# Latency & Throughput

Latency and Throughput → 2 most imp measures of the performance of system

A] LATENCY:

i) Latency is basically how long it takes for data to traverse a sys. i.e. get frm 1 pt. in sys to another pt. in sys.

eg:- latency of network req: how long it takes for 1 req. to go frm client to a server and then the processed response to go from server to client

eg 2:- time reqd. by a server to read data frm disk

ii) Diff parts of a sys hv diff. latencies
∴ While designin a sys we'll face a trade-off while optimizin it cuz we'll have

parts with lower latency | parts with higher latency

iii) Some comparisons.

| Part of sys + its $f^n$ | Latency |
|---|---|
| Readin 1MB from memory (RAM) | 250 $\mu s$ |
| Readin 1MB from SSD | 1000 $\mu s$ |
| Sendin 1MB over 1Gbps network | $10^4$ $\mu s$ → (sendin to computer next to us i.e. dist. is not considered |
| Readin 1MB from HDD | $2 \times 10^4$ $\mu s$ |
| Sendin packet (≈ 1055 B) over network from California to Netherlands and then back to California | $15 \times 10^4$ $\mu s$ |

eg: of sendin data over a network → API ∴ network 'req.

eg of ~~o~~ readin data frm memory → reading a variable in code

Takeaways :

i) Dependin on network and your PC hardware, sometimes ~~read~~ sendin, receivin data over network is faster than readin data frm HDD

ii) Sendin data around the world takes a lot longer than any other meth.
Reason : req. gonna ~~travel up~~, get converted into small locks (into binary data) → converted into freq. modulated radiowaves → sent to cell towers thru cables → → bounced to satellite passed around the world thru satellite comm.

passed to destina$^n$ cell tower

passed to destina$^n$ and reconverted back to original form

iii) Optimizin a sys → ↓ its overall latency

eg: - Video games gotta have to have really low latency and
Lag → delay in ac$^n$ passed from 1 user → server → receiving user. These ac$^n$s are passed as network req.s to . 's server so if you're pretty far away frm server, your PC will take more time to make network req /receive responses frm server

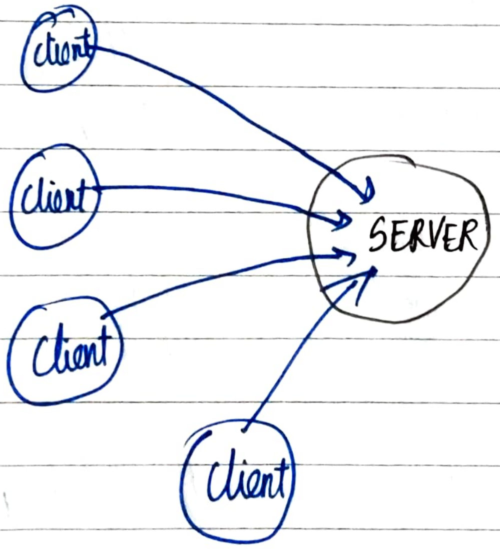
eg:- Websites → It's peace if they have low latency because in most cases, their priority is gettin the info displayed to be accurate, uptime to be cont. 24X7

## THROUGHPUT (TP)

Throughput is basically how much work a machine can perform in a given amt. of time → how much data can be transferred frm 1 pt. in sys to another pt, in given time

unit :- bytes/sec

eg:- 1Gbps network → network can support 1Gb per sec.



TP is how many req.s (each hvin some data) can this server handle in given time → how much data it can let thru per sec.

case(i)

To optimize sys. → pay to ↑ TP

case(ii) → in case just ↑ TP doesn't solve prob. as you might have a server hvin $10^3$ or $10^6$ req. issued to it per sec so no matter how much we ↑ TP, we'll still hv a bottleneck (only some data is let thru server) at server

Case (ii) soln : Have diff. servers for req.s so they don't clog at bottleneck

IMP : Latency and throughput are not corelated

eg:- you might have parts of sys with reall low latency (fast data transfer)

→ but then if you hv a part with v. less TP, then our advan. due to low latency in other parts gets cancelled out as req.s gotta wait (clog) at this part cuz of low TP

∴ You can't make assump$^{n}$s on latency or TP based on each other.
They affect each other but don't determine each other