AWS Global Infrastructure: Availability Zones, Regions, Edge Locations, Regional Edge Caches

Amazon Web Services is a global public cloud provider, and as such, it has to have a global network of infrastructure to run and manage its many growing cloud services that support customers around the world. In this post, we'll take a look at the components that make up the AWS Global Infrastructure.

The components are:

- Availability Zones (AZs)
- Regions
- Edge Locations
- Regional Edge Caches

If you are deploying services on AWS, you'll want to have a clear understanding of each of these components, how they are linked, and how you can use them within your solution to YOUR maximum benefit. Let's take a closer look.

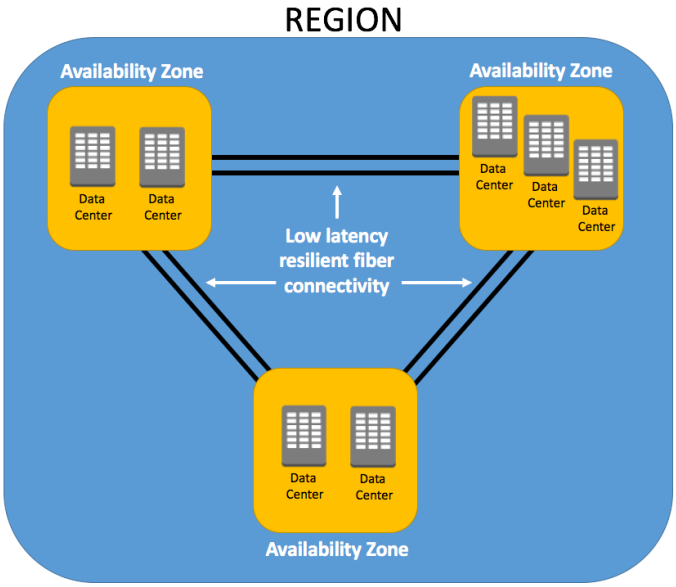
AWS Global Infrastructure: Availability Zones

Availability Zones and Regions are closely related. AZs are essentially the physical data centers of AWS. This is where the actual compute, storage, network, and database resources are hosted that we as consumers provision within our Virtual Private Clouds (VPCs). A common misconception is that a single availability zone is equal to a single data center. This is not the case. In fact, it's likely that multiple data centers located close together form a single availability zone.

Each AZ will always have at least one other AZ that is geographically located within the same area, usually a city, linked by highly resilient and very low latency private fiber optic connections. However, each AZ will be isolated from the others using separate power and network connectivity that minimizes impact to other AZs should a single AZ fail.

These low latency links between AZs are used by many AWS services to replicate data for high availability and resilience purposes. For example, when RDS (Relational Database Service) is configured for 'Multi-AZ' deployments, AWS will use synchronous replication between its primary and secondary database and asynchronous replication for any read replicas that have been created.

Often, there are three, four, or even five AZs linked together via these low latency connections. This localized geographical grouping of multiple AZs, which would include multiple data centers, is defined as an AWS Region.



AWS Global Infrastructure: Availability Zones, Regions, Edge Locations, Regional Edge Caches

Multiple AZs within a region allows you to create highly available and resilient applications and services. By architecting your solutions to utilize resources across more than one AZ ensures that minimal or no impact will occur to your infrastructure should an AZ experience a failure, which does happen).

Anyone can deploy resources in the cloud, but architecting them in a way that ensures your infrastructure remains stable, available, and resilient when faced with a disaster is a different matter. Making use of at least two AZs in a region helps you maintain high availability of your infrastructure and its always a recommended best practice.

AWS Global Infrastructure: Regions
As we now know, a Region is a collection of availability zones that are geographically located close to one other. This is generally indicated by AZs within the same city. AWS has deployed them across the globe to allow its worldwide customer base to take advantage of low latency connections. Every Region will act independently of the others, and each will contain at least two Availability Zones.

For example, if an organization based in London was serving customers throughout Europe, there would be no logical sense to deploy services in the Sydney Region simply due to the latency response times for its customers. Instead, the company would select the region most appropriate for them and their customer base, which may be the London, Frankfurt, or Ireland Region.

Having global regions also allows for compliance with regulations, laws, and governance relating to data storage (at rest and in transit). For example, you may be required to keep all data within a specific location, such as Europe. Having multiple regions within this location allows an organization to meet this requirement.

Similarly to how utilizing multiple AZs within a region creates a level of high availability, the same can be applied to utilizing multiple regions. Depending on the level of business continuity you require, you may choose to architect your AWS environment to support your applications and services across multiple regions, should an entire region become unavailable, perhaps due to a natural disaster.

You may want to use multiple regions if you are a global organization serving customers in different countries that have specific laws and governance about the use of data. In this case, you could even connect different VPCs together in different regions.

The number of regions is increasing year after year as AWS works to keep up with the demand for cloud computing services. At the time of publishing this article there are currently 20 Regions and 61 Availability Zones, with a further 4 Regions and 12 AZs planned.

Interestingly, not all AWS services are available in every region. This is a consideration that must be taken into account when architecting your infrastructure. Some services are classed as global services, such as AWS Identity & Access Management (IAM) or Amazon CloudFront, which means that these services are not tied to a specific region. However, most services are region-specific, and it's down to you to understand which services are available in which region.

Every single region is independent from other regions. The AWS GovCloud is an isolated region in the U.S. that is only available to U.S. government agencies and organizations in government-regulated industries, which must meet strict requirements.

AWS Global Infrastructure: Availability Zones, Regions, Edge Locations, Regional Edge Caches

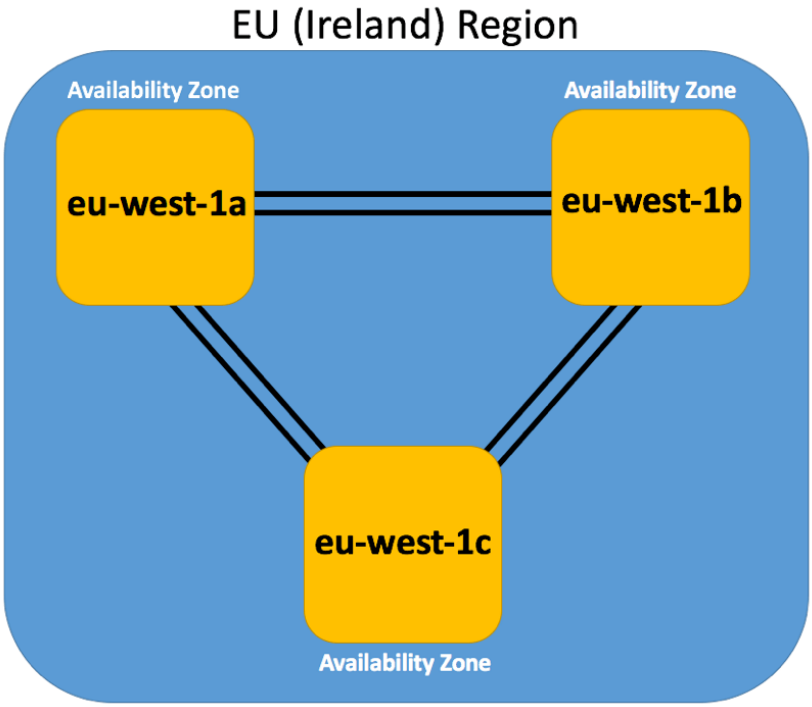
Region and Availability Zone Naming Conventions
AWS has a specific naming convention for both Regions and Availability Zones.
Depending on where you are viewing and using the Region name, it can be represented as two different names for the same Region.

Regions have both a 'friendly' name, indicating a location that can be viewed within the Management Console and a Code Name that is used when referencing regions programmatically, for example when using the AWS CLI.

Name	Code Name
US East (N. Virginia)	us-east-1
US East (Ohio)	us-east-2
US West (N. California)	us-west-1
US West (Oregon)	us-west-2
Canada (Central)	ca-central-1
EU (Ireland)	eu-west-1
EU (Frankfurt)	eu-central-1
EU (London)	eu-west-2
Asia Pacific (Tokyo)	ap-northeast-1
Asia Pacific (Seoul)	ap-northeast-2
Asia Pacific (Singapore)	ap-southeast-1
Asia Pacific (Sydney)	ap-southeast-2
Asia Pacific (Mumbai)	ap-south-1
South America (São Paulo)	sa-east-1

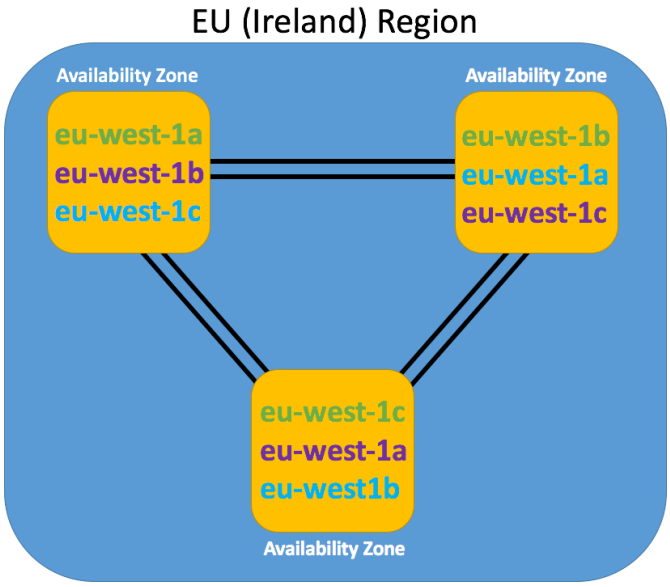
As you can see, the name in the first column is easier to associate to than that of the Code Name.
Availability Zones are always referenced by their Code Name, which is defined by the AZs Region Code Name that the AZ belongs to, followed by a letter. For example, the AZs within the eu-west-1 region (EU Ireland), are:

- eu-west-1a
- eu-west-1b
- eu-west-1c



AWS Global Infrastructure: Availability Zones, Regions, Edge Locations, Regional Edge Caches

An interesting point to be aware of here is that AWS maps these AZ letter identifiers to different physical AZs for different AWS accounts. This ensures that there is a more even distribution of resources across all AZs within a Region.



AWS Account:
111122223333
444455556666
777788889999

If you have multiple AWS accounts and you try to coordinate resources within the same AZ by selecting the same AZ Code Name, this may not necessarily mean that those resources are physically located within the same AZ as you can see in the image above.

AWS Global Infrastructure: Edge Locations
Edge Locations are AWS sites deployed in major cities and highly populated areas across the globe. They far outnumber the number of availability zones available.

While Edge Locations are not used to deploy your main infrastructures such as EC2 instances, EBS storage, VPCs, or RDS resources like AZs, they are used by AWS services such as Amazon CloudFront and AWS Lambda@Edge to cache data and reduce latency for end-user access by using the Edge Locations as a global Content Delivery Network (CDN).

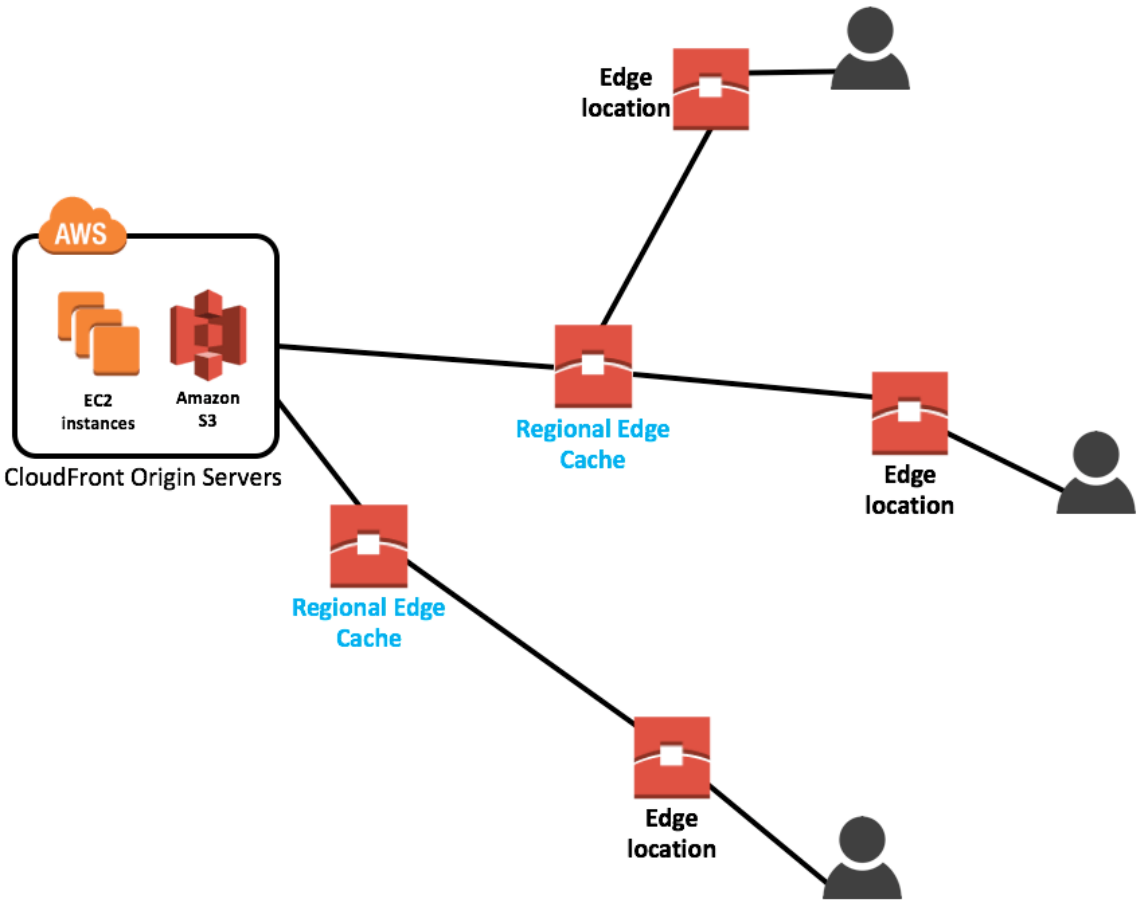
As a result, Edge Locations are primarily used by end users who are accessing and using your services.
For example, you may have your website hosted on EC2 instances and S3 (your origin) within the Ohio region with a configured CloudFront distribution associated. When a user accesses your website from Europe, they would be re-directed to their closest Edge Location (in Europe) where cached data could be read on your website, significantly reducing latency.

To understand more about how Amazon CloudFront achieves this, take a look at our courses and labs on the service.

AWS Global Infrastructure: Availability Zones, Regions, Edge Locations, Regional Edge Caches

AWS Global Infrastructure: Regional Edge Cache
In November 2016, AWS announced a new type of Edge Location, called a Regional Edge Cache. These sit between your CloudFront Origin servers and the Edge Locations. A Regional Edge Cache has a larger cache-width than each of the individual Edge Locations, and because data expires from the cache at the Edge Locations, the data is retained at the Regional Edge Caches.

Therefore, when data is requested at the Edge Location that is no longer available, the Edge Location can retrieve the cached data from the Regional Edge Cache instead of the Origin servers, which would have a higher latency.



I hope that this post has provided some clarity around the AWS global infrastructure of Availability Zones, Regions, Edge Locations, and Regional Edge Caches.

Understanding what each of these components can allow you to do will help you architect a resilient, highly available, secure, and low latency solution for you and your customers.

Stuart Scott
AWS Content Lead

@Stuart_A_Scott
https://uk.linkedin.com/in/stuartanthonyscott