What Is Infrastructure (Infra)?

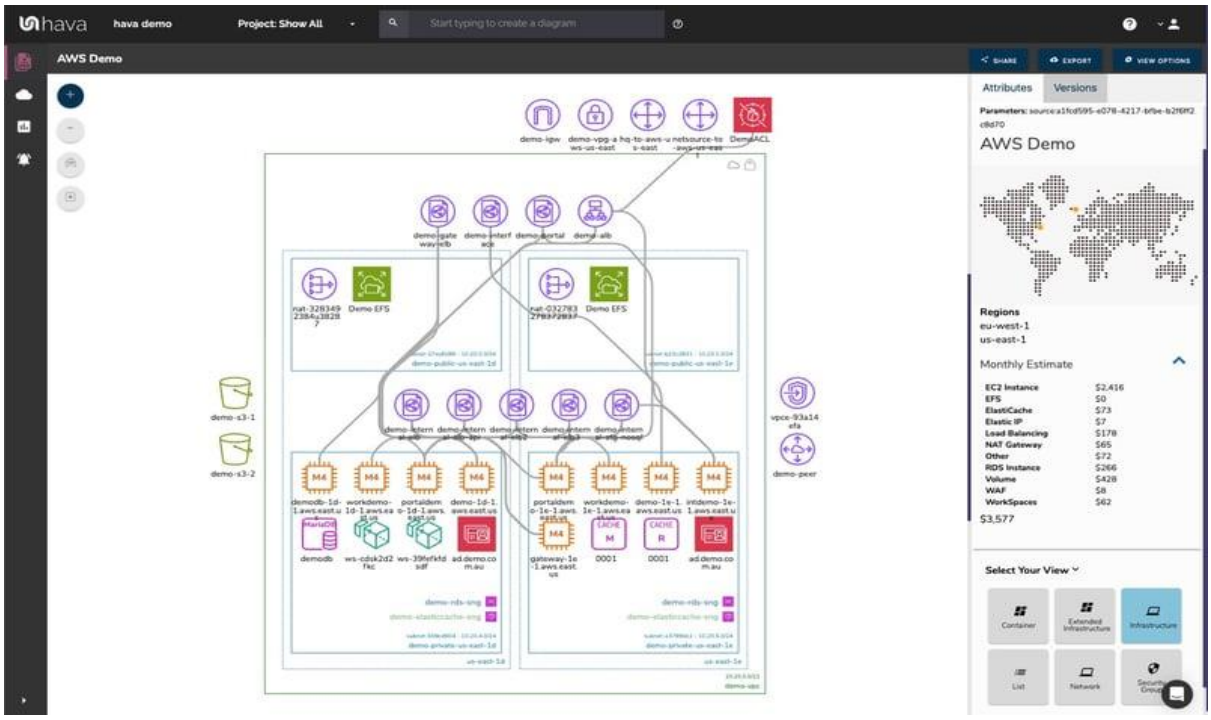
In the context of computing and cloud systems, **Infrastructure (Infra)** refers to all the foundational components required to run applications. Traditionally, this includes:

- Servers
- Storage systems
- Networking equipment
- Data centers
- Virtualization platforms
- Operating systems
- Security components

In AWS:

Infrastructure is delivered as on-demand cloud services. Instead of owning physical servers, companies “rent” resources from AWS.

This is why AWS is called *Infrastructure as a Service (IaaS)*—it provides compute, storage, and networking infrastructure via the cloud.



Define what is regions

A **Region** is a geographically distinct area with a collection of data centers. Each region operates independently, ensuring that a failure in one region doesn't affect others.

- **Location:** Each region is placed in different geographical locations around the world (e.g., US East, Europe West, Asia Pacific).
- **Redundancy:** Regions are isolated to prevent failures from spreading across them.
- **Data Residency:** Data hosted in a region typically remains within that geographical area, which helps with regulatory and compliance requirements.
- **Pricing:** Pricing can vary between regions due to operational costs, energy prices, and local taxes.

For example, AWS has regions like:

- **US East (N. Virginia)**
- **US West (Oregon)**
- **Europe (Frankfurt)**
- **Asia Pacific (Singapore)**

.

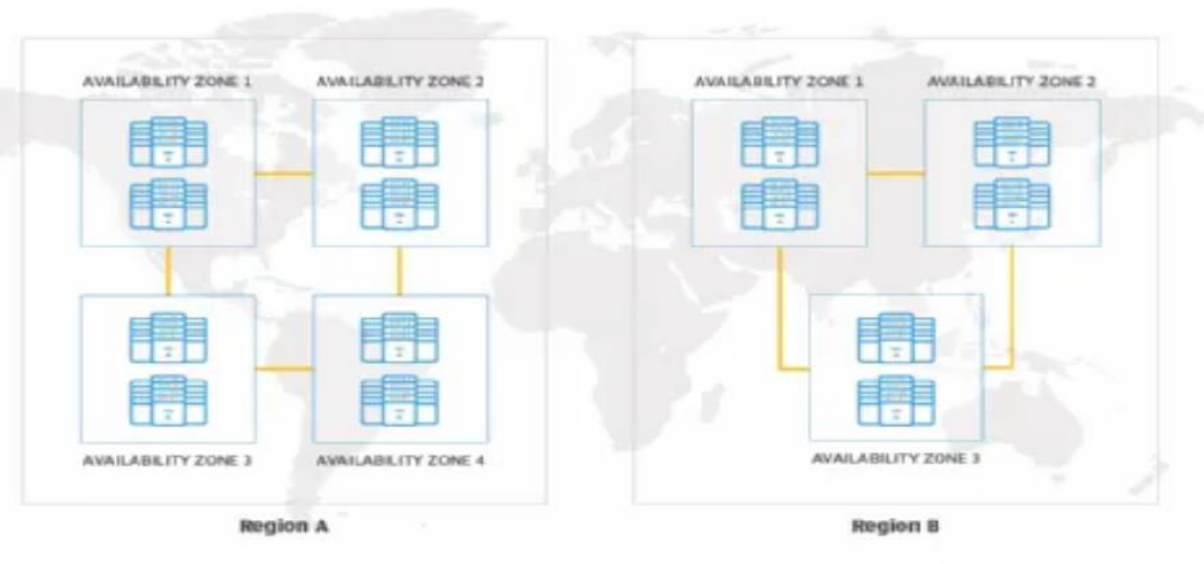

AWS Region
Availability zone 1
Data center
Availability zone 2
Data center
Availability zone 3
Data center

What are Availability Zones (AZs)?

An **Availability Zone** is a distinct physical location within a region. A region typically has multiple AZs, and each AZ consists of one or more data centers.

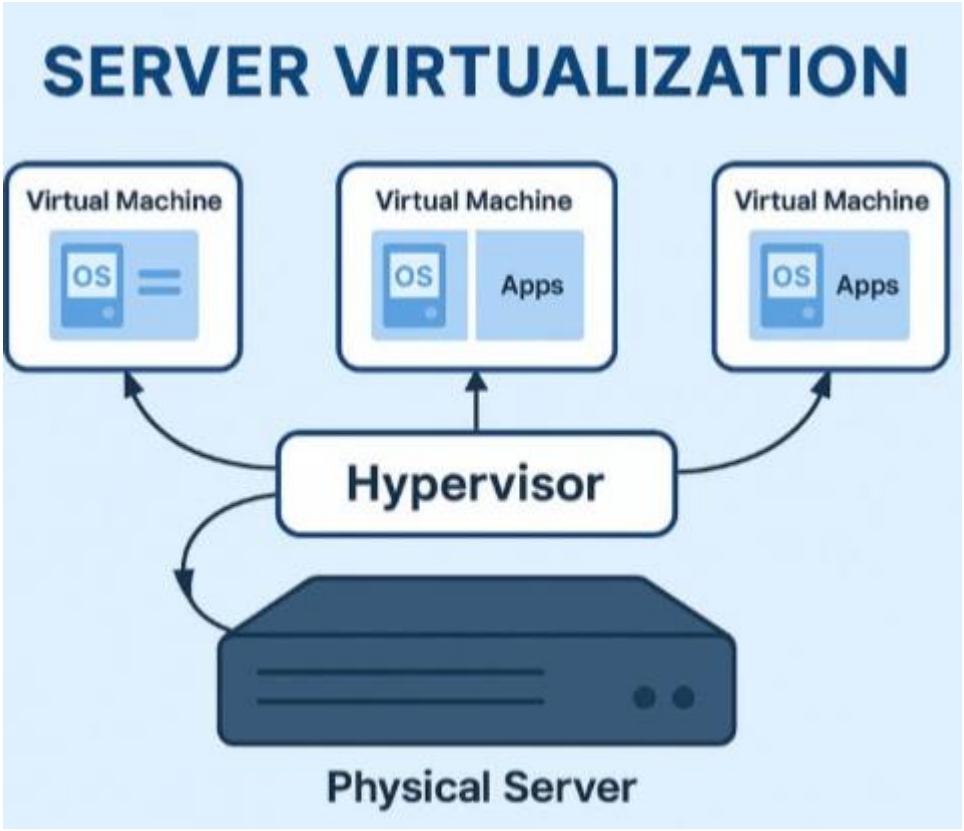
- **Fault Isolation:** AZs are isolated from each other to prevent one from being impacted by issues like power outages, network failures, or natural disasters. However, they are connected via low-latency networks.
- **Redundancy:** Since multiple AZs exist within a region, services can be deployed across AZs to ensure that if one AZ fails, others continue operating. This provides high availability.
- **Synchronization:** AZs in a region often share low-latency links, making it possible for applications to replicate data and failover quickly between AZs.

For example, the **US East (N. Virginia)** region has multiple AZs such as **us-east-1a**, **us-east-1b**, **us-east-1c**, etc.



How Virtualization Works

Virtualization is the **technology that allows multiple virtual machines (VMs) to run on a single physical machine** by sharing hardware resources efficiently.



A **hypervisor** is software that creates and manages **virtual machines**.

It sits between:

- **Hardware**
- **Virtual Machines**

Types of Hypervisors:

Type 1 (Bare-metal hypervisor)

Runs directly on hardware.

Examples:

- VMware ESXi
- Microsoft Hyper-V
- KVM

☞ Best for cloud platforms and production systems.

What is a Hypervisor?

A **Hypervisor** is a **software layer** that allows multiple virtual machines to share the same physical hardware.

It sits **between the hardware and the operating systems**.

Role of Hypervisor

- Allocates CPU, RAM, and storage
- Isolates virtual machines
- Starts, stops, and manages VMs
- Ensures security between VMs

Physical Hardware

↓

Hypervisor

↓ ↓ ↓

VM1 VM2 VM3