# **Elastic Compute Cloud**

Amazon EC2: Amazon Elastic Compute Cloud (Amazon EC2) is a web service that provides resizable compute capacity in the cloud. Amazon EC2 reduces the time required to obtain and boot new server instances to minutes, allowing you to quickly scale capacity, bout up and down, as your computing requirements change.

* It basically allows you to use based on the capacity of usage that you use for.
* There are different pricing options for EC2:
  + On Demand – allows you to pay a fixed rate by the hour (or by the second) with no commitment.
  + Reserved – Provide you with a capacity reservation, and offer a significant discount on the hourly charge for an instance. 1 Year or 3 Year Terms.
  + Spot – Enables you to bid whatever price you want for instance capacity, providing for even greater savings if your applications have flexible start and end times.
    - Eg: you can think of purchasing a stock in a stock market like you purchase when the stock goes to 90$ from 100$ and you want to close the sale when it reaches to 100$ again.
  + Dedicated Hosts – Physical EC2 Server dedicated for your use. Dedicated Hosts can help you reduce costs by allowing you to use your existing server-bound software licenses.
    - You can actually bring down your licenses to dedicated hosts for example for oracle, VM Ware, and SQL Server to reduce the cost and can use it on your dedicated host.
* Let’s get on to different use cases for all these type of pricings:
  + **On Demand:**
    - Perfect for users that want the low cost and flexibility of Amazon EC2 without any up-front payment or long-term commitment.
    - Applications with short term, spiky, or unpredictable workloads that cannot be interrupted.
    - Applications being developed or tested on Amazon EC2 for the first time
    - Basically in this course we will be using on Demand services
  + **Reserved:**
    - Applications with steady state or predictable usage, this may be like your web service for example.
    - Application that require reserved capacity.
    - Users can make up-front payments to reduce their total computing costs even further.
      * Standard RIs(Up to 75% off on-demand)
      * Convertible RIs (Up to 54% off on-demand) feature the capability to change the attributes of the RI as long as the exchange results in the creation of Reserved Instances of equal or greater value.
      * Scheduled RIs are available to launch within the time window you reserve. This option allows you to match your capacity reservation to a predictable recurring schedule that only requires a fraction of a day, a week, or a month.
  + **Spot:**
    - Applications that have flexible start and end times
    - Applications that are only feasible at very low compute prices.
      * Lot of genome (the branch of molecular biology concerned with the structure, function, evolution, and mapping of genomes.) companies and pharmaceutical companies use this kind of approach like 4 am on a Sunday morning that will reduce them to use same computing capacity whenever they need. There are lots of white papers out there on spot instances if you are interested you can go read out there.
    - Users with an urgent need for large amounts of additional computing capacity.
  + **Dedicated Hosts:**
    - This is useful for regulatory requirements that may not support multi-tenant virtualization.
      * Eg: if you are working in healthcare your country may say you cannot use multi-tenant virtualization potentially exposing your data to somebody else. You know that legislation always keeps up the technology.
    - Great for licensing which does not support multi-tenancy or cloud deployments.
    - It can be purchased on-demand(hourly)
    - It can be purchased as a reservation for up to 70% off the on-Demand Price.
* Let’s move on to different EC2 Instance Types:
  + You don’t have to memorise for this course it’s just that you need to know its only used at professional level.

|  |  |  |
| --- | --- | --- |
| Family | Speciality | Use Case |
| F1 | Field Programmable Gate Array | Genomics research, financial analytics, real-time video processing, big data etc. |
| I3 | High Speed Storage | NoSQL, DBs, Data Warehousing etc. |
| G3 | Graphics Intensive | Video Encoding/3D Application Streaming |
| H1 | High Disk Throughput | MapReduce-based workloads, distributed file systems such as HDF5 and MapR-F5 |
| T2 | Lowest Cost, General Purpose | Web Servers/ Small DBs |
| D2 | Dense Storage | Fileservers/Data Warehousing/Hadoop |
| R4 | Memory Optimized | Memory Intensive Apps/DBs |
| M5 | General Purpose | Application Servers |
| C5 | Compute Optimized | CPU Intensive Apps/DBs |
| P3 | Graphics/General Purpose GPU | Machine Learning, Bit Coin Mining etc |
| X1 | Memory Optimized | SAP HANA/Apache Spark etc |