AWS EC2

Instance Purchasing Options

- On-Demand Instances – Pay, by the second, for the instances that you launch.
- Reserved Instances – Purchase, at a significant discount, instances that are always available, for a term from one to three years.
    - Reserved Instances are not physical instances, but rather a billing discount applied to the use of On-Demand Instances in your account.
    - Reserved instance limit:
        - For region: 20
        - For each AZ: 60 (20*3)
        - Total: 20+60 = 80
    - RI Types:
        - Standard RIs: These provide the most significant discount (up to 75% off On-Demand) and are best suited for steady-state usage.
        - Convertible RIs: These provide a discount (up to 54% off On-Demand) and the capability to change the attributes of the RI
        - Scheduled RIs: These are available to launch within the time windows you reserve.
- Scheduled Instances – Purchase instances that are always available on the specified recurring schedule, for a one-year term.
- Spot Instances – Request unused EC2 instances, which can lower your Amazon EC2 costs significantly.
- Dedicated Hosts – Pay for a physical host that is fully dedicated to running your instances, and bring your existing per-socket, per-core, or per-VM software licenses to reduce costs.
- Capacity Reservations – Reserve capacity for your EC2 instances in a specific Availability Zone for any duration.

Instance Types:

Accelerated Computer Instances: (EBS optimized)

- Uses hardware accelerators or co-processors to perform some function more efficiently rather than software.
    - Eg:
        - Floating point calculation
        - Graphics Processing.
    - Types
        - GPU: Graphical Processing Unit.
        - GPU compute instances for general-purpose computing (P3/p2)
            - P3
                - Next gen parallel processing NVIDIA Volta GV100
                - P3 can only be used under VPC. Old EC2 Classic N/W can't be used.
            - P2
                - Used for General Purpose using CUDA and OpenCL
                - NVIDIA Tesla K80
            - Used for
                - Artificial intelligence (AI),
                - Machine learning (ML),
                - Deep learning (DL)
                - High performance computing (HPC) applications

        - GPU graphics instances for graphics intensive applications (G3)
            - G3
                - use NVIDIA Tesla M60 GPUs
                - Provide a high-performance platform for graphics applications using DirectX or OpenGL.
                -

        - FPGA programmable/Customizable hardware computes instances for advanced scientific workloads. (F1)
        - Field programmable gate arrays (FPGAs).
        - AFI: Amazon FPGA Image.
    - While using remote desktop use a different remote access tool, such as VNC.

Compute Optimized instances (EBS optimized)

- Designed for applications that benefit from high compute power.
- Used for
    - High-performance web servers
    - High-performance computing (HPC)
    - Distributed analytics and machine learning inference
- Can change the Processor State Control of C4.8XLarge instance.
- Types:
    - C5
    - C4
- C5 instance EBS as storage volume
- C5 instances access EBS volumes via PCI attached NVM Express (NVMe) interfaces.
- C5 instances use the Elastic Network Adapter (ENA) for networking
- C5 instances support a maximum for 27 EBS volumes for all Operating systems.

General Purpose instances

- M5:

Memory Optimized

- Memory-optimized instances offer large memory size for memory intensive applications
- Types
    - X1 & X1e: Used for running in-memory databases like SAP HANA, big data processing engines like Apache Spark
    - R5,R5a,R4,Z1d

Storage Optimized Instances (EBS-optimized instances)

- Used for
    - Applications that require high sequential read/write access and low cost storage for very large data sets

- Dense Storage (D2):
    - Designed for workloads that require high sequential read and write access to very large data sets
        - Eg:
            - Hadoop
            - log processing applications
    - Low price
    - HDD-based instance storage. So need external support for fault tolerance and redundancy.
    - Can be launched in both EC2-Classic and Amazon VPC
- H1
    - Can only be launched in Amazon VPC.
    - Used for:
        - Kafka
        - HDFS
        - log

- I3:
  - Provides Non-Volatile Memory Express (NVMe) SSD-backed instance storage optimized for
    - low latency
    - very high random I/O performance
    - high sequential read throughput and provide high IOPS at a low cost.
  - Use Cases:
    - No-SQL database
    - In-memmory database
  - I3 instances support TRIM (Command to wipe out unused data)

EBS

- Used to create storage volume and attach to EC2 instance.
- Placed in specific AZ to protect from failures.
- Not automatically replicated to different AZ.

EBS volume types

- General Purpose SSD (GP2)
    - General purpose balances both price and performance.

- Provisioned IOPS SSD (IO1)
    - Used for application such as large relational or NoSQL DB.

- Throughput Optimized HDD (ST1)
    - For frequently accessing
    - Can't be a boot Volume.
    - Magnetic storage.
    - Used for
        - Big Data
        - Log processing
        - Data warehousing
- Cold HDD (SC1)
    - Lowest cost storage for infrequently accessed records.
    - Can't be a boot Volume.

- Magnetic (Standard)
    - Bootable
    - Lowest cost/GB

- Can't Mount 1 EBS volume to multiple EC2 instances.

Amazon Machine Image (AMI)

- Snapshots of virtual machine.
- Types of virtualization
    a. HVM
    b. Para-virtual

LAB Points:

- One subnet = One AZ
- For SSH, we will use the public ip address.
- Types of Status Checks
    a. System Status Check: Monitor the AWS system where EC2 instance runs.
    b. Instance Status Check: Monitor the Software and N/W config. on our instance runs.
- We can't encrypt the root device volume by default (But you can do it).
- Termination Protection turned off by default.
- Default behavior is that once Ec2 instance is deleted, corresponding EBS volume will get delete.

Security Group Points:

- Any rule that added/removed to security group will reflect immediately.
- Security Groups are "state-full": Any rule added to inbound will auto reflect in outbound. No need to specify in outbound folder.
- All inbound traffics are "blocked" by default. We need to specify the rules to allow it.
- You can specify allow rules not deny rules.
- RDP port number: 3389
- MySQL : 3306
- We can't block specific IP address using Security Group. For that we need to use NAL.

EBS Volume

- We can't attach Ec2 instance in one AZ to EBS volumes from another AZ.
- For creating one EBS volume in another AZ, First need to create a snapshot from existing volume and then create EBS volume with another AZ.
- We can't modify only magnetic disk size on fly.

1.

EC2 Instance replacement		
First Step	AZ1 - AZ2	Region1 -Region2
Create Snapshot	Create new volume in AZ2	Do Copy

2.

EBS Volume Methods	Create Volume	Copy	Create Image
Create Snapshot			
	AZ1-AZ2	Region1 - Region2	Create new Ec2 instance

RAID Volumes (Redundant Array of In depended Disk)

- Used to improve DISK I/O performance.
- RAID types:
  - RAID-0: Striped, Good Performance, but no redundancy.
  - RAID-1: Mirrored , redundancy
  - RAID-5: Good for read, Bad for write,
  - RAID-10: RAID-1 + RAID-0
- Where to use RAID: Suppose any service that aws not support (Cassandra) and to use in your EC2 instance.
- Before RAID SnapShot:
  - Freeze the Filesystem.
  - Unmount the RAID array.

- Shutdown the EC2 instance.

AMI Types

- Instance Store
    a. We can't attach additional instance store volumes after launching Ec2 instance.
        i. But can attach EBS volume
    b. We can't stop the instance.
    c. Once failed, data will get lost.
- EBS
- You can reboot both.

ELB

- ELB Types
    - Application LB
    - Classic LB
- Inservice-outservice.
- We only get DNS name while creating the LB. We will not get any public IP address.
- Create health check html file (Index.html or any other)

Cloud Watch

- Pre-Instance Matrices in EC2
    - CPU based
    - Disk based
    - N/w based
    - Status based

Instance Metadata:

CURL http://169.254.169.254/latest/metadata/

CURL http://169.254.169.254/latest/metadata/public -ipv4

Launch Configuration and Auto Scaling

Placement Group:

- Logical Grouping of instances within a single AZ.
- Imp Points:
  - Can't span across multiple AZ.
  - Placement group name must be unique in our A/C.
  - Can't merge Placement group.
  - Can't move existing instance into placement group.

Elastic File System

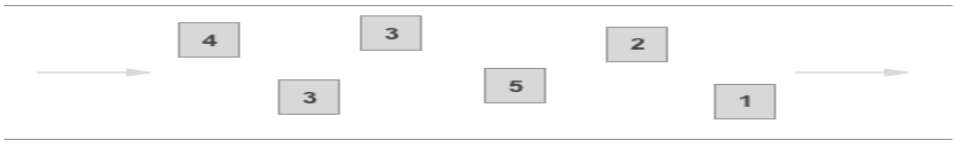
- File storage service.
- Data is stored across multiple AZ's in a region.
- Using as centralized repo. In between ec2 instances.
- EFS allows multiple instance to connect but EBS NOT.

AWS Application Services

Simple Queue Service

1. Used to store messages inside message queue.
2. SQS is always a PULL based system.
3. Message will last until reach the visibility time.
4. Types of queues:

a. Standard



b. FIFO

5. Comparison:

	Standard	FIFO
Default	Yes	No
Guarantee that Msg. will deliver	At least once	Only once
Duplicate	Yes	NO
Order	Try max to keep the same order that they sent	Strictly follow the order
Transactions		300 /sec

6. Msg. can kept in queue from 1 minute to 14 days. Default is 4 days.
7. Visibility time out: Amount of time that the message is invisible after read from queue.
    a. If job finishes before VTO expires, msg. will delete from the queue.
    b. Else, it will visible and chances are there to process the message by another job.
    c. Maximum VTO is 12 hours.
8. Polling types:
    a. Short poll:
        i. Default one.
        ii. Will return response even queue is empty. That may increase the cost.
    b. Long poll:
        i. Return only the queue is not empty or time out happens.

Simple Workflow Service (SWF)

9. Used for coordinate work across distributed application components.
10. Comparison between SQS and SWF:

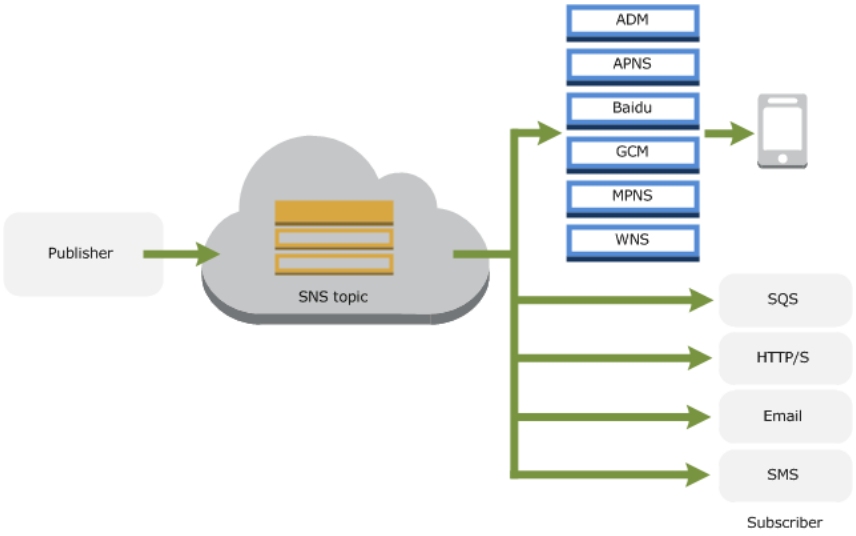
	SQS	**SWF**
Retention Period	14 Days	1 Year
API	Message oriented	Task oriented
Message processed	Need to ensure that only once	Only Once
Duplication	Need to handle duplicate message	Never
Keep Tracks	Need to create application-level tracking	Yes

11. SWF Actors:
    a. Workflow Starters: Application that start the workflow.
    b. Deciders: Control the flow of activity tasks.
    c. Activity workers: Do activity tasks.

Simple Notification Service

12. Used for sending notifications
13. Push notification services:
    a. Amazon Device Messaging (ADM)

b. Apple Push Notification Service (APNS) for both iOS and Mac OS X
c. Baidu Cloud Push (Baidu)
d. Google Cloud Messaging for Android (GCM)
e. Microsoft Push Notification Service for Windows Phone (MPNS)
f. Windows Push Notification Services (WNS)



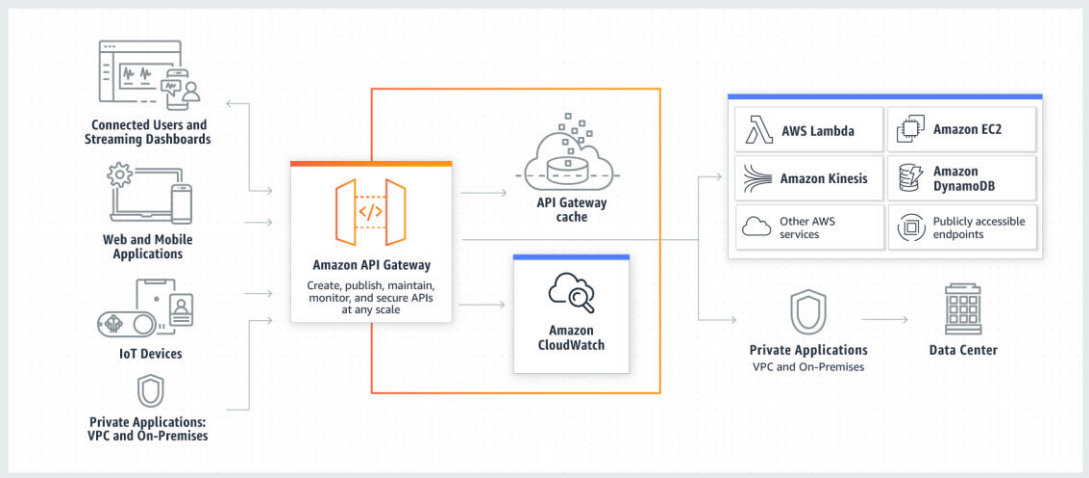
14. Also messages can be push to SQS, SMS, email, http/s and lambda functions**.**
15. AWS store messages inside **multiple AZ** to avoid data loss.
16. SNS is **PUSH** basted delivery.

Elastic Transcoder

1. Used to convert from one media format to other.

API Gateway

1. Create REST and WebSocket APIs that act as a “front door” for applications.



2. API caching: Once enabled, speed up the response by saving the response for specified time (TTL) and responds this response for subsequent requests.
3. Must enable the Cross origin resource sharing (CORS).

Kinesis

1. Used to load and analyses streaming data.
2. Services:
    a. Kinesis Streams

    b. Kinesis Firehose
    c. Kinesis Analytics

3. Kinesis Streams
    a. By default it will store data for 24 hours. Also can upgrade to 7days.
    b. Data stored in shards.

4. Kinesis Firehose
    a. No need to worry about shards and streams and fully automated.
    b. Data will send to S3 or redshift or elastic search cluster.

5. Kinesis Analytics
    a. Allow to run SQL queries and store the data into S3/Redshift/ElasticSearchCluster.