**Latency**is the time required to perform some action or to produce some result. Latency is measured in units of time like seconds, milli seconds, micro seconds, nanoseconds, etc.

**Throughput** is the number of such actions executed or results produced per unit of time. This is measured in units of time like requests per second. The term “memory bandwidth” is sometimes used to specify the throughput of memory systems.