Module Practice and Quiz

9.6.1

What did I learn in this module?

Network Transmission Quality

Voice and live video transmissions create higher expectations for quality delivery among users, and create a need for Quality of Service (QoS). Congestion occurs when multiple communication lines aggregate onto a single device such as a router, and then much of that data is placed on just a few outbound interfaces, or onto a slower interface. Congestion can also occur when large data packets prevent smaller packets from being transmitted in a timely manner. Without any QoS mechanisms in place, packets are processed in the order in which they are received. When congestion occurs, network devices such as routers and switches can drop packets. This means that time-sensitive packets, such as real-time video and voice, will be dropped with the same frequency as data that is not time-sensitive, such as email and web browsing. When the volume of traffic is greater than what can be transported across the network, devices queue (hold) the packets in memory until resources become available to transmit them. Queuing packets causes delay because new packets cannot be transmitted until previous packets have been processed. One QoS technique that can help with this problem is to classify data into multiple queues. Network congestion points are ideal candidates for QoS mechanisms to mitigate delay and latency. Two types of delays are fixed and variable. Sources of delay are code delay, packetization delay, queuing delay, serialization delay, propagation delay, and de-jitter delay. Jitter is the variation in the delay of received packets. Due to network congestion, improper queuing, or configuration errors, the delay between each packet can vary instead of remaining constant. Both delay and jitter need to be controlled and minimized to support real-time and interactive traffic.

Traffic Characteristics

Voice and video traffic are two of the main reasons for QoS. Voice traffic is smooth and benign, but it is sensitive to drops and delays. Voice can tolerate a certain amount of latency, jitter, and loss without any noticeable effects. Latency should be no more than 150 milliseconds (ms). Jitter should be no more than 30 ms, and voice packet loss should be no more than 1%. Voice traffic requires at least 30 Kbps of bandwidth. Video traffic is more demanding than voice traffic because of the size of the packets it sends across the network. Video traffic is bursty, greedy, drop sensitive, and delay sensitive. Without QoS and a significant amount of extra bandwidth, video quality typically degrades. UDP ports such as 554, are used for the Real-Time Streaming Protocol (RSTP) and should be given priority over other, less delay-sensitive, network traffic. Similar to voice, video can tolerate a certain amount of latency, jitter, and loss without any noticeable effects. Latency should be no more than 400 milliseconds (ms). Jitter should be no more than 50 ms, and video packet loss should be no more than 1%. Video traffic requires at least 384 Kbps of bandwidth. Data traffic is not as demanding as voice and video traffic. Data packets often use TCP applications which can retransmit data and, therefore, are not sensitive to drops and delays. Although data traffic is relatively insensitive to drops and delays compared to voice and video, a network administrator still needs to consider the quality of the user experience, sometimes referred to as Quality of Experience (QoE). The two main factors that a network administrator needs to ask about the flow of data traffic are if the data comes from an interactive application and if the data is mission critical.

Queuing Algorithms

The QoS policy implemented by the network administrator becomes active when congestion occurs on the link. Queuing is a congestion management tool that can buffer, prioritize, and, if required, reorder packets before being transmitted to the destination. This course focuses on the following queuing algorithms: First-In, First-Out (FIFO), Weighted Fair Queuing (WFQ), Class-Based Weighted Fair Queuing (CBWFQ), and Low Latency Queuing (LLQ). FIFO queuing buffers and forwards packets in the order of their arrival. FIFO has no concept of priority or classes of traffic and consequently, makes no decision about packet priority. When FIFO is used, important or time-sensitive traffic can be dropped when there is congestion on the router or switch interface. WFQ is an automated scheduling method that provides fair bandwidth allocation to all network traffic. WFQ applies priority, or weights, to identified traffic and classifies it into conversations or flows. WFQ classifies traffic into different flows based on packet header addressing, including such characteristics as source and destination IP addresses, MAC addresses, port numbers, protocol, and Type of Service (ToS) value. The ToS value in the IP header can be used to classify traffic. CBWFQ extends the standard WFQ functionality to provide support for user-defined traffic classes. With CBWFQ, you define traffic classes based on match criteria including protocols, access control lists (ACLs), and input interfaces. LLQ feature brings strict priority queuing (PQ) to CBWFQ. Strict PQ allows delay-sensitive packets, such as voice, to be sent before packets in other queues, reducing jitter in voice conversations.

QoS Models

There are three models for implementing QoS: Best-effort model, Integrated services (IntServ), and Differentiated services (DiffServ). The Best-effort model is the most scalable but does not guarantee delivery and does not give any packet preferential treatment. The IntServ architecture model was developed to meet the needs of real-time applications, such as remote video, multimedia conferencing, data visualization applications, and virtual reality. IntServ is a multiple-service model that can accommodate many QoS requirements. IntServ explicitly manages network resources to provide QoS to individual flows or streams, sometimes called microflows. It uses resource reservation and admission-control mechanisms as building blocks to establish and maintain QoS. The DiffServ QoS model specifies a simple and scalable mechanism for classifying and managing network traffic. The DiffServ design overcomes the limitations of both the best-effort and IntServ models. The DiffServ model can provide an “almost guaranteed” QoS, while still being cost-effective and scalable. DiffServ divides network traffic into classes based on business requirements. Each of the classes can then be assigned a different level of service. As the packets traverse a network, each of the network devices identifies the packet class and services the packet according to that class. It is possible to choose many levels of service with DiffServ.

QoS Implementation Techniques

There are three categories of QoS tools: classification and marking tools, congestion avoidance tools, and congestion management tools. Before a packet can have a QoS policy applied to it, the packet has to be classified. Classification and marking allows us to identify or “mark” types of packets. Classification determines the class of traffic to which packets or frames belong. Methods of classifying traffic flows at Layer 2 and 3 include using interfaces, ACLs, and class maps. Traffic can also be classified at Layers 4 to 7 using Network Based Application Recognition (NBAR). The Type of Service (IPv4) and Traffic Class (IPv6) carry the packet marking as assigned by the QoS classification tools. The field is then referred to by receiving devices which forward the packets based on the appropriate assigned QoS policy. These fields have 6-bits allocated for QoS. Called the Differentiated Services Code Point (DSCP) field, these six bits offer a maximum of 64 possible classes of service. The field is then referred to by receiving devices which forward the packets based on the appropriate assigned QoS policy. The 64 DSCP values are organized into three categories: Best-Effort (BE), Expedited Forwarding (EF), Assured Forwarding (AF). Because the first 3 most significant bits of the DSCP field indicate the class, these bits are also called the Class Selector (CS) bits. Traffic should be classified and marked as close to its source as technically and administratively feasible. This defines the trust boundary. Congestion management includes queuing and scheduling methods where excess traffic is buffered or queued (and sometimes dropped) while it waits to be sent out an egress interface. Congestion avoidance tools help to monitor network traffic loads in an effort to anticipate and avoid congestion at common network and internetwork bottlenecks before congestion becomes a problem. Cisco IOS QoS includes weighted random early detection (WRED) as a possible congestion avoidance solution. The WRED algorithm allows for congestion avoidance on network interfaces by providing buffer management and allowing TCP traffic to decrease, or throttle back, before buffers are exhausted. Traffic shaping and traffic policing are two mechanisms provided by Cisco IOS QoS software to prevent congestion.

9.6.2

Module Quiz - QoS Concepts

1. What is the term used to indicate a variation of delay?

○ latency
● jitter
○ serialization delay
○ speed mismatch

2. A network engineer performs a ping test and receives a value that shows the time it takes for a packet to travel from a source to a destination device and return. Which term describes the value?

○ priority
○ bandwidth
○ jitter
● latency

3. What role do network devices play in the IntServ QoS model?

○ Network devices provide a best-effort approach to forwarding traffic.
● Network devices ensure that resources are available before traffic is allowed to be sent by a host through the network.
○ Network devices use QoS on a hop-by-hop basis to provide excellent scalability.
○ Network devices are configured to service multiple classes of traffic and handle traffic as it may arrive.

4. Which device would be classified as a trusted endpoint?

○ switch
○ firewall
● IP phone
○ router

5. Under which condition does congestion occur on a converged network with voice, video, and data traffic?

○ if voice traffic latency begins to decrease across the network
● if the request for bandwidth exceeds the amount of bandwidth available
○ if video traffic requests more bandwidth than voice traffic requests
○ if a user downloads a file that exceeds the file limitation that is set on the server

6. Which type of traffic does Cisco recommend be placed in the strict priority queue (PQ) when low latency queuing (LLQ) is being used?

● voice
○ management
○ video
○ data

7. Which model is the only QoS model with no mechanism to classify packets?

○ IntServ
○ hard QoS
○ DiffServ
● best-effort

8. What happens when the memory queue of a device fills up and new network traffic is received?

● The network device will drop the arriving packets.
○ The network device queues the received traffic while sending previously received traffic.
○ The network device drops all traffic in the queue.
○ The network device sends the received traffic immediately.

9. What are two characteristics of voice traffic? (Choose two.)

□ It is bursty.
☑ It consumes few network resources.
□ It can tolerate latency up to 400 ms.
☑ It is delay sensitive.
□ It is insensitive to packet loss.

10. Which QoS model is very resource intensive and provides the highest guarantee of QoS?

○ best-effort
○ soft QoS
○ DiffServ
● IntServ

11. What happens when an edge router using IntServ QoS determines that the data pathway cannot support the level of QoS requested?

○ Data is forwarded along the pathway using IntServ but not provided preferential treatment.
○ Data is forwarded along the pathway using DiffServ.
● Data is not forwarded along the pathway.
○ Data is forwarded along the pathway using a best-effort approach.

12. In QoS models, which type of traffic is commonly provided the most preferential treatment over all other application traffic?

○ file transfers
● voice traffic
○ email
○ web traffic

13. Which queuing mechanism supports user-defined traffic classes?

○ FCFS
○ WFQ
○ FIFO
● CBWFQ

14. What mechanism compensates for jitter in an audio stream by buffering packets and then replaying them outbound in a steady stream?

○ digital signal processor
○ WFQ
○ voice codec
● playout delay buffer