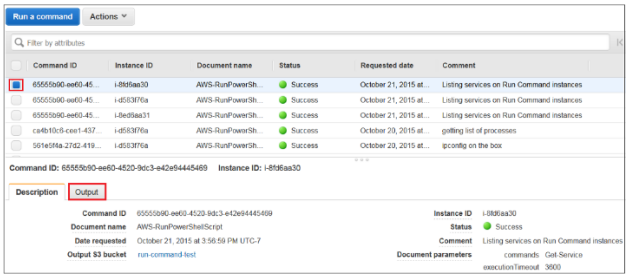
AMAZON EC2

WHAT IS EC2

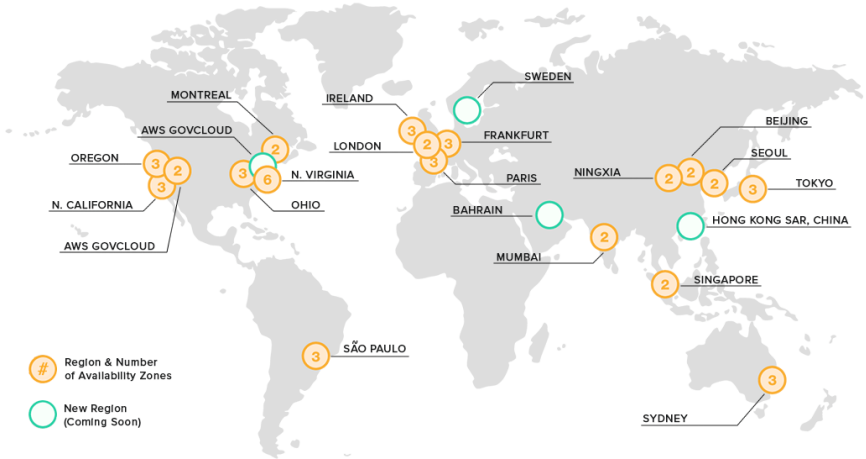
Amazon Elastic Compute Cloud (Amazon **EC2**) is a web service that provides resizable compute capacity in the cloud. Amazon EC2 reduces the time required to obtain and boot new server instances to minutes, allowing you to quickly scale capacity, both up and down, as your computing requirements change.

The following image shows the physical backbone of the AWS EC2 datacenters:

The following image shows how it looks from the AWS web console:



The following image shows AWS different datacenters spread across the globe:



An EC2 instance is a virtual server in Amazon's Elastic Compute Cloud (EC2) for running applications on the Amazon Web Services (AWS) infrastructure. It can be your virtual Linux or Windows server, and you can provision them on demand.

In the past the system administrators would have to order in advance the servers and network equipment required for the software, rack them up, connect them and install and configure them manually, configure security for them, networking, monitoring, high availability and scaling, a thing which could have taken months and sometimes even over a year to accomplish depending on the organization. Not only that, you would have to have a good estimate regarding to the compute capacity needed. If you would order too many servers it would be a waste, too few, then the software will not have enough compute power and you would then have to reorder more servers, hard drives, memory etc. It used to be a very frustrating process.

Then cloud computing came along, and changed everything. With AWS EC2, initially started off first in the year of 2006, you can provision servers on demand, that is, launch servers when you need them, and turn them off when they're no longer needed. Instead of paying for an entire server upfront, you pay only for the time you use it.

This was a game changer, allowing startups like Airbnb and Uber concentrate on developing and experimenting with their software without having to spend huge budgets on infrastructure. They could just pay by the hour for their server usage, and this literally changed the world.

Amazon EC2 changes the economics of computing by allowing you to pay only for capacity that you actually use. Amazon EC2 provides developers the tools to build failure resilient applications and isolate themselves from common failure scenarios. We will see how it's done in the labs to come.

EC2 INSTANCE PRICING OPTIONS

On Demand instances – allow you to pay a fixed rate by the hour (or by the second) with no commitment. On demand instances are designed mainly for:

- users who want both the low cost and the 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.

Reserved instances – provide you with a capacity reservation, and offer a significant discount on the hourly charge for an instance. 1 year or 3 years terms. Basically, it's paying upfront. The more you pay the more you get. Reserved instances are used for:

- Applications with steady state or predictable usage
- Applications that require reserved capacity
- Users are able to make upfront payments to reduce their total computing costs even further
    - Standard RI's – up to 75% off on demand
    - Convertible RI's – Up to 54% off on demand (e.g: convert Linux to Windows)

    - Scheduled RI's - available to launch in a predictable recurring schedule that only requires a fraction of a day, week or a month

Spot – enable you to bid whatever price you want for instance capacity, providing for even greater savings if your application has flexible start and end times. The price is determined by supply and demand of such instances and you can set a bid price of how much you're willing to pay by the hour. If the spot price is below your bid, then the instance will be launched. If it goes higher than your bid price, that instance will be terminated. Remember that if you terminate the spot instance you will pay for that hour. However, if AWS terminates it you get that hour for free. Sport instances are used mainly for:

- Applications that have flexible start and end times
- Application that are only feasible at very low compute prices (e.g: data processing on 3am on a Sunday morning)

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. They are used mainly for:

- Regulatory requirements that may not support multi-tenant virtualization
- Great for licensing which does not support multi-tenancy or cloud deployments
- Can be purchased on-demand (hourly.)
- Can be purchased as a reservation for up to 70% off the on-demand prices

You will not have to remember of the different pricing plans which may change anyway.

EC2 INSTANCE TYPES

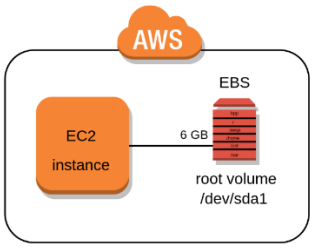
The following table describes the different instance types you can create:

Family	Speciality	Use case
D2	Dense Storage	Fileservers/Data Warehousing/Hadoop
R4	Memory Optimized	Memory Intensive Apps/DBs
M4	General Purpose	Application Servers
C4	Compute Optimized	CPU Intensive Apps/DBs
G2	Graphics Intensive	Video Encoding/ 3D Application Streaming
I2	High Speed Storage	NoSQL DBs, Data Warehousing etc
F1	Field Programmable Gate Array	Hardware acceleration for your code.
T2	Lowest Cost, General Purpose	Web Servers/Small DBs
P2	Graphics/General Purpose GPU	Machine Learning, Bit Coin Mining etc
X1	Memory Optimized	SAP HANA/Apache Spark etc

The type of instance you will choose depends on what it is that you're aiming to do with it.
For instance, D2 stands for Dense Storage, and the number "2" is the generation of that type (D1, D2 etc.). It will be used for Fileservers, Data warehousing or Hadoop etc. You don't need to remember all of them but it's important to understand them.

WHAT IS EBS?

Amazon Elastic Block Store (Amazon **EBS**) allows you to create storage volumes and attach them to Amazon EC2 instances. Once attached, you can create a file system on top of these volumes, run a database, or use them in any other way you would use a block device.



Amazon **EBS** volumes are placed in a specific Availability Zone, where they are automatically replicated to protect you from the failure of a single component. This means that the data is redundant so if a storage fails in the AWS datacenter then your data will not be lost.

EBS is a block-based storage as opposed to S3 which is an object-based storage so you can install an operating system, databases and applications on it, or in any other way you would use any block device.

You cannot mount one **EBS** volume to multiple EC2 instances, instead use **EFS**.

EBS VOLUME TYPES

SSD, General Purpose - GP2

- General Purpose, balances both price and performance
- Ratio of 3 IOPS per GB with up to 10,000 IOPS and the ability to burst up to 3000 IOPS for extended periods of time for volumes at 3334 GiB and above

SSD, Provisioned IOPS - IO1

- Designed for I/O intensive applications such as large relational or NoSQL databases
- Use if you need more than 10,000 IOPS.
- Can provision up to 20,000 IOPS per volume.

HDD, Throughput Optimized - ST1

- Sequential Data
    - Big data
    - Data warehouses
    - Log processing
- Cannot be a boot volume

HDD, Cold - SC1

- Lowest Cost Storage for infrequently accessed workloads
- File Server
- Cannot be a boot volume

HDD, Magnetic - Standard

- Lowest cost per gigabyte of all EBS volume types that is bootable. Magnetic volumes are ideal for workloads where data is accessed infrequently, and applications where the lowest storage cost is important.

SUMMARY

In this lecture we have learned what is EC2. We have covered the differences between On Demand, Spot, Reserved and Dedicated Hosts. We have covered the different type of instances. We have learned about EBS volumes, we covered the different options for EBS volumes, some as SSD volumes, and others as magnetic volumes.