Week 1 - Lesson 2 : Latency vs Throughput !

1 Definitions

Latency

Time taken for one request

Example

- You click a "like"
- Response comes in 200ms

Throughput

How many requests a system can handle per unit time

Example

- How many request i.e. A system handles 50,000 likes per second

2 Key insight

A system can have low latency but low throughput

A system can have hight throughput but high latency.

Thay are independent.

3 Real-world analogy

Bank example

- One teller
- Each customer takes 30 secs.

Latency = 30 sec

Throughput = 2 customer/minute

Add more tellers:

Latency stays ~ 30 sec

Throughput increases a lot

4 System example (Like Counter)

Senario A : Single Fast Server

- Responds in 50 ms
- Can handle 1000 req / sec

- Low latency - Low throughput

Senario B : Multiple servers

- Responds in 150 ms
- Handles 100,000 req / sec
- High latency : massive throughput

5 Instead of reducing latency

Increase throughput by removing bottlenecks.

6 Common mistakes

- Adding caching to reduce latency but DB still crashes
- Adding servers without load balancing
- Optimizing code ~~with~~ when bottleneck is disk

7 Mini Design Exercise

Task 2

You have a system where:

- Each request takes 100 ms
- One server handles 10 requests at a time.

Answer these:

1 What is latency ?

2. What is throughput ?

3. How do you increase throughput without reducing latency?