Network Optimization Strategies

by Sophia

WHAT'S COVERED

Regardless of how well a network is functioning, you should never stop trying to optimize its performance. This is especially true when latency-sensitive applications such as Voice-Over Internet Protocol (VoIP), streaming video, and web conferencing are implemented. In this lesson, you will learn some techniques you can use to ensure that these applications and services deliver on their promise of increased functionality.

Specifically, this lesson will cover the following:

1. Factors for Evaluating Network Performance

When many people try to access one network resource, like a valuable file server or a shared database, the systems can get as bogged down and clogged as a freeway at rush hour. This is why optimizing performance is in everyone's best interest—it keeps you and your network's users happily humming along.

Here are some ways that network optimization improves a network in practical, measurable ways.

1a. Latency Sensitivity

Most of us have clicked an icon to open an application or a web link only to have the computer just sit there staring back at us, helplessly hanging. That sort of lag comes when the resources needed to open the program or take us to the next page are not fully available. That kind of lag on a network is called **latency**—the time between when data is requested and the moment it actually gets delivered. The more latency, the longer the delay and the longer you have to stare blankly back at your computer screen, hoping something happens soon.

Latency affects some programs more than others. If you are sending an email, it may be annoying to have to wait a few seconds for the email server to respond, or for a video to begin playing, it's no big deal. Applications that *are* adversely affected by latency are said to have high **latency sensitivity**. A common example of this is online gaming where player reflexes can make a difference between a win and a loss.

TERMS TO KNOW

Latency
The time between when data is requested and the moment it actually gets delivered.

Latency Sensitivity
The degree to which an application is adversely affected by latency.

1b. Support for High-Bandwidth Applications

Many of the applications we now use over the network would have been totally unserviceable in the past because of the high amount of bandwidth they consume. And even though technology is constantly improving to give us more bandwidth, developers are in hot pursuit, developing new applications that gobble up that bandwidth as soon as it becomes available. For example:

VoIP Voice over Internet Protocol (VoIP)	Describes technologies that deliver voice (and sometimes video and data) communications over the internet or other data networks. Its most common application is video teleconferencing. But sadly, VoIP installations can be stressed heavily by things like really low bandwidth, latency issues, packet loss, jitter, security flaws, and reliability concerns.
Video Streaming	Although viewing digital media online is very common, it requires lots of bandwidth. And excessive use can cause traffic problems even on the most robust networks.

TERMS TO KNOW

Voice-Over Internet Protocol (VoIP)
Technologies that deliver voice (and sometimes video and data) communications over the internet or other data networks.

Video Streaming
Continuous transmission of video files from a server to a client.

1c. Support for Real-Time Services

Besides VoIP and streaming video, your network probably also supports other real-time services. For example, **presence** is a function of some collaboration apps that indicates a user's availability. If enabled across multiple communication tools, it also can help determine the communication channel on which the user is currently active and therefore the channel that provides the best possibility of an immediate response.

Multicast vs. Unicast

Transmissions that are **unicast** send data from one single device to another single device. In contrast, **multicast** transmissions send information from a single source to multiple recipients; it is commonly used for video streaming and conferencing. While unicast transmission creates a data connection and stream for each recipient, multicast uses the same stream for all recipients. This single stream is replicated as needed by multicast routers and switches in the network. The stream is limited to branches of the network topology that actually have subscribers to the stream. This greatly reduces the use of bandwidth in the network.

TERMS TO KNOW

Presence

A function of some collaboration apps that indicates a user's availability.

Unicast

Transmissions that send data from one single device to another single device.

Multicast

Transmissions send information from a single source to multiple recipients.

1d. Increased Uptime

Better network performance is not just about bandwidth and transfer rates; it's also about network availability. **Uptime** is the amount of time the system is up and accessible to your end users, so the more uptime you have the better. And depending on how critical the nature of your business is, you may need to provide **four-nine uptime** (99.99%) or **five-nine uptime** (99.999%) on your network.

TERMS TO KNOW

Uptime

The amount of time the system is up and accessible to end users.

Four-Nine Uptime

99.99% uptime.

Five-Nine Uptime

99.999% uptime.

2. Optimizing Performance

Here are some ways to make your network as responsive, reliable, and delay-free as possible for the clients you support.

2a. Quality of Service

Quality of Service (QoS) refers to the way the resources are controlled so that the quality of services is maintained. It's basically the ability to provide a different priority to one or more types of traffic over other levels for different applications, data flows, or users so that they can be guaranteed a certain performance level. QoS methods focus on one of five problems described in the table below that can affect data as it traverses a network cable.

Delay	Data can run into congested lines or take a less-than-ideal route to the destination, and delays like these can make some applications, such as VoIP, fail. This is the best reason to implement QoS when real-time applications are in use in the network—to prioritize delay-sensitive traffic.
Dropped Packets	Some routers will drop packets if they receive them while their buffers are full. If the receiving application is waiting for the packets but doesn't get them, it will usually request that the packets be retransmitted—another common cause of a service(s) delay.
Errors	Packets can be corrupted in transit and arrive at the destination in an unacceptable format, again requiring retransmission and resulting in delays.
Jitter	Not every packet takes the same route to the destination, so some will be more delayed than others if they travel through a slower or busier network connection. The variation in packet delay is called jitter, and this can have a nasty negative impact on programs that communicate in real time.
Out-of-Order	Out-of-order delivery is also a result of packets taking different paths through the network to their destinations. The application at the receiving end needs to put them back together in the right order for the message to be completed, so if there are significant delays or the packets are reassembled out of order, users will probably notice degradation of an application's quality.

QoS can ensure that applications with a required bit rate receive the necessary bandwidth to work properly. Clearly, on networks with excess bandwidth, this is not a factor, but the more limited your bandwidth is, the more important a concept like this becomes.

One of the methods of providing QoS is differentiated services code point (DSCP), or **DiffServ**. DiffServ uses a code in the IP header's Differentiated Services field (DS field) to classify packets. This enables software to assign priorities to various traffic classes.

In theory, a network could have up to 64 different traffic classes using different DSCPs, but most networks use the following traffic classifications:

- Default, which is typically best-effort traffic
- Expedited Forwarding (EF), which is dedicated to low-loss, low-latency traffic
- Assured Forwarding (AF), which gives assurance of delivery under prescribed conditions
- Class Selector, which maintains backward compatibility with the IP Precedence field

A second method of providing traffic classification and thus the ability to treat the classes differently is a 3-bit field called the **priority code point (PCP)** within an Ethernet frame header when frames with virtual local area network (VLAN) tag are used. The IEEE 802.1p standard describes eight different classes of service as expressed through the 3-bit PCP field:

Level	Description
0	Best effort
1	Background
2	Standard (spare)
3	Excellent load (business-critical applications)
4	Controlled load (streaming video)
5	Voice and video (interactive voice and video, less than 100 ms latency and jitter)
6	Layer 3 Network Control Reserved Traffic (less than 10 ms latency and jitter)
7	Layer 2 Network Control Reserved Traffic (lowest latency and jitter)

QoS levels are established per call, per session, or in advance of the session by an agreement known as a **service-level agreement (SLA)**.

TERMS TO KNOW

Quality of Service (QoS)

Controlling network traffic so that the quality of latency-sensitive traffic is maintained.

DiffServ

Short for differentiated services code point (DSCP), a method of categorizing packets using a code in the IP header's DS field.

Priority Code Point (PCP)

A method of categorizing packets using a field in an Ethernet frame header when VLAN-tagged frames are used.

Service-Level Agreement (SLA)

An agreement worked out in advance between a provider and customer regarding the expected QoS level.

2b. Unified Communications

Increasingly, workers and the organizations for which they work are relying on new methods of communicating and working together. **Unified communications (UC)** is the integration of real-time communication services such as instant messaging with non-real-time communication services such as unified messaging (integrated voicemail, email, SMS, and fax). UC allows an individual to send a message on one medium and receive the same communication on another medium.

UC systems are made of several components that make sending a message on one medium and receiving the same communication on another medium possible. The following table has a list of some of the components that may be part of a UC system.

UC Server	The UC server is the heart of the system. It provides call control mobility services and administrative functions. It may be a stand-alone device or in some cases a module that is added to a router.
UC Devices	UC devices are the endpoints that may participate in unified communications. This includes computers, laptops, tablets, and smartphones.
UC Gateways	UC gateways are used to tie together geographically dispersed locations that may want to make use of UC facilities. They are used to connect the IP-based network with the public switched telephone network (PSTN).

TERM TO KNOW

Unified Communications (UC)

The integration of real-time communication services such as instant messaging with non-real-time communication services such as unified messaging (integrated voicemail, email, SMS, and fax).

2c. Traffic Shaping

Traffic shaping, also called “packet shaping”, is another form of bandwidth optimization. It works by delaying packets that meet certain criteria to guarantee usable bandwidth for other applications. Traffic shaping is basically traffic triage—you’re really just delaying attention to some traffic so other traffic gets A-listed through. Traffic shaping uses bandwidth throttling to ensure that certain data streams don’t send too much data in a specified period of time as well as rate limiting to control the rate at which traffic is sent.

KEY CONCEPT

Most often, traffic shaping is applied to devices at the edge of the network to control the traffic entering the network, but it can also be deployed on devices within an internal network. The devices that control it have a traffic contract that determines which packets are allowed on the network and when. You can think of them as stoplights on busy freeway on-ramps, where only so much traffic is allowed onto the road at one time, based on predefined rules. Even so, some traffic (like carpools and emergency vehicles) is allowed on the road immediately.

TERM TO KNOW

Traffic Shaping

A form of bandwidth optimization that delays packets that meet a certain criteria to guarantee usable bandwidth for other applications.

2d. Load Balancing

Load balancing refers to a technique used to spread work out to multiple computers, network links, or other devices. Using load balancing, you can provide an active/passive server cluster in which only one server is active and handling requests.

REFLECT

Your favorite internet site might actually consist of 20 servers that all appear to be the same exact site because that site's owner wants to ensure that its users always experience quick access. You can accomplish this on a network by installing multiple redundant links to ensure that network traffic is spread across several paths and to maximize the bandwidth on each link. Think of this as having two or more different freeways that will both get you to your destination equally well—if one is really busy, just take the other one.

TERM TO KNOW

Load Balancing

A technique used to spread work out to multiple computers, network links, or other devices to optimize network performance.

2e. High Availability

High availability is a system-design protocol that guarantees a certain amount of operational uptime during a given period. The design attempts to minimize unplanned downtime—the time users are unable to access resources. In almost all cases, high availability is provided by duplicating equipment (multiple servers, multiple NICs, etc.). Organizations that serve critical functions obviously need this; after all, you really don't want to blaze your way to a hospital ER only to find that they can't treat you because their network is down!

KEY CONCEPT

One of the highest standards in uptime is the ability to provide five-nine availability. This means the network is accessible 99.999% of the time— impressive!

There's a difference between uptime and availability. Your servers may be up but not accessible if, say, a cable gets cut, and that outage would definitely count against your availability time.

TERM TO KNOW

High Availability

A system-design protocol that guarantees a certain amount of operational uptime during a given period.

2f. Caching Engines

A **cache** is a quick-access storage area that holds duplicate copies of data that may be needed soon. Computers use caches all the time to temporarily store information for faster access, and processors have both internal and external caches available to them, which speeds up their response times.

A **caching engine** is basically a database on a server that stores information people need to access quickly. The most popular implementation of this is with web servers and proxy servers, but caching engines are also used on internal networks to speed up access to things like database services.

TERMS TO KNOW

Cache

A quick-access storage area that holds duplicate copies of data that may be needed soon.

Caching Engine

A database on a server that stores information people need to access quickly.

2g. Fault Tolerance

Fault tolerance is the ability of a resource to continue to be available when a component fails. To implement fault tolerance, you need to employ multiple devices or connections that all provide a way to access the same resource(s).

KEY CONCEPT

A familiar form of fault tolerance is configuring an additional hard drive to be a mirror image of another so that if either one fails, there's still a copy of the data available to you. In networking, fault tolerance means that you have multiple paths from one point to another. What's really cool is that fault-tolerant connections can be configured to be available either on a standby basis only or all the time if you intend to use them as part of a load-balancing system.

TERM TO KNOW

Fault Tolerance

The ability of a resource to continue to be available when a component fails.

2h. Other Optimization to Consider

This lesson has only scratched the surface of the optimization techniques available to network engineers and administrators—and it also hasn't addressed optimization at a big-picture level. Many of the most effective ways to improve a network involve significant expense, such as upgrading outmoded hardware devices to newer, faster models or restructuring the network's architecture. Such big-picture decisions are typically not within the purview of the average network technician, but they are factors that the company's key IT decision makers might consider.

SUMMARY

In this lesson, you learned some ways to **evaluate a network's performance**. This included evaluating **latency sensitivity, support for high-band applications, support for real-time services**, and **increased uptime**. You also learned about several techniques for **optimizing performance**, including **QoS**, **Unified Communications, traffic shaping, load balancing, high availability, caching engines, fault tolerance**, and **other optimization to consider**.

Source: This content and supplemental material has been adapted from CompTIA Network+ Study Guide: Exam N10-007, 4th Edition. Source **Lammle: CompTIA Network+ Study Guide: Exam N10-007, 4th Edition - Instructor Companion Site (wiley.com)**

TERMS TO KNOW

Cache

A quick-access storage area that holds duplicate copies of data that may be needed soon.

Caching Engine

A database on a server that stores information people need to access quickly.

DiffServ

Short for differentiated services code point (DSCP), a method of categorizing packets using a code in the IP header's DS field.

Fault Tolerance

The ability of a resource to continue to be available when a component fails.

Five-Nine Uptime

99.999% uptime.

Four-Nine Uptime

99.99% uptime.

High Availability

A system-design protocol that guarantees a certain amount of operational uptime during a given period.

Latency

The time between when data is requested and the moment it actually gets delivered.

Latency Sensitivity

The degree to which an application is adversely affected by latency.

Load Balancing

A technique used to spread work out to multiple computers, network links, or other devices to optimize network performance.

Multicast

Transmissions send information from a single source to multiple recipients.

Presence

A function of some collaboration apps that indicates a user's availability.

Priority Code Point (PCP)

A method of categorizing packets using a field in an Ethernet frame header when VLAN-tagged frames are used.

Quality of Service (QoS)

Controlling network traffic so that the quality of latency-sensitive traffic is maintained.

Service-Level Agreement (SLA)

An agreement worked out in advance between a provider and customer regarding the expected QoS level.

Traffic Shaping

A form of bandwidth optimization that delays packets that meet a certain criteria to guarantee usable bandwidth for other applications.

Unicast

Transmissions that send data from one single device to another single device.

Unified Communications (UC)

The integration of real-time communication services such as instant messaging with non-real-time communication services such as unified messaging (integrated voicemail, email, SMS, and fax).

Uptime

The amount of time the system is up and accessible to end users.

Video Streaming

Continuous transmission of video files from a server to a client.

Voice-Over Internet Protocol (VoIP)

Technologies that deliver voice (and sometimes video and data) communications over the internet or other data networks.