Huawei Certification

HCDP-IENP

Improving Enterprise Network Performance



Huawei Technologies Co.,Ltd

Copyright © Huawei Technologies Co., Ltd. 2012. All rights reserved.

No part of this document may be reproduced or transmitted in any form or by any means without prior written consent of Huawei Technologies Co., Ltd.

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 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 Certification

HCDP-IENP Improving Enterprise Network Performance

Edition 1.6

Huawei Certification System

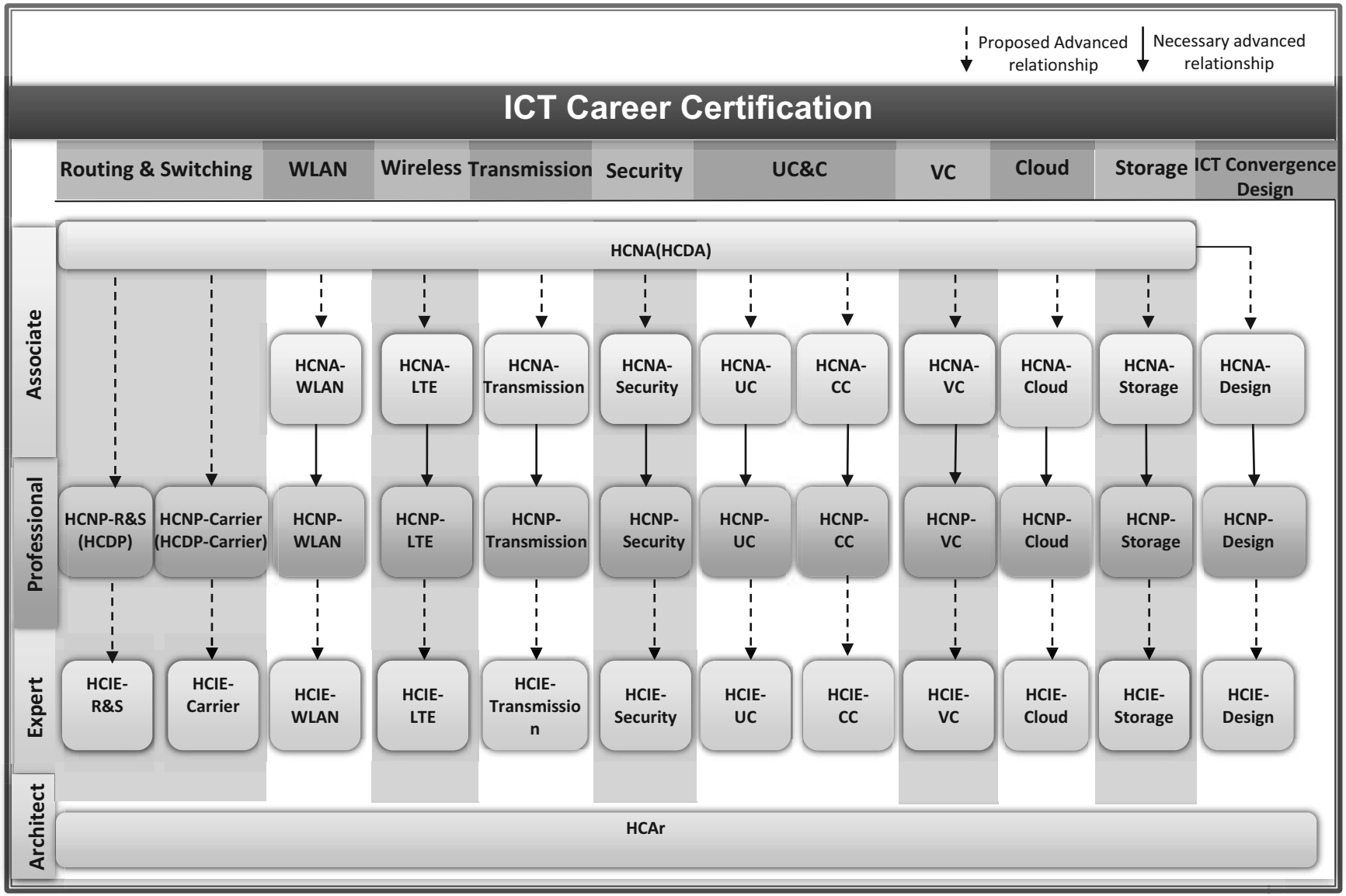
Relaying on its strong technical and professional training system, according to different customers at different levels of ICT technology, Huawei certification is committed to provide customs with authentic, professional certification.

Based on characteristics of ICT technologies and customers' needs at different levels, Huawei certification provides customers with certification system of four levels.

HCDA(Huawei Certification Datacom Associate) is primary for IP network maintenance engineers, and any others who want to learn the IP network knowledge.HCDA certification covers the TCP/IP basics, routing, switching and other common foundational knowledge of IP networks, together with Huawei communications products, versatile routing platform VRP characteristics and basic maintenance.

HCDP-Enterprise (Huawei Certification Datacom Professional-Enterprise) is aimed at enterprise-class network maintenance engineers, network design engineers, and any others who want to in depth grasp routing, switching, network adjustment and optimization technologies. HCDP-Enterprise is consist of IESN (Implementing Enterprise Switch Networks), IERN (Implementing Enterprise Routing Networks), and IENP (Improving Enterprise Network performance), which includes advanced IPv4 routing and switching technology principle, IP technology of network security, high availability and QoS, as well as the implementation in Huawei products.

HCIE-Enterprise(Huawei Certified Internetwork Expert-Enterprise) is designed to endue engineers with a variety of IP network technology and proficiency in maintenance, diagnostics and troubleshooting of Huawei products, which equips the engineers with competence in planning, design and optimization of large-scale IP network.


Proposed Advanced relationship
Necessary advanced relationship
ICT Career Certification
Routing & Switching
WLAN
Wireless Transmission
Security
UC&C
VC
Cloud
Storage
ICT Convergence Design
Associate
Professional
Expert
Architect
HCNA(HCDA)
HCNA-WLAN
HCNA-LTE
HCNA-Transmission
HCNA-Security
HCNA-UC
HCNA-CC
HCNA-VC
HCNA-Cloud
HCNA-Storage
HCNA-Design
HCNP-R&S (HCDP)
HCNP-Carrier (HCDP-Carrier)
HCNP-WLAN
HCNP-LTE
HCNP-Transmission
HCNP-Security
HCNP-UC
HCNP-CC
HCNP-VC
HCNP-Cloud
HCNP-Storage
HCNP-Design
HCIE-R&S
HCIE-Carrier
HCIE-WLAN
HCIE-LTE
HCIE-Transmission
HCIE-Security
HCIE-UC
HCIE-CC
HCIE-VC
HCIE-Cloud
HCIE-Storage
HCIE-Design
HCAr

Foreword

Outline

This book presents training courses for HCDP-IENP, targeting audience ranging from learners preparing for the HCDP-IENP examination to readers wishing not only to understand Huawei security products and technologies, High Availability (HA) technologies, and Quolity of Service (QoS) technologies, but also to comprehend implementations of these technologies in the Huawei Versatile Routing Platform (VRP).

Content

This book is composed of three modules that describe Huawei security products and technologies, HA technologies, and QoS technologies, as well as introduce configurations and implementations of these technologies in the VRP. Module 1 describes the functions and service features of Huawei Eudemon firewalls and will help you comprehensively understand network security concepts and Huawei security products. Module 2 describes HA technologies and will help familiarize you with principles and applications of each HA technology. Module 3 describes IP QoS technology and will help you understand the QoS principle and know how to configure QoS in the VRP. This book aims to guide you through the preceding concepts step by step so that you may fully understand the technologies and implentations of these technologies in Huawei products. You can also choose to read modules you are most interested in as desired.

Readers' Knowledge Background

Any of the following experience or knowledge will better your understanding of this book:

(1) Participating in the HCDA training

(2) Passing the HCDA examination

(3) Familiar with TCP/IP and obtaining basic IP network knowledge

(4) Familiar with multiple routing protocols, such as OSPF, and BGP

Icons Used in This Book



IPv6 Router



SOHO Router



Voice Router



Low-end Router



High-end Router



Core Router



Hub



Convergence Switch



Socket switch



Core Switch



Edge Switch



Cascade Switch



AP



AP Amplifier



Wireless Bridge



Wireless Network Card



Access Server



Audio Gateway



Firewall

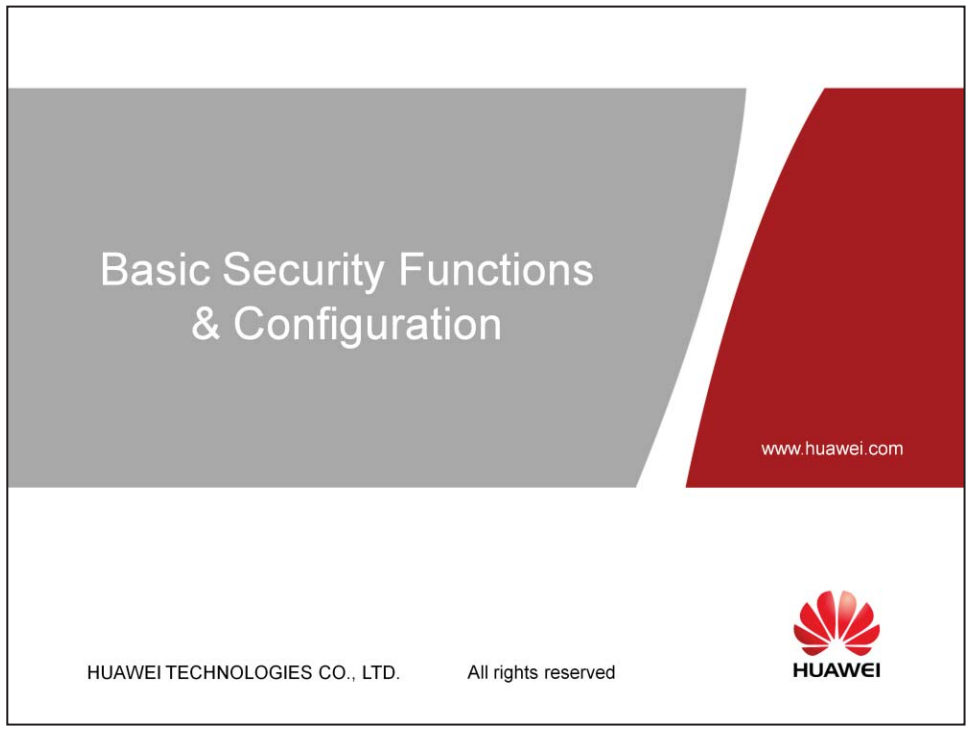


Internet Telephony

Table of Contents

Module 1 Security.....Page 1
Basic Security Functions &Configurations.....Page 3
NAT Functions & Configuration.....Page 74
Attack Defence & Configurations.....Page 90
USG Firewall Dual System and Configuration.....Page 122
Module 2 HA Technology .....Page 155
HA Technology .....Page 157
VRRP Principle and Configuration .....Page 185
Module 3 QoS.....Page 231
IP QoS Overview .....Page 233
Differentiated Service Model Overview .....Page 246
IP QoS Technologies.....Page 260
Class-based QoS.....Page 346
IP QoS Implementation on High-end Routers .....Page 376

Module 1
Security


Basic Security Functions
& Configuration
www.huawei.com
HUAWEI TECHNOLOGIES CO., LTD.
All rights reserved
HUAWEI



Foreword

This section introduces the main security technologies and features of USG series products, as well as device configuration. This includes firewall zones, modes, ASPF technology, NAT technology and extension techniques.

HUAWEI TECHNOLOGIES CO., LTD. All rights reserved





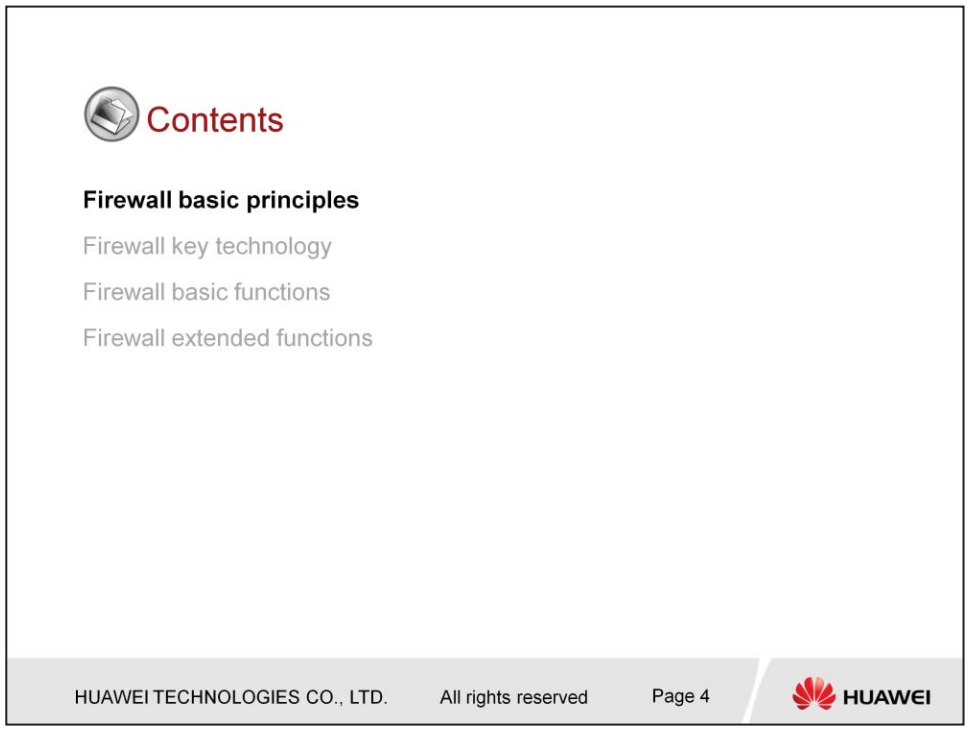
Objectives

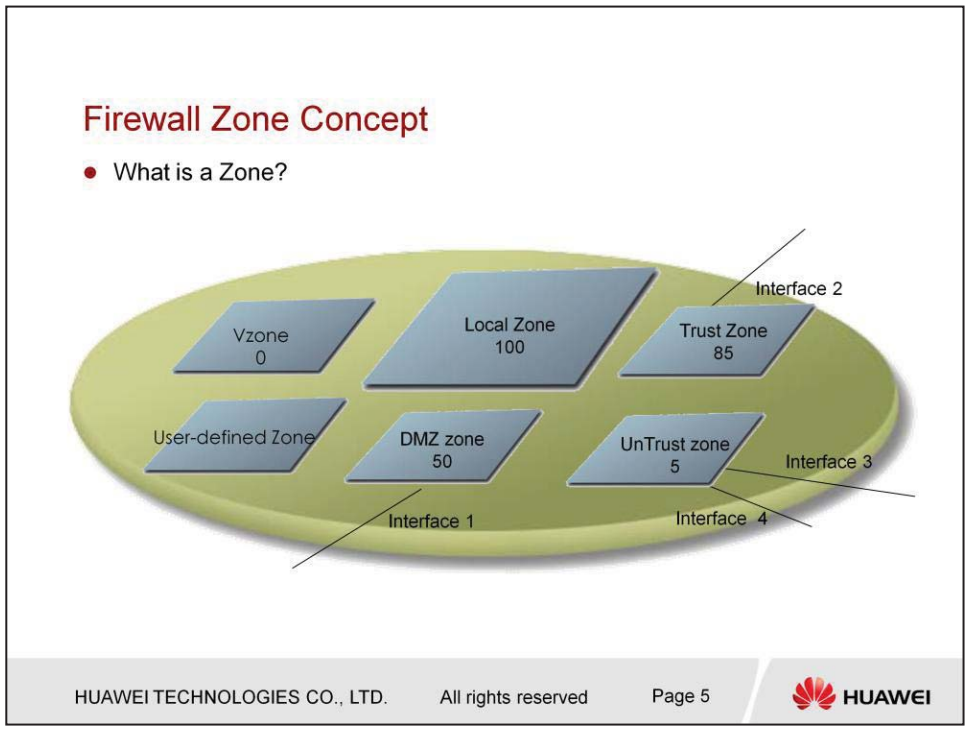
Upon completion of this section, you will be able to:

- Master USG product's main security technologies and features
- Master USG security feature configuration

HUAWEI TECHNOLOGIES CO., LTD. All rights reserved




Contents
Firewall basic principles
Firewall key technology
Firewall basic functions
Firewall extended functions
HUAWEI TECHNOLOGIES CO., LTD.
All rights reserved
Page 4
HUAWEI

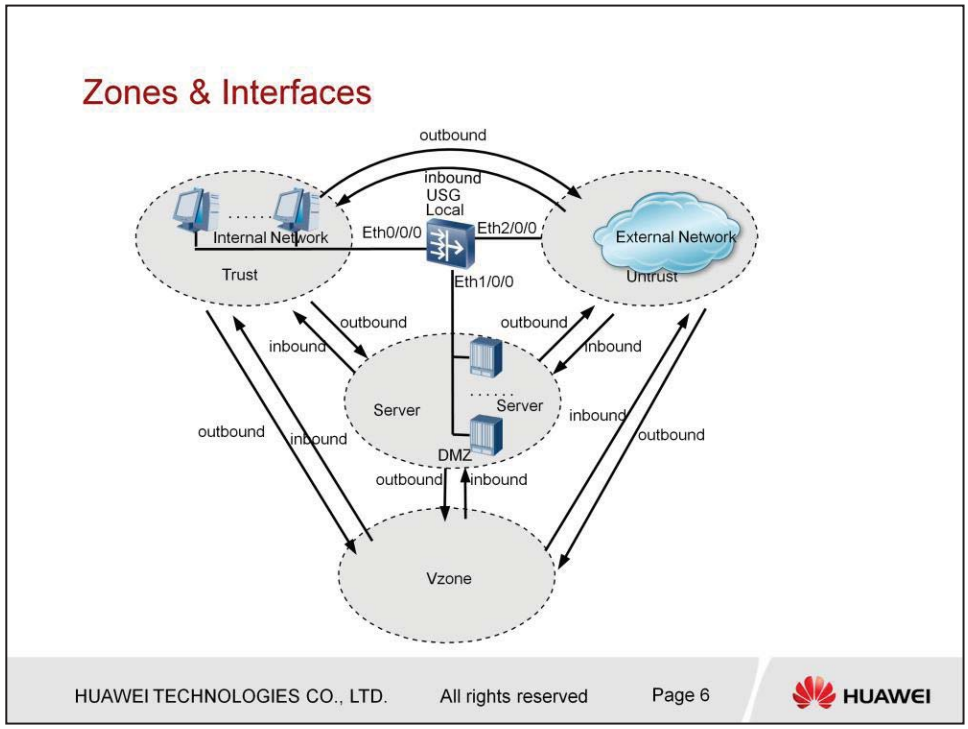


Zone is a concept introduced in firewalls, and is one of main features distinguishing the firewall from the router.

For the router, the network security check is performed on interfaces because the networks connected with each interface are equal in security. That is, there is no obvious difference between internal networks and external networks for the router. In this way, when a data stream unidirectionally passes through a router, it may be checked twice on both the inbound interface and the outbound interface to meet the separate security definitions on each interface. However, the firewall's situation is different, where internal networks and external networks are clearly defined. The firewall protects internal networks from illegal intrusion of external networks.

When a data stream passes through a firewall device, the security operation triggered varies with data stream direction. At this time, it is not suitable to check the security policy on the interface of the firewall, therefore the firewall introduces the security zone concept.

A security zone is composed of one or more interfaces with the same security level.



A security zone is composed of one or more interfaces with the same security level.

The features of the security zones are as follows:

The security level is denoted by an integer in the range of 0 to 100. The greater the number is, the higher the level is.

There are no two zones with the same security level.

The firewall security check will be enabled only when data is transmitted between interfaces with different security levels rather than the interfaces in the same security zone.

The direction of data flow from high priority zone to low priority zone is defined as outbound.

The direction of data flow from low priority zone to high priority zone is defined as inbound.

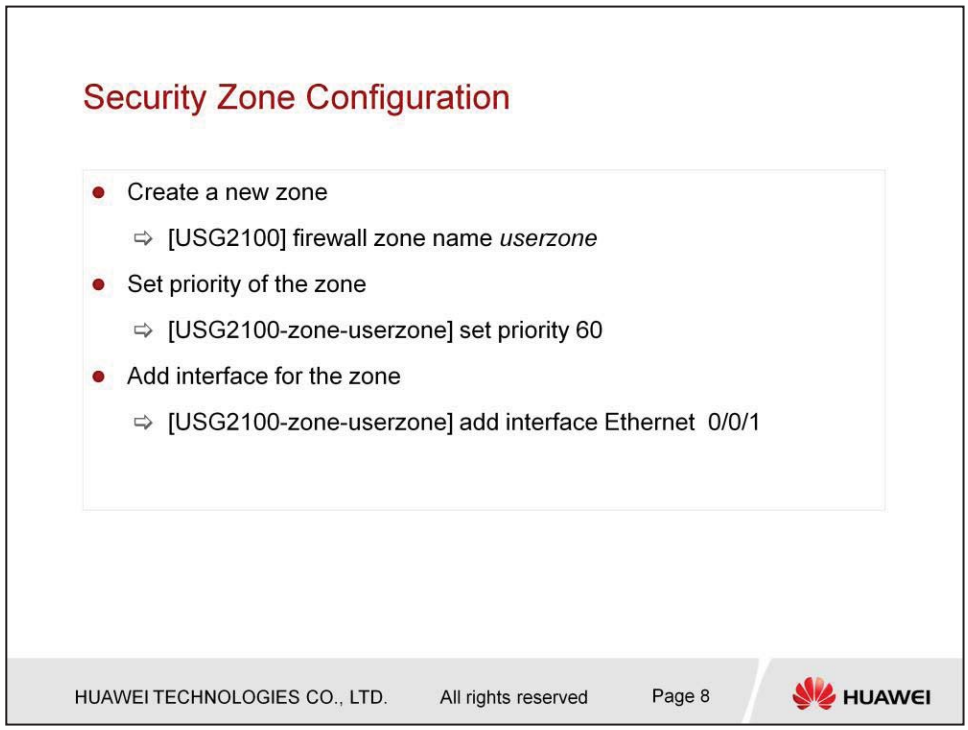
The Local zone has no interface. The firewall device is in the Local zone.

The Vzone has no interface, and is used for the traffic forwarding between the Virtual Private Network (VPN) instances.

The traffic of data flow between VPN instances needs to hop through their own Vzone.

For example, when a data flow moves from the Trust zone of the VPN1 to the DMZ zone of the VPN2, the data flow need to enter

from the Trust zone of the VPN1 to the Vzone of VPN1, and then it moves from the Vzone of the VPN2 to the DMZ zone of the VPN2. All VPN instances of the Vzone are inter-connected. Data can flow free of the restriction from interzone-filtering rule on the firewall.



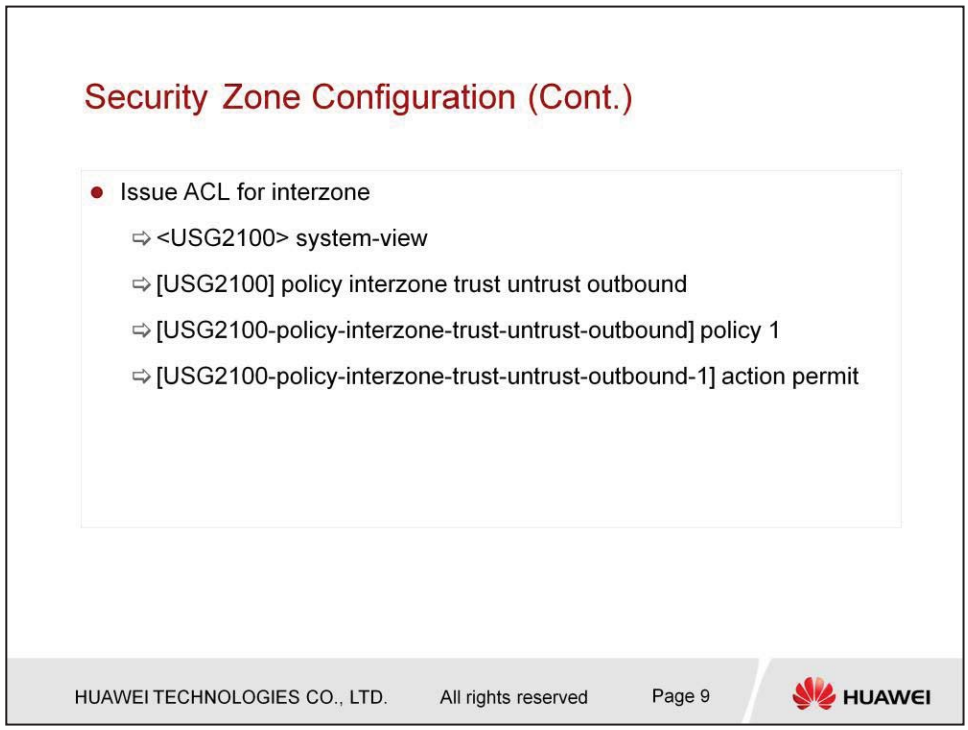
firewall zone [name] zone-name

By default, there are four reserved security zones, the local zone, the trust zone, the DMZ zone, and the untrust zone, which need neither be created nor be deleted.

The keyword name is required for creating a new security zone instead of entering a reserved or existing security zone view.

You can only configure priority for security zones that are created by yourself. The priorities of the default security zones (the local zone, the trust zone, the DMZ zone, and the untrust zone) are fixed to 100, 85, 50, and 5 respectively, which cannot be reconfigured. It is not possible to configure the same priority for two security zones in a system.

With the exception of the local zone, all the other security zones need to be associated with some interface of the firewall respectively, that is, to add the interface, which can be either a physical interface or a logic interface, into those zones. The command in this section can be used repeatedly to specify interfaces for security zones. A security zone can support 1024 interfaces at most.



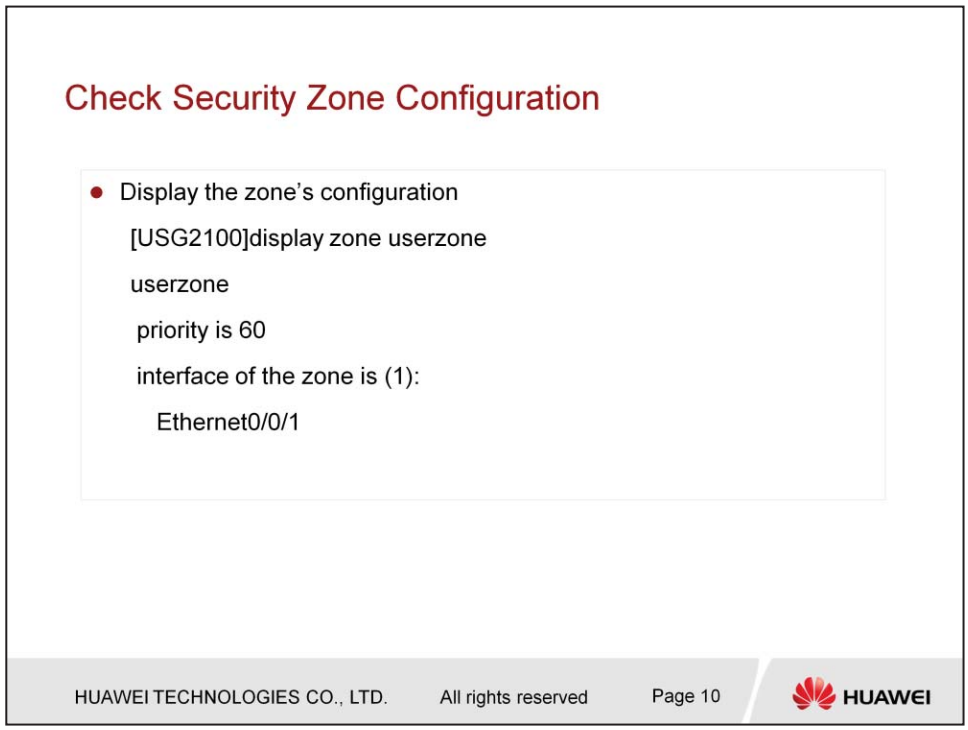
packet-filter acl-number { inbound | outbound }.

acl-number：ACL number, the range is 2000～3999、5000～5499.

inbound：filter packets coming from inbound direction.

outbound： filter packets coming from outbound direction.

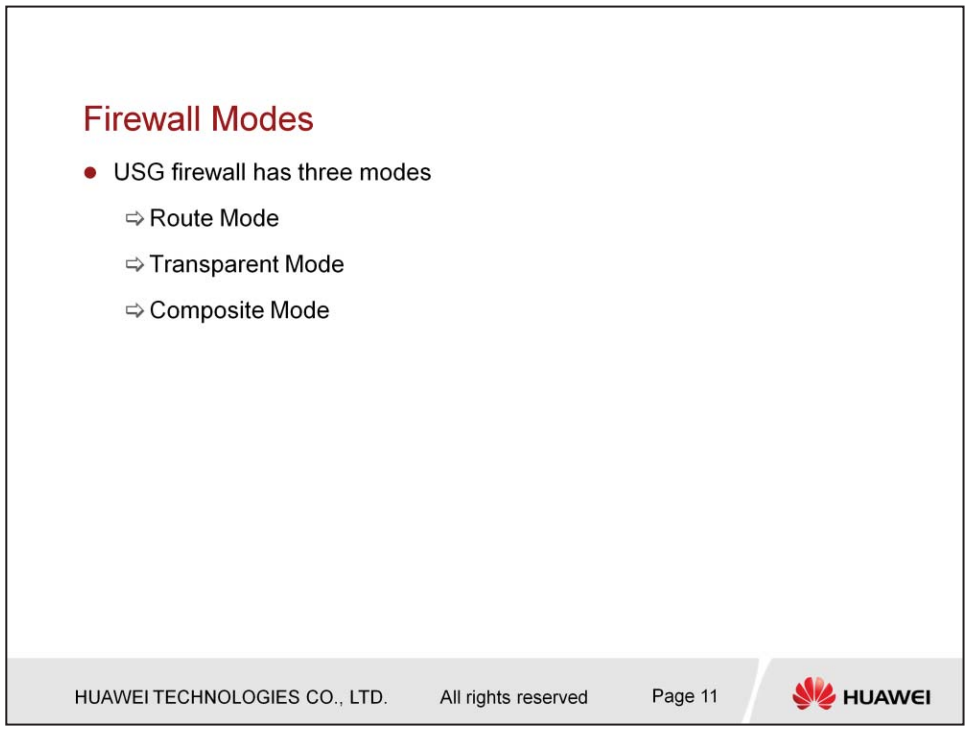
You can specify an ACL at the inbound and another one at the outbound in interzone. By default, no ACL is applied in interzone.



display zone [zone-name] [interface | priority].

zone-name：zone name.

priority：the priority of the zone.

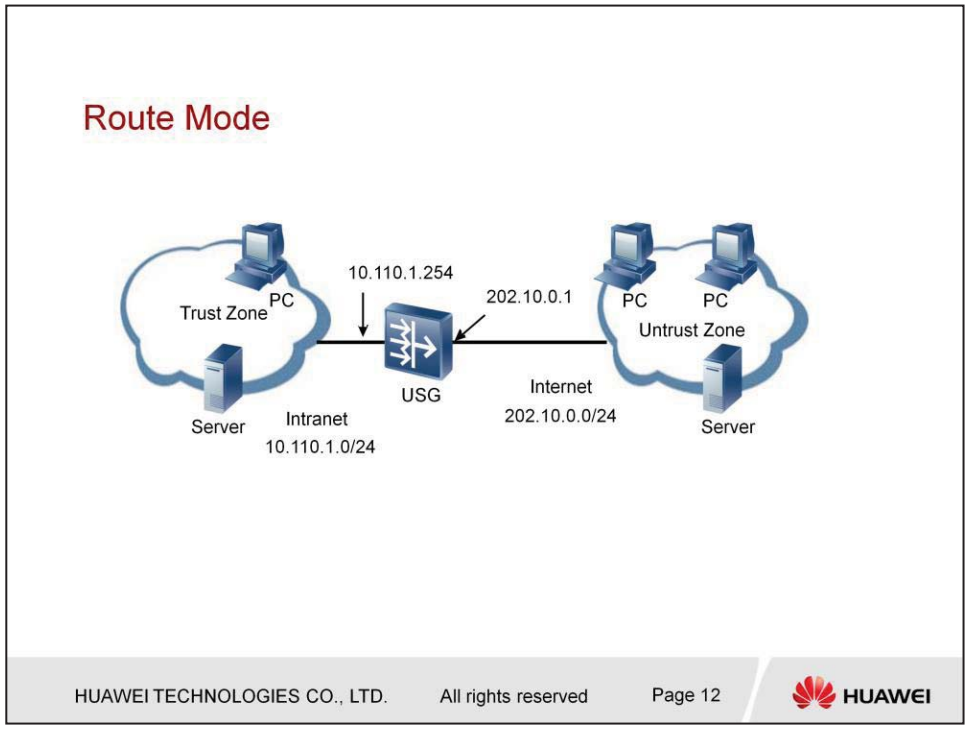


At present, the Eudemon can work in three modes: route mode, transparent mode, and composite mode.

If the firewall is connected to external networks through the interface that has an IP address, the firewall will work in route mode.

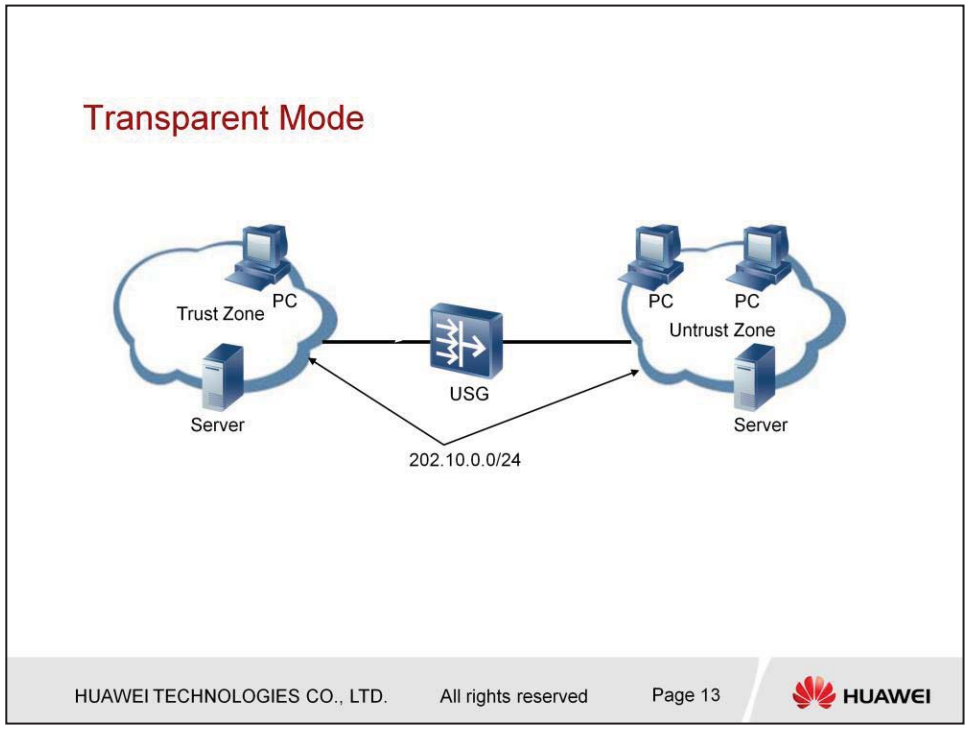
If the firewall is connected to external networks through the interface that has no IP address, the firewall will work in transparent mode.

If the firewall has both interfaces with IP addresses and interfaces without IP addresses, the firewall will work in composite mode.



When the USG is located between the internal network and the external network, you need to configure the interfaces, through which the firewall is connected with the internal network, the external network and DMZ, with IP addresses on different network segments. In this case, the firewall serves as a router.

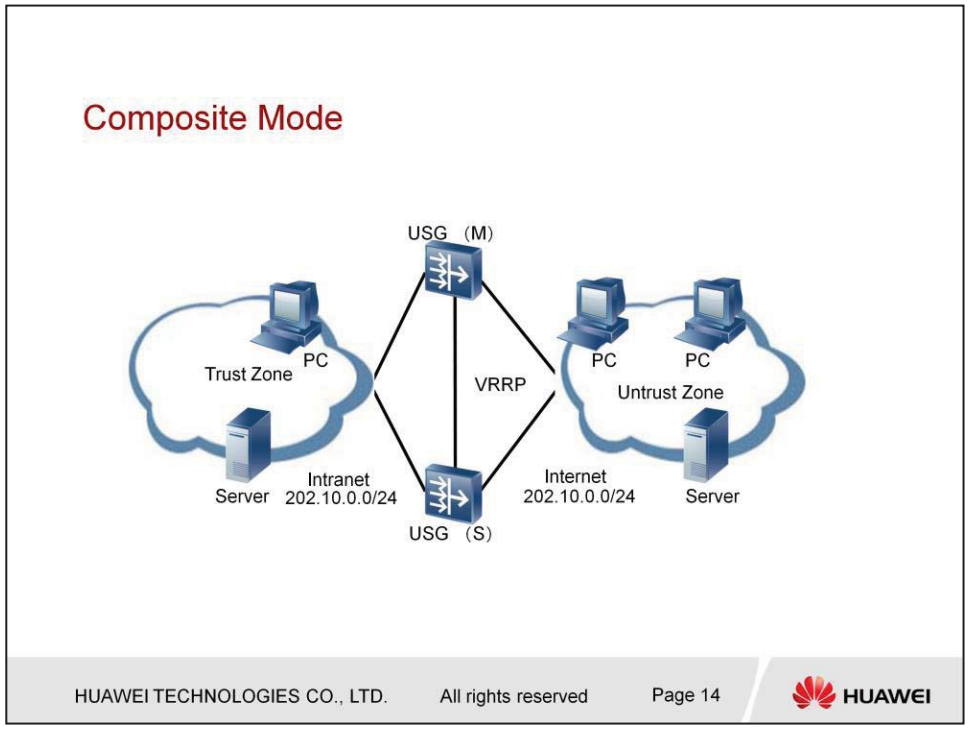
When working in route mode, the firewall can complete ACL packet filtering, ASPF dynamic filtering and NAT. However, the network topology needs to be changed, for example, internal network users need to change their gateways and routers' routing configurations, which can be troublesome. So it is necessary to weigh the advantages and disadvantages in using route mode.



If the USG firewall works in transparent mode, you do not need to change the network topology. In this case, the firewall is completely transparent to users in subnets and routers. That is, users are not fully aware of the existence of the firewall.

In transparent mode, you only need to place the Eudemon in the network like placing a bridge without the need to modify existing configuration. Similar to route mode, IP packets also need to be filtered and checked in transparent mode (while during this procedure the source or destination addresses in IP packets do not change), and internal users can be protected by the firewall.

The address on the firewall indicates, the interfaces on the both side of the firewall are in the same network just like switch. Firewall interfaces don't need to configure IP address.



If there are both interfaces working in route mode (such interfaces have IP addresses) and interfaces working in transparent mode (such interfaces have no IP address) in the USG, the firewall is working in composite mode.

Composite mode is applied in the case of dual-system hot backup with transparent mode. The interface on which VRRP is enabled needs to be configured with an IP address, and other interfaces do not.

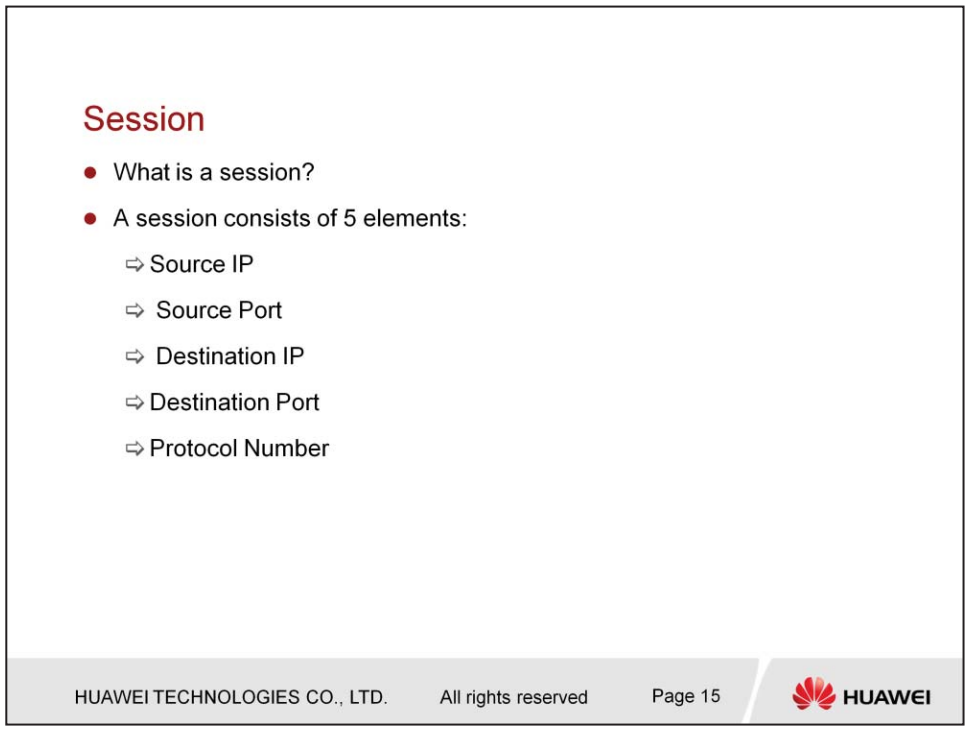
Master and backup USG firewalls are connected with the internal network through interfaces in the Trust zone, while they are connected with the external network through interfaces in the Untrust zone.

In addition, master and backup Eudemon firewalls:

Connect each other through hub or LAN switch.

Perform hot standby through VRRP.

Note that the internal network and the external network must reside in the same subnet.



Sessions have 5 items: source IP and source port, destination IP and destination port, and protocol number.

If it supports the virtual firewall feature, it will have another parameter, VPN ID.

Session (Cont.)

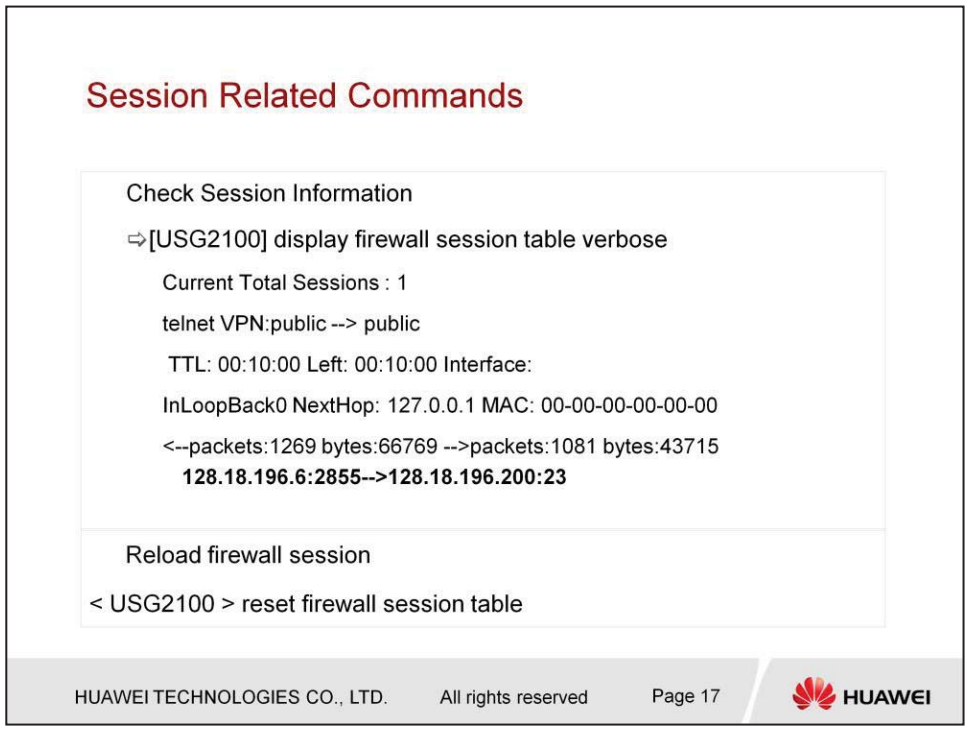
- Why does a firewall need to maintain session information?

Criteria		Operation
If the packet's 5 elements match the session table		Forward the packet
If the packet's 5 elements do not match the session table	Firewall Inter-zone ACL permit	Forward the packet, and establish a new session entry
	Firewall Inter-zone ACL deny	Abort the packet

HUAWEI TECHNOLOGIES CO., LTD. All rights reserved

Because the USG is a stateful firewall, a session entry will be created on the USG for each dynamically generated session in the security zone. Only the subsequent packets and reply packets that match with the session entry can pass through the Eudemon. Therefore, the inbound path and the outbound path of the same session must be consistent.

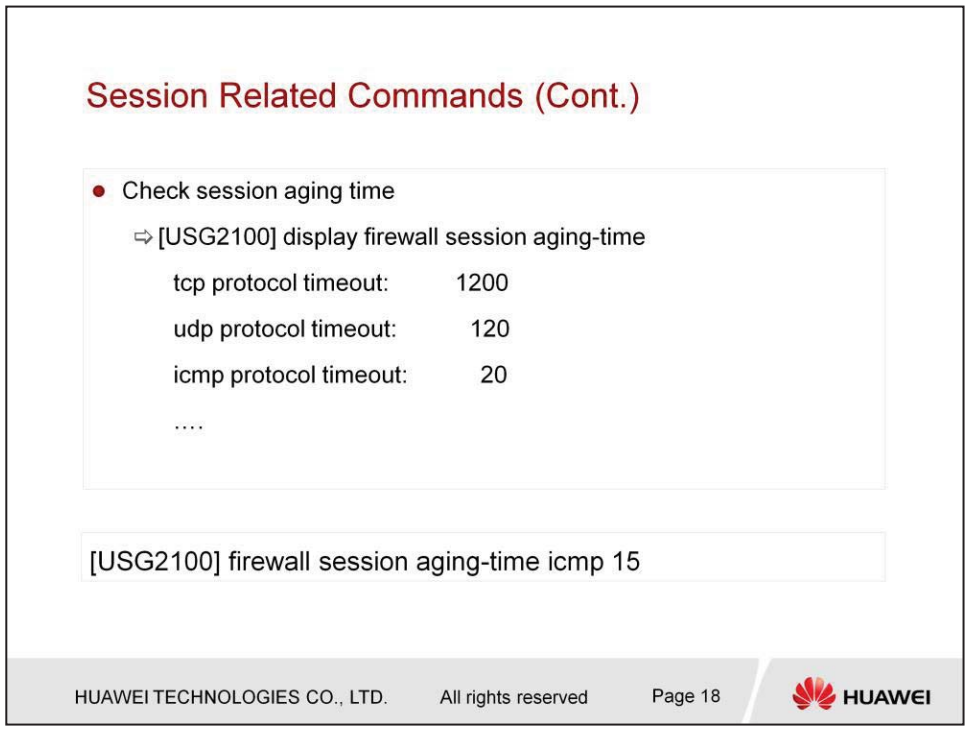
When compared with the ACL filter, the performance is improved.



display firewall session table [verbose].

Using the display firewall session table command, you can display the session tables.

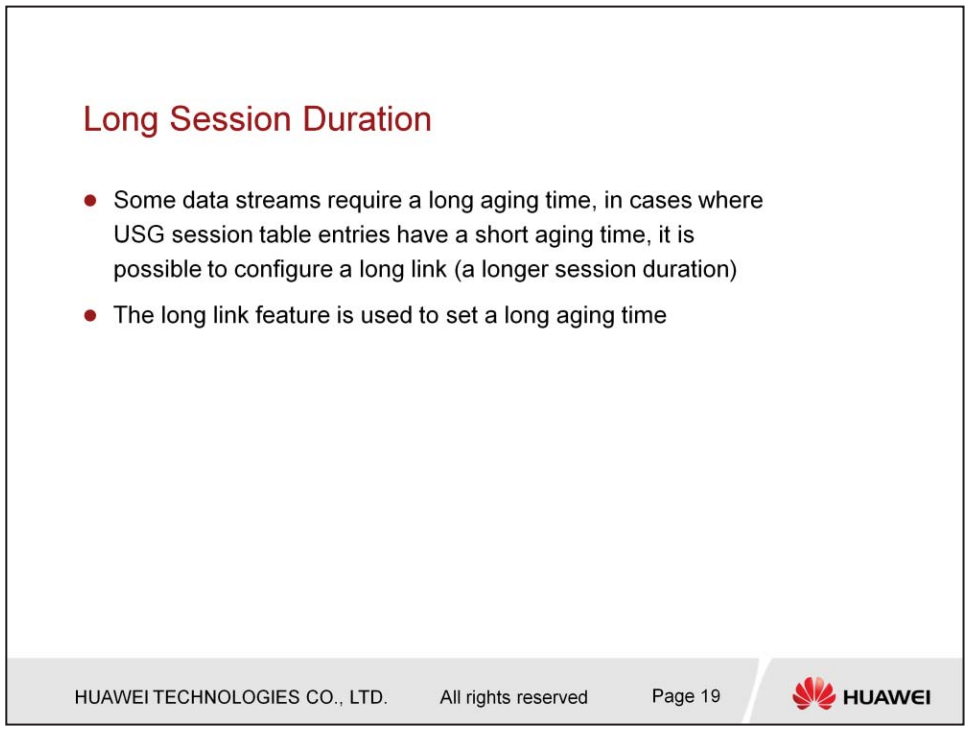
verbose: displays the details about the session table.



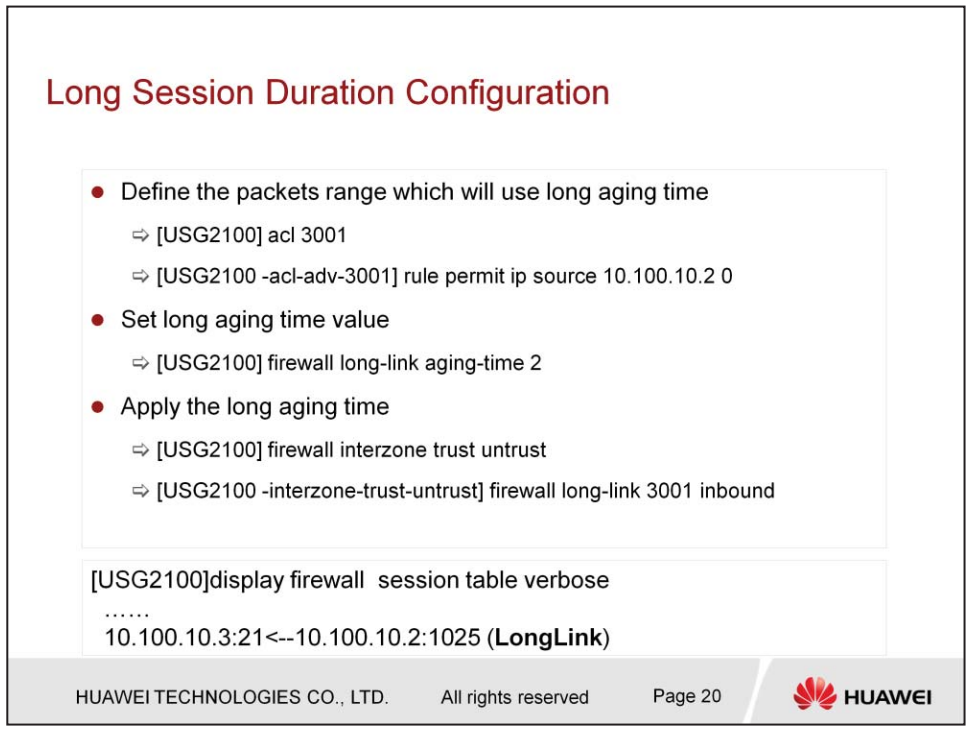
The session table has the aging time that is applicable to all the records in the table. The system deletes the records that are not updated within the aging time.

By default, the aging time for SYN, FIN/RST, TCP and UDP session tables is 5 seconds, 10 seconds, 1200 seconds and 120 seconds respectively.

Using the display firewall session aging-time command, you can view the aging time of the firewall session tables.



Data steams that need a long aging time are defined by ACL rules. The aging time of such data streams is beyond the control of the global aging time of the session table.



firewall long-link aging-time aging-time.

aging-time: specifies the value of the long link aging time. The value ranges from 1 to 480 hours. By default, the value of long link aging time is 168.

Using the firewall long-link aging-time command, you can set the aging time of long links.

long-link acl-number { inbound | outbound }

acl-number: specifies an ACL rule number in the range of 3000 to 3999.

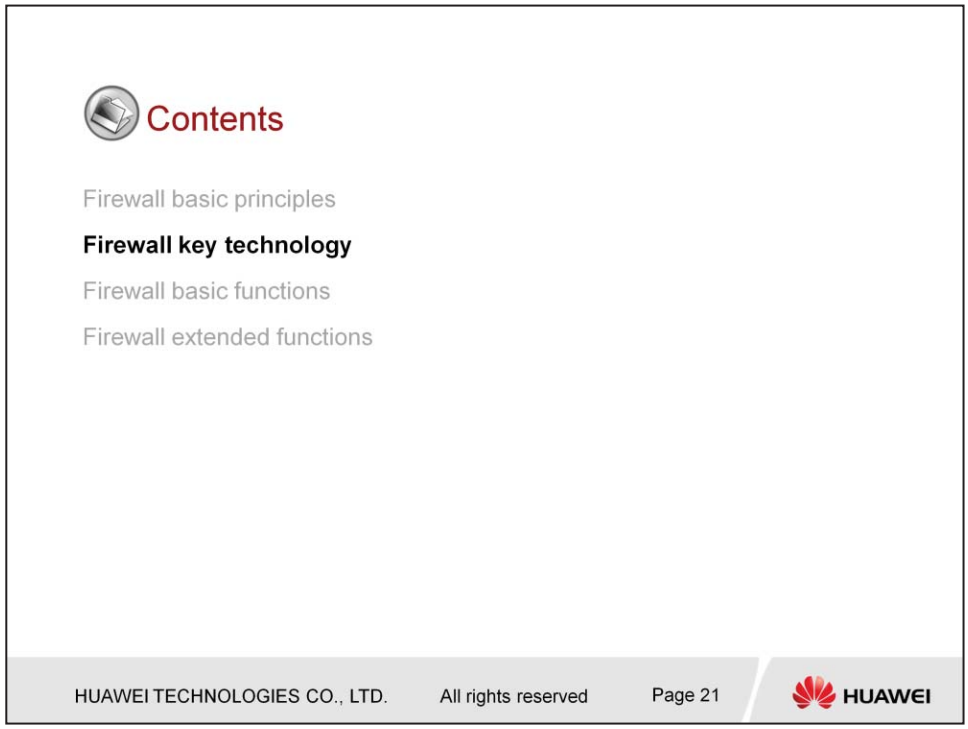
inbound: enables long link in incoming direction between two zones.

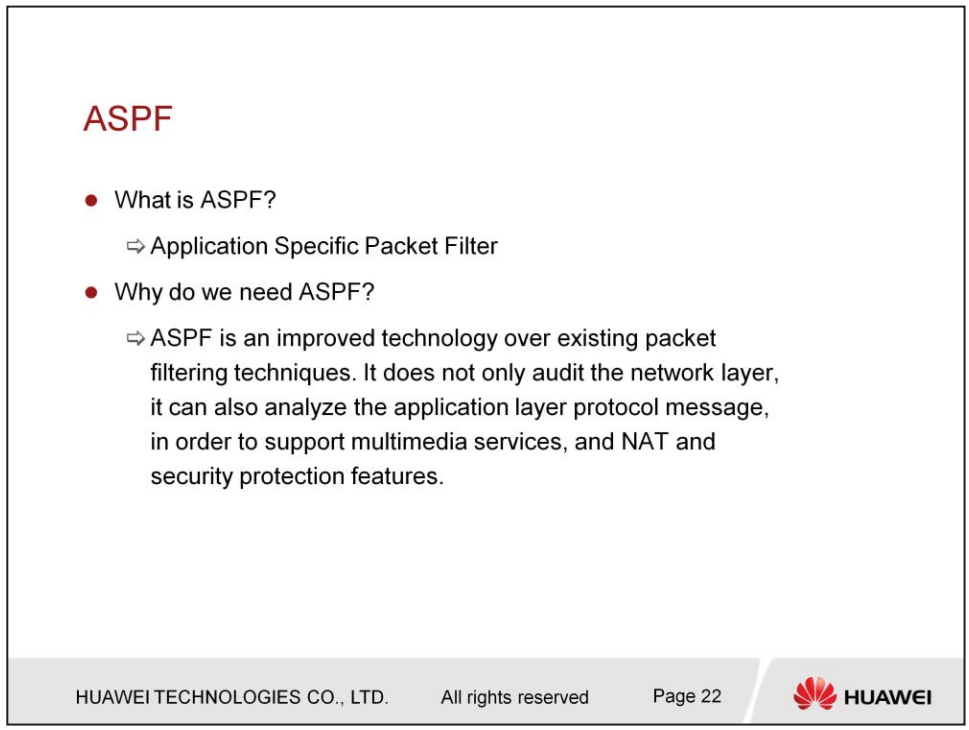
outbound: enables long link in outgoing direction between two zones.

Using the firewall long-link command, you can enable the long link function on the firewall.

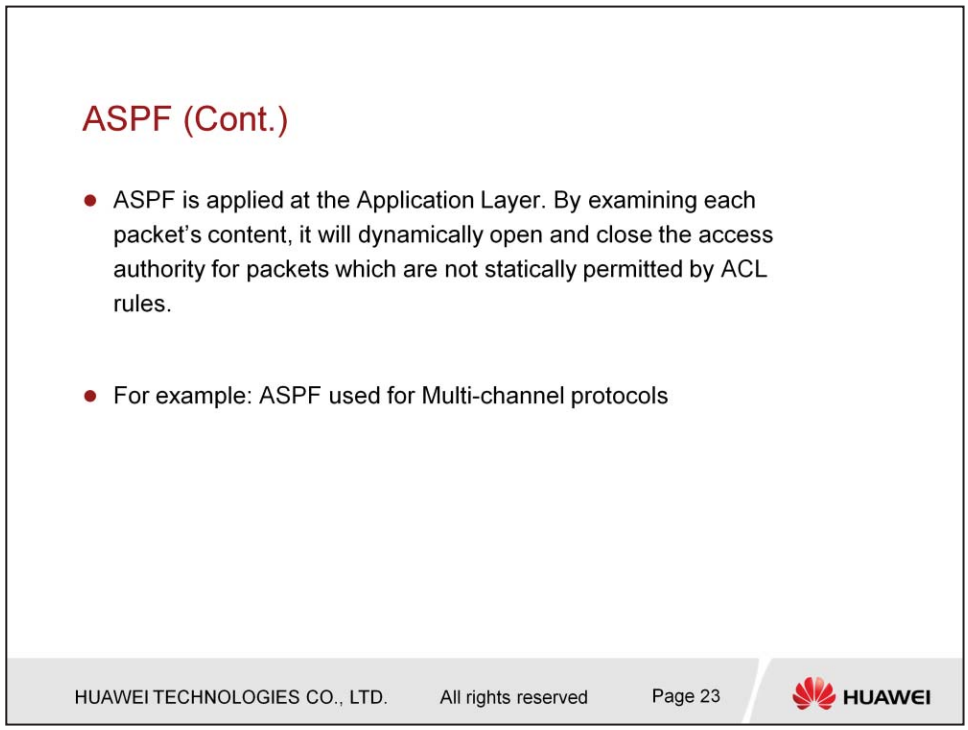
You can apply ACL rules to the incoming, outgoing or both direction packets between two zones. The applied ACL rules in two direction can be different from each other.

It is recommended not to apply an ACL rule with a large address range. otherwise, the firewall performance will be impacted.

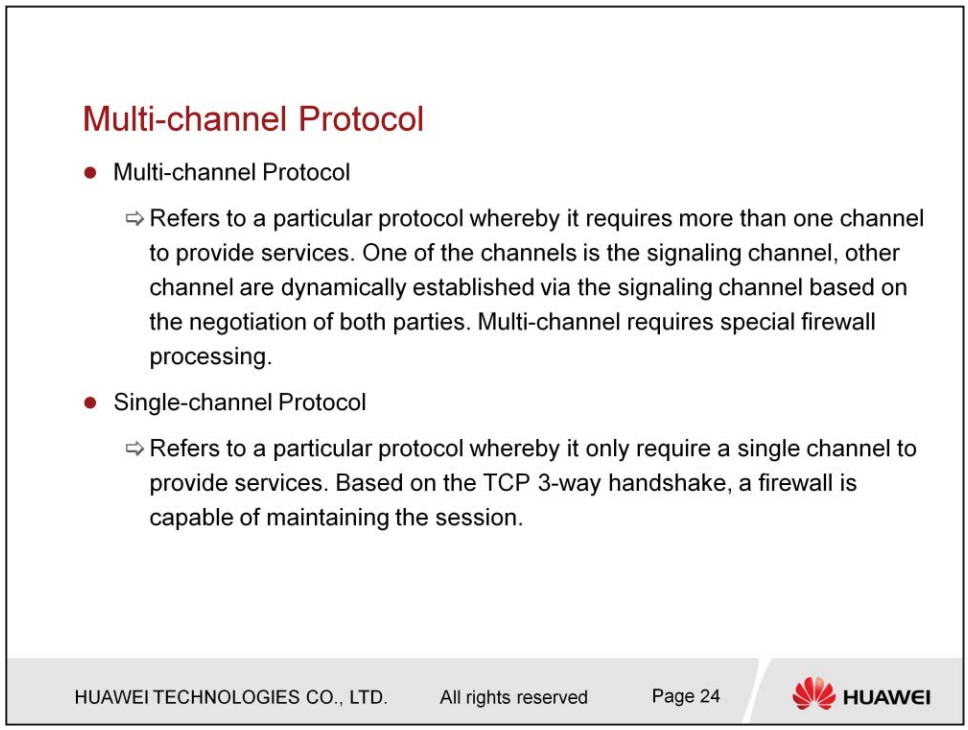

Contents
Firewall basic principles
Firewall key technology
Firewall basic functions
Firewall extended functions
HUAWEI TECHNOLOGIES CO., LTD.
All rights reserved
Page 21
HUAWEI



ACL-based packet-filtering will continue to work on Layer 2 and Layer 3, to prevent unauthorized access.

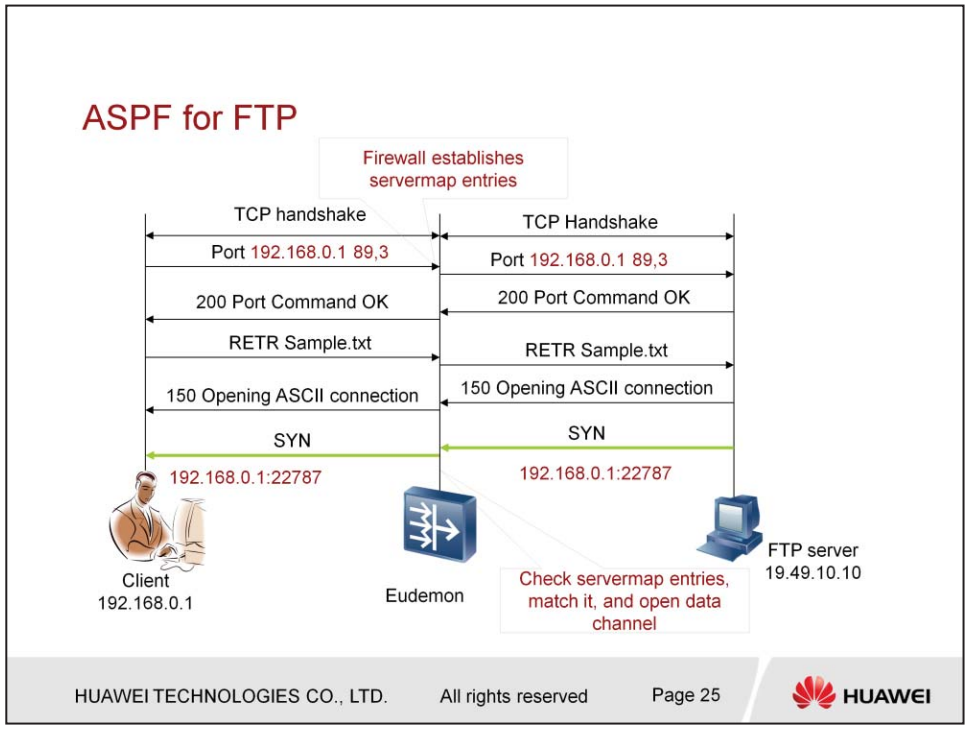


It supports protocols like: H.323, MGCP, SIP, H248, RTSP, HWCC (Huawei Conference control Protocol), ICMP, FTP, DNS, PPTP, NBT, ILS, HTTP, SMTP etc.



Multi-channel protocol: A multi-channel protocol establishes more than one channel for a user and transfers control messages and user data through different channels. FTP and RTSP are examples of multi-channel protocols.

Single-channel protocol: A single-channel protocol establishes only one channel to exchange both control messages and data for a user. SMTP is an example of a single-channel protocol.



FTP is a multi-channel protocol.

1.After the TCP three handshake, a TCP connection is established, input username and password, login to ftp server.

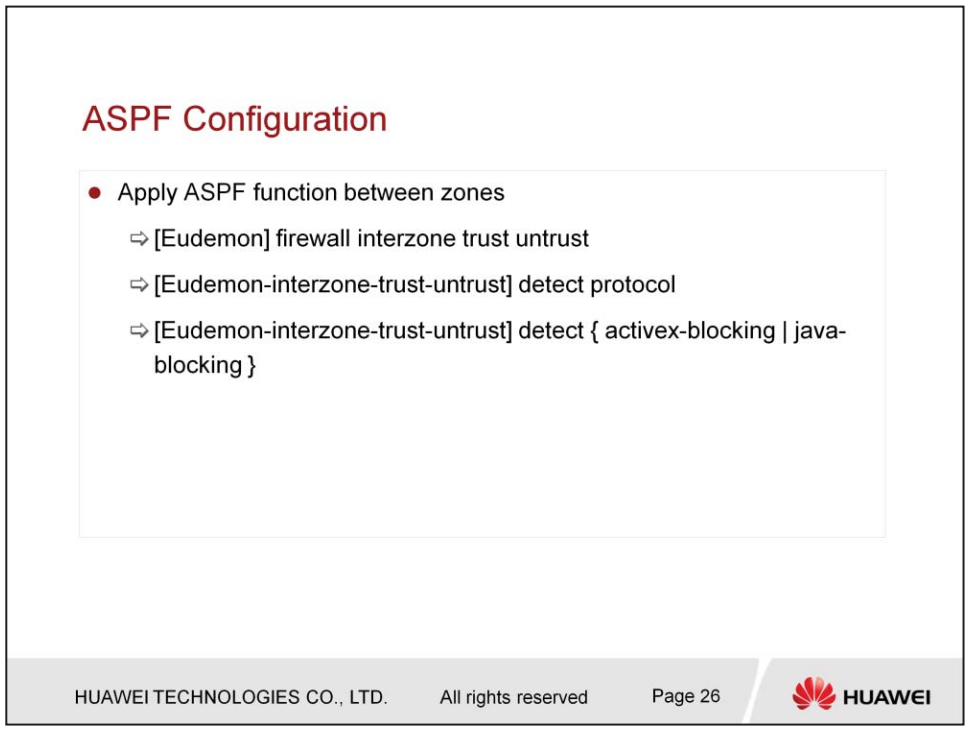
2.Client wants to download a file from server, use port command to notify server of the client's IP address and port information.

3.Server ACK.(200 port command ok).

4.Client requests the file name information (RETR Sample.txt).

5.Server initial the data connection (SYN).

Notes: the calculation method of port number is 89*256+3=22787.

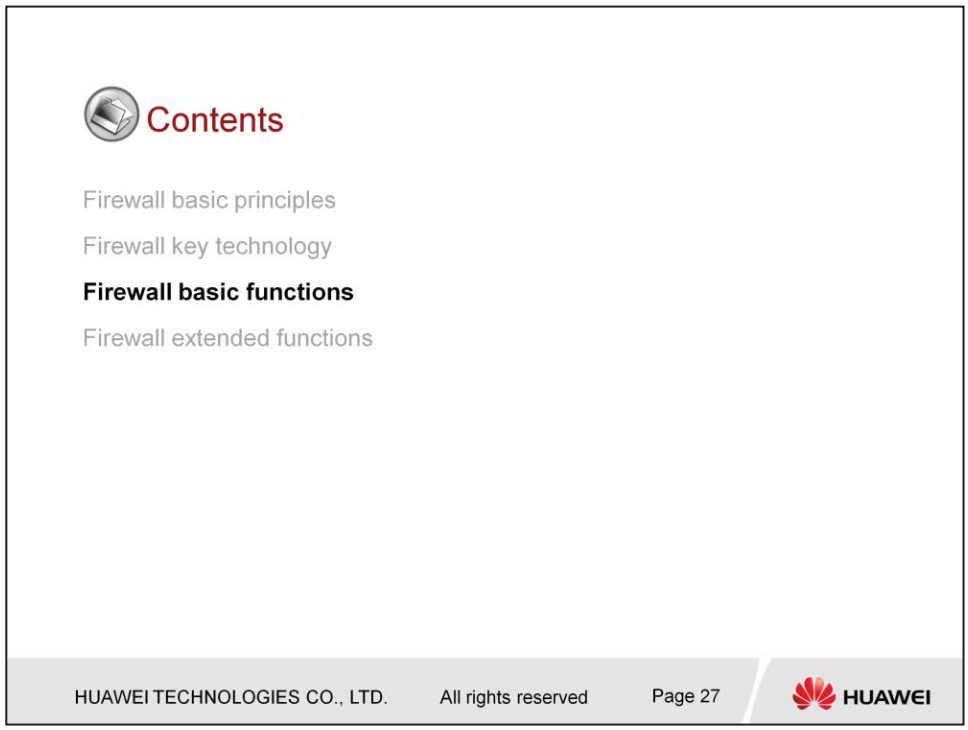


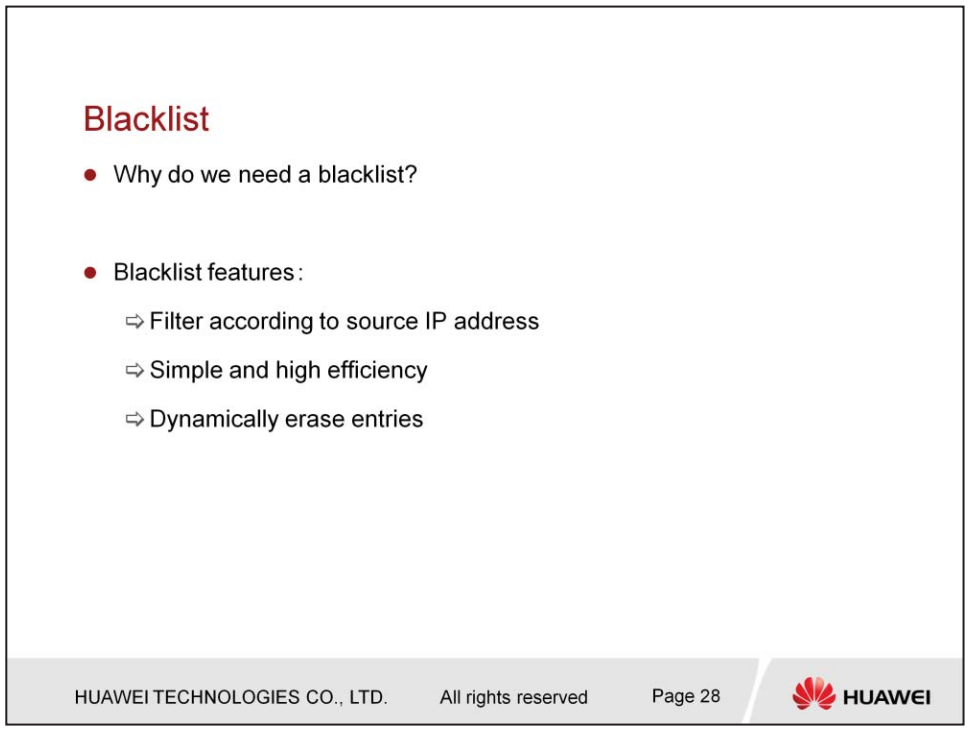
detect protocol [acl-number { inbound | outbound }]

protocol: refers to names of protocols that ASPF supports, such as FTP, HTTP, H323, HWCC, MSN, NETBIOS, PPTP, RTSP. protocol can even be all, activex-blocking or java-blocking which refers to the applying of the ACL for all ASPF detection, activex-blocking detection and java-blocking detection respectively.

inbound: filters inbound packets.

outbound: filters outbound packets.


Contents
Firewall basic principles
Firewall key technology
Firewall basic functions
Firewall extended functions
HUAWEI TECHNOLOGIES CO., LTD.
All rights reserved
Page 27
HUAWEI



Blacklist is one of the security features of the firewall. The most important feature of a blacklist is that it can be added dynamically by the firewall module. Compared with ACL-based packet filter, the zones for blacklist to match are much simpler, so it can filter packets at high speed, which effectively shields the packets sent from the specific IP address.

You can create black list entries in two ways:

Create blacklist entries through command lines

Automatically create black list entries through attack defense modules or the IDS intrusion detect modules

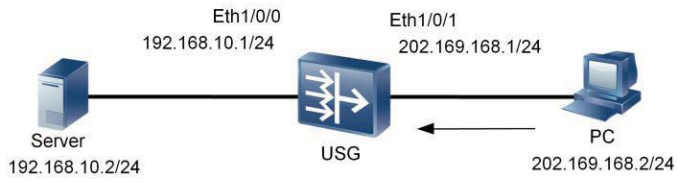
So far, the following firewall modules can insert entries into the blacklist:

Attack defense module

IDS intrusion detect module

When firewall discovers the attack attempt of a specific IP address based on the packet action, it can automatically modify its blacklist to filter all the packets sent from the specific address.

Static Blacklist Configuration



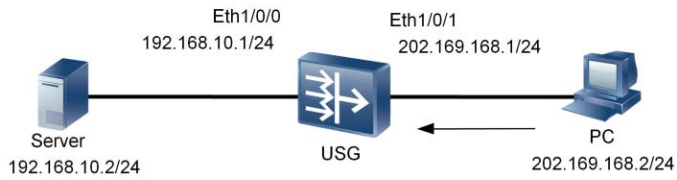

```
[USG2100] firewall blacklist item 202.169.168.2 timeout 100
[USG2100] firewall blacklist enable
```

HUAWEI TECHNOLOGIES CO., LTD. All rights reserved

Time out 100 minutes.

If timeout is not specified, timeout is permanent.

Dynamic Blacklist Configuration




```
[USG2100] firewall defend ip-sweep enable
[USG2100] firewall defend ip-sweep max-rate 1000
[USG2100] firewall defend ip-sweep blacklist-timeout 20
[USG2100] firewall blacklist enable
```

HUAWEI TECHNOLOGIES CO., LTD. All rights reserved

[Eudemon] firewall defend ip-sweep enable

Enables the IP sweeping attack defense function.

[Eudemon] firewall defend ip-sweep max-rate 1000

Indicates the threshold of the address sweeping rate.

[Eudemon] firewall defend ip-sweep blacklist-timeout 20

Indicates the timeout period of the source IP address of the IP sweeping attacker in the blacklist.

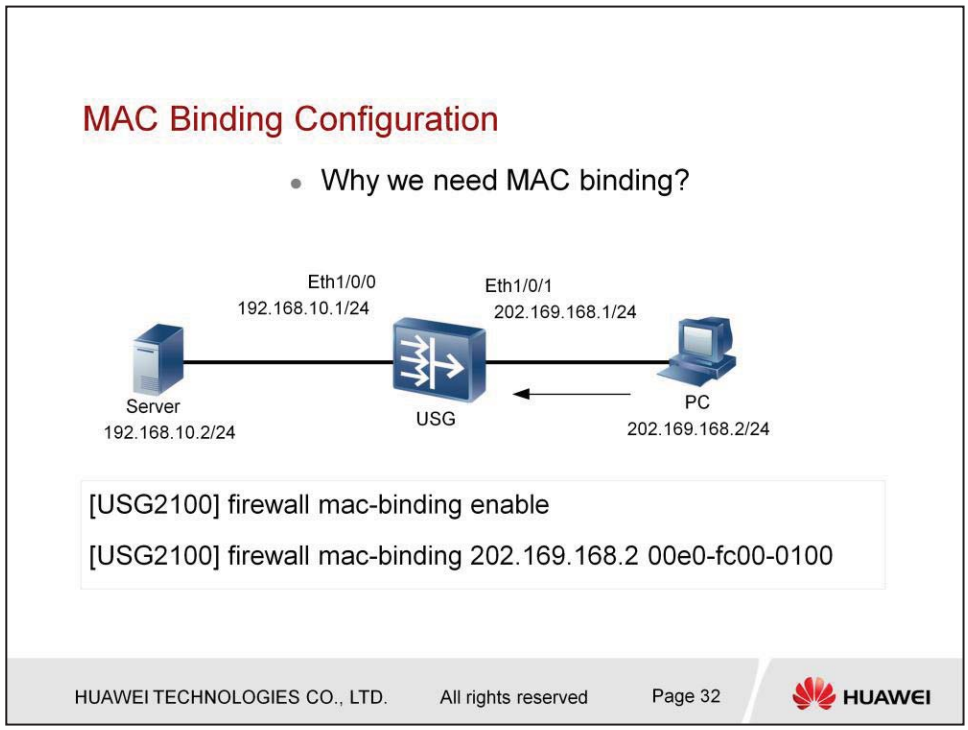
[Eudemon] firewall blacklist enable

Packets that match with entries in the blacklist are filtered only when the blacklist is enabled. Otherwise, those packets are not discarded by the firewall, even if source IP address of packets are in the blacklist.

Checking Blacklist

```
[USG2100] display firewall blacklist item
Total:1        Manual:1          IP Sweep:0        Port Scan:0
 IDS:0         Login Failed:0    PreAuthed:0       Get Flood:0
 tcp-illeage-session:0           Unknown:0
 IP              Reason     InsertTime    AgeTime  Vpn-instance
 --------------------------------------------------------------------------------
202.169.168.2    Manual     2009/05/12 17:47:35  Permanent
```

HUAWEI TECHNOLOGIES CO., LTD. All rights reserved



firewall mac-binding { enable | ip-address mac-address }

ip-address: refers to IP addresses of an address binding pair.

mac-address: refers to MAC addresses of an address binding pair.

Using the firewall mac-binding command, you can add or delete entries of address binding, or enable or disable address binding.

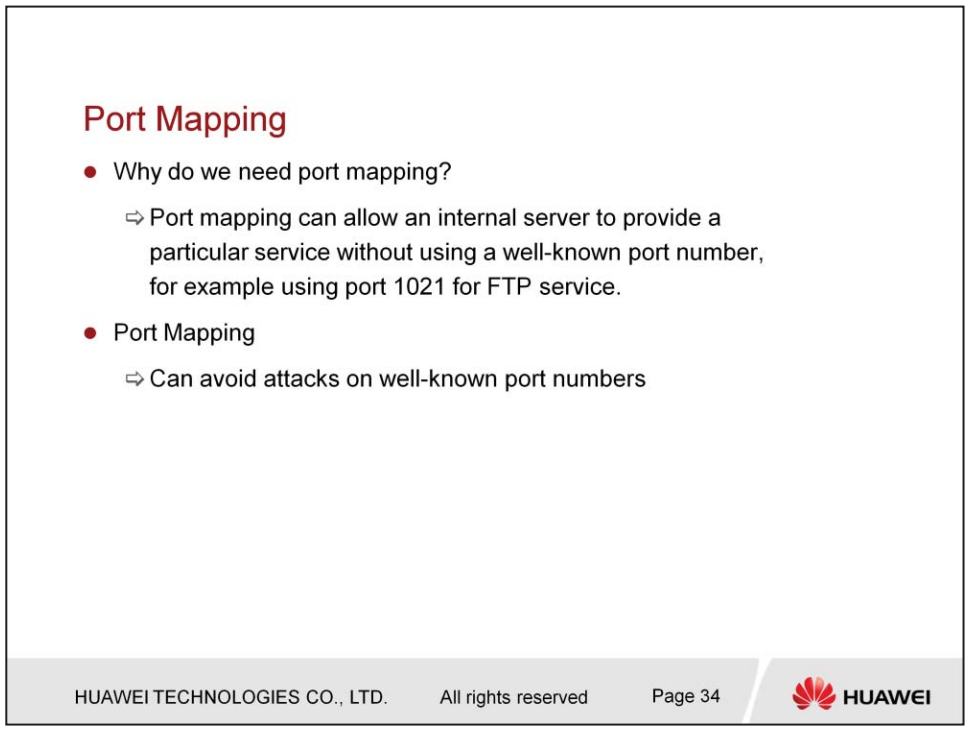
The IP address serving as an entry is added to address binding. If the same IP address is added to the address binding repeatedly by using the firewall mac-binding command, the latter entries overwrite the previous ones. One MAC address can associate with multiple IP addresses.

Adding or deleting an entry to or from the address binding is independent of the running status of the address binding. In other words, whether the address binding is enabled or not, you can add or delete an entry to or from the address binding.

Checking MAC Binding

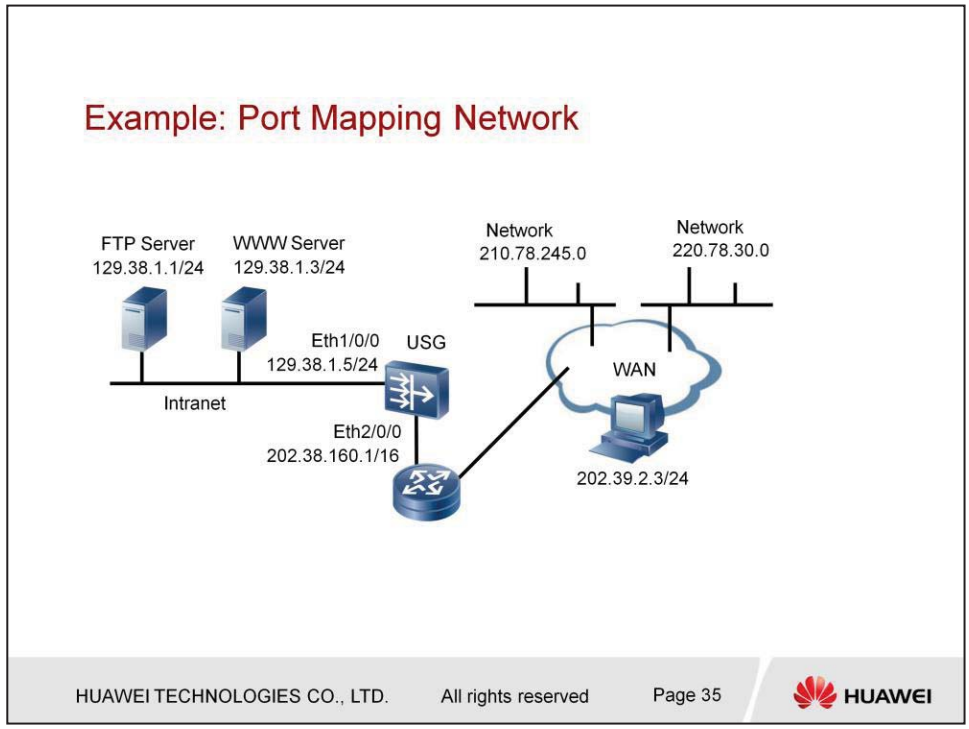
```
[USG2100] display firewall mac-binding item
Firewall Mac-binding items :
Current items : 1
IP ADDRESS      MAC ADDRESS       VPN-INSTANCE
202.169.168.2   00e0-fc00-0100
```

HUAWEI TECHNOLOGIES CO., LTD. All rights reserved



Application layer protocols use well-known port numbers for communication. Port mapping allows users to define a new set of port numbers for different applications. It can also specify the range of hosts who use non-well-known port numbers.

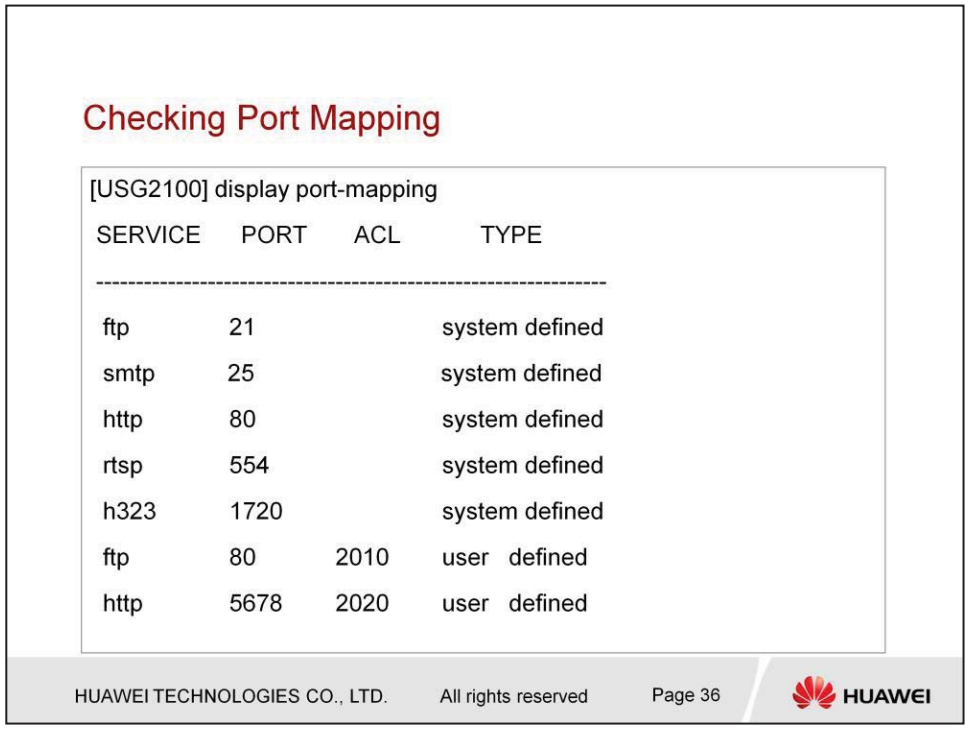
Port mapping applies to service-sensitive services such as Application Specific Packet Filter (ASPF) and Network Address Translation (NAT). For example, ASPF provides Java test for HTTP packets. By default, the port number of HTTP packets is 80. If the user uses another port, for example port 8080, as the port number, ASPF cannot perform the Java test on the packets at port 8080 because it only regards packets at port 80 as HTTP packets. In this case, use port mapping to map port 8080 so that the packets at port 8080 are regarded as HTTP packets. ASPF then recognizes packets at port 8080 as HTTP packets, and performs the Java test.



As shown in the above figure, a company wants to provide WWW and FTP service for users, the requirements are as follows:

Regard the packets destined for 129.38.1.1/24 whose port number is 80 as FTP packets.

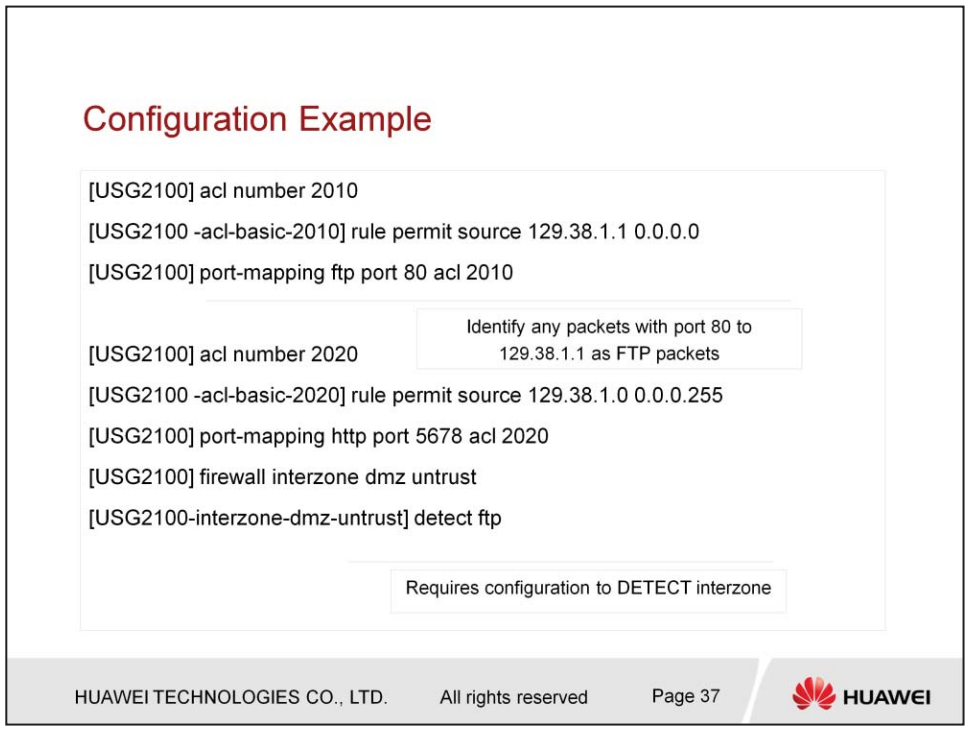
Regard the packets destined for 129.38.1.0/24 whose port number is 5678 as HTTP packets.



display port-mapping [protocol-name | port port-number]

application-name ： Indicates the application type, available applications include FTP, HTTP, H.323, SMTP, RTSP.

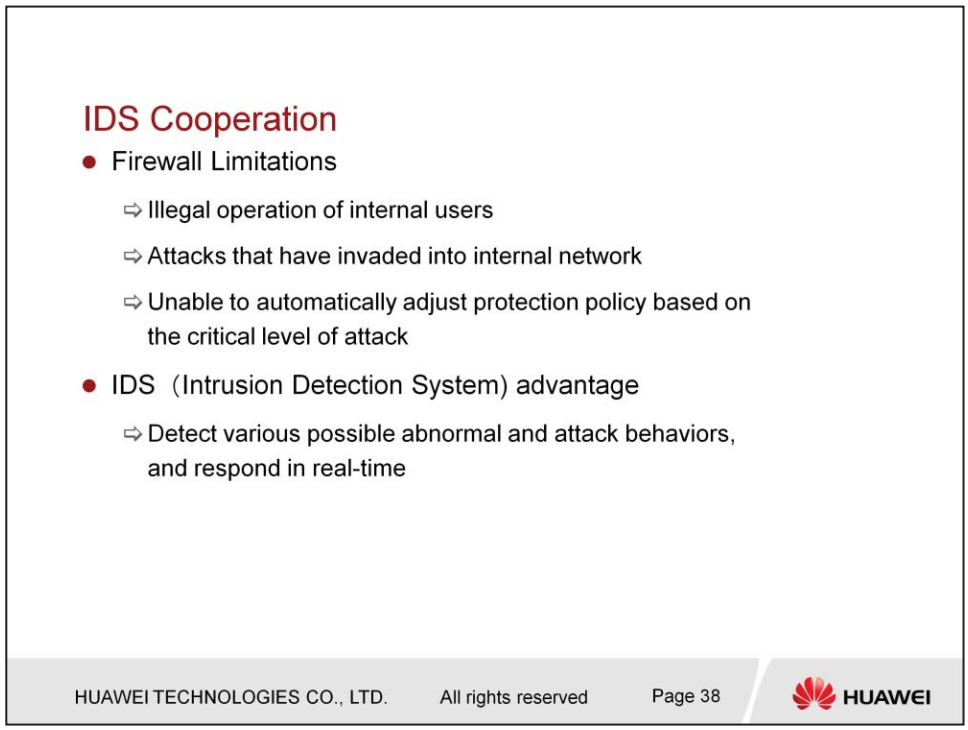
port port-number：the port number.



You cannot delete or modify system-defined port number. Instead, you can redefine it with basic ACL-based host port identification mechanism.

You can use the port-mapping command repeatedly to configure multiple identification ports for a protocol.

A port can be configured with different protocols in different host ranges defined by using the acl command. For instance, using the port-mapping ftp port 80 acl 2010 and port-mapping http port 5678 acl 2020 commands, you can define the packet from port 80 of the host defined by ACL 2010 as an FTP packet and the packet from port 5678 of the host defined based on ACL 2020 as an HTTP packet.



Usually, the USG is mainly used to:

Prevent users or information from entering some restricted sites. Monitor the access channel between the reliable network and unreliable ones to prevent threats from the external network from spreading into the internal network.

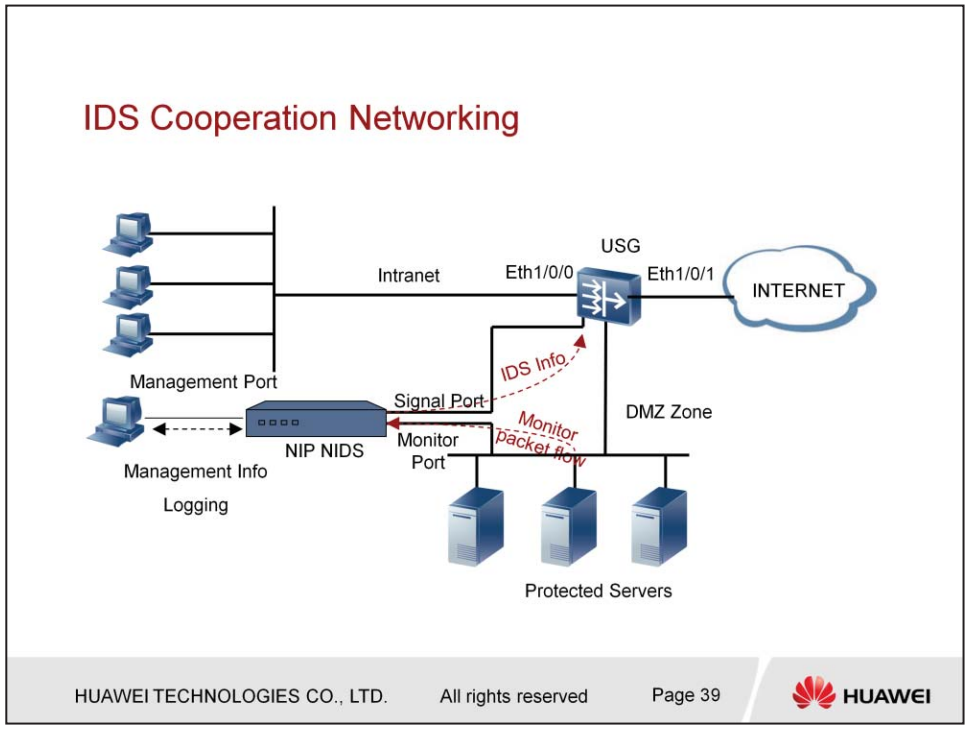
Prevents users or information from leaving the restricted sites. By effectively controlling external users' access to the internal resource, the security of information is guaranteed.

The firewall has a defect: Its detection granularity is rather broad; it cannot perform further analysis and detection on many protocols.

Therefore, the USG firewall opens some ports to link with other security software so as to construct a united security network. That is Intrusion Detection System (IDS) cooperation.

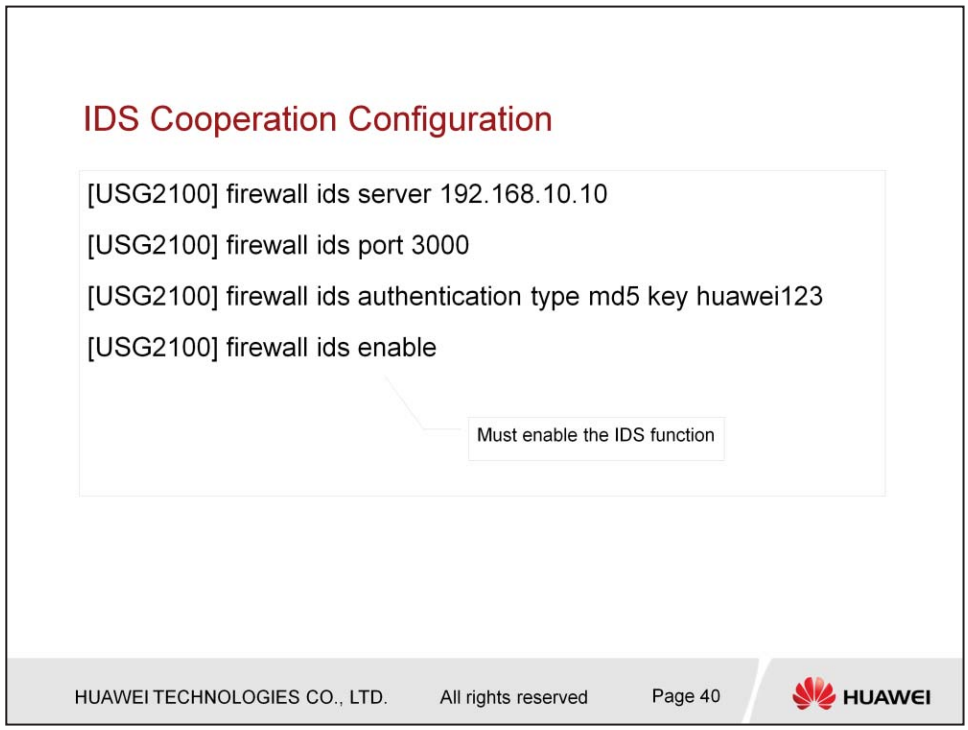
The IDS system in the network is just like a network analyzer installed in the network, that inspects the network transmission. The system knows the latest means of attack, and carefully detects each packet passed through, so doubtful network transmission can be processed in time. Measures to be taken are determined by the specific IDS system and configuration that users used.

Cooperating with the IDS system, the USG can make full use of functions of the IDS software to analyze and detect packets that flow across the network in detail and to probe various possible abnormal and attack behaviors and respond in real-time.



IDS monitors the mirrored traffic of DMZ Zone, and detects the traffic.

If they want to cooperate with the firewall, choose the signal port (for NIP is admin port) to communicate with firewall, add the attack source to blacklist dynamic.



firewall ids server ip-address

An IP address is assigned to the IDS server.

firewall ids port port-number

A port is configured through which the firewall communicates with the external IDS server.

firewall ids authentication type { { md5 | vip } [key key-string] | none }

The packet authentication mode between the firewall and the external IDS server is configured.

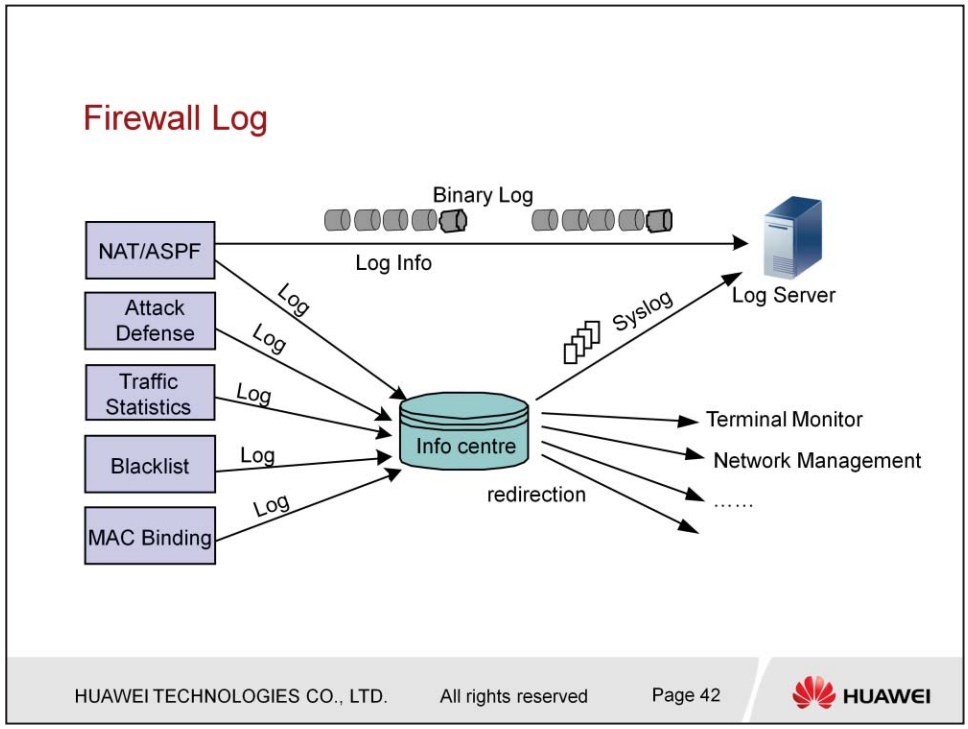
firewall ids enable

IDS cooperation is enabled.

Checking IDS Configuration

```
[USG2100] display firewall ids
Firewall IDS information:
  firewall IDS: enable
  debug flag: off
  server port: 3000
  authentication type: md5
  authentication string: huawei123
  client address 0: 192.168.10.10
```

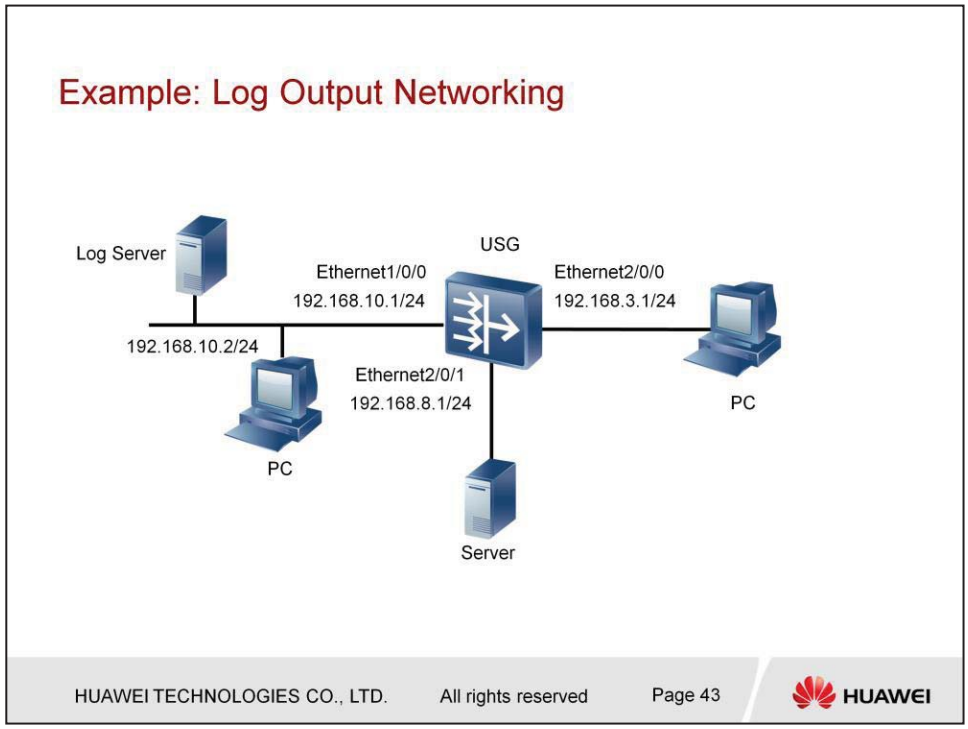
HUAWEI TECHNOLOGIES CO., LTD. All rights reserved



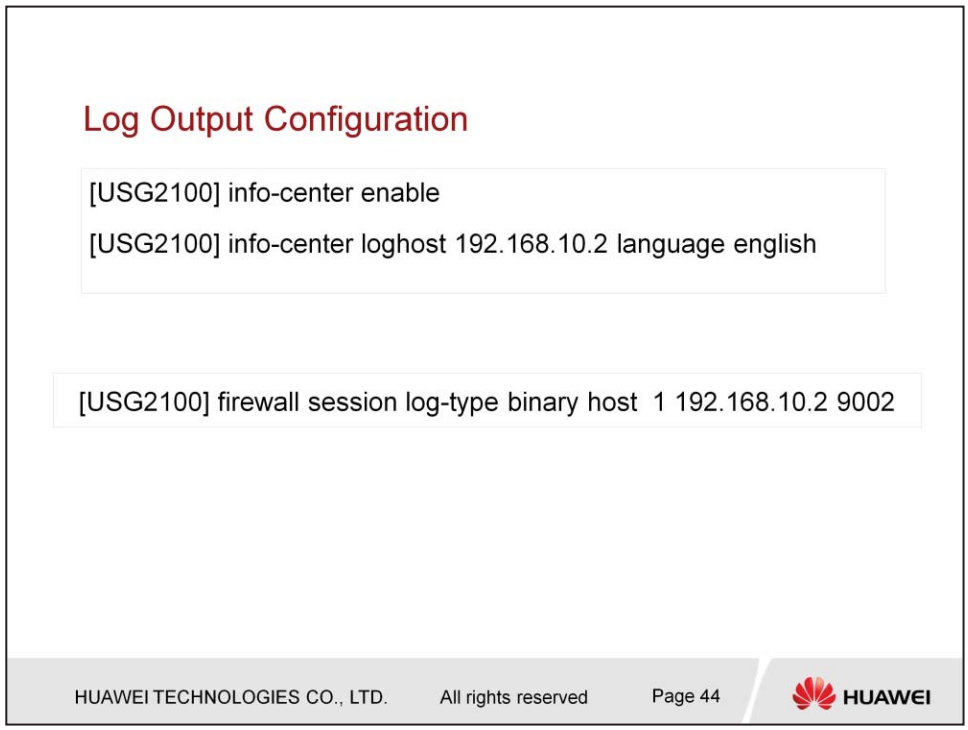
The firewall records its behavior and status, such as carrying out a firewall policy or detecting a network attack, and records the information into the log in real time.

By analyzing the firewall log, the administrator can detect the security hole of the firewall, the user who violates the security policy, and the type of the network attack. The real-time log can detect the ongoing attack.

When you need to record the behavior and the status of the firewall to detect the security hole and detect the attack and intrusion to the network, you can configure the firewall log.



USG Ethernet1/0/0 is connected to the Trust zone，Ethernet2/0/0 is connected to Untrust zone，Ethernet2/0/1 is connected to DMZ zone. The firewall outputs the attack defense log information to the log host.



[USG2100] info-center enable

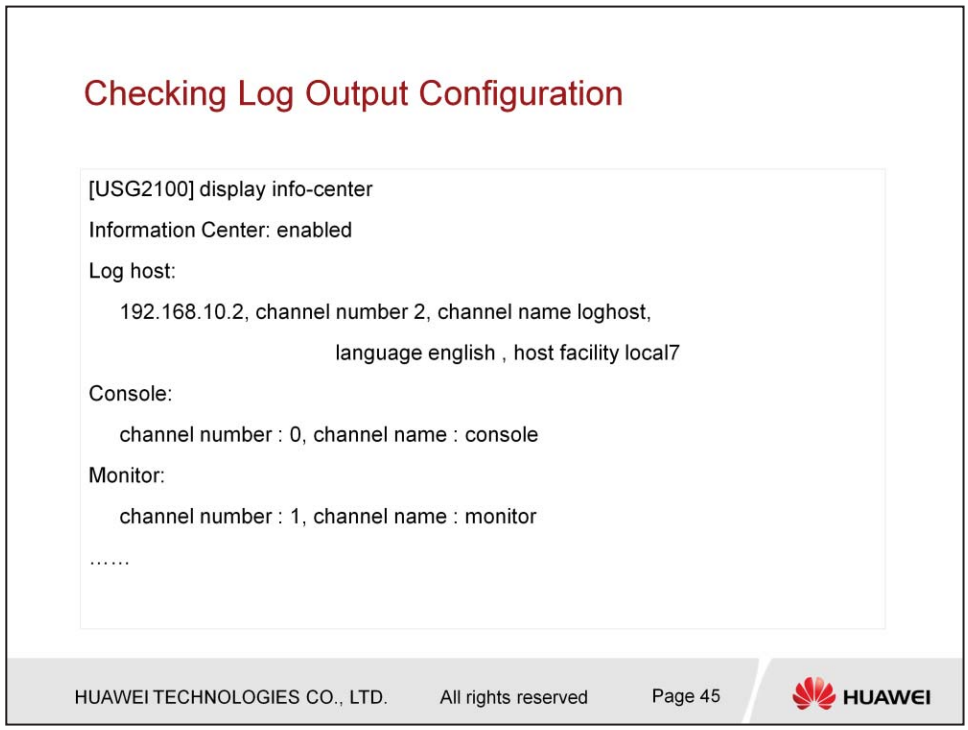
Enable information center.

[USG2100] info-center loghost 192.168.10.2 language english

the log host IP address is 192.168.10.1, the language is English.

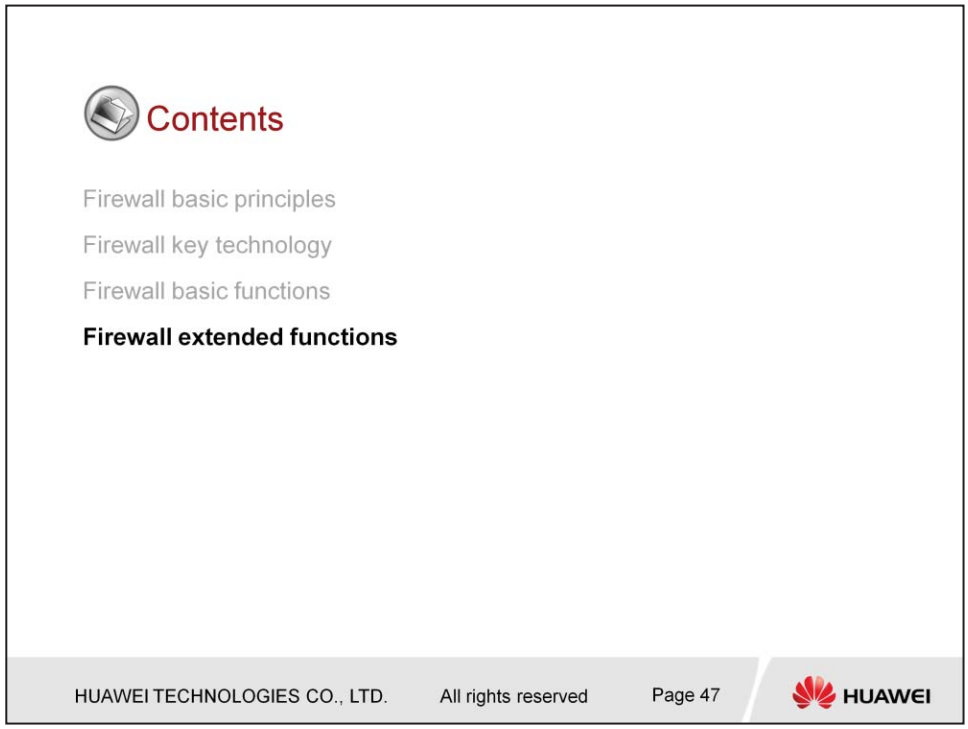
[USG2100] firewall session log-type binary host 1 192.168.10.2 9002

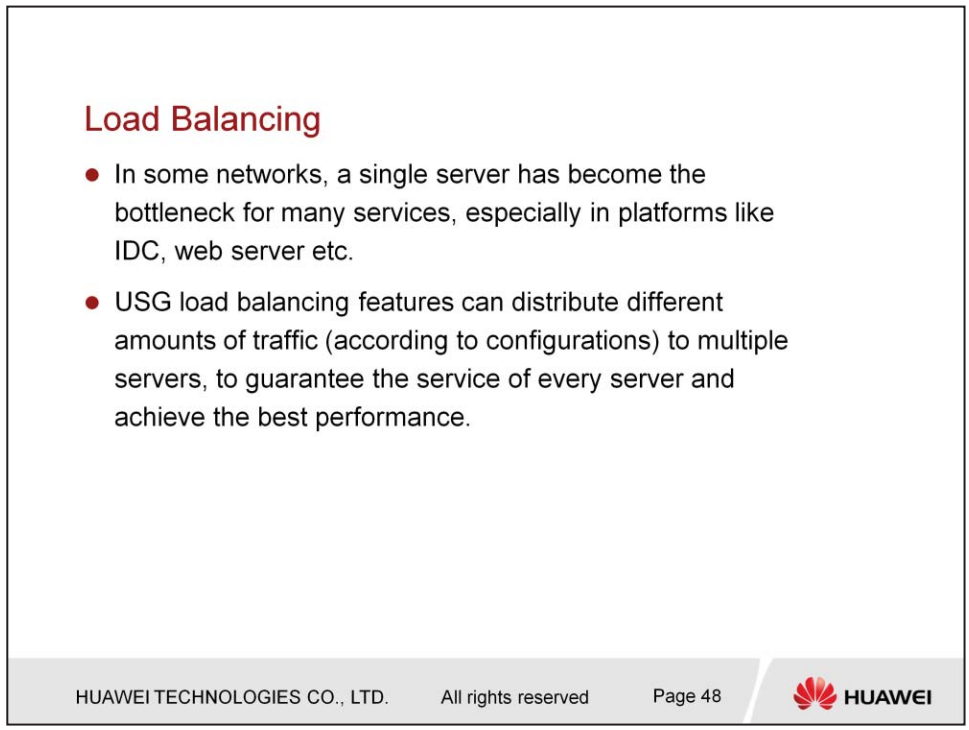
The IP address and receiving port are set for host logging.



[USG2100]display info-center
Information Center:enabled
Log host:
192.168.10.2, channel number 2, channel name loghost,
language english , host facility local7
Console:
channel number : 0, channel name : console
Monitor:
channel number : 1, channel name : monitor
SNMP Agent:
channel number : 5, channel name : snmpagent
Log buffer:
enabled,max buffer size 1024, current buffer size 512,
current messages 23, channel number : 4, channel name : logbuffer
dropped messages 0, overwritten messages 0 Trap buffer:
enabled,max buffer size 1024, current buffer size 256,
current messages 1, channel number:3, channel name:trapbuffer
dropped messages 0, overwritten messages 0
logfile:

channel number : 9, channel name : channel9, language : english
Information timestamp setting:
log - date, trap - date, debug - boot
Sent messages = 23, Received messages = 23


Contents
Firewall basic principles
Firewall key technology
Firewall basic functions
Firewall extended functions
HUAWEI TECHNOLOGIES CO., LTD.
All rights reserved
Page 47
HUAWEI



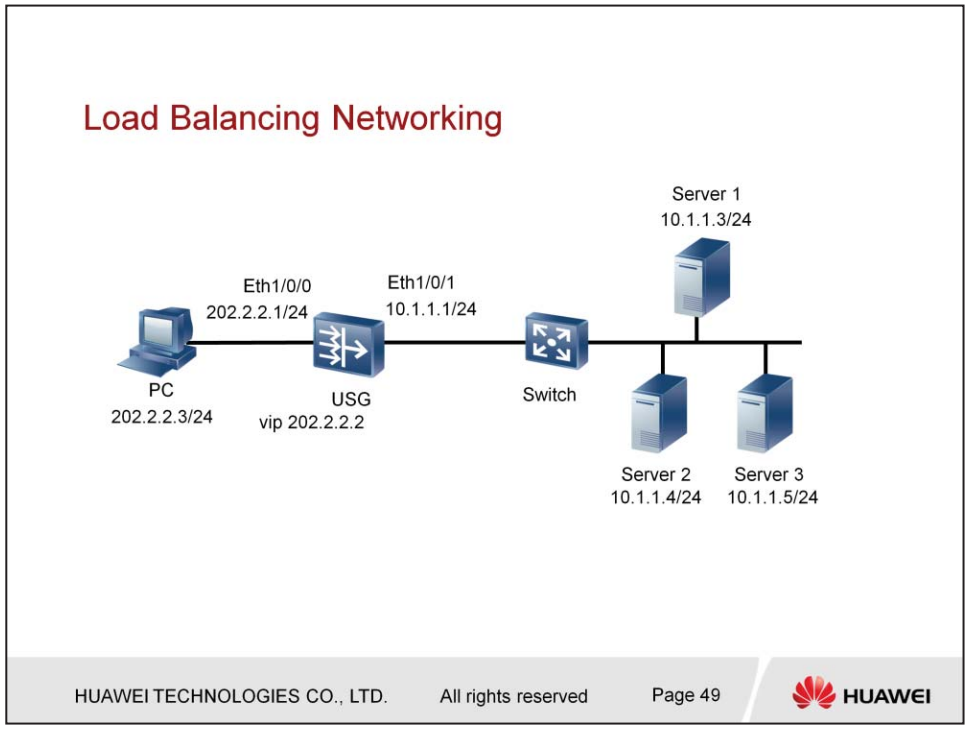
The load balancing of the USG is to distribute the user flow to several servers. In this way, the traffic is shared.

Based on a certain algorithm, the firewall can assign user traffic to different servers. In this way, the processing capability of each server can be fully used and achieve the optimized expandability.

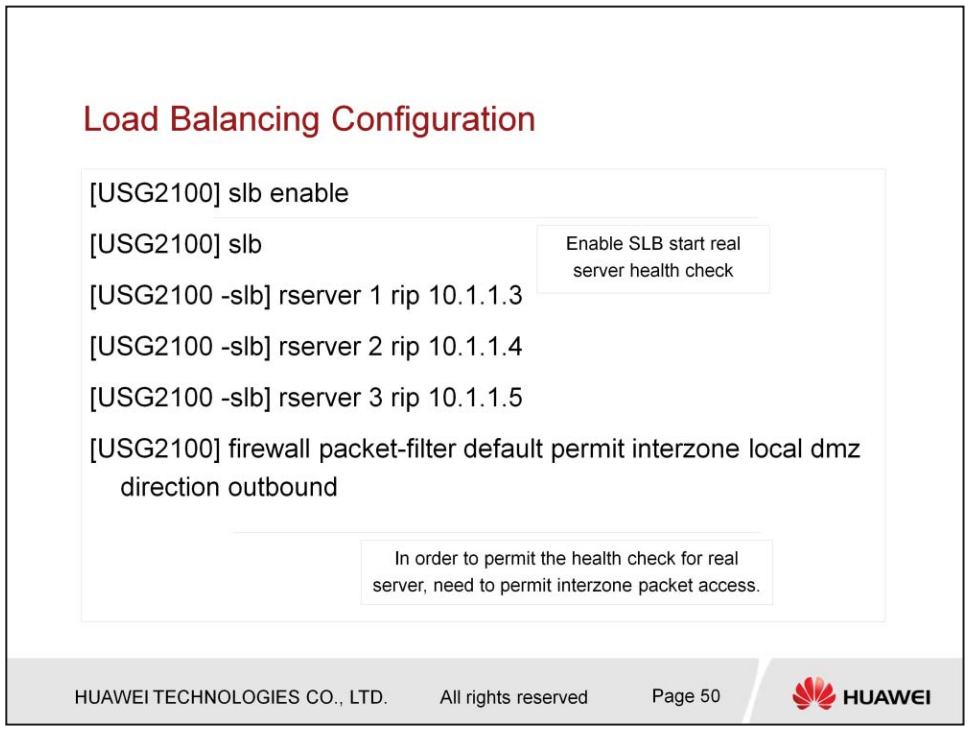
Using the following technologies, Server Load Balancing (SLB) assigns user flow to several servers.

After the firewall is configured with SLB, several servers share a public network IP address (the virtual IP address, VIP). These servers are called real servers. Users must use the virtual IP address to access these servers. Each real server uses different private network addresses (the real IP address, RIP). Multi-layer switches or firewalls assign the access traffic of VIP to each real server based on the pre-configured algorithm.

For the sake of management, a group is used to connect the virtual server and the real server. Group is a logic concept. The firewall uses a group to manage real servers and offers network services.



There are three servers inside DMZ zone, their IP addresses are 10.1.1.3/24, 10.1.1.4/24, 10.1.1.5/24 respectively. The virtual IP address is 202.2.2.2 .



rserver rserver-id [to end-rserver-id] rip ip-address

rserver-id：the real server id.

rip ip-address：the real server IP address.

Load Balancing Configuration (Cont.)

```
[USG2100-slb] group group1
[USG2100-slb-group- group1] metric roundrobin
[USG2100-slb-group- group1] addrserver 1
[USG2100-slb-group- group1] addrserver 2
[USG2100-slb-group- group1] addrserver 3
[USG2100-slb] vserver huawei vip 202.2.2.2 group group1
```

HUAWEI TECHNOLOGIES CO., LTD. All rights reserved

vserver server-name vip ip-address group group-name [tcp | udp] [vport port-number [rport port-number]]

vserver server-name: specifies the name of a virtual server. It is a string in a range of 1 character to 31 characters.

vip ip-address: refers to the IP address of the virtual server.

group group-name: specifies the name of a server group. It is a string in a range of 1 character to 31 characters.

tcp: uses the TCP protocol.

udp: uses the UDP protocol.

vport port-number: specifies the port number of virtual server in a range of 1 to 65535.

rport port-number: specifies the port number of the real server, in a range of 1 to 65535.

Load Balancing Check

```
[USG2100 -slb] display this
slb
  rserver 1 rip 10.1.1.3 weight 32 healthchk
  rserver 2 rip 10.1.1.4 weight 32 healthchk
  rserver 3 rip 10.1.1.5 weight 32 healthchk
 group group1
  metric roundrobin
  addrserver 1
  addrserver 2
  addrserver 3
 vserver huawei vip 202.2.2.2 group group1
```

HUAWEI TECHNOLOGIES CO., LTD. All rights reserved

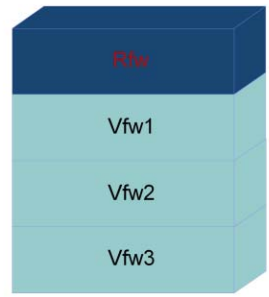
Load Balancing Result

```
[USG2100] display firewall session table
  icmp, (vpn: public -> public) 10.1.1.1:2048-->10.1.1.3:43
  icmp, (vpn: public -> public) 10.1.1.1:2048-->10.1.1.4:43
  icmp, (vpn: public -> public) 10.1.1.1:2048-->10.1.1.5:43
  FTP, (vpn: public -> public) 202.2.2.2:21[10.1.1.3:21]<-+202.2.2.3:1227
  FTP, (vpn: public -> public) 202.2.2.2:21[10.1.1.4:21]<-+202.2.2.3:1229
  FTP, (vpn: public -> public) 202.2.2.2:21[10.1.1.5:21]<-+202.2.2.3:1231
…<output omitted>…
 Current Total Sessions : 8
```

HUAWEI TECHNOLOGIES CO., LTD. All rights reserved

Virtual Firewall

- Establishing a virtual firewall on Eudemon (Vfw), could provide firewall renting services, implement subnet separation and avoid address overlapping problems.
- Every virtual firewall is a combination of a VPN-Instance, security-instance and configuration-instance, able to provide virtual firewall users with a private routing, security & configuration management platform



HUAWEI TECHNOLOGIES CO., LTD. All rights reserved

In recent years, small private networks are increasing. Such networks usually belong to small-scale enterprises, which have the following features:

High requirement on security

Cannot afford a dedicated security device

To meet the requirements of such customers, the network operator can adopt the USG firewall multi-instance solution of Huawei to logically divide one USG firewall into multiple VPN instances, so as to provide independent security services for multiple small private networks.

Each virtual firewall is a combination of one VPN instance, one security instance and one configuration instance. It can provide proprietary route forwarding plane, security service plane and configuration management plane for virtual firewall users.

VPN Instance

The VPN instance provides isolated VPN routes for the virtual firewall users. A security instance corresponds with one virtual firewall.

VPN routes support forwarding the packets from the interfaces bound with the VPN instance.

Security Instance

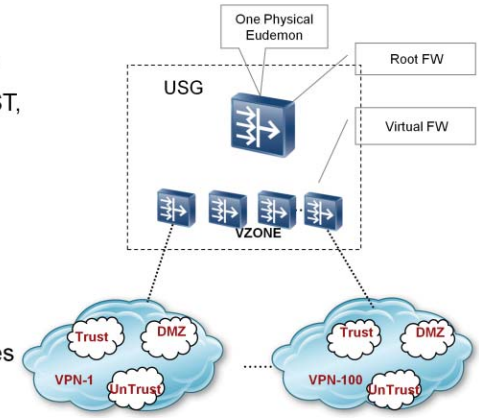
The security instance provides isolated security service for the virtual firewall users. A security instance corresponds with one virtual firewall.

Configuration Instance

The configuration instance provides isolated configuration management planes for virtual firewall users. A configuration instance corresponds with one virtual firewall. After virtual firewall users log on to the firewall, they have rights to manage and maintain the private VPN routes and security instances.

Virtual Firewall

- USG virtual firewall support.
  - Every Virtual firewall has an independent 5 (Local, TRUST, UNTRUST, DMZ, VZONE) zones and flexible interface binding.
  - Independent resources provide private security services, multi NAT instances and multi VPN instances



HUAWEI TECHNOLOGIES CO., LTD. All rights reserved

The security instance can provide private security services, including:

Address binding

Blacklist

Address translation

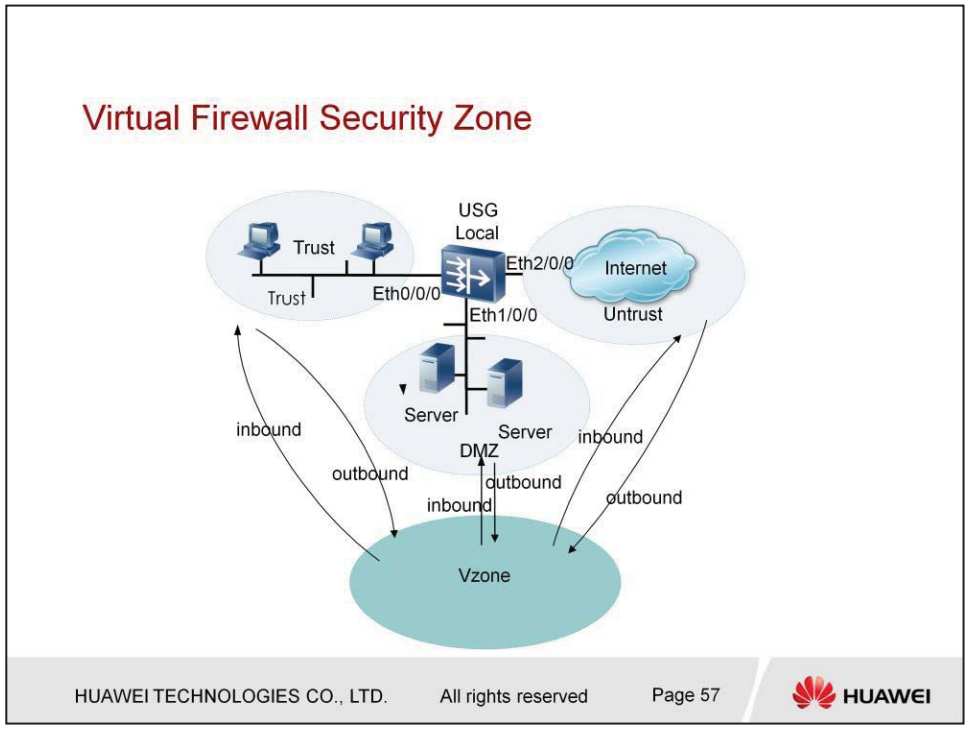
Packet filter

Statistics

Attack defense

ASPF

NAT ALG



Security Zone

Zone is a concept introduced in firewalls, which is one of main features distinguishing the firewall from the router.

For the router, the network security check is performed on interfaces because the networks connected with each interface are equal in security. That is, there is no obvious difference between internal networks and external networks for the router. In this way, when a data stream unidirectionally passes through a router, it may be checked twice on both the inbound interface and the outbound interface to meet the separate security definitions on each interface. However, the firewall's situation is different, where internal networks and external networks are clearly defined. The firewall protects internal networks from illegal intrusion of external networks.

When a data stream passes through a firewall device, the security operation triggered varies with data stream direction. At this time, it is not suitable to check the security policy on the interface of the firewall. Therefore, the firewall introduces the concept of security zone.

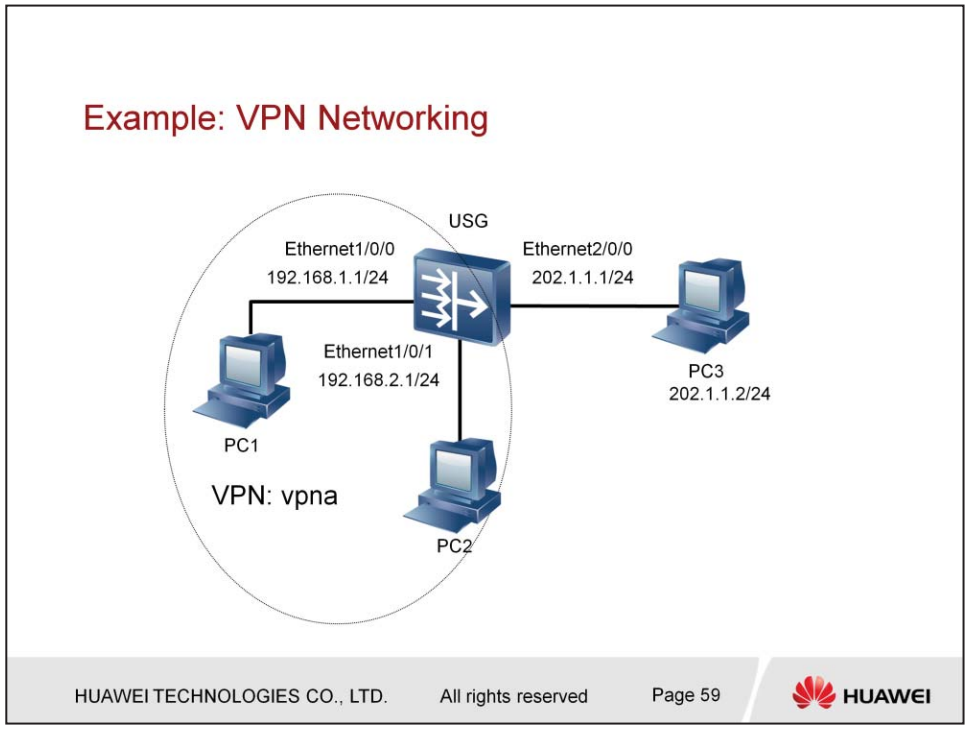
A security zone is composed of one or more interfaces with the same security level.

The features of the security zones are as follows:

The security level is denoted by an integer in the range of 0 to 100. The greater the number is, the higher the level is.

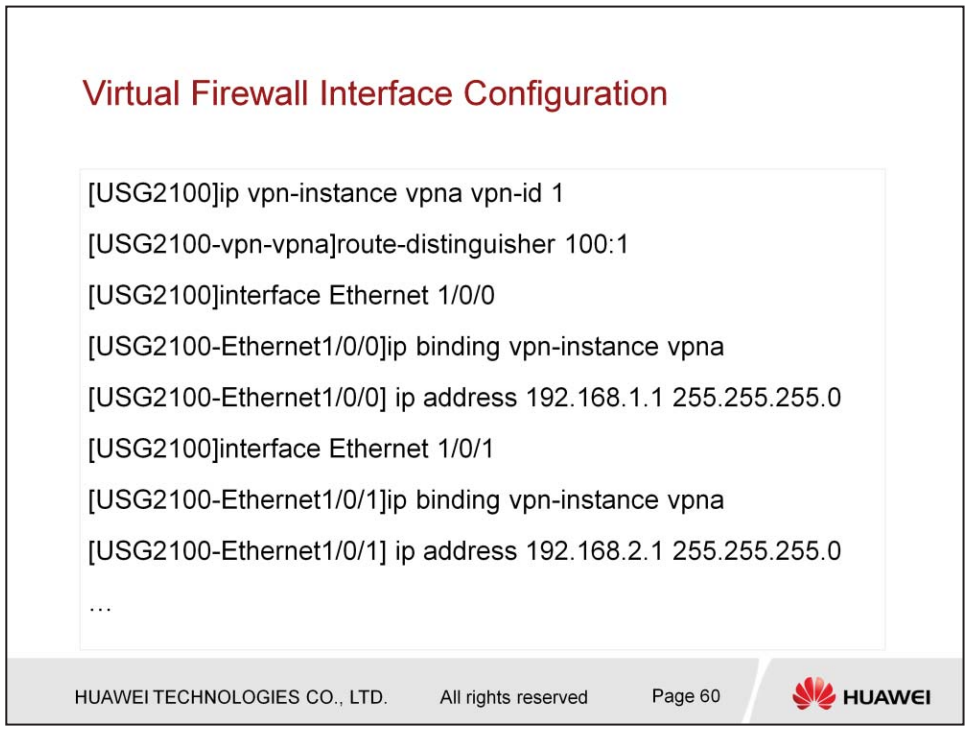
There are no two zones with the same security level.

The firewall security check will be enabled only when data is transmitted between interfaces with different security levels rather than the interfaces in the same security zone.

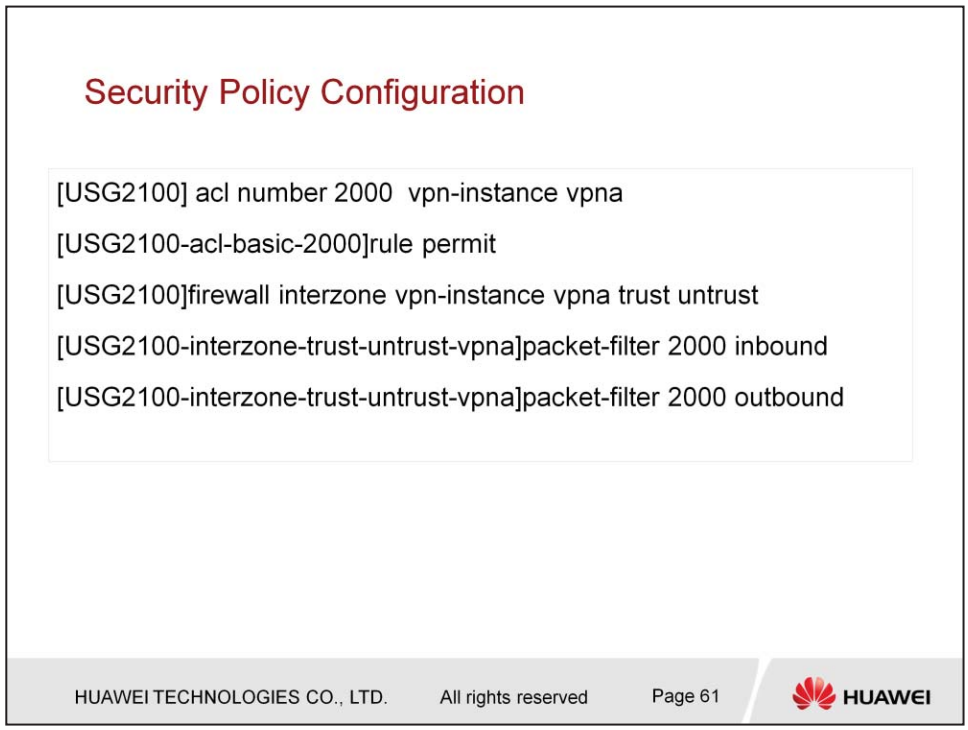


PC1 and PC2 belongs to vpna, PC3 belongs to public network, requirement is the user of vpna can access public PC3 after NAT.

IP address information is shown in figure the above.

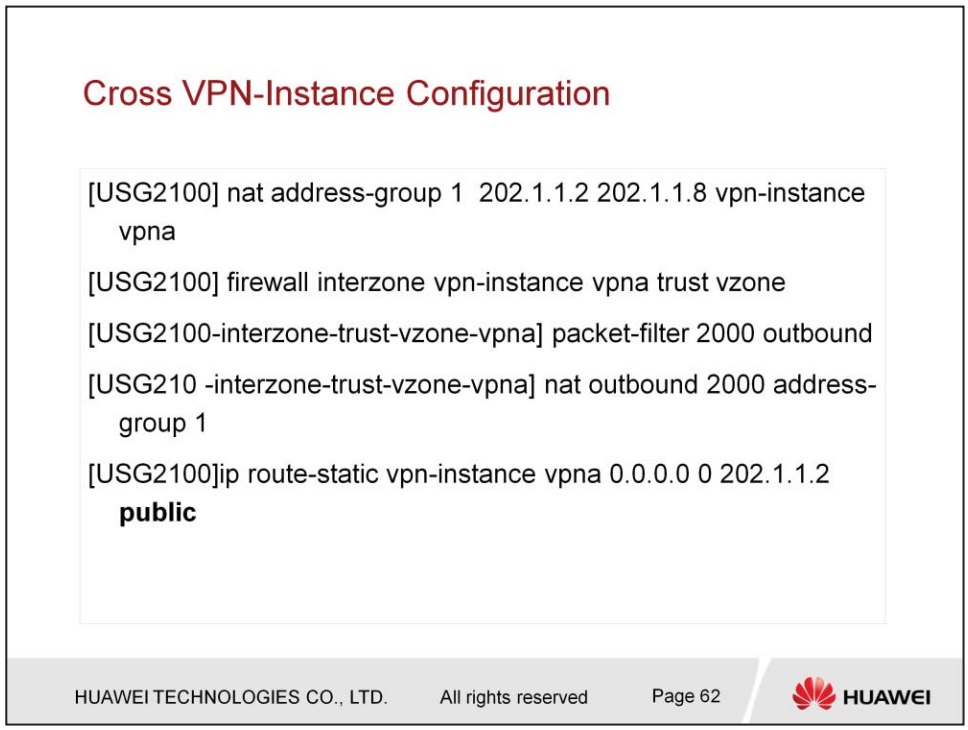


The configuration of virtual firewall for the interface configuration, should firstly bind to the vpn instance, and then configure the IP address.



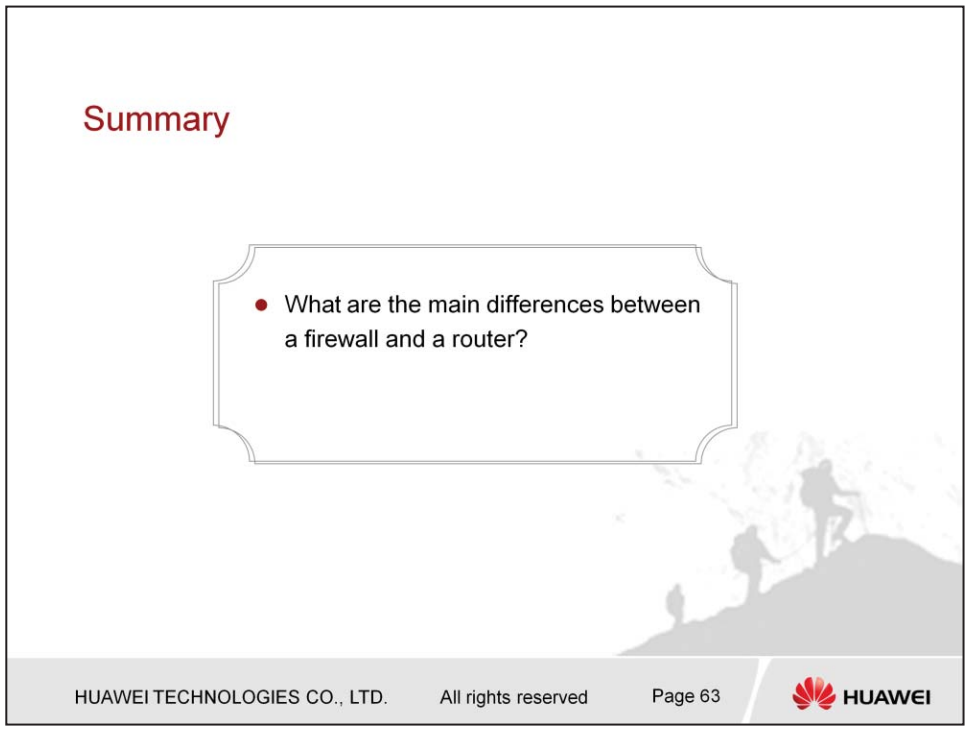
The acl used for vpn-instance, when created, should bind with the vpn instance.

Enter into vpna’s interzone view and apply the security policy.



For the traffic cross VPN-instance, for example vpna's trust should access root untrust, the data flow sequence cross zones is vpna's trust->Vzone->root untrust.

In order to access root zones (public network), for vpn instance vpna, we need to configure a static route, the next hop is a public IP address, means hop to public network.



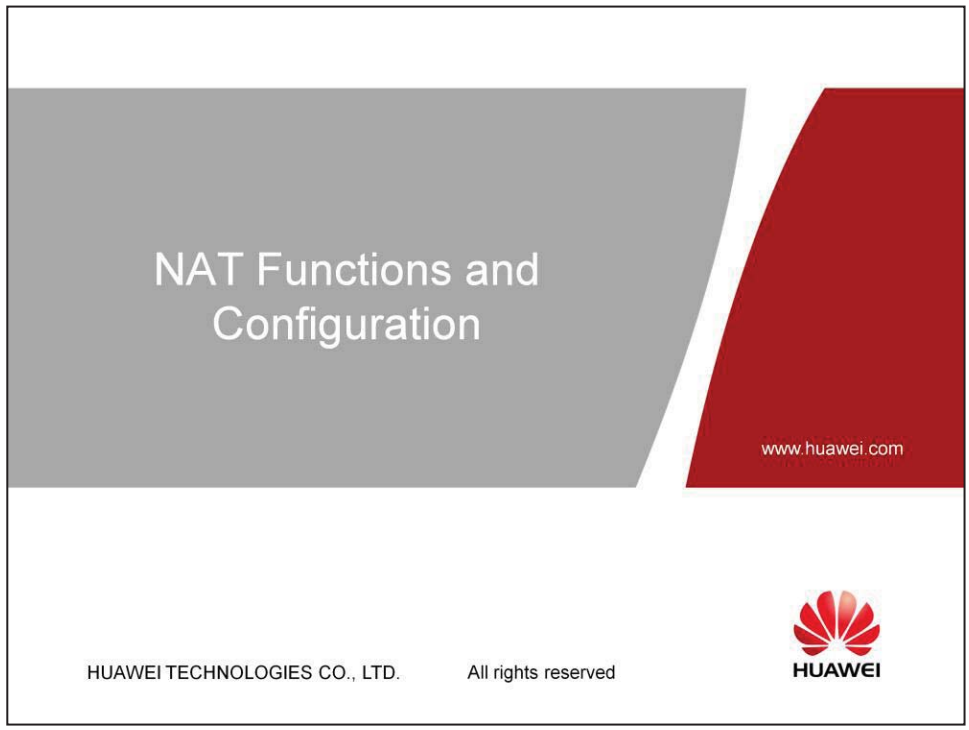
Reference:

A Firewall is stateful, and will maintain the session information, while a router is not.

A Firewall aims to provide communication control, router aims to provide interconnection.

A Firewall provides more detailed security filters than a router does for packets.

A Firewall has zones, while a router doesn't.


NAT Functions and Configuration
www.huawei.com
HUAWEI TECHNOLOGIES CO., LTD.
All rights reserved
HUAWEI

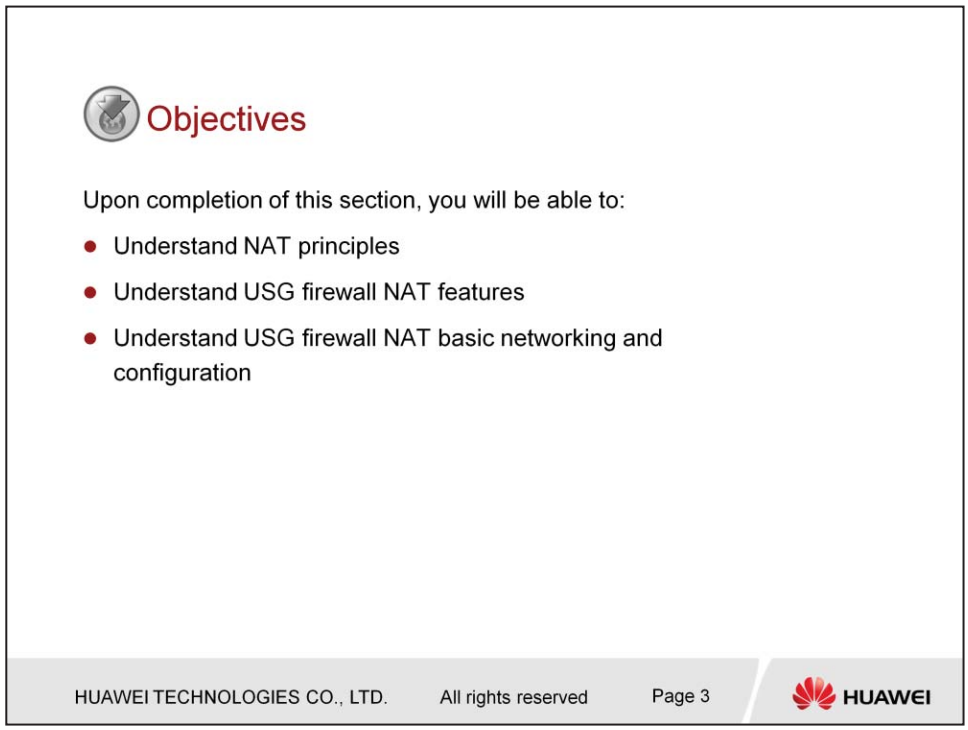


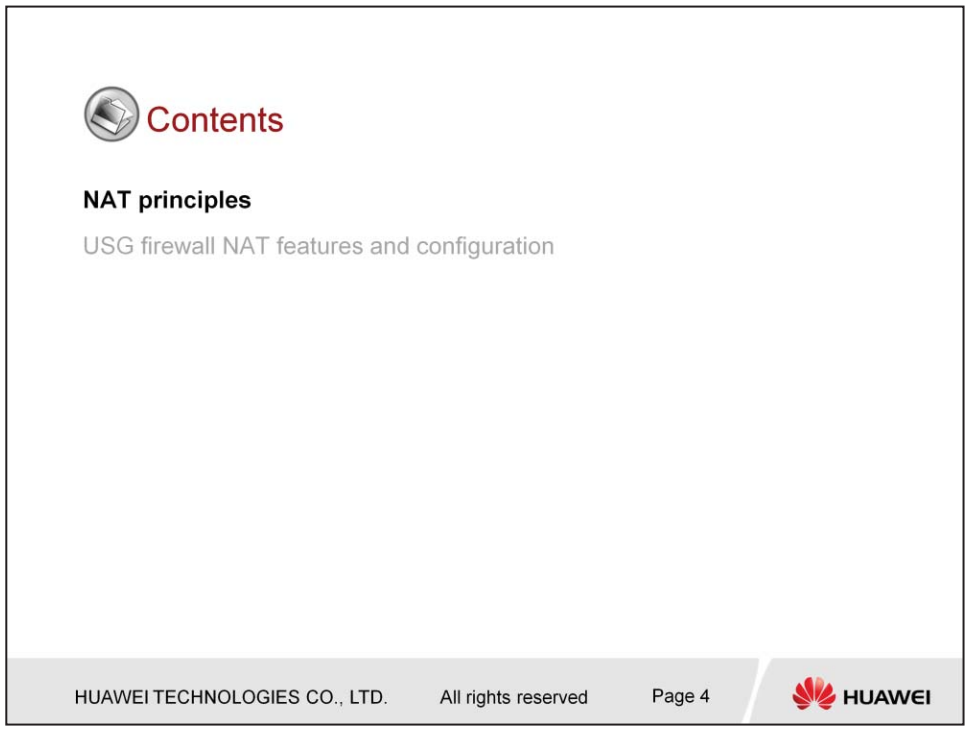
Foreword

With the rapid growth of the Internet, the two urgent problems have arisen, the lack of IP addresses and the expansion to the scale of routing. The long term and short term solutions have been developed, that is the network address translation (NAT) and IPv6 (the next generation of Internet protocol) technology. IPv6 technology is still evolving and has not completely replaced the existing IPv4 network. The short-term solution of NAT technology for alleviating current addressing limitations is still particularly important.

HUAWEI TECHNOLOGIES CO., LTD. All rights reserved

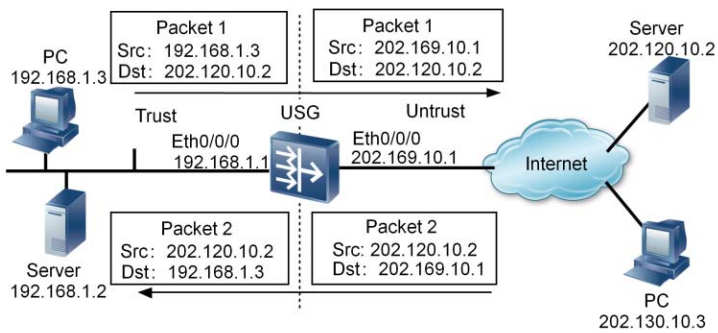



Objectives
Upon completion of this section, you will be able to:
Understand NAT principles
Understand USG firewall NAT features
Understand USG firewall NAT basic networking and configuration
HUAWEI TECHNOLOGIES CO., LTD. All rights reserved Page 3
HUAWEI


Contents
NAT principles
USG firewall NAT features and configuration
HUAWEI TECHNOLOGIES CO., LTD. All rights reserved Page 4
HUAWEI

Principles of NAT

- NAT (Network Address Translation) is a process whereby the IP address of an IP header is translated to another IP address



HUAWEI TECHNOLOGIES CO., LTD. All rights reserved HUAWEI

NAT mechanism can be divided into the following two parts:

Translating an IP address and port of a host in the internal network into an extranet address and port

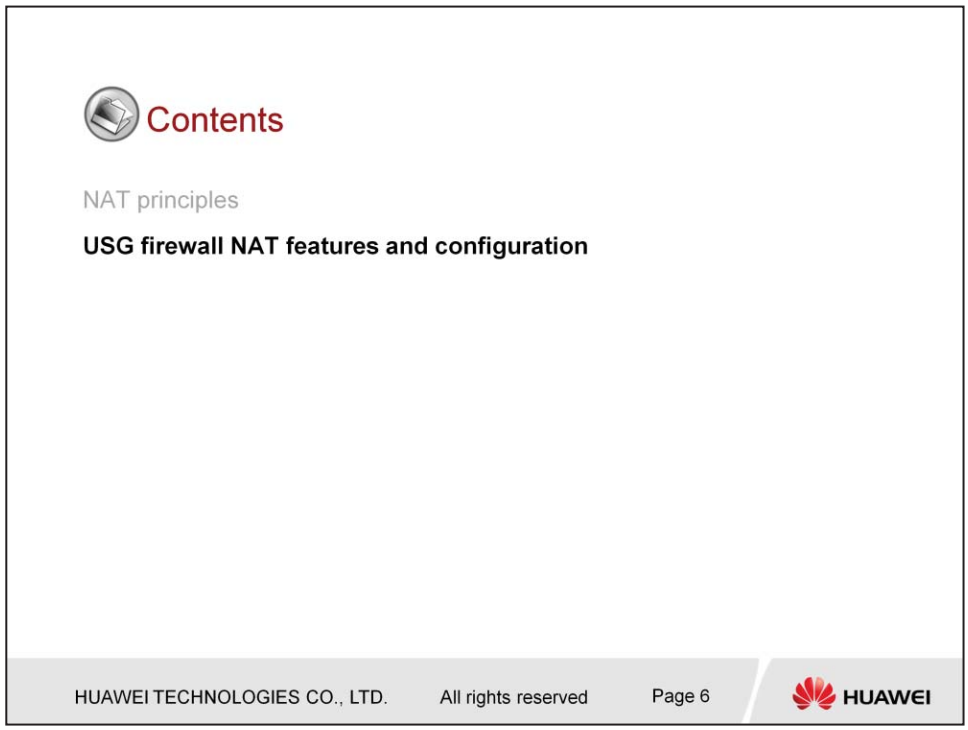
Translating the extranet address and port into the IP address and port of a host in the internal network

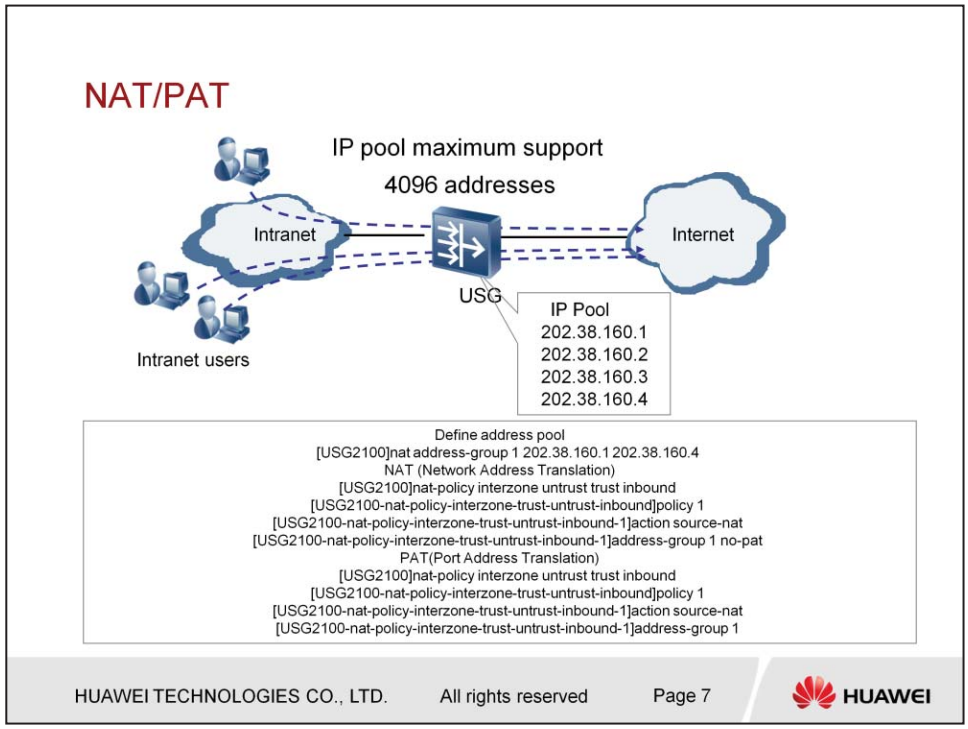
This process is called translation between private address or port and public address or port.

When data flow moves from one security zone to another, the USG checks the data packet to determine whether to perform the NAT. If necessary, the NAT is performed based on the following principles:

At the egress of the IP layer, the USG translates the source address from the private address to the public address and sends it to the external network.

At the ingress of the IP layer, the USG restores the destination address from the public address to the private address and sends it to the internal network.

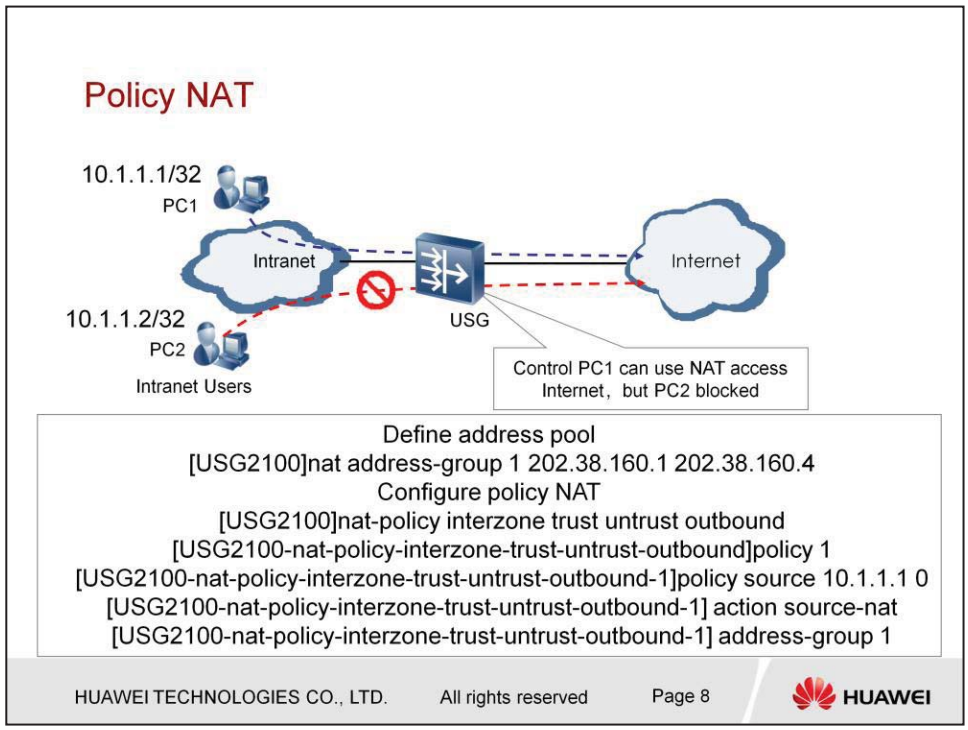

Contents
NAT principles
USG firewall NAT features and configuration
HUAWEI TECHNOLOGIES CO., LTD. All rights reserved Page 6
HUAWEI



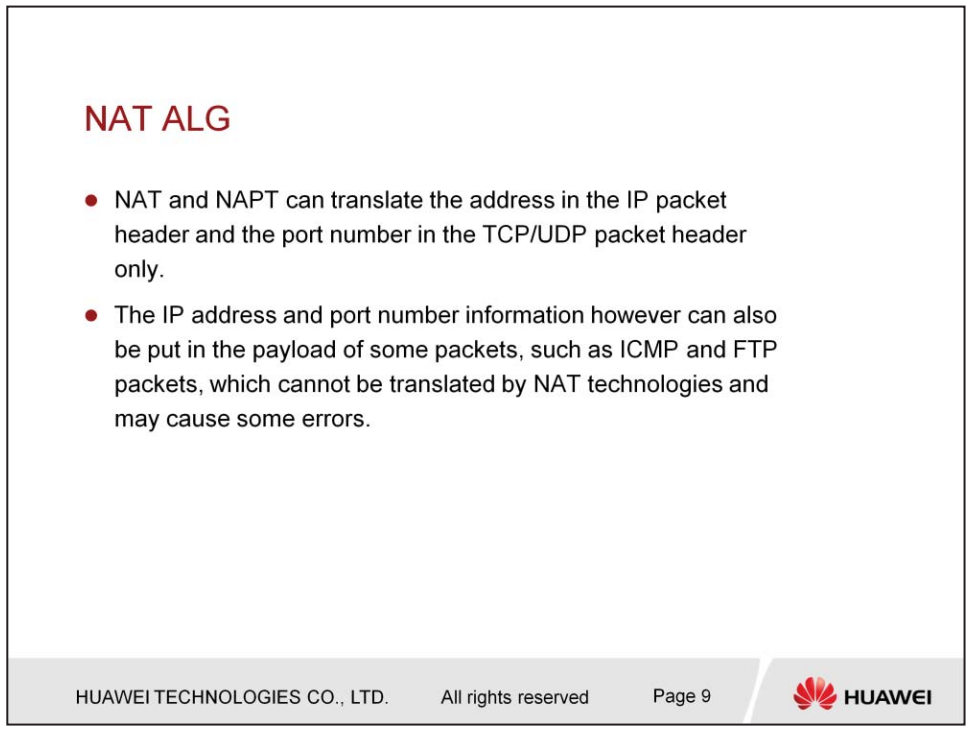
USG series firewall provides IP pool to dynamically assign public IP address for intranet users for NAT, to support more users to access internet simultaneously.

USG2200/5100/5500, USG2200/5100 BSR / HSR supports up to 1024 address pool, each address pool can be configured up to 4096 public network address.

USG2110-X/2100, USG2100 BSR / HSR supports up to 24 address pool eachaddress pool can be configured up to 4096 public network address.



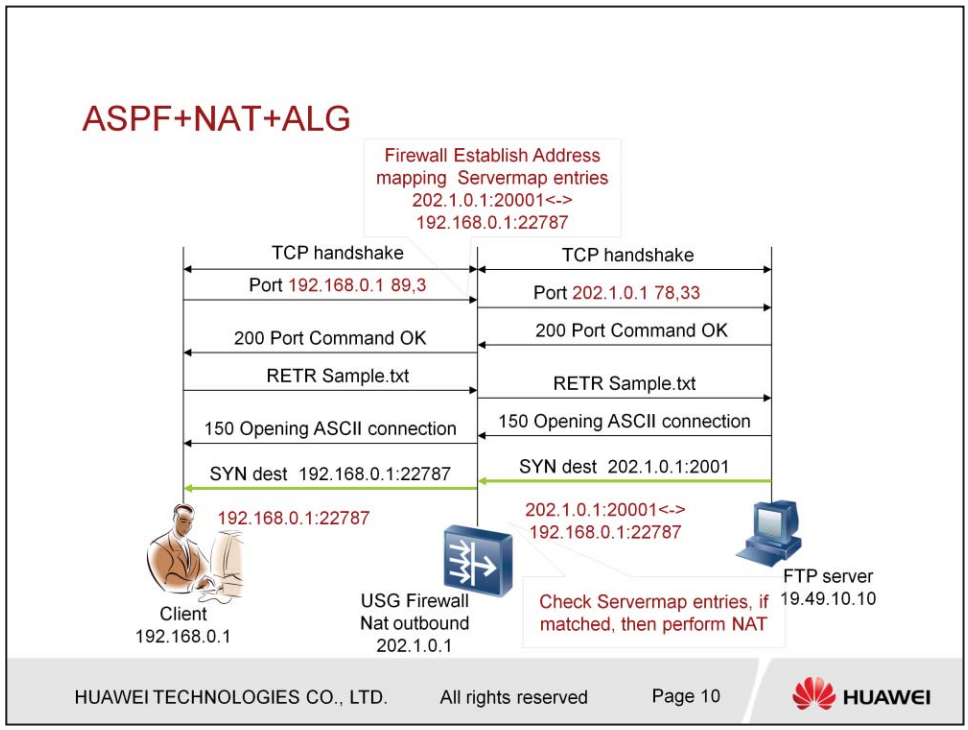
Firewall can use policies to decide which users are allowed to use NAT to access external networks, and block unauthorized group of users. NAT can be set flexibly according to time segments as well.



Adding Application Level Gateway (ALG) to NAT, you can solve the above problem. ALG is the translation proxy of some application protocols. It interacts with NAT to modify the specific data encapsulated in the IP packet based on the NAT state information and helps the application protocols to function in various ranges through other necessary processes.

For instance, a "destination unreachable" ICMP packet is that its data part contains the header of packet A which causes the error (Note because packet A has been translated over NAT, the current source address is not the real address of the internal host). If ICMP ALG is enabled, it will interact with NAT and open the ICMP packet before NAT forwards the packet. NAT translates the address in the header of packet A into the accurate format of the internal host address and forwards the ICMP packet after other necessary processes.

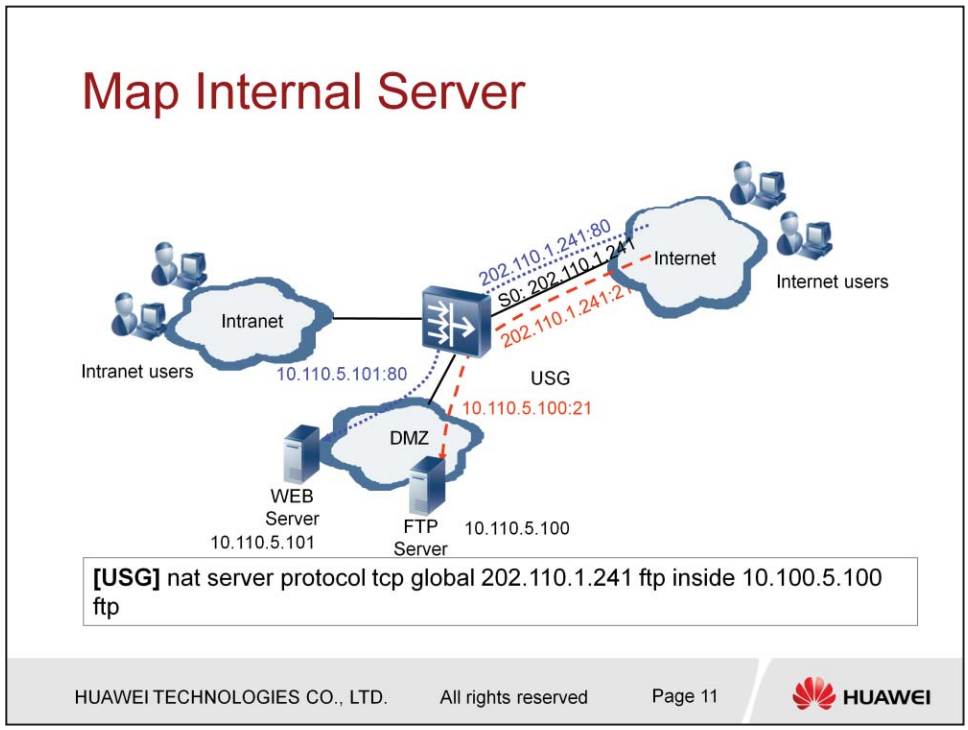
USG provides a perfect NAT ALG mechanism with good scalability, which can support various special application protocols without need of modifying the NAT platform.



Integrate with ASPF function, when NAT, not only translate the IP header's address and port information, some packets carry address and port information in application lay, such information also should be translated.

As shown in the picture, the packet carry port command, the information inside it has been changed, and of course the change mapping information maintained.

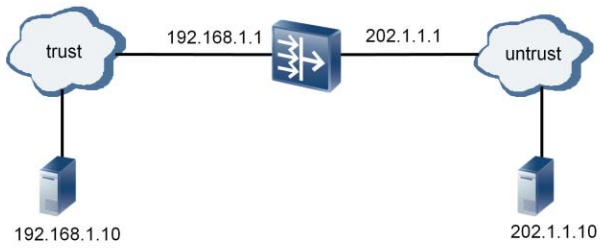
It is named NAT ALG.



USG series firewall can apply NAT to provide internal server mapping，1 public address can maximum map onto 256 servers' private address.

If NAT and internal server functions are configured on the USG at the same time, the address translation is processed based on the configuration of the internal server when the address needs translation. When NAT and internal server functions are configured for IP addresses in the same private network, the firewall translates the address based on the configuration of the internal server.

Bi-directional NAT



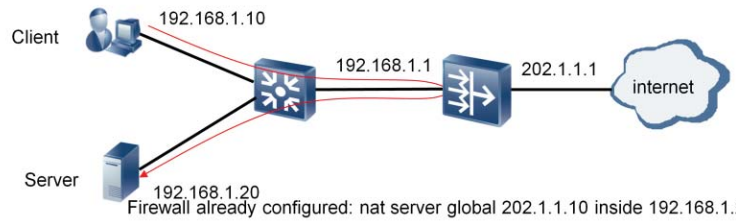
- **Configure NAT SERVER**
  [USG2100]nat server global 192.168.1.20 inside 202.1.1.10
  [USG2100]nat server global 202.1.1.20 inside 192.168.1.10

HUAWEI TECHNOLOGIES CO., LTD. All rights reserved

Server is located in Internet，using a firewall static NAT Server feature to map as an internal private IP address。 Intranet PC can directly configure an intranet IP address to represent the external server, as if the external server is placed within the intranet. This can avoid exposing the external server IP address to intranet users.

PC is located at high priority area, and server located in low priority area, requires to implement static NAT server from low priority to high priority NAT Server transition, as well as to support NAT ALG function.

Internal NAT



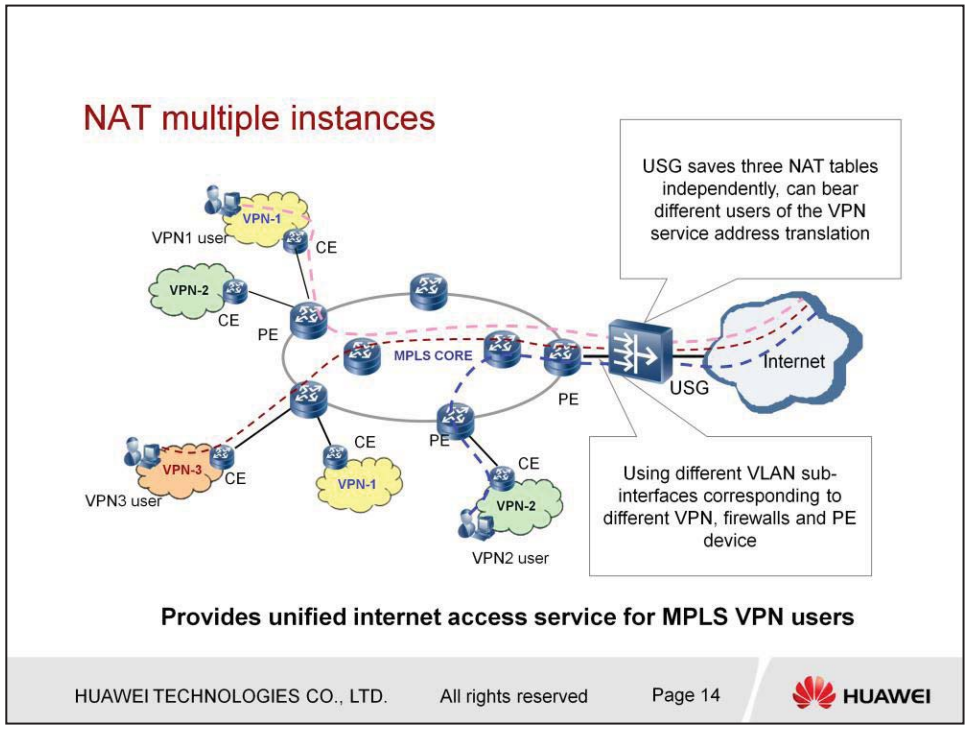
- Configure the same zone NAT

```
[USG] nat address-group 1 202.1.1.11 202.1.1.11
[USG] nat-policy zone trust
[USG-nat-policy-zone-trust] policy 0
[USG-nat-policy-zone-trust-0] policy source 192.168.1.0 0.0.0.255
[USG-nat-policy-zone-trust-0] action source-nat
[USG-nat-policy-zone-trust-0] address-group 1
```

HUAWEI TECHNOLOGIES CO., LTD. All rights reserved

Intranet server IP address is private IP, firewall can do a static NAT Server mapping onto a public IP address.

When intranet client trying to access the NAT SERVER public address, its IP packets containing destination public IP is changed into a private IP address, and then forward back into internal network server. The source IP address is also required to be changed to a public address range.



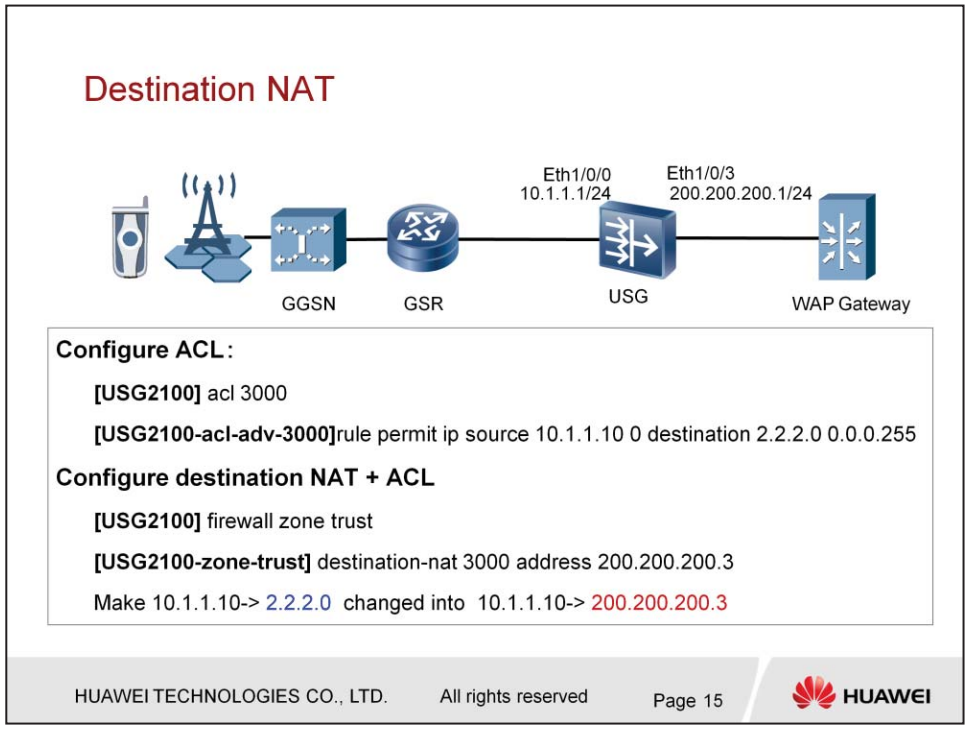
USG series firewall support virtual firewall, address translation multiple instance, multi-VRF VPN isolation technology and so on, so it is very suitable for USG firewall will be placed in a MPLS backbone network of export, as a MPLS VPN network access the Internet unified export equipment.

USG series firewall has some characteristics as follows, suitable for complete work are as follows:

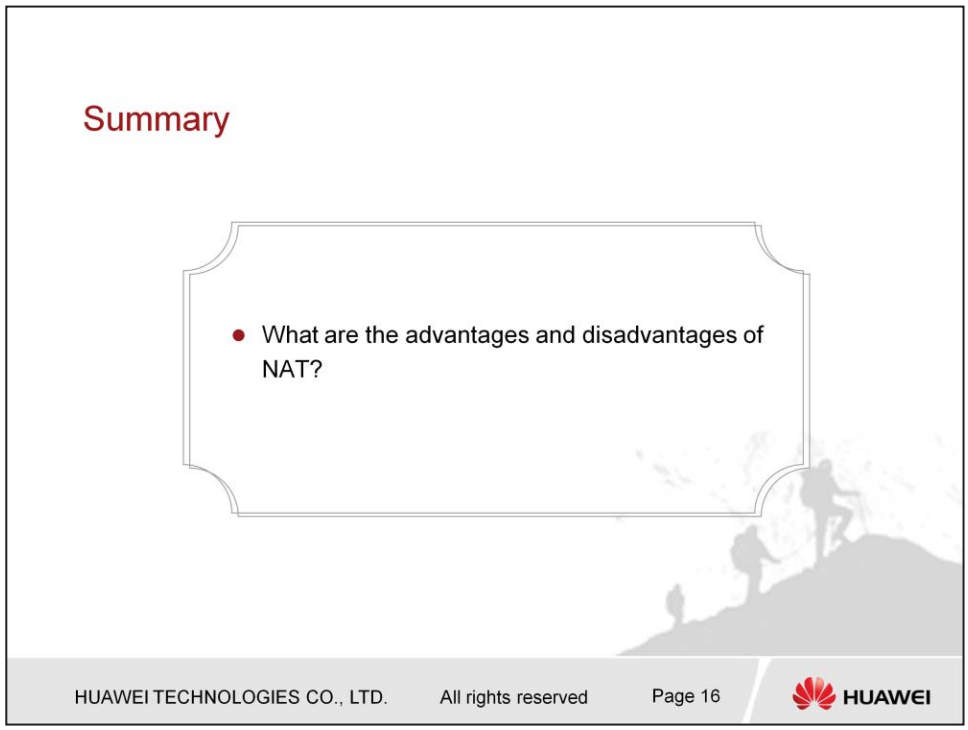
Support the perfect safety business, can avoid the Internet from attack, protect the whole MPLS network.

Support the perfect VLAN, VRF, OSPF, BGP technology, can solve the MPLS network and IP network integration problems well.

Support multiple instances business characteristics, the resources independent store, can solve the problem of overlapping private web address well.



For some mobile phones, the default WAP gateway is fixed. When the user roams to different countries, the WAP gateway should be changed. It is possible to use the firewall to change the WAP gateway IP address.



Reference:

Advantages: spare IP address, Hide the internal host avoid to be attacked.

Disadvantages: Internal network is hidden, difficult to trouble shooting, and for NAT device, it may be a performance bottleneck.


Attack Defense and Configurations
www.huawei.com
HUAWEI TECHNOLOGIES CO., LTD.
All rights reserved
HUAWEI



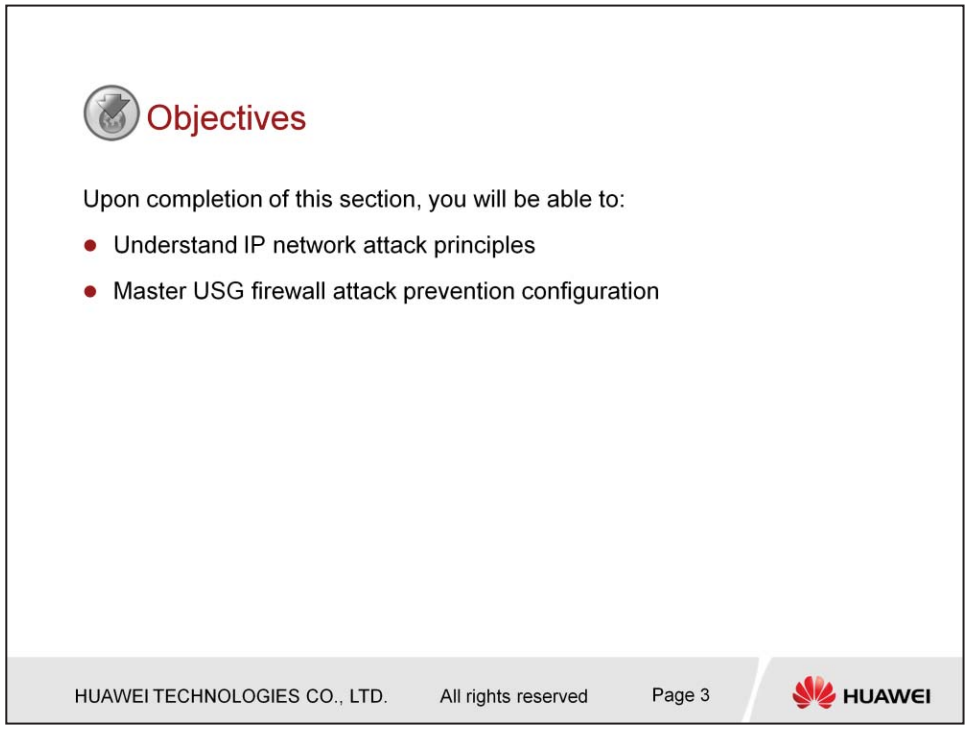
Foreword

The attack defense of the USG can detect various types of network attacks and take the measures to protect internal networks from malicious attacks. As a result, the Eudemon can assure the normal operations of the internal networks and systems.

This section mainly describes the principles and USG firewall preventive configurations for IP network attacks.

HUAWEI TECHNOLOGIES CO., LTD. All rights reserved




Objectives
Upon completion of this section, you will be able to:
Understand IP network attack principles
Master USG firewall attack prevention configuration
HUAWEI TECHNOLOGIES CO., LTD.
All rights reserved
Page 3
HUAWEI



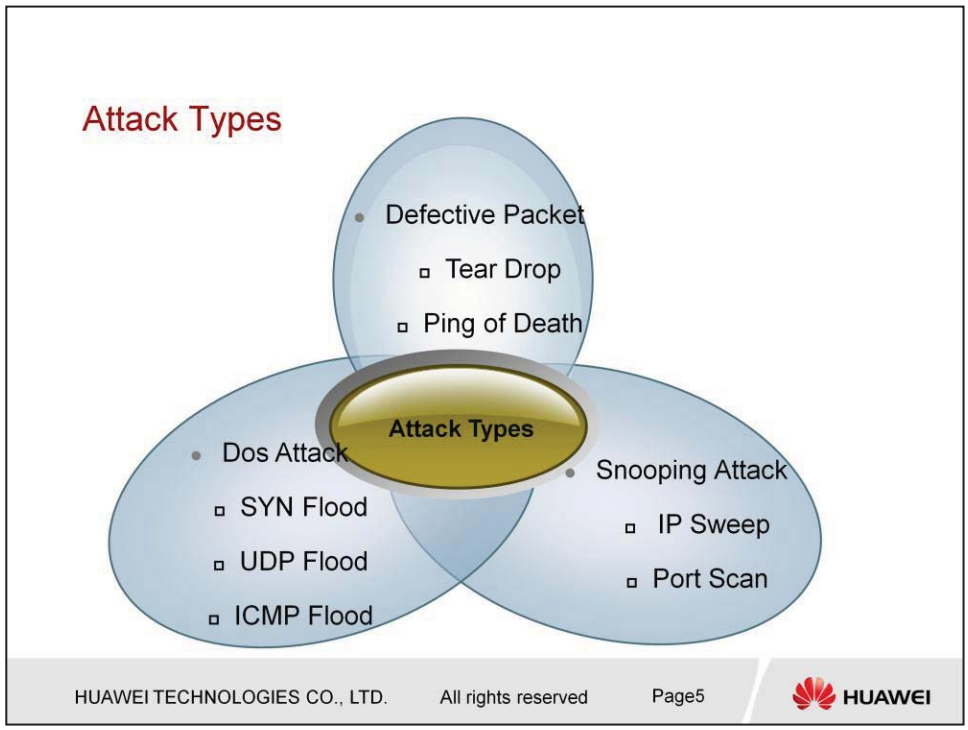
Contents

DDoS

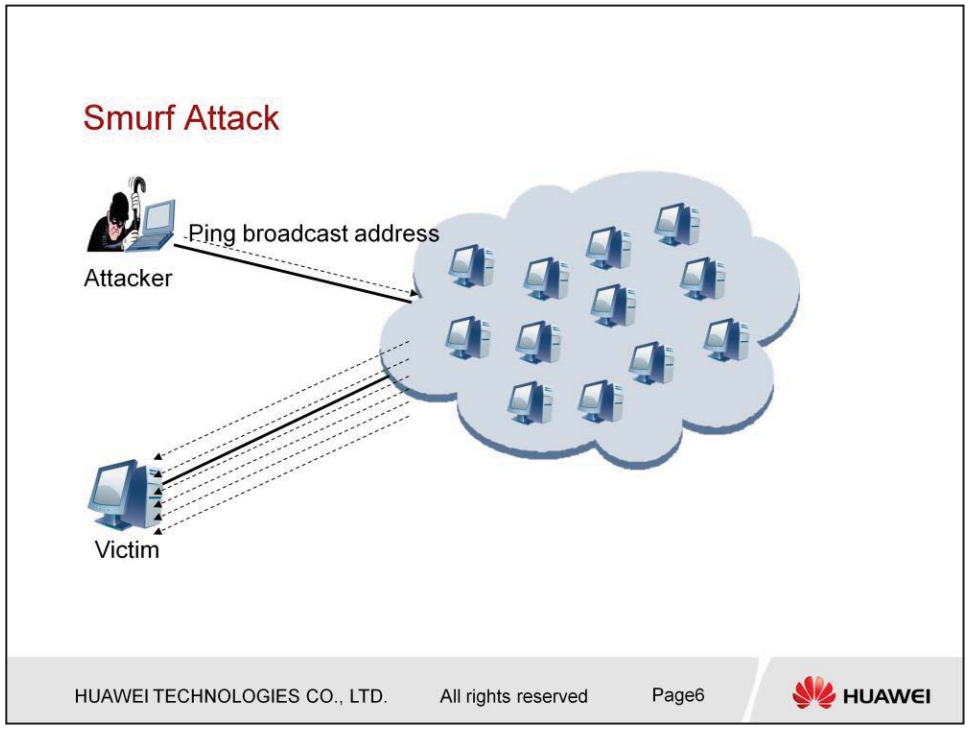
Deformity packet attack

IP sweep attack

HUAWEI TECHNOLOGIES CO., LTD. All rights reserved



There have many attacks, according to their features, we can classify them into several types, like Dos attacks, defective packet attacks and snooping attacks.



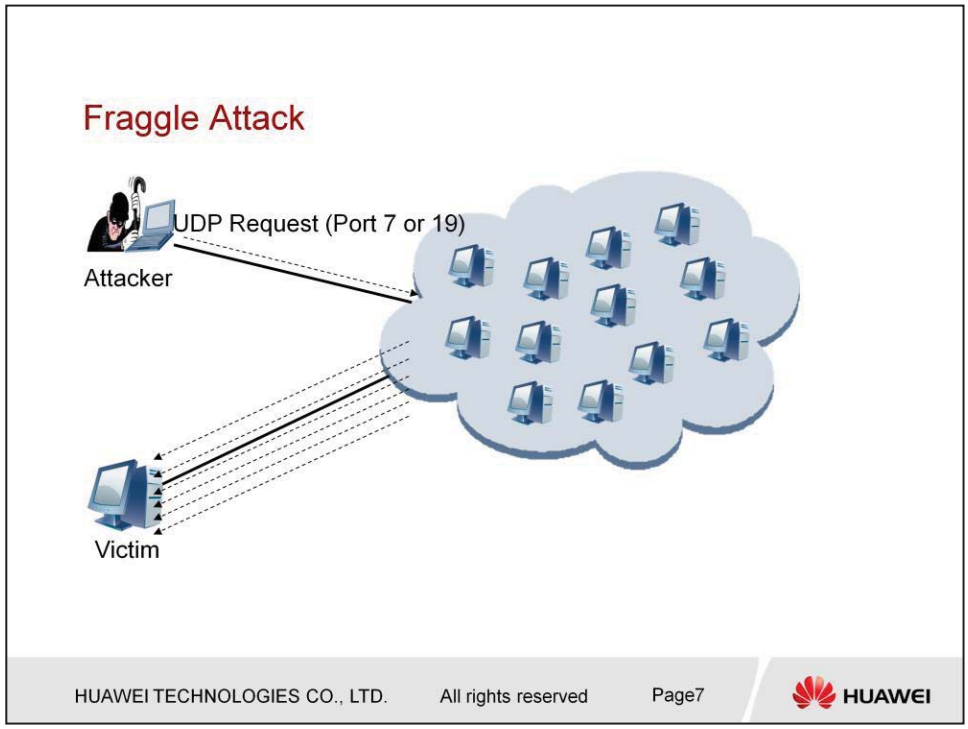
The simple Smurf attack is to attack a network by sending an ICMP request to the broadcast address of the target network. All the hosts in the network will respond to the request, which will generate the traffic 10 or 100 times more than the traffic of large ping packets. Network congestion thus occurs. The advanced Smurf attack is mainly used to attack the target host by configuring the source address of the ICMP packet to the address of the target host so as to make the host crash finally. It takes certain traffic and duration to send the attack packet to perform attack. Theoretically, the larger the number of the hosts is, the more obvious the effect will be. Another new form of the Smurf attack is the Fraggle attack.

Feature: Imitate the victim's IP address to send ping echo to a broadcast address.

Purpose: Cause the victim to be flooded by responses from hosts on the network.

Configuration: firewall defend smurf enable

Prevention: Discard the packets whose destination address is a broadcast address.

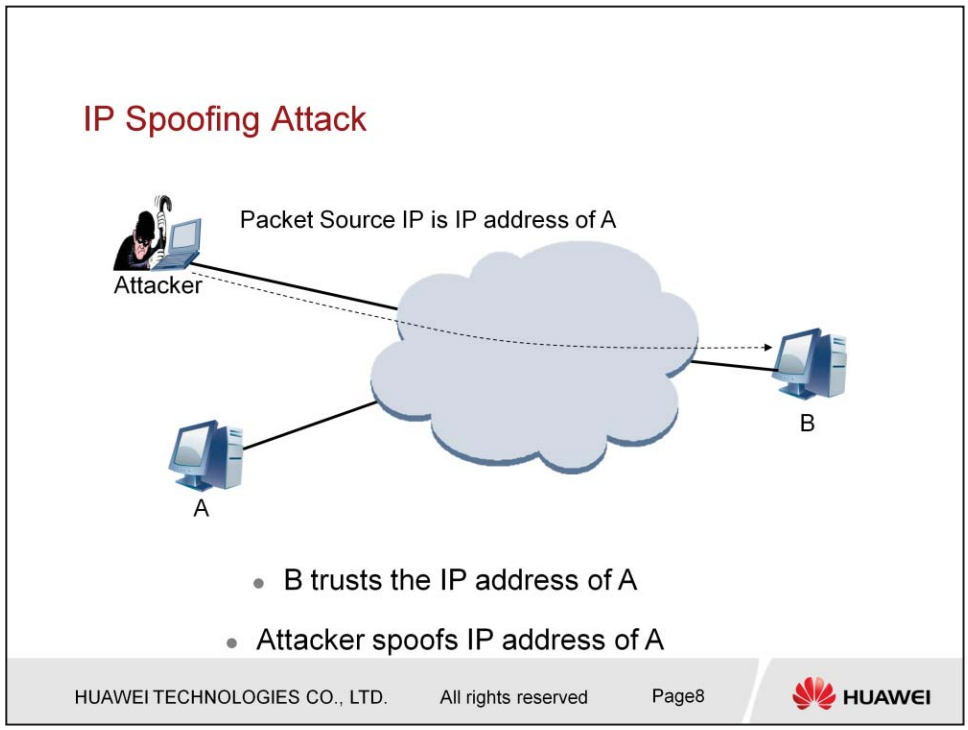


The Fraggle attack like smurf, smurf choose icmp ping service, fraggle choose udp echo (port 7) or udp chargen(udp 19) service.

The attacker sends, by imitating address of the victim, requests to the above port whose destination address is a broadcast address, so that the victim will be attacked by the widespread response packets. If both the source and destination addresses are broadcast addresses, the network bandwidth will be fully seized.

Configuration: firewall defend fraggle enable

Prevention: Filter UDP packets whose destination port number is 7 or 19.



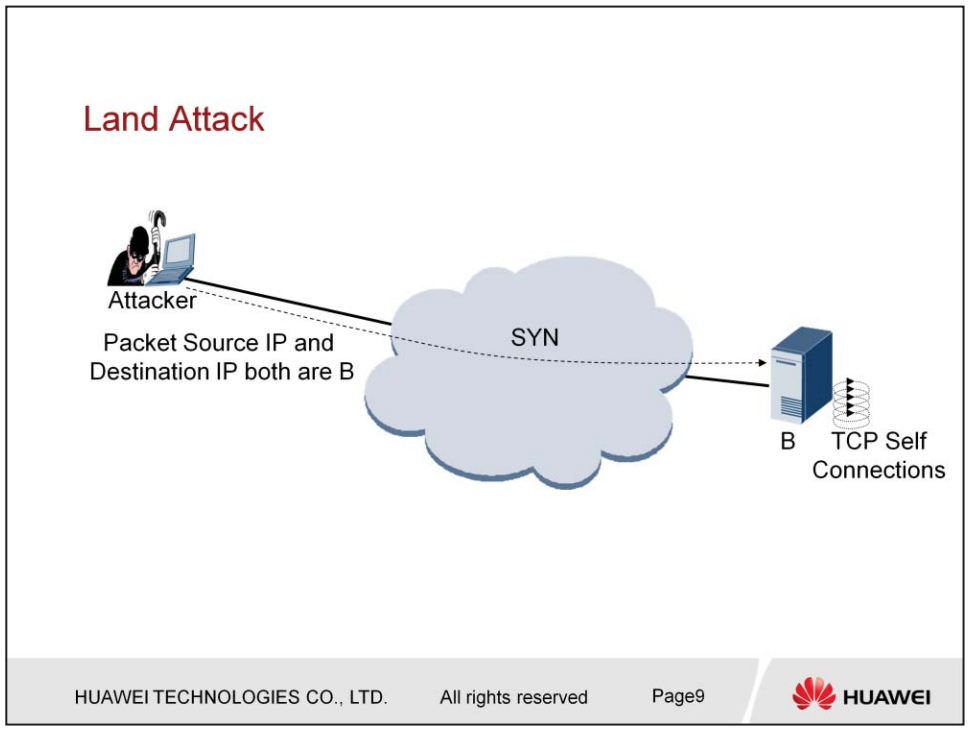
To get an access right, an intruder generates a packet carrying a bogus source address which can make an unauthorized client access the system applying the IP authentication even in the root authority. In this way, the system can also be destroyed even though the response packet does not reach the system. This is the IP Spoofing attack.

Feature: Address spoofing.

Purpose: Spoof an IP address to transmit packets.

Configuration: **firewall defend ip-spoofing enable**

Prevention: Search the routing table for the source address and discard the packet whose ingress interface is not the egress interface of the IP address.



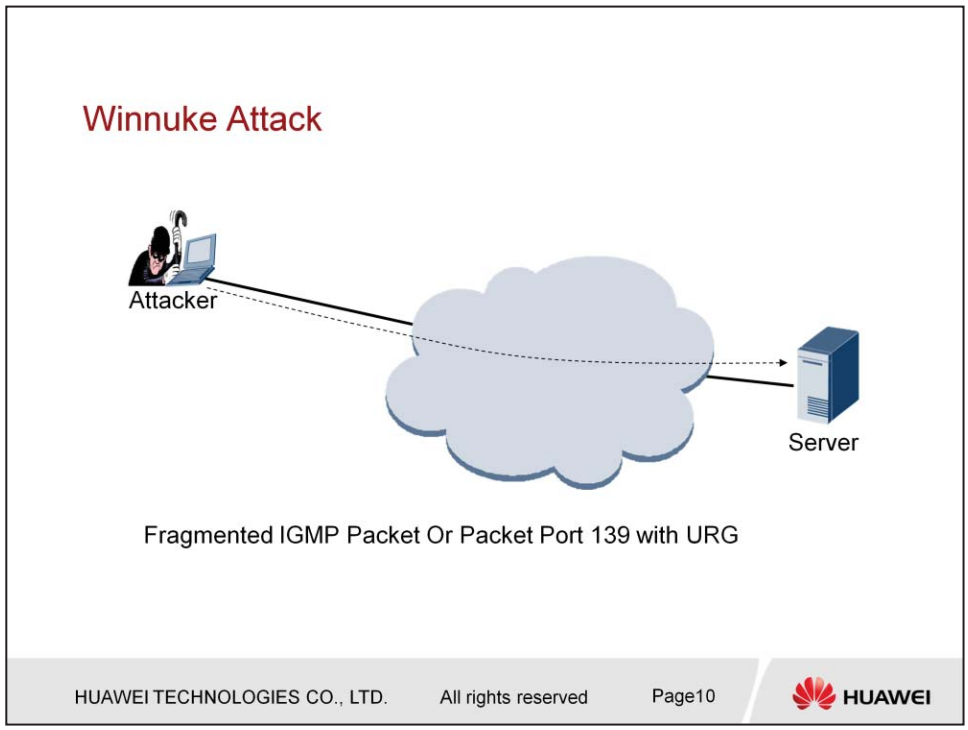
Land attack is to configure both the source address and the destination address of the TCP SYN packet to the IP address of the attack target. Thus, the target sends the SYN-ACK message and sends back the ACK message to itself, and then creates a null connection. Each of the null connections will be saved until it times out. Different attack targets have different responses to the Land attack. For instance, many UNIX hosts will crash and Windows NT hosts will slow down.

Feature: Both the source and destination addresses are the IP address of the victim, or the source address is in the “127” network segment.

Purpose: Let the attacked equipment send response packets to itself. This is used in the SYN flood attack.

Configuration: firewall defend land enable

Prevention: Discard the packets with the above feature.



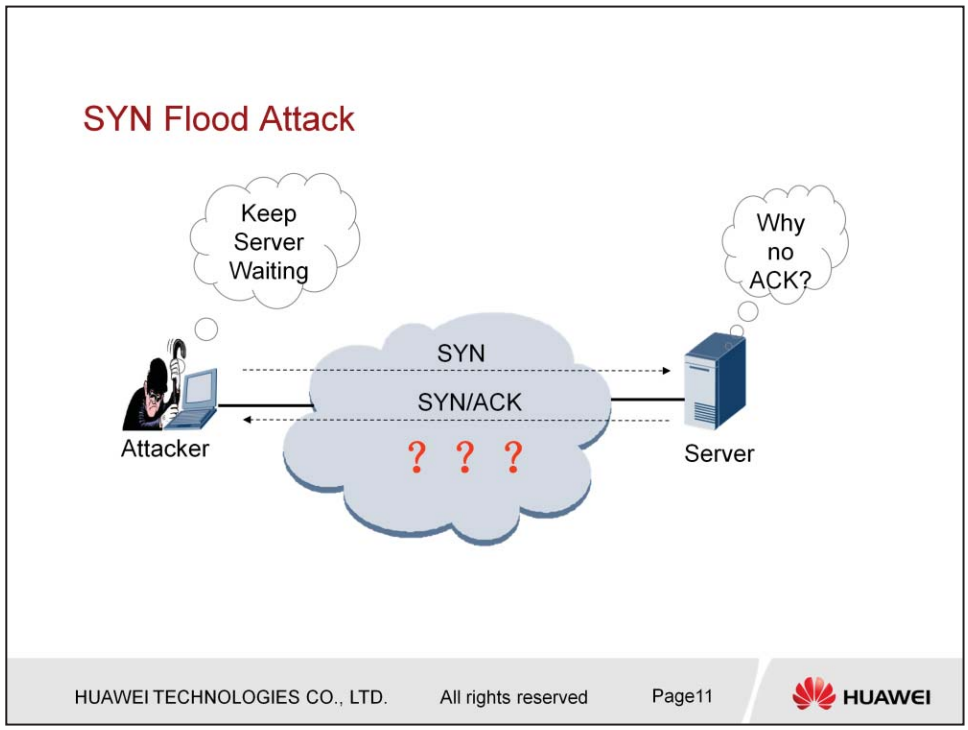
WinNuke attack is to cause a NetBIOS fragment overlap by sending Out-Of-Band (OOB) data packets to the NetBIOS port (139) of the specified target installed with the Windows system so as to make the target host crash. There are also Internet Group Management Protocol (IGMP) fragment packets. Because IGMP packets cannot be fragmented generally, few systems can solve the attack caused by IGMP fragment packets thoroughly. When the system receives IGMP fragment packets, you can guess there is attack.

Feature: IGMP packets are marked with fragmentation, or packets are marked with URG for port 139.

Purpose: Cause the attacked equipment to stop responding for improper processing.

Configuration: **firewall defend winnuke enable**

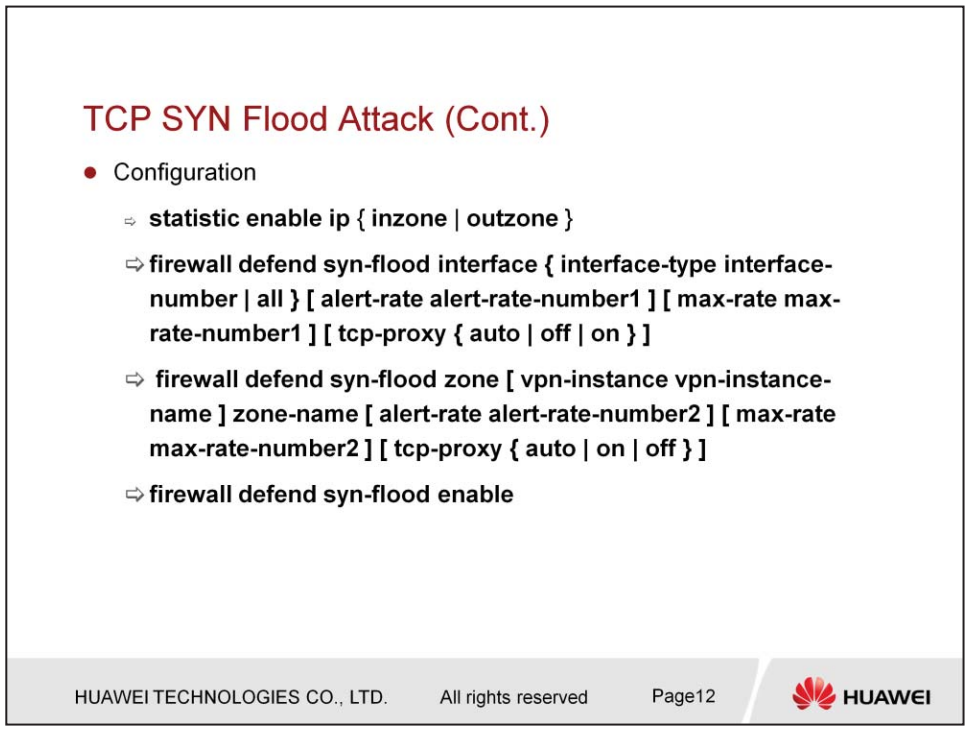
Prevention: Discard the packets with the above feature.



Because of the limited resources, TCP/IP stacks only permit a restricted number of TCP connections. Based on the above defect, the SYN Flood attack forges an SYN packet whose source address is a bogus or non-existent address and initiates a connection to the server. Accordingly, the server will not receive the ACK packet for its SYN-ACK packet, which forms a semi-connection. A large number of semi-connections will exhaust the network resources. As a result, valid users cannot access the network until the semi-connections time out. The SYN Flood attack also takes effect in the applications whose connection number is not limited to consume the system resources such as memory.

Feature: Send abundant TCP connection requests to the victim host.

Purpose: Use up all processing capability of the attacked equipment so that it cannot respond to the requests of normal users normally.



inzone: Collects statistics of the data packets in the inbound direction based on a security zone or an IP address. The data packets in the inbound direction indicate the data packets whose destination addresses are in the security zone and the data packets are collected according to the destination addresses in the case of IP-based statistics.

outzone: Collects statistics of the data packets in the outbound direction based on a security zone or an IP address. The data packets in the outbound direction indicate the data packets whose source addresses are in the security zone and the data packets are collected according to the source addresses in the case of IP-based statistics.

interface *interface-type interface-number*: Sets a protected interface.

ip *ip-address*: Sets the IP address for the protected host. Through the SYN Flood attack defense function, a maximum of 1000 IP addresses can be simultaneously protected.

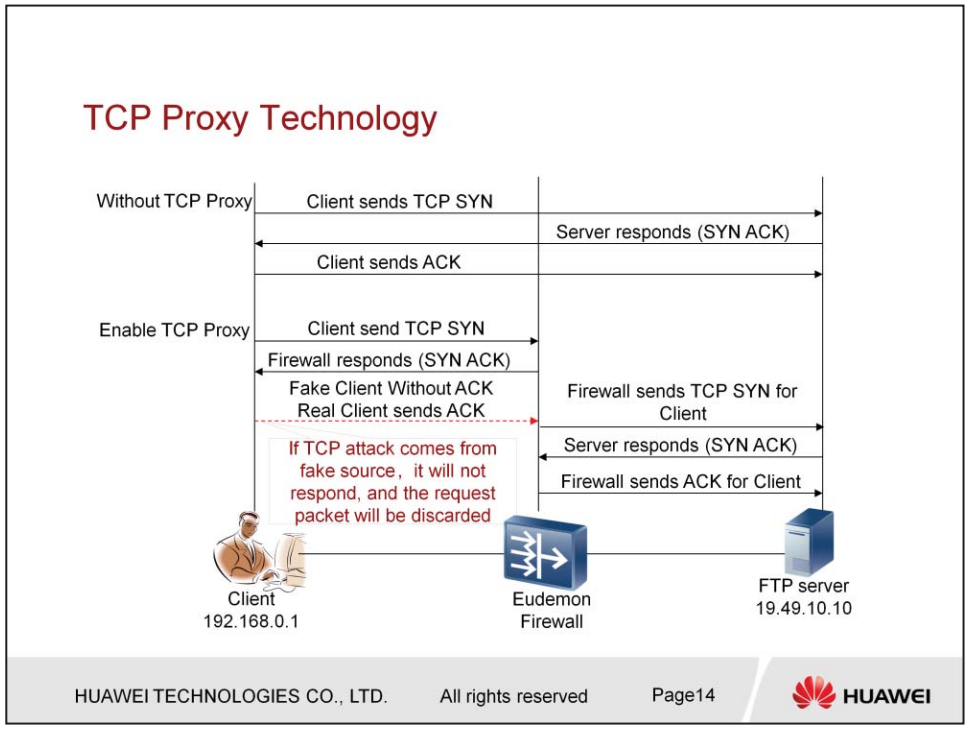
zone *zone-name*: Sets the name of a protected security zone.

max-rate *rate-number*: Sets the threshold of the SYN packet rate that connects the specific destination IP address, that is, the number of SYN packets per second. When the threshold is exceeded, the attack defense is enabled. For the interface-based case, the value of **rate-number** ranges from **1** to **1,000,000** (packets/second). For the IP or zone-based case, the value of **rate-number** ranges from **1** to **65535** (packets/second). The default value is **1000** (packets/second).

tcp-proxy { **auto** | **on** | **off** } : Sets the start state of TCP agent. **auto** indicates automatic enabling. When detecting that the protected host is attacked by SYN Flood, the TCP agent is automatically enabled. When the proportion of the received ACK packets to the received SYN packets exceeds 7/8, the TCP agent is automatically disabled. **on** means manual enabling, that is, the TCP agent is enabled no matter whether the protected host is attacked by SYN Flood. **off** means manual disabling, that is, the TCP agent is disabled no matter whether the protected host is attacked by SYN Flood. The default value of **tcp-proxy** is **auto**. That is, enabling or disabling the TCP agent depends on the system.

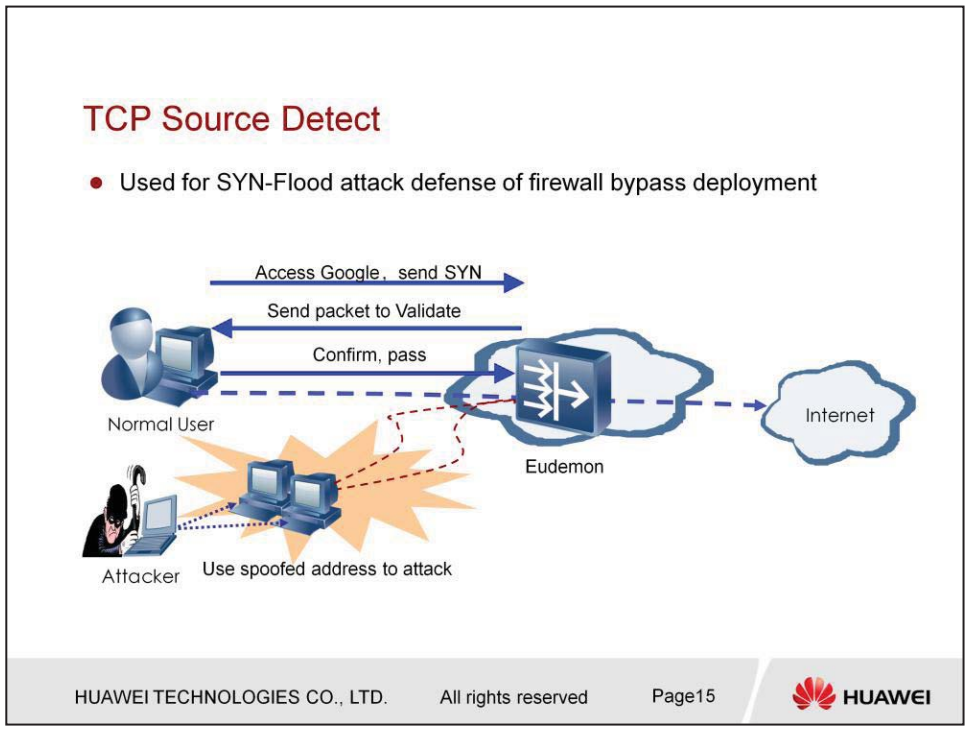
source-detect: Sets the enabling state of TCP reverse detection. **auto** indicates automatic enabling. When detecting that the protected host is attacked by SYN Flood, the TCP reverse detection is automatically enabled. When the proportion of the received RST packets to the received SYN packets exceeds 7/8, the TCP reverse detection is automatically disabled. **on** means manual enabling, that is, the TCP reverse detection is enabled no matter whether the protected host is attacked by SYN Flood. **off** means manual disabling, that is, the TCP agent is disabled no matter whether the protected host is attacked by SYN Flood. The default value of **source-detec**t is **auto**.

vpn-instance *vpn-instance-name*: Sets a protected VPN instance.



One way to defend against SYN flood attacks is using the TCP proxy function. The client sends TCP connection requests to the server through the firewall. If the firewall's proxy function is enabled, the firewall will intercept the packet sent from the client and pretend to be the server to send response to the client. When the client acknowledges the TCP connection, then the firewall acts as a client to request connection to the server. The firewall can protect the server from SYN flood attacking, the firewall should accelerate aging of the session so that the firewall's resources will not be used up.

TCP Proxy technology causes the firewall to do the TCP handshake before setting up the TCP connection to the server.



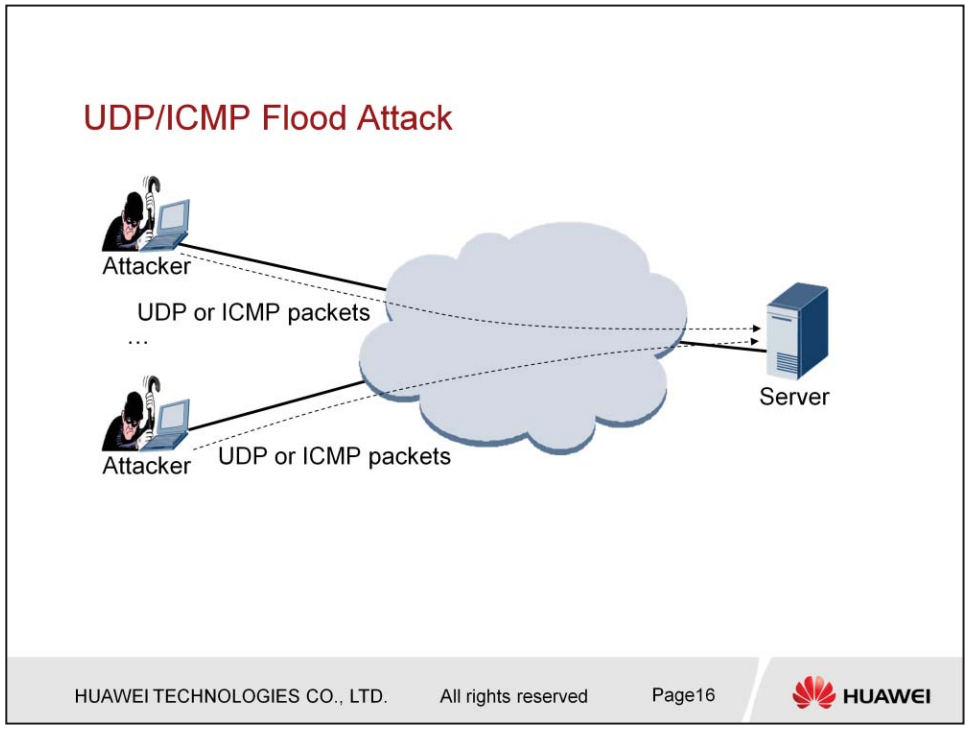
First, verify the source, and then verify the packet TTL.

Detailed process steps:

If the firewall has enabled the TCP source detect function, upon receiving a SYN packet, it will reply with a SYN-ACK packet with an ACK_sequence number, this ACK_sequence number taken from the HASH combines with the SYN packet's TTL. The ACK_sequence number should be the SYN packet's sequence number plus 1.

If the client of source IP exists, it will reply with a RST packet, the firewall checks whether the reply packet's sequence number is equal to the ACK_sequence number of previous SYN-ACK packet sent by firewall. If it is equal, recover the TTL from this sequence number and compare it with packet's TTL value, if these two numbers are same or only with little difference, it means the packets of SYN and RST are from the same source.

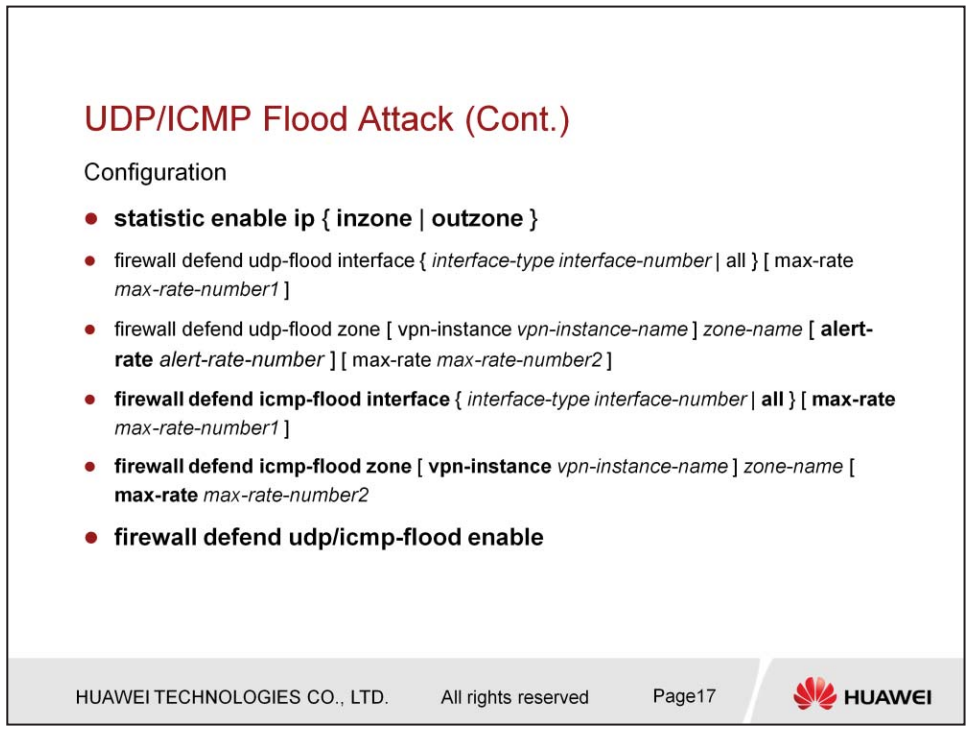
This source IP address' status in the source IP monitor table of firewall will be set to authenticated, the next packet can match the source IP monitor table, meaning pass the authentication, and be forwarded.



ICMP and UDP flood attack sends a large number of ICMP messages (such as ping) and UDP packets to the specific target in a short time. Thus, the target system is unable to transmit valid packets normally. Usually these attacks would be distributed (DDos attack).

Feature: Abundant UDP/ICMP packets are sent to the victim host.

Purpose: Use up all processing capability of the attacked equipment.



interface: Sets a protected interface.

ip ip-address: Sets a protected IP address. Through the UDP Flood attack defence function, a maximum of 1000 IP addresses can be simultaneously protected.

zone zone-name: Sets a protected security zone.

alert alert-rate: Sets the alarm rate for IP or security zone-based UDP packets. When the number of UDP packets per second exceeds the threshold, an attack occurs. Fingerprint filter is enabled. The value ranges from 1 to 30000. The default value is 500 packets/second.

max-rate rate-number: For the interface-based case, when the number of first UDP packets of each link per second exceeds the threshold, an attack occurs. The rate of first UDP packets of each link of the interface is restricted. For the IP or security zone-based case, when the number of UDP packets per second exceeds the threshold, the destination rate restriction is performed. For the IP or security zone-based case, the value ranges from 1 to 65535 (packets/second). For the interface-based case, the value ranges from 100 to 500000 (packets/second). For the IP or security zone-based case, the default value is 1000 (packets/second). For the interface-based case, the default value is 500000 (packets/second).

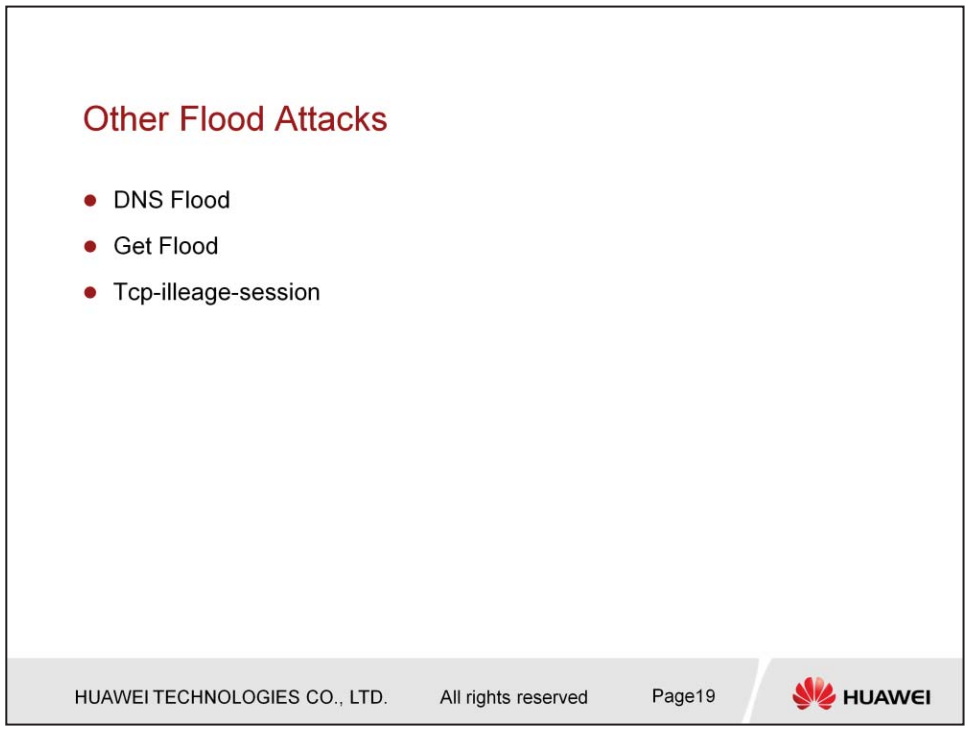
source-max-rate rate-number: Sets the threshold of the UDP packet rate of the source IP protection object. When IP or security zone-based attack defence is enabled, the command is valid and the source rate is

restricted. That is, when the number of UDP packets in the interval-value time exceeds the threshold, an attack occurs. The value ranges from 1 to 10,000. The default value is 300.

interval *interval-value*: Sets a time interval to collect statistics of UDP packets. The value ranges from **1** to **60** in seconds. The default value is **5** seconds.

vpn-instance *vpn-instance-name*: Sets a protected VPN instance.

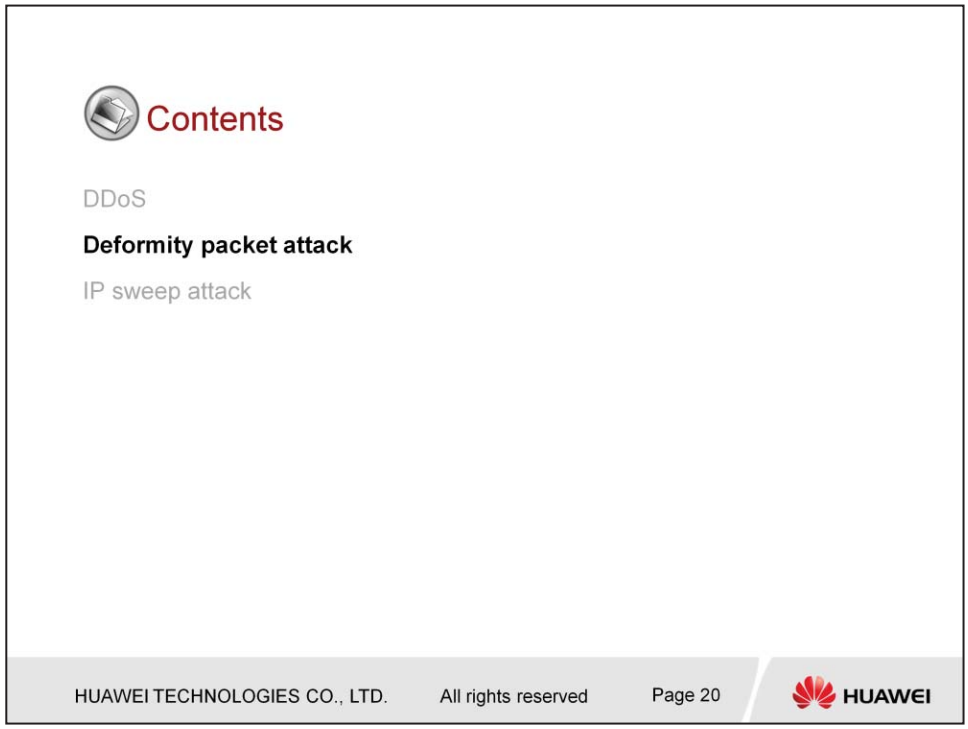
The configuration of ICMP Flood attack defense is somewhat similar to the UDP Flood attack defense.

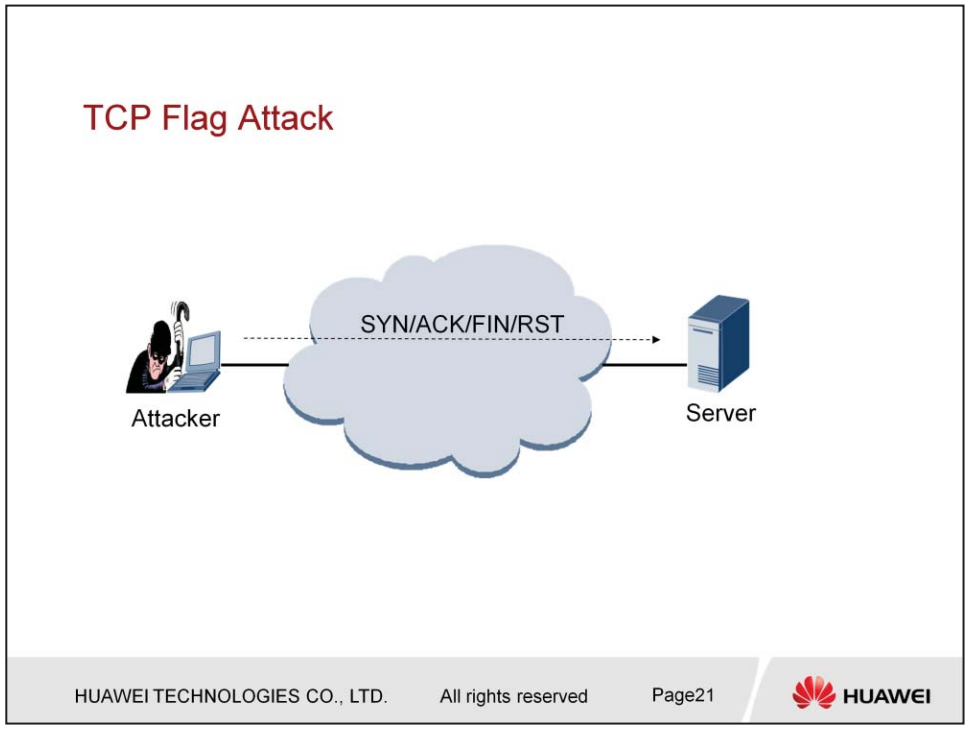


DNS flood used to attack DNS server.

Get flood used to attack the WEB server.

TCP-illeage-session used to use up the TCP connection of server.


Contents
DDoS
Deformity packet attack
IP sweep attack
HUAWEI TECHNOLOGIES CO., LTD.
All rights reserved
Page 20
HUAWEI

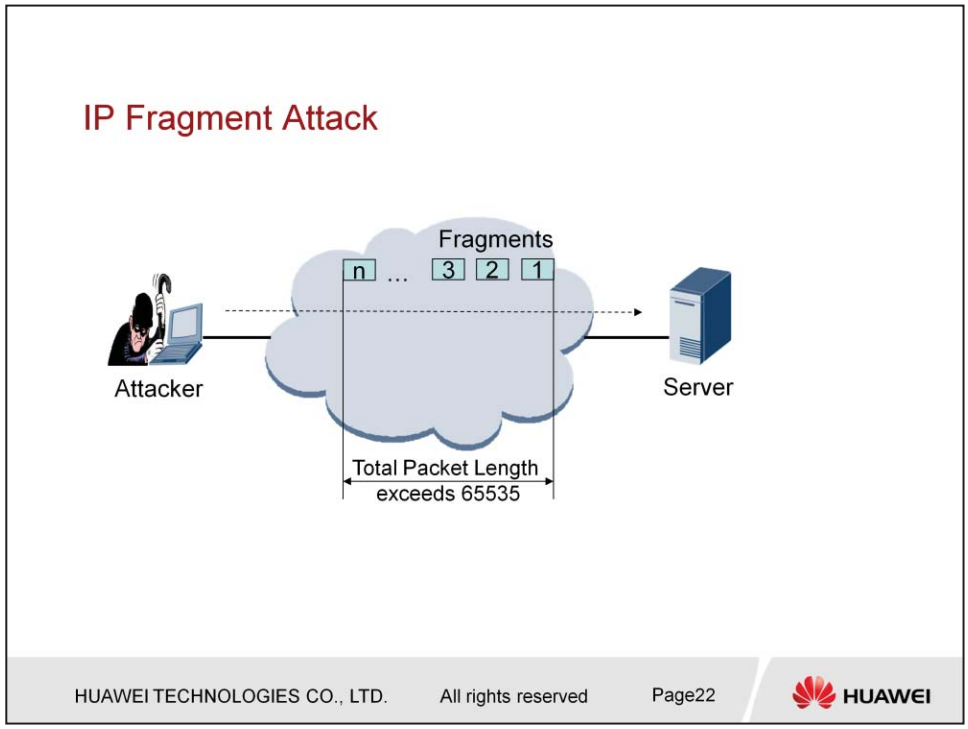


Feature: All settable flags of the packet are set in a conflicting way. For example, SYN, FIN, and RST are set at the same time.

Purpose: Cause the attacked host to stop responding for processing error.

Configuration: **firewall defend tcp-flag enable**

Prevention: Discard the packets with the above feature.

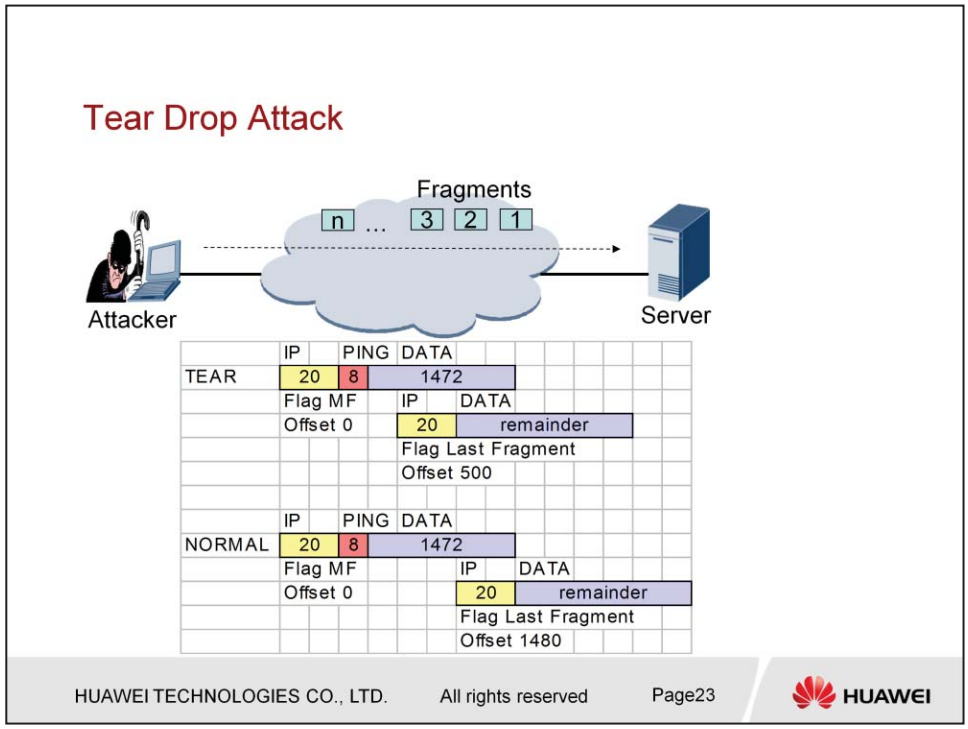


Feature: DF and MF flags are set at the same time, or the sum of offset value plus packet length exceeds 65535.

Purpose: Cause the attacked equipment to stop responding due to improper processing.

Configuration: **firewall defend ip-fragment enable**

Prevention: Discard the packets with the above feature.

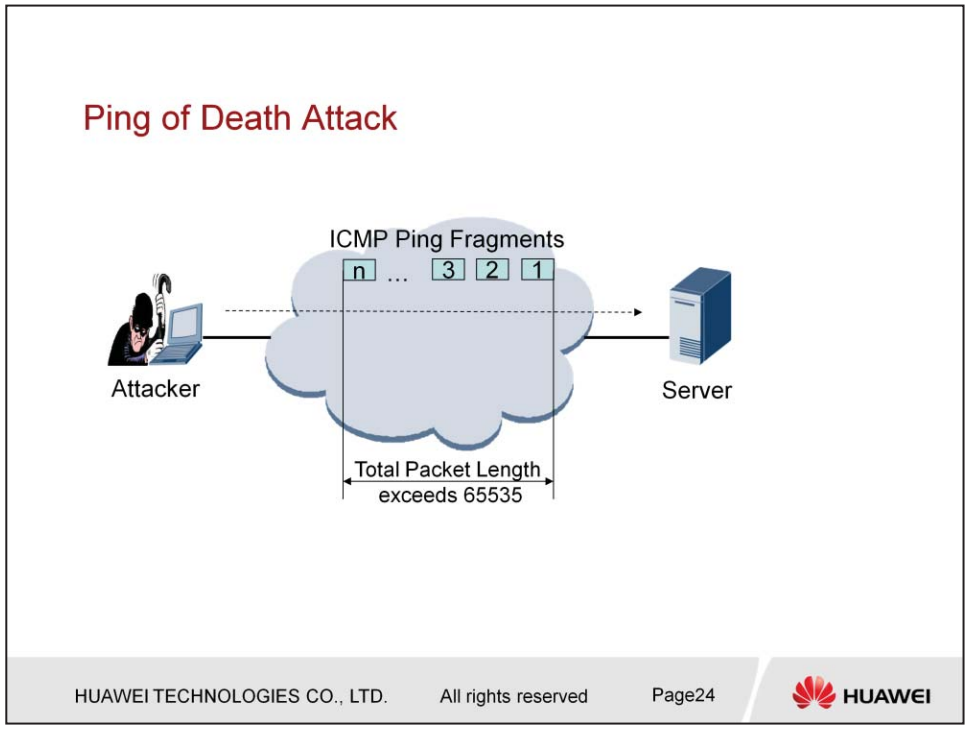


Feature: Overlapping of fragment packets.

Purpose: Causes the attacked equipment stops responding due to improper processing from causing the packet fragments to overlap.

Configuration: **firewall defend teardrop enable**

Prevention: Firewall establishes data structure for fragment packets and records the offset values of the fragment packets that pass the firewall, so that the packet will be discarded upon overlapping.



Ping of death is one type of IP fragment attacks, it uses icmp ping packets.

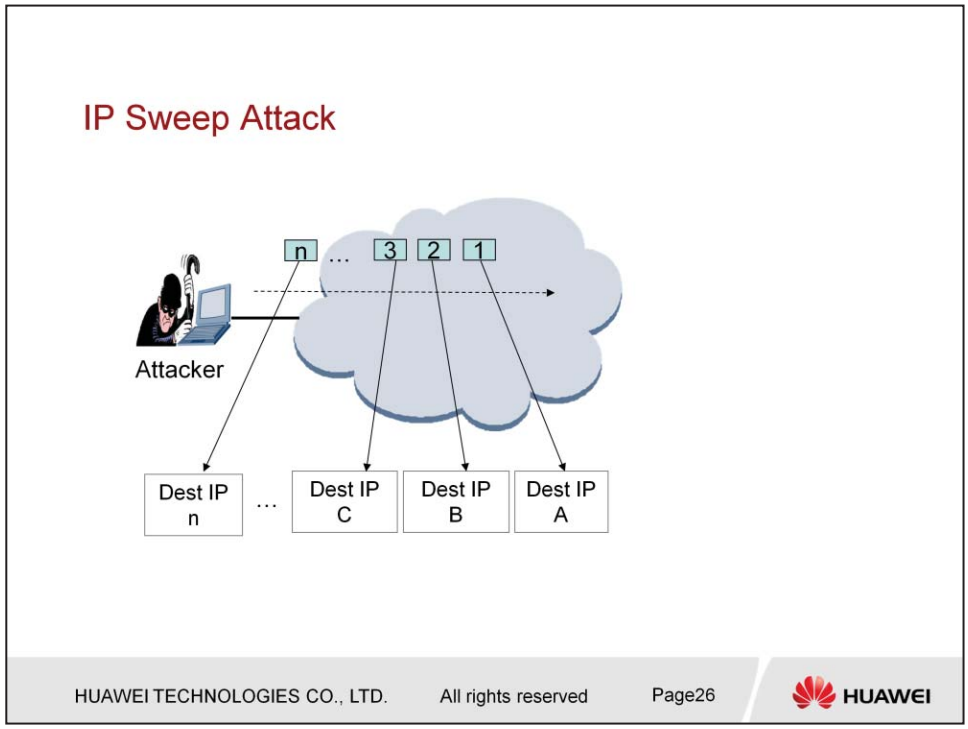
Feature: The entire ping packet is longer than 65535.

Configuration: **firewall defend ping-of-death enable**

Purpose: Causes the attacked equipment to stop responding duimproper processing.

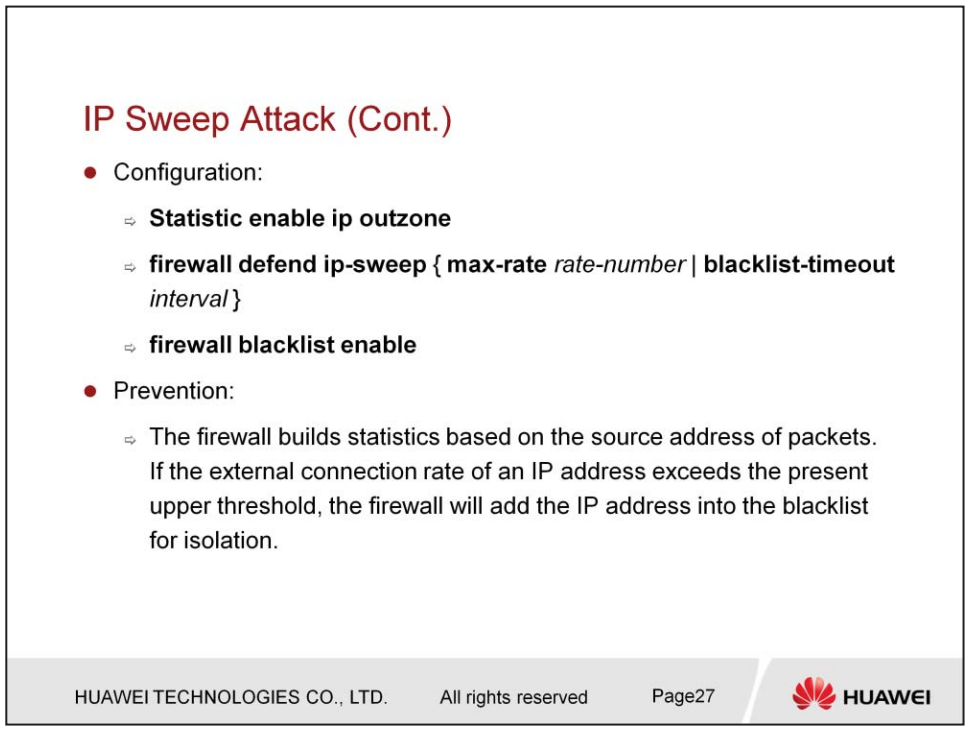
Prevention: Discard the packets with the above feature.


Contents
DDoS
Deformity packet attack
IP sweep attack
HUAWEI TECHNOLOGIES CO., LTD.
All rights reserved
Page 25
HUAWEI



Feature: IP sweeping or port scanning attack detects the IP addresses and ports of the target hosts by using scanning tools.

Purpose: The attacker is able to determine the hosts that exist on the target network according to the response and find the ports that are providing services.

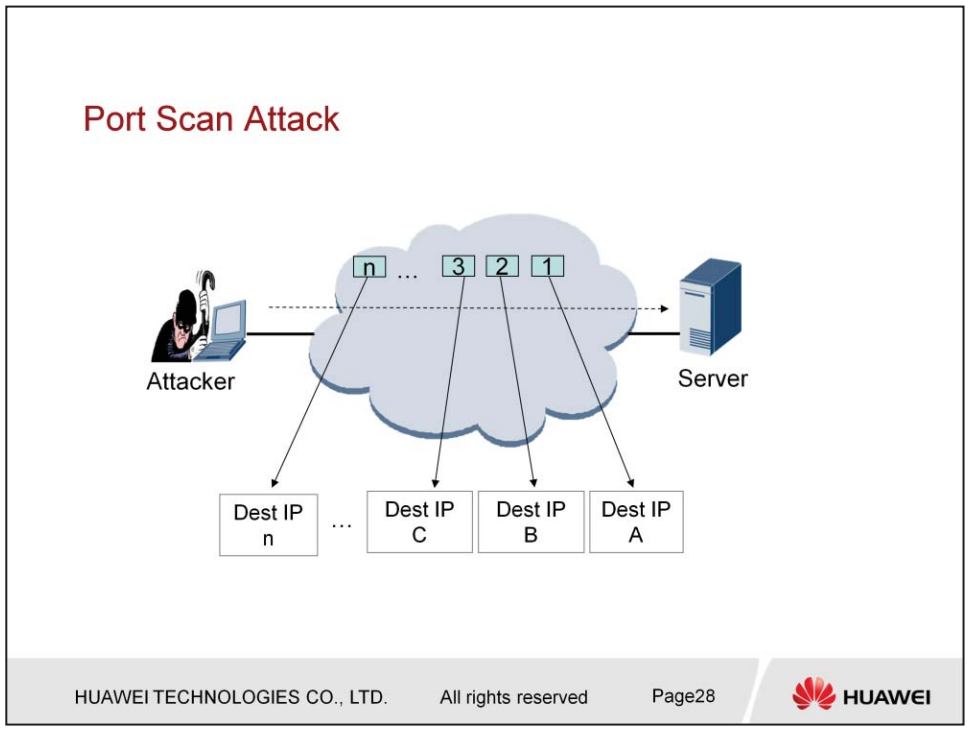


max-rate *rate-number*: Specifies the threshold for destination address changing rate of packets sent from the same source address. The default value of *rate-number* is 4000 times per second. The number ranges from 1 to 10,000 times per second.

blacklist-timeout *minutes*: Add the source address into the blacklist and set valid time for it. *minutes* is measured in minutes, ranging from 1 to 1000 minutes. The default value is 10.

Note: To enable the blacklist filter function, enable the blacklist first.

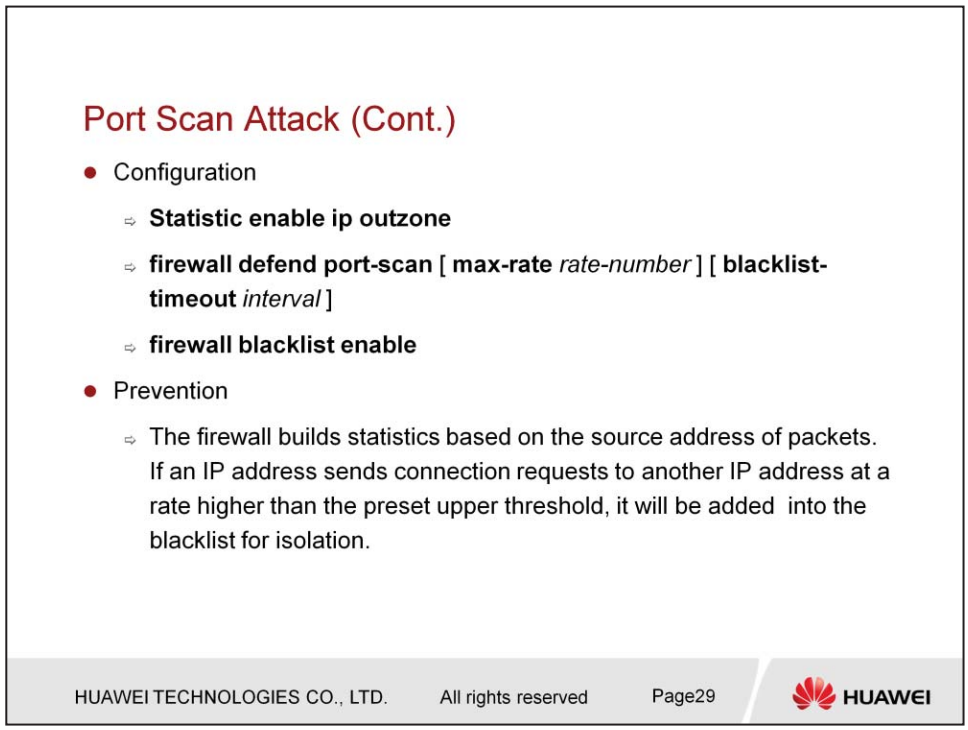
Configuration: **firewall blacklist enable**



Address or port scan attack is used to detect the target address and port to make sure the active system is connected with the target network. Its aim is to receive responses from the system and the port through which the host provides services.

Feature: Connection requests are sent to different ports

Purpose: Determine open services of the scanned host to prepare for subsequent attack.



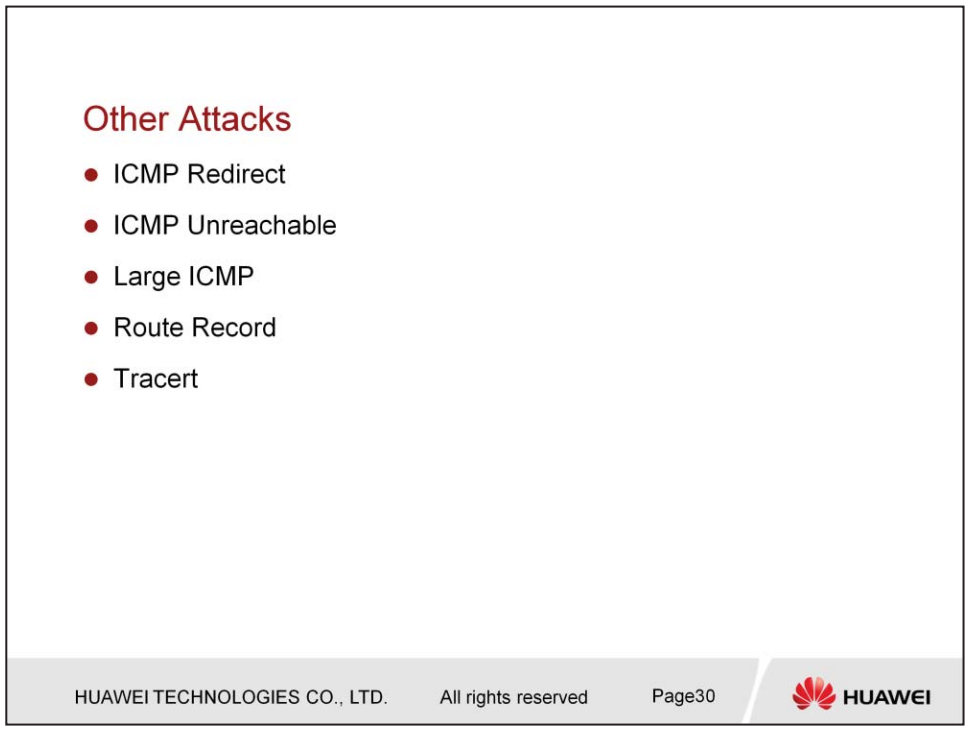
max-rate *rate-number*: Specifies the threshold for destination port changing rate of packets sent from the same source address. The default value of *rate-number* is 4000 times per second. The number ranges from 1 to 10,000 times per second.

blacklist-timeout *minutes*: Add the source address into the blacklist and set a valid time for it.

minutes is measured in minutes, ranging from 1 to 1000 minutes. The default value is 10.

Note: To enable the blacklist filter function, it is necessary to enable the blacklist first.

Configuration: **firewall blacklist enable**



ICMP Redirect Packet Attack

An ICMP redirect attack is similar to the ICMP unreachable attack.

A network device can send ICMP redirect packets to a host in the same subnet asking the host to change its routes.

Based on this mechanism, a malicious attacker can send a pseudo redirect packet to a target host in different network segments to change the host's routing table and disturb the process of sending packets.

ICMP Unreachable Packet Attack

After certain systems receive an ICMP packet that specifies that the network (code of 0) or the host (code of 1) is unreachable, they consider subsequent packets sent to this destination as unreachable.

Large ICMP packets attack

Generally speaking, the size of the ICMP packets is small, ICMP packets of a large size may cause errors for some operating systems.

Route-Record Packet Attack

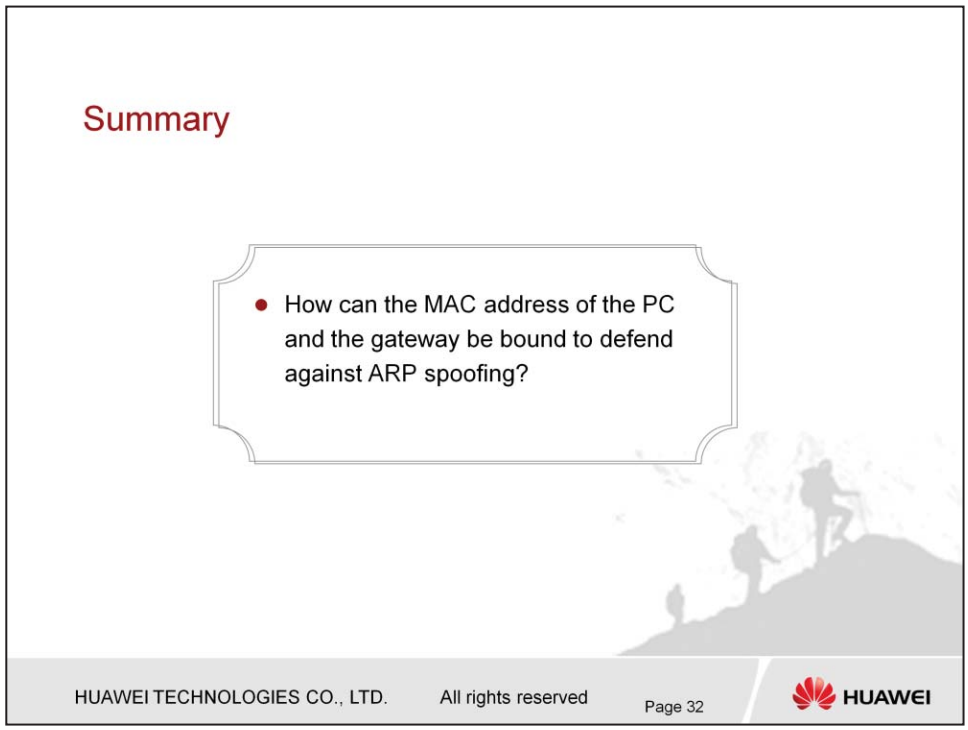
A network administrator uses the following kinds of IP packets to test or debug a network:

Route record is used to record the packets from the routers that pass through before arriving at the destination. These packets, however,

may be used by attackers to viciously probe the topology of a network. This results in potential risks.

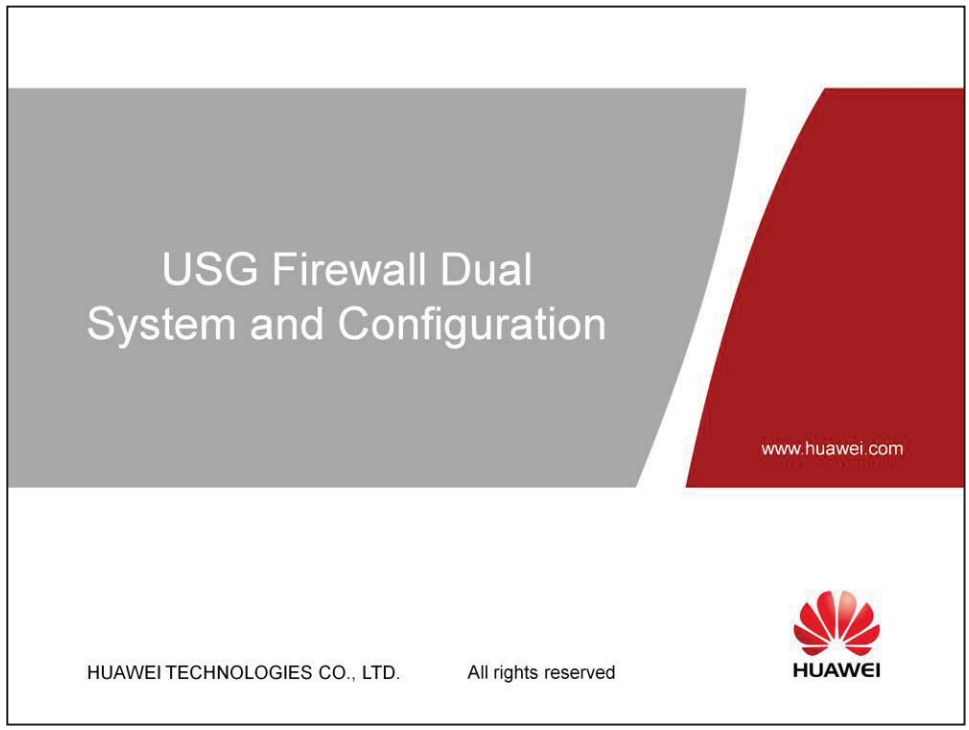
Tracert Attack

Tracert involves discovering the path of packet transmission based on ICMP timeout packets that are returned when Time To Live (TTL) value is 0, or on the returned ICMP port-unreachable packet. Based on this mechanism, an attacker can probe the architecture of a network. This results in potential risks.



How can the MAC address of the PC and the gateway be bound to defend against ARP spoofing?
Answer:

If for windows system, use command: **arp -s X.X.X.X XX-XX-XX-XX-XX-XX**, manually bind the gateway's IP and MAC, while you should know that if the system restarts, the bindngs will be lost, and should be reapplied.


USG Firewall Dual System and Configuration
www.huawei.com
HUAWEI TECHNOLOGIES CO., LTD.
All rights reserved
HUAWEI

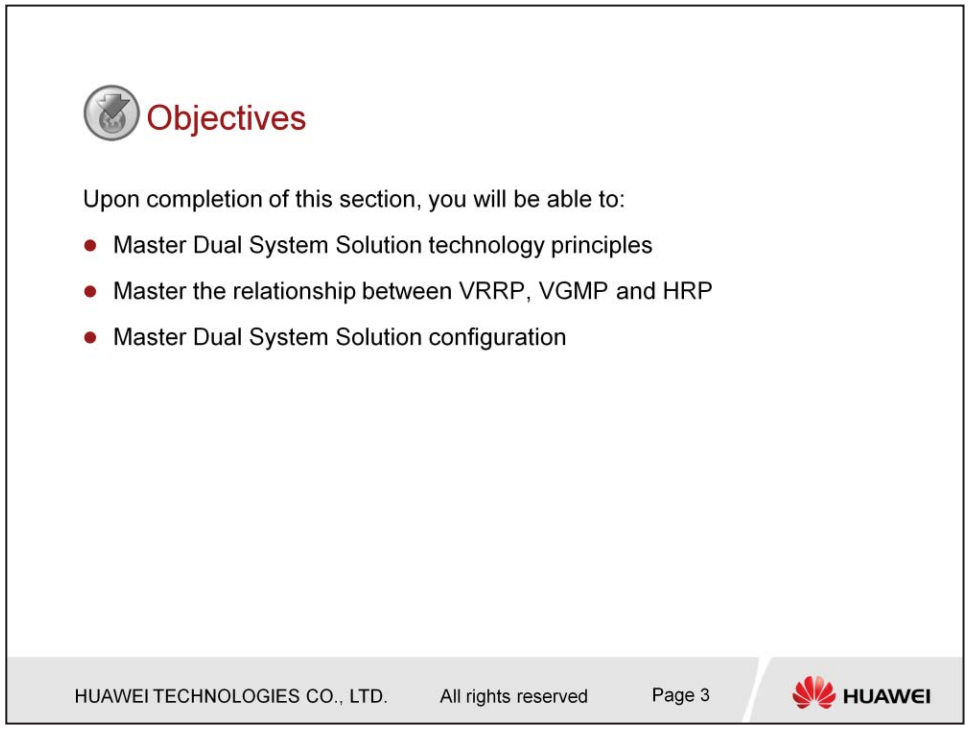


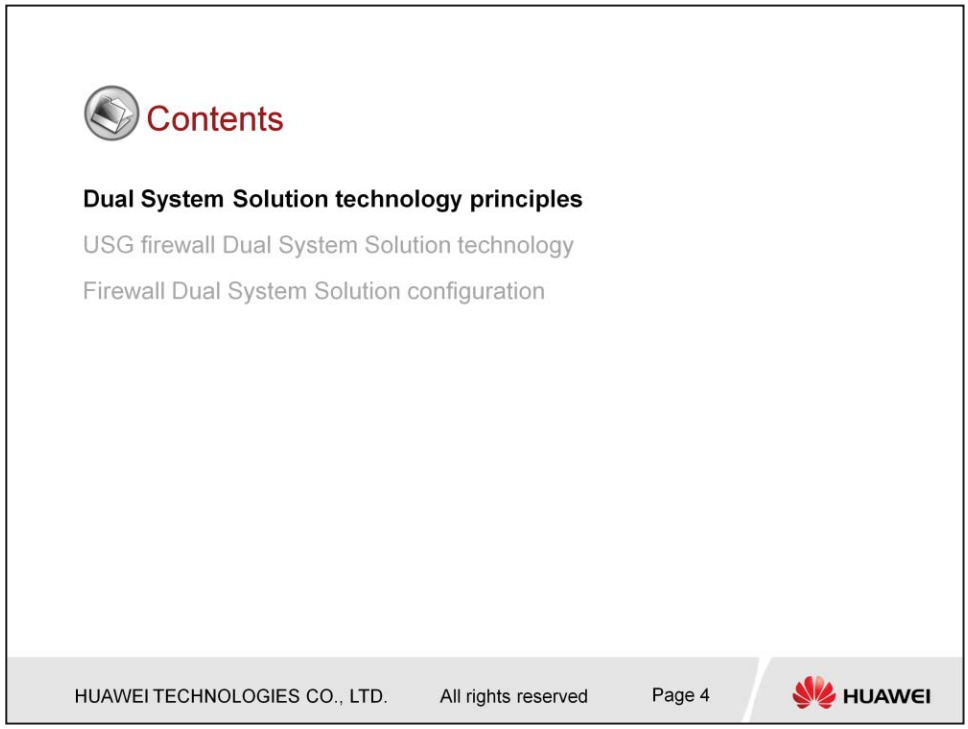
Foreword

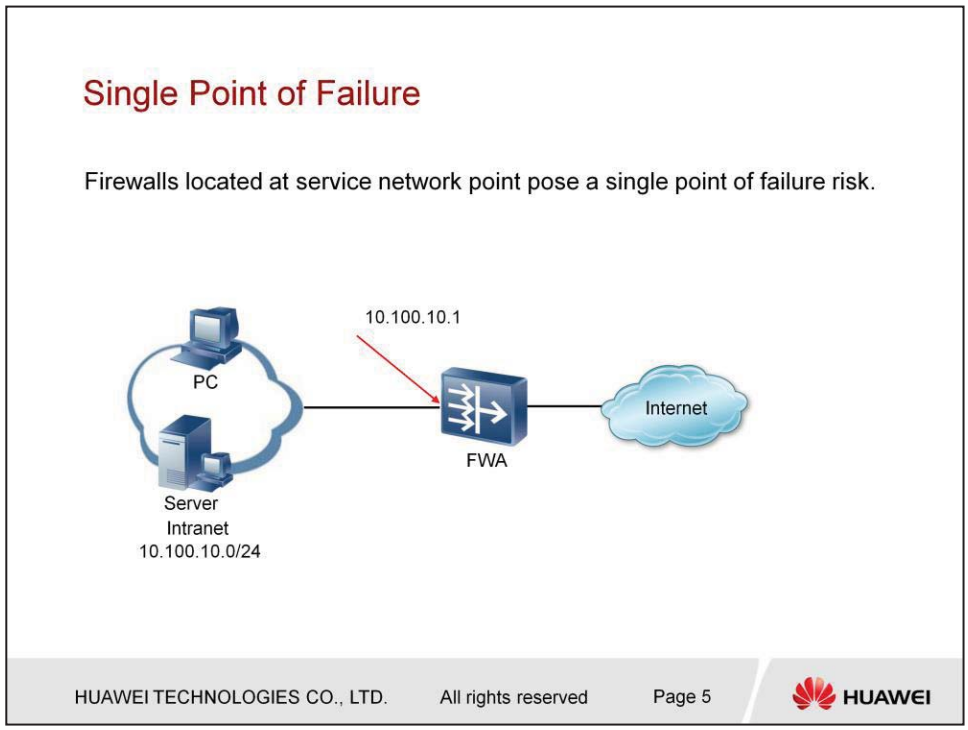
This section mainly introduces the firewall Dual System Solution technology principles and configuration, and as part of the USG firewall, three supporting protocols : VRRP, VGMP and HRP.

HUAWEI TECHNOLOGIES CO., LTD. All rights reserved



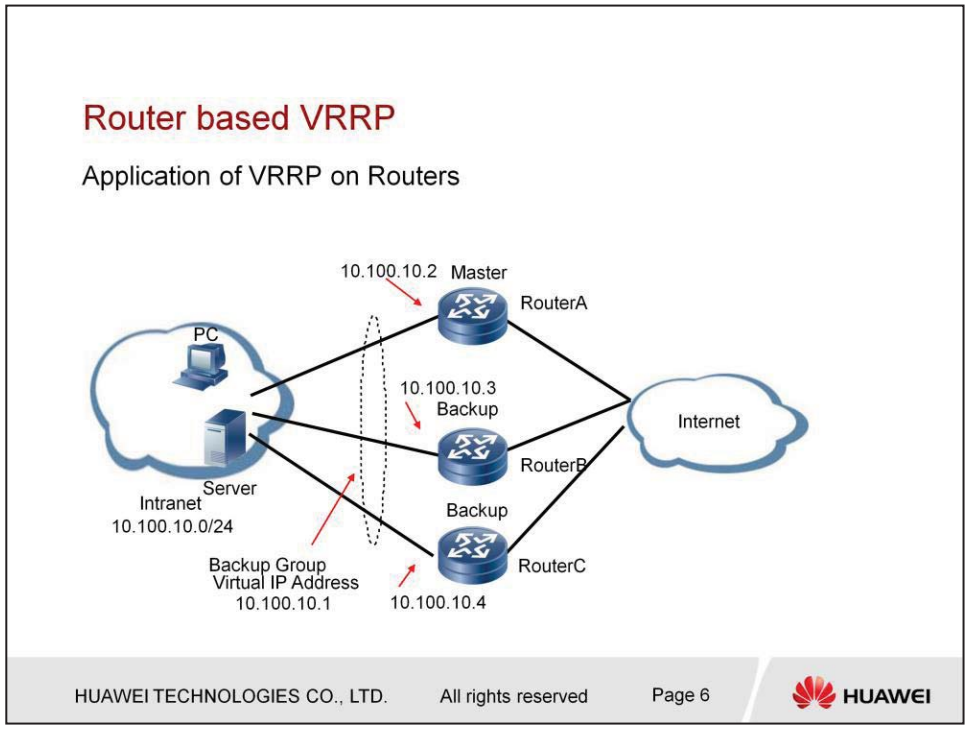

Objectives
Upon completion of this section, you will be able to:
Master Dual System Solution technology principles
Master the relationship between VRRP, VGMP and HRP
Master Dual System Solution configuration
HUAWEI TECHNOLOGIES CO., LTD.
All rights reserved
Page 3
HUAWEI


Contents
Dual System Solution technology principles
USG firewall Dual System Solution technology
Firewall Dual System Solution configuration
HUAWEI TECHNOLOGIES CO., LTD.
All rights reserved
Page 4
HUAWEI



Usually, each host on an internal network is configured with a default route to the next hop, which is the IP address of the egress firewall; that is, 10.100.10.1/24 as shown in Figure above.

The interactive packets between the internal and the external users all pass the firewall. When a firewall fails, all hosts (whose default next hop is firewall) on the internal network fail to communicate with the external networks, therefore communication is unreliable in the default route mode.

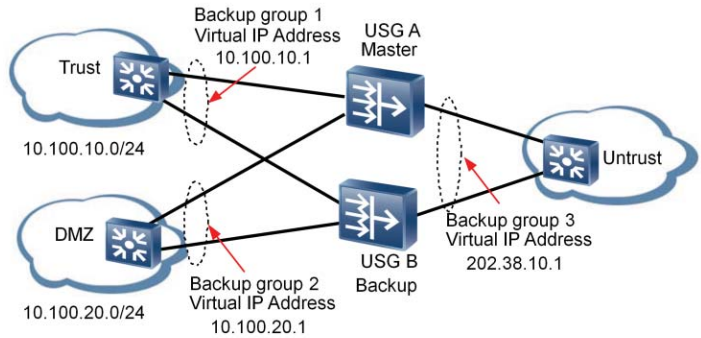


The mechanism of VRRP is to form a virtual router (a VRRP group) out of several real routers to provide a uniform virtual IP address and a virtual MAC address and select a master router according to the priorities. Only the master router can receive and forward the packets whose next-hop IP address is a virtual IP address. The backup router is in the monitoring state. This ensures that only one router is the master router and forwards users' packets. Therefore, if one router in the VRRP group is the master router, links do not break.

For detailed VRRP principle, please refer to VRRP principle section.

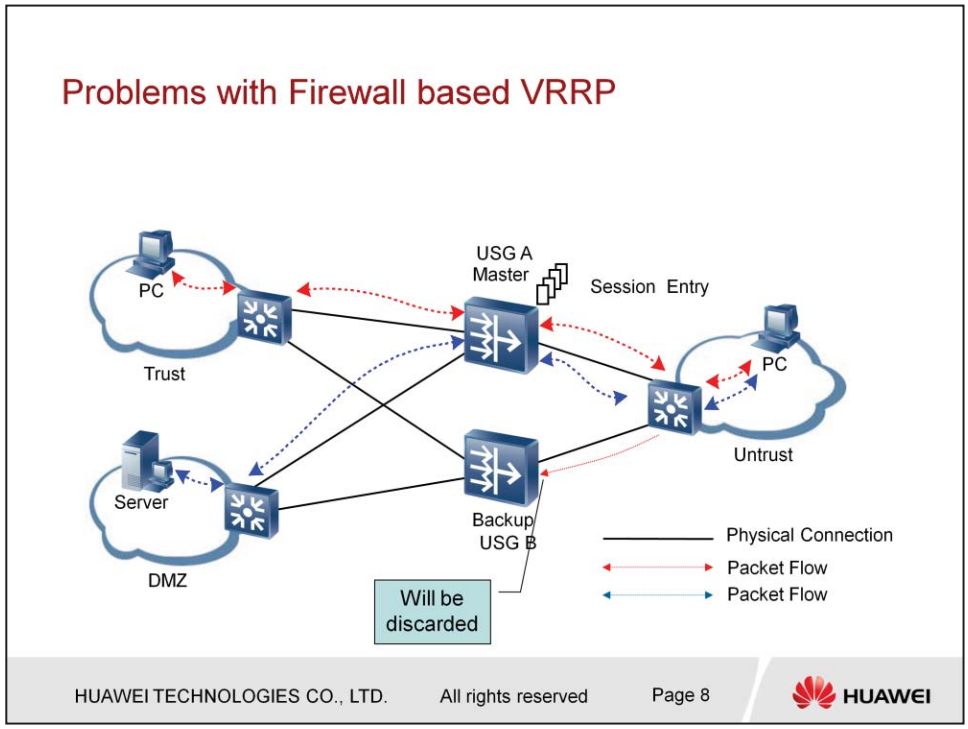
Firewall based VRRP

Firewalls are required to provide redundancy for multiple security zones, therefore require configuration of multiple VRRP backup groups.

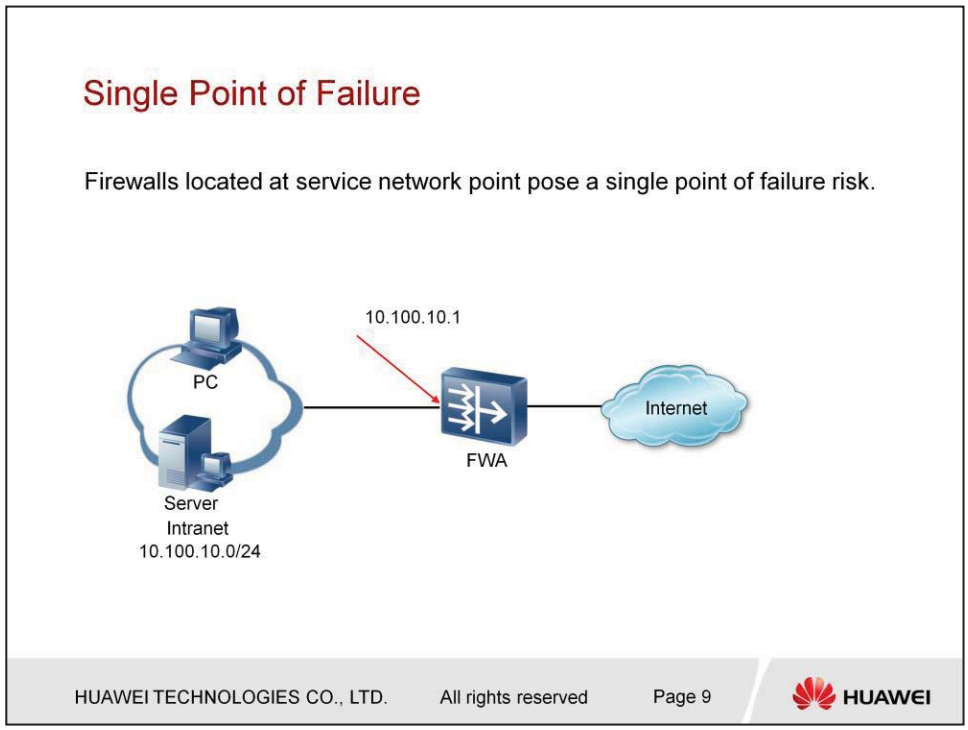


HUAWEI TECHNOLOGIES CO., LTD. All rights reserved Page 7 HUAWEI

For this figure, we have three security zones, and we need three VRRPs. For each security zone provide one virtual gateway IP address.

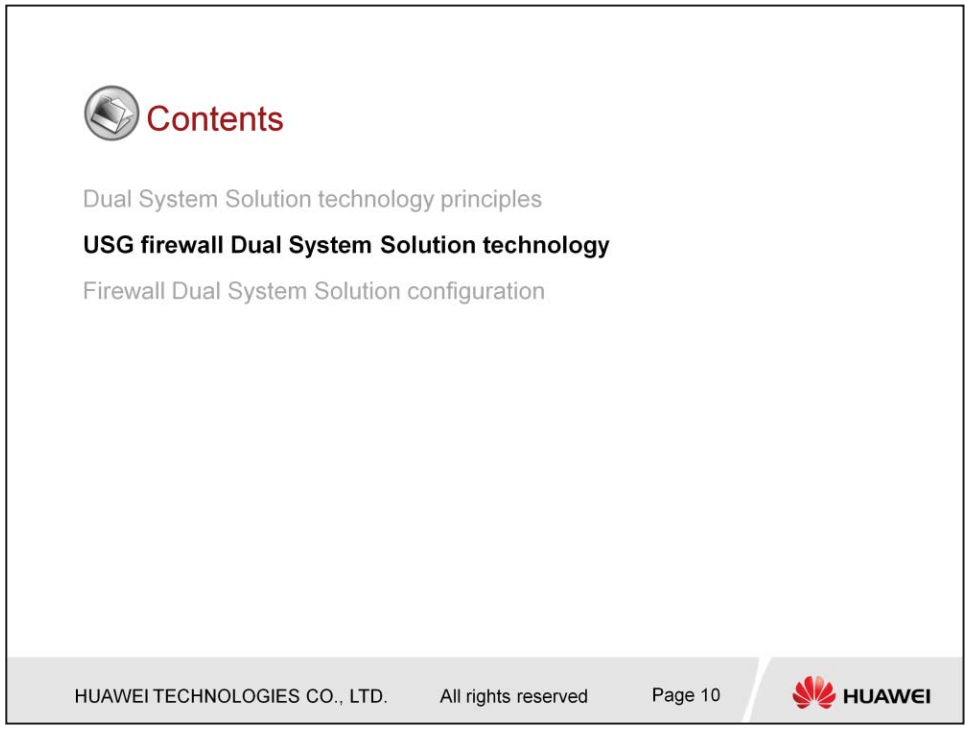


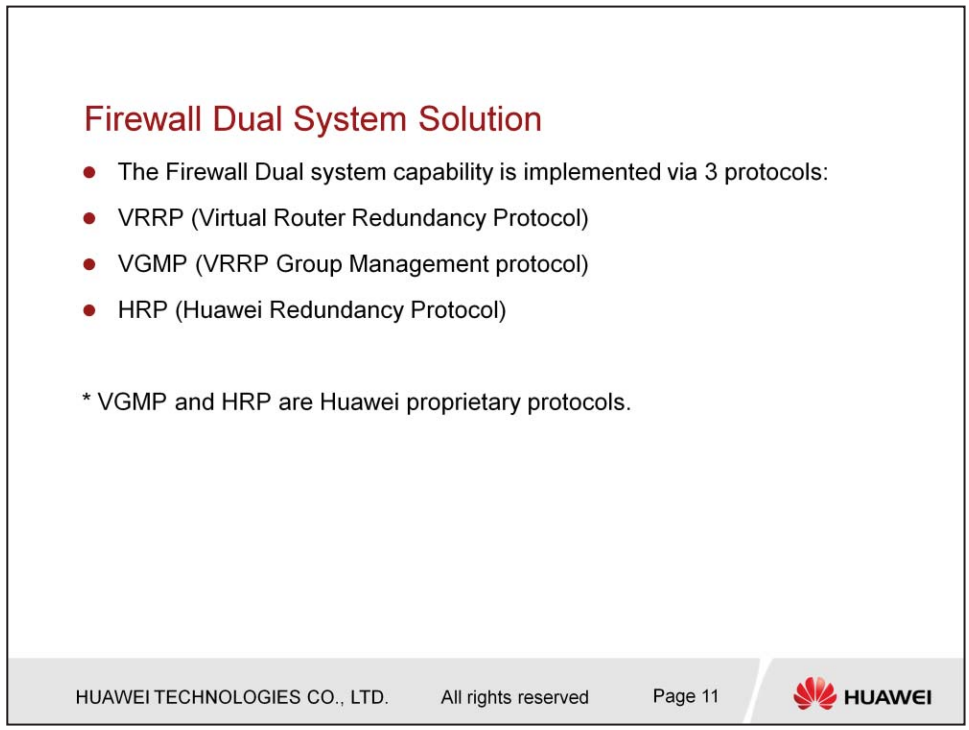
Because the USG firewall is a stateful firewall, it checks only the first packet of a session traffic, and a session entry will be created dynamically. Only the subsequent packets and return packets that match with the session entry can pass through the USG. Therefore, when the inbound path and the outbound path of the same session are inconsistent, the USG drops the subsequent packets. As shown in Figure, the return packets to USG B will be dropped.



Usually, each host on an internal network is configured with a default route to the next hop, which is the IP address of the egress firewall; that is, 10.100.10.1/24 as shown in Figure above.

The interactive packets between the internal and the external users all pass the firewall. When a firewall fails, all hosts (whose default next hop is firewall) on the internal network fail to communicate with the external networks, therefore communication is unreliable in the default route mode.


Contents
Dual System Solution technology principles
USG firewall Dual System Solution technology
Firewall Dual System Solution configuration
HUAWEI TECHNOLOGIES CO., LTD.
All rights reserved
Page 10
HUAWEI

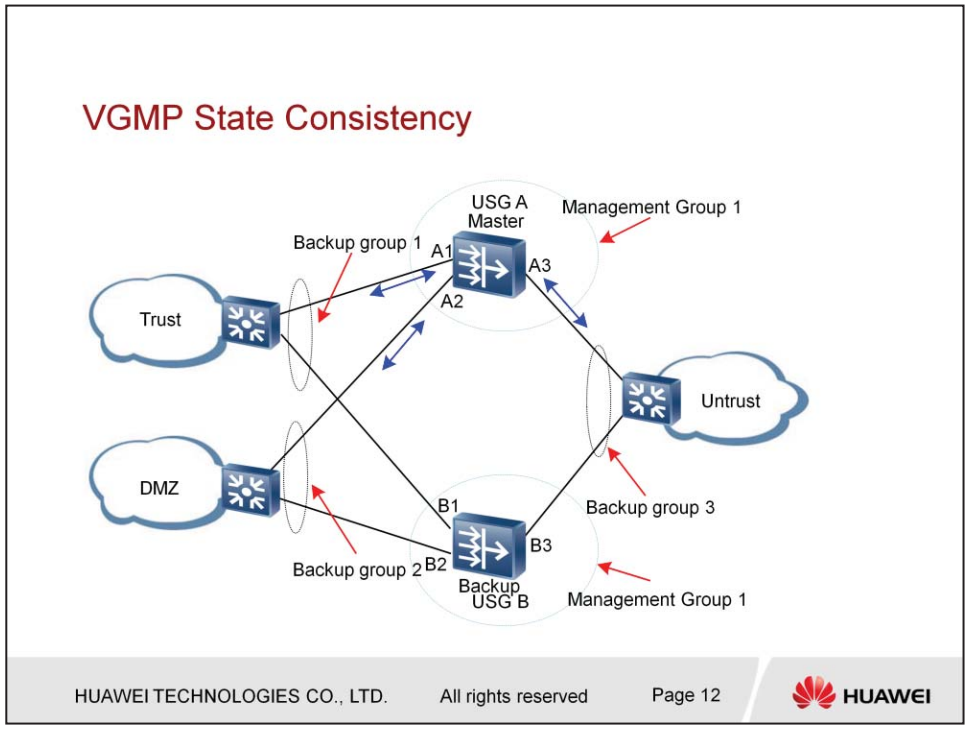


Three protocols work together, their main functions are as follows:

VRRP is used to provide a virtual gateway IP address for security zone.

VGMP used to control status consistency of VRRPs, to make them follow the status of VGMP. If VGMP is master, all VRRPs are master, if VGMP is backup, all VRRPs are backup.

HRP used to transfer the session and configuration information from master firewall to backup firewall.



Based on the traditional VRRP mechanism, VRRP in each backup group works in an independent state. Therefore, the state of VRRP on each interface on the same USG cannot maintain consistency. That is, the traditional VRRP cannot achieve VRRP state consistency with USG.

In this case, Huawei developed a protocol based on VRRP, named the VRRP Group Management Protocol (VGMP).

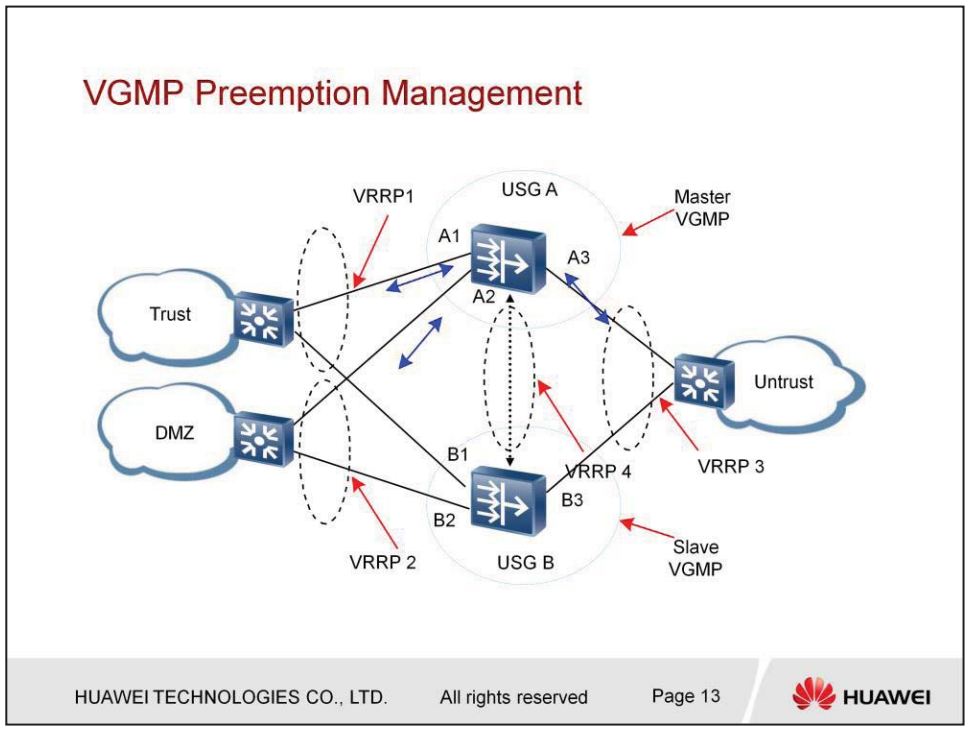
With the VGMP mechanism, the following management can be fulfilled over multiple VRRP backup groups (the virtual firewall):

State consistency management

Preemption management

Channel management

VGMP ensures the state of interfaces on one USG keeps consistency, to be master or backup at the same time, and achieves VRRP state consistency with USG.



The VRRP management groups on the master and the backup exchange information about the operational states of the firewalls through VGMP packets. In this way, the firewalls can maintain stability of their states and perform switchover between them when necessary.

VGMP packets are evolved from VRRP packets. VGMP packets mainly include Hello messages, state switchover request messages, response messages allowing state switchover, and response messages denying state switchover.

Hello message

Similar to VRRP groups, the VGMP group on the master regularly sends hello messages to the backup to inform the backup of the master's operational state, including the priority of the master and the states of the VRRP backup groups. When the backup receives a hello message, it returns an acknowledgement (ACK) message, which contains the priority of the backup and the states of the VRRP groups on the backup. The two firewalls exchange their state information through Hello messages.

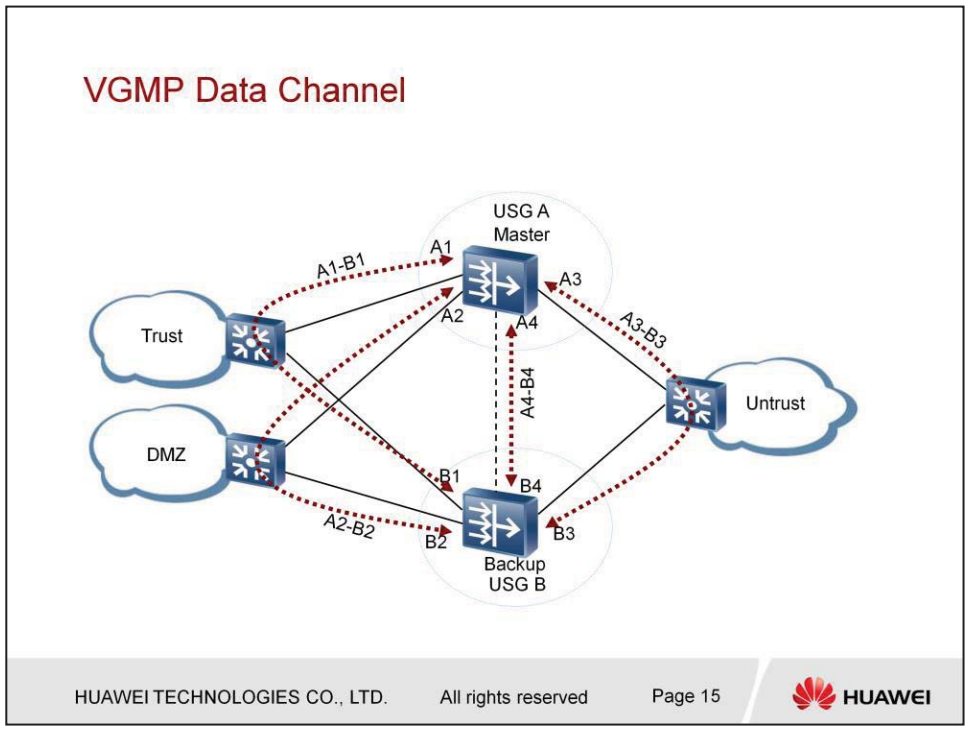
By default, the interval for sending VGMP hello messages is 1 second. When the backup fails to receive a hello message from the master after three hello message periods, it deems that the master has failed, and then becomes the master.

State switchover message

When a member in the backup group on the master fails, the VRRP management group can immediately sense the failure. In this case, the VRRP management group adjusts its own priority and sends a state switchover message to the peer. When the peer receives the message, it compares the priority in the message and its own priority.

If its own priority is higher than that in the message, the peer returns an ACK message to allow switchover and at the same time becomes the master. When the faulty firewall receives the ACK message, it immediately becomes the backup. When the state of the VRRP management group is switched, the states of all VRRP backup groups in the VGMP group are switched accordingly. If the priority of the peer is lower than that in the message, a NACK message is returned to deny state switchover, then the states of both firewalls do not change.

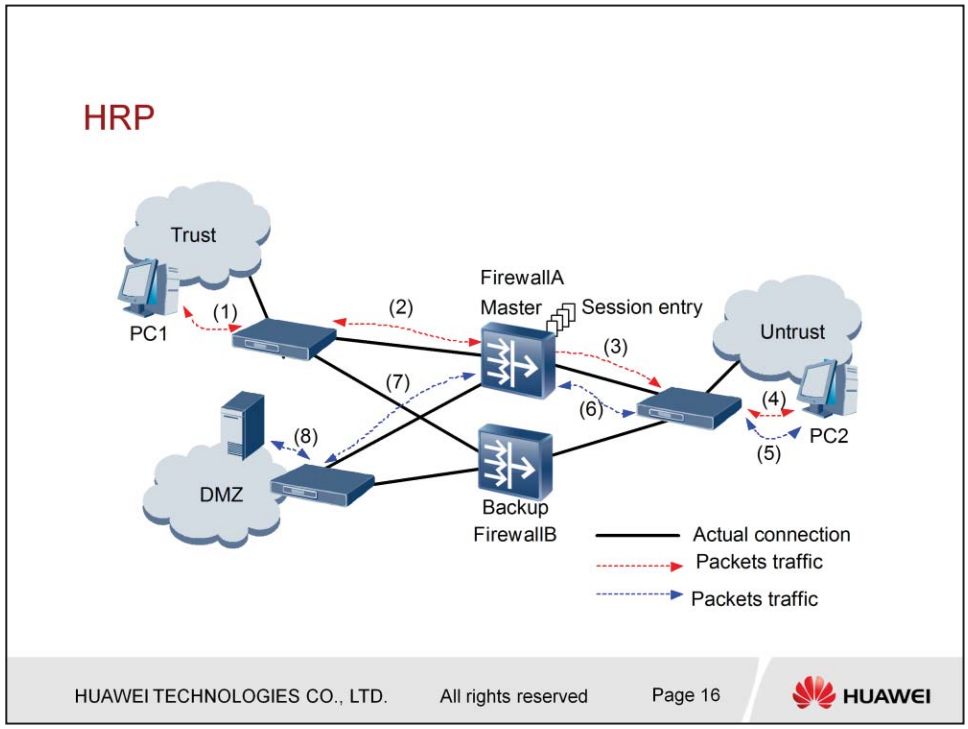
The VGMP group can proactively send a state switchover message immediately after it senses the fault on the interface. State switchover does not rely on timeout of three Hello messages. Therefore, the response to faults is much quicker.



The data transport of VGMP and HRP protocol, both choose the VGMP data channel, VGMP will choose some of its VRRP member link as its data channel.

From the figure above, it can have four data channel for VGMP.

The channel A4-B4, usually we will make it as a heart beat link. This channel can be set with parameter “transfer-only” to specify this link should act as heart beat link. For VGMP and HRP, it will choose such heart beat link channel as data channel preference, and such link usually are only used to transport VGMP and HRP traffic, no business traffic. So when such a heart beat link is down, it will not cause VGMP’s status change, because VGMP can choose other VRRP’s link to act as data channels. This feature aims to improve the stability of the system.

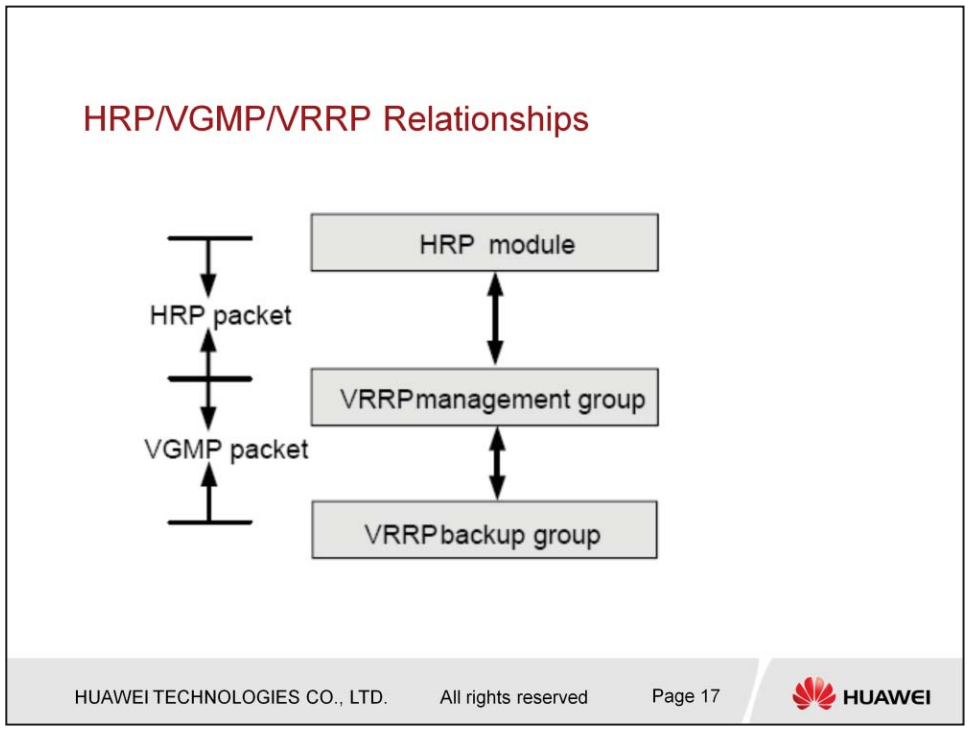


In master/backup mode, if Firewall A is the master device, it takes up all data transmission task and many dynamic session entries are set up on it; firewall B is the backup device, no data passes through it.

When errors occur on Firewall A or on associated links, firewall B will become the master firewall and begin to transfer data. However, if there is no backup session entry and configuration command on firewall B before the state switch, all sessions that have passed through firewall A before will be disconnected as a result of mismatch and services will be interrupted.

In order to make the backup firewall smoothly take over work from the master firewall when the master firewall breaks down, backup of the configuration commands and state information between the master firewall and the backup firewall are necessary.

Huawei Redundancy Protocol (HRP) is developed for this purpose. HRP is transmitted over VGMP packets on data channels in VRRP management group.



When the state of the VRRP management group changes, the system notifies HRP and the primary or secondary configuration device to change their states. In addition, the state of the VRRP management group is affected by the HRP state. When the state of the VRRP backup group changes, the VRRP management group determines whether to change the states of the following elements:

VRRP management group

HRP

Primary and secondary configuration devices



Contents

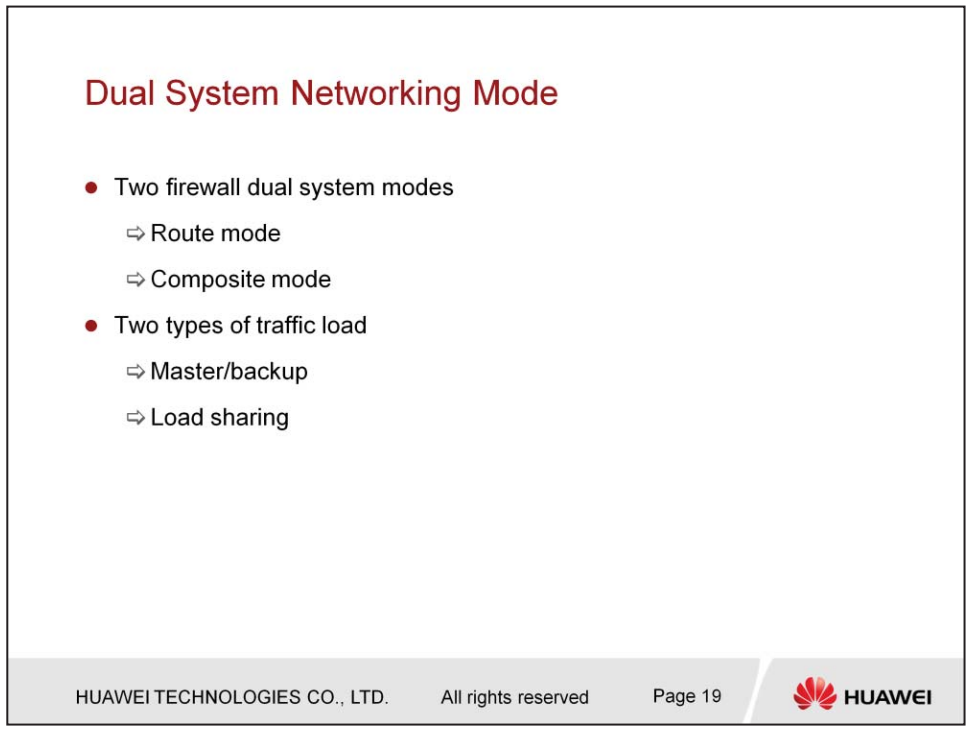
Dual System Solution technology principles

USG firewall Dual System Solution technology

Firewall Dual System Solution configuration

HUAWEI TECHNOLOGIES CO., LTD. All rights reserved



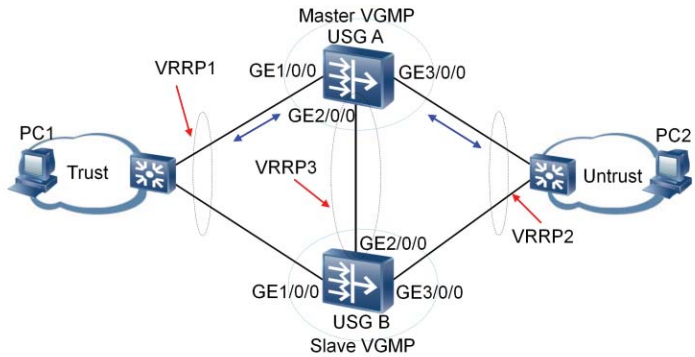


The VGMP mechanism can help realize active/standby backup.

Load balancing, also called mutual backup, is the process of distributing traffic over transmission lines so that they carry approximately the same density of traffic. The two USG working in load balancing mode are both functional in processing services. When a failure occurs in the master firewall, the backup firewall immediately serves as the master one and traffic is diverted to that firewall, ensuring service continuity and availability.

Since the two firewalls act as backup device for each other, the effect of the load balancing mechanism is equivalent to running two sets of master/backup mechanism. Therefore, there are two VRRP managements groups on each firewall. The following example assumes that the Master management group on the USG A and the Slave management group on the USG B constitute a master/backup mechanism, and the slave management group on the USG A and the Master management group on the USG B form the other set of master/backup mechanism.

Route Mode-Master/Backup



USG	VGMP	VRRP Member	Priority	Status	Traffic
A	Master	VRRP1，2，3	High	Master	100%
B	Slave	VRRP1，2，3	Low	Backup	0

HUAWEI TECHNOLOGIES CO., LTD. All rights reserved

USG A acts as master、USG B acts as backup, both working under route mode.

Networking as follows：

The network segment that needs to be protected is 10.100.10.0/24, which is connected with firewall interface GigabitEthernet 1/0/0, and belongs to the Trust zone.

External network connected with firewall interface GigabitEthernet 3/0/0, belongs to the Untrust zone

HRP backup interface GigabitEthernet 2/0/0 belongs to the DMZ zone

The connection of Trust and Untrust zone users and the firewall is achieved via a switch.

IP addressing is as follows：

The VRRP IP address of Trust is 10.100.10.1.

The VRRP IP address of Untrust is 202.38.10.1.

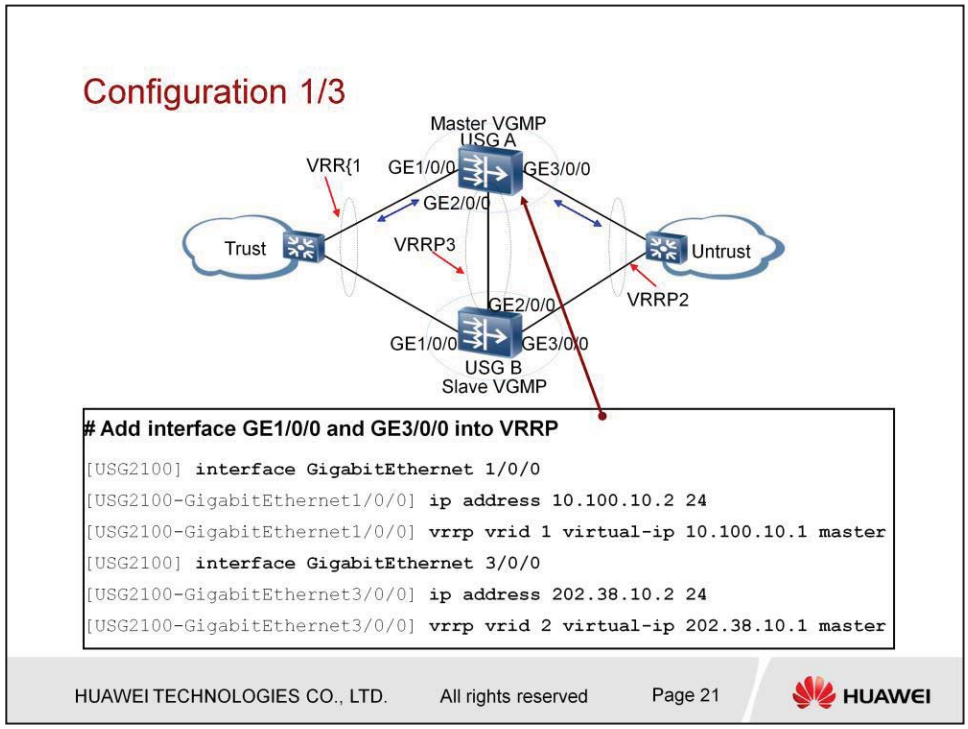
The VRRP IP address of DMZ is 10.100.20.1.

USG A

GE1/0/0： 10.100.10.2/24； GE2/0/0： 10.100.20.2/24； GE3/0/0： 202.38.10.2/24

USG B

GE1/0/0： 10.100.10.3/24； GE2/0/0： 10.100.20.3/24； GE3/0/0： 202.38.10.3/24



Add interface GigabitEthernet 1/0/0 into Trust zone

[USG2100] **firewall zone trust**

[USG2100-zone-trust] **add interface GigabitEthernet 1/0/0**

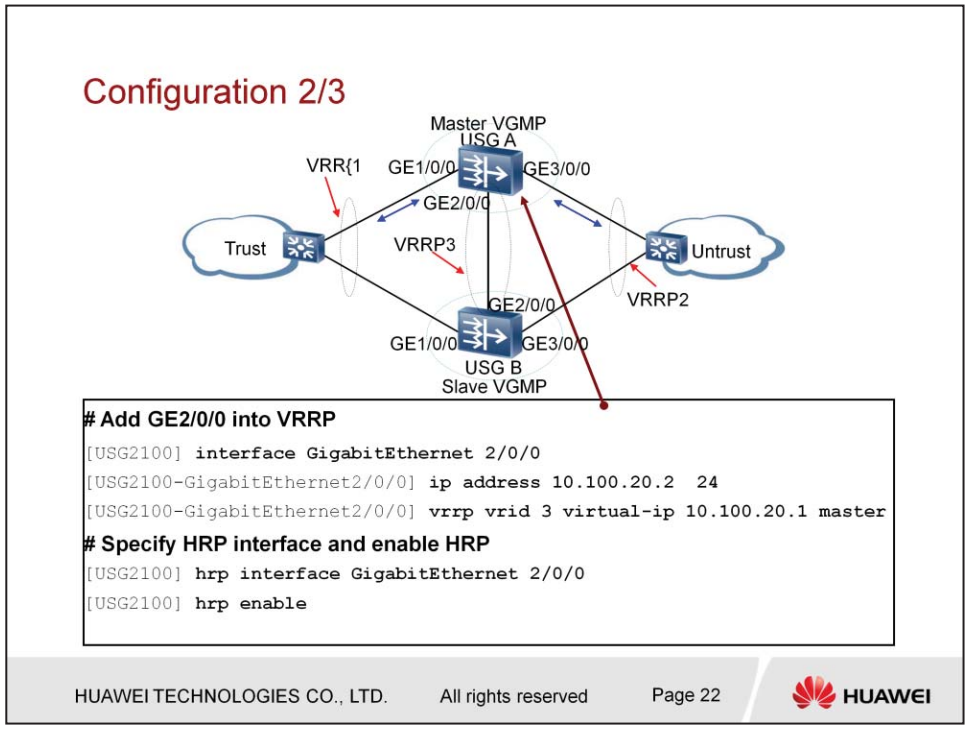
[USG2100-zone-trust] **quit**

Add interface GigabitEthernet 3/0/0 into Untrust zone

[USG2100] **firewall zone untrust**

[USG2100-zone-untrust] **add interface GigabitEthernet 3/0/0**

[USG2100-zone-untrust] **quit**



Add interface GigabitEthernet 2/0/0 to DMZ zone

[USG2100] **firewall zone dmz**

[USG2100-zone-dmz] **add interface GigabitEthernet 2/0/0**

[USG2100-zone-dmz] **quit**

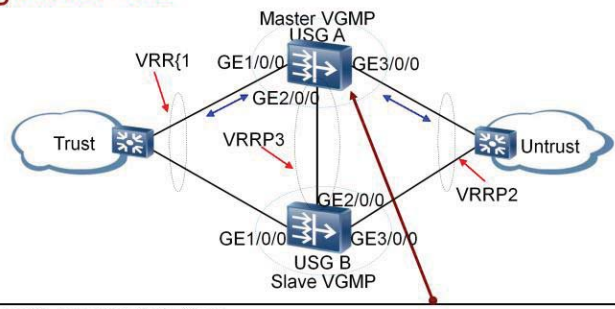
Configure USG B

Most of the configuration of USG B similar with USG A, the difference as follows:

The IP address of interface is different

USG B's VRRP management group should choose Slave

Configuration 3/3



Enable auto backup function

```
HRP_M[USG2100] hrp auto-sync config
```

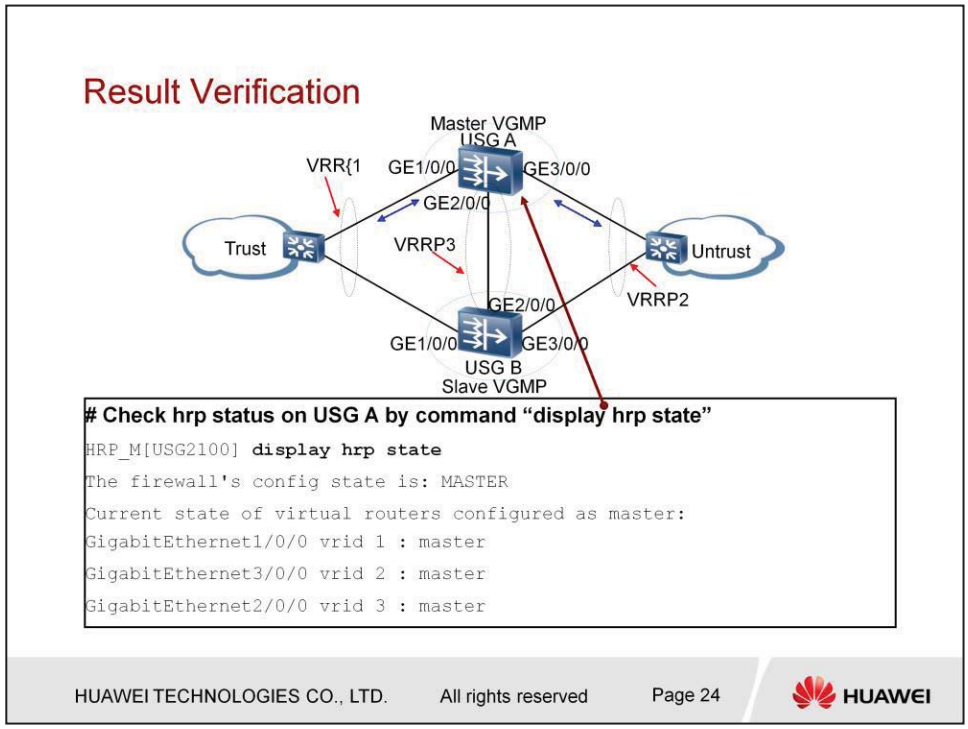
Enable interzone packet filter

```
HRP_M[USG2100] policy interzone trust untrust outbound
HRP_M[USG2100-policy-interzone-trust-untrust-outbound] policy 1
HRP_M[USG2100-policy-interzone-trust-untrust-outbound-1] action permit
HRP_M[USG2100-policy-interzone-trust-untrust-outbound-1] policy source
10.100.10.0 mask 24
```

HUAWEI TECHNOLOGIES CO., LTD. All rights reserved

After USG A and USG B both enable auto HRP and permit the heartbeat communication. The configuration on the master HRP will automatically backed up to the slave HRP firewall, there is no need to configure the slave HRP firewall, as the system forbids the configuration of some system commands.

hrp auto-sync config: This command is the startup configuration command automatic backup. This time through the ‘display current’ it is possible to see in firewall A, the configuration of the ACL synchronization to firewall B.



Check the status of VRRP by using "**display vrrp**" on USG A, the information shows as follows:

HRP_M[USG2100] **display vrrp**

GigabitEthernet1/0/0 | Virtual Router 1

state : Master

Virtual IP : 10.100.10.1

Virtual MAC : 0000-5e00-0101

Primary IP : 10.100.10.2

PriorityRun : 100

PriorityConfig : 100

MasterPriority : 100

Preempt : YES

Delay Time : 0

Timer : 1

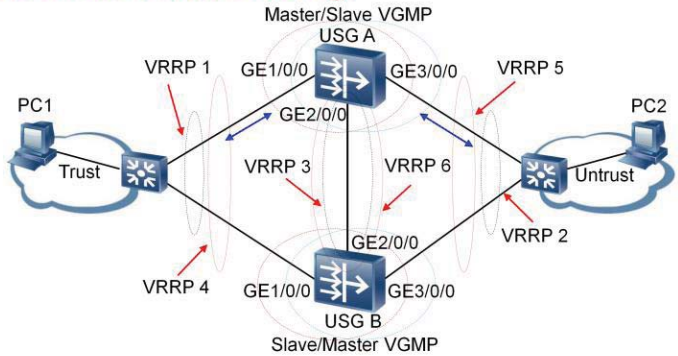
Auth Type : NONE

Check TTL : YES

GigabitEthernet3/0/0 | Virtual Router 2

state : Master
Virtual IP : 202.38.10.1
Virtual MAC : 0000-5e00-0102
Primary IP : 202.38.10.2
PriorityRun : 100
PriorityConfig : 100
MasterPriority : 100
Preempt : YES
Delay Time : 0
Timer : 1
Auth Type : NONE
Check TTL : YES
GigabitEthernet3/0/0 | Virtual Router 2
state : Master
Virtual IP : 202.38.10.1
Virtual MAC : 0000-5e00-0102
Primary IP : 202.38.10.2
PriorityRun : 100
PriorityConfig : 100
MasterPriority : 100
Preempt : YES
Delay Time : 0
Timer : 1
Auth Type : NONE
Check TTL : YES

Route Mode-Load Sharing



Firewall	VGMP	VRRP	Priority	Status	Traffic
A	Master	VRRP1，2，3	High	Master	Proportion
B	Slave	VRRP1，2，3	Low	Backup	0
A	Slave	VRRP4，5，6	Low	Backup	0
B	Master	VRRP4，5，6	High	Master	Proportion

HUAWEI TECHNOLOGIES CO., LTD. All rights reserved

USG A and USG B choose load sharing networking, both working under route mode.

Networking as follows：

The network segment that need to be protected is 10.100.10.0/24, is connected with firewall interface GigabitEthernet 1/0/0 and belongs to the Trust zone.

External network is connected with firewall interface GigabitEthernet 3/0/0 and belongs to the Untrust zone

HRP backup interface GigabitEthernet 2/0/0 belongs to the DMZ zone

The connection of Trust and Untrust zone users and the firewall is achieved via a switch

IP addressing is as follows：

Trust zone VRRP virtual IP address：

VRRP 1：10.100.10.1；VRRP 4：10.100.10.2

Untrust zone VRRP virtual IP address：

VRRP 2 ： 202.38.10.1；VRRP 5：202.38.20.2

DMZ zone VRRP virtual IP address：

VRRP 3：10.100.20.1；VRRP 6：10.100.20.2

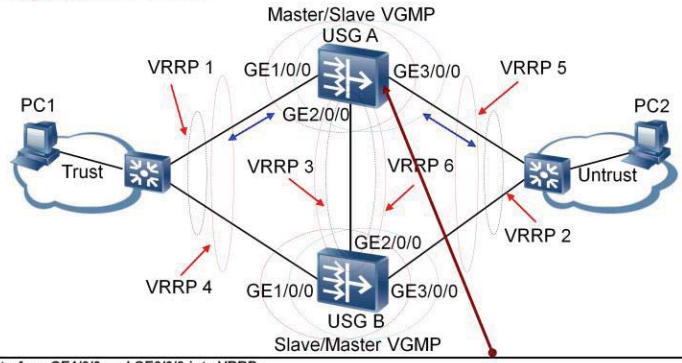
USG A

GE1/0/0： 10.100.10.3/24； GE2/0/0： 10.100.20.3/24；
GE3/0/0：202.38.10.3/24

USG B

GE1/0/0： 10.100.10.4/24； GE2/0/0： 10.100.20.4/24；
GE3/0/0：202.38.10.4/24

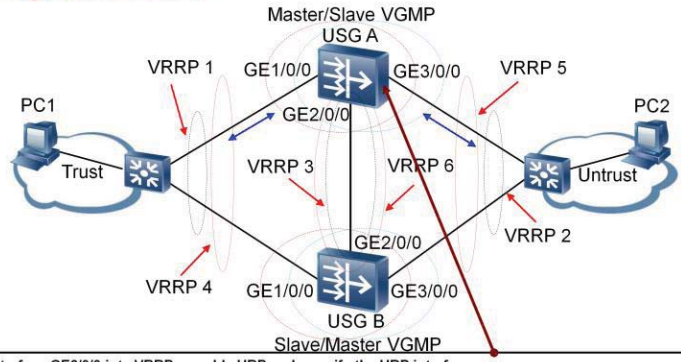
Configuration 1/3




```
# Add interface GE1/0/0 and GE3/0/0 into VRRPs
[USG2100] interface GigabitEthernet 1/0/0
[USG2100-GigabitEthernet1/0/0] ip address 10.100.10.3 24
[USG2100-GigabitEthernet1/0/0] vrrp vrid 1 virtual-ip 10.100.10.1 master
[USG2100-GigabitEthernet1/0/0] vrrp vrid 4 virtual-ip 10.100.10.2 slave
[USG2100] interface GigabitEthernet 3/0/0
[USG2100-GigabitEthernet3/0/0] ip address 202.38.10.3 24
[USG2100-GigabitEthernet3/0/0] vrrp vrid 2 virtual-ip 202.38.10.1 master
[USG2100-GigabitEthernet3/0/0] vrrp vrid 5 virtual-ip 202.38.10.2 slave
```

HUAWEI TECHNOLOGIES CO., LTD. All rights reserved

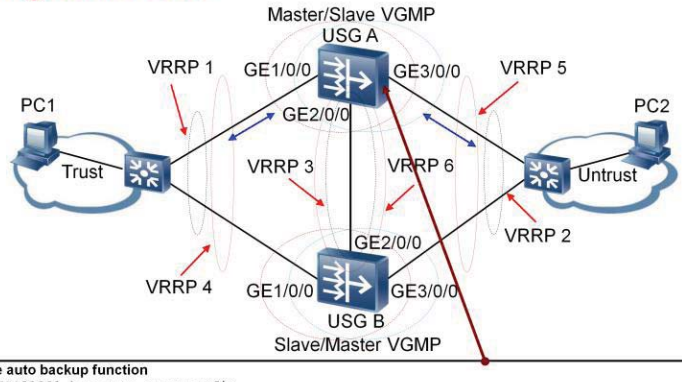
Configuration 2/3




```
# Add interface GE2/0/0 into VRRPs, enable HRP and  specify the HRP interface
[USG2100] interface GigabitEthernet 2/0/0
[USG2100-GigabitEthernet2/0/0] ip address 10.100.20.3  24
[USG2100-GigabitEthernet2/0/0] vrrp vrid 3 virtual-ip 10.100.20.1 master
[USG2100-GigabitEthernet2/0/0] vrrp vrid 6 virtual-ip 10.100.20.2 slave
[USG2100] hrp interface GigabitEthernet 2/0/0
[USG2100] hrp enable
```

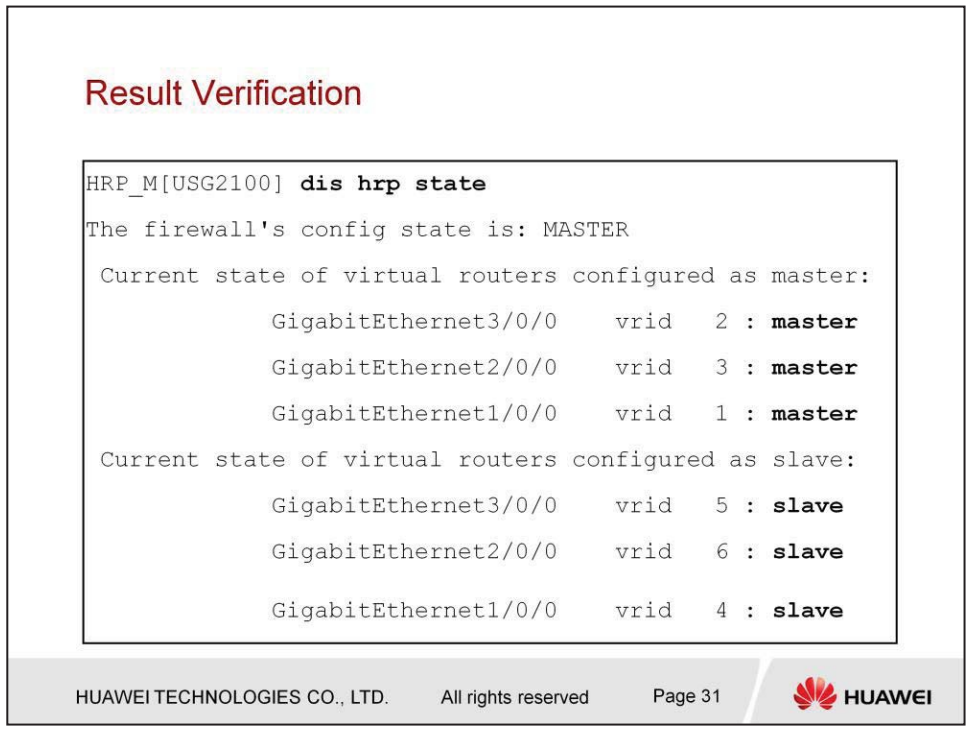
HUAWEI TECHNOLOGIES CO., LTD. All rights reserved

Configuration 3/3




```
# Enable auto backup function
HRP_M[USG2100] hrp auto-sync config
# Enable interzone packet filter
[USG2100] policy interzone trust untrust outbound
[USG2100-policy-interzone-trust-untrust-outbound] policy 1
[USG2100-policy-interzone-trust-untrust-outbound-1] action permit
[USG2100-policy-interzone-trust-untrust-outbound-1] policy source 10.100.10.0 mask 24
```

HUAWEI TECHNOLOGIES CO., LTD. All rights reserved



On USG A using **display vrrp** command, check the VRRP status and information.

HRP_M[USG2100]dis vrrp

GigabitEthernet3/0/0 | Virtual Router 2

state : Master

Virtual IP : 202.38.10.1

PriorityRun : 100

PriorityConfig : 100

MasterPriority : 100

Preempt : YES Delay Time : 0

Timer : 1

Auth Type : NONE

Check TTL : YES

GigabitEthernet3/0/0 | Virtual Router 5

state : Backup

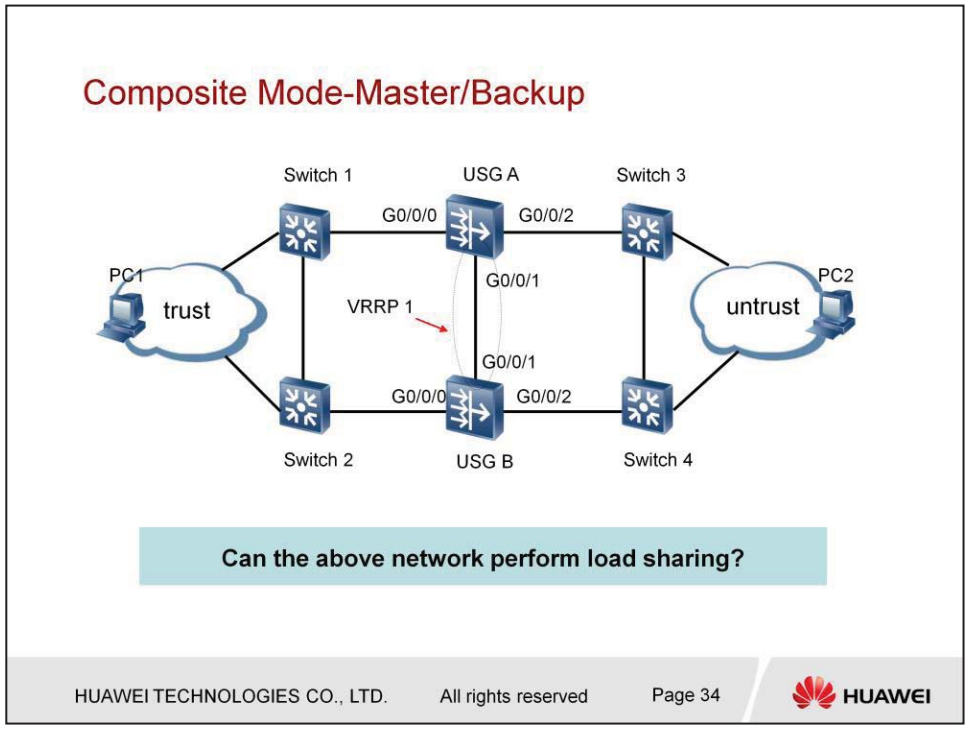
Virtual IP : 202.38.10.2

PriorityRun : 100

PriorityConfig : 100

MasterPriority : 100
Preempt : YES Delay Time : 0
Timer : 1
Auth Type : NONE
Check TTL : YES
GigabitEthernet2/0/0 | Virtual Router 3
state : Master
Virtual IP : 10.100.20.1
PriorityRun : 100
PriorityConfig : 100
MasterPriority : 100
Preempt : YES Delay Time : 0
Timer : 1
Auth Type : NONE
Check TTL : YES
GigabitEthernet2/0/0 | Virtual Router 6
state : Backup
Virtual IP : 10.100.20.2
PriorityRun : 100
PriorityConfig : 100
MasterPriority : 100
Preempt : YES Delay Time : 0
Timer : 1
Auth Type : NONE
Check TTL : YES
GigabitEthernet1/0/0 | Virtual Router 1
state : Master
Virtual IP : 10.100.10.1
PriorityRun : 100
PriorityConfig : 100
MasterPriority : 100

```
  Preempt : YES   Delay Time : 0
  Timer : 1
  Auth Type : NONE
  Check TTL : YES
GigabitEthernet1/0/0 | Virtual Router 4
  state : Backup
  Virtual IP : 10.100.10.2
PriorityRun : 100
  PriorityConfig : 100
  MasterPriority : 100
  Preempt : YES   Delay Time : 0
  Timer : 1
  Auth Type : NONE
  Check TTL : YES
```



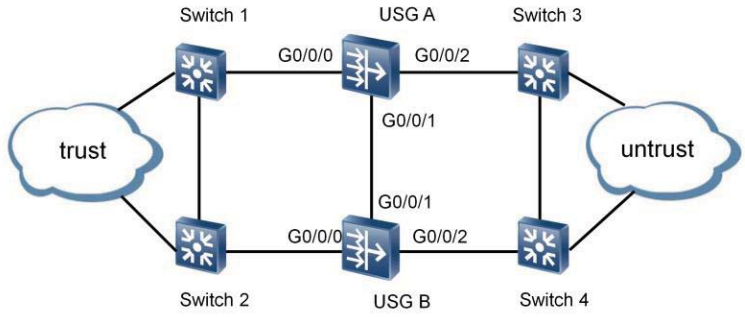
The upload and download devices are all layer 2 switches, and firewall choose master/backup networking mode.

The upload and download service port is under layer 2 mode, heart beat communication port is under layer 3 mode, and the two firewalls build a dual system.

USG A G0/0/0, G0/0/1, G0/0/2 belongs to related Trust, DMZ and Untrust zone, heartbeat interface GE0/0/1's IP address is 10.100.20.2/24, VRRP virtual IP address is 10.100.20.1/24 ;

USG B G0/0/0, G0/0/1, G0/0/2 belongs to related Trust, DMZ and Untrust zone, heartbeat interface GE0/0/1's IP address is 10.100.20.3/24, VRRP virtual IP address is 10.100.20.1/24.

Configuration 1/4



Add interface GE0/0/1 into VRRP

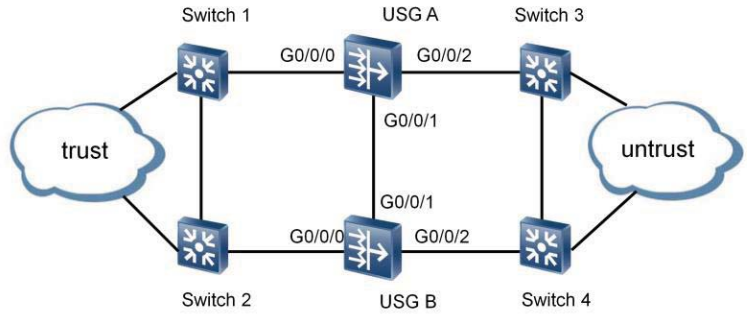
```
[USG2100] interface GigabitEthernet 0/0/1
[USG2100-GigabitEthernet0/0/1] ip address 10.100.20.2  24
[USG2100-GigabitEthernet0/0/1] vrrp vrid 1 virtual-ip 10.100.20.1 master
```

Add interface GE0/0/1 into DMZ

```
[USG2100] firewall zone dmz
[USG2100-zone-dmz] add interface gigabitethernet 0/0/1
```

HUAWEI TECHNOLOGIES CO., LTD. All rights reserved

Configuration 2/4

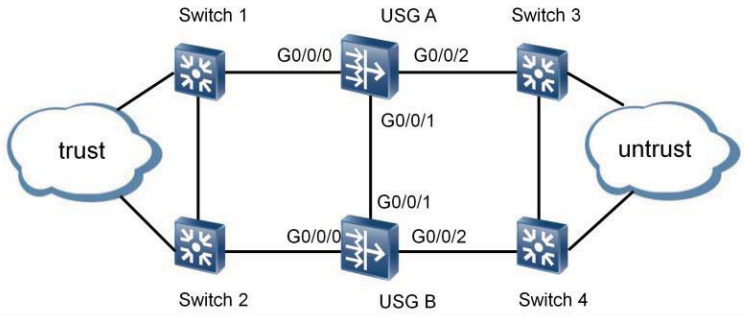


Set interface GE0/0/0 and GE0/0/2 to Layer 2 mode

```
[USG2100] interface gigabitethernet 0/0/0
[USG2100-GigabitEthernet 0/0/0] portswitch
[USG2100] interface gigabitethernet 0/0/2
[USG2100-GigabitEthernet 0/0/2] portswitch
```

HUAWEI TECHNOLOGIES CO., LTD. All rights reserved

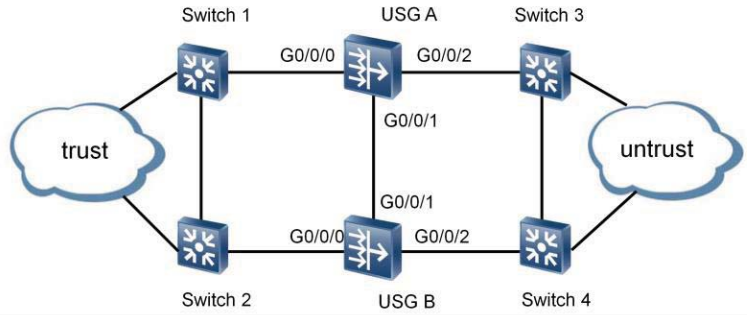
Configuration 3/4




```
# Create VLAN 2, add interface GE0/0/0 and GE0/0/2 to VLAN2, and HRP master VGMP track VLAN 2
[USG2100] vlan 2
[USG2100-vlan-2] port GigabitEthernet 0/0/0
[USG2100-vlan-2] port GigabitEthernet 0/0/2
[USG2100-vlan-2] hrp track master
```

HUAWEI TECHNOLOGIES CO., LTD. All rights reserved

Configuration 4/4



Enable HRP, and specify the HRP interface. The interface should be the VRRP channel interface

```
[USG2100] hrp enable
[USG2100] hrp interface gigabitethernet 0/0/1
```

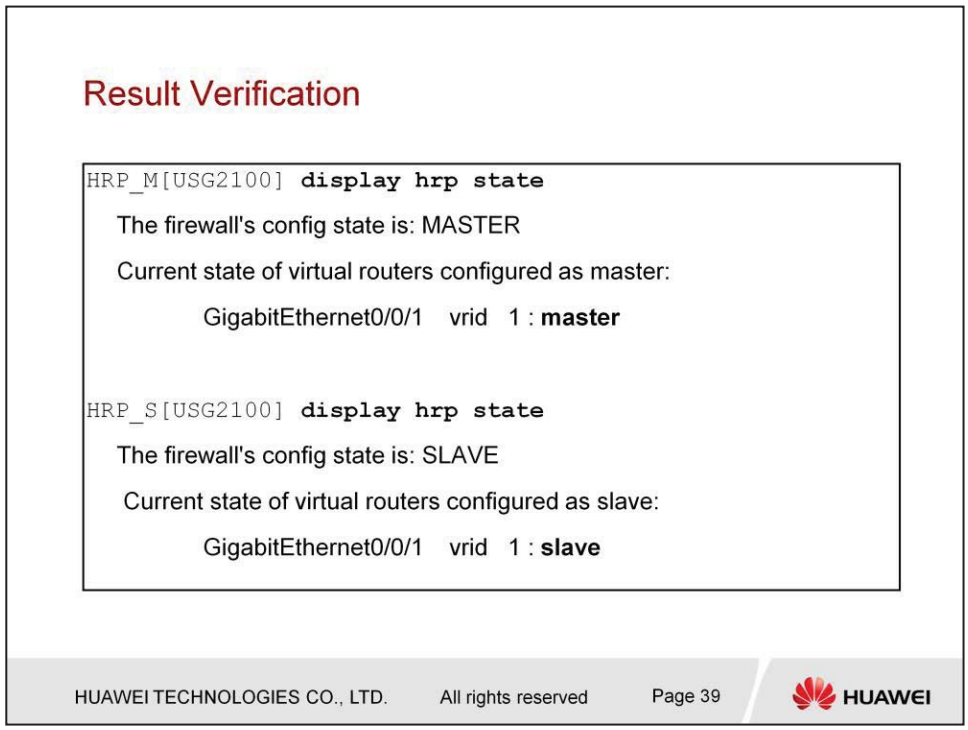
HUAWEI TECHNOLOGIES CO., LTD. All rights reserved

The configuration of USG B most similar with USG A, the difference is as follows:

The interface IP address of USG B is different with USG A.

On USG B the Slave VGMP track VLAN.

Caution: HRP backup channel interface can't be a layer 2 interface or VLANIF interface.



On USG A by using command **display vrrp**, to check the status of VRRP

HRP_M[USG2100]dis vrrp

GigabitEthernet0/0/1 | Virtual Router 1

state : Master

Virtual IP : 10.100.20.1

PriorityRun : 100

PriorityConfig : 100

MasterPriority : 100

Preempt : YES Delay Time : 0

Timer : 1

Auth Type : NONE

Check TTL : YES

On USG B by using command **display vrrp**, to check the status of VRRP

HRP_S[USG2100]dis vrrp

GigabitEthernet0/0/1 | Virtual Router 1

state : Backup

Virtual IP : 10.100.20.1

PriorityRun : 100

PriorityConfig : 100

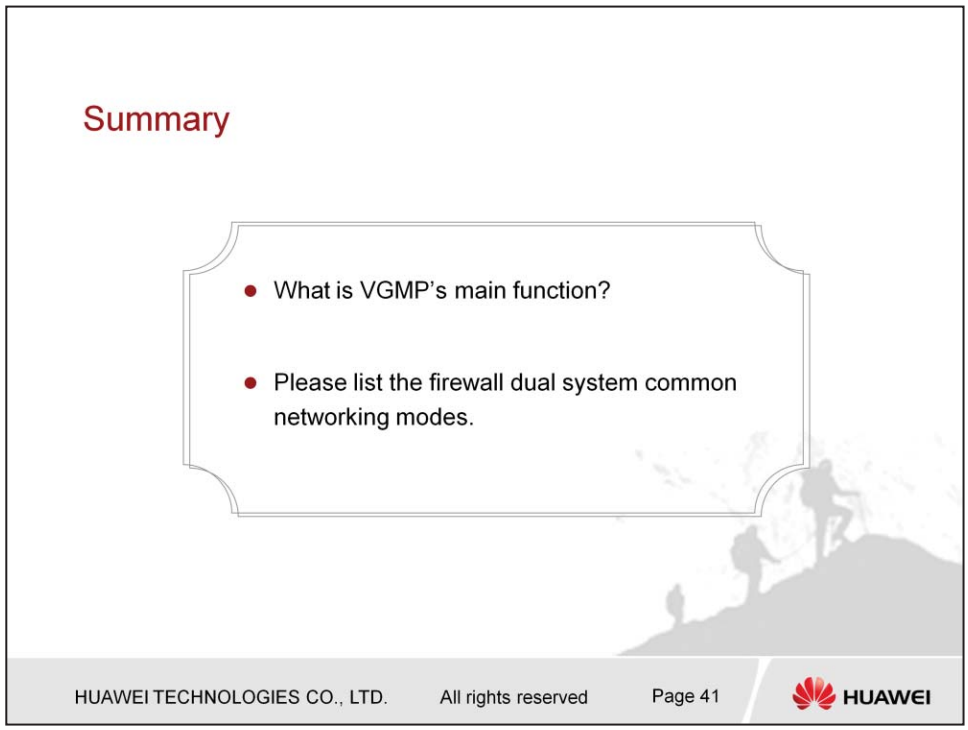
MasterPriority : 100

Preempt : YES Delay Time : 0

Timer : 1

Auth Type : NONE

Check TTL : YES



Answer:

1、

State consistency management

Preemption management

Channel management

2、

Route mode-master/backup

Route mode-load sharing

Composite-Master/backup

Composite-loadsharing

Module 2
HA Technology


HA Technology
www.huawei.com
HUAWEI TECHNOLOGIES CO., LTD. All rights reserved
HUAWEI



Foreword

Application growth and an increasing variety of value-added services being deployed has caused network bandwidth to also increase exponentially, resulting in network interruption for even a short period of time possibly affecting business and causing heavy loss. The network failure and service interruption due to inevitable factors invoke the need for improvements to system availability. It is feasible therefore to improve the fault-tolerance capability of the system, speed up recovery from faults, and reduce the impact of faults on the service. High availability (HA) supports this need for improved overall network stability.

.

HUAWEI TECHNOLOGIES CO., LTD. All rights reserved





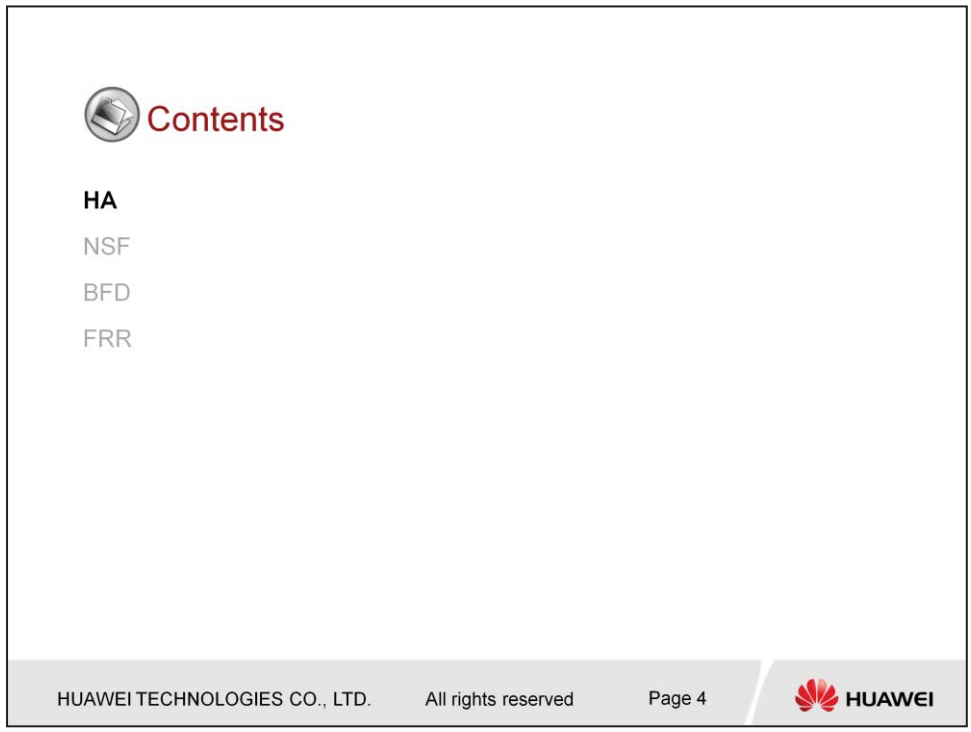
Objectives

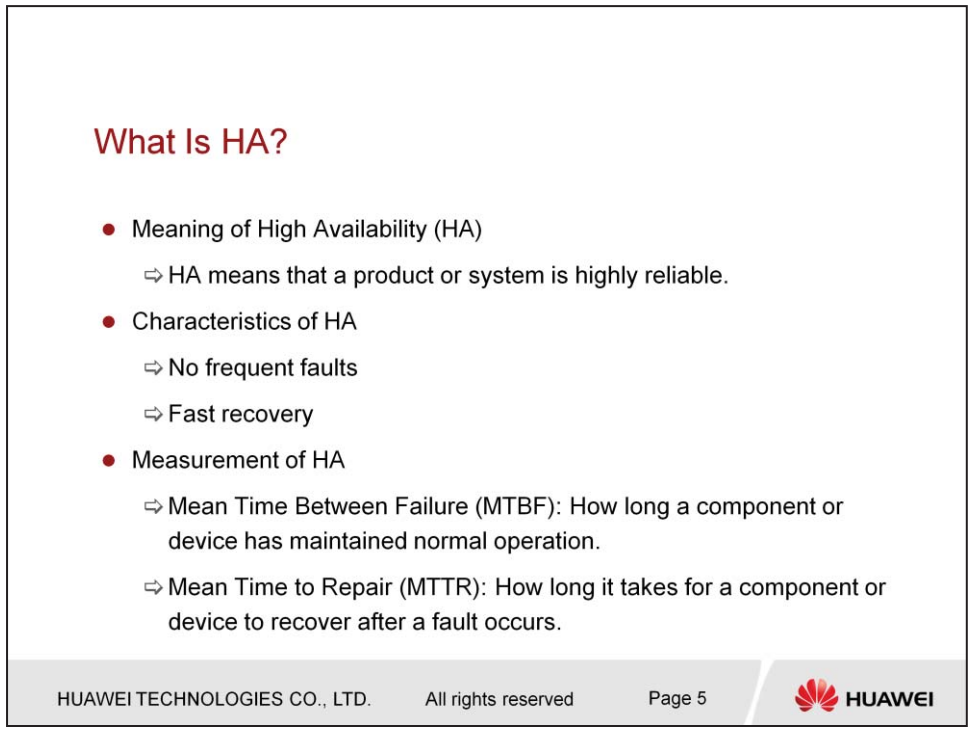
Upon completion of this section, you will be able to:

- Understand HA basic principles
- Master the evaluation of Internet availability

HUAWEI TECHNOLOGIES CO., LTD. All rights reserved

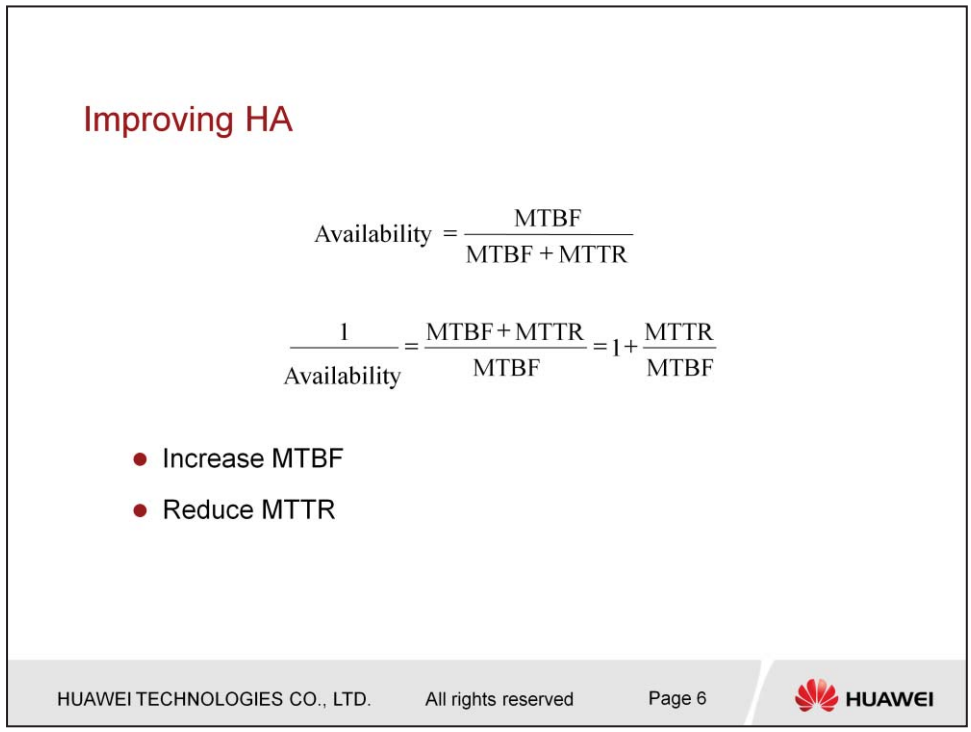



Contents
HA
NSF
BFD
FRR
HUAWEI TECHNOLOGIES CO., LTD. All rights reserved Page 4
HUAWEI



HA means that a product or system is highly reliable.

The network with HA has the following features: no frequent faults and fast recovery. The reliability of a product or system is measured by two indexes, MTBF and MTTR. MTBF refers to how long a component or device can work normally, and measured in hours. MTTR refers to how long a component or device can recover after a fault occurs. In broad sense, MTTR involves device management and customer service. It is an important index of device maintenance. The reliability of carrier-class devices must be equivalent to or higher than 99.999%. That is to say, the carrier-class device must be able to work normally in 99.999%. The faults cannot last more than 5.256 minutes per year.

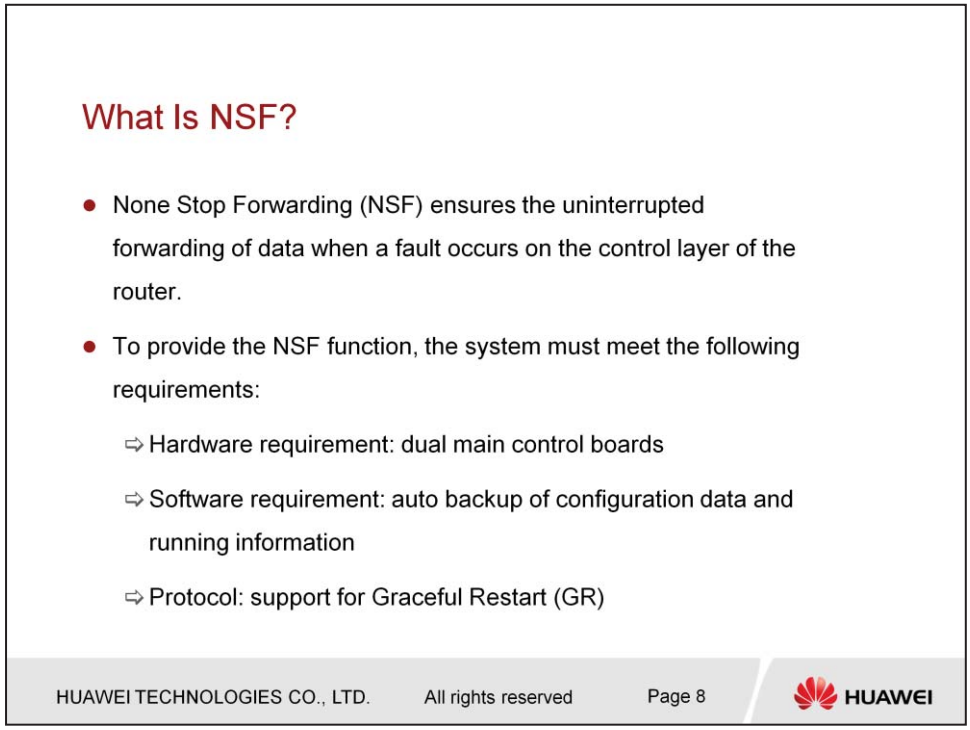


The formula of availability is as follows:

Availability = MTBF / (MTBF + MTTR) or 1 / Availability = 1 + MTTR / MTBF

The availability can be improved by increasing MTBF and reducing MTTR.


Contents
HA
NSF
BFD
FRR
HUAWEI TECHNOLOGIES CO., LTD. All rights reserved Page 7
HUAWEI

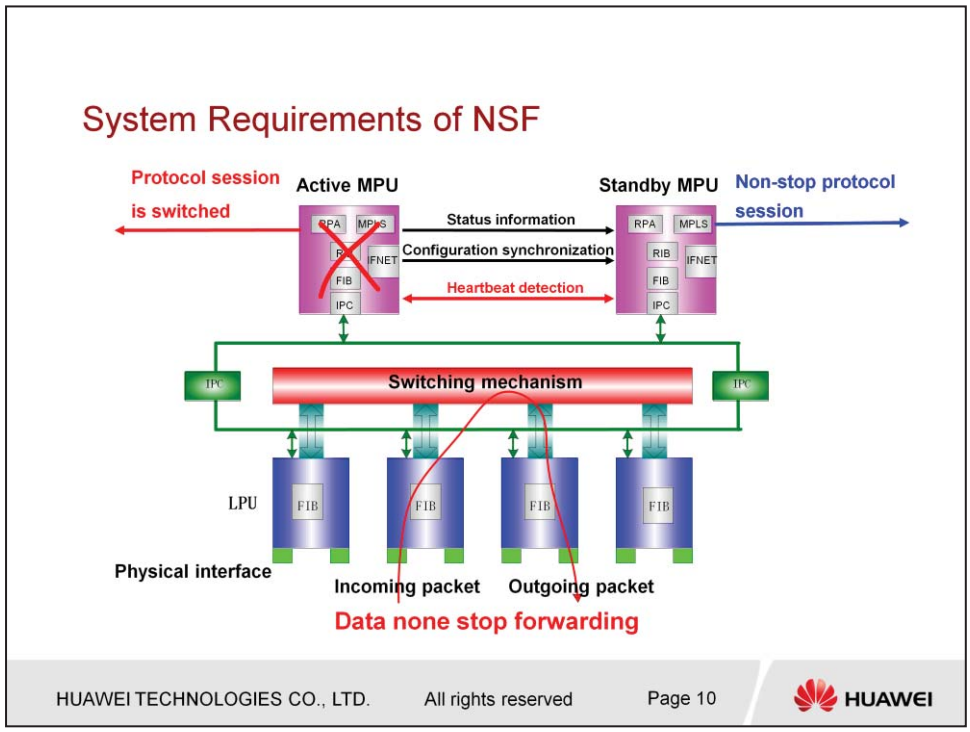


None Stop Forwarding (NSF) is a highly-reliable technology. It ensures the uninterrupted forwarding of data when a fault occurs on the control layer of the router. So the key services on the network are not affected by the faults. A fault may occur on the router control layer when the router is faulty or restarted due to software or hardware fault, or when the active/standby switchover occurs due to the running of configuration commands during the software upgrade. The router fault mentioned here refers to the resetting of the main control board of the router and with the active/standby switchover. Generally, when a router is faulty, the neighbor on the protocol layer can detect that the neighbor relationship is Down. After a certain period, the relationship is Up. This is called neighbor flapping. Such neighbor flapping will lead to route flapping. As a result, the black hole occurs on the restarted router or data bypass. The reliability of the network will decline. The purpose of NSF is to solve the problem of route f lapping. The following requirements should be satisfied:
Hardware requirement: Dual main control boards are provided.

One is active and the other is standby. When the active board is restarted, the other one becomes active. The distributed structure is adopted. That is, data forwarding is separated from control. A special line card (LPU) is used for data forwarding.

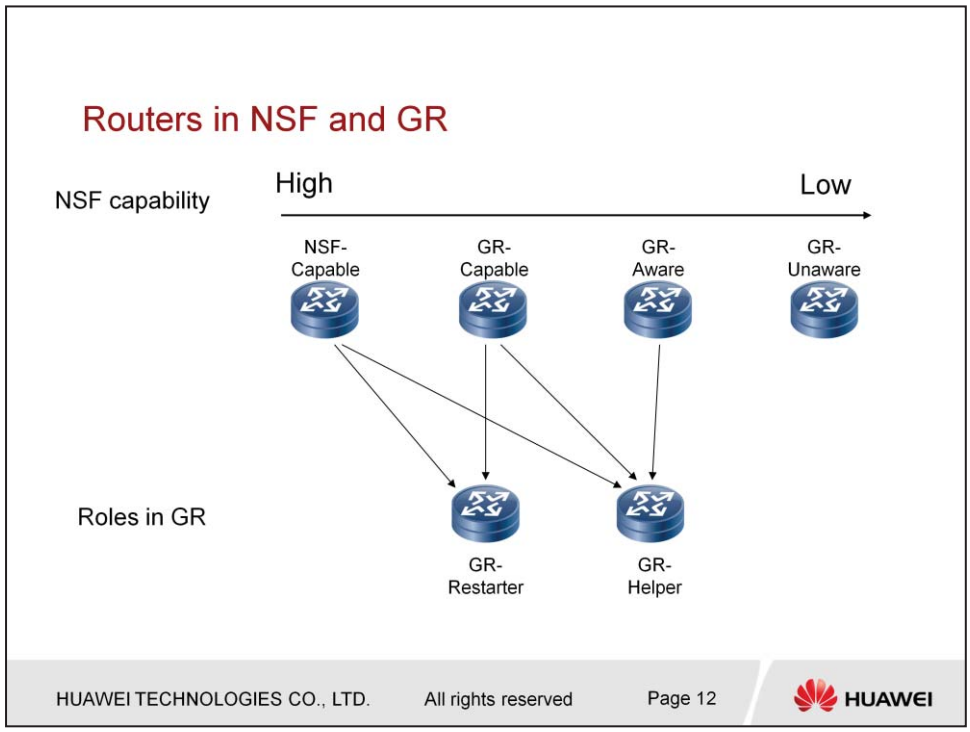
Software requirement: When the active board works, it backs up the data like configuration data and interface status to the standby board. If switchover occurs, the LPU with its interfaces still work normally, and the forwarding entry of the LPU is not deleted.

Protocol requirement: The network protocols like OSPF, IS-IS, BGP, LDP, and RSVP can be extended to support GR.



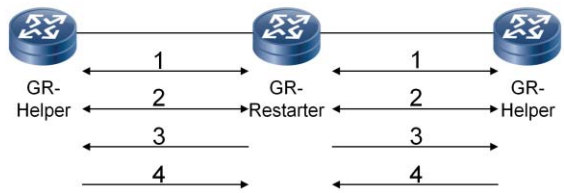
In the distributed model, the router consists of the main control board and the LPU. The main control board takes charge of the control layer and the management layer, for example, route learning, route calculation, and MPLS LSP setup. The LPU takes charge of the specific services, for example, IP packet forwarding, MPLS packet switching, and QoS guarantee. As shown in the above figure, the main control board of the high-end router provides the redundancy backup mechanism, namely, dual main control boards. One is called AMB, in running state; another is called SMB, in standby state. When the active board is running, It backs up all the static configuration information and some of the dynamic information to the standby board. The configuration data on the standby board and the active board can be synchronized. When a fault occurs on the main control board due to hardware or software fault, the SMB takes over as an AMB, and then restart the control layer and the management layer.

In addition, the system adopts the distributed structure in which the control and forwarding are separated, so the LPU is not restarted during the active/standby switchover. The forwarding entry of the LPU is kept and the service can still be forwarded.



Based on the NSF capability, the routers can be classified into:
NSF-Capable router: refers to the router that provides the NSF capability. The router must provide redundancy capability with dual main control boards and the
GR capability.
GR-Capable router: refers to the router that provides the GR capability.
GR-Aware router: refers to the router that is aware of the GR of neighbors and helps the neighbors complete GR. The
GR-Capable router is also a GR-Aware
router.
GR-Unaware router: refers to the router that is not aware of the GR of neighbors and cannot help the neighbors complete GR.
According to the roles, the routers during GR are classified into:
GR-Restarter router: the router performing GR.
GR-Helper router: the router helping the neighbors perform GR .

GR Communication



1. A GR session is established between the GR Restarter and the GR Helper.
2. The GR Helper discovers the fault of the GR Restarter and keeps the GR session. Before GR Time expires, the Helper keeps the routes learned from GR Restarter.
3. After the GR Restarter reboots, it sends the signal to the GR Helper and re-establishes the GR session.
4. The GR Restarter obtains the topology and routing information from neighbors, re-calculates the routes, and ages the old routes.

HUAWEI TECHNOLOGIES CO., LTD. All rights reserved

The communication mechanism between the GR Restarter and the GR Helper is as follows:

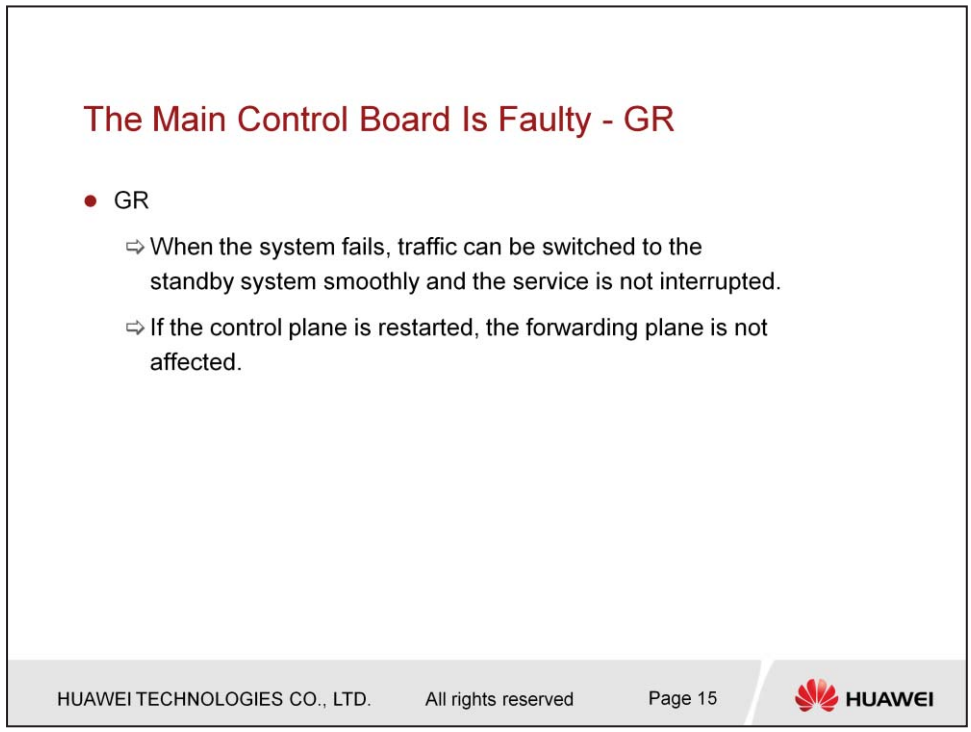
1. A GR session is established between the GR Restarter and the
2. GR Helper.

2. The GR Helper detects the status of the GR Restarter and maintains the status information with GR Time. If the GR Helper f inds that the peer is Down, it considers that the peer can recover in GR Time. Before the GR Time expires, the GE Helper does not end the session or delete the topology or routing information of the GR Restarter.

3. The GR Restarter has been recovered, it sends signal to every GR Helper and re-establishes the neighbor relationship list after receiving the response. The GR Restarter calculates its own routing table according to the topology or routing information which are learnt from all the GR Helpers by setting up sessions.

4.The GR Restarter and the GR trigger the aging of the routing table and forwarding table by the sign of Stale.

The wrong configuration by the administrator or a fault occurs on the router may cause the switchover of the GR Restarter and reboot. When a fault occurs on the GR Restarter, the route entries are not removed from the routing table and the forwarding table, and the LPU is not reset. The LPU still transmits services. Thus the packets are transmitted by the system without interruption. On the VRP5.30, GR supported by MPLS LDP (DU or DOD), OSPF (IPv4), IS-IS (IPv4/IPv6), BGP (IPv4/IPv6), and MP-BGP.



If the router reboots due to the fault on the main control board, GR can be started to ensure that service can be forwarded without interruption. Even though GR requires different hardware, system-level software, and application software, the purpose of GR is to transfer data to the standby system smoothly when the active system fails and the service is not interrupted. As a network system, a router requires the system-level software and the application software. The system-level software provides services that should be provided by the operating system, for example, managing processes, allocating memory space, and scheduling CPU. The application software can be considered as the protocol modules, like BGP, OSPF, ISIS, and MPLS. The services of the router are provided by the application software. Therefore, the most important thing of GR is to implement the application software. In the distributed router architecture, the application software is maintained by the control layer. So the smooth reboot of the software system is the smooth reboot of the control layer. The purpose of GR is to ensure that Non-stop Forwarding (NSF) of services can be implemented.

That is, when
the control layer reboots smoothly, the forwarding plane is not affected and service can be forwarded without interruption. When data is switched on the control layer, the forwarding layer can still forward data until the control layer is updated.



Contents

HA

NSF

BFD

FRR

HUAWEI TECHNOLOGIES CO., LTD. All rights reserved



BFD Functions and Features

- Bidirectional Forwarding Detection (BFD)
  - ⇨ Provides bidirectional detection for links and detects faults in several milliseconds.
  - ⇨ Provides asynchronous detection and query detection.
  - ⇨ Supports fault detection for one-hop links and multi-hop links.
  - ⇨ Provides the BFD results to other modules



HUAWEI TECHNOLOGIES CO., LTD. All rights reserved HUAWEI

BFD is developed from the basic transmission technology, therefore, it can detect the faults on all the layers of the network. it can detect the Ethernet, MPLS, routing encapsulation, and IPSec tunnel. Basically, BFD is a highspeed independent hello protocol. BFD detects the faults of communication between the forwarding engines. To be specific, BFD detects the connectivity of the data protocol between two systems on the same link. The link can be a physical link, logical link, or tunnel. BFD implements the following functions:

•Providing bidirectional detection for the links: At both ends of the bidirectional link, the BFD sends detection packets at the same time to detect the link status in dual directions. The BFD implements millisecond detection of link defects. A special example is the unidirectional link, such as an LSP. In this case, the BFD control packet needs to be sent only in one direction, and the peer can report the link status through other routes.

•Providing asynchronous detection and query detection: Asynchronous detection means that Each system sends BFD control packets based on the negotiated
period. If a system does not receive packets from the peer within the detection time, it sets the session to Down. Query detection means that Each system has an independent method to confirm whether it is connected to other systems. Once a BFD session is set

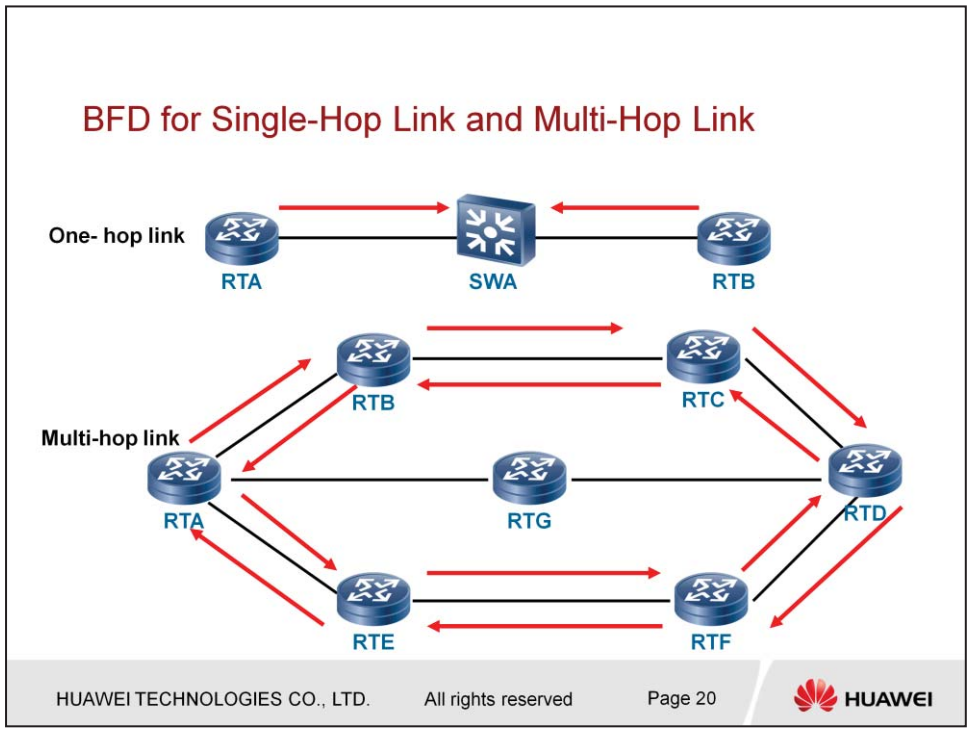
up, the system stops sending BFD control packets. Some routers need to explicitly verify the connectivity, then the system consecutively sends several BFD control packets whose P bit are 1, based on the negotiated period. If the system does not receive returned packets within the detection time, the session is set to Down. If the system receives returned packets within the detection time, it no longer sends BFD control packets. Instead, the system waits for triggering the next query.

•Changing BFD parameters dynamically: After the session is set up, you can dynamically change the relevant BFD parameters such as the minimum sending interval, the minimum receiving interval, enabling or disabling demand mode. Two ends adopt the new parameters after the negotiation with the session uninterrupted.

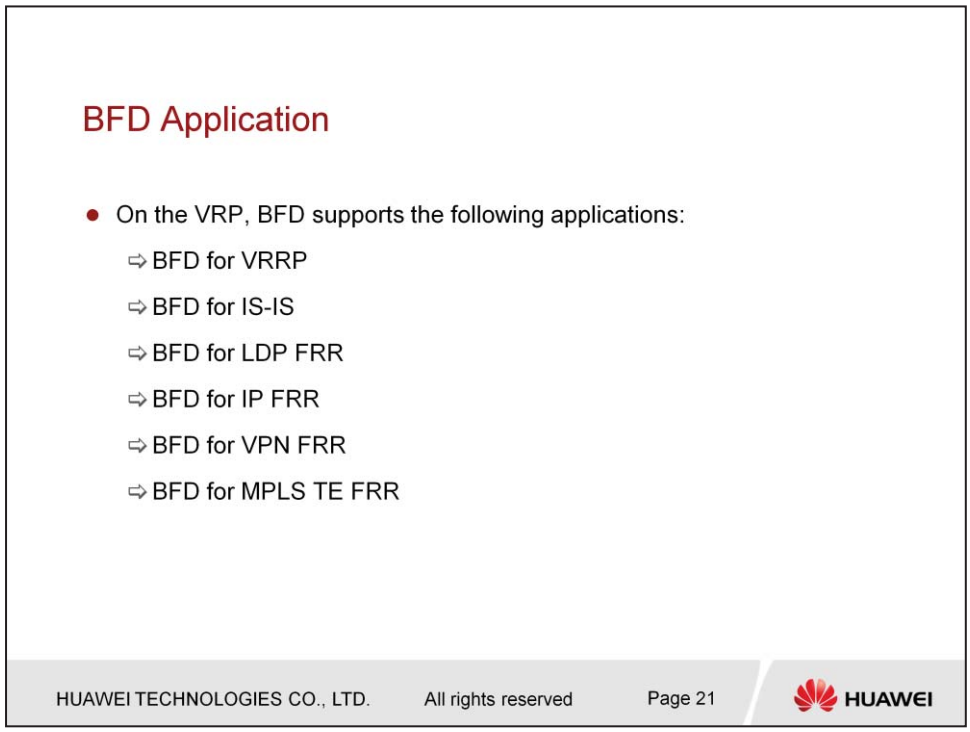
•Detecting One-hop and Multi-hop Links

•Informing Detection Results to Other Modules: The BFD provides the detection result and informs registered modules of the detection results so as to support the operation of other modules. The BFD mechanism is simple. Two systems set up the BFD session, and send the BFD control messages periodically.

If one end does not receive the BFD control message within the specified period, it considers the link faulty.



BFD for one-hop link means that detect the IP connectivity of the link that directly connects two systems. In this application, there will be only a single BFD session between two systems over a given interface (logical or physical) for a particular protocol. The BFD session must be bound to this interface. BFD Control packets must be transmitted in UDP packets with destination port 3 784. The source port must be in the range 49152 through 65535. All the BFD control messages of a session use the same source port. Single-hop BFD adopts a simple method to avoid spoofing. all BFD Control packets for the session must be sent with a TTL or Hop Count value of 255. All received BFD Control packets must be discarded if the received TTL or Hop Count is not equal to 255. BFD can also be useful on arbitrary paths between systems, which may span multiple network hops or overlap with another path. In multi-hop BFD, BFD Control packets must be transmitted in UDP packets with destination port 3784. The number of the source port ranges from 49152 to 65535. In the latest BFD draft, destination port number of UDP is 4784. The VRP5.30 complies with the earliest draft. The destination port number of multi-hop BFD is the same as that of one-hop BFD.



As a unified detection mechanism, the BFD can be used by many protocols. In VRP, BFD supports the following applications:

•BFD for VRRP

The BFD detects and monitors the link or IP routes forwarding at a fast pace. So VRRP fast switch is implemented.

•BFD for IS-IS

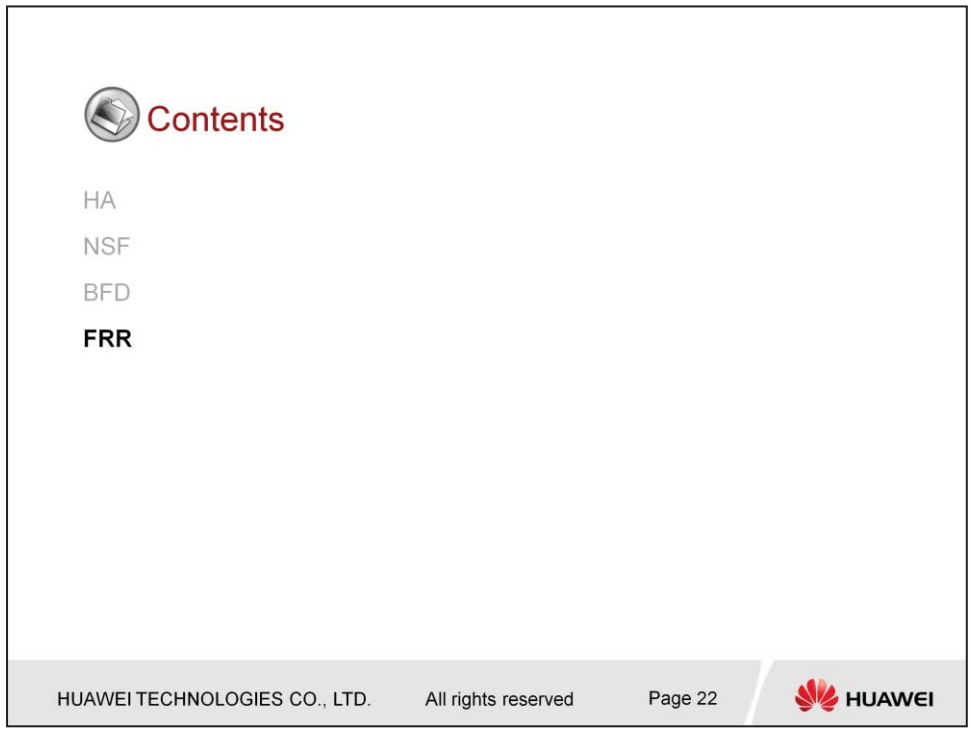
The BFD detects the link fault between ISIS neighbors at a fast pace and reports to the ISIS protocol. So ISIS fast convergence Is implemented. Because ISIS can
set up only one-hop ISIS adjacencies, the BFD limits its operations to one-hop
ISIS adjacencies.

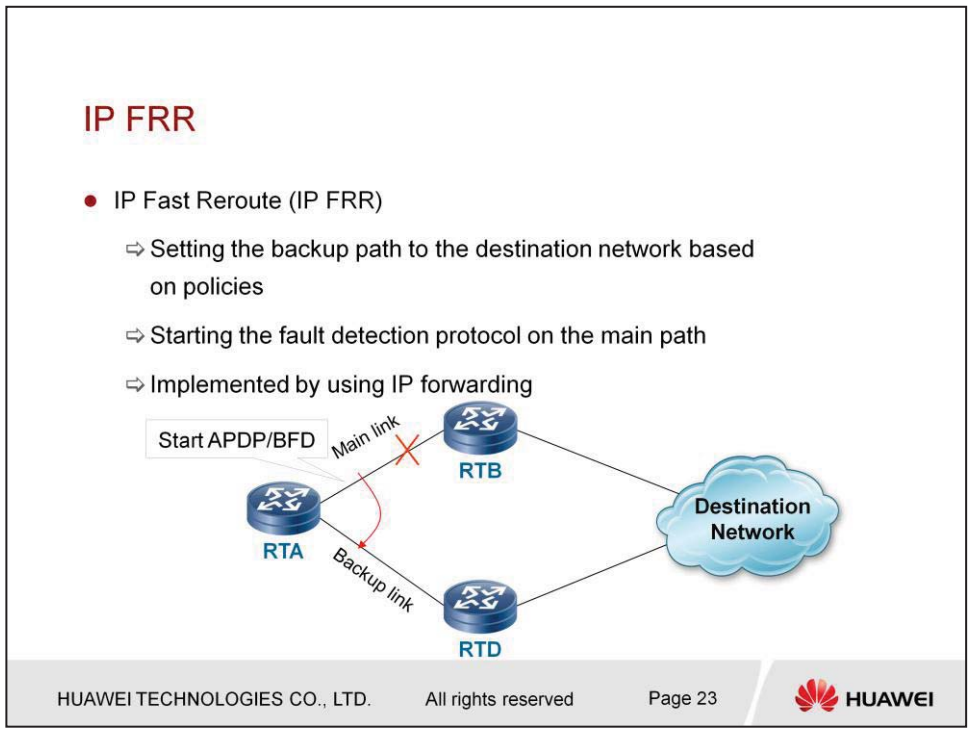
•BFD for LDP FRR

The BFD can detect the protected interfaces to trigger LDP FRR switch.

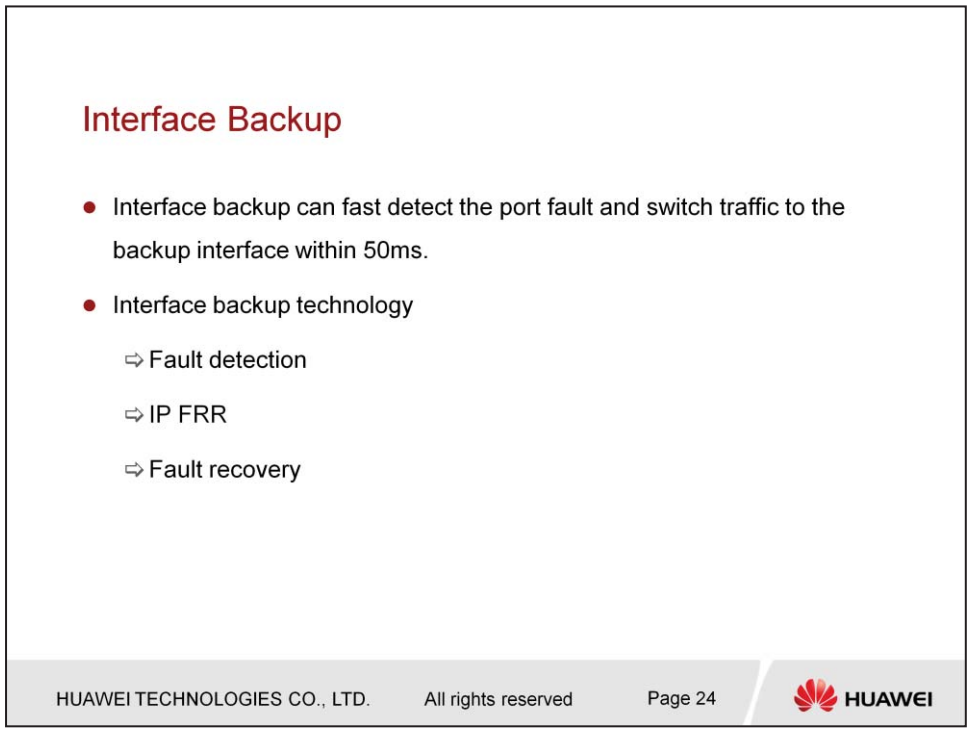
•BFD for IP FRR, VPN FRR, and MPLS TE FRR

The BFD can detected faults which can be reported to trigger IP FRR and VPN FRR.


Contents
HA
NSF
BFD
FRR
HUAWEI TECHNOLOGIES CO., LTD.
All rights reserved
Page 22
HUAWEI



IP FRR implements fast switching for the IP flows on the protected interface. The IP FRR completes within 50ms. The principle of IP FRR is that an interface is used as the backup of another interface. Both the main link and backup link are delivered to the FIB table. When the main link works normally, traffic is transmitted through the main link. When the main interface becomes invalid or the neighbor that is connected to the main interface becomes invalid, traffic is transmitted through the backup link. The local router is aware of the fault by using the hardware technology or other fault detection protocols like BFD. For example, BFD can trigger IP FRR. The principle is that when BFD is aware of the fault, it sets the port state table (PST) to Down. Then, IP FRR takes effect. Before routes are aggregated, traffic is switched to the backup interface. The next-hop router of the backup interface can reroute the IP flows, so packets will not lost. When the routes have been aggregated and the FIB table have been refreshed, IP FRR becomes invalid, then traffic is forwarded through the re-aggregate path. IP FRR can only be implemented using IP forwarding.



Interface backup technology has three steps, fault detection, rerouting, and fault recovery.

•Fault detection

When BFD detects the fault of link or port, it modifies the port status table to make the router aware of the fault.

•IP FRR

When a fault occurs on the link but the routes are not aggregated, IP traffic is switched to the link specified by the backup next hop. The IP FRR can be implemented. After the routes are aggregated, the new link is selected to forward traffic according to the routing.

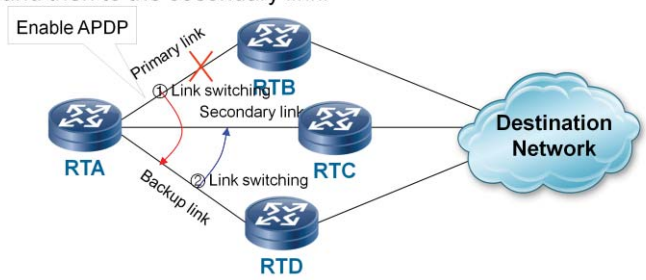
•Fault recovery

After fault recovery, the system remains the fault status for a while rather than switching traffic back to the primary link. This prevents the frequent traffic switchover between primary and standby links.

To ensure the reliability of the real-time services, the user should set up the protection links of which the routes are reachable in the networks that have the same destination address.

Interface Backup-IP FRR

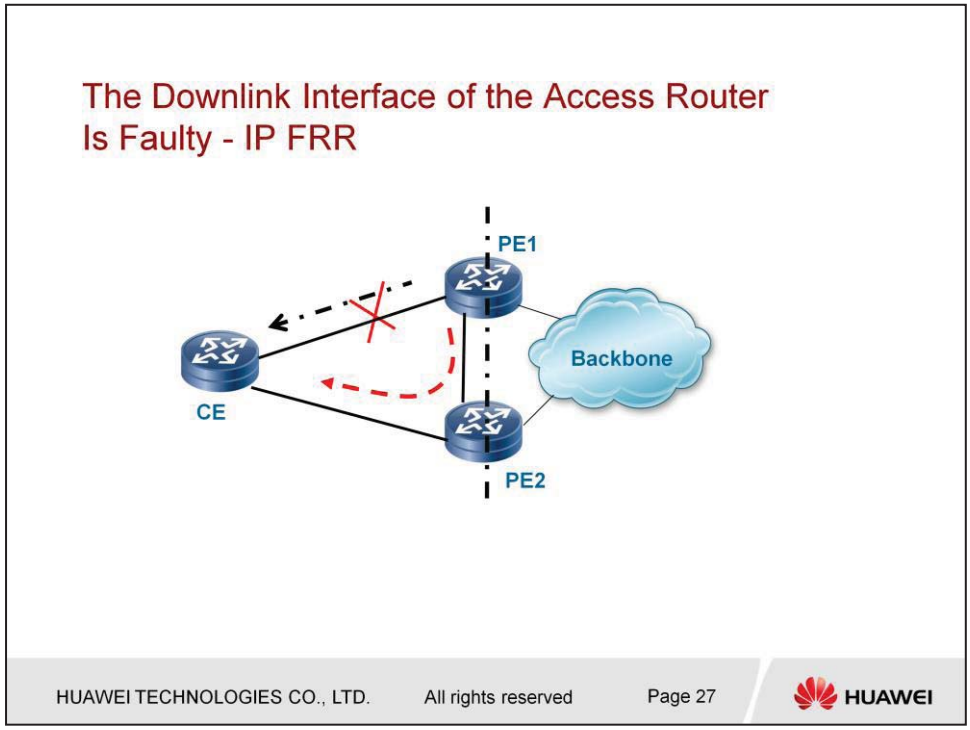
- Interface backup involves primary link, secondary link, and backup link
- When the primary link is faulty, traffic is switched to the backup link, and then to the secondary link.



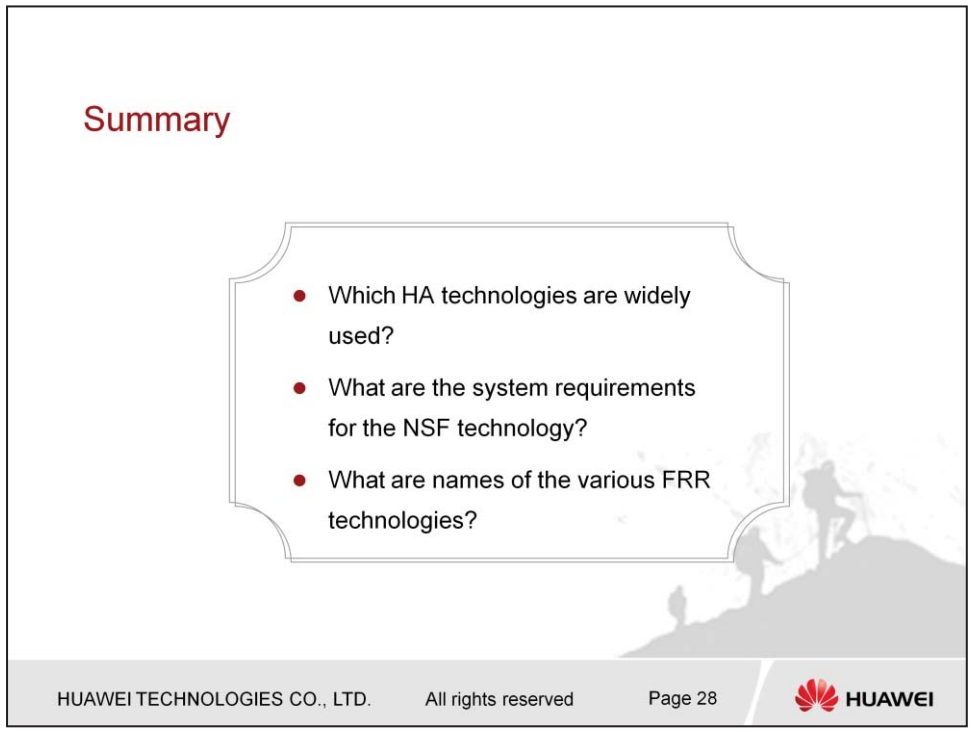
HUAWEI TECHNOLOGIES CO., LTD. All rights reserved

When a fault occurs on the link, the routes are aggregated within several seconds. Route aggregation may interrupt services for several seconds. This cannot be tolerated in the core network that has high service volume. To solve this problem, the system can perform IP FRR by specifying the backup next hop. Interface backup involves three types of links, primary link, secondary link, and backup link. The primary link is the optimal link. Service traffic is forwarded through the primary link when the network is stable and the routes are aggregated. The cost of the secondary link is higher than the cost of the primary link. When the primary link becomes invalid, routes are aggregated to the secondary link. The backup link is the link specified by the backup next hop. The cost values of the three types of links are different. When a fault occurs on the
link: The traffic is switched to the link specified by the backup next hop if the routes are not aggregated; the traffic is forwarded through the new link if the routes are aggregated. The function of the backup next hop is to fill the time slot between route aggregation, and thus the traffic can be switched to another link quickly. The service is not interrupted. In interface backup, the backup next hop should be specified on the primary link and BFD should be enabled on the primary link. If the primary link is fault,

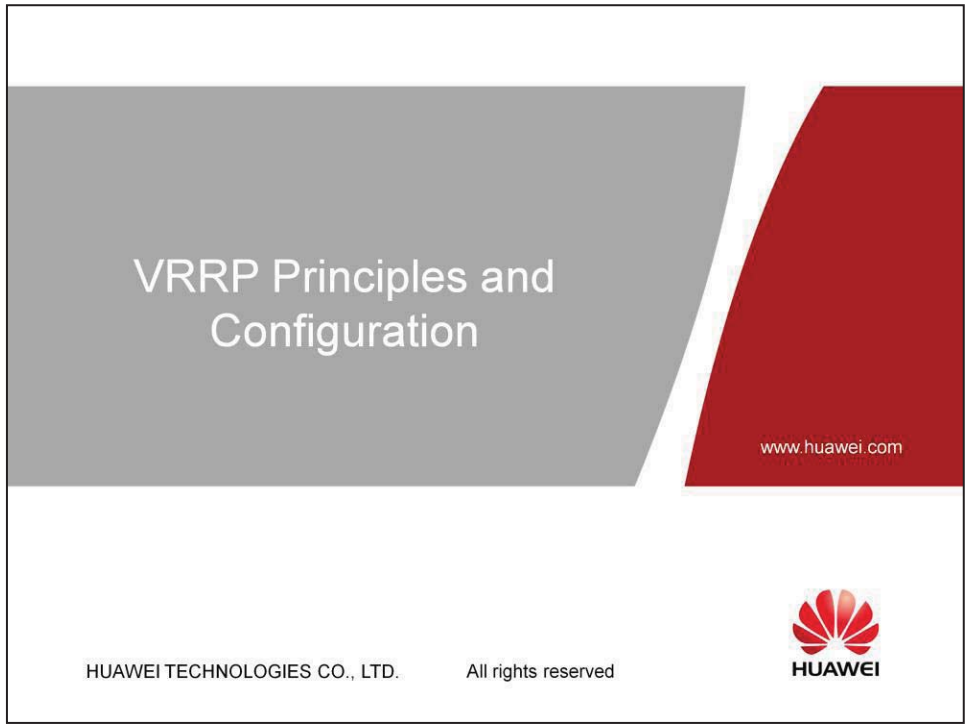
then the device can be aware of the fault on the primary link according to the port status table, and it re-selects the route according to the backup next hop. The packets are redirected to the backup link. Thus no packet is lost. In the following seconds, the routing protocol can be aware of the link fault. Rerouting is completed through the information interaction between routers. The route to the destination network is changed to the secondary link. Packets are switched to the secondary link. If the backup next hop is on the secondary link,
traffic is not switched. In this case, however, there are still two phases, that is the redirection phase and the secondary link phase. In the two phases, the packets
are forwarded through the same link.



As shown in the above figure, traffic to CE is forwarded by PE-1 (master PE). If the link from PE-1 to CE is faulty, we can use IP FRR to switch traffic to the link
from PE-1 to PE-2. The principle of FRR is that the forwarding plane stores the backup path for fast switchover. Similarly, when IP FRR is used, there are two paths from PE-1 to CE. One is the direct path, and the other is PE-1 – PE-2 – CE. Generally, the CE accesses the network through L3VPN, so the IP FRR is also used in the private network. We should set up a private network neighbor between PE-1 and PE-2. In this way, a master path and a backup path exist from PE-1 to the CE.



1. Which HA technologies are widely used?
Hardware redundancy on single device, like dual main control boards, hot plugging of boards, redundant power supplies, and redundant fans Link binding, like Ethernet link aggregation, MP, and MFR L2 redundant technologies, like Spanning Tree Protocol (STP) and Flex Link Redundant gateways, like VRRP, ECMP, floating static route, fast aggregation of dynamic routes (like fast hello and iSPF), NSF, MPLS FRR, Fast fault detection, like BFD
2. What are the system requirements for the NSF technology?
Hardware requirement: dual main control boards Software requirement: real-time backup of the configuration data and the running status Protocol requirement: Graceful Restart (GR)
3. What are names of the various FRR technologies?
IP FRR, MPLS LDP FRR, and VPN FRR.


VRRP Principles and Configuration
www.huawei.com
HUAWEI TECHNOLOGIES CO., LTD.
All rights reserved
HUAWEI



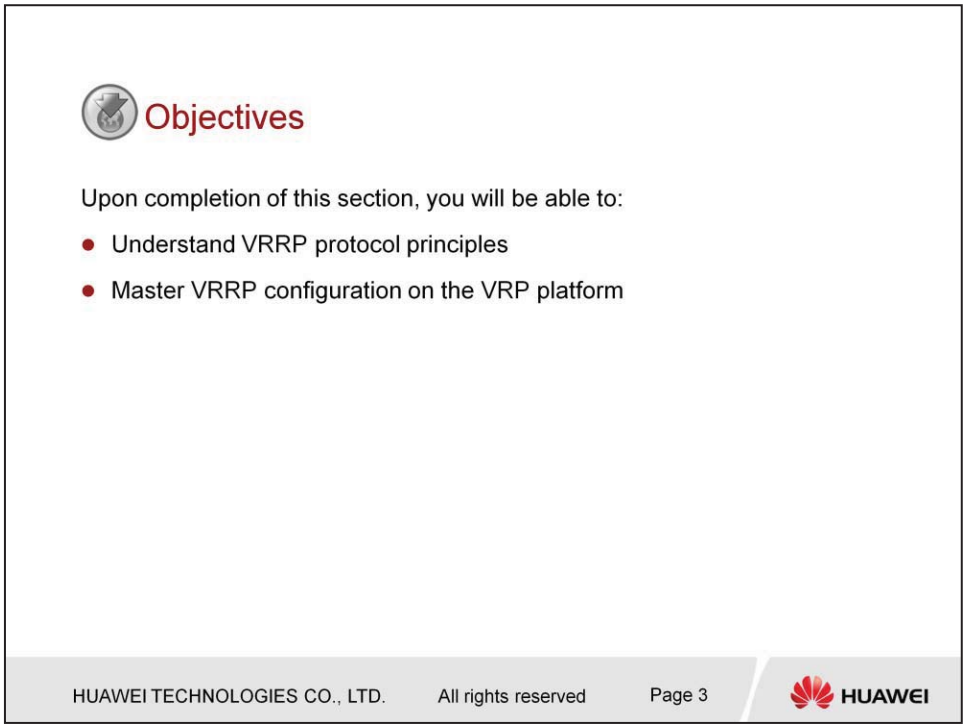
Foreword

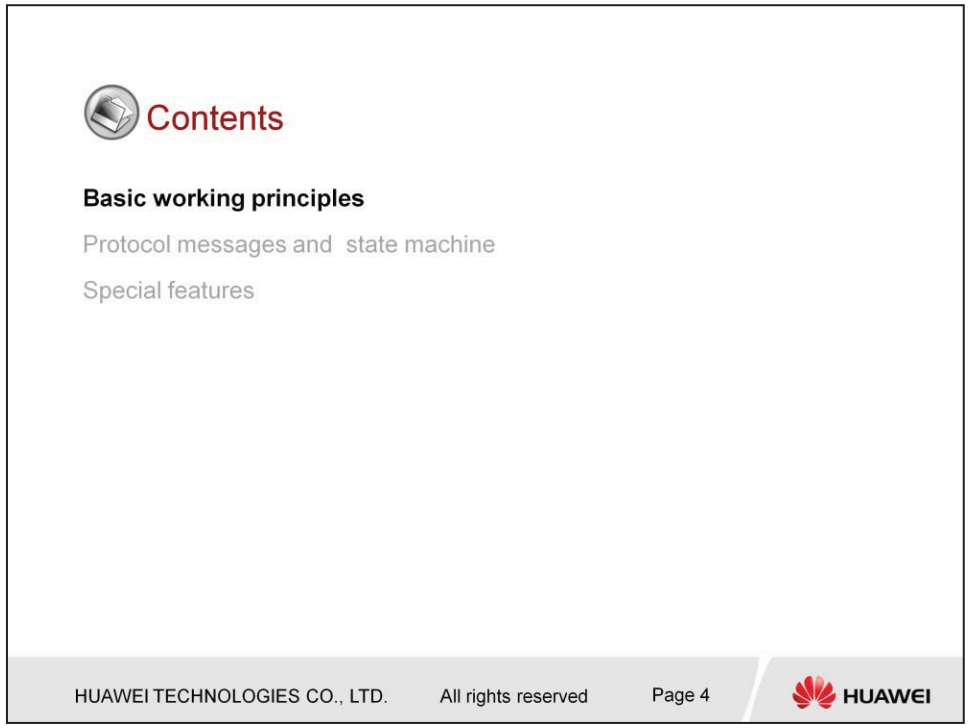
This section introduces virtual router redundancy protocol (VRRP) principles and configuration.

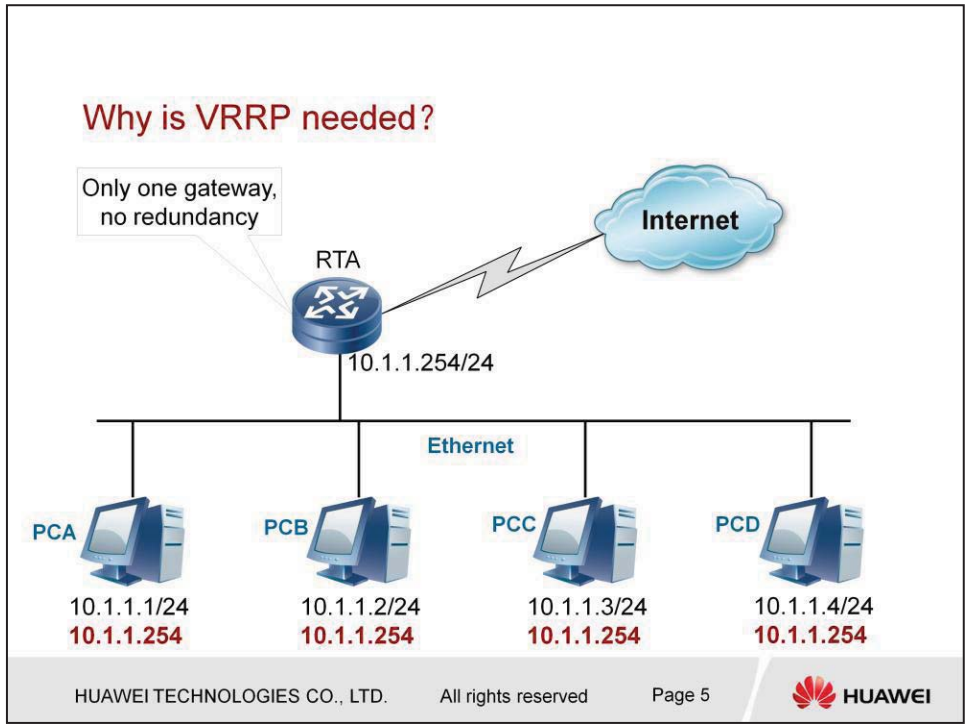
VRRP describes a dynamic election protocol, electing a Master router from a group of VRRP routers, and references it to a virtual router, as the gateway of the connected network.

HUAWEI TECHNOLOGIES CO., LTD. All rights reserved

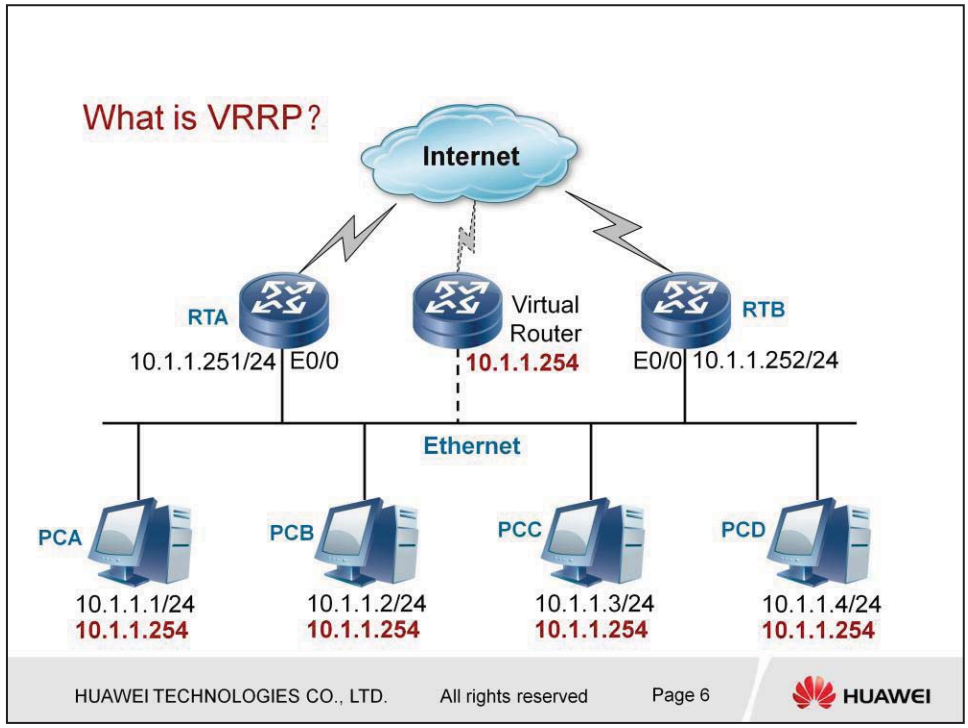



Objectives
Upon completion of this section, you will be able to:
Understand VRRP protocol principles
Master VRRP configuration on the VRP platform
HUAWEI TECHNOLOGIES CO., LTD.
All rights reserved
Page 3
HUAWEI


Contents
Basic working principles
Protocol messages and state machine
Special features
HUAWEI TECHNOLOGIES CO., LTD. All rights reserved Page 4
HUAWEI



In this example, there is only one router RTA in the LAN, which is used as the gateway by all the PCs. There is no redundancy provided, so if RTA fails then all the PCs will lose access to the Internet.

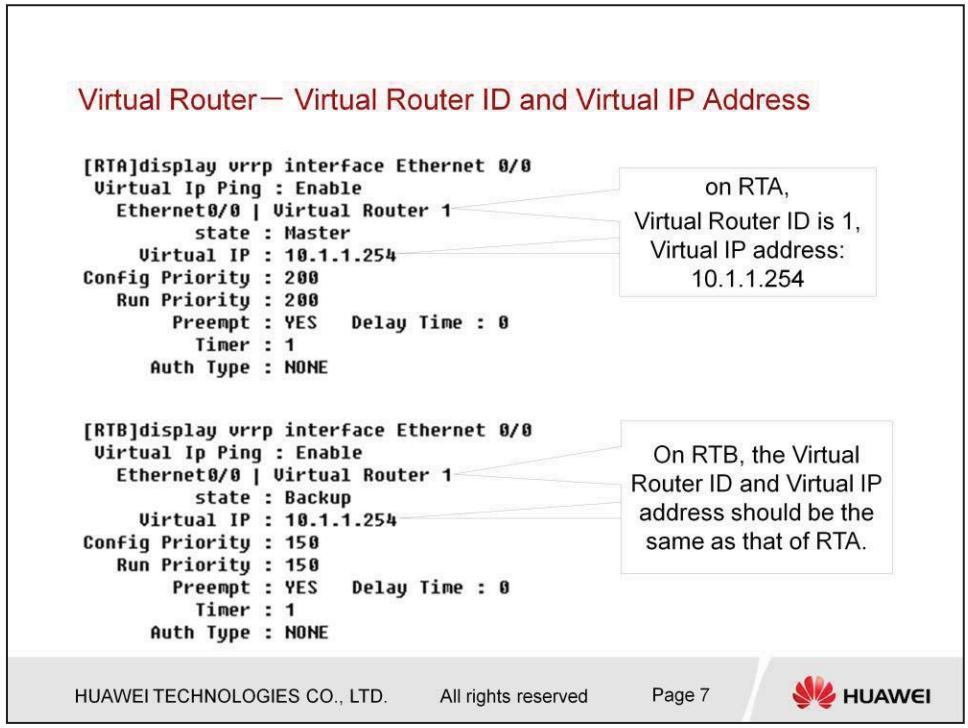


VRRP is designed to provide a virtual router on a LAN.

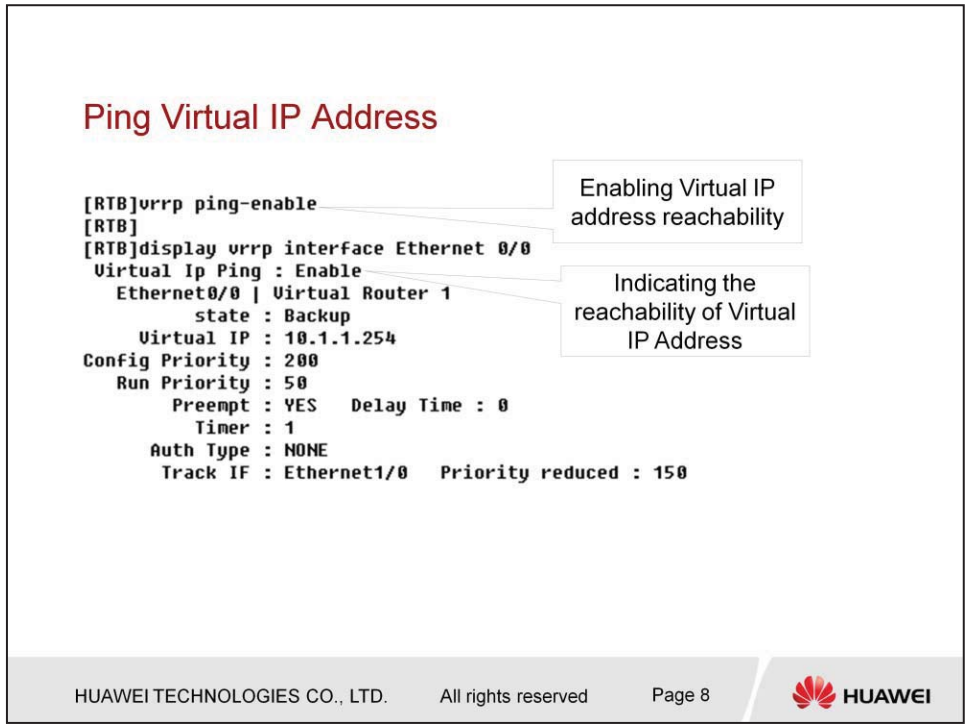
In this case:

There are two routers (RTA and RTB) on this LAN, RTA has physical IP address 10.1.1.251/24; RTB has physical IP address 10.1.1.252/24. RTA and RTB are configured to be associated with the same Virtual Router. This Virtual Router has a virtual IP address 10.1.1.254. All the PCs on this LAN can use the virtual IP address 10.1.1.254 as the default gateway, regardless the physical IP addresses of the two routers.

VRRP elects one router from the VRRP routers as the Master, and the Master processes all the packets sent to the virtual IP address. If the Master is down, VRRP elects a new Master from other VRRP routers.



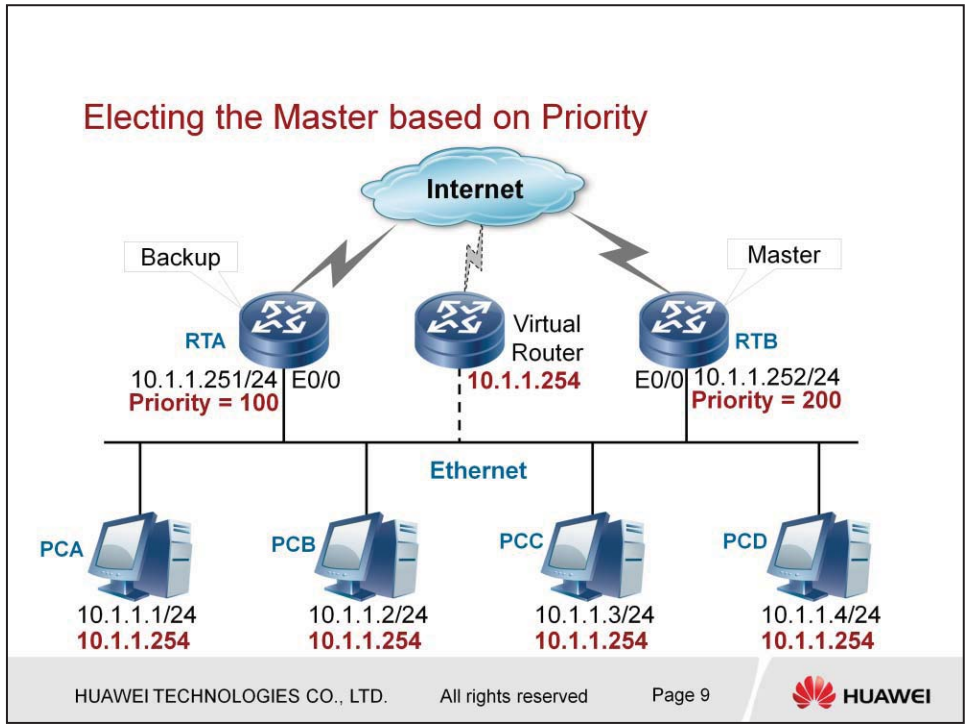
A Virtual Router is identified by both Virtual Router ID and associated Virtual IP Address. Multiple Virtual Routers could be configured on the same interface. A Virtual Router ID (VRID) is the identifier of a Virtual Router. Configurable item in the range 1-255 (decimal). The Virtual Router IDs configured on all the VRRP routers belonged to the same virtual group must be the same. A Virtual Router can be associated with more than one Virtual IP Addresses, however the Virtual IP Addresses configured for the VRRP routers belonged to the same Virtual Router should be the same. In VRP, If VRRP routers with the same VRIDs but different virtual IP addresses; or alternatively with same IP address but different VRIDs, they are regarded as different Virtual Routers.



By default, the ICMP - Echo message sent to the Virtual IP address will not be responded, even by the Master router. In the Master, under system view, the following commands could be used to enable or disable the function by which the ICMP -Echo message sent to Virtual IP address will be responded.

vrrp ping-enable

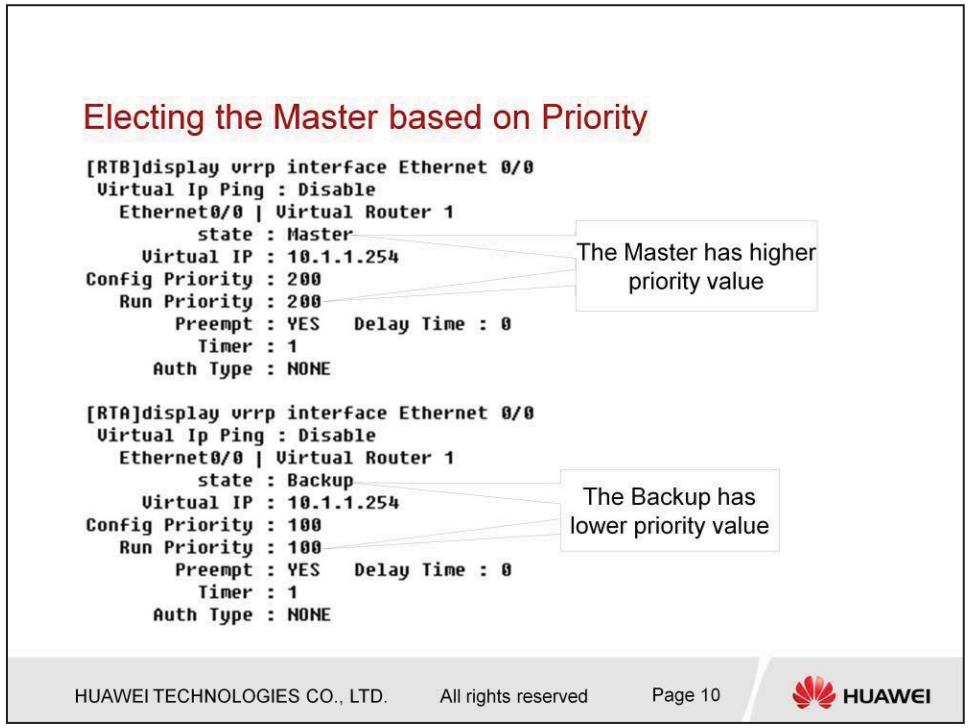
undo vrrp ping-enable



Master: The VRRP router that is assuming the responsibility of forwarding packets sent to the IP address associated with the virtual router, and answering ARP requests for these IP addresses.

Backup: The set of VRRP routers available to assume forwarding responsibility for a virtual router should the current Master fail.

The election of Master is based on the value of Priority. For the same interface, different Priority values could be assigned to different associated virtual routers.



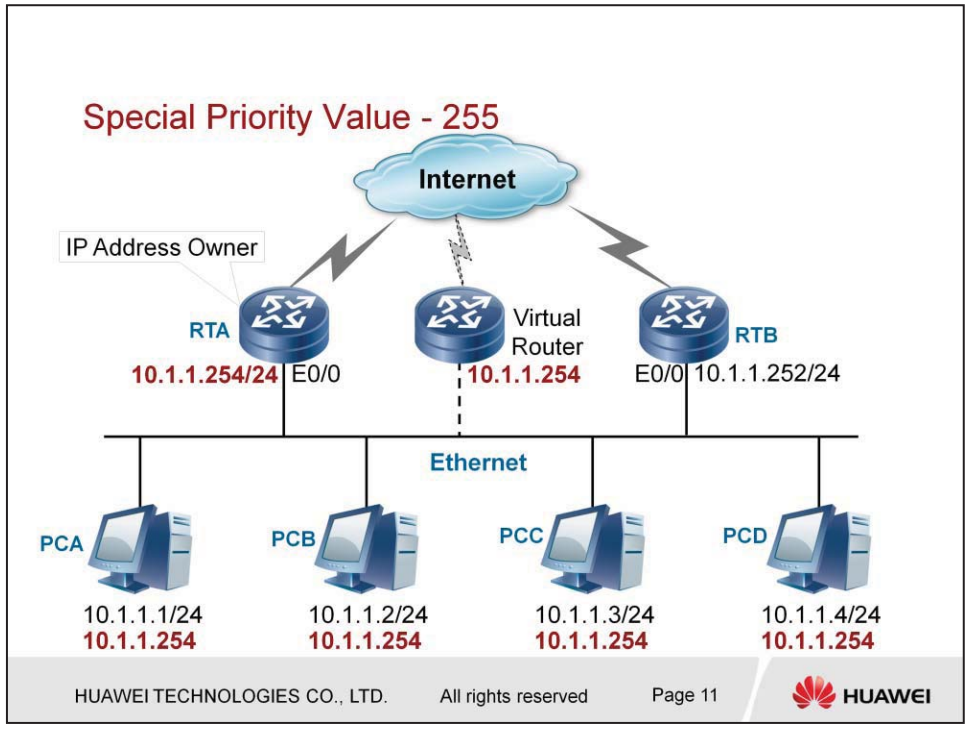
Config Priority：The configured Priority, the default value is 100.

Run Priority：The Priority used when the protocol is running; usually it is the same as Config Priority.

The Priority is in the range of 0-255. The value 255 is reserved for the IP address owner; and the VRRP packet with Priority 0 is used to trigger the immediate changeover from Backup to Master.

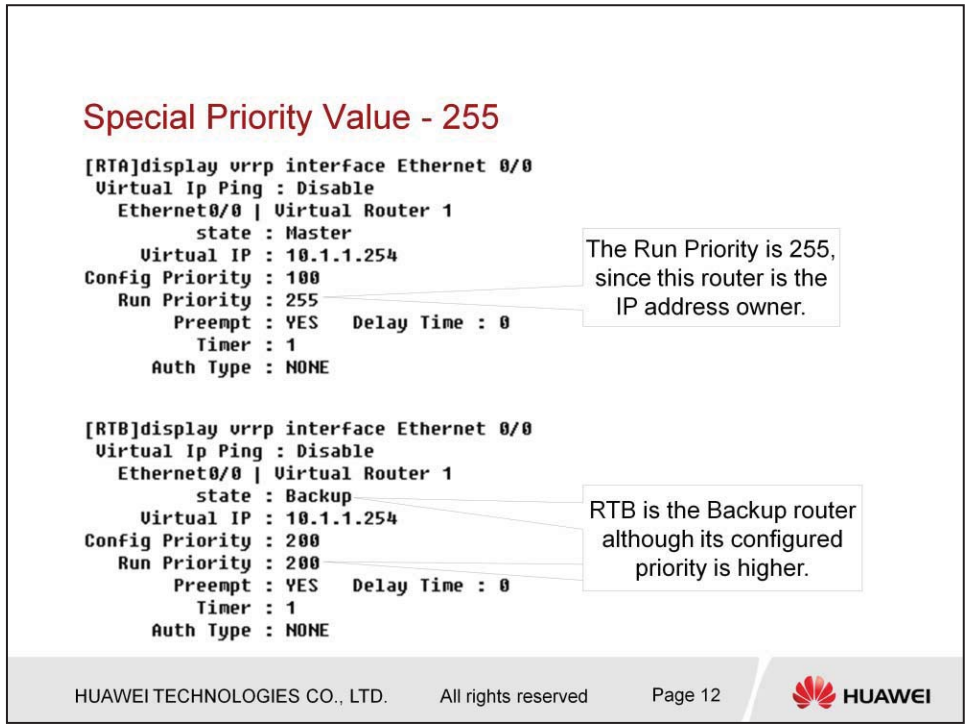
In this case:

Since the priority of RTA is 100, which is lower than the priority 200 of RTB, RTB will be the Master while RTA is the Backup.



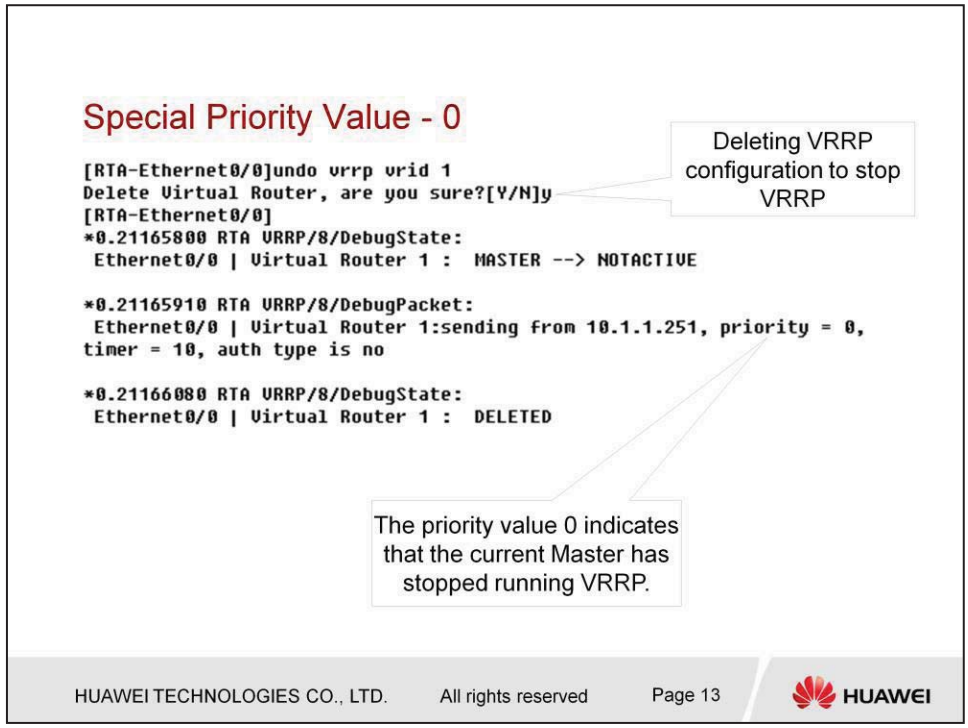
In this case:

There is a VRRP router that has the virtual router's IP address as real interface address. Such a router is called the IP Address Owner.

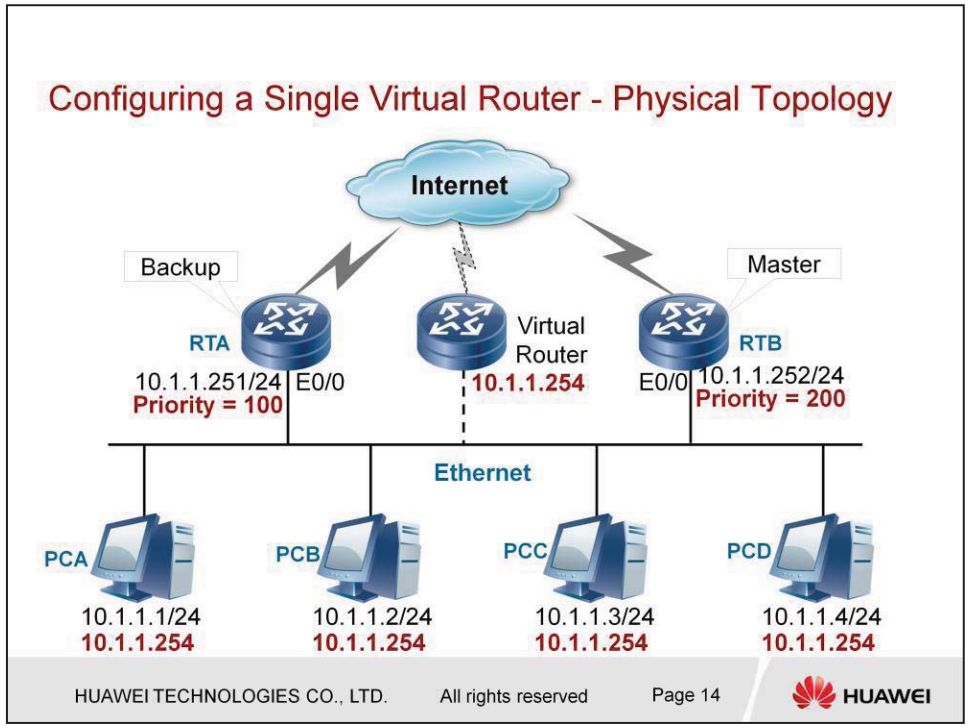


No matter what Config Priority is, the Run Priority of IP address owner is always 255. The IP address owner is always the Master.

Although the configured priority value of RTB is higher than that of RTA, the RTB is still the Backup, since its Run Priority is lower than that of RTA. Hence, when it comes to the election of Master, the contributing factor is the value of Run Priority instead of Config Priority.

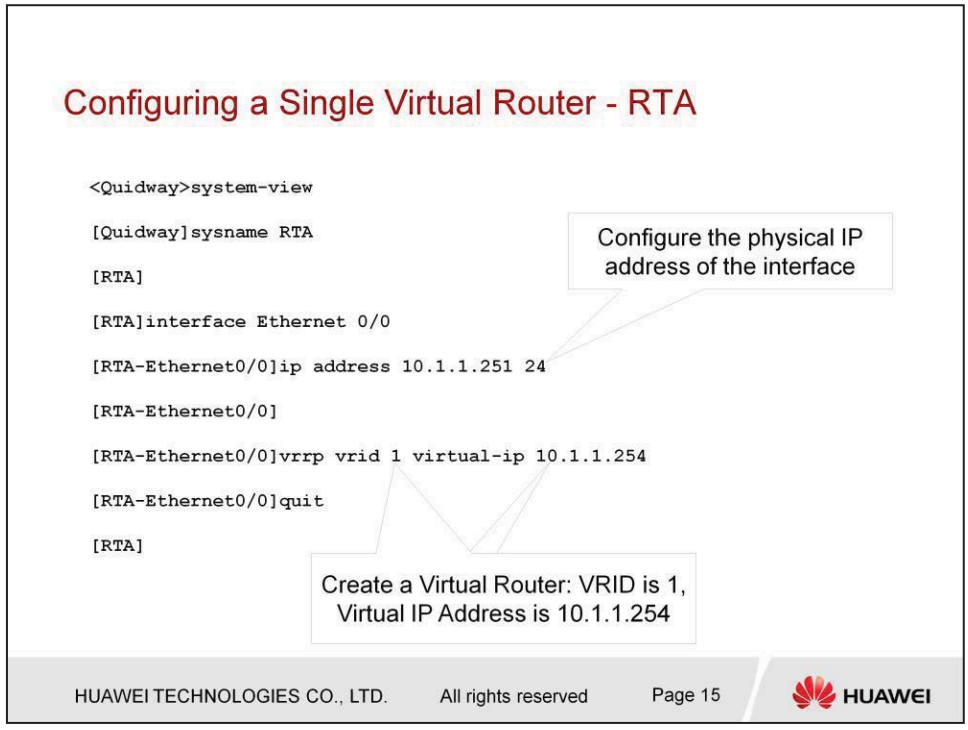


When the Master stops running VRRP, it will immediately send a VRRP advertisement with the value of 0 in Priority field. When the Backup receives such an advertisement, it will change from the Backup to Master state immediately.



In this case:

There are two routers in this LAN, RTA and RTB. A single Virtual Router is to be configured, with VRID 1 and Virtual IP Address 10.1.1.254. The Priority of RTB is to be configured as 200, and that of RTA as 100, so as to make RTB the Master.



The VRRP is configured under the interface view.

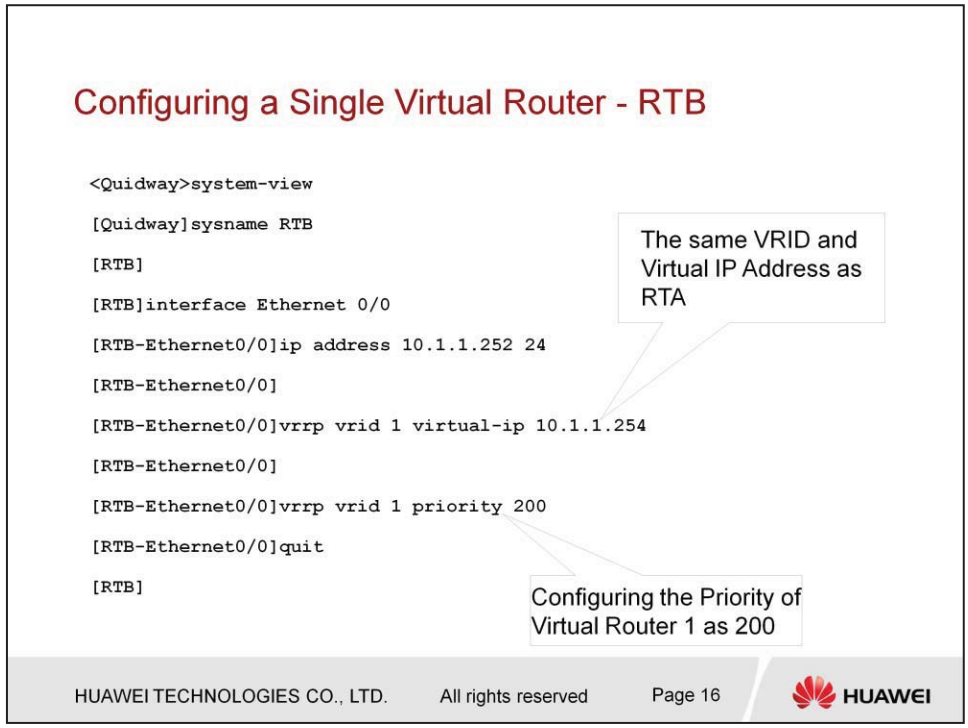
vrrp vrid virtual-router-ID virtual-ip virtual-address

undo vrrp vrid virtual-router-ID virtual-ip [virtual-address]

virtual-router-ID：The identifier of Virtual Router, in the range of 1-255.

virtual-address：Virtual IP address.

By default, if the Priority of virtual router is not designated, the default value is 100.



The VRID and Virtual IP Address should be the same as configured on RTA.

vrrp vrid virtual-router-ID priority priority-value

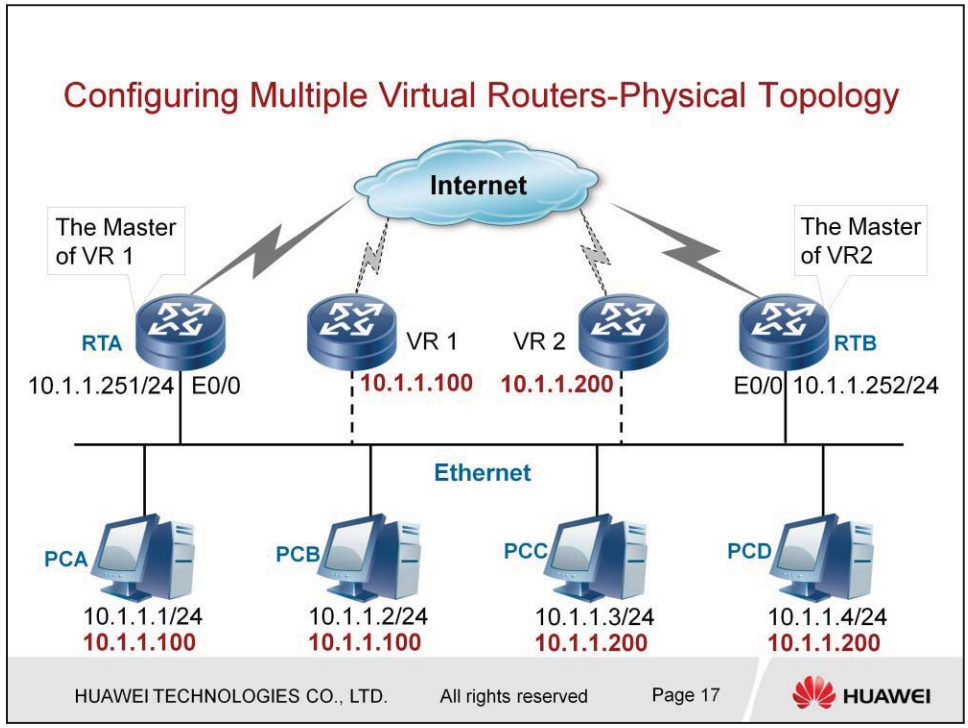
undo vrrp vrid virtual-router-ID priority

virtual-router-ID：The identifier of Virtual Router, in the range of 1-255.

priority-value：The value of Priority, with configured range from 1 to 254.

When configuring the priority, the VRID should be specified.

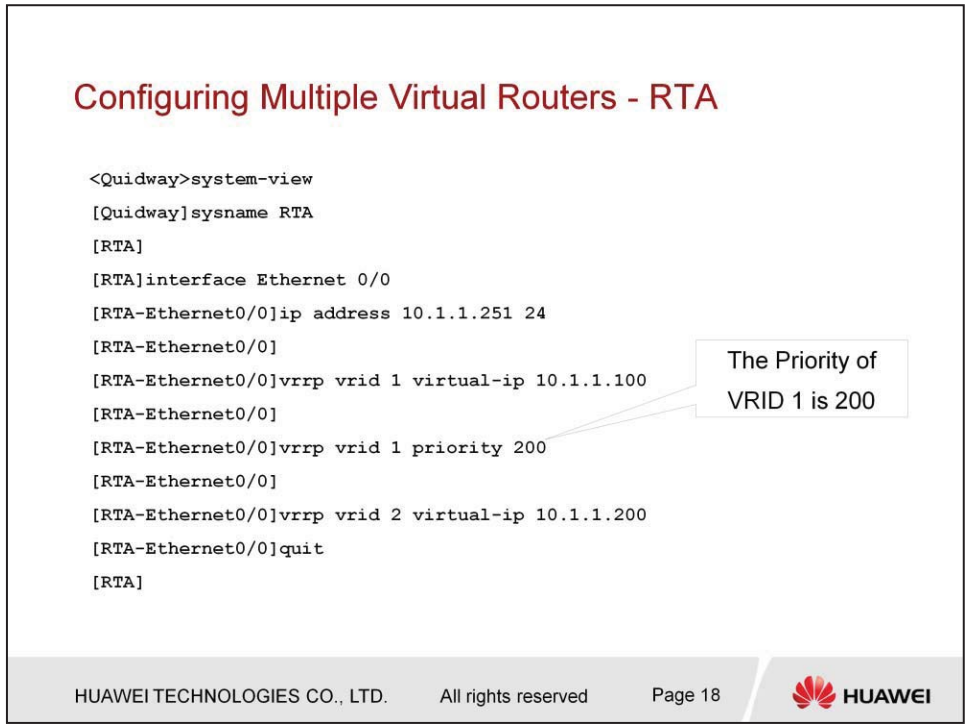
Different Virtual Routers can be configured with different Priority values.



In this case:

There are two routers on the LAN. Two Virtual Routers are to be configured. One of them is with VRID 1 and Virtual IP Address 10.1.1.100; the other with VRID 2 and Virtual IP Address 10.1.1.200. Configuring the Priority of Virtual Router 1 as 200 on RTA while 100 on RTB, so that in Virtual router 1, RTA is the Master. Configuring the Priority of Virtual Router 2 as 200 on RTB while 100 on RTA, so that in Virtual router 2, RTB is the Master.

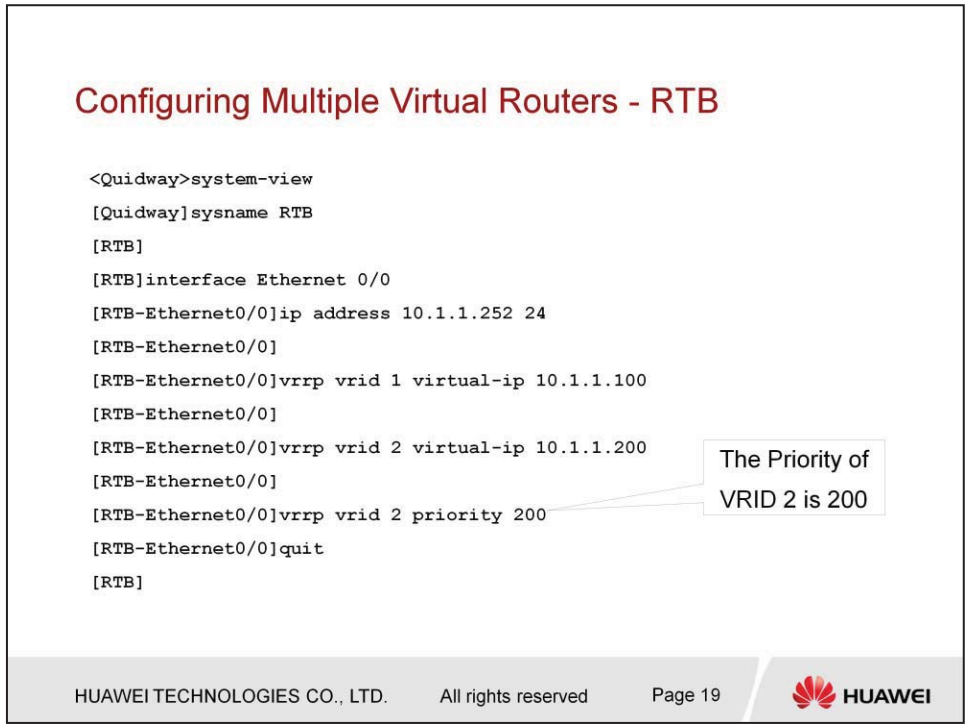
RTA is the Master of Virtual Router 1 and the Backup of Virtual Router 2; RTB is the Master of Virtual Router 2 and the Backup of the Virtual Router. In the LAN, PCs can use different Virtual IP addresses as the default gateway, so as to implement traffic sharing.



On RTA, configure two Virtual Routers as follows:

Virtual Router 1: Virtual IP address 10.1.1.100, Priority as 200;

Virtual Router 2: Virtual IP address 10.1.1.200, Priority as 100 (default).

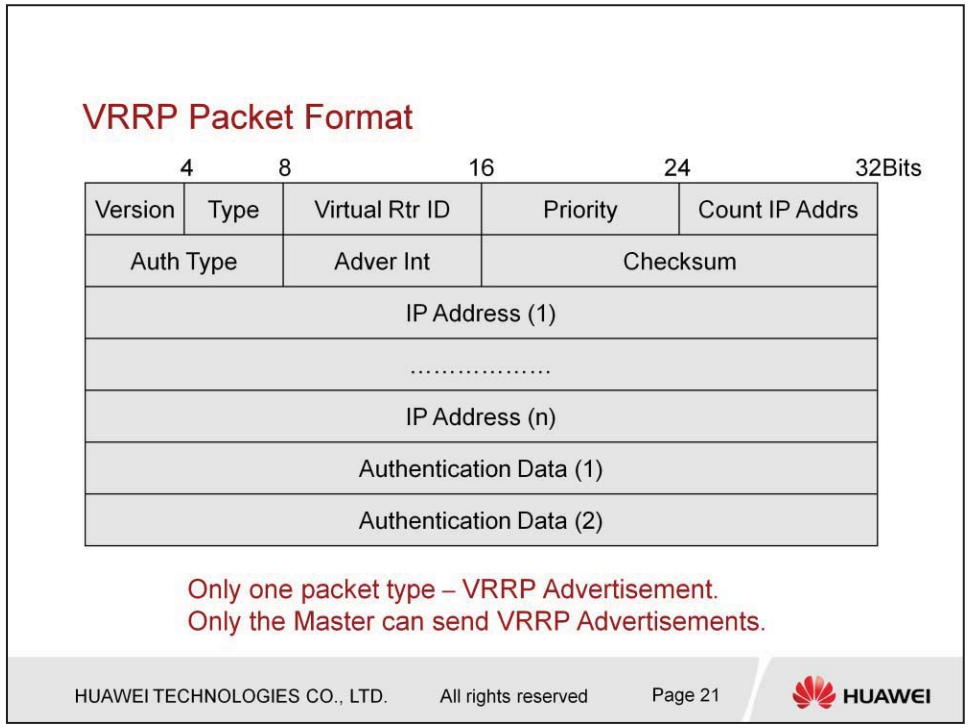


On RTB, configure two Virtual Routers as follows:

Virtual Router 1: Virtual IP address 10.1.1.100, Priority as 100 (default);

Virtual Router 2: Virtual IP address 10.1.1.200, Priority as 200.


Contents
Basic working principles
Protocol messages and state machine
Special features
HUAWEI TECHNOLOGIES CO., LTD. All rights reserved Page 20
HUAWEI



Version：The version field specifies the VRRP protocol version of this packet. By now, it is fixed to 2.

Type：The type field specifies the type of this VRRP packet.

The only packet type defined in this version of the protocol is:

1 Advertisement.

Virtual Rtr ID：The Virtual Router Identifier (VRID) field identifies the virtual router this packet is reporting status for.

Priority：The priority field specifies priority of the routers who sends the message.

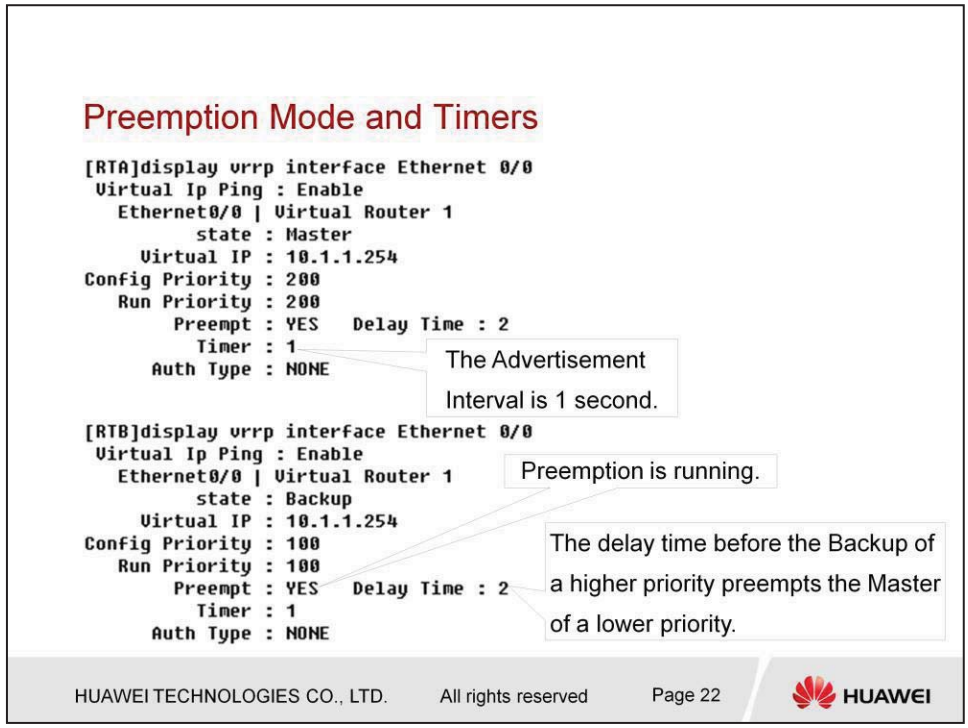
Count IP Addrs ： The number of IP addresses contained in this VRRP advertisement.

Auth Type：The authentication type field identifies the authentication method utilized.

Adver Int：The Advertisement interval indicates the time interval (in seconds) between ADVERTISEMENTS. The default is 1 second.

IP Address：One or more IP addresses that are associated with the virtual router.

Authentication Data： The password information needed for the authentication.



Advertisement_Interval：

Time interval between ADVERTISEMENTS (seconds), the default is 1 second. The Advertisement_Interval configured in all the VRRP routers associated to the same Virtual router should be the same. Or, in VRRP, it is regarded as the VRRP routers are associated to different Virtual Routers.

Preemption Mode：

If Preemption Mode is enabled, a Backup router with higher priority than the current Master router will change its Backup state to Master state. If Preemption Mode is disabled, even though a Backup router is of higher priority than the current Master router, it will not change its Backup state to Master state, the Preemption Mode is enabled by default.

Delay Time：

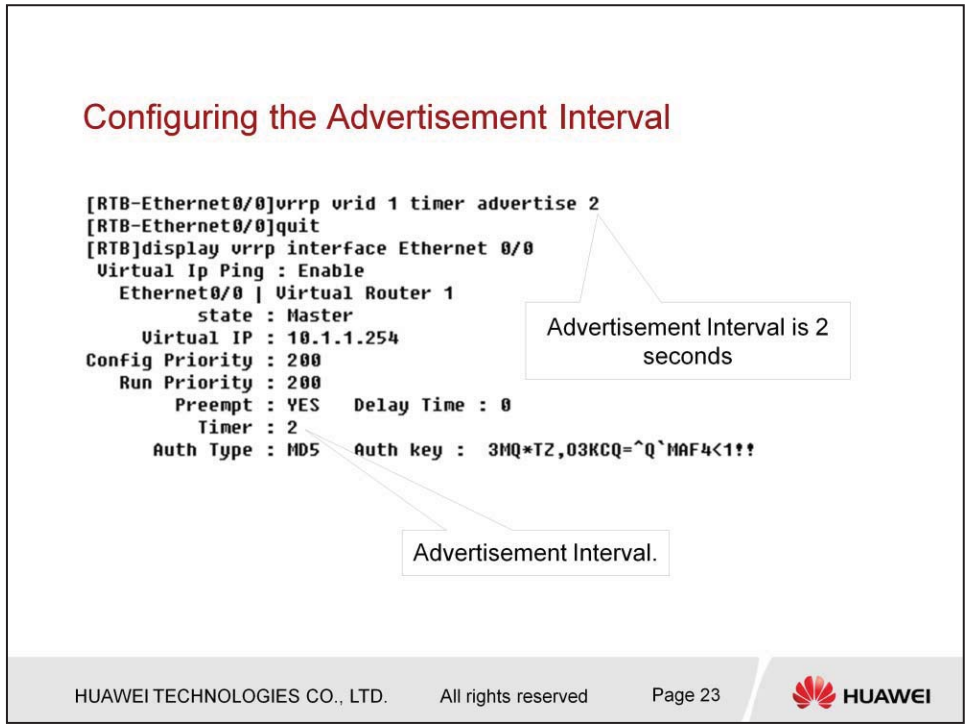
1：To delay the interval of Master failure. (Master_Down_Interval）

2：The delay before the Backup preempts the Master.

The default value is 0.

Master_Down_Interval：

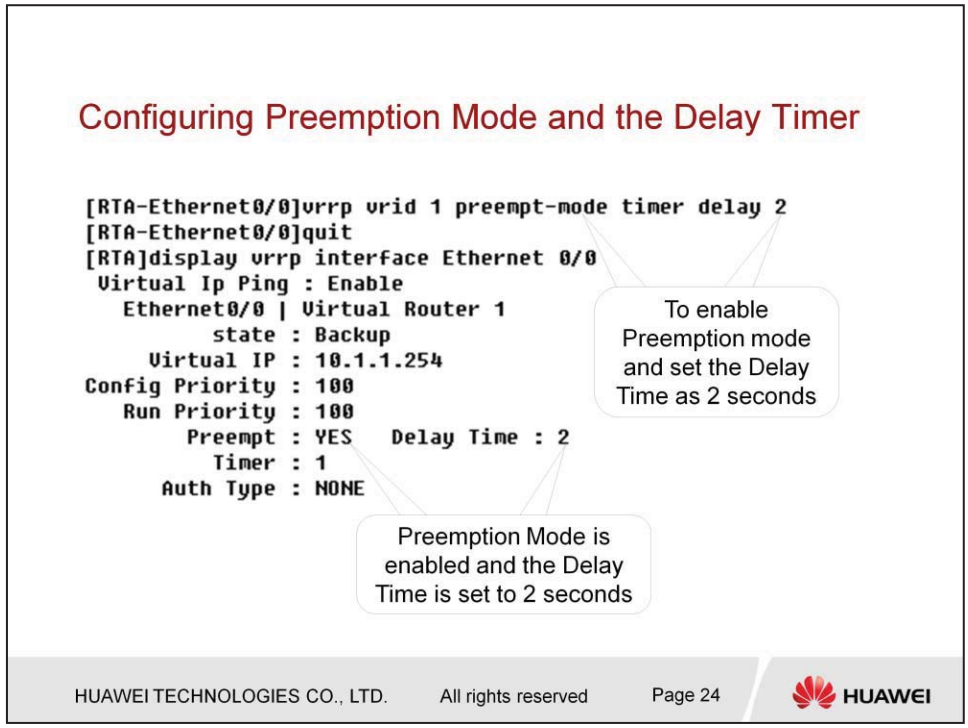
(3*Advertisement_Interval) + Delay Time.



The Advertisement_Interval configured in all the VRRP routers associated to the same Virtual router should be the same.

vrrp vrid virtual-router-ID timer advertise adver-interval undo vrrp vrid virtual-router-ID timer advertise virtual-router-ID： The identifier of Virtual Router, in the range of 1-255.

adver-interval：Time interval between ADVERTISEMENTS send by the Master, in the range of 1-255. Default value is 1 second. In this example, the Advertisement_Interval is modified to 2 seconds.



Preemption Mode：

Controls whether a higher priority Backup router preempts a lower priority Master.

vrrp vrid virtual-router-ID preempt-mode [timer delay delay-value]

undo vrrp vrid virtual-router-ID preempt-mode

virtual-router-ID：The identifier of Virtual Router, in the range of 1-255.

delay-value：It is the Delay Time, in seconds, in the range of 0-255. the preemption mode is enabled by default, and the delay time is 0 seconds. If the network is very busy, sometimes, the Backup cannot receive the Advertisement although the Master works properly. Under this circumstance, the Delay Time could be modified to prevent the Backup becoming the Master immediately, so as to reduce the network flapping. In the previous slide, the Advertisement_Interval is 2 seconds. In this example, the Delay Time is configured as 2 seconds. Hence, the Master_Down_Interval is 3*2+2=8 seconds.

Packet Authentication

Auth Type	Authentication Method
0	No Authentication
1	Simple Text Password
2	MD5 Password

HUAWEI TECHNOLOGIES CO., LTD. All rights reserved

No Authentication

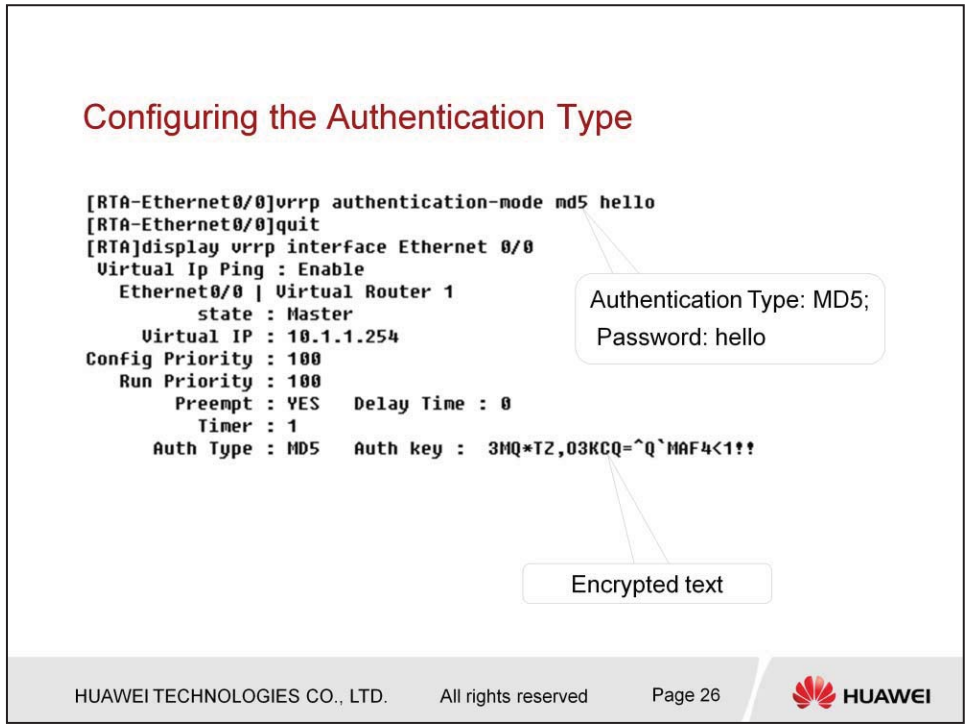
The use of this authentication type means that VRRP protocol exchanges are not authenticated. The contents of the Authentication Data field should be set to zero on transmission and ignored on reception.

Simple Text Password:

The use of this authentication type means that VRRP protocol exchanges are authenticated by a clear text password. The contents of the Authentication Data field should be set to the locally configured password.

MD5 Password:

The use of this authentication type means the VRRP protocol exchanges are authenticated using MD5 encrypted message. If the authentication type in the received VRRP packet is different from that locally configured, the received VRRP packet will be discarded.



Three types of authentication modes are supported in VRRP:

No Authentication, Simple Text Password, and MD5 Password. By default, there is no authentication. For VRRP routers associated to the same Virtual Router, the authentication mode and password information configured must be same.

vrrp authentication-mode { md5 key | simple key }

undo vrrp authentication-mode

simple： Means that VRRP protocol exchanges are authenticated by a clear text password.

md5：Means the VRRP protocol exchanges are authenticated using MD5 encrypted message.

key：Authentication word. When the authentication mode is simple, the key is in plain text, in the range of 1-8 bits; When the authentication mode is md5, the key is using MD5, if it is input in plain text, the length is in the range of 1-8 bits. (For example:1234567); if it is input using encrypted text, the length must be 24 bits. (For example: _(TT8F]Y\5SQ=^Q`MAF4<1!!.

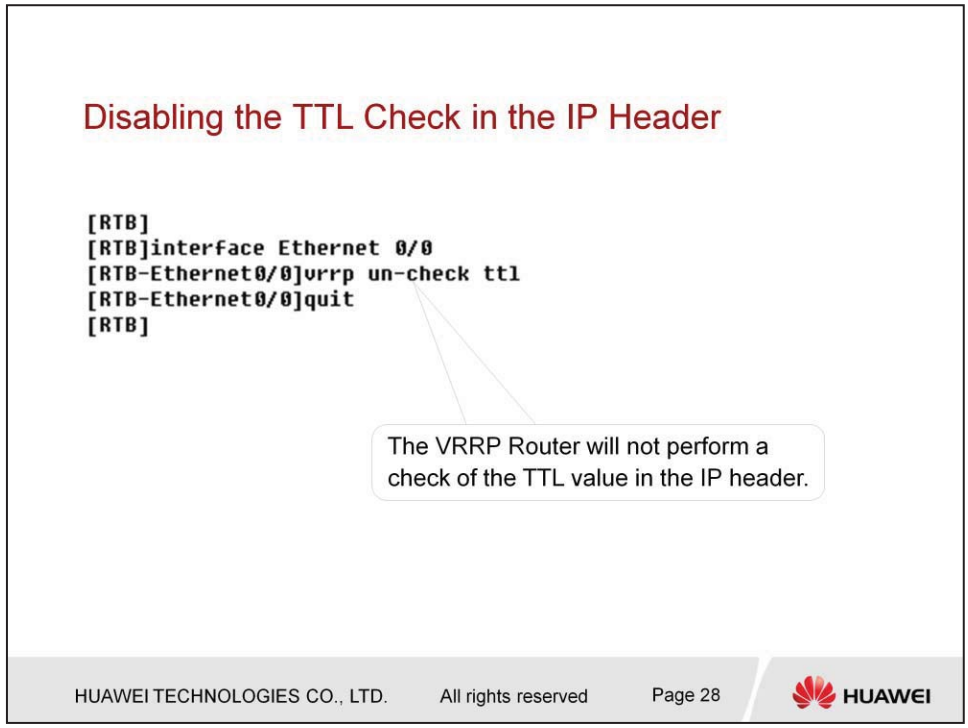
IP Fields of VRRP Packets

Field	Value
Source IP Address	Physical IP address of Master's interface sending the packet.
Destination IP Address	IP multicast address 224.0.0.18
TTL	Must be 255.
Protocol	The IP protocol number is 0x70 in VRRP.

HUAWEI TECHNOLOGIES CO., LTD. All rights reserved HUAWEI

The VRRP packet is directly encapsulated in the IP packet.

In the IP header, the source address is the IP address of the nterface the packet is being sent from. The Destination Address is a IP ulticast address: 224.0.0.18. This is a link local scope multicast address. Routers MUST NOT forward a datagram with this destination address regardless of its TTL. The TTL MUST be set to 255. A VRRP router receiving a packet with the TTL not equal to 255 MUST discard it.The IP protocol number assigned by the IANA for VRRP is 112 (decimal).

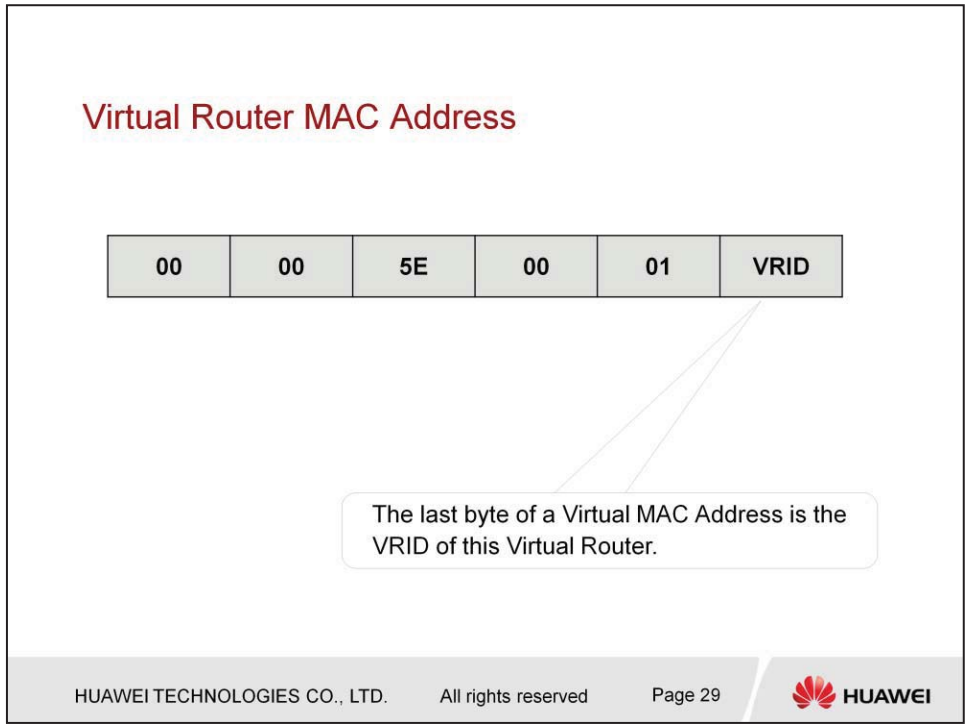


The VRRP router will by default check the TTL field of IP headerin the received VRRP packet, and the TTL value must be 255.

On the VRP platform, the function could be disabled by the following command.

vrrp un-check ttl

undo vrrp un-check ttl



Every virtual router has a virtual MAC address, in the format:

00-00-5E-00-01-{VRID}.

The first 5 bytes of the virtual MAC address is fixed. The final byte of the virtual MAC address is the VRID of the virtual router.

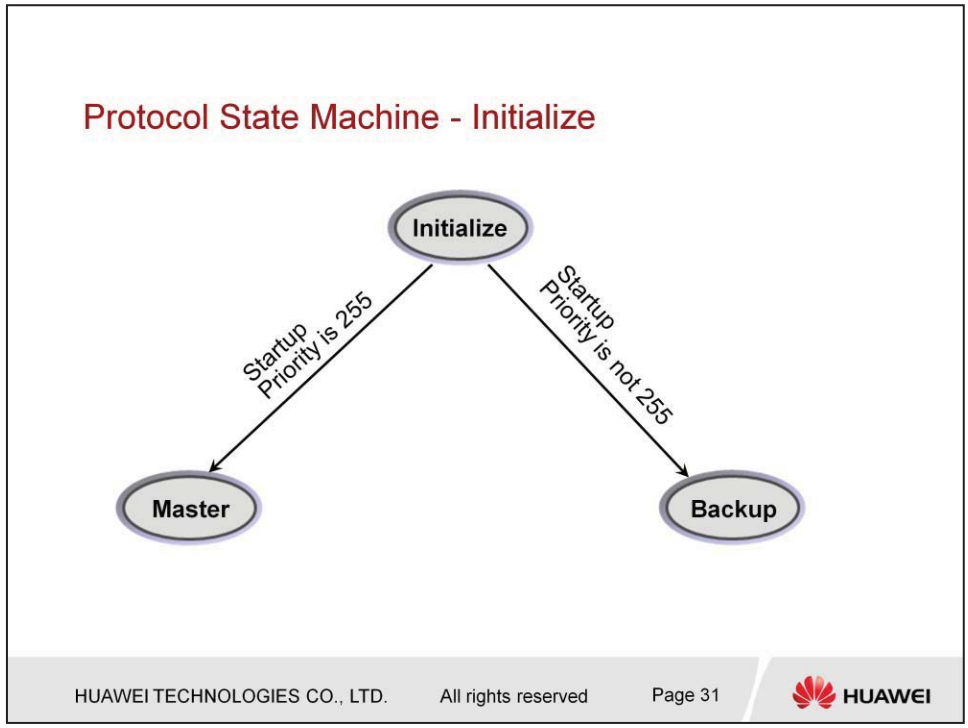
Layer 2 Fields of VRRP Packets

Field	Value
Destination MAC Address	Multicast MAC Address: 01-00-5E-00-00-12
Source MAC Address	The MAC Address of Virtual Router
Type	0x0800

HUAWEI TECHNOLOGIES CO., LTD. All rights reserved

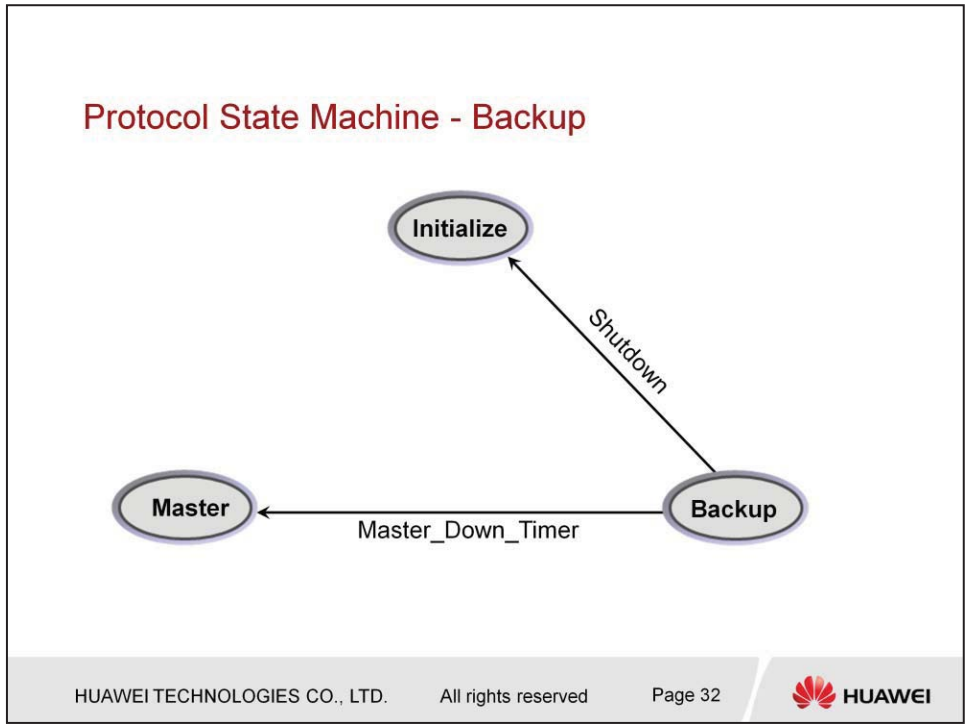
When the VRRP router sends out VRRP advertisement periodically:

The destination MAC address is the corresponding multicast MAC address of the Layer 3 IP address 224.0.0.18. The source MAC address is the virtual MAC address of the virtual router. The virtual router MAC address is used as the source in all periodic VRRP messages sent by the Master router to enable bridge learning in an extended LAN, to update the MAC table of the layer 2 switch through which the VRRP packet is passed through, and to ensure all the packets sent to the virtual router could be forwarded properly by the switch. The type field set to be 0x0800 indicates the packet encapsulated in the frame is the IP packet.



Initialize:

The purpose of this state is to wait for a Startup event. A Startup event could be triggered automatically as long as the VRRP is configured completely; also, if the VRRP has been already configured in some interface, the Startup event could be triggered by the change of underlying layer from bottom to top. If a Startup event is received, and the Priority = 255 (i.e., the router owns the IP address) associated with the virtual router, the state will be transitioned to Master state, and the router begins to send VRRP packet periodically. If a Startup event is received, and the Priority is not 255, the state will be transitioned to Backup state, and the router will begin to monitor the state of Master Router.



Backup:

The purpose of the Backup state is to monitor the availability and state of the Master Router. While in this state, a VRRP router MUST do the following: MUST NOT respond to ARP requests for the IP address associated with the virtual router; MUST discard packets with a destination link layer MAC address equal to the virtual router MAC address; MUST NOT accept packets addressed to the IP address associated with the virtual router.

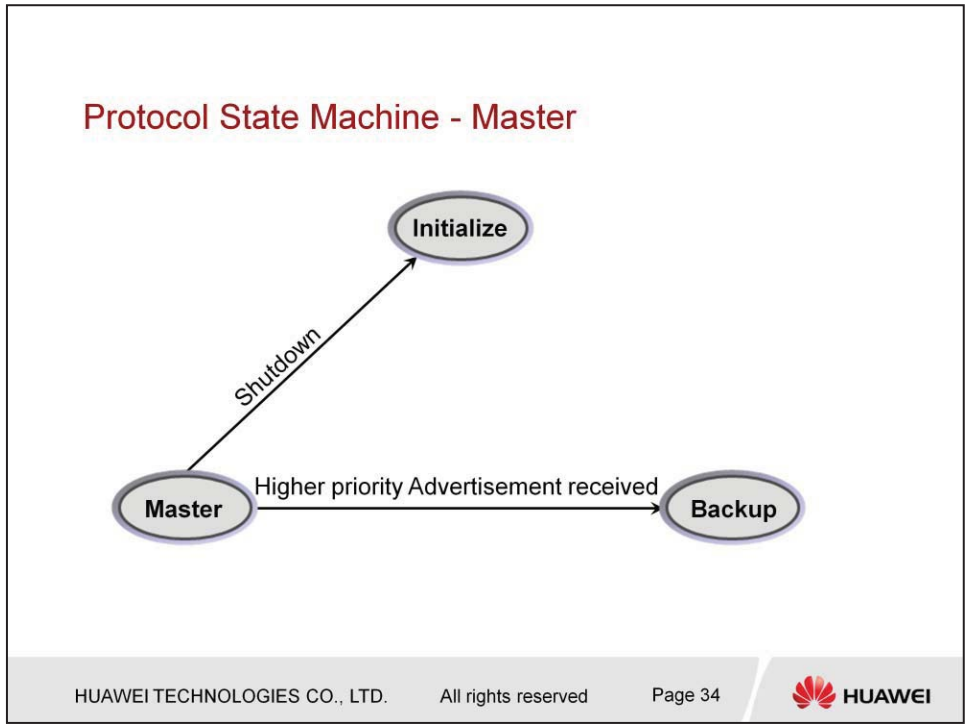
If a Shutdown event is received, such as the deletion of relative configuration of VRRP, then the router stops monitoring the statre of Master Router and transit to the Initialize state.

If the Master_Down_Timer is set, the state is changed to Master, and the VRRP advertisement begins to be sent periodically. If an ADVERTISEMENT is received whose Priority is Zero, the Master_Down_Timer is set to be Delay time. And if the Delay Time is 0, the Master_Down_Timer will fire immediately, and state transforms

to Master. If an ADVERTISEMENT is received whose Priority is greater than or equal to the local Priority; or If the Priority in the ADVERTISEMENT is lower than the local Priority, but Preempt_Mode is False, then the Master_Down_Timer will be reset to 0 and the router continues to monitor the state of Master Router.

If an ADVERTISEMENT is received whose Priority is lower than the local Priority, and Preempt_Mode is True, then the ADVERTISEMENT

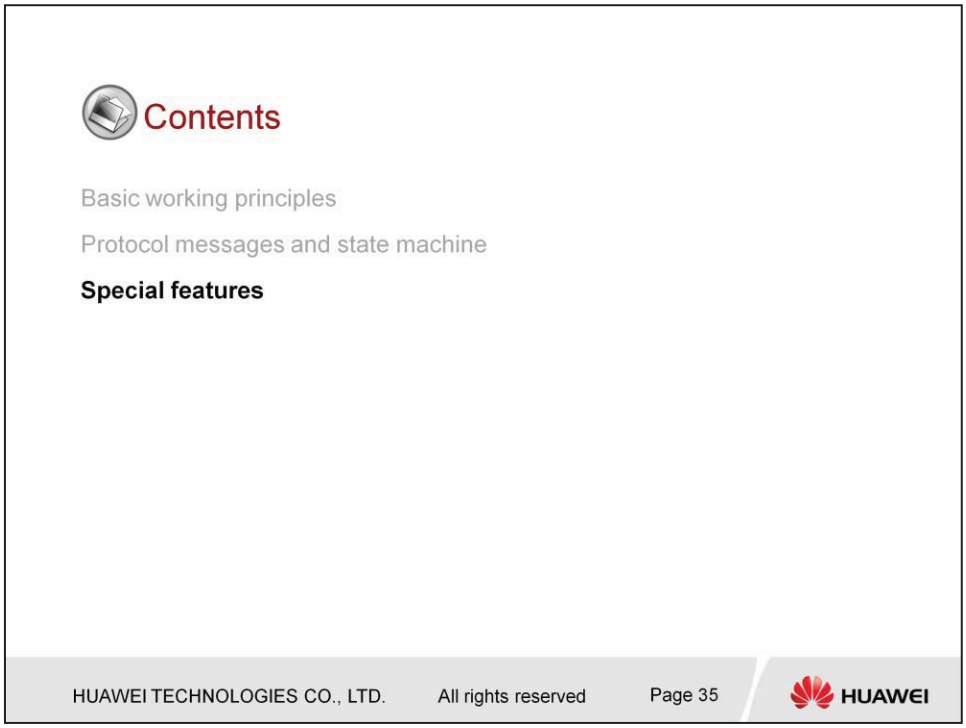
will be discarded. The Master_Down_Timer will be increased along the time. After the timer fires, the state will be changed to Master.

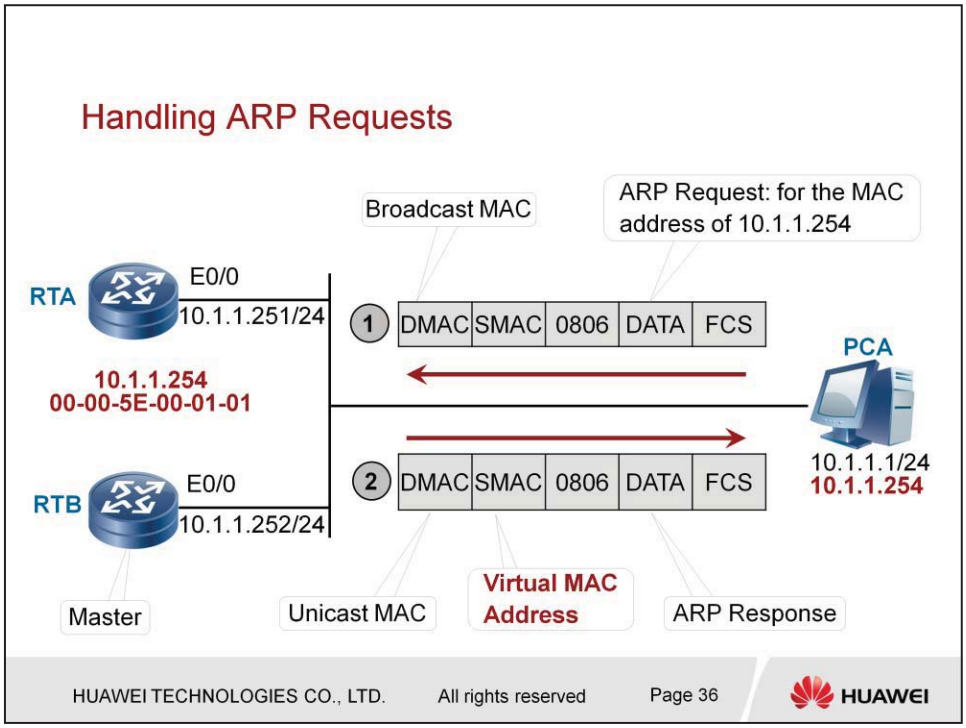


Master:

While in the Master state the router functions as the forwarding router for the IP address associated with the virtual router. While in this state, the VRRP router must respond to ARP requests for the IP address associated with the virtual router and must forward packets with a destination link layer MAC address equal to the virtual router MAC address. In the Master state, the VRRP router sends out VRRP Advertisement periodically. After every VRRP Advertisement is sent, the Advertise_Timer is reset to 0. And when the Advertise_Timer fires, the next VRRP Advertisement will be sent. If a Shutdown event is received, the Advertise_Timer is canceled, a VRRP Advertisement with Priority 0 is sent, and the state is changed to Initialize. If an ADVERTISEMENT is received whose Priority is greater than the local Priority; or If the Priority in the ADVERTISEMENT is equal to the local Priority, butthe remote IP address is higher than that of local, then the

Advertise_Timer is canceled, and a VRRP Advertisement with Priority 0 is sent, and the state is changed to Backup.


Contents
Basic working principles
Protocol messages and state machine
Special features
HUAWEI TECHNOLOGIES CO., LTD. All rights reserved Page 35
HUAWEI



In this case:

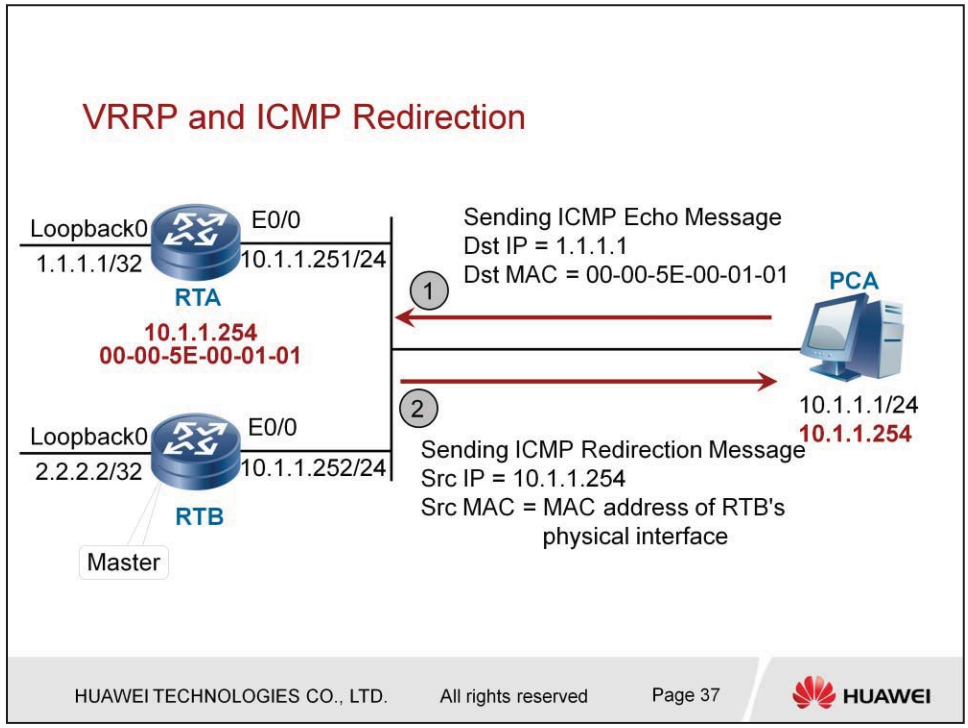
RTB is the Master, the Virtual IP Address is 10.1.1.254, and the Virtual MAC Address is 00-00-5E-00-01-01. The process of sending and receiving ARP packets is as follows:

PCA sends a ARP Request packet using Broadcast MAC Address as the Destination MAC Address, requesting the MAC Address for 10.1.1.254. The Type field set to 0x0806 indicates this is an ARP packet.

2. After that ARP Request packet is received, RTB (the Master)sends a ARP Response packet, using MAC Address of PCA as Destination MAC Address,

while the Source MAC Address is the Virtual MAC Address (00-00-5E-00-01-01). The Type field set to 0x0806 indicates this is an ARP packet. The ARP request could only be handled by the Master Router (RTB), instead of Backup Router (RTA).

There are only two situations under which the virtual MAC address is used as the source MAC address of the sent data frame: firstly, when the ARP response message is sent by the Master to respond the ARP request of the Virtual IP address; secondly, when the periodic VRRP advertisement is sent by the Master. In other cases, the MAC address of physical interface is always used as the source MAC address.

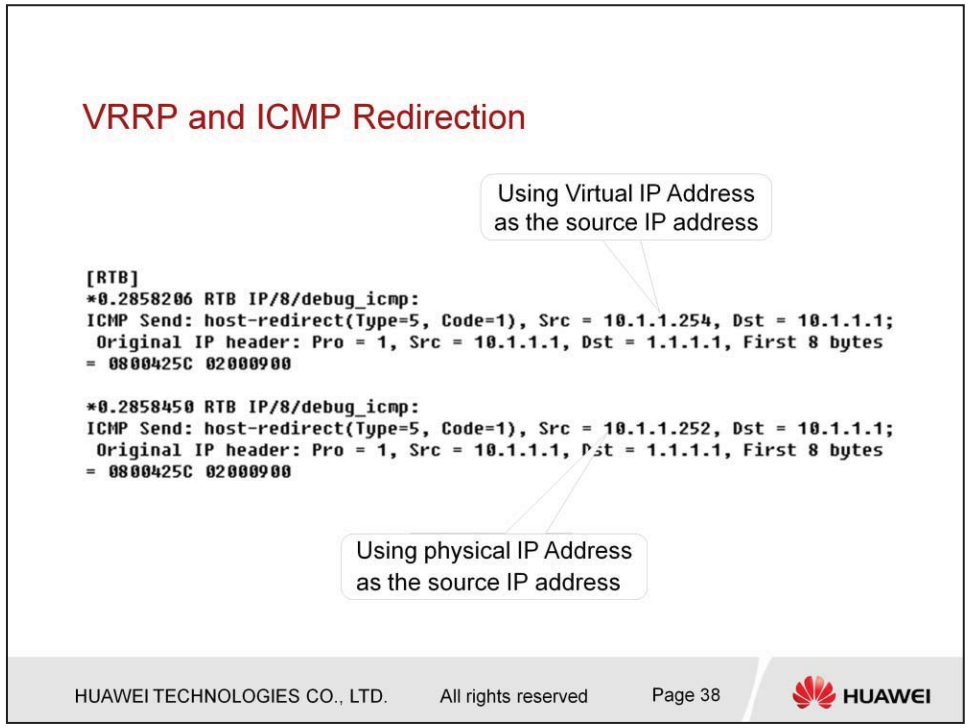


In this case:

RTA and RTB run OSPF to learn route information. VRRP is enabled on the Ethernet interface of RTA and RTB. The Virtual IP Address of the Virtual Router is 10.1.1.254, and the Virtual MAC Address is 00-00-5E-00-01-01. Through the configuration of Priority, RTB is set to be the Master Router.

When using command "ping 1.1.1.1" on PCA, a ICMP Echo Message is sent in which the Destination MAC is the Virtual Router' MAC Address, and the Destination IP address is 1.1.1.1. In routing-table of RTB, for the route item to 1.1.1.1/32, the outgoing interface is the same as the receiving interface of the Echo Message sent by PCA.

Hence, after the Echo Message is received by RTB (the Master), it will send a Redirect Message to PCA in order to let PCA send the ICMP Echo Message to RTA directly. The Source IP Address of the Redirect Message MUST be the Virtual IP Address of the Virtual Router, since the PC only processes the ICMP Redirect Message sent by the gateway.



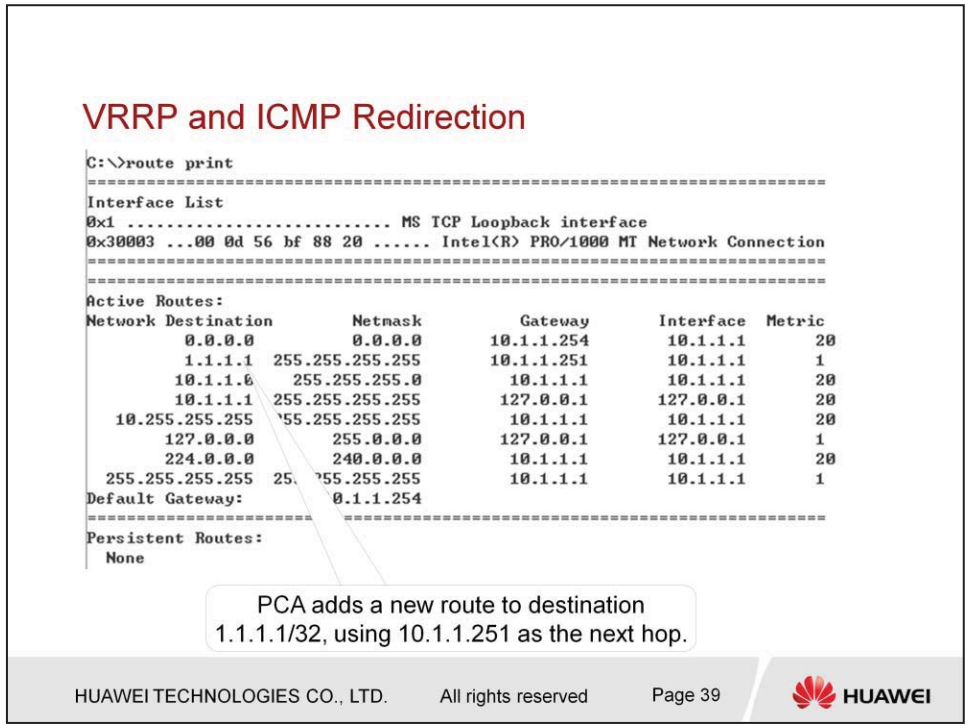
Since in the network it is possible to use the IP address of the physical interface of RTB as the default gateway, RTB will send two Redirect Messages.

One of them uses the Virtual IP Address as the source IP address, the other using the IP address of the physical interface as the source IP address.

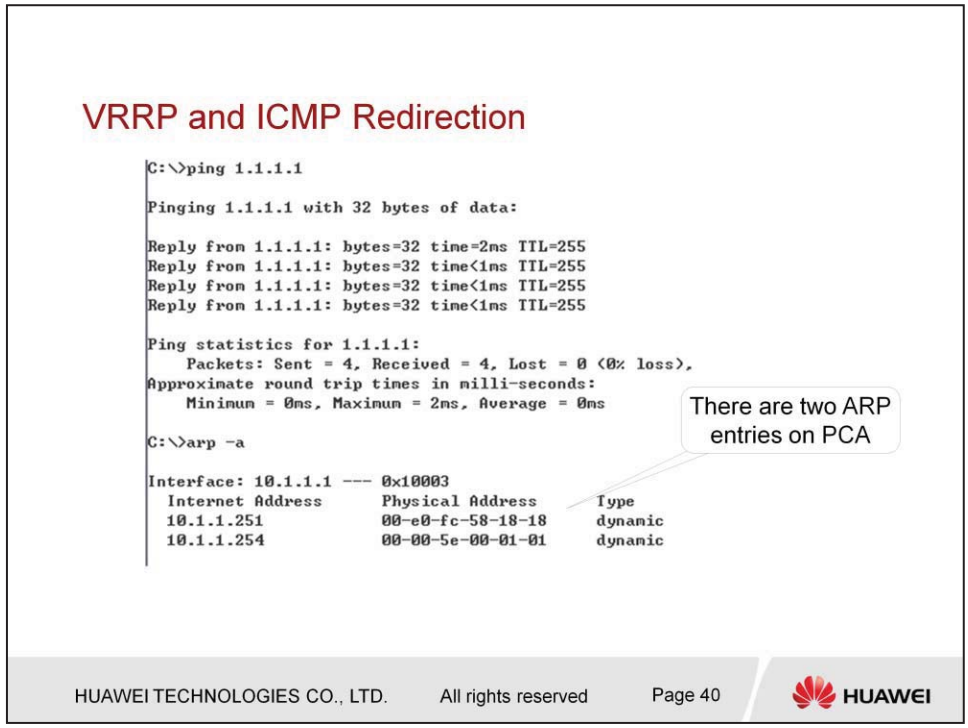
Which one is useful for PCA?

Only the ICMP Redirect Messages whose source IP address is set to the Virtual IP Address is useful for PCA, because a PC can only process Redirect Messages sent by the Default Gateway.

There is only one situation under which the virtual IP address is used as the source IP address: When the ICMP Redirect Message is sent by the VRRP router. In other cases, the source IP address is always the IP address of physical interface.

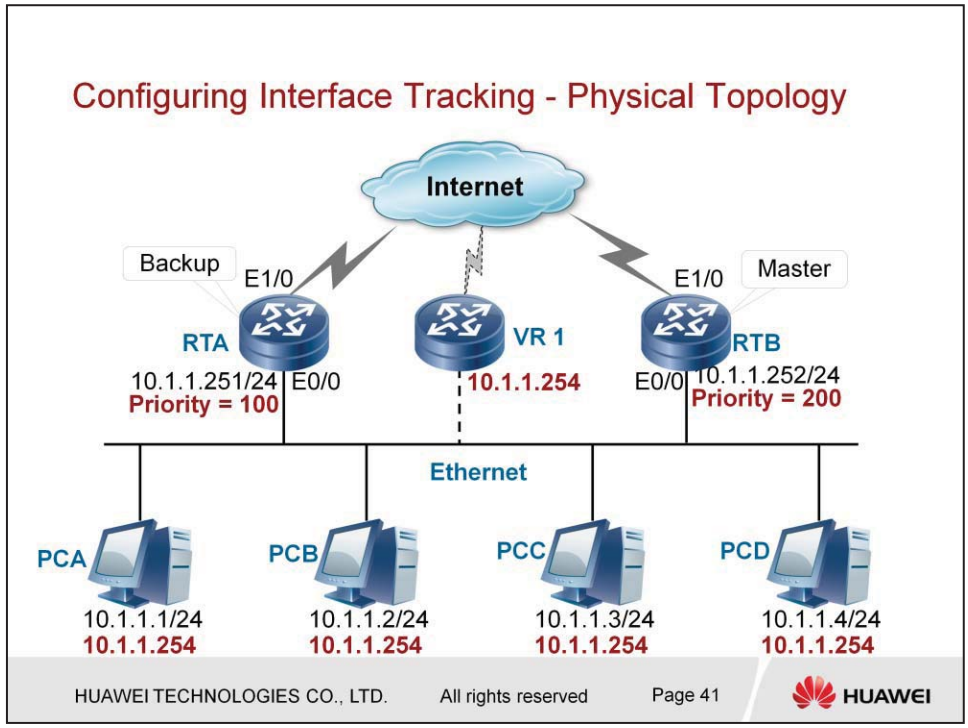


After the ICMP Redirect Message is processed by PCA, in the routing table of PCA, a new route item to destination 1.1.1.1/32 will be added. The next hop of the new route item is 10.1.1.251, which is the physical interface of RTA.



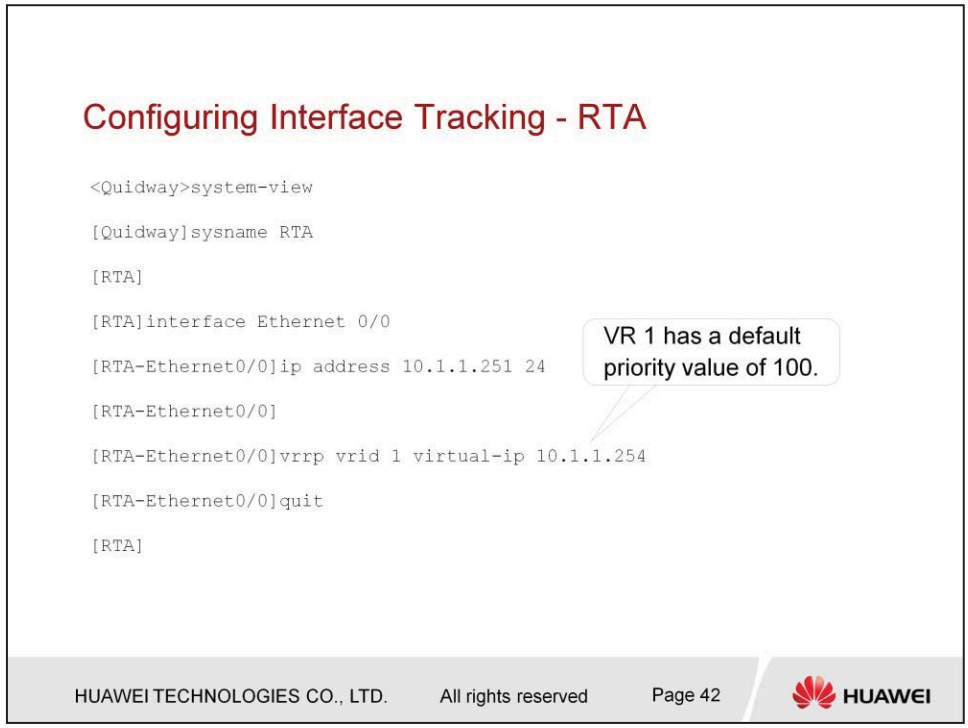
There are two ARP entries on PCA, one for the Virtual IP Address (default gateway), the other for the IP address of the physical interface of RTA (10.1.1.251).

It is because according to the routing table, the following data packets will be directly sent to the IP address 10.1.1.251, PCA needs request the MAC address of the IP address 10.1.1.251.

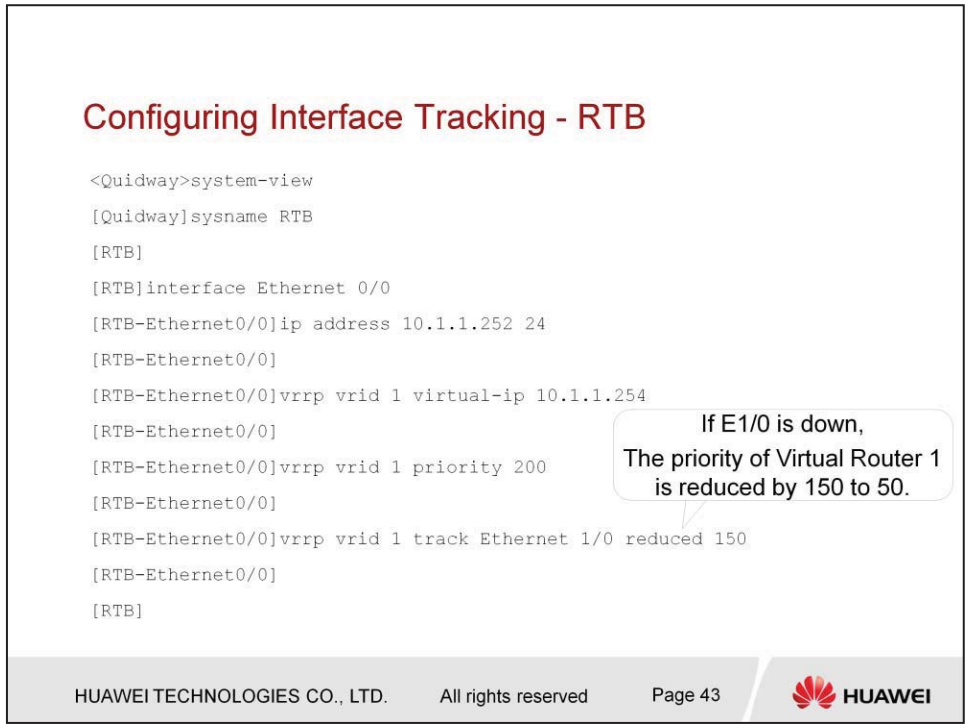


VRRP can track upstream interfaces. In this case:

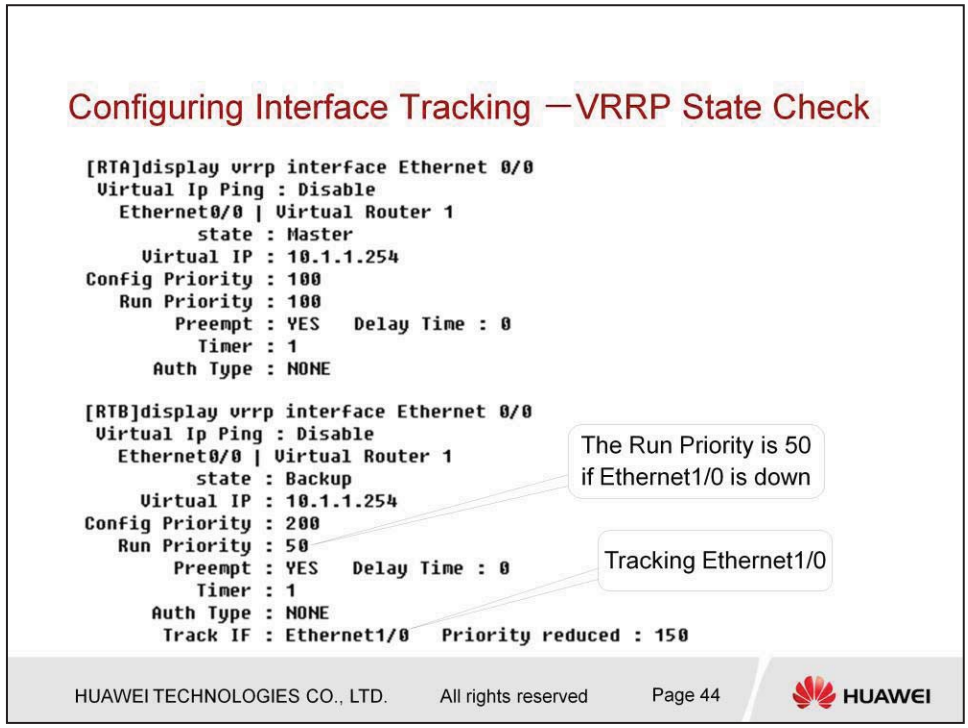
RTB is the Master Router. If the interface Ethernet 1/0 (WAN interface) of RTB is down, we hope RTA to be the new Master immediately. VRP supports such function by configuring on RTB to enable the Virtual Router tracking interface Ethernet 1/0. If the interface Ethernet 1/0 is down, the Priority of the Virtual Router would be reduced by a configured value to be a new value lower than that of RTA. Hence, RTA will be the new Master Router automatically. If the interface E1/0 of RTB recovers and works properly, the Priority of RTB will return to the original value, and RTB will be the Master again.



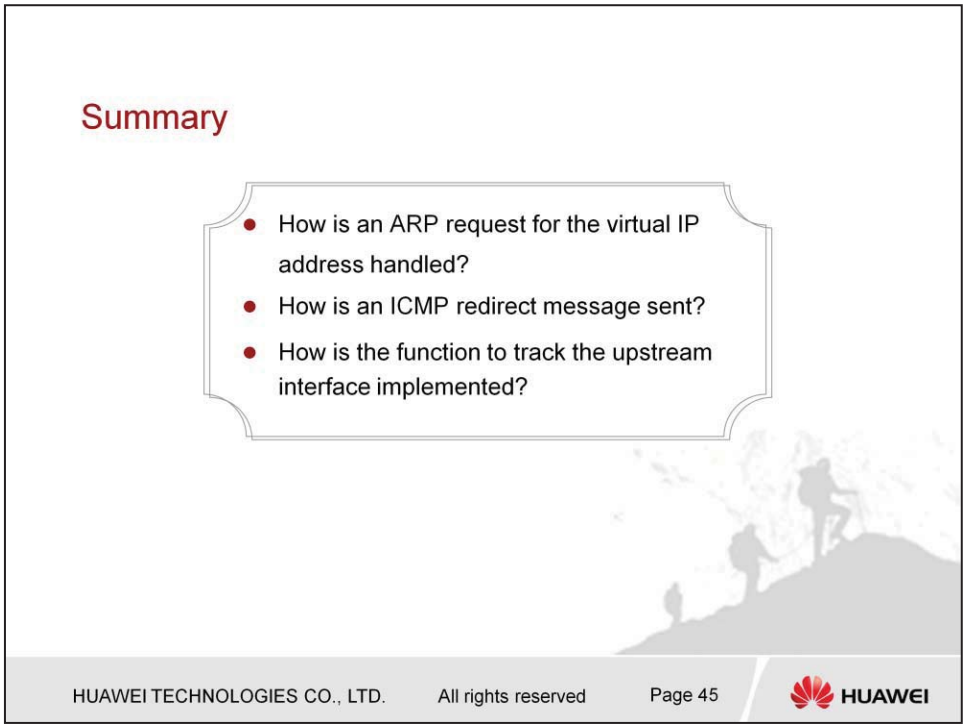
The configuration of RTA is the same as that of configuring a single Virtual Router, by default the Priority is 100.



Through configuring the Priority as 200, RTB recognized as the Master. Configuring tracking interface Ethernet 1/0 on RTB. If interface Ethernet 1/0 is down, the Priority is reduced by 150, and the new Priority is 50. Hence, RTA will be the new Master.



This is the VRRP States if the tracked interface is down. On RTB, although the Configured Priority is 200, the Running Priority is reduced to 50. Hence, RTA will become the Master.



This chapter covers the following points:

1. How is an ARP request for the virtual IP address handled?

When the VRRP interface receives the ARP request for the virtual IP address, only the Master could respond to the request, and in the responding message, the virtual MAC address is used as the source MAC address.

2. How is an ICMP redirect message sent?

When the ICMP redirect message needs to be sent on some VRRP interface, the virtual IP address should be used as the source IP address.

3. How is the function to track the upstream interface implemented?

By specific configuration, the value of Priority is made to change along with the change of state of the tracked interface.

Module 3
QoS


IP QoS Overview
www.huawei.com
HUAWEI TECHNOLOGIES CO., LTD.
All rights reserved
HUAWEI

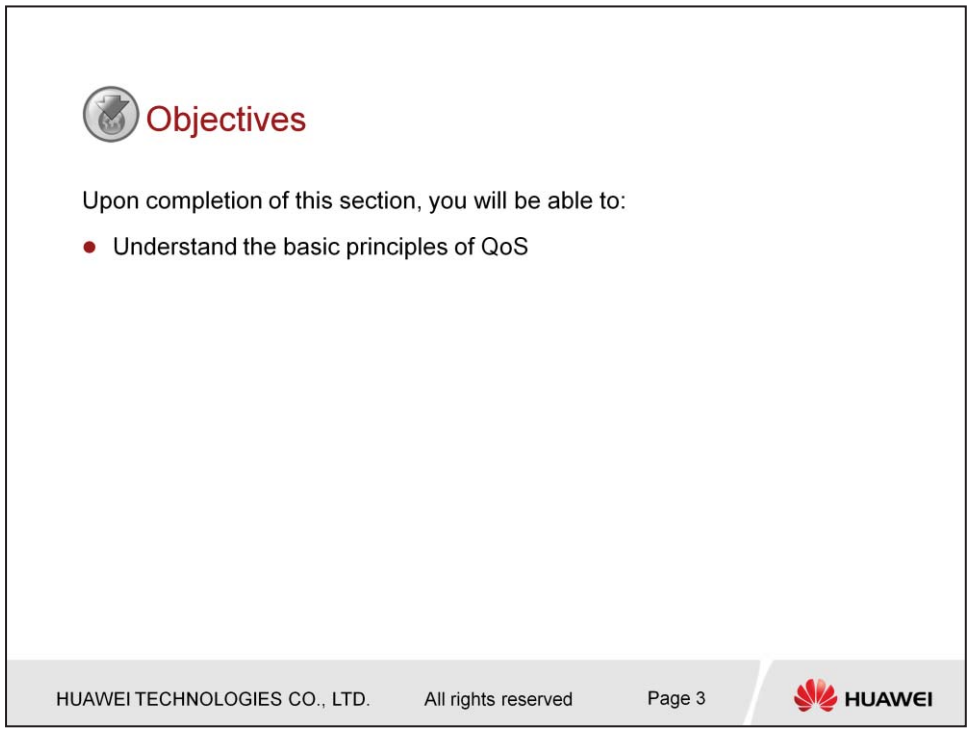


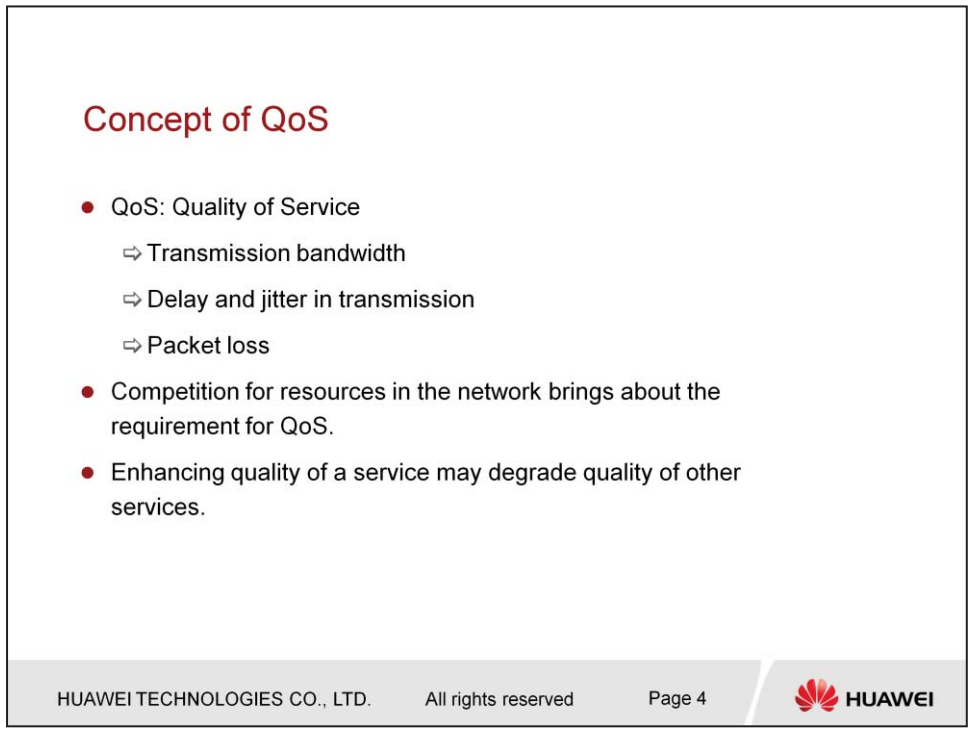
Foreword

This section mainly introduces basic concepts of IP QoS, including the influencing factors, and QoS models. At the same time puts forward the basic methods of guarantee supported by QoS.

HUAWEI TECHNOLOGIES CO., LTD. All rights reserved




Objectives
Upon completion of this section, you will be able to:
Understand the basic principles of QoS
HUAWEI TECHNOLOGIES CO., LTD. All rights reserved Page 3
HUAWEI



QoS stands for Quality of Service. Conventionally, quality of a network service includes the bandwidth, transmission delay, and packet loss ratio.

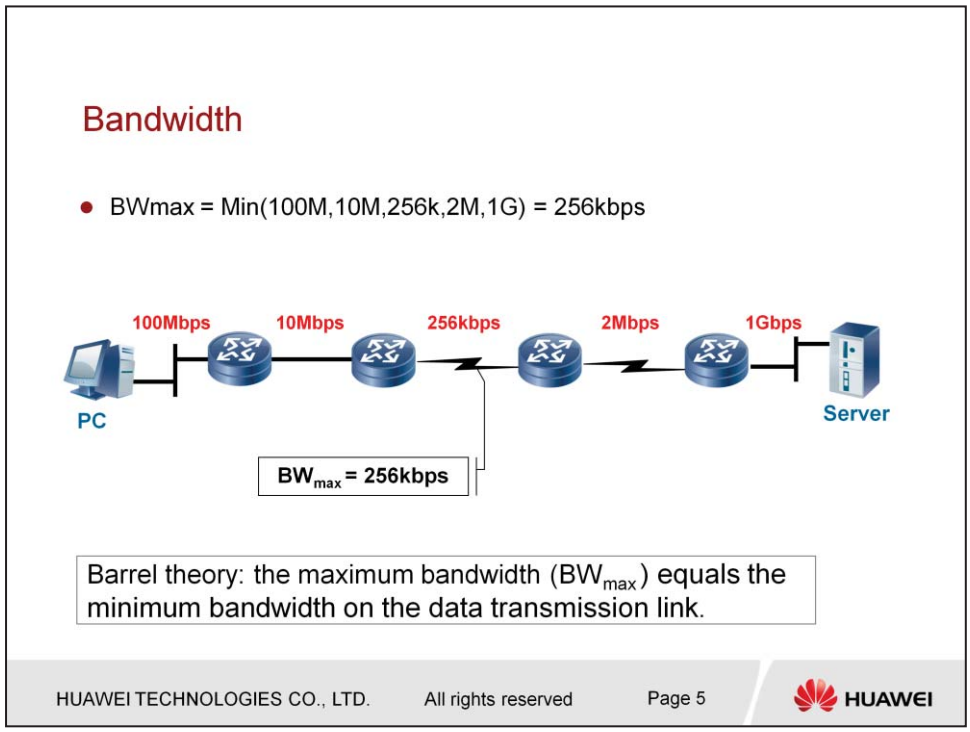
Therefore, to enhance the QoS is to ensure sufficient bandwidth for transmission, reduce the delay and jitter, and lower the packet loss ratio.

In a broad sense, QoS is influenced by various factors in network application. Any positive measure for network applications can improve the QoS. From this aspect, firewall, policy routing, and expedited forwarding are all measures to improve the QoS. However, QoS is assessed for a single network service. Enhancing quality of one service may degrade quality of other services. Network resources are limited, and competition for resources in the network brings about the requirement for QoS. For example, the total bandwidth is 100 Mbps and BT download service occupies 90 Mbps.

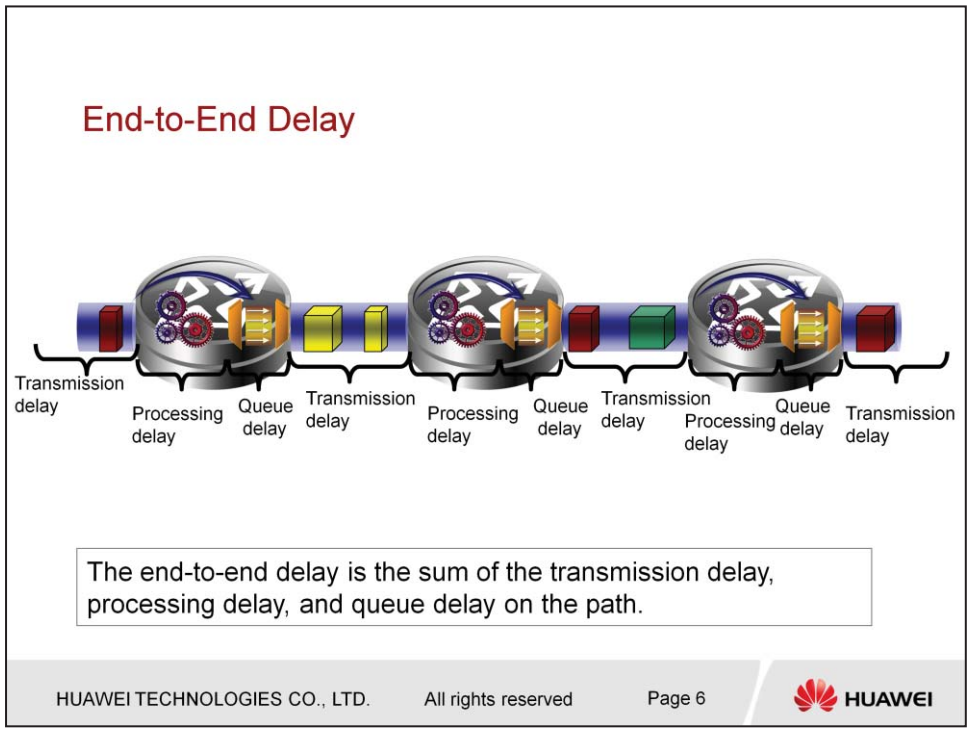
Therefore, only 10 Mbps bandwidth is left for other services.

If the bandwidth for BT download service is limited to 50 Mbps, other services can use at least 50 Mbps bandwidth.

In this way, Quality of other services is improved but quality of the BT service is degraded.

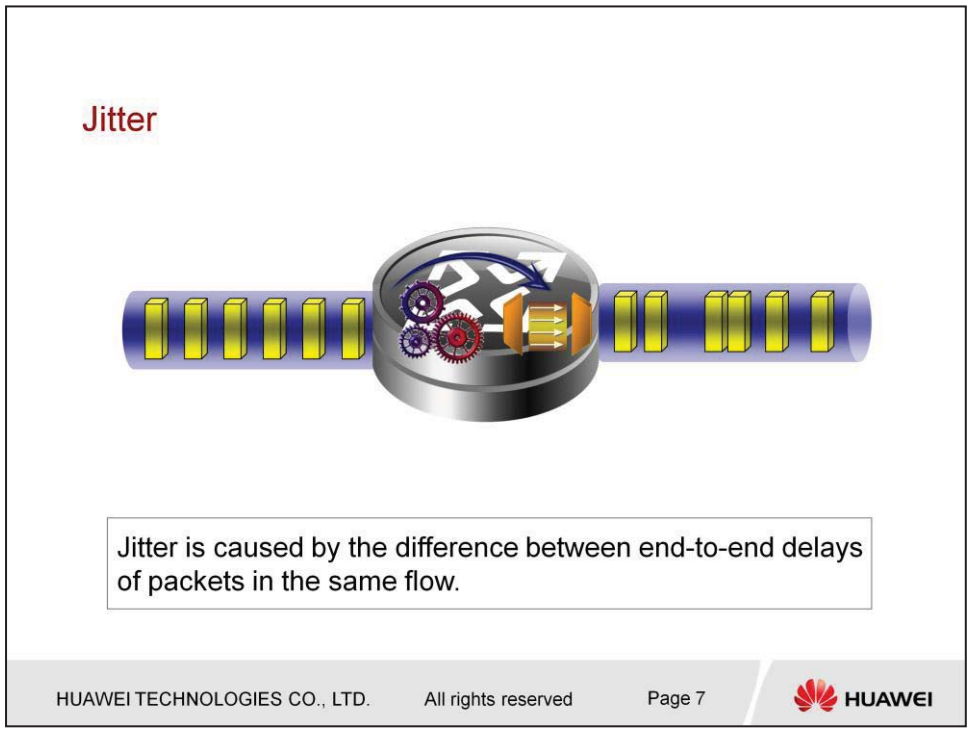


The bandwidth determines the data transmission rate. Theoretically, if the bandwidth is 100 Mbps, it indicates that the data can be transmitted at a rate of 100 Mbit/s. The bandwidth of a transmission path depends on the minimum bandwidth among all links on the path. As shown in the figure, although the maximum bandwidth on the path is 1 Gbps, the maximum transmission rate from the PC to the server is limited to 256 kbps. The reason is that the maximum transmission bandwidth is determined by the minimum bandwidth on this path. Therefore, the minimum bandwidth on a transmission path is the key factor that influences the transmission.



The end-to-end delay consists of the transmission delay, processing delay, and queue delay. The transmission delay depends on the physical feature and distance of the link. The processing delay is the periodduring which the router adds the packets from the incoming interface into the queue on the outgoing interface. The value of the process delay depends on the performance of the router.

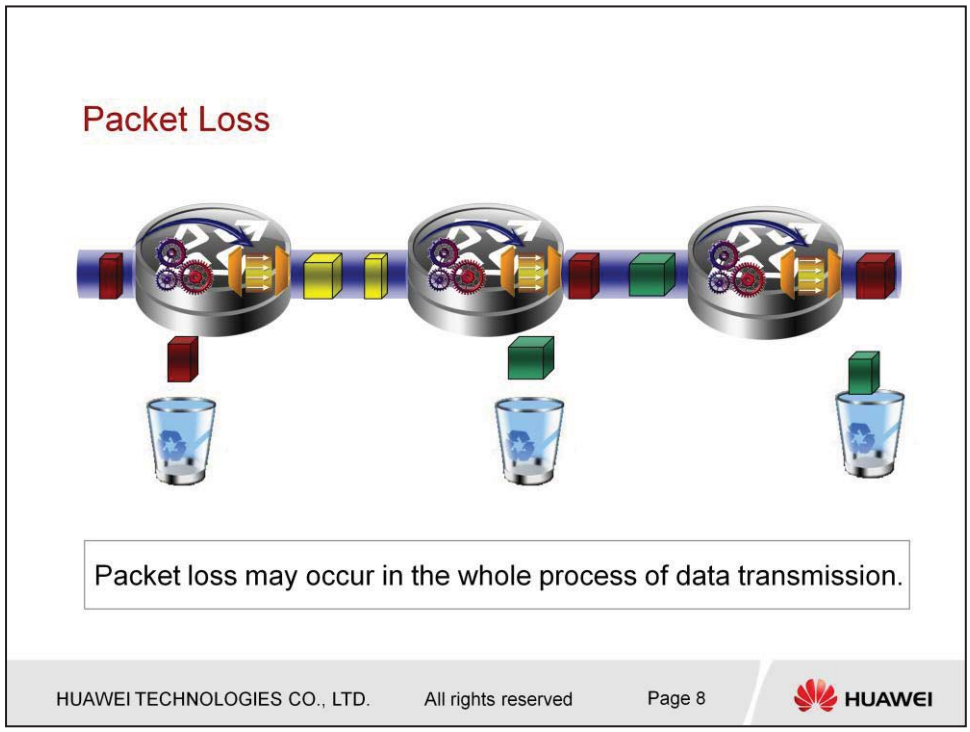
The queue delay is the period during which the packet stays in the queue on the outgoing interface. The value of the queue delay depends on the size and quantity of packets in the queue, the bandwidth, and queuing mechanism.



Jitter is caused by the difference between end-to-end delays of packets in the same flow.

As shown in the figure, the source end sends packets at equal intervals. The packets are transmitted with different end-to-end delays and they arrive at the destination end at unequal intervals, and thus jitter occurs. The jitter range is determined by the delay.

Shorter delays cause small jitter range.

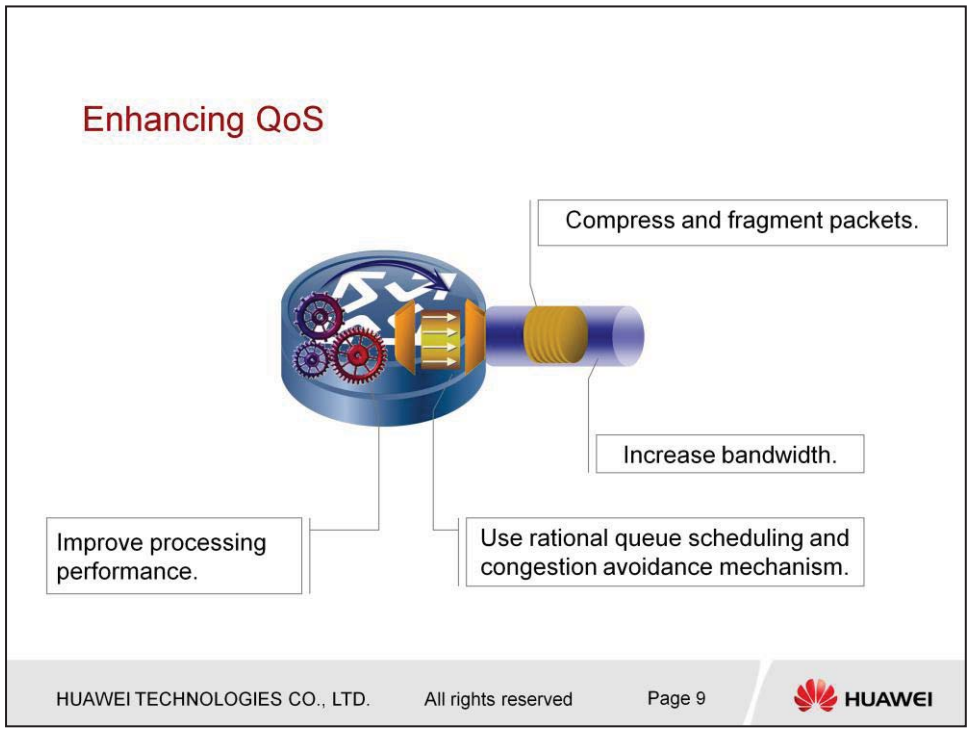


Packet loss may occur in the whole process of data transmission.

For example:

- When the router receives the packets, the CPU is busy and cannot process the packets. Packet loss occurs.
- In queue scheduling, if the queue is full, packet loss will occur.
- If link fails or collision occurs during data transmission, packet loss may occur.

In most cases, packet loss is caused by a full queue. When the queue is full, the packets that arrive subsequently are dropped.

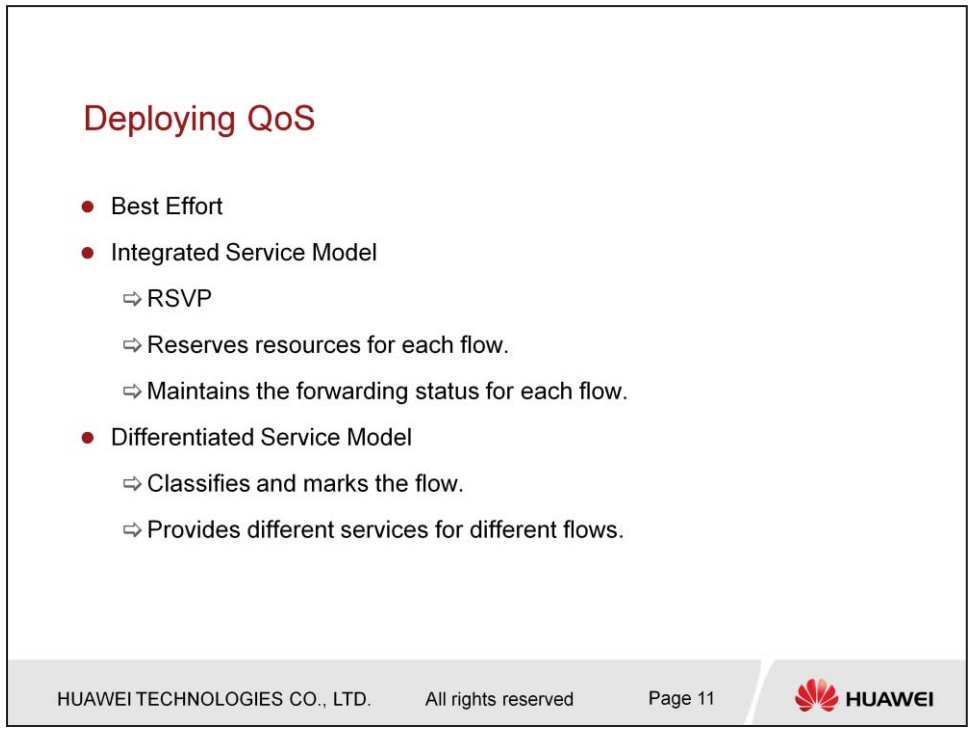


The network QoS can be enhanced with the following methods:
1. Increase the link bandwidth.
QoS of the network will be obviously enhanced when the link bandwidth increases. The available bandwidth increases with the link bandwidth, which ensures higher
traffic. Increase in link bandwidth also reduces the transmission delay and jitter. In addition, when the link bandwidth increase, the packet loss ratio is lowered, so
less packets are dropped. In terms of effect, increasing the link bandwidth is the best method of enhancing the QoS. However, it is impractical, because bandwidth increase requires more investment and time and may interrupt the network service.
2. Use rational queue scheduling and congestion avoidance mechanism.
The queue scheduling mechanism has the following advantages:
1) Data of various services are scheduled to different queues,
2) and thus the network bandwidth can be allocated more
3) rationally. This ensures sufficient bandwidth for the data that r
4) equires high bandwidth and avoids bandwidth waste.
2) The delay-sensitive data are added to the queue with higher priority so that the data can obtain the service with low delay.

3) Through congestion avoidance mechanism, packets are dropped randomly at a certain proportion according to the significance. This avoids congestion. A router supports various queuing mechanisms, such as custom queue and priority queue. You should configure proper queue according to the service requirement.

3. Improve the link efficiency by the link efficiency mechanism. Within certain bandwidth of a physical link, compressing the data can logically increase the available bandwidth of the link. You can choose to compress the header or payload of the data, but this may increase the CPU load and transmission delay. In addition, large packets can be fragmentized. Delay sensitive data can be transmitted between the fragments, and thus data delay sensitive data obtains the service with low delay.

4. Improve the processing performance of the equipment. To improve the processing performance of the equipment, you can improve the capability of the CPU, use the chip with higher performance, increase the memory, or adopt better implementation structure. A higher processing performance reduces the delay and packet loss but also increases the cost.



Best-effort service model

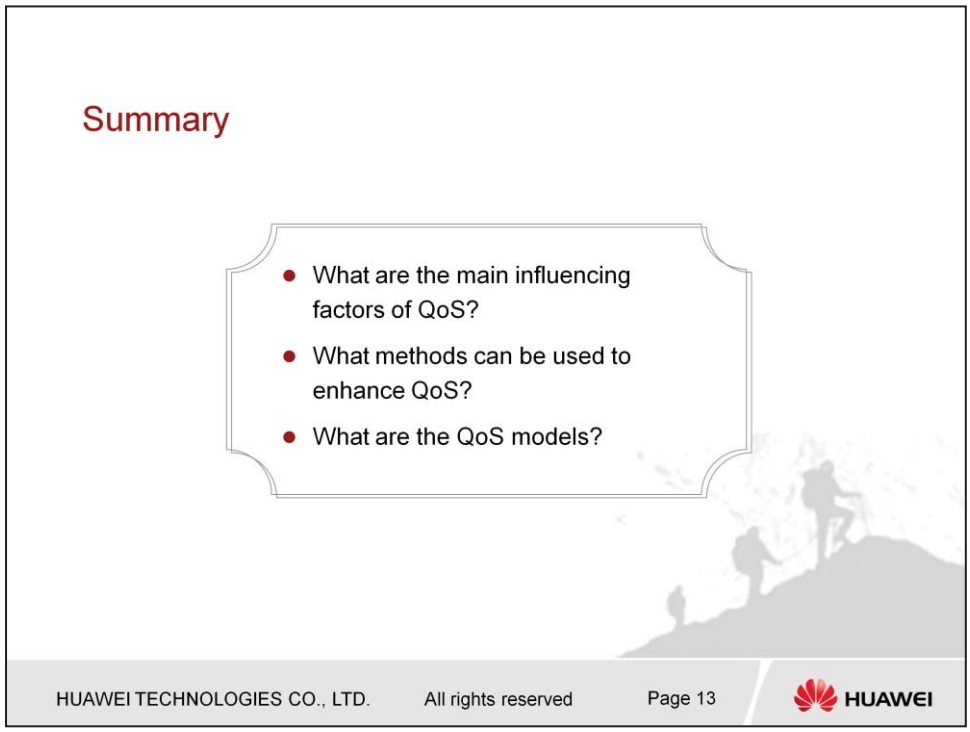
Data communication devices such as routers and switches select transmission path for each packet individually through the TCP/IP stack. This process uses statistical multiplexing, which does not involve dedicated connection, as time division multiplexing (TDM). The traditional IP network provides only one service type, namely the best-effort service. In this service type, all packets transmitted on the network have the same priority. Best effort means that the IP network transmits the packets to the destination as complete as possible, but it cannot avoid dropping, damage, repetition, disorder, or wrong transmission of packets. Besides, the IP network does not ensure features (such as delay and jitter) related to the transmission quality. Actually, the best-effort service does not belong to the QoS technology. However, the best-effort service model is the main service model used on the Internet, so it is necessary to know about. The phrase “best effort” does not have any derogatory sense. The best-effort service type enables the development of the Internet. However, with the development of the Internet, the best-effort service cannot meet the requirement for wide application of the Internet nowadays. ISPs need to provide more service types to meet the specific requirements.

Integrated services (IntServ) model

The IntServ model is developed by the IETF in 1993. It supports multiple service types in the IP network. The objective of the IntServ model is to support both the real-time service and traditional best-effort service. This model is based on the mechanism that reserves resources for each flow.

In the IntServ model, the source, destination hosts and all the nodes along the path exchange RSVP signaling messages to establish the forwarding status on every node along the transmission path between the source and destination hosts. The forwarding status must be maintained for each flow, so the expansibility of the IntServ model is poor. In addition, maintaining the status of millions of flows on the Internet consumes too many resources of the device. Therefore, the IntServ model has not really come into use. In recent years, RSVP is modified and can be used with the differentiated service model. Development of the MPLS VPN technology also promotes the development of RSVP. However, the IntServ model is still not widely used in QoS technology. Differentiated service (DiffServ) model to solve the problem of poor expansibility of the IntServ model, in 1995, ISPs and institutes began to develop a new mechanism that supports multiple service types and has good expansibility. In 1997, the IETF realized that the network service model used at that time could not satisfy the network operation. The IETF hoped to develop a more expansible model to classify information flows more easily and to differentiate specific users and applications. The IETF established a work group and developed the DiffServ model. This model is simpler and classifies information flows on the Internet.

The DiffServ model supports multiple types of applications in commercial mode.



Q: What are main influencing factors of QoS?

A: Bandwidth, delay and jitter, packet loss

Q: What methods can be used to enhance QoS?

A: 1. Increase the link bandwidth.

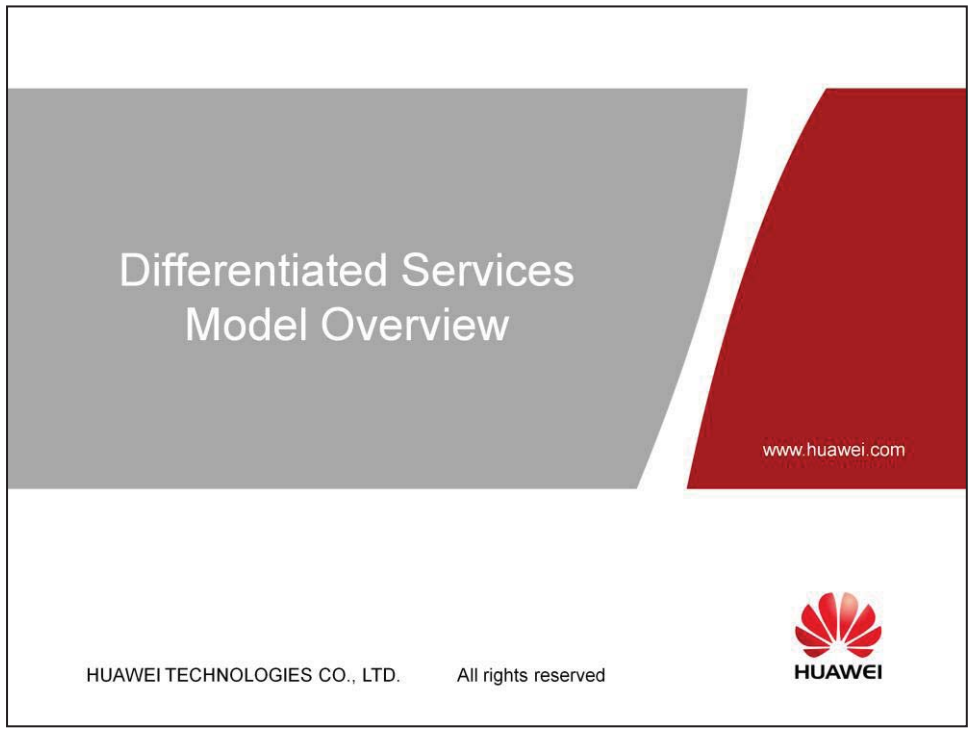
2. Use rational queue scheduling and congestion avoidance mechanism.

3. Adopt the link efficiency mechanism.

4. Improve the performance of the equipment.

Q: What are the QoS models?

A: Best-effort service model, integrated service model, differentiated service model


Differentiated Services
Model Overview
www.huawei.com
HUAWEI TECHNOLOGIES CO., LTD.
All rights reserved
HUAWEI



Foreword

This section introduces the characteristics, marking and Per Hop Behavior (PHB) of the Differentiated Services Model.

HUAWEI TECHNOLOGIES CO., LTD. All rights reserved





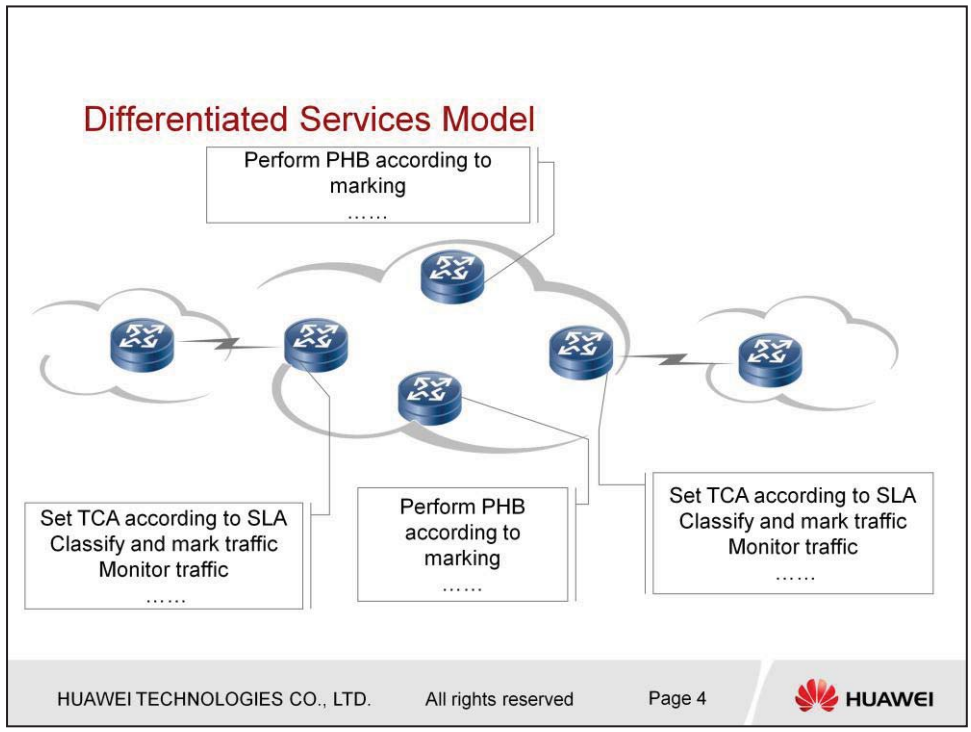
Objectives

Upon completion of this section, you will be able to:

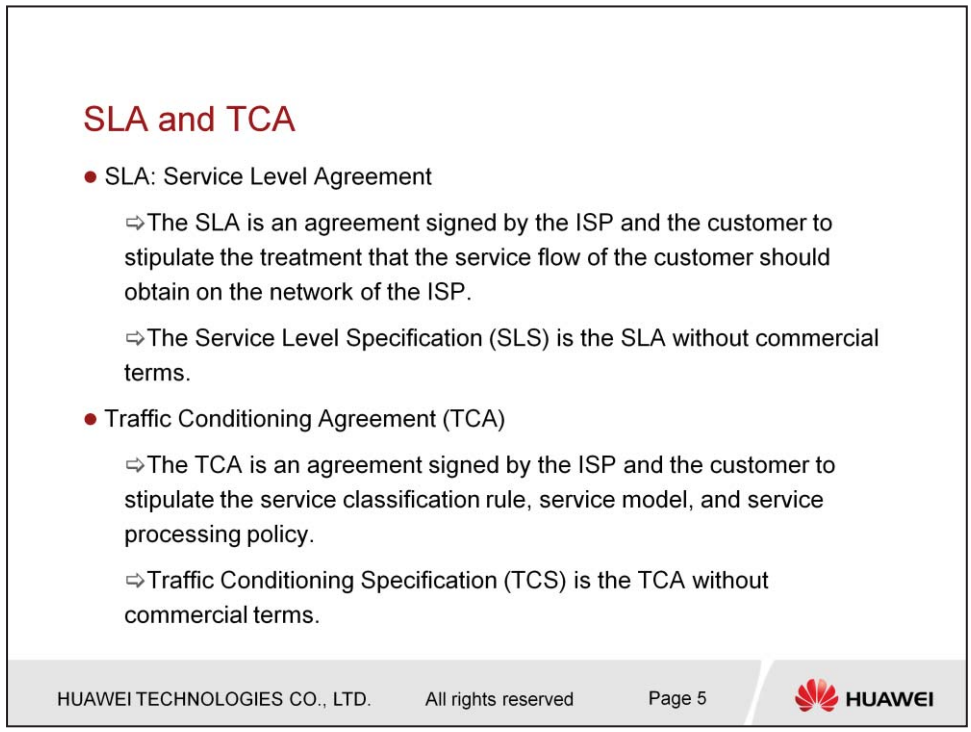
- Describe the basic principles of the differentiated services model
- Describe the marking schemes of the Type of Service field
- Describe the differences and features between various PHB

HUAWEI TECHNOLOGIES CO., LTD. All rights reserved





In the differentiated service (DiffServ) model, services are described by the traffic classifier. The flows are classified and marked on the ingress router in the DiffServ domain. The internal routers perform corresponding PHB according to the classification marking of the packets and need not perform complex traffic classification. PHB stands for per-hop behavior. It is the action performed to the traffic by a router, for example, expedited forwarding, re-marking, and dropping of packets. The traffic classification marking is contained in the packet header and transmitted in the network with the data. Therefore, the router need not maintain the status information for the flows. (In integrated service model, the router must maintain the status information for each flow.) The service that a packet can obtain is related to the marking of the packet. The ingress router and egress router of a DiffServ (DS) domain are connected to other DS domains or non-DS domains through links. Different administrative domains may apply different QoS policies, so the administrative domains must negotiate the Service Level Agreement (SLA) and establish the Traffic Conditioning Agreement (TCA). The inbound traffic to the ingress router and the outbound traffic to the egress router must comply with theTCA.



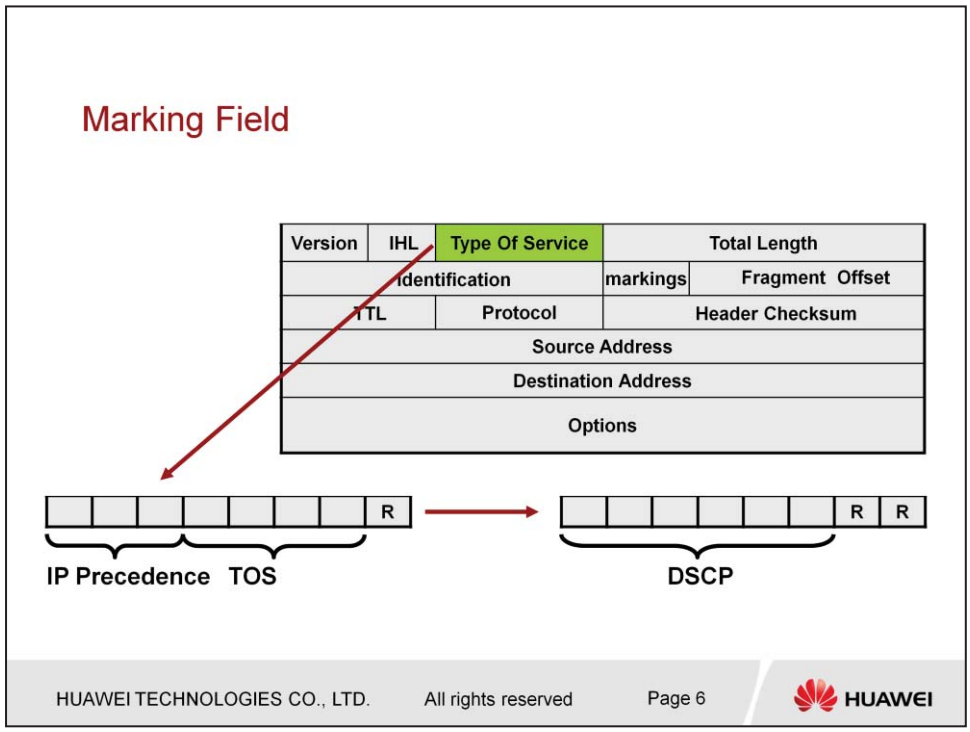
- Service Level Agreement (SLA)

The SLA is an agreement signed by the ISP and the customer to stipulate the treatment that the service flow of the customer should obtain on the network of the ISP.

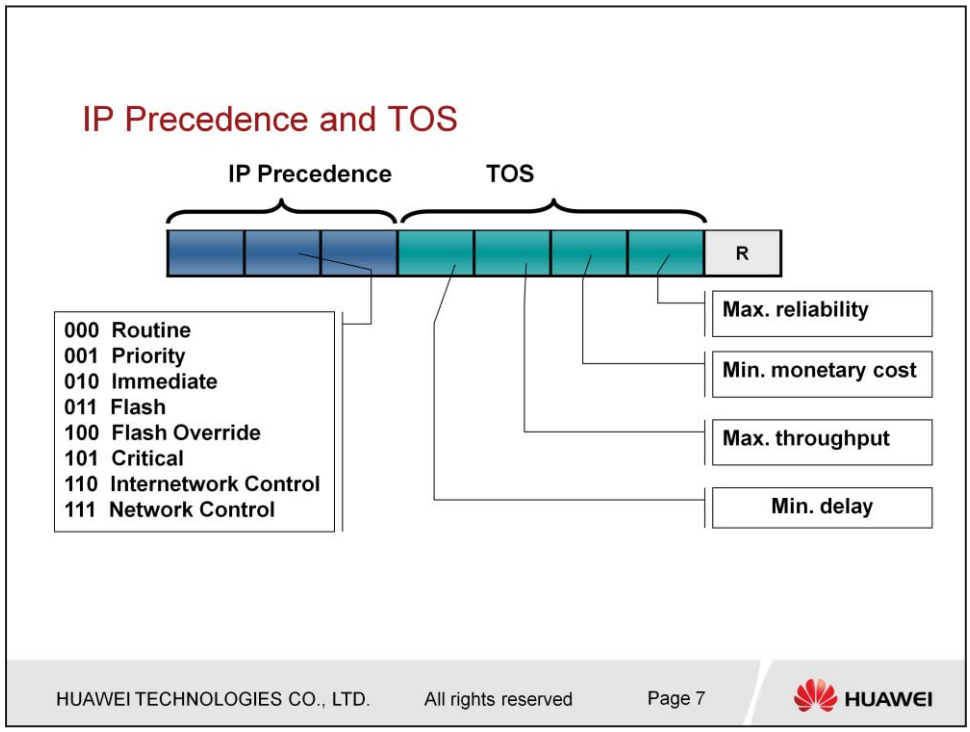
The SLA contains some commercial information, and the technical specifications can be described in the Service Level Specification (SLS). In many documents, SLA is used to specify the certain QoS. SLS is the SLA without commercial terms.

- Traffic Conditioning Agreement (TCA)

The TCA is an agreement signed by the ISP and the customer to stipulate the service classification rule, service model, and service processing policy. The technical specifications in the TCA can be described in the Traffic Conditioning Specification (TCS). The SLA can include the TCA. The SLA or SLS stipulates common requirements for service processing, such as the service processing mechanism. The TCA or TCS stipulates specific requirements, such as the bandwidth. TCS is the TCA without commercial terms.



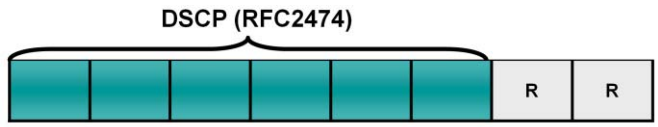
In the basic model of priority-based service classification, services are classified based on their priorities. The priority is contained in the in a certain field of the packet header. The network node determines the forwarding policy according to the priority in the packet header. Currently, several standards for priority-based classification are established. RFC 791 defines the IP Precedence field and classifies IP applications into eight categories. RFC 2474 redefines the TOS field and classifies services into 64 categories.



According to the characteristics of the IP applications, RFC 791 (Internet Protocol) classifies services into eight categories: Network Control, Internetwork Control, CRITIC/ECP, Flash Override, Flash, Immediate, Priority, and Routine, mapping eight priority levels. The Routine service has the lowest priority and the Network Control service has the highest priority.

RFC 1349 (Type of Service in the Internet Protocol Suite) defines 16 priority levels according the TOS. The TOS field occupies four bits, representing minimize delay, maximize throughput, minimize monetary cost, and maximize reliability respectively. RFC 1349 also provides the recommended TOS value for various IP applications. For example, the recommended TOS value for the FTP control packet is minimize delay.

DSCP: DiffServ Code Point

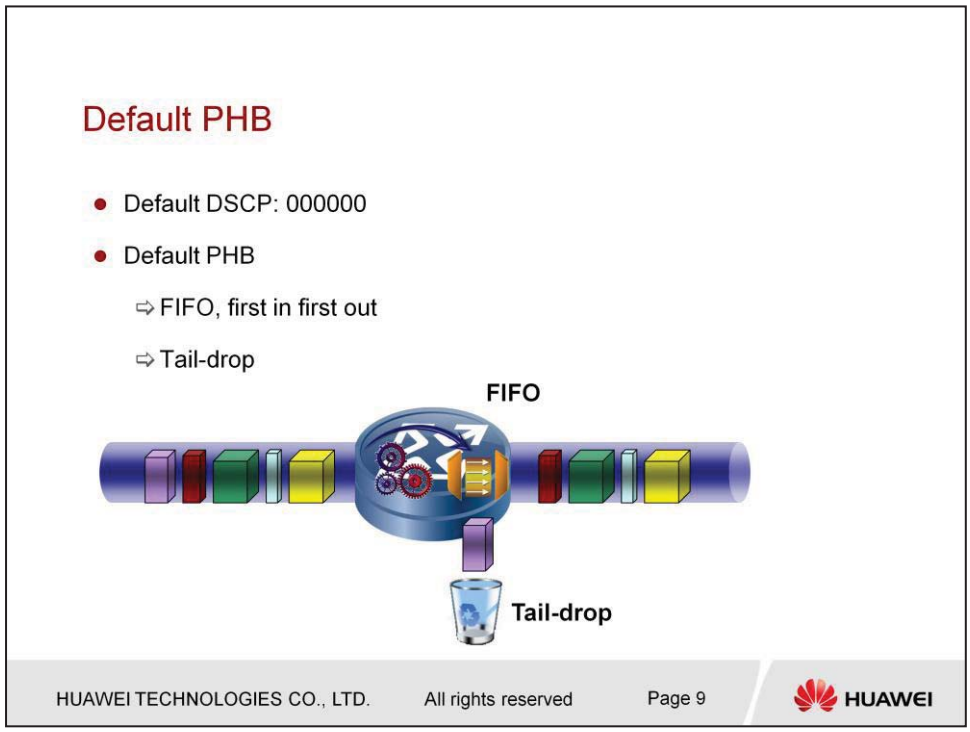


- Default PHB
- Class Selector (IP Precedence) PHB
- Expedited Forwarding (EF) PHB
- Assured Forwarding (AF) PHB

●Each DSCP value maps a BA.
●Each BA is assigned a PHB.
●Each PHB is implemented by a group of QoS mechanisms.

HUAWEI TECHNOLOGIES CO., LTD. All rights reserved

RFC 2474, Definition of the Differentiated Services Field (DS Field) in the IPv4 and IPv6 Header, redefines the TOS field. The first six bits (high-order bits) identifies the service type. The s ubsequent two bits (low-order bits) are reserved. Based on this definition, the service traffic can be classified into 64 ategories through DSCP. Each DSCP value maps to a Behavior Aggregate (BA). Each BA is assigned a PHB (such as forwarding, dropping, etc.). The PHB is implemented by some QoS mechanisms, such as traffic policing and queuing mechanism. The DiffServ model defines four types of PHB: EF PHB, AF PHB , CS PHB, and BE PHB. Expedited Forwarding (EF) PHB is applicable to preferential services with low delay, low packet loss, and guaranteed bandwidth. Assured Forwarding (AF) PHB consists of four classes, and each class has three drop precedence levels. Therefore, the AF PHB can subdivide services. Its QoS performance is lower than the EF PHB. The class selector (CS) PHB is derived from the TOS field. It consists of eight classes. The BE PHB (default PHB) is a special class of CS. The traffic of this class is not guaranteed anything. The traffic on the current IP network belongs to this class by default.



The default DSCP value is 0, which is compatible with the default value of the IP precedence 0. DSCP 0 maps the default PHB. The default PHB processes the traffic by the principle of first in first out (FIFO) and tail drop.

CS PHB

- Ensures the compatibility with the nodes not supporting DSCP.
- The three high-order bit maps the IP precedence value.

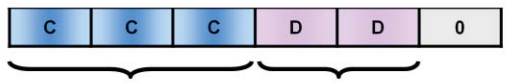
CS Class	DSCP Value	Precedence Value	Precedence Class
Default	000000	0	Routine
CS1	001000	1	Priority
CS2	010000	2	Immediate
CS3	011000	3	Flash
CS4	100000	4	Override Flash
CS5	101000	5	Critical
CS6	110000	6	Internetwork Control
CS7	111000	7	Network Control

HUAWEI TECHNOLOGIES CO., LTD. All rights reserved

DiffServ defines the Class Selector PHB (CS PHB) and the mapping DSCP value to ensure the compatibility with the IP precedence. The first three bits map to the IP precedence value. If a router supports only the IP precedence, it concerns only the first three bits of the DSCP marking when it receives a packet. Same as the IP precedence value, a larger DSCP value maps a higher priority. The last three bits of all the DSCP values in the tables are 000. But for a router that does not support DSCP, even if these bits are not 000, the meaning is the same. For example, 010000 and 010011 has the same meaning. Therefore, eight DSCP values maybe mapped to one IP precedence value.

AF PHB

- Guarantees certain bandwidth.
- Permits occupation of idle bandwidth.



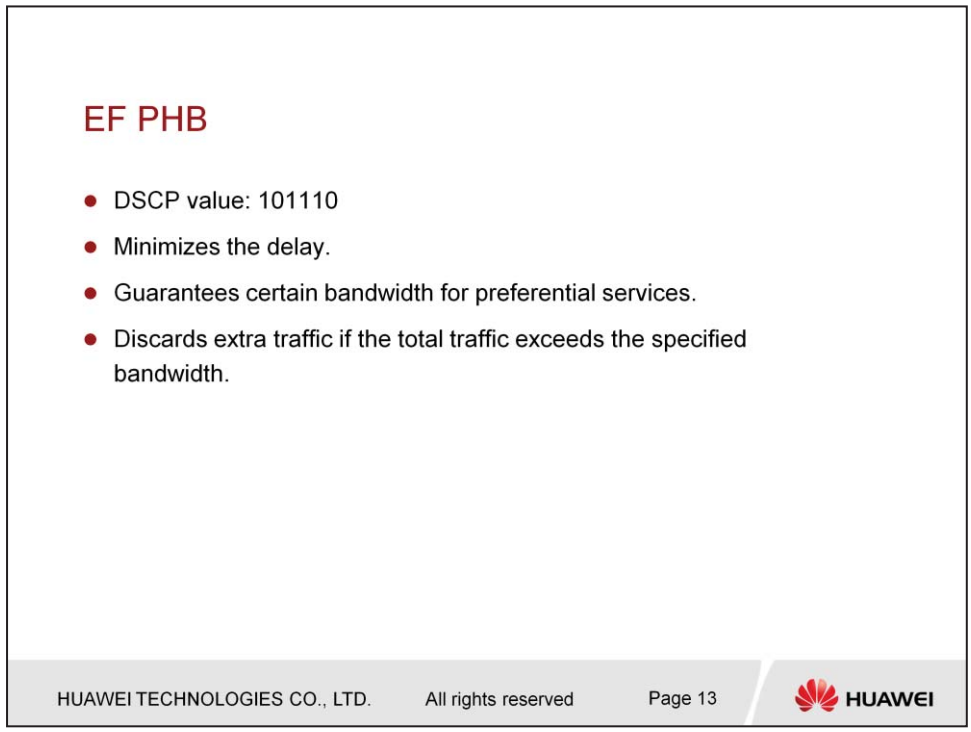
Class Drop probability

	Class1	Class2	Class3	Class4
Low drop precedence	001010	010010	011010	100010
Medium drop precedence	001100	010100	011100	100100
High drop precedence	001110	010110	011110	100110

HUAWEI TECHNOLOGIES CO., LTD. All rights reserved

AF PHB (the assured forwarding per-hop behaviors) is defined in RFC 2597. RFC 2597 defines 12 DSCP values, which are classified into four classes (based on the first three bits): Class1, Class2, Class3, and Class4. each class has three drop precedence levels (classified based on the fourth and fifth bits): low drop precedence, medium drop precedence, and high drop precedence. The data marking with DSCP AF are provided with certain guaranteed bandwidth. If idle bandwidth exists, the data can occupy the bandwidth. AF PHB is implemented through the queue scheduling and congestion avoidance mechanisms. Each class corresponds to a queue, which provides certain guaranteed bandwidth for the traffic of this class. The idle bandwidth of a class can be used by traffic of other classes. Note that the classes are treated at the same precedence. For example, Class2 cannot obtain more guarantee than Class1. The four classes are equal in priority. With a queue, the congestion avoidance mechanism (such as WRED) is adopted. This mechanism sets two thresholds. When the number of packets in the queue is less than the lower threshold, no packets are dropped. When the number of packets is between the lower threshold and higher threshold, packets are dropped at certain probability. The probability increases with the increase of packets.

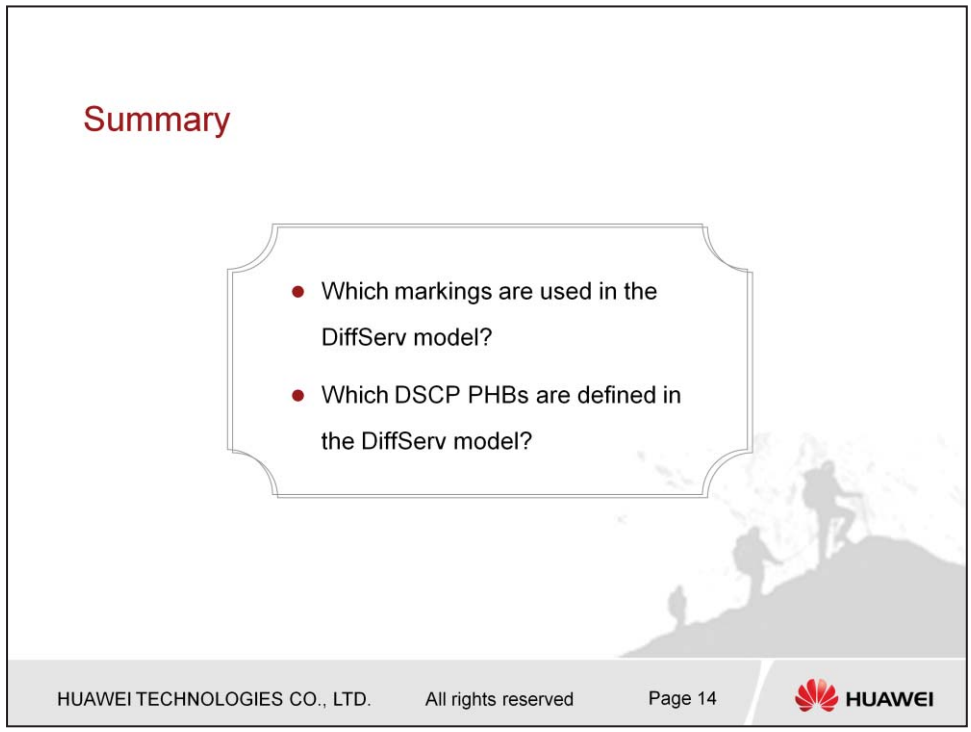
When the number of packets exceeds the higher threshold, the drop probability is 100%. AF PHB is generally implemented through the Class-based Queuing (CBQ) technology. In CBQ, four queues are defined to map four classes. Weighted Random Early Detection (WRED) is configured for each queue. CBQ and WRED mechanism will be described in the later sections.



EF PHB maps the DSCP value 101110. For a device that does not support DSCP, EF PHB is equivalent to IP precedence 5. The delay-sensitive data is tagged “101110”. This types of data should be forwarded as soon as possible and should obtain certain guaranteed bandwidth. To prevent the data from consuming all bandwidth, the router drops the extra packets when the traffic exceeds the guaranteed bandwidth. Two mechanisms must be defined to implement EF PHB. Firstly, a queue scheduling mechanism is required to ensure fastest scheduling of EF packets. Thus the EF packets are ensured with lowest delay and jitter. This mechanism can be implemented through strict priority queue, IP RTP queue, or LLQ queue. These queue scheduling mechanisms will be described in later sections.

Secondly, a traffic policing policy is required to specify certain bandwidth for the EF traffic. Within the specified bandwidth, the EF traffic can obtain the service with low delay.

However, if the traffic exceeds the bandwidth, the extra traffic is dropped.



Q: What are markings of the DiffServ model?

A: IP Precedence, ToS, and DSCP

Q: What DSCP PHBs are defined in the DiffServ model?

A:1.BE PHB; 2. CS PHB; 3.EF PHB; 4.AF PHB.


IP QoS Technologies
www.huawei.com
HUAWEI TECHNOLOGIES CO., LTD.
All rights reserved
HUAWEI



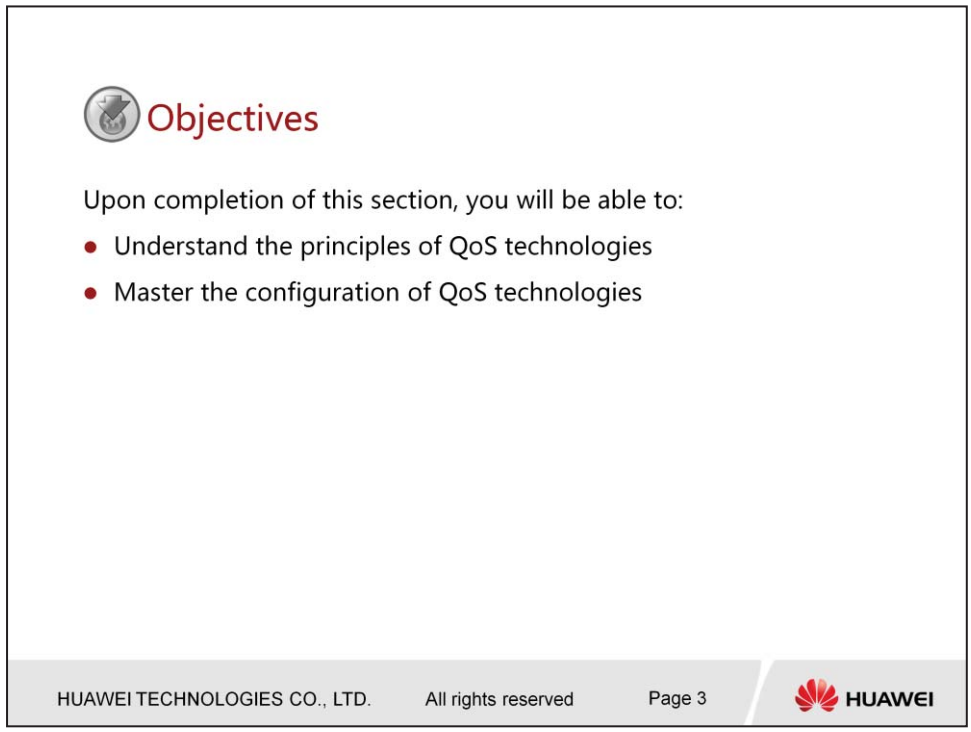
Foreword

Deployment of QoS in the IP network relies on the use of various forms of QoS technology, which includes: classification and marking, traffic policing and shaping, congestion management technology, congestion avoidance and link efficiency mechanisms.

This section introduces the principles and configuration of these technologies.

HUAWEI TECHNOLOGIES CO., LTD. All rights reserved




Objectives
Upon completion of this section, you will be able to:
Understand the principles of QoS technologies
Master the configuration of QoS technologies
HUAWEI TECHNOLOGIES CO., LTD.
All rights reserved
Page 3
HUAWEI



Contents

Classification and Marking

Policing and Shaping

Congestion Management

Congestion Avoidance Mechanisms

Link Efficiency Mechanisms

HUAWEI TECHNOLOGIES CO., LTD. All rights reserved



Traffic Classification and Marking

- Traffic classification and marking is the basis for QoS deployment.
- The traffic can be classified based on the ACL and packet length.
- The traffic can be marked with DSCP or IP Precedence (coloring)



HUAWEI TECHNOLOGIES CO., LTD. All rights reserved

Traffic classification is to classify the traffic into multiple precedence levels and service classes. If packets are marked by the first three bits (IP precedence) of the ToS field in the packet header, IP packets can

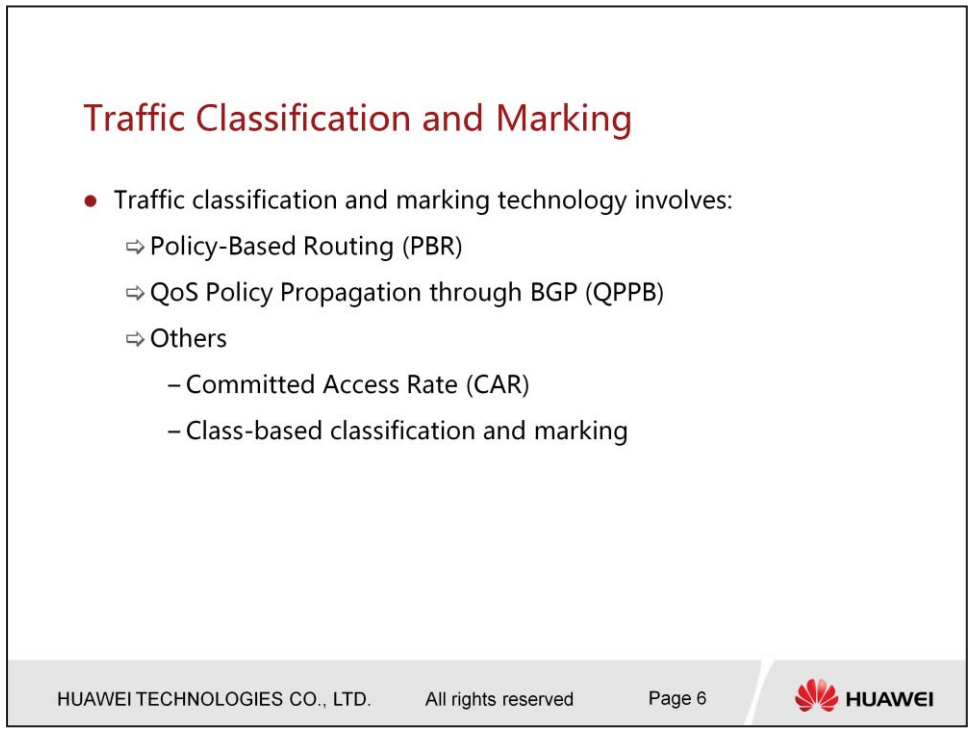
be classified into up to 2 to the power of 3, or 8 classes. If packets are marked by DSCP, which is the first six bits in the ToS field, IP packets can be classified into up to 2 to the power of 6, or 64 classes. After classification of packets, QoS features can be applied to different classes to implement class based congestion management and traffic shaping. The traffic can be classified according to almost all information contain in the packet, such as the source IP address, destination IP address, source port number, destination port number, and protocol ID. Although traffic classification can be performed according to almost all information in the packet, in most cases, the traffic is marked by the ToS field of the IP packet. Through traffic marking, the application system or device that processes the packets obtains the class of the packets and processes the packets according to the pre-defined policy (PHB). For example, the following classification and marking policy is defined on the network edge:

•All VoIP data packets belong to the EF service class. The IP precedence for these packets is 5 and the DSCP flag is EF.

•All VoIP control packets belong to the AF service class. The IP precedence for these packets is 4 and the DSCP flag is AF31.

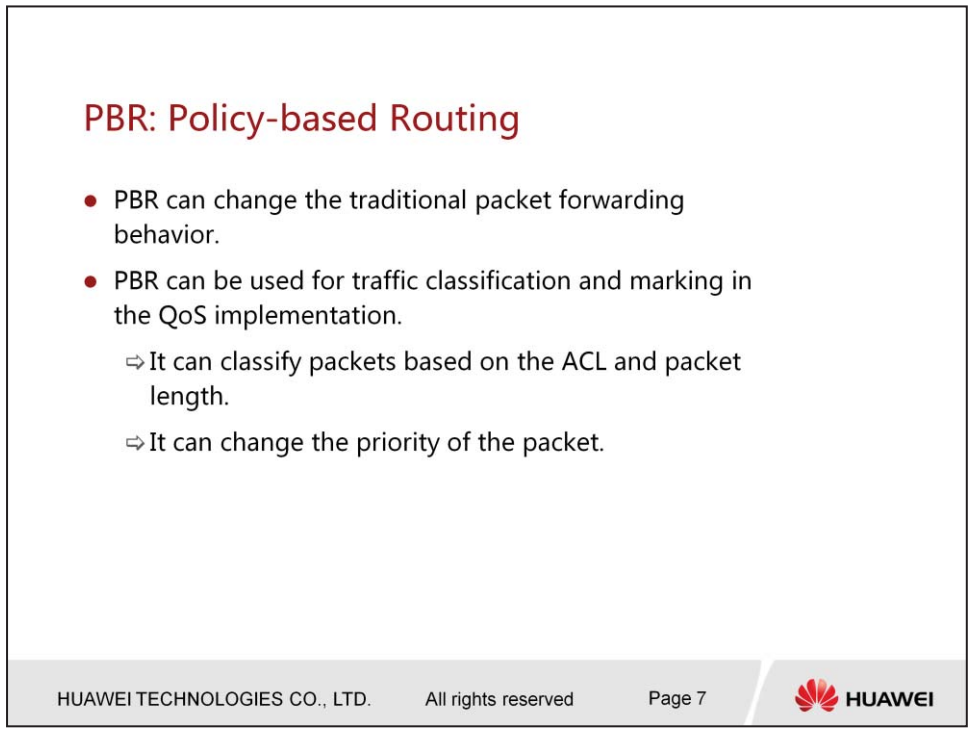
When packets are classified and marked on the network edge, the medium nodes in the network can provide differentiated services for various classes of traffic

according to DSCP flags. In the above example, the medium node ensures low delay and jitter for services of EF class and performs traffic policing. When congestion occurs, the medium node guarantee certain bandwidth for services of AF class.

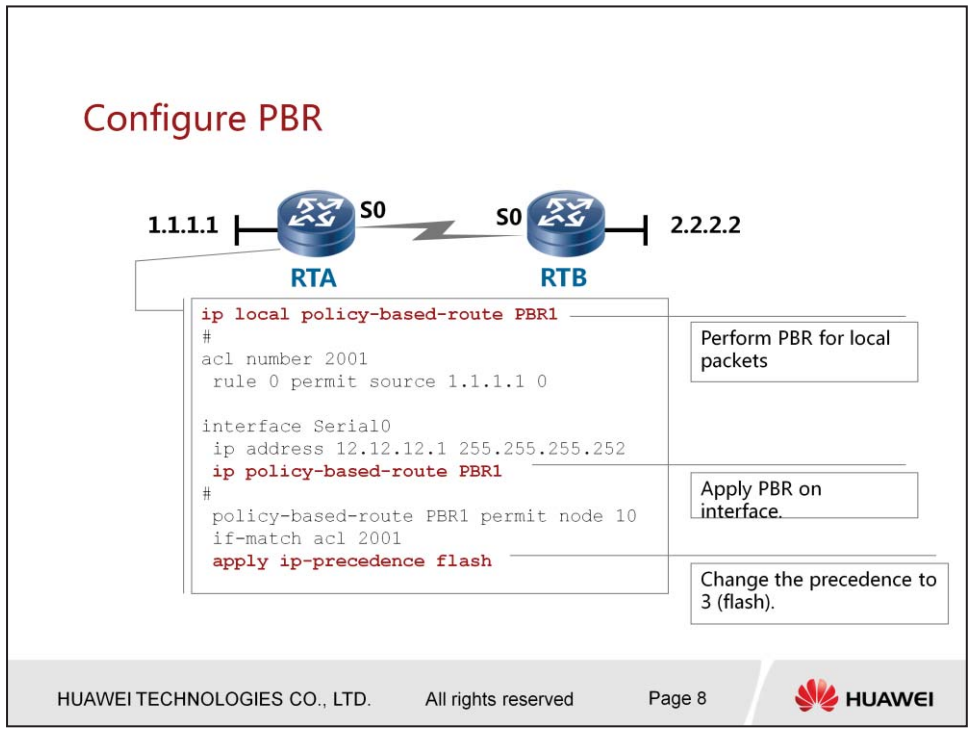


PBR changes the traditional forwarding behavior based on the destination address. PBR defines some "if-match" and "Apply" statements. The "if-match"

statement defines the match rule. The "Apply" statement defines the behavior that should be performed after matching. The behavior may be changing the next hop for forwarding or changing the marking field of the packet. QPPB is a mechanism for transferring the QoS policy by the BGP attribute. PBR and QPPB only classify and mark the traffic. Other traffic classification and marking technologies such as CAR and class-based classification and marking can also implement other QoS mechanisms. These technologies will be discussed later in this section.



Traditionally, a router calculates out the routing table by using the routing protocol and forwards the packets according to the destination IP address. Policy-based routing (PBR) has higher routing control capability and is more flexible. Based on PBR, a network administrator can select the forwarding path according to the destination address, routing protocol, packet length, application type, source IP address or other policies. The policy can be defined according to the application. A policy can be used to control load balancing among routers, guarantee QoS for packet forwarding on a single link, or satisfied some specific requirements. When deploying QoS, you can use PBR to classify and mark the traffic. The router classifies the packets based on the ACL rules (source IP address, destination IP address, protocol type, port number, and so on). After classification, the router changes the DSCP or IP Precedence field to mark the packets.



In the configuration example, PBR is used to change the IP precedence of packets to 3 for packets from 1.1.1.1.

Define the match rule for PBR.

1. Enter the system view: system-view

2. Create a policy or a policy node: policy-based-route policy-name { deny | permit } node node-id

3. Set the rule for matching the IP packet length: if-match packet-length minimum-length maximum-length

4. Set the rule for matching the IP address: if-match acl acl-number

A policy specified by the policy name can contain several policy nodes. The policy node is specified by the node ID. A policy node with a smaller node ID maps a higher priority, that is, the policy of this node is performed earlier. A policy can be used to import routers and forward IP packets through PBR. When you run the **policy-based-route** command with the same ID repeatedly, the new configuration overwrites the previous one. The contents of the policy are specified by the if-match and apply statements. **permit** indicates to perform PBR for packets that match the rule; **deny** indicates not to perform PBR for packets that match rule. The IP unicast PBR provides two methods to match the packets: based on the packet length and based on the ACL rules. A policy may contain multiple if-match statements, which are used in a combination. Define the PBR behavior. 1. Enter the system view: system-view

2. Create a policy or a policy node: policy-based-route policy-name { deny | permit } node node-id

3. Set the precedence of packets: apply ip-precedence precedence

4. Specify the default next hop for the packets: apply ip-address default next-hop ip-address1 [ip-address2]

5. Specify the default outgoing interface for the packets: apply default output interface interface-type1 interface-number1 [interface-type2 interface-number2]

6. Set the next hop of the packets: apply ip-address next-hop ip-address

7. Specify the outgoing interface for the packets: apply output-interface interfacetype1 interface-number1 [interface-type2 interface-number2]

8. Configure the access VPN instance: apply access-vpn vpn-instance vpninstance-name&<1-6>

The apply statement is used to set the packets meeting the match rule so that the packets are forwarded along the specified route. A policy may contain multiple apply statements, which are used in a combination. You can specify multiple next hops or configure multiple outgoing interfaces. In this case, packets are forwarded through load balancing. The traffic can be shared by only next hops or by only outgoing interfaces. If both outgoing interfaces and next hops are configured, the traffic is shared by only the outgoing interfaces.

Note that the outgoing interface cannot be a broadcast interface such as an Ethernet interface.

Apply the route policy. Enable the local PBR.

1. Enter the system view: system-view

2. Enable the local PBR: ip local policy-based-route policy-name

The local PBR is valid for only the packets generated on the local device. You can configure only one local route policy.

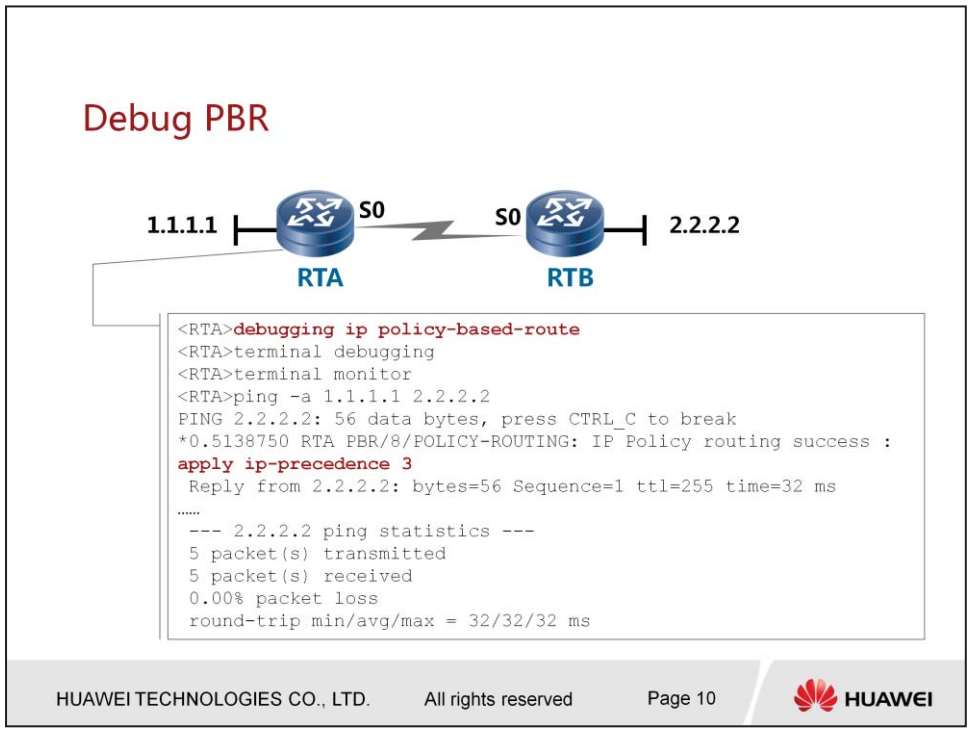
Enable the interface PBR.

1. Enter the system view: system-view

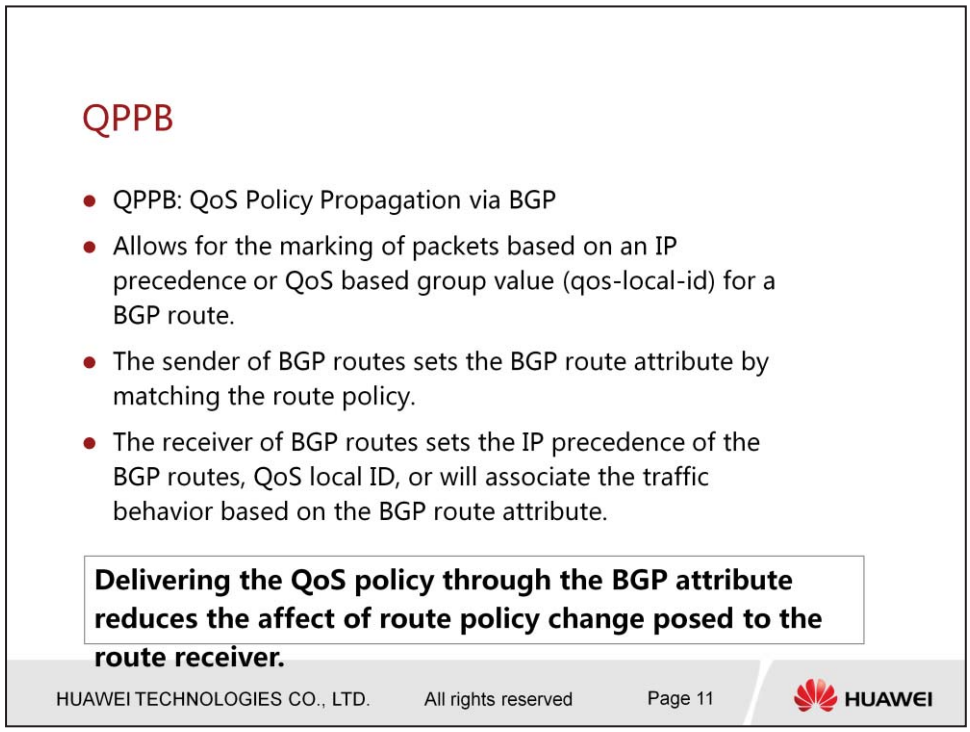
2. Enter the interface view: interface interface-type interface-number

3. Enable the interface PBR: ip policy-based-route policy-name

The interface PBR is valid for the packet received on this interface. Only one route policy can be configured on an interface.



Using the **debugging ip policy-based-route** command, you can view the implementation of the PBR. From the output information in the example, you can see that the IP precedence of the packets has been changed to 3 (flash).



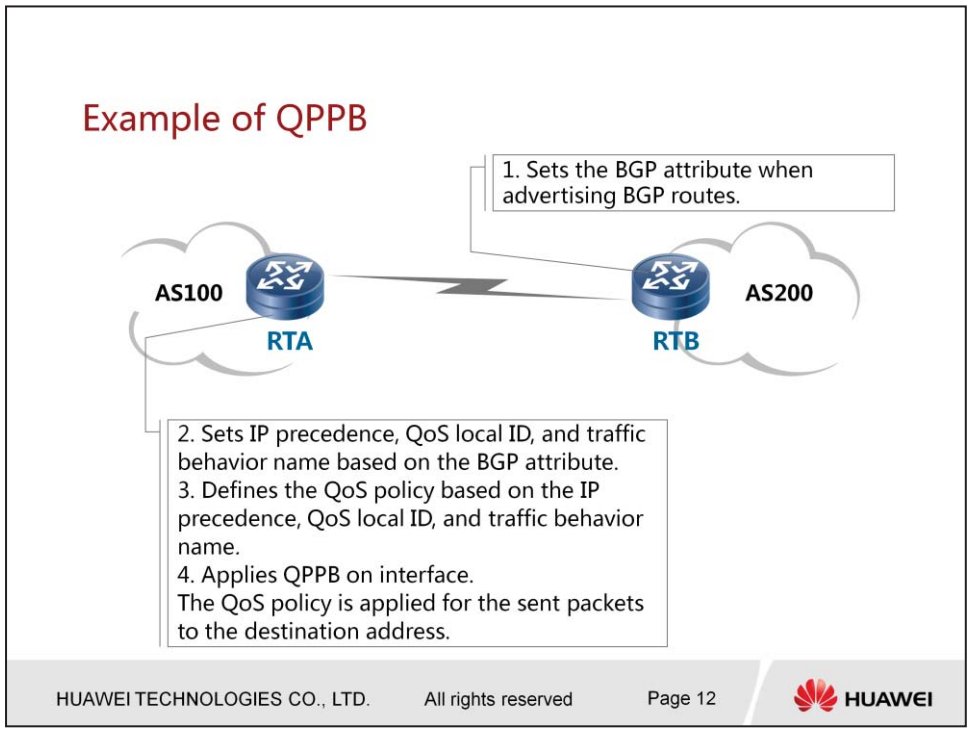
QPPB stands for QoS Policy Propagation Through the Border Gateway Protocol. The sender of BGP routes set the BGP route attribute by matching the route policy.

The receiver of BGP routes sets the IP precedence, QoS local ID, and traffic behavior name based on the BGP community list, cost, BGP AS paths list, ACL, or prefix list .

The IP precedence, QoS local ID, and trafficbehavior name are delivered to the FIB with the routing information. In packet forwarding, the

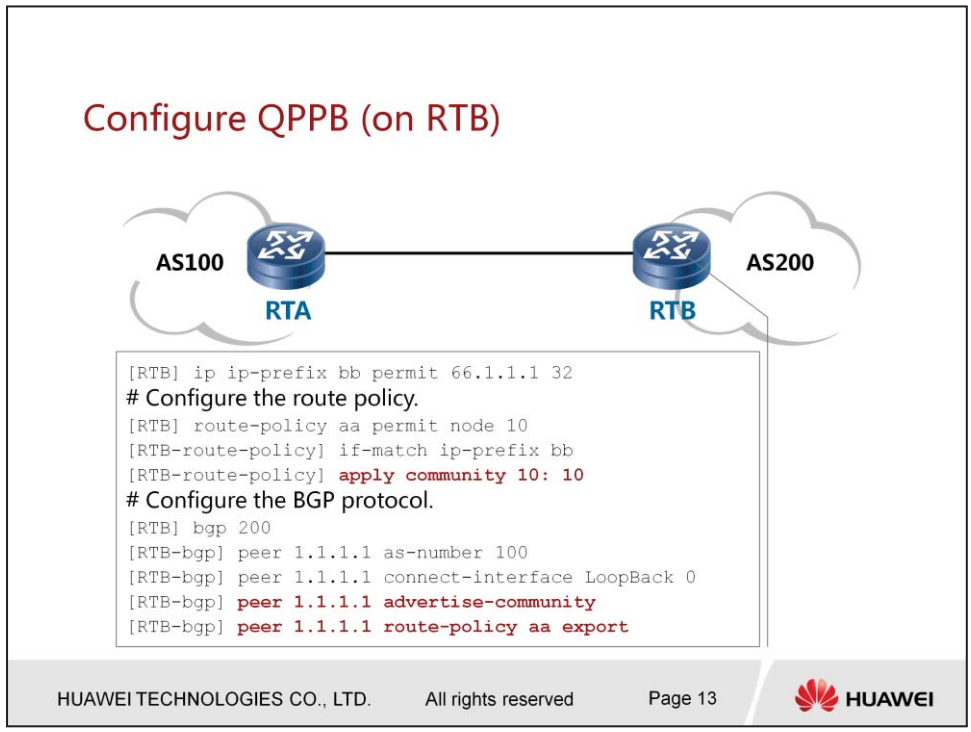
QoS policy can be applied according to the IP precedence, QoS local ID, and traffic behavior name.

The main advantage of QPPB is that the sender of BGP routes can classify the routes by setting the BGP attribute, and the receiver of BGP routes can apply the local QoS policy for the BGP routes based on the BGP attribute. In a complex networking environment, the classification policy often needs be dynamically changed. QPPB simplifies the change of the route policy on the route receiver. The route policy needs to be changed on only the route sender.



In this figure, RTB advertises the BGP routes with the BGP attribute to RTA. After receiving the BGP routes, RTA matches the BGP community list, ACL, BGP AS path list to set the IP precedence,

QoS local ID, and traffic behavior name for the BGP routes.
After QPPB is enabled on the interfaces connecting RTA and RTB, the corresponding QoS policy is applied for the packets sent to RTB.



In this example, when RTB sends the routing information to RTA, it adds the attribute 10: 10 to the route 66.1.1.1/32.

1. Configure the IP prefix **bb** to match the route 66.1.1.1/32.

2. Define a route policy **aa** and adds the community attribute 10: 10 to the routes matching IP prefix **bb**.

3. Use the advertise-community parameter in the command ***peer 1.1.1.1 …*** to configure the router to deliver the community attribute with the routing information to the neighbor. Apply route policy **aa** in the export direction.

The configuration commands are described as follows:

• Configure the route policy and set QoS parameters.

1. Enter the system view: system-view

2. Create a route policy node and enter the route policy view: route-policy *routepolicy-*

name { permit | deny } node *node-number*

3. Match the ACL: if-match acl *acl-number*

Or match the AS path in the routing information: if-match as-path *as-path-aclnumber*

Or match the community attribute in the BGP routing information:

if-match community { *basic-comm-filter-num* [whole-match] | *ext-comm-filternum* } &<1-16> Or match the route cost in the routing

information: if-match cost *value* Or match the ip prefix list: if-match

ip-prefix *ip-prefix-name*

4. Set the QoS local ID: apply qos-local-id *qos-local-id*

Or set the IP precedence: apply ip-precedence *ip-precedence*

Or associate the traffic behavior: apply behavior *behavior-name*

Caution: The IP precedence, QoS local ID, and the associated traffic behavior are QoS parameters for QPPB. They are mutually exclusive and you can configure only one of them for a route policy.
The traffic behavior must be configured beforehand.

•Apply the route policy with the QoS parameters.

1. Enter the system: system-view

2. Enable BGP and enter the BGP view: bgp *as-number*

3. Create a BGP peer: peer *ip-address* as-number *as-number*

4. Configure the policy for receiving the BGP routes: peer *ip-address* route-policy *route-policy-name* import

•Define the traffic classifier and match rule.

1. Enter the system view: system-view

2. Define the traffic classifier and enter the traffic classifier view: traffic classifier *classifier-name* [operator { and | or }]

3. Define the match rule for the classifier: if-match qos-local-id *qos-local-id* [source]

•Configure the class-based behavior.

•Define the traffic policy.

1. Enter the system view: system-view

2. Define the traffic policy and enter the traffic policy view: traffic policy *policyname*

3. Specify the behavior for the classifier in the traffic policy: classifier *classifiername*

behavior *behavior-name*

In this example, the community attribute of the routes is set to 10: 10 when RTB advertises route 66.1.1.1 to RTA.

The configuration commands are described as follows:

•Apply the traffic policy on the interface.

1. Enter the system view: system-view

2. Enter the interface view: interface interface-type *interface-number*

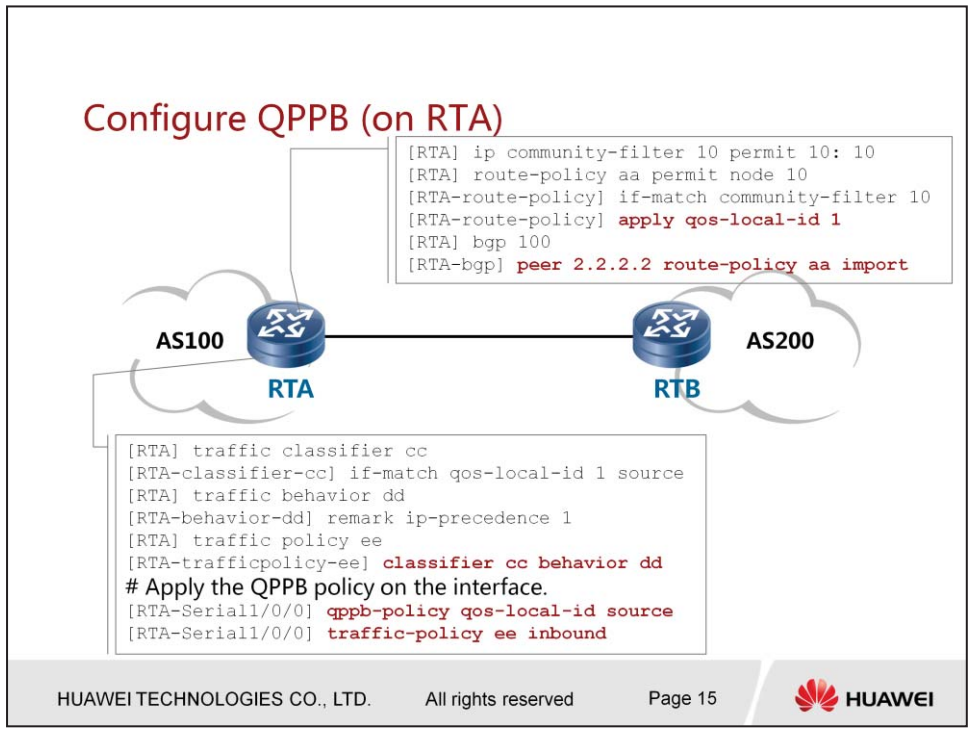
3. Apply the traffic policy: traffic-policy *policy-name* { inbound | outbound }

•Apply QPPB on the interface.

1. Enter the system view: system-view

2. Enter the interface view: interface *interface-type interface-number*

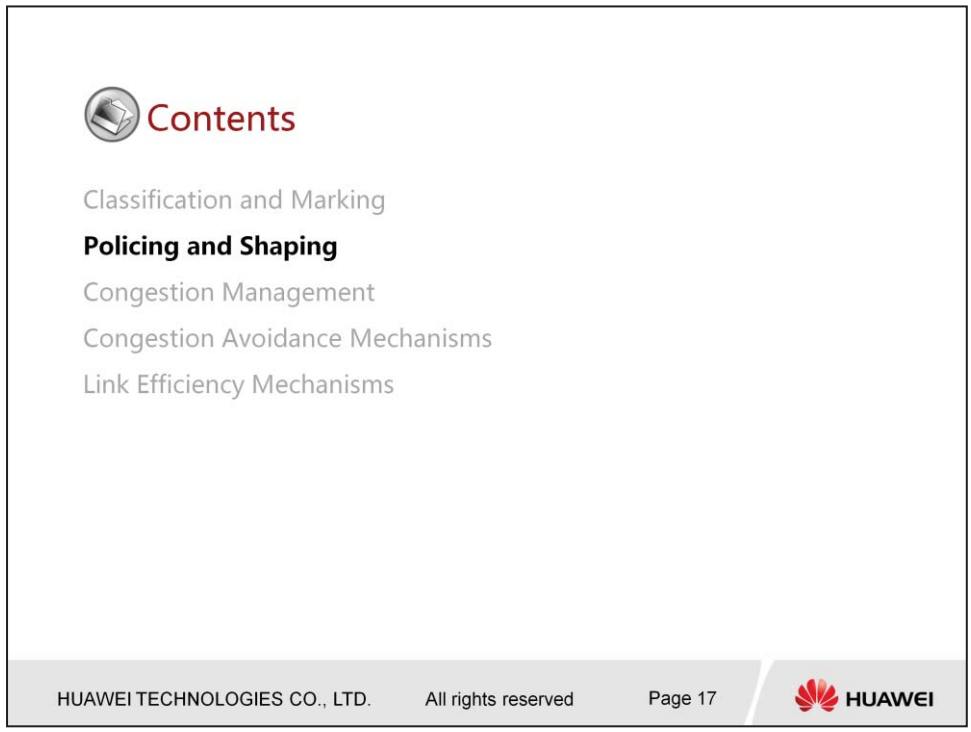
3. Apply QPPB: qppb-policy qos-local-id { source | destination }

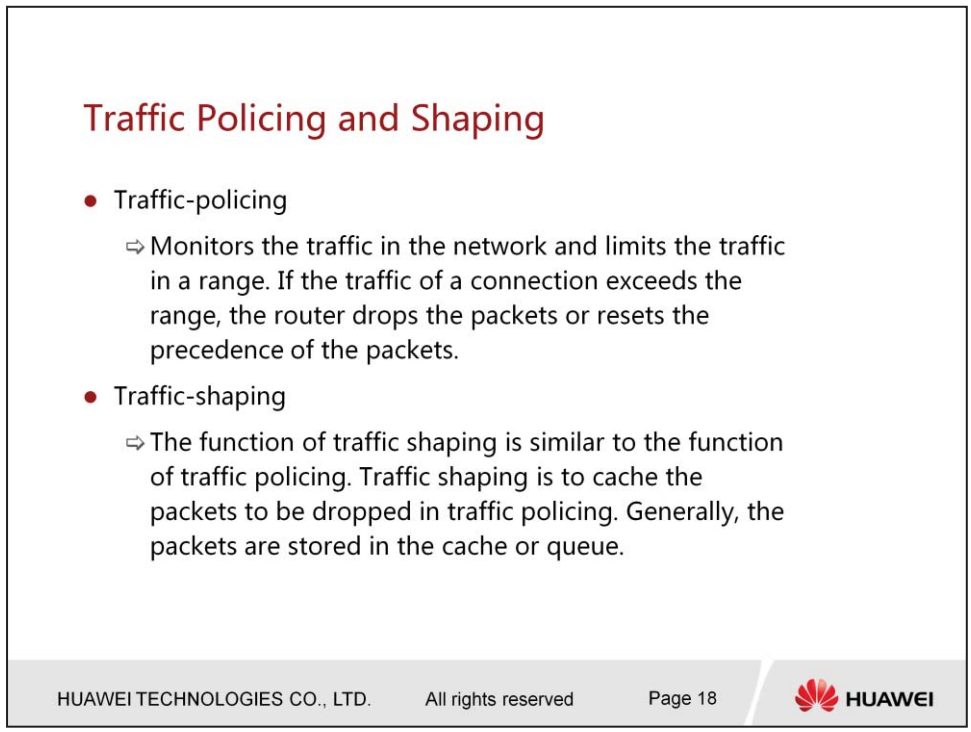


1. When RTA receives the route containing the 10: 10 community attribute, set the QoS local ID to 1 in the route policy.

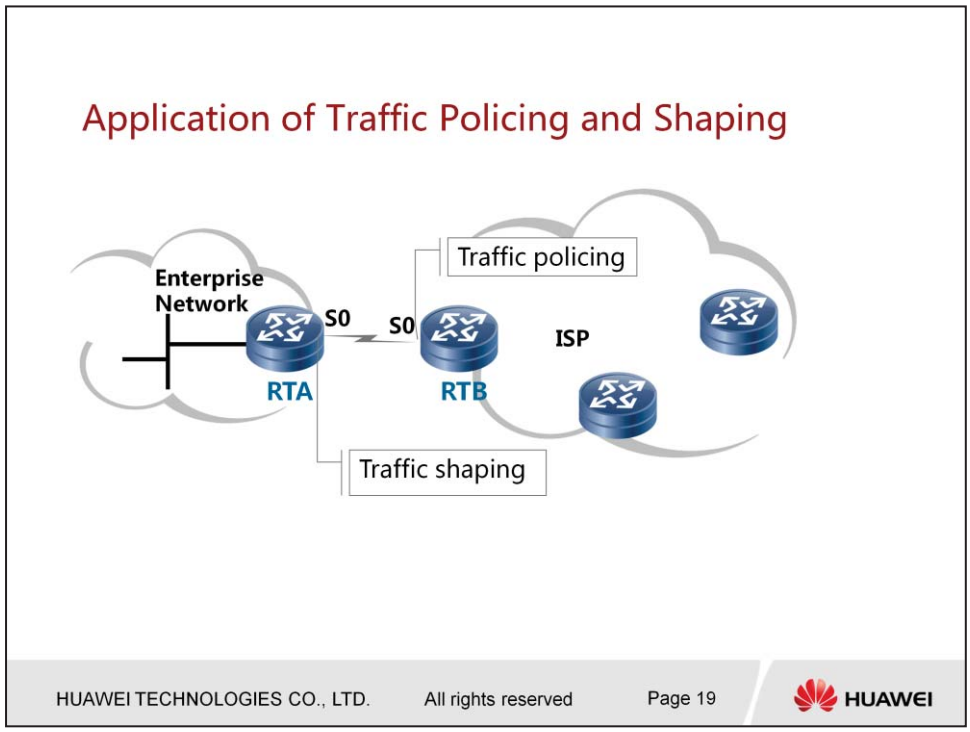
2. On RTA, configure the traffic policy to change the IP precedence of the route with QoS local ID 1 to 1. This is a re-mark behavior. You can also configure other QoS behaviors.

3. Apply the QPPB policy on the interface. After the configuration, RTA can automatically change the IP precedence of the packet to 1 (performs PHB) when it sends the packet to 66.1.1.1.


Contents
Classification and Marking
Policing and Shaping
Congestion Management
Congestion Avoidance Mechanisms
Link Efficiency Mechanisms
HUAWEI TECHNOLOGIES CO., LTD.
All rights reserved
Page 17
HUAWEI

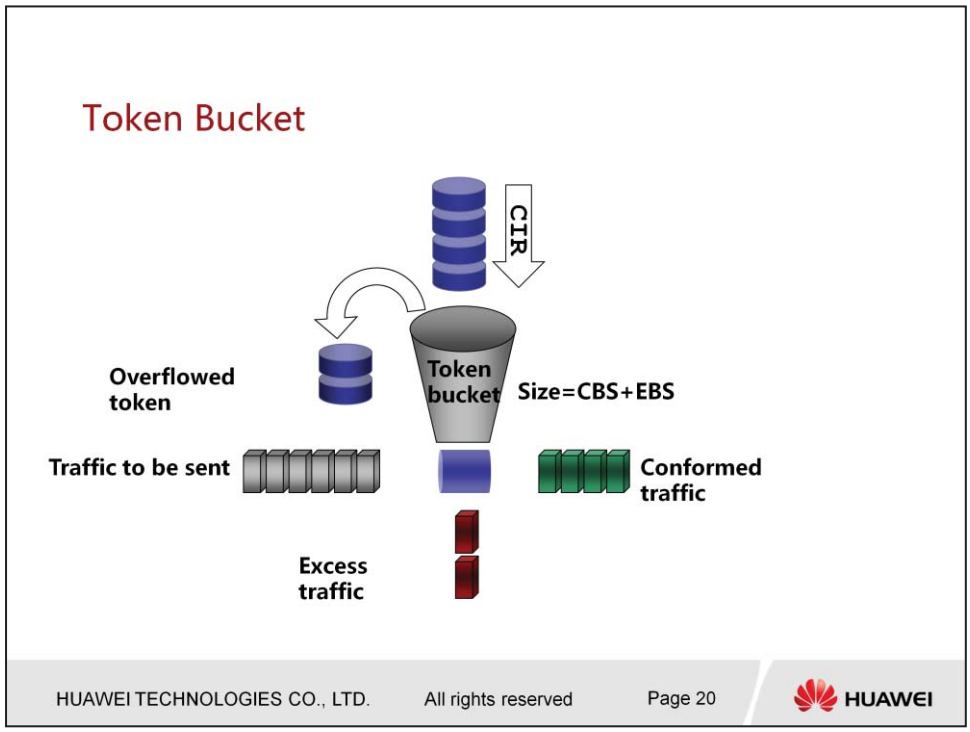


When data is transmitted from a high-speed link to a low-speed link, the incoming interface of the low-speed link becomes the bottleneck. This causes severe data loss and delay, especially for the data that requires low delay, such as the voice data or the data that requires low packet loss, such as the signaling data. A typical function of traffic policing is limiting traffic and burst size of the inbound and outbound packets in the network. If the packets meets certain condition, for example, the traffic of a connection exceeds the threshold, traffic policing carries out corresponding behavior to handle the excess packets. The packets may be dropped or the precedence of the packets may be changed. In general, CAR is used to limit the traffic of a certain type of packets. For example, the CAR can limit the bandwidth of HTTP packets to 50% of the total bandwidth. A typical function of traffic shaping is limiting the traffic and burst size of the outbound packets of a connection in the network. When the packet transmission rate exceeds the threshold, the packets are cached to the buffer. Under the control of the token bucket, the packets in the buffer are sent evenly.



When the adjacent network sends packet at a rate higher than the maximum rate that the local network can handle, traffic policing can be applied on the ingress of the network. Traffic policing on the egress is also supported but not commonly used.

If traffic policing is adopted in the upstream adjacent network, traffic shaping needs to be configured on the egress of the local network. Traffic shaping lowers the traffic and thus reduces the dropped packets and avoids congestion on the egress. Note that traffic shaping increases the transmission delay because of its caching mechanism.



The token bucket is used to assess whether the traffic exceeds the specified limit and carry out corresponding measures. The token bucket contains tokens instead of packets. A token is generated and added to the token bucket every Δ t period. When the token bucket is full, the new token is dropped. A token permit to send a single bit (or, in some cases, a byte) of traffic. A packet can pass through when there are enough tokens in the bucket to send this packet.

The number of tokens decreases accordingly, depending on the packet length. If there are not enough tokens, the packet is dropped and the number of tokens does not change. The assessment of whether the tokens in the bucket are enough for forwarding packets has two results: conform and excess.

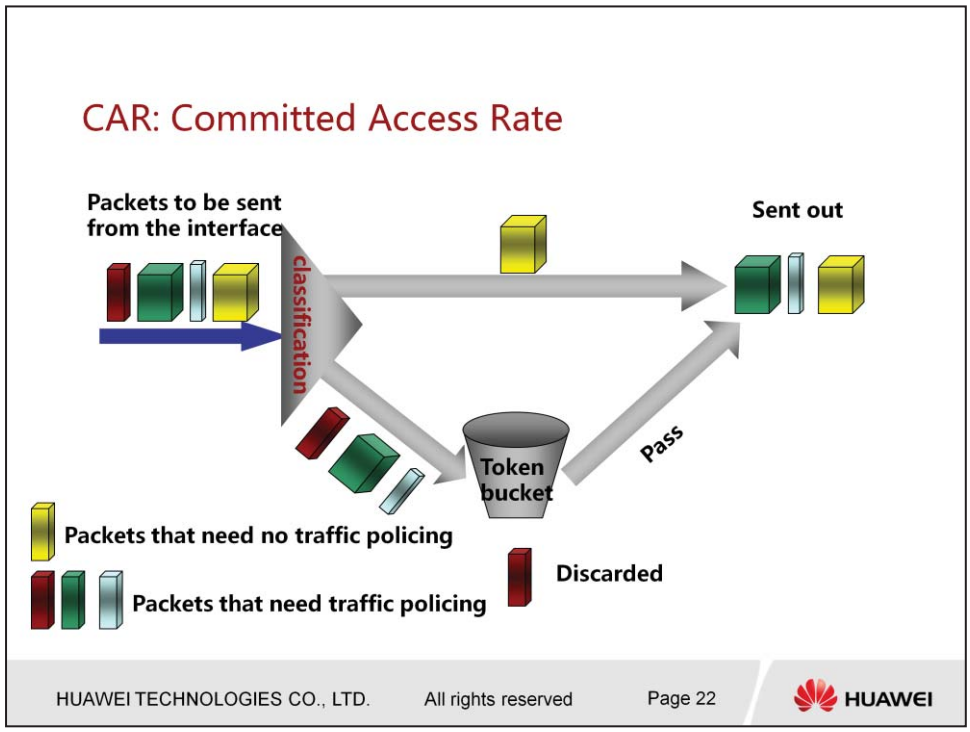
The parameters of the token bucket for assessing the traffic are as follows:

Committed Information Rate (CIR): the rate at which tokens are added to the bucket.

Committed Burst Size (CBS): capacity of the token bucket, namely the maximum size allowed for a traffic burst. The CBS must be larger than the packet length.

To measure more complex traffic and carry out more flexible control policy, you can configure two token bucket. For example, the policy of traffic policing involves three parameters: Committed Information Rate (CIR), Committed Burst Size (CBS), and Excess Burst Size (EBS). Two token buckets are used in this policy. The rates of both buckets are CIR, but their sizes are CBS and EBS respectively. The buckets are respectively called bucket C and bucket E for short. In traffic assessment, there are three cases: Bucket C has enough token;

bucket C has not enough tokens but bucket E has enough tokens; either bucket C or bucket E has enough tokens. Different traffic control policies can be adopted for these cases.

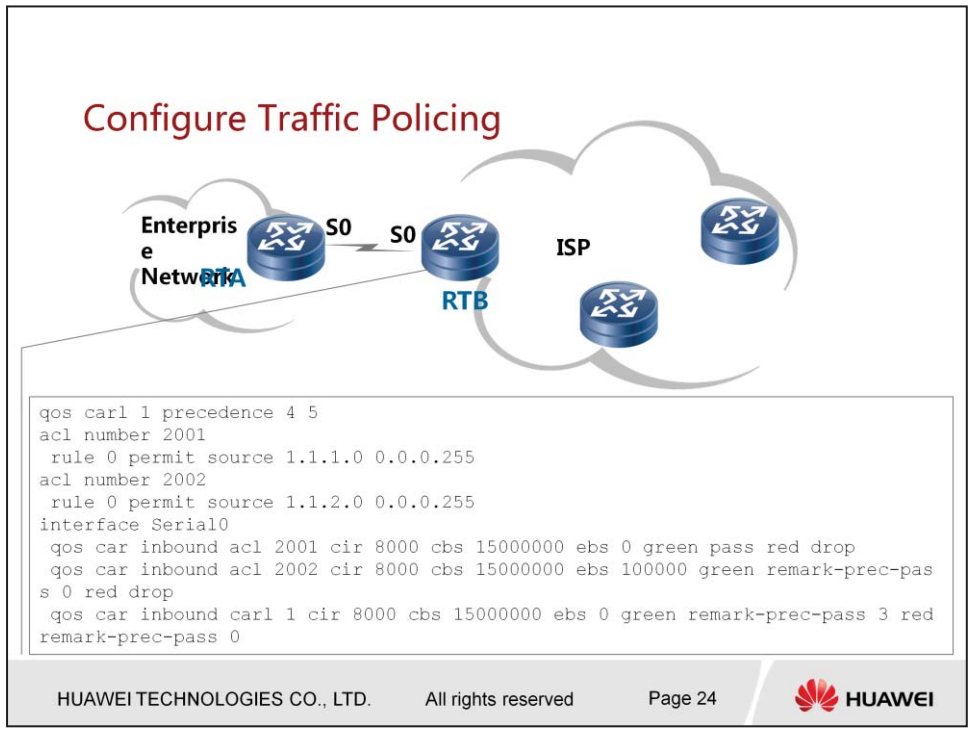


The CAR can be used in policing of specific traffic. The excess traffic is dropped or re-marked. Packets are classified according to the predefined match rule. If the packets do not need traffic policing, they are sent directly and not processed by the token bucket. If the packets need traffic policing, they are processed by the token bucket. Here we assume that the packet length is B and the number of tokens is TB. For the packets sent to the token bucket, if B –TB<0, the packets that obtain the tokens are marked green, regardless of the behavior performed for theses packets (dropping or forwarding the packets). At the same time, the number of tokens decreases accordingly, that is, TB=TB-B. The packets that do not obtain the tokens are marked red. The number of tokens does not change, regardless of behavior performed for these packets. If the behavior for the red packets is set to PASS, the packets can be sent out although they do not obtain tokens. For example, the length of the packets in the bucket is B=800 bits, and TB=30000 bits. 30000-800>0, so the packets are marked green. The number of tokens TB=30000-800.

When no token exists in the token bucket, the packets cannot be sent until new tokens are generated. Therefore, the traffic of the packets must be less than the rate at which the tokens are generated. In this way, the traffic is limited. Tokens are added to the token bucket at the set rate. The user can also set the capacity of the token bucket.

In the application, CAR setting in VRP of Huawei can be used not only to control the packet, but also to mark and re-mark the packets. More specifically, the CAR can be used to set or change the precedence of IP packets. For example, when the packets conform to the traffic feature, the precedence can be set to 5. When the packets do not conform to the traffic feature, the router drops the packets or changes the precedence to 1 and forwards the packets. In the following packet processing, the router tries to forward all packets with precedence 5. If

no congestion occurs, the router also sends the packets with precedence 1. When congestion occurs, the router drops the packets with precedence 1 before dropping the packets with precedence 5. CAR can be used to set the traffic feature and mark for different types of traffic. That is, packets are classified. Different types of packets have different traffic features and marks. You can use CAR policies jointly. For example, you can set a traffic limit for all packets. Within the traffic limit, you can set the traffic feature for some of packets. The CAR takes effect on the incoming and outgoing interfaces.



In this example, the CAR list is defined to match the packets with precedence 4 and 5. Two ACLs are defined to match the packets with the source IP addresses

in the ranges of 1.1.1.0-1.1.1.255 and 1.1.2.0-1.1.2.255.Apply CAR policies in the inbound direction of the serial0 interface on RTB. The first CAR policy limits the traffic of the packets with the source addresses in the range of 1.1.1.0-1.1.1.255. (The CIR is 8000 bps; the CBS is 15000000 bits; the EBS is 0.) The excess traffic is dropped. The second CAR policy limits the traffic of the packets with the source addresses in the range of 1.1.2.0-1.1.2.255. (The CIR is 8000 bps; the CBS is 15000000 bits; the EBS is 100000 bits.) For the traffic within the limit, the precedence is re-marked to 0. The excess traffic is dropped. The third CAR policy limits the traffic of the packets with precedence 4 and 5. (The CIR is 8000 bps; the CBS is 15000000 bits; the EBS is 0.) For the traffic within the limit, the precedence is re-marked to 3. For the traffic that exceeds the limit, the precedence is re-marked to 0. The configuration commands are as follows:

•Configure the CAR list.

1. Enter the system view: system-view

2. Configure the CAR list: qos carl *carl-index* { precedence *precedence-value*&<1-

8> | mac *mac-address* }

By repeating the command with different carl-index values, you can create multiple CAR lists. By repeating the command with the same carl-index, you can

change the parameters in the CAR list. That is, the new CAR list overwrites the previous one.

To match multiple precedence levels in a CAR list, you can specify multiple precedence-values.

•Configure the CAR policy.

1. Enter the system view: system-view

2. Enter the interface view: interface *interface-type interface-number*

3. Configure the CAR policy: qos car { inbound | outbound } { any | acl *acl-index* | carl *carl-index* } cir *cir* cbs *cbs* ebs *ebs* green *action* red *action*

car: applies the committed access rate to perform traffic policing.

inbound: limits the rate of the packets received by the interface.

outbound: limits the rate of the packets sent by the interface.

any: limits the rate of all IP packets.

acl: limits the rate of the packets matching the access control list (ACL).

acl-index: specifies the number of an ACL. The value ranges form 2000 to 3999.

carl: limits the rate of the packets matching the CAR list.

carl-index: specifies the number of a CAR list. The value ranges from 1 to 199 currently.

cir: indicates the committed information rate.

committed-information-rate: specifies the value of the committed information rate. The value ranges from 8000 bps to 155000000 bps.

cbs: indicates the committed burst size, namely the burst traffic generated when the average rate is within the committed rate.

committed-burst-size: specifies the value of the committed burst size. The value ranges from 15000 bits to 155000000 bits.

ebs: indicates excess burst size.

excess-burst-size: specifies the value of the excess burst size. The value ranges from 0 bits to 155000000 bit.

green: indicates the action taken when the data traffic conforms to the committed rate.

red: indicates the action taken when the data traffic does not conform to the committed rate.

action: specifies the action taken for the packets, including:

continue: leaves the packets to the next CAR policy.

drop: drops the packets.

pass: sends the packets.

remark-prec-continue: sets a new IP precedence *new-precedence* and leaves the packets to the next CAR policy. The value of *new-precedence* ranges from 0 to 7.

remark-prec-pass: sets a new IP precedence *new-precedence* and sends the packets to the destination address. The value of *new-precedence* ranges from 0 to 7.

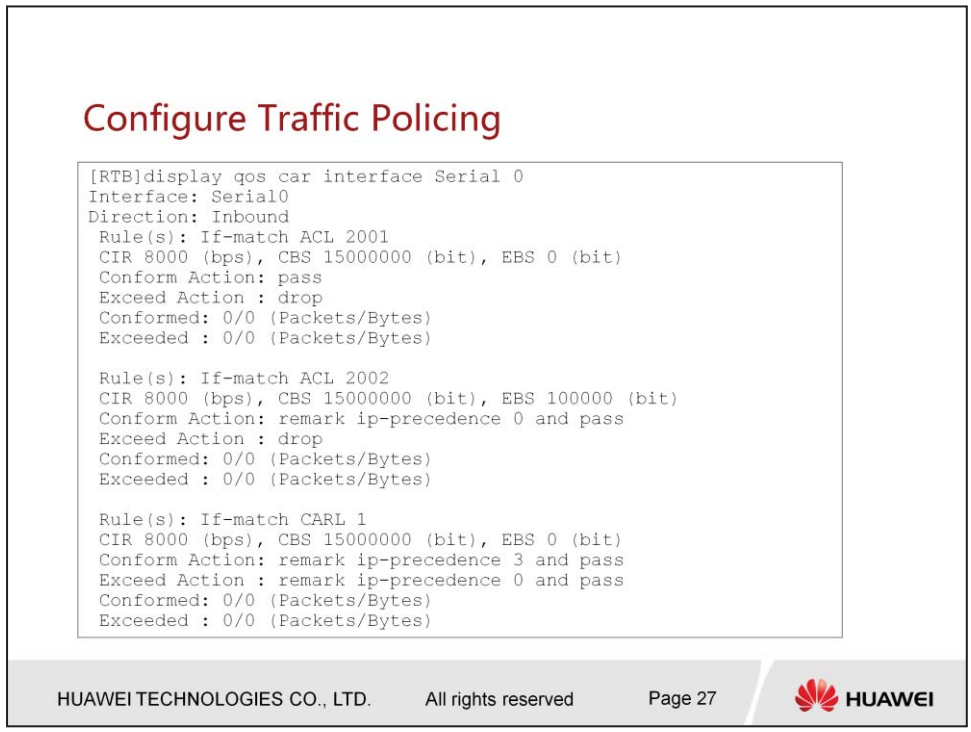
remark-mpls-exp-continue: Sets a new MPLS EXP value *new-mpls-exp* and leaves the packets to the next CAR policy. The value of *new-mpls-exp* ranges from 0 to 7.

remark-mpls-exp-pass: Sets a new MPLS EXP value *new-mpls-exp* and send the packets to the destination address. The value of *new-mpls-exp* ranges from 0 to 7.

The CAR policy is applicable to only the IP packets. It can be applied to the incoming interface and outgoing interface of packets. Multiple CAR policies can be configured on an interface.

If the **acl** keyword is used, you can set the CAR parameters for the flow matching an ACL. Or you can set CAR parameters for different flows by using different ACLs. If the **any** keyword is used, you can set the CAR parameters for all flows. If you repeat the command, the new settings overwrite the previous settings.

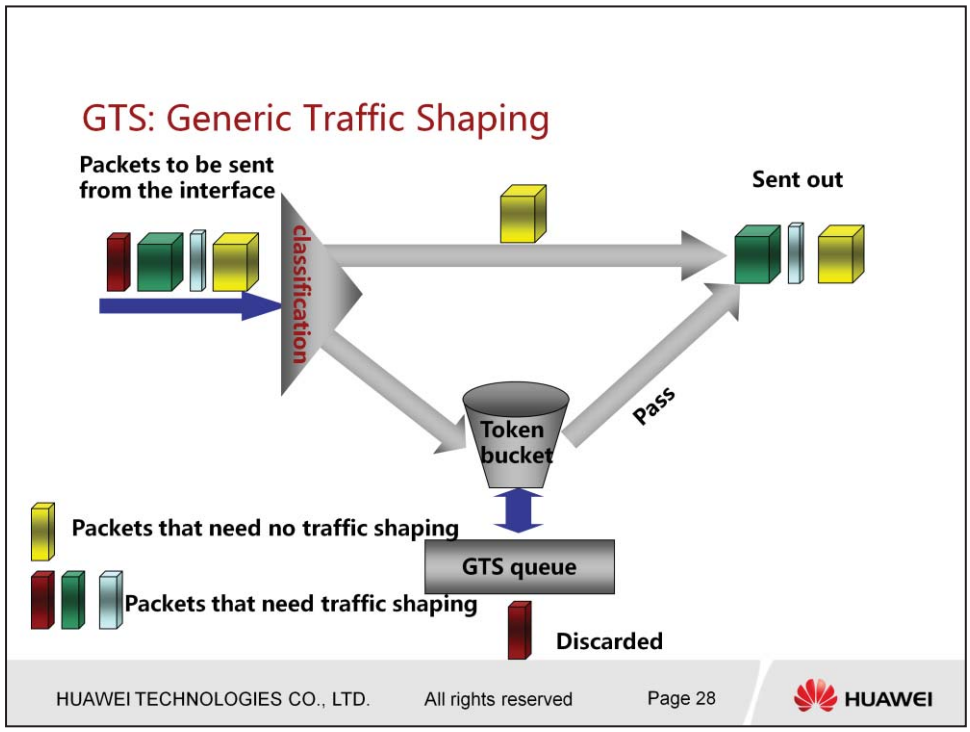
The **acl** and **any** keywords cannot be used at the same time.



In this example, for the data flow matching ACL 2001, the CIR is 8000 bps; the CBS is 15000000 bits; the EBS is 0. The traffic within the limit is forwarded and the traffic exceeding the limit is dropped.

For the data flow matching ACL 2002, the CIR is 8000 bps; the CBS is 15000000 bits; the EBS is 100000 bits.The traffic within the limit is forwarded after itsprecedence is re-marked to 0, the traffic exceeding the limit is dropped.

For the data flow with the precedence 4 and 5, the CIR is 8000 bps; the CBS is 15000000 bits; the EBS is 0. The traffic within the limit is forwarded after its precedence is re-marked to 3. The traffic exceeding the limit is forwarded after its precedence is re-marked to 0.

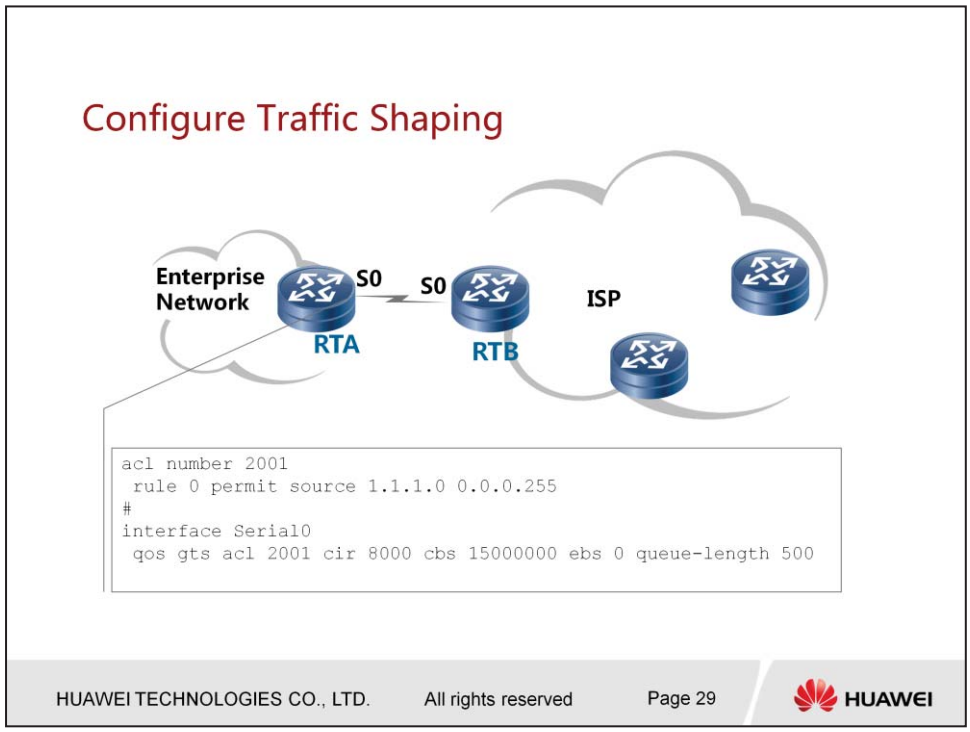


Generic Traffic Shaping (GTS) shapes abnormal traffic or the traffic that does not conform to any traffic feature. GTS guarantees the bandwidth allocation among upstream and downstream network nodes and avoids the congestion. Similar to CAR, GTS also adopts the token bucket to control the traffic. In traffic

control through CAR, the packets that do not conform the traffic feature are dropped. In GTS, the packets that do not conform to the traffic feature are cached.

This reduce packet loss and makes the packets conform to the traffic feature. This figure shows the GTS process. The queue used to cache the packets is call the GTS queue.

GTS can shape the specified packet flow or all packets.
The received packets are classified. If the packets do not need GTS, they are sent and not processed by the token bucket. The token bucket is the same as the bucket used in CAR control. If the packets need GTS, they are matched with the token in the tokenbucket. If the packet length B – TB <0, the packets are sent; otherwise, the packets are cached to the GTS queue (FIFO queue). This FIFO queue is not the same FIFO on the interface. The length of the queue is a certain value (measured by packets). When the number of packets to be cached is larger than the length of the GTS queue, the packets are dropped. When the GTS queue contains packets, GTS sends the packets at certain intervals. Each time GTS sends a packet in the queue, it compare the packet with the number of tokens. If there are enough tokens, the packet is sent; otherwise, the packet still stays inthe queue. In addition, GTS allows burst traffic. GTS takes effect only on the outgoing interface.



This configuration example is to perform traffic shaping for the packets with source addresses in the range of 1.1.1.0-1.1.1.255. The packets that exceeds the CAR limit (CIR: 8000 bps; CBS: 15000000 bits; EBS: 0) are cached to the GTS queue. The length of the GTS queue is 500 packets.

The configuration commands are as follows:

•Configure traffic shaping.

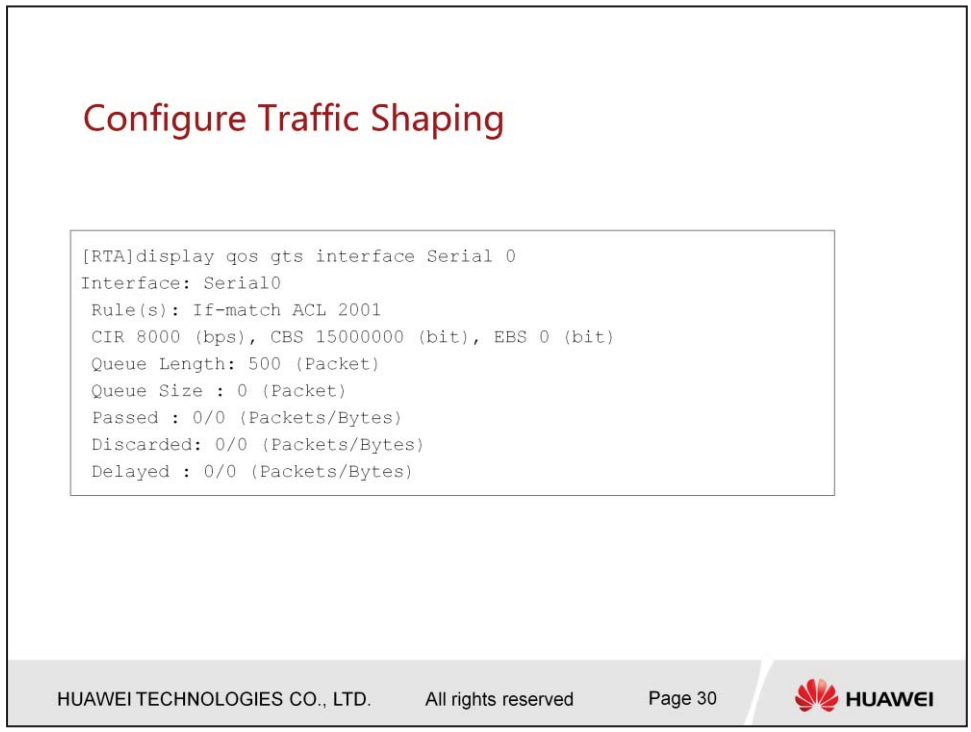
1. Enter the system view: system-view

2. Enter the interface view: interface *interface-type interface-number*

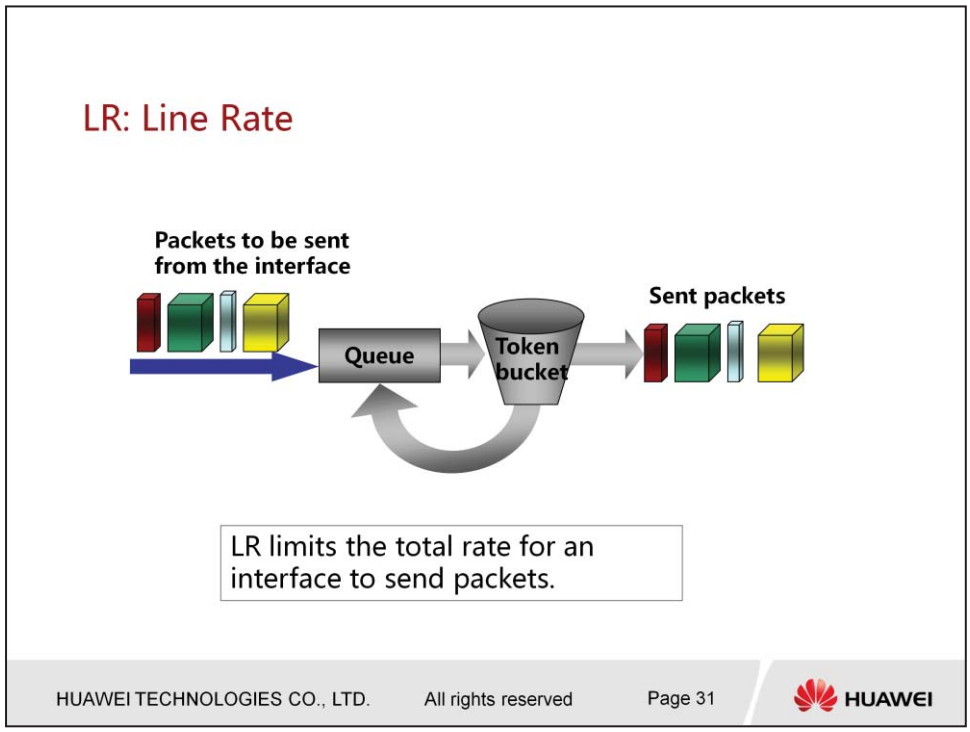
3. Configure the GTS policy: qos gts { any | acl *acl-index* } cir *cir* [cbs *cbs* [ebs *ebs* [queue-length *queue-length*]]]

If the **acl** keyword is used, you can set the GTS parameters for the flow matching an ACL. You can set GTS parameters for different flows by using different ACLs. If the **any** keyword is used, you can set the GTS parameters for all flows. If you repeat the command, the new settings overwrite the previous settings.

The **acl** and **any** keywords cannot be used at the same time.

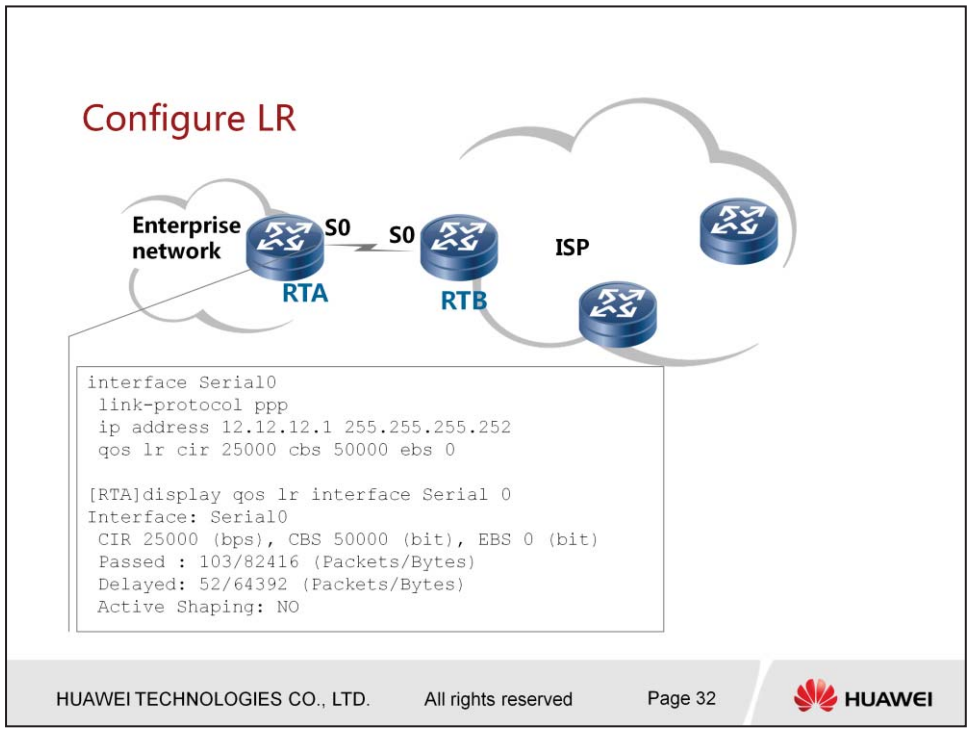


After the configuration, you can run the **display qos gts interface** command to check the effect of the configuration.



The line rate (LR) limits the total rate for a physical interface to send packets (including urgent packets). Traffic control through LR is also implemented through the token bucket. If the user set the LR on an interface to specify the traffic feature, all packets to be sent by this interface are processed by the token bucket. If the bucket has enough tokens to send packets, the packets are sent; otherwise, the packets are added to the QoS queue for congestion management. In this way, the traffic passing through the physical interface can be controlled. The figure shows the LR processing flow. When the token bucket is adopted to control the traffic, burst packets can be transmitted when there are tokens in the bucket. If the bucket has no token, no packet can be sent, until new tokens are generated. Therefore, the traffic of packets must be less than the rate at which tokens are generated. In this case, the traffic is limited and burst traffic is allowed to pass through. LR can limit the traffic of all packets passing through a physical interface. CAR and GTS are implemented on the IP layer and are invalid for non-IP datagram. Compared with GTS, the LR not only caches the packets exceeding the traffic limit, but also processes the packets by using the QoS queue. Therefore, LR uses more flexible queue scheduling mechanism. If the customer needs only to limits the traffic of all packets, LR is a simpler configuration method.

The network investor can hide the actual bandwidth for customers so that the customers can only use the bandwidth they have purchased.



This example is to limit the traffic on the serial0 interface of RTA and to add the excess traffic to the defined QoS queue for scheduling. (The CIR is 25000 bps;

the CBS is 50000 bits; the EBS is 0.) You can run the command **display qos lr interface** to view the information about traffic limit on the interface, including the traffic limit conditions, number of packets sent directly, and number of packets sent with a delay.

The configuration commands are described as follows:

•Configure the LR on the interface.

1. Enter the system view: system-view

2. Enter the interface view: interface *interface-type interface-number*

3. Configure the LR on the physical interface: qos lr cir *cir* [cbs *cbs* [ebs *ebs*]]

cir: indicates the committed information rate.

cir: specifies the value of the committed information rate. The value ranges from 8000 bps to 155000000 bps.

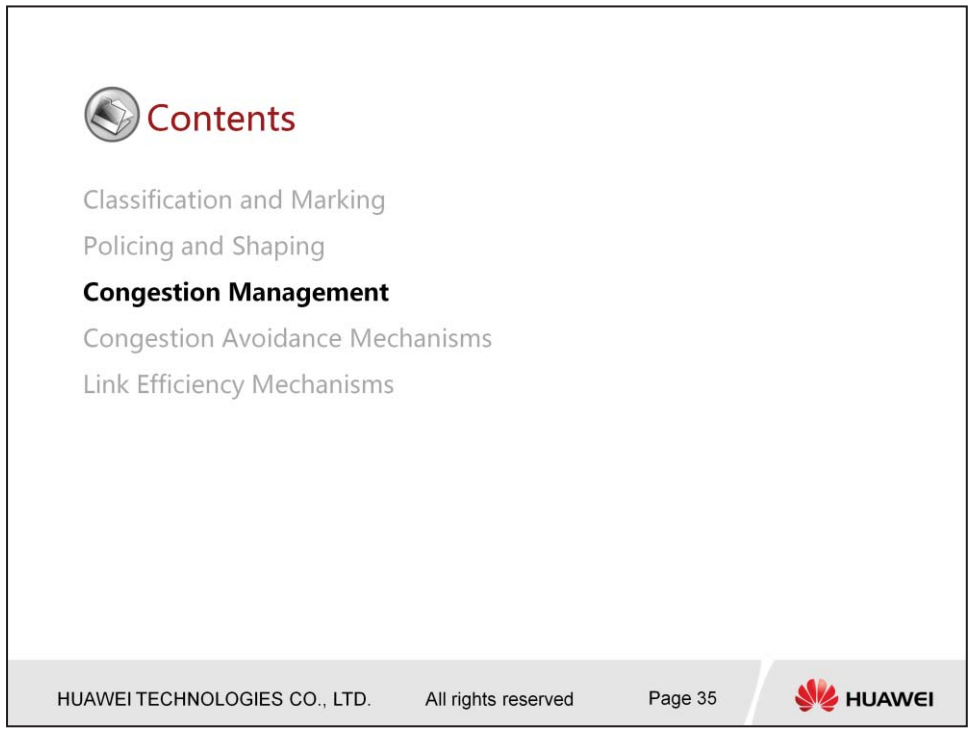
cbs: indicates the committed burst size, namely the burst size generated when the average rate is within the committed rate.

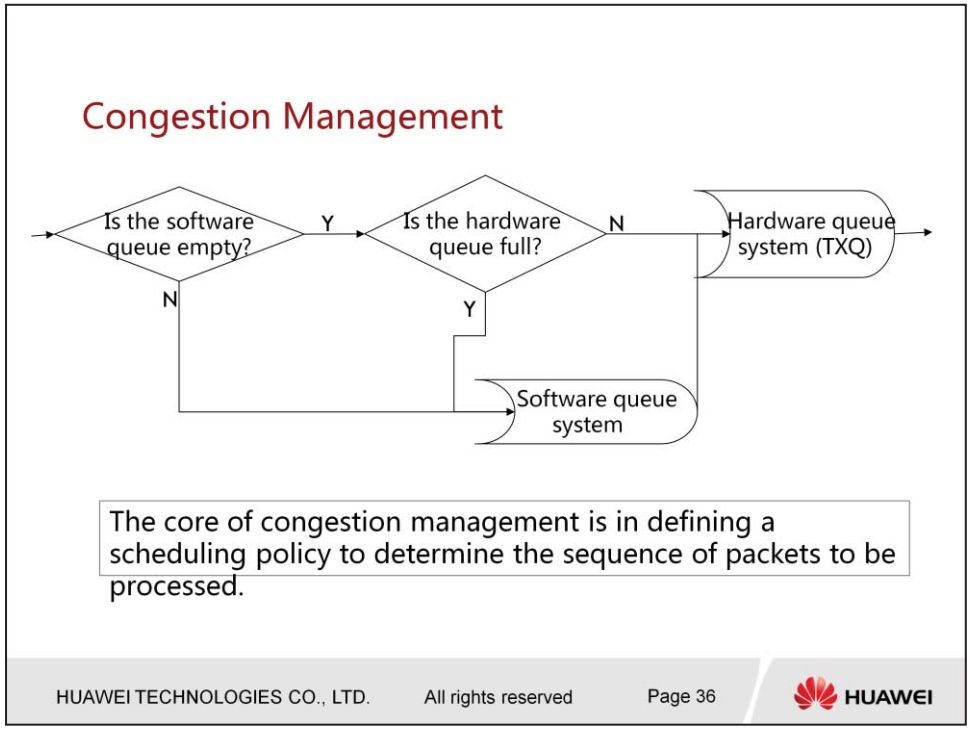
cbs: specifies the value of the committed burst size. The value ranges from 15000 bits to 155000000 bits. When cir>30000 bit/s, the default value of *cbs* is half of the cir value. When cir<30000 bit/s, the default value of *cbs* is 15000.

ebs: indicates the excess burst size.

ebs: specifies the value of the excess burst size. The value ranges from 0 to 155000000 bits.

By default, the value is 0, that is, only one token bucket is used. Command **qos lr** is used to limit the rate for an interface to send the data. Command **undo qos lr** is used to cancel the traffic limit. LR is also applicable to a tunnel interface. The traffic limit can be used with various queue scheduling algorithms to implement congestion management. Before configuring the queue on a tunnel interface, you must apply the LR on the tunnel interface. Before cancelling LR configured on the tunnel interface, you must delete the queue configured on the tunnel interface.

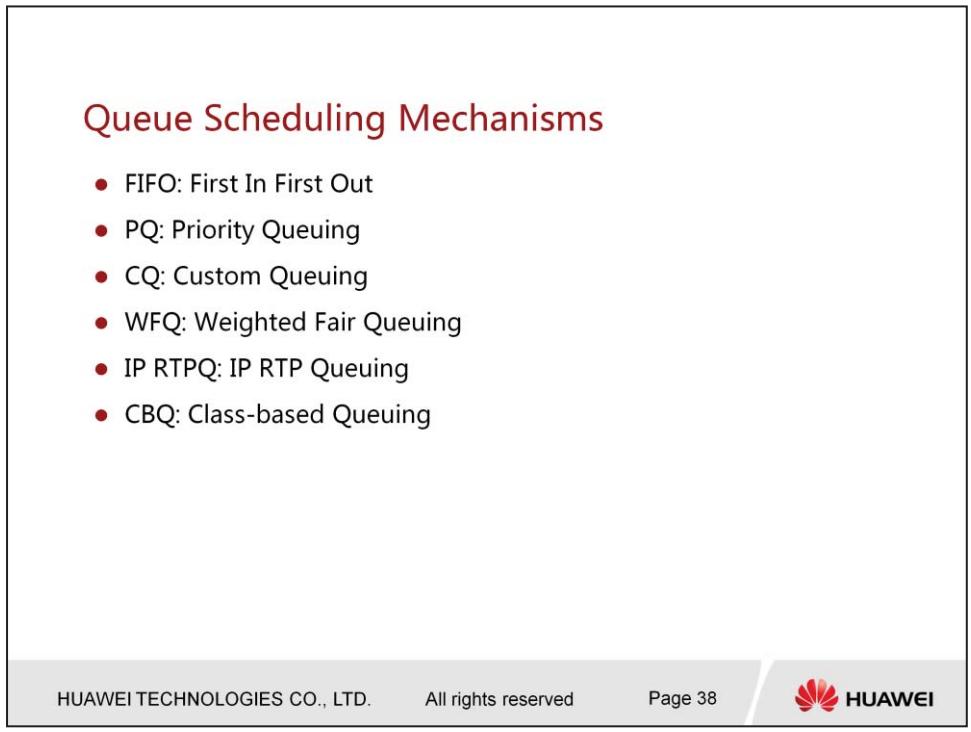

Contents
Classification and Marking
Policing and Shaping
Congestion Management
Congestion Avoidance Mechanisms
Link Efficiency Mechanisms
HUAWEI TECHNOLOGIES CO., LTD.
All rights reserved
Page 35
HUAWEI



For a network element, when packets are received faster than they are sent, congestion occurs on the interface. If the cache is insufficient for storing the packets, some packets will be lost. Packet loss may make the PC re-send the packets or the router retransmit the packets due to timeout.
This is a vicious circle.

When congestion occurs, many packets compete for network resources. Service packets that cannot get the resource are dropped. The situation is worse if the key service cannot be guaranteed with sufficient bandwidth, low delay, jitter, and other QoS parameters.
In this case, defining a resource scheduling policy todetermine the sequence of packets to be processed becomes the core of congestion management. Congestion management is implemented through thequeuing mechanism, which involves creation of queues, traffic classification that determines the queue for packets, and scheduling policy among queues. Initially, only the First In First Out (FIFO) scheduling policy was used. Later, other scheduling polices are designed to meet the requirements for different services. The queue scheduling mechanism involves two parts: the hardware queue and the software queue. The hardware queue is also called transmit queue (TxQ). It is used when the interface drive transmits packets one by one. The hardware queue is an FIFO queue. The software queue is used to schedule the packets in the hardware queue according to the requirement for QoS. The software queue can use various scheduling methods. Software queue is used only when the hardware queue is full. The length of the hardware queue relates to the bandwidth set on the

interface. If the bandwidth on the interface is large, the delay is relatively low. In this case, the queue should be long. If the bandwidth is small, the queue length should be short. A proper queue length is the key for congestion management. If the hardware queue is too long, the effect of the policy carried out by the software queue is reduced. If the hardware queue is too short, the scheduling efficiency is lowered, which lowers the link utilization and raises the CPU usage.



Listed above are the commonly used queue scheduling mechanisms that shall be discussed. CBQ willed be described in the Class-based QoS section.

Features of Queue Scheduling Mechanism

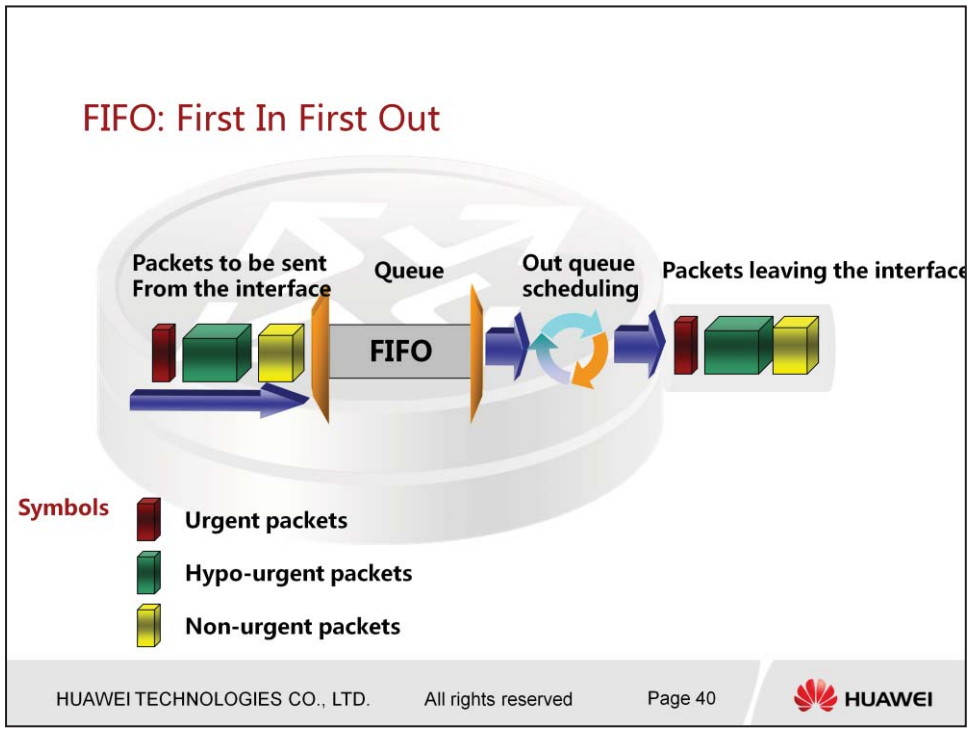
Feature	Description	Affected QoS Parameter
Classification	Check packets and determine the queue to which the packets are added.	None
Drop policy	Define the rule for the device to drop packets.	Packet loss
Scheduling mode for a single queue	In a queue, packets may be reordered. But in most cases, FIFO queuing is adopted.	Bandwidth, delay, jitter, packet loss
Scheduling mode among queues	Defines the queue that adds the packets to the TxQ (transmission Queue).	Bandwidth, delay, jitter, packet loss
Number of queues	Defines the granularity based on which the packets are classified.	None
Queue length	Defines the maximum number of packets that can be stored in a queue.	Packet loss, delay

HUAWEI TECHNOLOGIES CO., LTD. All rights reserved HUAWEI

The concept of queue is easy to understand. It also exists in our daily life. For example, we stand in a queue to buy tickets at the cinema. The first person in the queue get the ticket first. Sometimes, some one may buy the ticket first by showing a certificate, instead of waiting in the queue. This person has privilege and is treated with precedence.

Other people have to wait till this person gets the ticket. The queue scheduling mechanism is an important technology to manage congestion. When congestion occurs on the outgoing interface, proper queue scheduling mechanism can be adopted to guarantee the QoS parameters, such as the bandwidth, delay, and jitter, of certain type of packets. The queue mentioned here is the outgoing queue. It stores the packets until the interface has thecapability to send the packets. Therefore, the queue scheduling mechanism functions when congestion occurs on the outgoing interface. Another function of the queue scheduling mechanism is to reorder packet (FIFO queue has no such function). Main queuescheduling mechanisms are: FIFO, PQ, CQ, WFQ, CBQ, and IP RTPQ.

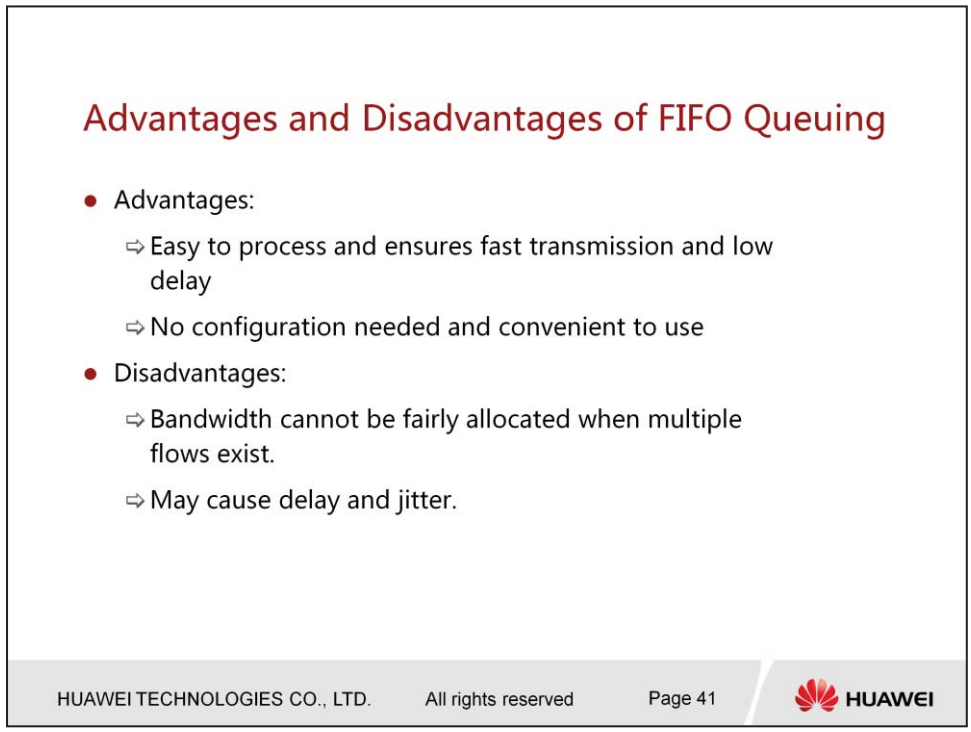
We can learn about these queuing technologies from the aspect of classification, drop policy, scheduling mode for a single queue, scheduling mode among queues, number of queues, and queue length.



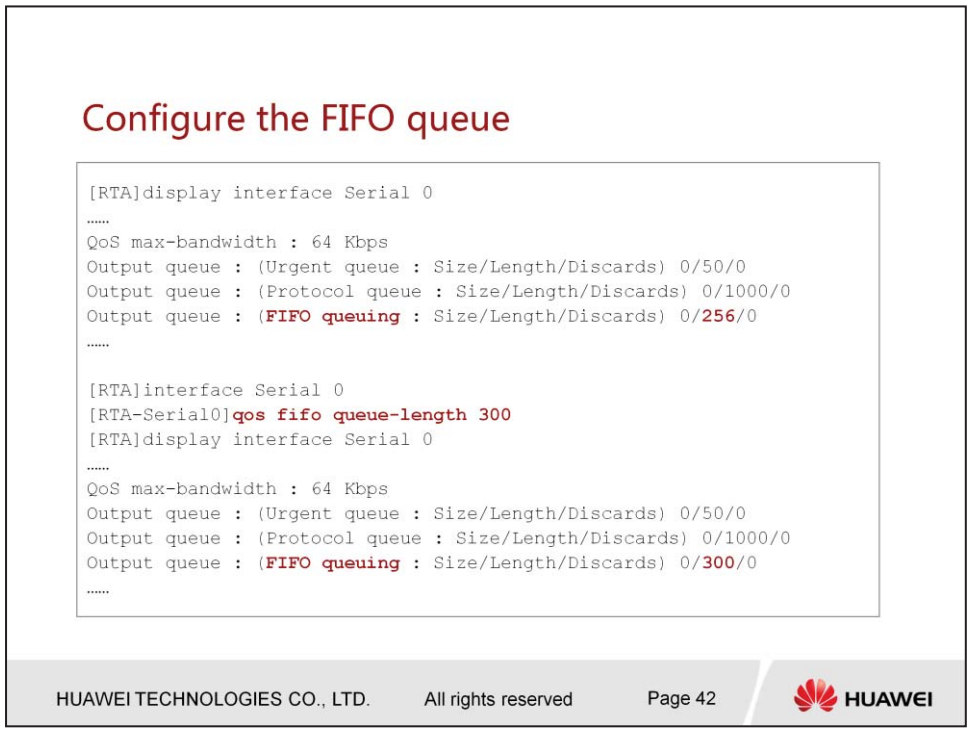
FIFO is the simplest queuing mechanism. Each interface can have only one FIFO queue. It seems that the FIFO queue does not provide any guarantee for QoS.

The fact is quite the contrary. Since there is only one queue on the interface, it is not necessary to determine to which queue certain type of packets should be added.

It also need not to determine which queue should the next packet be picked up and how many packets should be picked up. That is to say, FIFO queue does not need traffic classification and scheduling mechanism. In the FIFO queue, packets are sent in sequence, so FIFO queue does not need to reorder packets. The FIFO queue simplifies these processes and thus enhances the guarantee of low delay. The FIFO queuing mechanism concerns only the queue length, because the queue length influences the delay, jitter, and packet loss ratio. The queue length is limited, and a queue may be fully filled. So, drop policy is required in this mechanism. FIFO mechanism uses the tail drop policy. If the queue length is quite long, the queue is not easy to be fully filled and few packets will be dropped. However, a long queue causes long delay, and long delay usually increases the jitter. If the queue is quite short, low delay can be guaranteed, but more packets will be dropped. Other queuing mechanisms also have the similar problem. The tail drop policy specifies that if a queue is full, later packets are dropped. The later packets cannot replace the position of the packets in the queue.

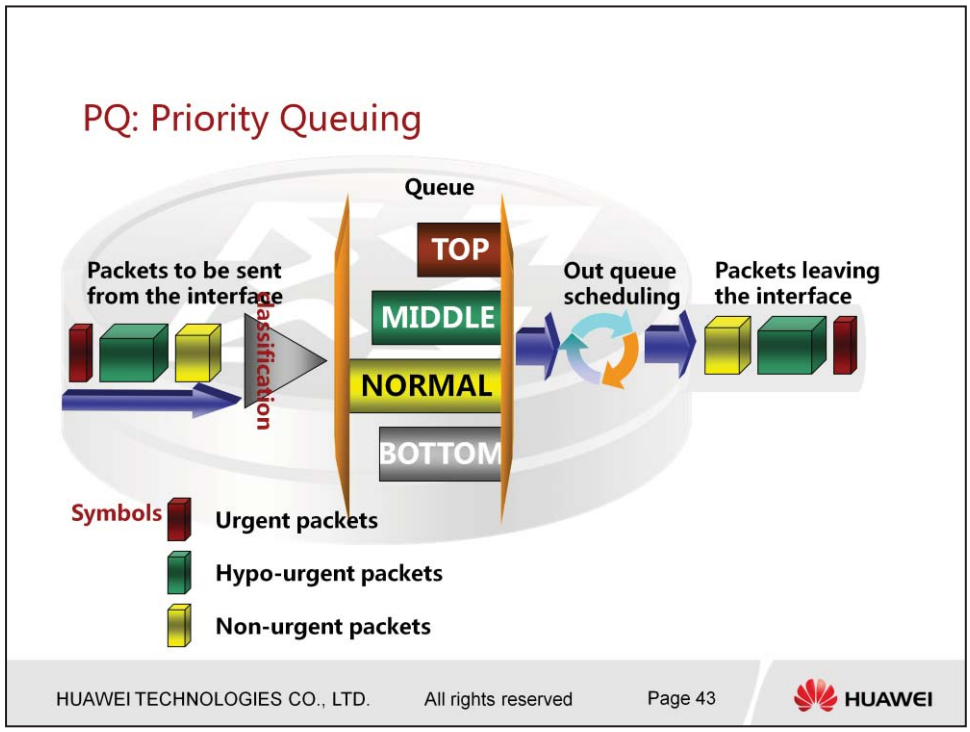


The main advantage of FIFO queuing is its simplicity and high speed, because it does not need any classification or complex scheduling. It is the default queuing mechanism for most interfaces and does not need extra configurations. When multiple flows need to be transmitted, FIFO cannot allocate the bandwidth fairly. Some flows may occupy much bandwidth because they may send large amount or large packets. In this case, high delay and jitter may be caused for delay-sensitive packets.



In this configuration example, the default queuing mechanism is FIFO. The length of the FIFO queue is 256. You can use command **qos fifo queue-length** to change the queue length.

If you set a long queue length, the queue is not easy to be filled and few packets will be dropped. However, a long queue causes high delay. If you set a short queue length, low delay can be ensured and burst packets can be prevented, however more packets will be dropped.



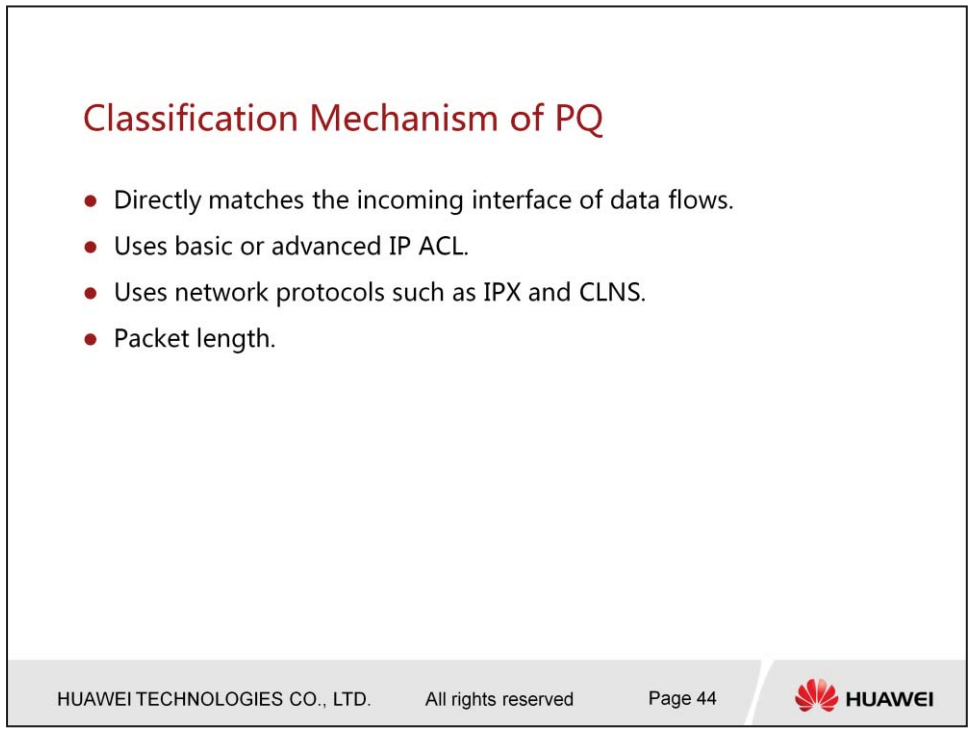
Priority queuing defines four queue levels: Top, Middle, Normal, and Bottom. Currently, most devices implement eight queuing priorities. Once the queue with higher priority has packets, the device always picks up packets in this queue. Given this, PQ has obvious advantage and disadvantage.

The advantage of PQ is that it ensures high bandwidth, low delay, and low jitter for the service with higher priority. Its disadvantage is that packets in the queue with lower priories cannot be scheduling in time or even not be scheduled.

PQ has the following features:

1. Packets can be classified through the ACL and can be added to proper queues as required.
2. Only the tail drop policy is adopted.
3. The queue length can set to 0, which indicates infinite queue length. That is, packets joining this queue will not be dropped unless the memory is exhausted.
4. FIFO logic is used in the queue.
5. Packets in the queue with higher priority are scheduled first.

The features of PQ shows that PQ ensures best service for a certain type of flows but dose not concern QoS of other flows.



PQ classifies packets based on the following factors:

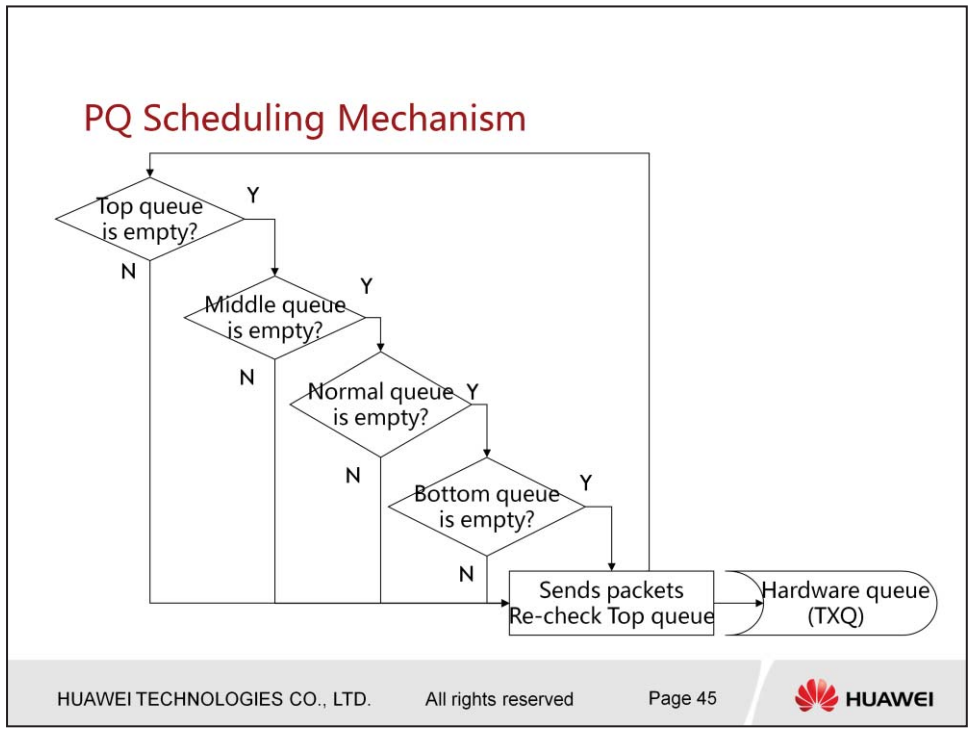
1. Incoming interface of data flows.

2. Basic or advanced ACL. The ACL can match the following parameters:

- Source IP address
- Destination IP address
- UDP/TCP source port number or port number range
- UDP/TCP destination port number or port number range
- IP precedence, namely the higher three bits of the ToS field
- DSCP value, namely the higher six bits of the ToS field.
- Packet fragments, which are identified by the fragmentation flag and offset value in the IP packets

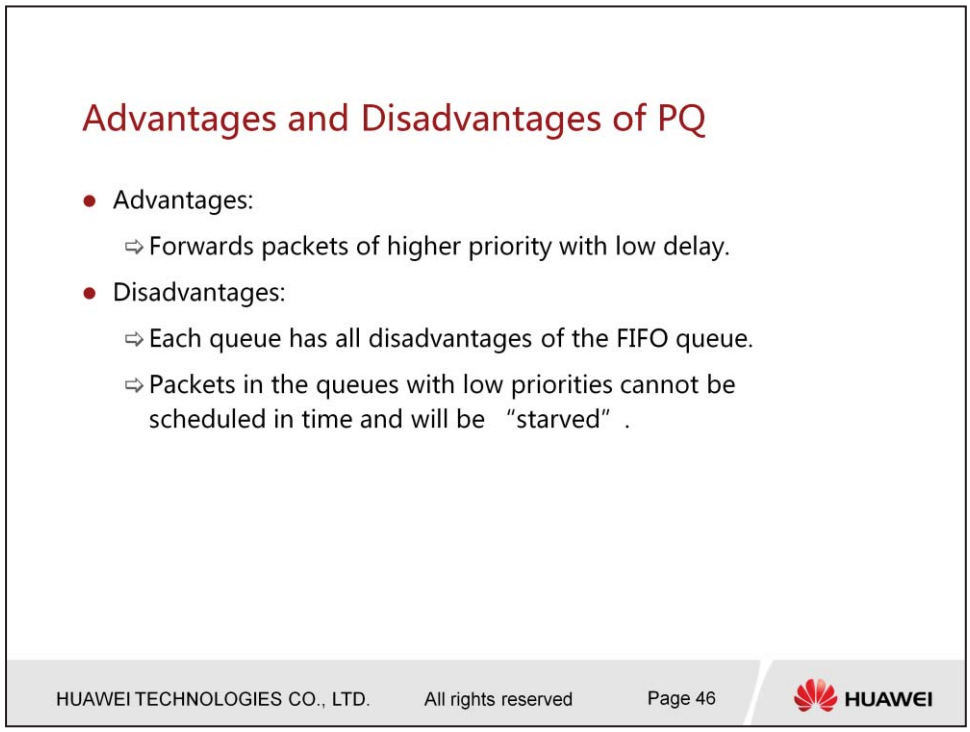
3. Network protocol, such as IPX and CLNS.

4. Packet length.



PQ performs strict priority scheduling, that is, schedules packets in the queue with higher priority first. As shown in the flowchart, the system first checks the Top queue. If the Top queue contains packets, the system schedules these packets, until the top queue is empty. Then the system checks the Middle queue. If the Middle queue contains packets, the system provides service for this queue. The system then checks the Normal and Bottom queues in sequence and provides service for them. The PQ mechanism has a defect, that is, packets in queues with low priories cannot be scheduled in time and will be starved.

Undefined or unidentifiable packets are added to the default queue (normal queue by default, and can be modified).



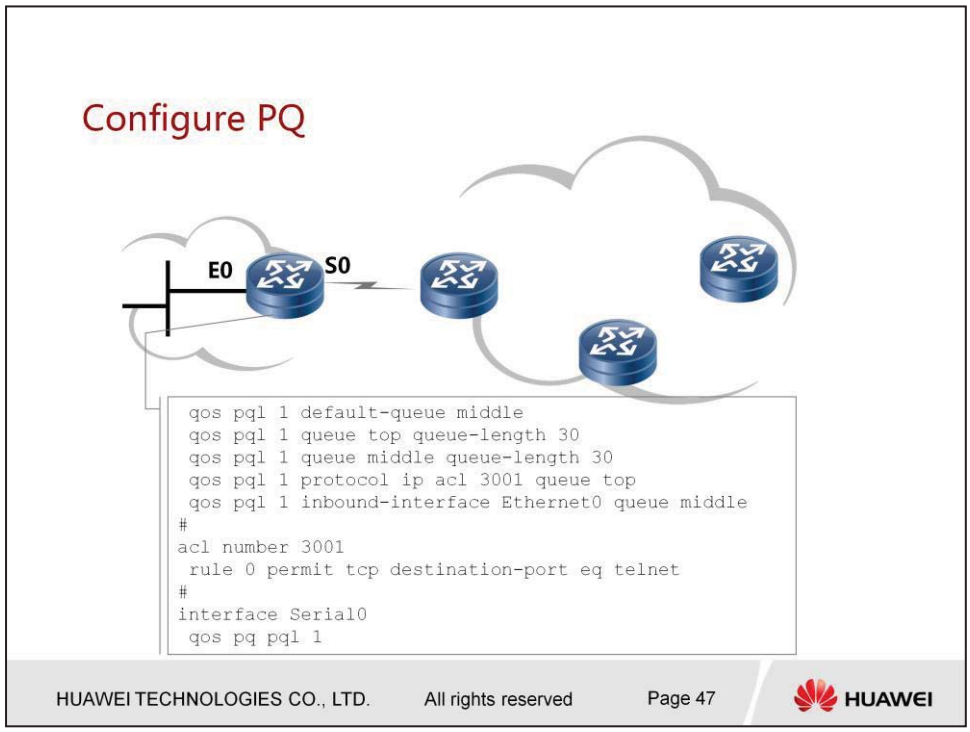
Advantages:

1. Forwards packets with low delay. Packets in a queue can be forwarded only after all the packets in the queues with higher priorities are forwarded. This ensures the higher priority packets with low delay in forwarding.

Disadvantages:

1. All the four queues use FIFO queuing mechanism within the queue, so each queue has all disadvantages of the FIFO queue.

2. If a queue with higher priority contains packets for a long time, packets in the queues with lower priories cannot be scheduled and will be ~~"s~~tarved".



In this example, the telnet traffic is added to the Top queue. The traffic from the interface eth0 is added to the Middle queue. The length of the Top and

Middle queues are both 30. The configuration commands are described as follows:

•Configure the PQ list.

1. Enter the system view:

system-view

2. Configure the PQ list based the network protocol:

qos pql *pql-index* **protocol** *protocol-name* **queue-key** *key-value* **queue** { **top** | **middle** | **normal** | **bottom** }

Or configure the PQ list based on the inbound interface of packets:

qos pql *pql-index* **inbound-interface** *interface-type interface-number* **queue** { **top** | **middle** | **normal** | **bottom**}

The system classifies the packets based on the protocol type or the inbound interface and adds the packets to different queues. By repeating this command with the same pql-index, you can set multiple rules for this PQ list. The system matches packets with the configuration sequence of the rules. If the packet matches a rule, the system stops the matching process.

•Configure the default queue.

1.Enter the system view:

system-view

2. Configure the default queue:

qos pql *pql-index* **default-queue** { **top** | **middle** | **normal** | **bottom** }

You can define multiple rules for a PQ list and then apply the rules to an interface. When a packet arrives at this interface (and will be sent by the interface), the system matches the packet with the rules at the configuration sequence. If a matching rule is find, the packet is added to the corresponding queue and the

matching process is complete. If the packets does not match any rule, it is added to the default queue. Configure a default queue for the packet that does not match any rule. By repeating this command with the same pql-index, you can overwrite the previous default queue. The default queue is normal.

•Set the queue length.

1. Enter the system view: system-view

2. Set the length of each of each queue: qos pql *pql-index* queue { top | middle |

normal | bottom } queue-length *queue-length*

The default length of the queues are: Top 20; Middle 40; Normal 60; Bottom 80.

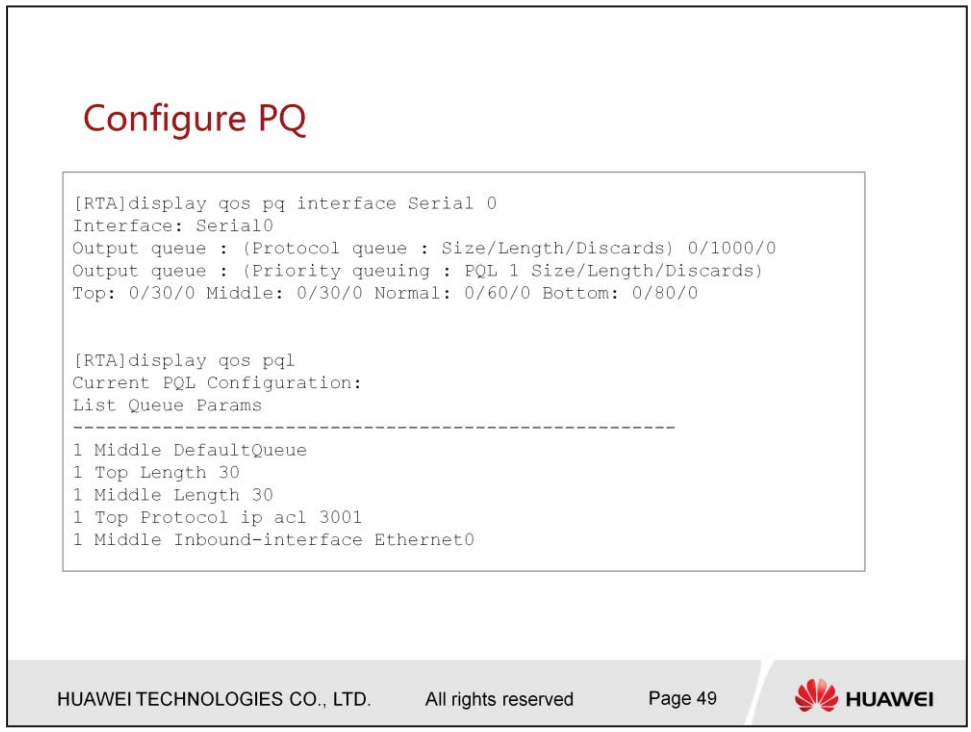
•Apply the PQ list to a interface.

1. Enter the system view: system-view

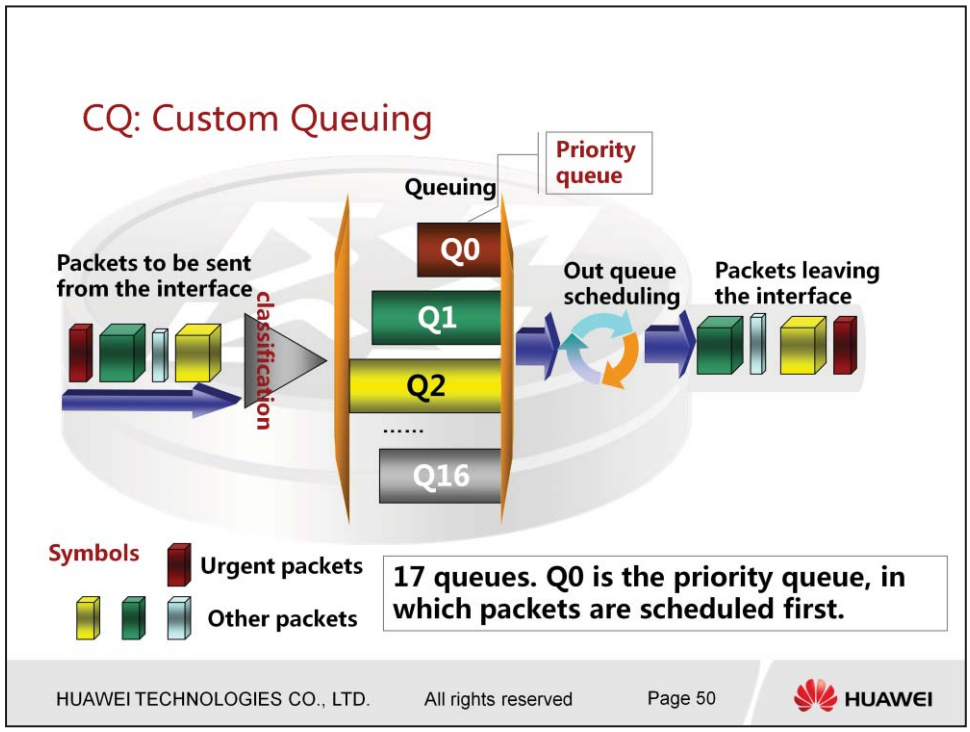
2. Enter the interface view: interface *interface-type interface-number*

3. Apply the PQ list to the interface: qos pq pql *pql-index*

This command applies a PQ list to the interface. You can repeat this command on the same interface to set a new PQ list for this interface.



After the configuration, you can run the **display qos pq interface** command to check the effect of configuration. You can see the length of each queue. You can use the **display qos pql** to view the configured PQ list.



CQ is similar to PQ in terms of traffic classification options and configuration. However, they use completely different scheduling mechanisms. CQ removes the defect of ~~-~~packet starvation" in PQ. CQ defines 17 queues, numbered from 0 to 16. Q0 is the priority queue. Other queues are processed only when Q0 has no packets. Q0 is usually used as the system queue. The bandwidth is allocated to Q1 to Q16 according to the proportion defined by the user. Round

Robin scheduling mode is adopted for packets leaving the queue. Certain bytes of packets are picked up from each queue. Within a queue, the tail drop policy is still used. Similar to PQ, CQ also classifies packets based on the following factors:

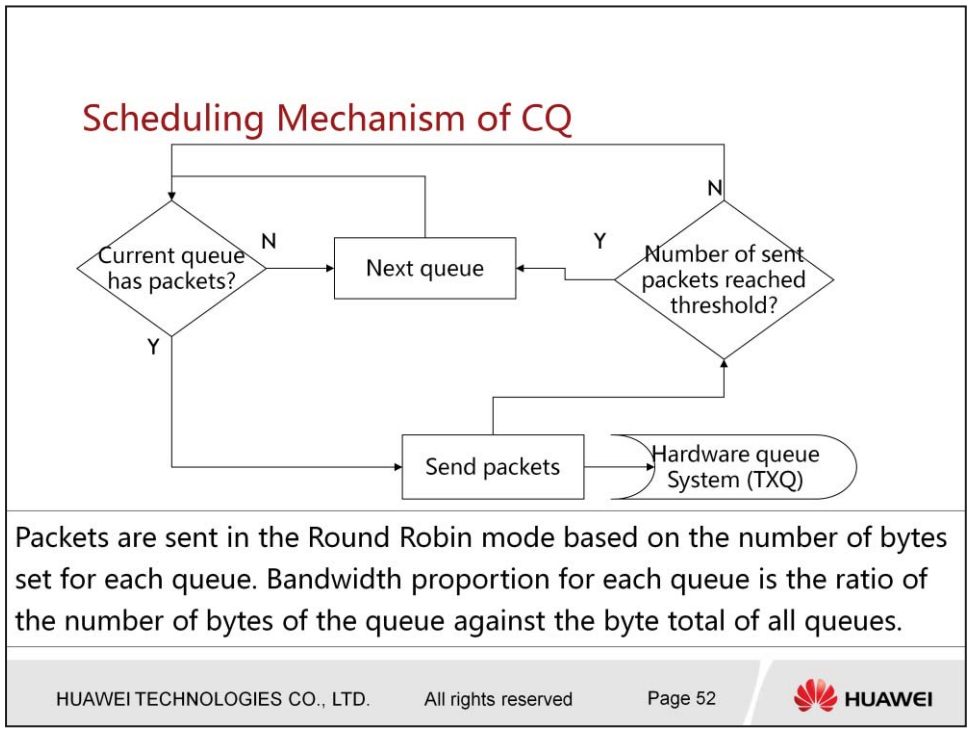
1. Incoming interface of the packets
2. Basic or advanced ACL. ACL can match the following parameters:

- Source IP address
- Destination IP address
- UDP/TCP source port number or port number range
- UDP/TCP destination port number or port number range
- IP precedence, namely the first three bits of the ToS field
- DSCP value, namely the first six bits of the ToS field.
- Packet fragments, which are identified by the fragmentation flag

and offset value in the IP packets

3. Network protocol, such as IPX and CLNS

4. Packet length

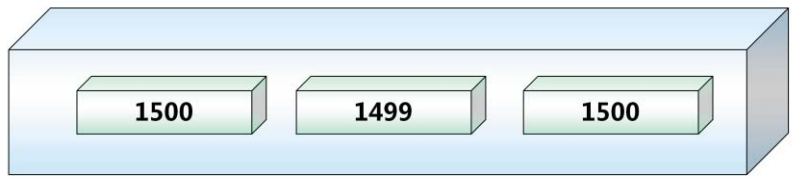


CQ uses the Round Robin scheduling mode. Beginning with queue1, certain number of packets are picked up from each queue. When the number of processed packets reaches the set threshold or there are no packets in the queue, the system processes the next queue in the same way. In CQ mechanism, the number of bytes for each queue is configured instead of the exact bandwidth proportion. You can calculate the link bandwidth for each queue by using the number of bytes each roll for this queue. The formula is:

Number of bytes in the queue/total number of bytes that all queues should have = link bandwidth proportion for this queue

If a queue keeps empty for a period, the bandwidth for this queue is allocated to other queues according to their bandwidth proportions. Assume that five queues are configured. The numbers of bytes for the queues are respectively 5000, 5000, 10000, 10000, and 20000. If all the five queues have enough packets to be sent, their bandwidth is allocated at the proportion of 10%, 10%, 20%, 20%, and 40%. If queue 4 has no packets to be sent for a period, that is, queue 4 is empty, the 20% bandwidth of queue 4 is allocated to the other four queues at the proportion of 1: 1: 2: 4. Therefore, within this period, the four queues occupy respectively 12.5%, 12.5%, 25%, and 50% of total bandwidth.

Disadvantage of CQ Scheduling

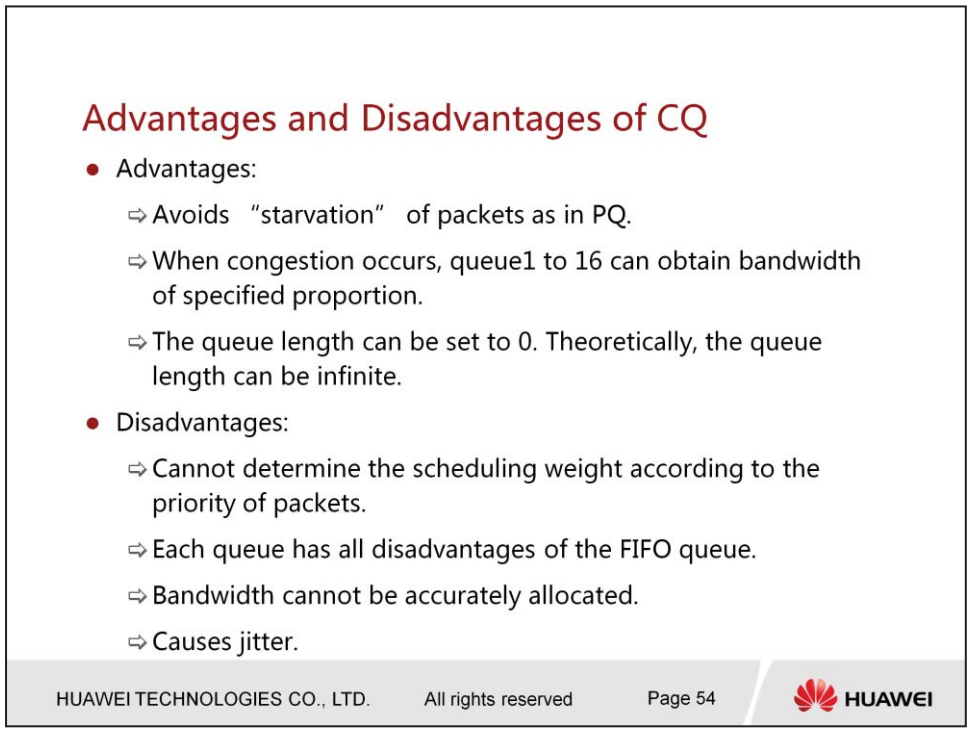


- Example
  - ⇨ The MTU of an interface is 1500 bytes.
  - ⇨ CQ schedules 3000 bytes of each queue at a time.
  - ⇨ Actually, 4499 bytes are scheduled and nearly 50% more bandwidth is occupied by this queue.

Inaccurate bandwidth allocation is the disadvantage of CQ scheduling.

HUAWEI TECHNOLOGIES CO., LTD. All rights reserved HUAWEI

The CQ mechanism cannot fragmentize packets. For example, 3000 bytes need to be picked out from queue1. The following condition may occur: the first packets is 1500bytes, and fetched out, then the second packet is 1499 bytes and only 1 byte should be picked up subsequently. However, the third packet is 1500 bytes, then 1500 + 1499 + 1500 =4499 bytes are picked out from Q1. For this queue, the bandwidth proportion is different from the expected value 3000.



Advantages:

1. Packets of various services can obtain different bandwidth. This ensures more bandwidth for key services and also provides certain bandwidth for non-key services.

That is, it avoids the ~~-s~~tarvation" of packets as in PQ mechanism.

2. When congestion occurs, queue1 to queue16 can obtain bandwidth of specified proportion.

3. The queue length can be set to 0. Theoretically, the queue length can be infinite .

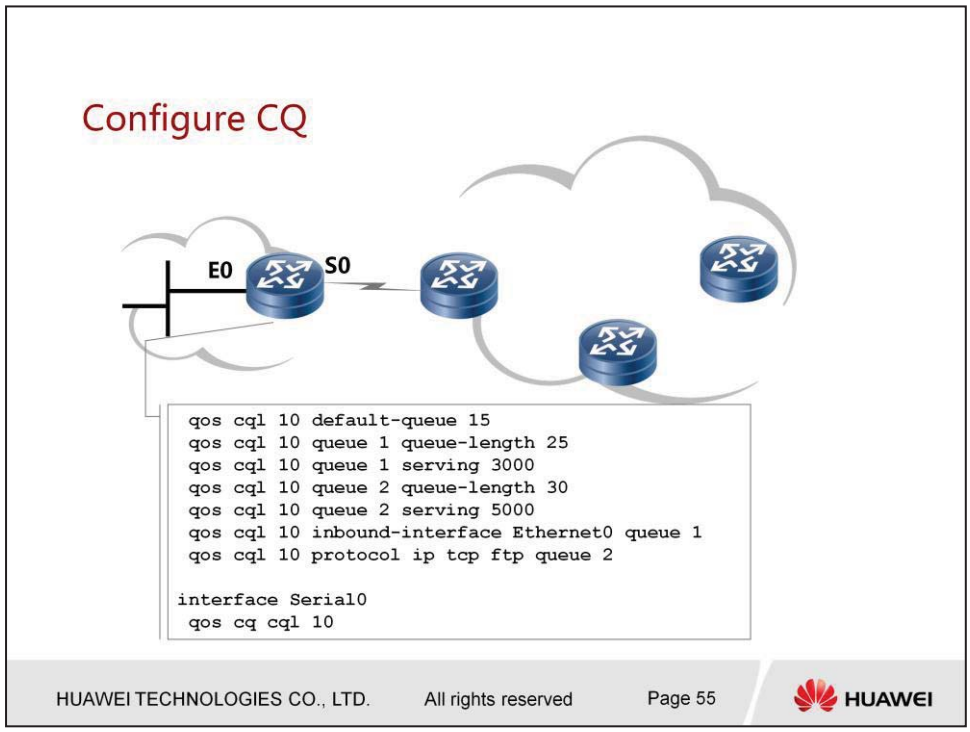
Disadvantages:

1. It cannot determine the scheduling weight according the precedence of packets. Packets with high priority cannot be processed first.

2. Each queue uses FIFO mechanism. That is, each queue has all disadvantages of the FIFO queue.

3. Bandwidth cannot be accurately allocated.

4. It causes jitter, so CQ is a scheduling mechanism applied in a network not having high requirement for jitter.



In the configuration example, length of queue1 is set to 25, and 3000 bytes in queue1 is scheduled each time. Length of queue2 is set to 30, and 5000 bytes in queue1 is scheduled each time. Packets from eth0 is added to queue1 and FTP data is added to queue2. queue15 is configured as the default queue.

The configuration commands are described as follows:

- Configure the CQ list.

1. Enter system view: system-view

2. Configure the CQ list based on the network protocol: qos cql *cql-index* protocol *protocol-name* queue-key *key-value* queue *queue-number* Or configure the CQ list based on the inbound interface of packets: qos cql *cqlindex* inbound-interface *interface-type interface-number* queue *queue-number*

The CQ list consists of 16 groups (1-16). Each group specifies the queues to which certain types of packets should be added, length of each queue, and number of bytes scheduled each time. Only one group can be applied to an interface.

Create the classification rule based on the inbound interface or the features of the packets. By repeating this command with the same cql-index, you can add new rules to this CQ list.

•Configure the default queue.

1. Enter the system view: system-view

2. Configure the default queue: qos cql *cql-index* default-queue *queue-number*

Configure a default queue for the packets that do not match any rule. You can define multiple rules for a CQ list and then apply the rules to an interface. When a packet arrives at this interface (and will be sent by the interface), the system matches the packet with the rules at the configuration sequence. If a matching rule is find, the packet is added to the corresponding queue and the matching process is complete. If the packets does not match any rule, they are added to the default queue.

By default, the default queue is queue1.

•Set the queue length.

1. Enter the system view: system-view

2. Set the queue length: qos cql *cql-index* queue *queue-number* queue-length

queue-length

This commands sets the length of a specified custom queue (namely the number of packets in the queue).

queue-length specifies the maximum length of a queue. The default value is 20.

•Set the number of bytes scheduled each time in a queue.

1. Enter the system view: system-view

2. Set the number of bytes scheduled each time in a queue: qos cql *cql-index* queue *queue-number* serving *byte-count*

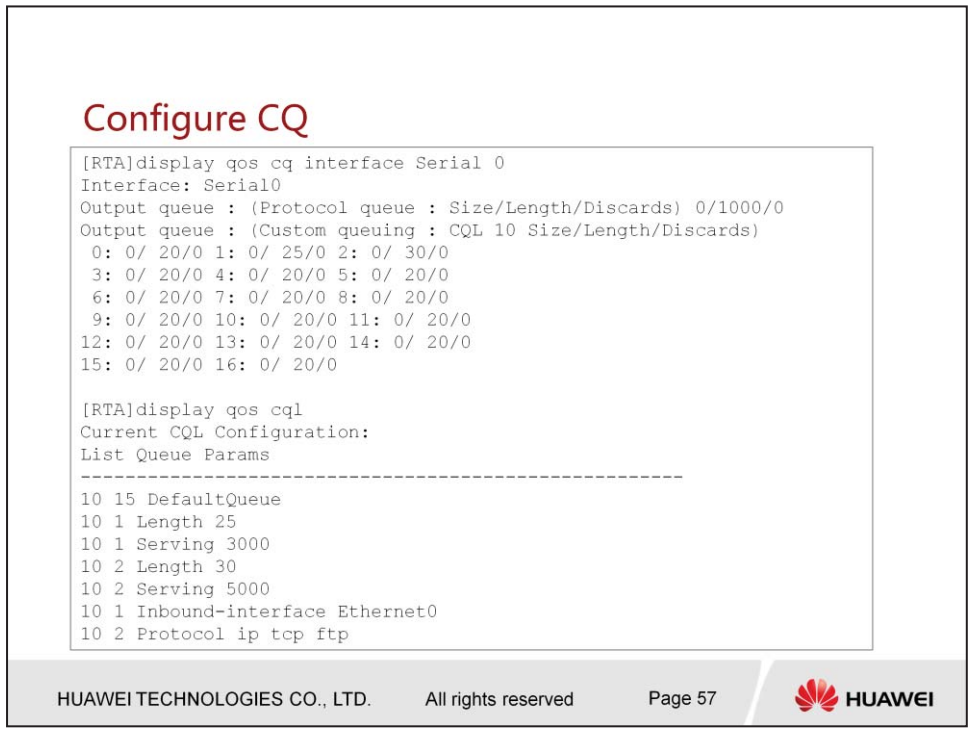
Byte-count specifies the number of bytes scheduled each time. When the router schedules the custom queues, it sends packets in a queue continuously, until the number of sent bytes reaches or exceeds the byte-count value of this queue or the queue is empty. Then the router begins to schedule the next queue. Therefore, the value of byte-count influences the bandwidth proportion of the queues. The value also determines the interval for the router to schedule the next queue. The default value of byte-count is 1500. The packets in the system queue should be sent first. When the system queue becomes empty, the router sends certain number of packets from queue1 to queue16 according to their bandwidth proportion.

•Apply the CQ list to the interface.

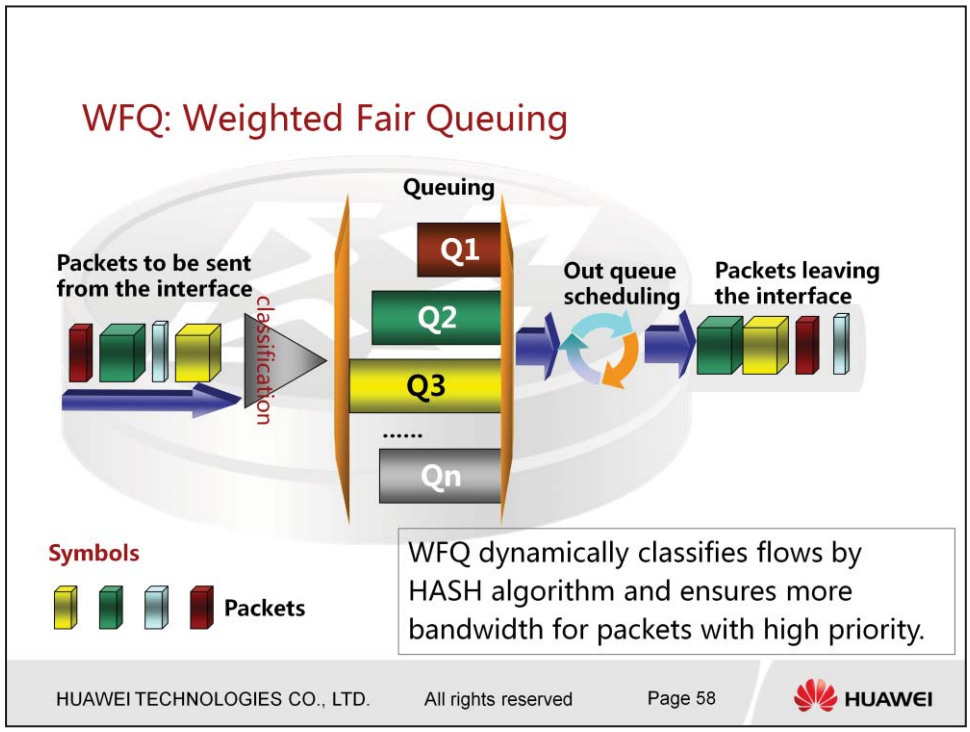
1. Enter the system view: system-view

2. Enter the interface view: interface interface-type *interface-number*

3. Apply the CQ list to the interface : qos cq cql *cql-index*



After the configuration, you can run command **display qos cq interface** to check the effect of configuration. You can see the information about all the 17 queues. You can use the **display qos cql** command to view the information for the configured CQ list.



The most obvious difference between WFQ and PQ or CQ is that WFQ does not allow packet classification based on the ACL, in stead, WFQ dynamically classifies packets based on flows. A flow is identified by the quintuple (source IP address, destination IP address, protocol number, source port number, and

destination port number) of packets. Packets with the same quintuple belong to the same flow, which is mapped to a queue through the Hash algorithm. In some cases, the ToS field is also used. The rigid classification method has some defects and needs to be optimized by another mechanism. In WFQ, the flows with lower volume and higher precedence are processed earlier than flows with larger volume and lower precedence. Because WFQ is based on flows and each flow maps a queue, WFQ must support a large number of queues. WFQ supports a maximum of 4096 queues on each interface. Differences between WFQ and CQ are:

1. CQ can define ACL rules to classify packets, while WFQ can only use the quintuple to classify packets.

2. Their queue scheduling mechanisms are different. Scheduling mechanism of CQ is Preemptive + WRR, while scheduling mechanism of WFQ is weighted fair queuing.

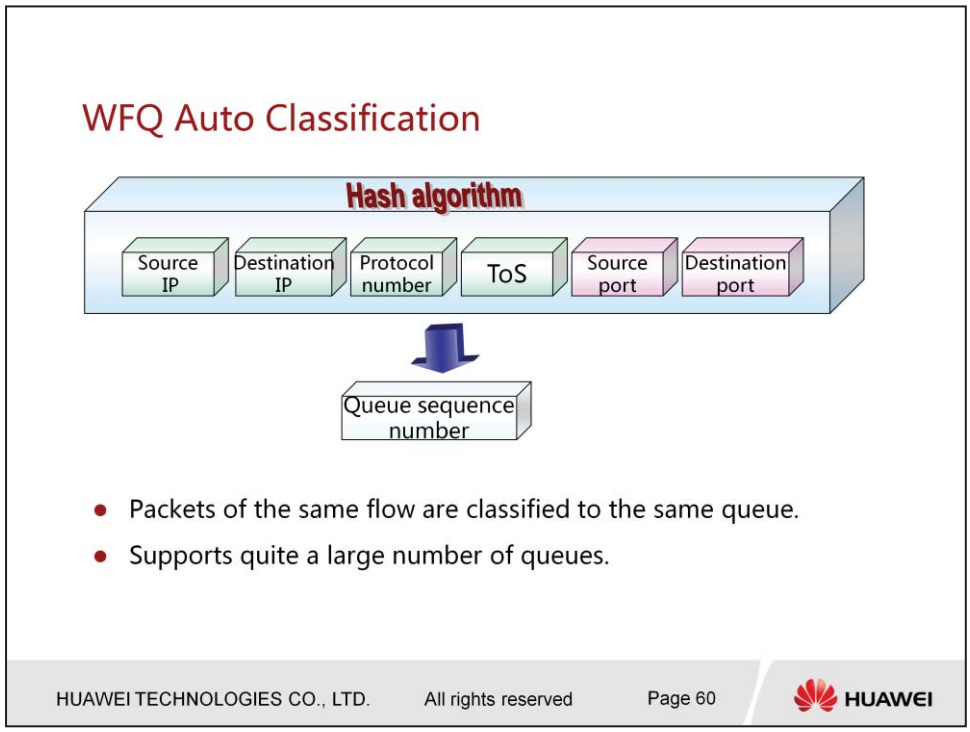
3. Their drop policies are different. CQ uses the tail drop policy, while WFQ uses WFQ drop policy, which is an improvement to the tail drop policy.

4. WFQ is based on flows. Each flow occupy a queue and each interface supports a maximum of 4096 queues. WFQ scheduling has two objectives. One is providing fair scheduling for flows. This is the meaning of F (fair) in WFQ. The other is guaranteeing more bandwidth

for flows with high precedence. This is the meaning of W (weighted). To provide fair scheduling for flows, WFQ provides the same bandwidth for each flow. For example, if there are 10 flows on an interface and bandwidth of this interface is 128 Kbps, then the bandwidth for each flow is 128/10 = 12.8 Kbps. In a sense, this mechanism is similar to time division multiplexing. WFQ allows other flows to use the remaining bandwidth of a flow. If the bandwidth of an interface is 128 kbps and there are 10 flows on the interface, then each flow has the bandwidth of 12.8kbps. It is possible that flow1 needs only 5 kbps and flow2 needs 20 kbps. Flow2 can use the remaining bandwidth of flow1:12.8 – 5 = 7.8 kbps.

The weighting of WFQ is based on the IP precedence of flows. WFQ allocates more bandwidth to flows with high IP precedence. The algorithm is (IP precedence+1)/Sum (IP precedence+1). Four example, four flows have the IP precedence 1, 2, 3, 4 respectively. The bandwidth for these flows should be

respectively 2/14, 3/14, 4/14, 5/14.

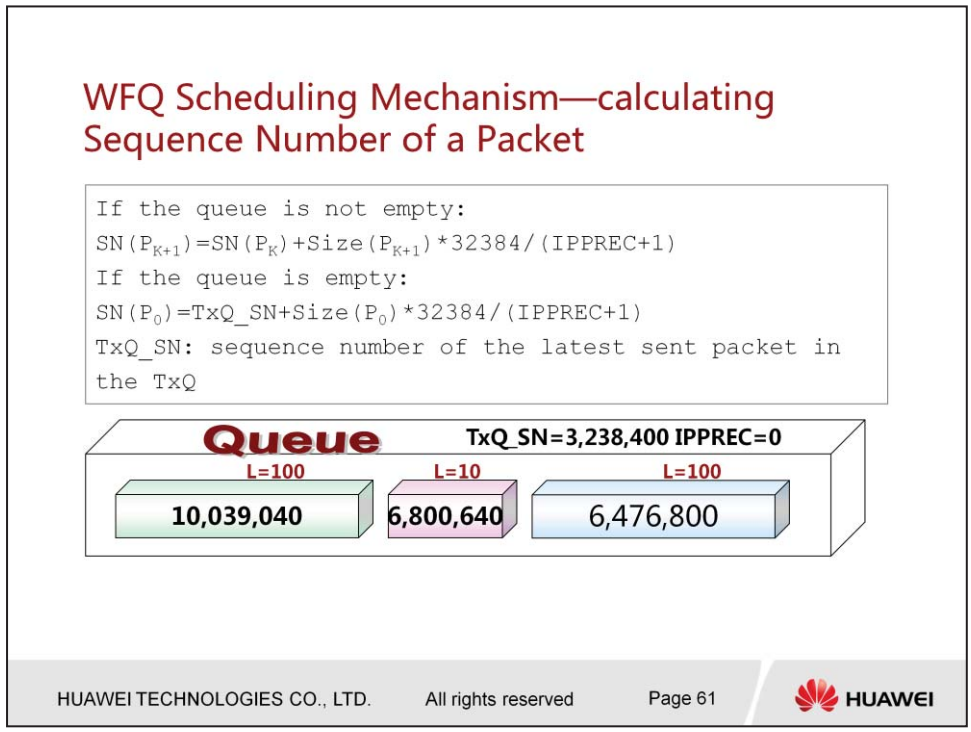


Packets with the same quintuple (source IP address, destination IP address, protocol number, source port number, and destination port number) belong to the same flow.

WFQ uses the quintuple and ToS field as the input of the Hash algorithm. The calculation result is a number, namely the sequence number of the queue to which the packet should join.

Packets in the same flow should be classified to the same queue. However, if the ToS field of a packet is different from the ToS field of other packets, this packets is added to another queue. This causes disorder of packets. To avoid disorder of packets, you should set the same ToS field for the packets in the same flow. By default, WFQ uses 256 queues. The number of queues can be set to a value ranging from 16 to 4096, but this value must be the power of 2 (2n). In addition, eight system queues are configured for system packets and up to 1000 RSVP queues can be configured. System packets and RSVP packets are mapped to specific queues for them.

To ensure that each flow is mapped to a queue, WFQ requires a large number of queues. If there are not enough queues, different flows may be mapped to the same queue.



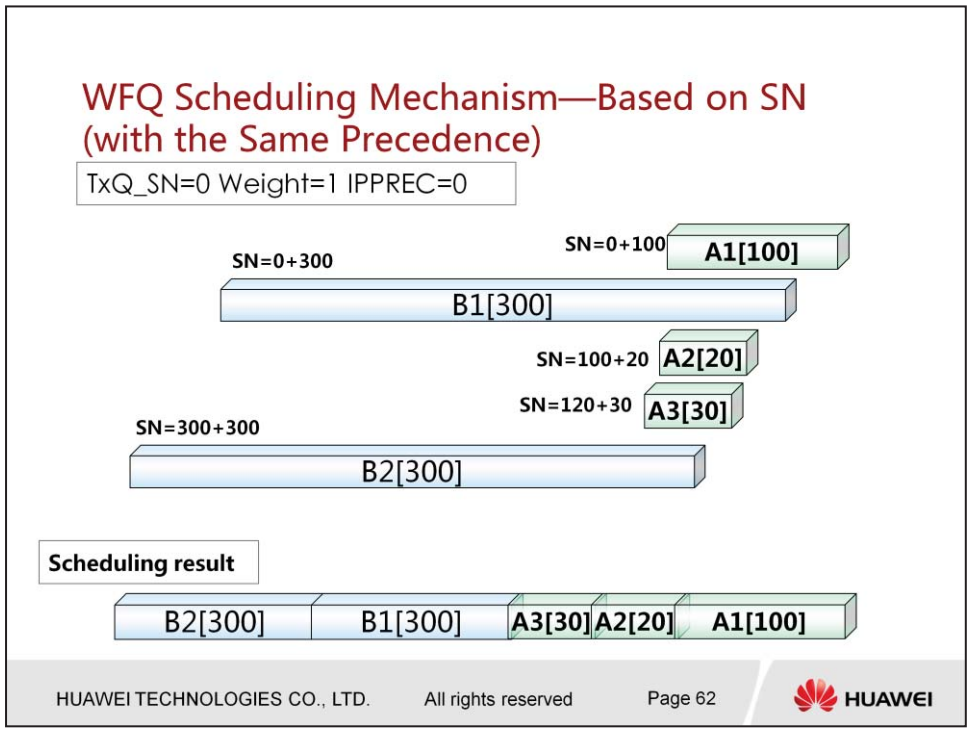
To understand the drop policy and queue scheduling mechanism of WFQ, you need to know an important concept: sequence number (SN). (Other parameters may be used in other documents.

No matter what parameter is used,
the flows with lower volume and higher IP precedence should be scheduled first). After classification, each packet is assigned

an SN before it is added to a queue or dropped.

The formula for calculating the SN is: SN = Previous_SN+weight*new_packet_length. The Previous_SN varies in the following cases:
If the queue to which the packet should be added is not empty, Previous_SN is the SN of the last packet that joins the queue. If the queue is empty, Previous_SN is the SN of the last packet that is sent by the TxQ. In WFQ scheduling, the packet with smaller SN is scheduled first. From the formula for calculating the SN, you can see that the weight is in reverse proportion to the precedence so that flows with higher IP precedence can obtain more bandwidth. Generally, the value of weight is 32384/(IPPREC+1). In the figure, when the first packets of 100 bytes joins the queue, the queue is empty. The SN of this packet is: 3238400+100*32384/(0+1)=6476800. When the second packet of 10 bytes joins the queue, the first packet is not sent yet, so the queue is not empty. The SN of the second packet is:

6476800+10*32384/(0+1)=6800640.
Similarly, the SN of the third packet (100 bytes) is 10039040.



Assume two queues A and B. Packets joining A and B all have the IP precedence 0. According to the formula: weight=32384/(IPPREC+1), the value of weight is 32384. For the convenience of description, we set the weight to 1. At the beginning, TxQ_SN is 0.

1. Packet A1 joins queue A, and its length is 100 bytes. Queue A is empty, so SN(P0)=TxQ_SN+Size(P0)*=100.

2. Packet B1 joins queue B and its length is 300 bytes. Queue B is empty, so SN(P0)=TxQ_SN+Size(P0)*Weight=300.

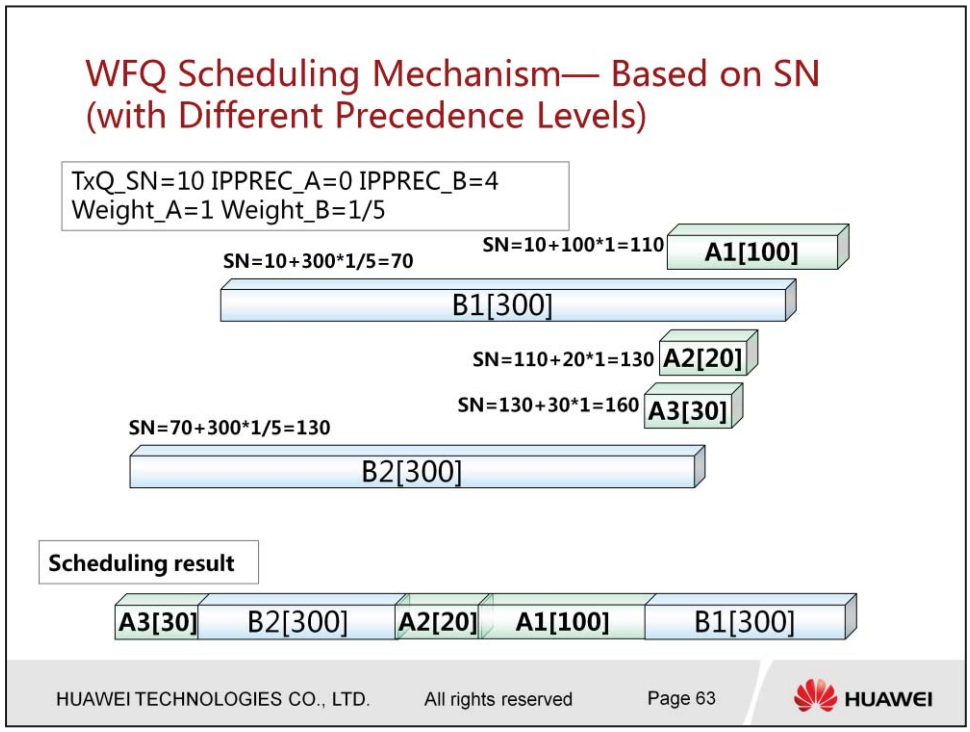
3. Packet A2 joins queue A. Now A1 is not sent yet, so queue A is not empty. SN(PK+1)=SN(PK)+Size(PK+1)*Weight=120.

4. Packet A3 joins queue A. SN=150.

5. Packet B2 joins queue B. Now queue B is not empty, so SN=600.

6. By comparing the SNs, you can obtain the scheduling sequence: A1, A2, A3, B1, B2.

In this example, the IP precedence is not taken into account, so queue A and queue B have the same weight. The scheduling sequence relates to the packet Length.



Assume two queues A and B. The IP precedence of packets joining queue A is 0. The IP precedence of packets joining queue B is 4. According to the formula weight=32384/(IPPREC+1), the weight of queue B is 1/5 of weight of queue A. For the convenience of description, we set the weight of queue A to 1 and weight of queue B to 1/5.
At the beginning, TxQ_SN is 10.

1. Packet A1 joins queue A, and its length is 100 bytes. Queue A is empty, so SN(P0)=TxQ_SN+Size(P0)*Weight=110.

2. Packet B1 joins queue B and its length is 300 bytes. Queue B is empty, so SN(P0)=TxQ_SN+Size(P0)*Weight=70.

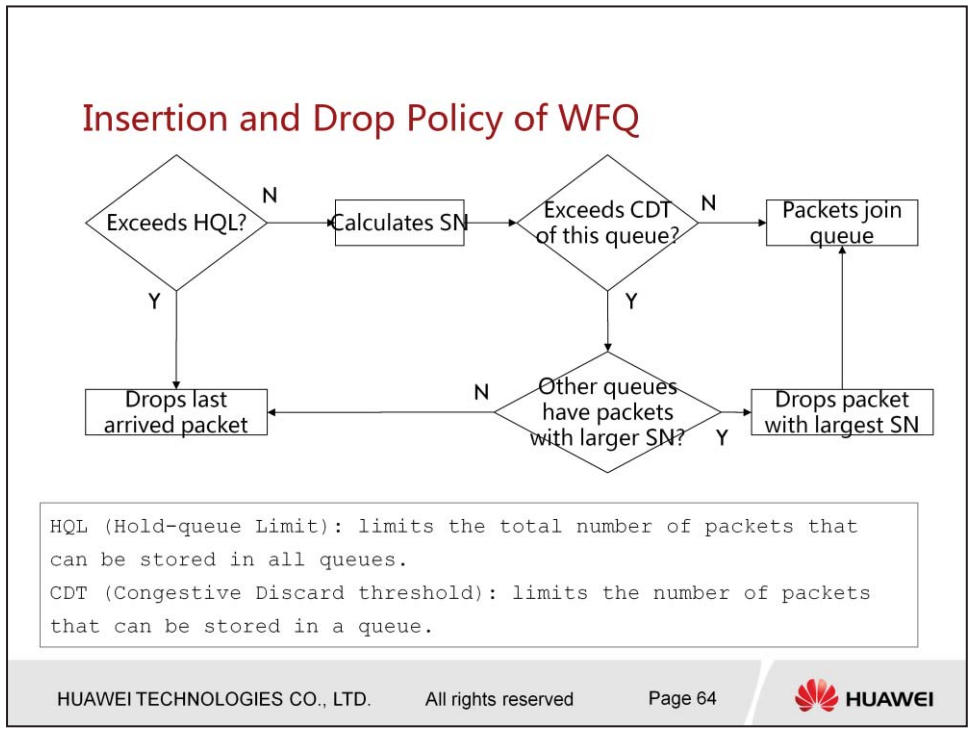
3. Packet A2 joins queue A. Now A1 is not sent yet, so queue A is not empty. SN(PK+1)=SN(PK)+Size(PK+1)*Weight=130.

4. Packet A3 joins queue A. SN=160.

5. Packet B2 joins queue B. Now queue B is not empty, so SN=130.

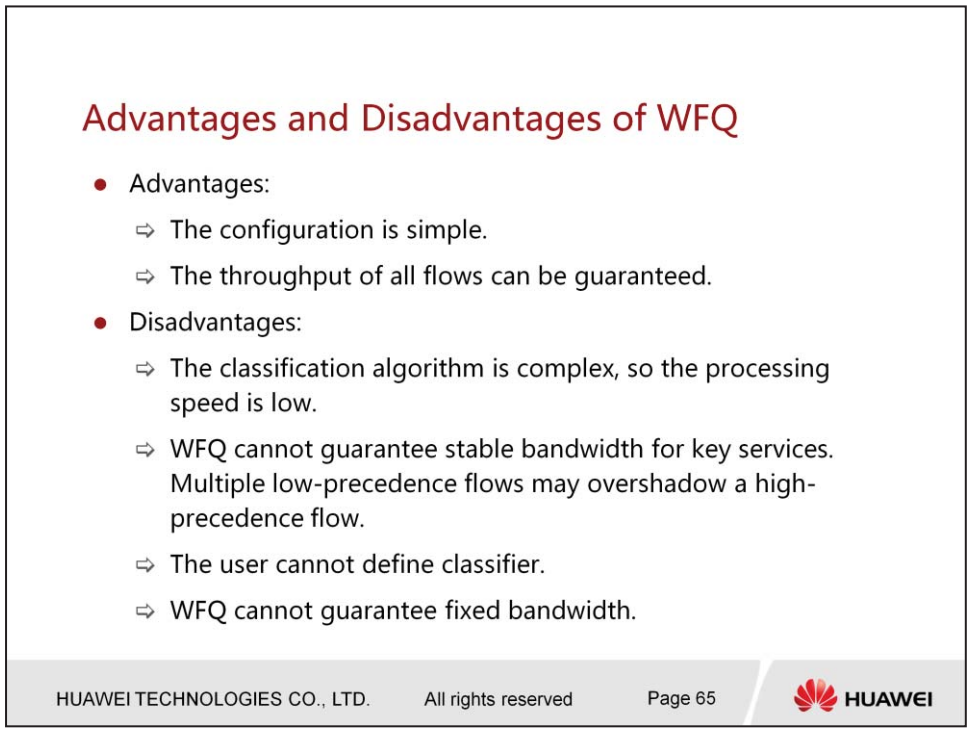
6 . By comparing the SNs, you can obtain the scheduling sequence: B1, A1, A2, B2, A3.

In this example, the IP precedence is taken into account, so the scheduling sequence relates to both the packet length and IP precedence.



WFQ uses WFQ drop policy, which is an improvement to the tail drop policy. Apart from the SN, WFQ also uses Hold-queue-Limit (HQL) and Congestive Discard threshold (CDT) to determine the packets to be dropped. If the HQL reaches the maximum value when a packet arrives, the packet is dropped directly.

If the HQL does not reach the maximum value, WFQ classifies this packet according to the classification rule, determines which queue should the packet join, and calculate the SN. The drop policy is also determined by the CDT, which is the drop threshold of queue. If the CDT does not reach the maximum value, the packet joins the queue. If the CDT reaches the maximum value, the system judges whether other queues have packets with larger SNs. If no packet with larger SN exists, this packet is dropped. If other queues have packets with larger SNs, WFQ drops the packet with the largest SN. In brief, if the number of packets in a queue is larger than the CDT of this queue, WFQ drops the packet with the largest SN among all the queues.



WFQ has the following advantages and disadvantages:

Advantages:

• The configuration is simple.

• The throughput of all flows can be guaranteed

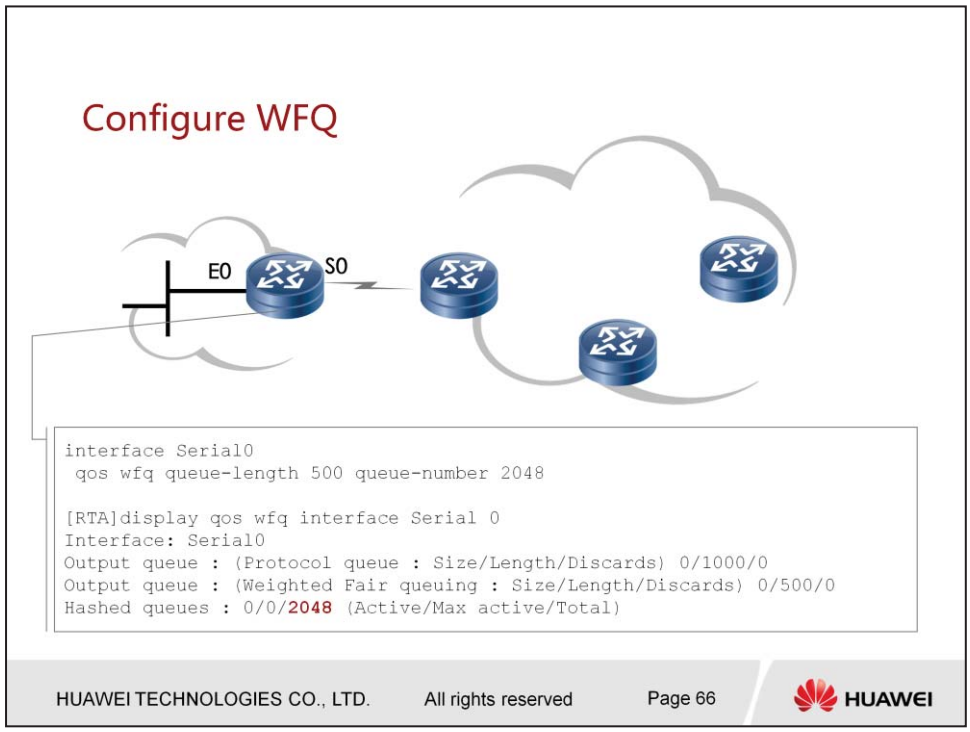
Disadvantages:

• The classification algorithm is complex, so the processing speed is low.

• WFQ cannot guarantee stable bandwidth for key services. Multiple low-precedence flows may overshadow a high-precedence flow.

• The user cannot define classifier.

• WFQ cannot guarantee fixed bandwidth.



In this example, WFQ is configured on interface serial0. The queue length is 500 packets. The interface has 2048 queues.

The configuration commands are described as follows:

•Configure WFQ.

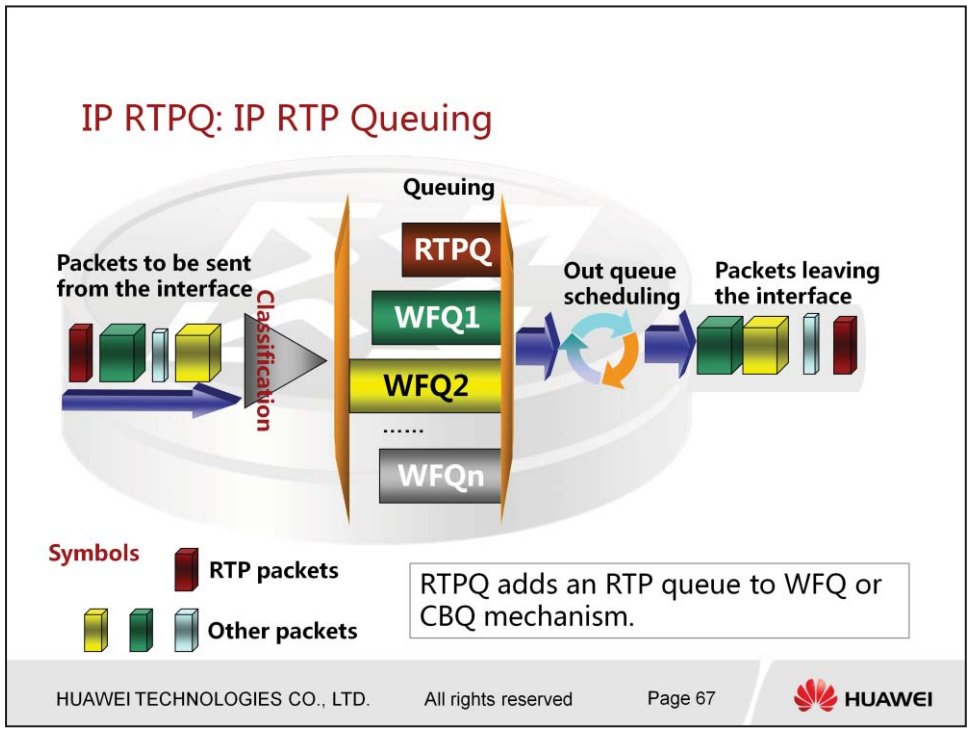
1. Enter the system view: system-view

2. Enter the interface view: interface *interface-type interface-number*

3. Configure WFQ on the interface: qos wfq [queue-length *max-queue-length*

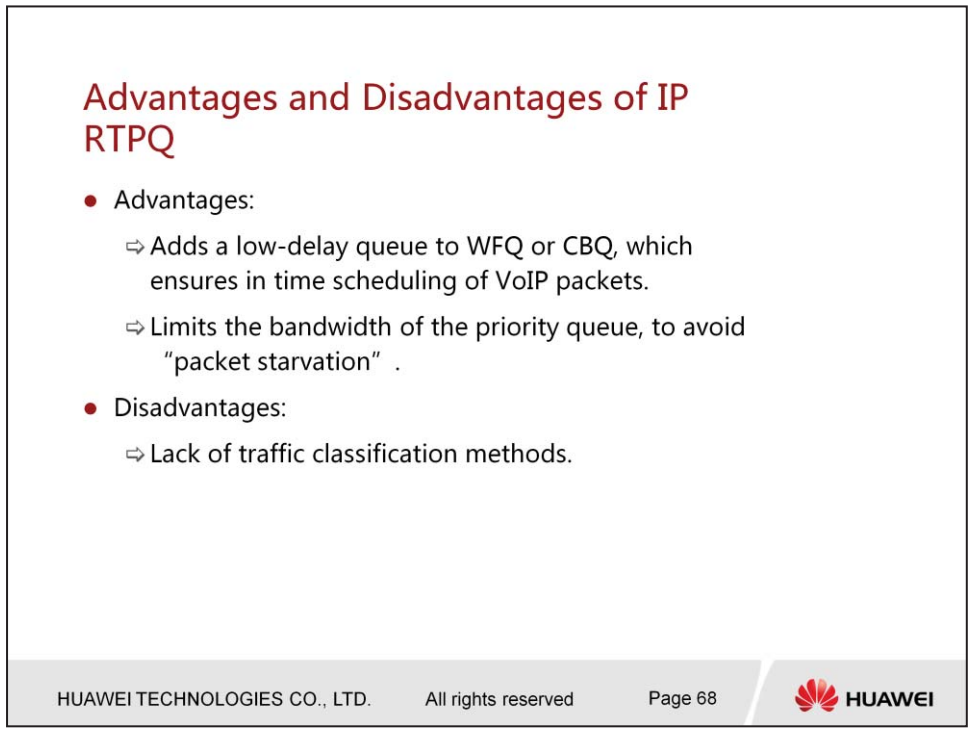
[queue-number *total-queue-number*]]

If no WFQ policy is applied to an interface, you can use this command to apply the WFQ policy to this interface and set WFQ parameters. If a WFQ policy is applied to this interface, you can use this command to change the WFQ parameters.

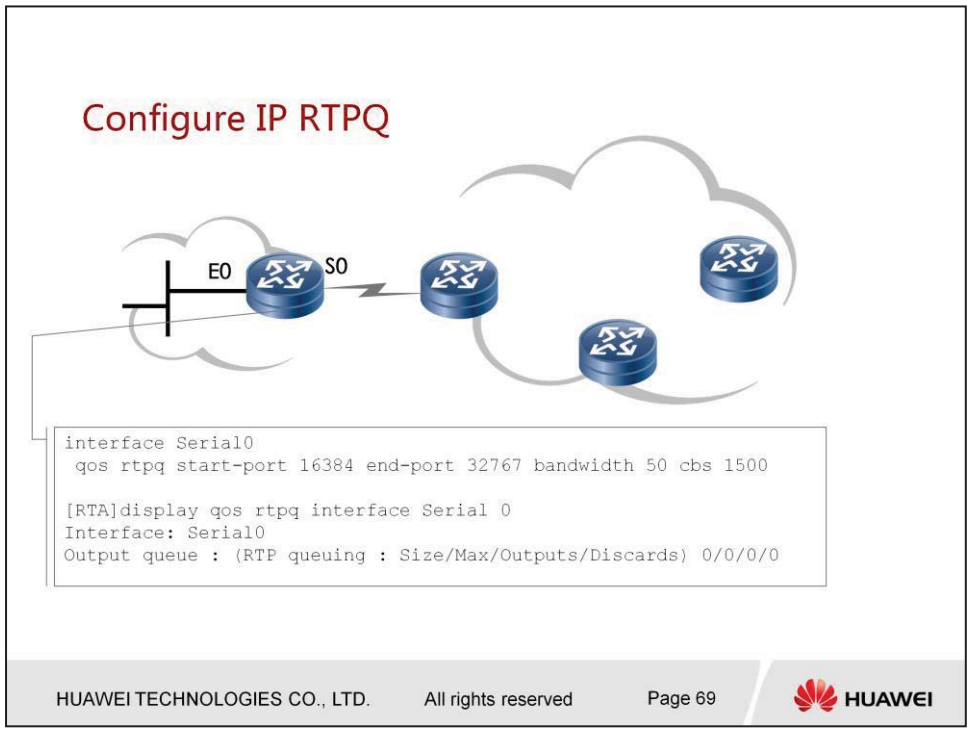


IP RTP queuing adds a strict priority queue to WFQ or CBQ. IP RTP queuing classifies VoIP packets by the destination port number of the UDP packet. It then selects the flows with the UDP destination port number being an even in a certain range. IP RTP queue is a strict priority queue, so it is scheduled first. Certain policy is adopted to prevent the strict priority queue from occupying too much bandwidth. That is, the bandwidth for the RTP queue is limited.

Traffic that exceeds the limit is dropped. RTP queue can be used with any other queues (FIFO, PQ, CQ, WFQ, or CBQ) and it has the highest precedence. However, LLQ of CBQ can also process the real-time service, so it is not recommended to use the RTP queue with the CBQ.



RTPQ can only classify packets according to the port number of RTP packets, so the traffic classification method is not flexible.



In the example, RTPQ is applied on the serial0 interface. Packets with the destination port number being an even in the range of 16384-32767 are added to the RTP queue. The bandwidth for the RTP queue is 50 kbps and the CBS is 1500.

The configuration commands are as follows:

•Apply RTPQ to the interface.

1. Enter the system view: system-view

2. Enter the interface view: interface *interface-type interface-number*

3. Apply RTPQ to the interface: qos rtpq start-port *min-rtp-port-number* end-port *max-rtp-port-number* bandwidth *bandwidth* [cbs *cbs*]

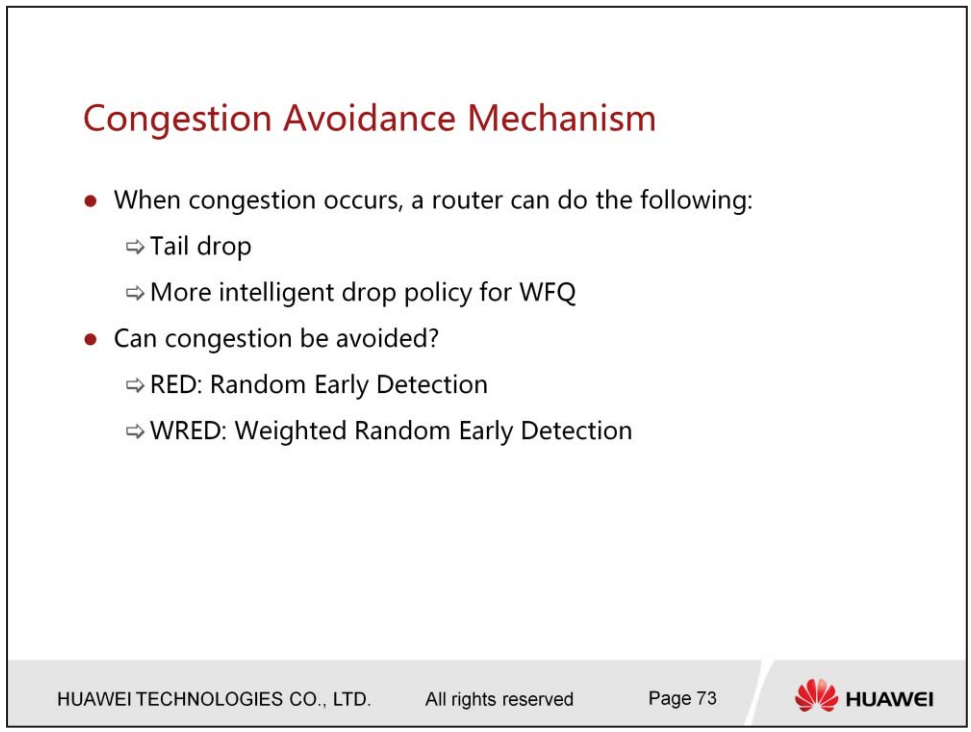
The **qos rtpq** command is used to enable RTPQ on an interface. The objective is to reserve a real-time service for the RTP packets with the UDP destination port number in a certain range. The **undo qos rtpq** command is used to disable RTPQ on the interface. This command creates a strict priority queue. Only the UDP packets with the destination port number being an even in the range of start-port to end-port.

By default, RTPQ is disabled.

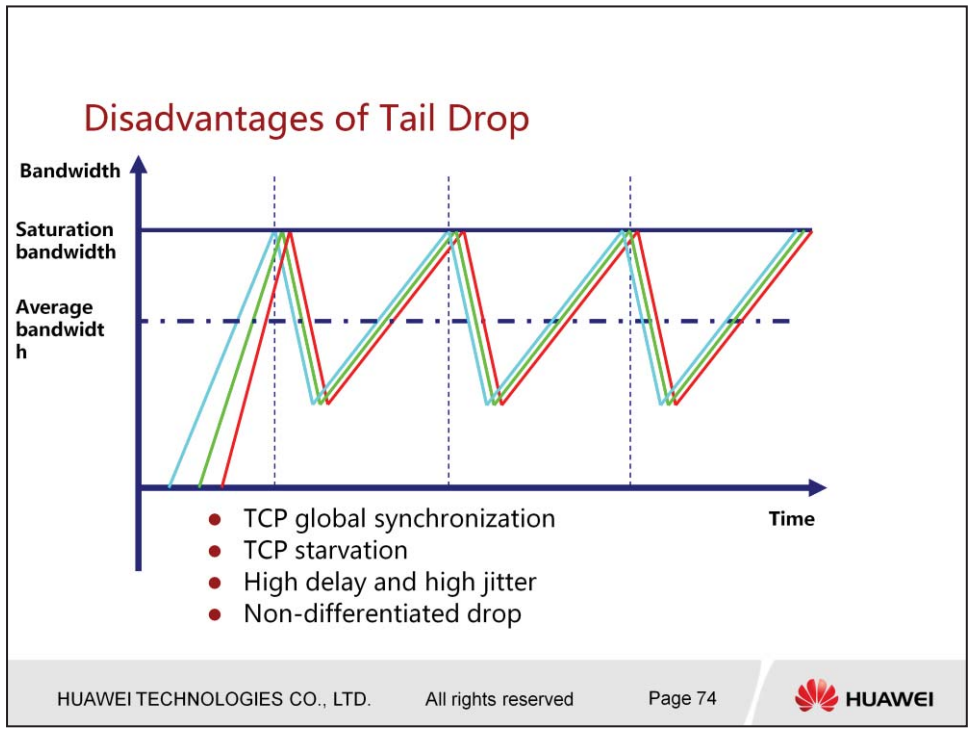
This command is applicable to the delay-sensitive applications, for example, real-time voice transmission. The **qos rtpq** command ensures that the voice service is served first. The *bandwidth* parameter should be larger than the total bandwidth for the real-time service to prevent the compact of burst traffic. However, the bandwidth for the real-time protocol (RTP) queue cannot exceed 75% of the valid bandwidth. To set the bandwidth larger than 75% of the valid bandwidth, run the **qos reservedbandwidth** command to change the default value of the maximum reserved bandwidth. In addition, 25% of

bandwidth must be reserved for the layer-2 packet header. The reason is that the IP header, UDP header, and RTP header are taken into account in bandwidth allocation, but the layer-2 frame header is not considered.


Contents
Classification and Marking
Policing and Shaping
Congestion Management
Congestion Avoidance Mechanisms
Link Efficiency Mechanisms
HUAWEI TECHNOLOGIES CO., LTD.
All rights reserved
Page 72
HUAWEI



When congestion occurs, the traditional drop policy (tail drop) is adopted. When the queue length reaches the maximum value, the new packets are dropped. If WFQ is configured, WFQ drop policy can be adopted. Too severe congestion greatly damages the network resource and must be eliminated by some measures. Congestion avoidance here means to monitor the usage of network resource (such as queues or memory buffer) and drops packets when the c congestion tends to worsen. It is a traffic control mechanism that eliminates network overload by adjusting the traffic. Congestion avoidance methods available now are Random Early Detection (RED) and Weighted Random Early Detection (WRED).



1. TCP global synchronization

The traditional method to react to the congestion is tail drop. When the queue length reaches the specified maximum value, all new packets are dropped. If large number of TCP packets are dropped, TCP timeout occurs. This initiates slow startup of TCP and congestion avoidance mechanism to reduce packets to be sent. When the queue drops all the new arrival packets of TCP sessions, slow startup and congestion avoidance mechanism are initiated for all the TCP sessions. This is called the TCP global synchronization. In this case, all these TCP sessions reduce the packets sent to the queue. Thus, the number of packets sent to the queue is less than the packet sending speed on the interface, and the link bandwidth utilization is reduced. In addition, the traffic sent to the queue is not stable, so the traffic on the link fluctuates between the lowest value and the saturation value. Tail drop also increase the delay and jitter of specific traffic.

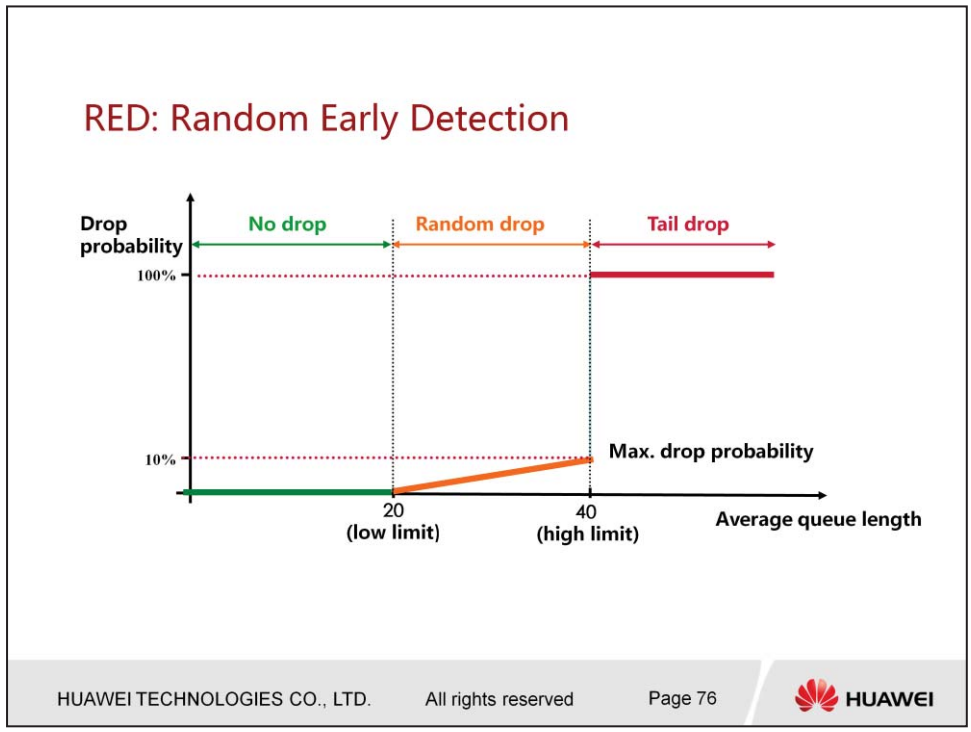
2. TCP starvation

Tail drop causes uneven bandwidth allocation among TCP flows. Some "greedy" flow may occupy most bandwidth, while common TCP flows cannot obtain bandwidth and are starved. The situation is still worse when both TCP flows and UDP flows exist in the network. Because of the sliding window mechanism, TCP flows release the bandwidth (because tail drop reduces the window). UDP does not use the sliding window mechanism, so the UDP flows quickly occupy the bandwidth released by the TCP flows.

The result is that the UDP flows occupy all the bandwidth, and TCP flows cannot obtain bandwidth and are starved.

3. The congestion increases the delay and jitter.

4. It is a non-differentiated drop method and does not classify packets based on the precedence.



To avoid the problems caused by tail drop, the system dropped packets before the congestion occurs on an interface. Random Early Detection (RED) is a mechanism to drop packets before congestion.

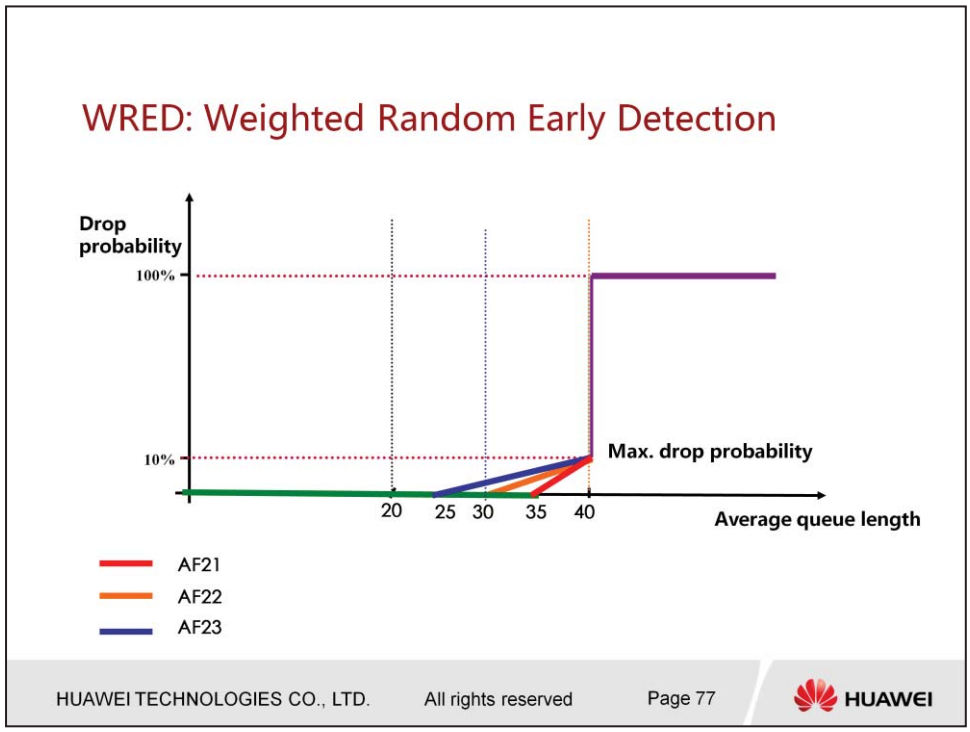
RED drops the packets that may cause congestion. It makes the TCP session release the bandwidth more slowly,

so large scale of TCP global synchronization and TCP starvation is avoided. RED also decreases the average queue length. RED uses three drop behaviors: not dropping green packets, randomly dropping yellow packets according to the drop probability, and dropping red packets. The drop behavior is determined by the low limit and high limit.

1. Green packets—when the average queue length is less than the low limit, the packets are marked green and not dropped.

2. Yellow packets—when the average queue length is between the low limit and high limit, the packets are marked yellow and are dropped according to the drop probability. The longer the queue is, the higher the drop probability will be.

3. Red packets—when the average queue length is larger then the high limit, the packets are marked red and are all dropped (tail drop).

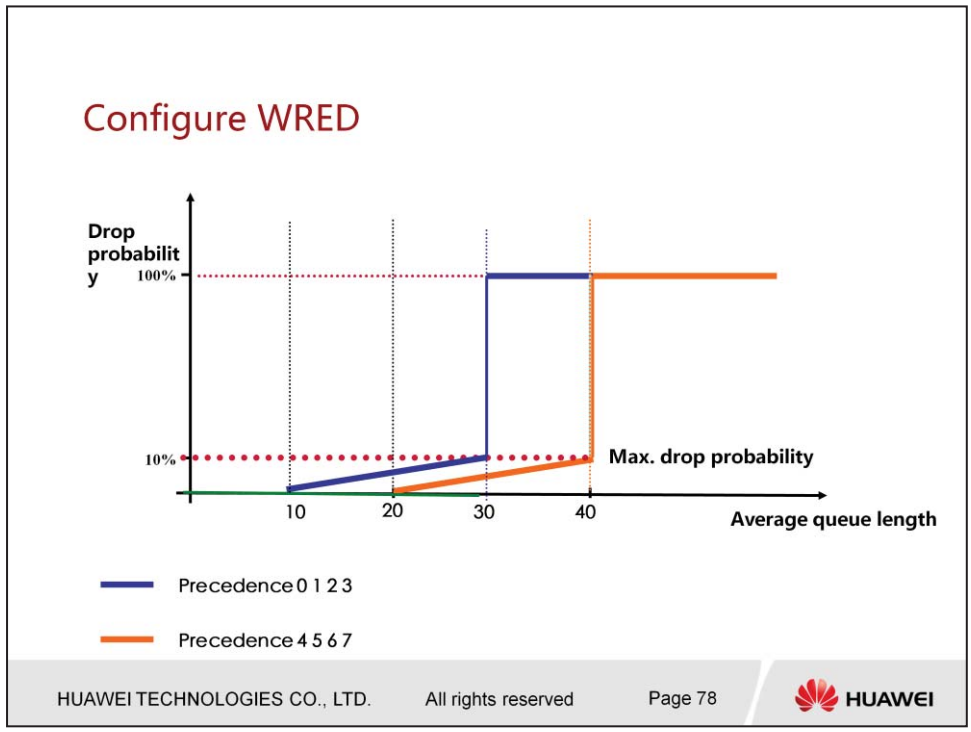


The difference between Weighted Random Early Detection (WRED) and RED is that WRED uses the precedence. Different drop policies are applied for various precedence levels.

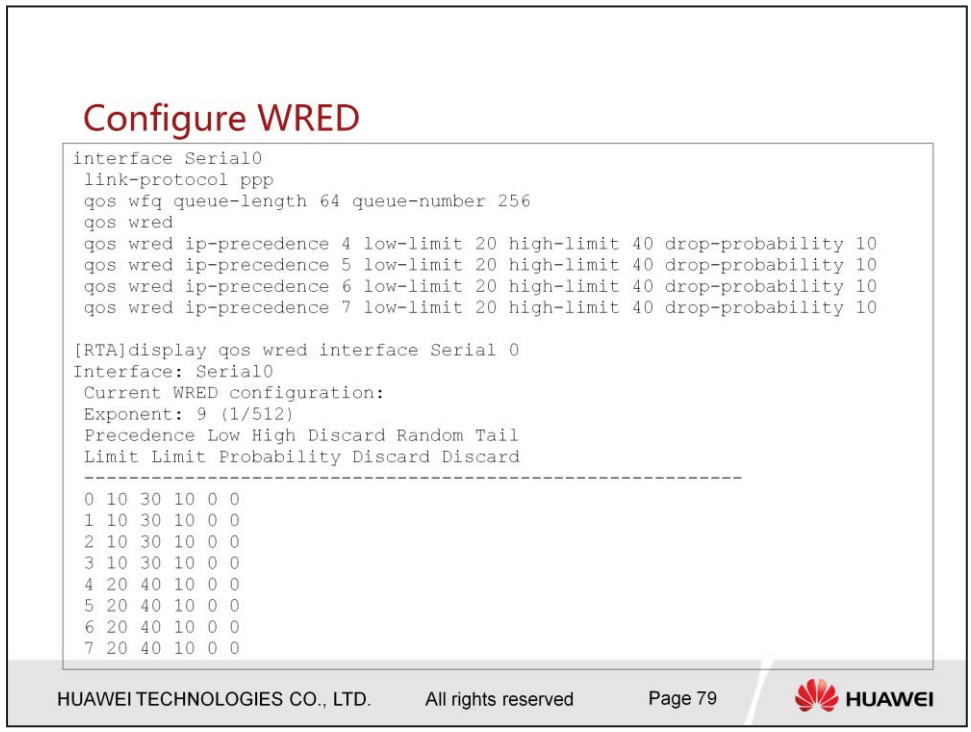
Each drop policy has three RED parameters: low limit, high limit, and maximum drop probability. Currently, WRED precedence is classified based on the DSCP value and IP precedence.

The drop probability for the packets with low precedence is larger than the drop probability for the packets with high precedence.

DSCP AF PHB is expressed as aaadd0. aaa‘ indicates the traffic class; dd‘ indicates the drop probability. For example, AF21(010010), AF22(010100), and AF23(010110) belong to the same class. Their drop probabilities are AF21<AF22<AF23 when congestion occurs. Thus, the WRED parameters can be set as shown in the figure. For the AF21 flow, the low limit is 35 and the high limit is 40. For the AF22 flow, the low limit is 30 and the high limit is 40. For the AF23 flow, the low limit is 25 and the high limit is 40. The drop probability is 10% when the traffic reaches high limit. Therefore, before congestion, packets in the AF23 flow are discarded first.



As shown in the figure, for the flows with precedence 0, 1, 2, 3, the low limit is 10 and the high limit is 30. For the flows with precedence 4, 5, 6, 7, the low limit is 20 and the high limit is 40.



Configure WFQ on the interface. (You need to configure WFQ before configuring WRED.) Set the queue length to 64 and the number of queues to 256. Then configure WRED. Set the low limit of the flows with the precedence 0,1, 2, 3 to 10 and set the high limit to 30 (default values on the VRP).
Set the low limit of the flows with the precedence 4, 5, 6, 7 to 20 and set the high limit to 40.

The configuration commands are described as follows:

•Enable WRED.

1. Enter the system view: system-view

2. Enter the interface view: interface *interface-type interface-number*

3. Enable WRED: qos wred

WRED can only be used with WFQ and CBQ. It cannot be used independently or with other queuing mechanisms. By default, WRED is disabled and the drop policy is tail drop. Set the exponent for calculating the average queue length.

1. Enter the system view: system-view

2. Enter the interface view: interface *interface-type interface-number*

3. Set the exponent for calculating the average queue length: qos wred weighting constant *exponent*

exponent: specifies the exponent for calculating the average queue length. The value ranges from 1 to 16 and the default value is 9. When *exponent* value is larger, the current queue length has greater influence on the average queue value. When *exponent* is 1, the average queue length equals the current queue length.

The command **qos wred weighting-constant** is used to set the

exponent for calculating the average queue length in WRED. The command **undo qos wred weighting-constant** is used to restore the default exponent.

You must apply WRED on the interface by using the command **qos wred** before setting WRED parameters.

•Set WRED parameters for various precedence levels.

1. Enter the system view: system-view

2. Enter the interface view: interface *interface-type interface-number*

3. Set WRED parameters for various precedence levels: qos wred ip-precedence *ip-precedence* low-limit *low-limit* high-limit *high-limit* drop-probability *dropprobability*

ip-precedence: specifies the precedence of the IP packet. The value ranges from 0 to 7.

low-limit: specifies the low limit of the flow with certain precedence. The value ranges from 1 to 1024, and the default value is 10.

high-limit: specifies the high limit of the flow with certain precedence. The value ranges from 1 to 1024, and the default value is 30.

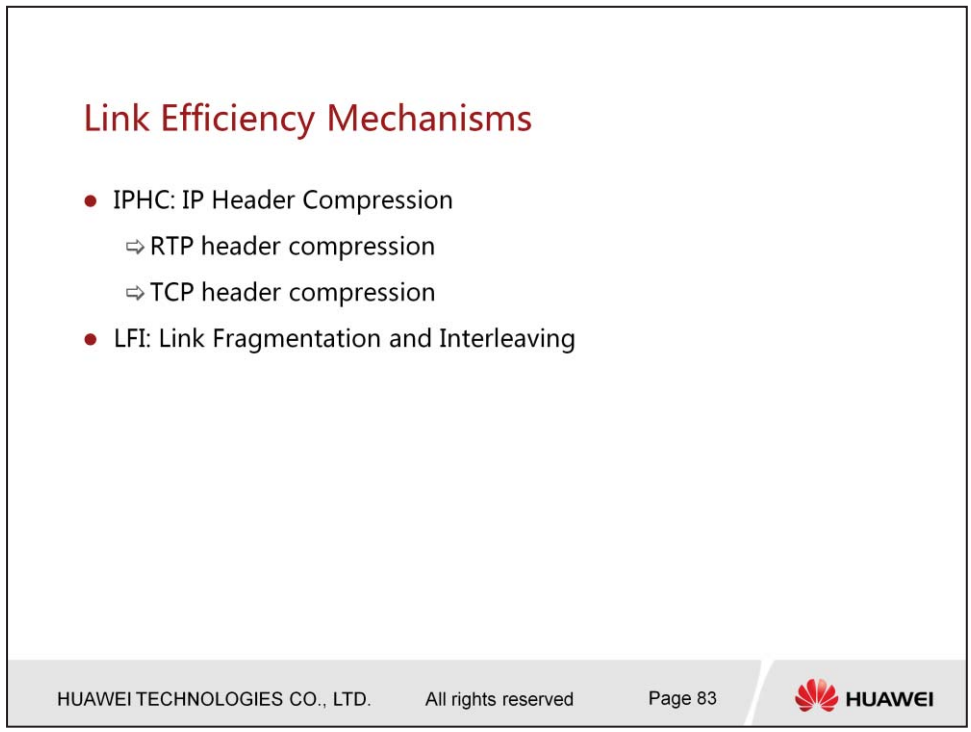
drop-probability: specifies the denominator of the drop probability. The value ranges from 1 to 255, and the default value is 10. The command **qos wred ipprecedence** is used to set the low limit, high limit, and denominator of the drop probability for the flows with certain precedence in WRED. The command **undo qos wred ip-precedence** is used to restore the default values of these parameters.



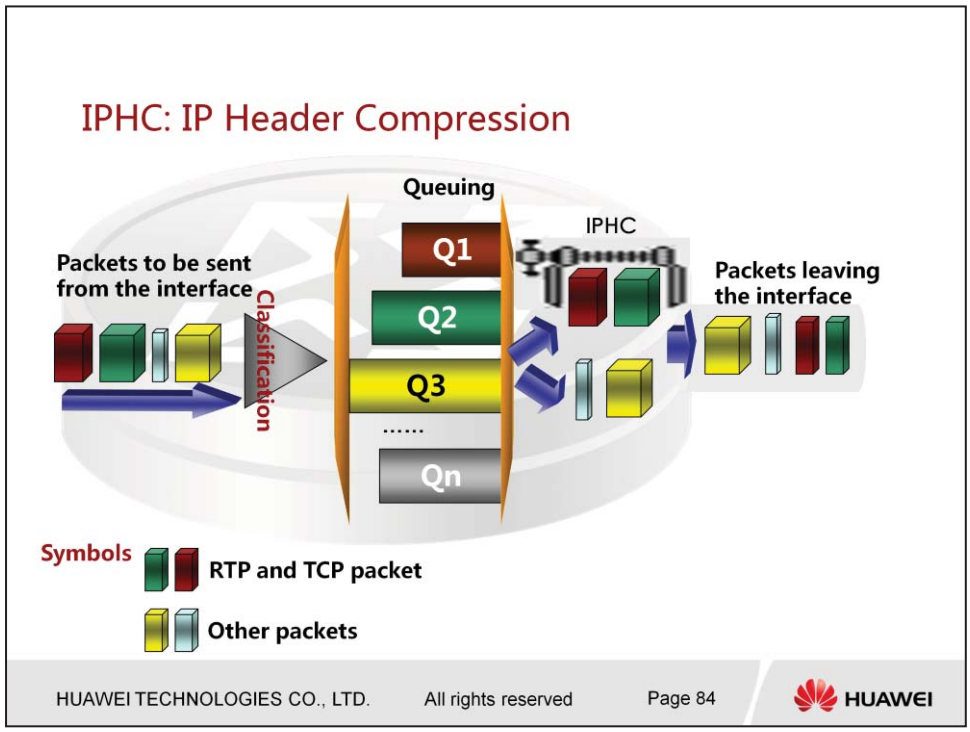
Contents

Classification and Marking
Policing and Shaping
Congestion Management
Congestion Avoidance Mechanisms
Link Efficiency Mechanisms

HUAWEI TECHNOLOGIES CO., LTD. All rights reserved



The VRP provides two mechanisms to raise the link efficiency: IP Header Compression (IPHC) and Link Fragmentation and Interleaving (LFI). The IP header compression protocol can compress the RTP and TCP headers. For the data of the same flow, most fields in the IP header are the same. These fields can be compressed to improve the link transmission efficiency. LFI is mainly applied to the low-speed link to reduce the delay and jitter of real-time data packets.



IP Header Compression (IPHC) is a host-to-host protocol. It is used to carry multimedia services such as voice and video over the IP network. IPHC is applied to low-speed links such as PPP link and FR link.

IPHC can compress RTP and TCP headers.

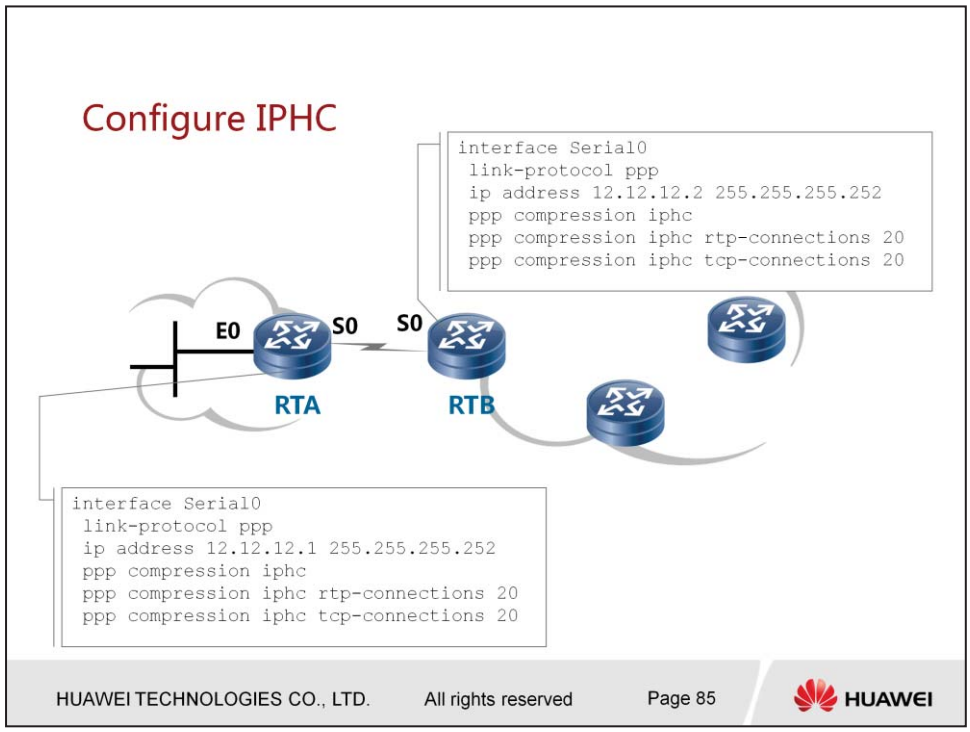
An RTP packet consists of the data and the header.

The data part of the RTP packet is relatively small and the header is relatively large. The 40-byte IP/UDP/RTP header consists of 12-byte RTP header, 20-byte IP header, and 8- byte UDP header.

A typical RTP packet carries 20-160 bytes data. To avoid unnecessary bandwidth consumption, you can apply the IPHC function to compress the header. IPHC compresses the 40-byte IP/UDP/RTP header to 2-4 bytes and the 40-byte payload is not compressed.

The compression ratio is (40+40)/(40+4), which is approximately 1.82. You can see that the compression is quite effective.

A TCP datagram contains 40-byte IP/TCP header. After TCP header compression, the header is 3-5 bytes.



In this example, IP header compression is enabled on two ends of the serial link.

The maximum number of connections for RTP and TCP header compression is 20.

The configuration commands are described as follows:

•Enable IP header compression.

1. Enter the system view: system-view

2. Enter the interface view: interface interface-type *interface-number*

3. Enable IP header compression : ppp compression iphc [nonstandard]

The command **ppp compression iphc** is used enable IP header compression on an interface. The command **undo ppp compression iphc** is used to disable IP header compression.

By default, IP header compression is disabled on the interface.

The command **ppp compression iphc** enables the IP/UDP/RTP header compression and IP/TCP header compression for RTP sessions.

You must configure this command on both ends of the link.

After the configuration, you must validate the command by running the commands **shutdown** and **undo shutdown** on the interface in sequence. If IPHC is applied to the MP, the commands **shutdown** and **undo shutdown** must be run on all interfaces bound to the MP.

•Set the maximum number of connections that allow TCP header compression.

1. Enter the system view: system-view

2. Enter the interface view: interface interface-type *interface-number*

3. Set the maximum number of connections for TCP header compression: ppp compression iphc tcp-connections *number* number: specifies the maximum number of connections for TCP header compression. The value ranges from 3 to 256 and the default value is 16. The command **ppp compression iphc tcp-connections** is used to set the maximum number of connections that allow TCP header compression. The command **undo ppp compression iphc tcp-connections** is used to restore the default value. After the configuration, you must validate the configuration by running **shutdown** and **undo shutdown** on the interface in sequence. If IPHC is applied to the MP, the command **shutdown** and **undo shutdown** must be run on all interfaces bound to the MP.

•Set the maximum number of connections that allow RTP header compression.

1. Enter the system view: system-view

2. Enter the interface view: interface interface-type *interface-number*

3. Set the maximum number of connections that allow RTP header compression :

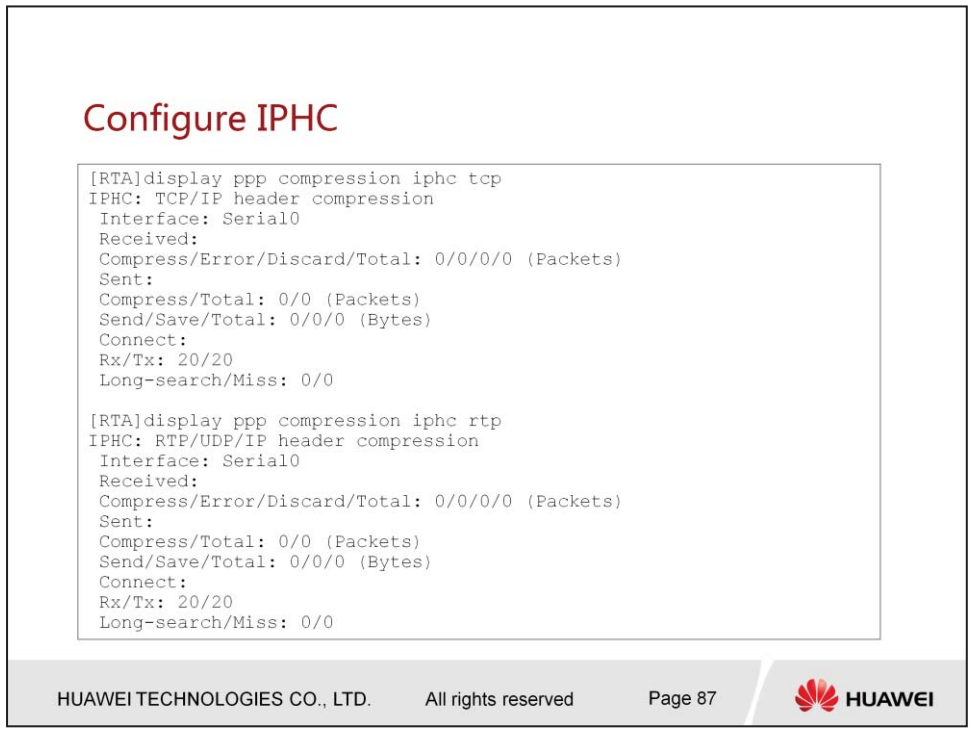
ppp compression iphc rtp-connections *number*

number: specifies the maximum number of connections that allow RTP header compression. The value ranges from 3 to 1000 and the default value is 16.

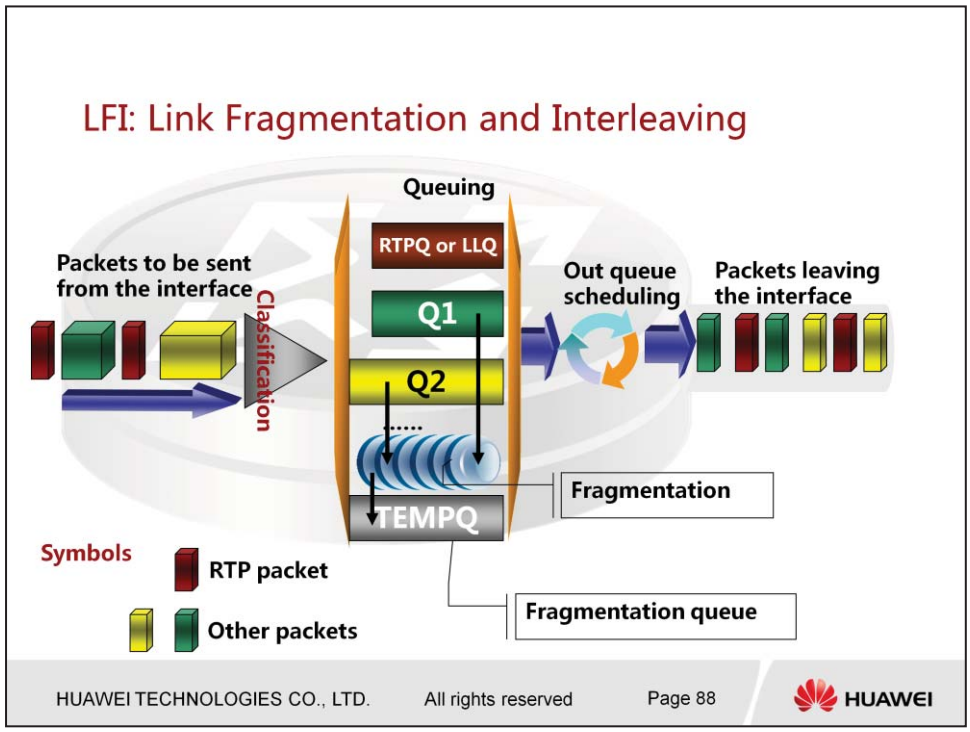
The command **ppp compression iphc rtp-connections** is used to set the maximum number of RTP connections that allow RTP header compression.

The command **undo ppp compression iphc rtp-connections** is used to restore the default value.

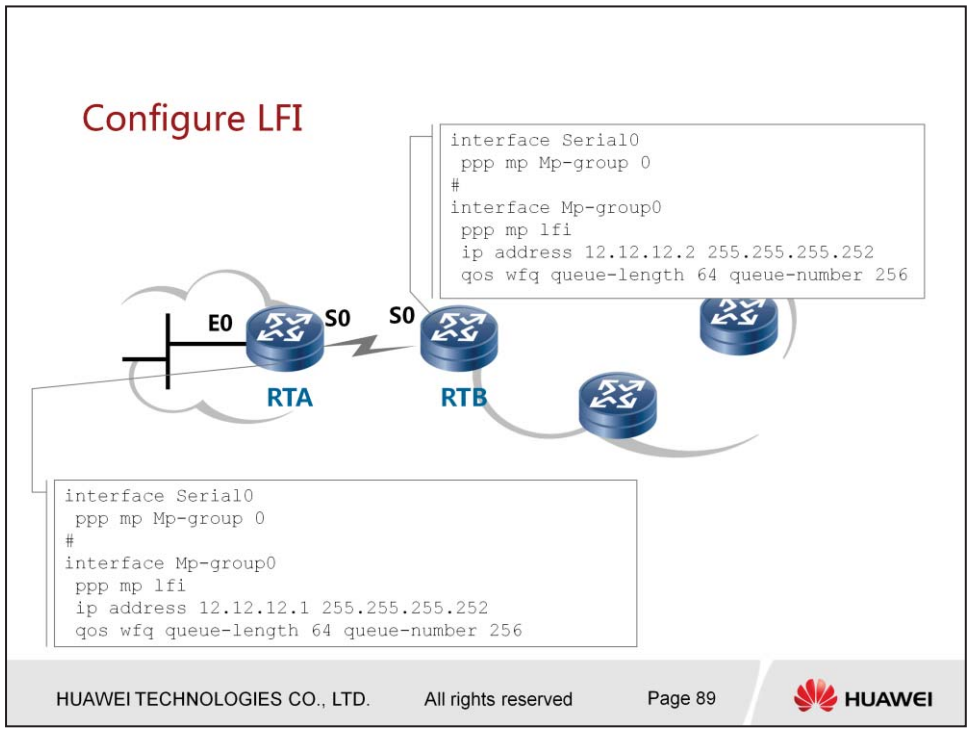
After the configuration, you must validate the configuration by running the commands **shutdown** and **undo shutdown** on the interface in sequence. If IPHC is applied to the MP, the commands **shutdown** and **undo shutdown** must be run on all interfaces bound to the MP.



After the configuration, you can run the command **display ppp compression iphc tcp** or **display ppp compression iphc rtp** to check the effect of IPHC.



Link fragmentation and interleaving (LFI) is applied to the low-speed link such as PPP link and FR link. On a low-speed serial link, real-time interactive communication such as Telnet and VoIP may cause congestion delay because of the transmission of large packets. For example, a voice packet arrives when a large packet is waiting for scheduling. The voice packet cannot be scheduled before the large packet is transmitted. For the real-time applications such as the interactive voice service, the congestion delay caused by large packets is too long. In this case, the peer end can only hear irregular intermittent audio. The interactive voice service requires that the end-to-end delay be within 100-150 ms. A packet of 1500 bytes (usually the MTU) cost 215 ms to pass through a 56 Kbps link, which is beyond the bearable delay of human being. To limit the delay of real-time packets on a relatively low-speed link such as a 56Kbps Frame Relay link or 64 Kbps ISDN B channel, you need to fragmentize large packets, and then add the small packets and fragments of large packets to the queue. LFI fragments large data frames to small frames and sends the small frames with other small packets. Thus, delay and jitter on a low-speed link is reduced. The figure shows the process of fragmentation and interleaving. Large packets and packets of real-time applications arrive at the interface at the same time. Except the packets in the RTPQ and LLQ, large packets in all queues are fragmented to some small packets and sent to the fragmentation queue. If RTPQ and LLQ have packets, these packets are scheduled first; otherwise, the fragmentation queue is scheduled. This method reduces the delay and jitter of real-time packets caused by large packets on a low-speed link.



In the example, LFI is configured on both ends of the serial link. The configuration commands are described as follows:

•Enable LFI.

1. Enter the system view: system-view

2. Enter the virtual template interface view: interface virtual-template *vtnumber* or enter the MP-group view: interface mp-group *number*

3. Enable LFI: ppp mp lfi

The **ppp mp lfi** command is used to enable the LFI function on an interface. The **undo ppp mp lfi** command is used to disable the LFI function. By default, after LFI is enabled on the virtual template interface, the fragmentation delay is 10 ms.

•Set the maximum delay of LFI fragmentation.

1. Enter the system view: system-view

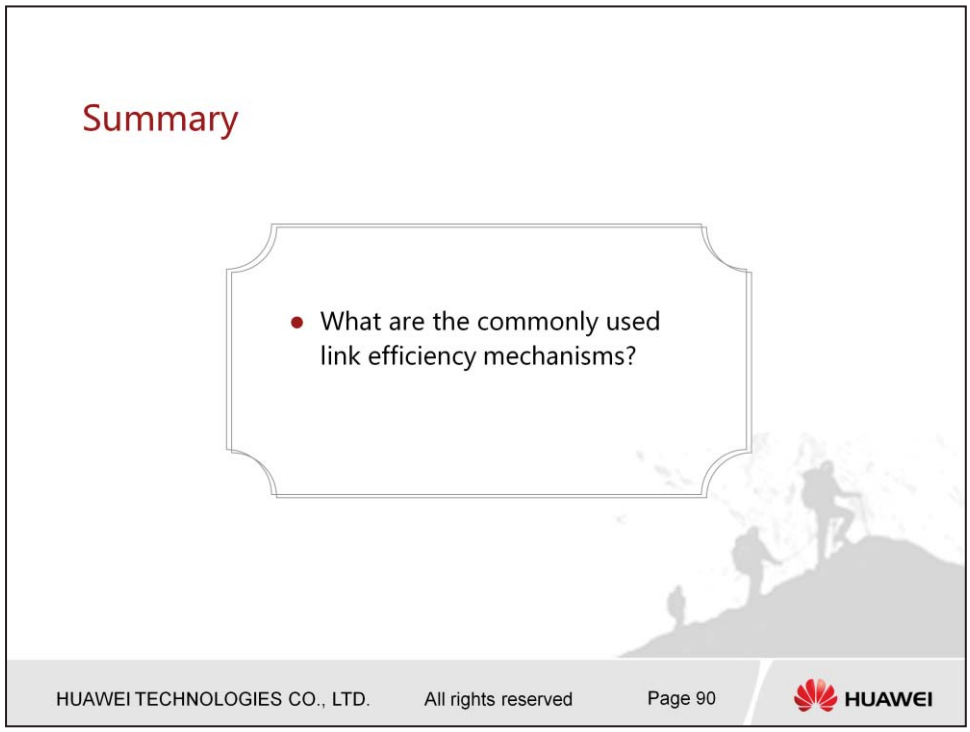
2. Enter the virtual template interface view: interface virtual-template *vt-number*

Or enter the MP-group: interface mp-group *number*

3. Set the maximum delay of LFI fragmentation: ppp mp lfi delay-per-frag *time*

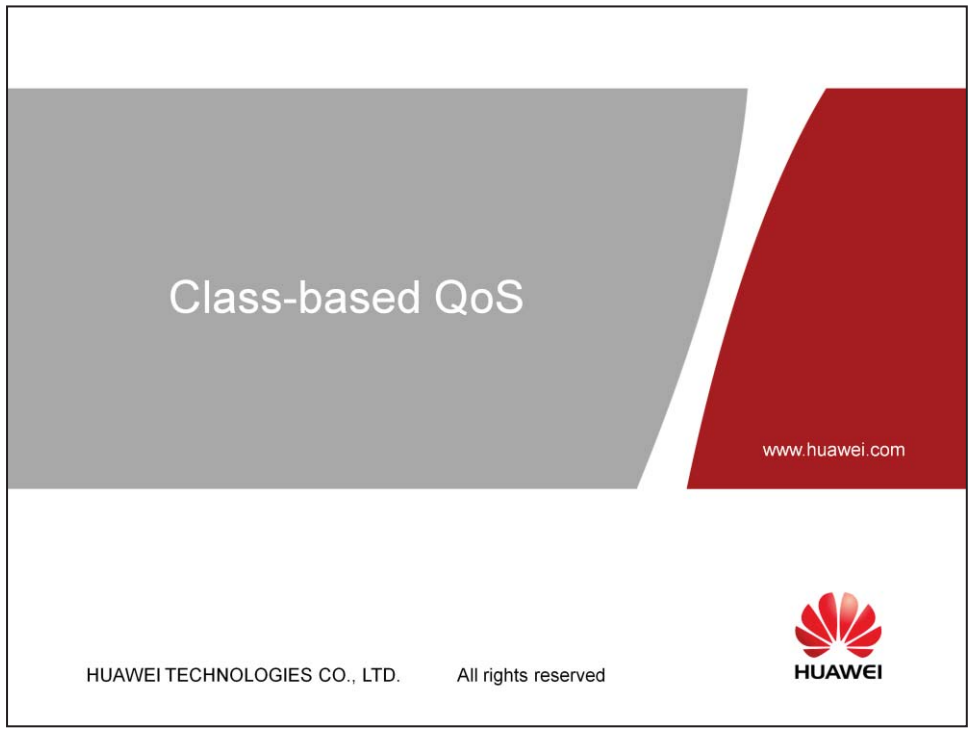
time: specifies the maximum delay of LFI fragmentation. The value ranges from 1 to 1000, in ms.

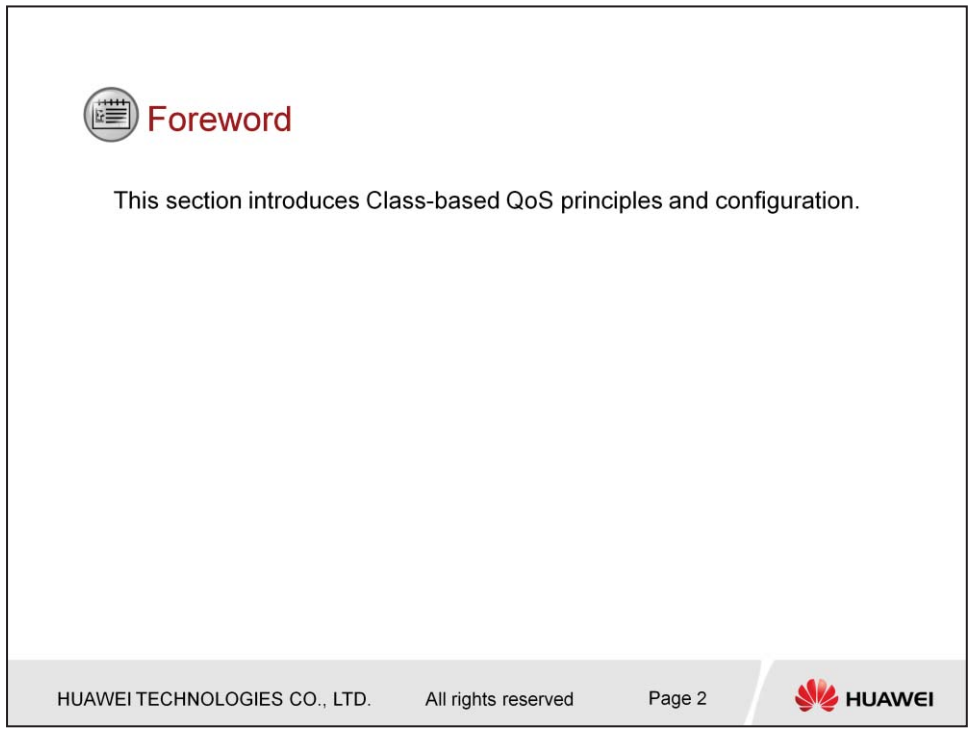
The command **ppp mp lfi delay-per-frag** is used to set the maximum delay for transmitting an LFI fragment. The command **undo ppp mp lfi delay-per-frag** is used to restore the default value of the maximum delay for transmitting an LFI fragment. By default, after LFI is enabled on the virtual template interface, the fragmentation delay is 10 ms.

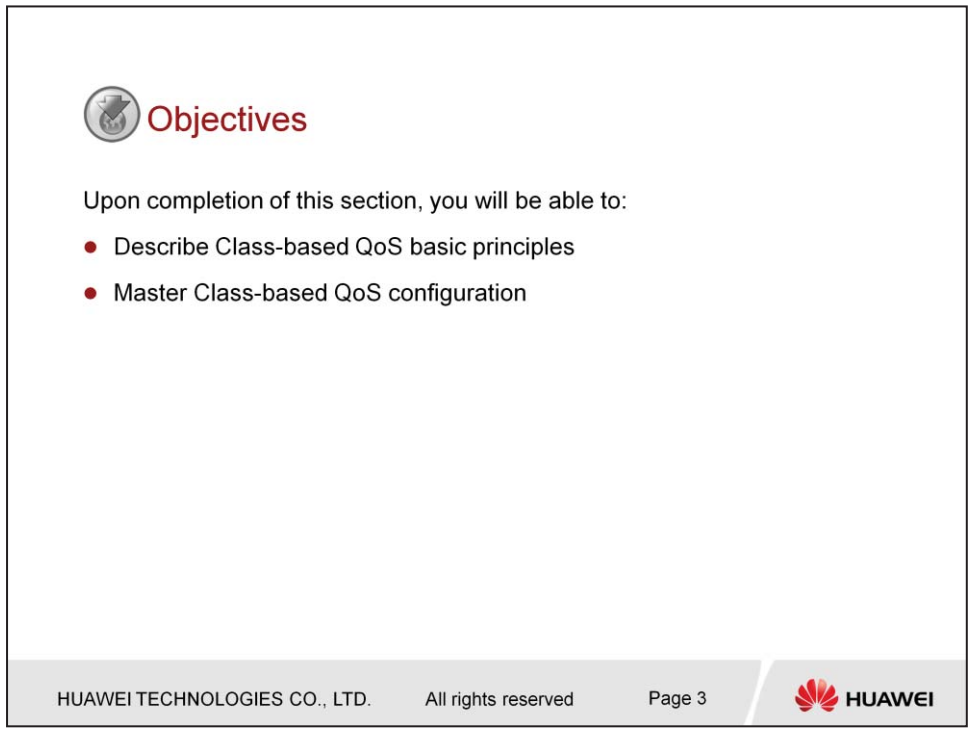


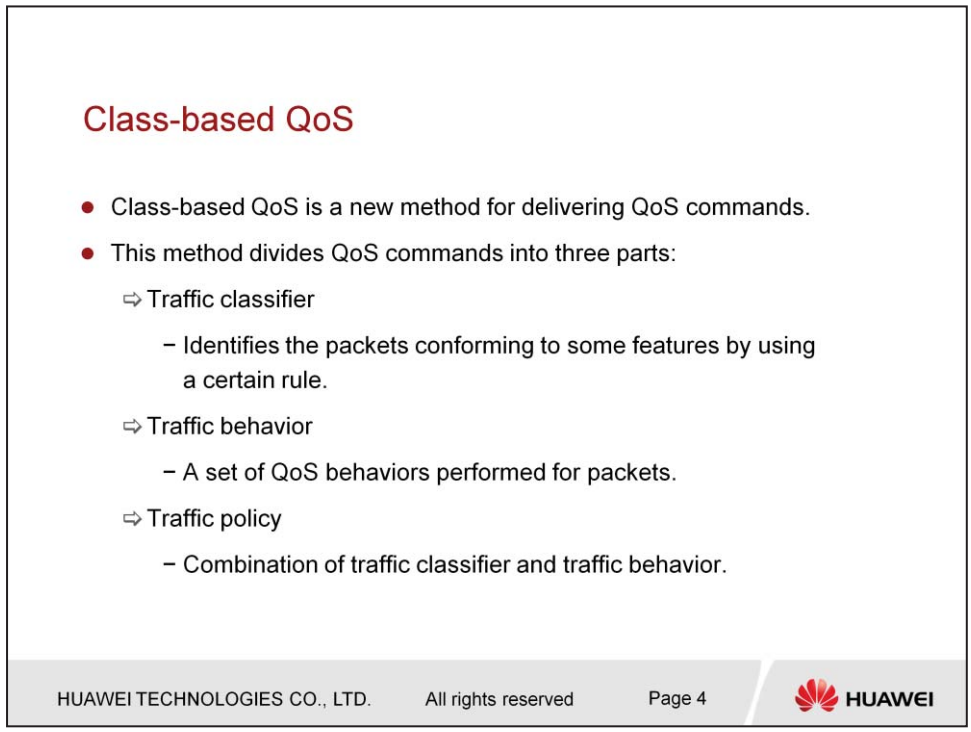
Q: What are the commonly used link efficiency mechanisms?

A: IP Header Compression and LFI.

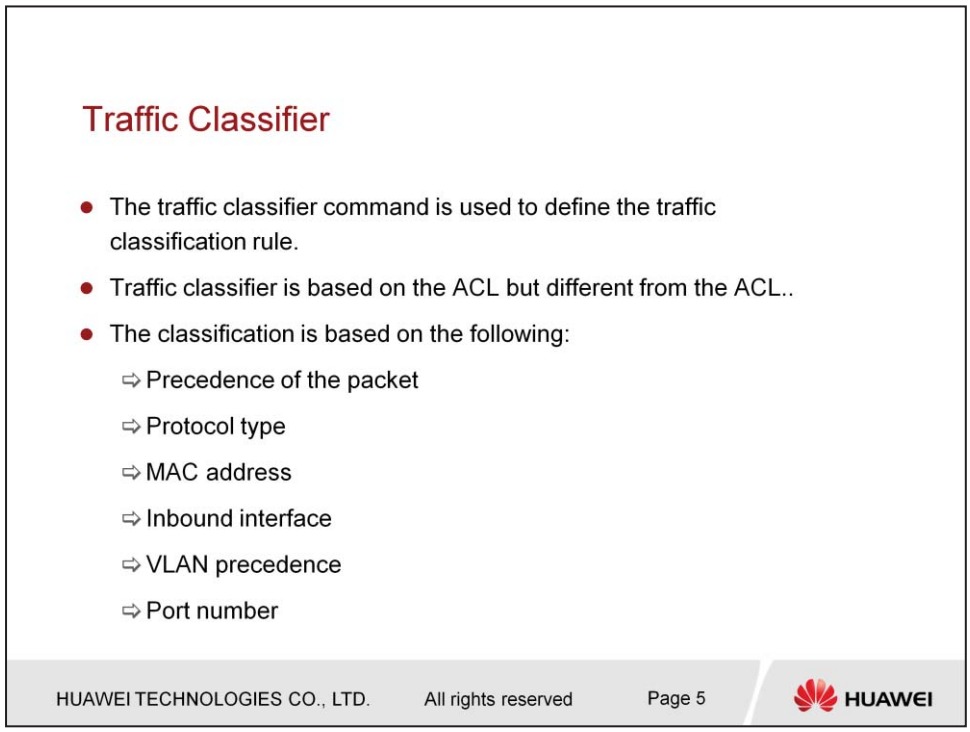

Class-based QoS
www.huawei.com
HUAWEI TECHNOLOGIES CO., LTD.
All rights reserved
HUAWEI


Foreword
This section introduces Class-based QoS principles and configuration.
HUAWEI TECHNOLOGIES CO., LTD. All rights reserved Page 2
HUAWEI


Objectives
Upon completion of this section, you will be able to:
Describe Class-based QoS basic principles
Master Class-based QoS configuration
HUAWEI TECHNOLOGIES CO., LTD.
All rights reserved
Page 3
HUAWEI



In class-based QoS, the traffic is classified according to a certain rule. The same type of traffic is associated with a behavior to form a traffic policy. This traffic policy is applied to implement class-based traffic policing, traffic shaping, congestion management, and precedence re-marking.



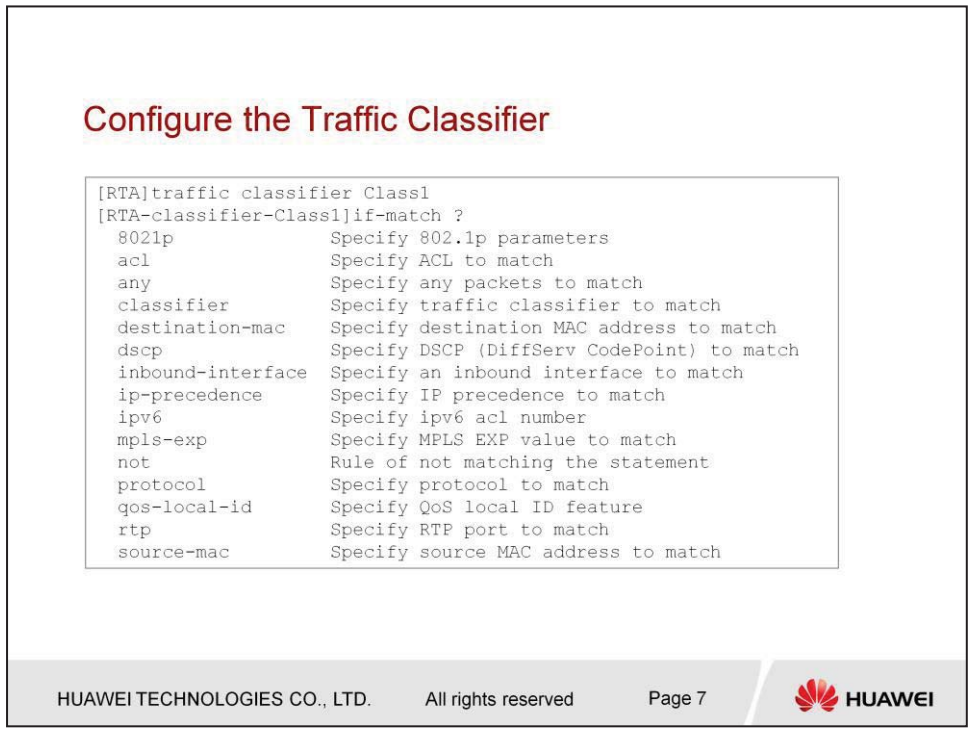
The traffic classifier uses certain rules to identify packets that conform to some features. It is the premise and basis for differentiated service. The traffic classifier uses the IP precedence or DSCP value in the ToS field of the IP header to identify the traffic with different precedence. The network administrator can also set the traffic classification policy. For example, the network administrator can define a traffic classifier based on the source IP address, destination IP address, MAC address, IP protocol port number and so on for application protocol. The classification result is not limited, meaning the result can be in a focused range determined by the quintuple (source IP address, source port number, protocol number, destination IP address, and destination port number). It can also be based on all packets in a network segment. The objective of traffic classification is to provide differentiated services. The traffic classifier is valid only when it is associated with some traffic control behavior or resource allocation behavior. The traffic control behavior depends on the service stage and current load of the network. For example, when packets arrive at the

network, traffic policing is performed based on the CIR. Before the packets leave a network node, traffic shaping is performed. When congestion occurs, queue scheduling is performed. When congestion worsens, the congestion avoidance method is adopted.

In class-based QoS, traffic classifier is based on the ACL but it is different from ACL. Traffic classifier only matches a behavior but does not define what behavior should be performed for the flows matching the classifier. The ACL defines the deny and permit behavior for access control.

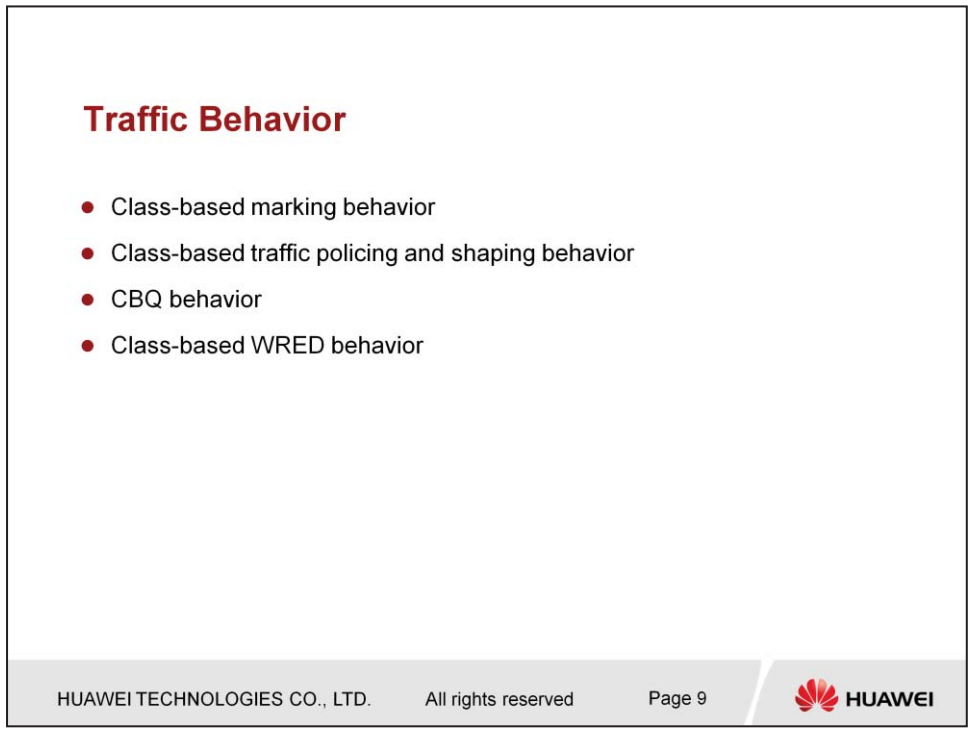
This is the most obvious difference between the traffic classifier and ACL. In addition, the traffic classifier and ACL have different match ranges. Currently, the match range of the traffic classifier is broader than or equals to the match range of the ACL.

We can say that the match range of ACL is a subset of the traffic classifier.



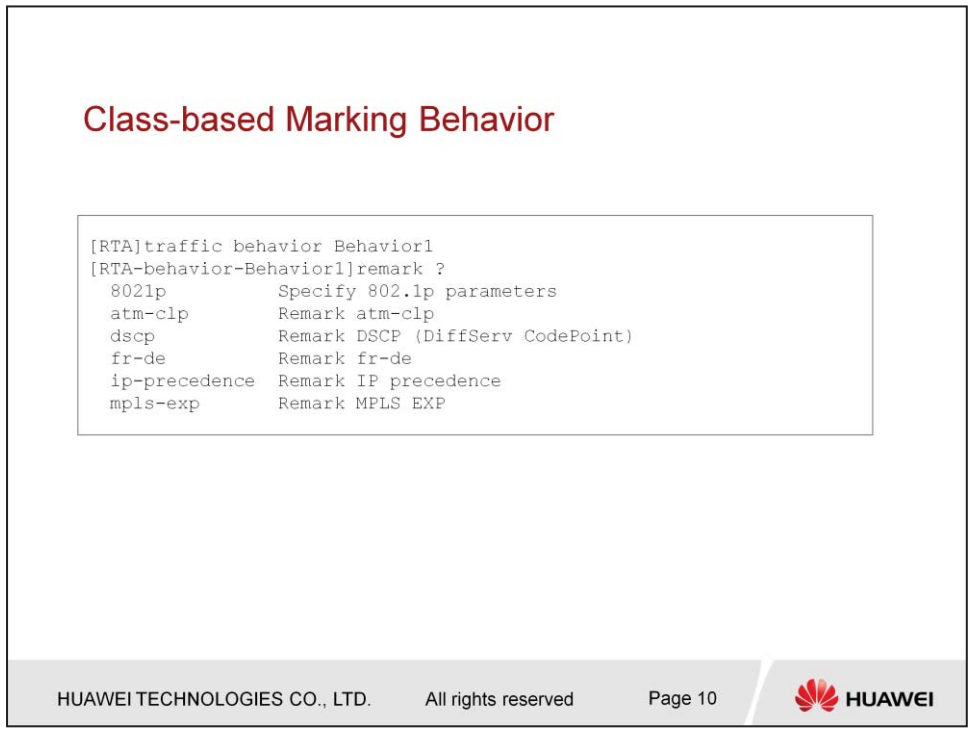
You can define a traffic classifier based on many conditions. The configuration commands are described as follows:
1. Enter the system view: system-view
2. Define a traffic classifier and enter the traffic classifier view: traffic classifier classifier-name [operator { and | or }]
3. Define the rule for matching all packets: if-match [not] any or define the classifier match rule: if-match [not] classifier classifier -name or define the match rule based on the ACL: if-match [not] acl access-list-number or define the match rule based on the IPv6 ACL: if-match [not] ipv6 acl access-list-number or define the match rule based on the MAC address: if-match [not] { destination-mac | source-mac } mac-address or define the match rule based on the inbound interface of the classifier: if-match [not] inbound-interface interface-type interface-number or define the match rule based on the DSCP value: if-match [not] dscp dscp-value &<1-8> or define the match rule based on the IP precedence: if-match [not] ip-precedence ip-precedence-value &<1-8> or define the match rule based on the MPLS or EXP field: ifmatch [not] mpls-exp mpls-experimental-value&<1-8> or match rule based on the VLAN 8021p: if-match [not] 8021p 8021p-value&<1-8> or define the match rule based on VLAN 8021p: if-match [not] protocol ip or define the match rule based

on the IPv6 protocol: if-match
[not] protocol ipv6 or define the match rule based on the RTP port number: if-match [not] rtp start-port min-rtp-port-number end-port max-rtp-port-number
The default value of the operator is **and**, that is, the relation among the match rules in the classifier view is logical AND.
The system predefines some classifiers and defines the universal rules for these classifiers. The name of the user-defined classifier cannot be the classifier predefined by the system. The user can directly use the predefined classifiers when defining the traffic policy. The predefined classifiers include the following:
(1) Default classifier
default-class: matches the default data flow.
(2) DSCP-based predefined classifier ef, af1, af2, af3, af4: respectively match the IP DSCP value ef, af1, af2, af3, and af4.
(3) IP-precedence-based predefined classifier ip-prec0,ip-prec1, …ip-prec7: respectively match the IP precedence levels 0,1,…7.
(4) MPLS-EXP-based predefined classifier mpls-exp0,mpls-exp1, …mpls-exp7: respectively match the MPLS EXP values 0,1,…7.
(5) VLAN-8021p-based predefined classifier vlan-8021p0,vlan-8021p1, vlan-8021p2,vlan-8021p3,vlan-8021p4,vlan-8021p5,vlan-8021p6, vlan-8021p7: respectively match the VLAN 8021p values 0,1,…7.
The classifier match rules cannot be used recursively. For example, if traffic classifier A defines the rule for matching traffic classifier B, traffic classifier B cannot directly or indirectly reference traffic classifier A. The match rule based on the destination MAC address is valid for only the outbound policy and the Ethernet interface. The match rule based on the source MAC address is valid for only the inbound policy and the Ethernet interface. List the IP precedence values in the same command; otherwise, the command **ifmatch ip-precedence** will overwrite the previous configuration. This restriction is also applicable when you configure the match rule based on the VLAN precedence or MPLS EXP field. The RTPQ takes precedence of CBQ, so RTP queue is scheduled first when the RTP queue and the scheduling queue based on RTP match rule are configured at the same time.



Traffic behavior is a set of QoS actions performed for packets. On the VRP the following traffic behaviors are used: class-based marking behavior, class-based traffic policing and shaping behavior, CBQ behavior, and class-based WRED behavior.

The class-based marking behavior can be associated with the classifier. It remarks the precedence or flag field of the packet to change the transmission status of the packet. The class-based traffic policing and shaping behavior implements traffic policy or traffic shaping. The CBQ behavior implements the class-based queue management. The class-based WRED behavior enables the WRED mechanism to cooperate with CBQ.



You can use the **traffic behavior** command to mark the traffic.
The marked field can be 802.1p, CLP of ATM cell, DSCP, DE field of FR, IP precedence, or MPLS
EXP field.
The configuration commands are as follows:
•Configure the behavior of re-marking the DSCP value of the
•packet.
1. Enter the system view system-view
2. Define a traffic behavior and enter the behavior view: traffic behavior *behavior–name*
3. Re-mark the DSCP value: remark dscp *dscp-value*
•Configure the behavior of re-marking the IP precedence of the
•packet.
1. Enter the system view: system-view
2. Define a behavior and enter the behavior view: traffic behavior *behavior–name*
3. re-mark the IP precedence of the packet: remark ip-precedence *ip-precedencevalue*
•Configure the behavior of re-marking the ED flag field of the FR
• packet.
1. Enter the system view: system-view
2. Define a behavior and enter the behavior view: traffic behavior *behavior–name*

3. Re-mark the DE flag field of the FR packet: remark fr-de *fr-de-value*

•Configure the behavior of re-marking the CLP flag field of the

• ATM cell.

1. Enter the system view: system-view

2. Define a behavior and enter the behavior view: traffic behavior *behavior–name*

3. Re-mark the CLP flag field of the ATM cell: remark atm-clp *atm-clp-value*

•Configure the behavior of re-marking the MPLS EXP field of the packet.

1. Enter the system view: system-view

2. Define a behavior and enter the behavior view: traffic behavior *behavior–name*

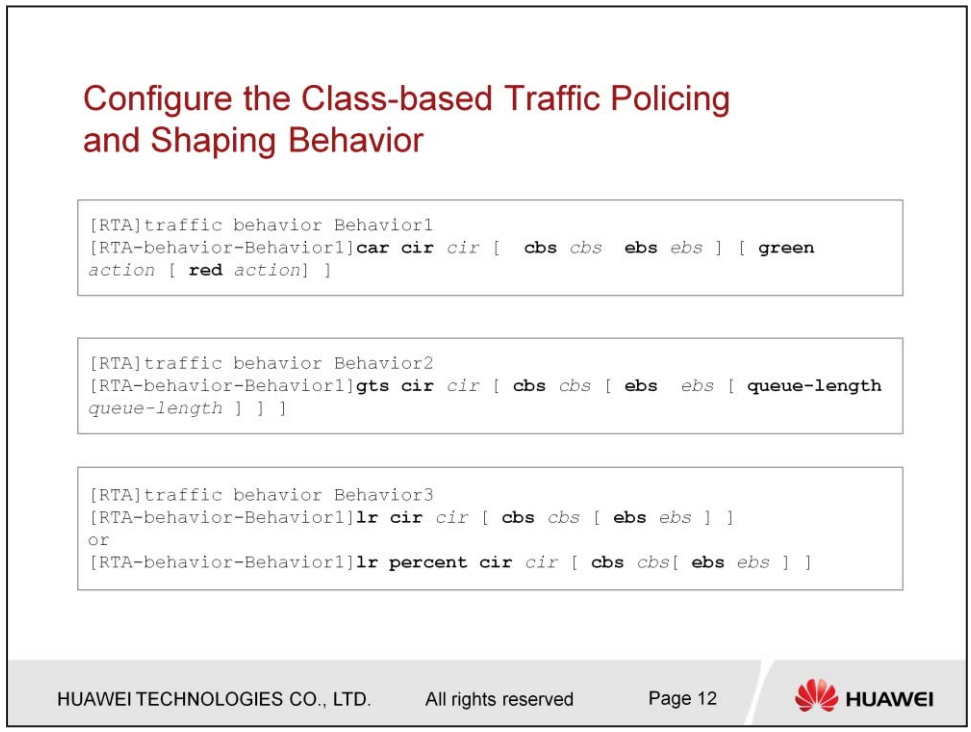
3. Re-mark the MPLS EXP field of the packet: remark mpls-exp *exp*

•Configure the behavior of re-marking the VLAN 8021P value.

1. Enter the system view: system-view

2. Define a behavior and enter the behavior view: traffic behavior *behavior–name*

3. Remark the behavior of re-marking the VLAN 8021P value: remark 8021p *8021p-value*



Traffic policing, traffic shaping, and CAR can also be configured for class-based QoS.

The configuration commands are as follows:

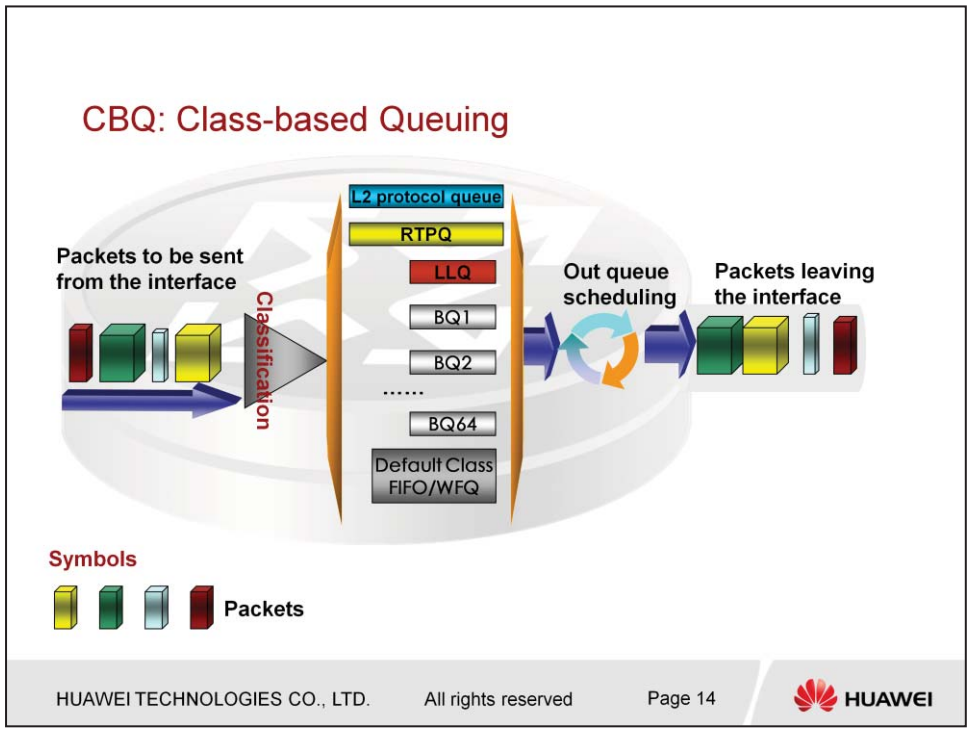
•Configure the class-based traffic policing behavior.

1. Enter the system view: system-view
2. Define a behavior and enter the behavior view: traffic behavior *behavior–name*
3. Configure the class-based traffic policing behavior: car cir *cir* [cbs *cbs* ebs *ebs*] [green *action* [red *action*]]

When the classifier in the traffic policy is associated with the behavior with the traffic policing feature, the policy can be applied in the inbound or outbound direction of an interface. When the classifier in the traffic policy is associated with the behavior with the traffic policing feature, this behavior overrides the behavior configured by the **qos car** command. If you repeat the command for the same behavior, the new configuration overwrites the previous one. If traffic policing behavior is configured but is not associated with the AF or EF classifier, the packet that passes the traffic policing detection can be sent. But if congestion occurs on the interface, the packets are added to the default queue.

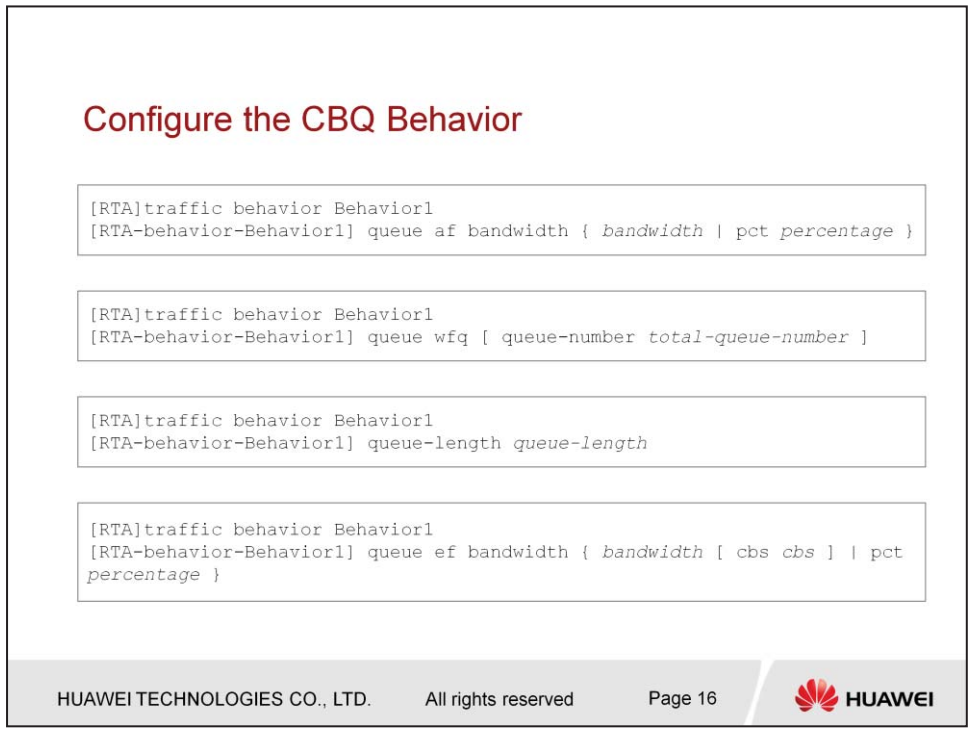
•Configure the class-based CAR behavior.

1. Enter the system view: system-view
2. Define a behavior and enter the behavior view: traffic behavior *behavior–name*
3. Configure the class-based CAR behavior: lr cir *cir* [cbs *cbs* [ebs *ebs*]]
Or configure the CAR behavior based on the CAR percentage: lr percent cir *cir* [cbs *cbs*[ebs *ebs*]]
If a policy contains the LR behavior, it can be applied in only the outbound direction of the interface. If you repeat the command for the same behavior, the new configuration overwrites the previous one. Note: this behavior can be configured on only the NE16E/08E/05 router.



The figure shows the process of class-based queuing (CBQ). In CBQ, packets are classified according to the IP precedence or DSCP value, inbound interface, and quintuple of the packet. Packets of different classes are added to different queues. The packets that do not match any class are added to the default queue. CBQ has a queue for low latency queuing (LLQ) to support the services of expedited forwarding (EF) class. These service flows are transmitted first and ensured with low delay. CBQ also has 64 queues for bandwidth queuing (BQ) to support the services of assured forwarding (AF) class. The bandwidth and controllable delay are ensured for each queue. CBQ has a queue for WFQ to support the services of the best effort (BE) class. These service flows are transmitted by the remaining bandwidth on the interface. CBQ classifies packets according to the inbound interface, ACL rule, IP precedence, DSCP, EXP, and label. Packets are added to corresponding q ueues after classification. The classification rule can be configured through the structural command line or the network management system. It can also be configured

automatically through the control plane of MPLS DiffServ-Aware TE.
Packets joining the LLQ and BQ are measured. Considering the link layer control packet, overhead of link layer encapsulation and physical overhead (for example, ATM cell tariff), we recommended that the bandwidth be occupied by the RTPQ, LLQ, and BQ not exceed 75% of the total bandwidth on the interface. LLQ can adopt only tail drop; BQ can adopt tail drop and WRED (based on IP precedence, DSCP, or MPLS EXP); WFQ can adopt tail drop and RED. CBQ can define scheduling policies (bandwidth, delay, and so on) for different services.
Because complex traffic classification is used, enabling the CBQ feature on a high-speed interface (such as a GE) resumes some system resources.



Configuring the CBQ behavior involves defining the bandwidth for the AF queue and EF queue, configuring the scheduling mode, and setting the queue length.

The configuration commands are described as follows:

•Configure the AF queue.

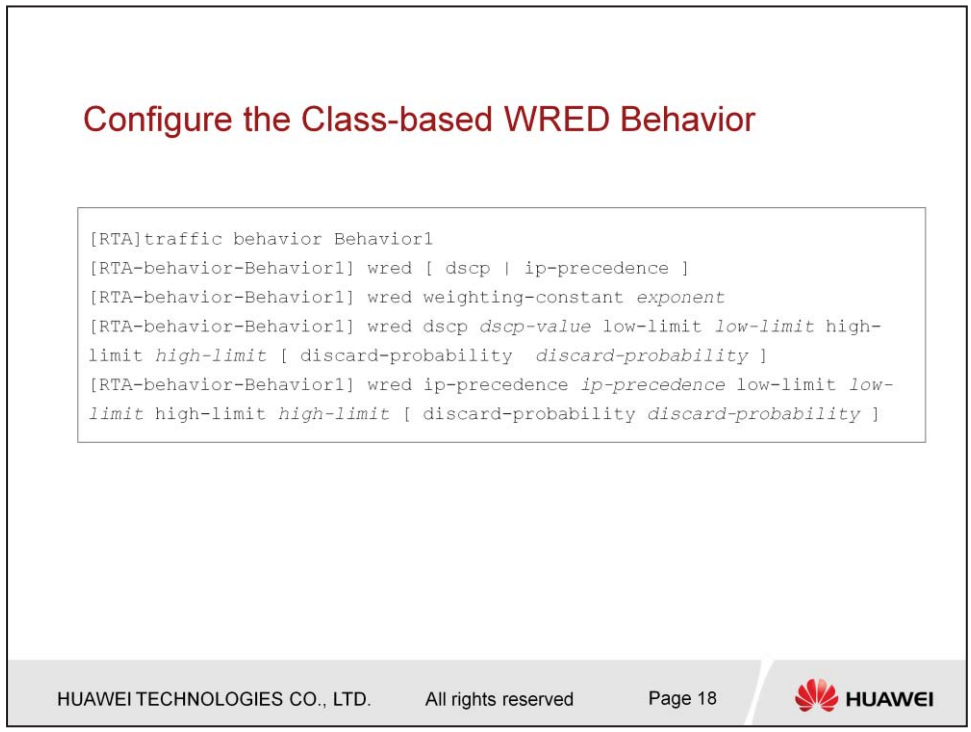
1. Enter the system view: system-view
2. Define a behavior and enter the behavior view: traffic behavior *behavior–name*
3. Configure the AF queue: queue af bandwidth { *bandwidth* | pct *percentage* } This configuration is applicable in only the outbound direction of an interface or ATM PVC. For the same policy, the EF queue and AF queue must use the same bandwidth unit, namely the absolute value of bandwidth or the percentage of bandwidth.

•Configure the WFQ.

1. Enter the system view: system-view
2. Define a behavior and enter the behavior view: traffic behavior *behavior–name*
3. Configure the WFQ : queue wfq [queue-number *total-queue-number*]

This configuration is applicable in only the outbound direction of an interface or ATM PVC. The traffic behavior with this feature

can be associated with only the default classifier.
•Set the maximum queue length.
1. Enter the system view: system-view
2. Define a behavior and enter the behavior view : traffic behavior *behavior–name*
3. Set the maximum queue length: queue-length *queue-length*
This command can be used only when the AF queue and WFQ are configured. The drop policy is tail drop.
•Configure the EF queue.
1. Enter the system view: system-view
2. Define a behavior and enter the behavior view: traffic behavior *behavior–name*
3. Configure the EF queue: queue ef bandwidth { *bandwidth* [cbs *cbs*] | pct *percentage* }
This command cannot be used with the **queue af**, **queue-length**, and **wred** commands in the behavior view. The default classifier cannot be associated with the behavior configured by this command. For the same policy, the EF queue and AF queue must use the same bandwidth unit, namely the absolute value of bandwidth or the percentage of bandwidth.



The procedure for configuring class-based WRED is similar to the procedure for configuring WRED in common QoS.

The configuration commands are described as follows:

•Configure the class-based WRED drop policy.

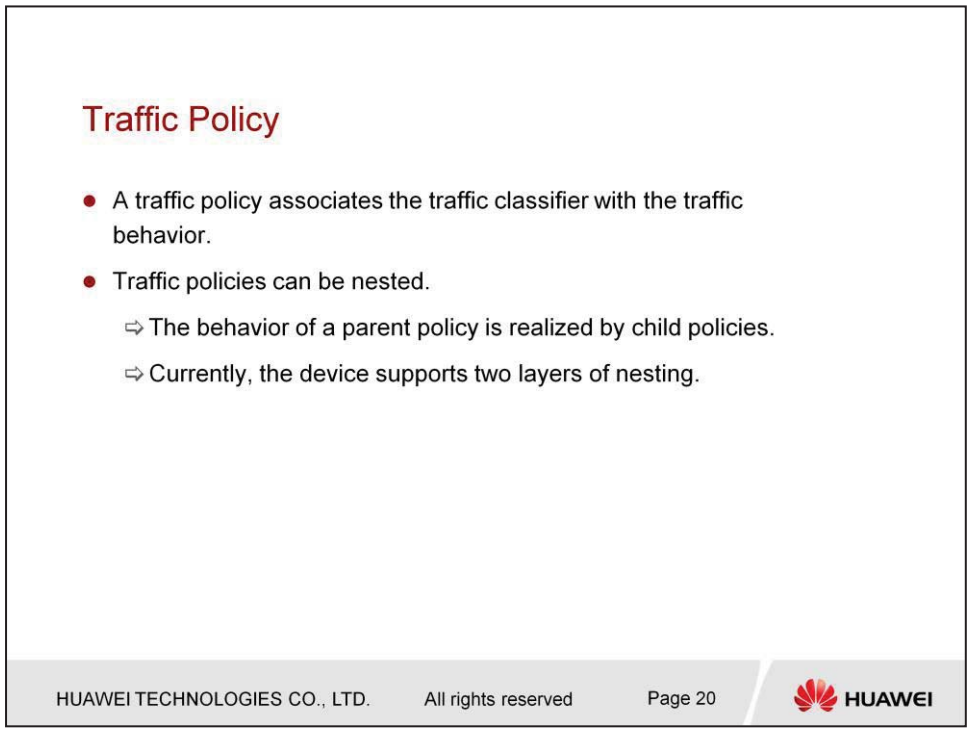
1. Enter the system view: system-view
2. Define a behavior and enter the behavior view: traffic behavior *behavior–name*
3. Configure the drop policy: wred [dscp | ip-precedence]

The drop policy can be configured only when the AF queue and WFQ are configured. The **wred** and **queue-length** command is mutually exclusive. When WRED drop policy is cancelled, other configurations for random drop are also cancelled. When the QoS policy containing the WRED feature is applied to an interface, the WRED in the QoS policy overrides the previous WRED configuration on the interface. The IP precedence or DSCP can be configured for the behavior associated with the default classifier.

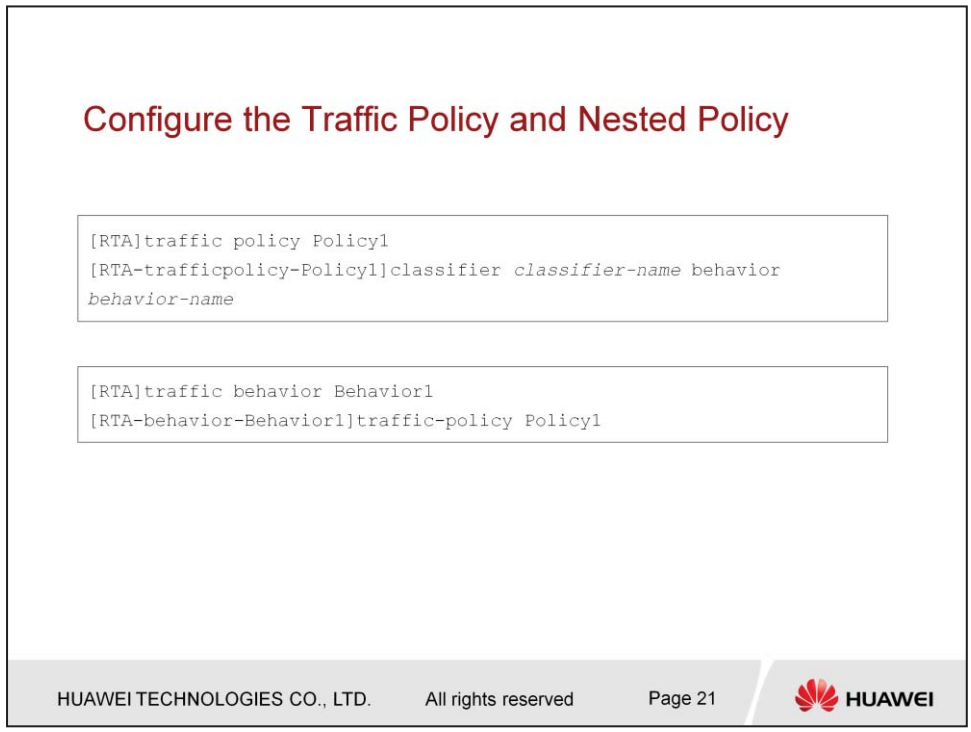
•Set the drop parameters for class-based WRED.

1. Enter the system view: system-view
2. Define a behavior and enter the behavior view: traffic behavior *behavior–name*
3. Set the exponent for calculating the average queue length for WRED: wred weighting-constant *exponent* Or set the low limit and high limit of flows with a certain DSCP value and the

denominator of the drop probability: wred dscp *dscp-value* low-limit *low-limit* hig\hlimit *high-limit* [discard-probability *discard-probability*] Or set the low limit and high limit of flows with a certain IP precedence level and the denominator of the drop probability: wred ip-precedence *ip-precedence* lowlimit *low-limit* high-limit *high-limit* [discard-probability *discard-probability*] Before setting the exponent for calculating the average queue length, you must configure the AF queue and enable WRED by using the **wred** command. Before setting the high limit and low limit for a DSCP value, you must configure the AF queue and enable DSCP-based WRED by using the **wred** command. Before setting the high limit and low limit for a precedence level, you must configure the AF queue and enable precedence-based WRED by using the **wred** command.



After defining the traffic classifier and traffic behavior, you need to configure the traffic policy to associate the traffic classifier with the traffic behavior. Policy nesting means that a QoS policy contains another QoS policy. The behavior of a parent policy is realized by child policies. After the behavior defined in the parent policy is performed for a flow, the flow is subdivided by the child policy and the behavior in the child policy is performed. Currently, the device supports two layers of nesting.



A traffic policy associates the traffic classifier with the traffic behavior. The configuration commands are described as follows:

•Define a policy and enter the policy view.

1. Enter the system view: system-view

2. Define a policy and enter the policy view: traffic policy *policy-name* The system predefines a policy. This policy specifies the predefined classifiers and associates them with predefined behaviors. This policy is named default and contains the default CBQ policy.

The rules of the default policy are as follows:

(1) The predefined classifier ef is associated with the predefined

(2) behavior ef.

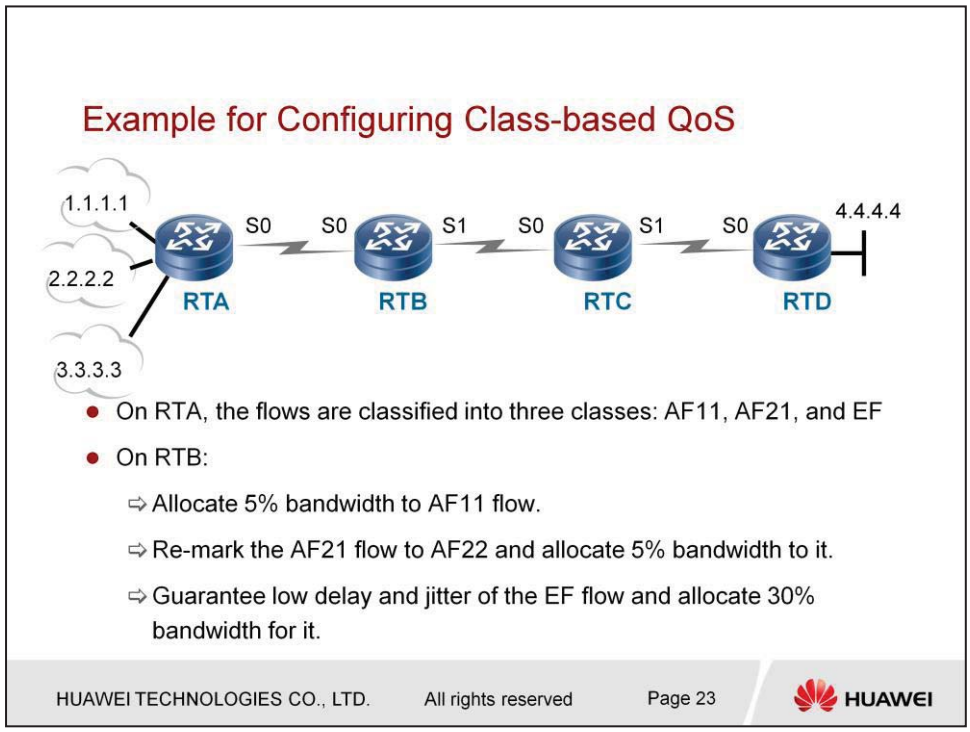
(2) The predefined classifiers af1 to af4 are associated with the predefined behavior af.

(3) The default classifier is associated with the predefine behavior be. Other policies cannot use the name of the policy predefined by the system. If a policy is applied to an interface, the policy cannot be deleted. To delete this policy, cancel the application of this policy on the interface, and then run the command **undo traffic policy** to delete the policy.

•Specify a traffic behavior for the classifier.

1. Enter the system view: system-view

2. Define a policy and enter the policy view: traffic policy *policy-name*
3. Specify a traffic behavior for the classifier: classifier *classifier-name* behavior *behavior-name*
•Configure the nested policy.
1. Enter the system view: system-view
2. Define a behavior and enter the behavior view: traffic behavior *behavior–name*
3. Configure the nested policy: traffic-policy *policy-name*
If the re-marking behavior is configured in both the parent policy and child policy, the re-marking behavior in the child policy overwrites that in the parent policy. If the CAR (or GTS) behavior is configured in both the parent policy and child policy, the CAR (or GTS) behavior is performed twice. The CAR (or GTS) behavior in the child policy is performed first. If the queuing behavior is configured in both the parent policy and child policy, packets exceeding the line rate (LR) are added to the queue specified in the child policy. Packets within the LR are added to the queue specified in the parent policy. If the EF queuing, AF queuing, and WFQ behaviors are configured in the child policy, that is, the child policy is a CBQ policy, the classifier in the parent policy must be associated with the LR behavior. The packets exceeding the LR are scheduled through the CBQ scheduling algorithm defined in the child policy. If the command **lr percent** is configured in the parent policy, the bandwidth for the queuing behavior in the child policy must be in the percentage form. In this case, the LR behavior cannot be configured in the child policy.



This is an example for configuring class-based QoS. RTA classifies the flows, and RTB limits the bandwidth and re-marks the flows.

Example for Configuring Class-based QoS—Configure RTA

```
acl number 2001
 rule 0 permit source 1.1.1.1 0
acl number 2002
 rule 0 permit source 2.2.2.2 0
acl number 2003
 rule 0 permit source 3.3.3.3 0

traffic classifier Class_af11 operator and
 if-match acl 2001
traffic classifier Class_af21 operator and
 if-match acl 2002
traffic classifier Class_ef operator and
 if-match acl 2003

traffic behavior Behavior_af11
 remark dscp af11
traffic behavior Behavior_af21
 remark dscp af21
traffic behavior Behavior_ef
 remark dscp ef

traffic policy Mark
 classifier Class_af11 behavior Behavior_af11
 classifier Class_af21 behavior Behavior_af21
 classifier Class_ef behavior Behavior_ef

interface Serial0
 ip address 12.12.12.1 255.255.255.252
 traffic-policy Mark outbound
```

HUAWEI TECHNOLOGIES CO., LTD. All rights reserved

RTA classifies the flows. It marks the flows as AF11, AF21, and EF according to their source address.

Example for Configuring Class-based QoS—Configure RTB

```
traffic classifier Class_af11 operator and
 if-match dscp af11
traffic classifier Class_af21 operator and
 if-match dscp af21
traffic classifier Class_ef operator and
 if-match dscp ef
#
traffic behavior Behavior_af11
 queue af bandwidth pct 5
traffic behavior Behavior_af21
 remark dscp af22
 queue af bandwidth pct 5
traffic behavior Behavior_ef
 queue ef bandwidth pct 30
#
traffic policy dscp
 classifier Class_af11 behavior Behavior_af11
 classifier Class_af21 behavior Behavior_af21
 classifier Class_ef behavior Behavior_ef
interface Serial1
 ip address 23.23.23.1 255.255.255.252
 traffic-policy dscp outbound
```

HUAWEI TECHNOLOGIES CO., LTD. All rights reserved HUAWEI

RTB carries out different QoS policies according to the flag of packets.

Example for Configuring Class-based QoS—Debugging (1)

```
[RTB]display traffic policy interface Serial 1
  Interface: Serial1
  Direction: Outbound
  Policy: dscp
   Classifier: default-class
     Matched : 211/14536 (Packets/Bytes)
     Rule(s) : if-match any
     Behavior: be
     Default Queue:
      Flow Based Weighted Fair Queuing
        Max number of hashed queues: 256
        Matched  : 0/0 (Packets/Bytes)
        Enqueued : 0/0 (Packets/Bytes)
        Discarded: 0/0 (Packets/Bytes)
        Discard Method: Tail
```

HUAWEI TECHNOLOGIES CO., LTD. All rights reserved HUAWEI

You can run the command **display traffic policy interface** to check the effect of the configuration. The display information includes the classifier, traffic behavior for each class, and number of matched packets.

Example for Configuring Class-based QoS—Debugging (2)

```
[RTB]display traffic policy interface Serial 1
  ......
   Classifier: Class_af11
     Matched : 5/420 (Packets/Bytes)
     Operator: AND
     Rule(s) : if-match dscp af11
     Behavior: Behavior_af11
 Assured Forwarding:
        Bandwidth 2 (Kbps)
        Matched  : 0/0 (Packets/Bytes)
        Enqueued : 0/0 (Packets/Bytes)
        Discarded: 0/0 (Packets/Bytes)
```

HUAWEI TECHNOLOGIES CO., LTD. All rights reserved HUAWEI

The display is the information about the policy for the AF11 class.

Example for Configuring Class-based QoS—Debugging (3)

```
[RTB]display traffic policy interface Serial 1
  ......
   Classifier: Class_af21
     Matched : 15/1260 (Packets/Bytes)
     Operator: AND
     Rule(s) : if-match dscp af21
     Behavior: Behavior_af21
      Marking:
        Remark DSCP af22
        Remarked: 15 (Packets)
      Assured Forwarding:
        Bandwidth 2 (Kbps)
        Matched  : 0/0 (Packets/Bytes)
        Enqueued : 0/0 (Packets/Bytes)
        Discarded: 0/0 (Packets/Bytes)
```

HUAWEI TECHNOLOGIES CO., LTD. All rights reserved HUAWEI

The display is the information about the policy for the AF21 class.

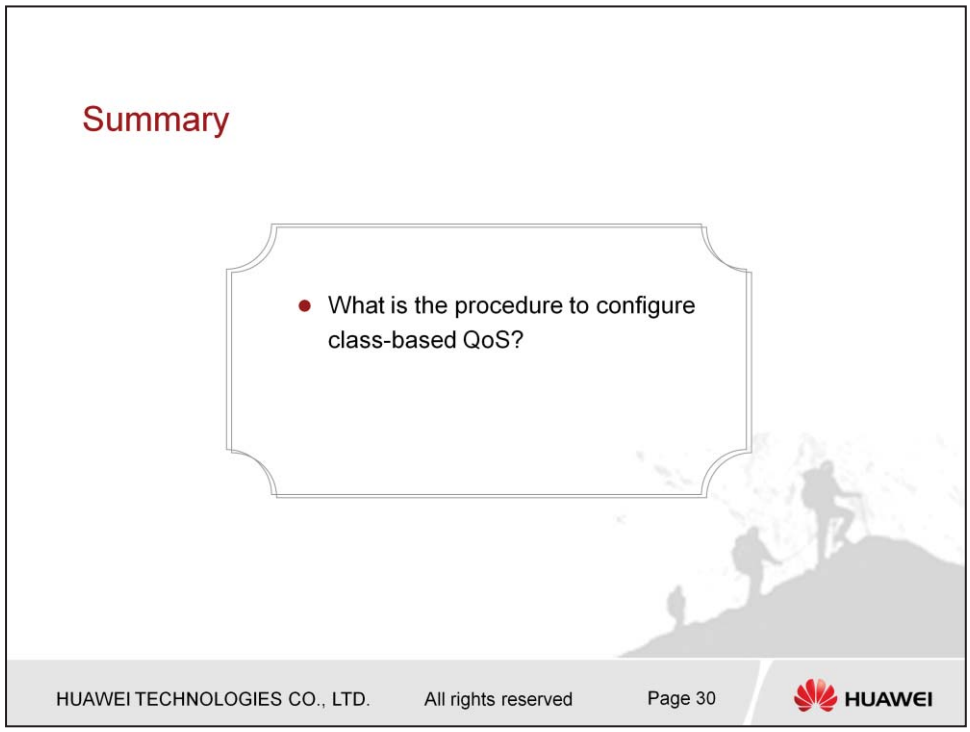
Example for Configuring Class-based QoS—Debugging (4)

```
[RTB]display traffic policy interface Serial 1
  ......
   Classifier: Class_ef
     Matched : 5/420 (Packets/Bytes)
     Operator: AND
     Rule(s) : if-match dscp ef
     Behavior: Behavior_ef
 Expedited Forwarding:
        Bandwidth 15 (Kbps), CBS 1500 (Bytes)
        Matched  : 0/0 (Packets/Bytes)
        Enqueued : 0/0 (Packets/Bytes)
        Discarded: 0/0 (Packets/Bytes)
```

HUAWEI TECHNOLOGIES CO., LTD. All rights reserved HUAWEI

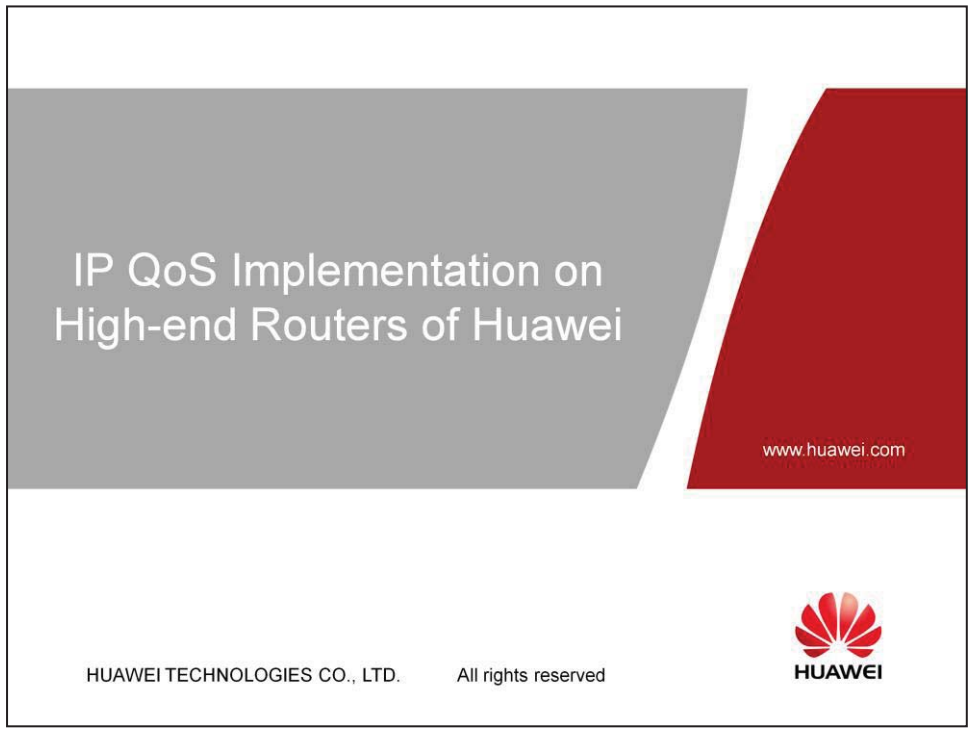
The display is the information about the policy for the EF class.

The packets marked as EF are expeditiously forwarded.



Q: What is the procedure to configure class-based QoS?

A: 1. Configure the traffic classifier. 2. Configure the traffic behavior. 3. Associate the traffic classifier with the traffic behavior. 4. Apply the traffic policy.


IP QoS Implementation on
High-end Routers of Huawei
www.huawei.com
HUAWEI TECHNOLOGIES CO., LTD.
All rights reserved
HUAWEI

MPLS packet. The value is an integer ranging from 0 to 7. Using the command **trust upstream** { *ds-domain-name* | **default** }, you can bind the DS domain to an interface so that the interface supports simple traffic classification. Then this interface can change the precedence of IP and MPLS packets. *ds-domain-name*: specifies the name of a DS domain. The value is a string of 1 to 8 characters. DS domain consists of a group of connected DS nodes. These nodes use the same service policy and the same PHB. **default**: specifies the default DS domain.



Foreword

IP QoS technology provides better service quality guarantee to the network, different products can provide the different levels of QoS.

HUAWEI TECHNOLOGIES CO., LTD. All rights reserved





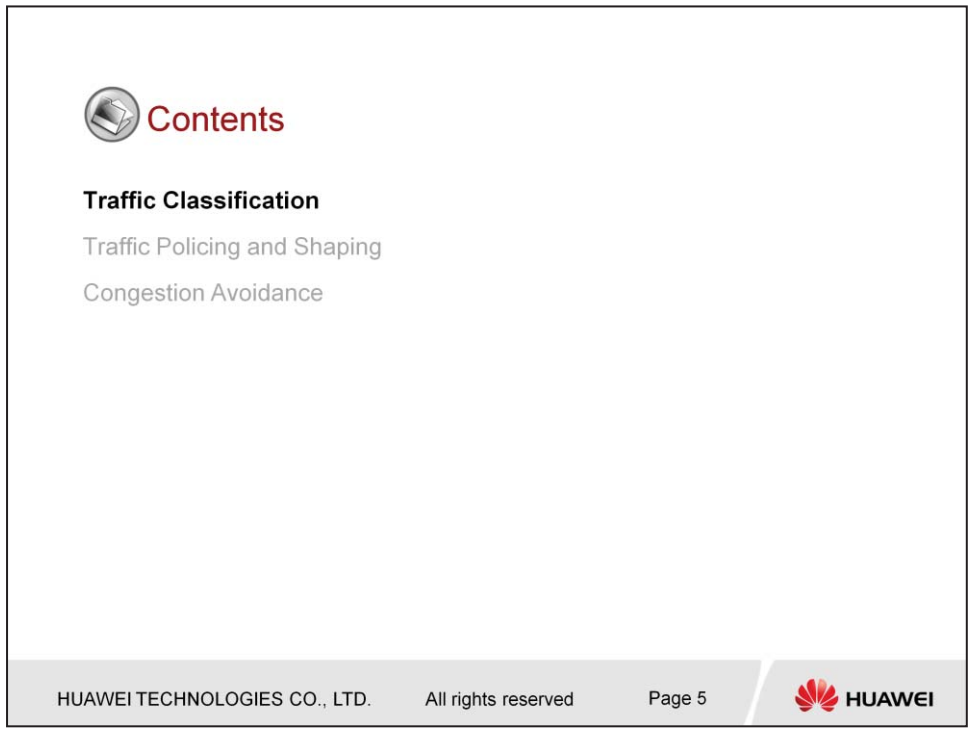
Objectives

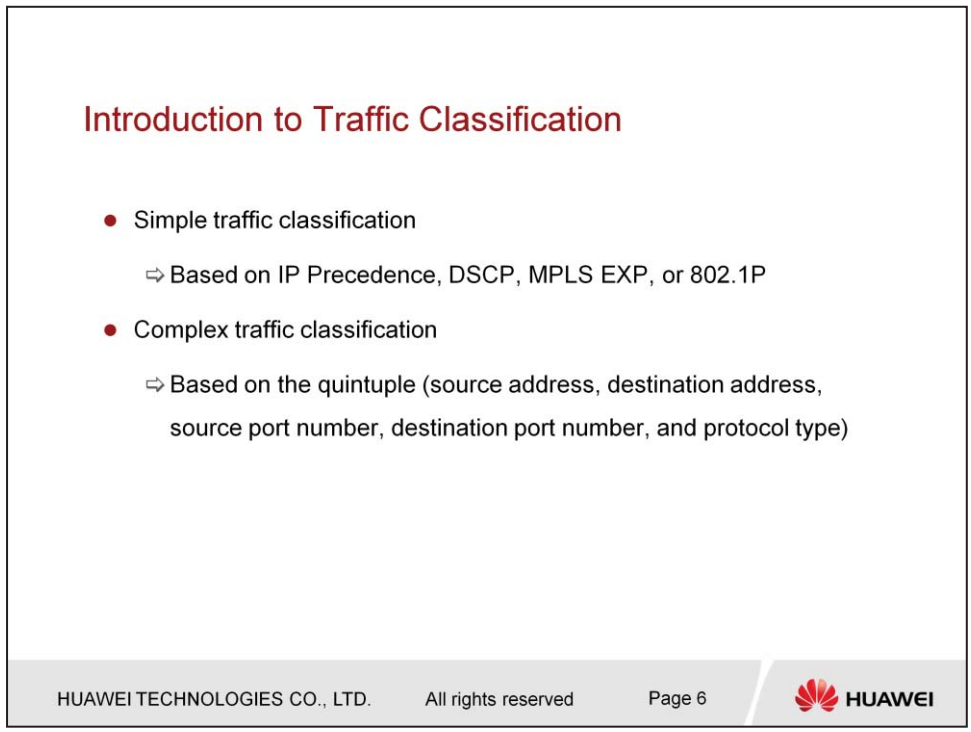
Upon completion of this section, you will be able to:

- Understand IP QoS principles on high-end products
- Master IP QoS configuration on high-end products

HUAWEI TECHNOLOGIES CO., LTD. All rights reserved

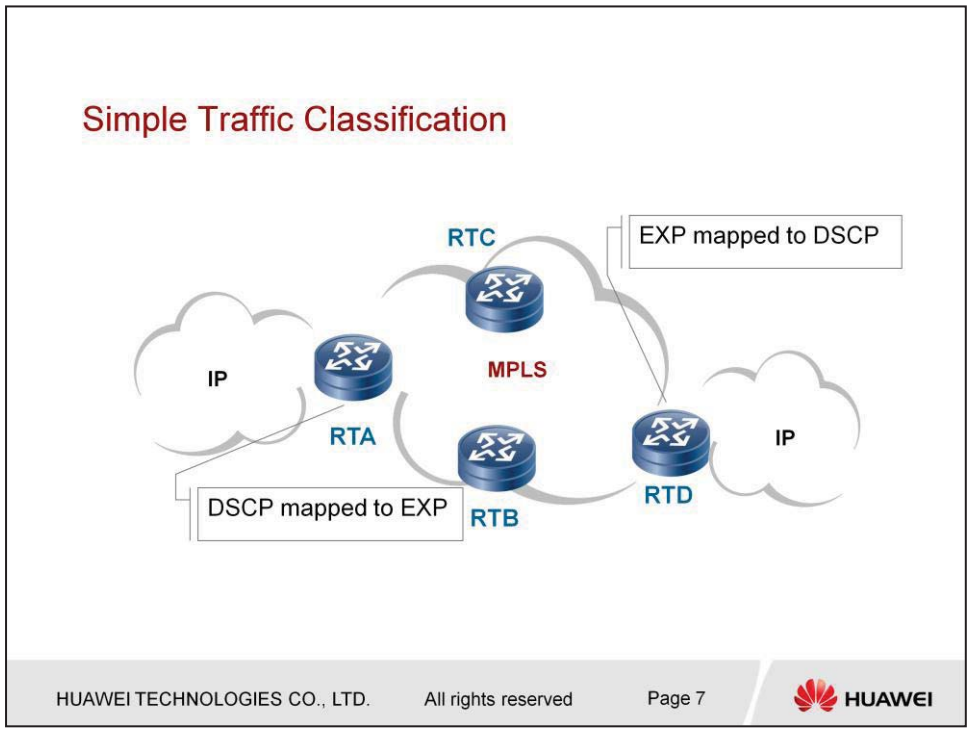



Contents
Traffic Classification
Traffic Policing and Shaping
Congestion Avoidance
HUAWEI TECHNOLOGIES CO., LTD. All rights reserved Page 5
HUAWEI

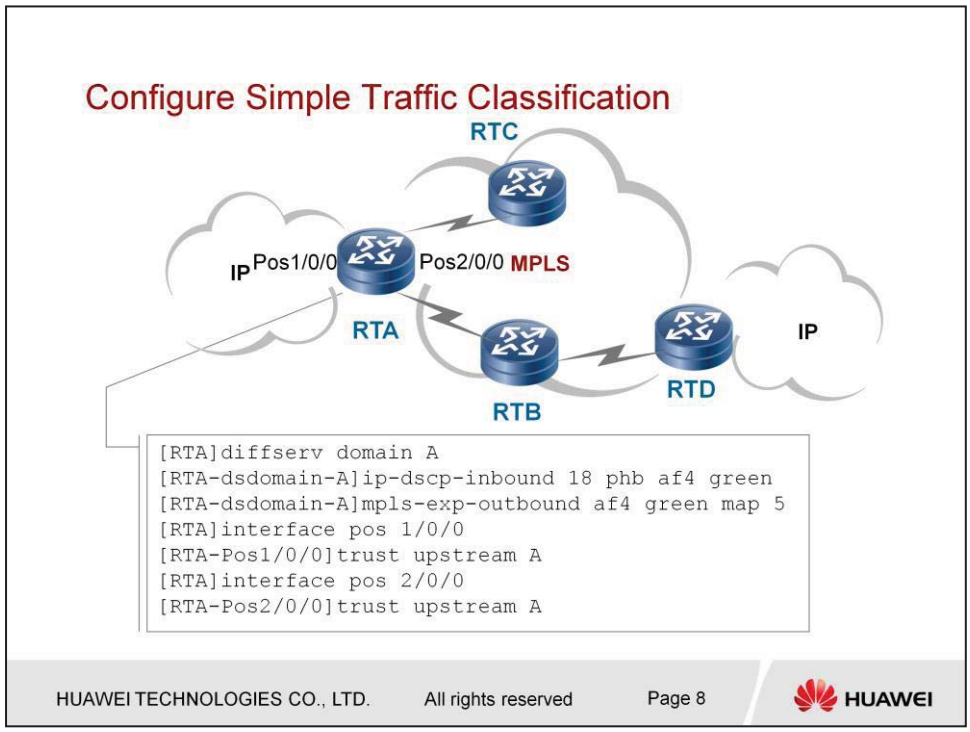


When the Diff-Serv model is adopted for QoS, a router must perform traffic classification to identify flows. QoS implementation varies with the location of a router in the network. The principles and implementation described in this course are applicable to high-end routers. Two traffic classification methods are used: simple traffic classification and complex traffic classification. Simple traffic classification is a process of identifying flows according to the IP precedence, DSCP, MPLS EXP, or 802.1P contained in packets. The packets belonging to the same classifier are called the behavior aggregate (BA). Complex traffic classification is a process of identifying packets according to the quintuple (source address, destination address, source port number, destination port number, and protocol type). Complex traffic classification is generally used on the network edge.

In general, a core router in the Diff-Serv domain needs to perform only simple traffic classification. Simple traffic classification is implemented on high-end routers. Mid-range and low-end routers implement traffic classification based on the traffic policy, meaning they implement complex traffic classification.



In simple traffic classification, packets are classified into various priorities or service classes according to the IP precedence, DSCP value, EXP value (for MPLS packets), or 802.1p value (for VLAN packets). When configuring the traffic policy based on simple traffic classification, you can map the precedence of one traffic type to that of another type so that the traffic can be transmitted based on the original precedence in the network of another type. Application of simple traffic classification When IP, MPLS, and VLAN packets are transmitted between domains, simple traffic classification implements mapping between DSCP, IP precedence, EXP, and 802.1P and keeps the service class of the same flow unchanged. Simple traffic classification is generally configured in the core of a network. Simple traffic application can be applied to both physical interfaces and logical interfaces. In enterprise networks, simple traffic classification is more extensively used on logical interfaces.



In this configuration example, RTA, RTB, RTC, and RTD run MPLS. RTA and RTD are connected to IP networks. Assume that MPLS configuration is complete on the four routers in the MPLS networks. The IP flow from RTA is forwarded to RTD through MPLS. The MPLS flow out of RTD can be transformed to the IP flow and its QoS remains unchanged.

Configure RTA as follows:

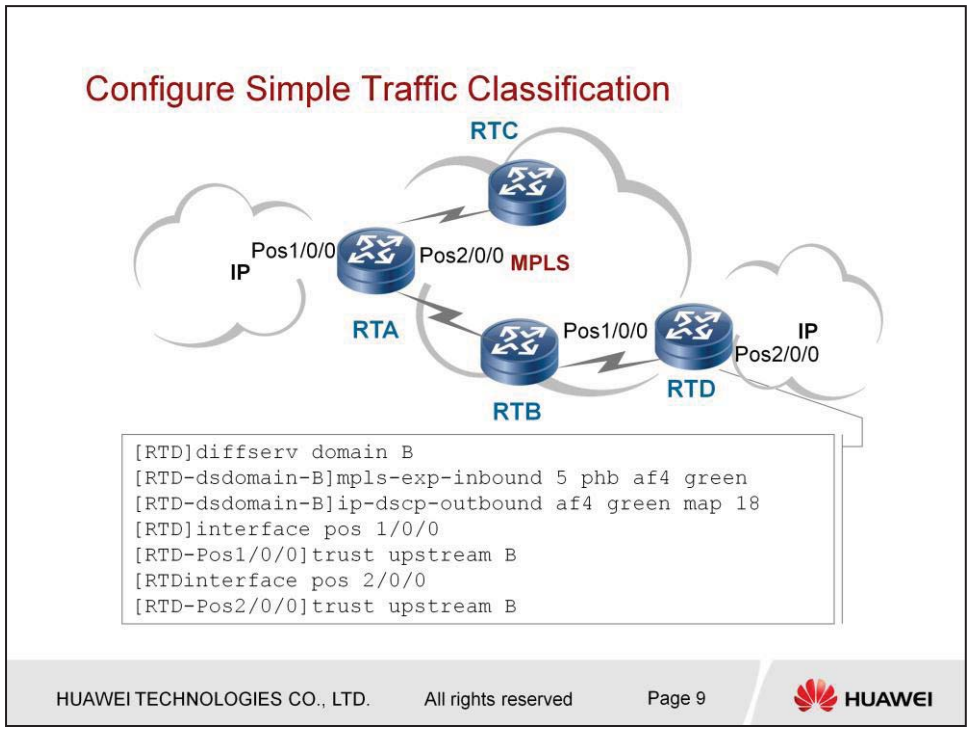
1. Create a Differserv domain.

2. Configure the PHB for IP packets with DSCP value 18 to AF4 and mark such packet in green.

3. Configure the MPLS EXP value for green packets with service class AF4 to 5.

4. Enable the Differserv domain on the IP interface and MPLS interface of RTA.

Using the command **diffserv domain** *ds-domain-name*, you can create a DS domain and enter the DS domain view. Using the command **ip-dscp-inbound** *dscp-value* **phb** *service-class* [*color*], you can configure the service class for a DSCP value of uplink IP packets and mark the packets with a certain color. By default, packets are marked in green. *dscp-value*: specifies the DSCP value of uplink IP packets. The value is an integer ranging from 0 to 63. **phb** *service-class*: specifies the service class. The value can be EF, AF1, AF2, AF3, AF4, BE, CS6, or CS7. *color*: specifies the packet color. The value can be green, yellow, or red. Using command the **mpls-exp-outbound** *service-class color* **map** *exp*, you can configure the EXP value mapping the specified service class and color. *service-class*: specifies the service class. The value can be EF, AF1, AF2, AF3, AF4, BE, CS6, or CS7. *Color*: specifies the packet color. The value can be green, yellow, or red. **map** *exp*: specifies the EXP value of the specified



When IP flows from RTA enter the MPLS network, they are forwarded through MPLS. RTD is connected to an MPLS network and an IP network, and so you need to configure the mapping between MPLS QoS and IP QoS. The configuration procedure is as follows:

1. Create a Differserv domain.

2. Configure the PHB for MPLS packets with EXP 5 to AF4 and mark such packets in green.

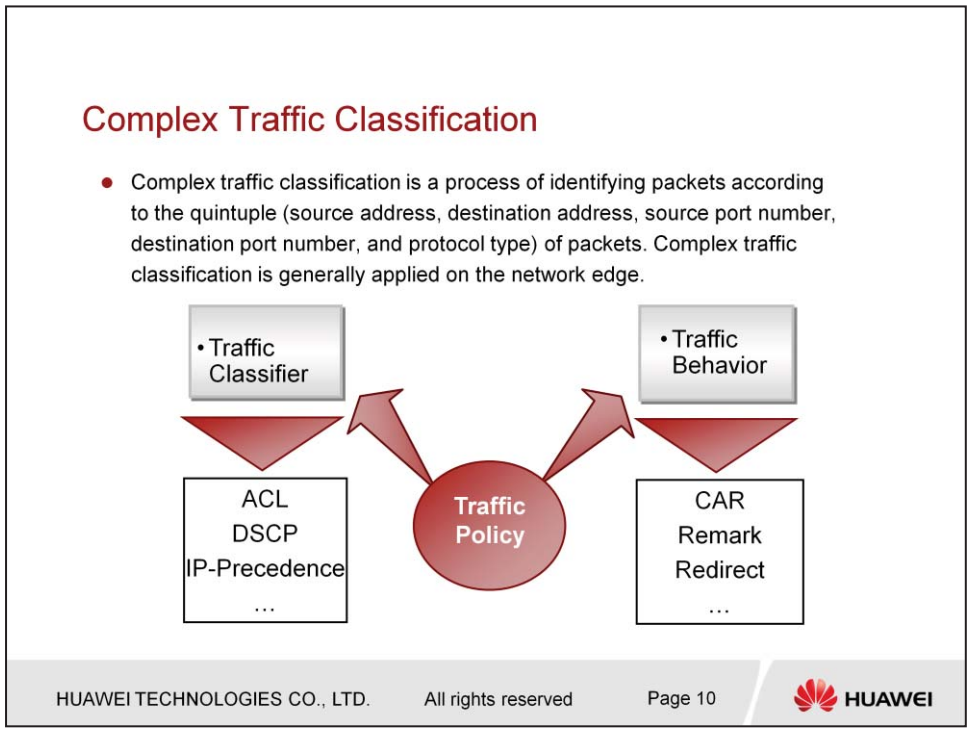
3. Configure the DSCP value for green packets with service class AF4 to 18.

4. Enable the Differserv domain on the IP interface and MPLS interface of RTD. Using the command **mpls-exp-inbound** *exp* **phb** *service-class* [*color*], you can

configure the service class for the MPLS packets with the specified EXP value and mark the packets with a certain color. By default, the color is green. *exp*: specifies the EXP value of MPLS packets. The value is an integer ranging from 0 to 7. **phb** *service-class*: specifies the service class. The value can be EF, AF1, AF2, AF3, AF4, BE, CS6, or CS7. *color*: specifies the packet color. The value can be green, yellow, or red. When the service class is CS6, CS7, EF, or BE, packets can only be marked green.

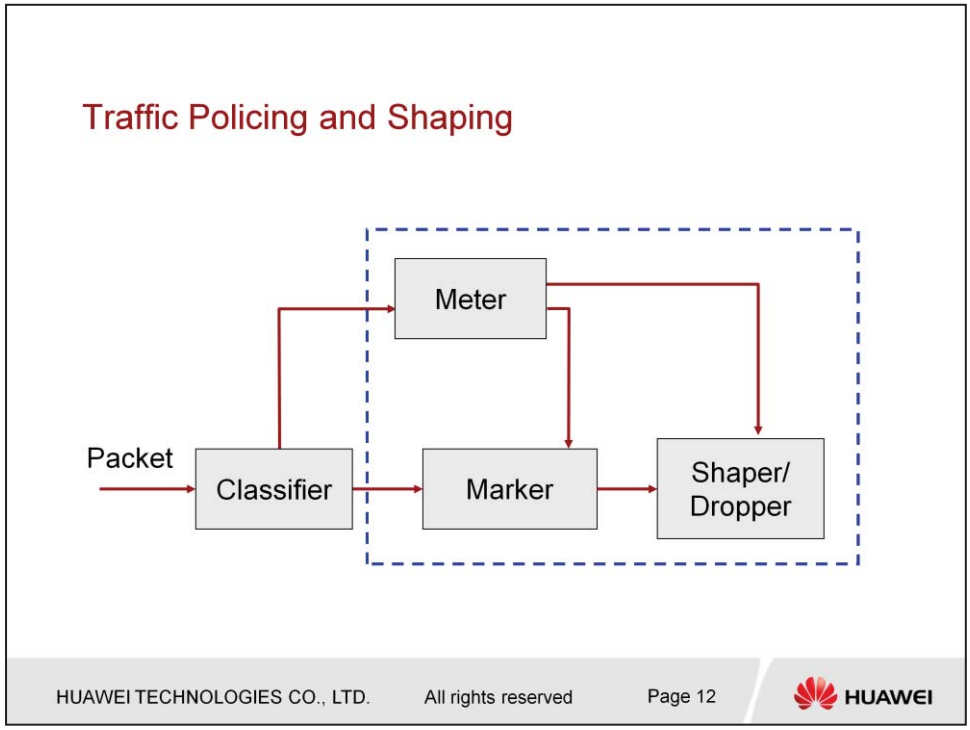
Using the command **ip-dscp-outbound** *service-class color* **map** *dscp-value*, you can configure the DSCP value mapping the specified service class and color of

downlink IP packets. *service-class:* specifies the service class for IP packets. The value can be EF, AF1, AF2, AF3, AF4, BE, CS6, or CS7. *color*: specifies the packet color. The value can be green, yellow, or red. ***map*** *dscp-value:* specifies the DHCP value. The value is an integer ranging from 0 to 63.

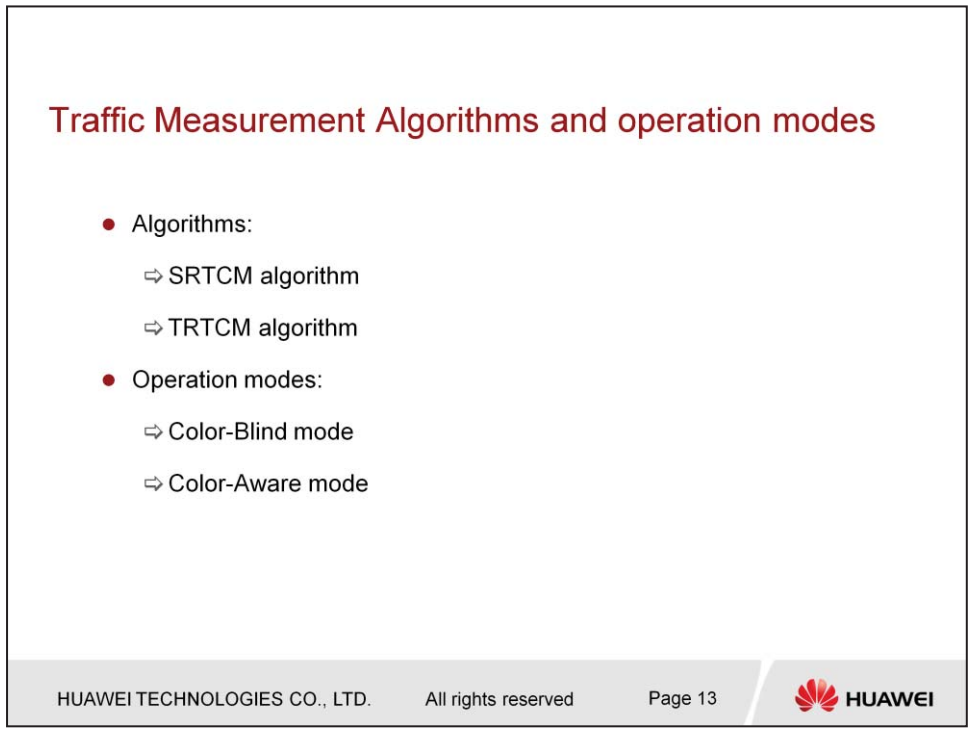


Complex traffic classification is a process of identifying packets according to the quintuple (source address, destination address, source port number, destination port number, and protocol type) of packets. Complex traffic classification is generally applied on the network edge. Traffic classification must be associated with traffic control or resource allocation to provide differentiated service for different services. You can configure the traffic policy to implement this function. In general, complex traffic classification is configured on edge routers, and simple traffic classification is configured on core routers. Complex traffic classification on Huawei high-end routers is the same as the class-based QoS described in the QoS course. Complex traffic classification also consists of the traffic classifier, traffic behavior, and traffic policy. Their configurations are also the same as those described in the QoS course and are not mentioned here.


Contents
Traffic Classification
Traffic Policing and Shaping
Congestion Avoidance
HUAWEI TECHNOLOGIES CO., LTD. All rights reserved Page 11
HUAWEI



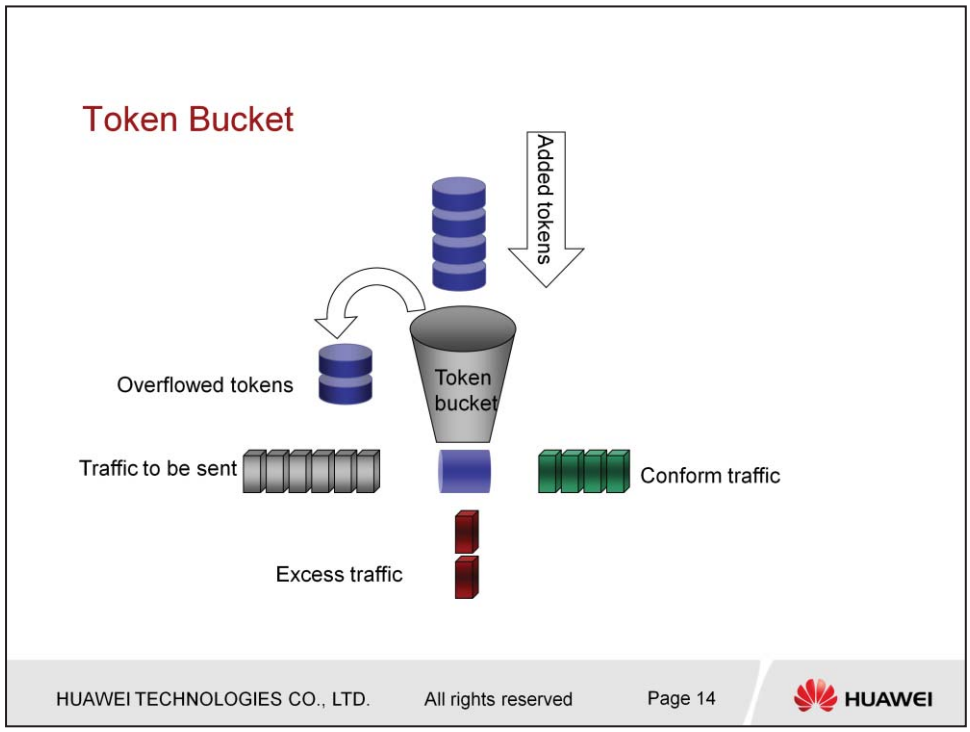
In a Diff-Serv domain, traffic policing and traffic shaping are implemented by the traffic conditioner. The traffic conditioner consists of the Meter, Marker, Shaper, and Dropper, as shown in the figure. The Meter is used to measure the traffic and judge whether the information flow conforms to the traffic specification defined in the TCS. According to the measurement, a router can perform the behavior such as marking, shaping and dropping. The Marker is used to re-mark the priority of the packets and adds the re-marked packets to specified BAs. The behavior can lower the service level of the flow that is out of profile according to the SLS or keeping the service class. The Shaper has a buffer for received packets. The buffer controls the traffic to be sent within the committed specification. The Dropper is a behavior for traffic policing. It drops some packets to control the traffic within the specification. You can configure the Dropper by setting the buffer of the Shaper to 0 or setting the buffer to a small value.



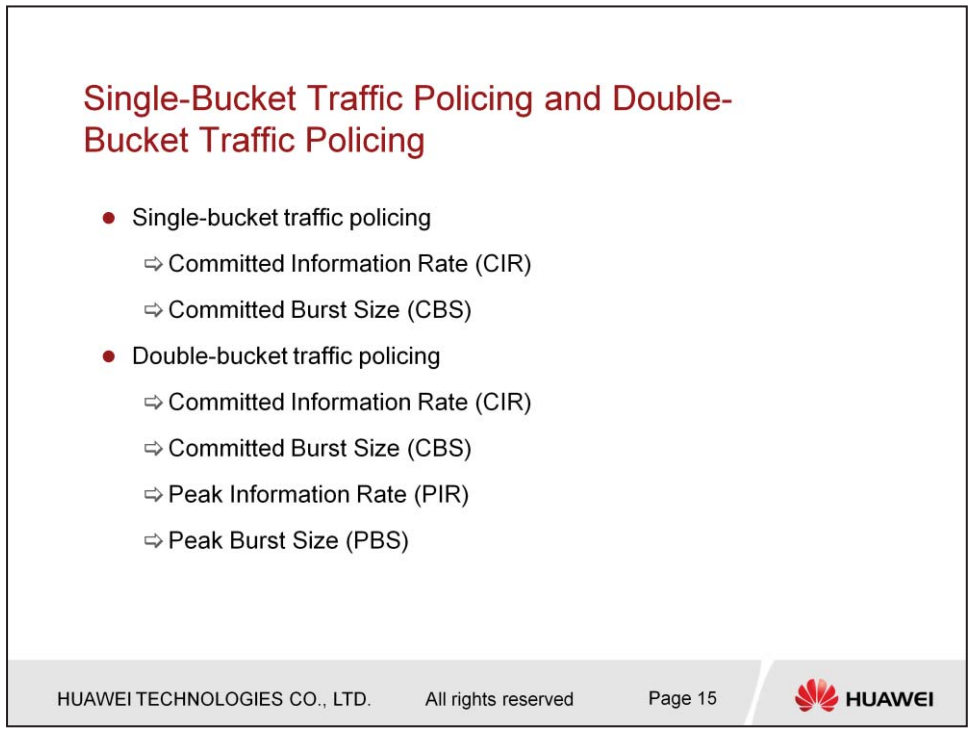
The srTCM (Single Rate Three Color Marker) algorithm is defined in RFC 2697, and the trTCM (Two Rate Three Color Marker) algorithm is defined in RFC2698. These algorithms are used to measure the traffic. Packets are marked in green, yellow, or red according to the measurement. The srTCM and trTCM both use two token buckets to assess packets. They allow burst traffic at a certain level. The srTCM focuses on burst of the packet size, while the trTCM algorithm focuses on burst of the rate. Both the srTCM algorithm and the trTCM algorithm have two operation modes:

Color-Blind mode and Color-Aware mode.

In color-Blind mode, the Meter assumes that packets are uncolored. In color-Aware mode, the Meter assumes that packets are pre-colored. The Color-Blind mode is more often used. The srTCM algorithm or the trTCM algorithm marks packets differently in different modes.



A token bucket judges whether the traffic exceeds the specified value. If the traffic exceeds the specified value, certain measures are carried out. In the bucket there are tokens instead of packets. Tokens are added to the token bucket at intervals. When the token bucket is full, the subsequent tokens are discarded. A token ensures that one or more bits/bytes can be forwarded. When a packet arrives, if there are enough tokens in the bucket, the packets are forwarded. At the same time, the number of tokens decreases accordingly. If tokens in the bucket are not enough for sending a packet, the packet is discarded and the number of tokens does not change. Depending on whether tokens in the bucket are enough for sending packets, the result of traffic assessment can be conformed or in excess.



Traffic policing can be classified into single-bucket traffic policing and double-bucket traffic policing.

Single-bucket traffic policing

Single-bucket traffic policing is used when traffic measurement is relatively simple. In this mode, the traffic type depends on whether tokens in the bucket are enough for forwarding packets. If there are enough tokens in the bucket, the traffic conforms to the specification, and the packets are marked in green. If there are not enough tokens in the bucket, the traffic is considered excess traffic, and the packets are marked in red. In single-bucket traffic policing, two parameters are used:

1. Committed Information Rate (CIR): indicates the speed at which tokens are added to the bucket, namely, the allowed average rate.

2. Committed Burst Size (CBS): indicates the capacity of the token bucket, namely, the maximum traffic size allowed in a burst. The CBS must be greater than the length of the largest packet.

Double-bucket traffic policing

In more complex cases, two token buckets can be used to carry out more flexible control policies.

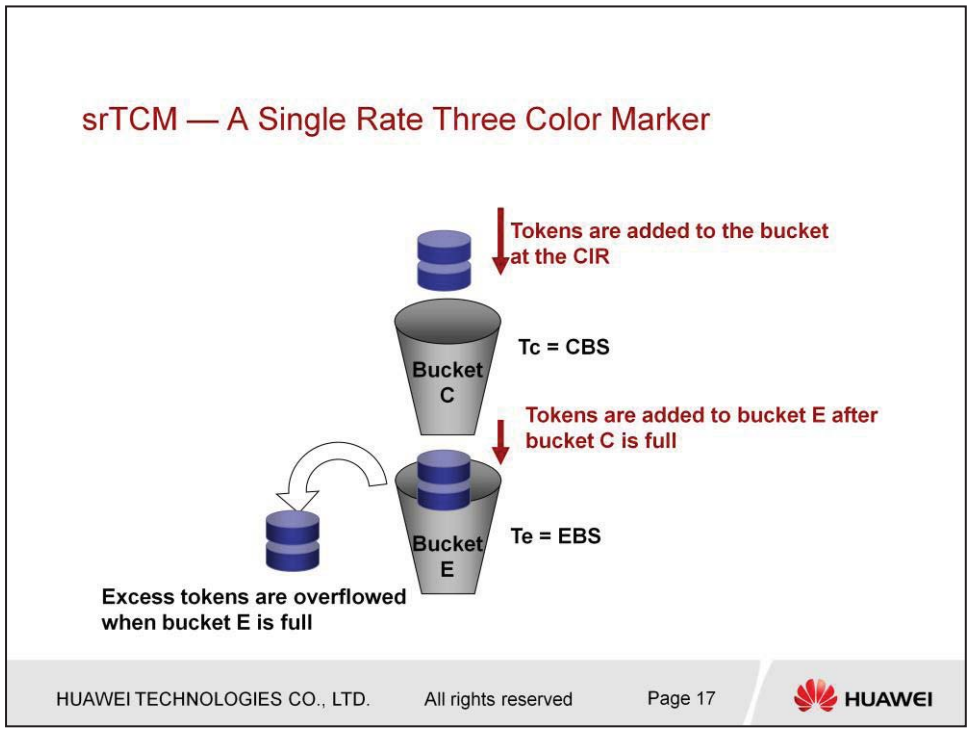
In this case, the following traffic parameters are used:

1. Committed Information Rate (CIR): indicates the speed at which tokens are added to bucket C, namely, the average rate allowed by bucket C.

2. Committed Burst Size (CBS): indicates the capacity of bucket C, namely, the maximum burst traffic size allowed by bucket C.

3. Peak Information Rate (PIR): indicates the speed at which tokens are added to bucket P, namely, the average rate allowed by bucket P.

4. Peak Burst Size (PBS): indicates the capacity of bucket P, namely, the maximum burst traffic size allowed by bucket P. Method of measuring and marking traffic varies with the algorithm used and the operation mode of the algorithm. The srTCM and trTCM algorithms both assess received packet by using two buckets.

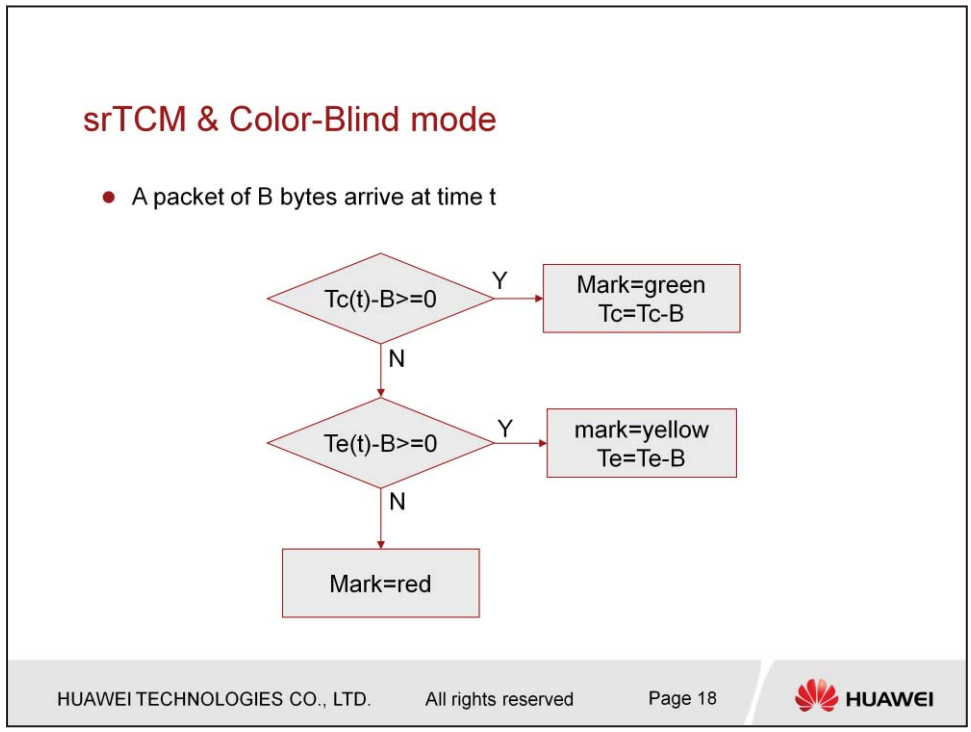


srTCM (A Single Rate Three Color Marker) uses two buckets to assess the traffic. Tokens are added to the two buckets at the Committed Information Rate (CIR). The sizes of the two buckets are respectively Committed Burst Size (CBS) and Excess Burst Size (EBS). For convenience of description, we call the two buckets bucket C and bucket E. The numbers of tokens in the bucket C and bucket E are represented by Tc and Te. Initially, Tc is CBS, and Te is EBS. Tc and Te are updated CIR times in a second. The update observes the following principles:

If Tc < CBS, Tc is increased by 1 each time.

If Tc >= CBC and Te < EBS, Te is increased by 1 each time. Otherwise, neither Tc nor Te is increased.

The srTCM algorithm is based on parameters CIR, CBS, and EBS. If only these parameters are set, it is considered that the srTCM algorithm is used.



As mentioned before, in Color-Blind mode, the Meter assumes that the packets are uncolored. When a packet of B bytes arrives at time t, the Marker marks the packet as follows:

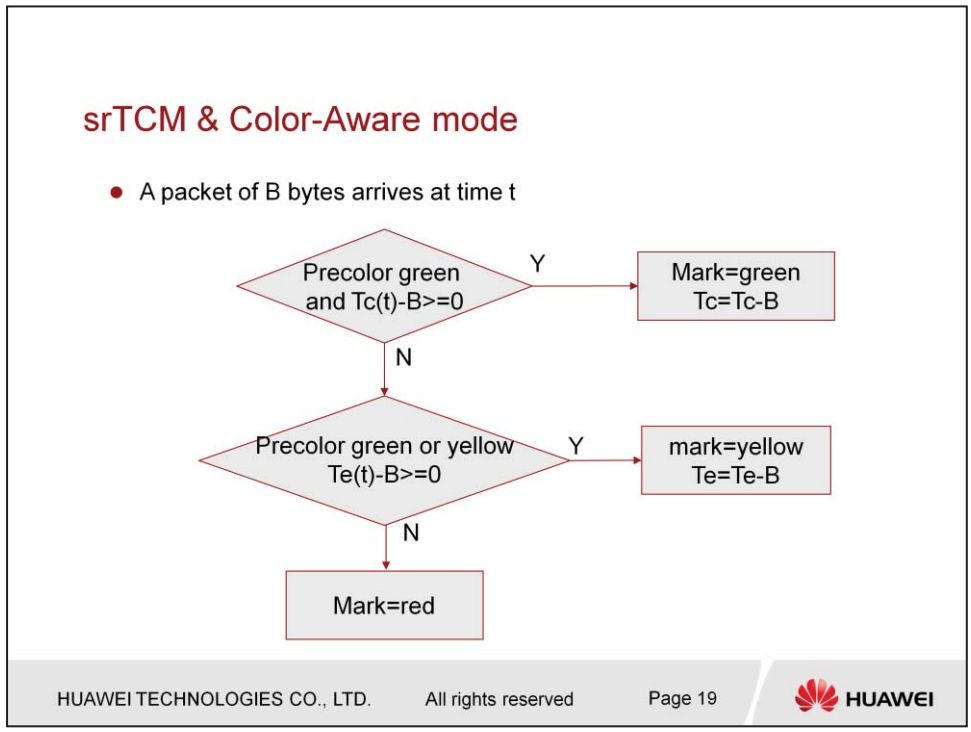
If Tc(t) - B >= 0, mark = green Tc = Tc – B; else

If Te(t) - B >= 0, mark = yellow Te = Te – B; else

mark = red

At time t, Tc(t)-B>=0 means that there are enough tokens in bucket C. In this case, the packet is marked in green. At the same time, the number of tokens in bucket C decreases by B; that is, Tc=Tc-B. If bucket C does not have enough tokens, the system checks bucket E. If Te (t)-B>=0, it indicates that tokens in bucket E are enough for forwarding the packet. In this case, the packet is marked in yellow and the number of tokens in bucket E decreases by B (Te=Te-B). If neither bucket C nor bucket E has enough tokens,

the packet is marked in red and the number of tokens does not decrease.



When the srTCM is configured to operate in Color-Aware mode, the Marker assumes that the packets are pre-colored. The 4th and 5th bits of the 6-bit DSCP field indicate the packet color. 00 and 01 indicate green; 10 indicates yellow; 11 indicates red. As shown in the figure, when a packet of B bytes arrives at time t, the Marker marks the packet as follows:

If the packet is precolored green and TC(t) - B >= 0

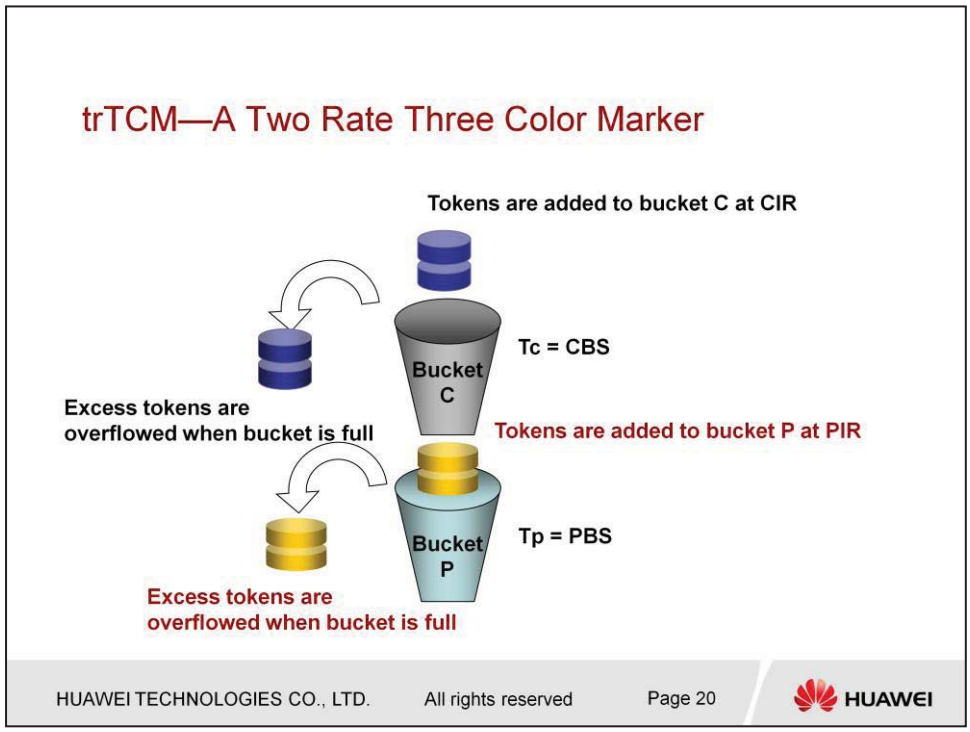
mark = green, Tc =Tc – B; else

If the packet is precolored green or yellow and Te(t) - B >= 0

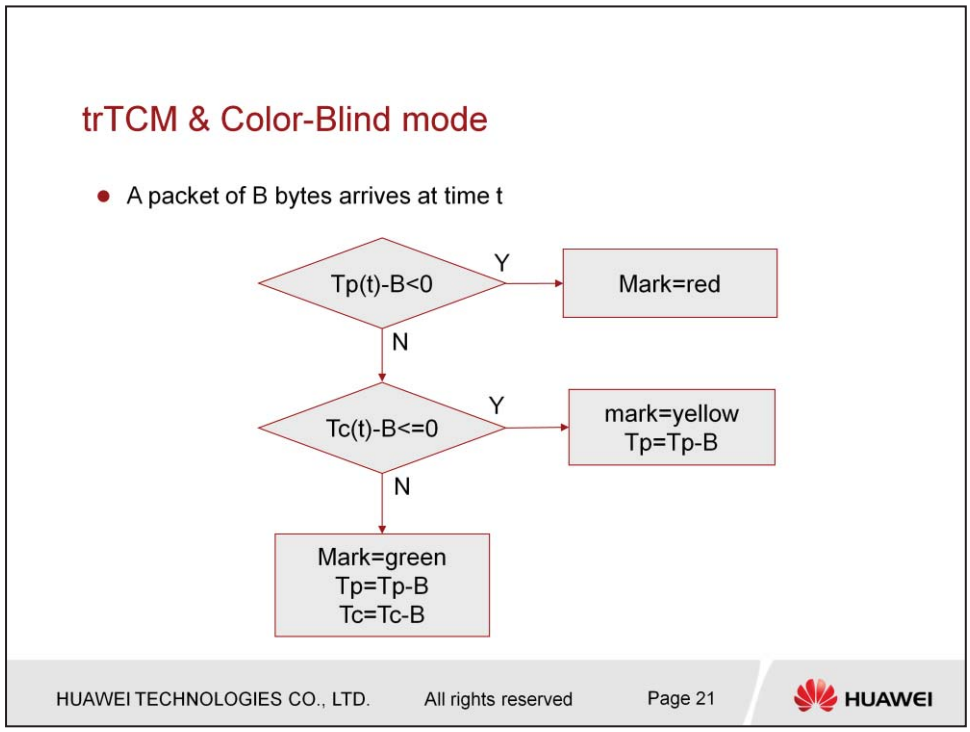
mark = yellow, Te = Te – B; else

mark = red

When the packet is precolored green and TC(t) - B >= 0, that is, bucket C has enough tokens to forward the packet, the packet is re-marked in green. At the same time, the number of tokens in bucket C decreases by B (mark = green, Tc =Tc – B). If tokens in bucket C are not enough for forwarding the green packet, or the packet is precolored yellow, and token E has enough tokens, the packet is remarked in yellow. At the same time, the number of tokens in bucket E decreases by B. If tokens in bucket E are not enough for forwarding the packet either, the packet is re-marked in red, regardless of the original color of this packet. In this case, the number of tokens does not decrease.



In the trTCM (A Two Rate Three Color Marker), tokens are added to the two buckets at different rates: Committed Information Rate (CIR) for one bucket and Peak Information Rate (PIR) for another. For convenience of description, the two buckets are called bucket C and bucket P. their sizes are respectively the Committed Burst Size (CBS) and Peak Burst Size (PBS). Tc and Tp represent the number of tokens in the two buckets. Initially, Tc is CBS and Tp is PBS. Tc is increased by CIR with the ceiling CBS every second and Tp is increased by PIR with the ceiling PBS every second. Same as the srTCM algorithm, in the trTCM algorithm, a token is added each time Tc or Tp is updated (till the bucket is full). The details are omitted here. The trTCM algorithm is based on parameters CIR, PIR, CBS, and PBS. If CIR, PIR, CBS, and PBS are set, it is considered that the trTCM is used.



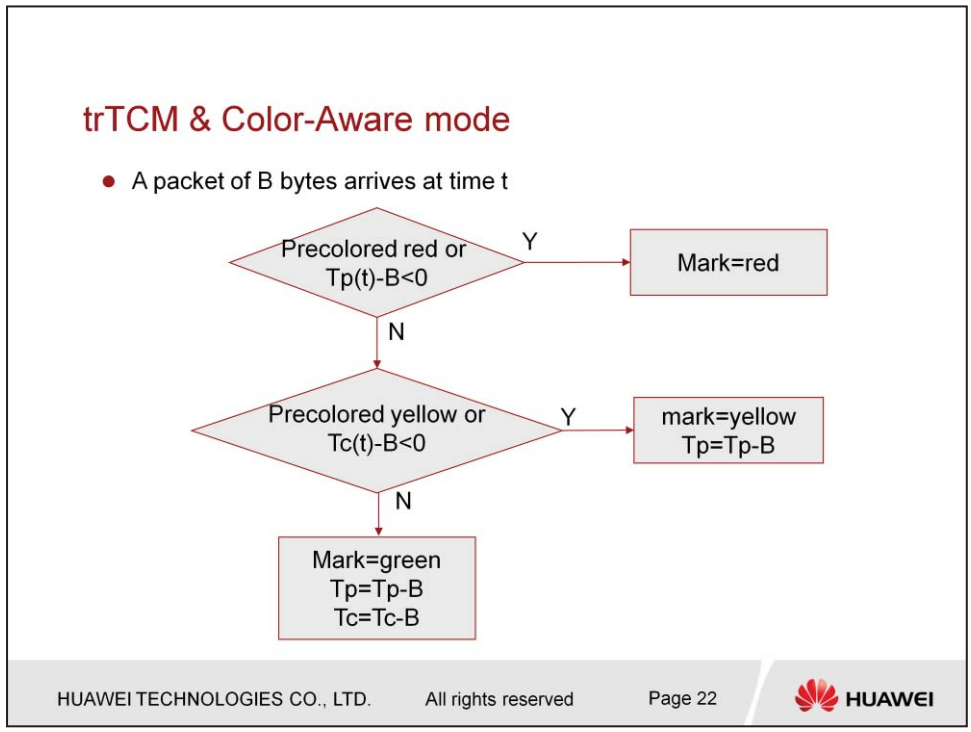
When the trTCM is configured to operate in Color-Blind mode, the Meter assumes that packets are uncolored. When a packet of B bytes arrives at time t, the Marker marks the packet as follows:

If Tp(t)-B<0, mark = red; else

If Tc(t)-B<=0, mark = yellow, Tp = Tp – B; else

mark = green ,Tp = Tp - B, Tc = Tc - B

If bucket P does not have enough tokens, the packet is marked in red. If bucket C does not have enough tokens but bucket P has enough tokens, the packet is marked in yellow. At the same time, the number of tokens in bucket P decreases by B. That is Tp=Tp-B. If bucket P and bucket C both have enough tokens, the packet is marked in green. At the same time, the number of tokens decreases by B in both buckets. That is, Tp=Tp-B, Tc=Tc-B. You can see that packets metering and marking starts at bucket P in the trTCM algorithm.



When the trTCM is configured to operate in Color-Aware mode, the Marker assumes that packets are precolored. The color of a packet is represented by the last two bits of the 6-bit DSCP field. 00 and 01 indicate green; 10 indicates yellow; 11 indicates red. As shown in the figure, when a packet of B bytes arrives at time t, the Marker marks the packet as follows:

If the packet is precolored red or Tp(t) - B < 0

mark = red;

else

If the packet is precolored yellow or Tc(t) - B < 0

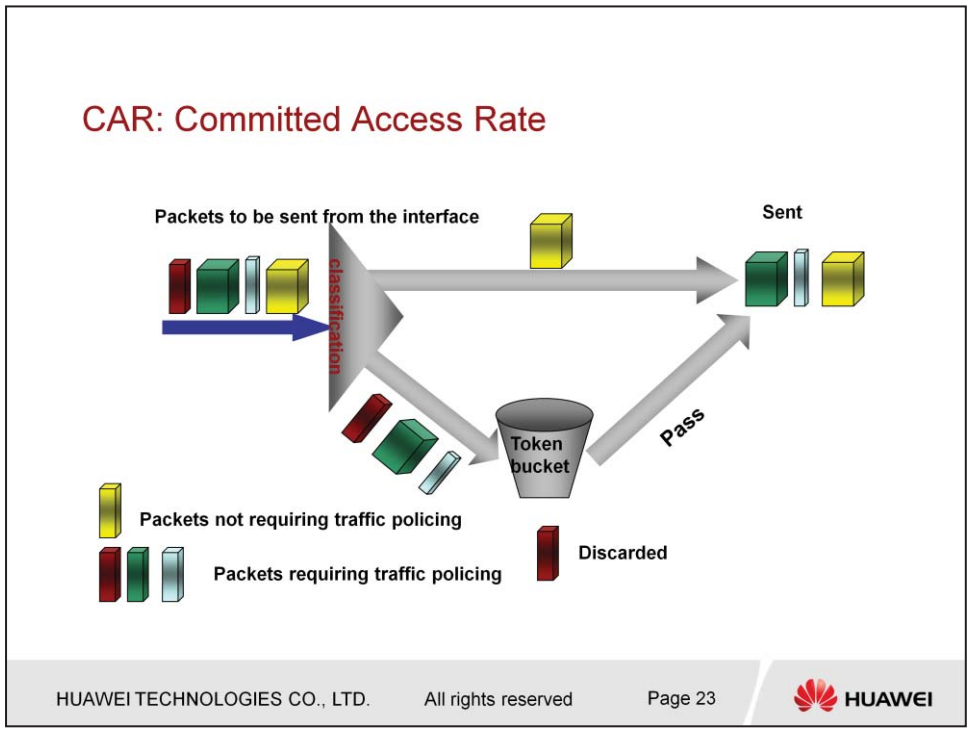
mark = yellow , Tp = Tp – B;

else

mark = green, Tp = Tp – B, Tc = Tc - B

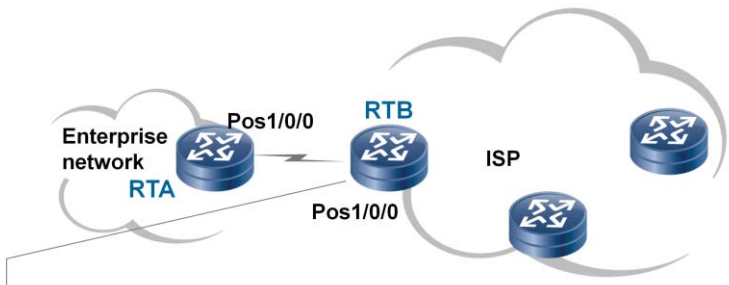
When the packet is precolored red or there are not enough tokens in bucket P, the packet is re-marked in red. When the packet is precolored yellow or there are not

enough tokens in bucket C, the packet is re-marked in yellow. At the same time, the number of tokens in bucket P decreases by B. If there are enough tokens in bucket C to forward the packet or the packet is precolored green, the packet is remarked in green. The number of tokens in bucket P and number of tokens in bucket P both decrease by B.



The CAR implements traffic policing through the token bucket. CAR can implement traffic policing for specified traffic. In traffic policing, pre-defined behaviors are carried out according to the assessment result. The following behaviors can be pre-defined: pass: forwards the packets for which the assessment result is conform. Such packets can also be marked with DSCP field and then forwarded. discard: discards the packets for which the assessment result is excess. remark: Re-marks the precedence of the excess packets, and then forwards them.

Configure Interface-based Traffic Policing




```
<RTB>system-view
[RTB]interface pos1/0/0
[RTB-pos1/0/0]qos car cir 100 pir 10000 green pass yellow
pass red discard inbound
```

HUAWEI TECHNOLOGIES CO., LTD. All rights reserved HUAWEI

Interface-based traffic policing controls traffic of all packets received by an interface but does not concern the packet type. Interface-based traffic policing is generally applied to core routers. In this example, CAR is configured on POS1/0/0 of RTB to implement traffic policing for all packets received on POS1/0/0. The CIR is 100kbit/s; the PIR is 10000kbit/s. Green and yellow packets are forwarded and red packets are discarded. The configuration procedure is as follows:

1. Run the command **system-view** to enter the system view.

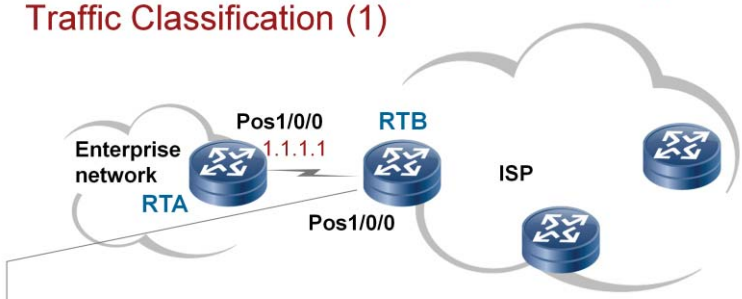
2. Run the command **interface** *interface-type interface-number* to enter the interface view.

3. Run the command **qos car cir** *cir-value* [**pir** *pir-value*] [**cbs** *cbs-value* **pbs** *pbs-value*] [**green** { **discard** | **pass** } [**yellow** { **discard** | **pass** } [**red** { **discard**

pass }]]] { **inbound** | **outbound** } to configure CAR on the interface. cir cir-value: specifies the Committed Information Rate (CIR), namely, the rate at which packets are ensured to pass through the interface. The value is an integer ranging from 100 to 10000000, in kbit/s. pir pir-value: specifies the Peak Information Rate (PIR), namely, the maximum rate allowed by the interface. The value is an integer ranging from 100 to 10000000, in kbit/s. The value of pir-value cannot be smaller than the value of cirvalue. cbs cbs-value: specifies the Committed Burst Size (CBS), namely, the instant committed traffic. The value is the capacity of the first bucket (assume it is bucket A). It is an integer ranging from 64 to 33554432, in bytes. The default value relates to the cir-value. If cir-value<=10000 kbit/s, the default value of cbs-value is 10000 byte. If cir-value>10000 kbit/s, the default value of cbs-value equals cir-value, in bytes. pbs pbs-value: specifies the Peak Burst Size (PBS), namely, the instant peak traffic. The value is the capacity of the second bucket (assume it is bucket B). It is an integer

ranging from 0 to 33554432, in bytes. The default value relates to pirvalue. If pir-value is not specified, the default value of pbs-value is 0. If pir-value is specified and pir-value<10000kbit/s, the default value of pbs-value is 10000 bytes. If pir-value is specified and pir-value>=10000 kbit/s, the default value of pbs-value is pir-value, in bytes. pass, discard: specifies the behavior for packets with the specified color (green, yellow, or red). green: specifies the behavior carried out when the traffic is within the CIR. The default value is pass. yellow: specifies the behavior carried out when the traffic exceeds the CIR but is less than the PIR. The default value is pass. red: specifies the behavior carried out when the traffic exceeds the PIR. The default value is discard. inbound: performs traffic policing in the inbound direction. outbound: performs traffic policing in the outbound direction. In this example, two CAR parameters are set: CIR and PIR. If CBS and PBS are not set, their default values are used. The trTCM algorithm is based on the CIR, PIR, CBS, and PBS, and so the trTCM algorithm is used in this example and it operates in Color-blind mode.

Configure Traffic Policing Based on Complex Traffic Classification (1)




```
[RTB] acl number 2001
[RTB-acl-basic-2001] rule permit source 1.1.1.1 0.0.0.0
[RTB] traffic classifier aa
[RTB-classifier-a] if-match acl 2001
[RTB] traffic behavior bb
[RTB-behavior-a] car cir 5000 pir 6000 green pass yellow pass red discard
[RTB-behavior-a] quit
```

HUAWEI TECHNOLOGIES CO., LTD. All rights reserved HUAWEI

In complex traffic classification, packets are classified based on the quintuple (source address, source port number, protocol number, destination address, destination port number). Normally, complex traffic classification is applied to the network edge. To control the traffic of packets of only one or more types, you can combine complex traffic classification with traffic policing. That is, configure the traffic policing policy based on complex traffic classification and apply the policy to the interface.

In this example, RTB controls the traffic of packets from 1.1.1.1. The CIR is 5000 kbit/s and PIR is 6000 klbit/s. Green and yellow packets are forwarded and red packets are discarded.

The procedure for configuring traffic policing based on complex traffic classification is as follows:

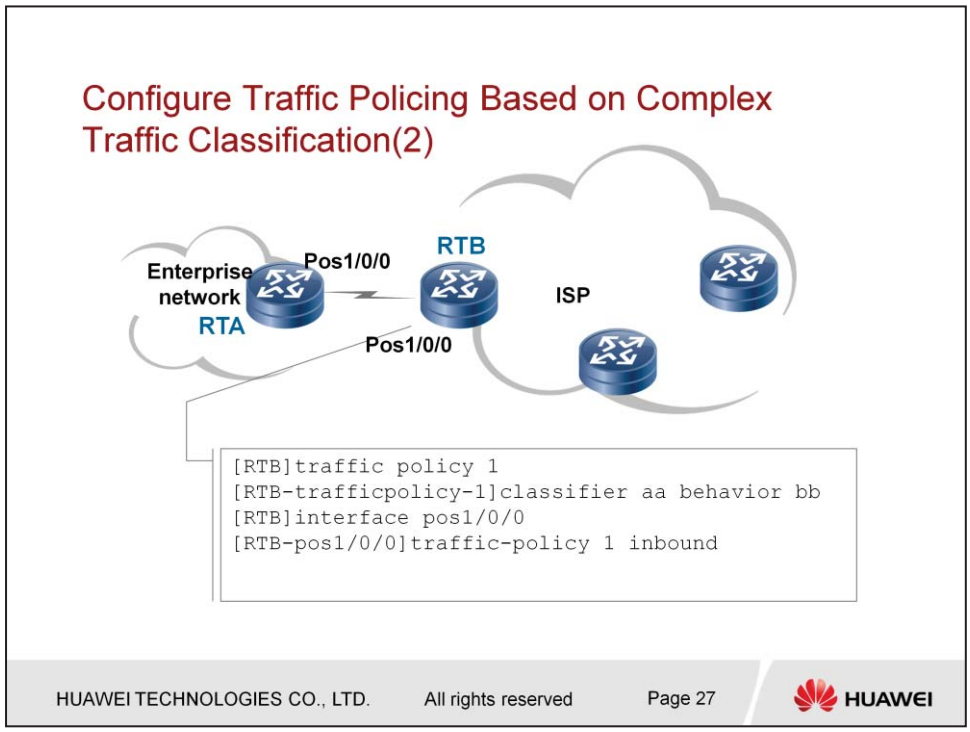
1. Run the command **system-view** to enter the system view.

2. Run the command **traffic classifier** *classifier-name* [**operator** { **and** | **or** }] to define a classifier and enter the classifier view.

3. Define the matching rule on the router based on the actual application. In this example, the router matches packets according to the source IP address of packets.

4. command Run the **traffic behavior** *behavior–name* to enter the behavior view.

5. Run the command **car cir** *cir-value* [**pir** *pir-value*] [**cbs** *cbs-value* **pbs** *pbsvalue*] [**green** { **discard** | **pass** } [**yellow** { **discard** | **pass** } [**red** { **discard** | **pass** }]]] to configure the traffic policing behavior.



6. Run the command **traffic policy** *policy-name* to configure the traffic policy and enter the traffic policy view.

7. Run the command **classifier** *classifier-name* **behavior** *behavior-name* to configure the behavior for the classifier.

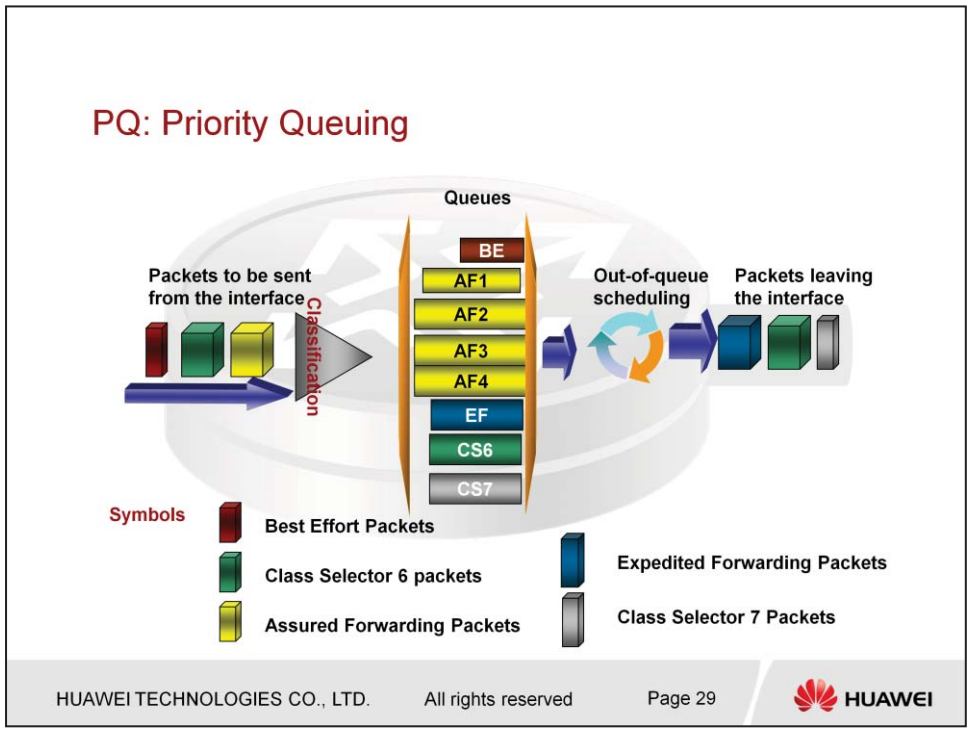
8. Run the command **interface** *interface-type interface-number* to enter the interface view.

9. Run the command **traffic-policy** *policy-name* { **inbound** | **outbound** } [**linklayer**] to apply the traffic policy on the interface.

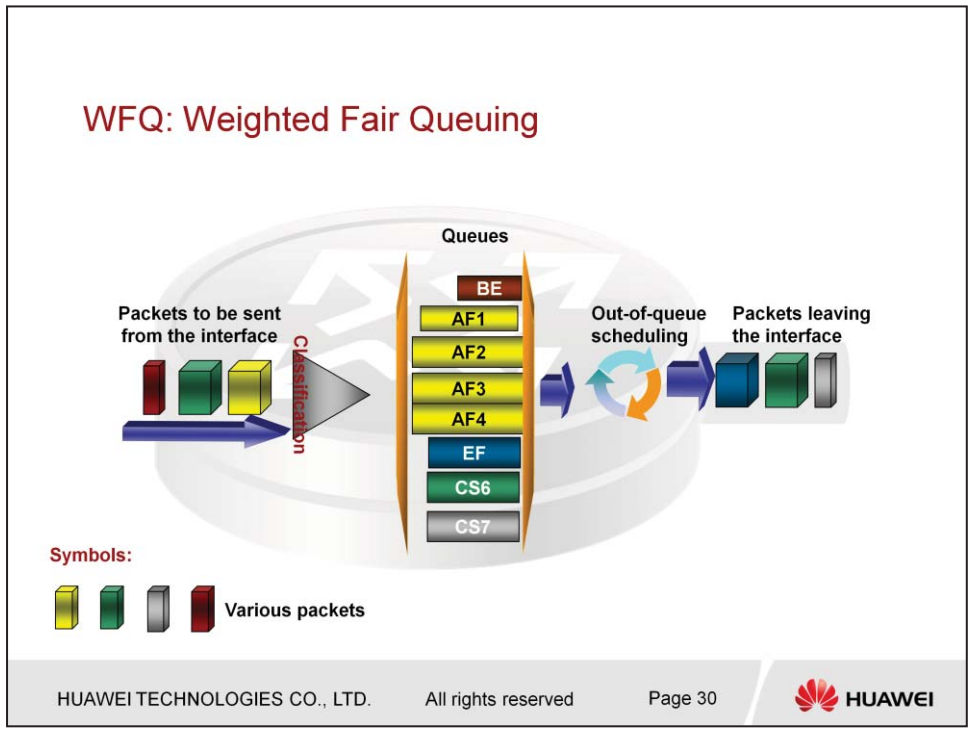
Congestion Management

- Service classes supported by an interface:
  - ⇨ BE, AF1, AF2, AF3, AF4, EF, CS6, CS7
- Queue scheduling algorithms:
  - ⇨ FIFO
  - ⇨ PQ
  - ⇨ WFQ

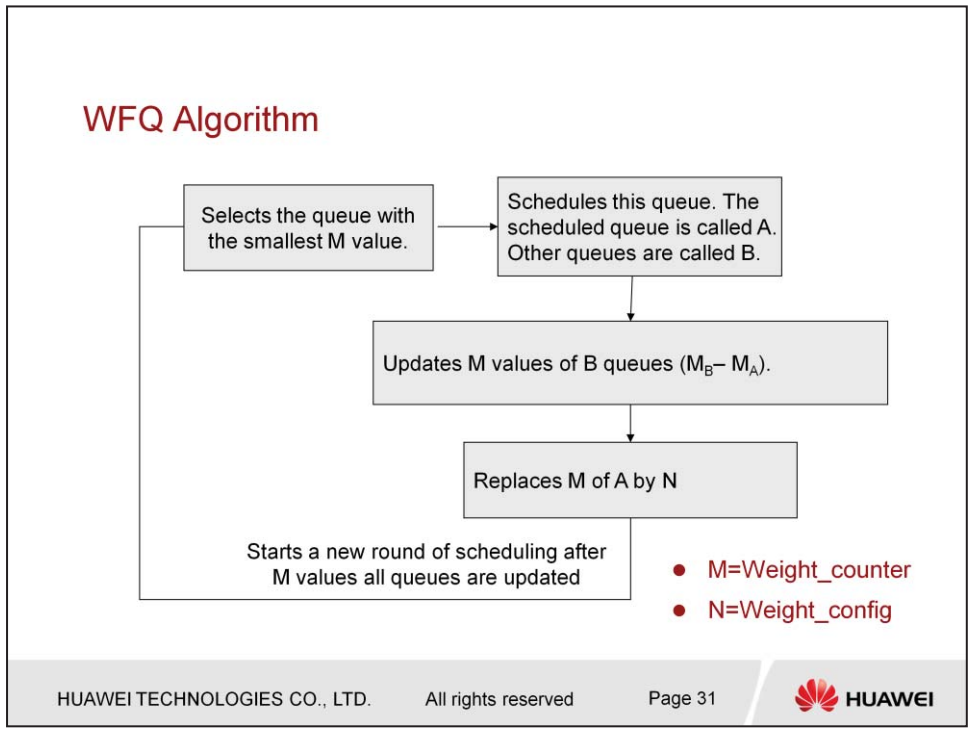
HUAWEI TECHNOLOGIES CO., LTD. All rights reserved



High-end routers of Huawei support eight service classes, which can use the PQ or WFQ algorithm. By default, the PQ algorithm is applied to the EF, CS6, and CS7 classes. You can also configure all the eight service classes to used the PQ algorithm. Packets in the same queue are scheduled with FIFO algorithm. PQ is based on the strict priority (SP) algorithm. Queues with higher priority are scheduled first. The priority sequence of the eight service classes is CS7, CS6, EF, AF4, AF3, AF2, AF1, BE, from the highest priority to the lowest. PQ always schedules the packet of class CS7 first. The queue with the second higher priority can be schedule only when the CS7 queue is empty. Once packets are added to the queue with higher priority, PQ will schedule that queue. Therefore, packets in queues with lower priorities may be starved. That is, queues with lower priorities cannot be scheduled as long as there are packets in the queue with the highest priority. Due to the problem of packets starvation, PQ usually cooperates with another scheduling algorithm, namely, WFQ. In this way, queues with lower priorities can also be scheduled fairly.

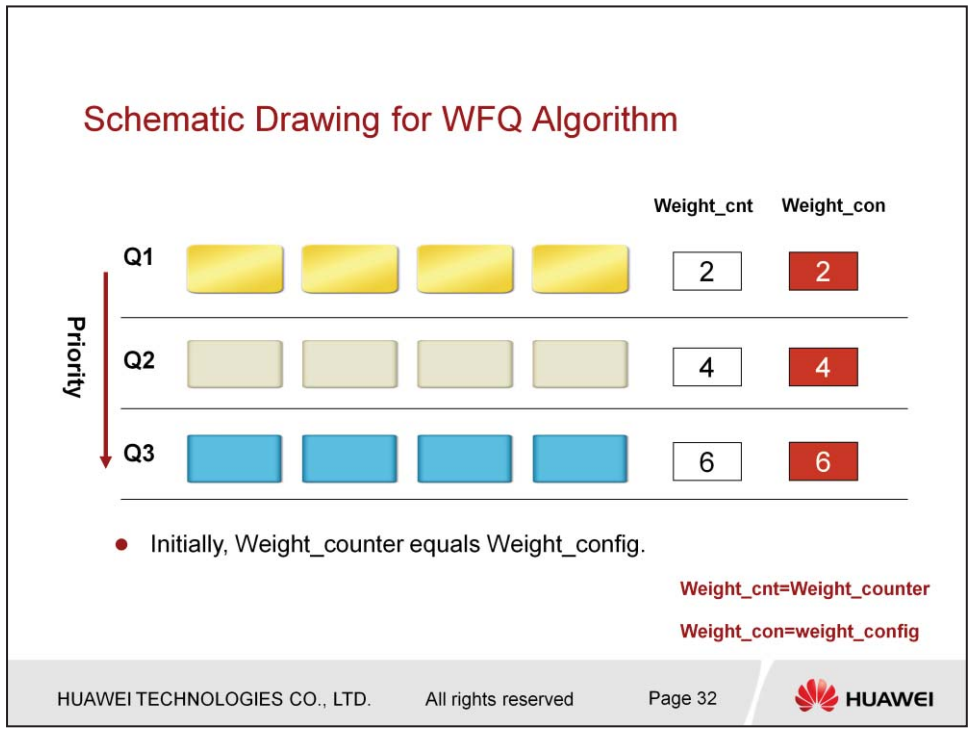


Weighted fair queuing (WFQ) is described in the course of IP QoS. The implementation of WFQ is a little different on high-end routers of Huawei. An interface supports eight service classes: BE, AF1, AF2, AF3, AF4, EF, CS6 and CS7. By default, CS7, CS6, and EF use the PQ algorithm, and BE, AF1, AF2, AF3, and AF4 use the WFQ algorithm. WFQ is based on the weight of each queue. The queues with smaller weight are more likely to be scheduled. The weight of a queue is configured through the CIR parameter in the **qos queue** command. By default, weighs of the eight classes are :10%, 10%, 10%, 15%, 15%, 10%, 5%, 5%.

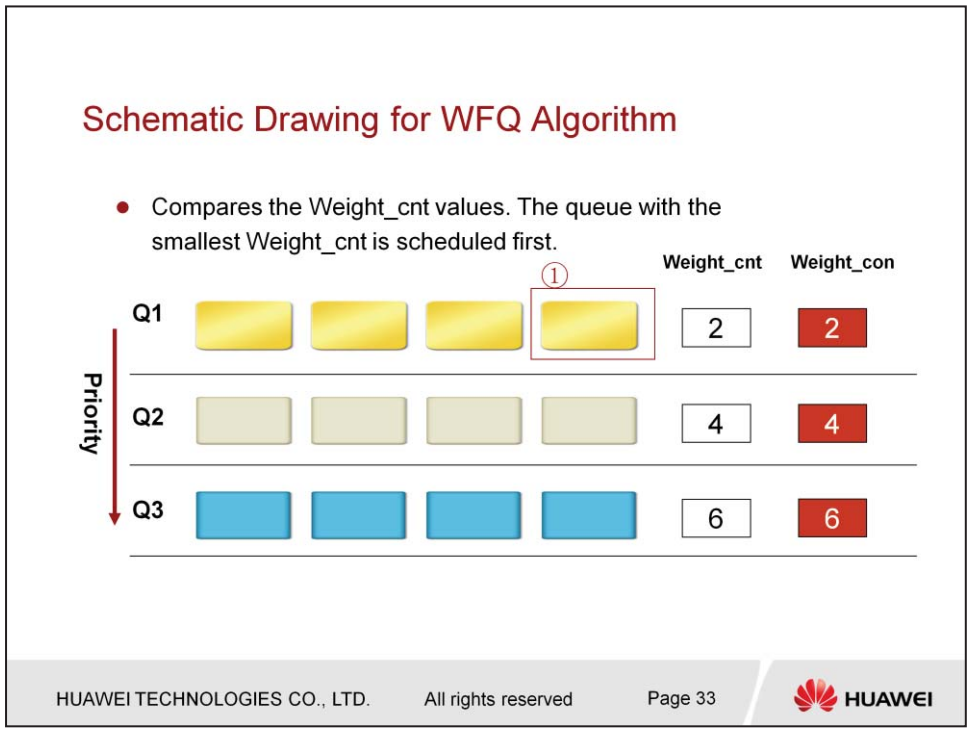


The WFQ algorithm involves two parameters: Weight_counter (weigh counter of a queue) and Weight_config (configured weight of a queue). This page describes the process of scheduling queues on an interface through WFQ. Here, M indicates the Weight_counter of a queue, and N indicates the weight_config of a queue. First, WFQ selects the queue with the smallest Weight_counter. This is the first step in the flowchart (selects the queue with the smallest M value). The queue with the smallest M value is scheduled first. Only one queue is scheduled at a time. The scheduled queue is called A, and other queues are called B. When A queue is scheduled, Weight_counter values of B queues are updated.

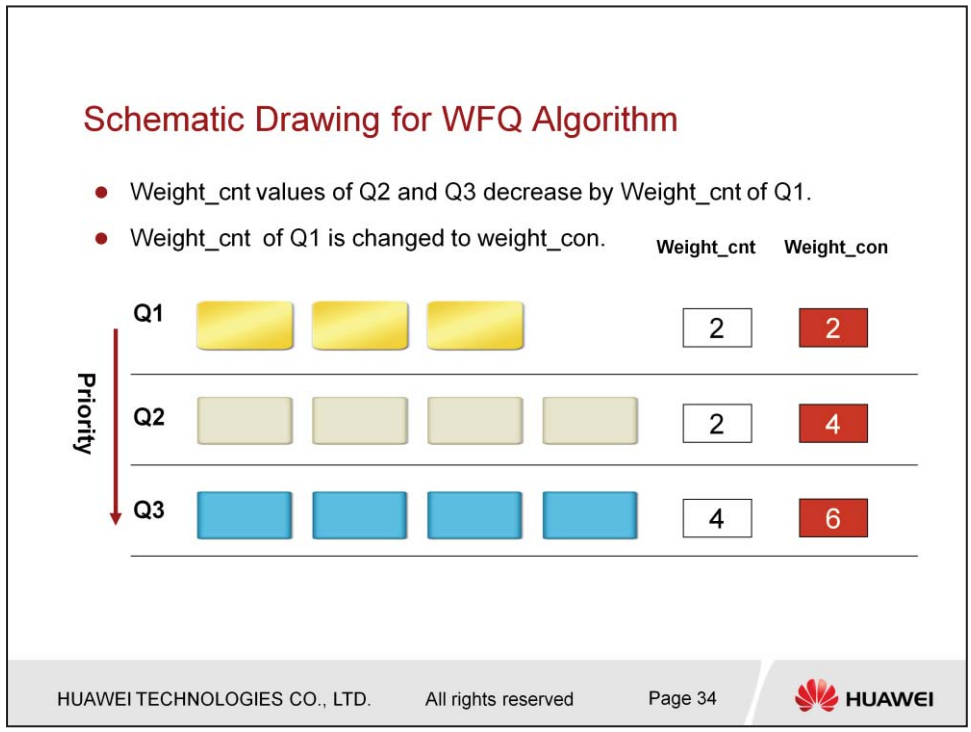
The new M value equals the initial M value minus the M value of A. That is, the new Weight_counter equals the initial Weight_counter value minus the Weight_counter of A. Then the Weight_counter of A is changed to the Weight_config. After Weight_counter values of all queues are updated, another round of scheduling begins.



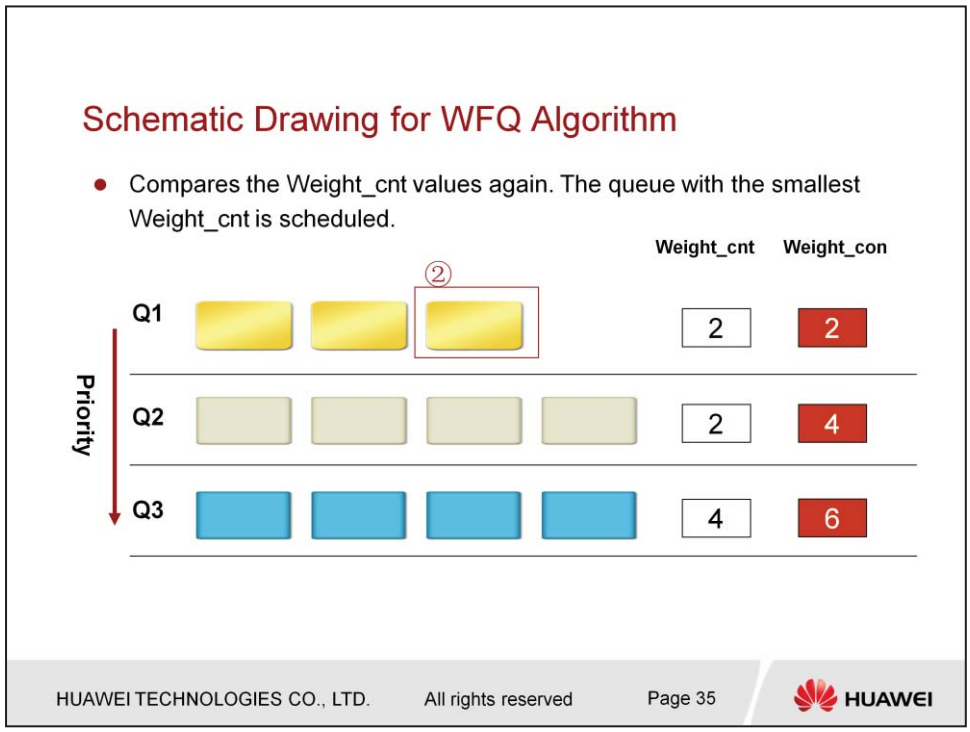
Here, an example is provided to explain the WFQ algorithm. Assume that three queues, Q1, Q2, and Q3, are configured on an interface. WFQ is adopted for the queues. The priority sequence of the queues is Q1, Q2, Q3, from the highest to the lowest. Their Weight_config values are 2, 4, 6 respectively. Initially, the weight_counter of each queue equals Weight_config, so their Weight_counter values are also 2, 4, 6.



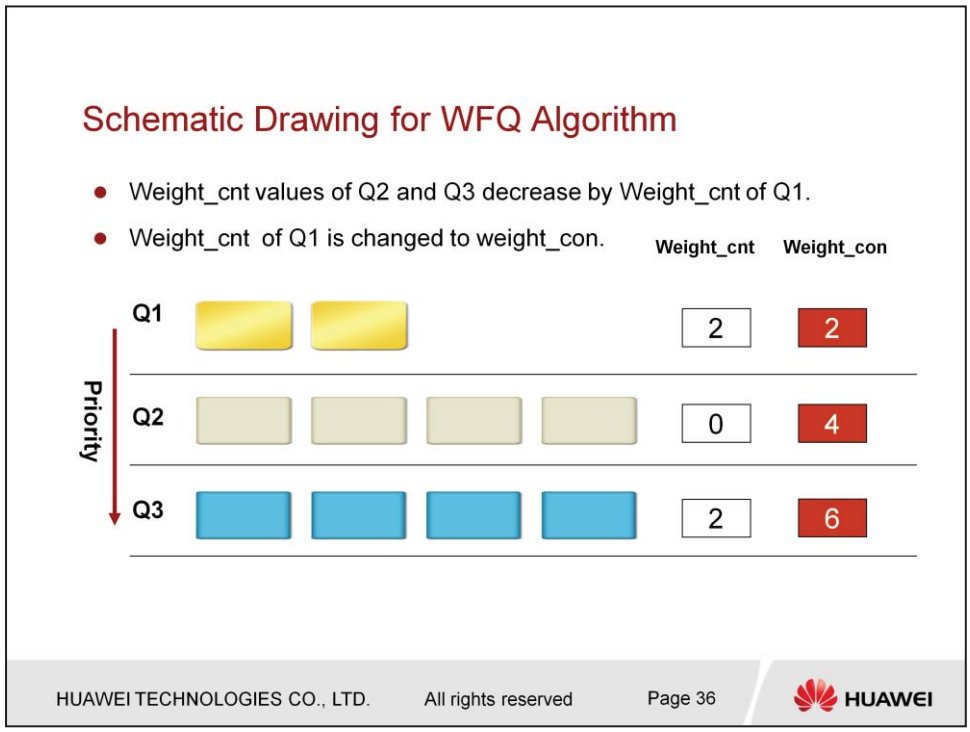
By comparison, Q1 is scheduled first, because its Weight_counter (2) is the smallest. Some of packets in Q1 is scheduled. ① represents the packets scheduled first.



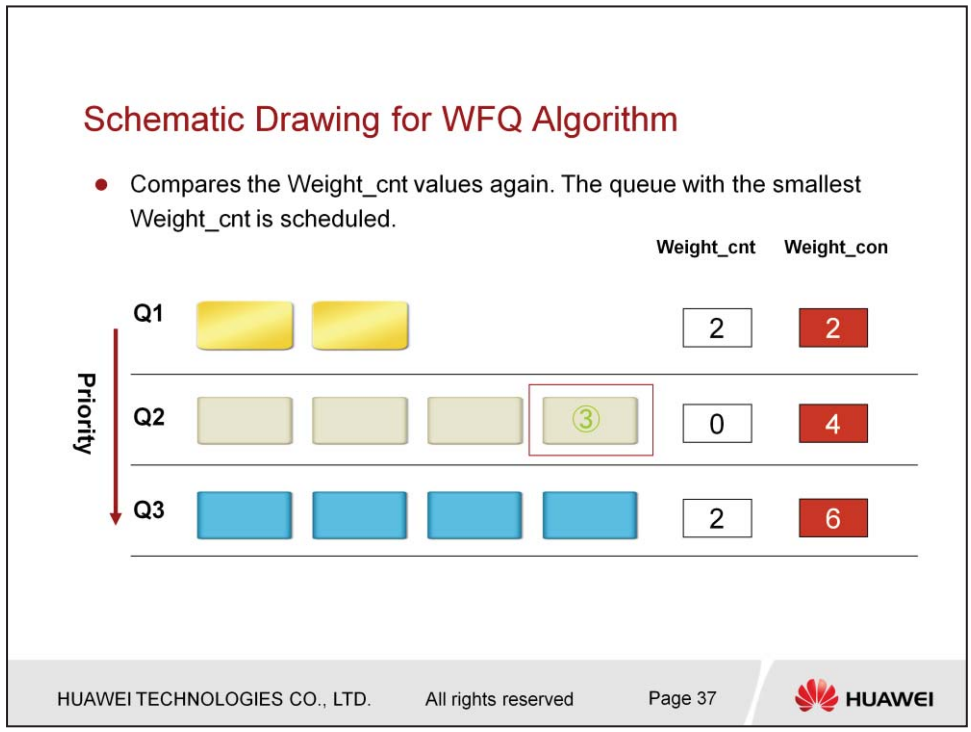
After Q1 is scheduled, Weight_counter values of all queues are updated. For Q2 and Q3, Mnew = MB – MA. That is, Weight_counter values of Q2 and Q3 decrease by Weight_counter of Q1. The Weight_counter of Q1 is changed to the Weight_config. After the first round of scheduling, Weight_counter of Q2 = 4 - 2 = 2. Weight_counter of Q3 = 6 – 2 = 4. Weight_counter of Q1 is 2. So, updated Weight_counter values of the queues are respectively 2, 2, 4.



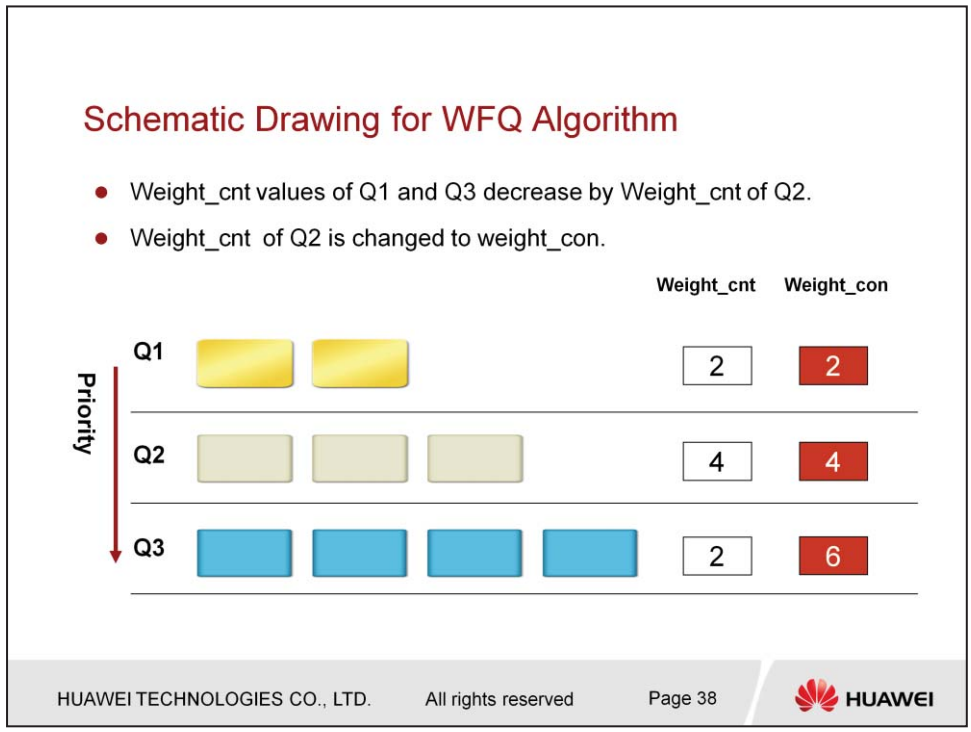
After Weight_counter values of all queues are updated, the second round of scheduling begins. The queue with the smallest Weight_counter is scheduled. Weight_counter values of Q1 and Q2 are equal. The WFQ algorithm specified that when two queues have equal Weight_counter values, the one with higher priority is scheduled first. The priority of Q 1 is higher than that of Q2, and so Q1 is still scheduled first in the second round scheduling. Some packets in Q1 are forwarded. These packets are represented by ②.



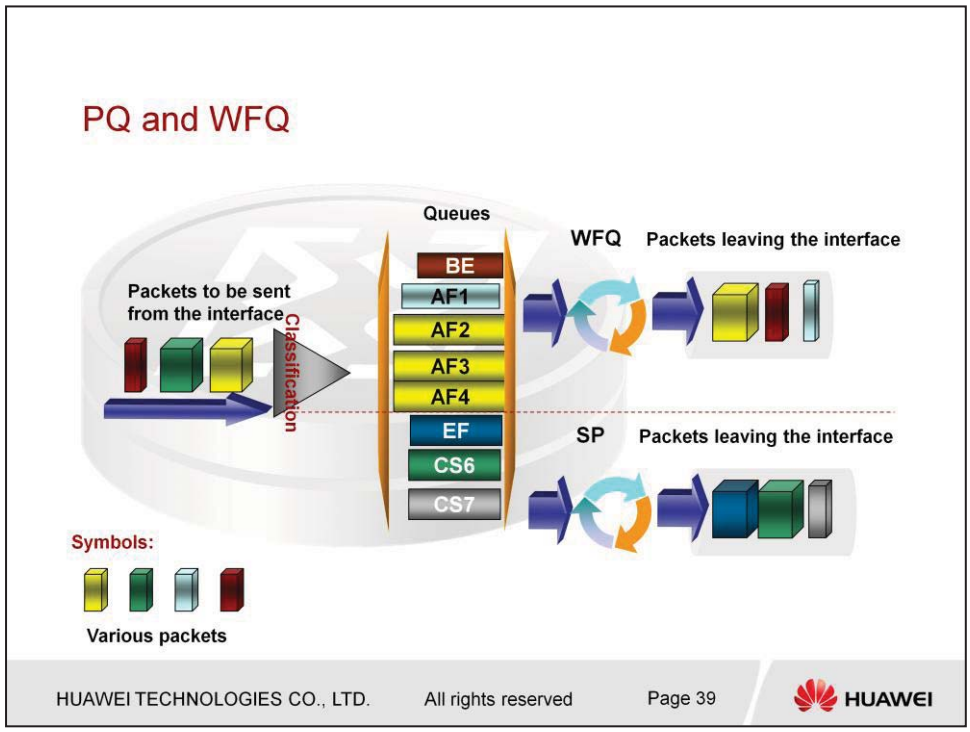
After the second round of scheduling, Q1 is scheduled twice. According to the principle for updating the Weight_counter values of these queues, Weight_counter values of Q1, Q2, and Q3 are updated to 2, 0, and 2 respectively.



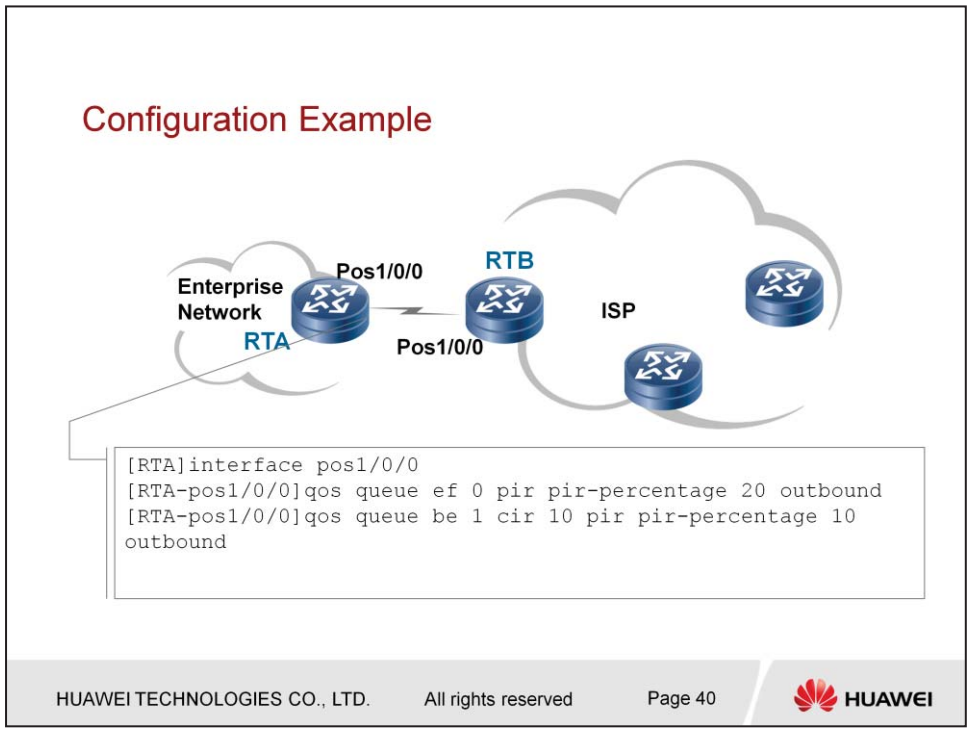
WFQ compares the Weight_cnt values again. Weight_cnt of Q2 is 0, which is the smallest, so Q 2 is scheduled. Scheduled packets in Q2 are represented by ③. After update, the Weight_counter may be a negative. In this case, the Weight_counter is regarded as 0.



After the third round of comparison, Q2 is scheduled. According to the principle for updating the Weight_counter values of these queues, Weight_counter values of Q1, Q2, and Q3 are updated to 2, 4, and 2 respectively. Next round of scheduling begins. This process lasts till packets in all packets in these queues are scheduled. Checking the scheduling rule of WFQ, you will find that WFQ adopts a weighted circular algorithm, rather than a simple circular algorithm among queues. The weight is set through the CIR parameter. If CIR values of all queues are set to 0, the PQ algorithm is used



On high-end routers of Huawei, an interface supports eight service classes. PQ and WFQ can be configured for these service classes. By default, BE, AF1, AF2, AF3, and AF4 use WFQ. EF, CS6, and CS7 use PQ. When PQ cooperates with WFQ, PQ have precedence over WFQ. That is, after PQ completes scheduling, WFQ uses the remaining bandwidth to schedule queues based on weights of queues. The cooperation of PQ and WFQ on an interface guarantees QoS of real-time services that are sensitive to the delay. Packets of preferred services have the highest precedence to occupy the bandwidth. In addition, bandwidth is allocated to flows with different priorities according to their weight. PQ is generally applied to delay-sensitive services.

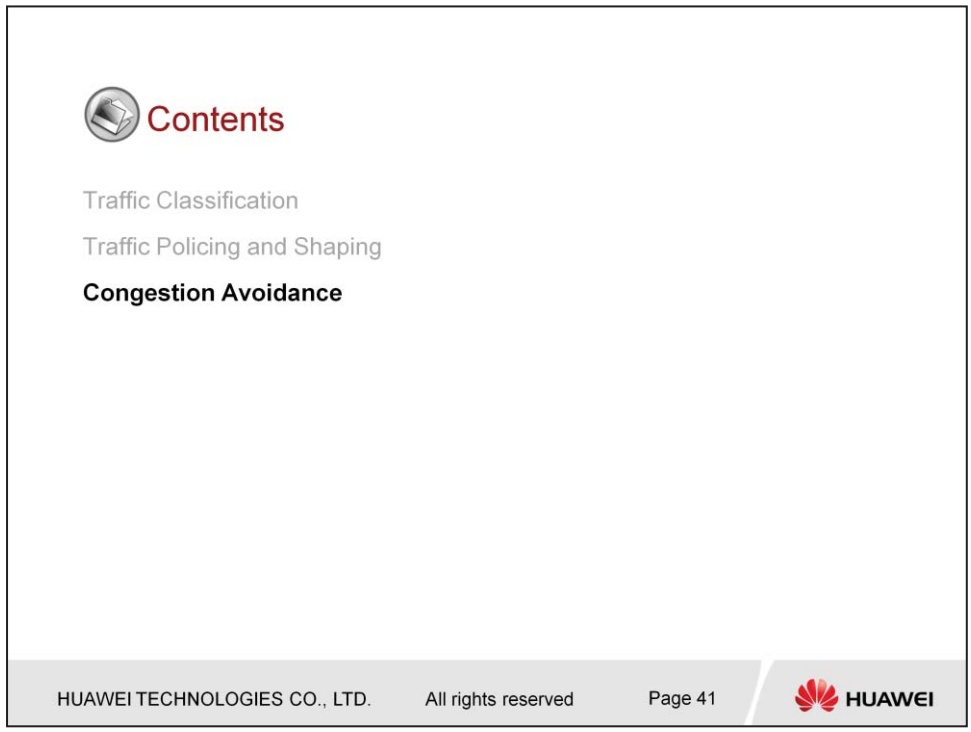


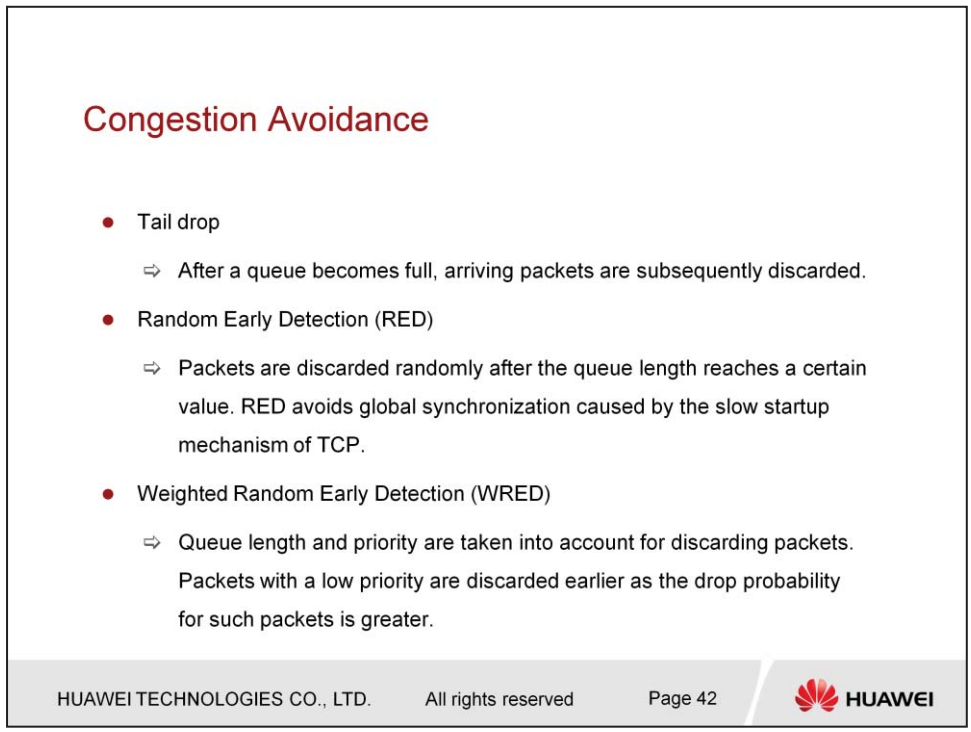
On high-end routers of Huawei, you can use the command **qos queue** to set the CIR, PIR, and scheduling priority of a specified service class on in the outbound direction (downlink). **qos queue** *service-class* [**priority** *priority*] [**cir** { *cir-value* | **cir-percentage** *percentage* }] [**pir** { *pir-value* | **pir-percentage** *percentage* }] **outbound** The meanings of the parameters in the command are:

service-class: specifies the service class. The value can be EF, AF1, AF2, AF3, AF4, BE, CS6, or CS7. priority *priority*: specifies the scheduling priority of a queue. The value is an integer ranging from 0 to 1. **0 indicates PQ scheduling, and 1 indicates WFQ scheduling**. By default, WFQ scheduling is configured for AF1-AF4 and BE queues, and PQ scheduling is configured for EF, CS6, and CS7 queues. *cir-value*: specifies the committed information rate. The value is an integer ranging from 0 to 10000, in Mbit/s. **In the WFQ** algorithm, the **CIR indicates the weight of a queue. By default, weights of the eight classes of queues are** 10%, 10%, 10%, 15%, 15%, 10%, 5%, and 5%. cir-percentage *percentage*:
specifies the percentage of the CIR against the interface bandwidth. The value is an integer ranging from 0 to 100. *pir-value*: specifies the peak information rate. The value is an integer ranging from 0 to 0 to 10000, in Mbit/s. pir-percentage *percentage*: specifies the percentage of the PIR against the interface bandwidth. The value is an integer ranging from 0 to 100. outbound: indicates that QoS scheduling is applied in the outbound direction of the interface.

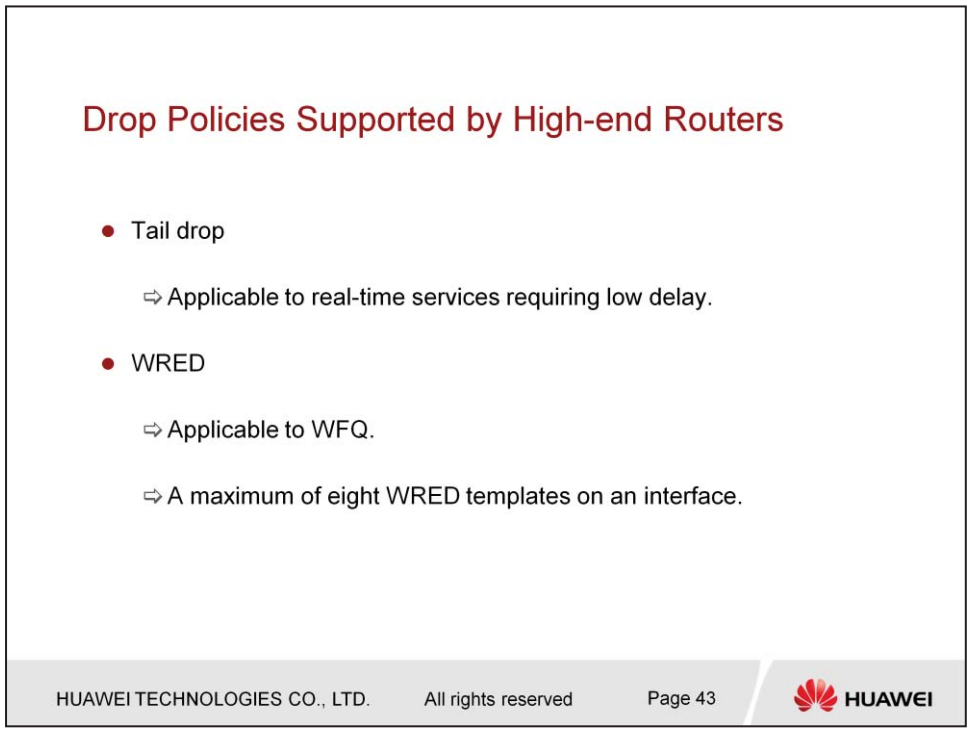
In this example, traffic policing is configured on pos1/0/0 of RTA. PQ scheduling is configured for the EF class. The PIR of the EF class makes 20% of the interface

bandwidth. WFQ scheduling is configured for the BE class. The weight of the BE class is 10, and its PIR makes 20% of the interface bandwidth.


Contents
Traffic Classification
Traffic Policing and Shaping
Congestion Avoidance
HUAWEI TECHNOLOGIES CO., LTD. All rights reserved Page 41
HUAWEI

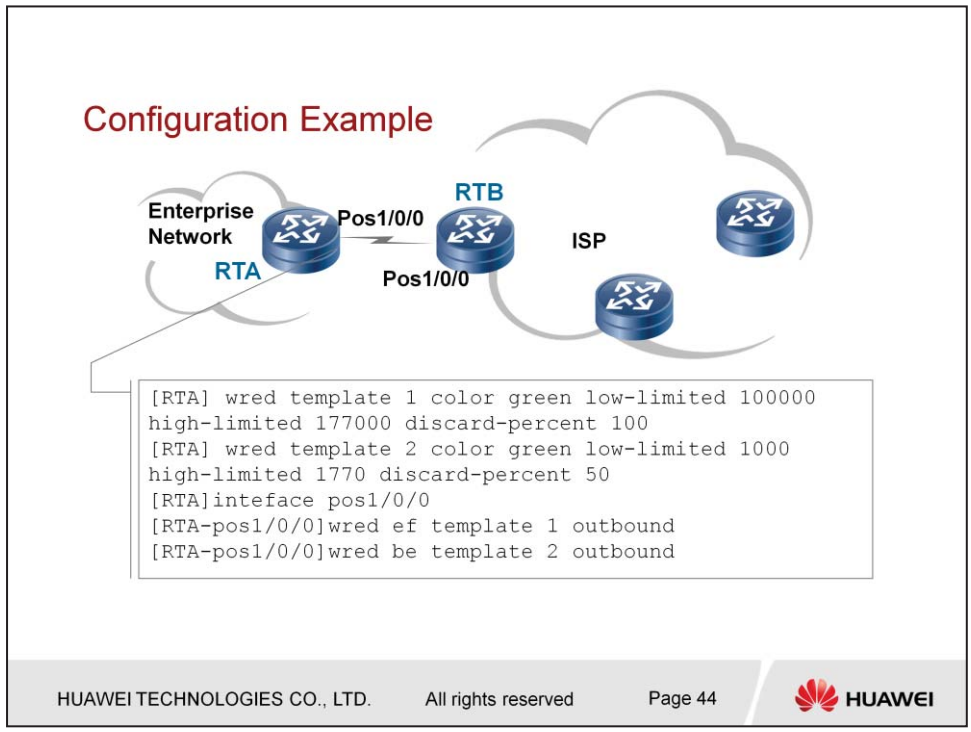


Tail drop is a traditional drop policy. When the length of a queue reaches the maximum value, subsequent packets are discarded. Tail drop may cause global TCP synchronization. When packets of many TCP sessions are discarded, these TCP sessions enter congestion avoidance and slow startup state at the same time to lower and adjust the traffic, then these sessions generate peak traffic at the same time. The process repeats, and thus the network traffic becomes unstable. To avoid global synchronization of TCP sessions, you can configure Random Early Detection (RED) or Weighted Random Early Detection (RED) on the router.



High-end routers of Huawei support tail drop and WRED. For PQ, the drop policy can be tail drop or WRED. In general, tail drop is applied to real-time services that require low delay. For WFQ, the drop policy can be tail drop or WRED. In general, WRED is used. On high-end routers of Huawei, WRED templates are used. The high-end routers adopt interface-based scheduling. Each interface can be configured with a maximum of eight WRED templates. Each template can process packets in three colors: red, yellow, and green. Generally, green packets have the lowest drop probability and the highest drop threshold. Red packets have the highest drop

probability and the lowest drop threshold. The drop threshold and drop probability for each color can be set flexibly.



For traffic shaping, POS1/0/0 of RTA is configured with two services classes: EF and BE, which use PQ and WFQ respectively. WRED template 1 is configured for the EF class and WRED template 2 is configured for the BE class.

The configuration procedure is as follows:

1. For the EF flow, set higher lower threshold and upper larger upper threshold. Set a lower drop probability so that more packets of the EF class can be cached. wred template 1 color green low-limited 100000 high-limited 177000 discard percent 100

2. On the interface, specify the WRED template for the EF class. [RTA-pos1/0/0]wred ef template 1 outbound

3. For the BE flow, set smaller lower threshold and upper threshold. Set a higher drop probability so that fewer packets of the BE class can be cached. wred template 2 color green low-limited 1000 high-limited 1770 discard-percent 50

4. On the interface, specify the WRED template for the BE class. [RTA-pos1/0/0]wred be template 2 outbound

The following command:

wred template *template-index* **color** { *green* | *yellow* | *red* } **low-limited** *low-limit* **high-limited** *high-limit* **discard-percent** *discard-percent* is used to configure the WRED template. Each template can process packets in three colors: red, yellow, and green. Generally, green packets have the lowest drop probability and the highest drop threshold. Yellow have packets higher drop probability and lower drop threshold than green packets. Red packets have the highest drop probability and the lowest drop threshold. By configuring the WRED template, you can set the drop threshold (including upper threshold and lower threshold) and the drop threshold.

When the queue length is smaller than the lower threshold, no packets are discarded. When the queue length is between the lower threshold and upper threshold, packets are discarded randomly (the longer the queue, the higher is the drop probability). When the queue length exceeds the upper threshold, all packets are discarded. The drop threshold and drop probability for each color can be set flexibly. If a queue occupies higher bandwidth, it indicates that this type of flow has higher priority. For such a queue, set a larger lower threshold and upper threshold and set a smaller drop probability. For a queue that occupies lower bandwidth, set smaller lower threshold and upper threshold and set larger drop probability.

Note the following:

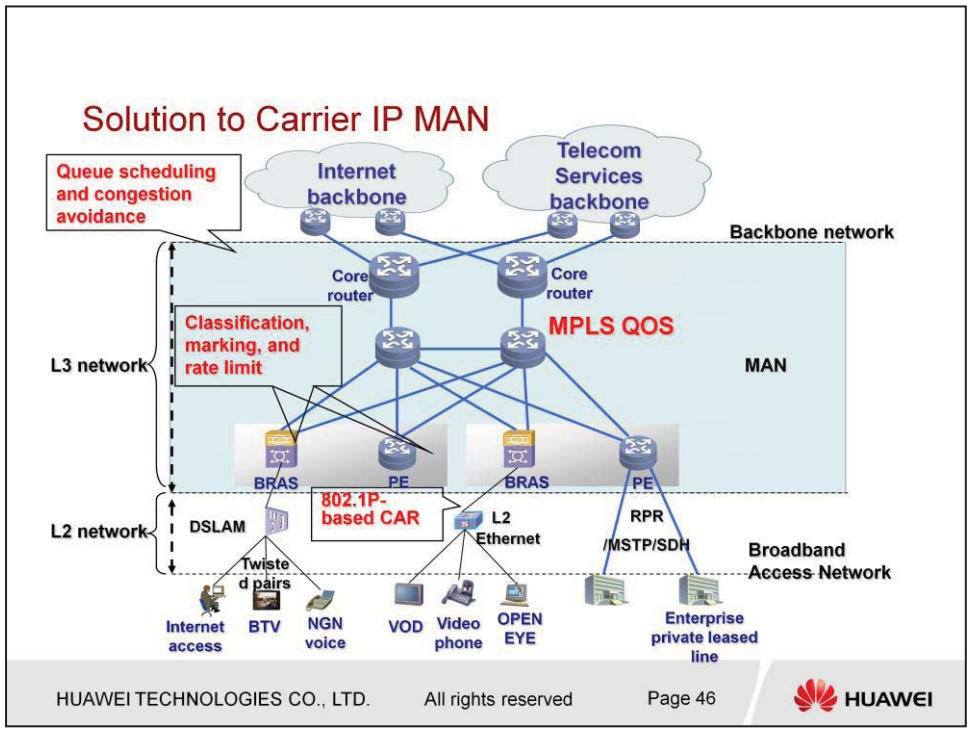
1. The WRED template only sets the drop probability for a color. It does not specify the drop probability for a certain service class. In actual application, you must specify the WRED templates in the outbound direction for queues of different service classes.

2. WRED is invalid for uplink queues. It is applicable only to the downlink queues.

The meanings of the parameters in the command are:

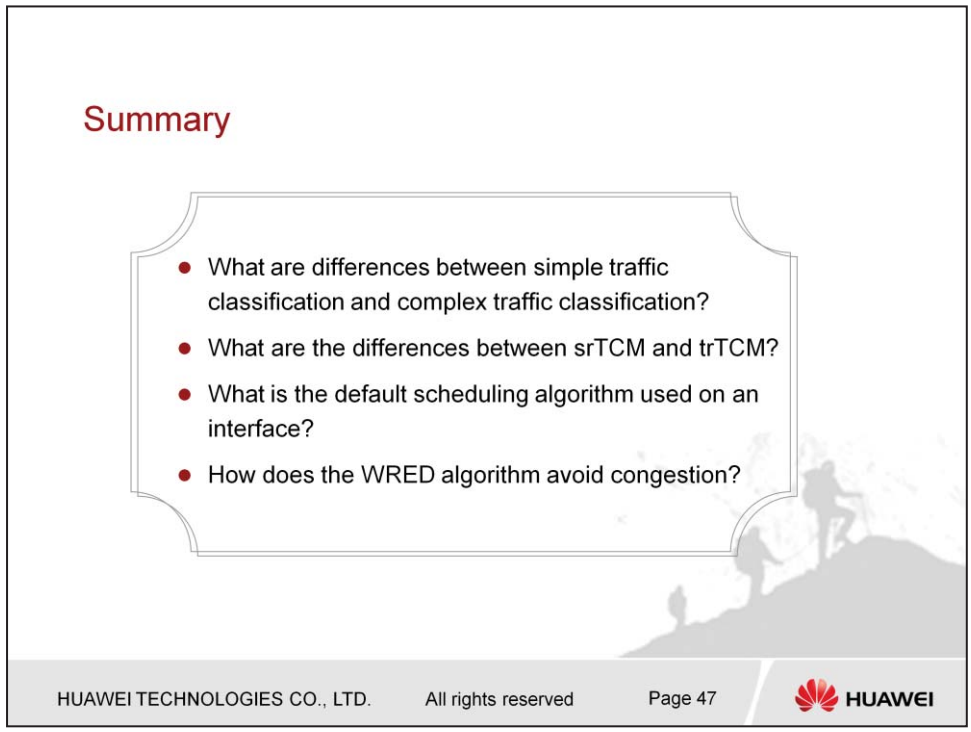
template-index: specifies the index of a WRED template. The value is an integer ranging from 0 to 6. color { *green* | *yellow* | *red* }: specifies the color of packets.

low-limited *low-limit*: specifies the lower drop threshold of a queue. The value is an integer ranging from 0 to 0 to 177216, in bytes. high-limited *high-limit*: specifies the upper drop threshold of a queue. The value is an integer ranging from 0 to 0 to 177216, in bytes. discard-percent *discard-percent:* specifies the drop probability of a queue. The value is an integer ranging from 1 to 100. The command **wred** *service-class* **template** *template-index* **outbound** is used to specify WRED templates for queues of different service classes in the outbound direction (downlink). *service-class*: specifies the service class. The value can be EF, AF1, AF2, AF3, AF4, BE CS6, or CS7. **template** *template-index*: specifies the index of a WRED template. The value is an integer ranging from 0 to 6. **outbound**: indicates to configure the WRED template in the outbound direction (downlink).



With the upgrading of network services, services on the IP network have expanded from pure Internet data service to the integrated service, streaming media, and IPTV service. These new services have comprehensive QoS requirements, including the delay, jitter, packet loss, and network reliability, rather than simple requirements for network reliability. The major objective of QoS technology is ensuring services when link congestion occurs. The main parameters in this aspect are delay and packet loss. The most commonly used QoS technologies are queuing and WRED. The QoS technology does not create any bandwidth, but it manages the bandwidth according to the network situation.

Generally, a carrier network consists of the broadband access network, Metropolitan Area Network (MAN), and backbone network. The broadband access network is a layer-2 network that consists of DSLAMs or layer-2 switches. For QoS, 802.1P can be deployed to differentiate users with different priorities. 802.1P-based CAR can be configured to limit the inbound traffic. In the MAN, the BRAS or PE router identifies service flows of the same user through traffic classification, and then selects proper scheduling algorithms and congestion avoidance methods for different flows. Thus various services are provided with different levels of QoS guarantee. QoS implementation varies with the change of network requirement. QoS is a complicated project. Besides queuing and congestion avoidance introduced here, in more complex networks, QinQ can be used to implement user-based and service-based QoS and hierarchical QoS (HQoS). If you are interested in knowledge about QinQ, refer to other technical documents.



1. What are differences between simple traffic classification and complex traffic classification?

Simple traffic classification is a process of identifying flows according to the IP precedence, DSCP, MPLS EXP, or 802.1P contained in packets. The packets belonging to the same classifier are called the behavior aggregate (BA). In general, a core router in the Diff-Serv domain needs to perform only simple traffic classification. Simple traffic classification is implemented on high-end routers normally. Mid-range and low-end routers implement traffic classification based on the traffic policy. That is, they implement complex traffic classification. Complex traffic classification is a process of identifying packets according to the quintuple (source address, destination address, source port number, destination port number, and protocol type). Complex traffic classification is generally used on the network edge.

2. What are differences between srTCM and trTCM?

srTCM: A Single Rate Three Color Marker. Tokens are added to a single bucket at the rate of CIR. trTCM: A Two Rate Three Color Marker. Tokens are added to bucket C and bucket P respectively at the rate of CIR and PIR.

3. What is the default scheduling algorithm on an interface?

By default, an interface supports eight service classes: BE, AF1, AF2, AF3 AF4, EF CS6, and CS7. By default, EF, CS6, and CS7 classes use the PQ scheduling algorithm, and other classes use the WFQ scheduling algorithm.

4. How does WRED algorithm avoid congestion?

A WRED template can process packets in three colors, namely, green, yellow, and red. Each color is configured with a drop probability, a lower drop threshold, and an upper drop threshold. Generally, green

packets have the lowest drop probability and the highest drop threshold. Yellow packets have higher drop probability and lower drop threshold than green packets. Red packets have the highest drop probability and the lowest drop threshold. When the queue length is smaller than the lower threshold, no packets are discarded. When the queue length is between the lower threshold and upper threshold, packets are discarded randomly (the longer the queue, the higher is the drop probability). When the queue length exceeds the upper threshold, all packets are discarded.