How Cloud Deploys our Application Across the Globe? – Part 1

This lesson discusses deploying an app in multiple data centers.


We'll cover the following ︿

- Overview
- Communication between regions, availability zones, and data centers

---

Overview

Cloud providers deploy our service across the globe in different *regions* and *availability zones* to keep the service available, having a contingency plan for disasters and placing the code near the end-user to cut down network latency.

A *region* is a geographic area. For instance, *Las Vegas, Delhi, Melbourne, and Toronto* are different regions. A *region* contains multiple *availability zones* also called *AZs*. Further, these *AZs* contain multiple data centers. *Regions* may also contain multiple *local zones* or *LZs*.
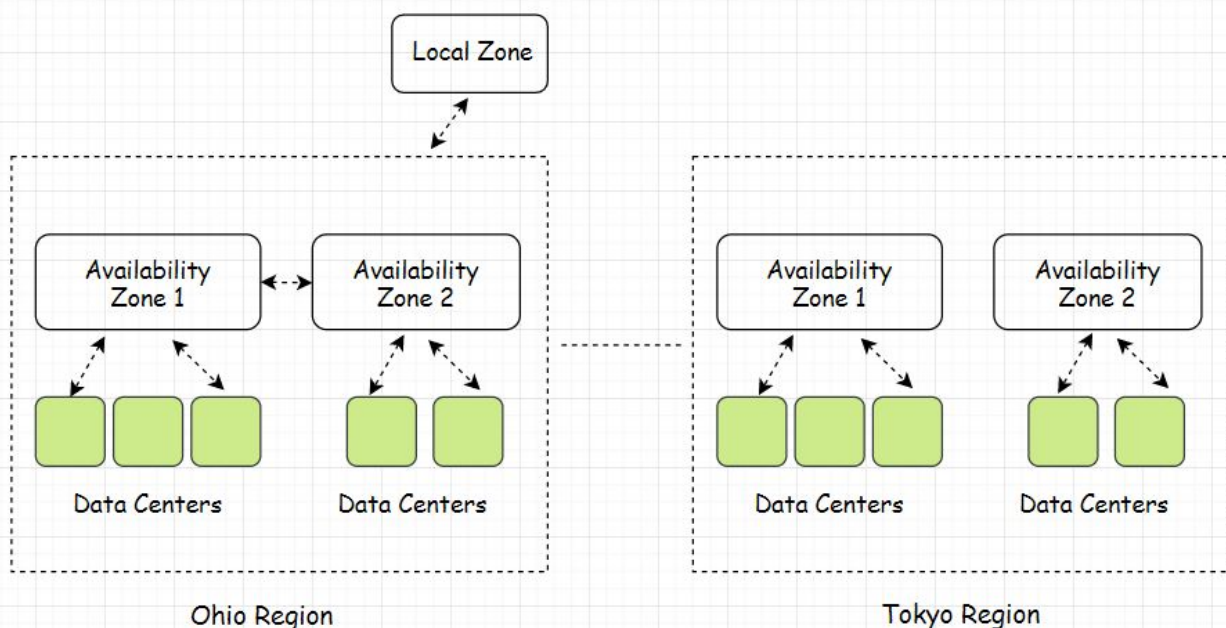
Regions, Availability Zones, Data Centers

Taking *AWS Global Infrastructure* as a reference; its *Ohio region* contains three *availability zones*, the *Tokyo region* contains four *availability zones*. Meanwhile, the *Oregon region* contains four *availability zones* and one *local zone*, and so on.

Local zones are an extension of a region in case additional computational power or other infrastructural resources are required in a *region. LZs* can be set up near the end-user to augment the *regions'* power and to cut down the latency. A *local zone* set up near the end-user helps deploy latency-sensitive services like stock trading, online multiplayer games, and so on.

If we have deployed our service in a *region* and want to cut network latency down, we can take advantage of *local zones*. Though not all the *regions* have the ability to set up *local zones*. The availability of an *LZ* in a *region* largely depends on the cloud provider.

Communication between regions, availability zones, and data centers

Regions are kept physically isolated and independent of each other in terms of power supply, water supply, hardware etc. This ensures service availability and fault-tolerance from a regional perspective. If any single region goes down, it doesn't impact the other regions.

Also, there are often legal constraints on companies dealing with sensitive data to keep it within a certain region. Having isolated deployments in different regions enables them to comply with the law. The government and other legal entities do not want their data to be replicated to other regions across the globe for obvious security reasons.

The *availability zones* contain multiple data centers. Every individual *availability zone* is run by its own set of data centers. Data centers are not shared between the availability zones to ensure service availability and fault-tolerance at the zonal level. If an *AZ* goes down due to a local disaster, the other availability zones are not impacted due to this. An availability zone is generally run by two to three data centers. A large availability can also be run by five to six data centers.

Within an availability zone, the data centers are connected to each other via a low latency network. Even the network is made redundant to cope with any sort of
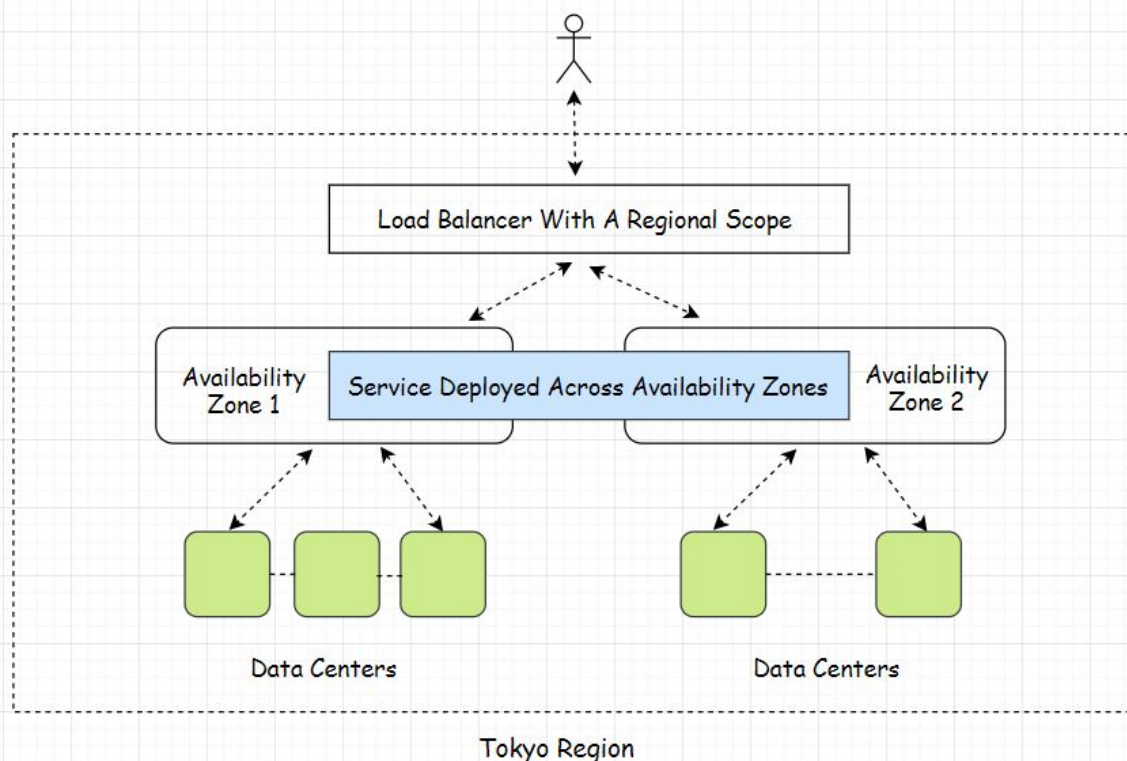
failure. When we deploy our application on the cloud, we are prompted by the cloud provider to select a region and an availability zone to run our application in.

Ideally, the region and the *AZ* nearest to our end users is selected to keep the network latency low.

The cloud deploys our application spanning multiple availability zones to avoid a single point of failure from an *AZ* standpoint. Even if an entire zone goes offline, the requests will be routed to a different zone. This is achieved with the help of a *load balancer* that runs at a regional level outside the zones. These load balancers have a regional scope.

This is the standard setup of *regions, availability zones*, and *data centers*. Businesses can always tweak their deployments based on their requirements with the help of cloud providers.

This is a good read on how AWS ensures service availability during natural or man-made disasters.

Additionally, here is a nice visual tool that gives insight into Google Cloud's global infrastructure.

> *Recommended read*: A closer look at Netflix's Christmas Eve Outage.

In the next lesson, let's discuss multi-regional deployments.