Module 2

Amazon Elastic Compute Cloud (Amazon EC2)

Amazon EC2 provide secure, resizable compute capacity in the cloud as Amazon EC2 instances.

Why to use Amazon EC2:

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

Amazon EC2 instance types

There are 5 types of instances:

1. General purpose instances:

It provide a balance of compute, memory, and networking resources. You can use them for :

1. Application servers
2. gaming servers
3. backend servers for enterprise applications
4. small and medium databases

2. Compute optimized instance:

It is 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.

The difference is compute optimized applications are ideal for high performance web servers, compute-intensive applications servers, and dedicated gaming servers.

3. Memory optimized instances:

These 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 data and instructions that a CPU needs to be able to complete actions. Before a computer program or application is able to run, it is loaded form storage into memory. This preloading process gives the CPU direct access to the computer program.

4. Accelerated computing instances:

It use hardware accelerators, or coprocessors, to perform such functions more efficiently than is possible in software running on CPUs.

Examples of these function include:

1. Floating point number calculations
2. graphics processing
3. data pattern matching

Accelerated computing instances are ideal for workloads such as graphics applications, game streaming, and application streaming.

5. Storage optimized instances:

These 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 instance include distributed file systems, data warehousing applications, and high-frequency **online transaction processing** (OLTP) systems.

Storage optimized instances are designed to deliver tens of thousands of low-latency, random **input/output operations per second** (IOPS) to applications.

Amazon EC2 Pricing Details:

On-demand:

On demand instances are ideal for short term, irregular workloads that cannot be interrupted. No upfronts costs or minimum contracts apply.

This instances run continuously until you stop them, and you pay for only the compute time you use.

▼ Sample use cases for On-Demand Instances include developing and testing applications and running applications that have unpredictable usage patterns.

Amazon EC2 Savings Plans:

Amazon EC2 Savings Plans are ideal for workloads that involve a consistent amount of compute usage over a 1-year or 3-year term.

Amazon EC2 Savings Plans enable you to reduce your compute costs by committing to a consistent amount of compute usage for a 1-year or 3-year term.

Reserved Instances:

These are billing discount applied to the use of on-demand instances in you account.

Spot Instances:

Spot instances are ideal for workloads with flexible start and end times, or that can withstand interruptions.

It use unused amazon EC2 computing capacity and offer you cost savings at up to 90% off on-demand prices.

Dedicated Hosts:

Dedicated hosts are physical servers with amazon EC2 instance capacity that is fully dedicated to your use.

Scaling:

Scalability involves beginning with only the resources you need and designing your architecture to automatically respond to changing demand by scaling out or in. As a result, you pay for only the resource you use.

Amazon EC2 Auto Scaling enables you to automatically add or remove Amazon EC2 instances in response to changing application demand.

Within Amazon EC2 Auto Scaling, you can use two approaches: dynamic scaling and predictive scaling.

- *Dynamic scaling* responds to changing demand.
- *Predictive scaling* automatically schedules the right number of Amazon EC2 instances based on predicted demand.

Elastic Load Balancing:

It is the AWS Service that automatically distributes incoming application traffic across multiple resources, such as amazon EC2 instance.

Although Elastic Load Balancing and Amazon EC2 Auto Scaling are separate services, they work together to help ensure that applications running in Amazon EC2 can provide high performance and availability.

Messaging and Queuing

Applications are made of multiple components. The components communicate with each other to transmit data, fulfill requests, and keep the application running.

These components might include databases, servers, the user interface, business logic, and so on. This type of architecture can be considered a **monolithic application**

.

To help maintain application availability when a single component fails, you can design your application through a microservices approach.

In a microservices approach, application components are loosely coupled.

Amazon Simple notification Service:

It is a publish/subscribe service. Using Amazon SNS, a publisher messages to subscribers.

In amazon SNS, subscribers can be web servers, email addresses, AWS lambda functions or several other options.

Amazon simple Queue Service (Amazon SQS)

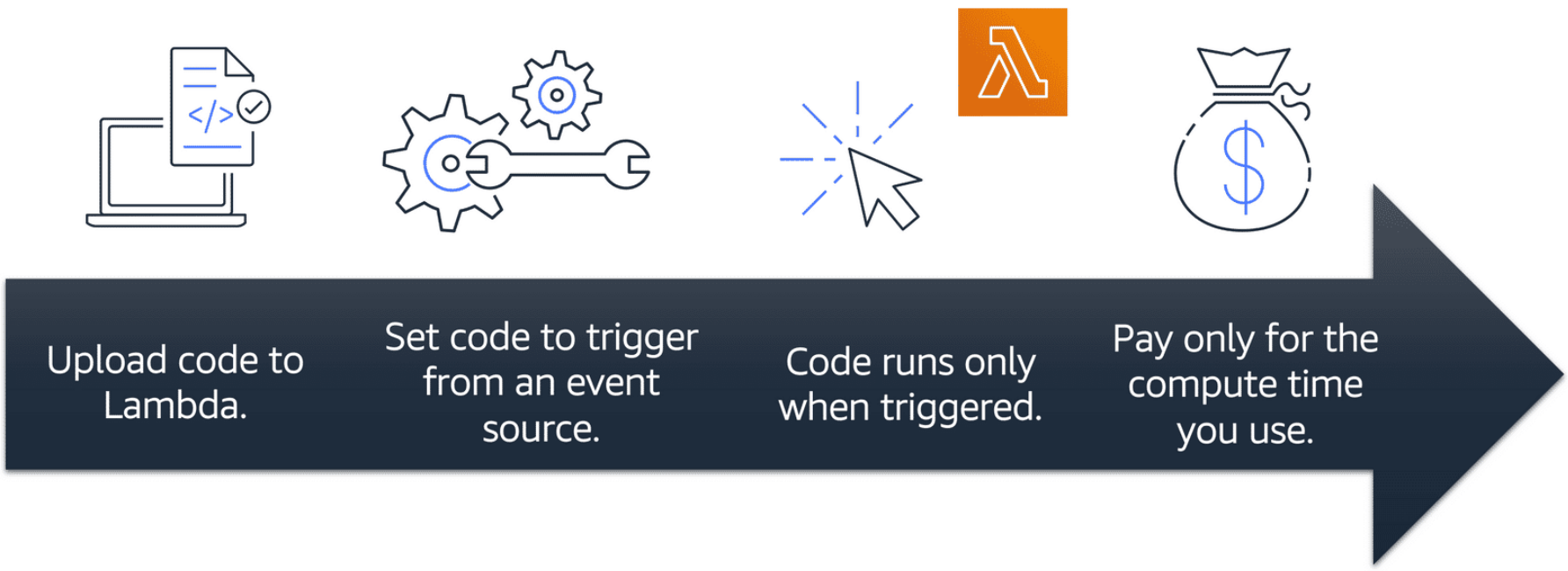
It is a message queuing service. Using Amazon SQS, you can send, store, and receive messages btw software components, without losing messages or requiring other services to be available.

AWS lambda

It is a service that lets you run code without needing to provision or manage servers.

While using lambda, you pay only for the compute time that you consume. Charges apply only when your code is running. You can also run code for virtually any type of application or backend service, all with zero administration.

How AWS lambda works:



1. Upload code to lambda
2. Set the code to trigger from an event source, such as AWS services, mobile applications or HTTP endpoints
3. Lambda runs your code only when triggered
4. You pay only for the compute time that you use.

Containers:

Containers provide you with a standard way to package your applications code and dependences into a single object.

You can also use containers for processes and workflows in which there are essential requirement for security, reliability, and scalability.

Amazon Elastic Container Service

It is a high performance container management system that enables you to run and scale containerized applications on AWS.

Amazon ECS supports Docker containers.

Docker is a software platform that enables you to build, test, and deploy applications quickly.

Amazon Elastic Kubernetes Service

Amazon EKS is a fully managed service that you can use to run Kubernetes on AWS.

Kubernetes is open-source software that enables you to deploy and manage containerized applications at scale.

AWS fargate

It is a serverless compute engine for containers. It works with both Amazon ECS and Amazon EKS. AWS fargate manages your server infrastructure for you.