ELASTIC COMPUTE CLOUD

http://clusterfrak.com/notes/certs/aws_saa_notes/

Backed by AWS provide the re-sizeable compute capacity in the cloud. It's designed to make web scale computing easier for developers.

POINTS TO REMEMBER

1. EC2 enable compute in the cloud. Amazon EC2 changes the economics of computing by allowing you to pay only for capacity that you actually use.
2. Once an Instance has been launched with instance store storage, you can not attach additional instance store volumes after the instance is launched, only EBS volumes
3. When using ephemeral storage, an underlying host failure will result in data loss
4. You can reboot both instance types (w/ephemeral and EBS volumes) and will not lose data, but again, an ephemeral volume based instance can NOT be stopped
5. You can reboot both instance types (w/ephemeral and EBS volumes) and will not lose data, but again, an ephemeral volume based instance can NOT be stopped.
6. You can poll an instances meta-data by using curl http://169.254.169.254/latest/meta-data/
7. You can get an instance's IP address by using curl http://169.254.169.254/latest/meta-data/public-ipv4
8. Can not encrypt root volumes, but you can encrypt any additional volumes that are added and attached to an EC2 instance.
9. You can have up to 10 tags per EC2 instance
10. AWS does not recommend ever putting RAID 5's on EBS
11. Termination protection is turned off by default, you must turn it on
12. **Roles:**
    - You can only assign an EC2 role to an instance on create. You cannot assign a role after the instance has been created and/or is running.
    - You can change the permissions on a role post creation, but can NOT assign a new role to an existing instance
    - Role permissions can be changed, but not swapped
    - Roles are more secure then storing your access key and secret key on individual EC2 instances.
    - Roles are easier to manager, You can assign a role, and change permissions on that role at any time which take effect immediately
    - Roles can only be assigned when that EC2 instance is being provisioned
    - Roles are universal, you can use them in any region.

13. Instance sizing:

a. T2 - Lowest Cost General Purpose - Web/Small DBs
b. M4 - General Purpose - App Servers
c. M3 - General Purpose - App servers
d. C4 - Compute Optimized - CPU Intensive Apps/DBs
e. C3 - Compute Optimized - CPU Intensive Apps/DBs
f. R3 - Memory Optimized - Memory Intensive Apps/DBs
g. G2 - Graphics / General Purpose - Video Encoding/Machine Learning/3D App Streaming
h. I2 - High Speed Storage - NoSQL DBs, Data Warehousing
i. D2 - Dense Storage - Fileservers/Data Warehousing/Hadoop
j. D - Density
k. I - IOPS
l. R - RAM
m. T - Cheap General Purpose
n. M - Main General Purpose
o. C - Compute
p. G – Graphics

14. Storage Types:

- **Instance Store (Ephemeral):**
  - Also referred to as ephemeral storage and is not persistent.
  - Instances using instance store storage cannot be stopped. If they are, data loss would result.
  - If there is an issue with the underlying host and your instance needs to be moved, or is lost, Data is also lost
  - Instance store volumes cannot be detached and reattached to other instances; They exist only for the life of that instance
  - Best used for scratch storage, storage that can be lost at any time with no bad ramifications, such as a cache store
- **EBS (Elastic Block Storage):**
  - Elastic Block Storage is persistent storage that can be used to procure storage to EC2 instances.
  - You can NOT mount 1 EBS volume to multiple EC2 instances instead you must use EFS .
  - Default action for EBS volumes is for the root EBS volume to be deleted when the instance is terminated.
  - By default, ROOT volumes will be deleted on termination, however with EBS volumes only, you can tell AWS to keep the root device volume.
  - EBS backed instances can be stopped, you will NOT lose any data.

- EBS volumes can be detached and reattached to other EC2 instances 3 Types of available EBS volumes can be provisioned and attached to an **EC2 instance:**
  - **General Purpose SSD (GP2):**
    - General Purpose up to 10K IOPS.
    - 99.999% availability.
    - Ratio of 3 IOPS per GB with up to 10K IOPS and ability to burst.
    - Up to 3K IOPS for short periods for volumes under 1GB.
  - **Provisioned IOPS SSD (I01)**
    - Designed for I/O intensive applications such as large relational or No-SQL DBs.
    - Use if need more than 10K IOPS
  - **Magnetic (Standard)**
    - Lowest cost per GB
    - Ideal for workloads where data is accessed infrequently and apps where the lowest cost storage is important.
    - Ideal for fileservers
- **Encryption**:
  - Root Volumes cannot be encrypted by default, you need a 3rd party utility.
  - Other volumes added to an instance can be encrypted.

- **AMIs:**
  - AMI's are simply snapshots of a root volume and is stored in S3.
  - AMI's are regional. You can only launch an AMI from the region in which it was stored.
  - You can copy AMI's to other regions using the console, CLI or Amazon EC2 API.
  - Provides information required to launch a VM in the cloud.
  - Template for the root volume for the instance (OS, Apps, etc).
  - Permissions that control which AWS accounts can use the AMI to launch instances.
  - When you create an AMI, by default it's marked private. You have to manually change the permissions to make the image public or share images with individual accounts.
  - Block device mapping that specifies volumes to attach to the instance when its launched.
  - Hardware Virtual Machines (HVM) AMI's Available.
  - Paravirtual (PV) AMI's Available
  - You can select an AMI based on:
    - Region
    - OS
    - Architecture (32 vs. 64 bit)
    - Launch Permissions
    - Storage for the root device (Instance Store Vs. EBS)

- **Security Groups:**
  - Act like virtual firewalls for the associated EC2 instance.
  - If you edit a security group, it takes effect immediately.
  - You cannot set any deny rules in security groups, you can only set allow rules.
  - There is an implicit deny at the end of the security group rules.
  - You don't need outbound rules for any inbound request. Rules are stateful meaning that any request allowed in, is automatically allowed out.
  - You can have any number of EC2 instances associated with a security group.
- **Snapshots:**
  - You can take a snapshot of a volume, this will store that volumes snapshot on S3.
  - Snapshots are point in time copies of volumes.
  - The first snapshot will be a full snapshot of the volume and can take a little time to create.
  - Snapshots are incremental, which means that only the blocks that have changes since your last snapshot are moved to S3.
  - Snapshots of encrypted volumes are encrypted automatically.
  - Volumes restored from encrypted snapshots are encrypted automatically.
  - You can share snapshots but only if they are not encrypted.
  - Snapshots can be shared with other AWS accounts or made public in the market place again as long as they are NOT encrypted
  - If you are making a snapshot of a root volume, you should stop the instance before taking the snapshot.
- **RAID Volumes:**
  - If you take a snapshot, the snapshot excludes data held in the cache by applications or OS. This tends to not be an issue on a single volume, however multiple volumes in a RAID array, can cause a problem due to interdependencies of the array.
  - Take an application consistent snapshot
    - Stop the application from writing to disk
    - Flush all caches to the disk
  - Snapshot of RAID array --> 3 Methods:
    - Freeze the file system
    - Unmount the RAID Array
    - Shutdown the EC2 instance --> Take Snapshot --> Turn it back on.

- **Placement Groups:**
  - A logical group of instance in a single AZ.
  - Using placement groups enables applications to participate in a low latency, 10Gbps network.
  - Placement groups are recommended for applications that benefit from low network latency, high network throughput or both.
  - A placement group can't span multiple AZ's so it is a SPoF.
  - Then name you specify for a placement group must be unique within your AWS account
  - Only certain types of instances can be launched in a placement group. Computer Optimized, GPU, Memory Optimized, and Storage Optimized.
  - AWS recommends that you use the same instance family and same instance size within the instance group.
  - You can't merge placement groups
  - You can't move an existing instance into a placement group
  - You can create an AMI from your existing instance and then launch a new instance from the AMI into a placement group

- **Pricing Models:**
  - On Demand:
    - Pay fixed rate by the hour with no commitment.
    - Users that want the low cost and flexibility of EC2.
    - Apps with short term, spiky or unpredictable workloads that cannot be interrupted.
    - Apps being developed or tested on EC2 for the first time.
  - Reserved:
    - Provide capacity reservation and offer significant discount on the hourly charge for an instance (1-3 year terms).
    - Applications have steady state, or predictable usage.
    - Apps that require reserved capacity.
    - Users able to make upfront payments to reduce their total computing costs even further.
  - Spot:
    - Bid whatever price you want for instance capacity by the hour.
    - When your bid price is greater than or equal to the spot price, your instance will boot
    - When the spot price is greater than your bid price, your instance will terminate with an hours notice.
    - Applications have flexible start and end times.
    - Apps that are only feasible at very low compute prices.
    - Users with urgent computing needs for large amounts of additional capacity.
    - If the spot instance is terminated by Amazon EC2, you will not be changed for a partial hour of usage.
    - If you terminate the instance yourself you WILL be charged for any partial hours of usage.

Amit Goswami

AWS – Architect: Notes (EC2)