# AWS Certification – Compute Services – Cheat Sheet

## EC2

- provides scalable computing capacity
- Features
  - Virtual computing environments, known as *EC2 instances*
  - Preconfigured templates for EC2 instances, known as *Amazon Machine Images (AMIs)*, that package the bits needed for the server (including the operating system and additional software)
  - Various configurations of CPU, memory, storage, and networking capacity for your instances, known as *Instance types*
  - Secure login information for your instances using *key pairs* (public-private keys where private is kept by user)
  - Storage volumes for temporary data that's deleted when you stop or terminate your instance, known as *Instance store volumes*
  - Persistent storage volumes for data using Elastic Block Store (EBS)
  - Multiple physical locations for your resources, such as instances and EBS volumes, known as *Regions and Availability Zones*

    - A firewall to specify the protocols, ports, and source IP ranges that can reach your instances using *Security Groups*
    - Static IP addresses, known as *Elastic IP addresses*
    - Metadata, known as *tags*, can be created and assigned to EC2 resources
    - Virtual networks that are logically isolated from the rest of the AWS cloud, and can optionally connect to on premises network, known as Virtual private clouds (VPCs)
- Amazon Machine Image
    - template from which EC2 instances can be launched quickly
    - does NOT span across across regions, and needs to be copied
    - can be shared with other specific AWS accounts or made public
- Purchasing Option
    - On-Demand Instances
        - pay for instances and compute capacity that you use by the hour
        - with no long-term commitments or up-front payments
    - Reserved Instances
        - provides lower hourly running costs by providing a billing discount
        - capacity reservation that is applied to instances
        - suited if consistent, heavy, predictable usage

        - provides benefits with Consolidate Billing
        - can be modified to switch Availability Zones or the instance size within the same instance type, given the instance size footprint (Normalization factor) remains the same
        - pay for the entire term regardless of the usage, so if the question targets cost effective solution and answer mentions reserved instances are purchased & unused, it can be ignored
    - Spot Instances
        - cost-effective choice but does NOT guarantee availability
        - applications flexible in the timing when they can run and also able to handle interruption by storing the state externally
        - AWS will give a two minute warning if the instance is to be terminated to save any unsaved work
    - Dedicated Instances, is a tenancy option which enables instances to run in VPC on hardware that's isolated, dedicated to a single customer
    - Light, Medium, and Heavy Utilization Reserved Instances are no longer available for purchase and were part of the Previous Generation AWS EC2 purchasing model
- Enhanced Networking
    - results in higher bandwidth, higher packet per second (PPS) performance, lower latency, consistency,

scalability and lower jitter
    - supported using Single Root I/O Virtualization (SR-IOV) only on supported instance types
    - is supported only with an VPC (not EC2 Classic), HVM virtualization type and available by default on Amazon AMI but can be installed on other AMIs as well
- Placement Group
    - provide low latency, High Performance Computing via 10Gbps network
    - is a logical grouping on instances within a Single AZ
    - don't span availability zones, can span multiple subnets but subnets must be in the same AZ
    - can span across peered VPCs for the same Availability Zones
    - existing instances cannot be moved into an existing placement group
    - for capacity errors, stop and start the instances in the placement group
    - use homogenous instance types which support enhanced networking and launch all the instances at once

EBS

# Elastic Load Balancer & Auto Scaling

- Elastic Load Balancer
    - Managed load balancing service and scales automatically
    - distributes incoming application traffic across multiple

EC2 instances

- is distributed system that is fault tolerant and actively monitored by AWS scales it as per the demand
- are engineered to not be a single point of failure
- need to Pre Warm ELB if the demand is expected to shoot especially during load testing
- supports routing traffic to instances in multiple AZs in the same region
- performs Health Checks to route traffic only to the healthy instances
- support Listeners with HTTP, HTTPS, SSL, TCP protocols
- has an associated IPv4 and dual stack DNS name
- can offload the work of encryption and decryption (SSL termination) so that the EC2 instances can focus on their main work
- supports Cross Zone load balancing to help route traffic evenly across all EC2 instances regardless of the AZs they reside in
- to help identify the IP address of a client
    - supports Proxy Protocol header for TCP/SSL connections
    - supports X-Forward headers for HTTP/ HTTPS connections
- supports Stick Sessions (session affinity) to bind a user's session to a specific application instance,
    - it is not fault tolerant, if an instance is lost the information is lost

        - requires HTTP/HTTPS listener and does not work with TCP
        - requires SSL termination on ELB as it users the headers
    - supports Connection draining to help complete the in-flight requests in case an instance is deregistered
    - For High Availability, it is recommended to attach one subnet per AZ for at least two AZs, even if the instances are in a single subnet.
    - cannot assign an Elastic IP address to an ELB
    - IPv4 & IPv6 support however VPC does not support IPv6
    - HTTPS listener does not support Client Side Certificate
    - for SSL termination at backend instances or support for Client Side Certificate use TCP for connections from the client to the ELB, use the SSL protocol for connections from the ELB to the back-end application, and deploy certificates on the back-end instances handling requests
    - supports a single SSL certificate, so for multiple SSL certificate multiple ELBs need to be created
- Auto Scaling
    - ensures correct number of EC2 instances are always running to handle the load by scaling up or down automatically as demand changes
    - cannot span multiple regions.
    - attempts to distribute instances evenly between the

AZs that are enabled for the Auto Scaling group
    - performs checks either using EC2 status checks or can use ELB health checks to determine the health of an instance and terminates the instance if unhealthy, to launch a new instance
    - can be scaled using manual scaling, scheduled scaling or demand based scaling
    - cooldown period helps ensure instances are not launched or terminated before the previous scaling activity takes effect to allow the newly launched instances to start handling traffic and reduce load
- Auto Scaling & ELB can be used for High Availability and Redundancy by spanning Auto Scaling groups across multiple AZs within a region and then setting up ELB to distribute incoming traffic across those AZs
- With Auto Scaling use ELB health check with the instances to ensure that traffic is routed only to the healthy instances