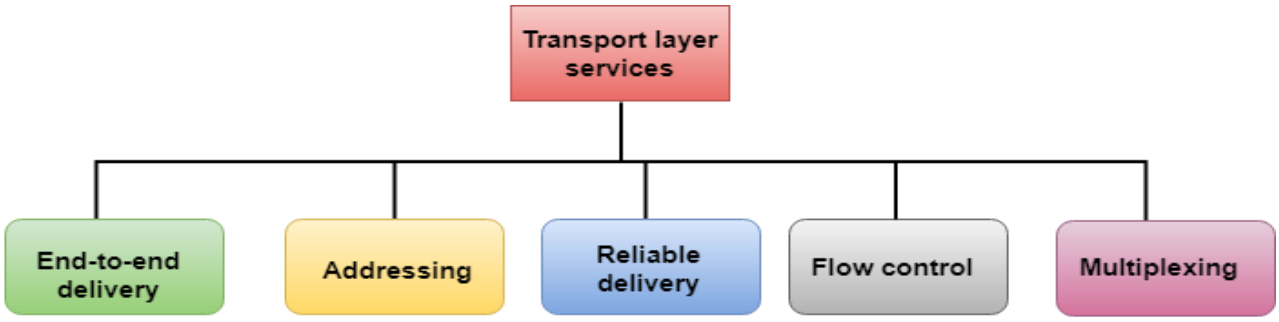
UNIT – IV
TRANSPORT LAYER

Introduction

The transport layer is the core of the OSI model. Protocols at this layer oversee the delivery of data from an application program on one device to an application program on another device. They act as a liaison between the upper-layer protocols (session, presentation, and application) and the services provided by the lower layers.

The services provided by the transport layer protocols can be divided into five categories:

End-to-end delivery
Addressing
Reliable delivery
Flow control
Multiplexing



End-to-end delivery:

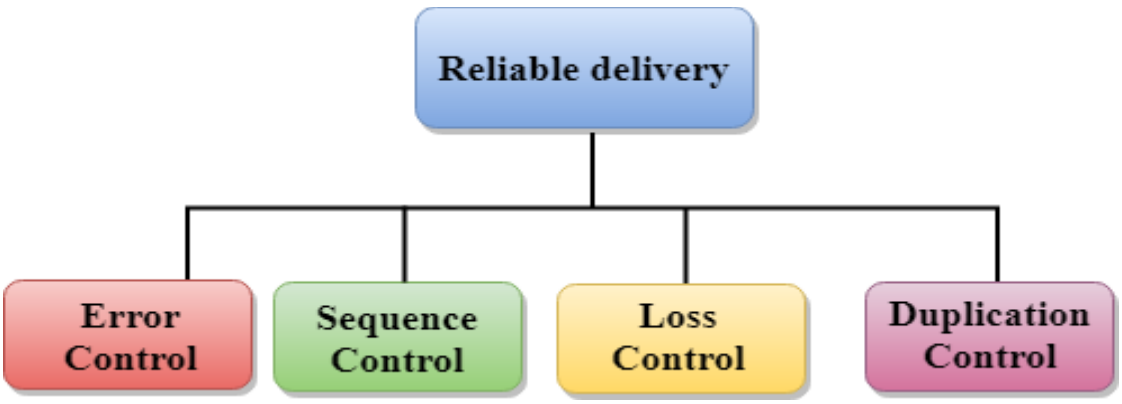
The transport layer transmits the entire message to the destination. Therefore, it ensures the end-to-end delivery of an entire message from a source to the destination.

Reliable delivery:

The transport layer provides reliability services by retransmitting the lost and damaged packets.

The reliable delivery has four aspects:

Error control
Sequence control
Loss control
Duplication control

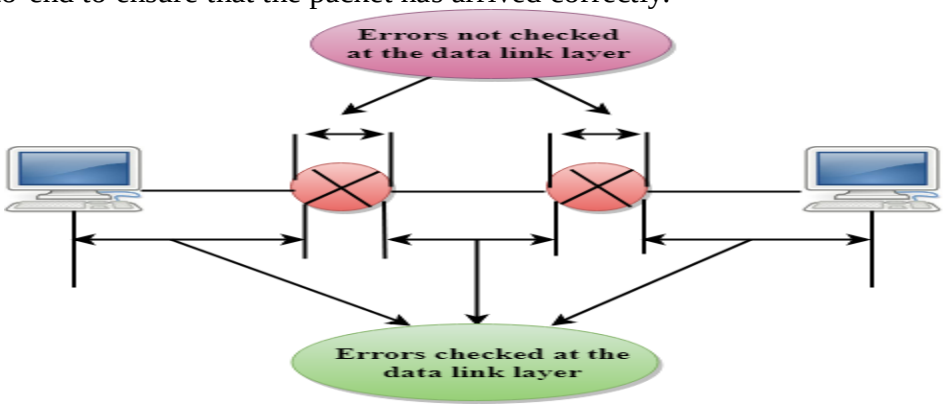


Error Control

The primary role of reliability is **Error Control**. In reality, no transmission will be 100 percent error-free delivery. Therefore, transport layer protocols are designed to provide error-free transmission.

The data link layer also provides the error handling mechanism, but it ensures only node-to-node error-free delivery. However, node-to-node reliability does not ensure the end-to-end reliability.

The data link layer checks for the error between each network. If an error is introduced inside one of the routers, then this error will not be caught by the data link layer. It only detects those errors that have been introduced between the beginning and end of the link. Therefore, the transport layer performs the checking for the errors end-to-end to ensure that the packet has arrived correctly.



Sequence Control

The second aspect of the reliability is sequence control which is implemented at the transport layer.

On the sending end, the transport layer is responsible for ensuring that the packets received from the upper layers can be used by the lower layers. On the receiving end, it ensures that the various segments of a transmission can be correctly reassembled.

Loss Control

Loss Control is a third aspect of reliability. The transport layer ensures that all the fragments of a transmission arrive at the destination, not some of them. On the sending end, all the fragments of transmission are given sequence numbers by a transport layer. These sequence numbers allow the receiver?s transport layer to identify the missing segment.

Duplication Control

Duplication Control is the fourth aspect of reliability. The transport layer guarantees that no duplicate data arrive at the destination. Sequence numbers are used to identify the lost packets; similarly, it allows the receiver to identify and discard duplicate segments.

Flow Control

Flow control is used to prevent the sender from overwhelming the receiver. If the receiver is overloaded with too much data, then the receiver discards the packets and asking for the retransmission of packets. This increases network congestion and thus, reducing the system performance. The transport layer is responsible for flow control. It uses the sliding window protocol that makes the data transmission more efficient as well as it controls the flow of data so that the receiver does not become overwhelmed. Sliding window protocol is byte oriented rather than frame oriented.

Multiplexing

The transport layer uses the multiplexing to improve transmission efficiency.

Multiplexing can occur in two ways:

Upward multiplexing: Upward multiplexing means multiple transport layer connections use the same network connection. To make more cost-effective, the transport layer sends several transmissions bound for the same destination along the same path; this is achieved through upward multiplexing.

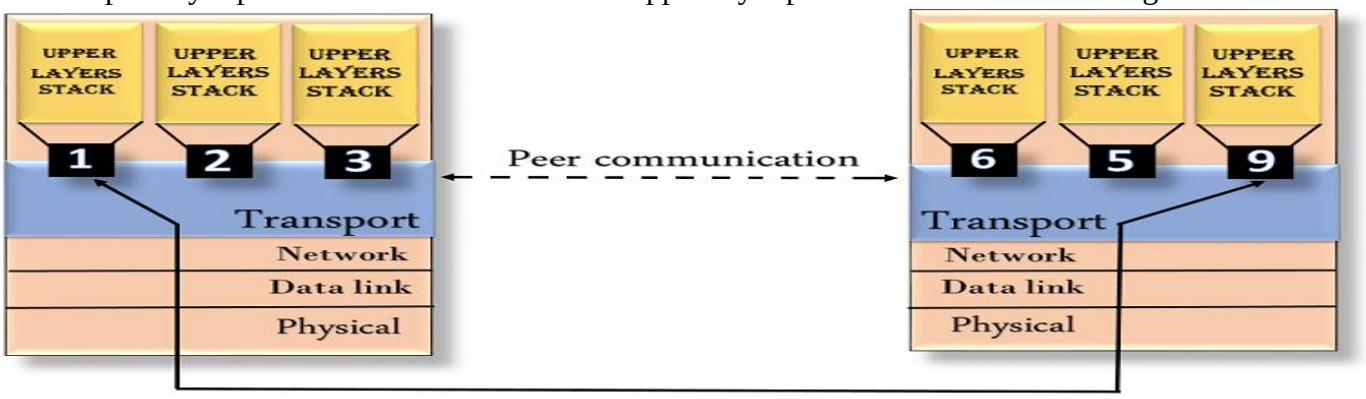
Downward multiplexing: Downward multiplexing means one transport layer connection uses the multiple network connections. Downward multiplexing allows the transport layer to split a connection among several paths to improve the throughput. This type of multiplexing is used when networks have a low or slow capacity.

Addressing

According to the layered model, the transport layer interacts with the functions of the session layer. Many protocols combine session, presentation, and application layer protocols into a single layer known as the application layer. In these cases, delivery to the session layer means the delivery to the application layer. Data generated by an application on one machine must be transmitted to the correct application on another machine. In this case, addressing is provided by the transport layer.

The transport layer provides the user address which is specified as a station or port. The port variable represents a particular TS user of a specified station known as a Transport Service access point (TSAP). Each station has only one transport entity.

The transport layer protocols need to know which upper-layer protocols are communicating.



Quality of Service

The transport protocol improves the QoS (Quality of Service) provided by the network layer. Following are the QoS parameters:

Connection establishment delay:

The connection establishment delay is the amount of time elapsing between a transport connection being requested and the confirmation being received by the user of the transport service. It includes the processing delay in the remote transport entity. As with all parameters measuring a delay, the shorter the delay, the better the service.

Connection establishment failure probability:

The connection establishment failure probability is the chance of a connection not being established within the maximum establishment delay time, for example, due to network congestion, lack of table space somewhere, or other internal problems.

Throughput:

The throughput parameter measures the number of bytes of user data transferred per second, measured over some time interval. The throughput is measured separately for each direction.

Transit delay:

The transit delay measures the time between a message being sent by the transport user on the source machine and its being received by the transport user on the destination machine. As with throughput, each direction is handled separately.

The Residual error ratio :

Measures the number of lost or garbled messages as a fraction of the total sent. In theory, the residual error rate should be zero, since it is the job of the transport layer to hide all network layer errors. In practice it may have some (small) finite value.

The Protection parameter provides a way for the transport user to specify interest in having the transport layer provide protection against unauthorized third parties (wiretappers) reading or modifying the transmitted data.

The Priority parameter provides a way for a transport user to indicate that some of its connections are more important than other ones, and in the event of congestion, to make sure that the high-priority connections get serviced before the low-priority ones.

End –to –end delivery

The network layer oversees the end-to-end delivery of individual packets but does not see any relationship between those packets, even those belonging to a single message.

It treats each as an independent entity. The transport layer, on the other hand, makes sure that the entire message (not just a single packet) arrives intact. Thus, it oversees the end-to-end (source –to-destination) delivery of an entire message.

Addressing

The transport layer interacts with the functions of the session layer. However, many protocols (or protocol stacks, meaning groups of protocols that interact at different levels) combine session, presentation, and application level protocols into a single packages, called an application. In these cases, delivery to the session layer functions is, in effect, delivery to the application. In these cases, delivery to the session layer functions is, in effect, delivery to the application. So communication occurs not just from end machine to end machine but from end application to end application. Data generated by an application on one machine must be received not just by the other machine but by the correct application on that other machine.

To ensure accurate delivery from service access point to service access point, we need another level of addressing in addition to those at the data link and network levels. Data link level protocols need to know which two computers within a network are communicating. Network level protocols need to know which two computers within an internet are communicating. But at the transport level, the protocol needs to know which upper-layer protocols are communicating.

Reliable Delivery

At the transport layer, reliable delivery has four aspects: error control, sequence control, loss control, and duplication control.

Error Control

When transferring data, the primary goal of reliability is error control.

But if we already have error handling at the data link layer, why do we need it at the transport layer? Data link layer functions guarantee error-free delivery node-to-node for each link. However, node-to-node reliability does not ensure end-to-end reliability.

Sequence Control

The second aspect of reliability implemented at the transport layer is sequence control. On the sending end, the transport layer is responsible for ensuring that data units received from the upper layers are usable by the lower layers. On the receiving end, it is responsible for ensuring that the various pieces of a transmission are correctly reassembled.

Segmentation and Concatenation

When the size of the data unit received from the upper layer is too long for the network layer datagram or data link layer frame to handle, the transport protocol divides it into smaller, usable blocks. The dividing process is called segmentation. When, on the other hand, the size of the data units belonging to a single

session are so small that several can fit together into a single datagram or frame, the transport protocol combines them into a single data unit. The combing process is called concatenation.

Sequence Numbers

Most transport layer services add a sequence number at the end of each segment. If a longer data unit has been segmented, the numbers indicate the order for reassembly. If several shorter units have been concatenated, the numbers indicate the end of each submit and allow them to be separated accurately at the destination. In addition, each segment carries a field that indicates whether it is the final segment of a transmission or a middle segment with more still to come.

Loss Control

The third aspect of reliability covered by the transport layer is loss control. The transport layer ensures that all pieces of a transmission arrive at the destination, not just some of them. When data have been segmented for delivery, some segments may be lost in transit. Sequence numbers allow the receiver's transport layer protocol to identify any missing segments and request redelivery.

Duplication Control

The fourth aspect of reliability covered by the transport layer is duplication control. Transport layer functions must guarantee that no pieces of data arrive at the receiving system duplicated. Just as they allow identification of lost packets, sequence numbers allow the receiver to identify and discard duplicate segments.

Flow Control

Like the data link layer, the transport layer is responsible for flow control. However, flow control at this layer is performed end-to-end rather than across a single link. Transport layer flow control also uses a sliding window protocol. However, the window at the transport layer can vary in size to accommodate buffer occupancy.

Multiplexing

To improve transmission efficiency, the transport layer has the option of multiplexing. Multiplexing at this layer occurs two ways: upward, meaning that multiple transport layer connections use the same network connection, or downward, meaning that one transport-layer connection uses multiple network connections.

Elements of Transport Protocols

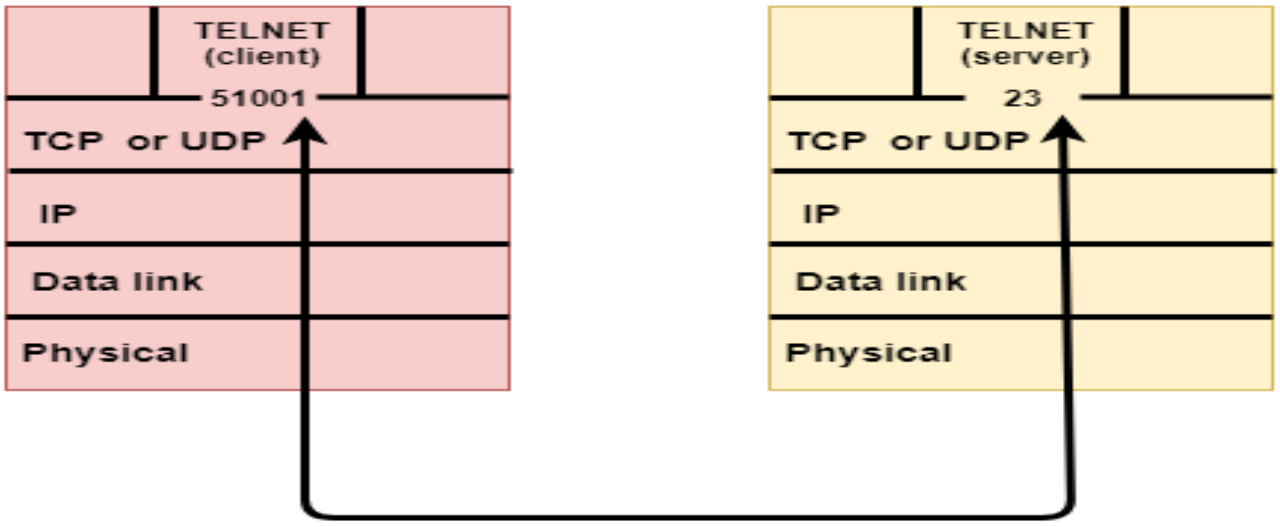
The transport service is implemented by a transport protocol used between the two transport entities. In some ways, transport protocols resemble the data link protocols. Both have to deal with error control, sequencing, and flow control, among other issues.

These important issues are as follow

- Addressing
- Connection Establishment
- Connection Release
- Flow Control and Buffering
- Multiplexing
- Crash Recovery

Transport Layer protocols

- The transport layer is represented by two protocols: TCP and UDP.
- The IP protocol in the network layer delivers a datagram from a source host to the destination host.
- Nowadays, the operating system supports multiuser and multiprocessing environments, an executing program is called a process. When a host sends a message to other host means that source process is sending a process to a destination process. The transport layer protocols define some connections to individual ports known as protocol ports.
- An IP protocol is a host-to-host protocol used to deliver a packet from source host to the destination host while transport layer protocols are port-to-port protocols that work on the top of the IP protocols to deliver the packet from the originating port to the IP services, and from IP services to the destination port.
- Each port is defined by a positive integer address, and it is of 16 bits.

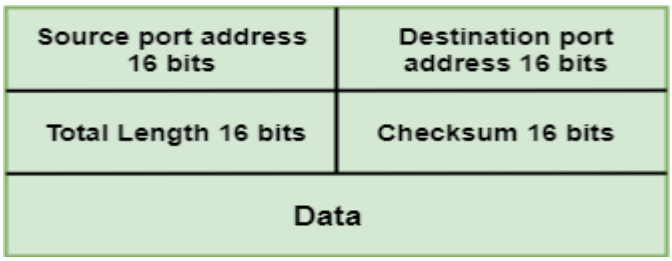


UDP

- UDP stands for **User Datagram Protocol**.
- UDP is a simple protocol and it provides nonsequenced transport functionality.
- UDP is a connectionless protocol.
- This type of protocol is used when reliability and security are less important than speed and size.
- UDP is an end-to-end transport level protocol that adds transport-level addresses, checksum error control, and length information to the data from the upper layer.
- The packet produced by the UDP protocol is known as a user datagram.

User Datagram Format

The user datagram has a 16-byte header which is shown below:



- **Source port address:** It defines the address of the application process that has delivered a message. The source port address is of 16 bits address.
- **Destination port address:** It defines the address of the application process that will receive the message. The destination port address is of a 16-bit address.
- **Total length:** It defines the total length of the user datagram in bytes. It is a 16-bit field.
- **Checksum:** The checksum is a 16-bit field which is used in error detection.

Disadvantages of UDP protocol

- UDP provides basic functions needed for the end-to-end delivery of a transmission.
- It does not provide any sequencing or reordering functions and does not specify the damaged packet when reporting an error.
- UDP can discover that an error has occurred, but it does not specify which packet has been lost as it does not contain an ID or sequencing number of a particular data segment.

TCP

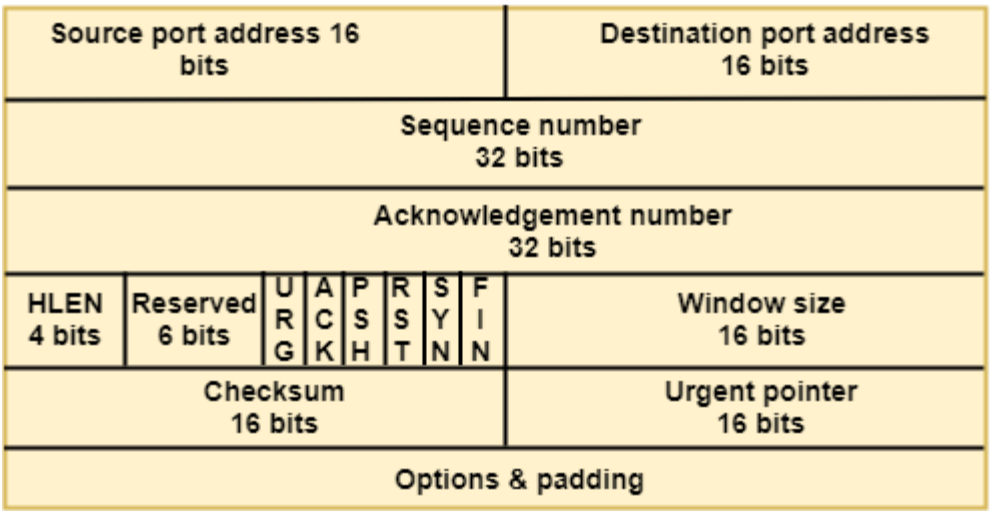
- TCP stands for Transmission Control Protocol.
- It provides full transport layer services to applications.
- It is a connection-oriented protocol means the connection established between both the ends of the transmission. For creating the connection, TCP generates a virtual circuit between sender and receiver for the duration of a transmission.

Features Of TCP protocol

- **Stream data transfer:** TCP protocol transfers the data in the form of contiguous stream of bytes. TCP group the bytes in the form of TCP segments and then passed it to the IP layer for transmission to the destination. TCP itself segments the data and forward to the IP.
- **Reliability:** TCP assigns a sequence number to each byte transmitted and expects a positive acknowledgement from the receiving TCP. If ACK is not received within a timeout interval, then the data is retransmitted to the destination.
  The receiving TCP uses the sequence number to reassemble the segments if they arrive out of order or to eliminate the duplicate segments.
- **Flow Control:** When receiving TCP sends an acknowledgement back to the sender indicating the number the bytes it can receive without overflowing its internal buffer. The number of bytes is sent in ACK in the form of the highest sequence number that it can receive without any problem. This mechanism is also referred to as a window mechanism.

- **Multiplexing:** Multiplexing is a process of accepting the data from different applications and forwarding to the different applications on different computers. At the receiving end, the data is forwarded to the correct application. This process is known as demultiplexing. TCP transmits the packet to the correct application by using the logical channels known as ports.
- **Logical Connections:** The combination of sockets, sequence numbers, and window sizes, is called a logical connection. Each connection is identified by the pair of sockets used by sending and receiving processes.
- **Full Duplex:** TCP provides Full Duplex service, i.e., the data flow in both the directions at the same time. To achieve Full Duplex service, each TCP should have sending and receiving buffers so that the segments can flow in both the directions. TCP is a connection-oriented protocol. Suppose the process A wants to send and receive the data from process B. The following steps occur:
- Establish a connection between two TCPs.
- Data is exchanged in both the directions.
- The Connection is terminated.

TCP Segment Format



- **Source port address:** It is used to define the address of the application program in a source computer. It is a 16-bit field.
- **Destination port address:** It is used to define the address of the application program in a destination computer. It is a 16-bit field.
- **Sequence number:** A stream of data is divided into two or more TCP segments. The 32-bit sequence number field represents the position of the data in an original data stream.
- **Acknowledgement number:** A 32-field acknowledgement number acknowledge the data from other communicating devices. If ACK field is set to 1, then it specifies the sequence number that the receiver is expecting to receive.

- **Header Length (HLEN):** It specifies the size of the TCP header in 32-bit words. The minimum size of the header is 5 words, and the maximum size of the header is 15 words. Therefore, the maximum size of the TCP header is 60 bytes, and the minimum size of the TCP header is 20 bytes.
- **Reserved:** It is a six-bit field which is reserved for future use.
- **Control bits:** Each bit of a control field functions individually and independently. A control bit defines the use of a segment or serves as a validity check for other fields.

There are total six types of flags in control field:

- **URG:** The URG field indicates that the data in a segment is urgent.
- **ACK:** When ACK field is set, then it validates the acknowledgement number.
- **PSH:** The PSH field is used to inform the sender that higher throughput is needed so if possible, data must be pushed with higher throughput.
- **RST:** The reset bit is used to reset the TCP connection when there is any confusion occurs in the sequence numbers.
- **SYN:** The SYN field is used to synchronize the sequence numbers in three types of segments: connection request, connection confirmation (with the ACK bit set), and confirmation acknowledgement.
- **FIN:** The FIN field is used to inform the receiving TCP module that the sender has finished sending data. It is used in connection termination in three types of segments: termination request, termination confirmation, and acknowledgement of termination confirmation.

Differences b/w TCP & UDP

Basis for Comparison	TCP	UDP
Definition	TCP establishes a virtual circuit before transmitting the data.	UDP transmits the data directly to the destination computer without verifying whether the receiver is ready to receive or not.
Connection Type	It is a Connection-Oriented protocol	It is a Connectionless protocol
Speed	slow	high
Reliability	It is a reliable protocol.	It is an unreliable protocol.
Header size	20 bytes	8 bytes

acknowledgement	It waits for the acknowledgement of data and has the ability to resend the lost packets.	It neither takes the acknowledgement, nor it retransmits the damaged frame.

Performance Problems in Computer Networks

Network performance refers to the quality and speed of a network's transmission of data between devices. It is typically measured by factors such as bandwidth, latency, and throughput.

Network performance is important because it determines how well devices can communicate with each other and access the resources they need, such as the internet or shared files. Poor network performance can lead to slow response times, reduced productivity, and other problems.

Five Common Potential Issues that can Affect Network Performance

- **Bandwidth bottlenecks** − If the network's available bandwidth is inadequate for the number and type of devices and applications using it, performance can suffer.
- **Interference** − Physical objects or other electronic devices can interfere with wireless signals, causing them to degrade and reducing network performance.
- **Congestion** − When too many devices are trying to use the network at the same time, congestion can occur, leading to slow performance.
- **Malware** − Malware, such as viruses and worms, can compromise the performance of individual devices and the network as a whole.
- **Outdated hardware or software** − Using outdated equipment or software can limit the network's capabilities and lead to poor performance.

Bandwidth Bottlenecks

A bandwidth bottleneck is a network performance issue that occurs when the available bandwidth of the network is not sufficient to handle the volume of data being transmitted. This can result in slow response times and decreased performance for devices on the network.

Bandwidth is the maximum amount of data that can be transmitted over a network at any given time, and it is typically measured in bits per second (bps). If the demand for data transmission exceeds the available bandwidth, a bottleneck can occur. This can be caused by a variety of factors, such as an increase in the number of devices on the network, the use of bandwidth-intensive applications or services, or the presence of bottlenecks at specific points in the network.

To address a bandwidth bottleneck, you may need to upgrade your network's infrastructure or optimize your use of bandwidth by prioritizing certain types of traffic or limiting the use of bandwidth-intensive applications. It may also be necessary to increase the available bandwidth by adding additional capacity to the network.

Interference

Interface-related network performance issues can occur when there are problems with the hardware or software interfaces that connect devices to the network. These issues can affect the ability of devices to communicate with each other and access network resources, leading to reduced performance.

Some common interface-related problems include −

- **Incorrectly configured interfaces** − If an interface is not configured properly, it may not be able to communicate with other devices on the network.
- **Faulty hardware** − Physical issues with an interface, such as a damaged connector or malfunctioning hardware, can prevent it from working properly.
- **Incompatible software** − If the software that controls an interface is not compatible with the rest of the network, it may not function correctly.

To troubleshoot interface-related network performance issues, you may need to check the configuration of the interfaces, replace faulty hardware, or update the software that controls the interfaces.

Congestion

Network congestion is a performance issue that occurs when there are too many devices trying to use the network at the same time. This can result in slow response times, dropped connections, and other problems.

There are several factors that can contribute to congestion on a network, including −

- An increase in the number of devices on the network − As more devices are added to a network, they compete for bandwidth and other resources, which can lead to congestion.
- The use of bandwidth-intensive applications − Applications that require a lot of bandwidth, such as streaming video or online gaming, can consume a large portion of the available bandwidth and cause congestion for other devices on the network.
- Limited bandwidth − If the available bandwidth of the network is insufficient to handle the volume of data being transmitted, congestion can occur.

To address congestion on a network, you may need to optimize your use of bandwidth by limiting the use of bandwidth-intensive applications or prioritizing certain types of traffic. You may also need to increase the available bandwidth by adding additional capacity to the network.

Malware

Malware is software that is specifically designed to harm or exploit computer systems. It can take many forms, including viruses, worms, Trojans, and ransomware. Malware can have a significant impact on network performance by consuming resources, slowing down devices, and disrupting communication between devices.

Some common ways in which malware can affect network performance include −

- **Consuming resources** − Malware can consume a large amount of a device's resources, such as CPU time and memory, which can lead to reduced performance.
- **Disrupting communication** − Malware can interfere with the normal operation of a device's network interface, preventing it from communicating with other devices on the network.
- **Spreading to other devices** − Malware can spread from one infected device to others on the network, potentially infecting multiple devices and causing widespread performance issues.

To address malware-related network performance issues, it is important to protect your devices with antivirus software and keep it up to date. You should also be careful about what you download and install, and avoid opening suspicious emails or visiting untrusted websites. If malware does infect your network, you may need to take steps to remove it, such as running antivirus software or restoring infected devices to a known good state.

Outdated Hardware or Software

Hardware and software issues can both affect network performance. Hardware problems can occur when there are physical issues with the devices or equipment that make up the network, while software problems can occur when there are issues with the programs and operating systems that run on the devices.

Some common hardware-related network performance issues include −

- **Faulty hardware** − If a device or piece of equipment is malfunctioning, it can disrupt communication on the network and reduce performance.
- **Incompatible hardware** − If the hardware on a device is not compatible with the rest of the network, it may not function correctly and could cause performance issues.
- **Insufficient hardware resources** − If a device does not have enough processing power, memory, or other resources, it may not be able to handle the demands placed on it, leading to reduced performance.

Some common software-related network performance issues include −

- **Outdated software** − If the software on a device is outdated, it may not be able to take advantage of newer technologies or may not be compatible with the rest of the network, leading to reduced performance.
- **Software bugs** − If the software on a device contains bugs or errors, it can cause performance issues or even crash the device.
- **Inefficient software** − If the software on a device is not optimized for performance, it may consume more resources than necessary, leading to reduced performance.

To address hardware and software-related network performance issues, you may need to upgrade or replace outdated or faulty equipment, update software to the latest version, or optimize the software on your devices to improve performance.

How do you measure Network Performance?

It is a qualitative and quantitative procedure that assesses and defines a network's performance level. Thus, it assists a network administrator in reviewing, evaluating, and improving network services.

Parameters Used to Measure Network Performance

The following parameters are used to measure Network Performance −

- Bandwidth
- Throughput
- Latency
- Packet Loss
- Jitter

Let us discuss each of these parameters in detail.

Bandwidth

The quantity of bandwidth allocated to the network is one of the most important conditions of a website's performance. The web server's bandwidth controls how quickly it can transfer the requested data. While there are many elements to consider regarding a site's speed, bandwidth is frequently the limiting issue.

The amount of data or information that can be transmitted in a given amount of time is referred to as bandwidth. The phrase can be applied in two ways, each having its own set of estimating values. The bandwidth of digital devices is measured in bits per second (bps) or bytes per second (bps). The bandwidth of analog devices is measured in cycles per second, or Hertz (Hz).

Throughput

The number of messages successfully delivered per unit time is referred to as throughput. Throughput is influenced by the available bandwidth, as well as the available signal-to-noise ratio and device limitations.

To separate the concepts of throughput and latency, throughput will be calculated from the arrival of the first bit of data reaching the receiver for this article. The terms 'throughput' and 'bandwidth' are frequently interchanged in discussions of this nature.

The Time Window refers to the time frame in which the throughput is calculated. The selection of a suitable time window will frequently determine whether or not latency affects throughput. Likewise, whether or not latency is taken into account will determine whether or not latency impacts throughput.

Latency

Latency is simply the time it takes for data to travel from one designated location to another regarding network performance evaluation. The term "delay" is sometimes used to describe this attribute. The latency of a network should be as low as possible.

Speed of light is the fundamental factor for latency, but packet queuing and refractive index of fiber optic cable are also two factors that can be used to reduce latency.

Packet Loss

Packet loss refers to the number of packets that fail to transfer from one destination to another regarding network performance measurement. This statistic can be measured by recording traffic data on both ends and then identifying lost packets and packet retransmission.

Network congestion, router performance, and software difficulties, among other things, can cause packet loss.

Jitter

The variance in time delay for data packets carried over a network is known as jitter. This variable denotes an interruption in data packet sequencing that has been identified. Jitter and latency are linked because jitter generates increased or uneven latency between data packets, which can damage network performance and cause packet loss and congestion.

While some jitter is to be expected and can typically be tolerated, quantifying network jitter is an integral part of measuring overall network performance.

Factors Affecting Network Performance

The following factors affect the performance of a network −

- Network Infrastructure
- Applications Used in the Network
- Network Issues
- Network Security

Network Infrastructure

Network hardware, such as routers, switches, and cables, networking software, security and operating systems, and network services, such as IP addressing and wireless protocols, are all part of the entire network infrastructure. Therefore, it is critical to characterize the network's overall traffic and bandwidth patterns from an infrastructure standpoint.

This network performance evaluation will reveal which flows are the most congested over time, perhaps posing an issue.

Instead of just responding to any performance crisis that may develop, identifying the over-capacity aspects of the infrastructure might lead to pre-emptive fixes or upgrades that can minimize future downtime.

Applications Used in the Network

While network hardware and infrastructure difficulties can directly impact a specific application's user experience, it's also crucial to consider the impact of applications as essential cogs in the overall network architecture. For example, poorly performing programs can eat up a lot of bandwidth and make the user experience poor.

As applications become more complicated, diagnosing and monitoring their performance becomes increasingly important. In addition, application characteristics like window sizes and keep-alive have an impact on network speed and capacity.

Network Issues

The network's intrinsic performance constraints are frequently the focus of attention. Several aspects of the network influence performance and flaws in any of these areas can lead to systemic issues. Because hardware requirements are so crucial incapacity planning, these components should be built to meet all expected system demands.

Network Security

Privacy, intellectual property, and data integrity are all protected by network security. As a result, the importance of solid cybersecurity is never in doubt. Device scanning, data encryption, virus prevention, authentication, and intrusion detection are all required for managing and mitigating network security challenges, all of which take valuable network bandwidth and can negatively influence performance.