UNIT 5: Compute- Hands-on Session

## Amazon EC2

- 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

## 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:
  - Network-attached (EBS & EFS)
  - 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

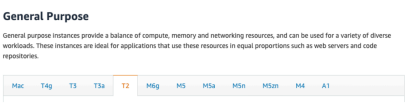
## EC2 Instance Types - Overview

- You can use different types of EC2 instances that are optimised for
- different use cases (https://aws.amazon.com/ec2/instance-types/)
- AWS has the following naming convention:
  m5.2xlarge
  - m: instance class
  - 5: generation (AWS improves them over time)
  - 2xlarge: size within the instance class

## EC2 Instance Types – General Purpose

Great for a diversity of workloads such as web servers or code repositories

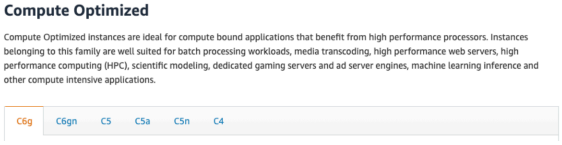
- Balance between:
  - Compute
  - Memory
  - Networking



## EC2 Instance Types – Compute Optimized

Great for compute-intensive tasks that require high performance processors:

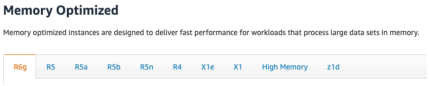
- Batch processing workloads
- Media transcoding
- High performance web servers
- High performance computing (HPC)
- Scientific modeling & machine learning
- Dedicated gaming servers




## EC2 Instance Types – Memory Optimized

Fast performance for workloads that process large data sets in memory

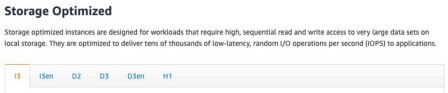
- Use cases:
  - High performance, relational/non-relational databases
  - Distributed web scale cache stores
  - In-memory databases optimized for BI (business intelligence)
  - Applications performing real-time processing of big unstructured data




## EC2 Instance Types – Storage Optimized

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

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



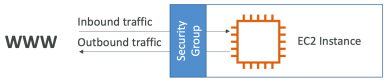

## EC2 Instance Types: example

| 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)

## Introduction to Security Groups

- Security Groups are the fundamental of network security in AWS
- They control how traffic is allowed into or out of our EC2 Instances.



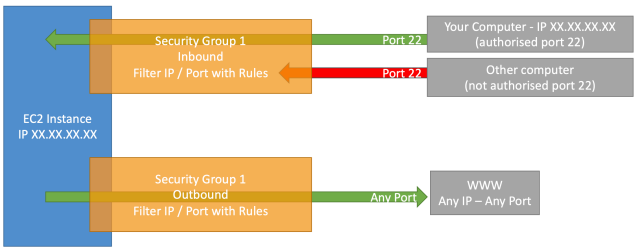

- Security groups only contain rules
- Security groups rules can reference by IP or by security group

## Security Groups Deeper Dive

- Security groups are acting as a “firewall” on EC2 instances
- They regulate:
  - Access to Ports
  - Authorized IP ranges – IPv4 and IPv6
  - Control of inbound network (from other to the instance)
  - Control of outbound network (from the instance to other)

| Type | Protocol | Port Range | Source | Description |
|---|---|---|---|---|
| HTTP | TCP | 80 | 0.0.0.0/0 | test http page |
| SSH | TCP | 22 | 122.149.196.85/32 | |
| Custom TCP Rule | TCP | 4567 | 0.0.0.0/0 | java app |

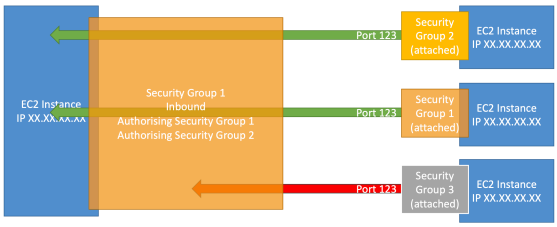
## Security Groups Diagram




## Security Groups:

- Can be attached to multiple instances
- Locked down to a region / VPC combination
- 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

## Referencing other security groups Diagram




## 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 Connect

- Connect to your EC2 instance within your browser
- No need to use your key file that was downloaded
- The “magic” is that a temporary key is uploaded onto EC2 by AWS

- Works only out-of-the-box with Amazon Linux 2
- Need to make sure the port 22 is still opened!

### EC2 Instances Purchasing Options

- On-Demand Instances – short workload, predictable pricing, pay by second
- Reserved (1 & 3 years)
- Reserved Instances – long workloads
- Convertible Reserved Instances – long workloads with flexible instances
- Savings Plans (1 & 3 years) –commitment to an amount of usage, long workload
- Spot Instances – short workloads, cheap, can lose instances (less reliable)
- Dedicated Hosts – book an entire physical server, control instance placement
- Dedicated Instances – no other customers will share your hardware
- Capacity Reservations – reserve capacity in a specific AZ for any duration

### EC2 On Demand

- 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

### EC2 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

- Commitment: 1 or 3 years
- Scope: Spend-based (applies across multiple instances, regions, and even AWS services)
- Flexibility: High (applies to different instance types, sizes, OS, and even services like Lambda & Fargate)
- Discount: Up to 66% based on committed spend per hour
- Types:
  - Compute Savings Plan ⟶ Works across any EC2 instance, Lambda, and Fargate
  - EC2 Instance Savings Plan ⟶ Specific to an EC2 family but allows flexibility in region & OS

Best for: Users who want more flexibility and don't want to be locked into a specific instance type.

**Key Differences**

| Feature | Reserved Instances (RIs) | Savings Plans |
|---|---|---|

| | | |
|---|---|---|
| **Commitment** | 1 or 3 years | 1 or 3 years |
| **Discount** | Up to 72% | Up to 66% |
| **Flexibility** | Low (instance-specific) | High (instance-family and service-wide) |
| **Scope** | Single instance type & region | Any instance, region, or even Lambda/Fargate |
| **Use Case** | Fixed, predictable workloads | Dynamic workloads with changing needs |

## EC2 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 jobs
  - Data analysis
  - Image processing
  - Any distributed workloads
  - Workloads with a flexible start and end time
  - Not suitable for critical jobs or databases

## EC2 Dedicated Hosts

• **Physical Server Control**: You get an entire physical server dedicated to your use.
• **Instance Placement**: You can control instance placement on the host.
• **Licensing Benefits**: Allows you to use your **own software licenses** (e.g., Windows, SQL Server) that require dedicated hardware.
• **Billing**: Charged **per host** (not per instance), regardless of how many instances you run.
• **Use Case**: Ideal for organizations needing **hardware-level isolation, compliance requirements, or specific licensing**.

## EC2 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)

## Shared Responsibility Model for EC2


- Infrastructure (global network security)
- Isolation on physical hosts
- Replacing faulty hardware
- Compliance validation

- Security Groups rules
- Operating-system patches and updates
- Software and utilities installed on the EC2 instance
- IAM Roles assigned to EC2 & IAM user access management
- Data security on your instance

## 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), Dedicated Host, Dedicated Instance

EC2 Instance Storage Section - EC2 Instance Storage Section

## EC2 Instance Types - EC2 Instance Types

• Generating custom Public Key and Private keys for EC2 instances
• Security groups
• Volumes and Snapshots
• Creating customized Amazon Machine Images

---

## Amazon EC2 - RAID Overview & RAID Configurations

RAID (**Redundant Array of Independent Disks**) is used in **Amazon EC2 instances** to improve **performance, fault tolerance, or both**. Since AWS **EBS (Elastic Block Store) volumes** do not support built-in RAID configurations, you need to configure RAID at the OS level within an EC2 instance.

## User Data and MetaData - User Data and Metadata

## Elastic Load Balancers & Health Checks - Elastic Load Balancers & Health Checks

• Auto Scaling Groups

• CloudWatch
• Creating Billing Alarm and EC2 instance alarms.
• AWS CLI&EC2 Roles
• Elastic File System
• AWS LightSail
• Elastic Beanstalk