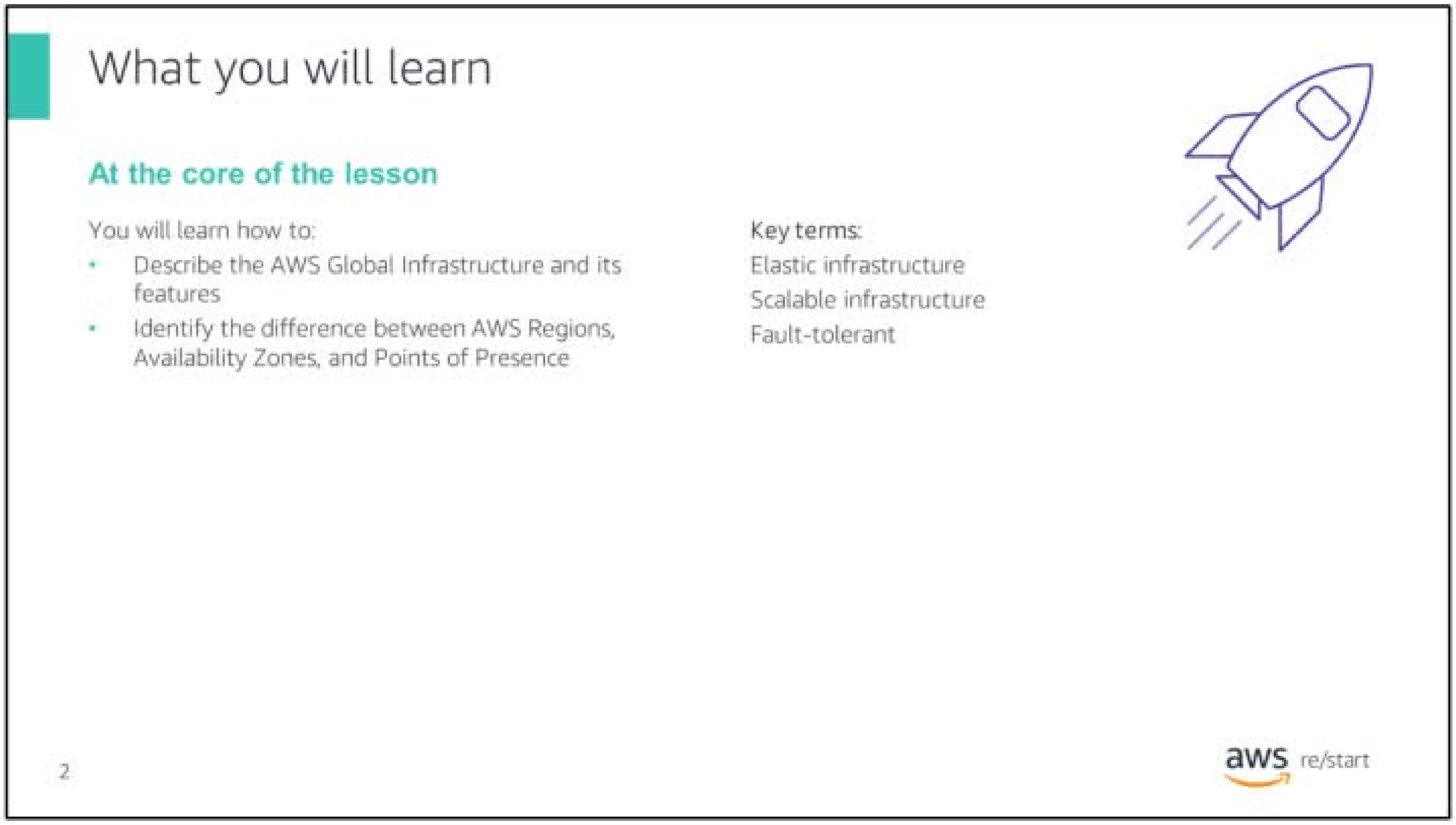
aws re/start

AWS Infrastructure Overview



What you will learn

At the core of the lesson

You will learn how to:

- Describe the AWS Global Infrastructure and its features
- Identify the difference between AWS Regions, Availability Zones, and Points of Presence

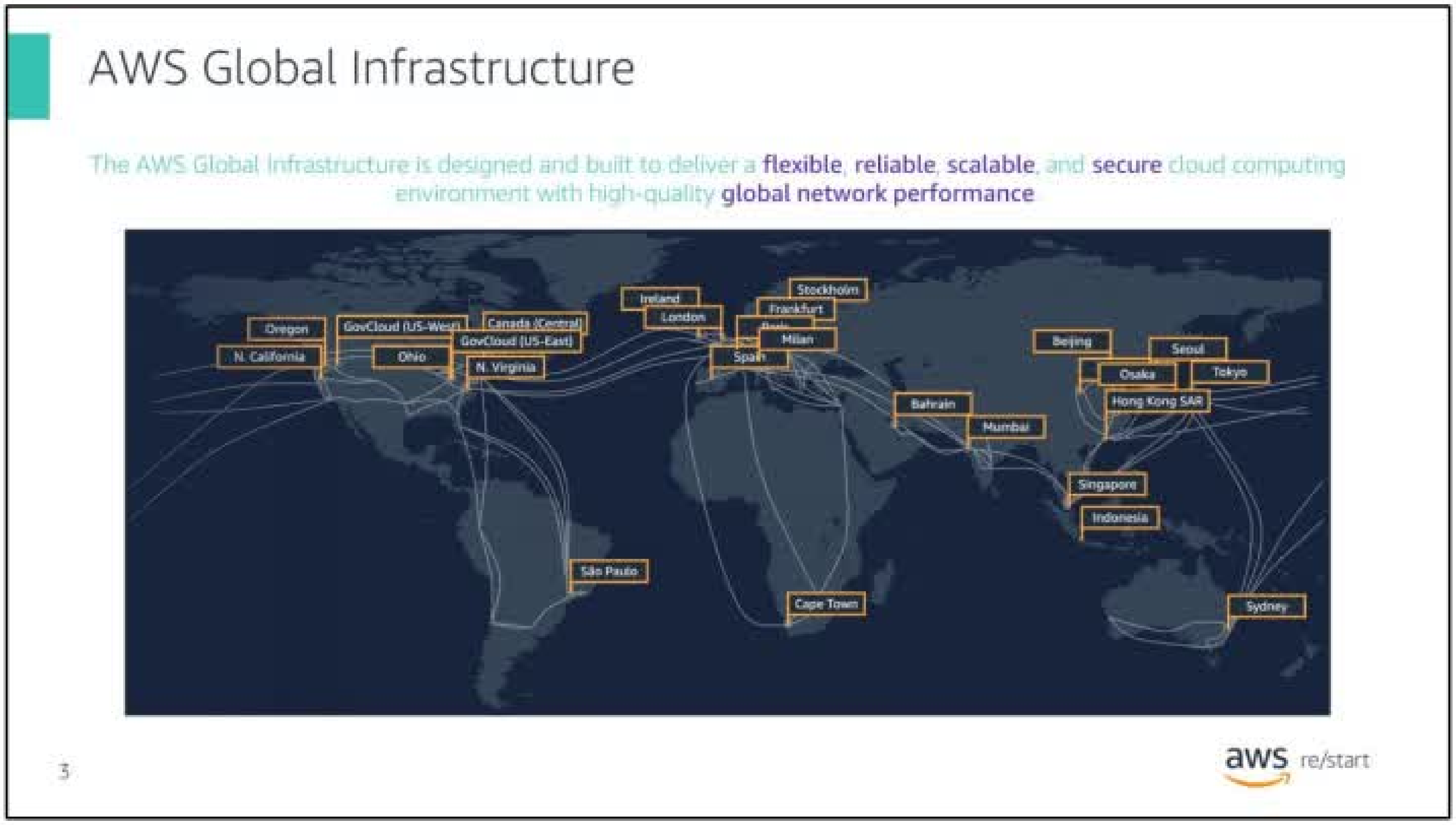
Key terms:

Elastic infrastructure

Scalable infrastructure

Fault-tolerant

In this module, you will review the AWS Global Infrastructure and its features. You will also learn how to identify the difference between AWS Regions, Availability Zones, and Points of Presence.



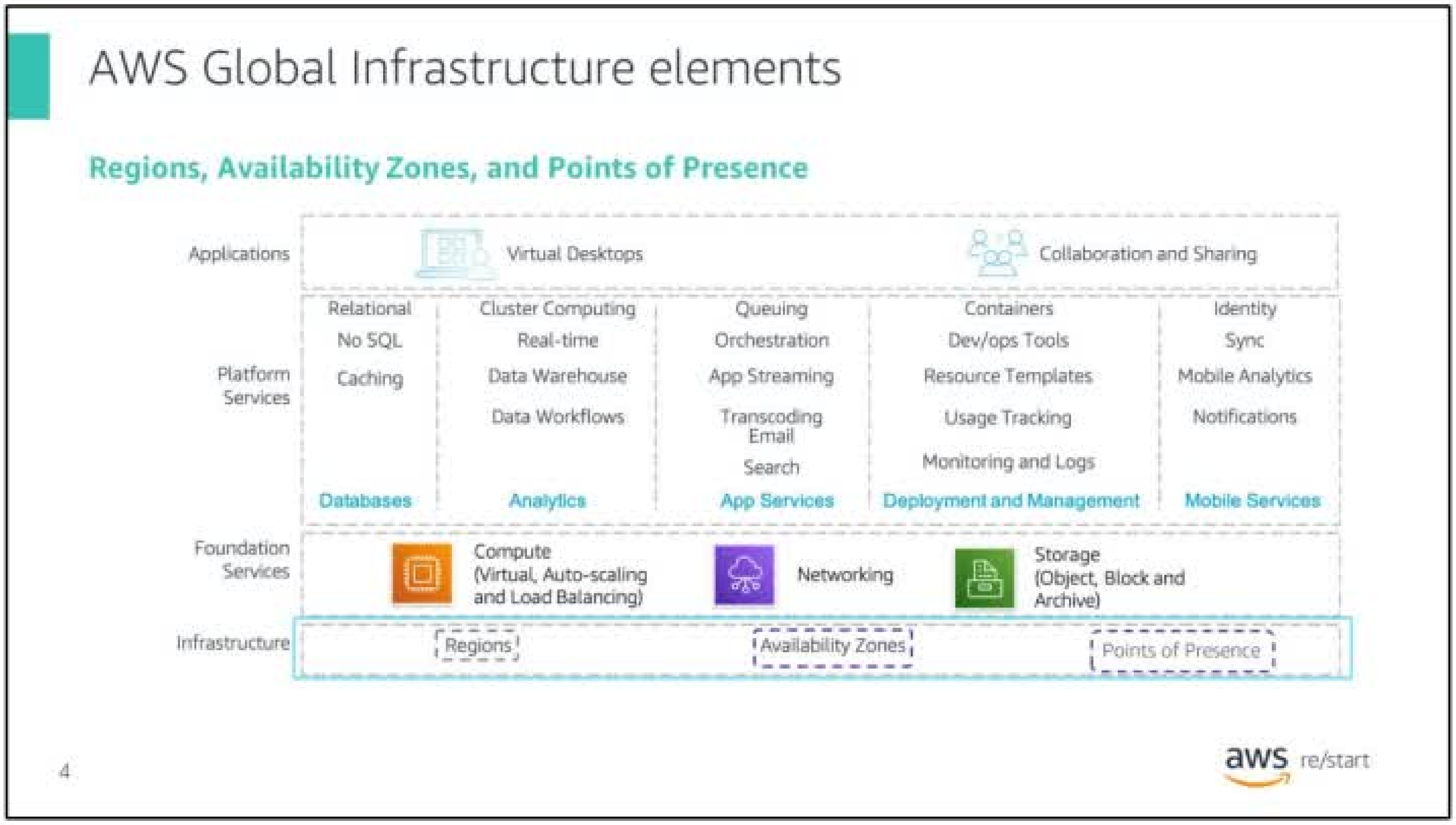
AWS Global Infrastructure

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.



AWS Global Infrastructure elements

Regions, Availability Zones, and Points of Presence

Applications: Virtual Desktops · Collaboration and Sharing

Platform Services:

Databases	Analytics	App Services	Deployment and Management	Mobile Services
Relational	Cluster Computing	Queuing	Containers	Identity
No SQL	Real-time	Orchestration	Dev/ops Tools	Sync
Caching	Data Warehouse	App Streaming	Resource Templates	Mobile Analytics
	Data Workflows	Transcoding Email	Usage Tracking	Notifications
		Search	Monitoring and Logs	

Foundation Services: Compute (Virtual, Auto-scaling and Load Balancing) · Networking · Storage (Object, Block and Archive)

Infrastructure: Regions · Availability Zones · Points of Presence

aws re/start

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

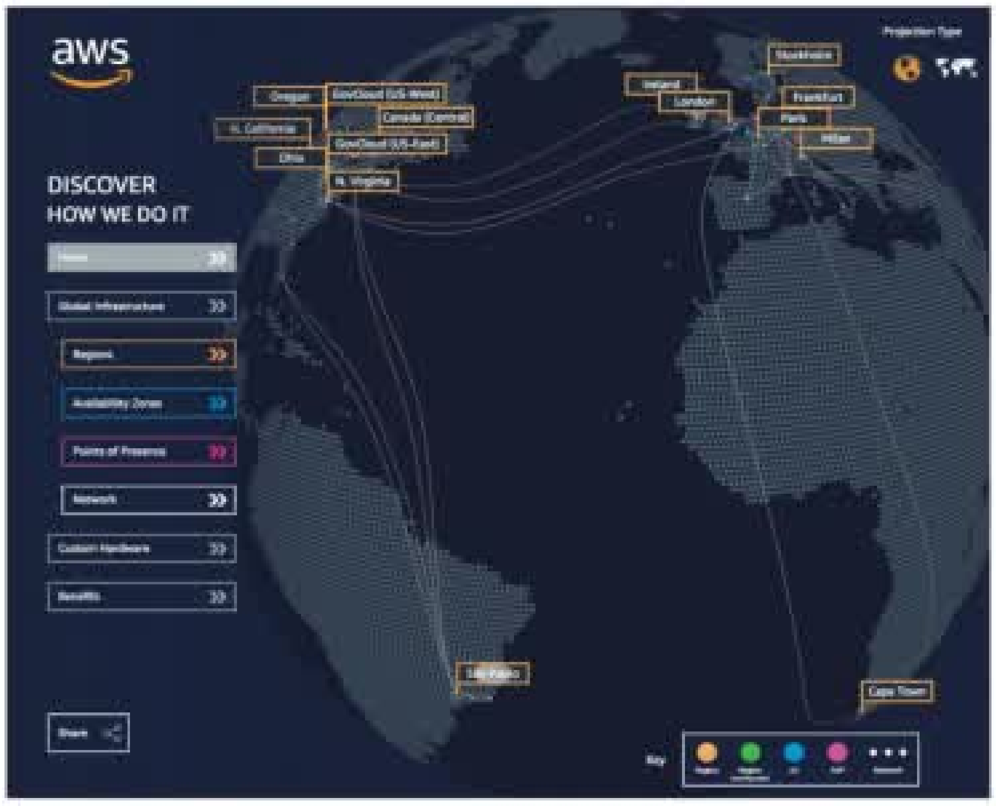
The AWS Global Infrastructure can be broken down into three elements: Regions, Availability Zones, and Points of Presence.

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

Demonstration

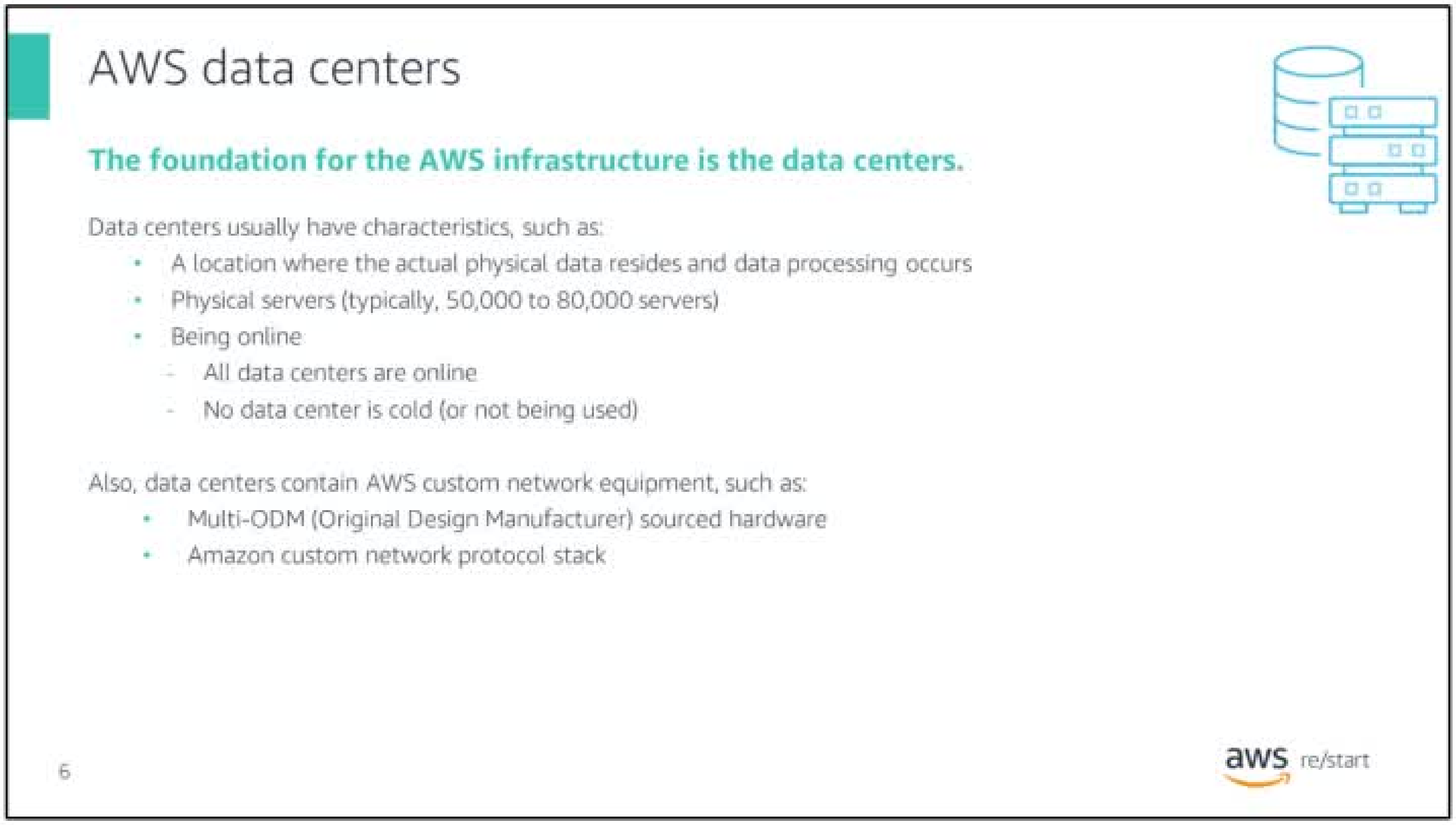
AWS Global Infrastructure

AWS Global Infrastructure tool



aws re/start

The educator might now choose to conduct a live demonstration of the AWS Global Infrastructure tool. This resource provides an interactive way to learn about the AWS Global Infrastructure. The remaining slides in this section cover many of the same topics and go into greater detail on some topics.



AWS data centers

The foundation for the AWS infrastructure is the data centers.

Data centers usually have characteristics, such as:

- A location where the actual physical data resides and data processing occurs
- Physical servers (typically, 50,000 to 80,000 servers)
- Being online
  - All data centers are online
  - No data center is cold (or not being used)

Also, data centers contain AWS custom network equipment, such as:

- Multi-ODM (Original Design Manufacturer) sourced hardware
- Amazon custom network protocol stack

aws re/start

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.

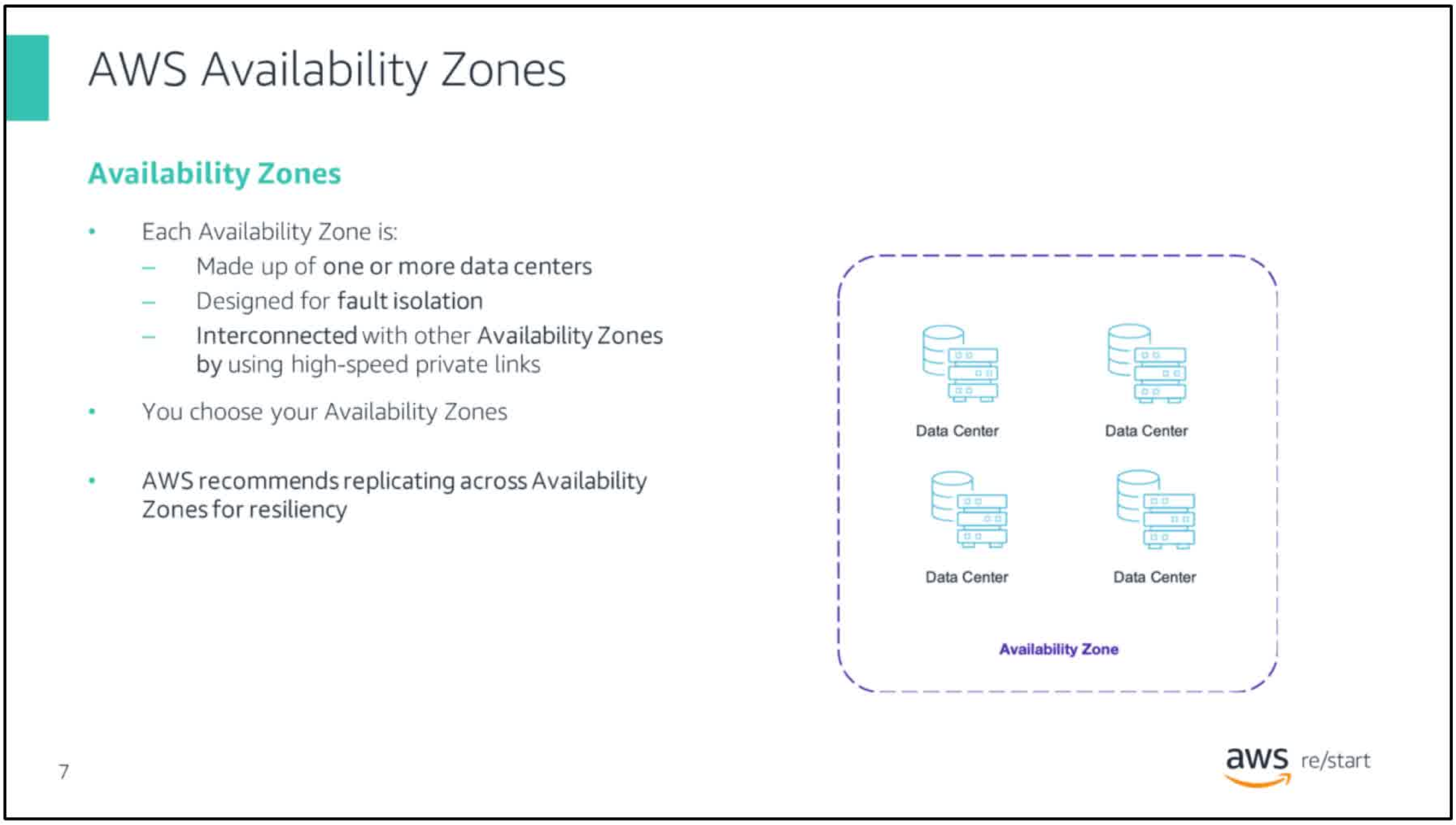
- Each location is carefully evaluated to mitigate environmental risk
- Data centers have a redundant design that anticipates and tolerates failure while maintaining service levels
- To ensure availability, critical system components are backed up across multiple isolated locations that are known as Availability Zones
- To 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.

To learn more about AWS data center security, refer to the AWS Data Centers page.



AWS Availability Zones

Availability Zones

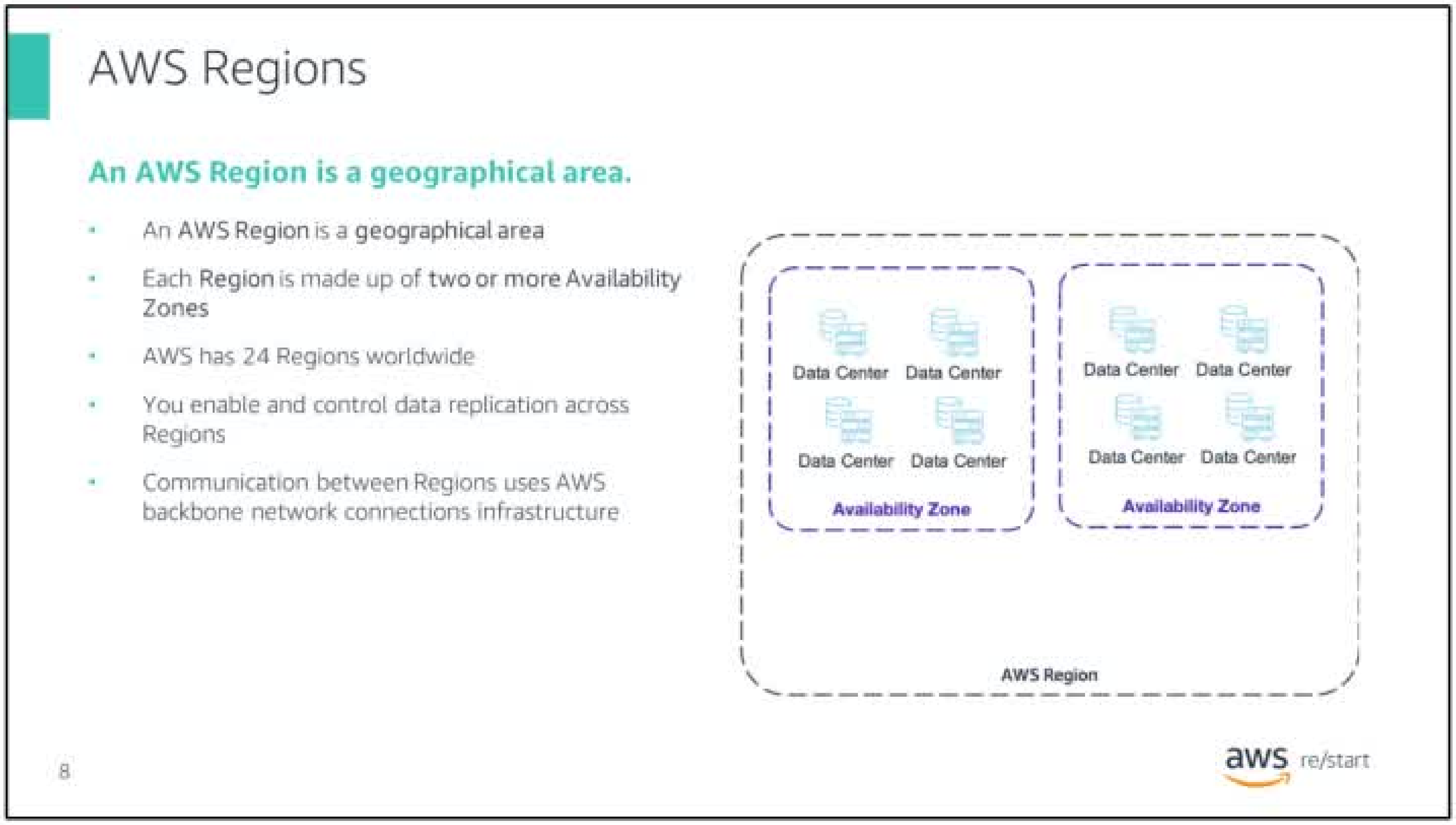
- Each Availability Zone is:
  - Made up of **one or more data centers**
  - Designed for **fault isolation**
  - **Interconnected** with other **Availability Zones** **by** using high-speed private links
- You choose your Availability Zones
- AWS recommends replicating across Availability Zones for resiliency

Data Center
Data Center
Data Center
Data Center
Availability Zone

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 via 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 enables them to remain resilient in most failure situations, including natural disasters or system failures.



AWS Regions

An AWS Region is a geographical area.

- An AWS Region is a geographical area
- Each Region is made up of two or more Availability Zones
- AWS has 24 Regions worldwide
- You enable and control data replication across Regions
- Communication between Regions uses AWS backbone network connections infrastructure

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

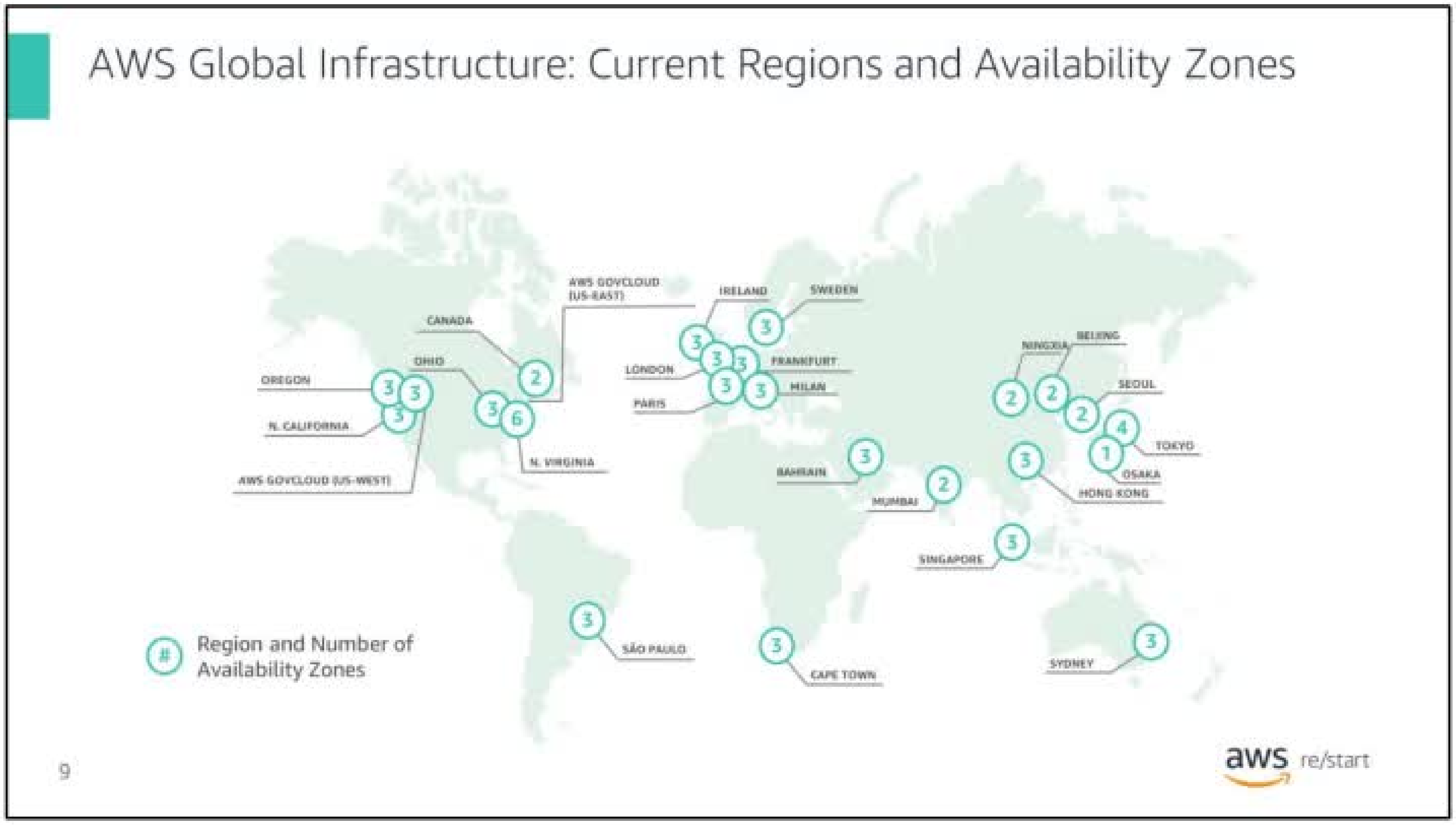
An AWS Region is a physical geographical location in the world where AWS has multiple Availability Zones. To achieve fault tolerance and stability, Regions are isolated from each other. Resources in one Region are not automatically replicated to other Regions. Each AWS Region contains two or more Availability Zones. As of August 2020, AWS has 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 easily 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 in the East Coast of the US. Say that your largest customer base is located 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 and with minimal cost.

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



AWS Global Infrastructure: Current Regions and Availability Zones

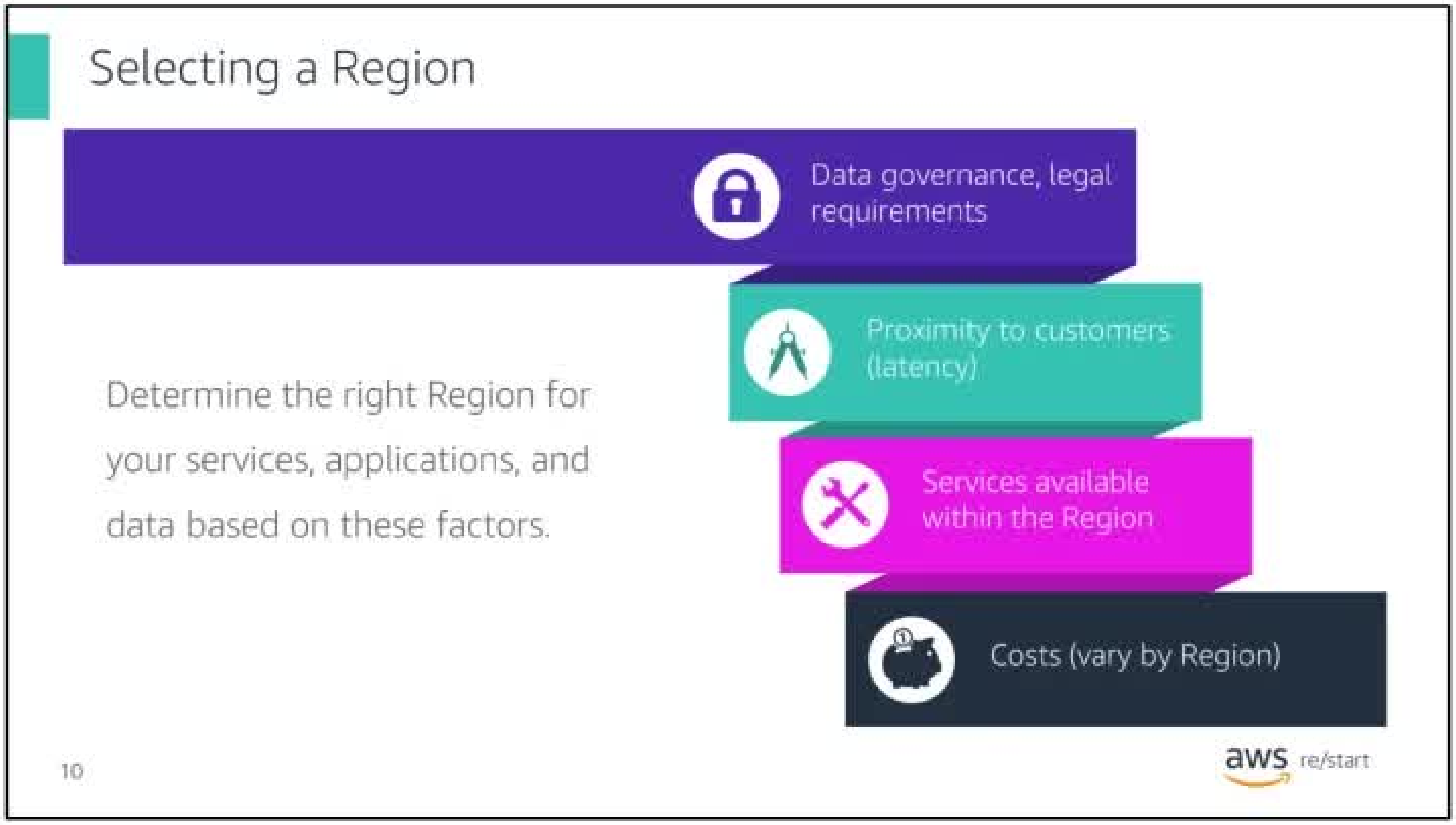
Region and Number of Availability Zones

9

aws re/start

As of August 2020, the AWS Global Infrastructure includes 24 Regions and 77 Availability Zones. AWS constantly expands its global infrastructure by steadily adding more Regions. By expanding infrastructure, AWS helps customers achieve lower latency and higher throughput, and helps customers ensure that their data resides only in the locations that they want.

To learn more about edge locations, refer to the Amazon CloudFront Key Features page.



Selecting a Region

Determine the right Region for your services, applications, and data based on these factors.

- Data governance, legal requirements
- Proximity to customers (latency)
- Services available within the Region
- Costs (vary by Region)

10

aws re/start

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

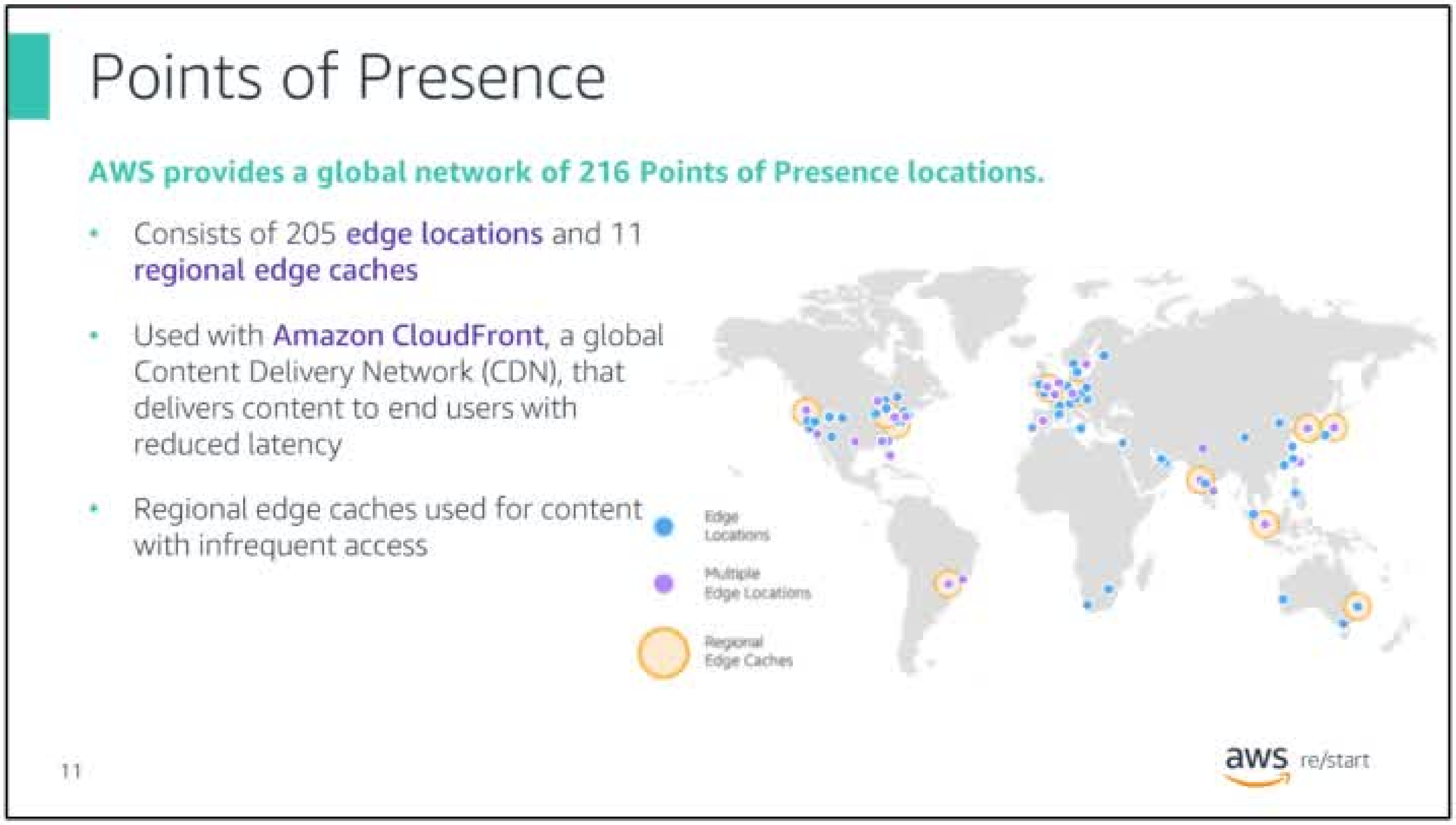
One essential consideration is **data governance and legal requirements**. Local laws might require that certain information be kept within geographical boundaries. Such laws might restrict the Regions where you can offer content or services. For example, consider the European Union (EU) Data Protection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

All else being equal, it's generally desirable to run your applications and store your data in a Region that's 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. To learn more about CloudPing, refer to the CloudPing website.

Keep in mind that not all services are available in all Regions. To learn more, refer to the AWS Regional Services page.

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, running an On-Demand t3.medium-size Amazon Elastic Compute Cloud (Amazon EC2) Linux Instance in the US East (Ohio) Region costs $0.0416 per hour. However, running the same instance in the Asia Pacific (Tokyo) Region costs $0.0544 per hour.

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 Points of Presence locations.

- Consists of 205 **edge locations** and 11 **regional edge caches**
- Used with **Amazon CloudFront**, a global Content Delivery Network (CDN), that delivers content to end users with reduced latency
- Regional edge caches used for content with infrequent access

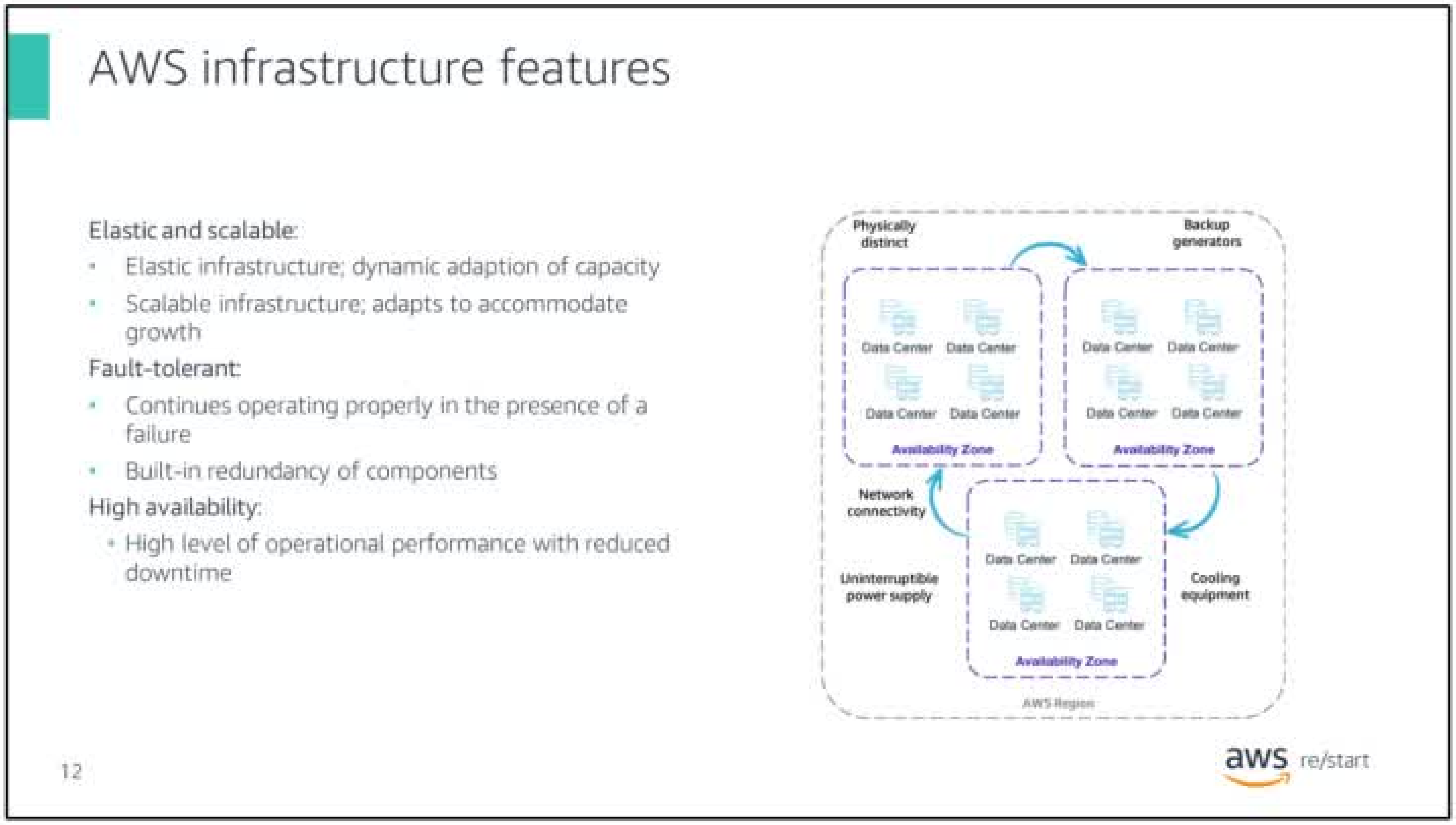
A Point of Presence is where end users access AWS services through either the **Amazon CloudFront** or the **Amazon Route 53** service.

As of August 2020, the global AWS infrastructure contains 216 Points of Presence (PoP) 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 Amazon CloudFront and Amazon Route 53.

Amazon CloudFront is a content delivery network (or CDN) used to distribute content to end users to reduce latency. Amazon 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 Amazon 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 that content having to be fetched from the origin server.

To learn more about the AWS Global Infrastructure, refer to the Global Infrastructure page.



AWS infrastructure features

Elastic and scalable:

- Elastic infrastructure; dynamic adaption of capacity
- Scalable infrastructure; adapts to accommodate growth

Fault-tolerant:

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

High availability:

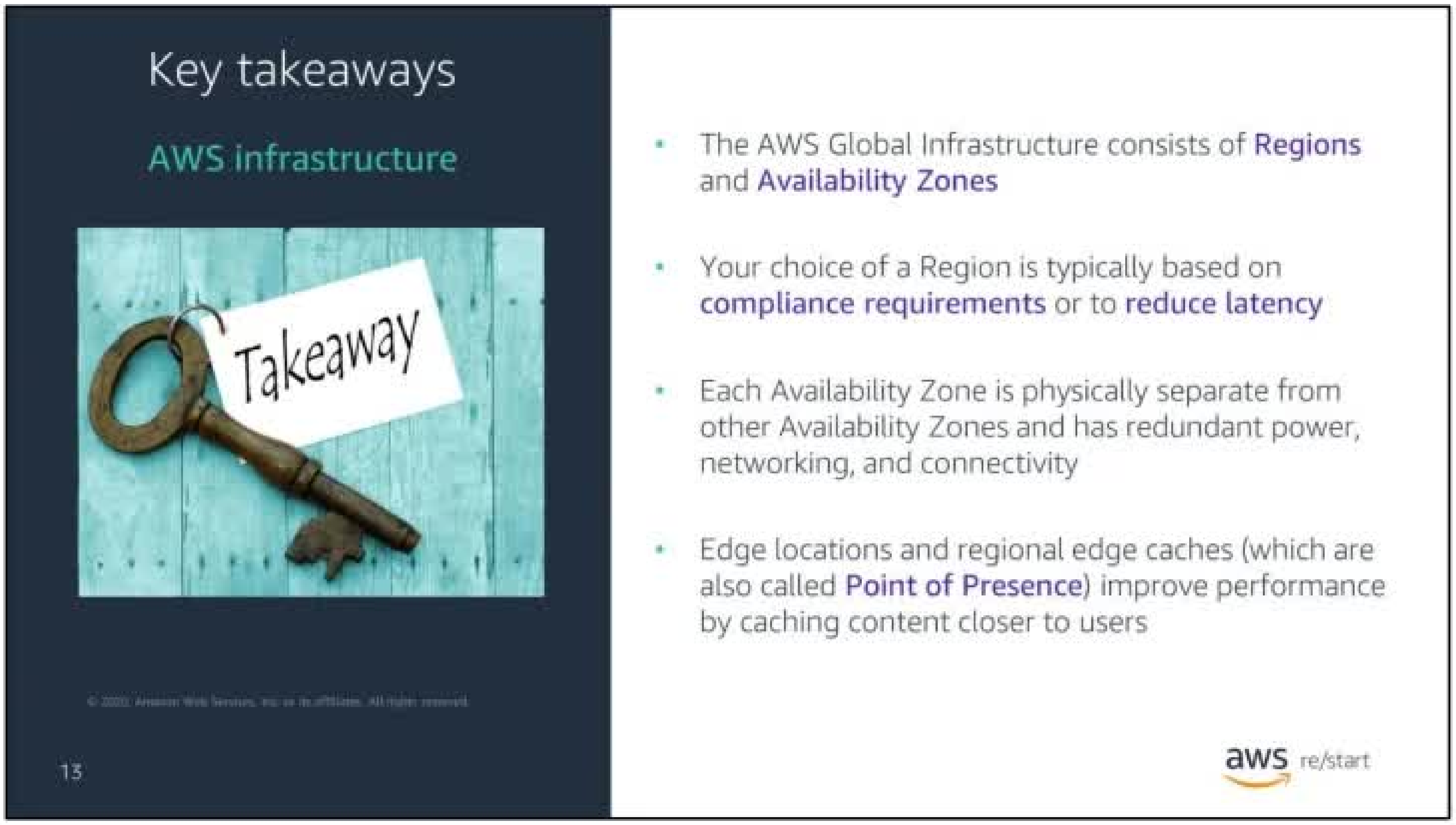
- High level of operational performance with reduced downtime

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

An Availability Zone is a data center or collection of data centers that are connected with low latency, high throughput, and highly redundant networking. Availability Zones are physically distinct and each has equipment like Uninterruptible Power Supplies, cooling equipment, backup generators, and security, to ensure uninterrupted operations.

This infrastructure has several valuable features:

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



Key takeaways

AWS infrastructure

- 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 **Point of Presence**) improve performance by caching content closer to users

Some key takeaways from this lesson include:

- 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 Point of Presence) improve performance by caching content closer to users