Week 1 - lesson2 - task 2

What is the latency?

latency is when one request takes 100 ms
i.e 100 ms per request

What is the throughput?

Each request = 100 ms
In one second (1000 ms)
It can finish 100 requests
∴ Throughput = 100 requests per second

How do you increase throughput without reducing latency?

we will need to do horizontal scaling

- Latency stays 100 ms
- Add more servers → more paralel requests
- Throughput increase linearly