* PART-1 **EC2:- ELASTIC COMPUTE CLOUD**

Amazon EC2 provides scalable computing capacity in the AWS cloud

You can use amazon ec2 to launch as many or as few virtual servers as you needs configure security and networking and manage storage

Amazon EC2 enables up to scale up or scale down the instance

Integrate with s3 and RDS

Amazon EC2 is having two storage options i.e, EBS & instance store

Pre-configured templates are available know as amazon machine image

By defaults when you create on 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**
* \*General purpose
* \*Compute optimized
* \*Memory optimized
* \*Storage optimized
* \*Accelerated computing
* \*High Memory
* \*Previous generation

**General Purpose**: - Balanced Memory & CPU

**Compute Optimized: -**More CPU than RAM

**Memory optimized**: -More RAM

**Accelerated Computing /GPU**: -Graphics Optimized

**Storage Optimized**:- low latency

**High Memory Optimized:**- High RAM Nitro system

PART-02

* **General Purpose Instances: -**

General purpose instance provides a balance of compute, memory and networking resources, and can be used for a variety of workloads

3 Series in General Purpose

A series M series T series

Medium large A1 large T2 T3 T3a

M4 M5 M5a M5ad M5d

Instances Available in four sizes – Nano, small, Medium, Large

* **A1 Instances**
* **A1: –** Instances are ideally suited for scale-out workloads that are supported by the Arm ecosystem

These instances are well-suited for the following Application

1.Webserver

2.containerized miccor

3.Distributed Data stores

4.Distributed data stores

5.Application that requires arm instruction set

M4 M5 M5a M5ad and M5d

* **M4 Instance:-**

The New my instance features a custom intel xeon E5-2676 v3 haswell processor optimized specifically for EC2

VCP:- 2 to 40 (max)

RAM:- 8GB TO 160 GB (Max.)

Instance Storage – EBS Only

* **M4 M5 M5a M5ad and M5d Instance**:- These instance provide on ideal cloud infra, offering a balance of compute, m
* emory, and networking resources
* **M5 M5a M5ad and M5d Instance:-** These instance provide on ideal cloud infra, offering a balance of compute, memory, and networking resources for a brought

Range of Application

Used in:-Gaming server, web server, Small and Medium databases

VCPU :- 2 to 96 (max.)

RAM :- 8 to 384 (Max.)

Instance Storage :- EBS & NVME SSD

* **T2 T3 and 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 (base line rahti hai 5% to 40%)

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 :-

1.website and web app

2.code repositories

3.Development, build, test

4.microservices

PART-03

* **Compute Optimized Instances**

Compute optimized Instances are ideal for compute-Bound application that benefit from high performance processors.

Three types are available: - c4, c5, c5n (c3 is previous instance)

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

VCPU – 2 to 36 RAM – 3.75 to 60GB

Storage – EBS only Network – BW – 10 Gbps

* **Use cases: -** Web server, Batch processing, MMO gaming video encoding
* **C5 Instance**: - C5 are optimized for compute-intensive workloads and deliver cost-effective high performance at a low price per compute ratio
* Powered by aws nitro system

VCPU – 2 to 72 RAM – 4 to 192GB

Network BW – upto 25 Gbps

Instance storage- EBS only & NVMe SSD

* **Use cases:-** Web server, Batch processing, MMO gaming video encoding

Note : C5 support max 25 EBS volumes

C5 use elastic network adoptor

C5 uses new EC2 hypervisor

Part 04

* **Memory optimized - R series ,X series , Z series**

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

* **R4, R5, R5a, R5ad and R5d**

High Performance relational (mysql) and no SQL (mangoDB, Cassondra) databases.

Distributed web scale cache stores that provide in-memory caching of key value type data

Used in financial services, Hadoop

VCPU -2 to 96

RAM – 16 to 768 GB

Instances storage – EBS only & NVMe SSD

* **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 3904 GB

Instance storage -SSD

* **Z1d Instance**

High frequency z1d delivers a sustained all core frequency of upto 40 GHz , the fastest on any cloud instances

AWS nitro system, xeon processor upto 1.8TB of instances storage

VCPU – 2 to 48

RAM – 16 to 384 GB

Storage – NVM SSD

Use Cases – Electronic Design automation and certain databases workloads with high per-core licensing cost

PART 05

* **Storage Optimized Instances:- T series, D series, H series**
* **D2 instances**

Well suited for the following

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 244 GB

Storage – SSD

* **H1 Instances**

This family features upto 16 TB of HDD based local storage high disk throughout and balance of compute & memory

Well suited for app requiring sequential access to large amounts of data on direct-attached instance storage

Application that requires high throughout access to large Quantities of data

VCPU – 8 to 64

RAM – 32 to 256 GB

Storage – HDD

* **I3 and I3en Instances**

Well suited for –

High frequency online transaction processing system (OLTP)

Relational Databases

* No SQL databases
* Distributed file system
* Data warehousing application

VCPU – 2 to 96

RAM – 16 to 768 GB

Local Storage – NVMe SSD

Networking performance – 25 Gbps to 100 Gbps

Sequential throughout

Read – 16 Gb/s

Write – 64 Gb/s (I3)

8 GB/s (I3en)

Part 06

* **Accelerated computing instances :-** P series, G series, F series

Accelerated computing instance families use hardware acclerators, or co-processor to perform some functions such as floating point number calculation graphics processing or data pattern matching move efficiently than is possible in software running on cpus

* **P2 & P3 instances**

It uses NVIDIA  tesla GPUs

Provide high bandwidth networking

Upto 32 GB of memory per GPUs which makes them ideal for deep learning & computational fluid dynamics

P2 instance

VCPU – 4 TO 64 VCPU – 8 to 96

GPU – 1 to 16 GPU – 1 to 8

RAM - 61 to 768 Gb RAM – 61 to 768GB

Storage – SSD & DBS Storage – SSD & EBS

Used in – Machine learing, databases seismic analysis, geromics, molecular modeling, AI, deep learing

Note:- P3 support CUDA9 & Open CL APIs P2 suppports CUDA & AND OPEN CL 12

* **G2 & G3 Instances**

Optimized for graphics intensive application

Well suited for app like 3D virtualizations

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 t0 488 GB

GPU Memory – 8 to 32 GB

Network performance – 25 Gbps

Used in – Video creation services 3D visualization, streaming Graphics- interitive application

* **F1 instance**

F1 instances offer customizable field programmable gate arrays (FPG)

Each FPGA contains 2.5 million logic elements & 6800 DSP engines

Designed to accelerate computationally intensive alogorithims such as data flow or highly parallel operations

F1 provides local NVM SSD Storage

VCPU – 8 to 64

FGPA – 1 to 8

RAM – 122 to 976 GB

Storage – NVMe SSD

Used in :- Genomics research, financial analytics, real time video processing & big data search

PART 07

* **High Memory Instances :- U series**

High memory Instances are purpose built to run large-in memory databases, including production developments of SAP HANA in the cloud

Features:-

1. Latest Generation intel xeon Pentium 8176 M processor
2. 6, 9, 12 TB of instance memory the largest of any EC2 instance
3. Powered by the AWS nitro system a combination of dedicated hardware & lightweight hypervisor
4. Bare metal performance with direct access to host hardware
5. EBS optimized by default at no additional cost

Model no – u-6tb1.metal, u-9tb1.metal & u-12tb1.metal

1. Network performance – 25 Gbps

Dedicated EBS Bandwidth – 14 Gbps

1. Each instance offer 448 logical processor

Note:- High memory instances are bare metal instances and do not un on a hypervisor

Only available under dedicate host purchasing category (for 3 yr term)

O.S directly on hardware

* **Previous Generation Instances :- T1, M1, C1, CC2, M2, CR1, CG1, i2, HS1, M3, C3 and R3**

PART 08

* **EC2 instance purchasing options**
* **On Demand**
* **Dedicated Instance**
* **Dedicated host**
* **Spot Instance**
* **Schedule instance**
* **Reserved instance:- Standard RI , convertible RI, Schedule RI**

There are four ways to pay for amazon EC2 instance i.e on-demand reserved instance and sopt instance

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

On demand Instance

Aws on demand instances are virtual servers that run in AWS or AWS relational database service (RDS) and are purchased at a fixed rate per hour

They are also suitable

* **On demand Instance**

Aws on demand instance are vitural servers that tun in AWS or AWS relational Database service (RDS) and are purchasedat a fixed rate per hour

Aws recommends using on-demand Instances for applications with short term irregular workloads that connot be interrupted

They are 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 form the cost and complexities of planning purchasing and maintaining hardware and transforms what are commonly large fixed costs into much smaller variable cost

Pricing is per instance-hour consumed for each instance from the time on instance is launched until it is terminated or stopped

Each partial instance hour consumed will be billed per second for linux instances and as a full hour for all other instance types

* **Didicated instance**

Dedicated instances are run in a VPC on hardware that is dedicated to a single customer

Your dedicated instance are physically isolated at the host hardware level from instances that belong to other AWS Account

Dedicated instance may share hardware with other instance form the same AWS account that are not dedicated instance

Pay for dedicated instances on demand save upto 70 % by purchasing reserved instances, or save upto 90% purc hasing spot instances

* **Dedicated host**

An amazon EC2 dedicated that 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-code or per-vm software license to reduce cost

* **Spot instance**

Amazon EC2 spot instance let you take advantage of unused EC2 capacity In the AWS cloud spot instances are available at upto 90% discount compared to on-demand prices

You can use spot-instance for various test & development workloads

You also have the options to hi bernate stop terminate your spot instances when EC2 reclaims the capacity back with two minutes of notice

* **Que:- When would my spot instance get interrupted ?**

Primary reaon would be amazon EC2 capacity requirement (eg. On-demand or reserved instances)

Secondary if you have chosed to set a max spot price and the spot price rises above this

* **Schedule instance**

Schedule reserve instance enable you to purchase capcity reservations that resue on daily weekly or monthly basis with a specified start time and duration for a 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

* **Reserved Instances**

Amazon EC2 RI provide a significant discount (upto75%) compared to on demand pricing and provide capacity reservation when used in a specific availability zone

Types of reserved instance

Standard RI , convertible RI ,schedule RI

Standard RI :- these provide the most significant discount (upto 75% off on demand) and one best suited for steady state usage

Convertible RI :- these provide a discount (upto 54%) and the capability to change the attributes of the RI as long as the exchange results in the creation of reserved instances of grater or equal value

Schedule RI :- These are available to launch withing the time window you reserve

**Q, can I transfer a convertible or standard RI from one region to another?**

**Q. how do I change the configuration of a convertible RI?**

**Ans:-** Yes can change the configuration of your convirtable RI using the EC2 management console or the get reserved instances management Quata API

**Q. Do I need to pay a fee when I exchange my convertible RI ?**

PART-09

* **EC2 Access :-**

To access instances, you need a key and key pair name

You can download the private key only ones

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 instnaces

Without key pair you connot access instances via RDP or SSH (Linux)

There is 20 EC2 instances soft limit per account , you can submit a request to AWS to increase it

* **EC2 Status Check**

By default AWS EC2 service performs automated status checks every one minute

This is done on every running EC2 instances to identifiey any H/W or software issue

Status check is buit 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 (enabled by defaults)

Enabled detailed monitoring is chargeable and sends metric in every 1 minutes

You are not charged for EC2 instances if they are stopped however attached EBS volume in charges

* **When you stop on EBS Backed EC2 Instance**

Instances perform a shutdown

State changes from running – stopping

EBS volumes remains attached to the instances

Any data cached in RAM or instance store volume is gone

Instance retain its private IPV4 or any IPV6 address

Instance releases its pubic IPV4 address back to AWS pool

* **EC2 Termination**

When you terminate a running instance the instance state changes as follow running – shutting down – Terminated

During the shutting down and terminated states, you do not incur charges

By defaults EBS root devices volumes are deleted automatically when the EC2 instances are terminated

Any additional (non boot/not volumes attached to the instances by default persisit ofter the instances is terminated

You can modify both behaviours by modifiying the Delete on termination attributed of any EBS volumes during instances launch or while running

Enable EC2 termination protection against accidental termination

* **EC2 metadeta**

This is instance data that you can use to configure or manage the instance

EX. IPV4 addresses, IPV6 addresses DNS hostname, AMI – id instances ID Instance Type, local hostname public keys security groups

Metadata can be only viewed from within the instance itself i.e you have to login to the instance

Metadata is not proteded 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

* **Instance user data**

Data supplied by the user at instance launch in the form of a script to be executed during the instance boot

User data is limited to 16 kb

You can change user data, by stopping EC2 encrypted

* **EC2 bare metal instance**

Non virutalised Enviroment

Operating system runs directly on hardware

Suitable for licensing restricted tier-1 business critical application

EX. I3 metal , i5 metal, i5 metal , z1d metal u-6tb1 metal

* **Elastic Block Storage**

Most common replicated with A-Z

EBS volumes attached at launch are deleted when instance terminate

EBS volumes attached to a running instance are not deleted when

Instance terminated 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 variable long term data