Amazon EC2 instance types

Amazon EC2 instance types are **optimized for different tasks**. When selecting an instance type, consider the specific needs of your workloads and applications. This might include requirements for compute, memory, or storage capabilities.

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

 application servers

 gaming servers

 backend servers for enterprise applications

 small and medium databases

Suppose that you have an application in which the resource needs for compute, memory, and networking are roughly equivalent. You might consider running it on a general purpose instance because the application does not require optimization in any single resource area.

**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.

However, the difference is compute optimized applications are ideal for high-performance web servers, **compute-intensive applications servers, and dedicated gaming servers**. You can also use compute optimized instances for **batch processing workloads** that require **processing many transactions** in a single group.

**Memory 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 data 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 than is possible in software running on CPUs. Examples of these functions include **floating-point number calculations, graphics processing, and data pattern matching.**

In computing, a **hardware accelerator** is a component that can expedite data processing. Accelerated computing instances are ideal for workloads such as **graphics applications, game streaming, and application streaming.**

**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.

Amazon EC2 pricing

With Amazon EC2, you pay only for the compute time that you use. Amazon EC2 offers a variety of pricing options for different use cases. For example, if your use case can withstand interruptions, you can save with Spot Instances. You can also save by committing early and locking in a minimum level of use with Reserved Instances.

**On-Demand Instances** are ideal for short-term, irregular workloads that cannot be interrupted. **No upfront costs or minimum contracts** apply. The 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. On-Demand Instances are not recommended for workloads that last a year or longer because these workloads can experience greater cost savings using Reserved Instances.

AWS offers **Savings Plans** for several compute services, including Amazon EC2. **Amazon EC2 Savings Plans** enable you to reduce your compute costs by **committing to a consistent amount of compute usage** (measured in dollars per hour) for a 1-year or 3-year term. This term commitment results in savings of up to 72% over On-Demand costs.

Any usage up to the commitment is charged at the discounted Savings Plan rate (for example, $10 an hour). Any usage beyond the commitment is charged at regular On-Demand rates.

**Reserved Instances** are a billing discount applied to the use of **On-Demand Instances** in your account, and these are suited for **steady-state workloads or ones with predictable usage**. You can purchase Standard Reserved and Convertible Reserved Instances for a 1-year or 3-year term, and Scheduled Reserved Instances for a 1-year term. You realize greater cost savings with the 3-year option.

At the end of a Reserved Instance term, you can continue using the Amazon EC2 instance without interruption. However, you are charged On-Demand rates until you do one of the following:

 Terminate the instance.

 Purchase a new Reserved Instance that matches the instance attributes (instance type, Region, tenancy, and platform).

**Spot Instances** are ideal for workloads with **flexible start and end times**, or that can withstand interruptions. They allow you to request spare Amazon EC2 computing capacity for up to 90% off of the On-Demand price. The catch here is that **AWS can reclaim the instance at any time** they need it, giving you a two-minute warning to finish up work and save state. You can always resume later if needed. So when choosing Spot Instances, **make sure your workloads can tolerate being interrupted.** A good example of those are **batch workloads**.

Suppose that you have a background processing job that can start and stop as needed (such as the data processing job for a customer survey).

**Dedicated Hosts** are physical servers with Amazon EC2 instance capacity that is fully dedicated to your use. These are usually for meeting certain compliance requirements and **nobody else will share tenancy of that host**.

You can use your existing per-socket, per-core, or per-VM software licenses to help maintain license compliance. You can purchase On-Demand Dedicated Hosts and Dedicated Hosts Reservations. Of all the Amazon EC2 options that were covered, Dedicated Hosts are the most expensive.