Throughput: System Design Concept

Have you ever observed the flow of water coming out of a pipe? The flow of water can vary, sometimes being less and sometimes being more, but there is a maximum capacity for the flow of water that a pipe can handle. This concept is similar to throughput in computer science and communication networks.

Throughput is an important concept when designing any computer system. It refers to the rate at which a system or network can process or transmit data. In this blog, we will discuss the importance of throughput in computer systems and how it is used in design. Let's begin!

Throughput meaning in System Design

Throughput is a measure of the rate at which something is processed. It is usually expressed in terms of the number of items that are processed in a given time period, such as the number of bits transmitted per second or the number of HTTP operations per day.

To calculate throughput, the total number of items that are processed is summed and then divided by the sample interval. While this is a common method for calculating throughput, it does not take into account variations in processing speed. This means that it may not accurately reflect the true rate of production or processing.

Suppose an assembly line is manufacturing cars. Let’s consider the factory can able to produce around 100 cars per day. So the Throughput of the line is **Throughput ~ 100 cars/day**.

Misconceptions with Latency

Latency is the amount of time that passes between making a request and receiving a response. It is measured in units of time and is often confused

with throughput. It is commonly assumed that systems with high throughput should also have low latency. But, this is not always the case. For example, data processing using disks may have high throughput but also suffer from high latency.

In networked connections, latency can also increase with throughput. As the throughput increases, more packets are transmitted on the network, which can increase latency. On the other hand, it is also possible to have systems with low throughput and low latency. So it is important to consider both latency and throughput when designing a system and selecting the appropriate combination based on the business requirements.

Factors affecting Throughput

Throughput of a system can be affected by several factors. These factors are: analog limitations, hardware processing power, service accessibility, network traffic, transmission errors, protocol overhead, etc. Protocol overhead refers to extra data that must be transmitted along with the actual message to ensure proper communication. This additional data can limit the maximum achievable throughput of a system.

Analog limitations

The physical medium of a networked communication system can have a significant impact on its maximum achievable throughput. This can set an upper bound on the amount of information that can be transmitted, which can affect the throughput of the system. In other words, analog characteristics of the medium can limit amount of data that can be transmitted at a given time.

Hardware limitations

Every computing and processing system has limitations that can impact its throughput. This is because these systems have limited processing power and can only handle a certain amount of workload at a time.

- If a complex query requires a significant amount of computation, it can slow down the processing speed and ultimately affect the throughput of the system.
- When the workload becomes too much for the system to handle, its ability to process data may decrease, and its throughput will be affected.

High accessibility or concurrent requests

Service accessibility can also impact its throughput. When multiple users share a single communication system at the same time, they may need to share resources, which can reduce the system's ability to process and transmit data efficiently. This, in turn, can affect the throughput of the service.

Increased accessibility to a service can also increase network traffic, which further decrease its throughput. For example, if demand for the service is high and multiple users are accessing it simultaneously, it can put strain on the system.

Other Factors

- I/O operations like reading or writing to a disk can affect throughput if they become a bottleneck.
- Throughput will get affected if a system does not have enough memory to store data. In this situation, system will need to constantly swap data in and out of disk.
- Use of inefficient algorithms can increase processing time and increase throughput.
- Due to lack of proper load balancing, some components become overloaded and affect throughput.
- If system relies on external services or APIs, the performance of these services can affect throughput.

How to increase throughput of a system?

- Upgrade hardware components like processors, memory, and storage to increase processing speed.
- Use proper load-balancing techniques to evenly distributed workload among different components.
- Increase network bandwidth or upgrade network components to improve data transmission speed.
- Write efficient code and use optimized algorithms to improve processing speed.
- Cache frequently used data in memory to reduce the time required for data retrieval.
- Break down a task into smaller sub-tasks and process them simultaneously (parallel processing).
- Minimize protocol overhead to increase the speed of data transmission.

Conclusion

Throughput is a critical concept in the design of any system. It is used to measure the capacity and performance of a system. As such, architects and designers often strive to increase throughput as much as possible in order to improve the system's capabilities. In this blog, we have discussed the various aspects of throughput and how it is used in system design.

Thanks Suyash for his contribution in creating the first version of this content. If you have any queries or feedback, please write us at contact@enjoyalgorithms.com. Enjoy learning, Enjoy system design!