Day - 03 | EC2: Amazon Elastic Compute Cloud | AWS Cloud Practitioner Certification

Created on 2024-07-17 03:59
Published on 2024-07-17 17:13

EC2: Virtual Machines

➤ What is Amazon EC2?

☞ EC2 sizing & configuration options

☞ EC2 User Data

☞ EC2 Instance Types - Overview

☞ General Purpose

☞ Compute Optimized

☞ Memory Optimized

☞ Storage Optimized

☞ EC2 Instance Types: example

➤ Introduction to Security Groups

☞ Deeper Dive

☞ Security Groups Diagram

☞ Good to know

➤ Classic Ports to know

➤ EC2 Instance Launch Types

☞ On Demand Instance

☞ Reserved Instances

☞ Savings Plans

☞ Spot Instances

☞ Dedicated Hosts

☞ Dedicated Instances

☞ Capacity Reservations

➤ Which purchasing option is right for me?

➤ Price Comparison Example – m4.large – us-east-1

➤ Shared Responsibility Model for EC2

➤ EC2 Section – Summary

What is Amazon EC2?

Amazon Elastic Compute Cloud (EC2) is a web service that provides resizable compute capacity in the cloud. It is designed to make web-scale cloud computing easier for developers. EC2 allows users to rent virtual computers on which they can run their applications.

- EC2 is one of the most popular of AWS offering
- EC2 = Elastic Compute Cloud = Infrastructure as a Service
- It mainly consists in the capability of :

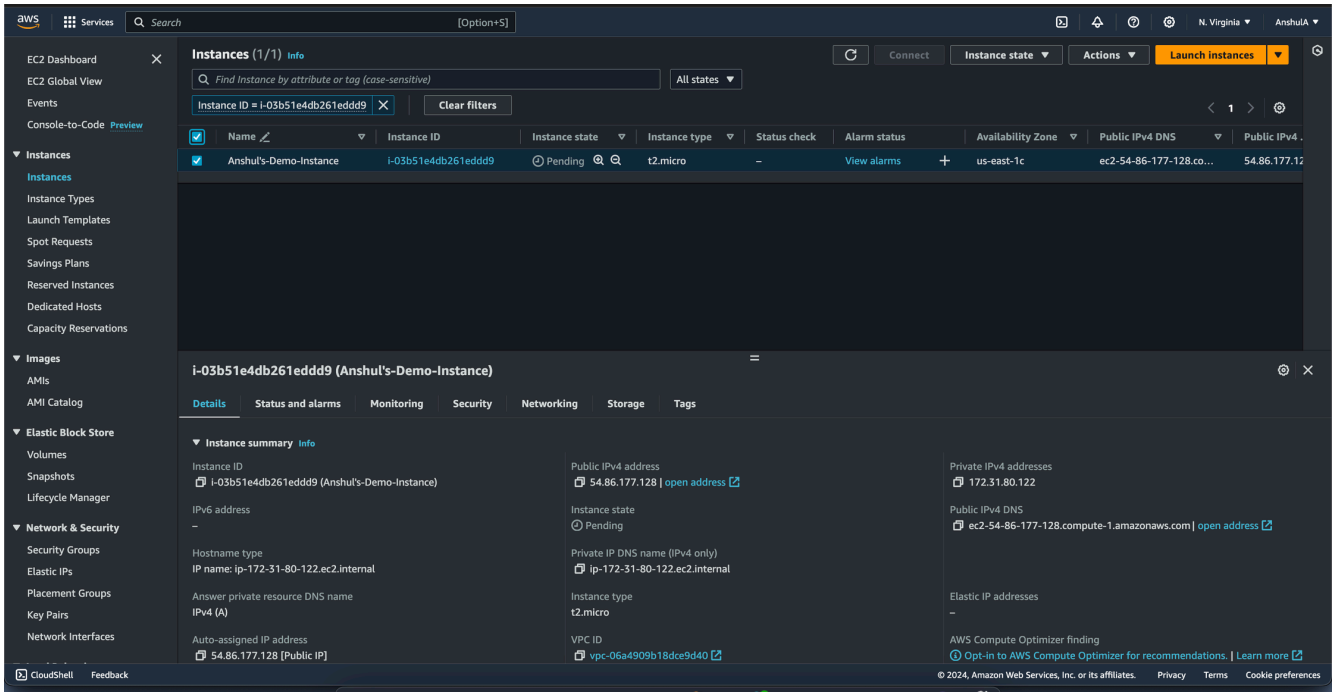
→ Renting virtual machines (EC2)

→ Storing data on virtual drives (EBS)

→ Distributing load across machines (ELB)

→ Scaling the services using an auto-scaling group (ASG)

- Knowing EC2 is fundamental to understand how the Cloud works



Amazon EC2

EC2 sizing & configuration options

- Operating System (OS): Linux, Windows or Mac OS

- How much compute power & cores (CPU)
- How much random-access memory (RAM)
- How much storage space: (a) Network-attached (EBS & EFS), (b) hardware (EC2 Instance Store)
- Network card: speed of the card, Public IP address
- Firewall rules: security group
- Bootstrap script (configure at first launch): EC2 User Data

EC2 User Data

- It is possible to bootstrap our instances using an **EC2 User data** script.
- **Bootstrapping** means launching commands when a machine starts
- That script is **only run once** at the instance **first start**
- EC2 user data is used to automate boot tasks such as:

⇾ Installing updates

⇾ Installing software

⇾ Downloading common files from the internet

⇾ Anything you can think of

- The EC2 User Data Script runs with the root user

EC2 Instance Types - Overview

EC2 instance types are categorized into different families based on their intended use cases:

- **General Purpose**: Balanced resources for a variety of workloads.
- **Compute Optimized**: High-performance processors for compute-intensive applications.
- **Memory Optimized**: Large amounts of memory for memory-intensive applications.
- **Storage Optimized**: High-performance local storage for I/O-intensive applications.

⇾ AWS has the following naming convention: m5.2xlarge

⇾ m: instance class

⇾ 5: generation (AWS improves them over time)

⇾ 2xlarge: size within the instance class

General Purpose

General Purpose instances provide a balance of compute, memory, and networking resources and can be used for a variety of diverse workloads. Examples include:

- **t2, t3**: Burstable performance instances.
- **m4, m5**: Balance of compute, memory, and network resources.

Compute Optimized

Compute Optimized instances are ideal for compute-bound applications that benefit from high-performance processors. Examples include:

- **c4, c5**: High CPU performance.
- Batch processing workloads, Media transcoding, High performance web servers, High performance computing (HPC), Scientific modeling & machine learning, Dedicated gaming servers

Memory Optimized

Memory Optimized instances are designed for workloads that require large amounts of memory. Examples include:

- **r4, r5**: High memory performance.
- **x1, x1e**: Extreme memory instances.
- High performance, relational/non-relational 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

Storage Optimized instances are designed for workloads that require high, sequential read and write access to large datasets on local storage. Examples include:

- **i3**: High I/O performance.
- **d2**: Dense storage instances.

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

EC2 Instance Types: example

Instance	vCPU	Mem (GiB)	Storage	Network Performance	EBS Bandwidth (Mbps)
t2.micro	1	1	EBS-Only	Low to Moderate	
t2.xlarge	4	16	EBS-Only	Moderate	
c5d.4xlarge	16	32	1 x 400 NVMe SSD	Up to 10 Gbps	4,750
r5.16xlarge	64	512	EBS Only	20 Gbps	13,600
m5.8xlarge	32	128	EBS Only	10 Gbps	6,800

t2.micro is part of the AWS free tier (up to 750 hours per month)

EC2 Instance Types: example

Introduction to Security Groups

Security groups act as virtual firewalls for your EC2 instances to control inbound and outbound traffic. You can specify allowed protocols, ports, and source IP ranges.

Deeper Dive

- Security groups are acting as a “firewall” on EC2 instances
- They regulate:

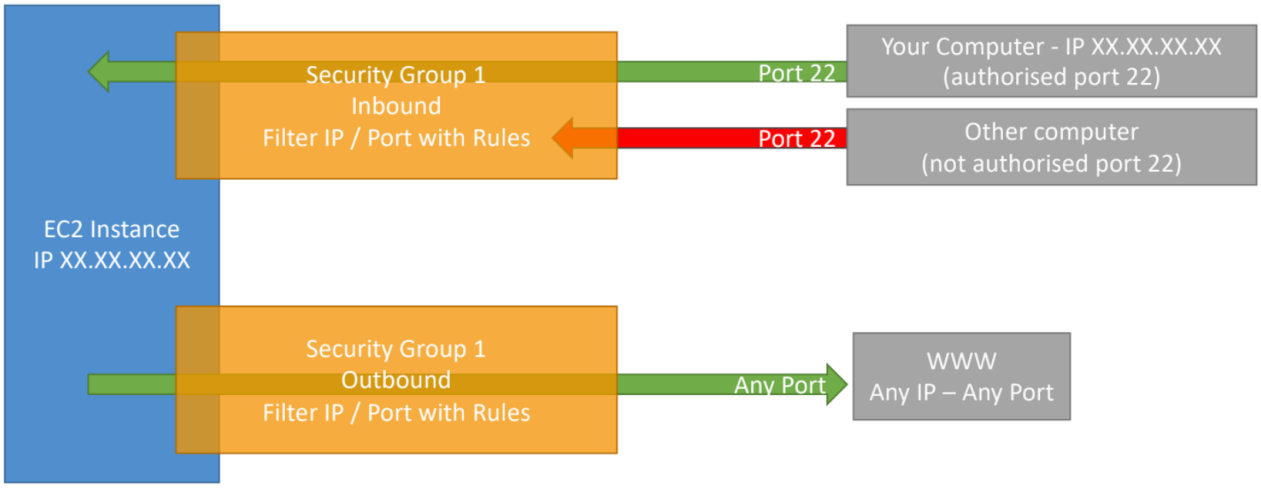
→ Access to Ports

→ Authorised IP ranges – IPv4 and IPv6

→ Control of inbound network (from other to the instance)

→ Control of outbound network (from the instance to other)

Security Groups Diagram



Security Groups Diagram

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 **authorized** by default

Classic Ports to Know

- **22 = SSH (Secure Shell) -** log into a Linux instance
- **21 = FTP (File Transfer Protocol)** – upload files into a file share
- **22 = SFTP (Secure File Transfer Protocol)** – upload files using SSH
- **80 = HTTP** – access unsecured websites
- **443 = HTTPS** – access secured websites
- **3389 = RDP (Remote Desktop Protocol)** – log into a Windows instance

---

EC2 Instance Launch Types

There are several purchasing options for EC2 instances, each suited for different use cases:

- **On Demand Instances**: Pay for compute capacity by the hour or second with no long-term commitments.
- **Reserved Instances**: Offer significant discounts (up to 75%) compared to On-Demand pricing. Requires a commitment of 1 or 3 years.
- **Savings Plans**: Flexible pricing model that offers significant savings compared to On-Demand, in exchange for a commitment to a consistent amount of usage (measured in $/hour) for 1 or 3 years.
- **Spot Instances**: Allow you to bid for unused EC2 capacity at a reduced cost. Instances can be terminated by AWS if the spot price exceeds your bid.

- **Dedicated Hosts**: Physical servers with EC2 instance capacity fully dedicated to your use.
- **Dedicated Instances**: Run in a VPC on hardware that's dedicated to a single customer.
- **Capacity Reservations**: Reserve capacity for your instances in a specific Availability Zone for any duration.

On Demand Instance

- 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 Instances

- 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 Instance

→ Can change the EC2 instance type, instance family, OS, scope and tenancy

→ Up to 66% discount

Savings Plans

- 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

- 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 jobsData analysisImage processing

⇾ Any distributed workloads

⇾ Workloads with a flexible start and end time

- Not suitable for critical jobs or databases

Dedicated Hosts

- A physical server with EC2 instance capacity fully dedicated to your use
- Allows you to 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

- 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 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

➤ Which purchasing option is right for me?

- **On Demand:** coming and staying in resort whenever we like, we pay the full price
- **Reserved:** like planning ahead and if we plan to stay for a long time, we may get a good discount.

- **Savings Plans:** pay a certain amount per hour for certain period and stay in any room type (e.g., King, Suite, Sea View, …)
- **Spot instances:** the hotel allows people to bid for the empty rooms and the highest bidder keeps the rooms. You can get kicked out at any time
- **Dedicated Hosts:** We book an entire building of the resort
- **Capacity Reservations:** you book a room for a period with full price even you don't stay in it

---

➤ Price Comparison Example – m4.large – us-east-1

Price Type	Price (per hour)
On-Demand	$0.10
Spot Instance (Spot Price)	$0.038 - $0.039 (up to 61% off)
Reserved Instance (1 year)	$0.062 (No Upfront) - $0.058 (All Upfront)
Reserved Instance (3 years)	$0.043 (No Upfront) - $0.037 (All Upfront)
EC2 Savings Plan (1 year)	$0.062 (No Upfront) - $0.058 (All Upfront)
Reserved Convertible Instance (1 year)	$0.071 (No Upfront) - $0.066 (All Upfront)
Dedicated Host	On-Demand Price
Dedicated Host Reservation	Up to 70% off
Capacity Reservations	On-Demand Price

Price Comparison Example

➤ Shared Responsibility Model for EC2

- **AWS**: Responsible for protecting the infrastructure that runs all the services offered in the AWS Cloud.

- **Customer**: Responsible for managing their data, identity, and access management, operating system, and network configurations.

AWS	USER
Infrastructure (global network security)	Security Groups rules
Isolation on physical hosts	Operating-system patches and updates
Replacing faulty hardware	Software and utilities installed on the EC2 instance
Compliance validation	IAM Roles assigned to EC2 & IAM user access management, Data security on your instance

Shared Responsibility Model for EC2

➤ EC2 Section – 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

Happy Learning !