AWS EC2

EC2 : Elastic Cloud Computing

means this service can be scaled up or down

EC2 basically means AWS offers a service which provides the user with a virtual machine from its data center, which is generated in its physical machine using virtualization & hypervisors.

A VM has
- CPU
- RAM
- and everything else including OS

Why AWS EC2 ?

Say you are a devops engineer and your team requests you 1000 virtual machines. Now, how plausible would it be to create a 1000 VMs using hypervisor software, when the fact that

each VM needs regular maintenance is still there.

Its not about using AWS EC2, but is about using a public cloud computing service

- Timely upgrade
- Security issues
- Server

AWS is also pay-as-you-go, so you dont have to pay during night, christmas, diwali and so on.

This is why people are moving towards AWS EC2 / cloud virtualization service.

Types of EC2 instances

- General
- Compute optimized
- Memory
- Storage
- Accelerated
- High performance computing

① General EC2 instance : Balanced compute, memory and networking resources, suitable for a wide range of applications

- M5/M6g : General-purpose instances with a balance of compute, memory and networking
- T3/T4g : Burstable performance instances, ideal for workloads that don't need sustained high CPU performance
- A1 : Cost effective instances powered by ARM-based AWS graviton processors, suitable for workloads

(2) Compute Optimized: Applications requiring high compute power, such as high performance web-servers, scientific modeling and batch processing.

(3) Memory Optimized Instances: Applications with high memory requirements, such as DB, big data processing, and in-memory caching

(4) Storage Optimized Instances: high, sequential read and write access to a very large data sets on local storage.

(5) Accelerated Computing Instances: Applications that benifit from hardware accelerators such as GPUs or FPGAs including machine learning, gaming and video encoding.

AWS EC2 Regions

AWS EC2 is hosted in multiple locations worldwide. These locations are composed of AWS regions, availability zones, local zones, AWS outposts and wavelength zones.

- Each region is a separate geographic area
- Availibility zones are isolated locations and each region.
- local zones provide you the ability to place resources such as compute and storage, in multiple locations closer to your end users
- AWS outposts brings native AWS services, infrastructure, and operating models to virtually any data center, co-location orspace or on premise facility
- Wavelength zones allow developers to build applic that deliver ultra-low latencies to 5G devices and end users. Wavelength deploys AWS standard commute & storage services to edge of telecommunication's 5G network

Let's see a detailed explanation:

- **Region** Geographic locations across the world
  - > Total 24 Regions & 76 AZ
  - > You can choose any region based on the workload
- Regions are connected via AWS backbone network (via 100 gbps redundant trans-oceanic cables)
- Every region has a corresponding name
  - N. Virginia → us-east-1
  - Sydney → ap-southeast-2
  - mumbai → ap-south-1
- Why so many regions?
  - → low latency to applications
  - → data regulatory/compliance
  - → Disaster Recovery Site

Availability Zone

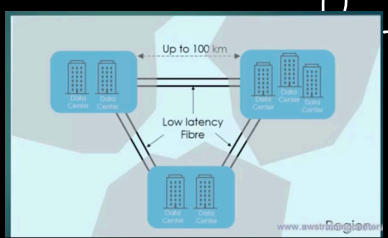
- Every region consists of two or more cluster of datacenters called "availability zones"
- Every AZ has a considerable code which

consists of region code followed by alphabets

Eg: mumbai has 3 AZ

- ap-south-1a
- ap-south-1b
- ap-south-1c

- Why AZ?



→ 1 region has at least 3 AZs

→ Each AZ is a cluster of data center

why?

→ different floodplains

→ Redundant power supply

→ Redundant network connectivity

Local Zones

Consider a situation where your organization hosts EC2 instance in North Virginia. Now, say a user tries to access your application, then, they may get 15ms of latency. Now, to make this faster, we can leverage a local zone near to my users in central america, and now the latency would decrease to 5ms. You cannot access all the services but can still access compute related services.

Note that any local zone is still connected to

- AZ.

So,

> Local zones are type of infrastructure deployment that places core services (compute, storage & database) and other selected AWS services close to large cities

> These are extension of parent region and are close to large population, industry & it centers

Outposts

These are on premise data center. If a company wants a data center within, then it can request for a outpost. This is hybrid cloud deployment model.

Wavelength Zone

Now we know all other AWS service zone operate with low latency fibres. But wavelength zones are a step ahead. They use existing telecommunication (CSP) service providers (or their data centres) for compute & storage. This reduces compute & storage latency very highly.

Finally, it is to the user to choose their region, AZ and if they want more they can choose their LZ or outpost or WZ.

AWS Regions

- Seperate geographic location
- contains multiple availability zones
- entire global distibution
- offer resilience and fault tolerance

Availibility Zones

- cluster of data centers 3 AZ min = 1 region
- isolated locations
- redundant power & disaster friendly
- low latency connections

Extension

Local Zone

- close to user-end
- child of AZ
- in big cities

Hybrid Extension

Outpost

- A small data center within company premises
- Hybrid cloud deployment model

Super extention

Wavelength zones

- Connect to AZ using 5G telecom network
- uses CSP's data centers