Detailed Types of EC2 Instances in AWS Cloud

General Purpose

- T4g, T3, T3a: Provide a baseline level of CPU performance with the ability to burst above the baseline. Ideal for web servers, small databases, and development environments.
- M7g, M6g, M5, M5a: Provide a balance of compute, memory, and networking resources. Best suited for general-purpose applications like app servers, gaming servers, and backend systems.

Compute Optimized

- C7g, C6g, C6i, C5: Designed for compute-intensive workloads. Use cases include high-performance web servers, scientific modeling, batch processing, and dedicated gaming servers.

Memory Optimized

- R7g, R6g, R6i, R5: Deliver fast performance for memory-intensive applications such as in-memory caching (e.g., Redis), real-time big data analytics, and high-performance databases.
- X2gd, X1e, u-6tb1: Offer high memory-to-vCPU ratios. Suitable for large-scale SAP workloads, in-memory databases like SAP HANA, and real-time analytics platforms.

Storage Optimized

- I4i, I3, D3: Provide low-latency, high-throughput local storage. Ideal for NoSQL databases (e.g., Cassandra), data warehousing, and Elasticsearch workloads.
- H1: Offer high disk throughput and dense HDD-based storage. Commonly used for data-intensive workloads like MapReduce, Hadoop, and log processing.

Accelerated Computing

- P4, P3: Equipped with GPUs. Designed for machine learning training, high-performance computing (HPC), and graphics-intensive applications like video rendering.

- Inf1, Trn1: Optimized for machine learning inference and training, respectively. Suitable for AI applications that require high throughput and low latency.

- F1: Feature customizable field programmable gate arrays (FPGAs). Ideal for hardware acceleration of custom computing applications such as genomics research and financial analytics.