EC2 Services

EC2

What is EC2?

- Abbreviation of EC2 is “Elastic Compute Cloud”
- Amazon EC2 is a web service that provides resizable compute capacity in the cloud.
- It is a pay-per-use, scalable platform to launch virtual machines (mostly Linux and Windows OS)
- In simple words, EC2 is nothing but a virtual machines in cloud. See below diagram
- This virtual machines are called as “**Instances”** in AWS. .
- These Instances are build using **Amazon Machine Image**(AMI) Templates.

1) Virtual Machine = Instance (Power ON or Power OFF)

2) Amazon Machine Image = Machine Disk Template (likeDVD)
(we can do import and export AMI)

3) Elastic Block Storage (EBS volume) = Instance Hard Disk

Why it is Used?

- It is a revolutionary or Innovative way of spinning up VMs in minutes instead of waiting for weeks in traditional way.(Instant Provisioning)
- It supports auto-scaling means you can increase the set of VM or decrease the set of VM (Auto-scaling)
- It is designed to make web-scale computing easier for developers.(self-service)
- Low cost
- Resizable
- CLoudWatch monitoring etc.…
- We can put anywhere in world as per our requirement (Availability Zone)

Understanding EC2 Pricing Models

1) **On-Demand** Instance pricing (Pay/hour) Model : User wants low-cost or flexibility of AWS EC2 without paying any upfront cost and without long term commitment.

   T2.micro(N Virginia) cost is $0.0116 per Hour => 1.16 cents => 1.16*70 paise = **81.2 Paisa** <1 Rs

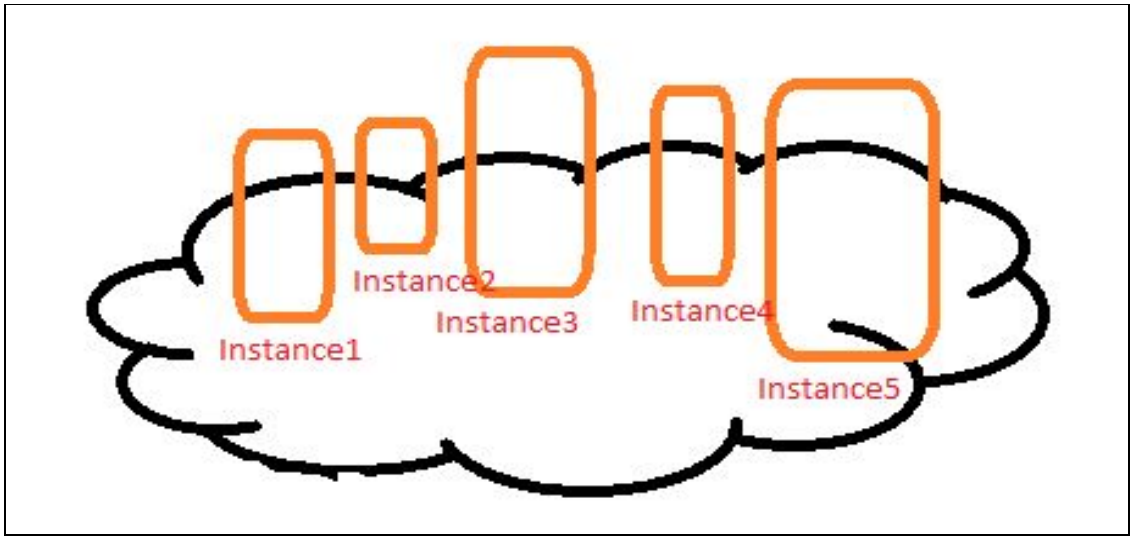
   T2.micro(Mumbai) cost is $0.0144 per Hour => 1.44 cents => 1.44*70 Paise

2) **Reserved** Instance - This type of instance you are blocking for 1 or 3 yr contract, It is very low -cost compared to On-demand Instances. It is used in the scenario where we know the usage of our applications and their needs(Light,Medium,Heavy).We have to give some upfront payment but the cost is very cheap because of that.

3) **Spot** Instance - We are bidding for this type of instance. It is used for the applications where start and end date is not a concern.It is very very cheap.

4) **Dedicated** Instance - Physical EC2 servers available as pay/hour basis. We can even use your existing server bound licenses.

Diagrammatic Representation:



AMI

What is AMI?

A pre-configured package (template) required to launch an EC2 instance. It includes an

1) Operating System
2) Software packages
3) Other required settings (root storage type and virtualization type)

Why it is used?

- AMI are used with Auto Scaling to quickly launch new servers on demand, and quickly recover from disaster.

AMI Categories

- You can create your own AMI also. Choose from a list of free AWS/ Community AMI's or choose from AWS marketplace.

1) Community AMI
2) MarketPlace AMI
3) My AMIs

EC2 INSTANCE TYPE

What is EC2 Instance Type?

- Instance types describes the “hardware” components that an EC2 instance will contain

1) Compute power (processor/vCPU)
2) Memory (ram)
3) Storage Options/ Optimisation (Hard Drive)
4) Network Performance (Bandwidth)

- As an Cloud Architect, it's important to use the proper instance type to handle your application’s workload

<u>Types of EC2 Instances (Flavor of Instances)</u>

Amazon EC2 instances are grouped into 5 families

1) General Purpose
2) Compute Optimized
3) Memory Optimized
4) Storage Optimized
5) GPU Optimized

General Purpose instances have memory to CPU ratios suitable for most general purpose applications and comes with fixed performance (M4 and M3 instances) or burstable performance (T2)

Compute Optimized instances (C4 and C3 instances have proportionally more CPU resources than memory (RAM) and are well suited for scale out compute-intensive applications and High Performance Computing (HPC) workloads.

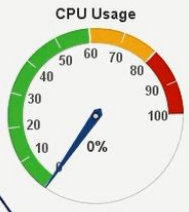
Memory Optimized instances (R3 and R4 instances) offer larger memory sizes for memory-intensive applications

There is a collection of pre-configured instance types that are grouped into families and types that you can choose from:

Type	Family	Notes
t2	General Purpose	“Burstable” Performance instances
M3	General Purpose	Nice balancing
c3/c4	Compute Optimized	For high traffic front end fleets, web servers
d2	Storage Optimized	For large-scale data warehouse or Parallel File systems
i2	Storage Optimized	For large-scale data warehouse or parallel File Systems
g2	GPU Optimized	For machine learning, high performance databases, rendering
p2	GPU Optimized	For machine learning, high performance databases, rendering
r3/r4	Memory Optimized	For databases, memcached, large deployments of enterprise applications
x1	Memory Optimized	For databases, memcached, large deployments of enterprise applications

EC2 Instance Types and Performance

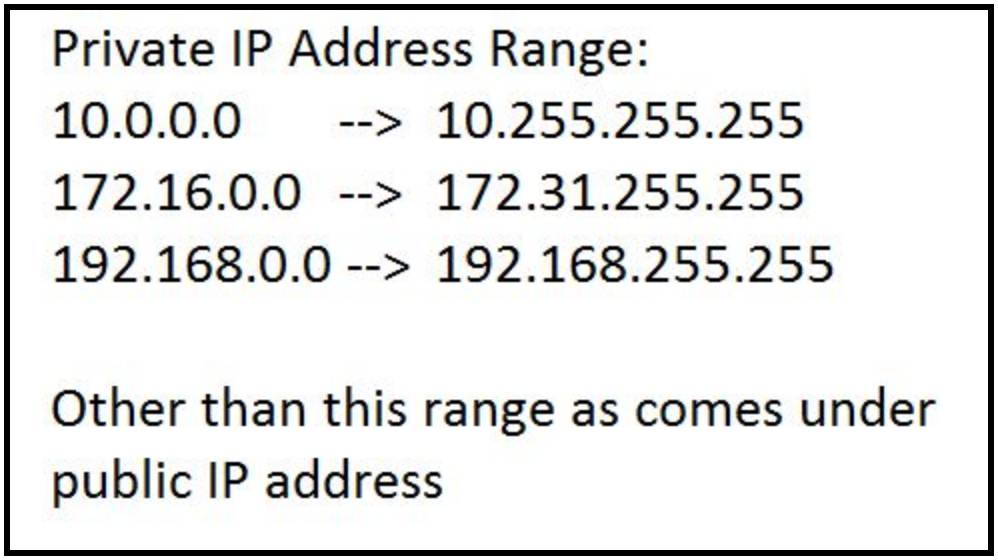
- It all starts from an AMI and becomes...
  - t2 - General Purpose, Burstable: Micro, Small, Medium
  - m3 - General Purpose, Fixed: Medium, Large, xlarge, 2xlarge
  - c3 - CPU Intensive: Large, xlarge, 2xlarge, 4xlarge, 8xlarge
  - g2 - GPU Intensive: 2xlarge
  - r3 - Memory Intensive: Large, xlarge, 2xlarge, 4xlarge, 8xlarge
  - i2 - Storage Optimized: Large, xlarge, 2xlarge, 4xlarge, 8xlarge



EC2 IP ADDRESS

Private IP Address:

- All EC2 instances are automatically created with a PRIVATE IP address.
- The private IP address is used for internal (inside the VPC)communication between instances.



Public IP Address:

- When creating an EC2 instance, you have the option to enable (or auto-assign) a public IP address
- A public IP address is required if you want the EC2 instance to have **direct** communication with resources across the open internet
  i.e. if you want to directly sshed with into the instance or have it directly server web traffic
- Auto-assigning is based on the setting for the selected subnet that you are provisioning the instance in.

Elastic IP Address:

- An EIP is a static IPv4 address designed for dynamic cloud computing
- An EIP is a public IPv4 address
- With an EIP, you can attach a public IP address to an EC2 instance that was created with only a private IP address OR
- You can mask the failure if an instance or software by rapidly remapping the address to another instance in your account (i.e detaching the EIP from one instance and attaching it to another)
- Attaching an EIP to an instance will replace its default public IP address for as long as it is attached.

BOOTSTRAPPING

Bootstrapping:

- Bootstrapping means **self-starting process** or means set of commands without external input
- With EC2, we can bootstrap the instance (during the creation process) with custom commands (such as installing software packages, running updates, and configuring other various settings)

User-Data:

- It is step or section during the EC2 instance creation process where you can include your own custom commands via a script (i.e bash script)
- Here is an example of bash script that will automate the process of updating the yum package installer, install Apache Web Server, and start the Apache service.

```
#!/bin/bash
yum update -y
yum install httpd -y
service httpd start
chkconfig httpd on
echo "This is Instance 1 " > /var/www/html/index.html
```

To view User-Data & Instance Metadata:
curl http://169.254.169.254/latest/user-data (displays bootstrapping commands)
curl http://169.254.169.254/latest/meta-data (displays AMI, instance type etc)

EBS (Elastic Block Storage)

EBS Basics

EBS Volumes are **persistent** (they can live beyond the life of EC2 instance they are attached to) - Its opposite to another type of storage called as instance store volume

- EBS backed volumes are network attached storage meaning they can be attached/detached to or from various EC2 instances.
- However, EBS volumes can only be attached to ONE EC2 instance at a time.
- We can also take a backup of these volumes, and these backups are called as **snapshots**

EBS Types

General Purpose SSD:

- Use for dev/test environments and smaller DB instances
- Performance of 3 IOPS/GiB of storage ie (burstable with baseline performance)
- Volume size of 1GiB to 16 TiB
- Considerations when using T2 instances with SSD root volumes (burstable vs Baseline performance)

Provisioned IOPS SSD:

- Used for mission critical applications that require sustained IOPS performance
- Large database workloads
- Volume size of 4GiB to 16 TiB

- Performs at provisioned level and can provision up to 20,000 IOPS per volume

Magnetic:

- Low storage cost
- Used for workloads where performance is not important or data is infrequently accessed
- Volume size of min 1GiB and Max is 1TiB

Security Group

- Itis nothing but a virtual firewall for an instance

EC2 Key Pairs

Key Pair Basics

- It is used for login authentications
- EC2 key pairs are two cryptographically secure keys that are used by AWS to authenticate a client when logging into and EC2 instance
- Each key pair consists of a public key and private key
- AWS stores the public key on the instance and you are responsible for storing the private key.

To login into an EC2 instance, you must create and authenticate with a key pair

For Linux instance: You don't need a password and you use a key pair and then using ssh
For Windows Instance: you use a key pair to obtain the Administrator password and then login using RDP

EBS SNAPSHOT

What is snapshot?
Snapshots are point-in-time backups of EBS volumes that can be stored in S3

Snapshot properties:
- Snapshots are incremental in nature
- A snapshots only stores the changes since the most recent snapshot, thus reducing costs
- However, if the “original” snapshots is deleted, all data is still available in all the other snapshots.
- Snapshots can be used to create fully restored EBS volumes.

NOTE: During snapshot process against the EBS volume, it can degrades performance so snapshot should occur during non-production or non-peak load hours.

EC2 Placement Group

What is Placement Group?
- A Placement Group is a cluster of instances within the same AZ
- It is used by the applications that require an extremely low latency network between them
- Instances within a placement group have a low-latency,i.e 10 Gbps network connection between them
- AWS attempts to place all the instances as close as physically possible in the data center to reduce latency.

Troubleshooting Placement Groups
- If an instance in a placement group is stopped, once it is started again it will continue to be a member of the placement group
- It is suggested to launch all the required instances within a placement group in a single request, and that the same instance type is used for all the instances within a placement group
- It is possible, if more instances are added at a later time to the placement group OR if a placement group instance is stopped and started again, to receive an “insufficient capacity error”
    - Resolve the capacity error by stopping all instances in the member group and attempting to start them again.

Facts about Placement Group
- Instances not originally launched/created in the placement group cannot be moved into the placement group
- Placement group cannot be merged together

- A Placement Group cannot span multiple Availability one
- Placement group names must be unique within your own AWS account
- Placement group can be connected
- Instances must have 10 gigabit network speeds in order to take advantage of placement groups (proper instance type)

Elastic File System

What is EFS?

- EFS is a storage which connects to mostly EC2 instances which allows easy scalability so it is called as “scalable storage”
- EFS storage capacity is “elastic” in nature
    - Elastic means the storage capacity will increase and decrease as you add or remove files
    - Applications running on an EC2 instance using EFS will always have the storage they need, without having to provision and attach larger storage devices

  (suppose if you use EBS volume, it has fixed size eg 20GB and later you need to increase the size of File system , then u need to take snapshot of that EBS volume and take a larger EBS volume and the restore from the snapshot)
- EFS is fully managed so there is no maintenance required
- Supports the Network File System version 4.0 and 4.1 (NFSv4) protocols when mounting
- Best performance when using an EC2 AMI with linux kernel 4.0 or newer

Benefits of EFS

- The EFS file system can be accessed by one or more EC2 instances at the same time.
    - Shared file access across all your EC2 instances
    - Application that span multiple EC2 instances can access the same data
- EFS can be mounted to on-premise servers (when connected to your VPC Via AWS Direct Connect)
    - This allows you to migrate data from on-premise servers to EFS and/or use it as a backup solution

- EFS can scale to petabytes in size, while maintaining low-latency and high level of throughput
- You pay only for the amount of storage you are using

Security:

- Control file system access through POSIX permissions
- VPC for network access control, and IAM for API access control
- Encryption data at rest using AWS key Management Service (KMS)

When to use?

- BIg data and Analytics
- Media processing workflows
- Web servicing and content management

00000000000000000