Latency And Throughput

Latency

Latency is the time taken by a piece of data to traverse through a system (i.e. Time taken by a piece of data to travel from one point of the system to another point

Eg: Network latency:

Time taken by a request to travel from the client to the server and back again to the client

Memory / Disk latency:

Time taken to retrieve / write data to the memory / disk

Latency for reading 1mb of data from

RAM	~ 250 µs
SSD	~ 1000 µs
HDD	~ 20,000 µs

Latency of sending a packet (~ 1500bytes) of data from California to Netherlands and back again to California ⇒ 150,000 µs

Throughput

Throughput is the amount of work performed by a system in a given period of time.

Generally measured in mb/s (or) Gbps

The system can handle 1mb of data per second

GB per second

Number of operations a system can handle properly per unit time

Expected

Low Latency High Throughput ☺