TIL

Dates	@2023년 1월 12일
Type	
Topic	

TOPIC

AWS - EC2 Basic (PPT - p. 39 ~ p.75)

Summary

- EC2 Instance: AMI (OS) + Instance Size (CPU + RAM) + Storage + security groups + EC2 User Data
- Security Groups: Firewall attached to the EC2 instance
- EC2 User Data: Script launched at the first start of an instance
- SSH: start a terminal into our EC2 Instances (port 22)
- EC2 Instance Role: link to IAM roles
- Purchasing Options: On-Demand, Spot, Reserved (Standard + Convertible + Scheduled), Dedicated Host, Dedicated Instance

What is the EC2(Elastic Compute Cloud)

EC2 is a Infrastructure as a Service

It mainly consists in the capability of:

- Renting virtual machines(EC2)
- Storing data on virtual drives (EBS)
- Distributing load across machines (ELB)

- Scaling the services using an auto-scailing group(ASG)

EC2 Sizing & Configuration Options

1. Operating System : Linux, Windows, Mac
2. CPU : How much computer power & cores
3. RAM : How much random-access memory
4. Storage
    a. Network-attached(EBS & EFS)
    b. hardware(EC2 instance Store)
5. Network card : speed of the card, Public IP address
6. Firewall rules : Security Group
7. Bootstrap script : EC2 User Data (It can configure what to do at first launch)
    a. That script is **only run once** at the instance first start **with the root user**
    b. Tasks we can decide
        i. Installing updates
        ii. Installing software
        iii. Downloading common files from the internet
        iv. Anything you can think of

EC2 Instance Types

Overview

AWS has the following naming convention

ex) m5.2xlarge

- m : instance class
- 5 : generation (AWS improves them over time)
- 2xlarge : size within the instance class

General Purpose

- Great for a **diversity of workloads** such as web servers or code repositories
- **Balance** between:
  - Compute
  - Memory
  - Networking

Compute Optimized

- Great for compute-intensive tasks that require high performance processors
  - Batch processing workloads
  - Media transcoding
  - High performance web servers
  - High performance computing (HPC)
  - Scientific modeling & machine learning
  - Dedicated gaming servers

Memory Optimized

- Fast performance for workloads that process large data sets in memory
  - High performance, relation/non-relation databases
  - Distributed web scale cache stores
  - In-memory databases optimized for BI (business intelligence)
  - Applications performing real-time processing of big unstructured data

Storage Optimized

- Grate for storage-intensive tasks that require high, sequential read and write access to large data sets on local storage

- High frequency online transaction processing(OLTP) systems
- Relational & NoSQL databases
- Cache for in-memory databases(ex. Redis)
- Data warehousing applications
- Distributed file systems

Security Groups

Introduction

- Security Groups are fundamental of network security in AWS
- They control how traffic is allowed into or out of our EC2 Instances
- Security groups **only contain allow rules**
- Security groups rules can reference by IP or by security group

Deep Dive

- Security groups are acting as a ‘firewall’ on EC2 instances
- They regulate
  - Access to Ports
  - Authorised IP ranges - IPv4 and IPv6
  - Control of inbound network
  - Control of outbound network

Good to know

- Can be attached to multiple instances
- Locked down to a region/VPC combination
- Does live “outside” the EC2 – if traffic is blocked the EC2 instance won’t see it
- It’s good to maintain **one separate security group for SSH access**

- If your application is not accessible (time out), then it’s a security group issue
- If your application gives a “connection refused“ error, then it’s an application error or it’s not launched
- All inbound traffic is blocked by default
- All outbound traffic is authorised by default

EC2 Instances Purchasing Options

- **On-Demand Instances** – short workload, predictable pricing, pay by second
  - Pay for what you use
    - Linux or Windows - billing per second, after the first minute
    - All other operating systems - billing per hour
  - Has the highest cost but no upfront payment
  - No long-term commitment
  - Recommended for short-term and un-interrupted workloads, where you can't predict how the application will behave
- **Reserved (1 & 3 years)**
  - **Reserved Instances** – long workloads
    - Up to 72% discount compared to On-demand
    - You reserve a specific instance attributes (Instance Type, Region,Tenancy, OS)
    - Reservation Period – 1 year (+discount) or 3 years (+++discount)
    - Payment Options – No Upfront (+), Partial Upfront (++), All Upfront (+++)
    - Reserved Instance’s Scope – Regional or Zonal (reserve capacity in an AZ)
    - Recommended for steady-state usage applications (think database)
    - You can buy and sell in the Reserved Instance Marketplace
  - **Convertible Reserved Instances** – long workloads with flexible instances
    - Can change the EC2 instance type, instance family, OS, scope and tenancy

        - Up to 66% discount
- **Savings Plans (1 & 3 years)** –commitment to an amount of usage, long workload
    - Get a discount based on long-term usage (up to 72% - same as RIs)
    - Commit to a certain type of usage ($10/hour for 1 or 3 years)
    - Usage beyond EC2 Savings Plans is billed at the On-Demand price
    - Locked to a specific instance family & AWS region (e.g., M5 in us-east-1)
    - Flexible across
        - Instance Size (e.g., m5.xlarge, m5.2xlarge)
        - OS (e.g., Linux, Windows)
        - Tenancy (Host, Dedicated, Default)
- **Spot Instances** – short workloads, cheap, can lose instances (less reliable)
    - Can get a discount of up to 90% compared to On-demand
    - Instances that you can “lose” at any point of time if your max price is less than the
      current spot price
    - The MOST cost-efficient instances in AWS
    - Useful for workloads that are resilient to failure
        - Batch jobs
        - Data analysis
        - Image processing
        - Any distributed workloads
        - Workloads with a flexible start and end time
    - Not suitable for critical jobs or databases
- **Dedicated Hosts** – book an entire physical server, control instance placement
    - A physical server with EC2 instance capacity fully dedicated to your use
    - Allows you address compliance requirements and use your existing server-bound software licenses (per-socket, per-core, pe—VM software licenses)
    - Purchasing Options

        - On-demand – pay per second for active Dedicated Host
        - Reserved - 1 or 3 years (No Upfront,Partial Upfront,All Upfront)
    - The most expensive option
    - Useful for software that have complicated licensing model (BYOL – Bring Your Own License)
    - Or for companies that have strong regulatory or compliance needs
- **Dedicated Instances** – no other customers will share your hardware
    - Instances run on hardware that's dedicated to you
    - May share hardware with other instances in same account
    - No control over instance placement (can move hardware after Stop / Start)
- **Capacity Reservations** – reserve capacity in a specific AZ for any duration
    - Reserve On-Demand instances capacity in a specific AZ for any duration
    - You always have access to EC2 capacity when you need it
    - No time commitment (create/cancel anytime), no billing discounts
    - Combine with Regional Reserved Instances and Savings Plans to benefit from billing discounts
    - You're charged at On-Demand rate whether you run instances or not
    - Suitable for short-term, uninterrupted workloads that needs to be in a specific AZ

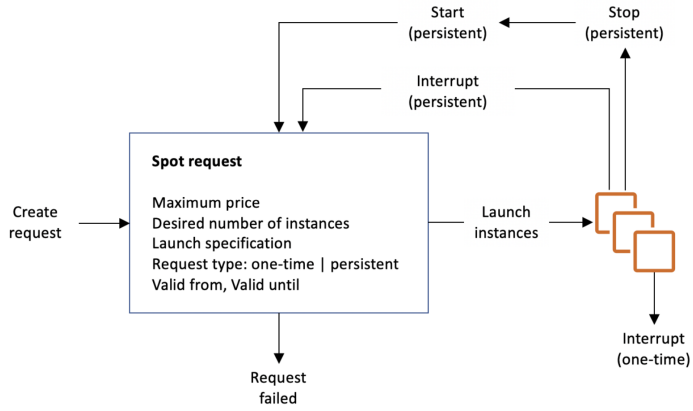
EC2 Spot Instance Requests

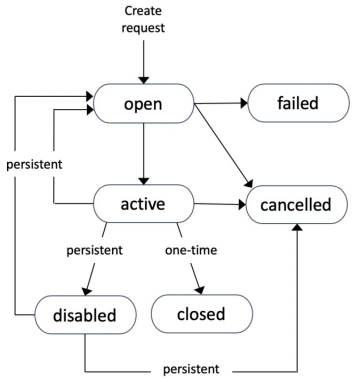
- Can get a discount of up to 90% compared to On-demand
- Define max spot price and get the instance while current spot price < max
    - The hourly spot price varies based on offer and capacity
    - If the current spot price > your max price you can choose to stop or terminate your
      instance with a 2 minutes grace period.
- Other strategy: Spot Block

- “block” spot instance during a specified time frame (1 to 6 hours) without interruptions
- In rare situations, the instance may be reclaimed

- Used for batch jobs, data analysis, or workloads that are resilient to failures.
- Not great for critical jobs or databases

How Terminate EC2 Instance





Spot Fleets

(**set of Spot Instances + (optional) On-Demand Instances)**

- The Spot Fleet will try to meet the target capacity with price constraints
  - Define possible launch pools: instance type (m5.large), OS, Availability Zone
  - Can have multiple launch pools, so that the fleet can choose
  - Spot Fleet stops launching instances when reaching capacity or max cost
- Strategies to allocate Spot Instances
  - lowestPrice: from the pool with the lowest price (cost optimization, short workload)
  - diversified: distributed across all pools (great for availability, long workloads)
  - capacityOptimized: pool with the optimal capacity for the number of instances

- Spot Fleets allow us to automatically request Spot Instances with the lowest price