Network Throughput, Latency, and Bandwidth

The relationship between Throughput, Latency, and Bandwidth

The relationship between throughput and latency is underpinned by the concept of bandwidth.

Bandwidth is the name given to **the number of packets that can be transferred throughout the network**.

If you were to think of a pipe, a physical pipe restricts the quantity of content that can transfer through the pipe. In the context of a network, this is how many packets can be transferred at once.

The time it takes for a packet to travel from the source to its destination is referred to as latency.

Latency indicates how long it takes for packets to reach their destination.

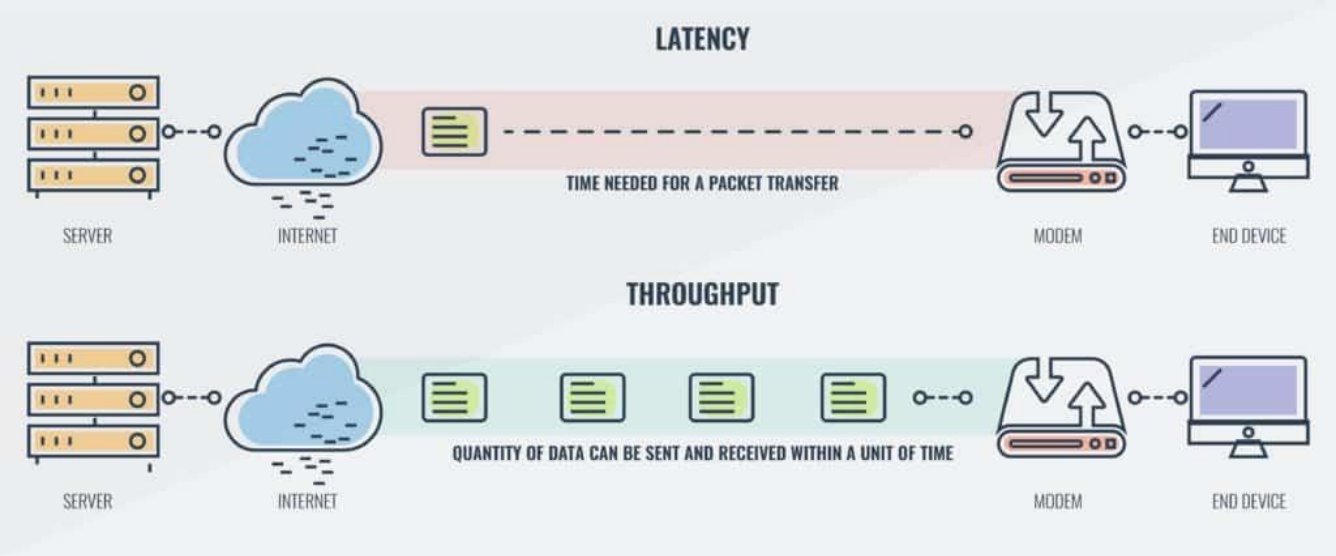
Throughput is the term given to **the number of packets that are processed within a specific period**. Throughput and latency have a direct relationship in the way they work within a network.

Putting it another way, the relationship between these three are as follows:

- The **maximum bandwidth** of a network specifies the **maximum number of conversations that the network can support**. Conversations are exchanges of data from one point to another.
- **Latency** is used to measure **how quickly these conversations take place**. The more latency there is, the longer these conversations take to hold.
- The level of latency determines the maximum throughput of a conversation. **Throughput is how much data can be transmitted within a conversation**.

Naturally, the amount of data that can be transmitted in a conversation decreases the more network latency there is.

This is because it takes longer for data to be transmitted within the conversation because packets take a longer time to reach their destination.



Latency and bandwidth are two very different concepts that have a close relationship with each other.

Latency measures the speed of packet transfers whereas bandwidth is used to refer to the maximum capacity of the network.

The simplest way to explain the relationship between the two is that **bandwidth refers to how big the pipe is**, and latency is used to measure how fast the contents of the pipe travels to its destination.

These two have a cause-and-effect relationship. For instance, the less bandwidth you have the longer it will take for your data to reach its destination, and the more latency you will have.

Likewise, the more bandwidth you have the faster packets will reach their destination. This is the case even if you have low latency.

What causes network latency?

Network latency can be caused by a range of issues but generally, it comes down to the state of routers and the distance between your network devices.

The more routers a packet must travel through the more latency there is because each router has to process the packet.

In most cases, this latency isn't noticeable but when traffic travels across the internet it can be more pronounced (because the number of routers the packet passes through increases).

The distance that a packet travel can also have a significant influence on the amount of latency within a network.

A packet that travels around the world would have at least 250 ms of latency. In enterprise-level networks, latency is present to a lesser extent.

When packets travel across a network to their destination, they rarely travel to the node in a straight line.

As such the amount of latency is dependent on the route that the packet takes.

On a well-designed network, efficient routes should be available so that packets arrive promptly at their destination.

If the network is poorly designed with indirect network paths, then latency is going to be much more pronounced.

What is network throughput?

As we said earlier, throughput is the term used to refer to the quantity of data being sent that a system can process within a specific time.

Throughput is a good way to measure the performance of the network connection because it tells you how many messages are arriving at their destination successfully.

If most messages are delivered successfully then throughput will be considered high. In contrast, a low rate of successful delivery will result in lower throughput.

The lower the throughput is, the worse the network is performing.

Devices rely on successful packet delivery to communicate with each other so if packets are not reaching their destination the result is going to be poor service quality.

Within the context of a **VoIP call**, low throughput would cause the callers to have a poor-quality call with audio skips.

What is a VoIP Phone?

VoIP phone refers to a device or program that utilizes Voice over Internet Protocol (VoIP) technology.

VoIP technology allows the user to make voice calls over broadband internet, rather than through a traditional, analog connection.

VoIP phone can look just like a traditional office desk phone.

The difference is behind the scenes.

Instead of transmitting through a physical pair of copper wires, VoIP utilizes the internet to transmit voice calls, in the form of data packets.

VoIP phone systems can also be a software application or app, coined softphone, and not require desk phone hardware.

Throughput vs Bandwidth

Bandwidth is a term used to describe the maximum amount of data that can be transferred throughout your network.

The maximum bandwidth of your network is limited to the standard of your internet connection and the capabilities of your network devices.

Think of bandwidth as the limits of your network connection. In contrast, throughput is the actual data transfer rate that occurs on your network.

It goes without saying that throughput is lower than bandwidth.

That is because bandwidth represents the maximum capabilities of your network rather than the actual transfer rate.

This is most important to note during peak periods or when performance issues are rampant as throughput will often be lower than bandwidth.

What causes poor network throughput?

Poor network throughput can be caused by several factors.

One of the main culprits is poor hardware performance.

If devices like routers are experiencing performance degradation, faults, or are simply outdated then you can end up with low throughput.

Likewise, if computer networks are congested with lots of traffic, then packet loss will occur. Packet loss is where data packets are lost in transit.

Low network throughput is often caused when packets are lost in transit.

How to measure latency and throughput

Latency is one of the most reliable ways to measure the speed of your network.

Latency is measured in milliseconds.

In the event that you want to measure the amount of data traveling from one point to another, you would use network throughput.

Throughput is measured in bits per second (bps) in the form of **megabits per second (Mbps)** or **gigabits per second (Gbps)**.

Throughput is the rate at which packets reach their destination successfully within a specific time period.

While you can calculate throughput numbers, it is simpler to measure it with bps rather than running a calculation.

Why are network latency and throughput important?

Both network latency and throughput are important because they influence how well your network is performing.

If latency is too high, then packets will take a longer amount of time to reach their destination.

The more time it takes for packets to reach their destination, the slower devices, services, and applications will operate within the network.

Likewise, the lower the amount of throughput, the lower the number of packets being processed in a specific time period.

The moment latency gets too high or throughput falls, then your network is going to grind to a halt.

This is the point at which services will start to perform sluggishly as packets fail to reach their destination at a speed that can sustain the full operation of your network.

It is **important to measure network latency and throughput** because it allows you to check that your network isn’t falling victim to poor performance.

There are many ways that you can measure latency and throughput but the simplest way is to use a network monitoring tool.

This type of tool will be able to tell you when latency and throughput have reached problematic levels.