# AWS Global Infrastructure

### AWS Global Infrastructure:

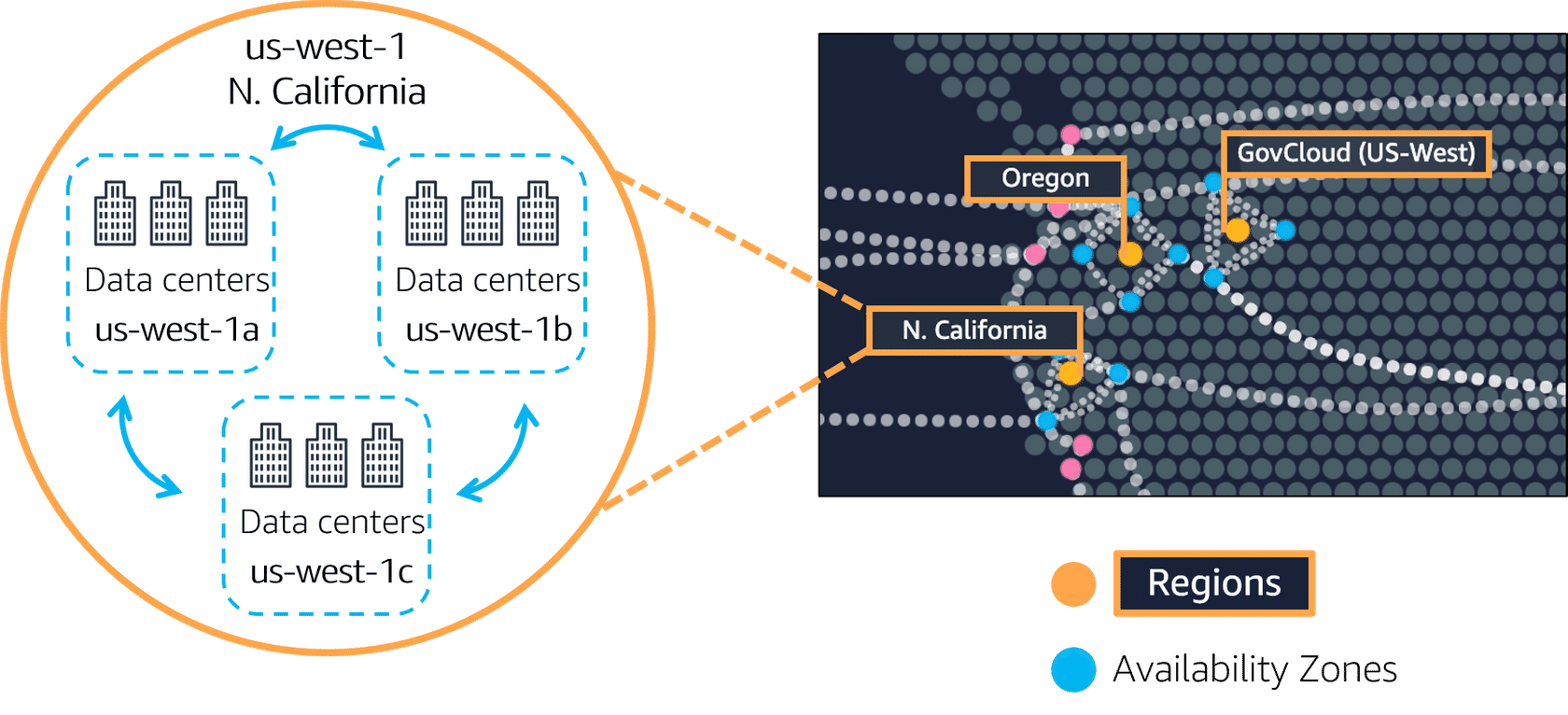
* Global infrastructure is set up to ensure high availability and fault tolerance no matter what the problem is.
* It is like chain of a business at various locations that is developed so that customers can access the products from nearer locations without having to reach main branch if they want anything.

#### **AWS Regions:**

* Key factors to choose a region are:
  1. Compliance with data governance and legal requirements:
     + Because of some reasons and restrictions, you may need to choose a region from a specific area.
     + For example, if you are working for Government of India, they may ask you to choose region that comes under Indian boundaries.
  2. Proximity to the customers:
     + Selecting a region that is close to the customers those will access your application or service.
     + For example, if you are streaming an IPL match then most of the viewers will be from India so you will definitely go for Indian region (ap-south-1) of infrastructure.
  3. Availability service within a region:
     + All the services of AWS are not readily available in each region as number of services keeps on increasing and some of them may not have ready infrastructure in each region.
     + For example, AWS Braket a quantum service by AWS whose required architecture may not be present everywhere so if you want to use that service then you need to choose the region in which it is available.
  4. Pricing:
     + AWS pricing is region specific so you will get charged differently if you setup infrastructure in different regions.
     + For example, if you want to run an app in Asia and charges for some services are less in other region than India then you may opt for it.

#### **Availability Zones:**

* When we build a business, we don’t want to have our business in a state that if a single centre faces some failure, then whole business gets disrupted instead, we can setup business at multiple centres to have availability in case of disaster.
* The same way AWS has multiple “Availability Zones” (data centres) in each region so that if a single data centre faces any disaster you can spin up your load to that region.
* An Availability Zone is a single data centre or a group of data centres within a Region.
* It is always suggested to develop any application in at least two regions to avoid application failure in case of emergency.
* AWS calls a single data centre or a group of data centres, an Availability Zone or AZ. Each Availability Zone is one or more discrete data centres with redundant power, networking, and connectivity. When you launch an Amazon EC2 instance, it launches a virtual machine on a physical hardware that is installed in an Availability Zone. This means each AWS Region consists of multiple isolated and physically separate Availability Zones within a geographic Region.
* Each of them is physically separated by a large distance so that if any disaster happens or connection is lost at a single location another AZ will always remail available.
* If you try to create a Load Balancer it will prompt to choose more than one AZ.

Image depicting regions in us-west and AZs in us-west-1 (N. California)

* Planning for failure and deploying applications across multiple Availability Zones is an important part of building a resilient and highly available architecture.

#### **CDN & Edge Locations:**

* Let’s take an example of a clothing company, if it gains fame across the country then they do not start their production across whole country instead they open their outlets in various cities.
* The same way company’s production plant is the main data centre where data is stored in AWS but edge locations are sites where cached copies of the data are present to reduce the latency and for faster delivery.
* Caching copies of data closer to the customers all around the world uses the concept of content delivery networks, or CDNs (Content Delivery Networks).
* CDNs are known as Amazon CloudFront in AWS.
* Amazon CloudFront is a service that helps deliver data, video, applications, and APIs to customers around the world with low latency and high transfer speeds.
* Amazon CloudFront uses what are called Edge locations, all around the world, to help accelerate communication with users, no matter where they are.
* Edge locations are separate from regions, so you can push content from inside a Region to a collection of Edge locations around the world, in order to accelerate communication and content delivery.
* AWS Edge locations, also run more than just CloudFront. They run a domain name service, or DNS, known as Amazon Route 53, helping direct customers to the correct web locations with reliably low latency.

\*Document is created while learning from AWS skill builder, so it contains some info from there and it belongs to them only.