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

**(1) Launch**

First, you launch an instance. Begin by selecting a template with basic configurations for 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, contact 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.

Multi-tenancy EC runs on top of physical hast machines managed by AS 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

General purpose instances provide a balance of compute, memory, and networking resources, You can use them for a variety of workloads, such as

1) application servers

2) gaining servers

3)O backend servers for enterprise applications

4)small and medium databases

Compute optimized instances are ideal for compute-bound applications that benefit from high-performance processors. Like general purpose Instances, you can use compute optimized instances for workloads such as web, application, and gaming servers.

optimized Instances are designed to deliver fast performance for workloads that process large datasets in memory. In computing, memory is a temporary storage area. It holds all the date and instructions that a central processing unit (CPU) needs to be able to complete actions. Before a computer program or application is able to run, it is loaded from storage into memory. This preloading process gives the CPU direct access to the computer program

Accelerated computing Instances use hardware accelerators, or coprocessors, to perform some functions more efficiently then is possible in software running on CPUs, Examples of these functions Include floating-point number calculations, graphics processing, and data pattern matching.

Memory optimized Instances are designed to deliver fast performance for workloads that

process large datasets In memory, In computing, memory 19 a temporary storage area. It

holds all the data and Instructions tha t a central processing unit (CPU) needs to be able to

complete actions. Before a computer program or application Is able to run, it 19 loaded from

storage into memory, This preloading process gives the CPU direct access to the computer

program.

Accelerated computing instances use hardware accelerators, or coprocessors, to perform

some functions more efficiently than 1s possible in software running on CPUs. Examples of

these functions Include Floating-point number calculations, graphics processing, and data

pattern` matching

Storage optimized Instances are designed for workloads that require high, sequential read

and write access to large datasets on local storage, Examples of workloads suitable for

storage optimized instances include distributed file systems, data warehousing

applications, and high-frequency online transaction processing (OLTP) systems.

**Types of Instances**

Storage optimized instances are designed for workloads that require high, sequential read and write access to large datasets on local storage. Examples of workloads suitable for storage optimized Instances Include distributed file systems, data warehousing applications, and high-frequency online transaction processing (LTP) systems.

On-Demand Instances are ideal for short-terms, irregular workloads that cant be Interrupted. No upfront costs sinimas contracts apply. The stances ran antinuously until you stop them, and you pay for only the compute time you ne

reserved Instances are a billing discount applied to the use of On-Demand Instances in your account, and these sy suited for steady-state workloads or ones with predictable unge. You can purchase Standard Reserved and Convertible merunt Instances for a 1-year or 3-year ters, and Scheduled teserved Instances for a 1-year ters as reslits greater cont savings with the 3-year option

splet Instances are ideal for workloads with this start and and Claes, or that can withstand Interruptions. They allow you to repiest spars 12 coving creacity for so to D off of the De-demand price. The catch here Se thet us can reclaia the Instance at any time they need it, giving you that wending to Finish up work

Amazon EC2 Auto Scaling enables you to automatically add or remove Amazon EC2

instances in response to changing application demand, By automatically scaling your

instances in and out as needed, you are able to maintain a greater sense of application

avaflability.

Within Amazon EC2 Auto Scaling, you can use two approaches: dynamic acaling and

predictive scaling,

O Dynanic scaling responds to changing demand.

O Predictive acaling automatically achedules the right number of Amazon EC2 instances

based on predicted demand,

Elastic Load Balancing

Elastic Load Balancing (ELD) is the AWS service that automatically distributes incoming

application traffic across multiple resources, such as Amazon EC2 instances.

A load balancer acts as a angle point of contact for all incoming web traffic to your Auto

Scaling group. This means that as you add or remove Amazon EC2 instances in response to

the amount of incoming traffic, these requests route to the load balancer first. Then, the

requests spread across multiple resources that will handle them, For example, if you have

multiple Amazon EC2 instances, Elastic Load Balancing distributes the workload across the

multiple instances so that no single instance has to carry the bulk of it

**VPC**

A networking service that you can use to establish boundaries around your AWS resources is Amazon Virtual Private Cloud (Amazon VPC).

A VPC, or Virtual Private Cloud, is essentially your own private network in AWS. A VPC allows you to define your private IP range for your ANS resources, and you place things like

EC2 instances and ELBS inside of your VPC, Amazon VPC enables you to provision an isolated

section of the Als Cloud. In this isolated section, you can launch resources in a virtual network that you define. These resources can be public facing so they have access to the internet, or private with no internet access, usually for backend services like databases or application servers.

Within a virtual private cloud (VPC), you can organize your resources into subnets.

A subnet is a section of a VPC that can contain resources such as Amazon EC2 instances.

Subnets are IP addresses in your VPC that allow you to group resources together

As Direct Connect is a service that enables you to establish a dedicated private connection between your data centre and a VPC,

A subnet is a section of a VPC in which you can group resources based on security or operational needs. Subnets can be public or private

Now, the only technical reason to use subnets in a VPC is to control access to the gateways.

The public subnets have access to the internet gateways the private subnets do not. Public subnets contain resources that need to be accessible by the public, such as an online store's website. Private subnets contain resources that should be accessible only through your private network, such as a database that contains customers' personal information and order histories.

In a VPC, subnets can communicate with each other. For example, you might have an application that involves Amazon EC2 instances in a public subnet communicating with databases that are located in a private subnet

**Network access control lists (ACLS)**

A network access control list (ACL) is a virtual firewall that controls inbound and outbound traffic at the subnet level. Stateless packet filtering

Network ACLS perform stateless packet filtering. They remember nothing and check packets that cross the subnet border each way: Inbound and outbound.

**Security groups**

To solve instance level access questions, we introduce security groups. A security group is a virtual firewall that controls inbound and outbound traffic for an Amazon EC2 instance. Every EC2 instance, when it's launched, automatically comes with a security group. And by default, the security group does not allow any traffic into the instance at all (1) security

group denies all inbound traffic and allows all outbound traffic). All ports are blocked, all IP addresses sanding packets are blocked. That's very secure, but perhaps not very useful State full packet filtering Security groups perform state full packet filtering. They remember previous decisions made for incoming packets.