# EC2 Fundamentals

Uday Manchanda

August 30, 2021

## 1 EC2 Instance Types

- Naming Convention: m5.2xlarge
  - m = instance class
  - 5 = generation
  - 2xlarge = size within instance class (larger size means more CPU and memory)
- General purpose = T, M, A
- Compute optimized = C
- Memory optimized = R, X, z
- Storage optimized = I, G, H

## 2 Security Groups

- Control how traffic is allowed in/out of EC2 instance
- only contain ALLOW rules, reference by IP or SG
- Lives "outside" the EC2 - if traffic is blocked the EC2 instance won't see it

## 3 Instance Launch Types

- On demand: short workload, predictable pricing
  - Pay for what you use
  - Highest cost but no upfront payment
  - Recommended for short term and un-interrupted workloads where you can't predict how the application will behave

- Reserved: minimum 1 year
  - 75 percent discount compared to on demand
  - Reserve a specific instance type
  - Recommended for steady state usage applications (think database)
- Spot: short workloads, cheap, can lose instances (less reliable)
  - highest discount compared to on-demand
  - can lose instances at any point if your max price is less than the current spot price
  - Most cost efficient, useful for workloads that are resilient to failure
  - like batch jobs, data analysis, image processing
- Dedicated hosts: book an entire physical server, control instance placement
  - allow you to use your existing server-bound software licenses
  - allocated for a 3 yr reservation
  - more expensive, useful for a complicated licensing model
  - if your company has specific regulatory requirements

## 3.1 Spot Instance Requests

- Can get up to 90 percent discount compared to on-demand
- Define max spot price and get the instance while current spot price ¡ max
- if that limit is breached you can stop or terminate instance
- Spot block - block the spot instance for a specified time frame (1 to 6 hr) w/out interruptions
- Batch jobs, data analysis, workloads resilient to failures
- Not great for critical jobs/databases
- How to terminate spot instance?
  - Define spot request - the specs for your instance(s)
  - can cancel spot request based on whether request type is one time or persistent
  - You must cancel the spot request that is open/active/disabled and then terminate the instances
- Spot fleets = set of spot instances + (optional) on-demand instances
- Spot fleet will try and meet target capacity with price constraints
- Spot fleets allow us to automatically request spot instances with the lowest price