Juan Pérez - 20231020028.

T8: What are cache servers and how to use them in an application?

In computing, a cache is a high-speed data storage layer that stores a subset of data, usually transient, so that future requests for that data are served more quickly than if the data had to be accesed from the main storage location. Caching allows for efficient reuse of previously retrieved or processed data.

Data in a cache is typically stored in fast-access hardware such as random access memory (RAM) and may also be used in conjunction with a software component. The primary purpose of a cache is to increase data retrieval performance to avoid having to access the slower underlying storage layer.

Trading capacity for speed, a cache typically stores a subset of data transiently, unlike databases whose items are typically complete and durable.

Caches can be applied and leveraged at multiple layers of technology, including operating systems, network layers including content delivery networks (CDN) and DNS, web applications, and databases.

You can use caching to significantly reduce latency and improve IOPS for many read-intensive application workloads such as Q&A portals, gaming, content sharing, and social networking. Cached information can include database query results, compute-intensive calculations, API requests/responses, web artifacts such as HTML, Javascript, and image files.