3. Elastic Compute Cloud

3. Elastic Compute Cloud

EC2 stands for Amazon **Elastic Compute Cloud**.

Your business, whether it be in healthcare, manufacturing, insurance, or delivering video content to millions of users all around the world, are also using this model to deliver products, resources, or data to your end users. And you're going to need servers to power your business and your applications. You need raw compute capacity to host your applications and provide the compute power that your business needs. When you're working with AWS, those servers are virtual. And the service you use to gain access to virtual servers is called EC2.

Using EC2 for compute is **highly flexible, cost effective, and quick when you compare it to running your own servers on premises** in a datacenter that you own. The time and money it takes to get up and running with on-premises resources is fairly high. When you own your own fleet of physical servers, you first have to do a bunch of research to see what type of servers you want to buy and how many you'll need. Then you purchase that hardware up front. You'll wait for multiple weeks or months for a vendor to deliver those servers to you. You then take them to a datacenter that you own or rent to install them, rack and stack them, and wire them all up. Then you make sure that they are secure and powered up and then they're ready to be used. Only then can you begin to host your applications on top of these servers. The worst part is, once you buy these servers you are stuck with them whether you use them or not.

With EC2, it's much easier to get started. AWS took care of the hard part for you already. AWS already built and secured the datacenters. AWS has already bought the servers, racked and stacked them, and they are already online ready to be used. AWS is constantly operating a massive amount of compute capacity. And you can use whatever portion of that capacity when you need it. All you have to do is request the EC2 instances you want and they will launch and boot up, ready to be used within a few minutes. Once you're done, you can easily stop or terminate the EC2 instances. You're not locked in or stuck with servers that you don't need or want. Your usage of EC2 instances can vary greatly over time. And you only pay for what you use. Because **with EC2, you only pay for running instances, not stopped or terminated instances.**

EC2 runs on top of physical host machines managed by AWS using virtualization technology. When you spin up an EC2 instance, you aren't necessarily taking an entire host to yourself. Instead, you are sharing the host with multiple other instances, otherwise known as virtual machines. And a hypervisor running on the host machine is responsible for sharing the underlying physical resources between the virtual machines. This idea of sharing underlying hardware is called multitenancy. The hypervisor is responsible for coordinating this multitenancy and it is managed by AWS. The hypervisor is responsible for isolating the virtual machines from each other as they share resources from the host. This means EC2 instances are secure. Even though they may be sharing resources, one EC2 instance is not aware of any other EC2 instances also on that host. They are secure and separate from each other.

Amazon Elastic Compute Cloud (Amazon EC2)

Amazon Elastic Compute Cloud (Amazon EC2) provides secure, resizable compute capacity in the cloud as Amazon EC2 instances.

Imagine you are responsible for the architecture of your company's resources and need to support new websites. With traditional on-premises resources, you have to do the following:

- Spend money upfront to purchase hardware.
- Wait for the servers to be delivered to you.
- Install the servers in your physical datacenter.
- Make all the necessary configurations.

By comparison, with an Amazon EC2 instance you can use a virtual server to run applications in the AWS Cloud.

- You can provision and launch an Amazon EC2 instance within minutes.
- You can stop using it when you have finished running a workload.
- You pay only for the compute time you use when an instance is running, not when it is stopped or terminated.
- You can save costs by paying only for server capacity that you need or want.

How Amazon EC2 works

To learn more, select each marker.

1   2   3

Launch   Connect   Use
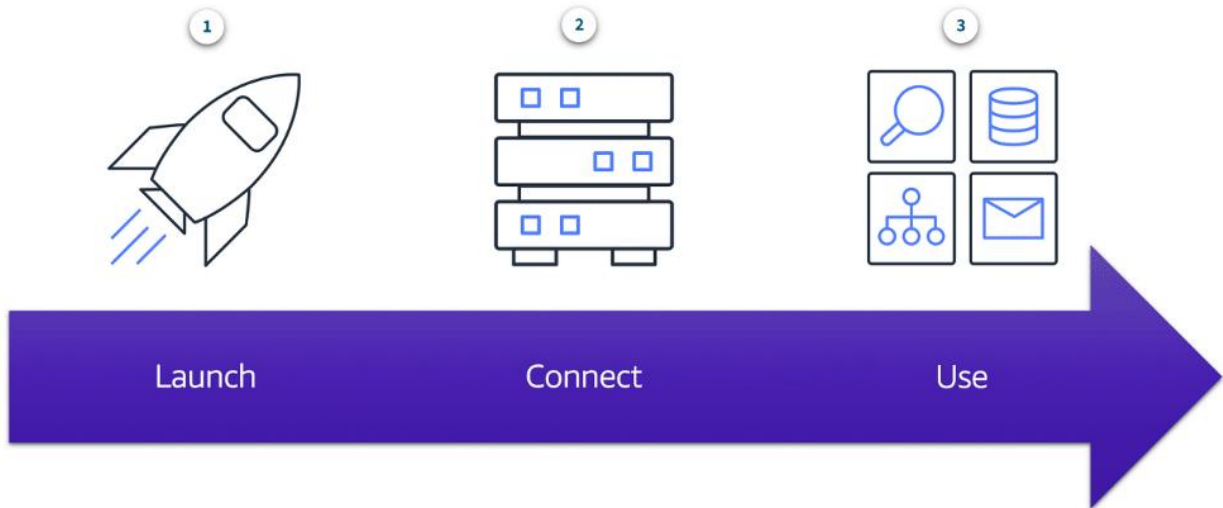
1.

Launch   < >

First, you launch an instance. Begin by selecting a template with basic configurations for your instance. These configurations include the operating system, application server, or applications. You also select the instance type, which is the specific hardware configuration of your instance.

As you are preparing to launch an instance, you specify security settings to control the network traffic that can flow into and out of your instance. Later in this course, we will explore Amazon EC2 security features in greater detail.

2.

Connect ‹ ›

Next, connect to the instance. You can connect to the instance in several ways. Your programs and applications have multiple different methods to connect directly to the instance and exchange data. Users can also connect to the instance by logging in and accessing the computer desktop.

3

Use ‹ ›

After you have connected to the instance, you can begin using it. You can run commands to install software, add storage, copy and organize files, and more.

Amazon EC2 is a service that provides reusable compute capacity in the cloud.

Free tier details:-

a. 12 MONTHS FREE
b. 750 hours per month of Linux, RHEL, or SLES t2.micro or t3.micro instance dependent on region.
c. 750 hours per month of Windows t2.micro or t3.micro instance dependent on region.

d. 30 GB volume per month free.

url – https://aws.amazon.com/ec2/pricing/

TechData-Infinity-Devops with MultiCloud

Benefits of amazon ec2-

1. Elastic web-scale computing, i.e. its scalable.
2. Completely controlled by the user.
3. Can be easily used for cloud hosting.
4. Can be configured with other AWS services.
5. Other benefits such as its Reliable, Secure, Inexpensive, easy to start.

Features of EC2-

1. Virtual computing service.
2. Preconfigured templates are available or else we are create our own templates.
3. Various config of CPU, Memory, storage etc. can be done.
4. Secure login using keypairs
5. Persistent volumes/storages are available for your data.
6. Multiple physical locations for your resources.
7. Firewall that enables you to specify the ports, protocols, IP etc.
8. Static IP addresses available.
9. Isolated Virtual networks possible.

Types/categories of instances-

1. General Purpose Instances.
2. Compute Optimized Instances.
3. GPU Graphic instances/ Accelerated computing
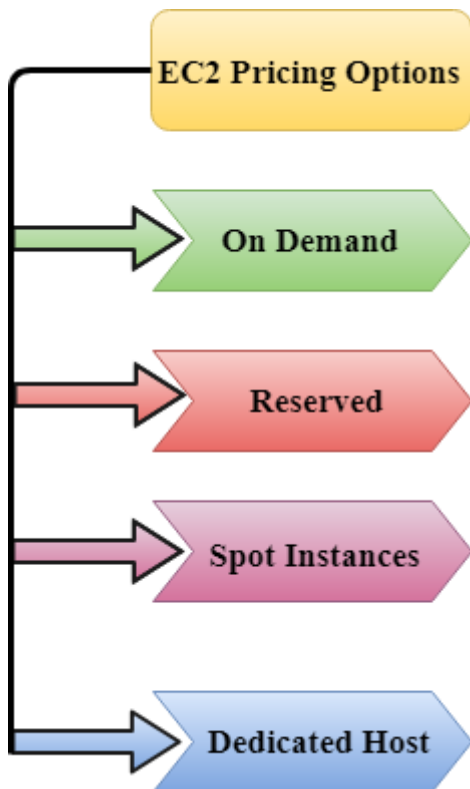4. Memory Optimized
5. Storage Optimized instances

url – https://aws.amazon.com/ec2/instance-types/

What all OS can we launch?

1. Windows
2. RHEL
3. Fedora
4. CentOs
5. Ubuntu
6. Suse Linux and many many more.

EC2 Pricing Options –

TechData-Infinity-Devops with MultiCloud

On Demand –

- It allows you to pay a fixed rate by the hour or even by the second with no commitment.
- Linux instance is by the second and windows instance is by the hour.
- On Demand is perfect for the users who want low cost and flexibility of Amazon EC2 without any up-front investment or long-term commitment.
- It is suitable for the applications with short term, spiky or unpredictable workloads that cannot be interrupted.
- It is useful for the applications that have been developed or tested on Amazon EC2 for the first time.
- On Demand instance is recommended when you are not sure which instance type is required for your performance needs.

Reserved –

- It is a way of making a reservation with Amazon or we can say that we make a contract with Amazon. The contract can be for 1 or 3 years in length.
- In a Reserved instance, you are making a contract means you are paying some upfront, so it gives you a significant discount on the hourly charge for an instance.
- It is useful for applications with steady state or predictable usage.
- It is used for those applications that require reserved capacity.
- Users can make up-front payments to reduce their total computing costs. For example, if you pay all your upfronts and you do 3 years contract, then only you can get a maximum discount, and if you do not pay all upfronts and do one year contract then you will not be able to get as much discount as you can get If you do 3 year contract and pay all the upfronts.

Spot Instances –

- It allows you to bid for a price whatever price that you want for instance capacity, and providing better savings if your applications have flexible start and end times.
- Spot Instances are useful for those applications that have flexible start and end times.
- It is useful for those applications that are feasible at very low compute prices.
- It is useful for those users who have an urgent need for large amounts of additional computing capacity.
- EC2 Spot Instances provide less discounts as compared to On Demand prices.
- Spot Instances are used to optimize your costs on the AWS cloud and scale your application's throughput up to 10X.
- EC2 Spot Instances will continue to exist until you terminate these instances.

Dedicated Hosts –

- A dedicated host is a physical server with EC2 instance capacity which is fully dedicated to your use.
- The physical EC2 server is the dedicated host that can help you to reduce costs by allowing you to use your existing server-bound software licenses. For example, Vmware, Oracle, SQL Server depending on the licenses that you can bring over to AWS and then they can use the Dedicated host.
- Dedicated hosts are used to address compliance requirements and reduces host by allowing to use your existing server-bound server licenses.
- It can be purchased as a Reservation for up to 70% off On-Demand price.

EC2 Instance Families-

1. **General Purpose Instances (e.g., t2, t3, m5)** –

These instances provide a balance of compute, memory, and network resources. They are suitable for a wide range of workloads, including web servers, small databases, and development environments.

2. **Compute Optimized Instances (e.g., c5, c6g)** –

These instances are designed for computationally intensive workloads that require high performance processors. They are suitable for applications such as scientific modeling, gaming servers, and high-performance web servers.

3. **Memory Optimized Instances (e.g., r5, x1e)** –

These instances are optimized for memory-intensive workloads that require high memory capacity. They are suitable for applications like in-memory databases, real-time big data analytics, and high-performance computing.

4. **Accelerated Computing Instances (e.g., p3, g4)** –

These instances are equipped with specialized hardware accelerators such as GPUs (Graphics Processing Units) or FPGAs (Field-Programmable Gate Arrays). They are suitable for tasks like machine learning, data processing, and video encoding.

5. **Storage Optimized Instances (e.g., i3, d2)** –

These instances are designed for applications that require high-speed, low-latency storage. They are suitable for data warehousing, distributed file systems, and large-scale analytics.