What is Cloud Computing? › Cloud Computing Concepts Hub › Networking & Content Delivery

What’s the Difference Between Throughput and Latency?

What’s the Difference Between Throughput and Latency?

Latency and throughput are two metrics that measure the performance of a computer network. Latency is the delay in network communication. It shows the time that data takes to transfer across the network. Networks with a longer delay or lag have high latency, while those with fast response times have lower latency. In contrast, throughput refers to the average volume of data that can actually pass through the network over a specific time. It indicates the number of data packets that arrive at their destinations successfully and the data packet loss.

Read about latency »

Why are throughput and latency important?

You can determine network speed by looking at how quickly a network can transfer data packets to their destinations. This speed is the result of network performance factors like latency and throughput.

Latency determines the delay that a user experiences when they send or receive data from the network. Throughput determines the number of users that can access the network at the same time.

A network with low throughput and high latency struggles to send and process high data volume, which results in congestion and poor application performance. In contrast, a network with high throughput and low latency is responsive and efficient. Users experience improved performance and increased satisfaction.

High-performing networks directly impact revenue generation and operational efficiency. In addition, certain use cases—like real-time streaming, Internet of Things (IoT) data analytics, and high-performance computing—require certain network performance thresholds to operate optimally.

Key differences: network latency vs. throughput

Although latency and throughput both contribute to a reliable and fast network, they are not the same. These network metrics focus on distinct statistics and are different from each other.

How to measure

You can measure network latency by measuring ping time. This process is where you transmit a small data packet and receive confirmation that it arrived.

Most operating systems support a *ping* command which does this from your device. The round-trip-time (RTT) displays in milliseconds and gives you an idea of how long it takes for your network to transfer data.

You can measure throughput either with network testing tools or manually. If you wanted to test throughput manually, you would send a file and divide the file size by the time it takes to arrive. However, latency and bandwidth impact throughput. Because of this, many people use network testing tools, as the tools report throughput alongside other factors like bandwidth and latency.

Read about RTT in networking »

Unit of measurement

You measure latency in milliseconds. If you have a low number of milliseconds, your network is only experiencing a small delay. The higher the number in milliseconds, the slower the network is performing.

Originally, you would measure network throughput in bits per second (bps). But, as data transmission technologies have improved, you can now achieve much higher values. Because of this, you can measure throughput in kilobytes per second (KBps), megabytes per second (MBps), and even gigabytes per second (GBps). One byte is equal to eight bits.

Impacting factors: latency vs throughput

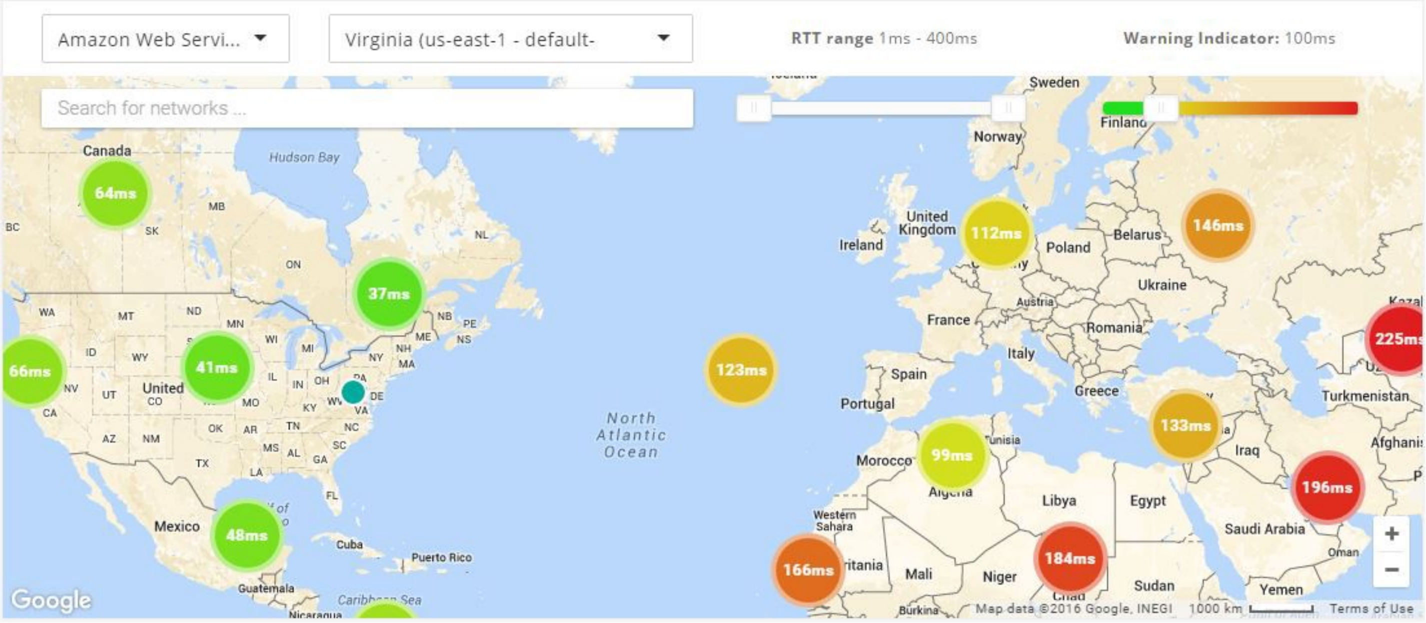
Different factors can impact your latency and throughput metrics.

Latency

Latency has several factors that contribute to it being high or low.

Location

One of the most important factors is the location of where data originates and its intended destination. If your servers are in a different geographical region from your device, the data has to travel further, which increases latency. This factor is called *propagation*.



Amazon Web Servi... ▾ | Virginia (us-east-1 - default- ▾

RTT range 1ms - 400ms

Warning Indicator: 100ms

Search for networks ...

Network congestion

Network congestion occurs when there is a high volume of data being transmitted over a network. The increased traffic on the network causes packets to take longer routes to their destination.

Protocol efficiency

Some networks require additional protocols for security. The extra handshake steps create a delay.

Network infrastructure

Network devices can become overloaded, which results in dropped packets. As packets are delayed or dropped, devices retransmit them. This adds additional latency.

Throughput

Throughput speeds are directly impacted by other factors.

Bandwidth

If your network capacity has reached the maximum bandwidth of your transmission medium, its throughput will never be able to go beyond that limit.

Processing power

Certain network devices have specialized hardware or software optimizations that improve their processing performance. Some examples are dedicated application-specific integrated circuits or software-based packet processing engines.

These optimizations enable the device to handle higher volumes of traffic and more complex packet processing tasks, which leads to higher throughput.

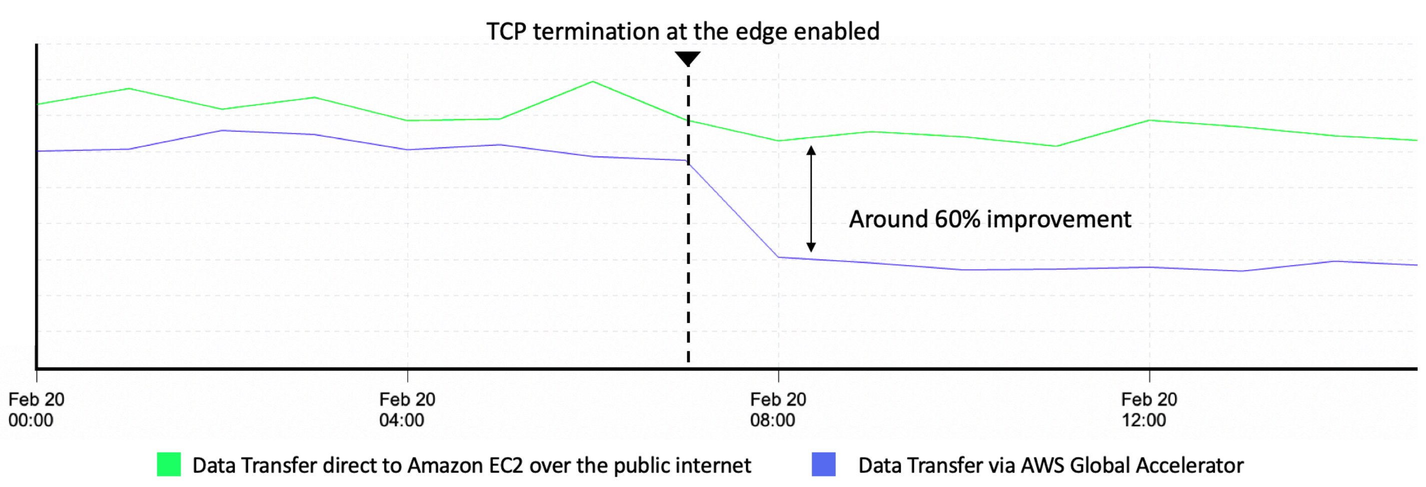
Packet loss

Packet loss can occur for a variety of reasons, including network congestion, faulty hardware, or misconfigured network devices. When packets are lost, they must be retransmitted. This results in delays and reduces the overall throughput of the network.

Network topology

Network topology refers to the number of network devices, the bandwidth of the network links, and the distance between devices in a network path.

A well-designed network topology provides multiple paths for data transmission, reduces traffic bottlenecks, and increases throughput. Networks with more devices or longer distances require complex network topologies to achieve high throughput.



Relationship between bandwidth, latency and throughput

As they are closely linked, you must monitor both latency and throughput to achieve high network performance.

Bandwidth and network throughput

Bandwidth represents the total volume of data that you can transfer over a network. Your total bandwidth refers to the theoretical maximum amount of data that you could transfer over a network. You measure it in megabytes per second (MBps). You can think of bandwidth as the theoretical maximum throughput of your network.

Bandwidth is how much data you can transfer, while throughput is the actual amount of data you transmit in any given moment based on real-world network limitations. A high bandwidth does not guarantee speed or a good network performance, but a higher bandwidth leads to higher throughput.

How can you improve latency and throughput?

To improve latency, you can shorten the propagation between the source and destination. You can improve throughput by increasing the overall network bandwidth.

Next, we give some suggestions to improve latency and throughput together.

Caching

Caching in networking refers to the process of storing frequently accessed data geographically closer to the user. For example, you can store data in proxy servers or content delivery networks (CDNs).

Your network can deliver data from the cached location much faster than if it had to be retrieved from the original source. And the user receives data much faster, improving latency. Additionally, because the data is retrieved from a cache, it reduces the load on the original source. This allows it to handle more requests at once, improving throughput.

Transport protocols

By optimizing the transport protocol that you use for specific applications, you can improve network performance.

For instance, TCP and UDP are two common network protocols. TCP establishes a connection and checks that you receive data without any errors. Because of its goal of reducing packet loss, TCP has higher latency and higher throughput. UDP does not check for packet loss or errors, transmitting several duplicate packets instead. So, it gives minimal latency but a higher throughput.

Depending on the application that you are using, TCP or UDP may be the better choice. For example, TCP is useful for transferring data, while UDP is useful for video streaming and gaming.

Quality of service

You can use a quality of service (QoS) strategy to manage and optimize network performance. QoS allows you to divide network traffic into specific categories. You can assign each category a priority level.

Your QoS configurations prioritize latency-sensitive applications. Some applications and users experience lower latency than others. Your QoS configurations can also prioritize data by type, reducing packet loss and increasing throughput for certain users

Summary of differences: throughput vs. latency

	Throughput	Latency
What does it measure?	Throughput measures the volume of data that passes through a network in a given period. Throughput impacts how much data you can transmit in a period of time.	Latency measures the time delay when sending data. A higher latency causes a network delay.
How to measure?	Manually calculate throughput by sending a file or using network testing tools.	Calculate latency by using ping times.
Unit of measurement	Megabytes per second (MBps).	Milliseconds (ms).
Impacting factors	Bandwidth, network processing power, packet loss, and network topology.	Geographical distances, network congestion, transport protocol, and network infrastructure.

How can AWS support your network performance requirements?

Amazon Web Services (AWS) has a number of solutions to reduce network latency and improve network throughput. You can implement any of the following services, depending on your requirements:

- Amazon CloudFront is a content delivery network service built for high performance, security, and developer convenience. You can use it to securely deliver content with low latency and high transfer speeds.
- AWS Direct Connect is a cloud service that links your network directly to AWS to deliver more consistent and lower network latency. When creating a new connection, you can choose a hosted connection that an AWS Direct Connect Delivery Partner provides, or choose a dedicated connection from AWS to deploy at over 100 AWS Direct Connect locations around the world.
- AWS Global Accelerator is a networking service that improves the performance of your users' traffic by up to 60% by using the AWS global network infrastructure. When the internet is congested, AWS Global Accelerator optimizes the path to your application to keep packet loss, jitter, and latency consistently low.
- AWS Local Zones are a type of infrastructure deployment that places compute, storage, database, and other select AWS services close to large population and industry centers. You can deliver innovative applications requiring low latency closer to end users and on-premises installations.

Get started with optimizing your throughput and latency on AWS by creating an account today.

Next Steps with AWS

Learn

What Is AWS?

What Is Cloud Computing?

What Is Agentic AI?

Cloud Computing Concepts Hub

AWS Cloud Security

What's New

Blogs

Press Releases

Resources

Getting Started

Training

AWS Trust Center

AWS Solutions Library

Architecture Center

Product and Technical FAQs

Analyst Reports

AWS Partners

Developers

Builder Center

SDKs & Tools

.NET on AWS

Python on AWS

Java on AWS

PHP on AWS

JavaScript on AWS

Help

Contact Us

File a Support Ticket

AWS re:Post

Knowledge Center

AWS Support Overview

Get Expert Help

AWS Accessibility

Legal

Back to top ↑