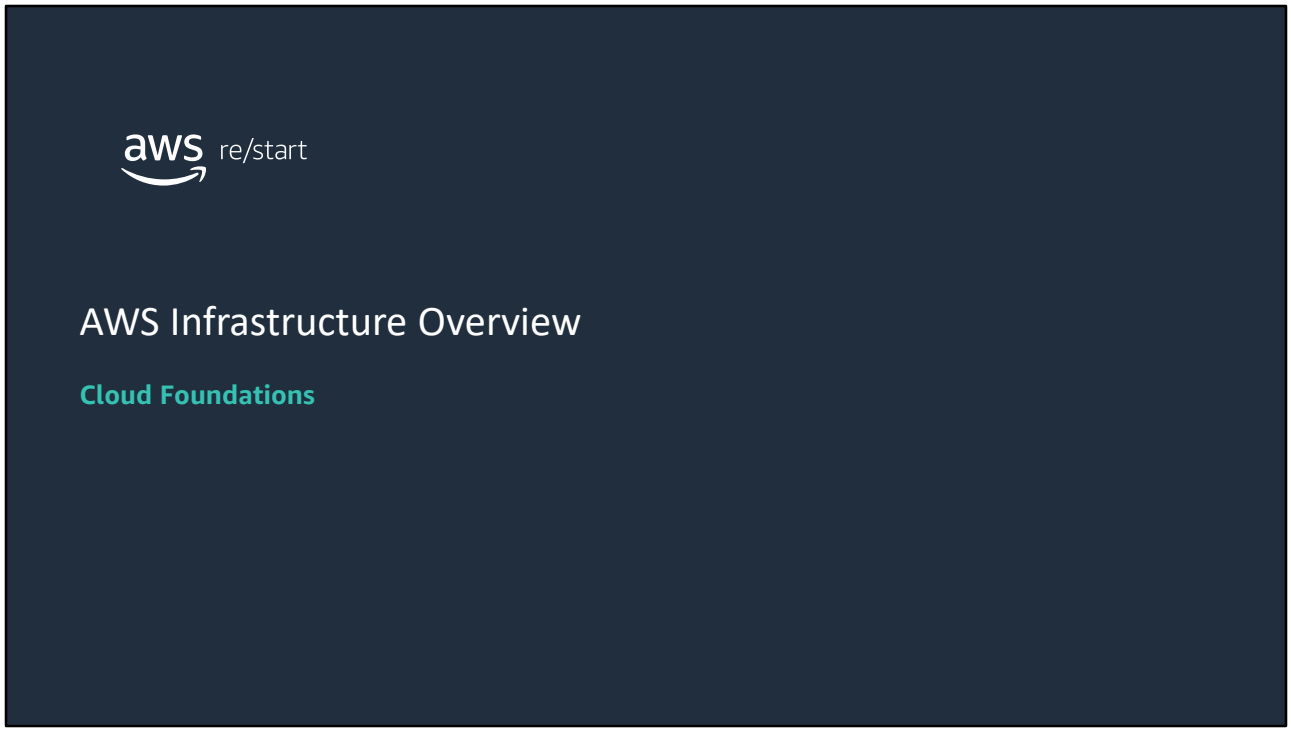


Welcome to AWS Infrastructure Overview.

What you will learn

At the core of the lesson

You will learn how to do the following:

- Describe the AWS Global Infrastructure and its features.
- Identify the difference between Amazon Web Services (AWS) Regions, Availability Zones, and points of presence (PoPs).

Key terms:

- Elastic infrastructure
- Scalable infrastructure
- Fault tolerance

aws re/start

In this module, you will review the AWS Global Infrastructure and its features. You will also learn how to identify the difference between Amazon Web Services (AWS) Regions, Availability Zones, and points of presence (PoPs).

AWS Global Infrastructure

The AWS Global Infrastructure is designed and built to deliver a flexible, reliable, scalable, and secure cloud computing environment with high-quality global network performance.

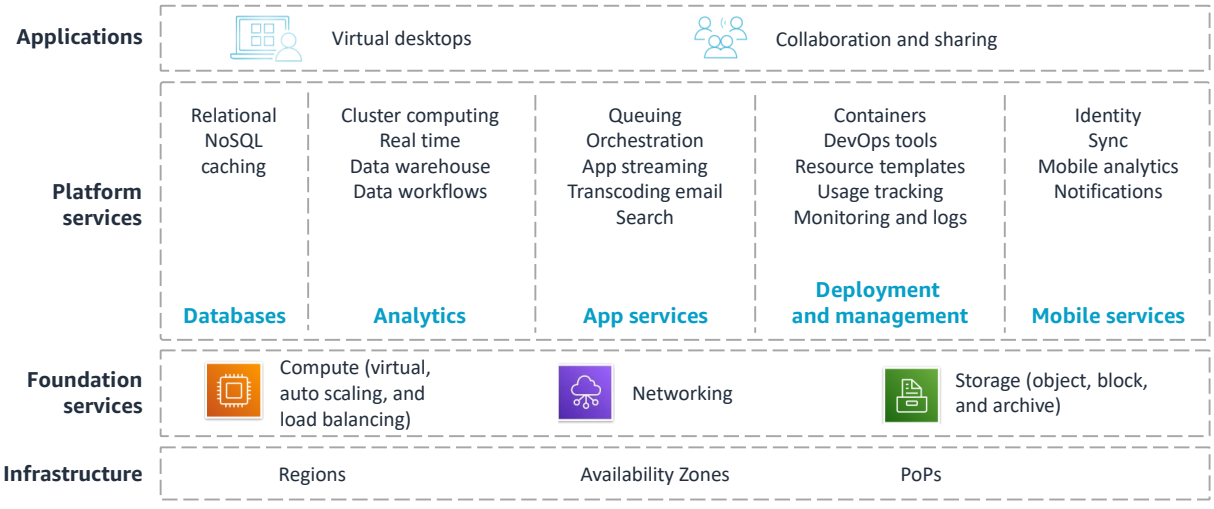
aws re/start

The diagram shows the 24 current AWS Regions in addition to a few Regions that will become available soon (as of August 2020).

To learn more about the current AWS Regions, refer to the Global Infrastructure page at https://aws.amazon.com/about-aws/global-infrastructure/?p=ngi&loc=0.

AWS Global Infrastructure elements

Regions, Availability Zones, and PoPs



aws re/start

As discussed earlier, AWS provides a broad set of services, such as compute, storage options, networking, and databases. They are delivered as an on-demand utility that is available in seconds with pay-as-you-go pricing. All these services reside on the AWS Global Infrastructure.

The AWS Global Infrastructure consists of three elements: Regions, Availability Zones, and points of presence (PoPs).

Next, you will take an in-depth look at the AWS Global Infrastructure and learn about these elements.

AWS Global Infrastructure (cont.)

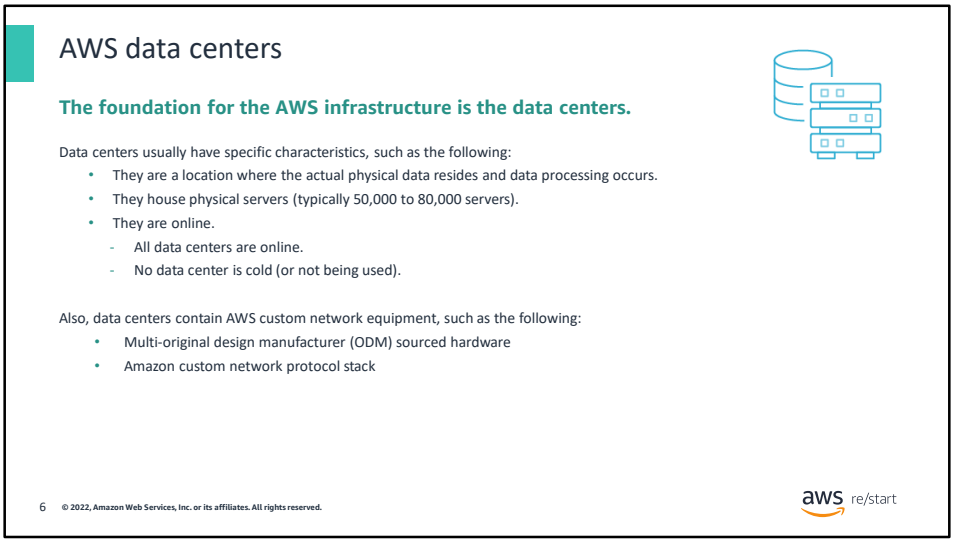
AWS Cloud infrastructure spans 84 Availability Zones in 26 geographic Regions around the world, with many more on the way.

aws re/start

AWS offers 7 Regions in North America, 1 Region in South America, 6 Regions in Europe, 1 Region in the Middle East, 1 Region in Africa, and 10 Regions in the Asia Pacific. Within each Region, there is one or more Availability Zone.

Benefits include security, availability, performance, a large global footprint, scalability, and flexibility.

For more information, see Regions and Availability Zones at https://aws.amazon.com/about-aws/global-infrastructure/regions_az/?p=ngi&loc=2 .



The foundation for the AWS infrastructure is the data centers. A data center is a location where the actual physical data resides and data processing occurs. AWS data centers are built in clusters in various global Regions.

Data centers are securely designed with several factors in mind:

- Each location is carefully evaluated to mitigate environmental risk.
- Data centers have a redundant design that anticipates and tolerates failure while maintaining service levels.
- To help ensure availability, critical system components are backed up across multiple isolated locations that are known as Availability Zones.
- To help ensure capacity, AWS continuously monitors service usage to deploy infrastructure to support availability commitments and requirements.
- Data center locations are not disclosed, and all access to them is restricted.
- In case of failure, automated processes move customer data traffic away from the affected area.

A single data center typically houses 50,000 to 80,000 physical servers.

All data centers are online and serving customers, so no data center is cold.

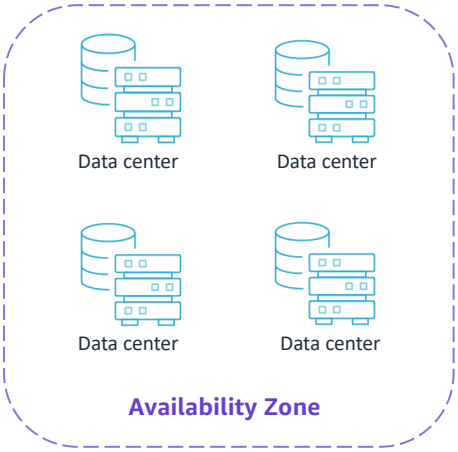
AWS uses custom, multi-ODM sourced network equipment. An original design manufacturer (or ODM) designs and manufactures products based on specifications from a second company. The second company then rebrands the products for sale.

For more information about AWS data center security, see the AWS Data Centers page at https://aws.amazon.com/compliance/data-center/.

AWS Availability Zones

Availability Zones

- Each Availability Zone is made up of **one or more data centers.**
- Availability Zones are designed for **fault isolation.**
- Availability Zones are **interconnected** with other Availability Zones by using high-speed private links.
- You choose your Availability Zones.
- **AWS recommends replicating across Availability Zones for resiliency.**

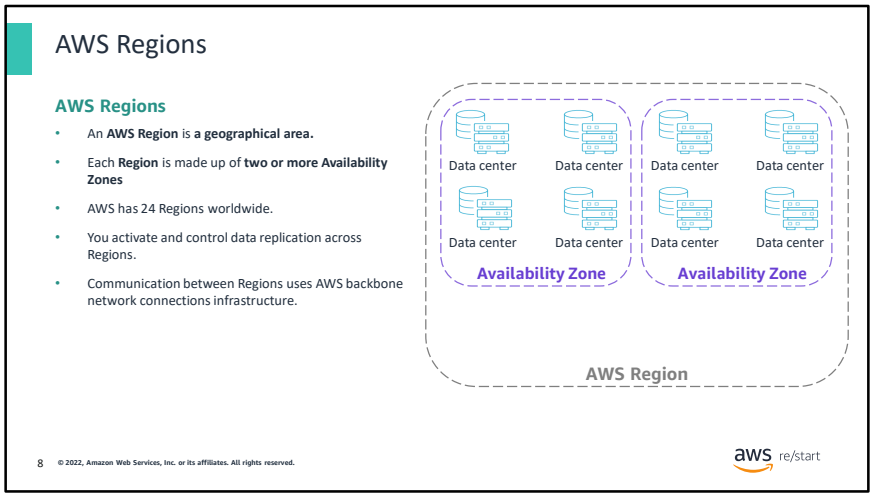


aws re/start

Availability Zones consist of one or more discrete data centers that are designed for fault isolation. They each have redundant power, networking, and connectivity resources that are housed in separate facilities. They are interconnected with other Availability Zones by using high-speed private links. Some Availability Zones have as many as six data centers. However, no data center can be part of two Availability Zones.

Each Availability Zone is designed as an independent failure zone. Availability Zones are physically separated in a typical metropolitan Region. They are located in lower-risk flood plains with specific flood-zone categorization that varies by Region. In addition to having a discrete, uninterruptible power supply and onsite backup generation facilities, they are each fed through different grids from independent utilities to further reduce single points of failure. Availability Zones are all redundantly connected to multiple tier-1 transit providers. Availability Zones in a Region are connected through low-latency links.

You are responsible for selecting the Availability Zones where your systems will reside. Systems can span across multiple Availability Zones. AWS recommends replicating across Availability Zones for resiliency. You should design your systems to survive temporary or prolonged failure of an Availability Zone if a disaster occurs. Distributing applications across multiple Availability Zones helps them remain resilient in most failure situations, including natural disasters or system failures.



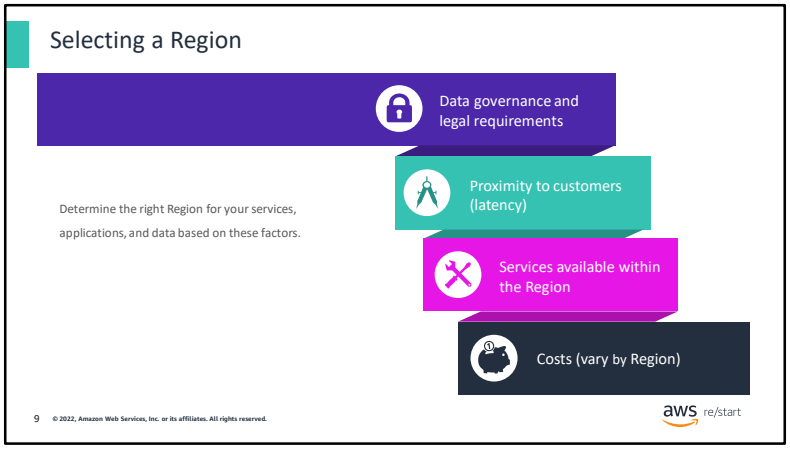
The AWS Cloud infrastructure is built around Regions and Availability Zones.

An AWS Region is a physical geographical location in the world where AWS has multiple Availability Zones. To achieve fault tolerance and stability, Regions are isolated from each other. Resources in one Region are not automatically replicated to other Regions. Each AWS Region contains two or more Availability Zones. As of August 2020, AWS had 24 Regions worldwide.

When you store data in a specific Region, it's not replicated outside that Region. AWS never moves your data out of the Region that you put it in. It's your responsibility to replicate data across Regions if your business needs require it. AWS provides information about the country and—where applicable—the state where each Region resides. You are responsible for selecting the Region to store data in based on your compliance and network latency requirements.

Consider these additional details. If you are using cloud computing services, you can deploy your application in multiple Regions. For instance, you can have an application in a Region that's nearest to your headquarters, such as San Diego on the West Coast of the US. You could then also have a deployable application in a Region on the East Coast of the US. Say that your largest customer base is in Virginia. With a few clicks, you can deploy in the US East Region to provide a better experience for your customers who are located there. You will reduce latency and increase agility for your organization within minutes with minimal cost.

Some Regions have restricted access. For example, the isolated AWS GovCloud (US) Region is designed so that US government agencies and customers can move sensitive workloads into the cloud by addressing their specific regulatory and compliance requirements.



You should consider a few factors when you select the optimal Region or Regions where you store data and use AWS services.

One essential consideration is **data governance and legal requirements**. Local laws might require that certain information be kept within geographical boundaries. Such laws might restrict the Regions where you can offer content or services. For example, consider the European Union (EU) Data Protection Directive.

All else being equal, it's generally desirable to run your applications and store your data in a Region that is as close as possible to the user and systems that will access them. This will help you **reduce latency**. CloudPing is one website that you can use to test latency between your location and all AWS Regions. For more information about CloudPing, see the CloudPing website at https://www.cloudping.info/.

Keep in mind that not all services are available in all Regions. For more information, see the AWS Regional Services page at https://aws.amazon.com/about-aws/global-infrastructure/regional-product-services/.

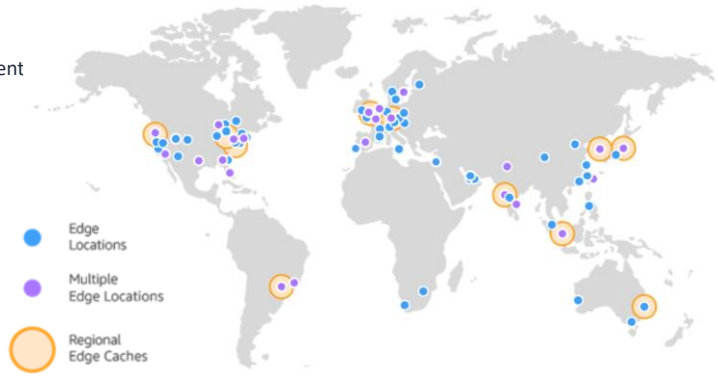
Finally, there is some variation in the **cost** of running services, which can depend on which Region you choose. For example, as of this writing, the per-hour cost to run a t3.medium Amazon Elastic Compute Cloud (Amazon EC2) On-Demand Linux Instance in the US East (Ohio) Region might differ from running the same instance in the Asia Pacific (Tokyo) Region.

In summary, when you select a Region, you should consider which Region offers the services that you need and where it's located. Doing so can help you optimize latency while reducing costs. It can also help you follow whatever regulatory requirements you might have.

Points of presence

AWS provides a global network of 216 PoP locations.

- The PoPs consist of 205 edge locations and 11 Regional edge caches.
- PoPs are used with Amazon CloudFront, a global content delivery network (CDN) that delivers content to end users with reduced latency.
- Regional edge caches are used for content with infrequent access.



aws re/start

A PoP is where end users access AWS services through either the **Amazon CloudFront** or the **Amazon Route 53** services.

As of August 2020, the global AWS infrastructure contained 216 PoPs, consisting of 205 edge locations and 11 Regional edge caches located in most of the major cities around the world. These PoPs serve requests for CloudFront and Route 53.

CloudFront is a content delivery network (or CDN) used to distribute content to end users to reduce latency. Route 53 is a Domain Name System (DNS) service. Requests going to either one of these services will be routed to the nearest edge location automatically.

Regional edge caches, used by default with CloudFront, are used when you have content that is not accessed frequently enough to remain in an edge location. Regional edge caches absorb this content and provide an alternative to the content having to be fetched from the origin server.

For more information about AWS Global Infrastructure, see the Global Infrastructure page at https://aws.amazon.com/about-aws/global-infrastructure/?p=ngi&loc=0.

AWS infrastructure features

Elastic and scalable:

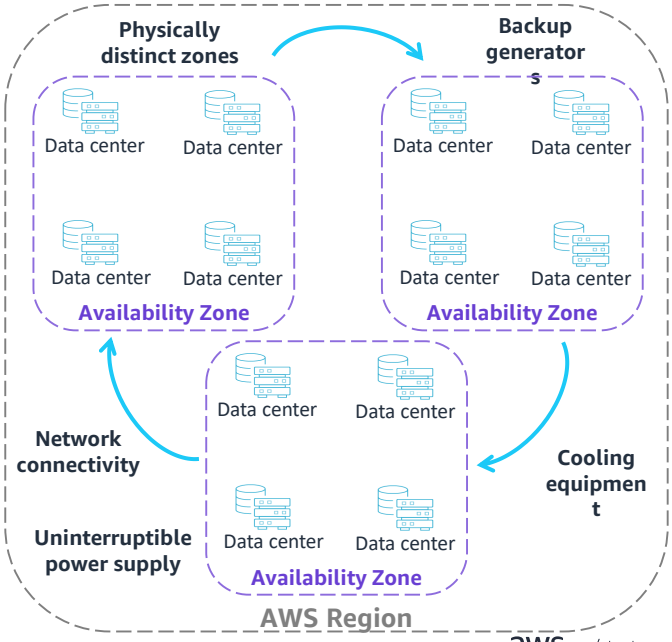
- Elastic infrastructure that dynamic adapts to capacity
- Scalable infrastructure that adjusts to accommodate growth

Fault-tolerant:

- Continues operating properly in the presence of a failure
- Includes built-in redundancy of components

Highly available:

- High level of operational performance with reduced downtime



aws re/start

The AWS Global infrastructure is built around Regions and Availability Zones. AWS Regions provide multiple physically separated, isolated Availability Zones. An AWS Region contains two or more Availability Zones.

An Availability Zone is a data center or collection of data centers. Availability Zones are connected with low-latency, high-throughput, highly redundant networking. Availability Zones are physically distinct. Each one has equipment like uninterruptible power supplies, cooling equipment, backup generators, and security, to help ensure uninterrupted operations.

This infrastructure has several valuable features:

- First, it is elastic and scalable. This means that resources can dynamically adjust to increases or decreases in capacity requirements. The infrastructure can also rapidly adjust to accommodate growth.
- Second, this infrastructure is fault tolerant, which means it has built-in component redundancy so that it can continue operations despite a failed component.
- Finally, it requires minimal to no human intervention while providing high availability with minimal downtime.

Key takeaways

- The AWS Global Infrastructure consists of **Regions** and **Availability Zones**.
- Your choice of a Region is typically based on **compliance requirements** or to **reduce latency**.
- Each Availability Zone is physically separate from other Availability Zones and has redundant power, networking, and connectivity.
- Edge locations and Regional edge caches (which are also called **points of presence**) improve performance by caching content closer to users.

aws re/start

This module includes the following key takeaways:

- The AWS Global Infrastructure consists of Regions and Availability Zones.
- Your choice of a Region is typically based on compliance requirements or to reduce latency.
- Each Availability Zone is physically separate from other Availability Zones and has redundant power, networking, and connectivity.
- Edge locations and Regional edge caches (which are also called points of presence) improve performance by caching content closer to users.

Thank you

Thank you for completing this module.