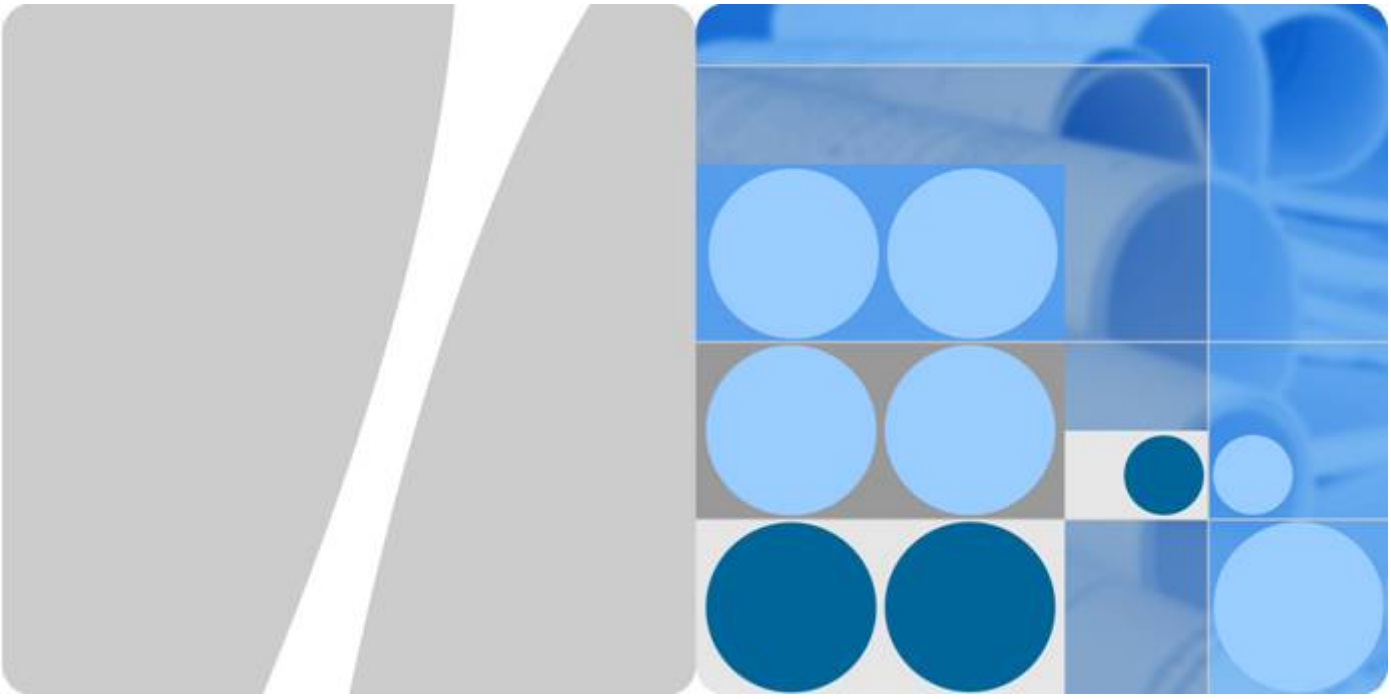


Quidway S6700 Series Ethernet Switches

V100R006C00

Feature Description - QoS

Issue 01

Date 2011-07-15

HUAWEI TECHNOLOGIES CO., LTD.

HUAWEI

Trademarks and Permissions


and other Huawei trademarks are trademarks of Huawei Technologies Co., Ltd.
All other trademarks and trade names mentioned in this document are the property of their respective holders.

Notice

The purchased products, services and features are stipulated by the contract made between Huawei and the customer. All or part of the products, services and features described in this document may not be within the purchase scope or the usage scope. Unless otherwise specified in the contract, all statements, information, and recommendations in this document are provided "AS IS" without warranties, guarantees or representations of any kind, either express or implied.

The information in this document is subject to change without notice. Every effort has been made in the preparation of this document to ensure accuracy of the contents, but all statements, information, and recommendations in this document do not constitute the warranty of any kind, express or implied.

Huawei Technologies Co., Ltd.

Address: Huawei Industrial Base
Bantian, Longgang
Shenzhen 518129
People's Republic of China

Website: http://www.huawei.com

Email: support@huawei.com

About This Document

Intended Audience

This document describes the QoS feature in terms of its overview, principle, and applications.

This document together with other types of document helps intended readers get a deep understanding of the QoS feature.

This document is intended for:

- Network planning engineers
- Commissioning engineers
- Data configuration engineers
- System maintenance engineers

Symbol Conventions

The symbols that may be found in this document are defined as follows.

Symbol	Description
DANGER	Indicates a hazard with a high level of risk, which if not avoided, will result in death or serious injury.
WARNING	Indicates a hazard with a medium or low level of risk, which if not avoided, could result in minor or moderate injury.
CAUTION	Indicates a potentially hazardous situation, which if not avoided, could result in equipment damage, data loss, performance degradation, or unexpected results.
TIP	Indicates a tip that may help you solve a problem or save time.
NOTE	Provides additional information to emphasize or supplement important points of the main text.

Command Conventions

The command conventions that may be found in this document are defined as follows.

Convention	Description
Boldface	The keywords of a command line are in **boldface**.
Italic	Command arguments are in *italics*.
[ ]	Items (keywords or arguments) in brackets [ ] are optional.
{ x | y | ... }	Optional items are grouped in braces and separated by vertical bars. One item is selected.
[x | y | ...]	Optional items are grouped in brackets and separated by vertical bars. One item is selected or no item is selected.
{ x | y | ... }*	Optional items are grouped in braces and separated by vertical bars. A minimum of one item or a maximum of all items can be selected.
[x | y | ...]*	Optional items are grouped in brackets and separated by vertical bars. Several items or no item can be selected.
&<1-n>	The parameter before the & sign can be repeated 1 to n times.
#	A line starting with the # sign is comments.

Change History

Updates between document issues are cumulative. Therefore, the latest document issue contains all updates made in previous issues.

Changes in Issue 01 (2011-07-15)

Initial commercial release.

Contents

1 QoS

About This Chapter

1.1 Introduction to QoS

Definition

QoS evaluates the ability of the service supplied to meet customers' requirements. On the Internet, QoS is used to evaluate the ability of the network to transmit packets. The network provides various services, and QoS evaluates services from different aspects. Generally, QoS is used to evaluate the ability of meeting the core requirements including the delay, jitter, and packet loss ratio during packet transmission.

- Bandwidth

  Bandwidth is also called the throughput, indicating the average rate of service flows in a certain period, in kbit/s.

- Delay

  Delay refers to the average time that a service flow spends on passing through a network. For a network device, service flows of different priorities require different delay levels. For example, there are two delay levels on a device. Through queue scheduling, services of high priorities are processed first, whereas services of low priorities can be processed only after all the services of high priorities are processed.

- Delay jitter

  Delay jitter refers to the variation in delay when service flows pass through a network.

- Packet loss ratio

  Packet loss ratio refers to the ratio of lost packets during transmission. The modern transmission system has a high reliability; therefore, packet loss is usually caused by network congestion. A common cause of packet loss is queue overflow.

Purpose

- Traditional packet transmission

  It is difficult to ensure QoS on the traditional IP network. Devices on the network process all packets equally and adopt the First In First Out (FIFO) method to transfer packets. Resources used to forward packets are allocated based on the arrival order of packets. All packets share network resources such as the bandwidth. The quantity of the resources that can be obtained depends on the arrival time of packets. This policy is called Best-Effort (BE). The device in this mode attempts to transmit packets to the destination. The BE mode, however, does not ensure the delay, jitter, packet loss ratio, or high reliability.

  The traditional BE mode applies only to the services that have no specific request for bandwidth or jitter, such as World Wide Web (WWW), file transfer, and email services.

- Requirements raised by new applications

  With the rapid development of the network, an increasing number of networks are connected to the Internet. The Internet expands greatly in size, scope, and user numbers. More and more users use the Internet as a platform for data transmission and implementation of various applications. In addition, service providers also want to develop new services. Apart from traditional applications such as WWW, email, and File Transfer Protocol (FTP), the Internet has expanded to encompass other services such as E-learning, telemedicine, videophone, video-conference, and video on demand (VoD). Enterprise users want to connect their branches in different areas through Virtual Private Network (VPN) technologies to process transaction, for example, to access corporate databases or to manage

remote devices through Telnet. The new applications have special requirements for transmission performance such as bandwidth, delay, and jitter. For example, video-conference and VoD require high bandwidth, low packet loss ratio, short delay, and low jitter. Key tasks such as transaction processing and Telnet stress on short delay and priority processing in case of congestion although they do not necessarily require high bandwidth.

New services put forward higher requirements for the service capability of the IP network. Users are no longer satisfied with transmission of packets to the destination. They need better services, for example, provision of dedicated-line bandwidth, decrease in the packet loss ratio, management and avoidance of network congestion, and control of network traffic.

These requirements demand better service capabilities from the network.

1.2 References

The following table lists the references of QoS.

Document	Description	Remarks
RFC 2474	Differentiated Services Field	-
RFC 2475	Architecture for Differentiated Services	-
RFC 2597	A Single Rate Three Color Marker	-
RFC 2598	An Expedited Forwarding PHB	-

1.3 Principles

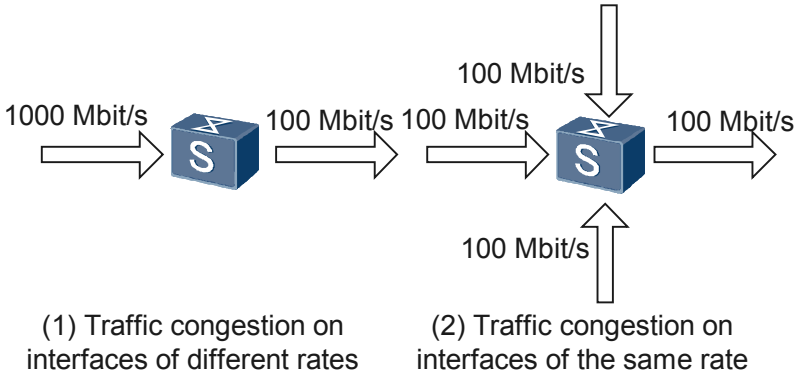
1.3.1 Cause and Impact of Congestion and Solutions to Congestion

On the traditional network, QoS is mainly affected by network congestion. Congestion means the low service rate resulting from the insufficiency of network resources.

Causes of Congestion

Congestion occurs frequently during packet transmission on the Internet.

Figure 1-1 Occurrence of congestion



The two conditions are described as follows:

- Packets enter the Switch at a high speed and are forwarded at a low speed.
- Packets reach the Switch from multiple interfaces of the same rate and are forwarded by a single interface of the same rate. If traffic reaches the Switch at wire speed, congestion will occur because of insufficiency of resources.

Apart from the insufficiency of link bandwidth, scarcity of any resource that is used to forward packets, such as the processor time, buffer, and memory, will cause congestion. In addition, network congestion may occur if the Switch fails to control the traffic in a certain period. As a result, the network resource is unavailable to process the excessive traffic.

Impact of Congestion

Congestion has the following impacts:

- Increases the delay and delay jitter of packet transmission. A long delay causes retransmission of packets.
- Lowers the throughput of the network and damages network resources.
- Occupies a large number of network resources, especially the storage resource. Improper resource allocation even causes deadlock of the system.

Congestion prevents traffic from obtaining resources immediately, which degrades the QoS. However, congestion often occurs in a complicated networking environment where packet transmission and provision of various services are both required.

Solutions to Congestion

Increasing the network bandwidth is a solution to solving the shortage of resources; however, the increasing bandwidth is used by a large amount of traffic. This method cannot solve all congestion problems. Enhancing the functions of traffic control and resource allocation to provide differentiated service for different flows and to allocate and use resources properly. This solution is more effective.

During resource allocation and traffic control, the system controls the direct and indirect factors that may cause network congestion to reduce the possibility of causing congestion. In addition, when congestion occurs, the system allocates resources according to the characteristics and requirements of services to minimize the impact of congestion on QoS.

1.3.2 Service Models

A service model is a combination of methods of guaranteeing end-to-end QoS. There are three service models: Best Effort model, integrated service (IntServ) model, and DiffServ model.

Best Effort Model

The Best Effort model is a unitary and simple service model. An application can send any number of packets at any time without any approval or notifying the network. In the Best Effort model, the network attempts to send packets, but cannot ensure the performance such as delay and reliability.

The Best Effort model is the default service model of the Internet and can be applied to various network applications, such as FTP and email. The Best Effort model can be implemented through the FIFO queue.

IntServ Model

In the IntServ model, a device must submit a request to the network before sending packets. The request is sent through a signaling protocol such as the Resource Reservation Protocol (RSVP). An application notifies the network of the QoS requirements, including the delay, bandwidth, and packet loss ratio, through the RSVP signaling. After receiving the RSVP request, the network nodes on the transmission path perform admission control to check the validity of the user and the availability of the resources. Then, the network nodes determine whether to reserve resources for the application.

When the resource is allocated to the application, the network ensures the QoS for the application as long as the packets of this application are controlled within the range specified by the traffic parameters. The network nodes on the reserved path can perform packet classification, traffic policing, and queue scheduling with a short delay to ensure the QoS. Combined with multicast, the IntServ model can be used in the real-time multimedia applications that require high bandwidth and low delay, such as video-conference and video on demand (VoD).

Currently, the RSVP-based IntServ model defines the following service types:

- Guaranteed service: provides the preset bandwidth and delay to satisfy the requirements of applications. For example, bandwidth of 10 Mbit/s is reserved and a delay of less than one second is guaranteed for the voice over IP (VoIP) service.
- Controlled-load service: provides QoS similar to the QoS provided by the BE model when network load is normal. Even if a network is overloaded (namely, during the network congestion), it can guarantee low delay and low packet loss ratio for certain applications.

The greatest advantage of the IntServ model is to provide end-to-end QoS guarantee, whereas its disadvantage is poor scalability. Network nodes must maintain soft state information for resource reservation. In the multicast application, network nodes need to periodically send resource requests and path update information to the network so that multicast members can join or leave a multicast group dynamically.

The preceding operations take the network nodes much time and occupy great memory space. When a network is expanded, the cost of maintenance increases greatly, which seriously degrades the performance of network nodes especially the core node that processes packets at wire speed. Therefore, the IntServ model is not applicable to backbone networks where the traffic aggregates.

DiffServ Model

To ensure that applications are provided with differentiated quality of service, the Internet Engineering Task Force (IEIF) defines the DiffServ model.

DiffServ is a multi-service model and can satisfy different QoS requirements. Unlike IntServ, an application in the DiffServ model does not need to send requests to the network devices on the transmission path for resource reservation. Instead, the application sets the precedence field in the packet header to notifies the network devices of its QoS requirements.

In the DiffServ model, network devices provide service for each packet according to the precedence field in the packet instead of maintaining the status of each flow. QoS can be specified based on various information, such as the IP precedence, source address, and destination address

of packets. Network devices perform traffic classification, traffic shaping, traffic policing, and queue scheduling according to these parameters.

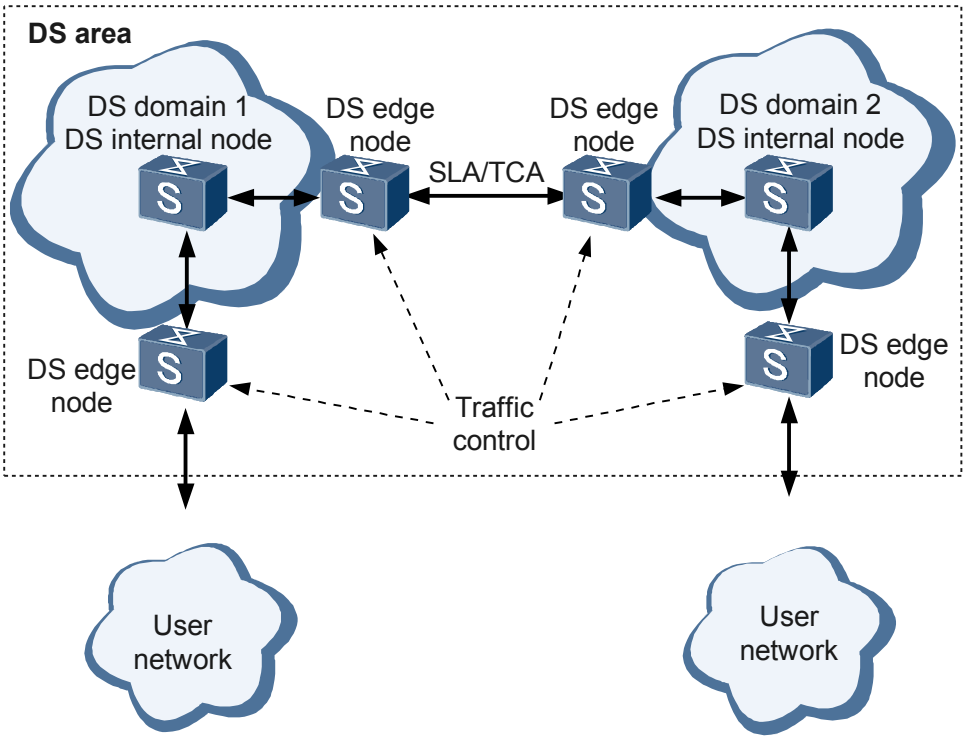
The DiffServ model provides the end-to-end QoS for important applications. In the DiffServ model, the edge device classifies packets according to information such as the source and destination addresses, and then sets the precedence of each type of packets in the packet headers. Other devices only need to schedule packets according to the precedence.

1.3.3 Implementation of the DiffServ Model

Architecture of the DiffServ Model

The architecture of the DiffServ model defines the system model and functional components that implement differentiated service. On a network node, differentiated service is implemented through functions such as per-hop behavior (PHB), traffic classification, traffic policing, traffic shaping, congestion avoidance, and congestion management. Differentiated service is provided in a DiffServ domain. The DiffServ domain consists of edge nodes and internal nodes. An edge node classifies service flows reaching the network, adjusts the traffic, marks the precedence of packets, and forwards the packets according to a PHB in the PHB group supported in the DiffServ domain. The internal nodes perform certain forwarding behavior according to the mapping between the PHB and the DSCP or 802.1p precedence of the packets and allocate bandwidth for the service flows.

Figure 1-2 Architecture of the DiffServ model



The architecture of the DiffServ model is shown in the figure.

- DS node

A DS domain is a network node that implements the DiffServ function. DS nodes can be classified into DS edge nodes and DS internal nodes.

- DS edge node

  A DS edge node connects a node in the other DiffServ domain or in a non-DiffServ domain. The DS edge node classifies the service flows entering the DiffServ domain, adjusts the traffic, marks the service flows properly, and forwards the flows according to a PHB in the PHB group supported in the DiffServ domain.

  A DS edge node can be the ingress node for the service flows in one direction and the egress node for the service flows in the other direction. Service flows enter the DiffServ domain from the ingress node and leave the DiffServ domain from the egress node. The ingress node ensures that the service flows entering the DiffServ domain conform to the Service Level Agreement (SLA) or Traffic Conditioning Agreement (TCA) between the local DiffServ domain and the DiffServ domain directly connected to the ingress node. The egress node adjusts the service flows that are forwarded to the directly connected DiffServ domain according to details of the TCA between the two domains.

- DS internal node

  A DS internal node connects to other DS internal nodes or DS edge nodes in the same DiffServ domain. The DS internal node in a DiffServ domain selects forwarding behaviors according to the DS field in the IP packet header or the PHB field defined in the 802.1p priority of packets in a VLAN. Both the DS edge nodes and DS internal nodes must forward packets according to the DSCP or 802.1p field of the packets.

- DS domain

  A DiffServ domain is composed of a set of interconnected DS nodes that use the same service policy and implement the same PHB. A DiffServ domain consists of DS edge nodes and DS internal nodes. The edge nodes form the boundary of the DiffServ domain, and the internal nodes form the core of the DiffServ domain.

- SLA

  An SLA is an agreement between a customer (individual user, enterprise, or neighbor ISP) and an Internet service provider (ISP) about the treatment of service flows on the network. The SLA stipulates a great number of terms such as the payment. The technical requirement in the SLA is called the Service Level Specification (SLS).

- TCA

  A TCA is an agreement between a customer and an ISP about service classification, service model, and service processing. A TCA without commercial clauses is called the Traffic Conditioning Specification (TCS). An SLA can contain a TCA. The SLA or SLS specifies the general clauses for service processing, such as the service processing mechanism. The TCA or TCS specifies more specific clauses, such as the requirement on bandwidth.

- DS area

  One or more adjacent DiffServ domains form a DS area. A DS area supports services crossing multiple DiffServ domains. DiffServ domains in a DS area may support different PHB groups and mappings from QoS fields to the PHBs. DiffServ domains can use different PHBs to implement different service policies. These DiffServ domains provide inter-domain services through the SLA and TCA. The SLA and TCA defines the rule for shaping the service flows from one DiffServ domain to another.

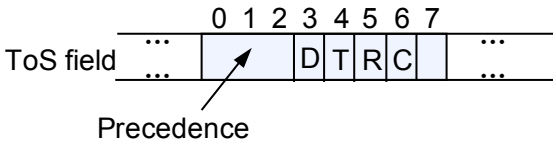
Field Containing QoS Information

Certain fields in the packet header or frame header are used to record QoS information to provide differentiated QoS for various services on the Internet. These fields include:

- Precedence field in an IP packet header

  Defined in RFC 791, the Precedence field in the Type of Service (ToS) field of an IP packet header identifies the IP precedence of the packet. **Figure 1-3** shows the Precedence field in an IP packet.

  Figure 1-3 Precedence field in an IP packet header



  Bits 0 to 2 indicate the precedence field. They encode eight transmission priorities, which rank in descending order of 7, 6, 5, 4, 3, 2, 1 and 0. The highest priority 7 or 6 is reserved for selecting routes or updating network control communications. User-level applications can use only the precedence levels from 0 to 5.

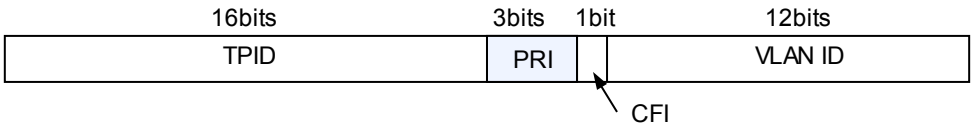
  Apart from the precedence field, a ToS field also contains the following sub-fields:

  - Bit D indicates the delay. The value 0 represents a normal delay and the value 1 represents a short delay.
  - Bit T indicates the throughput. The value 0 represents normal throughput and the value 1 represents high throughput.
  - Bit R indicates the reliability. The value 0 represents normal reliability and the value 1 represents high reliability.
  - Bits 6 and 7 in the ToS field are reserved.

- 802.1p priority of a VLAN frame header

  Layer 2 switches exchange VLAN frames. Based on the IEEE 802.1Q definition, the PRI field, which specifies the 802.1p priority, in a VLAN frame header identifies the requirement for QoS. **Figure 1-4** shows the PRI field in a VLAN frame.

  Figure 1-4 802.1p priority in a VLAN frame



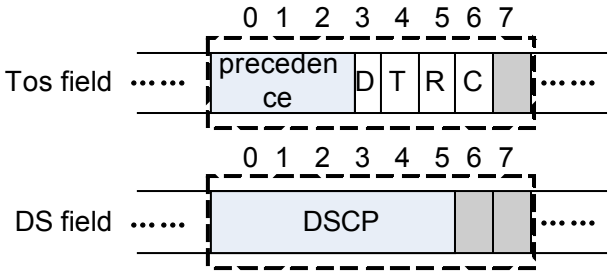
  The 802.1Q header contains a 3-bit PRI field. This PRI field specifies eight transmission priorities, which rank in descending order of 7, 6, 5, 4, 3, 2, 1 and 0. In addition, the header also contains the Tag Protocol Identifier (TPID) field, the Canonical Format Indicator (CFI) field, and the VLAN ID field.

- DSCP field

  In the DiffServ scheme, services are classified and traffic is controlled according to service requirements on the ingress of the network. The Differentiated Service Code Point (DSCP) field is also set. On the network, various types of communication are differentiated according to the DSCP field in packets, and services are provided, including resource allocation and packet dropping policy.

RFC 1349 re-defines the ToS field of an IP packet. It adds bit C, which encodes the monetary cost. After that, the IETF DiffServ working group re-defines bits 0 to 5 of a ToS field in RFC 2474 and uses them to encode the DSCP. In RFC 2474, the name of the field is changed from ToS to DS, that is, Differentiated Service. **Figure 1-5** shows the format of the DS byte.

Figure 1-5 Format of the DSCP field



In the DS byte, bits 0 to 5 are used to distinguish DSCP; bits 6 and 7 are reserved. The three leftmost bits of the DS field, that is, bits 0 to 2, are Class Selector Code Points (CSCPs). The CSCP defines the type of the DSCP. On the network, the devices that support DiffServ select forwarding behaviors according to the DSCP field.

Diff-Serv Model

- Introduction to the DiffServ model

  In order that differentiated service is provided for different services on the Internet, the IETF defines the DiffServ model.

  In the applications where the DiffServ model is used, the device notifies other devices of its QoS requirements by setting the preference field in the ToS field of the IP packet header. Then, the device sends the packet out. Devices along the transmission path obtain the service type of packets by analyzing the IP packet header. When carrying out the DiffServ model, access devices need to classify the packets and mark the service types at the packet headers. Downstream devices forward packets according to the service types in the packets. The DiffServ model, therefore, is a QoS scheme based on packet flows.

- Standard PHBs

  The IETF DiffServ working group defines the forwarding behaviors of network nodes as PHBs, such as, traffic scheduling and policing. Devices on the network select PHBs according to the DSCP field.

  Currently, the IETF defines four standard PHBs: class selector (CS), expedited forwarding (EF), assured forwarding (AF), and best-effort (BE). BE is defined as the default PHB.

  - CS

    CS stands for the class selector, representing the highest CoS. It is the same as the Precedence field. The value of the DSCP field is XXX000, where X is 0 or 1.

  - EF

    EF represents the highest QoS level on a DiffServ network. The EF PHB is applied to services that require low packet loss ratio, short delay, and high bandwidth. In any case, EF traffic can achieve a rate equal to or greater than the set rate. The value of the DSCP field is 101110.

- AF

  AF is applied to key data services that require assured bandwidth and short delay. For traffic within a reasonable limit, AF assures its forwarding quality. For traffic beyond the limit, AF degrades the service class before forwarding the traffic rather than discarding it.

  RFC 2597 defines four AFs, indicated by "AFI" (1 <= i <= 4), such as AF1, AF2, AF3, and AF4. Each type of AF is classified into three levels of drop precedence indicated by "AFij" (1 <= j <=3). The greater the value of j, the higher the drop precedence. **Table 1-1** shows the values of the DSCP field, corresponding to various types of AF services.

  Table 1-1 Values of the DSCP field corresponding to various types of AF services

Drop Precedence	AF1	AF2	AF3	AF4
Low	AF11 001010	AF21 010010	AF31 011010	AF41 100010
Middle	AF12 001100	AF12 001100	AF32 011100	AF42 100100
High	AF13 001110	AF23 010110	AF33 011110	AF43 100110

- BE

  BE is applied to best-effort services that require no strict assurance of QoS. The BE PHB focuses on only reachability, regardless of other aspects. For example, traditional IP packet transmission is in BE mode. In this case, the value of the DSCP field is 000000.

Functional Components of the DiffServ Model

The primary technologies for implementing Diffserv include traffic classification, traffic policing, traffic shaping, congestion management, and congestion avoidance. These technologies are described as follows:

- Traffic classification: Identifies data flows according to a certain matching rule. Traffic classification is the prerequisite for differentiated service.
- Traffic policing: Polices the traffic of certain specification entering the switch. When the traffic exceeds the specification, restrictive or punitive measures are taken to protect the commercial profit and network resources of the carrier.
- Traffic shaping: Adjusts the output speed of the traffic. This is a traffic control measure. Traffic shaping is used to adapt the traffic to the network resources that are provided by the downstream switch to prevent packet loss and congestion.
- Congestion management: Solves address resource competition when network congestion occurs. Usually, the packets are buffered in a queue and a scheduling algorithm is adopted to arrange the packet forwarding sequence.

Traffic classification is the basis of other functional components. It identifies packets according to certain matching rules and the other functional components control network traffic and resource allocation to provide differentiated service.

1.3.4 Class-based QoS

In the class-based QoS, traffic is classified according to certain rules. Then, a behavior is associated with the traffic of the same type to form a policy. After the policy is used, the S6700 performs traffic policing, priority re-marking, and redirection.

Traffic Classification

Traffic classification is used to identify the packets with certain features according to a rule, and is the prerequisite and basis for providing differentiated services. Traffic classification consists of complex traffic classification and simple traffic classification.

Simple Traffic Classification

The simple traffic classification mechanism sorts data packets into multiple classes with different priorities based on the ToS field.

- Rules of simple traffic classification

  The S6700 classifies traffic according to the following information:

  - DSCP priority of IP packets
  - 802.1p priority of packets in a VLAN

- Roadmap of deploying simple traffic classification

  The nodes in a DiffServ domain select forwarding behaviors according to the DS field in the IP packet header or the PHB field defined in the 802.1p priority of packets in a VLAN; therefore, the nodes in a DiffServ domain are required to support simple traffic classification.

  Simple traffic classification can be configured on the inbound interface or the outbound interface. If simple traffic classification is configured on the inbound interface, it is called upstream simple traffic classification. If simple traffic classification is configured on the outbound interface, it is called downstream simple traffic classification.

  - Upstream simple traffic classification is configured for incoming packets on an interface. It maps priorities of packets to PHBs and colors according to the mappings defined in the DiffServ domain.

    Upstream simple traffic classification is used to differentiate services including voice, video, and data services. During congestion management and queue scheduling, the packets of different services enter different queues for differentiated scheduling. For example, voice packets can enter the queue of a high priority so that a short delay is ensured.

    When upstream simple traffic classification is not configured, tagged packets enter queues based on their 802.1p priorities and untagged packets enter queues based on the default 802.1p priority of the interface

  - Downstream simple traffic classification is configured for outgoing packets on an interface. It maps PHBs and colors to priorities of packets according to the mappings defined in the DiffServ domain.

    When downstream simple traffic classification is configured, priorities of packets can be re-marked. The downstream device then ensures differentiated QoS for packets based on their re-marked priorities.

    When downstream simple traffic classification is not configured, the S6700 re-marks priorities of packets based on their default mappings.

Complex Traffic Classification

The complex traffic classification mechanism classifies packets according to Layer 2 and Layer 3 information carried by packets or by matching the following information with Access Control Lists (ACLs), and then providing the classified packets with specified QoS. The information includes:

- IP quintuple, which consists of the source IP address, the destination IP address, the source port number, the destination port number, and the packet type
- TCP SYN information
- Rules of complex traffic classification

  The S6700 can perform complex traffic classification according to the following Layer 2 information:

  - VLAN ID carried in the outer tag of packets in a VLAN
  - 802.1p priority in the outer tag of packets in a VLAN
  - Double tags of packets in a VLAN
  - Source MAC address
  - Destination MAC address
  - Incoming interface
  - Protocol type field encapsulated in Layer 2 packets

  The S6700 can also perform complex traffic classification according to the following Layer 3 information:

  - IP protocol type, that is, IPv4 or IPv6

  In addition, the S6700 can also perform complex traffic classification by matching the following information with ACLs:

  - IP precedence or DSCP priority
  - Prefix of the source IP address
  - Prefix of the destination IP address
  - Protocol number carried in an IP packet
  - Fragment flag
  - TCP SYN flag
  - TCP or UDP source port number or port number range
  - TCP or UDP destination port number or number range
- Roadmap of deploying complex traffic classification

  The nodes at the border of the DiffServ domain perform access control functions including traffic policing, defense against theft and Denial of Service (DoS) attacks, and implementation of traffic filtering; therefore, the nodes at the border of the DiffServ domain are required to support complex traffic classification to identify types of traffic.

Priority Mapping

The priority mapping mechanism is used to map precedence fields to CoS of the device. The CoS of the device is called the internal priority, including PHBs and colors.

Packets carry different precedence fields on various networks. For example, packets carry the 802.1p field in a VLAN, the DSCP field on an IP network. To provide differentiated QoS for

different packets, when the packets enter the S6700, the S6700 needs to map precedence fields of the packets to PHBs and colors; when the packets leave the S6700, the S6700 needs to map PHBs and colors to precedence fields of the packets. In this manner, downstream devices can provide corresponding quality of service according to precedence fields of the packets.

The priority mapping mechanism provides the mapping from precedence fields of packets to CoS or the mapping from CoS to precedence fields of packets based on simple traffic classification. In addition, the mechanism uses the DiffServ domain to manage and record the mapping of precedence fields of packets and CoS.

The S6700 providesthe following priority mapping modes:

- DSCP priority of IP packets

  When service traffic enters the S6700, the S6700 classifies packets according to DSCP priorities and searches the mapping table of DSCP priorities and CoS. Then, the S6700 maps CoS to the DSCP priorities of packets and provides differentiated QoS.

  When service traffic leaves the S6700, the S6700 searches the mapping table of CoS and DSCP priorities according to CoS. Then, the S6700 maps CoS to DSCP priorities of packets so that the device connected to the S6700 can provide differentiated QoS according to the DSCP priorities of packets.

- 802.1p priority of packets in a VLAN

  When service traffic enters the S6700, the S6700 classifies packets according to 802.1p priorities and searches the mapping table of 802.1p priorities and CoS. (The S6700 classifies untagged packets according to the default 802.1p priority of the interface.) Then, the S6700 maps CoS to the 802.1p priorities of packets and provides differentiated QoS.

  When service traffic leaves the S6700, the S6700 searches the mapping table of CoS and 802.1p priorities according to CoS. Then, the S6700 maps CoS to 802.1p priorities of packets so that the device connected to the S6700 can provide differentiated QoS according to the 802.1p priorities of packets.

There is the **default** DiffServ domain. Users can use the **default** DiffServ domain or re-define the DiffServ domain. On the S6700, a maximum of seven DiffServ domains can be created. You need to bind a DiffServ domain to an interface so that the system can perform priority mapping for incoming and outgoing traffic on the interface.

Traffic Behavior

Traffic behaviors refer to the actions taken on packets. Performing complex traffic classification is to provide differentiated QoS. Complex traffic classification takes effect only when it is associated with a traffic control action or a resource allocation action.

The S6700 provides the following traffic behaviors based on complex traffic classification:

- Permit/Deny
- Re-marking
- Traffic policing
- Flow mirroring
- Traffic statistics

The traffic behaviors except for **deny** can be used together.

- Deny/Permit

Deny/Permit is the simplest traffic control action. The S6700 controls network traffic by forwarding or discarding packets.

- Re-marking

  This traffic control action is used to set the precedence field in a packet. Packets carry different precedence fields on various networks. For example, packets carry the 802.1p field in a VLAN, the ToS field on an IP network. Therefore, the S6700 is required to mark the precedence fields of packets according to the network type.

  Generally, a device at the border of a network needs to re-mark the precedence fields of incoming packets; the device in the core of a network provides corresponding QoS services according to the precedence fields marked by the border device, or re-marks the precedence fields according to its own standard.

- Traffic policing

  This traffic control action is used to limit the traffic and the resource used by the traffic by monitoring the specifications of the traffic. Through traffic policing, the S6700 can discard, re-mark the color or precedence of, or perform other QoS measures over the packets that exceed the specifications.

- Flow mirroring

  This traffic control action is used to copy the specified data packets to a specified destination to detect and troubleshoot faults on a network.

- Traffic statistics

  The traffic statistics action is used to collect data packets of specified service flows, that is, passed and discarded packets and bytes of packets matching traffic classification rules.

  The actions of traffic statistics are not QoS measures but can be used together with other actions to improve the security of the network and packets.

Traffic Policy

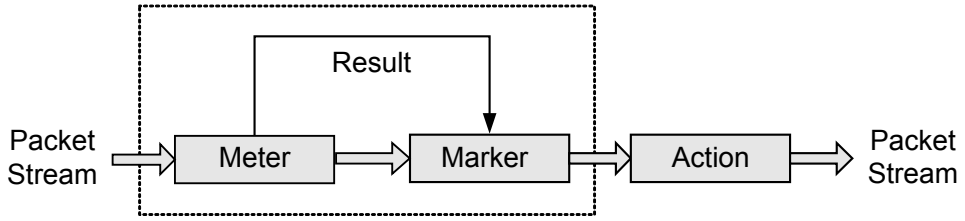
A traffic policy is a QoS policy in which traffic classifiers are bound to traffic behaviors. The traffic policy can be used on an interface, globally, on an LPU, or in a VLAN so that traffic classifiers bound to traffic behaviors in the traffic policy are used on the interface, globally, or in the VLAN at the same time.

1.3.5 Traffic Policing

Traffic policing is a traffic control mechanism that monitors the rate of incoming packets and controls the traffic that exceeds the set rate so that the traffic rate is limited within a proper range, and that both the network resources and carriers' interests are protected. On the S6700, traffic policing can be implemented by configuring and applying the QoS CAR template or the traffic policy.

Principle of Traffic Policing

Figure 1-6 Traffic policing components

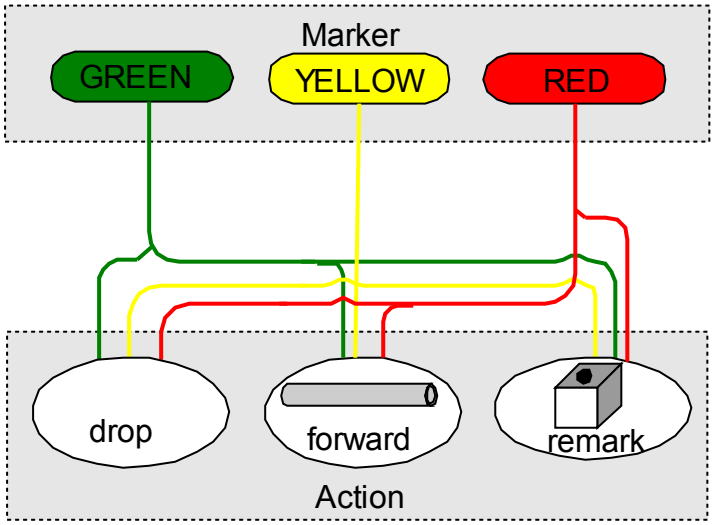


As shown in **Figure 1-6**, traffic policing on the S6700 is composed of the following items:

- Meter: Measures network traffic through the token bucket and reports the measuring result to Marker.
- Marker: Colors packets according to the measuring result. Packets are colored in green, yellow, or red.
- Action: Performs actions for packets according to the coloring, as follows:
  - forward: The packets with the measuring result of "conforming" are forwarded.
  - remark: The packets are forwarded after the internal priority of packets is changed.
  - drop: The packets with the measuring result of "not conforming" are discarded.

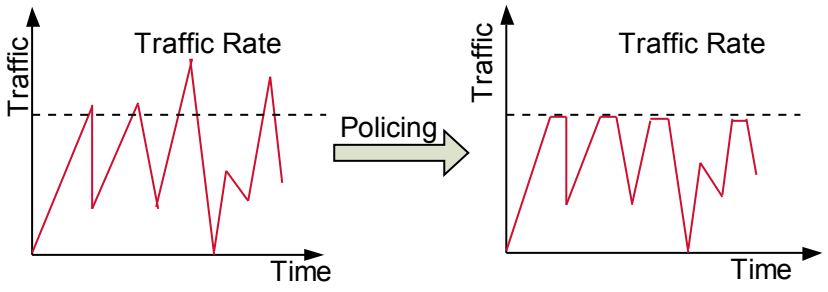
  By default, green and yellow packets are forwarded, whereas red packets are discarded.

Figure 1-7 Traffic policing actions



Through traffic policing, the S6700 reduces the priority of packets before forwarding or discarding them when the rate of the traffic exceeds the standard. By default, packets are discarded. **Figure 1-8** shows the curve of traffic in a specified rate range when traffic policing is used. The traffic whose rate exceeds the specified rate is deleted.

Figure 1-8 Curve of traffic on which traffic policing is implemented



1.3.6 Traffic Shaping

Overview

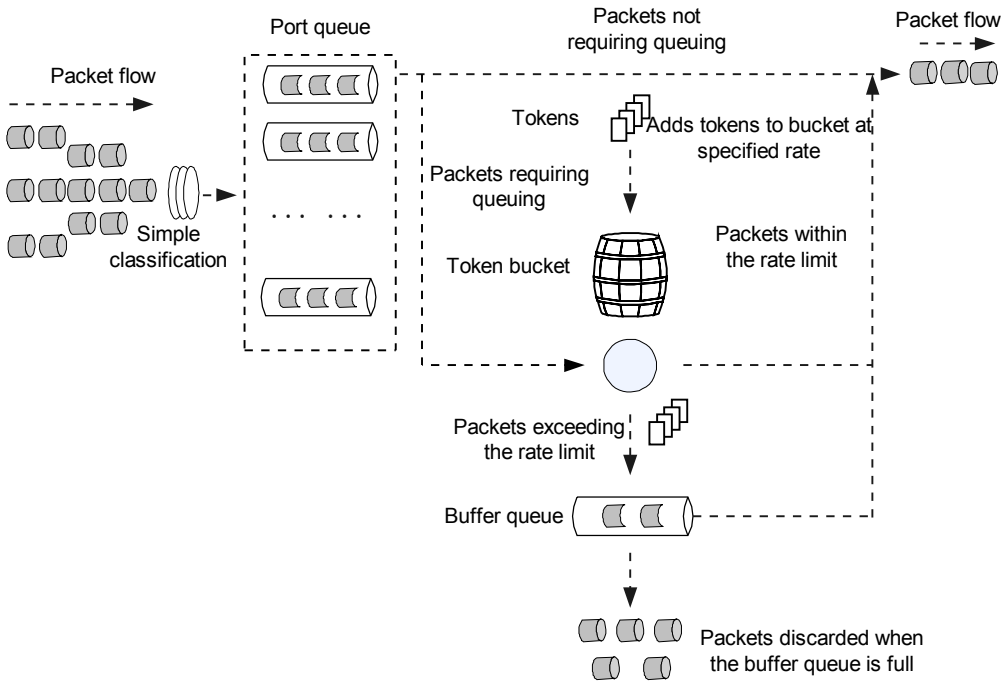
When the rate of the interface on the downstream device is smaller than the rate of the interface on the upstream device or the burst traffic occurs, traffic congestion may occur on the interface of the downstream device. In this case, you can configure traffic shaping on the interface of the upstream device at the outbound direction. The traffic shaping technology can make the uneven incoming traffic become even, and then a even traffic is displayed. The congestion problem is thus solved.

Processing Procedure

The traffic shaping technology is used on an interface or in an interface queue, and can limit the rate of all the packets on an interface or the packets of a certain type (based on simple traffic classification) passing through an interface.

The traffic shaping technology also uses the token bucket to control traffic. **Figure 1-9** shows the procedure of traffic shaping in an interface queue.

Figure 1-9 Procedure of traffic shaping



The details of the procedure are as follows:

1. When packets arrive, the Switch classifies packets so that the packets enter different interface queues.

2. If the interface queue that packets enter is not configured with traffic shaping, the packets of the interface queue are sent. Otherwise, proceed to the next step.
3. The system places tokens to the bucket at the rate (CIR) set by a user.
   - If there are sufficient tokens in the bucket, the Switch sends packets directly and the number of tokens decreases.
   - If there are insufficient tokens in the bucket, the Switch places packets into the buffer queue. If the buffer queue is full, packets are discarded.
4. When there are packets in the buffer queue, the system extracts the packets from the queue and sends them periodically. Each time the system sends a packet, it compares the number of packets with the number of tokens till the tokens are insufficient to send packets or all the packets are sent.

After traffic shaping is configured in an interface queue, the system needs to control the packets at the rate of traffic shaping set on an interface if traffic shaping is configured on the interface. The procedure is the same as the procedure of traffic shaping in an interface queue; however, you do not need to perform **1** and **2**.

Differences Between Traffic Shaping and Traffic Policing

The differences between traffic shaping and traffic policing are as follows:

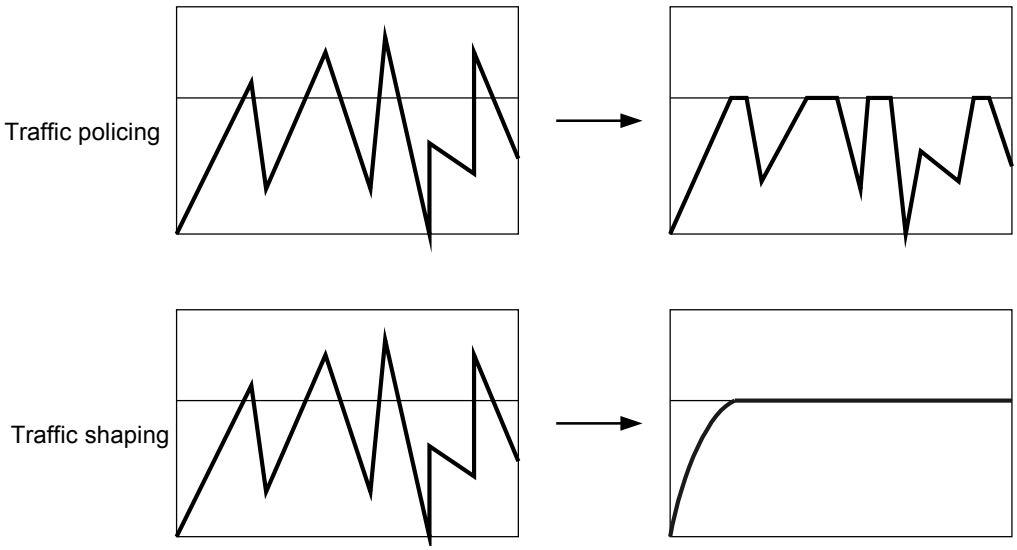
- Traffic policing is used to control traffic by discarding the packets whose rate does not meet the requirements. Traffic shaping, however, is used to buffer the packets whose rate does not meet the requirements. When there are sufficient tokens in the token bucket, the buffered packets are then forwarded at an even rate.
- Traffic shaping may increase the delay, but traffic policing does not lead to the extra delay.

Table 1-2 Comparison between traffic shaping and traffic policing

Type	Advantage	Disadvantage
Traffic shaping	Has less opportunities to discard packets.	Introduces the delay and jitter. More buffer resources are required to buffer packets.
Traffic policing	Supports the re-marking action. No extra buffer is required.	Discards more packets, which may cause retransmission.

Figure 1-10 describes the differences between traffic shaping and traffic policing.

Figure 1-10 Differences between traffic shaping and traffic policing



1.3.7 Congestion Management

If the network is congested intermittently when a delay-sensitive service requires higher QoS than a delay-insensitive service, you need to perform congestion management. The bandwidth needs to be increased if the network is always congested. Based on the queuing technology and scheduling algorithms, the S6700 sends packets in queues to provide congestion management.

Based on the queuing and scheduling policies, the congestion management technologies of the S6700 are classified into Priority Queuing (PQ), Weighted Round Robin (WRR), Deficit Round Robin (DRR), PQ+DRR, and PQ+WRR. Each scheduling technology solves the problem of certain traffic, and greatly influences bandwidth allocation, delay, and jitter.

On the S6700, there are eight queues at the outbound direction on each interface, which are identified by index numbers. The index numbers range from 7 to 0. In Priority Queue (PQ) scheduling, queue 7 is of the highest priority, whereas queue 0 is of the lowest priority. Based on the mapping between local priorities and queues, the S6700 sends the classified packets to queues, and then schedules the packets according to the queue scheduling mechanisms.

- PQ scheduling

  PQ scheduling is designed for core services, and is applied to the queues in descending order of priorities. A queue with lower priories are processed only after all the queues with higher priorities are empty. In PQ scheduling, the packets of core services are placed into a queue of a higher priority, and the packets of non-core services such as email services are placed into a queue of a lower priority. This ensures that core services are processed first, and non-core services are sent during the intervals when core services are not processed.

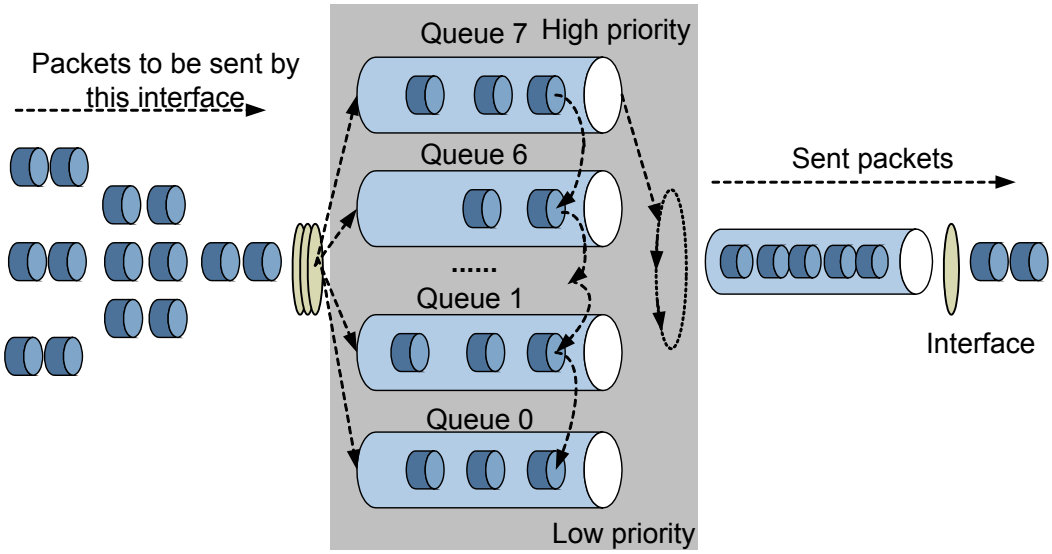
  As shown in **Figure 1-11**, the priorities of queues 7 to 0 are in descending order of priorities. The packets in queue 7 are processed first as long as packets can be transmitted on the link. The scheduler processes packets in queue 6 only after queue 7 becomes empty. The packets in queue 6 are sent at the link rate when packets in queue 6 need to be sent and queue 7 is empty. The packets in queue 5 are sent at the link rate when queue 6 and queue 7 are empty.

PQ scheduling is useful for short-delay services. Assume that data flow X is mapped to the queue of the highest priority on each node. When packets of data flow X reach a node, the packets are processed first.

The PQ scheduling mechanism, however, may result in starvation of packets in queues of lower priorities. For example, if data flows mapped to queue 7 arrive at 100% link rate in a period, the scheduler does not process flows in queue 6 and following queues.

To prevent starvation of packets, upstream devices need to define service features of data flows carefully. The service flows mapped to queue 7 cannot exceed a certain percentage of the link capacity. In this case, queue 7 is always in empty state. The scheduler can process packets in queues of lower priorities.

Figure 1-11 Networking diagram of PQ scheduling

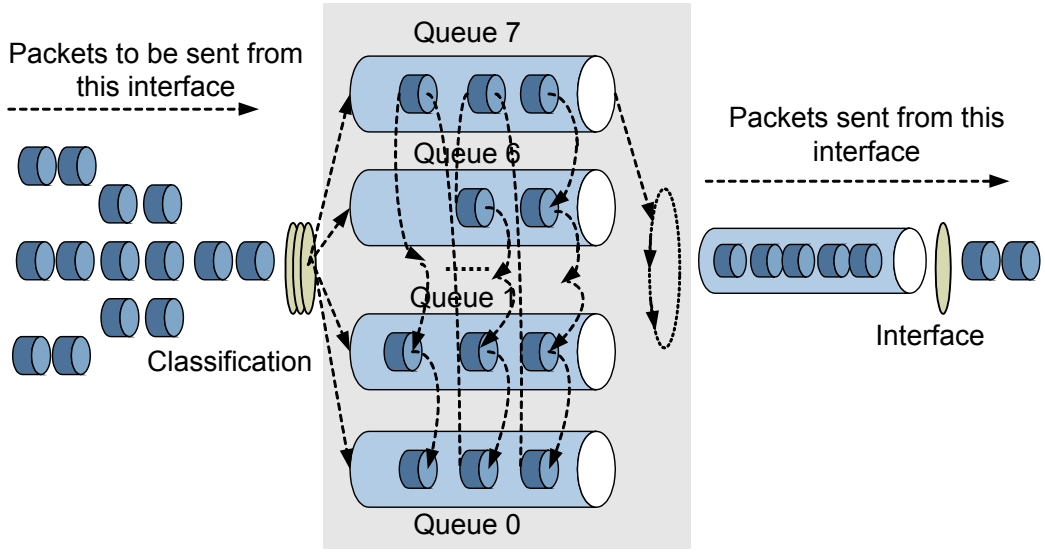


- WRR scheduling

  WRR scheduling is based on Round Robin (RR) scheduling. Packets in each queue are scheduled in a polling manner depending on the weight of the queue. RR scheduling equals WRR scheduling with the weight being 1.

  Figure 1-12 shows WRR scheduling.

Figure 1-12 Networking diagram of WRR scheduling



In WRR scheduling, the S6700 performs scheduling in a polling manner according to the weight of each queue. After one round of scheduling, the weight of all queues is decreased by 1. The queue whose weight is decreased to 0 cannot participate the scheduling. When the weight of all the queues is decreased to 0, the next round of scheduling starts. For example, the weights of eight queues on an interface are set to 4, 2, 5, 3, 6, 4, 2, and 1. For the WRR scheduling results, see **Table 1-3**.

Table 1-3 WRR scheduling results

Queue Index	Queue 7	Queue 6	Queue 5	Queue 4	Queue 3	Queue 2	Queue 1	Queue 0
Queue Weight	4	2	5	3	6	4	2	1
Queue in the first round of scheduling	Queue 7	Queue 6	Queue 5	Queue 4	Queue 3	Queue 2	Queue 1	Queue 0
Queue in the second round of scheduling	Queue 7	Queue 6	Queue 5	Queue 4	Queue 3	Queue 2	Queue 1	-

Queue Index	Queue 7	Queue 6	Queue 5	Queue 4	Queue 3	Queue 2	Queue 1	Queue 0
Queue in the third round of scheduling	Queue 7	-	Queue 5	Queue 4	Queue 3	Queue 2	-	-
Queue in the fourth round of scheduling	Queue 7	-	Queue 5	-	Queue 3	Queue 2	-	-
Queue in the fifth round of scheduling	-	-	Queue 5	-	Queue 3	-	-	-
Queue in the sixth round of scheduling	-	-	-	-	Queue 3	-	-	-
Queue in the seventh round of scheduling	Queue 7	Queue 6	Queue 5	Queue 4	Queue 3	Queue 2	Queue 1	Queue 0
Queue in the eighth round of scheduling	Queue 7	Queue 6	Queue 5	Queue 4	Queue 3	Queue 2	Queue 1	-

Queue Index	Queue 7	Queue 6	Queue 5	Queue 4	Queue 3	Queue 2	Queue 1	Queue 0
Queue in the ninth round of scheduling	Queue 7	-	Queue 5	Queue 4	Queue 3	Queue 2	-	-
Queue in the tenth round of scheduling	Queue 7	-	Queue 5	-	Queue 3	Queue 2	-	-
Queue in the eleventh round of scheduling	-	-	Queue 5	-	Queue 3	-	-	-
Queue in the twelfth round of scheduling	-	-	-	-	Queue 3	-	-	-

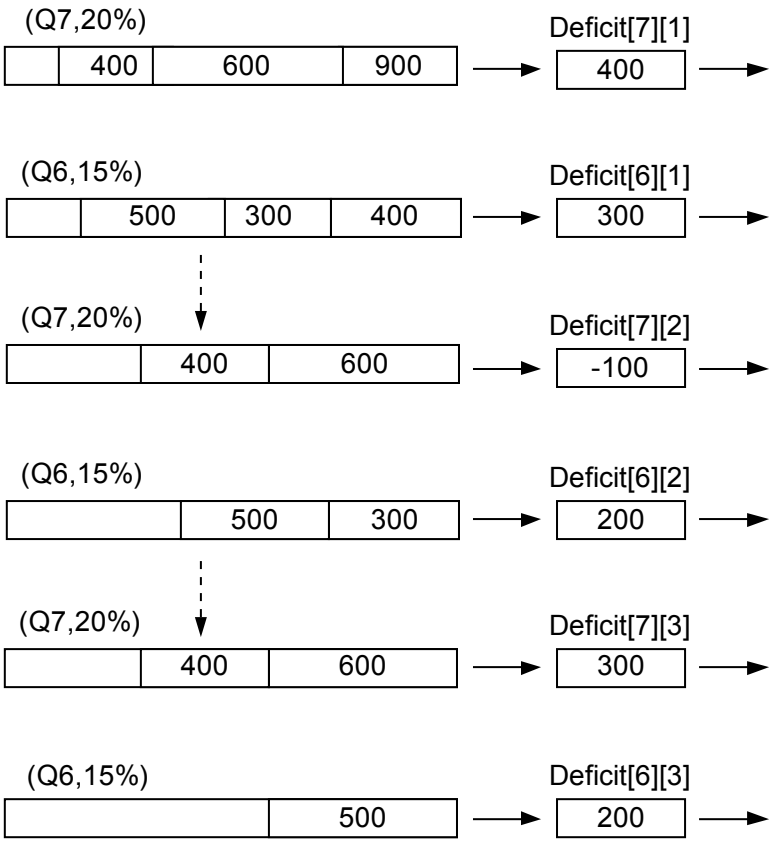
From the statistics, you can view that the number of times for the packets to be scheduled in each queue is in direct ratio to the weight of this queue. The higher the weight is, the more the number of times packets in the queue are scheduled. The unit for the WRR scheduling is packet; therefore, there is no fixed bandwidth for each queue. With an equal scheduling chance, the bandwidth obtained by packets of a large size is greater than the bandwidth obtained by packets of a small size.

WRR scheduling offsets the disadvantage of PQ scheduling where packets in queues of lower priories may be not processed for a long time. In addition, WRR scheduling can dynamically adjust the time for scheduling packets in a queue even though packets of multiple queues are scheduled in a polling manner. For example, if a queue is empty, WRR scheduling ignores this queue and starts to schedule the next queue. This ensures the efficient use of bandwidth. WRR scheduling, however, cannot ensure that short-delay services are scheduled in time.

- DRR scheduling

DRR scheduling is also based on RR. DRR scheduling solves the problem of WRR scheduling, which considers only packets. In the equal scheduling chance, the bandwidth obtained by packets of a large size is greater than the bandwidth obtained by packets of a small size. DRR scheduling, however, considers the packet length during scheduling, ensuring equality of scheduling packets. **Figure 1-13** shows DRR scheduling.

Figure 1-13 Networking diagram of DRR scheduling



Assume that the weights of Q7, Q6, Q5, Q4, Q3, Q2, Q1, and Q0 are set to 40, 30, 20, 10, 40, 30, 20, and 10, as shown in **Figure 1-13**. Q7 and Q6 can obtain 20% and 15% of the bandwidth respectively. In Q7, there are packets of 400 bytes, 600 bytes, and 900 bytes; in Q6, there are packets of 500 bytes, 300 bytes, and 400 bytes. In each scheduling, the system allocates bandwidth to each queue according to the weight. Assume that the bandwidth of Q7 is 400 bytes/s and the bandwidth of Q6 is 300 bytes/s. Deficit indicates the bandwidth deficit of each queue in each scheduling. Before scheduling is performed, if the bandwidth deficit of a queue is a negative value, the bandwidth deficit values of two queues are changed. That is, 400 and 300 are added to the two bandwidth values respectively according to the bandwidth allocation ratio. Then the scheduling can be performed.

- First scheduling

  The bandwidth is allocated first. Deficit [7][1] = 400, Deficit [6][1] = 300, packets of 900 bytes in Q7 and packets of 400 bytes in Q6 are sent. Then, Deficit [7][1] = -500, Deficit [6][1] = -100.

- Second scheduling

  If the bandwidth deficit is a negative value, the bandwidth value is added according to the weight of the queue. Deficit [7][2] = -500 + 400 = -100, Deficit [6][1] = -100 + 300

= 200, packets in Q7 are not scheduled because the deficit of Q7 is negative. Packets of 300 bytes in Q6 are sent. Then, Deficit [6][2] = -100.

- Third scheduling

  If the bandwidth deficit is still negative, the bandwidth is added according to the weight of the queue continuously. Deficit [7][3] = -100 + 400 = 300, Deficit [6][3] = -100 + 300 = 200, packets in Q7 are scheduled because the deficit of Q7 is positive. Packets of 600 bytes in Q7 and packets of 500 bytes in Q6 are sent. Then, Deficit [7][1] = -300, Deficit [6][1] = -200. Such a process is repeated until Q7 and Q6 obtain 20% and 15% of the bandwidth respectively. Therefore, you can obtain the required bandwidth by setting the weights.

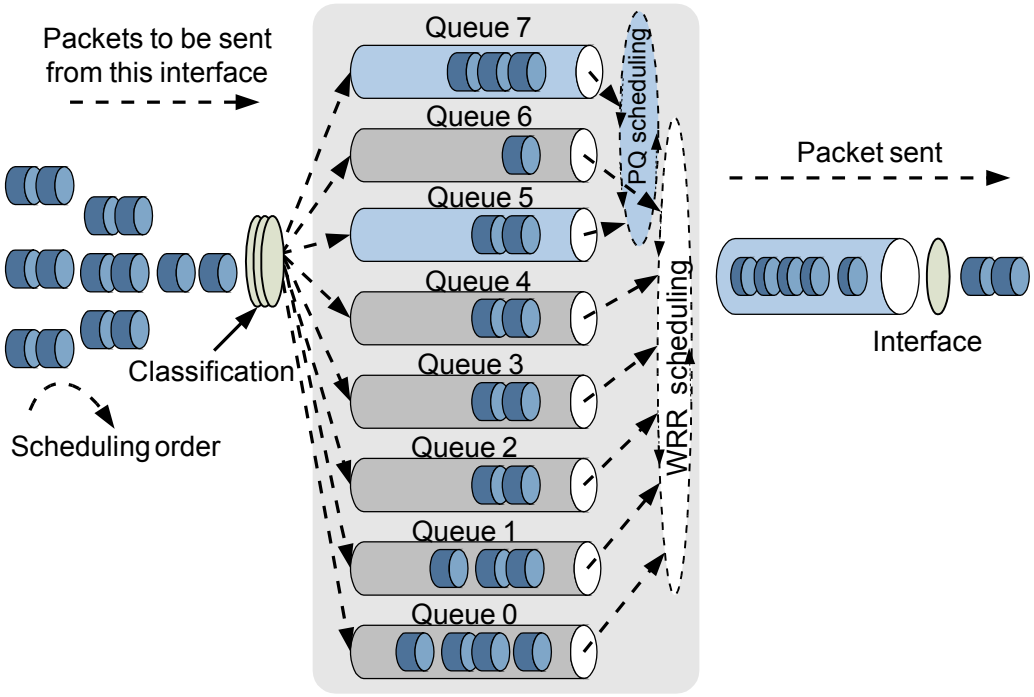
In DRR scheduling, short-delay services still cannot be scheduled in time.

- PQ+WRR scheduling

  PQ scheduling and WRR scheduling have advantages and disadvantages. To offset disadvantages of PQ scheduling or DRR scheduling, you can use PQ+WRR scheduling. In this manner, packets from queues of lower priorities can obtain the bandwidth through WRR scheduling and short-delay services can be first scheduled through PQ scheduling.

  On the S6700, you can set WRR parameters of the queue. The eight queues on each interface are classified into two groups. One group includes queues Queue 7, Queue 6, and Queue 4, and is scheduled in PQ mode; the other group includes Queue 5, Queue 3, Queue 2, Queue 1, and Queue 0, and is scheduled in WRR mode. **Figure 1-14** shows PQ+WRR scheduling.

Figure 1-14 PQ+WRR scheduling



During the scheduling, the S6700 first schedules the packet traffic in Queue 7 and Queue 5 in PQ mode. The S6700 circularly schedules the packet traffic in other queues in WRR mode only after the packet traffic in Queue 7 and Queue 5 are scheduled. Queue 6, Queue 4, Queue 3, Queue 2, Queue 1, and Queue 0 have their own weights. Important protocol

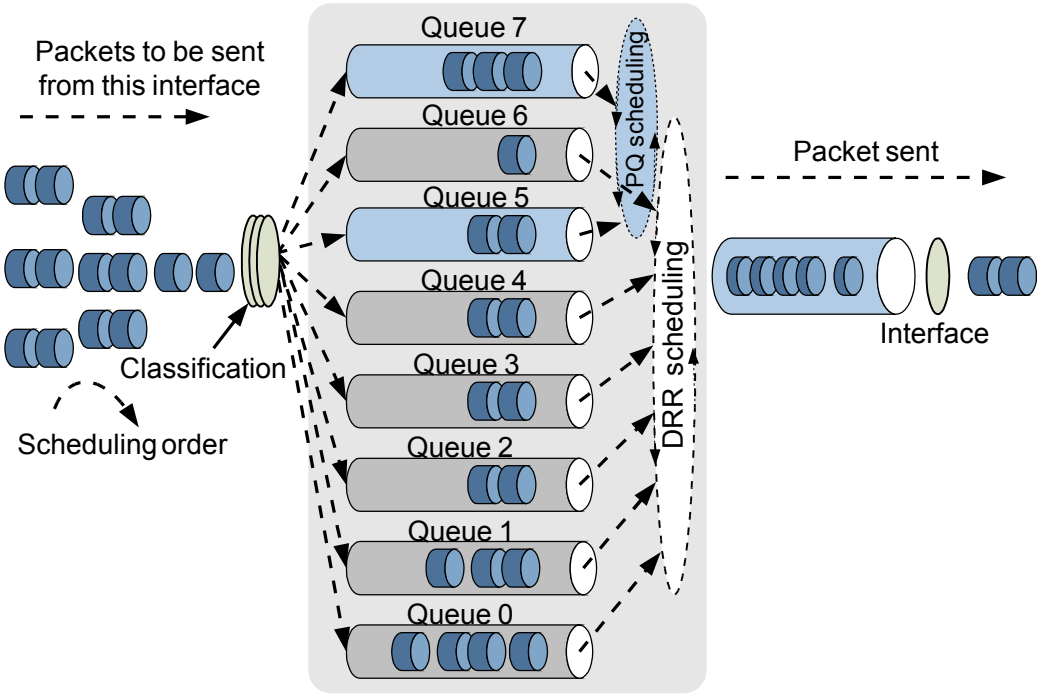
packets or short-delay service packets need to be placed in queues adopting PQ scheduling, so that they can be scheduled first. Other packets are placed in the queues adopting WRR scheduling.

- PQ+DRR scheduling

  Similar to PQ+WRR, PQ+DRR scheduling offsets disadvantages of PQ scheduling or DRR scheduling. If only PQ scheduling is used, packets in queues with lower priorities cannot obtain the bandwidth for a long time. If only DRR scheduling is used, short-delay services such as voice services cannot be scheduled first. PQ+DRR scheduling can make full use of the advantages of both PQ and DRR scheduling and can offset their disadvantages.

  Eight queues on the interface of the S6700 are classified into two groups. You can specify PQ scheduling for certain groups and DRR scheduling for other groups.

Figure 1-15 Networking diagram of PQ+DRR scheduling



As shown in **Figure 1-15**, the S6700 first schedules packet traffic in queue 7 and queue 5 in PQ mode, and then schedules packet traffic in queue 6, queue 4, queue 3, queue 2, queue 1, and queue 0 in DRR mode. Queue 6, queue 4, queue 3, queue 2, queue 1, and queue 0 have their ensured bandwidth and peak bandwidth.

Important protocol packets or short-delay service packets need to be placed in queues adopting PQ scheduling, so that they can be scheduled first. Other packets are placed in the queues adopting WRR scheduling.

1.3.8 Congestion Avoidance

Congestion avoidance is a traffic control mechanism used to remove network overload by adjusting network traffic. With this mechanism, the S6700 can monitor the usage of network resources and discard packets when the network congestion becomes more severe.

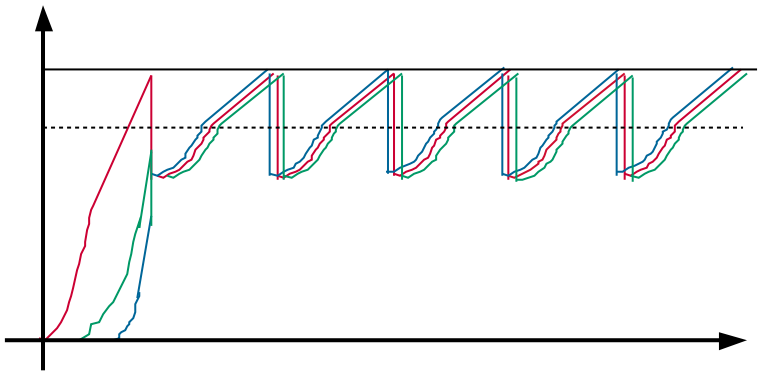
This section describes the basics of congestion avoidance:

- Traditional tail-drop policy

  The traditional packet loss policy uses the tail-drop method. When the queue reaches its maximum length, the packets that follow recently (stored at the tail of the queue) are discarded.

  This packet loss policy can cause global TCP synchronization. As a result, the TCP connection cannot be set up all the time. The three colors represent three TCP connections. When packets from multiple TCP connections are discarded, these TCP connections enter the congestion avoidance and slow start state at the same time. As a result, these TCP connections send less packets at a time but increase the number of packets later at another time. Thus, the volume of traffic varies greatly from time to time.

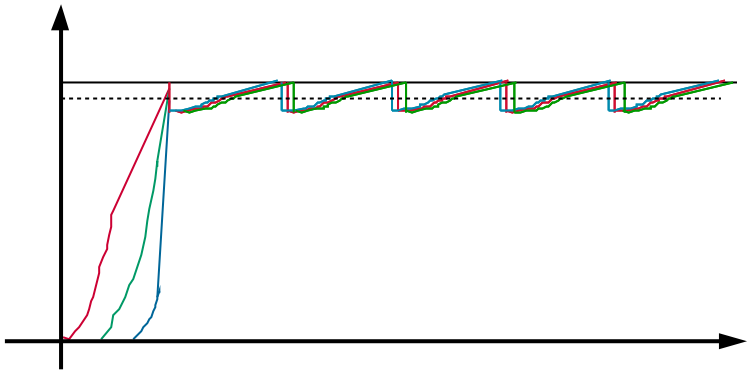
  Figure 1-16 Networking diagram of the tail-drop policy



- SRED

  To avoid global TCP synchronization, RED is used. RED randomly discards packets, so that multiple TCP connections do not reduce their transmission speed at the same time. Global TCP synchronization is thus prevented. The rate of TCP traffic and network traffic become stable.

  Figure 1-17 Networking diagram of RED



  Based on the RED technology, the S6700 provides WRED. WRED colors outgoing packets as green, yellow, red, and non-tcp according to the priorities and types of the packets on

an interface. You can set the upper threshold, lower threshold, and packet loss ratio for these four types of packets independently. When the number of packets reaches the lower threshold, the S6700 starts to discard packets. When the number of packets reaches the upper threshold, the S6700 discards all the packets. With the increase of the threshold, packet loss ratio increases. The greatest packet loss ratio does not exceed the set packet loss ratio. When the packet loss ratio reaches the set one, all the packets are discarded. In this manner, packets in queues are discarded according to the drop probability. Congestion is thus prevented to a certain degree.

The S6700 implements the Simple Random Early Detection (SRED) technology based on the RED technology. In a queue on an outbound interface, the S6700 colors the packets red or yellow according to the priorities of packets; the S6700 sets a threshold for discarding red packets, a threshold for discarding yellow packets, and the drop probability.

Based on SRED, the S6700 actively discards packets in the queue based on the drop probability to adjust the rate of outgoing traffic at the interface.

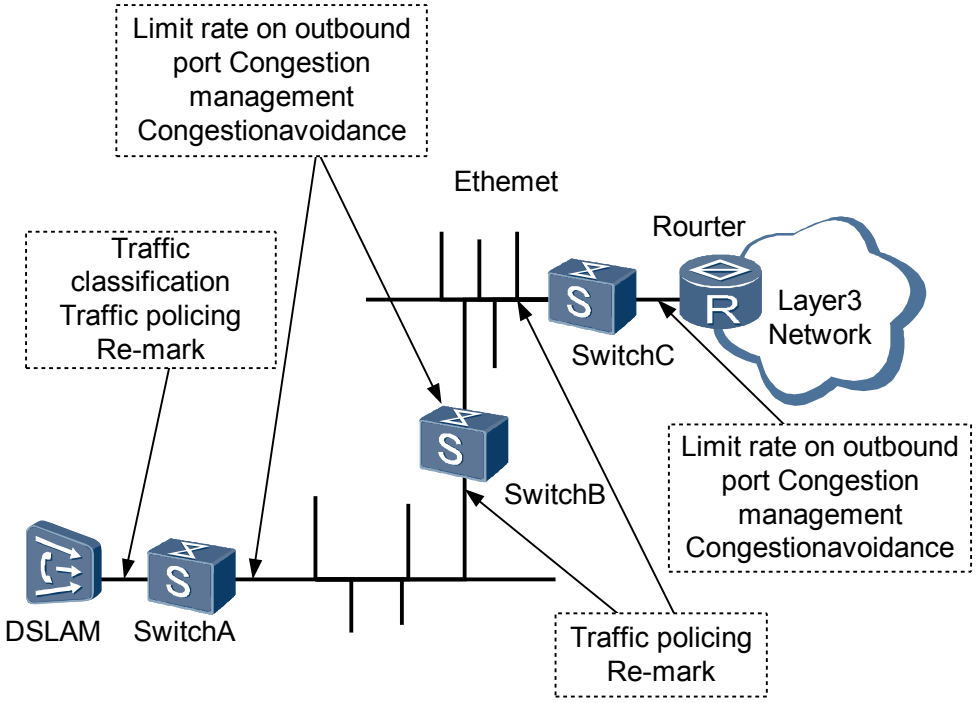
1.3.9 Limit Rate on an Interface at the Outbound Direction

Through line rate (LR), the total rate of sending packets on an interface can be limited. LR also uses the token bucket for traffic control. If the limit rate function is configured on an interface of the device, all the packets sent from this interface need to be processed by the token bucket of the LR first. If there are sufficient tokens in the token bucket, packets are sent; otherwise, packets are discarded. Unlike traffic policing, LR controls all the packets that pass through an interface. If the transmission of all the packets needs to be limited, LR is recommended, because it is easy to implement.

1.4 Applications

On an Ethernet, packet flows can be encapsulated into VLAN frames as required. The Switch can obtain the 802.1p priority from the VLAN frame header, which is the basis for providing differentiated QoS. **Figure 1-18** shows class-based QoS used on the Ethernet.

Figure 1-18 Networking diagram for configuring QoS on the Ethernet



Ingress of the switch at the Edge of the Network

On the ingress of the S6700 at the edge of the network, the service flows are classified according to QoS requirements of service flows. Then, the service flows are policed and the burst traffic that exceeds the rate limit is punished. At the same time, certain fields of the packet traffic are re-marked as required, so that the devices on the network classify packets according to the re-marked values.

Egress of the switch at the Edge of the Network

After classification, service flows are sent to different queues on the outbound interface of the S6700 according to priorities. On the egress of the S6700 at the edge of the network, congestion management is performed through implementation of various scheduling policies for queues. Before sending packets, the device restricts the transmission rate of the outgoing packets on the interface.

switch within the Network

On the incoming interfaces of the switches within the network, you can re-classify the traffic or classify the traffic according to the upstream remarking information. At the same time, you can re-mark the traffic again. Then, service flows of different types are sent to various queues. On the egress, you can perform congestion management, and rate limit on the outbound interface.

1.5 Terms and Abbreviations

Term

Term	Description
Differentiated service	Differentiated Service is called DiffServ for short. DiffServ is a multi-service model and can satisfy different QoS requirements. An application does not need to notify the communication device before sending packets. In addition, the device does not need to maintain the status of each flow on the network. The application provides special services according to specified QoS of each packet, including classification, traffic shaping, traffic policing, and queuing of packets. CAR and queuing technologies are used.
CAR	Committed Access Rate is CAR for short. If there are sufficient tokens in the token bucket to forward packets, it means that traffic is within the rate limit. If not, the traffic exceeds the rate limit. CAR supports single rate single bucket, srTCM, and trTCM. It evaluates the traffic according to the preset matching rules, and measures and polices the traffic.
CBS	Committed Burst Size is CBS for short. It indicates the capacity of the token bucket, or the maximum size of traffic that is allowed to burst at a time. The burst size must be greater than the maximum length of packets.
CIR	Committed Information Rate is CIR for short. It indicates the average rate at which tokens are put into the token bucket, or the average rate of the traffic flow allowed. Generally, the traffic rate is smaller than the CIR.

Abbreviation

Abbreviation	Full Spelling
QoS	Quality of Service
srTCM	Single Rate Three Color Marking
trTCM	Two Rate Three Color Marking
WFQ	Weighted Fair Queue
VLAN	Virtual Local Area Network
MQC	Modular QoS Command