ELASTIC COMPUTE CLOUD (EC2)

- ➤ Amazon EC2 provides scalable computing capacity in the AWS cloud.
- ➤ You cans use Amazon EC2 to lunch as many or as few virtual servers as you need, configure security and networking and manage storage.
- ➤ Amazon ec2 enables you to scale up or scale down the instance.
- ➤ Amazon ec2 is having two storage options i.e EBS and instance store.
- ➤ Preconfigured templates are available known as amazon machine image.
- ➤ By default when you create an ec2 account with amazon your account is limited to a maximum of 20 instances per ec2 region with two default high I/O instances.

Types of EC2 instances:

1. General purpose
2. Compute optimized
3. Memory optimized
4. Storage optimized
5. Accelerated computing or GPU
6. Storage optimized
7. High memory optimized

1. General purpose
   General purpose instances provide a balance of compute, memory and networking resources and can be used for a variety of workloads.
   There are 3 series are available in general purpose instance:
   a. A series: A1
   b. M series: M4, M5, M5a, M5d, M5ad (large)
   c. T series: T2 (free tier eligible), T3, T3a
      Instances are available in four sizes: Nano, Small, Medium, Large

   a. A series: A1-instances:
      A1 instances are ideally suited for scale out workloads that are supported by the ARM Ecosystem.
      ARM ecosystem of software provides customers a wide range of products to get to market faster than the competition. ARM development boards are the ideal platform for accelerating development and reducing the risk of new Soc design. Micro service is a distinct method to developing software systems that tries to focus on building single function modules with well-defined interfaces and operations. Cache is a high speed data storage layer which stores a subset of data typically transient in nature, so that future request fir that data are served up faster.
      These instances are well suited for the following applications:
      1. Web server
      2. Containerized micro services

3. Caching fleets
4. Distributed data stores
5. Application that requires ARM instruction set

b. M series: M4, M5, M5a, M5d, M5ad

M4 instance:
The new M4 instances features a custom Intel Xeon E5-2676 v3 Haswell processor optimized specifically for EC2.

vCPU- 2 to 40 (max)
RAM- 8GB to 160GB(max)

Instance storage: EBS only (root volume storage)

M5, M5a, M5d and M5ad instances:
These instances provide an ideal cloud infra, offering a balance of compute, memory and networking resources for a broad range of applications.

Used in : gaming server, webserver, small and medium database
vCPU- 2 to 96(max)
RAM- 8 to 384(max)
Instance storage- EBS and NVMe SSD

c. T series: T2, T3, T3a instances:
These instances provide a baseline level of CPU performance with the ability to burst to a higher level when required by your workload. An unlimited instances can sustain high CPU performance for any period of time whenever required.

vCPU- 2 to 8
RAM- 0.5 to 32 GB

Used for:
i.   Website and web app
ii.   Code repositories
iii.   Development, build, test
iv.   Micro services

2. Compute optimized
Compute optimized are ideal for compute bound applications that benefits from high performance processors.

C Series:
Three types are available: C4, C5, C5n [C3- previous instance]

C4:

C4 instances are optimized for compute intensive workloads and deliver very cost effective high performance at a low price per complete ratio.

vCPU- 2 to 36
RAM- 3.75 to 60GB
Storage- EBS only
Network BW- 10 Gbps

Usecase: web server, batch processing, MMO gaming, Video encoding

Note:
C5 support max 25 EBS volumes
C5 use Elastic Network Adaptor
C5 uses new EC2 Hypervisor

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

   There are 3 series are available:
   R series, X series, Z series

   A. R Series:
      R4, R5, R5a, R5ad, R5ad
      ➢ High performance, relational (MySQL) and NoSQL (MongoDB, Casssandra) databases.
      ➢ Distributed web scale cache stores that provide in memory caching of key volume type data.
      ➢ Used in financial services, Hadoop
         vCPU- 2 to 96
         RAM- 16 768GB
         Instance storage- EBS only and NVMe SSD

   B. X Series
      X1, X1e instances:
      ➢ Well suited for high performance database, memory intensive enterprise application, relational database workload, SAP HANA.
      ➢ Electronic design automation
         vCPU- 4 to 128
         RAM- 122 to 3904GB
         Instance storage- SSD

C. Z1d instance:
  ➢ High frequency Z1d delivers a sustained all core frequency of up to 4.0 GHz, the fastest of any cloud instances.
  ➢ AWS Nitro System, Xeon processor, up to 1.8 TB of instances storage.
    vCPU- 2 to 48
    RAM- 16 to 384 GB
    Storage- NVM SSD

    Use case: electronic design automation and certain database workloads with high per-core licensing cost.

4. 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 application.

   It is of three types:
   A. D series- D2 instance
   B. H series- H1 instance
   C. I series- I3 and I3en instance

A. D2 instance:
  ➢ Massive parallel processing (MPP) data warehouse.
  ➢ Map reduce and Hadoop distributed computing.
  ➢ Log or data processing app
    vCPU- 4 to 36
    RAM- 30.5 to 244GB
    Storage- SSD

B. H series- H1 instance:
  ➢ This family features up to 16GB of HDD based local storage, high disk throughput and balance of compute and memory.
  ➢ Well suited for app requiring sequential access to large amounts of data on direct attached instance storage.
  ➢ Application that requires high throughput access to large quantities of data.
    vCPU- 8 to 64
    RAM- 32 to 256GB
    Storage- HDD

C. I3 and I3en instances:
   - High frequency online transaction processing system (OLTP)
   - Relational databases:
     NoSQL database
     Distributed file system
     Data warehousing application

   vCPU- 2 to 96
   RAM- 16 to 768GB
   Local storage- NVMe SSD
   Networking performance- 25 Gbps to 100 Gbps
   Sequential throughput:

    Read- 16GBps
    Write- 6.4 GBps (I3)
     8GBps (I3en)

5. Accelerated Computing Instances

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

   It is of 3 types:
   A. F series- F1 instance
   B. P series- P2 and P3 instance
   C. G series- G2 and G3 instance

a. F1 instance:
   - F1 instances offers customizable hardware acceleration with field programmable gate arrays.(FPGA)
   - Each FPGA contains 2.5 million logic elements and 6800 DSP (Digital Processing Unit) engines.
   - Designed to accelerate computationally intensive algorithms such as data flow or highly parallel operations.
   - F1 provides local NVM SSD storage.

   vCPU- 8 to 64
   FPGA- 1 to 8
   RAM- 122 to 976GB
   Storage- NVMe SSd

Used in- genomics research, financial analytics, real time video processing and big data search.

b. P2 and P3 Instance:
➢ It uses NVIDIA Tesla GPUs.
➢ Provide high bandwidth networking.
➢ Up to 32GB of memory per GPUs which makes them ideal for deep learning and computational fluid dynamics.

P2 instance:	P3 instance
vCPU- 4 to 64	vCPU- 8 to 96
GPU- 1 to 16	GPU- 1 to 8
RAM- 61 to 731GB	RAM- 61 to 768GB
GPU RAM- 12 to 192 GB	storage- SSD and EBS
Network bandwidth- 25 GBps	

Used in- machine learning, databases, seismic analysis, genomics, molecular modeling, AI, deep learning

Note: P3 support CUDA9 and OPENCL APIs,
        P2 supports CUDA8 and OPENCL 1.2

c. G2 and G3 instances:
➢ Optimized for graphics intensive application.
➢ Well suited for app like 3D visualization.
➢ G3 instances use NVIDIA Tesla M60 GPU and provide a cost effective, high performance platform for graphics applications.

vCPU- 4 to 64
GPU- 1 to 4
RAM- 30.5 to 488GB
GPU memory- 8 to 32 GB
Network performance- 25GBps

Used in: video creation service, 3D visualization, streaming, graphic intensive application

6. High Memory Instance:
➢ High memory instances are purpose built to run large-in-memory databases, including production developments of SAP HANA in the cloud.
➢ It has only on series i.e U series.

Features:
➢ Latest generation intel Xeon Pentium 8176M processor.
➢ 6, 9, 12 TB of instance memory, the largest of any EC2 instance.

- Powered by the AWS Nitro System, a combination of dedicated hardware and light weight hypervisor.
- Bare metal performance with direct access to host hardware.
- EBs optimized by default at no additional cost.
- Model number- U-6tb1.metal, U-9tbi.metal, U-12tb1.metal
- Network performance- 25 GBps
- Dedicated bandwidth- 14GBps
- Each instance offer 448 logical processor

Note:

- High memory instances are bare metal instances and do not run on a hypervisor.
- Only available under dedicated host purchasing category. (for 3 years term)
- O.S directly on Hardware.

7. Previous generation instance:
   T1, M1, C1, CC2, M2, CR1, CG1, I2, HS1, M3, C3, R3: all instances are available now and we can purchase all of them.

EC2 PURCHASING OPTION:

There are 6 ways of purchasing options available for AWS EC2 instances, but there are 3 ways to pay for Amazon EC2 instance i.e On demand, Reserved instance and Spot instance.

You can also pay for dedicated host which provide you with EC2 instance capacity on physical servers dedicated for your use.

1. On demad
2. Dedicated instance
3. Dedicated Host
4. Spot instance
5. Scheduled instance
6. Reserved instance

1. On-Demand Instance:
   - AWS on demand instances are virtual servers that run in AWS of AWS relational database service (RDS) and are purchased at a fixed rate per hour.
   - AWS recommends using on demand instances for applications with short term irregular workloads that cannot be interrupted.
   - They also suitable for use during testing and development of applications on EC2.
   - With on demand instances you only pay for EC2 instances you use.

- The use of on demand instances frees you from the cost and complexities of planning, purchasing, and maintaining hardware and transforms what are commonly large fixed costs into mush smaller variable cost.
- Pricing is per instance hour consumed for each instance from the time an instance is launched until if it terminated of stopped.
- Each partial instance hour consumed will be billed per second for linux instances and as a full hour for all other instance types.

2. Dedicated Instance:
   - Dedicated instances are run in a VPC on hardware that is dedicated to a single customer.
   - Your dedicated instances are physically isolated at the host hardware level from instances that belong to other AWS account.
   - Dedicated instances may share hardware with other instances from the same account that are not dedicated instance.
   - Pay for dedicated instances on demand save up to 70% by purchasing reserved instance or save up to 90% by purchasing spot instances.

3. Dedicated Host:
   - An Amazon EC2 dedicated host is a physical server with EC2 instance capacity fully dedicated to your use.
   - Dedicated host can help you address compliance requirement and reduce costs by allowing you to use your existing server bound software licenses.
   - Pay for a physical host that is fully dedicated to running your instances and bring your existing per socket, per core, per VM software license to reduce cost.
   - Dedicated host gives you additional visibility and control over how instances are placed in a physical server and you can consistently deploy your instances to the same server over time.
   - As a result dedicated host enables you to use your existing server bound software license and address corporate compliance and regulatory requirements.
   - Instances that run on a dedicated host are the same virtualized instances that you had get with traditional EC2 instance3s that use the XEN Hypervisor.
   - Each dedicated host supports a single instance size and type (for e.g C3.XLARGE)
   - Only BYOL, Amazon linux and AWS marketplace AMIs can be launched onto dedicated hosts.

4. Spot Instances:
   - Amazon EC2 spot instances let you take advantage of unused EC2 capacity in the AWS cloud. Spot instances are available at up to 90% discount compared to on-demand prices.
   - You can use spot instances for various test and development workloads.
   - You can also have the options to hibernate, stop or terminate your spot instances when EC2 reclaims the capacity back with two minutes of notice.
   - Spot instances are spare EC2 capacity that can save you up 90% off of on-demand prices that AWS can interrupt with a 2 minute notification. Spot uses the same underlying EC2 instances as on-demand and reserved instances, and is best suited for flexible workloads.
   - You can request spot instances up to your spot limit for each region.
   - You can determine the status of your spot request via spot request status code and message. You can access spot request status information on the spot instance page of the EC2 console of the AWS management console.
   - In case of hibernate, your instance gets hibernated and RAM data persisted. In case of stop, your instance gets shutdown and RAM is cleared.
   - With hibernate, spot instances will pause and resume around any interruptions so your workloads can pick up from exactly where they left off.

     Question: when would my spot instance get interrupted?
     Ans: primary reason would be Amazon EC2 capacity requirement (e.g: on-demand or reserved instances). Secondarily, if you have chosen to set a 'max spot price' and the spot price raise above.

5. Scheduled Instance:
   - Scheduled reserve instances enable you to purchase capacity reservations that recur on a daily, weekly or monthly basis, with a specified start time and duration for one year term.
   - You reserve the capacity in advance so that you know it is available when you need it.
   - You pay for the time that the instances are scheduled even if you do not use them.
   - Scheduled instances are a good choice for workloads that do not run continuously but do run on a regular schedule.
   - Purchase instances that are always available on the specified recurring schedule for a one year term.
     For example: you can use schedule instances for an application that runs during business hours of for batch processing that run at the end of the week.

6. Reserved Instances:
   - Amazon EC2 RI provides a significant discount up to 75% compared to on-demand pricing and provide a capacity reservation when used in a specific availability zone.
   - Reserved instances give you the option to reserve a DB instance for a one or three year term and in turn receive a significant discount compared to the on-demand instance pricing for the DB instance.

   There are 3 types of RI are available such as
   a. Standard RI: these provide the most significant discount up to 75% off on-demand and are best suited for steady-state usage.
   b. Convertible RI: these provide a discount up to 54% and the capability to change the attributes of the RI as long as the exchange results in the creation of reserved instances of greater or equal value.
   c. Scheduled RI: these are available to lunch within the time window you reserve.

   Question: can I transfer a convertible or standard RI from one region to another? How Do I change the configuration of a convertible RI?
   Ans: you can change the configuration of your convertible RI using the EC2 management console of the get reserved instance management quota API.

   Question: do I need to pay a fee when I exchange my convertible RI?

EC2 ACCESS:

- To access instances you need a key and key-pair name.
- You can download the private key only once.
- The public key is saved by AWS to match it to the key pair name and private key when you try to login to the EC2 instances.
- Without key pair you cannot access instances via RDP or SSH (linux).
- There are 20 EC2 instances soft limit per account, you can submit a request to AWS to increase it.

EC2 STATUS CHECK:

- By default AWS EC2 instance performs automated status checks every one minute.
- This is done on every running instance to identify any h/w of s/w issue.
- Status check is built into the AWS EC2 instance.
- They cannot be configured, deleted or disable.
- EC2 services can send its metric data to AWS cloudwatch every 5 minutes (enable by default)
- Enable detailed monitoring is chargeable and sends metric in every 1 minute.

- You are not charged for EC2 instances if they are stopped however attached EBS
- You area not charged for EC2 instances if they are stopped however attached EBS volumes incur charges.

When you stop an EBS Backed EC2 instance:

- Instances perform a shutdown.
- State changes from running to stopping.
- EBS volumes remain attached to the instance.
- Any data cached in RAM or instances store volume is gone.
- Instances retain its private IPV4 or any IPV6 address.
- Instances releases its public IPV4 address back to AWS pool.
- Instances retain its elastic IP address.

EC2 TERMINATION:

- When you terminate a running instance, the instance states changes as follows: Running -> shutting down -> terminate
- During the shutting down and terminated states you do not incur charges.
- By default EBS root devices volumes are deleted automatically when the EC2 instances are terminated.
- Any additional (non boot/boot) volumes attached to the instances by default persist after the instance is terminated.
- You can modify both behaviors by modifying the "delete on termination" attributes of any EBS volumes during instance launch or while running.
- Enable "EC2 termination protection" against accidental termination

EC2 METADATA:

- This is the instance data that you can use to configure of manage the instance.
- e.g IPV4 addresses IPV6 addresses, DNS hostname, AMI-id, instance id, instance type, local hostname, public keys, and security groups.
- Metadata can be only viewed from within the instance itself i.e you have to login to the instance.
- Metadata is not protected by encryption; anyone that has access to the instance can view this data.
- To view instance metadata: GET http://169.254.169.254/latest/Metadata

INSTANCES USER DATA:

- ➢ Data supplied by the user at instances launch in the form of a script to be executed during the instance boot.
- ➢ User data is limited to 16kb.
- ➢ You can change user data, by stopping EC2 first.
- ➢ User data is not encrypted EC2 bare metal instances.
- ➢ Non virtualized environmental.
- ➢ Operating system runs directly on hardware.
- ➢ Suitable for licensing restricted tier-1 business critical application.
- ➢ E.g : I3.metal, I5.metal, R5.metal, Z1d.metal, U-6tb1.metal

ELASTIC BLOCK STORAGE:

- ➢ Most common replicated with A-Z EBS volumes attached at lunch are deleted when instance terminate.
- ➢ EBS volumes attached to a running instance are not deleted when instance is terminate but are detached with data interact.

INSTANCE STORAGE:

- ➢ Physically attach to the host server.
- ➢ Data not lost when OS is rebooted.

Data lost when:

- ➢ Underlying drive fails.
- ➢ Instance is stop or terminated.
- ➢ You can't detach or attach to another instance.
- ➢ Do not rely on for valuable long term data.

WRITTEN BY:

NAGARJUNA HOTA