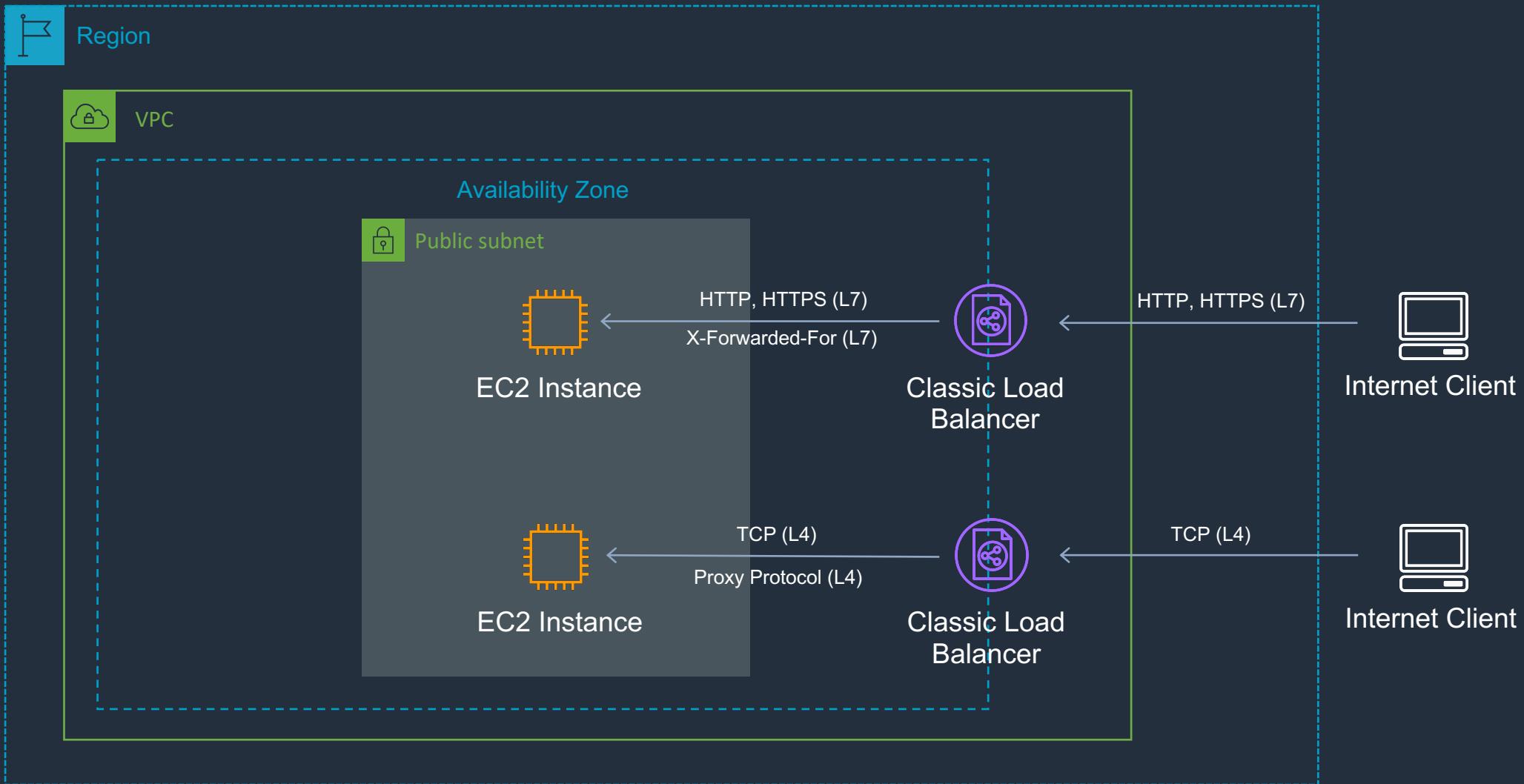
Section 4: Multi-Tier Web Architecture – Security Groups



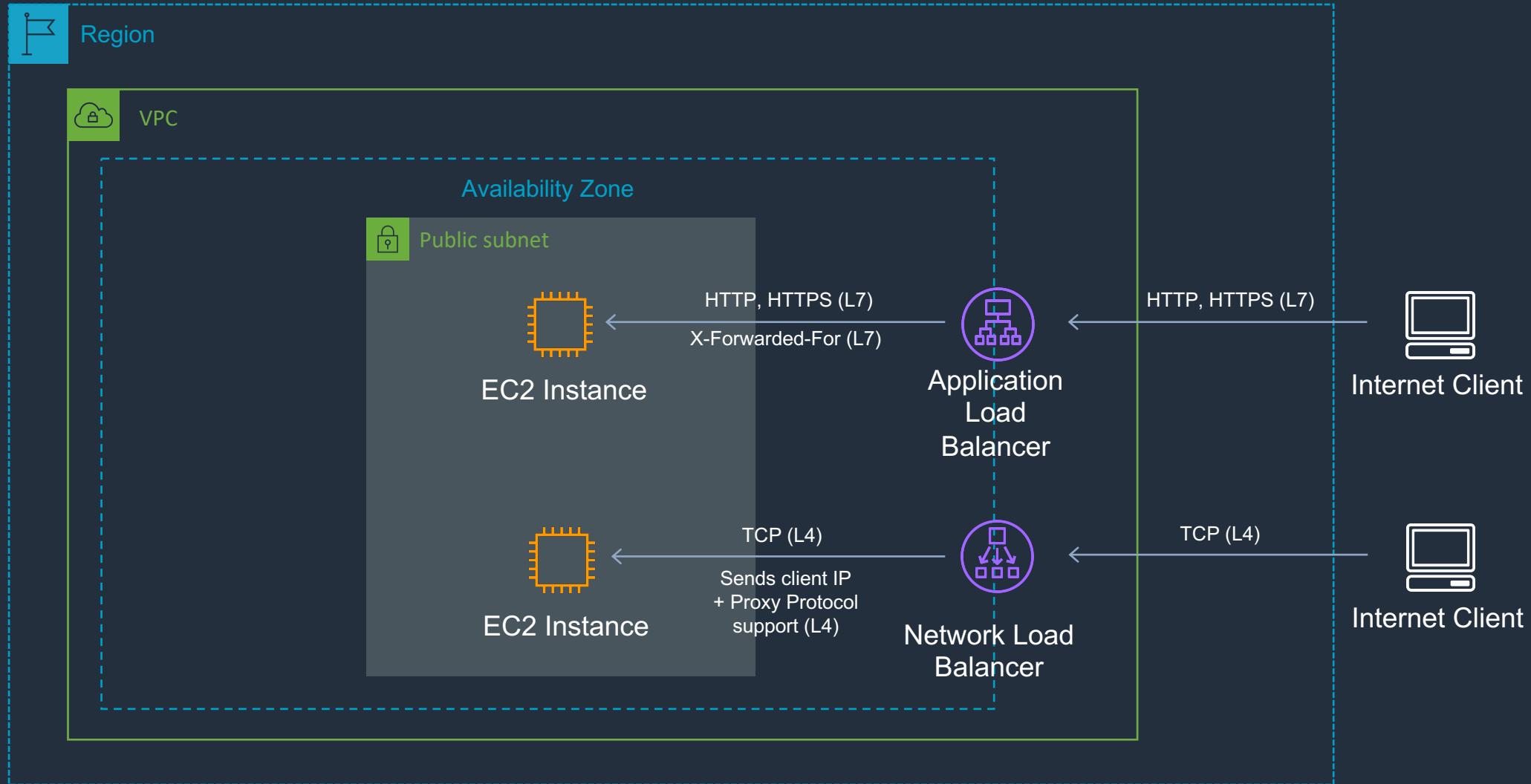
Section 4: ELB Connections and Logging



Section 4: CLB - Proxy Protocol and X-Forwarded-For



Section 4: ALB/NLB - Proxy Protocol, X-Forwarded-For and Access Logging

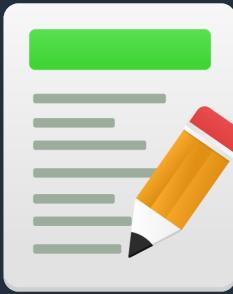


Section 4: Exam Cram

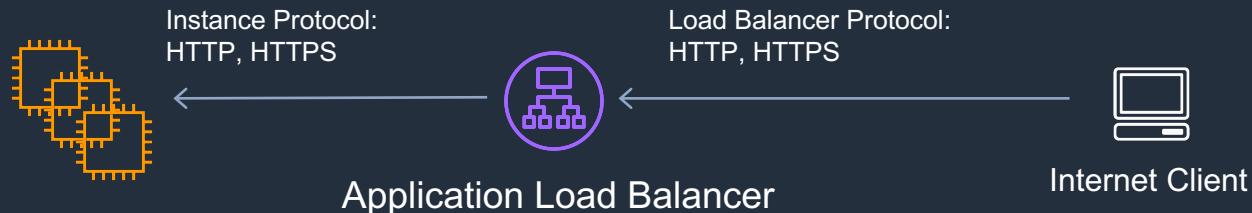
Elastic Load Balancing

There are three types of Elastic Load Balancer (ELB) on AWS:

- Classic Load Balancer (CLB) – this is the oldest of the three and provides basic load balancing at both layer 4 and layer 7.
- Application Load Balancer (ALB) – layer 7 load balancer that routes connections based on the content of the request.
- Network Load Balancer (NLB) – layer 4 load balancer that routes connections based on IP protocol data.

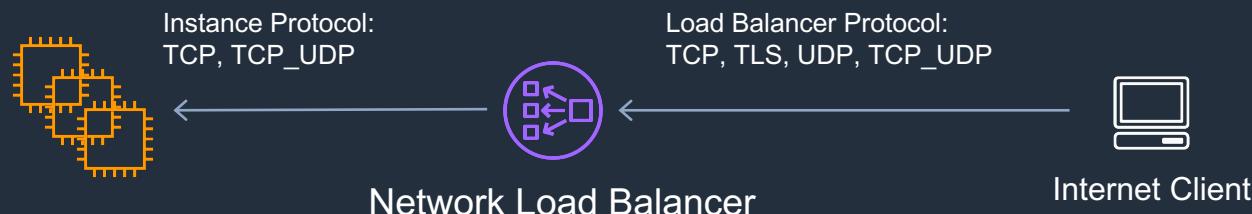


Section 4: Exam Cram



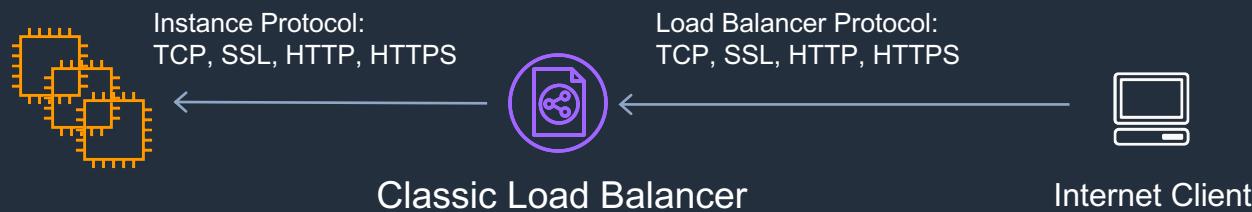
Application Load Balancer

- Operates at the request level
- Routes based on the content of the request (layer 7)
- Supports path-based routing, host-based routing, query string parameter-based routing, and source IP address-based routing
- Supports IP addresses, Lambda Functions and containers as targets



Network Load Balancer

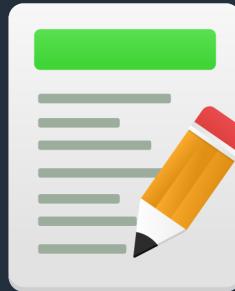
- Operates at the connection level
- Routes connections based on IP protocol data (layer 4)
- Offers ultra high performance, low latency and TLS offloading at scale
- Can have static IP / Elastic IP
- Supports UDP and static IP addresses as targets



Classic Load Balancer

- Old generation; not recommended for new applications
- Performs routing at Layer 4 and Layer 7
- Use for existing applications running in EC2-Classic

Section 4: Exam Cram



Elastic Load Balancing

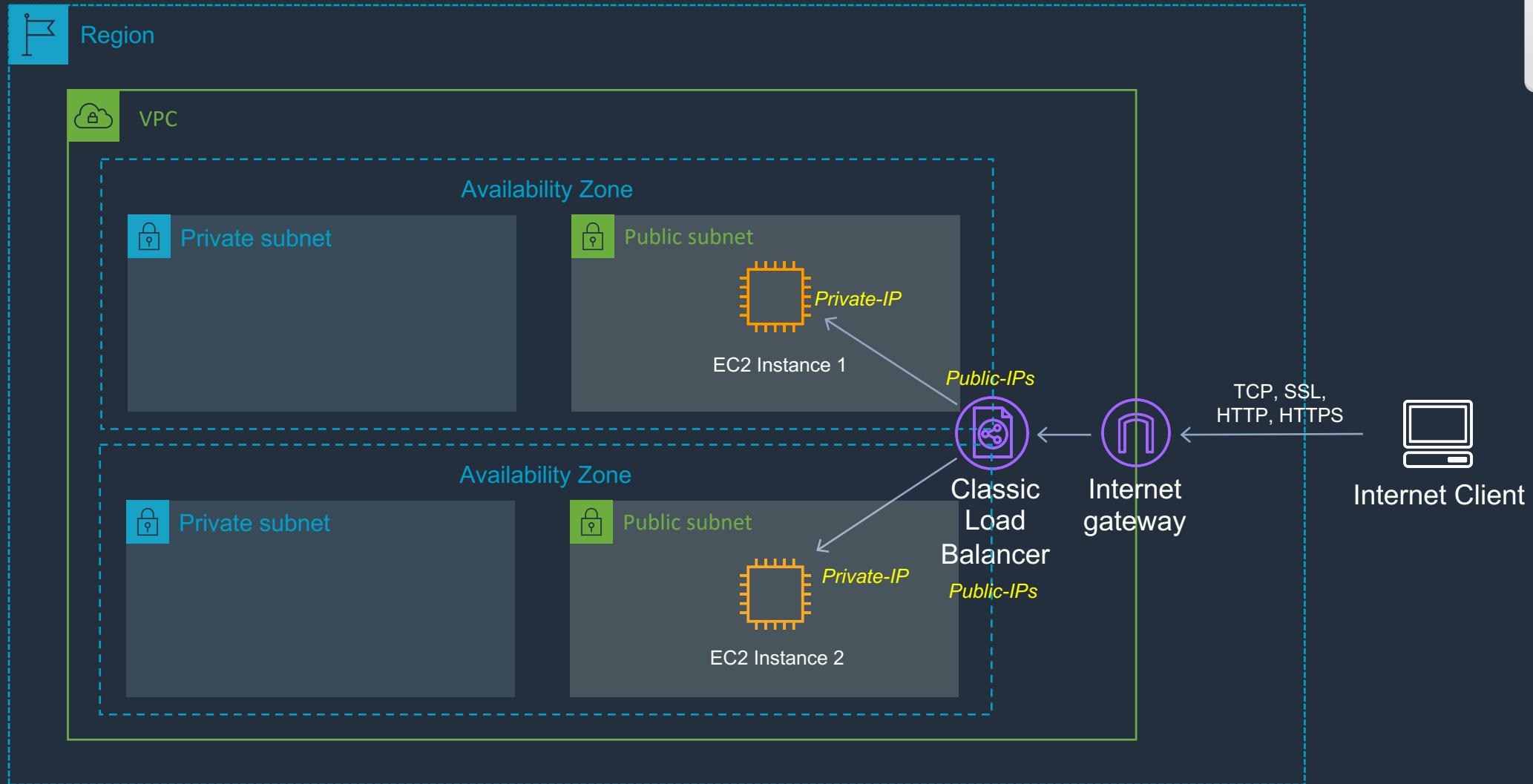
Internet facing ELB:

- ELB nodes have public IPs.
- Routes traffic to the private IP addresses of the EC2 instances.
- Need one public subnet in each AZ where the ELB is defined.
- ELB DNS name format: <name>-<id-number>.<region>.elb.amazonaws.com.

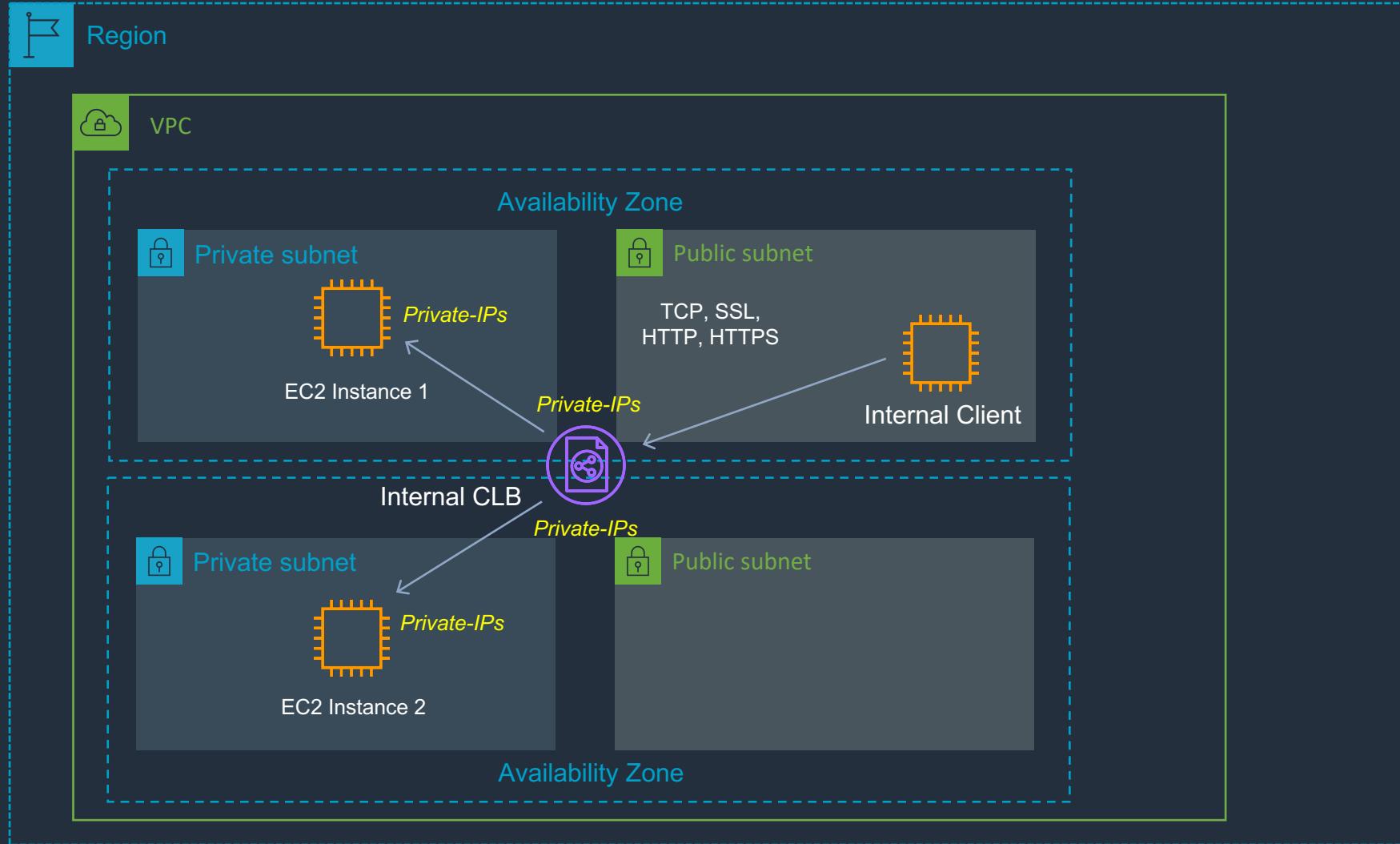
Internal only ELB:

- ELB nodes have private IPs.
- Routes traffic to the private IP addresses of the EC2 instances.
- ELB DNS name format: internal-<name>-<id-number>.<region>.elb.amazonaws.com.

Section 4: Classic Load Balancer (Internet-Facing)



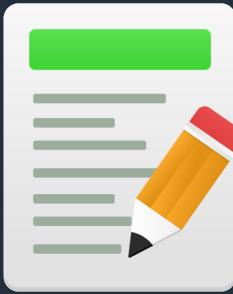
Section 4: Classic Load Balancer (Internal)



Section 4: Exam Cram

Elastic Load Balancing

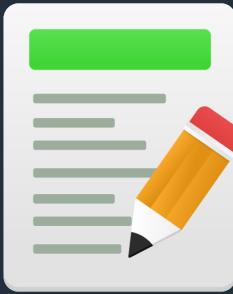
- EC2 instances and containers can be registered against an ELB.
- ELB nodes use IP addresses within your subnets, ensure at least a /27 subnet and make sure there are at least 8 IP addresses available in order for the ELB to scale.
- An ELB forwards traffic to eth0 (primary IP address).
- An ELB listener is the process that checks for connection requests:
 - Listeners for CLB provide options for TCP and HTTP/HTTPS.
 - Listeners for ALB only provide options for HTTP and HTTPS.
 - Listeners for NLB only provide TCP as an option.



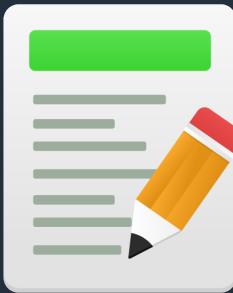
Section 4: Exam Cram

ELB Security Groups

- Security groups control the ports and protocols that can reach the front-end listener.
- You must assign a security group for the ports and protocols on the front-end listener.
- You need to also allow the ports and protocols for the health check ports and back-end listeners.



Section 4: Exam Cram



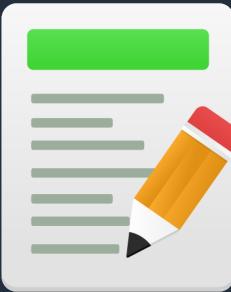
ELB Monitoring

- CloudWatch – every 1 minute:
 - ELB service only sends information when requests are active.
- Access Logs:
 - Disabled by default.
 - Includes information about the clients (not included in CloudWatch metrics).
 - Can identify requester, IP, request type etc.
 - Can be optionally stored and retained in S3.
- CloudTrail:
 - Can be used to capture API calls to the ELB.
 - Can be stored in an S3 bucket.

Section 4: Exam Cram

EC2 Auto Scaling

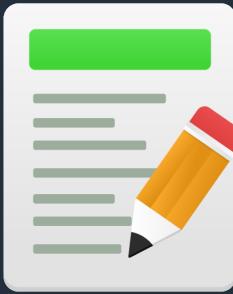
- Amazon EC2 Auto Scaling dynamically and automatically ensures the correct number of instances for an application.
- You create collections of EC2 instances, called Auto Scaling groups.
- Automatically provides horizontal scaling (scale-out) for your instances.
- Triggered by an event of scaling action to either launch or terminate instances.
- Auto Scaling is a region-specific service.
- Auto Scaling can span multiple AZs within the same AWS region.
- There is no additional cost for Auto Scaling, you just pay for the resources (EC2 instances) provisioned.



Section 4: Exam Cram

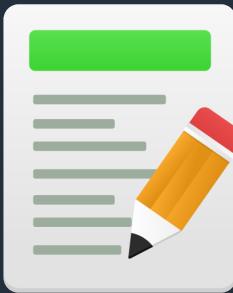
EC2 Auto Scaling

- You can attach one or more classic ELBs to your existing ASG.
- You can attach one or more Target Groups to your ASG to include instances behind an ALB.
- The ELBs must be in the same region.
- Launch configuration is the template used to create new EC2 instances and includes parameters such as instance family, instance type, AMI, key pair and security groups.
- You cannot edit a launch configuration once defined.
- You can use a launch configuration with multiple Auto Scaling Groups (ASG).



Section 4: Exam Cram

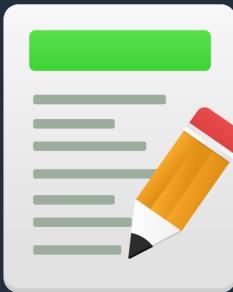
EC2 Auto Scaling – Scaling Option



Scaling Option	What it is	When to use
Maintain	Ensures the required number of instances are running	Use when you always need a known number of instances running at all times
Manual	Manually change desired capacity via the console or CLI	Use when your needs change rarely enough that you're OK to make manual changes
Scheduled	Adjust min/max instances on specific dates/times or recurring time periods	Use when you know when your busy and quiet times are. Useful for ensuring enough instances are available <i>before</i> very busy times
Dynamic	Scale in response to system load or other triggers using metrics	Useful for changing capacity based on system utilization, e.g. CPU hits 80%

Section 4: Exam Cram

EC2 Auto Scaling – Scaling Types (associated with Dynamic Scaling Policies)



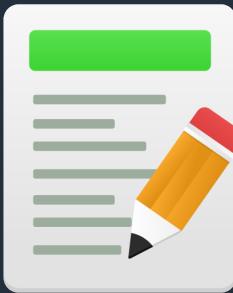
Scaling	What it is	When to use
Target Tracking Policy	The scaling policy adds or removes capacity as required to keep the metric at, or close to, the specified target value	A use case is that you want to keep the aggregate CPU usage of your ASG at 70%
Simple Scaling Policy	Waits until health check and cool down period expires before re-evaluating	This is a more conservative way to add/remove instances. Useful when load is erratic. AWS recommend step scaling instead of simple in most cases
Step Scaling Policy	Increase or decrease the current capacity of your Auto Scaling group based on a set of scaling adjustments, known as step adjustments	Useful when you want to vary adjustments based on the size of the alarm breach



Section 4: Exam Cram

EC2 Auto Scaling

- Can also scale based on an Amazon Simple Queue Service (SQS) queue.
- This comes up as an exam question for SAA-C02.
- Uses a custom metric that's sent to Amazon CloudWatch that measures the number of messages in the queue per EC2 instance in the Auto Scaling group.
- Then use a target tracking policy that configures your Auto Scaling group to scale based on the custom metric and a set target value. CloudWatch alarms invoke the scaling policy.
- Use a custom “backlog per instance” metric to track not just the number of messages in the queue but the number available for retrieval.
- Can base off the SQS Metric “`ApproximateNumberOfMessages`”.



Section 4: Exam Cram

EC2 Auto Scaling – Termination Policy

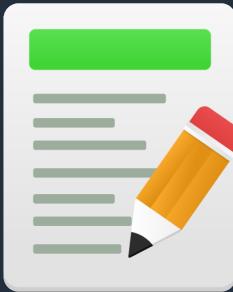
- Termination policies control which instances are terminated first when a scale-in event occurs.
- There is a default termination policy and options for configuring your own customized termination policies.
- The default termination policy is designed to help ensure that instances span Availability Zones evenly for high availability.
- The default policy is kept generic and flexible to cover a range of scenarios.



Section 4: Exam Cram

EC2 Auto Scaling

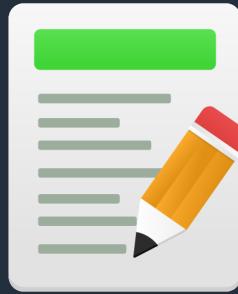
- You can define Instance Protection which stops Auto Scaling from scaling in and terminating the instances.
- Auto Scaling can perform rebalancing when it finds that the number of instances across AZs is not balanced.
- Auto Scaling rebalances by launching new EC2 instances in the AZs that have fewer instances first, only then will it start terminating instances in AZs that had more instances.
- Imbalances occur due to removing AZs/subnets, manually terminating instances, EC2 capacity issues, or Spot termination etc.



Section 4: Exam Cram

EC2 Auto Scaling – Health Checks

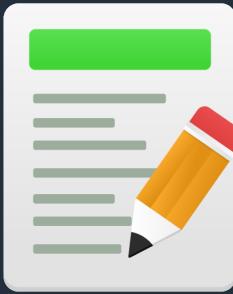
- By default uses EC2 status checks.
- Can also use ELB health checks and custom health checks.
- ELB health checks are in addition to the EC2 status checks.
- If any health check returns an unhealthy status the instance will be terminated.
- With ELB an instance is marked as unhealthy if ELB reports it as OutOfService.
- A healthy instance enters the InService state.
- If an instance is marked as unhealthy it will be scheduled for replacement.
- If connection draining is enabled, Auto Scaling waits for in-flight requests to complete or timeout before terminating instances.



Section 4: Exam Cram

EC2 Auto Scaling

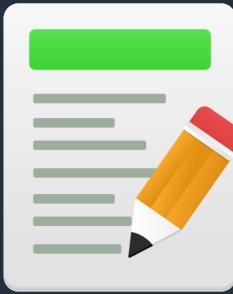
- Unlike AZ rebalancing, termination of unhealthy instances happens first, then Auto Scaling attempts to launch new instances to replace terminated instances.
- You can suspend and then resume one or more of the scaling processes for your Auto Scaling group.
- Suspending scaling processes can be useful when you want to investigate a configuration problem or other issue with your web application and then make changes to your application, without invoking the scaling processes.
- You can manually move an instance from an ASG and put it in the standby state.
- Standby state can be used for performing updates/changes/troubleshooting etc. without health checks being performed or replacement instances being launched.



Section 4: Exam Cram

EC2 Auto Scaling

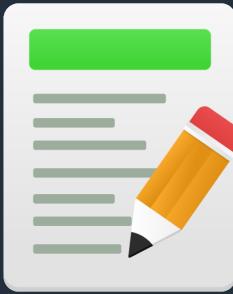
- The cooldown period is a configurable setting for your Auto Scaling group that helps to ensure that it doesn't launch or terminate additional instances before the previous scaling activity takes effect.
- The default value is 300 seconds.
- The warm-up period is the period of time in which a newly created EC2 instance launched by ASG using step scaling is not considered toward the ASG metrics.



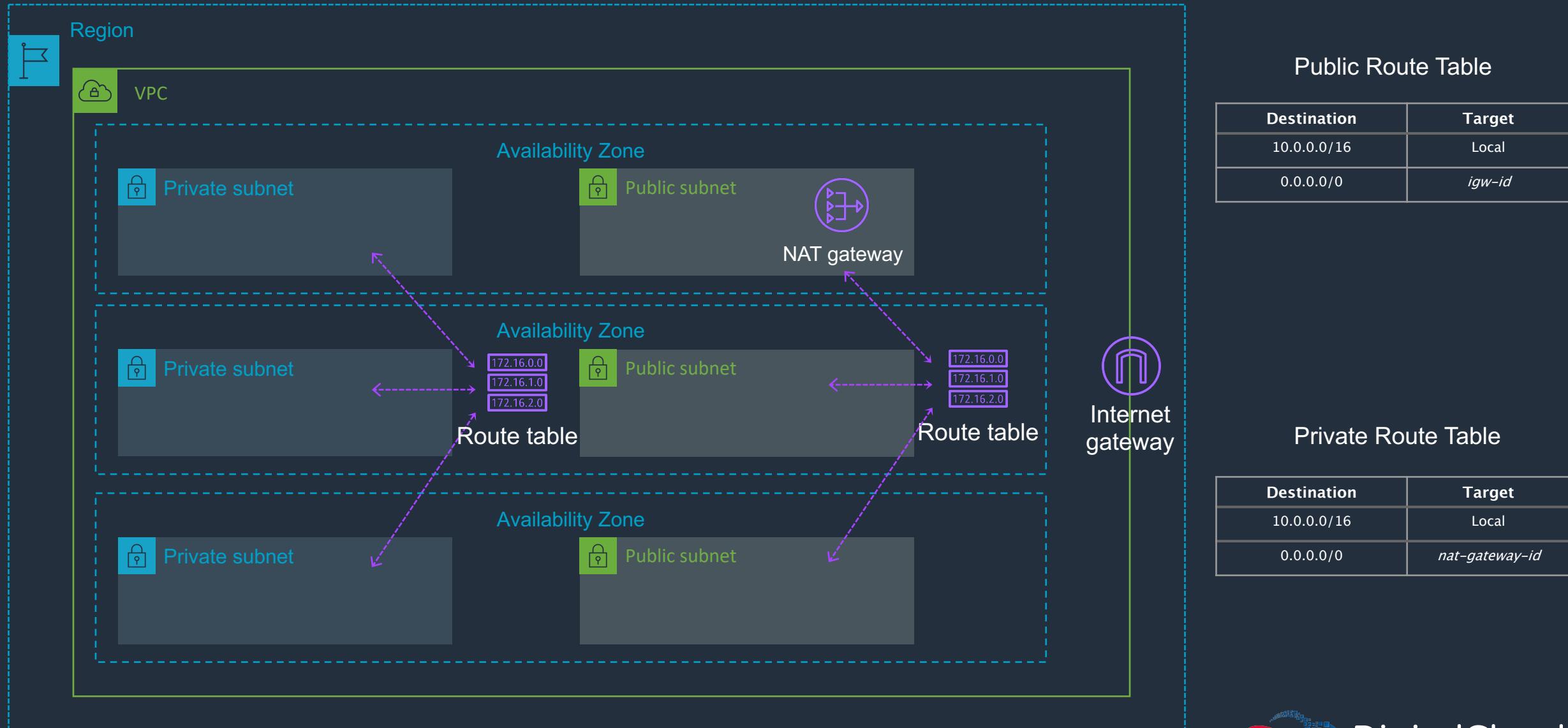
Section 4: Exam Cram

EC2 Auto Scaling – Monitoring

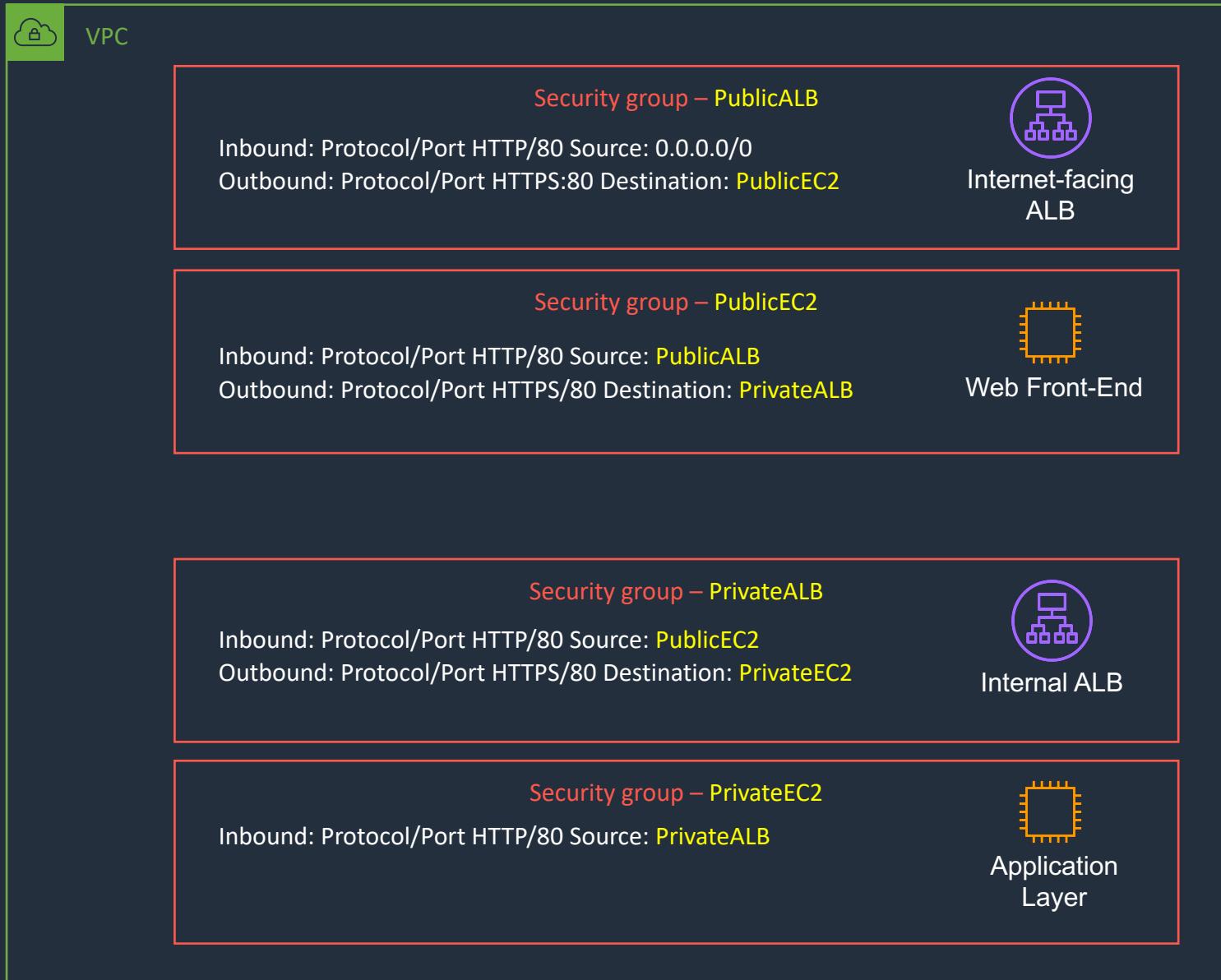
- Basic monitoring sends EC2 metrics to CloudWatch about ASG instances every 5 minutes.
- Detailed can be enabled and sends metrics every 1 minute (chargeable).
- When the launch configuration is created from the console basic monitoring of EC2 instances is enabled by default.
- When the launch configuration is created from the CLI detailed monitoring of EC2 instances is enabled by default.



Section 5: Creating a Custom VPC



Section 5: Security Groups



Default Security Group

Inbound:

Source	Protocol	Port
Security Group ID	All	All

Outbound:

Destination	Protocol	Port
0.0.0.0/0	All	All
::/0	All	All

Custom Security Group

Inbound:

Source	Protocol	Port

Outbound:

Destination	Protocol	Port
0.0.0.0/0	All	All
::/0	All	All

Section 5: Network Access Control Lists (NACLS)



Default NACL

Inbound:

Protocol	Port	Source	Action
All	All	0.0.0.0/0	ALLOW
All	All	::/0	ALLOW

Outbound:

Protocol	Port	Source	Action
All	All	0.0.0.0/0	ALLOW
All	All	::/0	ALLOW

Custom NACL

Inbound:

Protocol	Port	Source	Action
All	All	0.0.0.0/0	DENY
All	All	::/0	DENY

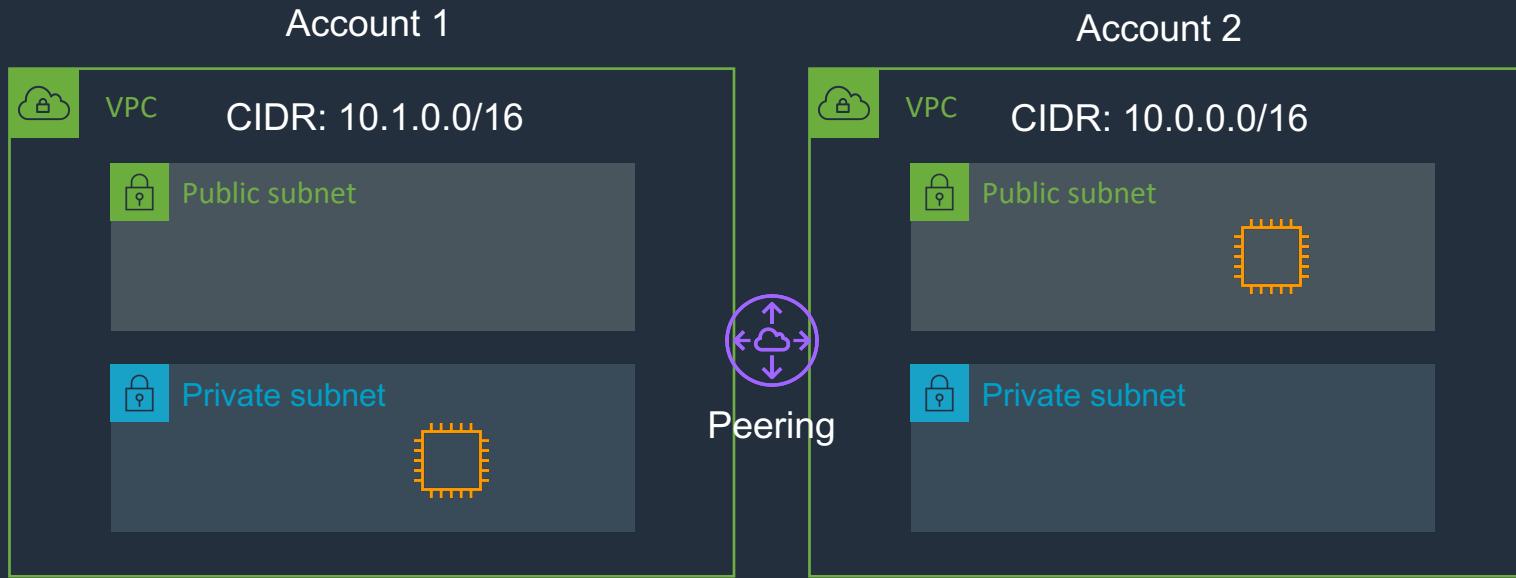
Outbound:

Protocol	Port	Source	Action
All	All	0.0.0.0/0	DENY
All	All	::/0	DENY

Section 5: Security Groups vs Network ACLs

Security Group	Network ACL
Operates at the instance level	Operates at the subnet level
Supports allow rules only	Supports allow and deny rules
Stateful	Stateless
Evaluates all rules	Processes rules in order
Applies to an instance only if associated with a group	Automatically applies to all instances in the subnets its associated with

Section 5: VPC Peering



Security Group

Inbound:

Source	Protocol	Port
Security Group ID	All ICMP v4	All

Route Table

Destination	Target
10.0.0.0/16	<i>peering-connection-id</i>

Security Group

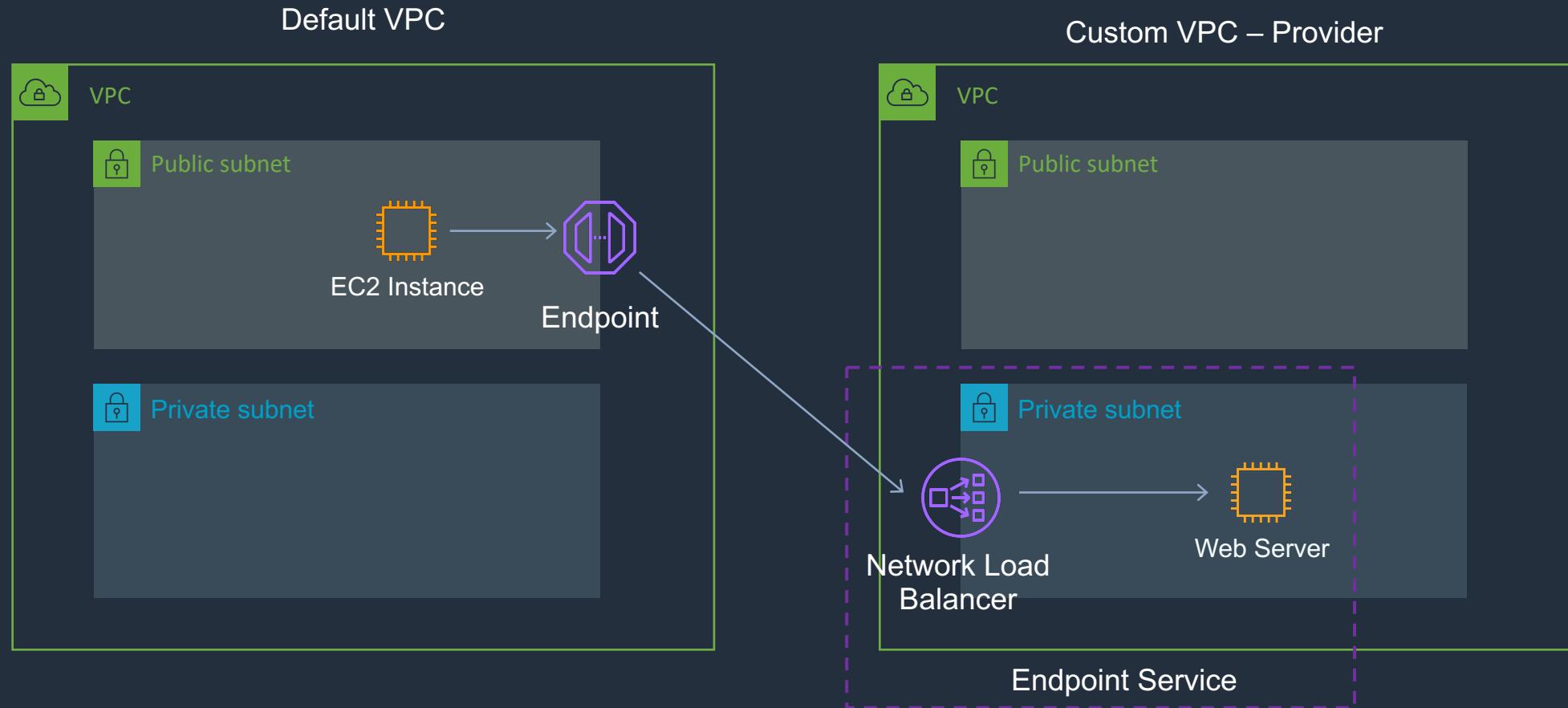
Inbound:

Source	Protocol	Port
0.0.0.0/0	TCP	22

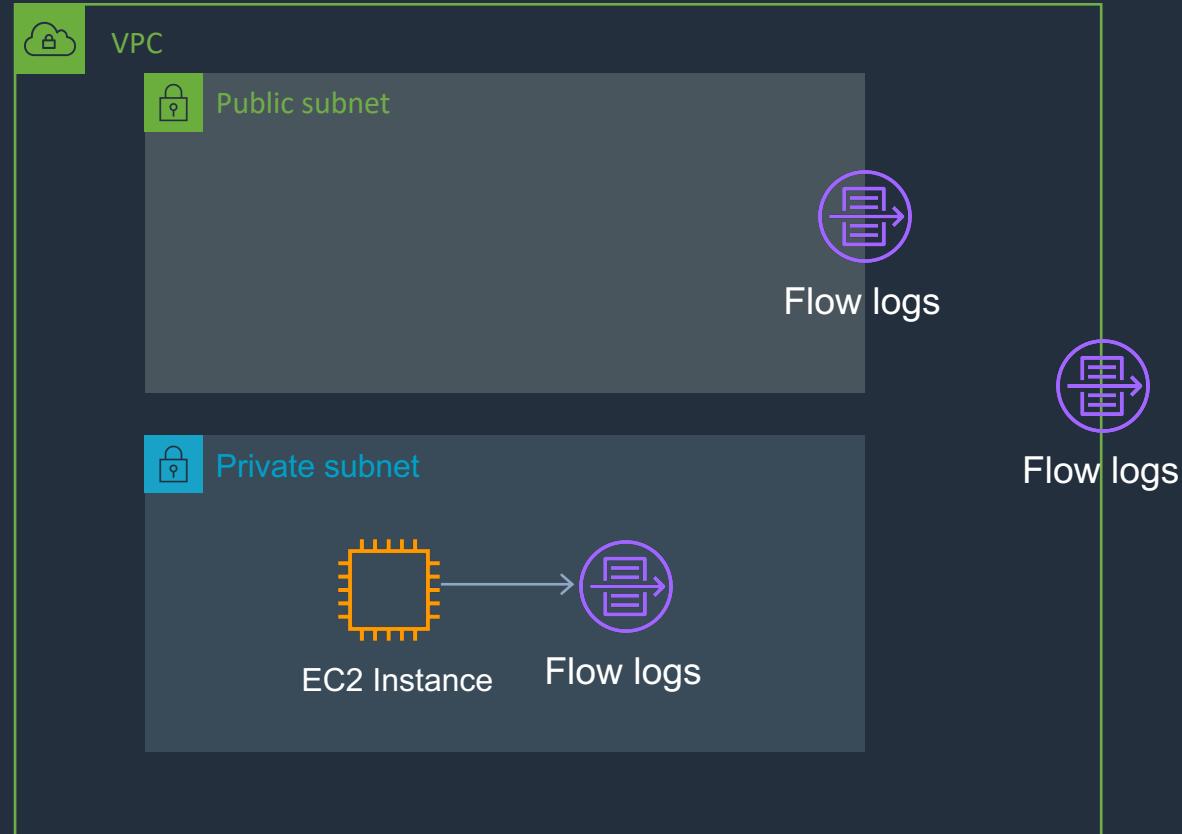
Route Table

Destination	Target
10.1.0.0/16	<i>peering-connection-id</i>

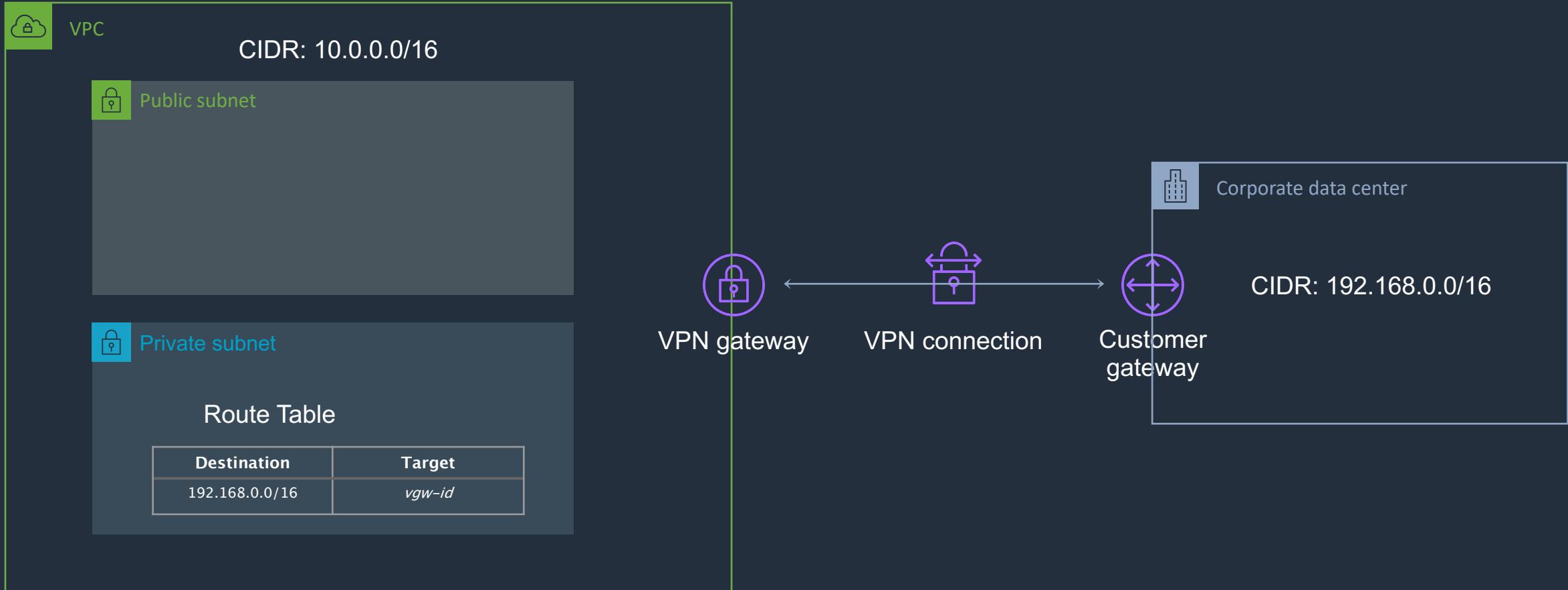
Section 5: VPC Endpoint Services



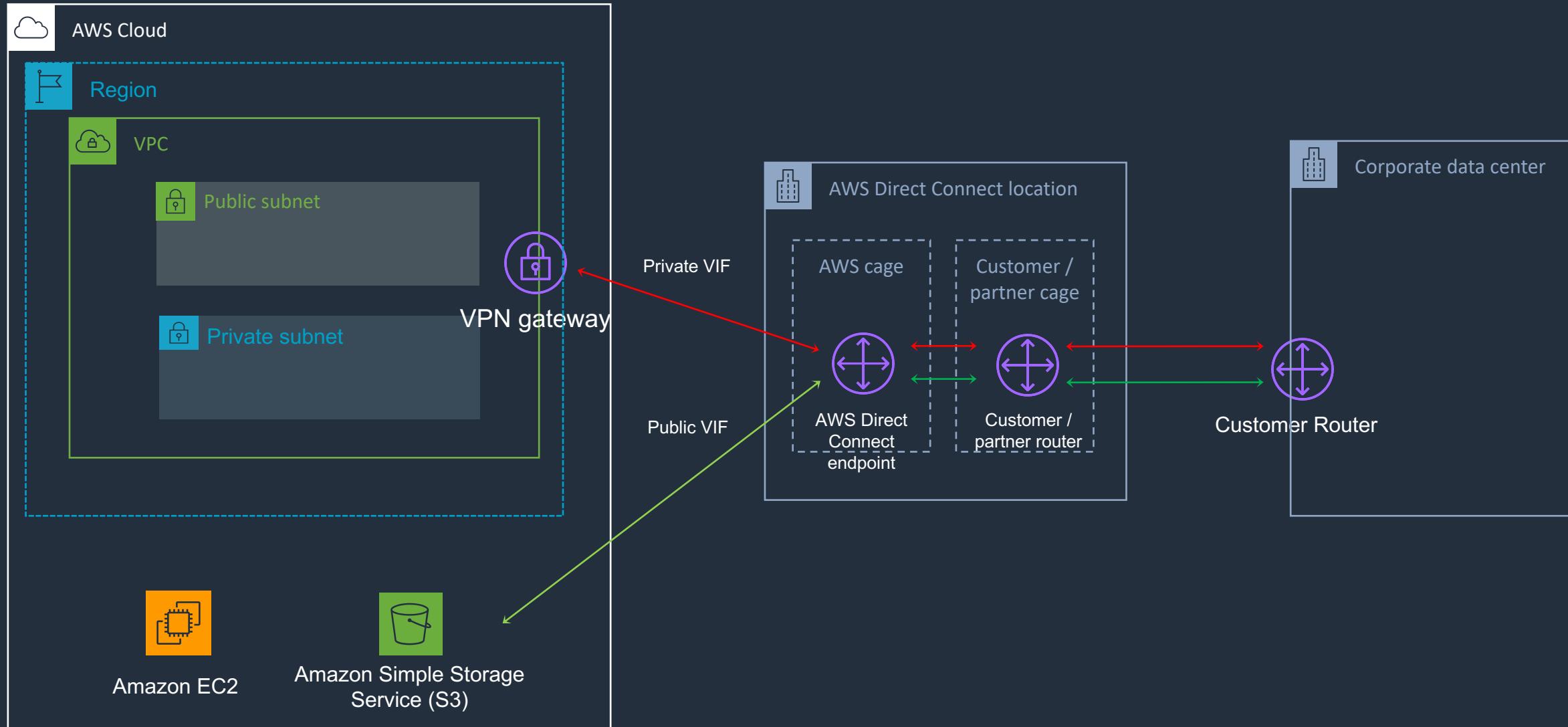
Section 5: VPC Flow Logs



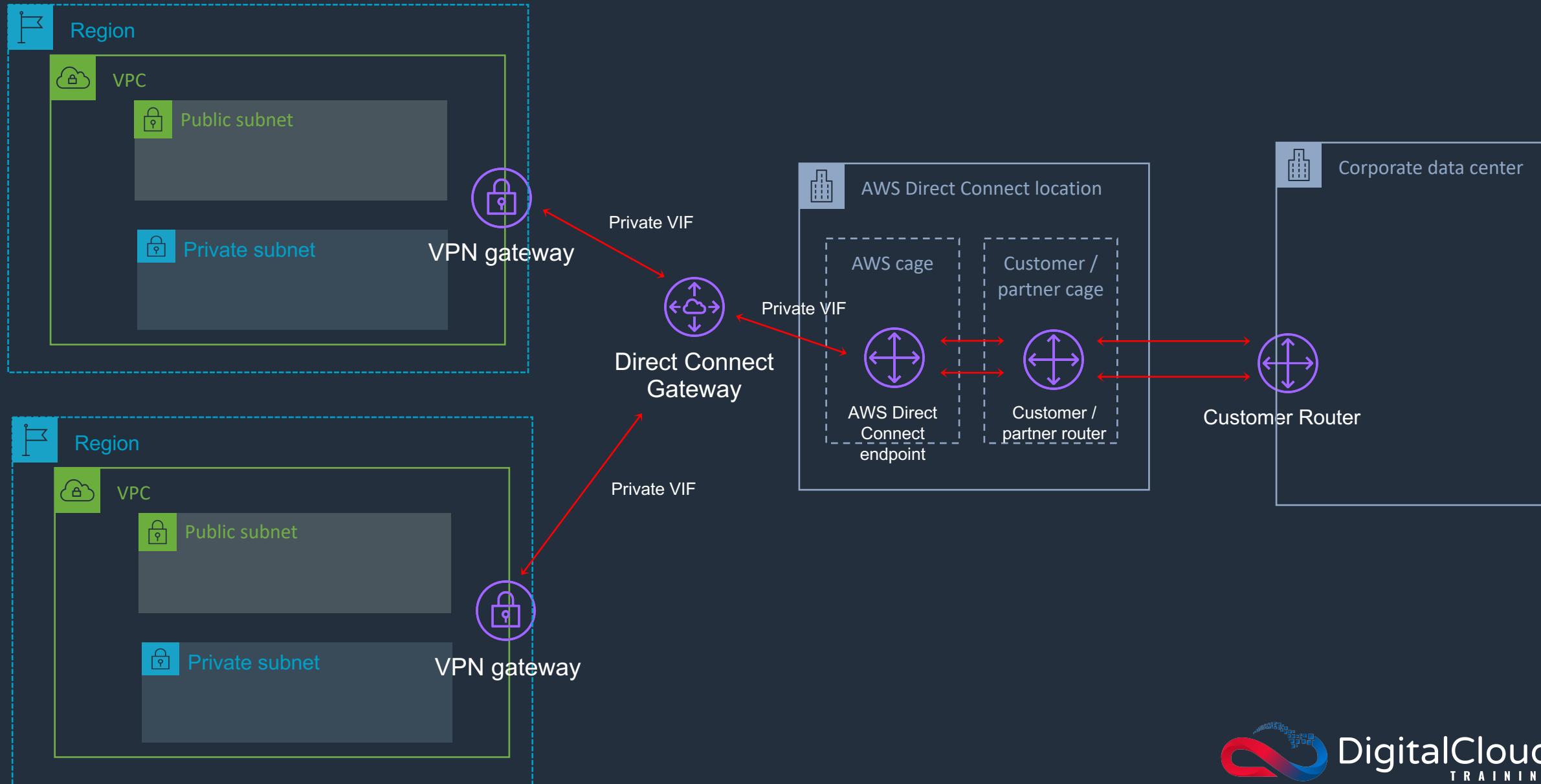
Section 5: Virtual Private Networks (VPN)



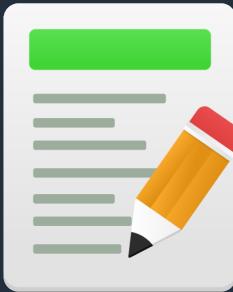
Section 5: AWS Direct Connect



Section 5: AWS Direct Connect Gateway



Section 5: Exam Cram



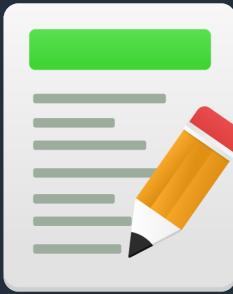
Amazon VPC

- A Virtual Private Cloud (VPC) is logically isolated from other VPCs on AWS.
- VPCs are region wide.
- A default VPC is created in each region with a subnet in each AZ.
- You can define dedicated tenancy for a VPC to ensure instances are launched on dedicated hardware (overrides the configuration specified at launch).
- The default VPC has all-public subnets.

Section 5: Exam Cram

Amazon VPC

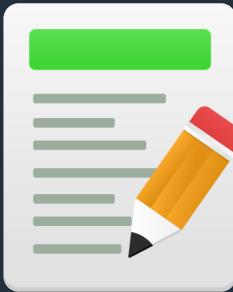
- Public subnets are subnets that have:
 - “Auto-assign public IPv4 address” set to “Yes”.
 - The subnet route table has an attached Internet Gateway.
- Instances in the default VPC always have both a public and private IP address.
- AZs names are mapped to different zones for different users (i.e. the AZ “ap-southeast-2a” may map to a different physical zone for a different user).



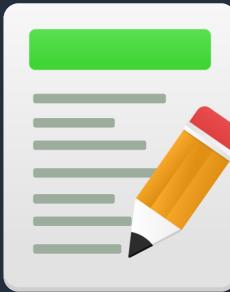
Section 5: Exam Cram

Amazon VPC – Components

- **A Virtual Private Cloud:** A logically isolated virtual network in the AWS cloud. You define a VPC's IP address space from ranges you select.
- **Subnet:** A segment of a VPC's IP address range where you can place groups of isolated resources (maps to a single AZ).
- **Internet Gateway:** The Amazon VPC side of a connection to the public Internet.
- **NAT Gateway:** A highly available, managed Network Address Translation (NAT) service for your resources in a private subnet to access the Internet.
- **Hardware VPN Connection:** A hardware-based VPN connection between your Amazon VPC and your datacenter, home network, or co-location facility.
- **Virtual Private Gateway:** The Amazon VPC side of a VPN connection.
- **Customer Gateway:** Your side of a VPN connection.



Section 5: Exam Cram



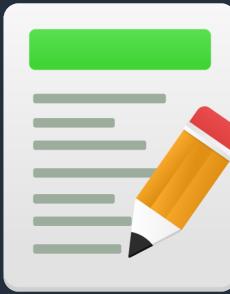
Amazon VPC – Components

- **Router:** Routers interconnect subnets and direct traffic between Internet gateways, virtual private gateways, NAT gateways, and subnets.
- **Peering Connection:** A peering connection enables you to route traffic via private IP addresses between two peered VPCs.
- **VPC Endpoints:** Enables private connectivity to services hosted in AWS, from within your VPC without using an Internet Gateway, VPN, Network Address Translation (NAT) devices, or firewall proxies.
- **Egress-only Internet Gateway:** A stateful gateway to provide egress only access for IPv6 traffic from the VPC to the Internet.

Section 5: Exam Cram

Amazon VPC – Routing

- The VPC router performs routing between AZs within a region.
- The VPC router connects different AZs together and connects the VPC to the Internet Gateway.
- Each subnet has a route table the router uses to forward traffic within the VPC.
- Route tables also have entries to external destinations.



Public Route Table

Destination	Target
10.0.0.0/16	Local
0.0.0.0/0	<i>igw-id</i>

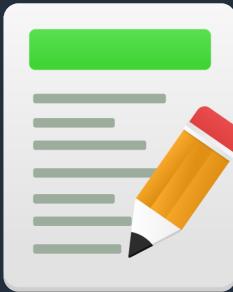
Private Route Table

Destination	Target
10.0.0.0/16	Local
0.0.0.0/0	<i>nat-gateway-id</i>

Section 5: Exam Cram

Amazon VPC – Subnets

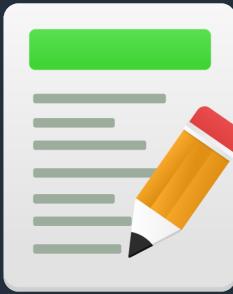
- Types of subnet:
 - If a subnet's traffic is routed to an internet gateway, the subnet is known as a public subnet.
 - If a subnet doesn't have a route to the internet gateway, the subnet is known as a private subnet.
 - If a subnet doesn't have a route to the internet gateway, but has its traffic routed to a virtual private gateway for a VPN connection, the subnet is known as a VPN-only subnet.



Section 5: Exam Cram

Amazon VPC – Subnets and Addressing

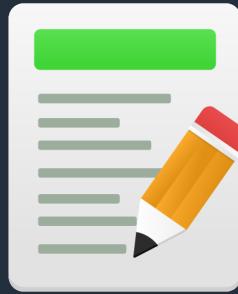
- The VPC is created with a master address range (CIDR block, can be anywhere from 16-28 bits), and subnet ranges are created within that range.
- New subnets are always associated with the default route table.
- Once the VPC is created you cannot change the CIDR block.
- You cannot create additional CIDR blocks that overlap with existing CIDR blocks.
- You cannot create additional CIDR blocks in a different RFC 1918 range.
- Subnets with overlapping IP address ranges cannot be created.
- The first 4 and last 1 IP addresses in a subnet are reserved.
- Subnets are created within availability zones (AZs).
- Subnets map 1:1 to AZs and cannot span AZs.



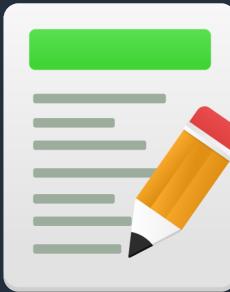
Section 5: Exam Cram

Amazon VPC – Internet Gateways

- An Internet Gateway serves two purposes: .
 - To provide a target in your VPC route tables for internet-routable traffic.
 - To perform network address translation (NAT) for instances that have been assigned public IPv4 addresses.
- Internet Gateways (IGW) must be created and then attached to a VPC, be added to a route table, and then associated with the relevant subnet(s).
- No availability risk or bandwidth constraints.
- You cannot have multiple Internet Gateways in a VPC.
- Egress-only Internet Gateway provides outbound Internet access for IPv6 addressed instances.



Section 5: Exam Cram



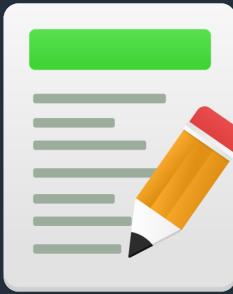
Amazon VPC – Security Groups

- Security groups act like a firewall at the instance (network interface) level.
- Can only assign permit rules in a security group, cannot assign deny rules.
- All rules are evaluated until a permit is encountered or continues until the implicit deny.
- Can control ingress and egress traffic.
- Security groups are stateful.
- By default, custom security groups do not have inbound allow rules (all inbound traffic is denied by default).
- By default, default security groups do have inbound allow rules (allowing traffic from within the group).

Section 5: Exam Cram

Amazon VPC – Security Groups

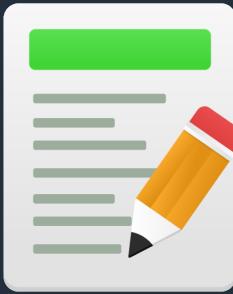
- All outbound traffic is allowed by default in custom and default security groups.
- You cannot delete the security group that's created by default within a VPC.
- You can use security group names as the source or destination in other security groups.
- You can use the security group name as a source in its own inbound rules.
- Security group membership can be changed whilst instances are running.
- Any changes made will take effect immediately.
- You cannot block specific IP addresses using security groups, use NACLs instead.



Section 5: Exam Cram

Amazon VPC – Network ACLs

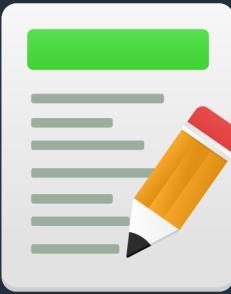
- Network ACL's function at the subnet level.
- With NACLs you can have permit and deny rules.
- Network ACLs contain a numbered list of rules that are evaluated in order from the lowest number until the explicit deny.
- Network ACLs have separate inbound and outbound rules and each rule can allow or deny traffic.
- Network ACLs are stateless so responses are subject to the rules for the direction of traffic.
- NACLs only apply to traffic that is ingress or egress to the subnet not to traffic within the subnet.



Section 5: Exam Cram

Amazon VPC – Network ACLs

- A VPC automatically comes with a default network ACL which allows all inbound/outbound traffic.
- A custom NACL denies all traffic both inbound and outbound by default.
- All subnets must be associated with a network ACL.
- You can create custom network ACL's. By default, each custom network ACL denies all inbound and outbound traffic until you add rules.
- You can associate a network ACL with multiple subnets; however a subnet can only be associated with one network ACL at a time.
- Network ACLs do not filter traffic between instances in the same subnet.



Section 5: Exam Cram

Amazon VPC – Network ACLs

- NACLs are the preferred option for blocking specific IPs or ranges.
- Security groups cannot be used to block specific ranges of IPs.
- NACL is the first line of defence, the security group is the second line.
- Changes to NACLs take effect immediately.

Default NACL

Inbound:

Protocol	Port	Source	Action
All	All	0.0.0.0/0	ALLOW
All	All	::/0	ALLOW

Outbound:

Protocol	Port	Source	Action
All	All	0.0.0.0/0	ALLOW
All	All	::/0	ALLOW

Custom NACL

Inbound:

Protocol	Port	Source	Action
All	All	0.0.0.0/0	DENY
All	All	::/0	DENY

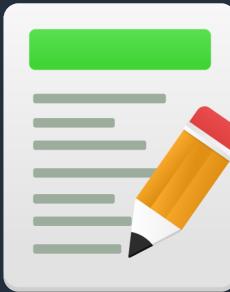
Outbound:

Protocol	Port	Source	Action
All	All	0.0.0.0/0	DENY
All	All	::/0	DENY



Section 5: Exam Cram

Security Group vs Network ACL



Security Group	Network ACL
Operates at the instance level	Operates at the subnet level
Supports allow rules only	Supports allow and deny rules
Stateful	Stateless
Evaluates all rules	Processes rules in order
Applies to an instance only if associated with a group	Automatically applies to all instances in the subnets its associated with

Section 5: Exam Cram

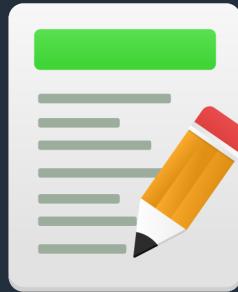
Amazon VPC – Connectivity

- There are several methods of connecting to a VPC. These include:
- AWS Managed VPN.
- AWS Direct Connect.
- AWS Direct Connect plus a VPN.
- AWS VPN CloudHub.
- Software VPN.
- Transit VPC.
- VPC Peering.
- AWS PrivateLink.
- VPC Endpoints.



Section 5: Exam Cram

AWS Managed VPN



What	AWS Managed IPSec VPN Connection over your existing Internet
When	Quick and usually simple way to establish a secure tunnelled connection to a VPC; redundant link for Direct Connect or other VPC VPN
Pros	Supports static routes or BGP peering and routing
Cons	Dependent on your Internet connection
How	Create a Virtual Private Gateway (VPG) on AWS, and a Customer Gateway on the on-premises side

Section 5: Exam Cram

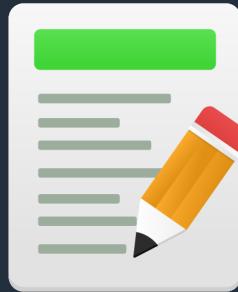
AWS Direct Connect



What	Dedicated network connection over private lines straight into the AWS backbone
When	Requires a large network link into AWS; lots of resources and services being provided on AWS to your corporate users
Pros	More predictable network performance; potential bandwidth cost reduction; up to 10 Gbps provisioned connections; supports BGP peering and routing
Cons	May require additional telecom and hosting provider relationships and/or network circuits; costly; takes time to provision
How	Work with your existing data networking provider; create Virtual Interfaces (VIFs) to connect to VPCs (private VIFs) or other AWS services like S3 or Glacier (public VIFs)

Section 5: Exam Cram

AWS Direct Connect Plus VPN



What	IPSec VPN connection over private lines (Direct Connect)
When	Need the added security of encrypted tunnels over Direct Connect
Pros	More secure (in theory) than Direct Connect alone
Cons	More complexity introduced by VPN layer
How	Work with your existing data networking provider

Section 5: Exam Cram

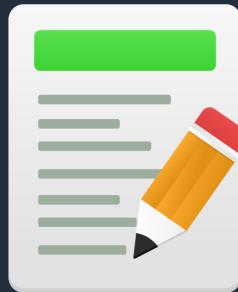
AWS VPN CloudHub



What	Connect locations in a hub and spoke manner using AWSs Virtual Private Gateway
When	Link remote offices for backup or primary WAN access to AWS resources and each other
Pros	Reuses existing Internet connections; supports BGP routes to direct traffic
Cons	Dependent on Internet connection; no inherent redundancy
How	Assign multiple Customer Gateways to a Virtual Private Gateway, each with their own BGP ASN and unique IP ranges

Section 5: Exam Cram

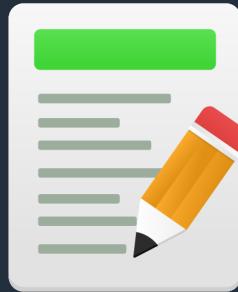
Software VPN



What	You provide your own VPN endpoint and software
When	You must manage both ends of the VPN connection for compliance reasons or you want to use a VPN option not supported by AWS
Pros	Ultimate flexibility and manageability
Cons	You must design for any needed redundancy across the whole chain
How	Install VPN software via Marketplace appliance or on an EC2 instance

Section 5: Exam Cram

Transit VPC



What	Common strategy for connecting geographically dispersed VPCs and locations in order to create a global network transit center
When	Locations and VPC-deployed assets across multiple regions that need to communicate with one another
Pros	Ultimate flexibility and manageability but also AWS-managed VPN hub-and-spoke between VPCs
Cons	You must design for any needed redundancy across the whole chain
How	Providers like Cisco, Juniper Networks, and Riverbed have offerings which work with their equipment and AWS VPC

Section 5: Exam Cram

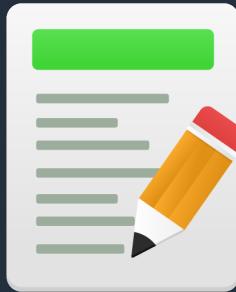
VPC Peering



What	AWS-provided network connectivity between two VPCs
When	Multiple VPCs need to communicate or access each other's resources
Pros	Uses AWS backbone without traversing the Internet
Cons	Transitive peering is not supported
How	VPC peering request made; accepter accepts request (either within or across accounts)

Section 5: Exam Cram

AWS PrivateLink



What	AWS-provided network connectivity between VPCs and/or AWS services using interface endpoints
When	Keep Private Subnets truly private by using the AWS backbone to reach other AWS or Marketplace services rather than the public Internet
Pros	Redundant; uses the AWS backbone
Cons	
How	Create endpoint for required AWS or Marketplace service in all required subnets; access via the provided DNS hostname

Section 5: Exam Cram

VPC Endpoints



	Interface Endpoint	Gateway Endpoint
What	Elastic Network Interface with a Private IP	A gateway that is a target for a specific route
How	Uses DNS entries to redirect traffic	Uses prefix lists in the route table to redirect traffic
Which services	API Gateway, CloudFormation, CloudWatch etc.	Amazon S3, DynamoDB
Security	Security Groups	VPC Endpoint Policies

Section 5: Exam Cram

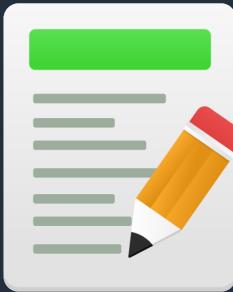
VPC Sharing

You can allow other AWS accounts to create their application resources, such as EC2 instances, Relational Database Service (RDS) databases, Redshift clusters, and Lambda functions, into shared, centrally-managed Amazon Virtual Private Clouds (VPCs).

VPC sharing enables subnets to be shared with other AWS accounts within the same AWS Organization.

Benefits include:

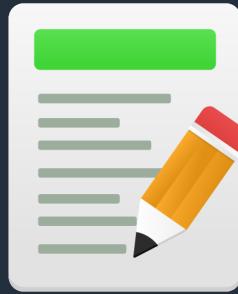
- Separation of duties: centrally controlled VPC structure, routing, IP address allocation.
- Application owners continue to own resources, accounts, and security groups.
- VPC sharing participants can reference security group IDs of each other.
- Efficiencies: higher density in subnets, efficient use of VPNs and AWS Direct Connect.
- Hard limits can be avoided, for example, 50 VIFs per AWS Direct Connect connection through simplified network architecture.
- Costs can be optimized through reuse of NAT gateways, VPC interface endpoints, and intra-Availability Zone traffic.



Section 5: Exam Cram

VPC Flow Logs

- Flow Logs capture information about the IP traffic going to and from network interfaces in a VPC.
- Flow log data is stored using Amazon CloudWatch Logs.
- Flow logs can be created at the following levels:
 - VPC.
 - Subnet.
 - Network interface.



Section 5: Exam Cram

Guidance on High Availability for Networking in AWS

- Create subnets in the available AZs, to create Multi-AZ presence for your VPC.
- Best practice is to create at least two VPN tunnels into your Virtual Private Gateway.
- Direct Connect is not HA by default, so you need to establish a secondary connection via another Direct Connect (ideally with another provider) or use a VPN.
- For Multi-AZ redundancy of NAT Gateways, create gateways in each AZ with routes for private subnets to use the local gateway.



Section 6: Route 53 Overview



Amazon Route 53



Section 6: Route 53 DNS Record Types

Supported DNS records

- A (address record)
- AAAA (IPv6 address record)
- CNAME (canonical name record)
- Alias (an Amazon Route 53-specific virtual record)
- CAA (certification authority authorization)
- MX (mail exchange record)
- NAPTR (name authority pointer record)
- NS (name server record)
- PTR (pointer record)
- SOA (start of authority record)
- SPF (sender policy framework)
- SRV (service locator)
- TXT (text record)

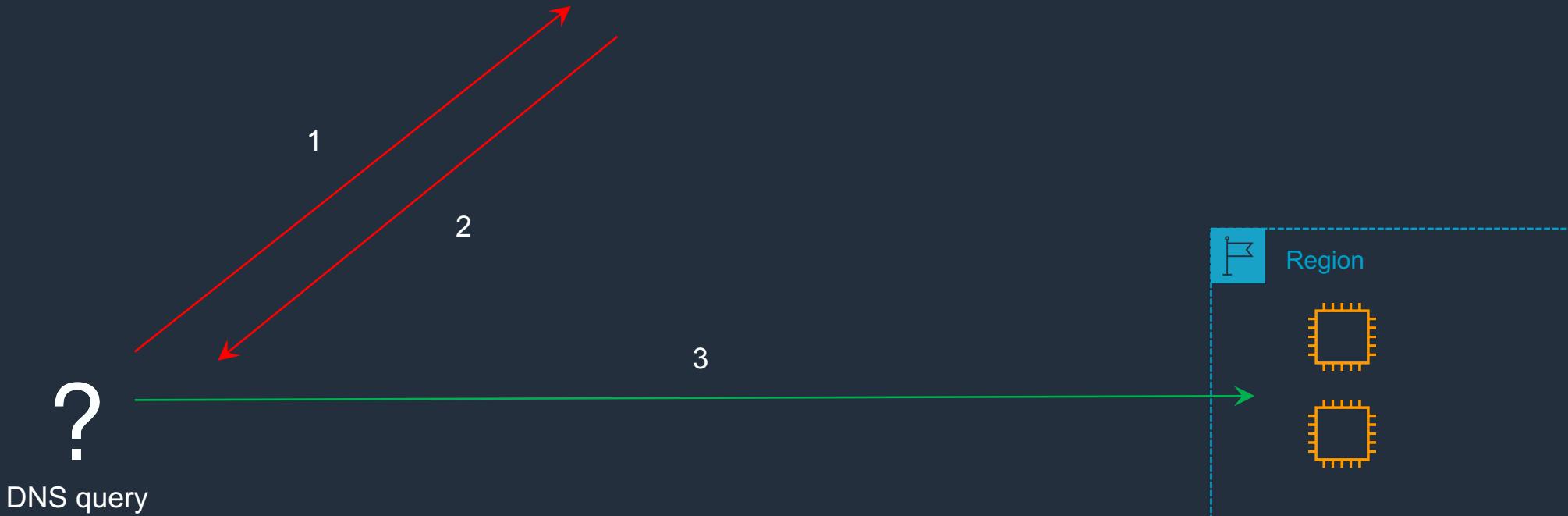
CNAME	Alias
Route 53 charges for CNAME queries	Route 53 doesn't charge for alias queries to AWS resources
You can't create a CNAME record at the top node of a DNS namespace (zone apex)	You can create an alias record at the zone apex (however you can't route to a CNAME at the zone apex)
A CNAME can point to any DNS record that is hosted anywhere	An alias record can only point to a CloudFront distribution, Elastic Beanstalk environment, ELB, S3 bucket as a static website, or to another record in the same hosted zone that you're creating the alias record in

Section 6: Route 53 - Simple Routing Policy

Name	Type	Value	TTL
simple.dctlabs.com	A	1.1.1.1	60
		2.2.2.2	
simpler.dctlabs.com	A	3.3.3.3	60

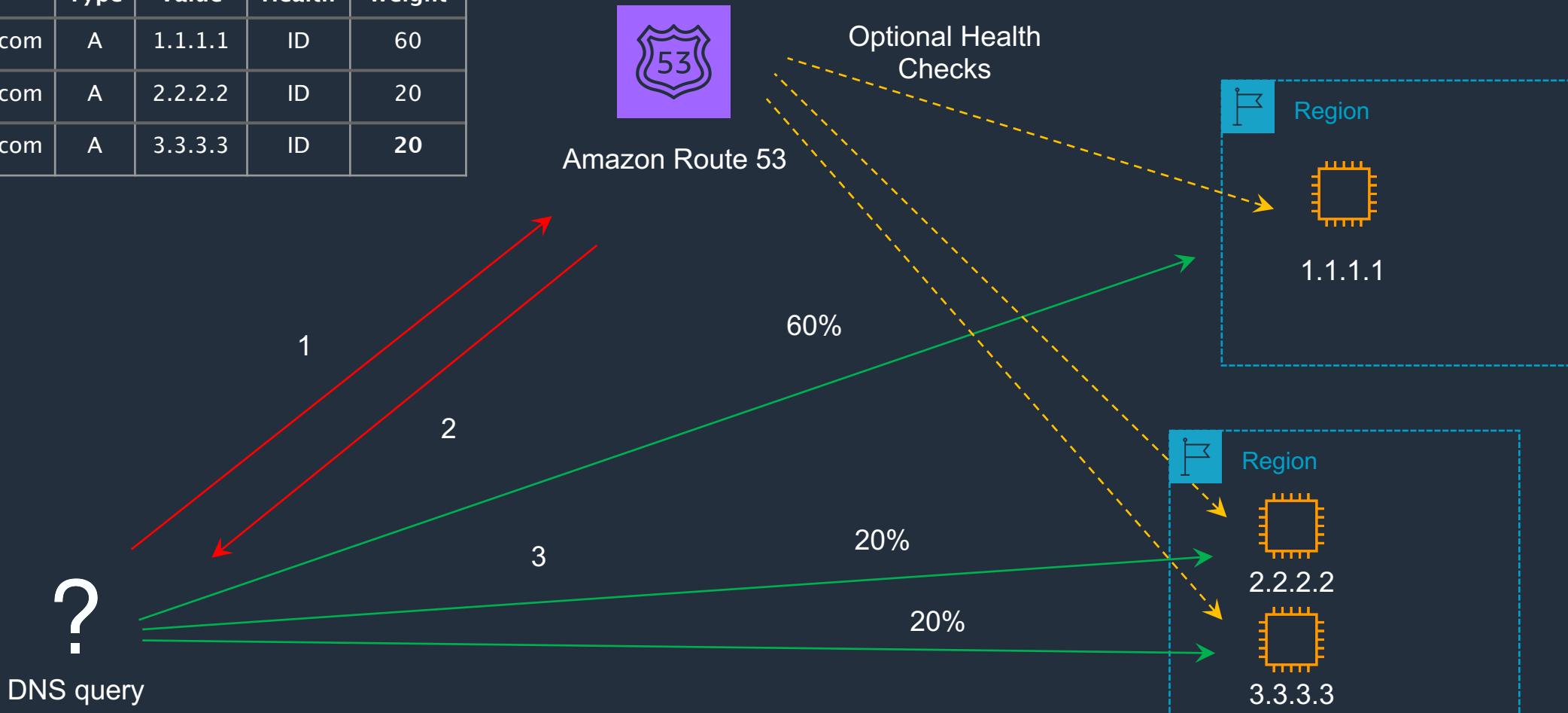


Amazon Route 53



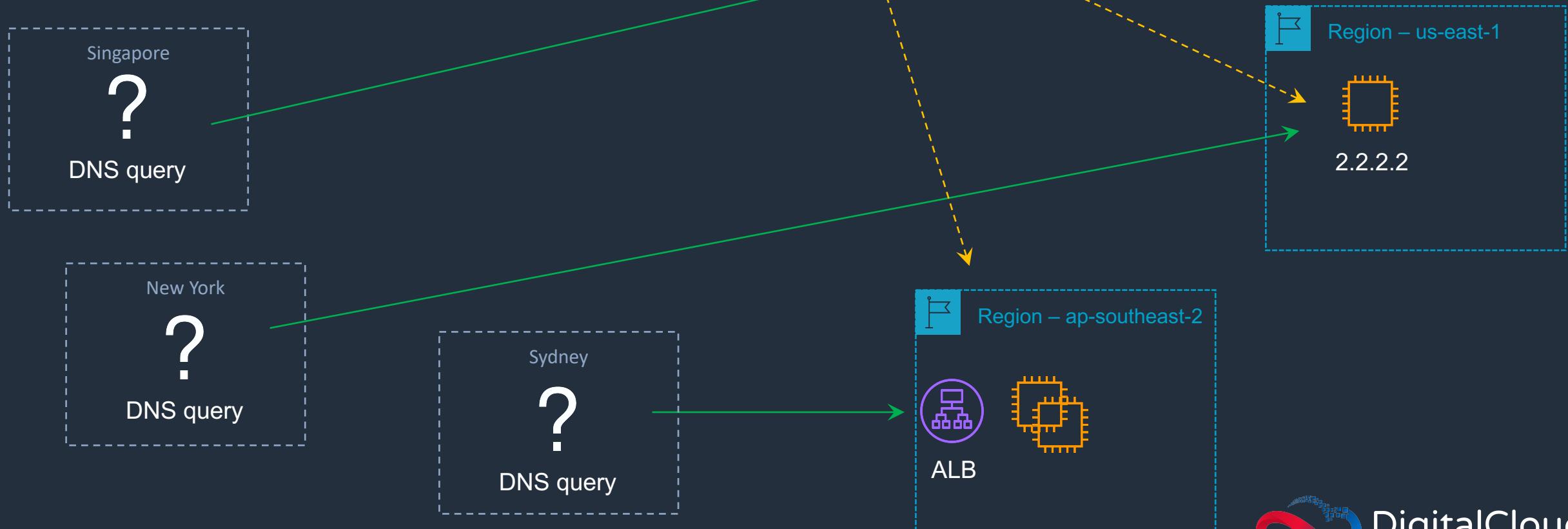
Section 6: Route 53 - Weighted Routing Policy

Name	Type	Value	Health	Weight
weighted.dctlabs.com	A	1.1.1.1	ID	60
weighted.dctlabs.com	A	2.2.2.2	ID	20
weighted.dctlabs.com	A	3.3.3.3	ID	20



Section 6: Route 53 - Latency Routing Policy

Name	Type	Value	Health	Region
latency.dctlabs.com	A	1.1.1.1	ID	ap-southeast-1
latency.dctlabs.com	A	2.2.2.2	ID	us-east-1
latency.dctlabs.com	A	<i>alb-id</i>	ID	ap-southeast-2



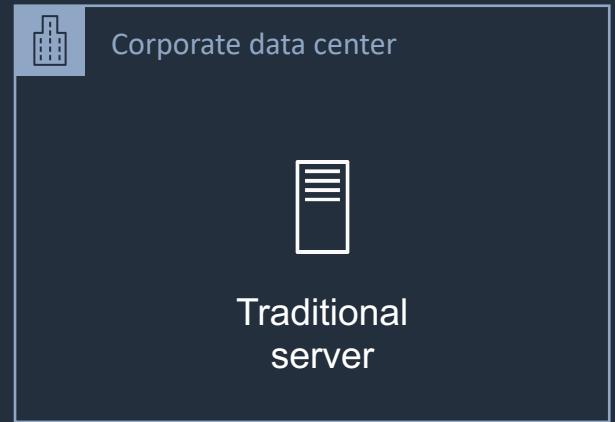
Section 6: Route 53 - Failover Routing Policy

Name	Type	Value	Health	Record Type
failover.dctlabs.com	A	1.1.1.1	ID	Primary
failover.dctlabs.com	A	<i>alb-id</i>		Secondary

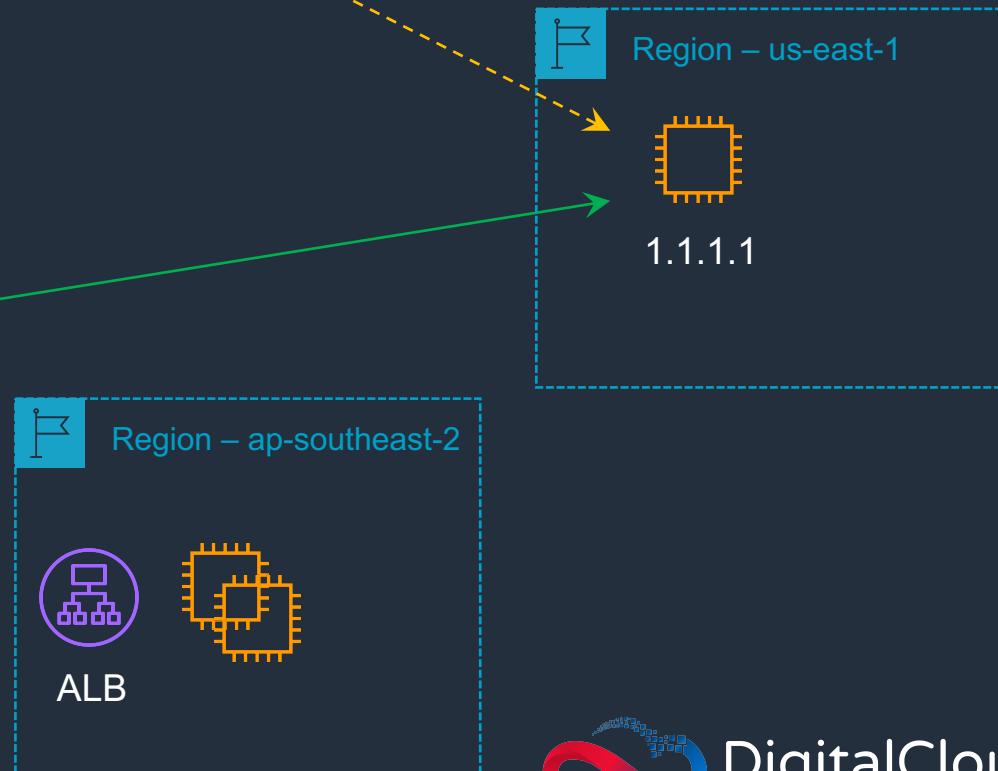


Amazon Route 53

Health Check
required on
Primary

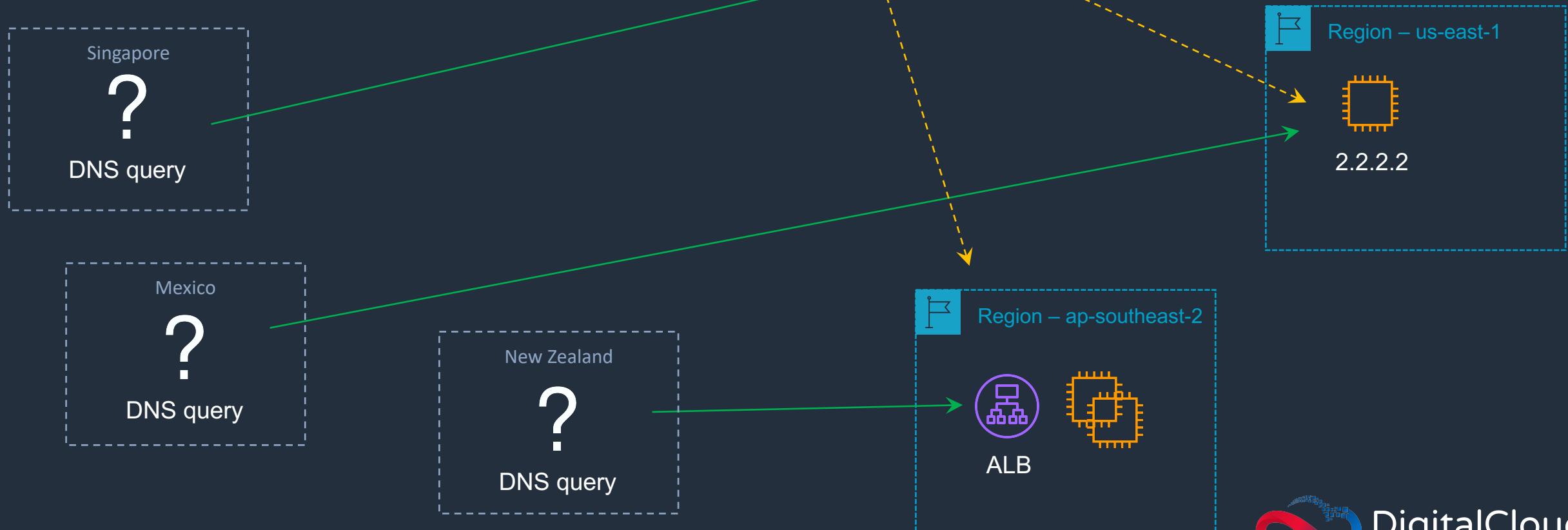


DNS query



Section 6: Route 53 - Geolocation Routing Policy

Name	Type	Value	Health	Geolocation
geolocation.dctlabs.com	A	1.1.1.1	ID	Singapore
geolocation.dctlabs.com	A	2.2.2.2	ID	Default
geolocation.dctlabs.com	A	<i>a/b-id</i>	ID	Oceania



Section 6: Route 53 - Multivalue Routing Policy

Name	Type	Value	Health	Multi Value
multivalue.dctlabs.com	A	1.1.1.1	ID	Yes
multivalue.dctlabs.com	A	2.2.2.2	ID	Yes
multivalue.dctlabs.com	A	3.3.3.3	ID	Yes

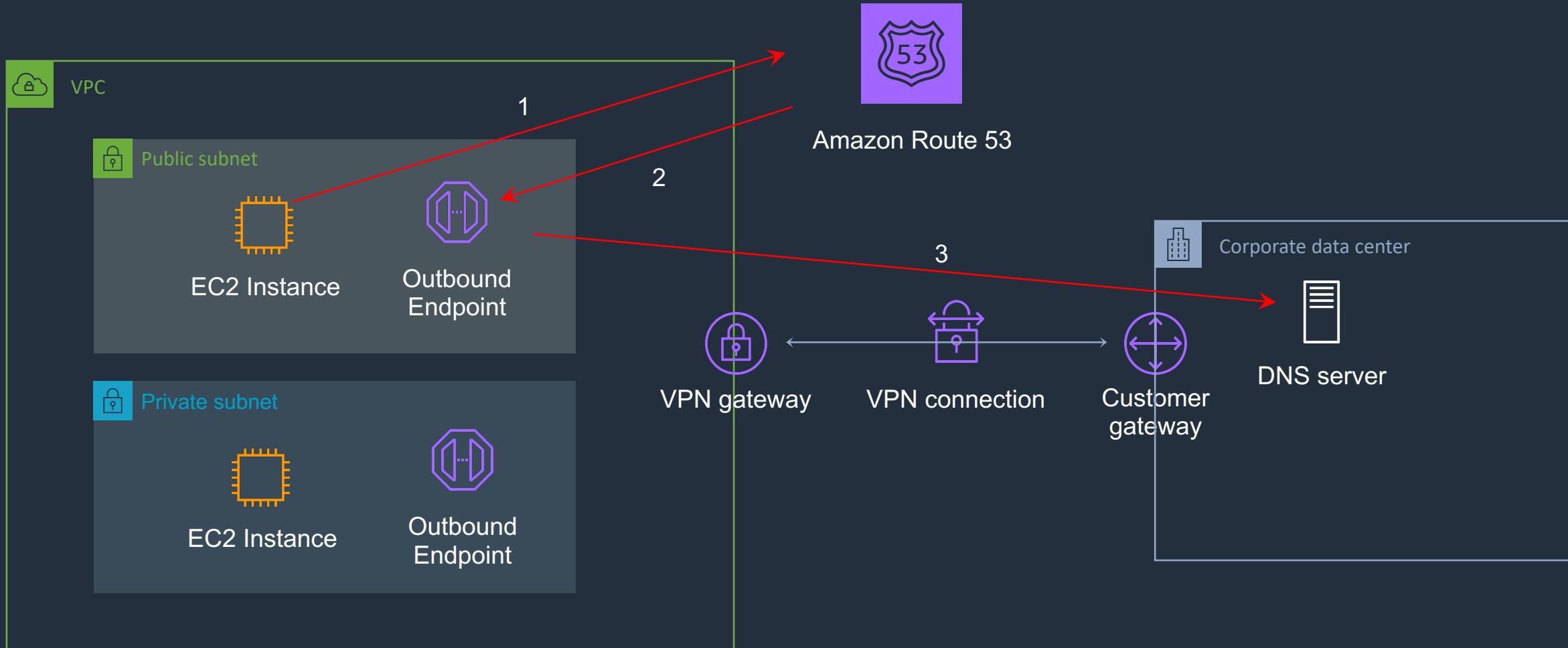


Amazon Route 53

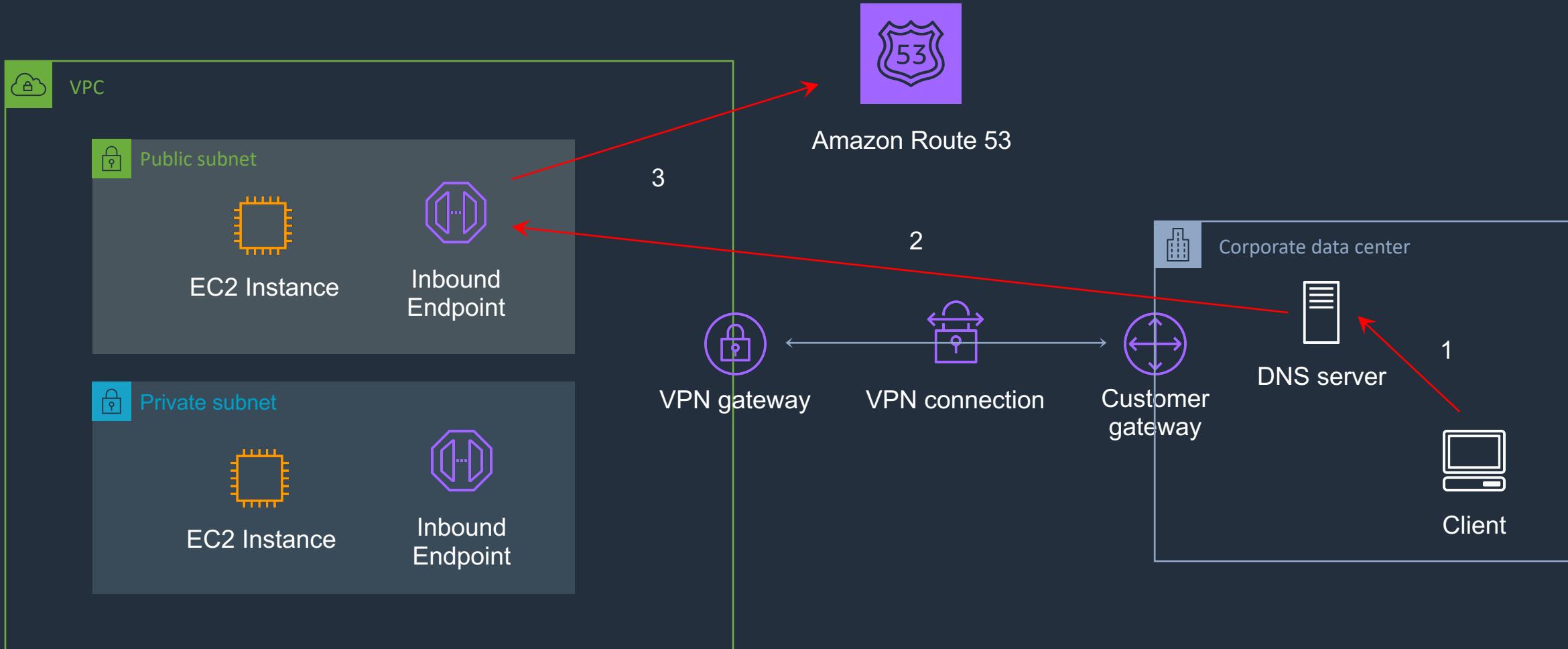
Health Checks:
returns healthy
records only



Section 6: Route 53 Resolver – Outbound Endpoints



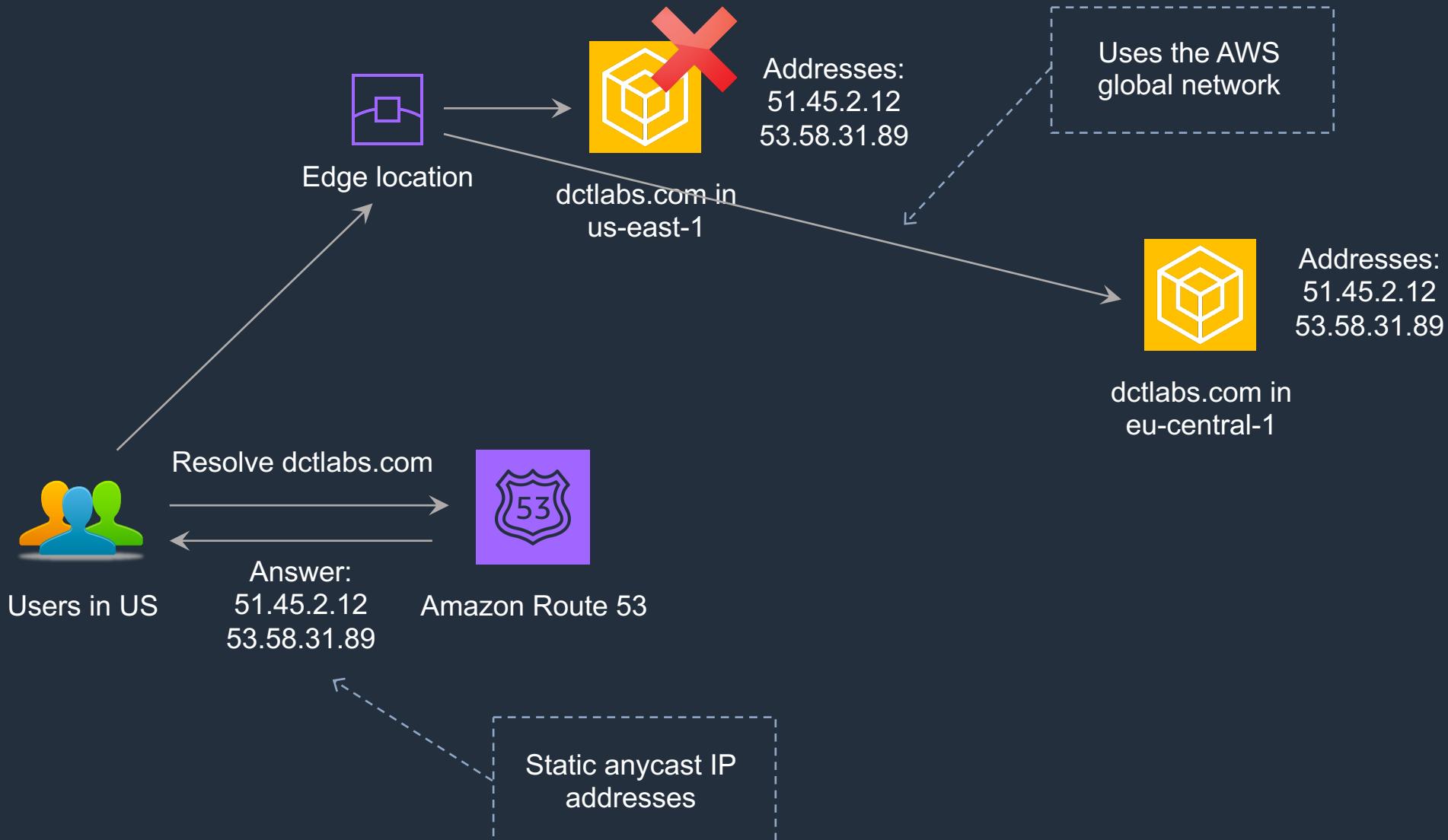
Section 6: Route 53 Resolver – Inbound Endpoints



Section 6: AWS Global Accelerator

- AWS Global Accelerator is a service that improves the availability and performance of your applications with local or global users. It provides static IP addresses that act as a fixed entry point to your application endpoints in a single or multiple AWS Regions, such as your Application Load Balancers, Network Load Balancers or Amazon EC2 instances.
- AWS Global Accelerator uses the AWS global network to optimize the path from your users to your applications, improving the performance of your TCP and UDP traffic. AWS Global Accelerator continually monitors the health of your application endpoints and will detect an unhealthy endpoint and redirect traffic to healthy endpoints in less than 1 minute.

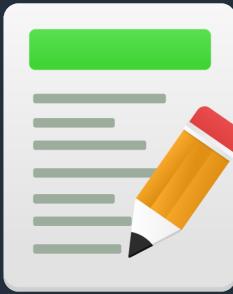
Section 6: AWS Global Accelerator



Section 6: Exam Cram

AWS Route 53

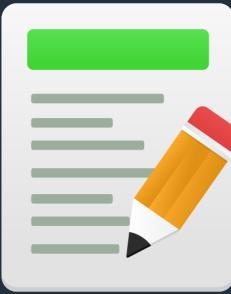
- Route 53 offers the following functions:
 - Domain name registry.
 - DNS resolution.
 - Health checking of resources.
- Route 53 is located alongside all edge locations.
- When you register a domain with Route 53 it becomes the authoritative DNS server for that domain and creates a public hosted zone.
- You can transfer domains to Route 53 only if the Top Level Domain (TLD) is supported.
- You can transfer a domain from Route 53 to another registrar by contacting AWS support.



Section 6: Exam Cram

AWS Route 53

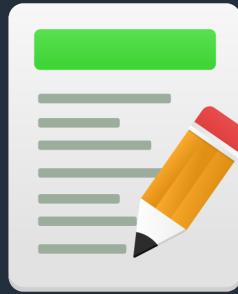
- You can transfer a domain to another account in AWS however it does not migrate the hosted zone by default (optional).
- It is possible to have the domain registered in one AWS account and the hosted zone in another AWS account.
- Private DNS is a Route 53 feature that lets you have authoritative DNS within your VPCs without exposing your DNS records (including the name of the resource and its IP address(es)) to the Internet.
- You can use the AWS Management Console or API to register new domain names with Route 53.



Section 6: Exam Cram

AWS Route 53 Hosted Zones

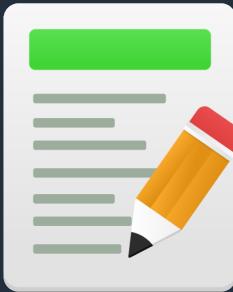
- A hosted zone is a collection of records for a specified domain.
- A hosted zone is analogous to a traditional DNS zone file; it represents a collection of records that can be managed together.
- There are two types of zones:
 - Public host zone – determines how traffic is routed on the Internet.
 - Private hosted zone for VPC – determines how traffic is routed within VPC (resources are not accessible outside the VPC).
- For private hosted zones you must set the following VPC settings to “true”:
 - enableDnsHostname.
 - enableDnsSupport.
- You also need to create a DHCP options set.



Section 6: Exam Cram

AWS Route 53 Health Checks

- Health checks check the instance health by connecting to it.
- Health checks can be pointed at:
 - Endpoints.
 - Status of other health checks.
 - Status of a CloudWatch alarm.
- Endpoints can be IP addresses or domain names.

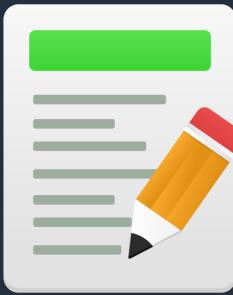


Section 6: Exam Cram

AWS Route 53 Records

➤ Amazon Route 53 currently supports the following DNS record types:

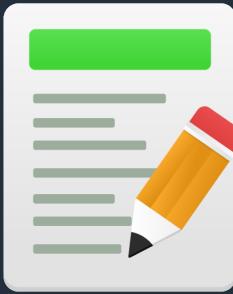
- A (address record).
- AAAA (IPv6 address record).
- CNAME (canonical name record).
- CAA (certification authority authorization).
- MX (mail exchange record).
- NAPTR (name authority pointer record).
- NS (name server record).
- PTR (pointer record).
- SOA (start of authority record).
- SPF (sender policy framework).
- SRV (service locator).
- TXT (text record).
- Alias (an Amazon Route 53-specific virtual record).



Section 6: Exam Cram

AWS Route 53 Records

- The Alias record is a Route 53 specific record type.
- The Alias is pointed to the DNS name of the service.



Section 6: Exam Cram

AWS Route 53 CNAME vs Alias



CNAME	Alias
Route 53 charges for CNAME queries	Route 53 doesn't charge for alias queries to AWS resources
You can't create a CNAME record at the top node of a DNS namespace (zone apex)	You can create an alias record at the zone apex (however you can't route to a CNAME at the zone apex)
A CNAME can point to any DNS record that is hosted anywhere	An alias record can only point to a CloudFront distribution, Elastic Beanstalk environment, ELB, S3 bucket as a static website, or to another record in the same hosted zone that you're creating the alias record in

Section 6: Exam Cram

AWS Route 53 Routing Policies



Policy	What it Does
Simple	Simple DNS response providing the IP address associated with a name
Failover	If primary is down (based on health checks), routes to secondary destination
Geolocation	Uses geographic location you're in (e.g. Europe) to route you to the closest region
Geoproximity	Routes you to the closest region within a geographic area
Latency	Directs you based on the lowest latency route to resources
Multivalue answer	Returns several IP addresses and functions as a basic load balancer
Weighted	Uses the relative weights assigned to resources to determine which to route to

Section 6: Exam Cram

AWS Route 53 Routing Policies



- Simple:
 - An A record is associated with one or more IP addresses.
 - Uses round robin.
 - Does not support health checks.
- Failover:
 - Failover to a secondary IP address.
 - Associated with a health check.
 - Used for active-passive.
 - Can be used with ELB.

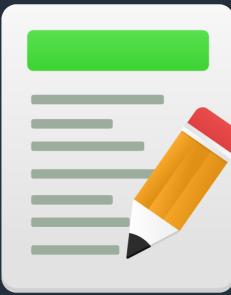
Section 6: Exam Cram

AWS Route 53 Routing Policies



- Geo-location:
 - Caters to different users in different countries and different languages.
 - Contains users within a particular geography and offers them a customized version of the workload based on their specific needs.
 - Geolocation can be used for localizing content and presenting some or all of your website in the language of your users.
 - Can also protect distribution rights.
 - Can be used for spreading load evenly between regions.
 - If you have multiple records for overlapping regions, Route 53 will route to the smallest geographic region.

Section 6: Exam Cram

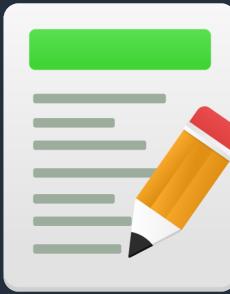


AWS Route 53 Routing Policies

- Geo-proximity routing policy (requires Route Flow):
 - Use for routing traffic based on the location of resources and, optionally, shift traffic from resources in one location to resources in another.
- Latency based routing:
 - AWS maintains a database of latency from different parts of the world.
 - Focussed on improving performance by routing to the region with the lowest latency.
 - You create latency records for your resources in multiple EC2 locations.

Section 6: Exam Cram

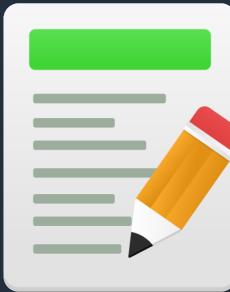
AWS Route 53 Routing Policies



- Multi-value answer routing policy:
 - Use for responding to DNS queries with up to eight healthy records selected at random.
- Weighted:
 - Similar to simple but you can specify a weight per IP address.
 - You create records that have the same name and type and assign each record a relative weight.
 - Numerical value that favours one IP over another.
 - To stop sending traffic to a resource you can change the weight of the record to 0.

Section 6: Exam Cram

AWS Route 53 Traffic Flow

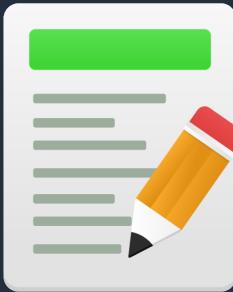


- Route 53 Traffic Flow provides Global Traffic Management (GTM) services.
- Traffic flow policies allow you to create routing configurations for resources using routing types such as failover and geolocation.
- Create policies that route traffic based on specific constraints, including latency, endpoint health, load, geo-proximity and geography.
- Scenarios include:
 - Adding a simple backup page in Amazon S3 for a website.
 - Building sophisticated routing policies that consider an end user's geographic location, proximity to an AWS region, and the health of each of your endpoints.

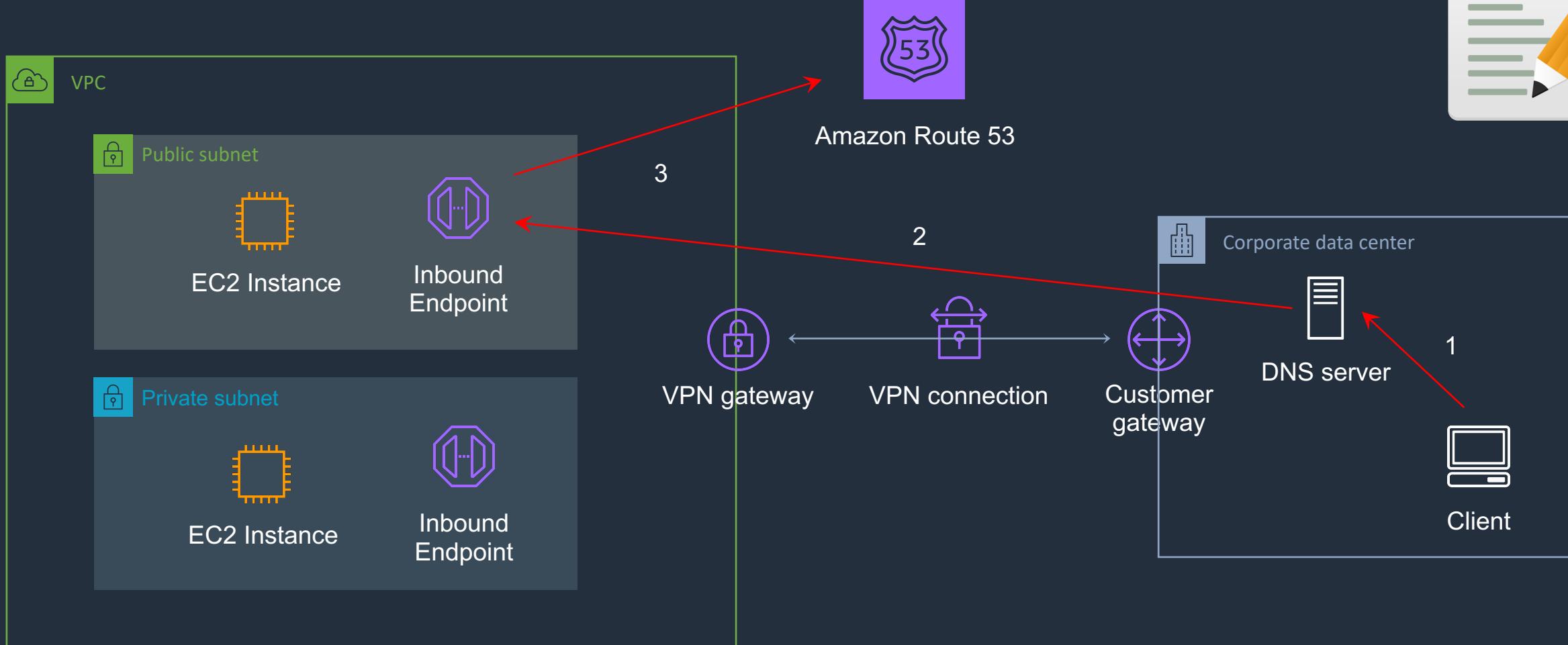
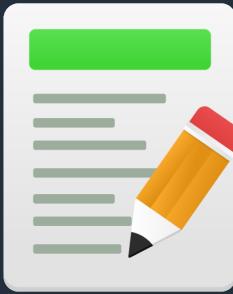
Section 6: Exam Cram

AWS Route 53 Resolver

- Route 53 Resolver is a set of features that enable bi-directional querying between on-premises and AWS over private connections.
- Used for enabling DNS resolution for hybrid clouds.
- Route 53 Resolver Endpoints.
 - Inbound query capability is provided by Route 53 Resolver Endpoints, allowing DNS queries that originate on-premises to resolve AWS hosted domains.
 - Connectivity needs to be established between your on-premises DNS infrastructure and AWS through a Direct Connect (DX) or a Virtual Private Network (VPN).
 - Endpoints are configured through IP address assignment in each subnet for which you would like to provide a resolver.

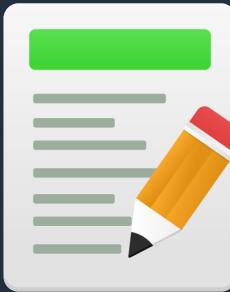


Section 6: Route 53 Resolver – Inbound Endpoints



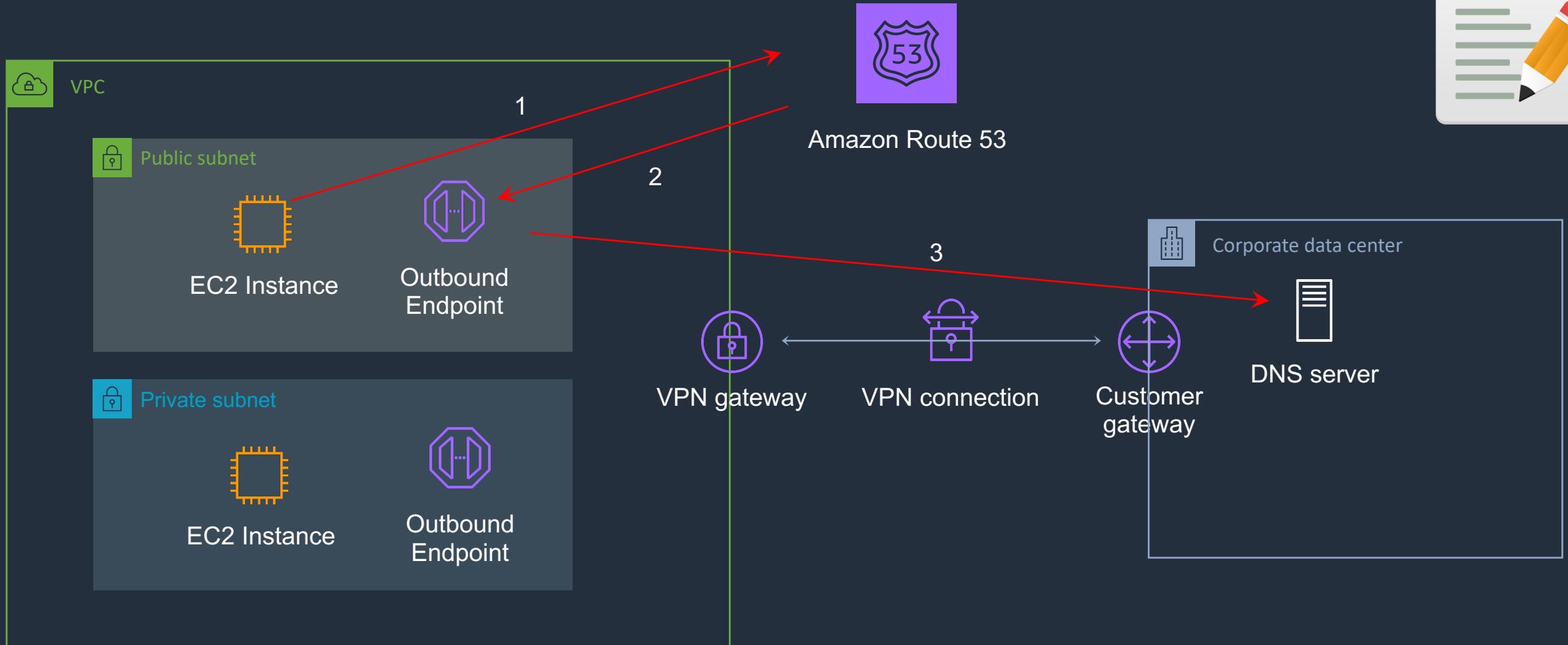
Section 6: Exam Cram

AWS Route 53 Resolver



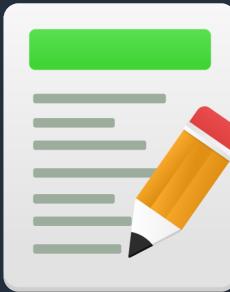
- Conditional forwarding rules:
 - Outbound DNS queries are enabled through the use of Conditional Forwarding Rules.
 - Domains hosted within your on-premises DNS infrastructure can be configured as forwarding rules in Route 53 Resolver.
 - Rules will trigger when a query is made to one of those domains and will attempt to forward DNS requests to your DNS servers that were configured along with the rules.
 - Like the inbound queries, this requires a private connection over DX or VPN.

Section 6: Route 53 Resolver – Outbound Endpoints



Section 6: Exam Cram

AWS Global Accelerator (SAA-C02 exam only)

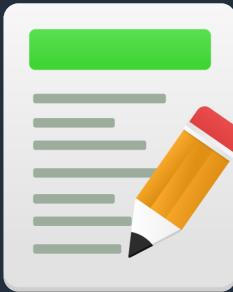


- AWS Global Accelerator is a service that improves the availability and performance of applications with local or global users.
- It provides static IP addresses that act as a fixed entry point to application endpoints in a single or multiple AWS Regions, such as Application Load Balancers, Network Load Balancers or EC2 instances.
- Uses the AWS global network to optimize the path from users to applications, improving the performance of TCP and UDP traffic.
- AWS Global Accelerator continually monitors the health of application endpoints and will detect an unhealthy endpoint and redirect traffic to healthy endpoints in less than 1 minute.

Section 6: Exam Cram

AWS Global Accelerator (SAA-C02 exam only)

- Uses redundant (two) static anycast IP addresses in different network zones (A and B).
- The redundant pair are globally advertized.
- Uses AWS Edge Locations – addresses are announced from multiple edge locations at the same time.
- Addresses are associated to regional AWS resources or endpoints.
- AWS Global Accelerator's IP addresses serve as the frontend interface of applications.
- Intelligent traffic distribution: Routes connections to the closest point of presence for applications.
- Targets can be Amazon EC2 instances or Elastic Load Balancers (ALB and NLB).



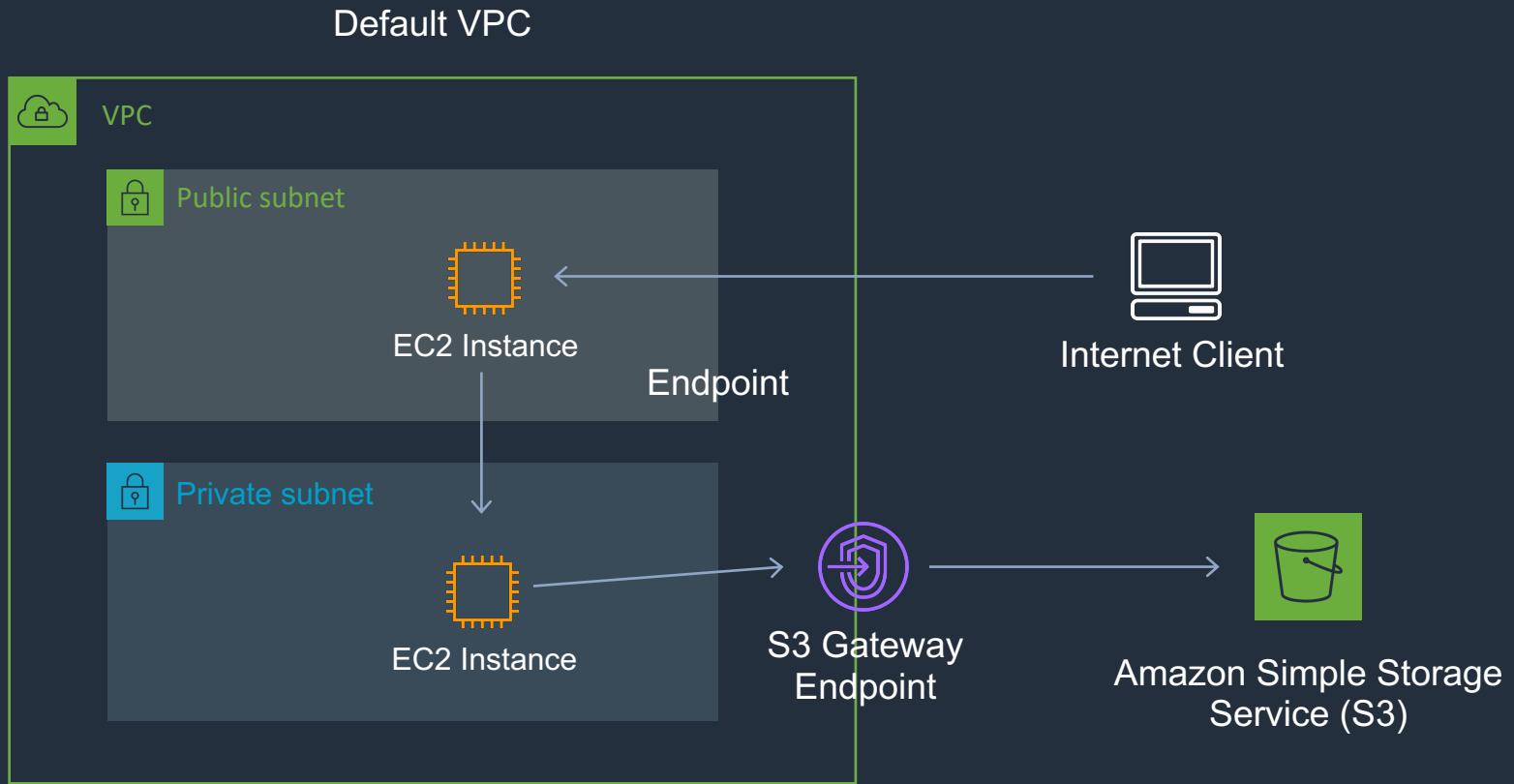
Section 6: Exam Cram

AWS Global Accelerator (SAA-C02 exam only)



- By using the static IP addresses, you don't need to make any client-facing changes or update DNS records as you modify or replace endpoints.
- The addresses are assigned to your accelerator for as long as it exists, even if you disable the accelerator and it no longer accepts or routes traffic.
- Uses the vast, congestion-free AWS global network to route TCP and UDP traffic to a healthy application endpoint in the closest AWS Region to the user.
- If there's an application failure, AWS Global Accelerator provides instant failover to the next best endpoint.

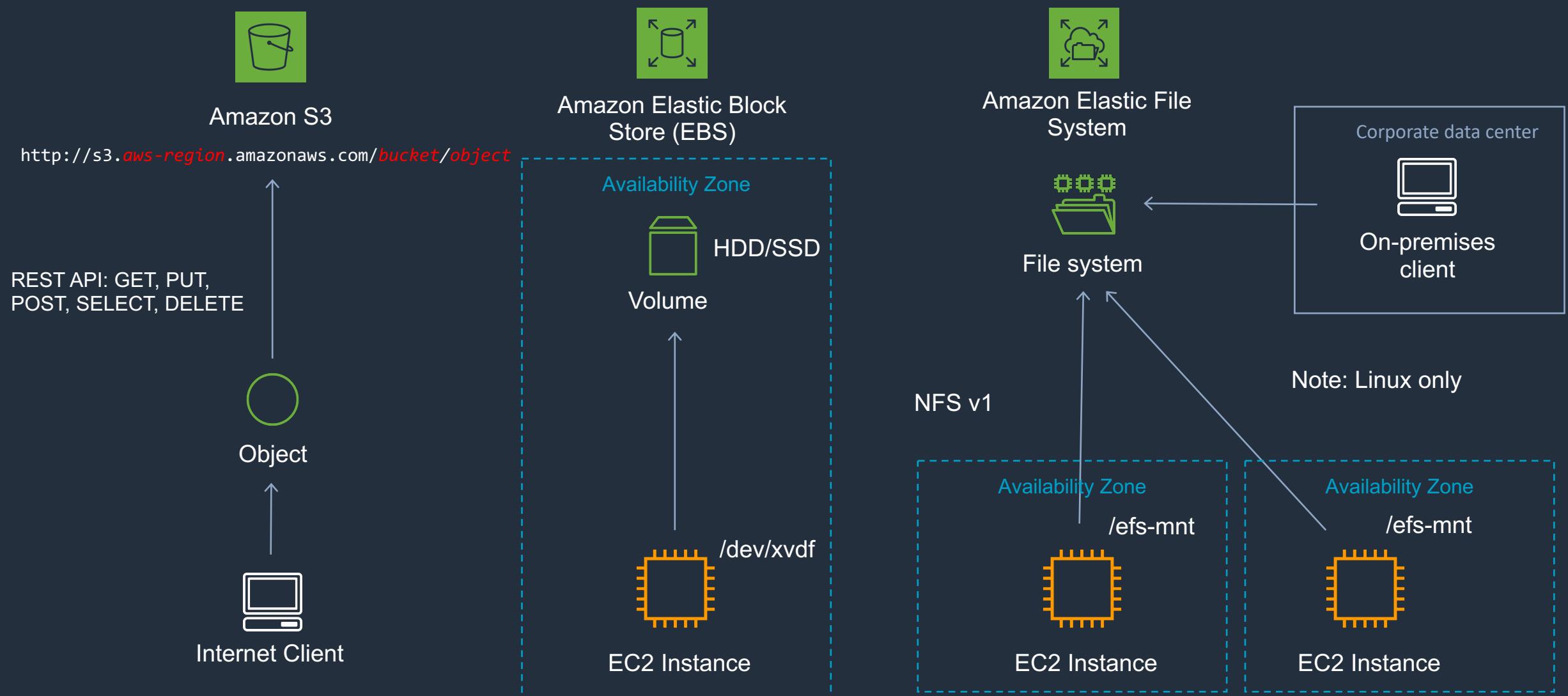
Section 7: S3 Gateway Endpoints



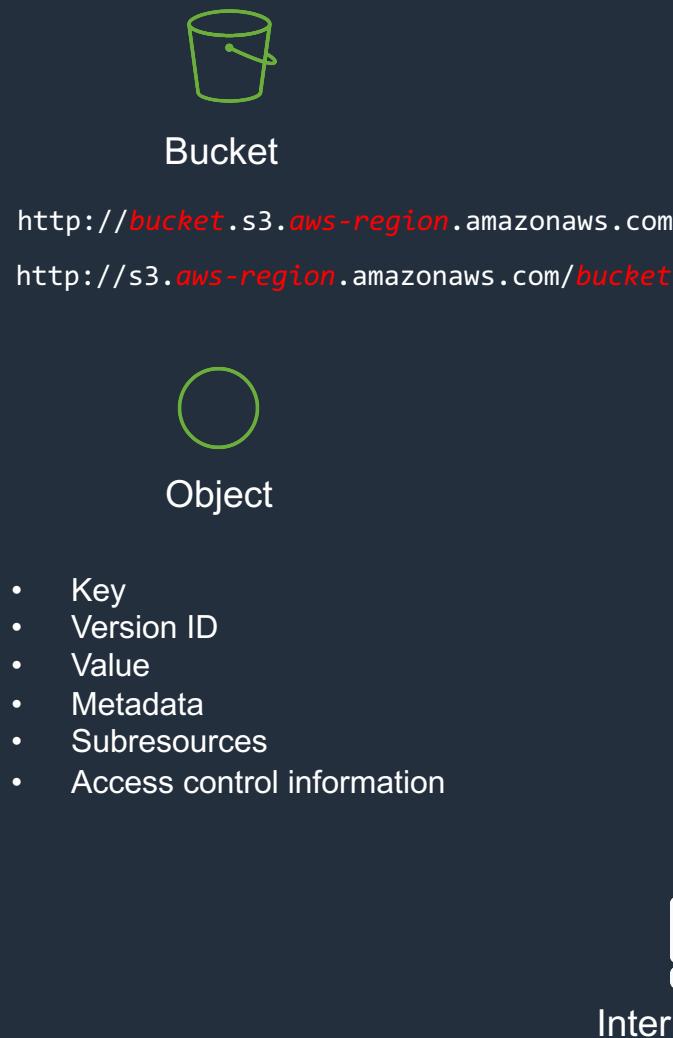
Route Table

Destination	Target
<code>pl-6ca54005 (com.amazonaws.ap-southeast-2.s3, 54.231.248.0/22, 54.231.252.0/24, 52.95.128.0/21)</code>	<code>vpce-ID</code>

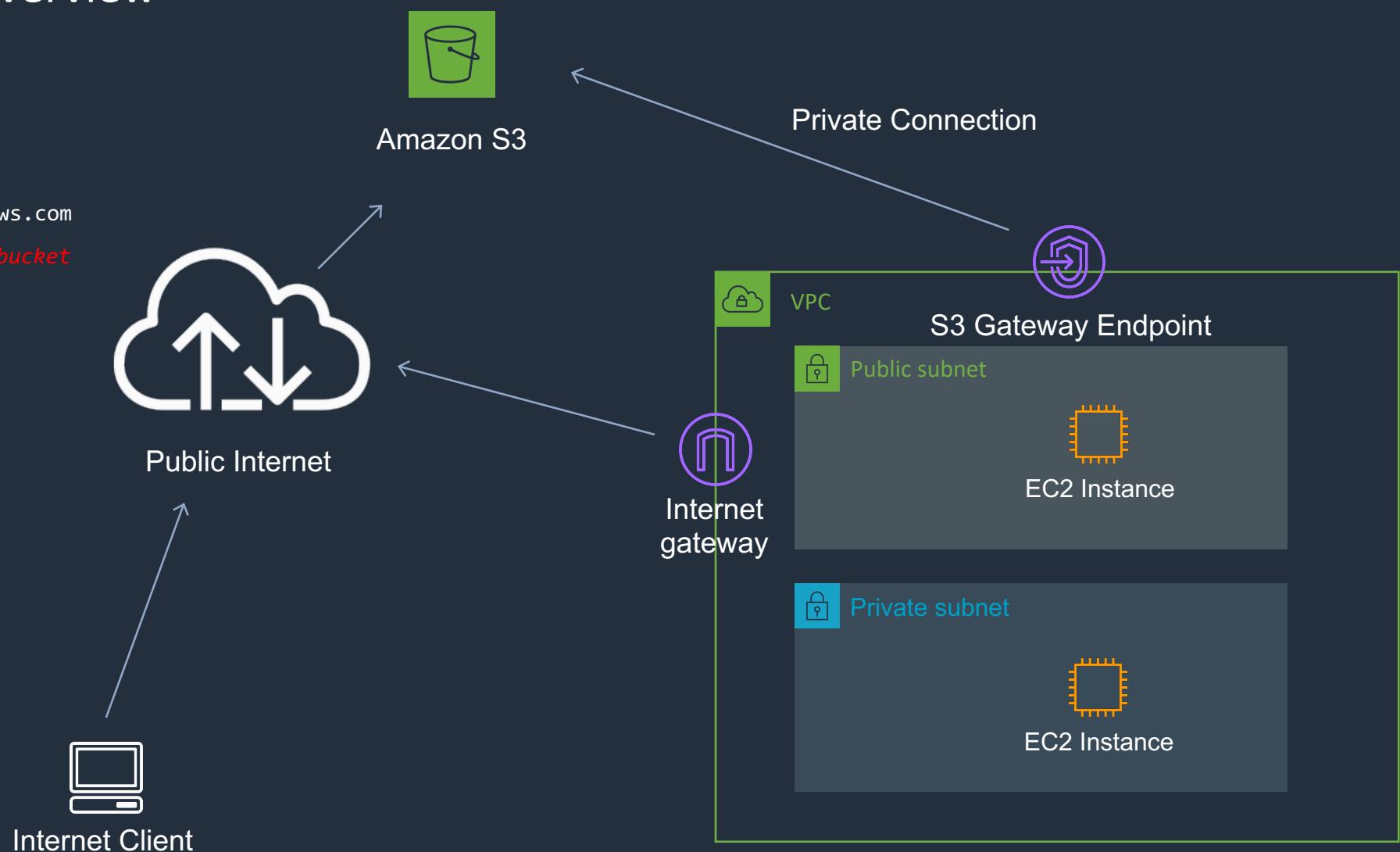
Section 7: Block, Object and File Storage



Section 7: Amazon S3 Overview



- Key
- Version ID
- Value
- Metadata
- Subresources
- Access control information



Section 7: Identity-Based and Resource-Based Policies

Example Policy

```
{  
  "Version": "2012-10-17",  
  "Statement": [  
    {  
      "Sid": "SeeBucketListInTheConsole",  
      "Action": ["s3>ListAllMyBuckets"],  
      "Effect": "Allow",  
      "Resource": ["arn:aws:s3:::*"]  
    }  
  ]  
}
```

Identity-based policies



IAM Role Inline Policy

Resource-based policy



Bucket Policy

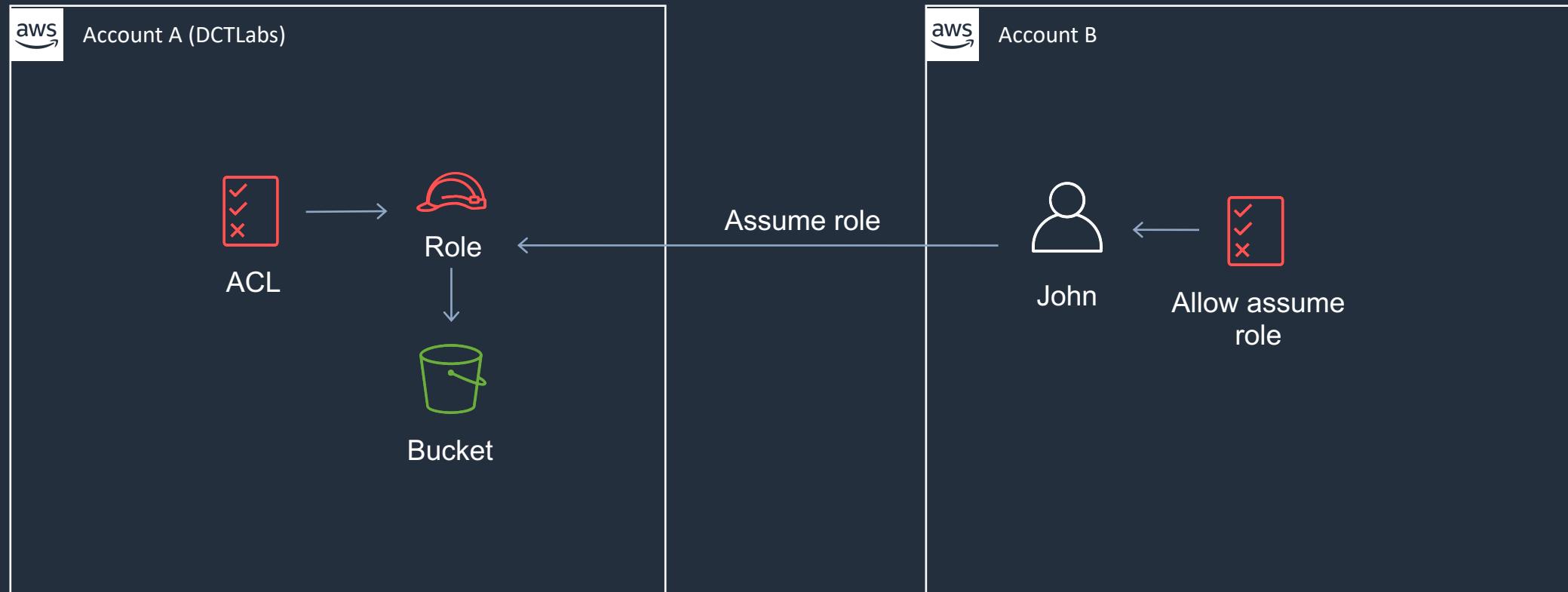


IAM User Inline Policy



IAM Group Policy

Section 7: Cross-Account Access

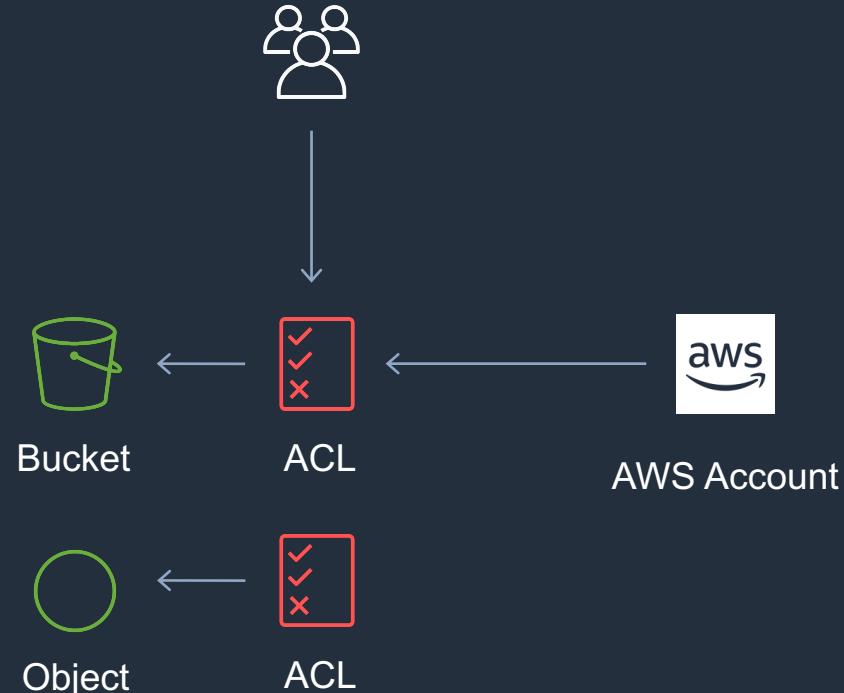


Section 7: Access Control Lists

Example ACL

```
... <AccessControlPolicy>
<Owner>
  <ID> AccountACanonicalUserID </ID>
  <DisplayName> AccountADisplayName </DisplayName>
</Owner>
<AccessControlList>
...
  <Grant>
    <Grantee xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
xsi:type="CanonicalUser">
      <ID> AccountBCanonicalUserID </ID>
      <DisplayName> AccountBDisplayName </DisplayName>
    </Grantee>
    <Permission> WRITE </Permission>
  </Grant>
...
</AccessControlList>
</AccessControlPolicy>
```

- S3 Predefined Group
- Authenticated Users
 - All Users
 - Log Delivery Group



Section 7: Access Control List Permissions

Permissions	When granted on a bucket	When granted on an object
READ	Allows grantee to list the objects in the bucket	Allows grantee to read the object data and its metadata
WRITE	Allows grantee to create, overwrite, and delete any object in the bucket	Not applicable
READ_ACP	Allows grantee to read the bucket ACL	Allows grantee to read the object ACL
WRITE_ACP	Allows grantee to write the ACL for the applicable bucket	Allows grantee to write the ACL for the applicable object
FULL_CONTROL	Allows grantee the READ, WRITE, READ_ACP, and WRITE_ACP permissions on the bucket	Allows grantee the READ, READ_ACP, and WRITE_ACP permissions on the object

Section 7: Choosing Access Control Options

Identity-based policies

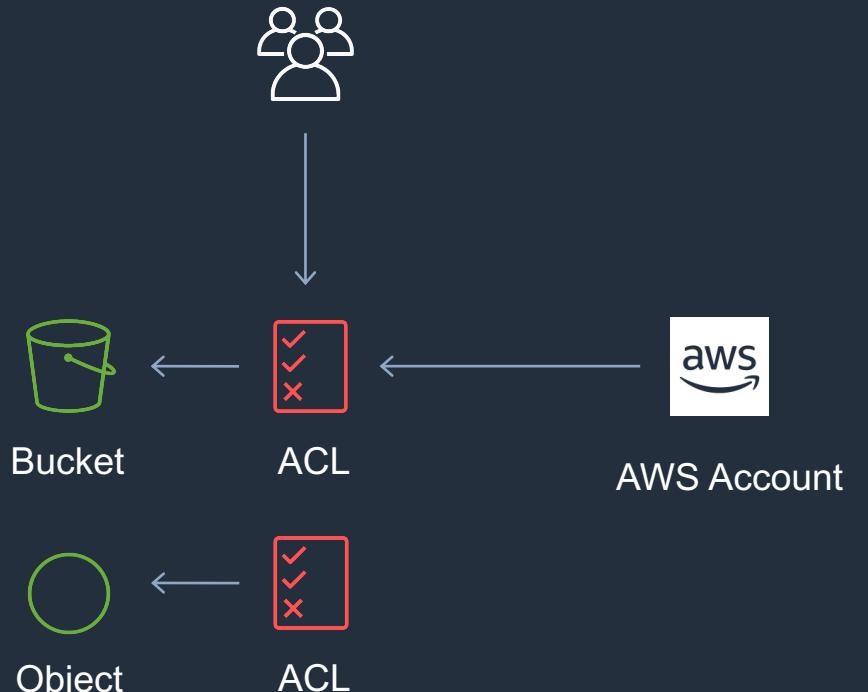


Resource-based policy

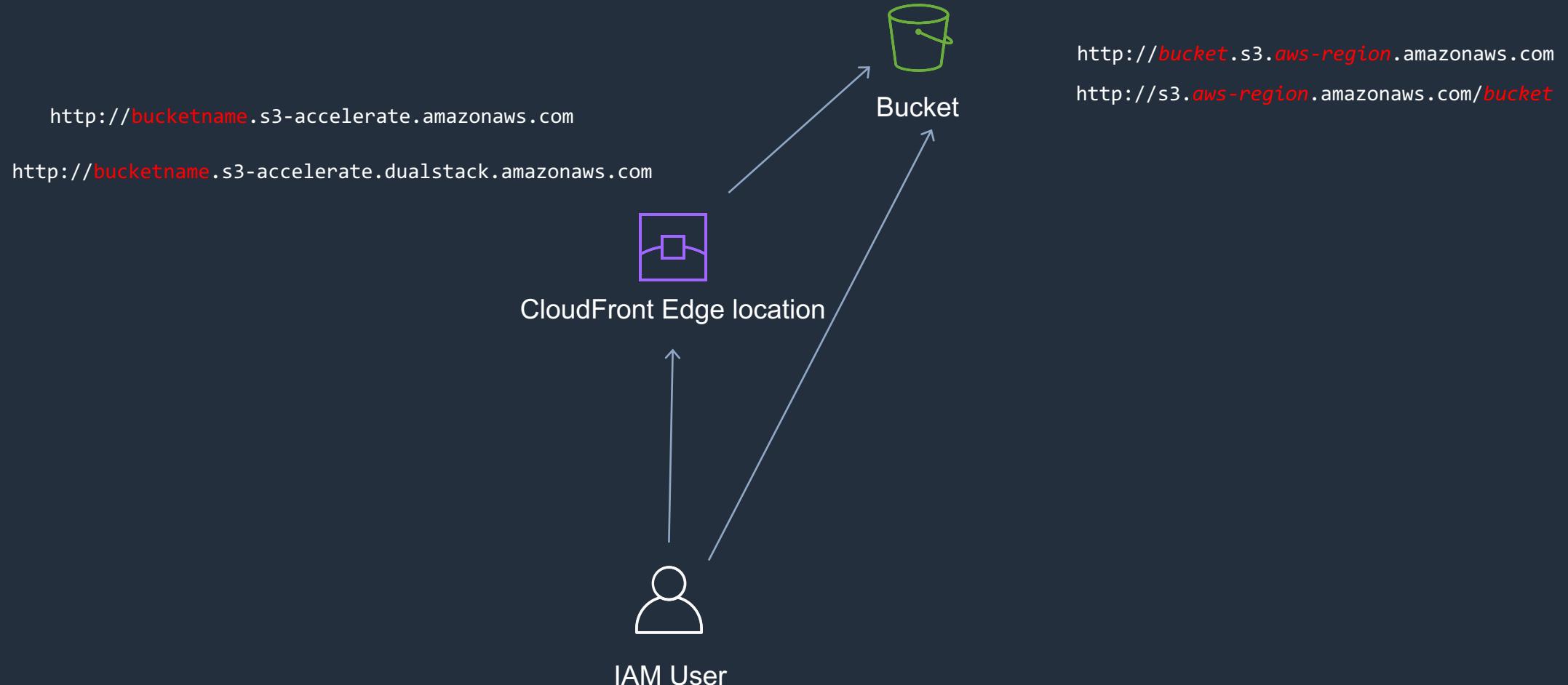


S3 Predefined Group

- Authenticated Users
- All Users
- Log Delivery Group



Section 7: Transfer Acceleration



Section 7: S3 Encryption

Server-side encryption with S3 managed keys (SSE-S3)

- S3 managed keys
- Unique object keys
- Master key
- AES 256



Encryption / decryption



Server-side encryption with AWS KMS managed keys (SSE-KMS)



- KMS managed keys
- Customer master keys
- CMK can be customer generated



Encryption / decryption



Server-side encryption with client provided keys (SSE-C)



Encryption / decryption



- Client managed keys
- Not stored on AWS

Client side encryption

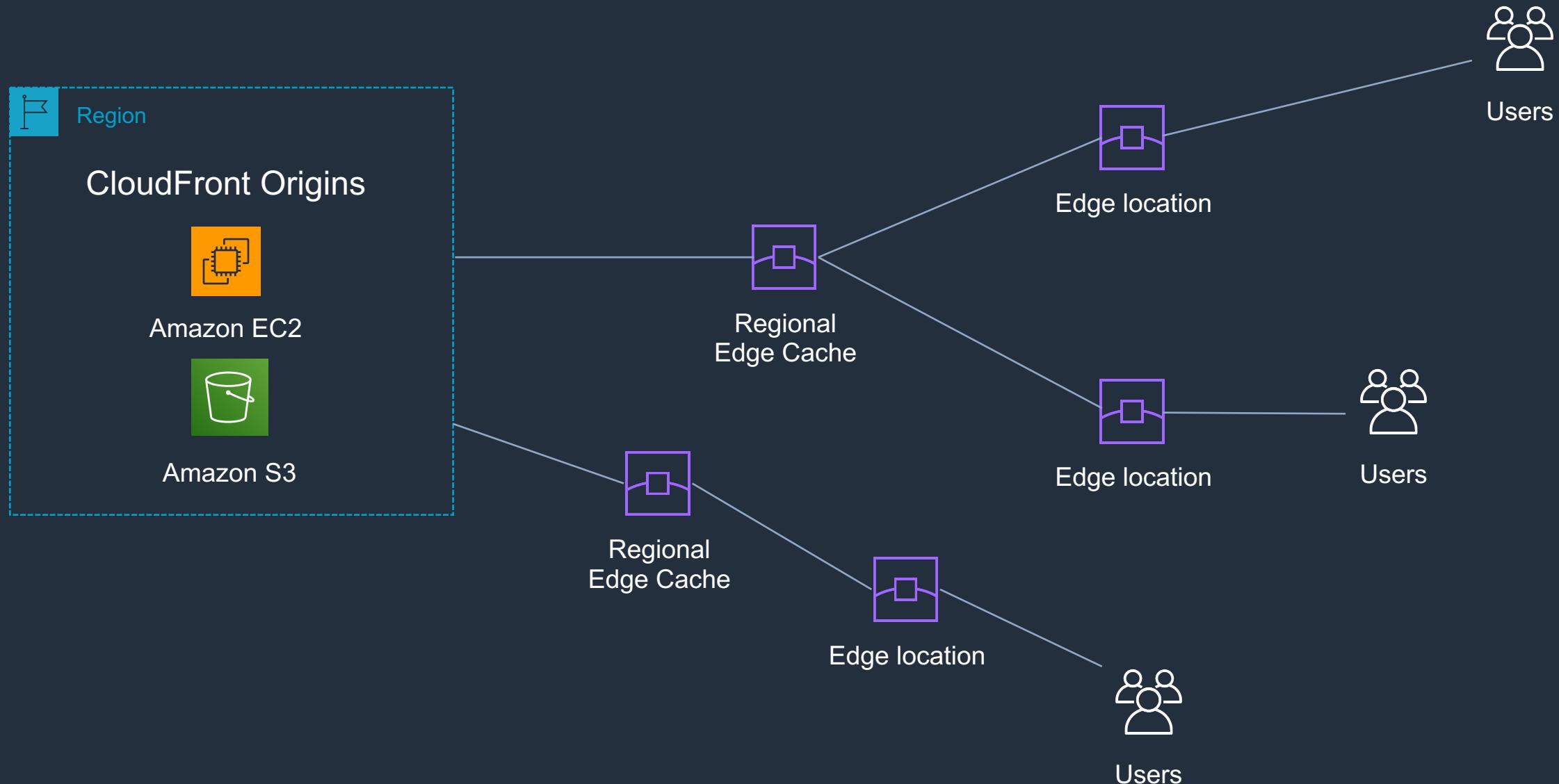


Encryption / decryption

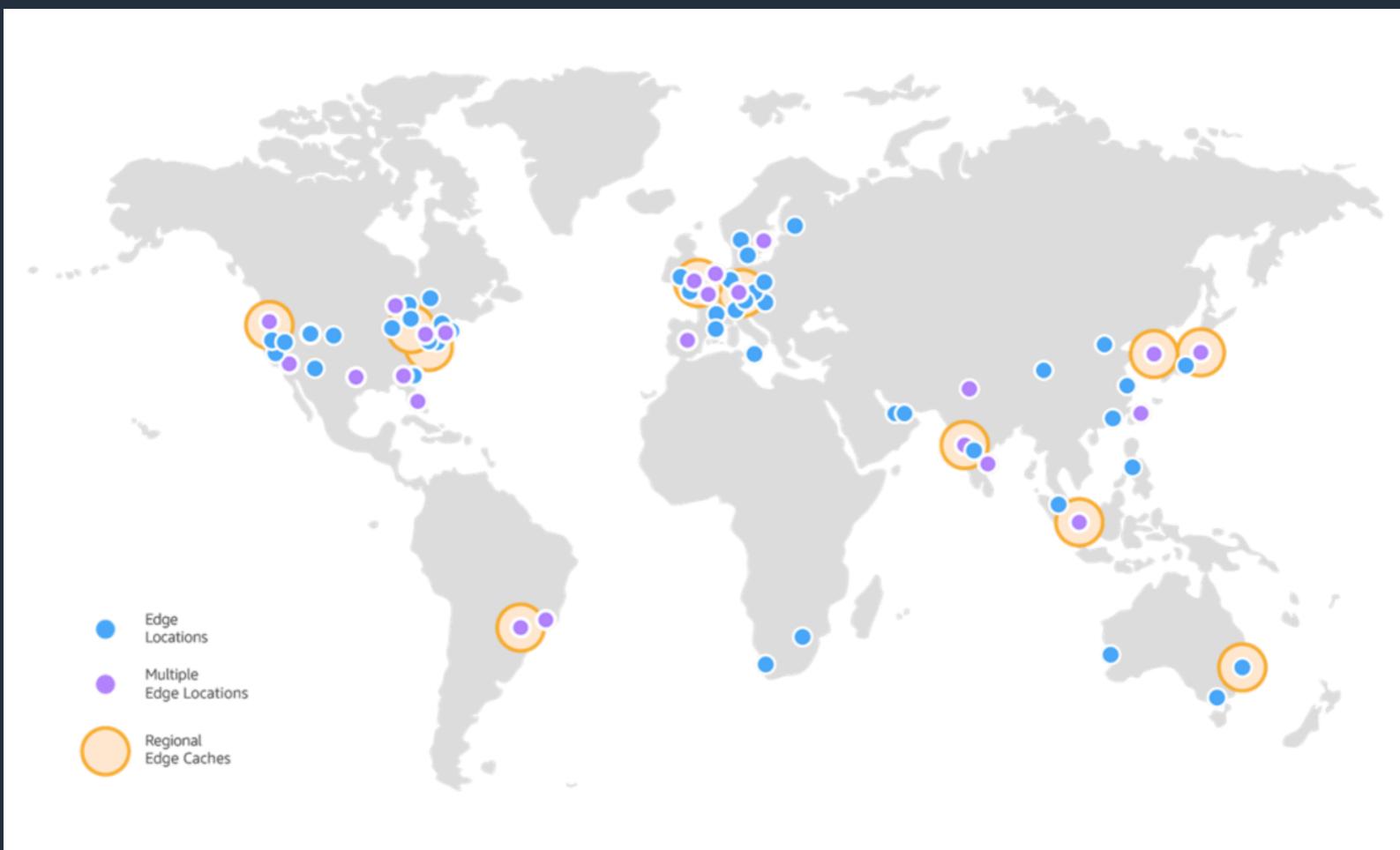


- Client managed keys
- Not stored on AWS

Section 7: CloudFront Overview

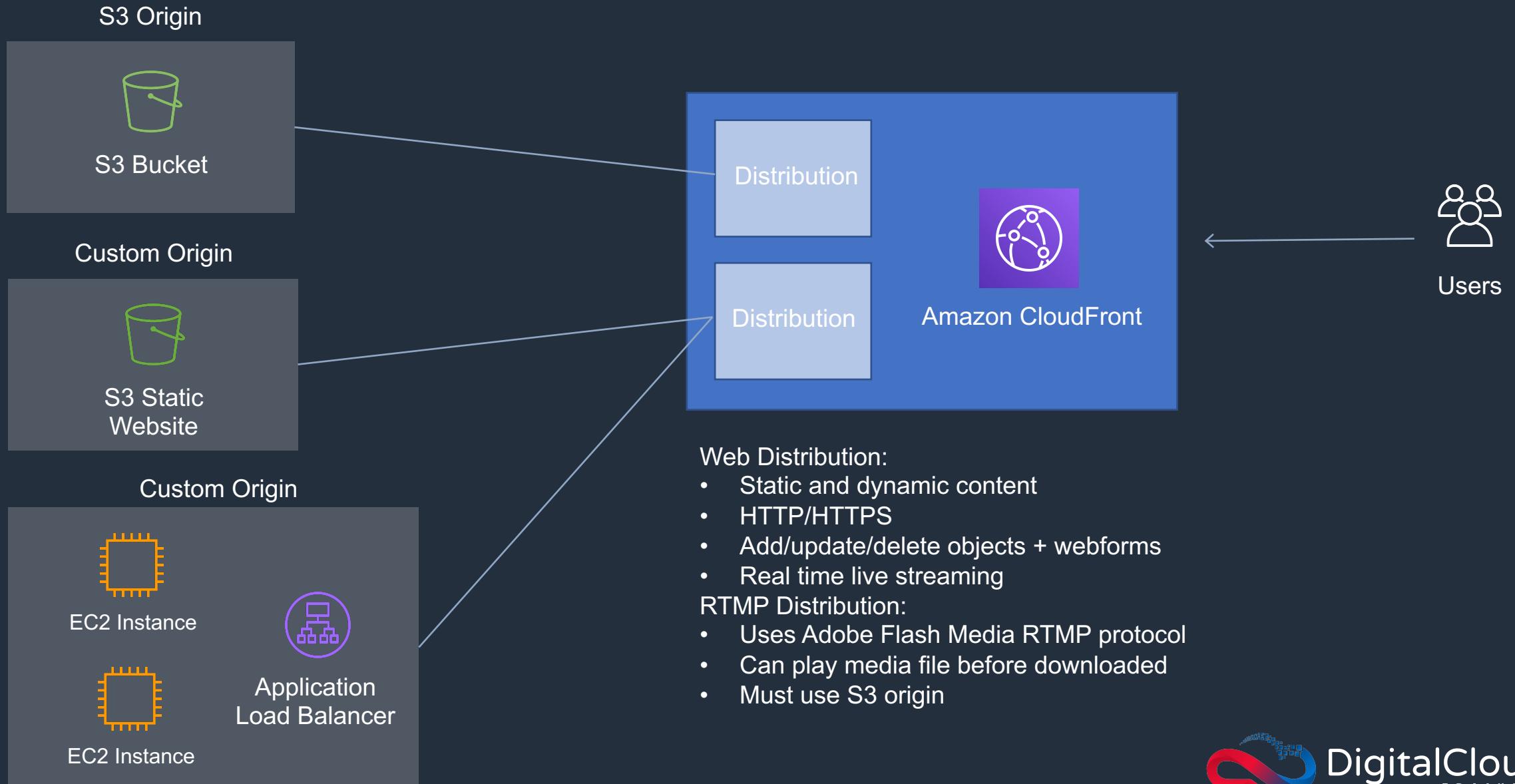


Section 7: CloudFront – Points of Presence

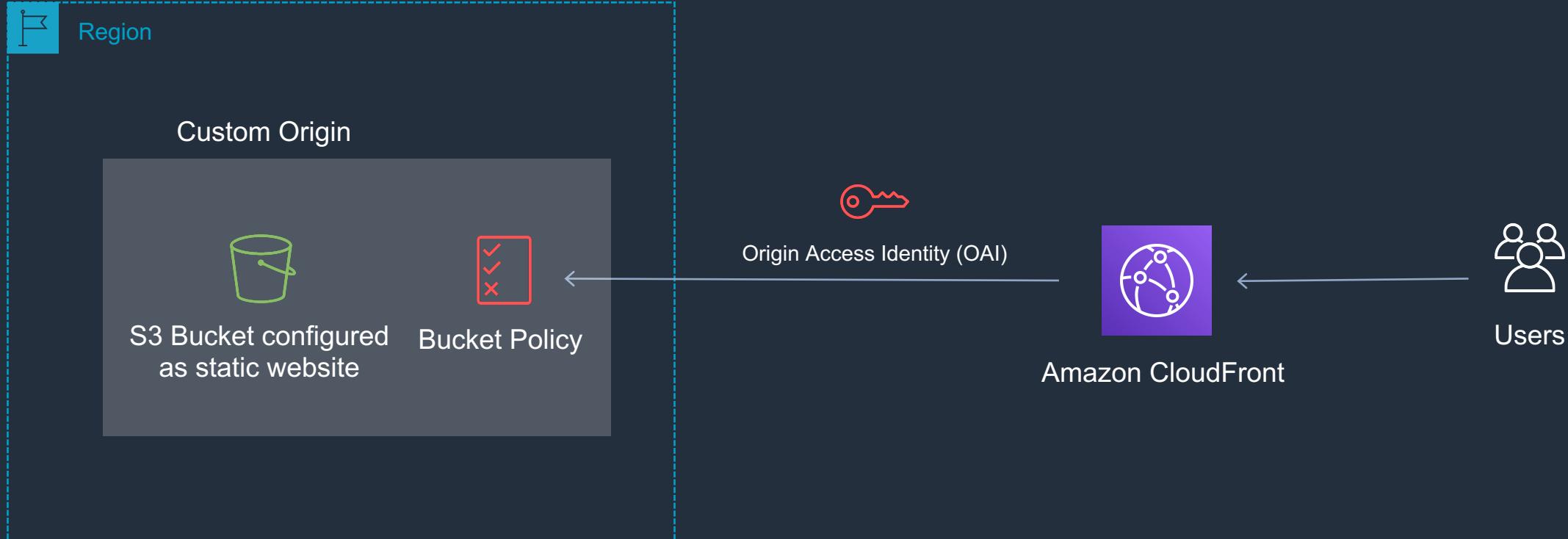


- Points of Presence:
- 176 Edge Locations
 - 11 Regional Edge Caches
 - 69 cities
 - 30 countries

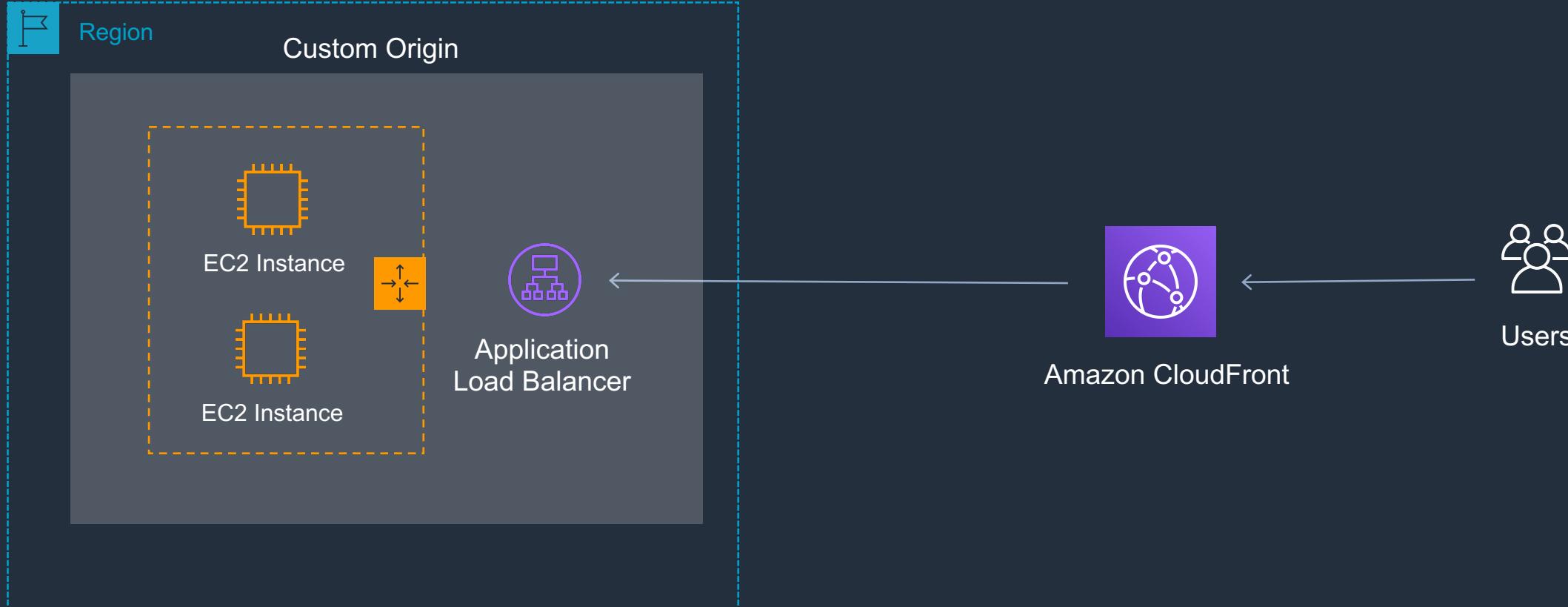
Section 7: CloudFront Distribution and Origins



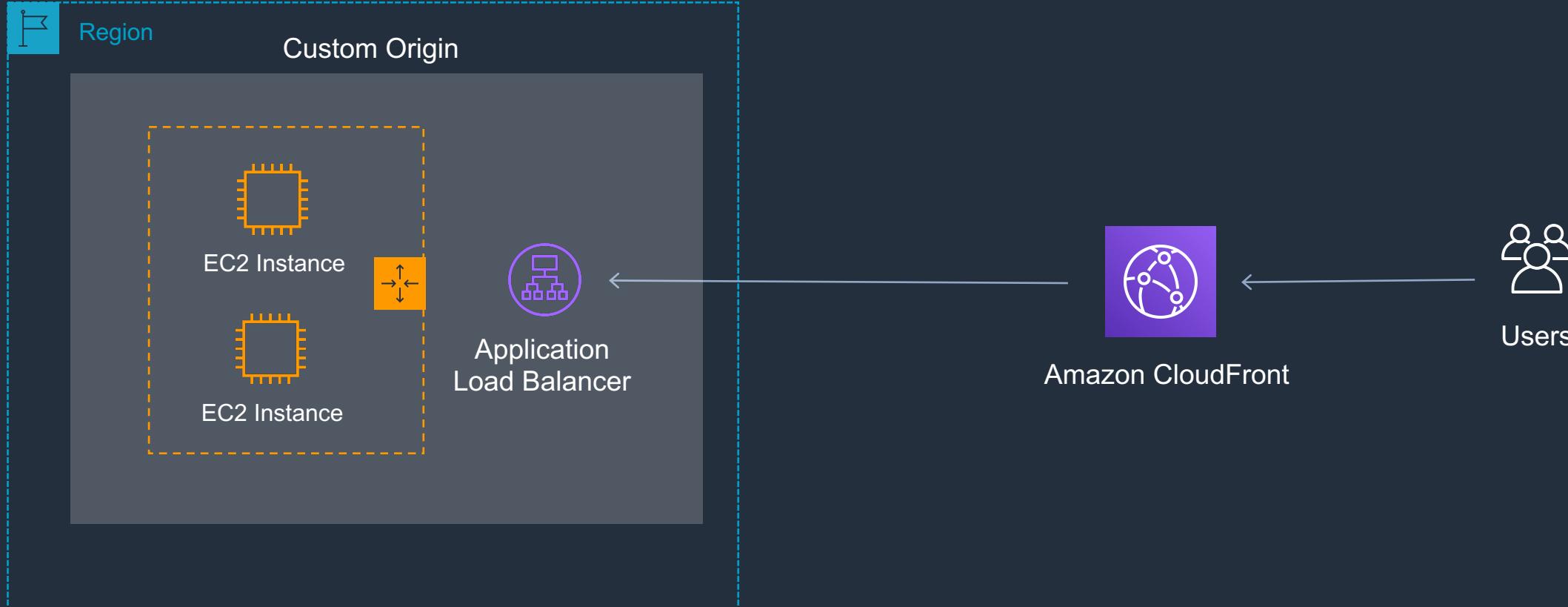
Section 7: CloudFront with S3 Static Website



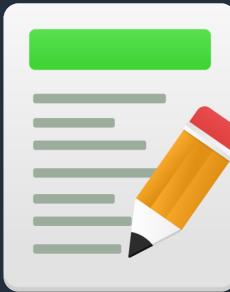
Section 7: CloudFront with ALB and EC2 Custom Origin



Section 7: CloudFront with Lambda@Edge



Section 7: Exam Cram



Amazon S3

- Amazon S3 is a distributed architecture and objects are redundantly stored on multiple devices across multiple facilities (AZs) in an Amazon S3 region.
- Amazon S3 is a simple key-based object store.
- Amazon S3 provides a simple, standards-based REST web services interface that is designed to work with any Internet-development toolkit.
- Files can be from 0 bytes to 5TB.
- The largest object that can be uploaded in a single PUT is 5 gigabytes.
- For objects larger than 100 megabytes use the Multipart Upload capability.

Section 7: Exam Cram



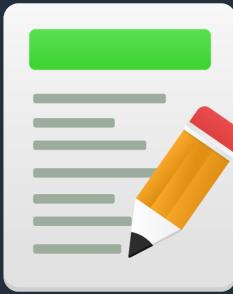
Amazon S3

- Event notifications for specific actions, can send alerts or trigger actions.
- Notifications can be sent to:
 - SNS Topics.
 - SQS Queue.
 - Lambda functions.
 - Need to configure SNS/SQS/Lambda before S3.
 - No extra charges from S3 but you pay for SNS, SQS and Lambda.
- Provides read after write consistency for PUTS of new objects.
- Provides eventual consistency for overwrite PUTS and DELETES (takes time to propagate).

Section 7: Exam Cram

Amazon S3

- S3 data is made up of:
 - Key (name).
 - Value (data).
 - Version ID.
 - Metadata.
 - Access Control Lists.



Section 7: Exam Cram

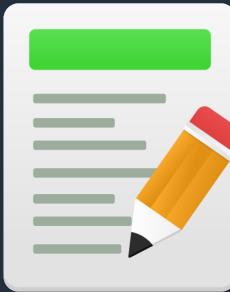
Amazon S3 Additional Capabilities



Additional S3 Capability	How it Works
Transfer Acceleration	Speed up data uploads using CloudFront in reverse
Requester Pays	The requester rather than the bucket owner pays for requests and data transfer
Tags	Assign tags to objects to use in costing, billing, security etc.
Events	Trigger notifications to SNS, SQS, or Lambda when certain events happen in your bucket
Static Web Hosting	Simple and massively scalable static website hosting
BitTorrent	Use the BitTorrent protocol to retrieve any publicly available object by automatically generating a .torrent file

Section 7: Exam Cram

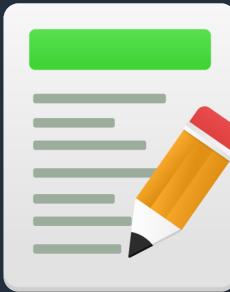
Amazon S3



- Typical use cases include:
 - **Backup and Storage** – Provide data backup and storage services for others.
 - **Application Hosting** – Provide services that deploy, install, and manage web applications.
 - **Media Hosting** – Build a redundant, scalable, and highly available infrastructure that hosts video, photo, or music uploads and downloads.
 - **Software Delivery** – Host your software applications that customers can download.
 - **Static Website** - you can configure a static website to run from an S3 bucket.

Section 7: Exam Cram

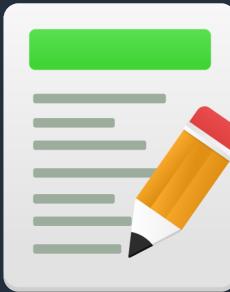
Amazon S3 Buckets



- Files are stored in buckets:
 - A bucket can be viewed as a container for objects.
 - A bucket is a flat container of objects.
 - It does not provide a hierarchy of objects.
 - You can use an object key name (prefix) to mimic folders.
- 100 buckets per account by default.
- You can store unlimited objects in your buckets.
- You can create folders in your buckets (only available through the Console).
- You cannot create nested buckets.
- An S3 bucket is region specific.

Section 7: Exam Cram

Amazon S3 Objects

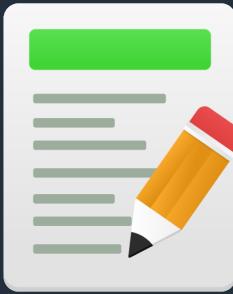


- Each object is stored and retrieved by a unique key (ID or name).
- An object in S3 is uniquely identified and addressed through:
 - Service end-point.
 - Bucket name.
 - Object key (name).
 - Optionally, an object version.
- Objects stored in a bucket will never leave the region in which they are stored unless you move them to another region or enable cross-region replication.
- You can define permissions on objects when uploading and at any time afterwards using the AWS Management Console.

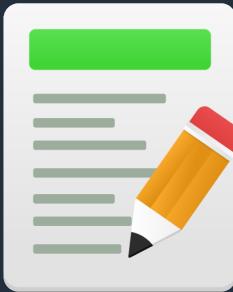
Section 7: Exam Cram

Amazon S3 Sub-resources

- Sub-resources (configuration containers) associated with buckets include:
 - Lifecycle - define an object's lifecycle.
 - Website - configuration for hosting static websites.
 - Versioning - retain multiple versions of objects as they are changed.
 - Access Control Lists (ACLs) - control permissions access to the bucket.
 - Bucket Policies - control access to the bucket.
 - Cross Origin Resource Sharing (CORS).
 - Logging.
- **NOTE:** Please check training notes / ebook for details of the above



Section 7: Exam Cram

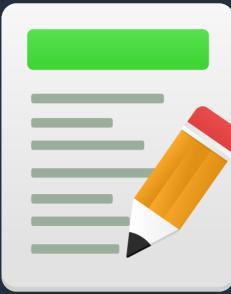


Amazon S3 Storage Classes

- There are six S3 storage classes.
 - S3 Standard (durable, immediately available, frequently accessed).
 - S3 Intelligent-Tiering (automatically moves data to the most cost-effective tier).
 - S3 Standard-IA (durable, immediately available, infrequently accessed).
 - S3 One Zone-IA (lower cost for infrequently accessed data with less resilience).
 - S3 Glacier (archived data, retrieval times in minutes or hours).
 - S3 Glacier Deep Archive (lowest cost storage class for long term retention).

Section 7: Exam Cram

Amazon S3 Storage Classes

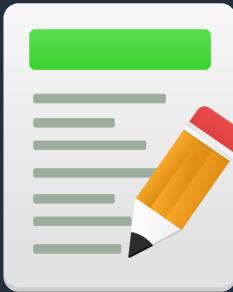


	S3 Standard	S3 Intelligent-Tiering*	S3 Standard-IA	S3 One Zone-IA†	S3 Glacier	S3 Glacier Deep Archive
Designed for durability	99.999999999% (11 9's)					
Designed for availability	99.99%	99.9%	99.9%	99.5%	99.99%	99.99%
Availability SLA	99.9%	99%	99%	99%	99.9%	99.9%
Availability Zones	≥3	≥3	≥3	1	≥3	≥3
Minimum capacity charge per object	N/A	N/A	128KB	128KB	40KB	40KB
Minimum storage duration charge	N/A	30 days	30 days	30 days	90 days	180 days
Retrieval fee	N/A	N/A	per GB retrieved	per GB retrieved	per GB retrieved	per GB retrieved
First byte latency	milliseconds	milliseconds	milliseconds	milliseconds	select minutes or hours	select hours
Storage type	Object	Object	Object	Object	Object	Object
Lifecycle transitions	Yes	Yes	Yes	Yes	Yes	Yes

Section 7: Exam Cram

Amazon S3 Multipart upload

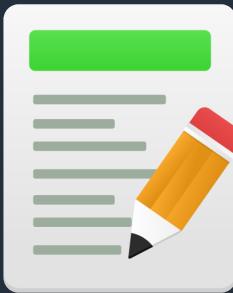
- Multipart upload uploads objects in parts independently, in parallel and in any order.
- Performed using the S3 Multipart upload API.
- It is recommended for objects of 100MB or larger.
- Can be used for objects from 5MB up to 5TB.
- Must be used for objects larger than 5GB.



Section 7: Exam Cram

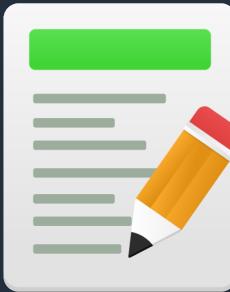
Amazon S3 Copy

- You can create a copy of objects up to 5GB in size in a single atomic operation.
- For files larger than 5GB you must use the multipart upload API.
- Can be performed using the AWS SDKs or REST API.
- The copy operation can be used to:
 - Generate additional copies of objects.
 - Renaming objects.
 - Changing the copy's storage class or encryption at rest status.
 - Move objects across AWS locations/regions.
 - Change object metadata.



Section 7: Exam Cram

Amazon S3 Transfer Acceleration



- Amazon S3 Transfer Acceleration enables fast, easy, and secure transfers of files over long distances between your client and your Amazon S3 bucket.
- S3 Transfer Acceleration leverages Amazon CloudFront's globally distributed AWS Edge Locations.
- Used to accelerate object uploads to S3 over long distances (latency).
- Transfer acceleration is as secure as a direct upload to S3.
- You are charged only if there was a benefit in transfer times.
- Need to enable transfer acceleration on the S3 bucket.
- Cannot be disabled, can only be suspended.

Section 7: Exam Cram

Amazon S3 Encryption

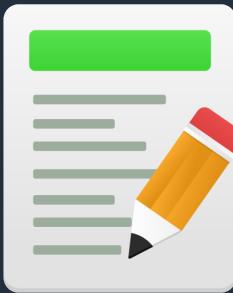


Encryption Option	How it Works
SSE-S3	Use S3's existing encryption key for AES-256
SSE-C	Upload your own AES-256 encryption key which S3 uses when it writes objects
SSE-KMS	Use a key generated and managed by AWS KMS
Client-Side	Encrypt objects using your own local encryption process before uploading to S3

Section 7: Exam Cram

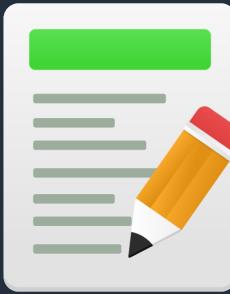
Amazon S3 Performance

- Measure Performance.
- Scale Storage Connections Horizontally.
- Use Byte-Range Fetches.
- Retry Requests for Latency-Sensitive Applications.
- Combine Amazon S3 (Storage) and Amazon EC2 (Compute) in the Same AWS Region.
- Use Amazon S3 Transfer Acceleration to Minimize Latency Caused by Distance.
- **NOTE:** Check the training notes and AWS documentation for details.



Section 7: Exam Cram

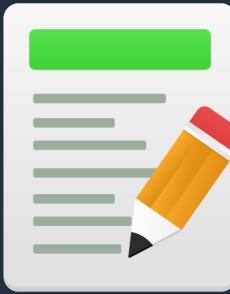
Amazon CloudFront



- CloudFront is a web service that distributes content with low latency and high data transfer speeds.
- Used for dynamic, static, streaming, and interactive content.
- CloudFront is a global service:
 - Ingress to upload objects.
 - Egress to distribute content.
- You can use a zone apex DNS name on CloudFront.
- CloudFront supports wildcard CNAME.
- Supports wildcard SSL certificates, Dedicated IP, Custom SSL and SNI Custom SSL (cheaper).

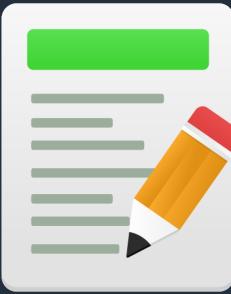
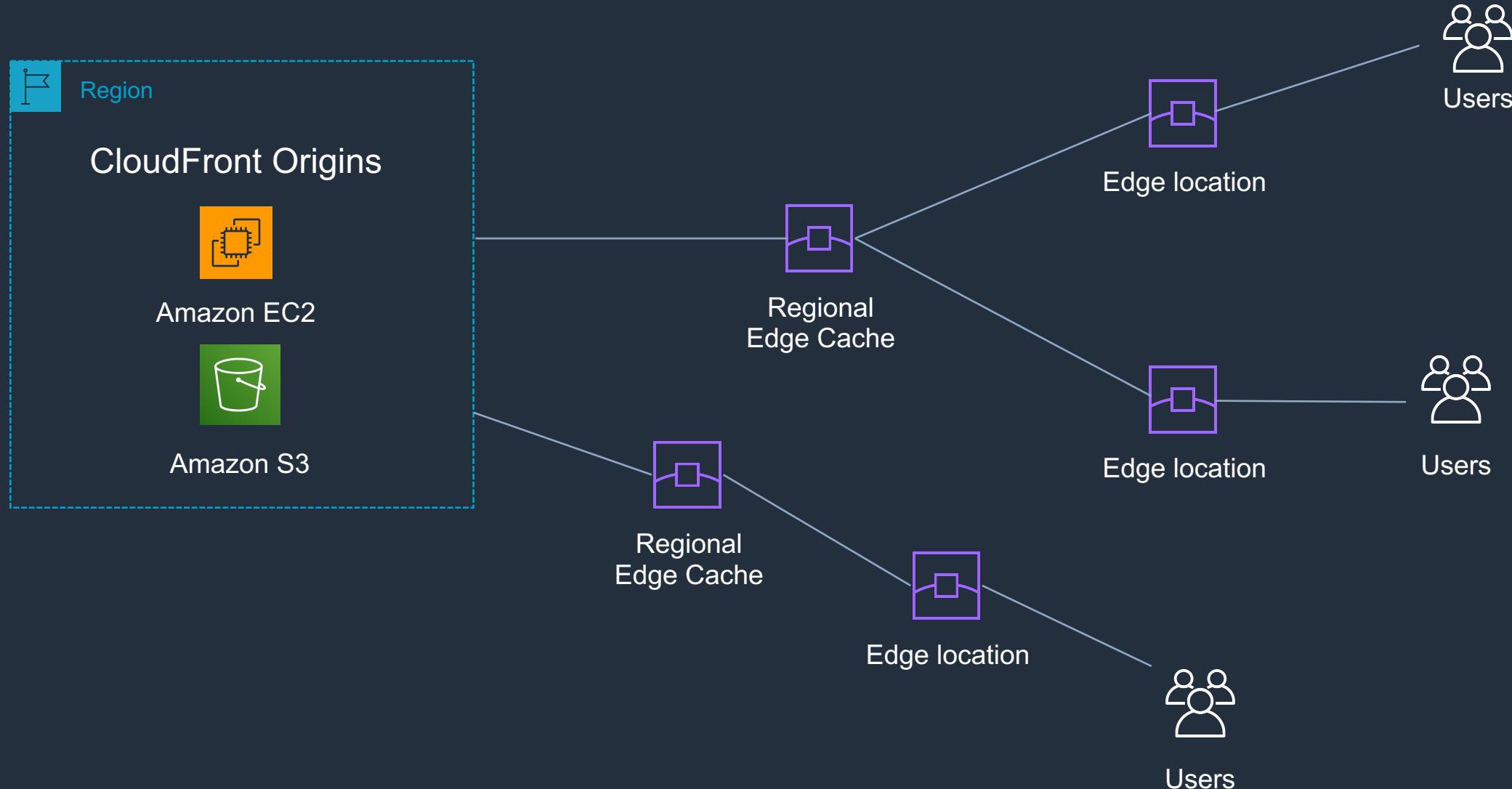
Section 7: Exam Cram

Amazon CloudFront Edge Locations and Regional Edge Caches



- An edge location is the location where content is cached
- Requests are automatically routed to the nearest edge location.
- Edge locations are not tied to Availability Zones or regions.
- Regional Edge Caches are located between origin web servers and global edge locations and have a larger cache.
- Regional Edge Caches have larger cache-width than any individual edge location, so your objects remain in cache longer at these locations.
- Regional Edge caches aim to get content closer to users.

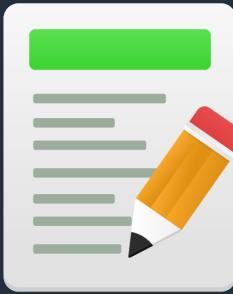
Section 7: CloudFront Overview



Section 7: Exam Cram

Amazon CloudFront Origins

- An origin is the origin of the files that the CDN will distribute.
- Origins can be either an S3 bucket, an EC2 instance, an Elastic Load Balancer, or Route 53 – can also be external (non-AWS).
- A custom origin server is a HTTP server which can be an EC2 instance or an on-premise/non-AWS based web server.
- Amazon EC2 instances are considered custom origins.
- Static websites on Amazon S3 are also considered custom origins.



Section 7: Exam Cram



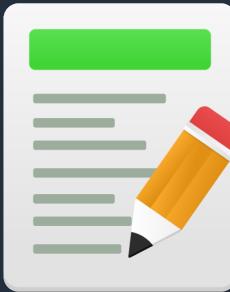
Amazon CloudFront Distributions

- There are two types of distribution.
- Web Distribution:
 - Static and dynamic content including .html, .css, .php, and graphics files.
 - Distributes files over HTTP and HTTPS.
 - Add, update, or delete objects, and submit data from web forms.
 - Use live streaming to stream an event in real time.
- RTMP:
 - Distribute streaming media files using Adobe Flash Media Server's RTMP protocol.
 - Allows an end user to begin playing a media file before the file has finished downloading from a CloudFront edge location.
 - Files must be stored in an S3 bucket.

Section 7: Exam Cram

Amazon CloudFront

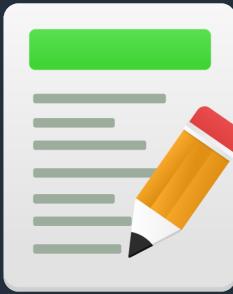
- You can restrict access to content using the following methods:
 - Restrict access to content using signed cookies or signed URLs.
 - Restrict access to objects in your S3 bucket.
- A special type of user called an Origin Access Identity (OAI) can be used to restrict access to content in an Amazon S3 bucket.
- By using an OAI you can restrict users so they cannot access the content directly using the S3 URL, they must connect via CloudFront.



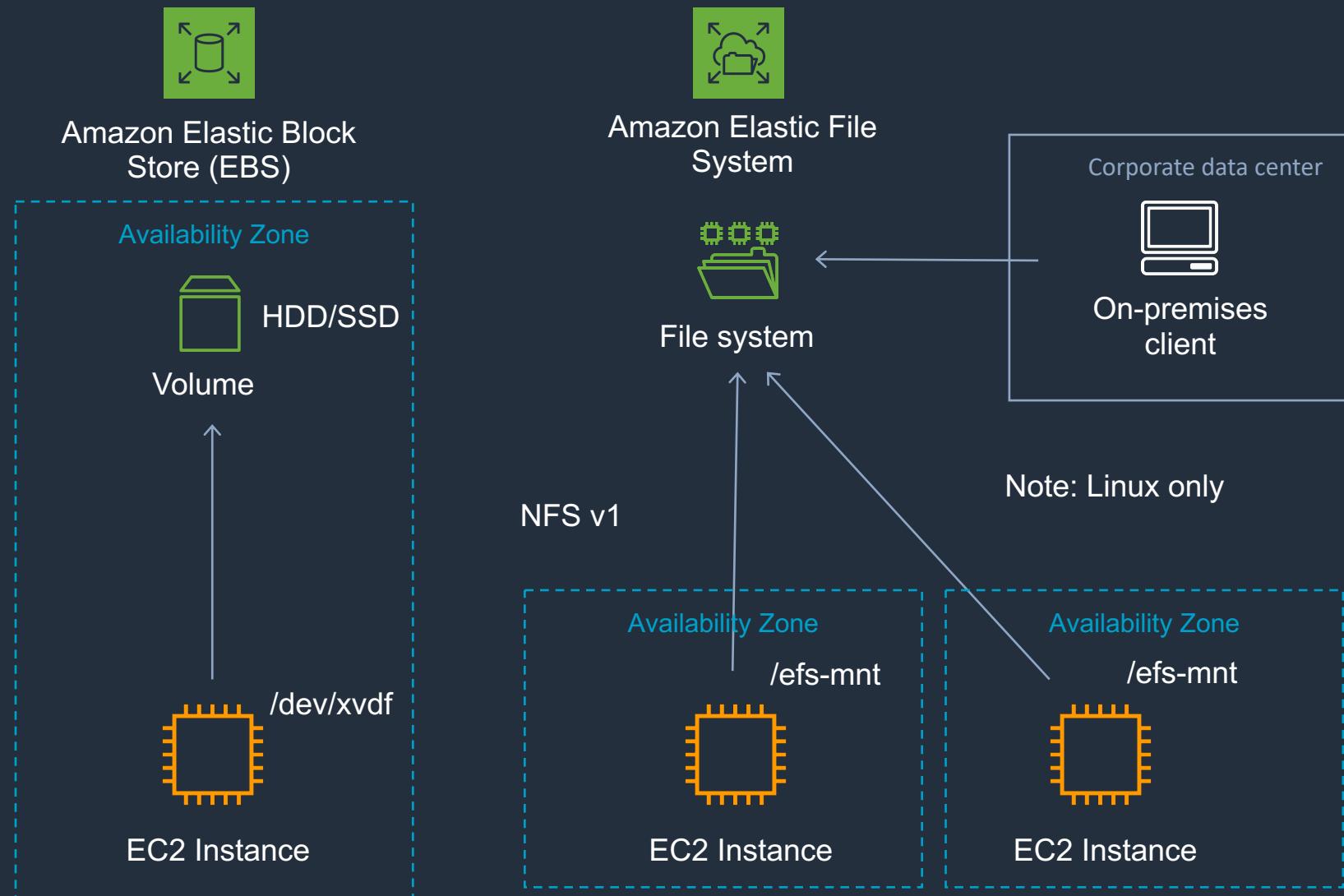
Section 7: Exam Cram

Amazon CloudFront Charges

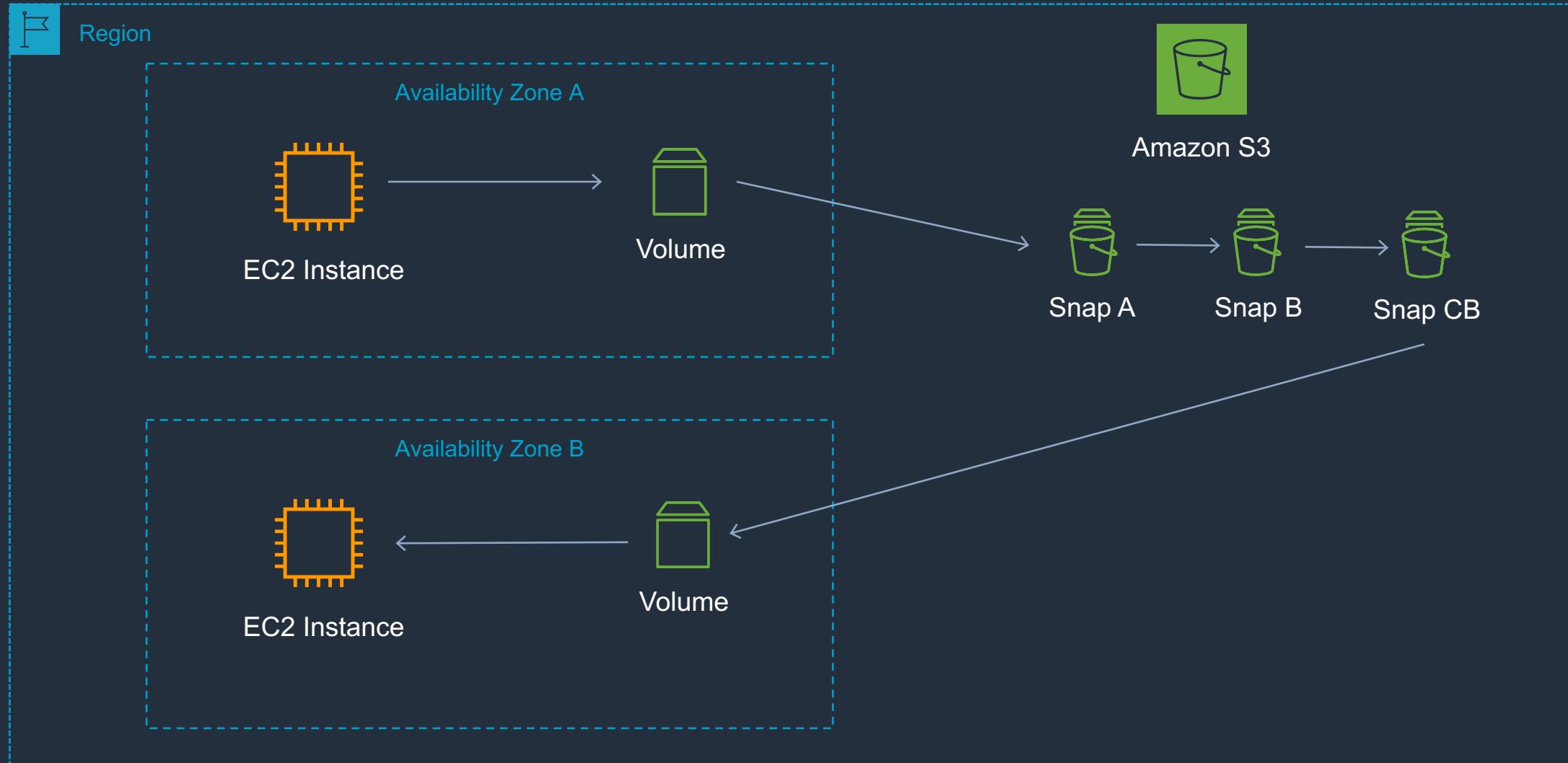
- You pay for:
 - Data Transfer Out to Internet.
 - Data Transfer Out to Origin.
 - Number of HTTP/HTTPS Requests.
 - Invalidation Requests.
 - Dedicated IP Custom SSL.
 - Field level encryption requests.
- You do not pay for:
 - Data transfer between AWS regions and CloudFront.
 - Regional edge cache.
 - AWS ACM SSL/TLS certificates.
 - Shared CloudFront certificates.



Section 8: EBS and EFS Overview



Section 8: EBS Snapshots



Section 8: EBS Copying, Sharing and Encryption



- Encryption state retained
- Same region



- Can be encrypted
- Can change regions



- Can be encrypted
- Can change AZ



- Cannot be encrypted
- Can be shared with other accounts
- Can be shared publicly



- Can change encryption key
- Can change regions



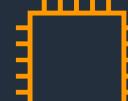
- Can change encryption key
- Can change AZ



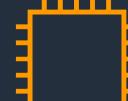
- Block devices remain encrypted
- Cannot be shared with other accounts if using AWS CMK
- Cannot be shared publicly



- Block devices remain encrypted
- Can change regions

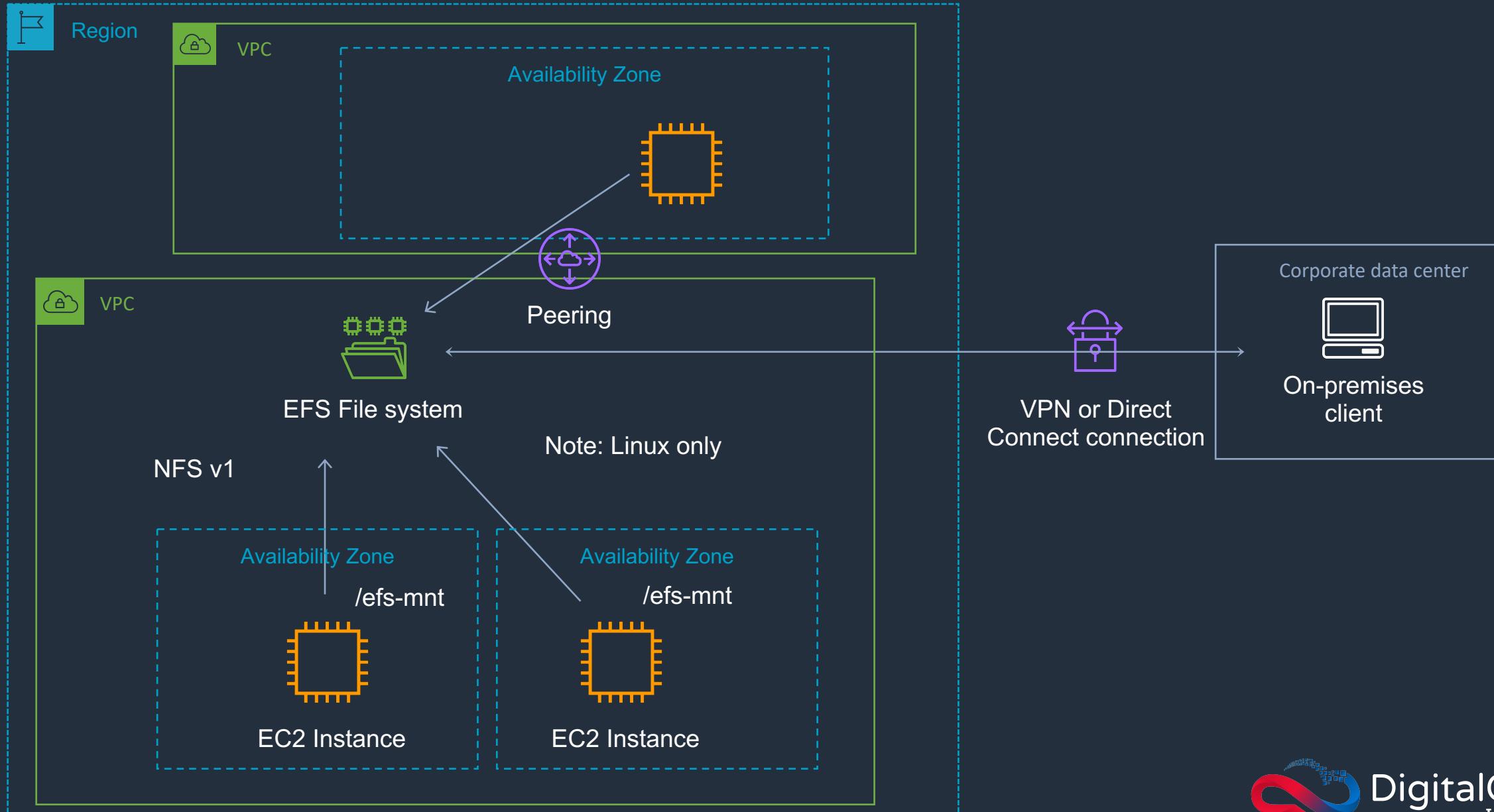


- Can change encryption key
- Can change AZ



- Can change encryption state
- Can change AZ

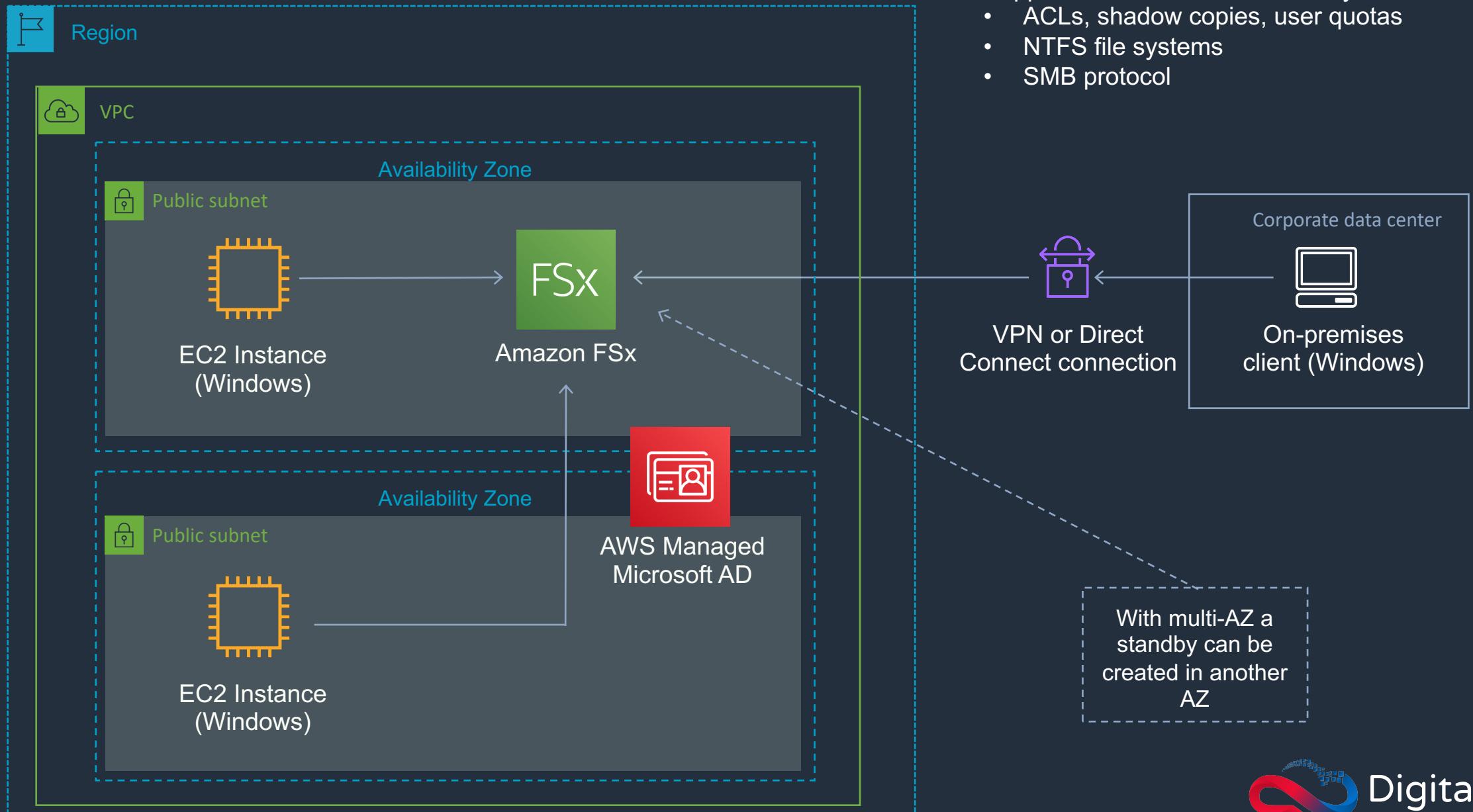
Section 8: EFS Overview



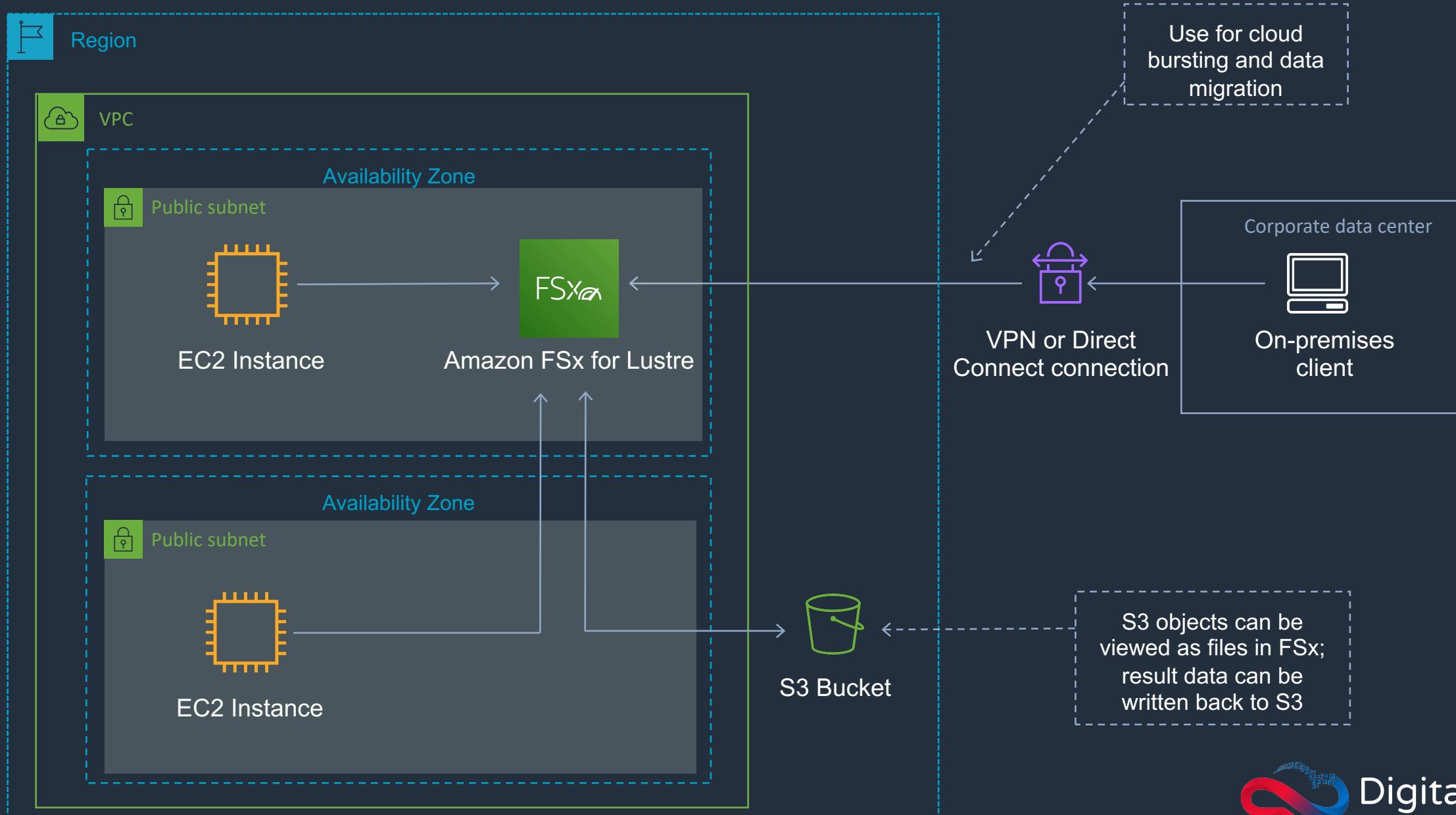
Section 8: Amazon FSx

- Amazon FSx provides fully managed third-party file systems.
- Amazon FSx provides you with the native compatibility of third-party file systems with feature sets for workloads such as Windows-based storage, high-performance computing (HPC), machine learning, and electronic design automation (EDA).
- Amazon FSx provides you with two file systems to choose from:
 - Amazon FSx for Windows File Server for Windows-based applications
 - Amazon FSx for Lustre for compute-intensive workloads.

Section 8: Amazon FSx for Windows File Server



Section 8: Amazon FSx for Lustre



Section 8: Exam Cram

Amazon Elastic Block Store (EBS)



- EBS volumes are network attached storage that can be attached to EC2 instances.
- EBS volume data persists independently of the life of the instance.
- EBS volumes do not need to be attached to an instance.
- You can attach multiple EBS volumes to an instance.
- You cannot attach an EBS volume to multiple instances (use Elastic File System instead).
- EBS volume data is replicated across multiple servers in an AZ.
- EBS volumes must be in the same AZ as the instances they are attached to.
- Root EBS volumes are deleted on termination by default.
- Extra non-boot volumes are not deleted on termination by default.
- The behavior can be changed by altering the “DeleteOnTermination” attribute.

Section 8: Exam Cram

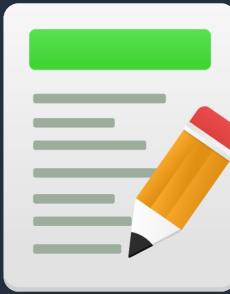
Amazon Instance Store

- An instance store provides temporary (non-persistent) block-level storage for your instance.
- This is different to EBS which provides persistent storage but is also a block storage service that can be a root or additional volume.
- Instance store storage is located on disks that are physically attached to the host computer.
- Instance store is ideal for temporary storage of information that changes frequently, such as buffers, caches, scratch data, and other temporary content, or for data that is replicated across a fleet of instances, such as a load-balanced pool of web servers.
- The instance type determines the size of the instance store available and the type of hardware used for the instance store volumes.
- Instance store volumes are included as part of the instance's usage cost.



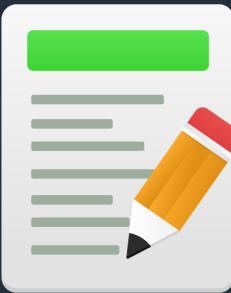
Section 8: Exam Cram

Amazon EBS vs Instance Store



- EBS-backed means the root volume is an EBS volume and storage is persistent.
- Instance store-backed means the root volume is an instance store volume and storage is not persistent.
- Instance store backed instances cannot be stopped. If the underlying host fails the data will be lost.
- Instance store volume root devices are created from AMI templates stored on S3.
- EBS backed instances can be stopped. You will not lose the data on this instance if it is stopped (persistent).
- EBS volumes can be detached and reattached to other EC2 instances.
- EBS volume root devices are launched from AMI's that are backed by EBS snapshots.
- Instance store volumes cannot be detached/reattached.
- When rebooting the instances for both types data will not be lost.

Section 8: Exam Cram

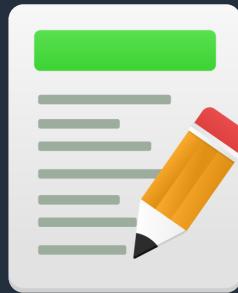


	Solid State Drives (SSD)		Hard Disk Drives (HDD)	
Volume Type	EBS Provisioned IOPS SSD (io1)	EBS General Purpose SSD (gp2)	Throughput Optimized HDD (st1)	Cold HDD (sc1)
Short Description	Highest performance SSD volume designed for latency-sensitive transactional workloads	General Purpose SSD volume that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

Section 8: Exam Cram

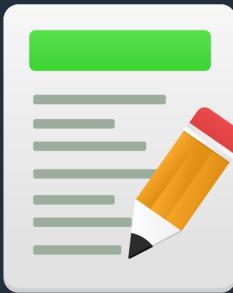
Amazon EBS Snapshots

- Snapshots capture a point-in-time state of an instance.
- Cost-effective and easy backup strategy.
- Can be used to migrate a system to a new AZ or region.
- Can be used to convert an unencrypted volume to an encrypted volume.
- Snapshots are stored on Amazon S3.
- Does not provide granular backup (not a replacement for backup software).
- Even though snapshots are saved incrementally, the snapshot deletion process is designed so that you need to retain only the most recent snapshot in order to restore the volume.
- Snapshots can only be accessed through the EC2 APIs.
- EBS volumes are AZ specific but snapshots are region specific.



Section 8: Exam Cram

Amazon EBS Encryption

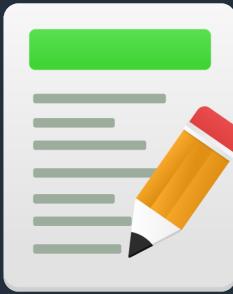


- You can encrypt both the boot and data volumes of an EC2 instance. When you create an encrypted EBS volume and attach it to a supported instance type, the following types of data are encrypted:
 - Data at rest inside the volume.
 - All data moving between the volume and the instance.
 - All snapshots created from the volume.
 - All volumes created from those snapshots.
- Encryption is supported by all EBS volume types.
- Expect the same IOPS performance on encrypted volumes as on unencrypted volumes.
- All instance families support encryption.

Section 8: Exam Cram

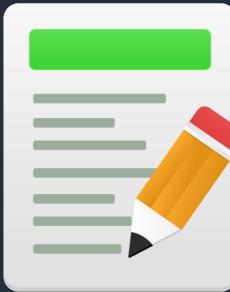
Amazon EBS AMIs

- An Amazon Machine Image (AMI) is a special type of virtual appliance that is used to create a virtual machine within the Amazon Elastic Compute Cloud ("EC2").
- An AMI includes the following:
 - A template for the root volume for the instance (for example, an operating system, an application server, and applications).
 - Launch permissions that control which AWS accounts can use the AMI to launch instances.
 - A block device mapping that specifies the volumes to attach to the instance when it's launched.
- AMIs are either instance store-backed or EBS-backed.
- You can copy an AMI within or across an AWS region using the AWS Management Console, the AWS AWS CLI or SDKs, or the Amazon EC2 API.



Section 8: Exam Cram

Amazon Elastic File System (EFS)

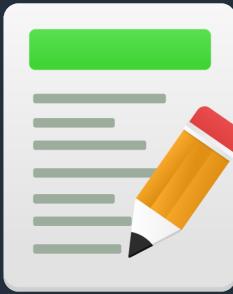


- EFS is a fully-managed service that makes it easy to set up and scale file storage in the Amazon Cloud.
- Implementation of an NFS file share and is accessed using the NFSv4.1 protocol.
- Elastic storage capacity and pay for what you use (in contrast to EBS with which you pay for what you provision).
- Multi-AZ metadata and data storage.
- Can configure mount-points in one, or many, AZs.
- Can be mounted from on-premises systems ONLY if using Direct Connect or a VPN connection.
- Alternatively, use the EFS File Sync agent.
- EFS is elastic and grows and shrinks as you add and remove data.
- Can scale up to petabytes.

Section 8: Exam Cram

Amazon Elastic File System (EFS)

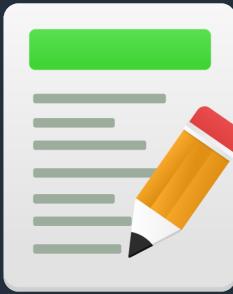
- Can concurrently connect 1 to 1000s of EC2 instances, from multiple AZs.
- Can choose General Purpose or Max I/O (both SSD).
- EFS provides a file system interface, file system access semantics (such as strong consistency and file locking).
- Data is stored across multiple AZ's within a region.
- Read after write consistency.
- Need to create mount targets and choose AZ's to include (recommended to include all AZ's).



Section 8: Exam Cram

Amazon EFS Access Control

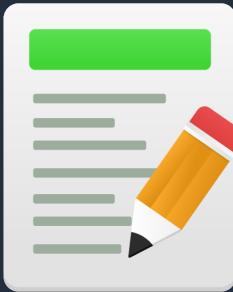
- You can control who can administer your file system using IAM.
- You can control access to files and directories with POSIX-compliant user and group-level permissions.
- POSIX permissions allow you to restrict access from hosts by user and group.
- EFS Security Groups act as a firewall, and the rules you add define the traffic flow.



Section 8: Exam Cram

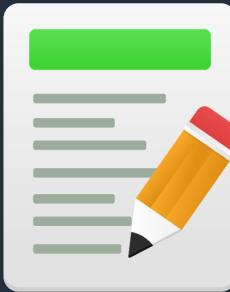
Amazon EFS Encryption

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



Section 8: Exam Cram

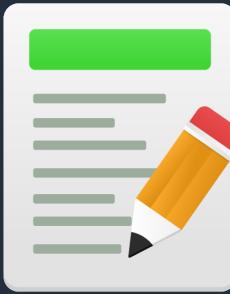
Amazon EFS File Sync



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

Section 8: Exam Cram

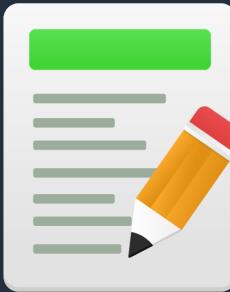
Amazon FSx (SAA-C02 exam only)



- Amazon FSx provides fully managed third-party file systems.
- Amazon FSx provides you with the native compatibility of third-party file systems with feature sets for workloads such as Windows-based storage, high-performance computing (HPC), machine learning, and electronic design automation (EDA).
- You don't have to worry about managing file servers and storage, as Amazon FSx automates the time-consuming administration tasks such as hardware provisioning, software configuration, patching, and backups.
- Amazon FSx integrates the file systems with cloud-native AWS services, making them even more useful for a broader set of workloads.
- Amazon FSx provides you with two file systems to choose from:
 - Amazon FSx for Windows File Server for Windows-based applications
 - Amazon FSx for Lustre for compute-intensive workloads.

Section 8: Exam Cram

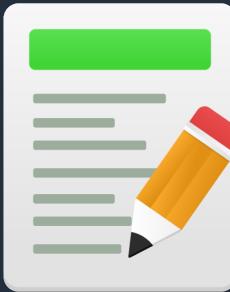
Amazon FSx for Windows File Server (SAA-C02 exam only)



- Amazon FSx for Windows File Server provides a fully managed native Microsoft Windows file system
- Built on Windows Server, Amazon FSx provides the compatibility and features that Microsoft applications rely on, including full support for the SMB protocol, Windows NTFS, and Microsoft Active Directory (AD) integration.
- Amazon FSx uses SSD storage to provide fast performance with low latency.
- This compatibility, performance, and scalability enables business-critical workloads such as home directories, media workflows, and business applications.
- Supports Windows-native file system features:
 - Access Control Lists (ACLs), shadow copies, and user quotas.
 - NTFS file systems that can be accessed from up to thousands of compute instances using the SMB protocol.

Section 8: Exam Cram

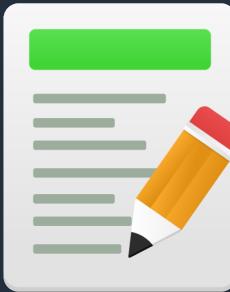
Amazon FSx for Windows File Server (SAA-C02 exam only)



- Amazon FSx can connect file systems to Amazon EC2, VMware Cloud on AWS, Amazon WorkSpaces, and Amazon AppStream 2.0 instances.
- Amazon FSx also supports on-premises access via AWS Direct Connect or AWS VPN, and access from multiple VPCs, accounts, and regions using VPC Peering or AWS Transit Gateway.
- Amazon FSx automatically encrypts your data at-rest and in-transit.
- Assessed to comply with ISO, PCI-DSS, and SOC certifications, and is HIPAA eligible.
- Amazon FSx automatically replicates your data within an Availability Zone (AZ) it resides in (which you specify during creation) to protect it from component failure.
- Amazon FSx offers a multiple availability (AZ) deployment option, designed to provide continuous availability to data, even in the event that an AZ is unavailable. Multi-AZ file systems include an active and standby file server in separate AZs, and any changes written to disk in your file system are synchronously replicated across AZs to the standby.

Section 8: Exam Cram

Amazon FSx for Lustre (SAA-C02 exam only)

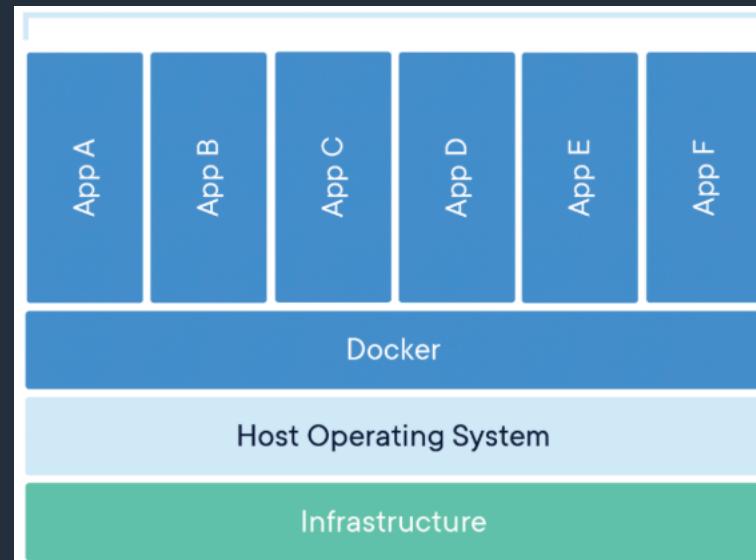


- Amazon FSx for Lustre provides a high-performance file system optimized for fast processing of workloads such as machine learning, high performance computing (HPC), video processing, financial modeling, and electronic design automation (EDA).
- These workloads commonly require data to be presented via a fast and scalable file system interface, and typically have data sets stored on long-term data stores like Amazon S3.
- Amazon FSx works natively with Amazon S3, letting you transparently access your S3 objects as files on Amazon FSx to run analyses for hours to months.
- You can then write results back to S3, and simply delete your file system. FSx for Lustre also enables you to burst your data processing workloads from on-premises to AWS, by allowing you to access your FSx file system over Amazon Direct Connect or VPN.
- You can also use FSx for Lustre as a standalone high-performance file system to burst your workloads from on-premises to the cloud.

Section 9: Elastic Container Services Overview

Definition of containers (from Docker):

A container is a standard unit of software that packages up code and all its dependencies so the application runs quickly and reliably from one computing environment to another. A Docker container image is a lightweight, standalone, executable package of software that includes everything needed to run an application: code, runtime, system tools, system libraries and settings.



Section 9: Elastic Container Services Overview



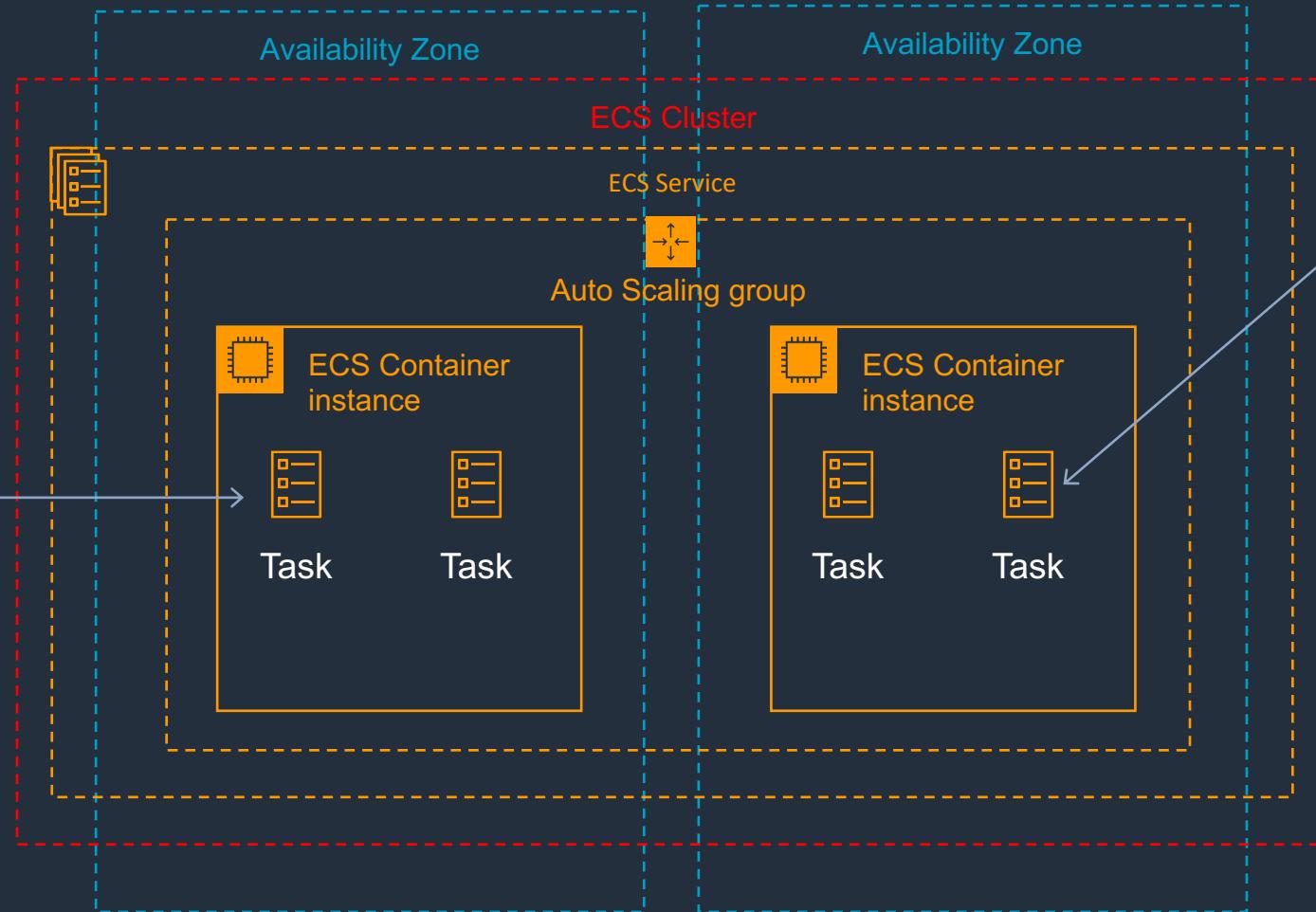
Amazon Elastic Container Registry



Amazon Elastic Container Service

Task Definition

```
{
  "containerDefinitions": [
    {
      "name": "wordpress",
      "links": [
        "mysql"
      ],
      "image": "wordpress",
      "essential": true,
      "portMappings": [
        {
          "containerPort": 80,
          "hostPort": 80
        }
      ],
      "memory": 500,
      "cpu": 10
    }
  ]
}
```



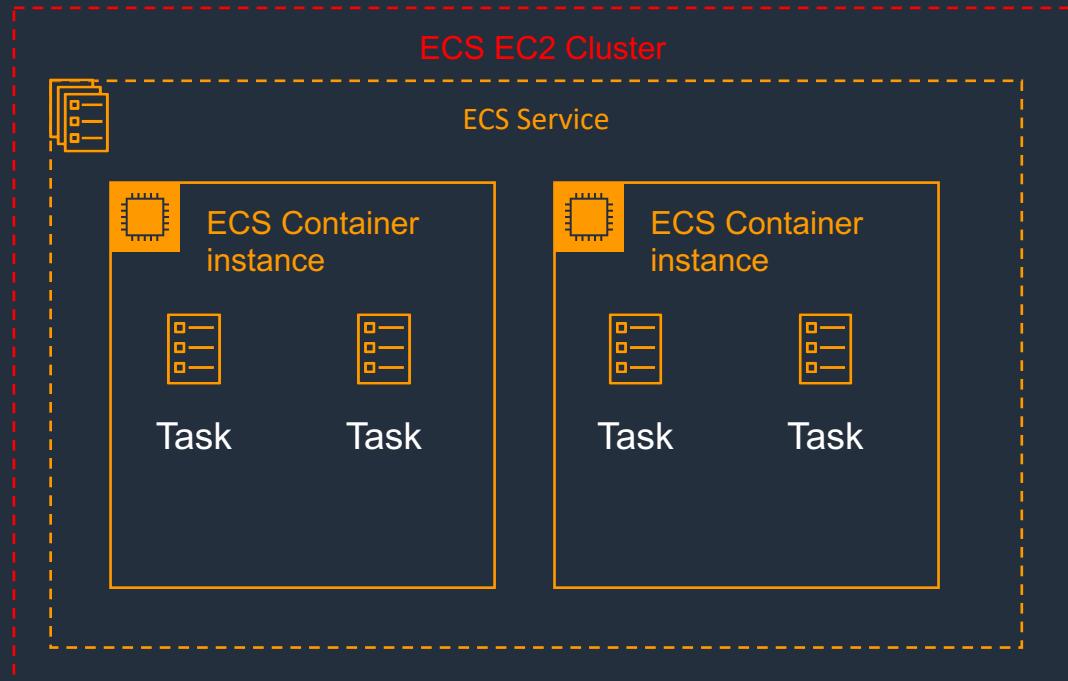
Section 9: ECS Terminology

Elastic Container Service (ECS)	Description
Cluster	Logical grouping of EC2 instances
Container instance	EC2 instance running the the ECS agent
Task Definition	Blueprint that describes how a docker container should launch
Task	A running container using settings in a Task Definition
Service	Defines long running tasks – can control task count with Auto Scaling and attach an ELB

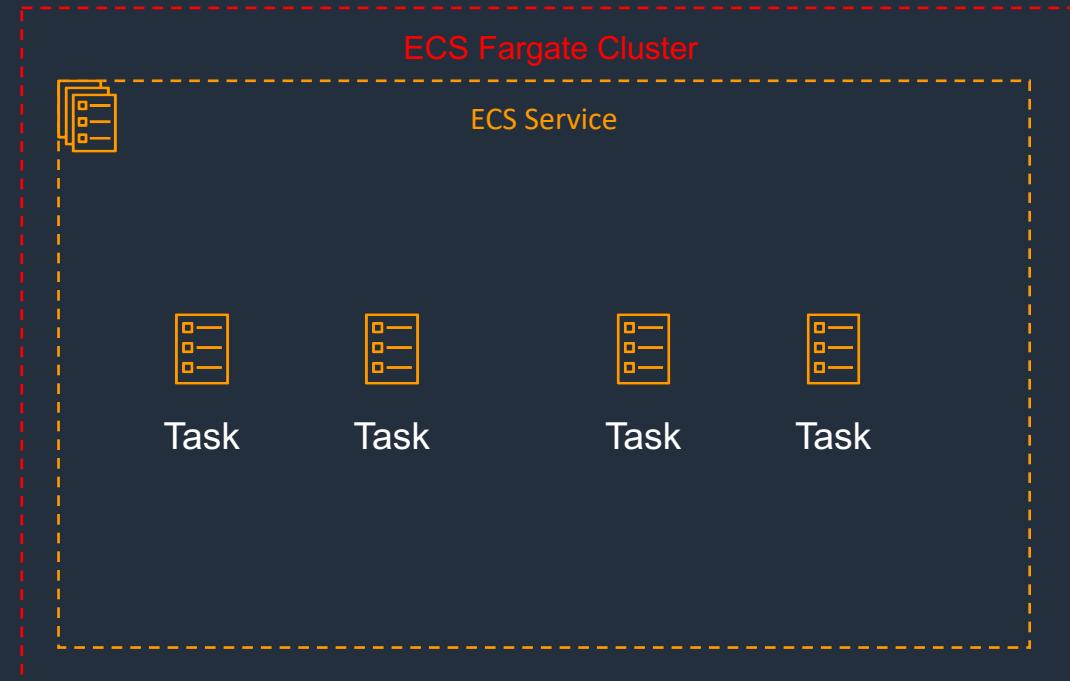
Section 9: Launch Types – EC2 and Fargate



Registry:
ECR, Docker Hub, Self-hosted



Registry:
ECR, Docker Hub



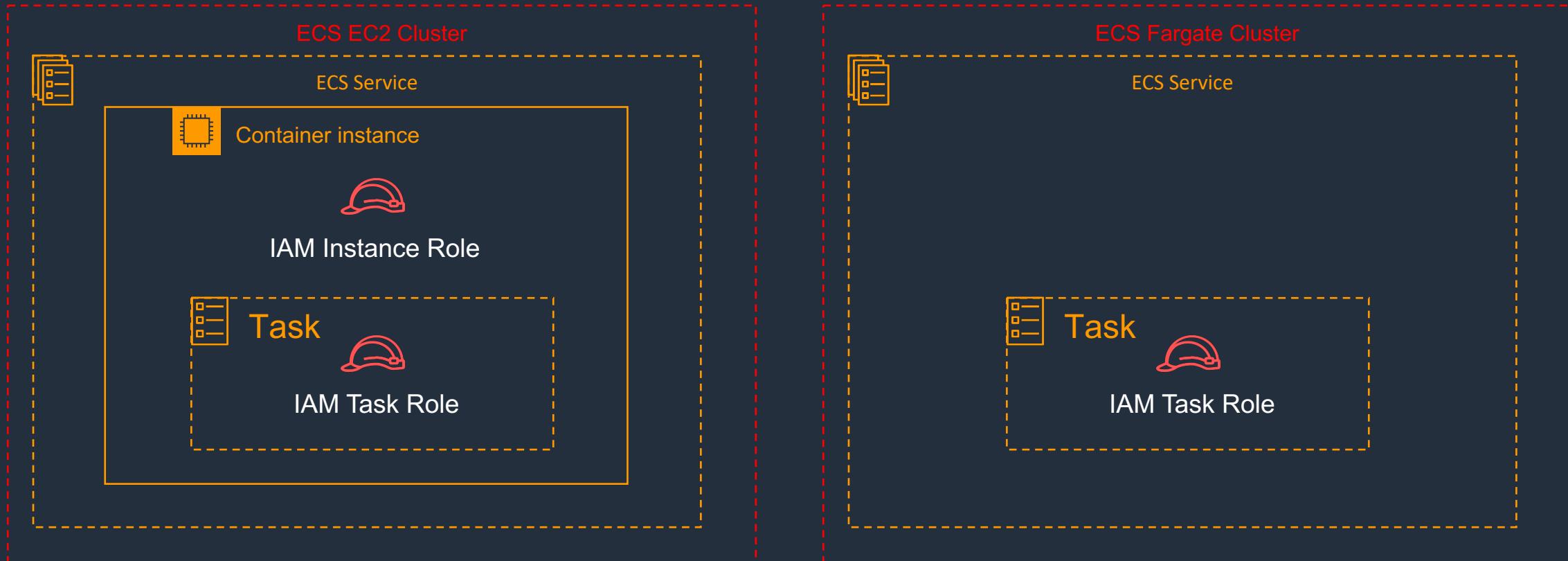
EC2 Launch Type

- You explicitly provision EC2 instances
- You're responsible for managing EC2 instances
- Charged per running EC2 instance
- EFS and EBS integration
- You handle cluster optimization
- More granular control over infrastructure

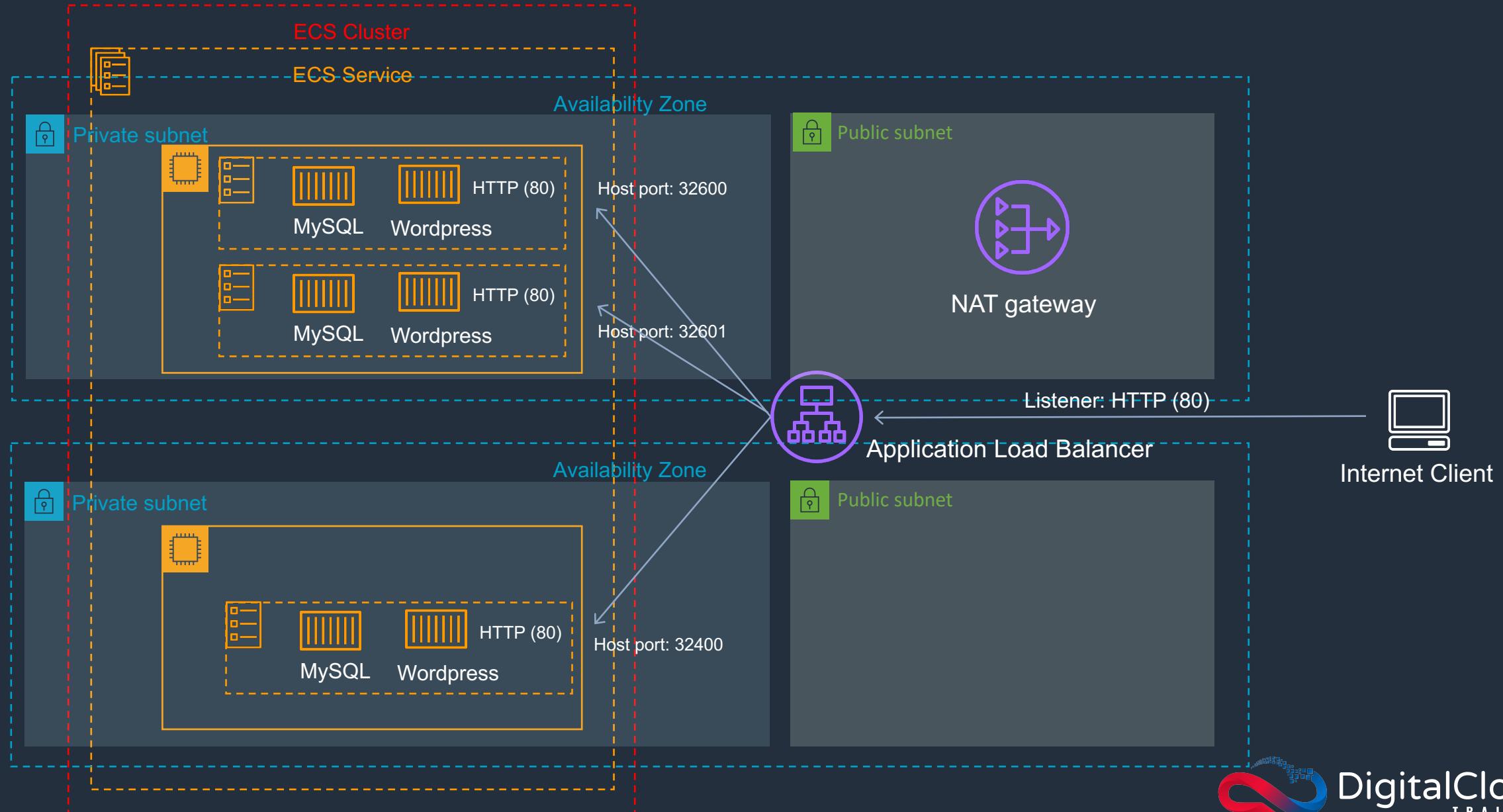
Fargate Launch Type

- Fargate automatically provisions resources
- Fargate provisions and manages compute
- Charged for running tasks
- No EFS and EBS integration
- Fargate handles cluster optimization
- Limited control, infrastructure is automated

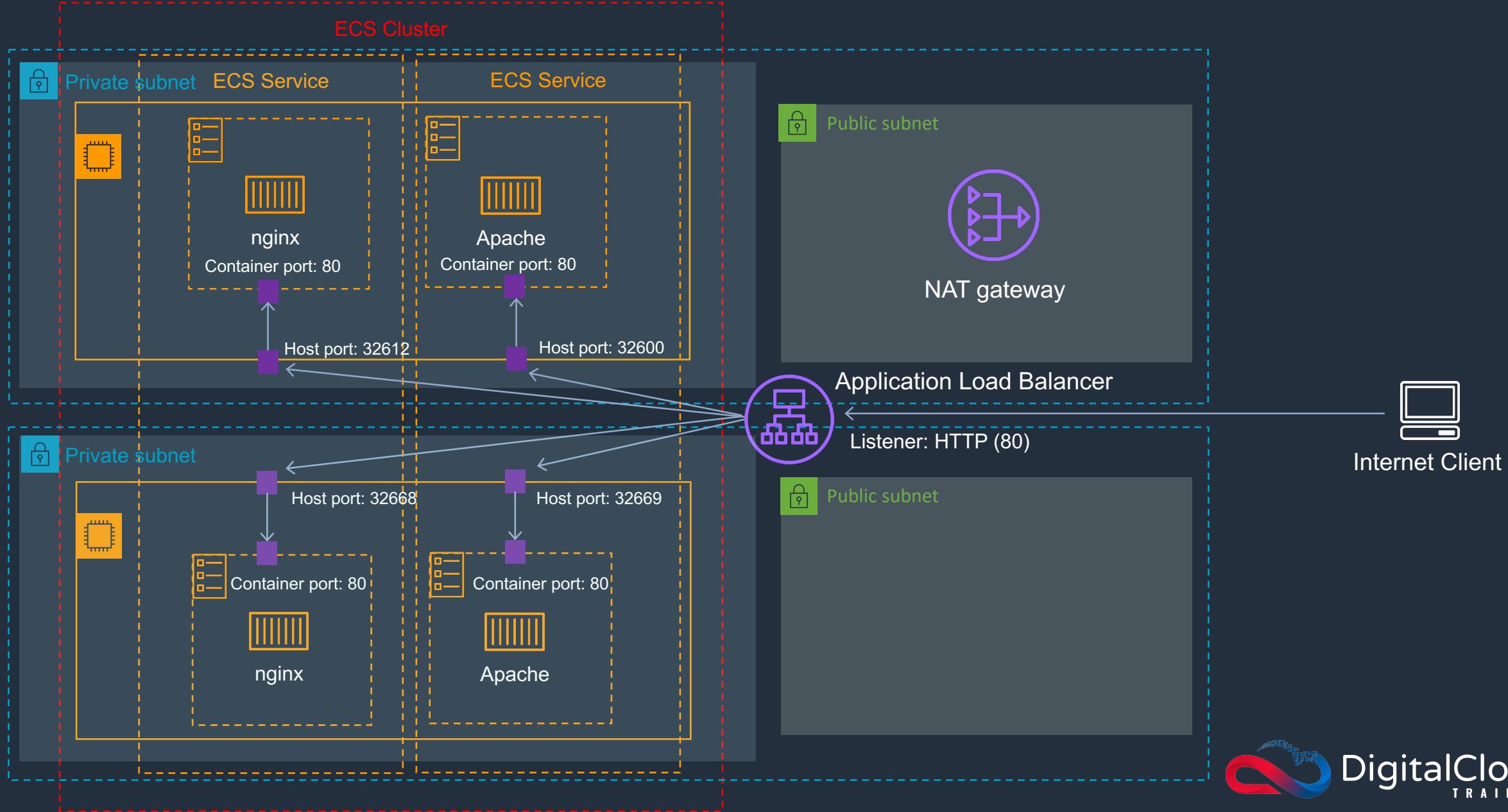
Section 9: IAM Roles



Section 9: ECS with Application Load Balancer (ALB)



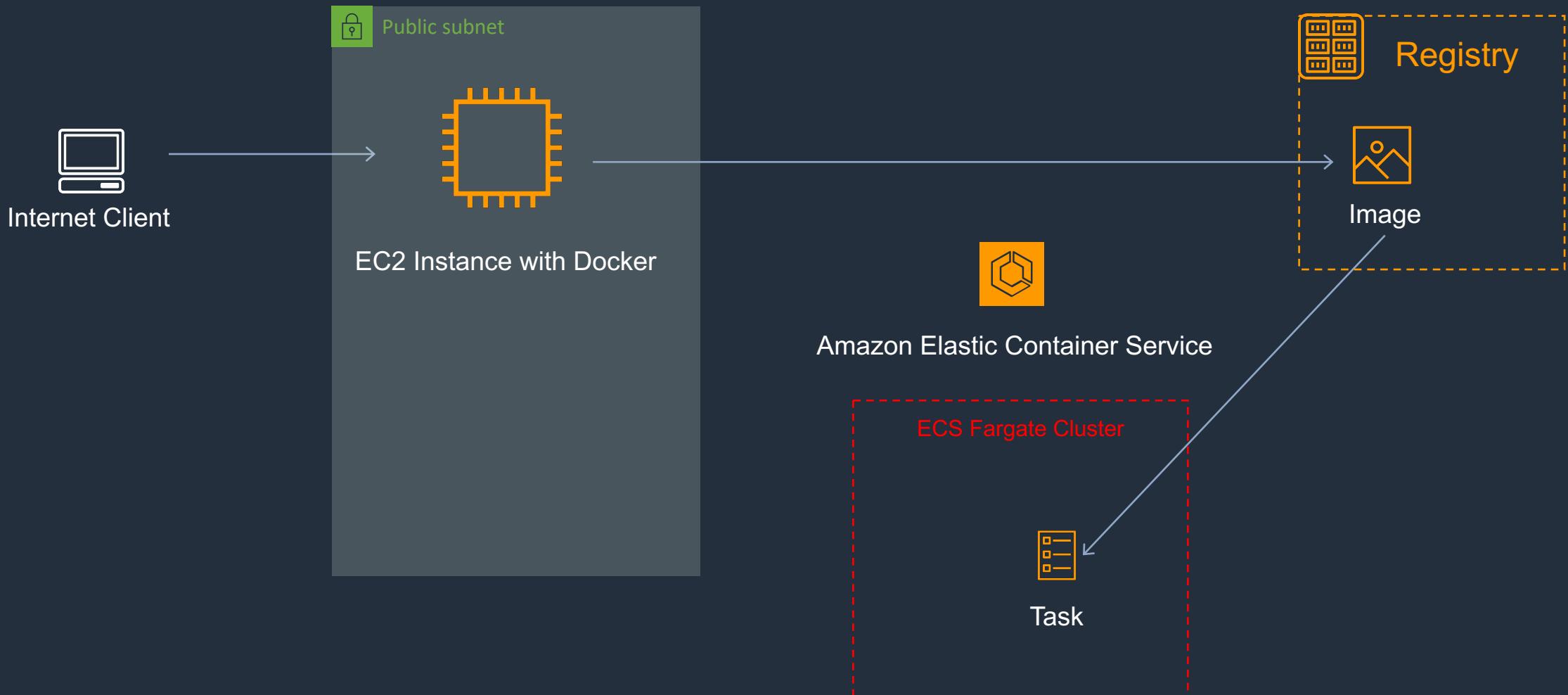
Section 9: ECS with Application Load Balancer (ALB)



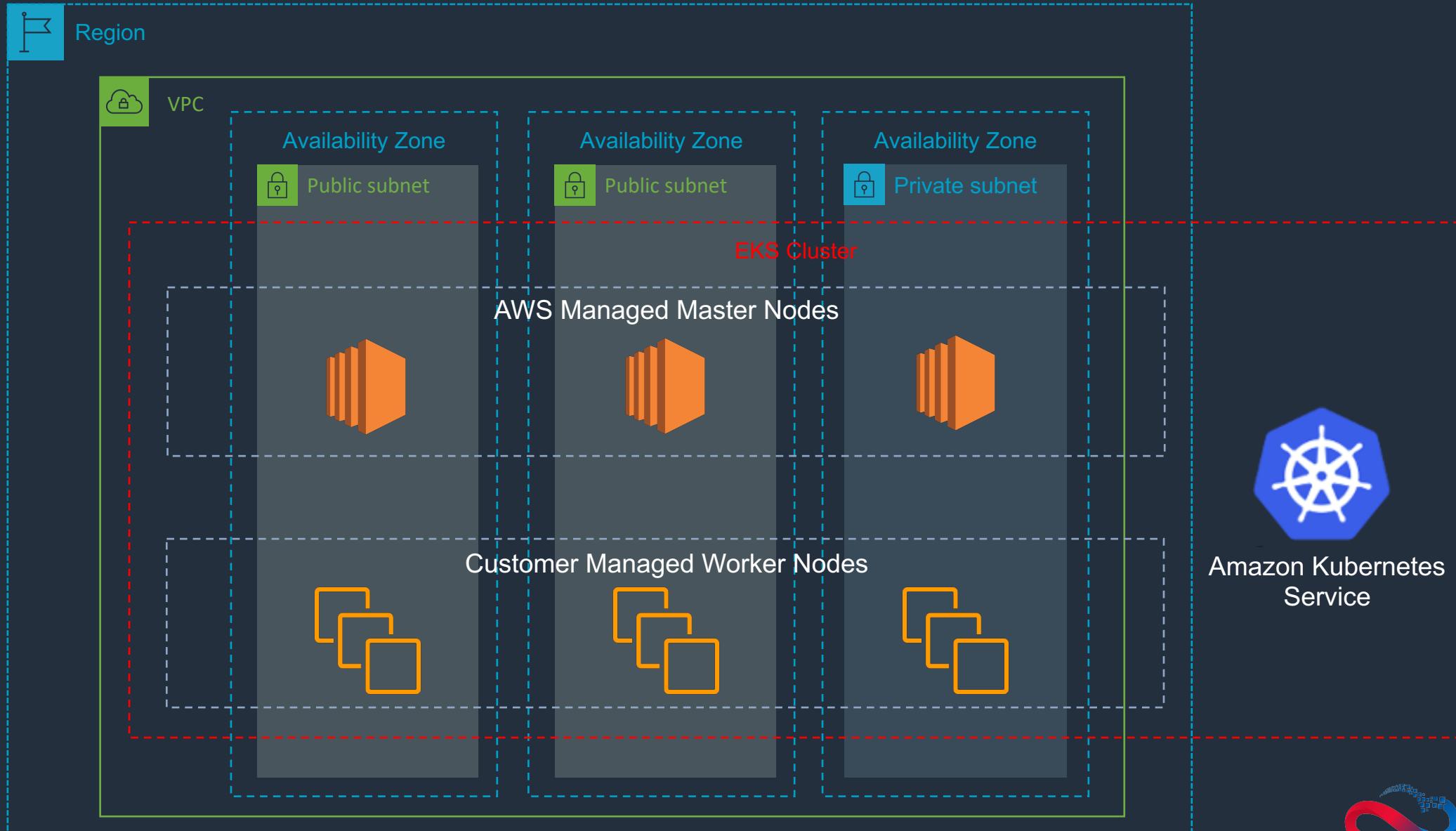
Section 9: Elastic Container Registry



Amazon Elastic Container Registry

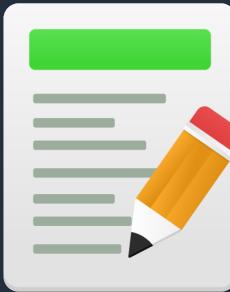


Section 9: Elastic Kubernetes Service



Section 9: Exam Cram

Amazon Elastic Container Service (ECS)



- Amazon Elastic Container Service (ECS) is a highly scalable, high performance container management service that supports Docker containers.
- Amazon ECS eliminates the need for you to install, operate, and scale your own cluster management infrastructure.
- Amazon ECS can be used to schedule the placement of containers across clusters based on resource needs and availability requirements.
- There is no additional charge for Amazon ECS. You pay for AWS resources (e.g. EC2 instances or EBS volumes) you create to store and run your application.
- It is possible to associate a service on Amazon ECS to an Application Load Balancer (ALB) for the Elastic Load Balancing (ELB) service.

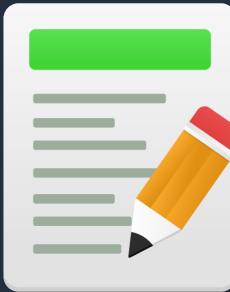
Section 9: Exam Cram

Amazon ECS vs EKS



Amazon ECS	Amazon EKS
Managed, highly available, highly scalable container platform	
AWS-specific platform that supports Docker containers	Compatible with upstream Kubernetes so it's easy to lift and shift from other Kubernetes deployments
Considered simpler to learn and use	Considered more feature-rich and complex with a steep learning curve
Leverages AWS services like Route 53, ALB, and CloudWatch	A hosted Kubernetes platform that handles many things internally
“Tasks” are instances of containers that are run on underlying compute but more or less isolated	“Pods” are containers collocated with one another and can have shared access to each other
Limited extensibility	Extensible via a wide variety of third-party and community add-ons

Section 9: Exam Cram



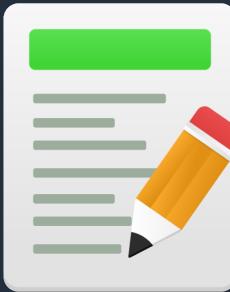
Amazon ECS Launch Types

- An Amazon ECS launch type determines the type of infrastructure on which your tasks and services are hosted.
- There are two launch types and the table below describes some of the differences between the two launch types:

Amazon EC2	Amazon Fargate
You explicitly provision EC2 instances	The control plane asks for resources and Fargate automatically provisions
You're responsible for upgrading, patching, care of EC2 pool	Fargate provisions compute as needed
You must handle cluster optimization	Fargate handles cluster optimization
More granular control over infrastructure	Limited control, as infrastructure is automated

Section 9: Exam Cram

Amazon ECS Images

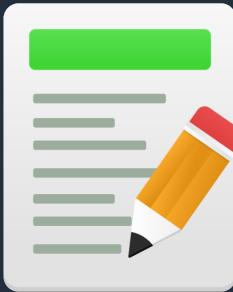


- Containers are created from a read-only template called an image which has the instructions for creating a Docker container.
- Images are built from a Dockerfile.
- Only Docker containers are currently supported.
- An image contains the instructions for creating a Docker container.
- Images are stored in a registry such as DockerHub or AWS Elastic Container Registry (ECR).
- ECR is a managed AWS Docker registry service that is secure, scalable and reliable.

Section 9: Exam Cram

Amazon ECS Tasks

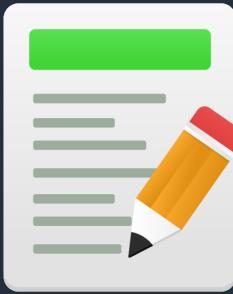
- A task definition is required to run Docker containers in Amazon ECS.
- A task definition is a text file in JSON format that describes one or more containers, up to a maximum of 10.
- Task definitions use Docker images to launch containers.
- You specify the number of tasks to run (i.e. the number of containers).



Section 9: Exam Cram

Amazon ECS Clusters

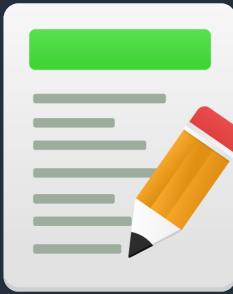
- ECS Clusters are a logical grouping of container instances the you can place tasks on.
- A default cluster is created but you can then create multiple clusters to separate resources.
- ECS allows the definition of a specified number (desired count) of tasks to run in the cluster.
- Clusters can contain tasks using the Fargate and EC2 launch type.
- For clusters with the EC2 launch type clusters can contain different container instance types.
- Each container instance may only be part of one cluster at a time.
- "Services" provide auto-scaling functions for ECS.
- Clusters are region specific.
- You can create IAM policies for your clusters to allow or restrict users' access to specific clusters.



Section 9: Exam Cram

Amazon ECS Container Agent

- The ECS container agent allows container instances to connect to the cluster.
- The container agent runs on each infrastructure resource on an ECS cluster.
- The ECS container agent is included in the Amazon ECS optimized AMI and can also be installed on any EC2 instance that supports the ECS specification (only supported on EC2 instances).
- Linux and Windows based.
- For non-AWS Linux instances to be used on AWS you must manually install the ECS container agent.



Section 9: Exam Cram

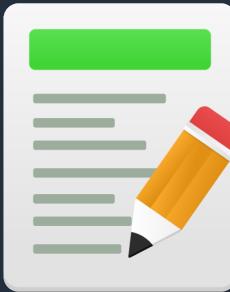
Amazon ECS Auto Scaling



- Service Auto Scaling
 - Amazon ECS service can optionally be configured to use Service Auto Scaling to adjust the desired task count up or down automatically.
 - Service Auto Scaling leverages the Application Auto Scaling service to provide this functionality.
 - Supports the following types of scaling policies:
 - Target Tracking Scaling Policies.
 - Step Scaling Policies.

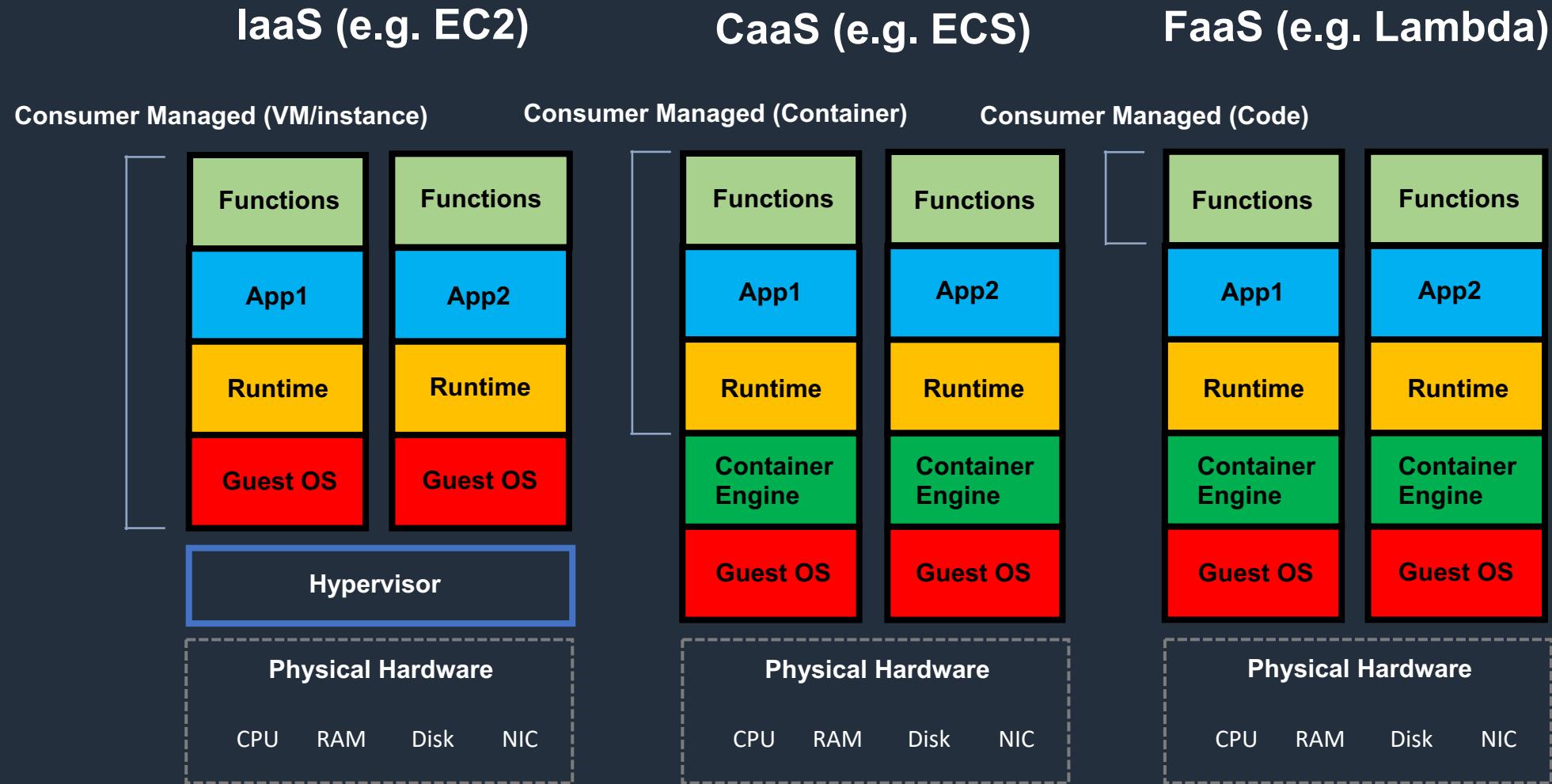
Section 9: Exam Cram

Amazon ECS Auto Scaling



- Cluster Auto Scaling
 - This is a new feature released in December 2019. It is unlikely that this will appear on the SAA-C01 exam but could appear on the SAA-C02 exam.
 - Uses a new ECS resource type called a Capacity Provider.
 - A Capacity Provider can be associated with an EC2 Auto Scaling Group (ASG).
 - When you associate an ECS Capacity Provider with an ASG and add the Capacity Provider to an ECS cluster, the cluster can now scale your ASG automatically by using two new features of ECS:
 - Managed scaling.
 - Managed instance termination protection.

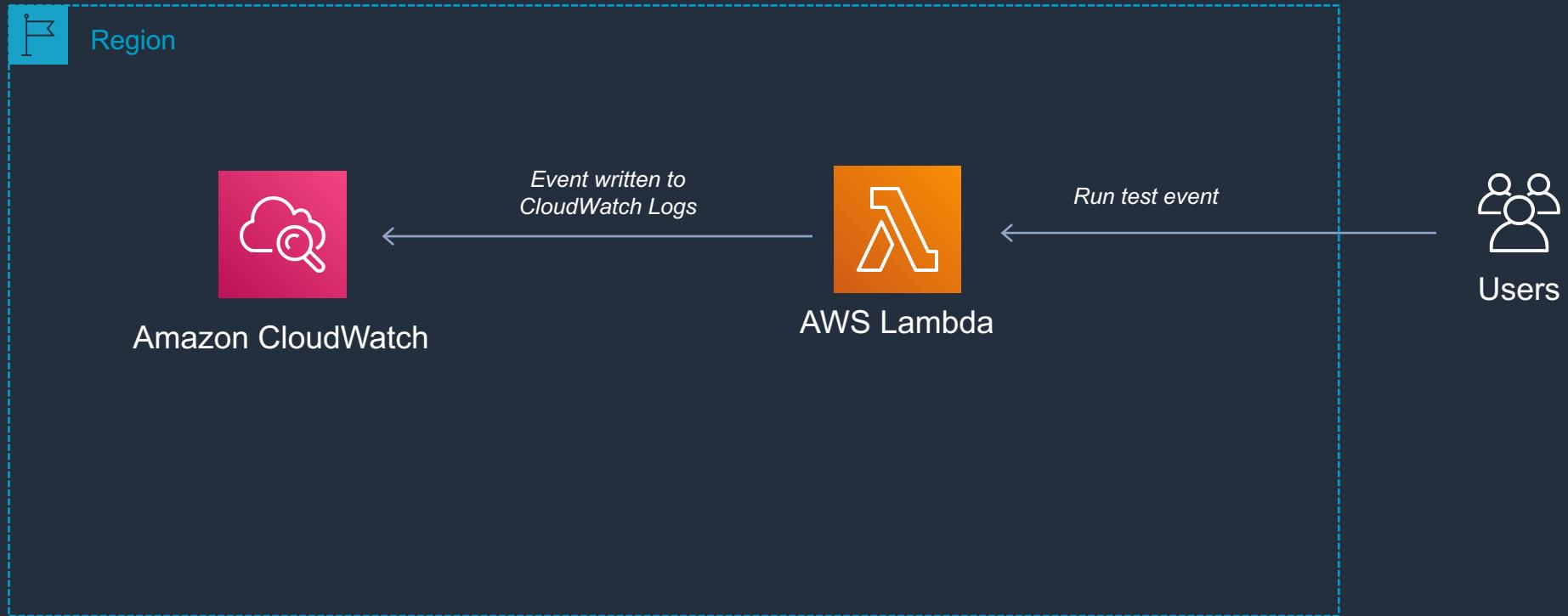
Section 10: Comparing IaaS, CaaS, and FaaS



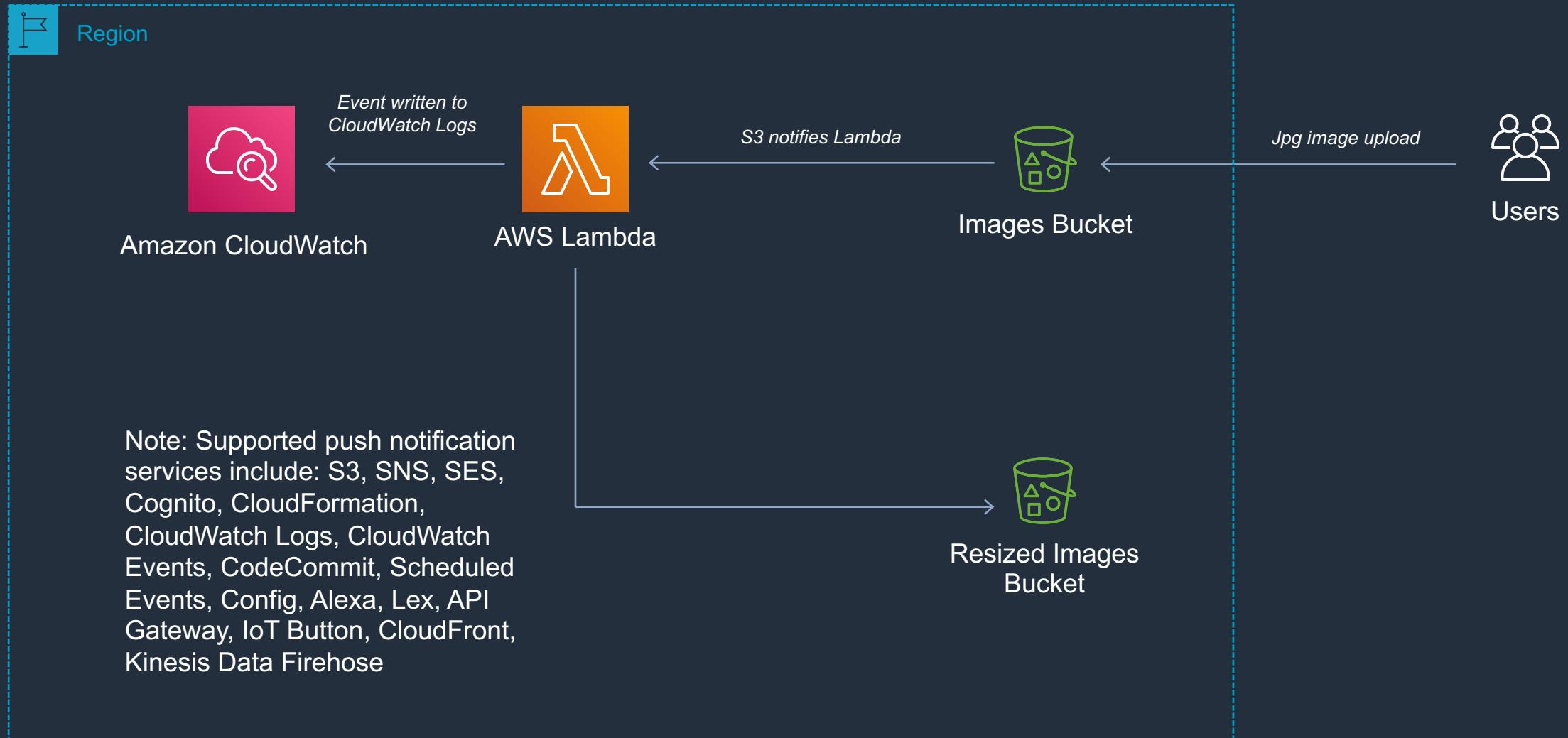
Section 10: Comparing Compute Options

EC2	ECS (EC2 Launch Type)	ECS (Fargate Launch Type)	Lambda
You manage the operating system	You manage container instance (EC2) and the containers (tasks)	You manage the containers (tasks)	You manage the code
Scale vertically – more CPU/Mem/HDD or scale horizontally (automatic) with Auto Scaling	Manually add container instances or use ECS Services and EC2 Auto Scaling	AWS scales the cluster automatically	Lambda automatically scales concurrent executions up to default limit (1000)
Use for traditional applications and long running tasks	Use for microservices and batch use cases where you need containers and need to retain management of underlying platform	Use for microservices and batch use cases	Use for ETL, infrastructure automation, data validation, mobile backends
No timeout issues	No timeout issues	No timeout issues	Limited to 900 seconds execution time for single execution (3 second default)
Pay for instance run time based on family/type	Pay for instance run time based on family/type	Pay for container run time based on allocated resources	Pay only for execution time based on memory allocation

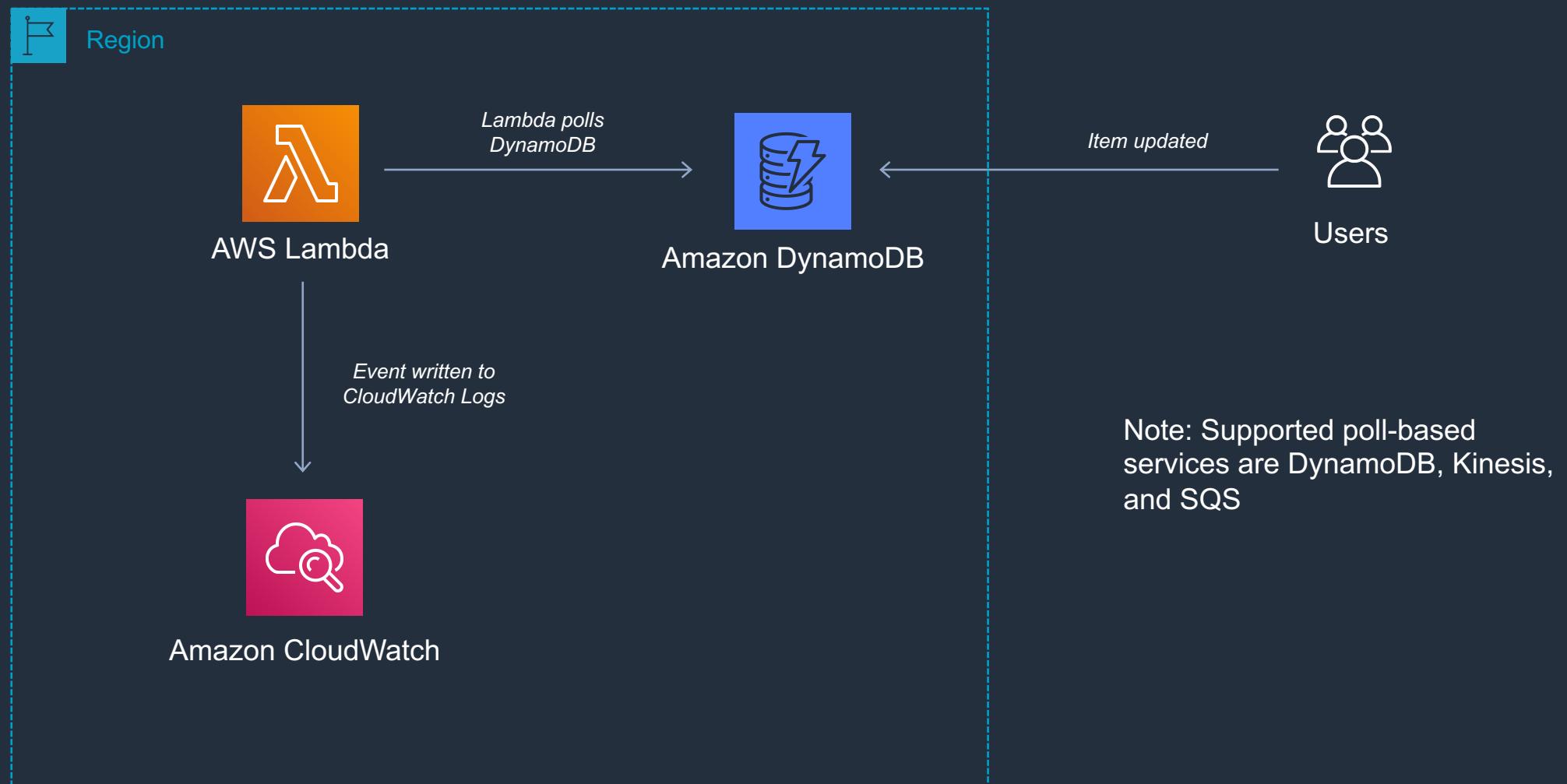
Section 10: AWS Lambda – Hello World



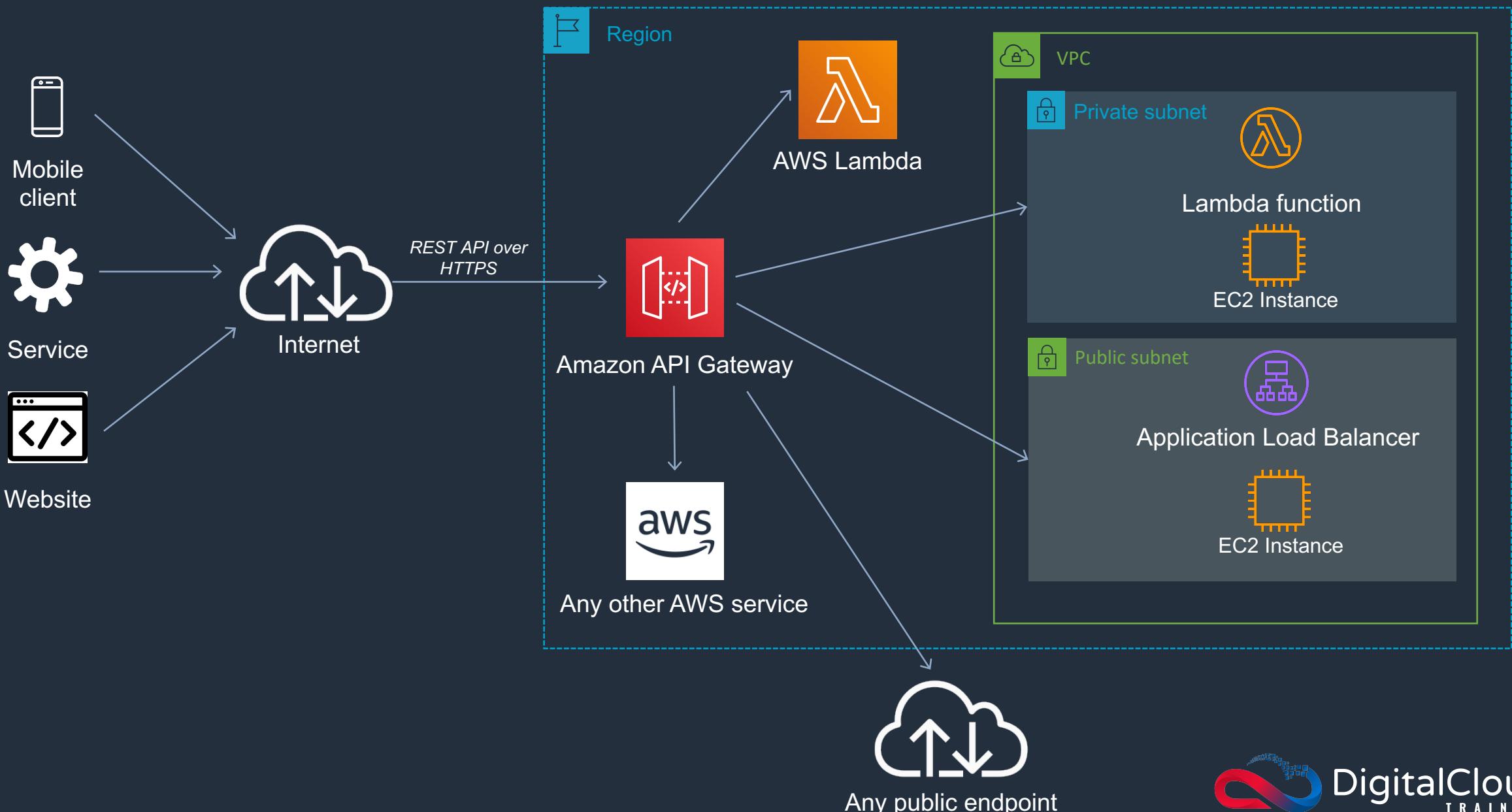
Section 10: AWS Lambda – S3 Event Source Mapping



Section 10: AWS Lambda – DynamoDB Event Source Mapping

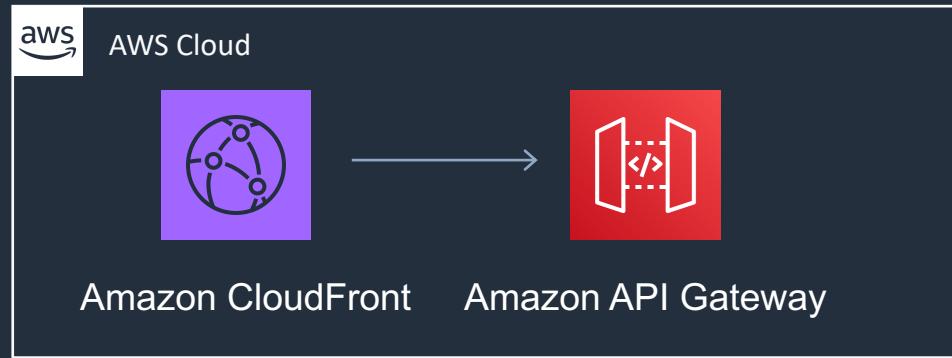


Section 10: API Gateway Overview



Section 10: API Gateway Endpoint Types

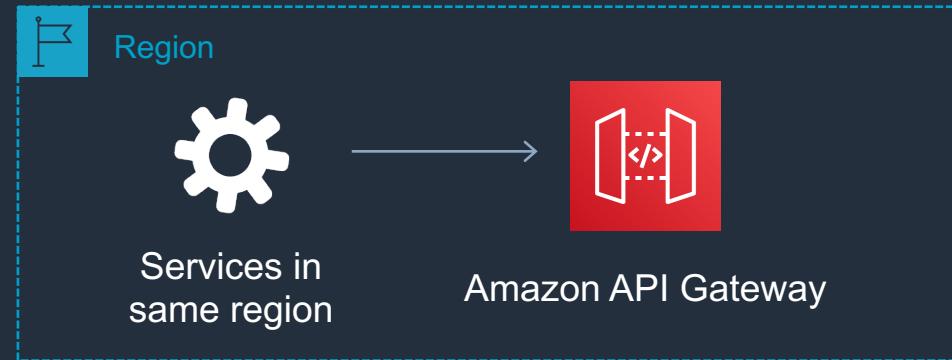
Edge-optimized endpoint:



Key benefits:

- Reduced latency for requests from around the world

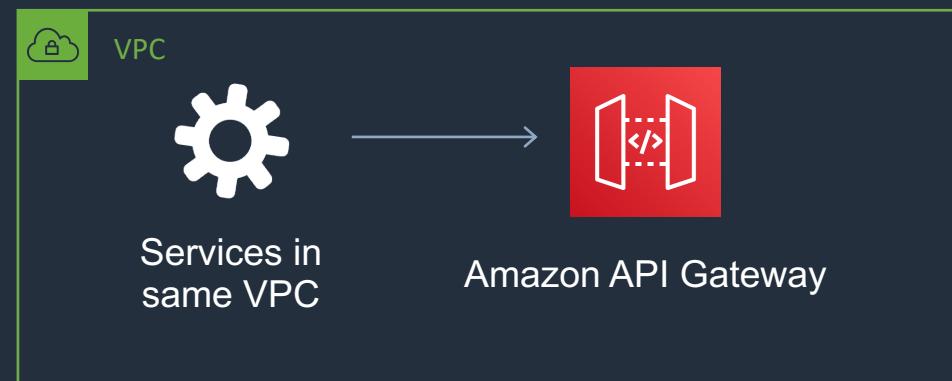
Regional endpoint:



Key benefits:

- Reduced latency for requests that originate in the same region
- Can also configure your own CDN and protect with WAF

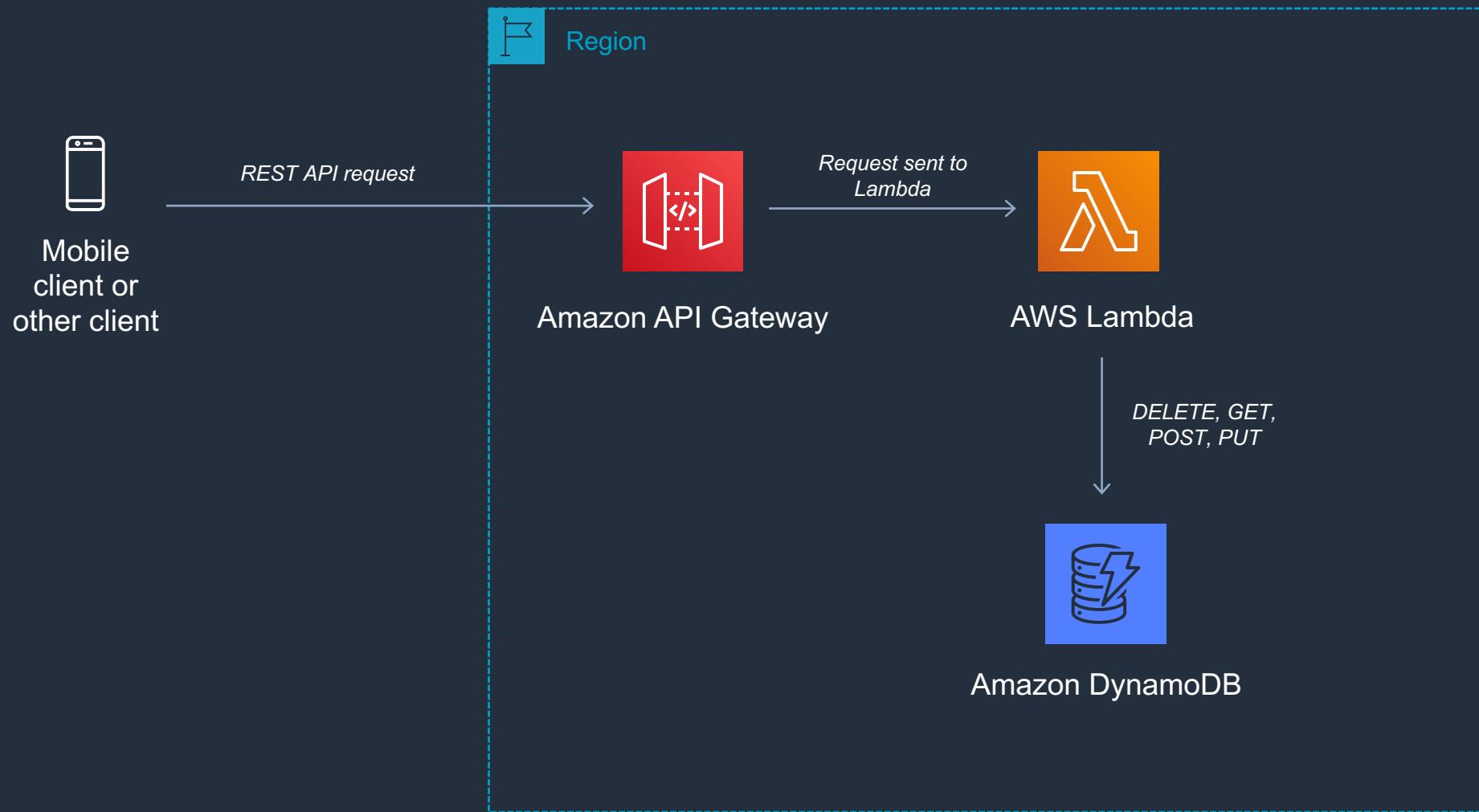
Private endpoint:



Key benefits:

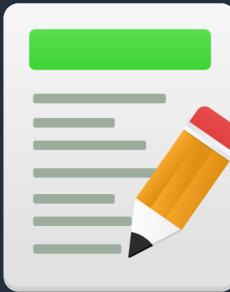
- Securely expose your REST APIs only to other services within your VPC or connect via Direct Connect

Section 10: AWS Lambda – Microservice with Lambda, API Gateway and DynamoDB



Section 10: Exam Cram

AWS Lambda



- AWS Lambda lets you run code as functions without provisioning or managing servers.
- Lambda-based applications (also referred to as serverless applications) are composed of functions triggered by events.
- With serverless computing, your application still runs on servers, but all the server management is done by AWS.
- You specify the amount of memory you need allocated to your Lambda functions.
- AWS Lambda allocates CPU power proportional to the memory you specify using the same ratio as a general purpose EC2 instance type.

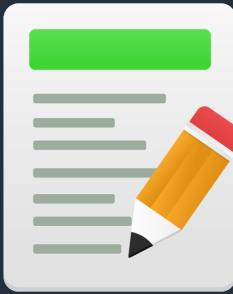
Section 10: Exam Cram

AWS Lambda



- Functions can access:
- AWS services or non-AWS services.
- AWS services running in VPCs (e.g. RedShift, ElastiCache, RDS instances).
- Non-AWS services running on EC2 instances in an AWS VPC.
- Compute resources:
 - You can request additional memory in 64MB increments from 128MB to 3008MB.
 - Functions larger than 1536MB are allocated multiple CPU threads, and multi-threaded or multi-process code is needed to take advantage.

Section 10: Exam Cram

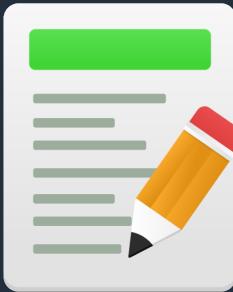


AWS Lambda

- There is a maximum execution timeout.
 - Max is 15 minutes (900 seconds), default is 3 seconds.
 - You pay for the time it runs.
 - Lambda terminates the function at the timeout.
- Lambda is an event-driven compute service where AWS Lambda runs code in response to events such as changes to data in an S3 bucket or a DynamoDB table.
- An event source is an AWS service or developer-created application that produces events that trigger an AWS Lambda function to run.
- Event sources are mapped to Lambda functions.
- Event sources maintain the mapping configuration except for stream-based services (e.g. DynamoDB, Kinesis) for which the configuration is made on the Lambda side and Lambda performs the polling.

Section 10: Exam Cram

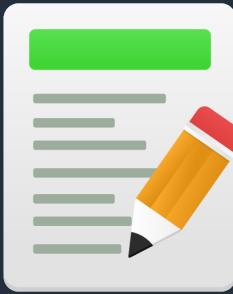
AWS Lambda – Building Lambda Apps



- You can deploy and manage your serverless applications using the AWS Serverless Application Model (AWS SAM).
- AWS SAM is a specification that prescribes the rules for expressing serverless applications on AWS.
- This specification aligns with the syntax used by AWS CloudFormation today and is supported natively within AWS CloudFormation as a set of resource types (referred to as “serverless resources”).
- You can automate your serverless application’s release process using AWS CodePipeline and AWS CodeDeploy.
- You can enable your Lambda function for tracing with AWS X-Ray.

Section 10: Exam Cram

Amazon API Gateway



- An Amazon API Gateway is a collection of resources and methods that are integrated with back-end HTTP endpoints, Lambda functions or other AWS services.
- API Gateway is a fully managed service that makes it easy for developers to publish, maintain, monitor, and secure APIs at any scale.
- API Gateway provides developers with a simple, flexible, fully managed, pay-as-you-go service that handles all aspects of creating and operating robust APIs for application back ends.
- API Gateway handles all of the tasks involved in accepting and processing up to hundreds of thousands of concurrent API calls.
- API calls include traffic management, authorisation and access control, monitoring, and API version management.

Section 10: Exam Cram

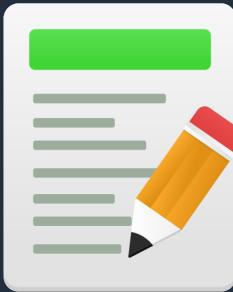
Amazon API Gateway



- Together with Lambda, API Gateway forms the app-facing part of the AWS serverless infrastructure.
- Back-end services include Amazon EC2, AWS Lambda or any web application (public or private endpoints).
- CloudFront is used as the public endpoint for API Gateway.
- All of the APIs created with Amazon API Gateway expose HTTPS endpoints only (does not support unencrypted endpoints).
- An API endpoint type refers to the hostname of the API.
- The API endpoint type can be edge-optimized, regional, or private, depending on where the majority of your API traffic originates from.

Section 10: Exam Cram

Amazon API Gateway



- API Gateway provides several features that assist with creating and managing APIs:
 - **Metering** – Define plans that meter and restrict third-party developer access to APIs.
 - **Security** – API Gateway provides multiple tools to authorize access to APIs and control service operation access.
 - **Resiliency** – Manage traffic with throttling so that backend operations can withstand traffic spikes.
 - **Operations Monitoring** – API Gateway provides a metrics dashboard to monitor calls to services.
 - **Lifecycle Management** – Operate multiple API versions and multiple stages for each version simultaneously so that existing applications can continue to call previous versions after new API versions are published.

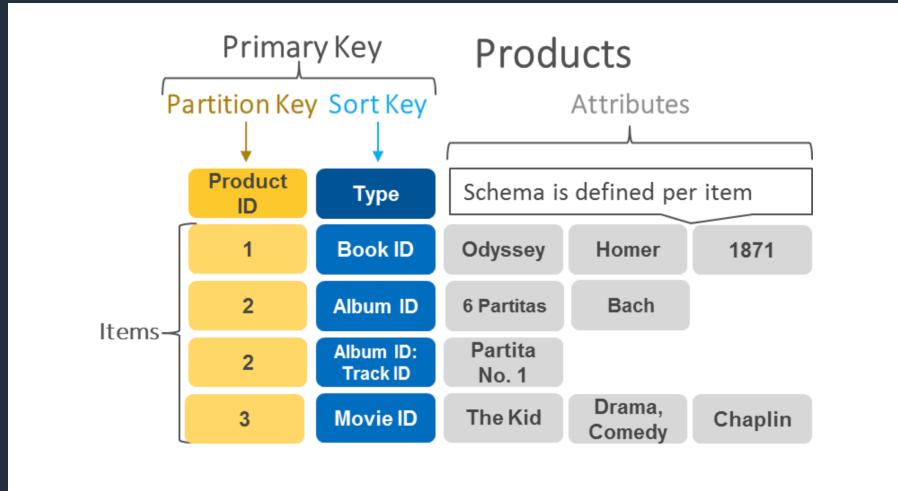
Section 11: Database Types – Relational vs Non-Relational

Key differences are how data are **managed** and how data are **stored**

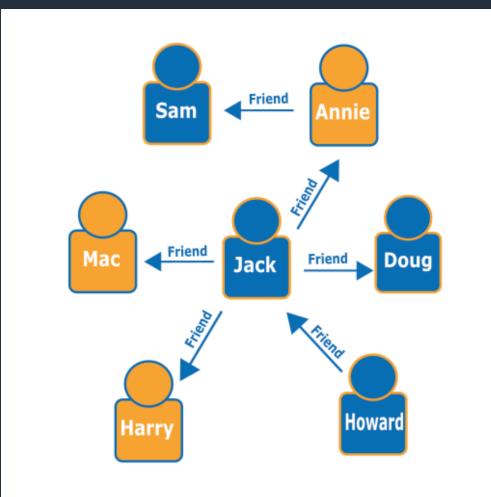
Relational	Non-Relational
Organized by tables, rows and columns	Varied data storage models
Rigid schema (SQL)	Flexible schema (NoSQL) – data stored in key-value pairs, columns, documents or graphs
Rules enforced within database	Rules can be defined in application code (outside database)
Typically scaled vertically	Scales horizontally
Supports complex queries and joins	Unstructured, simple language that supports any kind of schema
ACID (Atomicity, Consistency, Isolation, Durability) compliance typically enforced	Performance is typically prioritised
Amazon RDS, Oracle, MySQL, IBM DB2, PostgreSQL	Amazon DynamoDB, MongoDB, Redis, Neo4j

Section 11: Types of Non-Relational DB

Key-value – e.g. Amazon DynamoDB



Graph – e.g. Amazon Neptune



Document – e.g. MongoDB

```
JSON
1 [           ]
2 {           }
3   "year" : 2013,
4   "title" : "Turn It Down, Or Else!",
5   "info" : {
6     "directors" : [ "Alice Smith", "Bob Jones" ],
7     "release_date" : "2013-01-18T00:00:00Z",
8     "rating" : 6.2,
9     "genres" : [ "Comedy", "Drama" ],
10    "image_url" : "http://ia.media-imdb.com/images/N/09ERWAU7FS797AJ7LU8HN09AMUP908RLlo5JF90EWR7LJKQ7@._V1_SX400_.jpg",
11    "plot" : "A rock band plays their music at high volumes, annoying the neighbors.",
12    "actors" : [ "David Matthewman", "Jonathan G. Neff" ]
13  },
14 },
15 {
16   "year": 2015,
17   "title": "The Big New Movie",
18   "info": {
19     "plot": "Nothing happens at all.",
20     "rating": 0
21   }
22 }
23 ]
```

Section 11: Database Types – Operational vs Analytical

Key differences are **use cases** and how the database is **optimized**

Operational / transactional	Analytical
Online Transaction Processing (OLTP)	Online Analytics Processing (OLAP) – the source data comes from OLTP DBs
Production DBs that process transactions. E.g. adding customer records, checking stock availability (INSERT, UPDATE, DELETE)	Data warehouse. Typically, separated from the customer facing DBs. Data is extracted for decision making
Short transactions and simple queries	Long transactions and complex queries
Relational examples: Amazon RDS, Oracle, IBM DB2, MySQL	Relational examples: Amazon RedShift, Teradata, HP Vertica
Non-relational examples: MongoDB, Cassandra, Neo4j, HBase	Non-relational examples: Amazon EMR, MapReduce

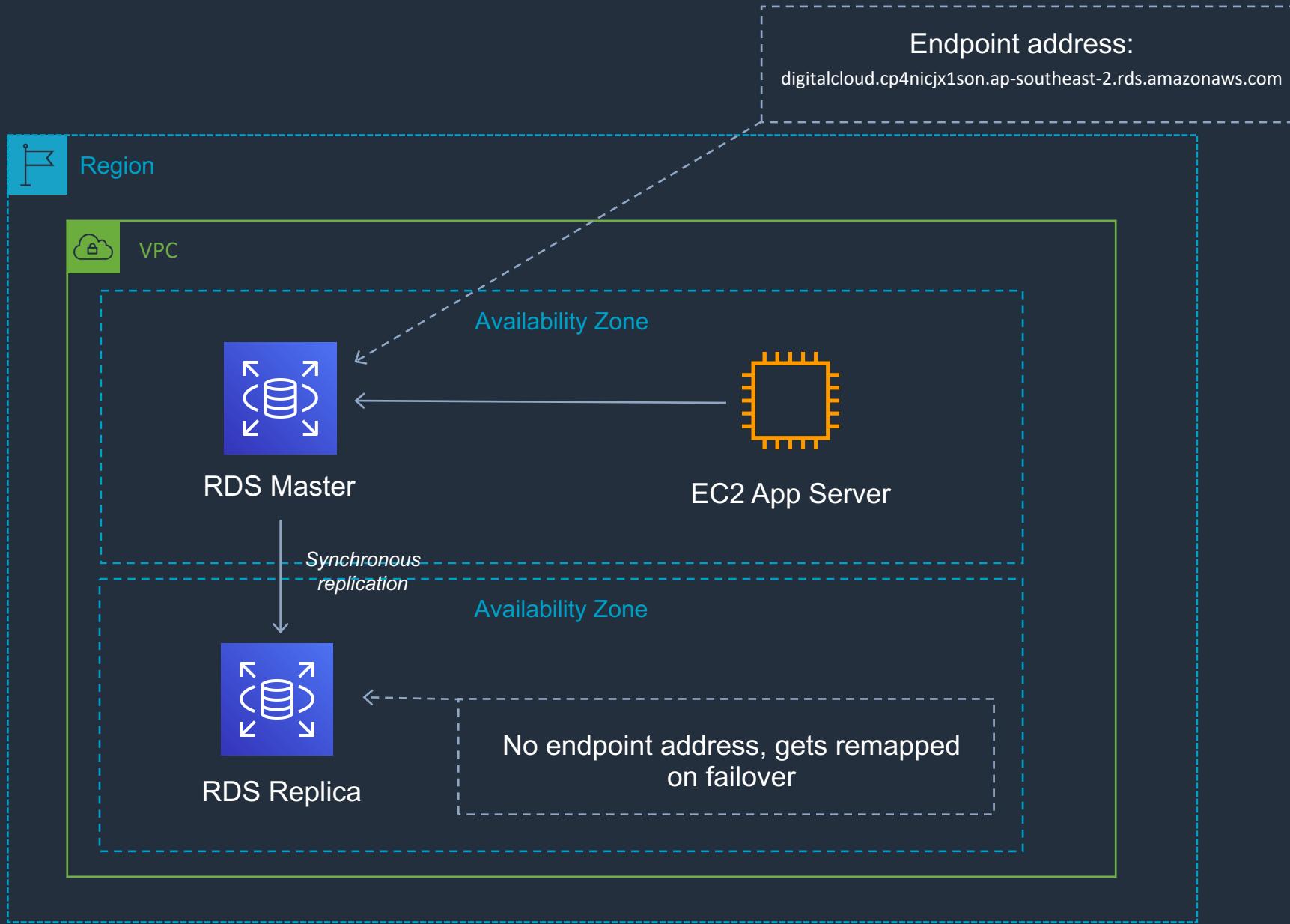
Section 11: Databases –Architecture Discussion

Data Store	When to Use
Database on EC2	<ul style="list-style-type: none">• Full control over instance and database• Preferred DB not available under RDS
Amazon RDS	<ul style="list-style-type: none">• Need traditional relational database for OLTP• Your data is well-formed and structured
Amazon DynamoDB	<ul style="list-style-type: none">• Name/value pair data• Unpredictable data structure• In-memory performance with persistence• High I/O needs• Require dynamic scaling
Amazon RedShift	<ul style="list-style-type: none">• Data warehouse for large volumes of aggregated data• Primarily OLAP workloads
Amazon Neptune	<ul style="list-style-type: none">• Relationships between objects are of high value
Amazon ElastiCache	<ul style="list-style-type: none">• Fast temporary storage for small amounts of data• Highly volatile data (non-persistent)

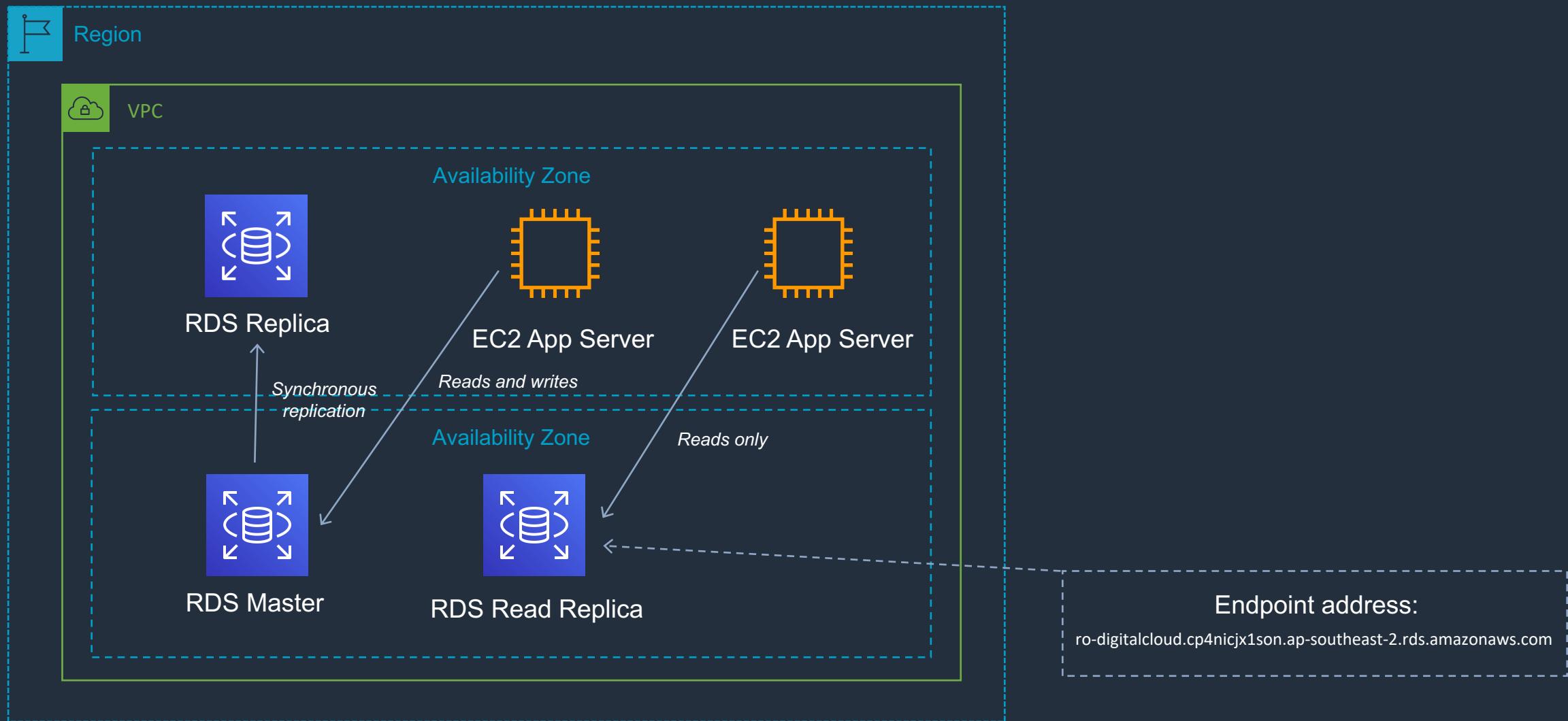
Section 11: Amazon RDS – Multi-AZ and Read Replicas

Multi-AZ Deployments	Read Replicas
Synchronous replication – highly durable	Asynchronous replication – highly scalable
Only database engine on primary instance is active	All read replicas are accessible and can be used for read scaling
Automated backups are taken from standby	No backups configured by default
Always span two Availability Zones within a single Region	Can be within an Availability Zone, Cross-AZ, or Cross-Region
Database engine version upgrades happen on primary	Database engine version upgrade is independent from source instance
Automatic failover to standby when a problem is detected	Can be manually promoted to a standalone database instance

Section 11: Amazon RDS Multi-AZ



Section 11: Amazon RDS Read Replicas



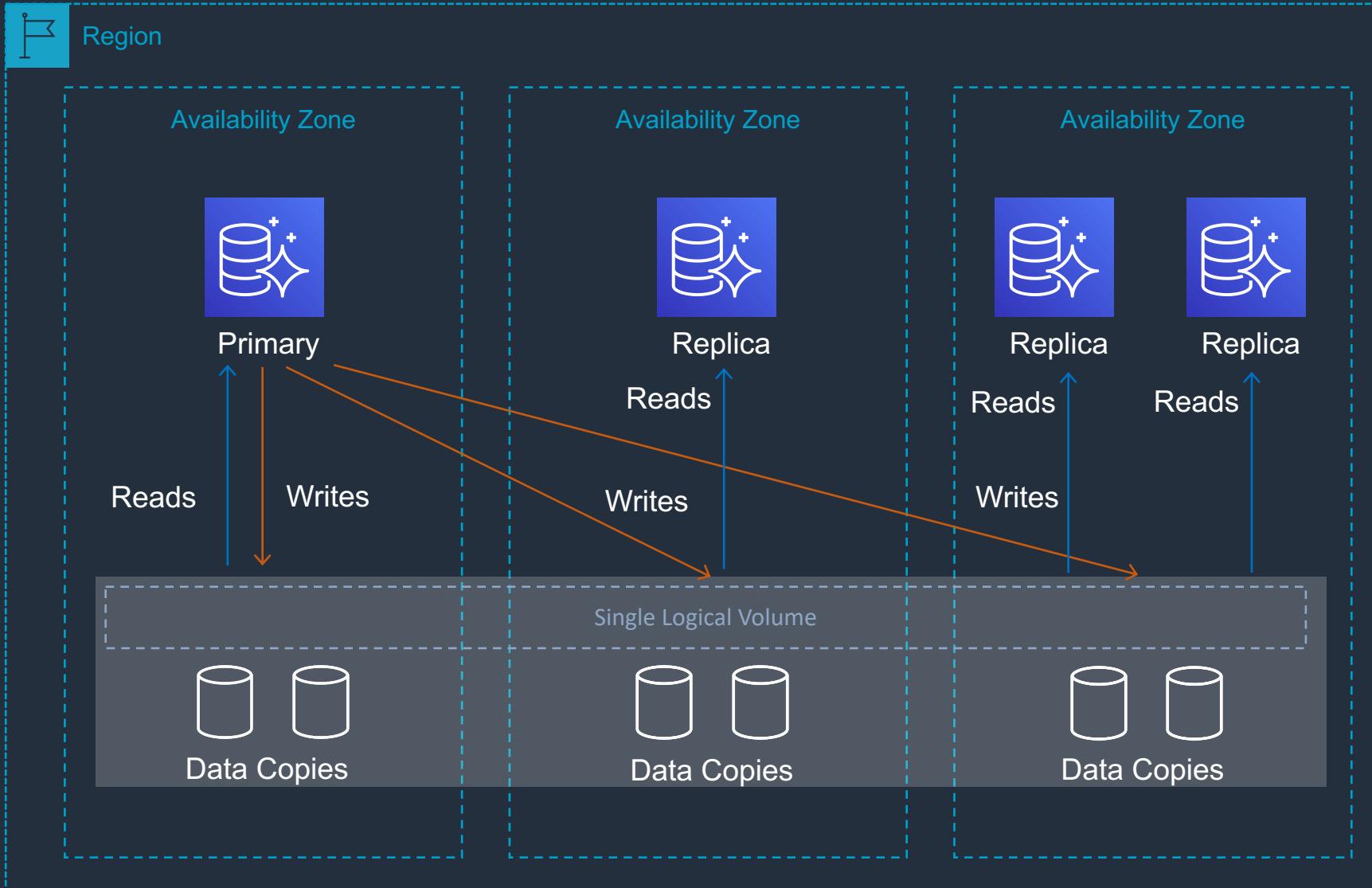
Section 11: Amazon RDS Aurora Key Features

Aurora Feature	Benefit
High performance and scalability	Offers high performance, self-healing storage that scales up to 64TB, point-in-time recovery and continuous backup to S3
DB compatibility	Compatible with existing MySQL and PostgreSQL open source databases
Aurora Replicas	In-region read scaling and failover target – up to 15 (can use Auto Scaling)
MySQL Read Replicas	Cross-region cluster with read scaling and failover target – up to 5 (each can have up to 15 Aurora Replicas)
Global Database	Cross-region cluster with read scaling (fast replication / low latency reads). Can remove secondary and promote
Multi-Master	Scales out writes within a region. In preview currently and will not appear on the exam
Serverless	On-demand, autoscaling configuration for Amazon Aurora - does not support read replicas or public IPs (can only access through VPC or Direct Connect - not VPN)

Section 11: Amazon RDS Aurora Replicas

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

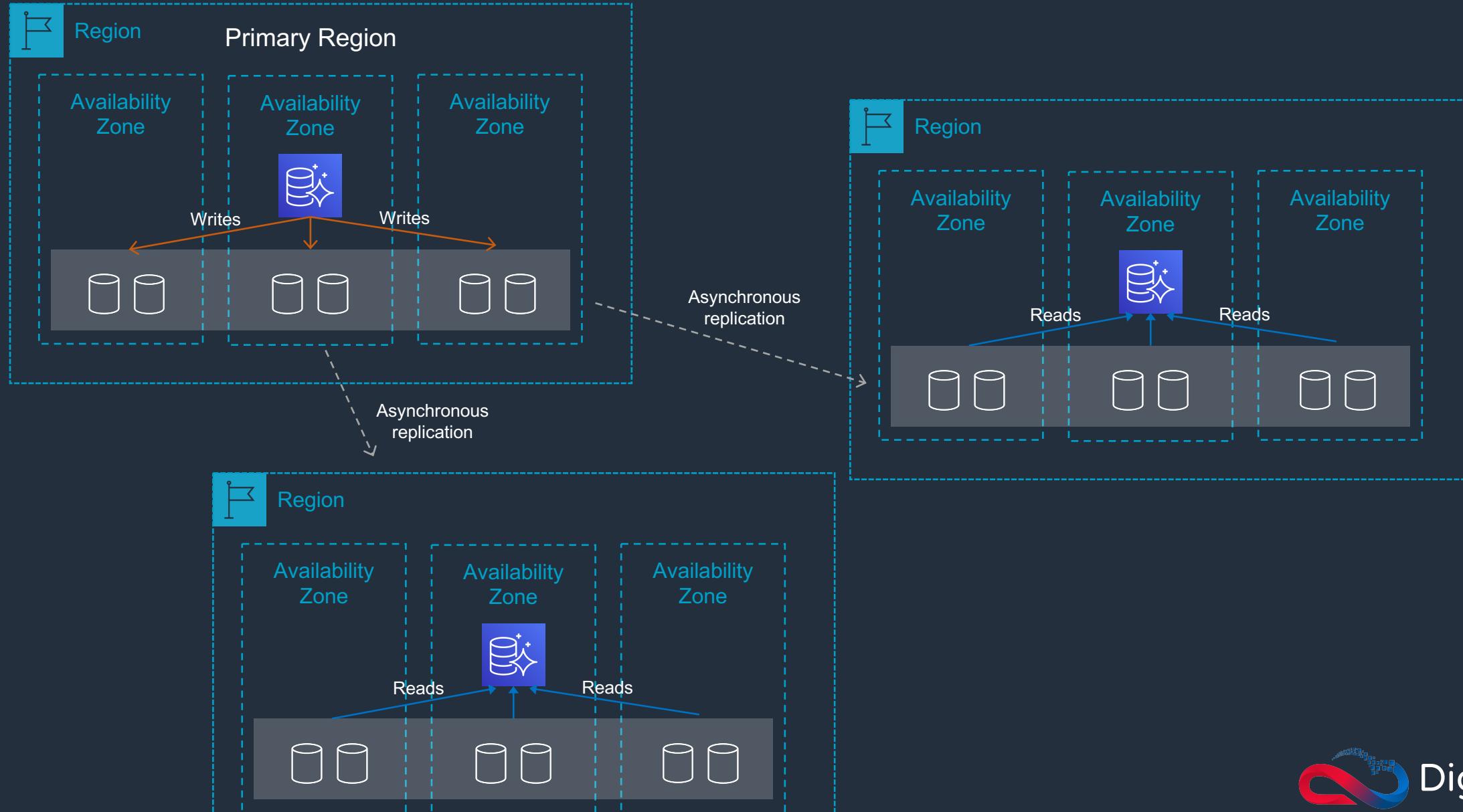
Section 11: Aurora Fault Tolerance and Aurora Replicas



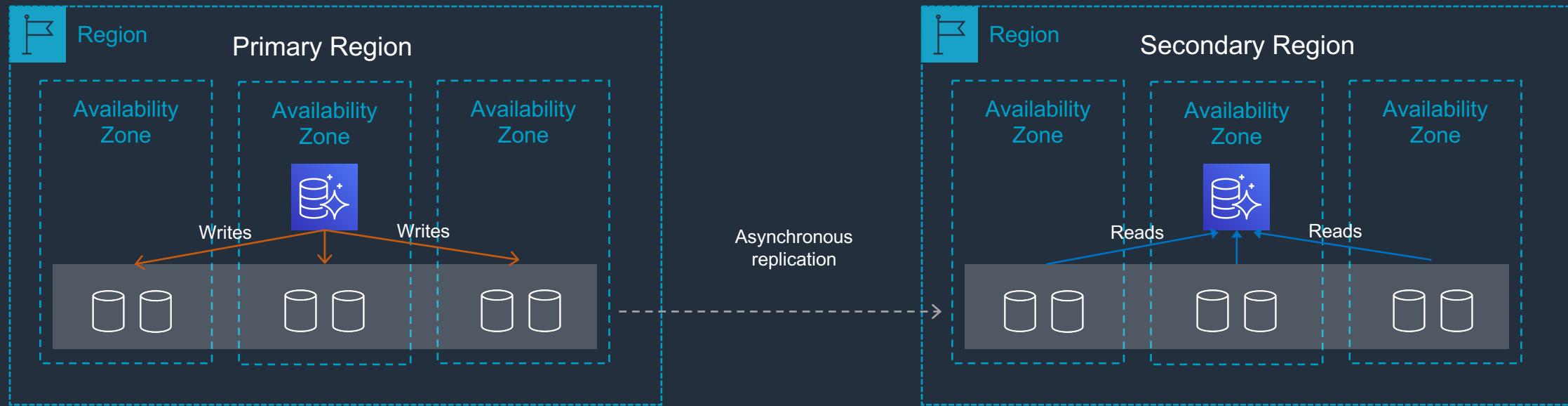
Aurora Fault Tolerance

- Fault tolerance across 3 AZs
- Single logical volume
- Aurora Replicas scale-out read requests
- Up to 15 Aurora Replicas with sub-10ms replica lag
- Aurora Replicas are independent endpoints
- Can promote Aurora Replica to be a new primary or create new primary
- Set priority (tiers) on Aurora Replicas to control order of promotion
- Can use Auto Scaling to add replicas

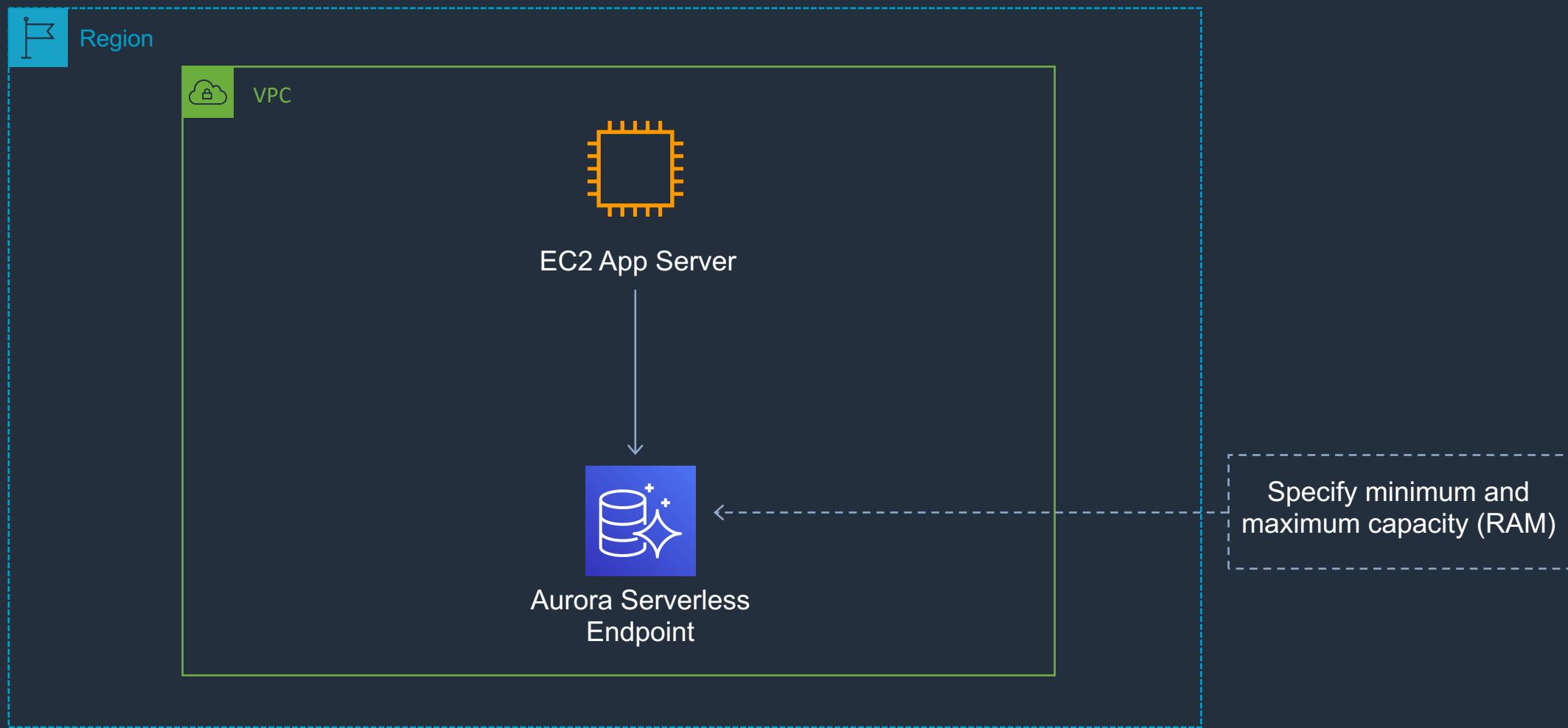
Section 11: Cross-Region Replica with Aurora MySQL



Section 11: Aurora Global Database



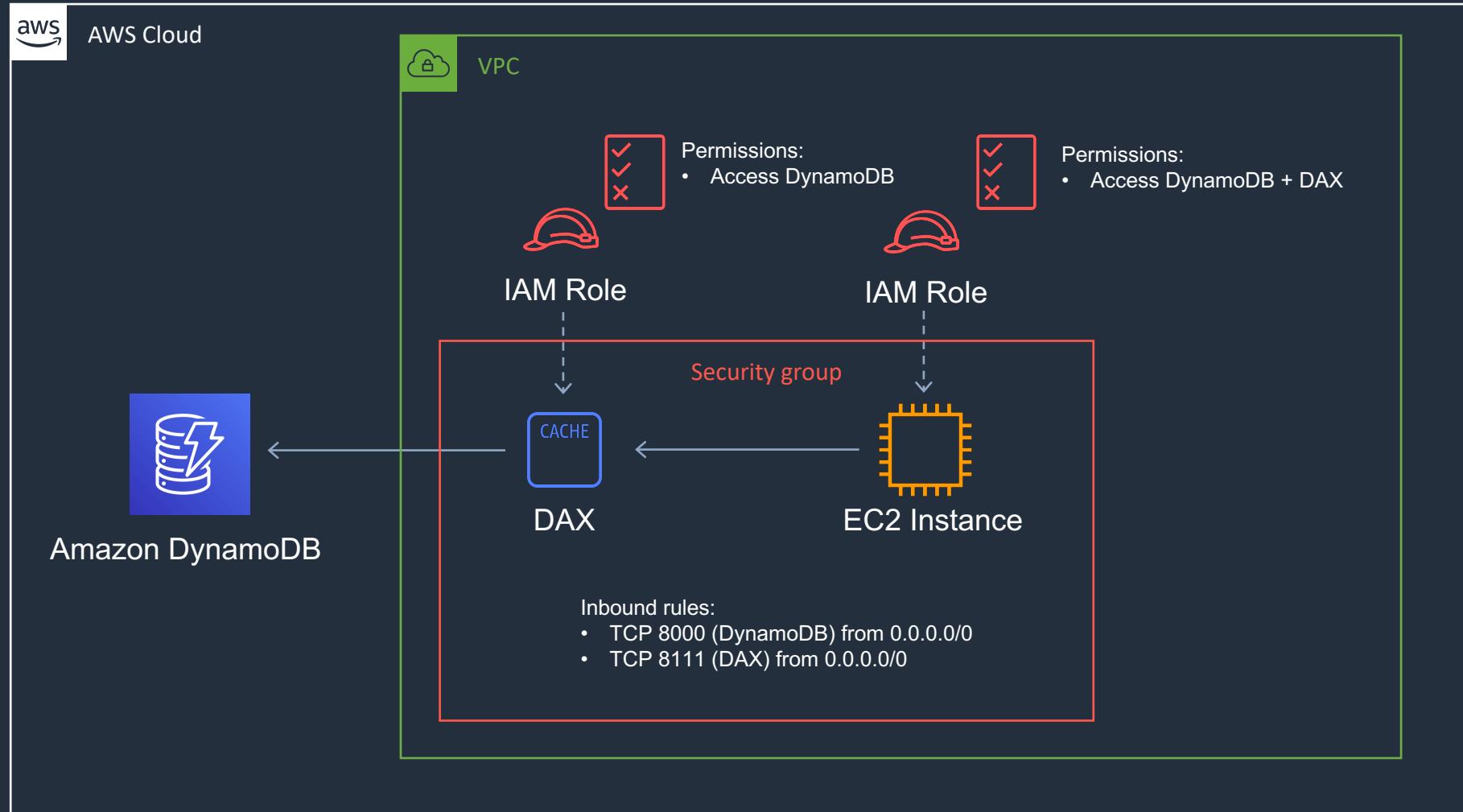
Section 11: Aurora Serverless



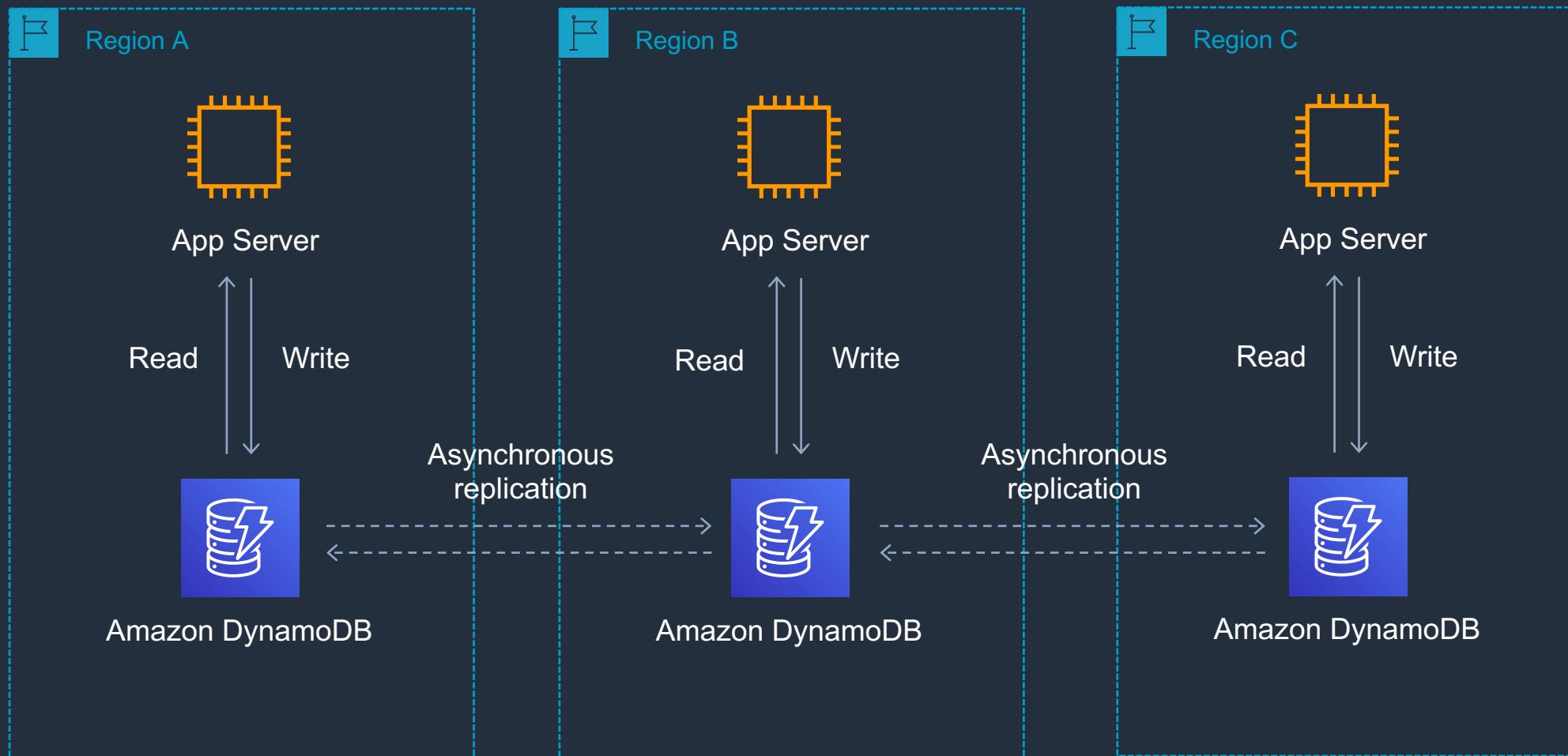
Section 11: DynamoDB Overview

DynamoDB Feature	Benefit
Serverless	Fully managed, fault tolerant, service
Highly available	99.99% availability SLA – 99.999% for Global Tables!
NoSQL type of database with Name / Value structure	Flexible schema, good for when data is not well structured or unpredictable
Horizontal scaling	Seamless scalability to any scale with push button scaling or Auto Scaling
DynamoDB Streams	Captures a time-ordered sequence of item-level modifications in a DynamoDB table and durably stores the information for up to 24 hours. Often used with Lambda and the Kinesis Client Library (KCL)
DynamoDB Accelerator (DAX)	Fully managed in-memory cache for DynamoDB that increases performance (microsecond latency)
Transaction options	Strongly consistent or eventually consistent reads, support for ACID transactions
Backup	Point-in-time recovery down to the second in last 35 days; On-demand backup and restore
Global Tables	Fully managed multi-region, multi-master solution

Section 11: DynamoDB Accelerator (DAX)



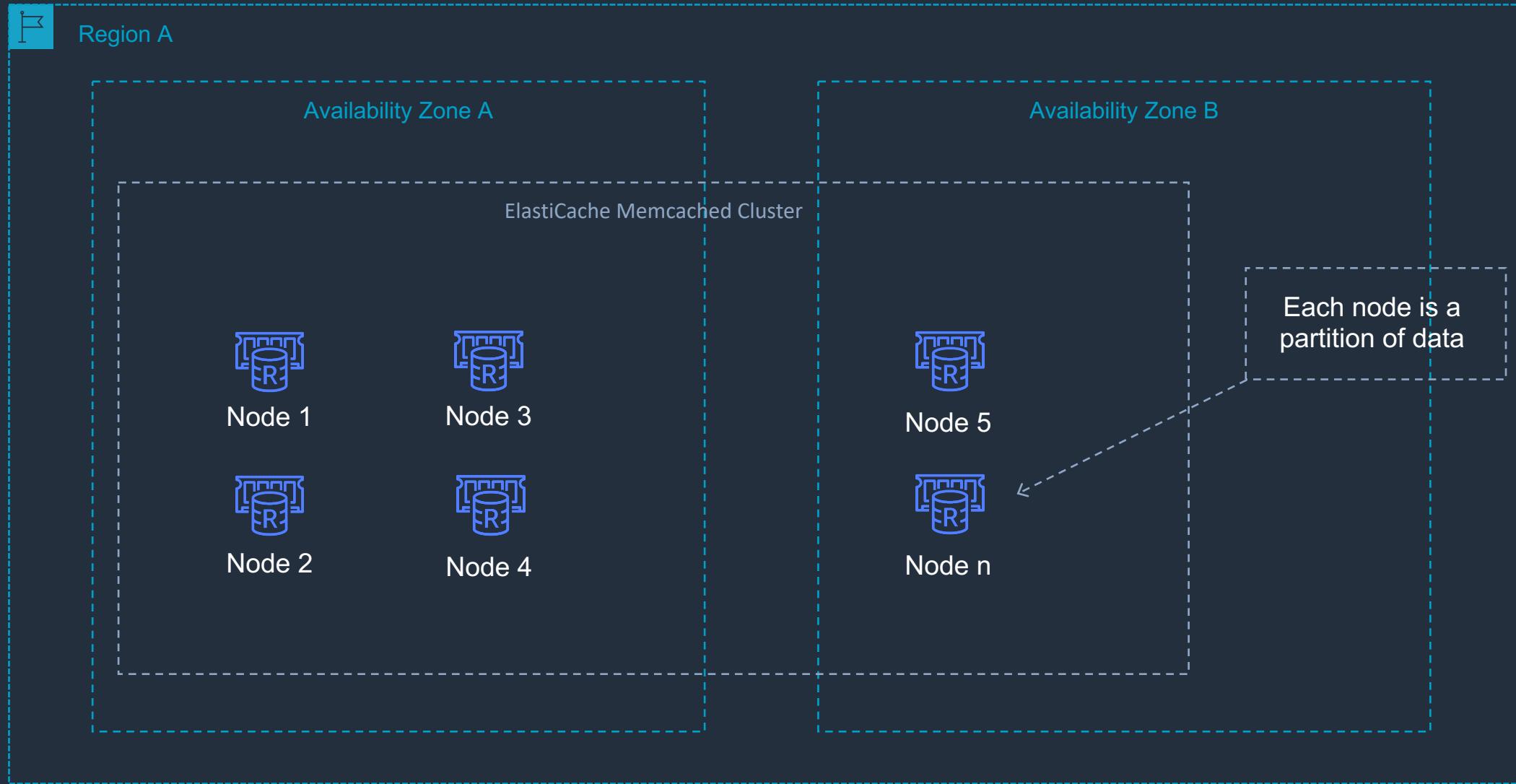
Section 11: DynamoDB Global Tables



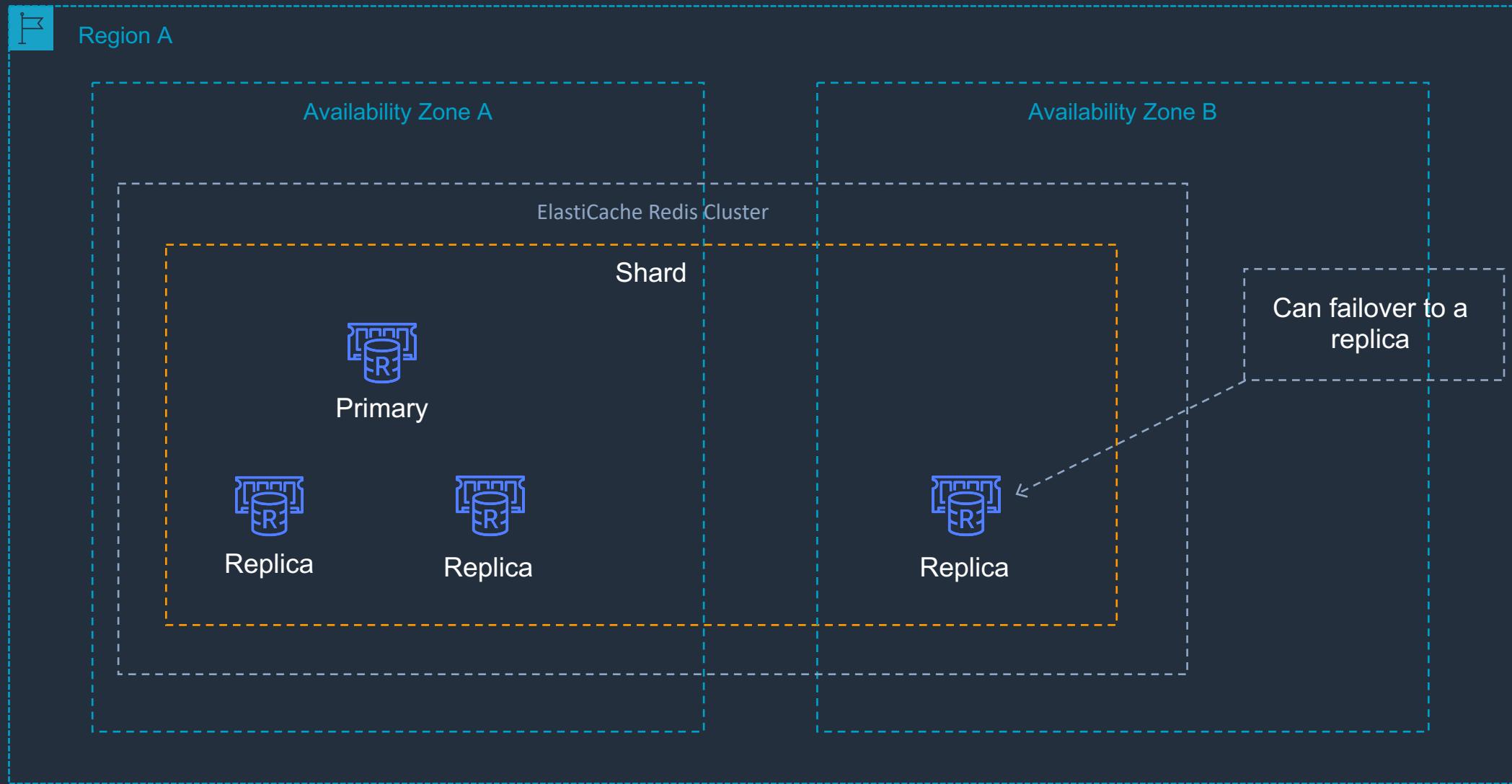
Section 11: ElastiCache Overview

Feature	Memcached	Redis (cluster mode disabled)	Redis (cluster mode enabled)
Data persistence	No	Yes	Yes
Data types	Simple	Complex	Complex
Data partitioning	Yes	No	Yes
Encryption	No	Yes	Yes
High availability (replication)	No	Yes	Yes
Multi-AZ	Yes, place nodes in multiple AZs. No failover or replication	Yes, with auto-failover. Uses read replicas (0-5 per shard)	Yes, with auto-failover. Uses read replicas (0-5 per shard)
Scaling	Up (node type); out (add nodes)	Single shard (can add replicas)	Add shards
Multithreaded	Yes	No	No
Backup and restore	No (and no snapshots)	Yes, automatic and manual snapshots	Yes, automatic and manual snapshots

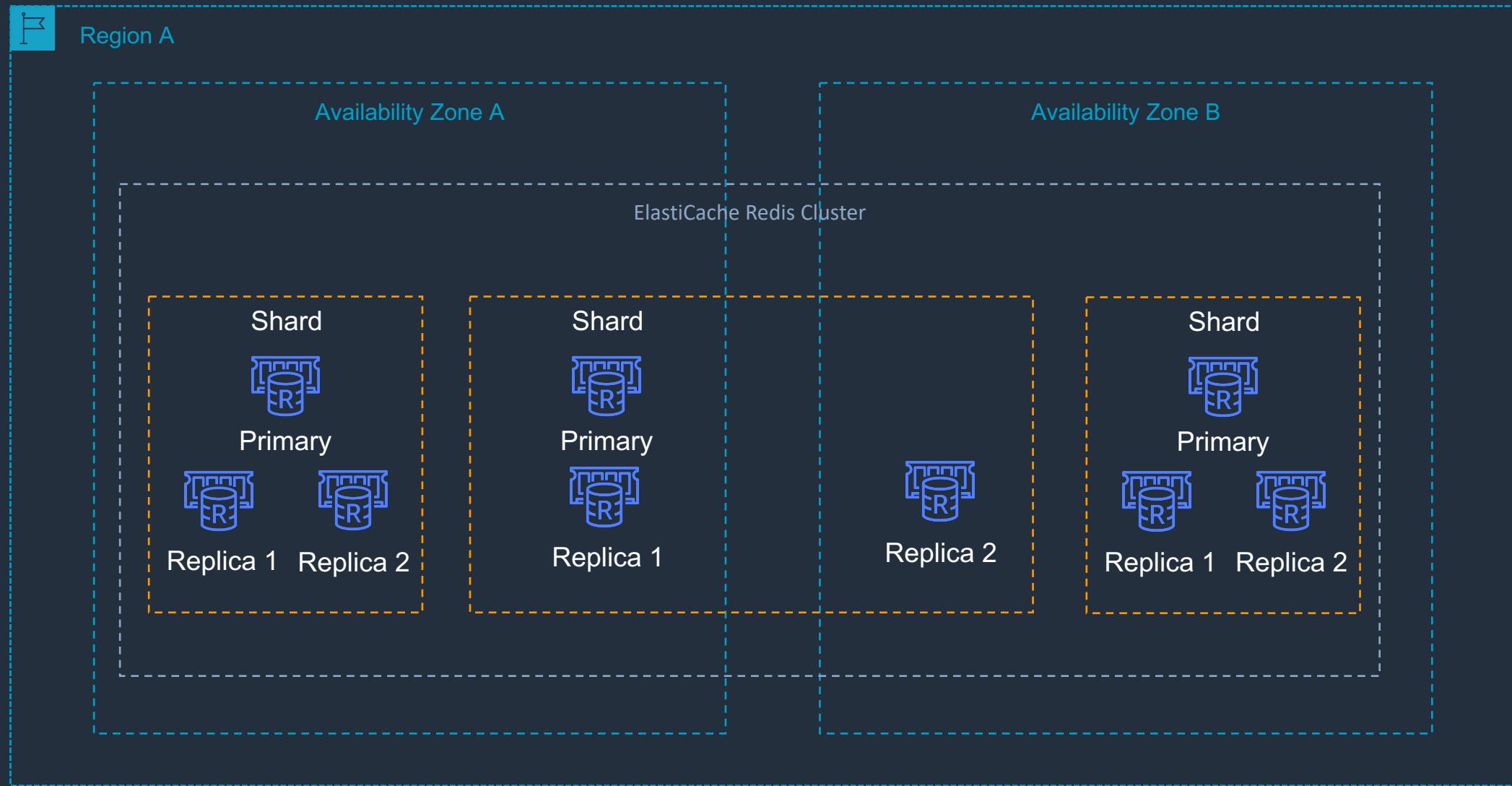
Section 11: ElastiCache Memcached



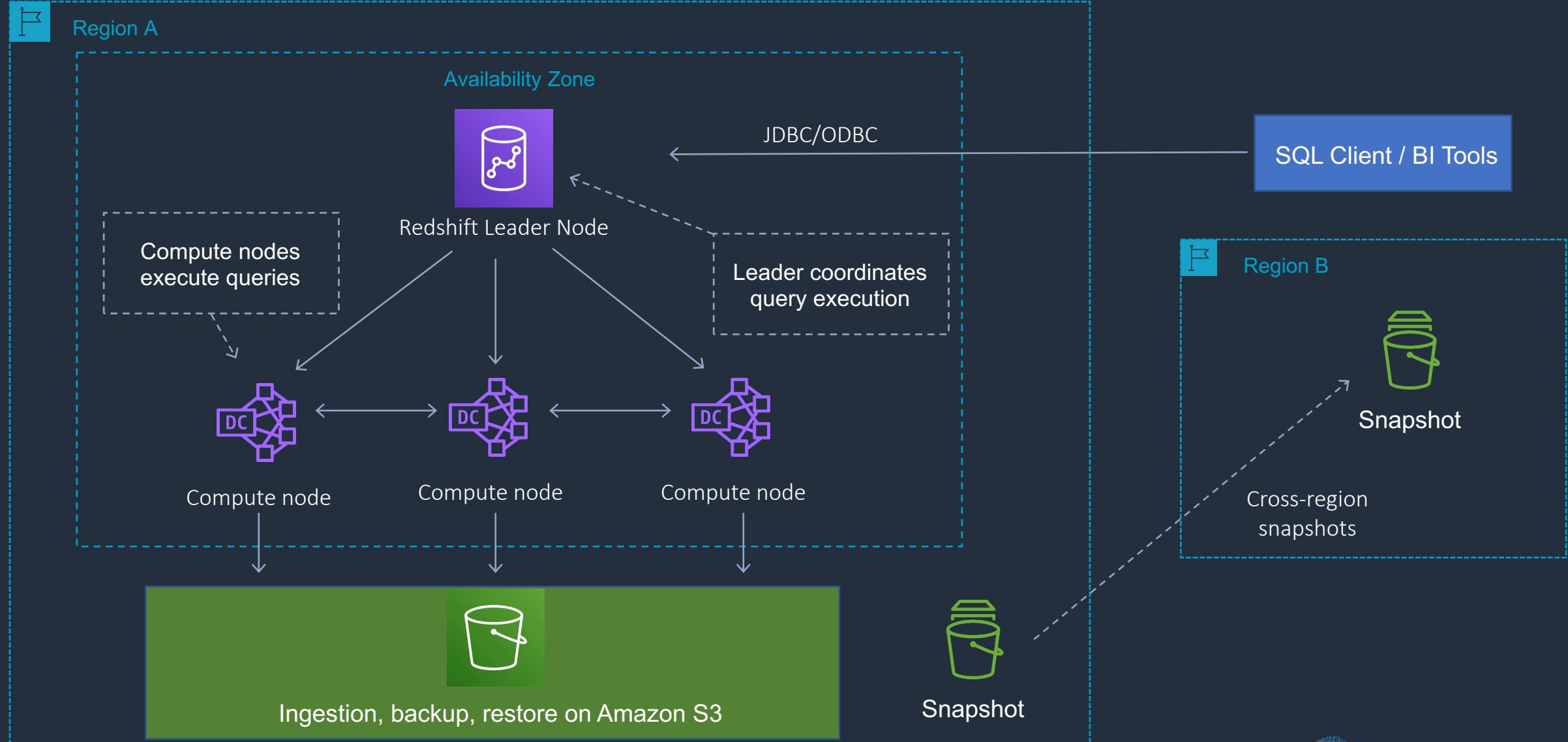
Section 11: ElastiCache Redis (Cluster mode disabled)



Section 11: ElastiCache Redis (Cluster mode enabled)

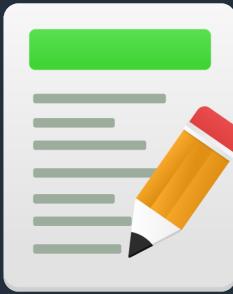


Section 11: Amazon RedShift



Section 11: Exam Cram

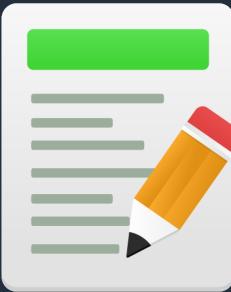
Amazon RDS



- Amazon Relational Database Service (Amazon RDS) is a managed service that makes it easy to set up, operate, and scale a relational database in the cloud.
- RDS is an Online Transaction Processing (OLTP) type of database.
- Primary use case is a transactional database (rather than analytical).
- Best for structured, relational data store requirements.
- Aims to be a drop-in replacement for existing on-premise instances of the same databases.
- Automated backups and patching applied in customer-defined maintenance windows.
- Push-button scaling, replication and redundancy.

Section 11: Exam Cram

Amazon RDS DB Engines

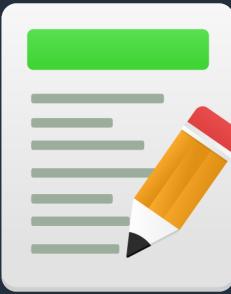


- Amazon RDS supports the following database engines:
 - Amazon Aurora.
 - MySQL.
 - MariaDB.
 - Oracle.
 - SQL Server.
 - PostgreSQL.

Section 11: Exam Cram

Amazon RDS

- RDS is a managed service and you do not have access to the underlying EC2 instance (no root access).
- The RDS service includes the following:
 - Security and patching of the DB instances.
 - Automated backup for the DB instances.
 - Software updates for the DB engine.
 - Easy scaling for storage and compute.
 - Multi-AZ option with synchronous replication.
 - Automatic failover for Multi-AZ option.
 - Read replicas option for read heavy workloads.



Section 11: Exam Cram

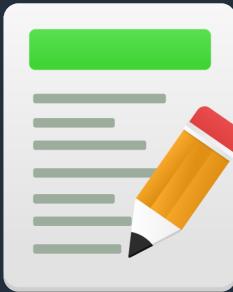
Amazon RDS Alternatives



- If your use case isn't supported on RDS, you can run databases on Amazon EC2. .
- Consider the following points when considering a DB on EC2:
 - You can run any database you like with full control and ultimate flexibility.
 - You must manage everything like backups, redundancy, patching and scaling.
 - Good option if you require a database not yet supported by RDS, such as IBM DB2 or SAP HANA.
 - Good option if it is not feasible to migrate to AWS-managed database.

Section 11: Exam Cram

Amazon RDS Encryption



- You can encrypt your Amazon RDS instances and snapshots at rest by enabling the encryption option for your Amazon RDS DB instance.
- Encryption at rest is supported for all DB types and uses AWS KMS.
- When using encryption at rest the following elements are also encrypted:
 - All DB snapshots.
 - Backups.
 - DB instance storage.
 - Read Replicas.
- You cannot encrypt an existing DB, you need to create a snapshot, copy it, encrypt the copy, then build an encrypted DB from the snapshot.

Section 11: Exam Cram

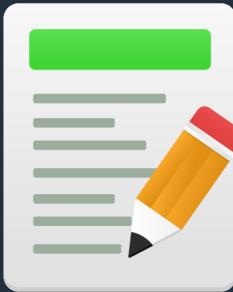
Amazon RDS Encryption



- Data that is encrypted at rest includes the underlying storage for a DB instance, its automated backups, Read Replicas, and snapshots.
- A Read Replica of an Amazon RDS encrypted instance is also encrypted using the same key as the master instance when both are in the same region.
- If the master and Read Replica are in different regions, you encrypt using the encryption key for that region.
- You can't have an encrypted Read Replica of an unencrypted DB instance or an unencrypted Read Replica of an encrypted DB instance.

Section 11: Exam Cram

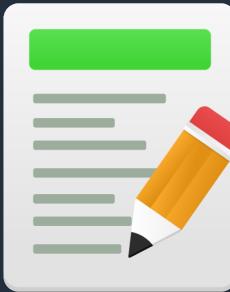
Amazon RDS Charges



- AWS Charge for:
 - DB instance hours (partial hours are charged as full hours).
 - Storage GB/month.
 - I/O requests/month – for magnetic storage.
 - Provisioned IOPS/month – for RDS provisioned IOPS SSD.
 - Egress data transfer.
 - Backup storage (DB backups and manual snapshots).
- Backup storage for the automated RDS backup is free of charge up to the provisioned EBS volume size.

Section 11: Exam Cram

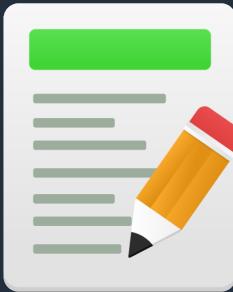
Amazon RDS Charges



- AWS Charge for:
 - DB instance hours (partial hours are charged as full hours).
 - Storage GB/month.
 - I/O requests/month – for magnetic storage.
 - Provisioned IOPS/month – for RDS provisioned IOPS SSD.
 - Egress data transfer.
 - Backup storage (DB backups and manual snapshots).
- Backup storage for the automated RDS backup is free of charge up to the provisioned EBS volume size.
- However, AWS replicate data across multiple AZs and so you are charged for the extra storage space on S3.

Section 11: Exam Cram

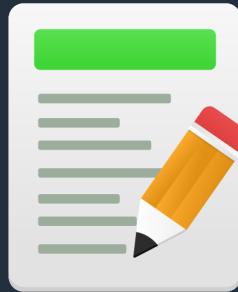
Amazon RDS Charges



- For multi-AZ you are charged for:
 - Multi-AZ DB hours.
 - Provisioned storage.
 - Double write I/Os.
- For multi-AZ you are not charged for DB data transfer during replication from primary to standby.
- Oracle and Microsoft SQL licences are included or you can bring your own (BYO).
- On-demand and reserved instance pricing available.

Section 11: Exam Cram

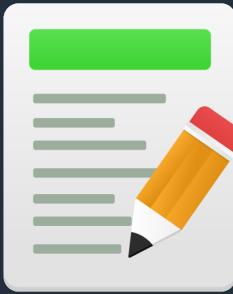
Amazon RDS Scalability



- You can only scale RDS up (compute and storage).
- You cannot decrease the allocated storage for an RDS instance.
- You can scale storage and change the storage type for all DB engines except MS SQL.
- For MS SQL the workaround is to create a new instance from a snapshot with the new configuration.
- Scaling storage can happen while the RDS instance is running without outage however there may be performance degradation.
- Scaling compute will cause downtime.
- You can choose to have changes take effect immediately, however the default is within the maintenance window.

Section 11: Exam Cram

Amazon RDS Performance



- Amazon RDS uses EBS volumes (never uses instance store) for DB and log storage.
- There are three storage types available: General Purpose (SSD), Provisioned IOPS (SSD), and Magnetic.
- Magnetic is not recommended anymore, available for backwards compatibility.

Section 11: Exam Cram

Amazon RDS Multi-AZ and Read Replicas



- Multi-AZ and Read Replicas are used for high availability, fault tolerance and performance scaling.

Multi-AZ Deployments	Read Replicas
Synchronous replication – highly durable	Asynchronous replication – highly scalable
Only database engine on primary instance is active	All read replicas are accessible and can be used for read scaling
Automated backups are taken from standby	No backups configured by default
Always span two Availability Zones within a single Region	Can be within an Availability Zone, Cross-AZ, or Cross-Region
Database engine version upgrades happen on primary	Database engine version upgrade is independent from source instance
Automatic failover to standby when a problem is detected	Can be manually promoted to a standalone database instance

Section 11: Exam Cram

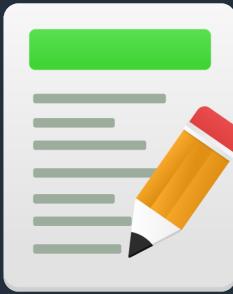
Amazon RDS Multi-AZ



- Multi-AZ RDS creates a replica in another AZ and synchronously replicates to it (DR only).
- There is an option to choose multi-AZ during the launch wizard.
- AWS recommends the use of provisioned IOPS storage for multi-AZ RDS DB instances.
- Each AZ runs on its own physically distinct, independent infrastructure, and is engineered to be highly reliable.
- You cannot choose which AZ in the region will be chosen to create the standby DB instance.
- You can view which AZ the standby DB instance is created in.

Section 11: Exam Cram

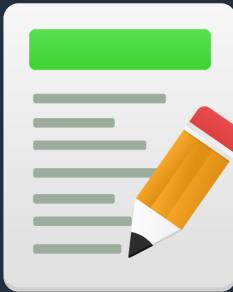
Amazon RDS Multi-AZ



- Read Replica Support for Multi-AZ:
 - Amazon RDS Read Replicas for MySQL and MariaDB support Multi-AZ deployments.
 - Combining Read Replicas with Multi-AZ enables you to build a resilient disaster recovery strategy and simplify your database engine upgrade process.
 - A Read Replica in a different region than the source database can be used as a standby database and promoted to become the new production database in case of a regional disruption.
 - This allows you to scale reads whilst also having multi-AZ for DR.
 - Note that RDS for PostgreSQL does not yet support this feature.

Section 11: Exam Cram

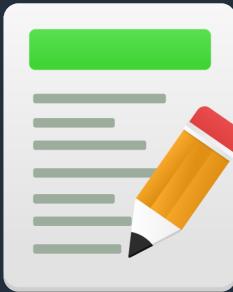
Amazon RDS Read Replicas



- Read replicas are used for read heavy DBs and replication is asynchronous.
- Read replicas are for workload sharing and offloading.
- Read replicas provide read-only DR.
- Read replicas are created from a snapshot of the master instance.
- Must have automated backups enabled on the primary (retention period > 0).
- Only supported for transactional database storage engines (InnoDB not MyISAM).
- Read replicas are available for MySQL, PostgreSQL, MariaDB, Oracle and Aurora (not SQL Server).
- You can enable automatic backups on MySQL and MariaDB read replicas.
- You can enable writes to the MySQL and MariaDB Read Replicas.

Section 11: Exam Cram

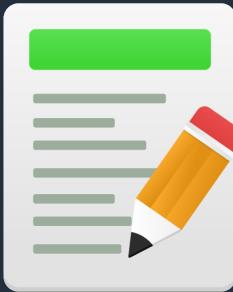
Amazon RDS Read Replicas



- In a multi-AZ failover the read replicas are switched to the new primary.
- You can have 5 read replicas of a production DB.
- You cannot have more than four instances involved in a replication chain.
- You can have read replicas of read replicas for MySQL and MariaDB but not for PostgreSQL.
- You can promote a read replica to primary.

Section 11: Exam Cram

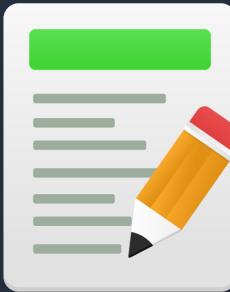
Amazon RDS DB Snapshots



- DB Snapshots are user-initiated and enable you to back up your DB instance in a known state as frequently as you wish, and then restore to that specific state.
- Snapshots are stored on S3.
- Snapshots remain on S3 until manually deleted.
- Backups are taken within a defined window.
- Restored DBs will always be a new RDS instance with a new DNS endpoint.
- You cannot restore from a DB snapshot to an existing DB - a new instance is created when you restore.

Section 11: Exam Cram

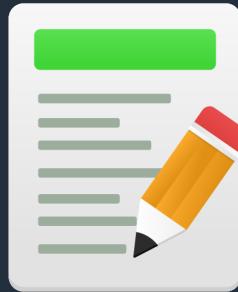
Amazon RDS Migration



- AWS Database Migration Service helps you migrate databases to AWS quickly and securely.
- Use along with the Schema Conversion Tool (SCT) to migrate databases to AWS RDS or EC2-based databases.
- The source database remains fully operational during the migration, minimizing downtime to applications that rely on the database.
- The AWS Database Migration Service can migrate your data to and from most widely used commercial and open-source databases.
- Schema Conversion Tool can copy database schemas for homogenous migrations (same database) and convert schemas for heterogeneous migrations (different database).

Section 11: Exam Cram

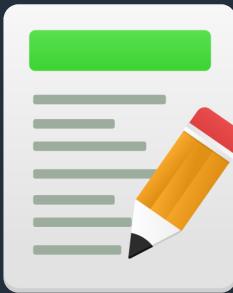
Amazon Aurora



- Aurora is an AWS proprietary database.
- Fully managed service.
- High performance, low price.
- Scales in 10GB increments.
- Scales up to 32vCPUs and 244GB RAM.
- 2 copies of data are kept in each AZ with a minimum of 3 AZ's (6 copies).
- Can handle the loss of up to two copies of data without affecting DB write availability and up to three copies without affecting read availability.

Section 11: Exam Cram

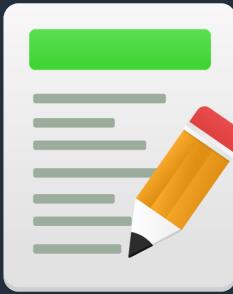
- Amazon Aurora Replicas:
 - There are two types of replication: Aurora replica (up to 15), MySQL Read Replica (up to 5).
 - Aurora Replica is in-region and MySQL Replica is cross-region.
- Amazon Aurora Cross-Region Replicas:
 - Cross-region read replicas allow you to improve your disaster recovery posture, scale read operations in regions closer to your application users, and easily migrate from one region to another.
 - Cross-region replicas provide fast local reads to your users.
 - Each region can have an additional 15 Aurora replicas to further scale local reads.
 - You can choose between Global Database, which provides the best replication performance, and traditional binlog-based replication.



Section 11: Exam Cram

Amazon Aurora Global Database

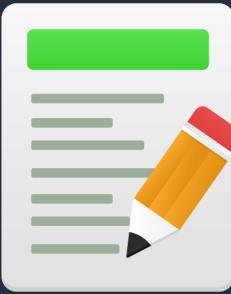
- For globally distributed applications you can use Global Database, where a single Aurora database can span multiple AWS regions to enable fast local reads and quick disaster recovery.
- Global Database uses storage-based replication to replicate a database across multiple AWS Regions, with typical latency of less than 1 second.
- You can use a secondary region as a backup option in case you need to recover quickly from a regional degradation or outage.
- A database in a secondary region can be promoted to full read/write capabilities in less than 1 minute.



Section 11: Exam Cram

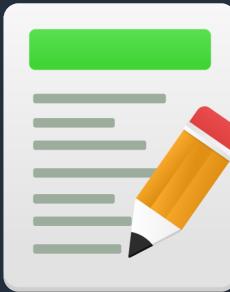
Amazon Aurora Multi-Master

- Amazon Aurora Multi-Master is a new feature of the Aurora MySQL-compatible edition that adds the ability to scale out write performance across multiple Availability Zones, allowing applications to direct read/write workloads to multiple instances in a database cluster and operate with higher availability.
- Aurora Multi-Master is designed to achieve high availability and ACID transactions across a cluster of database nodes with configurable read after write consistency.



Section 11: Exam Cram

Amazon Aurora Serverless – features on SAA-C02

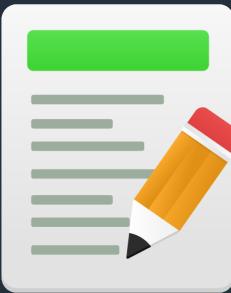


- Amazon Aurora Serverless is an on-demand, auto-scaling configuration for Amazon Aurora.
- Available for MySQL-compatible and PostgreSQL-compatible editions.
- The database automatically starts up, shuts down, and scales capacity up or down based on application needs.
- It enables you to run a database in the cloud without managing any database instances. It's a simple, cost-effective option for infrequent, intermittent, or unpredictable workloads.
- You simply create a database endpoint and optionally specify the desired database capacity range and connect applications.
- With Aurora Serverless, you only pay for database storage and the database capacity and I/O your database consumes while it is active.
- Pay on a per-second basis for the database capacity you use when the database is active.

Section 11: Exam Cram

Amazon DynamoDB

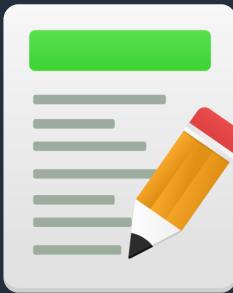
- Amazon DynamoDB is a fully managed NoSQL database service that provides fast and predictable performance with seamless scalability.
- Multi-AZ NoSQL data store with Cross-Region Replication option.
- Push button scaling means that you can scale the DB at any time without incurring downtime.
- Defaults to eventual consistency reads but can request strongly consistent read via SDK parameter.
- Can achieve ACID compliance with DynamoDB Transactions.
- SSD based and uses limited indexing on attributes for performance.
- Data is synchronously replicated across 3 facilities (AZs) in a region.
- DynamoDB is schema-less.
- DynamoDB can be used for storing session state.



Section 11: Exam Cram

Amazon DynamoDB

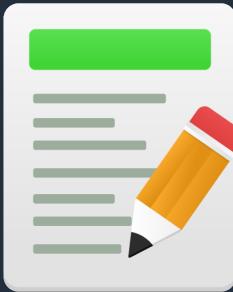
- Provides two read models.
- Eventually consistent reads (Default):
 - The eventual consistency option maximises your read throughput (best read performance).
 - An eventually consistent read might not reflect the results of a recently completed write.
 - Consistency across all copies reached within 1 second.
- Strongly consistent reads:
 - A strongly consistent read returns a result that reflects all writes that received a successful response prior to the read (faster consistency).



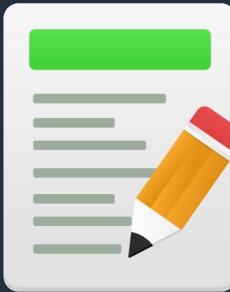
Section 11: Exam Cram

Amazon DynamoDB Streams

- DynamoDB Streams help you to keep a list of item level changes or provide a list of item level changes that have taken place in the last 24hrs.
- Amazon DynamoDB is integrated with AWS Lambda so that you can create triggers—pieces of code that automatically respond to events in DynamoDB Streams.
- If you enable DynamoDB Streams on a table, you can associate the stream ARN with a Lambda function that you write.



Section 11: Exam Cram



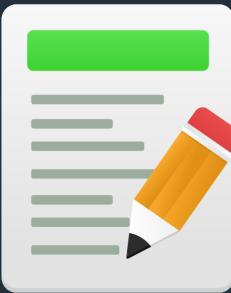
Amazon DynamoDB DAX

- Amazon DynamoDB Accelerator (DAX) is a fully managed, highly available, in-memory cache for DynamoDB that delivers up to a 10x performance improvement .
- Improves performance from milliseconds to microseconds, even at millions of requests per second.
- DAX does all the heavy lifting required to add in-memory acceleration to your DynamoDB tables, without requiring developers to manage cache invalidation, data population, or cluster management.
- You do not need to modify application logic, since DAX is compatible with existing DynamoDB API calls.
- Provisioned through clusters and charged by the node (runs on EC2 instances).

Section 11: Exam Cram

Amazon DynamoDB Cross Region Replication with Global Tables

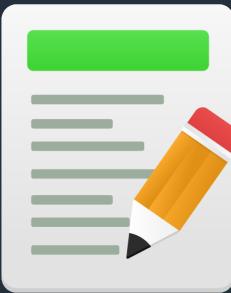
- Amazon DynamoDB global tables provide a fully managed solution for deploying a multi-region, multi-master database.
- When you create a global table, you specify the AWS regions where you want the table to be available.
- DynamoDB performs all of the necessary tasks to create identical tables in these regions and propagate ongoing data changes to all of them.
- DynamoDB global tables are ideal for massively scaled applications, with globally dispersed users.
- Global tables provide automatic multi-master replication to AWS regions world-wide, so you can deliver low-latency data access to your users no matter where they are located.



Section 11: Exam Cram

Amazon DynamoDB Auto Scaling

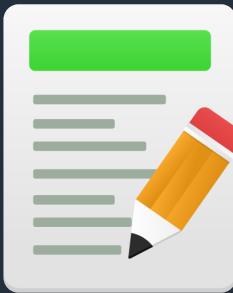
- DynamoDB auto scaling uses the AWS Application Auto Scaling service to dynamically adjust provisioned throughput capacity on your behalf, in response to actual traffic patterns.
- This enables a table or a global secondary index to increase its provisioned read and write capacity to handle sudden increases in traffic, without throttling.
- When the workload decreases, Application Auto Scaling decreases the throughput so that you don't pay for unused provisioned capacity.



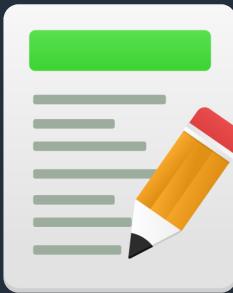
Section 11: Exam Cram

Amazon DynamoDB Charges

- DynamoDB charges for reading, writing, and storing data in your DynamoDB tables, along with any optional features you choose to enable.
- There are two pricing models for DynamoDB:
 - **On-demand capacity mode:** DynamoDB charges you for the data reads and writes your application performs on your tables. You do not need to specify how much read and write throughput you expect your application to perform because DynamoDB instantly accommodates your workloads as they ramp up or down.
 - **Provisioned capacity mode:** you specify the number of reads and writes per second that you expect your application to require. You can use auto scaling to automatically adjust your table's capacity based on the specified utilization rate to ensure application performance while reducing cost.



Section 11: Exam Cram



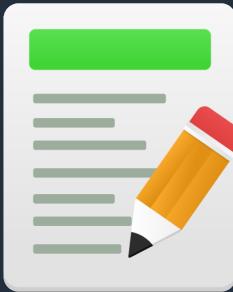
Amazon ElastiCache

- Fully managed implementations of two popular in-memory data stores – Redis and Memcached.
- ElastiCache is a web service that makes it easy to deploy and run Memcached or Redis protocol-compliant server nodes in the cloud.
- The in-memory caching provided by ElastiCache can be used to significantly improve latency and throughput for many read-heavy application workloads or compute-intensive workloads .
- Best for scenarios where the DB load is based on Online Analytics Processing (OLAP) transactions.
- Push-button scalability for memory, writes and reads.
- In-memory key/value store – not persistent in the traditional sense.
- Elasticache can be used for storing session state.

Section 11: Exam Cram

Amazon ElastiCache

- A node is a fixed-sized chunk of secure, network-attached RAM and is the smallest building block.
- Each node runs an instance of the Memcached or Redis protocol-compliant service and has its own DNS name and port.
- Failed nodes are automatically replaced.
- Access to ElastiCache nodes is controlled by VPC security groups and subnet groups (when deployed in a VPC).
- Subnet groups are a collection of subnets designated for your Amazon ElastiCache Cluster.



Section 11: Exam Cram

Amazon ElastiCache Use Cases



Use Case	Benefit
Web session store	In cases with load-balanced web servers, store web session information in Redis so if a server is lost, the session info is not lost, and another web server can pick it up
Database caching	Use Memcached in front of AWS RDS to cache popular queries to offload work from RDS and return results faster to users
Leaderboards	Use Redis to provide a live leaderboard for millions of users of your mobile app
Streaming data dashboards	Provide a landing spot for streaming sensor data on the factory floor, providing live real-time dashboard displays

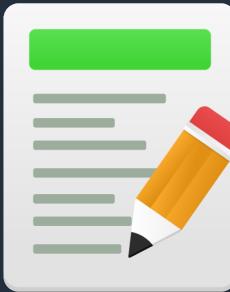
Section 11: Exam Cram

Amazon ElastiCache Requirements



Memcached	Redis
Simple, no-frills	You need encryption
You need elasticity (scale out and in)	You need HIPAA compliance
You need to run multiple CPU cores and threads	Support for clustering
You need to cache objects (e.g. database queries)	You need complex data types
	You need HA (replication)
	Pub/Sub capability
	Geospatial Indexing
	Backup and restore

Section 11: Exam Cram



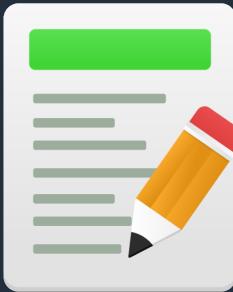
Amazon ElastiCache Memcached

- Not persistent.
- Cannot be used as a data store.
- Supports large nodes with multiple cores or threads.
- Scales out and in, by adding and removing nodes.
- Ideal front-end for data stores (RDS, Dynamo DB etc.).
- Scales out/in (horizontally) by adding/removing nodes.
- Scales up/down (vertically) by changing the node family/type.
- Does not support multi-AZ failover or replication.
- Does not support snapshots.
- You can place nodes in different AZs.

Section 11: Exam Cram

Amazon ElastiCache Redis

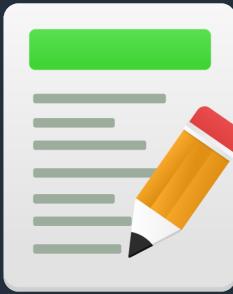
- Data is persistent.
- Can be used as a datastore.
- Not multi-threaded.
- Scales by adding shards, not nodes.
- A Redis shard is a subset of the cluster's keyspace, that can include a primary node and zero or more read-replicas.
- Multi-AZ is possible using read replicas in another AZ in the same region.



Section 11: Exam Cram

Amazon RedShift

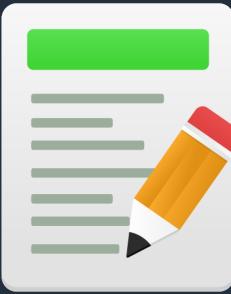
- Amazon Redshift is a fast, fully managed data warehouse that makes it simple and cost-effective to analyze all your data using standard SQL and existing Business Intelligence (BI) tools.
- Clustered peta-byte scale data warehouse.
- RedShift is a SQL based data warehouse used for analytics applications.
- RedShift is an Online Analytics Processing (OLAP) type of DB.
- RedShift is used for running complex analytic queries against petabytes of structured data, using sophisticated query optimization, columnar storage on high-performance local disks, and massively parallel query execution.
- RedShift is ideal for processing large amounts of data for business intelligence.



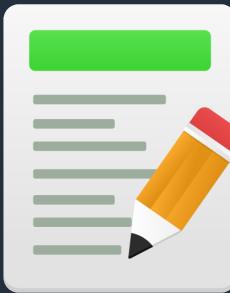
Section 11: Exam Cram

Amazon RedShift

- Features parallel processing and columnar data stores which are optimized for complex queries.
- Option to query directly from data files on S3 via RedShift Spectrum.
- RedShift uses EC2 instances so you need to choose your instance type/size for scaling compute vertically, but you can also scale horizontally by adding more nodes to the cluster.
- You cannot have direct access to your AWS RedShift cluster nodes as a user, but you can through applications.



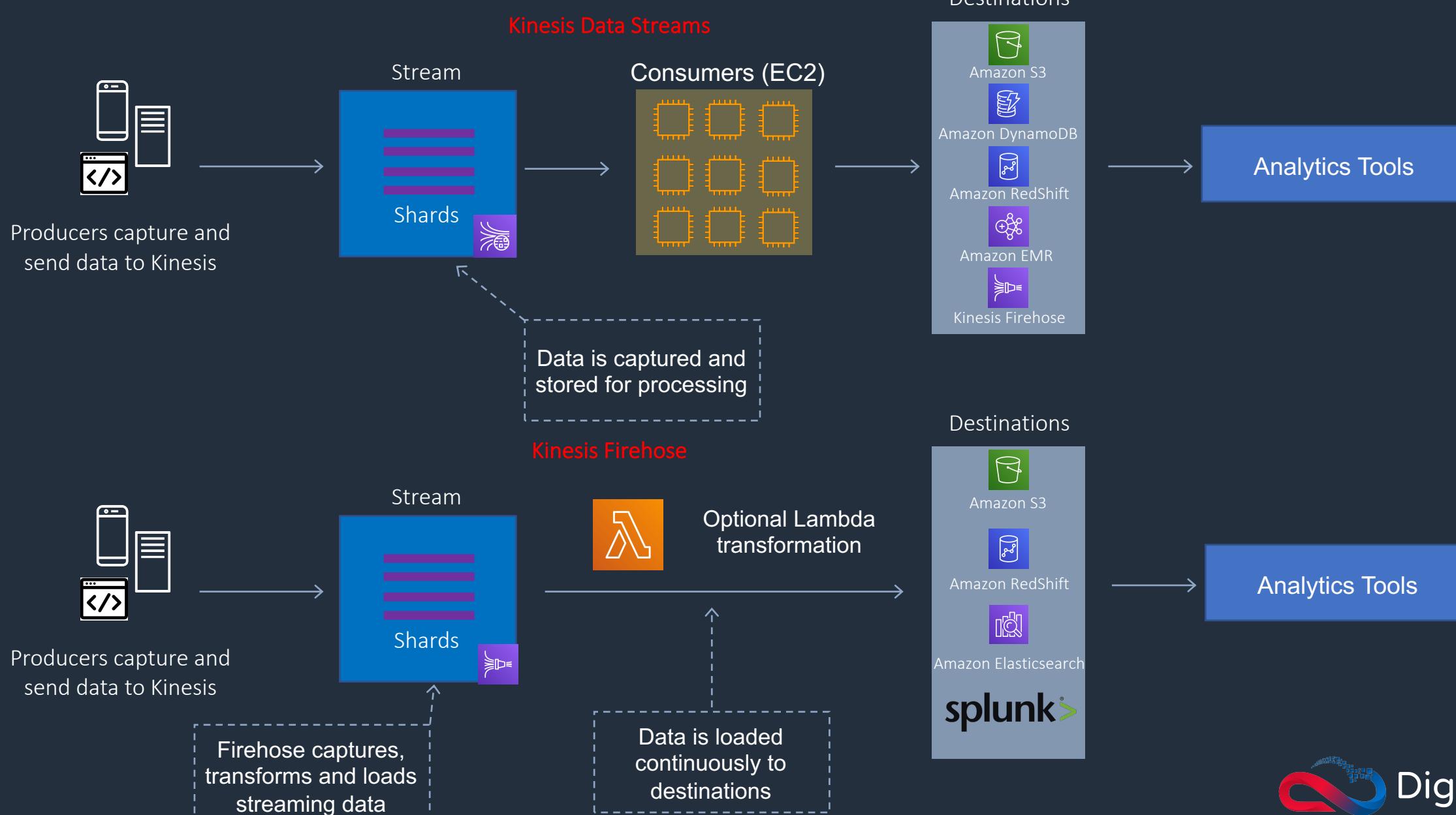
Section 11: Exam Cram



Amazon RedShift Availability and Durability

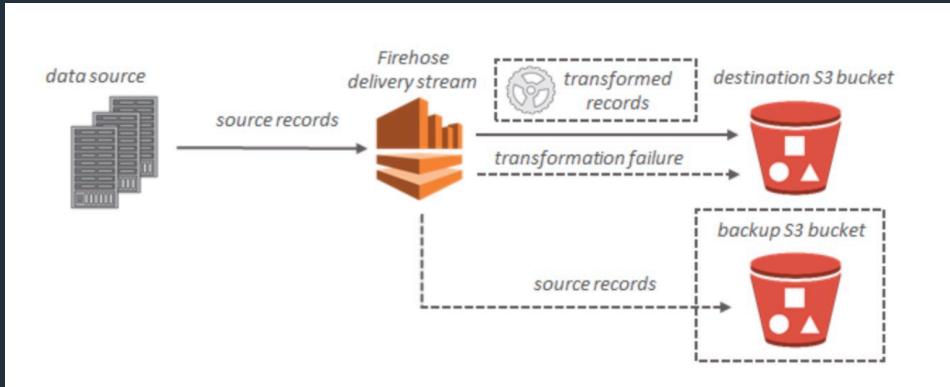
- RedShift uses replication and continuous backups to enhance availability and improve durability and can automatically recover from component and node failures.
- Only available in one AZ but you can restore snapshots into another AZ.
- Alternatively, you can run data warehouse clusters in multiple AZ's by loading data into two Amazon Redshift data warehouse clusters in separate AZs from the same set of Amazon S3 input files.
- Redshift replicates your data within your data warehouse cluster and continuously backs up your data to Amazon S3.
- RedShift always keeps three copies of your data:
 - The original.
 - A replica on compute nodes (within the cluster).
 - A backup copy on S3.

Section 12: Amazon Kinesis Data Streams and Firehose

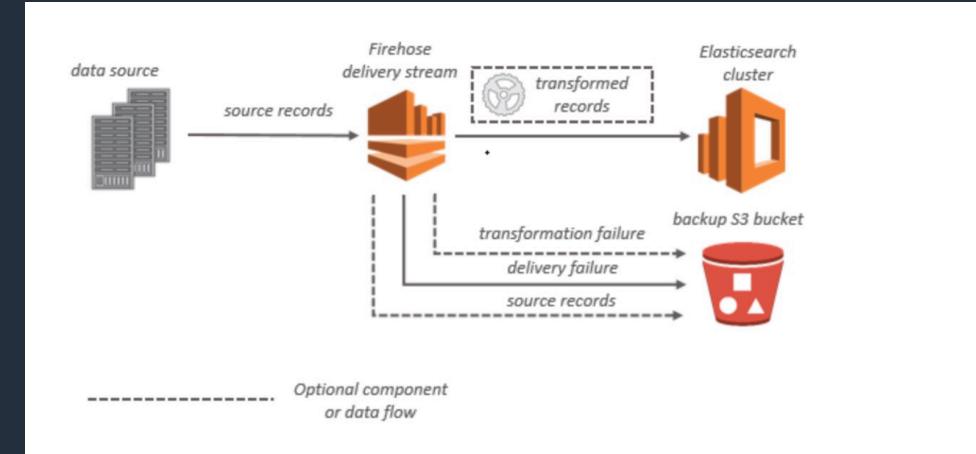


Section 12: Amazon Kinesis Firehose Destinations

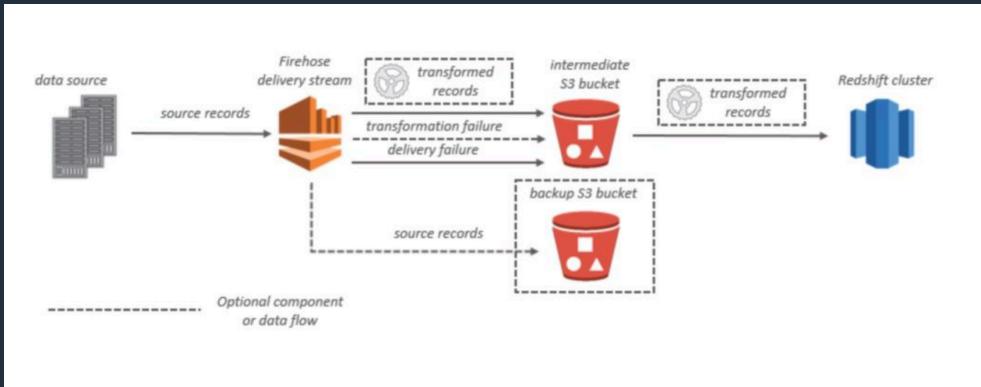
Amazon S3: delivered to S3 bucket, optional backup



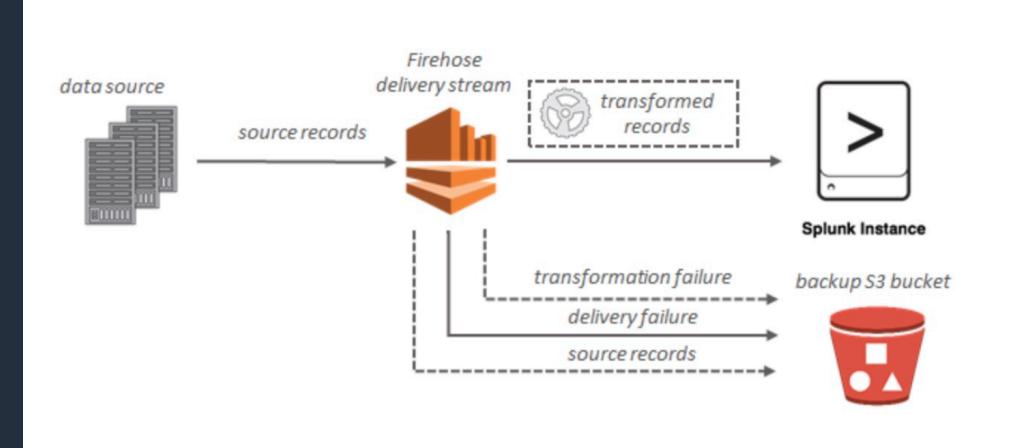
Amazon Elasticsearch: delivered to ES and optionally to S3



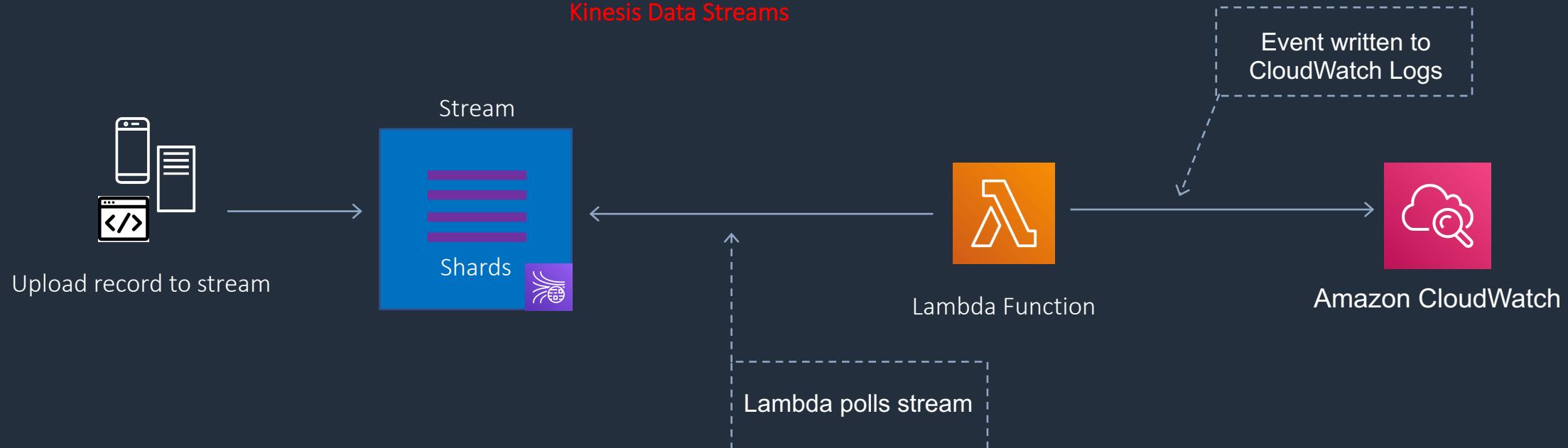
Amazon RedShift: delivered to S3 bucket first, then RedShift



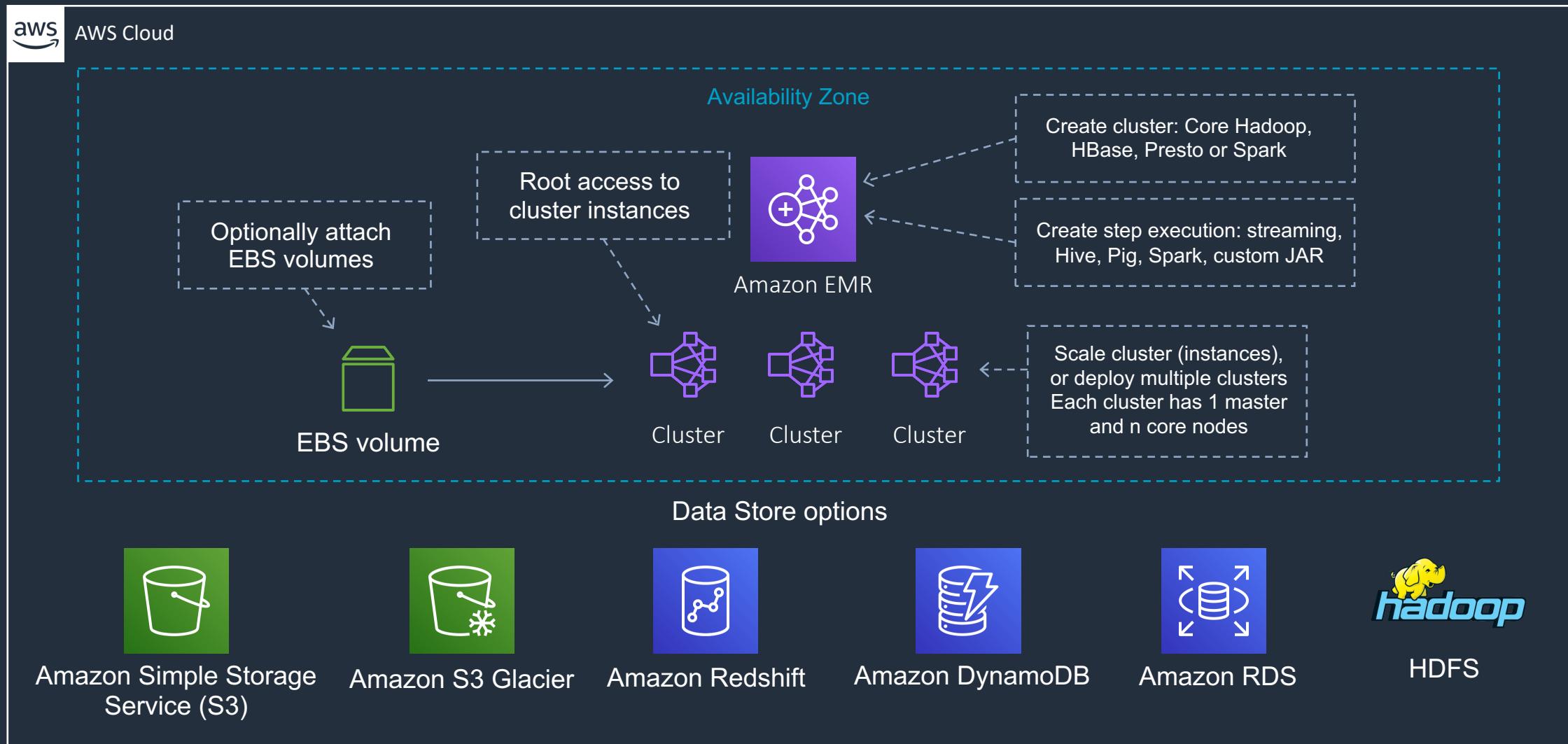
Splunk: delivered to Splunk and optionally to S3



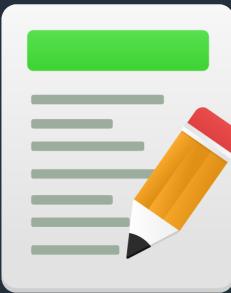
Section 12: AWS Lambda and Kinesis Stream



Section 12: Amazon EMR



Section 12: Exam Cram



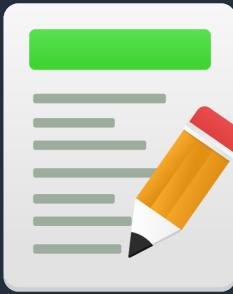
Amazon Kinesis

- Amazon Kinesis makes it easy to collect, process, and analyze real-time, streaming data so you can get timely insights and react quickly to new information.
- Collection of services for processing streams of various data.
- Data is processed in “shards” – with each shard able to ingest 1000 records per second.
- There is a default limit of 500 shards, but you can request an increase to unlimited shards.
- A record consists of a partition key, sequence number, and data blob (up to 1 MB).
- Transient data store – default retention of 24 hours but can be configured for up to 7 days.

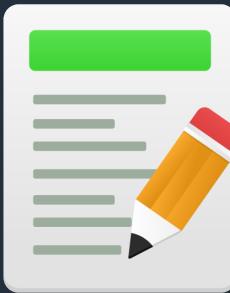
Section 12: Exam Cram

Amazon Kinesis Data Streams

- Kinesis Data Streams enables you to build custom applications that process or analyze streaming data for specialized needs.
- Kinesis Data Streams enables real-time processing of streaming big data.
- Kinesis Data Streams is useful for rapidly moving data off data producers and then continuously processing the data.
- Kinesis Data Streams stores data for later processing by applications (key difference with Firehose which delivers data directly to AWS services).



Section 12: Exam Cram



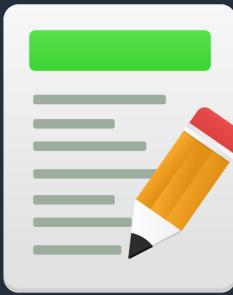
Amazon Kinesis Data Streams

- A producer creates the data that makes up the stream.
- Producers can be used through the following:
 - Kinesis Streams API.
 - Kinesis Producer Library (KPL).
 - Kinesis Agent.
- Consumers are the EC2 instances that analyze the data received from a stream.
- Consumers are known as Amazon Kinesis Streams Applications.
- A shard is the base throughput unit of an Amazon Kinesis data stream.
- One shard provides a capacity of 1MB/sec data input and 2MB/sec data output.
- Each shard can support up to 1000 PUT records per second.
- Scale by adding shards.

Section 12: Exam Cram

Amazon Kinesis Data Firehose

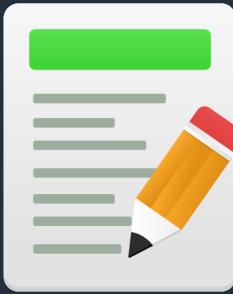
- Kinesis Data Firehose is the easiest way to load streaming data into data stores and analytics tools.
- Captures, transforms, and loads streaming data.
- Enables near real-time analytics with existing business intelligence tools and dashboards.
- Kinesis Data Streams can be used as the source(s) to Kinesis Data Firehose.
- You can configure Kinesis Data Firehose to transform your data before delivering it.
- Firehose can batch, compress, and encrypt data before loading it.
- Firehose synchronously replicates data across three AZs as it is transported to destinations.
- Each delivery stream stores data records for up to 24 hours.



Section 12: Exam Cram

Amazon Kinesis Data Firehose

- Firehose Destinations include:
 - Amazon S3.
 - Amazon Redshift.
 - Amazon Elasticsearch Service.
 - Splunk.
 - Producers provide data streams.
- No shards, totally automated.



Section 12: Exam Cram



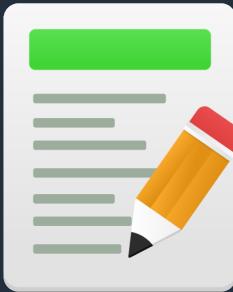
Amazon Kinesis Data Analytics

- Amazon Kinesis Data Analytics processes and analyzes real-time, streaming data.
- Can use standard SQL queries to process Kinesis data streams.
- Provides real-time analysis.
- Use cases:
 - Generate time-series analytics.
 - Feed real-time dashboards.
 - Create real-time alerts and notifications.
- Quickly author and run powerful SQL code against streaming sources.
- Can ingest data from Kinesis Streams and Kinesis Firehose.
- Output to S3, RedShift, Elasticsearch and Kinesis Data Streams.
- Sits over Kinesis Data Streams and Kinesis Data Firehose.

Section 12: Exam Cram

Amazon EMR

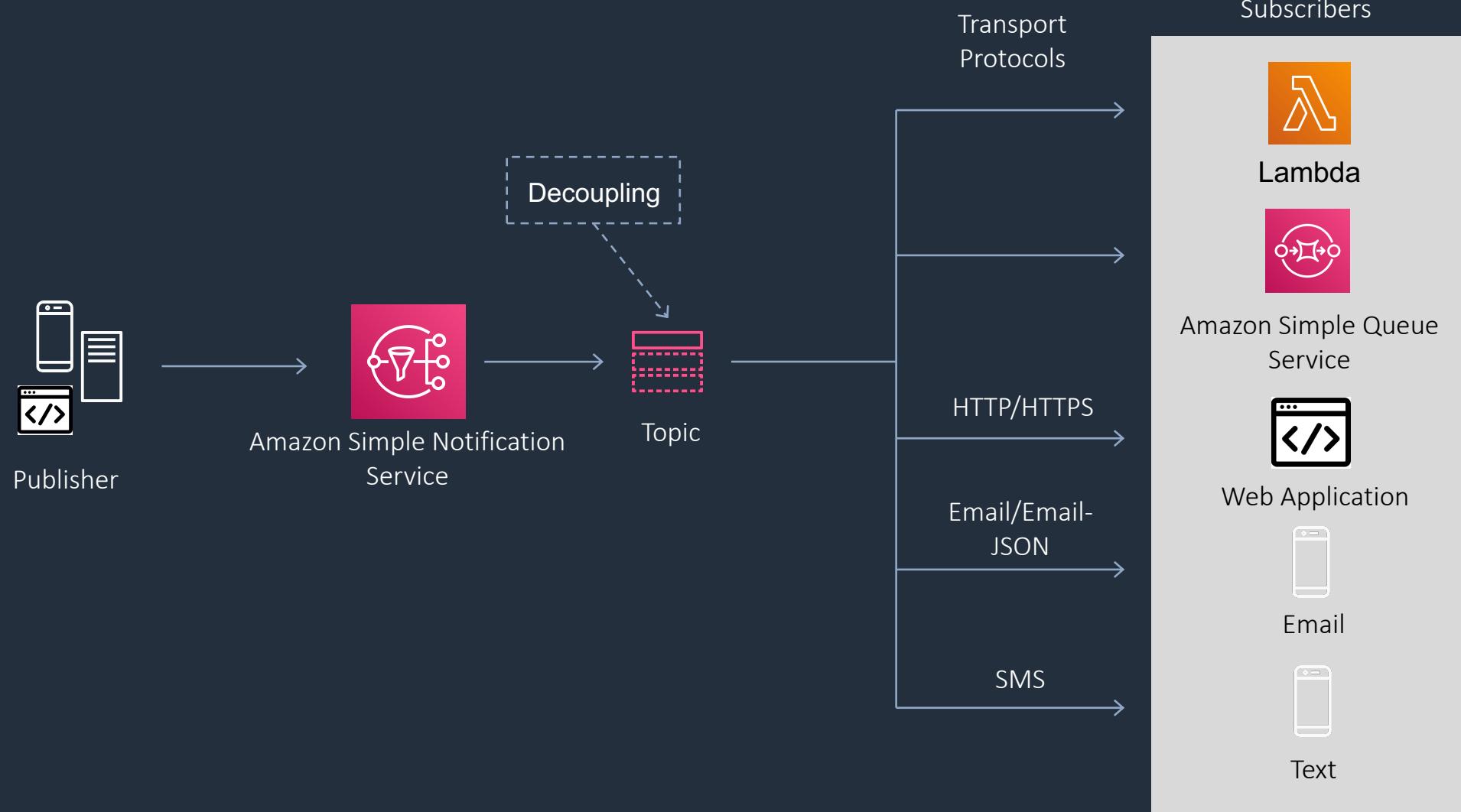
- Amazon EMR is a web service that enables businesses, researchers, data analysts, and developers to easily and cost-effectively process vast amounts of data.
- EMR utilizes a hosted Hadoop framework running on Amazon EC2 and Amazon S3.
- Managed Hadoop framework for processing huge amounts of data.
- Also support Apache Spark, HBase, Presto and Flink.
- Most commonly used for log analysis, financial analysis, or extract, translate and loading (ETL) activities.
- A Step is a programmatic task for performing some process on the data (e.g. count words).
- A cluster is a collection of EC2 instances provisioned by EMR to run your Steps.



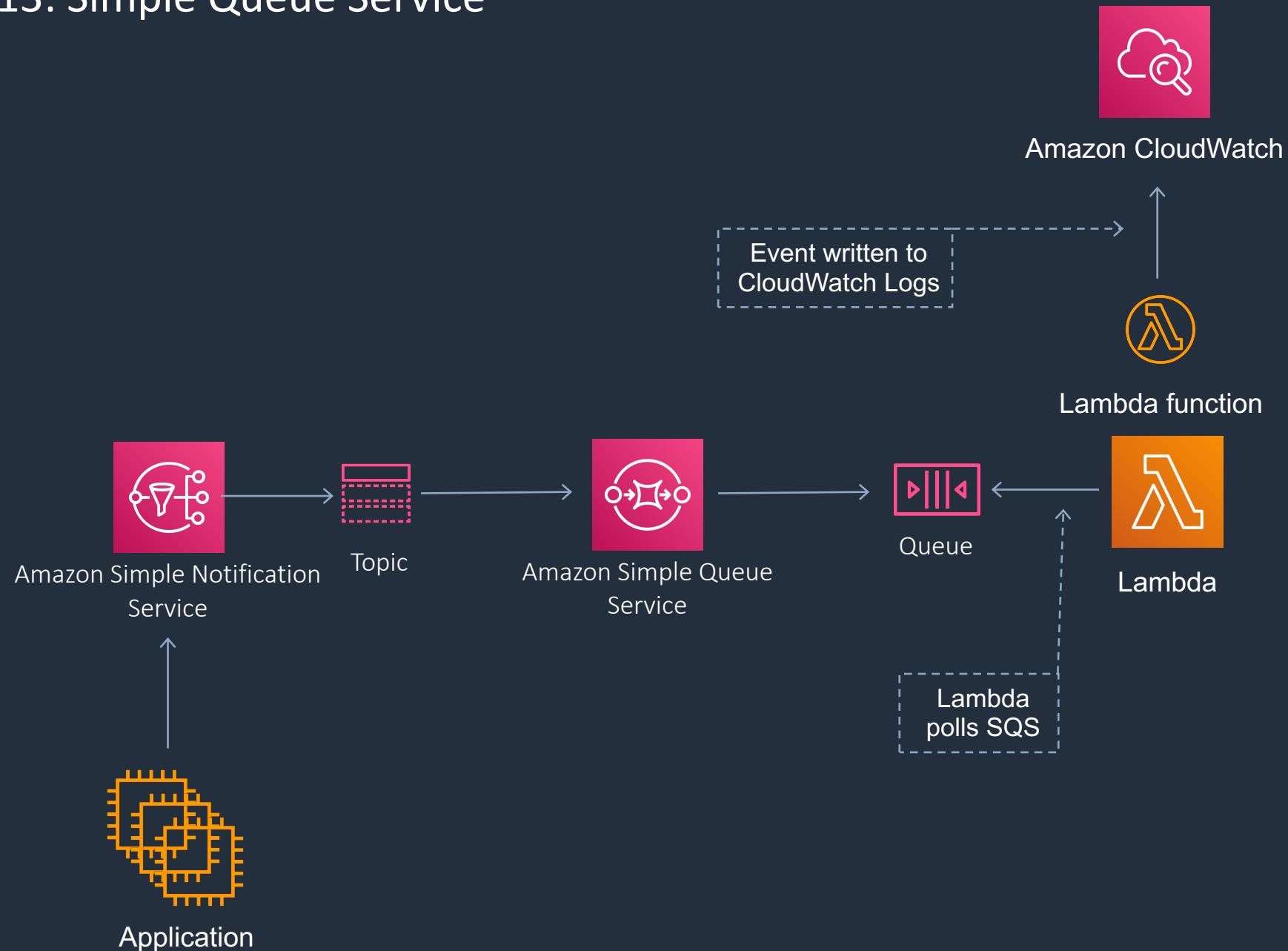
Section 13: Application Integration Services

Service	What it does	Example use cases
Simple Notification Service	Set up, operate, and send notifications from the cloud	Send email notification when CloudWatch alarm is triggered
Step Functions	Out-of-the-box coordination of AWS service components with visual workflow	Order processing workflow
Simple Workflow Service	Need to support external processes or specialized execution logic	Human-enabled workflows like an order fulfilment system or for procedural requests AWS recommends that for new applications customers consider Step Functions instead of SWF
Simple Queue Service	Messaging queue; store and forward patterns	Building distributed / decoupled applications
Amazon MQ	Managed message broker based on Apache MQ	Easy low-hassle path to migrate from existing message brokers to AWS

Section 13: Simple Notification Service



Section 13: Simple Queue Service



Section 13: Simple Queue Service Queue Types

Standard Queues

Unlimited Throughput: Standard queues support a nearly unlimited number of transactions per second (TPS) per API action.

At-Least-Once Delivery: A message is delivered at least once, but occasionally more than one copy of a message is delivered.

Best-Effort Ordering: Occasionally, messages might be delivered in an order different from which they were sent.



You can use standard message queues in many scenarios, as long as your application can process messages that arrive more than once and out of order, for example:

- Decouple live user requests from intensive background work: Let users upload media while resizing or encoding it.
- Allocate tasks to multiple worker nodes: Process a high number of credit card validation requests.
- Batch messages for future processing: Schedule multiple entries to be added to a database.

FIFO Queues

High Throughput: By default, FIFO queues support up to 300 messages per second (300 send, receive, or delete operations per second). When you batch 10 messages per operation (maximum), FIFO queues can support up to 3,000 messages per second. To request a limit increase, [file a support request](#).

Exactly-Once Processing: A message is delivered once and remains available until a consumer processes and deletes it. Duplicates aren't introduced into the queue.

First-In-First-Out Delivery: The order in which messages are sent and received is strictly preserved (i.e. First-In-First-Out).



FIFO queues are designed to enhance messaging between applications when the order of operations and events is critical, or where duplicates can't be tolerated, for example:

- Ensure that user-entered commands are executed in the right order.
- Display the correct product price by sending price modifications in the right order.
- Prevent a student from enrolling in a course before registering for an account.

Section 13: Exam Cram

Amazon SNS

- Amazon Simple Notification Service (Amazon SNS) is a web service that makes it easy to set up, operate, and send notifications from the cloud.
- Amazon SNS is used for building and integrating loosely-coupled, distributed applications.
- Provides instantaneous, push-based delivery (no polling).
- Flexible message delivery is provided over multiple transport protocols.
- SNS supports a wide variety of needs including event notification, monitoring applications, workflow systems, time-sensitive information updates, mobile applications, and any other application that generates or consumes notifications.

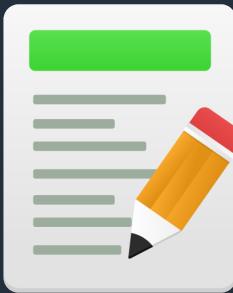


Section 13: Exam Cram

Amazon SNS Subscribers

- SNS Subscribers:

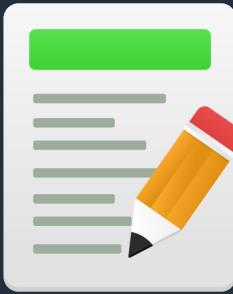
- HTTP.
- HTTPS.
- Email.
- Email-JSON.
- SQS.
- Application.
- Lambda.



Section 13: Exam Cram

Amazon SNS Transport Protocols

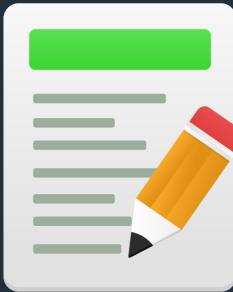
- SNS supports notifications over multiple transport protocols:
 - HTTP/HTTPS – subscribers specify a URL as part of the subscription registration.
 - Email/Email-JSON – messages are sent to registered addresses as email (text-based or JSON-object).
 - SQS – users can specify an SQS standard queue as the endpoint.
 - SMS – messages are sent to registered phone numbers as SMS text messages.



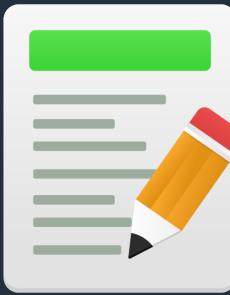
Section 13: Exam Cram

Amazon Step Functions

- AWS Step Functions makes it easy to coordinate the components of distributed applications as a series of steps in a visual workflow.
- You can quickly build and run state machines to execute the steps of your application in a reliable and scalable fashion.
- Managed workflow and orchestration platform.
- Scalable and highly available.
- Define your app as a state machine.
- Create tasks, sequential steps, parallel steps, branching paths or timers.



Section 13: Exam Cram



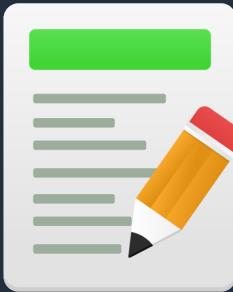
Amazon SWF

- Amazon Simple Workflow Service (SWF) is a web service that makes it easy to coordinate work across distributed application components.
- Create distributed asynchronous systems as workflows.
- Supports both sequential and parallel processing.
- Tracks the state of your workflow which you interact and update via API.
- Best suited for human-enabled workflows like an order fulfilment system or for procedural requests.
- AWS recommends that for new applications customers consider Step Functions instead of SWF.

Section 13: Exam Cram

Amazon SQS

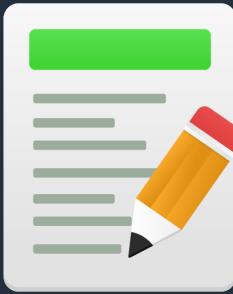
- Amazon Simple Queue Service (Amazon SQS) is a web service that gives you access to message queues that store messages waiting to be processed.
- SQS offers a reliable, highly-scalable, hosted queue for storing messages in transit between computers.
- SQS is used for distributed/decoupled applications.
- SQS can be used with RedShift, DynamoDB, EC2, ECS, RDS, S3 and Lambda.
- SQS uses a message-oriented API.
- SQS uses pull based (polling) not push based.
- Messages are 256KB in size.
- Messages can be kept in the queue from 1 minute to 14 days (default is 4 days).



Section 13: Exam Cram

Amazon SQS

- The visibility timeout is the amount of time a message is invisible in the queue after a reader picks up the message.
- If a job is processed within the visibility timeout the message will be deleted.
- If a job is not processed within the visibility timeout the message will become visible again (could be delivered twice).
- The maximum visibility timeout for an Amazon SQS message is 12 hours.



Section 13: Exam Cram

Amazon SQS vs Kinesis Data Streams

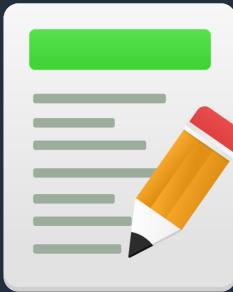


Kinesis Data Streams	Amazon SQS
Routing related records to the same record processor	Messaging semantics such as message-level ack/fail and visibility timeout
Maintaining the order of records	Individual message delay of up to 15 minutes
Connecting multiple consumers to a stream concurrently	Seamless and automatic scalability (Kinesis requires planning and provisioning shards)
Store records for up to 7 days and then consume whilst maintaining order	

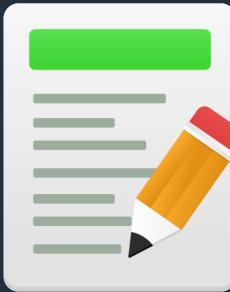
Section 13: Exam Cram

Amazon SQS Polling

- SQS uses short polling and long polling.
- Short polling:
 - Does not wait for messages to appear in the queue.
 - It queries only a subset of the available servers for messages (based on weighted random execution).
 - Short polling is the default.
 - ReceiveMessageWaitTime is set to 0.
 - More requests are used, which implies higher cost.



Section 13: Exam Cram



Amazon SQS Polling

- Long polling:
 - Uses fewer requests and reduces cost.
 - Eliminates false empty responses by querying all servers.
 - SQS waits until a message is available in the queue before sending a response.
 - Requests contain at least one of the available messages up to the maximum number of messages specified in the ReceiveMessage action.
 - Shouldn't be used if your application expects an immediate response to receive message calls.
 - ReceiveMessageWaitTime is set to a non-zero value (up to 20 seconds).
 - Same charge per million requests as short polling.

Section 13: Exam Cram



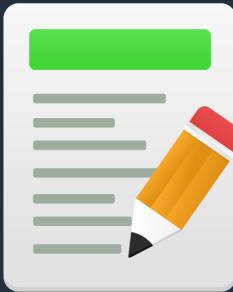
Amazon SQS Queues

- Queues can be either standard or first-in-first-out (FIFO).
- Standard queues provide a loose-FIFO capability that attempts to preserve the order of messages.
- Because standard queues are designed to be massively scalable using a highly distributed architecture, receiving messages in the exact order they are sent is not guaranteed.
- Standard queues provide at-least-once delivery, which means that each message is delivered at least once.
- FIFO (first-in-first-out) queues preserve the exact order in which messages are sent and received.
- If you use a FIFO queue, you don't have to place sequencing information in your message.
- FIFO queues provide exactly-once processing, which means that each message is delivered once and remains available until a consumer processes it and deletes it.

Section 13: Exam Cram

Amazon SQS Limits

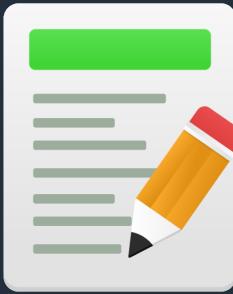
- In-flight messages are messages that have been picked up by a consumer but not yet deleted from the queue.
- Standard queues have a limit of 120,000 in-flight messages per queue.
- FIFO queues have a limit of 20,000 in-flight messages per queue.
- Queue names can be up to 80 characters.
- Messages are retained for 4 days by default up to 14 days.
- FIFO queues support up to 3000 messages per second when batching or 300 per second otherwise.
- The maximum messages size is 256KB.



Section 13: Exam Cram

Amazon SQS Scalability and Durability

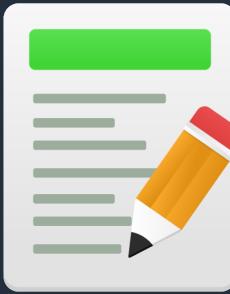
- You can have multiple queues with different priorities.
- Scaling is performed by creating more queues.
- SQS stores all message queues and messages within a single, highly-available AWS region with multiple redundant AZs.



Section 13: Exam Cram

Amazon SQS Security

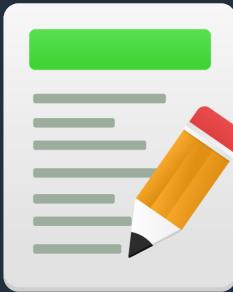
- You can use IAM policies to control who can read/write messages.
- Authentication can be used to secure messages within queues (who can send and receive).
- SQS supports HTTPS and supports TLS versions 1.0, 1.1, 1.2.
- SQS is PCI DSS level 1 compliant and HIPAA eligible.
- Server-side encryption (SSE) lets you transmit sensitive data in encrypted queues (AWS KMS).



Section 13: Exam Cram

Amazon MQ

- Amazon MQ is a managed message broker service for ActiveMQ that makes it easy to set up and operate message brokers in the cloud, so you can migrate your messaging and applications without rewriting code.
- Amazon MQ supports industry-standard APIs and protocols so you can migrate messaging and applications without rewriting code.
- Amazon MQ provides cost-efficient and flexible messaging capacity – you pay for broker instance and storage usage as you go.
- It's a managed implementation of Apache ActiveMQ.
- Fully managed and highly available within a region.



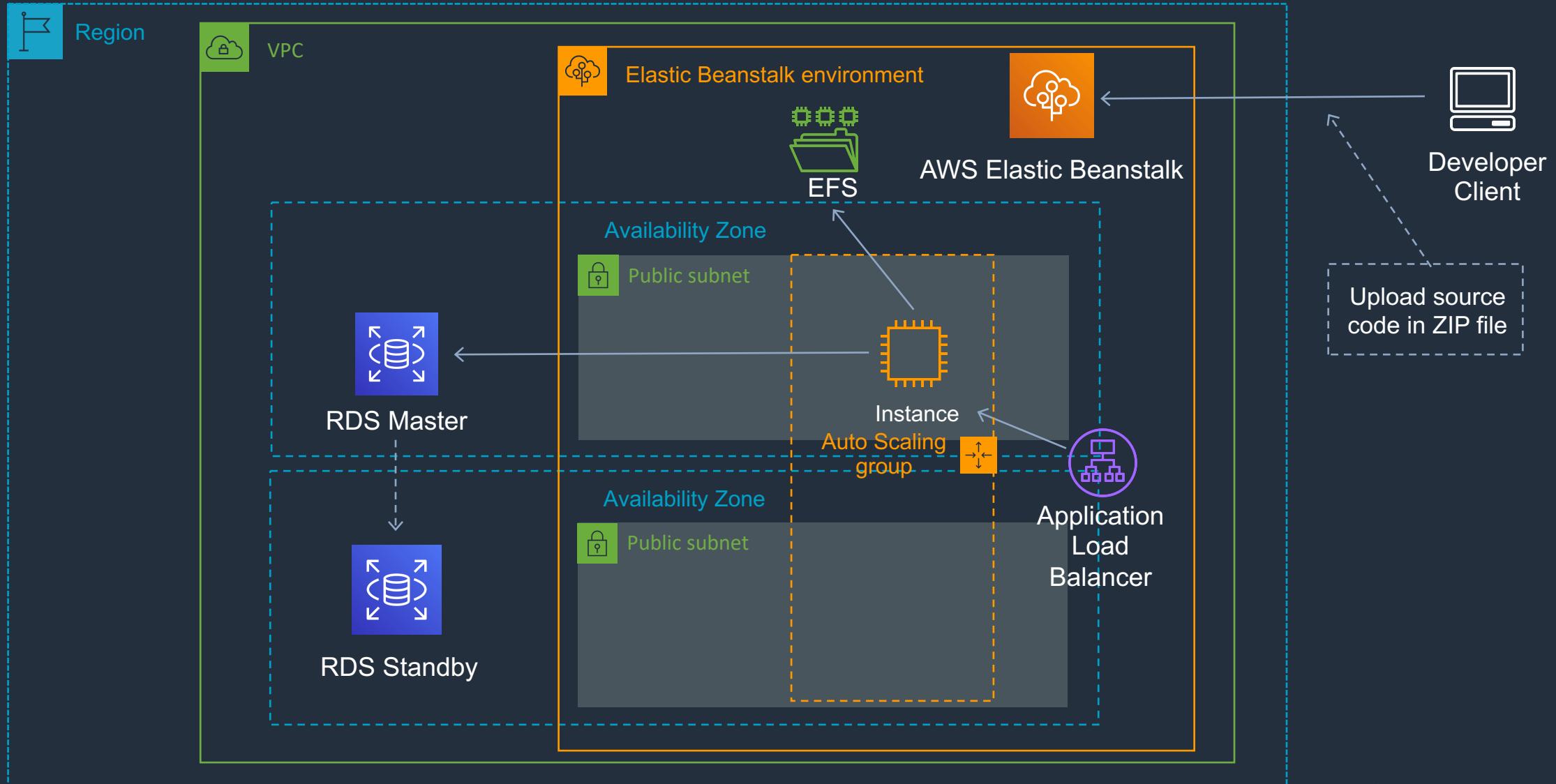
Section 14: Infrastructure as Code and PaaS

CloudFormation	Elastic Beanstalk
“Template-driven provisioning”	“Web apps made easy”
Deploys infrastructure using code	Deploys applications on EC2 (PaaS)
Can be used to deploy almost any AWS service	Deploys web applications based on Java, .NET, PHP, Node.js, Python, Ruby, Go, and Docker
Uses JSON or YAML template files	Uses ZIP or WAR files (or Git)
CloudFormation can deploy Elastic Beanstalk environments	Elastic Beanstalk cannot deploy using CloudFormation
Similar to Terraform	Similar to Google App Engine

Section 14: HA Wordpress using CloudFormation



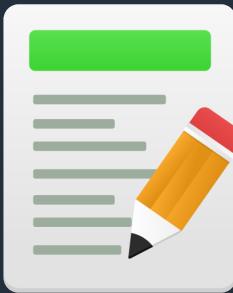
Section 14: HA WordPress with Elastic Beanstalk and RDS



Section 14: Exam Cram

AWS CloudFormation

- AWS CloudFormation provides a common language for you to describe and provision all the infrastructure resources in your cloud environment.
- CloudFormation can be used to provision a broad range of AWS resources.
- Think of CloudFormation as deploying infrastructure as code.
- Elastic Beanstalk is more focussed on deploying applications on EC2 (PaaS).
- CloudFormation can deploy Elastic Beanstalk-hosted applications however the reverse is not possible.
- Logical IDs are used to reference resources within the template.
- Physical IDs identify resources outside of AWS CloudFormation templates, but only after the resources have been created.



Section 14: Exam Cram

AWS CloudFormation

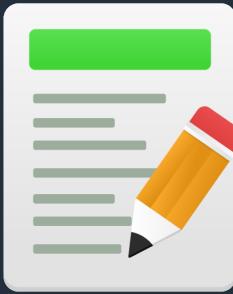


Element	Description
Templates	The JSON or YAML text file that contains the instructions for building out the AWS environment
Stacks	The entire environment described by the template and created, updated, and deleted as a single unit
Change Sets	A summary of proposed changes to your stack that will allow you to see how those changes might impact your existing resources before implementing them

Section 14: Exam Cram

AWS CloudFormation

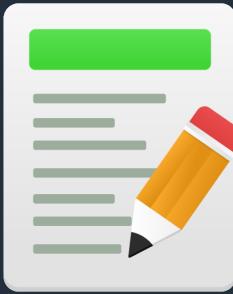
- Updating stacks:
 - AWS CloudFormation provides two methods for updating stacks: direct update or creating and executing change sets.
 - When you directly update a stack, you submit changes and AWS CloudFormation immediately deploys them.
 - Use direct updates when you want to quickly deploy your updates.
 - With change sets, you can preview the changes AWS CloudFormation will make to your stack, and then decide whether to apply those changes.



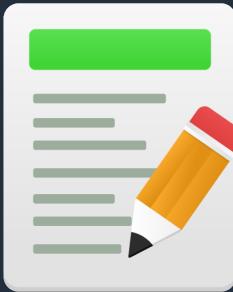
Section 14: Exam Cram

AWS CloudFormation

- StackSets.
 - AWS CloudFormation StackSets extends the functionality of stacks by enabling you to create, update, or delete stacks across multiple accounts and regions with a single operation.
 - Using an administrator account, you define and manage an AWS CloudFormation template, and use the template as the basis for provisioning stacks into selected target accounts across specified regions.



Section 14: Exam Cram



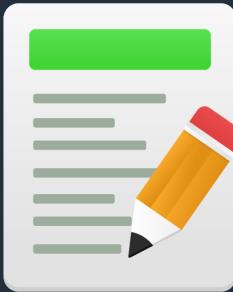
AWS Elastic Beanstalk

- AWS Elastic Beanstalk can be used to quickly deploy and manage applications in the AWS Cloud.
- Developers upload applications and Elastic Beanstalk handles the deployment details of capacity provisioning, load balancing, auto-scaling, and application health monitoring.
- AWS Elastic Beanstalk leverages Elastic Load Balancing and Auto Scaling to automatically scale your application in and out based on your application's specific needs.
- In addition, multiple availability zones give you an option to improve application reliability and availability by running in more than one zone.
- Considered a Platform as a Service (PaaS) solution.
- Supports Java, .NET, PHP, Node.js, Python, Ruby, Go, and Docker web applications.

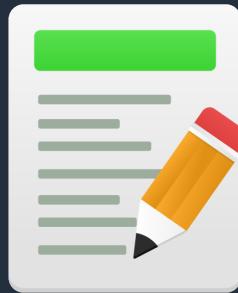
Section 14: Exam Cram

AWS Elastic Beanstalk

- Can provision most database instances.
- Allows full control of the underlying resources.
- Stores your application files and, optionally, server log files in Amazon S3.
- Application data can also be stored on S3.
- Multiple environments are supported to enable versioning.
- Changes from Git repositories are replicated.



Section 14: Exam Cram



CloudFormation	Elastic Beanstalk
“Template-driven provisioning”	“Web apps made easy”
Deploys infrastructure using code	Deploys applications on EC2 (PaaS)
Can be used to deploy almost any AWS service	Deploys web applications based on Java, .NET, PHP, Node.js, Python, Ruby, Go, and Docker
Uses JSON or YAML template files	Uses ZIP or WAR files (or Git)
CloudFormation can deploy Elastic Beanstalk environments	Elastic Beanstalk cannot deploy using CloudFormation
Similar to Terraform	Similar to Google App Engine

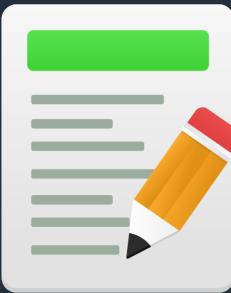
Section 15: Monitoring and Logging Overview

CloudWatch	CloudTrail
Performance monitoring (operations)	Auditing (security)
Log events across AWS services – think operations	Log API activity across AWS services – think activities
Higher-level comprehensive monitoring and eventing	More low-level granular
Log from multiple accounts	Log from multiple accounts
Logs stored indefinitely	Logs stored to S3 or CloudWatch indefinitely
Alarms history for 14 days	No native alarming; can use CloudWatch alarms

Section 15: Exam Cram

AWS CloudWatch

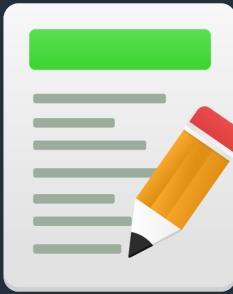
- Amazon CloudWatch is a monitoring service for AWS cloud resources and the applications you run on AWS.
- Used to collect and track metrics, collect and monitor log files, and set alarms.
- Automatically react to changes in your AWS resources.
- With CloudWatch you can monitor resources such as:
 - EC2 instances.
 - DynamoDB tables.
 - RDS DB instances.
 - Custom metrics generated by applications and services.
 - Any log files generated by your applications.



Section 15: Exam Cram

AWS CloudWatch

- Monitor application performance.
- Monitor operational health.
- CloudWatch is accessed via API, command-line interface, AWS SDKs, and the AWS Management Console.
- CloudWatch integrates with IAM.



Section 15: Exam Cram

AWS CloudWatch Logs

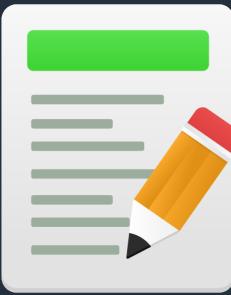


- CloudWatch Logs:
 - Amazon CloudWatch Logs lets you monitor and troubleshoot your systems and applications using your existing system, application and custom log files.
 - You can use Amazon CloudWatch Logs to monitor, store, and access your log files from Amazon Elastic Compute Cloud (Amazon EC2) instances, AWS CloudTrail, Route 53, and other sources.
 - CloudWatch Logs can be used for real time application and system monitoring as well as long term log retention.
 - CloudWatch Logs keeps logs indefinitely by default.
 - CloudTrail logs can be sent to CloudWatch Logs for real-time monitoring.
 - CloudWatch Logs metric filters can evaluate CloudTrail logs for specific terms, phrases or values.

Section 15: Exam Cram

AWS CloudWatch Metrics

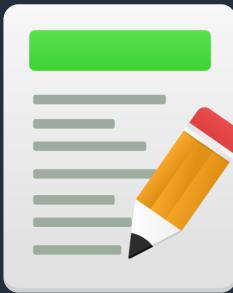
- CloudWatch retains metric data as follows:
 - Data points with a period of less than 60 seconds are available for 3 hours. These data points are high-resolution custom metrics.
 - Data points with a period of 60 seconds (1 minute) are available for 15 days.
 - Data points with a period of 300 seconds (5 minute) are available for 63 days.
 - Data points with a period of 3600 seconds (1 hour) are available for 455 days (15 months).



Section 15: Exam Cram

AWS CloudTrail

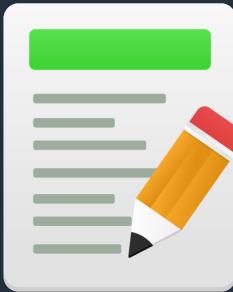
- AWS CloudTrail is a web service that records activity made on your account
- A CloudTrail trail can be created which delivers log files to an Amazon S3 bucket.
- CloudTrail is about logging and saves a history of API calls for your AWS account.
- Provides visibility into user activity by recording actions taken on your account.
- API history enables security analysis, resource change tracking, and compliance auditing.
- Logs API calls made via:
 - AWS Management Console.
 - AWS SDKs.
 - Command line tools.
 - Higher-level AWS services (such as CloudFormation).



Section 15: Exam Cram

AWS CloudTrail

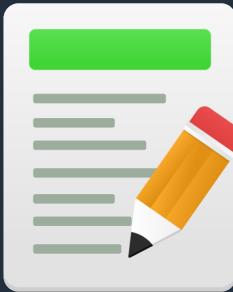
- CloudTrail records account activity and service events from most AWS services and logs the following records:
 - The identity of the API caller.
 - The time of the API call.
 - The source IP address of the API caller.
 - The request parameters.
 - The response elements returned by the AWS service.



Section 15: Exam Cram

AWS CloudTrail

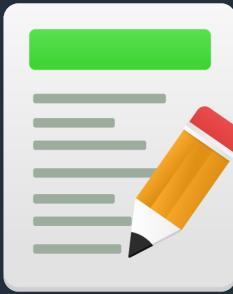
- CloudTrail records account activity and service events from most AWS services and logs the following records:
 - The identity of the API caller.
 - The time of the API call.
 - The source IP address of the API caller.
 - The request parameters.
 - The response elements returned by the AWS service.



Section 15: Exam Cram

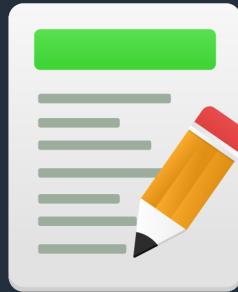
AWS CloudTrail

- Trails can be enabled per region or a trail can be applied to all regions.
- CloudTrail log files are encrypted using S3 Server-Side Encryption (SSE).
- You can also enable encryption using SSE KMS for additional security.



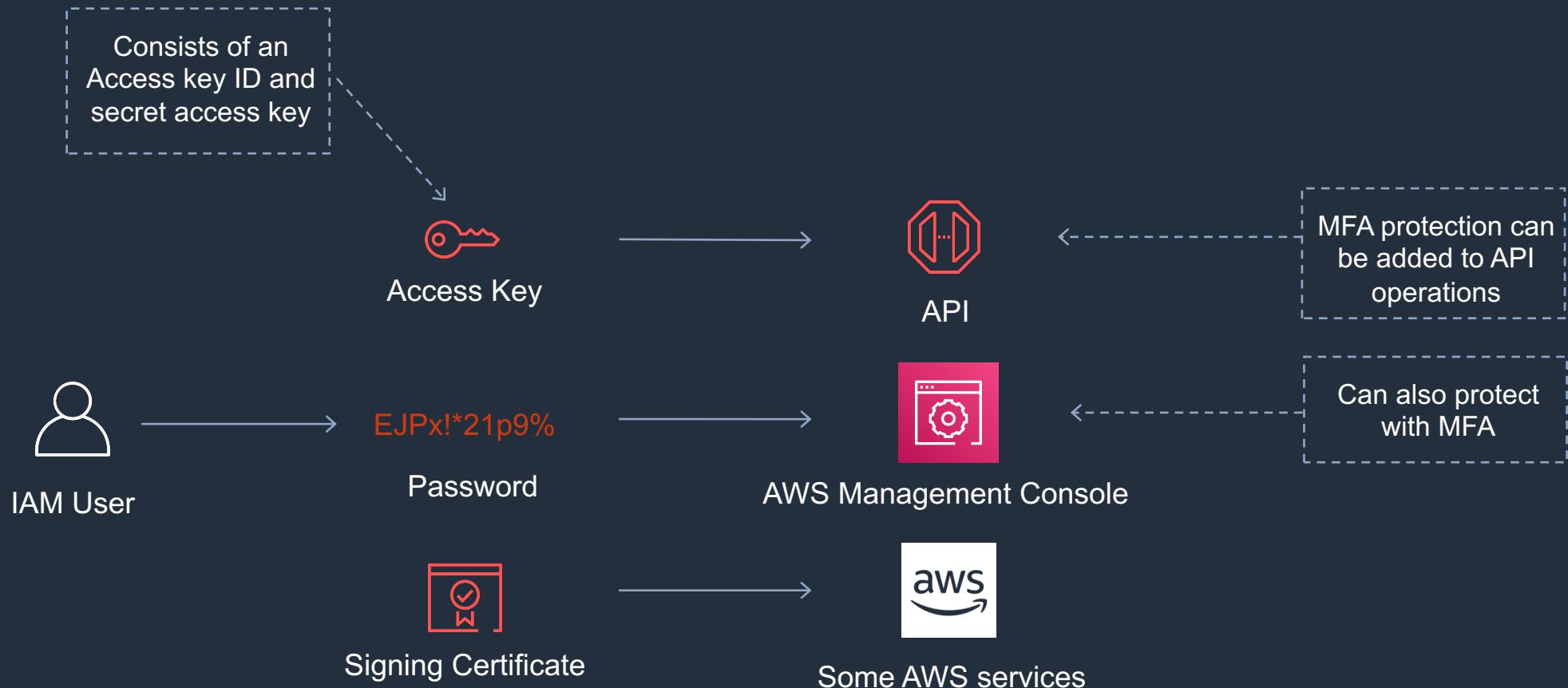
Section 15: Exam Cram

AWS CloudWatch vs CloudTrail

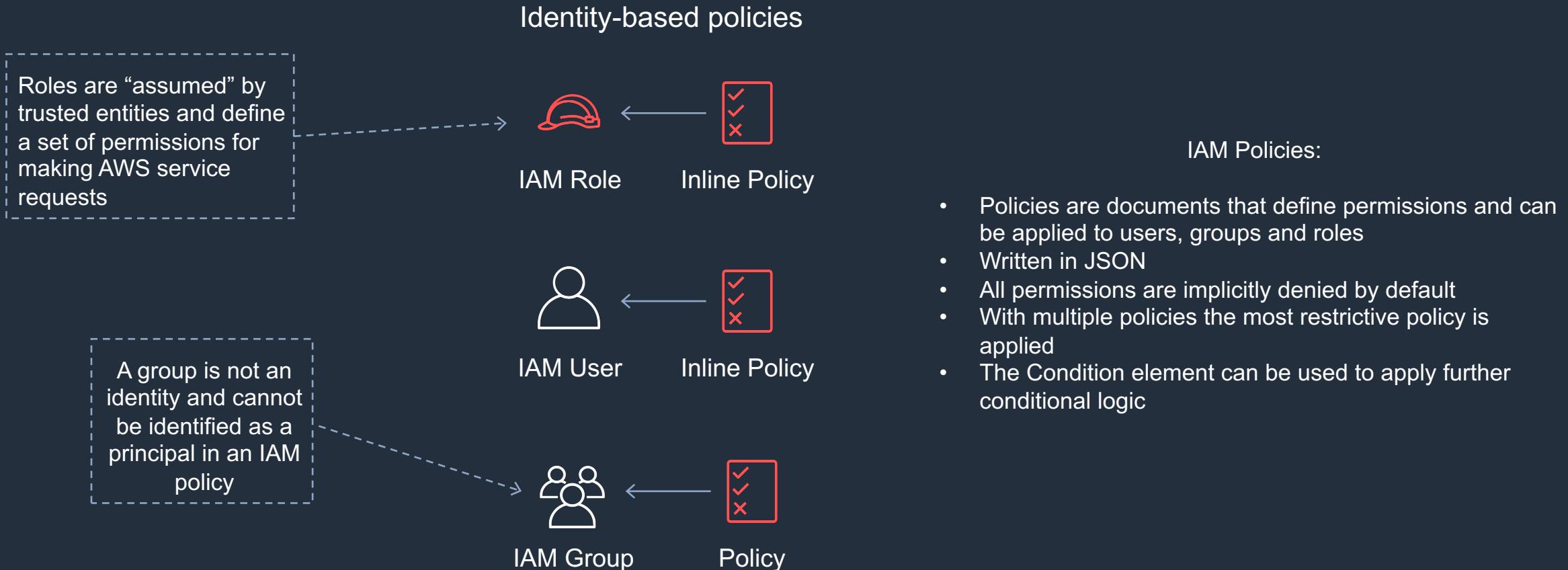


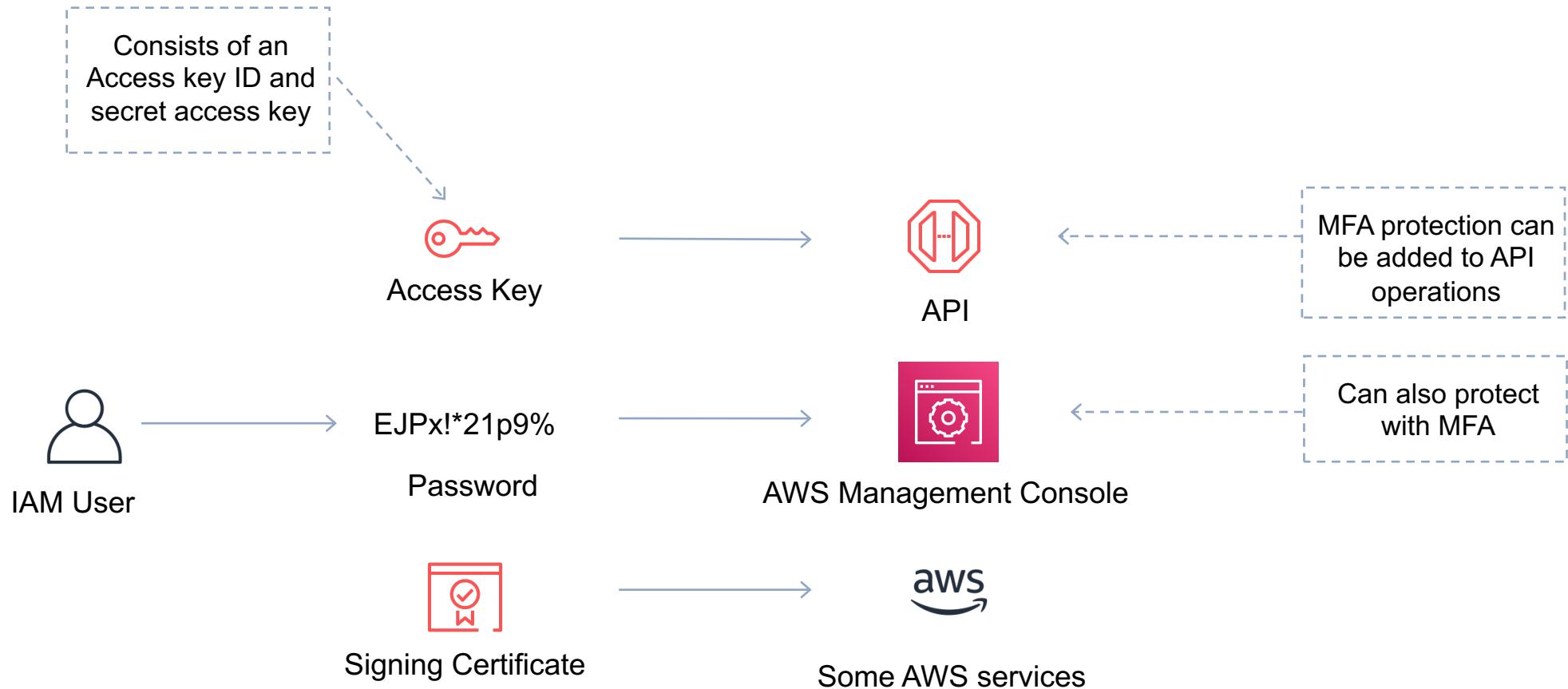
CloudWatch	CloudTrail
Performance monitoring (operations)	Auditing (security)
Log events across AWS services – think operations	Log API activity across AWS services – think activities
Higher-level comprehensive monitoring and eventing	More low-level granular
Log from multiple accounts	Log from multiple accounts
Logs stored indefinitely	Logs stored to S3 or CloudWatch indefinitely
Alarms history for 14 days	No native alarming; can use CloudWatch alarms

Section 16: IAM Authentication Methods



Section 16: IAM Policies – Roles, Users and Groups

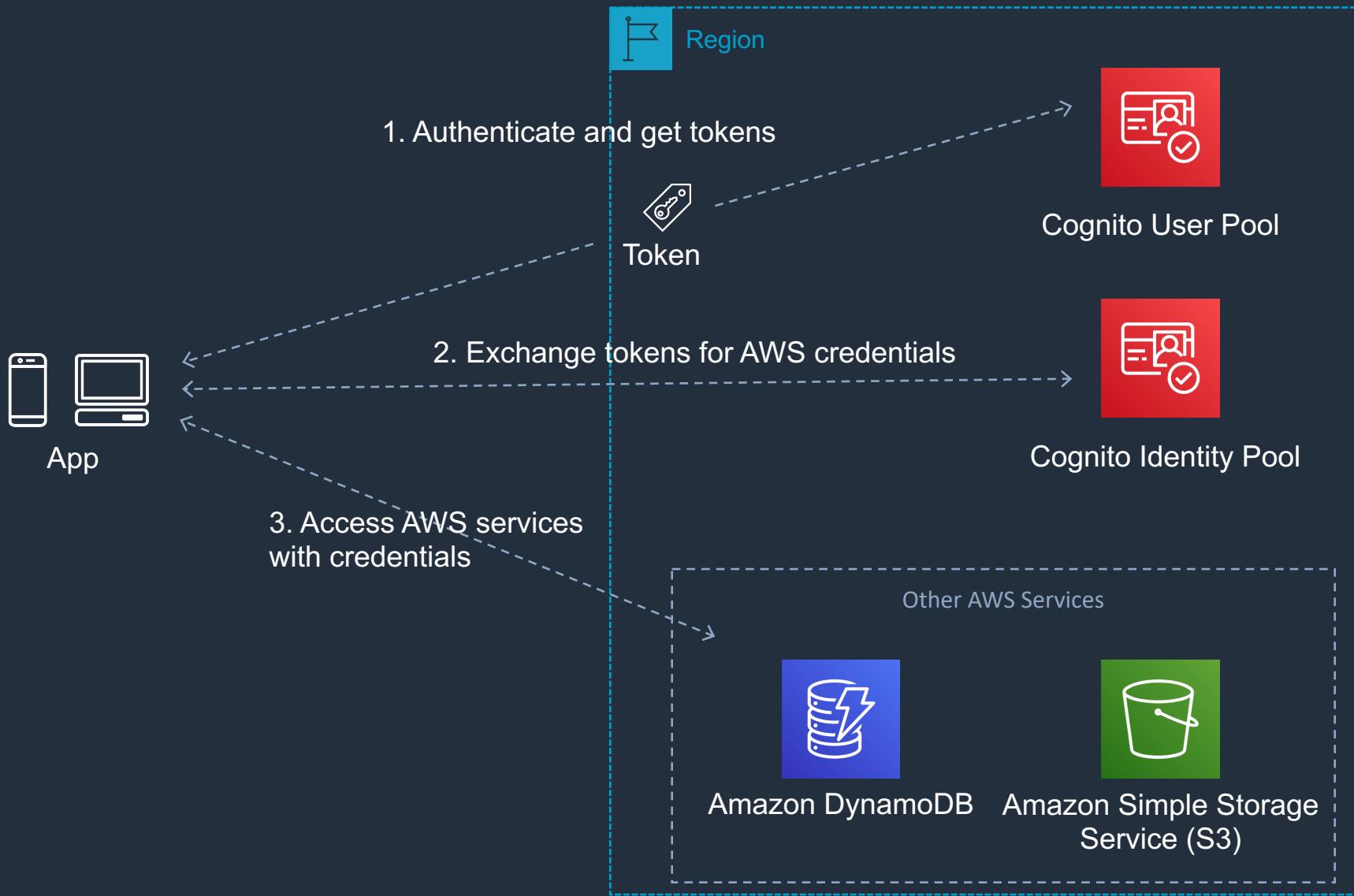




Section 16: IAM Best Practices

- Lock away the AWS root user access keys
- Create individual IAM users
- Use AWS defined policies to assign permissions whenever possible
- Use groups to assign permissions to IAM users
- Grant least privilege
- Use access levels to review IAM permissions
- Configure a strong password policy for users
- Enable MFA
- Use roles for applications that run on AWS EC2 instances
- Delegate by using roles instead of sharing credentials
- Rotate credentials regularly
- Remove unnecessary credentials
- Use policy conditions for extra security
- Monitor activity in your AWS account

Section 16: Amazon Cognito



Section 16: KMS Customer Master Keys (CMKs)

Type of CMK	Can view	Can manage	Used only for my account
Customer managed CMK	Yes	Yes	Yes
AWS managed CMK	Yes	No	Yes
AWS owned CMK	No	No	No

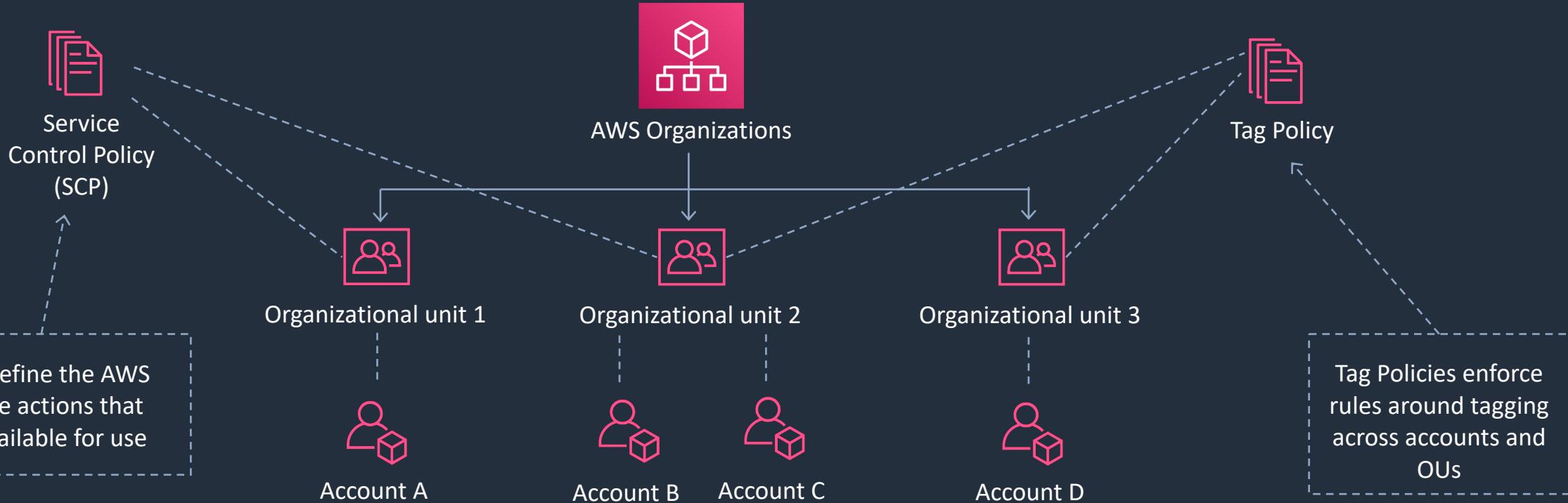
Section 16: Old and New CloudHSM

	"Classic" CloudHSM	Current CloudHSM
Device	safeNET Luna SA	Proprietary AWS
Pricing	Upfront cost required (\$5000)	No upfront cost, pay per hour
High Availability	Have to buy a second device	Clustered
FIPS 140-2	Level 2	Level 3

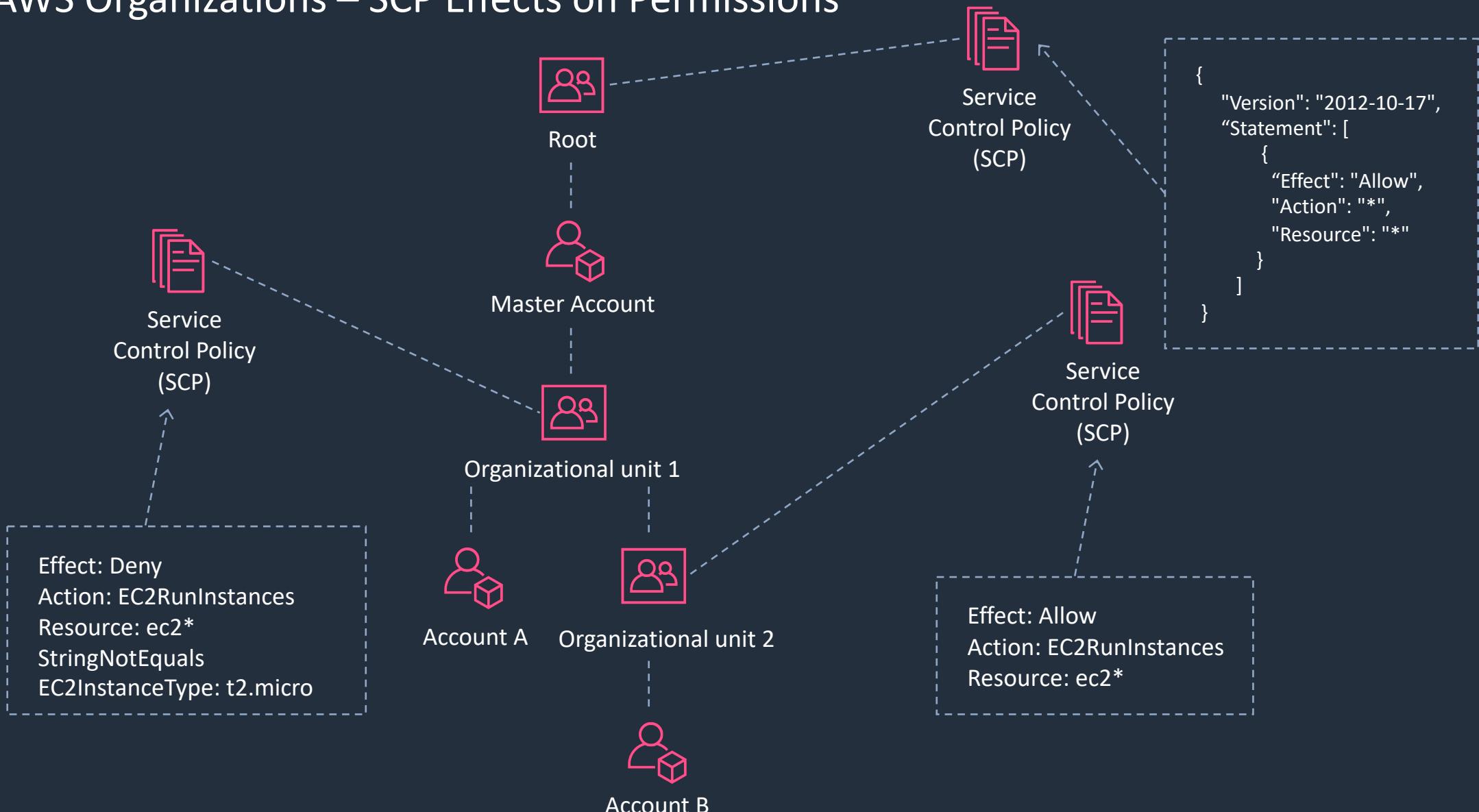
Section 16: AWS Organizations

- AWS Organizations helps you centrally govern your environment as you grow and scale your workloads on AWS.
- Organizations helps you to centrally manage billing; control access, compliance, and security; and share resources across your AWS accounts.
- Using AWS Organizations, you can automate account creation, create groups of accounts to reflect your business needs, and apply policies for these groups for governance.
- You can also simplify billing by setting up a single payment method for all of your AWS accounts.
- Through integrations with other AWS services, you can use Organizations to define central configurations and resource sharing across accounts in your organization.
- Available in two feature sets:
 - Consolidated billing.
 - All features.

Section 16: AWS Organizations – SCPs and Tag Policies



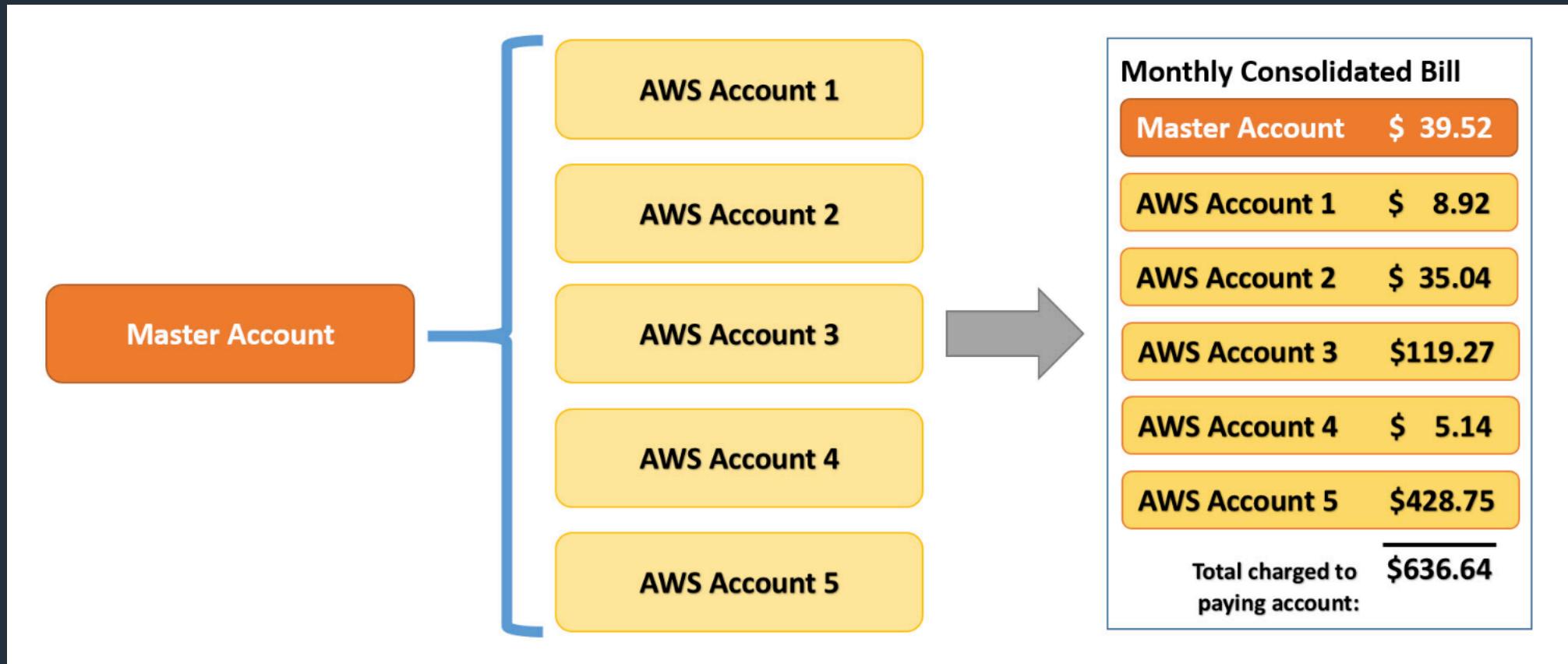
Section 16: AWS Organizations – SCP Effects on Permissions



Section 16: AWS Organizations – Restrict EC2 Instance Types



Section 16: AWS Organizations – Consolidated Billing



Section 16: Exam Cram

AWS IAM

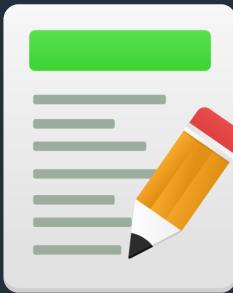
- IAM is used to securely control individual and group access to AWS resources.
- IAM can be used to manage:
 - Users.
 - Groups.
 - Access policies.
 - Roles.
 - User credentials.
 - User password policies.
 - Multi-factor authentication (MFA).
 - API keys for programmatic access (CLI).



Section 16: Exam Cram

AWS IAM

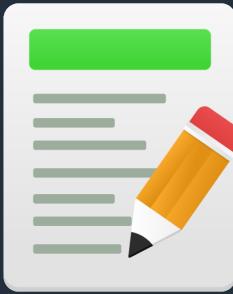
- By default new users are created with NO access to any AWS services – they can only login to the AWS console.
- Permission must be explicitly granted to allow a user to access an AWS service.
- IAM is not used for application-level authentication.
- Identity Federation (including AD, Facebook etc.) can be configured allowing secure access to resources in an AWS account without creating an IAM user account.
- Multi-factor authentication (MFA) can be enabled/enforced for the AWS account and for individual users under the account.
- It is a best practice to use MFA for all users and to use U2F or hardware MFA devices for all privileged users.



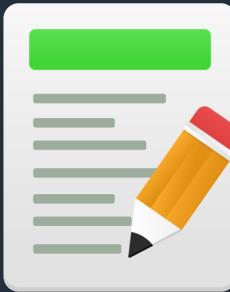
Section 16: Exam Cram

AWS IAM

- IAM is universal (global) and does not apply to regions.
- The "root account" is the account created when you setup the AWS account. It has complete Admin access and is the only account that has this access by default.
- It is a best practice to not use the root account for anything other than billing.



Section 16: Exam Cram



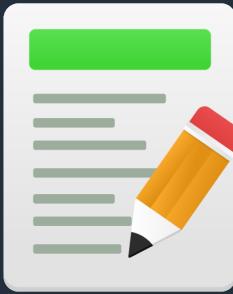
AWS IAM Authentication Methods

- Console password:
 - A password that the user can enter to sign into interactive sessions such as the AWS Management Console.
- Access Keys:
 - A combination of an access key ID and a secret access key.
 - These can be used to make programmatic calls to AWS when using the API in program code or at a command prompt when using the AWS CLI or the AWS PowerShell tools.
- Server certificates:
 - SSL/TLS certificates that you can use to authenticate with some AWS services.
 - AWS recommends that you use the AWS Certificate Manager (ACM) to provision, manage and deploy your server certificates.

Section 16: Exam Cram

AWS IAM Users

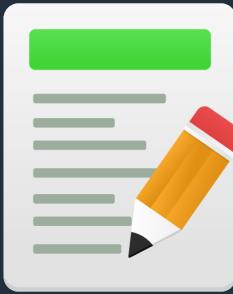
- An IAM user is an entity that represents a person or service.
- Can be assigned:
 - An access key ID and secret access key for programmatic access to the AWS API, CLI, SDK, and other development tools.
 - A password for access to the management console.
- IAM users can be created to represent applications and these are known as "service accounts".



Section 16: Exam Cram

AWS IAM Groups

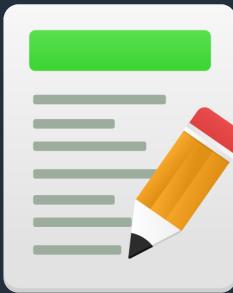
- Groups are collections of users and have policies attached to them.
- A group is not an identity and cannot be identified as a principal in an IAM policy.
- Use groups to assign permissions to users.
- Use the principle of least privilege when assigning permissions.
- You cannot nest groups (groups within groups).



Section 16: Exam Cram

AWS IAM Roles

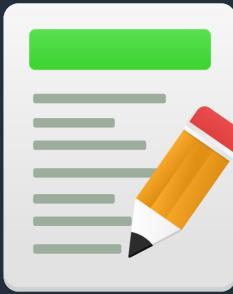
- Roles are created and then "assumed" by trusted entities and define a set of permissions for making AWS service requests.
- With IAM Roles you can delegate permissions to resources for users and services without using permanent credentials (e.g. user name and password).
- IAM users or AWS services can assume a role to obtain temporary security credentials that can be used to make AWS API calls.
- You can delegate using roles.
- There are no credentials associated with a role (password or access keys).



Section 16: Exam Cram

AWS IAM Policies

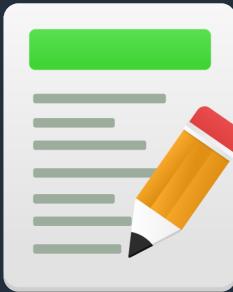
- Policies are documents that define permissions and can be applied to users, groups and roles.
- Policy documents are written in JSON (key value pair that consists of an attribute and a value).
- All permissions are implicitly denied by default.
- The most restrictive policy is applied.



Section 16: Exam Cram

AWS IAM Best Practices

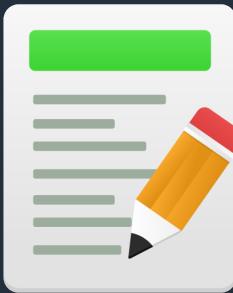
- Lock Away Your AWS Account Root User Access Keys.
- Create Individual IAM Users.
- Use Groups to Assign Permissions to IAM Users.
- Grant Least Privilege.
- Get Started Using Permissions with AWS Managed Policies.
- Use Custom Managed Policies Instead of Inline Policies.
- Use Access Levels to Review IAM Permissions.
- Configure a Strong Password Policy for Your Users.
- Enable MFA.
- Use Roles for Applications That Run on Amazon EC2 Instances.
- Use Roles to Delegate Permissions.
- Do Not Share Access Keys.
- Rotate Credentials Regularly.
- Remove Unnecessary Credentials.
- Use Policy Conditions for Extra Security.
- Monitor Activity in Your AWS Account.



Section 16: Exam Cram

AWS Cognito

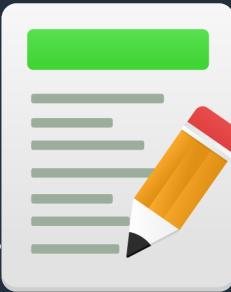
- Amazon Cognito lets you add user sign-up, sign-in, and access control to web and mobile apps.
- Amazon Cognito provides authentication, authorization, and user management for web and mobile apps.
- Your users can sign in directly with a username and password, or through a third party such as Facebook, Amazon, or Google.
- The two main components of AWS Cognito are user pools and identity pools:
 - User pools are user directories that provide sign-up and sign-in options for your app users.
 - Identity pools enable you to grant your users access to other AWS services.
- You can use identity pools and user pools separately or together.
- AWS Cognito works with external identity providers that support SAML or OpenID Connect, social identity providers (such as Facebook, Twitter, Amazon).
- Cognito Identity provides temporary security credentials to access your app's backend resources in AWS or any service behind Amazon API Gateway.



Section 16: Exam Cram

AWS KMS

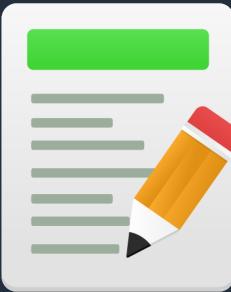
- AWS Key Management Store (KMS) is a managed service that enables you to easily encrypt your data.
- AWS KMS provides a highly available key storage, management, and auditing solution for you to encrypt data within your own applications and control the encryption of stored data across AWS services.
- AWS KMS allows you to centrally manage and securely store your keys. These are known as customer master keys or CMKs.
- You can generate CMKs in KMS, in an AWS CloudHSM cluster, or import them from your own key management infrastructure.
- These master keys are protected by hardware security modules (HSMs) and are only ever used within those modules.
- You can submit data directly to KMS to be encrypted or decrypted using these master keys.
- You set usage policies on these keys that determine which users can use them to encrypt and decrypt and data under which conditions.



Section 16: Exam Cram

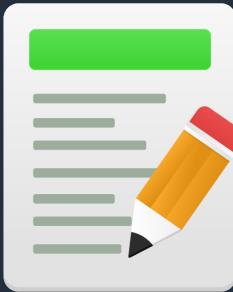
AWS CloudHSM

- The AWS CloudHSM service helps you meet corporate, contractual and regulatory compliance requirements for data security by using dedicated Hardware Security Module (HSM) instances within the AWS cloud.
- A Hardware Security Module (HSM) provides secure key storage and cryptographic operations within a tamper-resistant hardware device.
- HSMs are designed to securely store cryptographic key material and use the key material without exposing it outside the cryptographic boundary of the hardware.
- You can use the CloudHSM service to support a variety of use cases and applications, such as database encryption, Digital Rights Management (DRM), Public Key Infrastructure (PKI), authentication and authorization, document signing, and transaction processing.



Section 16: Exam Cram

AWS CloudHSM vs KMS

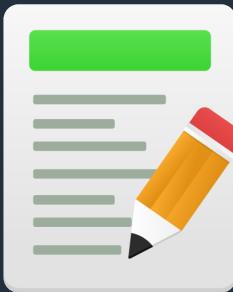


	CloudHSM	AWS KMS
Tenancy	Single-tenant HSM	Multi-tenant AWS service
Availability	Customer-managed durability and available	Highly available and durable key storage and management
Root of Trust	Customer managed root of trust	AWS managed root of trust
FIPS 140-2	Level 3	Level 2 / Level 3 in some areas
3rd Party Support	Broad 3 rd Party Support	Broad AWS service support

Section 16: Exam Cram

AWS WAF & Shield

- AWS WAF and AWS Shield help protect your AWS resources from web exploits and DDoS attacks.
- AWS WAF is a web application firewall service that helps protect your web apps from common exploits that could affect app availability, compromise security, or consume excessive resources.
- AWS Shield provides expanded DDoS attack protection for your AWS resources. Get 24/7 support from our DDoS response team and detailed visibility into DDoS events.
- AWS WAF is tightly integrated with Amazon CloudFront and the Application Load Balancer (ALB), services.
- When you use AWS WAF on Amazon CloudFront, rules run in all AWS Edge Locations, located around the world close to end users.



Section 16: Exam Cram

AWS WAF & Shield

- AWS Shield is a managed Distributed Denial of Service (DDoS) protection service that safeguards applications running on AWS.
- AWS Shield provides always-on detection and automatic inline mitigations that minimize application downtime and latency, so there is no need to engage AWS Support to benefit from DDoS protection.
- There are two tiers of AWS Shield - Standard and Advanced:
 - **Standard:** All AWS customers benefit from the automatic protections of AWS Shield Standard, at no additional charge.
 - **Advanced:** Provides higher levels of protection against attacks targeting applications running on Amazon Elastic Compute Cloud (EC2), Elastic Load Balancing (ELB), Amazon CloudFront, AWS Global Accelerator and Amazon Route 53 resources.

