Amazon EC2

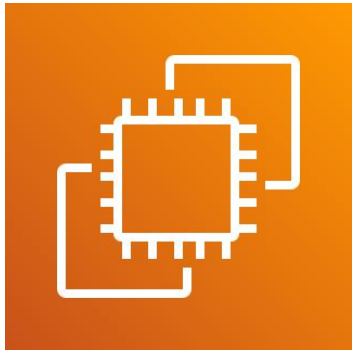
Contents

- EC2 Introduction
- EBS, Snapshot introduction
- AMI type (EBS and Instance store)
- Security Groups
- EFS
- ENI, ENA, EFA
- EC2 Userdata
- EC2 Spot and Spot Fleet
- Placement Group
- EC2 Hibernate

EC2 Introduction

What is EC2?

- EC2 stand for **E**lastic **C**ompute **C**loud
- One of the most important service of AWS offering
- Providing Compute capability in the cloud.



EC2 pricing model

1

On-Demand

you pay for compute capacity by the second with fixed rate with no long-term commitments

2

Reserved

Provide capacity reservation.
Offering significant discount price compare On-Demand.
Contract period is 1 ~ 3 years

3

Spot

Allow you to set bid price.
Offering significant discount price depend on Spot price (Providing from Amazon EC2)

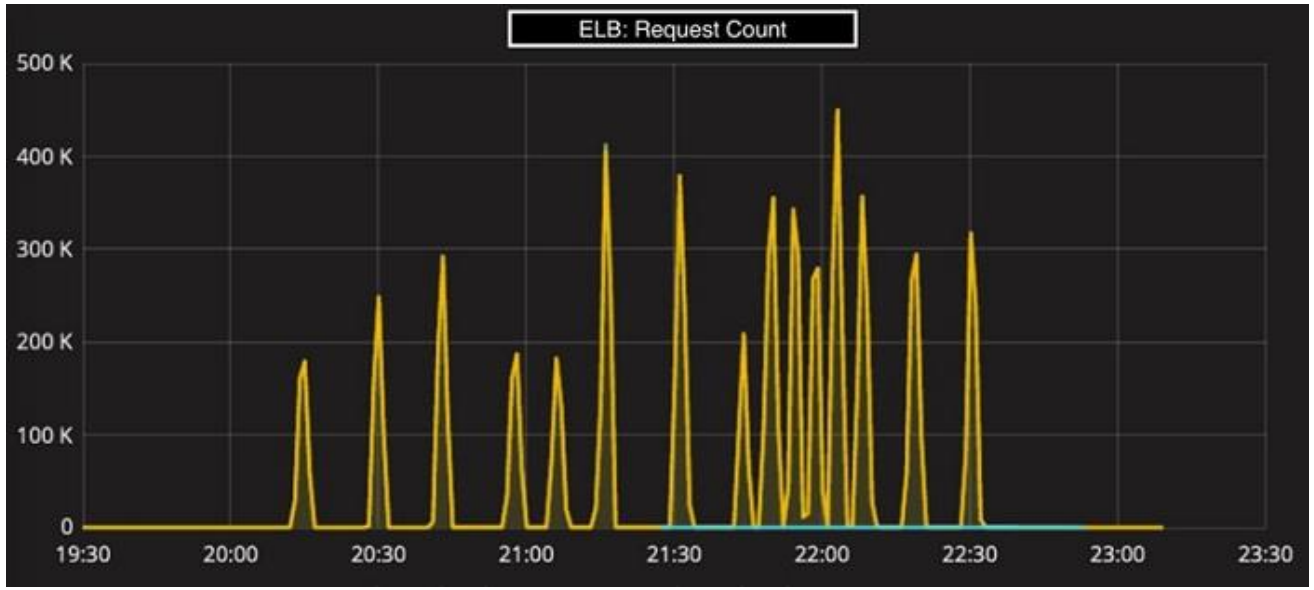
4

Dedicated Host

Physical server fully dedicated for your use.
For compliance with licence model from 3rd party (Microsoft, Oracle…)

On-demand use case

- Applications with short term, spiky, unpredictable workload
- For testing purpose, Lab
- No commitment, no Up-front payment



Reserved Instance (RI) use case

- Application with stable, predictable workload
- Need to make up-front payment
- Need to sign commitment term at least 1 year

Standard RI

Convertible RI

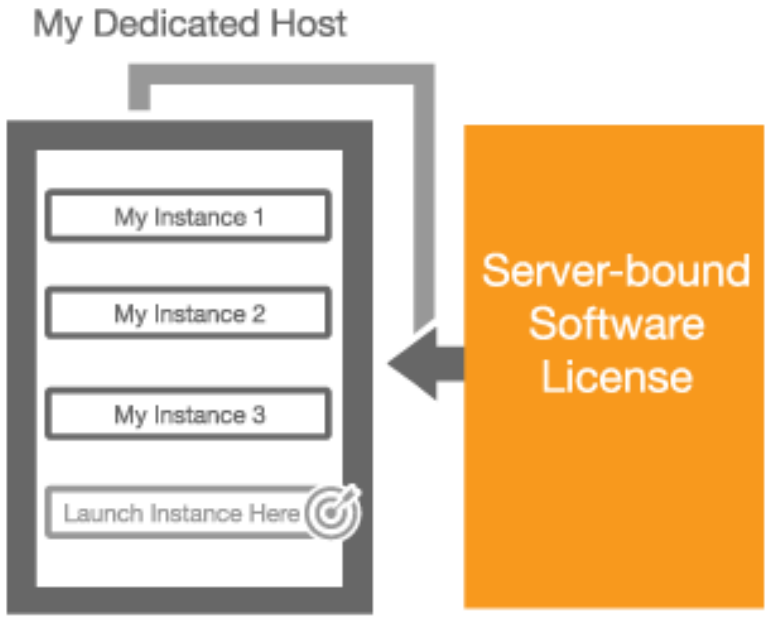
Scheduled RI

Spot Instance use case

- Can get a discount of up to 90% compare to On-Demand
- Short workload, can lose instances, downtime is acceptable
- Running high compute workload with low cost

Dedicated Hosts use case

- Physical dedicated EC2 server
- Licencing does not support multi-tenant virtualization or cloud deployments
- Can be purchased by hourly (On-Demand) or Reserved.



EC2 instance types

- Instance type format: family.type (t2.micro, m5.large...)
- Each family for specific purpose
- Example

Family	Speciality	Use case
R5	Memory Optimized	Application, DBs
C5	Compute Optimized	High CPU apps, DBs
T2/T3	General purpose, cheap	Web
Mac1	For MacOS	For MacOS user

EC2 summary

- EC2 is AWS service providing compute capability
- Pricing model need to be considered during provisioning infra on AWS

  - On-demand

  - RI

  - Spot

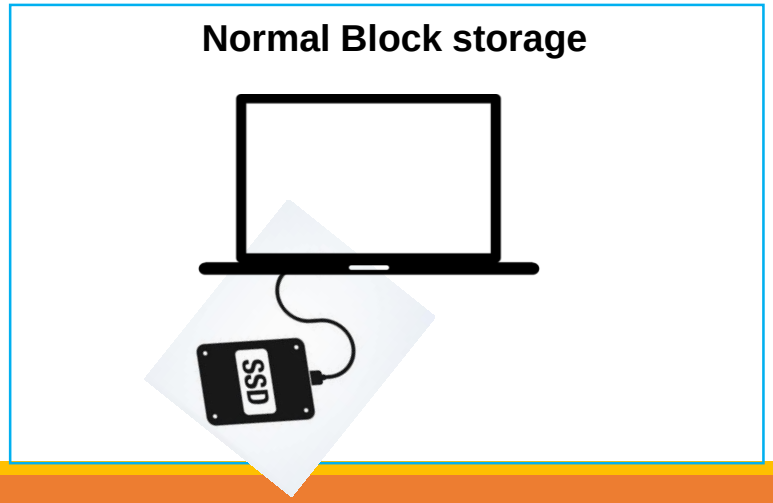
  - Dedicated Hosts

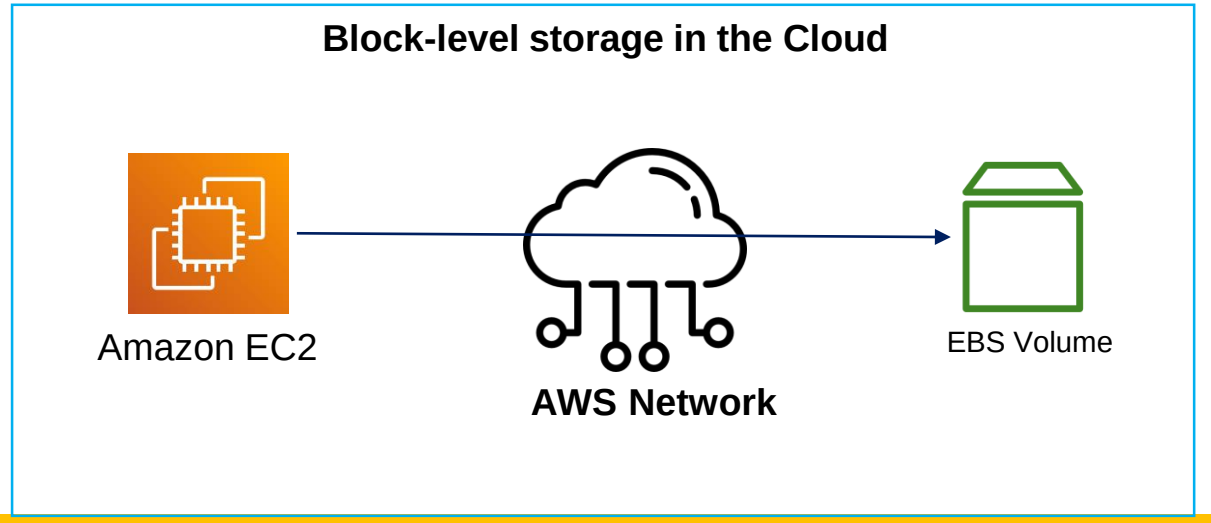
- Each instance type for specific purpose (High Mem, CPU, GPU, I/O…)

EBS Introduction

What is EBS volume?

- EBS stand for **E**lastic **B**lock **S**torage
- EBS provides block-level storage in the cloud
- EBS can easy to attach to (detach from) instances
- EBS is locked Availibity Zone scope





EBS volume type

Solid state drives (SSD)			Hard disk drives (HDD)		
Volume Type	General Purpose SSD	Provisioned IOPS SSD	Throughput Optimized HDD	Cold HDD	Magnetic
Description	Provides a balance of price and performance. For most common workloads	Provides high performance for mission-critical, low-latency, or high-throughput workloads.	A low-cost HDD designed for frequently accessed, throughput-intensive workloads.	The lowest-cost HDD design for less frequently accessed workloads.	Previous generation volume types
Name	gp2, gp3	io1, io2	st1	sc1	standard
Use Cases	**Boot volumes, low-latency interactive applications, dev, test.**	**I/O intensive NoSQL and relational databases.**	**Big data** **Data warehouses** **Log processing**	**File Servers**	**Workloads where data is infrequently accessed**
Volume size	1 GiB - 16 TiB	4 GiB - 64 TiB	125 GiB - 16 TiB	125 GiB - 16 TiB	1 GiB-1 TiB
Max ThoughPut/volume	250 MiB/s (gp2) 1000 MiB/s (gp3)	1000 MiB/s (io1, io2) 4000 MiB/s (io2 express)	500 MiB/s	250 MiB/s	40–90 MiB/s
Max IOPS/volume	16000	64000	500	250	40 - 200

Exam tips

- EBS is block-level storage (not Object storage)
- EBS is in AZ scope. Cannot mount an instance to an EBS volume in difference AZ
- With GP2 type
  - Volume size < 5,334 GiB => IOPS = Volume size * 3
  - Volume size 5,334 GiB ~ 16TiB => IOPS = 16,000
  - Burst IOPS = 3000 IOPS for volume size < 1000 GiB (Having credit balance machenisism)

Snapshot

What is Snapshot?

- Snapshot is photograph of a disk (EBS volume)
- Snapshots are point in time copies of volumes and stored in S3 (transparent with users)
- Snapshots are incremental (Only blocks changed from previous snapshot are uploaded to S3)
- You should stop the instance before taking snapshot (for consistent data)
- Taking snapshot needs I/O, shouldn't take snapshot while instance is handling a lot of traffic
- AMI can create from a snapshot

AMI type (EBS vs Instance store)

AMI

- AMI Stand for **A**mazon **M**achine **I**mage
- AMI uses to run an EC2 instance
- AMI can select based one:
  - Region
  - Operating System
  - Architecture
  - Storage for Root Devices (EBS backed volume or Instance Store)

AMI

AMI type

EBS backed Volume

- Root device is an **Amazon EBS** volume created from an EBS snapshot
- EC2 instance can be stopped
- Data on root device is persistent

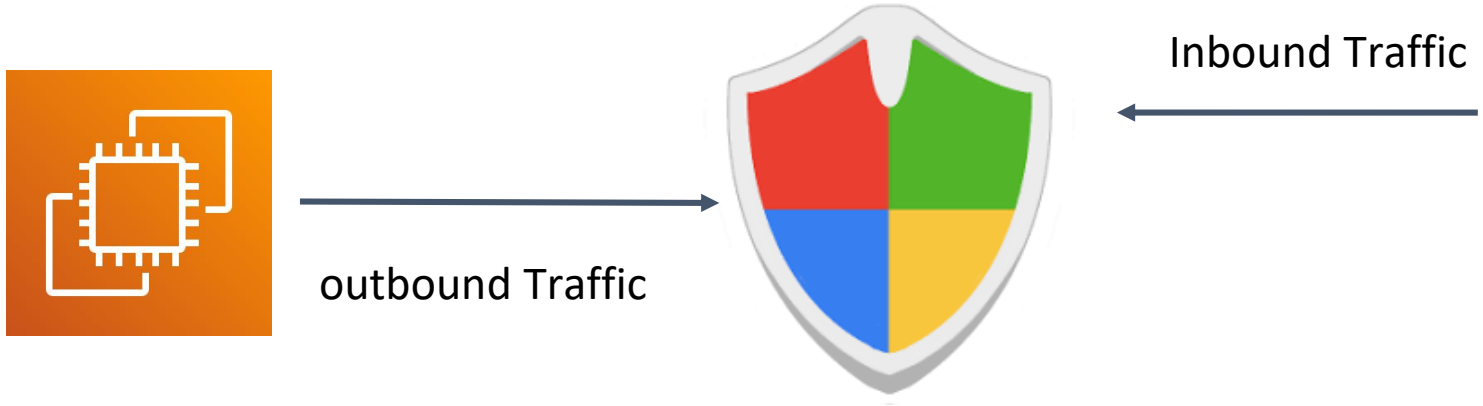
Instance store Volume

- Root device is an Amazon **Instance Store** volume created from a template stored in S3
- EC2 instance cannot be stopped
- Underlying host fails, data in Root device will be lost

Security Groups

What is Security Groups?

- Security Groups work as a virtual firewall
- Control Inbound/Outbound for EC2 instances



Security Groups features

- SGs are **STATEFUL** firewall
- Only allow rules, but not deny rules
- Cannot block specific IP address, using NACL instead
- Changes to SGs take effect immediately
- One SGs can attach to many instances, one instance can use many SGs
- By default

  - All inbound traffic is blocked

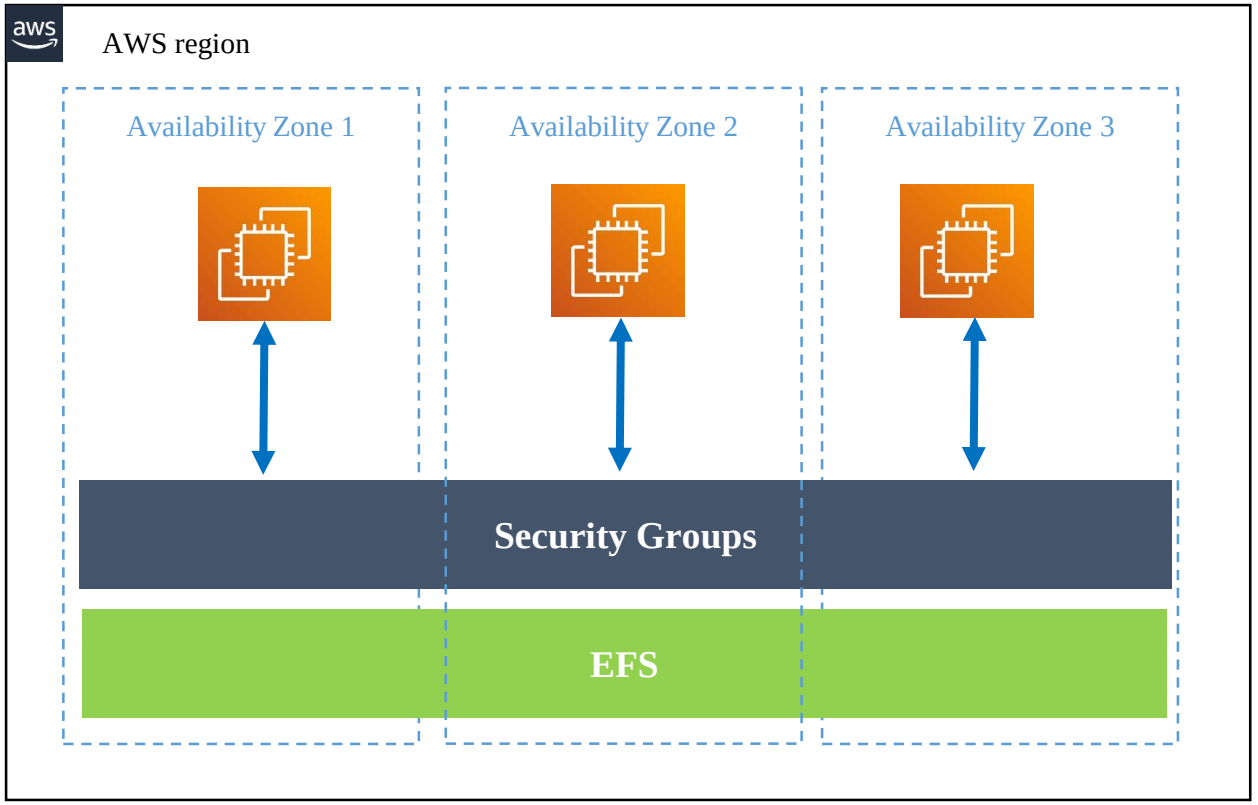
  - All outbound traffic is allowed

EFS Introduction

What is EFS? (cont.)

- EFS stands for **E**lastic **F**ile **S**ystem.
- AWS managed NFS (Network File System)
- EFS spans across multiple Avalibility Zones (regional scope)
- High Availibility, Scalable. Pay for storage usage
- EFS can attach Security Groups to control access

What is EFS?

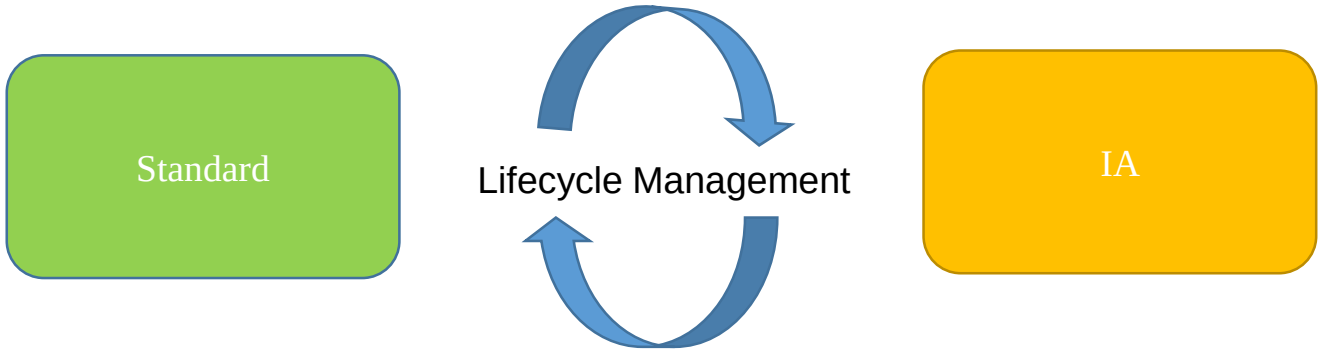

aws
AWS region
Availability Zone 1
Availability Zone 2
Availability Zone 3
Security Groups
EFS

What is EFS? (cont.)

- Using for sharing data purpose (Bigdata, wordpress, data sharing, web…)
- POSIX compliance
- Only for Linux instances (not for Window instances)
- Data can be encrypted at rest using KMS

EFS storage class

- **Standard**
  - For files frequent access
  - Pay for storage
- **Infrequent Access (IA)**
  - For files not frequent access
  - Pay for storage + retrieve files



EFS performance mode

- **General Purpose**
  - Latency-sensitive use cases
  - For web serving, CMS, home directories, and general file serving
- **Max I/O**
  - Higher throughput, parallel but higher latency
  - For high parallel, throughput workload (Bigdata Analytic, Media Proccessing)

ENI, ENA and EFA

What is ENI, ENA and EFA?

- **ENI**
  - **E**lastic **N**etwork **I**nterface – Virtual Network Card
- **ENA**
  - **E**nhanced **N**etworking **A**dapter.
  - Providing high performance networking capability for supported instance type
- **EFA**
  - **E**lastic **F**abric **A**dapter – Network device
  - Accelerate for High Performance Computing (HPC) and ML

ENI

- One primary private IP and one or more secondary private IP from VPC IP range
- Can attach multiple SGs to one ENI
- Use case
  - Host multiple websites in same instances. Each SSL for each private IP
  - Using for standby instances by detach/attach secondary ENI
  - For Network appliances (Firewall, Loadbalancer)

Elastic network adapter

ENA

- Using new network device virtualization method (SR-IOV)
- Enhance networking capability up to 100 Gbps for supported instance type
- Use case
  - Application needs high network throughput capability

Elastic network adapter

EFA

- Can attach to EC2 instance to accelerate HPC, ML.
- Support OS bypass and Support Linux instance only
- Use case
  - HPC (High Performance Computing), Machine Learning

Userdata

Userdata

- Run commands, setup config when instance starts
- Userdata script only run once at the instance first launch
- Userdata script run with root user privilege
- Use case:
  - Software update
  - Install package
  - Config (ssh, user, application)

Instance spot and Spot Fleet

Spot Instance

- Up to 90% discount compared to On-Demand prices
- Use cases
  - For Batch jobs, Data analysis, workloads are resilient to failures
  - Not good for critical job, Database

Spot Price

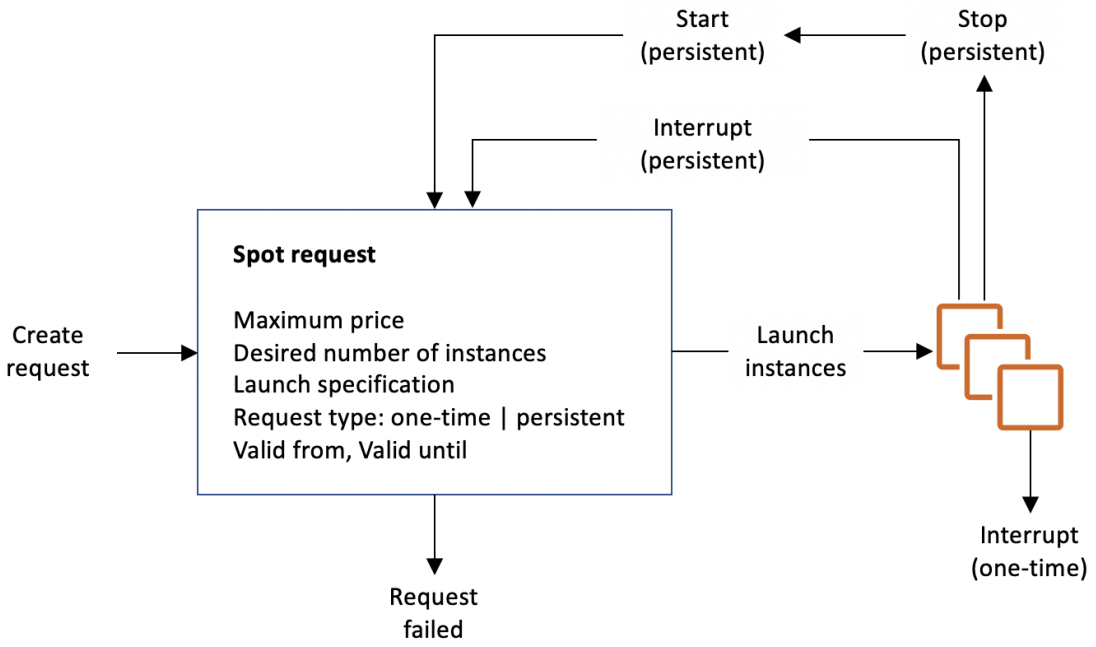
- You define **bid price** (max spot price)
- Instance launch when **current spot price** < **bid price** and
- Instance will be stopped or terminated when **current spot price** > **bid price** (2 minutes for grace period)
- Hourly **spot price** varies based on demand and capacity (the law of supply and demand)

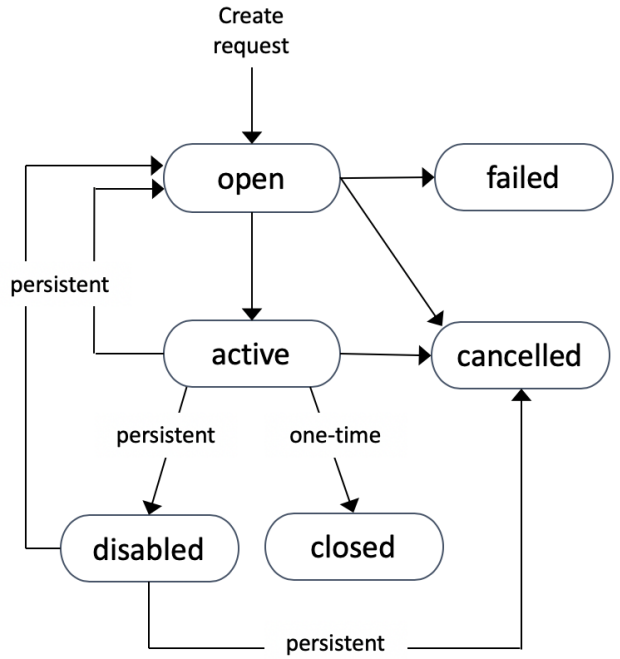
Spot Price (cont.)

Spot Block

- Using **Spot Block** to stop instance from being terminated when **bid price < spot price**
- Spot Block can be set from one to six hours

How to request Spot Instance





Spot Fleet

- Spot Fleets = collection of Spot Instances + optionally On-Demand Instances
- Launch Spot + On-Demand instances to meet target capacity with price restraints
  - You can have multiple Spot Instance pool (Same instance type, AZ, Network)
  - Spot fleet will pick pools to meet the defined strategy
  - Fleet stop launching instances when max budged is reached or capacity match

Spot Fleet (cont.)

- Allocation strategies
  - Lowest price (default): Pick the pool that has lowest price
  - Diversified: Distributed across the pools
  - Capacity Optimized: Pick the pool that has optimal capacity for number of instances

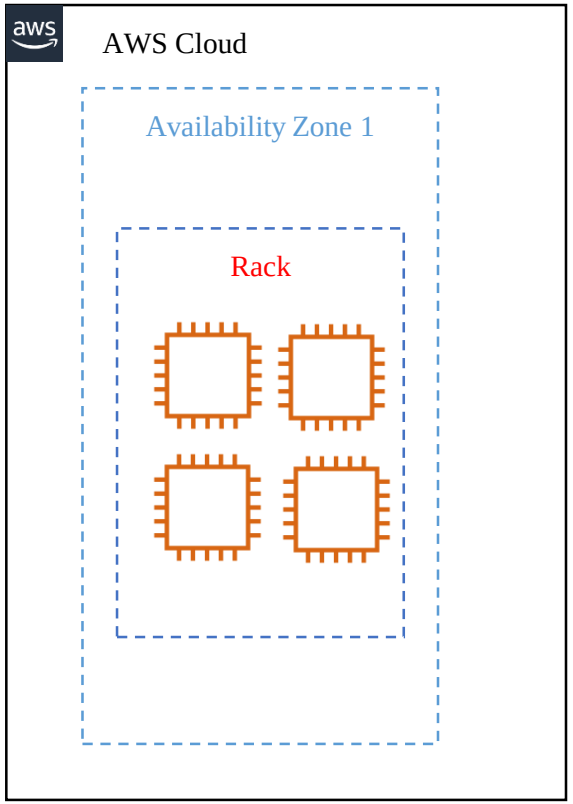
Placement Groups

Placement Groups?

- By default, EC2 service places your instances in the way to minimize correlated failures
- Placement Groups help you to control over this default way
- Using for EC2 instances are tightly-couple, need Low-latency and High Throughput
- Placement Groups Strategy
  - Cluster
  - Partition
  - Spread

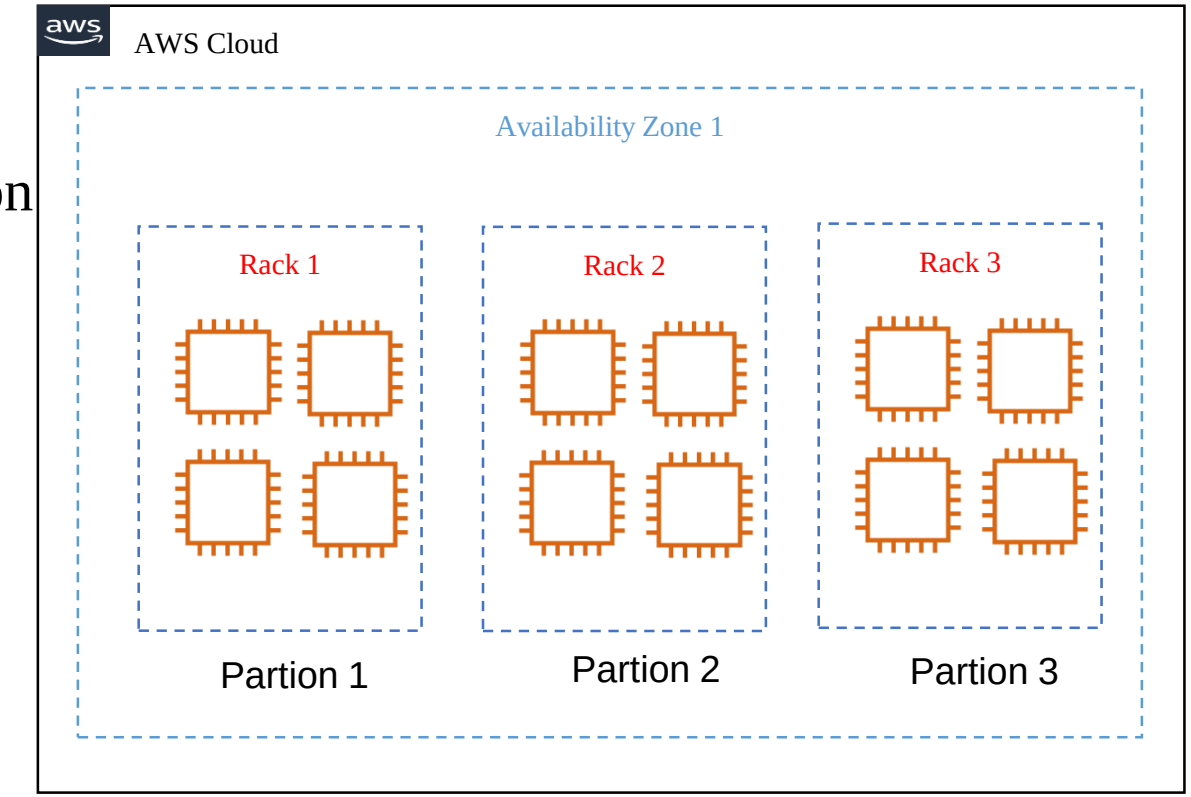
Placement Groups – Cluster

- Places all EC2 instances are same **Rack (underlying hardware)**
- High throughput (~ 10Gbps), Low latency
- Risk to loss all cluster when rack got failure
- Usa case: Bigdata, Low-latency app



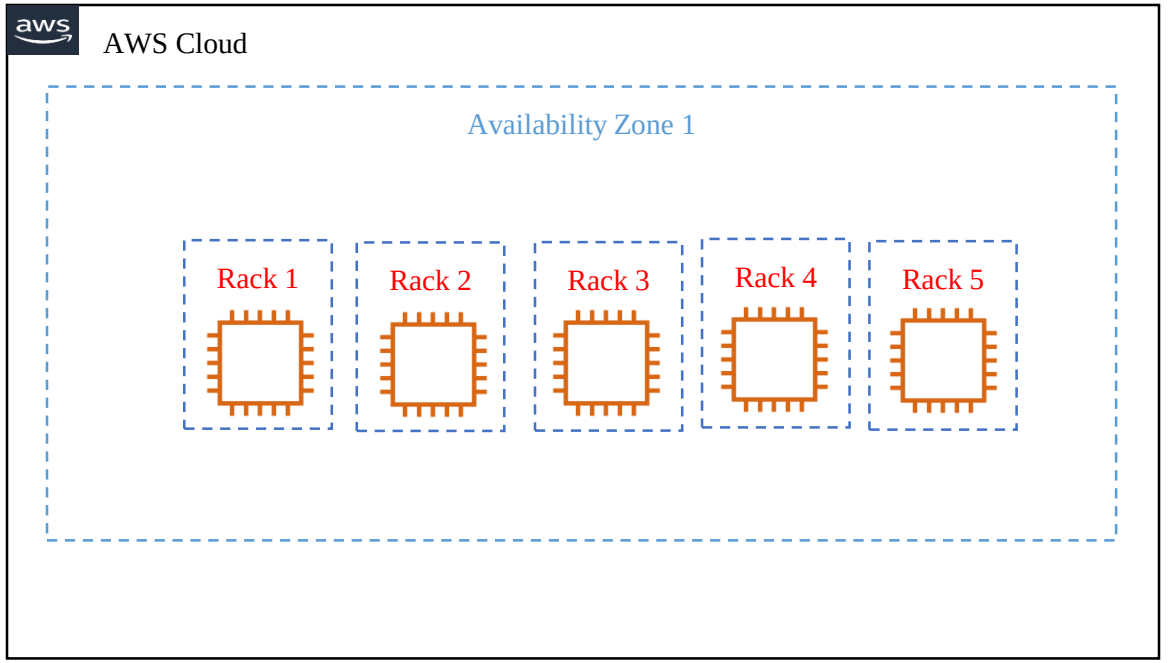
Placement Groups – Partition

- Instances in difference partion don’t share Rack
- A partition failure won’t affect other paritions
- A partition can locate in any AZs in same region
- Up to 7 parttions in one AZ
- Use case: HDFS, Cassandra…



Placement Groups – Spread

- Instances are placed in distinct Rack
- Reduces the risk of simultaneous failures
- Can span multiple AZs in the same region
- Maximum 7 instances per AZ per group
- Use case
  - Application needs maximize High Availibility
  - Critical Application, each instance failure will not allow to affect others



EC2 hibernate

What is EC2 hibernate?

When we start instances, the following happens

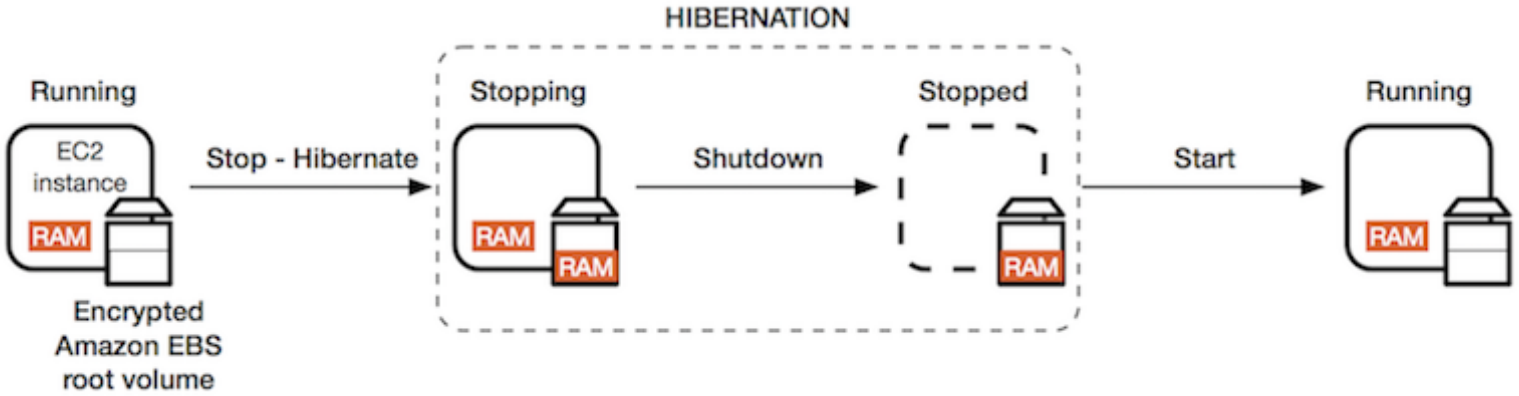
- OS boots up
- Run userdata script (in case of first start)
- Start and load applications into RAM.

What is EC2 hibernate?

- The contents from RAM is saved to EBS root volume
- Persist EBS root volume and any attached volumes
- EBS root volume must be encrypted
- The instance boot is much faster

Start instances with EC2 hibernate

- The EBS root volume is restored to its previous state
- The RAM contents are reloaded
- The processes that were previously running on the instance are resumed
- Previously attached data volumes are reattached and the instance retains its instance ID



EC2 Hibernate use case

- Long-running process
- Services that take time to initialize

Exam tips

- Instance RAM size < 150 GB
- Supported instance families: C3, C4, M3, M4, R3, R4, R5
- AMI: Amazon Linux 2, Linux AMI, Ubuntu and Window
- Root Volume must be EBS and encrypted
- Available for On-Demand and RI
- Cannot hibernated more than 60 days