Elastic Compute Cloud – EC2

Ponnam Phani Krishna
PONNAM.PHANI@GMAIL.COM

Elastic Compute Cloud – EC2

- EC2 is a web service which provides secure, resizable compute capacity in the cloud
- EC2 interface allows you to obtain & configure capacity with minimal friction
- EC2 offers the broadest and deepest compute platform with choice of processor, storage, networking, operating system, and purchase model.
- Amazon offer the fastest processor in the cloud and they are the only cloud with 400 Gbps ethernet networking
- Amazon have the most powerful GPU instances for machine learning and graphic workloads.

Reliable, Scalable, Infrastructure on Demand:

- Increase or decrease capacity with in minutes, not hours or days
- SLA commitment of 99.99% availability for each amazon EC2 region. Each region consists of atleast 3 availability zones
- Region/AZ model is recognized by gartner as the recommended approach for running enterprise applications that require high availability.

AWS Supports 89 security standards & compliance certifications including:

1. PCI-DSS
2. HIPAA/HITECH
3. FedRAMP
4. GDPR
5. FIPS
6. NIST etc..

Features of Amazon EC2:

- Virtual Computing instances, known as instances
- Pre configured templates for your instances, Known as AMI, which contains Operating System, Configuration and softwares.
- Various configurations of CPU, Memory, Storage, Networking Capacity for your instances, known as instance types
- Secure login information for your instances using keypairs (AWS Stores Public key, and you store the private key)
- Storage volumes for temporary data that's deleted when you stop, hibernate or terminate your instances, known as Instance store volumes
- Persistent storage volumes for your data using elastic block storage, known as amazon EBS Volumes
- Multiple physical locations for your resources, such as instance & EBS volumes, known as regions and availability zones.
- A firewall that enables you to specify the protocols, ports and source IP ranges that can reach your instance using security groups

- Static IPV4 addresses for dynamic cloud computing known as Elastic IP Address

Amazon EC2 Provides the following purchasing options:

1. On-Demand
2. Spot Instances
3. Reserved Instances
4. Savings Plan

On-Demand Instances:

- You pay for compute capacity by the hour or the second depending on which instances you run
- No long-term commitment or upfront payments are needed
- You can increase or decrease your compute capacity depending on the demands of your application and only pay the specified per hourly rates for the instance you use.

On Demand Instances are Recommended for

- Users that prefer the low cost and flexibility of amazon EC2 without any upfront payment or longterm commitment
- Applications with short term, spiky or unpredictable workloads that cannot be interrupted.
- Applications being developed or tested on amazon EC2 for the first time

Spot Instances:

- Amazon EC2 spot instances allow you to request spare amazon EC2 computing capacity for upto 90% of on-demand price

Spot instances are recommended for

- Applications that have flexible start and end times
- Applications that are only feasible at very low compute prices
- No guarantee for 24x7 uptime

Reserved Instances:

- Reserved Instances provide you with a significant discount (upto 75%) compared to on-demand instance pricing
- For applications that have steady state or predictable usage, reserved instance can provide significant savings compared to using on-demand instances

Recommended for:

- Applications with steady state usage
- Applications that may require reserved capacity
- Customers that can commit to using EC2 over a 1 or 3 year term to reduce their total computing costs

Savings Plan:

Savings plans are a flexible pricing model that offer low prices on EC2 in exchange for a commitment to a consistent amount usage for a 1 or 3 year term. Discount upto 72%

Amazon EC2 Instance Types:

Amazon EC2 provides a wide selection of instance types optimized to fit different use cases. Instance types comprise varying combinations of CPU, memory, storage, and networking capacity Each instance type includes one or more instance sizes, allowing you to scale your resources to the requirements of your target workload.

1. General Purpose
2. Compute Optimized
3. Accelerated Computing (GPU Optimized)
4. Memory Optimized
5. Storage Optimized

General Purpose:

General purpose instances provide a balance of compute, memory and networking resources, and can be used for a variety of diverse workloads. These instances are ideal for applications that use these resources in equal proportions such as web servers and code repositories.

Ex: Mac, T4g, T3, T3a, T2, M6g, M5, M5a, M5n, M5zn, M4, A1

Compute Optimized:

Compute Optimized instances are ideal for compute bound applications that benefit from high performance processors. Instances belonging to this family are well suited for batch processing workloads, media transcoding, high performance web servers, high performance computing

Ex: C6g, C6gn, C5, C5a, C5n, C4

Memory Optimized:

Memory optimized instances are designed to deliver fast performance for workloads that process large data sets in memory.

Use case: Memory-intensive applications such as open-source databases, in-memory caches, and real time big data analytics

Ex: R6g, R5, R5a, R5b, R5n, R4, X2gd, X1e, X1, u, Z1d

Accelerated Computing:

Accelerated computing instances use hardware accelerators, or co-processors, to perform functions, such as floating point number calculations, graphics processing, or data pattern matching, more efficiently than is possible in software running on CPUs.

UseCase: Machine learning, high performance computing, computational fluid dynamics, computational finance, seismic analysis, speech recognition, autonomous vehicles, and drug discovery.

Ex: P4, P3, P2, Inf1, G4dn, G3, F1

Storage Optimized:

Storage optimized instances are designed for workloads that require high, sequential read and write access to very large data sets on local storage. They are optimized to deliver tens of thousands of low-latency, random I/O operations per second (IOPS) to applications.

Ex: I3, I3en, D2, D3, D3en, H1

Instance Features:

Amazon EC2 instances provide a number of additional features to help you deploy, manage, and scale your applications.

- Burstable Performance instances
- Multiple Storage Options
- EBS Optimized Instances
- Cluster Networking

Burstable Performance Instances: Amazon EC2 allows you to choose between Fixed Performance Instances (e.g. M5, C5, and R5) and Burstable Performance Instances (e.g. T3). Burstable Performance Instances provide a baseline level of CPU performance with the ability to burst above the baseline.

For example, a t2.small instance receives credits continuously at a rate of 12 CPU Credits per hour. This capability provides baseline performance equivalent to 20% of a CPU core (20% x 60 mins = 12 mins). If the instance does not use the credits it receives, they are stored in its CPU Credit balance up to a maximum of 288 CPU Credits. When the t2.small instance needs to burst to more than 20% of a core, it draws from its CPU Credit balance to handle this surge automatically.

Multiple Storage Options: Amazon EC2 allows you to choose between multiple storage options based on your requirements. Amazon EBS is a durable, block-level storage volume that you can attach to a single, running Amazon EC2 instance.

Amazon EBS provides three volume types to best meet the needs of your workloads: General Purpose (SSD), Provisioned IOPS (SSD), and Magnetic.

EBS Optimized instances: For an additional, low, hourly fee, customers can launch selected Amazon EC2 instances types as EBS-optimized instances. For M6g, M5, M4, C6g, C5, C4, R6g, P3, P2, G3, and D2 instances, this feature is enabled by default at no additional cost. EBS-optimized instances enable EC2 instances to fully use the IOPS provisioned on an EBS volume.

EBS-optimized instances deliver dedicated throughput between Amazon EC2 and Amazon EBS, with options between 500 and 4,000 Megabits per second (Mbps) depending on the instance type used.

Cluster Networking: Select EC2 instances support cluster networking when launched into a common cluster placement group. A cluster placement group provides low-latency networking between all instances in the cluster. The bandwidth an EC2 instance can utilize depends on the instance type and its networking performance specification.

EC2 Tenancy Model:

AWS offers 3 different types of tenancy model for your EC2 instances. this relates to what underlying host your EC2 instance will reside on

- Shared Tenancy
- Dedicated Instance
- Dedicated Host

Shared Tenancy : this option will launch your EC2 instance on any available host with the specified resources required for your selected instance type. Regardless of which other customers and users also have EC2 instances running on the same host. Means We are going to share the physical resources with other customers.

AWS implement advanced security mechanisms to prevent one EC2 instance from accessing another on the same host.

Dedicated Instance: Dedicated instances are hosted on hardware that no other customer can access. It can only be accessed by your own AWS account. You may be required to launch your instances as a dedicated instance due to internal security policies or external compliance controls.

Dedicated instances do incur additional charges due to the fact you are preventing other customers from running EC2 instances on the same hardware and so there will likely be unused capacity remaining.

Dedicated Host: A dedicated host is effectively the same as dedicated instances. However they offer additional visibility and control, how you can place your instances on the physical host. They also allow you to use your existing licenses, such as PA-VM license or Windows Server licenses Etc. Using dedicated hosts give you the ability to use the same host for a number of instances that you want to launch and align with any compliance and regulatory requirements.

Following are the list of important terms need to know before creating EC2 instances:

- Amazon Machine Image (AMI)
- Instance Type
- Network
- Subnet
- Public IP
- Elastic IP
- Private IP
- Placement Group
- Root Volume
- Security Group
- KeyPair

Amazon Machine Image: An Amazon Machine Image (AMI) Provides the information required to launch an instance. An AMI Includes the following, one or more Elastic Block Store snapshot, a template for the root volume of the instance (for example Operating system, software, configurations etc.)

Instance Type: Instance types comprise varying combinations of CPU, Memory, storage & Networking capacity and give you the flexibility to choose the appropriate mix of resources for your applications.

Subnet: Subnet is a subnetwork in your virtual network of your Amazon Network. By default there is one subnet per availability zone.

Public IP: A public IP is an IP Address which can be used to access internet and allow the communication over the internet. Public IP will be assigned by amazon and it is dynamic. If you stop and start your EC2 instance, The public IP will change.

Elastic IP(EIP): Elastic IP is a kind of Fixed Public IP address which we can attach to our Instances. Elastic IP will not change if we stop & Start our EC2 instances. We need to request EIP from amazon and it will be free if we attach to any instances, if you keep this EIP unused in your account then it will be charged after initial 1st hour.

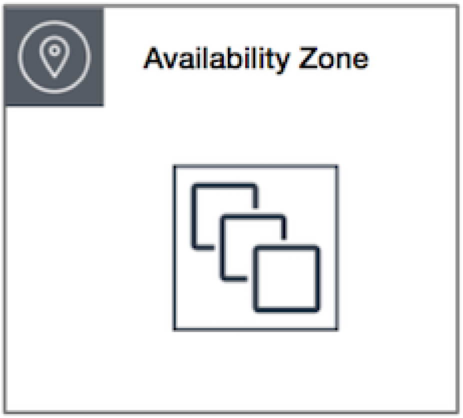
Private IP: Private IP can be used to establish the communication with in the same network only, Private (internal) addresses are not routed on the Internet and no traffic can be sent to them from the Internet, means no internet access will be available over private address.

Placement group: is a logical grouping of instances with in a single availability zone. AWS provides three types of placement groups

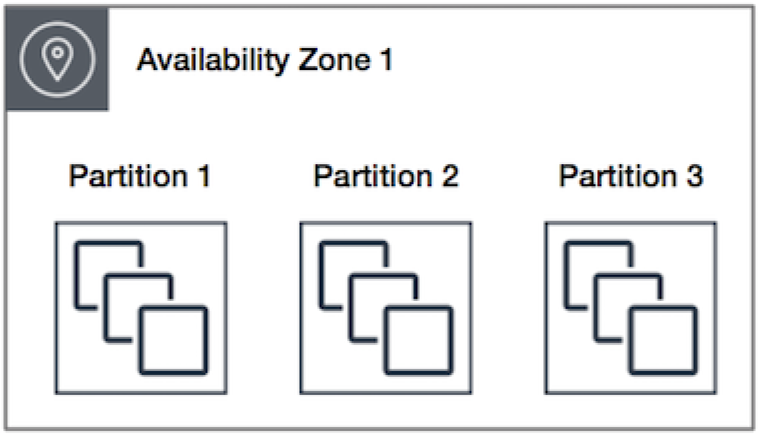
- Cluster
- Partition
- Spread

Cluster: A cluster placement group is a logical grouping of instances within a single Availability Zone. Instances in the same cluster placement group enjoy a higher per-flow throughput limit for TCP/IP traffic and are placed in the same high-bisection bandwidth segment of the network.

The following image shows instances that are placed into a cluster placement group.

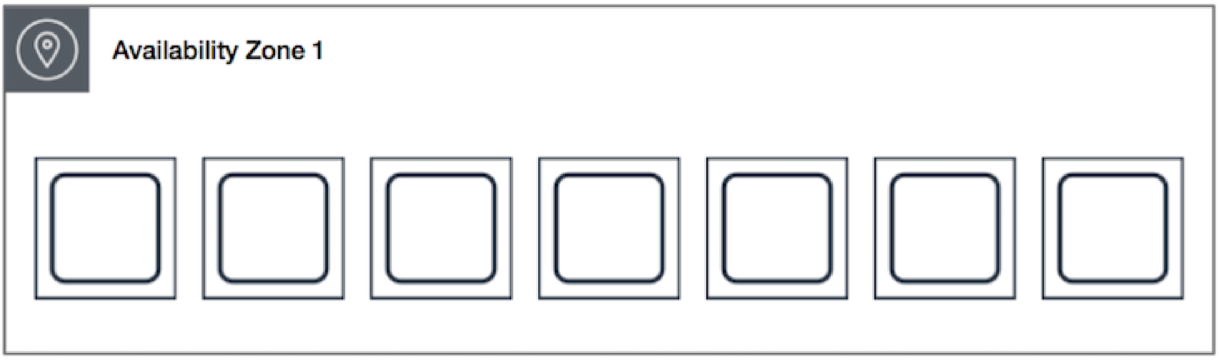


Partition Placement Group: Partition placement groups help reduce the likelihood of correlated hardware failures for your application. When using partition placement groups, Amazon EC2 divides each group into logical segments called partitions. Amazon EC2 ensures that each partition within a placement group has its own set of racks. Each rack has its own network and power source. No two partitions within a placement group share the same racks, allowing you to isolate the impact of hardware failure within your application.



Spread: A spread placement group is a group of instances that are each placed on distinct racks, with each rack having its own network and power source.

The following image shows seven instances in a single Availability Zone that are placed into a spread placement group. The seven instances are placed on seven different racks



Spread placement groups are recommended for applications that have a small number of critical instances that should be kept separate from each other.

A spread placement group can span multiple Availability Zones in the same Region. You can have a maximum of seven running instances per Availability Zone per group.

Root Volume: The storage which we used to install Operating system for instance is called as root volume (Ex: C:\ Drive). The following volume types are supported as root volumes: General purpose SSD, Provisioned IOPS SSD, Magnetic.

Security Group: A Security group acts as a virtual firewall for your instance to control incoming & Outgoing traffic. Security groups to be attached and we can attach 5 security groups to each instance.

Following are the basic characteristics of a security groups:

- You can specify allow rules, but not deny rules
- You can specify separate rules for inbound and outbound traffic.
- Security group rules enable you to filter traffic based on protocols and port numbers.
- Security groups are stateful — if you send a request from your instance, the response traffic for that request is allowed to flow in regardless of inbound security group rules.
- When you create a new security group, it has no inbound rules.
- By default, a security group includes an outbound rule that allows all outbound traffic.
- By default we can create 2500 security groups per region, can be increased upto 5000 per region
- 60 inbound and outbound rules per security group

KeyPair: Key pair is a combination of public key and private key which can be used to encrypt and decrypt the data, is a set of security credentials that you use to prove your identity when connecting to an instance. Amazon EC2 stores the public key and user stores the private key.

Elastic Compute Cloud (EC2) Lab:

Prerequisite:

- Amazon Account access
- Putty & puttygen tools on your computer

ToDo List 1:

1. Launch Windows Server 2016 EC2 instance in N.Virginia Region
    a. While creating EC2 instance open RDP port in the security group from your IP only
    b. Decrypt the password using keystore file which we used while creating the EC2 instance
    c. Access Windows server using Remote desktop connection tool
2. Launch Amazon Linux2 EC2 instance in N.Virginia Region
    a. While creating EC2 instance open SSH port in the security group from anywhere
    b. Generate Private key using puttygen tool from the keypair which we used while creating the EC2 instance
    c. Access Amazon Linux EC2 instance using putty tool
3. Install webserver on Amazon Linux2 EC2 and host a website
4. Create Custom AMI from the Amazon Linux EC2 in which we hosted the website.
5. Launch New EC2 instance from the custom AMI in N.Virginia Region
6. Copy the AMI from N.Virginia to Mumbai Region
7. Launch New Instance from Custom AMI in Mumbai Region
8. Share custom AMI with a specific Amazon Account & launch New EC2 instance in the other amazon account
9. Share custom AMI with public