The Complete Guide to AWS EC2 Instance Types and Their Use Cases

What is instances

An **instance in AWS** is a **virtual server** that runs applications in the cloud, within the Amazon Elastic Compute Cloud (EC2) service. It functions much like a physical server, but as a virtual machine, it provides the flexibility, scalability, and efficiency of cloud computing without the need to manage the underlying hardware.

What is EC2

EC2 (Elastic Compute Cloud) is a core Amazon Web Services (AWS) offering that provides scalable, resizable virtual servers, called "instances," in the cloud, allowing users to run applications without investing in physical hardware. It offers flexible computing power with various configurations, letting you pay only for what you use, and easily scale resources up or down based on demand, making it ideal for web hosting, big data, development, and machine learning.

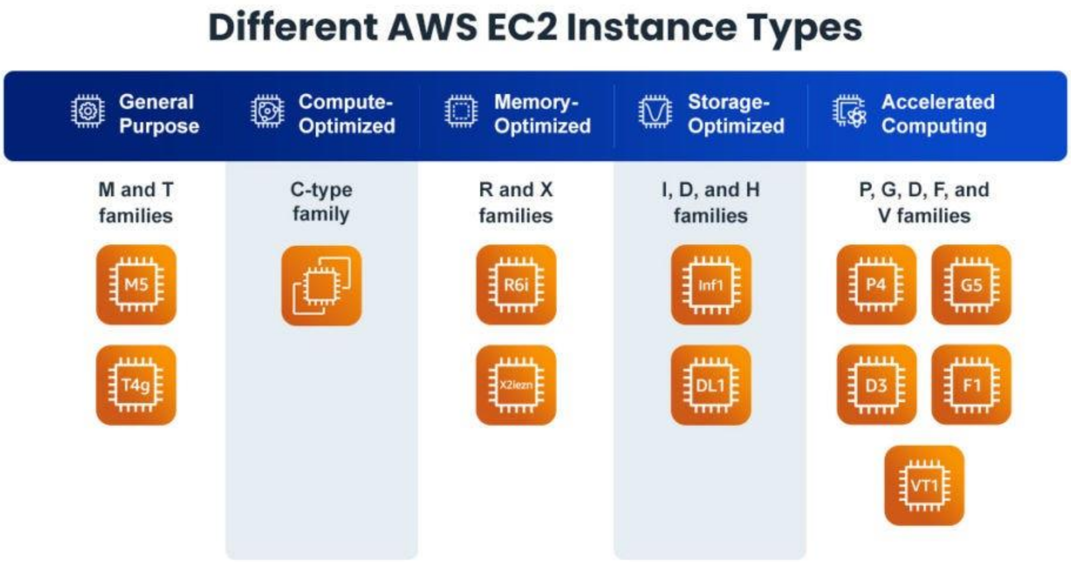
The Definitive Guide to AWS EC2 Instance Types

Amazon allows you to run different AWS EC2 instance types in the AWS cloud, giving you much needed flexibility. You can select an EC2 instance that best meets your requirements at the right price. This blog post explains AWS EC2 instance types and categories as well as provides come recommendations to help you make the right decision when you need to run an instance in the Amazon cloud.

AWS EC2 instances are divided into multiple categories called instance types. Each type is optimized for specific use cases.

There are 5 AWS EC2 instance types:

- General Purpose
- Compute Optimized
- Memory Optimized
- Storage Optimized
- Accelerated Computing



Instance Category	Key Features	Use Cases	Examples
General Purpose Instances	Balanced CPU, memory, and networking resources	Web servers, small databases, development environments	T3, M5, M6g
Compute Optimized Instances	High CPU power compared to memory	Batch processing, gaming servers, machine learning inference	C5, C6g
Memory Optimized Instances	More memory compared to CPU	High-performance databases, big data analytics, real-time processing	R5, X1, Z1d
Storage Optimized Instances	Designed for workloads needing high, fast local storage	Data warehousing, big data, log processing	I3, D2, H1
Accelerated Computing Instances	Specialized hardware like GPUs or FPGAs for faster processing	Machine learning, graphics rendering, scientific simulations	P3, G4, F1

1.General Purpose instance types

General-purpose instances can be used in different scenarios and provide a balance of CPU (central processing unit), RAM (random access memory), and networking resources.

Key Features:

- Balanced CPU, memory, and network capabilities
- Versatile for many different workloads
- Cost-effective option for common use cases

EC2 General-Purpose Instance Types

Here are several general-purpose examples from which we can pick:

Families:

- **M Series (e.g., M7g, M6i, M5):** Balanced resources, good for small to medium databases, enterprise applications, web servers, and development/test environments. Newer generations (like M7g with Graviton3 processors) offer improved price-performance.
- **T Series (e.g., T4g, T3, T2):** Burstable Performance Instances. They provide a baseline CPU performance with the ability to "burst" above the baseline when needed (using CPU credits). Ideal for applications with variable or low-to-moderate CPU usage, such as small web servers, microservices, development environments, and CI/CD pipelines.
- **A1 Series:** ARM-based instances powered by AWS Graviton processors, offering a good price-performance ratio for scale-out workloads and ARM-compatible applications.
- **Mac Series:** Mac mini computers used as EC2 instances for macOS development and testing

Applications

1. **Web Servers:** The web servers can be hosted in General-purpose instances. EC2 instances provide a flexible and scalable platform for web applications.
2. **Development and Test Environment:** The developers can use these General-purpose instances to build, test and deploy the applications. It is a cost-effective solution for running this environment.
3. **Content delivery:** The hosting of content delivery networks (CDNs) that distribute content to users all over the world is possible using general-purpose instances. EC2 instances can be set up to provide content with low latency and great performance.

2. Compute Optimized Instances

Compute Optimized Instances are special types of cloud servers designed for tasks that need a lot of processing power. They have strong CPUs (the “brain” of the computer) to handle heavy calculations quickly. These instances are perfect for applications like gaming servers, scientific modeling, machine learning, and batch processing, where fast computing is very important

Key Features:

- Balanced CPU, memory, and networking resources.
- Flexible and suitable for a wide variety of workloads.
- Good for applications that don’t need extreme CPU or memory power.
- Often cost-effective and scalable for growing needs.

Use Cases:

- Web servers and application servers.
- Small to medium databases.
- Development and testing environments.
- Content management systems and backend servers.
- Business applications with moderate resource needs.

Families:

- **C Series (e.g., C7g, C6i, C5):** Excellent for batch processing workloads, high-performance web servers, scientific modeling, media transcoding, dedicated gaming servers, and machine learning inference. Graviton-powered C7g instances provide significant price-performance improvements.

3. Memory Optimized Instances

Memory Optimized Instances are designed to provide a large amount of RAM relative to CPU power. They are ideal for applications that require fast, efficient processing of large datasets stored in memory.

Key Features:

- High memory capacity with fast access
- Low latency memory performance
- Enhanced networking and storage support

Use Cases:

- High-performance databases (SQL, NoSQL)
- Big data analytics and processing (Apache Spark, Hadoop)
- Real-time data streaming and processing
- In-memory caches (Redis, Memcached)

Families:

- **R Series (e.g., R8g, R7g, R6g, R5):** Best for high-performance databases (like in-memory databases such as SAP HANA), big data analytics, large in-memory caches (e.g., Redis, Memcached), and enterprise applications requiring substantial memory.
- **X Series (e.g., X2gd, X1e):** Provide extremely high memory capacity and are designed for very large-scale, memory-intensive enterprise workloads.
- **Z1d Series:** Offer high compute capacity combined with high memory, suitable for Electronic Design Automation (EDA) and relational databases.

4. Storage Optimized Instances

Storage Optimized Instances are designed to deliver high, fast, and low-latency local storage. They are perfect for workloads that require heavy read/write access to large amounts of data stored on the instance.

Key Features:

- High-speed, low-latency local storage (often NVMe SSDs)
- Optimized for large sequential I/O operations
- Enhanced networking capabilities for fast data transfer

Use Cases:

- Data warehousing and big data analytics
- High-frequency online transaction processing (OLTP)
- Distributed file systems
- Log or data processing applications

Families:

- **I Series (e.g., I4i, I3en, I3):** Optimized for low-latency, high-IOPS transactional workloads. Ideal for NoSQL databases (Cassandra, MongoDB), relational databases, data warehousing, and real-time analytics.
- **D Series (e.g., D3en, D2)**: Offer high-density HDD storage for data-intensive workloads. Suitable for distributed file systems (HDFS), large-scale parallel processing (MapReduce), and log processing.
- **H1 Series:** Provide high disk throughput for large-scale data processing and distributed file systems

5. Accelerated Computing Instances

Accelerated Computing Instances include specialized hardware like GPUs (Graphics Processing Units) or FPGAs (Field Programmable Gate Arrays) to perform specific tasks faster than standard CPUs. These instances are designed for workloads that require heavy computation, such as graphics rendering, machine learning, and scientific simulations.

Key Features**:**

- Equipped with GPUs or FPGAs for faster processing
- High parallel processing power
- Optimized for compute-intensive and graphics-heavy tasks

Use Cases:

- Machine learning training and inference
- Video rendering and transcoding
- Scientific modeling and simulations
- Financial risk analysis and high-performance computing (HPC)

Families:

- **P Series (e.g., P5, P4d, P3):** Equipped with NVIDIA GPUs, primarily for machine learning training, high-performance computing (HPC), and deep learning.
- **G Series (e.g., G6, G5, G4dn):** Also use NVIDIA GPUs, but often for graphics-intensive applications (3D rendering, video encoding, virtual workstations), machine learning inference, and game streaming.
- **Inf/Trn Series (e.g., Inf2, Trn1):** Feature AWS Inferentia or Trainium chips, purpose-built for high-performance machine learning inference and training at scale.
- **F1 Series:** Use FPGAs (Field-Programmable Gate Arrays) for custom hardware acceleration, suitable for genomics, financial modeling, and real-time video processing.