A complete guide to understanding, monitoring and fixing network latency.

Table of Contents

What is network latency?

Network latency, or lag, is the term used to describe delays in communication over a network. In networking, it is best thought of as the amount of time taken for a packet of data to travel through multiple devices, then be received at its destination and decoded.

When delays in transmission are small, it's referred to as a low-latency network (desirable) and longer delays are called a high-latency network (not so desirable).

Long delays that occur in high-latency networks create bottlenecks in communication. In the worst cases, it's like traffic on a four-lane highway trying to merge into a single lane. High latency decreases communication bandwidth, and can be temporary or permanent, depending on the source of the delays.

Latency is measured in milliseconds, or during speed tests, it's referred to as a ping rate. The lower the ping rate the better the performance. A ping rate of less than 100ms is considered acceptable but for optimal performance, latency in the range of 30-40ms is desirable. Obviously, zero to low latency in communication is what we all want. However, standard latency for a network is explained slightly differently in various contexts, and latency issues also vary from one network to another.

Causes of network latency?

1. Distance

One of the main causes of network latency is distance, or how far away the device making requests is located from the servers responding to those requests.

For example, network latency between cities: if a website is hosted in a data center in Trenton, New Jersey, it will respond faster to requests from users in Farmingdale, NY (100 miles away), or most likely within 10-15 milliseconds. On the other hand, users in Denver, Colorado (about 1,800 miles away) will face longer delays of up to 50 milliseconds.

The amount of time it takes for a request to reach a client device is referred to as Round Trip Time (RTT). While an increase of a few milliseconds might seem negligible, there are other considerations that can increase latency.

- There's the to-and-fro communication necessary for the client and server to make that connection in the first place.
- The total size and load time of the page
- Problems with network hardware which the data passes through along the way.

Data travelling back and forth across the internet often has to cross multiple Internet Exchange Points (IXPs), where routers process and route the data packets, often having to break them up into smaller packets. All this additional activity adds a few milliseconds to RTT.

2. Website construction

The way webpages are constructed makes a difference latency. Webpages that carry heavy content, large images, or load content from several third- party websites may perform more slowly, as browsers need to download larger files to display them.

3. End-user issues

Network problems might appear to be responsible for latency, but sometimes RTT latency is the result of the end-user device being low on memory or CPU cycles to respond in a reasonable timeframe.

4. Physical issues

In a physical context, common network latency causes are the components that move data from one point to the next. Physical cabling such as routers, switches and WiFi access points. In addition, latency can be influence by other network devices like application load balancers, security devices, firewalls and Intrusion Prevention Systems (IPS).

Latency vs bandwidth vs throughput

Latency, bandwidth and throughput are all equal contributors to the quality of communications. While these three factors work together, they have different meanings. To understand it better, you could imagine that data packets flow through a pipe:

Bandwidth is the width of the pipe. The narrower the pipe, the less data allowed to travel back and forth through it. The wider the communication band, the more data that can flow through it simultaneously.

Latency is how fast the data packets inside the pipe travel from client to server and back. Packet latency is dependent on the physical distance that data must travel through cords, networks and the like to reach its destination.

Throughput is the volume of data that can be transferred over a specified time period.

Low latency and low bandwidth means that throughput will also be low. This means that while data packets should technically be delivered without delay, a low bandwidth means there can still be considerable congestion. But with high bandwidth, low latency, then throughput will be greater and the connection much more efficient.

Other types of latency

Now that we have determined the meaning of global latency and its effects on smooth communications, the following describes two other examples of the effects of latency.

Fiber optic latency

In the case of fiber optic networks, latency refers to the time delay that affects light as it travels through the fiber optic network. Latency increases over the distance traveled, so this must also be factored in to compute the latency for any fiber optic route.

Based on the speed of light (299,792,458 meters/second), there is a latency of 3.33 microseconds (0.000001 of a second) for every kilometer covered. Light travels slower in a cable which means the latency of light traveling in a fibre optic cable is around 4.9 microseconds per kilometer.

The quality of fiber optic cable is an important factor in reducing latency in a network.

VoIP latency

The reasons behind audio latency are based on the speed of sound. Latency in VoIP is the difference in time between when a voice packet is transmitted and the moment it reaches its destination. A latency of 20 ms is normal for VoIP calls; a latency of up to 150 ms is barely noticeable and therefore acceptable.

Any higher than that, however, and quality starts to diminish. At 300 ms or higher, it becomes completely unacceptable

High latency in VoIP can severely affect call quality, resulting in:

- Slow and interrupted phone conversations
- Overlapping noises, with one speaker interrupting the other
- Echo
- Disturbed synchronization between voice and other data types, especially during video conferencing

Reasons behind VoIP latency and how to address them:

Insufficient bandwidth – with a slow internet connection, insufficient bandwidth means that data packets take more time reach their destination, and often arrive in the wrong order.

Firewall blocking traffic – to prevent bottlenecks, always allow clearance for your VoIP applications within your firewall software.

Wrong codecs – codecs encode voice signals into digital data ready to be transmitted. This is often an issue that your provider needs to solve, however some VoIP apps allow you to tweak codecs.

Outdated hardware – Sometimes the mix of old hardware and new software can cause latency problems. Changing your

telephone adaptor or other VoIP-specific software can help. Even your headset can cause latency.

Signal conversion – If your system is converting your signal to or from analog and digital, this could cause latency.

Best practices for monitoring and improving network latency

The slowing of your network can be extremely problematic in the business world, where time is such a precious commodity. As your network grows bigger, having additional connections means more points where delays and issues can happen.

Problems can increase again as more and more organizations connect to cloud servers, use more applications and expand to accommodate remote workers extra branch offices.

Everyone has experienced latency in various aspects of daily business, and it can severely threaten deadlines, expected outcomes and eventually ROI. This is where comprehensive network monitoring and troubleshooting comes into its own. Network monitoring and troubleshooting can quickly and accurately diagnose and identify the root causes of latency and put solutions in place to reduce and improve the problem.

Before you can do anything to improve your network latency, you need to know how to calculate and measure it. By becoming familiar with your latency, you're far better equipped to troubleshoot.

How to Check Network Latency

If you feel that your network is running slow, you can check your latency manually by using Windows. Open a command

prompt and type tracert followed by the destination you'd like to query, such as cloud.google.com.

How to Measure Network Latency

Network monitoring and management tools will get this information automatically, but here's how to do it manually. Once you type in the tracert command, you'll see a list of all routers on the path to that website address, followed by a time measurement in milliseconds (ms). Add up all the measurements, and the resulting quantity is the latency between your machine and the website in question.

Latency can either be measured as the Round Trip Time (RTT) or the Time to First Byte (TTFB):

- ***RTT*** - the amount of time it takes a packet to get from the client to the server and back.
- ***TTFB*** - the amount of time it takes for the server to receive the first byte of data when the client sends a request.

How to Reduce Network Latency

One simple way to improve network latency is to check that others on your network aren't unnecessarily using up your bandwidth, or increasing your latency with excessive downloads or streaming. Then, check application performance to determine whether applications are acting unexpectedly and potentially placing pressure on the network.

Subnetting is another way to help reduce latency across your network, by grouping together endpoints that communicate most frequently with each other.

Additionally, you could use traffic shaping and bandwidth allocation to improve latency for the business-critical parts of your network.

Finally, you can use a load balancer to help offload traffic to parts of the network with the capacity to handle some additional activity.

How to Troubleshoot Network Latency Issues

Manually troubleshooting issues across a large network can become complex, which highlights again the importance of network monitoring and troubleshooting tools.

To check if any of the devices on your network are specifically causing issues, you can try disconnecting computers or network devices and restarting all the hardware. You'll need to ensure that you have network monitoring deployed.

If you still have latency problems after checking all your local devices, it's the issues could be coming from the destination you're trying to connect to.

How to Test Network Latency

Testing network latency can be done by using ping or traceroute (tracert), although, comprehensive network monitoring and performance managers can test and check latency more accurately.

Maintaining a reliable network is an important part of a smoothly operating business. Network issues can become worse if they're not managed properly.

What Tools Help Improve Network Latency?

Network monitoring and troubleshooting tools are the best way to keep tabs on latency, as well as the other most troubling network problems, packet loss and jitter. You can typically set network standard expectations for latency and create alerts when the network latency reaches a certain threshold above this baseline.

Network monitoring tools can help you set up data comparisons between different metrics. This can help you identify performance issues, such as application performance or **errors also affecting network latency.**

A network mapping tool can also help you pinpoint *where* within the network latency the performance issues are occurring, which allows you to troubleshoot problems more quickly.

Specific traceroute tools monitor packets and how they move across an IP network, including how many "hops" the packet took, the roundtrip time, best time (in milliseconds), as well as the IP addresses and countries the packet traveled through.

By improving your network speed and reducing latency, your business processes will also make leaps and bounds towards efficiency and high performance.

Summary - addressing network latency

This comprehensive guide has been created to define network latency and to help identify, understand and troubleshoot the most common problems related to latency in computer networks.

The key takeaways are that network latency, jitter, and network packet loss can severely impede clear communication and universally affect your user experience (UX). Ensuring a low latency often means a positive UX whilst a high latency can affect this negatively and result in poor UX.

For further insightful information on network performance complications, download our additional guides on the full explanation of latency, jitter and packet loss:

› **What is Network Packet Loss? A Complete Guide to Understanding, Monitoring and Fixing Network Packet Loss.**

› **What is Network Jitter? A Complete Guide to Understanding, Monitoring and Fixing Jitter.**

Prognosis UC Assessor is a 100% software-based solution that can find and fix problems before migration without the need for network probes.

- Ensure a positive end-user experience with one-click troubleshooting for all network issues affecting UC performance. Deployment and getting started is quick, generating insights within minutes of installation across multiple sites within your environment.
- You can improve IT efficiency with the ability to operate and troubleshoot your entire multi-vendor UC environment from a single viewing point.
- Reduce costly outages and service interruptions with automated, intelligent alerts.