Module 10. The AWS Global Infrastructure

This module covers the following subjects:

- **Regions:** This section describes the concept of regions and what they contain. Although regions are constantly being added, this section gives you a look at the regions that exist and the naming conventions used for them.
- **Availability Zones:** This section provides valuable information about the Availability Zones within regions.
- **Connections:** This section details connection technologies for your AWS global infrastructure interactions.

A critical advantage that Amazon provides through AWS is a high-tech, high-speed global infrastructure of advanced data centres worldwide. This module describes how the AWS global infrastructure is organised and how you can connect to the infrastructure.

FOUNDATION TOPICS

REGIONS

AWS serves over a million active customers in more than 190 countries. Amazon is steadily expanding their global infrastructure to help customers achieve lower latency and higher throughput and ensure their data resides only in the region they specify.

Amazon builds the AWS Cloud infrastructure around regions and Availability Zones (AZs):

- A region is a physical location in the world with multiple AZs. Note that a region, by design, must have at least two or more AZs, never just one.
- At the time of this update, 34 regions worldwide and 108 AZs exist. Note that these numbers are now increasing at a faster rate than ever. See the following link for the latest information:

https://aws.amazon.com/about-aws/global-infrastructure/

Figure 10-1 shows a sample structure of the AWS Global Infrastructure of Regions and Availability Zones.

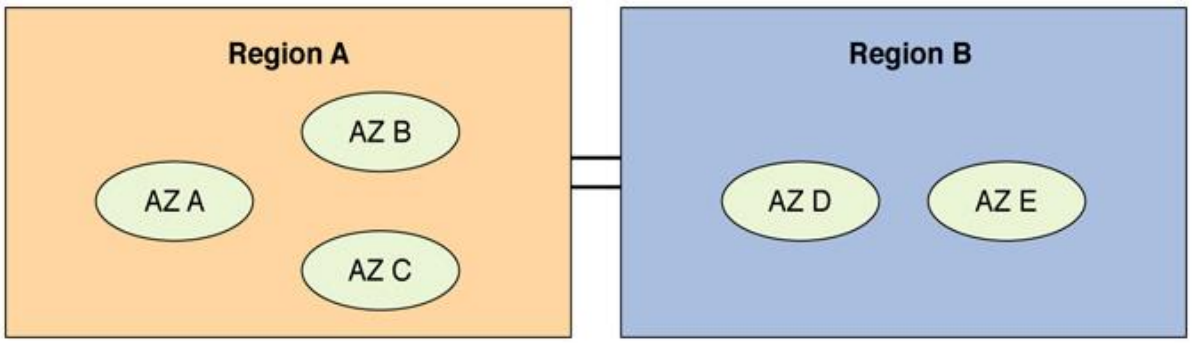

Figure 10-1 The AWS Global Infrastructure

Each Amazon region is designed to be completely isolated from the other Amazon regions. This isolation achieves the highest possible fault tolerance and stability. While each AZ is isolated from a fault tolerance perspective, Amazon connects the AZs within a region through low-latency links.

To give you a sense of regions in AWS, examine the details for some of the North American and European regions:

US East (Ohio) Region
Availability Zones: 3
Launched 2016

US West (Oregon) Region
Availability Zones: 4
Launched 2011
Local Zone: 1
Launched 2019

US West (Northern California) Region
Availability Zones: 3*
Launched 2009

GovCloud (US-West) Region
Availability Zones: 3
Launched 2011

GovCloud (US-East) Region
Availability Zones: 3
Launched 2018

Canada (Central) Region**
Availability Zones: 3
Launched 2016

Learn more at AWS Canada

Europe (Frankfurt) Region
Availability Zones: 3
Launched 2014

Europe (London) Region
Availability Zones: 3
Launched 2016

Europe (Paris) Region
Availability Zones: 3
Launched 2017

Middle East (Bahrain) Region
Availability Zones: 3
Launched 2019

Europe (Ireland) Region
Availability Zones: 3
Launched 2007

Europe (Milan) Region
Availability Zones: 3
Launched 2020

Europe (Stockholm) Region
Availability Zones: 3
Launched 2018

AWS Africa (Cape Town) Region
Availability Zones: 3
Launched 2020

When you reference regions in your AWS code, you use standardised names created by Amazon for each region. Table 10-2 lists some regions and their official AWS names.

Table 10-2 Regions and Their Names in AWS

Region	Name
US West (Oregon) Region	us-west-2
US West (N. California) Region	us-west-1
US East (Ohio) Region	us-east-2
US East (N. Virginia) Region	us-east-1
Asia Pacific (Mumbai) Region	ap-south-1
Asia Pacific (Seoul) Region	ap-northeast-2
Asia Pacific (Singapore) Region	ap-southeast-1
Asia Pacific (Sydney) Region	ap-southeast-2
Asia Pacific (Tokyo) Region	ap-northeast-1
Canada (Central) Region	ca-central-1
China (Beijing) Region	cn-north-1
EU (Frankfurt) Region	eu-central-1
EU (Ireland) Region	eu-west-1
EU (London) Region	eu-west-2
EU (Paris) Region	eu-west-3
South America (São Paulo) Region	sa-east-1
AWS GovCloud (US)	us-gov-west-1

The AWS infrastructure also hosts AWS Edge Locations. AWS CloudFront uses these locations to deliver content at low latency to local clients requesting the data. There are too many Edge Locations to list here, but you can see them on the AWS Region Tables.

AVAILABILITY ZONES

Be sure to remember these facts regarding Availability Zones:

- AZs consist of one or more discrete data centres housed in separate facilities, each with redundant power, networking, and connectivity.
- These AZs enable you to operate production applications and databases that are more highly available, fault-tolerant, and scalable than would be possible from a single data centre.

AWS provides the flexibility to place instances and store data within multiple geographic regions and across multiple Availability Zones within each region. Amazon designs each Availability Zone as an independent failure zone. This independence means that Amazon physically separates Availability Zones within a typical metropolitan region. Amazon chooses lower-risk floodplains in each region.

In addition to discrete uninterruptible power supplies (UPSs) and onsite backup generation facilities, AZs are each fed via different grids from independent utilities to further reduce single points of failure. AZs are all redundantly connected to multiple Tier-1 transit providers. Some AZs have their power substations; as I write this, a majority are creating their own power!

EDGE LOCATIONS

AWS Edge Locations are a critical component of Amazon Web Services (AWS) global infrastructure, designed to deliver services with minimal latency. These locations are essentially data centres strategically placed close to users in various geographic areas, allowing for faster and more efficient content delivery. They play a pivotal role in services such as Amazon CloudFront and AWS Global Accelerator, which utilize Edge Locations to cache content and route traffic more effectively. By leveraging these Edge Locations, AWS ensures that data is transmitted securely and swiftly worldwide, providing users with a seamless and responsive experience.

Local Zones

AWS Local Zones are a revolutionary step in cloud computing, bringing AWS services closer to users and reducing latency for many applications. These zones enable customers to run applications with single-digit millisecond latencies, meeting the needs of real-time gaming, live streaming, and other latency-sensitive tasks. They also support hybrid cloud migrations and comply with local data residency requirements, making them a versatile solution for modern IT challenges.

CONNECTIONS

There are many options for transparent and effective connectivity to the AWS Global Infrastructure and any Virtual Private Clouds (VPCs) you might be implementing.

Direct Connect

AWS Direct Connect makes establishing a dedicated network connection from your premises to AWS easy. Features are numerous and include the following:

- Establish private connectivity between AWS and your data centre, office, or colocation environment; keep in mind that you typically work with an AWS partner data centre, so the privacy of the connection is still relative.
- Potential reduction of your network costs (through savings of the transfer-out fee).
- Potential increase in bandwidth throughput.
- Typically, a more consistent network experience than Internet-based connections.
- Use 802.1Q VLANs, which enable you to partition the connection into multiple virtual interfaces that can access different resources.

VPC Endpoints

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

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

There are two types of VPC endpoints: interface endpoints and gateway endpoints. You should create the type of VPC endpoint required by the supported service.

Interface Endpoints (Powered by AWS PrivateLink)

An interface endpoint is an elastic network interface with a private IP address that serves as an entry point for traffic destined to a supported service. The following services are supported:

- API Gateway
- CloudWatch
- CloudWatch Events
- CloudWatch Logs
- CodeBuild
- Config
- EC2 API
- Elastic Load Balancing API
- Key Management Service
- Kinesis Data Streams

- SageMaker Runtime
- Secrets Manager
- Security Token Service
- Service Catalog
- SNS
- Systems Manager
- Endpoint services hosted by other AWS accounts
- Supported AWS Marketplace partner services

Gateway Endpoints

A gateway endpoint is a gateway that is a target for a specified route in your route table and is used for traffic destined to a supported AWS service. The following AWS services are supported:

- S3
- DynamoDB

VPC Peering

An AWS VPC peering connection connects two VPCs and enables you to route traffic between them privately. Instances in either VPC can communicate as if they are within the same network. You can create a VPC peering connection between your VPCs, with a VPC in another AWS account or a VPC in a different AWS region.

AWS uses the existing infrastructure of a VPC to create a VPC peering connection; it is neither a gateway nor a VPN connection and does not rely on a separate piece of physical hardware. There is no single point of failure for communication or a bandwidth bottleneck. Figure 10-2 shows the configuration of a VPC peering in AWS.

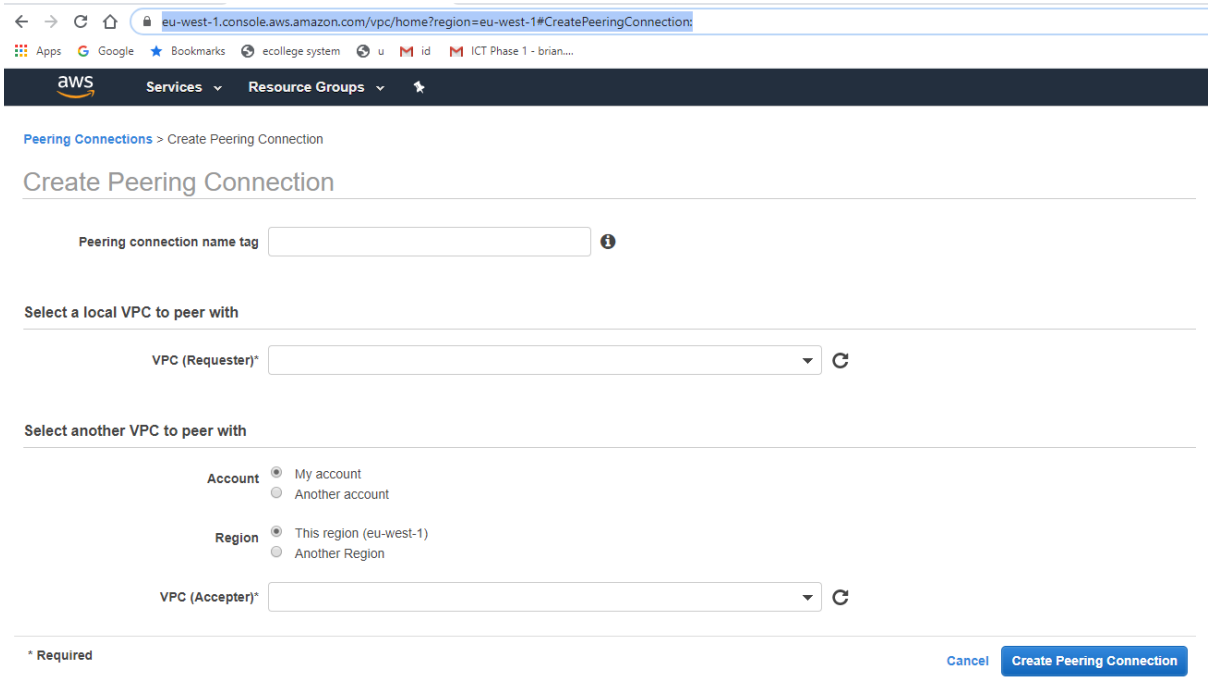

Figure 10-2 Configuring a VPC Peering

ClassicLink (deprecated)

This section is for interest only – it is unlikely to be on the exam

https://aws.amazon.com/blogs/aws/ec2-classic-is-retiring-heres-how-to-prepare/

Classic Link allows you to link your EC2-Classic instance to a VPC in your account within the same region. This allows you to associate the VPC security groups with the EC2-Classic instance, enabling communication between your EC2-Classic instance and instances in your VPC using private IPv4 addresses.

ClassicLink removes the need for public IPv4 or Elastic IP addresses to enable communication between instances in these platforms. It is available to all users with accounts that support the EC2-Classic platform and can be used with any EC2-Classic instance.

There is no additional charge for using ClassicLink. Standard charges for data transfer and instance usage apply.

Note

EC2-Classic instances cannot be enabled for IPv6 communication. You can associate an IPv6 CIDR block with your VPC and assign IPv6 addresses to resources in your VPC; however, communication between a ClassicLinked instance and resources in the VPC is over IPv4 only.

https://aws.amazon.com/blogs/aws/ec2-classic-is-retiring-heres-how-to-prepare/

EXAM PREPARATION TASKS

- **REVIEW ALL TOPICS**
- **DEFINE ALL KEY TERMS AND CHECK ANSWERS IN THE GLOSSARY.**
- **DO THE QUIZ – REPEAT UNTIL YOU PASS IT (100% PASSMARK).**

DEFINE KEY TERMS

Define the following key terms from this module and check your answers in the Glossary:

Region

Availability Zone

Direct Connect

VPC endpoint

VPC peering

ClassicLink (ClassicLink and EC2 classic are deprecated)

Q&A

1. Describe the AWS Global Infrastructure, from the largest component to the smallest.

2. How does AWS decide on the location of Availability Zones inside a region?

3. What AWS component permits you to allow traffic flows between your VPCs in your AWS account?