Throughput

You have to organise a birthday party in a cafe. Suppose cafe A has 2 sittings and serves food in 15 minutes and cafe B has 5 sittings and serves food in 20 minutes. Which one will you prefer? Confused?
Let's calculate the rate of providing food for each cafe.
For cafe A, rate of serving food = number of sittings/serving time = 2/15 = 0.134 person/minute
For cafe B, rate of serving food = number of sittings/serving time = 5/20 = 0.25 person/minute
It is very clear now that cafe B has a higher rate of serving which means it serves more people per minute than cafe A. So, one should choose cafe B.
This rate helped us determine the better cafe's throughput.

You must have understood how important throughput is. Let's understand it better.

What is throughput?

Throughput is the amount of data transmitted per unit of time. It is the process flow rate.
Throughput is measured in bits per second i.e. bps.

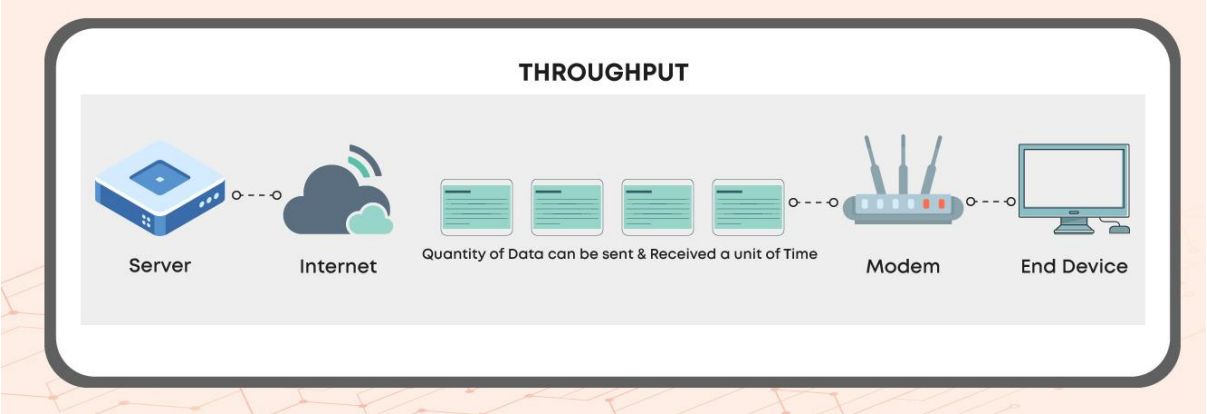

Fig: Throughput

Throughput Formula

Throughput = Inventory / Flow Time

Where,

Inventory: Number of units contained in a process
Flow Time: Total time a unit spends in the process
Throughput: Flow Rate. The number of units passing per unit time.

Example:
Suppose a car manufacturing company makes 100 cars. One car is manufactured in 10 days. Find the throughput.

Inventory (total number of cars): 100
Flow Time (Manufacturing time of one car): 10 days
Throughput : 100/10 = 10 cars/day
Therefore,

Flow Rate or Throughput = 10 cars/day

Causes of Low Throughput

1. Congestion:
   Similarly to the traffic jam, when multiple cars try to cross the road but at the end due to crossing paths, none is able to and the process is stuck. A higher number of process requests at the same time can cause congestion in the process.
2. Protocol overhead:
   If there is a requirement like handshakes or to-fro communication this leads to overloading of the protocol overheads instead of the content.
3. Latency:
   Higher latency (slow transmission of data) will also reduce the amount of data transmitted.

Throughput for different architectures

Monolithic Architecture:

- Limited resources and threads because of a single codebase handling all layers (single machine is limited in means of hardware and machine resources).
- The number of machines cannot be increased.

Limited Resources => Limited Throughput

Distributed System:

- Multiple machines are connected over a network.
- No limit on the availability of machines or resources available.
- Load balances help to distribute the load better. It avoids the situation where some nodes are completely ideal and some are highly overloaded.

Ample Resources => Higher Throughput

Improving Throughput

1. Improving the hardware/software or machine resources:
   Updating and improving the software/hardware or machine resources increase the throughput every time by improving the performance.
2. Improving the performance by using CDN(Content Delivery Network):
   CDN decreases the distance between the user and the server. This helps in shortening the response time since the distance between user and server is decreased. This improves the network performance.
3. Improving the performance by caching:
   Database Caching helps to reduce the retrieving latency which increases the throughput.
4. Monitoring and fixing all the performance bottlenecks
   Performance bottlenecks refer to network overloading which is a cause of low throughput as mentioned above. Monitoring and fixing these issues helps to improve the network performance.
5. Distributed computation using load balancers
   Load Balancing improves both the response time (latency) and resource utilisation. It avoids the situation where some nodes are completely ideal and some are highly overloaded.