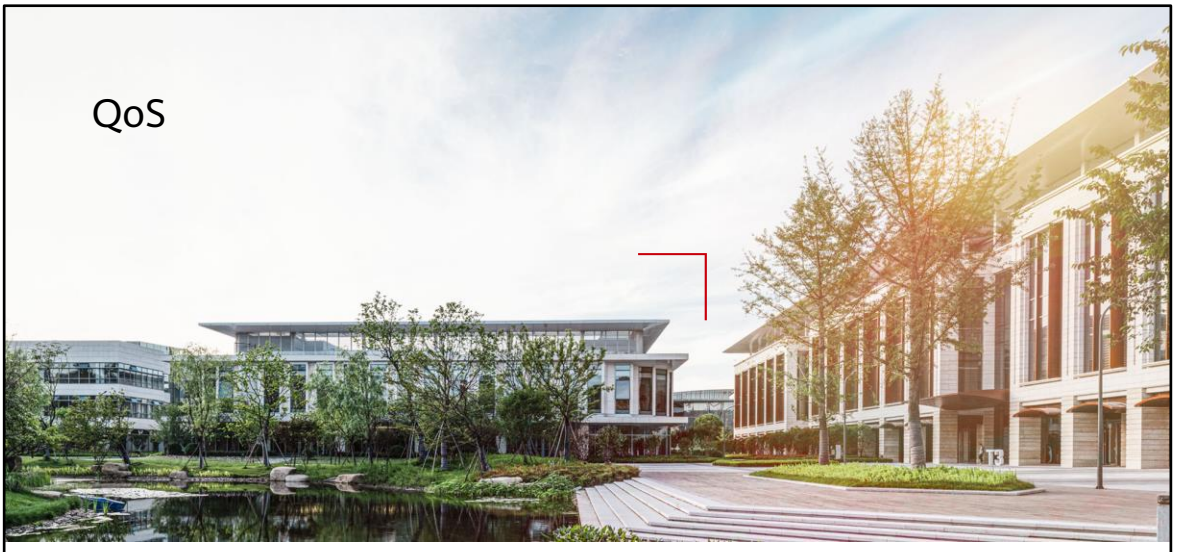
QoS



Foreword

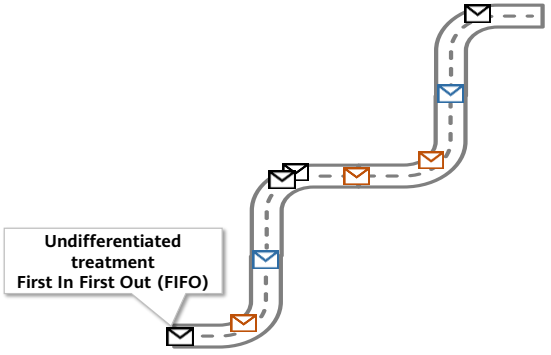
- With continuous development of networks, the network scale and traffic types increase continuously. As a result, Internet traffic increases sharply, network congestion occurs, the forwarding delay increases, and even packet loss occurs. In this case, the service quality deteriorates or even services are unavailable. To deploy real-time and non-real-time services on the IP network, network congestion must be resolved. The commonly used solution is to increase the network bandwidth. However, this solution is not ideal considering the network construction cost.
- Quality of service (QoS) is introduced in this situation. At limited bandwidth, QoS uses a "guaranteed" policy to manage network traffic and provides different priority services for different traffic.
- This course describes QoS fundamentals.

Objectives

- Upon completion of this course, you will be able to:
  - Describe the QoS background.
  - Describe QoS types.
  - Describe HQoS fundamentals.

"Best-Effort" Traditional Network

- When the IP network emerges, there is no QoS guarantee.
- You only know that the packets have been sent out. Whether the packets can be received and when the packets can be received are unknown.



- On the traditional IP network, each network device handles all packets in an undifferentiated manner and follows the First In First Out (FIFO) rule to transmit packets. The devices transmit packets to the destination in best-effort (BE) mode, but the BE mode cannot ensure the performance such as delay and reliability.

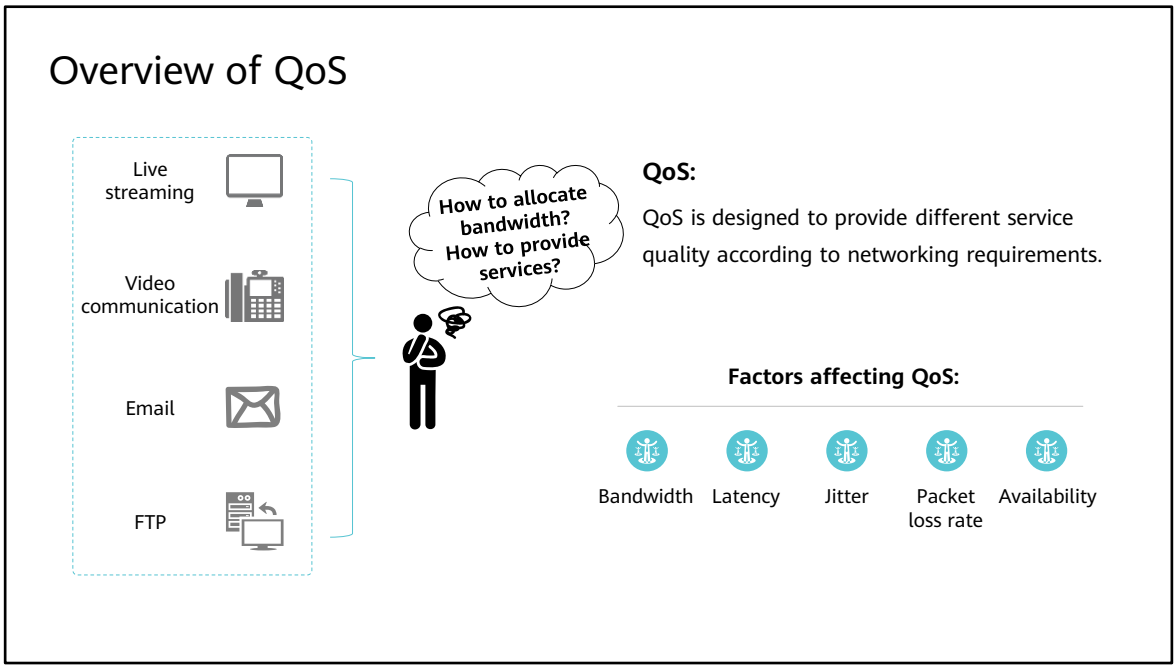
QoS Background

- With continuous technology improvement and fierce product competition, users have increasingly higher requirements on the network quality.

High-definition image quality and high network speed
Good signal quality

Poor image quality, low network speed, and frame freezing
Poor signal quality

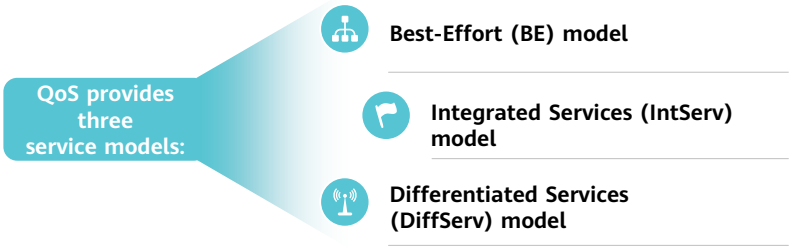
- With the emergence of new applications on IP networks, new requirements are raised to QoS of IP networks.



- To support voice, video, and data services of different requirements, the network is required to distinguish different communication types before providing corresponding QoS.
  - For example, real-time services such as Voice over IP (VoIP) demand shorter latency. A long latency for packet transmission is unacceptable. Email and the File Transfer Protocol (FTP) services are comparatively insensitive to the latency.
- To support voice, video, and data services of different requirements, the network is required to distinguish different communication types before providing corresponding QoS.
  - The BE mode of traditional IP networks cannot identify and distinguish various communication types on the networks. This distinguishing capability is the premise for providing differentiated services. The BE mode cannot satisfy application requirements, so QoS is introduced.
- What is QoS?

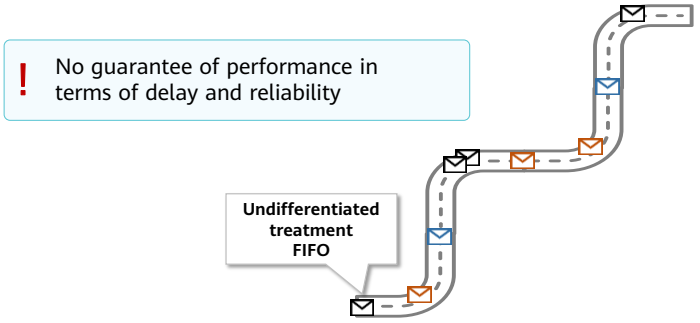
- QoS is designed to provide different service quality according to networking requirements. Example:
  - The bandwidth used by FTP on the backbone network can be limited, and database access can be given a higher priority.
  - For an ISP, its users may transmit voice, video, or other real-time services. QoS enables the ISP to differentiate these packets and provide different services.
  - QoS can provide bandwidth and low delay guarantee for time-sensitive multimedia services, and other services on the network do not affect these time-sensitive services.
- Which factors affect QoS?
  - Bandwidth: indicates the transfer speed of IP packets on a network. It can be the average value or peak value. — Bandwidth competition can be resolved by increasing the bandwidth. However, the bandwidth cannot be increased infinitely.
  - Latency: indicates the round trip time (RTT) of an IP packet between two nodes on a network. — Delay-sensitive traffic, such as video and voice traffic
  - Jitter: indicates the change in the latencies of different packets which are in the same data stream and transferred in the same direction. — It is related to the latency. If the latency is short, the jitter range is small, which has a great impact on real-time services such as voice and video services.
  - Packet loss rate: indicates the allowed maximum packet loss rate when a service is transmitted on a network. — It is used to measure the network reliability. A small number of lost packets have little impact on services, but a large number of lost packets severely affect the transmission efficiency.
  - Availability: indicates the availability of a connection between a user and the IP service, including the connection setup time and holding time.

QoS Service Models



BE Model

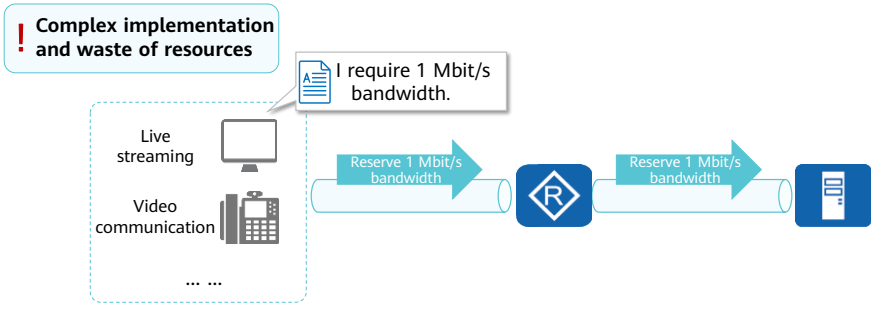
- An application can send any number of packets at any time.
- The network then makes the best effort to transmit the packets.



- The BE model is the simplest service model in which an application can send any number of packets at any time without obtaining approval or notifying the network.
- The network then makes the best effort to transmit the packets but provides no guarantee of performance in terms of delay and reliability.
- The BE model is the default service model for the Internet and applies to various network applications, such as the File Transfer Protocol (FTP) and email. It uses FIFO queues.

IntServ Model

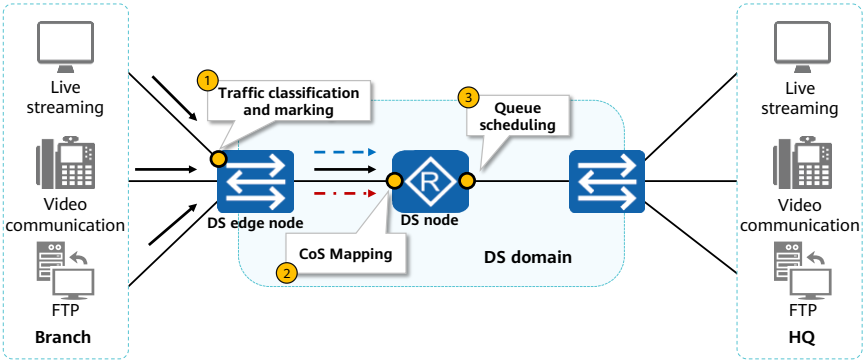
- Before sending packets, an application needs to apply for specific services through signaling.
- After receiving a resource request from an application, the network reserves resources for each information flow by exchanging RSVP signaling information.



- The IntServ model is a comprehensive service model to meet various QoS requirements.
- Before sending packets, an application needs to apply for specific services through signaling. This request is sent through RSVP. RSVP applies for network resources for an application before the application starts to send packets.
- Once the network determines to allocate resources to the application, the network maintains a state for each flow (determined by IP addresses, port numbers, and protocol numbers at both ends), and performs packet classification, traffic policing, queuing, and scheduling based on the state. After receiving the acknowledgment message from the network (the application confirms that the network has reserved resources for the packets of the application), the application starts to send packets. As long as packets of the application are controlled within the range described by traffic parameters, the network promises to meet QoS requirements of the application.
- Example: If you want to reserve a vehicle, you need to apply for a service in advance and reserve resources when resources are sufficient.
- However, the vehicle service vendor has to maintain a large number of booking information.
- Disadvantage: The implementation of the IntServ model is complex. When no traffic is transmitted, the bandwidth is still exclusively occupied, and the usage is low. This solution requires that all end-to-end nodes support and run the RSVP protocol.

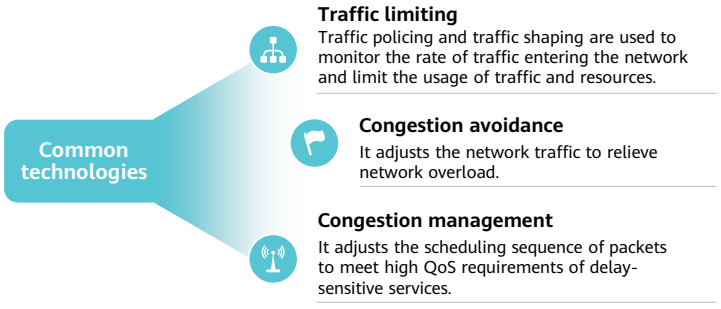
DiffServ Model

- Traffic on a network is classified into multiple classes, and a corresponding processing behavior is defined for each class, so that the traffic has different forwarding priorities, packet loss rates, and delays.



- DiffServ is a multi-service model and can satisfy different QoS requirements. Currently, this model is widely used on IP networks.
- Before sending a packet, the application does not need to notify the network to reserve resources for it. In the DiffServ model, the network does not need to maintain the status of each flow. Instead, it provides specific services based on precedence fields of packets (for example, the DS field in the IP header).
- The DiffServ model classifies network traffic into multiple classes for differentiated processing. To be specific, the DiffServ model implements traffic classification first and allocates different identifiers to different classes of packets. After a network node receives these packets, it simply identifies these identifiers and processes packets based on the actions corresponding to these identifiers.
- There is an analogy between the DiffServ model and train ticket service system. A train ticket marks the service that you book: soft sleeper, hard sleeper, hard seat, or no seat. You get on a train and enjoy the specific service marked in your ticket. On an IP network, an identifier is to a packet as a train ticket is to a passenger.
- In addition to traffic classification and marking, the DiffServ model provides the queuing mechanism. When network congestion occurs on a device, the device buffers packets in queues. The device sends the packets out of queues when network congestion is relieved.

Common QoS Technologies (DiffServ Model)



- Rate limiting: Traffic policing and traffic shaping monitor the rate of traffic entering the network to limit the traffic and resource usage, providing better services for users.
- Congestion avoidance and congestion management: When congestion occurs on a network, the device determines the sequence in which packets are forwarded according to a certain scheduling policy so key services are processed first. Or, the device proactively adjusts traffic to relieve network overload by discarding packets.
  - Traffic shaping: is a traffic control measure that initiatively adjusts the output speed of traffic. Traffic shaping enables the traffic to adapt to the network resources that can be provided by the downstream device to prevent packet loss and congestion. Traffic shaping is usually applied to the outbound direction of an interface.

Quiz

1. (Multiple-answer question) Which of the following service models are provided by QoS?
   A. DiffServ model
   B. IntServ model
   C. BE model
   D. Network service model

- 1. ABC

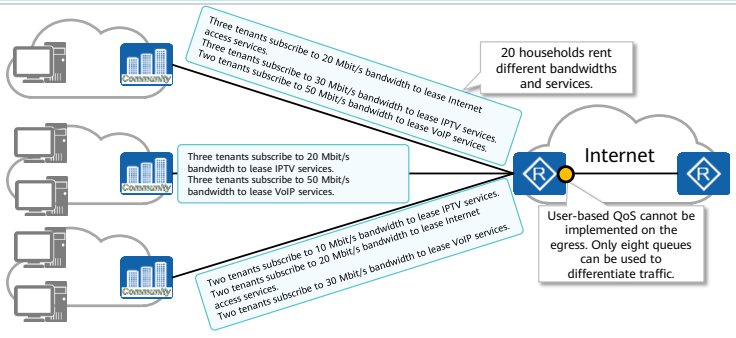
Section Summary

- QoS service models include the DiffServ, IntServ, and BE models.
- The DiffServ model is the most commonly used QoS model. It provides rate limiting, congestion avoidance, and congestion management.

Limitations of QoS

- Traditional QoS distributes a flow into only eight queues for scheduling and control. Therefore, it has great limitations in multi-tenant scenarios.

QoS limitations in home broadband scenarios



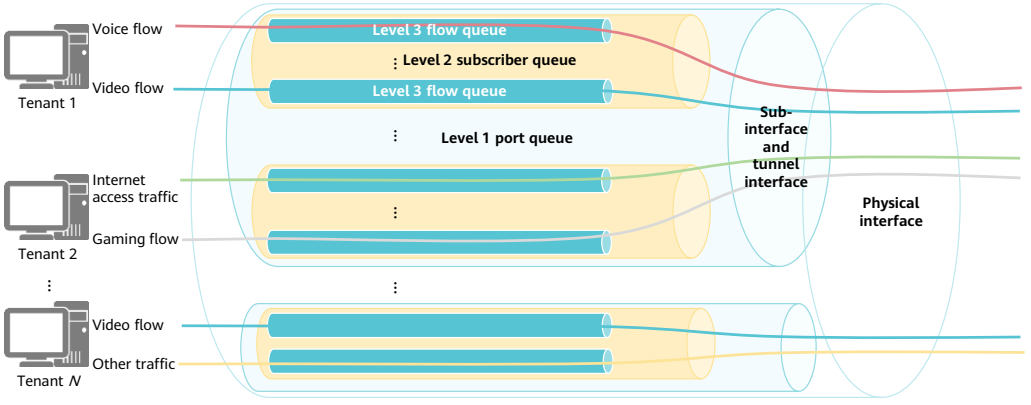
- In home broadband scenarios, different families may rent different network bandwidths and network services. Therefore, QoS cannot manage these families in a refined manner.

HQoS Overview

- Traditional QoS schedules traffic based on interfaces. An interface can only differentiate service priorities. The traffic of the same priority uses the same interface queue and competes for the same queue resources. Therefore, traditional QoS technology cannot provide differentiated services based on types of traffic and users.
- HQoS meets this requirement by implementing hierarchical scheduling based on multiple levels of queues, differentiating both services and users to provide refined QoS guarantee.
- Different devices provide different HQoS features. This section describes HQoS features supported by the CPE (AR series router).

Introduction to HQoS Queues

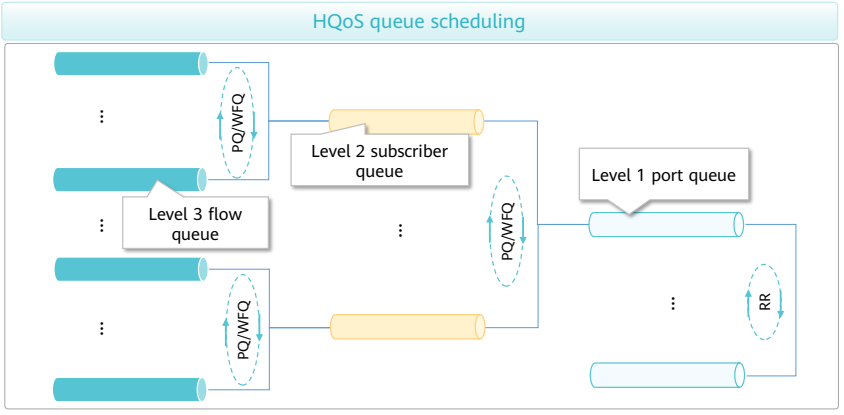
- The CPE supports three-level queues: flow queue (level 3), subscriber queue (level 2), and port queue (level 1).



- Flow queue
  - The same type of services of a user is taken as a service flow. HQoS schedules queues based on service flows. Flow queues correspond to service types and are classified into EF, AF, and BE queues. You can set scheduling modes for flow queues.
- Subscriber queue
  - Services from a user are placed into a subscriber queue. HQoS allows all services in the subscriber queue to share the bandwidth.
- Port queue
  - Each port corresponds to a queue and port queues are scheduled in RR mode. You can configure only interface-based traffic shaping, but cannot configure scheduling modes.

Introduction to HQoS Queue Scheduling

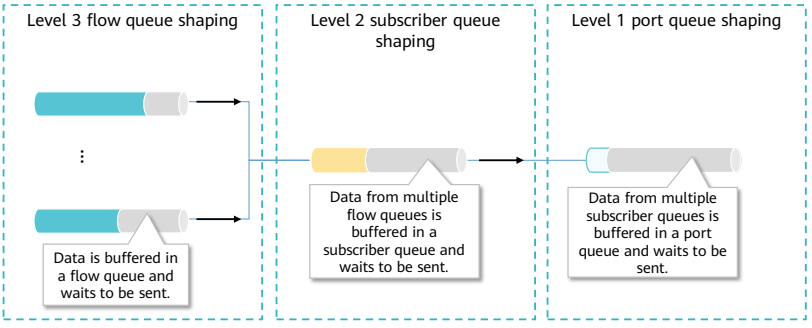
- The flow queue and subscriber queue support PQ scheduling, WFQ scheduling, and PQ+WFQ scheduling. The port queue uses RR scheduling.



- HQoS deployment for enterprise users is used as an example. Enterprise users have VoIP, video conference, and data services. Each subscriber queue corresponds to one enterprise user and each flow queue corresponds to a type of services. By deploying HQoS, the device can control the following items:
  - Traffic scheduling among three types of services of a single enterprise user
  - Total bandwidth of three types of services of a single enterprise user
  - Bandwidth allocation between multiple enterprise users
  - Total bandwidth of multiple enterprise users

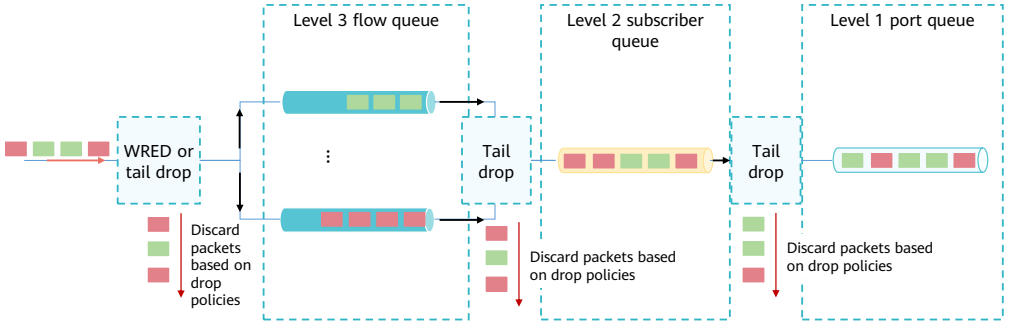
Introduction to HQoS Traffic Shaping

- The HQoS shaper buffers packets and limits the rate of packets. The device supports three levels of shapers, that is, flow queue shaper, subscriber queue shaper, and port queue shaper. After packets enter the device, the device buffers the packets in queues and sends the packets at the limited rate. Shapers can ensure the CIR and limit the maximum rate of packets by using the rate limiting algorithm.



Introduction to the HQoS Dropper

- The HQoS dropper discards packets based on a drop policy before packets are sent to queues.
- The three types of queues supported by HQoS support different drop modes. The port queue and subscriber queue support tail drop; the flow queue supports tail drop and WRED.



Quiz

1. (True or false) HQoS cannot distinguish users or services.
   A. True
   B. False
2. (Multiple-answer question) What are three types of HQoS queues?
   A. Flow queue
   B. Subscriber queue
   C. Data queue
   D. Port queue

- 1. B
- 2. ABD

Section Summary

- HQoS can ensure services with finer granularities.
- HQoS has three levels of queues: flow queue, subscriber queue, and port queue. Traffic shaping can be deployed for the three types of queues. Flow queues are scheduled in PQ+WFQ mode, subscriber queues are scheduled in PQ+WFQ mode, and interface queues are scheduled in RR mode.

Summary

- QoS is an important means to ensure service quality. Generally, the DiffServ model is used on the live network.
- This model uses rate limiting, congestion avoidance, and congestion management.
- HQoS is used in complex scenarios with finer granularity. Flow queues, subscriber queues, and port queues can be used to distinguish different users and different services of the same user.