Huawei Certification

HCDP-IERN

Implementing Enterprise Routing Network



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-IERN Implementing Enterprise Routing Network

Edition 1.6

Huawei Certification System

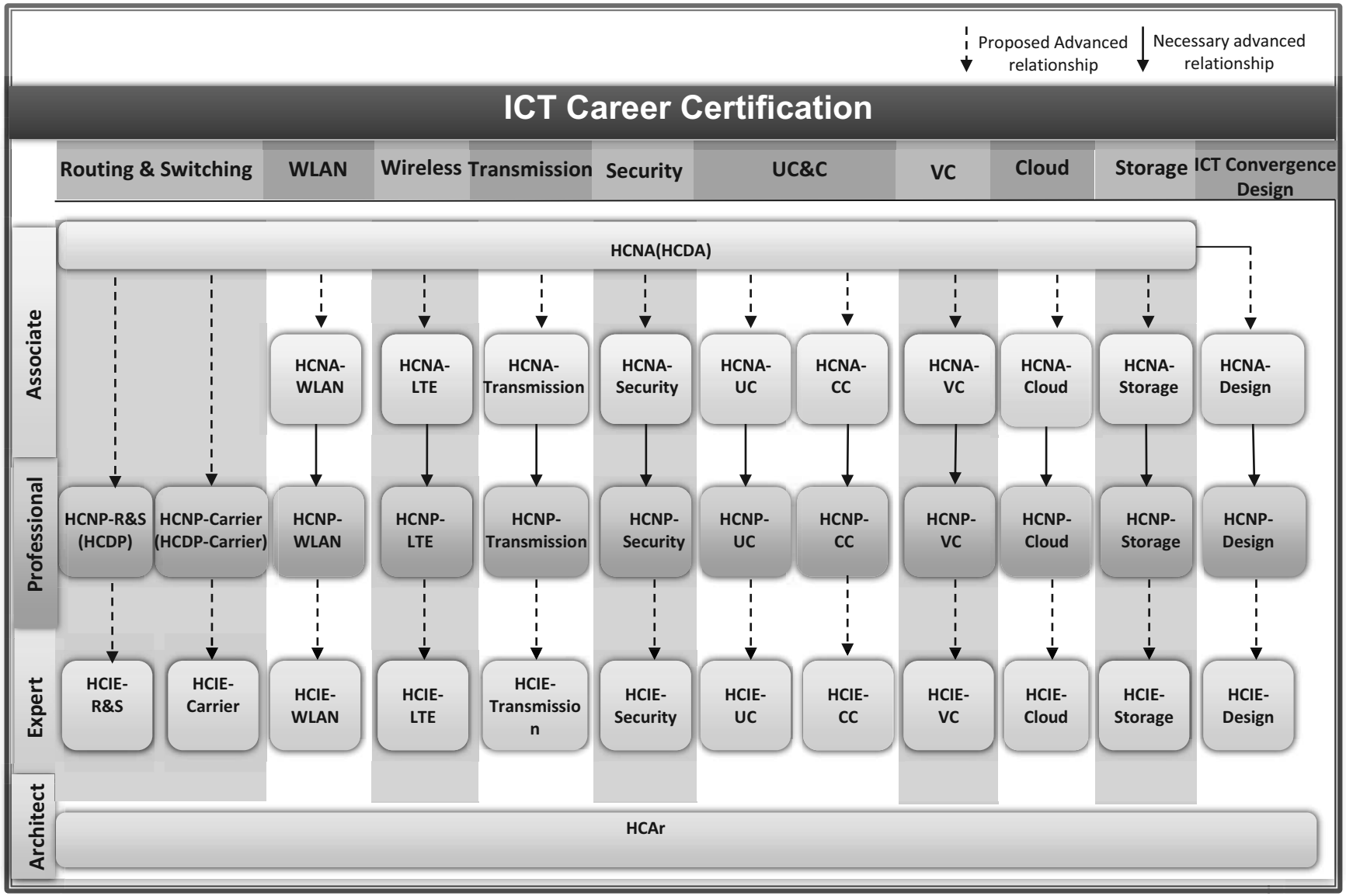
Relaying on its strong technical and professional training system, according to different customers at different levels of ICT technology, Huawei certification is committed to provide customs with authentic, professional certification.

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 (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 (Huawei Certified Internetwork Expert) 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-IERN, targeting audience

ranging from learners preparing for the HCDP-IERN examination to readers wishing not only to comprehend principles of common routing protocols, but also to understand implementations of these routing protocols in the Huawei Versatile Routing Platform (VRP).

Content

This book is composed of five modules that introduce the principles of common routing protocols, configurations of these protocols in the VRP, and characteristics and applications of the Huawei router series starting from the basic concepts towards the more detailed ones.

Module 1 introduces IPv4 address planning and sub-net division and will help you consolidate basic knowledge.

Modules 2, 3 systematically describe operating principles of one typical

Internal Gateway Protocol, OSPF, and the Border Gateway Protocol (BGP).

In addition, these modules describe the configurations and implementations of these protocols in the VRP, which will assist you in comprehensively understanding IPv4 routing protocols.

Module 4 provides various case studies to illustrate how to flexibly employ multiple tools to enable routing control and selection, which will assist you in improving skills in comprehensively planning and flexibly deploying routing protocols.

Module 5 briefly introduces basic information about multicast, such as the multicast address, Internet Group Management Protocol (IGMP), Protocol Independent Multicast-Dense Mode (PIM-DM), and Protocol Independent Multicast-Sparse Mode (PIM-SM), which will assist you in understanding basic multicast principles and applications.

This book aims to guide you through the preceding concepts step by step so that you may fully understand the technologies. 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:

- Participating in the HCDA training
- Passing the HCDA examination
- Familiar with the TCP/IP protocol stack and IP addresses

ge5

Icons Used in This Book



IPv6 Router



SOHO Router



Voice Router



Low-end Router



High-end Router



Core Router



Hub



Convergence 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 Advanced IP.....Page 1

Advanced IP Address Planning.....Page 3

Module 2 OSPF.....Page 37

OSPF Routing Protocol Basis.....Page 39

OSPF Neighbor and Adjacency.....Page 53

OSPF Protocol Packet and LSA.....Page 87

Establishing OSPF Neighbor&Adjacency.....Page 107

Calculating OSPF Intra-area Routes.....Page 127

OSPF Inter-area Routing.....Page 170

OSPF External Routing.....Page 188

OSPF Special Areas.....Page 214

OSPF Troubleshooting.....Page 239

OSPF Extension.....Page 276

Module 3 BGP.....Page 291

BGP Overview.....Page 293

Working Principle of BGP.....Page 323

BGP Route Selection.....Page 347

BGP Route Aggregation.....Page 379

BGP Routing Policies.....Page 405

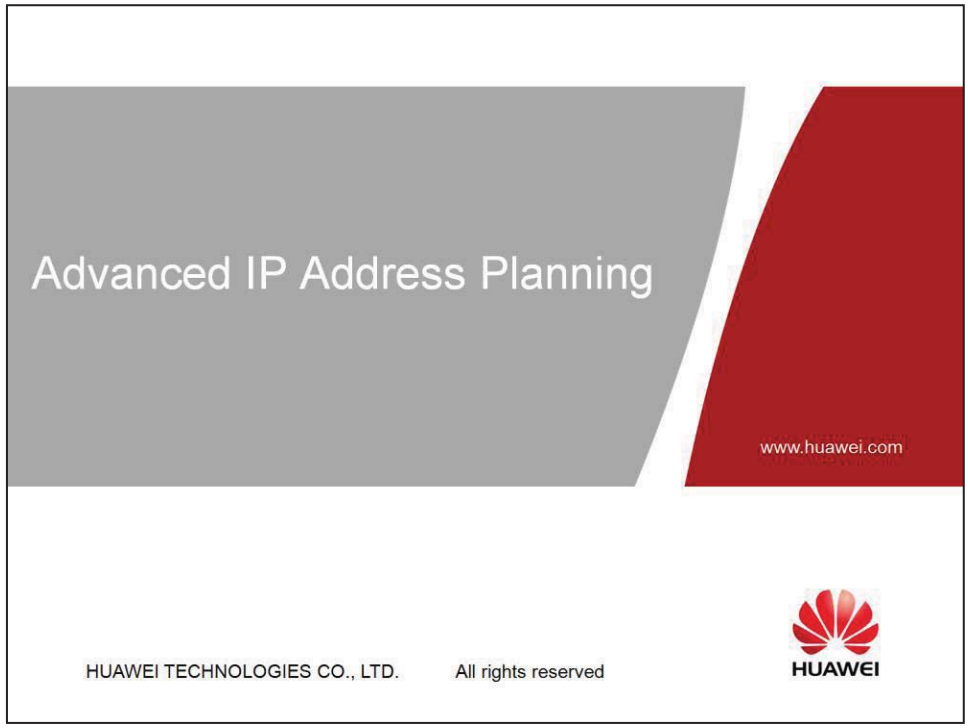
BGP Route Reflection and Confederation.....Page 445

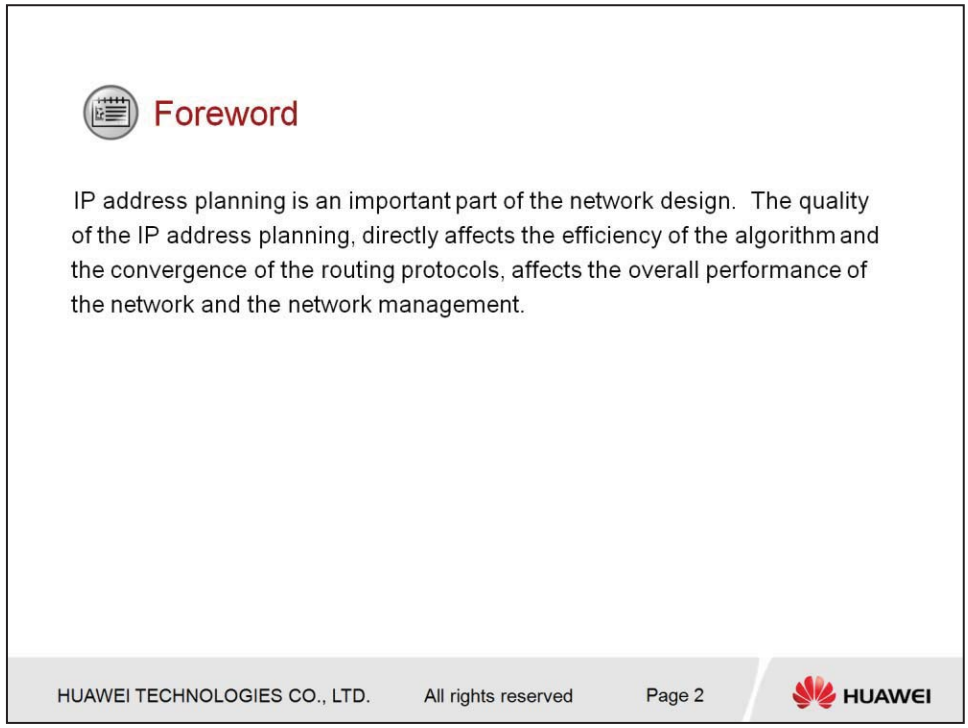
BGP Multi-homed.....Page 476

BGP Troubleshooting.....Page 500

Module 4 Route Selection & Control ..... Page 553
Route Selection Tools ..... Page 555
Routing Policy ..... Page 574
Policy-based Route Selection ..... Page 625
Module 5 Multicast ..... Page 635
IP Multicast Basis ..... Page 637
IGMP Protocol Principles ..... Page 672
PIM-DM Protocol Principles ..... Page 714
PIM-SM Protocol Principles ..... Page 739

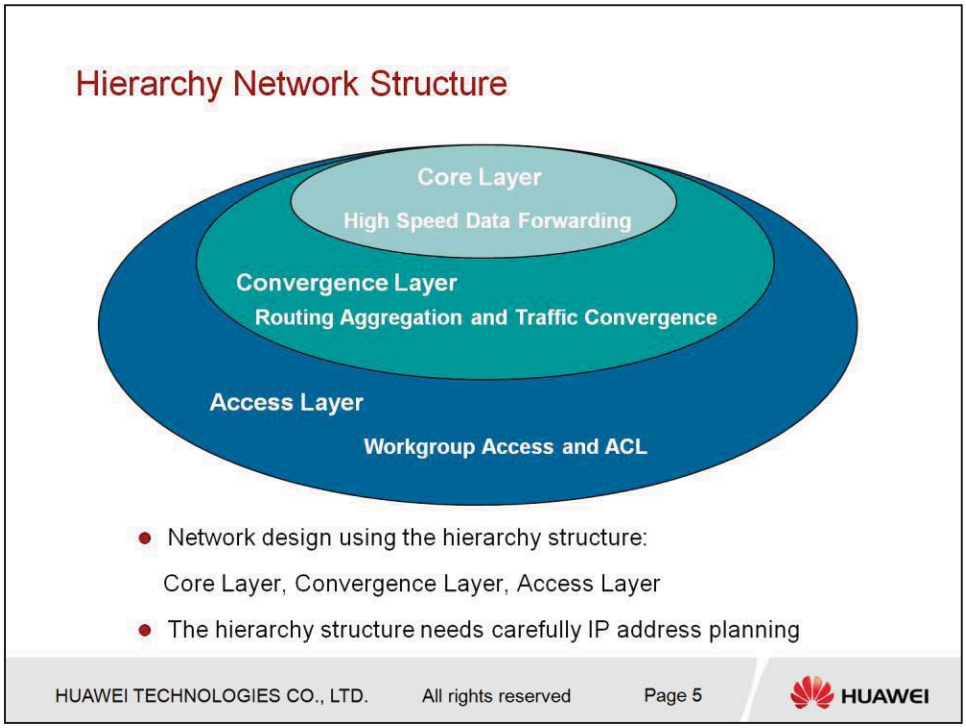
Module 1
Advanced IP


Advanced IP Address Planning
www.huawei.com
HUAWEI TECHNOLOGIES CO., LTD.
All rights reserved
HUAWEI


Foreword
IP address planning is an important part of the network design. The quality of the IP address planning, directly affects the efficiency of the algorithm and the convergence of the routing protocols, affects the overall performance of the network and the network management.
HUAWEI TECHNOLOGIES CO., LTD. All rights reserved Page 2
HUAWEI

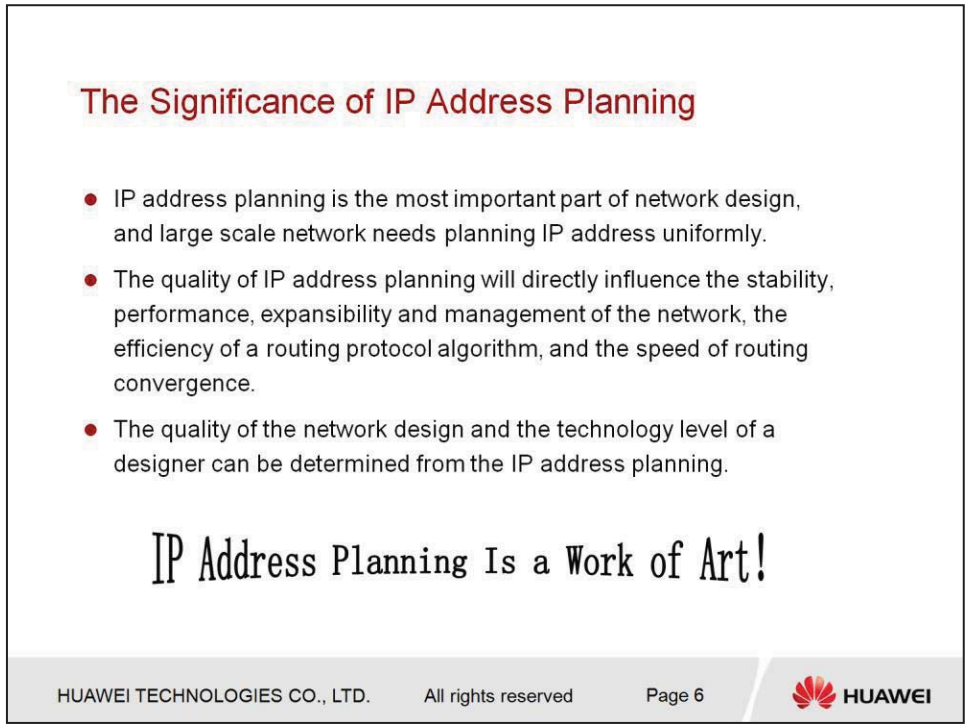

Objectives
Upon completion of this section, you will be able to :
Master VLSM
Master route aggregation and CIDR
HUAWEI TECHNOLOGIES CO., LTD.
All rights reserved
Page 3
HUAWEI


Contents
The importance of the IP address planning
IP address planning with VLSM
Route aggregation and CIDR
HUAWEI TECHNOLOGIES CO., LTD.
All rights reserved
Page 4
HUAWEI



With the extension of network scale, a hierarchical network structure appears.

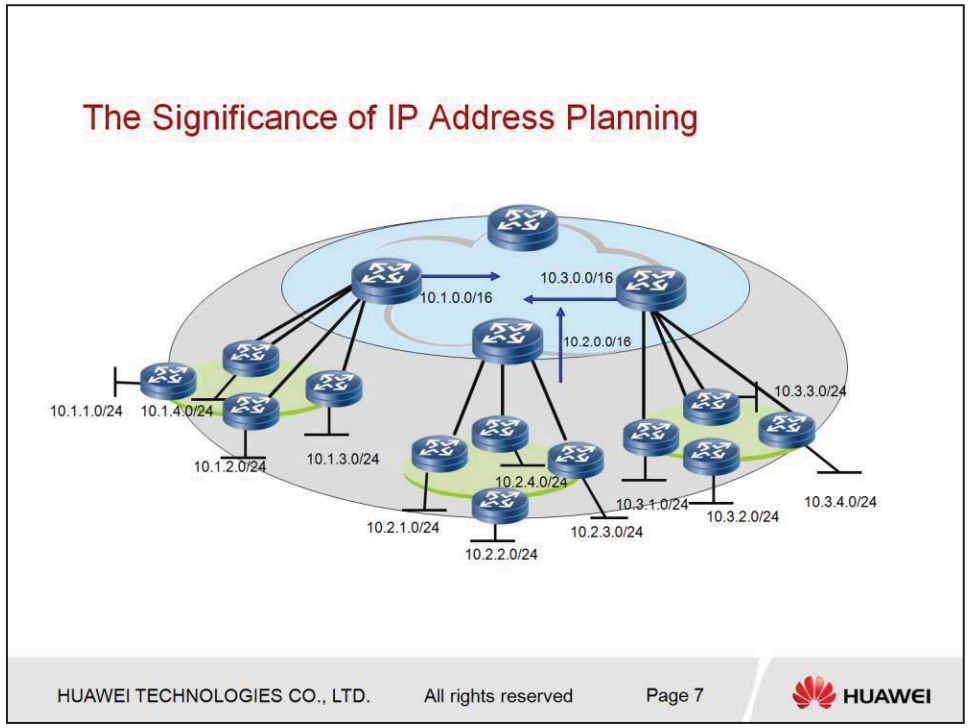
Basically, the network can be classified into the core layer, convergence layer and access layer. The core layer implements packet switching and high-speed traffic transmission. It is required that the devices on the core layer have large capacity, fast forwarding speed and high stability. The convergence layer separates the network topology, controls the size of routing table and the network convergence, and implements various service features. The access layer connects terminal users to the network. The access layer provides abundant of interfaces and powerful access capability, and implements various service features. The hierarchical network structure requires a specific IP address plan.



IP address planning is the most importance in network design. In the large network, the IP addresses should be planned uniformly.

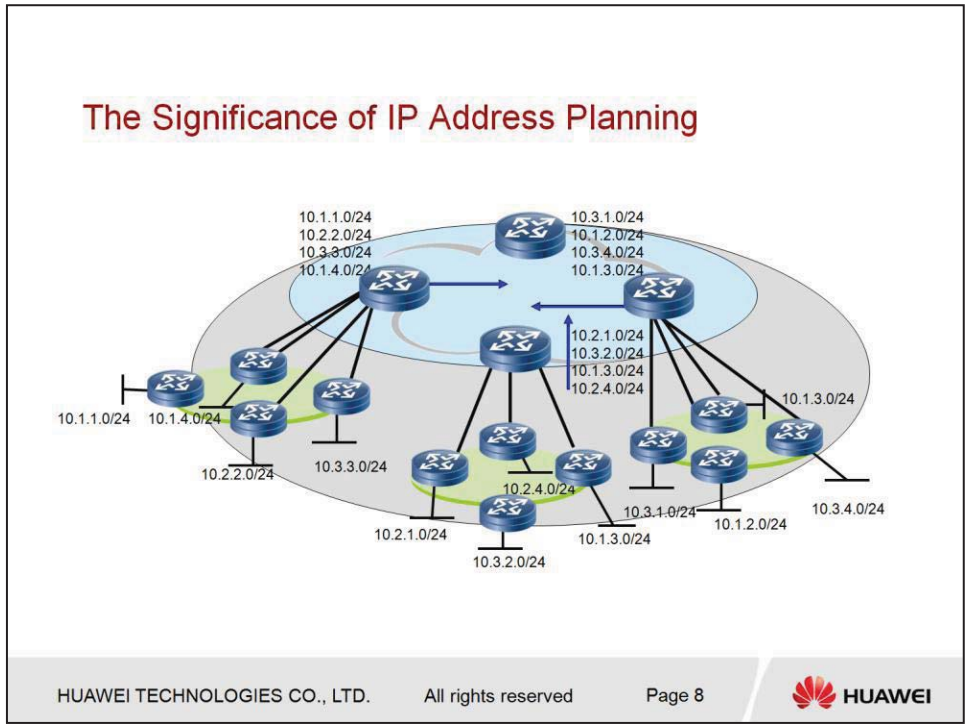
The IP address planning influences the efficiency of routing protocol calculation, the network performance, extensibility and management. Moreover, the IP address planning influences the development of network application.

To check the quality of the network design or the skill of a network engineer, just check the quality of IP address planning.



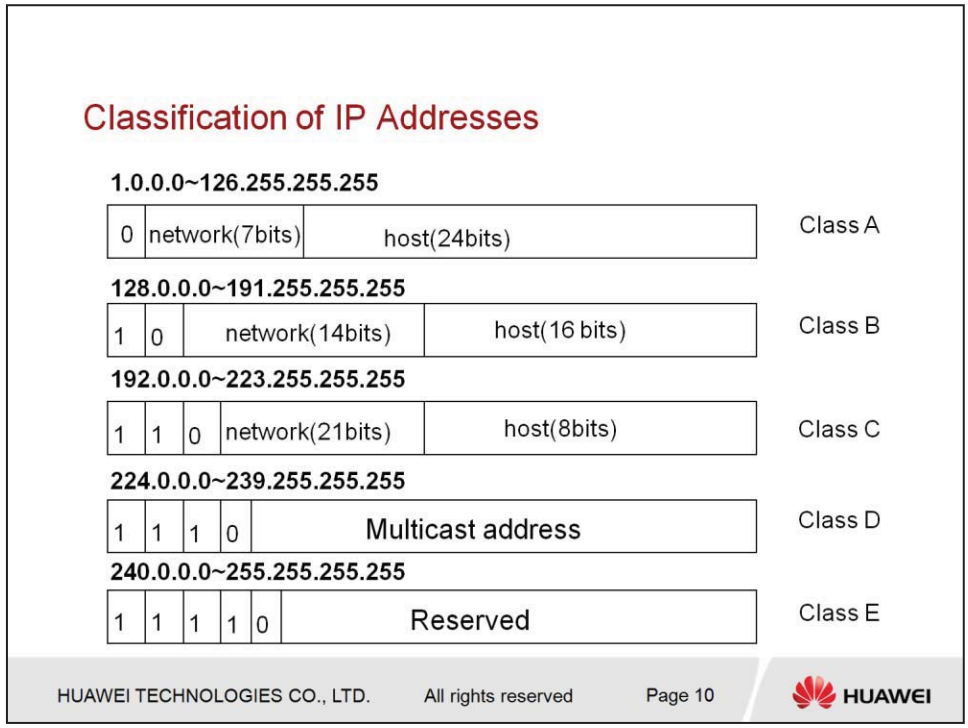
As shown in the figure above, each area has four network segments. The routes of a network segment are aggregated on the core layer, and the aggregated route is forwarded to other areas.

The number of routes on each router is 6, 4 own area routes and 2 other area aggregated routes.



Compared with the preceding network, this one does not plan the IP address properly. The routes cannot be aggregated on the core router. All the routes of the network segment will be sent to other areas. Each router has 12 routes, which are two times of the routes on the router in the preceding network.


Contents
The importance of the IP address planning
IP address planning with VLSM
Route aggregation and CIDR
HUAWEI TECHNOLOGIES CO., LTD.
All rights reserved
Page 9
HUAWEI



An IP address contains a network ID, which identifies a network segment uniquely or identifies the aggregation of multiple network segments. Devices in the same network segment use the same network ID.

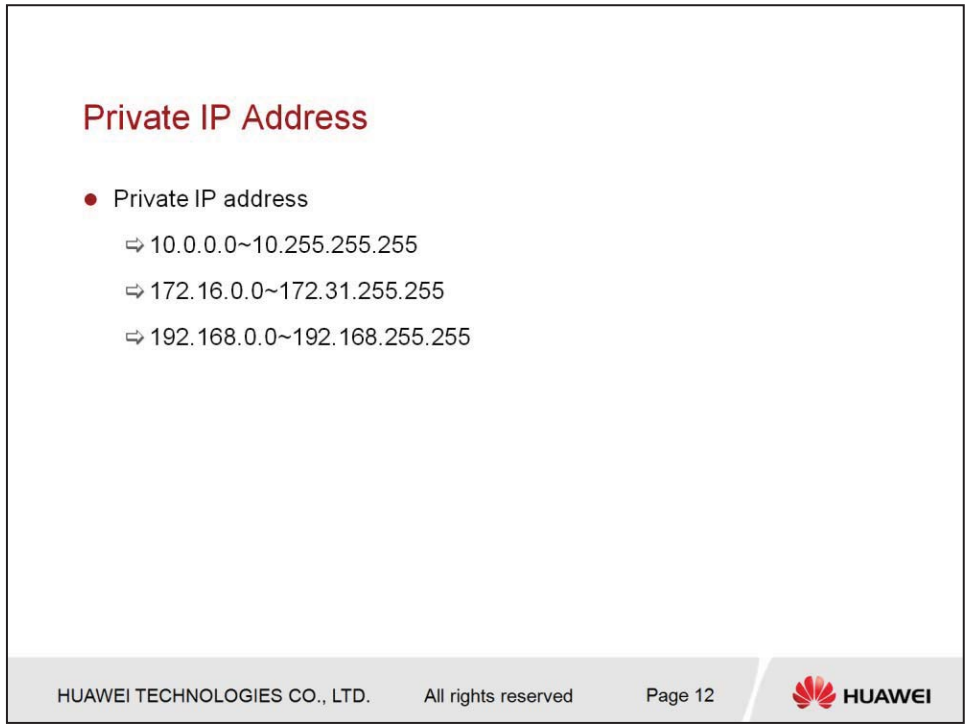
An IP address also contains a host ID, which identifies a device in the network segment uniquely.

How to distinguish the network ID and the host ID? The Internet designer classifies the IP addresses into five classes according to the size of the network, namely, class A, class B, class C, class D, and class E.

The network ID of the IP address of class A is the first octet, and the first binary digit of the first binary octet is 0. Therefore, the number of valid bits in class A address is 7. The first octet of class A address ranges from 1 to 126 (127 is reserved). For example 10.1.1.1 and 126.2.4.78 are class A addresses. The

host ID of the class A address is the last three octets, namely, the last 24 bits. The IP address of class A ranges from 1.0.0.0 to 126.255.255.255. Each class A network has 224 IP addresses.

The network ID of the class B address is the first two octets. The first binary digit of the first octet is 1 and the second digit is 0. Therefore, the number of valid binary digits of the class B

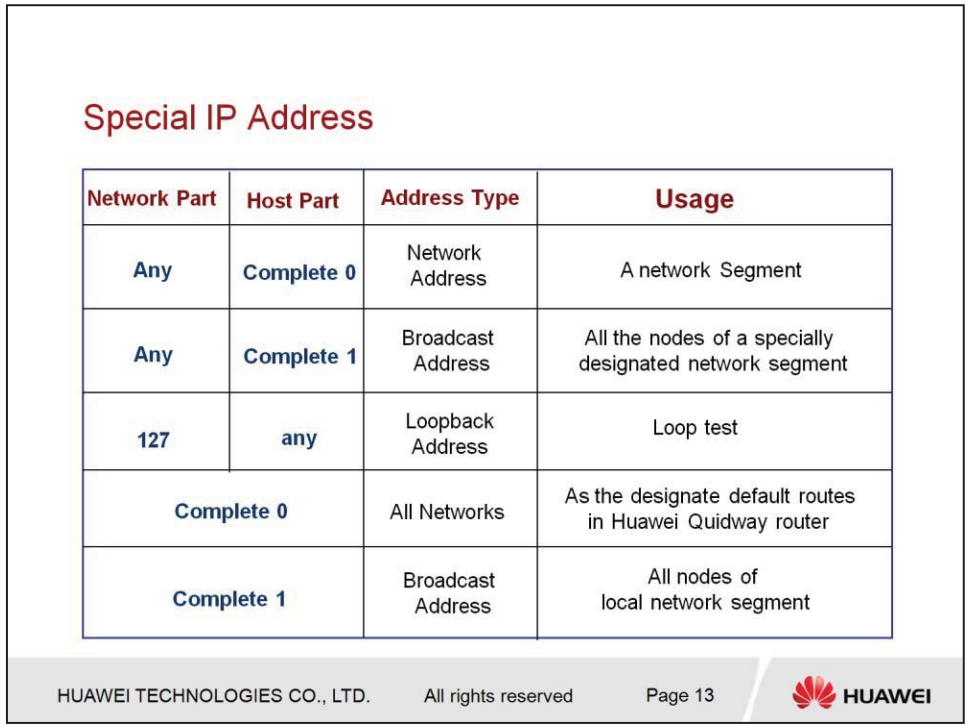


When deploying IP addresses, we usually use private IP addresses. Private IP addresses, reserved by the InterNIC, can be freely used by companies. The private IP addresses cannot be used to access the Internet. The reason is that there are no routes to the private IP addresses on the public network and the IP addresses may conflict. When a user needs to access the Internet with a private IP address, the private IP address must be translated into a public address that can be identified by the public network through the Network Address Translation (NAT) technique. The InterNIC reserves the following network

segments as the private IP addresses: class A: 10.0.0.0-10.255.255.255; class B: 172.16.0.0-172.31.255.255; class C: 192.168.0.0-192.168.255.255.

By using the private IP addresses, the enterprises reduces the cost of buying the public addresses and the IP addresses are saved. However, the private IP addresses cannot solve the problem of IP address shortage completely. This problem

can be solved by IPv6 addresses. An IPv6 address consists of 128 binary digits and up to 2128 IP addresses can exist.



An IP address uniquely identifies a device on the network. However, some IP addresses cannot be used to identify devices, because they are used for some special purposes.

The IP address with the all 0s host ID is called network address. The network address identifies a network segment. For example, class A address 1.0.0.0/8 and private addresses 10.0.0.0/8 and 192.168.1.0/24 are network addresses.

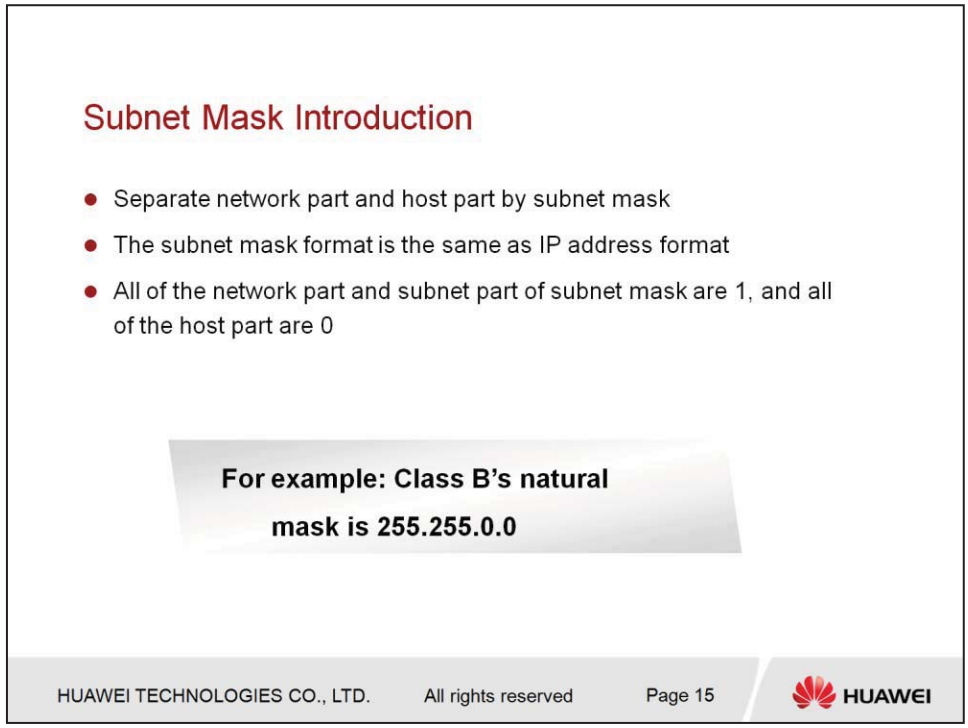
The IP address with the all 1s host ID is called broadcast address. A broadcast address identifies all the hosts on a network. For example, 10.255.255.255/8 and 192.168.1.255/24 are broadcast addresses. The router can send broadcast

packets on the 10.0.0.0/8 or 192.168.1.0/24 network segment. If the router sends the packet to the broadcast address, all the nodes on the network segment can receive the packet.

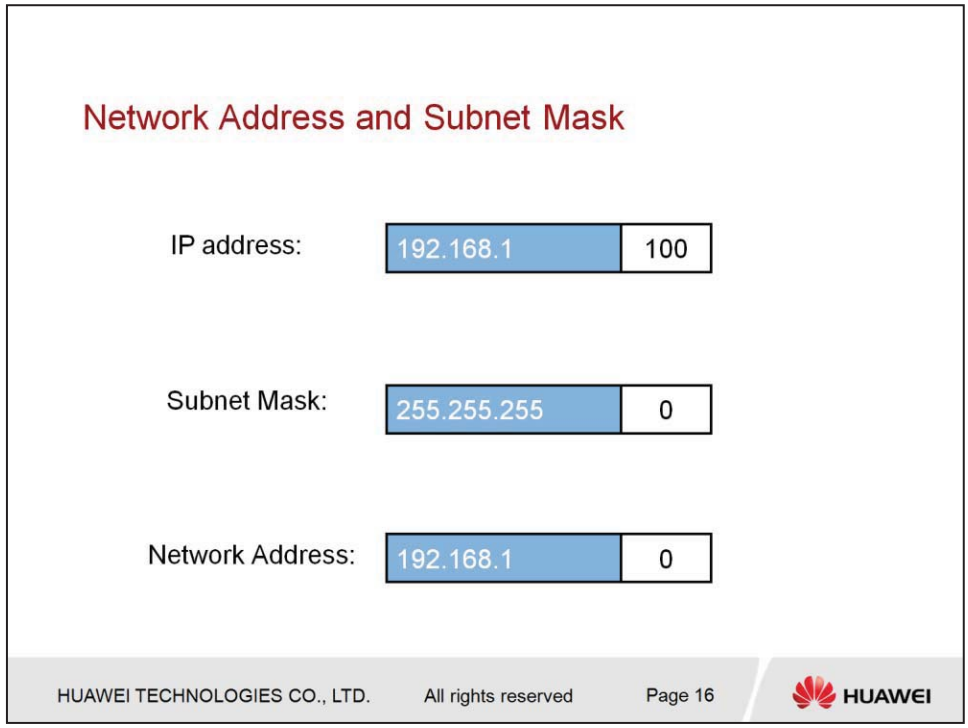
The IP address with the network ID being 127 is the loopback address, for example, 127.0.0.1 is used for loopback test.

The IP address of all 0s indicates all the hosts. On the Huawei ARG3 routers, IP address 0.0.0.0 specifies the default route.

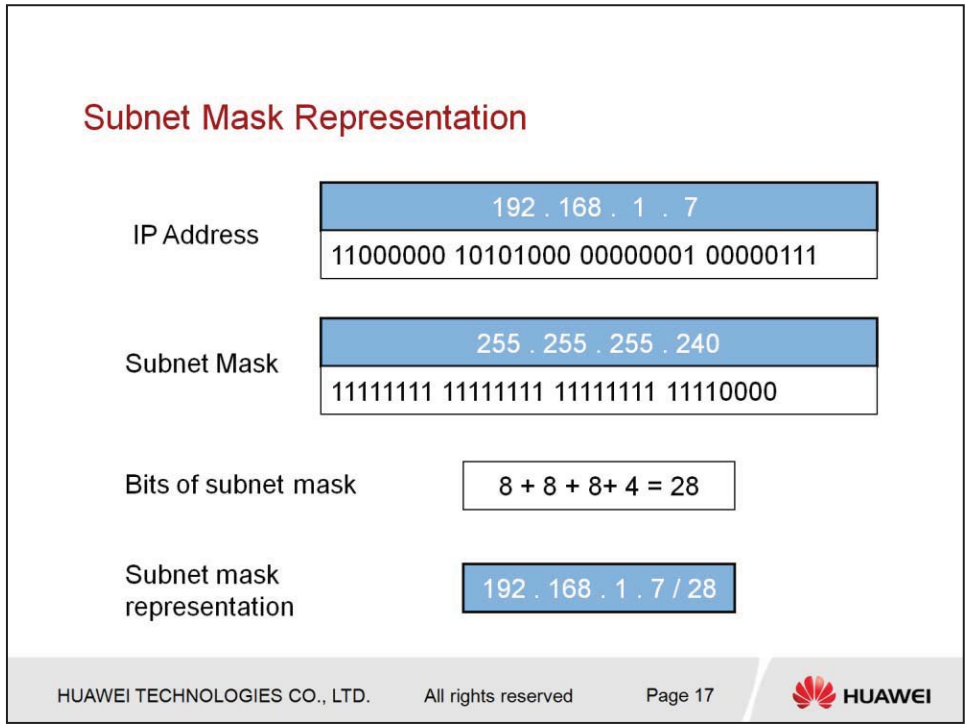
IP address 255.255.255.255 is also a broadcast address, but it stands for all hosts and is used to send packets to all the nodes



By default, the subnet is not divided. The natural mask of class A network is 255.0.0.0, the natural mask of class B network is 255.255.0.0, and the natural mask of class C network is 255.255.255.0. The IP addresses can be used more effectively with the subnets. Multiple subnets are divided within the network.

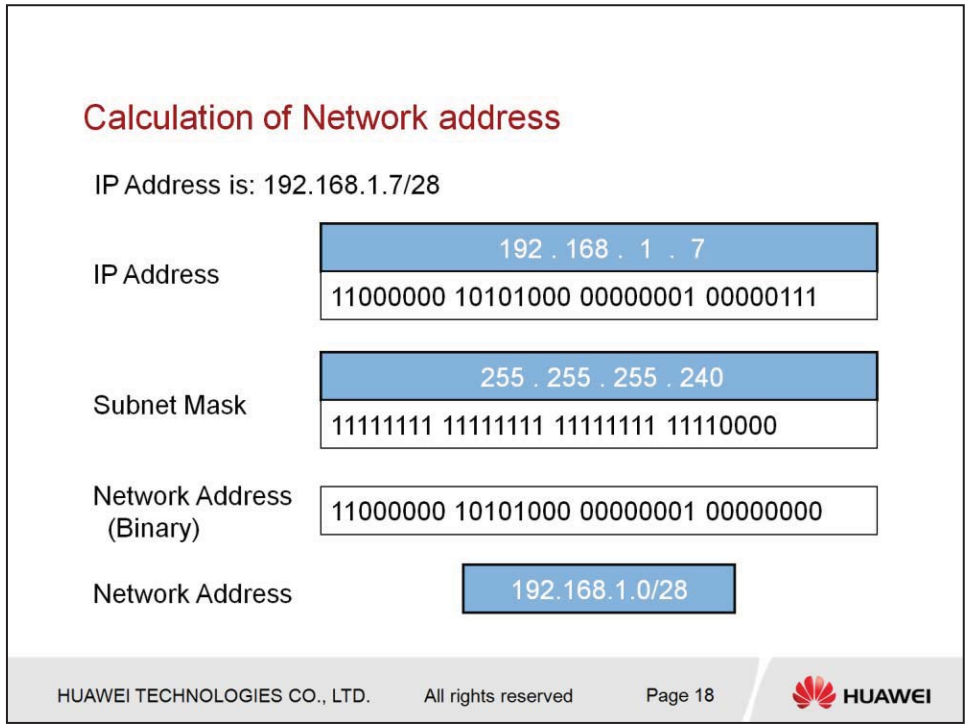


This is a class C address. The first 24 bits stand for the network ID and the last 8 bits stand for the host ID.



As shown in the figure above, the subnet can be represented in two ways.

255.255.255.240 and /28 indicate that the first 28 bits are the network ID.



As shown in the figure above , the IP address and subnet mask are already known.

Then, the network address is obtained from the AND operation between the IP address and the subnet mask. The AND operation is 1&1=1, 1&0=0, and 0&0=0.

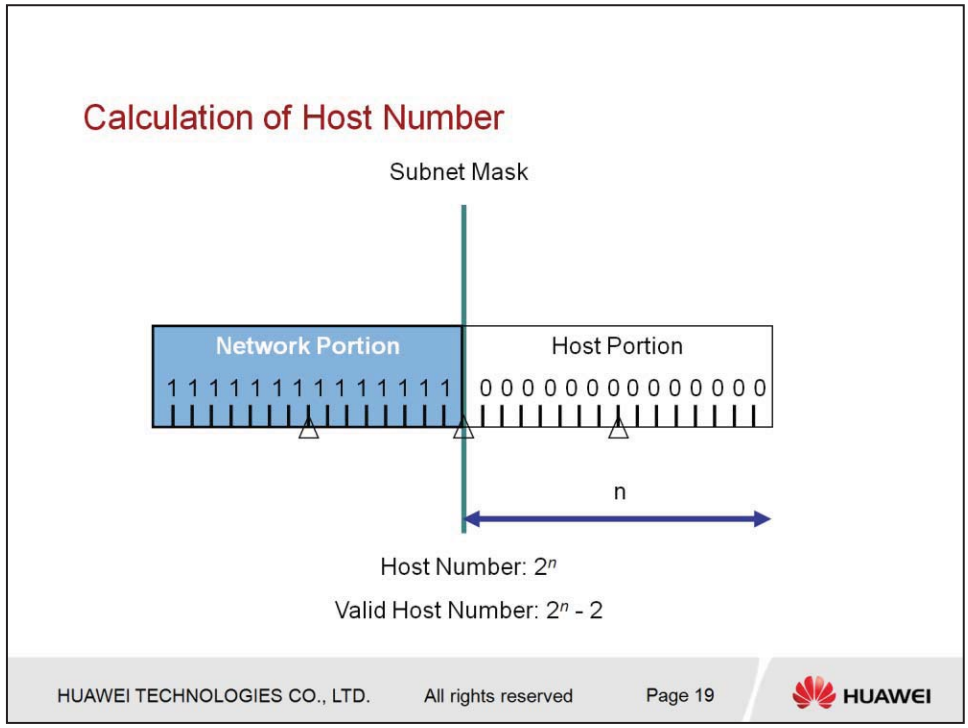
Therefore, the calculation of the AND operation is as follows:

11000000, 10101000, 00000001, 00000111

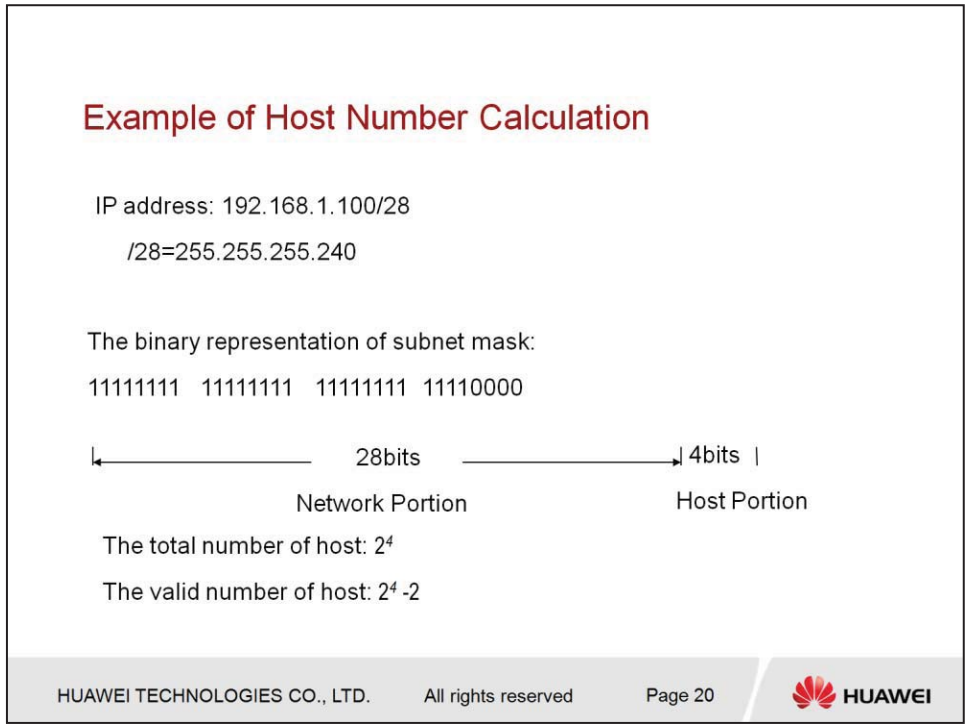
& 1111111. 11111111, 11111111, 11110000

11000000, 10101000, 00000001, 00000000

The calculation result is the network address.



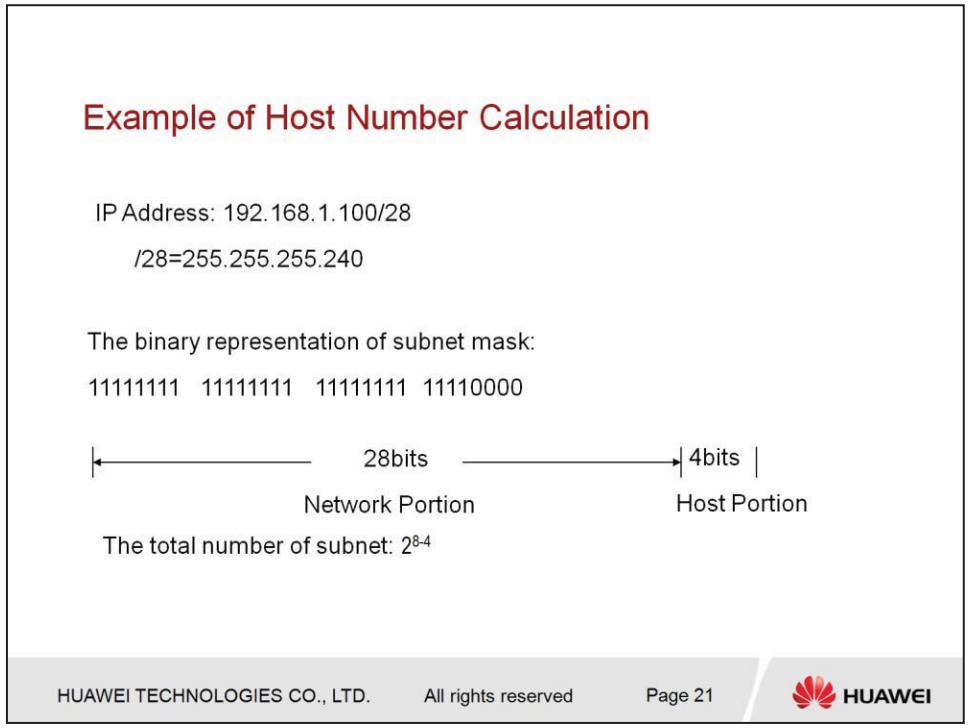
The number of hosts is calculated through the subnet mask. First, we should know that how many 0s are there in the last several bits. As shown in the above figure, if there are n 0s, then, the number of hosts is 2n. The number of IP addresses that can be allocated is 2n–2 (deducting the network address containing all 0s and the broadcast address containing all 1s).



This example shows the calculation of host quantity.

The subnet mask of class A address is 255.0.0.0, namely, the 24-bit host ID. The subnet mask of class B address is 255.255.0.0, namely, the 16-bit host ID. The subnet mask of class C address is 255.255.255.0, namely, the 8-bit host ID.

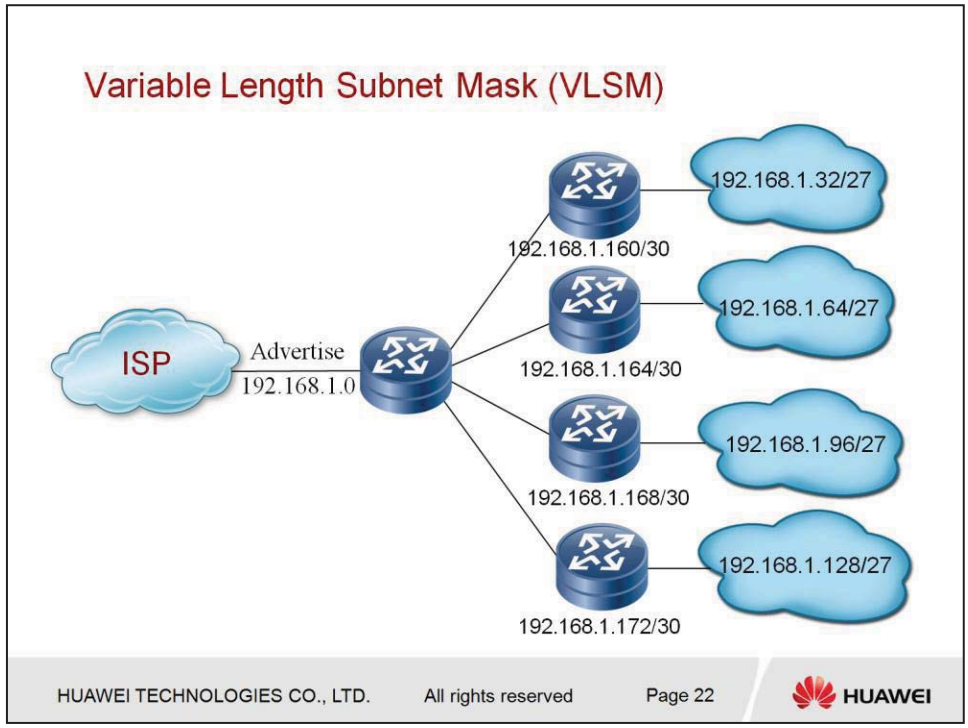
Take class C address for example. The subnet mask has an 8-bit host ID, in which the first 4 bits are also used as the subnet mask. Then, the maximum number of hosts is 28-4. 8 refers to the bits of host ID in the natural mask, and 4 refers to the bits of the subnet mask. The number of host ID bits is 8-4, and the valid number of hosts is 24-2 .



As we known, the natural mask of class A address is 255.0.0.0, namely, the host ID has 24 bits; the natural mask of class B address is 255.255.0.0, namely, the host ID has 16 bits; the natural mask of class C address is 255.255.255.0, namely, the host ID has 8 bits.

Take class C address for example. The standard subnet mask has 8-bit 0s (host ID). Then, the number of subnets is 28-4. 8 refers to the number of 0s (host ID) in the subnet mask, and 4 refers to the bits of the host ID. The number of subnet is

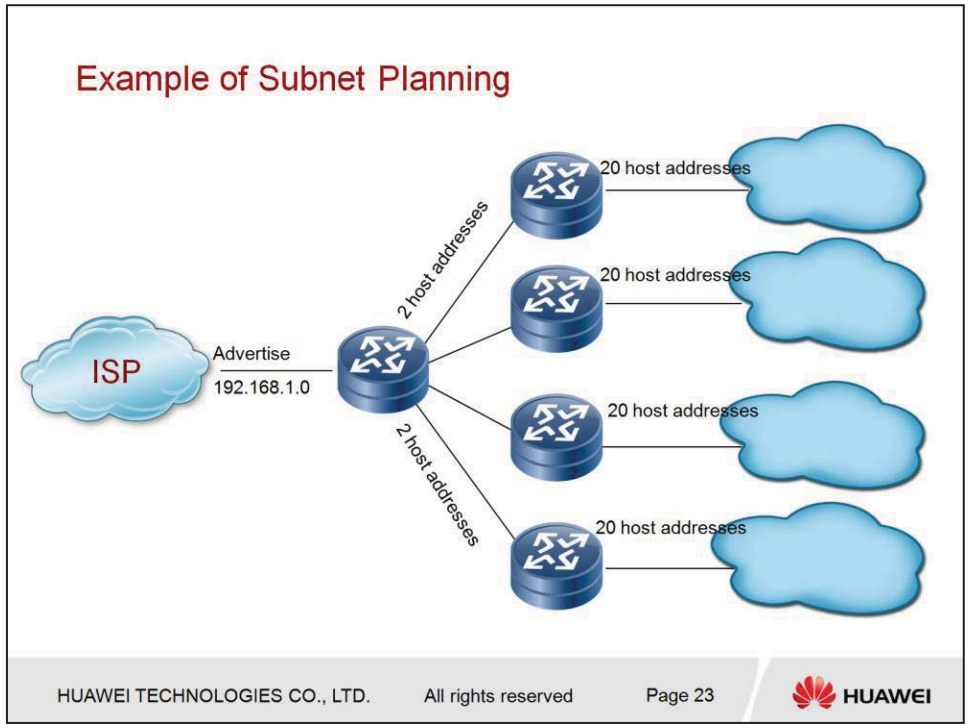
28-4. The number of class B addresses can be calculated similarly.



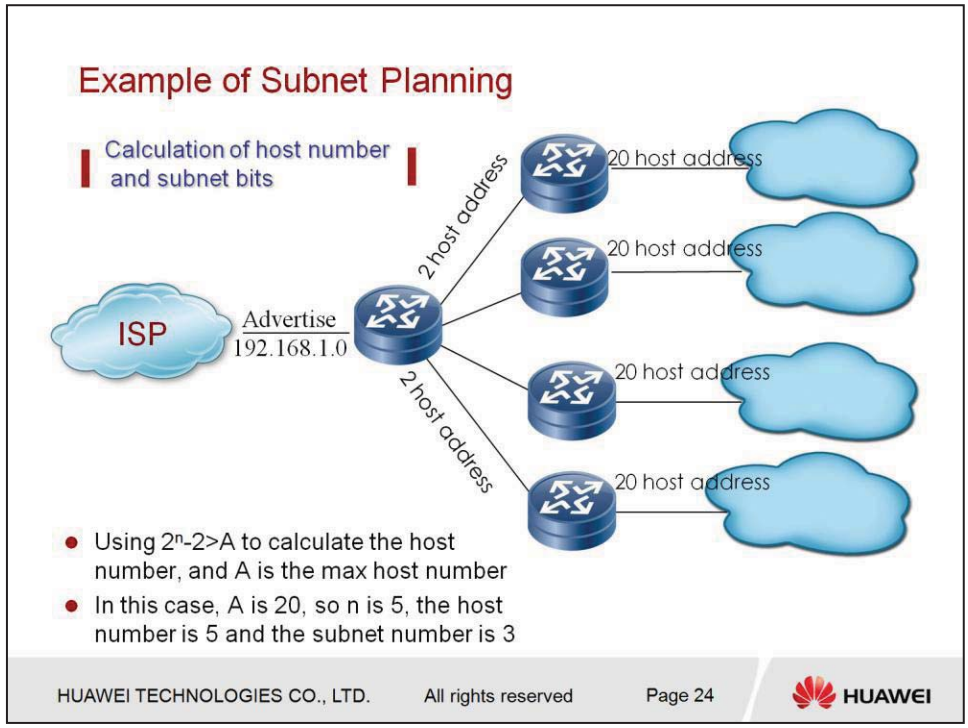
A network can be divided into multiple subnets, and each subnet uses a unique ID.

The numbers of hosts in the subnets may be different. If the length of subnet mask is fixed and the number of IP addresses in the subnets is the same, lots of IP addresses are wasted.

In this case, the variable length subnet masking (VLSM) technique can be used. If the subnet has lots of nodes, the subnet mask is short. The IP address with short subnet mask represents less networks/subnets, but more IP addresses can be allocated. If the subnet has a few nodes, the subnet mask is long. The IP address with long subnet mask represents more logical networks/subnets, but less IP addresses can be allocated. Such addressing scheme can save lots of IP addresses, which can be used in other subnets.



As shown in the above figure, a company deploys the IP addresses with class C address 192.168.1.0/24. The company has bought five routers. A router, which works as the gateway of the intranet, is connected to the local ISP. The other four routers are connected to four branch offices. Each office has 20 PCs, so each office needs 20 host address. How to plan the IP addresses?

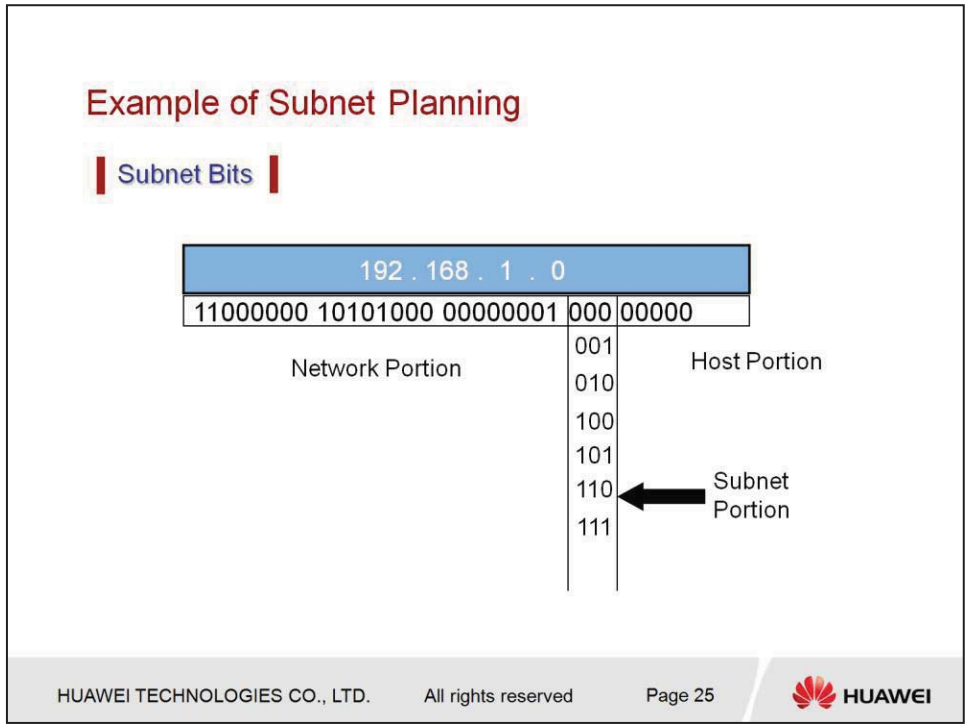


Step 1: Get to know how many subnets are needed and how many hosts are required in each subnet. Calculate the subnet bits and host bits with the formula 2n-2>A (A refers to the maximum number of host).

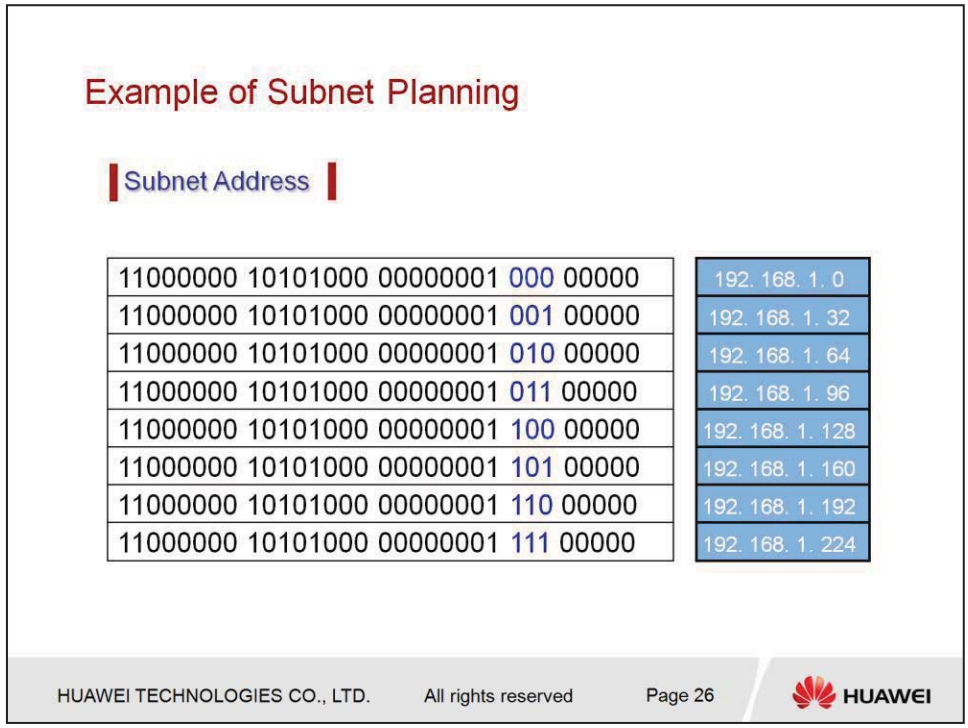
As shown in the above figure, eight subnets are required. Four offices need 21 IP addresses (including a router interface). The 4 network segments connected with the gateway need 2 IP addresses. In this example, the IP addresses of the four

offices are planned first, and then, the IP addresses of the interfaces between the four routers and the gateway are planned.

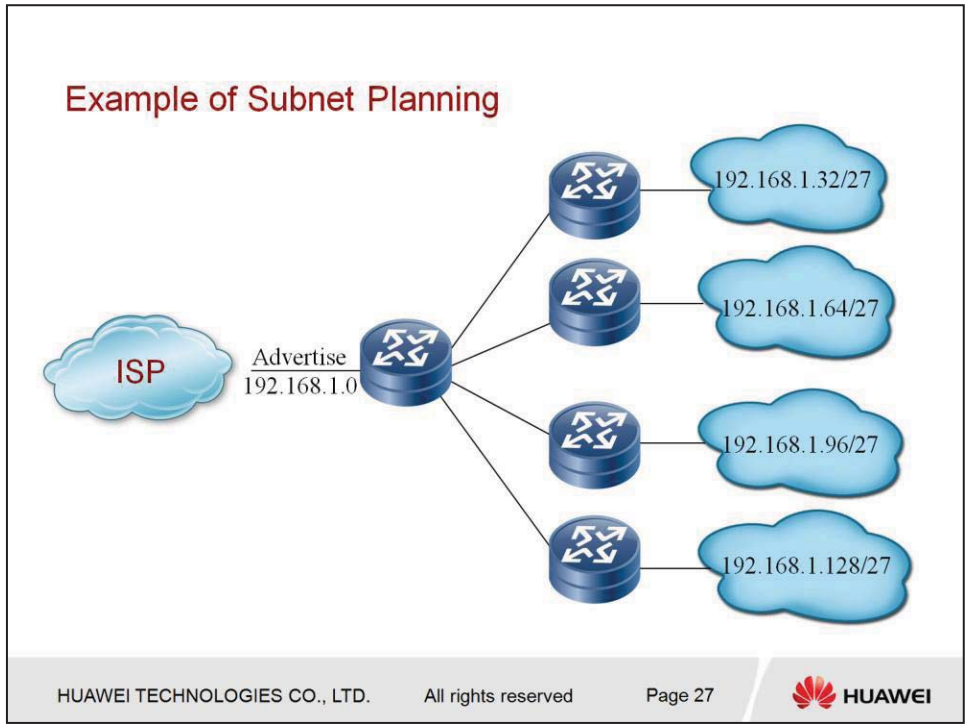
According to the formula 2n-2>A, if A is 20, then, n is 5. That is, the number of host bits is 5 and the number of subnet bits is 3. Therefore, the number of host bits is 5 in each office.



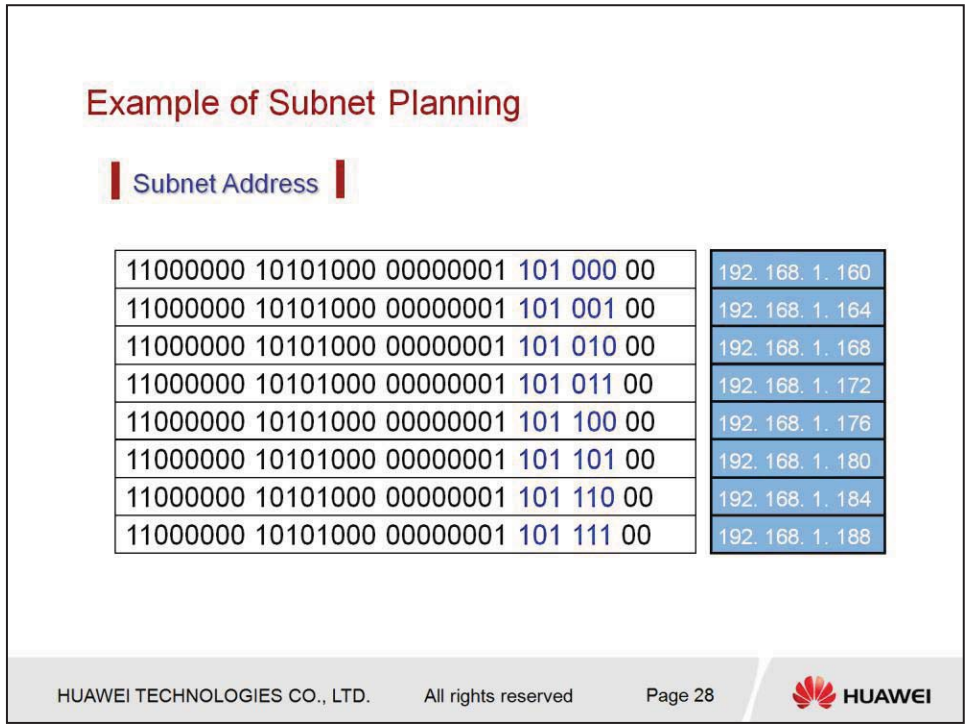
Step 2: In the calculation of subnet portion, the host portion of IP address 192.168.1.0 is divided into subnet portion and new host portion. According to the calculation result, the subnet portion contains three bits, which is represented in binary notation. The vertical lines mark the subnet space. All the combinations of the subnet portion are listed from 000.



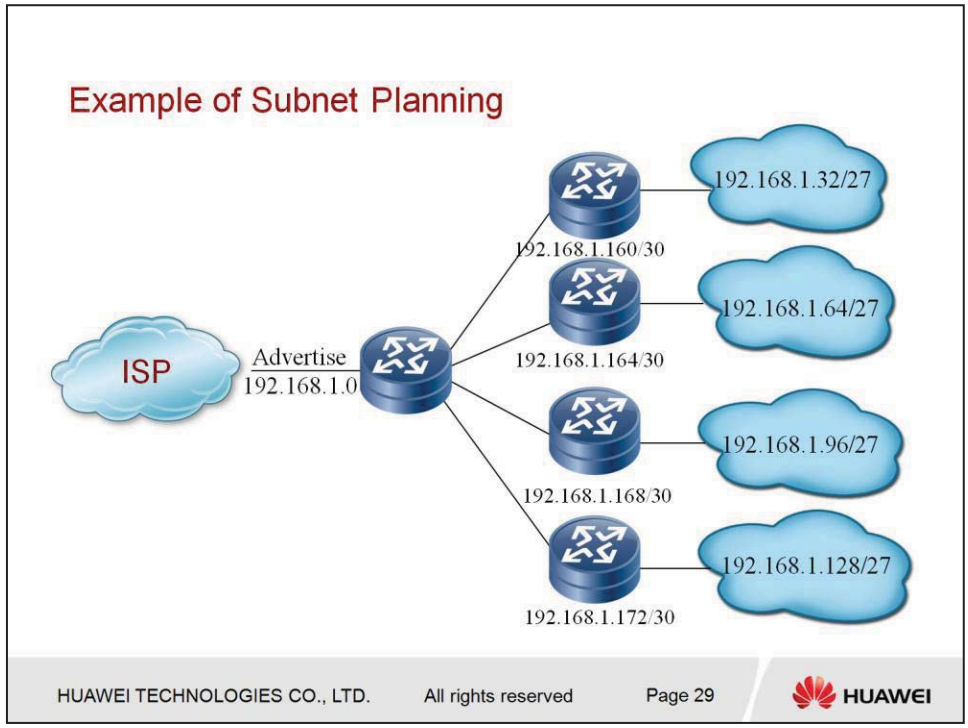
Step 3: As shown in the figure above, the network segment addresses in the right part are the calculation result represented in dotted decimal notation.



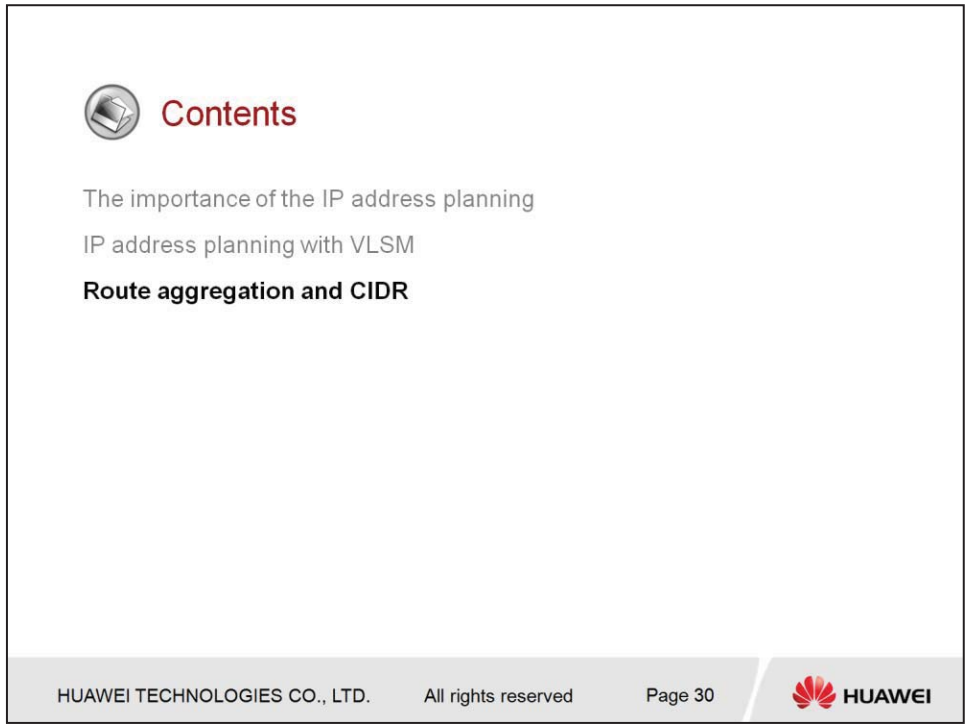
Select some of the network segment addresses calculated in Step 3 as the final result. In this example, network segments 192.168.1.32/27, 192.168.1.64/27, 192.168.1.96/27 and 192.168.1.128/27 are selected.

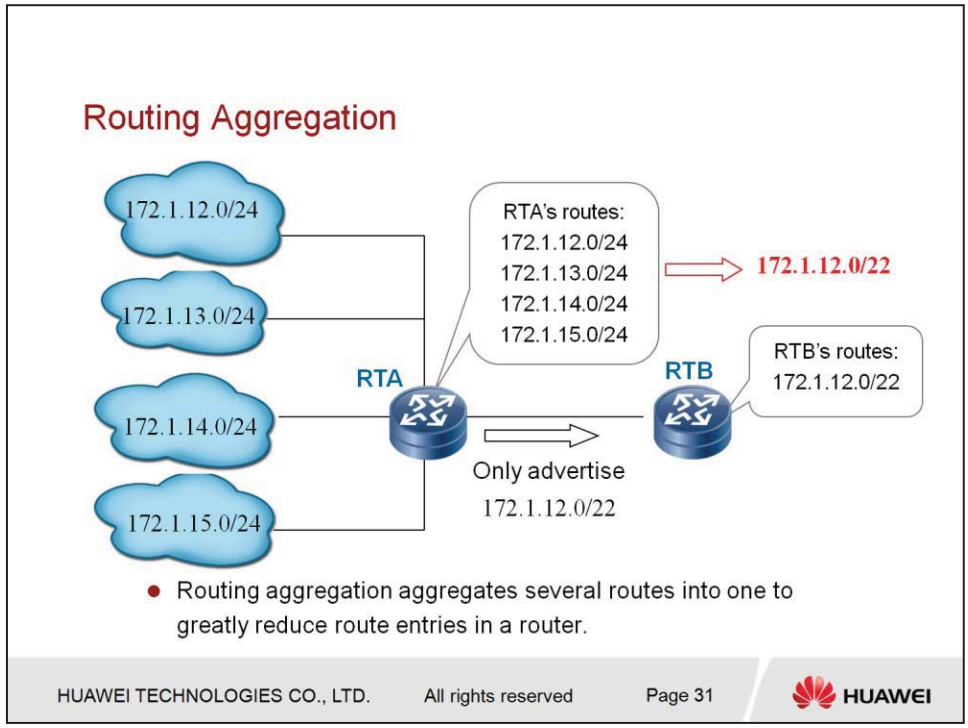


Select network segment 192.168.1.160 to plan new subnet. The IP address is used as the subnet address between the four offices and the gateway. Through the calculation, the subnet address is as shown in the figure above .



The figure above shows the final subnet planning.

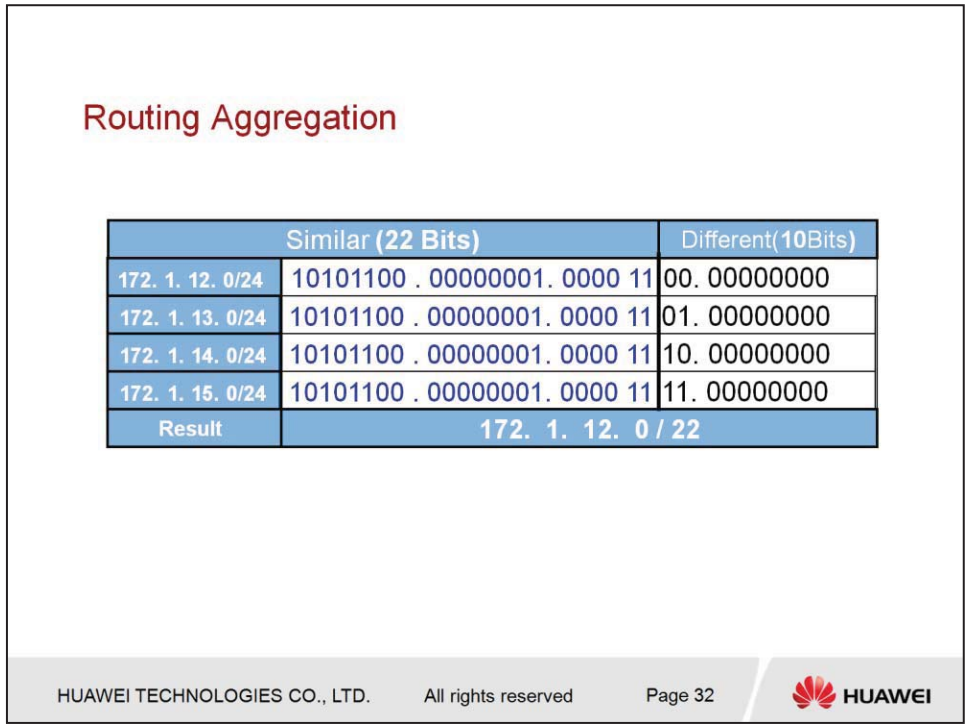

Contents
The importance of the IP address planning
IP address planning with VLSM
Route aggregation and CIDR
HUAWEI TECHNOLOGIES CO., LTD.
All rights reserved
Page 30
HUAWEI



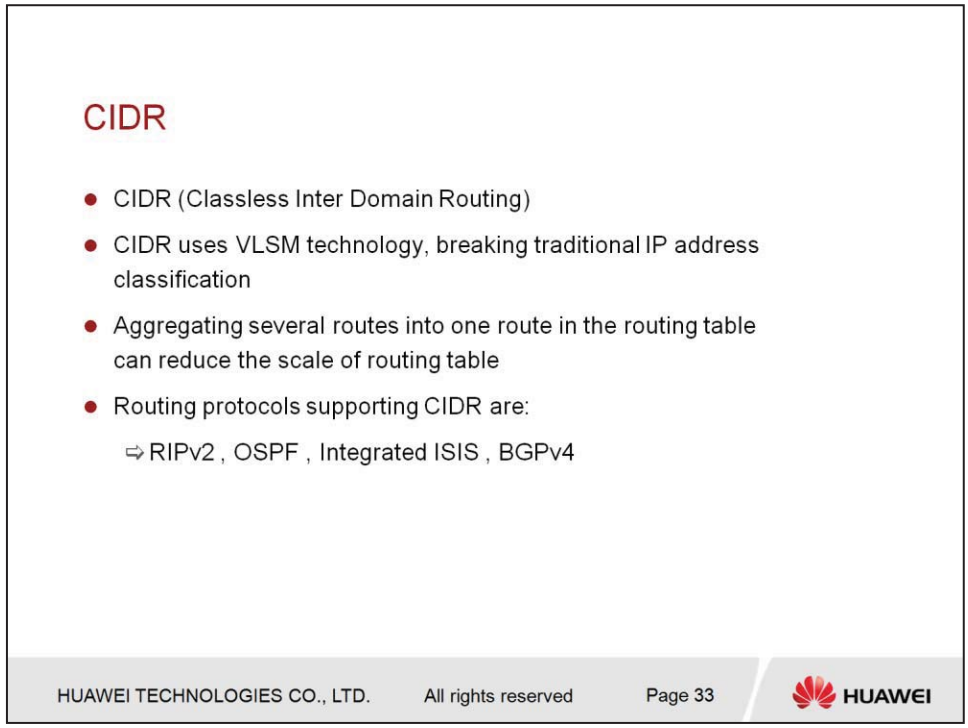
As shown in the above figure, RTA is connected to four network segments 172.1.12.0/24, 172.1.13.0/24, 172.1.14.0/24 and 172.1.15.0/24. Then, RTA stores the routes of the four network segments, and the routes are aggregated on RTA.

The routes can be aggregated into a single-hop route 172.1.12.0/22. Then, RTA advertises only route 172.1.12.0/22 to RTB. It can greatly reduce the number of routes.

Routing aggregation is to aggregate multiple routes into one . Routing aggregation can greatly reduce the number of routes, reduce the workload of route maintenance, and improve the utilization of the network.

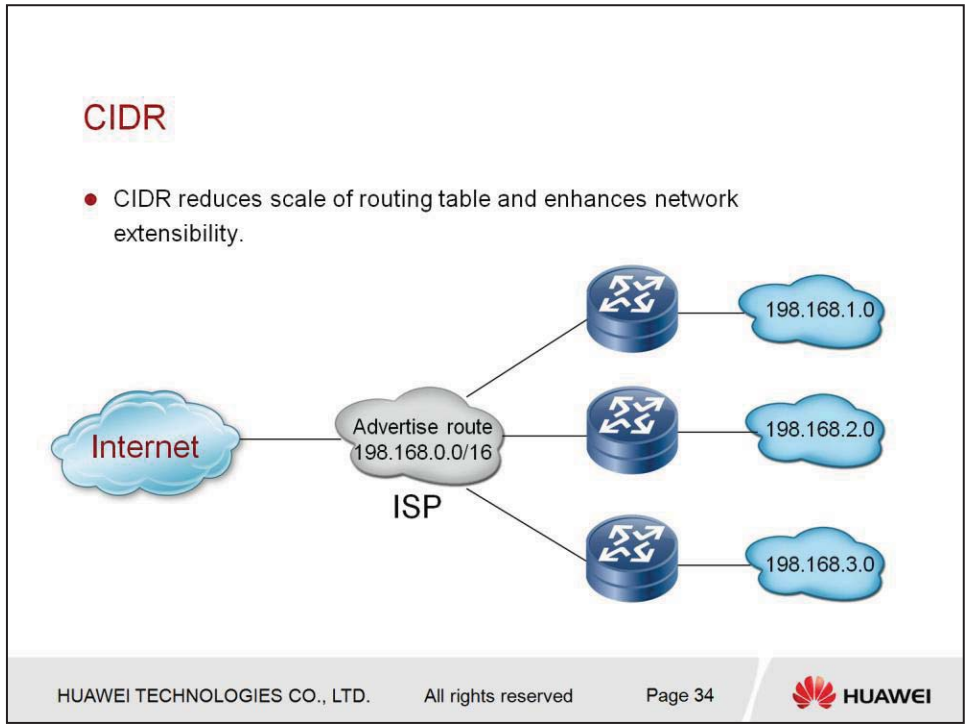


As shown in the figure above, routes 172.1.12.0/24, 172.1.13.0/24, 172.1.14.0/24 and 172.1.15.0/24 can be aggregated into route 172.1.12.0/22.



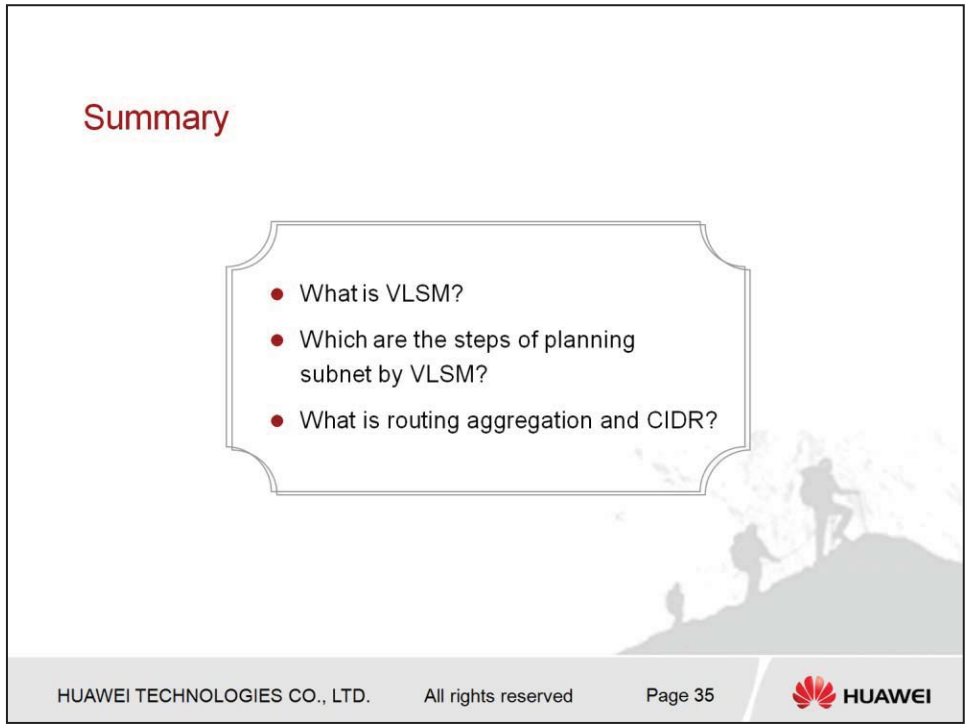
Classless Inter Domain Routing (CIDR), defined by RFC 1817, uses the VLSM technique to aggregate multiple routes into one. It does not adhere to the IP address classification rule. This can minimize the size of the routing table and improve the extensibility of the router.

The routing protocols supported by CIDR are RIPv2, OSPF, Integrated ISIS, and BGPv4.



As shown in the above figure, some class C networks are allocated to the ISP, 198.168.0.0-198.168.255.0. The ISP allocates the class C networks to the user groups. At present, three class C networks have been allocated to user groups. If the CIDR technique is not used, the routing table of the ISP's router has three routes connected to the downlink network segments, and the routes are advertised to the routers on the Internet. With the CIDR technique, the three routes 198.168.1.0/24, 198.168.2.0/24, and 198.168.3.0/24 can be aggregated into route 198.168.0.0/16. In this case, the ISP's router advertises only route 198.168.0.0/16 to the Internet, so the number of entries in the routing table is reduced.

Notes: that the number of bits in the network addresses aggregated by CIDR must be the same. As shown in the above figure, if the ISP is connected to network segment 72.178.1.0/24, the routes of the network segments cannot be aggregated and the CIDR technique cannot be implemented.



What is VLSM?

VLSM refers to the variable length subnet masking technique. If the subnet has lots of nodes, the subnet mask is short. The IP address with short subnet mask represents less networks/subnets, but more IP addresses can be allocated. If the subnet has a few nodes, the subnet mask is long. The IP address with long subnet mask represents more logical networks/subnets, but less IP addresses can be allocated. Such addressing scheme can save lots of IP addresses, which can be used in other subnets.

What are the steps of planning subnet by VLSM?

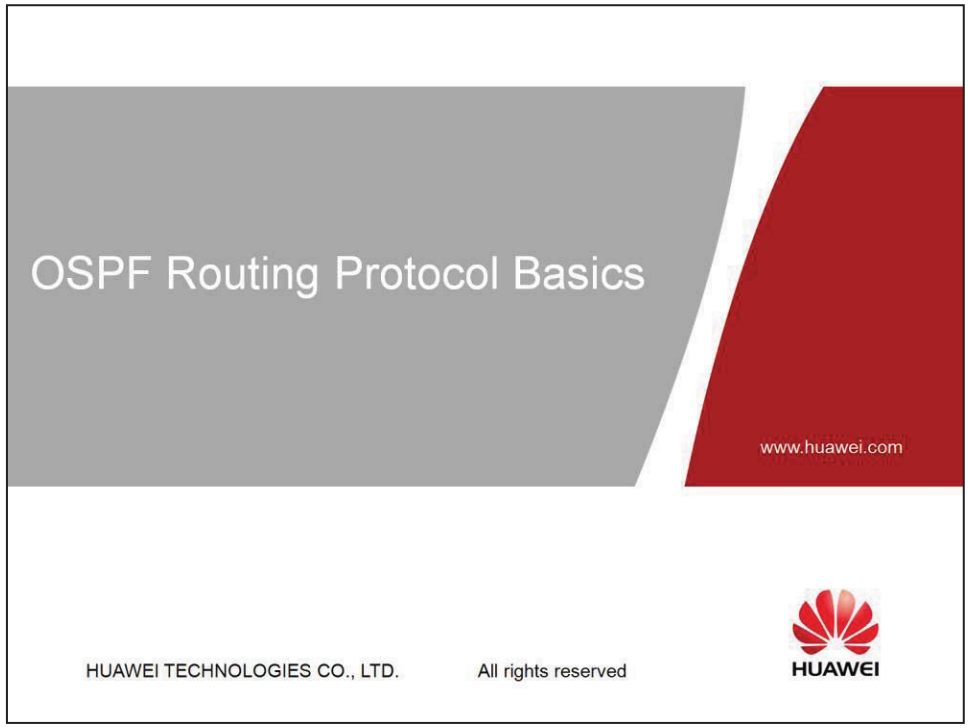
Four steps. Step 1: Get to know the number of subnet bits and host bits. Step 2:

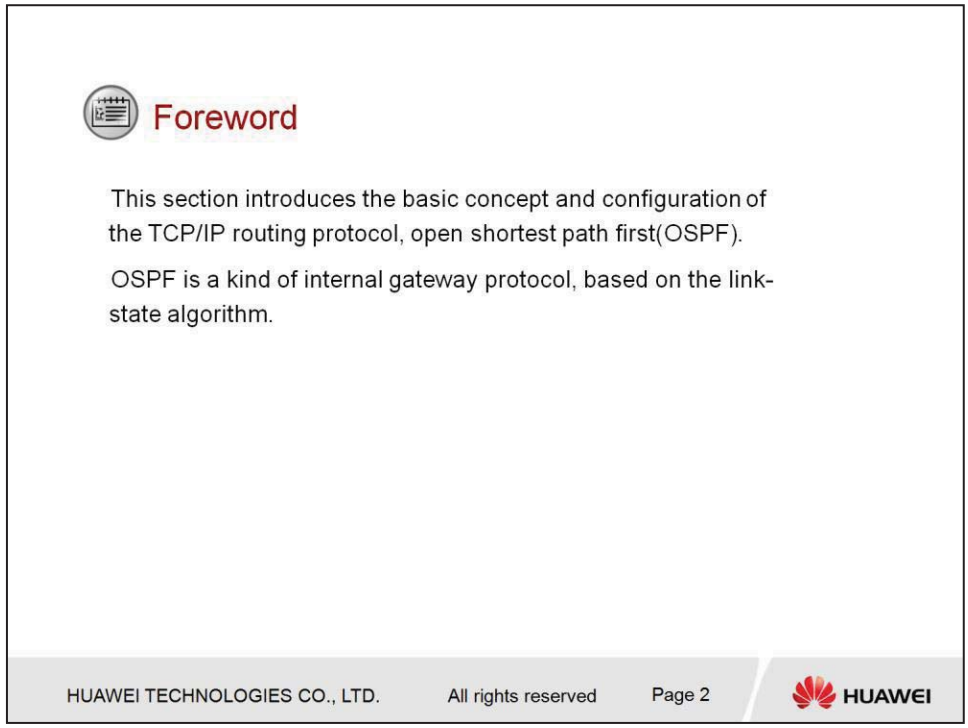
Calculate the subnet bits. Step 3: Calculate the subnet address. Step 4: Select the subnet address and obtain the final result.

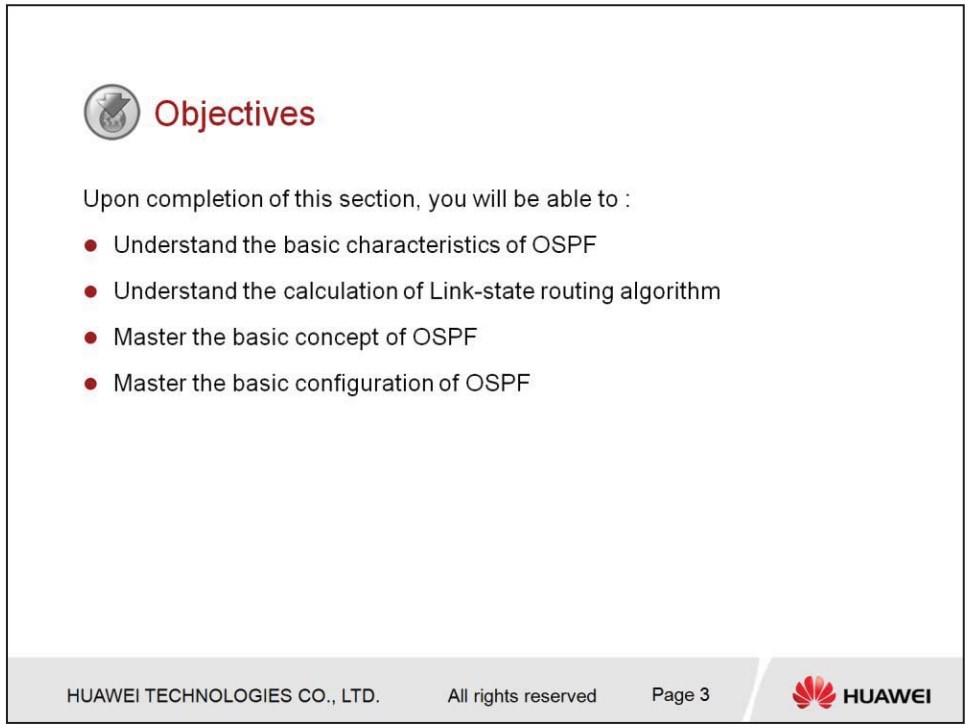
What are routing aggregation and CIDR?

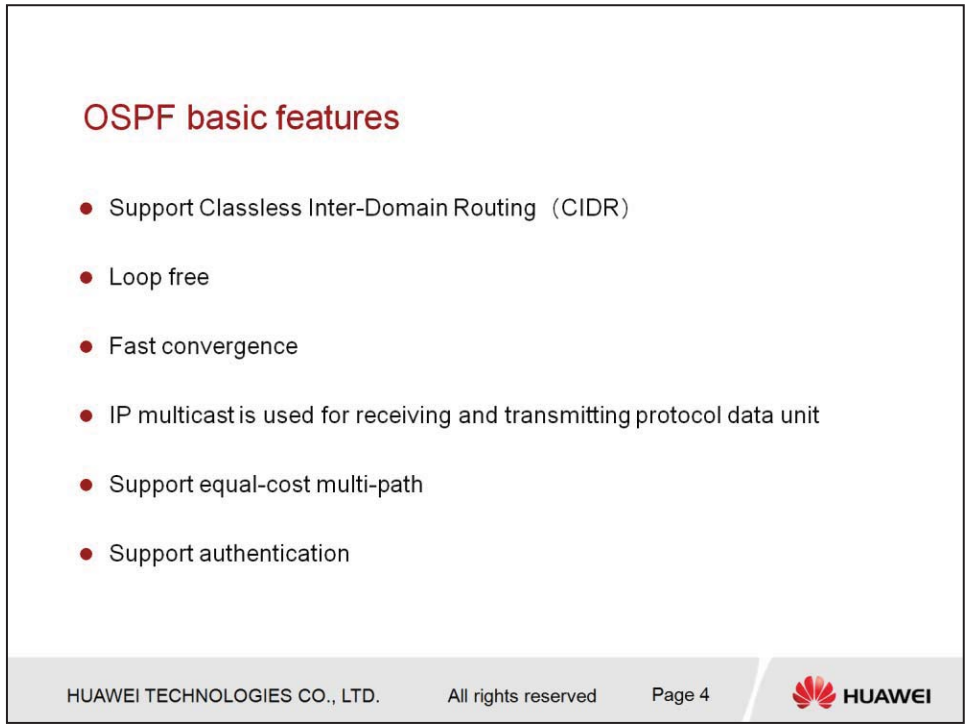
Routing aggregation is to aaggregation can greatly reduce the number of routes, reduce the workload of route maintenance, and improve the utilization of the network.

Module 2
OSPF


OSPF Routing Protocol Basics
www.huawei.com
HUAWEI TECHNOLOGIES CO., LTD.
All rights reserved
HUAWEI


Foreword
This section introduces the basic concept and configuration of the TCP/IP routing protocol, open shortest path first(OSPF).
OSPF is a kind of internal gateway protocol, based on the link-state algorithm.
HUAWEI TECHNOLOGIES CO., LTD.
All rights reserved
Page 2
HUAWEI


Objectives
Upon completion of this section, you will be able to :
Understand the basic characteristics of OSPF
Understand the calculation of Link-state routing algorithm
Master the basic concept of OSPF
Master the basic configuration of OSPF
HUAWEI TECHNOLOGIES CO., LTD.
All rights reserved
Page 3
HUAWEI



OSPF has the following features：

Support Classless Inter-Domain Routing（CIDR）：

OSPF is the routing protocol developed for TCP/IP network environment, which support Classless Inter-Domain Routing (CIDR) and Variable Length Subnet Mask (VLSM)

Loop Free:

The route calculation is based on detailed link state information, so the routes produced by OSPF are loop free.

Fast Convergence：

The route update is triggered by topology changes. Once the network topology is changed, the new link state information is flooded

Sending and receiving protocol data using multicast：

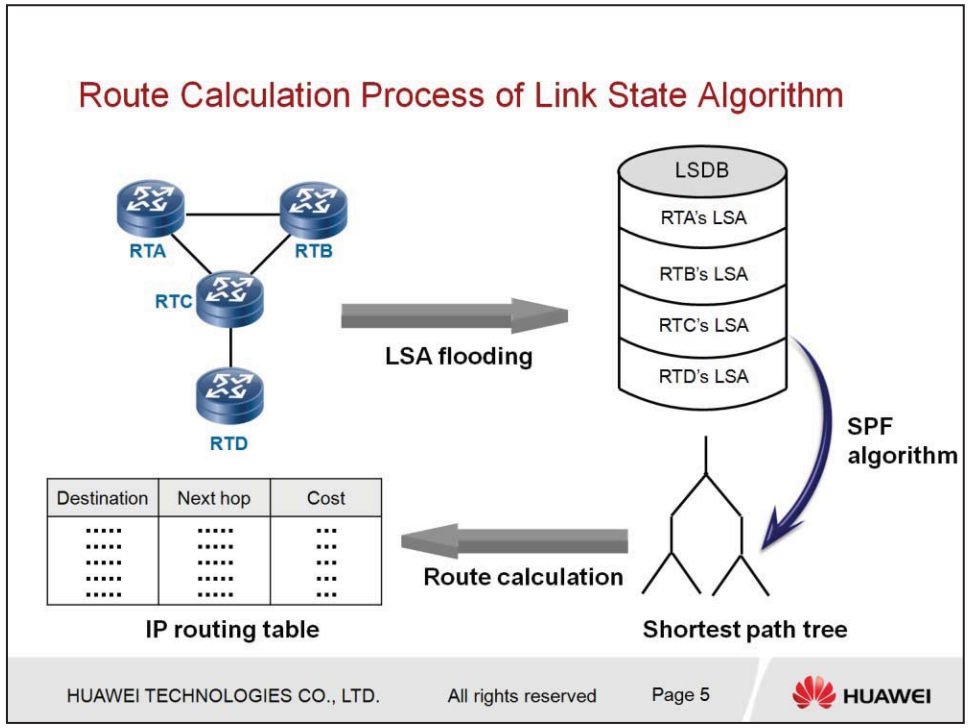
OSPF routers use multicast to send and receive protocol data, which take up the network resources is very small.

Support equal-cost multi-path：

If multiple equal-cost routes to a destination exist, the traffic load is shared equally on these paths.

Support authentication:

Each OSPF packet exchange is authenticated.



The most distinctive feature of OSPF is that it is based on Link State Algorithm, which is different from Distance Vector Algorithm used by previous routing protocols. In this course, the basic calculation process of Link State Algorithm is first introduced to better understand OSPF principles.

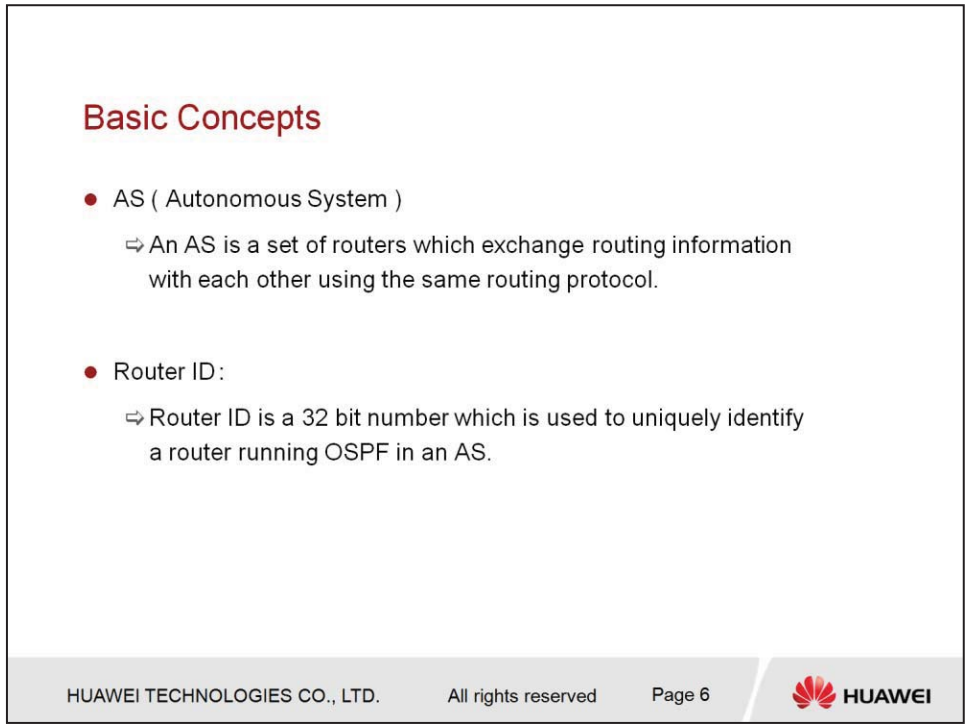
Each router advertises local link state information by flooding LSA, the link state information includes: available interfaces, reachable neighbors and attached networks etc.

Each router collects LSAs generated by other routers and itself to form a Link State Database (LSDB). LSDB gives the detailed information about network topology of routing domain.

All router's LSDB are the same.

According to LSDB, each router treats itself as the root and other routers as leaves to calculate the Shortest Path Tree.

The Shortest Path Tree calculated by each router gives route table entries destined for other routers on the network.



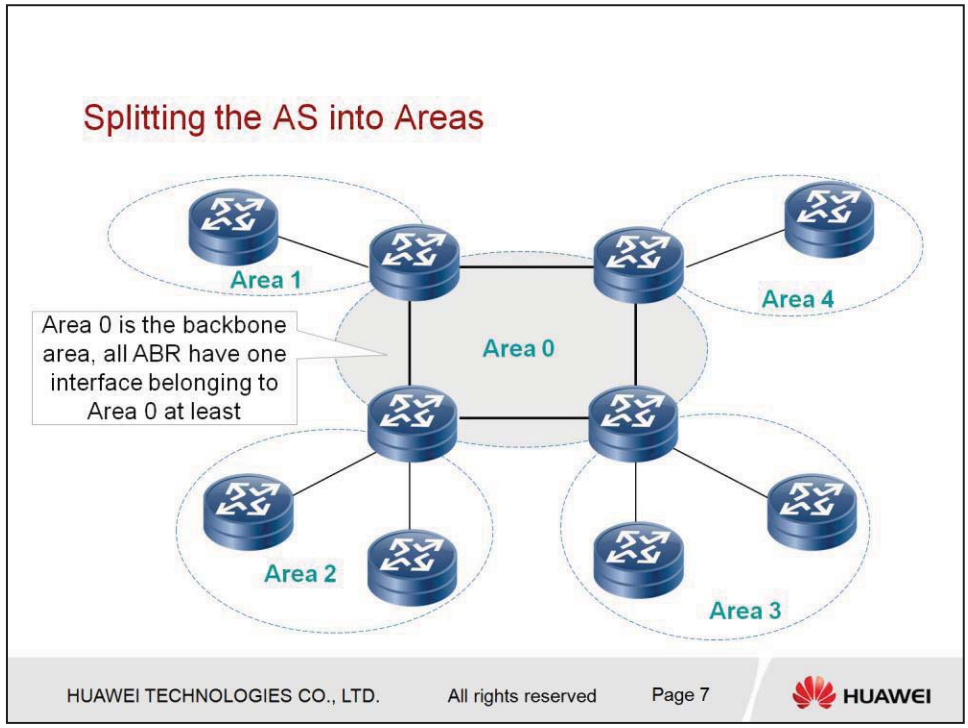
In OSPF, there are two important concepts, one is Autonomous System (AS), which is also called OSPF routing domain; the other is Router ID.

In OSPF, Autonomous System (AS)is a set of routers that exchange routing information using the same routing protocol.

LSDB describes the topology of the whole network, including all the routers on the network, so each router on the network should be identified uniquely.

Router ID is 32 bit number that is used to uniquely identify a router running OSPF in an AS. Each router running OSPF must have a Router ID.

Router ID uses the same format with that of IP address, it is recommended to use IP address of Loopback interface as Router ID.



Area is a set of networks.

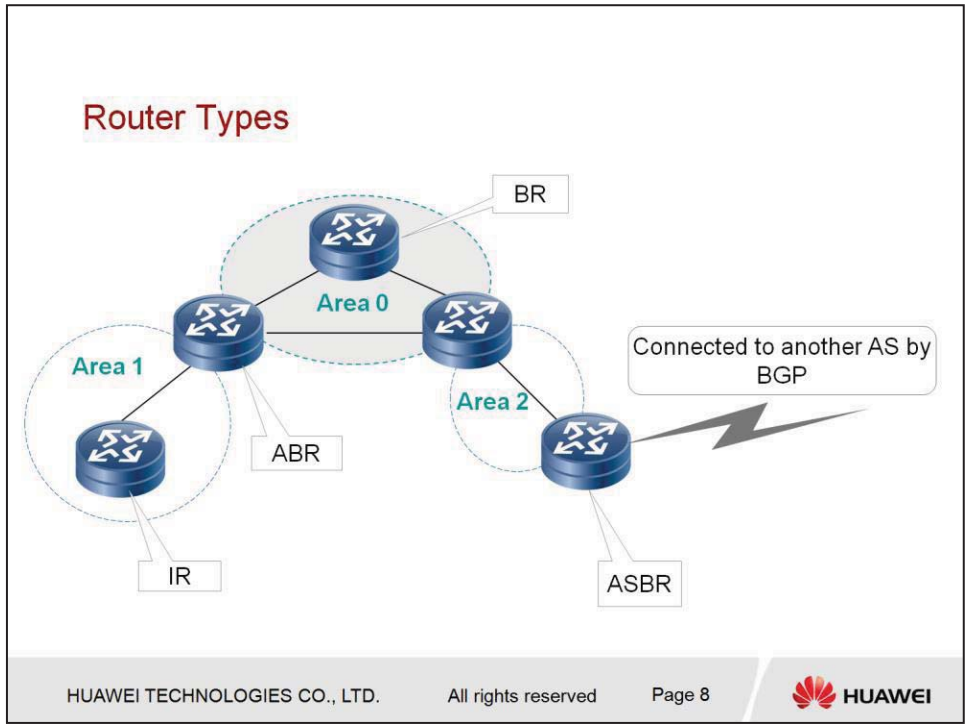
OSPF supports the combination of a group of networks, this group can be called an Area, that is to say, Area is a set of networks. Area division reduces the size of LSDB, thus reduce the network traffic The topology information within an area is not flooded into other areas. Abstract routing information instead of detailed link state information describing topology is transmitted between areas. Each area has it's own LSDB, different areas have different LSDB.

The router maintains a single LSDB for each area to which the router is attached.

The size of LSDB is greatly reduced because the detailed link state information is not advertised out of the area.

Area 0 is called Backbone Area, which is responsible for advertising routing information summarized by ABR (not detailed link state information) between non-backbone areas. In order to avoid inter-area route loops, non-backbone areas are not allowed to advertise inter-area routing information directly to each other.

Thus, all ABRs are required to have at least one interface belonging to Area 0, that is to say, each non-backbone area must be attached to the Backbone Area.



Internal Router:

Internal Router (IR) is the router whose attached networks are all within the same area. The Internal Routers within the same area maintain the same LSDB.

Area Border Router:

Area Border Router (ABR) is the router that is attached to more than one areas.

ABR maintains one LSDB for each attached area.

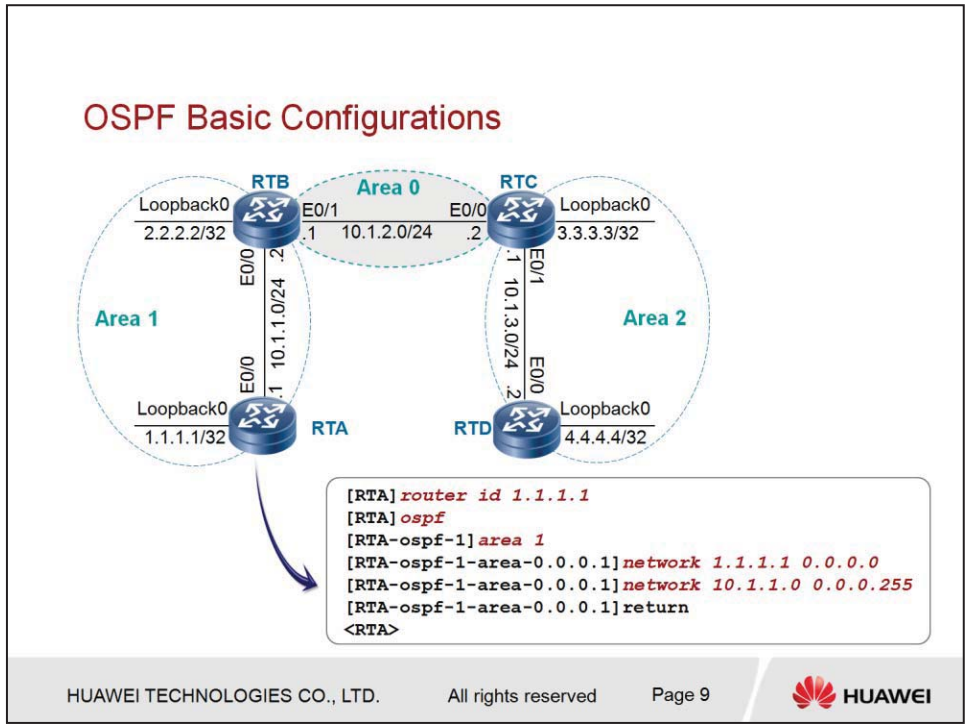
Backbone Router:

Backbone Router is the router that has more than one interfaces attached to the Backbone Area. All ABRs and the routers whose all interfaces are within Backbone Area are Backbone Routers.

AS Boundary Router:

AS Boundary Router (ASBR) is the router that exchanges routing information with routers in other AS. ASBR advertises AS external routing information into the whole routing domain.

An ASBR can be either an IR or an ABR, it can belong to either Backbone Area or non-backbone area.



Topology description：

There are four routers on the network, each router uses interface IP address of Loopback 0 as its Router ID, the whole routing domain is splitted into three areas.

RTB and RTC are ABRs.

IP address configuration is omitted here.

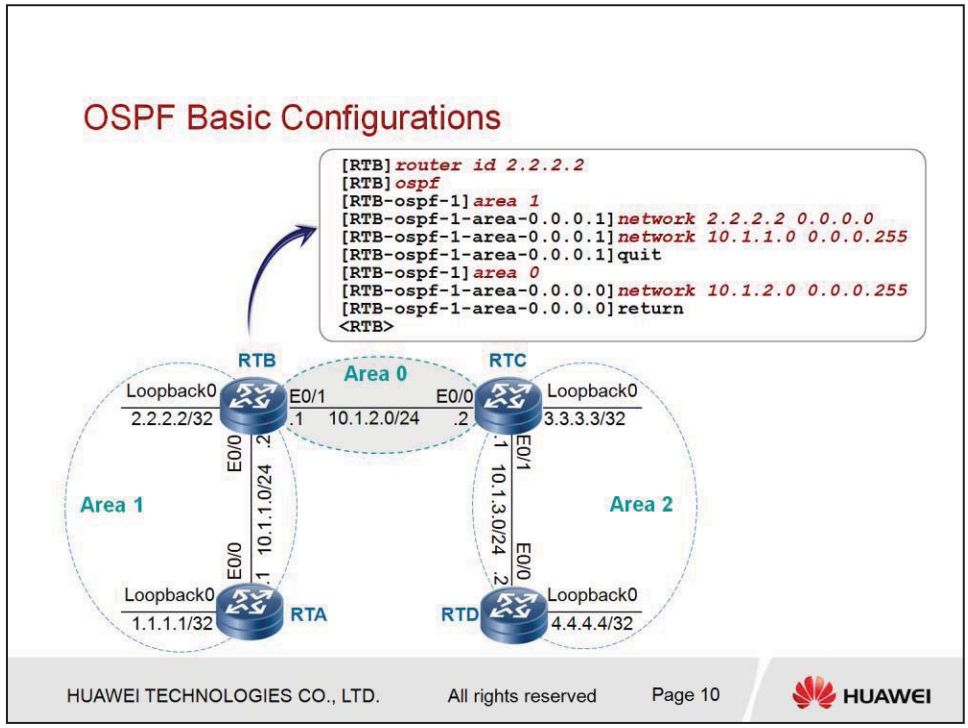
OSPF basic configuration include:

router id router-id : to specify router's Router ID. If Router ID is not specified manually, the numerically highest IP address on any of Loopback interfaces is used, if no Loopback interfaces are configured with IP addresses, the numerically highest IP address on any of physical interfaces is used.

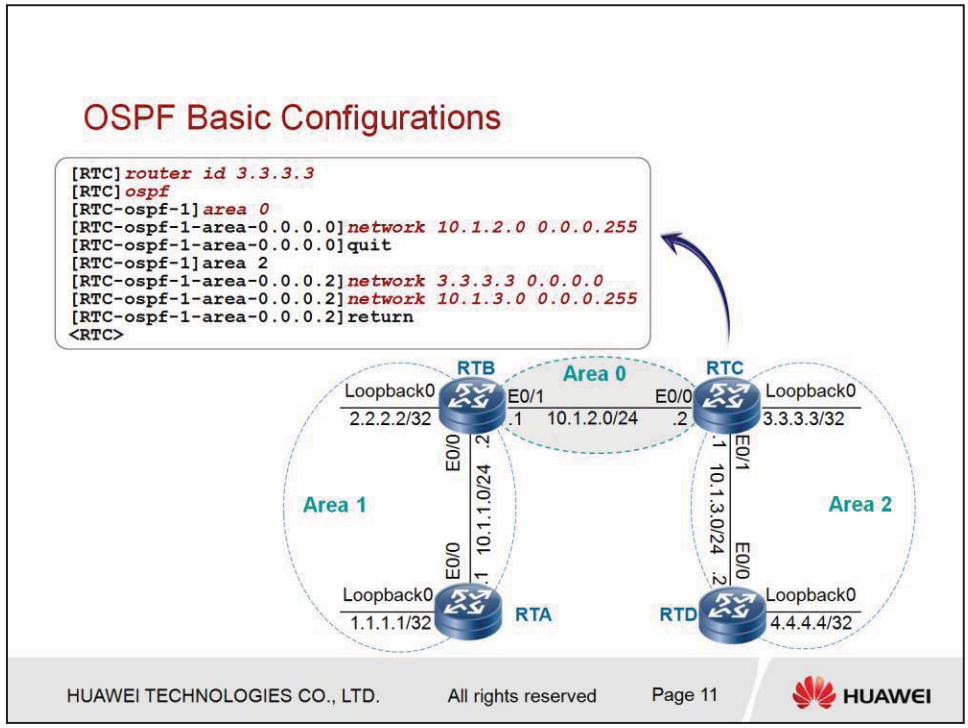
ospf process-id: to enable OSPF. OSPF supports multiple process, if no process ID is specified, the default process ID 1 is used.

area area-id: to enter Area view.

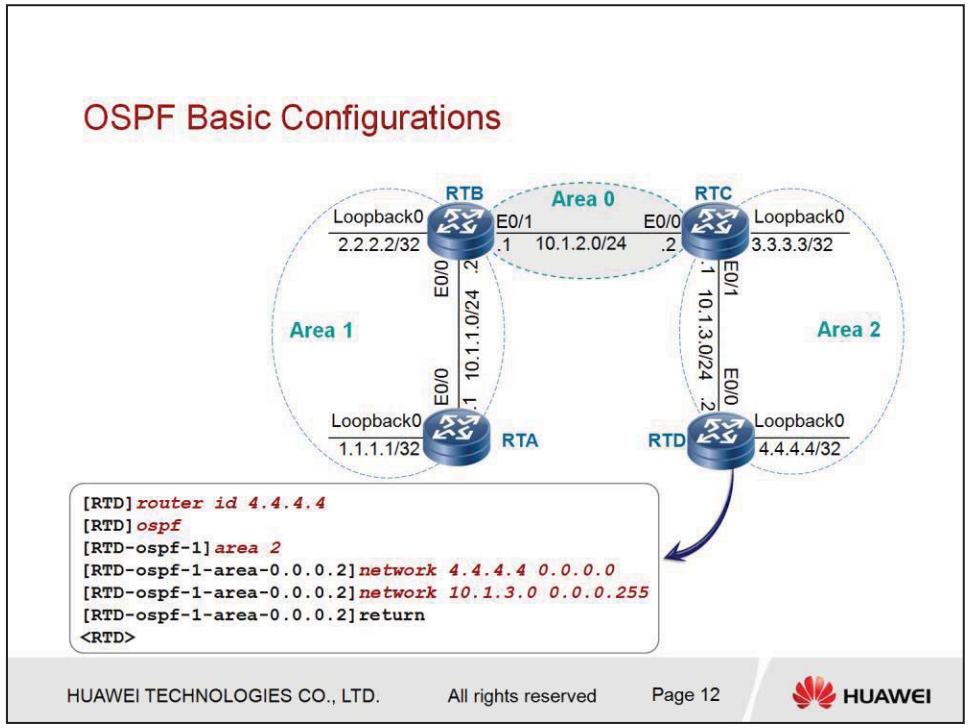
network ip-address wildcard: to specify which networks are attached to the area, the masks of networks should be inverse masks.



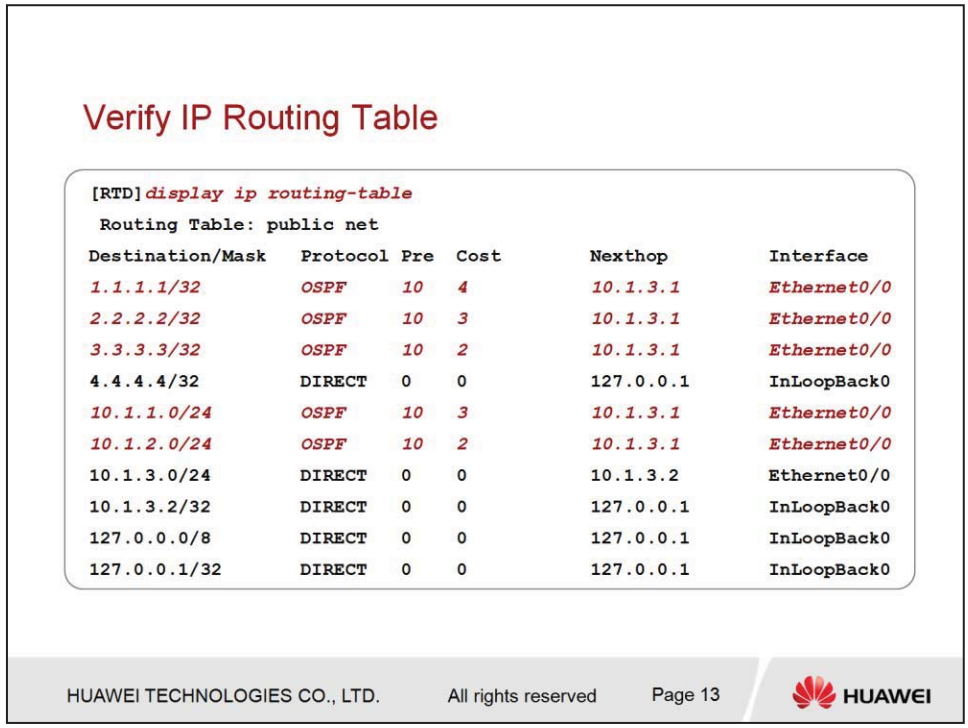
On RTB, two areas are configured, one is Backbone Area, the other is nonbackbone area.



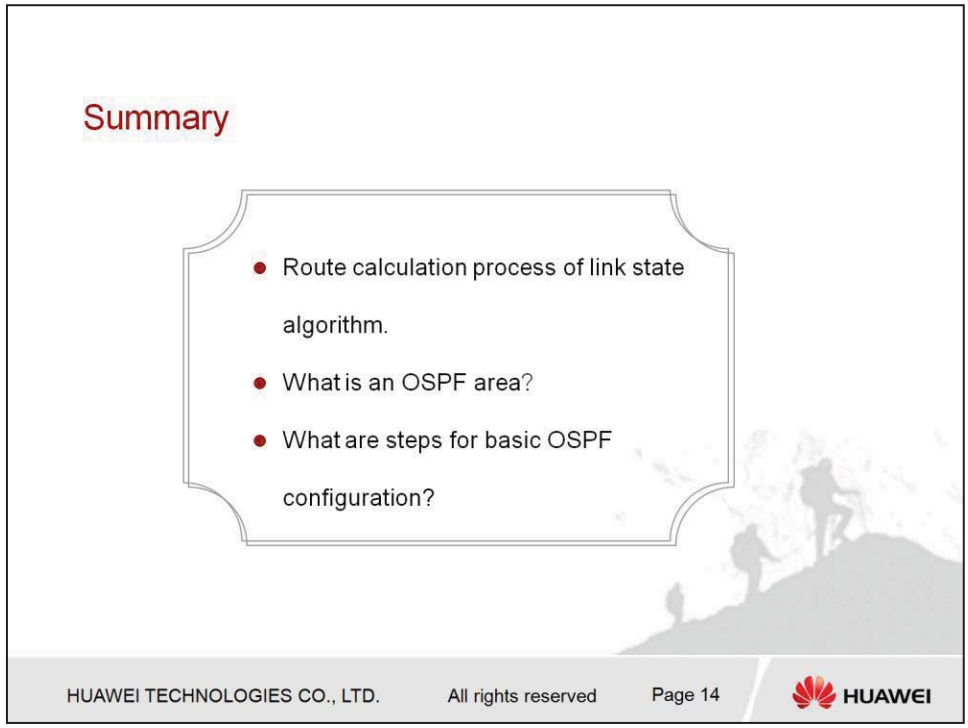
On RTC, two areas are configured, one is Backbone Area, the other is nonbackbone area.



There is only one area configured on RTD.



There are 5 route entries learned by OSPF in IP routing table.



What's the route calculation process of link state algorithm?

Each router advertise local link state information within the area, at the same time, collect the link state information advertised by other routers, thus the LSDB describing the network topology is formed. According to the LSDB, routers get a Shortest Path Tree by using Shortest Path First algorithm, the tree presents the routes reaching other routers within the area.

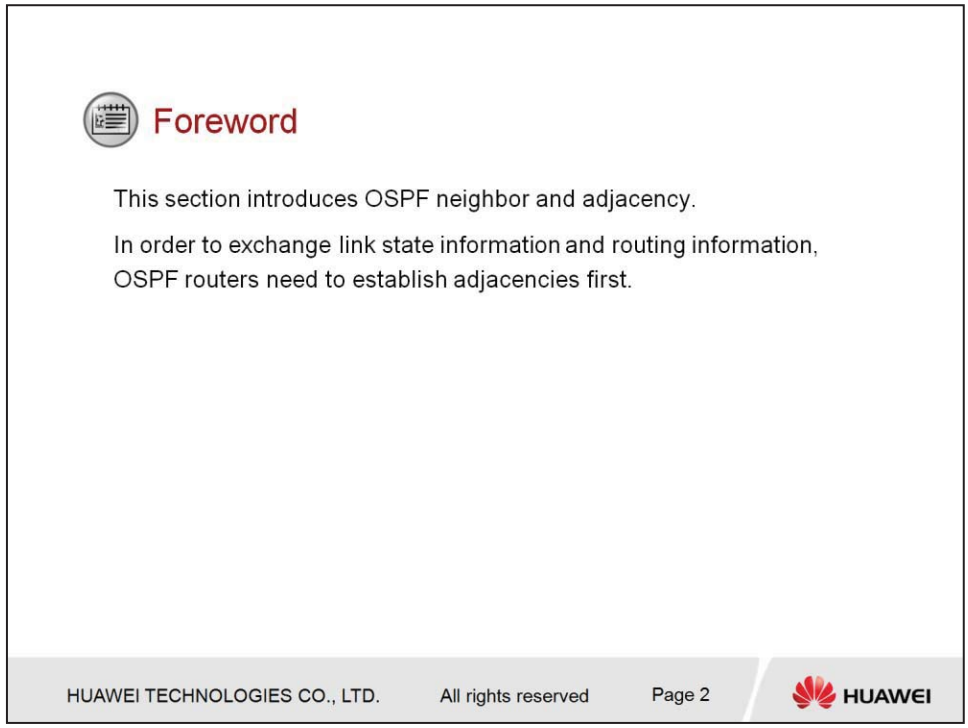
What's OSPF area?

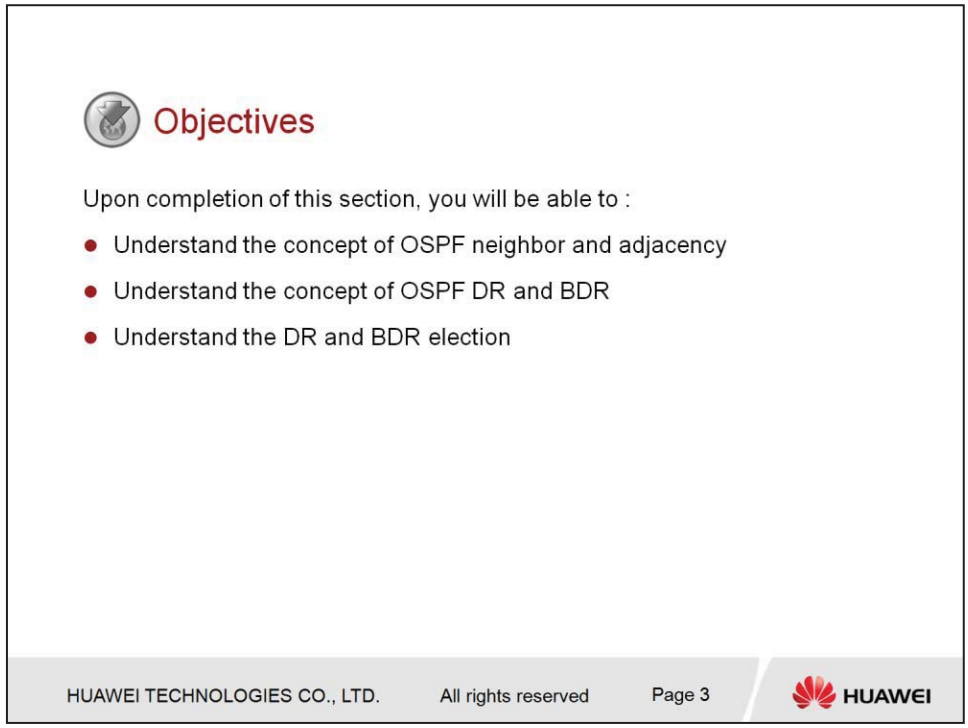
An OSPF area is a set of networks.

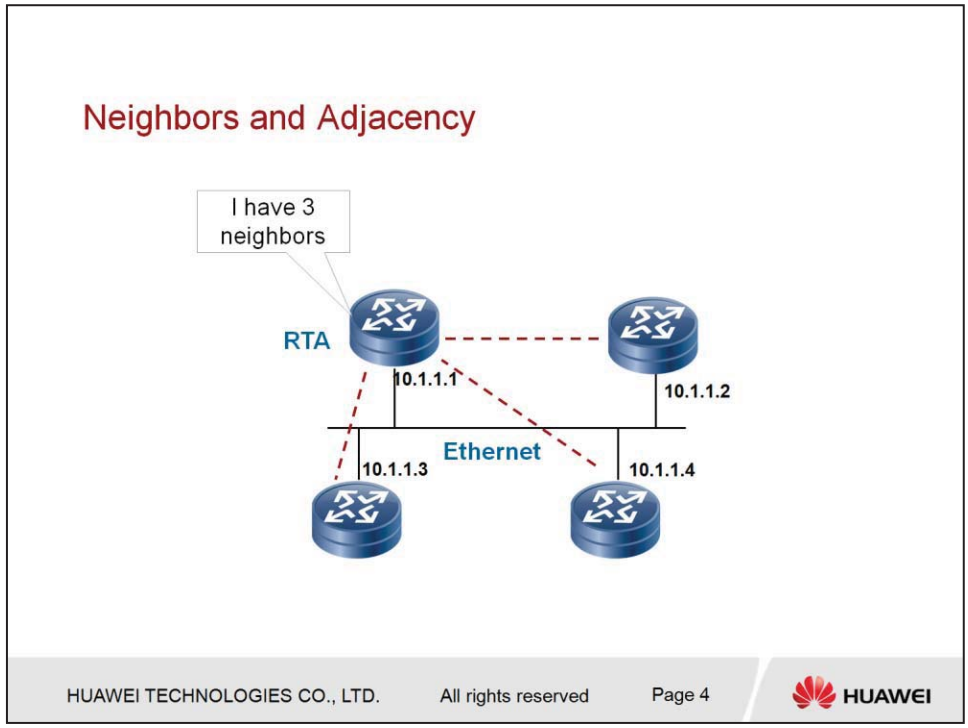
What are steps for OSPF basic configuration?

First to enable OSPF, then create OSPF areas, finally specify networks contained by each area.


Understanding OSPF
Neighbors and Adjacency
www.huawei.com
HUAWEI TECHNOLOGIES CO., LTD.
All rights reserved
HUAWEI


Foreword
This section introduces OSPF neighbor and adjacency.
In order to exchange link state information and routing information, OSPF routers need to establish adjacencies first.
HUAWEI TECHNOLOGIES CO., LTD.
All rights reserved
Page 2
HUAWEI


Objectives
Upon completion of this section, you will be able to :
Understand the concept of OSPF neighbor and adjacency
Understand the concept of OSPF DR and BDR
Understand the DR and BDR election
HUAWEI TECHNOLOGIES CO., LTD.
All rights reserved
Page 3
HUAWEI



OSPF is a type of dynamic routing protocol, routers running OSPF need to exchange link state information and routing information, adjacency relationship must be established before exchanging these routing information.

Neighbor:

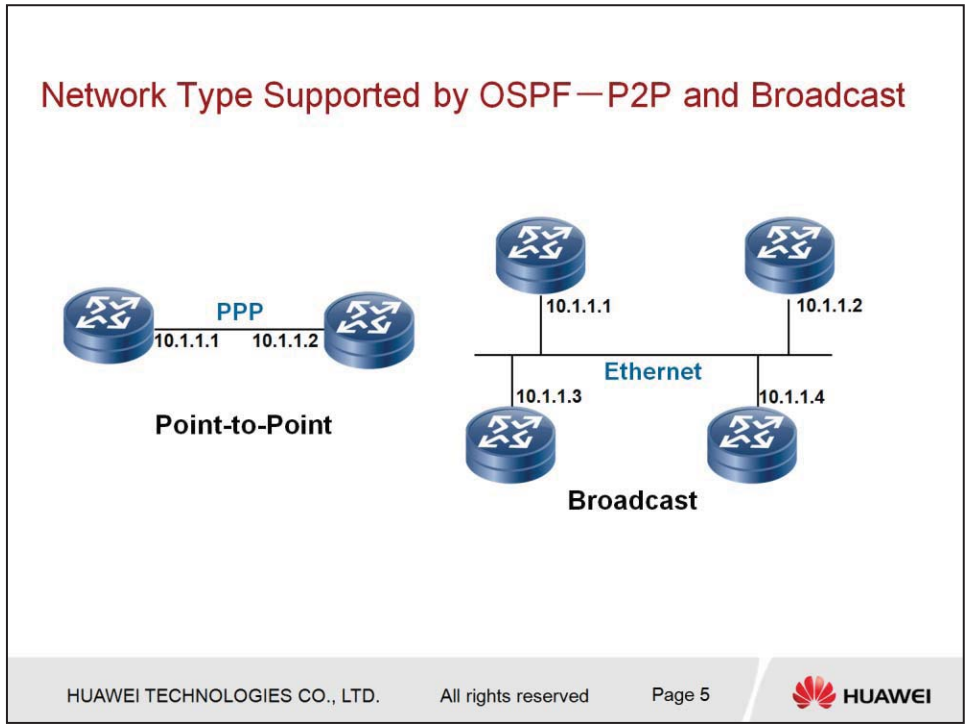
Two routers become neighbors if they have interfaces attached to the same network. The neighbor relationship is maintained by Hello Protocol.

Adjacency:

A relationship formed between selected neighboring routers for the purpose of exchanging routing information.

Not every two neighboring routers will become adjacent. The rule of establishing adjacency relationship varies with network types.

In the figure above, RTA has three neighbors.



As we know, not every two neighboring routers will become adjacent to exchange link state information and routing information, adjacency relationship establishment depends on network type. The network type is the layer two link type of network running OSPF.

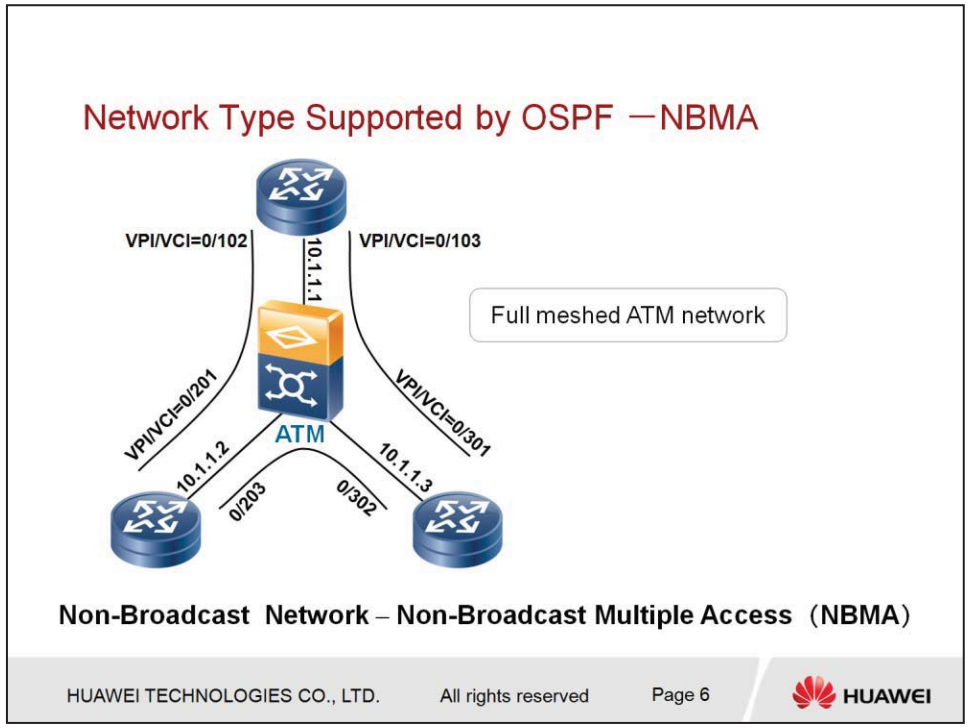
OSPF defines four network types: Point-to-point, Broadcast, NBMA and Point-to-MultiPoint.

Point-to-point network is a network that joins a single pair of routers.

A 64Kb serial line is an example of a point-to-point network.

Broadcast network is a network supporting more than two routers, together with the capability to broadcast.

An Ethernet having 4 routers is an example of broadcast network.



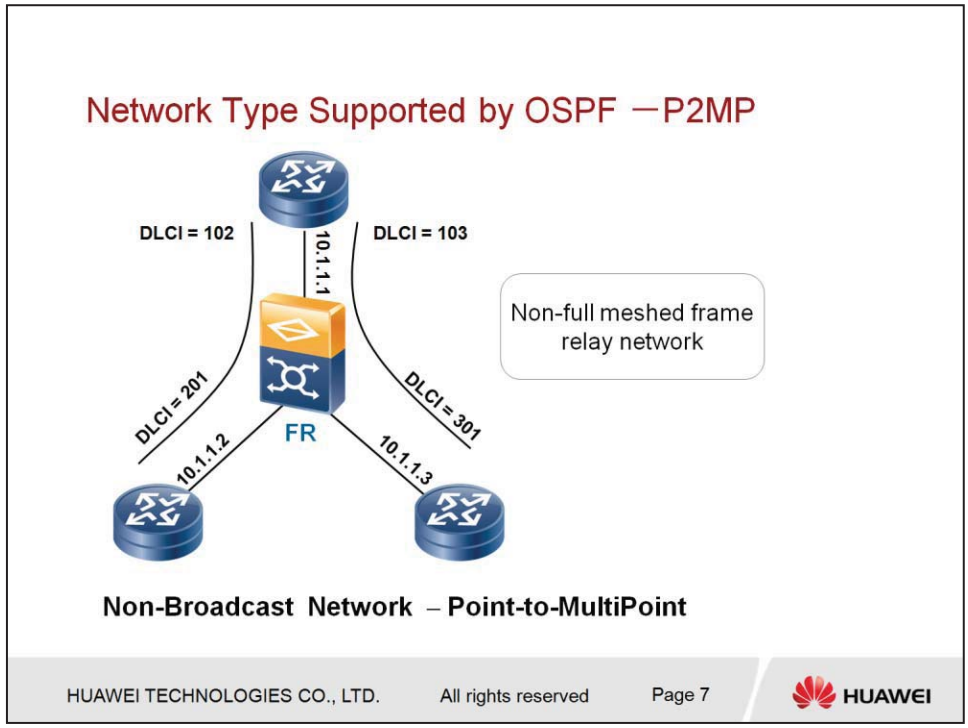
Non-broadcast network is a network supporting more than two routers, but having no broadcast capability.

OSPF runs in one of two modes over non-broadcast networks, one mode is called non-broadcast multi-access or NBMA, the other mode is called Pointto-MultiPoint or P2MP.

Non-Broadcast Multi-Access (NBMA):

NBMA mode simulates the operation of OSPF on a broadcast network, but each neighbor must be manually configured.

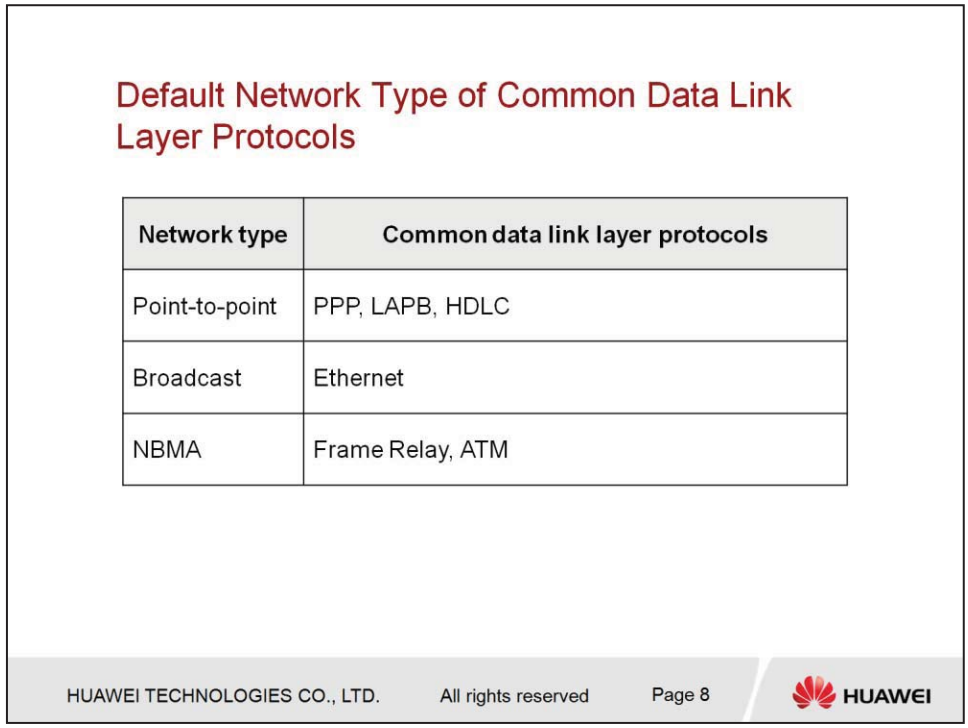
NBMA mode requires that all routers on the network are fully meshed. E.g. the ATM using SVC.



Point-to-MultiPoint:

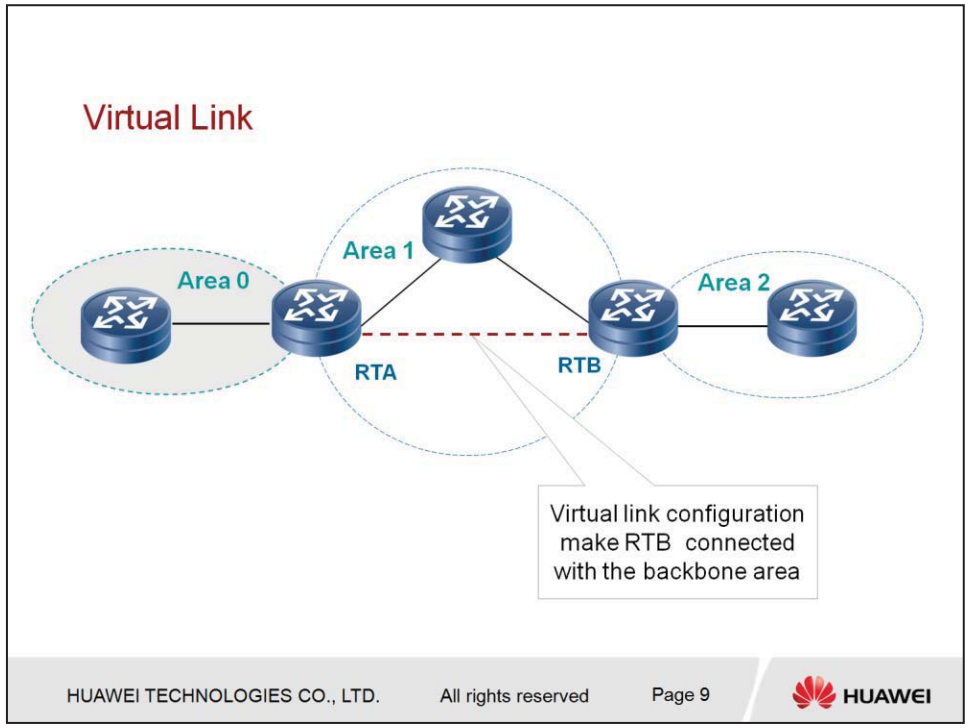
P2MP treats the non-broadcast network as a collection of point-to-point links. The neighbors can be discovered by using some lower-level protocols such as Inverse ARP.

If the network is not fully meshed, P2MP mode should be used. E.g. the frame relay that is not fully meshed.



The table list default network types for common data link layer protocols.

Point-to-MultiPoint is not a default network type.



There is another type of network except four network types just mentioned, that is Virtual Link.

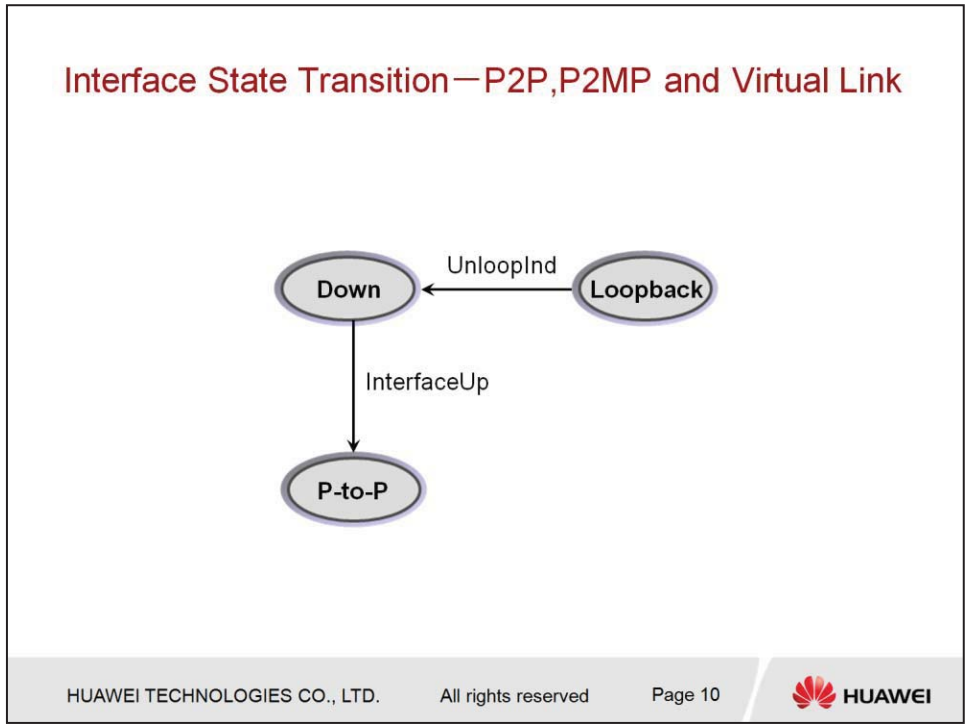
The Backbone Area must be continuous, if not, virtual links can be used to make it logically continuous.

Virtual links can be established between any two ABRs as long as these two ABRs have interfaces attached to the same non-backbone area, the nonbackbone area is called Transit Area.

According to the figure above, RTB is an ABR that is not attached to the Backbone Area, a virtual link can be established between RTA and RTB to connect RTB to the Backbone Area. Area 1 is the Transit Area of the virtual link.

Theoretically speaking, virtual links can allow the Backbone Area to be discontinuous, but it is not recommended in practice.

Virtual link is a logical link belonging to the Backbone Area (Area 0).



Interface states:

Down：

This is the initial interface state. In this state, the lower-level protocols have indicated that the interface is unusable. All interface timers are disabled.

Loopback：

In this state, the router's interface to the network is looped back. Interface in this state is advertised as a Stub network.

Point-to-point（P-to-P）：

In this state, the interface is operational, and connects either to a physical point-to-point network or to a virtual link. Upon entering this state, the routers attempts to form an adjacency with the neighboring router. Hello packets are sent to the neighbor every Hello Interval seconds.

Events causing interface state changes:

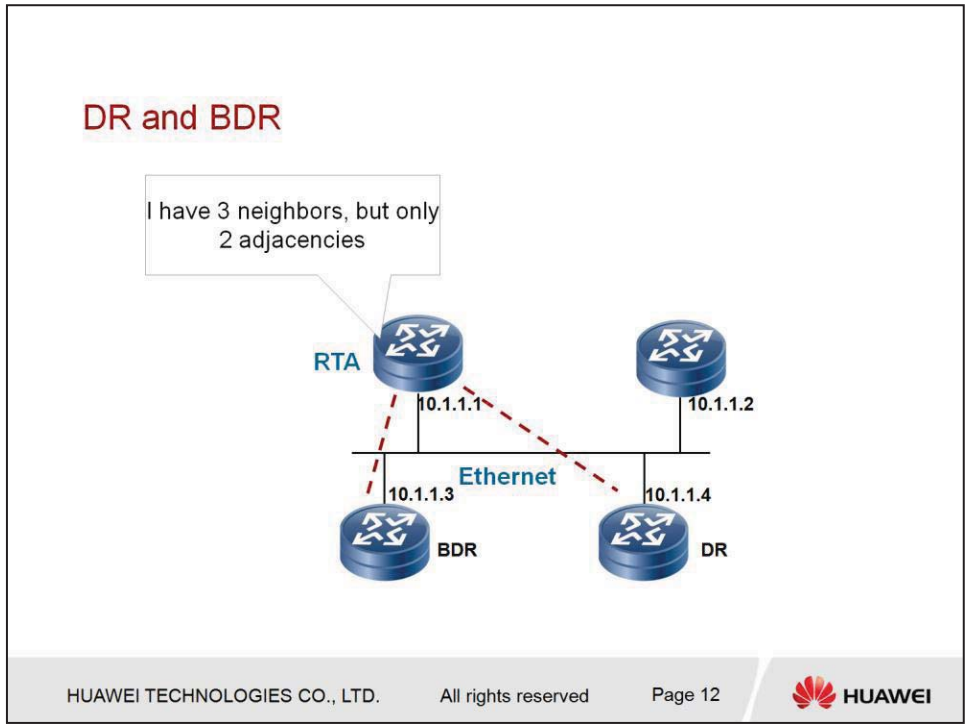
UnloopInd：

An indication has been received that the interface is no longer looped back. If this indication is received by interface in Loopback state, the new interface state will be Down.

InterfaceUp：

Lower-level protocols have indicated that the network interface is operational.

It enables the interface to transition out of Down state to point-to-point on P2P or Virtual links when an Interface Up event occurs. PIP,PIMP and virtual-link have simple interface state transition because these networks do not choose DR and BDR.



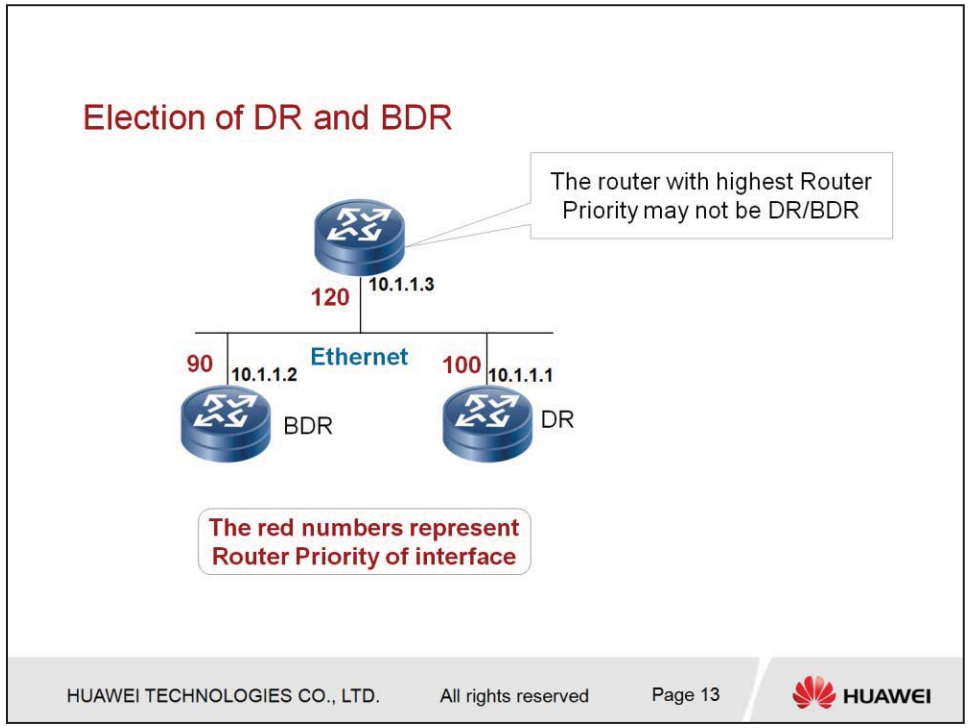
Every broadcast and NBMA networks containing more than two routers has a Designated Router (DR) and Backup Designated Router (BDR).

Functions of DR and BDR:

1. Reduce the number of adjacencies, thus reduce the frequency of exchanging link state information and routing information to save the bandwidth and lower the demand for routers capability. The router that is neither DR nor BDR only establish adjacency relationship with DR and BDR to exchange link state information and routing information, thus the number of adjacencies is greatly reduced on broadcast and NBMA network.

In the figure above, although RTA has 3 neighbors, but it has only 2 adjacencies.

2. the DR originates a network –LSA describing a broadcast or NBMA network.



DR and BDR are elected by the Hello Protocol. The election result is dependent on interface's Router Priority.

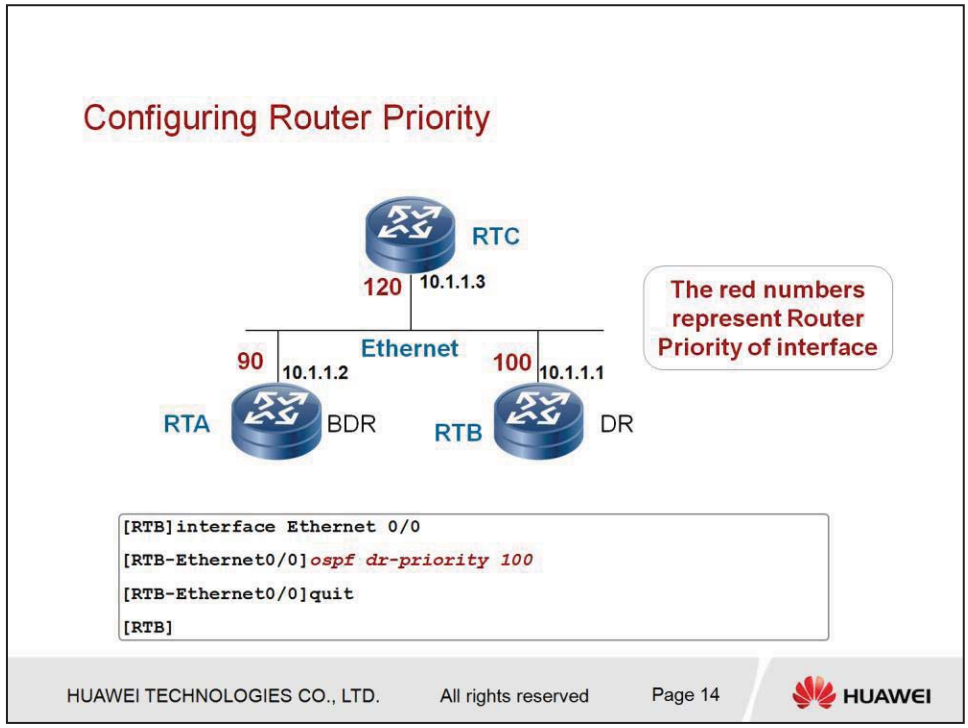
Routers with a priority of 0 are ineligible to become DR or BDR.

The higher the Router Priority, the more preferable the router is

In order to make adjacency relationships stable, if a router's interface to a network first become functional, it checks to see whether there is currently a DR for the network. If there is, it accepts that DR, regardless of its Router Priority.

If the current DR doesn't work, the BDR becomes the new DR and the new BDR is elected; if the current BDR doesn't work, the new BDR is elected while the DR is not changed.

The election mechanism tries to make the transition to a new DR smoother and maintain stable adjacency relationships, thus reduce the impact of the topology change on the adjacency relationship.

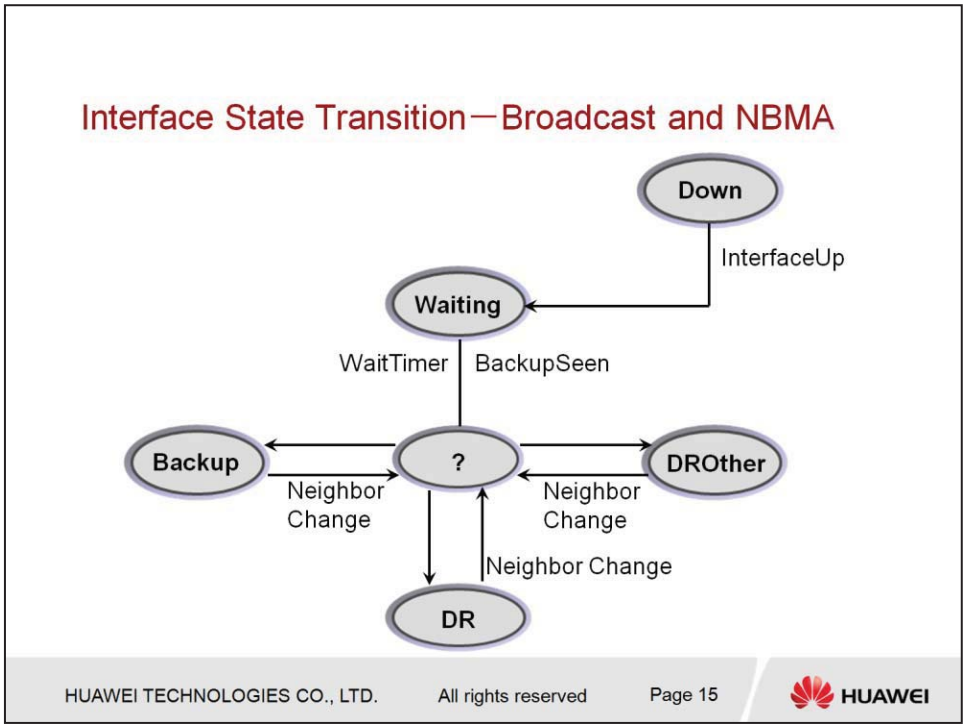


ospf dr-priority value：set interface's Router Priority.

Router Priority ranges from 0 to 255, the default priority on VRP is 1.

If two routers have the same Router Priority, then Router IDs are compared, the larger Router ID indicates higher priority.

If you modify the Router Priority, need to restart the ospf process to re-participate the DR and BDR election.



Interface states and events:

Waiting：

In this state, the router is trying to determine the identity of the DR and BDR for the network. To do this, the router monitors the Hello Packets it receives. The router is not allowed to elect a DR or a BDR until it transitions out of Waiting state. This prevents unnecessary changes of DR and BDR.

Backup：

In this state, the router itself is the BDR on the attached network. The router establishes adjacencies to all other routers attached to the network.

DR：

In this state, this router itself is the DR on the attached network. Adjacencies are established to all other routers attached to the network.

DROther：

The interface is to a broadcast or NBMA network on which another router has been selected to be the DR. In this state, the router itself has not been selected BDR either. The router form adjacencies to both the DR and the BDR.

BackupSeen：

The router has detected the existence or non-existence of a BDR for the network.

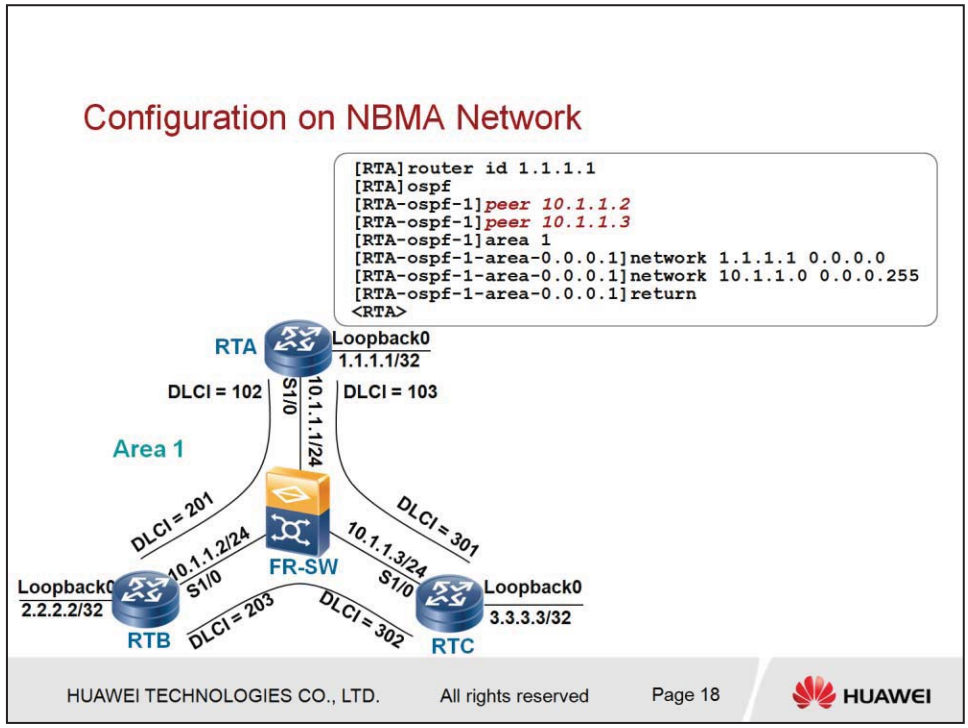
Before electing a DR and BDR for broadcast or NBMA network, the routers will wait for RouterDeadInterval to detect whether DR or BDR has existed for the network. If DR and BDR exists, the selection is disabled and the routers become DROther directly. So the router with the highest Router Priority may not be DR, the router with the second highest Router Priority may not be BDR either.

Whether Neighbors Will Become Adjacent

Network type	Whether neighbors will become adjacent
Point-to-point	Neighbors always become adjacent
Point-to-MultiPoint	Neighbors always become adjacent
Virtual link	Neighbors always become adjacent
Broadcast NBMA	DR is always adjacent to all the other routers including BDR; BDR is always adjacent to all the other routers including DR; Routers whose state is DROther is adjacent to only DR and BDR.

HUAWEI TECHNOLOGIES CO., LTD. All rights reserved HUAWEI

Routers connected by point-to-point networks, Point-to-MultiPoint networks and virtual links always become adjacent. On broadcast and NBMA networks, all routers become adjacent to both the DR and BDR, the number of adjacencies are smaller than that of neighbors.



Topology description:

Three routers are connected each other by a frame relay switch. All networks belong to Area 1.

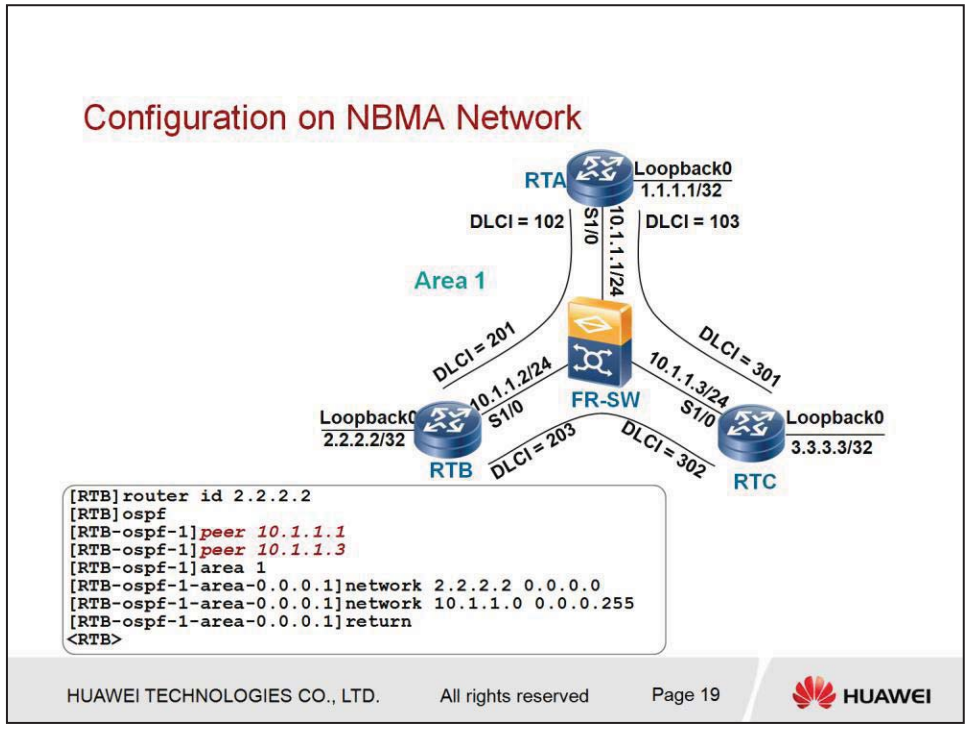
Inverse ARP is not needed on NBMA network, but neighbors must be manually configured.

RTA has two neighbors 10.1.1.2 and 10.1.1.3.

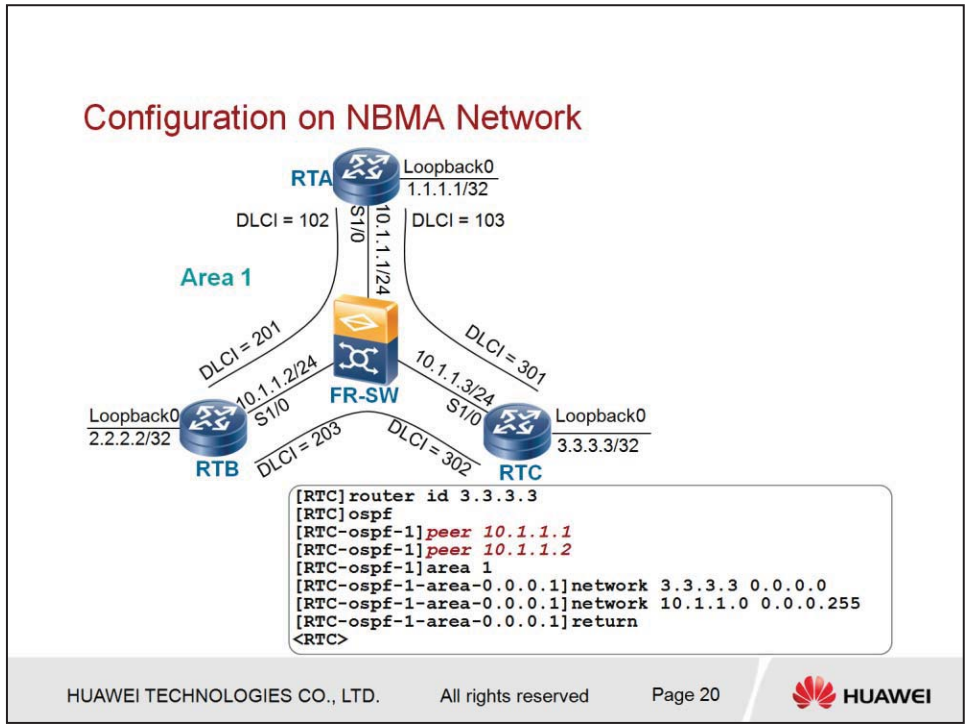
The IP address of the neighbor's interface to the network is used to identify this neighbor.

peer ip-address [dr-priority dr-priority-number]

dr-priority-number：neighbor's Router Priority, the default priority is 1.



Configure two neighbors on RTB, 10.1.1.1 and 10.1.1.3.



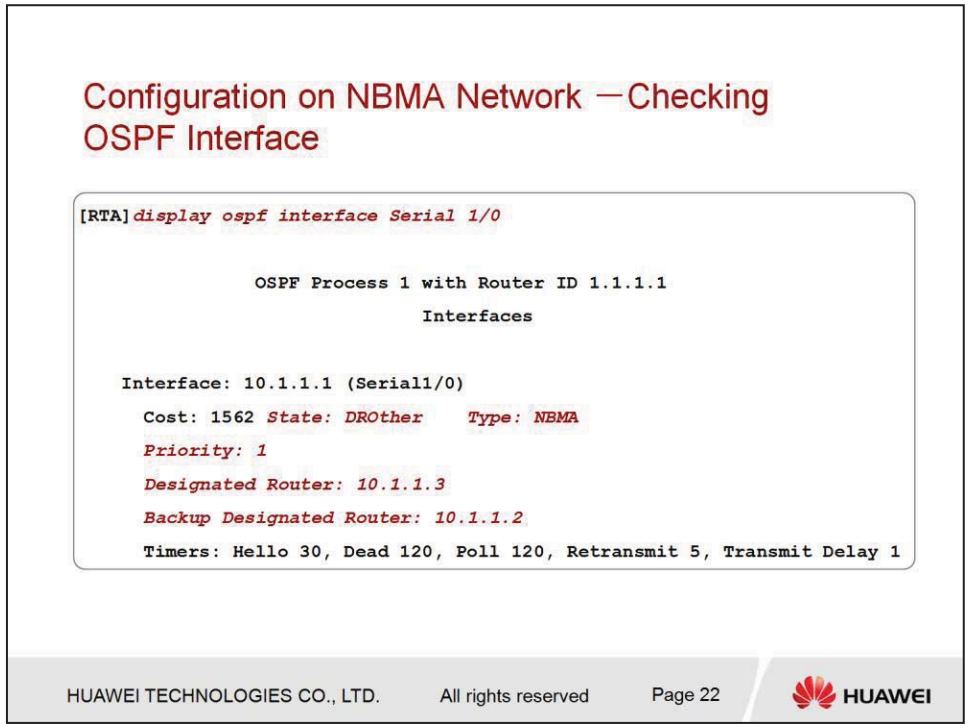
Configure two neighbors on RTC, 10.1.1.1 and 10.1.1.2.

Configuration on NBMA Network —Checking IP Routing Table

```
[RTA]display ip routing-table
 Routing Table: public net
Destination/Mask    Protocol  Pre  Cost      Nexthop         Interface
1.1.1.1/32          DIRECT    0    0         127.0.0.1       InLoopBack0
2.2.2.2/32          OSPF      10   1563      10.1.1.2        Serial1/0
3.3.3.3/32          OSPF      10   1563      10.1.1.3        Serial1/0
10.1.1.0/24         DIRECT    0    0         10.1.1.1        Serial1/0
10.1.1.1/32         DIRECT    0    0         127.0.0.1       InLoopBack0
10.1.1.2/32         DIRECT    0    0         10.1.1.2        Serial1/0
10.1.1.3/32         DIRECT    0    0         10.1.1.3        Serial1/0
127.0.0.0/8         DIRECT    0    0         127.0.0.1       InLoopBack0
127.0.0.1/32        DIRECT    0    0         127.0.0.1       InLoopBack0
```

HUAWEI TECHNOLOGIES CO., LTD. All rights reserved HUAWEI

The routes destined for other routers' Loopback interface are learned from OSPF.

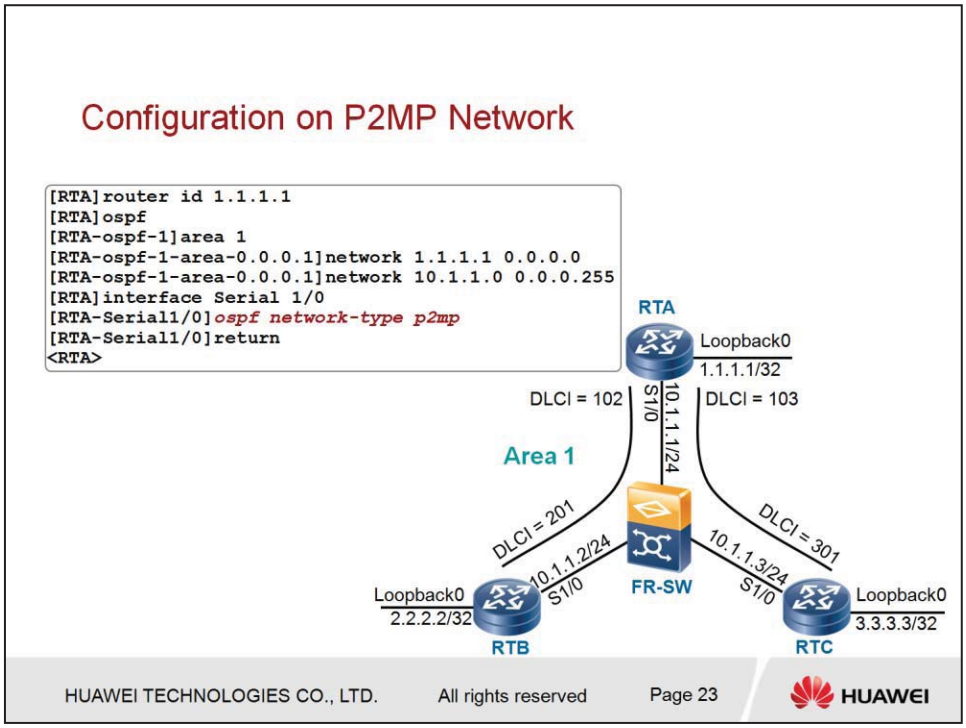


State: interface state

Type: network type of interface

Priority：interface's Router Priority, used for election DR and BDR

Designated Router: DR interface IP address Backup Designated Router: BDR interface IP address



In this example:

RTA connects to other two routers, but RTB and RTC are not connected.

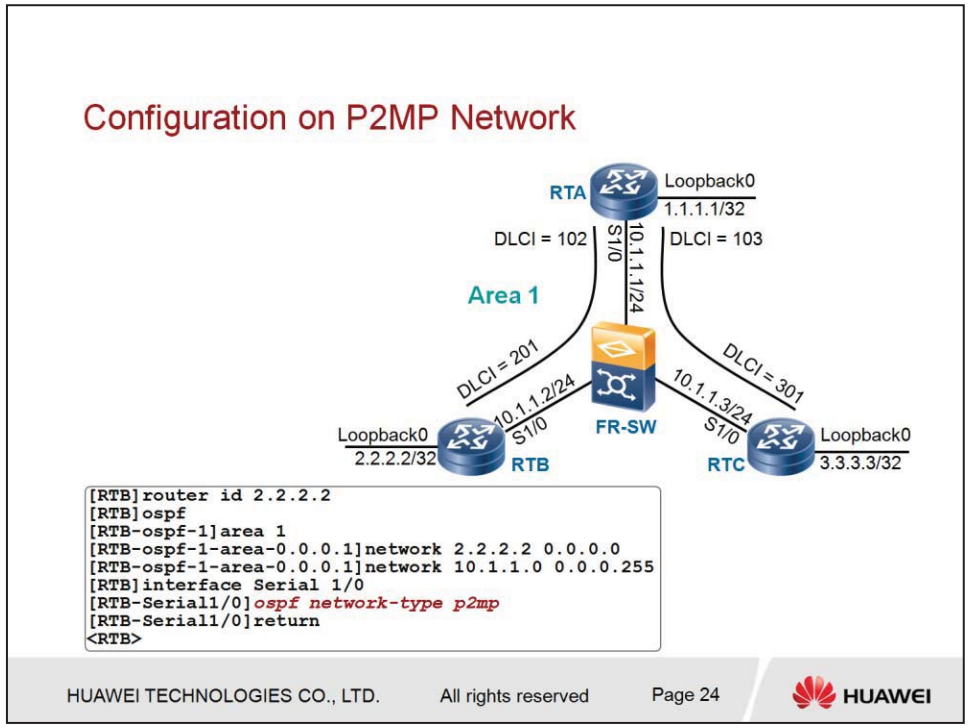
OSPF network type on frame relay network must be manually configured as Point-to-MultiPoint, because the network is not fully meshed.

Inverse ARP should be enabled on interface on frame relay network.

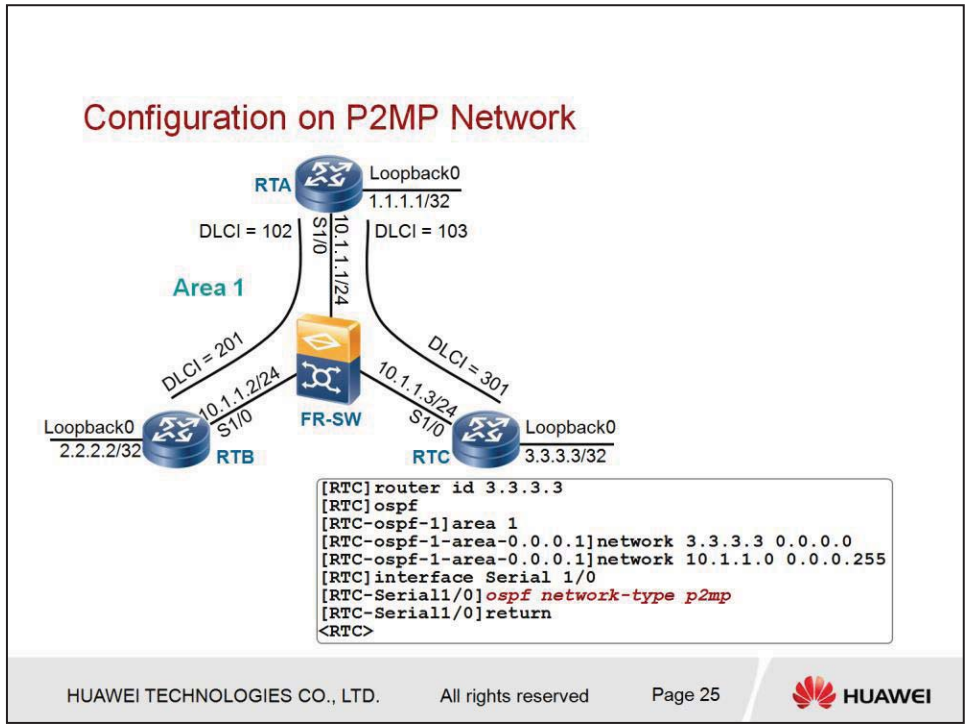
All RTA's attached networks are configured to be within the Area 1.

Manually configure the network type of Serial 1/0 as Point-to-MultiPoint ospf network-type { broadcast | nbma | p2mp | p2p }

4 network types are available.



All RTB's attached networks are configured to be within Area 1, the network type of Serial 1/0 is configured as Point-to-MultiPoint.



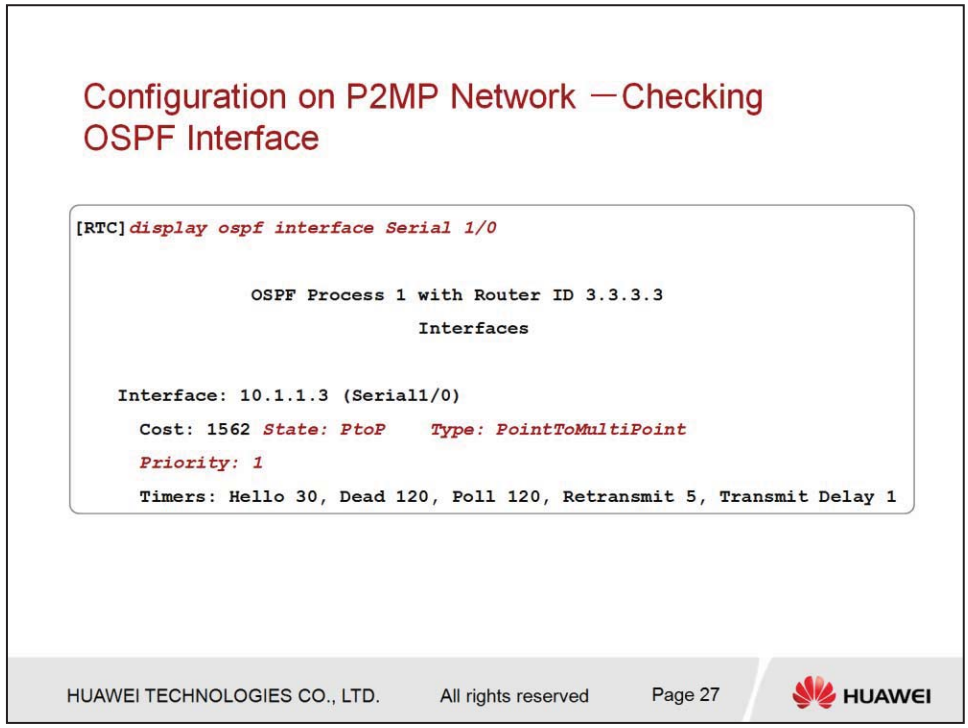
All RTC's attached networks are configured to be within Area 1, the network type of Serial 1/0 is configured as Point-to-MultiPoint.

Configuration on P2MP Network —Checking IP Routing Table

```
[RTC]display ip routing-table
 Routing Table: public net
Destination/Mask    Protocol  Pre  Cost      Nexthop         Interface
1.1.1.1/32          OSPF      10   1563      10.1.1.1        Serial1/0
2.2.2.2/32          OSPF      10   3125      10.1.1.1        Serial1/0
3.3.3.3/32          DIRECT    0    0         127.0.0.1       InLoopBack0
10.1.1.0/24         DIRECT    0    0         10.1.1.3        Serial1/0
10.1.1.1/32         DIRECT    0    0         10.1.1.1        Serial1/0
10.1.1.2/32         OSPF      10   3124      10.1.1.1        Serial1/0
10.1.1.3/32         DIRECT    0    0         127.0.0.1       InLoopBack0
127.0.0.0/8         DIRECT    0    0         127.0.0.1       InLoopBack0
127.0.0.1/32        DIRECT    0    0         127.0.0.1       InLoopBack0
```

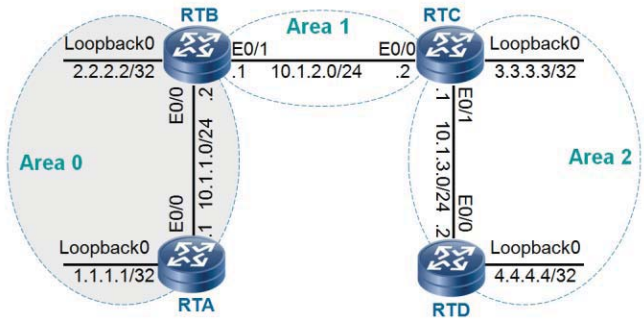
HUAWEI TECHNOLOGIES CO., LTD. All rights reserved HUAWEI

In RTC's routing table, the two routes destined for Loopback 0 are learned from OSPF, the route destined for RTB's physical interface is also learned from OSPF.



The network type is Point-to-MultiPoint, the stable interface state is point-to-point for Point-to-MultiPoint network type.

Configuring Virtual Link




```
[RTA]router id 1.1.1.1
[RTA]ospf
[RTA-ospf-1]area 0
[RTA-ospf-1-area-0.0.0.0]network 1.1.1.1 0.0.0.0
[RTA-ospf-1-area-0.0.0.0]network 10.1.1.0 0.0.0.255
[RTA-ospf-1-area-0.0.0.0]return
<RTA>
```

HUAWEI TECHNOLOGIES CO., LTD. All rights reserved HUAWEI

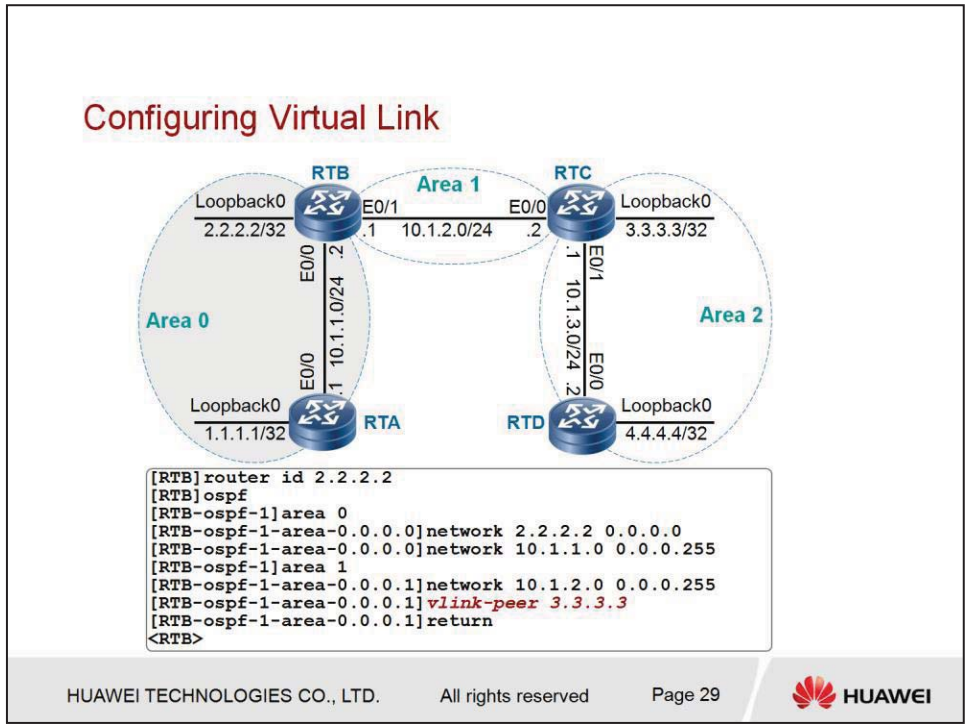
In this example:

RTC is ABR, but RTC is not attached to the Backbone Area,

So configure virtual link between RTB and RTC.

Area 1 is the Transit Area of virtual link.

All RTA's attached networks are configured to be within Area 0.

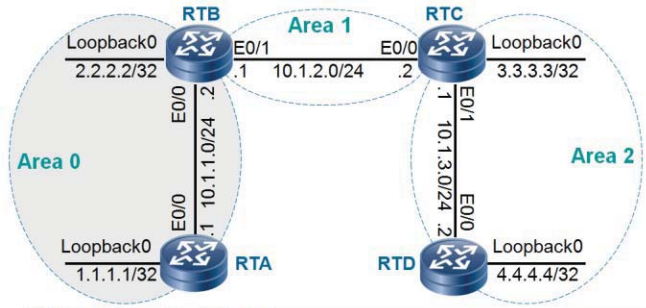


When configuring virtual link, the Router ID of the other endpoint must be configured.

Virtual link is configured in the area view of transit area.

vlink-peer router-id： the Router ID of the other endpoint is used to identify the peer router.

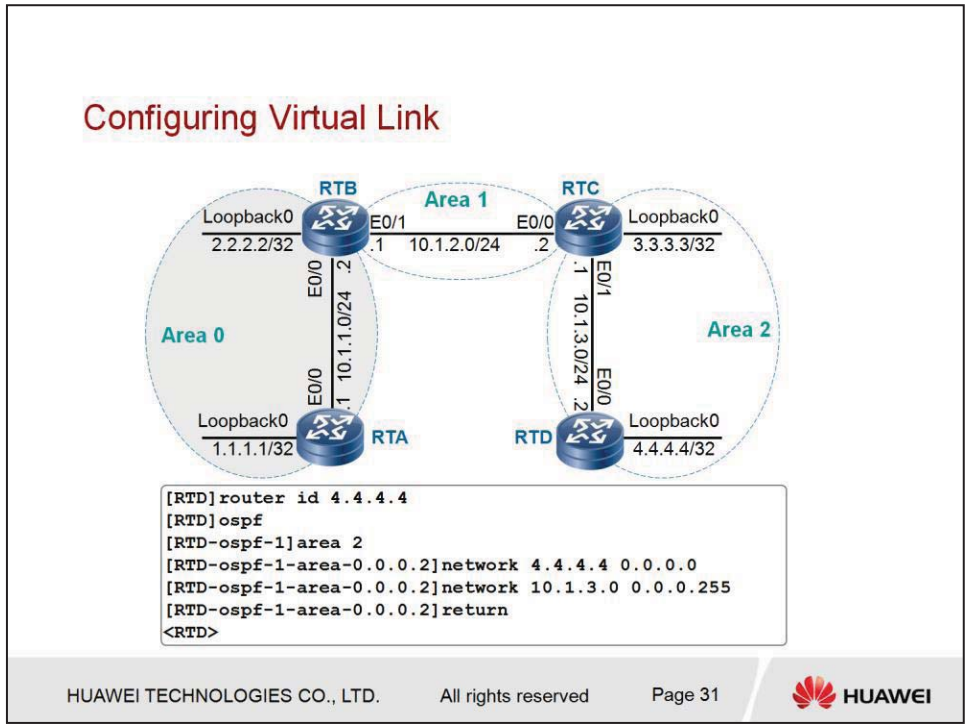
Configuring Virtual Link




```
[RTC]router id 3.3.3.3
[RTC]ospf
[RTC-ospf-1]area 1
[RTC-ospf-1-area-0.0.0.1]network 10.1.2.0 0.0.0.255
[RTC-ospf-1-area-0.0.0.1]vlink-peer 2.2.2.2
[RTC-ospf-1]area 2
[RTC-ospf-1-area-0.0.0.2]network 3.3.3.3 0.0.0.0
[RTC-ospf-1-area-0.0.0.2]network 10.1.3.0 0.0.0.255
[RTC-ospf-1-area-0.0.0.2]return
<RTC>
```

HUAWEI TECHNOLOGIES CO., LTD. All rights reserved HUAWEI

Virtual link must be configured on both ABRs.



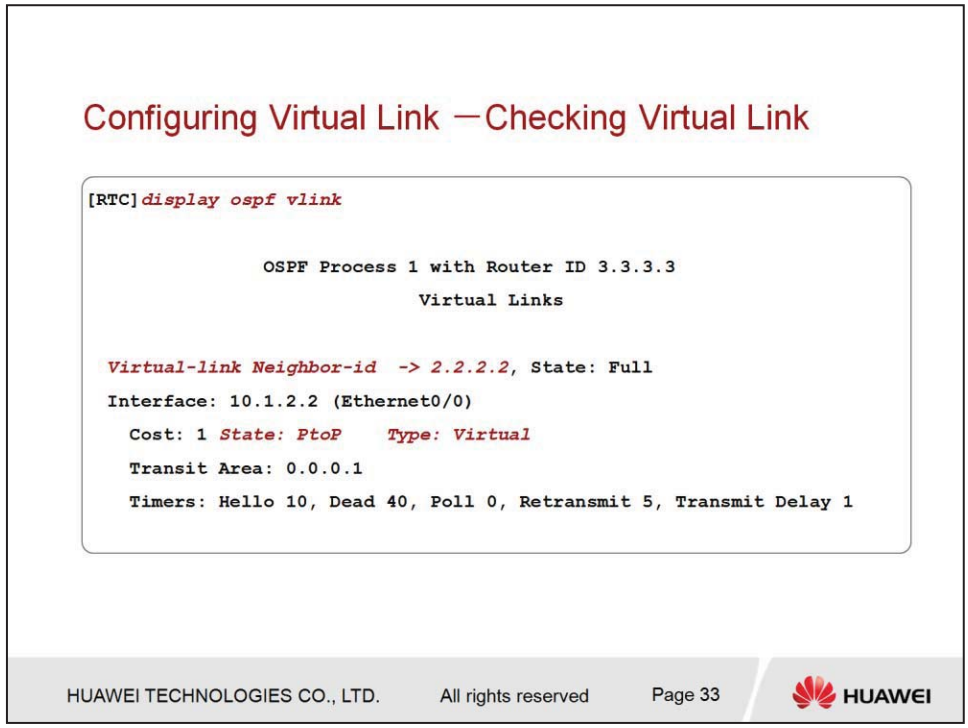
All RTD's attached networks are configured to be within Area 2.

Configuring Virtual Link —Checking IP Routing Table

```
[RTD]display ip routing-table
 Routing Table: public net
Destination/Mask    Protocol  Pre  Cost      Nexthop         Interface
1.1.1.1/32          OSPF      10   4         10.1.3.1        Ethernet0/0
2.2.2.2/32          OSPF      10   3         10.1.3.1        Ethernet0/0
3.3.3.3/32          OSPF      10   2         10.1.3.1        Ethernet0/0
4.4.4.4/32          DIRECT    0    0         127.0.0.1       InLoopBack0
10.1.1.0/24         OSPF      10   3         10.1.3.1        Ethernet0/0
10.1.2.0/24         OSPF      10   2         10.1.3.1        Ethernet0/0
10.1.3.0/24         DIRECT    0    0         10.1.3.2        Ethernet0/0
10.1.3.2/32         DIRECT    0    0         127.0.0.1       InLoopBack0
127.0.0.0/8         DIRECT    0    0         127.0.0.1       InLoopBack0
127.0.0.1/32        DIRECT    0    0         127.0.0.1       InLoopBack0
```

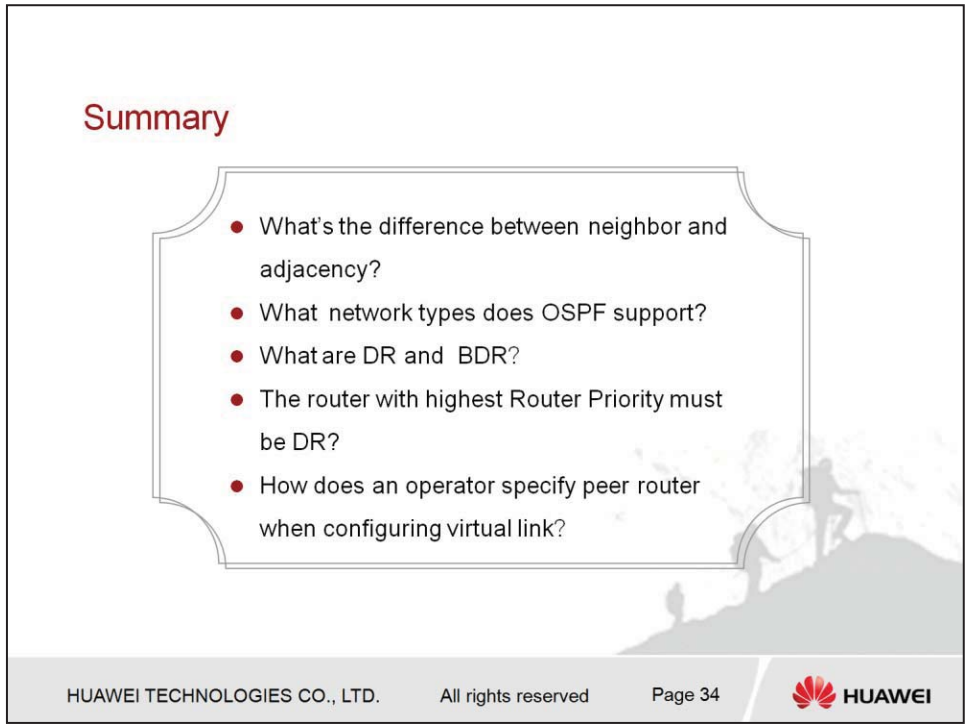
HUAWEI TECHNOLOGIES CO., LTD. All rights reserved HUAWEI

RTD learned 5 routes by OSPF.



Neighbor-id:

Peer Router ID of the virtual link.



What's the difference between neighbor and adjacency?

Any two routers that are attached to the same network can become neighbors, while adjacency indicates the neighbors which can exchange link state information, only part of neighbors can become adjacent.

What types of network does OSPF support?

Point-to-Point, broadcast, NBMA and Point-to-MultiPoint

What are DR and BDR?

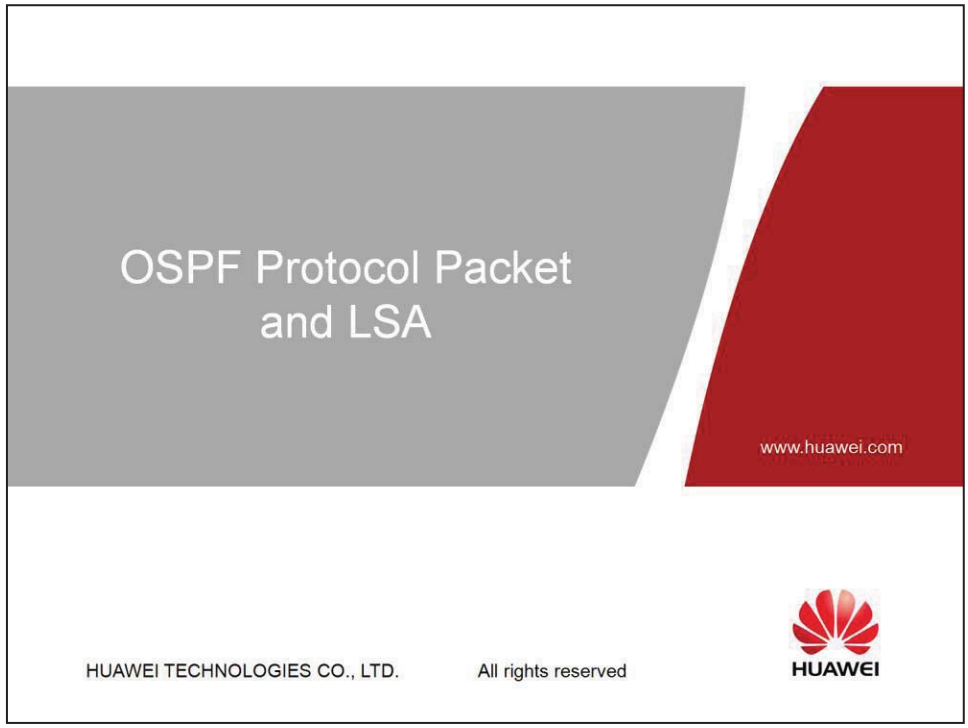
DR is the designated router on broadcast or NBMA network, it is used to exchange routing information with other routers on the network.

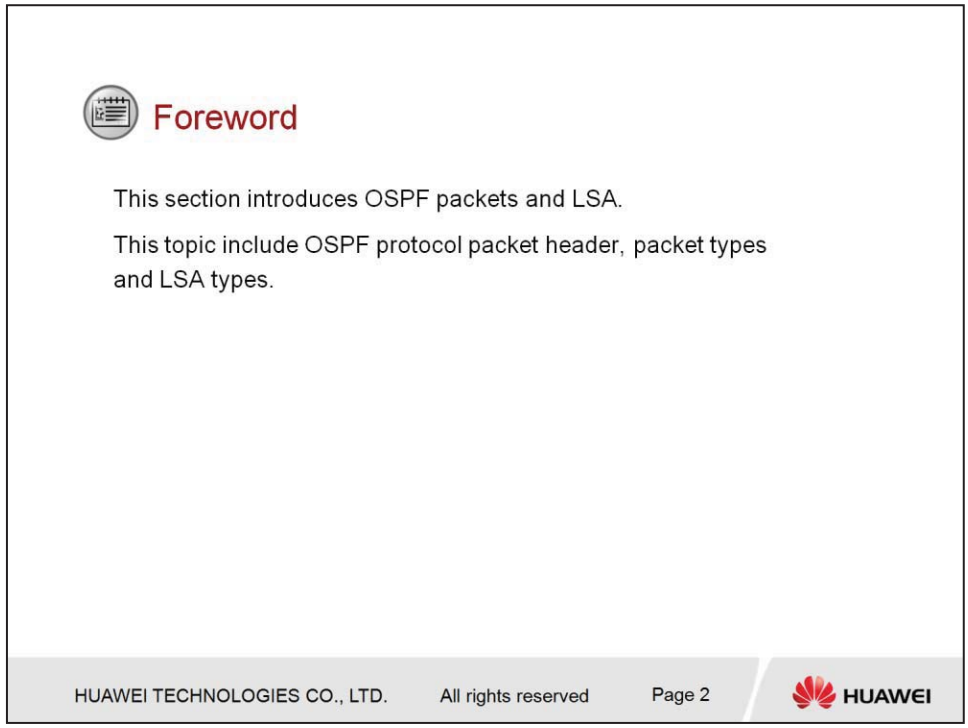
The router with biggest router ID should be DR absolutely, is that right?

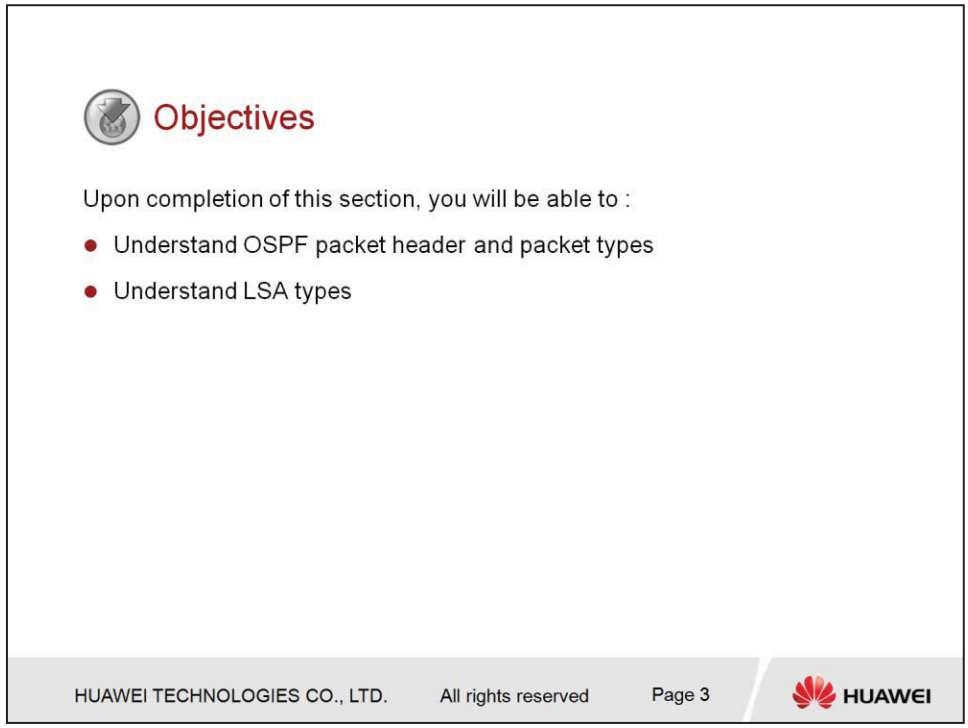
The router with highest Router ID may not be the DR. In order to maintain network stability, topology change does not lead to reelection of DR and BDR.

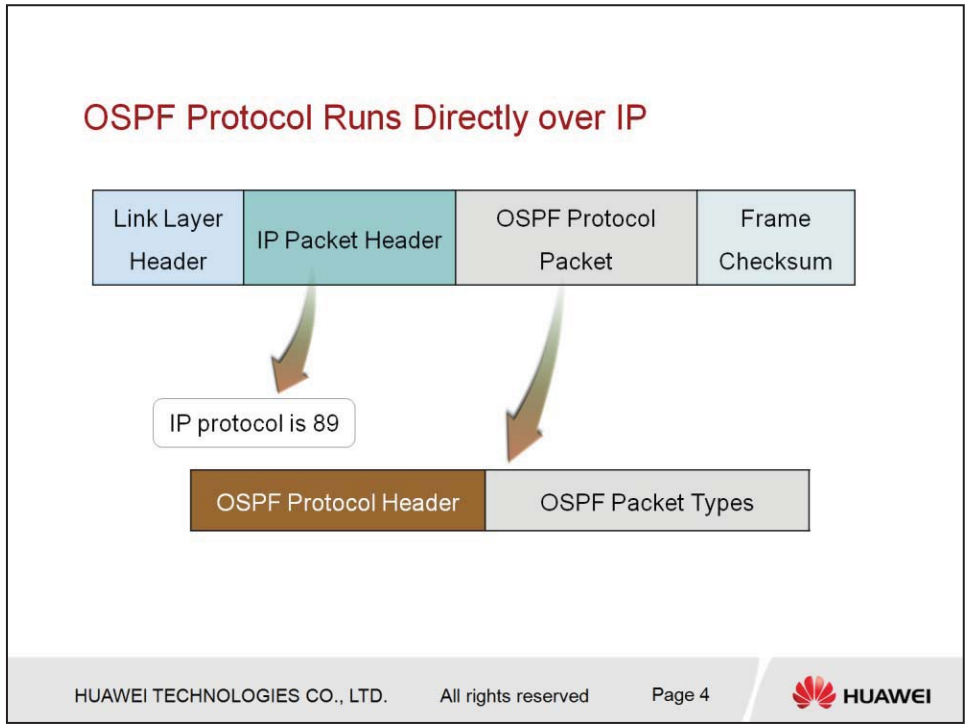
How does an operator specify peer router when configuring virtual link?

Use the peer's Router ID.


OSPF Protocol Packet and LSA
www.huawei.com
HUAWEI TECHNOLOGIES CO., LTD.
All rights reserved
HUAWEI


Foreword
This section introduces OSPF packets and LSA.
This topic include OSPF protocol packet header, packet types and LSA types.
HUAWEI TECHNOLOGIES CO., LTD.
All rights reserved
Page 2
HUAWEI

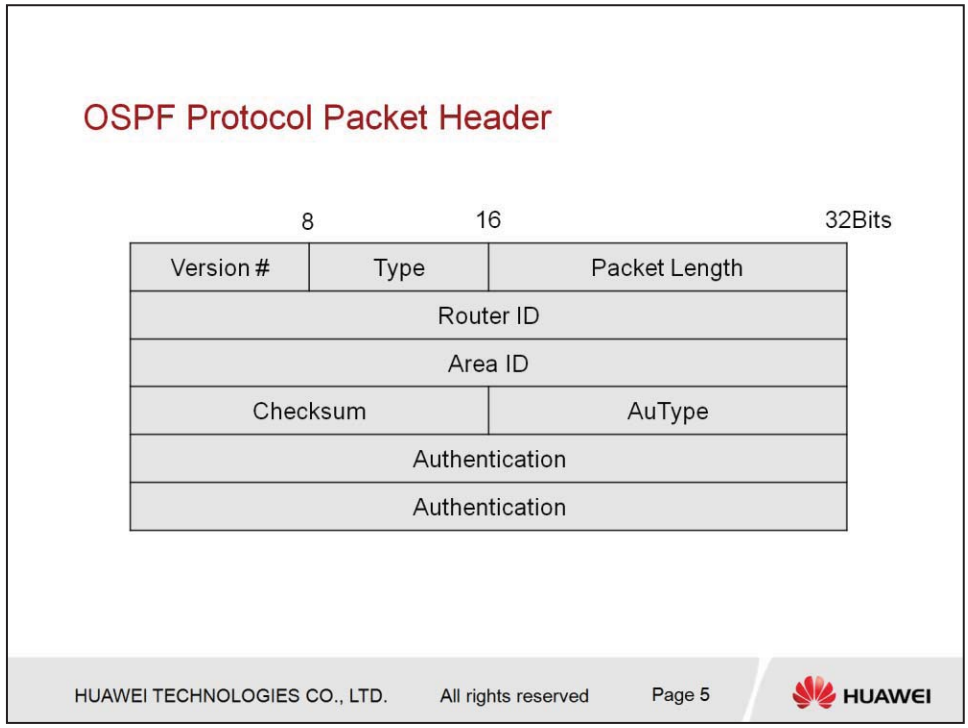

Objectives
Upon completion of this section, you will be able to :
Understand OSPF packet header and packet types
Understand LSA types
HUAWEI TECHNOLOGIES CO., LTD.
All rights reserved
Page 3
HUAWEI



The OSPF protocol runs directly over IP, using IP protocol 89.

OSPF has 5 types of protocol packets, but all these packets share a common

protocol header.



All OSPF protocol packets share a common protocol header.

Version #：

OSPF version number, which should be 2.

Type：

OSPF packet type, there are 5 type of packets.

Packet length：

The length of the OSPF protocol packet in bytes. This length includes the standard OSPF header.

Router ID：

The Router ID of the packet's source.

Area ID：

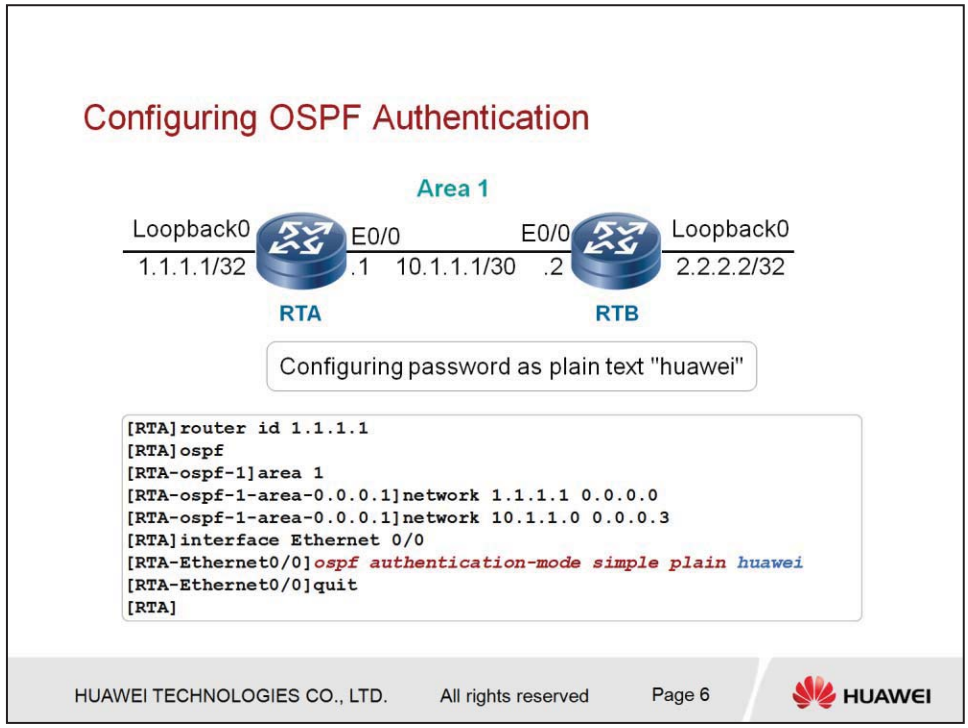
A 32-bit number identifying the area that this packet belongs to.

AuType：

Identifies the authentication procedure to be used for the packet.

Authentication：

A 64-bit field for use by the authentication scheme.



OSPF protocol packet authentication:

On VRP system, OSPF supports area authentication and interface authentication.

When using area authentication, all routers' authentication mode and password must match within the same area; interface authentication is used to authenticate between neighboring routers, which overrides area authentication.

In this example, there is only one area containing two routers. Interface authentication is configured, authentication mode is "simple", encryption type is "plain", the password between RTA and RTB is "huawei".

On RTA, in interface view, configure authentication mode as "simple", encryption type as "plain", password as "huawei".

If area authentication is used, the following command is used:

Authentication mode is "simple"

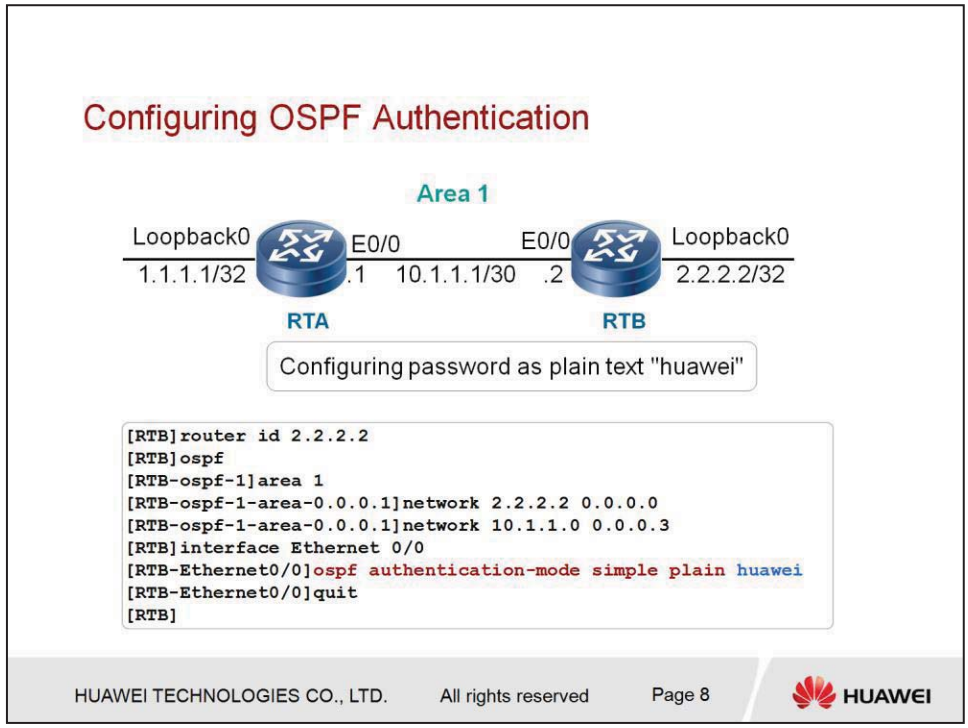
authentication-mode simple { [plain] plain-text | cipher cipher-text }

plain：encryption type is "plain".

cipher：encryption type is "cipher".

Authentication mode is “MD5”

authentication-mode md5 key-id { [plain] plain-text | cipher cipher-text }



On RTB, in interface view, configure authentication mode as “simple”, encryption type as “plain”, password as “huawei”.

OSPF Protocol Packet Type

Type	Packet name	Packet function
1	Hello	Discover /maintain neighbor relationships
2	Database Description	Summarize database contents
3	Link State Request	Database download
4	Link State Update	Database update
5	Link State Ack	Flooding acknowledgment

HUAWEI TECHNOLOGIES CO., LTD. All rights reserved HUAWEI

OSPF has 5 type of protocol packets.

Hello packet is used to discover and maintain neighbor relationship. On broadcast and NBMA network, Hello packet can also be used to elect DR and BDR.

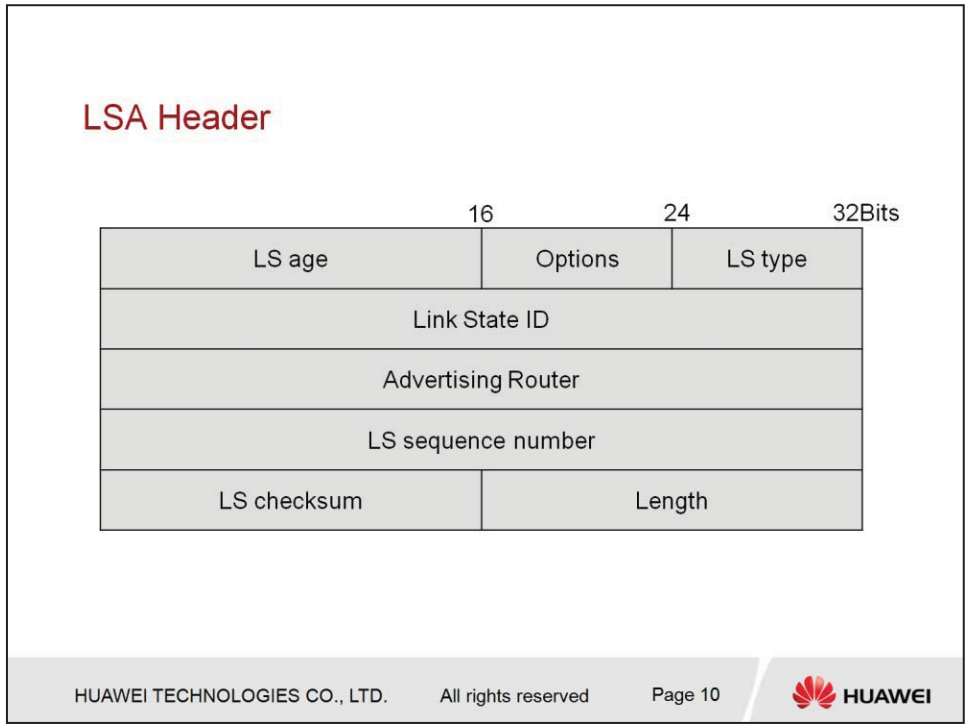
DD packet is used to summarize database contents.

LS Request packet is used to request the pieces of the neighbor's database that are more up-to-date.

LS Update is used to synchronize database by flooding of LSAs.

LS Ack is used to make the flooding of LSAs reliable by acknowledging flooded LSAs.

All packets are sent only between adjacent routers except Hellos.



All OSPF protocol packets carry LSA information except Hellos.

LS age：

The time in seconds since the LSA was originated.

LS type：

The type of the LSA. There are 5 type of common LSAs.

Link State ID：

This field identifies the portion of the internet environment that is being described

by the LSA. E.g. Router ID

Advertising Router：

The Router ID of the router that originated the LSA.

LS sequence number：

Detects old or duplicate LSAs.

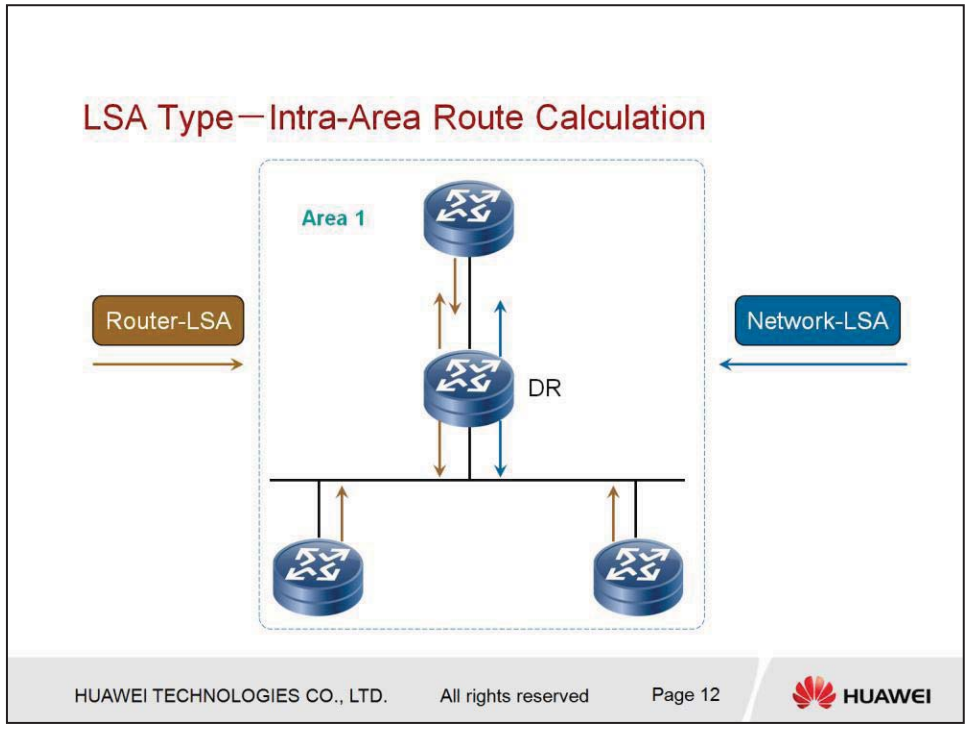
LS type，Link State ID and Advertising Router together uniquely identify the LSA.

LSA Type—Intra-Area Route Calculation

LS Type	LSA Name	LSA Description
1	Router-LSA	Generated by each router. Describing the collected states of the router's interfaces, flooded only within the area where this LSA is generated.
2	Network-LSA	Generated by DR, available in broadcast and NBMA network. Describing the set of routers attached to the network. flooded only within the area where this LSA is generated.

HUAWEI TECHNOLOGIES CO., LTD. All rights reserved HUAWEI

Router-LSA and Network-LSA are used to calculate intra-area routes, these two types of LSA describe detailed link state information.



Each router advertises Router-LSA to other routers in the same area.

Only DR advertises Network-LSA within the area.

LSA Type—Inter-Area Route Calculation

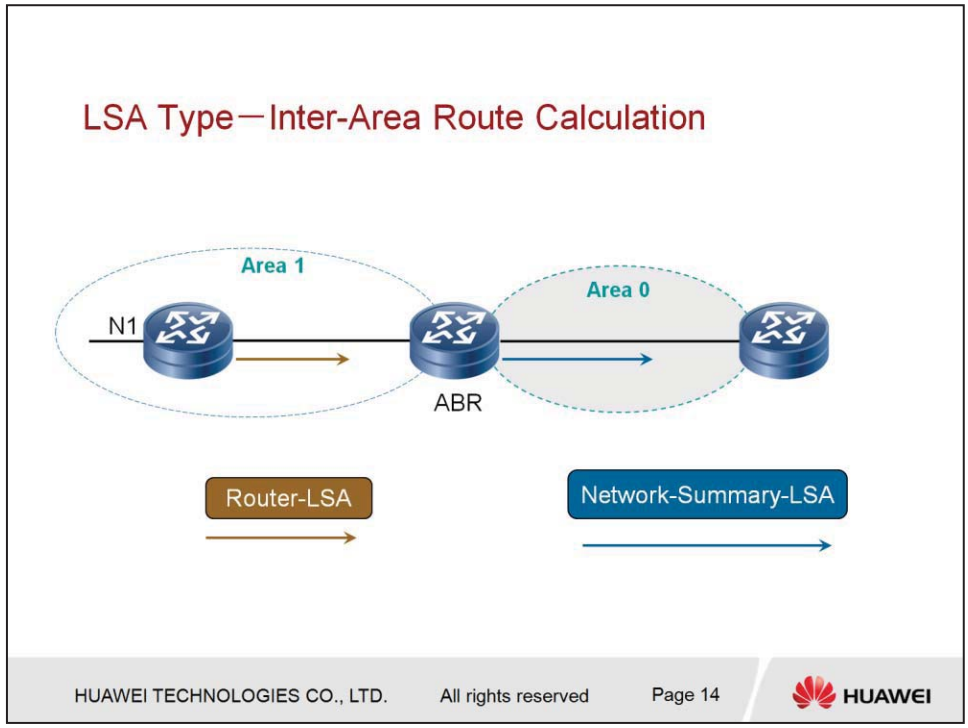
LS Type	LSA Name	LSA Description
3	Network-Summary-LSA	Generated by ABR. Describing routes to networks that are outside the area but inside the AS, flooded within the area where this LSA is generated

HUAWEI TECHNOLOGIES CO., LTD. All rights reserved

Network-Summary-LSA is used to calculate inter-area routes.

This type of LSA describes brief routing information instead of detailed link state information.

The default route can also be advertised by Network-Summary-LSA.



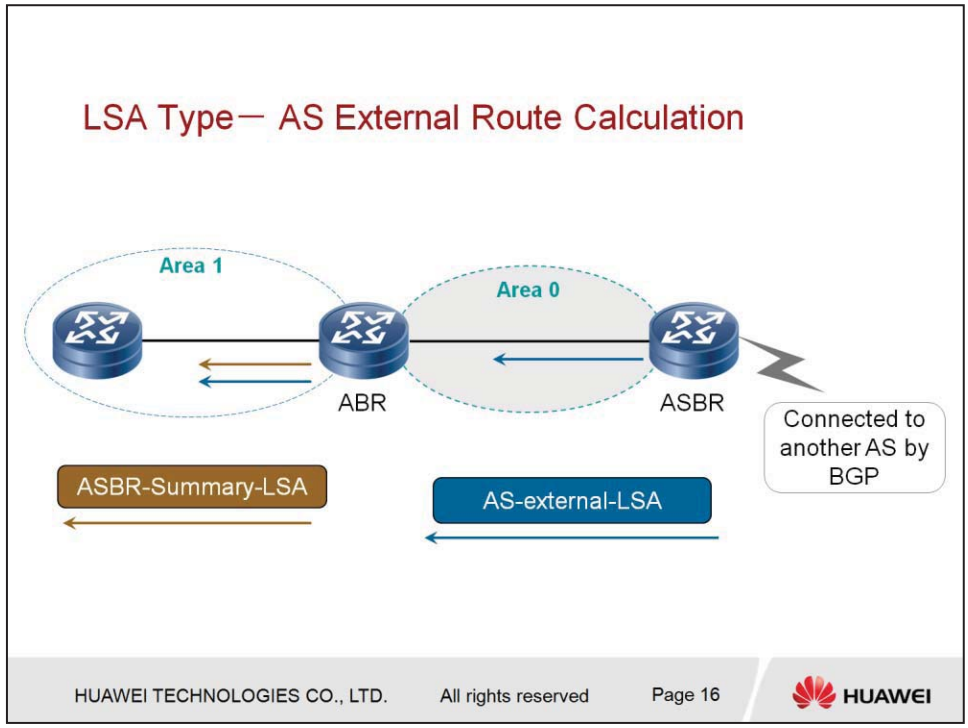
The route destined for network N1 is advertised to ABR by Router-LSA, ABR abstracts routing information from link state information, then advertises it to other areas by Network-Summary-LSA.

LSA Type－ AS External Route Calculation

LS Type	LSA Name	LSA Description
4	ASBR-Summary-LSA	Generated by ABR. Describing routes to ASBRs, flooded within the area where this LSA is generated.
5	AS-external-LSA	Generated by ASBR. Describing routes to destinations external to the As, flooded within the whole AS.

HUAWEI TECHNOLOGIES CO., LTD. All rights reserved HUAWEI

Type 4 LSA describes how to reach ASBR, type 5 LSA describes how to reach a certain destination outside the AS. These two types of LSA can be used together to calculate the AS external routes.



ASBR uses type 5 LSAs to describe AS external route, these LSAs are flooded within the whole AS.

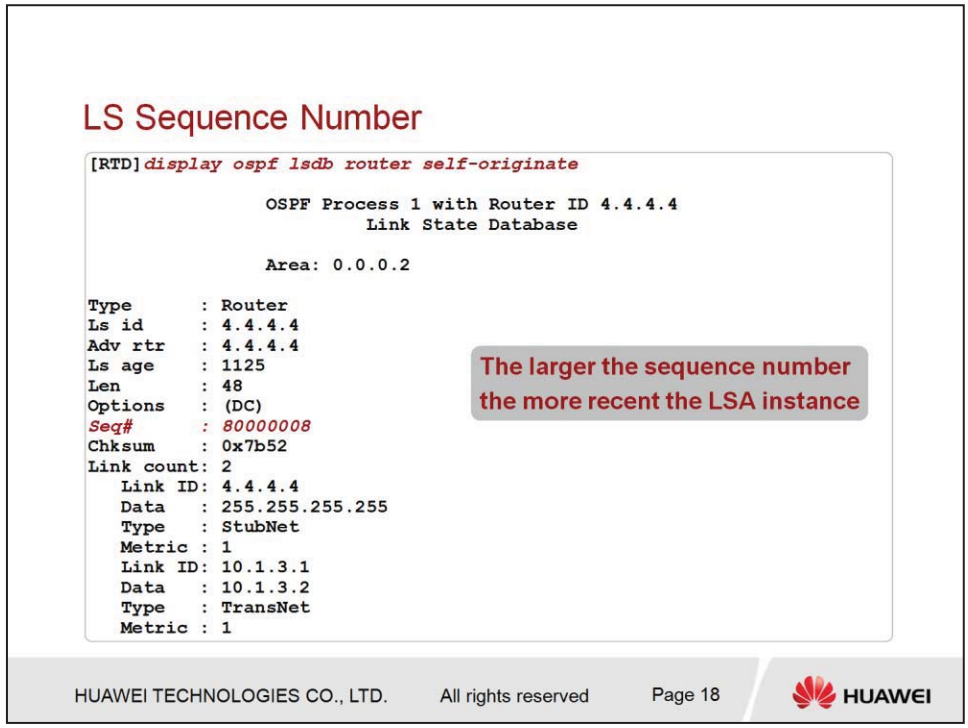
When advertising type 5 LSA received to other areas, ABR generates a type 4 LSA describing how to reach ASBR. Type 4 LSA can only be flooded within the area, when type 5 LSA is flooded into a certain area, the ABR associated with this area must create a new type 4 LSA for the area.

Link State ID

LSA Name	Link State ID
Router-LSA	The originating router's Router ID
Network-LSA	The IP interface address of the network's DR
Network-Summary-LSA	The destination network's IP address
ASBR-Summary-LSA	The Router ID of the described ASBR
AS-External-LSA	The destination network's IP address

HUAWEI TECHNOLOGIES CO., LTD. All rights reserved HUAWEI

Link State ID identifies a piece of the routing domain that is being described by the LSA. Depending on the LSA's LS type, the Link State ID takes on the different values.



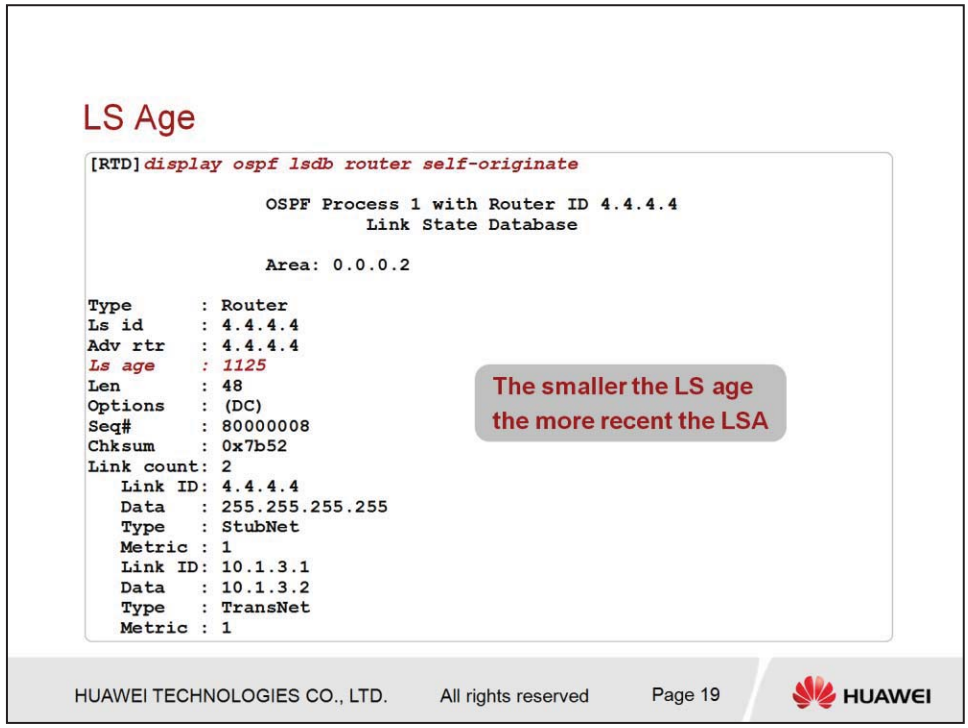
The combination of LS type, Link State ID and Advertising Router uniquely identifies one LSA, one LSA may have multiple instances. LS sequence number is used to detect which instance is more recent.

LS Sequence Number：

LS Sequence Number is a signed 32-bit integer. It is used to detect old and duplicate LSAs.

The sequence number 0x80000000(231) is reserved. This leaves 0x80000001 (231+1) as the smallest sequence number. This sequence number is used by router the first time it originates any LSA. Afterwards, the LSA's sequence number is incremented each time the router originates a new instance of LSA. The larger the sequence number the more recent the LSA.

If the router received a self-originated LSA which is newer than the last instance that the router actually originated, the router must then advance the LSA's LS sequence number one past the received LS sequence number, and originate a new instance of the LSA.



LS Age：

This field is the age of the LSA in seconds. LSAs are also aged as they are held in each router's database.

LS age must be incremented by InfTransDelay on every hop of the flooding procedure.

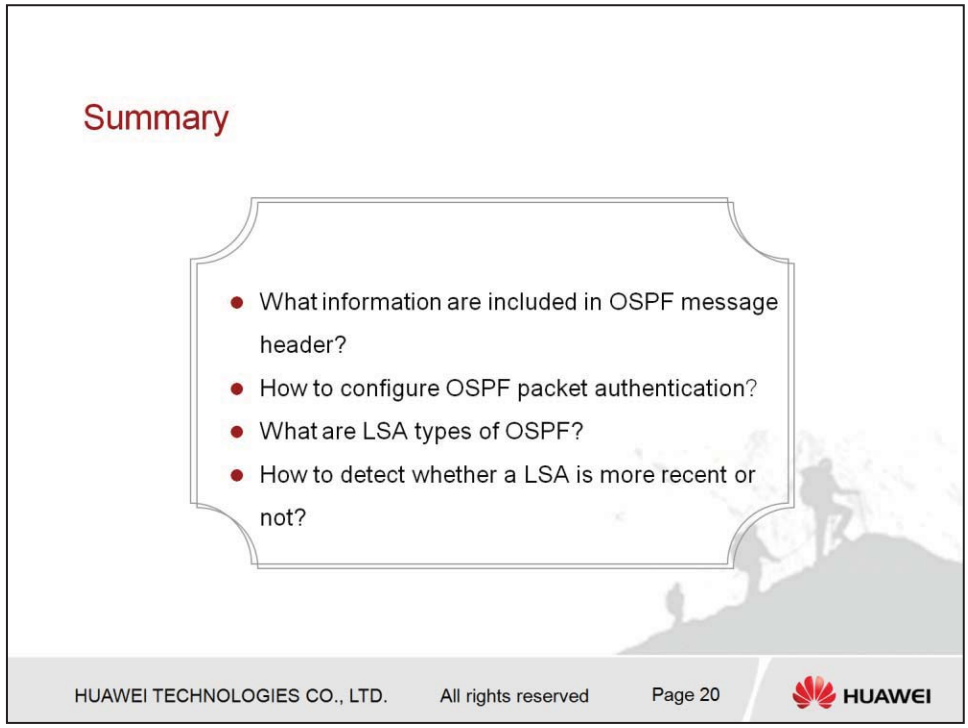
If the LS age field of one of the router's self-originated LSAs reaches the value LSRefreshTime, a new instance of the LSA is originated.

A MaxAge LSA must be removed from the router's LSDB.

The smaller the LS Age, the more recent the LSA.

If the router is to delete the self-originated LSA, it create a new LSA instance whose LS Age is set as Max Age.

If the router receives MaxAge LSA, the LSA in LSDB is deleted (if the LSA exists in LSDB).



What information are included in OSPF message header?

Check version number, Area ID, AuType and Authentication

How to configure OSPF packet authentication?

supports area authentication and interface authentication.

When using area authentication, all routers' authentication mode and password must mach within the same area; interface authentication is used to authenticate between neighboring routers, which overrides area authentication.

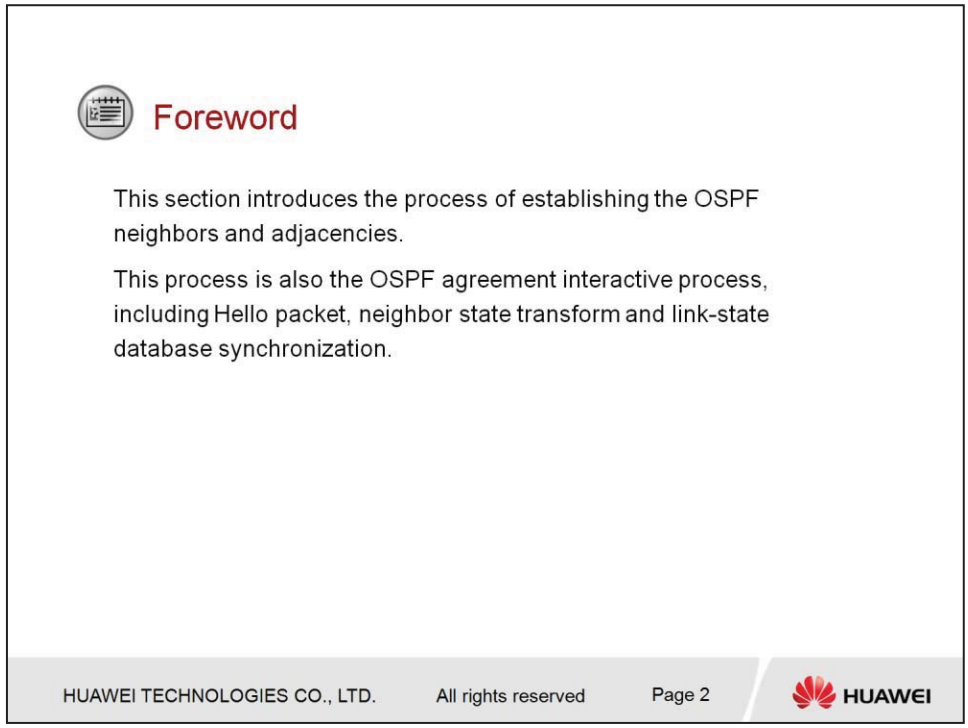
What are LSA types of OSPF?

Common LSAs include: Router-LSA，Network-LSA，Network-Summary-LSA,ASBR-Summary-LSA and AS-External-LSA.

How to detect whether a LSA is more recent or not?

Using LS Sequence Number and LS age. The larger the Sequence Number, the more recent the LSA, if the Sequence Number is the same, LS age is compared, the smaller the LS age, the more recent the LSA.


Establishing OSPF
Neighbors & Adjacencies
www.huawei.com
HUAWEI TECHNOLOGIES CO., LTD.
All rights reserved
HUAWEI


Foreword
This section introduces the process of establishing the OSPF neighbors and adjacencies.
This process is also the OSPF agreement interactive process, including Hello packet, neighbor state transform and link-state database synchronization.
HUAWEI TECHNOLOGIES CO., LTD.
All rights reserved
Page 2
HUAWEI

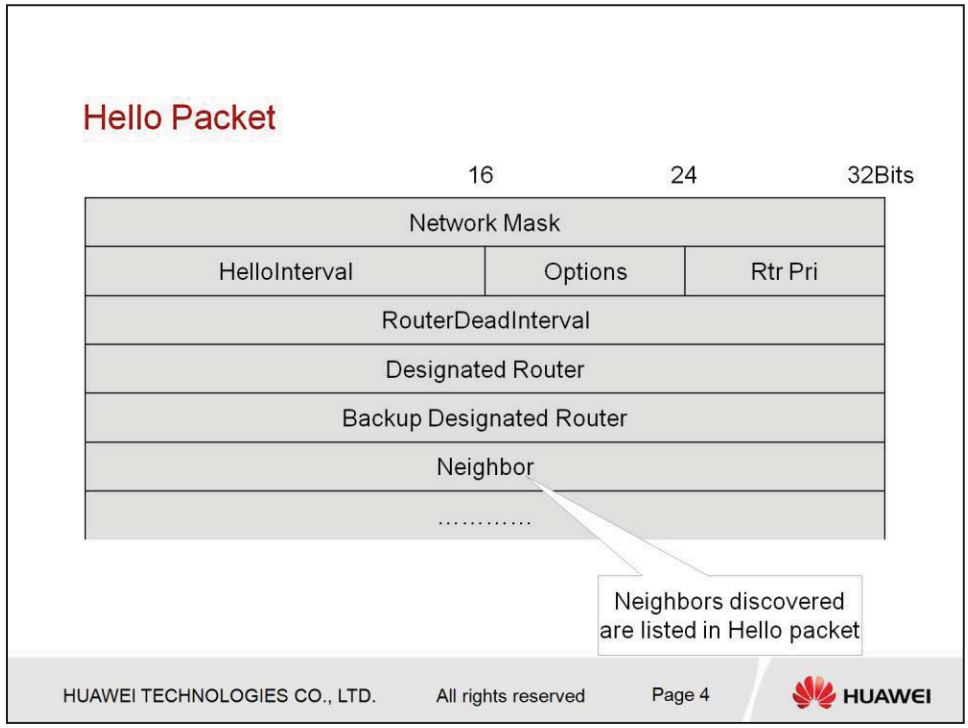


Objectives

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

- Understand the function of Hello packet
- Understand OSPF neighbor state transform stages
- Understand the process of establishing OSPF neighbor relationships and adjacencies
- Understand the process involved in LSDB synchronization

HUAWEI TECHNOLOGIES CO., LTD. All rights reserved



Key fields：

Network Mask：the network mask associated with this interface

HelloInterval：the number of seconds between this router's Hello packets.

Options：the optional capabilities supported by the router, detailed information is beyond the scope of this course.

Rtr Pri：this router's Router Priority. Used in DR and BDR election.

RouterDeadInterval：the number of seconds before the router's neighbors will declare it down, when they stop hearing the router's Hello Packets. The common value is 4 times

HelloInterval.

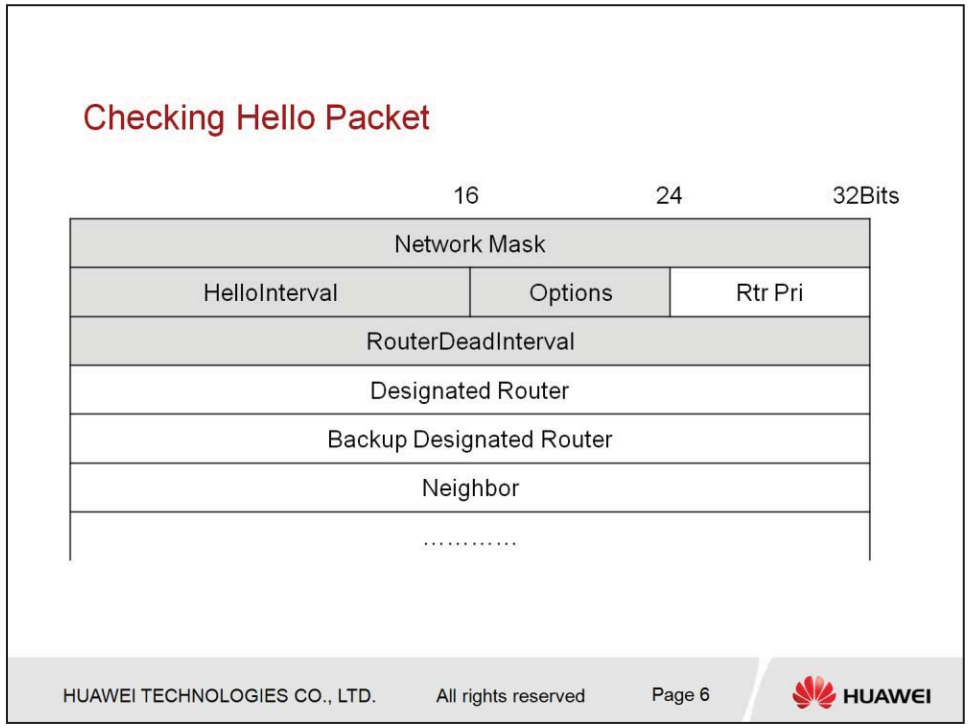
Designated Router：the identity of the Designated Router for this network, in the view of the sending router. The Designated Router is identified here by its IP interface address on

the network. Set to 0.0.0.0 if there is no Designated Router.

Backup Designated Router：the identity of the Backup Designated Router for this network, in the view of the sending router. The Backup Designated Router is identified here by its IP

interface address on the network. Set to 0.0.0.0 if there is no Backup Designated Router.

Neighbor：the Router IDs of each router from whom valid Hello packets have been seen recently on the network.



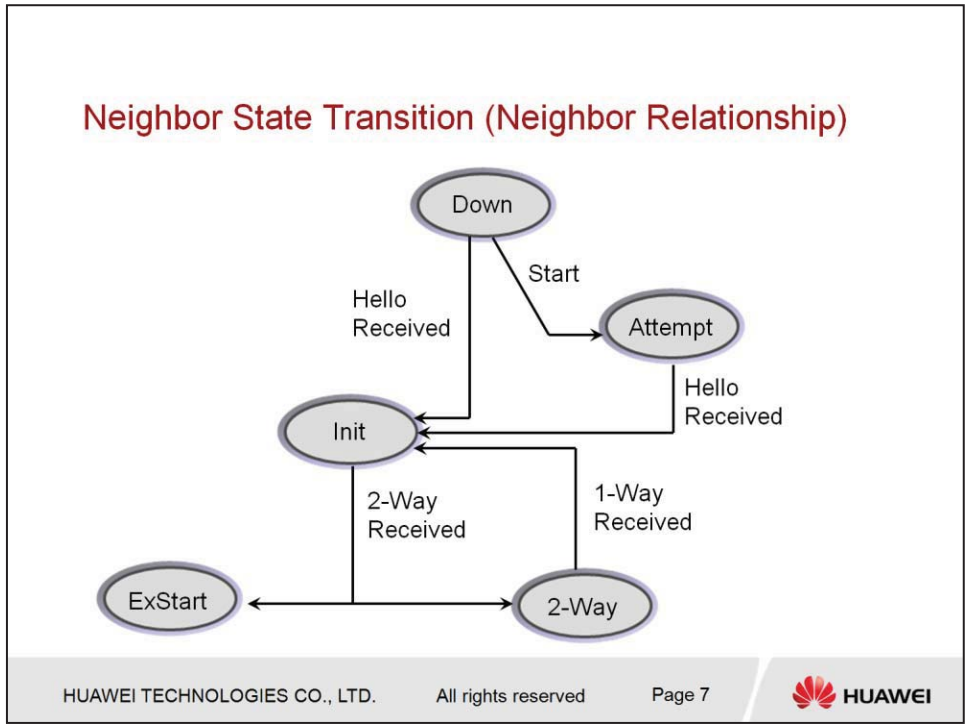
The validity of OSPF packet must be checked before checking the validity of Hello packet. When checking the validity of Hello packet, the following requirements are considered:

1.The values of the Network Mask, HelloInterval, and RouterDeadInterval fields in the received Hello packet must be checked against the values configured for the receiving interface. Any mismatch causes processing to stop and the packet to be dropped. However there is one exception to the above rule: on point-to-point networks and on virtual links, the Network Mask in the received Hello Packet should be ignored.

2.The setting of the E-bit (indicate whether accept AS external route) found in the Hello Packet's Options field must match this area's ExternalRoutingCapability.

The detailed information is covered in < OSPF Special Areas>.

If the router finds that its Router ID is included in the neighbor list of Hello Packet received, neighbor relationship is considered to be established.



This figure shows the process of establishing neighbor relationship and the transition of neighbor states.

Down：this is the initial state of a neighbor conversation. It indicates that there has been no recent information received from the neighbor. On NBMA networks, Hello packets may still be sent to “Down” neighbors every PollInterval seconds, generally this interval is equal to RouterDeadInterval.

Attempt：this state is only valid for neighbors attached to NBMA networks. It indicates that no recent information has been received from the neighbor, but that a more concerted effort should be made to contact the neighbor. This is done by

sending the neighbor Hello packets at intervals of HelloInterval. If no Hello packet is heard from neighbor in RouterDeadInterval, the state will transit to Down.

*On NBMA networks some configuration information may be necessary for the operation of the Hello Protocol. Each router that may potentially become Designated Router has a list of all other routers attached to the network. A router, having Designated Router potential, sends Hello Packets to all other potential Designated Routers when its interface to the NBMA network first becomes operational. This is an attempt to find the

Designated Router for the network. If the router itself is elected Designated Router, it begins sending Hello Packets to all other routers attached to the network. *

Init: in this state, a Hello packet has recently been seen from the neighbor.

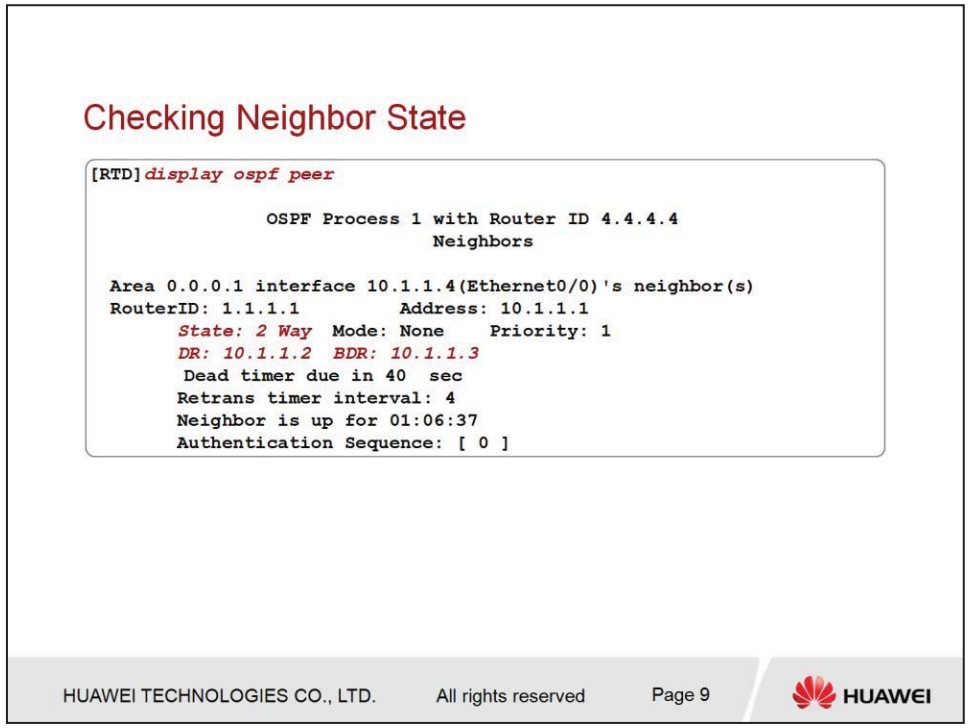
However the router itself did not appear in the neighbor's Hello packet indicating that bidirectional communication has not yet been established with the neighbor.

Neighbor in Init state should be contained in Hello packet sent by the router itself.

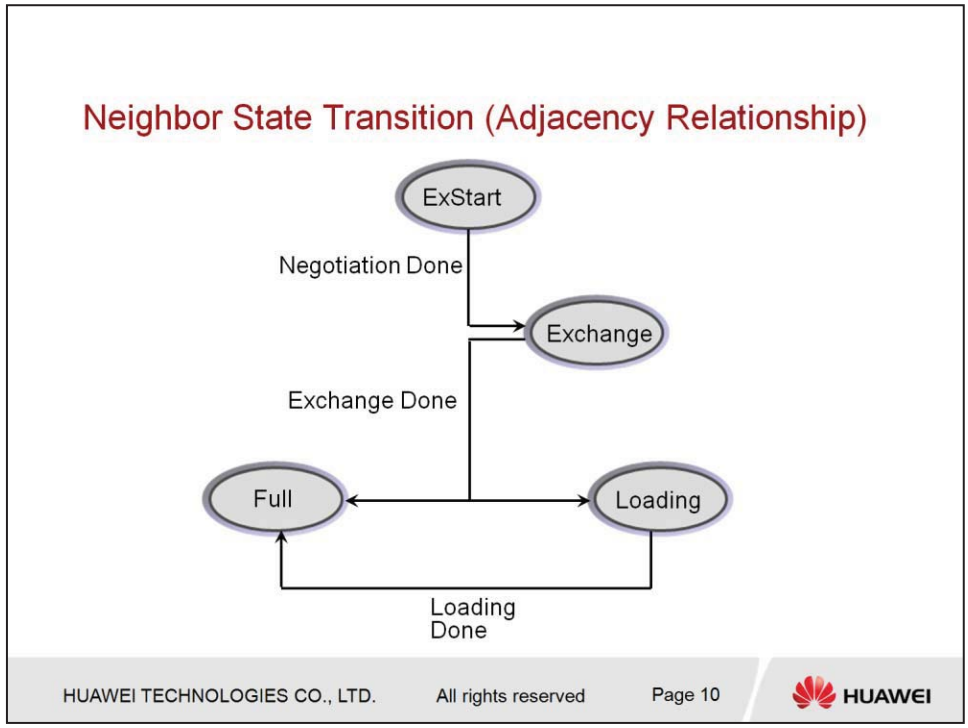
2-WayReceived: Bidirectional communication has been realized between the two neighboring routers. This is indicated by the router seeing itself in the neighbor's Hello packet.

2-Way: in this state, communication between the two routers is bidirectional, but adjacency is not established. This is the most advanced state short of beginning adjacency establishment.

1-WayReceived：a Hello packet has been received form the neighbor, in which the router is not mentioned. This situation is usually caused by restarting of neighbor.



In this example, DR for the network is 10.1.1.2, BDR for the network is 10.1.1.3, RTD and 1.1.1.1 are all DRothers, so RTD can not establish adjacency with 1.1.1.1, their stable neighbor state is 2 Way.



DD Sequence Number：

Each DD packet has a sequence number which is used for acknowledgement of DD packet. The sequence number is 2 bytes .

Master/Slave：

Master and Slave relationship should be established before sending DD packets to exchange LSDB information. If Master and Slave relationship is determined, Master is responsible for sending DD packet, Slave can only send a DD packet whose sequence number must equal to that of DD packet sent by Master.

Related neighbor states:

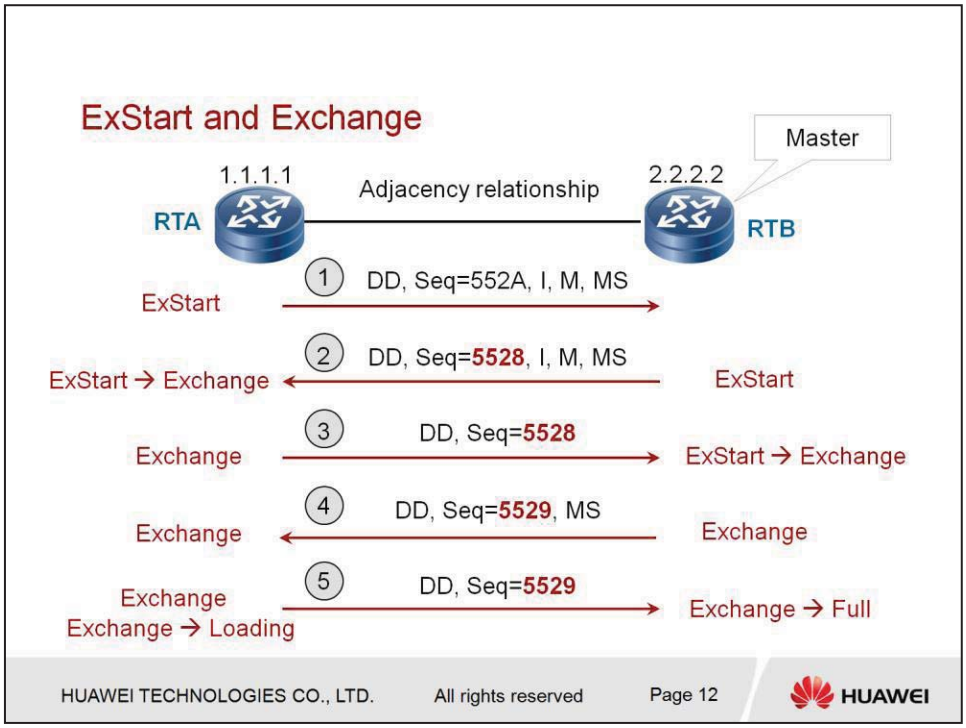
ExStart ： this is the first step in creating an adjacency between the two neighboring routers. The goal of this step is to decide which router is the master, and to decide upon the initial DD sequence number. Link state information is not included in DD packet sent by neighbor in this state.

Exchange：in this state the routers are sending DD packets containing abstract of link state information to each other.

Loading：in this state, LS Request packets are sent to the

neighbor asking for the more recent LSAs that have been discovered but not yet received in the exchange state; LS Update packets are sent to the neighbor sending LS Request packets.

Full：in this state, the neighboring routers' LSDBs are fully synchronized.



1. When the neighbor state transiting to ExStart, RTA sends the first DD packet to RTB, in this packet the DD sequence number is set to 552A, Initialize bit is set to indicate it's the first DD packet, More bit is set to indicate that there are more DD packets to be sent, Master bit is set to declare itself the Master.

2. When the neighbor state transiting to ExStart, RTB sends the first DD packet to RTA, in this packet the DD sequence number is set to 5528. RTB should be the Master, for RTB's Router ID is bigger than that of RTA. After finishing comparison of Router ID, RTA will generate the event called NegotiationDone to transit the state from ExStart to Exchange.

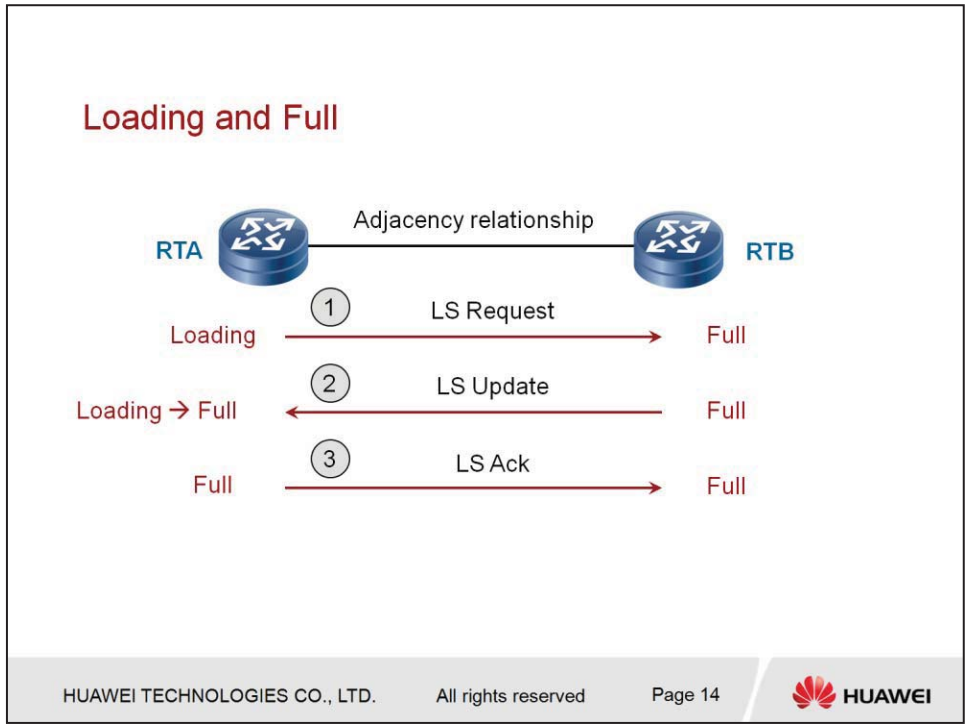
3. When the neighbor state transiting to Exchange, RTA sends the new DD packet, this packet contains abstract of LSDB, the DD Sequence number is equal to that of RTB, More bit is cleared indicating this is the last DD packet, Master bit is cleared indicating RTA declares itself the Slave. After receiving this packet, RTB will generate the event called NegotiationDone to transit the state to Exchange.

4. When the neighbor state transiting to Exchange, RTB sends a new DD packet, this packet contains abstract of LSDB, the DD Sequence number is set to 5529 (increments the previous

Sequence number).

5. Even RTA doesn't need any DD packet to describe its LSDB, but as the Slave, RTA must acknowledge every DD packet sent by RTB.

So RTA send a new empty DD packet with Sequence number equal to 5529. After sending the last DD packet, RTA generates the event called ExchangeDone and transits the neighbor state to Loading. After receiving the last DD packet, RTB transiting the neighbor state to Full. (Assuming that RTB's LSDB is the most recent and complete, it doesn't need to request for update from RTA)

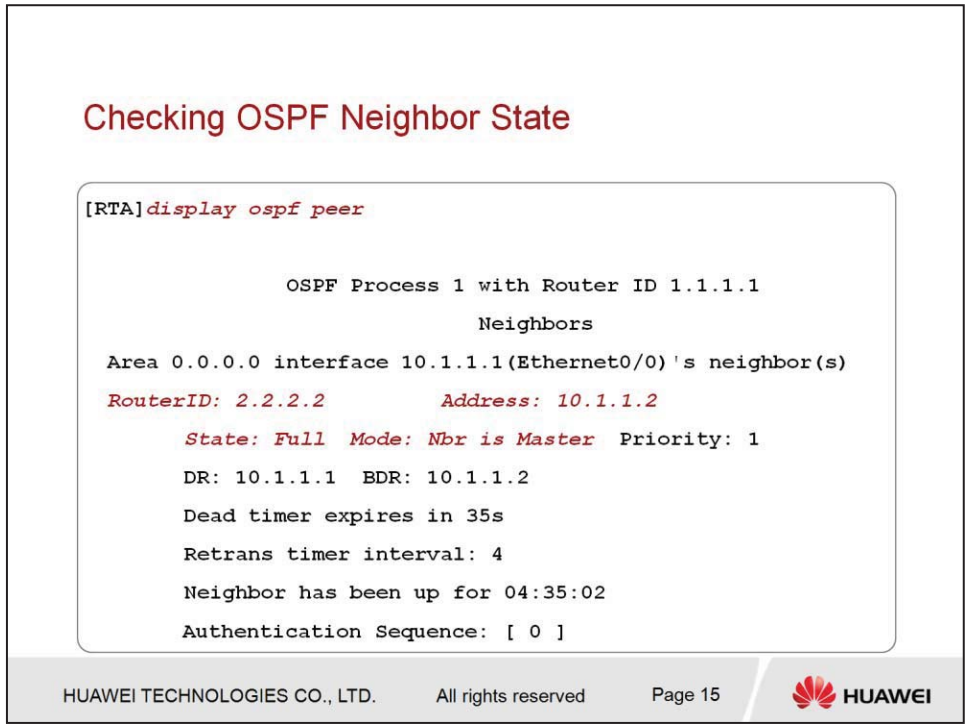


1. When the neighbor state transiting to Loading, RTA start sending LS request packet to RTB, requesting for link state information that is not contained in local LSDB but found by DD packets.

2. After receiving LS Request packet from RTA, RTB sends LS Update packet to RTA, this packet contains the detailed link state information requested. After receiving LS Update packet, RTA transits the neighbor state from Loading to Full.

3. RTA sends LS Ack to RTB to ensure the transmission reliability. LS Ack packets are used for acknowledgement of received LSA.

The neighbor state becomes Full, indicating that neighboring routers are fully adjacent.



RouterID：neighbor’s Router ID

Address：neighbor’s interface IP address on the network

State：neighbor’s state, Full state indicates full adjacency

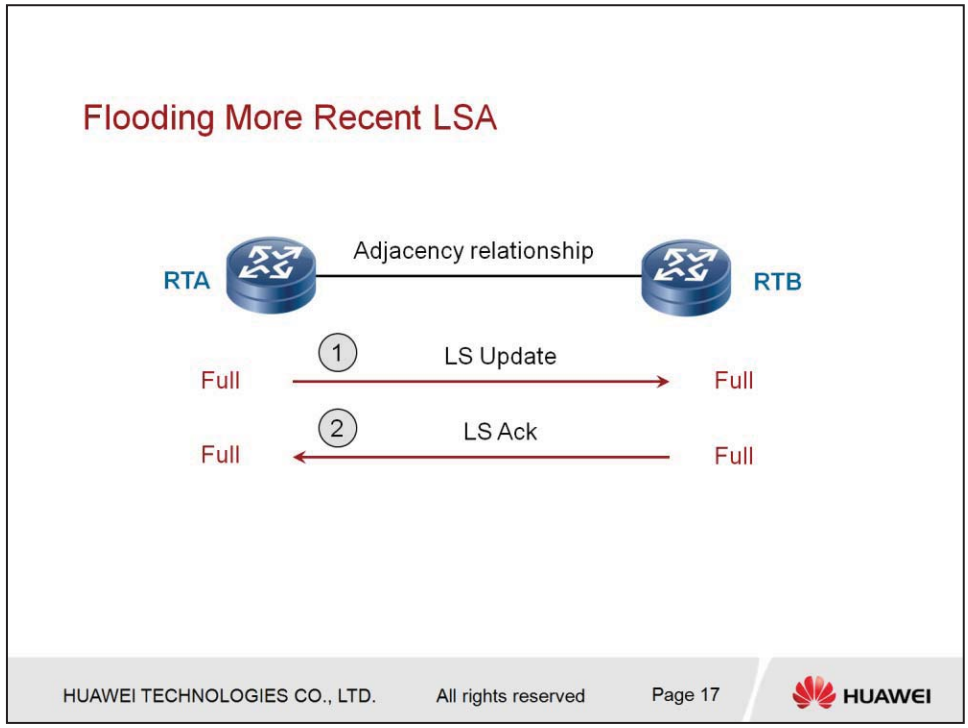
Mode：when exchanging DD packets, whether the neighbor is Master or Slave.

LSA Information Contained in Various Packets

Packet Type	LSA Information
Database Description	LSA header information, including LS Type, LS ID, Advertising Router and LS Sequence Number
LS Request	LS Type, LS ID and Advertising Router
LS Update	Complete LSA information, including LSA header and detailed link state information
LS Ack	LSA header information, including LS Type, LS ID, Advertising Router and LS Sequence Number

HUAWEI TECHNOLOGIES CO., LTD. All rights reserved HUAWEI

Different OSPF protocol packets have different parts of LSA information.



When new LSA is received or generated, this LSA should be flooded.

LSA flooding only need LS Update packet and LS Ack packet.

1. If RTA is to flood new LSA, RTA sends a LS Update packet to RTB, the new LSA is contained in this packet.

2. After receiving the new LSA, RTB send a LS Ack packet to RTA for acknowledgement.

When flooding new LSA between two fully adjacent routers, their neighbor states are not affected.

OSPF Packet Destination Address

	Hello	Database Description	Link State Request	Link State Update	Link State Ack
Point-to-point	224.0.0.5	224.0.0.5	224.0.0.5	224.0.0.5	224.0.0.5
NBMA	unicast	unicast	unicast	unicast	unicast
Virtual link	unicast	unicast	unicast	unicast	unicast

HUAWEI TECHNOLOGIES CO., LTD. All rights reserved HUAWEI

All packets are sent to the multicast address 224.0.0.5 on Point-to-Point network.

All packets are sent to the unicast address on NBMA network, the packet destination is the munally configured neighbor.

Packets on virtual link are sent as a unicast.

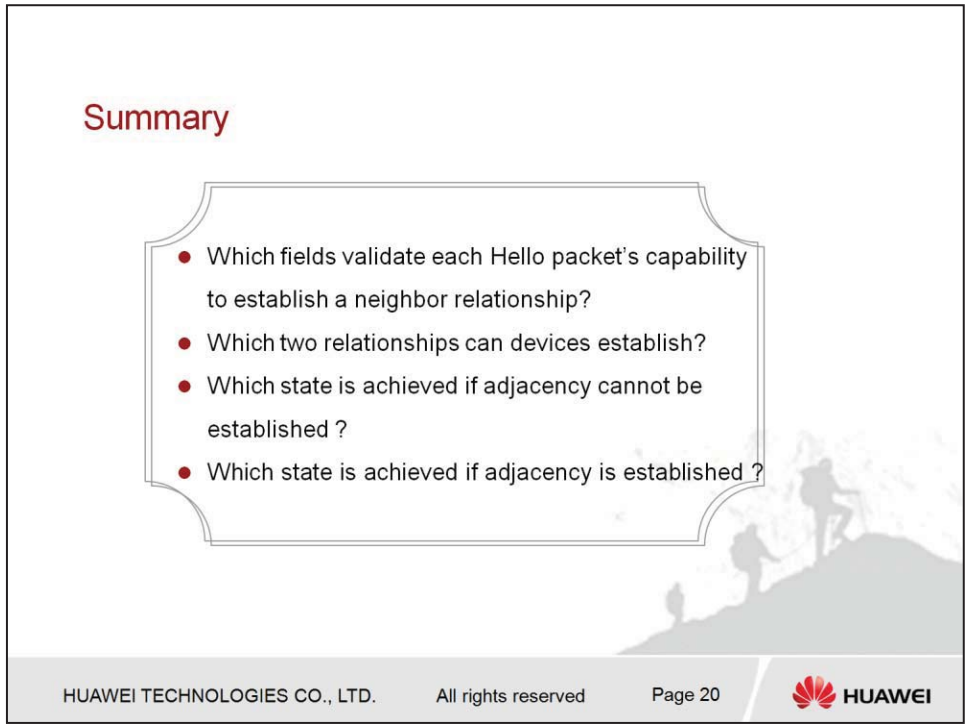
OSPF Packet Destination Address

	Hello	Database Description	Link State Request	Link State Update	Link State Ack
Broadcast	224.0.0.5	unicast	unicast	224.0.0.5 or 224.0.0.6	224.0.0.5 or 224.0.0.6
Point-to-MultiPoint	224.0.0.5	unicast	unicast	224.0.0.5 or unicast	224.0.0.5

HUAWEI TECHNOLOGIES CO., LTD. All rights reserved Page 19 HUAWEI

On broadcast network, DR and BDR send LS Update packet and LS Ack packet to the destination 224.0.0.5 (AllSPFRouters), other routers send LS Update packet and LS Ack packet to the destination 224.0.0.6 (AllDRouters).

On Point-to-MultiPoint network, if LS Update packet is sent for the purpose of acknowledgement of the LS Request packet, the LS Update packet is sent as a unicast, if LS Update packet is sent for flooding new LSA, the LS Update packet is sent to the destination 224.0.0.5 (AllSPFRouters).



How to check the valid Hello packet?

Check Network Mask，HelloInterval，RouterDeadInterval and E-bit in Options field.

What are stages of neighbor state transition?

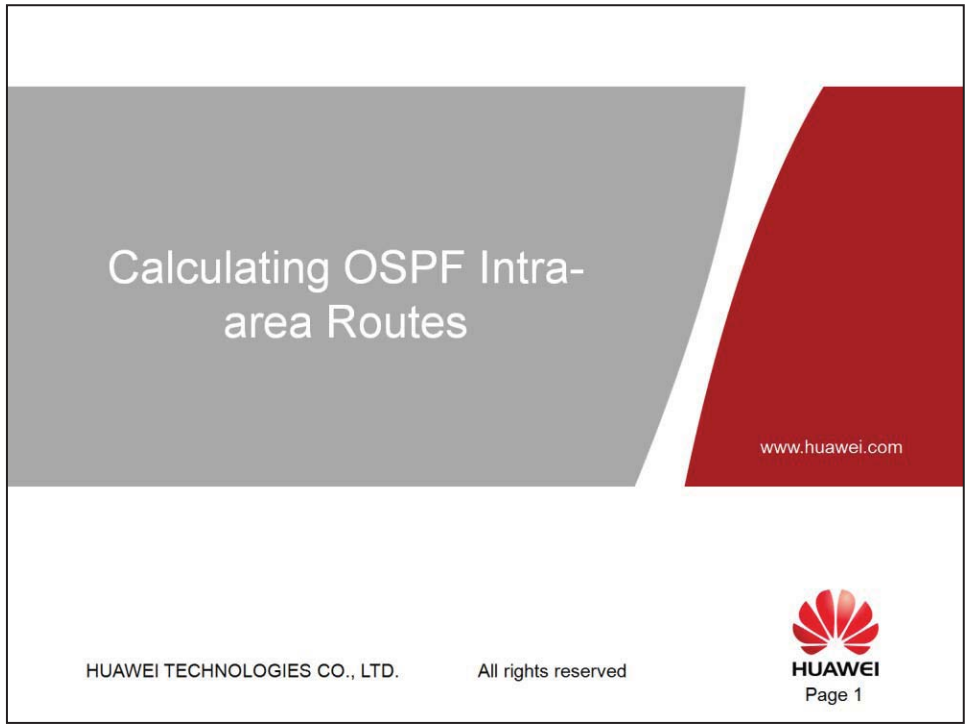
Two stages: the first stage is neighbor establishment, the second stage is adjacency establishment.

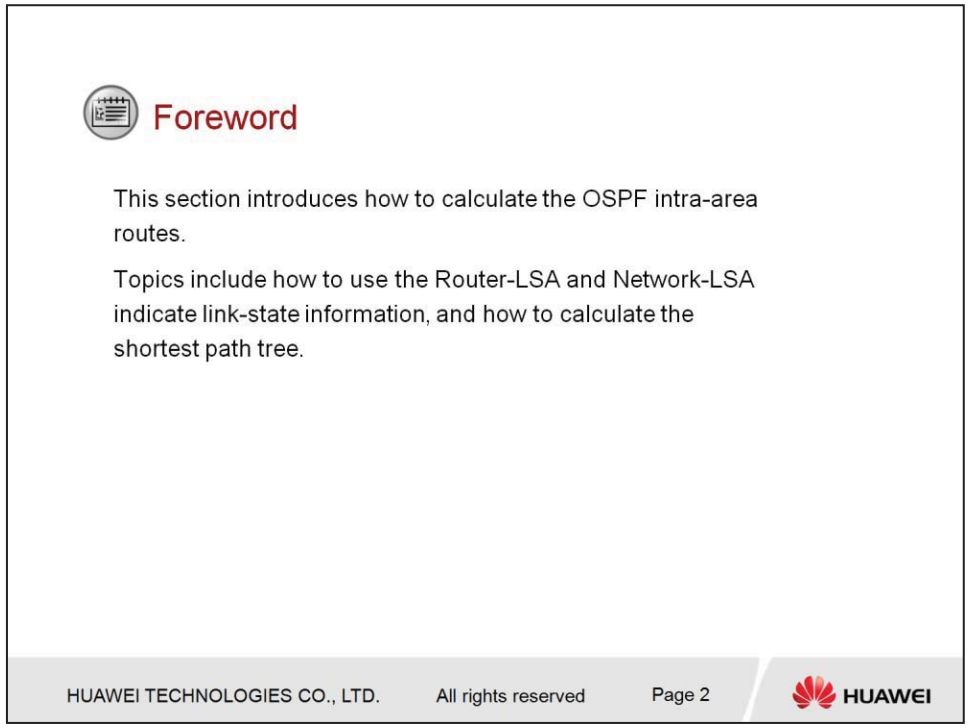
What's the stable neighbor state if neighboring routers can not establish adjacency relationship?

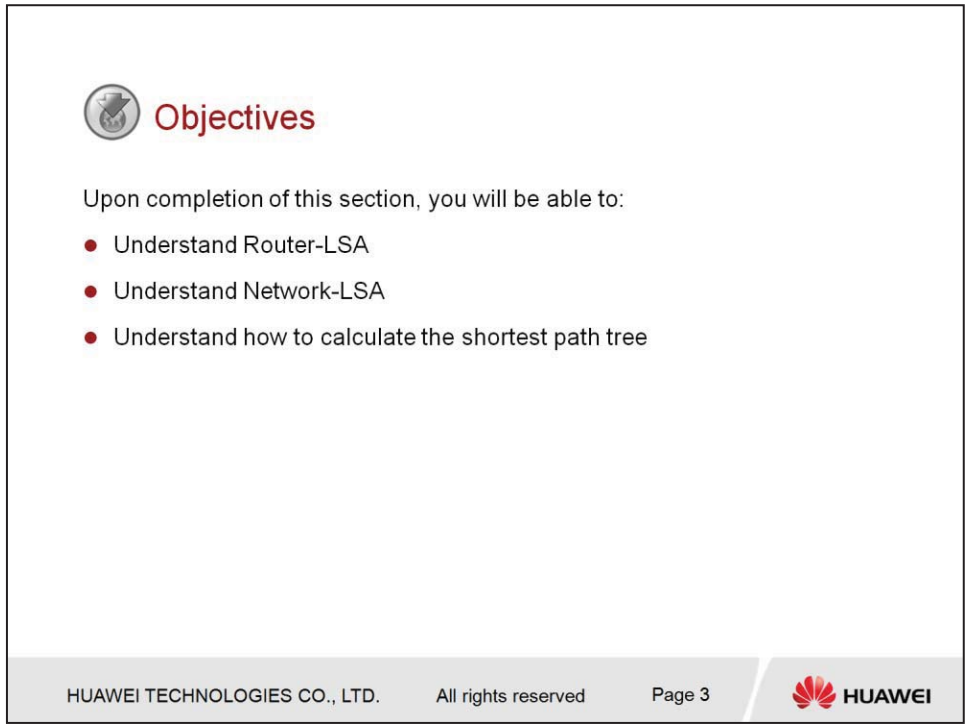
The stable neighbor state is 2 way.

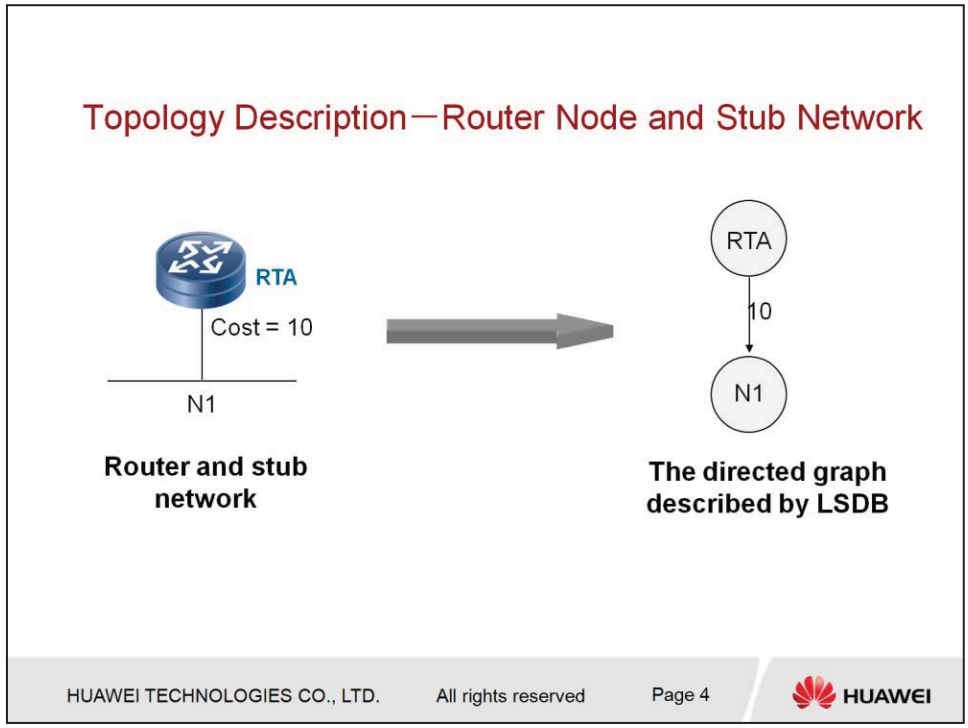
What's the stable neighbor state if neighboring routers can establish adjacency relationship?

The stable neighbor state is Full.


Calculating OSPF Intra-area Routes
www.huawei.com
HUAWEI TECHNOLOGIES CO., LTD.
All rights reserved
HUAWEI
Page 1


Foreword
This section introduces how to calculate the OSPF intra-area routes.
Topics include how to use the Router-LSA and Network-LSA indicate link-state information, and how to calculate the shortest path tree.
HUAWEI TECHNOLOGIES CO., LTD.
All rights reserved
Page 2
HUAWEI


Objectives
Upon completion of this section, you will be able to:
Understand Router-LSA
Understand Network-LSA
Understand how to calculate the shortest path tree
HUAWEI TECHNOLOGIES CO., LTD.
All rights reserved
Page 3
HUAWEI

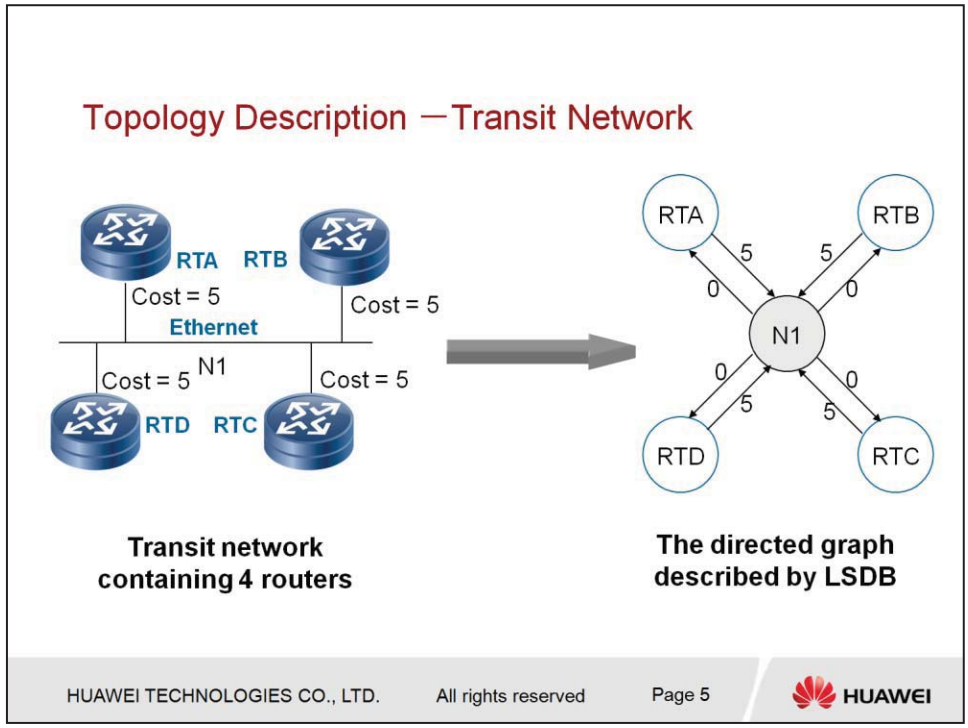


LSDB describes network topology by using a directed graph. The graph's vertices are routers, transit networks and stub networks.

A stub network's vertex has only incoming edges. E.g. the Loopback interface is a stub network.

This picture describes how to represent routers and stub networks.

Cost indicates the cost from one vertex to another, it is configurable on OSPF interface, representing the interface output cost.

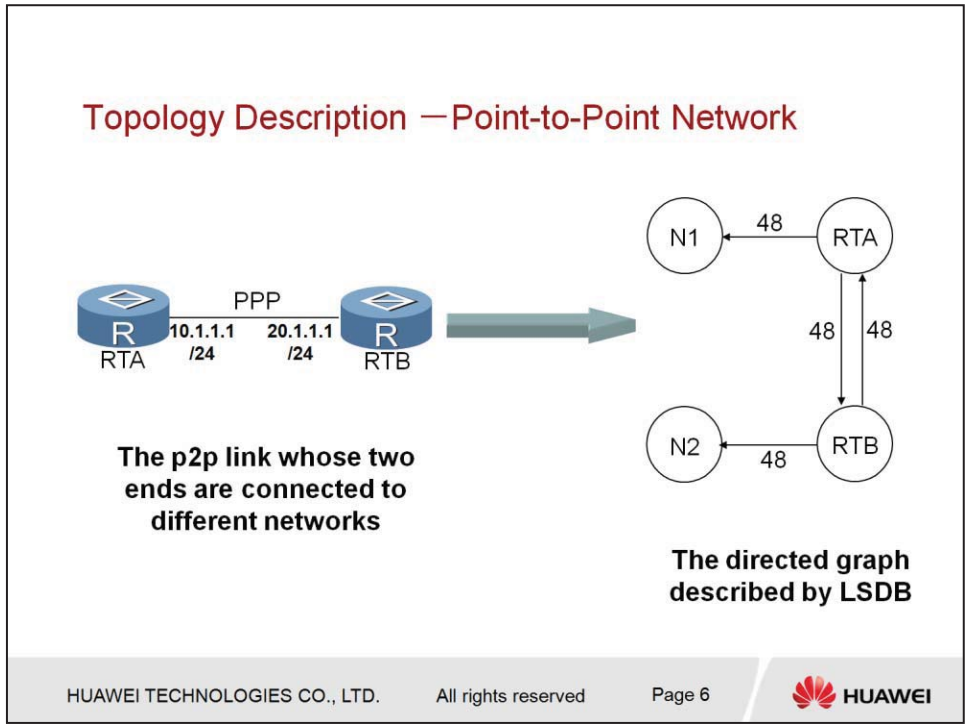


Transit networks are those capable of carrying data traffic that is neither locally originated nor locally destined.

A broadcast network or NBMA with more than two routers is a transit network.

The cost from the router to its attached Transit network is the cost of the interface attached to this network.

The cost from the router to its attached Transit network is 0.



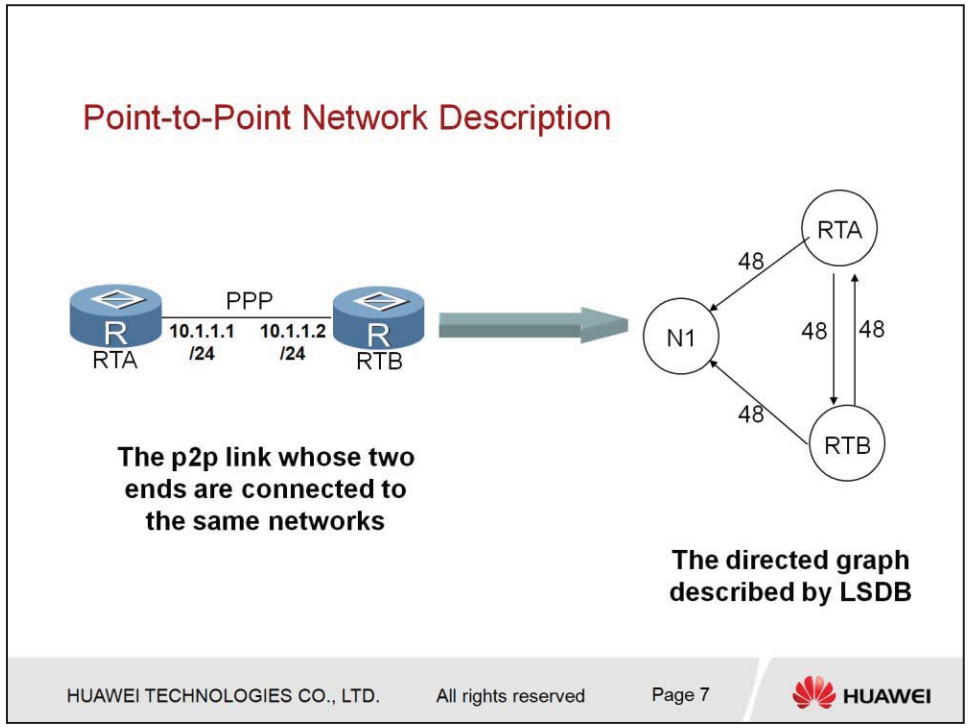
In this example, the costs of both interfaces are configured as 48. In the directed graph, N1 indicates 10.1.1.0/24, N2 indicates 20.1.1.0/24.

For Point-to-Point link whose two interfaces are attached to different networks, the rules for describing this type of link are:

Two routers are directly connected with two directed lines, one line for each direction.

Both interfaces' networks are stub network.

Each router advertises a stub link to the attached network.



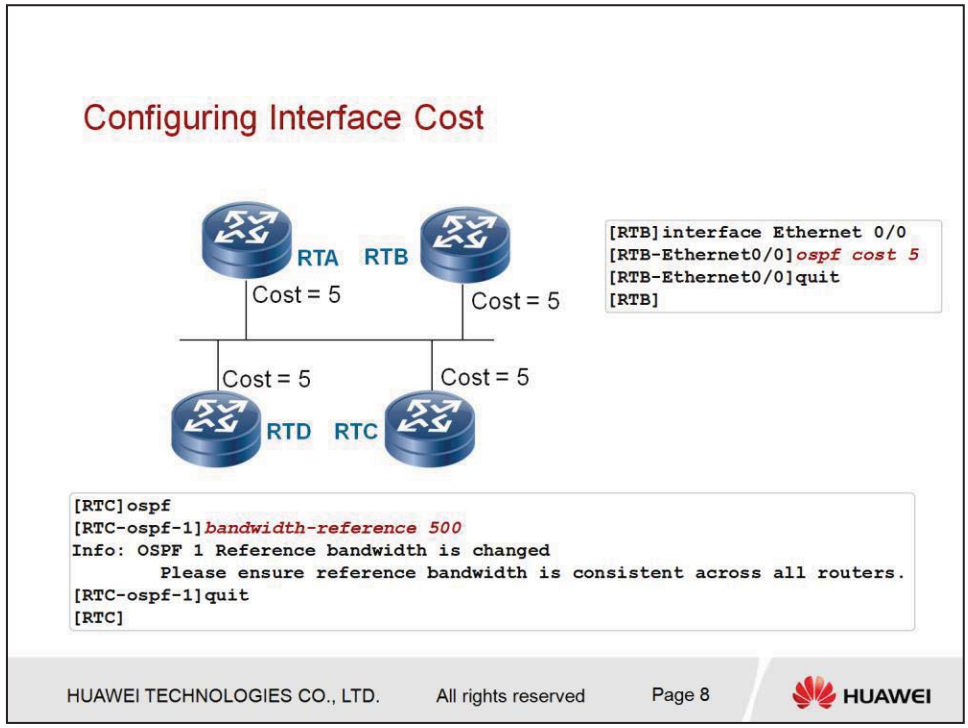
In this example, the costs for both interfaces are 48. In the directed graph, N1 indicates 10.1.1.0/24.

For Point-to-Point link whose two interfaces are attached to the same network, the rules for describing this type of link are:

Two routers are directly connected with two directed lines, one line for each direction.

The network connecting two interfaces is stub network.

Both routers advertise a stub link to the attached network.



By default, OSPF interface cost is associated with interface bandwidth, the equation is :

bandwidth-reference / bandwidth

The default value for bandwidth-reference is 100M, so the default value for FE interface cost is 1.

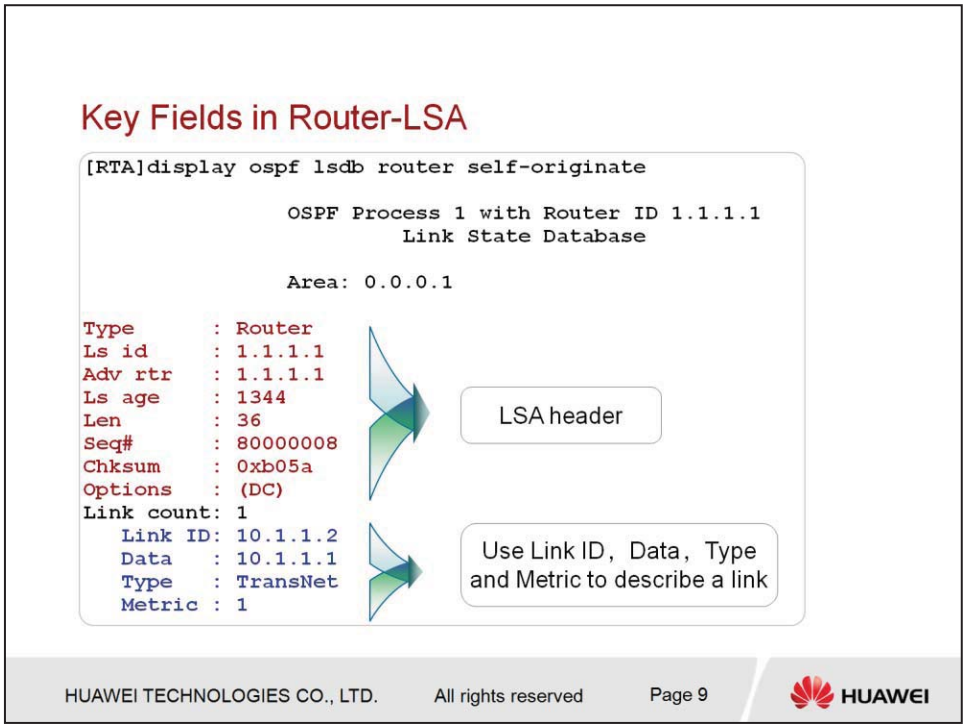
There are two ways to change interface cost, one way is to change the cost by the command “ospf cost” in interface view, the other way is to change bandwidth reference in OSPF view, then the system will calculate the cost automatically.

ospf cost cost

cost：OSPF interface cost, ranges from 1～65535.

bandwidth-reference value

value：reference value for calculating OSPF interface cost, the unit is Mbit/s, the range is 1～2147483648



A router-LSA is generated by each router in the area. It describes the state of the router's interfaces to the area. A router-LSA can describe multiple links, each link is described by Link ID, Data, Type and Metric.

1. Type：a quick description of the router link (not the network types supported by OSPF), Router-LSA describes 4 kinds of link type:

Point-to-Point: point-to-point connection to another router

TransNet：connection to a transit network, such as broadcast or NBMA network.

StubNet：connection to a stub network (such as Loopback interface).

Virtual：virtual link

2. Link ID：identifies the object that this router link connects to. Value depends on the link's Type.

3. Data：additional information describing the link, value again depends on the link's Type field.

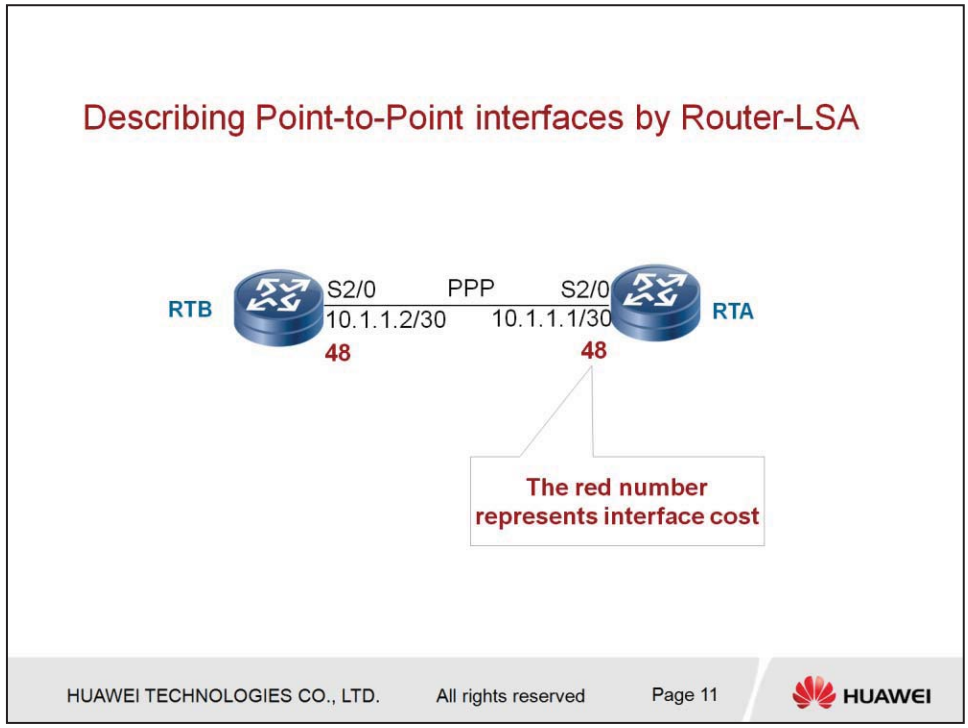
4. Metric：the cost of using this router link.

Key Fields in Router-LSA

Type	Link ID	Data
Point-to-point	Neighbor Router ID	The IP interface address of the originating router
TransNet	Interface address of DR	The IP interface address of the originating router
StubNet	IP network number	Attached network's address mask
Virtual	Virtual neighbor Router ID	The IP interface address associated with the virtual link

HUAWEI TECHNOLOGIES CO., LTD. All rights reserved HUAWEI

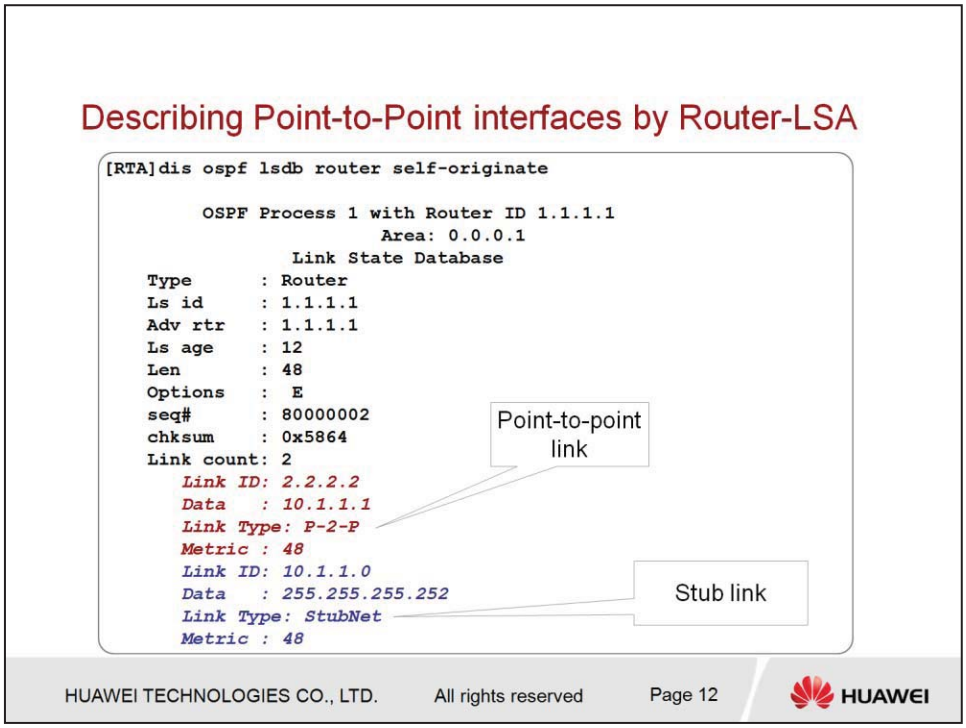
The contents of Link ID and Data depends on the link's Type field. This table shows their relationships.



In this example:

RTA's Router ID is 1.1.1.1; RTB's Router ID is 2.2.2.2

Both interfaces are connected to the same network, the network is within Area 1.



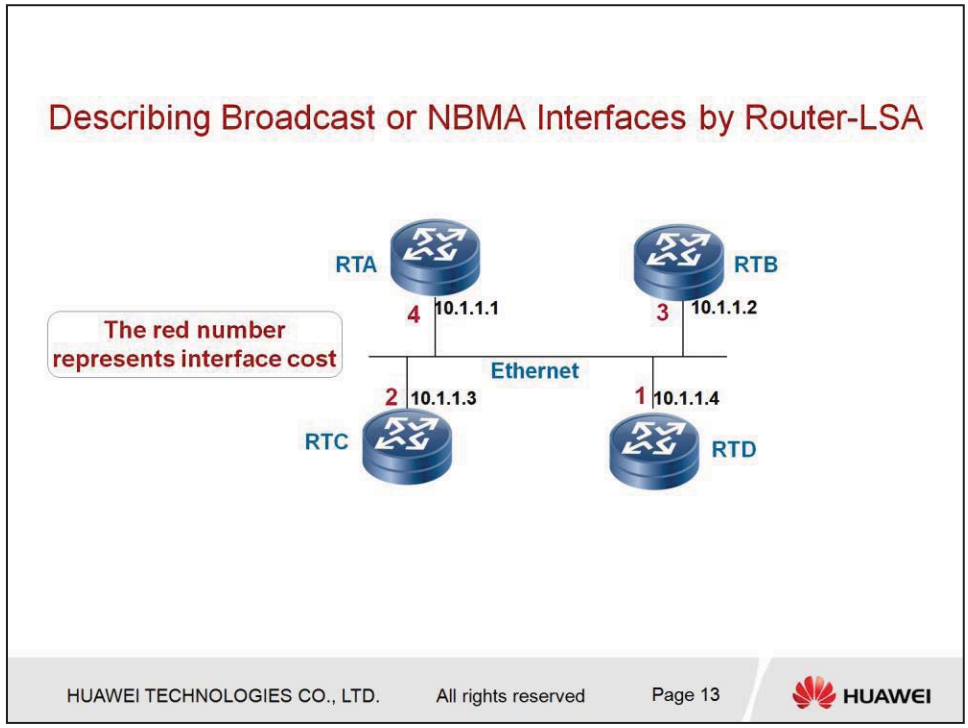
This figure shows Router-LSA generated by RTA.

This Router-LSA describes a PPP link:

1. Describe a point-to-point link, the Link ID should be set to the Router ID of the neighboring router, the Data field should specify the IP interface address.

2. Describe a stub network with Link ID set to the IP network number of the attached network, Link Data set to the attached network's address mask.

3. Both costs equal to the interface's configured output cost.

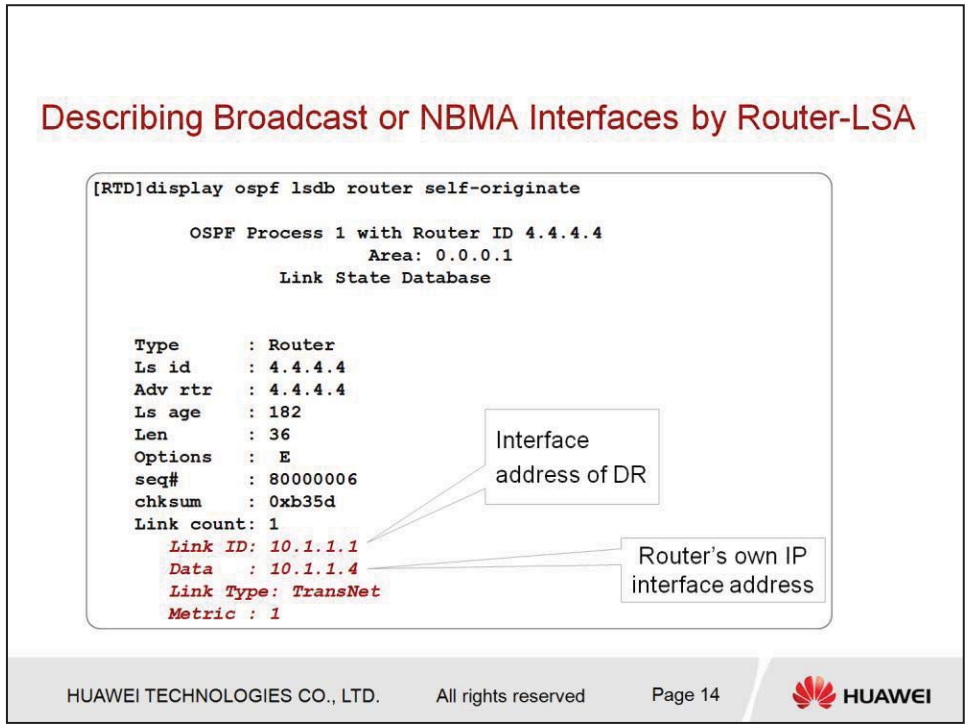


In this example:

RTA's Router ID is 1.1.1.1, RTB's Router ID is 2.2.2.2, RTC's Router ID is 3.3.3.3, RTD's Router ID is 4.4.4.4

Configure RTA to be DR by changing interface's Router Priority

Four interfaces are attached to the same network, the network is within Area 1.



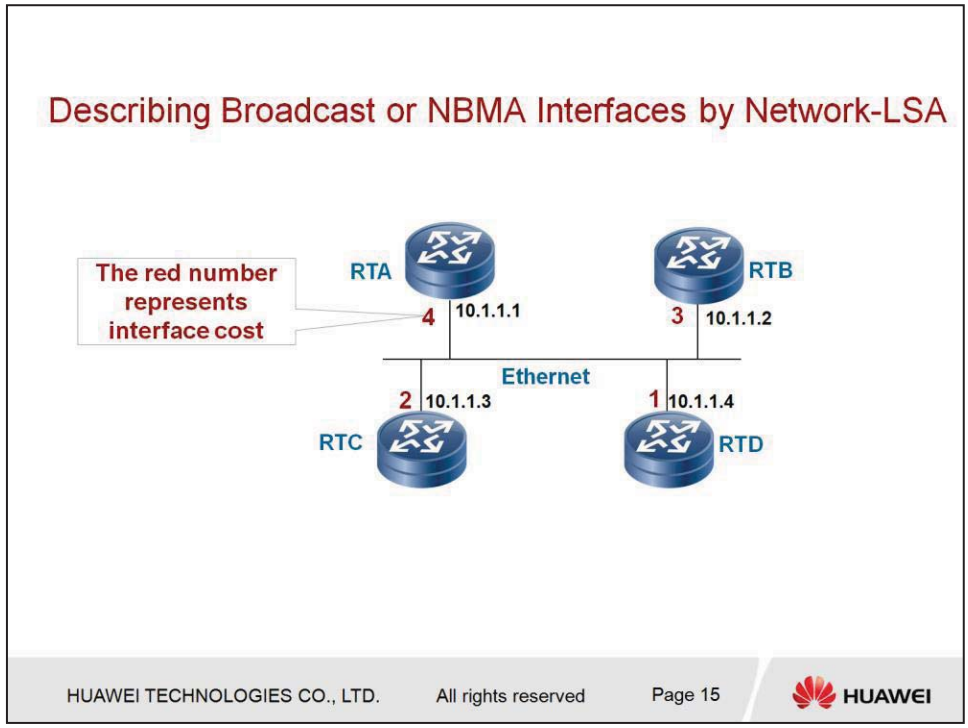
This is RTD's Router-LSA.

For Router-LSA describing broadcast and NBMA interfaces:

1. If the state of the interface is Waiting, or there is only one OSPF router exists in this network, or there is no DR on the network, then a stub network description is added to the router-LSA with Link ID set to the IP network number of the attached network, Link Data set to the attached network's address mask.

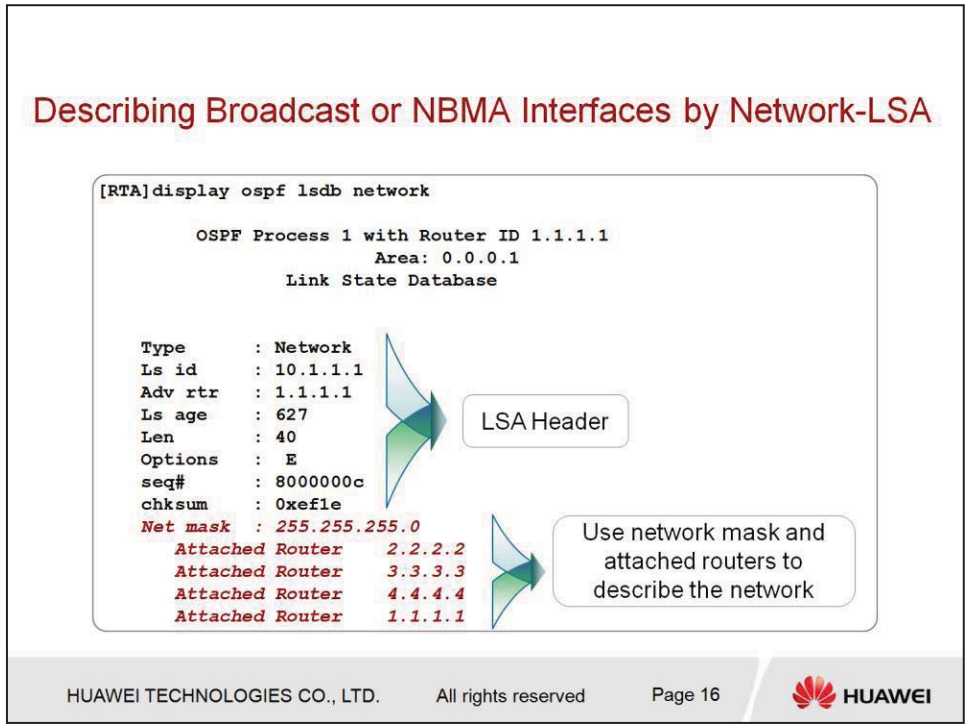
Else, there has been a Designated Router elected for the attached network. A transit network description is added to the router-LSA with Link ID set to the IP interface address of the attached network's Designated Router, Link Data set to the router's own IP interface address.

2. The cost is equal to interface's configured output cost.



In this example:

RTA's Router ID is 1.1.1.1, RTB's Router ID is 2.2.2.2, RTS's Router ID is 3.3.3.3, RTD's Router ID is 4.4.4.4 The network is within Area 1.



This is Area 1's network-LSA.

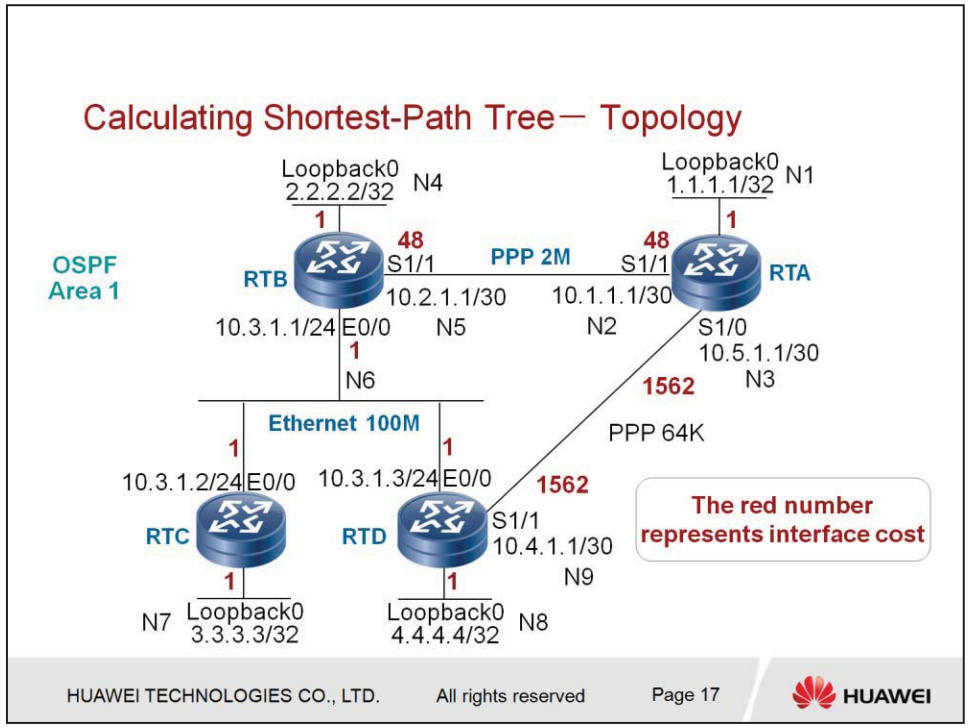
A network-LSA is generated for every transit broadcast or NBMA network. The network LSA describes all the routers that are attached to the network. In a network-LSA:

Link State ID for a network-LSA is the IP interface address of the DR Net mask for a network-LSA is the network's address mask.

Link State ID masked by the Net mask yields the network's IP address.

The network-LSA also contains an attached routers' list.

There is no cost for the link between transit network and attached routers.

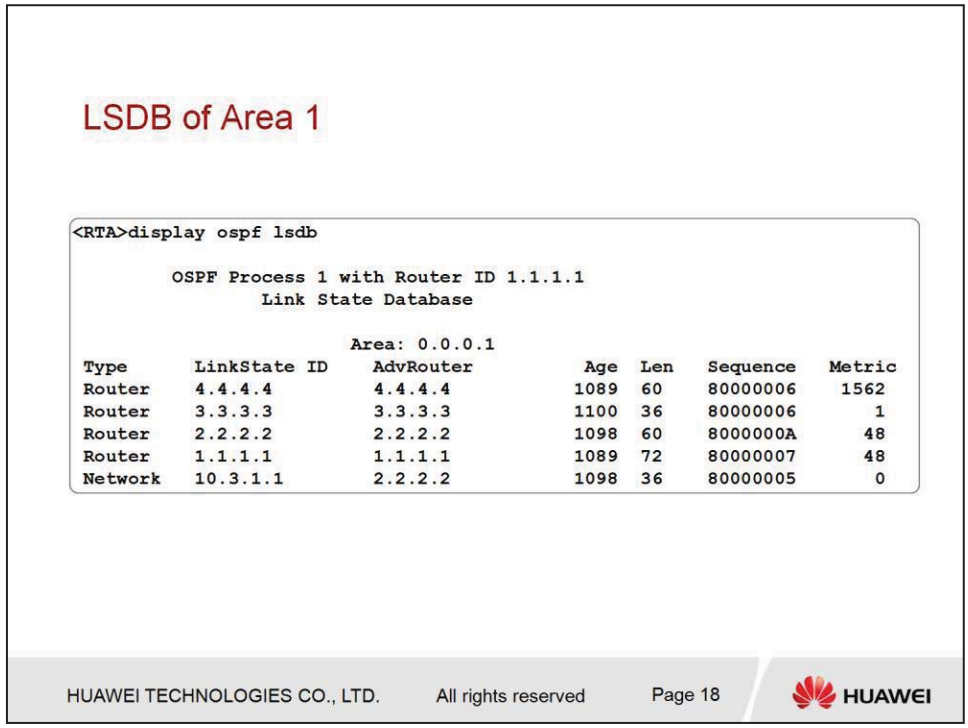


In this example:

All networks are within Area 1.

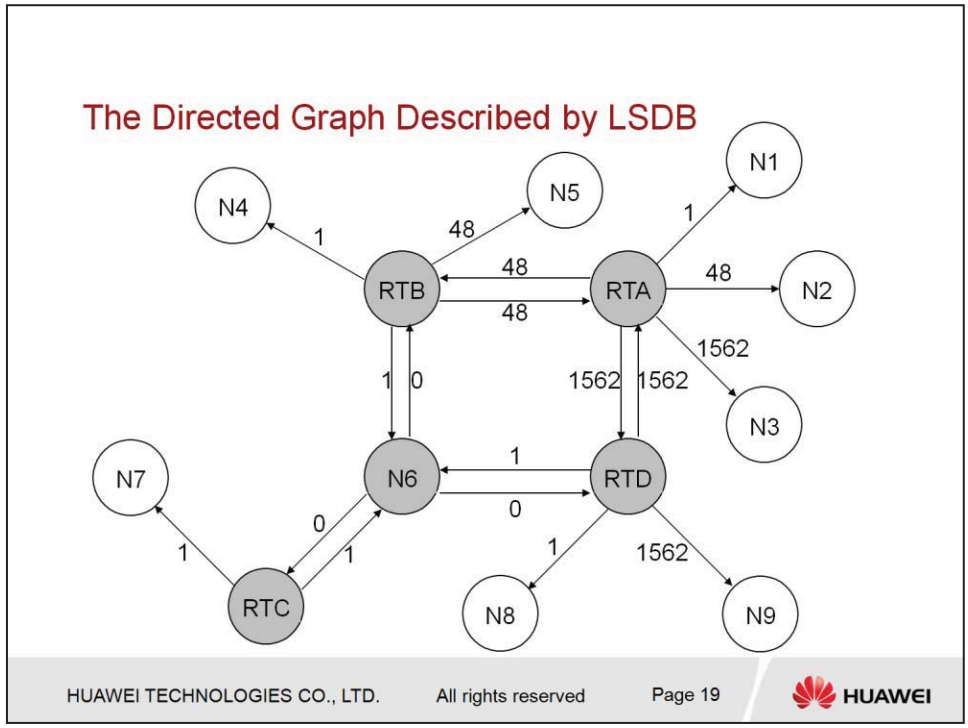
There are two Point-to-Point networks and a broadcast network

Each router’s Router ID is router’s Loopback 0’s interface IP address.



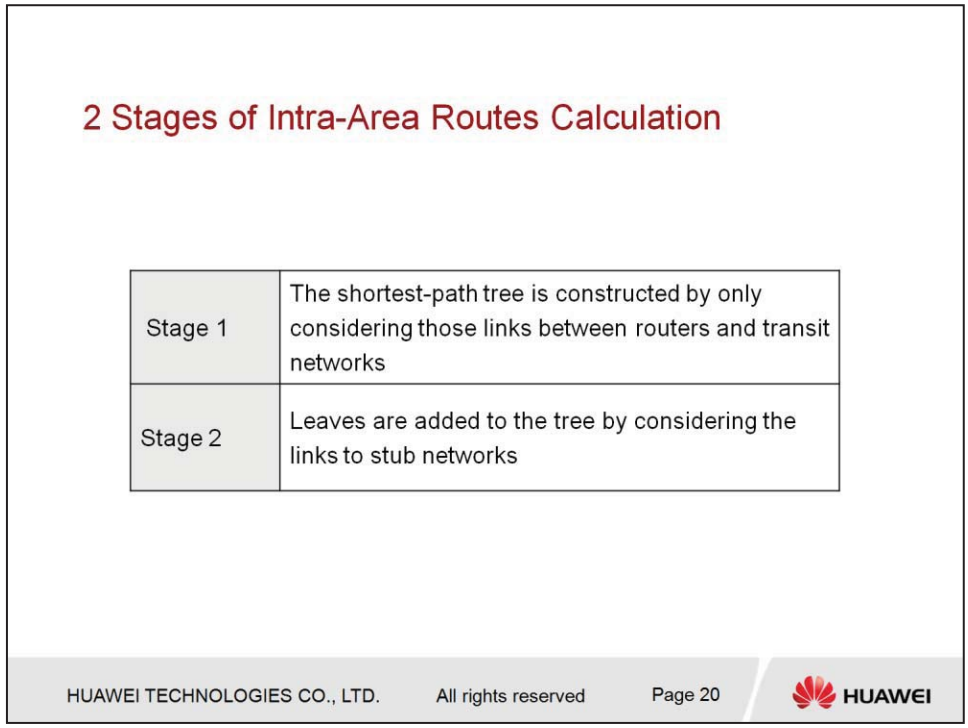
This is Area 1's LSDB.

In this LSDB, there are 4 Router-LSAs describing 4 routers' active link, one Network-LSA describing broadcast network.



This is the directed graph described by LSDB.

There are a transit network and 5 Loopback interfaces are described as stub network.

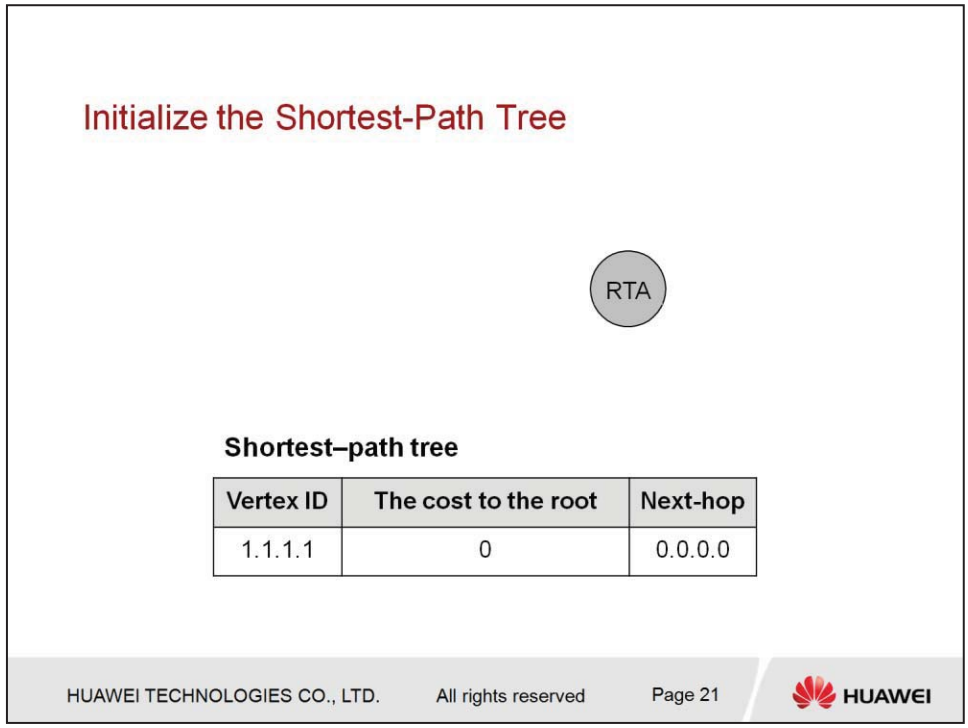


Each router calculates the shortest-path tree using itself as the root.

The formation of the shortest path tree is done here in two stages.

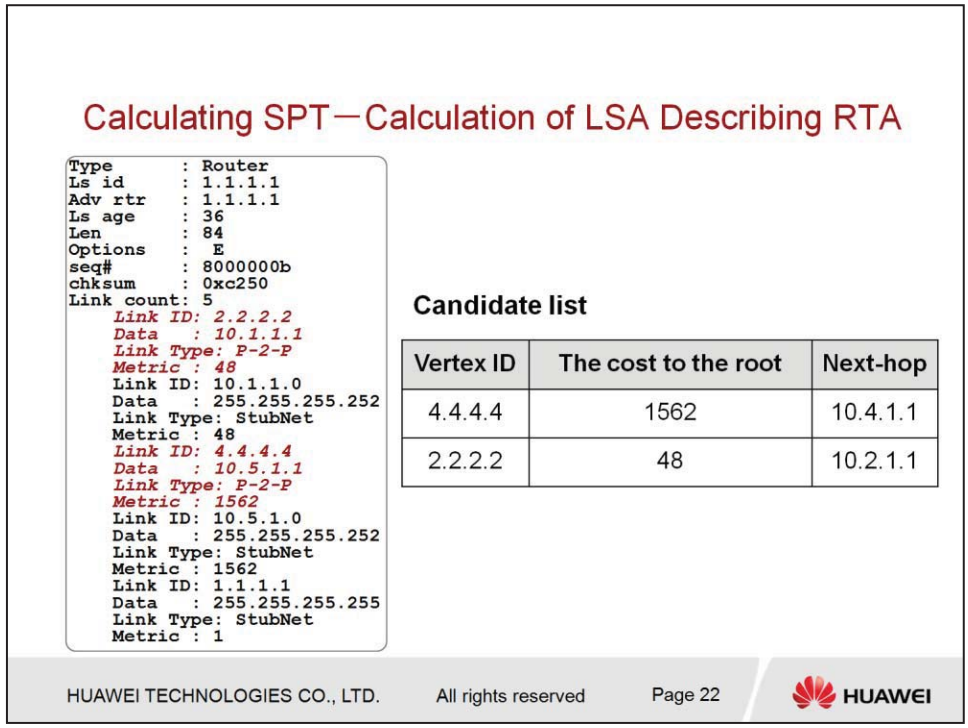
In the first stage, only links between routers and transit networks are considered.

In the second stage, leaves are added to the tree by considering the links to stub networks.



In this example, how RTA calculates the shortest-path tree regarding itself as the root is described in detail.

The first step is to initialize the shortest-path tree, RTA treat itself as the root, this root is added to the shortest-path tree.



After being added to the shortest-path tree, RTA examines Router-LSA originated by itself, each link described by Router-LSA except stub link is added to a list of candidate vertices, the vertex ID is Link ID, the cost to the root is the metric in LSA. In this example, the vertex 4.4.4.4 and 2.2.2.2 are added to the candidate list.

Calculating SPT—Calculation of LSA Describing RTA

Shortest-path tree

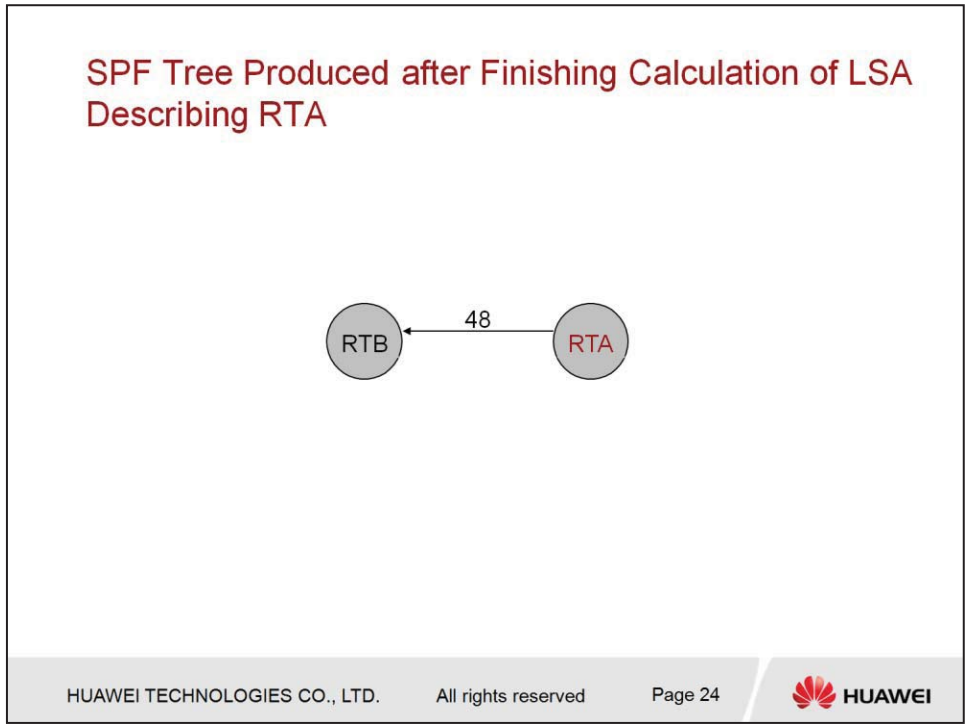
Vertex ID	The cost to the root	Next-hop
1.1.1.1	0	0.0.0.0
2.2.2.2	48	10.2.1.1

Candidate list

Vertex ID	The cost to the root	Next-hop
4.4.4.4	1562	10.4.1.1

HUAWEI TECHNOLOGIES CO., LTD. All rights reserved HUAWEI

The candidate vertex with the smallest cost is added to the shortest-path tree, removed from the candidate list. In this example, the vertex 2.2.2.2 is added to the shortest-path tree and removed from the candidate list.



This figure shows the current shortest-path tree described by the directed graph.

RTB is the newly added vertex.

Calculating SPT—Calculation of LSA Describing RTB

```
Type       : Router
Ls id      : 2.2.2.2
Adv rtr    : 2.2.2.2
Ls age     : 470
Len        : 72
Options    :  E
seq#       : 8000000e
chksum     : 0xbfa4
Link count: 4
   Link ID: 1.1.1.1
   Data   : 10.2.1.1
   Link Type: P-2-P
   Metric : 48
   Link ID: 10.2.1.0
   Data   : 255.255.255.252
   Link Type: StubNet
   Metric : 48
   Link ID: 10.3.1.1
   Data   : 10.3.1.1
   Link Type: TransNet
   Metric : 1
   Link ID: 2.2.2.2
   Data   : 255.255.255.255
   Link Type: StubNet
   Metric : 1
```

Candidate list

Vertex ID	The cost to the root	Next-hop
4.4.4.4	1562	10.4.1.1
10.3.1.1	48+1=49	10.2.1.1

HUAWEI TECHNOLOGIES CO., LTD. All rights reserved HUAWEI

If a new vertex is added to the shortest-path tree, examine the LSA whose LS ID is the new vertex's vertex ID. In this example, examine the LSA whose LS ID is 2.2.2.2.

If a Link ID described by the LSA already appears on the shortest-path tree, then this link is ignored. In this example, the link with Link ID 1.1.1.1 is ignored, only the vertex 10.3.1.1 is added to the candidate list.

The cost to the root is the sum of the metric of the link (here is 1) and the metric (here is 48) between parent vertex (here is 2.2.2.2)and the root.

Calculating SPT—Calculation of LSA Describing RTB

Shortest-path tree

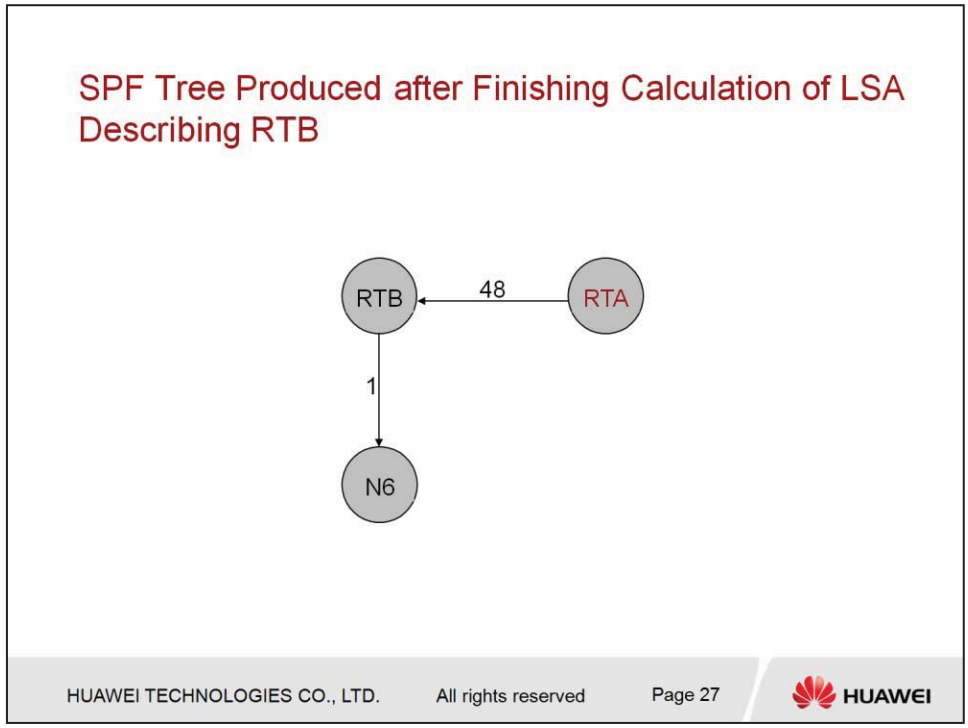
Vertex ID	The cost to the root	Next-hop
1.1.1.1	0	0.0.0.0
2.2.2.2	48	10.2.1.1
10.3.1.1	49	10.2.1.1

Candidate list

Vertex ID	The cost to the root	Next-hop
4.4.4.4	1562	10.4.1.1

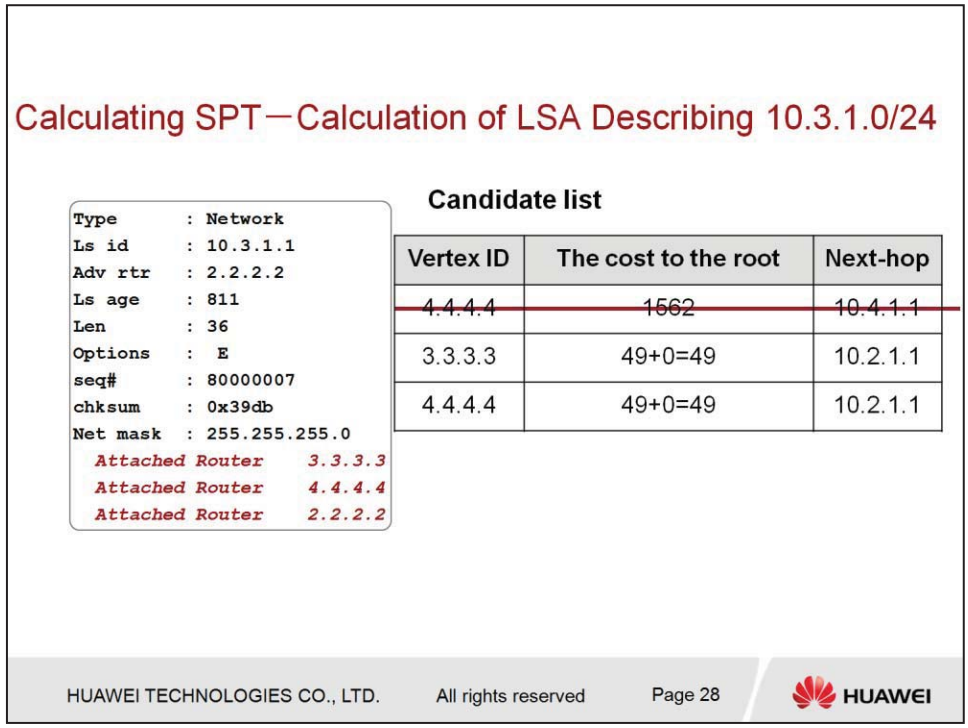
HUAWEI TECHNOLOGIES CO., LTD. All rights reserved HUAWEI

The candidate vertex whose cost to the root is smallest is added to the shortest path tree, removed from the candidate list. In this example, the vertex 10.3.1.1 is added to the shortest-path tree and removed from the candidate list.



This figure shows the current shortest-path tree described by the directed graph.

N6 is the newly added vertex.



Examine the LSA whose LS ID is the new vertex's vertex ID. In this example, examine the LSA whose LS ID is 10.3.1.1.

The link with Link ID is 2.2.2.2, the vertex 3.3.3.3 and 4.4.4.4 are added to candidate list.

The cost from transit network to attached router is 0.

If the candidate list has two same vertex ID with different cost to the root, the vertex with higher cost to the root is removed form the candidate list.

Calculating SPT—Calculation of LSA Describing 10.3.1.0/24

Shortest-path tree

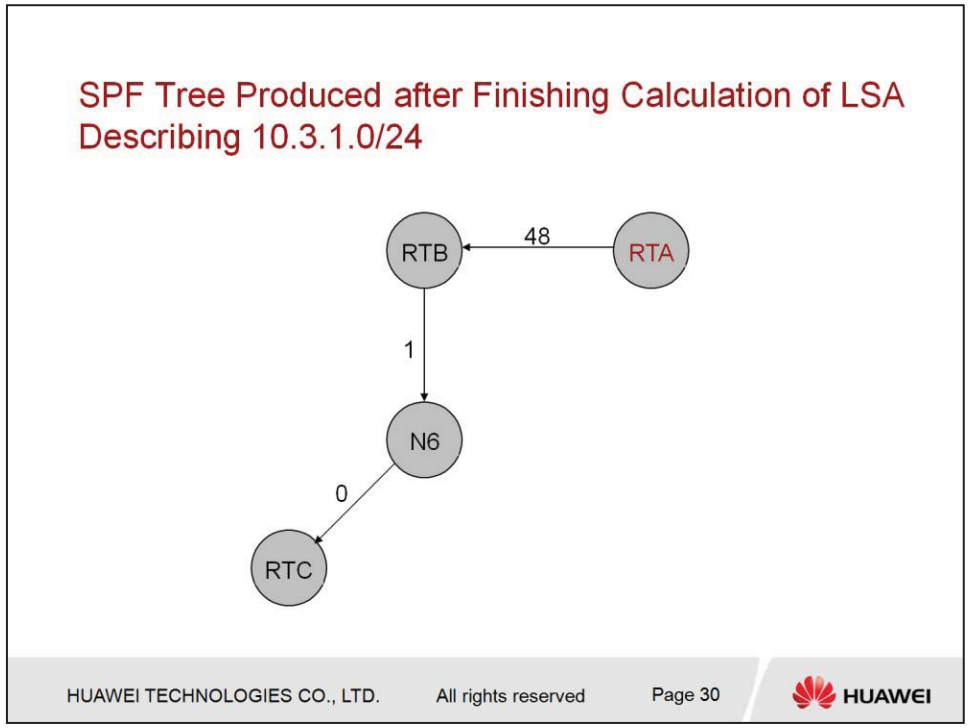
Vertex ID	The cost to the root	Next-hop
1.1.1.1	0	0.0.0.0
2.2.2.2	48	10.2.1.1
10.3.1.1	49	10.2.1.1
3.3.3.3	49	10.2.1.1

Candidate list

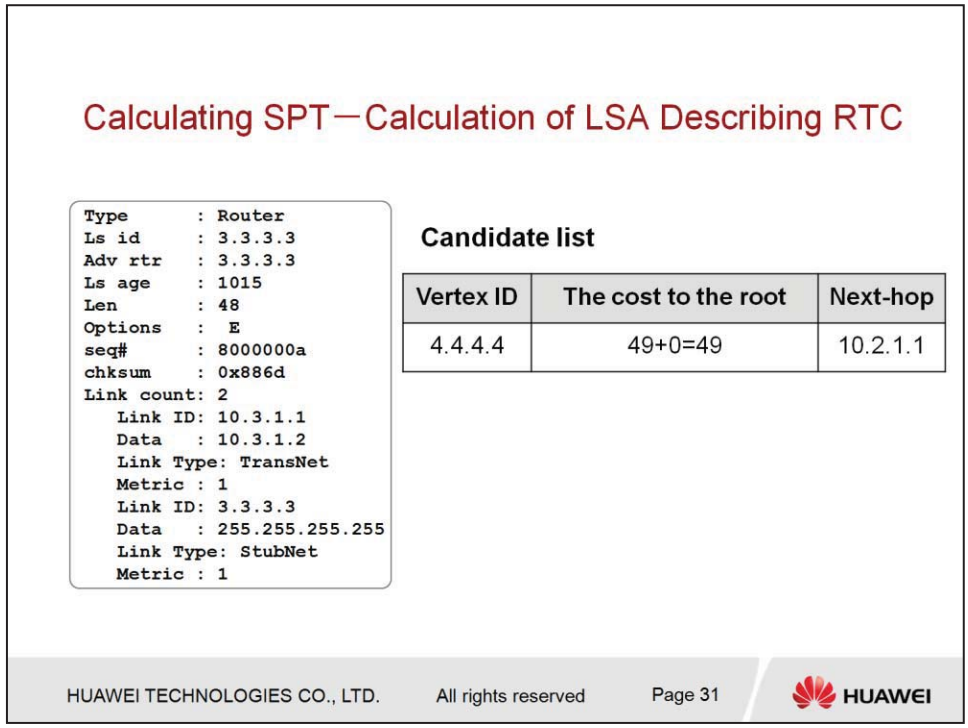
Vertex ID	The cost to the root	Next-hop
4.4.4.4	49+0=49	10.2.1.1

HUAWEI TECHNOLOGIES CO., LTD. All rights reserved HUAWEI

The candidate vertex with the smallest cost to the root is added to the shortestpath tree. In this example, the vertex 3.3.3.3 is added to the shortest-path tree.



This figure shows the shortest-path tree described by the directed graph. RTC is the newly added vertex.



Examine the LSA whose LS ID is the new vertex's vertex ID. In this example, examine the LSA whose LS ID is 3.3.3.3, no new vertex is added to the candidate list.

Calculating SPT—Calculation of LSA Describing RTC

Shortest-path tree

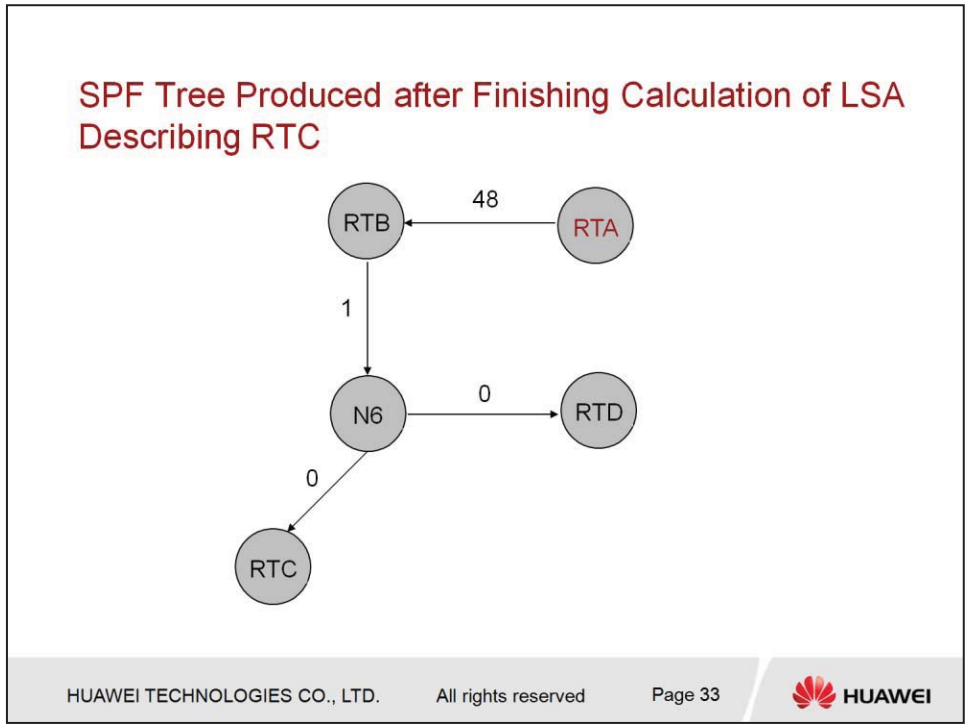
Vertex ID	The cost to the root	Next-hop
1.1.1.1	0	0.0.0.0
2.2.2.2	48	10.2.1.1
10.3.1.1	49	10.2.1.1
3.3.3.3	49	10.2.1.1
4.4.4.4	49	10.2.1.1

Candidate list

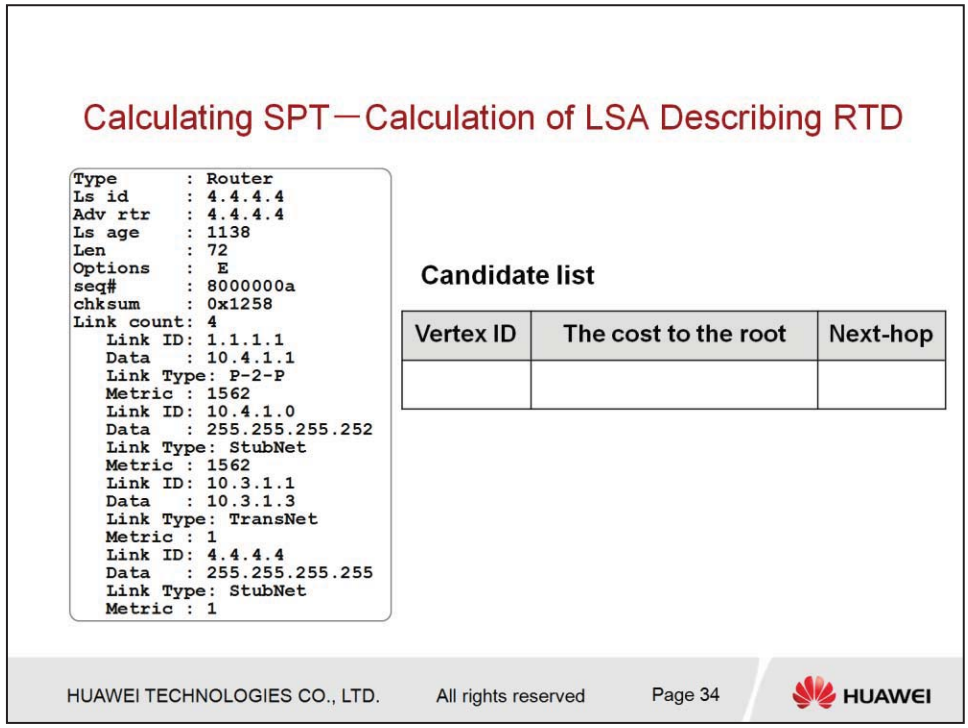
Vertex ID	The cost to the root	Next-hop

HUAWEI TECHNOLOGIES CO., LTD. All rights reserved

The candidate vertex with the smallest cost to the root is added to the shortest path tree. In this example, the vertex 4.4.4.4 is added to the shortest-path tree.



This figure shows the current shortest-path tree described by directed graph. RTD is the newly added vertex.



Examine the LSA whose LS ID is the new vertex's vertex ID. In this example, examine the LSA whose LS ID is 4.4.4.4, no new vertex is added to the candidate list.

If the candidate at this step list is empty, the shortest-path tree has been completely built and this stage of the procedure terminates.

Calculating SPT—Calculation of Stub Network RTA Connects to

```
Type       : Router
Ls id      : 1.1.1.1
Adv rtr    : 1.1.1.1
Ls age     : 36
Len        : 84
Options    :  E
seq#       : 8000000b
chksum     : 0xc250
Link count: 5
    Link ID: 2.2.2.2
    Data   : 10.1.1.1
    Link Type: P-2-P
    Metric : 48
    Link ID: 10.1.1.0
    Data   : 255.255.255.252
    Link Type: StubNet
    Metric : 48
    Link ID: 4.4.4.4
    Data   : 10.5.1.1
    Link Type: P-2-P
    Metric : 1562
    Link ID: 10.5.1.0
    Data   : 255.255.255.252
    Link Type: StubNet
    Metric : 1562
    Link ID: 1.1.1.1
    Data   : 255.255.255.255
    Link Type: StubNet
    Metric : 1
```

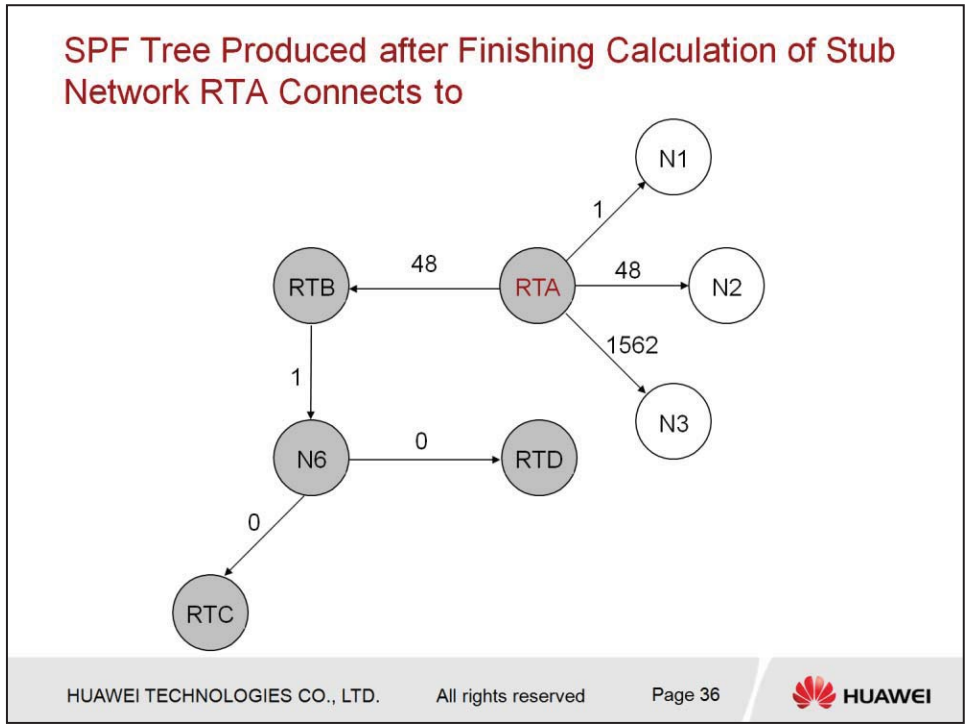
Stub network

Vertex ID	The cost to the root	Next-hop
10.5.1.0	1562+0=1562	0.0.0.0
10.1.1.0	48+0=48	0.0.0.0
1.1.1.1	1+0=1	0.0.0.0

HUAWEI TECHNOLOGIES CO., LTD. All rights reserved HUAWEI

The stub network are added to the tree in the procedure's second stage. In this stage, the Router-LSA of every router's vertex is examined, then stub network link appearing in the LSA is examined.

In this example, RTA's Router-LSA is first examined, three stub networks are included in the LSA.



This figure shows current shortest-path tree described by the directed graph,three stub networks N1, N2 and N3 are added to the tree.

Calculating SPT—Calculation of Stub Network RTB Connects to

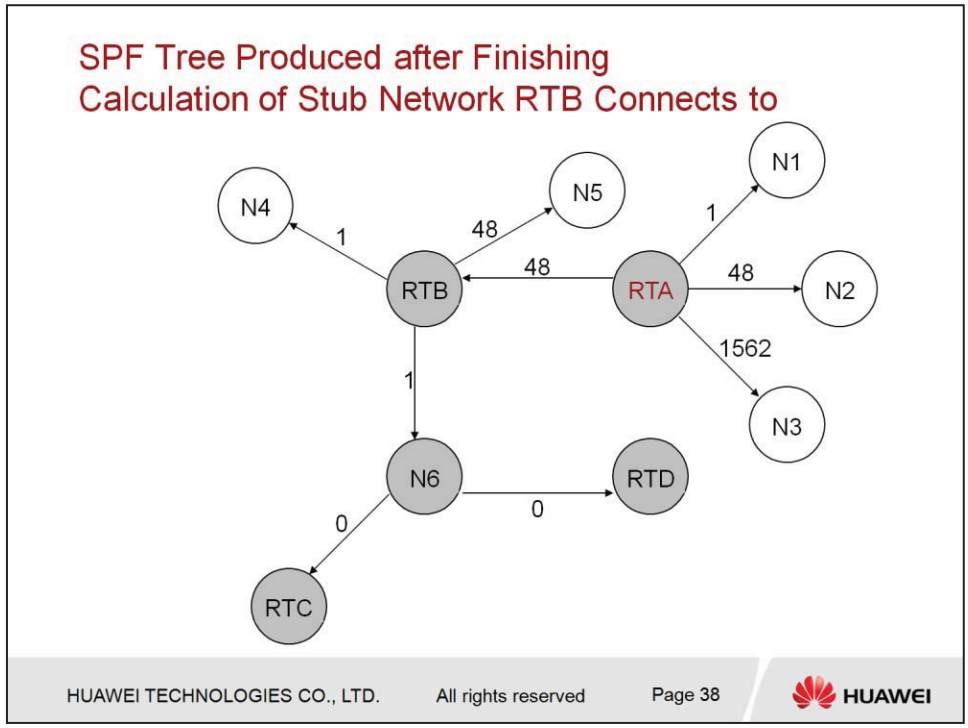
```
Type       : Router
Ls id      : 2.2.2.2
Adv rtr    : 2.2.2.2
Ls age     : 470
Len        : 72
Options    :  E
seq#       : 8000000e
chksum     : 0xbfa4
Link count: 4
   Link ID: 1.1.1.1
   Data   : 10.2.1.1
   Link Type: P-2-P
   Metric : 48
   Link ID: 10.2.1.0
   Data   : 255.255.255.252
   Link Type: StubNet
   Metric : 48
   Link ID: 10.3.1.1
   Data   : 10.3.1.1
   Link Type: TransNet
   Metric : 1
   Link ID: 2.2.2.2
   Data   : 255.255.255.255
   Link Type: StubNet
   Metric : 1
```

Stub network

Vertex ID	The cost to the root	Next-hop
10.5.1.0	1562+0=1562	0.0.0.0
10.1.1.0	48+0=48	0.0.0.0
1.1.1.1	1+0=1	0.0.0.0
2.2.2.2	48+1=49	10.2.1.1
10.2.1.0	48+48=96	10.2.1.1

HUAWEI TECHNOLOGIES CO., LTD. All rights reserved HUAWEI

Examine RTB's Router-LSA, two stub networks are included in the LSA.



This figure shows the shortest-path tree described by the directed graph, two stub networks N4, N5 are added to the tree.

Calculating SPT—Calculation of Stub Network RTC Connects to

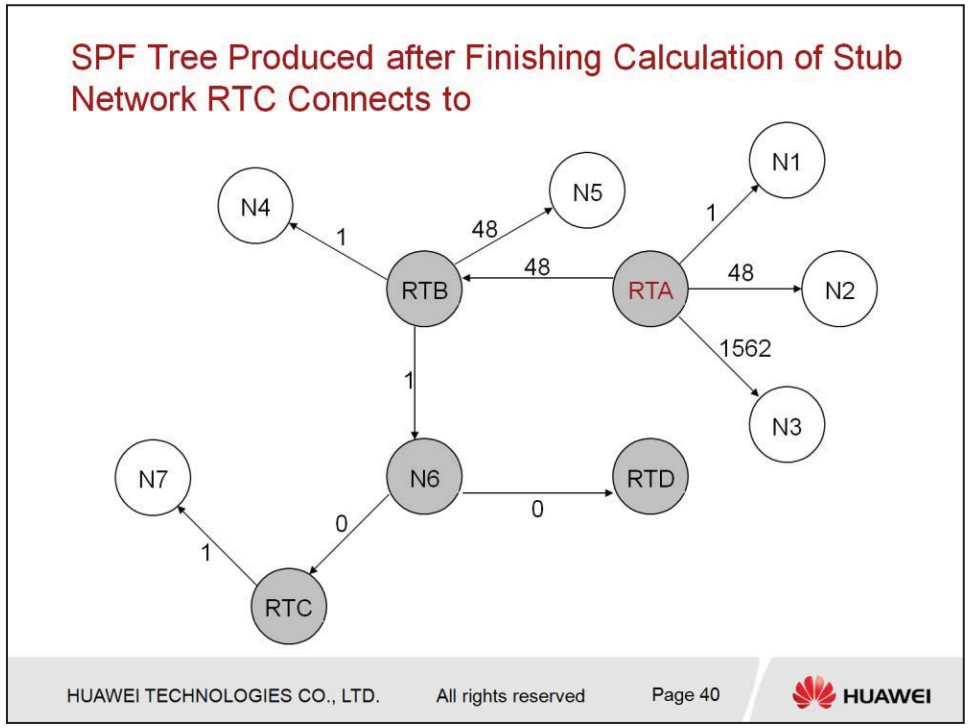
```
Type       : Router
Ls id      : 3.3.3.3
Adv rtr    : 3.3.3.3
Ls age     : 1015
Len        : 48
Options    :  E
seq#       : 8000000a
chksum     : 0x886d
Link count: 2
   Link ID: 10.3.1.1
   Data   : 10.3.1.2
   Link Type: TransNet
   Metric : 1
   Link ID: 3.3.3.3
   Data   : 255.255.255.255
   Link Type: StubNet
   Metric : 1
```

Stub network

Vertex ID	The cost to the root	Next-hop
10.5.1.0	1562+0=1562	0.0.0.0
10.1.1.0	48+0=48	0.0.0.0
1.1.1.1	1+0=1	0.0.0.0
2.2.2.2	48+1=49	10.2.1.1
10.2.1.0	48+48=96	10.2.1.1
3.3.3.3	49+1=50	10.2.1.1

HUAWEI TECHNOLOGIES CO., LTD. All rights reserved HUAWEI

Examine RTC's Router-LSA, only stub network is included in the LSA.



This figure shows the current shortest-path tree described by the directed graph, stub network N7 is added to the tree.

Calculating SPT—Calculation of Stub Network RTD Connects to

```
Type       : Router
Ls id      : 4.4.4.4
Adv rtr    : 4.4.4.4
Ls age     : 1138
Len        : 72
Options    :  E
seq#       : 8000000a
chksum     : 0x1258
Link count: 4
   Link ID: 1.1.1.1
   Data   : 10.4.1.1
   Link Type: P-2-P
   Metric : 1562
   Link ID: 10.4.1.0
   Data   : 255.255.255.252
   Link Type: StubNet
   Metric : 1562
   Link ID: 10.3.1.1
   Data   : 10.3.1.3
   Link Type: TransNet
   Metric : 1
   Link ID: 4.4.4.4
   Data   : 255.255.255.255
   Link Type: StubNet
   Metric : 1
```

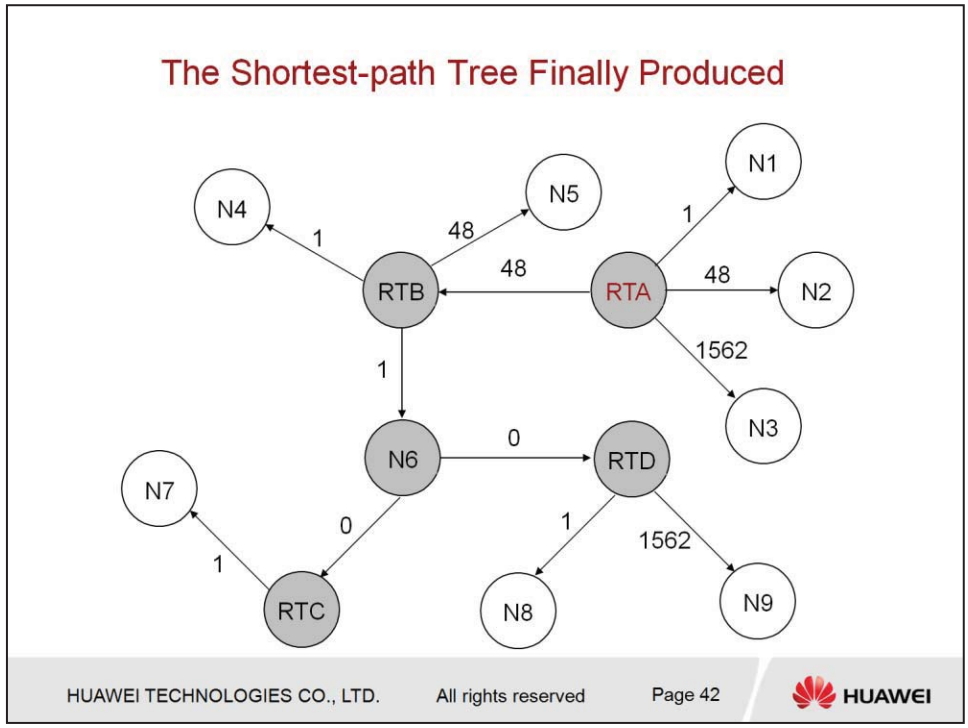
Stub network

Vertex ID	The cost to the root	Next-hop
10.5.1.0	1562+0=1562	0.0.0.0
10.1.1.0	48+0=48	0.0.0.0
1.1.1.1	1+0=1	0.0.0.0
2.2.2.2	48+1=49	10.2.1.1
10.2.1.0	48+48=96	10.2.1.1
3.3.3.3	49+1=50	10.2.1.1
4.4.4.4	49+1=50	10.2.1.1
10.4.1.0	49+1562=1611	10.2.1.1

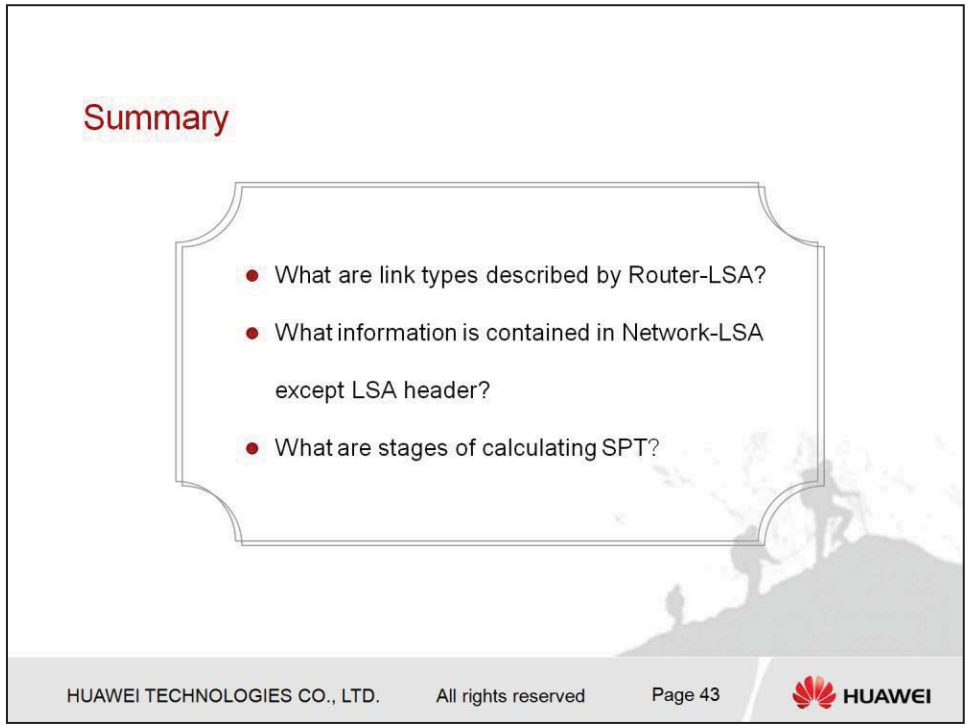
HUAWEI TECHNOLOGIES CO., LTD. All rights reserved HUAWEI

Examine RTD's Router-LSA, two stub networks are included in the LSA.

The shortest-path tree and stub network table show the routes destined for all networks.



This figure shows the final shortest-path tree described by the directed graph, two stub network N8, N9 are added to the tree.



What are link types described by Router-LSA?

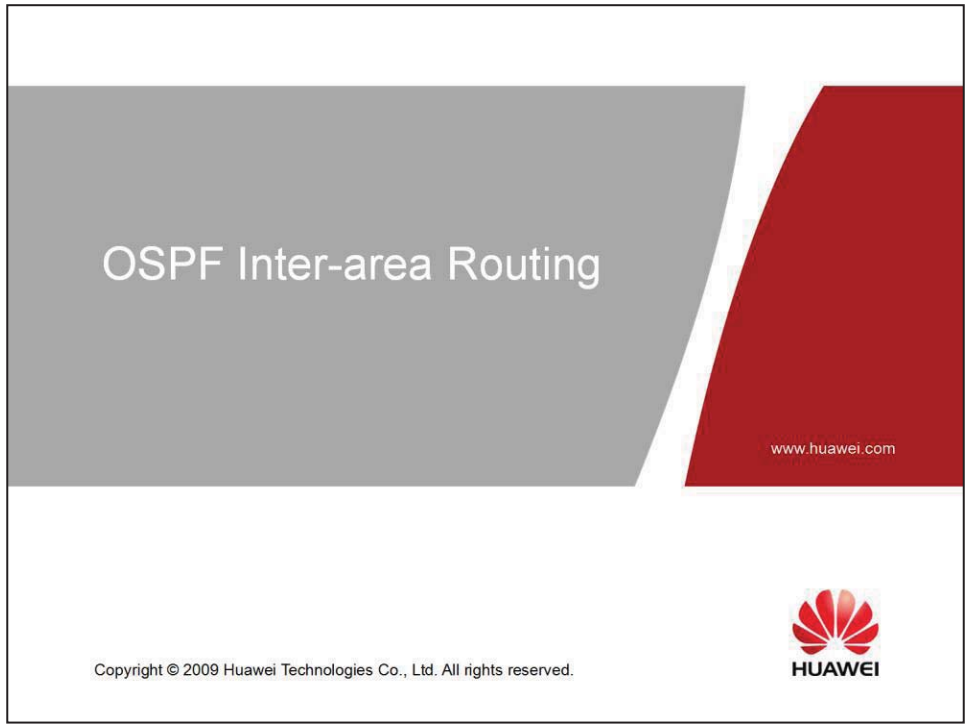
Four types: point-to-point, transit, stub and virtual link.

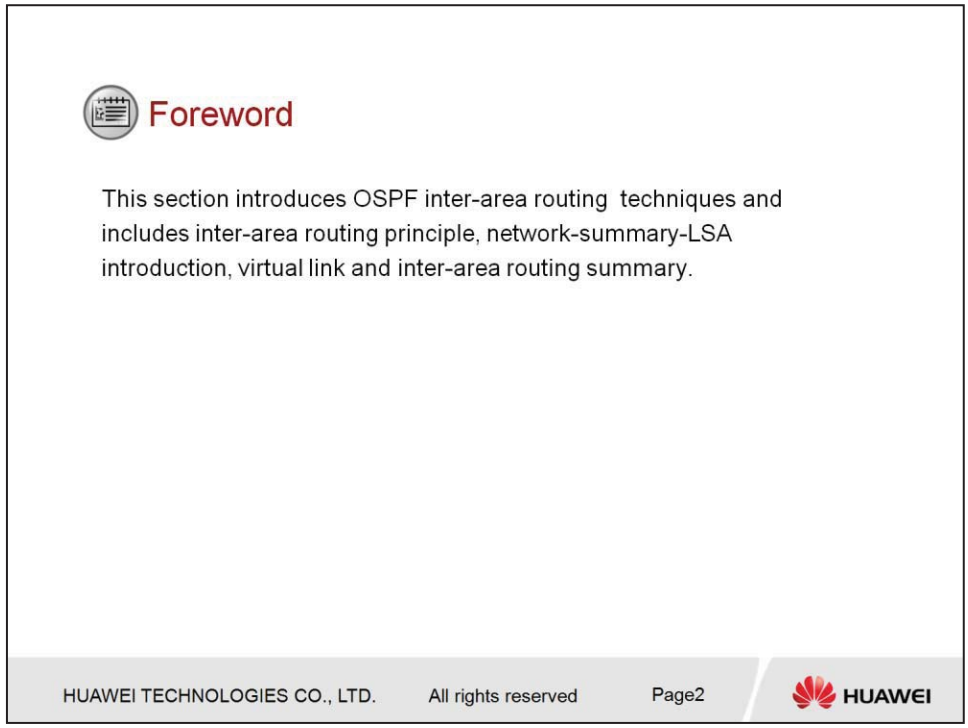
What information is contained in Network-LSA except LSA header?

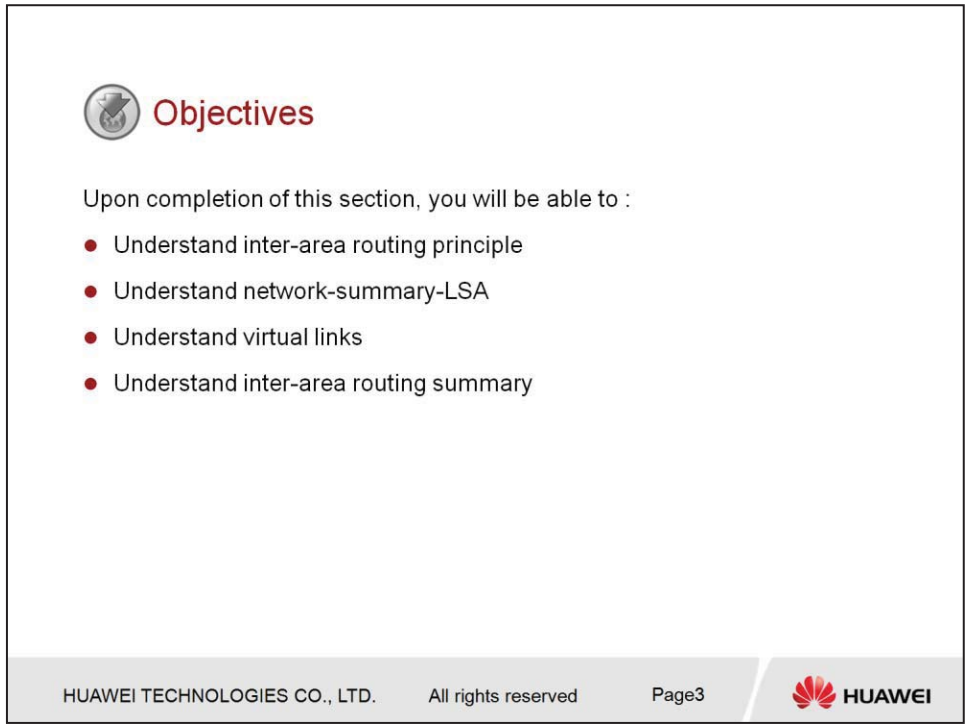
The network address mask and list of routers attached to this network.

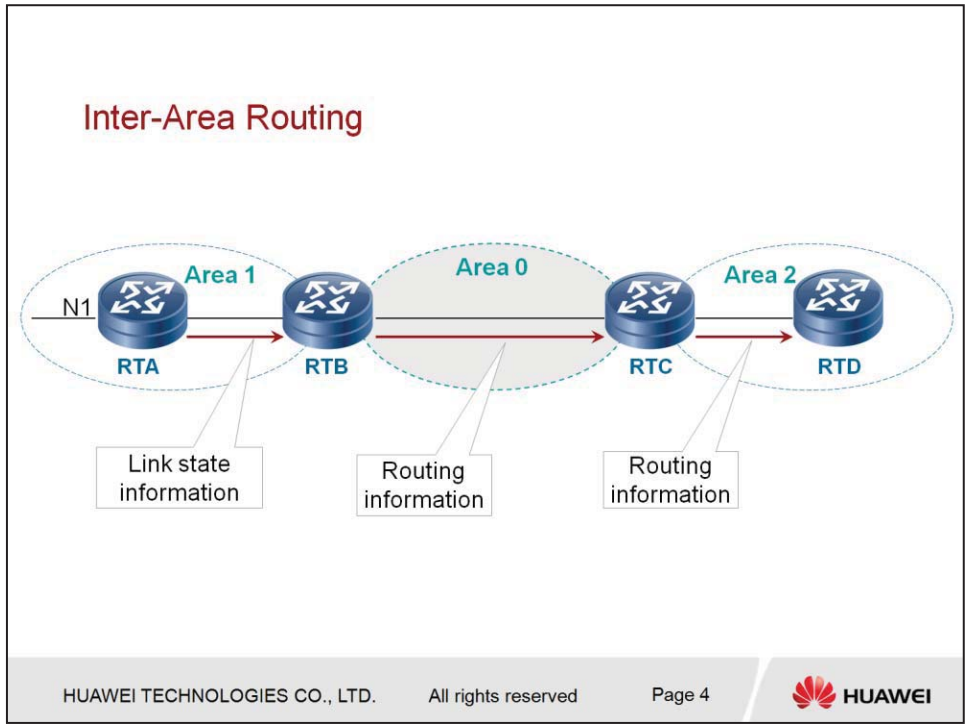
What are stages of calculating SPT?

Two stages: the first stage is to calculate links between routers and transit networks, the second stage is to calculate the links to stub networks.


OSPF Inter-area Routing
www.huawei.com
Copyright © 2009 Huawei Technologies Co., Ltd. All rights reserved.
HUAWEI


Foreword
This section introduces OSPF inter-area routing techniques and includes inter-area routing principle, network-summary-LSA introduction, virtual link and inter-area routing summary.
HUAWEI TECHNOLOGIES CO., LTD.
All rights reserved
Page2
HUAWEI


Objectives
Upon completion of this section, you will be able to :
Understand inter-area routing principle
Understand network-summary-LSA
Understand virtual links
Understand inter-area routing summary
HUAWEI TECHNOLOGIES CO., LTD.
All rights reserved
Page3
HUAWEI



There are several LSDBs on Area Border Router (ABR) , ABR maintains a separate LSDB for each area .

ABR gets the routing information from the non-backbone area's link state information , and advertises them into backbone area. So backbone area advertises it to the other non-backbone areas.

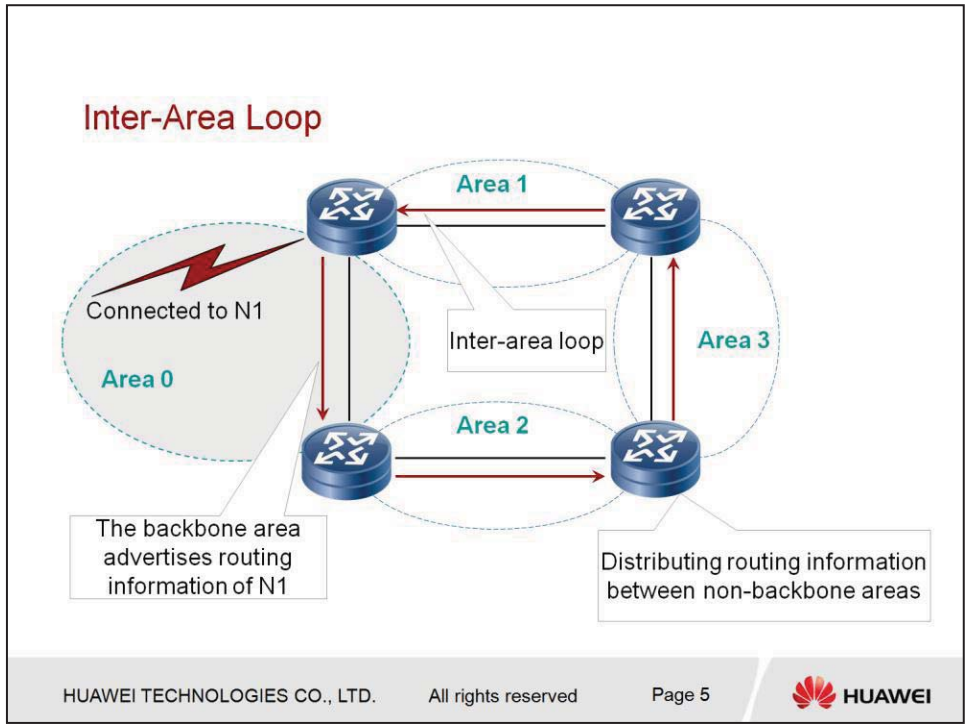
ABR needs to get backbone area's link state information and calculate routes about backbone based on LSDB, then advertises it to all non-backbone areas .

In this case:

RTA generates the link state information about N1 , and flood it to RTB .

RTB receives the routing information about N1 , and flood it into backbone area .

RTC receives the routing information ,and flood it to RTD .

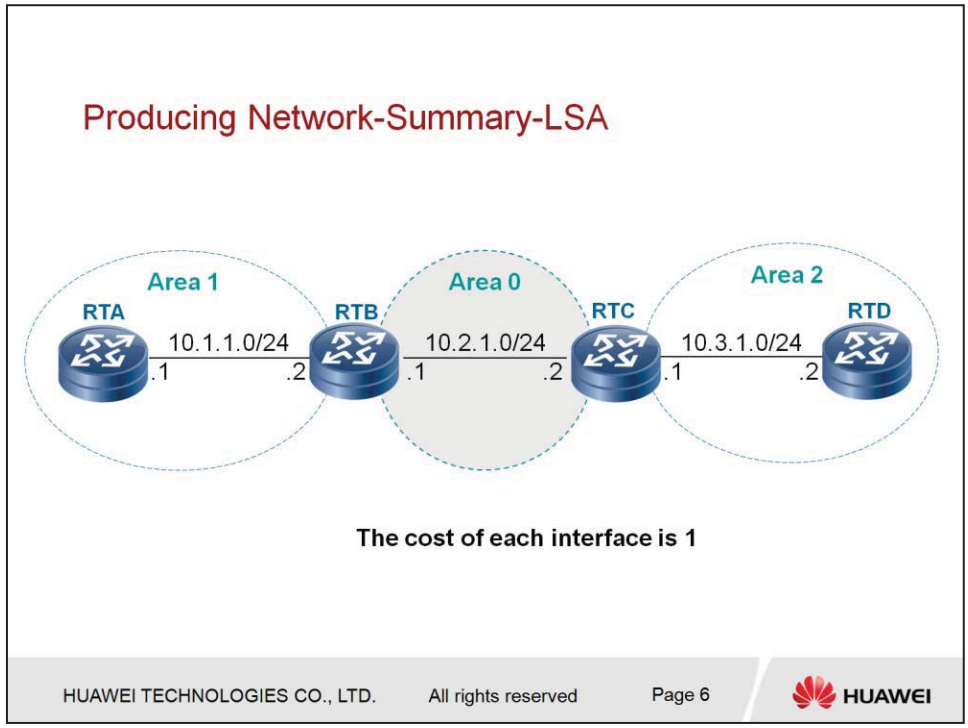


In this case :

If sending routing information between Area2 and Area3 is allowed , inter-area loop might happen.

In order to avoid loop between areas , It is strictly prohibited to exchange the routing information between non-backbone areas . Only send routing information within one area or between backbone area and non-backbone areas is allowed .

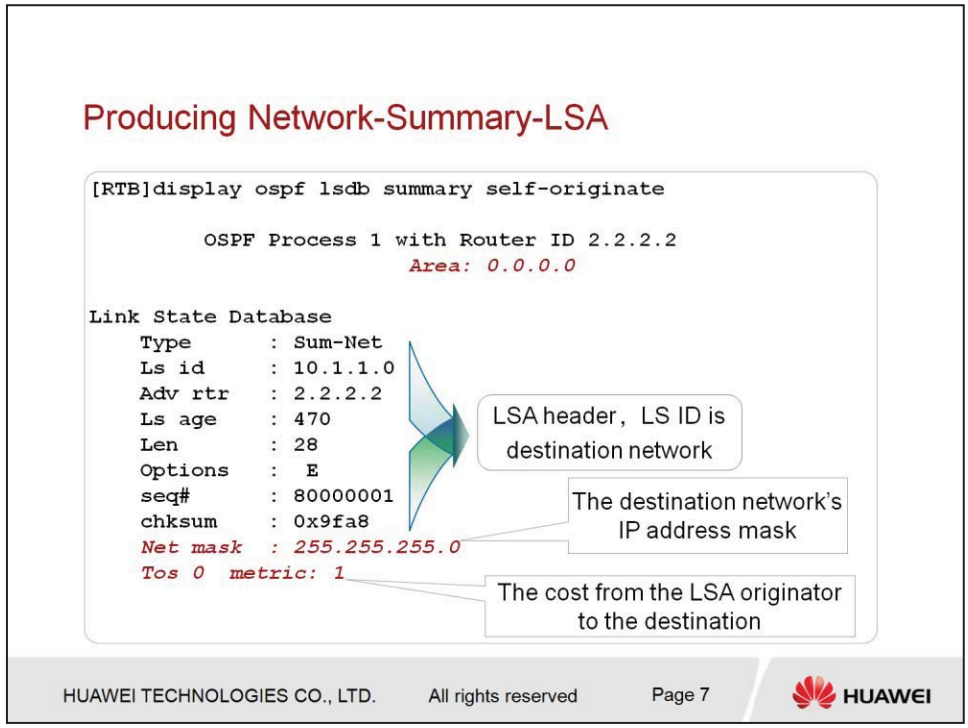
Thus, each ABR must connect to backbone area directly .



In this case:

RTA ’s router id is 1.1.1.1 , RTB is 2.2.2.2 , RTC is 3.3.3.3 , RTD is 4.4.4.4 .

All port’s cost is 1 .



Network-Summary-LSA contains the information:

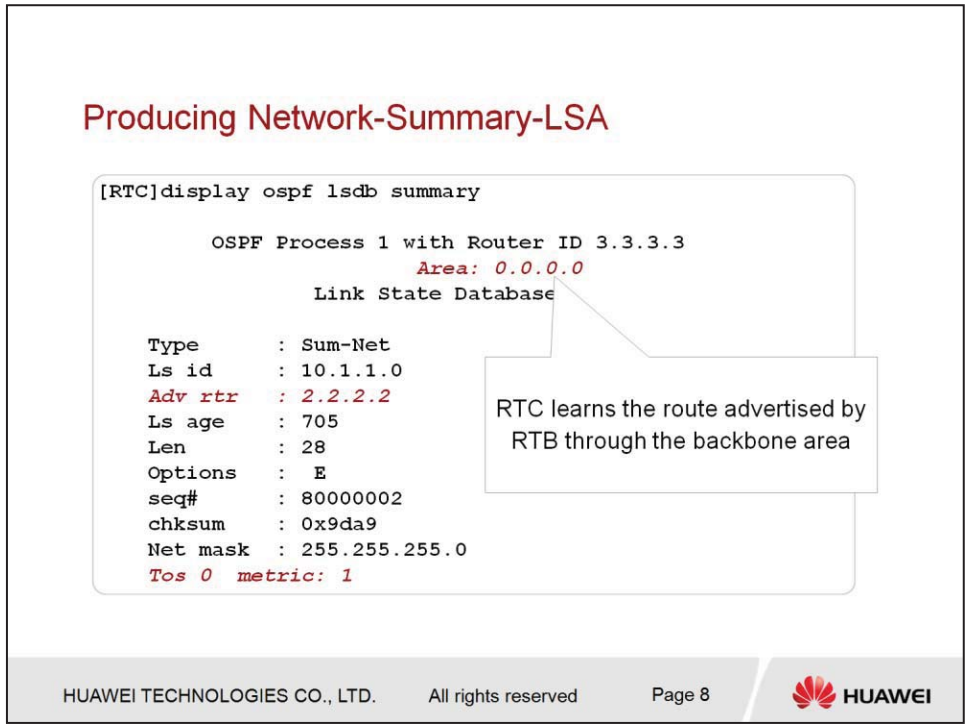
Link State ID stands for the destination IP address.

Net mask stands for the destination mask.

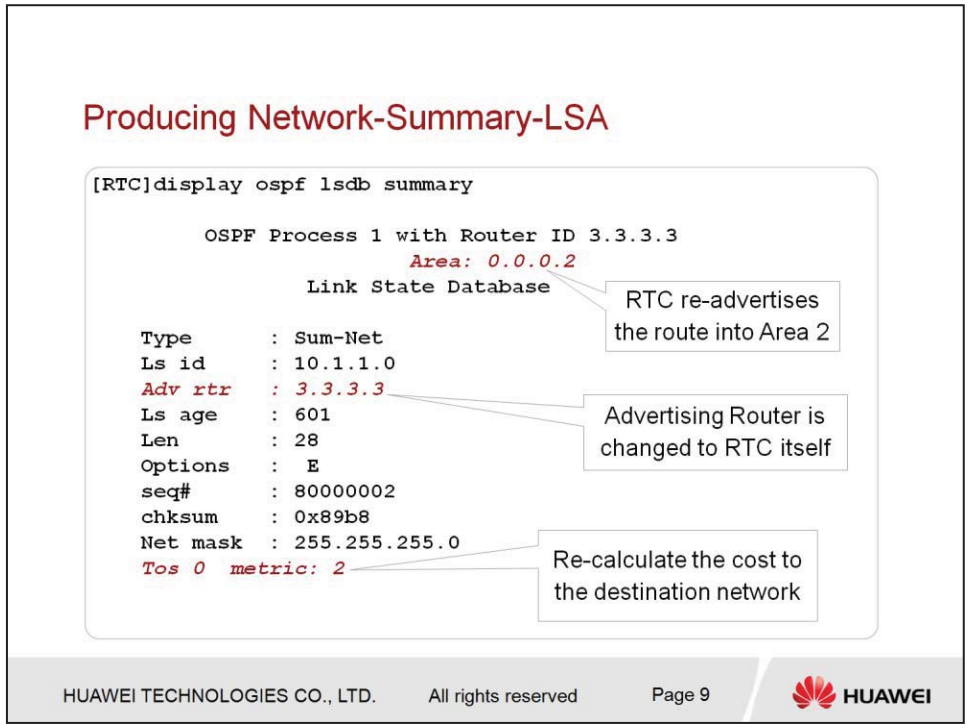
Metric stands for the cost from this ABR to destination network .

Take network 10.1.1.0/24 for instance , procedure of routing advertisement between areas act as follow :

First , RTB(ABR of area 1) advertise the routing information of this network into backbone area .



Then , RTC learned the routing information of 10.1.1.0/24 from RTB through backbone area.

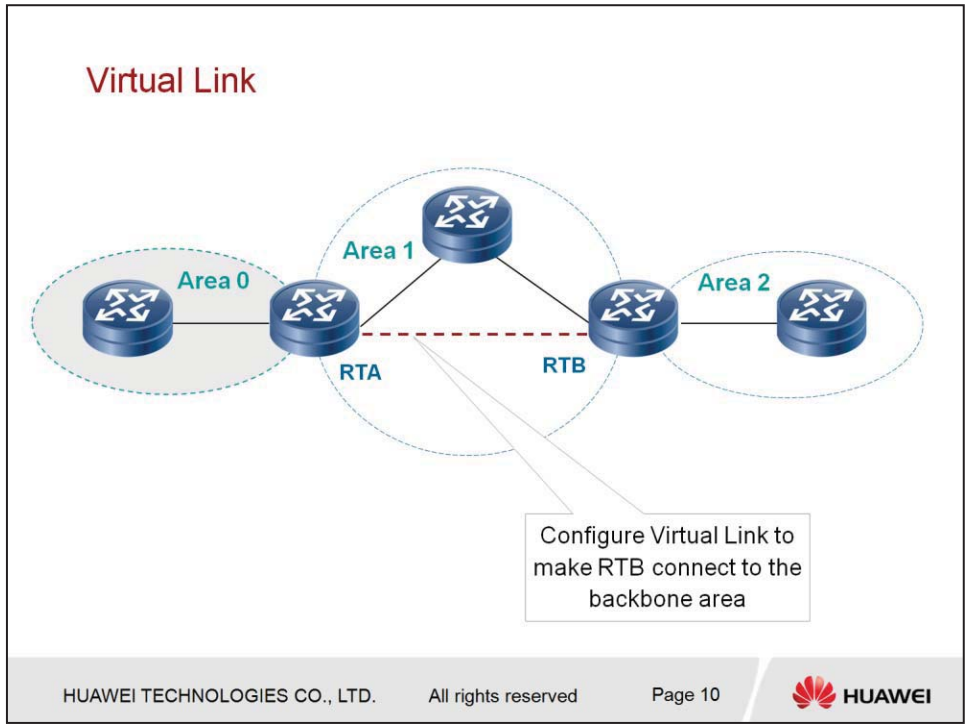


At last , according to Network-Summary-LSA which has learned from backbone area , RTC generated a new Network-Summary-LSA ,and flood that into area2 .

in this new LSA :

Advertising Router changes to RTC 's Router ID .

The cost to destination must be recalculate , change the cost into the sum cost from RTC to destination .

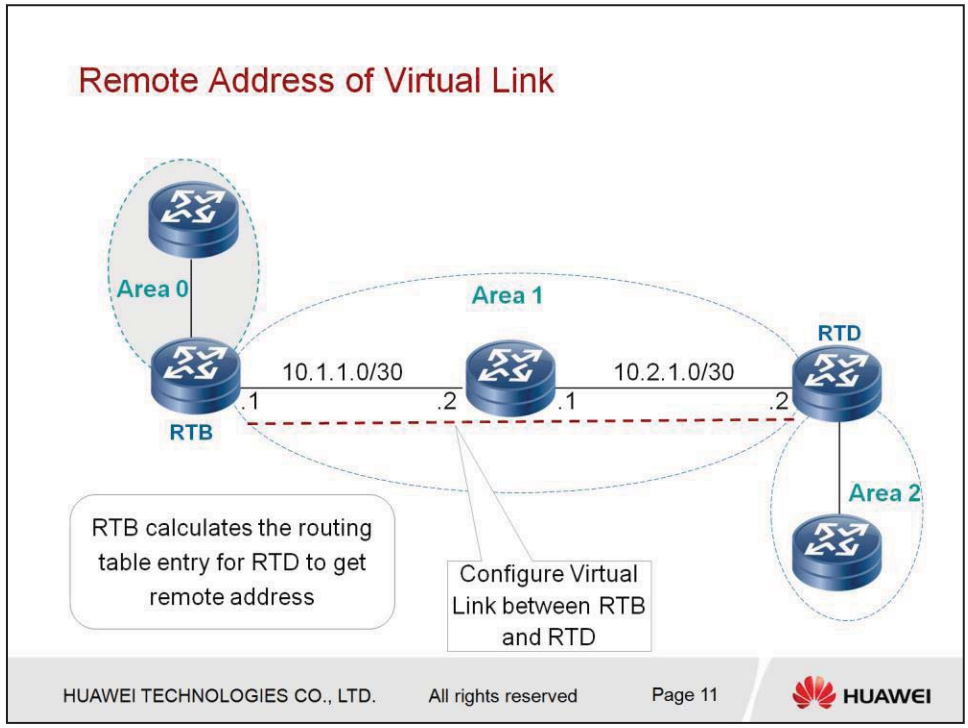


Backbone area must be continuously , but it doesn't need to make a physical continuous connection, virtual link can be used to make a logical continuous connection .

Virtual link can be built up between two discretional ABRs , each ABRs must have one port connected to the same non-backbone area .

Virtual link belongs to backbone area (Area 0)

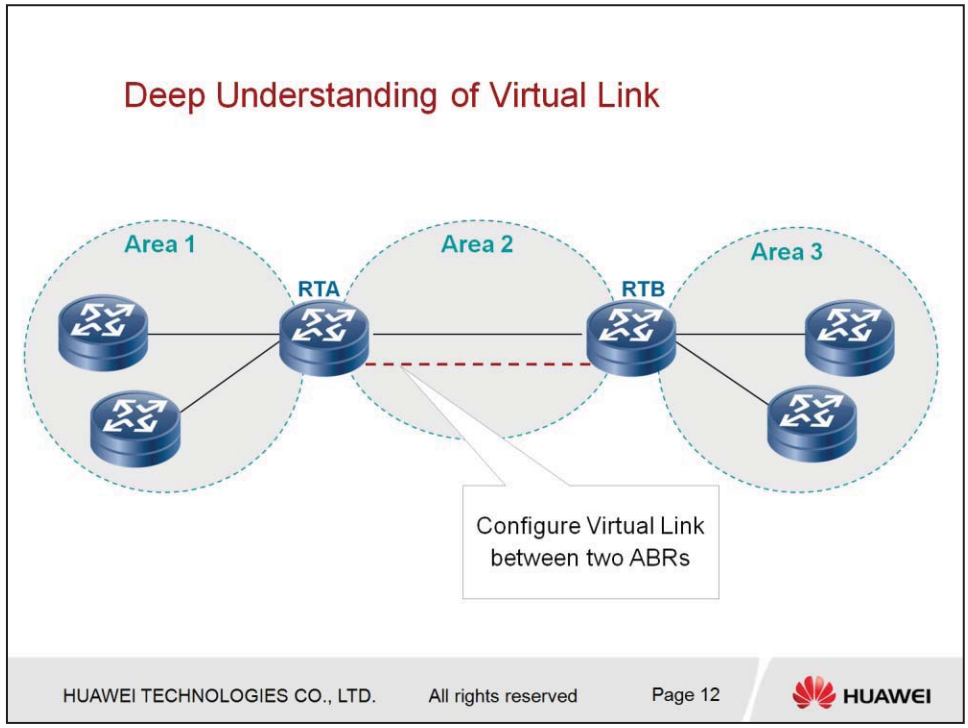
In this case , between RTA and RTB a virtual link is built up , which makes RTB connect to backbone area .



Two endpoints of Virtual link have to exchange protocol packets , but neighborhood of virtual link is identified by router id , which can't act as the destination address of protocol packet . How to identify the destination IP address with protocol packet ?

Each endpoint of virtual link has to calculate two shortest-path tree , as shown on the picture , RTB which is the endpoint of virtual link , has to calculate shortest path tree of area1 with root itself. On the other hand, RTB has to calculate the shortest-path tree in area 1 which regard RTD as its root (the shortest path tree of virtual link neighbor) . So as RTD .

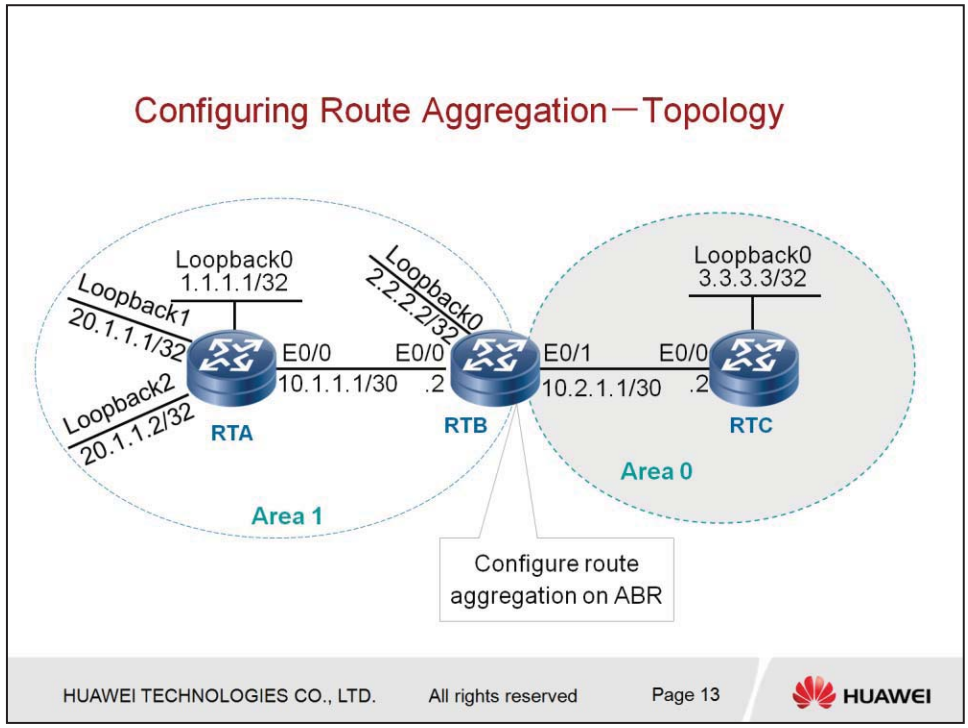
After calculating the shortest-path tree of virtual link , search how to reach to local router on the shortest path tree (use router id as it's identifier) , the IP address on virtual link neighbor which connected to local area router is the protocol packet destination IP address .



In this case, there is no backbone area, its area ID is 0. How do these three areas exchange their routing information each other?

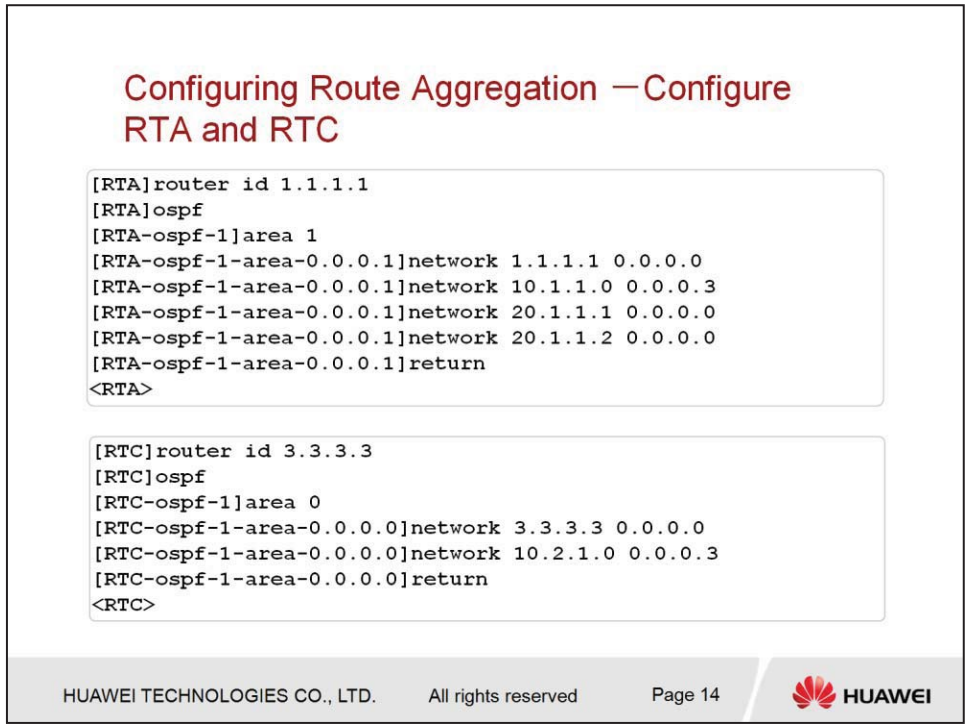
This problem is done by configuring a virtual link between two ABRs, RTA and RTB.

Virtual link is a part of backbone area , all the virtual links belong to area 0 .



In this case :

On RTB , after configuring route aggregation, routing information 20.1.1.1/32 and 20.1.1.2/32 can be aggregated into 20.1.1.0/24. When RTB advertises routing information to backbone area by Network-Summary-LSA , only 20.1.1.0/24 which has already been aggregated is advertised out and detail routing information is not advertised.

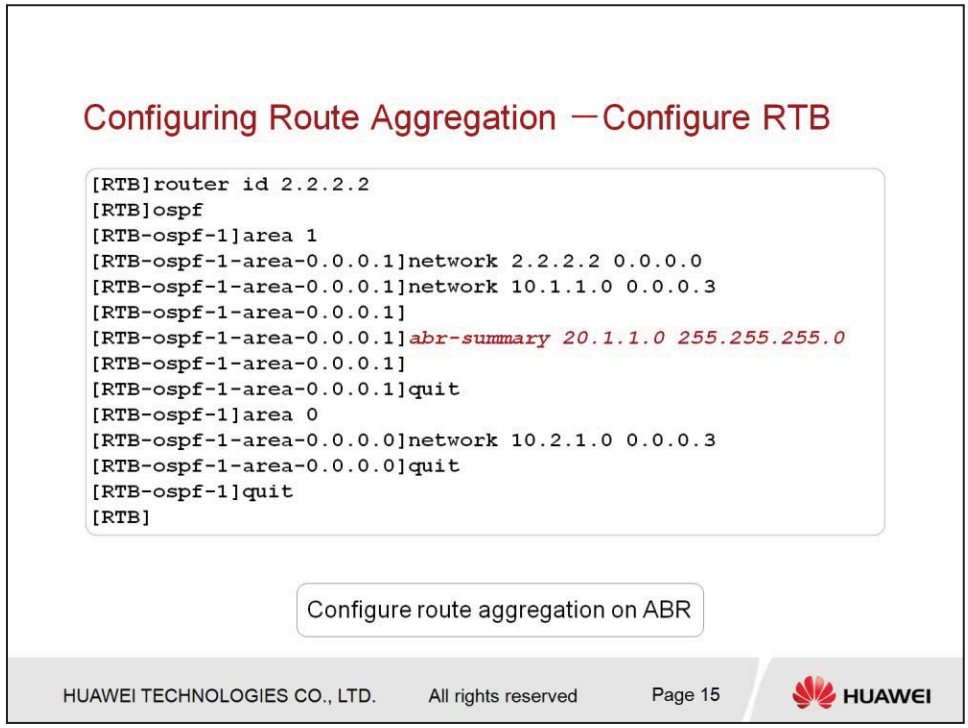


Configure these four networks of RTA into area 1 .

Don’t configure route aggregation on RTA , because configure route aggregation on RTA is meaningless , between RTA and RTB detail link state information is advertised , but not routing information , moreover RTA can’t generate Network-

Summary-LSA.

Configure these two networks of RTC into area 0 .



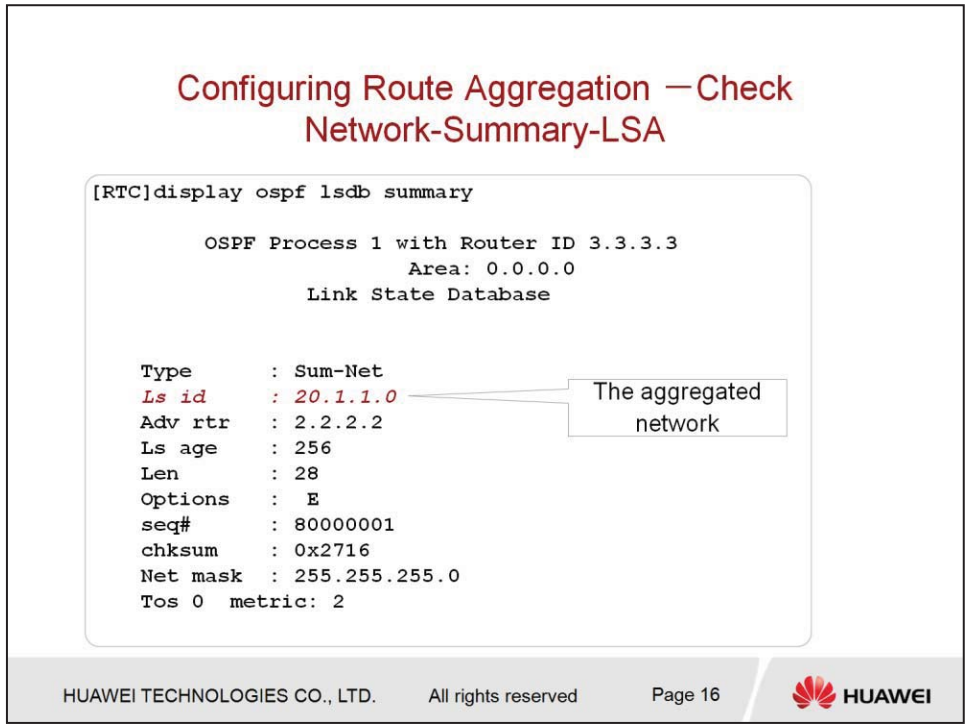
Configure route aggregation in RTB' s area view (the area that detail routing information is generated)

abr-summary ip-address mask [advertise | not-advertise]

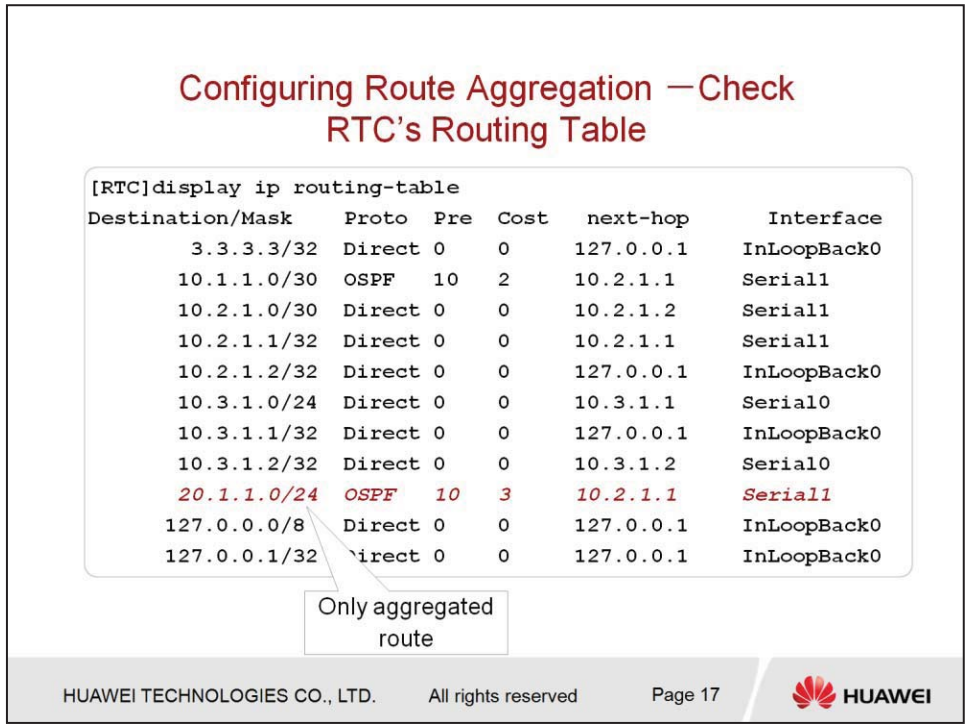
advertise：Advertises the summarized route or not.

Not-advertise： Not advertises the summarized route.

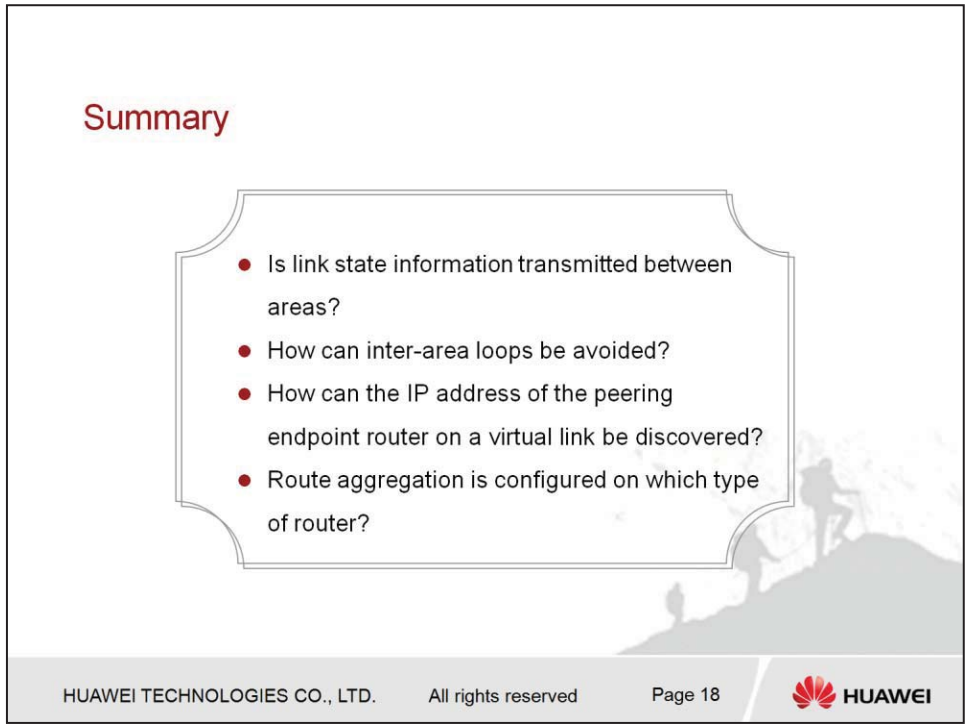
By default ,only aggregated route can be advertised .



When advertising the area1 routing information into backbone area , RTB uses a Network-Summary-LSA to describe the aggregated routing information 20.1.1.0/24 of area1, network 20.1.1.1/32 and 20.1.1.2/32 can not be flooded by Network-Summary-LSA .



On RTC's routing table , only aggregated routing information 20.1.1.0/24 exists, no detail routing information, 20.1.1.1/32 and 20.1.1.2/32 .



Is it link state information that is transmitted between areas ?

No , the one transmitted between areas is routing information ,but not detail link state information.

How to avoid inter-area loop ?

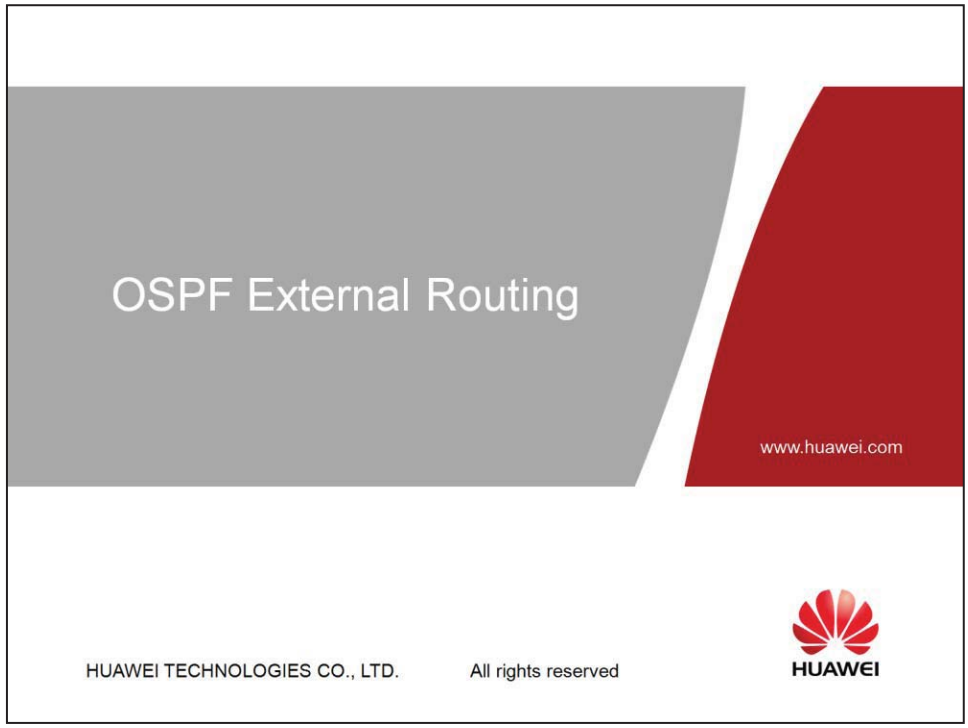
Only exchange routing information between backbone and non-backbone is allowed , exchange routing information between non-backbone areas is prohibited.

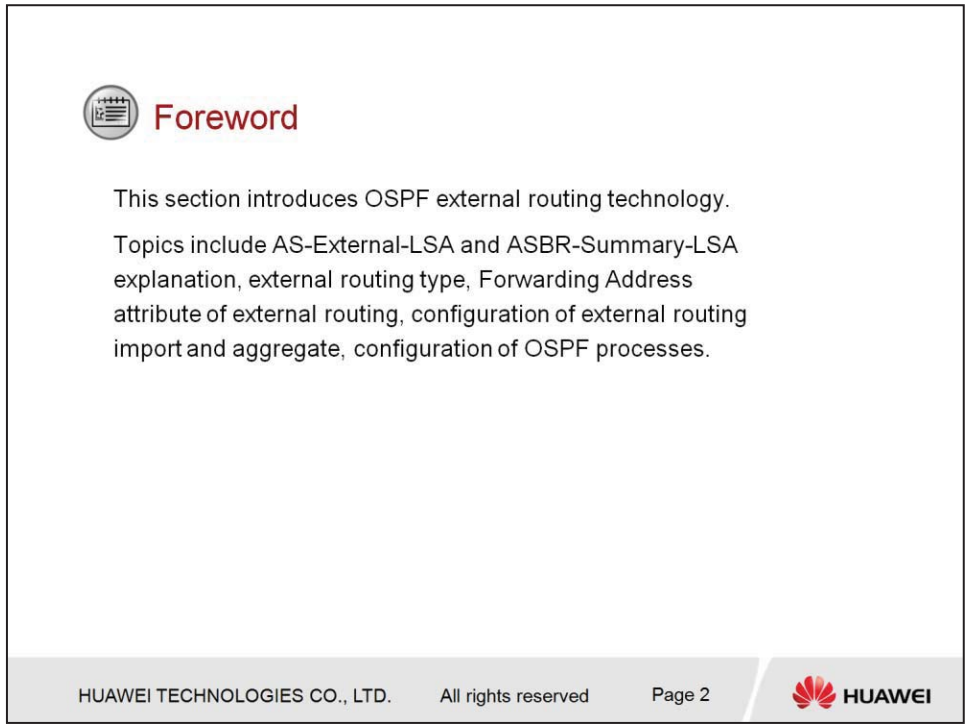
How to get IP address of the other endpoint router on virtual link?

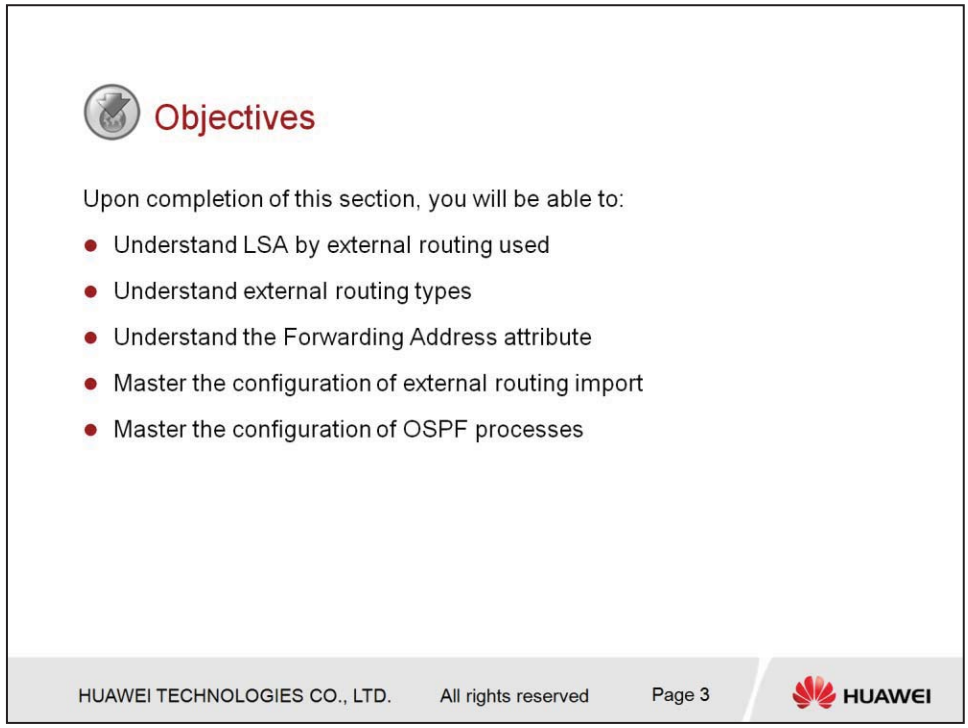
Find the IP address of virtual link through shortest path tree .

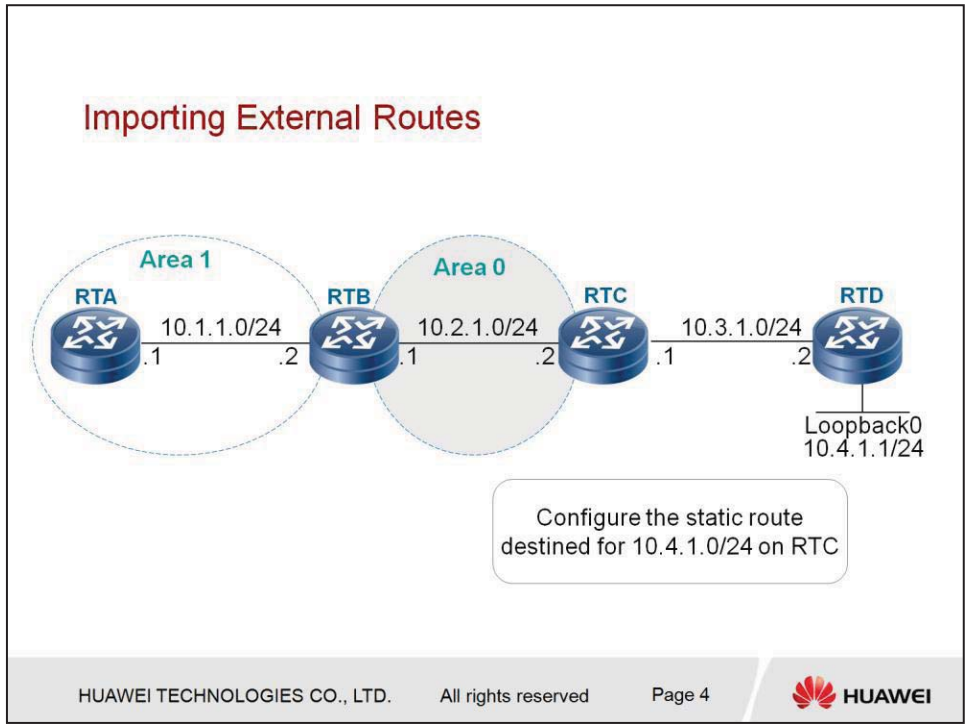
The route aggregation is configured on what type of router ?

On area border router (ABR).


OSPF External Routing
www.huawei.com
HUAWEI TECHNOLOGIES CO., LTD.
All rights reserved
HUAWEI


Foreword
This section introduces OSPF external routing technology.
Topics include AS-External-LSA and ASBR-Summary-LSA explanation, external routing type, Forwarding Address attribute of external routing, configuration of external routing import and aggregate, configuration of OSPF processes.
HUAWEI TECHNOLOGIES CO., LTD.
All rights reserved
Page 2
HUAWEI


Objectives
Upon completion of this section, you will be able to:
Understand LSA by external routing used
Understand external routing types
Understand the Forwarding Address attribute
Master the configuration of external routing import
Master the configuration of OSPF processes
HUAWEI TECHNOLOGIES CO., LTD. All rights reserved Page 3
HUAWEI



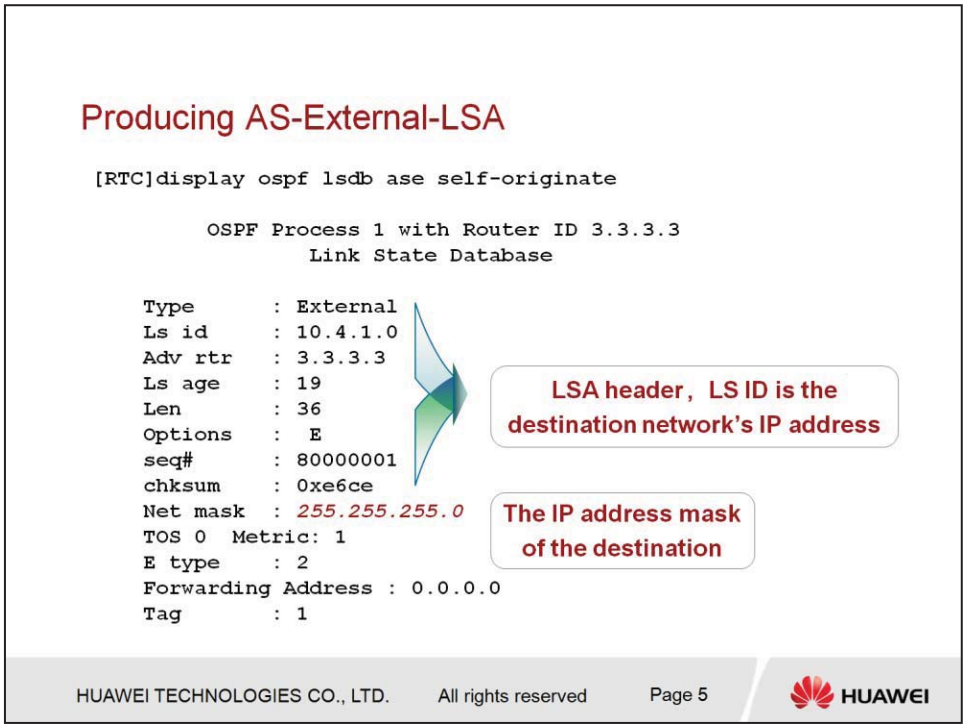
In this case :

Configuring a static route to 10.4.1.0/24 on RTC , import that static route to OSPF as an external route .

RTC will generate an AS-External-LSA to describe the imported route , RTB will generate an ASBR-Summary-LSA to describe the path to ASBR (RTC).

AS-External-LSA describes how to get to the external destination from ASBR.

ASBR-Summary-LSA describes how to get to the ASBR from ABR.



This is an AS-External-LSA generated by RTC .

LSA header of AS-External-LSA is shown as follow :

LS ID stands for destination IP address .

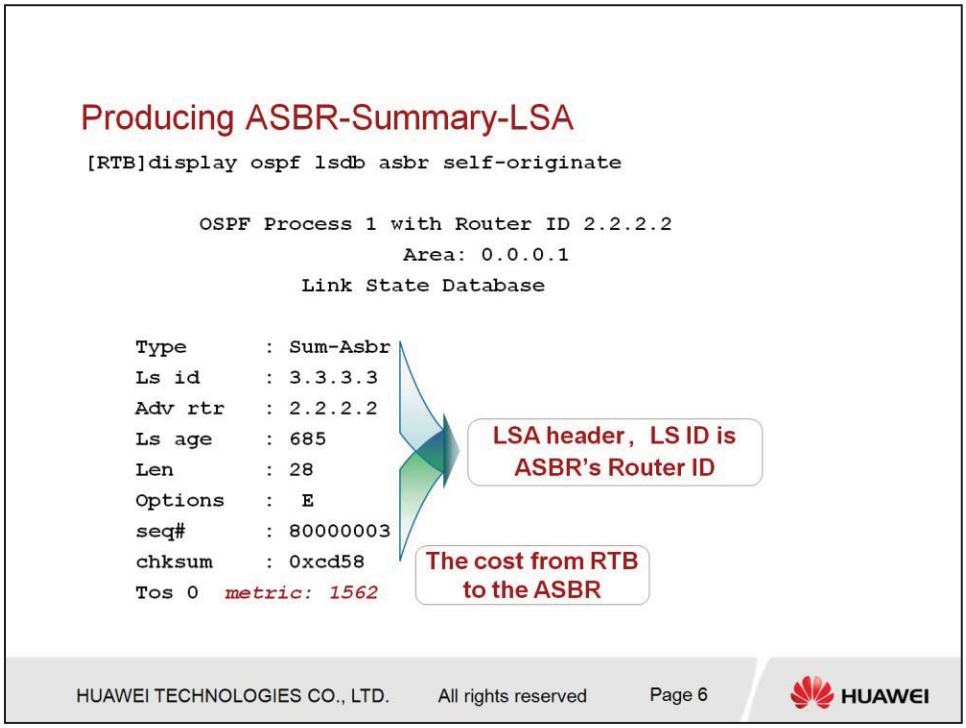
Adv rtr stands for ASBR's router ID .

Other fields is shown as follow:

Net mask stands for destination network mask .

Metric can be configured when import an external route , by default is 1 .

External route information can take a tag. Tag is used to take additional route information. Tag is normally used by route policy , and its default value is 1 .



This is an ASBR-Summary-LSA in area 1 which is generated by RTB .

When ABR floods an AS-External-LSA to the other areas, it will generate an ASBR-Summary-LSA at the same time , which describes the location of ASBR ,and floods it to the other areas .

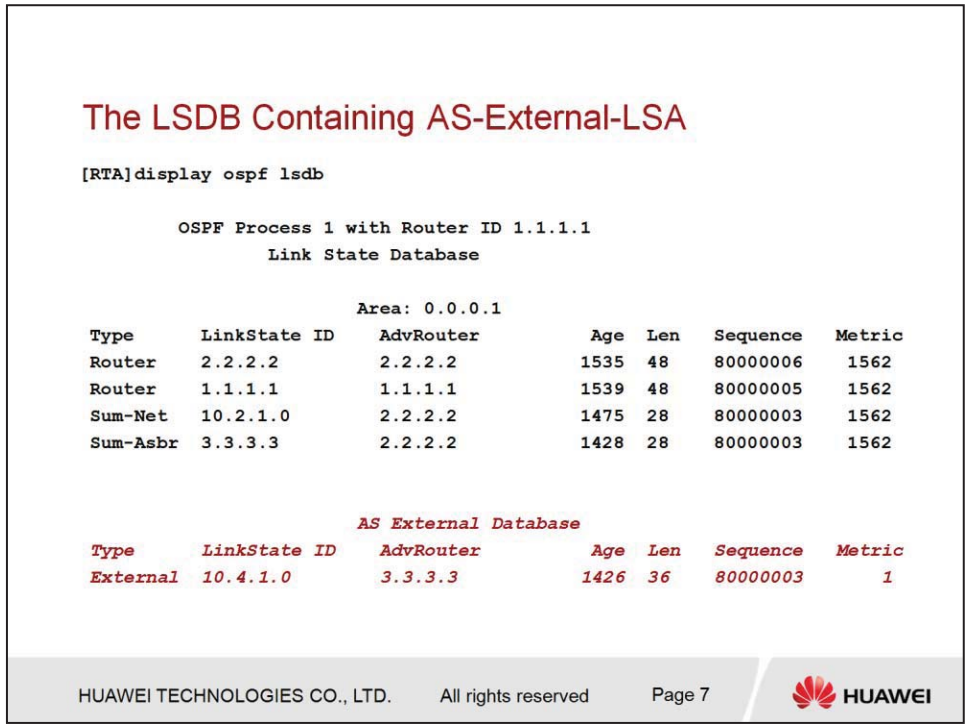
In ASBR-Summary-LSA :

LS ID stands for ASBR's Router ID ;

Adv rtr stands for ABR's Router ID;

Metric stands for the OSPF cost from ABR to ASBR.

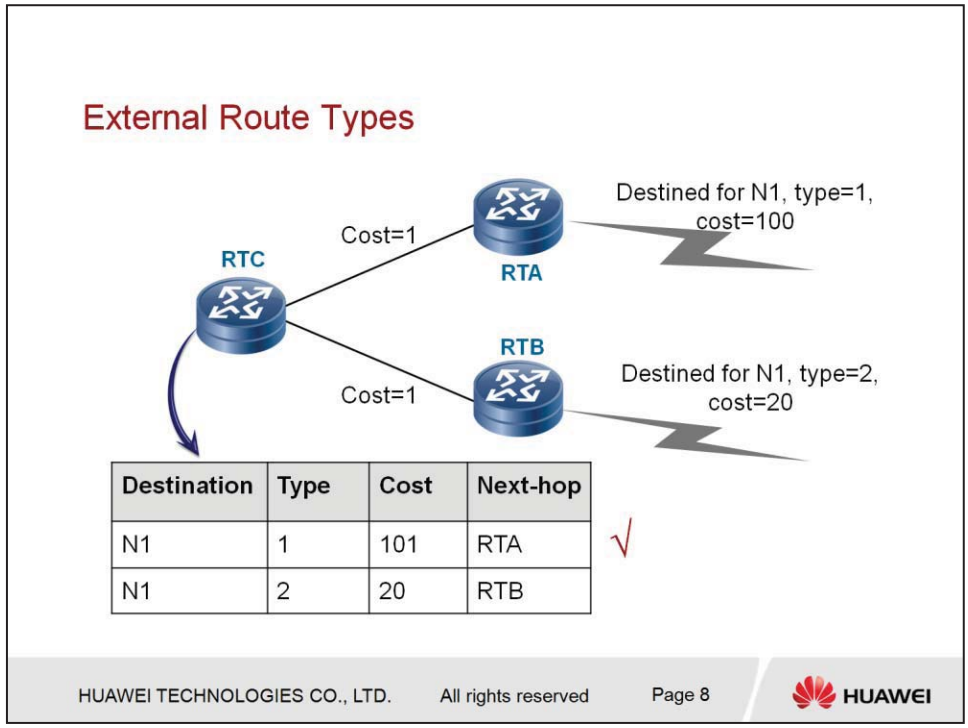
Type 4 LSA only can be flooded within one area , when type 5 LSA flooded to one area , interrelated ABR will generated a new type 4 LSA , to describe how to reach to ASBR , thus , there are many type 4 LSA to describe the same ASBR . But the advertising router and metric fields are different , it means the ASBR information is from different ABR .



This is RTA's LSDB .

In this LSDB , there is an ASBR-Summary-LSA and an AS-External-LSA In LSDB , the external routing information and the other link state information are separated .

AS-External-LSA dose not belong to any area.



There are two types of external route, type 1 external route (E1) and type 2 external route (E2) Type 1 external route (E1): the cost of a Type-1 external route equals the cost for the router to reach the corresponding ASBR plus the cost for ASBR to reach the destination.

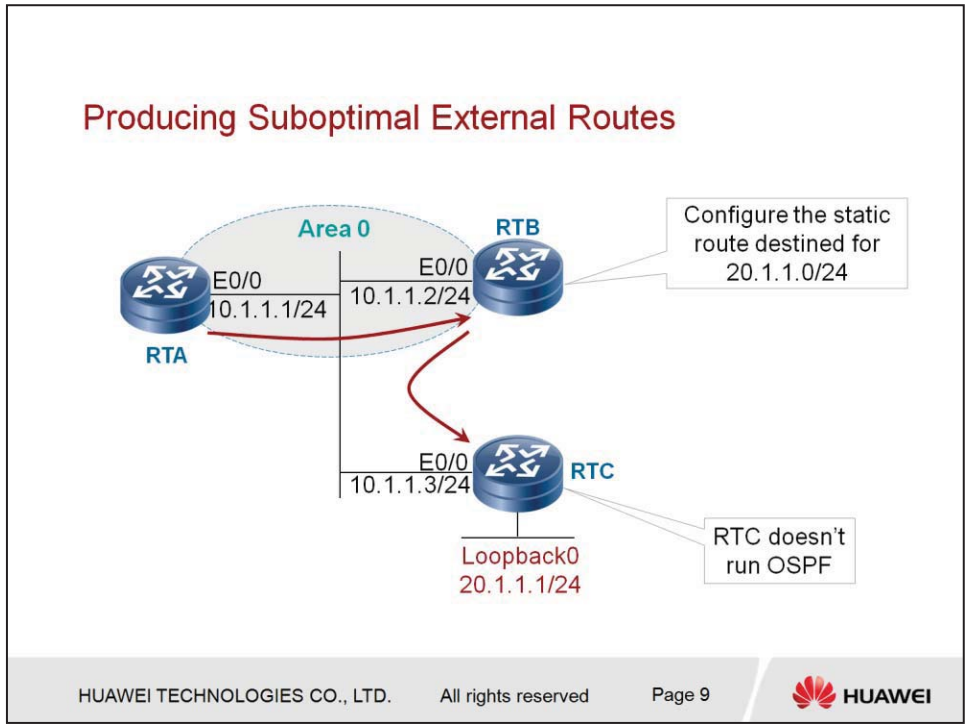
Type 2 external route(E2): the cost of a Type-2 external route equals the cost for ASBR to reach the destination.

Type 1 external route (E1) is prior to Type 2 external route(E2) forever.

The default external route type in VRP platform is Type 2(E2).

In this case :

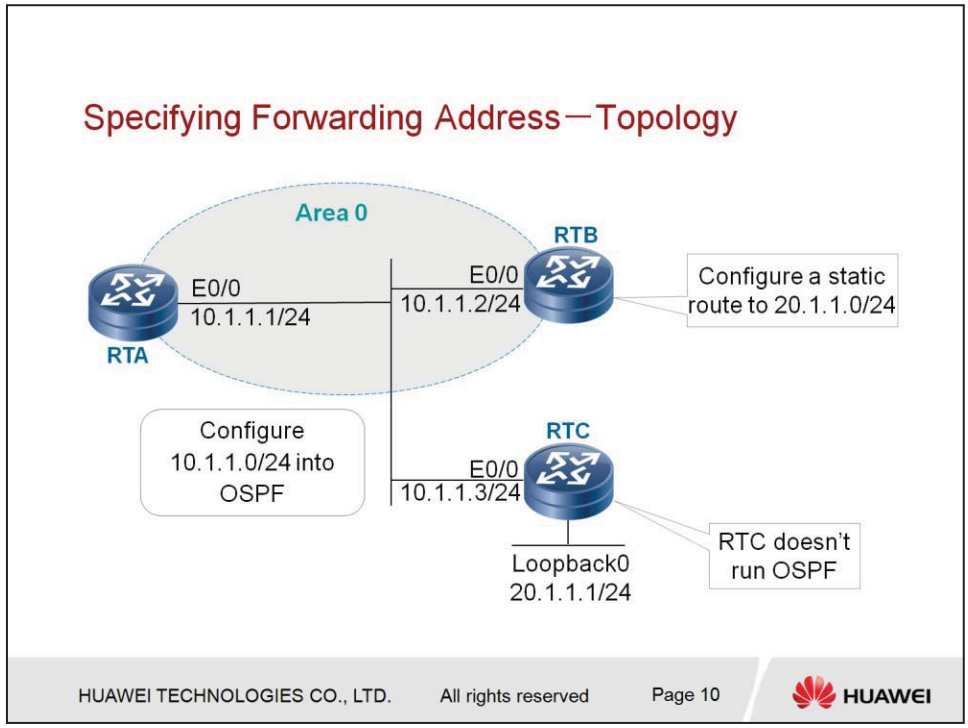
RTA advertise a AS external route to N1 , with type E1 , cost is 100 RTB advertise a AS external route to N1 , with type E2 , cost is 20 RTC receives type 5 LSA from RTA and RTB , RTA's declaration is E1 , RTC considers the cost from RTA to N1 is 100+1=101 , RTB's declaration is E2 , so RTC considers the cost from RTB to N1 is 20 (ignore the cost inside) , due to E1 has more priority than E2 ,so RTC chooses RTA as the next-hop to N1 , despite the cost is larger .



According to the topology ,10.1.1.0/24 belongs to OSPF domain ; RTC doesn't run OSPF .

Configure a static route on RTB to RTC's loopback 0 , and import it to OSPF as a external route. RTA learns the external route through OSPF , but next-hop is RTB , thus this route is suboptimal route, the best next-hop is RTC's interface E0/0.

OSPF can set forwarding address to resolve that problem .



In this case :

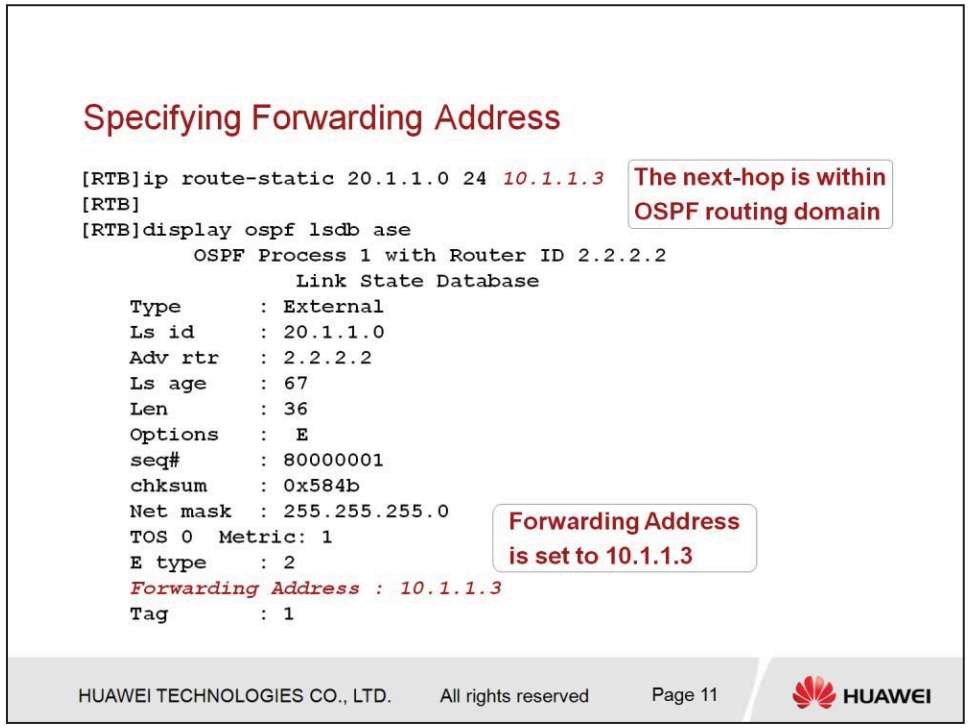
Run OSPF on RTA and RTB , configure 10.1.1.0/24 in area 0 .

Configure RTA 's router id 1.1.1.1 , RTB is 2.2.2.2 .

RTC doesn't run OSPF .

On RTB , define a static route to RTC's loopback 0 20.1.1.0/24 , and import that into OSPF .

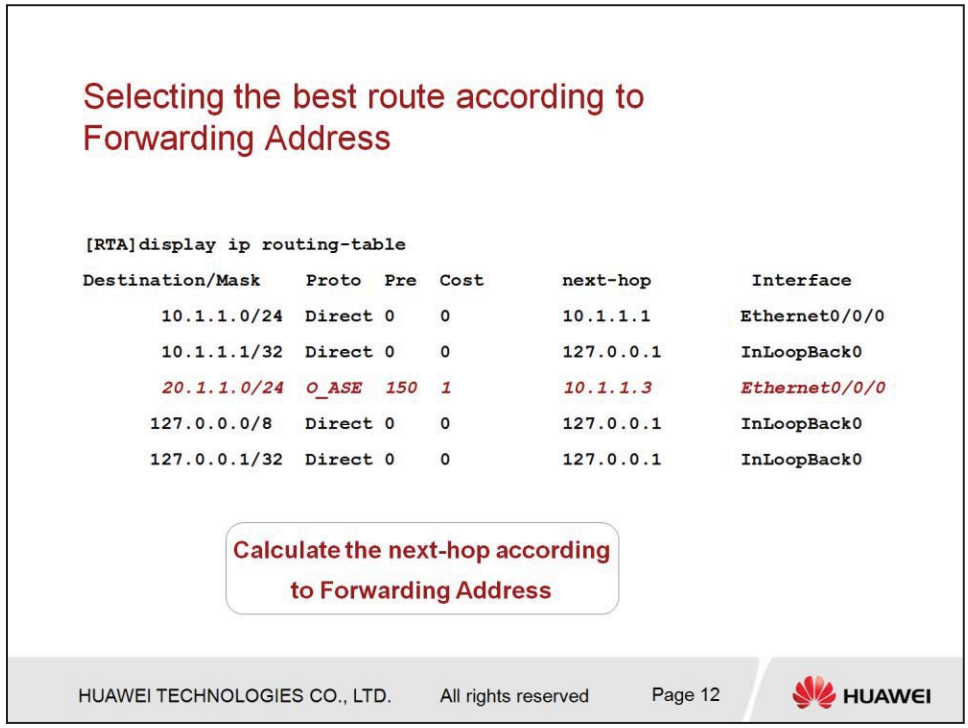
Basing on this case , using forwarding address to select the best route is discussed.



By default , when generate AS-External-LSA ,forwarding address is set to 0.0.0.0 .

If the import route's next-hop is in OSPF domain , the forwarding address which describe the AS-External-LSA must be set as the next-hop of ASBR in routing table.

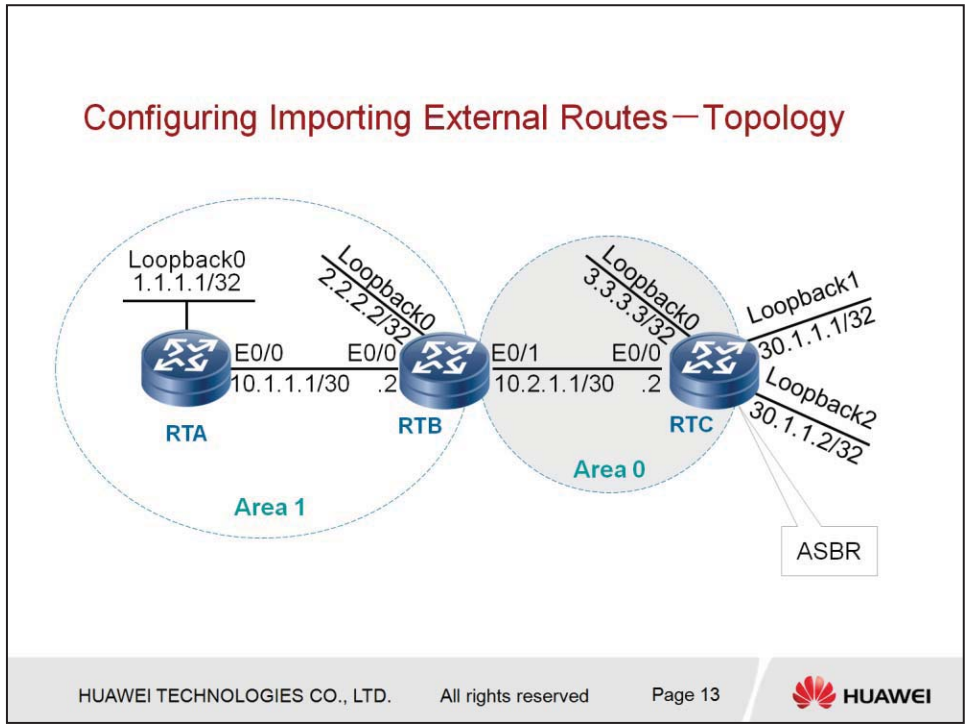
In this case , RTB defines the next-hop of the static route is 10.1.1.3 , in OSPF domain , the forwarding address in RTB's AS-External-LSA is set to 10.1.1.3 .



After receiving that LSA , RTA calculates AS external route .

If forwarding address doesn't set to zero , the next-hop of the route is the forwarding Address next-hop in routing table .

In this case , forwarding address is set to 10.1.1.3 , after calculation the external route is 10.1.1.3 either .

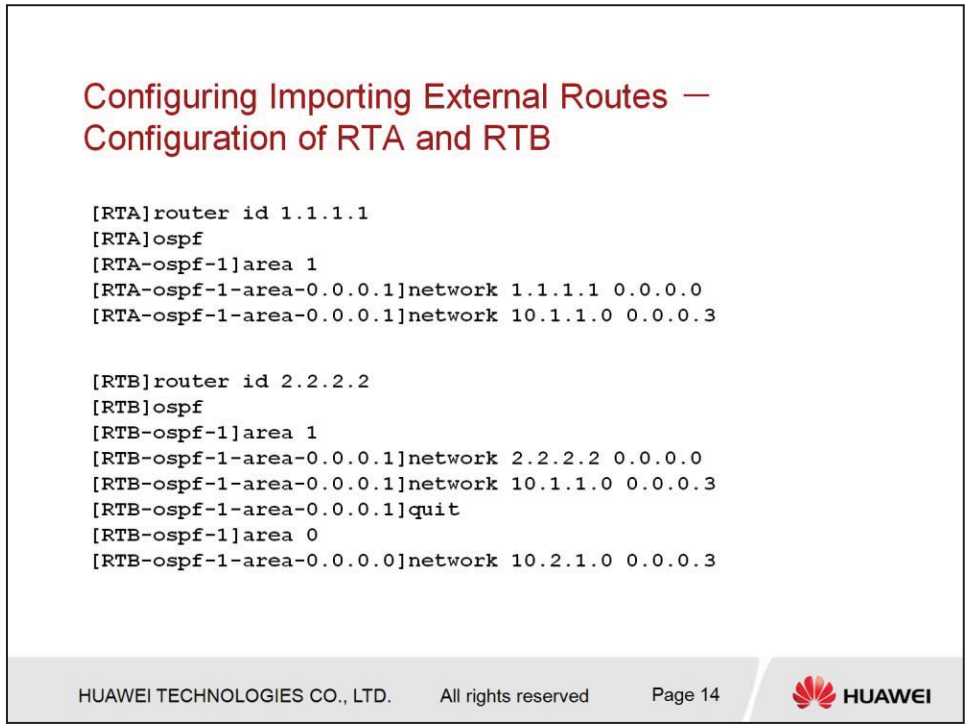


In this case :

RTC is an ASBR.

Configure static route on RTC and import that into OSPF , thus network 30.1.1.1/32 and 30.1.1.2/32 are acting as external route of OSPF.

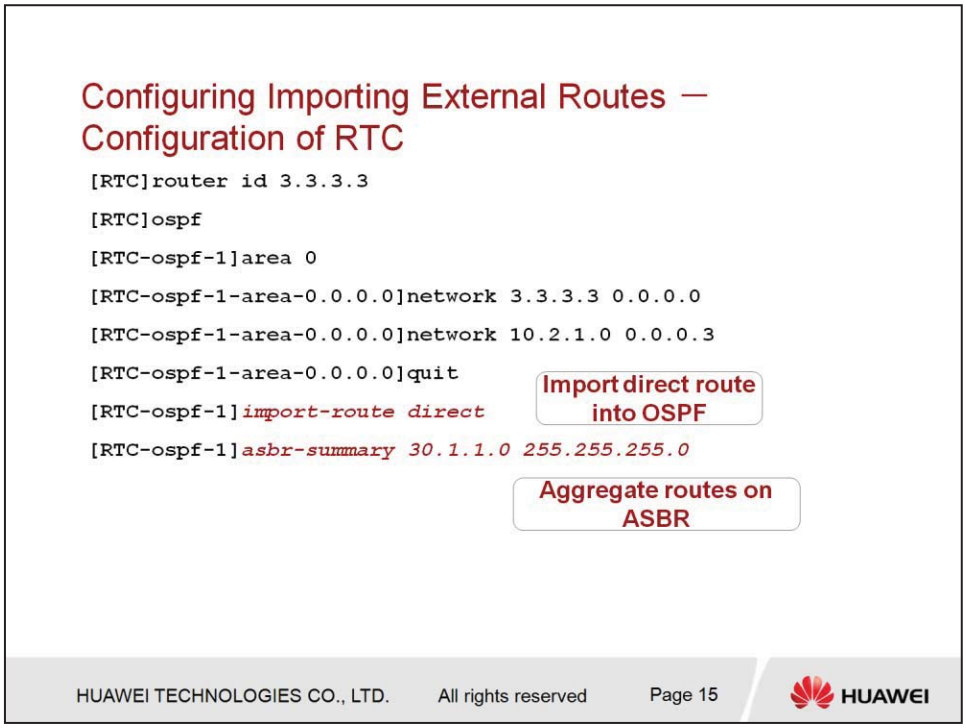
Configure route aggregation on RTC , and aggregate 30.1.1.1/32 and 30.1.1.2/32 into 30.1.1.0/24 . After configuration , RTC advertises a route to 30.1.1.0/24, does not advertise the detail route information 30.1.1.1/32 and 30.1.1.2/32 out.



On RTA , two networks are advertised in area 1.

RTB advertises 2.2.2.2/32 and 10.1.1.0/30 on area 1 , 10.2.1.0/30 on area 0 .

RTB is the ABR of area 1. When RTB advertises an AS-External-LSA in area 1 , it will generate an ASBR-Summary-LSA at the same time , in order to describe the ASBR information (RTC) .



On RTC , import direct route , and then , configure external route aggregation .

Import external routes :

import-route protocol [allow-ibgp] [cost value] [type value] [tag value]

[route-policy route-policy-name]

When import external routes , route cost , external route type and tag can be configured manually .

allow-ibgp：when protocol is BGP , allow-ibgp is optional key word. import-route bgp means only EBGP routes can be imported , import-route bgp allow-ibgp means both EBGP and IBGP can be imported .

External routes aggregation :

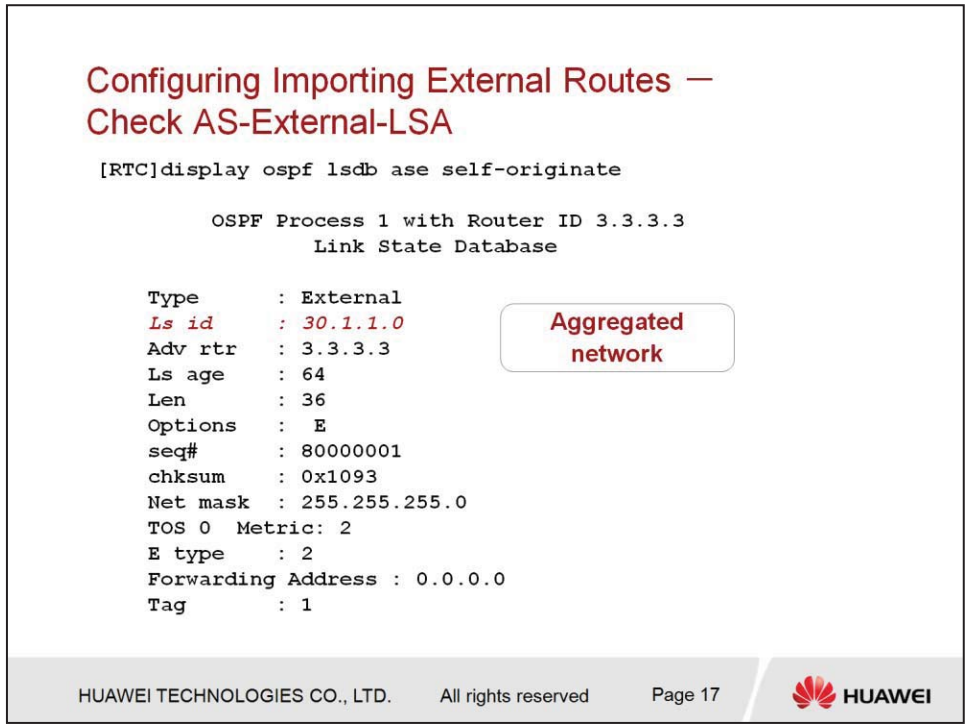
asbr-summary : ip-address mask [not-advertise | tag tag-value]

not-advertise：not advertise routes matching the specified IP address or mask. If it is not specified, the summarized route will be advertised.

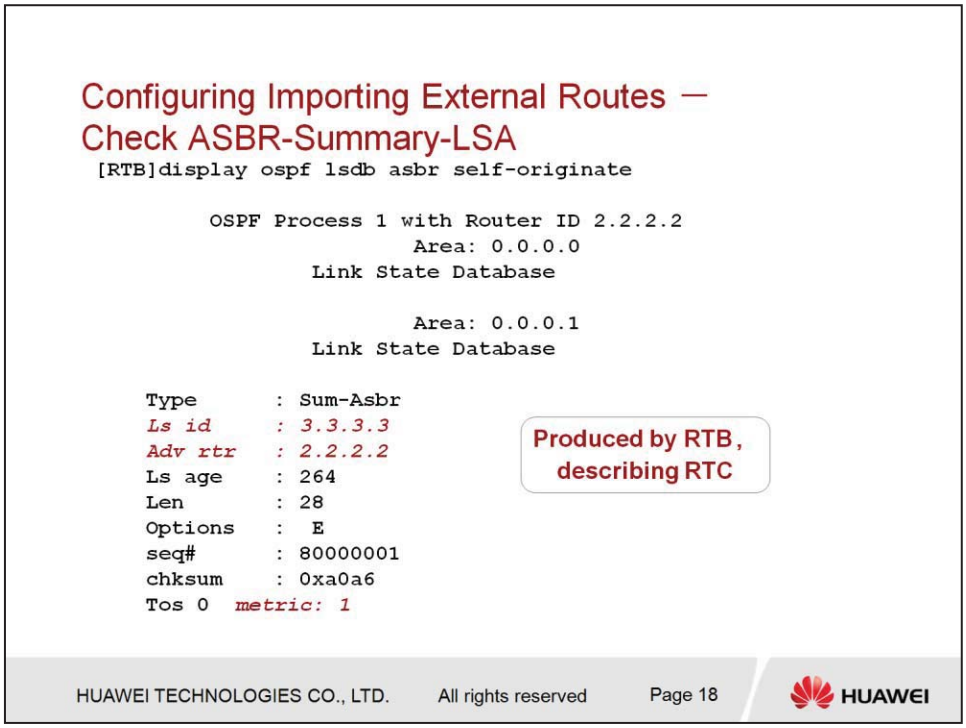
tag-value：specifies the tag value that is mainly used to control advertisement of routes through route-policy. The value ranges

from 0 to 4294967295. By default, the value is 1.

By default ,only aggregated route can be advertised .



RTC generates type 5 LSA , in this LSA only 30.1.1.0/24 exists , there is no detail,routing information about 30.1.1.1/32 and 30.1.1.2/32 .



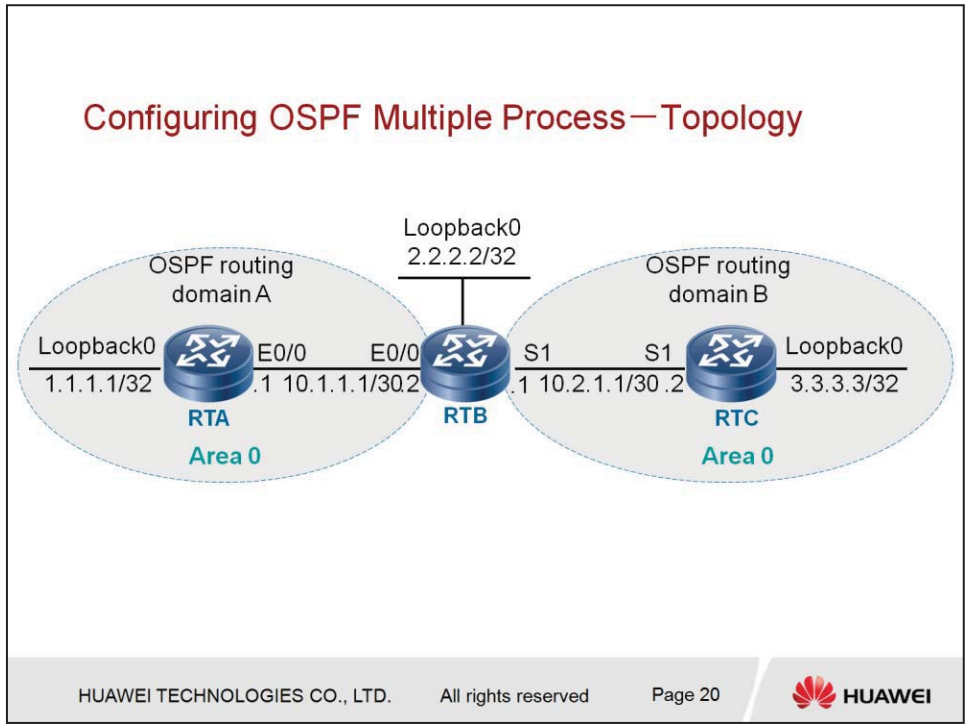
RTB generates a type 4 LSA , describes how to reach to ASBR (RTC) from RTB , and floods that information in area 1 .

Configuring Importing External Routes — Check RTA's Routing Table

```
[RTA]display ip routing-table
Routing Tables: Public
         Destinations : 17        Routes : 17
Destination/Mask    Proto   Pre  Cost      next-hop         Interface
        1.1.1.1/32  Direct  0    0         127.0.0.1        InLoopBack0
        3.3.3.3/32  O_ASE   150  1         10.1.1.2         Ethernet0
       10.1.1.0/24  O_ASE   150  1         10.1.1.2         Ethernet0
       10.1.1.0/30  Direct  0    0         10.1.1.1         Ethernet0
       10.1.1.1/32  Direct  0    0         127.0.0.1        InLoopBack0
       10.2.1.0/30  OSPF    10   2         10.1.1.2         Ethernet0
       10.2.1.1/32  O_ASE   150  1         10.1.1.2         Ethernet0
       10.3.1.0/24  O_ASE   150  1         10.1.1.2         Ethernet0
       10.3.1.2/32  O_ASE   150  1         10.1.1.2         Ethernet0
       10.5.1.0/30  Direct  0    0         10.5.1.1         Serial1
       10.5.1.1/32  Direct  0    0         127.0.0.1        InLoopBack0
       10.5.1.2/32  Direct  0    0         10.5.1.2         Serial1
       20.1.1.1/32  Direct  0    0         127.0.0.1        InLoopBack0
       20.1.1.2/32  Direct  0    0         127.0.0.1        InLoopBack0
       30.1.1.0/24  O_ASE   150  2         10.1.1.2         Ethernet0
      127.0.0.0/8   Direct  0    0         127.0.0.1        InLoopBack0
      127.0.0.1/32  Direct  0    0         127.0.0.1        InLoopBack0
```

HUAWEI TECHNOLOGIES CO., LTD. All rights reserved HUAWEI

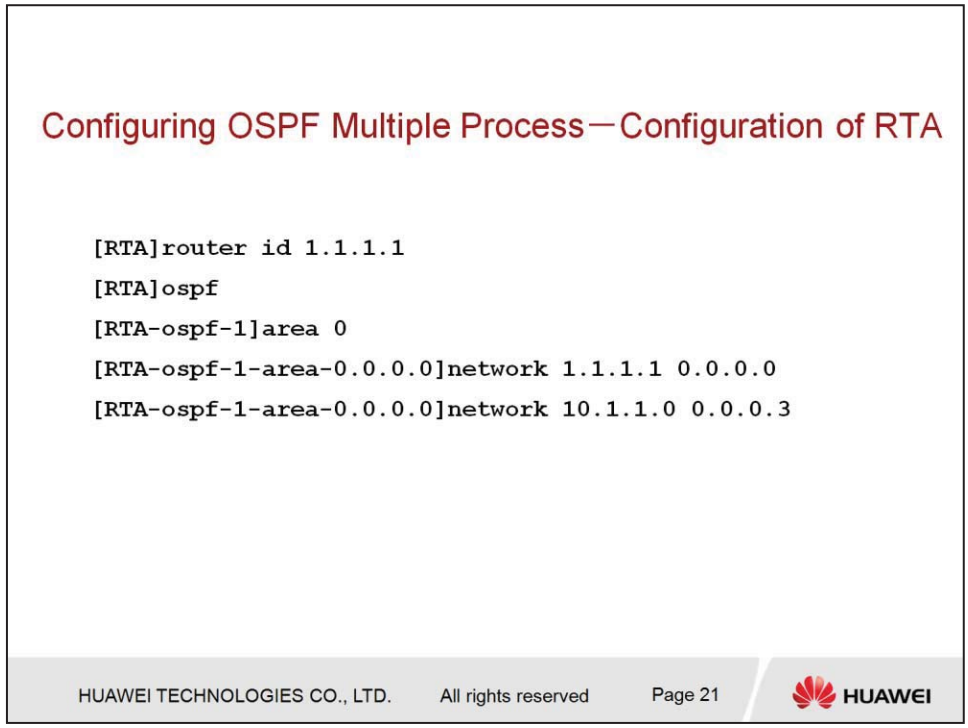
Routing table inside AS , there is no detail routing information , only aggregated route 30.1.1.0/24 exists.



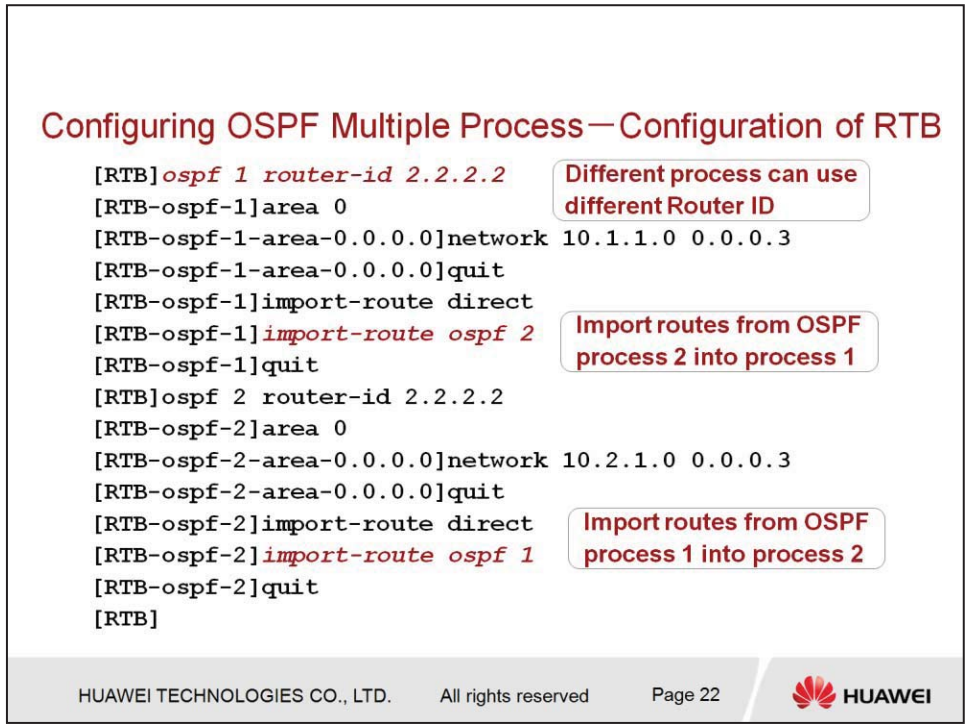
In this case :

There are two OSPF domains in this network . RTB is the ASBR of the two domains .

RTB must run two OSPF processes .

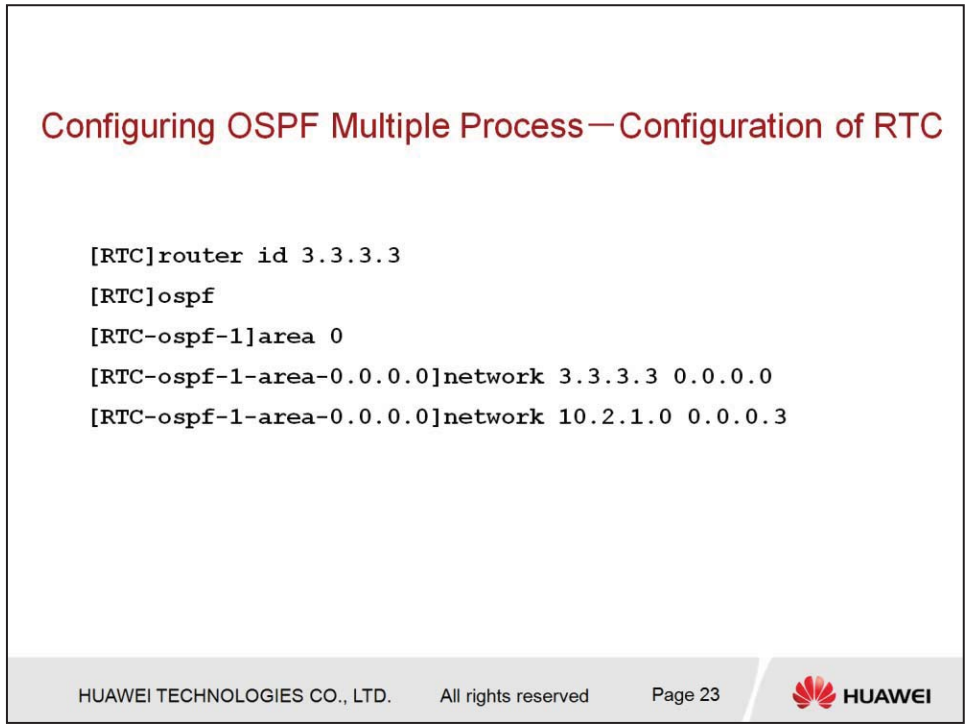


On RTA , there is only one OSPF process , advertise all networks in area 0 .



RTB runs two OSPF processes , configure different router id for each OSPF process .

When configure each process , import the routing information which belongs to the other process into itself .



On RTC , there is only one OSPF process , advertise all networks in area 0 .

Configuring OSPF Multiple Process—Check IP Routing Table

```
[RTC]display ip routing-table
Routing Tables: Public
         Destinations : 12        Routes : 12
Destination/Mask    Proto   Pre  Cost      next-hop         Interface
        1.1.1.1/32  O_ASE   150  1         10.2.1.1         Serial1
        3.3.3.3/32  Direct  0    0         127.0.0.1        InLoopBack0
       10.1.1.0/30  O_ASE   150  1         10.2.1.1         Serial1
       10.2.1.0/30  Direct  0    0         10.2.1.2         Serial1
       10.2.1.1/32  Direct  0    0         10.2.1.1         Serial1
       10.2.1.2/32  Direct  0    0         127.0.0.1        InLoopBack0
      127.0.0.0/8   Direct  0    0         127.0.0.1        InLoopBack0
      127.0.0.1/32  Direct  0    0         127.0.0.1        InLoopBack0
```

The routes learned from other process are considered as AS external routes

HUAWEI TECHNOLOGIES CO., LTD. All rights reserved

All the routing information learned by the other OSPF process is AS external route .

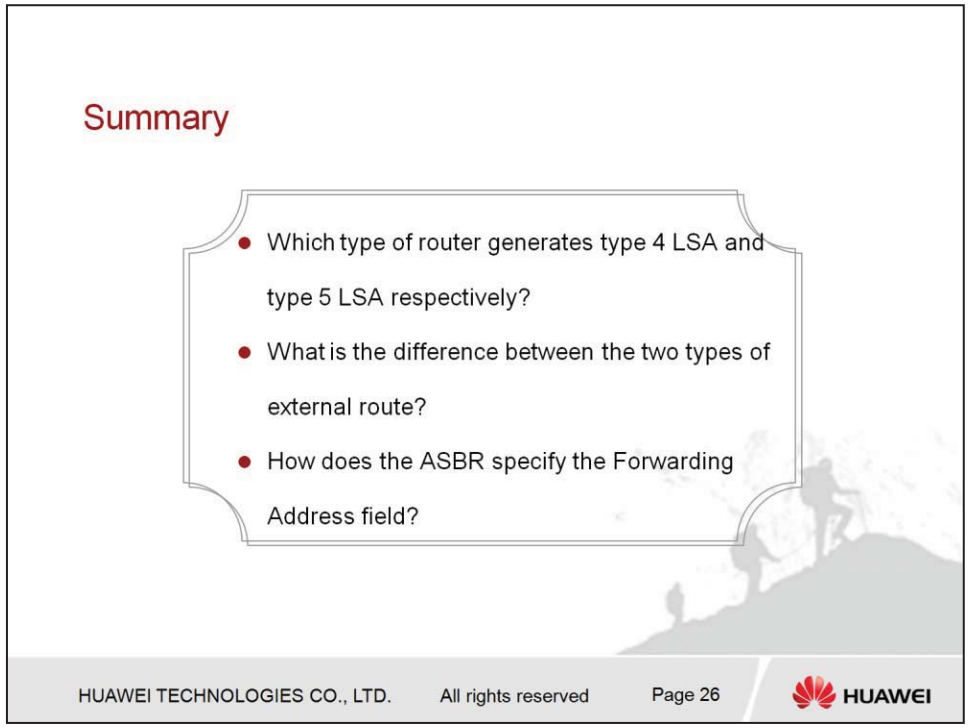
Configuring OSPF Multiple Process —Check LSDB

```
[RTB]display ospf lsdb
         OSPF Process 1 with Router ID 2.2.2.2
                   Link State Database
                         Area: 0.0.0.0
 Type      LinkState ID    AdvRouter  Age  Len   Sequence   Metric
 Router    2.2.2.2         2.2.2.2    136  36    80000006       1
 Router    1.1.1.1         1.1.1.1    469  36    80000003       1
 Network   10.1.1.2        2.2.2.2    463  32    80000002       0
                 AS External Database
 Type      LinkState ID    AdvRouter  Age  Len   Sequence   Metric
 External  10.2.1.0        2.2.2.2    136  36    80000001       1

         OSPF Process 2 with Router ID 2.2.2.2
                   Link State Database
                         Area: 0.0.0.0
 Type      LinkState ID    AdvRouter  Age  Len   Sequence   Metric
 Router    3.3.3.3         3.3.3.3    200  48    80000024       1
 Router    2.2.2.2         2.2.2.2    120  48    80000022       1
                 AS External Database
 Type      LinkState ID    AdvRouter  Age  Len   Sequence   Metric
 External  10.1.1.0        2.2.2.2    120  36    80000001       1
```

HUAWEI TECHNOLOGIES CO., LTD. All rights reserved HUAWEI

There are two LSDBs on RTB , each process maintains its own LSDB.



Which type of router generates type 4 LSA and type 5 LSA respectively?

Type 4 LSA is generated on ABR , each LSA is flooded within area , type 5 LSA is generated on ASBR , and flooded within AS

What's the difference between two types of external routes?

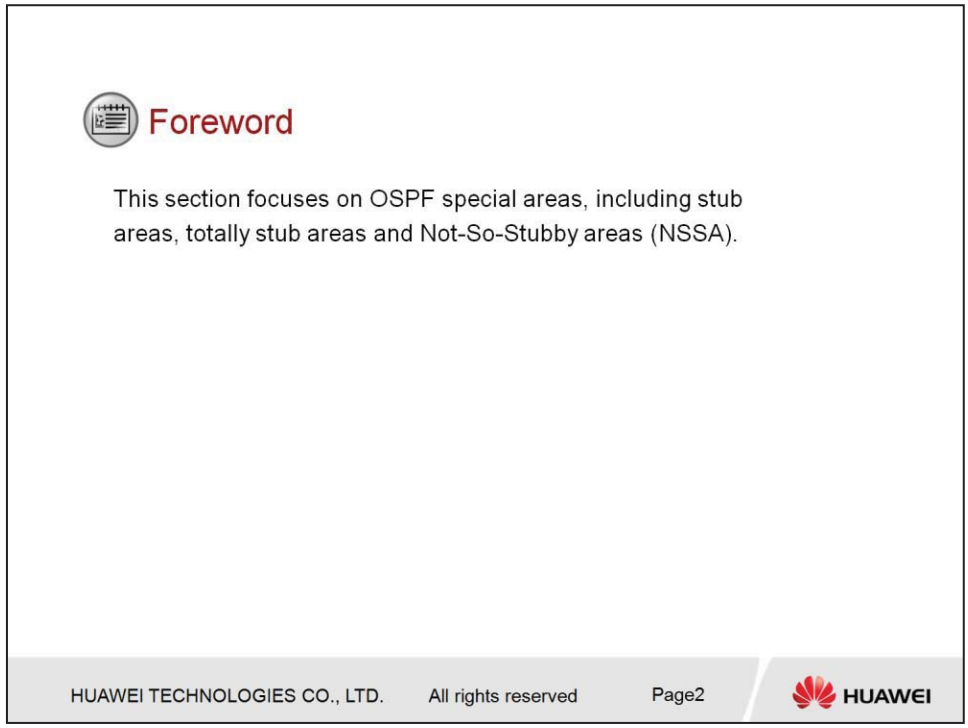
Cost calculation method is different , type 1 external route (E1) cost is the sum of AS inside cost and AS outside cost .

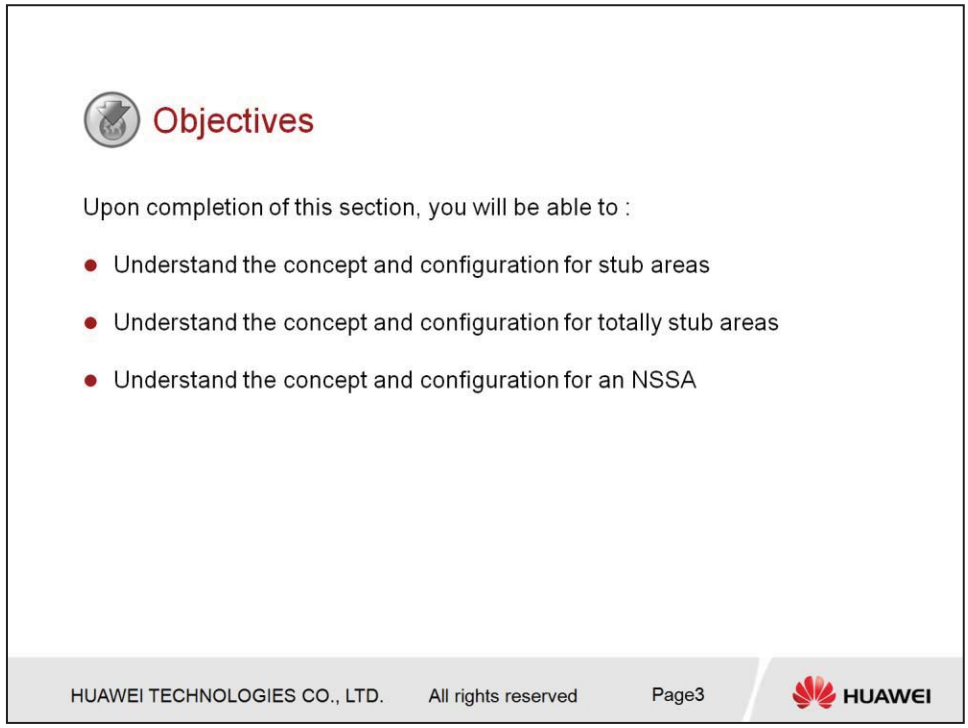
Type 2 external route(E2) cost is the AS outside cost , AS inside cost is ignored .

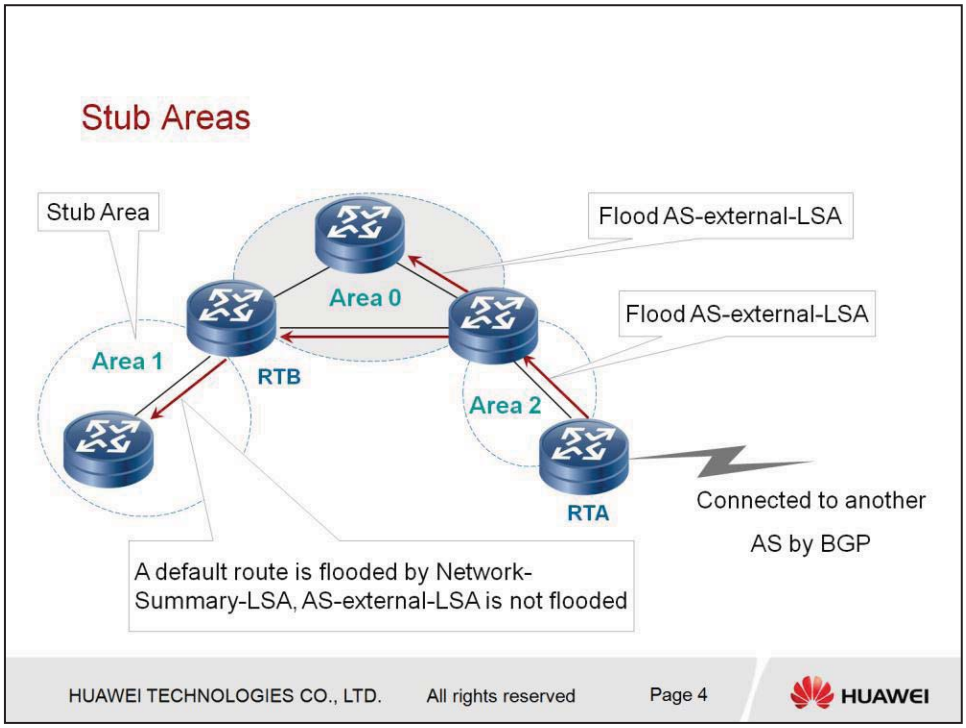
How to specify the Forwarding Address field?

On ASBR , if the next-hop of imported external route is in the OSPF domain , the forwarding address must specify as the next-hop of the external route , if the nexthop of the imported external route is outside the OSPF domain , the forwarding address is specified as 0.0.0.0 .


OSPF Special Areas
www.huawei.com
HUAWEI TECHNOLOGIES CO., LTD.
All rights reserved
HUAWEI


Foreword
This section focuses on OSPF special areas, including stub areas, totally stub areas and Not-So-Stubby areas (NSSA).
HUAWEI TECHNOLOGIES CO., LTD. All rights reserved Page2
HUAWEI


Objectives
Upon completion of this section, you will be able to :
Understand the concept and configuration for stub areas
Understand the concept and configuration for totally stub areas
Understand the concept and configuration for an NSSA
HUAWEI TECHNOLOGIES CO., LTD. All rights reserved Page3
HUAWEI



Special area is allowed to configure into stub area in OSPF .

AS-external-LSA can not be flooded into stub area . If the inside router wants to reach outside , there is a default route generated by ABR .

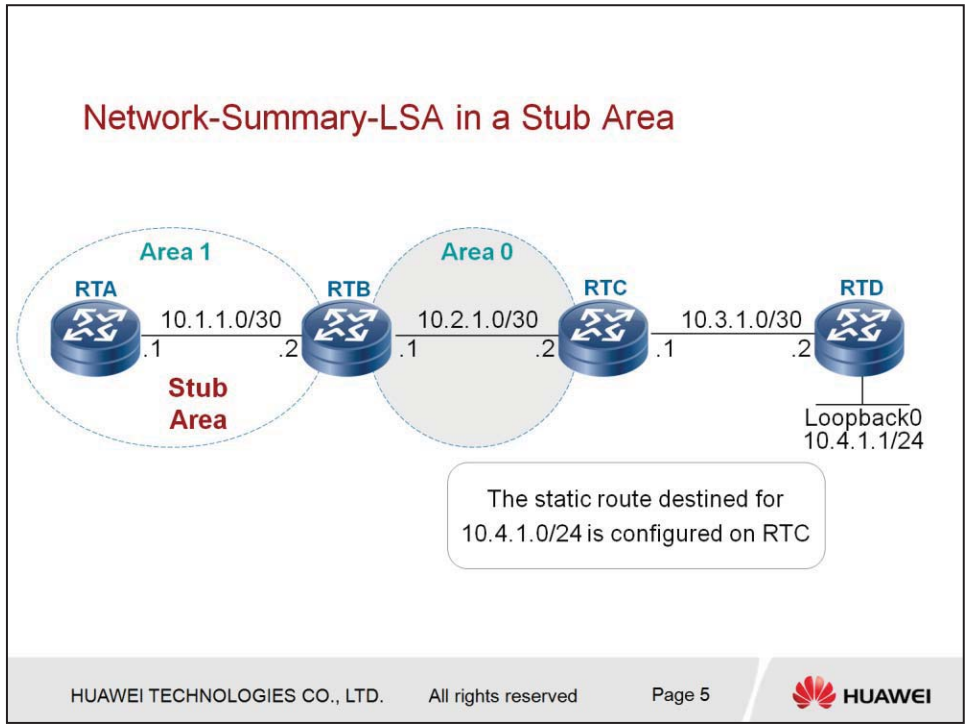
Using stub area can reduce the requirement of the memories, CPU usage and the size of LSDB .

In this case :

RTA is an ASBR , configure area 1 into a stub area , RTB is area 1's ABR .

RTA floods the external route into AS inside through AS-external-LSA .

RTB only advertises one default route into area 1 through Network-Summary-LSA , but not AS-external-LSA .

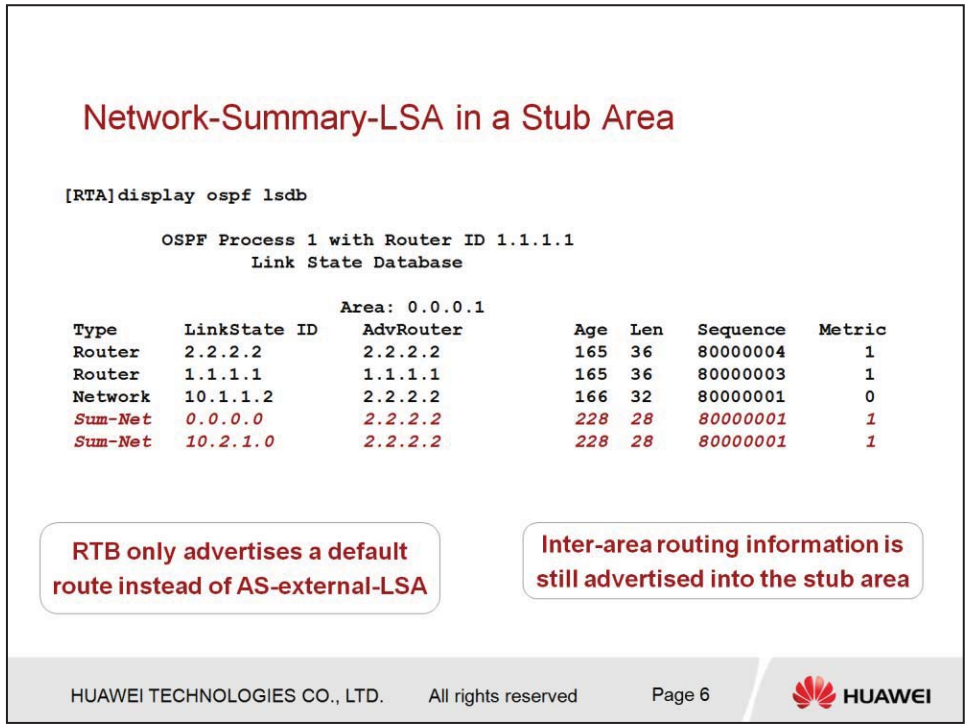


In this case :

A static route to 10.4.1.0/24 is configured on RTC , which is imported into OSPF , so RTC is an ASBR .

RTA uses 1.1.1.1 as its router id ; RTB uses 2.2.2.2 as its router id ; RTC uses 3.3.3.3 as its router id .

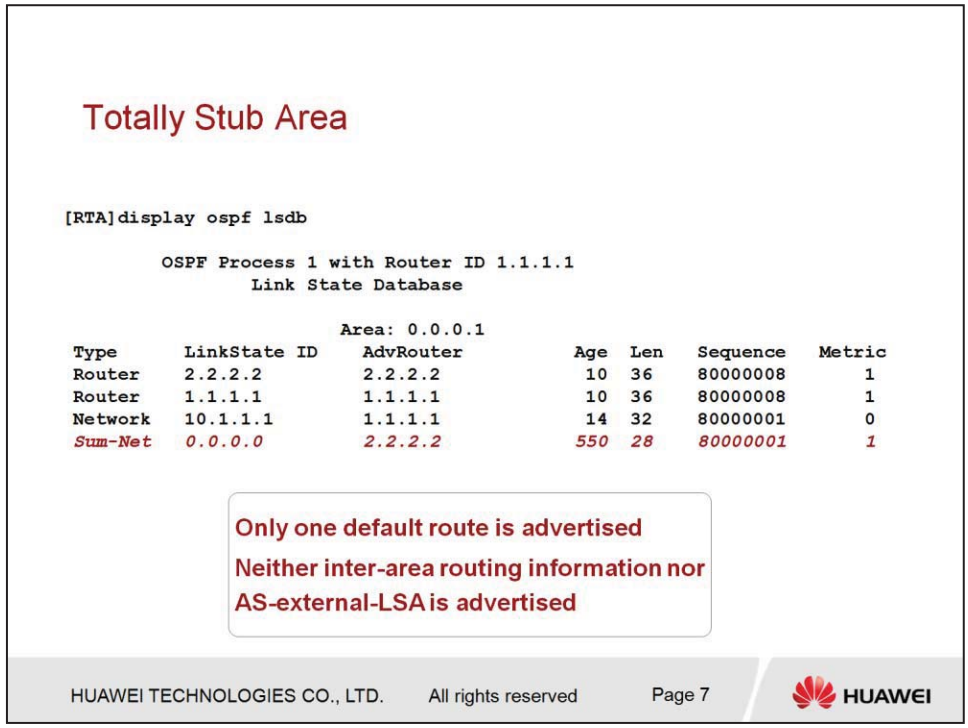
Configure area 1 as a stub area .



This is a Network-Summary-LSA in area 1 which generated by RTB .

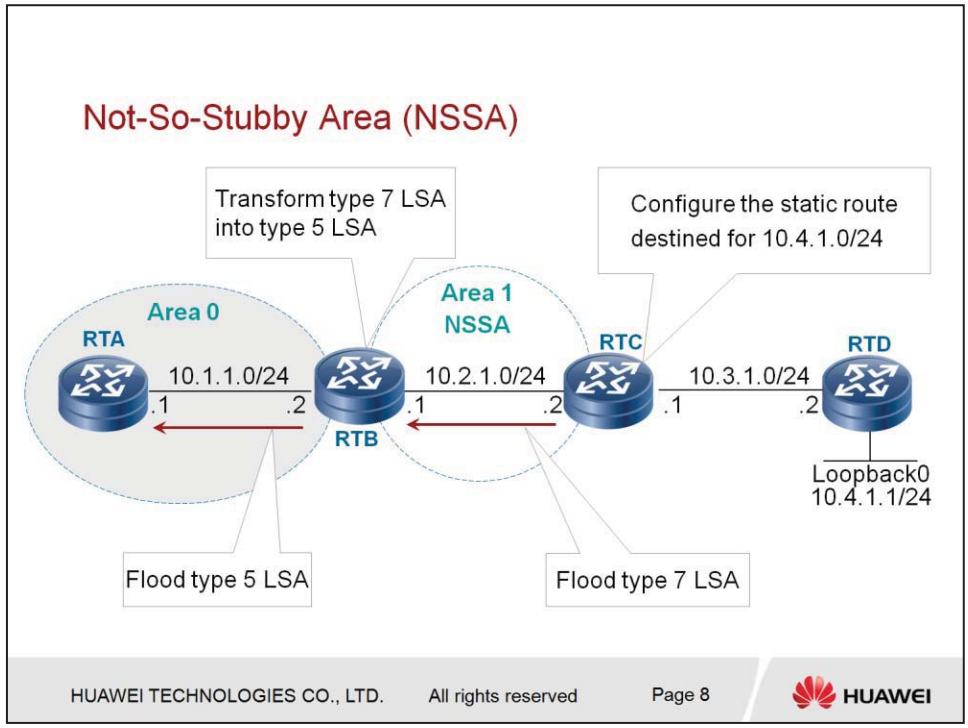
RTB describes a default route through a Network-Summary-LSA, and floods it within area 1.

Routing information between areas still need to be advertised into stub area by type 3 LSA.



A stub area can be configured as totally stub area by configuring ABR .

If an area is configured into a totally stub area , only one default route is advertised into this area through Network-Summary-LSA . There isn't any routing information between areas ,or outside AS . In another word , totally stub area can't flood Summary-LSA (except default route) or AS-external-LSA .



NSSA means Not So Stubby Area.

NSSA and Stub area are analogous. Both of them can't receive AS-external-LSA, but there is another way for NSSA to advertise external LSA.

In this case :

Create loopback0 on RTA, RTB and RTC , with the IP address

1.1.1.1 ,2.2.2.2 ,3.3.3.3 .

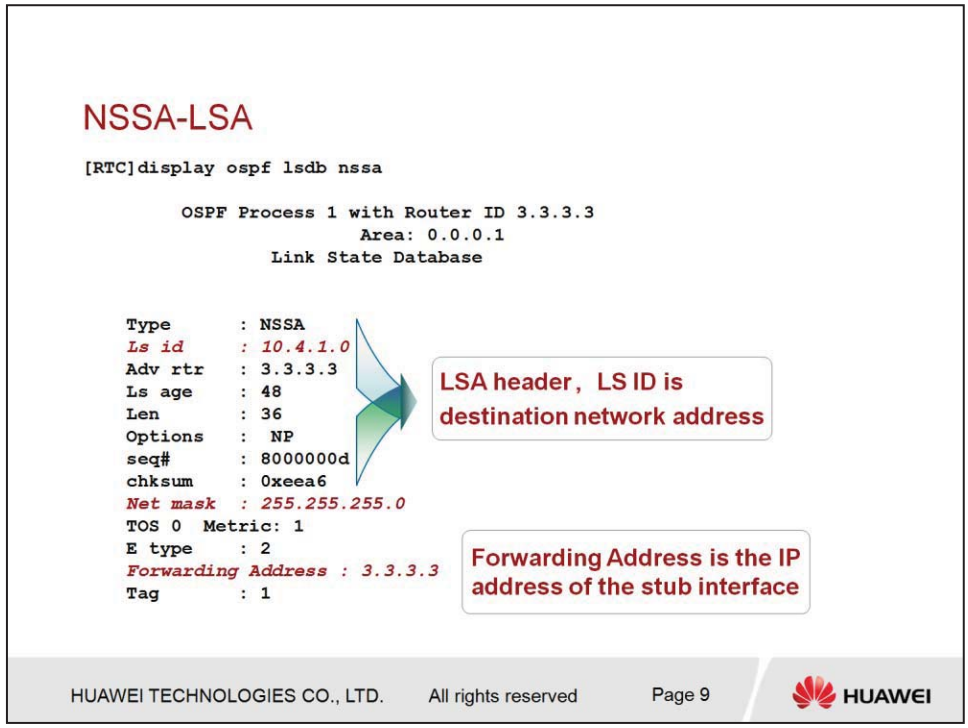
RTA's router id is 1.1.1.1 , RTB's router id is 2.2.2.2 , RTC's router id is 3.3.3.3 .

Configure network 1.1.1.1/32, 10.1.1.0/24 in area 0 . Configure network 2.2.2.2/32, 10.2.1.0/24 in area 1.

Configure static route to 10.4.1.0/24 on RTC with the next-hop is 10.3.1.2 , import the static route into OSPF , and configure area 1 into NSSA .

RTC advertises one NSSA-LSA (type 7) into area 1 , this LSA is used to describe the external route .

Type 5 LSA can't be flooded within NSSA area. But Type 7 LSA can be flooded inside NSSA area, so ABR of NSSA need to translate the NSSA-LSA into Asexternal-LSA, and flood that to the other areas.



This is a NSSA-LSA generated by RTC (Type 7 LSA).

In this case , RTC imports a static route into OSPF ,RTC generates a NSSA-LSA and describes a routing information to 10.4.1.0/24 which is outside AS , the key word is set as follow :

LS ID is set as destination IP address ;

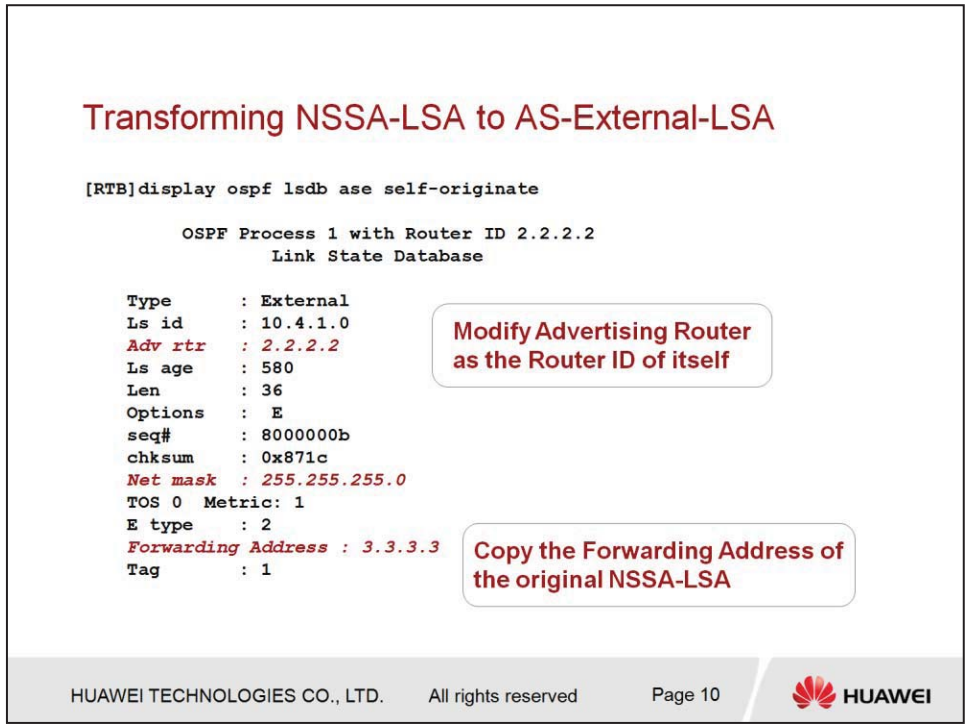
Options field shows this LSA can be translated into an AS-external-LSA by ABR;

Default metric is 1;

E type (external route type) type 2 by default ;

The rule of setting NSSA-LSA ‘s Forwarding Address is described as follow:

If options field shows this LSA can be translated into type 5 LSA , the forwarding address can’t be set to 0.0.0.0 . If the next-hop of the import route is in the OSPF domain , the forwarding address is set to the next-hop of the imported external route ; If the next-hop of the import route isn’t in the OSPF domain , the forwarding address is set to the stub interface address which belongs to the ASBR inside the OSPF domain (loopback 0 for example). If there are many stub networks , the largest IP address is selected .



When RTB (the ABR of area 1)receives that NSSA-LSA , the LSA is translated into an AS-external-LSA .

The key words is set as follow :

LS ID and Network Mask are copied from the original NSSA-LSA , describe the routing information to 10.4.1.0/24 , which is outside the AS .

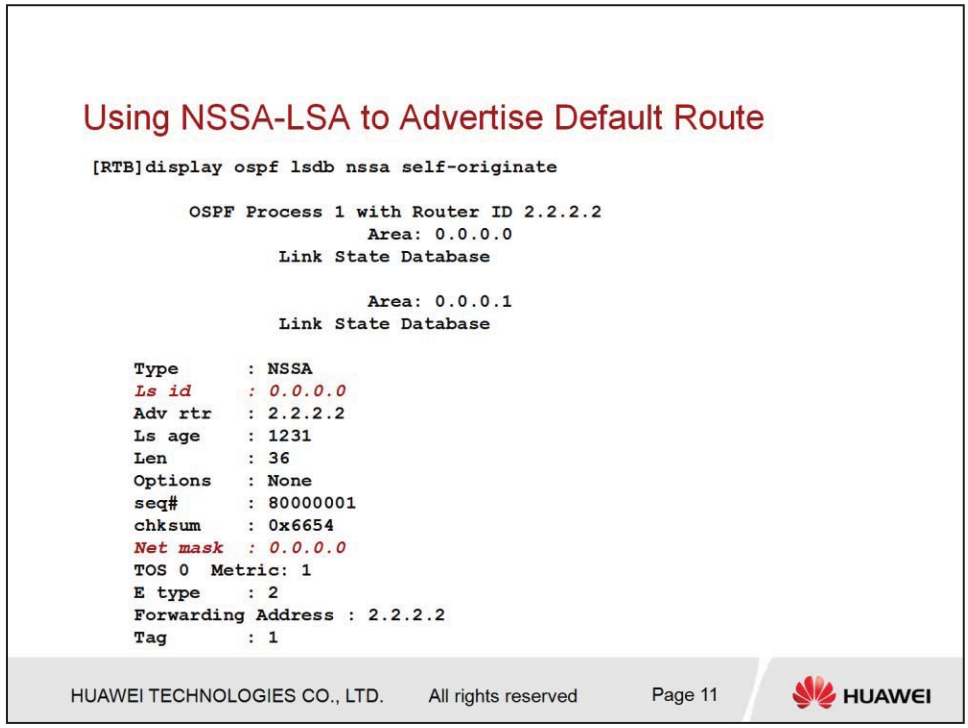
If not recalculate the metric , or not set metric when configure , the metric is 1 by default .

Forwarding address is copied from the original NSSA-LSA , without changes .

Thus , the router outside the NSSA area calculate next-hop , just according to the forwarding address of AS-external-LSA .

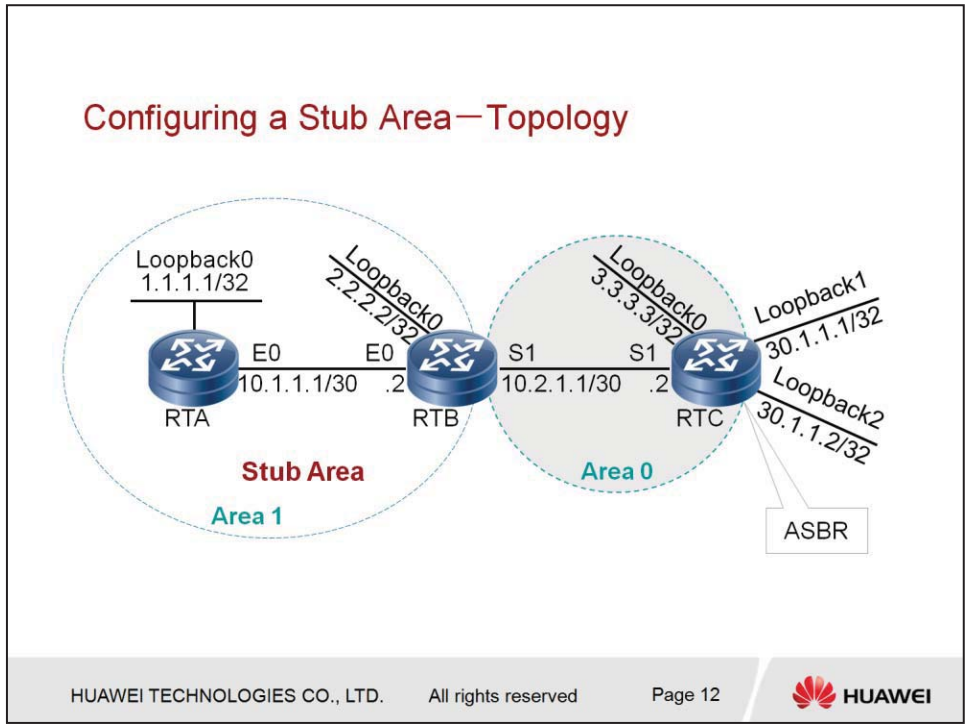
In order to prevent to duplication of external routing information , if there are several ABR inside the NSSA area , only one ABR can translate the NSSA-LSA into AS-external-LSA , so this ABR is called the translator of NSSA .

Translator is elected by router id . NSSA area ‘s ABR will identify itself with one bit in Router-LSA , with checking the Router-LSA within one area , every NSSA ABR can maintain a list of ABR , and elect the one with lager router id as the translator .



A default route can be advertised into NSSA through configure NSSA's ABR , this default route is described by a NSSA-LSA , but not Network-Summary-LSA.

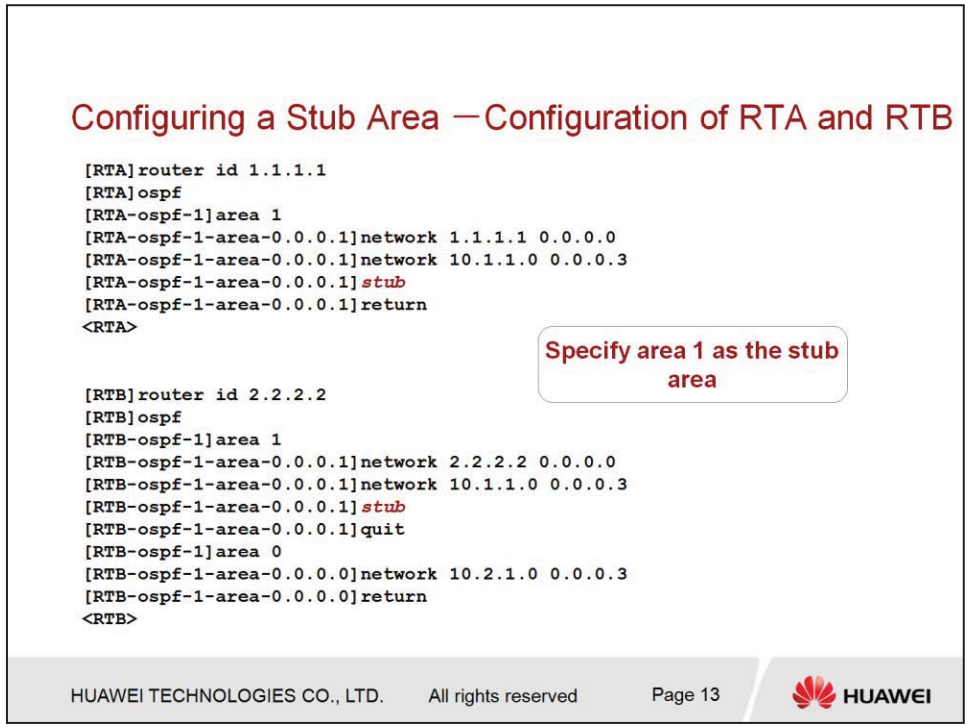
In this case , RTB generates a default route ,with LS ID is 0.0.0.0 , Network mask is 0.0.0.0.



In this case :

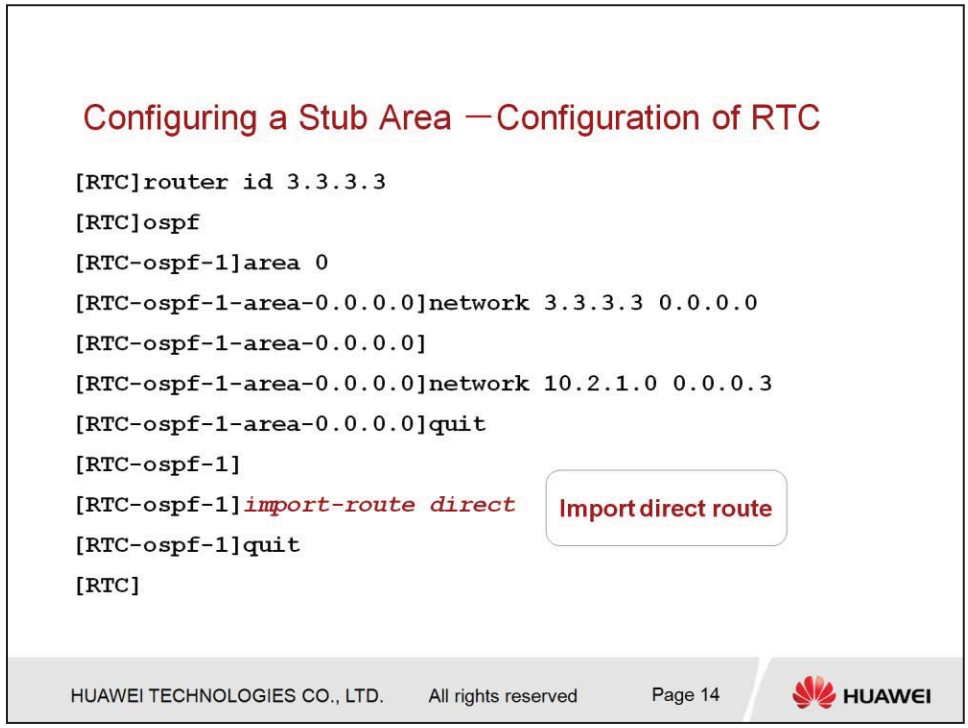
Importing direct route in OSPF on RTC , RTC is an ASBR , 30.1.1.1/32 and 30.1.1.2/32 are imported as external routes .

Configure area 1 as a stub area , so RTB can’t flood type 5 LSA into area 1 , only default route can be advertised into area 1.

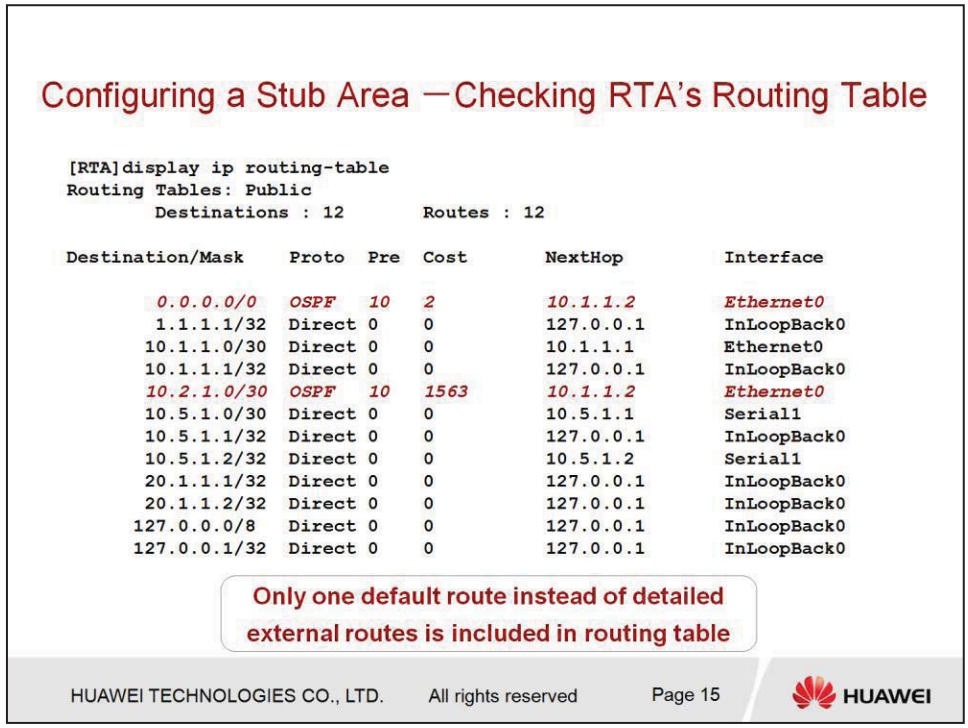


On RTA , all the networks are configured into area 1 , stub area must be configured on every router in the stub area .

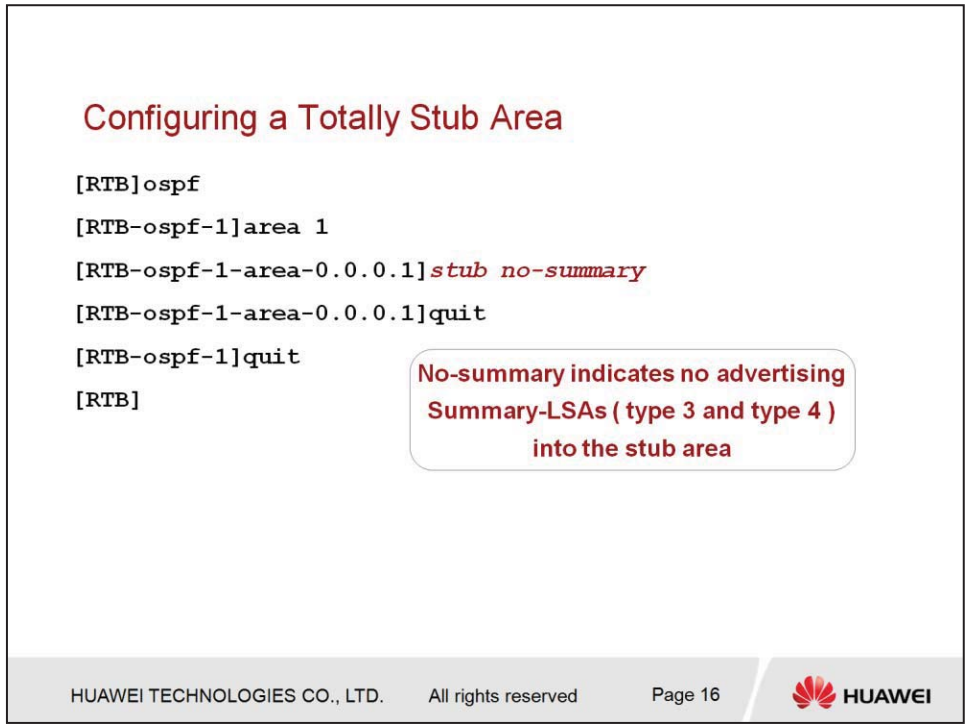
On RTB , configures area 1 as a stub area in area view .



On RTC , imports direct route .



In RTA's routing table , only default route and AS internal routes which advertised by RTB can be learned by OSPF , there is no AS external route in the routing table .



After configuring stub area on ABR (RTB), stub no-summary command is used to configure area 1 into a totally stub area .

After configuring the first area into totally stub area , the ABR inside that area will not generate and transmit type 3 , type 4 and type 5 LSA , so no inter-area routes or AS external routes can be advertised into a stub area .

Configuring a Totally Stub Area —Checking RTA's Routing Table

```
[RTA]display ip routing-table
Routing Tables: Public
        Destinations : 11       Routes : 11

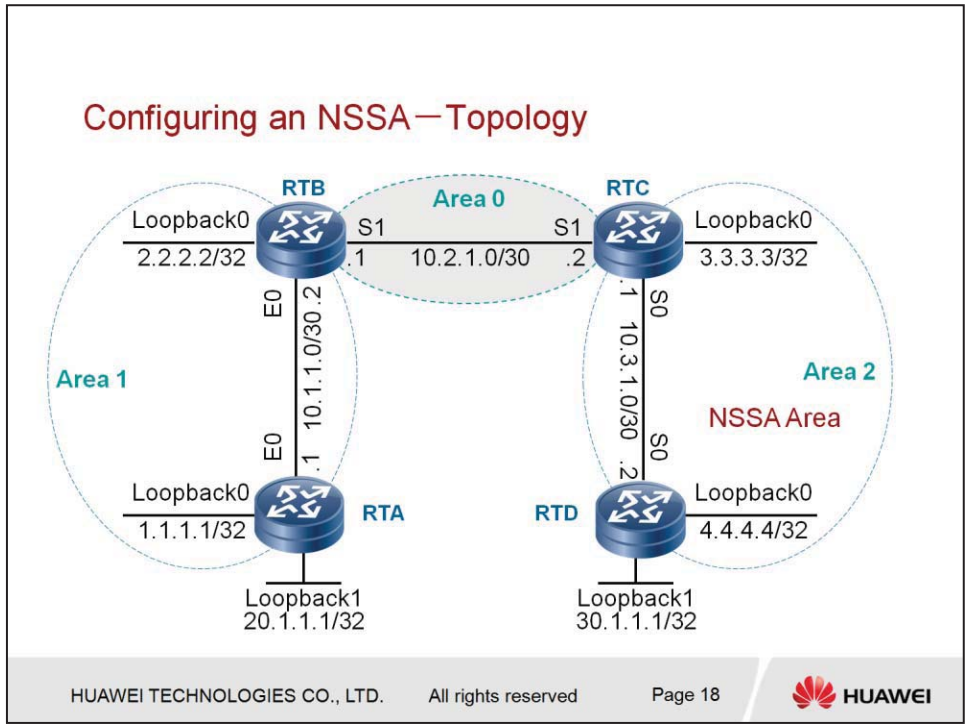
Destination/Mask    Proto   Pre  Cost      NextHop         Interface

        0.0.0.0/0   OSPF    10   2         10.1.1.2        Ethernet0
        1.1.1.1/32  Direct  0    0         127.0.0.1       InLoopBack0
       10.1.1.0/30  Direct  0    0         10.1.1.1        Ethernet0
       10.1.1.1/32  Direct  0    0         127.0.0.1       InLoopBack0
       10.5.1.0/30  Direct  0    0         10.5.1.1        Serial1
       10.5.1.1/32  Direct  0    0         127.0.0.1       InLoopBack0
       10.5.1.2/32  Direct  0    0         10.5.1.2        Serial1
       20.1.1.1/32  Direct  0    0         127.0.0.1       InLoopBack0
       20.1.1.2/32  Direct  0    0         127.0.0.1       InLoopBack0
      127.0.0.0/8   Direct  0    0         127.0.0.1       InLoopBack0
      127.0.0.1/32  Direct  0    0         127.0.0.1       InLoopBack0
```

Only one default route instead of detailed external routes or internal routes is included in routing table

HUAWEI TECHNOLOGIES CO., LTD. All rights reserved HUAWEI

On RTA's (totally stub area router's)routing table , only a default route and the routes within it's own area exist in OSPF routing table, there is no external route or inter-area route .

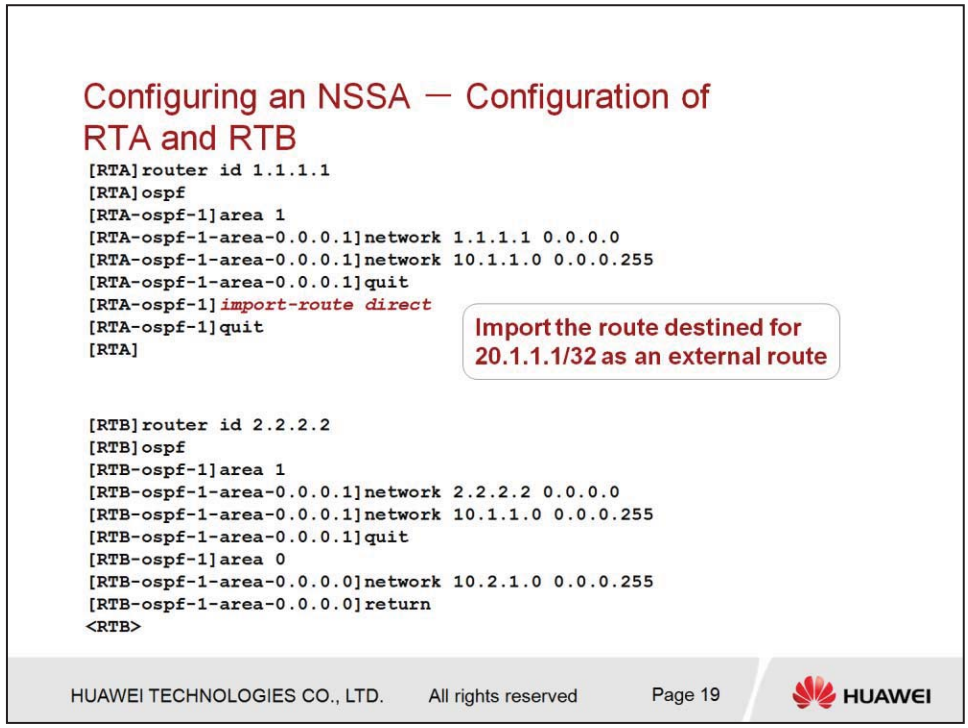


In this case :

IP addresses of loopback interface are Imported on RTA and RTD , so RTA and RTD became ASBRs .

Configure area 2 into NSSA. Enable RTC to advertize default route in NSSA area, it is necessary to import external route on RTD, and flood the external route to entire AS, but the external route imported from RTA can’t be flooded into this area .

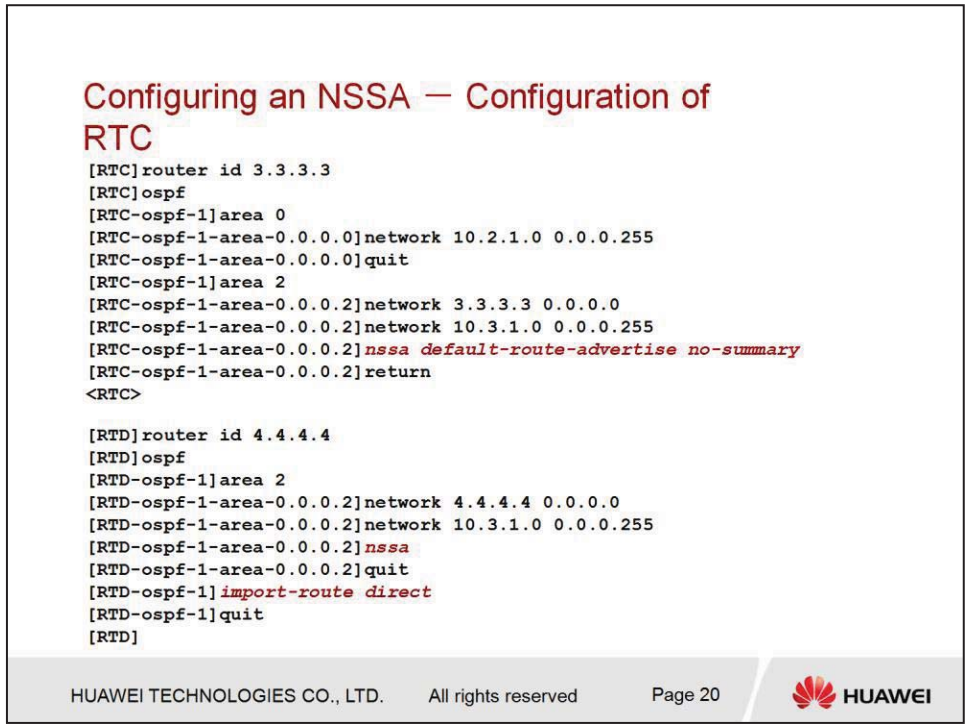
Configure area 1 as a normal area , so the both RTA and RTD's external route can be flooded into area 1.



On RTA , configure 1.1.1.1/32 and 10.1.1.0/24 in area 1 .

Import direct route 20.1.1.1/32 as an AS-external route .

Configure RTB as an ABR.



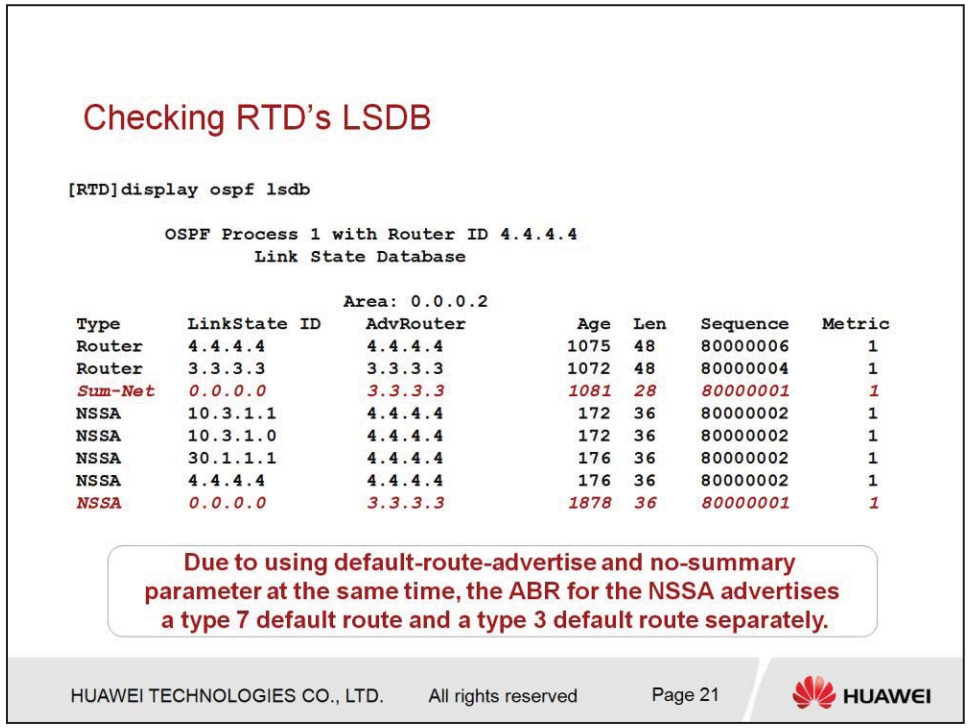
Configuring area 2 as a NSSA ,RTC is an ABR of NSSA (Area 2). Due to type 5 LSA can't be advertised into this area , it is necessary to advertise a default route for ABR of NSSA. On RTD , configure area 2 into NSSA , and import direct route 30.1.1.1/32 as external route .

nssa [default-route-advertise] [no-import-route] [no-summary]

default-route-advertise：Specifies generating default Type-7 LSA to the NSSA area. Only can be used on ABR or ASBR in NSSA , after configuration , if it's an ABR , no matter default route exist or not in the routing table , a default route will always be advertised into area through Type-7 LSA ; if it's an ASBR , only default route in the routing table , an default route can be advertised into area through Type-7 LSA .

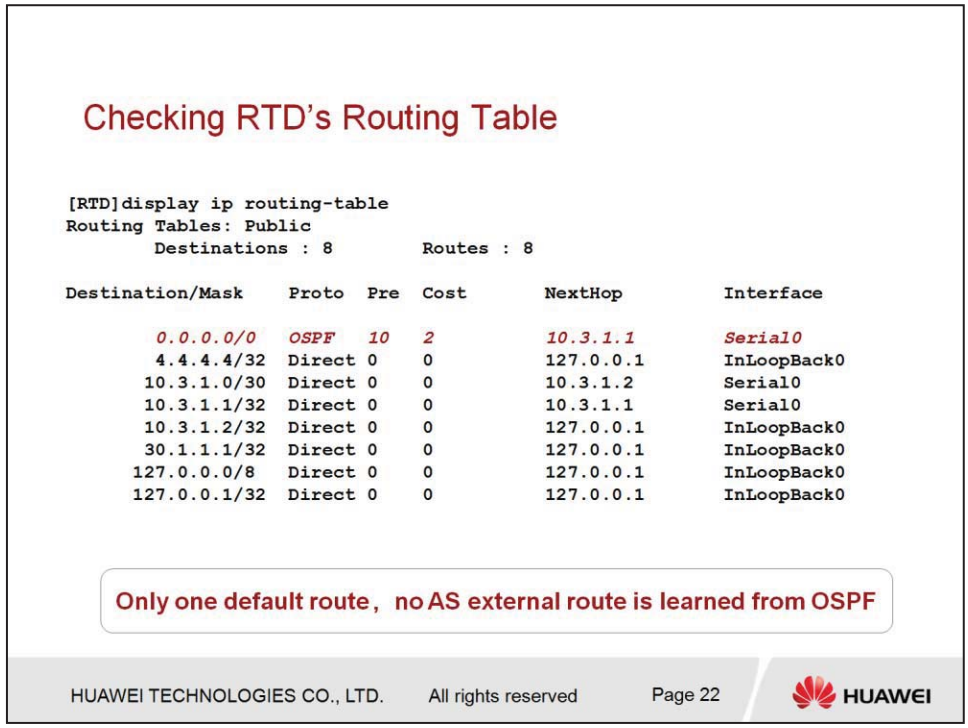
no-import-route：Specifies to disable import route to the NSSA area. only can be used on the router , which is an ABR of NSSA , also an ASBR of OSPF domain , ensure all the external route can get into OSPF domain .

no-summary：Specifies ABR is disabled to transmit type 3 and type 4 LSAs to the NSSA area. After using that parameter ABR will generate a default route into NSSA with type 3 LSA .



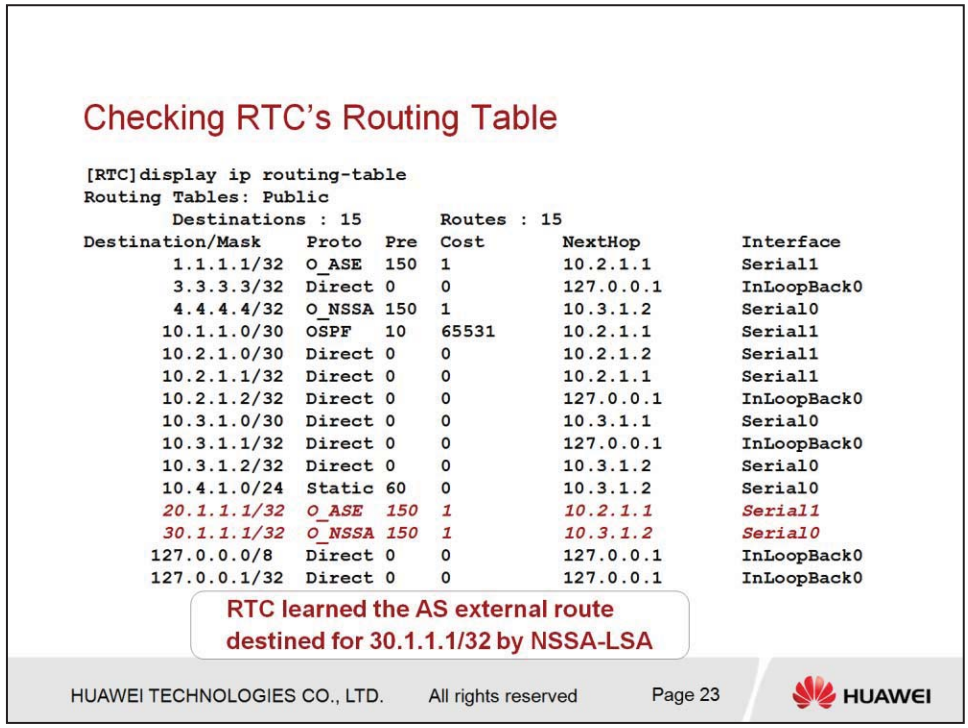
Due to using default-route-advertise and no-summary parameter at the same time, ABR of NSSA advertises type 7 default route and type 3 default route separately.

In this case, type 3 default route is selected. (type 7 default route is considered as an AS external route with the precedence 150).



Besides default route , there is no type 3 , type 4 and type 5 LSA advertised into this area , so there is no detail external routing information learnt by OSPF .

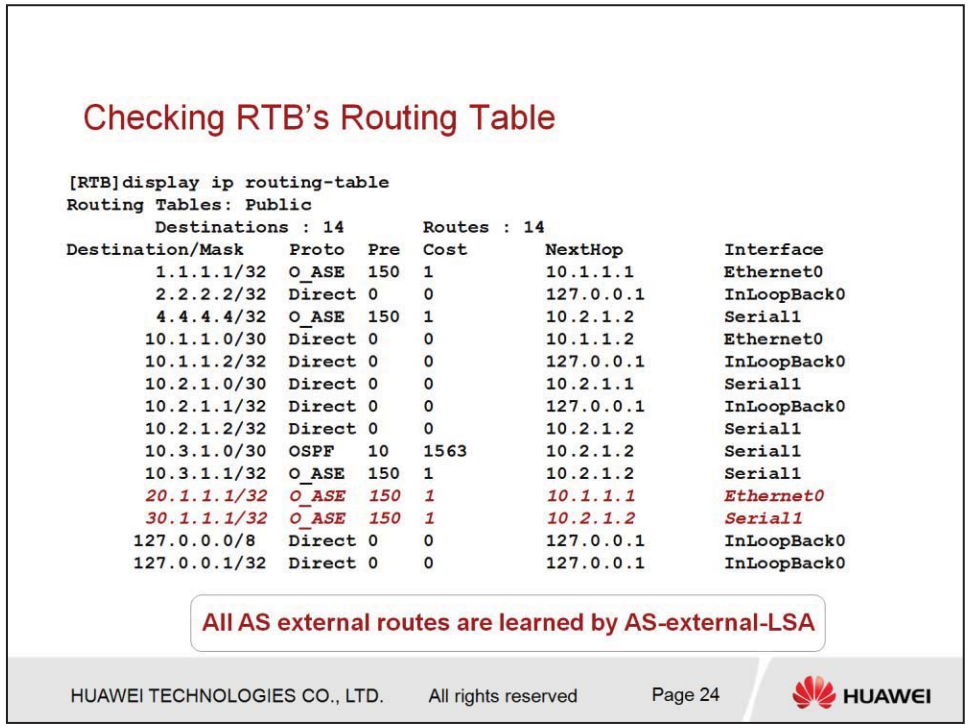
Neither imported route to 20.1.1.1/32 on RTA nor inter-area routes is learnt .



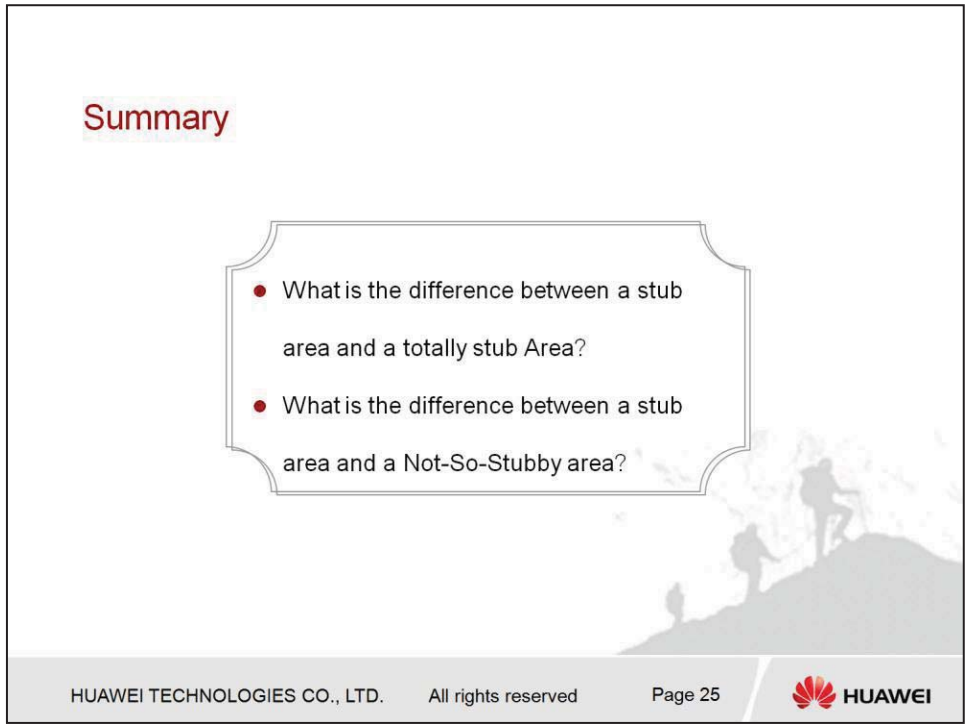
On RTC's routing table :

30.1.1.1/32 is learnt from type 7 LSA which advertised by RTD .

20.1.1.1/32 is learnt from type 5 LSA which advertised by RTA .



ABR(RTC) of NSSA , convert type 7 LSA into type 5 LSA , so in RTB's routing table , all the external route is learnt through type 5 LSA , including imported route to 20.1.1.1/32 on RTA and imported route to 30.1.1.1/32 on RTD .



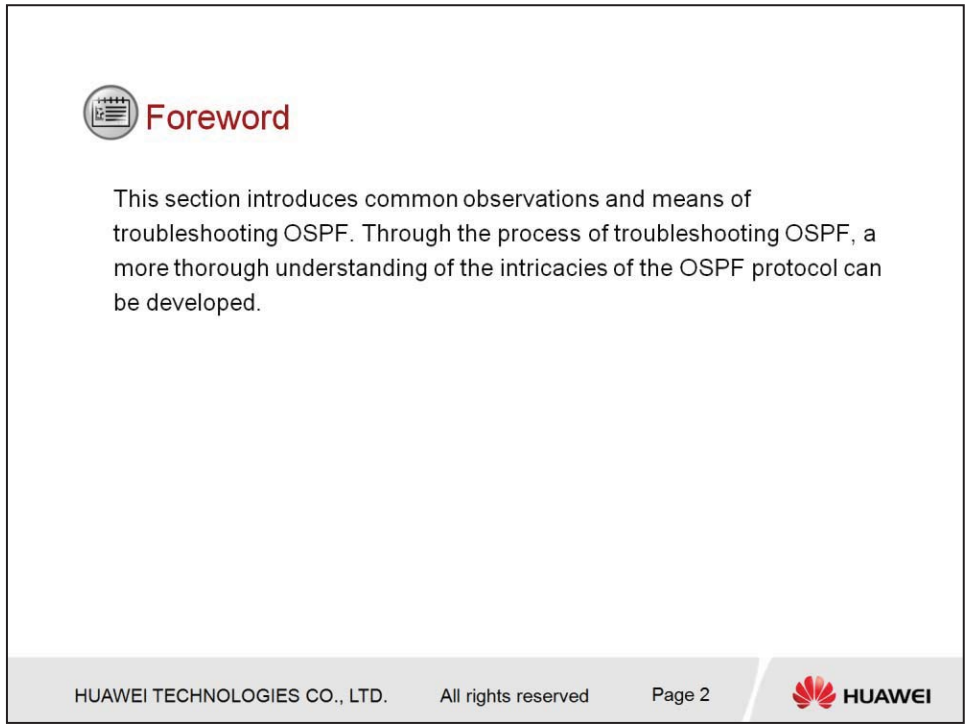
What's the difference between Stub Area and Totally Stub Area?

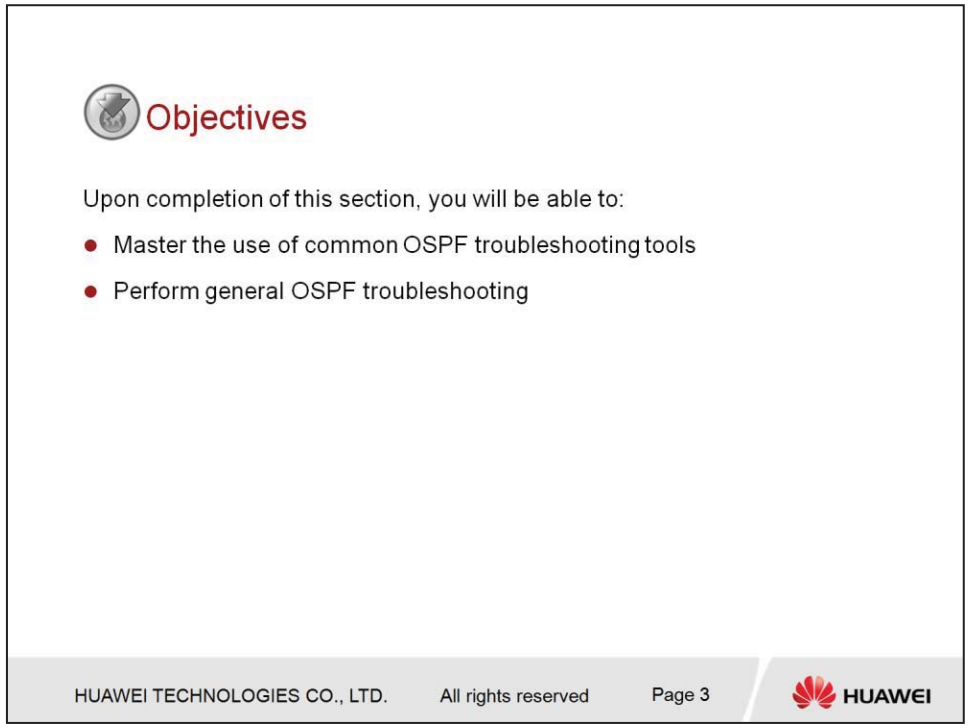
Stub area can flood type 3 , totally stub area can't flood type 3 and type 4 LSA (except default route)

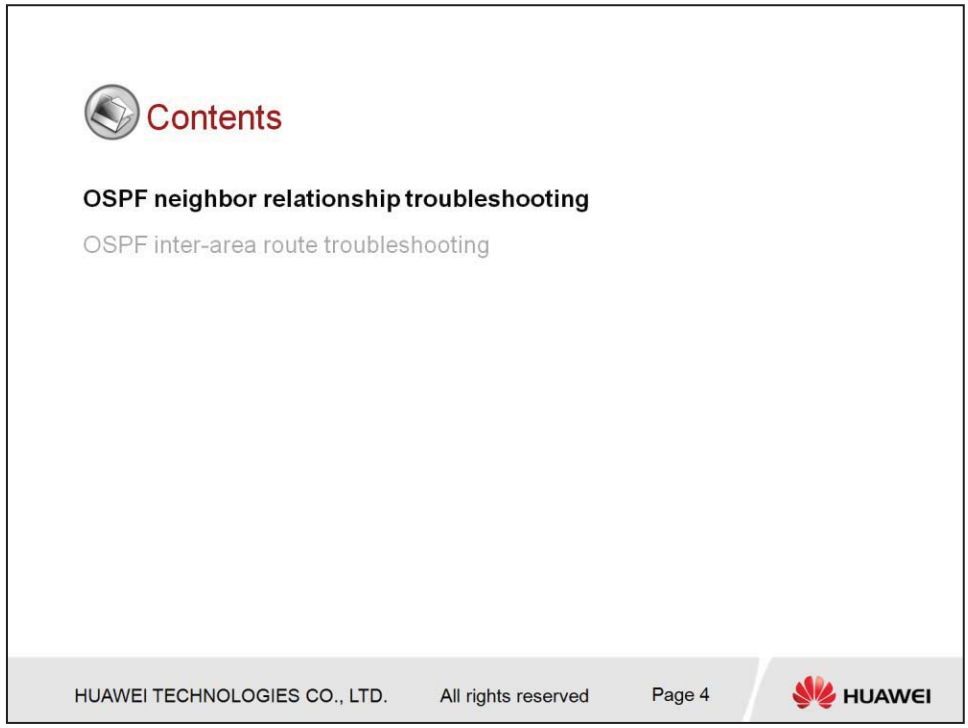
What's the difference between Stub Area and NSSA Area?

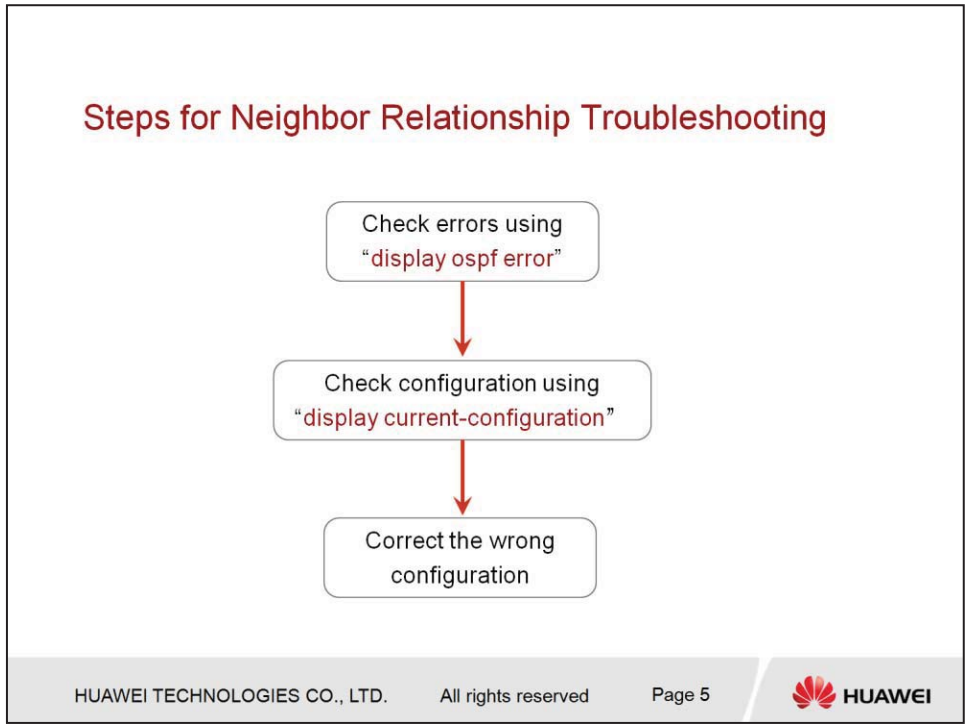
Stub area can't import external route , NSSA area can import external route .


OSPF Troubleshooting
www.huawei.com
HUAWEI TECHNOLOGIES CO., LTD.
All rights reserved
HUAWEI


Foreword
This section introduces common observations and means of troubleshooting OSPF. Through the process of troubleshooting OSPF, a more thorough understanding of the intricacies of the OSPF protocol can be developed.
HUAWEI TECHNOLOGIES CO., LTD.
All rights reserved
Page 2
HUAWEI

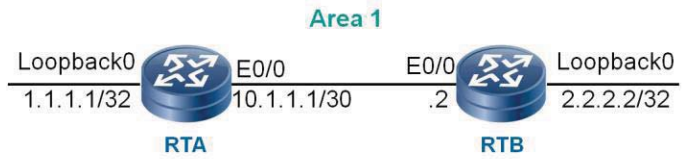

Objectives
Upon completion of this section, you will be able to:
Master the use of common OSPF troubleshooting tools
Perform general OSPF troubleshooting
HUAWEI TECHNOLOGIES CO., LTD.
All rights reserved
Page 3
HUAWEI


Contents
OSPF neighbor relationship troubleshooting
OSPF inter-area route troubleshooting
HUAWEI TECHNOLOGIES CO., LTD.
All rights reserved
Page 4
HUAWEI



Failures within area are mainly caused by neighbors relationship. It is not difficult to deal with that problem. Firstly, “display OSPF error” command is used to show errors information; Secondly, “display current-configuration” is used to show configuration of OSPF; Lastly modify wrong configurations.

Problem Description—No Intra-Area Routes



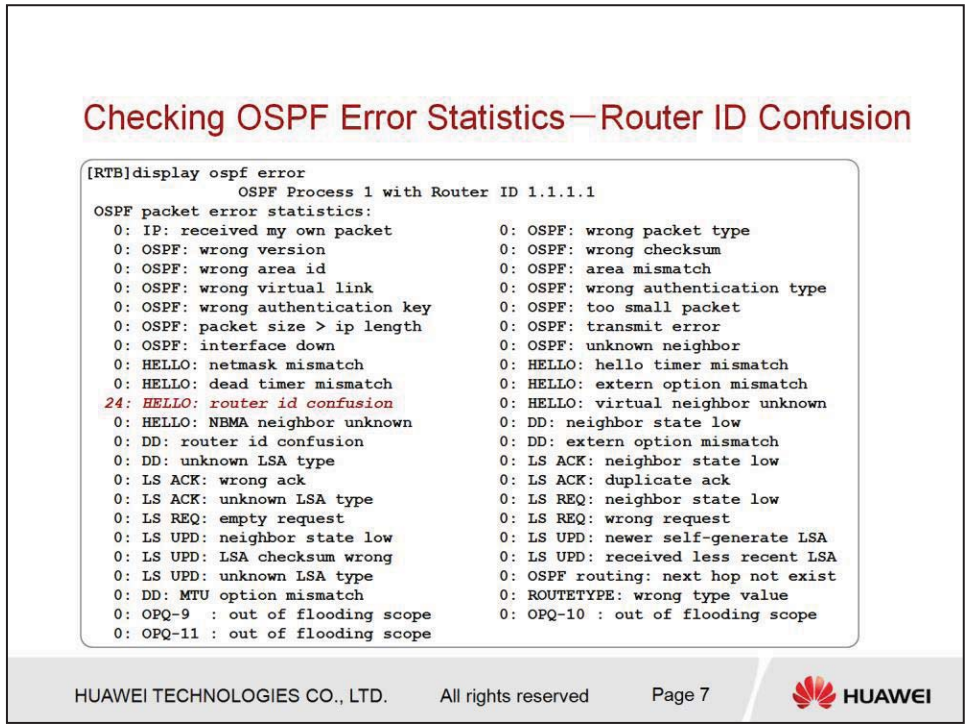

```
[RTB]display ip routing-table
 Routing Table: public net
Destination/Mask    Protocol  Pre  Cost      Nexthop         Interface
2.2.2.2/32          DIRECT    0    0         127.0.0.1       InLoopBack0
10.1.1.0/24         DIRECT    0    0         10.1.1.2        Ethernet0/0
10.1.1.2/32         DIRECT    0    0         127.0.0.1       InLoopBack0
127.0.0.0/8         DIRECT    0    0         127.0.0.1       InLoopBack0
127.0.0.1/32        DIRECT    0    0         127.0.0.1       InLoopBack0
```

HUAWEI TECHNOLOGIES CO., LTD. All rights reserved HUAWEI

Two routers are directly connected , all the networks are configured within one area .

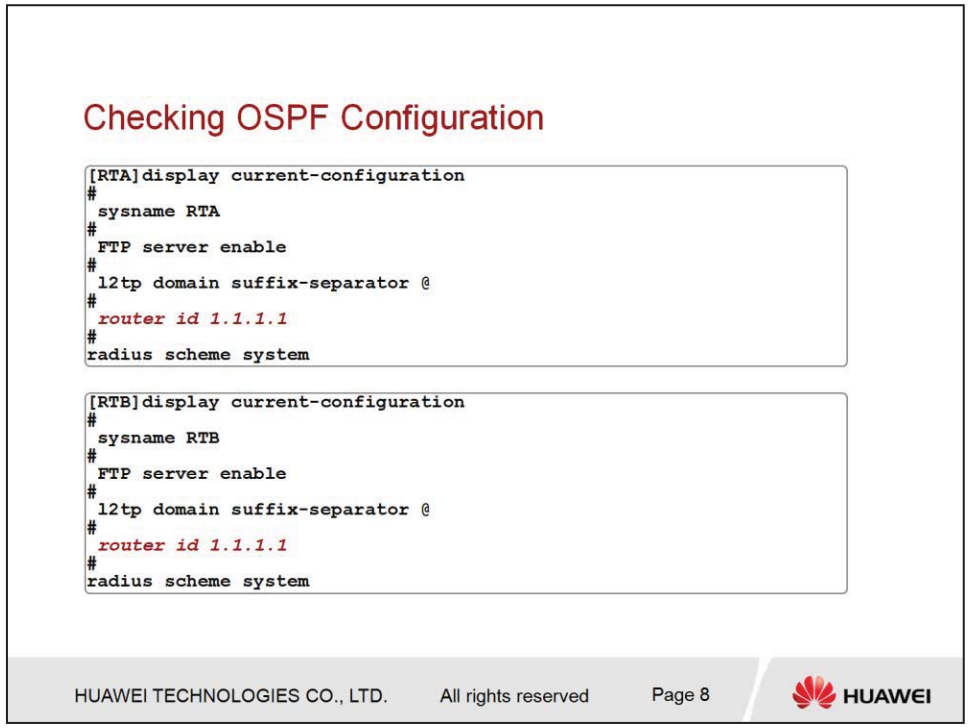
After completion of configuration , check the routing table of RTB , there is no routing information of RTA's loopback address .

This failure is usually caused by the relationship of neighbors , the main way is to check the error information of OSPF , check configurations according to the error information , and correct the wrong configurations .



"Router id confusion"is found through "display ospf error " command .

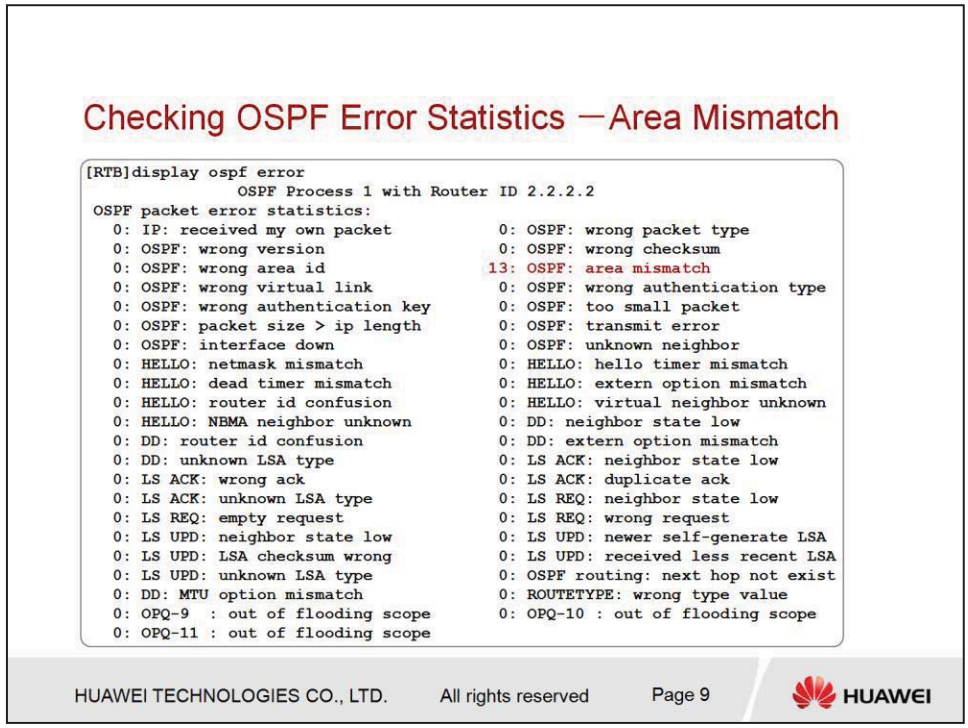
OSPF requires all the routers must have their own unique router id, which can’t be overlapped .



Check the configuration of RTA and RTB , both the router id are configured as 1.1.1.1 .

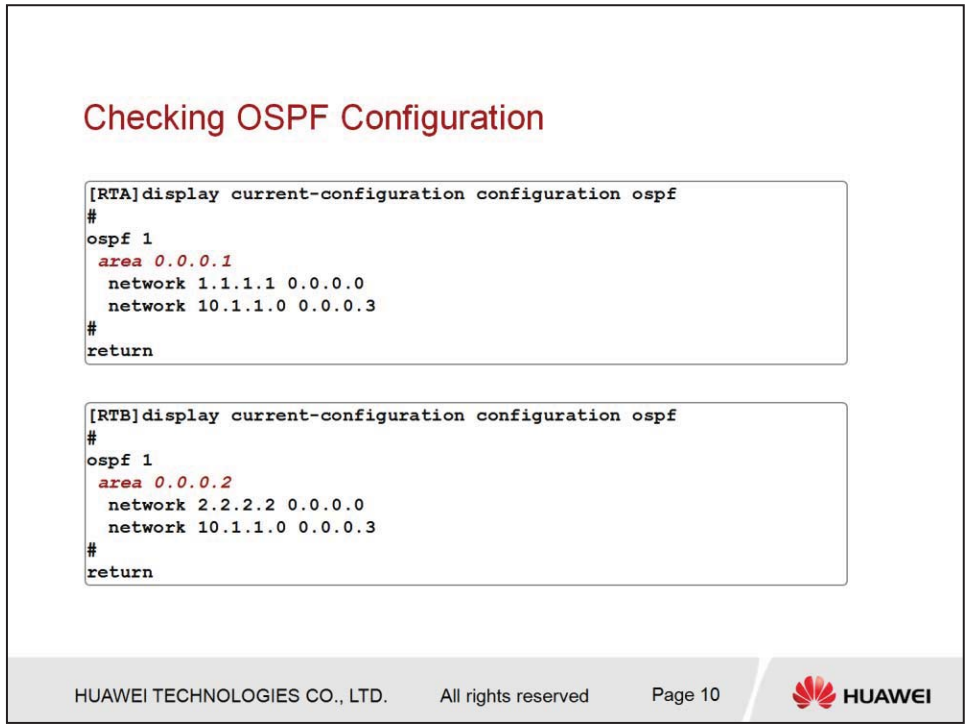
Change RTB's router id into 2.2.2.2 (loopback address) , the failure is

corrected .

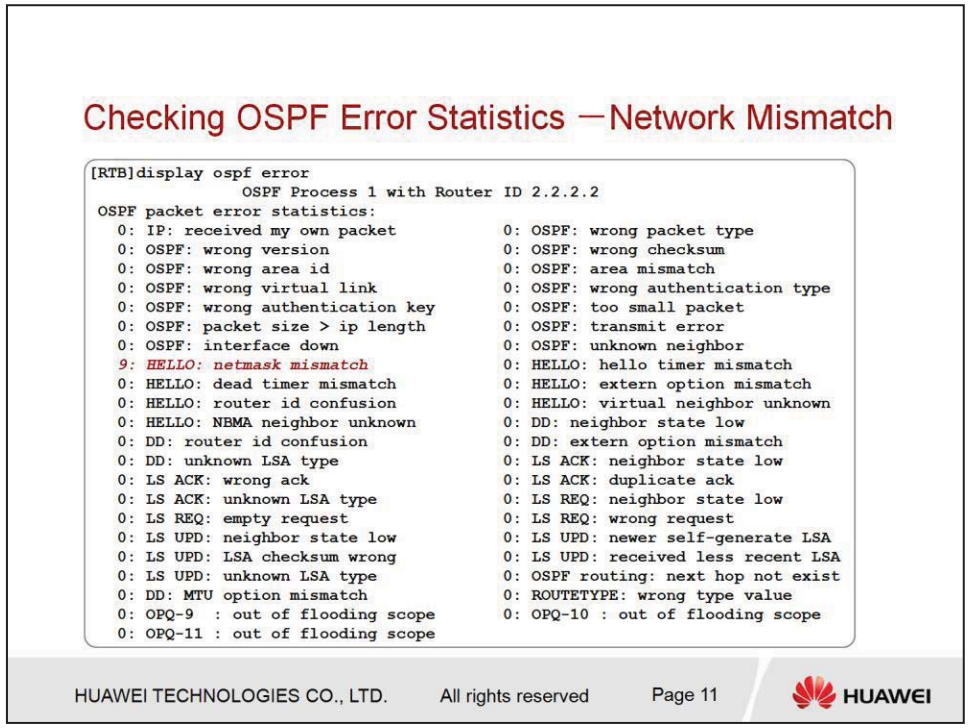


"Area mismatch" is found through "display ospf error" command .

OSPF prescribes that the area is a group of network , all the interface must be configured into one area , otherwise the neighbourship can’t be built up .

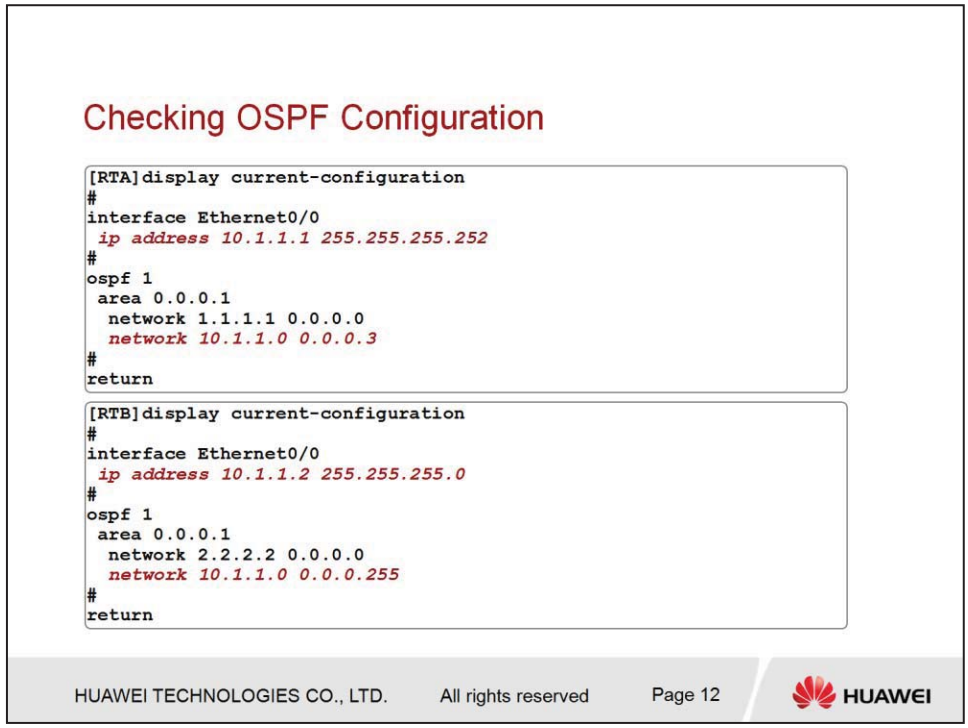


After check the configuration of OSPF , two interfaces are configured into different areas ,after change the configuration of RTB , all the networks are configured into area 1 , the failure is corrected .



"Netmask mismatch" is found through "display ospf error " command .

OSPF prescribes the network mask must be the same in broadcast , NBMA , Pto- MP network , otherwise the neighbourship can’t be built up.



Checking the configuration of OSPF , the network mask length is 30 on RTA E0/0 , however RTB's network mask length is 24 , if the network mask length mismatches , the neighbouring can’t be built up .

Change the network mask length into 30 bit , change the wild mask into 0.0.0.3 , the failure is corrected .

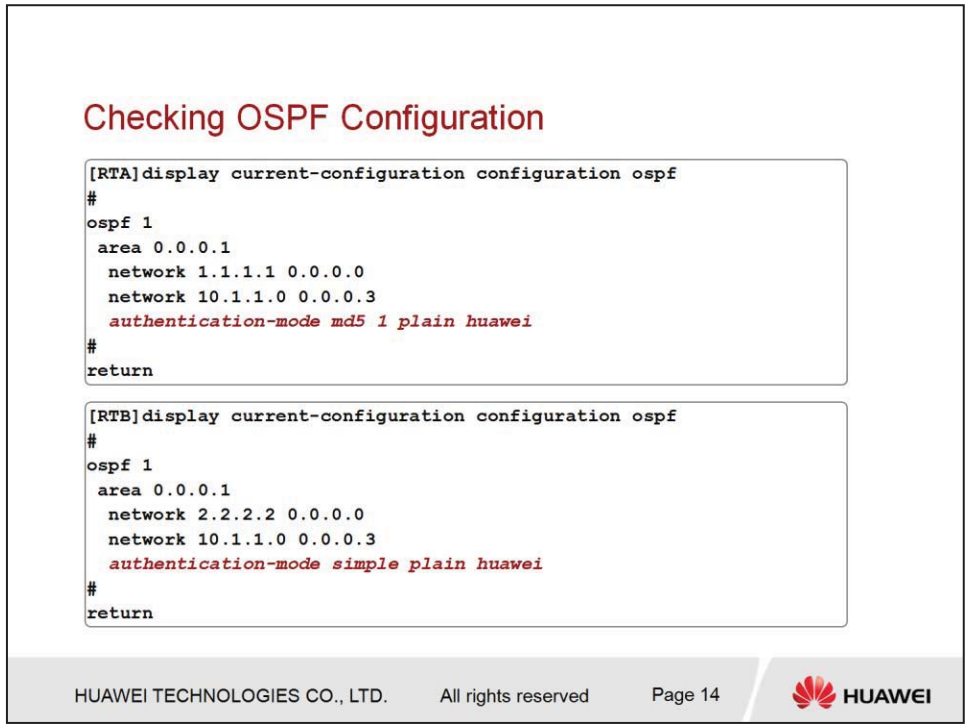
Checking OSPF Error Statistics —Wrong Authentication Type

```
[RTB]display ospf error
                OSPF Process 1 with Router ID 2.2.2.2
 OSPF packet error statistics:
   0: IP: received my own packet           0: OSPF: wrong packet type
   0: OSPF: wrong version                  0: OSPF: wrong checksum
   0: OSPF: wrong area id                  0: OSPF: area mismatch
   0: OSPF: wrong virtual link             2: OSPF: wrong authentication type
   0: OSPF: wrong authentication key       0: OSPF: too small packet
   0: OSPF: packet size > ip length        0: OSPF: transmit error
   0: OSPF: interface down                 0: OSPF: unknown neighbor
   0: HELLO: netmask mismatch              0: HELLO: hello timer mismatch
   0: HELLO: dead timer mismatch           0: HELLO: extern option mismatch
   0: HELLO: router id confusion           0: HELLO: virtual neighbor unknown
   0: HELLO: NBMA neighbor unknown         0: DD: neighbor state low
   0: DD: router id confusion              0: DD: extern option mismatch
   0: DD: unknown LSA type                 0: LS ACK: neighbor state low
   0: LS ACK: wrong ack                    0: LS ACK: duplicate ack
   0: LS ACK: unknown LSA type             0: LS REQ: neighbor state low
   0: LS REQ: empty request                0: LS REQ: wrong request
   0: LS UPD: neighbor state low           0: LS UPD: newer self-generate LSA
   0: LS UPD: LSA checksum wrong           0: LS UPD: received less recent LSA
   0: LS UPD: unknown LSA type             0: OSPF routing: next hop not exist
   0: DD: MTU option mismatch              0: ROUTETYPE: wrong type value
   0: OPQ-9  : out of flooding scope       0: OPQ-10 : out of flooding scope
   0: OPQ-11 : out of flooding scope
```

HUAWEI TECHNOLOGIES CO., LTD. All rights reserved Page 13 HUAWEI

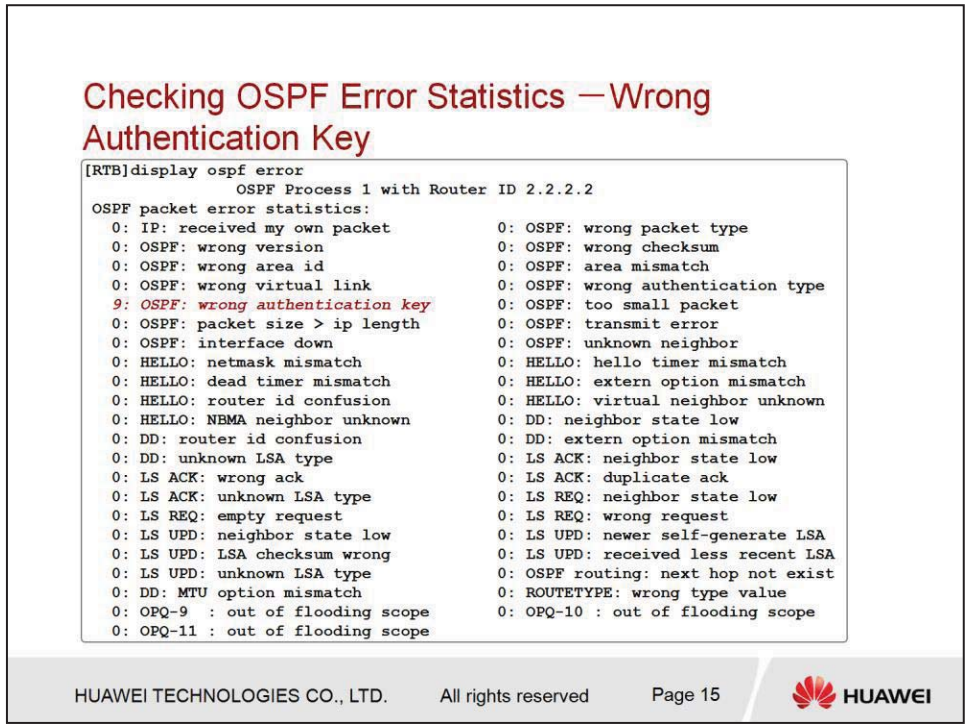
"Wrong authentication type" is found through "display ospf error " command .

Same authentication type must be used with in one area ; Configuring authentication type in area view .



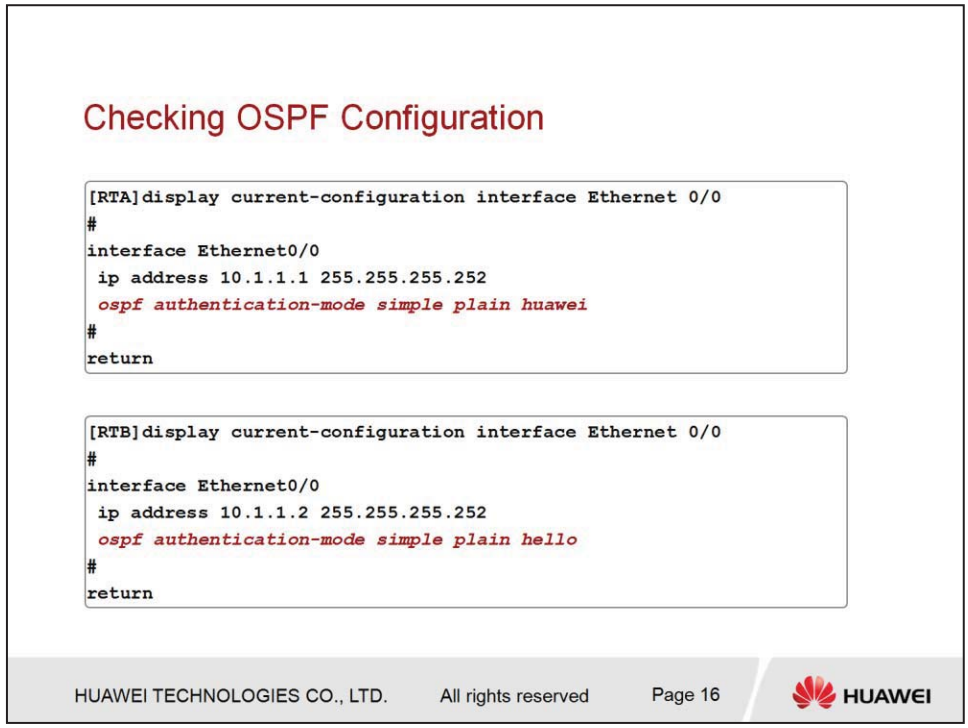
Checking the configuration of OSPF , RTA uses MD5 as its authentication method , RTB uses clear text as its authentication method , authentication method mismatch , the neighbourship can’t be built up .

Change the authentication method of RTA into "simple".



Changing the authentication method , routing information also can’t be learnt , check the OSPF error information . "wrong authentication key" is found .

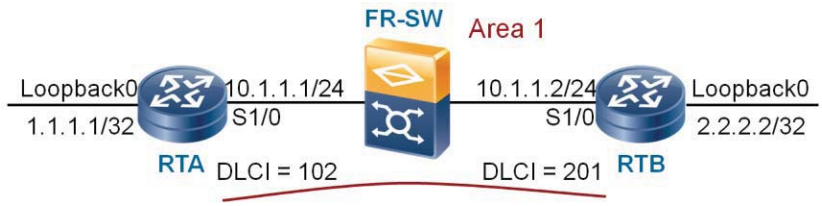
Password must be configured on the interface , OSPF prescribes the password must be same on different sides of the link .



After check the configuration , password mismatch is found .

Change the password into “huawei” on both sides , the failure is corrected .

Problem Description—No Routes Learned on NBMA Network




```
[RTA]display ip routing-table
 Routing Table: public net
Destination/Mask    Protocol  Pre  Cost      Nexthop         Interface
1.1.1.1/32          DIRECT    0    0         127.0.0.1       InLoopBack0
10.1.1.0/24         DIRECT    0    0         10.1.1.1        Serial1/0
10.1.1.1/32         DIRECT    0    0         127.0.0.1       InLoopBack0
10.1.1.2/32         DIRECT    0    0         10.1.1.2        Serial1/0
127.0.0.0/8         DIRECT    0    0         127.0.0.1       InLoopBack0
127.0.0.1/32        DIRECT    0    0         127.0.0.1       InLoopBack0
```

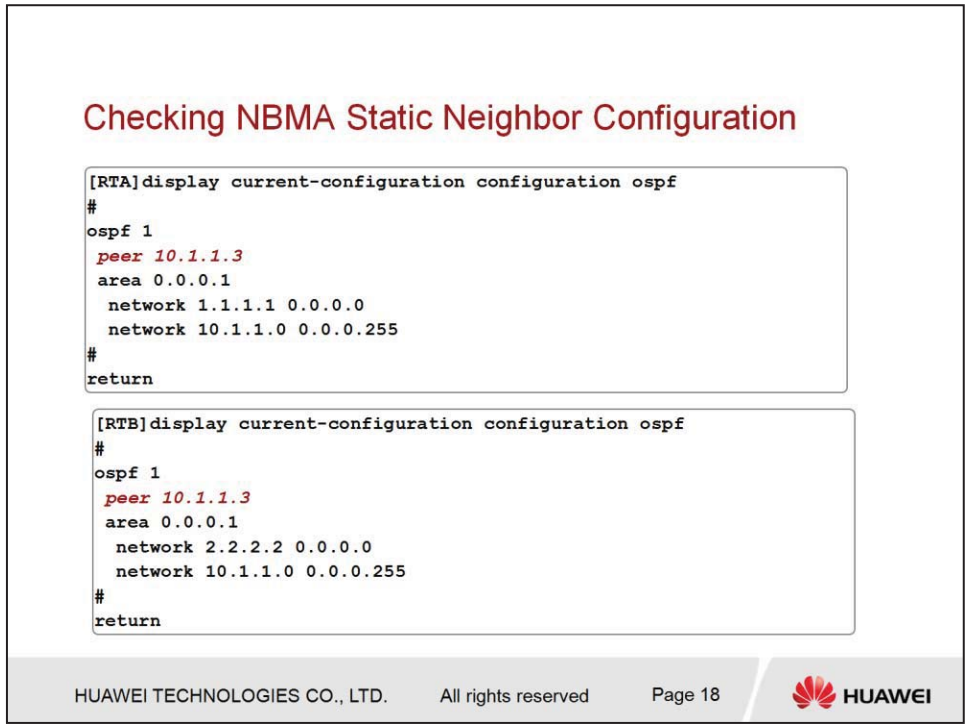
HUAWEI TECHNOLOGIES CO., LTD. All rights reserved HUAWEI

RTA and RTB are connected through frame relay switch , the network type is NBMA .

After checking the routing table , RTA can’t learn routes through OSPF .

Analysis of the routing table shows the local interface's information and remote interface’s information exist , which means frame relay and IP address are correctly configured, they can make a communication.

Neighbors on NBMA network can’t be auto discovered, only through static manually configuring. In this case, it is necessary to designate neighbor manually.



Check the configuration of NBMA static neighbor , IP address error is found .

Use 10.1.1.2 as RTA's remote neighbor identifier ; Use 10.1.1.1 as RTB‘s remote neighbor identifier .

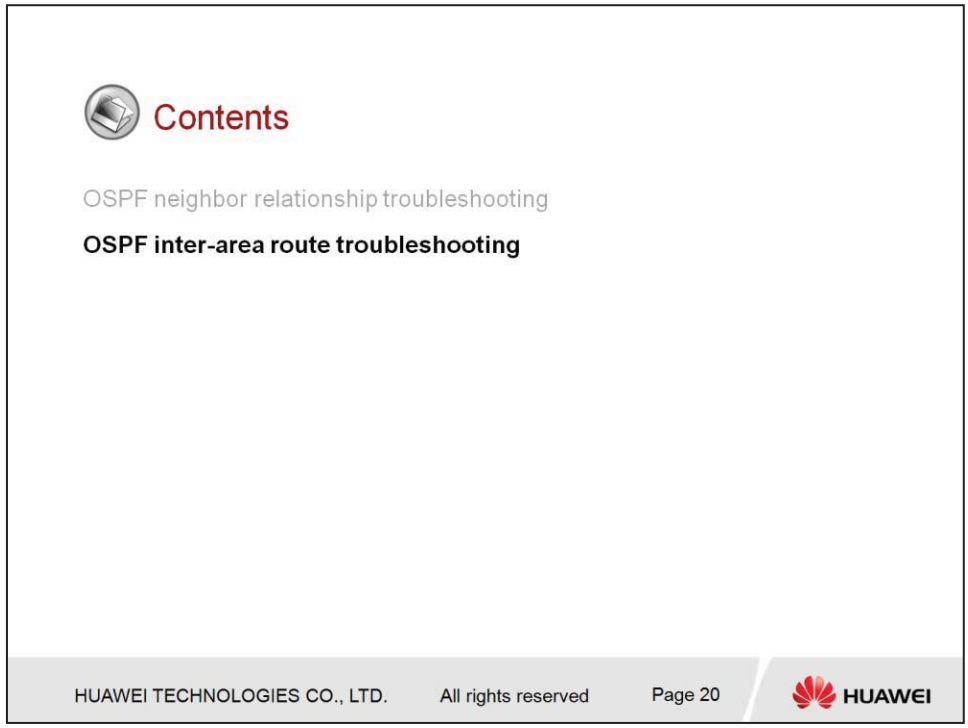
Configure a static neighbor, the failure is corrected.

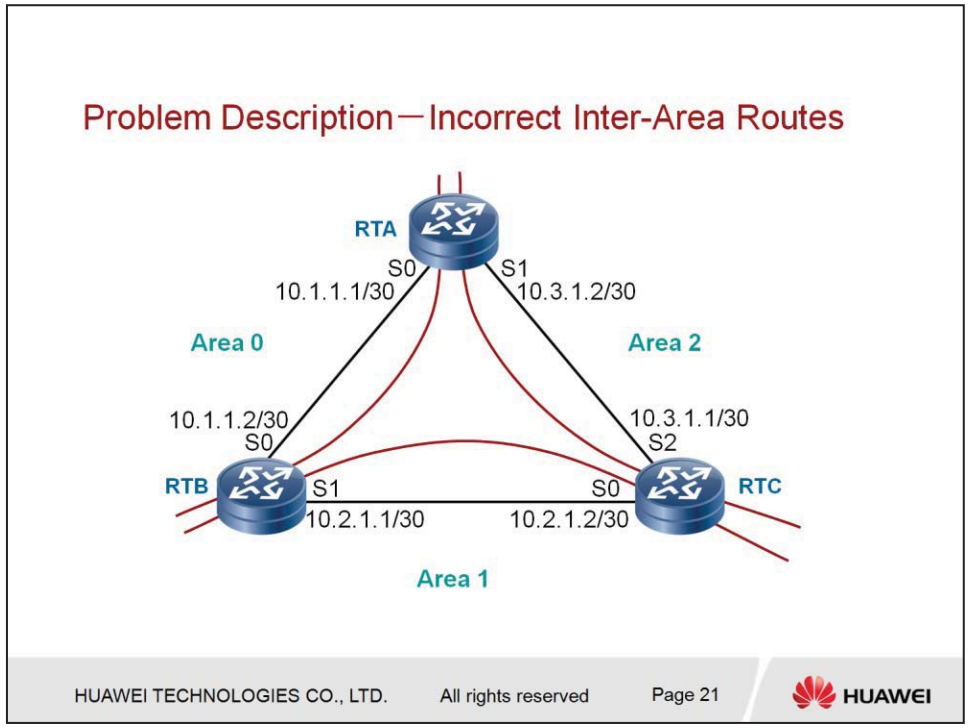
Summary for Neighbor Relationship Troubleshooting

Parameter	Requirement
router id	Each router's router id must be unique
area id	All the interfaces on the same network must belong to the same area
network mask	All the interfaces on the same network must have the same network mask, except P2P network.
authentication type	Authentication mode must match in the same area
authentication data	Authentication key must match on the same network
extern option	When configuring stub or NSSA, the command must be configured on all routers in the area
peer	Neighbors on NBMA network must be manually configured

HUAWEI TECHNOLOGIES CO., LTD. All rights reserved

This slide makes a summary of the neighbor failures , P-to-P network don't need to consider the network mask , manually configured peer only exists on NBMA network , other configurations are suitable for all network types .

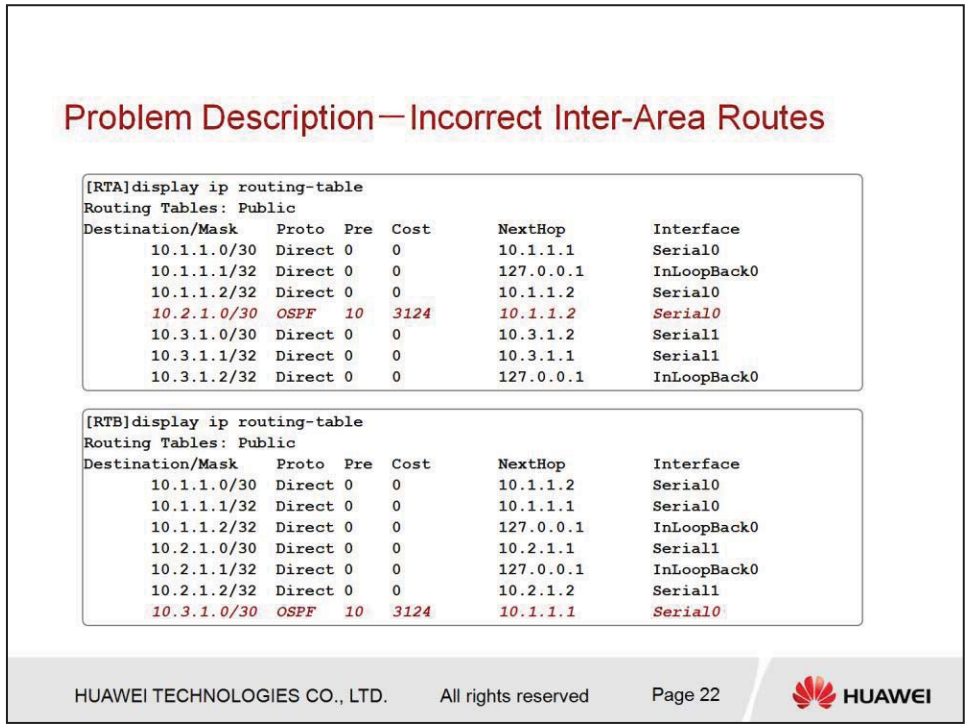

Contents
OSPF neighbor relationship troubleshooting
OSPF inter-area route troubleshooting
HUAWEI TECHNOLOGIES CO., LTD.
All rights reserved
Page 20
HUAWEI



In this case :

These three routers connect with each other through serial link , RTA uses 1.1.1.1 as its router id ; RTB uses 2.2.2.2 as its router id ; RTC uses 3.3.3.3 as its router id .

Network 10.1.1.0/30 is in area 0 , Network 10.2.1.0/30 is in area 1, Network 10.3.1.0/30 is in area 2 , all links use the same bandwidth , so each router has two equivalent path to the opposite network .

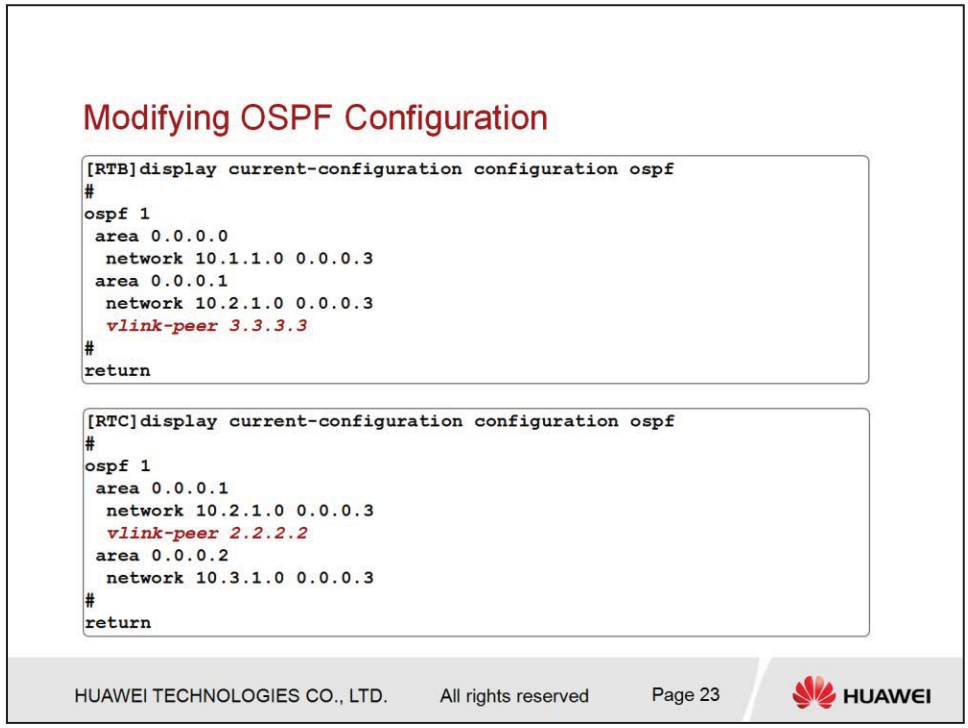


In RTA's routing table , there is only one route to 10.2.1.0/30 through backbone area ,with the next-hop is 10.1.1.2.

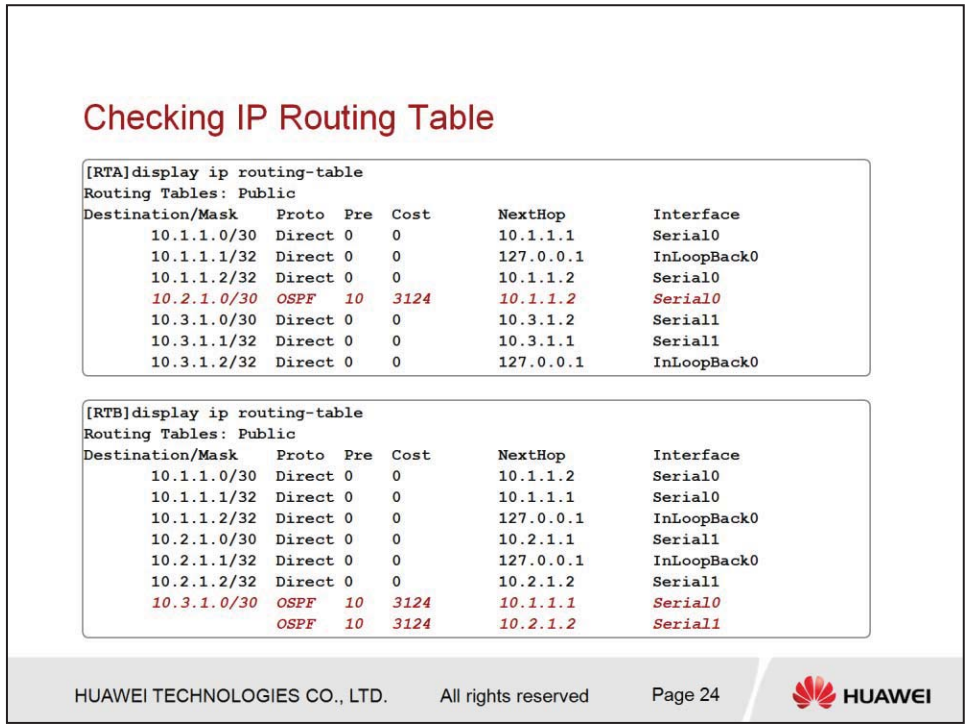
Trouble analysis:

Routing information inter-areas only can be advertised through area 0 , can’t be advertised between non-backbone area , thus , RTC can’t advertise the routing information 10.2.1.0/30 to RTA directly , so only area 1's ABR（RTB）can advertise the routing information 10.2.1.0/30 to RTA , in RTA's routing table , with the next-hop only 10.1.1.2 (RTB).

In RTB's routing table , the routing information reach to 10.3.1.0/30 has only one next-hop 10.1.1.1 (RTA).



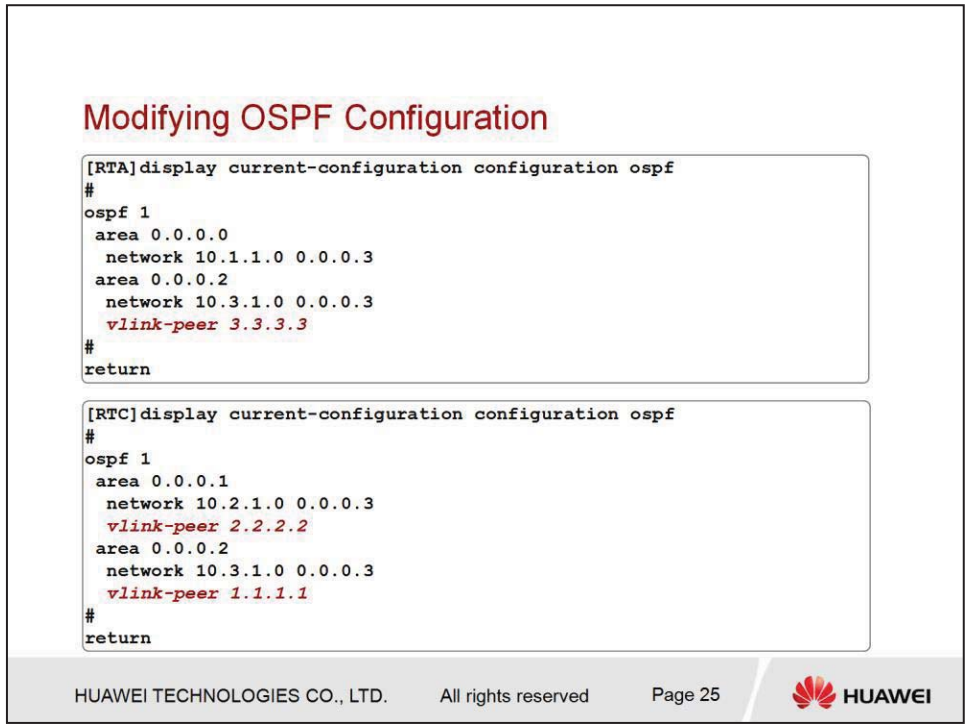
Configure a virtual link between RTB and RTC , virtual link use area 1 as transit area , the aim of configure virtual link is to make RTC connect to backbone area , so that , RTC can advertise routing information to backbone area .



After change the configuration , check RTA and RTB's routing table ,RTB‘s routing table is correct , in RTB's routing table , there are two next-hops to 10.3.1.0/30, which means equivalent route . But there is only one next-hop to 10.2.1.0/30 on RTA .

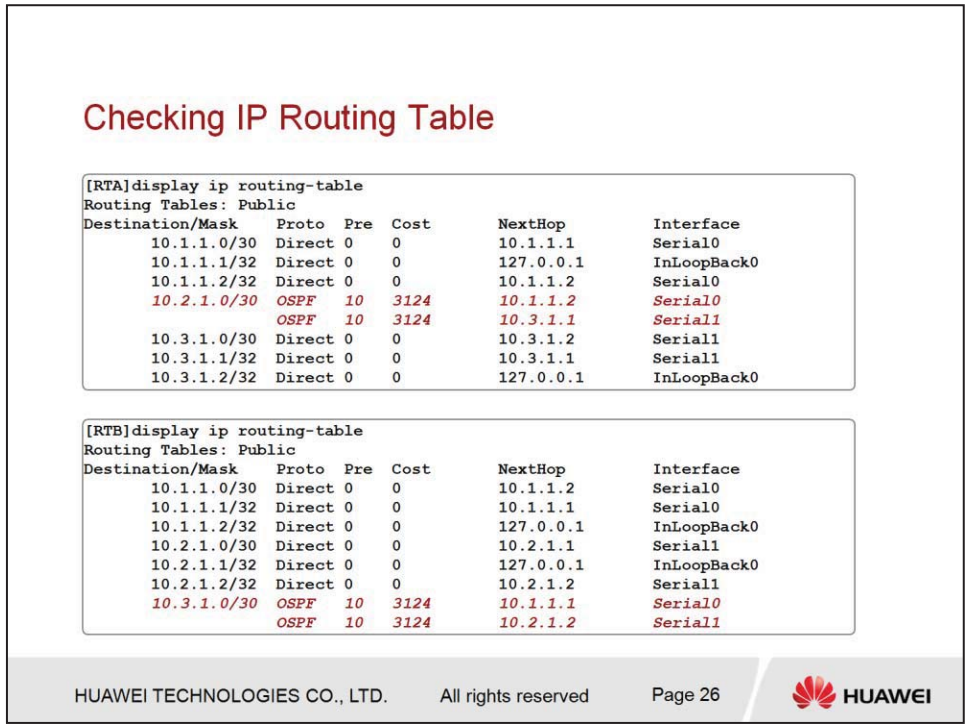
Trouble analyse:

After configure virtual link between RTB and RTC , area 1 is transit area , so RTC only can advertise area 2's routing information to RTB through area 1 , RTC can not advertise area 1's routing information to RTA through area 2 , so in RTA ‘s routing table , there is only one next-hop to 10.2.1.0/30 .



Configure a virtual link between RTA and RTC , the transit area of virtual link is area 2 .

Virtual link is use to make RTC connect to backbone area , RTC can advertise area 1 and area 2's information to backbone area at the same time .



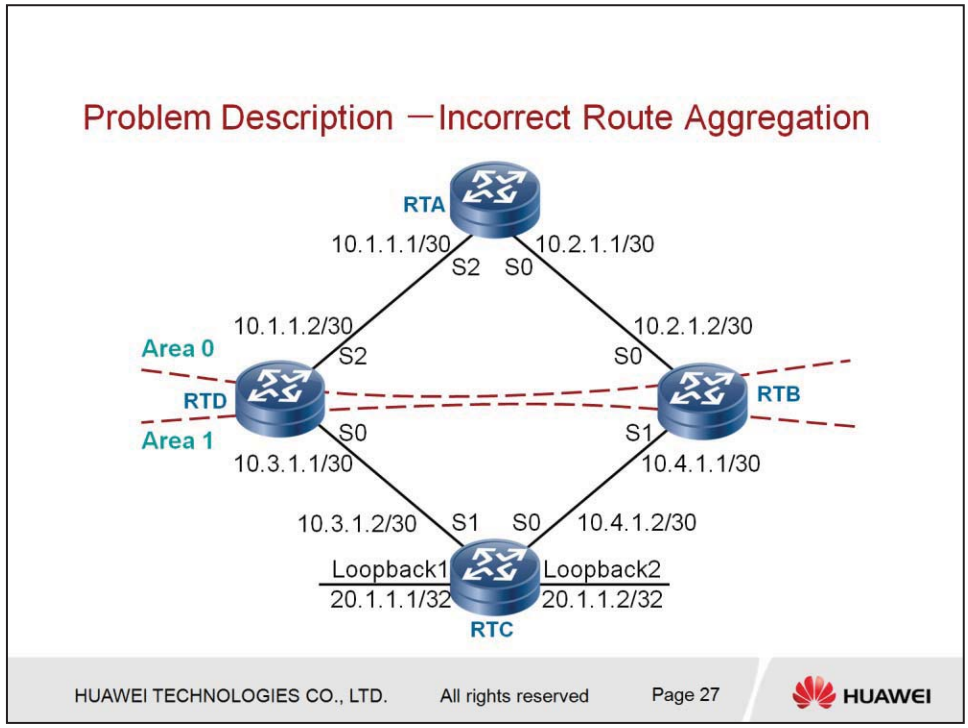
Check RTA and RTB's routing table again , route item is correct , the equivalent route exist .

Summary :

When configure OSPF areas , if a router connect more than two areas , one of them must be area 0 , this link can be physical link , also can be virtual link .

Make sure ABR advertise all the route directly connect from non-backbone area to backbone area .

When build a network , avoid to use virtual link , use backbone area as hub node , non-backbones area as spoke node , don’t make a connection between nonbackbone areas .



In this case :

All the networks on RTC are configured into area 1. Configuring route aggregation on RTB, aggregating 20.1.1.1/32 and 20.1.1.2/32 into 20.1.1.0/24, when RTB advertises routing information to area 0 , only advertises 20.1.1.0/24 , detail routing information 20.1.1.1/32 and 20.1.1.2/32 are restrained .

RTA 's router id is 1.1.1.1 ; RTB 's router id is 2.2.2.2 ; RTC 's router id is 3.3.3.3 ; RTD 's router id is 4.4.4.4 .

Problem Description —Incorrect Route Aggregation

```
[RTA]display ip routing-table
Routing Tables: Public
        Destinations : 13        Routes : 13

Destination/Mask    Proto   Pre  Cost      NextHop         Interface

       10.1.1.0/30  Direct  0    0         10.1.1.1        Serial2
       10.1.1.1/32  Direct  0    0         127.0.0.1       InLoopBack0
       10.1.1.2/32  Direct  0    0         10.1.1.2        Serial2
       10.2.1.0/30  Direct  0    0         10.2.1.1        Serial0
       10.2.1.1/32  Direct  0    0         127.0.0.1       InLoopBack0
       10.2.1.2/32  Direct  0    0         10.2.1.2        Serial0
       10.3.1.0/30  OSPF    10   3124      10.1.1.2        Serial2
       10.4.1.0/30  OSPF    10   3124      10.2.1.2        Serial0
       20.1.1.0/24  OSPF    10   3125      10.2.1.2        Serial0
       20.1.1.1/32  OSPF    10   3125      10.1.1.2        Serial2
       20.1.1.2/32  OSPF    10   3125      10.1.1.2        Serial2
      127.0.0.0/8   Direct  0    0         127.0.0.1       InLoopBack0
      127.0.0.1/32  Direct  0    0         127.0.0.1       InLoopBack0
```

HUAWEI TECHNOLOGIES CO., LTD. All rights reserved HUAWEI

Check RTA's routing table , routing information is incorrect , the detail routing information uses RTD as next-hop , the aggregated routing information uses RTD as next-hop , they are all existing .

Trouble analyse :

Because the detail routing information select RTD as next-hop , which means RTD only advertises detail routing information ,while not aggregated routing information ,maybe it doesn’t do routing aggregation . Aggregated route selects RTB as next-hop , which means RTB advertise routes correctly .

When configure route aggregation on many of ABRs , route aggregation should be configured on all the ABRs , not only on part of them .

Problem Description —Incorrect Route Aggregation

```
[RTD]display ip routing-table
Routing Tables: Public
        Destinations : 13       Routes : 13

Destination/Mask    Proto  Pre  Cost      NextHop         Interface
       10.1.1.0/30  Direct 0    0         10.1.1.2        Serial2
       10.1.1.1/32  Direct 0    0         10.1.1.1        Serial2
       10.1.1.2/32  Direct 0    0         127.0.0.1       InLoopBack0
       10.2.1.0/30  OSPF   10   3124      10.1.1.1        Serial2
       10.3.1.0/30  Direct 0    0         10.3.1.1        Serial0
       10.3.1.1/32  Direct 0    0         127.0.0.1       InLoopBack0
       10.3.1.2/32  Direct 0    0         10.3.1.2        Serial0
       10.4.1.0/30  OSPF   10   3124      10.3.1.2        Serial0
       20.1.1.0/24  OSPF   10   4687      10.1.1.1        Serial2
       20.1.1.1/32  OSPF   10   1563      10.3.1.2        Serial0
       20.1.1.2/32  OSPF   10   1563      10.3.1.2        Serial0
      127.0.0.0/8   Direct 0    0         127.0.0.1       InLoopBack0
      127.0.0.1/32  Direct 0    0         127.0.0.1       InLoopBack0
```

HUAWEI TECHNOLOGIES CO., LTD. All rights reserved HUAWEI

There is an aggregated route in RTD's routing table , this route item is advertised through backbone area , but it is meaningless to RTD .

Why there existing that invalid route , because route aggregation is just configured on RTB but not on RTD .

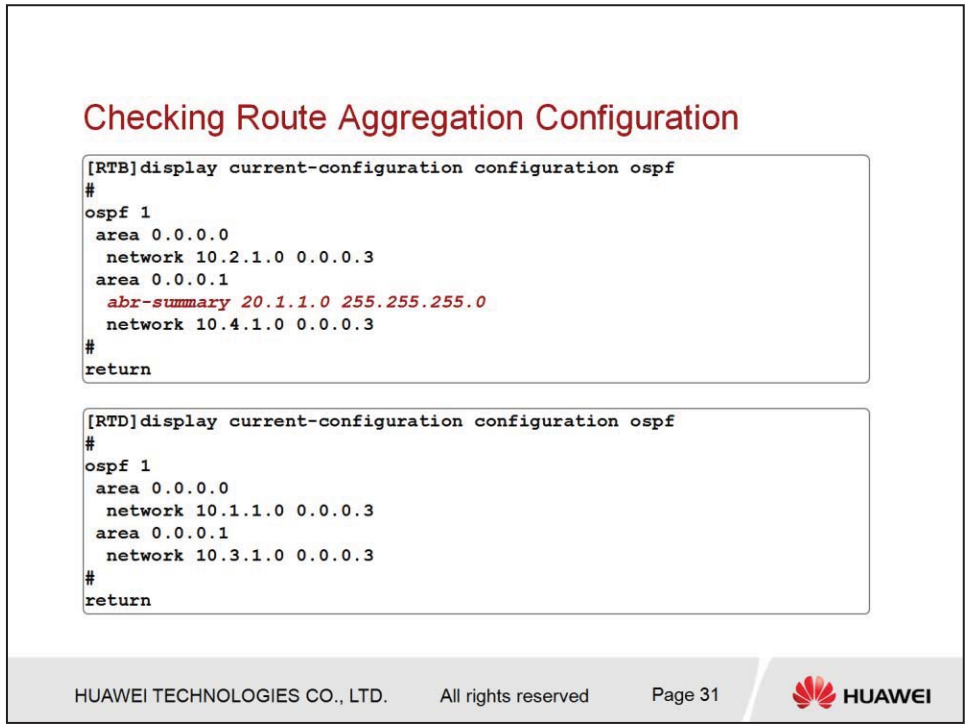
Problem Description —Incorrect Route Aggregation

```
[RTC]display ip routing-table
Routing Tables: Public
         Destinations : 13        Routes : 13

Destination/Mask    Proto   Pre  Cost      NextHop         Interface
       10.1.1.0/30  OSPF    10   3124      10.3.1.1        Serial1
       10.2.1.0/30  OSPF    10   3124      10.4.1.1        Serial0
       10.3.1.0/30  Direct  0    0         10.3.1.2        Serial1
       10.3.1.1/32  Direct  0    0         10.3.1.1        Serial1
       10.3.1.2/32  Direct  0    0         127.0.0.1       InLoopBack0
       10.4.1.0/30  Direct  0    0         10.4.1.2        Serial0
       10.4.1.1/32  Direct  0    0         10.4.1.1        Serial0
       10.4.1.2/32  Direct  0    0         127.0.0.1       InLoopBack0
       20.1.1.0/24  OSPF    10   6249      10.3.1.1        Serial1
       20.1.1.1/32  Direct  0    0         127.0.0.1       InLoopBack0
       20.1.1.2/32  Direct  0    0         127.0.0.1       InLoopBack0
      127.0.0.0/8   Direct  0    0         127.0.0.1       InLoopBack0
      127.0.0.1/32  Direct  0    0         127.0.0.1       InLoopBack0
```

HUAWEI TECHNOLOGIES CO., LTD. All rights reserved HUAWEI

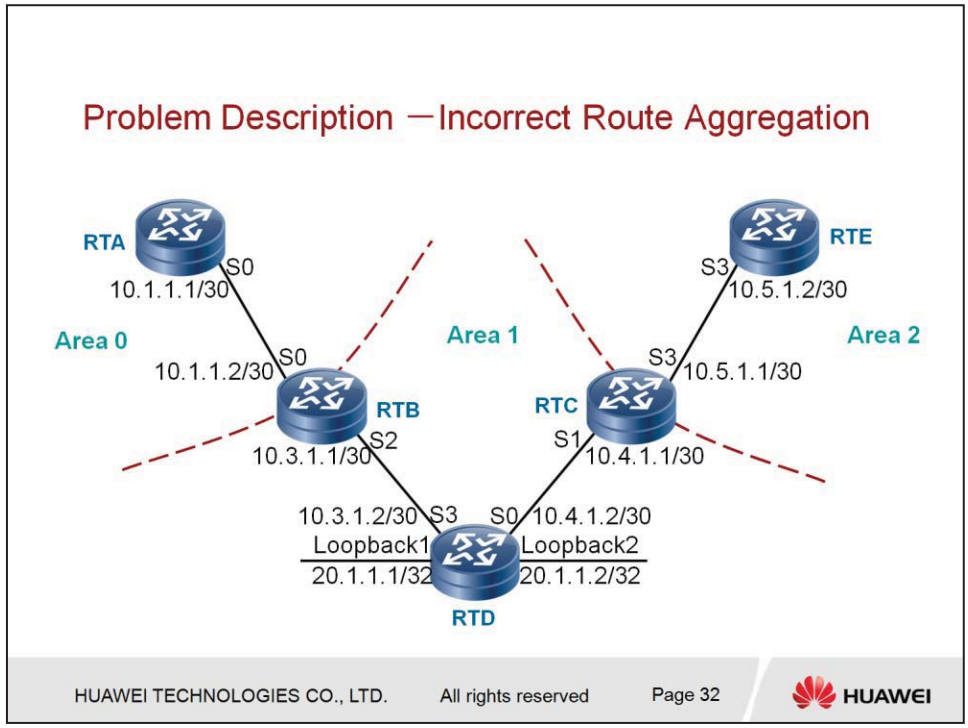
There is also an invalid aggregated route received through backbone on RTC‘s routing table. RTD advertised this aggregated route to RTC .



Checking the OSPF configuration on these two ABRS , route aggregation is correctly configured on RTB , but there is no route aggregation on RTD .

Correctly configure route aggregation on RTD can resolve that problem .

It shows , if the route aggregation is not properly configured ,it can cause route loop .



In this case :

Configuring virtual link on RTB and RTC , which makes area 2 connected to backbone area . Configuring route aggregation on area 1, and aggregate 20.1.1.1/32 and 20.1.1.2/32 into 20.1.1.0/24 on RTB .

RTA's router id is 1.1.1.1 ; RTB's router id is 2.2.2.2 ; RTC's router id is 3.3.3.3 ; RTD's router id is 4.4.4.4 ;RTE's router id is 5.5.5.5 .

Problem Description —Incorrect Route Aggregation

```
[RTA]display ip routing-table protocol ospf
Public Routing Table : OSPF
Destination/Mask    Proto   Pre  Cost      NextHop         Interface
       10.3.1.0/30  OSPF    10   3124      10.1.1.2        Serial0
       10.4.1.0/30  OSPF    10   4686      10.1.1.2        Serial0
       10.5.1.0/30  OSPF    10   6248      10.1.1.2        Serial0
       20.1.1.0/24  OSPF    10   3125      10.1.1.2        Serial0
       20.1.1.1/32  OSPF    10   6249      10.1.1.2        Serial0
       20.1.1.2/32  OSPF    10   6249      10.1.1.2        Serial0
```

```
[RTE]display ip routing-table protocol ospf
Public Routing Table : OSPF
Destination/Mask    Proto   Pre  Cost      NextHop         Interface
       10.1.1.0/30  OSPF    10   6248      10.5.1.1        Serial3
       10.3.1.0/30  OSPF    10   4686      10.5.1.1        Serial3
       10.4.1.0/30  OSPF    10   3124      10.5.1.1        Serial3
       20.1.1.1/32  OSPF    10   3125      10.5.1.1        Serial3
       20.1.1.2/32  OSPF    10   3125      10.5.1.1        Serial3
```

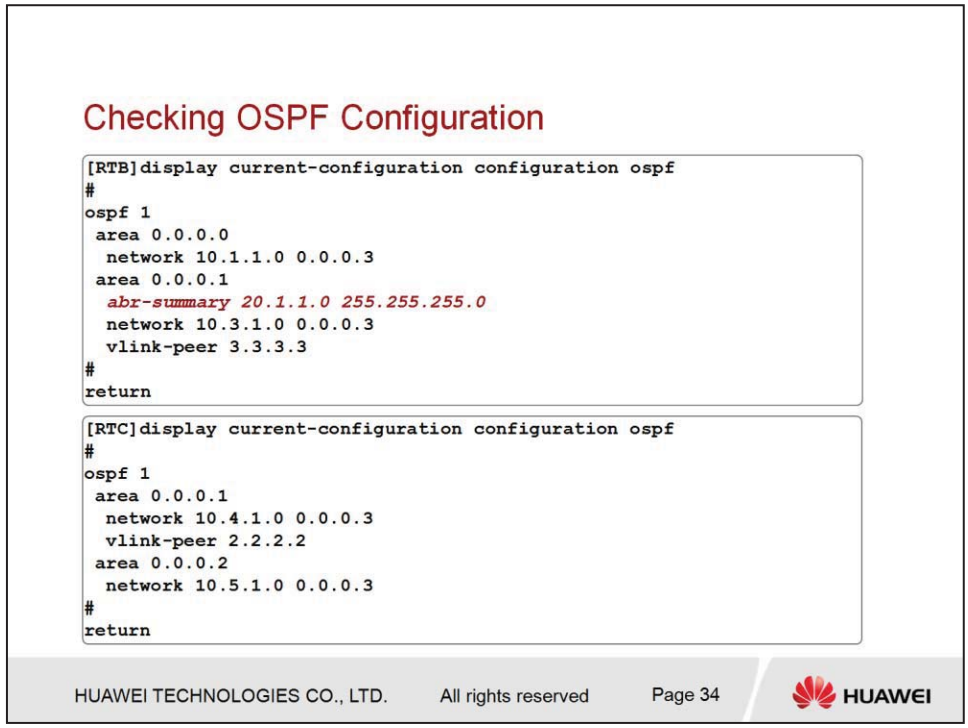
HUAWEI TECHNOLOGIES CO., LTD. All rights reserved HUAWEI

RTA's routing table contains both aggregated route and detail routes ; but there is only detail route in RTE's routing table .

Except that , other routes of OSPF are normal .

Trouble analyses :

RTA's routing table exists both aggregated route and detail routes , due to route aggregation isn’t configured on all area 1's ABR .



Checking the OSPF configuration on RTB and RTC , RTB has configured route aggregation , RTC didn’t configure route aggregation , that's why RTA's routing table contains both aggregated route and detail routes , because RTB advertises

aggregated route to RTA through backbone area , however , RTC advertises detail routes to backbone through virtual link .

But , why is there no aggregated route , only detail routing on RTE?

Because , when configure virtual link on a router , if the router learns a type 3 LSA through virtual link transit area , this LSA can be processed. In another word , the LSA must exist in both backbone area's and transit area's LSDB , this LSA can be processed.

Checking RTC's LSDB—Area 0

```
[RTC]display ospf lsdb

        OSPF Process 1 with Router ID 3.3.3.3
                Link State Database

                    Area: 0.0.0.0
Type      LinkState ID    AdvRouter       Age  Len   Sequence   Metric
Router    3.3.3.3         3.3.3.3         408  36    8000000A   3124
Router    2.2.2.2         2.2.2.2         413  60    8000001A   1562
Router    1.1.1.1         1.1.1.1         231  48    8000000C   1562
Sum-Net   10.3.1.0        2.2.2.2         409  28    8000000A   1562
Sum-Net   10.3.1.0        3.3.3.3         408  28    8000000B   3124
Sum-Net   10.4.1.0        2.2.2.2         409  28    8000000A   3124
Sum-Net   10.4.1.0        3.3.3.3         408  28    8000000B   1562
Sum-Net   20.1.1.1        3.3.3.3         408  28    8000000B   1563
Sum-Net   20.1.1.0        2.2.2.2         1658 28    80000001   1563
Sum-Net   20.1.1.2        3.3.3.3         408  28    8000000B   1563
Sum-Net   10.5.1.0        3.3.3.3         393  28    80000006   1562
```

HUAWEI TECHNOLOGIES CO., LTD. All rights reserved

RTC learns aggregated type 3 LSA through backbone area (virtual link) , this LSA is advertised by RTB .

Checking RTC's LSDB—Area 1 and Area 2

```
[RTC]display ospf lsdb
        OSPF Process 1 with Router ID 3.3.3.3
                Link State Database

                        Area: 0.0.0.1
Type      LinkState ID    AdvRouter       Age  Len   Sequence   Metric
Router    4.4.4.4         4.4.4.4         387  96    8000001C   1562
Router    3.3.3.3         3.3.3.3         413  48    80000018   1562
Router    2.2.2.2         2.2.2.2         414  48    80000014   1562
Sum-Net   10.5.1.0        3.3.3.3         395  28    80000006   1562
Sum-Net   10.1.1.0        2.2.2.2         425  28    8000000A   1562

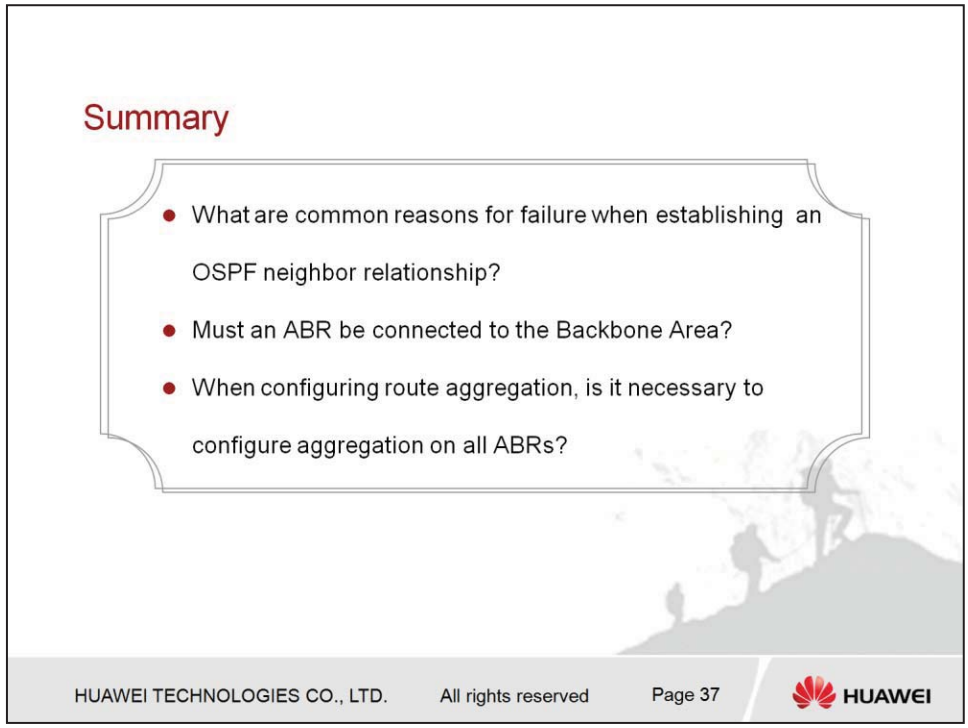
                        Area: 0.0.0.2
Type      LinkState ID    AdvRouter       Age  Len   Sequence   Metric
Router    5.5.5.5         5.5.5.5         367  48    8000000B   1562
Router    3.3.3.3         3.3.3.3         371  48    80000007   1562
Sum-Net   10.3.1.0        3.3.3.3         395  28    80000006   3124
Sum-Net   10.4.1.0        3.3.3.3         395  28    80000006   1562
Sum-Net   20.1.1.1        3.3.3.3         395  28    80000006   1563
Sum-Net   20.1.1.2        3.3.3.3         395  28    80000006   1563
Sum-Net   10.1.1.0        3.3.3.3         395  28    80000006   4686
```

HUAWEI TECHNOLOGIES CO., LTD. All rights reserved HUAWEI

This virtual link's transit area is area 1, but RTB aggregates area 1's routes , so RTB will not advertise aggregated route to area 1 , that ‘s why there's no

aggregated type 3 LSA in area 1's LSDB , and also , RTC will not process the aggregated route learnt from virtual link .

RTC can’t process the aggregate routes learnt from backbone area , so RTC can’t advertise aggregated routes into area 2, there is only detail routes on RTE , but no aggregated route .



What are common reasons for failure of establishing OSPF neighbor relationship?

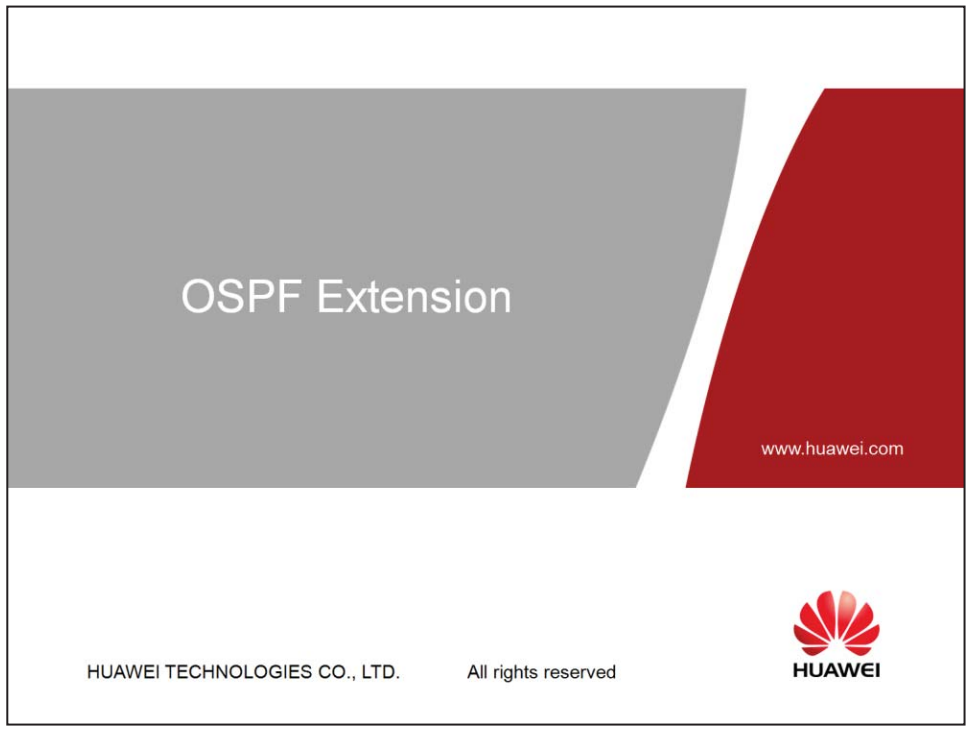
Router id confusion , area id mismatch , network mask mismatch, authentication type and authentication password mismatch, external routing capability mismatch , wrong static neighbor configuration on NBMA network .

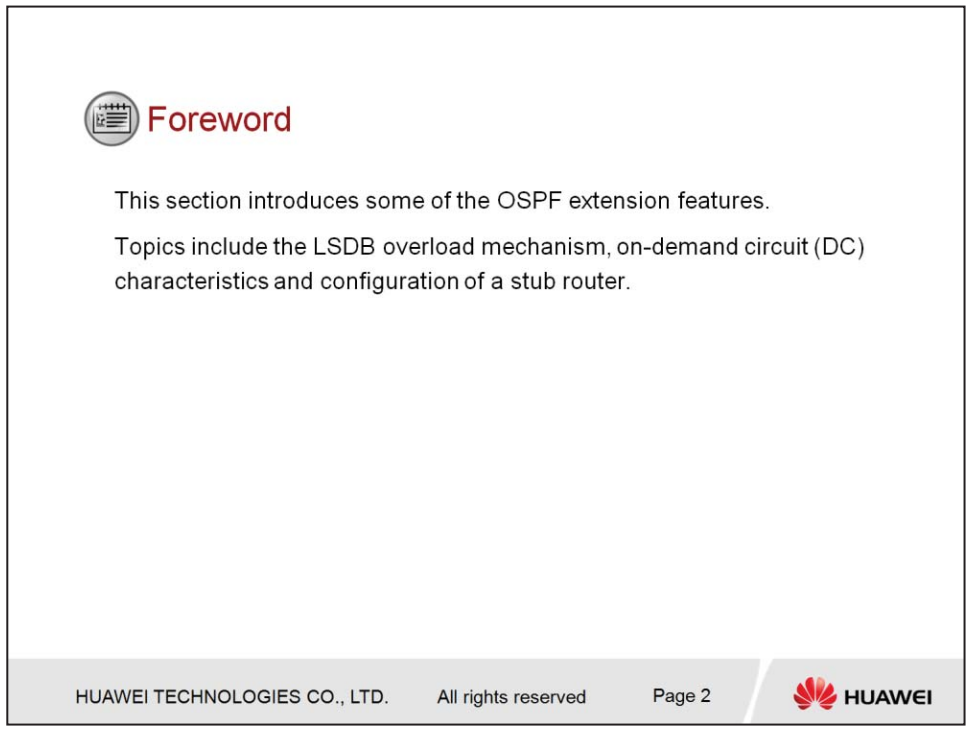
Must ABR be connected to the Backbone Area?

ABR must be connected to the Backbone Area (area 0) , both physical link and virtual link can be used .

When configuring route aggregation, is it necessary to configure aggregation on all ABRs?

When configuring route aggregation , you must configure aggregation on all ABRs.


OSPF Extension
www.huawei.com
HUAWEI TECHNOLOGIES CO., LTD.
All rights reserved
HUAWEI

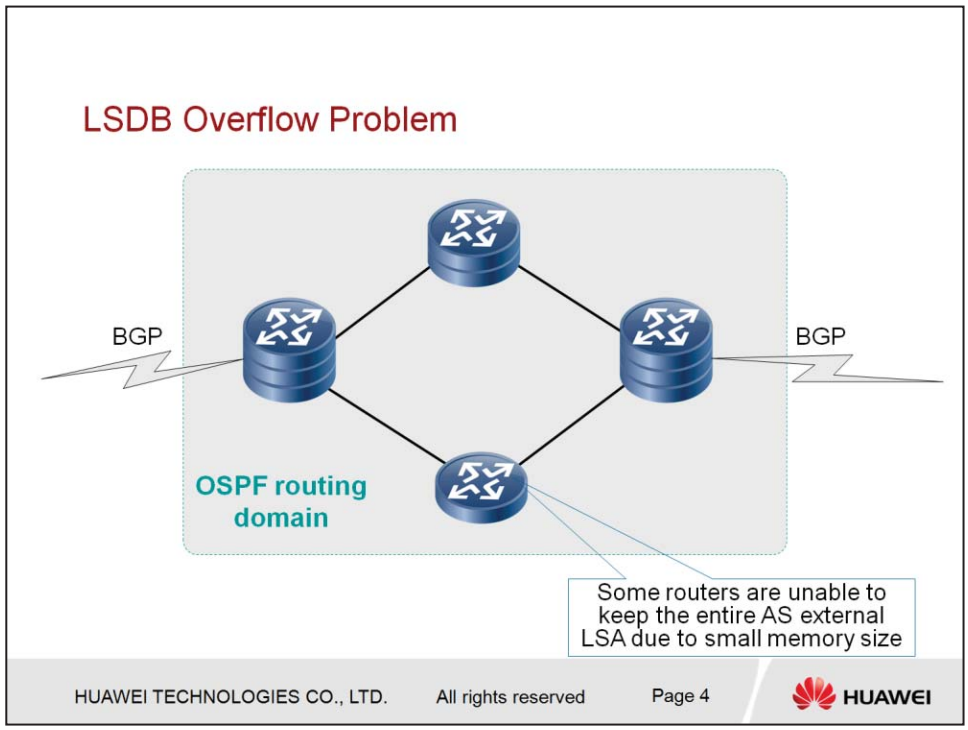

Foreword
This section introduces some of the OSPF extension features.
Topics include the LSDB overload mechanism, on-demand circuit (DC) characteristics and configuration of a stub router.
HUAWEI TECHNOLOGIES CO., LTD. All rights reserved Page 2
HUAWEI

Objectives

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

- Understand the LSDB overload mechanism
- Understand OSPF features for on-demand circuits
- Understand stub router functionality

HUAWEI TECHNOLOGIES CO., LTD. All rights reserved

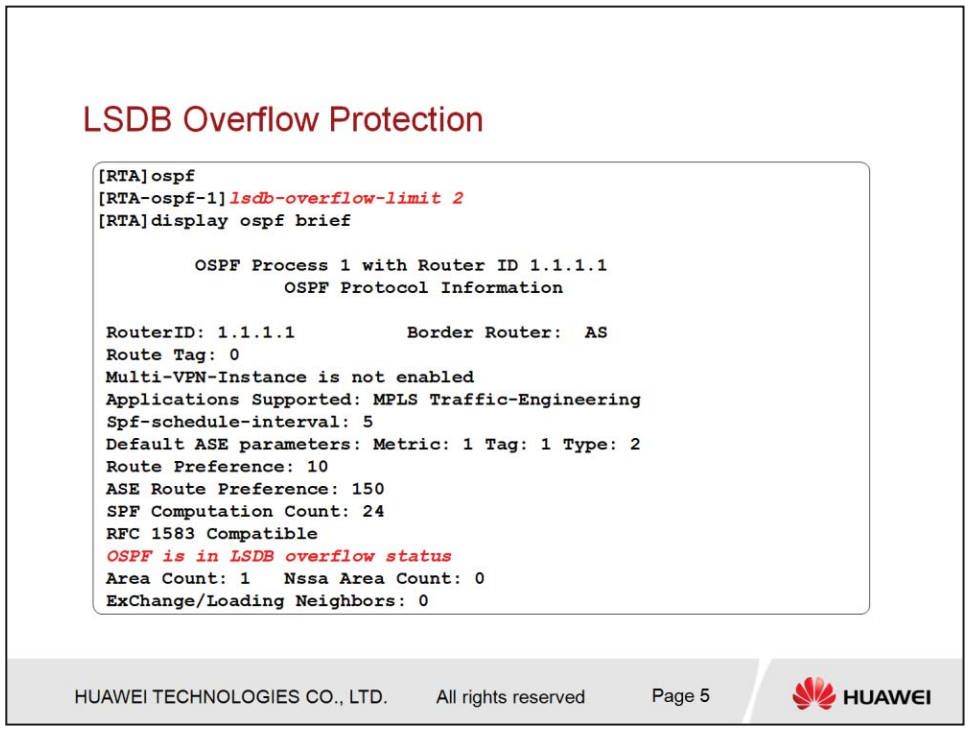


Due to the storage memory limitation , if LSDB is very large , some router can't store the entire LSDB , this is called LSDB overflow .

LSDB overflow always cause by AS external LSA (type 5 LSA) which has a very large number .

The size of LSDB can be reduced through configuring area into stub or totally stub area .

This chapter introduces another method to deal with overload of type 5 LSA , without configuring stub area or NSSA .



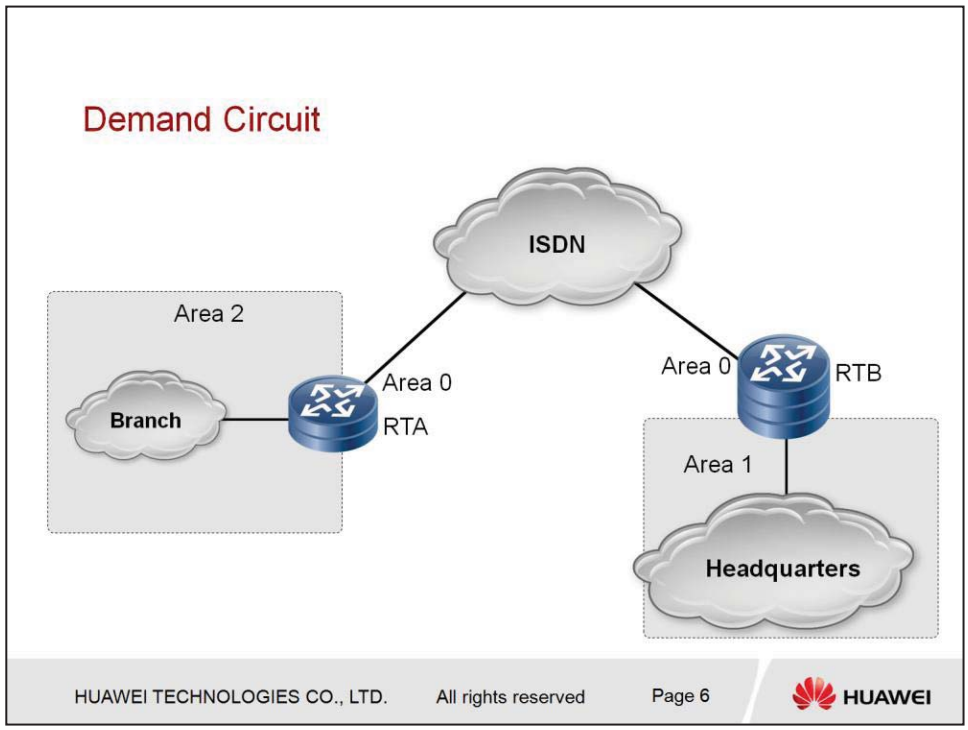
RFC1765 defines a new OSPF parameter : ospfExtLsdbLimit , which means the maximum quantity of type 5 LSA, the configuration of VRP platform .

lsdb-overflow-limit number

number： The number of type 5 LSA in LSDB , range is 1~1000000 , by default this function is disabled .

When type 5 LSA in LSDB is overloaded , router gets into overflow state , the type 5 LSA generate by itself is aging . and no more type 5 LSA can be generated .

The overloaded OSPF router can receive type 5 LSA generated by rother outer , but the quantity can’t be larger than the limitation , if receive a new type 5 LSA when overload , the LSA will be discarded .



Demand circuit can be built up when there has traffic , if there is no traffic , the link is down , in order to reduce the cost.

In this case , branch uses an ISDN line connect to headquarter , using OSPF routing protocol , we hope if there is traffic , the branch can dial up to headquarter , if no traffic , the link disconnects automatically ,so the cost is reduced .

OSPF can't sent hello packet periodically , also can't flood LSA periodically , it need to extend the OSPF working mechanism .

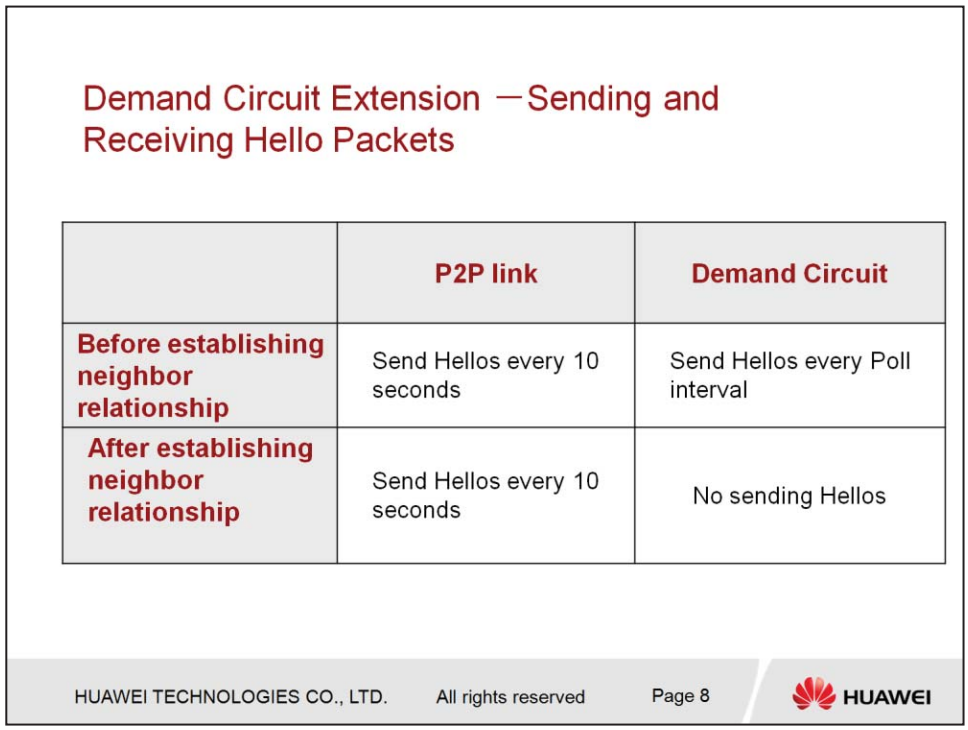
Demand Circuit Extension—Modifying Options Field

1st bit	2nd bit	DC bit	4th bit	5th bit	6th bit	7th bit	8th bit

Packet type	DC bit function	Where to set DC bit
Hello Packet	Negotiate with the other endpoint whether to enable extension features over DC	Hello Packets sent out an interface on Demand Circuit
DD Packet	Negotiate with the other endpoint whether to enable extension features over DC	DD Packets sent out an interface on Demand Circuit
LSA	Notify other routers of extension features supported by originator	All router's self-originated LSAs

HUAWEI TECHNOLOGIES CO., LTD. All rights reserved HUAWEI

The first extension is that OSPF's hello packet , DD packet and LSA's Options field add a DC bit, which means this router supports Demand Circuit or not .



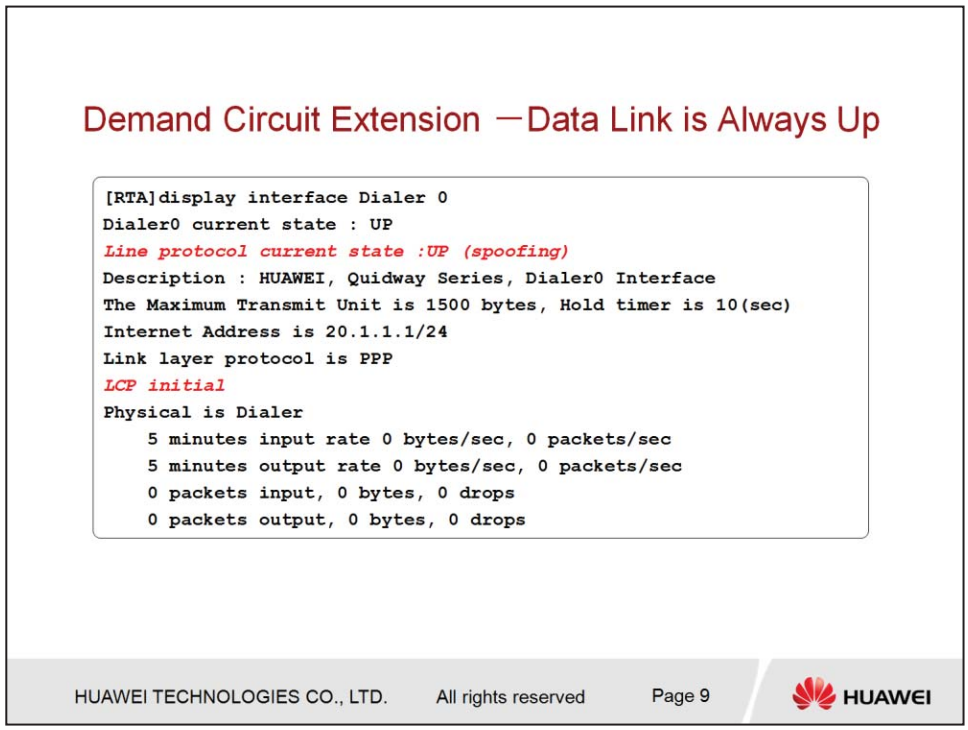
When the router which supports demand circuit to send a hello packet , DC bit must be set to 1 .

If a router receives a hello packet with DC bit set to 1 , it will consider this link as a demand circuit link , even though this side is not configured into a demand circuit link .

On a normal P-to-P link , no matter neighbor is built up or not , OSPF router will send a hello packet every 10 seconds , but on demand circuit link , OSPF changes this mechanism :

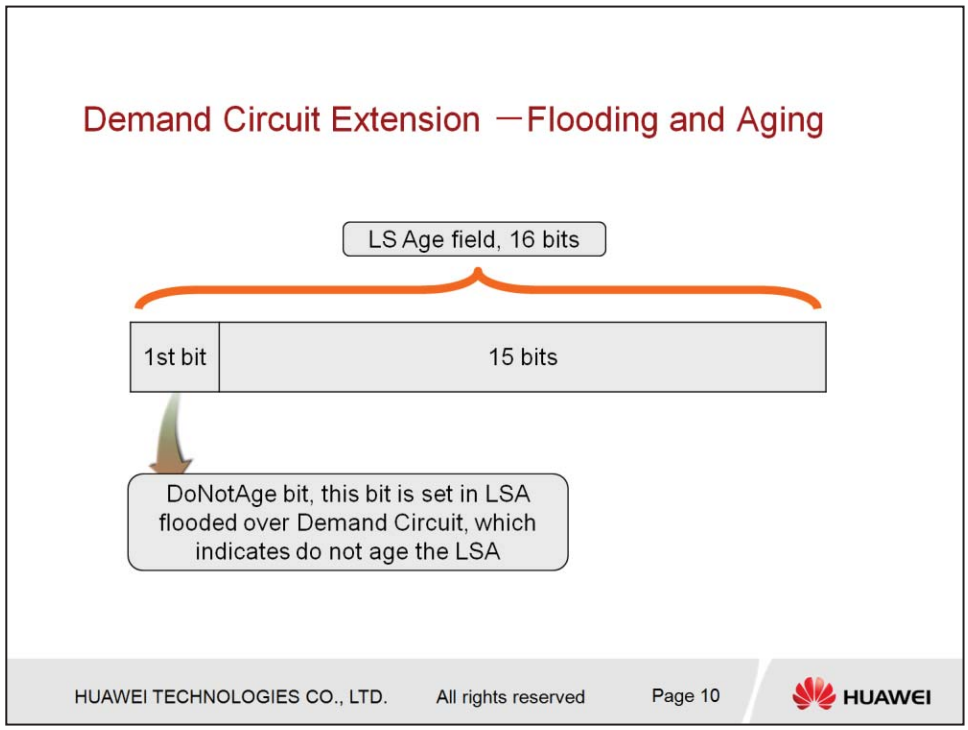
Before the neighbourship built up , a hello packet is sent every poll interval (120 seconds by default) , which is used to discover neighbors .

After building up a neighbourship , no hello is sent. The router considers its peer is active all the time .



Line protocol current state is always up , means the data link layer is up .So the routing protocol believes this link is always up , advertise this link's network .

If this physical link is down due to no data transfer , but the data link layer is still up , all the networks can be advertised outside through routing protocol .



Flooding LSA on demand circuit :

1. Content in LSA is changed , not only the sequence number or LS Age .

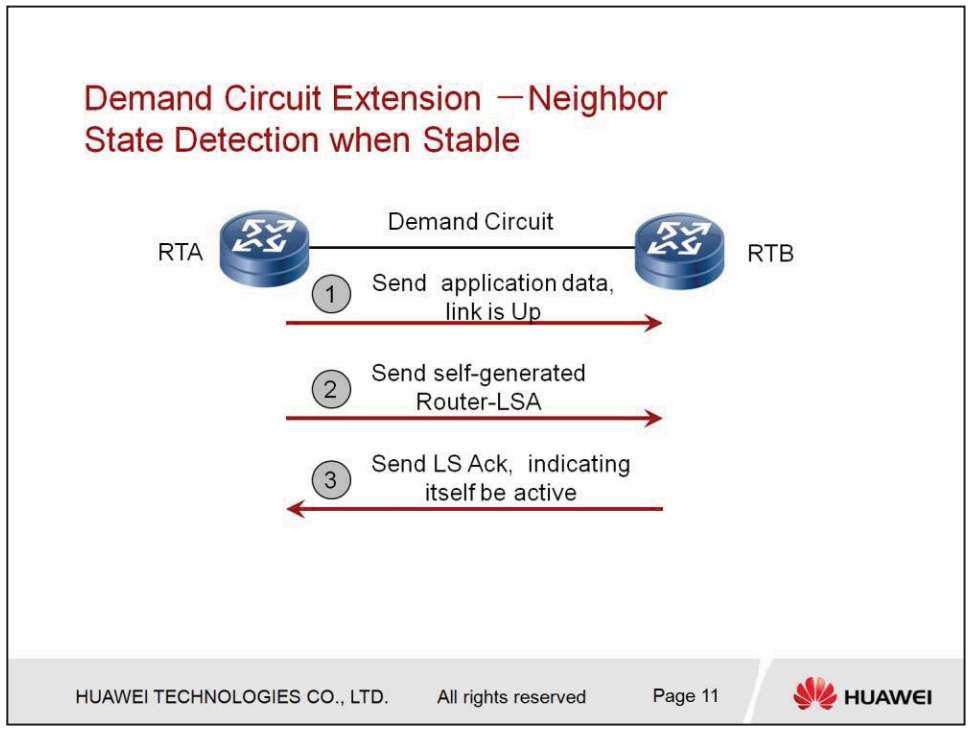
2. When LS Age reaches to MaxAge (the LSA needs to be deleted .)

The LSA on demand circuit can't be renewed periodically , it defines on demand circuit extension , the first bit of LS Age (DoNotAge bit)is set to 1 , which means the LSA can't age , when a router receives a LSA with DoNotAge set to 1 , it can't age this LSA .

A DoNotAge LSA will be removed if BOTH of the following conditions are met:

1) The LSA has been in the router's database for at least MaxAge seconds.

2) The originator of the LSA has been unreachable (means that the demand circuit remains in operative for longer than the maxage.) for at least MaxAge seconds.

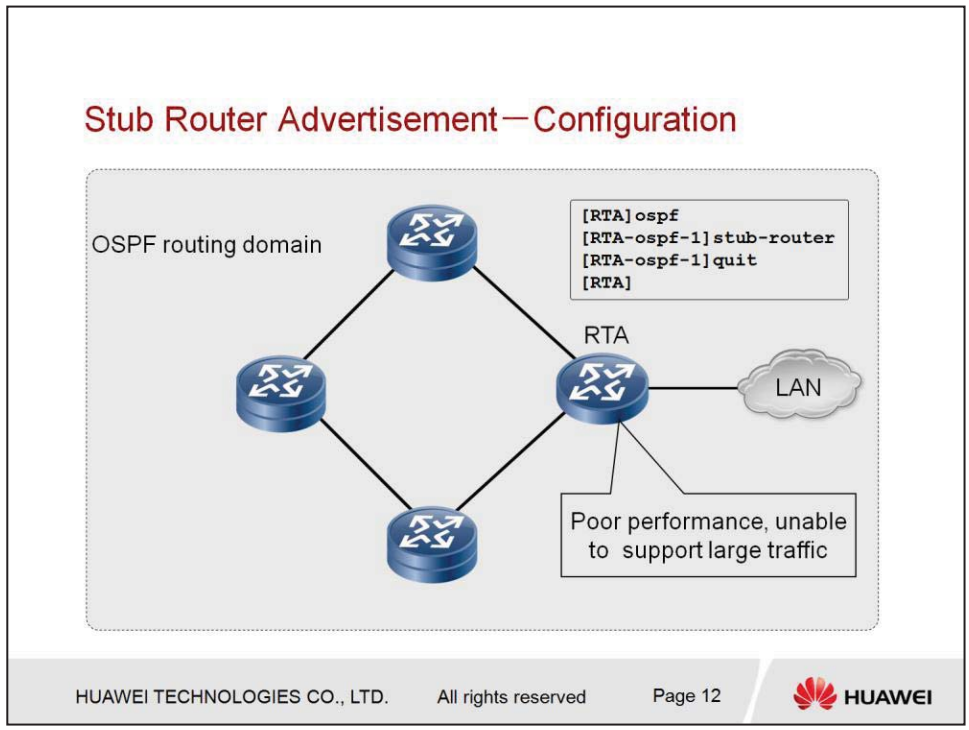


The mechanism of checking the opposite side on demand circuit is active or not in steady state (the neighbor is built up) .

1. If there is data , the demand circuit is active .

2. After the demand circuit is active , generate Router-LSA to the other side .

3. If the other side response LS Ack , means the neighbor is active , if no LS Ack is received in transit interval (5 seconds by default) , means the neighbor is not active .

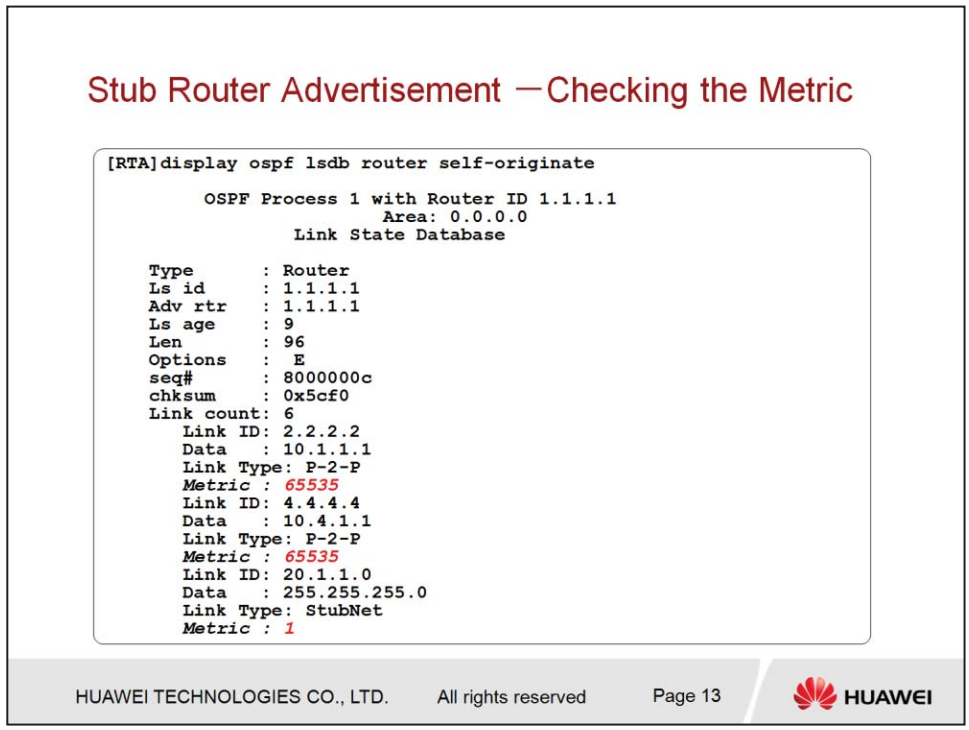


In this case , RTA is a low-end router , in some situation , the CPU is overloading , or memory utilization is very high ,the administrator hopes the great traffic don‘t pass through RTA also hopes the networks¡¯traffic directly connected to RTA can’t be interrupted .

VRP platform support a stub router , when a router is configured as stub router, the Router-LSA of this router , the metric of non-stub connection network is set to a very large value (65535) , so this link can’t be chosen , the cost of a stub connection network is not change , so the data sending to a stub network can’t be interrupted .

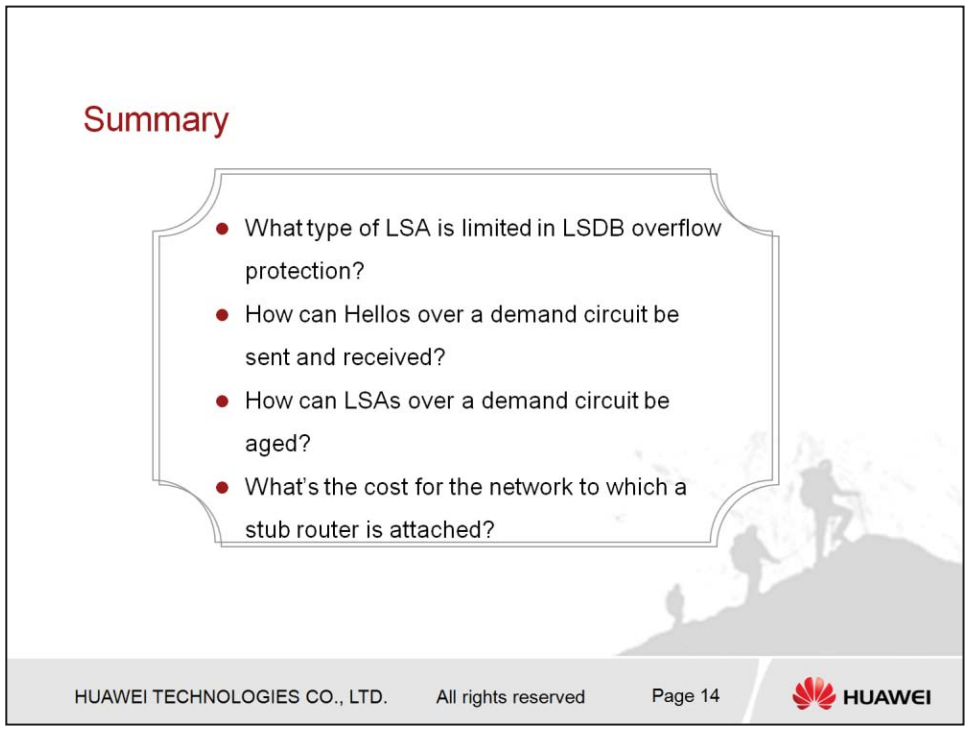
stub-router

stub-router command is used to configure this router into a stub router .



When RTA is configured as a stub router , in the Router-LSA generated by itself , cost of non-stub network is set to 65535 , stub network's cost is not changed .

Due to non-stub network's cost is very large , the other router will keep away from this router , so the traffic on this router is reduced .



What type of LSA is limited in LSDB overflow protection ?

Type 5 LSA .

How to send and receive Hellos over Demand Circuit ?

Before establishing neighbor relationship , Sending Hellos every Poll interval.

After establishing neighbor relationship , No sending Hellos .

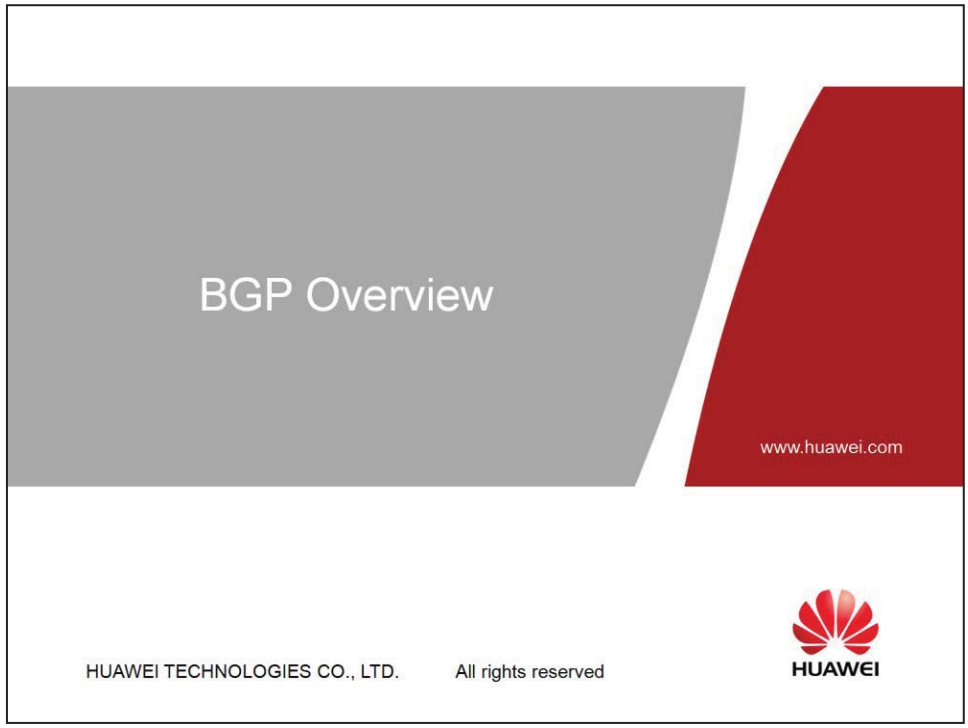
How to age LSA over Demand Circuit ?

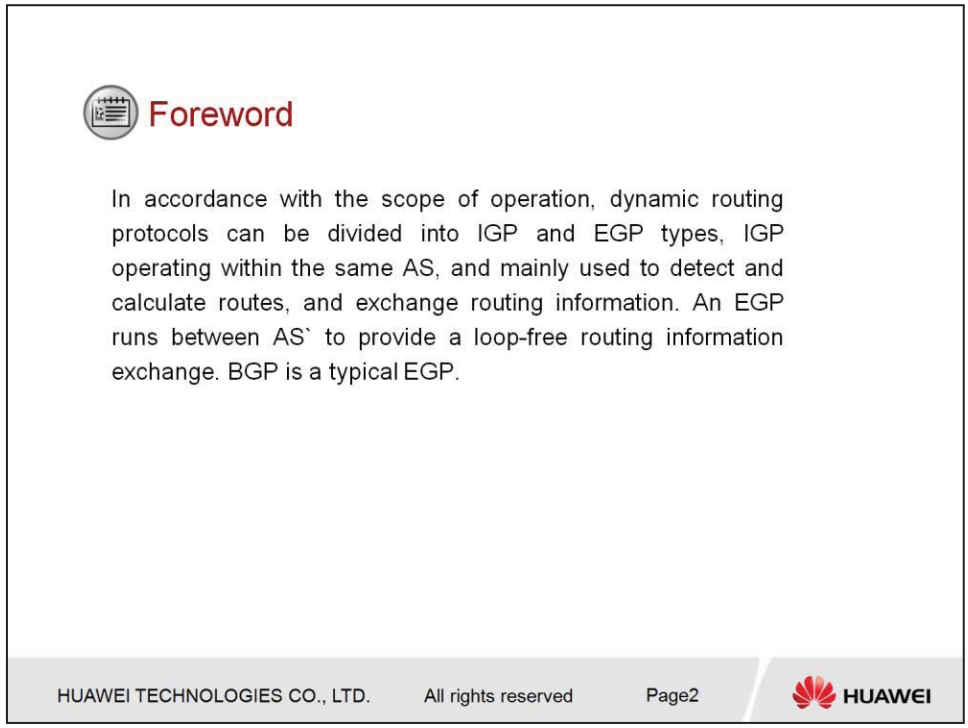
Set the first bit of LS Age（DoNotAge bit）to one .

What's the cost for the network to which Stub Router is attached ?

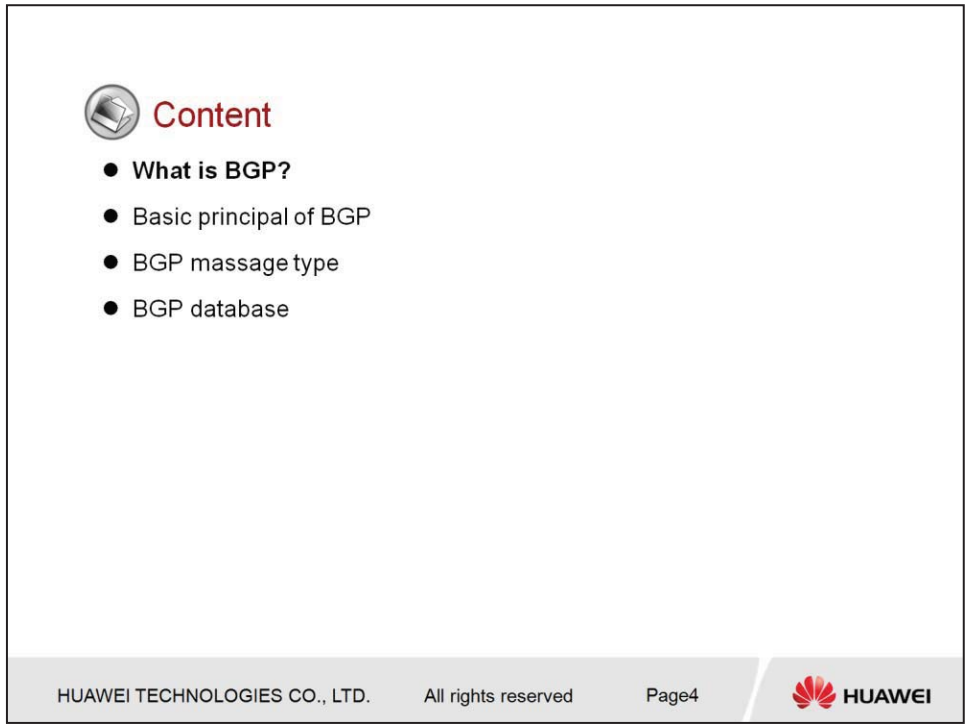
Non-stub network's cost is 65535 , stub network's cost is not changed .

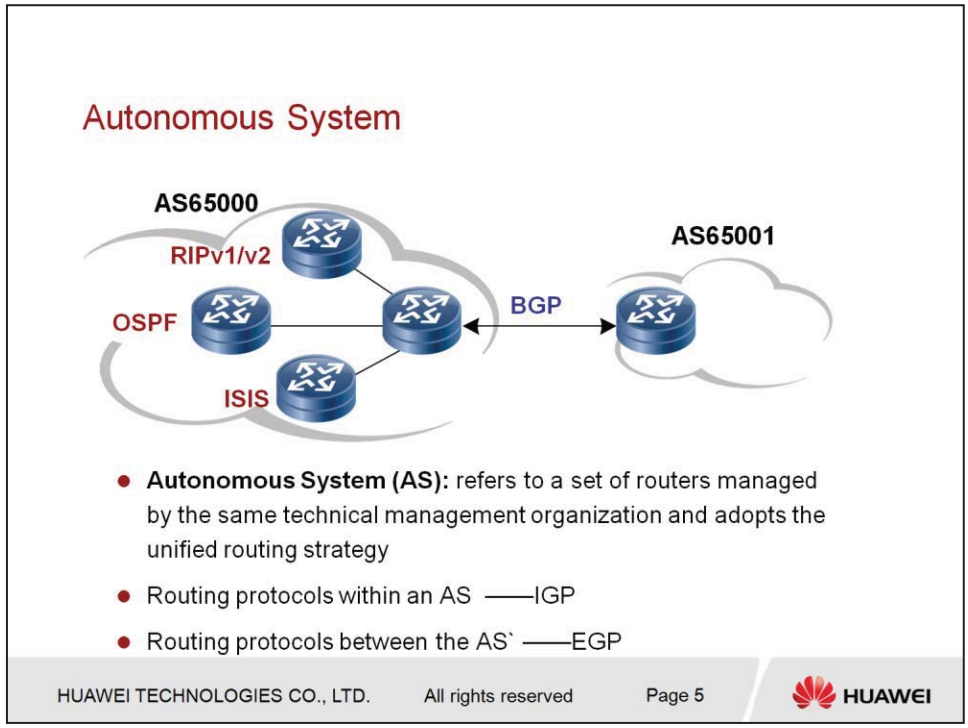
Module 3
BGP


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


Foreword
In accordance with the scope of operation, dynamic routing protocols can be divided into IGP and EGP types, IGP operating within the same AS, and mainly used to detect and calculate routes, and exchange routing information. An EGP runs between AS` to provide a loop-free routing information exchange. BGP is a typical EGP.
HUAWEI TECHNOLOGIES CO., LTD. All rights reserved Page2
HUAWEI


Objectives
Upon completion of this section, you will be able to :
Understand BGP`s operational scope
Understand BGP functions
Understand BGP features
HUAWEI TECHNOLOGIES CO., LTD. All rights reserved Page3
HUAWEI


Content
What is BGP?
Basic principal of BGP
BGP massage type
BGP database
HUAWEI TECHNOLOGIES CO., LTD. All rights reserved Page4
HUAWEI

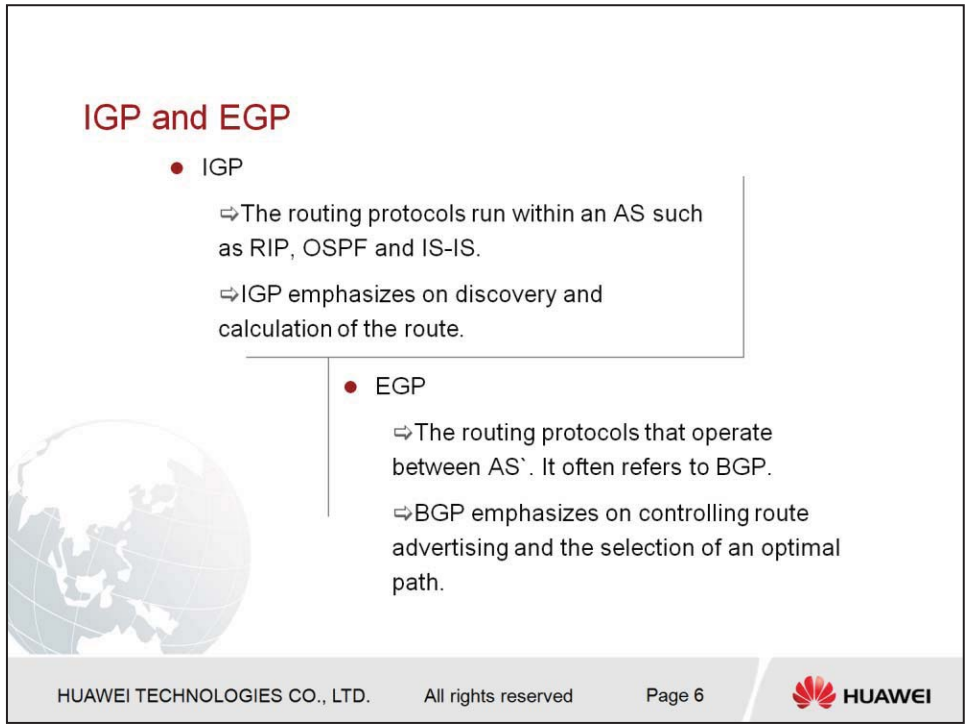


Autonomous system refers to a set of routers managed by the same technical organization and adopts the unified routing strategy.

Each autonomous system has a unique AS number which is allocated by IANA.

We distinguish different autonomous systems by using different numbers. When the network administrator does not want his/her communication data to pass some autonomous systems, these AS numbers become very useful. For example, the network administrator wants to avoid some autonomous systems managed by his competitor or avoid some autonomous systems which are lack of security mechanism. In this case, network administrator can specify the path for data transmission by using routing protocol, routing policy and AS number.

The AS numbers range from 1 to 65535. Among them, AS numbers 1 to 64511 are the registered Internet numbers while those from 64512 to 65535 are reserved for private use.

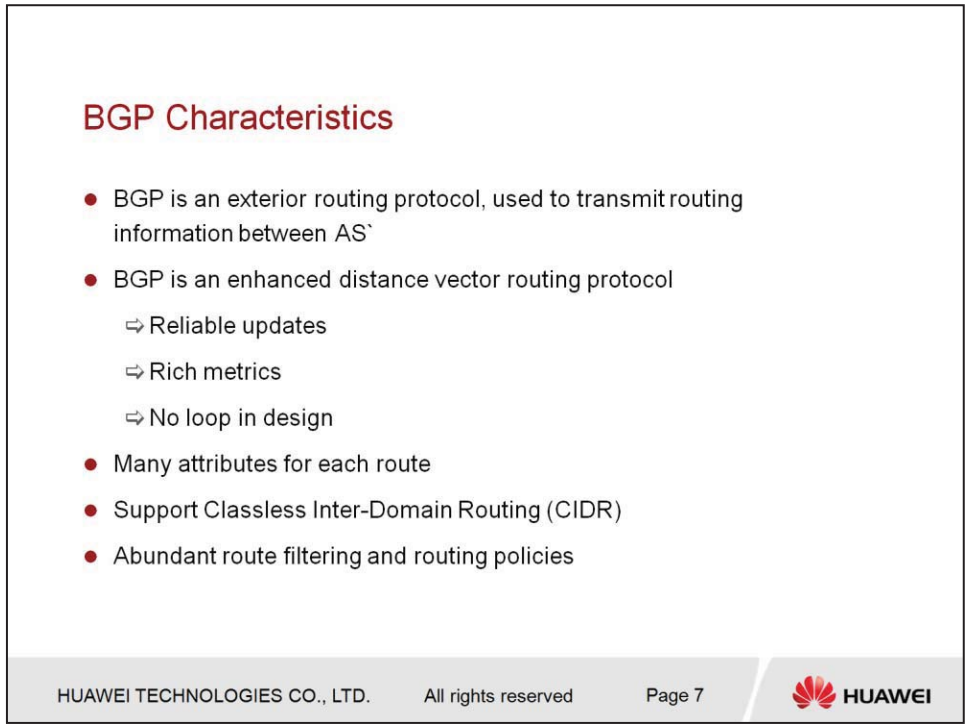


The differences between the IGP and EGP are:

1. IGP is the routing protocols that run within an autonomous system such as RIP, OSPF, and IS-IS. It emphasizes on discovery and calculation of the route.

2. EGP is the routing protocols that run between the autonomous systems.

Nowadays, it is often referred to BGP. BGP emphasizes on control of route advertising and selection of optimal path.



BGP (Border Gateway Protocol) is a dynamic routing protocol that runs between the autonomous systems. Its basic function is to automatically exchange the loop-free network reachability information between the ASs. This network reachability information contains the list of autonomous system that the reachability information traverses. This reachability information is sufficient to construct the topology map of the AS from which the routing loops may be avoided and some routing policies at the AS level can be enforced. Protocols like OSPF and RIP are interior gateway protocols (IGPs) that run inside an autonomous system while BGP is exterior gateway protocol (EGP) that run between ISPs.

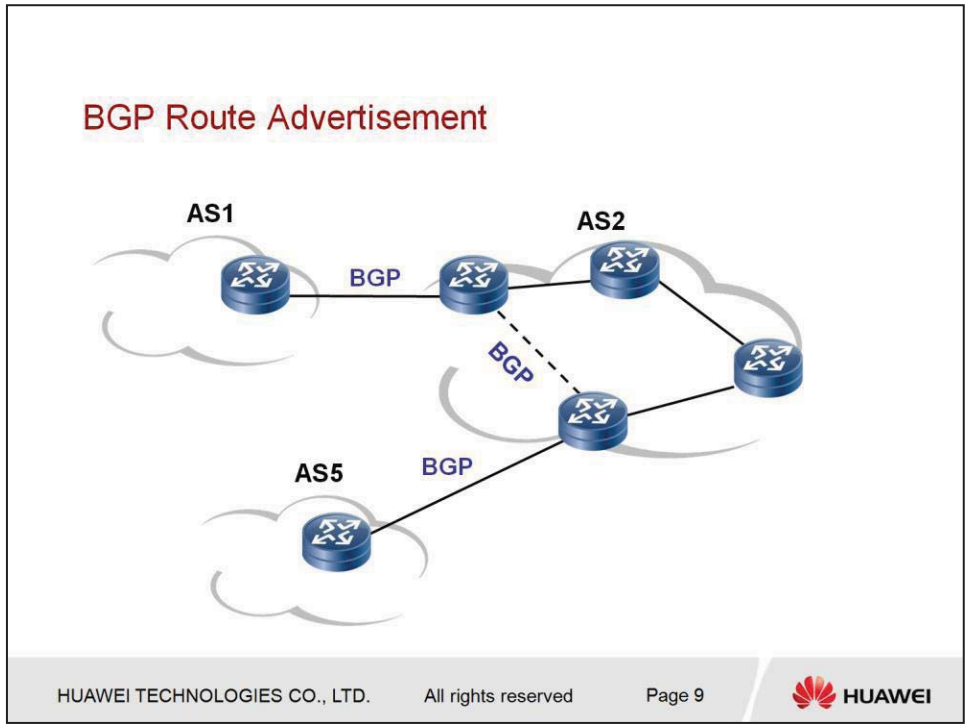
BGP was first introduced in 1988. The earliest versions of BGP are RFC1105 (BGP-1), RFC1163 (BGP-2) and RFC1267 (BGP-3). The current version of BGP is RFC4271/RFC1771 (BGP-4). BGP-4 has become the standard routing protocol for Internet.

The features of the BGP are as follow:

BGP provides the exchange of loop-free routing information between the autonomous systems (loop-free routing is guaranteed by using the AS-PATH attribute). BGP is policybased routing protocol. It enforces the policy through abundant BGP

route attributes. It works on application layer and uses TCP as the transport layer protocol (BGP exchanges the route between the neighbors on top of the reliable TCP connection). BGP is a distance vector routing protocol. This means that it will announce to its neighbors those networks that it can reach by itself. However, the BGP route selection is not solely based on the distance (route selection is based on the bandwidth for most of the routing protocols). The BGP route selection is based on the abundant route attributes. These attributes attached to the reachable IP subnets.

Therefore, we called BGP a distance vector routing protocol. It is easier to understand BGP as a distance vector routing protocol when we treat the whole AS as a single router. Apart from that, BGP has some features of the link state protocol. For example, incremental updated, advertising route with IP subnet mask a, etc..



To establish the TCP connection, two routers of each connection must know the IP address of each other. A router can learn the IP address of another router via direct connection, static route, or IGP.

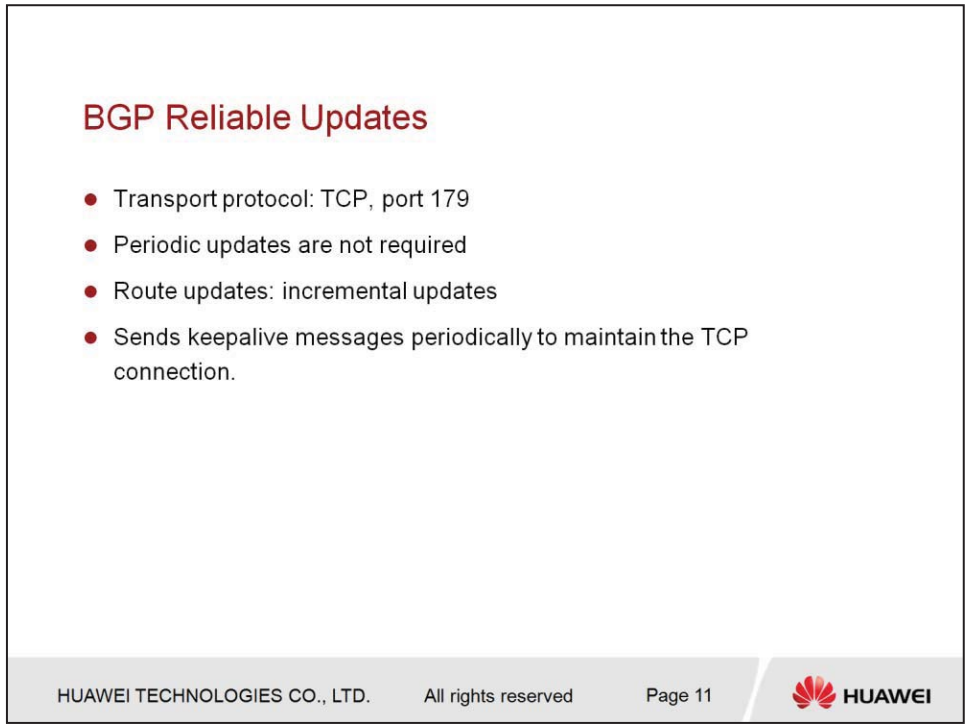
The border router of the ISP will try to establish the TCP connection after successfully learning the IP address of the other end. If the connection is not established successfully, the routers will try to re-establish the connection. This process will be repeated until the connection is established successfully.

When the TCP connection is successful, two routers will exchange some information to verify the capability of the other end or determine the next action to be performed. This is necessary because any equipment that support IP protocol stack can support the establishment of TCP connection. However, not all the equipments that support IP protocol stack can support BGP. Therefore, the exchange of information is to guarantee the capability of the router in supporting BGP. After confirming the capability of the routers, information from the BGP tables is exchanged.

The two routers that establish the BGP connection form the peer

relationship with each other. To guarantee the normal operation of the BGP process, the two ends of the peers will send the keepalive message periodically to ensure the validity of the connection.

If one end of the peer can't receive any keepalive message from its peer within the hold time interval, the BGP process can be considered has been stopped in the neighbor. Therefore, the TCP connection will be closed and all the routes learnt from the neighbor will be removed.



BGP uses TCP (port 179) as its transport protocol. TCP can ensure the reliable transmission of BGP.

Periodic updates is not required.

For route updating, BGP sends incremental routes only (for example new reachable route, changed route or withdrawn route). This greatly reduces the bandwidth occupied by BGP

route advertising. Therefore, BGP is suitable for advertising a large volume of routing information over the Internet.

During the initialization, BGP router sends all the routes to its peer and at the same time it saves the routes which have been sent in its local database. When the local BGP receives a

new route, it will compare this new route with the saved information which has been sent out.

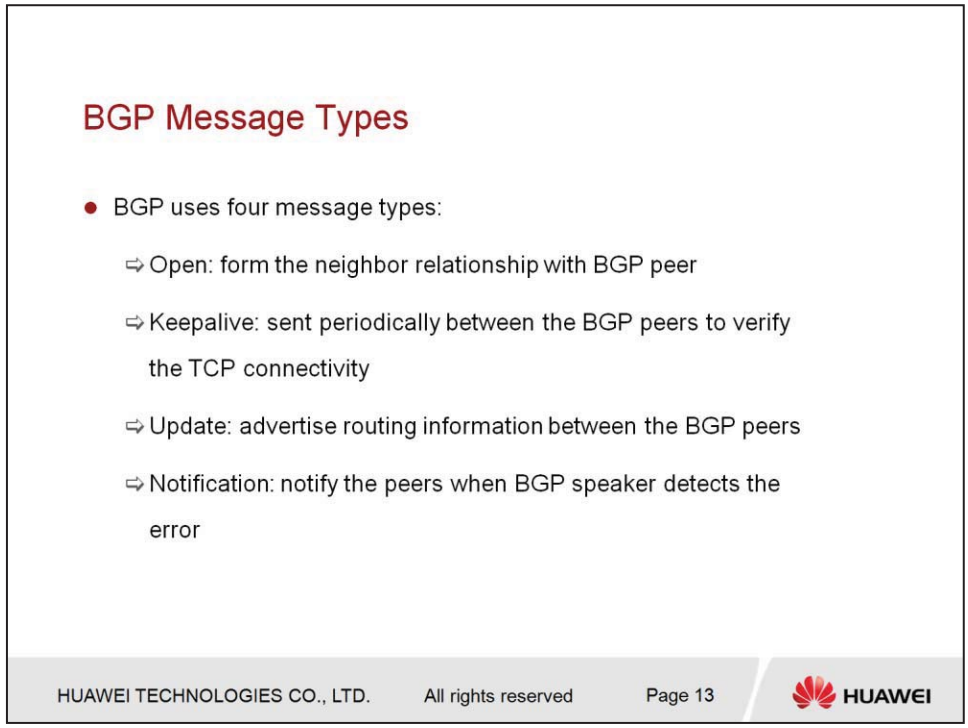
If the local BGP has not sent this new route before, it will send it out. Otherwise, the local BGP will compare this new route with the route already sent. If the new route is better, then local BGP will send out this new route and at the same time update the route already sent.

Else, if the new route is worse, it will not send the new route. What will the local BGP do if it finds out that one of the local

roules fails (for example the corresponding port fails)

and this route was once sent? In this case, the local BGP will send a route withdrawal message to the BGP peer. In conclusion, BGP does not necessarily broadcast all the routing information every time. It only sends the incremental routes after the initialization, which ensures the minimum communication between the BGP and its peer.

In addition, BGP sends and receives the keepalive message to verify the TCP connectivity.



Routers that run a BGP routing process are often referred to as BGP speakers. Four types of messages that are exchanged between the BGP speakers are Open, Keepalive, Update, and Notification. Among them, Open, Keepalive and Notification messages are used to establish and maintain the neighbor relationship.

Open Message: includes the BGP version, AS number of the sender and etc. After the establishment of TCP connection, two routers that try to establish the neighbor relationship will exchange the Open message and check whether the neighbor relationship can be established.

Keepalive Message: Keepalive message is exchanged periodically to maintain the neighbor relationship. It is used to verify the connectivity of the peer.

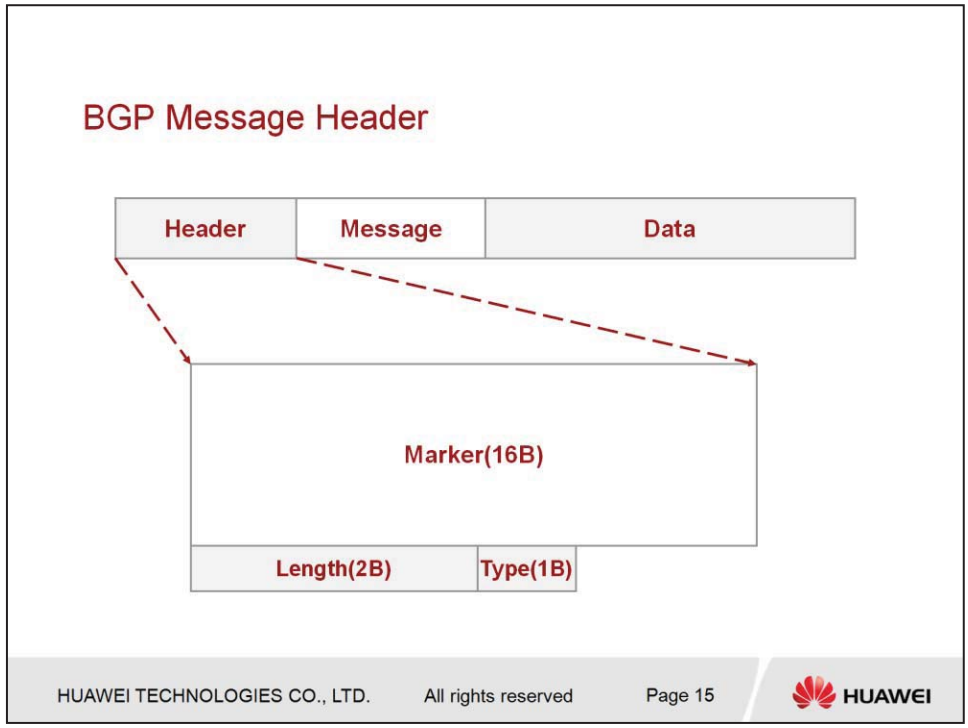
Update Message: Update message is used to exchange the routing information between peers.

It consists of withdrawn route information, network layer reachability information and various paths attribute information. Among the four message types, update message is the most important message for BGP.

Notification Message: Notification message is the error checking

mechanism used in BGP.

BGP speaker will send the notification message when an error occurs. This will always cause the BGP connection to be closed.



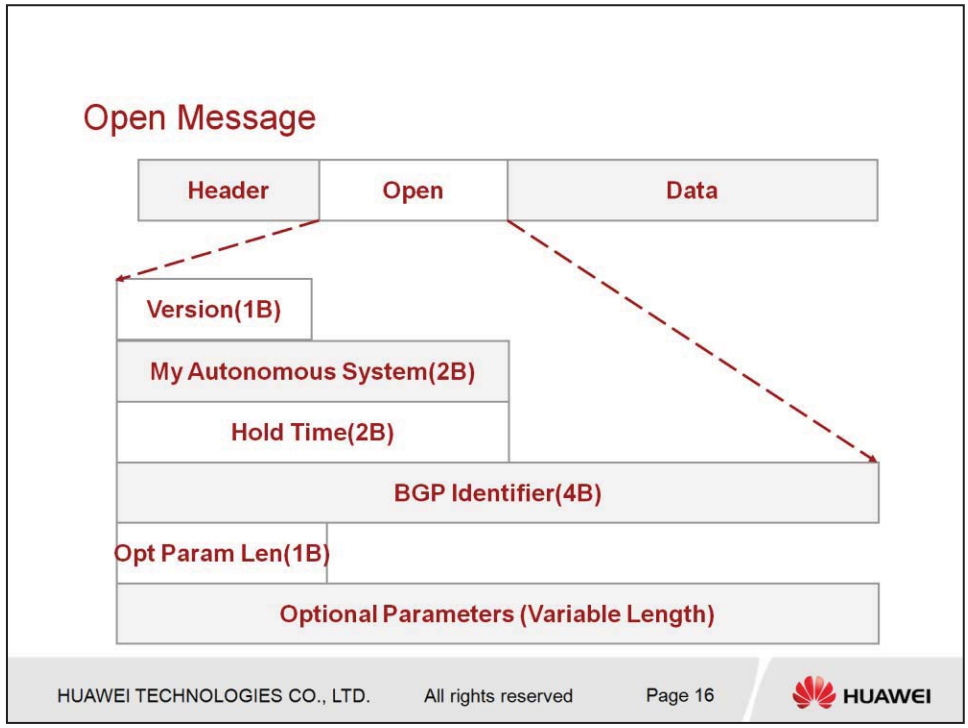
Marker：This 16-octet field contains a value that the receiver of the message can predict. In open message without authentication,then the Marker must be all ones. Otherwise, the value of the marker can be predicted by some a computation specified as part of the authentication mechanism used. The Marker can be used to detect loss of synchronization between a pair of BGP peers, and to authenticate incoming BGP messages. Length: This 2-octet unsigned integer indicates the total length of the message, including the header, in octets. Type： This 1-octet unsigned integer indicates the type code of the message. The following type codes are defined:

1 –Open

2 –Update

3 –Notification

4 –Keepalive



After a transport protocol connection is established, the first message sent by each side is an OPEN message. If the OPEN message is acceptable, a KEEPALIVE message confirming the OPEN is sent back. Once the OPEN is confirmed, UPDATE, KEEPALIVE, and NOTIFICATION messages may be exchanged.

The following describe each of the Open message fields:

Version ： This 1octet unsigned integer number indicates the BGP version number of the originator.

My Autonomous System ： This 2-octet unsigned integer indicates the Autonomous System number of the sender.

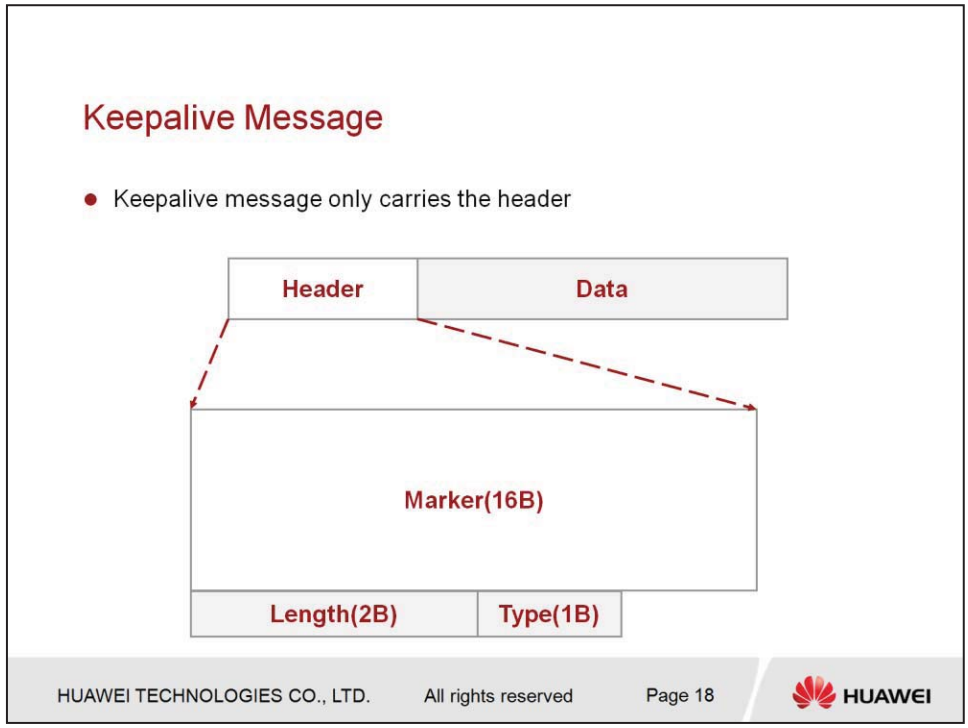
Hold Time： This 2-octet unsigned integer indicates the number of seconds that the sender proposes for the value of the Hold Timer. Upon receipt of an OPEN message, a BGP speaker MUST calculate the value of the Hold Timer by using the smaller of its configured Hold Time and the Hold Time received in the OPEN message. It is the maximum number of seconds that may elapse between the receipt of successive KEEPALIVE and/or UPDATE messages. The value of hold timer increases from

0 to the hold time value. The hold timer will be reset to 0 when the Keepalive or Update message is receipt. The neighbor will be declared dead when the hold timer expired.

BGP Identifier： indicates the router ID of the sender. This value is determined during the handshake operation between the BGP peers. The value of the BGP Identifier is the same for every local interface and every BGP peer.

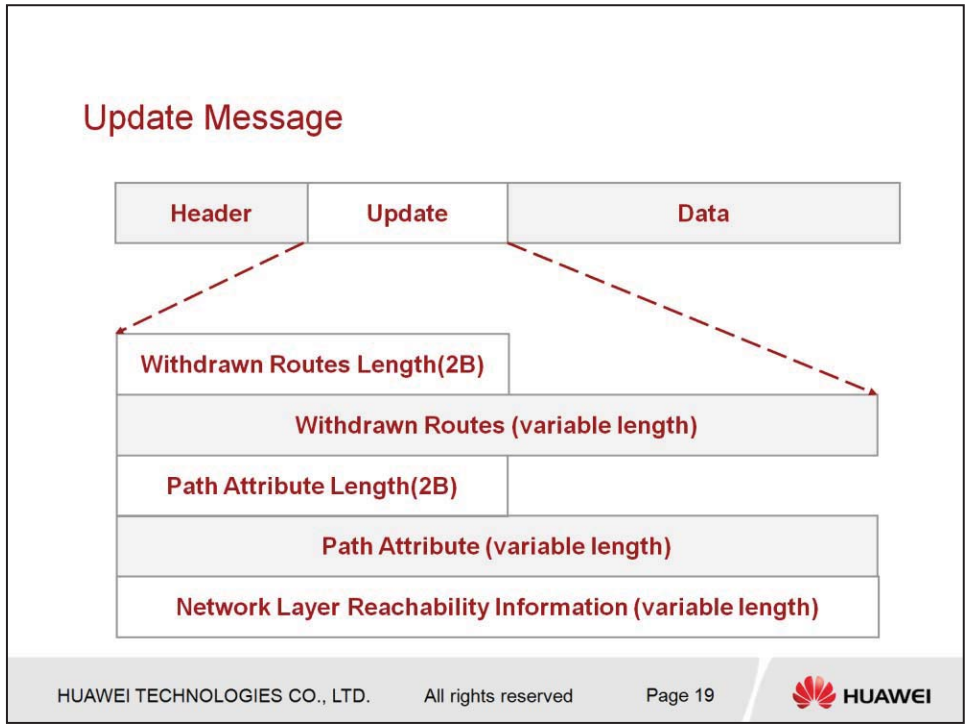
Optional Parameters Len： indicates the total length of the optional parameters in bytes. A length value of 0 indicates that no optional parameters are present.

Optional Parameters: indicates a list of optional parameters used in BGP neighbor session negotiation. This field is represented by one or several triplet <Parameters Types, Parameter Length, Parameter Value> with lengths of 1 byte, 1 byte and variable length, respectively.You can refer the optional parameters from RFC3392.



KEEPALIVE messages are sent periodically between BGP neighbors to ensure that the connection is valid. Keepalive message consists of only message header and has a length of 19 octets. The KEEPALIVE messages are sent at a rate that ensures that the hold time will not expire When the BGP connection has been formed between a router and its neighbor, the Keepalive message will be sent periodically to the peer at every keepalive interval. This is to ensure the availability of the connection.

The recommended KEEPALIVE rate is one-third of the Hold Timer value. By default, the Keepalive interval is 60s while the hold time interval is 180s. The value of hold timer increases from 0 to the hold time value. The hold timer will be reset to 0 when the Keepalive or Update message is receipt. The neighbor will be declared dead when the hold timer expired.



Update messages are used to transfer routing information between the BGP peers. Update message consists of the following field:

Withdrawn Routes Length: (2 bytes unsigned integer) indicates the length of the withdrawn route. A withdrawn Routes Length of 0 indicates that no routes are to be withdrawn and that no Withdrawn Routes field is included in the message.

Withdrawn Routes: (Variable Length) indicates a list of routes to be withdrawn. Each route in the list is described with a (Length, Prefix) tuple in which the Length is the length of the prefix and the Prefix is the IP address prefix of the withdrawn route. For example, <19, 198.18.160.0> indicates network 198.18.160.0 255.255.224.0.

Path Attribute Length: (2 bytes unsigned integer) indicates the total length of the Path Attribute field in octets. A value of zero indicates that the path attributes field is empty.

Path Attributes: (variable length) lists the attributes associated with the NLRI. Each path attribute is a variable-length triple of (Attribute Type, Attribute Length, Attribute Value).

Network Layer Reachability Information: (variable length) consists of a list of (Length, Prefix) tuples in which the format is

the same as withdrawn route field. The Length

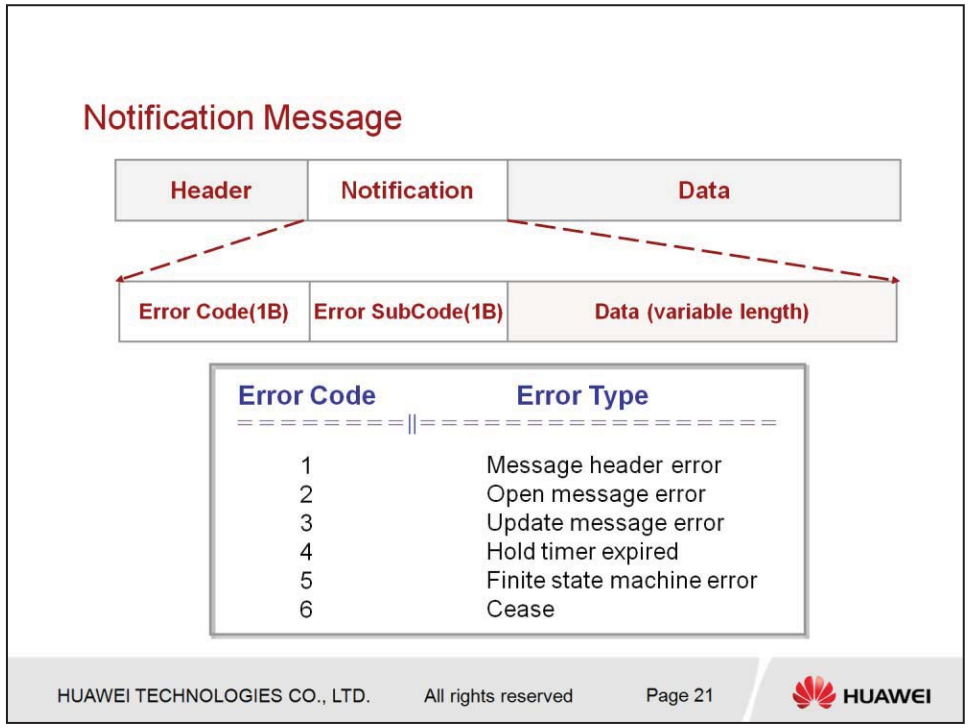
indicates the length in bits of the following prefix, and the Prefix is the IP address prefix of the NLRI.

The smallest Update information is 23 bytes (19 bytes of header + 2 bytes of withdrawn route length + path attribute length). This type of update information is called End-of-RIB and it is used in BGP GR.

One UPDATE message can advertise one route only at one time, but it can also carry multiple path attributes.

One UPDATE message can also advertise multiple routes at one time, but the path attributes must be the same.

One UPDATE message can list multiple withdrawn routes at one time.



A notification message is used when error occurs or the peer connection is stopped. This message carries various error codes (e.g. timer expiry), error subcode and error information.

Errorcode: A 1 byte field indicaties the type of error. Every errors is identified by the unique error code. Every error code can contain one or more error sub codes. If no appropriate

Error Sub code is defined, a zero value is used for the Error Subcode field.

Errsubcode：

Message Header Error Sub Codes:

1 – Connection not synchronized.

2 – Incorrect message length

3 – Incorrect message type

Open Message Error Sub Codes:

1 – Unsupported Version Number.

2 – Incorrect Peer AS.

3 – Incorrect BGP Identifier.

4 – Unsupported Optional Parameter.

5 – RFC1771 defines it as Authentication Failure. It is deprecated in RFC4271.Please refer to RFC1771/RFC4271

6 – Unacceptable Hold Time.

Update Message Error Sub Codes:

1 – Malformed Attribute List.

2 – Unrecognized Well-known Attribute.

3 – Missing Well-known Attribute.

4 – Attribute Flags Error.

5 – Attribute Length Error.

6 – Invalid ORIGIN Attribute

7 – RFC1771 defines it as AS Routing Loop. It is deprecated in RFC4271.Please refer to RFC1771/RFC4271.

8 – Invalid NEXT_HOP Attribute.

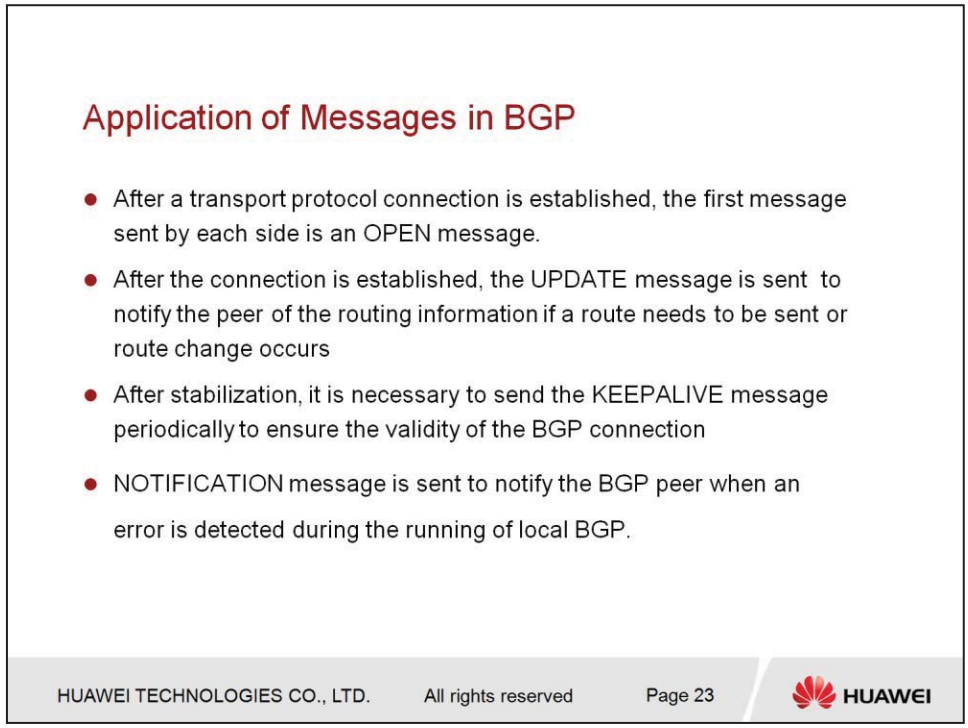
9 – Optional Attribute Error

10 – Invalid Network Field.

11 – Malformed AS_PATH.

Data: This variable-length field is used to diagnose the reason for the NOTIFICATION. The contents of the Data field depend upon the Error Code and Error Subcode. Note that the length of the Data field can be determined from the message Length field by the formula:

Message Length = 21 + Data Length. The minimum length of the NOTIFICATION message is 21 octets (including message header).



BGP uses TCP port 179 to establish the connection with its peer. Similar to the establishment of TCP connection, BGP uses a series of session and handshakes to establish the BGP connection. TCP uses the handshake negotiation to advertise the parameters like port. The handshake negotiation parameters of BGP include BGP version, hold timer of BGP connection, local router ID, authentication information and so on. These parameters are included in the Open message.

After the BGP connection is formed successfully, the Update message is sent to advertise the routing information to the peer if there is a route to be sent. Update message carries the attribute of the route when it is used to distribute the routing information to the peer. This attribute information can help the peer to select the best route. Update message can also be used to inform the changes to the BGP peer when the route of the local BGP changes.

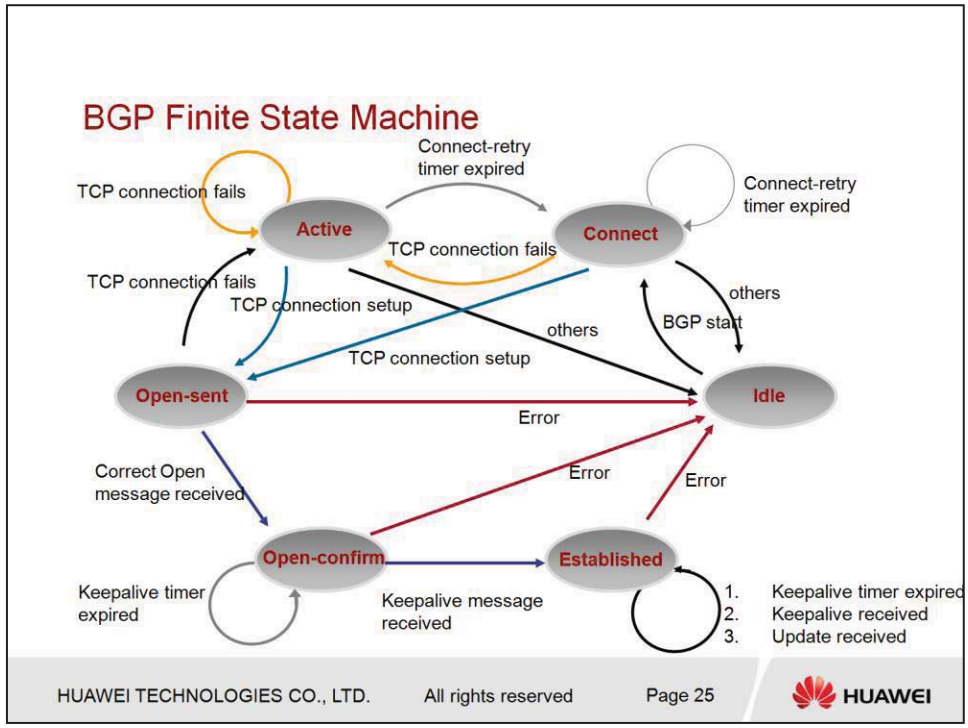
After exchanging the routing information for a period of time between the local BGP and the peer BGP, the status become stable when no new route to be advertised. At this moment,

Keepalive message is sent periodically to verify the validity of the BGP connection. When the hold time for a particular peer is

expired and the local BGP still doesn't receive any

BGP message from its peer, this BGP connection will be regarded as invalid. As a result, the BGP connection is closed and the local BGP will withdraw all the BGP routes learnt from that BGP peer.

A Notification message is sent to notify the BGP peer when an error is detected during the running of the BGP. For example, the local BGP does not support the BGP version of the peer; the local BGP receives the Update message with illegal structure from the peer and so on. Besides, the local BGP that exits the BGP connection will also send a Notification message.



Idle： This is the first state of the BGP connection in which BGP is waiting for a start event.

Example of start event are establishing a BGP session through router configuration or resetting an already existing session. After the Start event, BGP initializes its resources, resets a ConnectRetry timer, initiates a TCP connection, and starts listening for a connection that may be initiated by a remote peer. BGP then transitions to a Connect state.

In case of errors, BGP falls back to the Idle state.

Connect: In this state, BGP establish the firs TCP connection. If the TCP connection is successful, the state transitions to OpenSent (this is where the OPEN message is sent). If the connection is fail, the state transitions to Active. If the Connect Retry timer expires, the state remains in the Connect stage, the timer is reset, and a TCP connection is initiated again.

Active: In this state, BGP always attempt to establish the TCP connection. If the Connectretry timer expires, it will return to the connect state. Otherwise, it will enter the OpenSent state. If the TCP connection fails, it will remain in the Active state and keep initiating the TCP connection. In addition, BGP continues to listen for a connection that might be initiated from another peer.

The state might go back to Idle in case of other events, such as a Stop event initiated by the system or the operator.

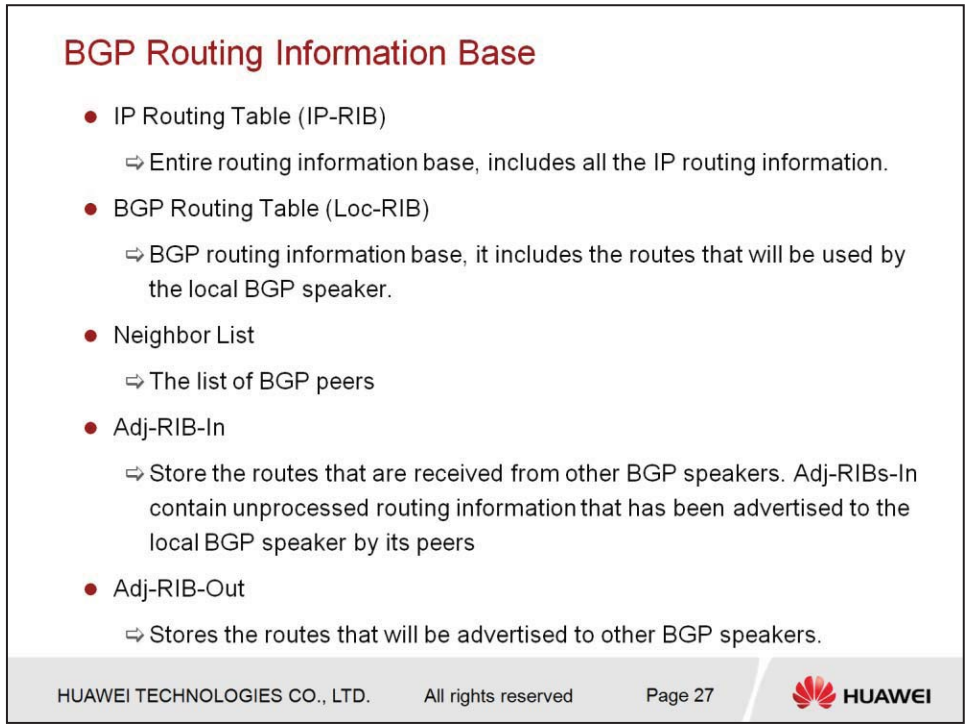
OpenSent: In this state, BGP connection has been established. The Open message has been sent, and BGP is waiting to hear an Open message from its neighbor. The OPEN message is checked for correctness. In case of errors, such as a bad version number or an unacceptable AS, the system sends an error NOTIFICATION message and goes back to Idle. If there are no errors, BGP starts sending KEEPALIVE messages and resets the KEEPALIVE timer. Meanwhile, it will enter the OpenConfirm state.

OpenConfirm： In this state, the BGP process waits for a Keepalive or Notification message. If a Keepalive is received, the state transitions to Established. If a Notification is received, or a TCP disconnect is received, the state transitions to Idle.If the Hold timer expires, an error is detected, or a Stop event occurs, a Notification is sent to the neighbor and the BGP connection is closed, changing the state to Idle.

Established：This is the final stage in the neighbor negotiation. At this stage, BGP starts exchanging UPDATE packets with its peers. The Hold Timer restarts at the receipt of an UPDATE or KEEPALIVE message. If the system receives any NOTIFICATION message (if an error has occurred), the state falls back to Idle. The UPDATE messages are checked for errors, such as missing attributes, duplicate attributes, and so on. If errors are found, a NOTIFICATION message is sent to the peer, and the state falls back to Idle. If the Hold Timer expires, or a disconnect notification is received from the transport protocol, or a Stop event, the system falls back to the Idle state.

By using"display bgp peer" command, we always observe these 2 states: Active and Established.

The neighbor state Active indicates that TCP connection fail to establish. This could be due to the inability of a neighbor to reach the IP address of its peer or mistakes in the configuration. As a result, the neighbors not able to exchange the routing information.



•IP Routing Table (IP-RIB)

Entire routing information base, includes all the IP routing information.

•BGP Routing Table (Loc-RIB)

BGP routing information base, it includes the routes that will be used by the local BGP speaker.

•Neighbor List

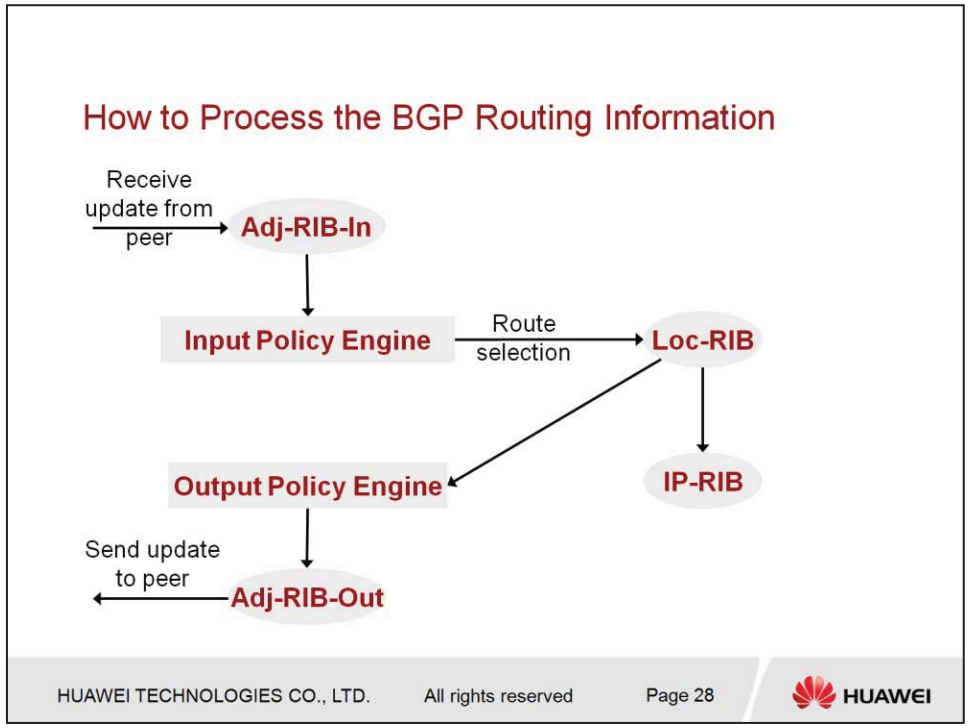
The list of BGP peer

•Adj-RIB-In

Store the routes that are received from other BGP speakers. Adj-RIBs-In contain unprocessed routing information that has been advertised to the local BGP speaker by its peers.

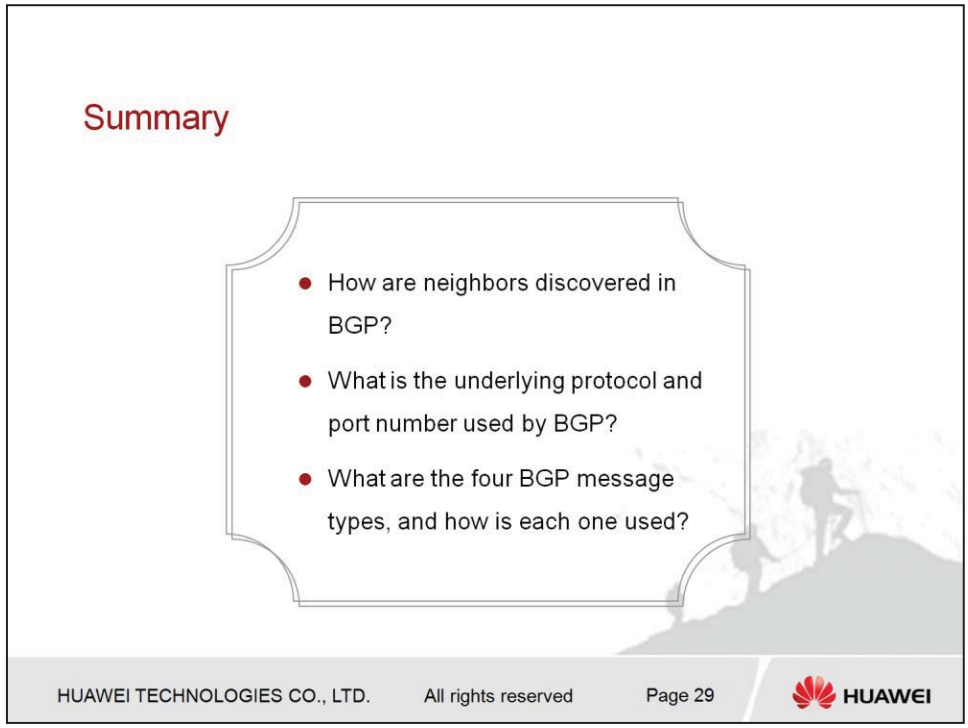
Adj-RIB-Out

Store the routes that will be advertised to other BGP speakers.



Upon receiving the update data packet from the peer, the BGP speaker will store this update data packet into the BGP Routing Information Base (RIB) for route selection. The Adj-RIB-In which is associated with each individual peer of the BGP speaker is specified. Then, the update data packets are being manipulated or filtered by the Input Policy Engine associated with the peer. The router will then execute the route selection criteria and a best path is selected for every IP prefix. The Loc-RIB contains only the preferred routes that have been selected as the best path to each available destination. This best route is then sent to the local IP-RIB, and it is under the installation consideration.

When multiple routes to the same IP subnet exist, the best path and all the equal cost paths are sent to the IP-RIB for consideration. In addition to the best route received from the BGP peer, LOC-RIB also contains route that the local router originates (if configured to do so) about the network inside its autonomous systems. This is how an AS advertises its internal networks to the outside world. Before the contents of the LOC-RIB are advertised to other BGP peers, it must be processed by the output policy engine. Only the route that has been filtered by the output policy engine can be installed in the Adj-RIBOut.



Q. How to discover the neighbor in BGP？

A. BGP does not use any neighbor discovery mechanism. Therefore, we have to specify all the neighbors manually.

2. Name the underlying protocol and the port number that BGP used.

A. BGP uses TCP port 179.

3. What are the four BGP message types, and how is each one used?.

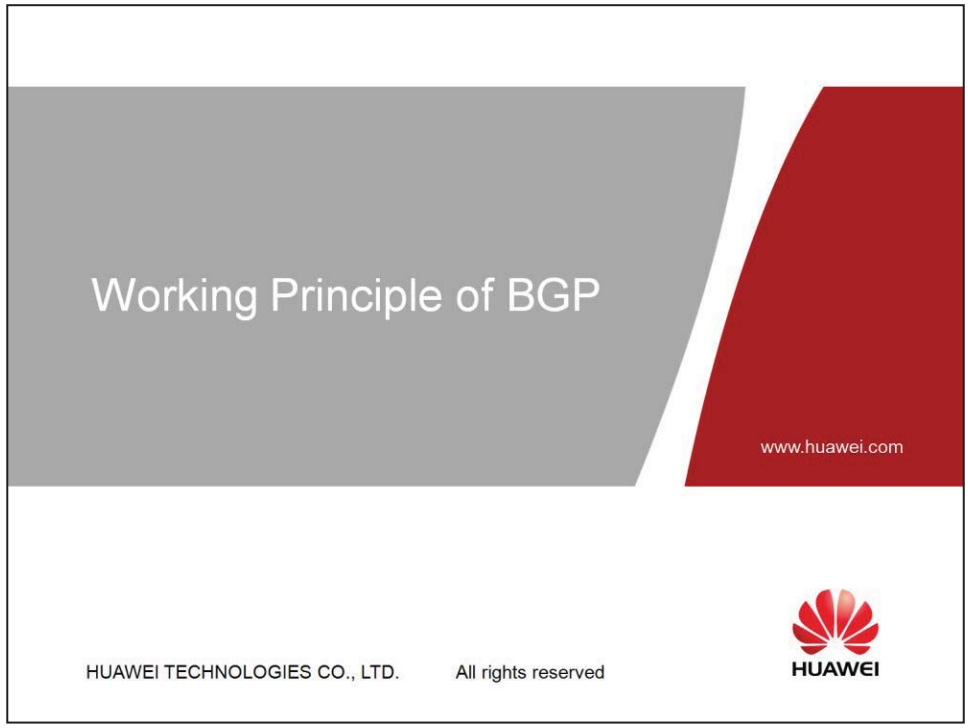
A: OPEN: The Open message includes BGP version number, AS number of the originator and so on. After the TCP session is established, both neighbors send Open messages to each other and determine whether the neighbor relationship can be formed.

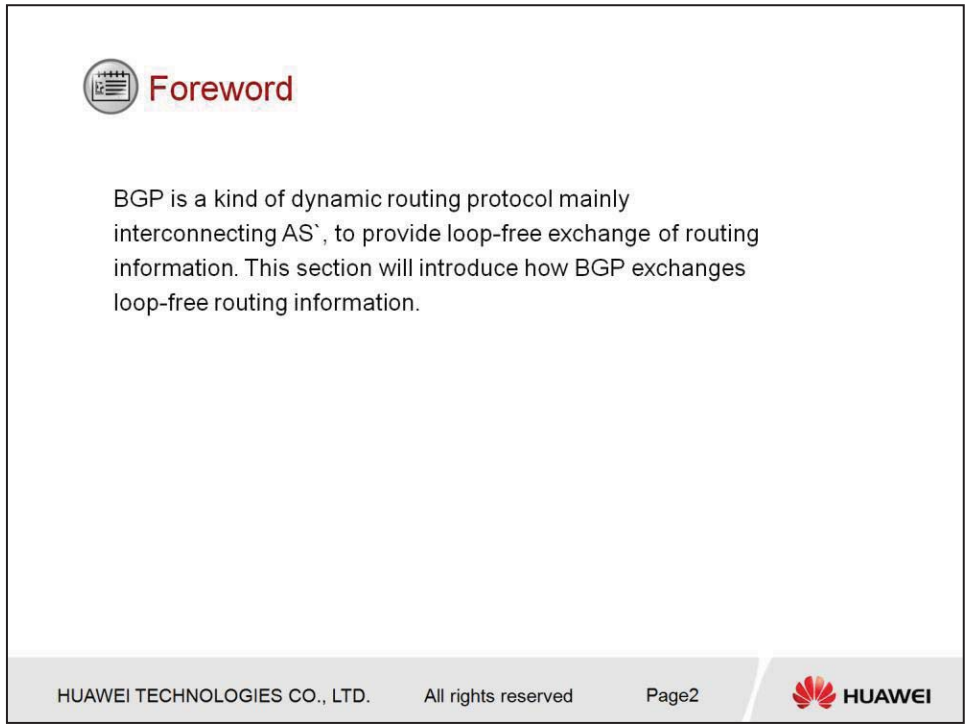
KEEPALIVE: Keepalive message is exchanged periodically to maintain the neighbor relationship. It is used to verify the connectivity of the peer.

NOTIFICATION: Notification message is the error checking mechanism used in BGP. BGP speaker will send the notification message when an error occurs. This will always cause the BGP connection to close.

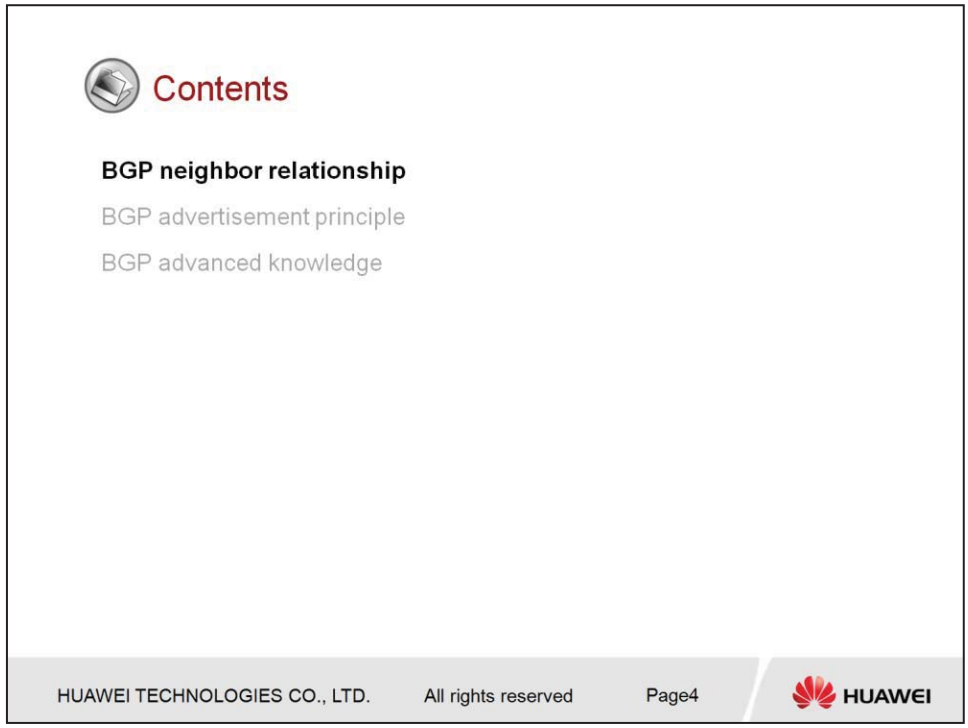
UPDATE: Among the 4 message types, update message is the most important message in the BGP system. Update message is used to exchange the routing information between the peers. It consists of all the information used by BGP to form the loop-free network structure.

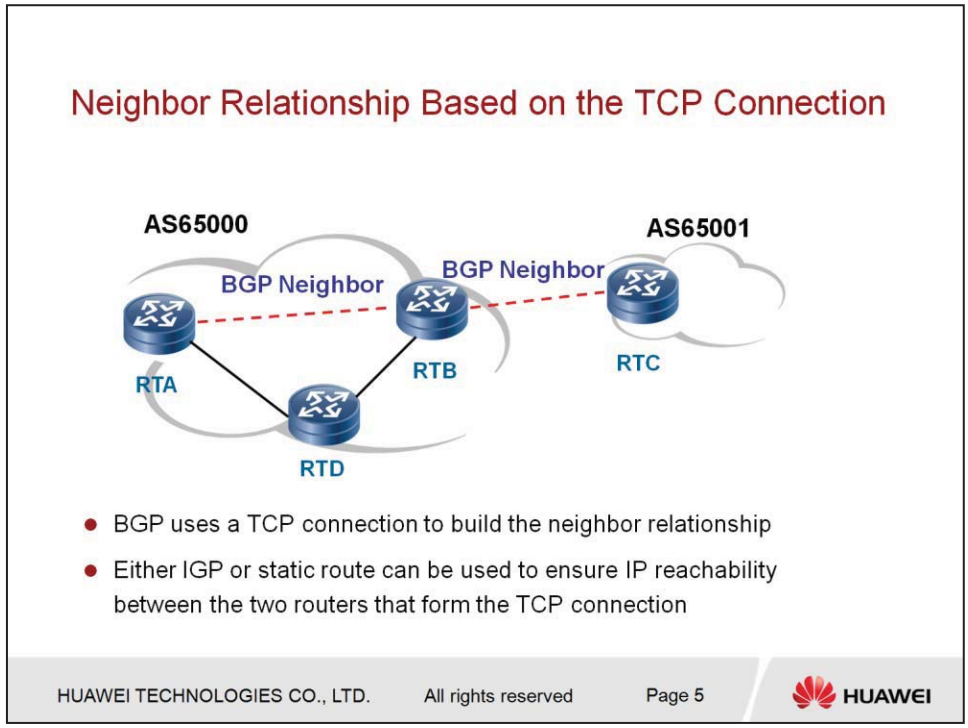
It comprises of Network Layer Reachability Information (NLRI), path attributes and withdrawn route fields.


Working Principle of BGP
www.huawei.com
HUAWEI TECHNOLOGIES CO., LTD.
All rights reserved
HUAWEI


Foreword
BGP is a kind of dynamic routing protocol mainly interconnecting AS`, to provide loop-free exchange of routing information. This section will introduce how BGP exchanges loop-free routing information.
HUAWEI TECHNOLOGIES CO., LTD. All rights reserved Page2
HUAWEI

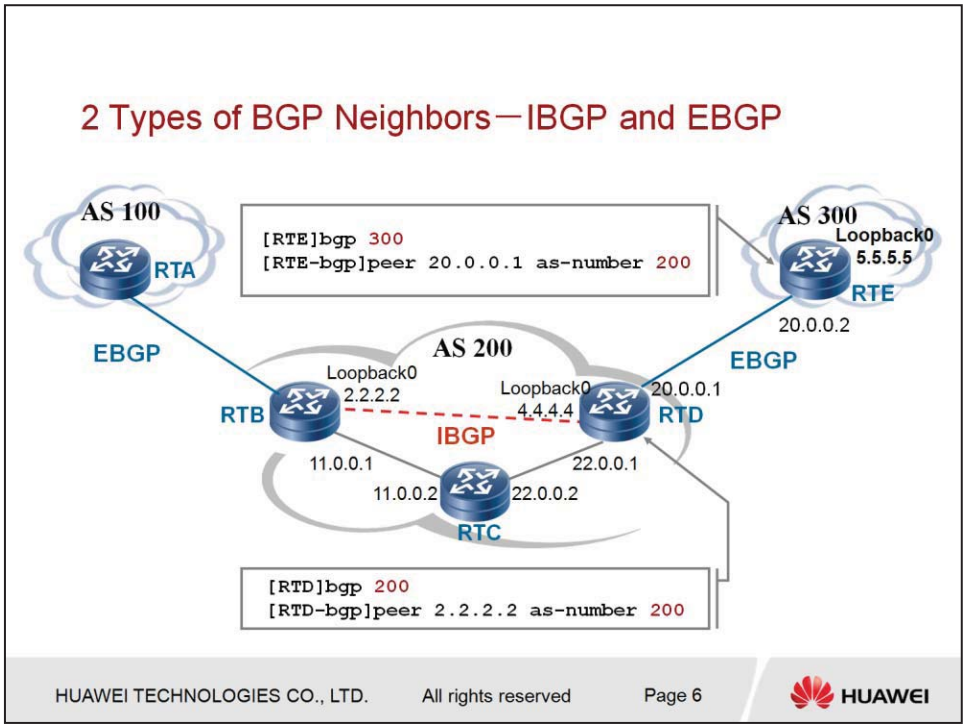

Objectives
Upon completion of this section, you will be able to :
⇨ Identify the two kinds of neighbor relationship in BGP protocol
⇨ Understand the BGP advertisement principle.
⇨ Understand how BGP advertises routing information.
HUAWEI TECHNOLOGIES CO., LTD. All rights reserved Page3
HUAWEI


Contents
BGP neighbor relationship
BGP advertisement principle
BGP advanced knowledge
HUAWEI TECHNOLOGIES CO., LTD.
All rights reserved
Page4
HUAWEI



Similar to OSPF and IS-IS, BGP learns the route through neighbor. The route learning process is slightly different between them. In OSPF and IS-IS, the neighbor relationship is established automatically. However, in BGP, we have to manually specify the IP address of the neighbor. Therefore, in BGP, 2 routers that are connected directly via physical link might not necessarily form the neighbor relationship. In contrast, 2 routers that are not connected directly via physical link might form the neighbor relationship.

To establish the neighbor relationship, BGP rely on the TCP connection. The IP connectivity between 2 BGP routers that form the TCP connection must be assured. We can use the protocol other than BGP to realize the IP connectivity. In other words, either IGP or static route can be used for this purpose. For convenience, we call the use of either IGP or static route to implement the IP connectivity as IGP connectivity or IGP reachability.



BGP runs in the following two modes: IBGP (Internal BGP), EBGP (External BGP)

•If two peers that exchange BGP messages belong to the same AS, they are Internal BGP (IBGP), such as RTB and RTD.

•If two peers that exchange BGP messages belong to different AS, they are External BGP (EBGP), such as RTD and RTE.

Although BGP runs between ASs, it is necessary to establish BGP connection between different border routers of an AS. Only in this way, the routing information can be transmitted to the whole network. To establish the communication between AS100 and AS300, we need to establish the IBGP connection between RTB and RTD.

The basic configuration of BGP is as follow:

Enable BGP (specify the local AS number), enter BGP view

[Router A] bgp as-number

Using the bgp command, you can enable BGP and enter BGP view.

By default, the BGP is disabled.

One router can run in only one AS. This means that one router

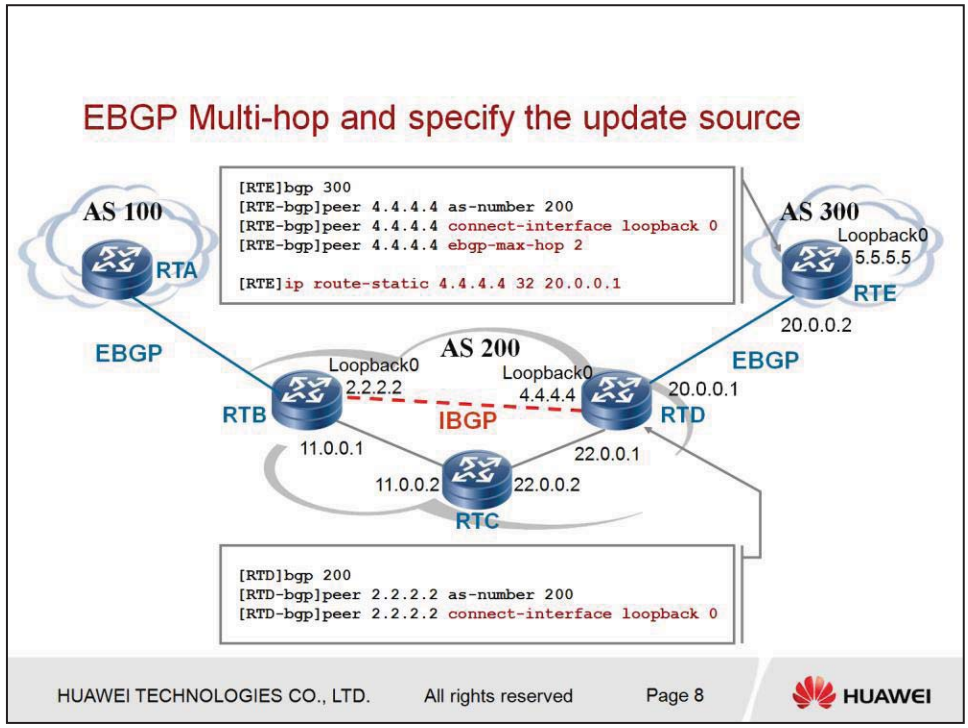
can be specified with only one AS-number.

Specified the peer IP address and its AS number

[Router A-bgp] peer { group-name | ipv4-address | ipv6-address} asnumber as-number

By using peer as-number command, you can configure the AS number of a peer or peer group. Using the undo peer as-number command, you can delete the AS number of a peer or peer group

By default, the opposite end of the peer group has no AS number.



The direct physical connection is not necessarily exists between the IBGP peers.

However, the TCP connection must be established between them. We usually specify the IBGP neighbor by using the loopback interface to ensure the reliability of the route advertisement of IBGP peers. In the case where loopback interface is used, we must specify the source interface of route update packet.

peer { group-name | peer-address } connect-interface interface-name

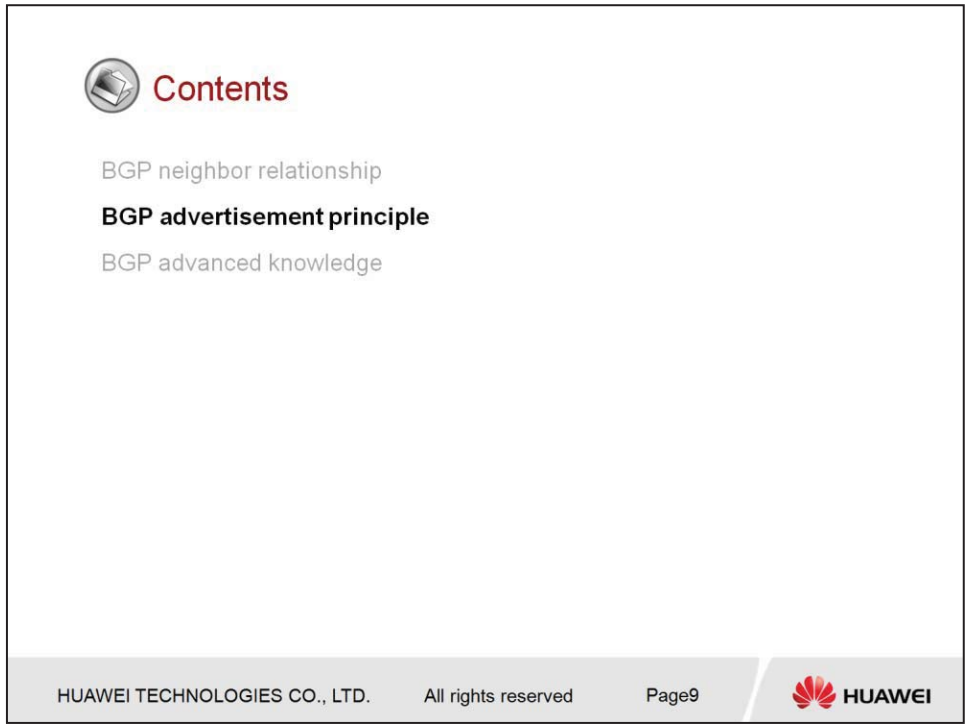
In most of the cases, there is a direct physical link between the EBGP peers.

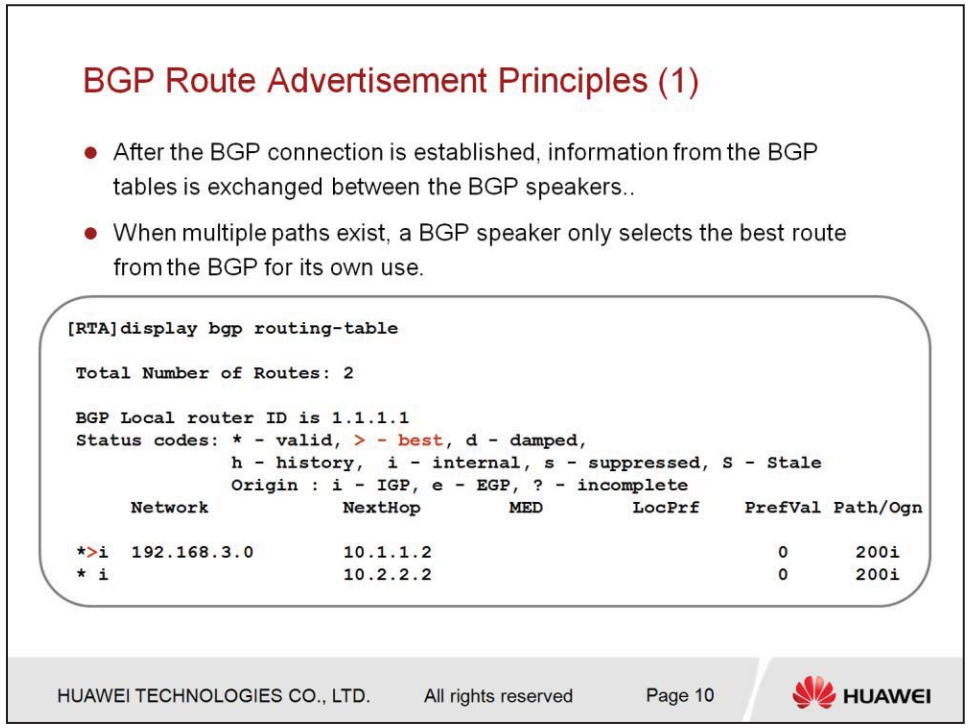
However, it is hard to realize this in some cases. To establish the EBGP connection with the peer on the indirectly connected network, we need to modify the maximum hop of EBGP packet by using the following command:

peer { group-name | peer-address } ebgp-max-hop [ttl]

“ttl” is the maximum hop value. It is in the range of 1 to 255. If you specify the maximum hop as 1, you can not establish the EBGP connection with the peer on the indirectly connected network.

Besides the “ebgp-max-hop” command, we need to ensure that the loopback interface of the BGP peer is reachable.


Contents
BGP neighbor relationship
BGP advertisement principle
BGP advanced knowledge
HUAWEI TECHNOLOGIES CO., LTD.
All rights reserved
Page9
HUAWEI



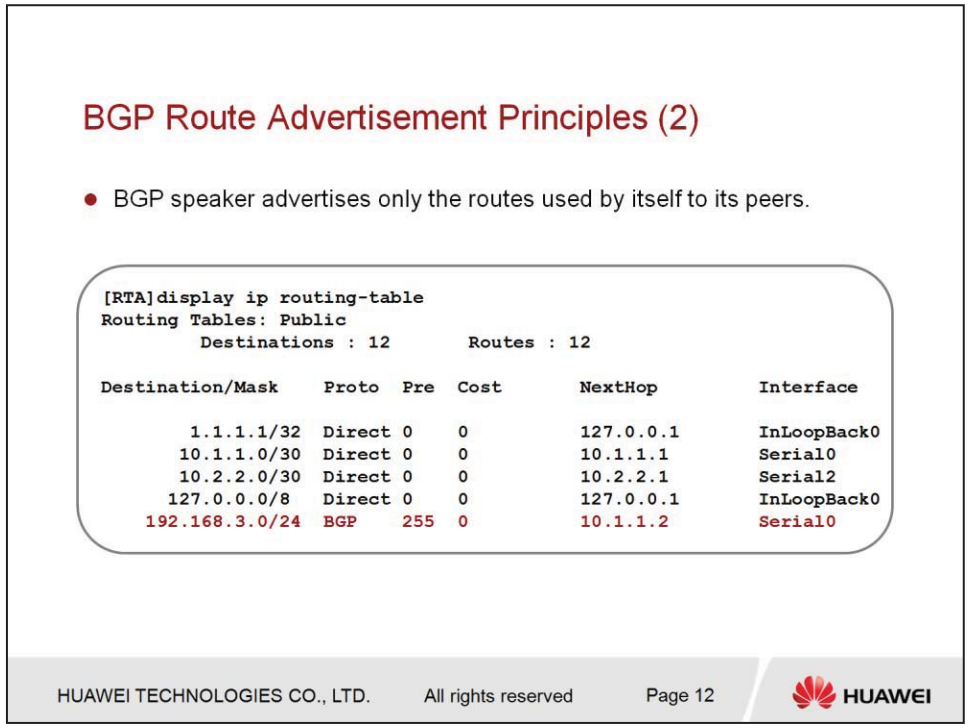
Under the normal circumstance, when there is more than one alternatives route to the same IP subnet, the BGP speaker will select the best route for its own use.

The best route is the candidate for installation in the IP routing table. However, before a route can be installed, the router will check if there is there is any other routing protocol that has information about the same subnet. If the subnet is known via different sources, the router uses the route preference to determine which source is more trustworthy. The router will install the route with smaller route preference value. That is to say the router will select also the best route for its own use and the best route of BGP speaker might not the best route for the router. For example, there are 2 routes to the same IP subnet. One is the best route selected by the BGP speaker and another is the static route. In this case, router will install the static route into the IP routing table because it has lower route preference value compare to BGP. Therefore, the best route of BGP will not be installed into the IP routing table.

As shown in the slide above, there are 2 routes towards network 192.168.3.0 on RTA. The next hops are 10.1.1.2 and 10.2.2.2 respectively.

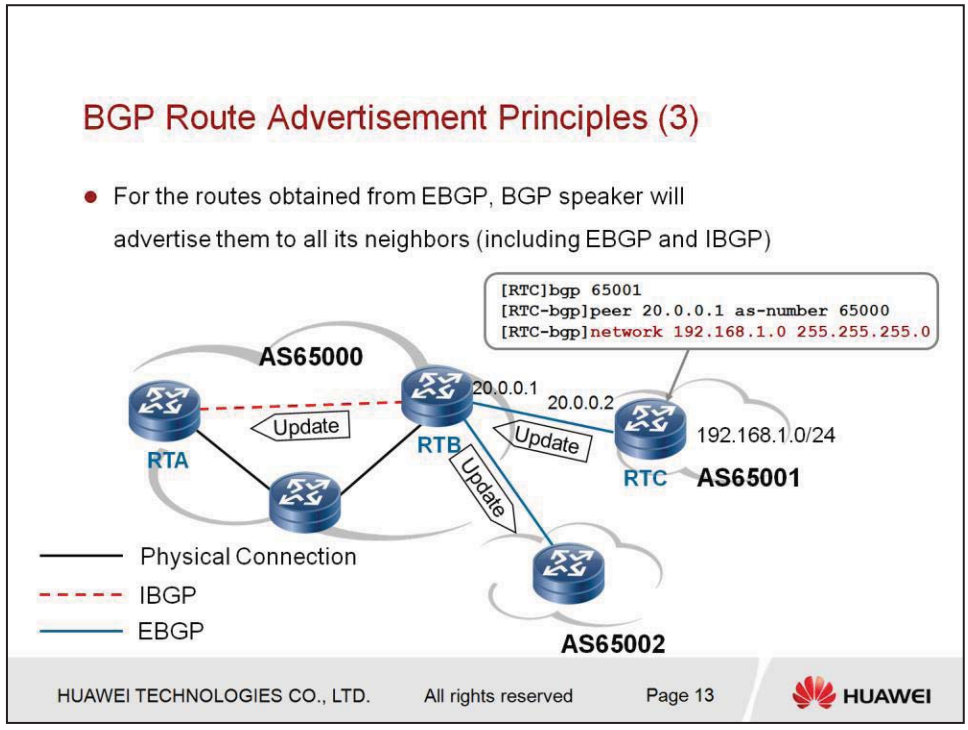
Base on the route selection criteria (will learn later), BGP will select a best route which is indicated with “>” sign. After that, the router will check if any better routes of other routing protocols exist. If exists, the better route will be installed in the IP routing table. Else, the best route of BGP will be installed into the IP routing table.

For example, the static route to 192.168.3.0 is better compare to BGP route to the same IP subnet (the route preference of static route is 60 while the route preference for BGP is 255, lower route preference value is more trustworthy).

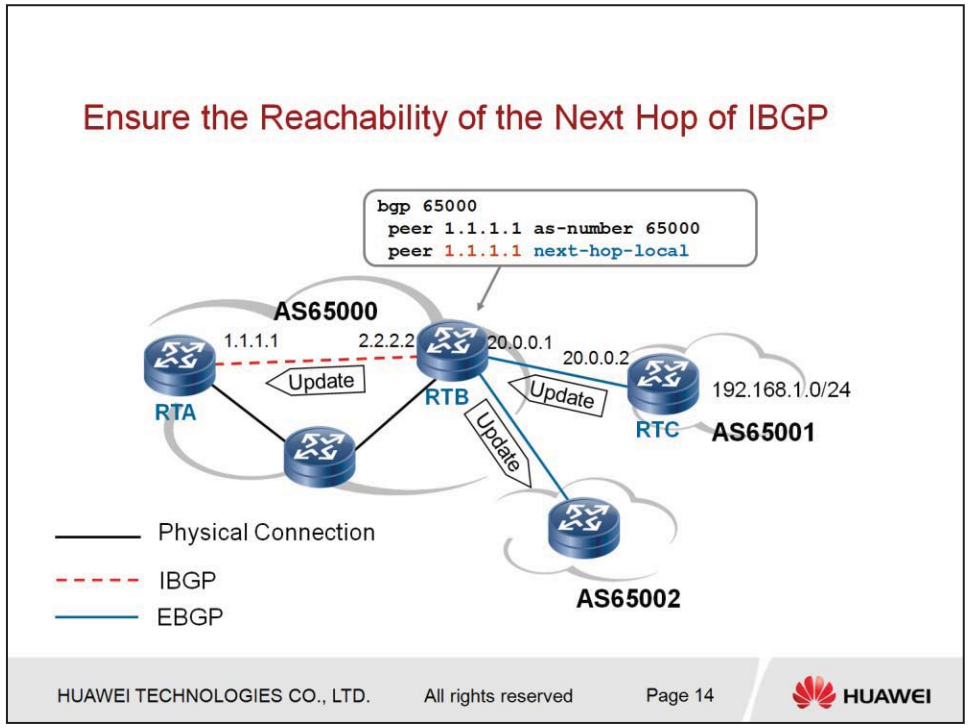


BGP speaker advertises only the best routes used by itself to its peers. This means that it only advertises the BGP routes which are installed in the IP routing table to its peers.

Once the best route (“>”) that has been selected by BGP is installed in the IP routing table, BGP will send Update message which consists of that best route entry to other BGP peer. However, BGP will never sends the updates back on the same BGP session upon which it was received.



For the routes obtained from EBGP, BGP speaker will advertise them to all its neighbors (including EBGP and IBGP).



IGPs are used to exchange the routing information between the routers.

Therefore, the next hop of any router is the interface IP address of the router that announces the route. This can be easily understood. However, the BGP is used to exchange the loop-free routing information between the autonomous systems.

The autonomous system can be regarded as an “abstract router” in BGP. So, RTB will directly send the update (which has been received from RTC) to RTA without making any changes to the information in the route update packet. As a result, the next hop for RTA to reach the network 192.168.1.0/24 is therefore

20.0.0.2. It is possible that RTA does not know the route 20.0.0.2. This causes unreachability of RTA to route 192.168.1.0/24.

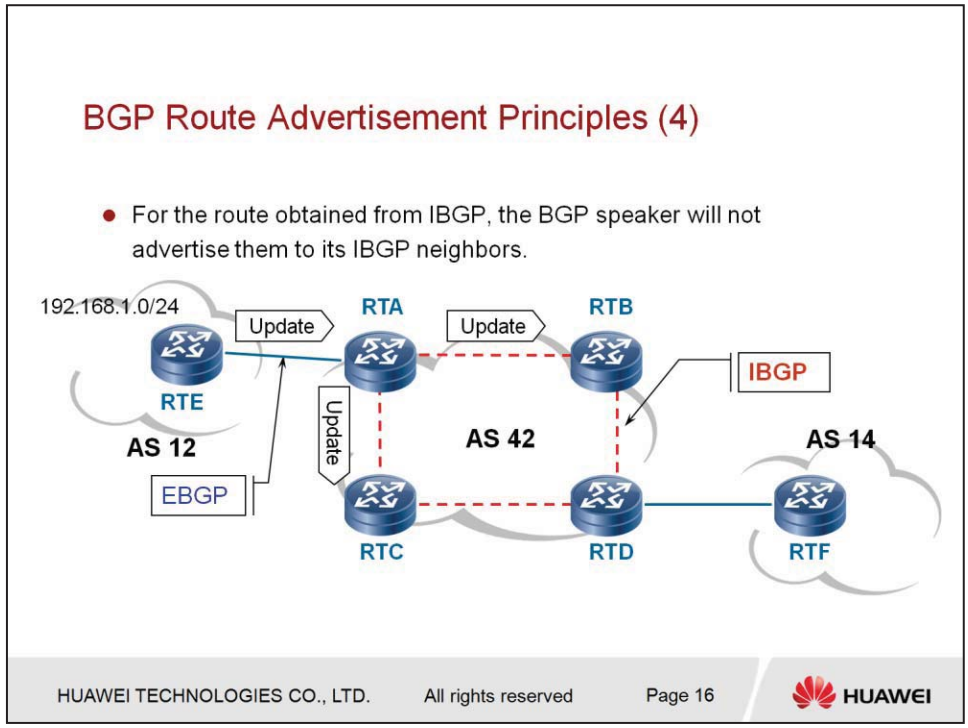
BGP provides a command that can be used to force the next hop to be the IP address of the border IBGP neighbor. This is to ensure that the IBGP neighbor can find the correct next hop information.

Configuration Command:

peer { group-name | ipv4-address } next-hop-local

By using the undo peer next-hop-local, you can restore to the default configuration.

By default, when the BGP notifies the EBGP peer of the route, it change the nexthop attribute to its local IP address. When the BGP notify the IBGP peers of the routes obtained from EBGP, it does not change the next-hop attribute of the route.



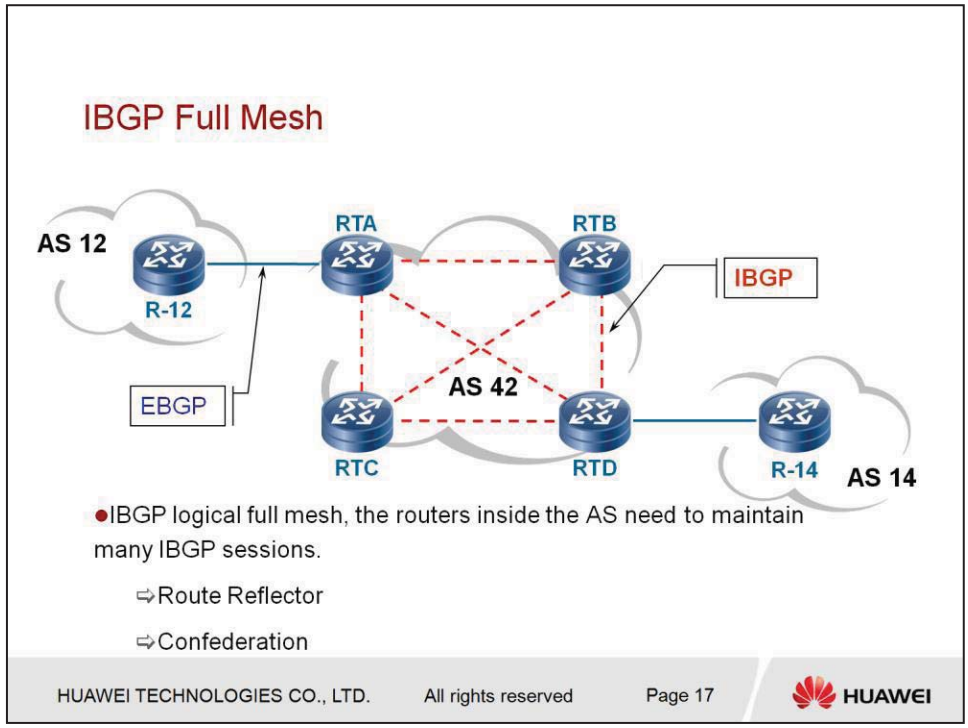
RTC learns the route update from its IBGP RTA. Assume that this rule doesn't enforced. RTC will send the route update learnt from IBGP RTA to its IBGP RTD.

RTD will send the route update learnt from IBGP RTC to its IBGP RTB. Similarly, RTB will send the route update learnt from IBGP RTD to its IBGP RTA. As a result, routing loop is generated.

Therefore, this rule is used to prevent routing loop inside an autonomous system.

However, the enforcement of this rule introduces a new problem to the network:

RTD will not able to receive the BGP route from AS12. To resolve this problem, we often establish the IBP logical full mesh connection. This means that we need to create 2 more IBGP connections for the network above. One of the IBGP connection is created between RTA-RTD and another one is created between RTB-RTC.

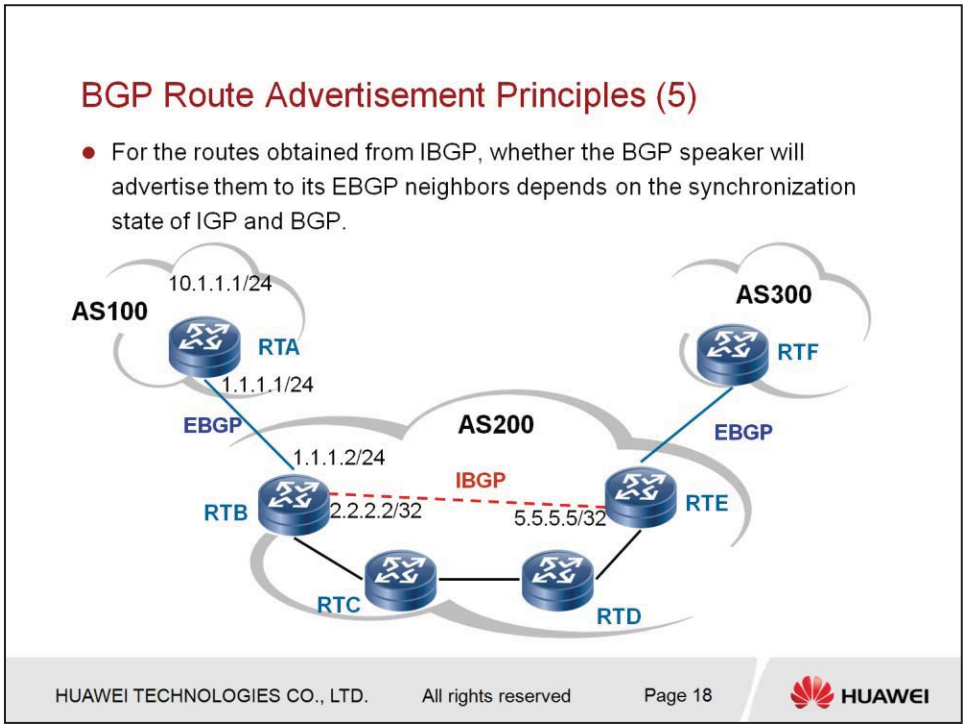


IBG Full Mesh is one of the solution for route advertisement problem caused by the IBGP split horizon rule as described in the previous slide. The disadvantage of this method is that the router need to maintain a large number of IBGP sessions and this consume a lot of resources.

Besides that, BGP provides 2 other methods for the problem caused by the IBGP split horizon rule.

Route-Reflector -- RFC 2796

Confederation -- RFC 3065



The concept of synchronization between BGP and IGP: BGP speaker will not advertise the routing information learnt from the IBGP peer to its EBGP peer unless all routers within the AS had learned about that route through the IGP.

This is known as synchronization. If a router knows about these destinations via an IGP, it assumes that the route has already been propagated inside the AS, and internal reachability is assured.

One of the main responsibilities of BGP is to transmit the network reachability information of its autonomous system to other autonomous system. AS shown in the diagram above, RTB will encapsulate the routing information towards network 10.1.1.0/24 into the BGP update message and advertise it to RTE via the TCP connection established by RTB and RTE. If RTE does not take the synchronization into account, it will directly accept this routing information and send it to RTF. If RTE or RTF has data packet to be sent to network 10.1.1.0/24, this data packet must pass RTD and RTC to reach the destination 10.1.1.0/24.

Since synchronization of IGP and BGP was not taken into account earlier, the routing tables of RTD and RTC do not contain the routing information to network 10.1.1.0/24 and result

in the data packet to be discarded when it reaches RTD.

Therefore, BGP must synchronize with IGP (RIP, OSPF and so on). In other words, a router when receives route update information from its IBGP peer will attempt to verify the internal reachability for that route before advertising it to other EBGP peers (check whether that route contains in the routing table of IGP, whether non-BGP routers can deliver traffic to that route). If that route is known via an IGP, the router announces it to other EBGP peers.Otherwise, the router treats the route as not being synchronized with the IGP and does not advertise it.

There are many solutions for the synchronization problem. The simplest one is for RTB to redistribute the BGP routing information to the IGP routing table and then the IGP routing protocol will advertise it to RTE. As a result, the synchronization is achieved. However, this is not recommended because the BGP routing table is very huge. This results in the high consumption of memory and CPU on the IGP

routers. The low and middle end equipments that cannot handle this burden will fail.

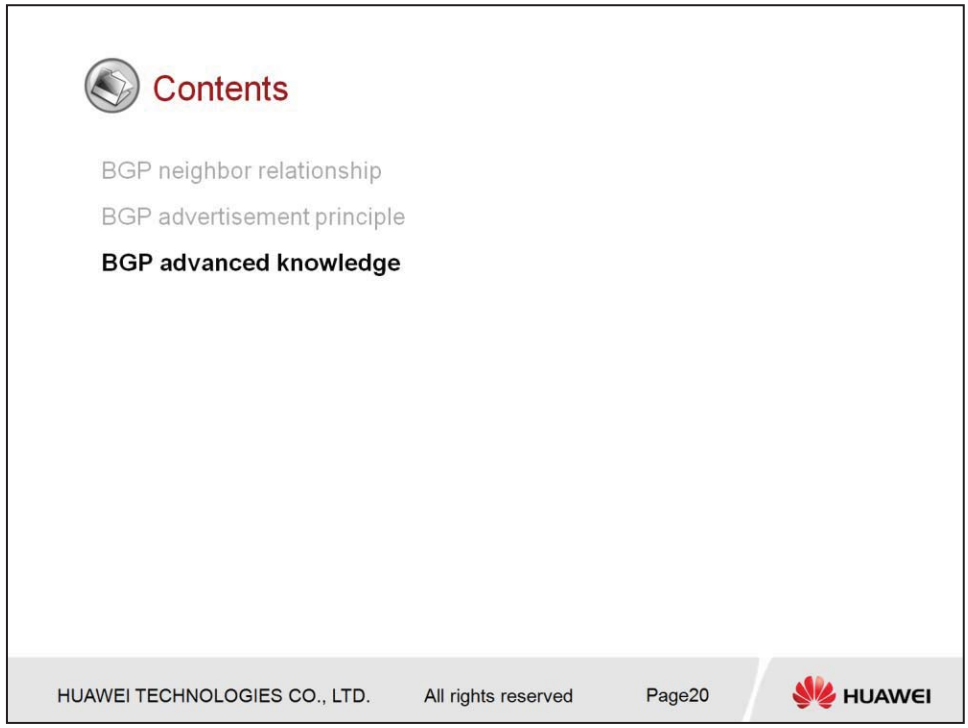
Another solution is to configure a static route to 10.1.1.0/24 on RTB, and then redistribute it into IGP routing table. The synchronization is therefore achieved.

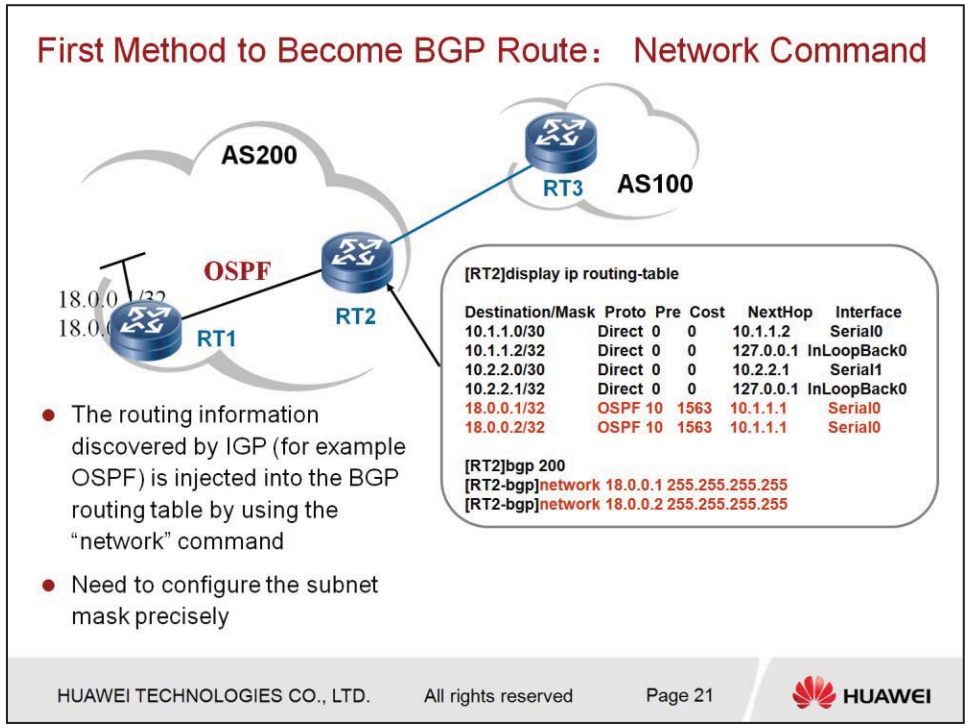
No matter what kind of method you are using, it is not suitable to be used in the large network. .By default, the VRP platform adopts “Undo synchronization” and it cannot be changed.

Certain conditions must be met in order to cancel the synchronization. We can cancel the synchronization when all the routers in the ASs establish an IBGP full connection. This means that the routers between RTB-RTC, RTB-RTD, RTB-RTE, RTC-RTD, RTC-RTE, and RTD-RTE has established the IBGP neighbor relationship via TCP connection. Since the IBGP neigbor relationship has been formed between RTB-RTD, RTD has learnt the BGP route towards 10.1.1.0/24 from RTB. when the data arrive at RTD, RTD will in turn transmit the data to RTC.

Similarly, RTC has learnt the BGP route towards network 10.1.1.0/24 from RTB because the IBGP neighbor relationship has been formed between RTB-RTC.

Therefore, RTC will in turn send the data to RTB. As a result, the data will not lost during the data transmission.


Contents
BGP neighbor relationship
BGP advertisement principle
BGP advanced knowledge
HUAWEI TECHNOLOGIES CO., LTD.
All rights reserved
Page20
HUAWEI



The main function of BGP is to transmit the loop-free routing information between the autonomous systems instead of discovering and calculating routing information. Therefore, the routing information of BGP needs to be redistributed into BGP using command.

To become BGP route, there are 2 methods : using network command or import command

In fact, we can use the aggregate command to redistribute the IGP aggregated route into the BGP routing table. The recondition to use this method is the BGP routing table must contain those detailed route first. Therefore, we do not consider this method as one of the solutions.

By using the Network command: the router will inject the routing information of IP routing table into the BGThen, the injected routing information will be sent to other peer via BGP. The precondition is that the routing information to be injected into the BGP routing table via network command must be existed in the IP routing table.

The related command:

network ipv4-address [mask | mask-length] [route-policy

route-policy-name]

ipv4-address: Specified the IPv4 network address advertised by BGP in dotted decimal format.

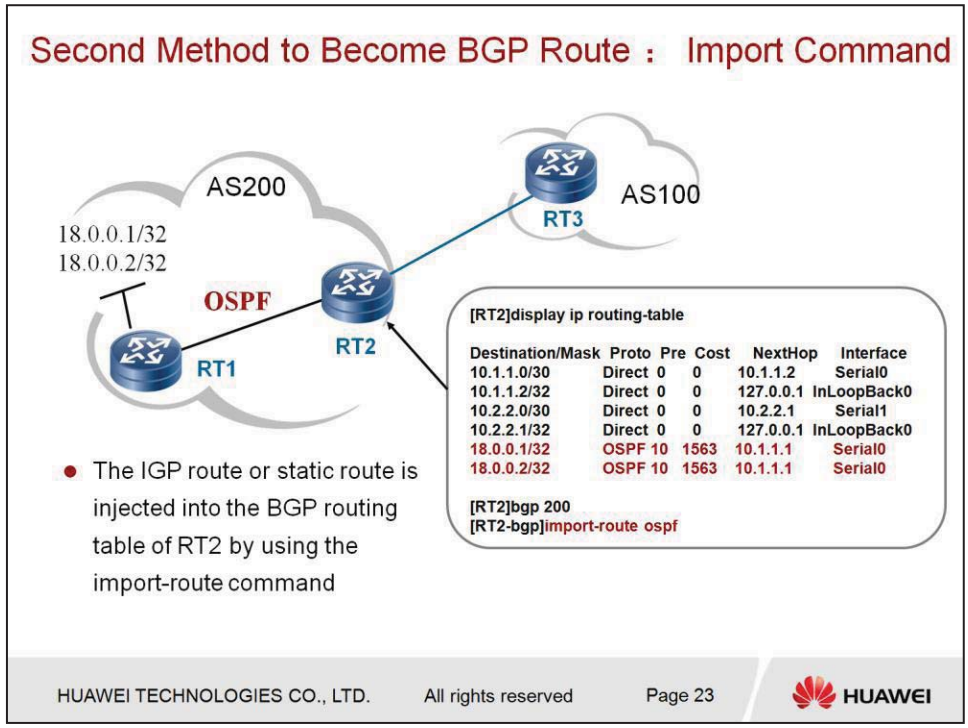
mask/mask-length: Specifies the network mask or the length of the network mask.

Default subnet mask will be used if it is not specified.

route-policy-name: Specifies the routing policy applied to the advertised routes.

By default, BGP do not advertise any of the local routes.

P routing table according to the Network specified.



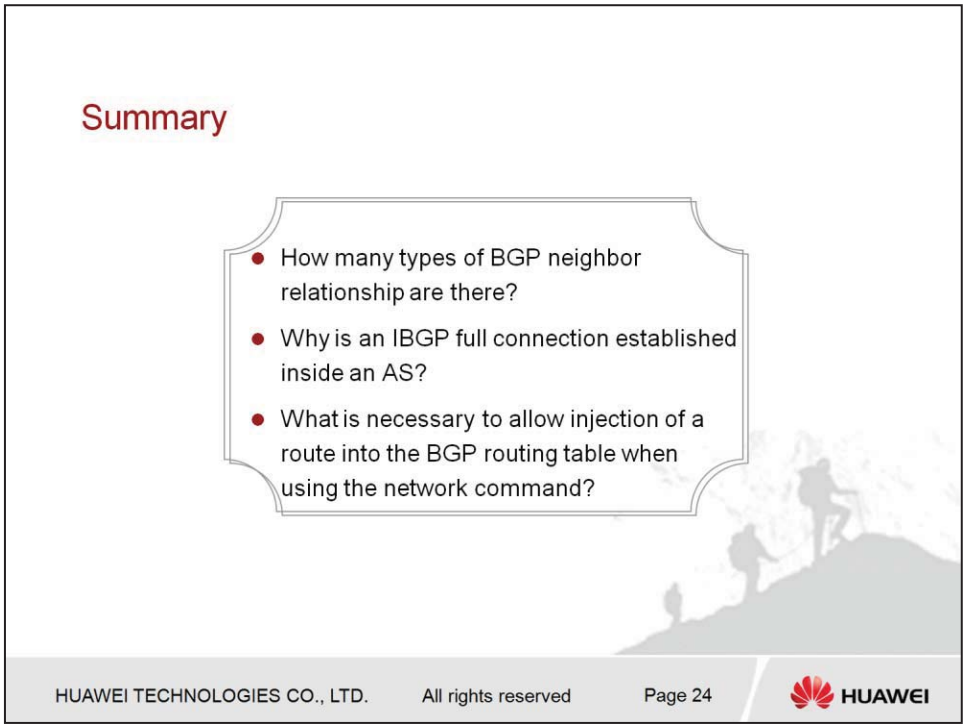
The second method is redistributing other routing protocols into BGP routing table by using the import command. Besides, we can specify the routing policy for the imported route.

import-route protocol [process-id] [med med | route-policy route-policy-name]

protocol: specifies the importable external routing protocols. At present, the importable external routing protocols include isis, ospf, static, direct and rip.

process-id: specifies the process id if the imported routing protocol are isis, ospf or rip.

Med: specifies the med metric of the imported route. The value is ranging from 0 to 65535. route-policy-name: When the routes are imported from other routing protocol, will filter those routes by using the specified routing policy.



1. How many types BGP relationship?

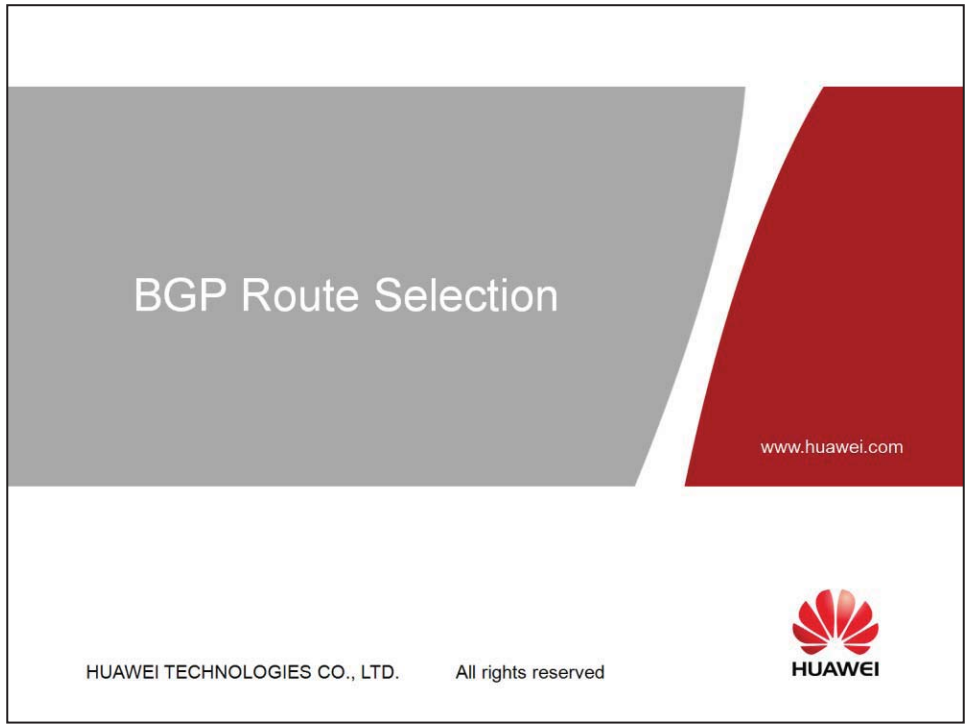
A：IBGP and EBGP. If two peers that exchange BGP messages belong to the same AS, they are Internal BGP (IBGP). If two peers that exchange BGP messages belong to different AS, they are External BGP (EBGP).

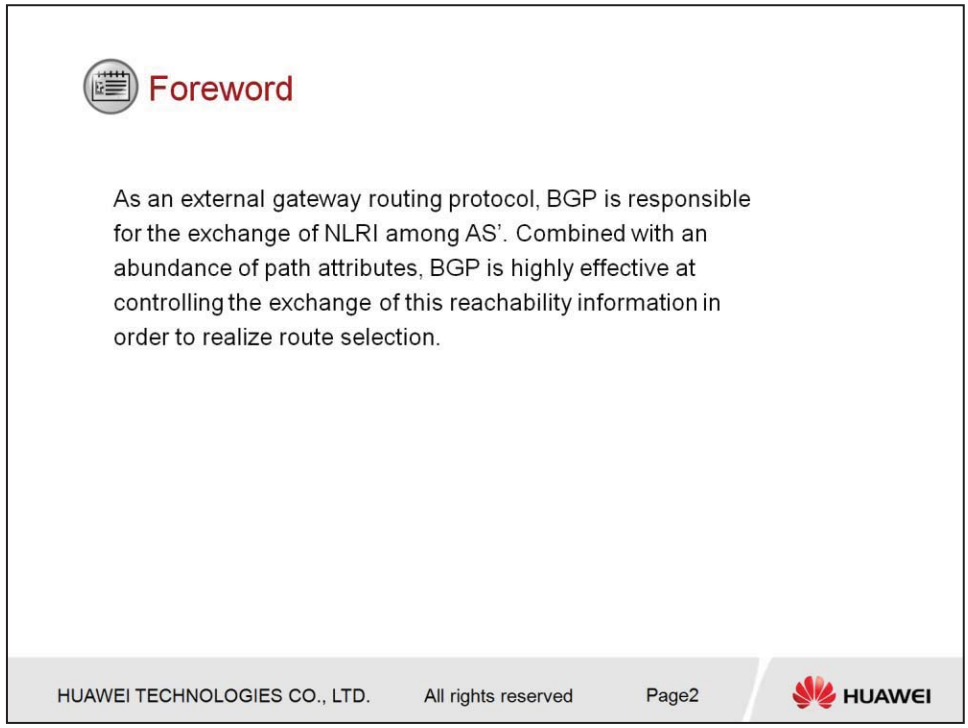
2. Why the IBGP full connection is established inside an AS?

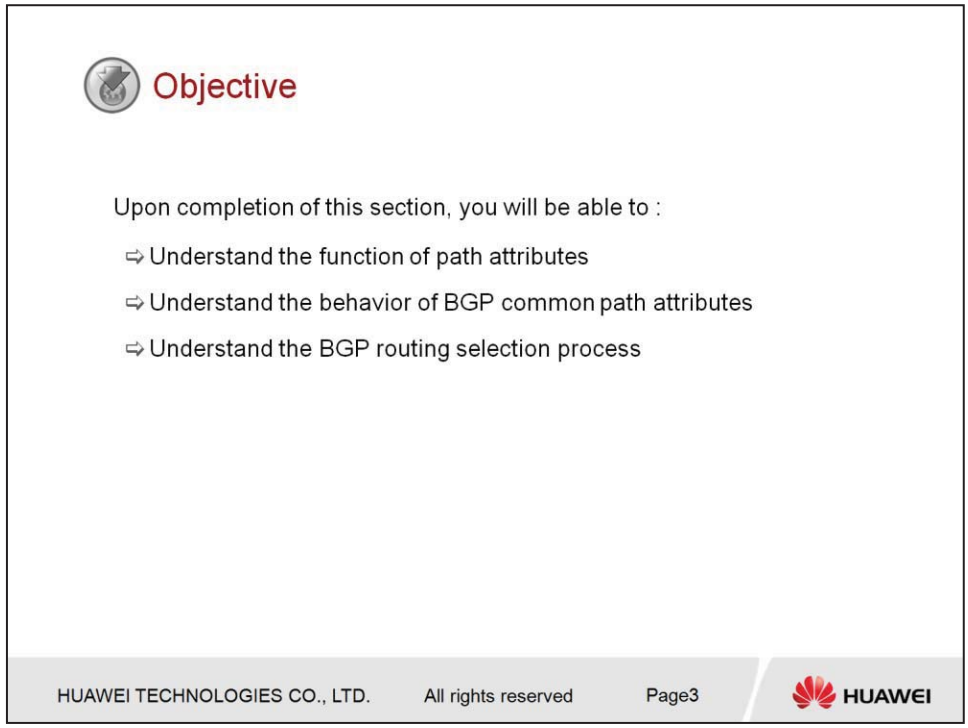
A：For the route obtained from IBGP, the BGP speaker will not advertise them to its IBGP neighbors. This is to prevent the routing loop inside an AS. So, IBGP full connection must be established between the BGP speakers inside an AS in order to transmit the routing information to all other IBGP peers successfully.

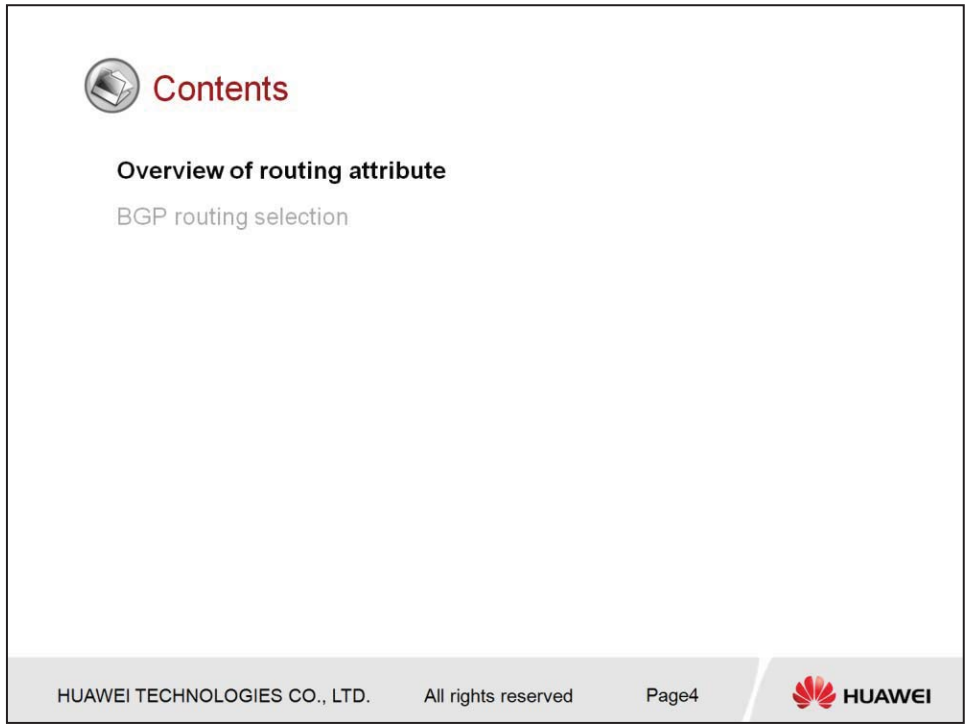
3. What is the requirement to inject the route into the BGP routing table by using network command.

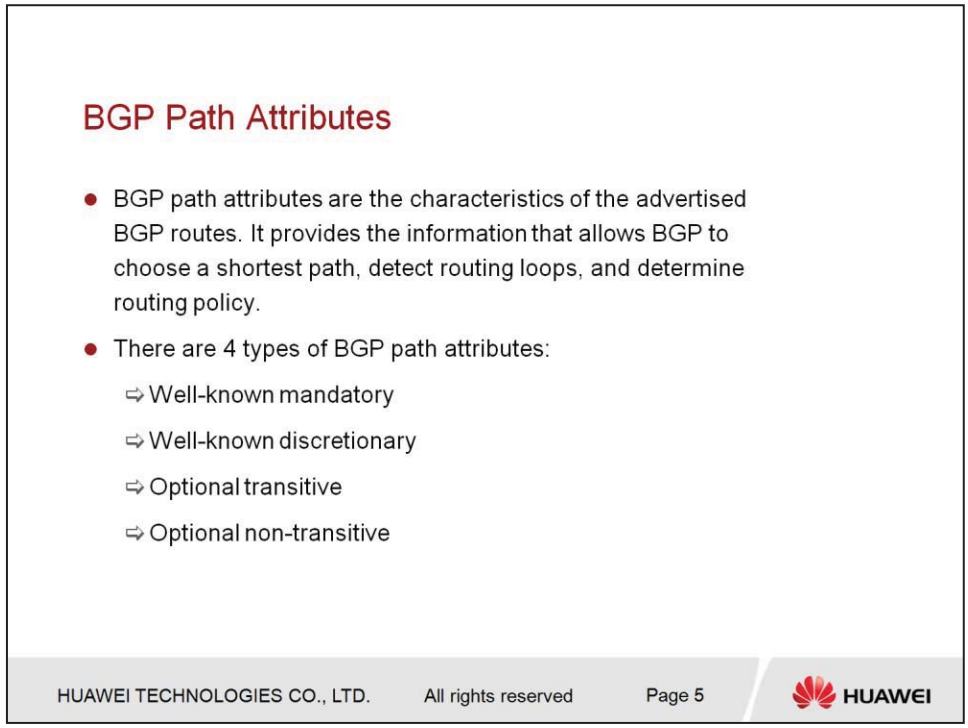
A：The routes to be injected into the BGP routing table via network command must be available inside the IP routing table. In addition, we must specify the length of the subnet mask precisely.


BGP Route Selection
www.huawei.com
HUAWEI TECHNOLOGIES CO., LTD.
All rights reserved
HUAWEI


Foreword
As an external gateway routing protocol, BGP is responsible for the exchange of NLRI among AS'. Combined with an abundance of path attributes, BGP is highly effective at controlling the exchange of this reachability information in order to realize route selection.
HUAWEI TECHNOLOGIES CO., LTD.
All rights reserved
Page2
HUAWEI


Objective
Upon completion of this section, you will be able to :
Understand the function of path attributes
Understand the behavior of BGP common path attributes
Understand the BGP routing selection process
HUAWEI TECHNOLOGIES CO., LTD. All rights reserved Page3
HUAWEI


Contents
Overview of routing attribute
BGP routing selection
HUAWEI TECHNOLOGIES CO., LTD.
All rights reserved
Page4
HUAWEI



The enterprises and service providers are often concerned about such questions, Example: How to filter some BGP routes？ How to control the BGP route selection? These questions can be solved by using abundant BGP path attributes.

BGP path attributes is a set of parameters and it further describes a specific route. We usevarious path attributes when configure the route policy.

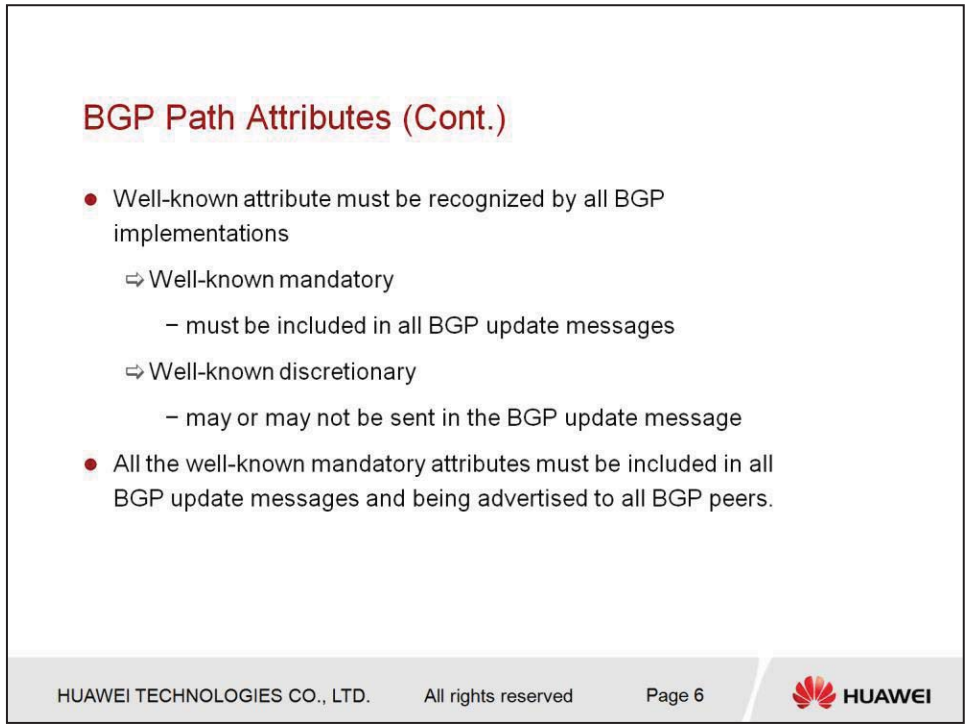
There are 4 types of BGP path attributes：

Well-known mandatory

Well-known discretionary

Optional transitive

Optional non-transitive



BGP implementations must recognize all the well-known attributes. The wellknown mandatory attributes must be included in every update messages while the well-known discretionary attributes may or may not be included in the update messages. Once the updates of BGP peer carry the well-known mandatory attribute, BGP peer must advertise it to other peers.

Well-known attributes must be recognized by all BGP implementation

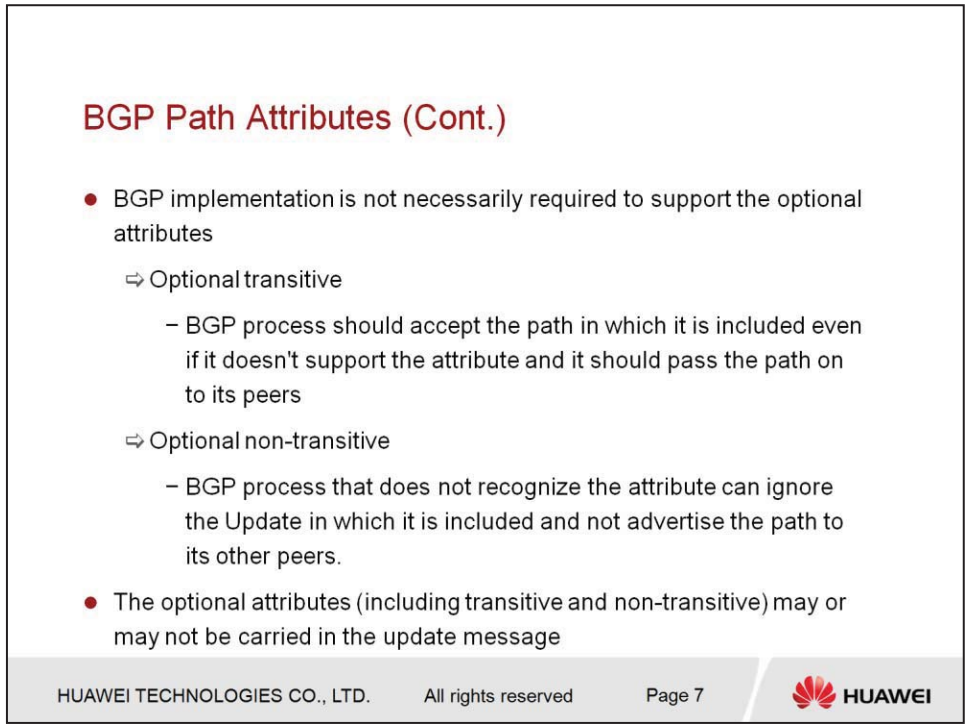
•Well-known mandatory

must be included in all BGP update messages,

example: NEXT_HOP attribute.

•Well-known discretionary

may or may mot be sent in the BGP update message, example: LOCAL_PREF attribute



In addition to the well-known attribute, each update messages can include one or more optional attributes. Not every BGP speakers are required to support these optional attributes.

BGP implementations is not necessarily required to support the optional attributes:

•Optional transitive

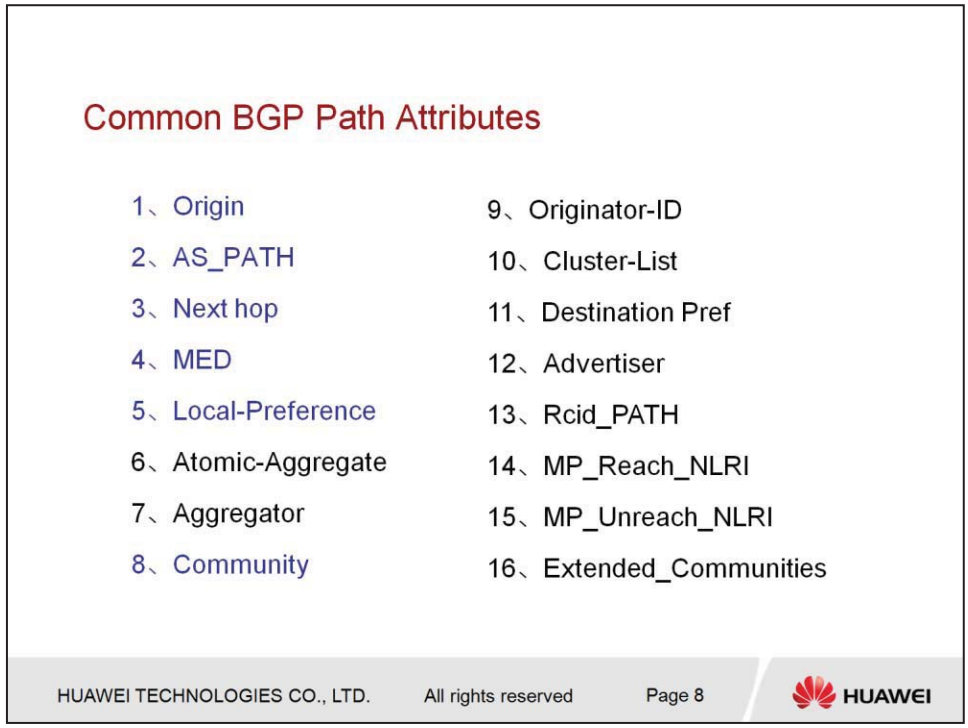
BGP process should accept the path in which it is included even if it doesn't support the attribute and it should pass the route on to its peers.

Example: AGGREGATOR attribute

•Optional non-transitive

BGP process that does not recognize attribute can ignore the Update in which it is included and not advertise the route to its other peers.

Example: MED attribute



Attributes	Well-known/Optional	Mandatory/Discretionary	Transitive/Non-transitive
Origin	Well-known	Mandatory	--
AS_PATH	Well-known	Mandator	--
Next-hop	Well-known	Mandatory	--
Local-preference	Well-known	Discretionary	--
MED	Optional	--	Non-transitive
Community	Optional	--	transitive

Origin：Origin is a well-known mandatory attribute that is used to define the origin of the routing information. It indicates how a route becomes BGP route, such as IGP, EGP and INCOMPLETE.

As_PATH: A well-known mandatory attribute that lists a sequence of the autonomous systems passed by a route. It is used to prevent the routing loop. Besides, it can be used for route filtering and selection.

Next hop: A well-known mandatory attribute that defines the IP address of the border router that should be used as the next hop to the destinations listed in the Network Layer Reachability field of the UPDATE message.

MED：MED is an optional non-transitive attribute. When some AS has multiple

entries, the MED attribute can be used to help other external AS in selecting a better entry path. The smaller the MED value of a route, the higher its precedence.

Local-Preference：Local preference is a well-known discretionary attribute. It is used only inside an autonomous system and not passed to other autonomous systems. If an internal BGP speaker receives multiple routes to the same destination, it compares the local preference attributes of the routes.

This reflects the preference level of the BGP speaker for each external route.

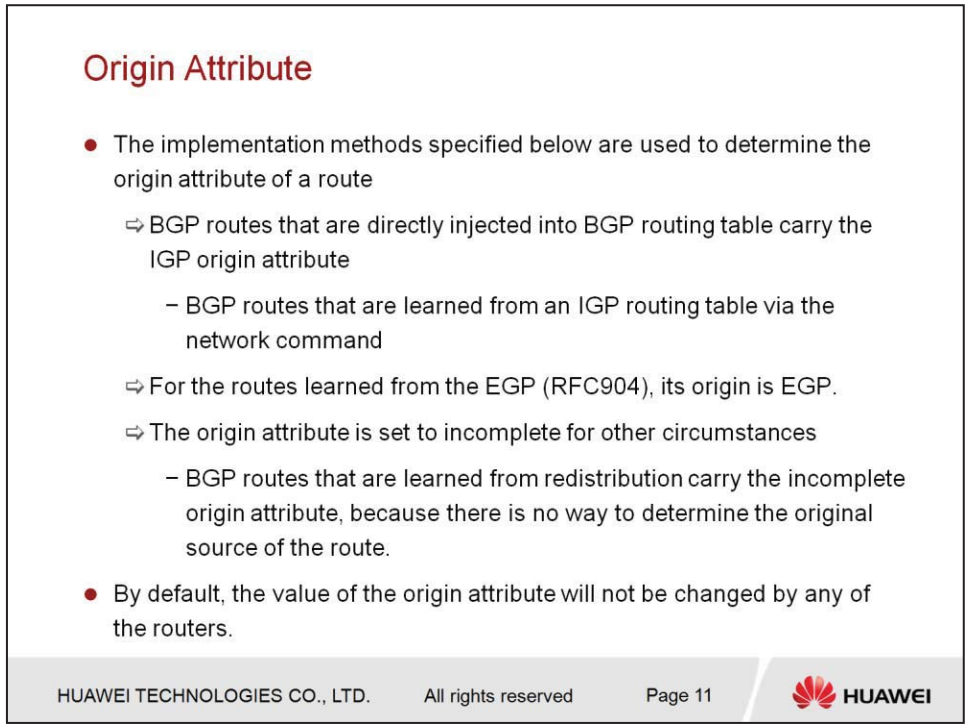
The larger the local preference value, the higher the preference level of the route.

Community：It is an optional transitive attribute that is designed to simplify policy enforcement. It marks a group of routing information that has the same feature which is irrelevant to its IP subnet or autonomous system.

Common BGP Path Attributes (Cont.)

BGP Attribute	Class
1. Origin	(well-known mandatory)
2. AS_Path	(well-known mandatory)
3. Next_Hop	(well-known mandatory)
4. Multi_Exit_Disc	(optional non-transitive)
5. Local_Pref	(well-known discretionary)
6. Atomic_Aggregate	(well-known discretionary)
7. Aggregator	(optional transitive)
8. Community	(optional transitive)
9. Originator ID	(optional non-transitive)
10. Cluster List	(optional non-transitive)

HUAWEI TECHNOLOGIES CO., LTD. All rights reserved



Origin attribute specify the origin of the BGP path information. In fact, it is the methods for BGP speaker to generate the BGP route. BGP considers three types of origins:

• IGP: The route with origin IGP is marked with “i” in BGP routing table (by using the “display bgp routing-table” command).The origins are IGP for the routes internal to the AS and are advertised via the network command. This method is also called as semi dynamic redistribution of BGP information. The

network advertised via the network command is dynamically discovered and calculated by IGP (including static route). Some of the routing information is selectively redistributed into the BGP system via network command. That’s why it is called as “semi dynamic”.

• EGP: The route with origin EGP is marked with “E” in BGP routing table. The origin “EGP” was used when the Internet when the routes are redistributed from EGP into the BGP routing table. It is used when the Internet was migrating from EGP to BGP. It is rather difficult to encounter the route with origin EGP in the real network. This is because EGP protocol is basically obsolete and not used anymore.

•Incomplete: The route with origin Incomplete is marked with “?”

in BGP routing table. The route with origin incomplete is learned by some other means. It means that the information for determining the origin of the route is incomplete. Routes that BGP learnt through redistribution from IGP or static route carry the incomplete origin attribute. Injecting the IGP routes into BGP dynamically or semi dynamically is based on the dependency of the BGP routes on the IGP routes.

Unconditionally injecting the IGP routes into BGP might bring some drawbacks.

The unnecessary or wrong information might leak into the BGP routing table. For example, the IGP might consist of some special address for internal AS only or some unregistered address. Apart from that, it might also cause the fluctuation of BGP (because the BGP route is base on the IGP route). BGP use a process called route dampening to penalize and ultimately discontinue advertisement of fluctuating routes, depending on their degree of instability. We will not discuss this in detail here.

In this circumstance, we must adopt some filtering strategies, and ascertain those networks can be redistributed from IGP into BGP. For the protocol that can be used to distinguish between the internal route and external route (for example OSPF), we can ensure that only the internal route is redistributed into the BGP

by using the appropriate configuration. By default, Huawei equipments will inject only the OSPF routes into the BGP. The OSPF external routes will not be injected. Besides, the static route can be redistributed into BGP and this method

can increase the stability of the route. The static routes will never disappear from the IP routing table and hence will always be advertised.

The precedence order of the 3 origin values are IGP>EGP>INCOMPLETE.

These 3 origin values are used to control the selection of BGP routes.

Origin Attribute (Cont.)

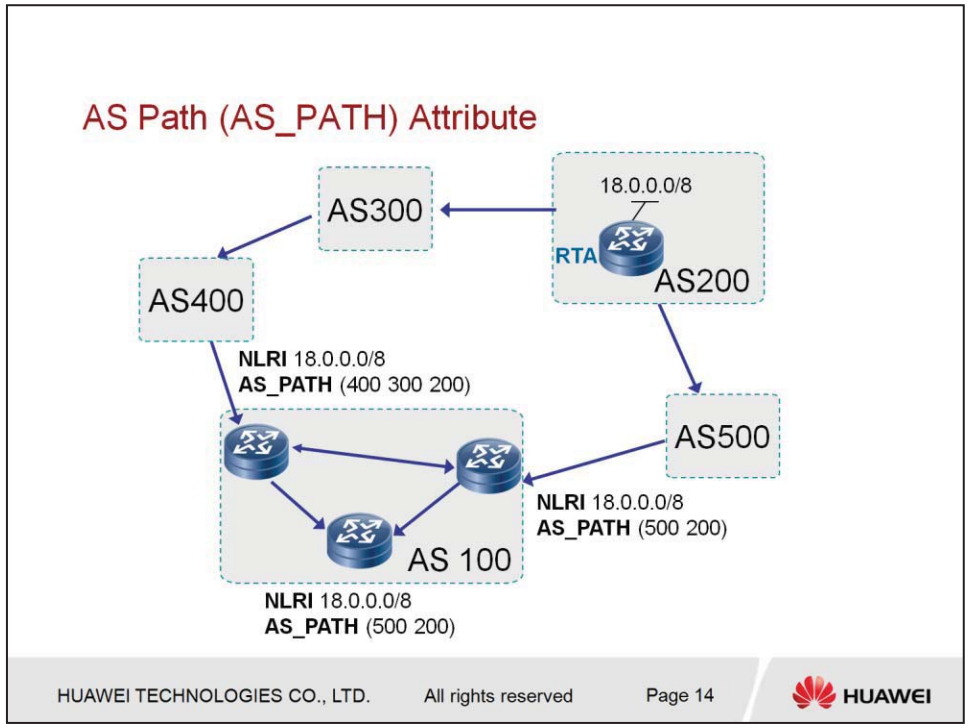
```
[RTB]display bgp routing-table

 Total Number of Routes: 2

 BGP Local router ID is 192.168.2.1
 Status codes: * - valid, > - best, d - damped,
               h - history,  i - internal, s - suppressed, S - Stale
               Origin : i - IGP, e - EGP, ? - incomplete
      Network            NextHop        MED        LocPrf    PrefVal Path/Ogn

 *>   192.168.1.0        10.1.1.1        0                      0      100i
 *    192.168.2.0        10.1.1.1        0                      0      100i
```

HUAWEI TECHNOLOGIES CO., LTD. All rights reserved



AS_PATH is an important well known mandatory attribute in BGP. It contains a sequence of autonomous system numbers that represent the path a route has traversed. To prevent routing loop, BGP will not accept the routing update that contains its local AS number in the AS_PATH attribute. Therefore, the egress router will append its AS number to the AS_PATH attribute each time the route is advertised by an egress router to the EBGP peer in another AS. This is to record the path that has been traversed by a particular route. If the local AS number is found in AS_PATH attribute of the routing update information, it indicates that the update has already passed this AS or the update is originated in this AS. This update information need to be discarded to prevent routing loop.

In addition, AS_PATH attribute is one of the important parameter used to determine the best route to reach a destination. When 2 or more routes toward the same destination exist on a router, the AS_PATH attribute can be used for route selection. A shorter AS_PATH is always preferred over a longer one. In case of a tie, other attributes are used to determine the best path to the destination. Note: In the most of the actual network implementations, AS_PATH is used to determine the best route

when multiple paths towards the same destination exist.

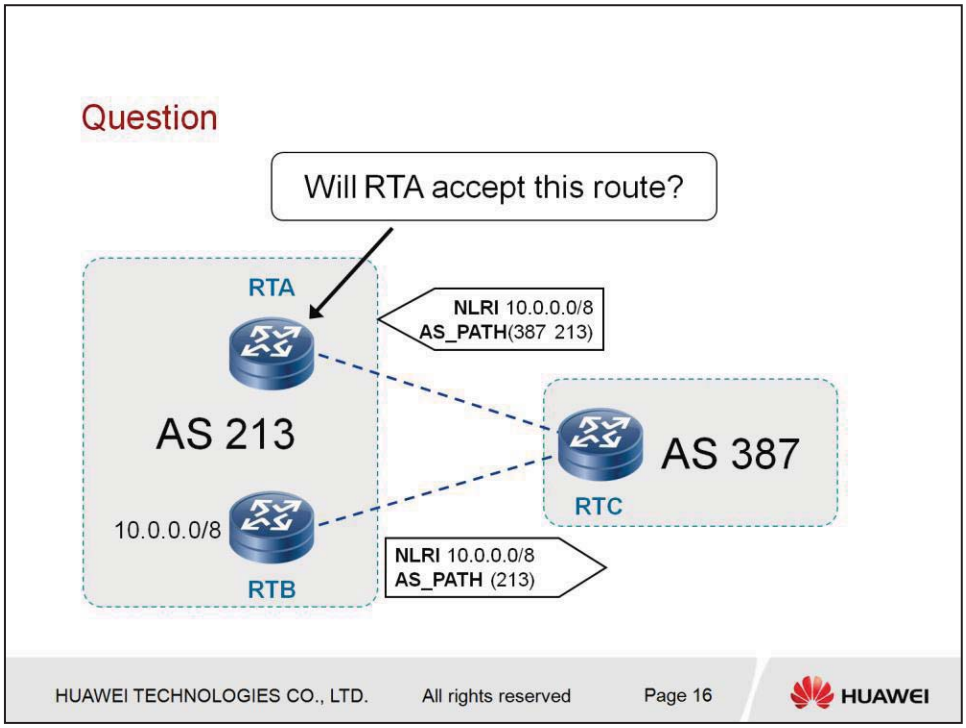
As shown in the diagram above, the BGP route for network 18.0.0.0/8 in AS 200 passes through AS200, AS300, and AS400 in order to reach AS100. Its AS_PATH is recorded as (400, 300, 200). Another path exist by passing AS200 and AS500 in order to reach AS100. Its AS_PATH is recorded as (500, 200). In this scenario, BGP will select the route with shortest AS_PATH. Therefore, route with AS_PATH “(500, 200)” will be selected.

After perform the BGP route aggregation, granularity that exists in the specific routes that form the aggregate is lost by default.

Assume that an AS is advertising an aggregate address representing addresses in several autonomous systems. In this case, the AS that originates the aggregate route include only its own number in the AS_PATH. The AS_PATH

information of the specific route is not included. This causes the lost of the path information to some of the more-specific prefixes. Therefore, the aggregated route might be sent back to the AS where the specific route reside and this generate the routing loop. We will discuss this problem in detail when learn the topic of route aggregation.

For most of the circumstances, route filtering base on AS _PATH list provides more flexible control over the route filtering base on the prefix list.



By default, BGP detects the routing loop via AS number. As shown in the diagram above, the EBGP relationship is established between RTA-RTC and RTB-RTC.

RTB will append its local AS number (213) when sends the update about network 10.0.0.0/8 to RTC. When RTA receives the same update from RTC, the AS_PATH attribute of the route contains its local AS number (213). Therefore, RTA will ignore this update information.

In some special implementation like the case of hub and spoke, we need accept the BGP route with the repeated AS number. In this case, we can use the command below to force the router accept the BGP route with repeated AS number.

peer { group-name | ipv4-address } allow-as-loop [number]

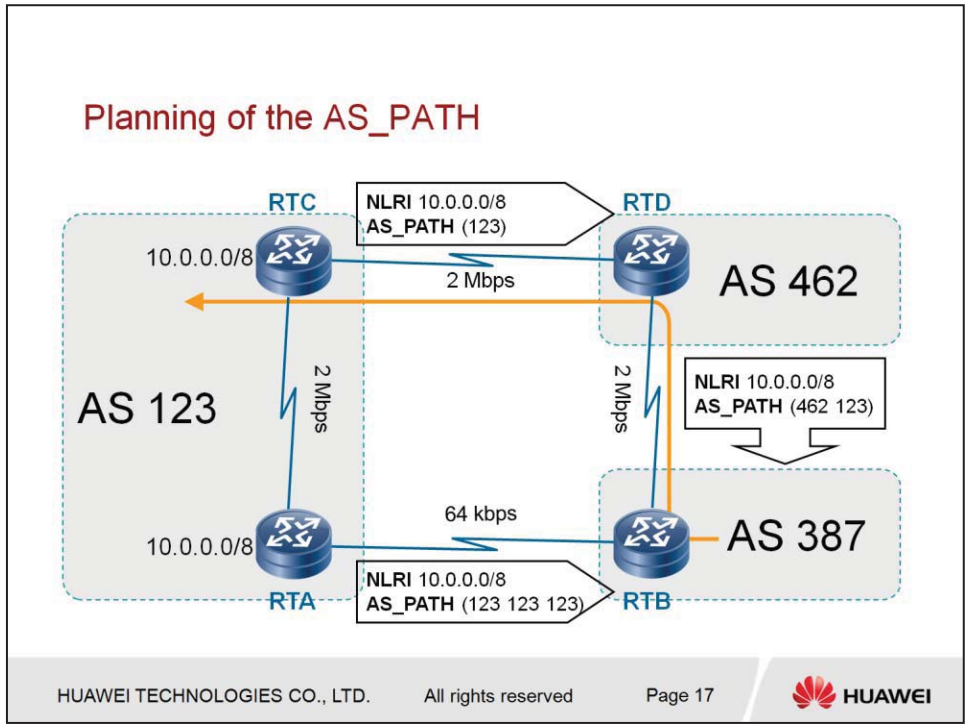
parameters：

group-name：specify the name of the peer group.

ipv4-address：specify the IPv4 address of the peer

number：specify the repeating times of the AS-number that are allowed.

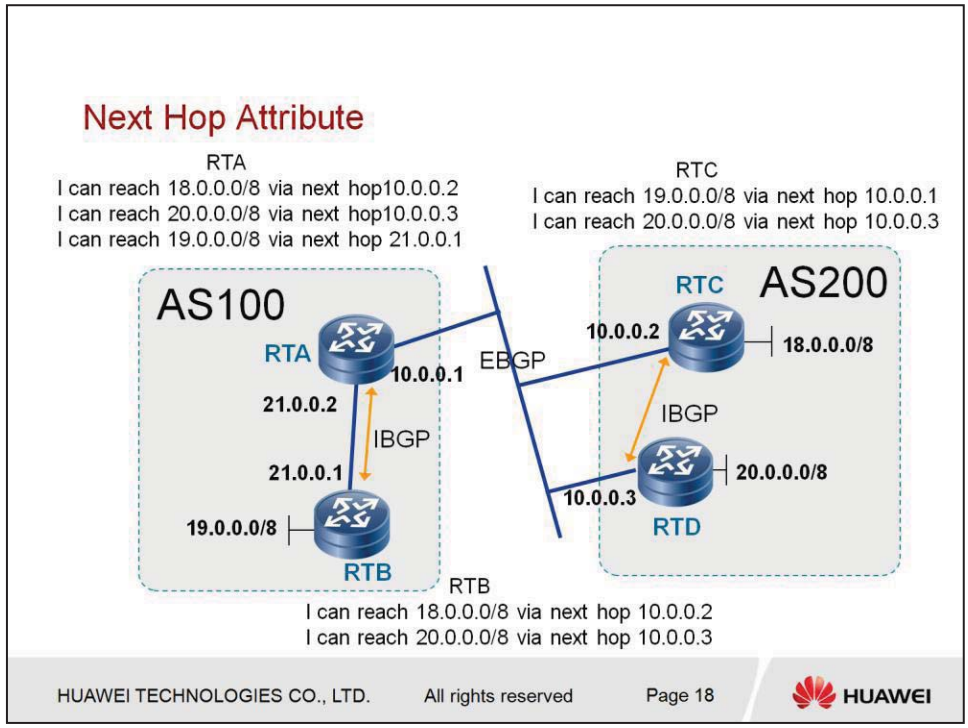
It is ranging from 1 to 1. The default value is 1.



The length of the AS_PATH can be extended to influence the route selection.

For this example, we can configure the RTA in AS123 that sends the EBGP updates of network 10.0.0.0/8 to AS 387 by appending the AS_PATH with 2 copies of the AS number 123. After the route has been advertised to RTB, its AS_PATH will be (123, 123,123). The second alternative routes to reach network

10.0.0.08 that RTB received is from AS462 which originates from AS123. Its AS_PATh is (462,123). RTB will perform the route selection process and determine the route to be used in order to reach network 10.0.0.0/8. In this case, RTB will select the best route base on the length of the AS_PATH. RTB will prefer the route with shorter AS_PATH: (462 123). The data is therefore forwarded to AS123 via AS462. As a result, the high speed link between RTC and RTD is selected while the low speed link between RTA and RTB is used for backup.



Next hop is well-known mandatory attribute that defines the IP address of the border router that should be used as the next hop to the destinations listed in the Network Layer Reachability field of the UPDATE message. The next hop processing of BGP is very complex. The following 3 rules apply:

(Note: RTA and RTC establish the EBGP neighbor relationship through the directly connected Ethernet interface. RTA and RTB establish the IBGP neighbor relationship through the directly connected interface. RTC and RTD establish the IBGP neighbor relationship through the directly connected Ethernet interface between 10.0.0.2 and 10.0.0.3).

1. When BGP advertise update message to the EBGP peer, the NEXT_HOP is the IP address of the advertising router's interface. When BGP advertise update message to the IBGP peer, and the NLRI of the update refers to a destination within the same AS, the NEXT_HOP is the IP address of the neighbor that advertised the route. AS shown in the diagram above, the next hop is 10.0.0.2 when RTC advertise the route 18.0.0.0/8 to its EBGP RTA. The next hop is 21.0.0.1 when RTB advertise the route 19.0.0.0/8 to its IBGP RTA.

2. The next hop processing for multi-access network (broadcast

network or NBMA network) is different. As shown in the diagram above: RTD advertise the route 20.0.0.0/8 to RTC with next hop IP as 10.0.0.3. If RTC advertises the route 20.0.0.0/8 to RTA with next hop 10.0.0.2, the packets from RTA to network 20.0.0.0/8 will have to crossed the shared LAN twice. To optimise the packet forwarding in the multi-access network, RTC will use 10.0.0.3 as next hop instead of 10.0.0.2 when advertising the route 20.0.0.0/8 to EBGP RTA.

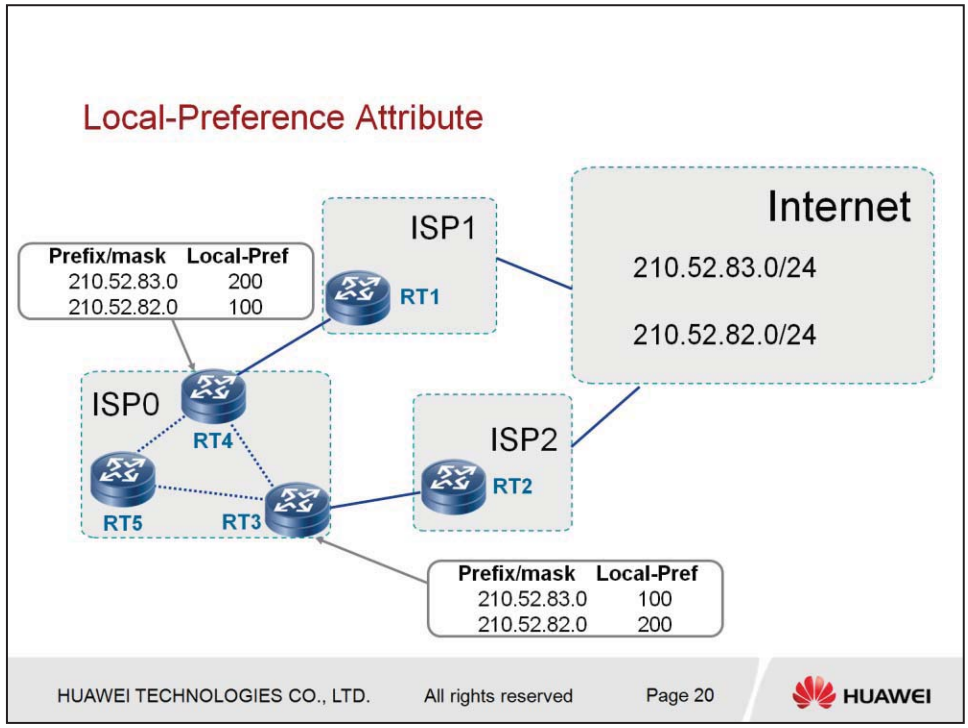
3. When BGP advertise update message to IBGP peer and the NLRI of the update refers to a destination in a different AS, the NEXT_HOP is the IP address of the external peer from which the route was learned. As shown in the diagram above, the next hop remain 10.0.0.2 when RTA advertise the route 18.0.0.0/8

learnt from EBGP peer to its IBGP peer RTB. This default behavior of BGP might bring some potential problems: If the next-hop address towards 10.0.0.2 for RTB is unreachable, the packets for the destination 18.0.0.0/8 will be dropped.

Solution：

First Method：Issue the command “import route direct” in the BGP view of RTA.

Second Method ：Issue the command peer { group-name | ipv4-address } next-hop-local on RTA. By using this command, the next hop IP address is set to the local address when advertising the route to the peer or peer group.



Local preference is a well-known discretionary attribute. In certain circumstance, one single ISP can access to the Internet by connecting to 2 larger ISPs using the high speed link. As shown in the diagram, ISP0 is connecting to ISP1 and ISP2 respectively via 2 separate links.

In this case, how ISP0 distributes the traffic evenly on 2 of the uplinks? Assume that 2 routes that exist in the Internet are 210.52.83.0/24 (represented as 83 for the explanation below) and 210.52.82.0/24 (represented as 82 for the discussion).

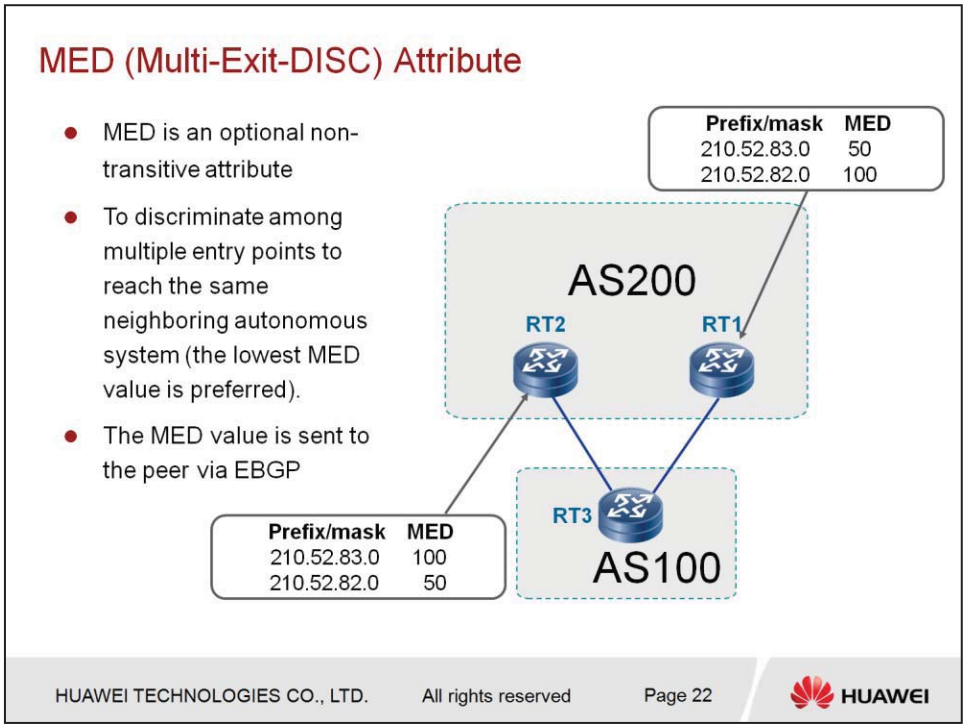
Our objective is to direct the traffic toward network 83 to the link of ISP1 and direct the traffic toward network 82 to the link of ISP2.

The internal network structure of ISP0 is as follows: Full mesh IBGP peer relationship is established for RT3, RT4 and RT5. RT3 establishes the EBGP peer relationship with ISP2 while RT4 establishes the EBGP relationship with ISP1. As a result, both RT3 and RT4 will receive the route 82 and 83 from their respective EBGP peer. Both RT3 and RT4 will in turn advertise the route 82 and 83 to their respective IBGP peer. Finally, RT5 learns the route 82 and 83 from two different sources. Therefore, we need to modify the attribute of the source on RT3 and RT4

in order to implement load balancing.

How to implement the load balancing? In this case, BGP can add the local preference attribute to the routes. If an internal BGP speaker receives multiple routes to the same destination, it compares the local preference attributes of the routes. The route with the highest local preference is selected.

When RT3 receives the routes 82 and 83 from ISP2, the local preference of route 83 is set to 100 (default local preference is 100, not require to configure it) while the local preference of route 82 is set to 200. Similarly, when RT4 receives the routes 82 and 83 from ISP1, the local preference of route 82 is set to 100 while the local preference of route 83 is set to 200. As a result, RT5 will receive 2 routes from 2 different sources. These 2 routes carry different local preference value and toward the same destination. RT5 will carry out the route selection based on the value of the local preference. Finally, the traffic from route 83 is sent to ISP1 while the traffic from 82 is sent to ISP2.



We have introduced how the local preference is used to control the traffic leaving the AS. In certain circumstance, we need to control the incoming traffic into an AS. MED attribute can be used to achieve this purpose. This can be shown from the example above. In this network, AS100 has 2 uplink connections to 2 different routers in AS200. Assume that 2 routes that exist in the Internet are 210.52.83.0/24 (represented as 83 for the explanation below) and 210.52.82.0/24 (represented as 82 for the discussion). These 2 routes are advertised to the border router RT3 in AS100 through BGP.

The network administrator of AS200 would like to achieve the objectives as follow:

The data traffic from AS100 to reach 82 must pass through RT2 while the data traffic from AS100 to reach 83 must pass through RT1. From here, we can observe the difference between local preference and MED. Local preference is used to control the outgoing data traffic from an AS while MED is used to control the incoming data traffic into an AS.

The objective above can be achieved by providing a MED value for the advertised route. When the peer at the other end receives multiple routes toward a destination, it will make the selection

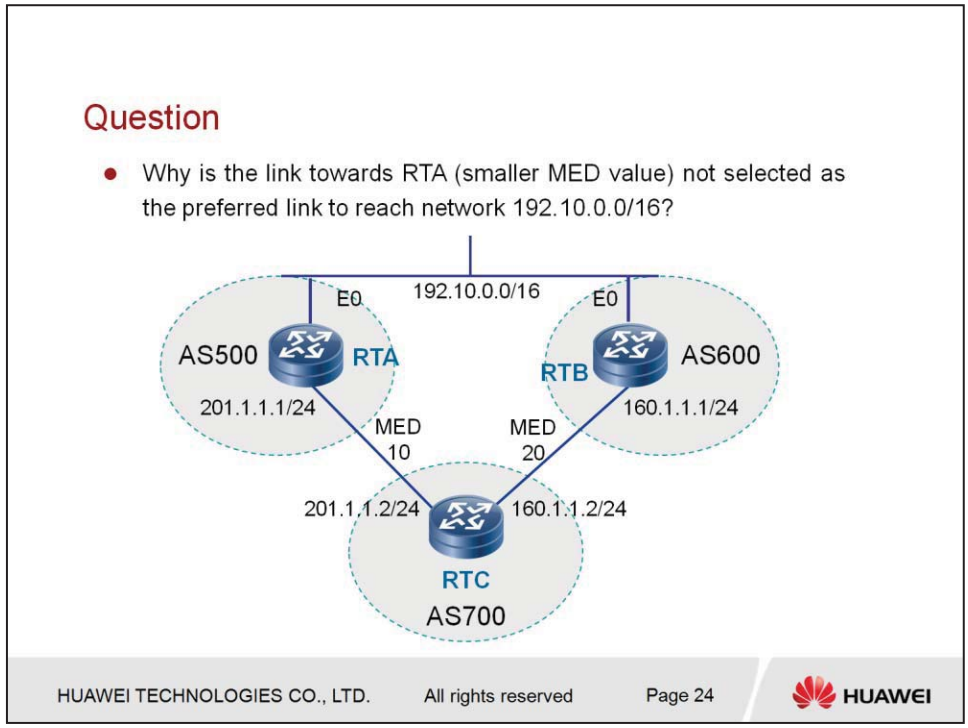
base on the MED value.

1. When the border router RT1 in AS200 advertises the routes 82 and 83 to RT3 in AS 100, it will mark the route 83 with MED value 50 and mark the route 82 with MED value 100.

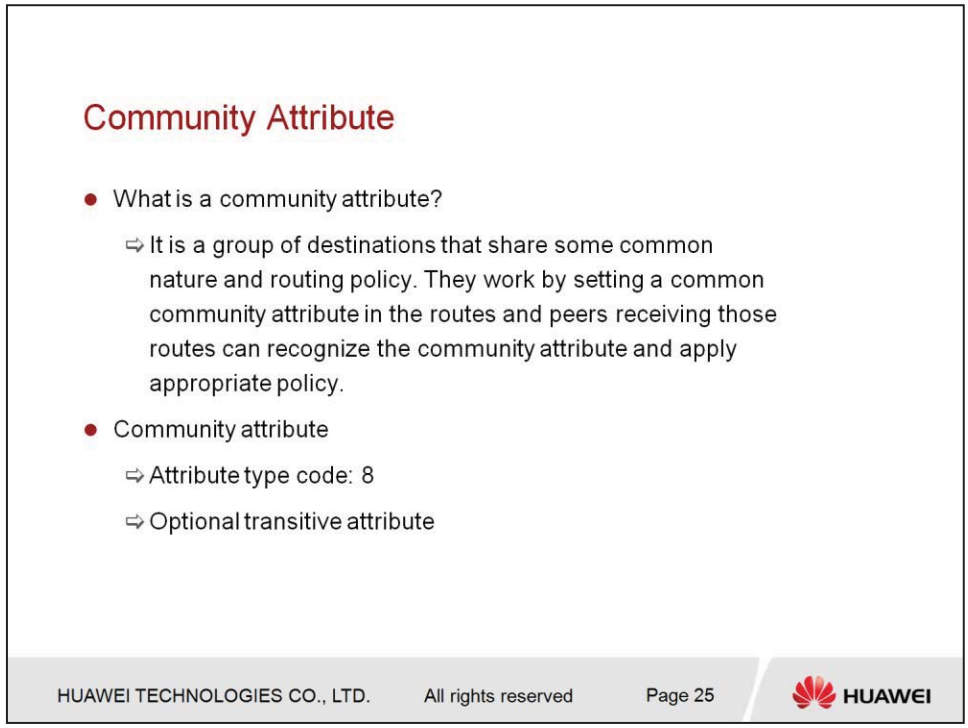
2. When the border router RT2 in AS200 advertise the routes 82 and 83 to RT3 in AS100, it will mark the route 82 with MED value 50 and mark the route 83 with MED value 100.

3. When RT3 in AS 100 has learnt the same routes from EBGP peer RT2 and RT1, it will select RT1 as the next hop to reach 83 and RT2 as the next hop to reach 82.

We prefer the lowest MED value. This is because MED value is considered as a metric. Therefore, the metric with lowest value or lowest distance is preferred.



By default, the BGP router only compare the MED values of the routes from the same AS (different peers). However, we can use the command comparedifferent- as-med to tell BGP to compare MEDs from different ASs for the same route. This command is used only when different ASs adopt same routing policies and routing protocol.

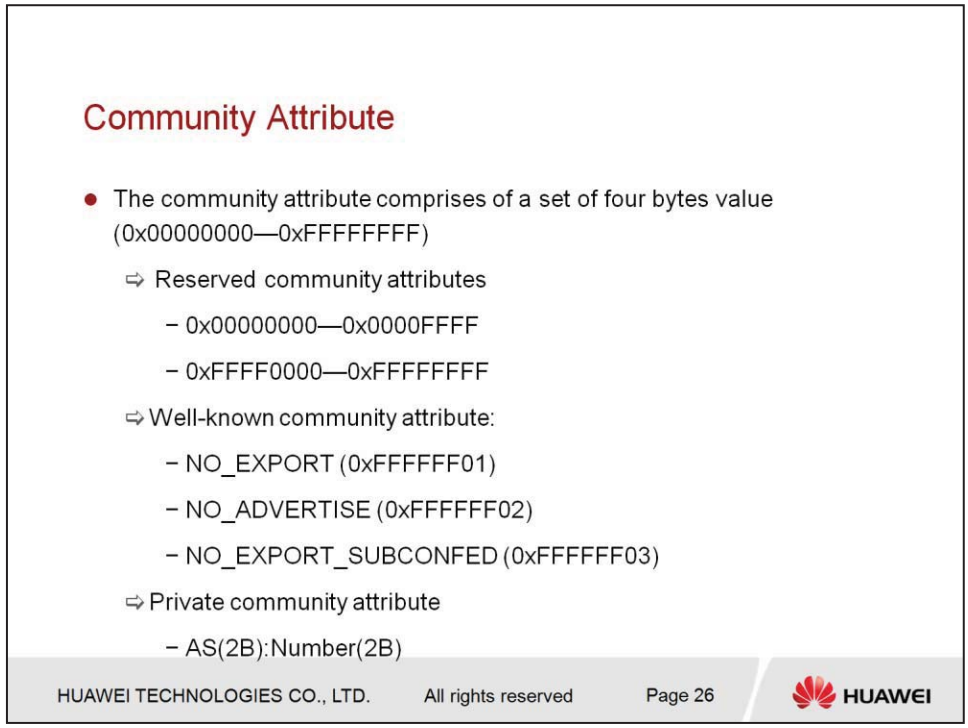


In BGP, a community is a group of destination that share the common nature.

RFC1997 defined the community attribute as variable length optional transitive attribute.

Each autonomous system administrator may define which communities a destination belongs to. By default, all destinations belong to the general Internet community. Received routes belonging to Internet community are advertised freely.

A single route can have more than one community attribute values. The BGP router that sees multiple community attribute values in one route can apply the appropriate policy according to one or more or all of those attribute values. The router can add or modify the community attribute values before it transmits the route to other peers.



The community attribute consists of a set of four octet values, each of which specify a community. All routes with this attribute belong to the communities listed in the attribute.

The community attribute values ranging from 0x0000000 through 0x0000FFFF and 0xFFFF0000 through 0xFFFFFFFF are reserved.

Well-known community attributes are recognized and have global significance.

The following are well-known community attributes:

NO_EXPORT(0xFFFFFFF01)：Routes received carrying this value cannot be advertised to EBGP peers or, if a confederation is configured, the routescannot be advertised outside of the confederation.

NO_ADVERTISE(0xFFFFFFF02)：All routes received carrying a communities attribute containing this value MUST NOT be advertised to o t h er BGP peers.

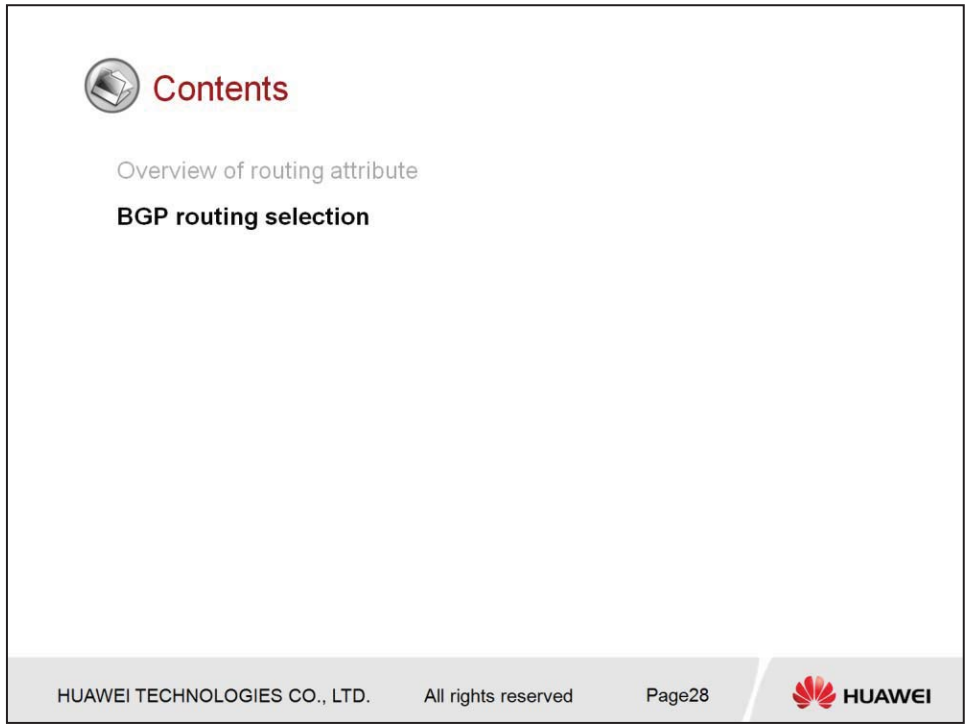
NO_EXPORT_SUBCONFED(0xFFFFFFF03)： All routes received carrying a communities attribute containing this value MUST NOT be advertised to external BGP pe e rs (this includes peers in other members autonomous systems inside a BGP

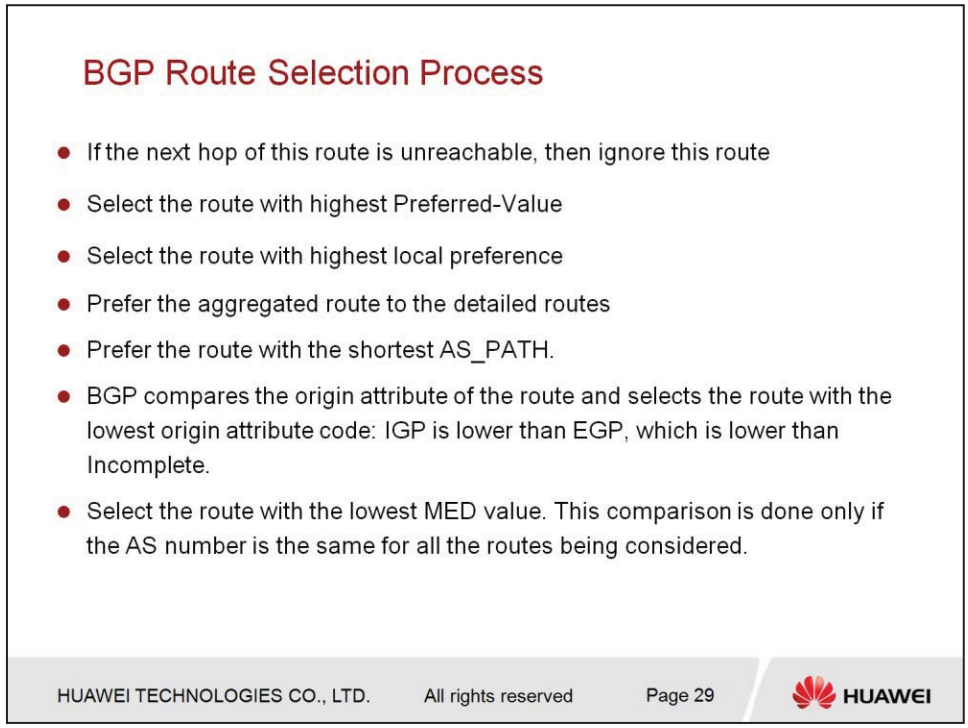
confederation). It is also called as LOCAL_AS attribute.

In addition to the well-known community attribute, private community attributes can be defined for special uses. These attributes are identified by some numerical value. A common practice is to use the first 2 bytes of the community

attribute for the AS number and the last 2 bytes for an arbitrary value ranging from 0 to 65535. (e.g. AS 690 may define research, educational and commercial community values that may be used for policy routing as defined by the operators of that AS using community attribute values 0x02B20000 through

0x02B2FFFF (690:0~65535))


Contents
Overview of routing attribute
BGP routing selection
HUAWEI TECHNOLOGIES CO., LTD.
All rights reserved
Page28
HUAWEI



1. If the next hop of this route is unreachable, then ignore this route.

2. Select the route with highest Preferred-Value. This new parameter is added in VRP5. It can be used to assign a preference value for a peer.

The higher the number, the more preferable the route.

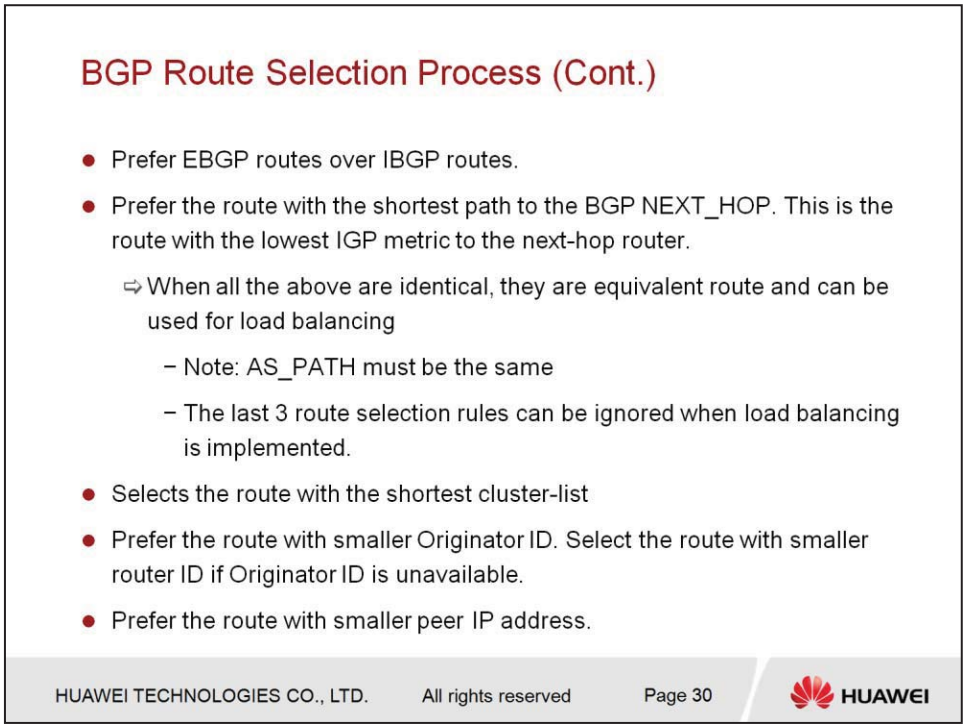
3. Select the route with a higher local preference

4. Prefer the aggregated route to the detailed routes

5. Prefer the route with the shortest AS_PATH.

6. BGP compares the origin attribute of the route and selects the route with the lowest origin attribute code: IGP is lower than EGP, which is lower than Incomplete.

7. Select the route with the lowest MED value. This comparison is done only if the AS number is the same for all the routes being considered.



8. Prefer EBGP routes over IBGP routes.

9. Prefer the route with the shortest path to the BGP NEXT_HOP. This is the route with the lowest IGP metric to the next-hop router.

When all the above are identical, they are equivalent route and c an be used for load balancing.

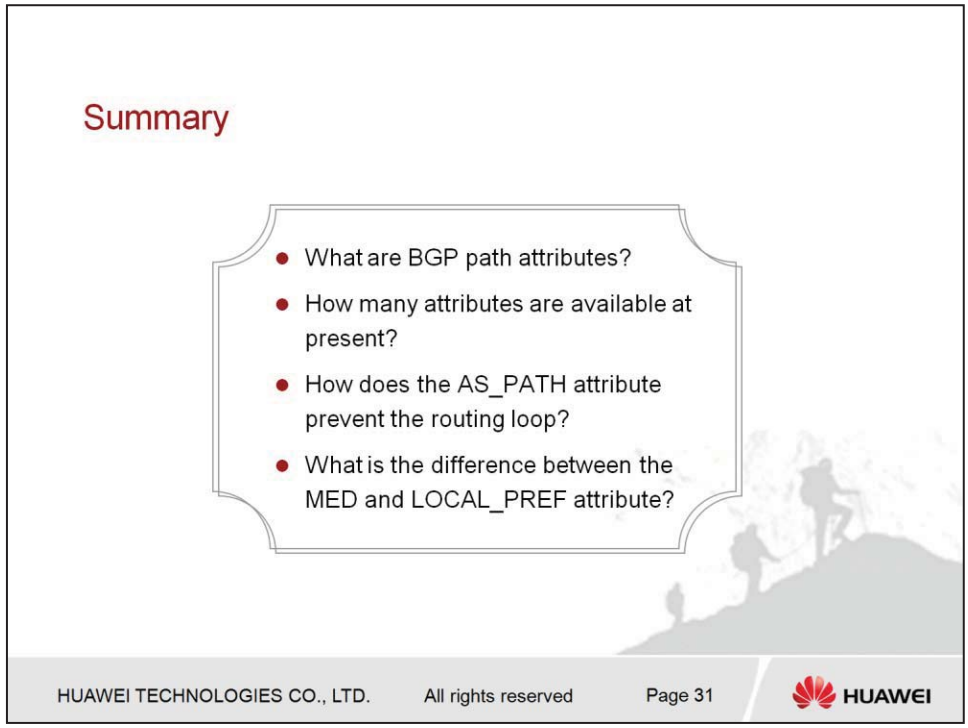
Note: AS_PATH must be the same

The last 3 route selection rules can be ignored when load balancing is implemented.

10. Selects the route with the shortest cluster-list.

11. Prefer the route with smaller Originator ID. Select the route with smaller router ID if Originator ID is unavailable.

12. Prefer the route with smaller peer IP address.



1. What are BGP path attributes? How many attributes are available at present?

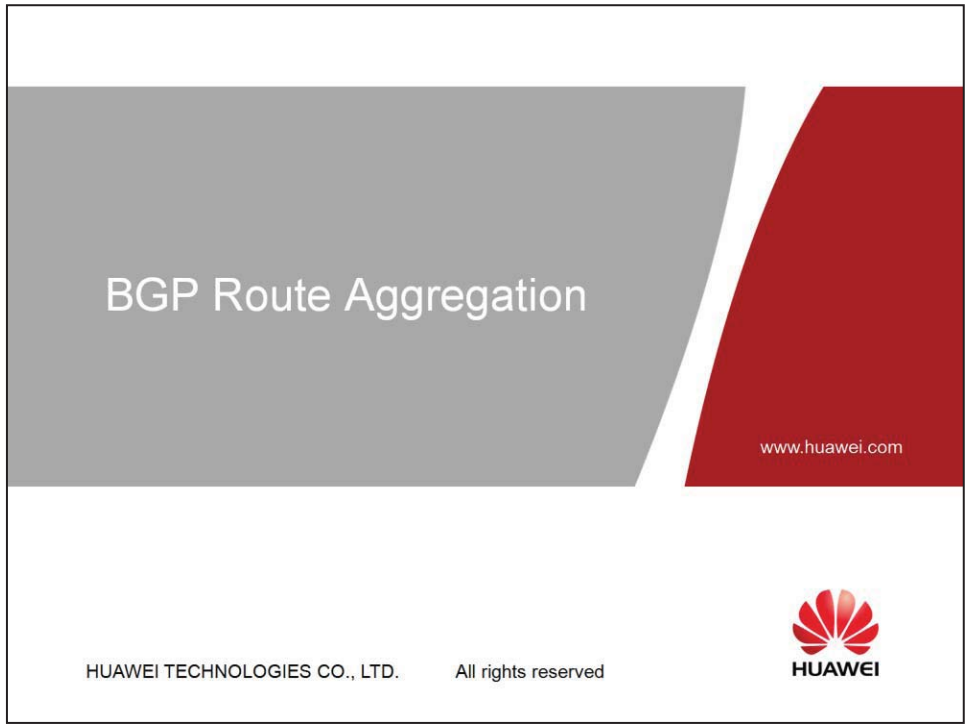
A: BGP is a policy tool for route selection. Different from IGP, the main function of BGP is to transmit the routing information instead of route discovering. BGP path attributes describe the characteristic of a BGP route. It can be used to control the route advertisement. At present, BGP4 has 16 types of path attributes.

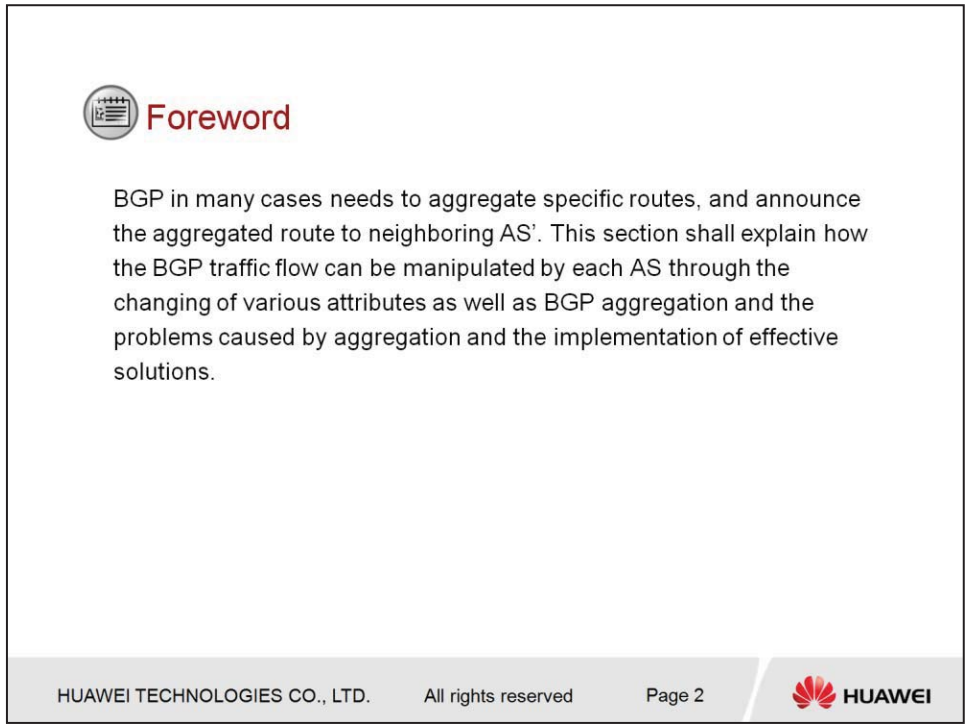
2. How the AS_PATH attribute prevent the routing loop?

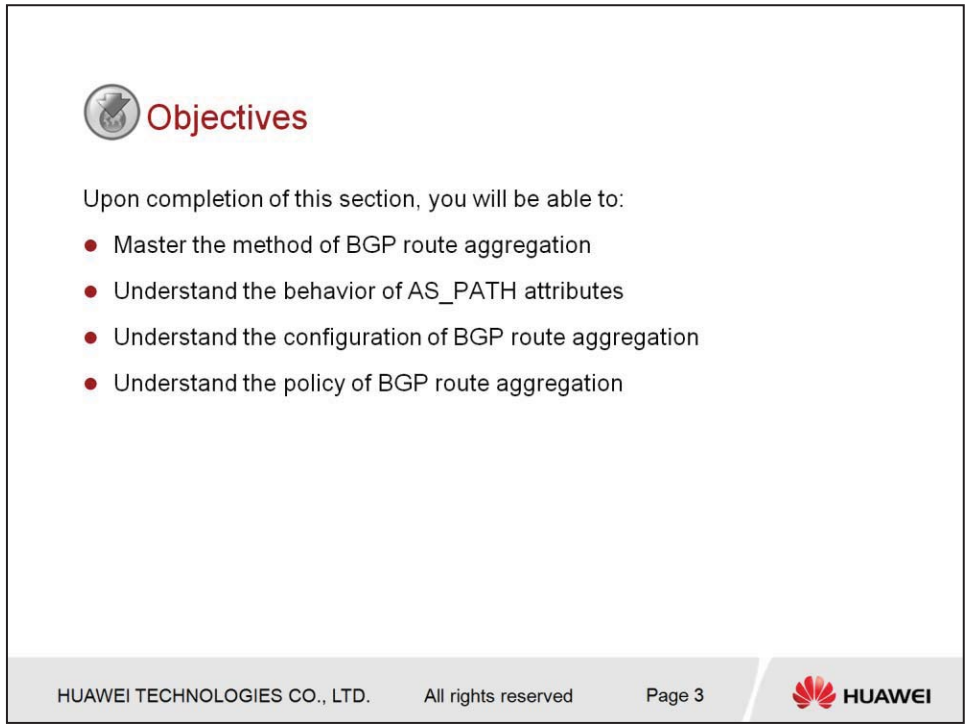
A： AS_PATH attribute list is a sequence of autonomous systems that must be passed through in order to reach the indicated network. These sequence of autonomous systems are appended from the right to the left. Therefore, the AS that originally injected the route into BGP is always found in the rightmost end of the AS_PATH. Each time the routing update crossed the boundary of an AS, the AS number of the sender is appended to the front part of the AS_PATH. A router that have received the BGP update will check the AS_PATH attribute and look for its own AS number. If it is found in the AS_PATH, then the route has already crossed the AS. This route information will be discarded to prevent the routing loop in this situation.

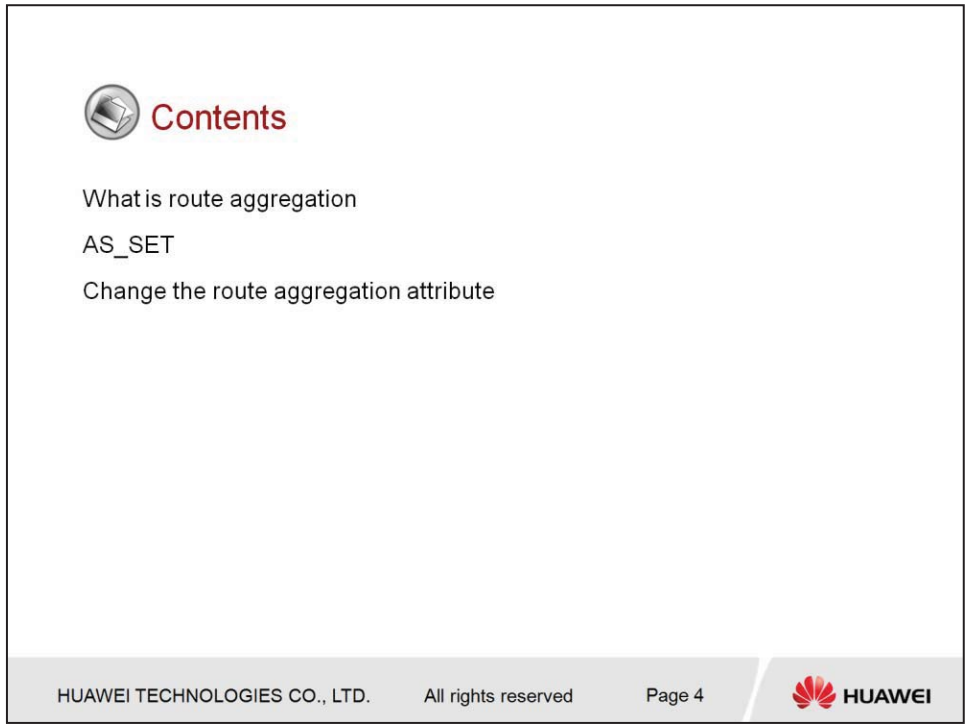
3. What is the difference between the MED and LOCAL_PREF attribute?

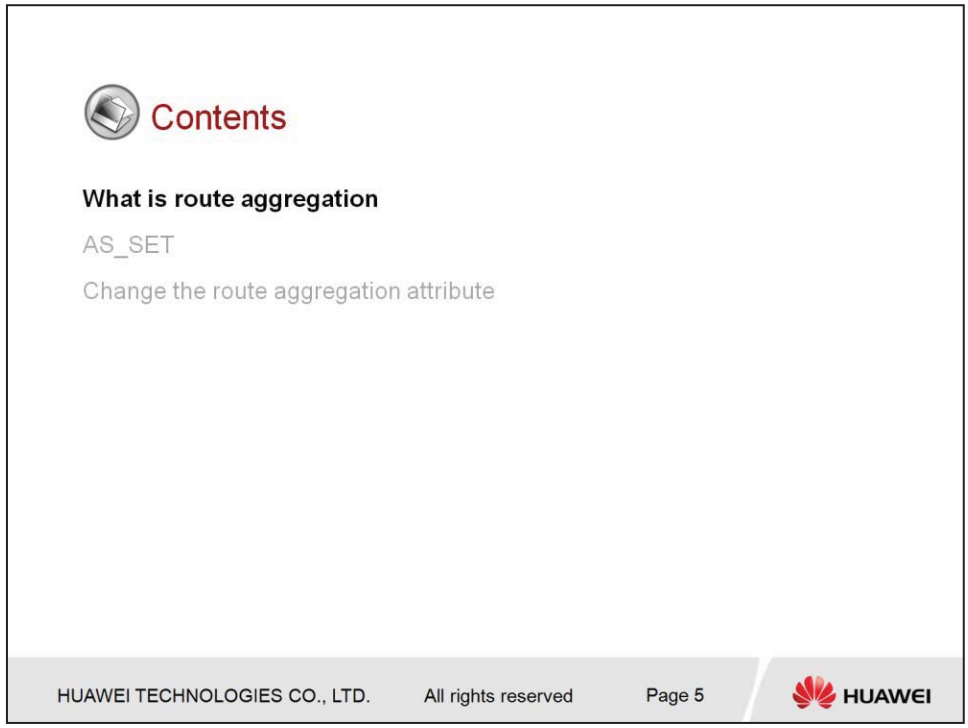
A：MED is used by the EBGP peer while LOCAL_PREF is used by the IBGP peer. In other words, EBGP speaker can used the MED to inform the neighboring AS of the preferred link for incoming traffic when multiple links exist between 2 autonomous systems. The LOCAL_PREF attribute can be used to identify the preferred route when multiple IBGP speaker are advertising the same route within an autonomous system.

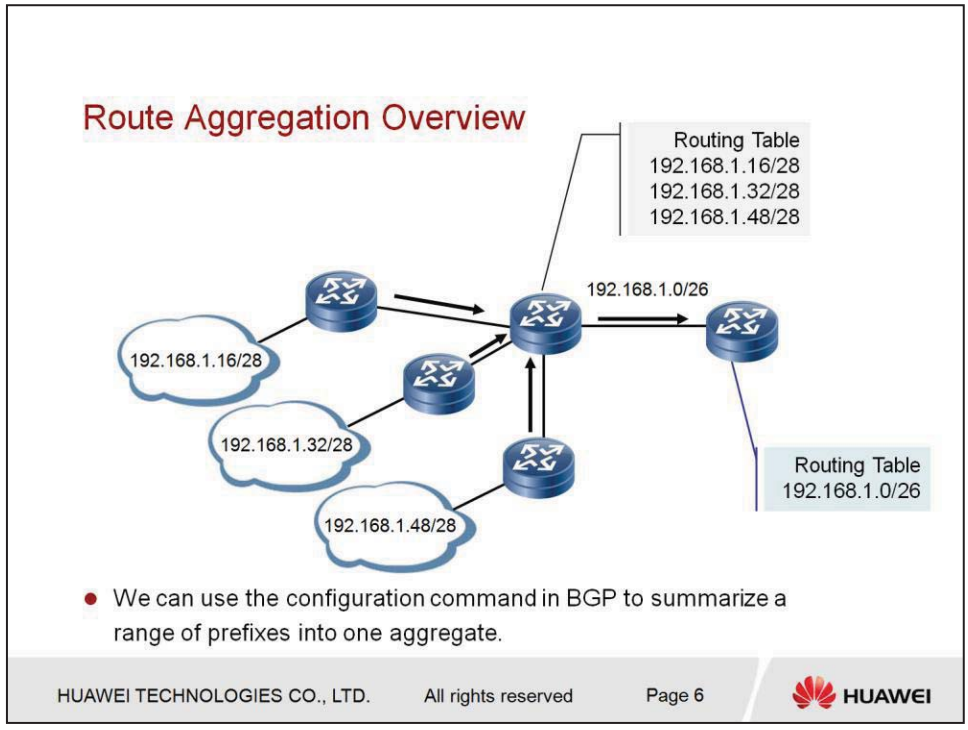

BGP Route Aggregation
www.huawei.com
HUAWEI TECHNOLOGIES CO., LTD.
All rights reserved
HUAWEI

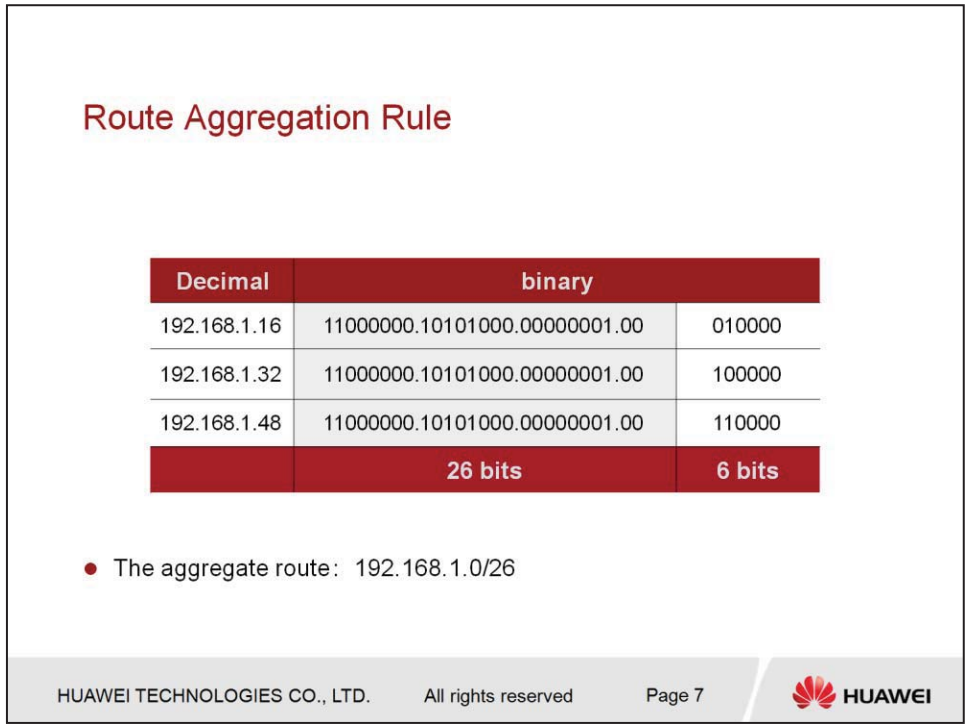

Foreword
BGP in many cases needs to aggregate specific routes, and announce the aggregated route to neighboring AS'. This section shall explain how the BGP traffic flow can be manipulated by each AS through the changing of various attributes as well as BGP aggregation and the problems caused by aggregation and the implementation of effective solutions.
HUAWEI TECHNOLOGIES CO., LTD.
All rights reserved
Page 2
HUAWEI


Objectives
Upon completion of this section, you will be able to:
Master the method of BGP route aggregation
Understand the behavior of AS_PATH attributes
Understand the configuration of BGP route aggregation
Understand the policy of BGP route aggregation
HUAWEI TECHNOLOGIES CO., LTD.
All rights reserved
Page 3
HUAWEI

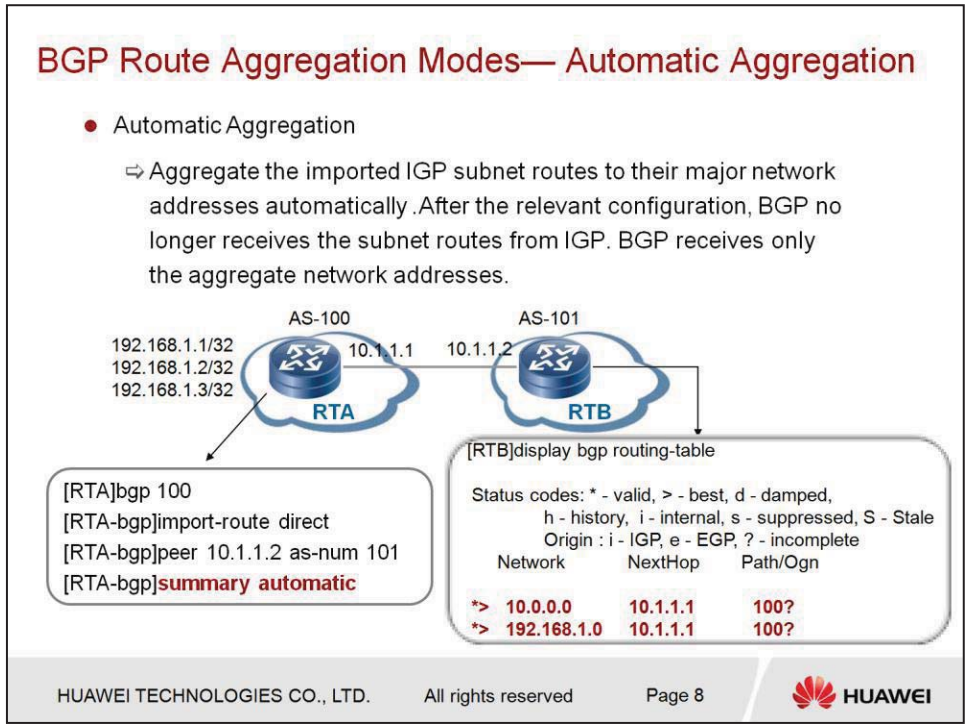

Contents
What is route aggregation
AS_SET
Change the route aggregation attribute
HUAWEI TECHNOLOGIES CO., LTD.
All rights reserved
Page 4
HUAWEI


Contents
What is route aggregation
AS_SET
Change the route aggregation attribute
HUAWEI TECHNOLOGIES CO., LTD.
All rights reserved
Page 5
HUAWEI


Route Aggregation Overview
Routing Table
192.168.1.16/28
192.168.1.32/28
192.168.1.48/28
192.168.1.0/26
192.168.1.16/28
192.168.1.32/28
192.168.1.48/28
Routing Table
192.168.1.0/26
We can use the configuration command in BGP to summarize a range of prefixes into one aggregate.
HUAWEI TECHNOLOGIES CO., LTD. All rights reserved Page 6
HUAWEI



Route Aggregation Rule: Adopt the method of longest match subnet mask.



Automatic Aggregation: Aggregate the imported IGP subnet routes to their major network addresses automatically. Once the automatic aggregation is enabled, BGP will not receive the subnet routes that are redistributed from IGP. BGP will

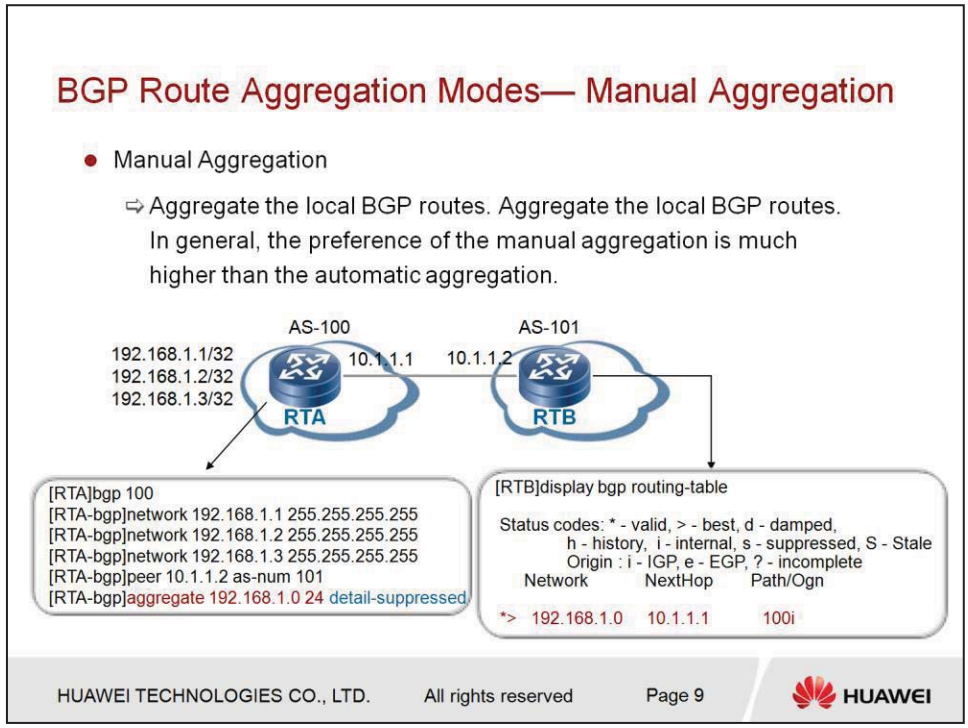
receive only the aggregate major network addresses.

The summary automatic command is used to aggregate the imported routes to their major network addresses automatically. The undo summary automatic is used to disable to the configuration. By default, the summary automatic function

is not enabled.

After the summary automatic function is enabled, BGP no longer receives the subnet routes from an IGP. BGP receives only the aggregate network addresses.

This can reduce the number of routes in the routing table



Manual Aggregation: Aggregate the local BGP routes. In general, the preference of the manual aggregation is much more higher than the automatic aggregation.

By default, both of the specific route and aggregate route will be advertised when manual aggregation is used.

The aggregate command is used create an aggregate record in the BGP routing table. The nn command is use disable the function. By default, no route aggregation is preformed.

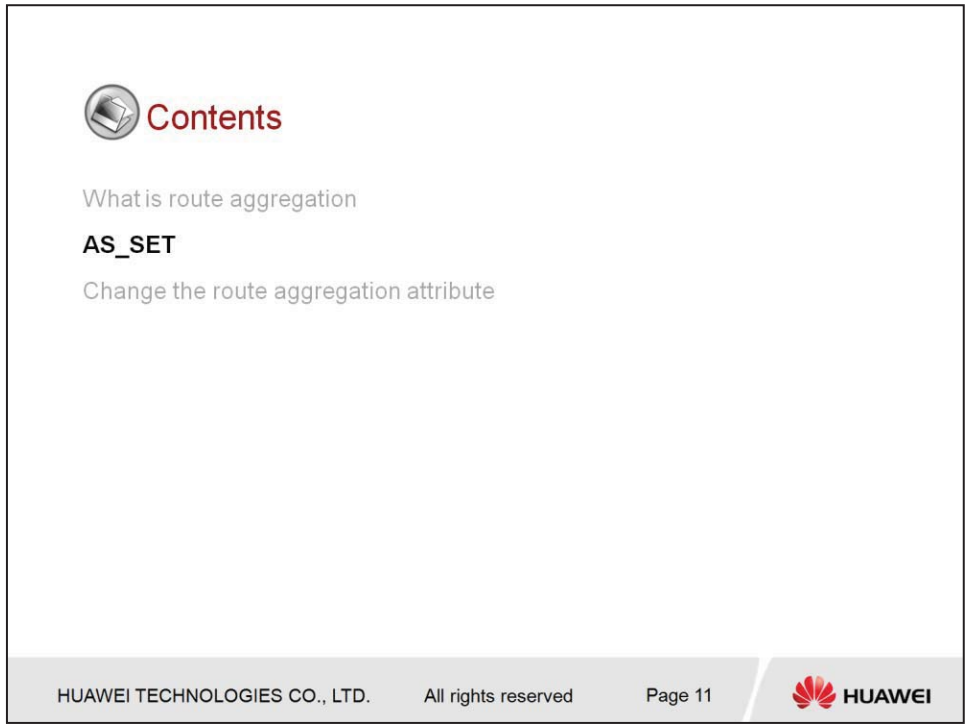
Multiple BGP specific BGP routes can be summarized into a single aggregate route by using “ aggregate” command. The aggregate route is then advertised to other peers. Different from the command “summary automatic”, the “aggregate”

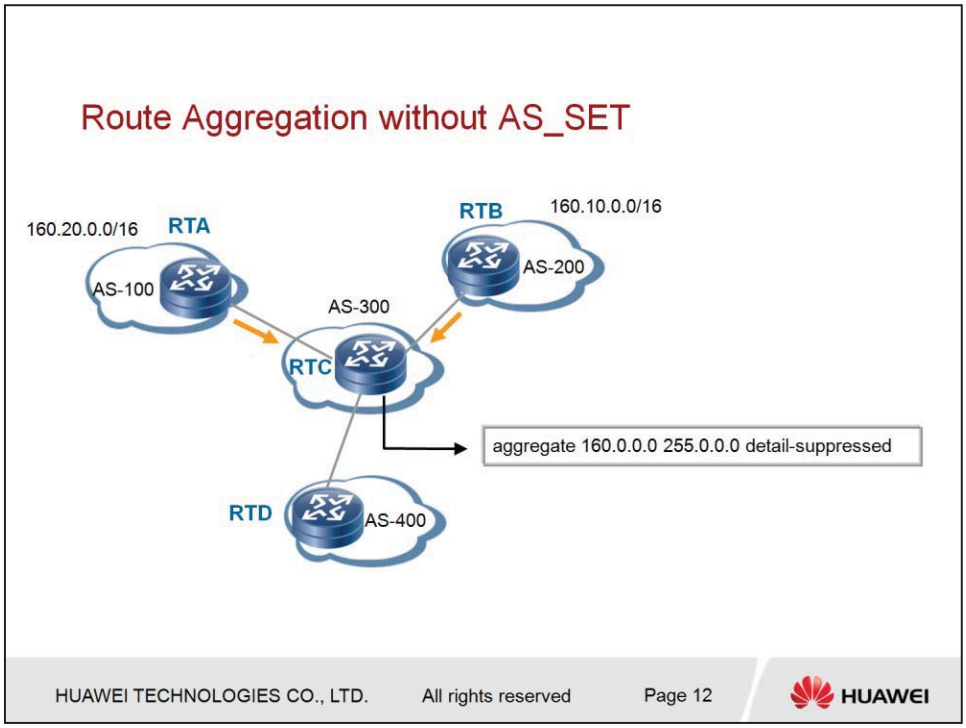
command requires us to manually specify the required aggregate prefix ip address mask [as-set | attribute-policy route-policy-name1 | detail suppressed | origin-policy route-policy-name2 | suppress-policy route-policyname3

BGP Route Aggregation Design Considerations

- BGP Route Aggregation Design Considerations
  - ⇨ The advertisement of specific route
  - ⇨ Inherited BGP Path Attribute
    - AS-Path
    - Origin
    - Community
    - … …

HUAWEI TECHNOLOGIES CO., LTD. All rights reserved


Contents
What is route aggregation
AS_SET
Change the route aggregation attribute
HUAWEI TECHNOLOGIES CO., LTD.
All rights reserved
Page 11
HUAWEI



The AS_PATH attribute has been learnt so far consists of an ordered sequence of AS numbers that describes the path to a particular destination.

The VRP platform of Huawei products support 4 types of AS_PATH attributes:

•AS_SEQUENCE

•AS_SET

•AS_CONFED_SEQUENCE

•AS_CONFED_SET

AS_SEQUENCE is the ordered list of AS numbers that a route has traversed.

Each time the route passes an AS, the AS number will be recorded in the front part of the list.

AS_SET is often used in route aggregation. It is an unordered list of the AS numbers along a path to a destination.

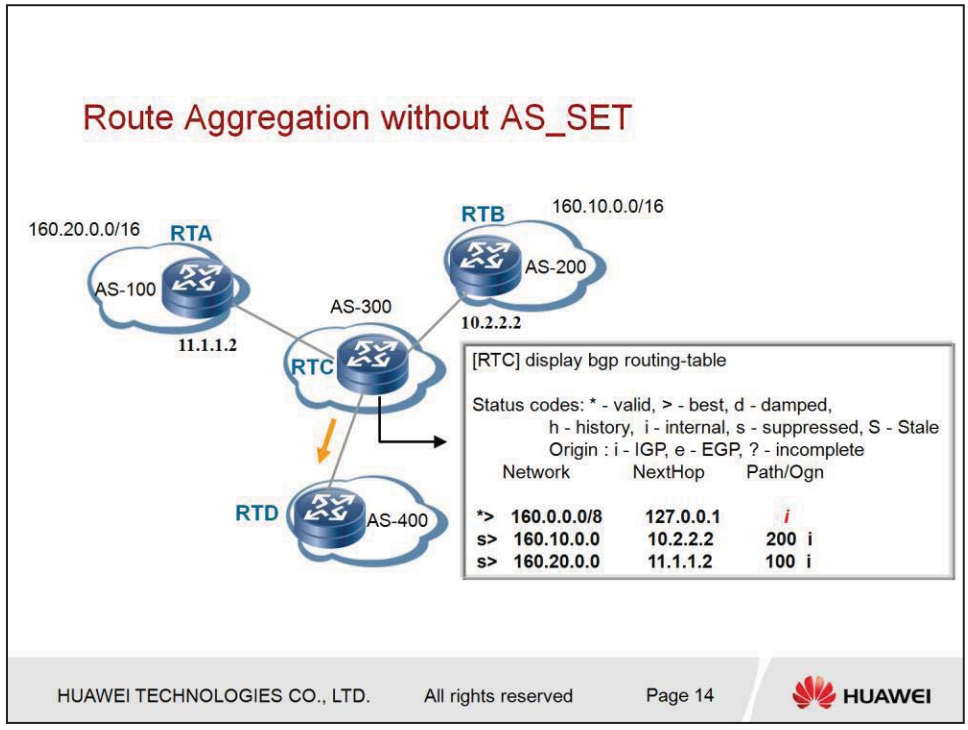
AS_CONFED_SEQUENCE is an ordered list of AS numbers along a path to a destination. It is used in the same way as the AS_SEQUENCE, except that the AS numbers in the list belong to autonomous systems within the local confederation.

l AS_CONFED_SET is an unordered list of AS numbers along a path to a destination. It is used in exactly the same way as the AS_SET, except that the AS numbers in the list belong to autonomous systems within the local confederation.

AS_CONFED can be used only in BGP confederation. When an update is sent to a peer external to the confederation, the AS_CONFED information is removed.

In the diagram above, RTA and RTB advertise their local network to RTC. RTC will summarized those networks and advertised it to RTD by using command:

aggregate 160.0.0.0 255.0.0.0 detail-suppressed . As a result, the routes 162.20.0.0/16 and 160.10.0.0/16 are summarized into a single route 160.0.0.0/8.



Observe the routing table of RTC by using the command “display bgp routing table”

Status codes: * - valid, > - best, d - damped,

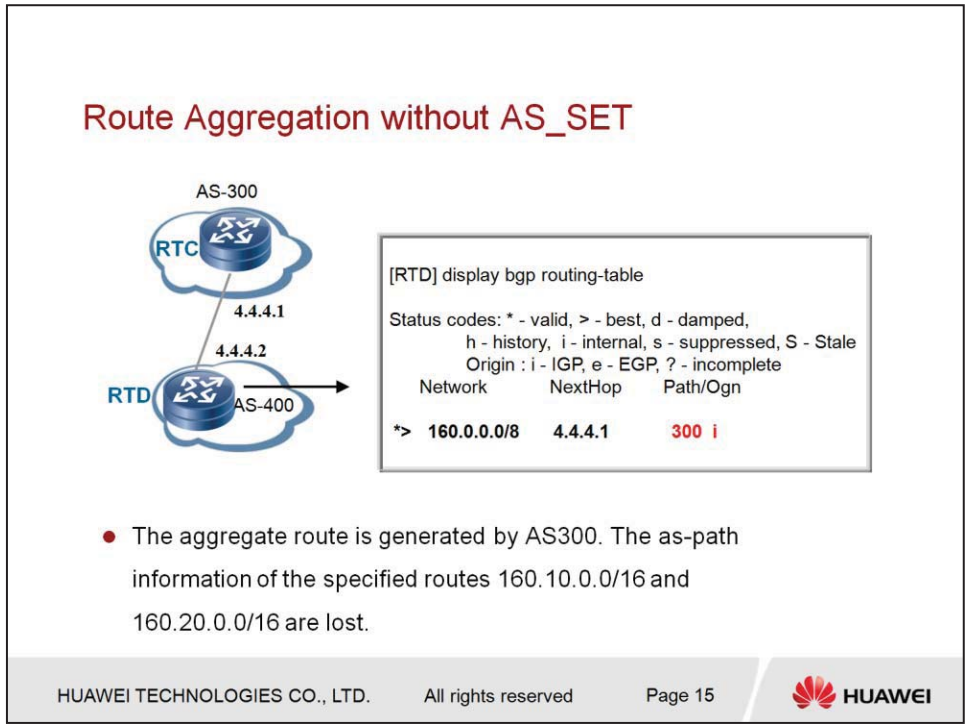
h - history, i - internal, s - suppressed, S - Stale

Origin : i - IGP, e - EGP, ? - incomplete

Network	NextHop	Path/Ogn
*> 160.0.0.0/8	127.0.0.1	i
s> 160.10.0.0	10.2.2.2 200	i
s> 160.20.0.0	11.1.1.2 100	i

It is shown that an additional aggregate route exists in the BGP routing table after the configuration . Besides, the AS_PATH attribute of the aggregate route not carry any AS information. This indicates that the aggregate route without AS_SET

is generated by RTC.



RTC aggregate the route 160.20.0.0/16 from AS 100 and the route 160.10.0.0/16 from AS200. Only the aggregate route is advertised to RTD since RTC has issued the parameter detail-suppressed. The specified routes 160.10.0.0/16 and 160.20.0.0/16 are suppressed. Below is the BGP routing table information of RTD.

Please take note of the as-path attribute of the aggregate route.

[RTD] display bgp routing-table

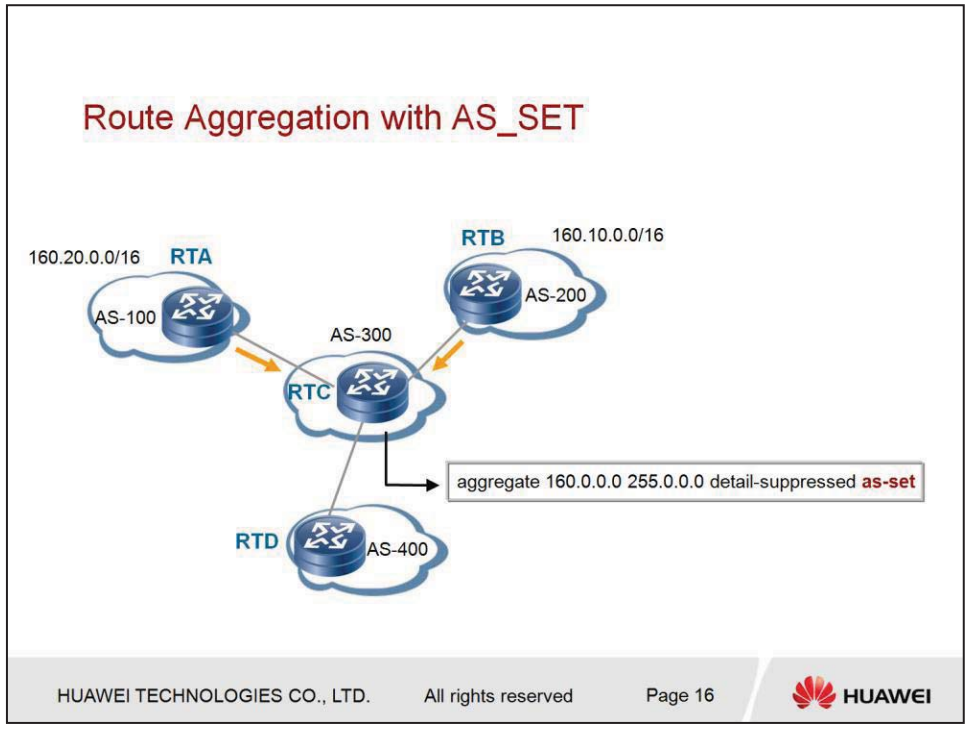
Status codes: * - valid, > - best, d - damped,

h - history, i - internal, s - suppressed, S - Stale

Origin : i - IGP, e - EGP, ? - incomplete

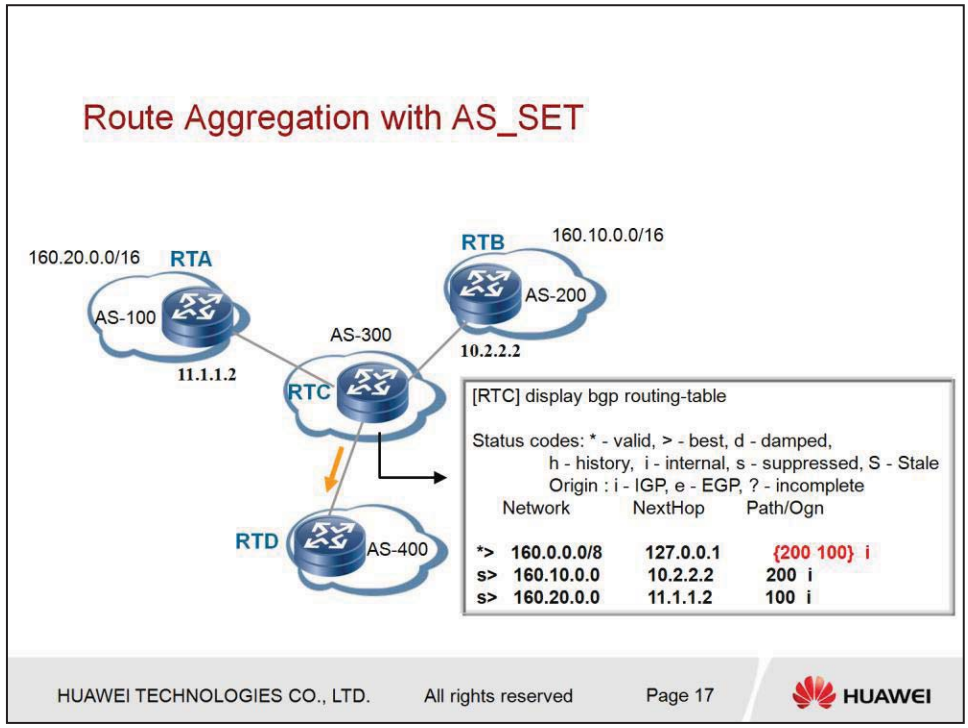
Network	NextHop	Path/Ogn
*> 160.0.0.0/8	4.4.4.1	300 i

The aggregate route 160.0.0.0/8 is generated by AS300 with origin IGP. The aspath information of the specified route 160.10.0.0/16 and 160.20.0.0/16 are lost.



Now, we add the parameter as-set in the aggregate command of RTC. The configuration is as below.

aggregate 160.0.0.0 255.0.0.0 detail-suppressed as-set.



The BGP routing table of RTC after the modification of configuration is as below:

[RTC] display bgp routing-table

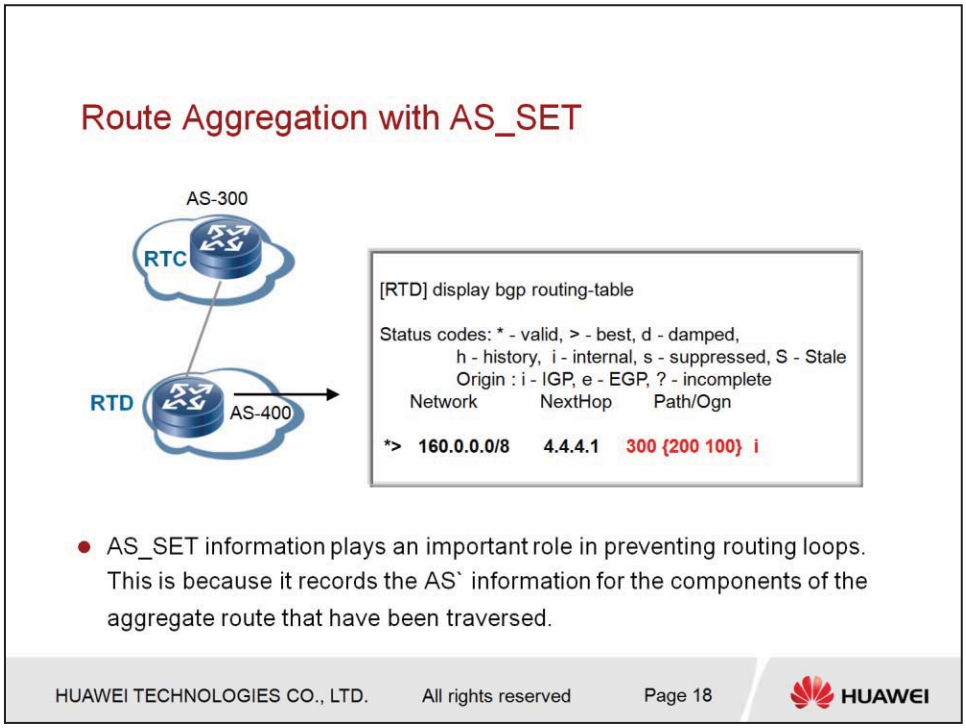
Status codes: * - valid, > - best, d - damped,

h - history, i - internal, s - suppressed, S - Stale

Origin : i - IGP, e - EGP, ? - incomplete

Network	NextHop	Path/Ogn
*> 160.0.0.0/8	127.0.0.1	{200 100} i
s> 160.10.0.0	10.2.2.2	200 i
s> 160.20.0.0	11.1.1.2	100 I

After configuring the as-set parameter, the as-path information of the BGP routing table on RTC is shown as {200,100}. This indicates the router has aggregated the routes from A200 and AS100.



The BGP routing table of RTD after the modification in configuration is as below:

[RTD] display bgp routing-table

Status codes: * - valid, > - best, d - damped,

h - history, i - internal, s - suppressed, S - Stale

Origin : i - IGP, e - EGP, ? - incomplete

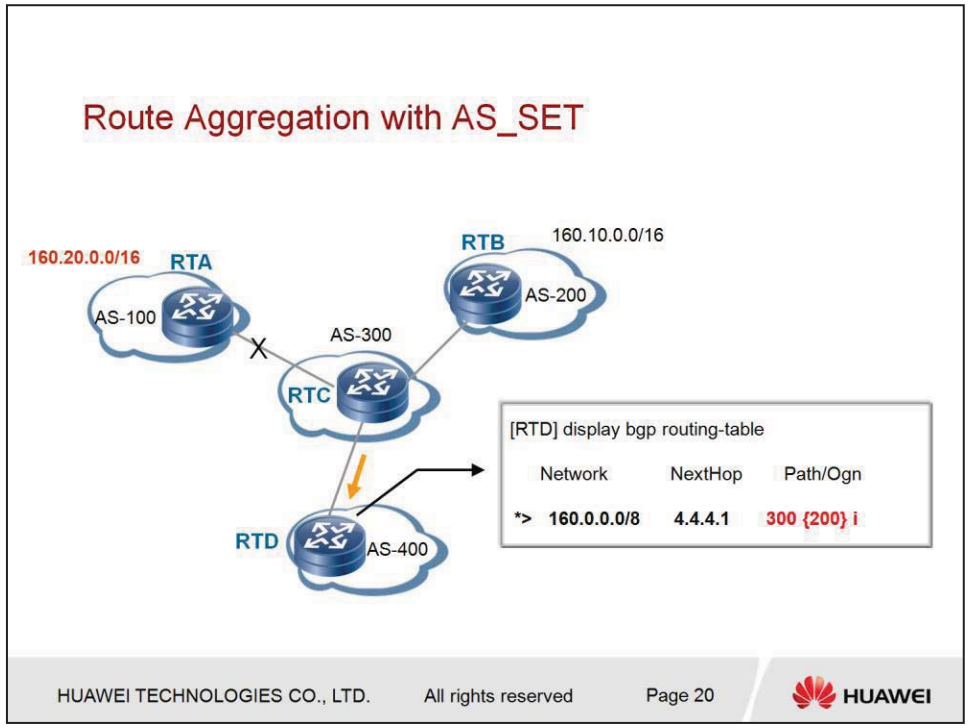
Network	NextHop	Path/Ogn
*> 160.0.0.0/8	4.4.4.1 300	{200 100} i

AS_SET information plays an important role in preventing routing loop. This is because it records the AS' information for the components of the aggregate route that have been traversed. An aggregate route might be re-entered into any AS' listed in the AS_SET through BGP route.

This will result in routing loop. Once BGP loop detection mechanism finds its AS number is in the list of the AS_SET attribute of the aggregate route received, it will ignore that route. This will prevent the routing loop.

Note: After the as-set parameter is used, the as-path information of the aggregate route contains the as-path information for each

of the specific routes being aggregated. This information is updated according to the changes occurring in the specific routes being aggregated. When the route 160.20.0.0/16 in the example above fails, the as-path information of the aggregate route will be changed from 300{200，100} to 300 {200}. This result in the attribute of the aggregate route change. The aggregate route will be in the unstable state if the router aggregates thousands of routes and at the same times problems keep happening to those specific routes being aggregated.

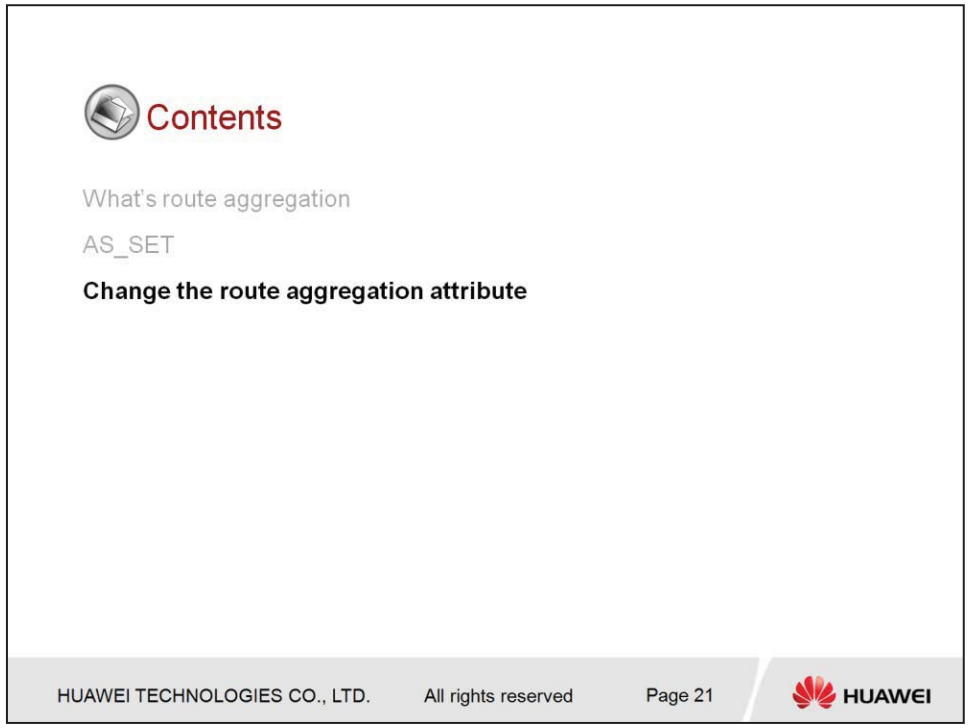


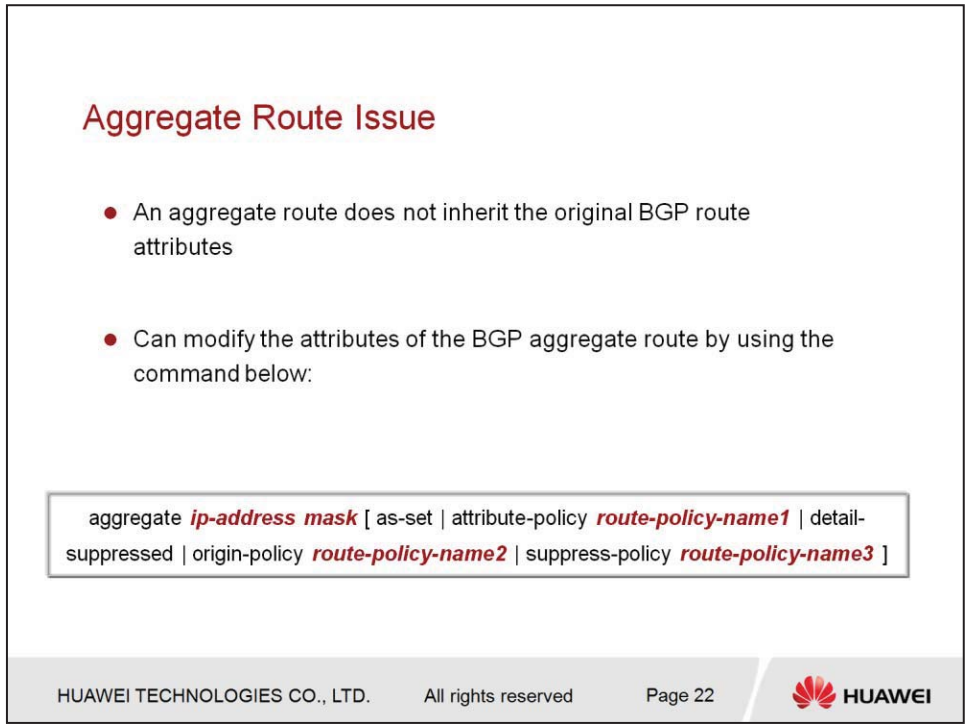
.

This infoAfter the as-set parameter is used, the as-path information of the aggregate route contains the as-path information for each of the specific routes being aggregated,

information is updated according to the changes occurring in the specific routes being aggregated. When the route 160.20.0.0/16 in the example above fails, the as-path information of the aggregate route will be changed from 300{200，100} to 300 {200}. This results in the attribute of the aggregate route

change. The aggregate route will be in the unstable state if the router aggregates thousands of routes and at the same times problems keep happening to those specific routes being aggregated.

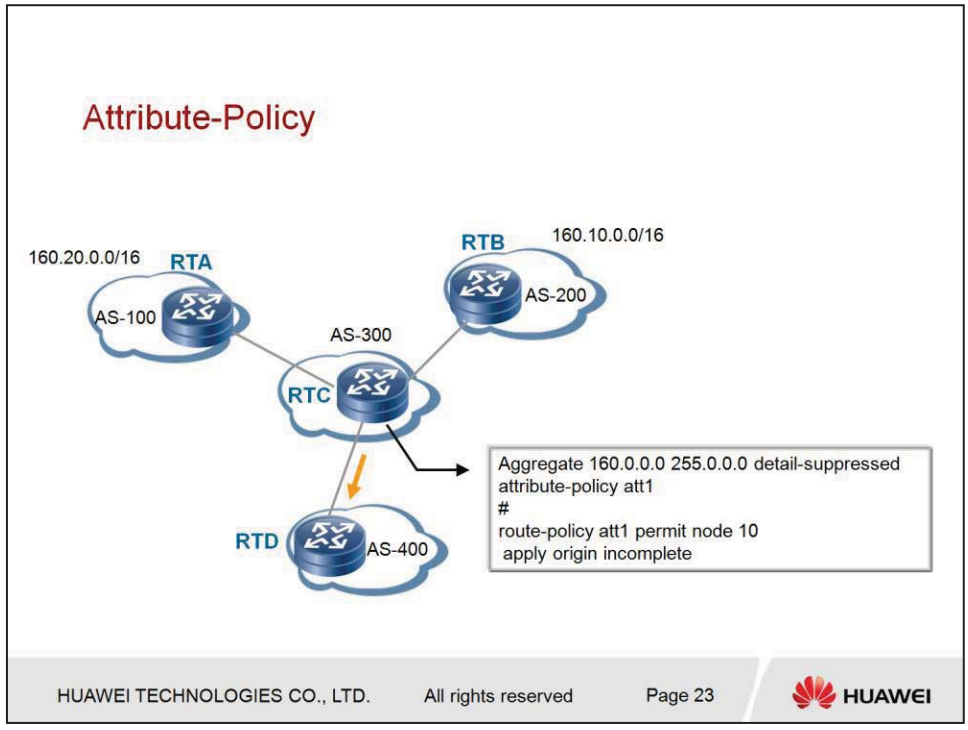

Contents
What's route aggregation
AS_SET
Change the route aggregation attribute
HUAWEI TECHNOLOGIES CO., LTD.
All rights reserved
Page 21
HUAWEI



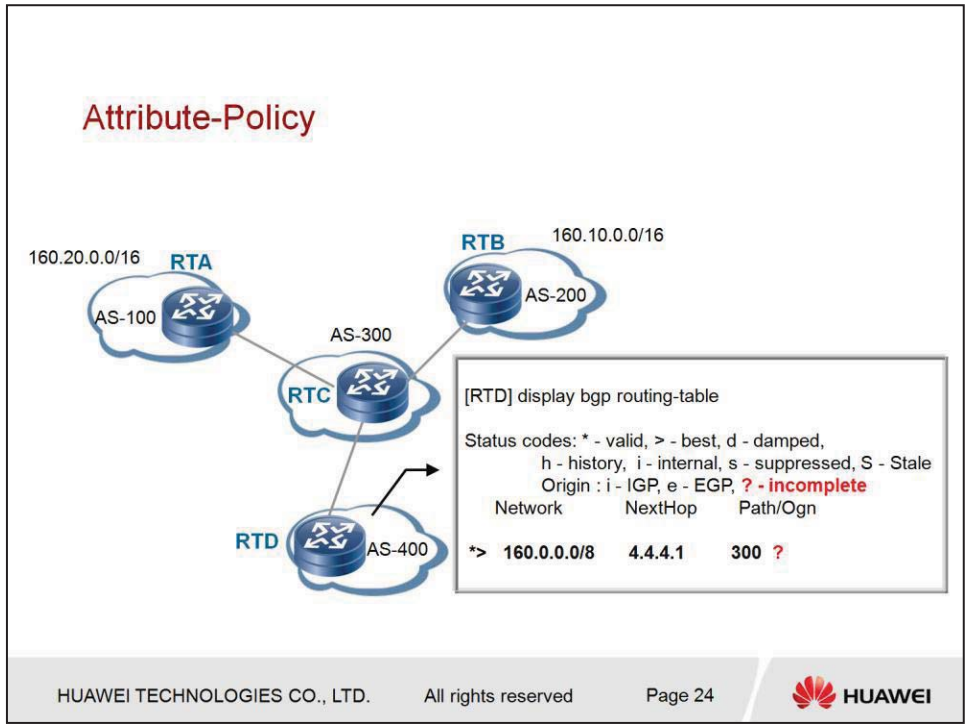
Aggregate route does not inherit the original BGP route attributes by default.

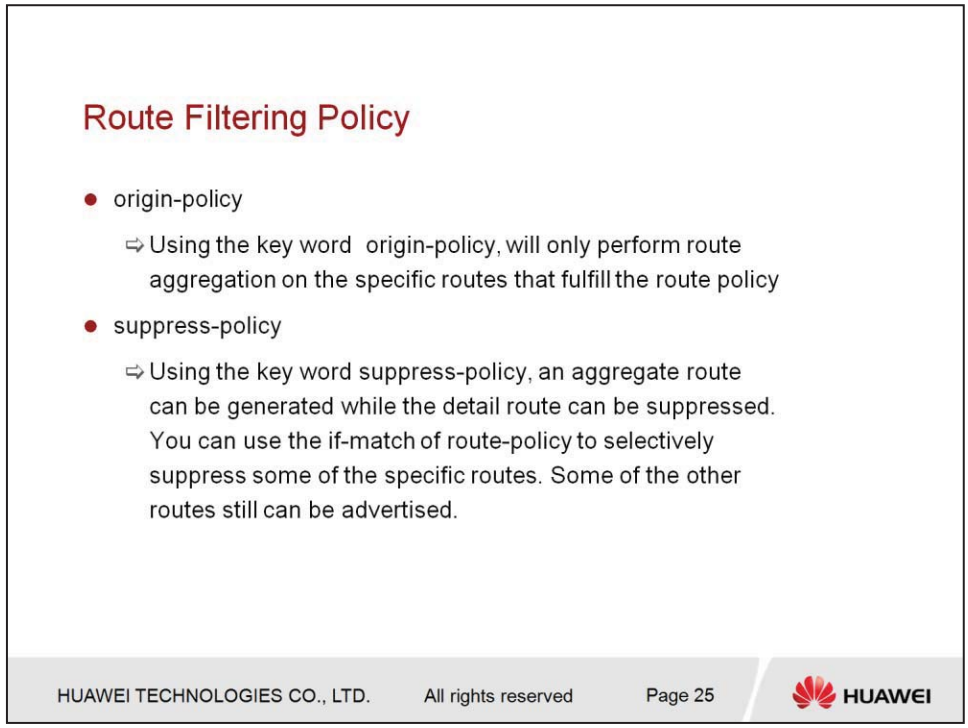
The command aggregate ip-address mask [as-set | attribute-policy routepolicy- name1 | detail-suppressed | origin-policy route-policy-name2 |

suppress-policy route-policy-name3] is used to modify the attributes of the BGP aggregate route.



Using the key word attribute policy, you can set the attribute of the aggregate route.


Attribute-Policy
160.20.0.0/16
RTA
AS-100
RTB
160.10.0.0/16
AS-200
AS-300
RTC
RTD
AS-400
[RTD] display bgp routing-table
Status codes: * - valid, > - best, d - damped,
h - history, i - internal, s - suppressed, S - Stale
Origin : i - IGP, e - EGP, ? - incomplete
Network NextHop Path/Ogn
*> 160.0.0.0/8 4.4.4.1 300 ?
HUAWEI TECHNOLOGIES CO., LTD. All rights reserved Page 24
HUAWEI



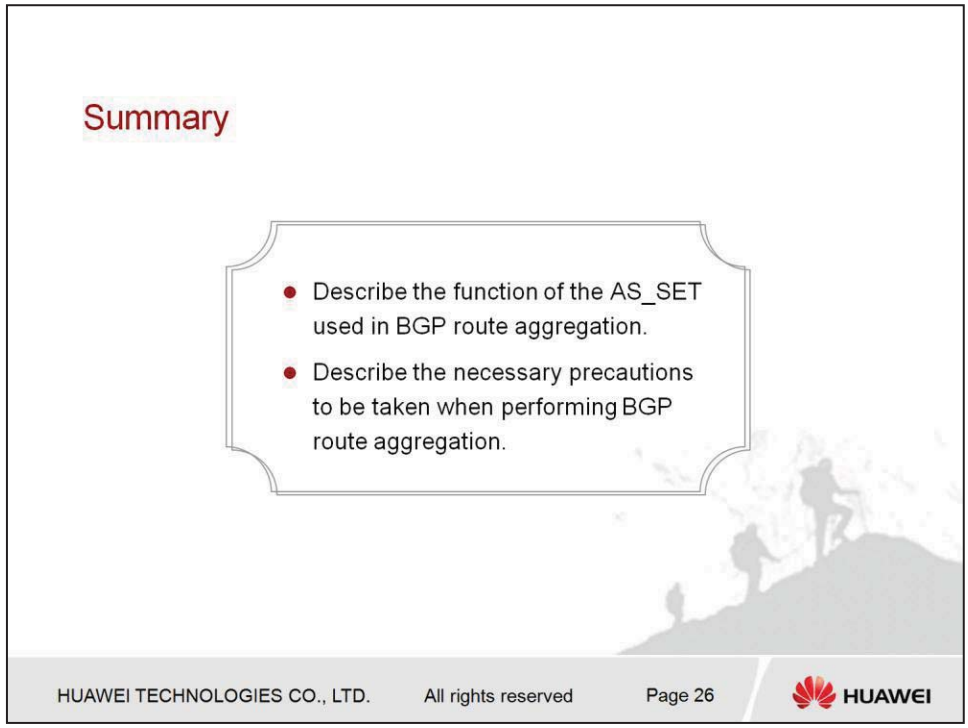
origin-policy

Using the key word origin-policy, will enable aggregation of only routes that match the route policy.

suppress-policy

Using the key word suppress-policy, an aggregate route can be generated while the detail route can be suppressed. You can use the ifmatch of route-policy to selectively suppress some of the specific routes.

Some of the other routes still can be advertised.

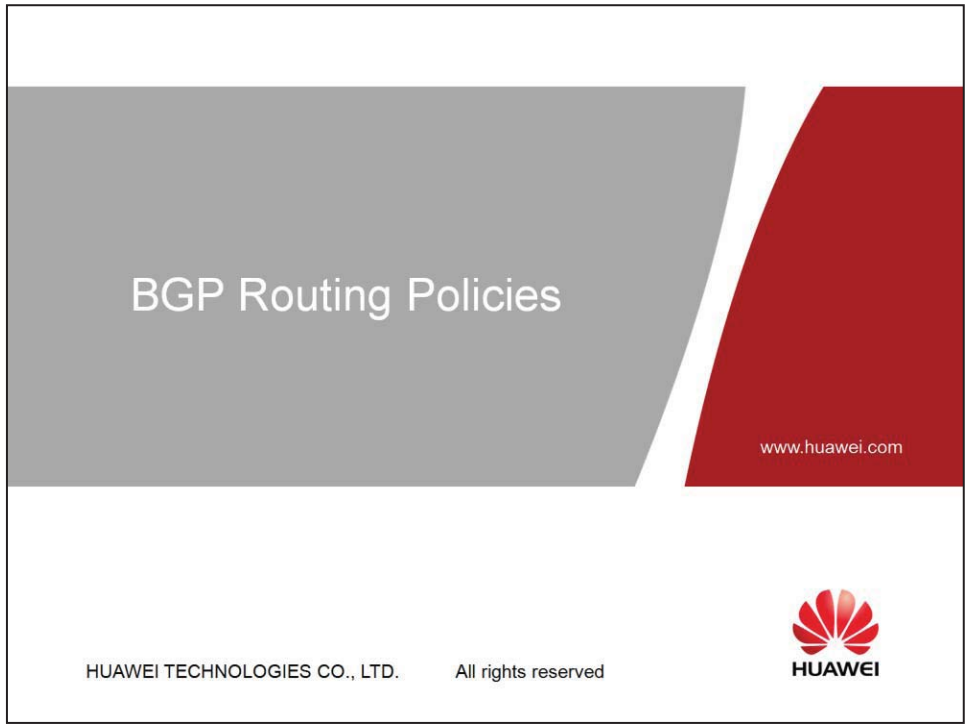


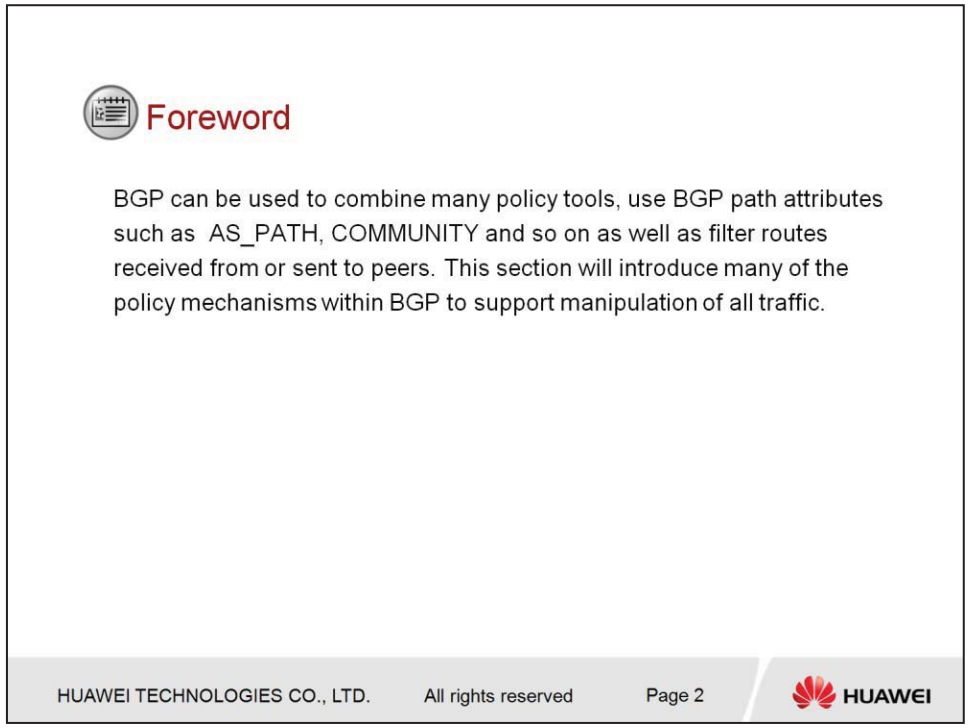
1. Describe the function of AS_SET used in BGP aggregate route.

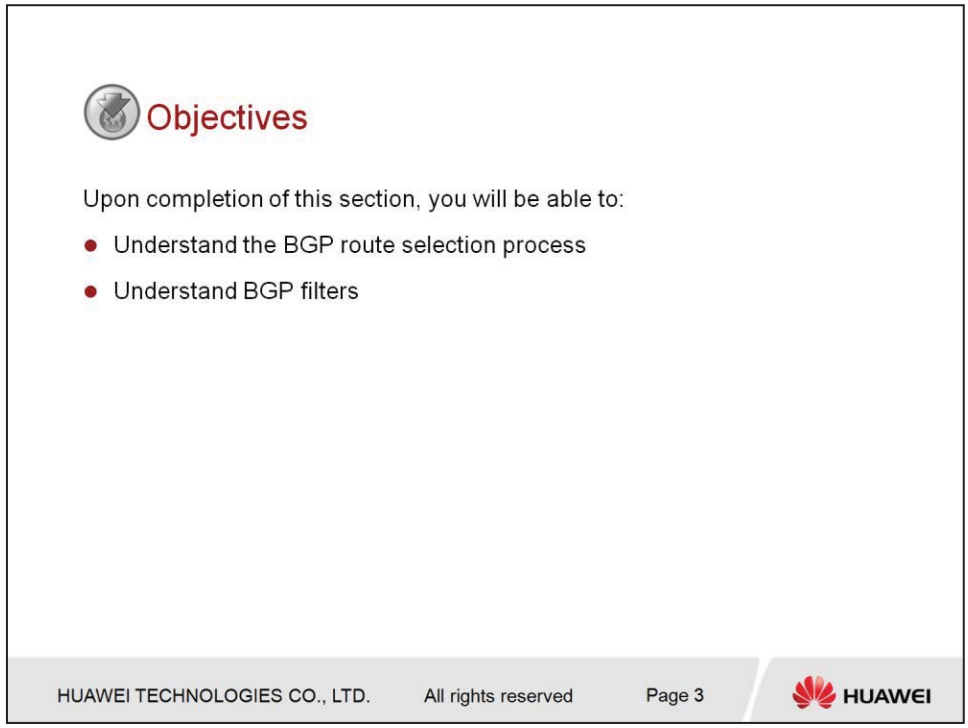
A：Generates aggregate route that contains the AS' information for the components of the aggregate route that have been traversed.

2. Describe the precaution when perform BGP route aggregation.

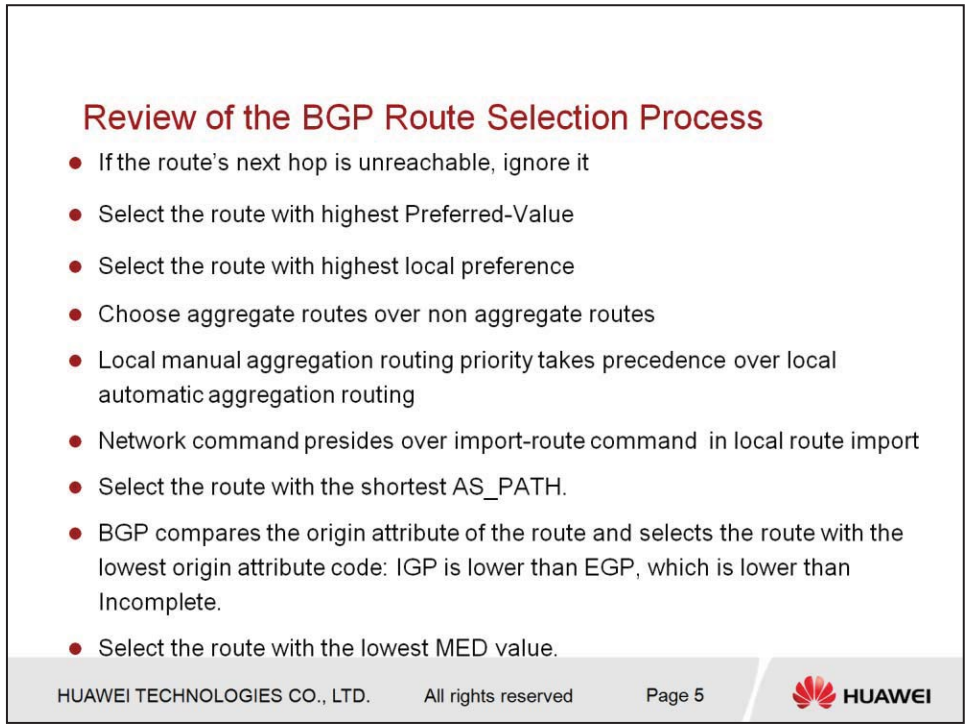
A：Take note the change in the attributes of the aggregate route. For example, the change happen in the AS_PATH attribute. If the AS_SET parameter is not included when configure the route aggregation, system will indicate that the aggregate route is generated by the aggregator. Therefore, only the local AS number of the aggregator is included in the as-path attribute. In addition, we need to take note the change for other BGP attributes as well.


BGP Routing Policies
www.huawei.com
HUAWEI TECHNOLOGIES CO., LTD.
All rights reserved
HUAWEI

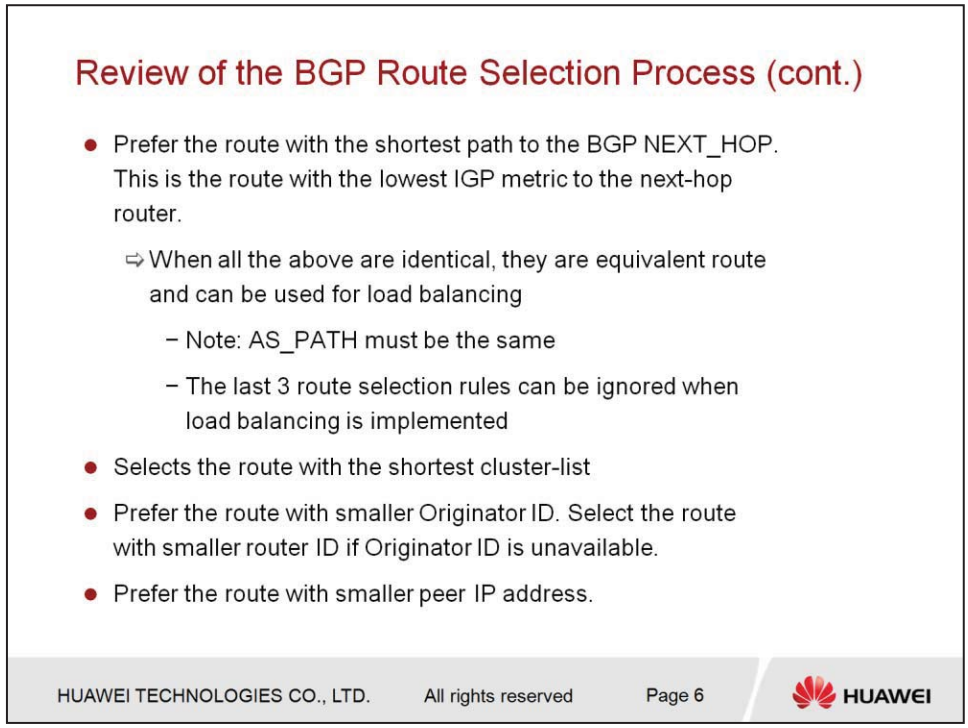

Foreword
BGP can be used to combine many policy tools, use BGP path attributes such as AS_PATH, COMMUNITY and so on as well as filter routes received from or sent to peers. This section will introduce many of the policy mechanisms within BGP to support manipulation of all traffic.
HUAWEI TECHNOLOGIES CO., LTD.
All rights reserved
Page 2
HUAWEI


Objectives
Upon completion of this section, you will be able to:
Understand the BGP route selection process
Understand BGP filters
HUAWEI TECHNOLOGIES CO., LTD.
All rights reserved
Page 3
HUAWEI


Contents
Review of the BGP Route Selection Process
BGP path selection
HUAWEI TECHNOLOGIES CO., LTD.
All rights reserved
Page 4
HUAWEI



1. Select the route with highest Preferred-Value
2. Select the route with highest local preference
3. Prefer the aggregated route to the detailed routes
1. Prefer the route with the shortest AS_PATH
2. BGP compares the origin attribute of the route and selects the route with the lowest origin attribute code: IGP is lower than EGP, which is lower than Incomplete.
3. Select the route with the lowest MED value.
4. Prefer EBGP routes over IBGP routes



8. Prefer the route with the shortest path to the BGP NEXT_HOP. This is the route with the lowest IGP metric to the next-hop router.

When all the above are identical, they are equivalent route and can be used for load balancing.

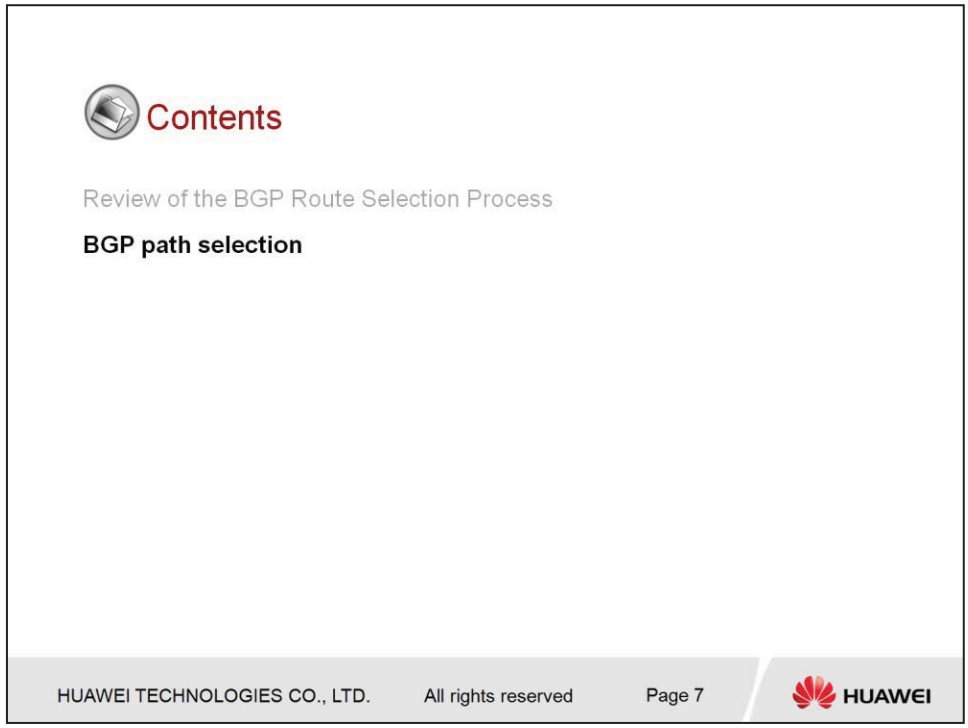
Note: AS_PATH must be the same

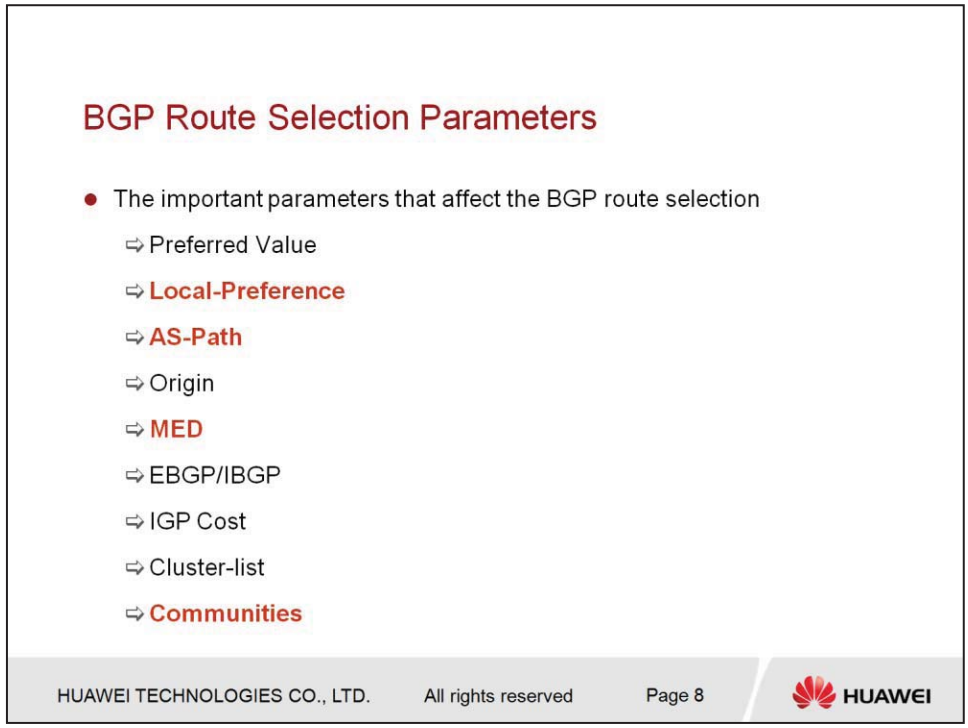
The last 3 route selection rules can be ignored when load balancing is implemented

9. Selects the route with the shortest cluster-list

10. Prefer the route with smaller Originator ID. Select the route with smaller router ID if Originator ID is unavailable.

11. Prefer the route with smaller peer IP address.


Contents
Review of the BGP Route Selection Process
BGP path selection
HUAWEI TECHNOLOGIES CO., LTD.
All rights reserved
Page 7
HUAWEI



Base on the BGP route selection process and the common BGP path attributes, 9 important BGP parameters that affect the BGP route selection are as follow:

•Preferred Value

•LOCAL_PREF

•AS_PATH

•ORIGIN

•MED

•Is the neighbor type EBGP or IBGP?

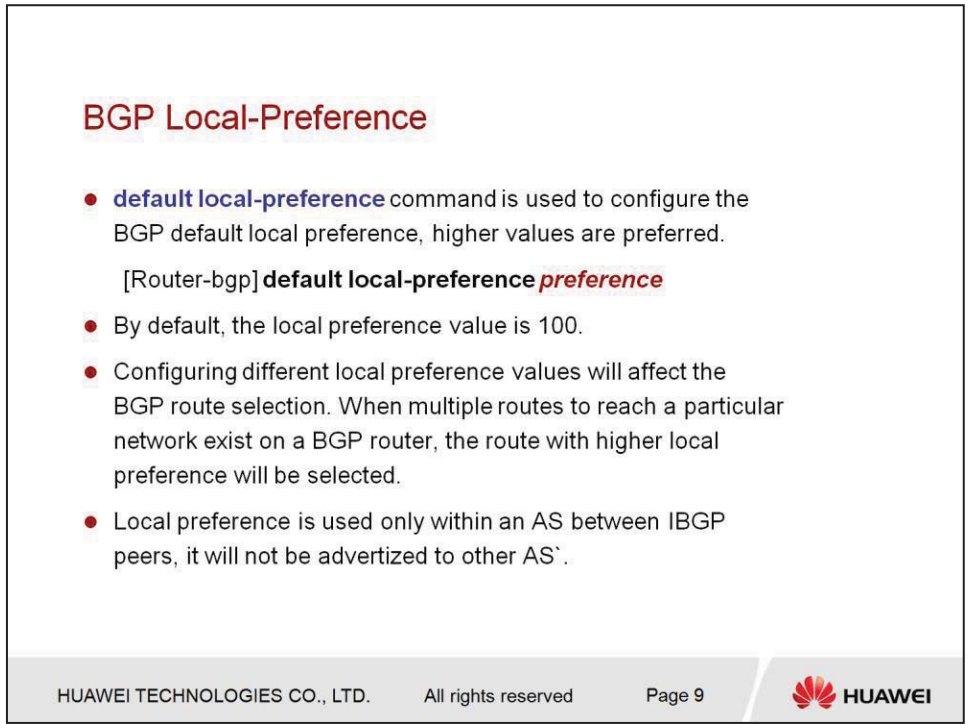
•Internal cost of IGP

•Cluster-list

•COMMUNITY

The parameters specified above can directly affect the BGP route selection.

Among them, the most common used parameters are LOCAL_PREF, AS_PATH and MED. The function and the configuration of these parameters will be discussed in detail in the following slides.

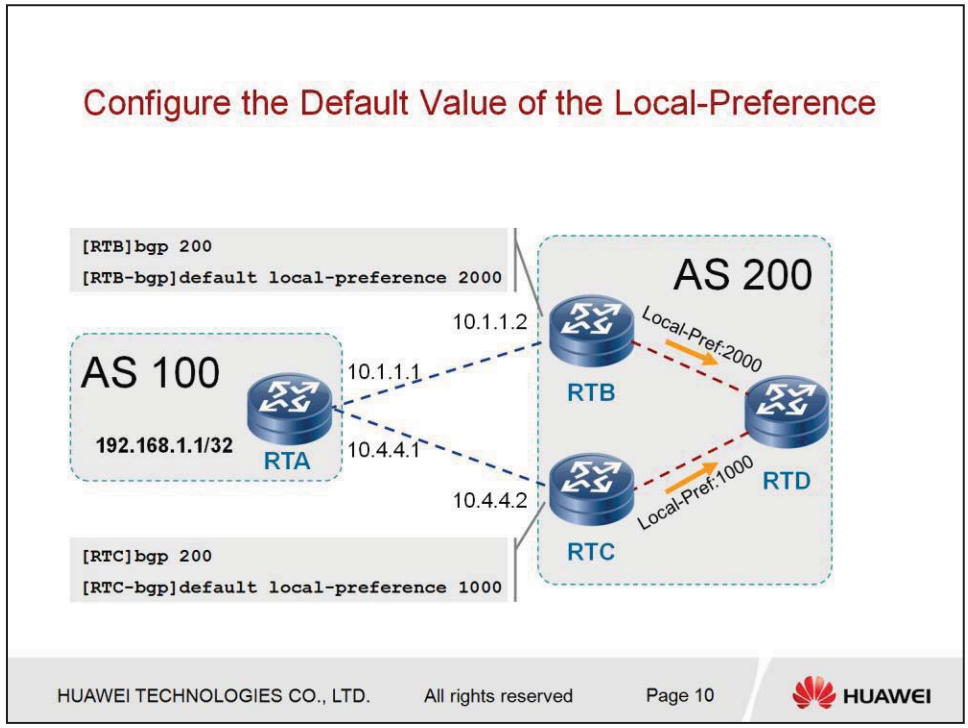


LOCAL_PREF attribute is used within an AS between IBGP peers to determine the best path for the data traffic to exit an AS. LOCAL_PREF attributes become very useful when there are multiple routes to leave the AS and reach an outside

network.

default local-preference command is used to configure the BGP default local preference, higher values are preferred.

The range of the LOCAL_PREF value on Huawei equipments are from 0 to 4294967295. Its default value is 100

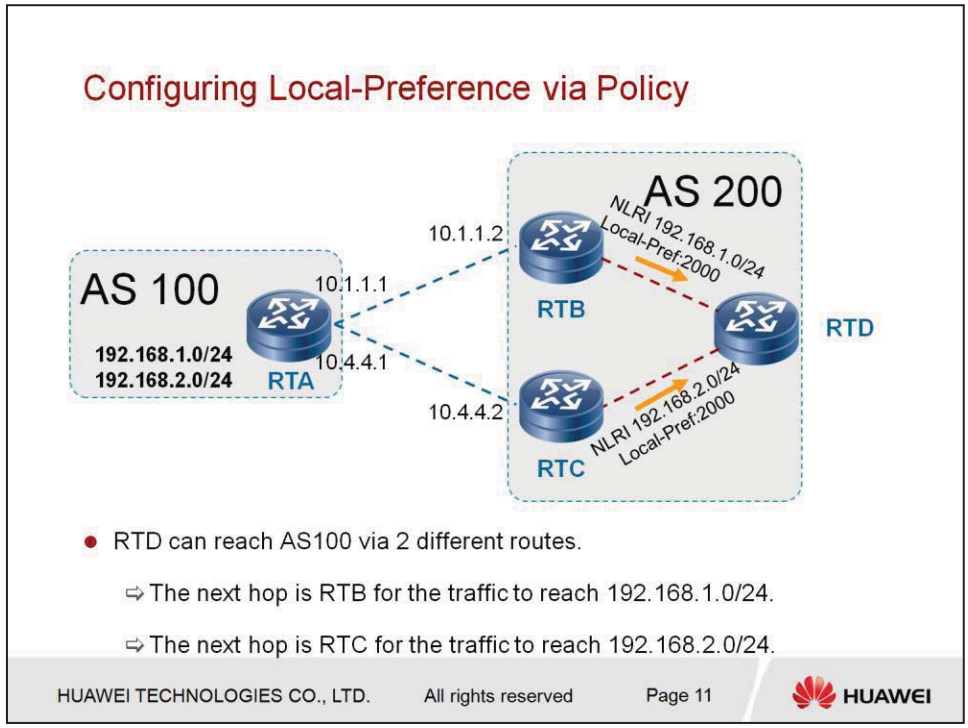


As shown in the diagram above, router D will receive 2 different update information to reach the network 192.168.1.1/32 from RTB and RTC respectively.

By default, RTD will perform the BGP route selection. The selected best route will be responsible to send the data to the destination 192.168.1.1/32. Both RTB and RTC change their default local preference value by using the “default localpreference”

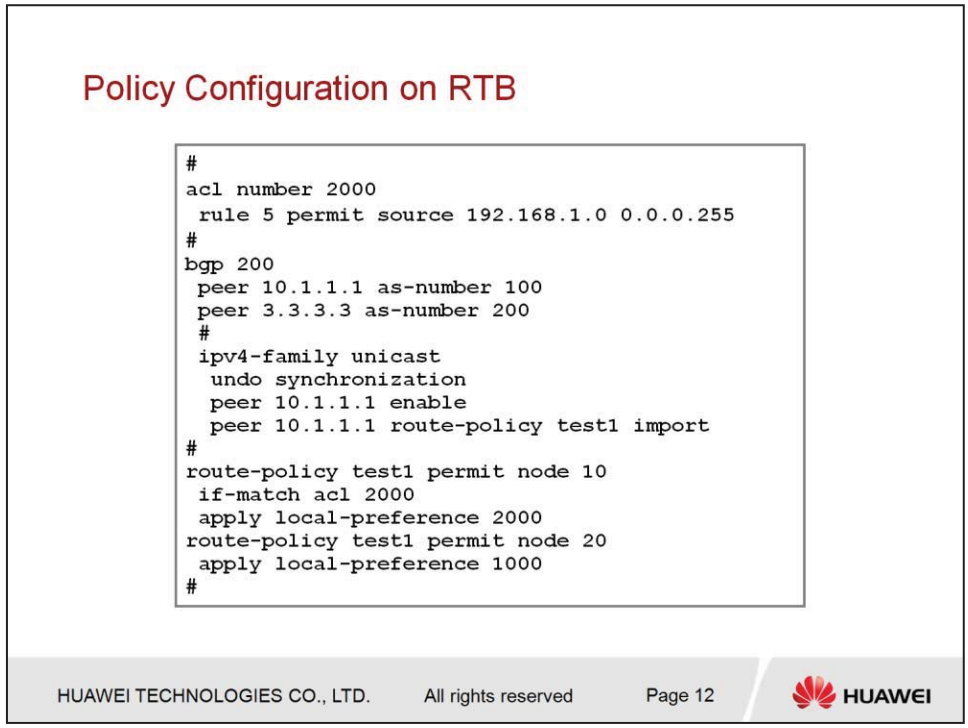
command . The local preference of RTB has been changed to 2000 while the local preference of RTC has been changed to 1000. As a result, RTD receives 2 routes that carry different local preference value toward 192.168.1.1/32.

One of the routes carries the local preference 2000 with the IP prefix of RTB as the next hop IP address. Another route carries the local preference 1000 with the IP prefix of RTC as the next hop IP address. In this circumstance (all other parameters adopt the default value), the route with RTB as next hop IP address will be selected as the best route which in charge of the data transmission to reach 192.168.1.1/32. This is because the highest local preference is preferred.



LOCAL_PREF is an important parameter that affect the BGP route selection within an AS. BGP can work together with some policy tools to implement load balancing in some complex network.

As shown in the diagram above, RTD can reach 192.168.1.0/24 and 192.168.2.0 in AS100 via 2 different routes. We can set the next hop IP address as RTB for the data to reach 192.168.1.0/24 and set the next hop IP address as RTC for the data to reach 192.168.2.0/24. These can be achieved by configuring the local preference via policy.



Explanation of the configuration:

#

acl number 2000

rule 5 permit source 192.168.1.0 0.0.0.255

#

Specify the range of IP addresses that match the ACL 2000.

bgp 200

peer 10.1.1.1 as-number 100

peer 3.3.3.3 as-number 200

#

ipv4-family unicast

undo synchronization

peer 10.1.1.1 enable

peer 10.1.1.1 route-policy test1 import

#

Apply route policy test1 for the route information received from peer 10.1.1.1.

#

route-policy test1 permit node 10

if-match acl 2000

apply local-preference 2000

route-policy test1 permit node 20

apply local-preference 1000

#

Node 10 in the routing policy assigns LOCAL_PREF 2000 to the network that

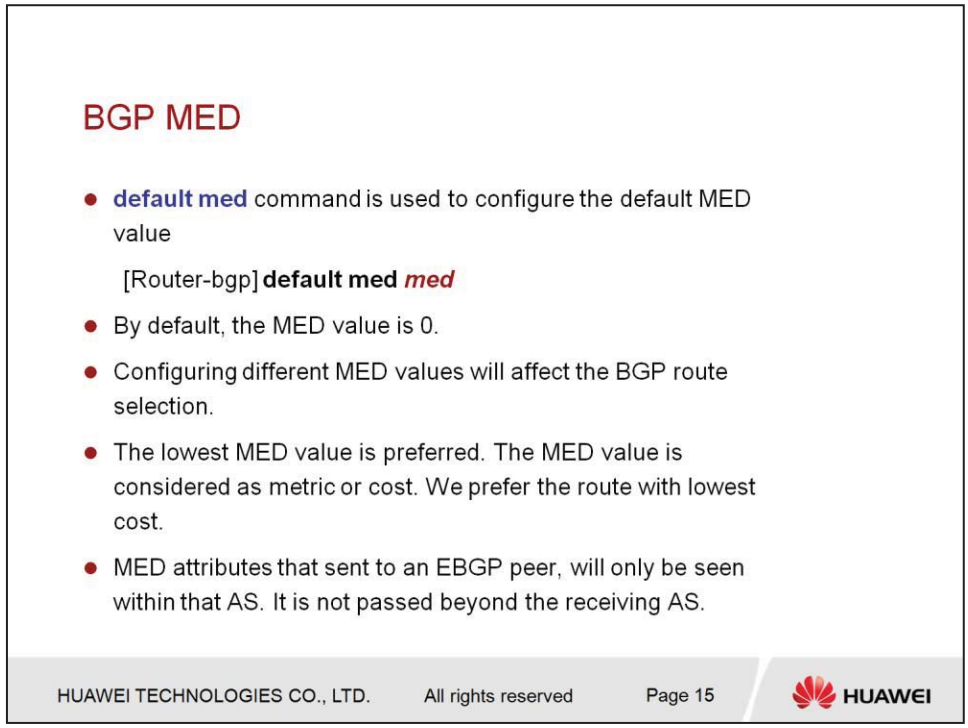
match ACL2000. Node 20 in the routing policy assigns LOCAL_PREF 1000 to

those networks that do not match ACL2000.

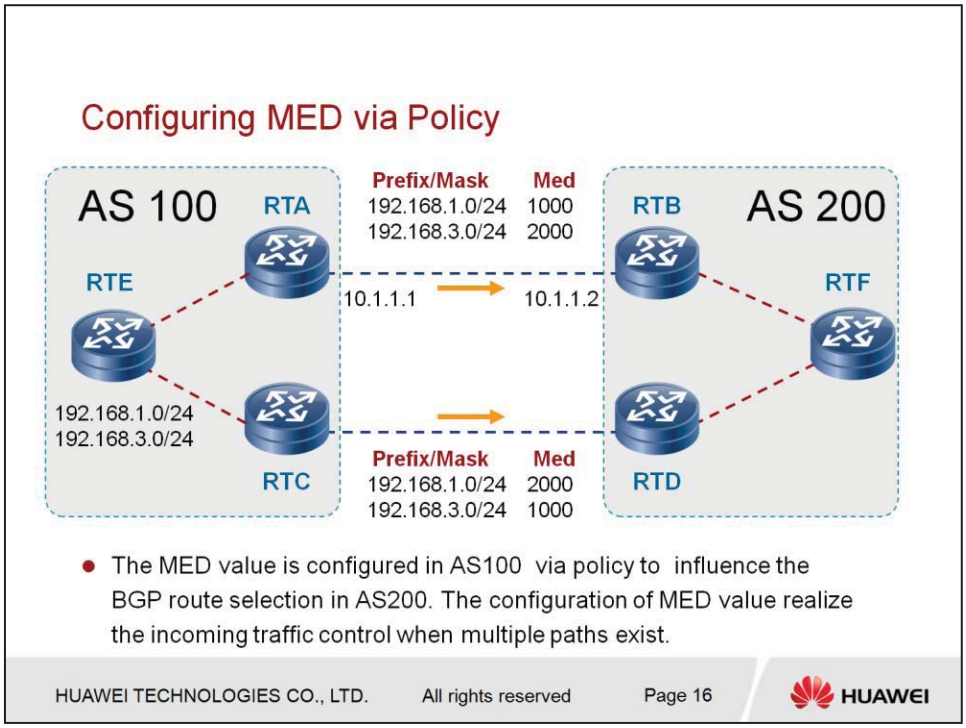
Policy Configuration on RTC

```
#
acl number 2000
 rule 5 permit source 192.168.2.0 0.0.0.255
#
bgp 200
 peer 10.4.4.1 as-number 100
 peer 2.2.2.2 as-number 200
 #
 ipv4-family unicast
  undo synchronization
  peer 10.4.4.1 enable
  peer 10.4.4.1 route-policy test1 import
#
route-policy test1 permit node 10
 if-match acl 2000
 apply local-preference 2000
route-policy test1 permit node 20
 apply local-preference 1000
#
```

HUAWEI TECHNOLOGIES CO., LTD. All rights reserved



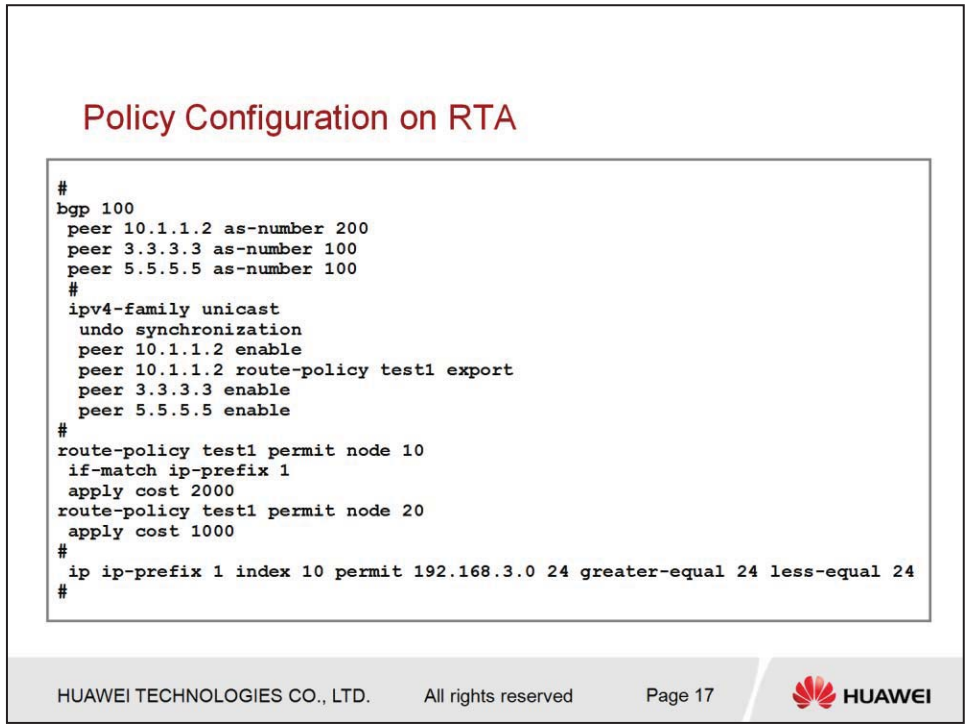
MED attribute is used to determine the entry point to an AS. It is used when multiple paths exist between 2 AS' and one AS is trying to influence the entry point from another AS. MED attribute is considered as metric or cost. Its value is in the range of 0 to 4284967295. The lowest MED value is preferred. By default, the MED value is 0.



The MED value is configured in AS100 via policy to influence the BGP route selection in AS200. The configuration of MED value realize the incoming traffic control when multiple paths exist.

There are 2 networks in AS100: 192.168.1.0/24 and 192.168.3.0/24. The routing updates of these 2 networks is sent to AS200 via RTA and RTC. On RTA, the MED value is configured as 1000 for the data to reach 192.168.1.0/24 and the

MED value is configured as 2000 for the data to reach 192.168.3.0/24. On RTC, the MED value is configured as 2000 for the data to reach 192.168.1.0/24 and the MED value is configured as 1000 for the data to reach 192.168.3.0/24. After the configuration, the traffic of RTF in AS200 can reach 192.168.3.0 via RTC and the data traffic of RTF in AS200 can reach 192.168.1.0 via RTA. In AS100, the load balancing is implemented base on the incoming traffic.



Explanation of the configuration:

#

bgp 100

peer 10.1.1.2 as-number 200

peer 3.3.3.3 as-number 100

peer 5.5.5.5 as-number 100

#

ipv4-family unicast

undo synchronization

peer 10.1.1.2 enable

peer 10.1.1.2 route-policy test1 export \\apply rout e

policy test1 for the route sent to peer 10.1.1.2

peer 3.3.3.3 enable

peer 5.5.5.5 enable

#

route-policy test1 permit node 10

\\route-policy test1，node 10

if-match ip-prefix 1

\\if match ip-prefix 1, apply cost (also called as MED) 2000

apply cost 2000

route-policy test1 permit node 20

\\route-policy test1，node 20

apply cost 1000

\\for all other routes that do not match ip-prefix 1, apply cost 2000

#

ip ip-prefix 1 index 10 permit 192.168.3.0 24 greater-equal 24 less-equal 24

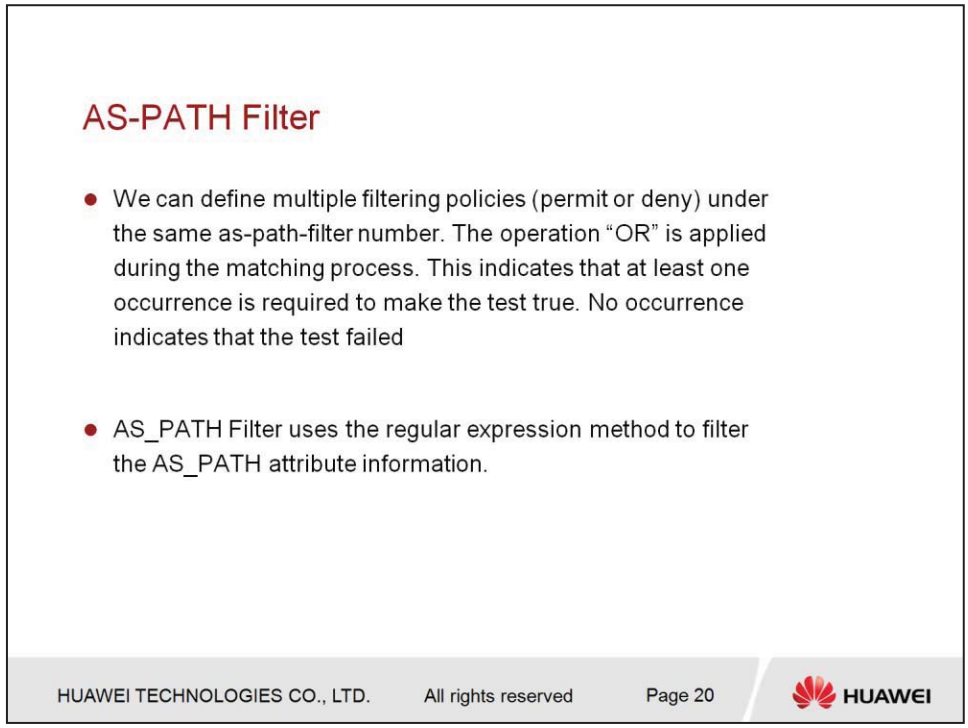
\\defines the range of IP address as 192.168.3.0/24 through the ip-prefix list

#

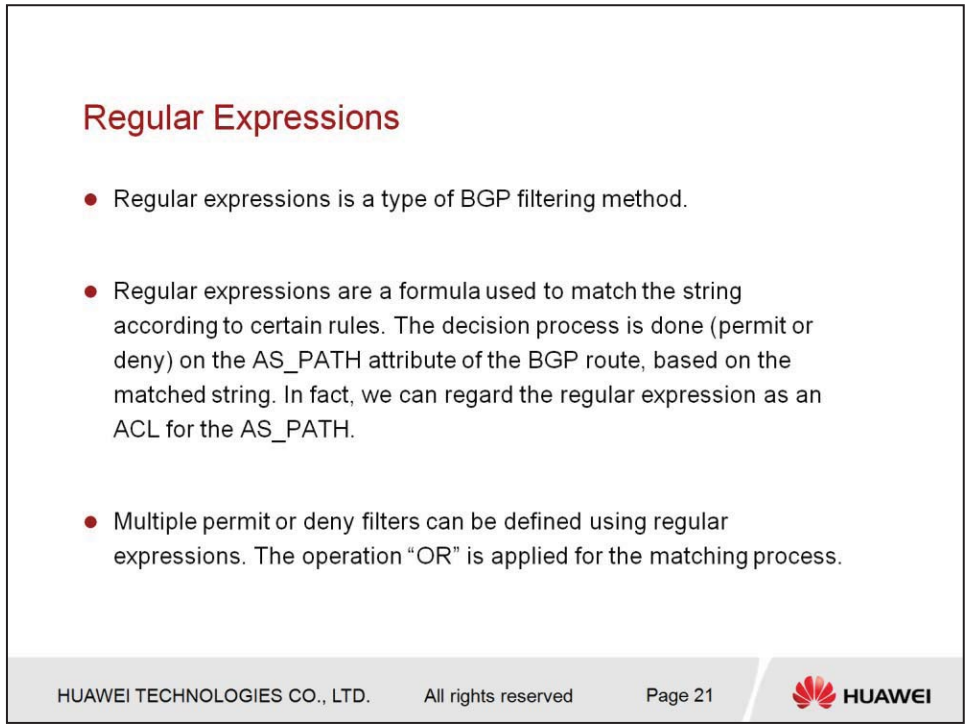
Policy Configuration on RTC

```
#
bgp 100
 peer 10.4.4.1 as-number 200
 peer 1.1.1.1 as-number 100
 peer 5.5.5.5 as-number 100
 #
 ipv4-family unicast
  undo synchronization
  peer 10.4.4.1 enable
  peer 10.4.4.1 route-policy test1 export
  peer 1.1.1.1 enable
  peer 5.5.5.5 enable
#
route-policy test1 permit node 10
 if-match ip-prefix 1
 apply cost 2000
route-policy test1 permit node 20
 apply cost 1000
#
 ip ip-prefix 1 index 10 permit 192.168.1.0 24 greater-equal 24 less-equal 24
#
```

HUAWEI TECHNOLOGIES CO., LTD. All rights reserved



We can define multiple filtering policies (permit or deny) under the same as-pathfilter number. The operation “OR” is applied during the matching process. This indicates that at least one occurrence is required to make the test true. No occurrence indicates that the test failed AS_PATH Filter uses the regular expression method to filter the AS_PATH attribute information.

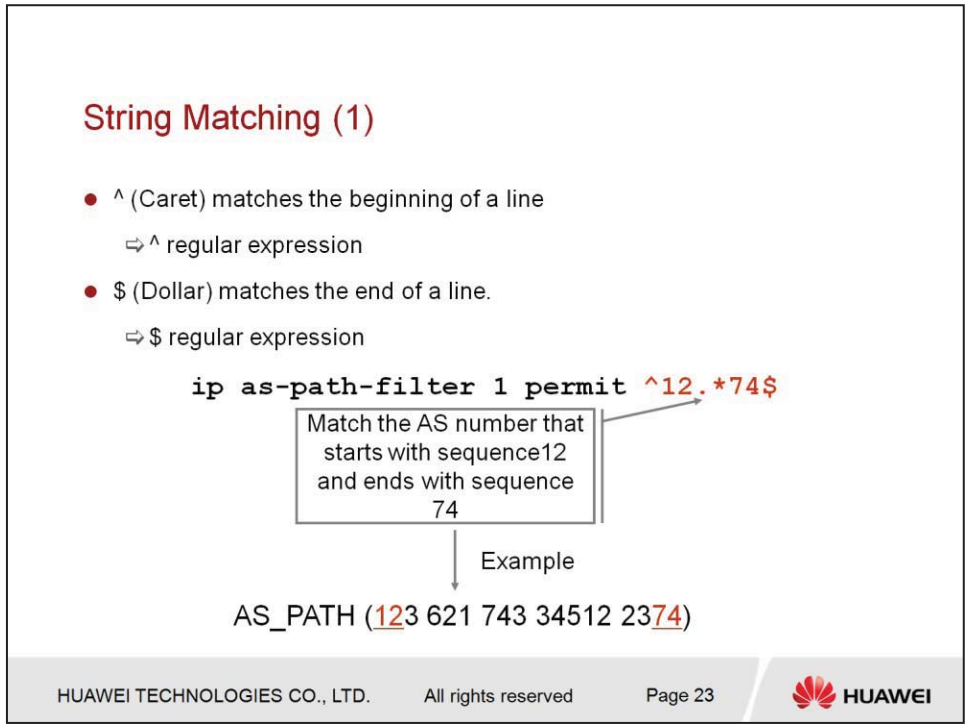


Regular expression is a formula used to match the string according to certain rules. Next, the decision process is done (permit or deny) on the AS_PATH attribute of the BGP route base on the matched string.

Regular Expressions

Characters	Explanation
^	Match the beginning of a string. For example “^200” indicates match only if the first value of the AS_PATH is 200.
$	Match the end of a string. For example “200$” indicates match only if the last value of the AS_PATH is 200.
.	Match any single character, including white space.
+	Matches one or more sequences of the pattern.
_	Match any delimiters. For example, comma, bracket, white space and so on.
*	Match zero or more sequences of the pattern.
()	To group smaller regular expressions into larger regular expressions, ”()” is often used together with “|”.
|	It is an operator that means “or”.
[ ]	Enable you to specify a range of single characters., ”[]” is often used together with “-”.
-	Separates the end points of a range.

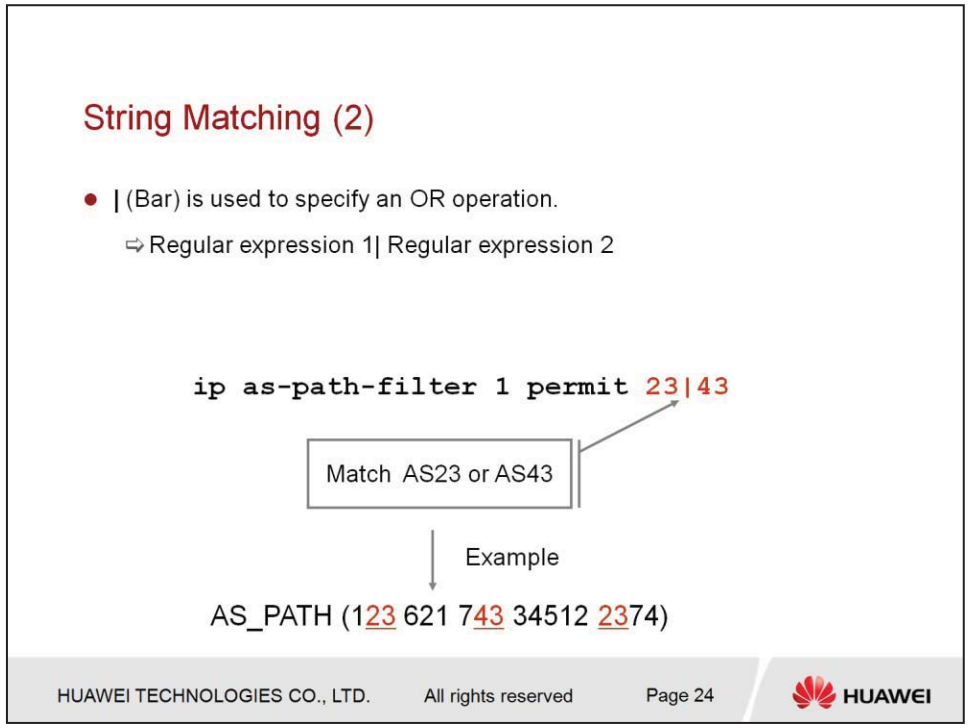
HUAWEI TECHNOLOGIES CO., LTD. All rights reserved



"^" represents start of a line. The regular expression ^12 can match sequence 12 of the first AS number 123 in AS_PATH (123 621 743 34512 2374). In other words, the AS number of the routing information must be start with sequence 12.

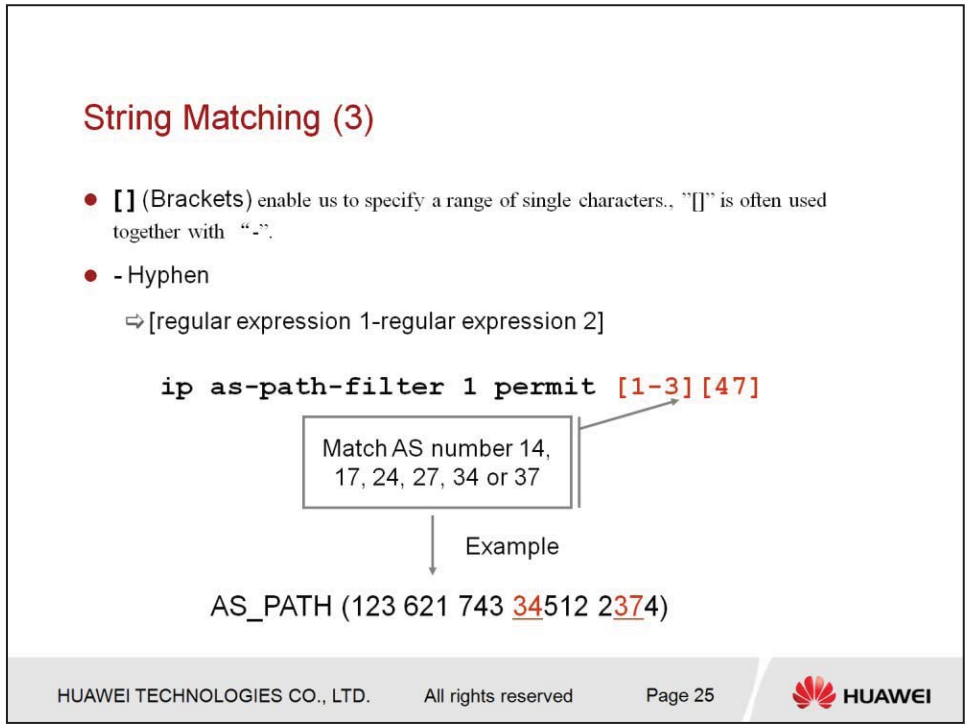
"$" represents the end of a line. The regular expression 74$ can match sequence 74 of the last AS number 2374 in AS_PATH (123 621 743 34512 2374). In other words, the AS number of the routing information must be end with sequence 74.

"^" is used at the front of a group of matching symbols.

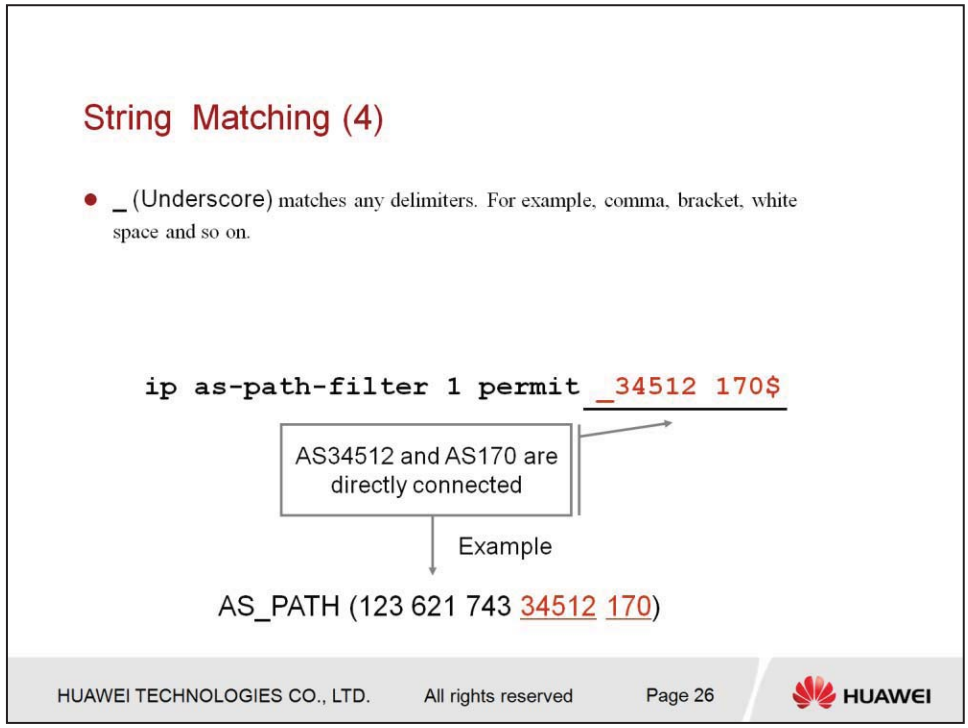


Regular expression 23|43 matches AS23 or AS43.

AS shown in the example above, the regular expression 23|43 matches AS_PATH (123 621 743 34512 2374) for 3 times. Regular expression 23|43 can match some characters of the whole AS number. For example, it matches 43 of the AS743. In addition, it matches the 23 of the AS123 and AS2374.

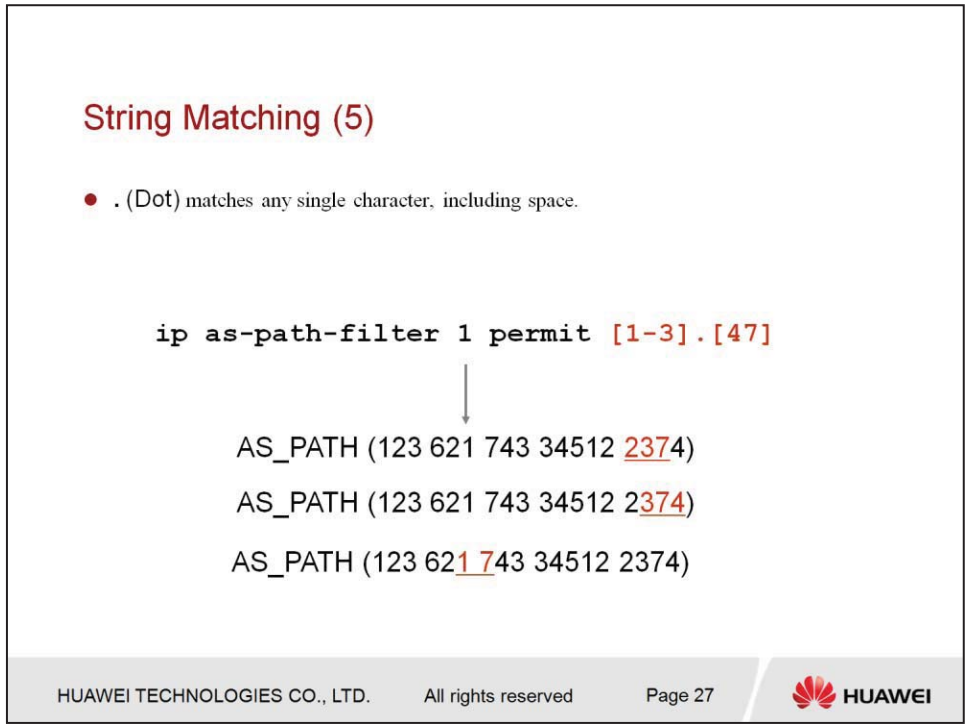


Regular expression [1-3] matches the character "1", "2" or "3". Regular expression [47] matches the character "4" or "7". When used together, regular expression [1- 3][47] matches the AS number of 14, 17, 24, 27, 34 or 37.

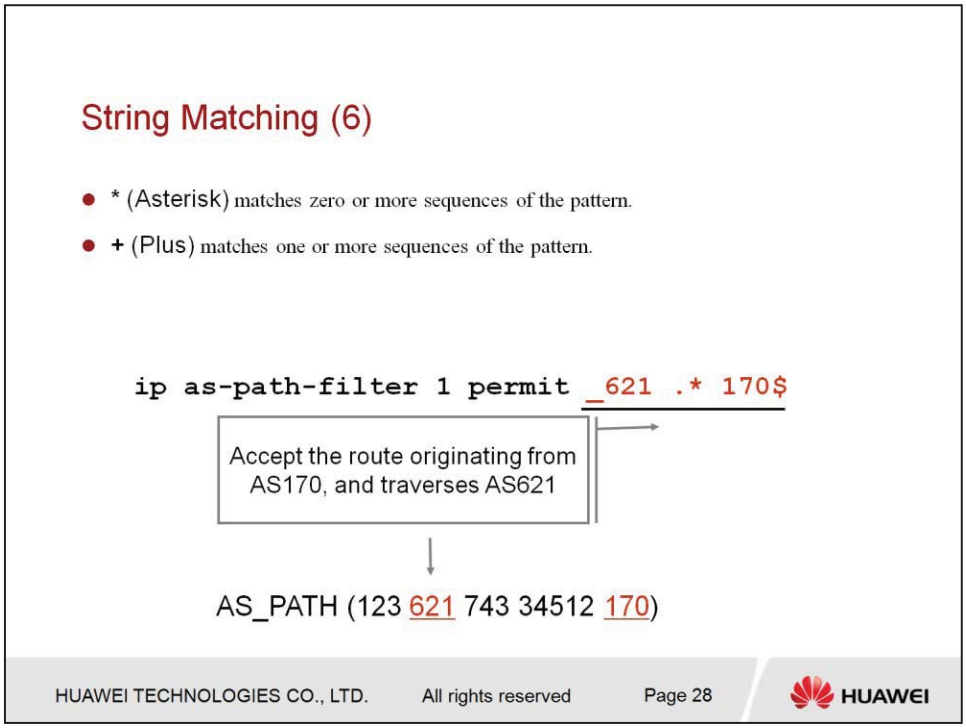


“.” and “_” are different in that “.” can be used to match any single character, including white space while “_” is used to match any delimiters. Example of the delimiters are comma, left parenthesis, right parenthesis, the beginning of theinput string, the end of the input string, space and so on.

"_34512 170$"among AS-PATH(123 621 743 34512 170) indicates that AS34512 and AS170 are indicates connected ."_" indicates any delimiters. The underscore in thisregular expression represents the white space in between the "743 34512".



The "."sign in the middle of the regular expression [1-3].[47] can match any single character, including space. So the AS_PATH (123 621 743 34512 2374) matches the regular expression [1-3].[47] 3 times: “1 7”, “237”and “374”.

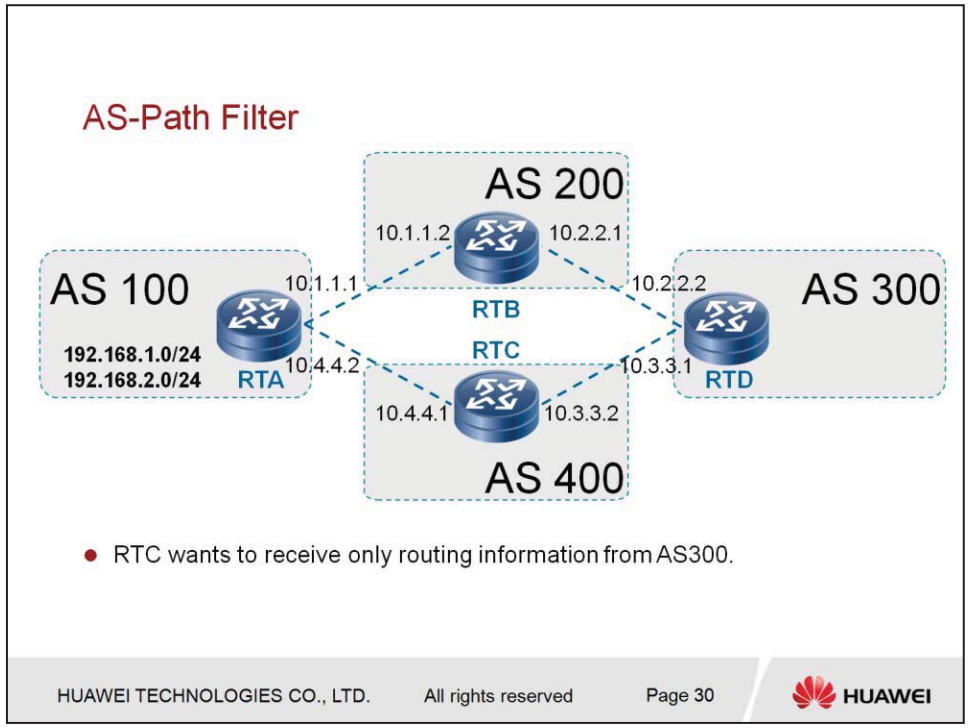


In the diagram above, the most difficult part is the ".+" sign in the middle of the regular expression. As per discussed earlier, "." sign in the regular expression matches any single character including white space while "+" sign matches one or more sequences of the pattern. Therefore, ".+" sign matches multiple single characters including white space or sequence of the pattern. In the example above, the ".+" sign matches " 743 34512 " of the AS_PATH (123 621 743 34512 170).

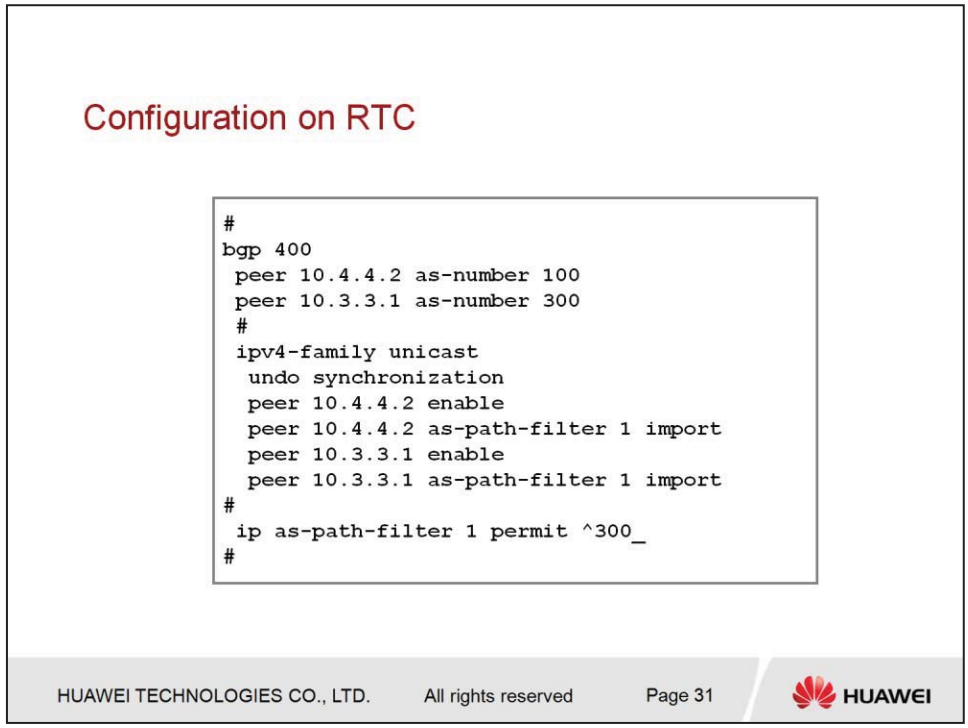
Common Used Regular Expressions

Regular Expressions	Meaning
^$	Match the routes originated in local AS
.*	Match all routes
_10_	Match the routes traversing AS10
^10$	Match the routes with only AS10 in AS_PATH
^10_	Match the routes received from AS10 directly
^[0-9]+$	Match the routes with only one AS in AS_PATH

HUAWEI TECHNOLOGIES CO., LTD. All rights reserved



Base on the requirement specified in the diagram above, we need to configure the filtering policies on RTC. These filtering policies are implemented to the routes coming from the EBGP peer RTA and EBGP peer RTD.



Explanation of the configuration:

#

bgp 400

peer 10.4.4.2 as-number 100

peer 10.3.3.1 as-number 300

#

ipv4-family unicast

undo synchronization

peer 10.4.4.2 enable

peer 10.4.4.2 as-path-filter 1 import \\apply as-path-filter to the route coming from the peer RTA

peer 10.3.3.1 enable

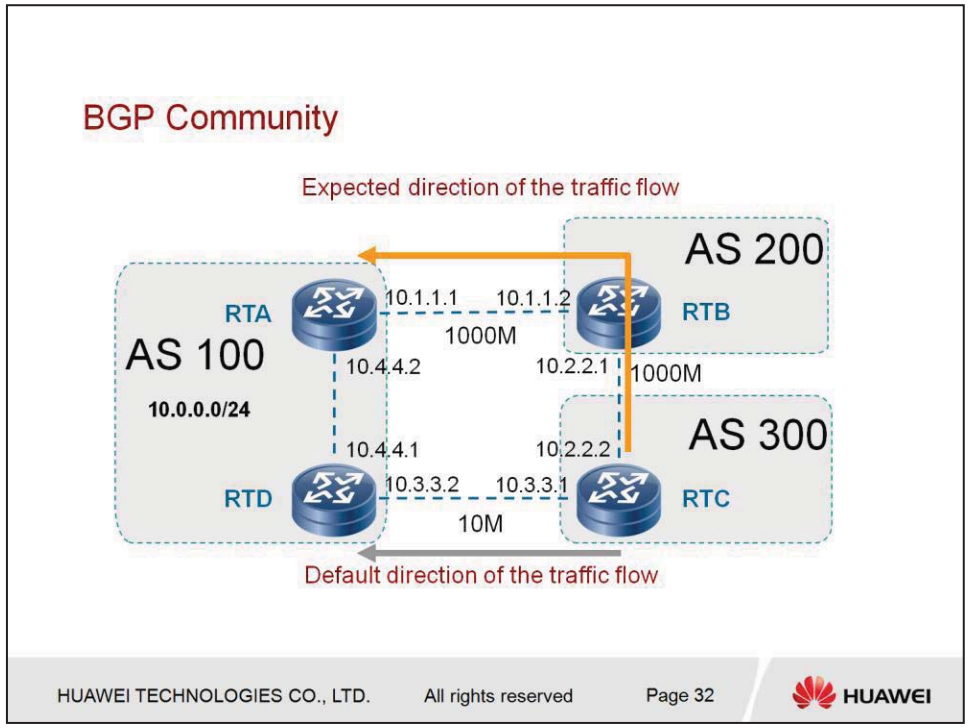
peer 10.3.3.1 as-path-filter 1 import \\apply as-path-filter to the route

coming from the peer RTD

#

ip as-path-filter 1 permit ^300_

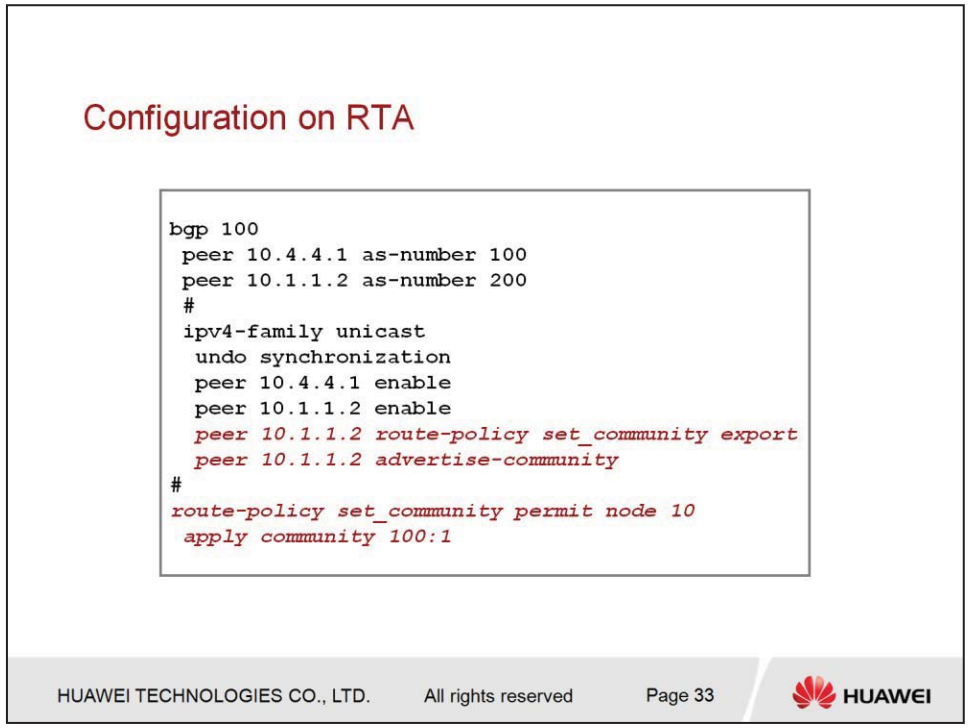
\\receive only the routing information from AS300



In the example above, ASBR RTA and ASBR RTD in AS100 announce the network 10.0.0.0/24 to all other nodes in the network.

As a result, both RTB and RTC have 2 different paths to reach AS100. By default, the routers will select the best path to reach AS100. For example, RTC will select the path between RTC and RTD to reach the network 10.0.0.0/24. From the diagram above, the bandwidth of the link between RTC and RTD is 10M only. In contrast, the bandwidth of the links between RTA-RTB and RTB-RTC are 1000M.

Therefore, we would like to direct the traffic to reach the network 10.0.0.0/24 by using the path via RTC-RTB-RTA.



Explanation of the configuration:

bgp 100

peer 10.4.4.1 as-number 100

peer 10.1.1.2 as-number 200

#

ipv4-family unicast

undo synchronization

peer 10.4.4.1 enable

peer 10.1.1.2 enable

peer 10.1.1.2 route-policy set community export

\\apply routing policy “set community” to the route advertise to the peer.

peer 10.1.1.2 advertise-community

\\advertise community attributes to a peer or peer group. By default, no community attribute is advertised to any peer or peer group.

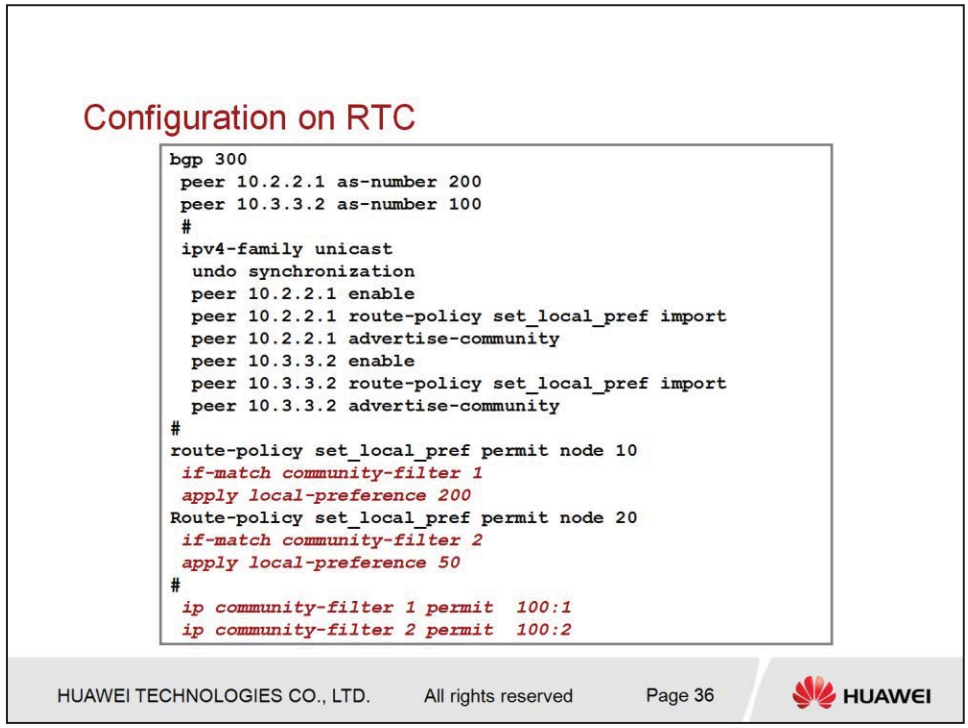
#

route-policy set community permit node 10
apply community 100:1
\\apply community attribute as 100:1

Configuration on RTD

```
bgp 100
 peer 10.4.4.2 as-number 100
 peer 10.3.3.1 as-number 300
 #
 ipv4-family unicast
  undo synchronization
  peer 10.4.4.2 enable
  peer 10.3.3.1 enable
  peer 10.3.3.1 route-policy set_community export
  peer 10.3.3.1 advertise-community
#
route-policy set_community permit node 10
 apply community 100:2
```

HUAWEI TECHNOLOGIES CO., LTD. All rights reserved



Explanation on the configuration:

#

ip community-filter 1 permit 100:1 \\configure the filtering list for community attribute

ip community-filter 2 permit 100:2

#

ip community-filter basic-comm-filter-num { deny | permit } [communitynumber| aa:nn] * &<1-16> [internet | no-export-subconfed | no-advertise | no-export]

ip community-filter adv-comm-filter-num { deny | permit } regular-expression In the basic community attribute list, we can specify only the community number or well-known community attributes.

In the advanced community attribute list, we can specify the regular expression that is used for string matching.

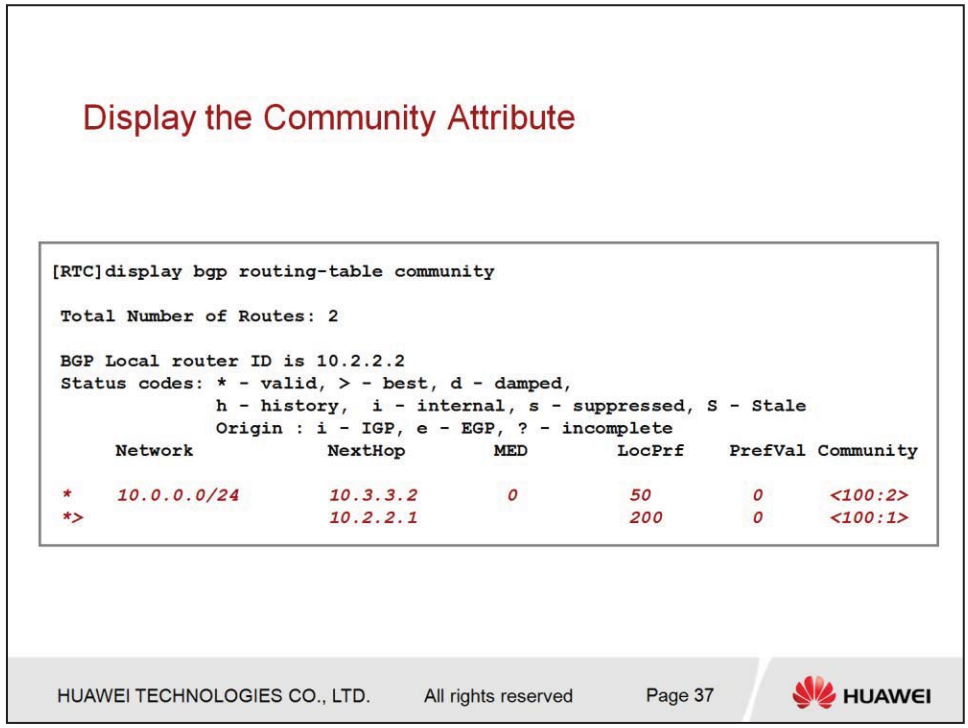
Example:

The basic community list with list number 1

[Quidway] ip community-filter 1 permit internet

The advanced community attribute list with list number

[Quidway] ip community-filter 100 permit ^10



display bgp routing-table community [aa:nn &<1-13>] [no-advertise | noexport | no-export-subconfed][whole-match]

community：display the routing information that belong to a specified BGP community.

aa:nn：specify the community number.

no-advertise ： specify the BGP routes that carry no-advertise community attribute. Routes received carrying this value

cannot be advertised at all, to either EBGP or IBGP peers.

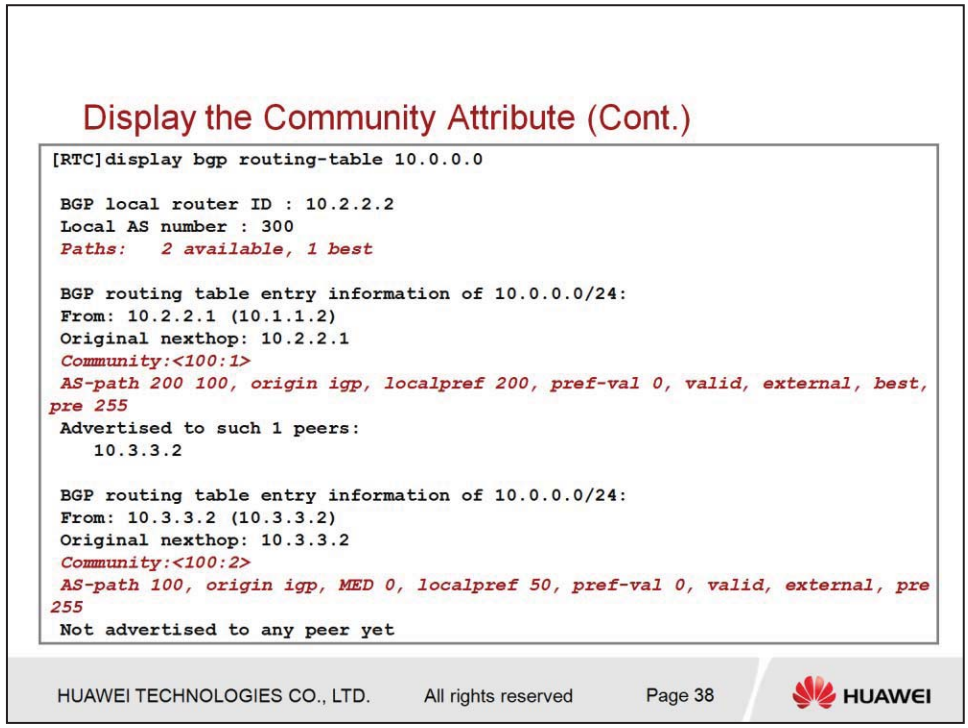
no-export：specify the BGP routes that carry no-export community attribute.

Routes received carrying this value cannot be advertised to EBGP peers or, if a confederation is configured, the routes

cannot be advertised outside of the confederation.

no-export-subconfed：specify the BGP routes that carry no-export-subconfed community attribute. All routes received carrying a communities attribute containing this value MUST NOT be advertised to external BGP peers (this includes peers in other members autonomous systems inside a BGP

confederation). whole-match：indicates the exact matching.

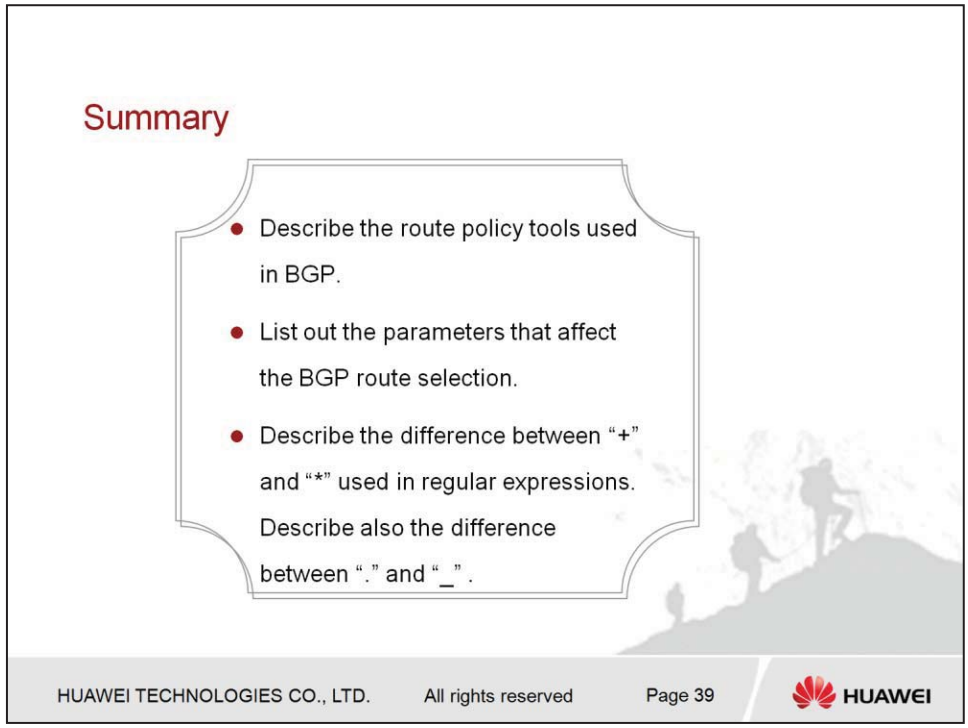


display bgp routing-table [ipv4-address] [{ mask | mask-length } [longerprefixes]]

ipv4-address：display the network address in IPv4 format.

mask/mask-length：specify the subnet mask in dotted decimal format/the length of the subnet mask

longer-prefixes：allow for the longer prefix match.



1. Describe the route policy tools used in BGP.

The route policy tools used in BGP are mainly consist of ACL, IP-PREFIX list, Filter-List, route policy. In addition, the policy tools AS-PATH-FILTER and COMMUNITY-FILTER are solely designed for BGP.

2. List out the parameters that affect the BGP route selection.

The important parameters that affect the BGP route selection are as follow:

Preferred Value

Local-Preference

AS-Path

Origin

MED

EBGP/IBGP

IGP Cost

CLUSTER ID

Communities

Among them, the commonly used parameters are Preferred

Value，Local-Pref，

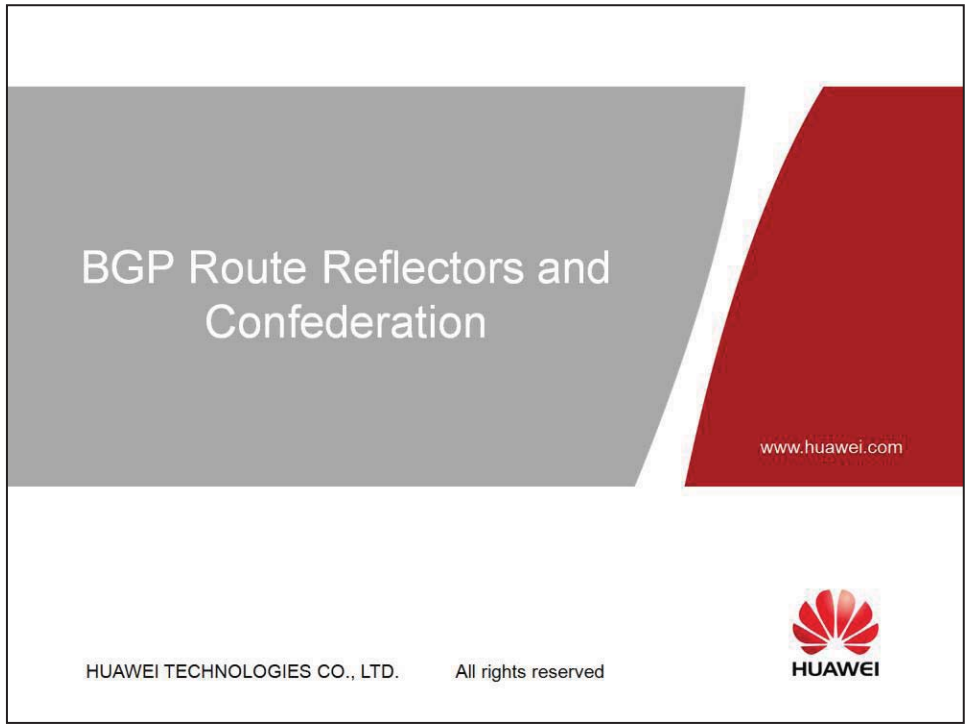
AS_PATH，MED and Community.

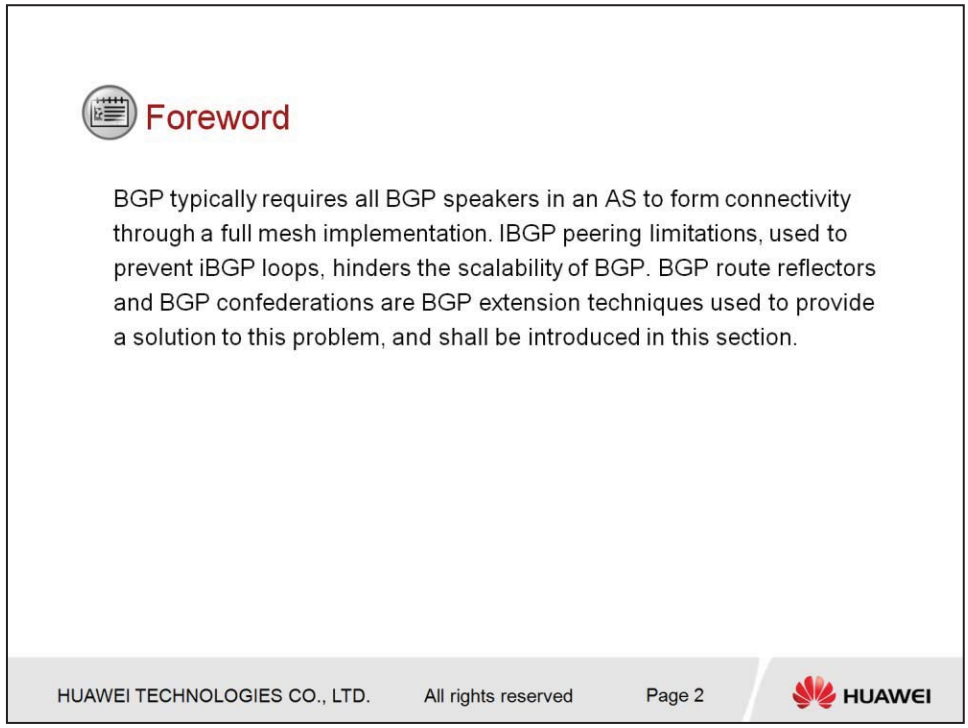
3. Describe the difference between "+" and"*" used in regular expressions.

Describe also the difference between"." and "_" .

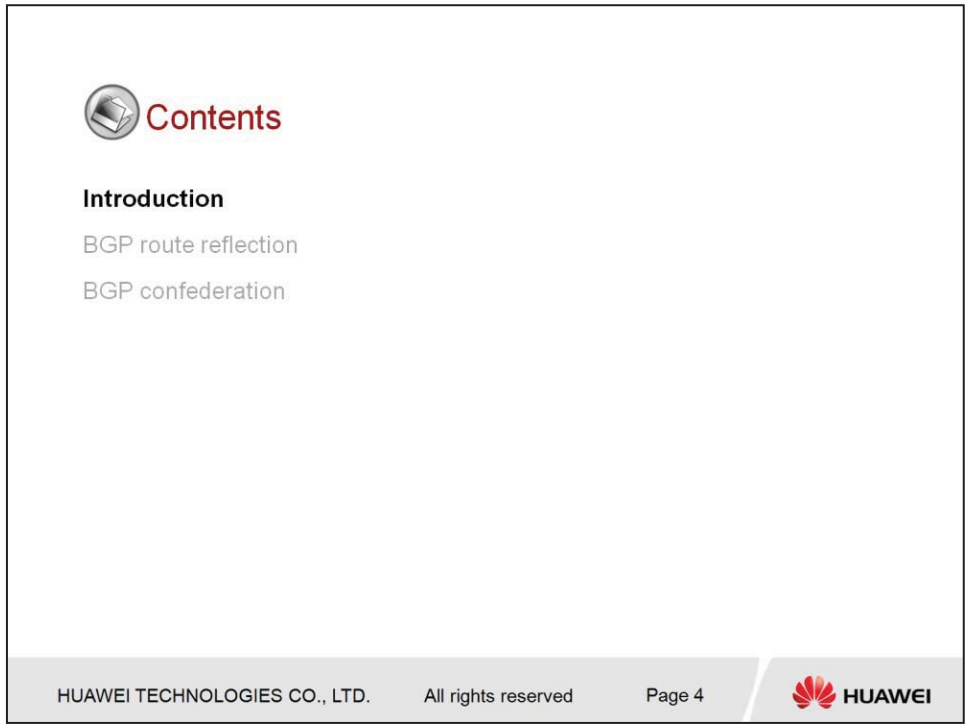
"+" matches one or more sequences of the pattern whereas "*" matches zero or more sequences of the pattern.

"." matches any single character, including white space while"_" matches any delimiters like comma, bracket, white space and so on.


BGP Route Reflectors and Confederation
www.huawei.com
HUAWEI TECHNOLOGIES CO., LTD.
All rights reserved
HUAWEI


Foreword
BGP typically requires all BGP speakers in an AS to form connectivity through a full mesh implementation. IBGP peering limitations, used to prevent iBGP loops, hinders the scalability of BGP. BGP route reflectors and BGP confederations are BGP extension techniques used to provide a solution to this problem, and shall be introduced in this section.
HUAWEI TECHNOLOGIES CO., LTD.
All rights reserved
Page 2
HUAWEI


Objectives
Upon completion of this section, you will be able to:
Understand BGP route reflection principles
Understand BGP confederation principles
HUAWEI TECHNOLOGIES CO., LTD.
All rights reserved
Page 3
HUAWEI


Contents
Introduction
BGP route reflection
BGP confederation
HUAWEI TECHNOLOGIES CO., LTD.
All rights reserved
Page 4
HUAWEI

The Extensibility Problem in IBGP

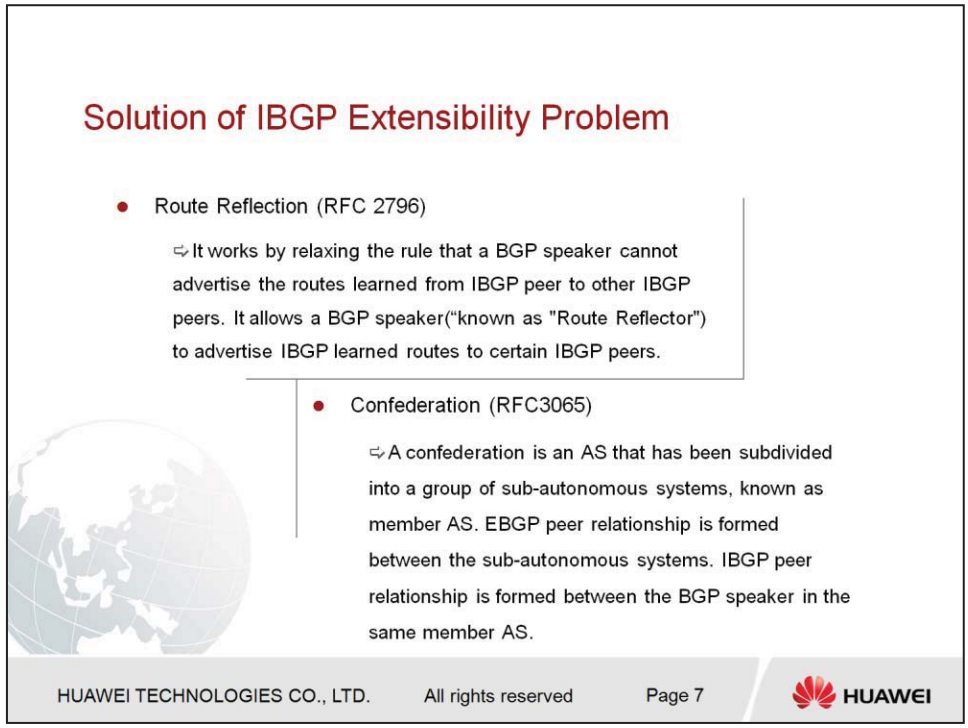
- How does BGP prevent the routing loops?
  - ⇨ EBGP
    - − EBGP prevents the routing loop through AS-PATH attribute. Router that receives an update containing its local AS number in the AS_PATH attribute from EBGP peer knows that a routing loop has occurred. To prevent routing loops, the router will ignore the update.
  - ⇨ IBGP
    - − BGP router will not announce the update information received from its IBGP peer to other IBGP peers.

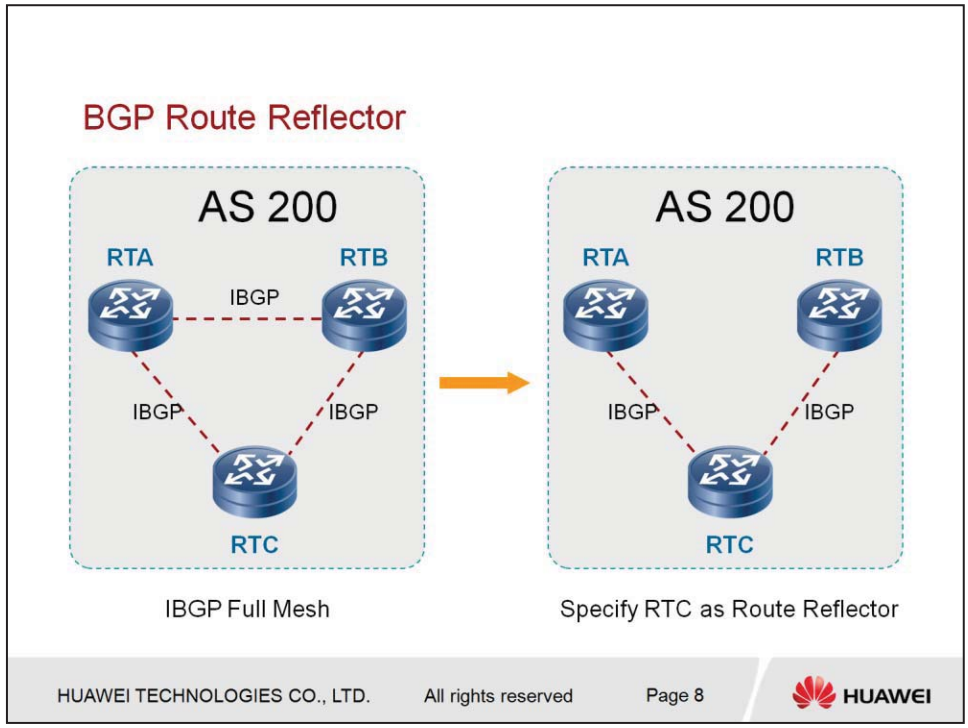
HUAWEI TECHNOLOGIES CO., LTD. All rights reserved

The Extensibility Problem in IBGP

- The loop prevention mechanism in IBGP introduces the problem below
  - ⇨ To ensure all the IBGP peers can receive the update information
    - − IBGP Speaker must be fully meshed
      - ▪ requires to maintain n*(n-1)/2 unique IBGP sessions
- Solution
  - ⇨ Route Reflection (RFC2796)
  - ⇨ Confederation (RFC3065)

HUAWEI TECHNOLOGIES CO., LTD. All rights reserved

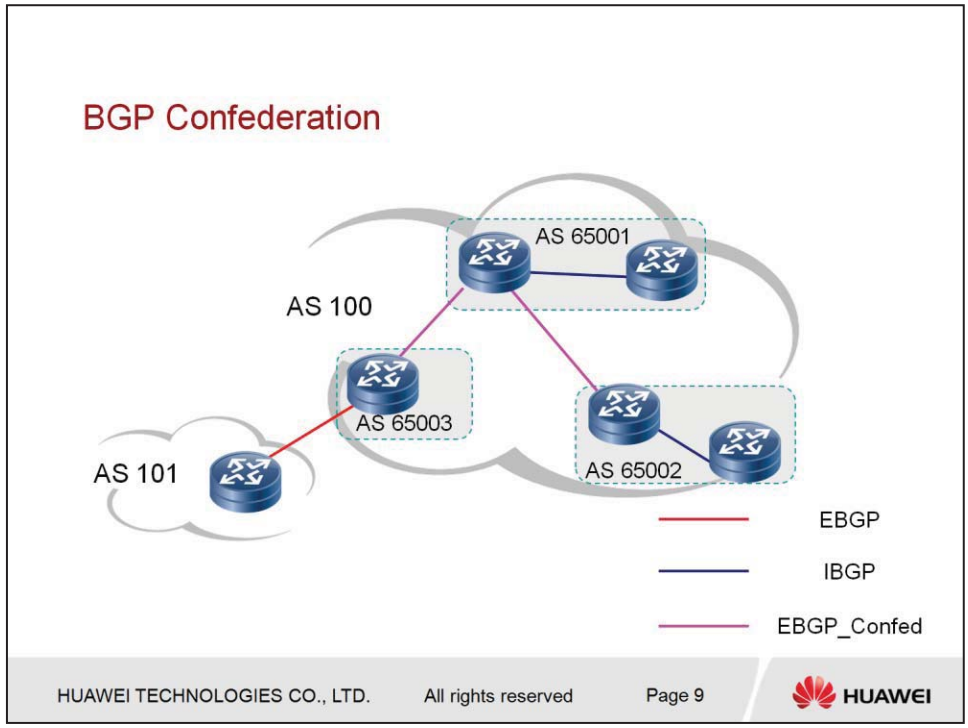

Solution of IBGP Extensibility Problem
Route Reflection (RFC 2796)
It works by relaxing the rule that a BGP speaker cannot advertise the routes learned from IBGP peer to other IBGP peers. It allows a BGP speaker("known as "Route Reflector") to advertise IBGP learned routes to certain IBGP peers.
Confederation (RFC3065)
A confederation is an AS that has been subdivided into a group of sub-autonomous systems, known as member AS. EBGP peer relationship is formed between the sub-autonomous systems. IBGP peer relationship is formed between the BGP speaker in the same member AS.
HUAWEI TECHNOLOGIES CO., LTD. All rights reserved Page 7
HUAWEI



There are 3 routers in AS200 namely RTA, RTB and RTC. In the diagram above, assume that RTA receives an update from an external peer and this update has been selected as the best route by RTA. By default, RTA will forward the update to its two internal peers, RTB and RTC. Since RTB and RTC form the IBGP peer relationship, they will not advertise the route learnt from their respective IBGP peer to other IBGP peers.

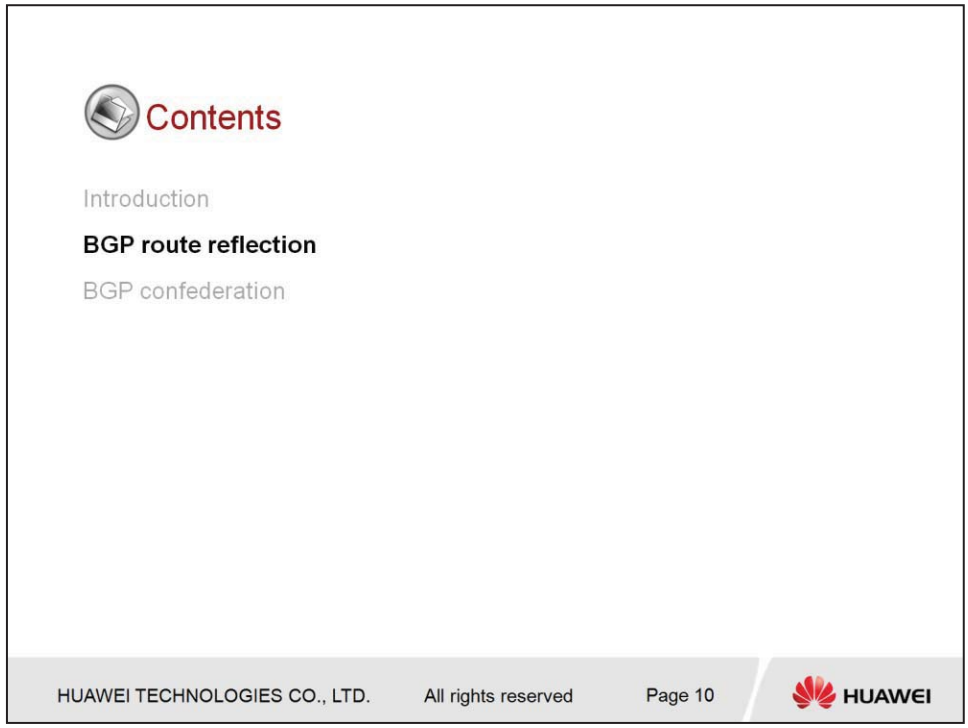
Next, assume that RTC has been configured as the route reflector. It works by relaxing the rule that a BGP speaker cannot advertise routes learned from IBGP peer to other IBGP peers. After the configuration, RTC is allowed to advertise the

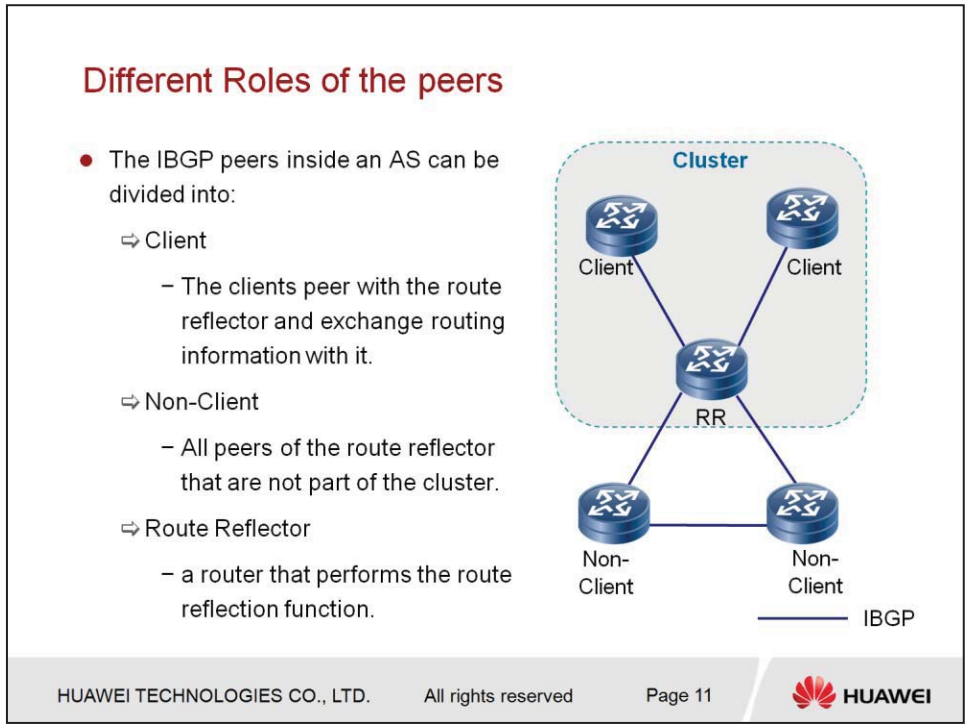
route learnt from RTA to other IBGP peers. As a result, the IBGP session between RTA and RTB can be cancelled.



Confederation controls the large numbers of IBGP peers by sub-dividing an autonomous system into a group of sub-autonomous systems, called member autonomous systems. Since the EBGP session is formed between the subautonomous systems, no full mesh connection is required between them.

However, IBGP full mesh is required between the BGP speakers within a sub-autonomous system.


Contents
Introduction
BGP route reflection
BGP confederation
HUAWEI TECHNOLOGIES CO., LTD.
All rights reserved
Page 10
HUAWEI



We use the term "Route Reflection" to describe the operation of a BGP speaker advertising an IBGP learned route to another IBGP peer. Such a BGP speaker is said to be a "Route Reflector" (RR), and such a route is said to be a reflected route.

The internal peers of a RR are divided into two groups:

1) Client peers

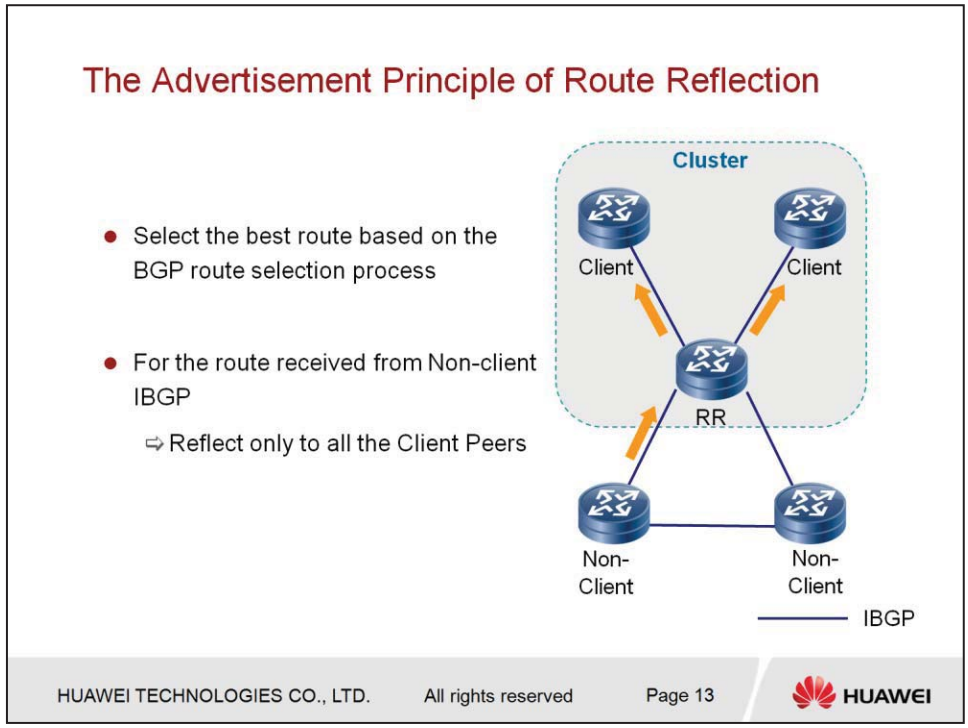
2) Non-Client peers

A RR reflects routes between these groups, and may reflect routes among client peers. A RR along with its client peers form a Cluster. We will discuss the concept of cluster in detail in other slides. The Non-Client peer must be fully meshed but the Client peers need not be fully meshed. Diagram above depicts a simple example outlining the basic RR components using the terminology mentioned here.

The Relationship between Peers

- Client needs to maintain only the IBGP session between the client and RR
- IBGP full mesh is required between the RRs
- IBGP full mesh is required between the Non-clients
- IBGP full mesh is required between the RR and Non-Client

HUAWEI TECHNOLOGIES CO., LTD. All rights reserved



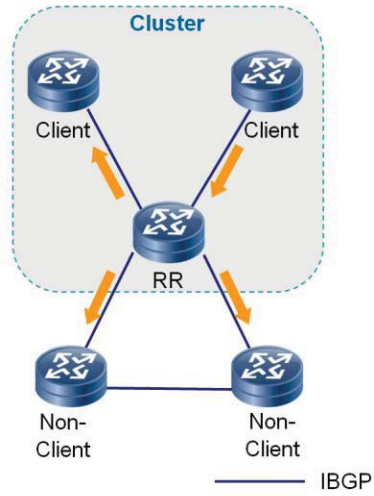
When a RR receives a route from an IBGP peer, it selects the best route based on the BGP route selection process. After the best route is selected, it must do the following depending on the type of the peer it is receiving the best route from:

1) A Route from a Non-Client IBGP peer Reflect to all the Clients.

2) A Route from a Client peer Reflect to all the Non-Client peers and also to the Client peers. (Hence the Client peers are not required to be fully meshed.)

The Advertisement Principle of Route Reflection (Cont.)

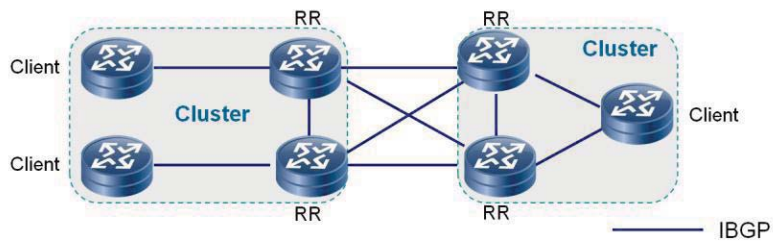
- For the route learnt from Client IBGP
  - ⇨ Reflect to all Clients and Non-Clients



HUAWEI TECHNOLOGIES CO., LTD. All rights reserved

Route Reflection Cluster

- Inside an AS, multiple RRs might exist to provide redundancy for the Clients. As a result, the routing update between the RRs might generate a routing loop. We use the concept of cluster to prevent this from happen.

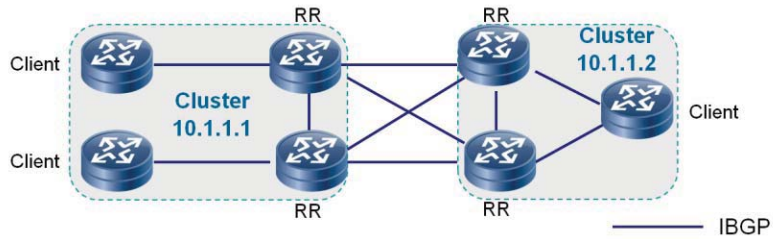


HUAWEI TECHNOLOGIES CO., LTD. All rights reserved

Inside an AS, multiple RRs might exist to provide redundancy for the Clients. As a result, the routing update between the RRs might generate a routing loop. We use the concept of cluster to prevent this from happen.

What is a Cluster?

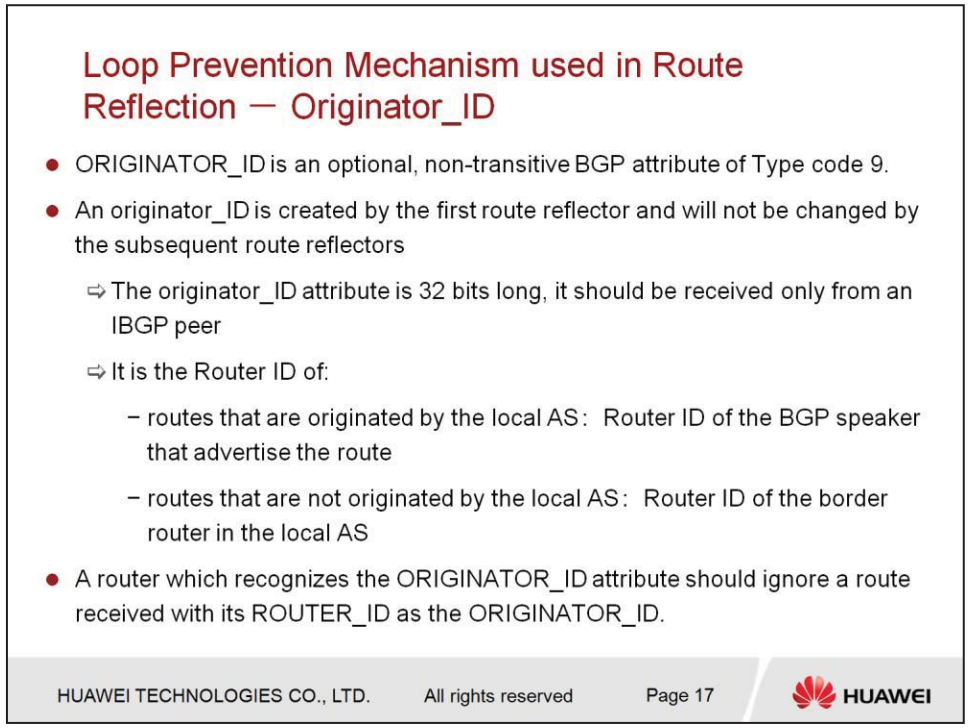
- A cluster is identified by the 4 bytes Cluster_ID. A loopback address is usually used as the Cluster_ID.
- It is possible to have multiple RRs in the same cluster; one client can belong to multiple clusters at the same time
- If the route reflector receives a routing update that contains the local CLUSTER_ID value, the update message will be discarded.



HUAWEI TECHNOLOGIES CO., LTD. All rights reserved

Usually a cluster of clients will have a single RR. In that case, the cluster will be identified by the BGP ROUTER-ID of the RR. However, this represents a single point of failure so to make it possible to have multiple RRs in the same cluster, all

RRs in the same cluster must be manually configured with a 4-byte CLUSTER_ID so that an RR can discard routes from other RRs in the same cluster.



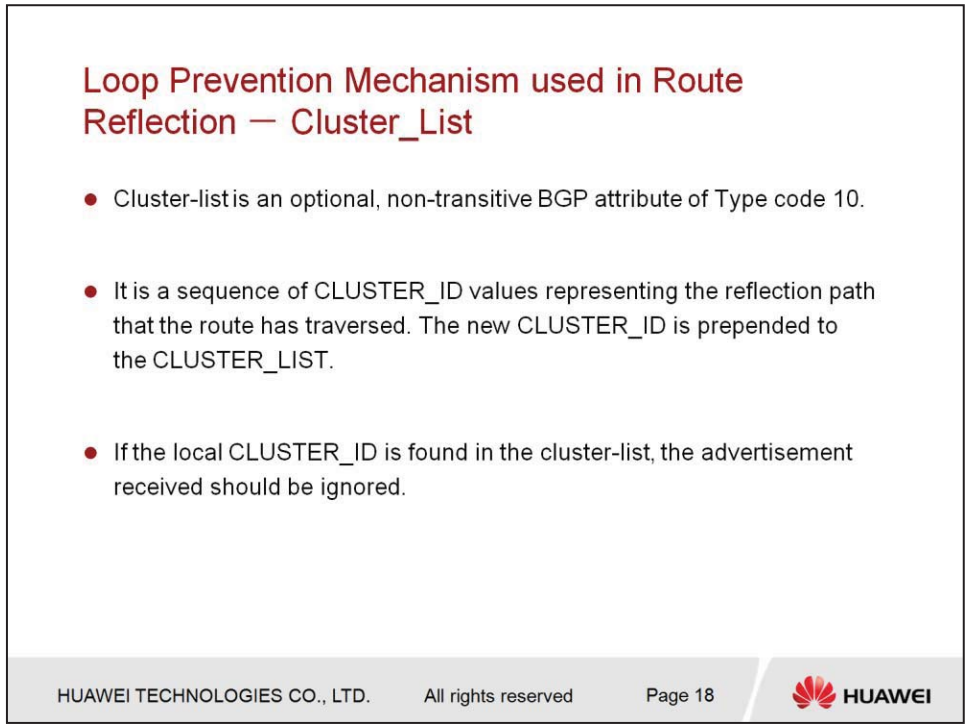
When a route is reflected, routing loop might be generated due to the misconfiguration.

So, the Route Reflection method defines the following attributes to detect and avoid the looping of routing information:

1) ORIGINATOR_ID

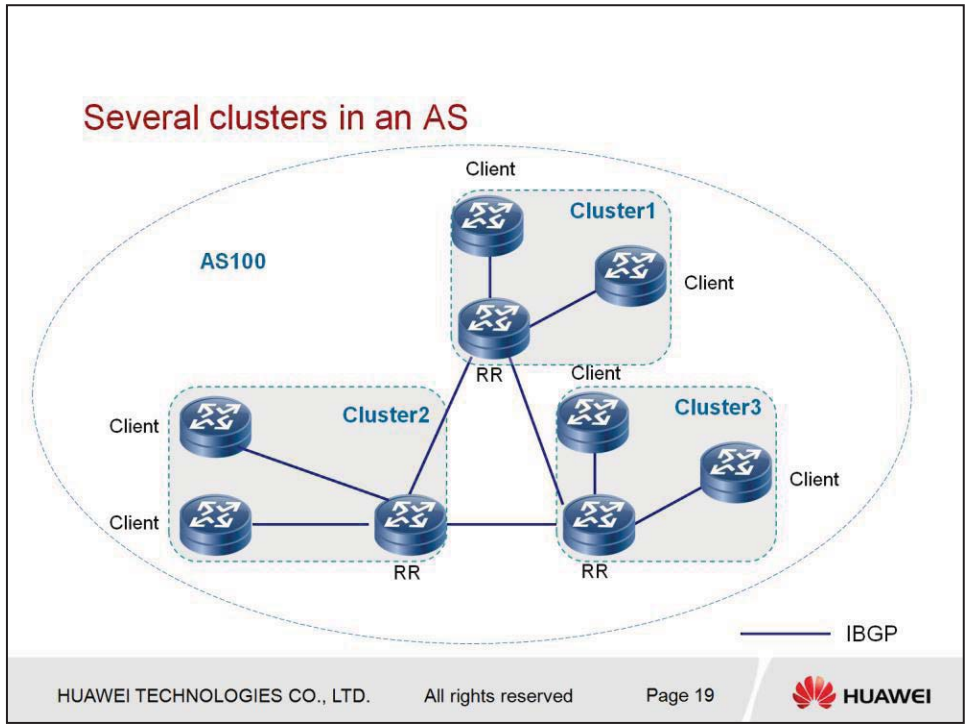
ORIGINATOR_ID is an optional, non-transitive BGP attribute of Type code 9. This attribute is 32 bits long and it will be created by a RR in reflecting a route.

Normally, this attribute will carry the ROUTER_ID of the originator of the route in the local AS. Once the ORIGINATOR_ID has been created, the subsequent BGP speakers are not allowed to add, delete or modify it. When a BGP speaker receives an update that carry the ORIGINATOR_ID attribute, the BGP speaker will match it with its local ROUTER_ID. The update will be discarded if match.



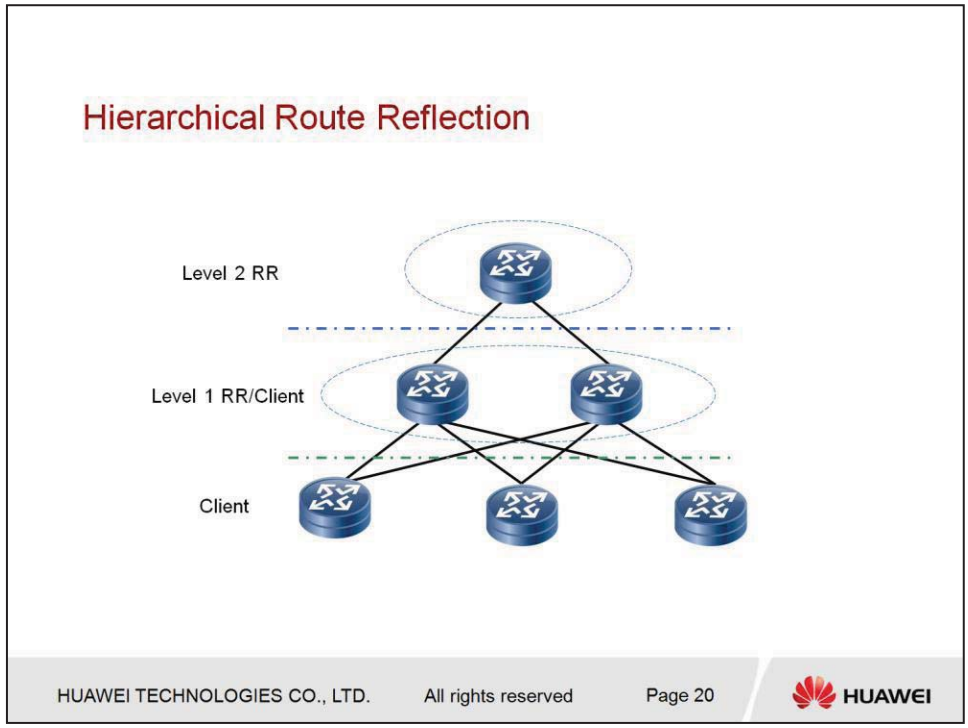
2) CLUSTER_LIST

Cluster-list is an optional, non-transitive BGP attribute of Type code 10. It is a sequence of CLUSTER_ID values representing the reflection path that the route has passed. When a RR reflects a route, it must prepend the local CLUSTER_ID to the CLUSTER_LIST. If the CLUSTER_LIST is empty, it must create a new one. Using this attribute an RR can identify if the routing information is looped back to the same cluster due to mis- configuration. If the local CLUSTER_ID is found in the cluster-list, the advertisement received should be ignored.



The AS may be has many Clusters. each RR is IBGP relationship, one RR can put another RR configuration as his own client or the non-client. So it can be flexible configuration the relationship between the cluster in AS.

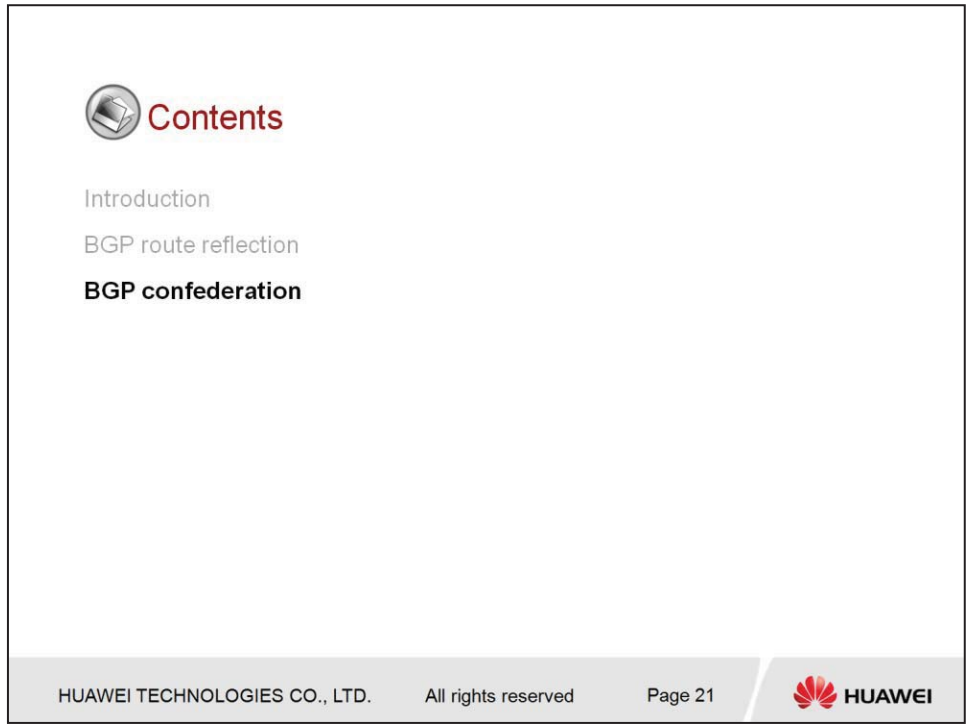
As shown in figure, one AS divided into several reflection cluster, each RR configured other RR to the non-client, all the RR established between each other. Each client only establish IBGP connection with the RR in the same cluster. So all the BGP routers in AS will receive reflection routing information.

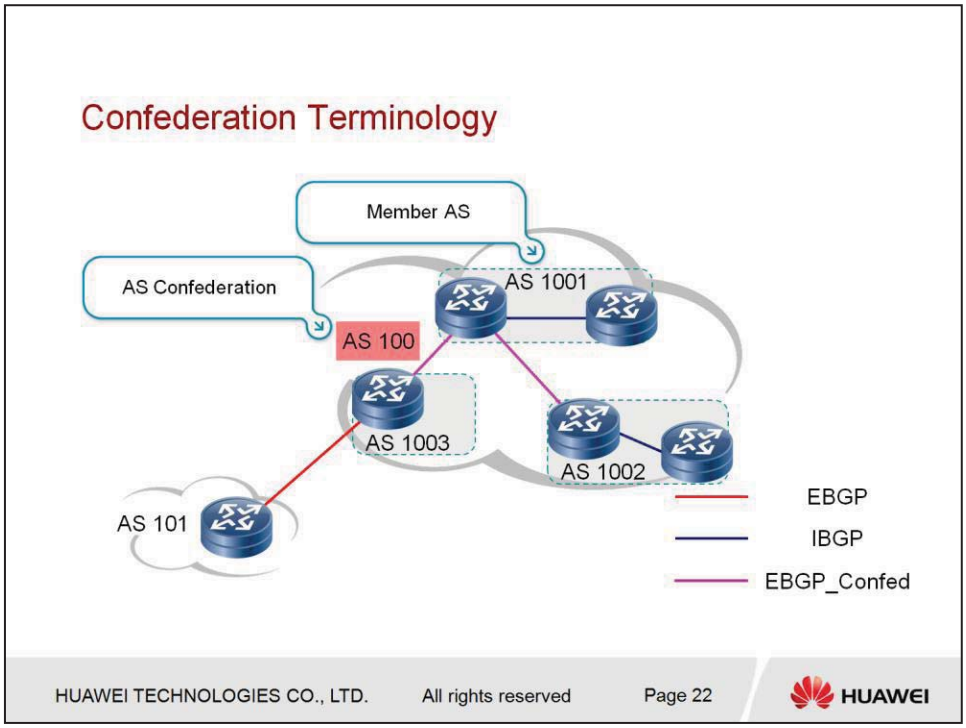


Route reflection can effectively reduce the total number. However, RRs required to be fully meshed in some circumstances. As a result, RRs require to maintain a

large number of IBGP session especially in a large network. Therefore, the hierarchical route reflection is introduced to further reduce the number of IBGP sessions.

Depend on the network requirement, the number of levels in the hierarchical route reflection can be increased gradually. Normally, 2 levels hierarchy or 3 levels hierarchy is sufficient for current network deplayment.


Contents
Introduction
BGP route reflection
BGP confederation
HUAWEI TECHNOLOGIES CO., LTD.
All rights reserved
Page 21
HUAWEI



To deal with the IBGP full mesh problem within an AS, confederation is used to subdivide an AS into a group of sub-autonomous systems, known as member autonomous systems. The BGP speakers within the confederation speak EBGP

to peers in other sub-autonomous systems. Therefore, full mesh connection is not require between them. However, BGP speakers within the confederation speak IBGP to peers in the same sub-autonomous system and full meshed IBGP

connection is required for all routers inside the sub-autonomous system.

EBGP sessions are formed between the sub-autonomous systems inside a confederation. These EBGP sessions behave differently from the conventional EBGP sessions and are therefore identified as intra-confederation EBGP session

to differentiate them from conventional EBGP sessions. Intra-confederation EBGP sessions, while having EBGP-like properties (for example, updating the AS-PATH attribute when BGP route is propagated), still run inside a real AS and

share some properties with IBGP sessions. Similar to IBGP sessions, LOCAL_PREF, MED and NEXT_HOP are not

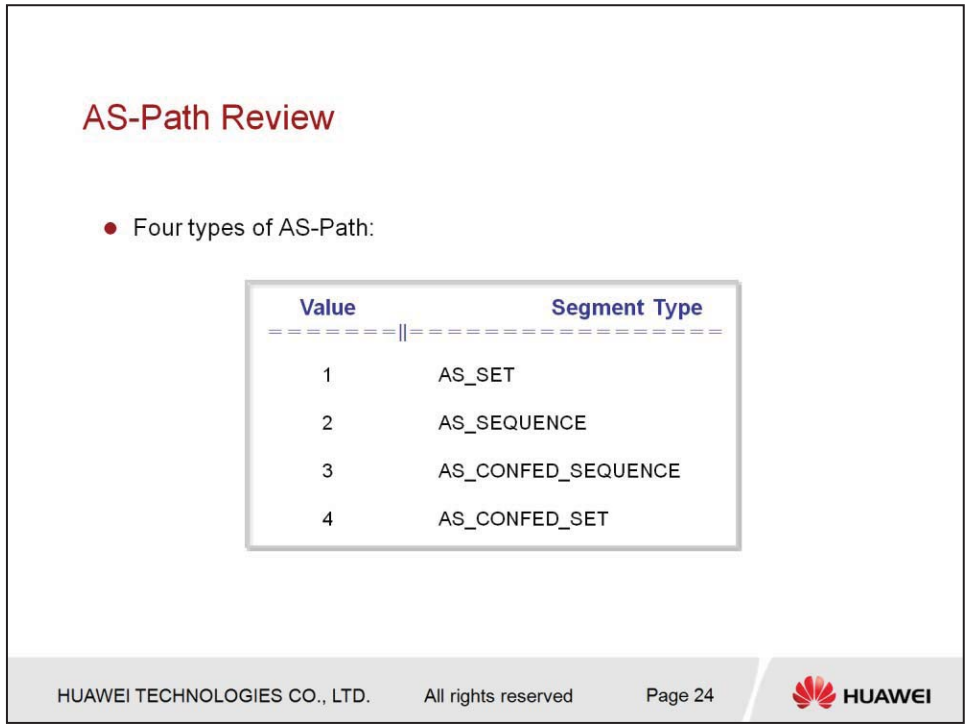
changed in updates propagated across the intra-confederation EBGP sessions.

External peers do not see the internal structure of the confederation. Instead, they will see the whole confederation as a single AS. This means that the AS_PATH information that has been modified inside an confederation is removed when the update information is sent to the conventional EBGP neighbor.

All BGP routers inside the sub-autonomous systems must be fully meshed.

Alternatively, we can implement the route reflection. One of the advantages of implementing confederation is that sub-autonomous systems are not required to use the same IGP. Each of the sub-autonomous systems do not require to

advertise their own internal topology to other sub-autonomous systems. However, when different IGPs are used, each of the sub-autonomous system must ensure the reachability of the next hop of BGP.



Currently, BGP specifies that the AS_PATH attribute is a well-known mandatory attribute that is composed of a sequence of AS path segments. Each AS path segment is represented by a triple <path segment type, path segment length, path segment value>. In BGPv4, the path segment type is a 1-octet long field with the two following values defined:

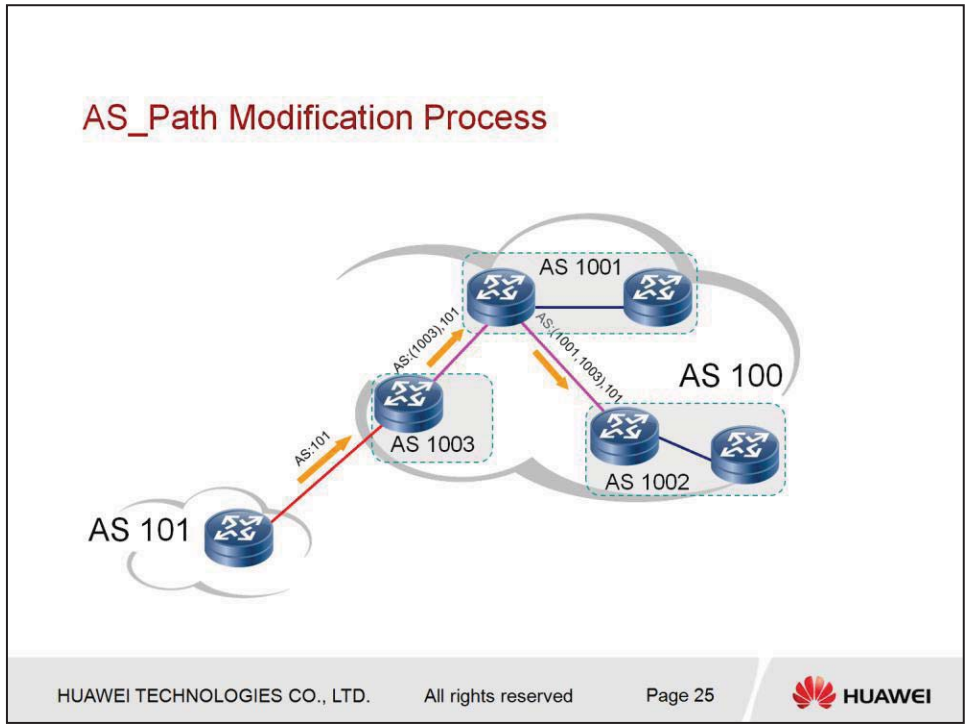
Value Segment Type

1 AS_SET: unordered set of ASs a route in the UPDATE message has traversed

2 AS_SEQUENCE: ordered set of ASs a route in the UPDATE message has traversed

3 AS_CONFED_SEQUENCE: ordered set of Member AS Numbers in the local confederation that the UPDATE message has traversed

4 AS_CONFED_SET: unordered set of Member AS Numbers in the local confederation that the UPDATE message has traversed



Inside a confederation, AS_CONFED is used to prevent the routing loop between the sub-autonomous systems.

The AS_PATH modification process inside the confederation:

1. Intra-confederation EBGP session

• Sub-autonomous system number is added in the leftmost of the AS_PATH as AS_PATH attribute type 4, AS_CONFED_SEQUENCE.

2. IBGP session

• Not modify the AS-PATH attribute

3. EBGP session with external peer

• the sub-autonomous system number is removed from the AS_PATH attribute, and the confederation ID is prepended to the leftmost of the AS_PATH.

When a BGP speaker propagates a route which it has learned from another BGP speaker's UPDATE message, it shall modify the route's AS_PATH attribute based on the location of the BGP speaker to which the route will be sent:

a) When a given BGP speaker advertises the route to another

BGP speaker located in its own autonomous system, the advertising speaker shall not modify the AS_PATH attribute associated with the route.

b) When a given BGP speaker advertises the route to a BGP speaker located in a neighboring autonomous system that is a member of the local autonomous system confederation, then the advertising speaker shall update the AS_PATH attribute as follows:1. if the first path segment of the AS_PATH is of type

AS_CONFED_SEQUENCE, the local system shall prepend its own AS number as the last element of the sequence (put it in the leftmost position).

2. if the first path segment of the AS_PATH is not of type

AS_CONFED_SEQUENCE the local system shall prepend a new path segment of type AS_CONFED_SEQUENCE to the AS_PATH, including its own confederation identifier in that segment.

c)： When a given BGP speaker advertises the route to a BGP speaker located in a neighboring autonomous system that is not a member of the current autonomous system confederation, the advertising speaker shall update the AS_PATH attribute as follows:1. if the first path segment of the AS_PATH is of type

AS_CONFED_SEQUENCE, that segment and any immediately

following segments of the type AS_CONFED_SET or

AS_CONFED_SEQUENCE are removed from the AS_PATH

attribute, leaving the sanitized AS_PATH attribute to be operated on by steps 2, or 3.

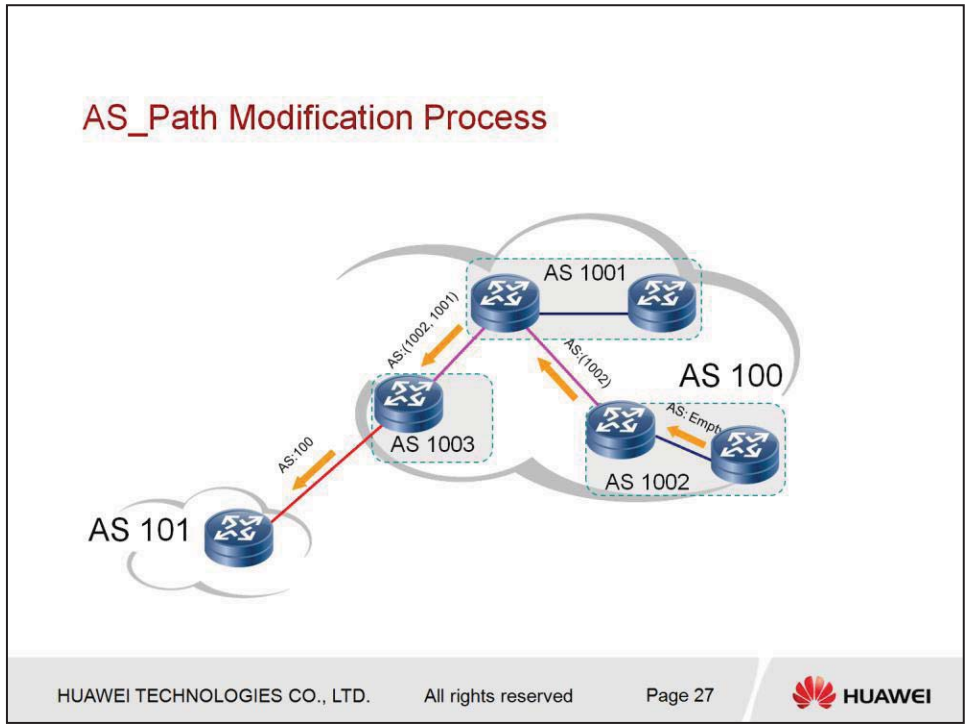
2. if the first path segment of the remaining AS_PATH is of type

AS_SEQUENCE, the local system shall prepend its own

confederation ID as the last element of the sequence (put it in the leftmost position).

3. if there are no path segments following the removal of the first

AS_CONFED_SET/AS_CONFED_SEQUENCE segments, or if the first path segment of the remaining AS_PATH is of type AS_SET the local system shall prepend a new path segment of type AS_SEQUENCE to the AS_PATH, including its own confederation ID in that segment.



When an update is sent to a peer external to the confederation, the AS_CONFED_SEQUENCE and AS_CONFED_SET information is stripped from the AS_PATH attribute, and the confederation ID is pretended to the AS_PATH. Because of this, external peers see the confederation as a single AS rather than as a collection of autonomous systems.

Confederation vs. Route Reflection

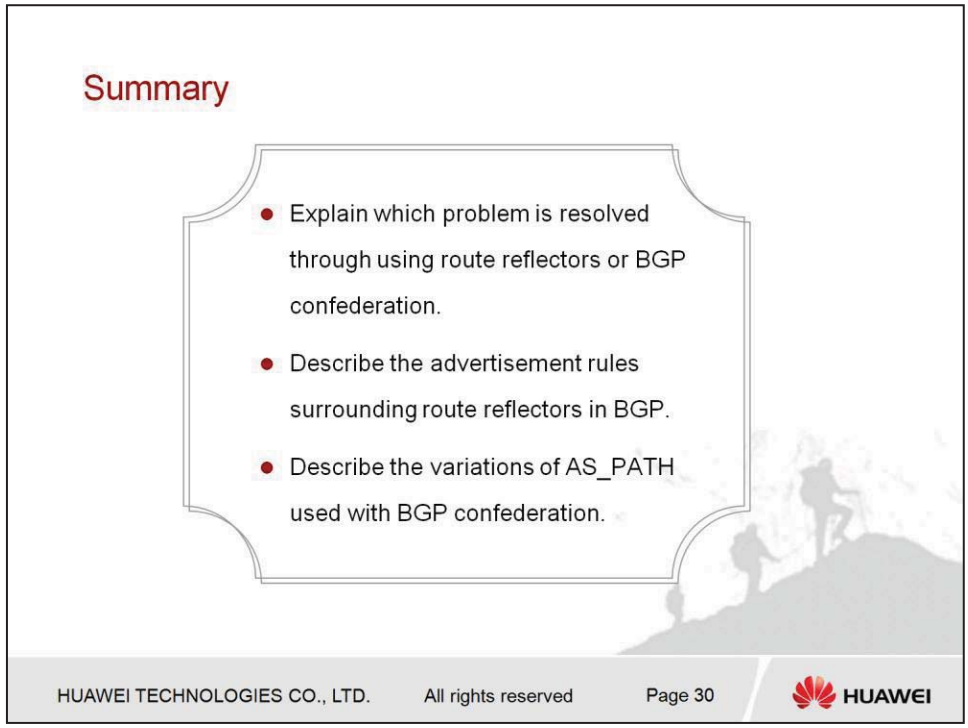
Reference Factor	Comparison
Multi-level Support	Both of the methods can further increase their extensibility by supporting multiple-level function. Route reflector supports multiple level reflection structures. Confederation support the used of route reflection in a sub-autonomous system.
Policy Control	Both of the methods support route selection by using policy control. However, the use of policy control in confederation is much more flexible.
Complexity of Conventional IBGP Migration	Migrating from a non-route reflector to a route reflector design is easy. This is because the changes that happen to the whole network as a result of the migration is less. However, the migration from non-confederation to confederation design requires major reconfiguration of the routers and a major change in the logical topology.
Capability Support	All routers in a confederation must understand and support confederations. This is because all the routers must understand the AS_PATH attribute used for confederation. In contrast, only the route reflectors themselves must understand route reflection. **However, the client is also required to understand the reflector attribute in new design of cluster division.**

HUAWEI TECHNOLOGIES CO., LTD. All rights reserved

Confederation vs. Route Reflection

Reference Factor	Comparison
Control of IGP Expansion	Inside an AS, route reflection requires only a single IGP. On the other hand, confederation could be used to run an IGP in one sub-AS independently of IGPs in other sub-AS, and provides a possible advantage of using confederation over route refection. When an IGP grows exponentially, management can become difficult, the confederation can be used to reduce the size of the IGP routing table.
Implementation Experience	Most of the service providers have deployed route reflection rather than confederation. Therefore, a lot of experience in route reflection has been obtained.
AS Combination	In fact, AS combination is irrelevant with the IBGP expandability. It is mentioned here because it is one of the advantages of confederation. An AS can be combined with an existing confederation. This can be realized by taking the new AS as the sub-AS of the confederation.

HUAWEI TECHNOLOGIES CO., LTD. All rights reserved



1. Describe the problem solved by BGP route reflector and BGP confederation.

A: Because of the BGP route advertisement behavior, IBGP peers must be interconnect to each other.

This result in the IBGP full meshes connection. The IBGP full meshes connection can solve the problem caused by the BGP advertisement behavior. However, it brings another problem to the network. The BGP speaker must therefore maintain

a large numbers of IBGP sessions. So BGP introduces route reflection and confederation.

2. Describe the advertisement principle of BGP route reflector.

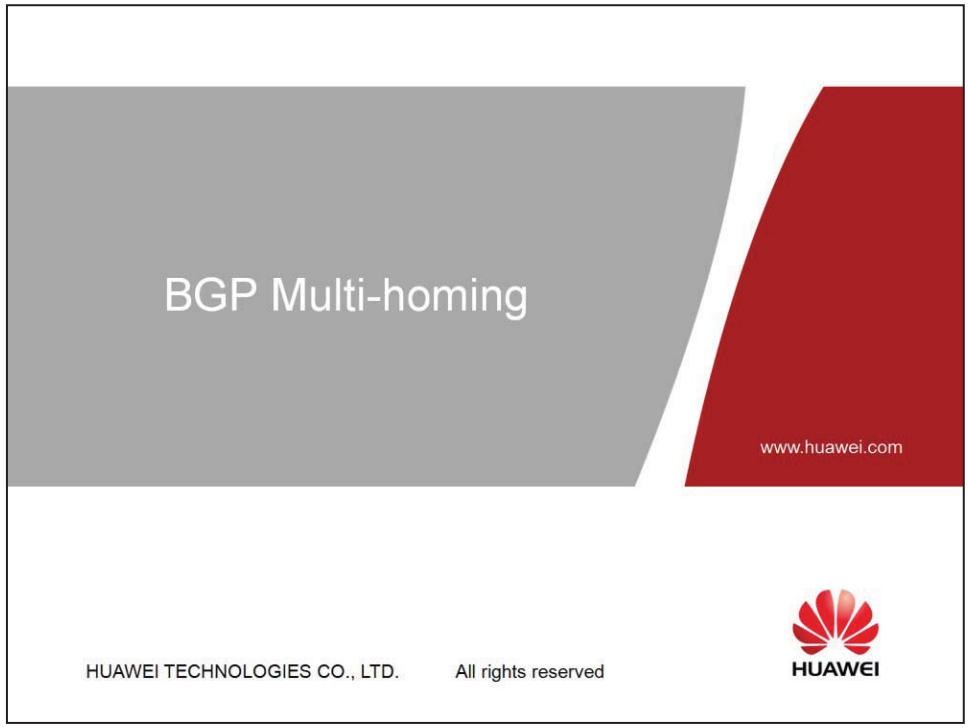
A: 1). Select the best route based on the BGP route selection process 2). For the route received from Non-client IBGP, reflect only to all the Client Peers

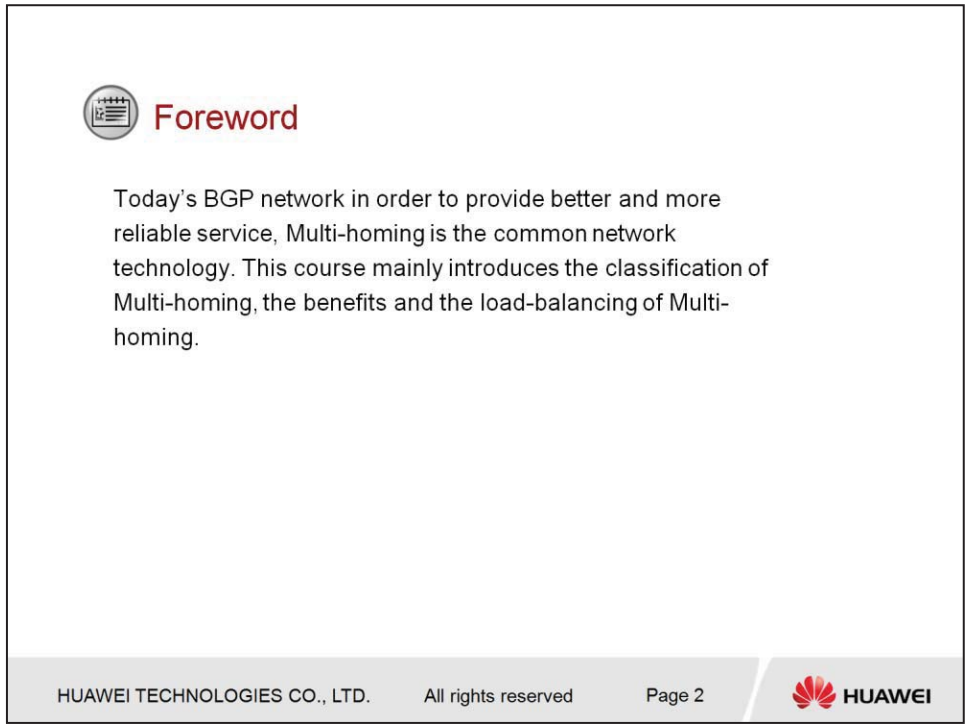
3). For the route learnt from Client IBGP, reflect to all Clients and Non- Clients 3. Describe the AS_PATH modification in BGP confederation.

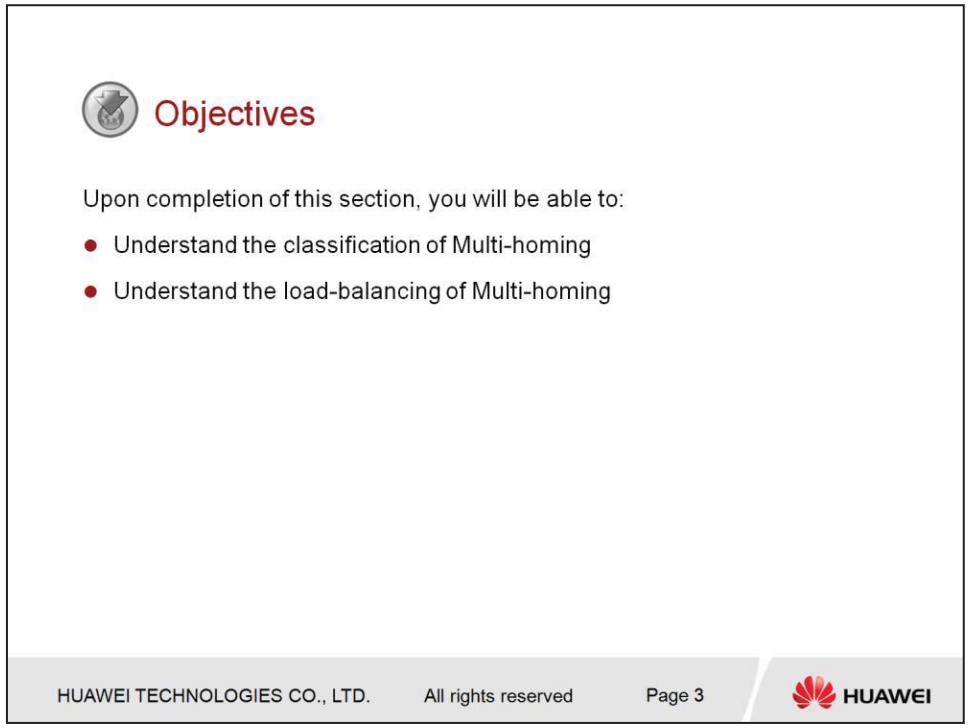
A: The confederation technology will generate a lot of sub-autonomous systems.

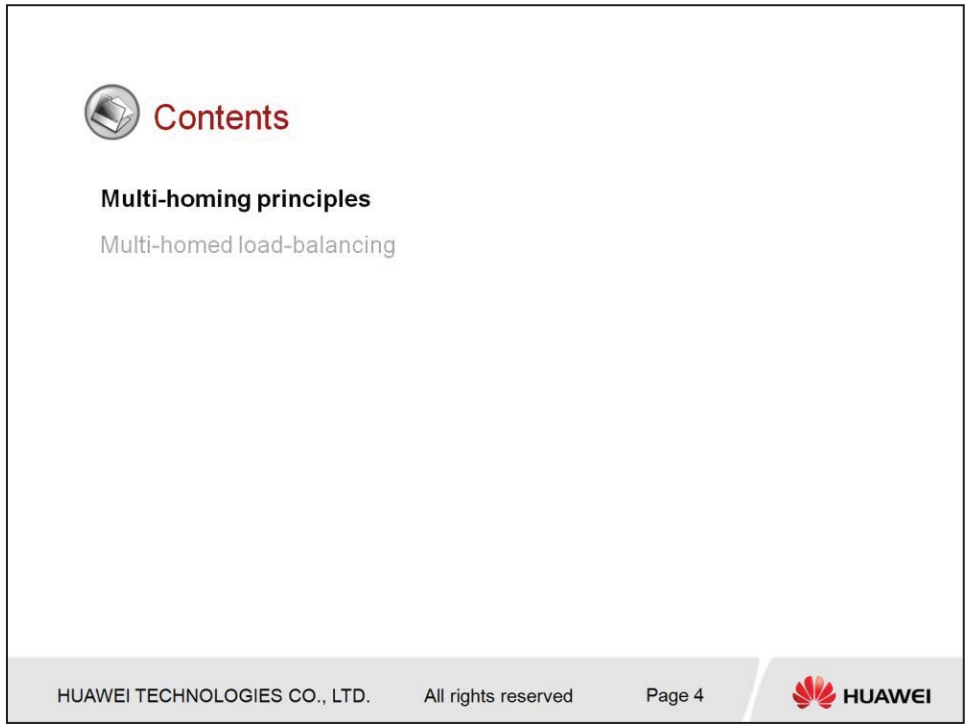
Therefore, 2 new AS_PATHs have been introduced for

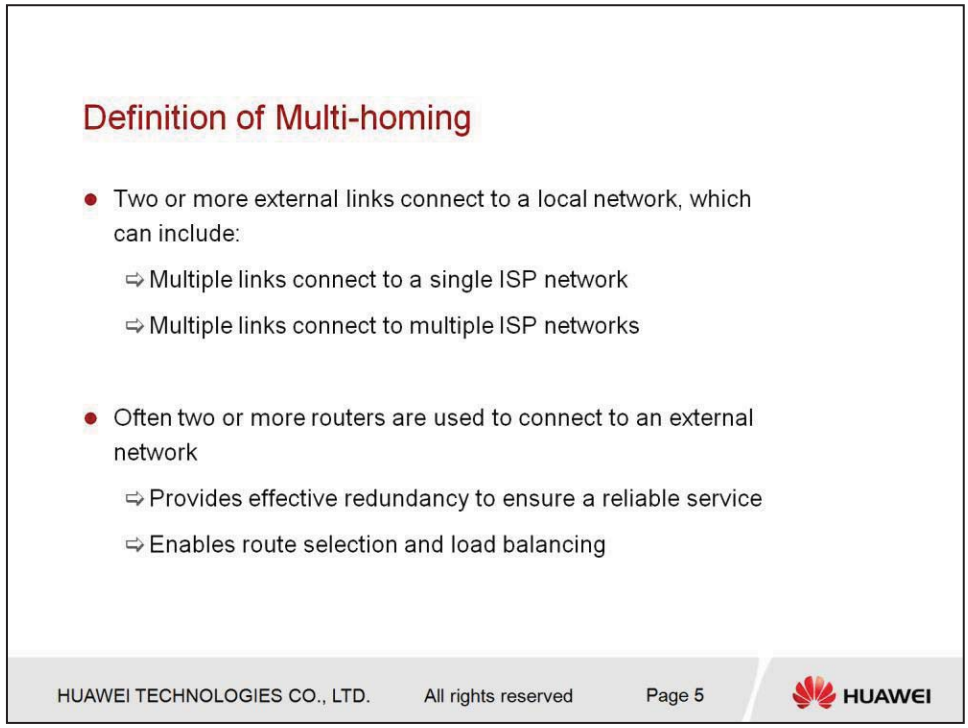
confederation. There are AS_CONFED_SEQUENCE and AS_CONFED_SET. The update is transmitted between the sub-autonomous systems. Each time the update information passes through a sub-autonomous system, the sub-autonomous system number will be included in the leftmost position of the AS_CONFED_SEQUENCE. When an update is sent to a peer external to the confederation, the AS_CONFED_SEQUENCE information is removed from the AS_PATH attribute, and the confederation ID is prepended to the AS_PATH. Because of this, external peers see the confederation as a single AS rather than as a collection of autonomous systems. Apart from that, when a confederation receives an update from the AS outside the confederation, the confederation will keep the AS_PATH of that AS outside the confederation and create an AS_CONFED_SEQUENCE for used inside the confederation


BGP Multi-homing
www.huawei.com
HUAWEI TECHNOLOGIES CO., LTD.
All rights reserved
HUAWEI


Foreword
Today's BGP network in order to provide better and more reliable service, Multi-homing is the common network technology. This course mainly introduces the classification of Multi-homing, the benefits and the load-balancing of Multi-homing.
HUAWEI TECHNOLOGIES CO., LTD.
All rights reserved
Page 2
HUAWEI


Objectives
Upon completion of this section, you will be able to:
Understand the classification of Multi-homing
Understand the load-balancing of Multi-homing
HUAWEI TECHNOLOGIES CO., LTD.
All rights reserved
Page 3
HUAWEI


Contents
Multi-homing principles
Multi-homed load-balancing
HUAWEI TECHNOLOGIES CO., LTD.
All rights reserved
Page 4
HUAWEI



The network is multi-homed if it has more than one exit point to the outside world.

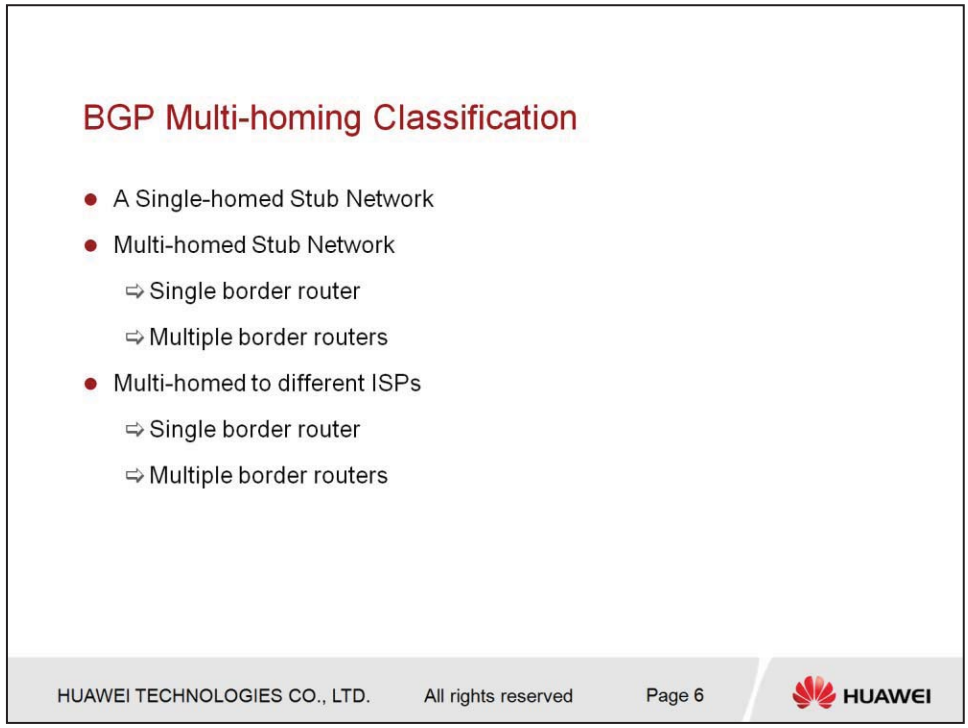
For example, the network can be multi-homed to a single provider or multiple providers.

The purpose of multi-homed is to provide the reliability or achieve the load balancing via redundant links.

Load balancing allows a router to distribute the incoming traffic or outgoing traffic among multiple paths. Multiple paths can be learnt through either static route or dynamic routing protocol such as RIP, OSPF and so on.

By default, BGP allow only a single best path and does not perform load balancing.

This lesson will introduce the methods how to perform load balancing in different scenarios using BGP.



BGP multi-homed can be divided into the following types:

•Single-homed Stub Network

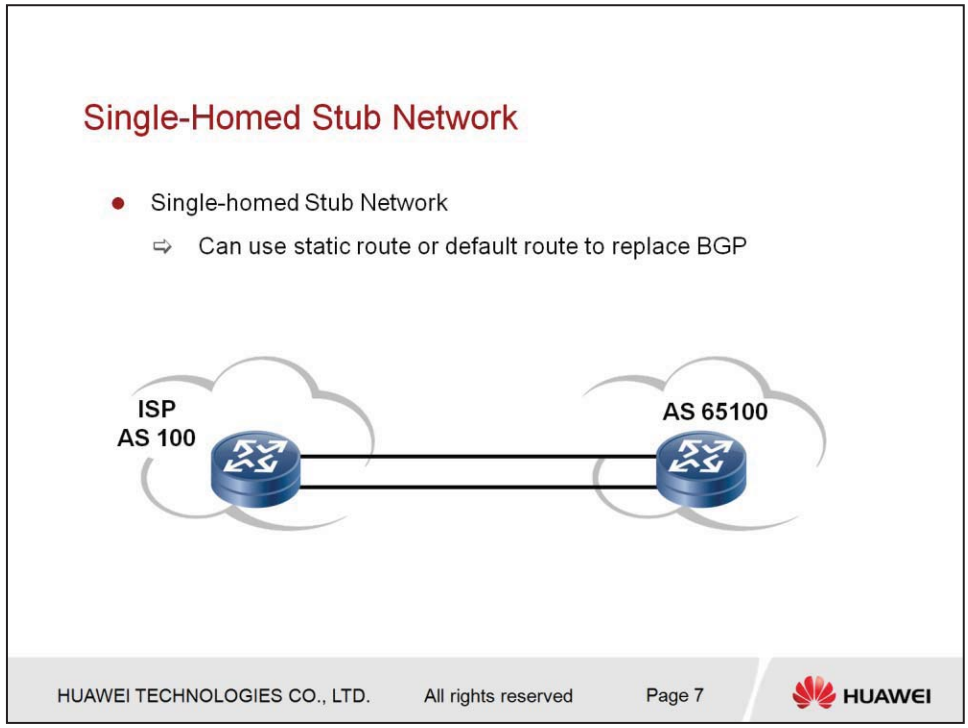
•Multi-home Stub Network

Single border router

Multiple border routers

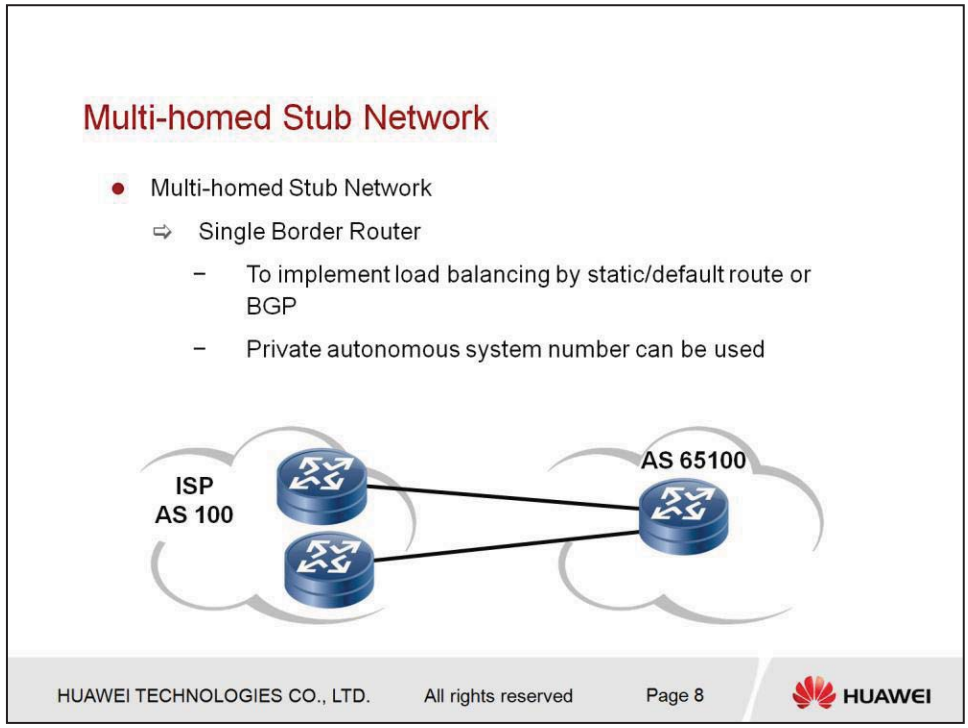
•Multi-homed to different ISPs

This lesson will discuss the types of BGP multi-homed network listed above in detail.



The scenario refers to single-homed Stub Network where 2 different autonomous systems are connected though their respective border router in their own autonomous system. To achieve load balancing, multiple links are used between 2 border routers. Current VRP platform can support a maximum of 8 equal cost paths for load balancing.

Single-homed network can not provide high redundancy and is therefore suitable for small network. As shown in the diagram above, customer in AS can reach ISP100 via 2 links. In this kind of network, BGP is not necessary for the customer network in AS65100 . The customer only need to add a static default route to the border router and redistributes the route into its own AS. Similarly, the ISP need to add a static route pointing to the customer's address range and advertises these route into its AS.



Multi-homed Stub Network for a single border router can provide the redundancy by protecting not only failure of a single link but also protect failure of a single ISP router.

When implement this kind of design solution, BGP can be used to provide an effective control for load balancing that might be implemented.

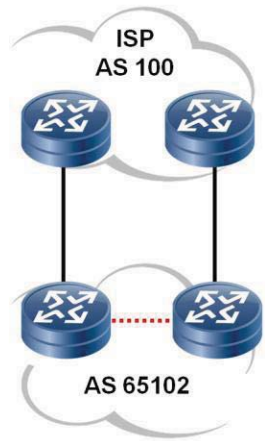
Private autonomous system number can be assigned to the customer network for the single upstream service provider solution as shown in the diagram above. The service provider will send a default route to customer via BGP. The customer does not require to obtain a public unique registered autonomous system number that is visible in the Internet from IRR. The upstream service provider will remove the

private autonomous system number from the received update information.

The use of BGP in this design solution allows the enterprise to have a better control the incoming and outgoing traffic flow. It is very useful when both of the links have different link bandwidth. This is because route selection policy can be used for proportionally traffic load balancing according to the bandwidth of the link.

Multi-homed Stub Network (Cont.)

- Multi-homed Stub Network
  - ⇨ Multiple border routers
    - Can better control the outgoing traffic and proportional traffic load balancing can be implemented according to the bandwidth of the link
    - IBGP session and IGP session need to be run between the border routers



HUAWEI TECHNOLOGIES CO., LTD. All rights reserved

The use of single enterprise border router might result in single point of failure.

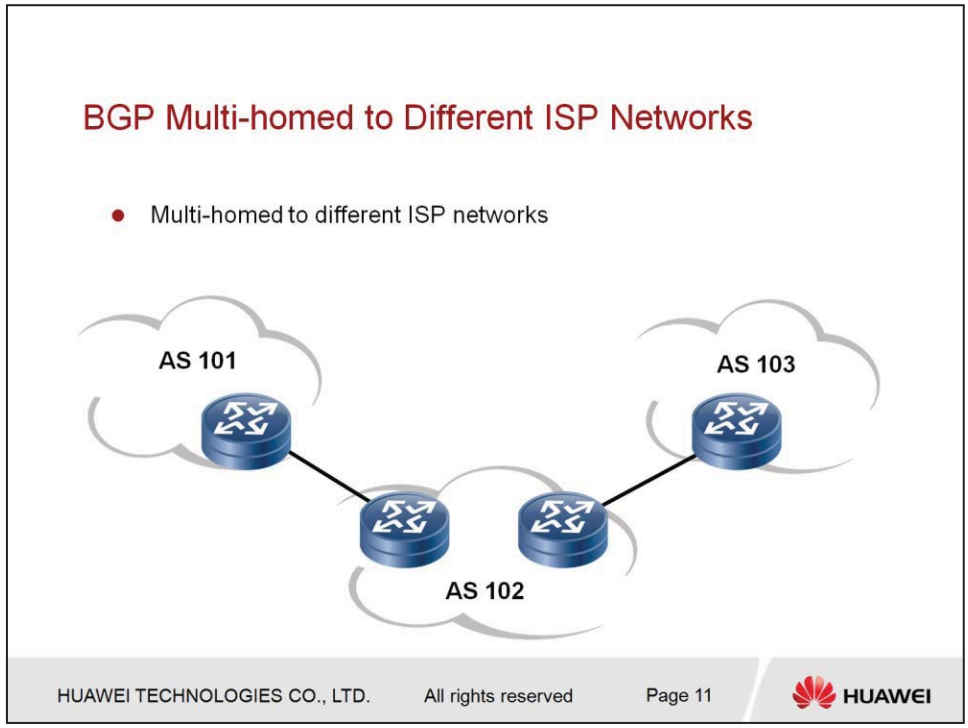
This single point of failure can be avoided by adding one or more enterprise border routers. Each of the enterprise border routers is connected to one or several links of the upstream service provider. This kind of solution still connects to single upstream service provider.

The enterprise is still allowed to use the private autonomous system number in this design. The advantage of using private autonomous system number is that the enterprise does not require to obtain a new public autonomous system number.

The use of BGP provides additional support for enterprise to define the route policies for incoming and outgoing traffic. In addition to the EBGP session with upstream service provider, the enterprise needs to establish full mesh IBGP sessions between all the border routers and all the layer 3 devices that might provide the cross-over to border routers. This is to ensure that the traffic will not be sent to devices that do not have routing information of the intended destination.

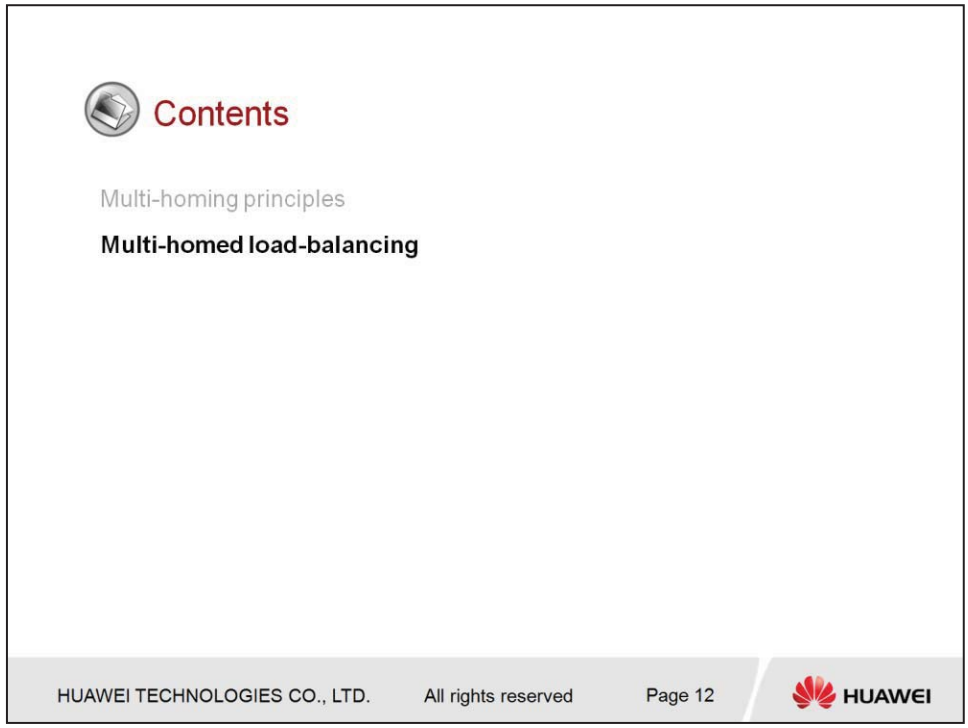
The enterprise network should distribute the default route from each of the border routers. The default route can be advertised

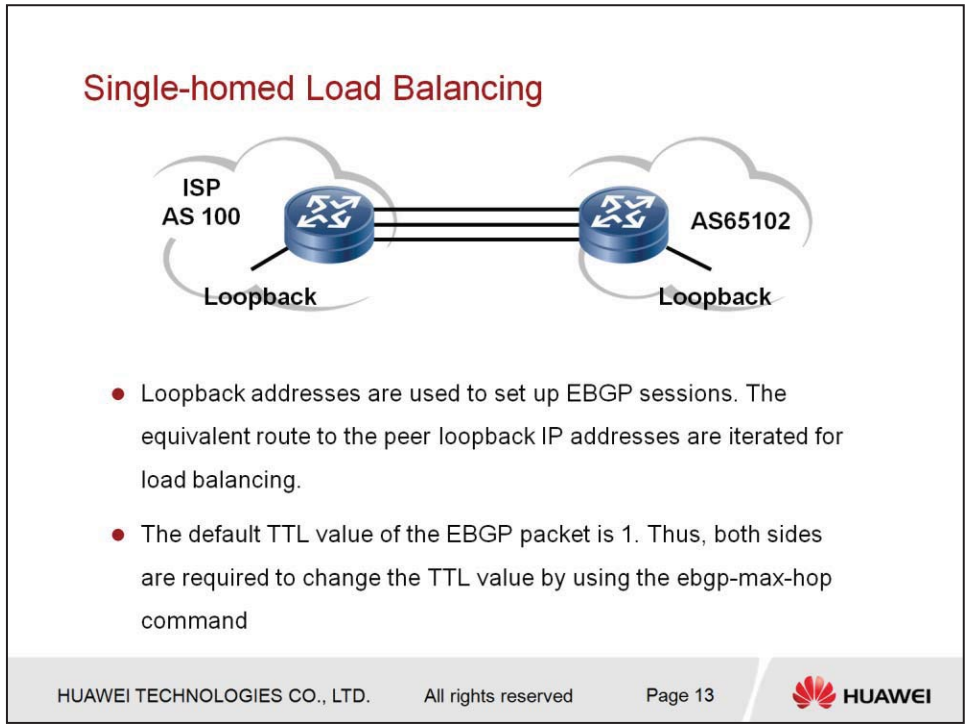
only when the uplink is connected and activated. This prevents the traffic from going to the border router along the default route in the case where the uplink of the border router fails. This kind of conditional advertisement can be achieved by the static default route pointing to the interface, or by importing the default static route received from BGP to IGP. In this context, the conditional advertisement here is not referred to the BGP conditional advertisement feature. Other route information received from the upstream service provider should not be imported to any of the IGP process running on the border router.



In this design, the border routers need to establish the EBGP peer relationship with the routers of their upstream service provider. Besides, the full IBGP sessions are established between all the border routers and any of the layer 3 devices that might provide cross-over to the border routers. The received routing information might include only the default route, or even the complete routing table. In this case, the routing information received is the same as that of single router.

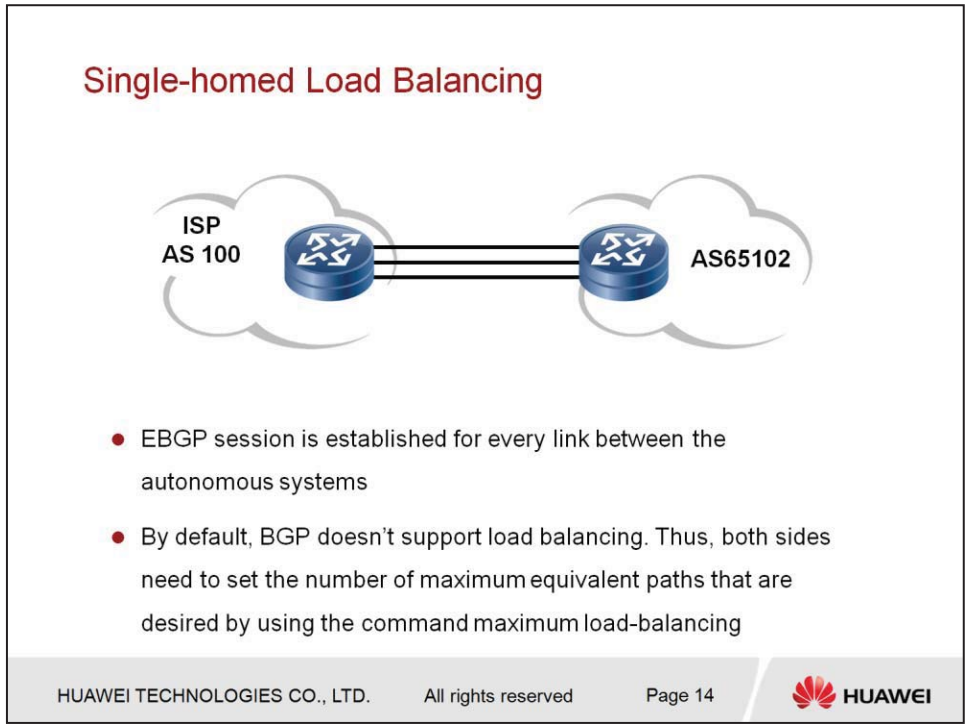
The most common load balancing mechanism involves the use of partial route selection information. This means that the enterprise network requests some of the routing information from its upstream service provider and use it together with the default route, or request the full routing table and modify the inbound or outbound filtering policies for reasonable load balancing. Finally, the implementation method is dependant on the objective of the enterprise. The easiest method is to use one of the links as the primary connection while the other links are solely used for backup purpose. The most difficult task is to achieve load balancing evenly among multiple paths.


Contents
Multi-homing principles
Multi-homed load-balancing
HUAWEI TECHNOLOGIES CO., LTD.
All rights reserved
Page 12
HUAWEI



A single EBGP session is established between 2 routers. The EBGP session above use loopback address instead of the physical address as the source. Each of the directly connected physical interfaces is configured with a static route pointing to the remote loopback address. This solution resolves the next hop IP address and the iterated route of next hop IP address is used to achieve load balancing.

Please note that we must manually modify the TTL value of the EBGP. Its value must be equal or greater than 2. Else, the BGP session cannot be established.

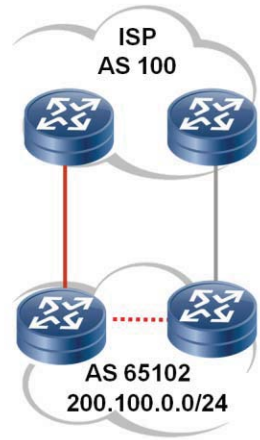


EBGP multi-path provides another solution for traffic load balancing on multi-path links. An EBGP session is established for each link between 2 routers. These EBGP sessions are directly bound to the physical interface addresses. The result

is that both of the routers receive multiple path information with one link representing one path information. EBGP multi-path allows all paths to be installed on the router up to a maximum of 8 paths.

Multi-homed to the same ISP — Primary and Backup Solution

- Primary and backup link solution
  - ⇨ Advertise /24 aggregate route on each links
    - Both primary and backup links advertise the route
    - On backup link, the MED value of the route on outbound direction is increased while the local preference value of the route on inbound direction is decreased
  - ⇨ AS 100 removes the private AS number when it advertises the route to external neighbor. However, the private AS number is still visible inside AS 100



HUAWEI TECHNOLOGIES CO., LTD. All rights reserved HUAWEI

The local AS65102 can reach the upstream AS100 via 2 different links. One of the link is configured as the primary link (the link in red color is primary link) and its responsibility is to transmit all the traffic. Another link is configured as the backup link (the link in grey color is backup link). When the primary link fails, all the traffic will be switched over to the backup link.

As shown in the diagram above, both primary and backup links advertise the route 200.100.0.0/24. However, the MED value of the route 200.100.0.0/24 is increased through the routing policy on the backup link.

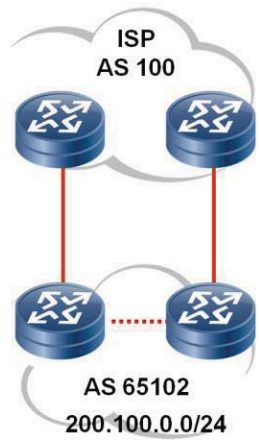
On the other hand, the local preference value of all the received IP prefix on inbound direction of the backup link is decreased through routing policy. By doing so, the incoming and outgoing traffic can be distributed on the primary link.

The private autonomous system number can be used within the local AS65102.

AS 100 will remove the private AS number when it advertises the route to the external neighbor.

Multi-homed to the same ISP— Load Balancing Solution

- Load balancing base on inbound traffic
  - ⇨ Advertise /24 route on each link
  - ⇨ Split /24 route into two /25 routes. Advertise the /25 prefixes on one of the routers
  - ⇨ Can further split the /25 route to implement the "50/50" load balancing
- Load balancing based on outbound traffic
  - ⇨ Default route can be used
  - ⇨ Accept the routing information from upstream
    - To implement load balancing base on outbound traffic via "the nearest exit" point.



HUAWEI TECHNOLOGIES CO., LTD. All rights reserved HUAWEI

This solution uses multiple local routers to provide multiple links that connect to the upstream autonomous system. In the diagram above, 2 of the routers connect to the upstream AS100 through EBGP peer respectively. Both of the links

implement load balancing for the inbound and outbound traffic of local network 200.100.0.0/24 Both of the local routers advertise the route 200.100.0.0/24 as normal. At the same time, the /24 route is subnetted into two /25 prefixes. Then, the /25 prefixes are advertised to the upstream AS100 through one of the routers. Please take note that the /25 prefixes may not be installed in the IP routing table of the device.

Therefore, we need to add a static route on advertising router: ip route-static 200.100.0.0 25 null 0. This is done to ensure the /25 prefixes can be advertised to the external peer successfully.

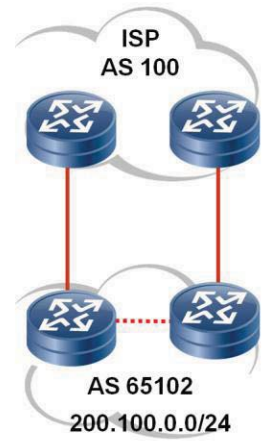
The purpose of performing the subnetting is to allow the upstream equipments match the routes as detailed as possible and implement the load balancing base on the inbound traffic. We can further subnet the /25 prefix if the perfect load

balancing can not be achieved by using the /25 prefixes. This action can be continue until the perfect load balancing which is close to 50/50 is achieved.

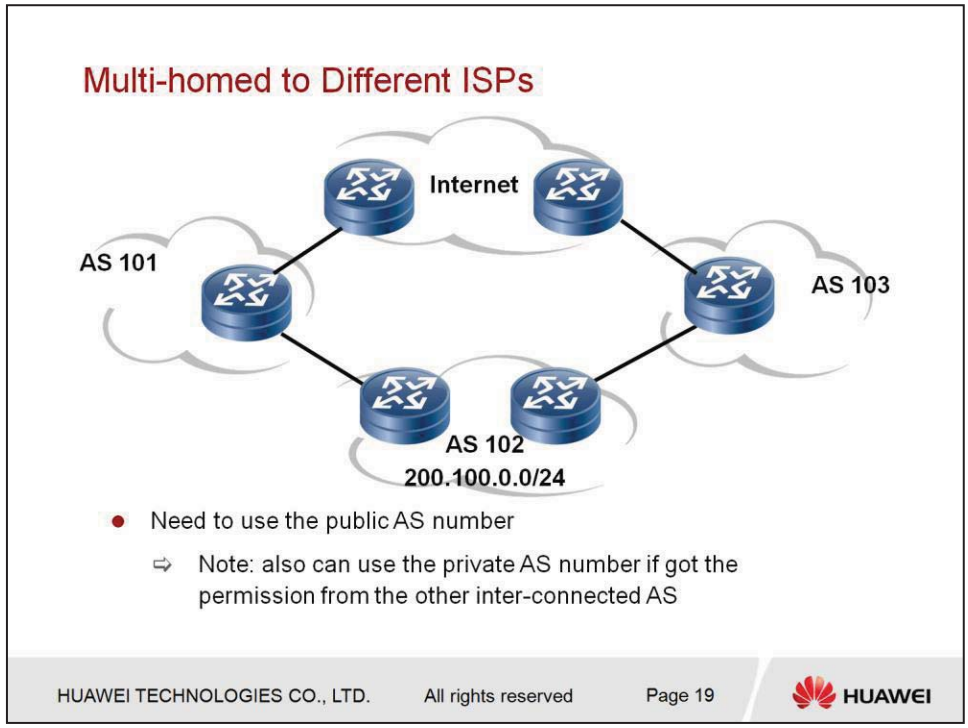
How to implement the load balancing for outbound traffic? We can receive only partial routing information which is desired by using route policies. Load balancing can be achieved by configuring route policy based on different destination IP addresses.

Multi-homed to the same ISP— Load Balancing Solution (Cont.)

- Load Balancing Solution (Continue)
  - ⇨ The border router of AS100 accept only the routing information from its downstream
  - ⇨ AS100 removes the subnet prefixes information from downstream AS65102
  - ⇨ AS100 removes the private autonomous system number of the route when advertising it to the external neighbor

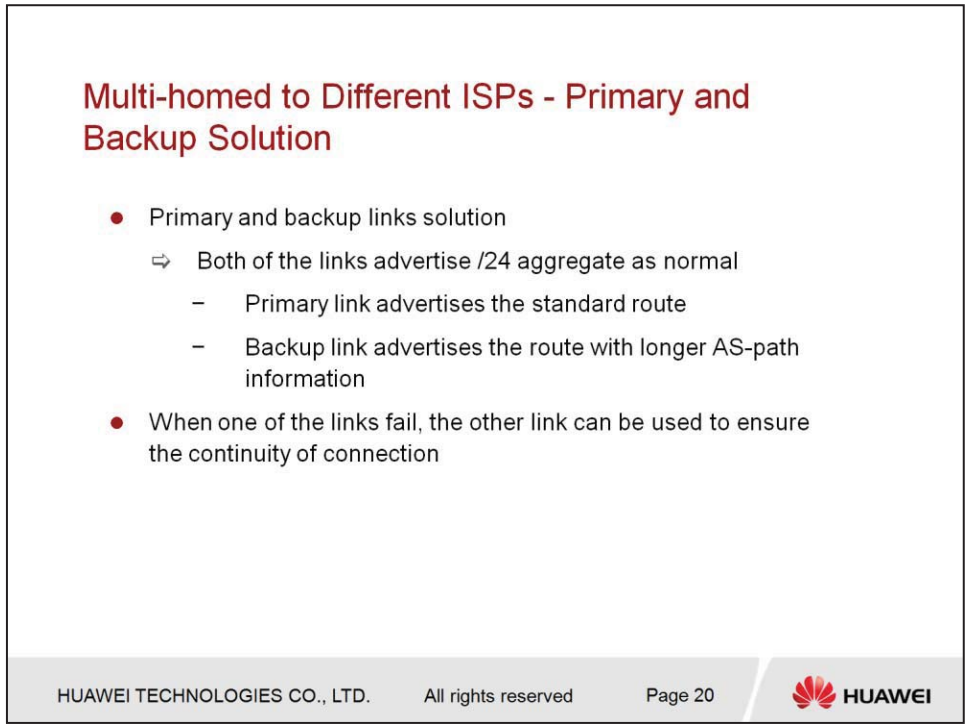


HUAWEI TECHNOLOGIES CO., LTD. All rights reserved



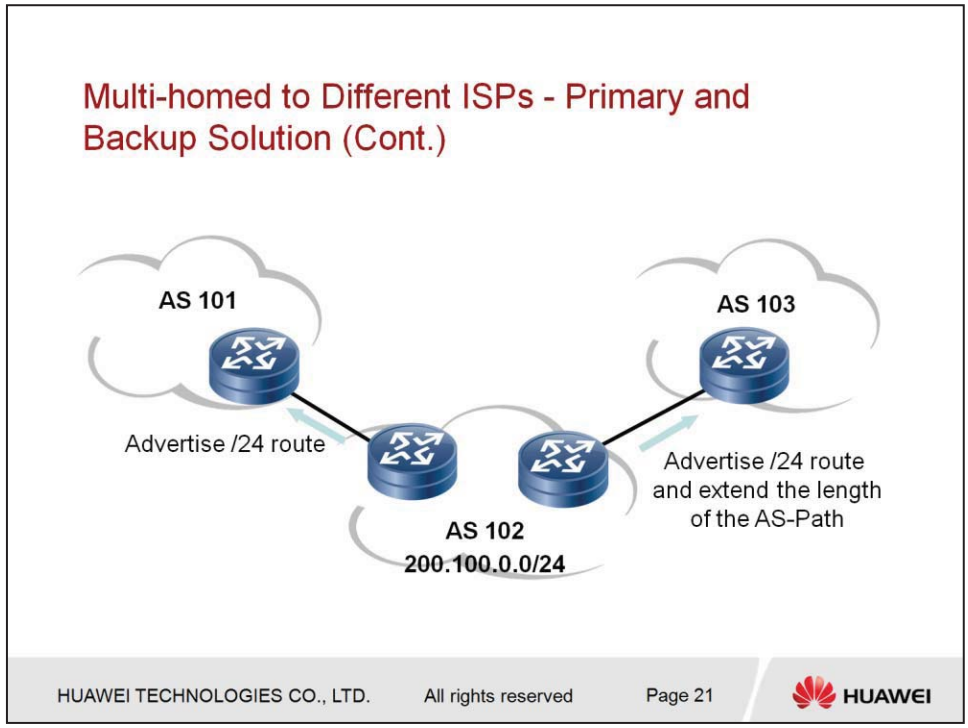
In this kind of design, the EBGP peer relationship is established between the border routers of the enterprise and their upstream ISP. In addition, the full mesh IBGP session is established between all the border routers and all the layer 3 equipment that might provide the cross-over to the border routers. The received routing information might include only the default route, or even the complete routing table. In this case, the received routing information is the same as that of single router.

The most common load balancing mechanism involves the use of partial route selection information. This means that the enterprise network requests some of the routing information from its upstream service provider and uses it together with the default route, or request the complete routing table and modify the inbound or outbound filtering policies for reasonable load balancing. Finally, the implementation method is dependant on the objective of the enterprise. The easiest method is to use one of the link as the primary connection while the other links are solely used for backup purpose. The most difficult task is to achieve load balancing evenly among multiple paths.

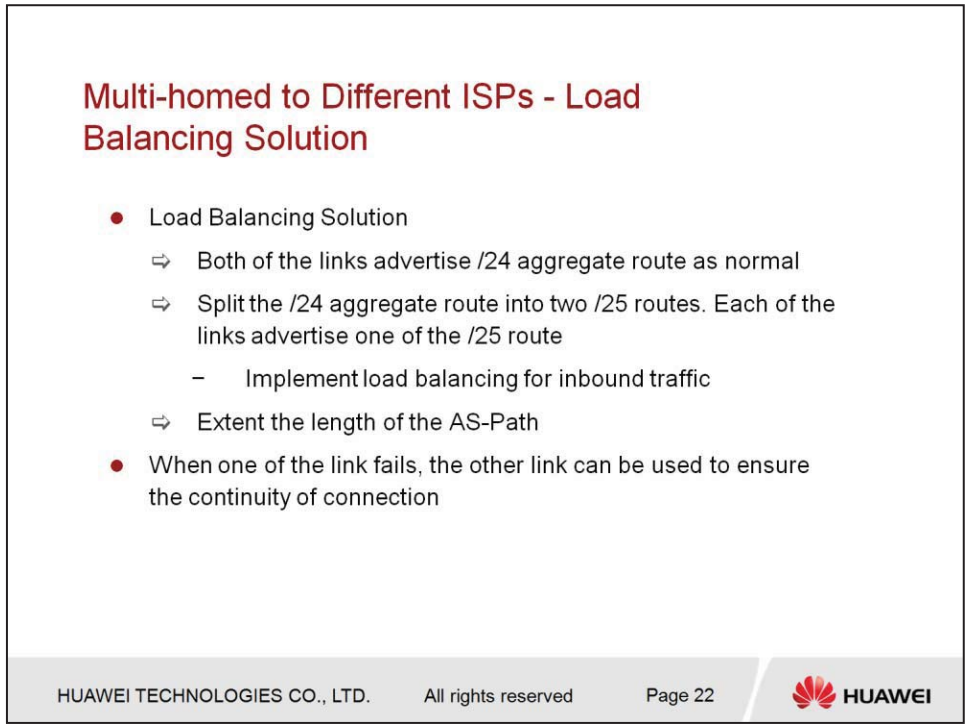


This solution is similar to the implementation of multi-homed to the same ISP.

The traffic flow which enters into the enterprise network through the optimized route depend on AS-Path attribute.



As shown in the diagram above, AS102 advertises the route 200.100.0.0/24 on both up-links as normal. The backup link also advertises the route and extends the length of the AS-Path. Therefore, the inbound traffic enters the enterprise network through the master link. At the same time, the backup can provide redundancy.

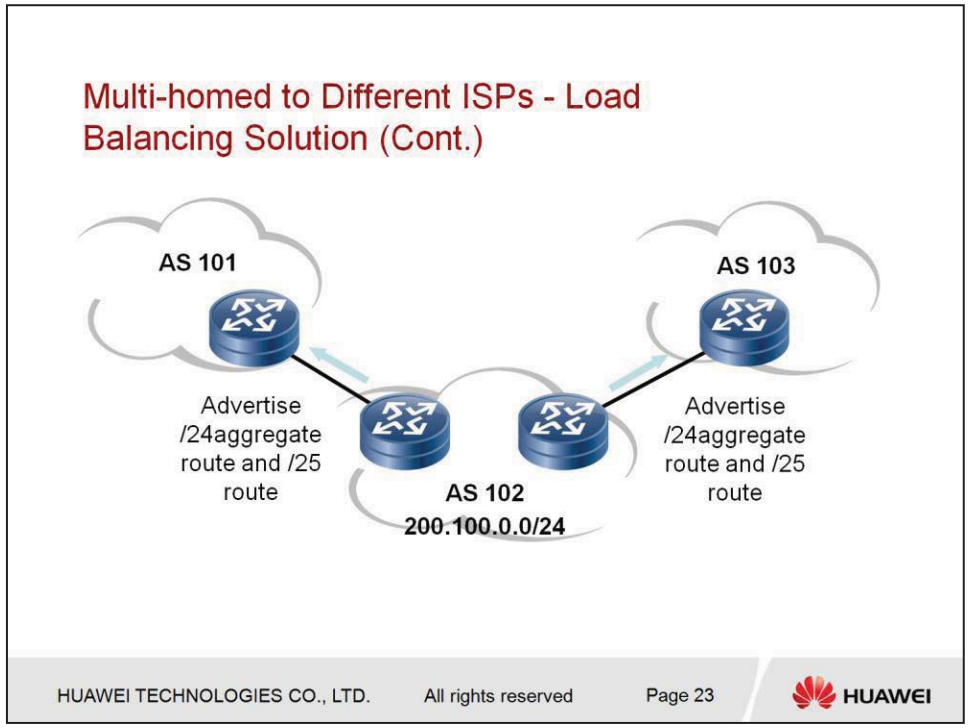


This solution is similar to the implementation of multi-homing to the same ISP.

The purpose for splitting the prefix is to allow upstream equipments to match the specific route that have been splitted. By doing so, load balancing base on inbound traffic can be implement.

But this method can not always be used. In this situation, to extend the length of the AS-Path also can implement load balancing. After the routes have been advertised to different ISPs, the traffic utilization rate on the links are monitored.

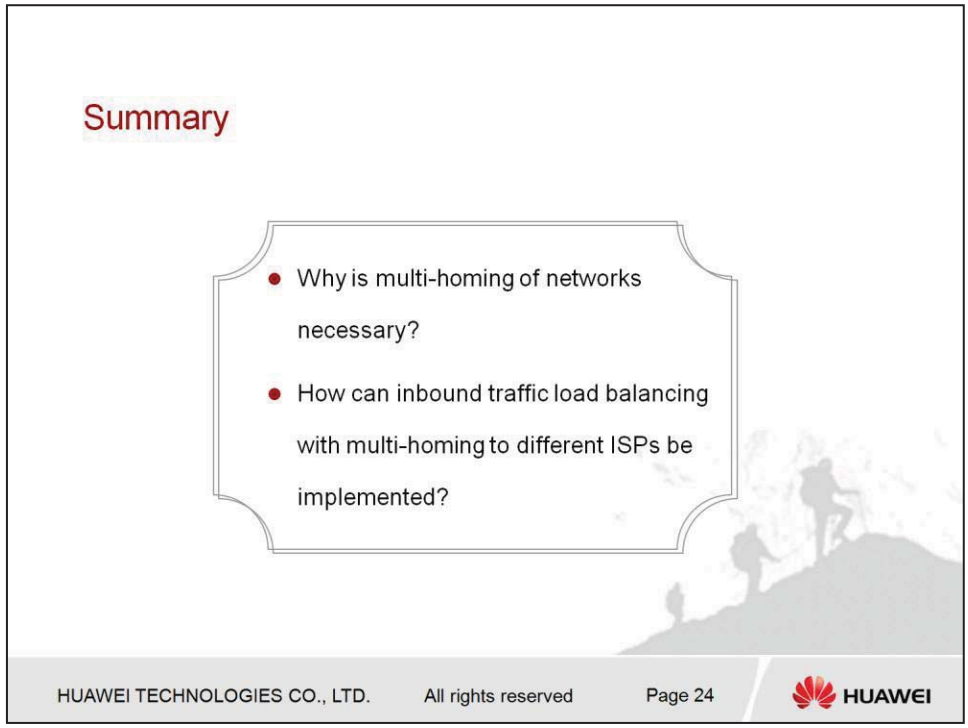
If the traffic flows on one of the links is obviously heavier compare to another link, then we can extend the AS-Path length on the link that has higher utilization rate.



As shown in the diagram above, AS102 advertise the route 200.100.0.0/24 on both up-links as normal. At the same time, one /24 prefix is split into two /25 prefixes.

One of /25 prefixes is advertised to one side and one of the /25 prefixes advertised to the other side. After the routes have been advertised to different ISPs, the traffic utilization rate on the links are monitored. If the traffic flows on one of the links is still obviously heavier compared to another link, then we can

extend the AS-Path length on the link that has higher utilization rate.



Why is multi-homing of networks necessary?

A: Multi-homed solution provides the network with better redundancy, and abundant traffic control solution.

2. How can inbound traffic load balancing with multi-homing to different ISPs be implemented?

A: For multi-homing to different ISPs, the inbound traffic load balancing can be achieved by increasing the length of the AS_PATH. It is recommended that the length of the AS_PATH should be increased by one each time. This is because we can control the traffic flow each time the length of the AS_PATH is increased by one. Therefore, we should observe the traffic flow and increase it gradually.


BGP Troubleshooting
www.huawei.com
HUAWEI TECHNOLOGIES CO., LTD.
All rights reserved
HUAWEI

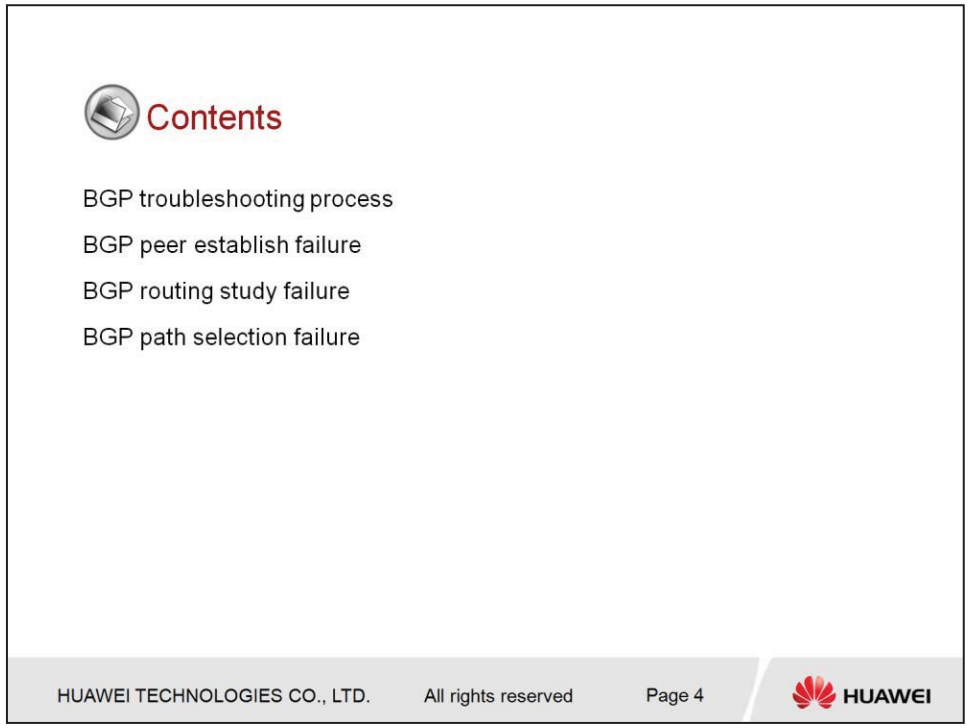


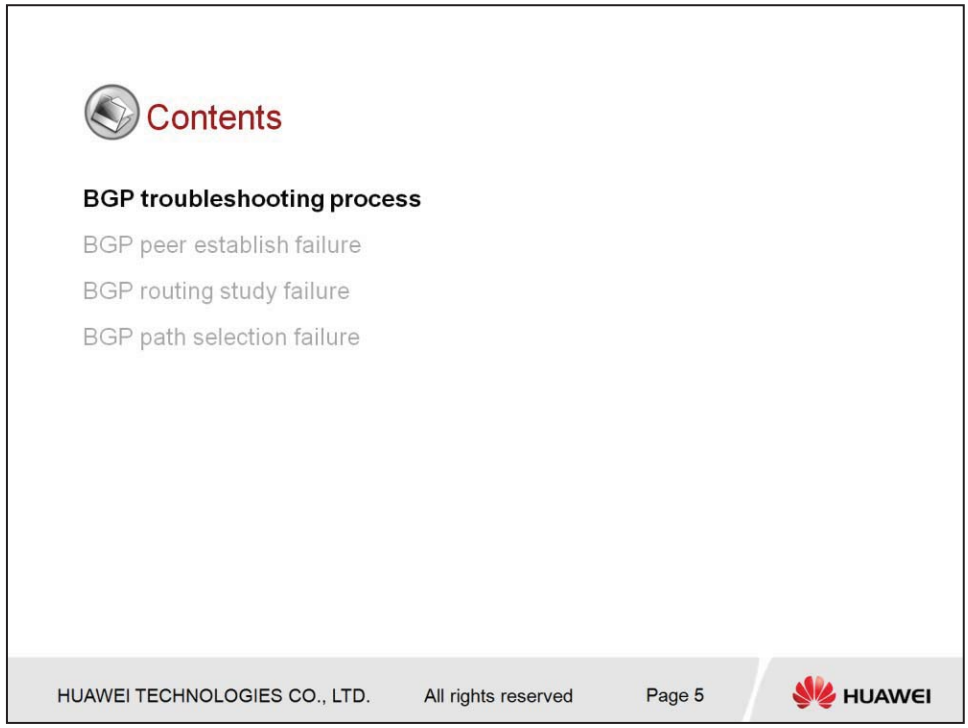
Foreword

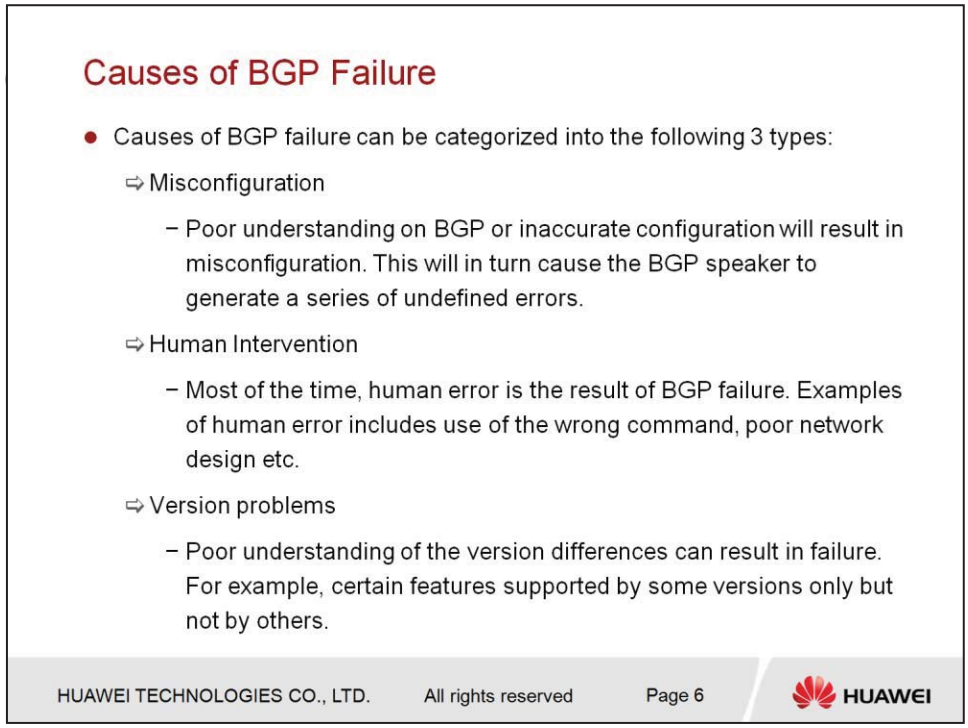
BGP as a complex inter-domain routing protocol often comes up a variety of failures, and to locate the causes and troubleshoot accurately, requires establishment of a strong understanding of the operation of the protocol. This section aims to improve the capability to handle failures by deepening the understanding of a BGP instance.

HUAWEI TECHNOLOGIES CO., LTD. All rights reserved


Objectives
Upon completion of this section, you will be able to:
Understand the basic steps for troubleshooting
Understand basic troubleshooting techniques for BGP
HUAWEI TECHNOLOGIES CO., LTD.
All rights reserved
Page 3
HUAWEI


Contents
BGP troubleshooting process
BGP peer establish failure
BGP routing study failure
BGP path selection failure
HUAWEI TECHNOLOGIES CO., LTD.
All rights reserved
Page 4
HUAWEI


Contents
BGP troubleshooting process
BGP peer establish failure
BGP routing study failure
BGP path selection failure
HUAWEI TECHNOLOGIES CO., LTD.
All rights reserved
Page 5
HUAWEI



Causes of BGP failure can be categorized into the following 3 types:

Mis-configuration

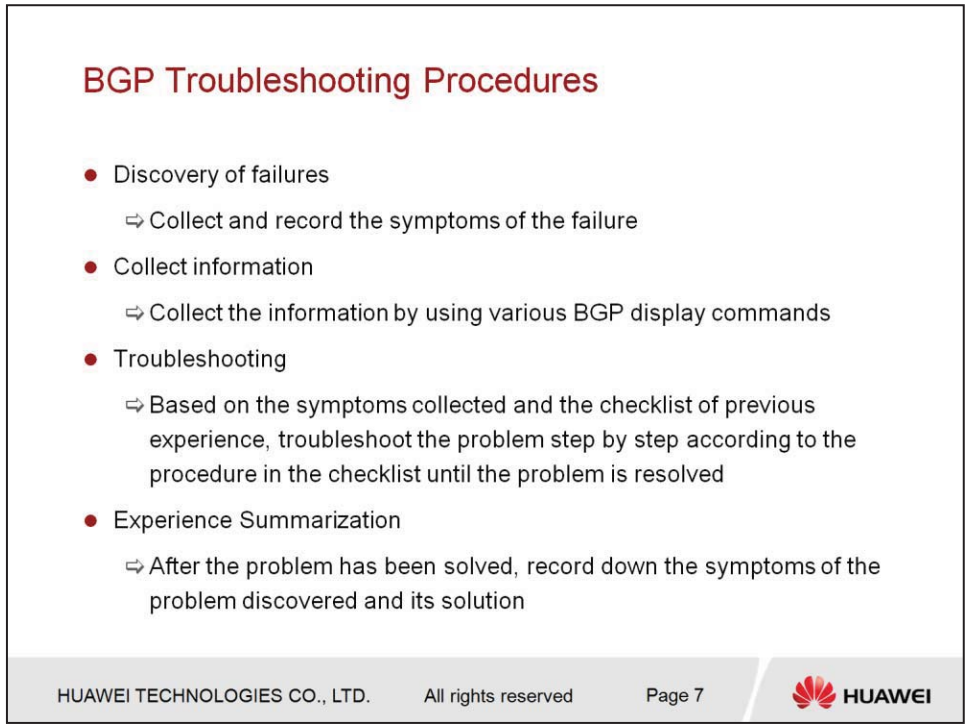
Poor understanding on BGP or inaccurate configuration script will cause mis-configuration. This will in turn result in a series of undefined errors between the BGP speakers.

Human being problems

Human being problems can result in BGP failure regardless of the types of routing protocol you are using. The following are example of human being problems: use of wrong command, poor network design and so on

Version Problems

Poor understanding in the version information can result in failure. For example, certain features supported by some version only but not by others



Troubleshooting Procedure：

1. Discover the failure

Collect and record the symptom of the failure

2. Collect information

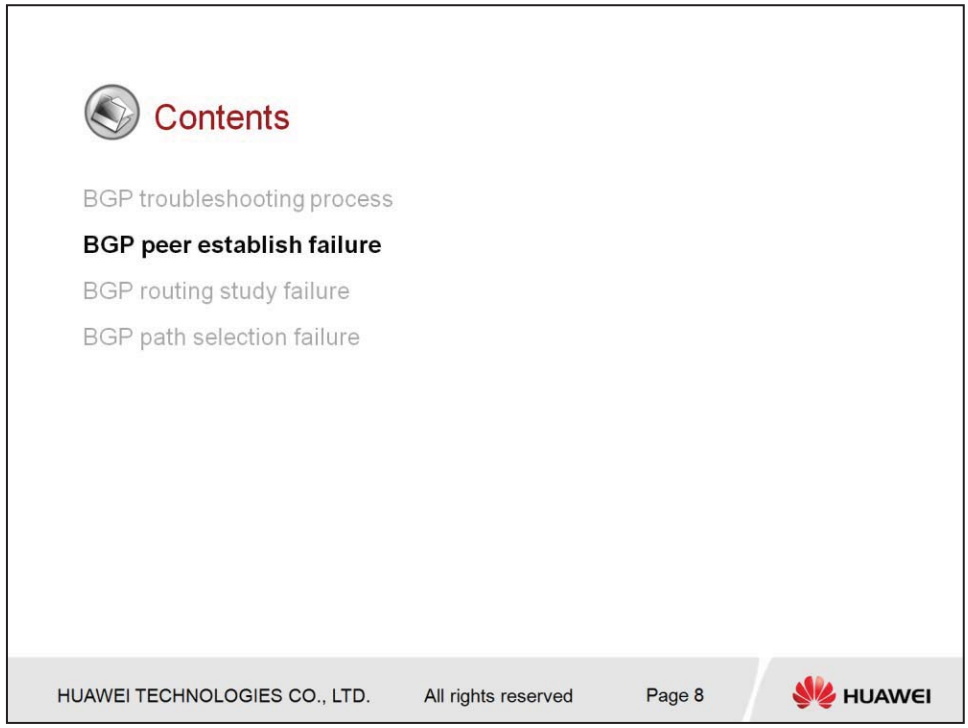
Collect the information by using various BGP display command

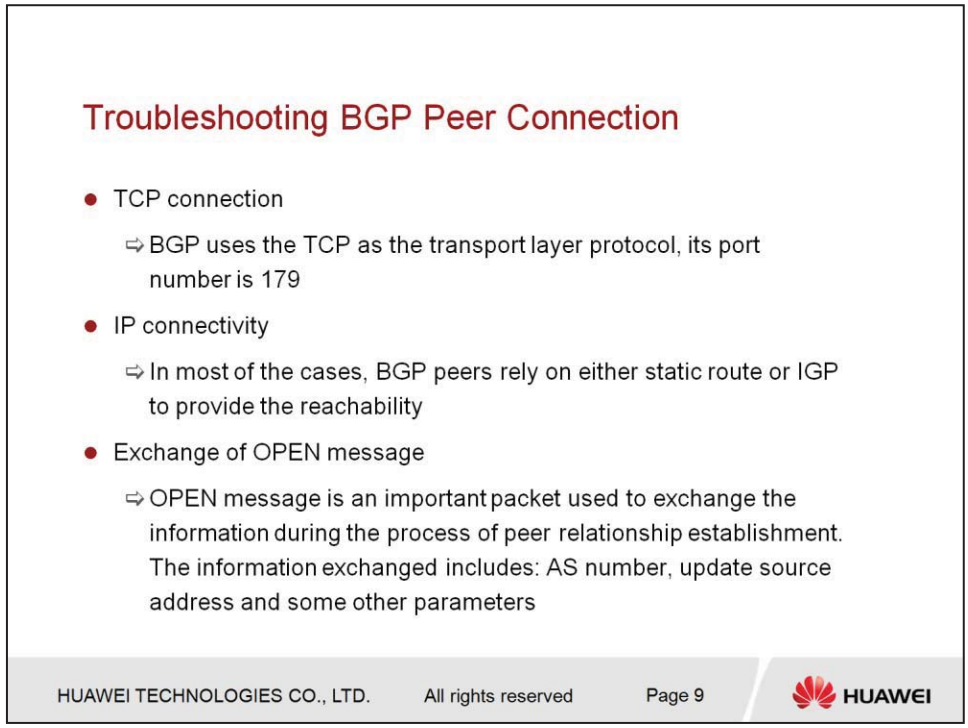
3. Troubleshooting

Base on the symptoms collected and the checklist of previous experience, troubleshoot the problem step by step according to the procedure in the checklist until the problem is solved. Contact the technical support if problem cannot be solved.

4. Summarize the experience

After the problem has been solved, record down the symptoms of the problem discovered and its solution. The objective of doing so is to share the experience with the engineers who provide the support in future.


Contents
BGP troubleshooting process
BGP peer establish failure
BGP routing study failure
BGP path selection failure
HUAWEI TECHNOLOGIES CO., LTD.
All rights reserved
Page 8
HUAWEI



BGP will not be able to form the neighbor relationship when error occurs on one of the three points listed below:

TCP connection BGP uses the TCP as the transport layer protocol, its port number is 179.

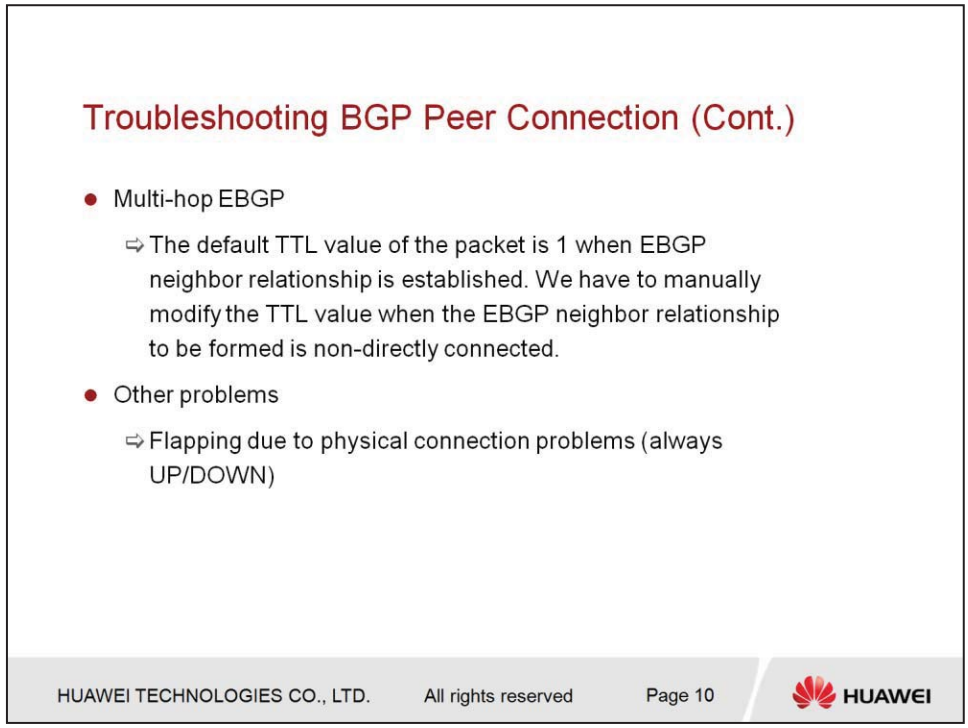
If port 179 is blocked, BGP will not be able to form the neighbor relationship.

IP connectivity

In most of the cases, BGP peers rely on either static route or IGP to provide the reachability

Exchange of OPEN message

OPEN message is an important packet used to exchange the information during the process of peer relationship establishment. The information exchanged includes: AS number, update source address and capabilities of the BGP peer



In addition to pay attention to the problems with regarding to the TCP connection, IP connectivity, and exchange of OPEN message, configuration error is another common reason that causes the failure of BGP.

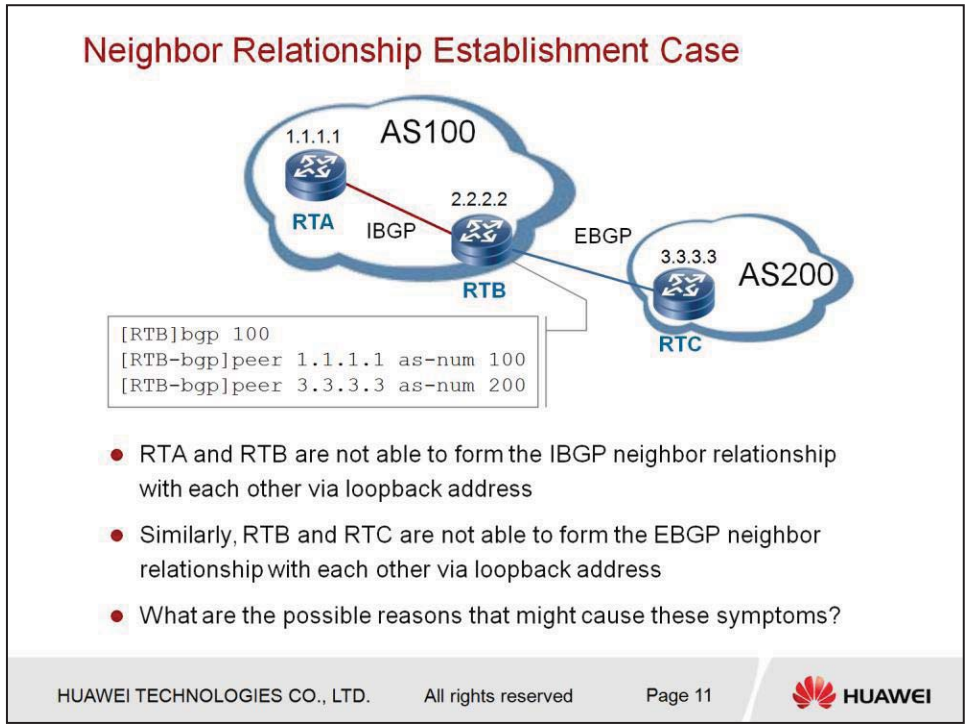
By default, the TTL value for the EBGP update packet is 1. Therefore, we need to manually modify the TTL value by using command when BGP forms the neighbor relationship through loopback address or non directly connected interface.

Besides, we need to aware that the mismatch of the source interface used to form the neighbor relationship will result in the failure of EBGP neighbor relationship establishment.

In addition to the establishment of EBGP peer relationship, we have to pay attention to the establishment of IBGP peer relationship as well. Similarly, the mismatch of the source interface used to form the neighbor relationship will result

in the failure of the IBGP neighbor relationship establishment.

Next, we need to pay attention to the other problems such as the physical connectivity problem. The physical connectivity problems will result in flapping of the link.



As shown in the diagram above, RTB in AS100 would like to form IBGP neighbor relationship with RTA in the same AS . Besides, RTB in AS100 would like to form EBGP neighbor relationship with RTC in AS200.

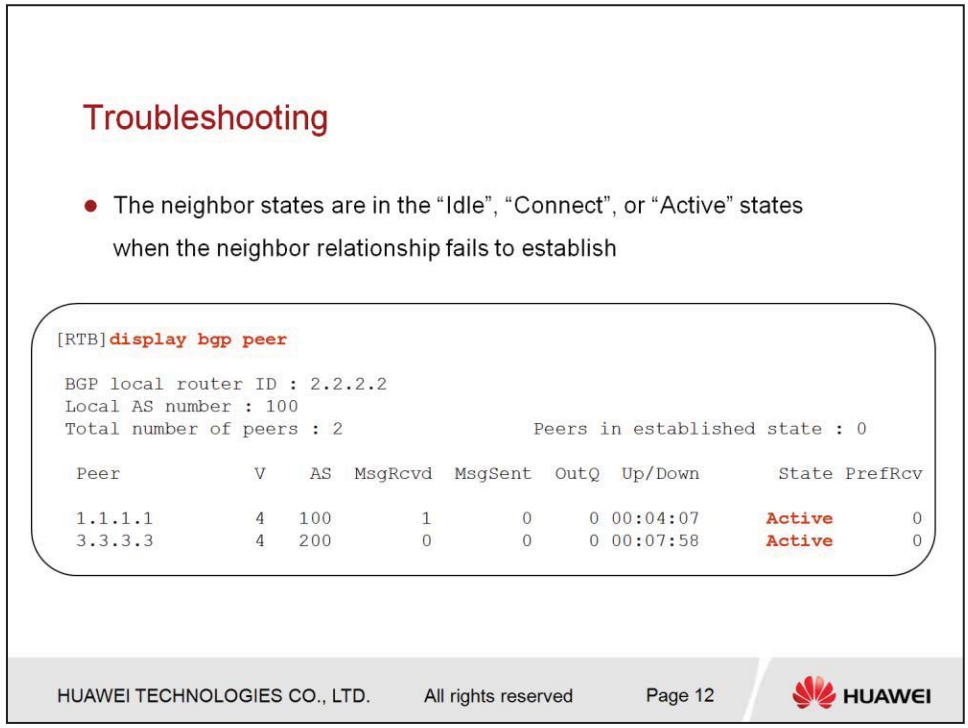
The symptoms of the failure are as below:

1) RTA and RTB are not able to form the IBGP neighbor relationship with each other through loopback interface

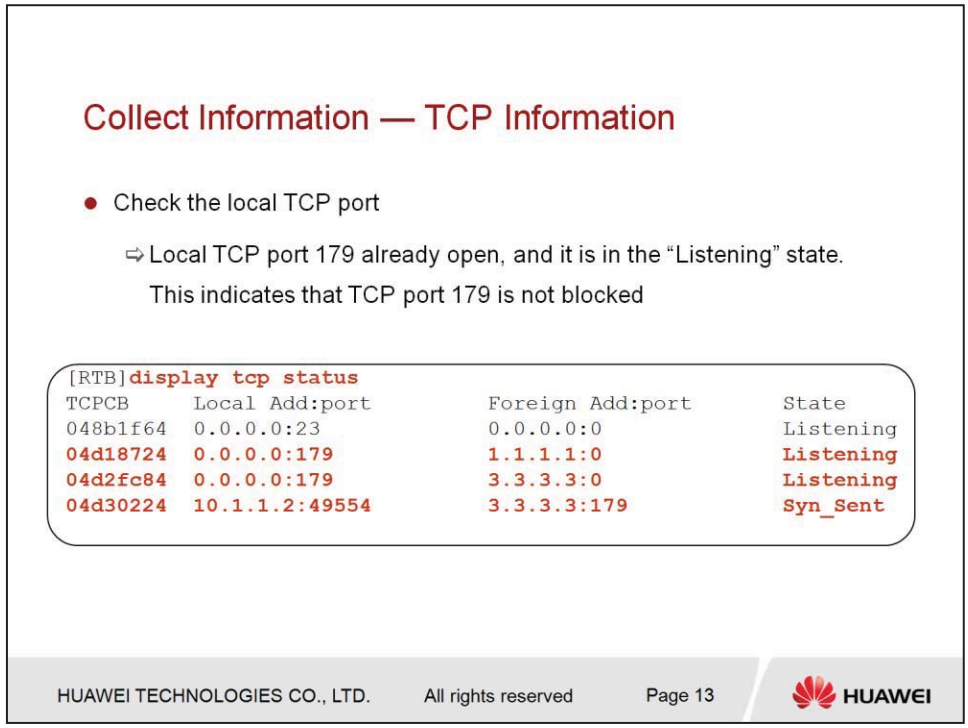
2) RTB and RTC are not able to form the EBGP neighbor relationship with each other through loopback interface

Based on the previous analysis that we have learnt, the reasons that might cause the failure in neighbor relationship establishment are as follow:

- TCP Port 179 is blocked
- No IP connectivity
- Parameters abnormality in the OPEN message
- Configuration error in EBGP/IBGP
- Failure in physical layer or others

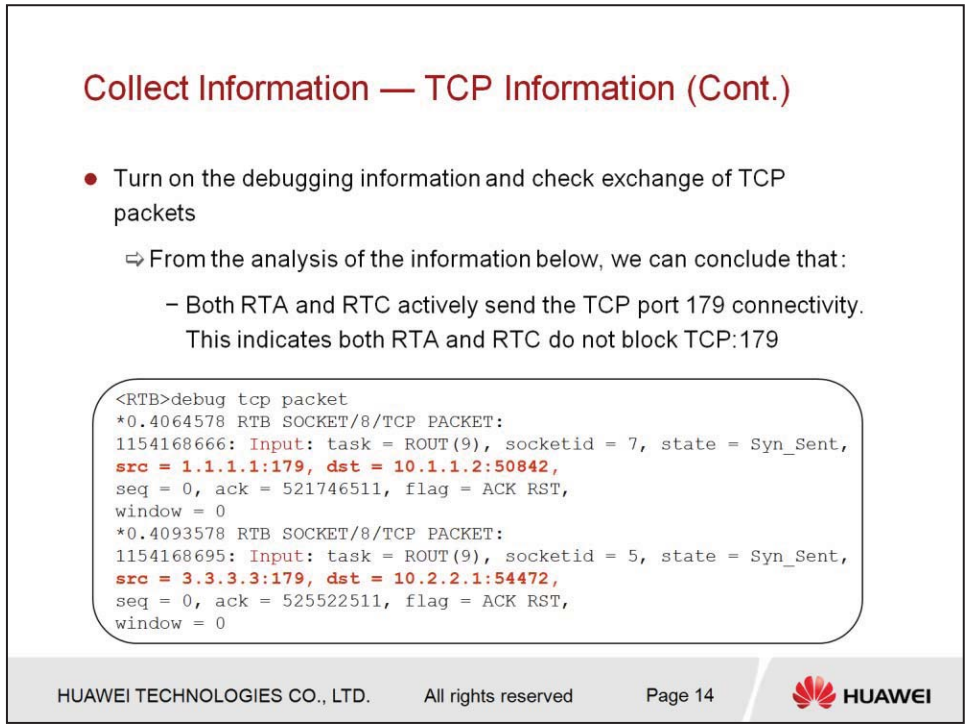


First of all, issue the "display bgp peer" command on RTB to check the BGP peer information. Please note that the state of both of its peers is “Active”. This indicates that the TCP connection has not been formed.



The “display tcp status” command can be used to check the TCP port status of the local router.

The TCP port 179 of the 2 lines as shown above is in the “Listening” state. This indicates that TCP port 179 is not blocked.



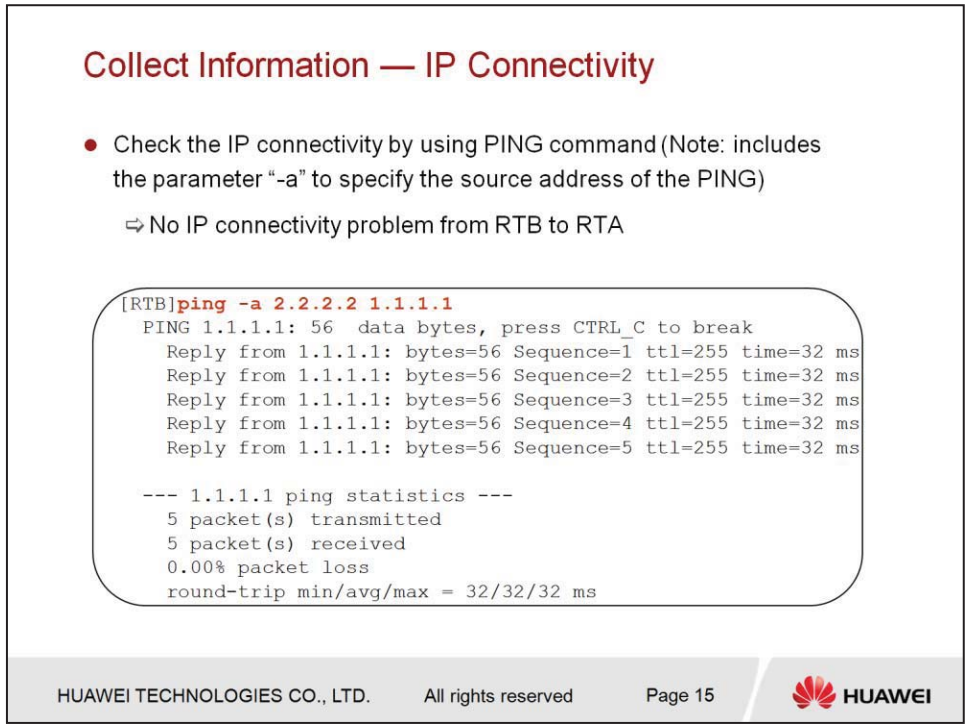
How can we determine whether the TCP port of the remote peer is being blocked or not by using command line?

We can use the debug command to turn on the debugging information. As shown in the diagram above, 2 TCP packets with port 179 are received from 1.1.1.1 and 3.3.3.3 respectively. This indicates that the port 179 of the remote peers 1.1.1.1

and 3.3.3.3 is not blocked.

After some careful analysis, we notice that the source address is 1.1.1.1 while the destination address is 10.1.1.2. This indicates that RTA forms the neighbor relationship with the physical interface of RTB through loopback address. Similarly,

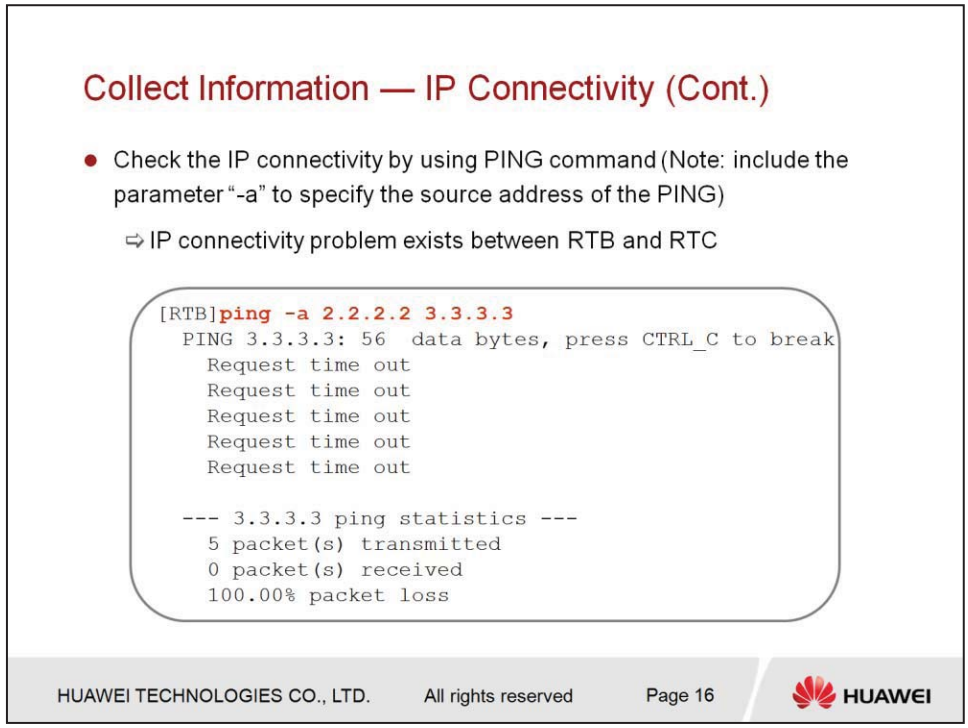
RTC forms the neighbor relationship with physical interface of RTB through loopback address as well.



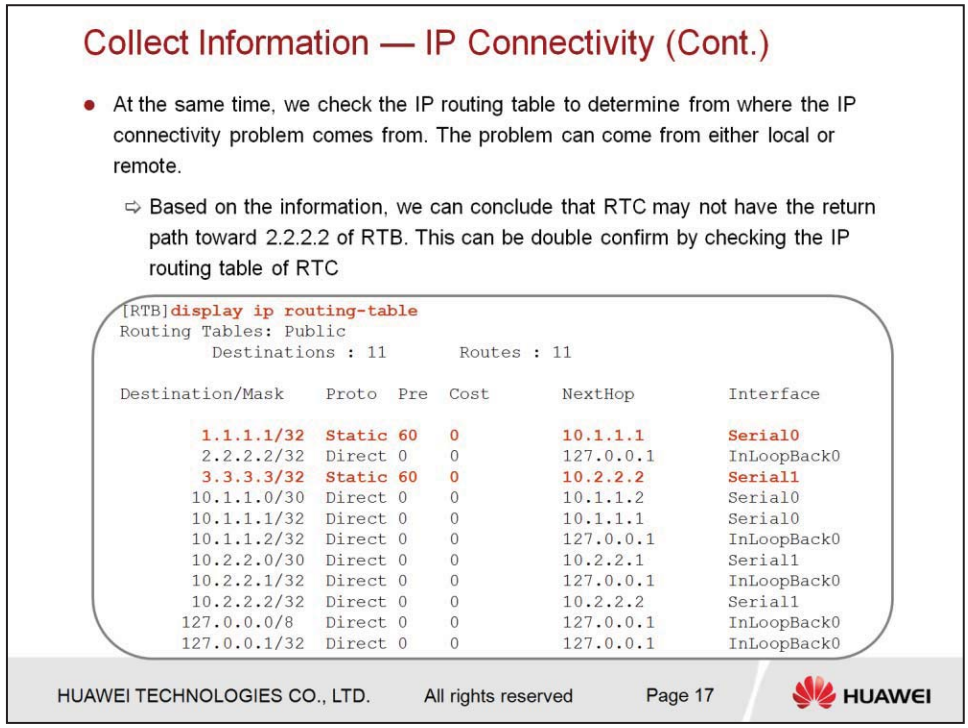
We have confirmed that the TCP port is not blocked. Next, we have to ensure that the IP connectivity exists between the BGP speakers. The common method to check the IP connectivity is by using "ping" command.

To check the reachability of 2 ends of the ports more precisely, we can specify the source address of the ping by including the "-a" parameter right after the "ping" command.

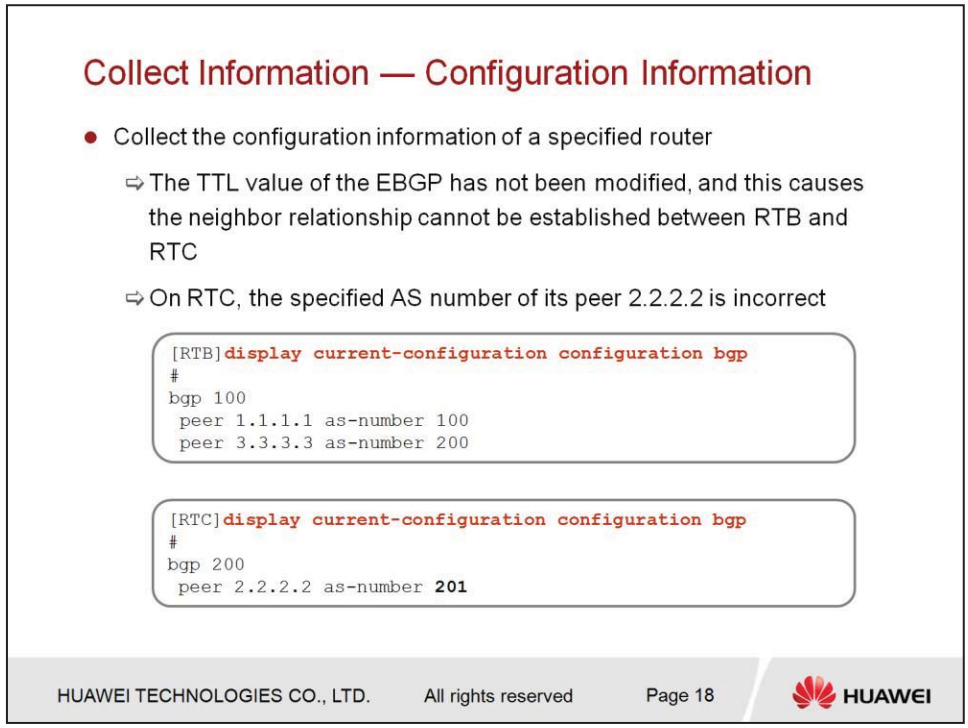
As shown in the diagram above, the IP connectivity is working well from source IP 2.2.2.2 to destination 1.1.1.1.



There is an IP connectivity problem between RTB and RTC. The ping result is time out from loopback address of RTB to loopback address of RTC.



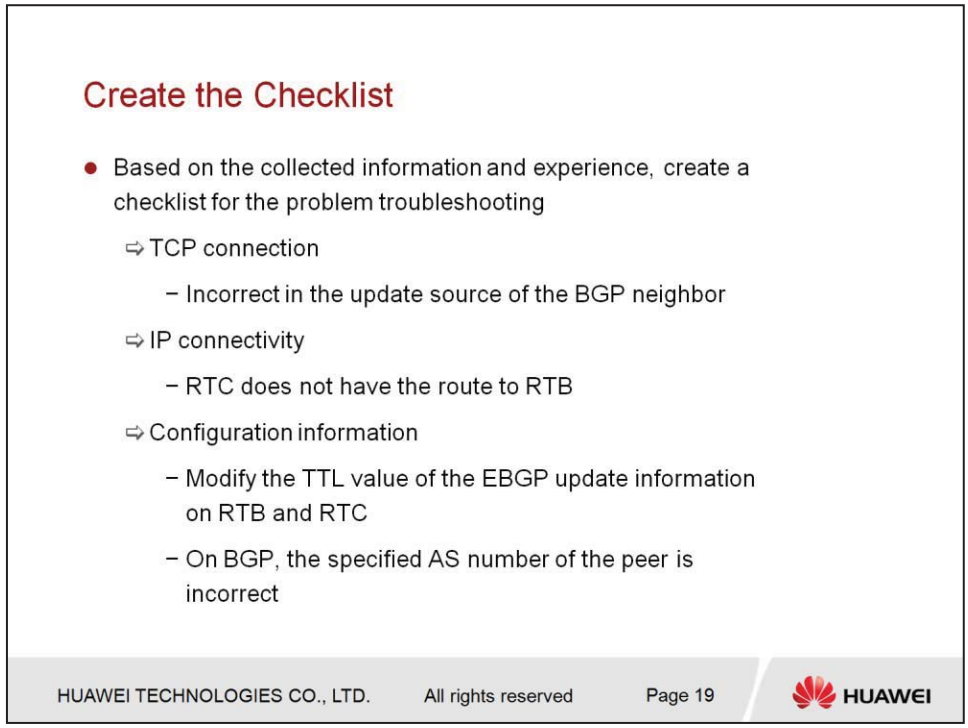
From the IP routing table of RTB, we can observe that there are 2 static routes with destination IP 1.1.1.1 and 3.3.3.3 respectively. This indicates the problem is not on RTB. This problem is due to RTC does not have the return path back to RTB. This can be double confirm by checking the IP routing table of RTC.



Next, check the configuration script of RTB and RTC.

The TTL value of the EBGP has not been modified, and this causes the neighborrelationship cannot be established between RTB and RTC.

On RTC, the specified AS number of its peer 2.2.2.2 is incorrect.



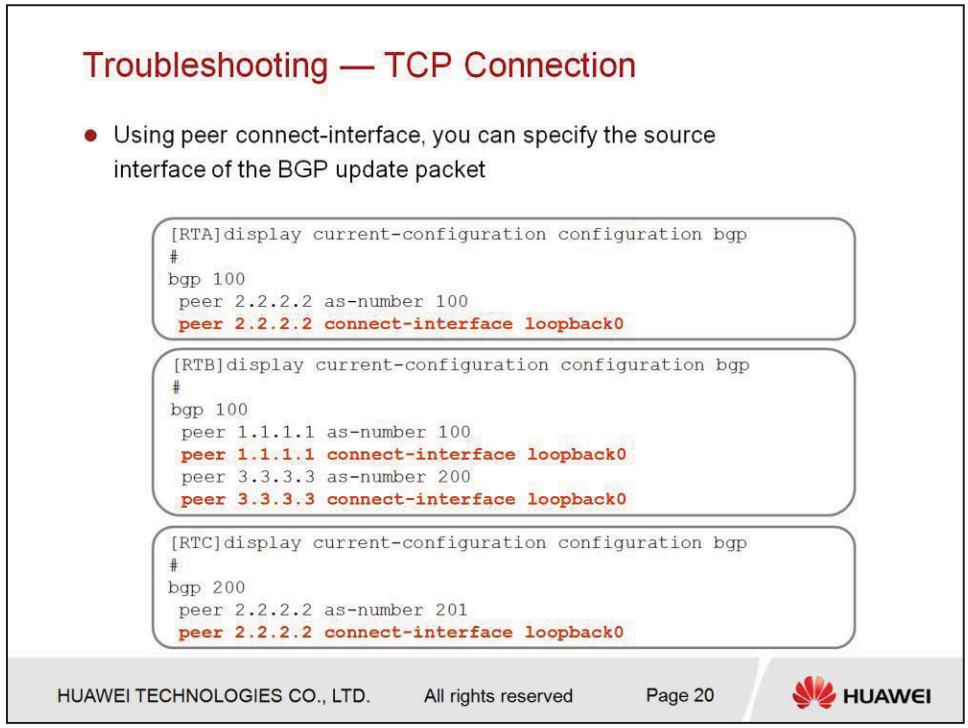
We can create a checklist base on the collected information.

TCP connection Incorrect in the update source of BGP neighbor

IP connectivity of RTC does not have the route to RTB

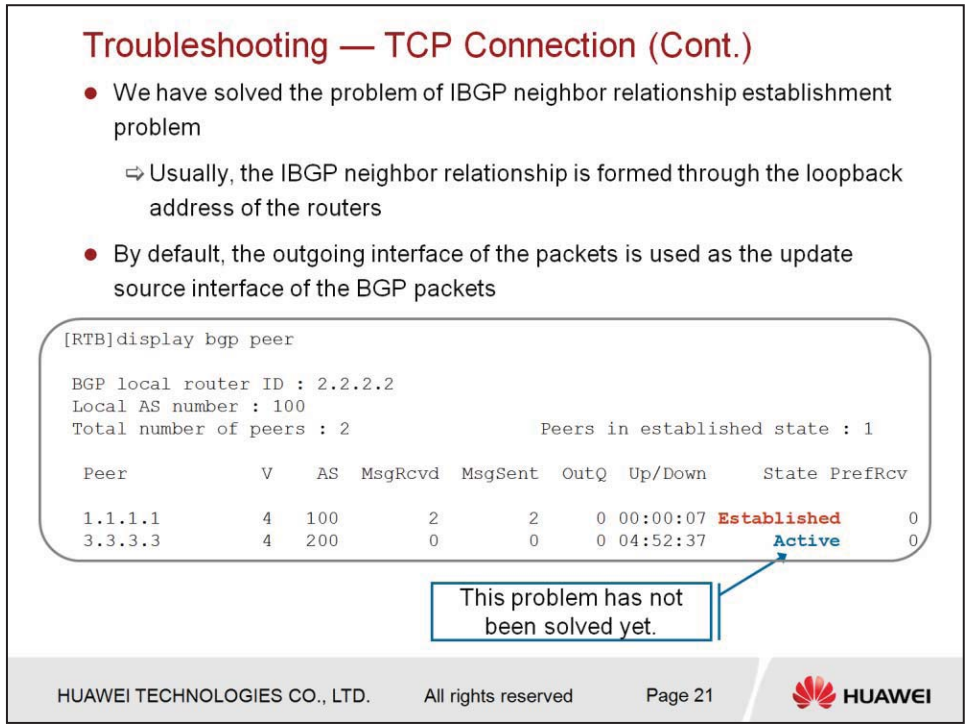
Configuration information Modify the TTL value of the EBGP update information on RTB and RTC.

On BGP, the specified AS number of the peer is incorrect.

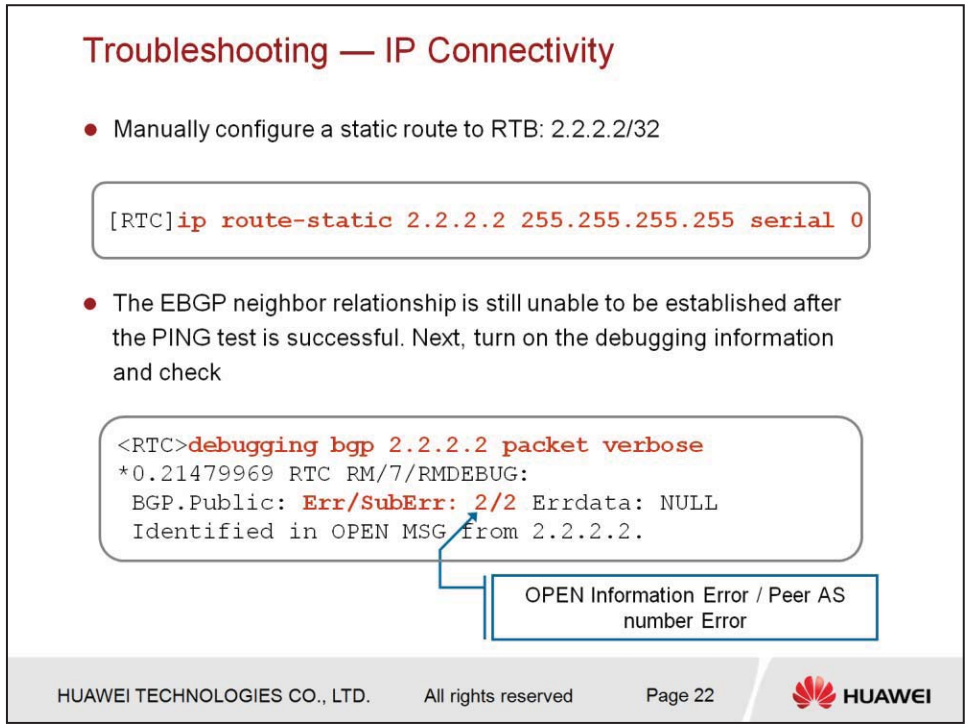


The TCP connection problem has to be solved first.

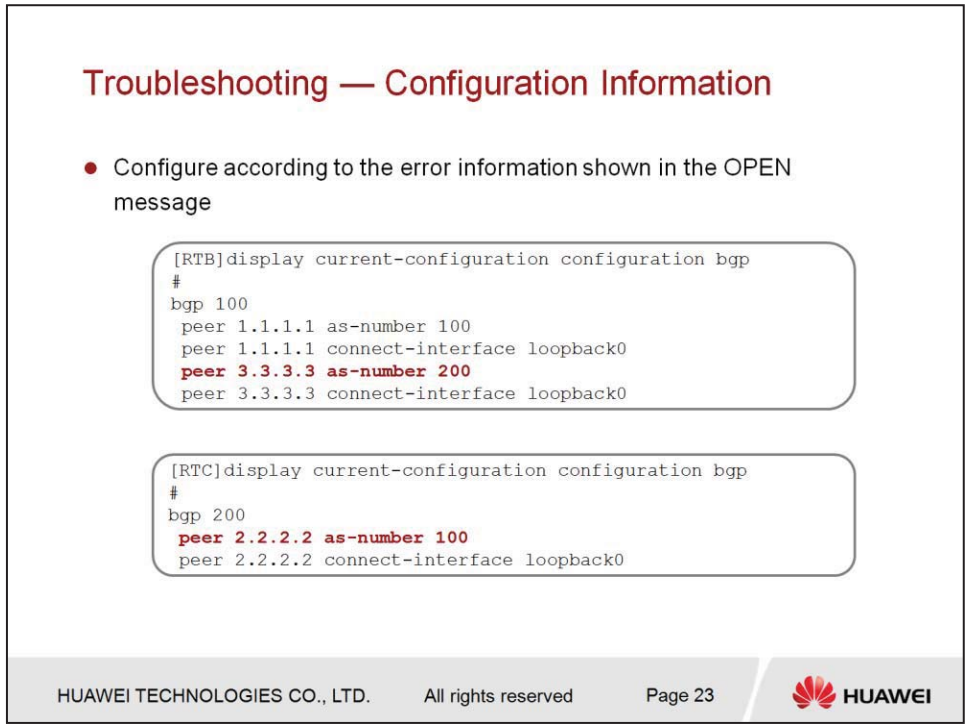
By default, the BGP peer relationship is formed by using the best interface IP address of the source. However, the neighbor relationship is formed by using loopback address in this case. Therefore, we need to modify it by using “peer connect-interface” command.



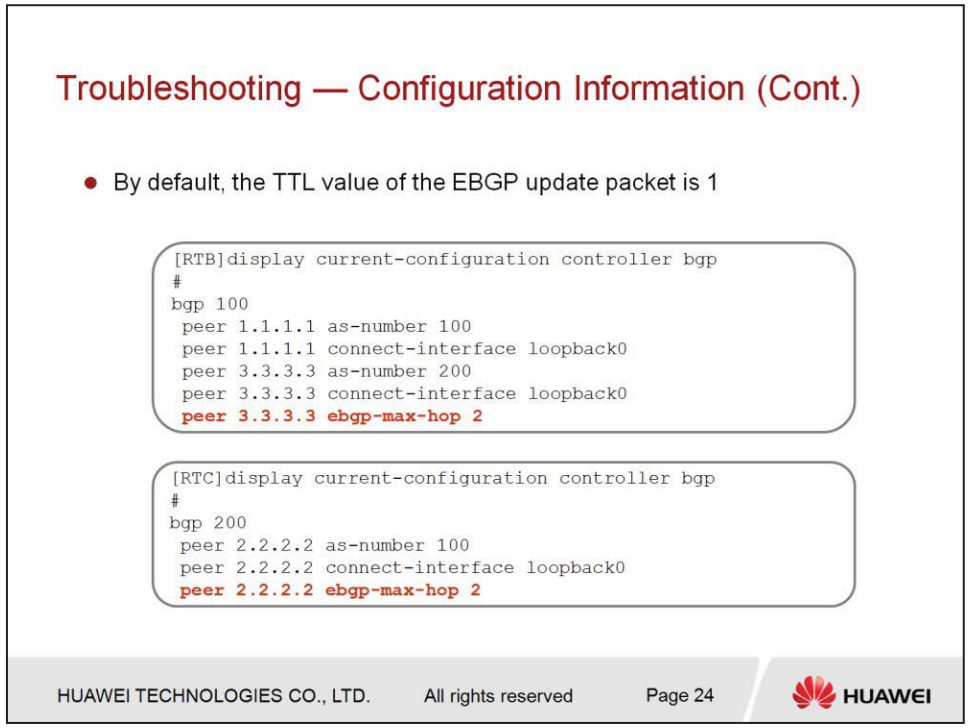
After the modification, IBGP neighbor relationship has been established successfully. However, the EBGP relationship is still in the “Active” state.



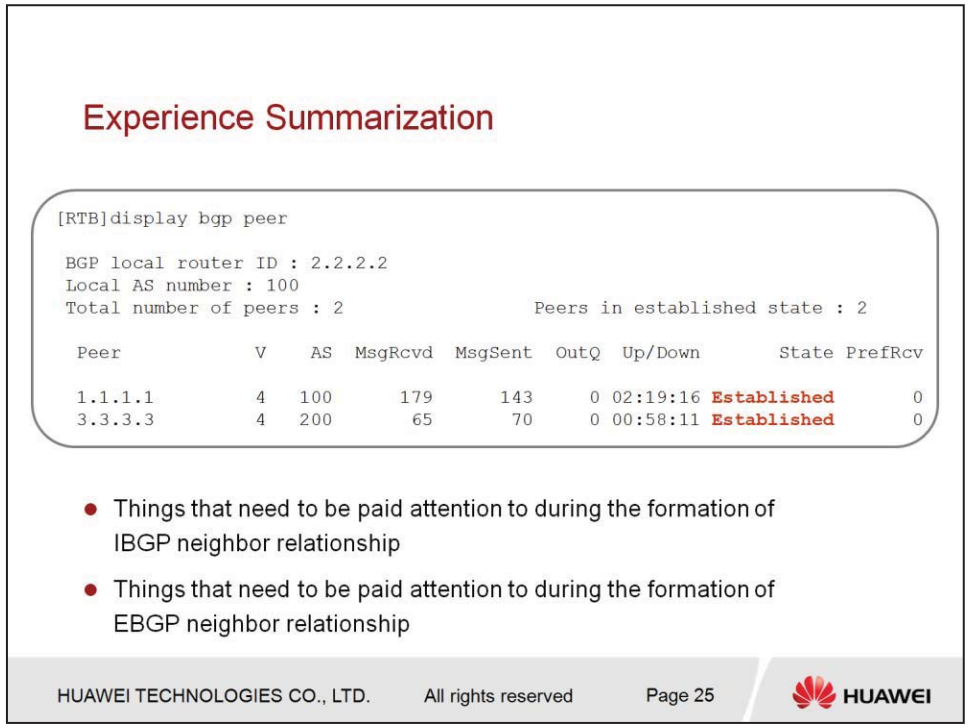
For the second step of the checklist, add a static route to destination 2.2.2.2 on RTC in order to solve the IP connectivity problem. However, the EBGP peer relationship still has not been established. From the debugging information, the specified AS number of the peer is incorrect.



After the modification on RTC, the problem still persists.



We can change the TTL value of the EBGP update packet to 2 by adding the “peer ebgp-max-hop” command. The problem is solved after the modification.



Base on the troubleshooting case in the previous example, we can conclude that:

• Take note to the common things below:

1. Is TCP port 179 being blocked?

2. IP Connectivity

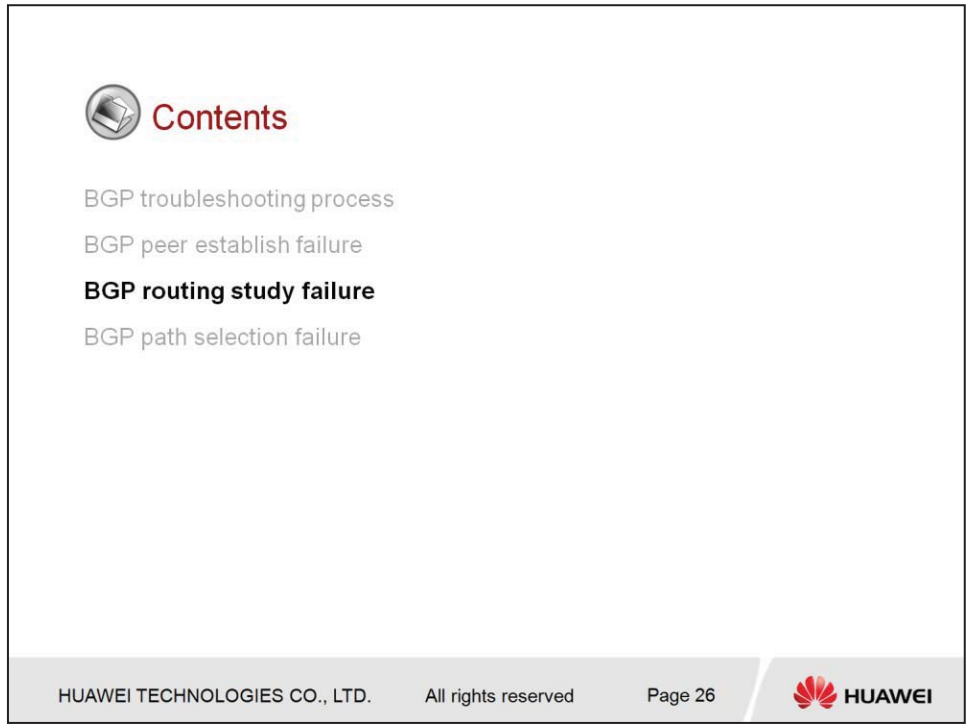
• For IBGP neighbor relationship establishment, take note to the following:

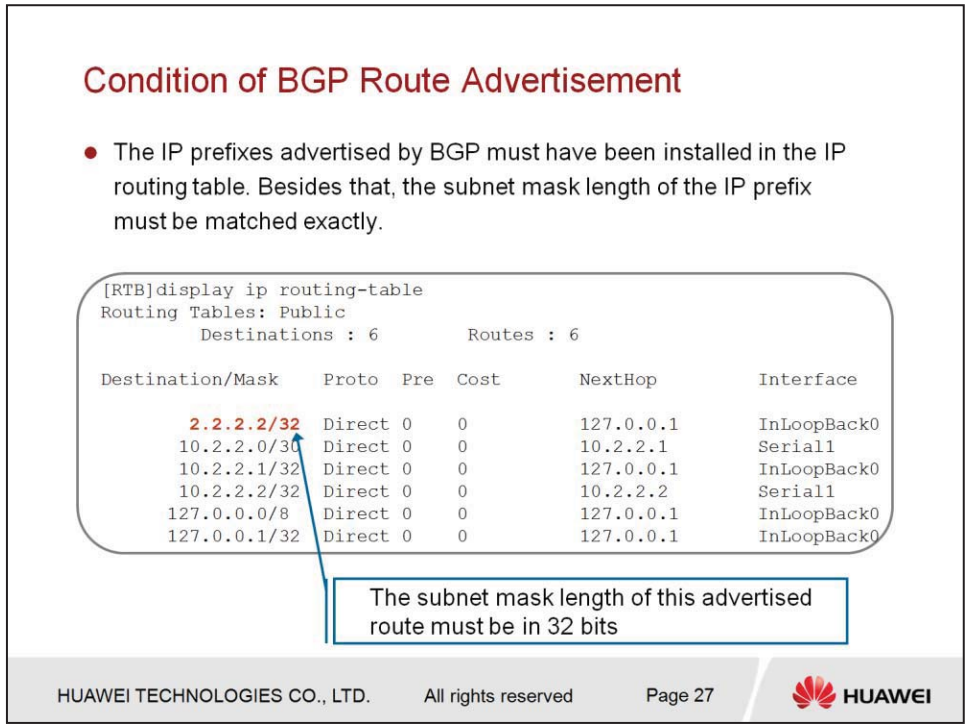
1. Specified the update source address

• For EBGP neighbor relationship establishment, take note to the following:

1. Problem with regarding to the multi-hop EBGP

2. Specified the update source address


Contents
BGP troubleshooting process
BGP peer establish failure
BGP routing study failure
BGP path selection failure
HUAWEI TECHNOLOGIES CO., LTD.
All rights reserved
Page 26
HUAWEI



BGP only advertises the IP prefixes that have been installed in the IP routing table.

Besides this, the subnet mask length of the IP prefix must be matched exactly.

As shown in the diagram above, route 2.2.2.2/32 advertised by BGP must be carried along with a 32 bit subnet mask, otherwise BGP will use the default subnet mask according to the class of that IP address.

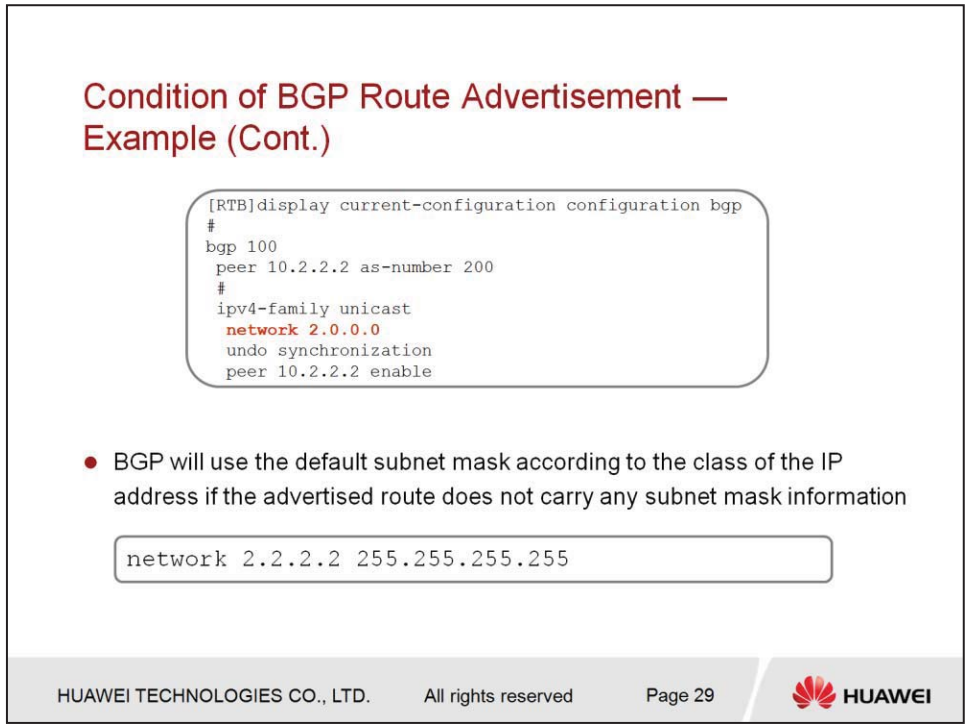
Condition of BGP Route Advertisement — Example



- The EBGP neighbor relationship between RTB and RTC is established successfully. However, RTC does not have the route towards 2.2.2.2/32 of RTB.

HUAWEI TECHNOLOGIES CO., LTD. All rights reserved

Problem symptom: The EBGP neighbor relationship between RTB and RTC is established successfully. However, RTC does not have the route towards 2.2.2.2/32 of RTB.



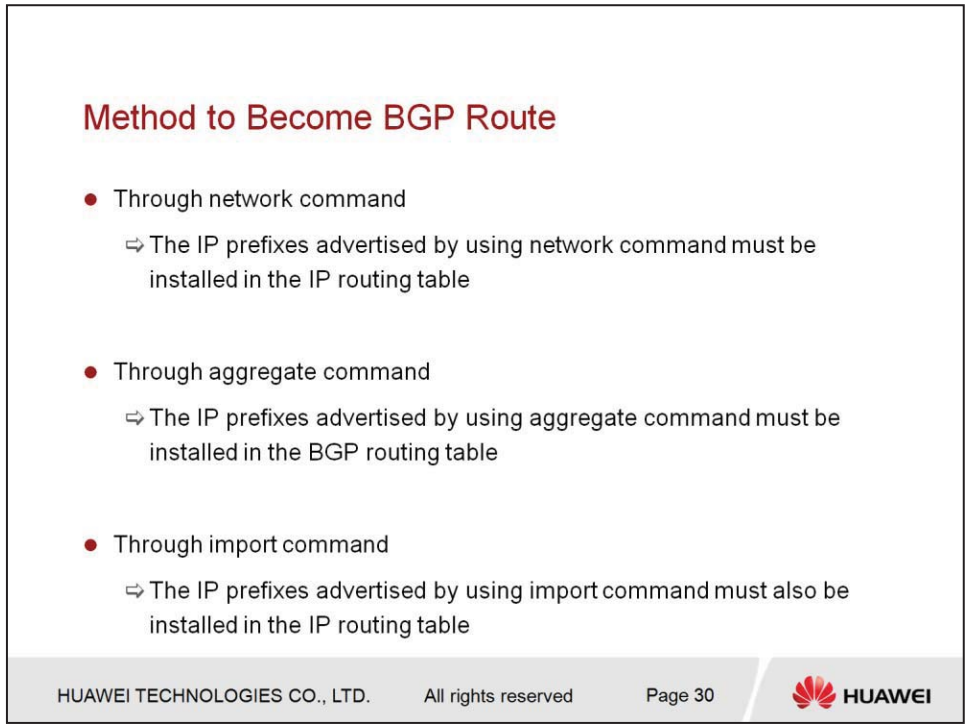
Check the BGP configuration by using “display cu configuration bgp” command.

From the configuration, we can observe that RTB has advertised routing information 2.0.0.0 on BGP 100 routing process. In this case, BGP will use the default subnet mask automatically. The 8 bits default subnet mask that has been assigned automatically is not match with the 32 bits subnet mask in the IP routing

table. As a result, BGP will not advertise this routing information entry to its peer.

This problem can be solved by advertising the route together with the subnet mask information. The modification can be made by using the following command:

"network 2.2.2.2 255.255.255.225".



How to become BGP route? There are 3 methods as follow:

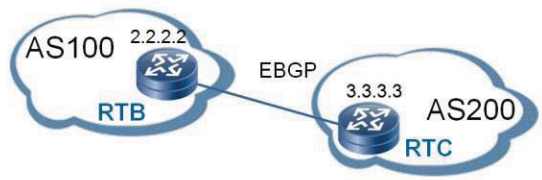
Through network command The IP prefixes to be advertised by using network command must be installed in the IP routing table. In addition, the subnet mask length of the IP prefix to b advertised and the subnet mask length of the route in the IP

routing table must be exactly matched. We can configure a static route with next hop point to the null interface if the IP prefix to be advertised is not available in the IP routing table.

Through aggregate command .The IP prefixes to be advertised by using aggregate command must be installed in the BGP routing table.

Through import command. The IP prefixes to be advertised by using import must be installed in the IP routing table.

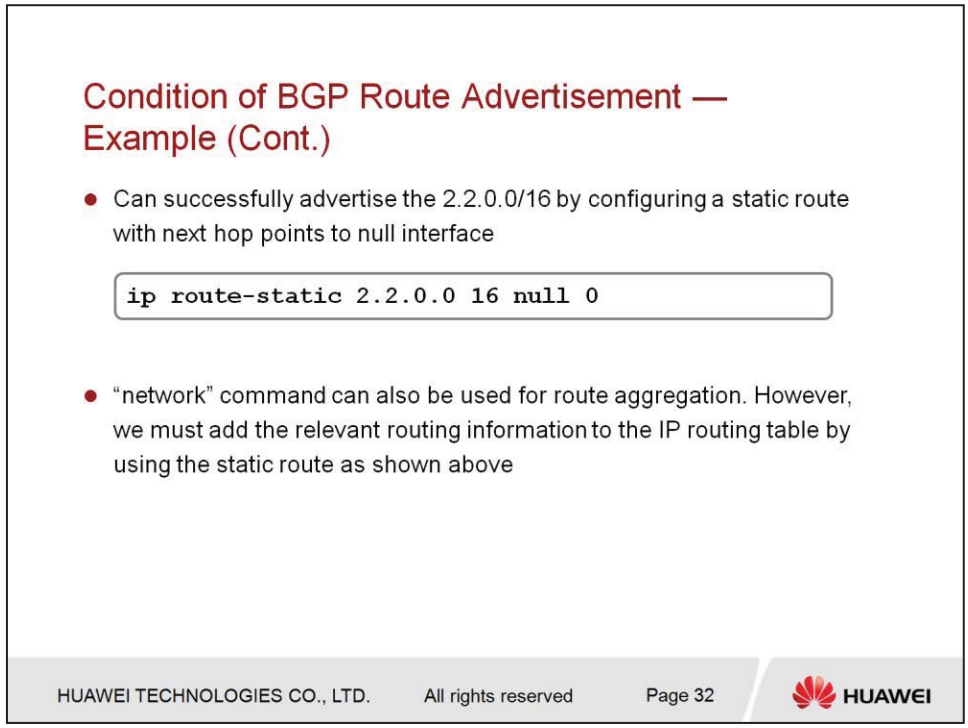
Method to Become a BGP Route — Example



- Import the route into BGP routing table via network, aggregate and import command respectively.
- RTB advertises network 2.2.0.0/16 to RTC. However, RTC does not receive the routing information about this network.

HUAWEI TECHNOLOGIES CO., LTD. All rights reserved

RTB and RTC from the previous example is used for the explanation. Assume that RTB does not want to advertise 2.2.2.2/32. Instead, RTB would like to advertise only 2.2.0.0/16.

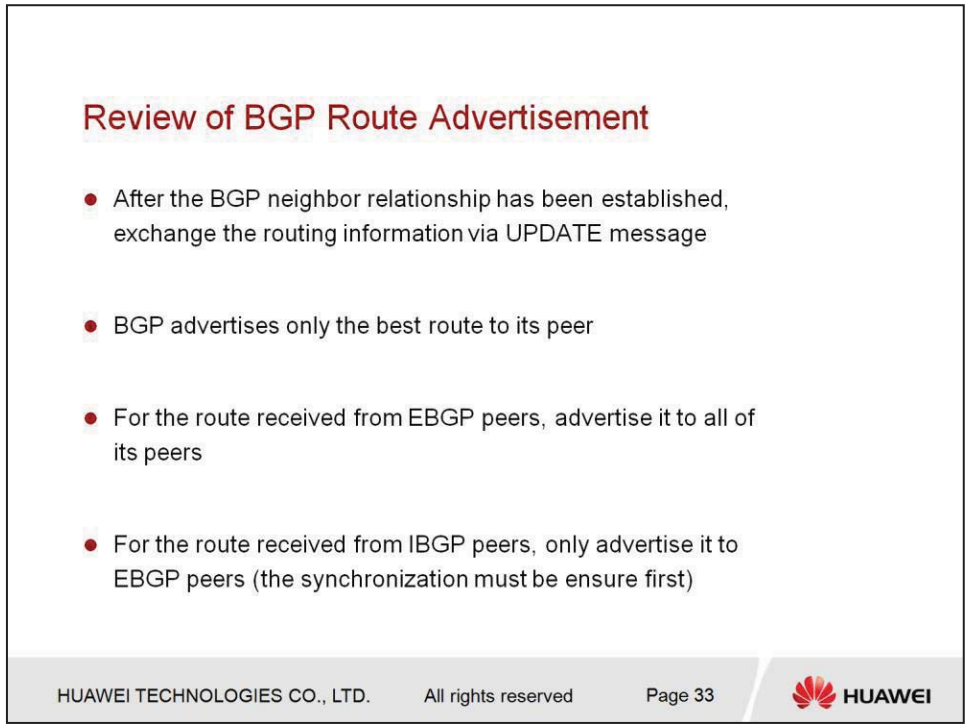


Configure a static route: ip route-static 2.2.0.0 16 null 0. The purpose of doing so is to add a routing entry with 16 bits subnet mask in the IP routing table. The entry 2.2.0.0/16 will be added to the IP routing table after the static route command is issued.

As a result, RTB can successfully advertise 2.2.0.0/16 to RTC.

On the other hand, we can also achieve the route aggregation result by using the "network" command together with the static route as shown above. Even though we can perform the aggregation by using the "network" command together with

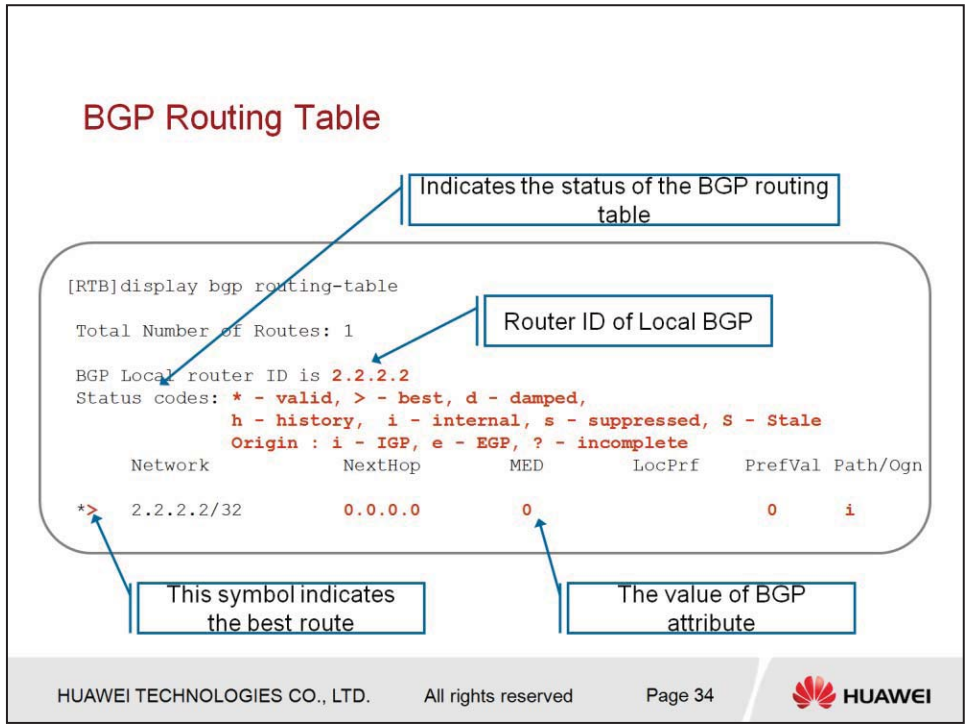
the static route, it introduces administrative burden on the configuration task for the network administrator. Besides, the functionality by using this method is not as perfect as the functionality offered by the "aggregate" command.



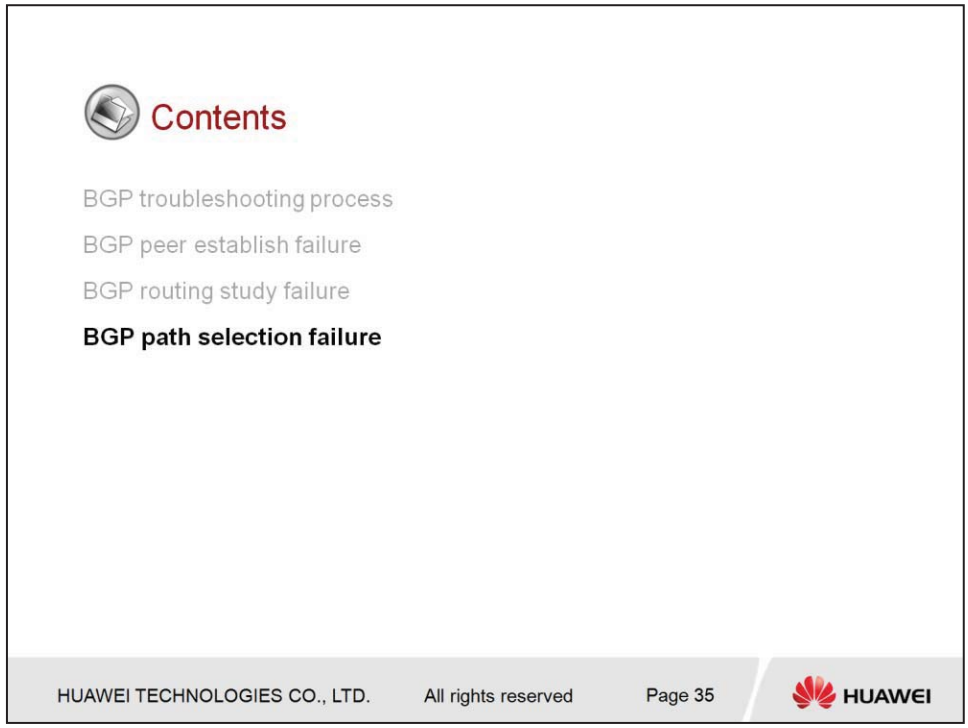
After the BGP neighbor relationship has been established, exchange the routing information via UPDATE message

BGP advertises only the best route to its peer

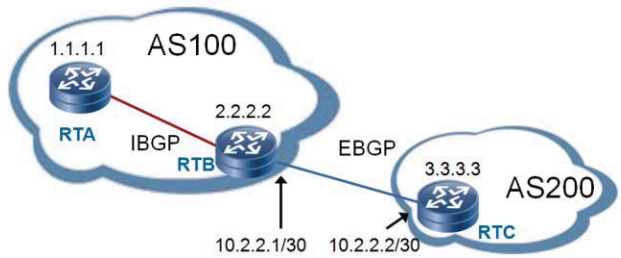
For the route received from EBGP peers, advertise it to all of its peers For the route received from IBGP peers, only advertise it to EBGP peers (the synchronization must be ensure first)



As shown in the diagram above, BGP routing table includes the BGP router ID, the meaning of the status code, routing entries and the value of the BGP attributes.


Contents
BGP troubleshooting process
BGP peer establish failure
BGP routing study failure
BGP path selection failure
HUAWEI TECHNOLOGIES CO., LTD.
All rights reserved
Page 35
HUAWEI

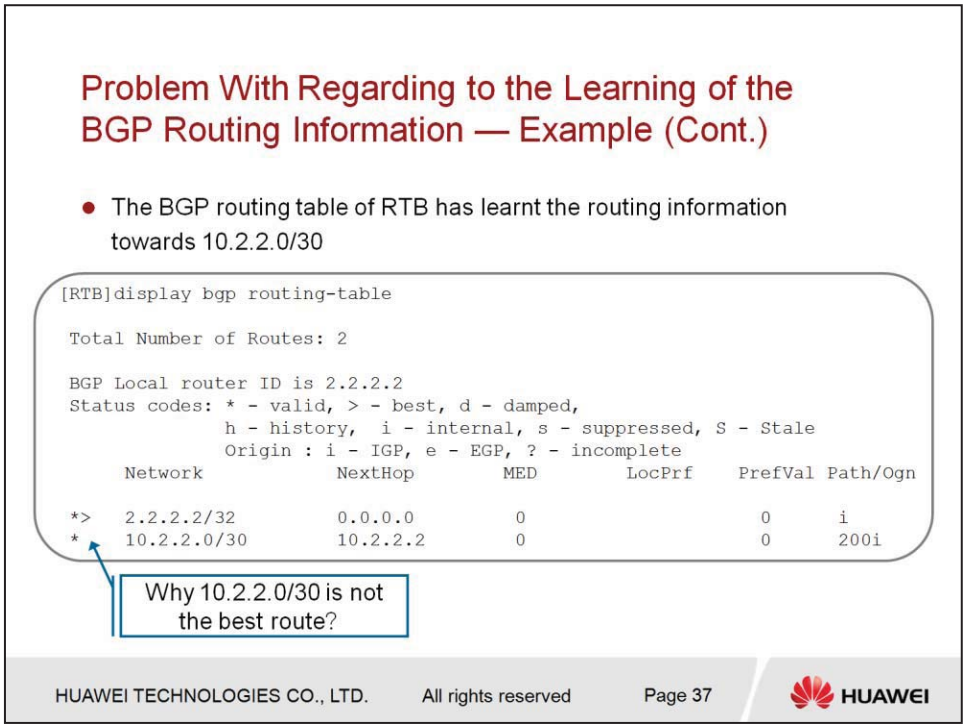
Problem With Regarding to the Learning of the BGP Routing Information — Example



- RTC advertise its directly connected route 10.2.2.0/30 to RTB. After RTB receives the routing update, it will not advertise it to RTA.

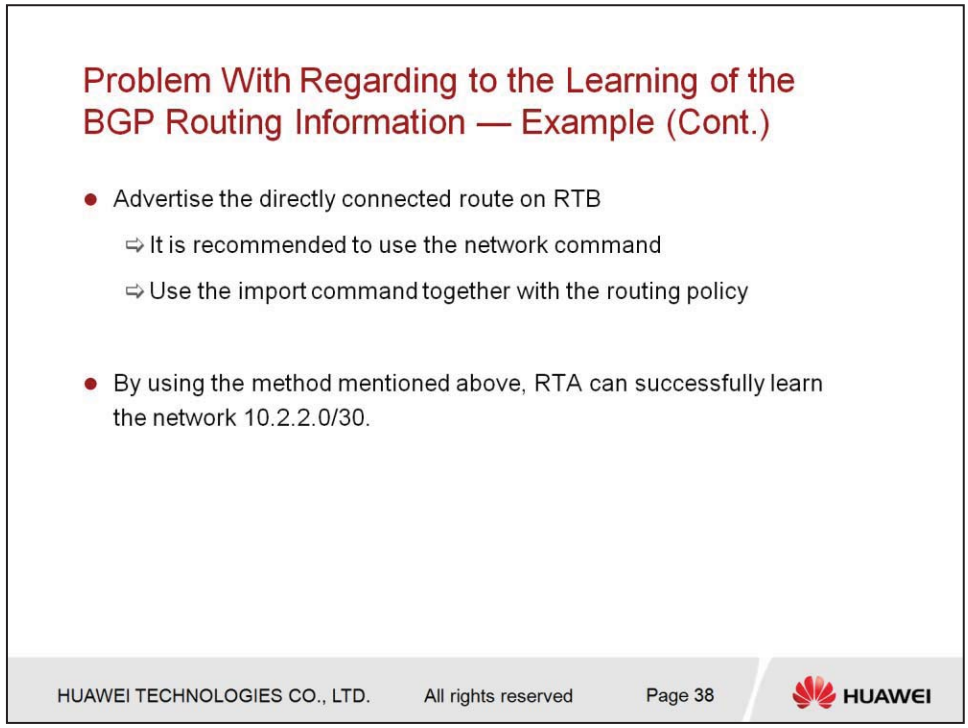
HUAWEI TECHNOLOGIES CO., LTD. All rights reserved

RTC advertise its directly connected route 10.2.2.0/30 to RTB. After RTB receives the routing update, it will not advertise it to RTA.



From the routing table of RTB, we can observe that RTB has learnt the routing information towards 10.2.2.0/30 from RTC. However, 10.2.2.0/30 is not selected as the best route. Therefore, BGP will not send the route 10.2.2.0/30 to IP routing table nor advertise it to other BGP peer.

Why the route 10.2.2.0/30 learnt from RTC is not selected as the best route? The reason is fairly simple. Network 10.2.2.0/30 is used to connect the link in between RTB and RTC. This implies that there is a direct route with preference value 0 on RTB. The direct route has the highest preference value in the routing table. As a result, RTB will not use the BGP routing update learnt from RTC.



The solution is very simple, we can re-advertise the route 10.2.2.0/30 by using “network” command on RTB. As a result, RTA can successfully learn the network 10.2.2.0/30.

Problem With Regarding to the Learning of the BGP Routing Information — Summarization

- Take note to the following when BGP advertise the routing information:
  - ⇨ The routing information advertised by using network command must be installed in the IP routing table
  - ⇨ The routing information advertised by using aggregate command must be installed in the BGP routing table
  - ⇨ The route which is redistributed into the BGP routing table by using the import command is often used together with the routing policy
- Full mesh IBGP connection
  - ⇨ Reduce the number of IBGP sessions by using route reflector and confederation

HUAWEI TECHNOLOGIES CO., LTD. All rights reserved

Review of BGP Route Selection process

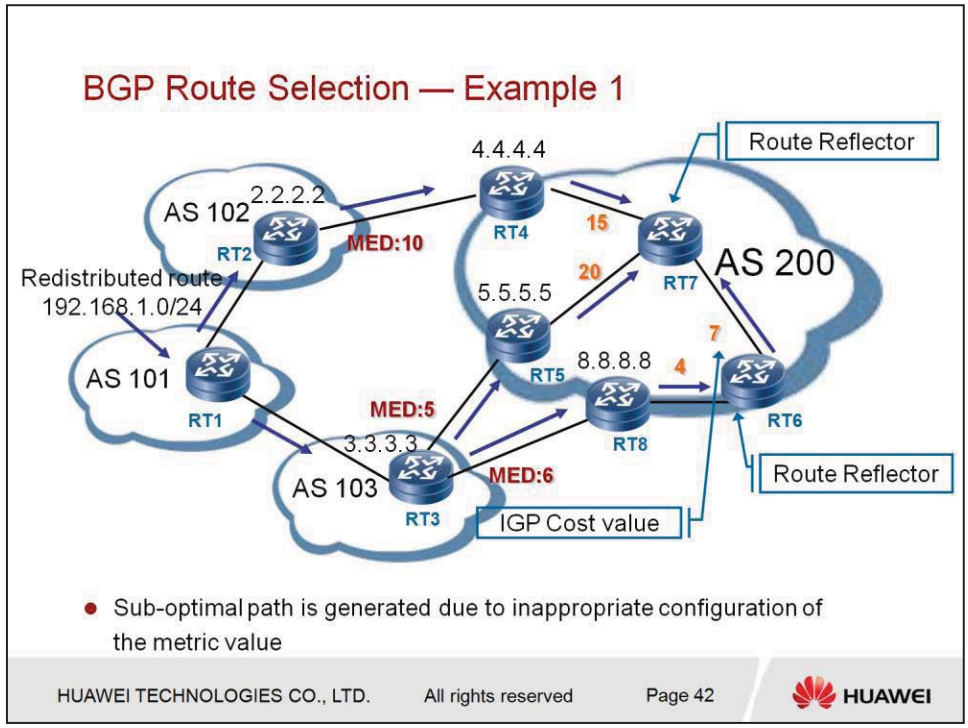
- If the next hop of this route is unreachable, then ignore this route
- Select the route with highest Preferred-Value
- Select the route with highest local preference
- Prefer the route that was originated locally on the router. That is, prefer a route that was learned from an IGP on the same router (If the local preference is the same).
- Prefer the route with the shortest AS_PATH.
- BGP compares the origin attribute of the route and selects the route with the lowest origin attribute code: IGP is lower than EGP, which is lower than Incomplete.
- Select the route with the lowest MED value. This comparison is done only if the AS number is the same for all the routes being considered.

HUAWEI TECHNOLOGIES CO., LTD. All rights reserved

Review of BGP Route Selection process (Cont.)

- Prefer EBGP routes over IBGP routes.
- Prefer the route with the shortest path to the BGP NEXT_HOP. This is the route with the lowest IGP metric to the next-hop router.
  - ⇨ When all the above are identical, they are equivalent route and can be used for load balancing
    - Note: AS_PATH must be the same
    - The last 3 route selection rules can be ignored when load balancing is implemented.
- Selects the route with the shortest Cluster_ID
- Prefer the route with smaller Originator ID. Select the route with smaller router ID if Originator ID is unavailable.
- Prefer the route with smaller peer IP address.

HUAWEI TECHNOLOGIES CO., LTD. All rights reserved



The sub-optimal path is generated due to inappropriate configuration of the metric value. Among others metric value that can affect the BGP route selection are AS_PATH, MED and IGP cost value.

RT7 has received 3 update information from RT4, RT5 and RT6 respectively.

These 3 update information carry different metric value as indicated in the diagram above. In addition, the “next-hop-local” command is issued on RT4, RT5 and RT8 respectively.

Process of Route Selection (Case 1)

- Assume that RT7 has received the update information from either RT5 or RT6 first, follow by the update information from RT4

Routing Table of RT7

Path	BGP Next Hop	AS-PATH	MED	IGP Metric
1	RT5	103 101	5	20
2	RT8	103 101	6	11
3	RT4	102 101	10	15

HUAWEI TECHNOLOGIES CO., LTD. All rights reserved

BGP will compare the routing information that has been received one by one according to their receiving order. The path that has been selected by BGP at the end is regarded as the best path.

First of all, the comparison is made between path 1 and path 2. Since the AS_PATH for path 1 and path 2 are the same. So, we will choose the path with lower MED value. According to the BGP route selection process, path 1 is better than path 2 because its MED value is lower.

Next, the comparison is made between path 1 and path 3. By default, BGP will not compare the MED value when the AS_PATH of path 1 and path 3 is the same. In the case, BGP will compare the IGP metric for path1 and path 3. At the end, path 3 is selected as the best path since it has the lower IGP metric.

Process of Route Selection (Case 2)

- Assume that RT7 has received the update information from RT4 first, follow by the update information from RT6 and RT5

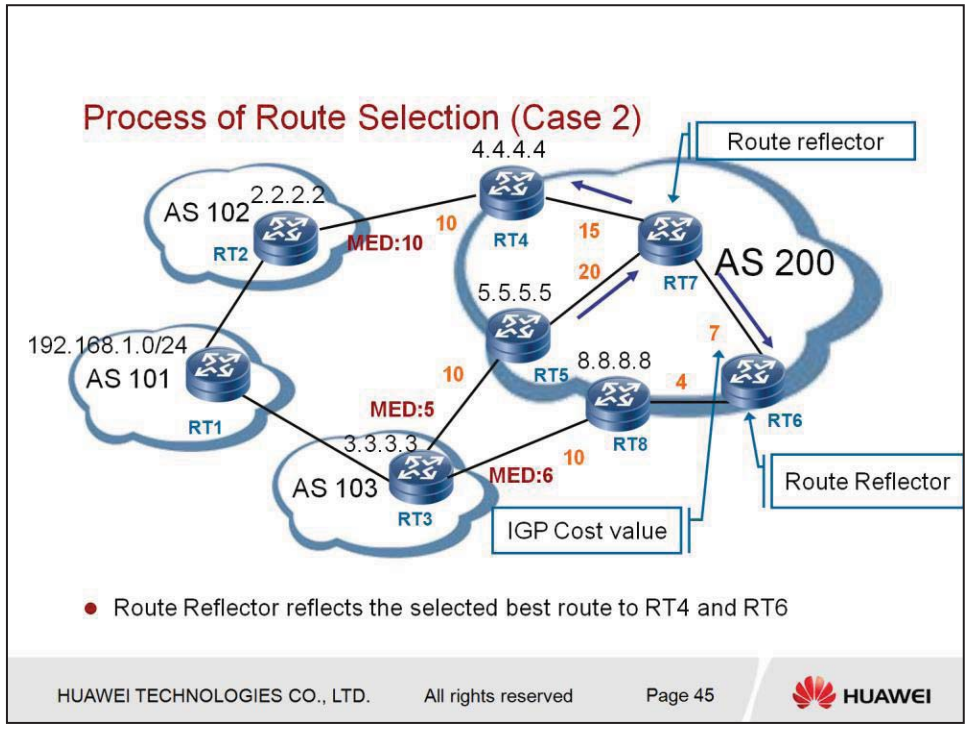
Routing Table of RT7

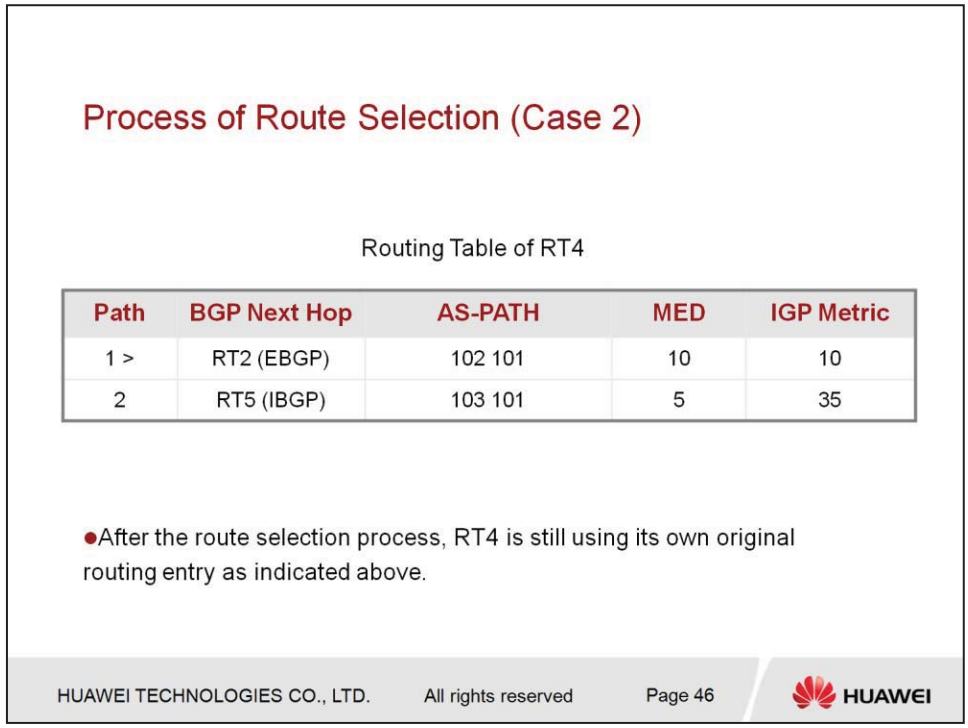
Path	BGP Next Hop	AS-PATH	MED	IGP Metric
1	RT4	102 101	10	15
2	RT8	103 101	6	11
3	RT5	103 101	5	20

HUAWEI TECHNOLOGIES CO., LTD. All rights reserved HUAWEI

Similarly, make the comparison between path 1 and path 2 according to the route selection process mentioned in the previous slide. In this case, path 2 has been selected as the best route. Next, the comparison is made between path 2 and path 3. Ultimately, path 3 has been selected as the best route.

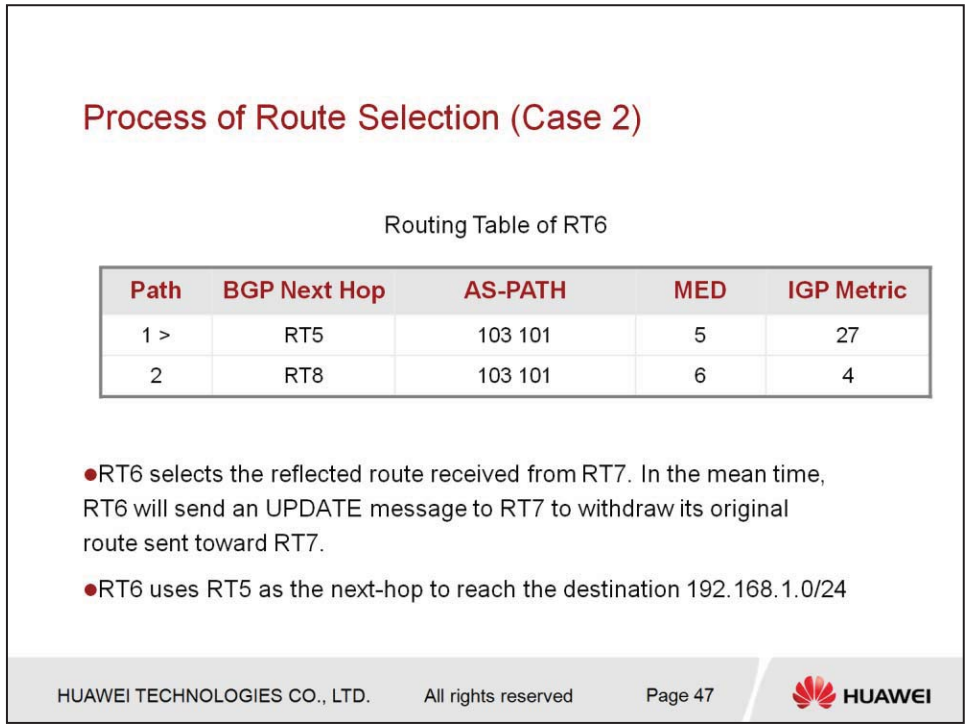
Since RT7 is the route reflector, the selected best route will be reflected to RT4 and RT6.


Process of Route Selection (Case 2)
Route reflector
4.4.4.4
AS 102
2.2.2.2
RT2
MED:10
10
RT4
15
RT7
AS 200
20
5.5.5.5
RT5
7
8.8.8.8
4
RT6
192.168.1.0/24
AS 101
RT1
10
MED:5
3.3.3.3
RT8
10
MED:6
AS 103
RT3
Route Reflector
IGP Cost value
Route Reflector reflects the selected best route to RT4 and RT6
HUAWEI TECHNOLOGIES CO., LTD.
All rights reserved
Page 45
HUAWEI



Both RT4 and RT6 receive the reflected routing information from RT7. According to the BGP route selection process, BGP prefer EBGP routes over IBGP routes.

Therefore, RT4 is still using its original routing entry which learnt from RT2 instead of the reflected route from RT5.



RT6 receives the reflected route from RT7 and a route from RT8. RT6 will make the comparison between these 2 routes. Since the AS_PATH between these 2 routes are the same, RT6 will check the MED value of these routes. The route with lower MED value is preferred over the route with higher MED value. As a result,

the reflected route from RT7 has been selected as the best route by RT6.

Once the path 1 has been selected as the best route, RT6 will send an UPDATE message to withdraw its original route sent toward RT7.

Process of Route Selection (Case 2)

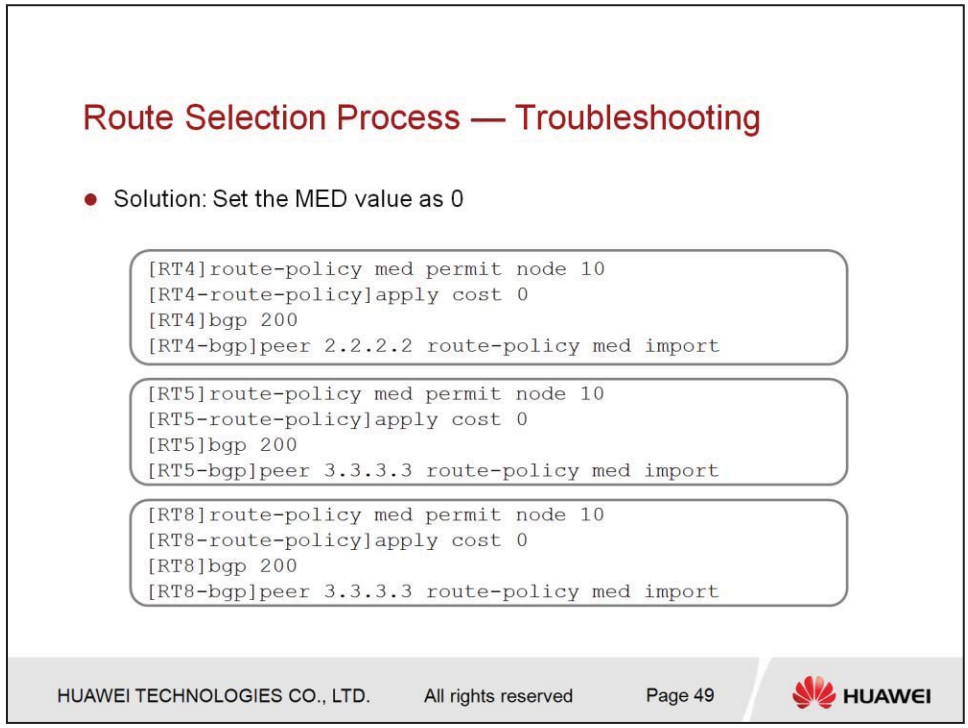
- After receiving the UPDATE message from RT6, RT7 withdraws the route from its routing table.

Routing Table of RT7

Path	BGP Next Hop	AS-PATH	MED	IGP Metric
1 >	RT5	103 101	5	20
2	RT4	102 101	10	15
~~3~~	~~RT8~~	~~103 101~~	~~6~~	~~11~~

HUAWEI TECHNOLOGIES CO., LTD. All rights reserved

Table above shows the routing table of RT7 after the route received from RT6 has been withdrawn



From the example above, we can observe that the modified MED value of the route is sent from external As to local AS. By default, the local BGP will not modify the MED value of the route received from the external AS. This will cause unpredictable result during the route selection process. How to solve this problem?

We can set the MED value of the received route as 0 so that the MED value will not affect the BGP route selection. As a result, IGP cost will be used to affect the BGP route selection instead of MED. This is to ensure the BGP uses the nearest exit point to transmit the data to external AS.

Please note that this is only one of the solutions. Since BGP is a policy tool, it has many methods that be used to affect the BGP route selection. We will not discuss each of the methods in detail here.

Route Selection Process — Troubleshooting

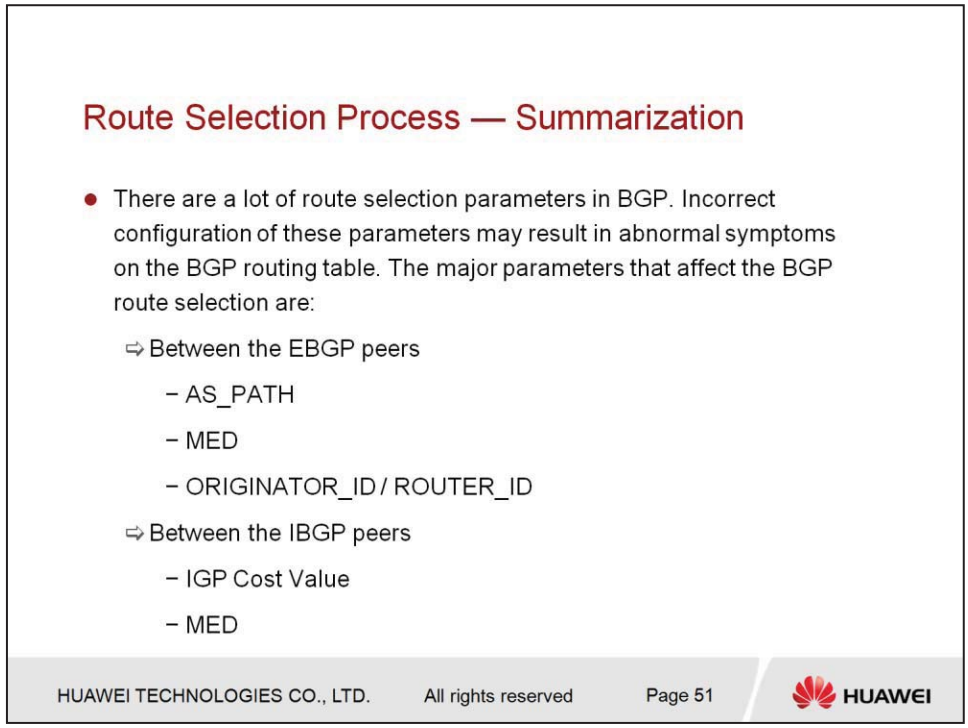
```
[RT7]display bgp routing-table

 Total Number of Routes: 3

 BGP Local router ID is 10.1.1.1
 Status codes: * - valid, > - best, d - damped,
               h - history,  i - internal, s - suppressed, S - Stale
               Origin : i - IGP, e - EGP, ? - incomplete
      Network            NextHop        MED        LocPrf    PrefVal Path/Ogn

 *>i  192.168.1.0/24     8.8.8.8                   100         0      103 101i
 * i                     4.4.4.4                   100         0      102 101i
 * i                     5.5.5.5                   100         0      103 101i
```

HUAWEI TECHNOLOGIES CO., LTD. All rights reserved



There are a lot of route selection parameters in BGP. Incorrect configuration of these parameters might result in abnormalities in the BGP routing table. The major parameters that affect the BGP route selection are:

Between EBGP peers

AS_PATH

MED

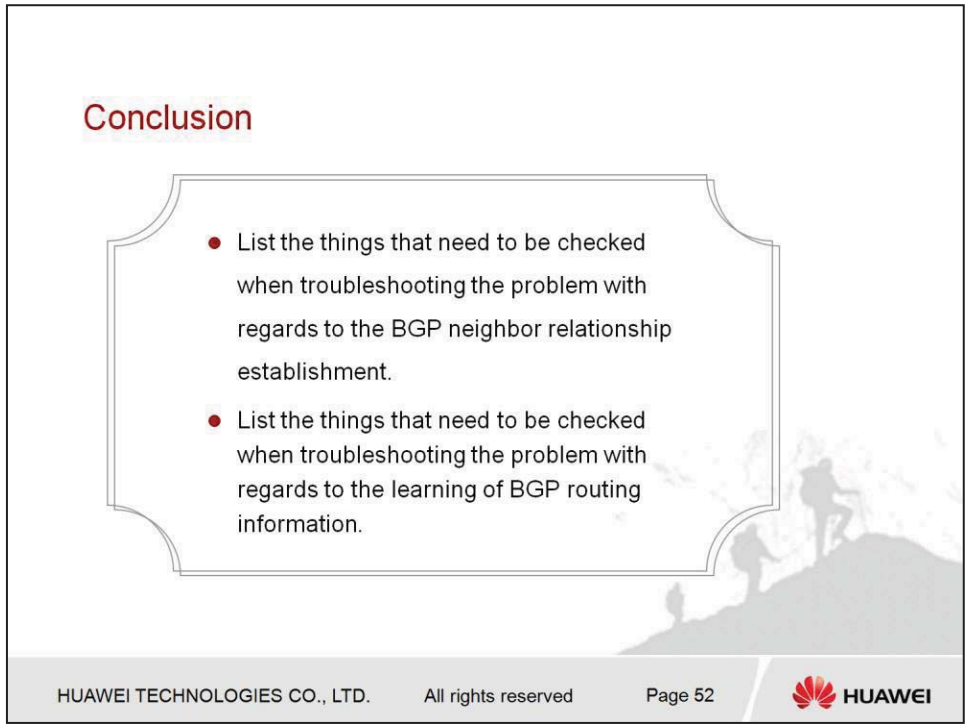
ORIGINATOR_ID / ROUTER_ID

Between IBGP peers

IGP Cost value

MED

In addition to the parameters mentioned above, there are other BGP parameters that can also affect the BGP route selection, for example: community attribute, Local_Pref and so on.



1. List down the things that need to be checked when troubleshoot the problem with regarding to the BGP neighbor relationship establishment.

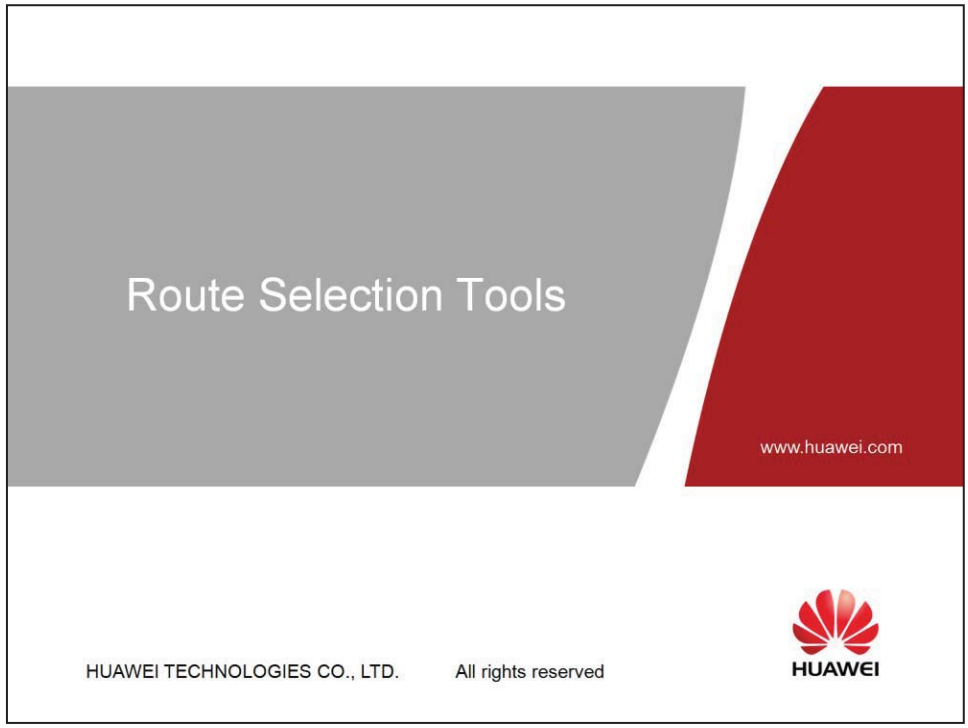
A: When troubleshooting the problem with regards to the establishment of the BGP neighbor relationship, we need to check whether or not the TCP port 179 is being blocked. Besides this, we must check whether the existence of IP connectivity between the neighbors. We also need to pay attention to the problem with regards to the EBGP/IBGP neighbor relationship establishment. For example, the problems with regards to the multi-hop EBGP and EBGP/IBGP update source. In addition, we need to pay attention to the parameters in the OPEN message such as the correctness of the AS number and the configuration of the ROUTER ID. The problems of incorrect configuration due to human error might contribute to the failure of establishment of the BGP neighbor relationship.

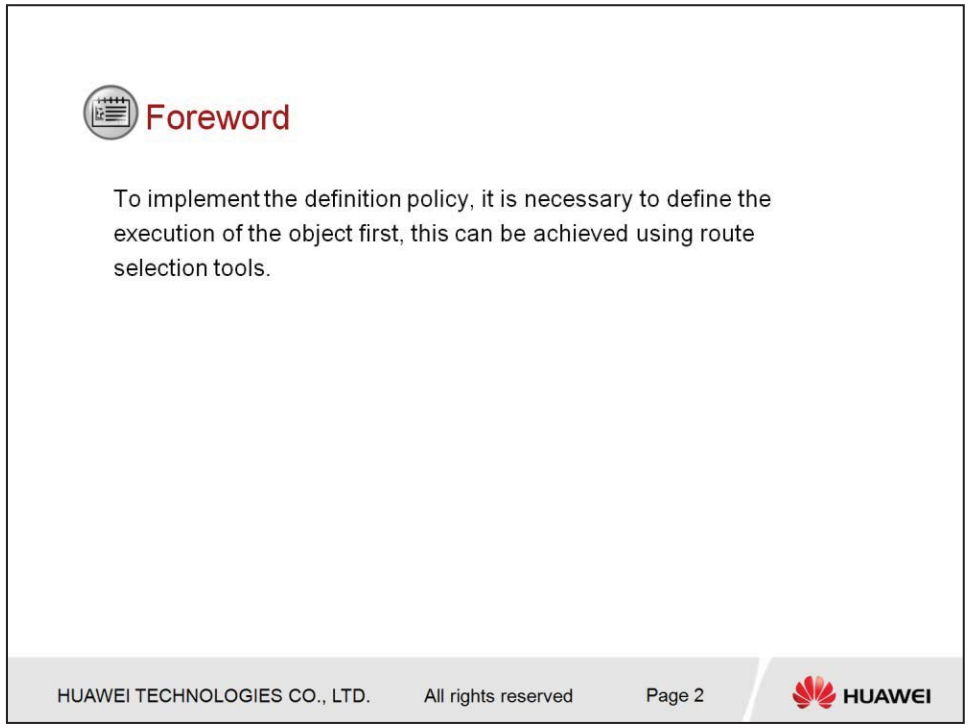
2. List the things that need to be checked when troubleshooting the problem with regards to the learning of BGP routing information.

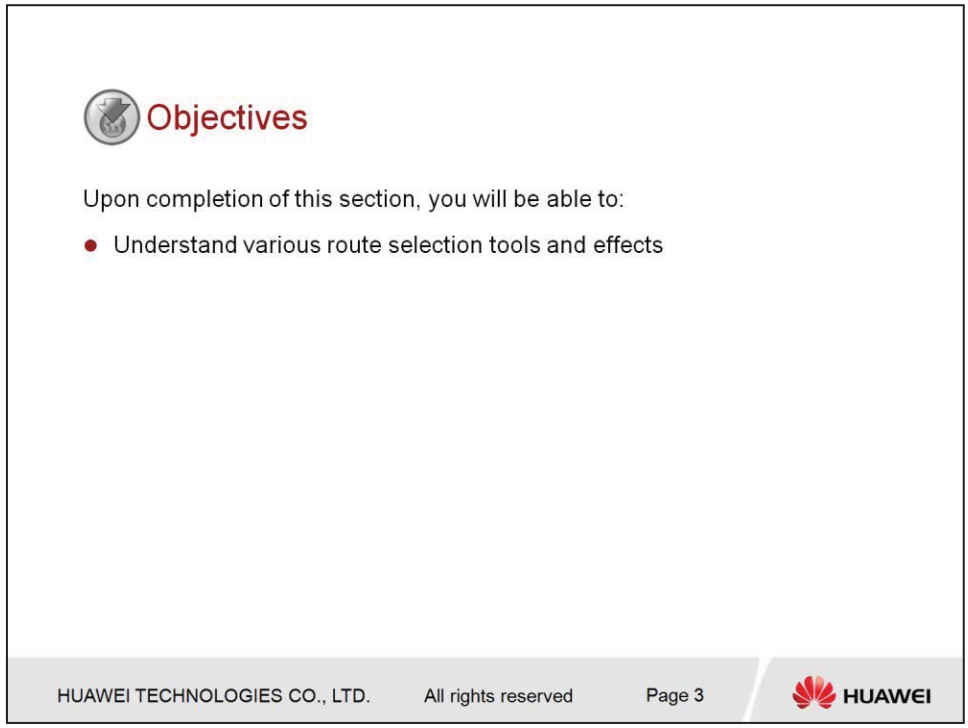
A: The neighbor will learn the routing information by exchanging the UPDATE message once the neighbor relationship has been

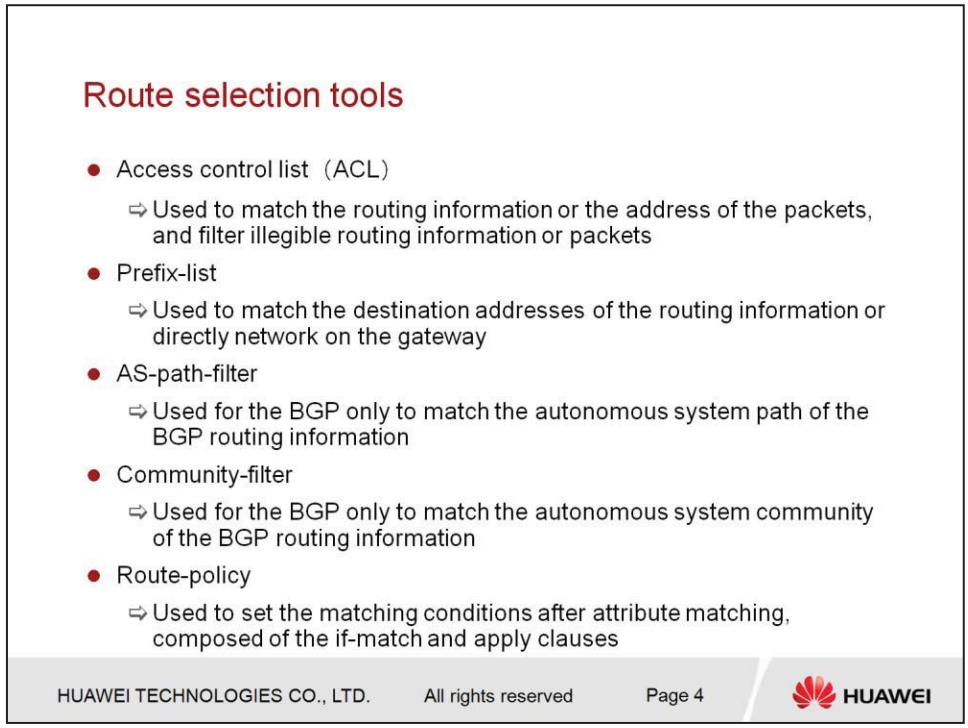
formed. We have to pay attention to the principle of BGP route advertisement. The IP prefixes to be advertised must be installed in the IP routing table. In addition, the subnet mask length of the IP prefix to be advertised and the subnet mask length of the route in the IP routing table must exactly match. The IP prefixes advertised by using the aggregate command must be installed in the BGP routing table. On the other hand, we must issue the "peer next-hop-local" command if the IBGP next hop is not reachable.

Module 4
Route Selection & Control


Route Selection Tools
www.huawei.com
HUAWEI TECHNOLOGIES CO., LTD.
All rights reserved
HUAWEI


Foreword
To implement the definition policy, it is necessary to define the execution of the object first, this can be achieved using route selection tools.
HUAWEI TECHNOLOGIES CO., LTD.
All rights reserved
Page 2
HUAWEI


Objectives
Upon completion of this section, you will be able to:
Understand various route selection tools and effects
HUAWEI TECHNOLOGIES CO., LTD. All rights reserved Page 3
HUAWEI

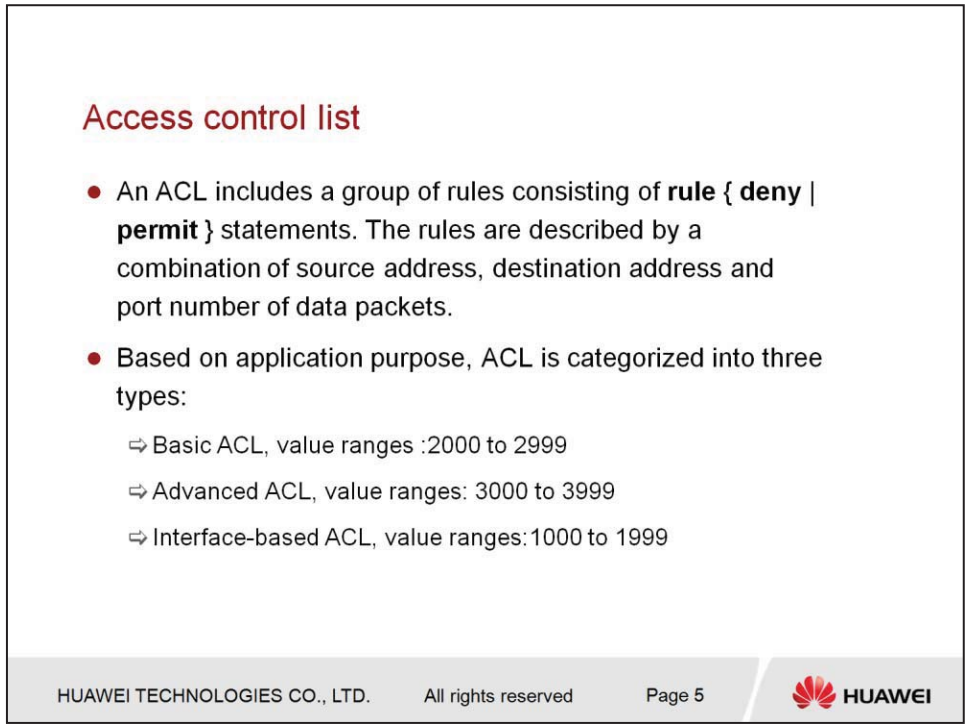


When we implement routing policy or policy-based route, we should first filter the specified route information or data packet. Different filtering tools can be used according to different filtering objects. Generally, access control list and prefix-list can be used to match IP address, but prefix-list can not filter data packet, it can only filter routing information. So we should first know the match object, route or data, and then choose the proper tool.

As-PATH-FILTER is used to match AS-PATH attribute in BGP route information, so it can only filter BGP route.

Community-filter is used to match community attribute in BGP route information, so it only can filter BGP route .

Route-policy is a powerful filter tool, it is not only a filter but also a policy tool. As a filter, it can match route and data packet by using if-match sentence, if-match sentence can also transfer other filters. As a policy tool, it can modify route attribute or forwarding behavior of data packet by using apply sentence.

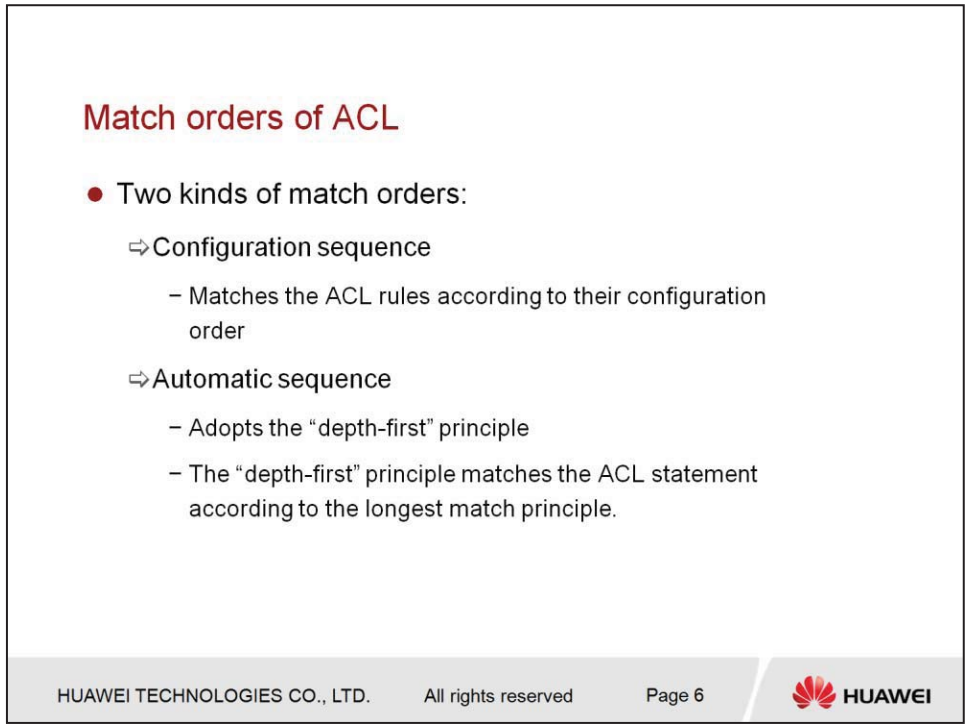


Access control list can match not only data packet but also route information.

Basic access control list can match source IP address.

Advanced access control list can match source IP address, destination IP address, source port number, destination port number and protocol number.

Interface-based Access control list can match interface.

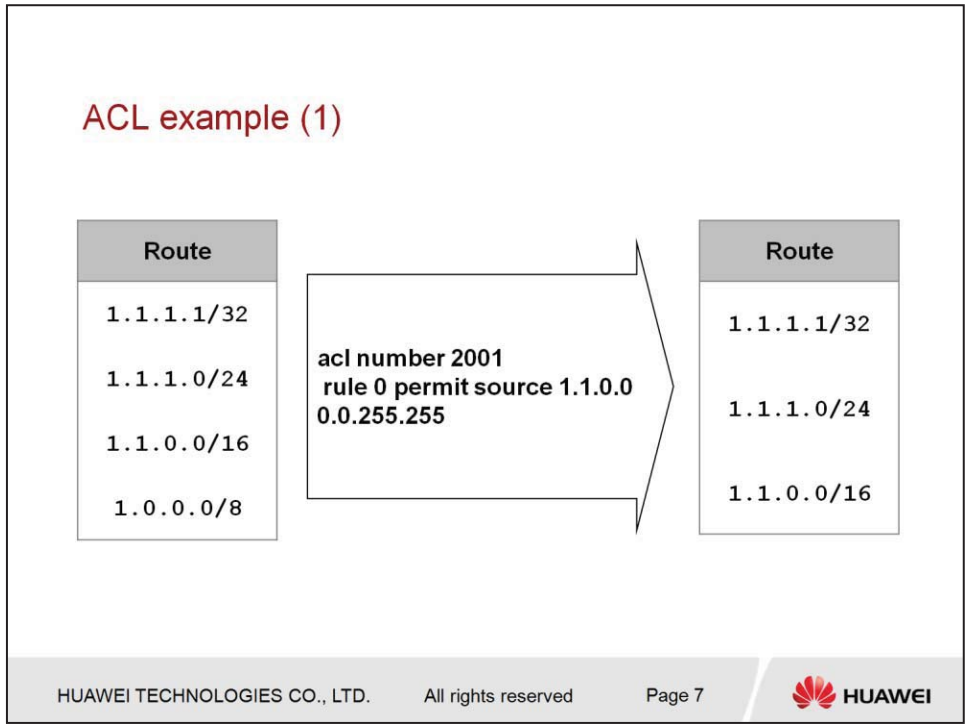


There are two kinds of match orders: Configuration sequence and Automatic sequence.

Configuration sequence: Matches the ACL rules according to their configuration order Automatic sequence adopts depth-first principle.

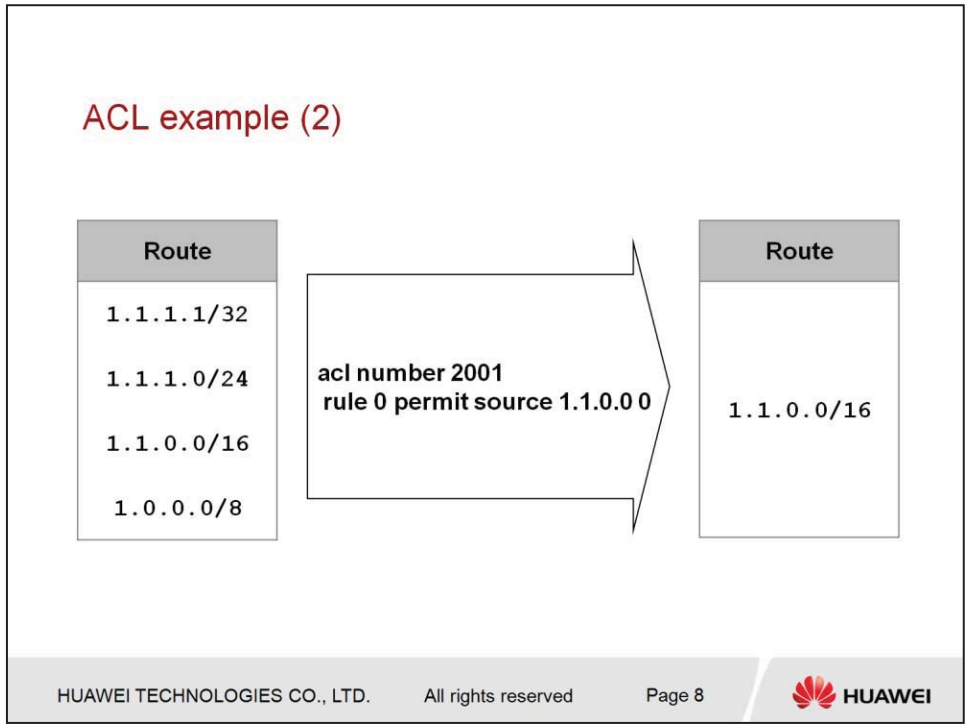
The "depth-first" principle matches the ACL statement according to the longest match principle. For example, 129.102.1.1 0.0.0.0 indicates a host: 129.102.1.1, but 129.102.1.1 0.0.0.255 indicates a network segment: 129.102.1.1 ～ 129.102.1.255. Obviously, the former rule will be matched first in access control list. The detailed standard is:

For the basic access control list. The router matches ACL rule according to the source wildcard mask. Then match the configuration sequence if two rule got the same source wildcard mask. For the interface-based access control list. The rule "any" should be back, the other rule can be according to configuration sequence. For advanced access control list, compare wildcard of source address first, then compare wildcard of destination address, at last compare port range, the smaller should be ahead. If the port range is the same, then adopt configuration sequence.

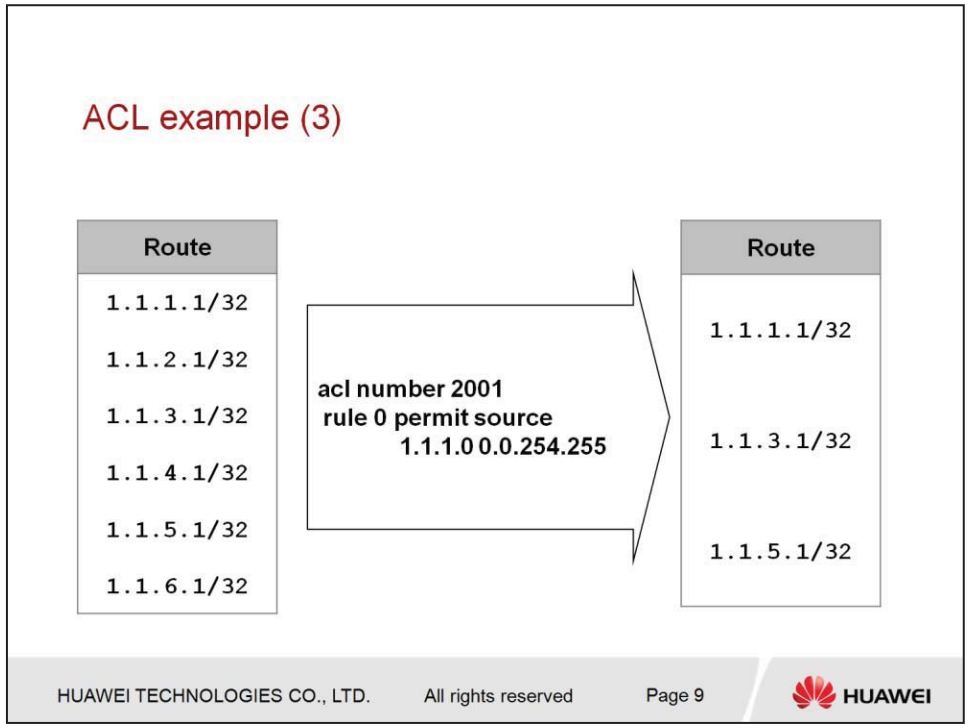


The result of this ACL is to only match the route entity with the first 2 bytes "1.1".

The last 2 bytes will be ignored, so the network 1.1.1.1/32, 1.1.1.0/24 and 1.1.0.0/16 are matched.

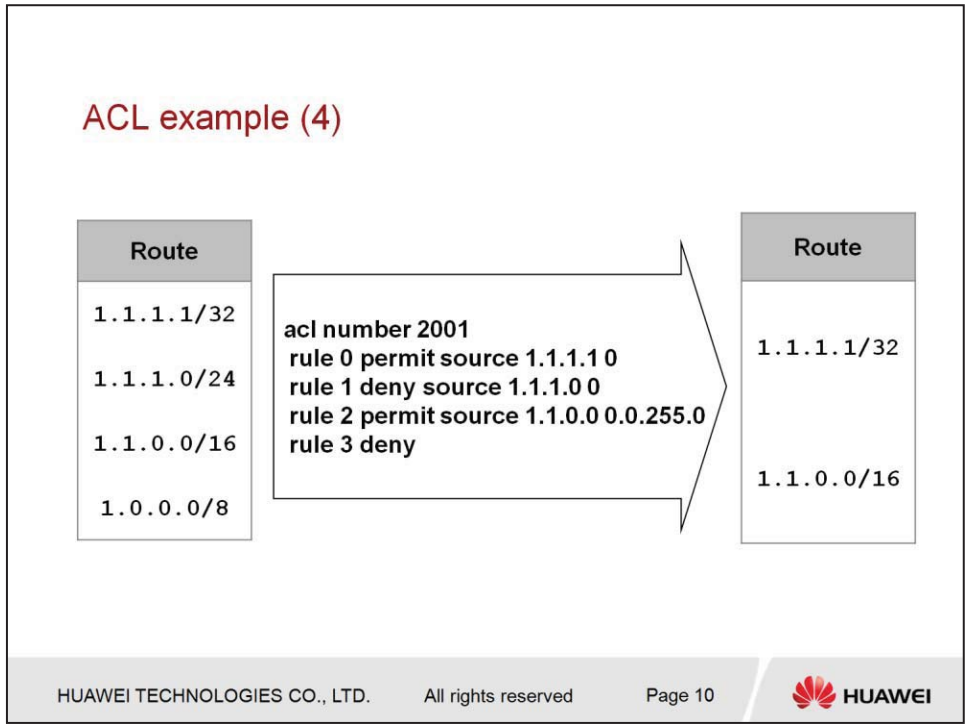


This ACL means that all the 32 bits of route must be match. So finally only one route entity 1.1.0.0/16 is matched.



Please remember that “0” means the routing entity must strictly match and “1” means the routing entity can be ignored. So convert decimal wildcard mask 0.0.254.255 into binary that is “00000000.00000000.11111110.11111111”. This binary wildcard mask means that the first 16 bits and the 24th bit must be strictly match, and the others are ignored. The conclusion is that the first 16 bits must be “1.1”

and the 24th bit must be “1” also. So network 1.1.1.1/32, 1.1.3.1/32, 1.1.5.1/32 are matched, the other routing entity are not matched.



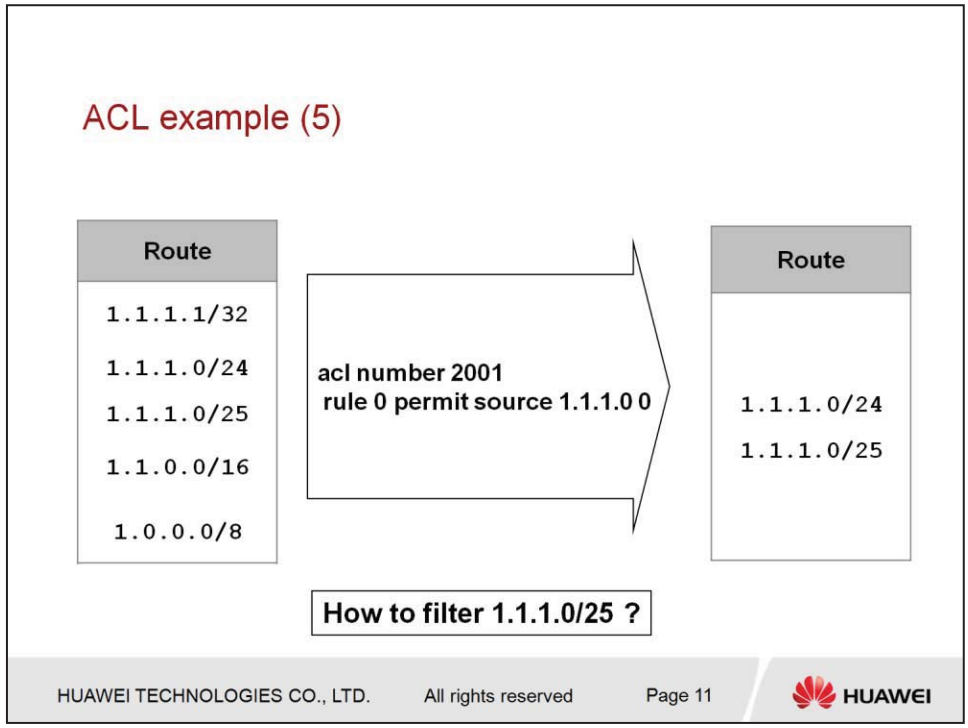
There is actually more than one rule that can be configured in an ACL. For this example, we configured 4 rules in ACL 2001.

“1.1.1.1/32” matches “rule 0”, permit.

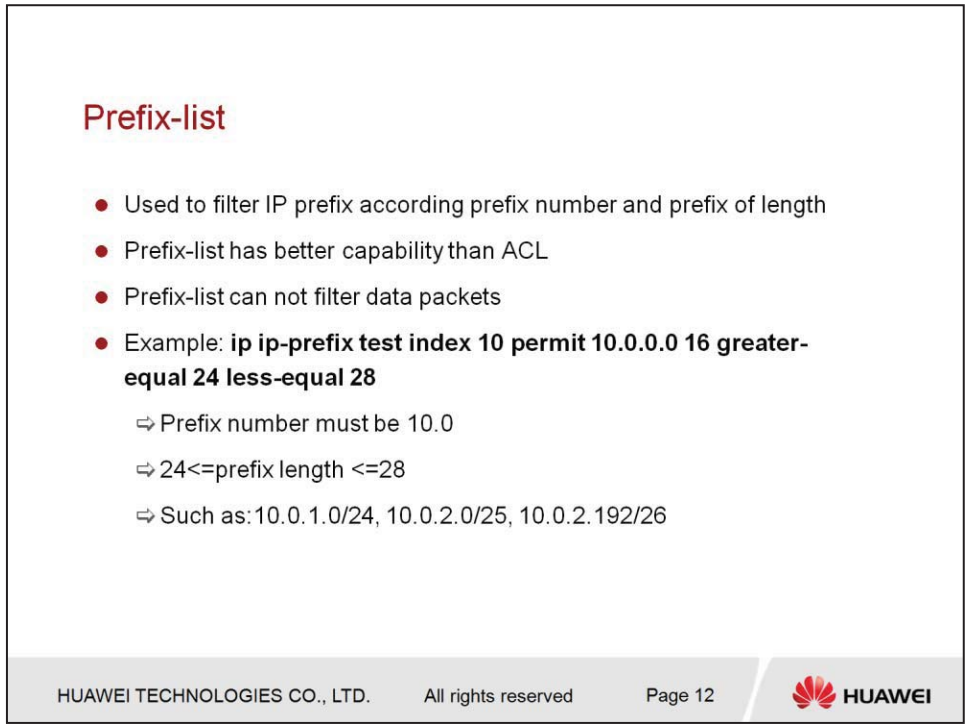
“1.1.1.0/24” matches “rule 1”, denied.

“1.1.0.0/16” matches “rule 2”; permit.

“1.0.0.0/8” can not match the first three match conditions, so it is filtered by “rule 3”.



Using an ACL is is possible to match the prefix of a route, but for routes with same prefix, but different masks, how can they be distinguished ? In this case, a prefix list can be used.



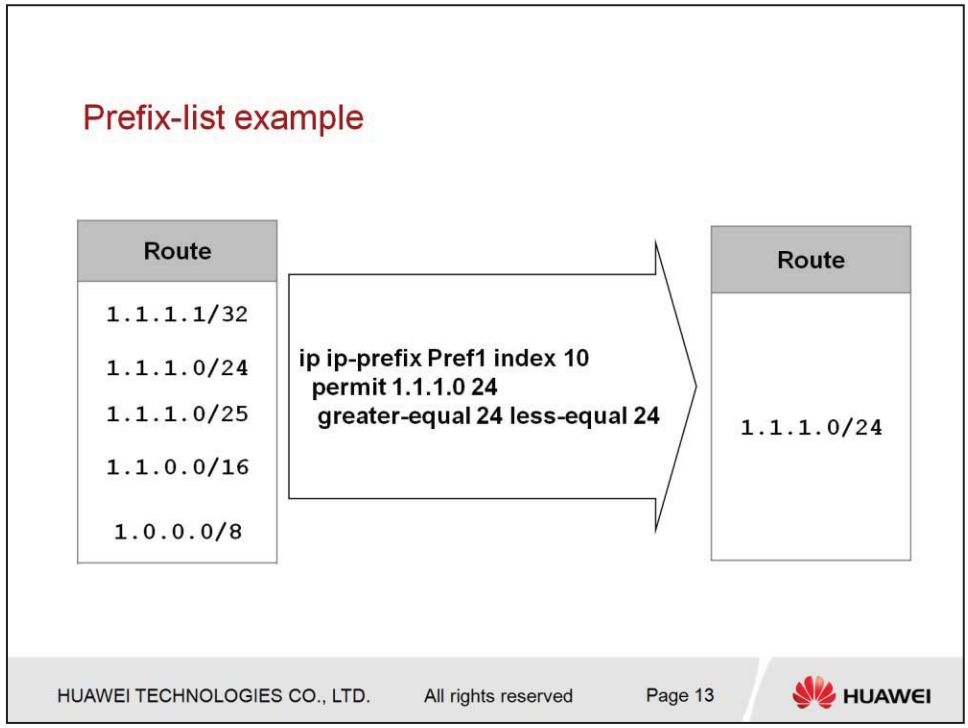
Prefix list can filter IP prefix, match prefix number and prefix length; the performance of prefix list is higher than access control list. But the prefix list can not filter data packet.

For example：ip ip-prefix test index 10 permit 10.0.0.0 16 greater-equal 24 less-equal 28

Prefix number must be 10.0

24<=prefix length <=28

The routes that satisfy the condition are as followings: 10.0.1.0/24, 10.0.2.0/25, 10.0.2.192/26

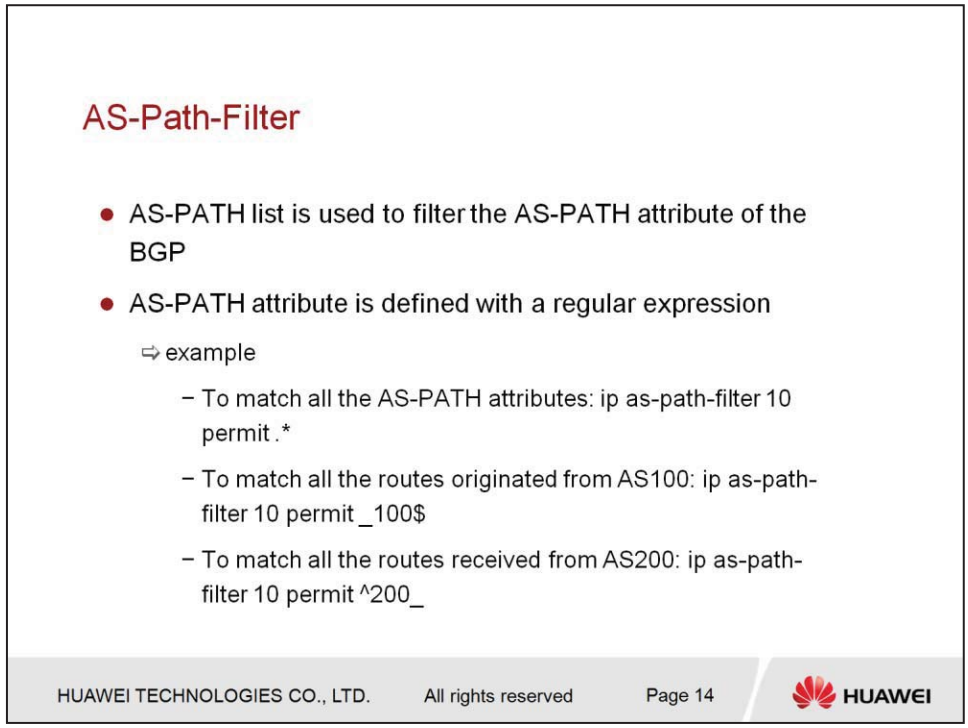


For this prefix list, “index 10” defines two match condition: one is “1.1.1.0 24 ”, the other is “greater-equal 24 less-equal 24”.

“1.1.1.0 24 ” indicates that the first three bytes of prefix must be “1.1.1”;

“greater-equal 24 less-equal 24”indicates that mask length of route must be 24 bits. So, only “1.1.1.0/24” satisfies the match condition.

Notes: prefix lists can define multiple “index”



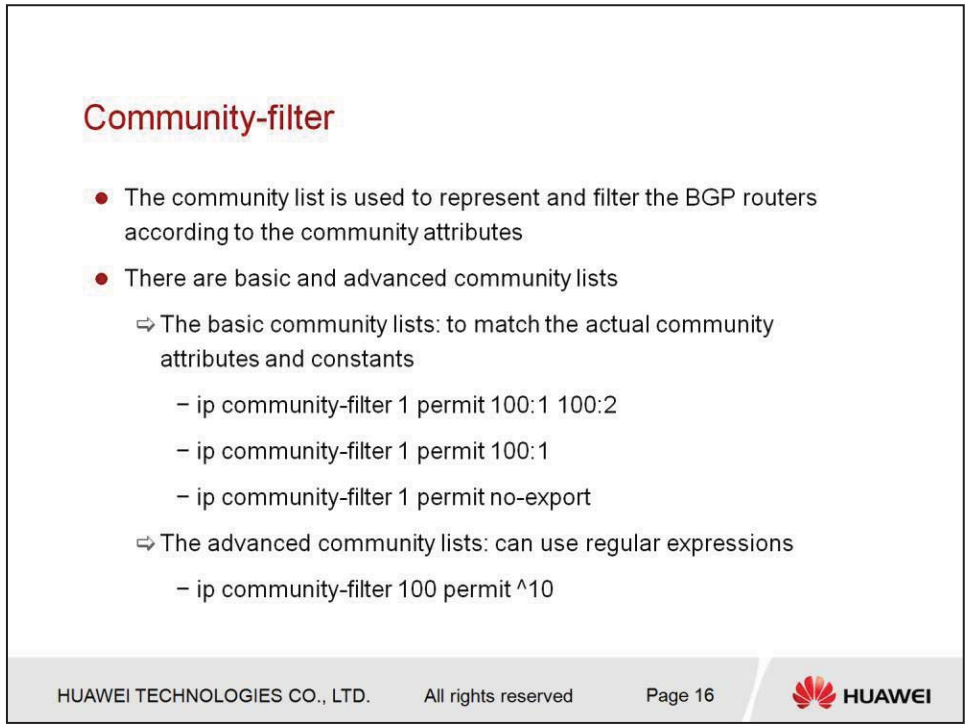
AS-PATH attribute is used to record the AS the route passes during the transmission. If one route originates from AS100, passes through AS300, AS200, AS500， and reaches AS600 at last, then in AS600, the AS-PATH attribute of the route is ‘500 200 300 100’. AS-PATH attribute is a string in fact, so we can use regular expression. Regular expression uses some special symbols to express special signification.

AS-Path-Filter and regular expression

name	symbol	meaning
period	.	Match any single character
asterisk	*	Match 0 or more sequences in the mode
plus	+	Match 1 or more sequences in the mode
Question mask	?	Match 0 or 1 mode appearance
Padding	^	Match the start of the input character string
Dollar	$	Match the end of the input character string
Underline	_	Match the start and end of comma, parenthesis, character string and space
Square Brackets	[range]	Indicate the range of single-character mode
Hyphen	-	Separate end points of a range

HUAWEI TECHNOLOGIES CO., LTD. All rights reserved

Regular expression is very flexible, the same signification can be expressed in many ways.

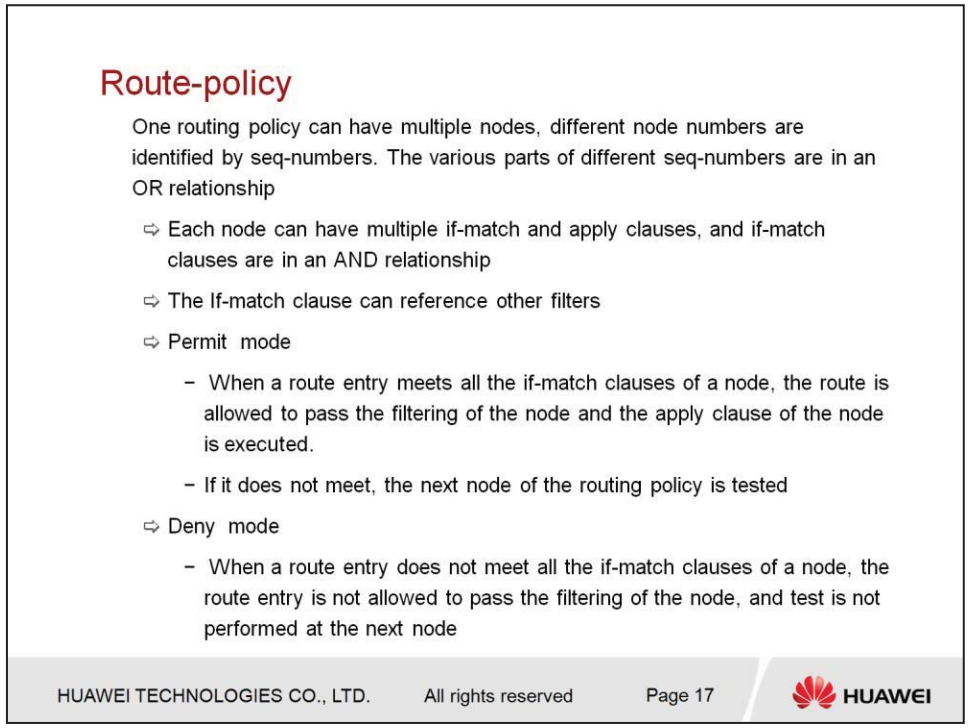


Community attribute includes Well-known community attribute and private community attribute.

Well-known community attribute includes: internet, no-advertise, no-export, noexport-subconfed.

Private community attribute is manually defined, it adds management tag to prefix in order to constitute corresponding policy and the format is AS:NUMBER.

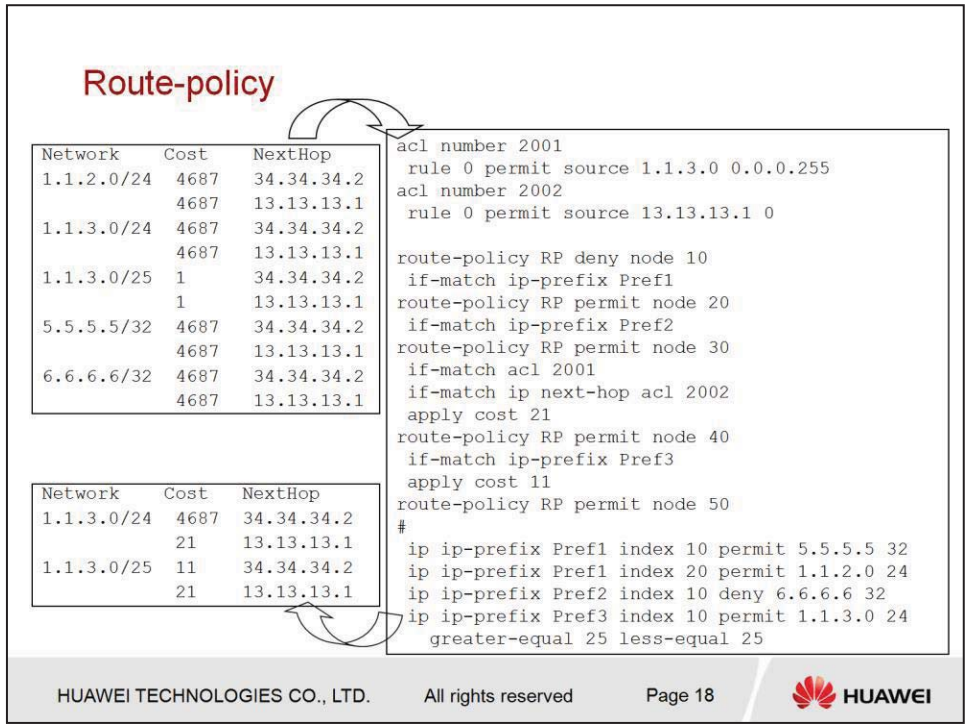
Advanced community list can use regular expression to match community attribute.



Route-policy is a powerful filter tool and policy tool. Each Route-policy node comprises a set of if-match and apply clauses. If-match sentence can match ACL, ip-prefix, as-path-filter, community-filter, interface, ip, extcommunity-filter, cost,

mpls-label, route type, tag and so on. Apply sentence can modify route attribute.

Route-policy is used to control routing information such as route filtering or route attributes.



The route in routing table matches the node of route policy according to the sequence.

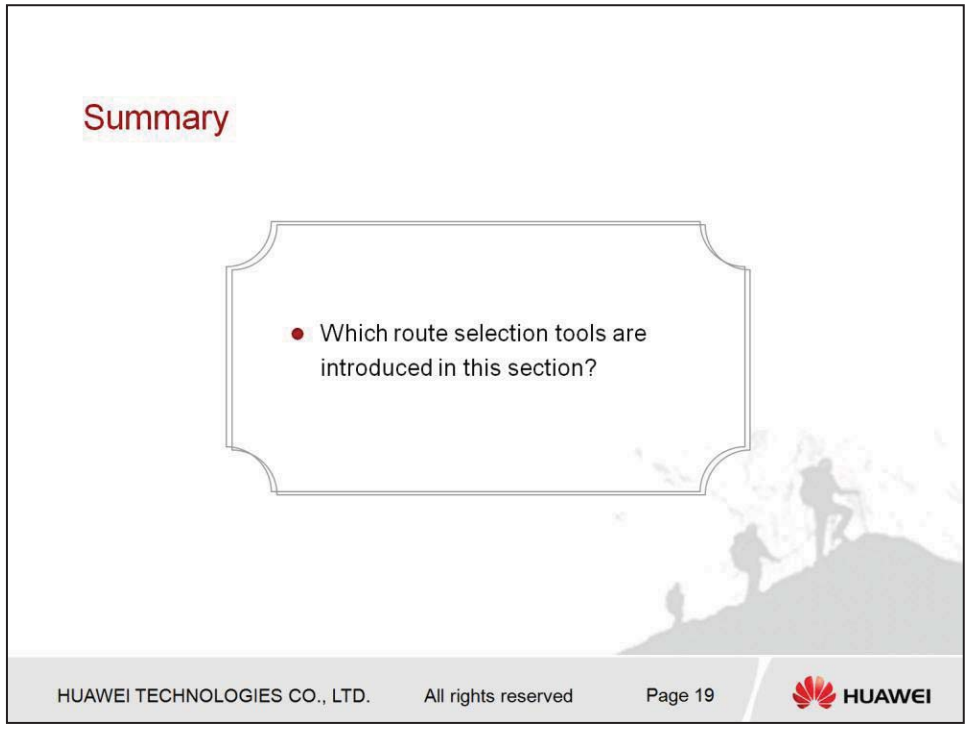
First, it defines prefix list Pref1 to match 5.5.5.5/32 and 1.1.2.0/24, the item that can match the prefix list is denied by node 10 of route policy. So in the filtered routing table, 5.5.5.5/32 and 1.1.2.0/24 can not be seen.

Prefix list Pref2 is used to filter 6.6.6.6/32, although the node 20 of route policy is “permit”, 6.6.6.6/32 is filtered.

Node 30 of route policy defines two if-match sentences, it should match ACL 2001 and the next-hop should match ACL 2002 at the same time; the route item that matches ACL 2001 includes: 1.1.3.0/24, 1.1.3.0/25 and so on 4 routes in total. But if it want to satisfy node 30, it should satisfy that the next-hop is 13.13.13.1. So only two routes satisfy the condition, the cost of them are modified to 21 by the apply sentence.

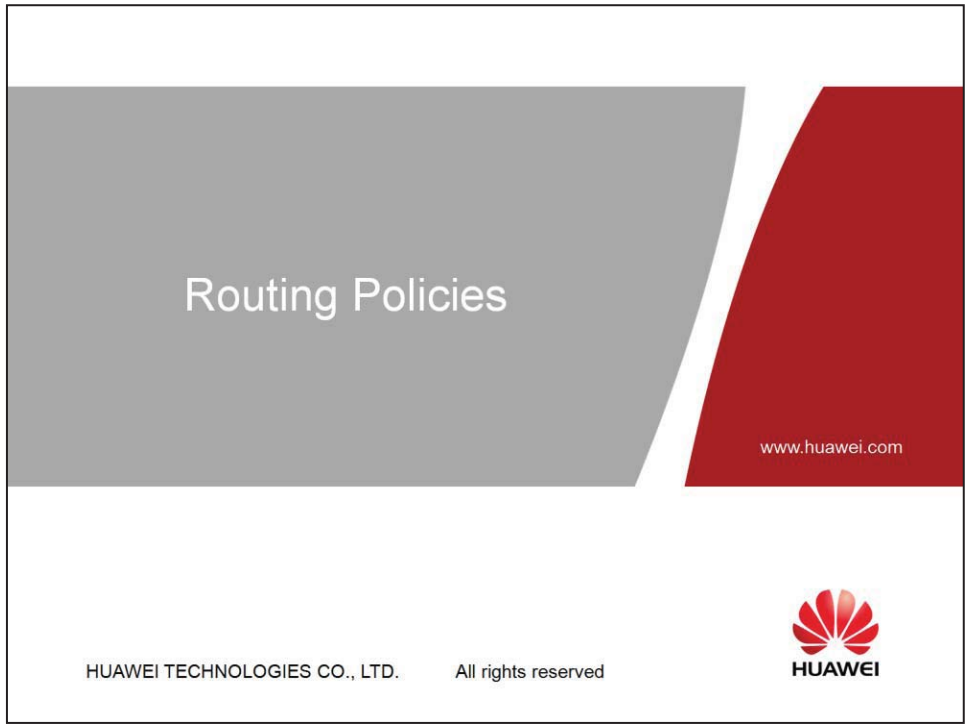
The left routes are 1.1.3.0/24 and 1.1.3.0/25 whose next-hop is 34.34.34.2, they try to match node 40. Route 1.1.3.0/25 can match prefix list Pref3,and the cost is modified to 11.

The last route 1.1.3.0/24 will be reserved by node 50.



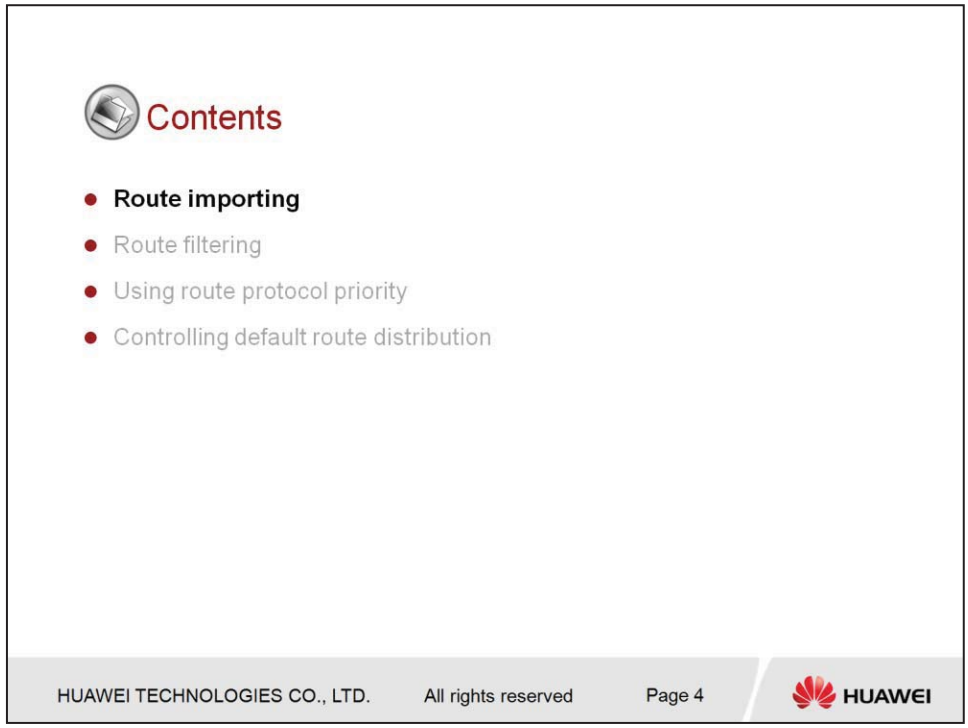
Q：Which route selection tools are introduced in this section?

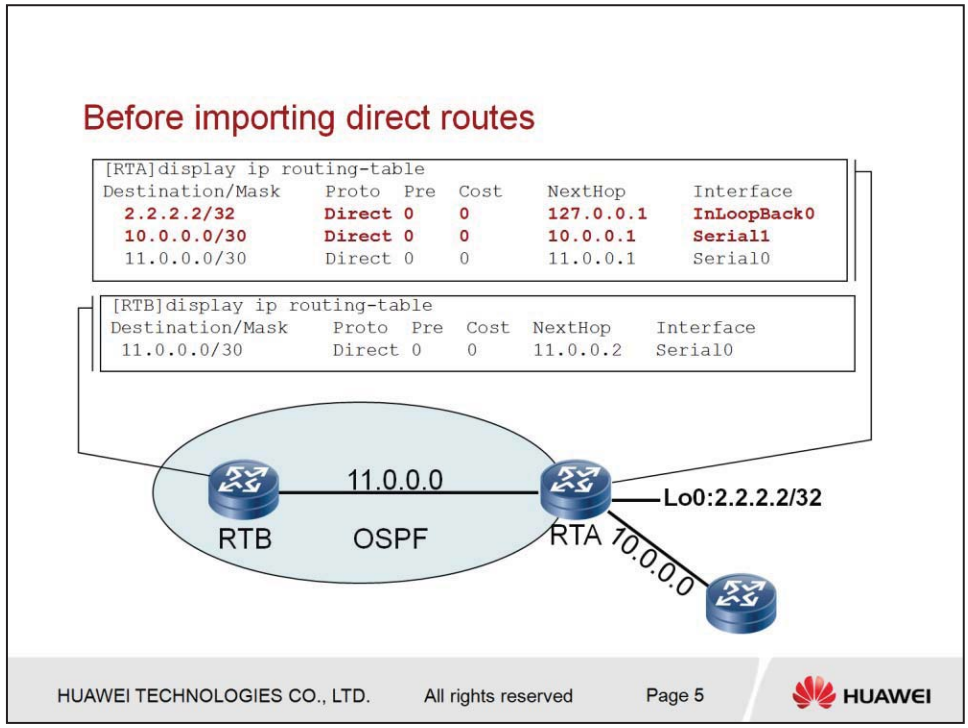
A：ACL; IP-Prefix list; AS-Path-Filter； Community Filter； Routing-Policy.


Routing Policies
www.huawei.com
HUAWEI TECHNOLOGIES CO., LTD.
All rights reserved
HUAWEI

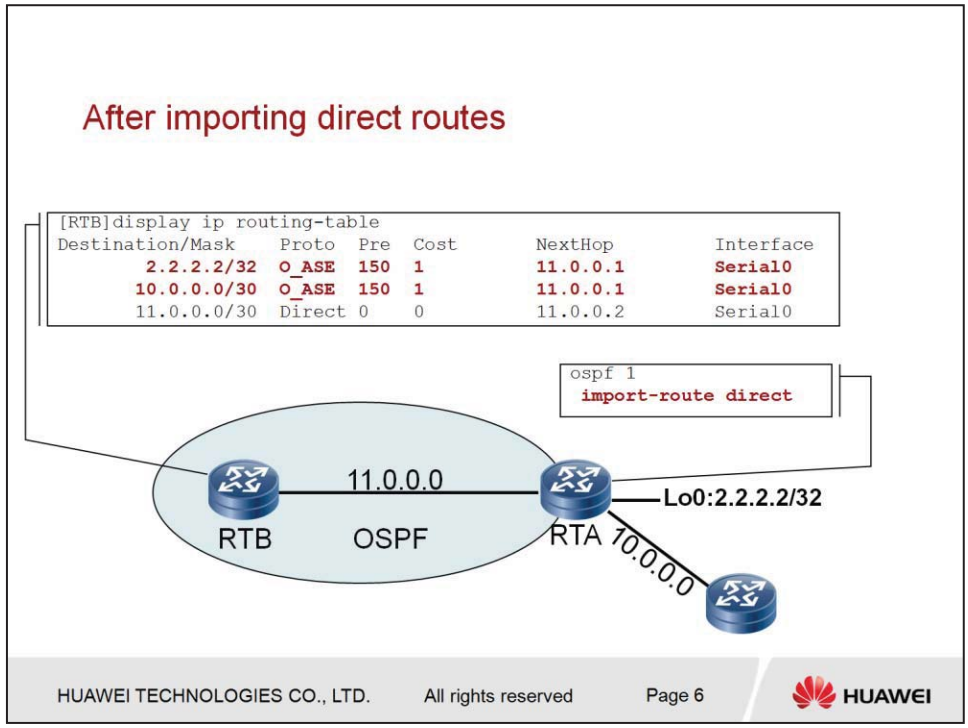

Foreword
We can use routing policies to control the route selecting and importing process.
HUAWEI TECHNOLOGIES CO., LTD.
All rights reserved
Page 2
HUAWEI


Objectives
Upon completion of this section, you will be able to:
Master the use of routing policies to control route receiving process
Master the use of routing policies to control route importing process
Master the use of route priority to select routes
HUAWEI TECHNOLOGIES CO., LTD.
All rights reserved
Page 3
HUAWEI

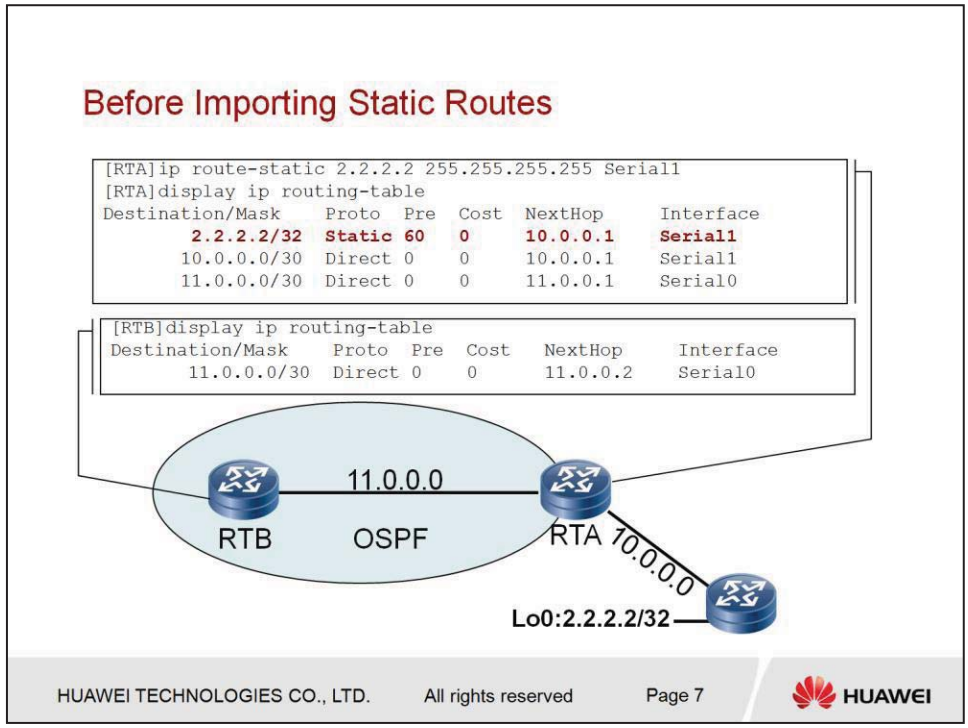

Contents
Route importing
Route filtering
Using route protocol priority
Controlling default route distribution
HUAWEI TECHNOLOGIES CO., LTD.
All rights reserved
Page 4
HUAWEI



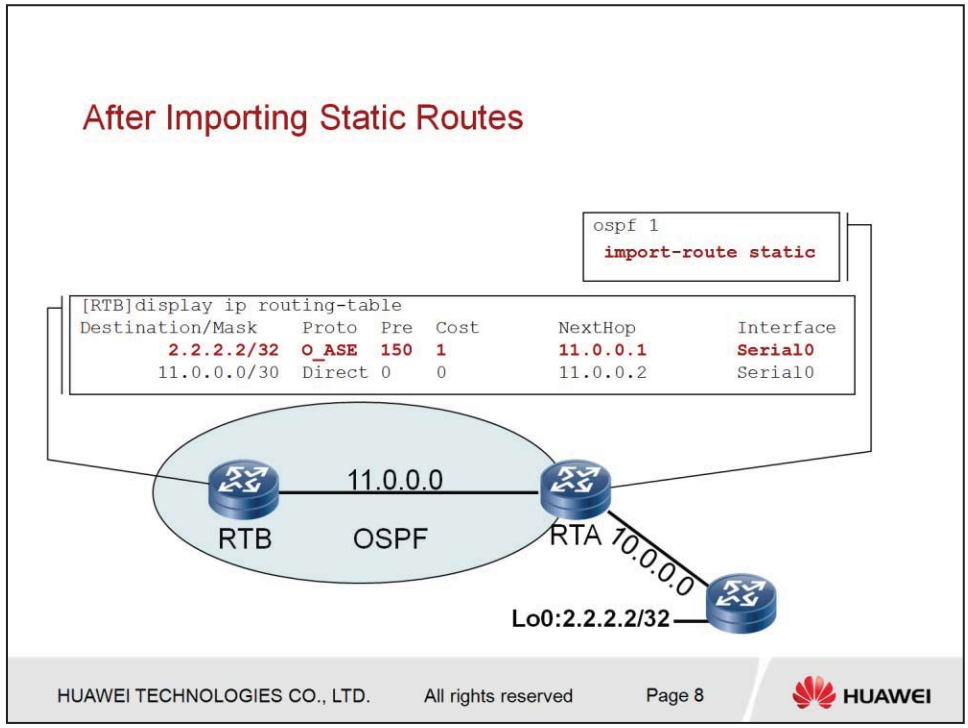
On RTA, OSPF runs only over network segment 11.0.0.0, in RTA's routing table, 2.2.2.2 and 10.0.0.0 are direct network segment; before configuring route import, 2.2.2.2 and 10.0.0.0 will not be advertised to RTB; in RTB's routing table, there is no any information about 2.2.2.2 and 10.0.0.0 network segment.



If import direct route into OSPF on RTA, RTA will advertise route information about network segment 2.2.2.2 and 10.0.0.0 to RTB, so, on RTB, we can see route about network segment 2.2.2.2 and 10.0.0.0, they are learnt by OSPF (Proto filed is O_ASE, namely , OSPF external route).

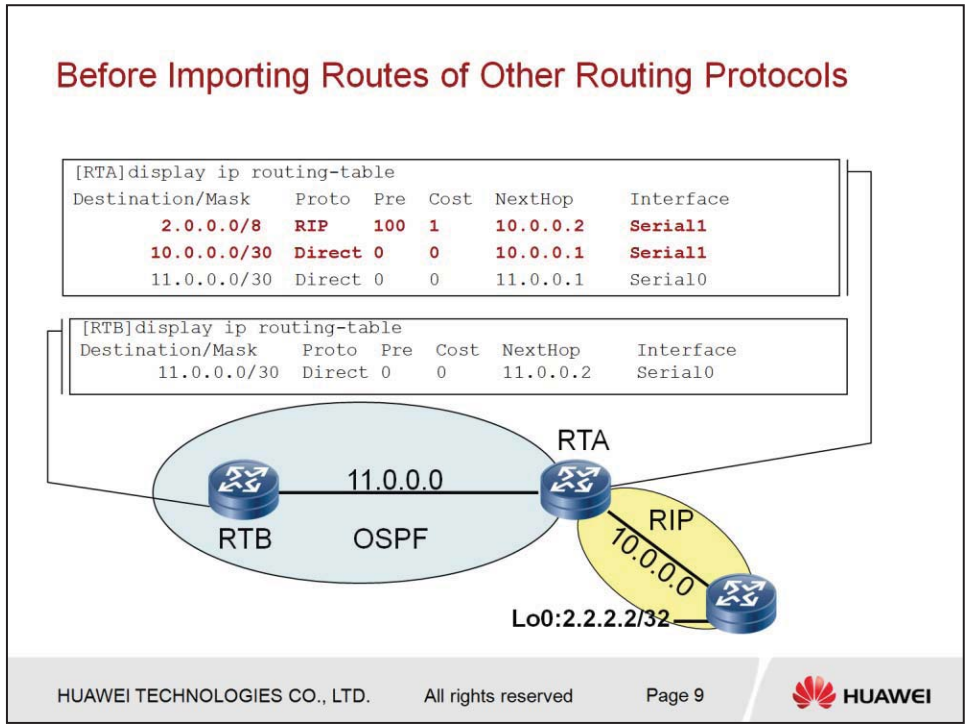


On RTA, we configure a static route destined for 2.2.2.2. if route import is not configured on RTA, this static route will not be advertised into OSPF routing domain automatically.

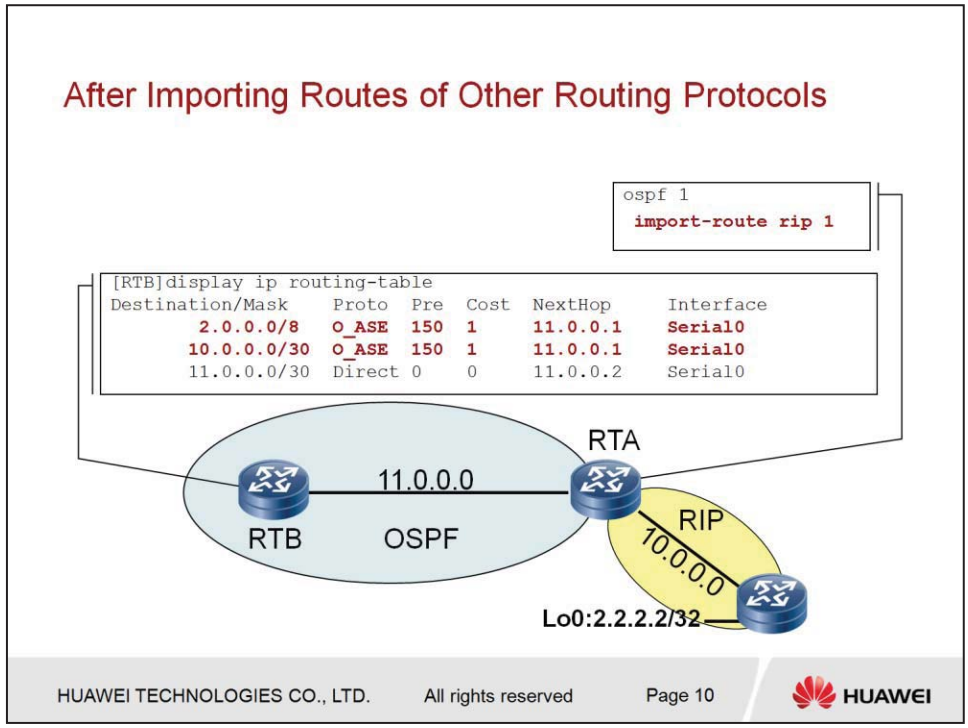


On RTA, importing static route is configured, network segment 2.2.2.2 will be advertised to RTB through OSPF.

OSPF does not run over the network segment 10.0.0.0, and network segment 10.0.0.0 is not static route but direct route. So 10.0.0.0 will not be advertised to RTB.



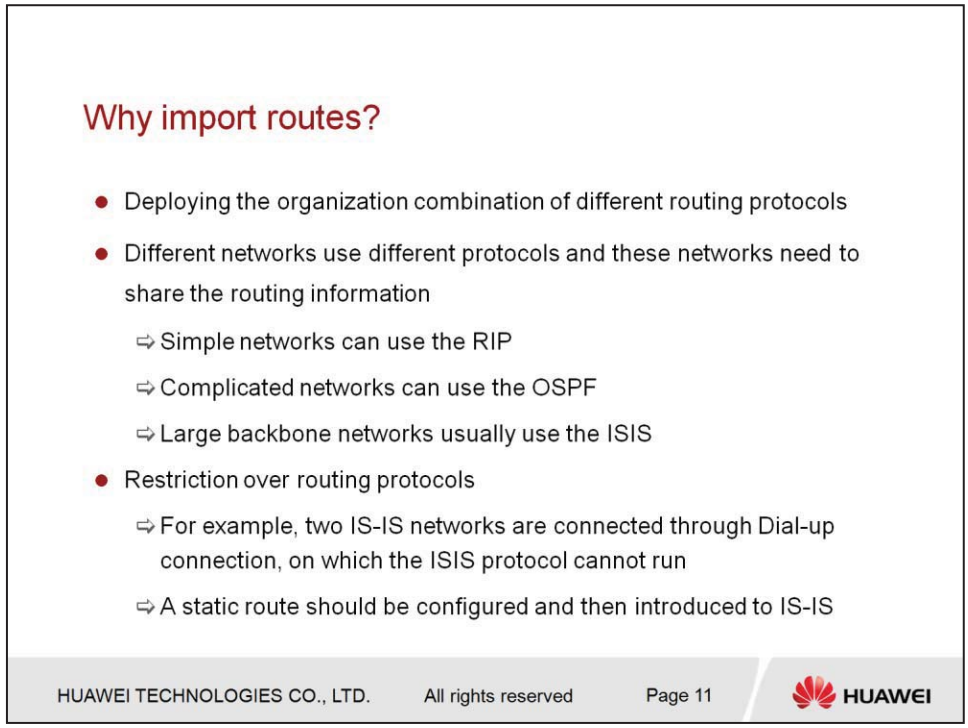
In this network, OSPF and RIP run at the same time, OSPF runs over network segment 11.0.0.0, RIP runs over network segment 10.0.0.0 and 2.0.0.0. RTB doesn't have any route information about network segment 2.0.0.0 and 10.0.0.0 before importing route.



Import RIP into OSPF on RTA, then RTB can learn route information about network segment 2.0.0.0 and 10.0.0.0.

Note:

In RTA routing table, although 10.0.0.0 is a direct connected network, RIP is running on network segment, so it is imported into OSPF.



In general, one protocol is enough. But in some cases, we should configure route import.

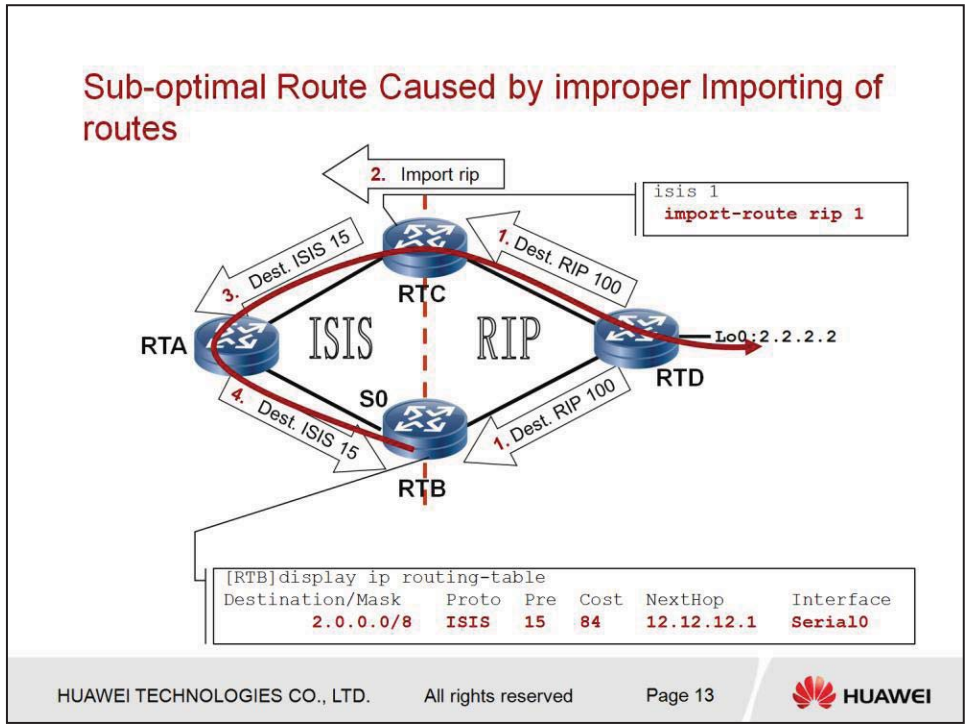
1. The mergence of organizations implementing different protocols For example, company A implements OSPF, while company B implements ISIS and A and B will combine into company C. The network of A and that of B want to communicate each other, in this case, route import is needed.

2. Different networks use different protocols, and the networks need to share route information A large-scale network may be made of many small-scale networks, the complexity of the small-scale networks is different, some run RIP, some run OSPF (OSPF can support more network types than ISIS), the other run ISIS. To implement the intercommunication of small-scale networks, route import is needed.

3、The limitation of network protocol

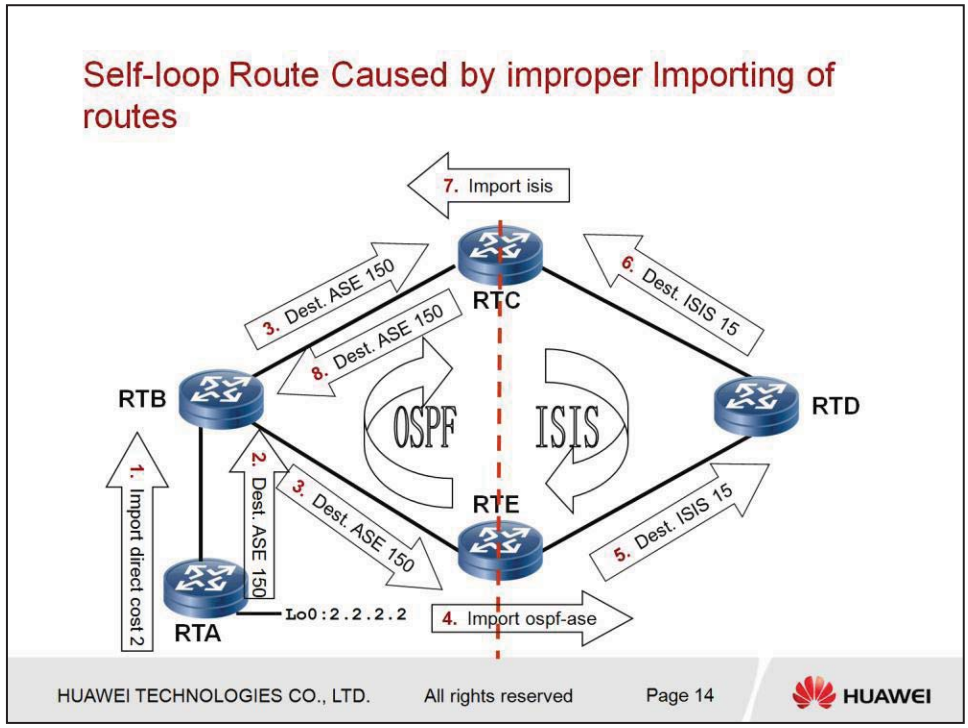
Dial-up network is accounted by time, so generally, Dial-up network is used as backup link; if the master link is normal, Dial-up link does not work. In Dial-up network, ISIS protocol is not suitable (OSPF protocol has the special design for Dial-up link), because ISIS sends packet periodically, which will lead that the

Dial-up link is UP state even though the master link is normal. In general, static route is configured on Dial-up link, and then import static route into ISIS.



In the network, ISIS and RIP are running at the same time. RTC and RTB are ASBRs. On RTC, import RIP into ISIS. Network segment 2.2.2.2 will be advertised to RTA by ISIS, and then RTA will advertise it to RTB. RTB learns route 2.2.2.2 from RIP and ISIS at the same time, then RTB will compare the preference of ISIS and RIP, the preference of ISIS is 15, the preference of RIP is 100, RTB will choose the route advertised by ISIS.

RTB will choose the sub-optimal path RTB-RTA-RTC-RTD to send data packet to 2.2.2.2. So when configure route import, sub-optimal route is avoided. But how to avoid? It is introduced in the following course .



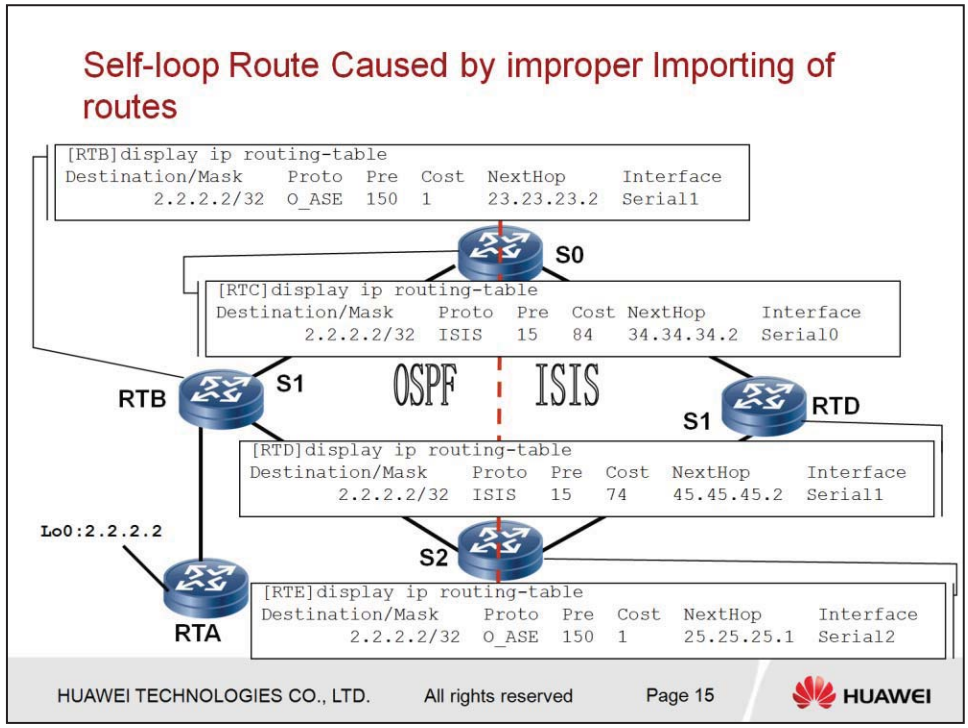
RTA imports direct route 2.2.2.2 into OSPF, OSPF will advertises it to RTB, and RTB will advertise to RTC and RTE by the ASE route (preference is 150). On RTE, importing of OSPF-ASE

is configured, and imports 2.2.2.2 into ISIS. On RTC, importing ISIS is configured, and imports the ISIS routes into OSPF. The network segment 2.2.2.2 is advertised into OSPF by ISIS, which is called route feedback. RTB learns route about DEST

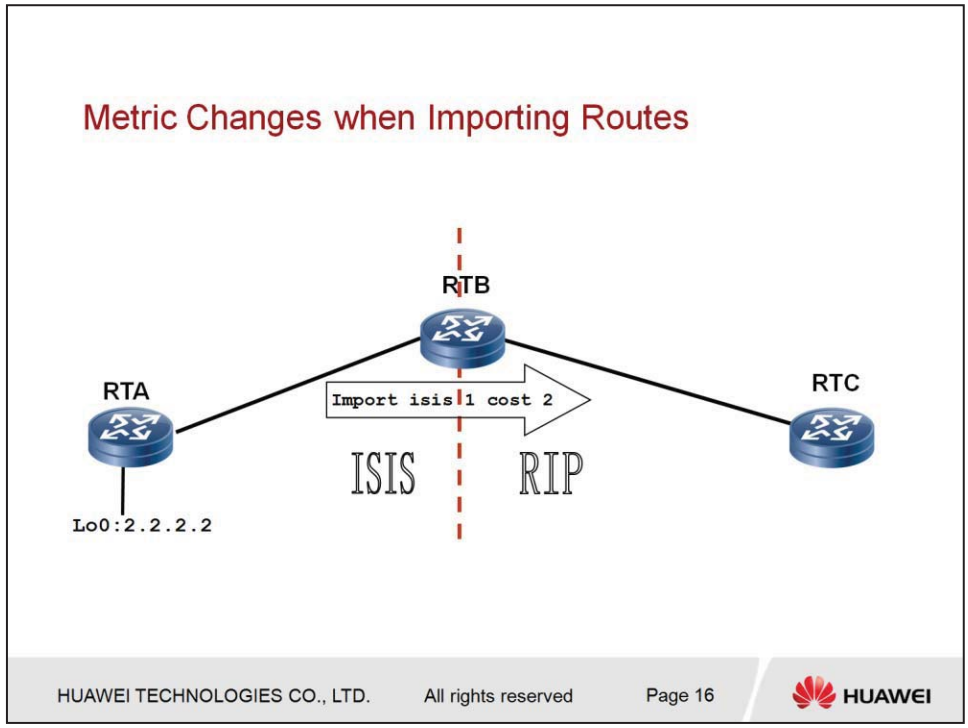
from RTA and now also from RTC, and because the preferences are the same (all are OSPF ASE routes), it compares the metric value. If RTB chooses the route advertised by RTC, a loop will occur. For example, RTD sends a packet to 2.2.2.2, the packet

is sent to RTE and RTB, and then RTB forwards it to RTC, then RTD, at last, the packet goes back to the originator.

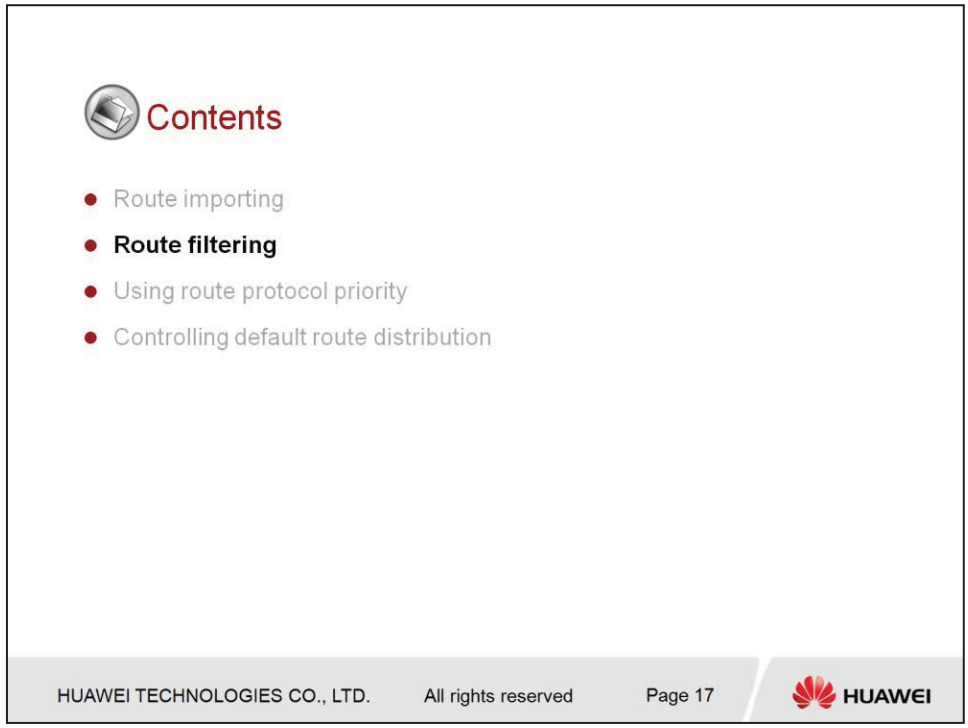
In complex environment, loop should be avoided carefully, but how to avoid? It is introduced in the following section.

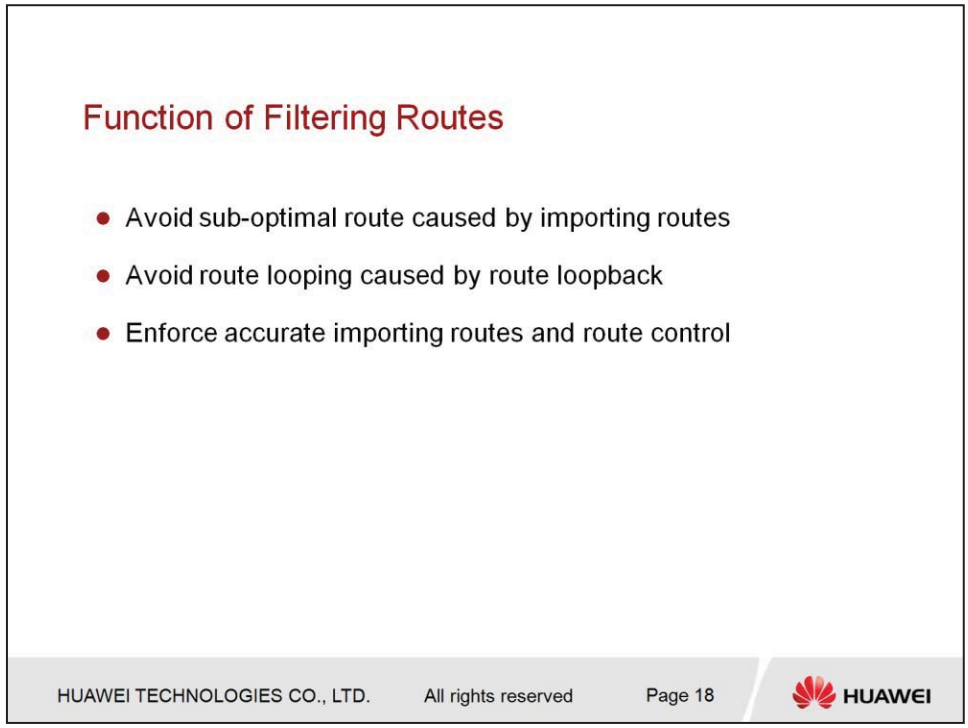


From the routing table, we can see that loop has occurred.

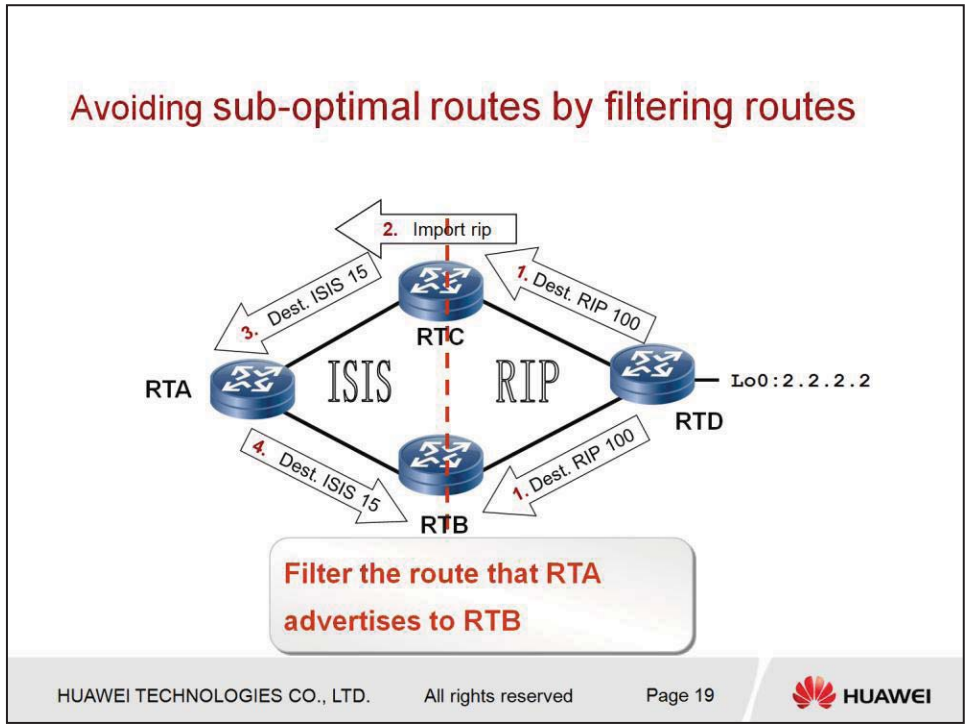


Different routing protocols calculate route metric according to different elements, the metric value is different. The metric range is also different. The metric of ISIS and OSPF can be based on bandwidth, and the value range is large; the metric of RIP is based on hop, the range is small, so when we configure ISIS and RIP import or OSPF and RIP import, we must be careful (Fortunately on VRP, when import OSPF or ISIS into RIP, if cost is not appointed, the value defaults to 1. Even so, we should configure metric manually so as to reflect real topology).


Contents
Route importing
Route filtering
Using route protocol priority
Controlling default route distribution
HUAWEI TECHNOLOGIES CO., LTD.
All rights reserved
Page 17
HUAWEI



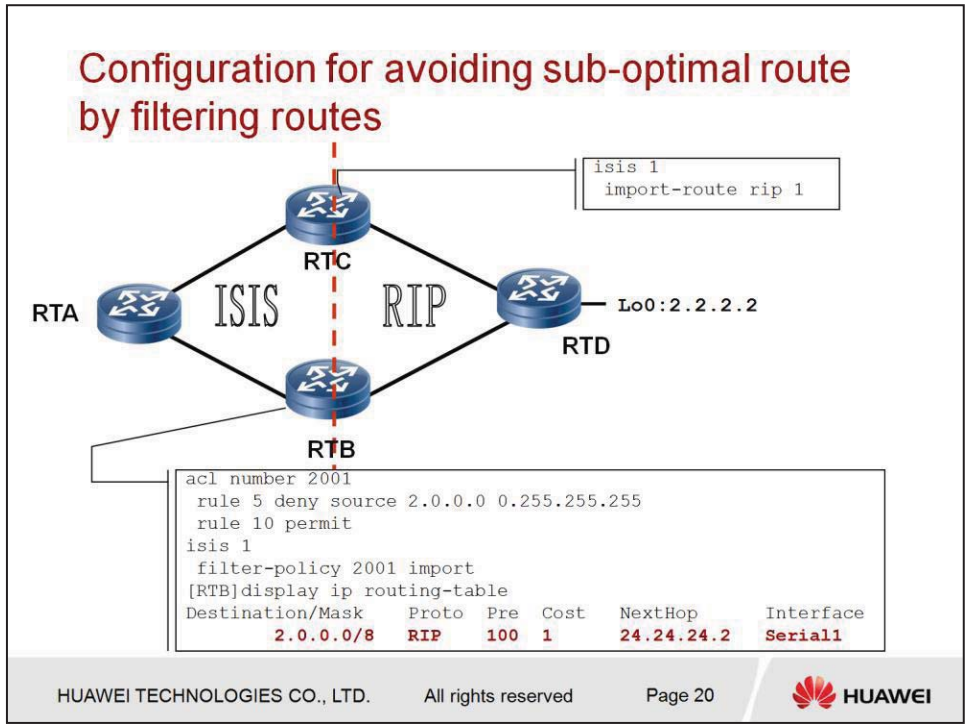
As referred before, route import may lead to sub-optimal route and route loop, then, we can adopt route filter to avoid the problem. Besides, route filter can ensure the route import and advertisement precisely.



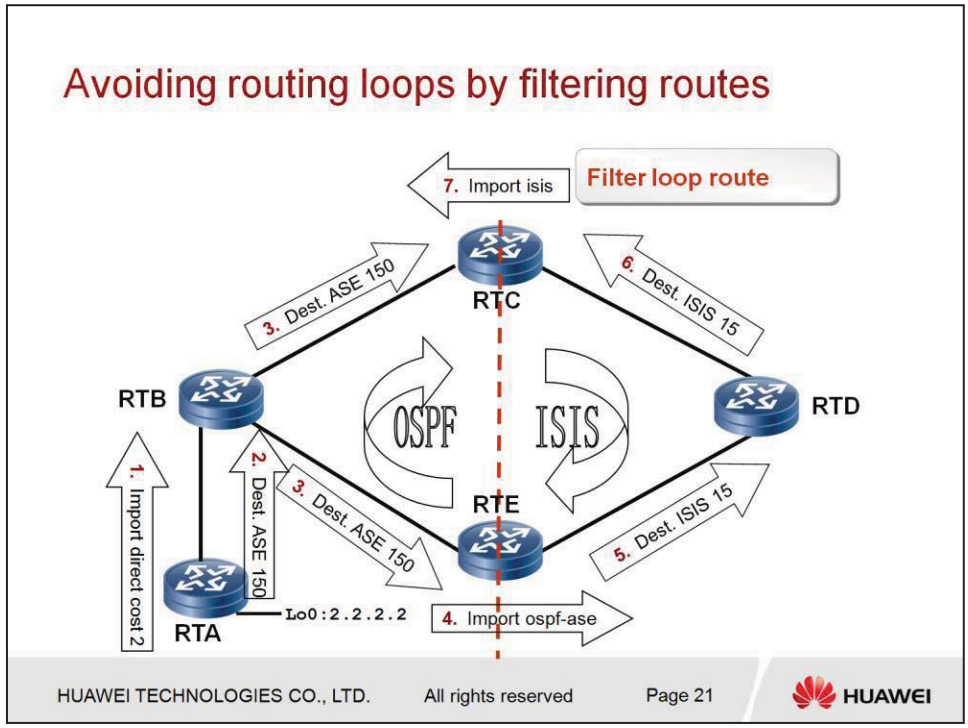
In this network, ISIS and RIP are running at the same time. RTB and RTC are ASBR. On RTC, import RIP into ISIS. Network segment 2.2.2.2 is advertised to RTA by ISIS, then RTA advertises it to RTB. RTB learns route 2.2.2.2 from RIP

and ISIS at the same time, it will compare the preferences of ISIS and RIP, the preferences of ISIS is 15, the preferences of RIP is 100, so RTB will choose the route advertised by ISIS. RTB sends packet to 2.2.2.2 along the sub-optimal path RTB-RTA-RTC-RTD.

Here, the reason causing sub-optimal route is that RTB learns route 2.2.2.2 from ISIS and RIP, and it choose the route learnt from ISIS. We can configure route filter on RTB to filter ISIS route, in this way, RTB will use RIP route to forward data packet so as to avoid sub-optimal route.

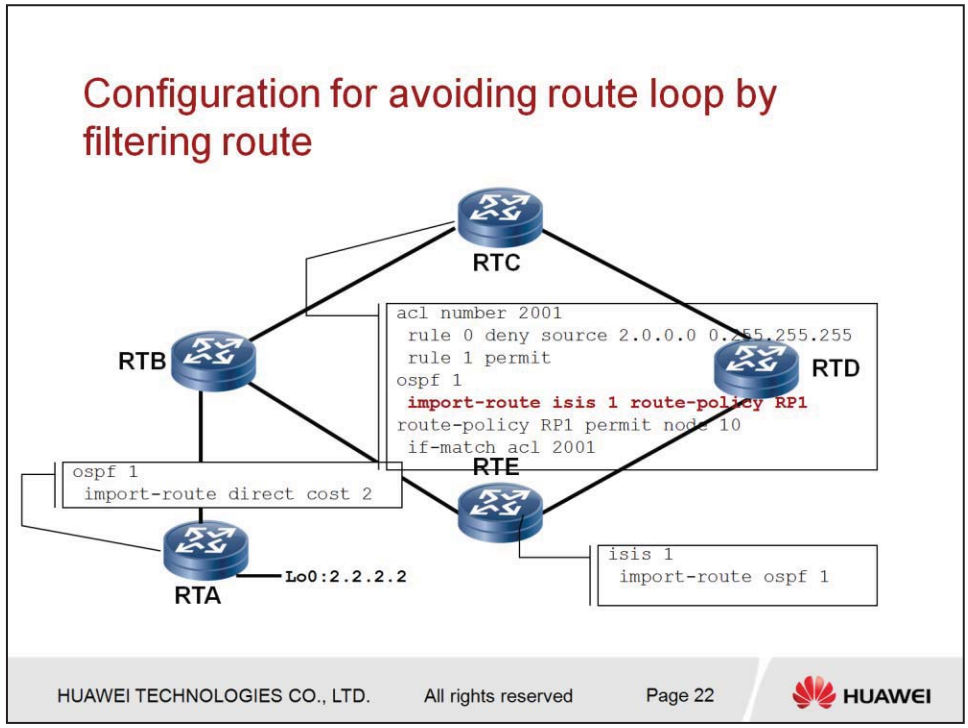


Configure filter-policy on RTB to filter ISIS route 2.0.0.0, then RIP route 2.0.0.0 is listed in routing table.

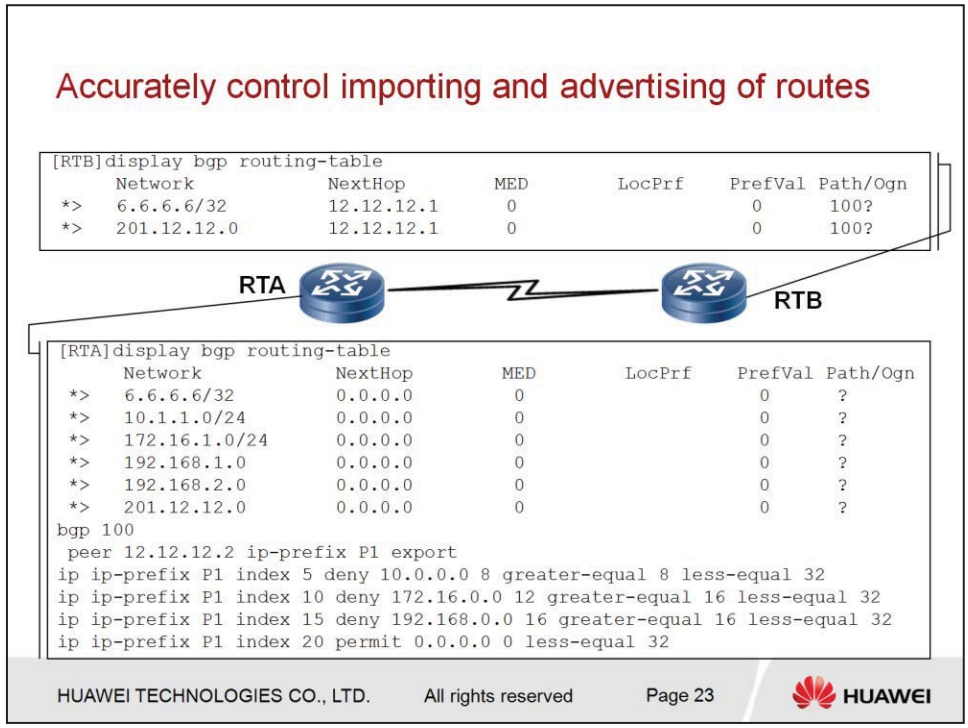


RTA imports direct route 2.2.2.2 into OSPF, and set the metric as 2, OSPF advertises 2.2.2.2 to RTB, RTC, RTE as ASE route (preference is 150). On RTE, importing OSPF-ASE is configured to import 2.2.2.2 into ISIS. On RTC, importing ISIS is configured to import ISIS route into OSPF, network segment 2.2.2.2 is advertised to OSPF by ISIS, which is called route feedback. In this case, RTB learns route 2.2.2.2 from RTA and RTC at the same time. Because the preference is the same (OSPF ASE route), it compares metric value, if RTB choose the route advertised by RTC, loop occurs. For example, RTD sends packet to 2.2.2.2, packet is first sent to RTE, and then to RTB; because RTB choose RTC route, RTB sends packet to RTC, RTD, and at last, reaches the originator.

Here, route loop is caused by route feedback. So, as long as we filter 2.2.2.2 when configuring route import on RTC, route loop can be avoided.



On RTC, route policy can be used to filter route 2.0.0.0 when importing IS-IS routes into OSPF, so as to avoid routing loops.

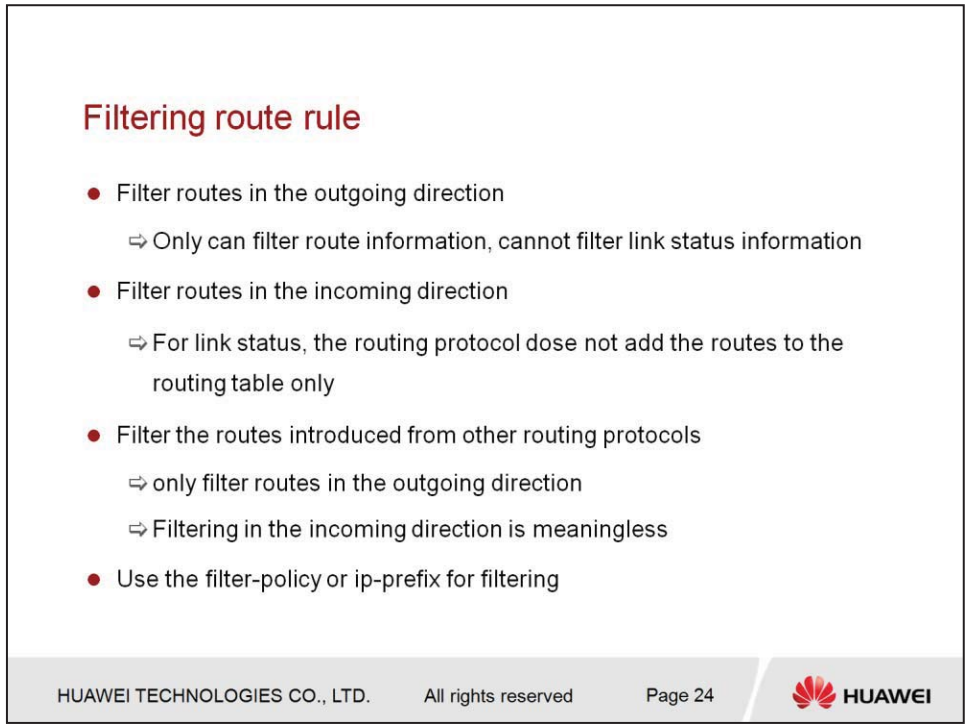


When we advertise routes, it is not expected to advertise private route to public network, sometimes, it is necessary to hide the route information of private network.

Route filter can be used to control the advertisement of route information precisely.

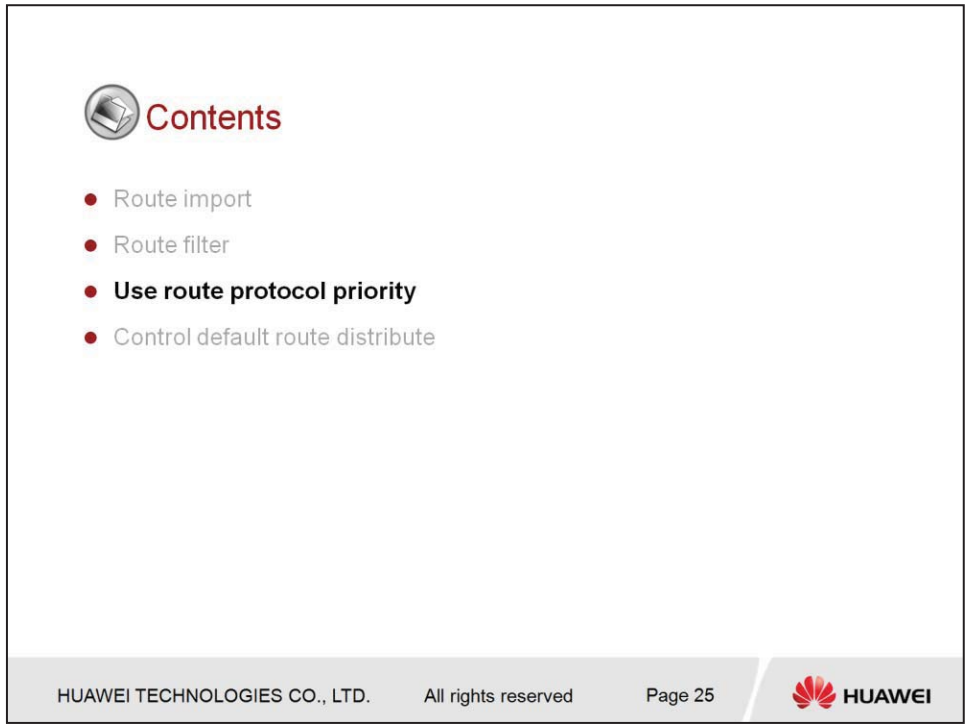
When import route, it is not expected to import all the routes but only some special ones, route filter can be used to control route import precisely.

In this example, routing table has three types of private routes, we need to define prefix list to filter private route. "ip ip-prefix P1 index 5 deny 10.0.0.0 8 greaterequal 8 less-equal 32" can filter private route 10.0.0.0~10.255.255.255; "ip ipprefix P1 index 10 deny 172.16.0.0 12 greater-equal 16 less-equal 32" can filter route 172.16.0.0~172.31.255.255; "ip ip-prefix P1 index 15 deny 192.168.0.0 16 greater-equal 16 less-equal 32"can filter route 192.168.0.0~192.168.255.255; "ip ip-prefix P1 index 20 permit 0.0.0.0 0 less-equal 32"permits other routes to pass.



Route filter can only filter route information, link state information can not be filtered. For OSPF, only type 3,5,7 routes can be filtered. The filter direction can be ingress and egress. For link state routing protocol, for example OSPF and ISIS, filtering route at ingress can not prevent the transmission of link state information, it can only prevent the route from being added to local routing table, its neighbors still can receive complete link state information and calculate the route.

Route filter can filter route imported from other protocols, for example, importing RIP route into OSPF, OSPF can use route filter to filter some RIP route.

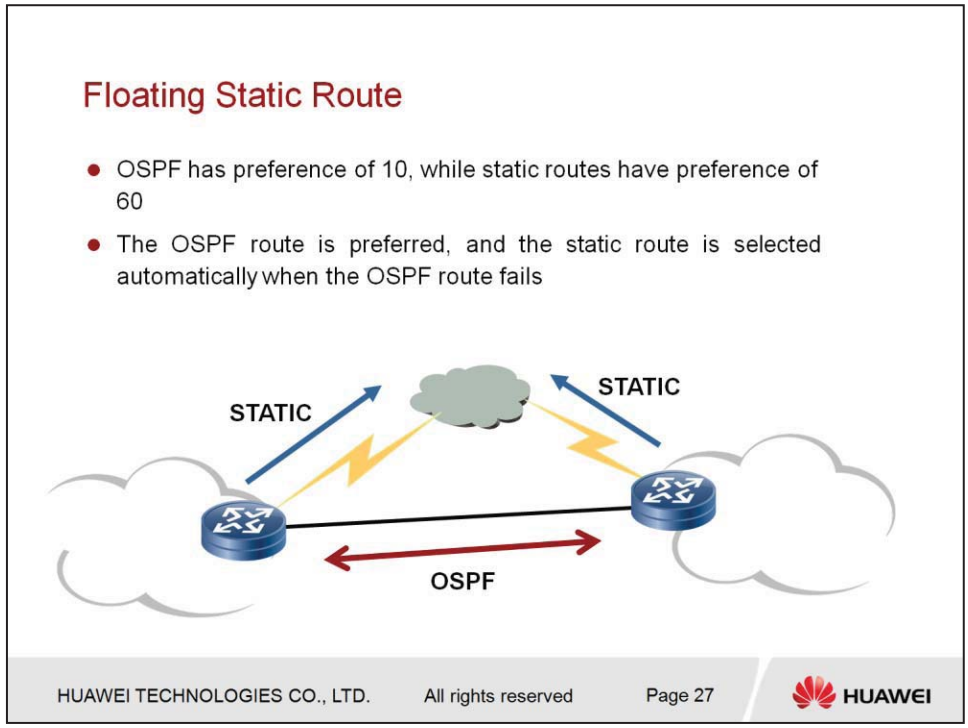

Contents
Route import
Route filter
Use route protocol priority
Control default route distribute
HUAWEI TECHNOLOGIES CO., LTD.
All rights reserved
Page 25
HUAWEI

Preference of Routing Protocol

Protocol	Preference
Direct	0
OSPF	10
IS-IS	15
Static	60
RIP	100
OSPF ASE	150
BGP	255

HUAWEI TECHNOLOGIES CO., LTD. All rights reserved

The routing protocol preference defined by different manufacturers is different. The function of preference is to distribute preference to routes discovered by different protocols, then when a router learns the same route from different routing protocols, it can choose the route according to the preference.

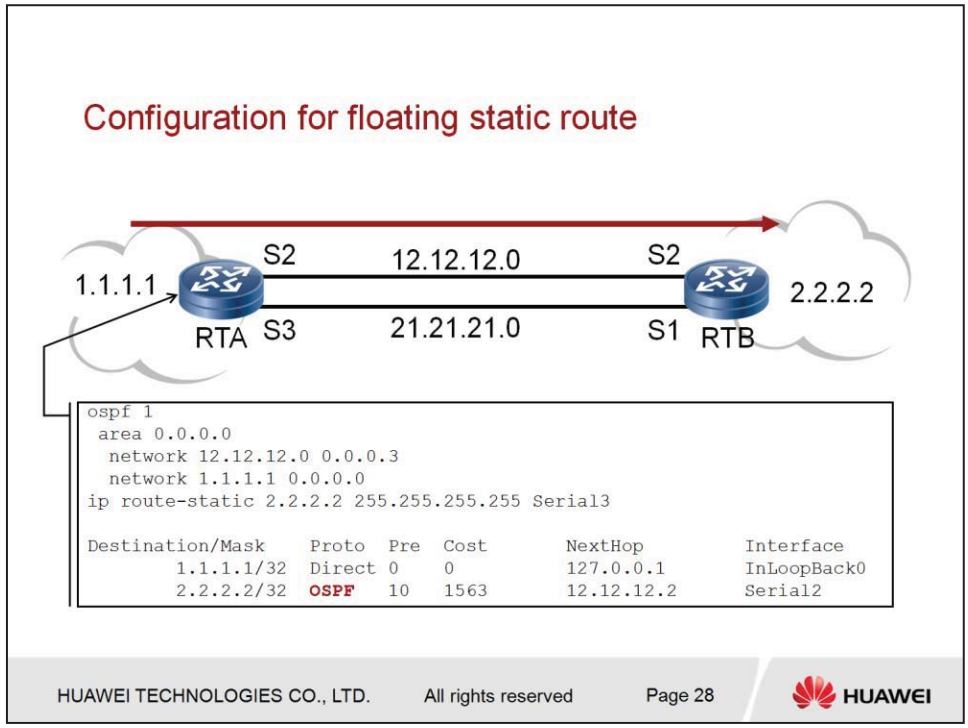


Float static route is a typical application for preference of routing protocol. In many cases, master/backup links are used to connect remote network. Dynamic routing protocol runs on the master link, for example OSPF,ISIS and so on. Backup link is

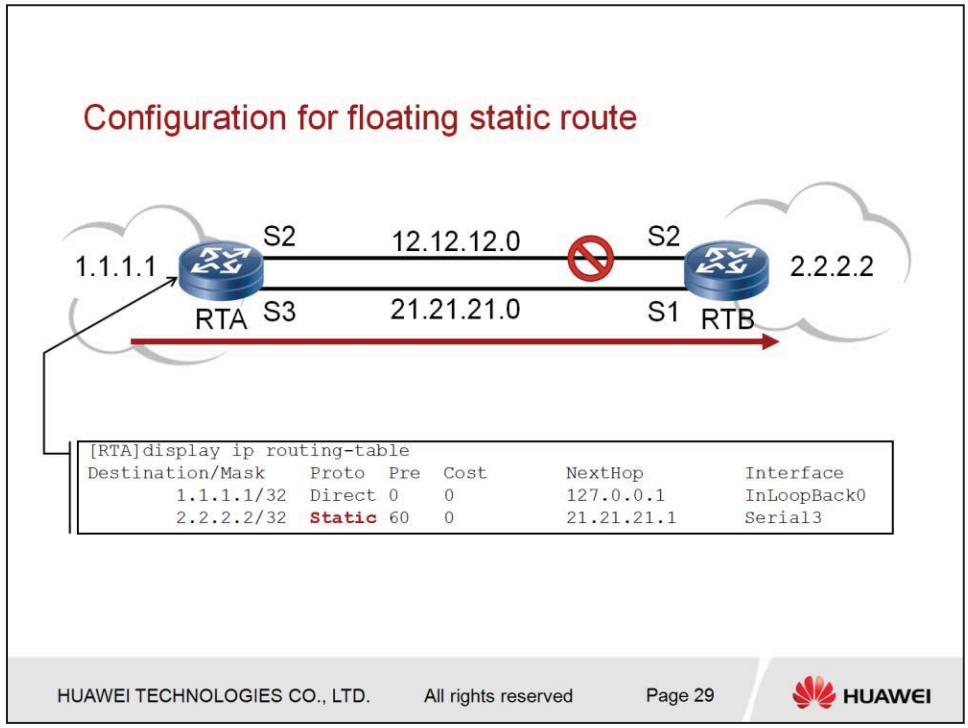
Dial-up link, it is expensive and accounted according to the connection time, so on the Dial-up link, we do not configure dynamic routing protocol but configure a static route to the remote network. If the master link is normal, router can learn the same route from OSPF and static configuration. But the preference of static route is lower than that of OSPF, router chooses OSPF route and data packet will is forwarded

through master link. If the master link is down, OSPF neighbors are disconnected, the OSPF route is invalid and cleared from the routing table. In this case, the static route is valid and added into routing table, the data packet is forwarded though backup link. When the master link restores OSPF neighbor relation is established again, the OSPF route replaces the static route and the backup link is down again.

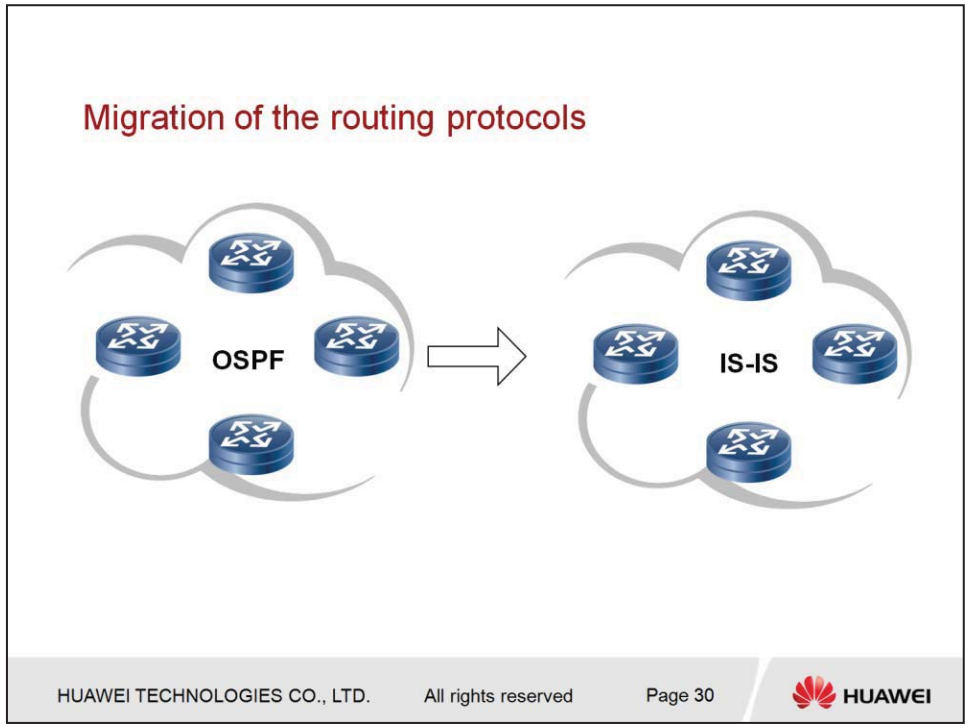
This kind of configuration not only saves cost, but also strengths reliability of the network. But it can not implement load balance.



In this example, RTA has two links connected to RTB and interface S2 establishes neighbor relationship with RTB and learns the route about network segment 2.2.2.2; besides, the static route destined for 2.2.2.2 is configured and the egress of the static route is S3. If the two links can works normally, RTA choose OSPF route to 2.2.2.2. In the routing table, we can see the route destined for network segment 2.2.2.2 is learnt by OSPF.



If there is something wrong with the link above, RTA and RTB cannot exchange hello packet, so the neighbor relation is invalid and the route 2.2.2.2 learnt by OSPF in the routing table of RTA is valid. In this case, the static route configured before occurs in the routing table and RTA can access 2.2.2.2 through interface S3 .



Sometimes, we need to use new routing protocol to replace the current routing protocol and as much as possible to reduce the broken time of network during transition.

For example, the network is running OSPF, which is to be transitioned to ISIS. The two type of protocols can run at the same time on every router. Modify the preference of the two protocols, so that ISIS can only run backstage at first. After confirming that ISIS neighbor relationship and LSDB are in good condition, modify the preference of ISIS and make it replace the current routing protocol.

The backstage steps includes:

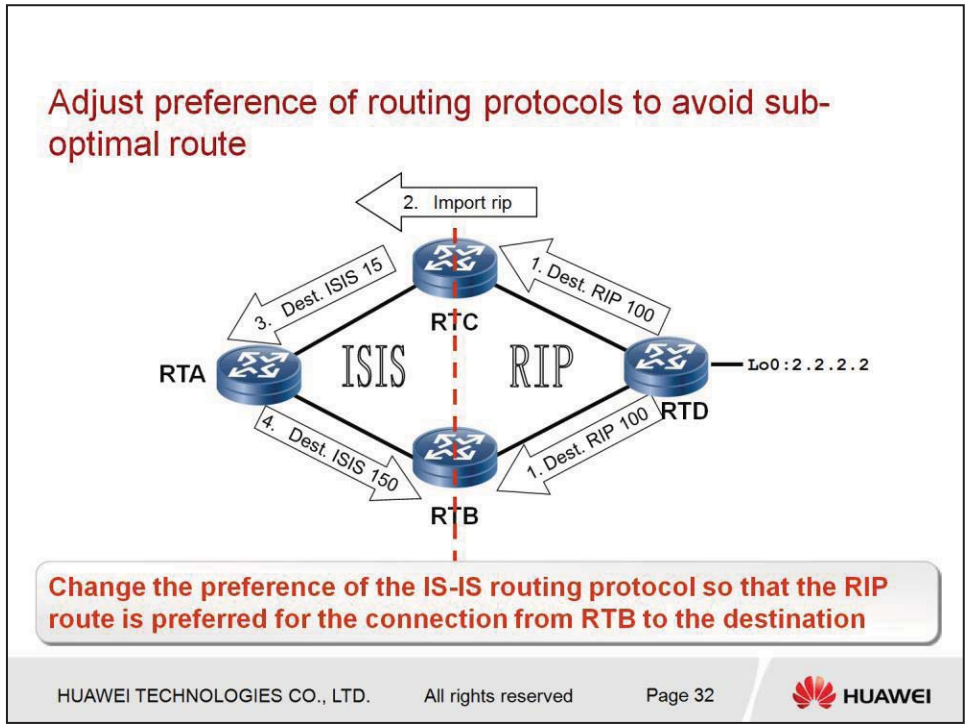
1、Only OSPF runs in the network, check both hardware and software , upgrade to insure hardware and software can support transfer.

2、Configure ISIS, set proper preference, insure that ISIS only runs backstage; in this case, ISIS link state database is established on every router, but routing table and forwarding table dose not change. In this phase, verify the ISIS running state, only each router has established LSDB and all the expected LSP is existed in database, ISIS route that can reflect IP routing table can already be generated,

this phase can be ended.

3、Change preference and make ISIS replace the primary IGP, so that ISIS can run front stage.

4、Delete the primary IGP after the network work normally.



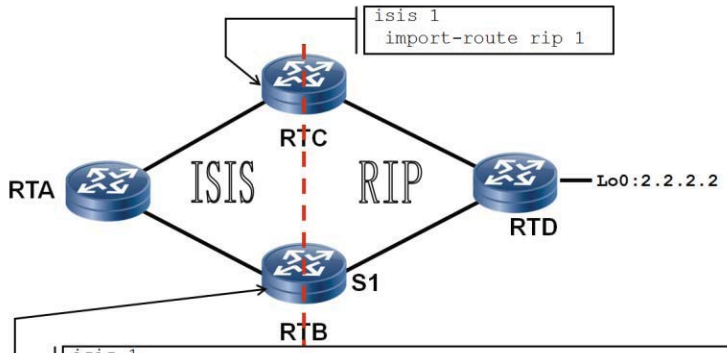
Adjusting preference of a routing protocol properly can avoid sub-optimal route.

In this topology, change preference of ISIS on RTB and then RTB chooses RIP route preferentially to avoid sub-optimal route.

Note:

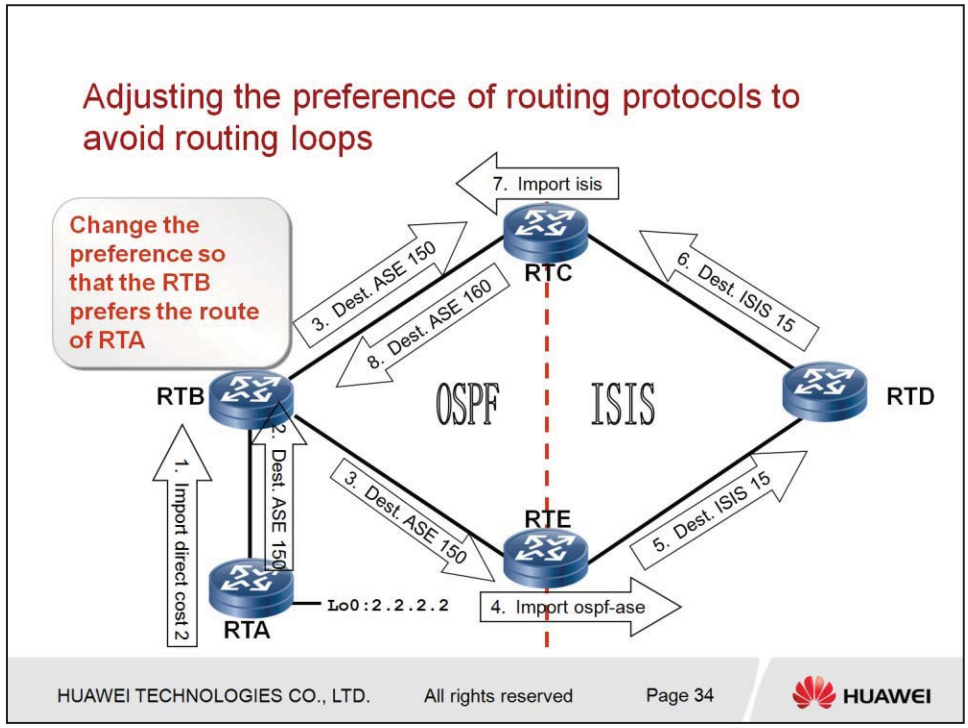
When adjusting protocol preference, we must be careful and avoid new problem that causes route confusion.

Configuration for adjusting preference of routing protocols to avoid sub-optimal routing




```
isis 1
 preference 150
[RTB]display ip routing-table
Destination/Mask    Proto   Pre  Cost      NextHop         Interface
        2.0.0.0/8   RIP     100  1         24.24.24.2      Serial1
```

HUAWEI TECHNOLOGIES CO., LTD. All rights reserved



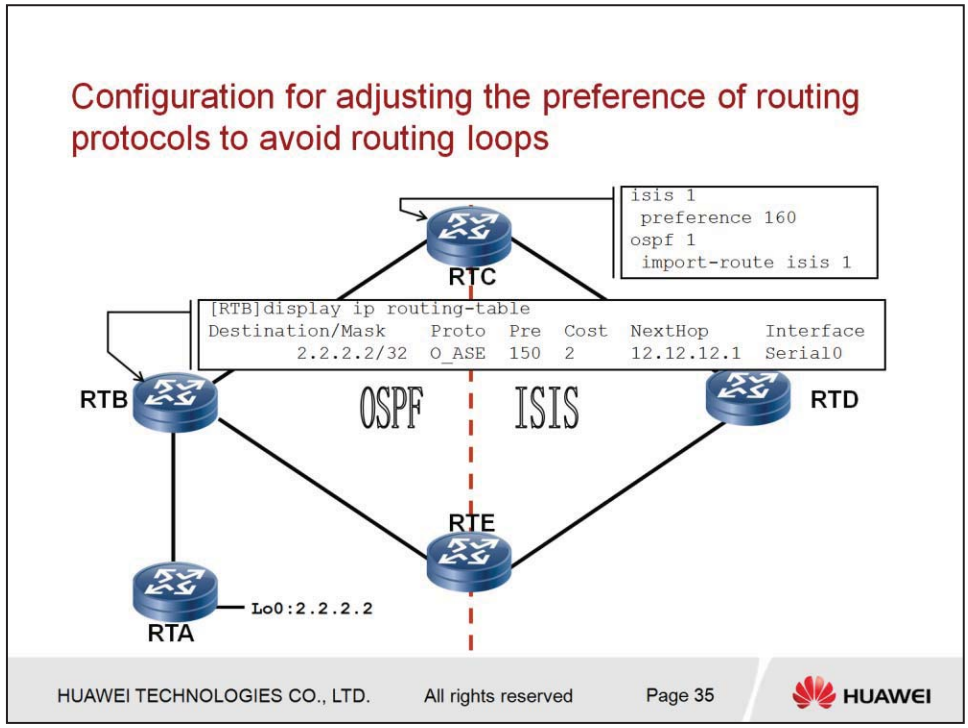
In this example, direct routes are imported into OSPF in RTA.

1.RTB advertises 2.2.2.2/32(ospf-ase pref 150) to both RTC and RTE.

2.In RTE, 2.2.2.2/32(ospf-ase pref 150) is imported to ISIS. So 2.2.2.2/32 will be in RTD's routing table,but it is an ISIS routing with pref 15 3. RTC will receive 2.2.2.2/32(isis pref 15) from RTD. Also please remember at step1 RTC received 2.2.2.2/32(ospf-ase pref 150) from RTB.RTC choose

2.2.2.2/32(isis pref 15) as best route because of lowest pref.

4. Just suppose now we import ISIS into OSPF in RTC now, there will be two 2.2.2.2/32 (both are ospf-ase pref 150)items in RTB's routing table.But one is from RTC with RTC as next hop(this one will cause loop), another one is from RTA with RTA as next hop.



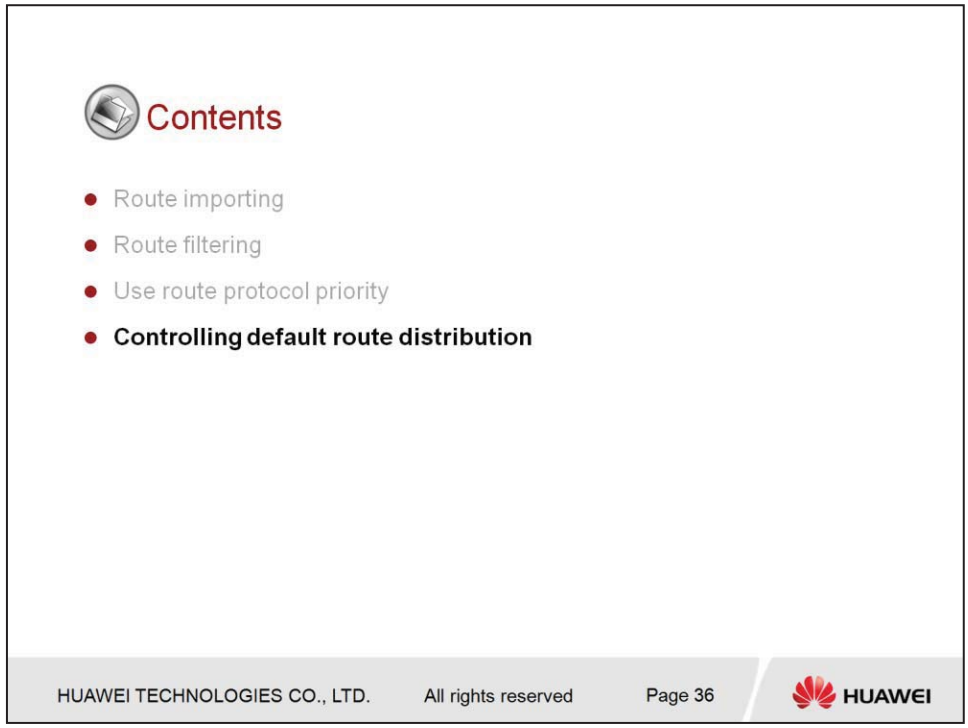
Now if we change isis preference in RTC as shown in the slide. Let's look at what will happen.

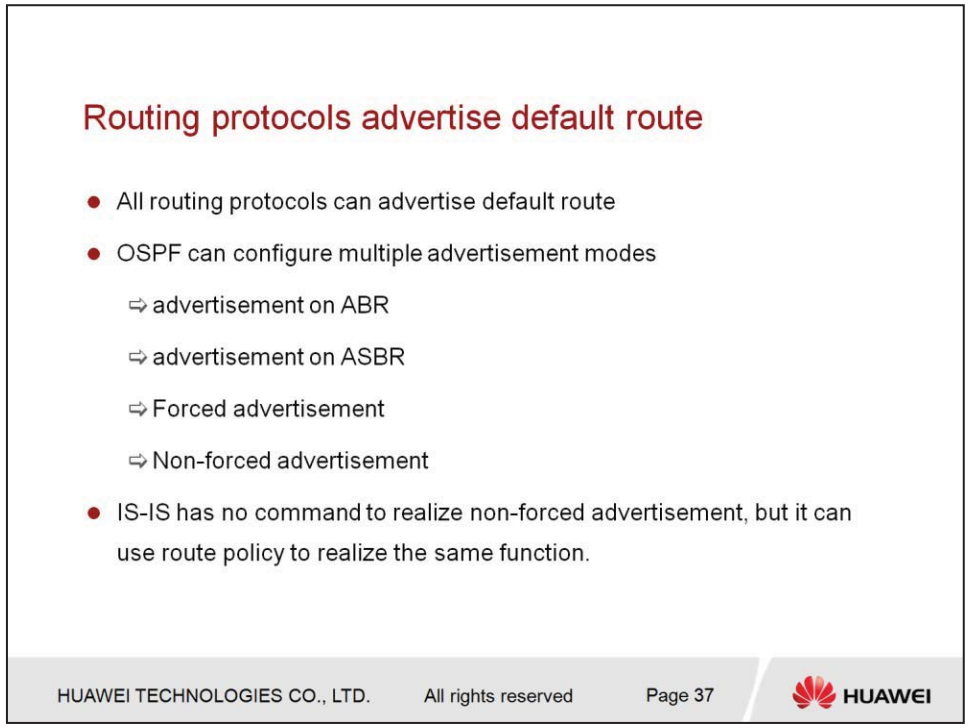
Step1 and Step2 are same.

3. RTC will receive 2.2.2.2/32(isis pref 15) from RTD. But we already changed the preference in RTC. So RTC will think the preference of 2.2.2.2/32(isis) is 160. Also please remember at step1 RTC received 2.2.2.2/32(ospf-ase pref 150) from RTB.RTC choose 2.2.2.2/32(ospf-ase pref 150) as best route because of lowest pref.

4. Just suppose now we import ISIS into OSPF in RTC now, just because now in RTC's routing table,2.2.2.2/32 is ospf-ase routing, it will not be 're-imported'.

So finally, RTC choose 2.2.2.2/32 (ospf-ase pref 150) with RTB as next-hop, RTB will use2.2.2.2/32 (ospf-ase pref 150) with RTA as next-hop, RTD choose 2.2.2.2/32 (isis pref 15) with RTE as next-hop. No loop occurs.


Contents
Route importing
Route filtering
Use route protocol priority
Controlling default route distribution
HUAWEI TECHNOLOGIES CO., LTD.
All rights reserved
Page 36
HUAWEI



It is complex for OSPF to distribute default route. Different area types and configuration position distribute default route differently.

The area type of OSPF includes: ordinary area, STUB area, total STUB area, NSSA area.

The area defaults to general area. General area can be standard area or backbone area. Standard area is the most common area, it carries intra-area route, inter-area route and external area route. Backbone area connects other OSPF areas together. In general area, default route is not be generated in default.

We can configure forced advertisement default route or non-forced advertisement TYPE 5 route on ASBR.

If non-forced advertisement is configured, a router advertises a default route only when an active default route that is not advertised by the it's own OSPF process, exists in the routing table. it advertises default route. At the same time, router learns TYPE5 default route advertised by other OSPF routers.

If the default route advertised by other OSPF router is prior than the active default route in the local routing table, the router will use the route advertised by OSPF route and stop advertising

TYPE5 default route. If forced advertisement is configured, no matter whether default route exists in the routing table, the router advertises default route. At the same time, it does not learn TYPE5 default route advertised by other OSPF routers.

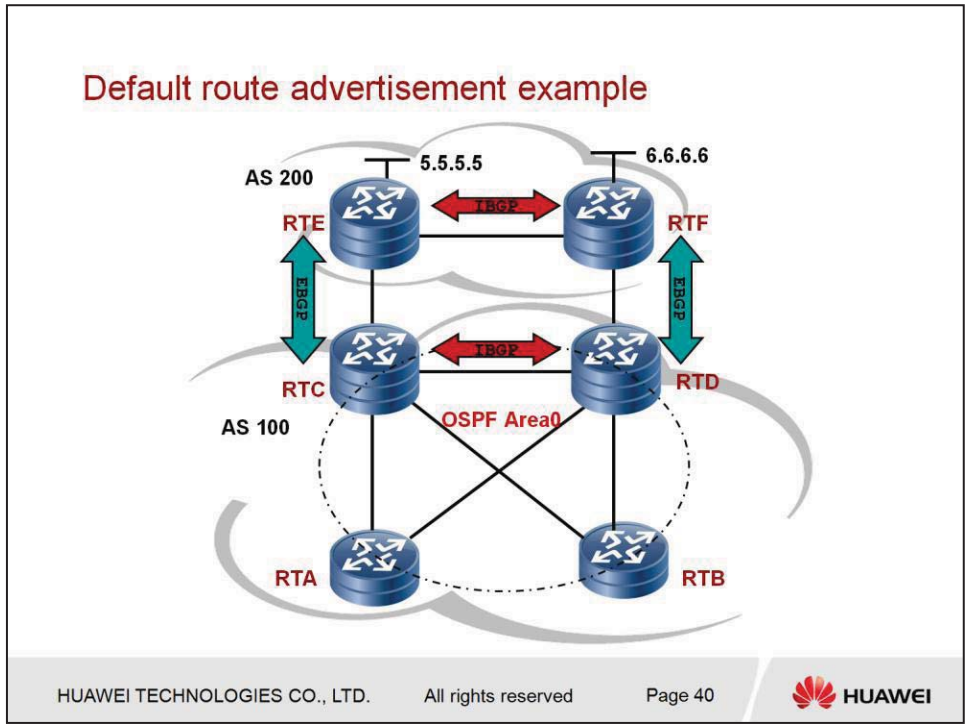
TYPE5 LSA is flooded in the whole routing domain.

OSPF advertisement of default route

Area type	Auto	command	generator	Existing default route	LSA type	scope
ordinary		default-route-advertise	ASBR	√	5	Route domain
		default-route-advertise always	ASBR	√	5	Route domain
STUB	√		ABR		3	STUB
Total STUB	√		ABR		3	STUB
NSSA		nssa default-route-advertise	ABR		7	NSSA
		nssa default-route-advertise	ASBR	√	7	NSSA

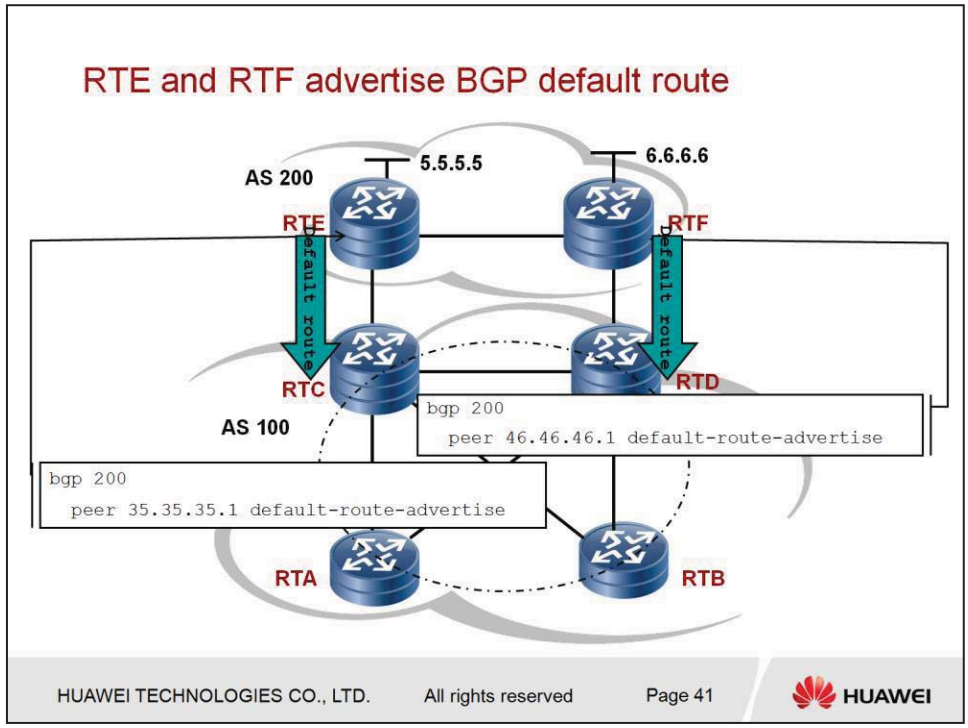
HUAWEI TECHNOLOGIES CO., LTD. All rights reserved

Distribution of default route for OSPF is complex, for the details, please refer to the table above.

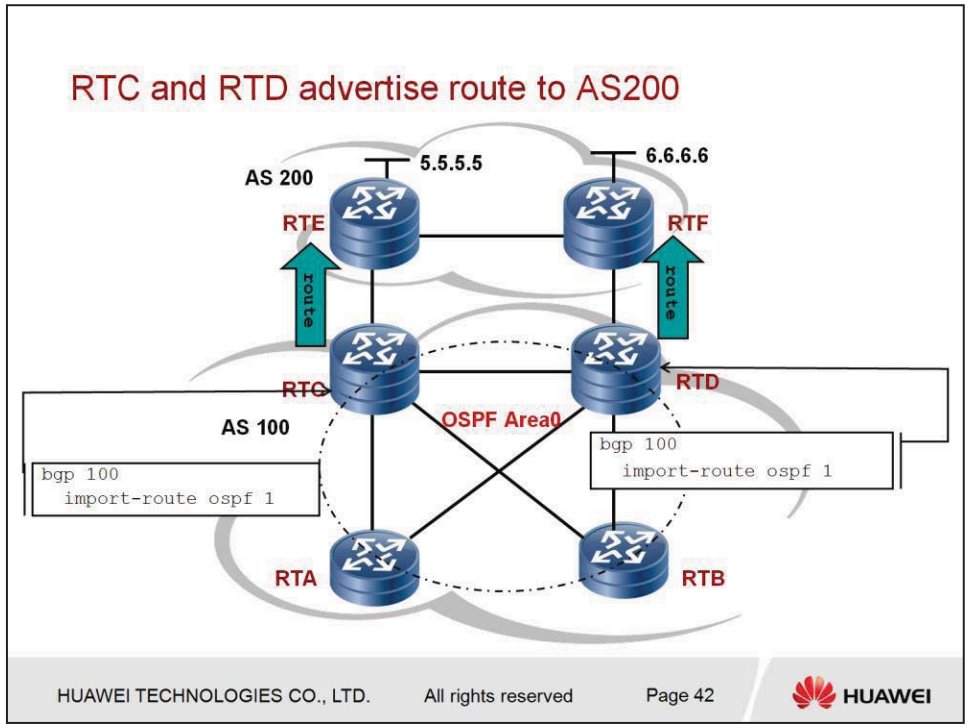


The network above is popular in the real network. AS100 has two egress routers RTC and RTD that connect to AS200, EBGP neighbor relationship is established between RTE and RTC, as well as RTF and RTD; IBGP neighbor relation is established between RTE and RTF, as well as between RTC and RTD.

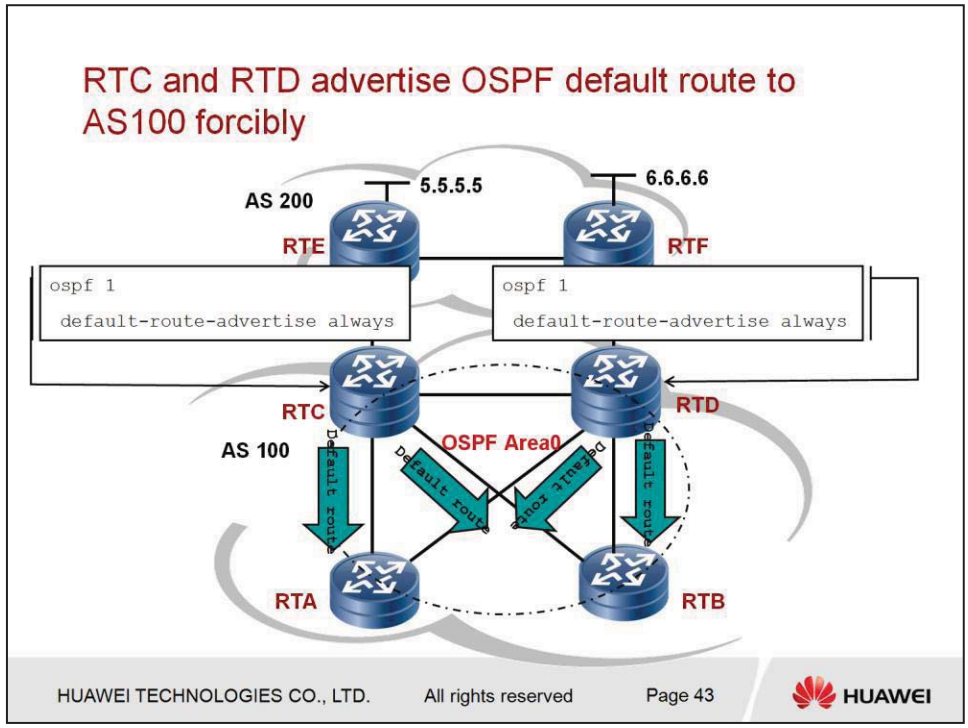
AS100 is an OSPF routing domain.



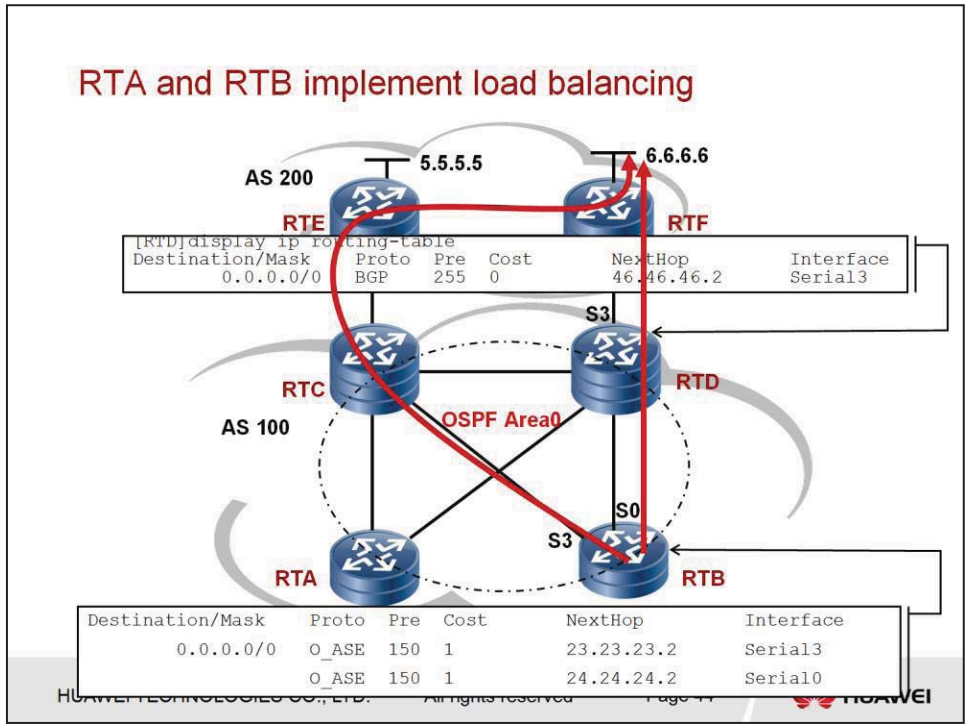
AS100 does not need to know all the BGP routes of AS200, so distribute the BGP default route to RTC and RTD from RTE and RTF.



Import OSPF route into BGP and then AS200 can forward the data packet to AS100.

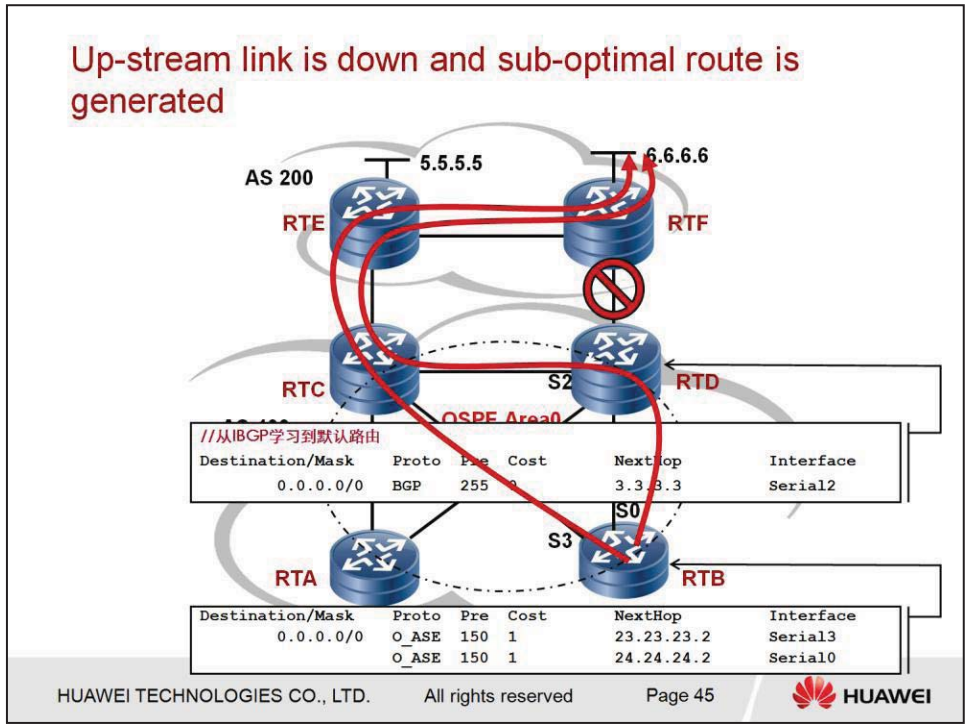


RTA and RTB do not run the BGP protocol, so they can not access AS200. RTC and RTD distribute OSPF default route forcibly and then RTA and RTB can use default route to send data packet to egress router RTC and RTD , RTC and RTD can forward data packet to AS200 according to BGP default route.



RTA and RTB learn OSPF default route from RTC and RTD, to form load balance, which can be verified according to the routing table.

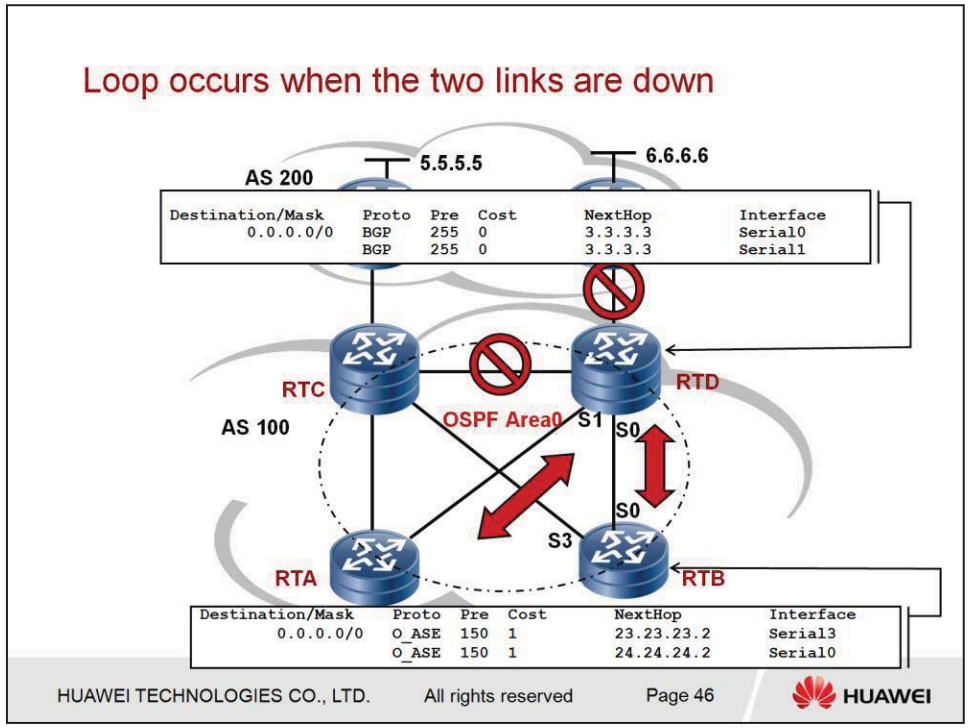
When RTA or RTB send packets to AS200, some packets choose RTC as the egress, while others choose RTD.



This kind of network can work well if link is normal, but once the link is down, problems occur.

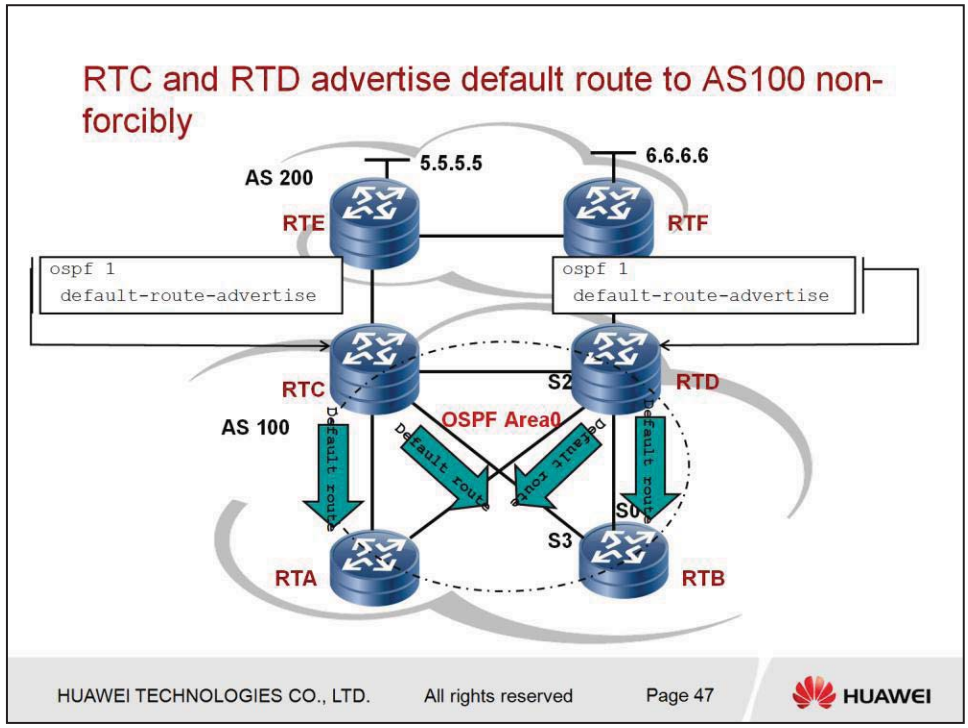
In the network, the up link of RTD is down, RTD can not learn BGP default route from RTF, but it can learn default route from RTC (the next-hop of RTD default route is 3.3.3.3, this is the loopback address of RTC). RTA and RTB can still access AS200, but partial traffic go through the link between RTC and RTD,

which is the sub-optimal route. moreover, if the network is not well designed, congestion occurs between RTC and RTD.



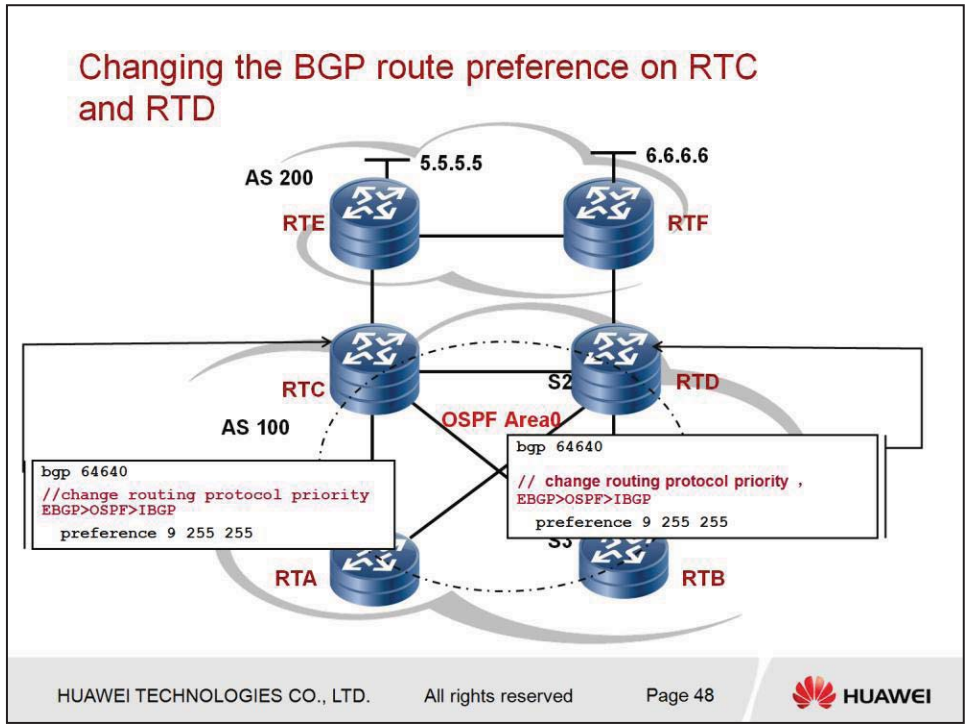
The more serious problem is that the up link of RTD and link between RTC and RTD are both down at the same time. In this case, RTD can neither receive EBGP default routes from RTF, not IBGP default route of RTC through the link between RTC and RTD. But the IBGP neighbor relationship between RTC

and RTD is valid, so RTD can learn IBGP default route advertised by RTC through RTA and RTB, and implement load balancing. When RTB sends data packets to AS200, because of the load balancing, partial packets are sent to RTC, which sends the packets correctly according to EBGP default route; other packets are sent to RTD, so problem occurs now, RTD checks its routing table and finds two default routes, one is destined for RTA, and the other is destined for RTB, RTD also implements load balancing and sends partial packets to RTB, thus a loop occurs.

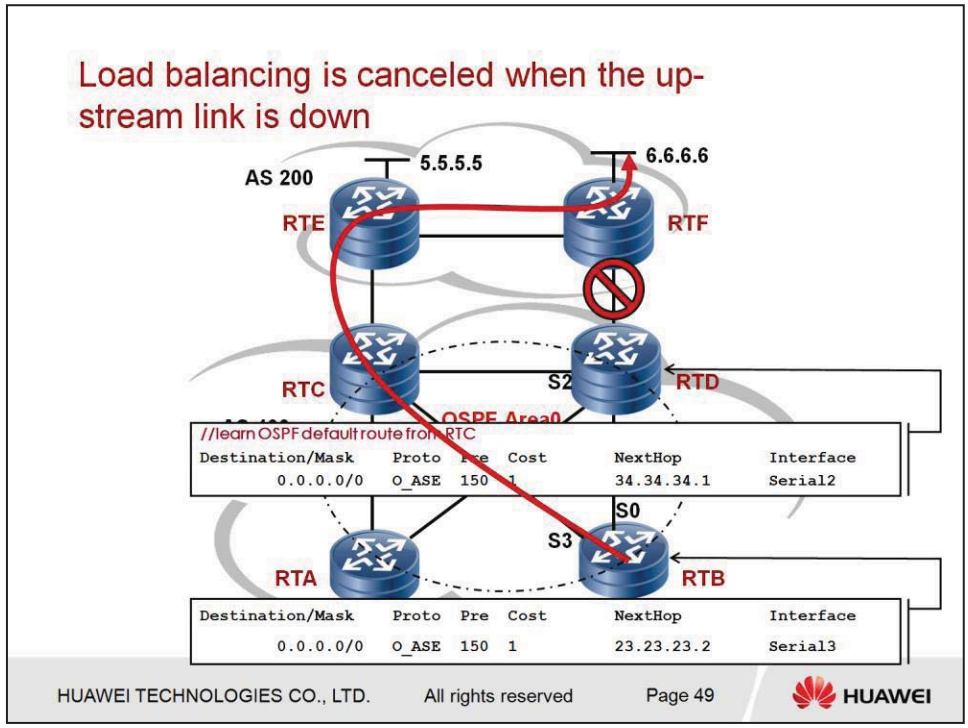


How can this problem be avoided? Configure advertisement default route on RTC and RTD non-forcibly.

To solve the problem mentioned before, configure non-forced advertisement default route on both RTC and RTD.

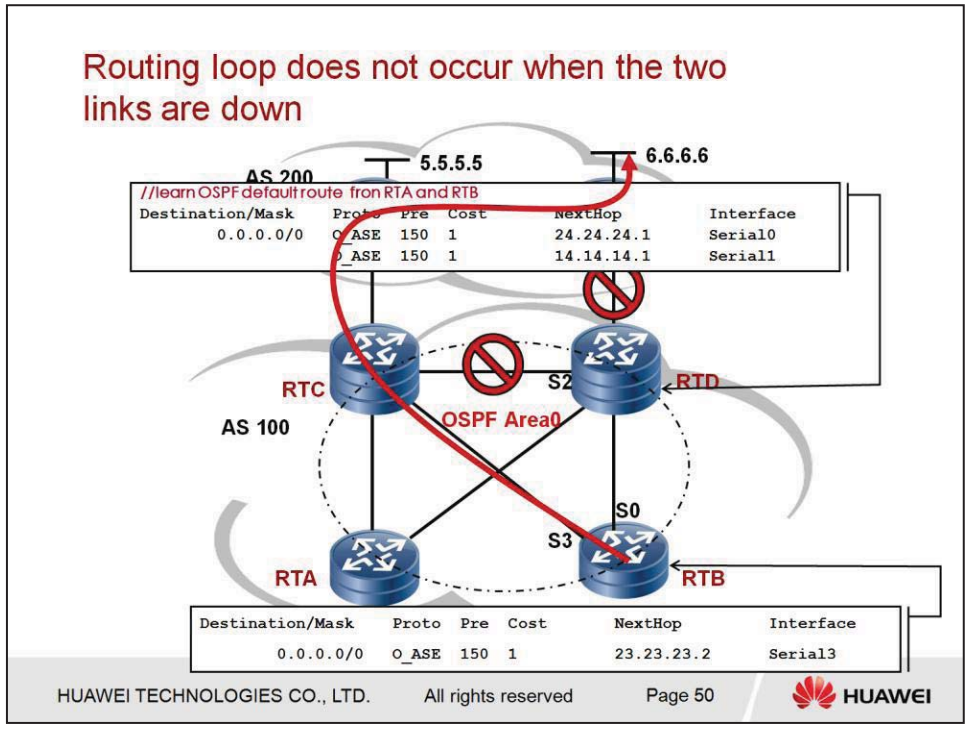


Besides, adjust preference of routing protocol on RTC and RTD to make the EBGP route preference higher than that of OSPF, and OSPF route preference higher than that of IBGP.

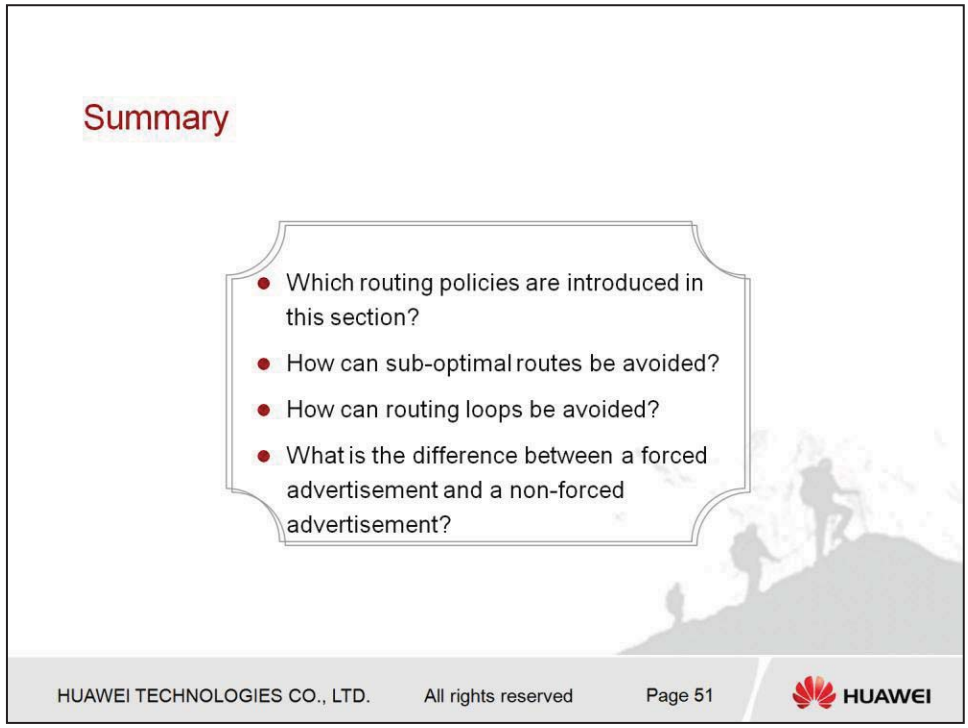


When the uplink of RTD is down, RTD cannot learn EBGP default route from RTF, but it can learn IBGP default route from RTC. In the routing table of RTD, only OSPF default route exists (the preference of OSPF is higher that that of IBGP).

According to the condition of non-forced advertisement default route, that exists in it's routing table , OSPF default route learnt from other routers, RTD dose not advertise any default route. On RTA and RTB, only one default route pointing to RTC exists. RTA and RTB will take RTC as the egress to send packets to AS200.



When the two links are down, RTD does not advertise a default route, therefore the problem can be avoided.



Q：Which routing policies are introduced in this section?

A：Route import; route filter; adjusting of the priority of a routing protocol; default route advertisement.

Q：How can sub-optimal routes be avoided?

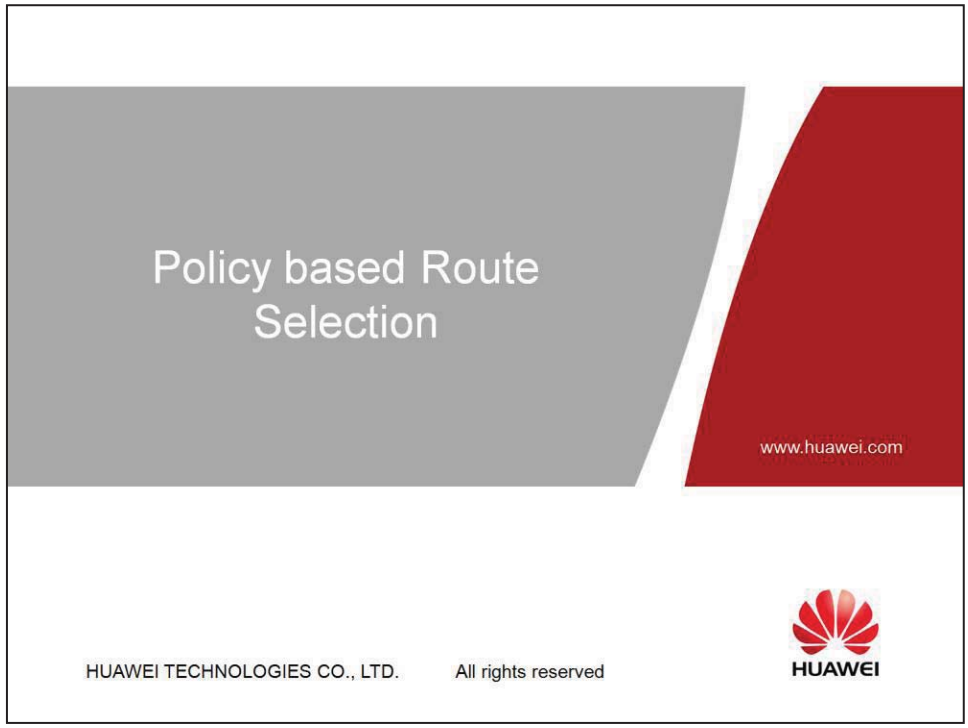
A：Through route filter and the changing priority of routing protocol can avoid sub-optimal routes.

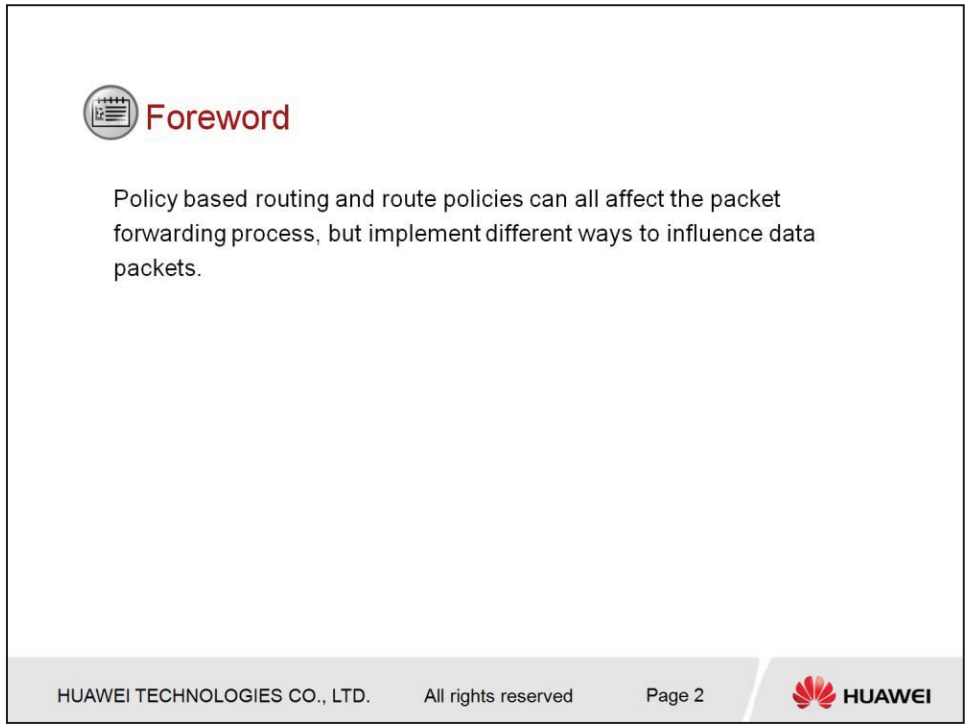
Q：How can routing loops be avoided?

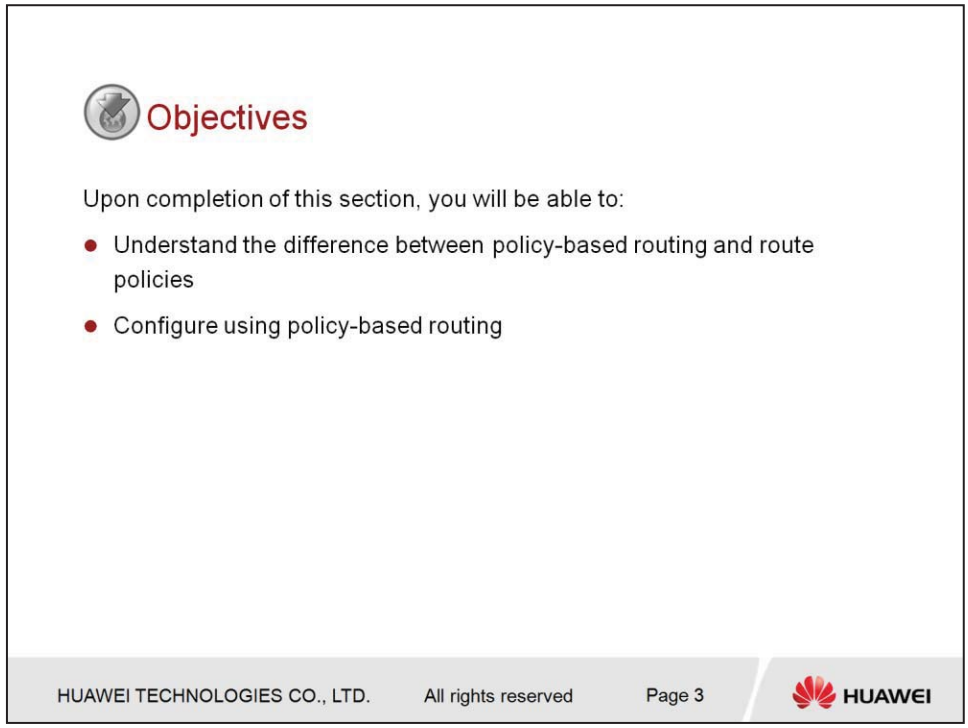
A：Route filtering and by changing priority of routing protocol can be used to avoid routing loops.

Q：What is the difference between a forced advertisement and a non-forced advertisement?

A：A “Non-forced advertisement” should satisfy some conditions (for example, a specified route exists), a “forced advertisement” can advertise a default route without satisfying any condition.


Policy based Route Selection
www.huawei.com
HUAWEI TECHNOLOGIES CO., LTD.
All rights reserved
HUAWEI


Foreword
Policy based routing and route policies can all affect the packet forwarding process, but implement different ways to influence data packets.
HUAWEI TECHNOLOGIES CO., LTD.
All rights reserved
Page 2
HUAWEI


Objectives
Upon completion of this section, you will be able to:
Understand the difference between policy-based routing and route policies
Configure using policy-based routing
HUAWEI TECHNOLOGIES CO., LTD.
All rights reserved
Page 3
HUAWEI

Policy-based route selection

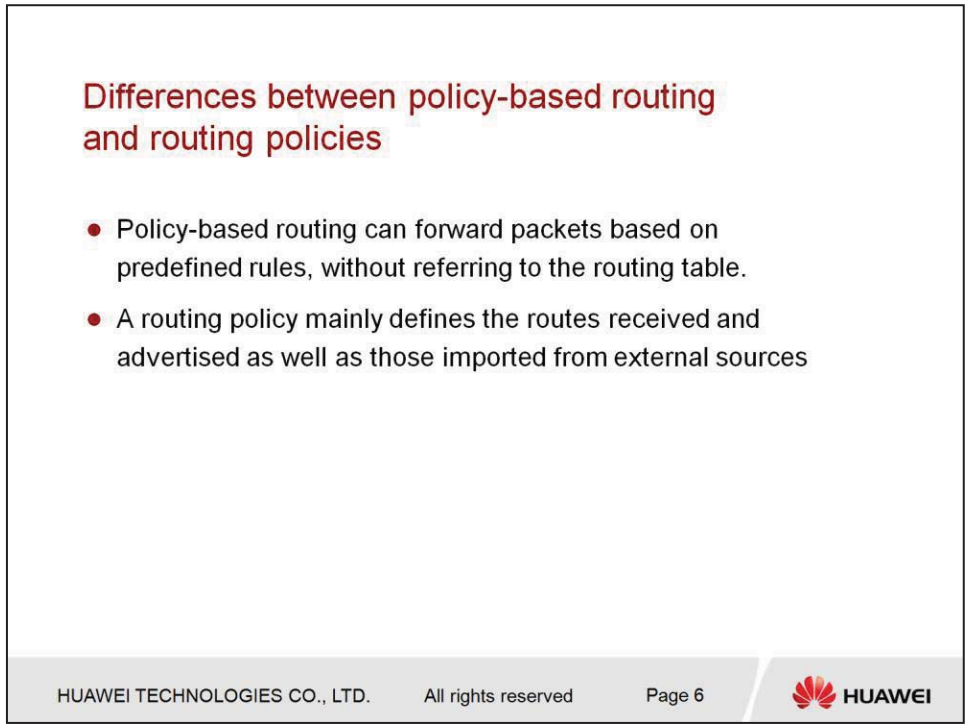
- Policy-based routing is a mechanism for packet forwarding path selection according to the customized policy.
- Unlike the forwarding that is performed by a lookup in the routing table according to the IP destination address, policy based routing is used in security, QoS and load balancing.
- Policy-based routing supports to designate routes flexibly according to the ACL, address length and so on..

HUAWEI TECHNOLOGIES CO., LTD. All rights reserved

Function of policy-based routing

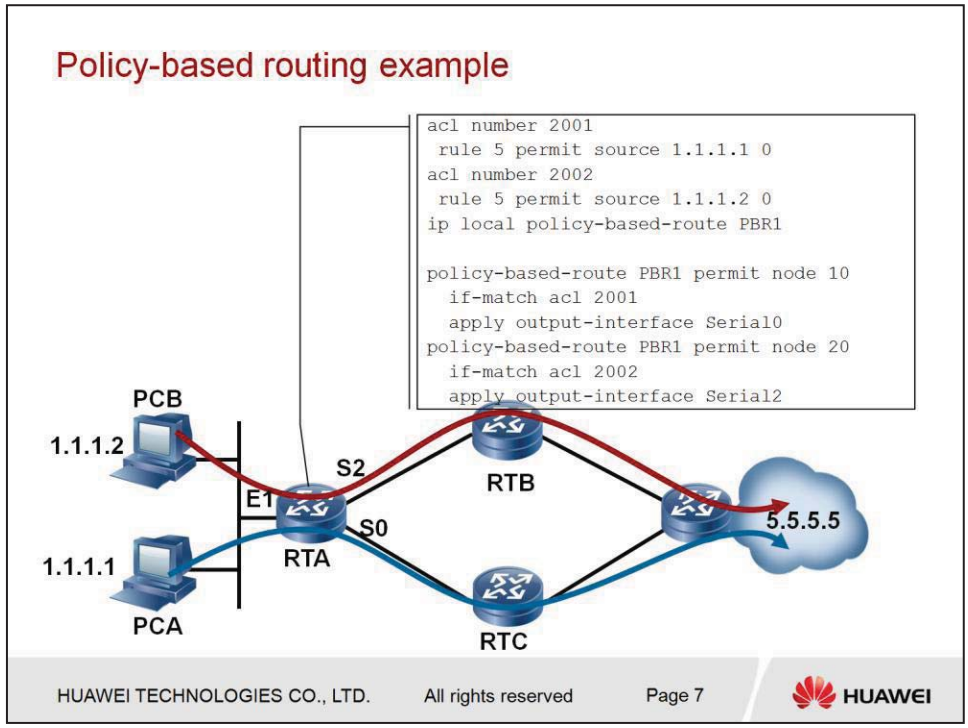
- Compared with traditional routing protocols, policy-based routing can provide network administrators with more powerful control over the forwarding and storing of packets.
  - ⇨ Network administrators can select the forwarding paths according to not only the destination addresses but also protocol types, packet sizes, applications, IP source addresses or other policies
  - ⇨ Policies can control load balancing between multiple routers, the QoS of packet forwarding on a single link or to meet a specific need.

HUAWEI TECHNOLOGIES CO., LTD. All rights reserved



Policy-based routing mainly controls the forwarding of packets, namely, the router can forward the packet according to the routing policy instead of the routing table (because the general forwarding needs to lookup the forwarding table, but after configure the policy, it does not need to lookup forwarding table anymore).

Routing policy mainly controls the route information import (which route information can be imported, which can not), distribution (which information can be advertised out, which can not) and receiving (control the receiving or discarding of information).



In this topology, if there is no load balancing, the route for destination 5.5.5.5 in RTA routing table goes though RTB or RTC. We can configure policy-based route on RTA to make the packet from 1.1.1.1 be forwarded to 5.5.5.5 though RTC and the packet from 1.1.1.2 be forwarded to 5.5.5.5 though RTB.

Define the match rule of the policy-based route

1、Enter into system view：system-view

2、Create policy or policy node：policy-based-route po licy-name { deny | permit } node node-id

3 、Configure match condition of IP packet length ： if-match packet-length minimum-length maximum-length

4、Configure match condition of IP address：if-match acl ac l-number Policy include multiple policy nodes, which are appointed by sequence number node-id. The smaller the value, the higher the prior is, the corresponding policy is implemented first. If create policy-based-route repeatedly, the new configure will

covers the old one. The particular content of policy is appointed by if-match and apply sentence.

“Permit” means to implement the policy-based route on the packet that satisfies the match condition, “deny” means

oppositely. IP unicast policy-based route provides two methods to define packet: match according to the packet length and ACL rule.

One item of policy can include multiple if-match sentences.

Define the policy-based route

1、Enter system view：system-view

2、Create policy or policy node：policy-based-route po licy-name { deny | permit } node node-id

3、Configure the priority of a packet：apply ip-precedence p recedence

4、Configure the default next-hop of packet：apply ip-address default next-hop ip-address1 [ip-address2]

5、Configure default egress of a packet：apply default output-interface interfacetype1 interface-number1 [interface-type2 interface-number2]

6、Configure the next-hop of a packet：apply ip-address next-hop i p-address

7、Configure egress of a packet：apply output-interface interface-type1 interfacenumber1 [interface-type2 interface-number2]

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

“Apply” sentence is used to guide the forwarding of the packet that satisfies the match rule. One policy can include multiple “apply” sentences.

If we configure multiple next-hops or egresses, it adopts load balance to forward packets. The traffic can implement load balance among multiple nexthops or egresses. If egress and next-hop are configured at the same time, it only implements load balance among egresses.

Note:

Egress cannot be ethernet interface or other broadcast interfaces.

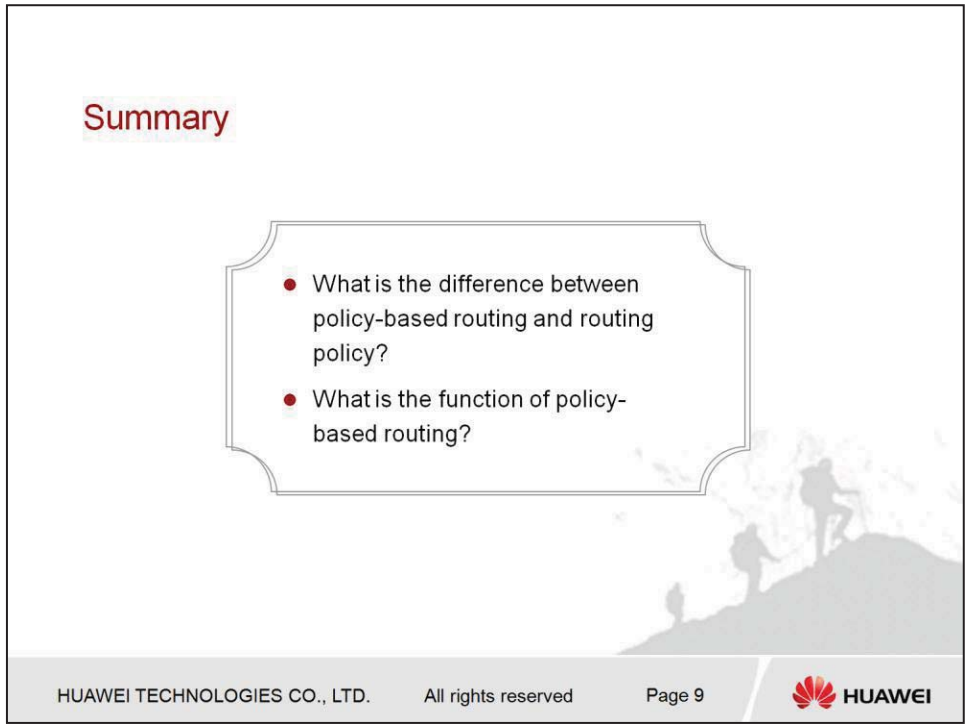
Apply policy-based route

apply policy-based route in system view, in this case, it only takes effect on local generated packet;

Start local policy-based route

1、Enter into the system view：system-view

2、Enable local policy-based route：ip local policy-based-route po licy-name Local policy route can only take effect on local generated packet. Only one local policy can be configured.



Q：What is the difference between policy-based route and routing policy?

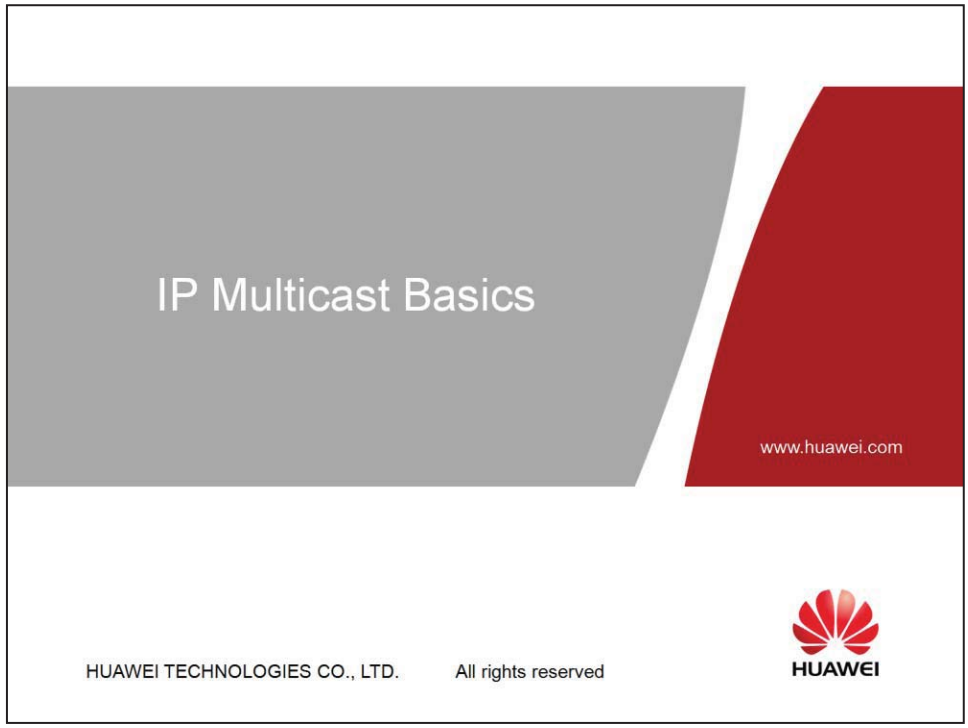
A： Policy-based route controls the forwarding of packets, namely, the router can forward the packet according to the routing policy instead of the routing table (because the general forwarding needs to lookup the forwarding table and after configure the policy, it does not need to lookup forwarding table any more).

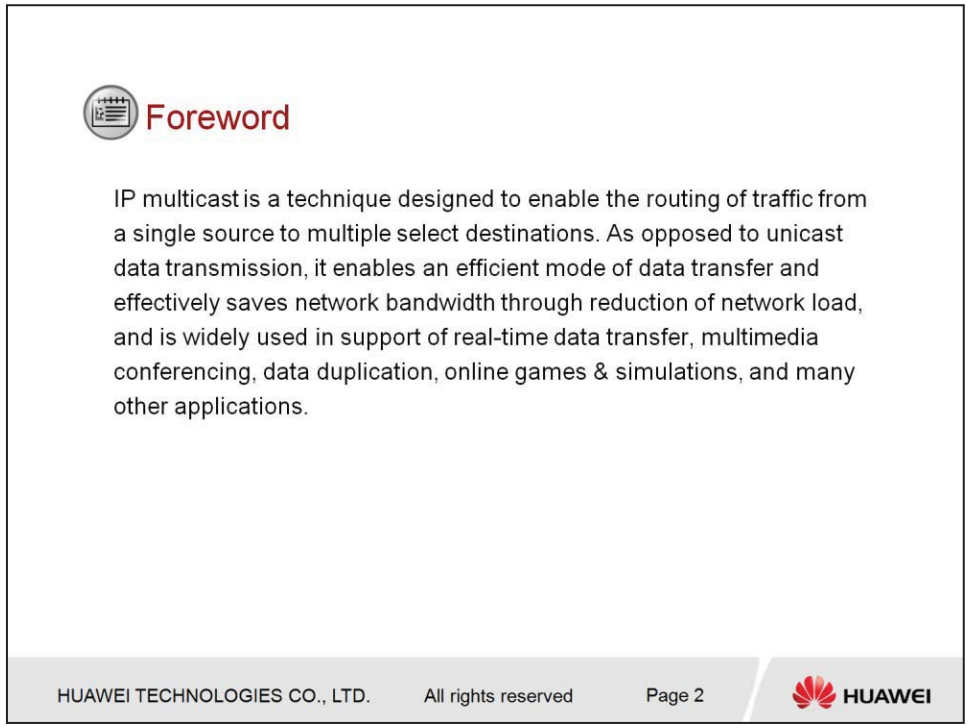
Routing policy controls the route information import (which route information can be imported, which one can not), distribution (which information can be advertised out, which one can not) and receiving (control the receiving or discarding of information).

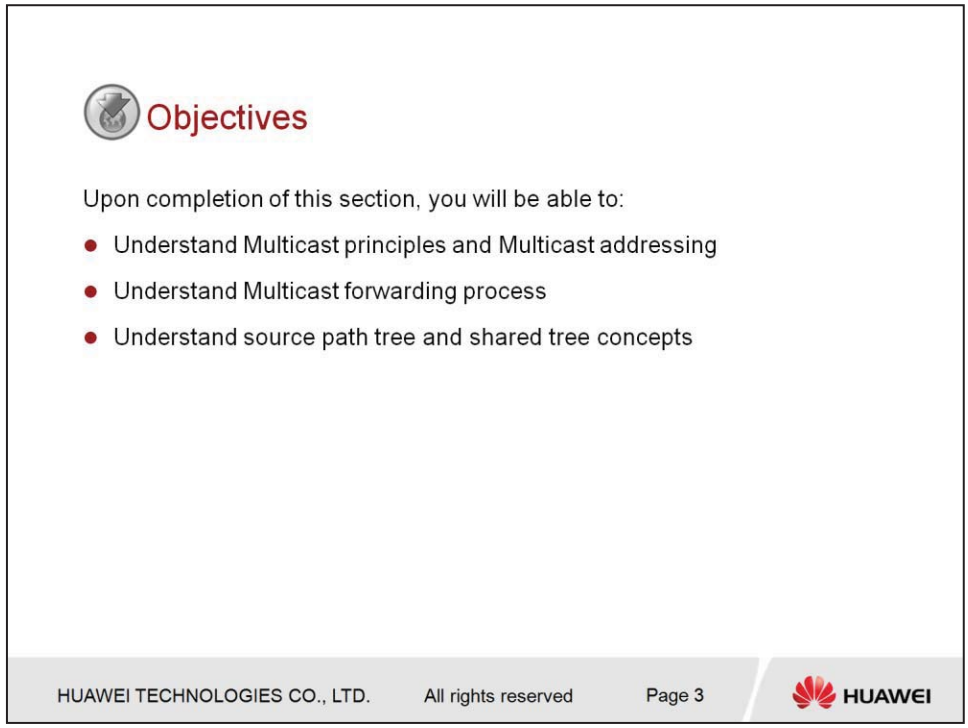
Q: What is the function of policy-based route?

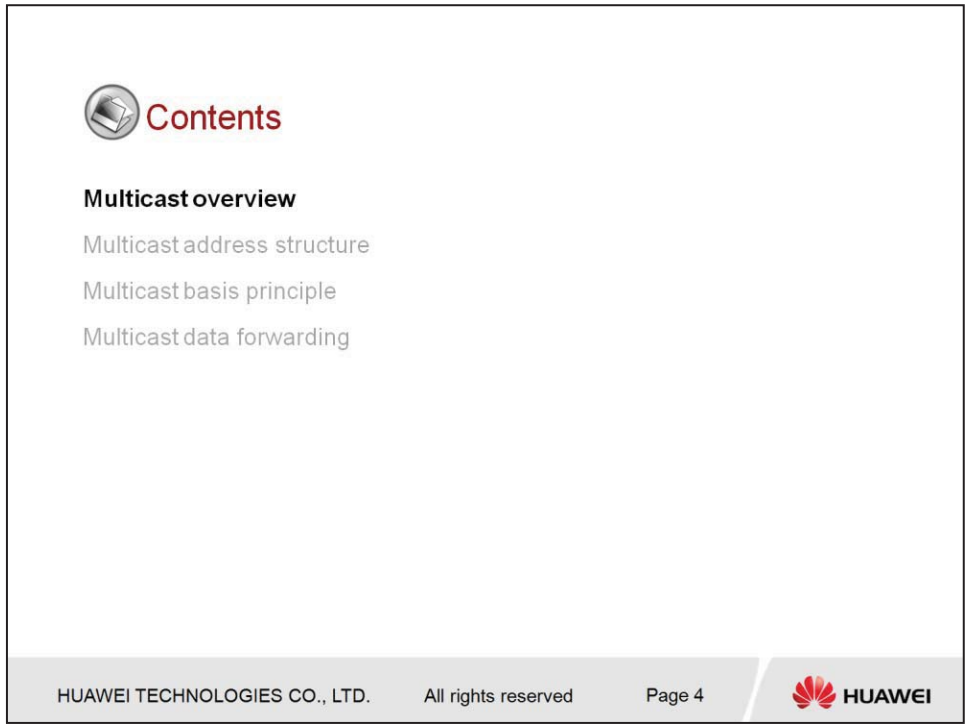
A: Compared with traditional routing protocol, it can provide more powerful control of forwarding and storing packets for network management. The network manager can choose the forwarding path according to not only destination address but also the protocol type, packet size, application, source IP address or other parameters. It can also control load balance among multiple routers, QoS for packet forwarding on single link and so on.

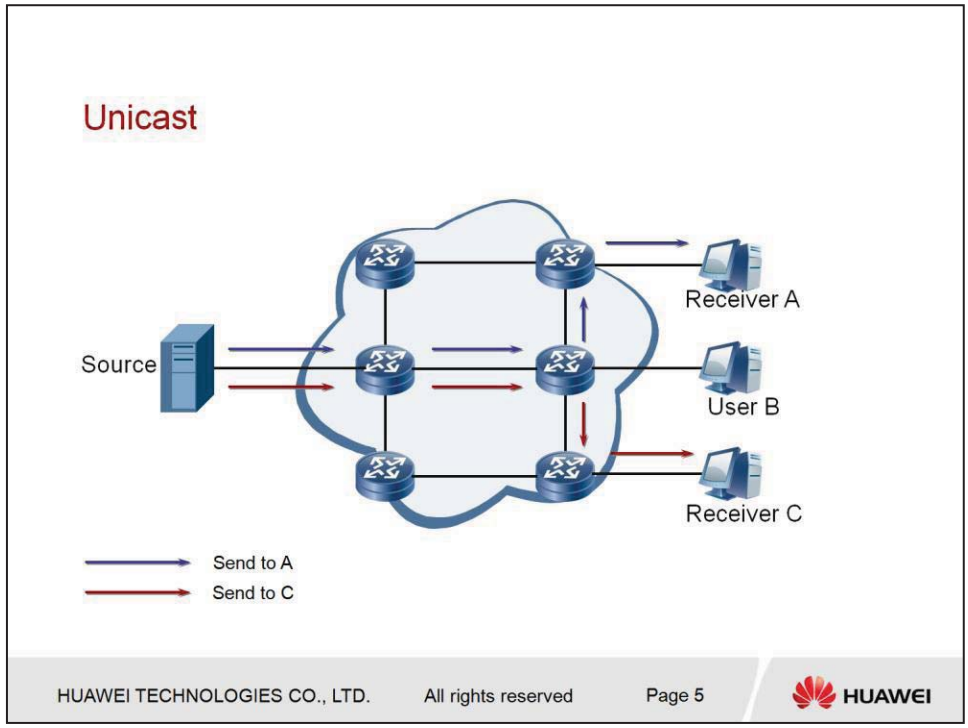
Module 5
Multicast


IP Multicast Basics
www.huawei.com
HUAWEI TECHNOLOGIES CO., LTD.
All rights reserved
HUAWEI


Foreword
IP multicast is a technique designed to enable the routing of traffic from a single source to multiple select destinations. As opposed to unicast data transmission, it enables an efficient mode of data transfer and effectively saves network bandwidth through reduction of network load, and is widely used in support of real-time data transfer, multimedia conferencing, data duplication, online games & simulations, and many other applications.
HUAWEI TECHNOLOGIES CO., LTD.
All rights reserved
Page 2
HUAWEI


Objectives
Upon completion of this section, you will be able to:
Understand Multicast principles and Multicast addressing
Understand Multicast forwarding process
Understand source path tree and shared tree concepts
HUAWEI TECHNOLOGIES CO., LTD.
All rights reserved
Page 3
HUAWEI


Contents
Multicast overview
Multicast address structure
Multicast basis principle
Multicast data forwarding
HUAWEI TECHNOLOGIES CO., LTD.
All rights reserved
Page 4
HUAWEI



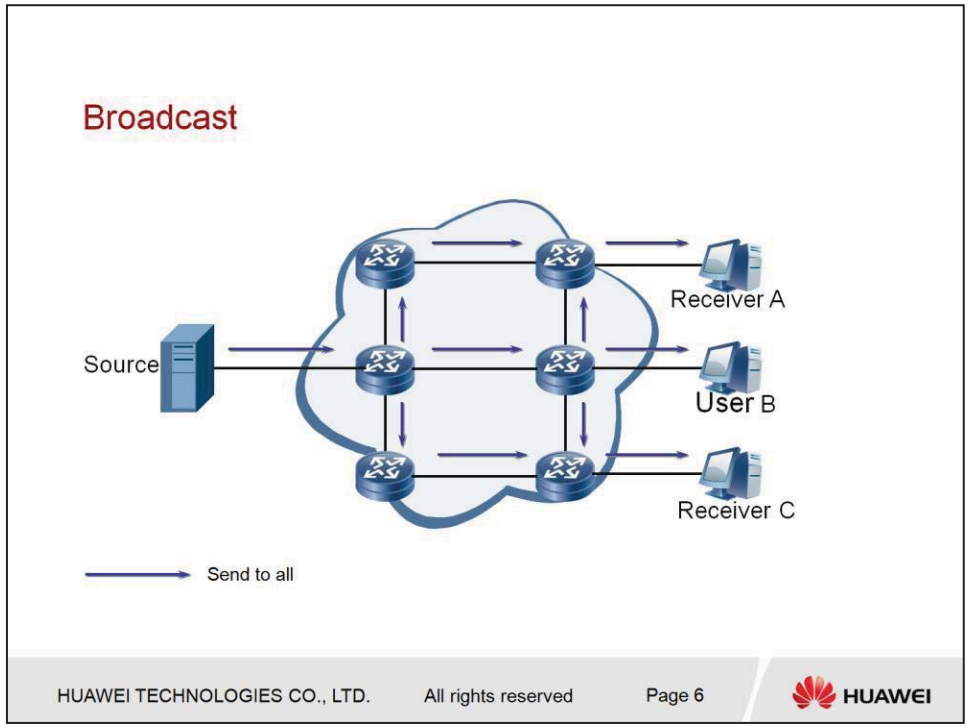
When transmitting with the unicast mode, the system creates a data transmission channel and sends a copy of the information for each subscriber who requested for the information. Thus, when a lot of users request for the same information

at the same time, there is load of duplicate flows in the network. While unicast is a simple mechanism for one-to-one communication, for one-to-many communication

it brings the network to its knees due to its huge bandwidth demands.

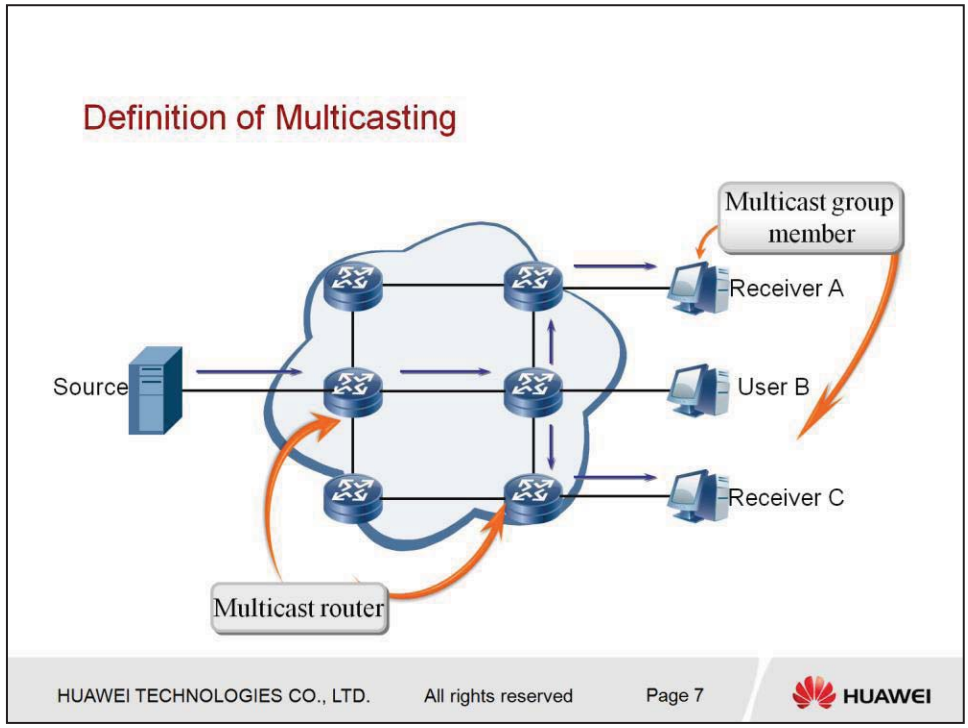
When transmitting with the unicast mode one receiver receives one copy from the sender. The bottleneck effect on bandwidth is a big problem as the number of transmissions made will increase as the number of users increases. The unicast

mode is better applied to networks with fewer users instead of environments where messages are sent on a larger scale.



In broadcasting, the system sends the information to all the users in a network whether they need it or not, which undermines information security and impairs charging efficiency. Moreover, if only a few receivers request the data, the network resources are used and the bandwidth is wasted.

So, the broadcast mode is suitable for networks with a high density of users. If the number of users who request for the same information is not clear, neither the unicast nor the broadcast mode is a wise choice.



The information sender or the multicast source sends only one copy of the information to the network and creates tree-shaped routes by the multicast routing protocol to let the network intelligently replicate the packet only where it needs to,

which helps conserve bandwidth and network resources both on the sending and the receiving end of a transmission.

Multicast is based on the group. Receivers in one group use the same IP multicast address as the destination address. As shown in the diagram above, Receiver A and Receiver C joined the multicast group and then they can receive data.

Compared with the unicast mode, multicast does not dramatically increase the burden of the network when the number of users hikes. Whatever how many receivers are there in the network, there is only one data stream with the same information. And unlike the broadcast scheme which requires the source to send

an individual copy to each receiver, multicast conserves network resources by sending packets only to receivers who are interested in the information.

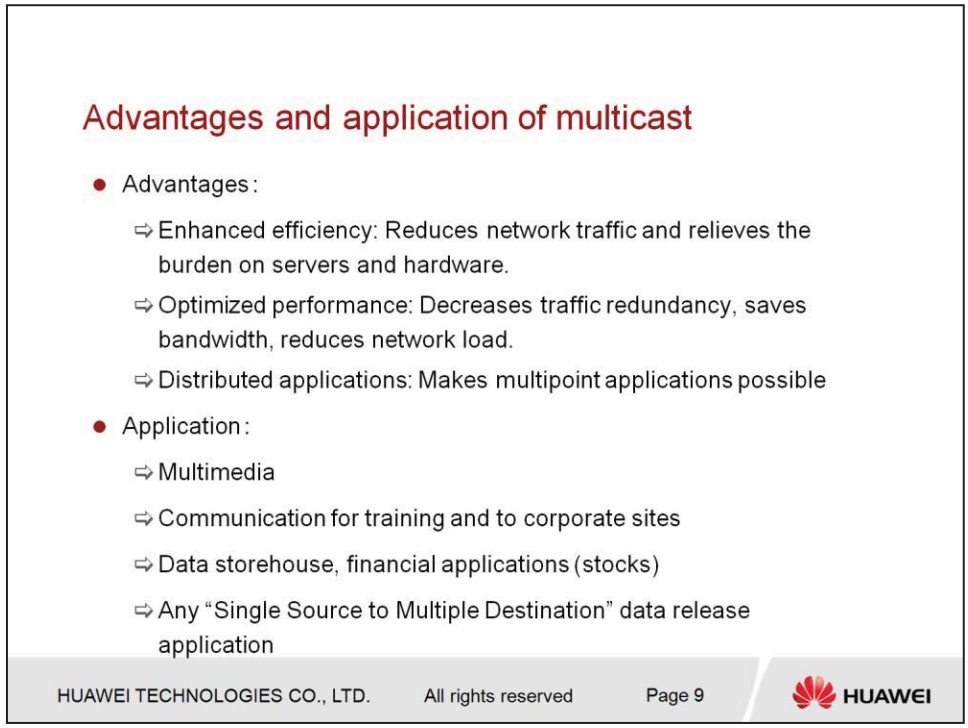
In the multicast mode:

The information sender is called the “multicast source”.

Receivers that get the same information from a multicast group and each of the receivers are a “multicast group member”.

Routers that provide the multicasting service are “multicast routers”.

Multicast routers do not only provide multicasting services, but also carry out management on group members. Moreover, a multicast router itself can be a member of one or more multicast groups. And those groups do not have any physical or geographical boundaries—the hosts can be located anywhere on the Internet.



The advantages of multicast lie in:

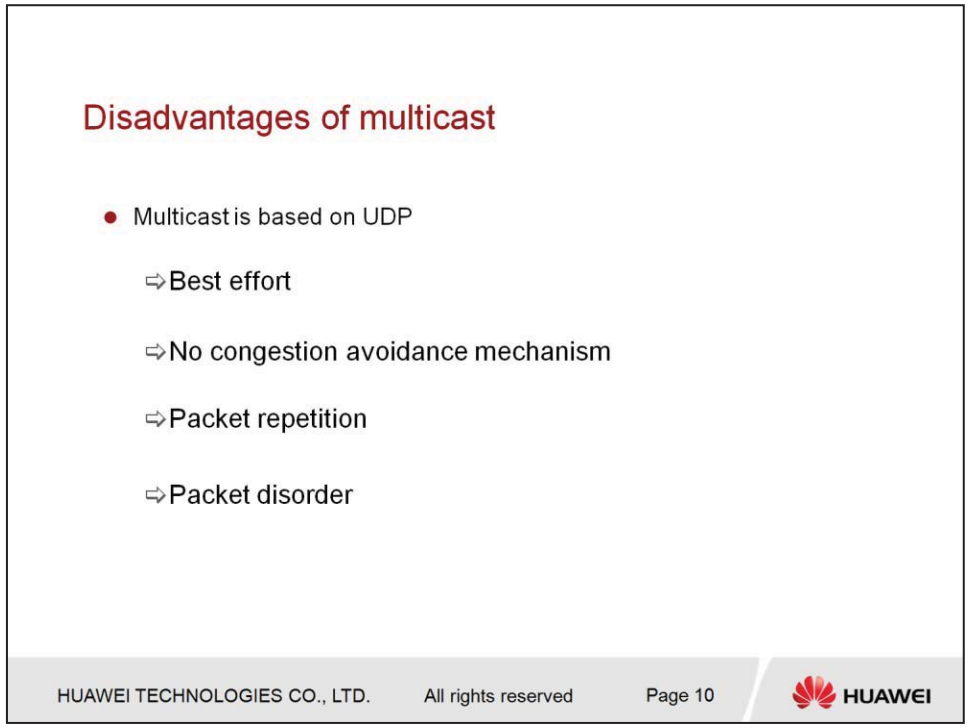
•Enhanced efficiency: it reduces network traffic and relieves the burden for servers and the hardware.

•Optimized performance: it reduces traffic redundancy, saves network bandwidth and lightens the network load.

•Distributed applications: it makes multipoint applications possible.

The multicast technology provides a satisfactory solution for transmitting information from a single point to multiple points which makes an efficient one-to-many transmission in IP networks a reality.

The multicast technology brings us versatile value-added services from on-line live broadcasting, Web TV, distance education, tele-healthcare to on-line radio broadcasting, and real-time audio/video conferencing.



The multicast technology provides a satisfactory solution to one-to-many transmission which makes an efficient point-to-multipoint communication in IP networks a reality. While the plus points of the multicast technology is obvious there are also disadvantages since it is based on the UDP.

•Best effort

Since packet loss is inevitable, the multicasting application program should be designed to overcome packet loss instead of depending on the multicast network for reliability. “Reliable multicasting” is still under research.

•No congestion avoidance mechanism

•For lack of mechanisms like the TCP window mechanism and the slow start mechanism, multicast could cause traffic congestion on the network. If possible, multicasting application program is expected to avoid congestion by certain checking attempts.

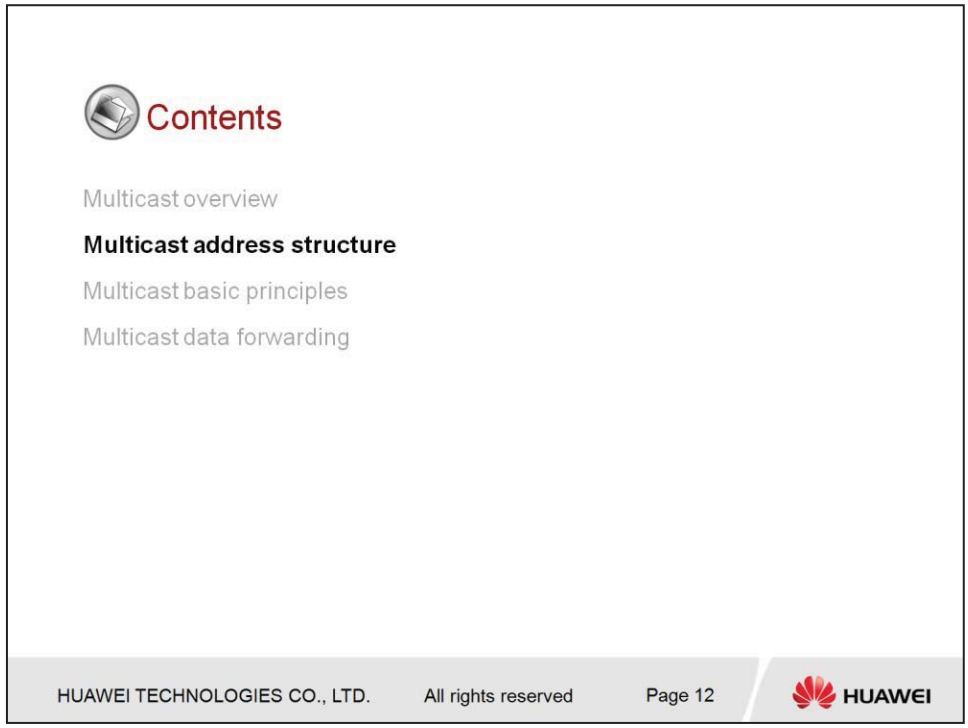
•Packet repetition

Some mechanisms of certain multicasting protocols such as the Assert mechanism and the SPT switch mechanism may cause packet repetition occasionally and the multicasting application

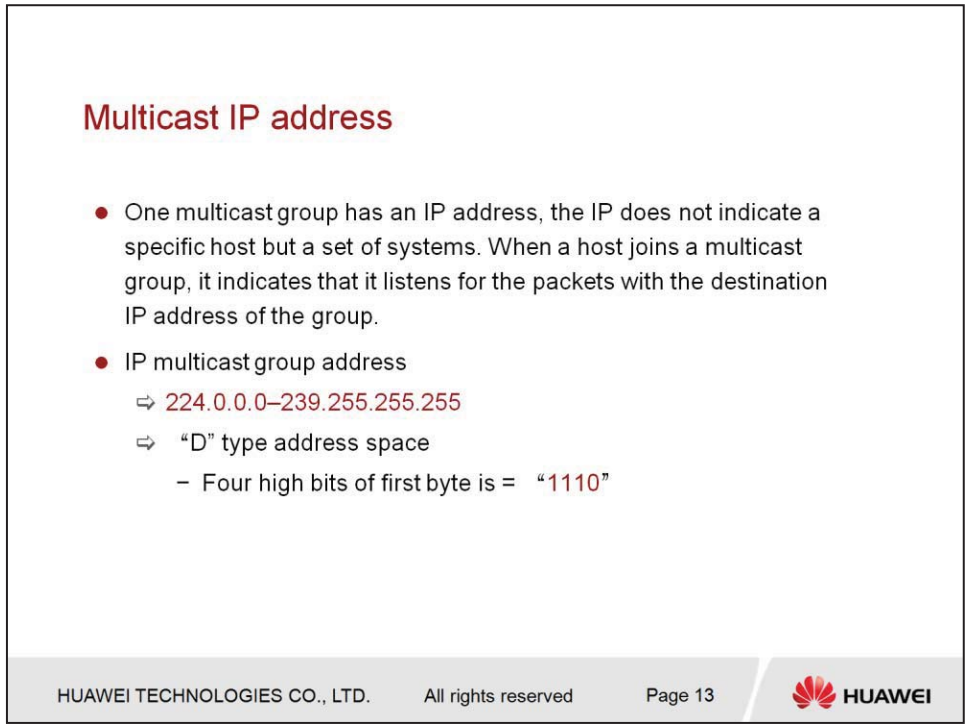
program should tolerate that .

•Packet disorder

Multicasting protocols may lead to packet disorder when packets reach their destination and the multicasting application program should take methods such as the buffer pool mechanism to rectify it.



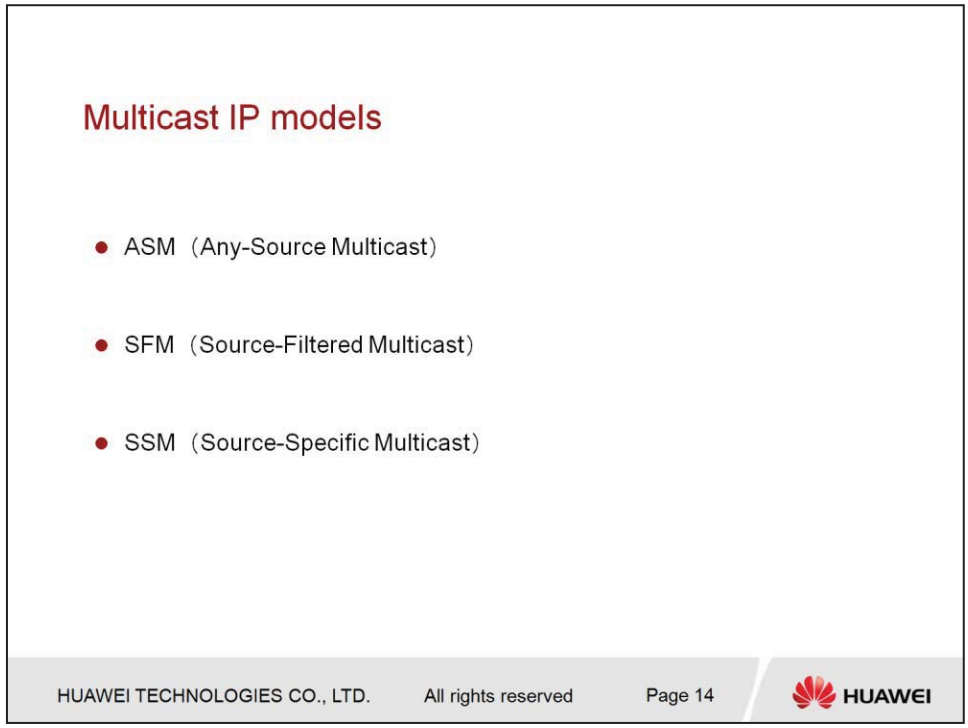
This chapter presents multicast address structure, address classification and the differences between multicast MAC address and unicast MAC address.



In unicast, a datagram travels from the source address to the destination address as the hop-by-hop theory defines. While in multicast, the destination address of a datagram is a group of addresses instead of a unique one. All the receivers that request the data are put into one group and as soon as they join the group, data destined for the address of the group is sent to them

Every member in the group, as it is called “the multicast group”, can receive the datagram.

The Internet Assigned Numbers Authority (IANA) controls the assignment of IP Multicast Addresses. IANA has assigned the Class D address space to be used for IP Multicast. This means that all IP multicast-group addresses fall in this range: 224.0.0.0 - 239.255.255.255. The Class D address cannot be put into the source IP address field of an IP packet.



•ASM

In the ASM model, any information sender can be a multicast source to send information to a multicast address. Hosts can receive all the information sent to the multicast group after they join the group. In this model, receivers have no idea about where the multicast source is before they get information from the source and they can join or withdraw from the group at any time.

•SFM

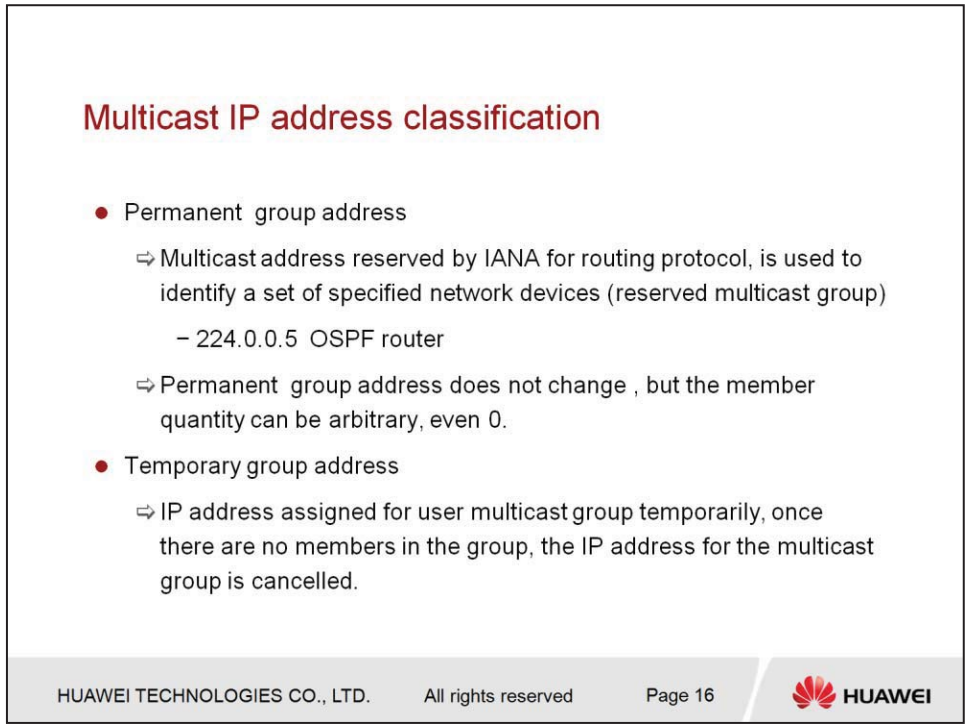
The SFM model is similar to the ASM model as in both the two modes, the members of a multicast group are totally equal from the perspective of the information sender. However, the SFM model is an extension of the ASM model as it can filter multicast sources. In the SFM model, the upper layer software checks the source address of multicast packets and decides whether to let the packets from a specific multicast source to pass. Thus, receivers can only get data from some of the sources. To receivers, multicast sources are filtered and only some of them are valid.

•SSM

In daily life, users are interested in information sent by particular

sources and are reluctant to get information from other sources. The SSM model provides a service that enables users to choose sources at the client.

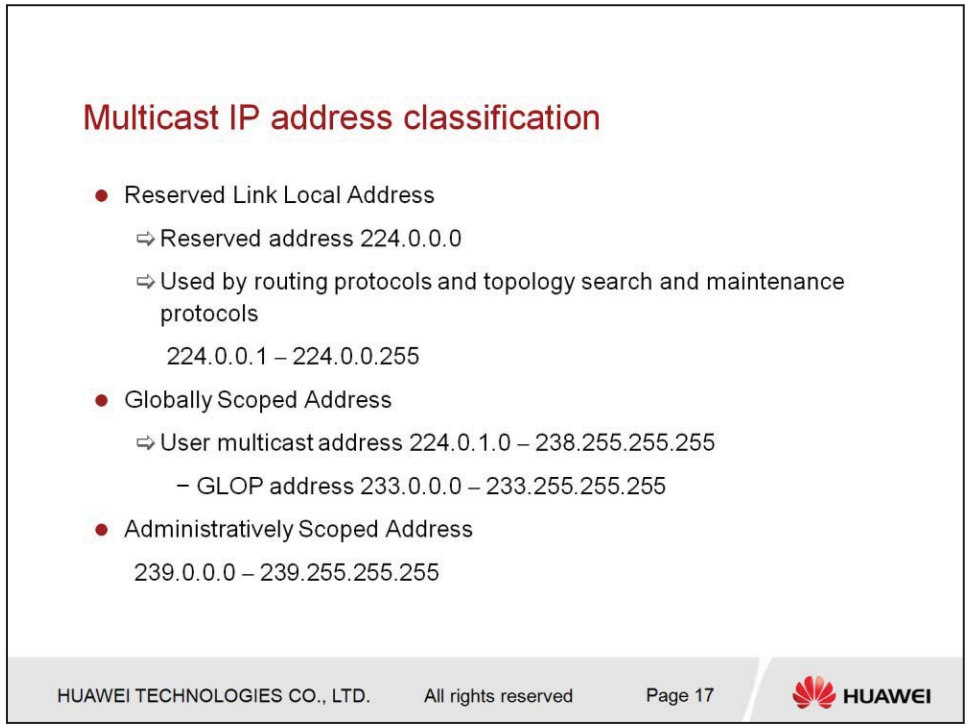
The substantial difference between the SSM model and the ASM model is that in the SSM model, receivers get to know the position of the multicast source beforehand. The SSM model uses multicast addresses in a different range from the ASM model and sets up special forwarding paths between receivers and specific resources.



Permanent group address: multicast address reserved by IANA for routing protocols. The permanent group address is used to identify a set of specified network devices (reserved multicast group). The permanent group address never change, however, the number of its members can be any, even 0. 224.0.0.5

is the multicast address reserved for OSPF routers.

Temporary group address: an IP address allocated to a multicast group for temporary use. A temporary group address is discarded when the number of members in the group turns to 0.



•Reserved Link Local Address

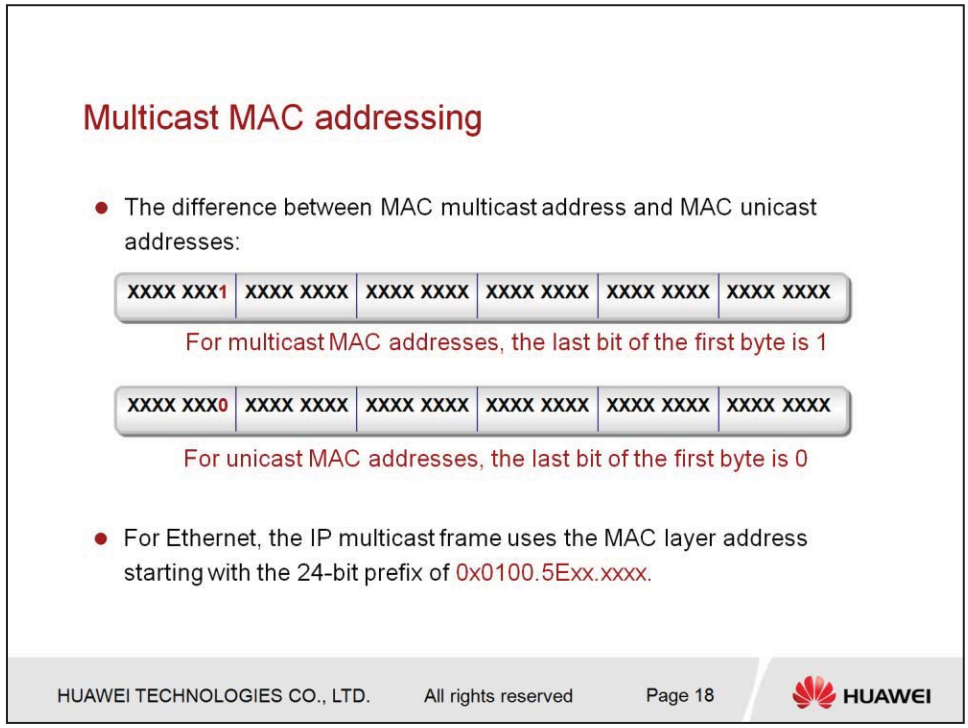
The IANA reserves addresses in the range of 224.0.0.0 and 224.0.0.255 to be used by routing protocols, topology search and maintenance protocols. Packets with these addresses should never be forwarded by a router whatever their time-to- live (TTL) values are. The address 224.0.0.0 is reserved and will not be allocated.

•Globally Scoped Address

The range of addresses from 224.0.1.0 to 238.255.255.255 is called Globally Scoped Address and they are multicast addresses that can be used globally.

233/8 is the address range for GLOP which is an address allocation mechanism works for AS'. An AS defines its range by using it's AS number derived from the bits of the middle two octets of the multicast address, the fourth octet is used by the AS to represent it's obtained 255 multicast addresses.

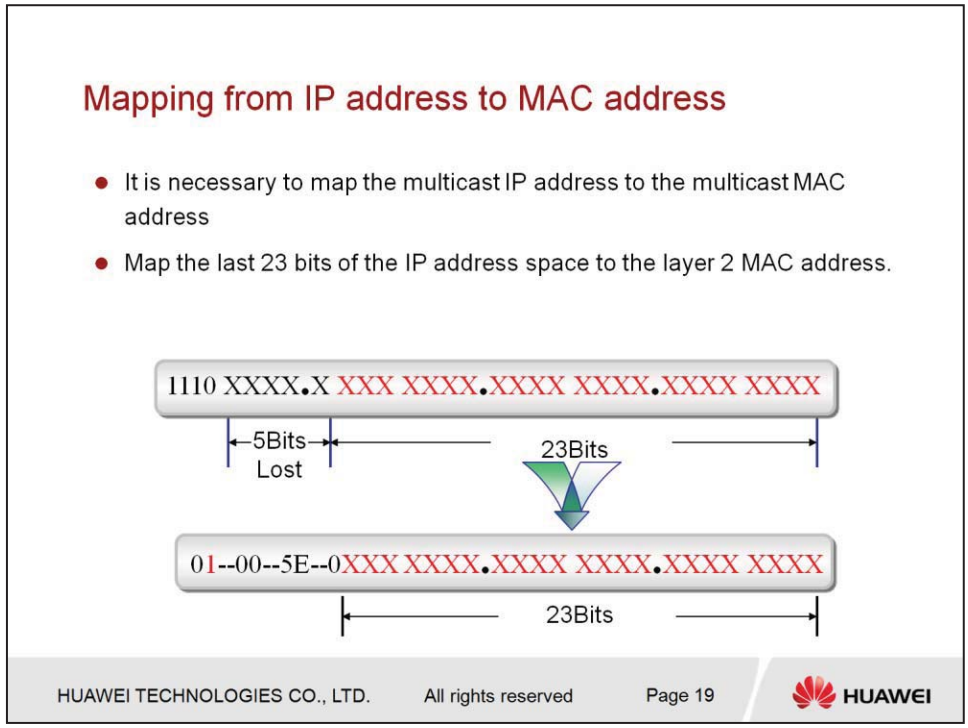
•Administratively Scoped Address The range of addresses from 239.0.0.0 to 239.255.255.255 is reserved for enterprises for internal use. These addresses are similar to reserved addresses of common IP address but can only be used within enterprises.



The MAC address of the destination is the MAC address of the receiver when packets are transmitted in the unicast mode on an Ethernet. However, when packets are transmitted in the multicast mode, the destination turns to a group of hosts instead of one receiver. In this case, we use the MAC multicast address.

The MAC multicast address is used to label receivers that belong to the same multicast group at the data-link layer.

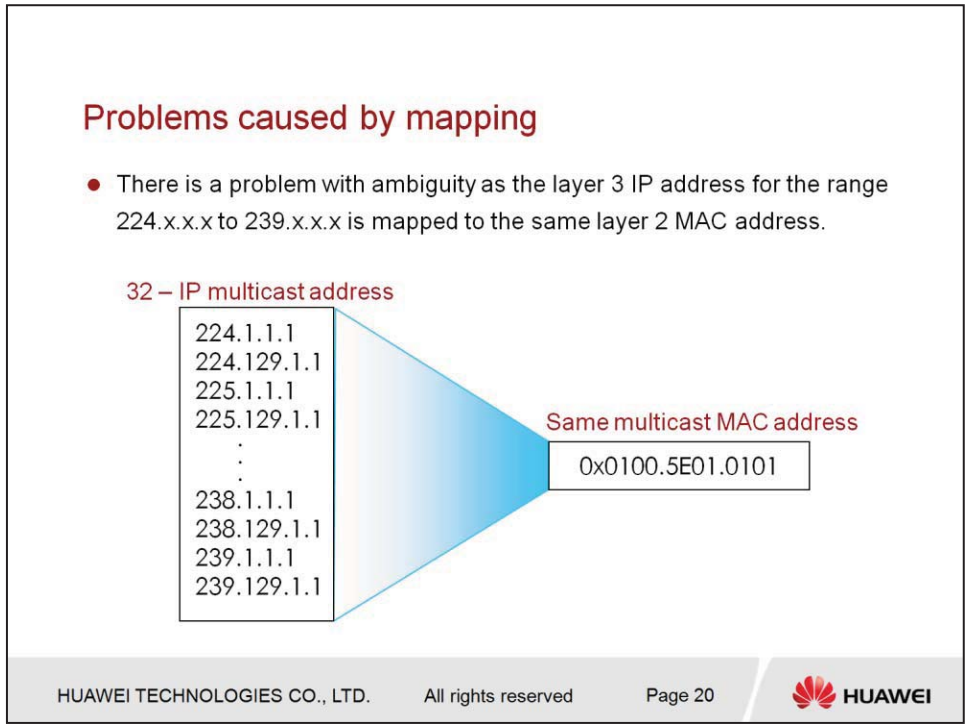
As defined by IANA, the first 24 bits of the MAC multicast address is 0x01005e and the lower 23 bits are the lower 23 bits of the multicast IP address.



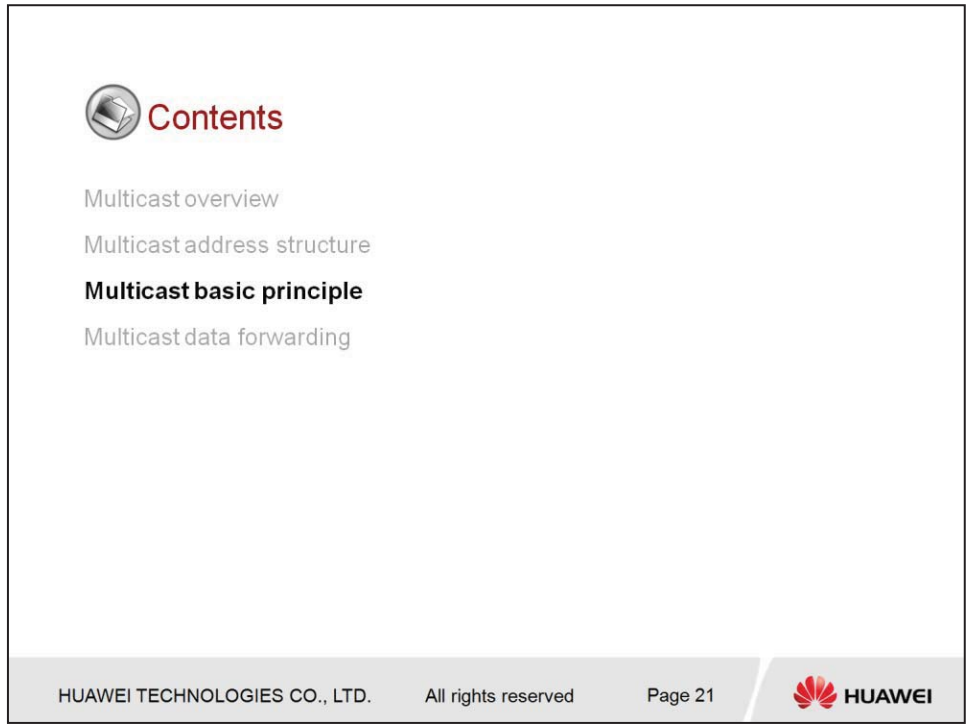
The mapping of an IP address to its MAC address is illustrated by the above figure.

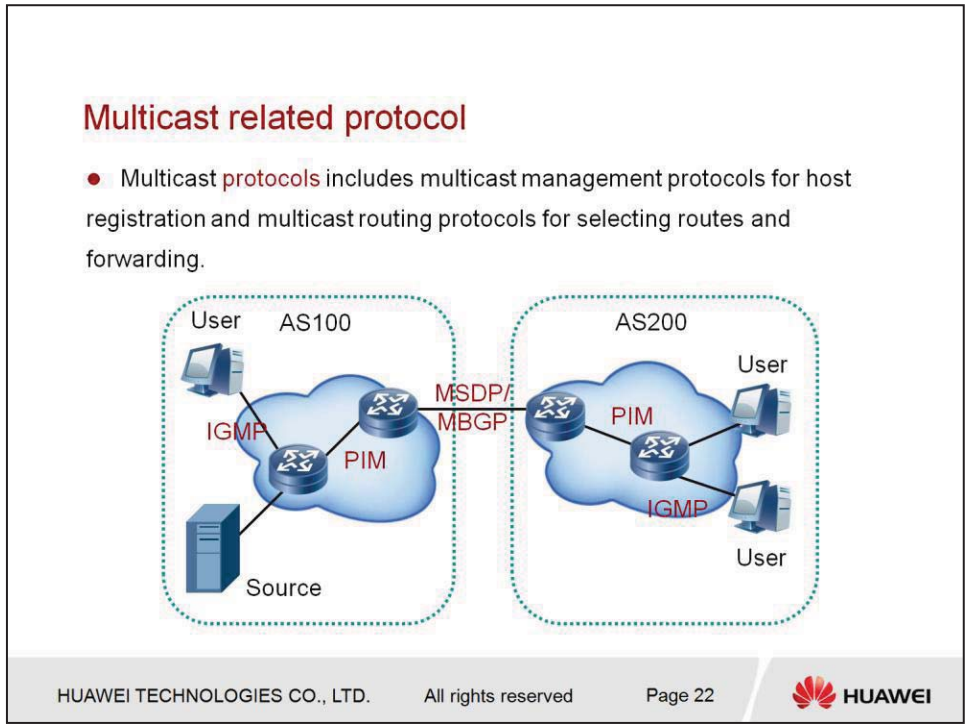
The first 24 bits of a MAC multicast address is 0x01005e.

The first four bits of an IP multicast address are 1110 which indicate that the address is a multicast address and in the last 28 bits, the final 23 bits are mapped to the MAC address.



The first four bits of an IP multicast address is 1110 which indicates the address is a multicast address. And 23 bits of the last 28 bits of the IP multicast address are mapped to the MAC address which means five bits are dropped during the process. The direct outcome is 32 IP multicast addresses are all mapped to the same MAC address.


Contents
Multicast overview
Multicast address structure
Multicast basic principle
Multicast data forwarding
HUAWEI TECHNOLOGIES CO., LTD.
All rights reserved
Page 21
HUAWEI



Multicast protocols consist of multicast management protocols for host registration and multicast routing protocols for route selection.

The applications of every Multiple multicast protocol are marked in the above figure.

Internet Group Management Protocol (IGMP) works between hosts and multicast routers. This protocol provides a mechanism for establishing and maintaining the relations between hosts and multicast routers.

Multicast routing protocols are applied to multicast routers. These protocols create and maintain routes for multicasting and make packet forwarding more reliable and efficient.

For the ASM model, multicast routing can be classified into the inter-domain routing and the intra-domain routing.

Intra-domain multicast routing protocols are used to search multicast sources and build multicast trees to send information to receivers within autonomous systems.

Intra-domain routing protocols include DVRMP, MOSPF and PIM.

DVRMP stands for Distance Vector Multicast Routing Protocol

and it is a protocol of dense mode. This protocol defines the maximum hops to be 32.

MOSPF is an extension of OSPF and it supports multicasting by defining new LSAs.

Protocol Independent Multicast (PIM) is a typical intra-domain multicast routing protocol and it comes in two modes, namely, the Dense Mode (DM) and the SM (Sparse Mode). DM is suitable when receivers are densely dotted in a network;

while SM is fit for networks where receivers scatter sparsely. PIM must work together with unicast routing protocols.

Inter-domain multicast routing protocols are used to transfer information between ASs.

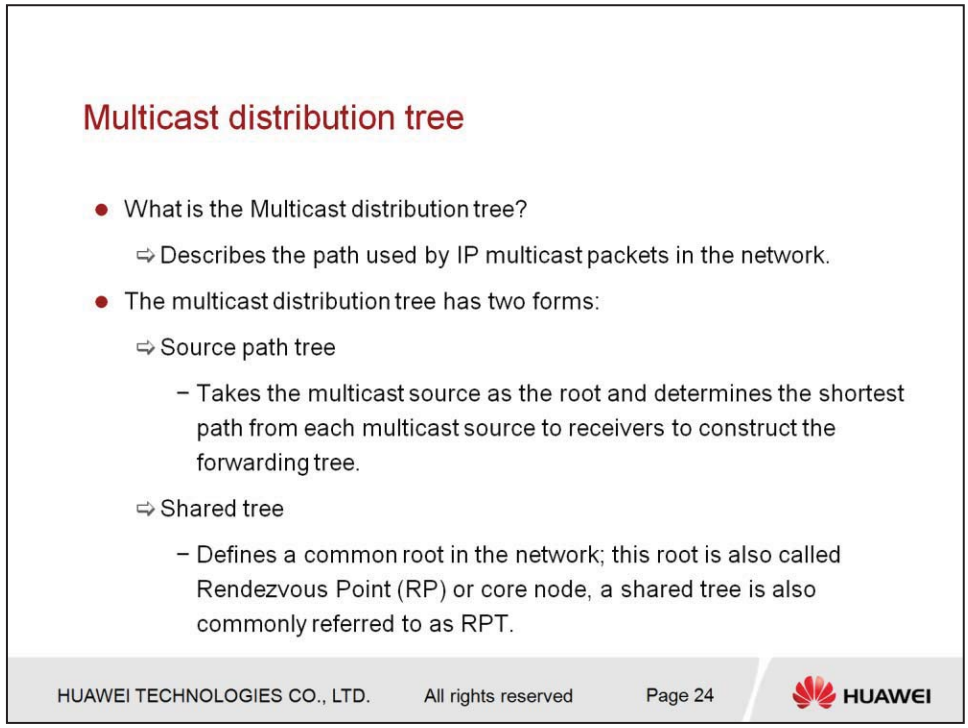
Multicast Source Discovery Protocol (MSDP) is able to spread news about multicast sources across ASs.

MultiProtocol Border Gateway Protocol (MPBGP) is an extension of MBGP and it can advertise multicast routes across ASs.

For the SSM model, there is no concept of the intra-domain and the inter-domain.

As the receivers have the knowledge of where the multicast sources are beforehand, the router can create paths for multicasting directly with the function of PIM-SM.

In this section, we focus on the intra-domain multicast routing protocols.



The one-way routes between a source and multiple receivers for multicasting form the multicast distribution tree.

The two basic types of multicast distribution trees are source path trees and shared trees.

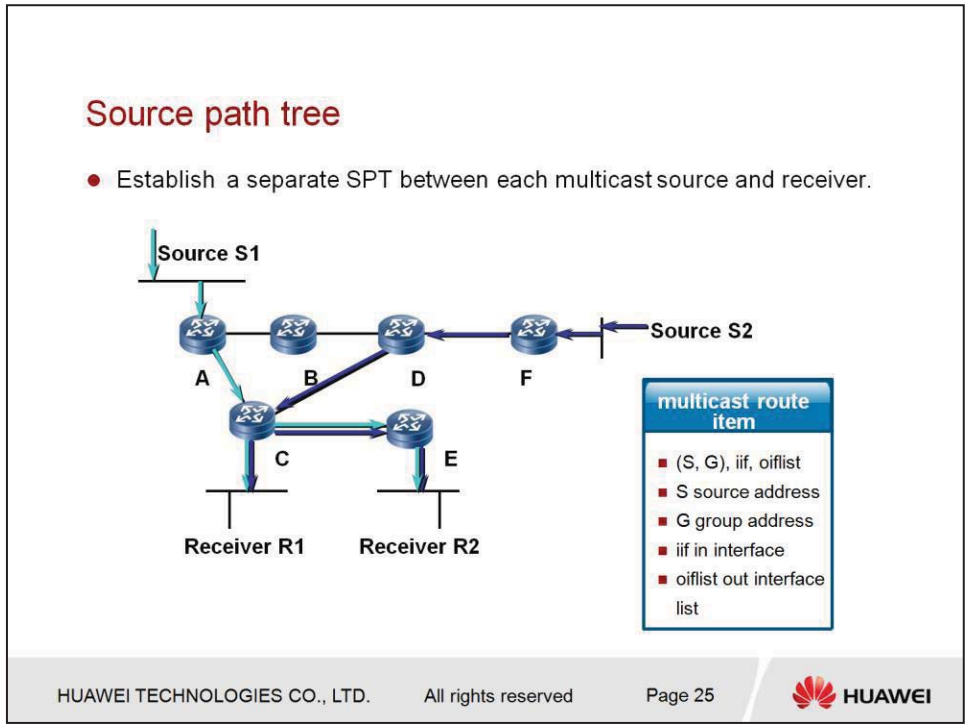
•Source Path Tree

Source path tree is a multicast distribution tree with its root at the source and branches forming a spanning tree through the network to the receivers. Because this tree uses the shortest path through the network, it is also referred to as a shortest path tree (SPT). The network is responsible for building a tree for each source that sends multicast datagram.

•Shared Tree

Unlike source trees that have their root at the source, shared trees use a router as the root. This shared root is called a Rendezvous Point (RP) and all the shortest paths between the RP and its receivers form a distribution tree.

When using a shared tree, there is only one tree in the network. All the sources in the network use the tree to send their traffic to the root and then the traffic is forwarded down the shared tree to reach all receivers.

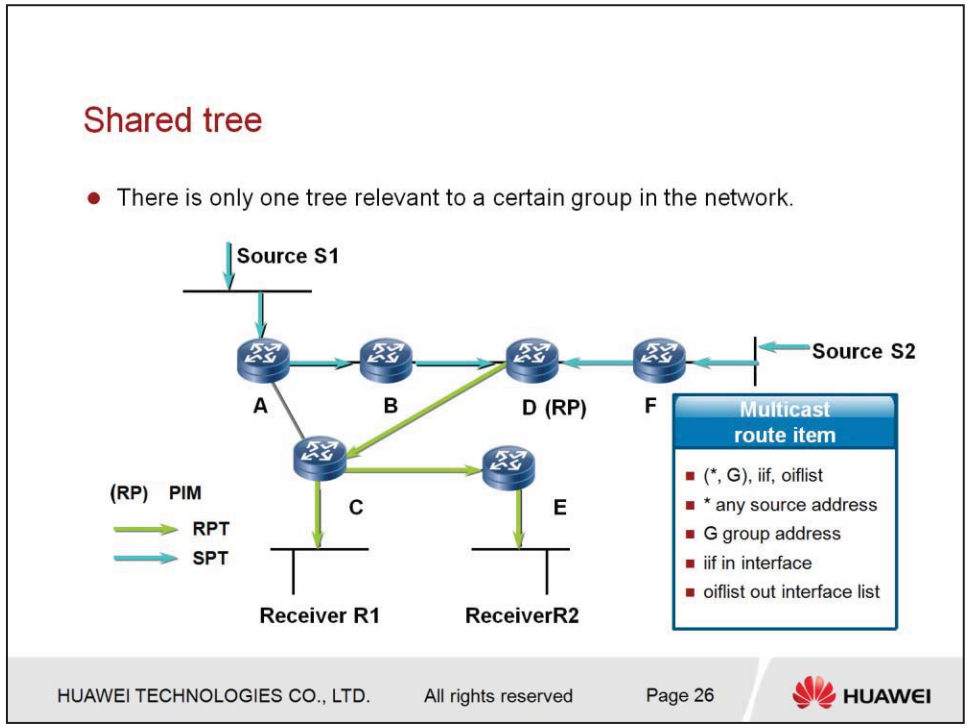


Source path tree is a multicast distribution tree with its root at the source and branches forming a spanning tree through the network to the receivers. Because this tree uses the shortest path through the network, it is also referred to as a shortest path tree (SPT). The network is responsible for building a tree for each source that sends multicast datagram.

The figure above shows an example of two STPs in a network since there are two multicast sources, namely, Source 1 and Source 2. R1 and R2 are the two receivers that get information from the two sources. The two STPs are:

S1—A---C (R1) -----E (R2)

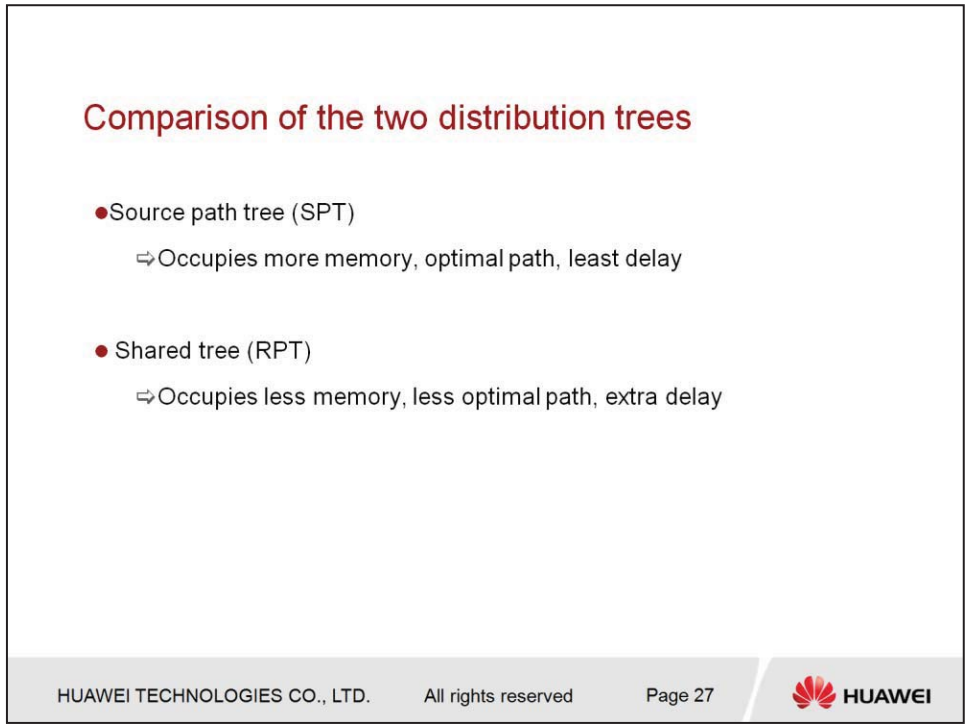
S2---F----D---C (R1)------E (R2)



Unlike source trees that have their root at the source, shared trees use a router as the root. This shared root is called a Rendezvous Point (RP) and all the shortest paths between the RP and its receivers form a distribution tree.

When using a shared tree, there is only one tree in the network. All the sources in the network use the tree to send their traffic to the root and then the traffic is forwarded down the shared tree to reach all receivers.

In the example above, two sources, S1 and S2, share the tree D (RP) ----C (RT) -- --E (R2).

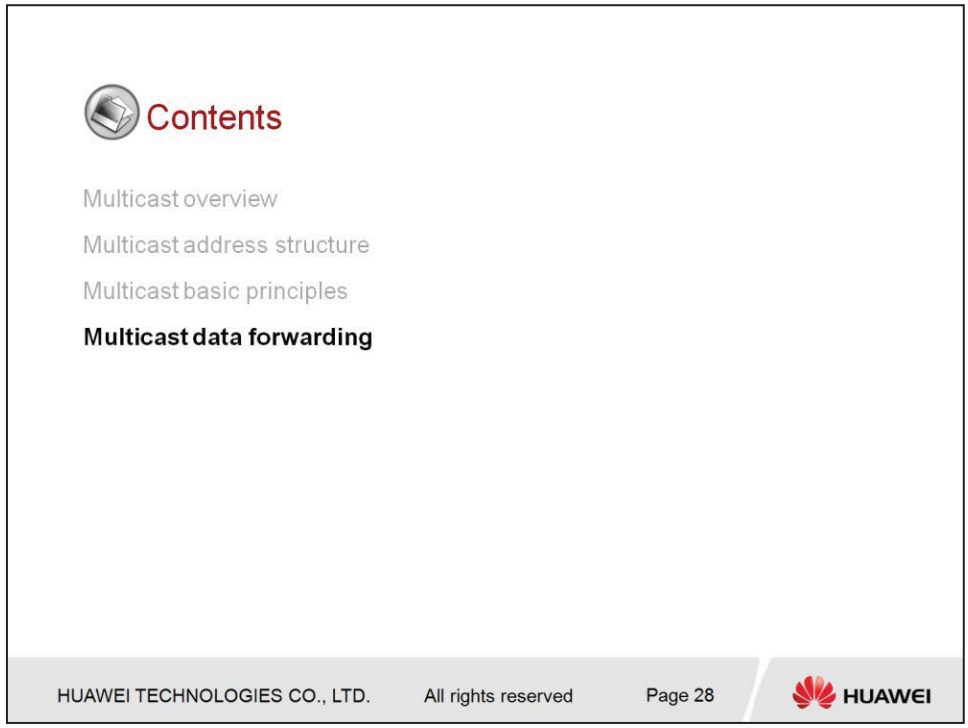


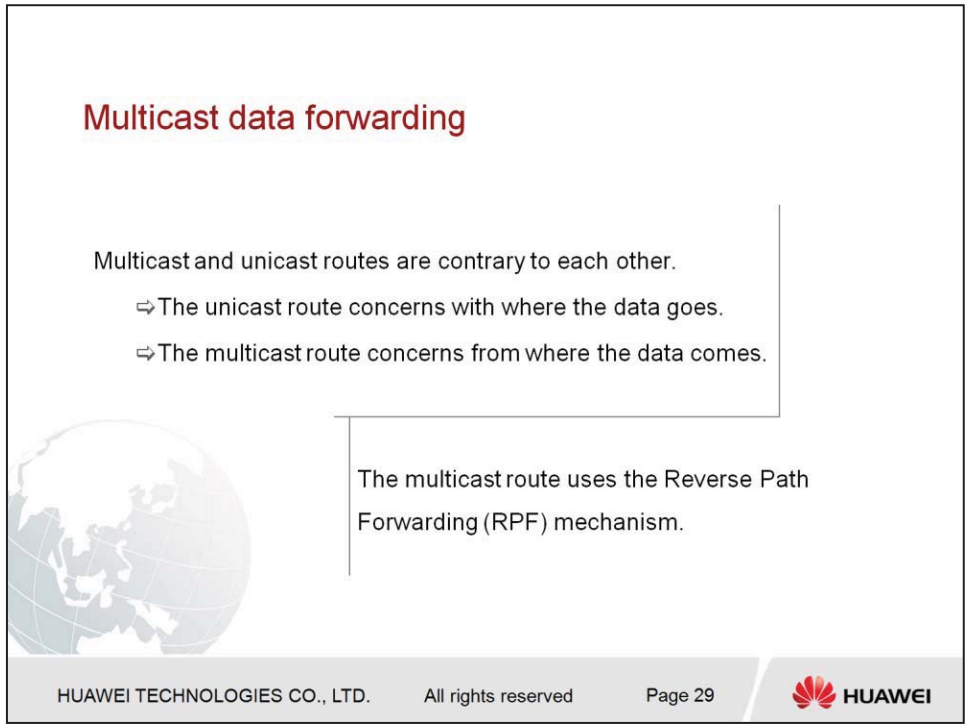
Shortest Path Trees have the advantage of creating the optimal path between the source and the receivers. This guarantees the minimum amount of network latency for forwarding multicast traffic. This optimization does come at a price. The

routers must maintain path information for each source. In a network that has thousands of sources and thousands of groups this can quickly become a serious problem. Memory consumption from the size of the multicast routing table is a

factor that network designers must take into consideration.

Shared Trees have the advantage of requiring the minimum amount of state in each router. The disadvantage of shared trees is that under certain circumstances the paths between the source and receivers might not be the optimal paths as the packets are sent to the RP first and then forwarded to the receivers. And it is also a very tough test for the reliability and capability of the RP.

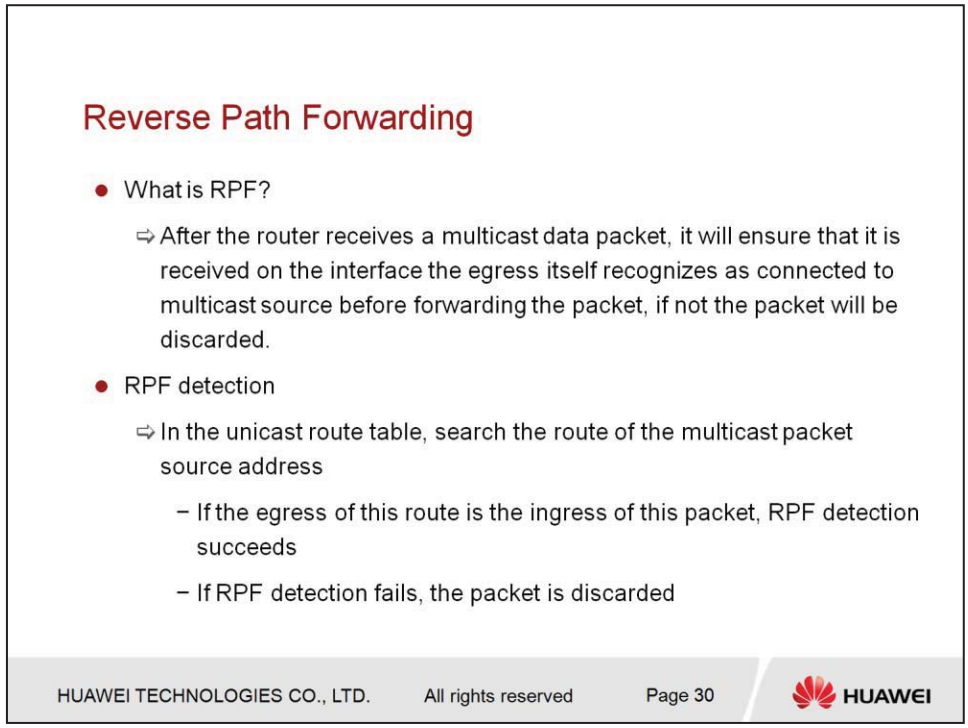

Contents
Multicast overview
Multicast address structure
Multicast basic principles
Multicast data forwarding
HUAWEI TECHNOLOGIES CO., LTD.
All rights reserved
Page 28
HUAWEI



In unicast routing, traffic is routed through the network along a single path from the source to the destination host. A unicast router does not actually care about the source address. It is the destination address and how to forward the traffic towards that destination that a unicast router really concerns about.

In multicast routing, the source sends traffic to an arbitrary group of receivers that are represented by a multicast group address. The multicast router must determine which direction is upstream (towards the source) and which direction (or directions)

is downstream and then forwards packets down the downstream. The concept of forwarding multicast traffic away from the source, rather than to the receiver, is called Reverse Path Forwarding (RPF).



RPF makes use of the existing unicast routing table to determine the upstream and downstream neighbors. A router only forwards a multicast packet when it is received on the upstream interface (RPF interface).

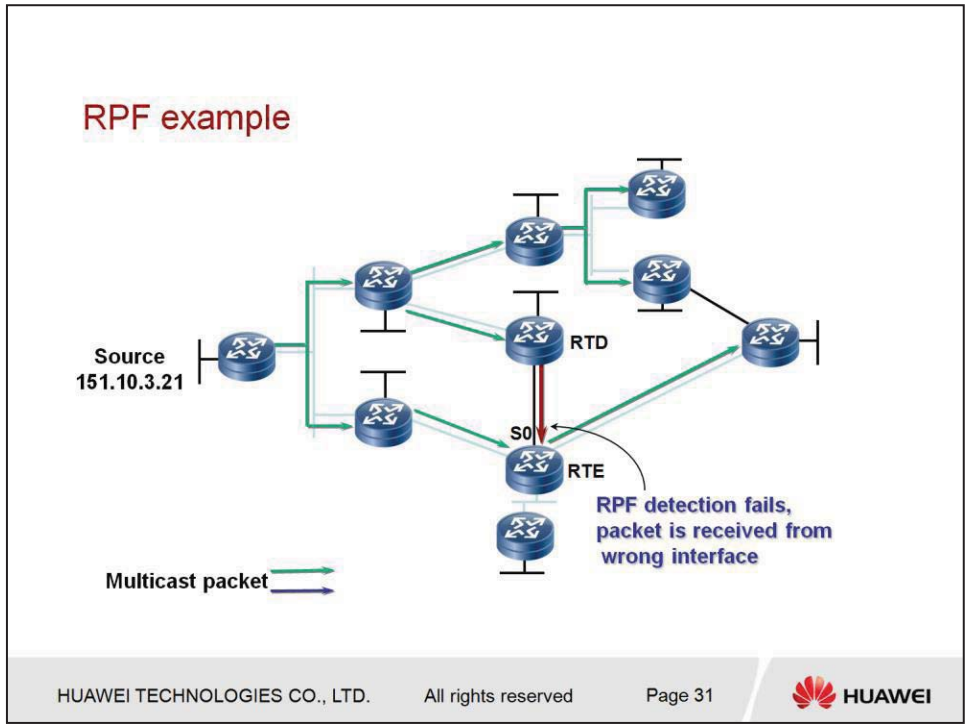
RPF not only enables routers to correctly forward multicast traffic down the distribution tree, but also helps to guarantee that the distribution tree is loop free which is very important for multicasting.

RPF detection is the essence of the RPF mechanism. When a multicast packet arrives at a router, the router performs the RPF detection on the packet. The packet is forwarded only after it passes the check. Otherwise it is dropped. For traffic flowing down a source tree, the RPF detection procedure works as follows:

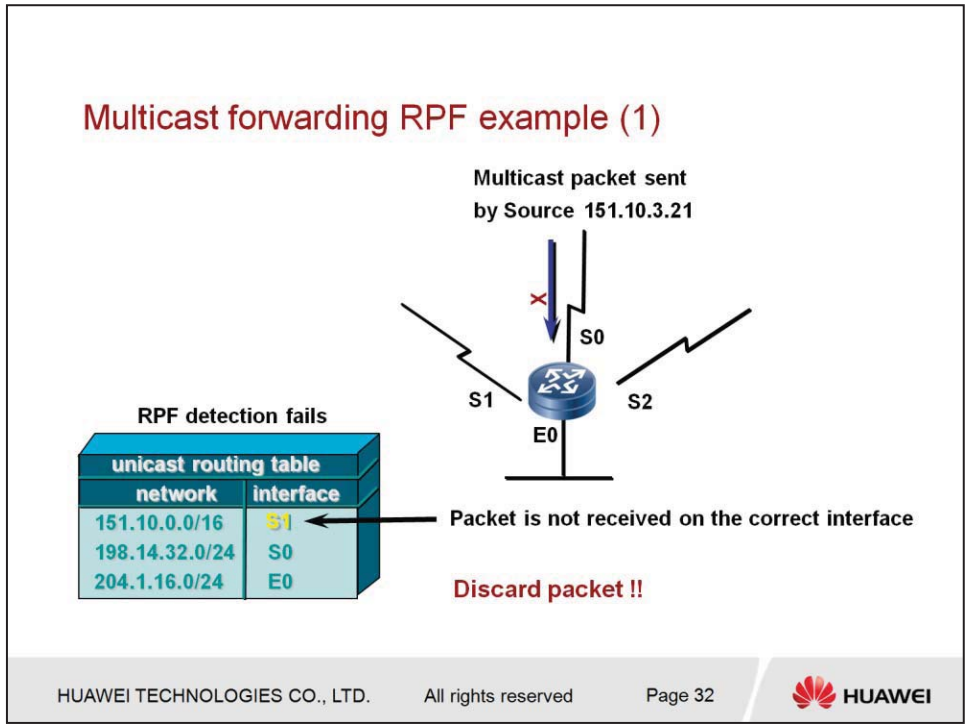
Step 1. The router looks up the RPF interface of the multicast source in a SPT environment or the RPF interface of the RP when a shared tree is used. The RPF interface is the egress where the router sends out the packet.

Step 2. If a packet is received from the RPF interface, then it passes RPF detection and is forwarded.

Step 3. If the RPF detection in Step 2 fails, the packet is dropped.



In the figure above, Router RTE performs RPF detection after it received a multicast packet from S0. The RPF detection discovers that the packet is received from a wrong interface, so Router RTE discards the packet.

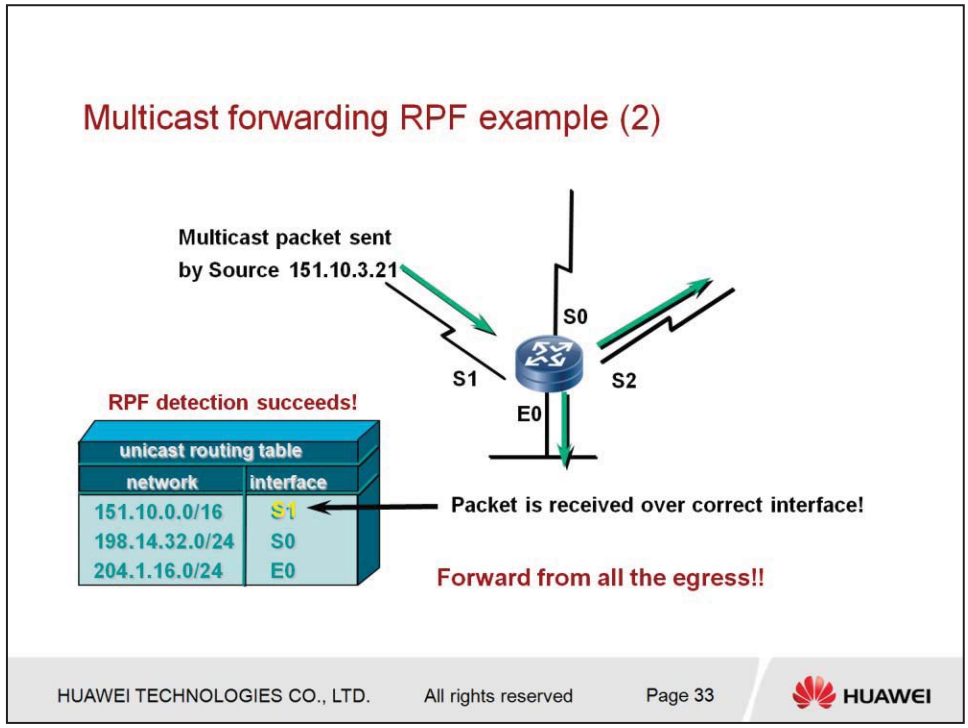


The process of a RPF detection is actually a process of consulting the unicast routing table.

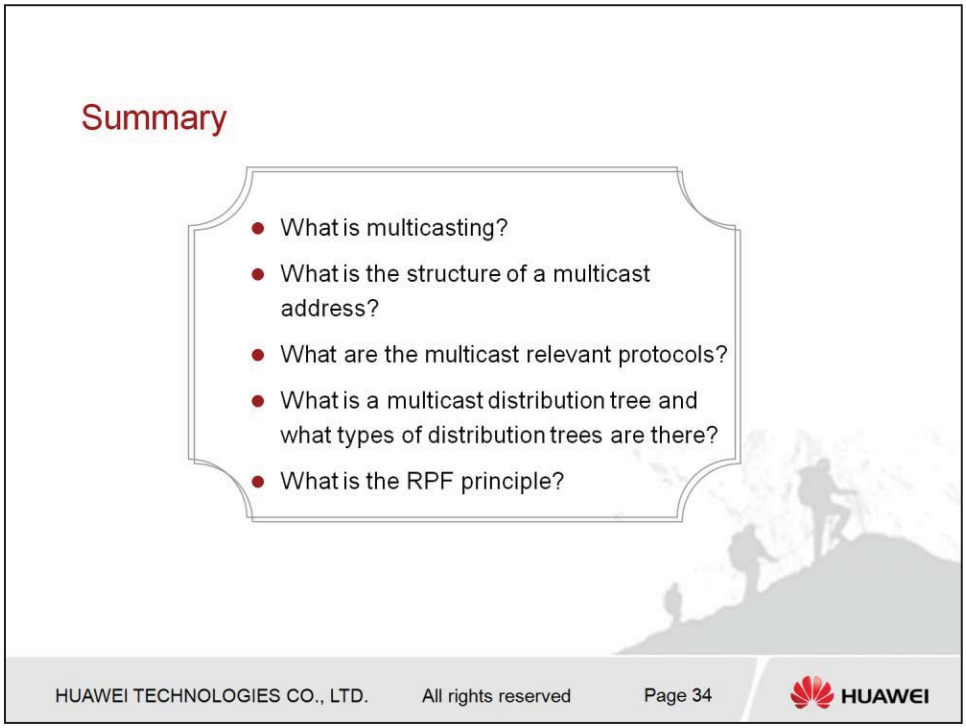
Routers look up the unicast routing table to see if the egress to the source is the interface that receives packets after they get multicast packets. If the egress to the source is not the interface that receives packets, then RFP detection fails and packets is dropped.

In the example above, the S0 interface of the router enables RPF detection after it receives a multicast packet. After consulting the unicast routing table, it finds the egress to the source 151.10.0.0/16 is S1 which is not the interface that received

the packet. So it decides to discard the packet since the packet is not received from the correct interface.



In this example the network remains the same as the previous example. The router performs RPF detection after it receives a multicast packet. After performing a lookup of the unicast routing table, the router finds the egress towards the source is S1 which is also the ingress of the packet, so the packet passes the detection and is forwarded along the distribution tree.



1. What is multicast?

A multicast group is identified by an IP multicast address. Receivers can receive any data sent to the group after they joins the multicast group. The information sender or the multicast source sends only one copy of the information to the

network and creates tree-shaped routes by multicast routing protocols to let the network intelligently replicate the packet only where it needs to, which helps conserve bandwidth and network resources.

2. What is the structure of a multicast address?

The Class D multicast addresses fall in the range from 224.0.0.0 to 239.255.255.255. The addresses can be classified into permanent group address and temporary group address according to their longevity. Permanent group addresses are multicast addresses reserved by IANA for routing protocols. They

are used to identify a specific group of network equipments and cannot be changed. The number of members in the group can be anything, even 0. The temporary group addresses are IP addresses allocated to multicast groups for temporary use. The address of a group is discarded as soon as there is no member

in the group. Upon their usage scope, multicast addresses can be divided into reserved link local address, globally scoped address and administratively scoped address.

3. What are the multicast relevant protocols?

Multicast protocols consist of IGMP that is for host registration and multicast routing protocols for route selection. Intra-domain multicast routing protocols include PIM-SM, PIM-DM and DVMRP. Inter-domain protocols include MSDP and

MBGP.

4. What is a multicast distribution tree and what types of distribution trees are there?

One-way paths between a source and its multiple receivers form a distribution tree.

There are two types of distribution trees: Source Path Tree and Shared Tree.

Source path tree is a multicast distribution tree with its root at the source and branches forming a spanning tree through the network to the receivers. Because this tree uses the shortest path through the network, it is also referred to as a shortest path tree (SPT). The shared tree takes a router as its root and all the

shortest paths between the RP and its receivers as branches.

5. What is the RPF principle?

RPF makes use of the existing unicast routing table to determine the upstream and downstream neighbors. A router only forwards a multicast packet when it is received on the upstream interface (RPF interface). The procedure of RPF detection is: routers look up the route to the source in the unicast routing table when they receive packets. If the egress of the route is the ingress of the packets, the packets pass the detection and are forwarded; otherwise, the packets are discarded.


IGMP Protocol Prinicples
www.huawei.com
HUAWEI TECHNOLOGIES CO., LTD.
All rights reserved
HUAWEI

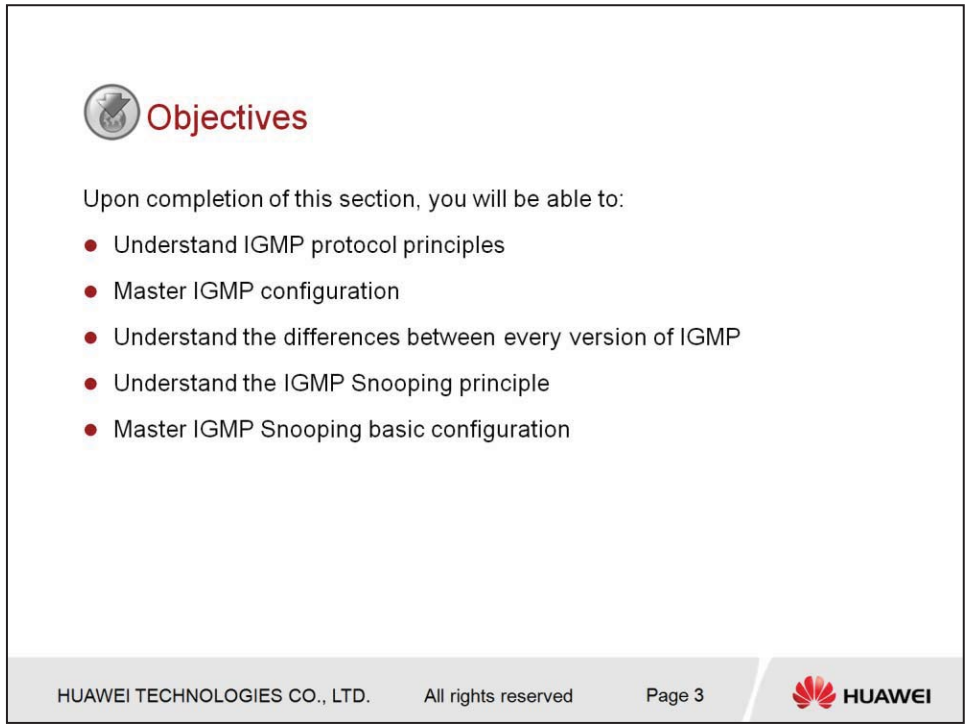


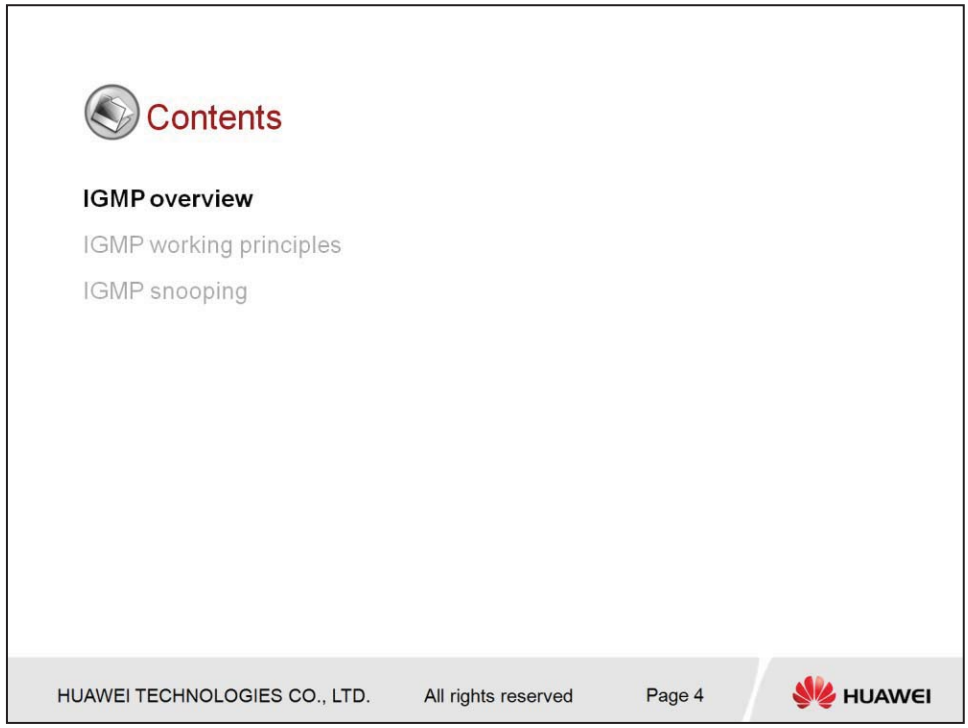
Foreword

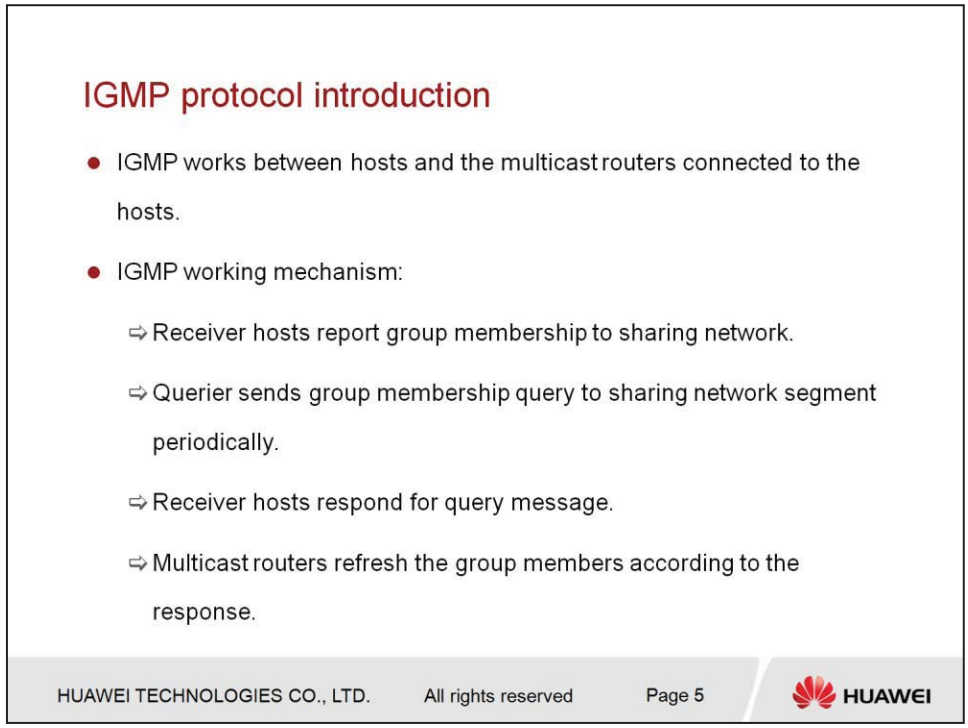
The Internet Group Management Protocol (IGMP) is used by IP hosts to report their multicast group memberships to any immediately-neighboring multicast routers. It is an integral part of IP and is required to be implemented by all hosts wishing to receive IP multicasts.

HUAWEI TECHNOLOGIES CO., LTD. All rights reserved




Objectives
Upon completion of this section, you will be able to:
Understand IGMP protocol principles
Master IGMP configuration
Understand the differences between every version of IGMP
Understand the IGMP Snooping principle
Master IGMP Snooping basic configuration
HUAWEI TECHNOLOGIES CO., LTD.
All rights reserved
Page 3
HUAWEI


Contents
IGMP overview
IGMP working principles
IGMP snooping
HUAWEI TECHNOLOGIES CO., LTD.
All rights reserved
Page 4
HUAWEI



IGMP is a signaling mechanism used by a host to communicate with a router on the end network for IP multicast. IGMP functions on the host side and the router side.

The working mechanism of IGMP is as follows:

1. The receiver host reports the group membership to the shared network.

2. Among all IGMP-enabled multicast routers in a network segment, a router is elected as the querier. The election mechanism varies with IGMP version. The

querier periodically sends the information about group members to the shared network segment.

3. After receiving the information sent by the querier, the receiver host sends a response to report the membership.

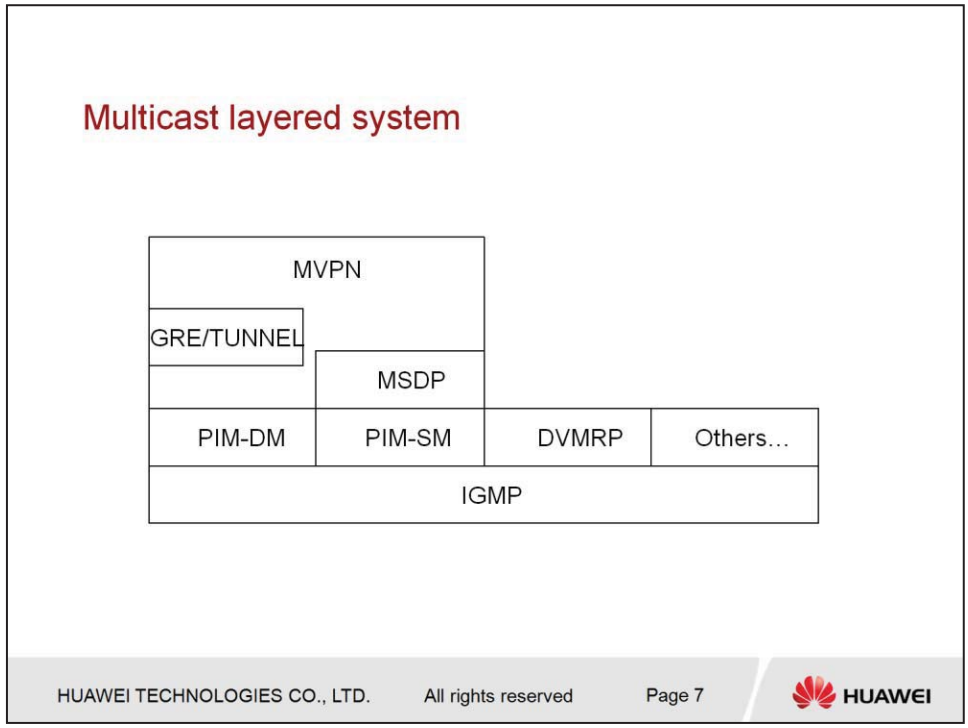
4. The multicast routers in this network segment refresh the member information according to the received response. If the routers do not receive any response within the timeout duration, the routers considers that no member of this group exists in this network segment. Thus, the routers stop forwarding the multicast data of this group.

All receiver hosts taking part in multicast transmission must apply

IGMP. A host can join or leave a multicast group at any time on any location, regardless of the number of members in the multicast group.

A multicast router cannot and also need not store the membership of all hosts.

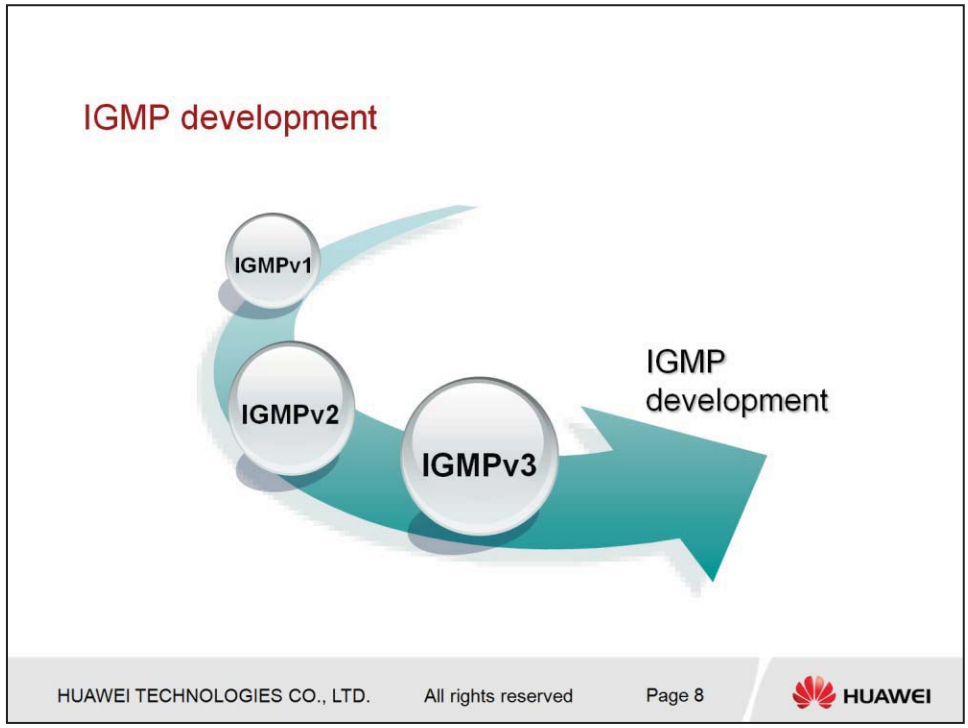
The router only needs to check whether the network segment connected to each interface has receivers of a multicast group, namely the members of a multicast group. The hosts need only to store the information about the multicast groups they have join.



From this figure, you can see that IGMP is on the lowest layer of the multicast protocol system. IGMP is the base of other multicast protocols. Among all multicast protocols, only IGMP is directly related to the host that demands a program. The

IGMP-enabled router manages joining and leaving of host in the multicast group.

The router maintains the user information and sends the multicast data to the host.



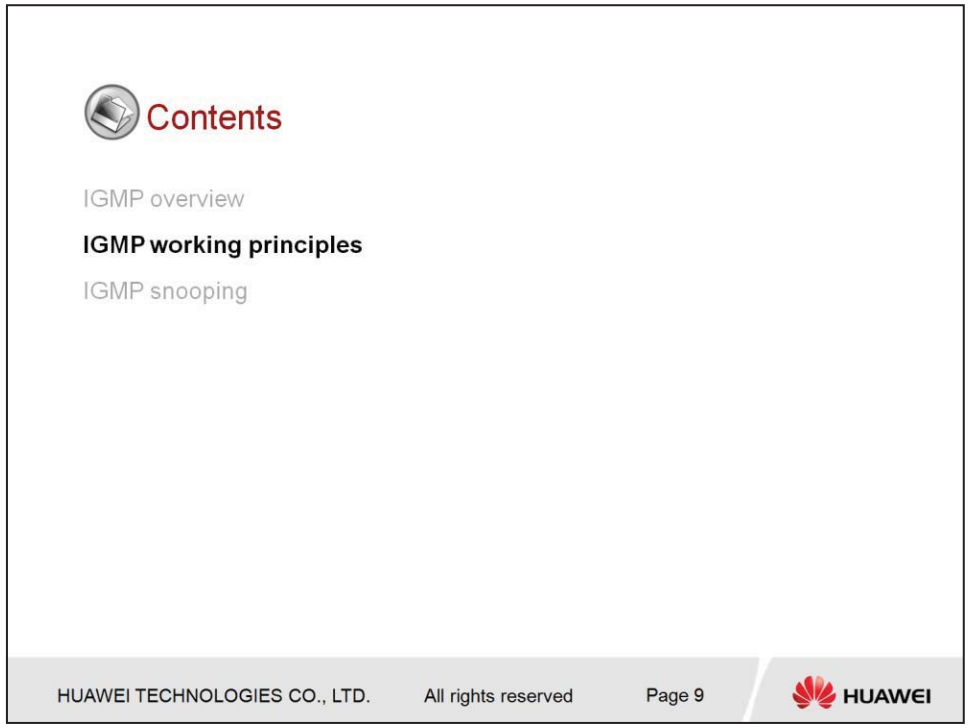
Currently, IGMP has three versions: IGMPv1, IGMPv2, and IGMPv3. All versions support the any-source multicast (ASM) model. IGMPv3 can be applied to the source-specific multicast (SSM) model independently. IGMPv1 and IGMPv2 need

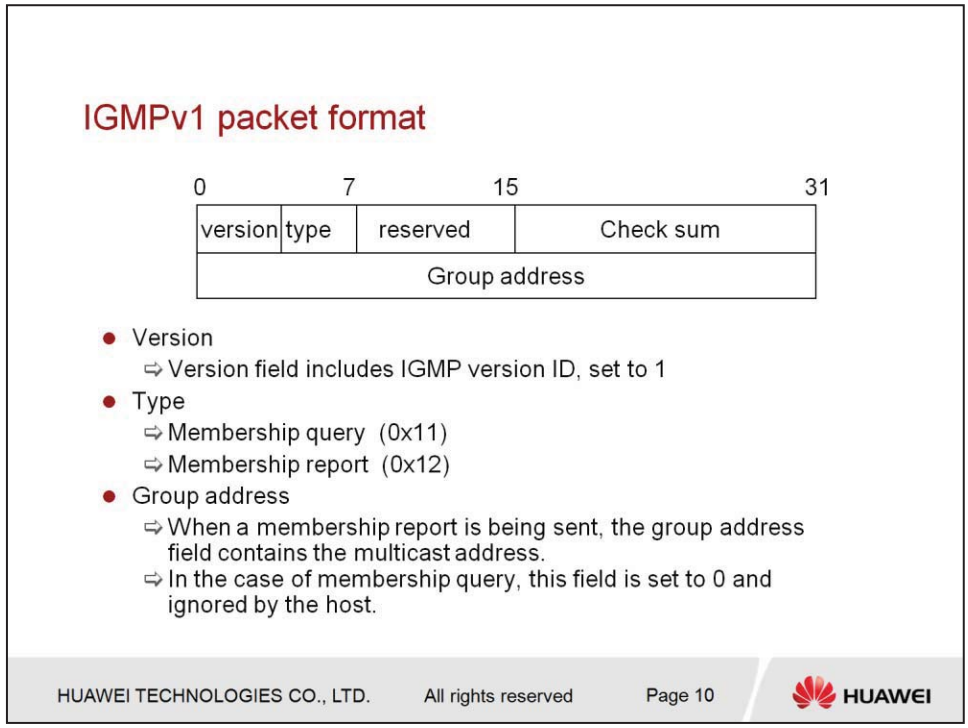
the support of SSM-mapping technology for the application in the SSM model.

IGMPv1 (RFC1112) defines the basic process of member query and report.

IGMPv2 (RFC2236) adds the mechanism for members to leave a group quickly.

IGMPv3 can specify the packets from certain multicast sources that the multicast members can receive or cannot receive.


Contents
IGMP overview
IGMP working principles
IGMP snooping
HUAWEI TECHNOLOGIES CO., LTD.
All rights reserved
Page 9
HUAWEI



The IGMPv1 packet format is shown in the figure. The packet contains the following fields:

Version: indicates the IGMP version. In IGMPv1, the value is 1. In IGMPv2, the packet does not contain this field.

Type: indicates the type of the IGMP packet.

IGMPv1 supports two types of packets:

Membership query (0x11): A router periodically sends member query messages to check whether members of a multicast group exist. The default query interval is 60 seconds.

Membership report (0x12): A host sends this packet to join a multicast group.

The membership report message can be sent passively or actively.

If a host wants to join a multicast group after it receives the membership query message, it sends the membership report message.

In this case, the membership report message is sent passively.

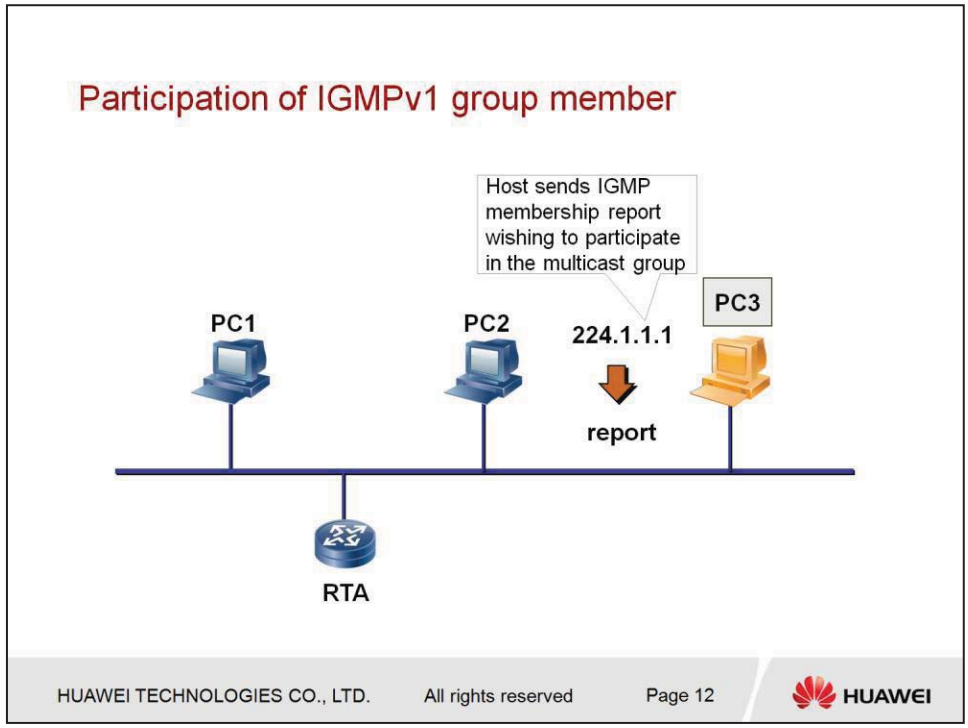
If a host want to join a multicast group, it can actively sends the membership report message without waiting for the member

Query message.

Group address: varies with the type of IGMP packets.

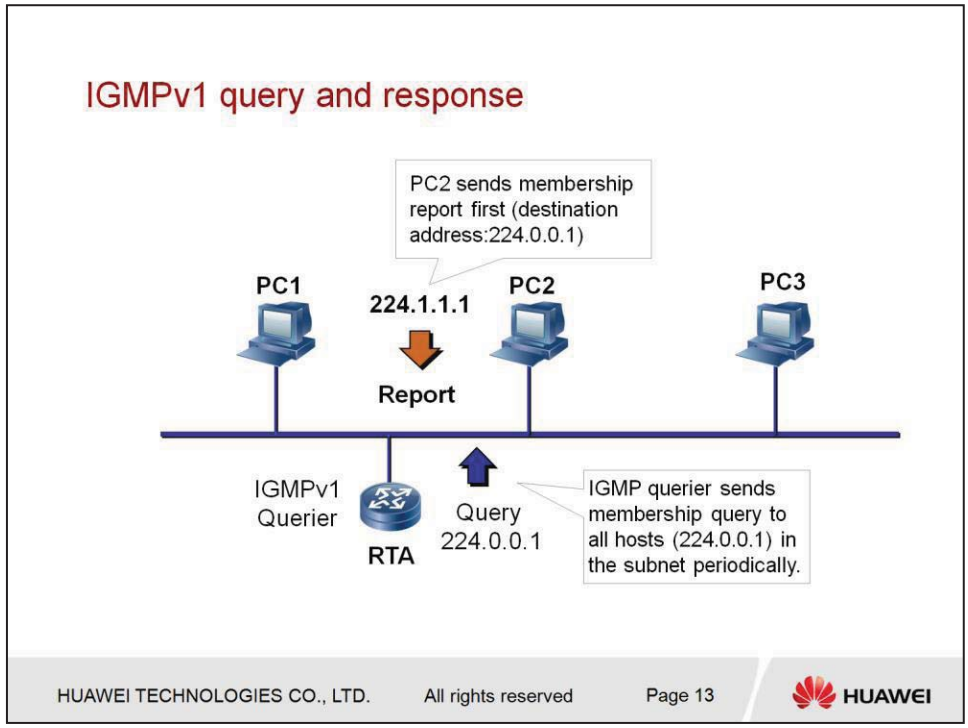
In a membership report message, the group address is a specified multicast address.

In a membership query message, the group address is 0.



As shown in the figure, the host sends an IGMP membership report message actively to indicate that it wants to join a multicast group.

The group address in this packet is the address of the multicast group the host wants to join.



The IGMPv1 query and response process is as follows:

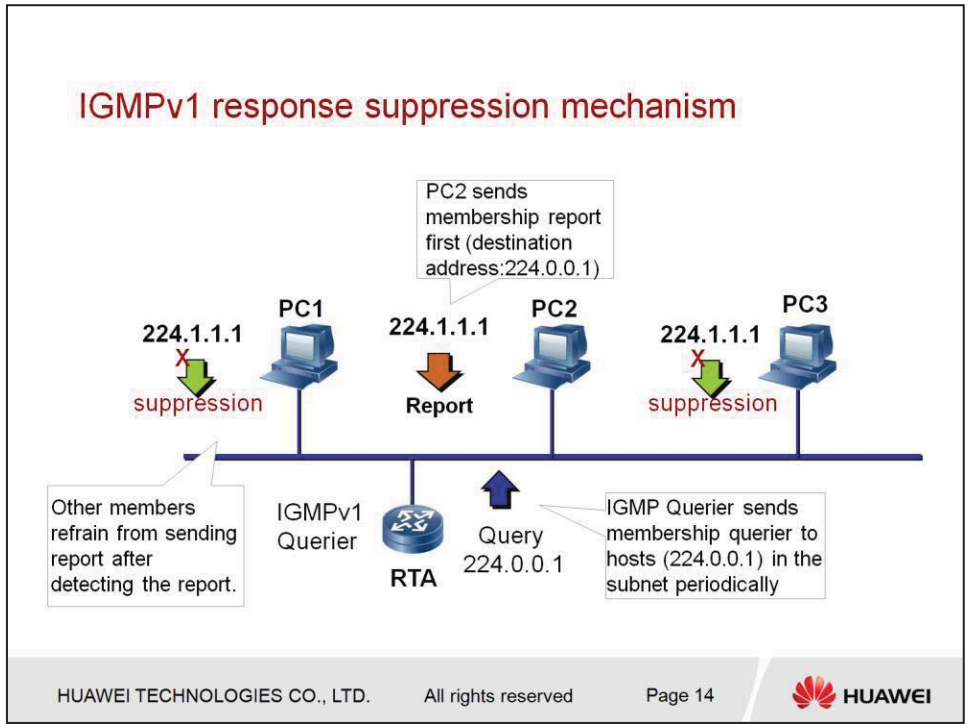
1. The IGMP querier periodically sends membership query messages (in which the group address is 0) to all hosts in the shared network segment in multicast mode. The destination of the membership query messages is 224.0.0.1.

2. All hosts in the network segment receive the query message. If some hosts (for example, PC1, PC2, and PC3) want to join multicast group G, they send membership report messages in multicast mode to announce that they will join group G. The membership report message contains the address of multicast group G. Assume that PC2 sends the membership report first.

3. After the query and response, the IGMP router finds the receiver of group G in the network. Then the router generates a (*, G) multicast entry and forwards multicast packets based on this entry.

* Indicates a multicast group member, and G indicates multicast group.

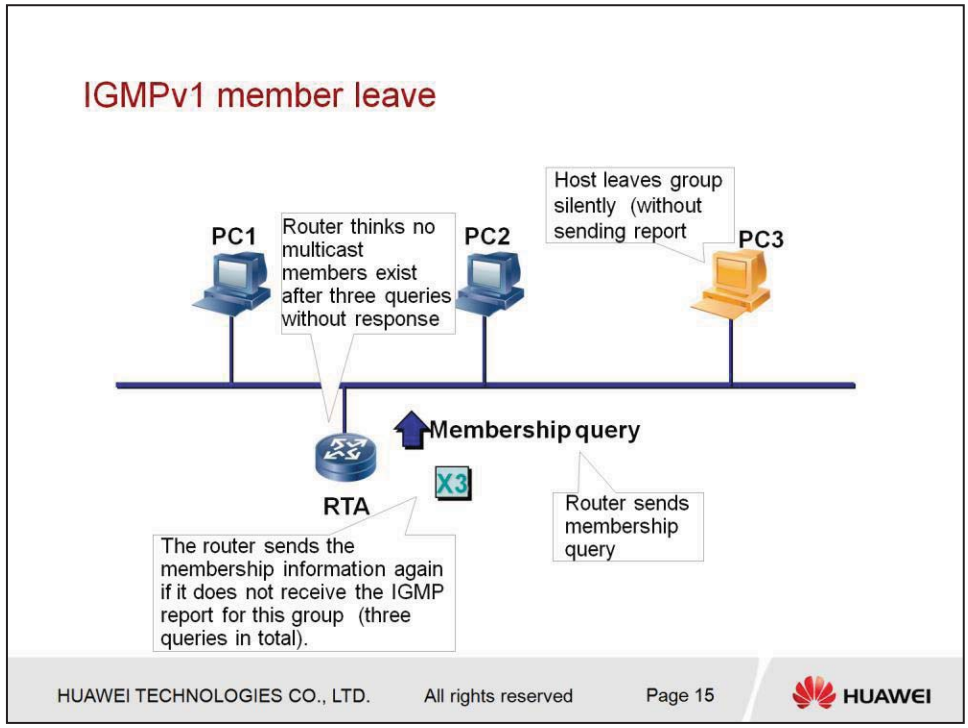
As specified in IGMPv1, when multiple routers exist in the shared network, a querier is elected based on the multicast routing protocol. The election mechanism varies with the multicast routing protocol.



The destination of the IGMP membership query message is 224.0.0.1. That is, all hosts in this network segment can receive this packet. However, not all receivers respond to this query. In this example, Only one host responds with the membership report message. Other hosts suppress transmission of membership report messages.

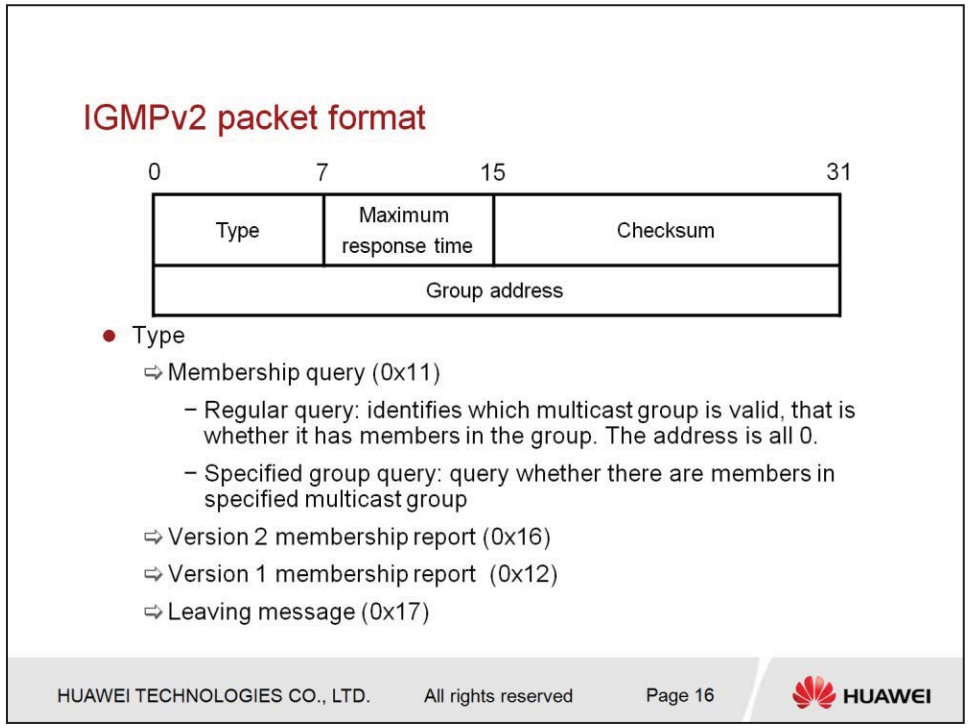
When a host receives the IGMP membership query message, the host starts a count down timer for each multicast group it has joined. In IGMPv1, the timer value is fixed to 10 seconds. When a timer expires, the host actively sends a membership report message. The group address is the address of the corresponding multicast group, and the destination address is 224.0.0.1. Thus all other hosts in this network segment receive the membership report message. The hosts suppress membership report messages and reset the timer.

When the router sends the membership query messages again, each host starts the timer again to repeat the process of query, response, and suppression.



IGMPv1 does not define the packet for leaving a multicast group, so a host leaves a multicast without sending any packets. Then how does the multicast router know the user has left the group? The answer is response timeout. IGMPv1 is implemented based on response timeout.

After members leave the group without sending any packets, the router still sends membership query messages every 60 seconds. If the router sends membership query messages three times but does not receive any membership report messages, the router considers that no member exists in this group and stop forwarding multicast packet to this network segment.



Different from the IGMPv1 packet, the IGMPv2 packet does not contain the version field but has a maximum response time field.

IGMPv2 supports three types of packets:

Type=0x11 membership query message, which falls into the following two subtypes:

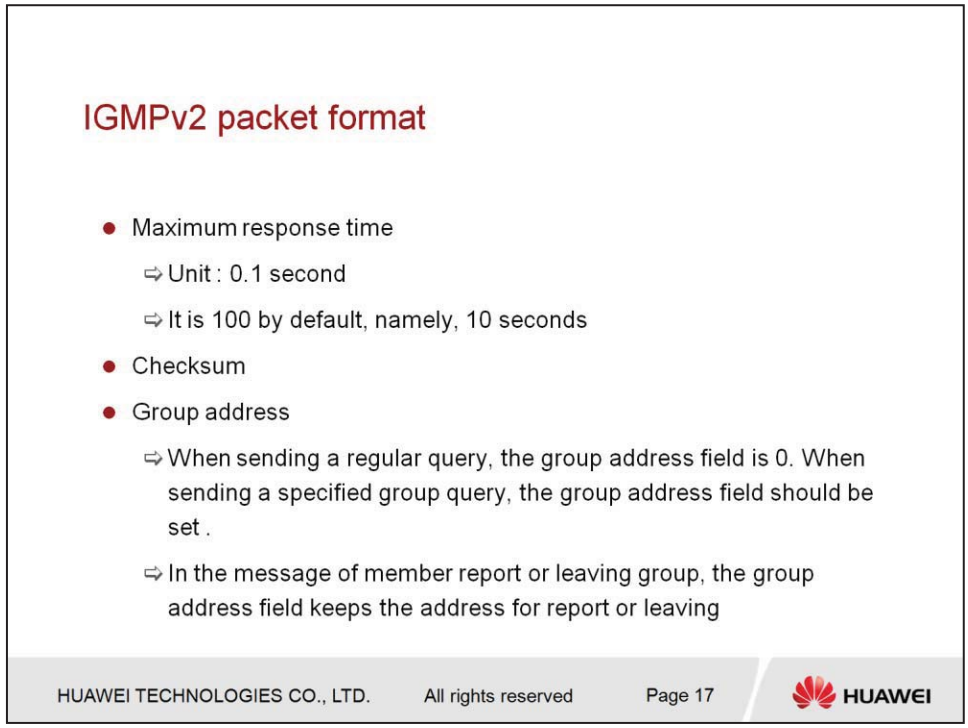
•Regular query: determines whether each multicast group is valid, that is, whether the group has members. The group address in a regular query message contains all 0s.

•Specified group query: checks whether a specified multicast group has members. The group address is a specified multicast address.

Type=0x16 IGMPv2 membership report

To be compatible with IGMPv1, IGMPv2 defines an additional packet type 0x12 = IGMPv1 membership report

Type=0x17 leaving message, sent by a host actively



The maximum response time field is used only in the membership query message.

This field indicates the period for the host to respond to the query message. The IGMPv1 packet does not contain this field.

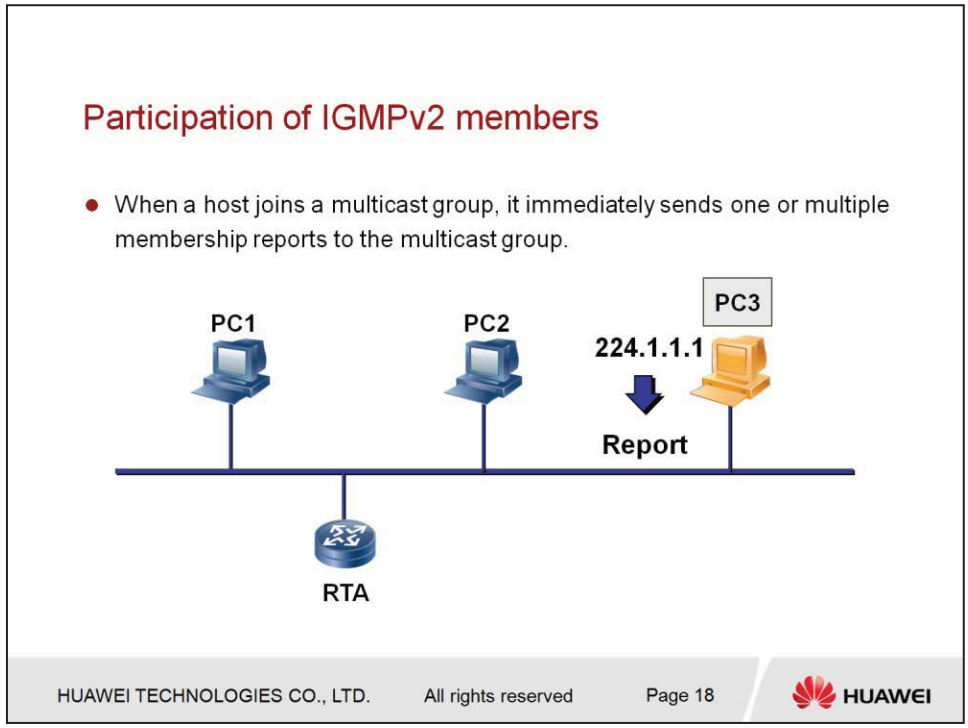
Group address field:

In the regular query message, the group address is 0.

In the specified group query message, the group address is the address of the group to be queried.

When a member host responds with a membership report message or leaving message, the group address field is the address of the target group.

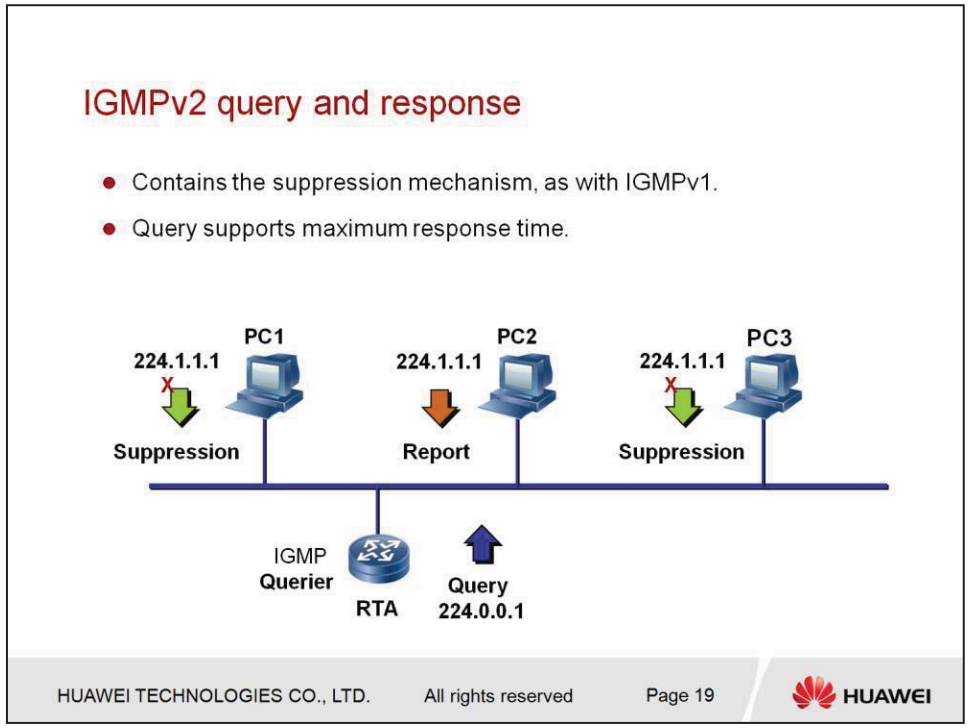
The checksum field is a 16-bit field used to check the length of the IGMP message, namely, the valid load of an IP packet.



When a host joins a multicast group for the first time, the host sends the membership report message immediately. To avoid loss or damage of the initial membership report, it is recommended to set the host to send the message once

or twice, again in a short interval. The interval recommended in RFC2326 is 10 seconds.

An IGMPv2 host also supports the membership report message of IGMPv1.

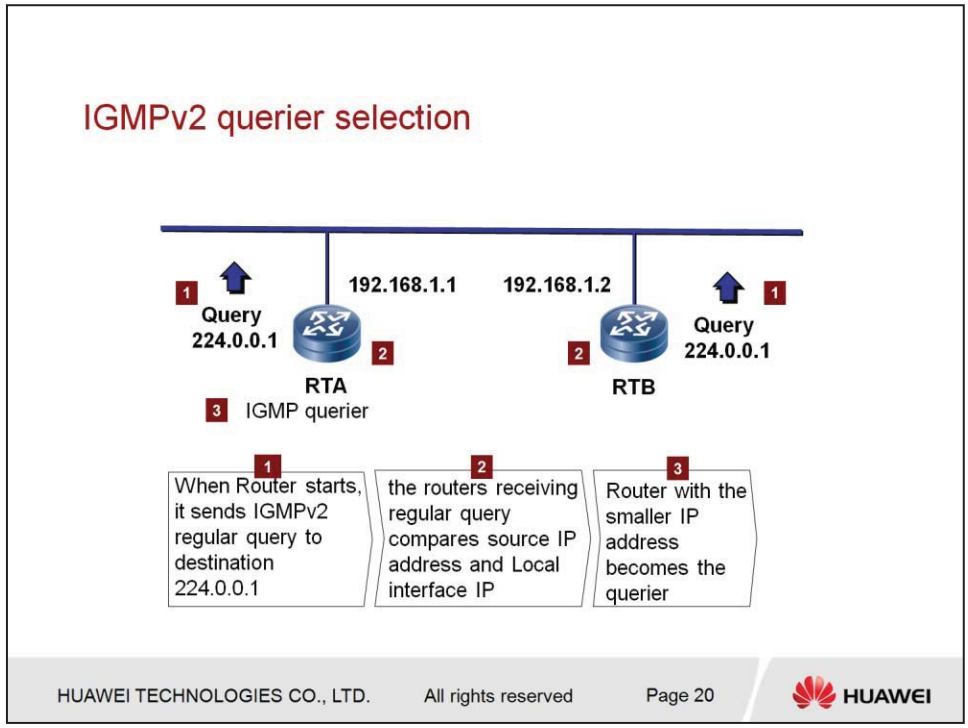


The maximum response time field is a new field in IGMPv2. As described before, a host starts a timer for each multicast group it has joined when it receives a membership query message. The host sends the membership report after the timer expires. In IGMPv2, the timer is a random value ranging from 1 to the maximum response time.

The maximum response time field is used to adjust the duration for a host to respond to the membership query message and specifies the maximum allowed time before sending a responding report in units of 1/10 of a second. In all other messages, it is set to zero by the sender and ignored by receivers.

In IGMPv1, the membership query message initiated by the multicast router is sent to all multicast groups in the network segment. Such query message is a regular query message.

IGMPv2 defines a specified group query message. The destination address is the specified group query message is the IP address of a multicast group. Only the members of this group respond to this query message. Members of other groups do not send response messages.



All routers running the IGMP on a shared network can receive the membership report from hosts. So, only one router needs to send the membership query message. Thus, a mechanism is needed to elect a router as the querier. The IGMPv1 uses the multicast routing protocol to elect the querier. The IGMPv2

specifies the multicast router with the smallest IP address on a shared network as the IGMP querier.

Only the querier sends membership query messages. In IGMPv1, the querier is elected through the multicast routing protocol.

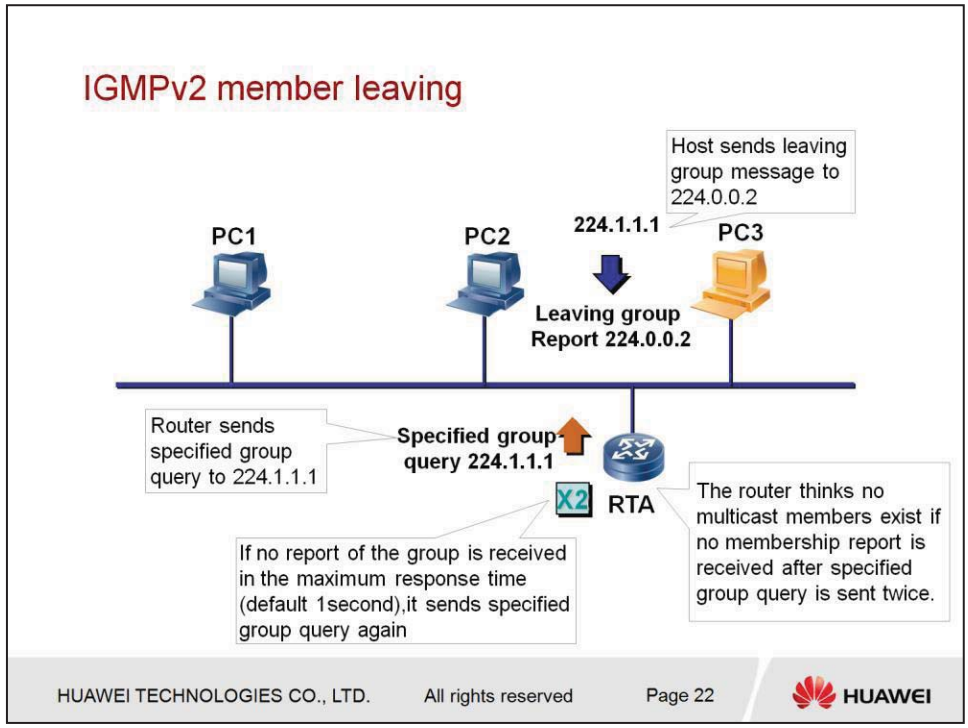
IGMPv2 modifies the election mechanism. IGMPv2 stipulates that the multicast router with the smallest IP address is elected as the querier.

When a router starts, it sends an IGMP regular query message to destination address 224.0.0.1.

When a router receives the regular query message, it compares the source IP address of this message with the IP address of the receiving interface. The router with the smallest IP address is elected as the querier.

Sometimes the querier is Down and another router becomes the

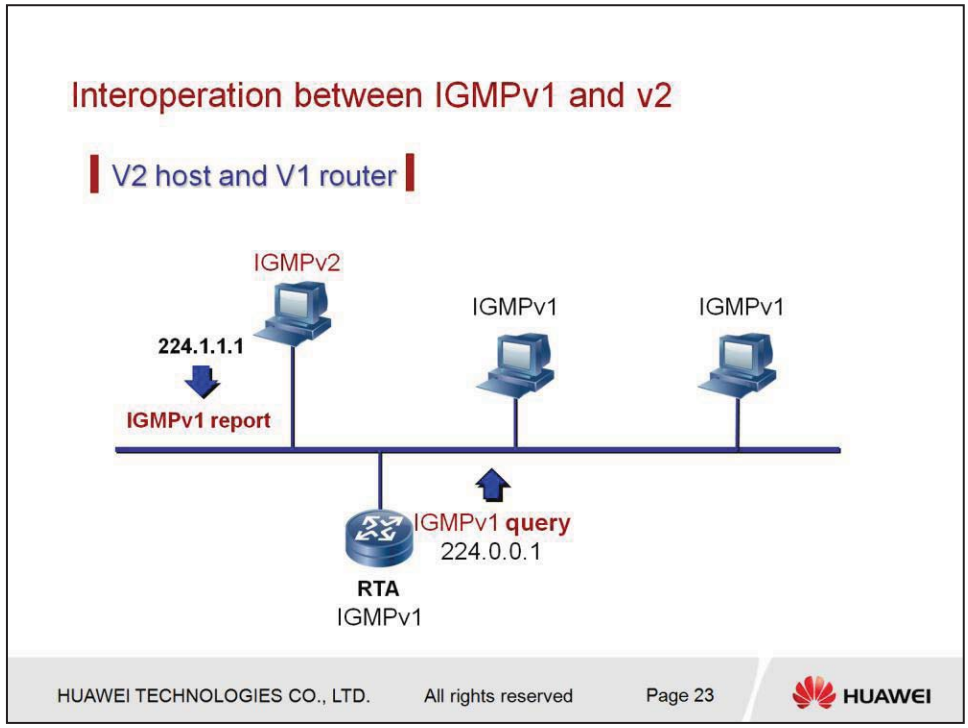
querier. All routers except the current querier start a query timer to periodically check the status of the IGMP querier. The check interval is 120 seconds. The interval can set by the timer other-querier-present interval command.



In IGMPv1, a host does not send any message to the multicast router when leaving a multicast group. The multicast router has to judge whether a member leaves the group by the response timeout.

In IGMPv2, if a host decides to leave a group, it sends the leaving message to all multicast routers in the network in multicast mode. That is, the leaving message is sent to the destination address (224.0.0.2). To check whether this multicast group contains other member hosts, the multicast router sends the specified group query message to the network. If the router does not receive the response within the maximum response time (1 second by default), the router sends the specified

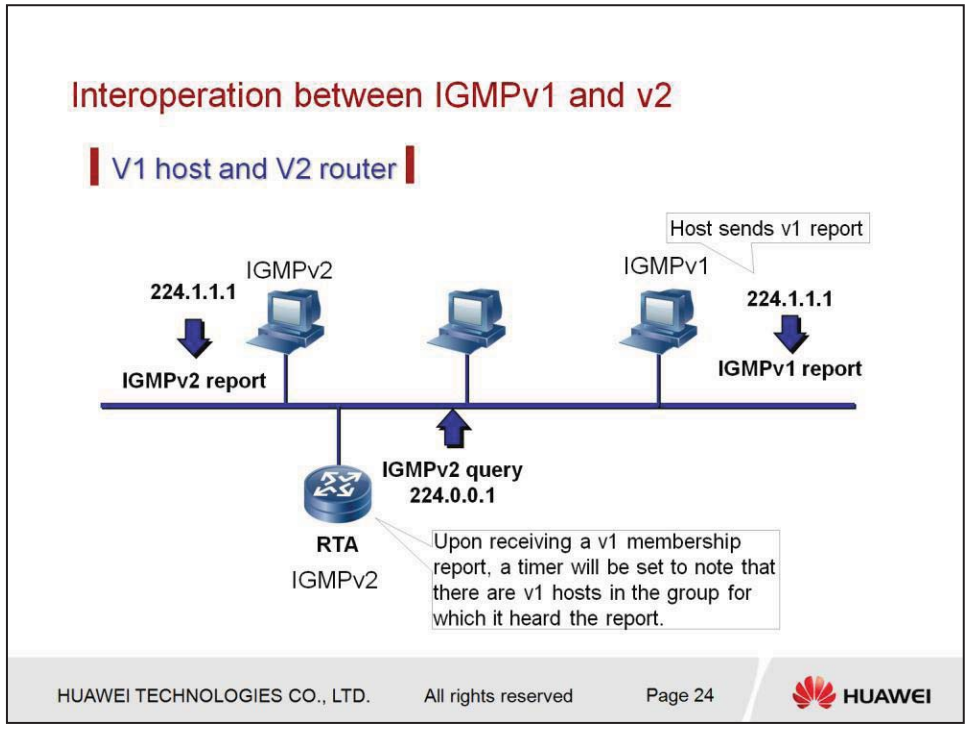
group query message again. If the router still does not receive any response, it considers that all members have left this group.



A router running IGMPv1 regards IGMPv2 messages as invalid and ignores such messages. When the IGMP querier runs IGMPv1, hosts that run IGMPv2 must send IGMPv1 messages.

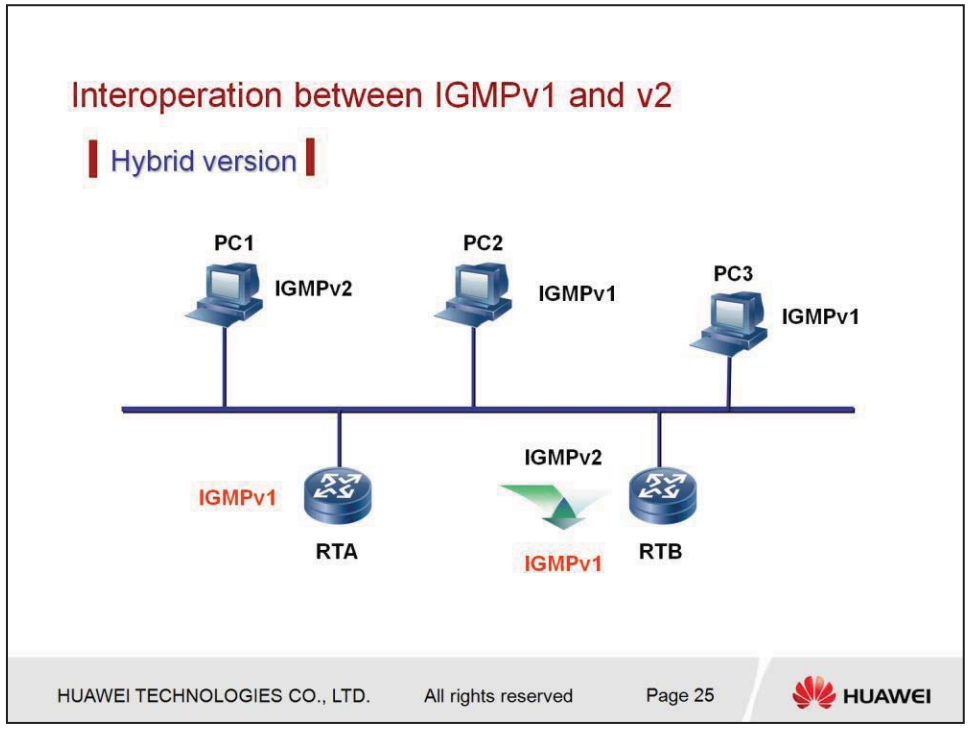
When an IGMPv2 host detects that the querier is an IGMPv1 router, the host must respond with IGMPv1 messages. In this case, the IGMPv2 host can also suppress the leaving message. To maintain the interface status, whenever the interface receives an IGMPv1 query message, the IGMPv2 host starts 400-second timer. When the next IGMPv1 query message is received, the timer is reset. When the timer expires, this interface is restored to an IGMPv2 interface and sends IGMPv2 message again.

An IGMPv2 host must allow its membership report to be suppressed by IGMPv1 or IGMPv2 membership report.

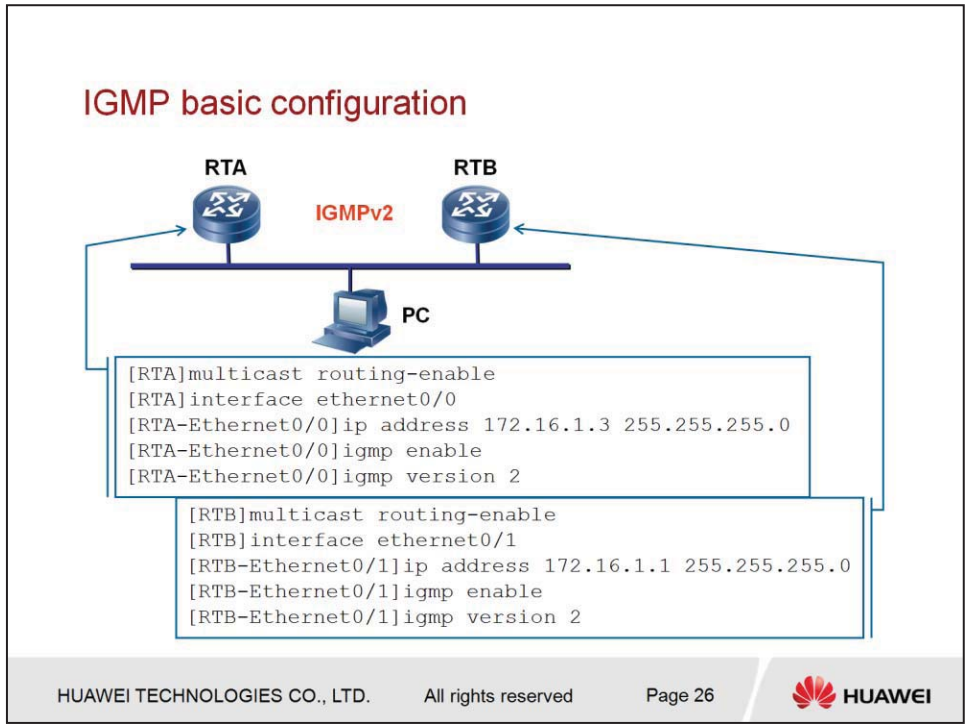


If the IGMP querier is an IGMPv2 router, and the IGMPv2/host is in the same group, all IGMPv1 messages of this group are received. IGMPv2 messages are not suppressed on IGMPv1 hosts. An IGMPv1 host cannot analyze IGMPv2 messages, so it ignores such messages.

Whenever an IGMPv1 host joins the group, IGMPv2 leaving message is suppressed.



If some IGMPv1 routers exist in the subnet, IGMPv1 must be enabled on all routers in this subnet.



IGMP is applied to the network that connects the router and users. Both the router and user host must run IGMP. Here, only the IGMP configuration on the router is mentioned.

Before configuring IGMP, you must enable IP multicast routing, because IP multicast routing is the prerequisite to all multicast functions. If IP multicast routing is disabled, all related configurations are canceled.

[Huawei]multicast routing-enable

Enable multicast routing.

To enable IGMP on the interface connected to the user host, you need to configure the matching IGMP version on the router and host, because different versions support different packets.

[Huawei-Ethernet0/0] igmp enable

Enable IGMP on an interface.

To add the hosts in the network connected to the interface to specified groups and to enable the interface to receive packets from these groups, you can set an ACL on this interface to limit the range of multicast groups the interface serves.

The IGMP version can be configured in the following modes:

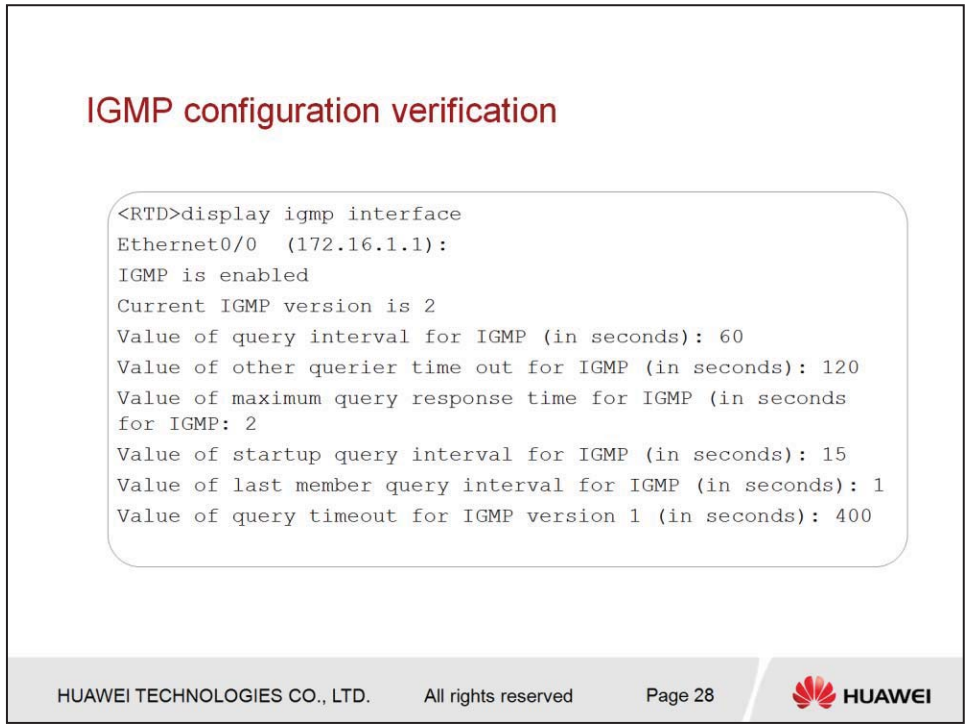
1. Interface configuration
2. [Huawei-Ethernet0/0]igmp version 2

The configuration on the interface takes precedence of the global configuration. If the version is not configured on the interface, the global configuration takes effect.

2. Global configuration

[Huawei] igmp //Enable IGMP globally.

[Huawei-igmp] igmp version: 2 //Globally configure the IGMP version.



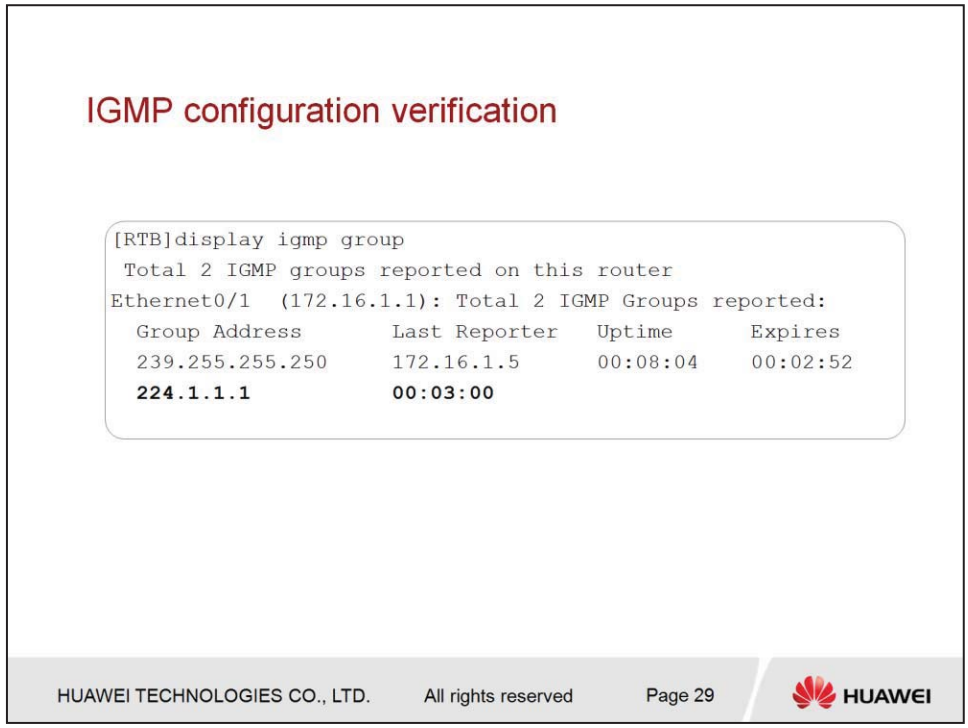
The display igmp interface command is used to display the configuration and running information of IGMP on an interface. If this command is used in global mode, the IGMP running information of all IGMP-enabled interfaces is displayed. If this command is used on an interface, the IGMP running information of this interface is displayed.

If IGMP is configured properly, you can view the IP address, IGMP version, and IGMP parameters on an IGMP-enabled interface through this command.

Value of query interval for IGMP (in seconds): indicates the interval of regular query messages. The default value is 60 seconds.

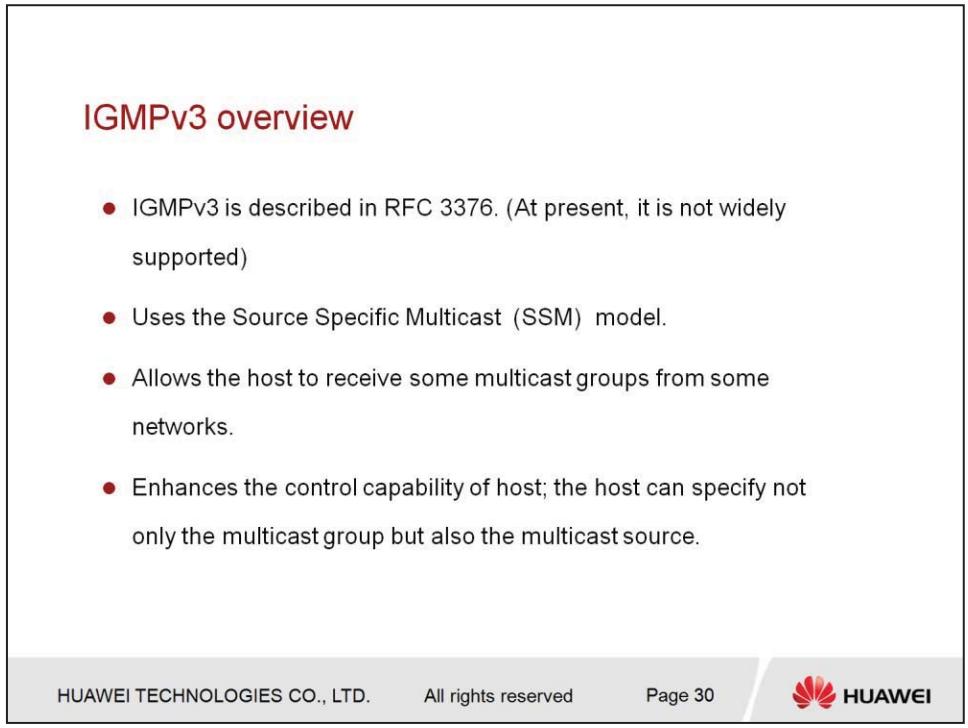
Value of other querier time out for IGMP (in seconds): 120 indicates that the timeout duration of the IGMP querier is 120 seconds.

Other parameters related to IGMP are not mentioned here. For details, refer to the VRP configuration guide.



The display igmp group command is used to display the information about multicast groups. The display contains the multicast that the router joins dynamically through the membership report and the group that the router joins statically through command lines.

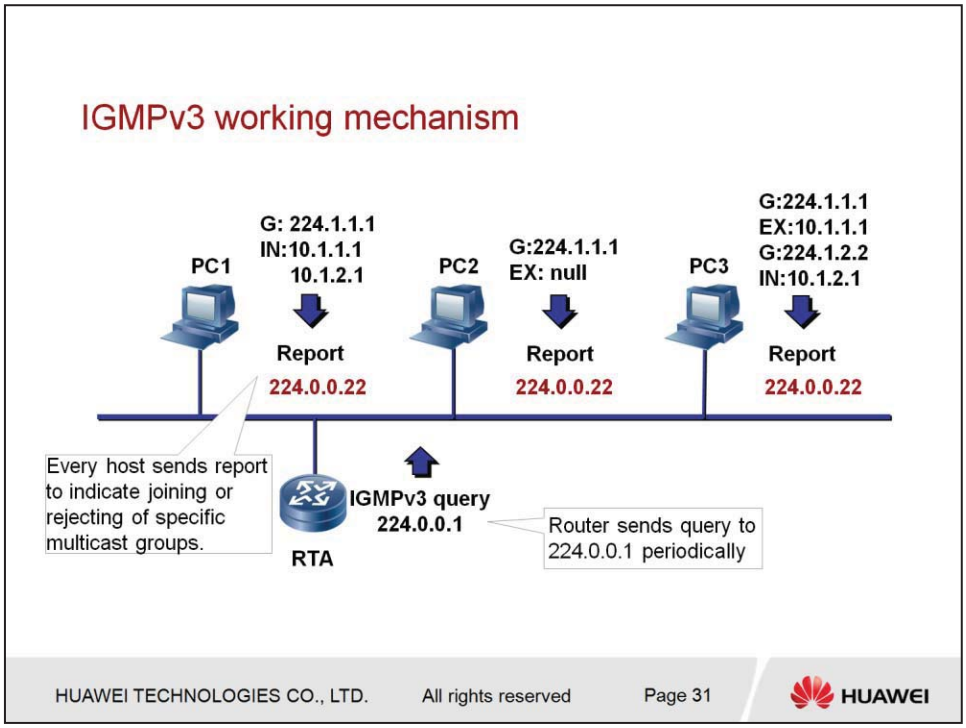
The bold line in the display shows the information about the multicast group that the router has joined.



IGMPv3 is compatible with IGMPv1 and v2. In addition, IGMPv3 enhances the capability of controlling hosts. It can specify the multicast group (G) a host can join and the multicast source (S) from which the host can receive messages.

If a host need to receive the information from specified sources, it can set the Filter-Mode field in the IGMP message to include and specify the multicast source address Sources in this IGMP message. Thus the host can receive multicast packets from the specified sources. To facilitate the description, we express this field as Include Sources (S1, S2, …).

If the host does not want to receive messages from specified sources, it can set the corresponding field as Exclude Sources (S1, S2, …). Thus, the host can receive packets from all sources except the specified source addresses.

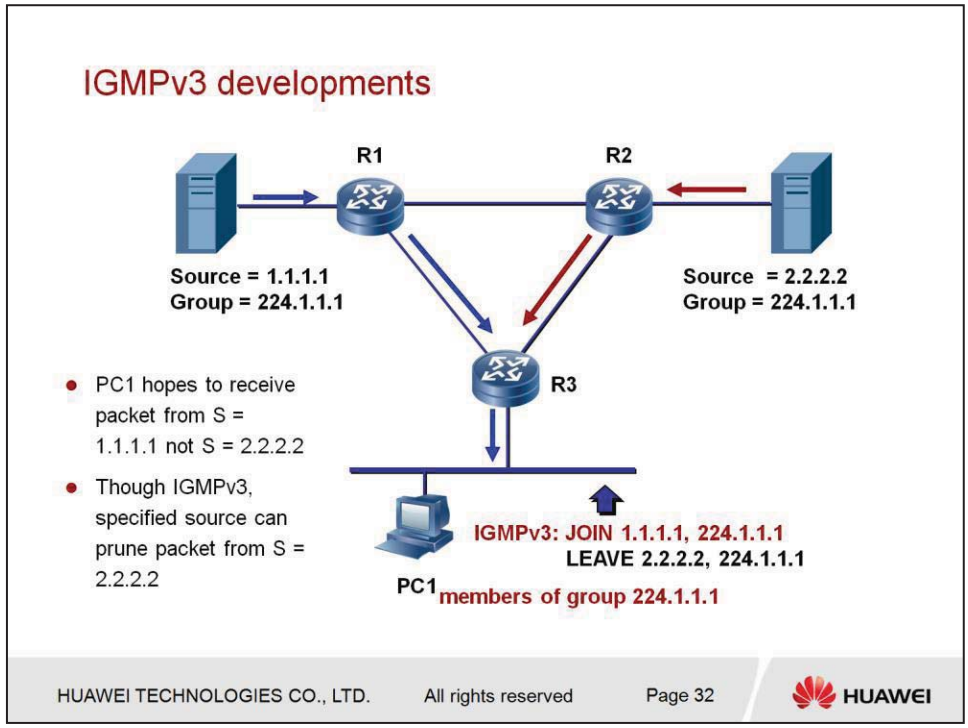


In IGMPv1 and v2, the response message and query message have the same format. That is, the messages contains only the group address. The IGMPv3 response message contains group address 224.0.0.22 and contains the record of one or more groups. Each group record contains a multicast group address and some source addresses. Multicast records are divided into the following types:

Current status record: records the current receiving state of the interface. The value can be Include or Exclude. Include indicates that the host receives messages from the specified source addresses. Exclude indicates that the host receives messages from all source addresses except the specified addresses.

Filter mode change record: records the change from Include state to Exclude state, or from Exclude state to Include state.

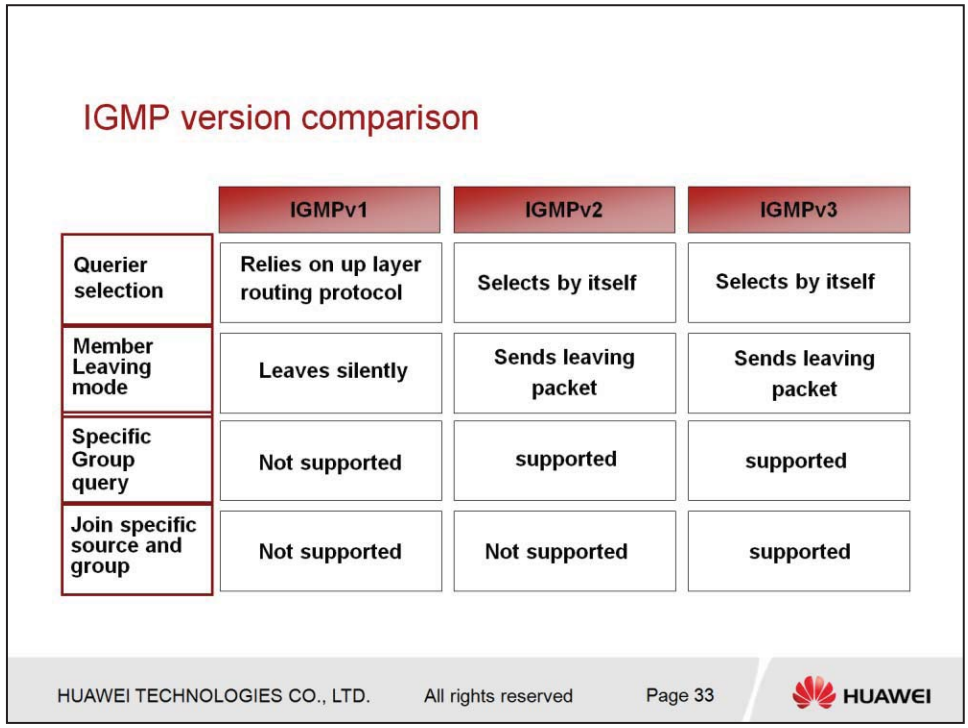
Source address list change record: records the new source addresses or deleted resource addresses.



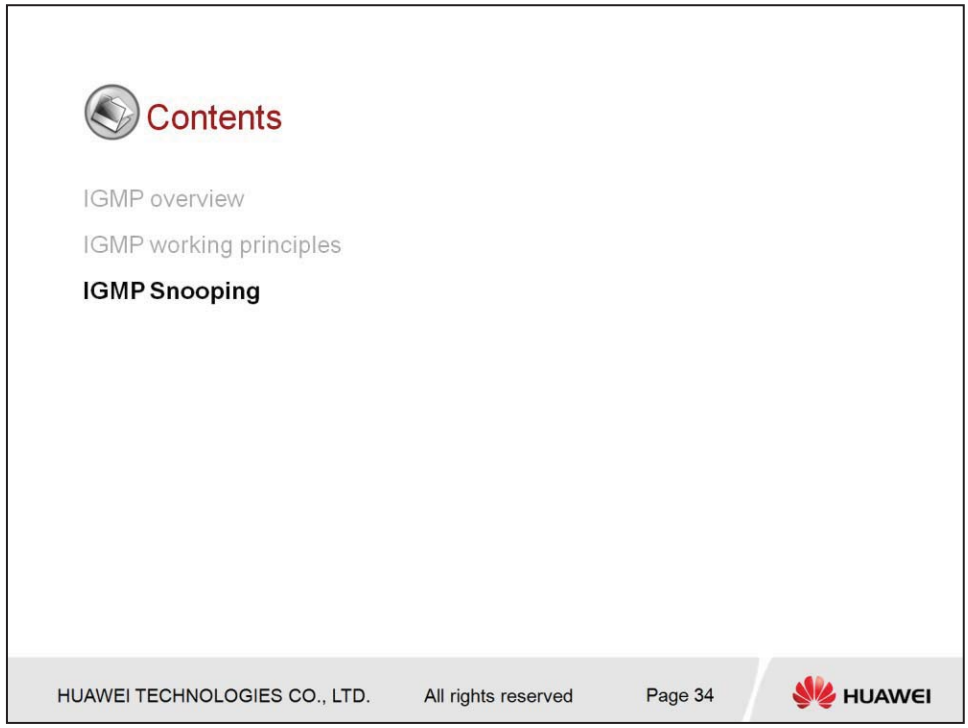
IGMPv3 supports the regular query message of IGMPv1, the specified group query message of IGMPv2. In addition, IGMPv3 supports the specified source/group query message. The IGMPv3 message contains the multicast source address and control fields (such as the robustness coefficient and query interval).

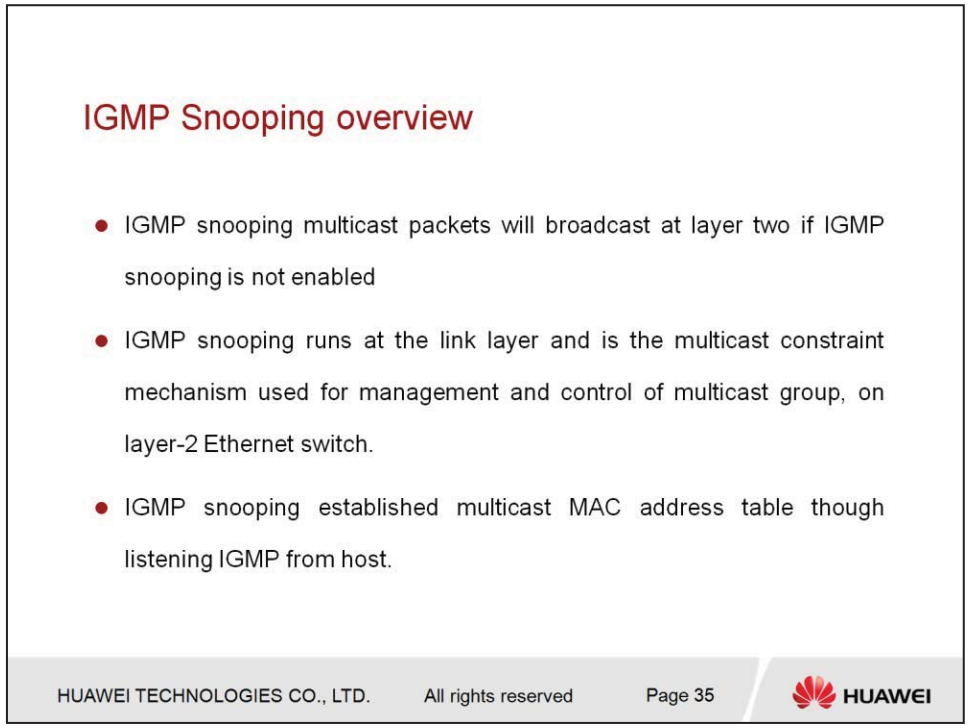
The regular query message does not contain the group address or source address.

The specified group query message contains the group address but does not contain the source address. The specified source/group query message contains the group address and one or more source addresses.



This table shows the comparison of three IGMP versions. IGMPv1 does not support querier election. The querier is elected through the multicast routing protocol.


Contents
IGMP overview
IGMP working principles
IGMP Snooping
HUAWEI TECHNOLOGIES CO., LTD.
All rights reserved
Page 34
HUAWEI

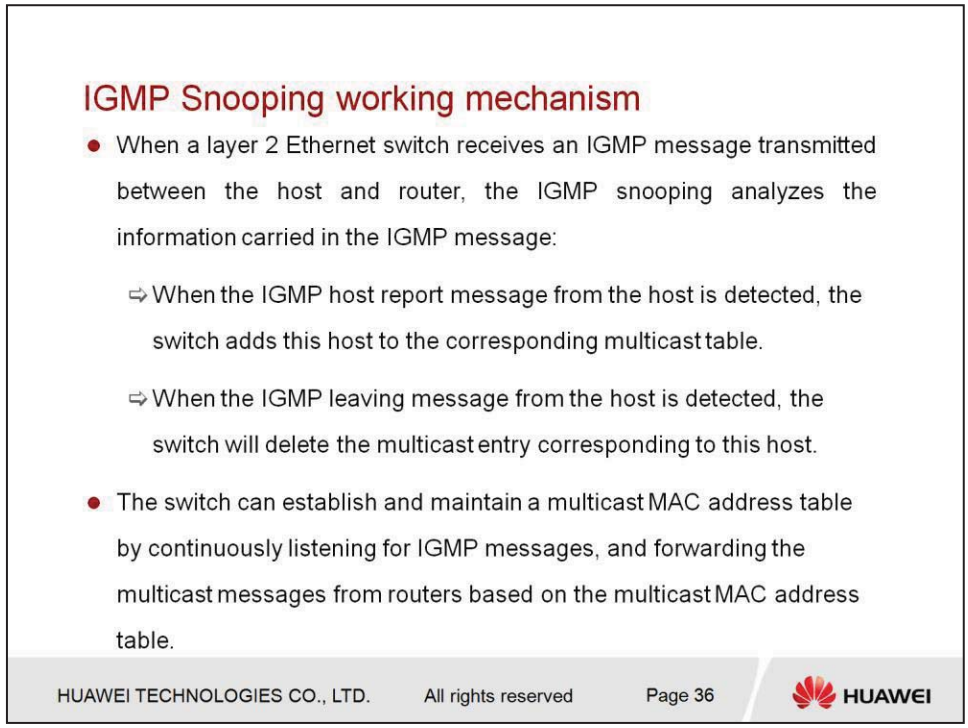


IGMP snooping runs at the link layer. It is a multicast constraint mechanism used to manage and control multicast groups.

When a layer-2 switch receives the IGMP packet transmitted between the host and router, the switch analyzes the contents of the packet through IGMP snooping.

If the switch detects an IGMP membership report message from the host, it adds the host to the matching multicast table. If the switch detects the leaving message from the host, it deletes the multicast entry for this host. By snooping IGMP packets, the switch establishes and maintains a multicast MAC address table on layer 2. Then the switch can forward multicast packets delivered by the router based on the MAC multicast address table.

If IGMP snooping is not enabled, multicast packets are transmitted on layer 2 through broadcast. After IGMP snooping is enabled, packets are transmitted on layer 2 through multicast.

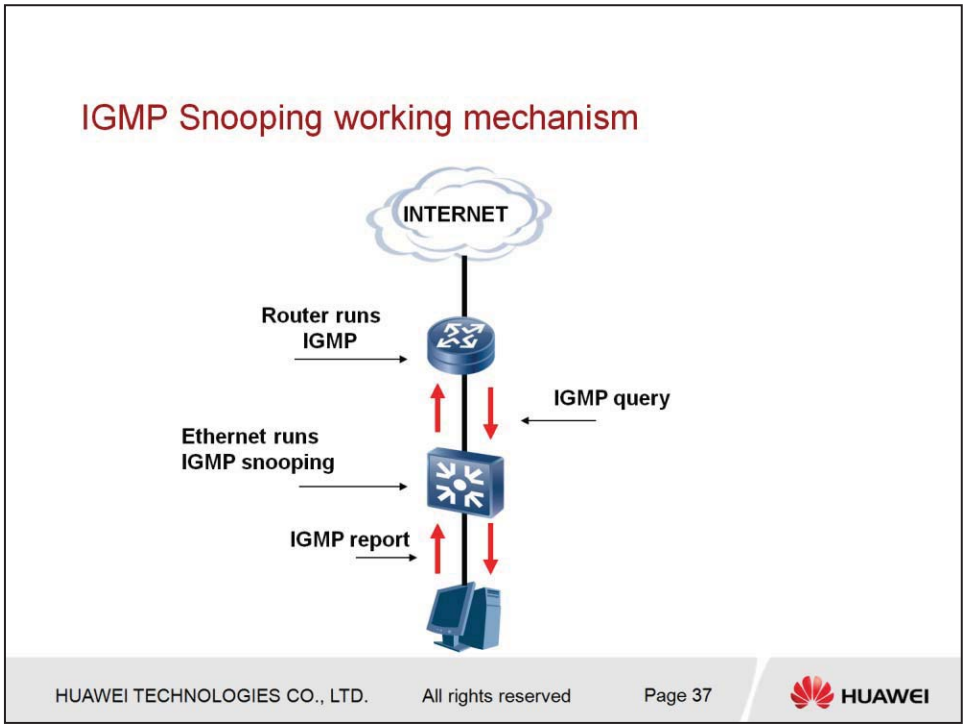


IGMP snooping runs at the link layer. It is a multicast constraint mechanism used to manage and control multicast groups.

When a layer-2 switch receives the IGMP packet transmitted between the host and router, the switch analyzes the contents of the packet through IGMP snooping.

If the switch detects an IGMP membership report message from the host, it adds the host to the matching multicast table. If the switch detects the leaving message from the host, it deletes the multicast entry for this host. By snooping IGMP packets, the switch establishes and maintains a multicast MAC address table on layer 2. Then the switch can forward multicast packets delivered by the router based on the multicast MAC address table.

If IGMP snooping is not enabled, multicast packets are transmitted on layer 2 through broadcast. After IGMP snooping is enabled, packets are transmitted on layer 2 through multicast.



Through IGMP snooping, a LAN switch detects IGMP messages and establishes the mapping between the host, related interface, and the multicast group address.

The working mechanism of IGMP snooping on a layer-2 switch is shown in the figure.

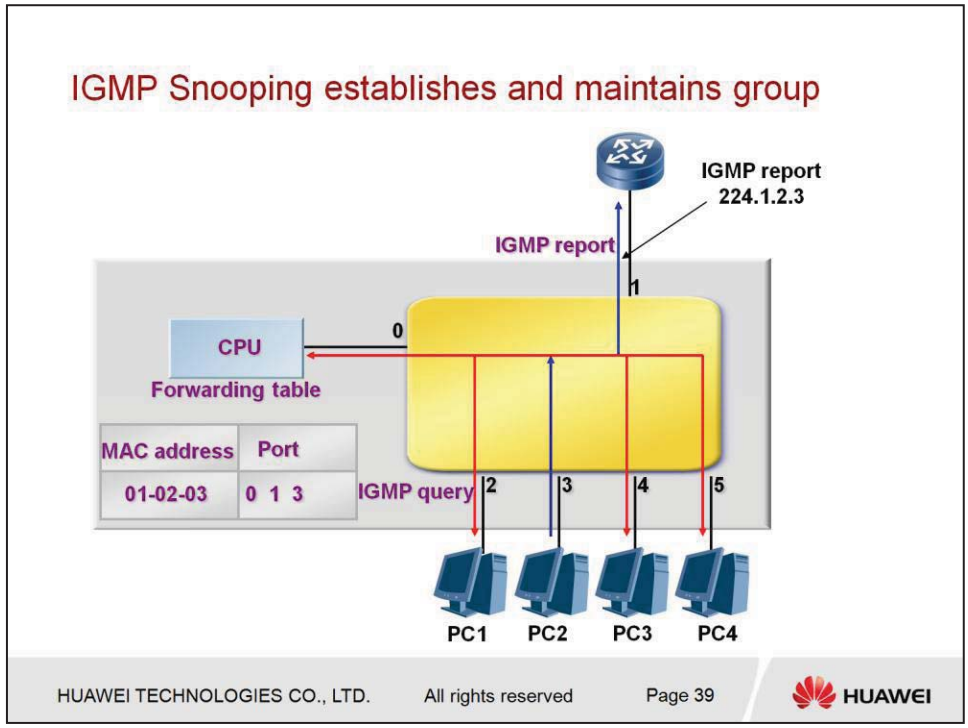
When receiving an IGMP regular query message, the switch checks the source port. If the source port is the port of the router, the switch reset the aging timer for this port; otherwise, the switch starts the aging timer for the router port.

When receiving an IGMP specified group query message, the switch sends the specified group query message only to the IP multicast group to be queried.

When receiving an IGMP report message, the switch checks whether the MAC multicast group mapping the IP multicast group the host wants to join exists. If the MAC multicast group does not exist, the switch creates a MAC multicast group and adds the port that receives this report message to this group. At the same time, the switch starts the aging timer of this port and adds all router ports in the same VLAN to this MAC multicast group. The switch also creates an IP multicast group and adds the port that receives this message to the group. If the MAC multicast group

exists, but the port that receives this message is not in the MAC multicast group, the switch adds the port to the MAC multicast group and starts the aging timer for this port. Then the switch checks whether the mapping IP multicast group exists. If the IP multicast group does not exist, the switch create an IP multicast group and adds the port that receives this message to this group. If the IP multicast group exists, the switch adds the port that receives this packet to the IP multicast group.If the MAC multicast exists, and the port that receives this message is in the MAC multicast group, the switch resets the aging timer for the port that receives this message.

When the switch receives a leaving message, it sends the specified group query to the port that receives this message to check whether other hosts connected to this port are the member of the multicast group the host wants to leave. At the same time, the switch starts a query response timer. If the switch does not receive the report message from this multicast group, the switch deletes this port from the MAC multicast group. If the MAC multicast group does not contain any member ports, the switch requests the router to delete this branch from the multicast tree.



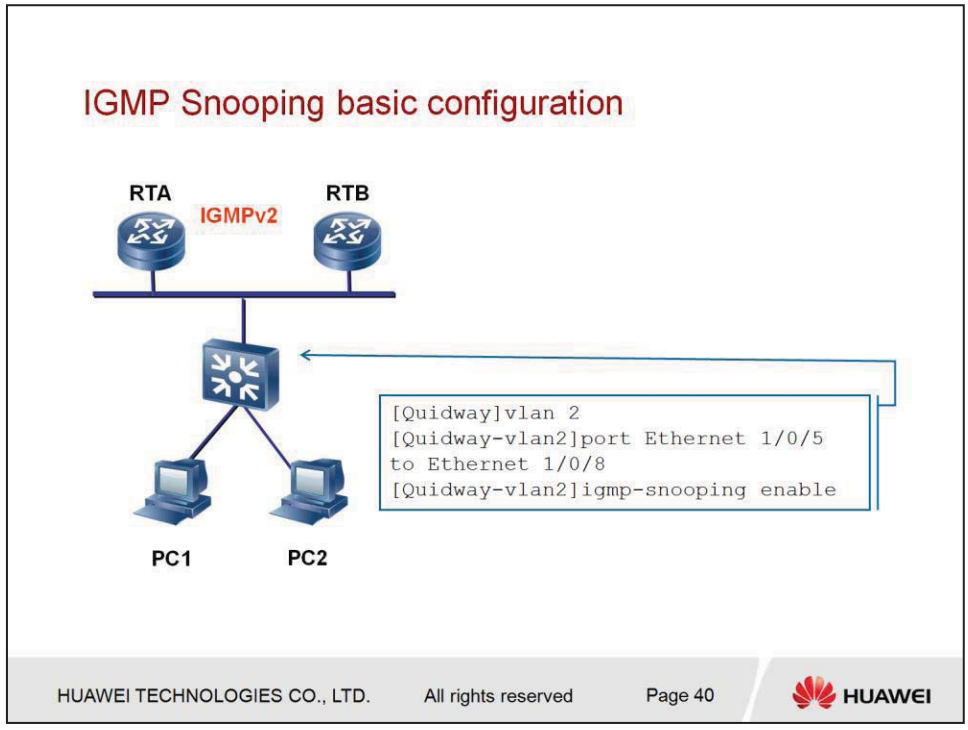
This figure shows how the information about multicast groups is established and maintained through IGMP snooping.

1. PC2 wants to join multicast group 224.1.2.3, so PC sends an IGMP membership report to the multicast group. The report message contains MAC address 0100.5e01.0203. Initially, the forwarding table does not contain the entry with this MAC address. Therefore, this report message is flooded to all ports of the switch, including internal port 0 connected to the CPU.

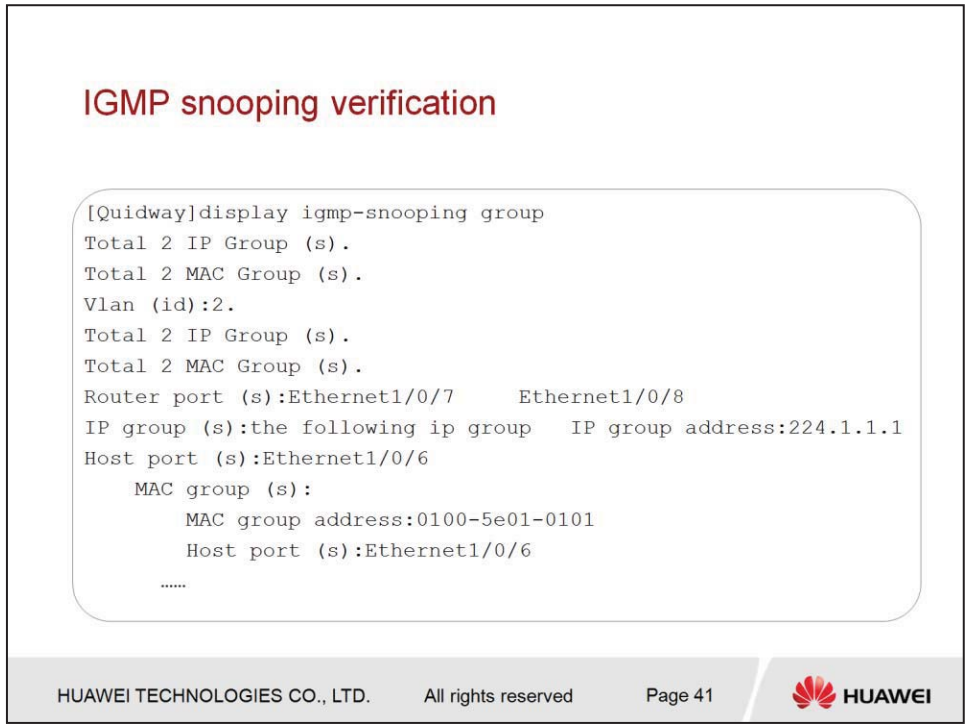
2. When the CPU receives the report message from PC2, the CPU establishes a forwarding entry according to the information in the IGMP report. This entry contains the port of PC2, the port of the connected router, and the internal port connected to the CPU.

3. After the entry is established, all multicast frames to destination address 0100.5e01.0203 are suppressed on ports 0, 1, 3 and cannot spread to other ports.

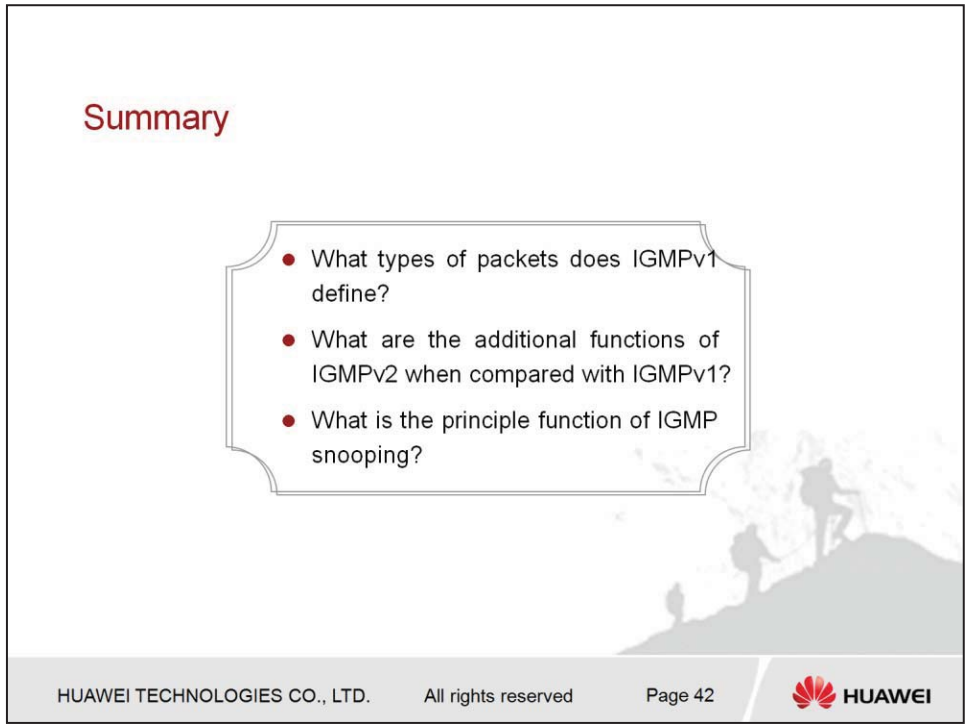
4. Assume that PC4 wants to join multicast group 224.1.2.3 and sends an IGMP report to this group. The switch forwards the report to port 1 and port 3 according to the forwarding entry. The CPU of the switch also receives this report and it adds a port (port 5) to the entry with MAC address 0100.5e01.0203.



The configuration of IGMP snooping on a layer-2 switch is quite simple. You only need to enable IGMP snooping in the system view or VLAN view.



The display igmp-snooping group command is used to display the IGMP information on a switch, including the port connected to the multicast router, IP address of the multicast group, and MAC address of the user host.



1. What types of packets does IGMPv1 define?

IGMPv1 defines the membership query packet and membership report message.

2. What are the additional functions of IGMPv2 when compared with IGMPv1?

IGMPv2 adds the member leaving packet. A multicast member sends a leaving packet to leaving the group. IGMPv2 also adds the specified group query packet.

The packet is used to check whether the multicast group that a member leaves contains other members.

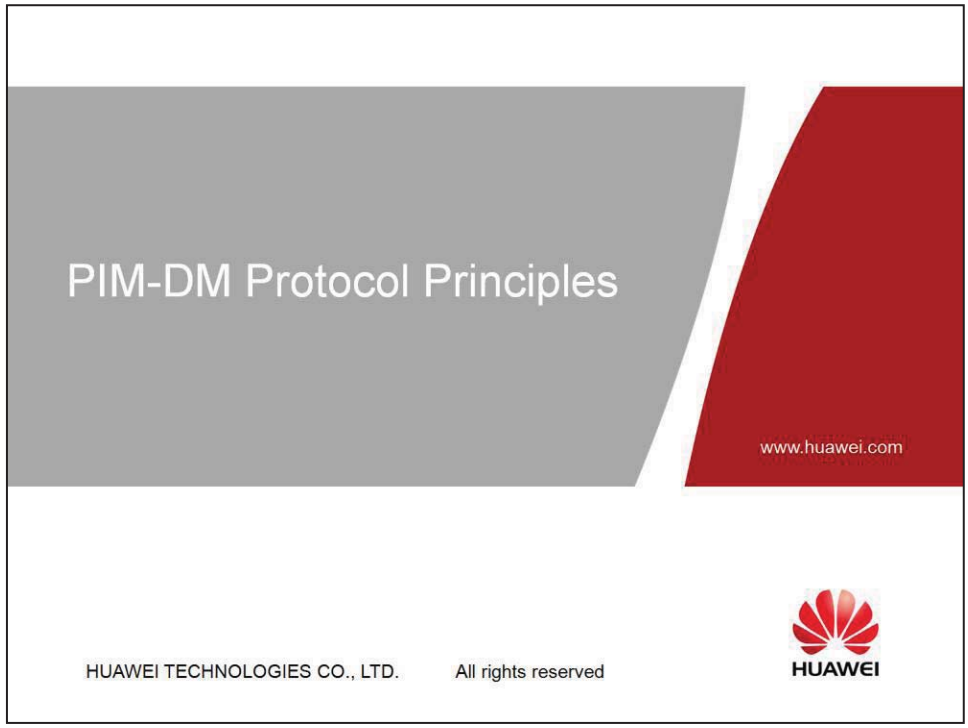
3. What is the principle function of IGMP snooping?

IGMP snooping runs on the link layer. It is the multicast restraint mechanism used to manage and control multicast groups.

When a layer-2 switch receives the IGMP packet transmitted between a host and a router, the switch analyzes the contents of the packet through IGMP snooping.

By detecting the IGMP packets, the switch establishes and maintains a Multicast MAC address table on layer 2. Then the switch can forward multicast packets

delivered by the router based on multicast MAC address table.


PIM-DM Protocol Principles
www.huawei.com
HUAWEI TECHNOLOGIES CO., LTD.
All rights reserved
HUAWEI



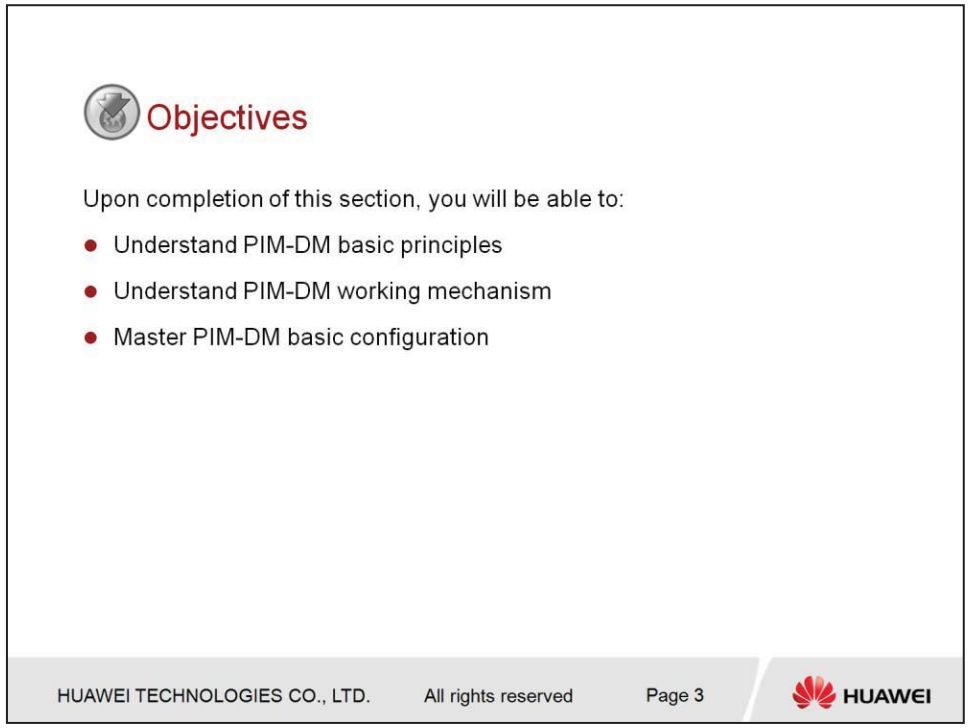
Foreword

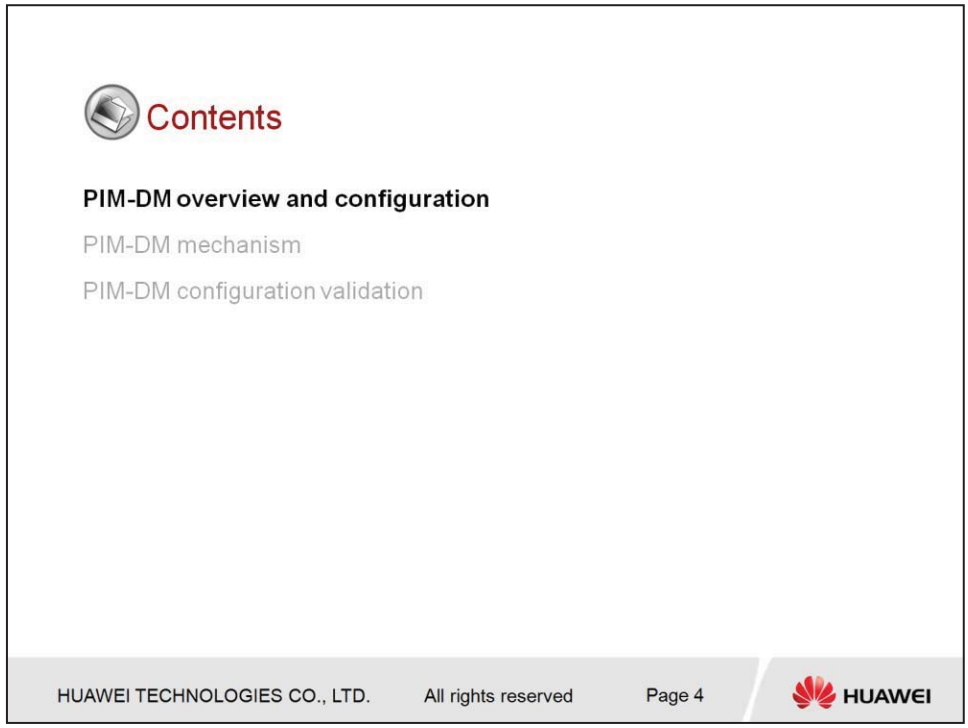
Multicast routing protocols are used between multicast routers to establish and maintain multicast routing, and forward of multicast packets accurately and efficiently.

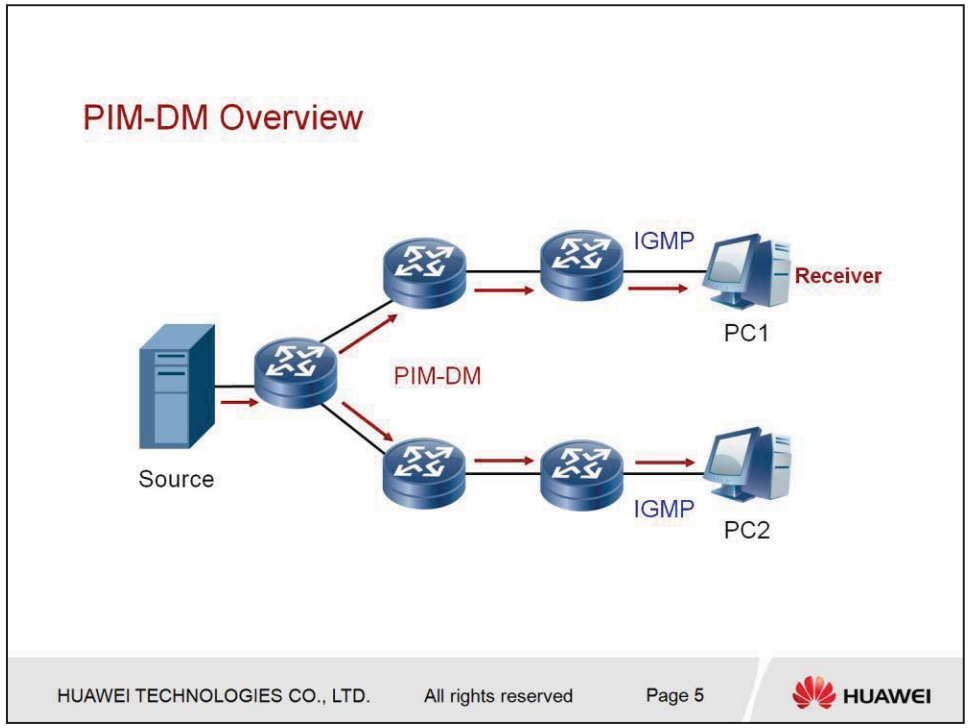
PIM (Protocol Independent Multicast) is a typical inter-domain Multicast routing protocol, divided into DM (Dense Mode) and SM (Sparse Mode) models.

HUAWEI TECHNOLOGIES CO., LTD. All rights reserved




Objectives
Upon completion of this section, you will be able to:
Understand PIM-DM basic principles
Understand PIM-DM working mechanism
Master PIM-DM basic configuration
HUAWEI TECHNOLOGIES CO., LTD.
All rights reserved
Page 3
HUAWEI


Contents
PIM-DM overview and configuration
PIM-DM mechanism
PIM-DM configuration validation
HUAWEI TECHNOLOGIES CO., LTD.
All rights reserved
Page 4
HUAWEI

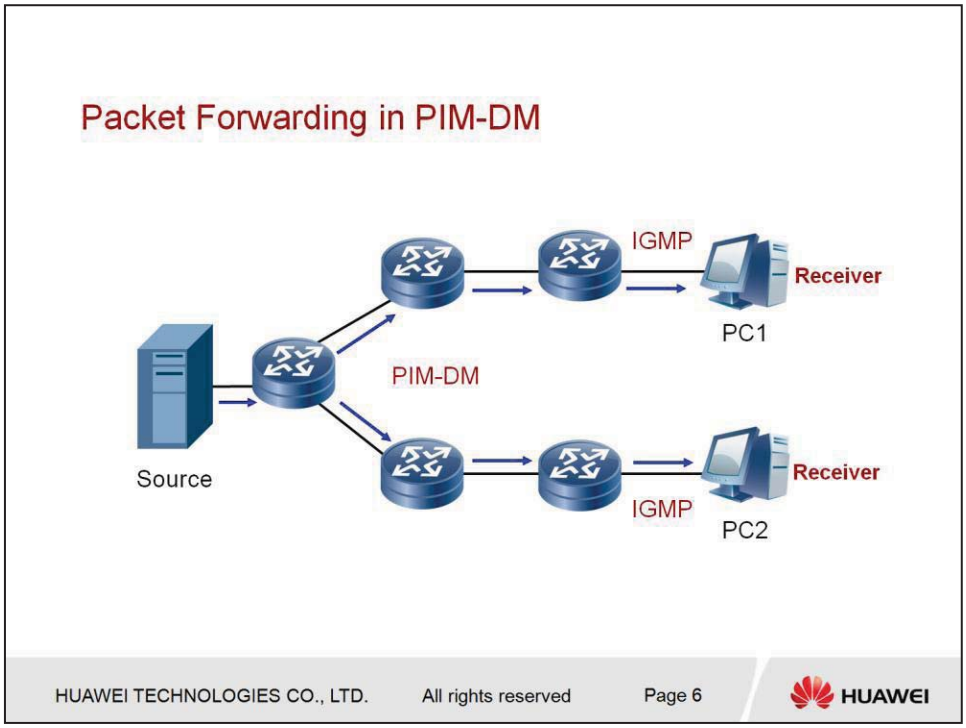


Protocol Independent Multicast (PIM) use all unicast routing protocols, to provide routing information for reverse with forwarding.The multicast routing is irrelevant with the unicast routing protocols, if only the unicast routing protocol can generate corresponding routing entries.Protocol Independent Multicast Dense Mode (PIMDM) is a dense mode multicast routing protocol applicable to small-sized networks, in which multicast receivers are located densely.

PIM-DM is based on the assumption that every subnet in the network has at least one receiving node of the multicast source. In this case, multicast packets are flooded to all nodes in the network, and thus related resources (bandwidth and router CPU) are consumed.

To reduce the consumption of network resources, the dense mode multicast routing protocol prunes the branches that do not need to forward multicast packets and only keeps the branches that contain the receivers. PIM-DM will sent graft message to restore forwarding of multicast packets on the pruned branches when pruned branches need to forward multicast packets.

Periodical packet flooding and pruning is the feature of the dense mode multicast routing protocol. Such protocol is applicable only to small-sized LANs.



In dense mode, the forwarding paths of multicast form a source tree.

The multicast source is the root of the source tree and the multicast receivers are leaves of the source tree. The source uses the shortest path from the multicast source to the receivers, so the tree is also called the shortest path tree (SPT). In

this figure, the arrows show the paths from the source to the receivers.

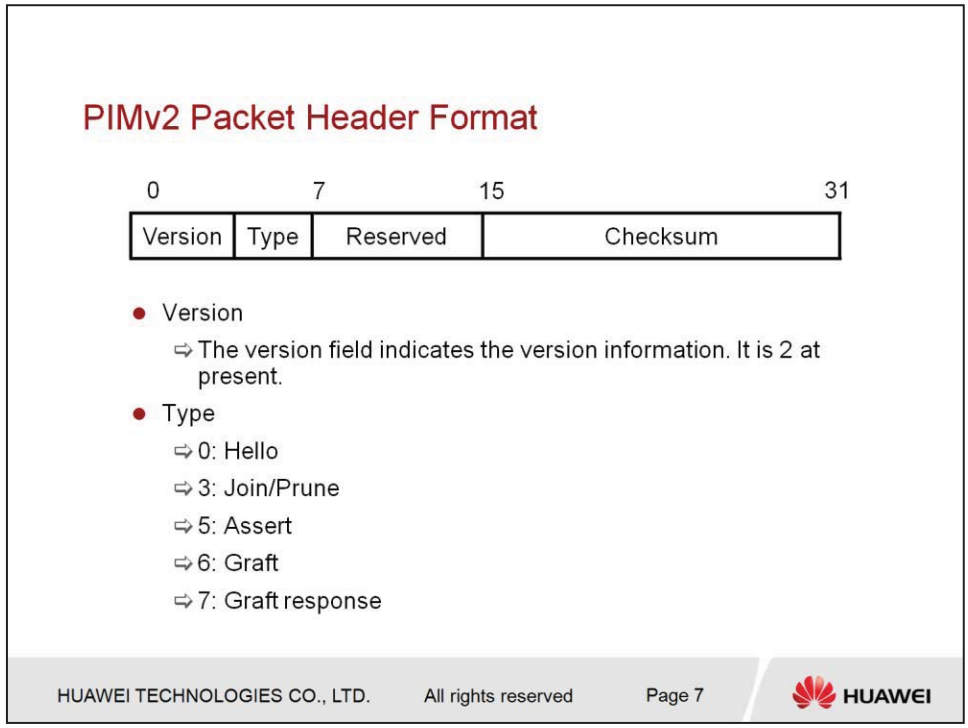
PIM-DM does not depend on specific unicast routing protocol. In stead, PIM-DM performs RPF check by using the existing unicast routing table.

Forwarding of packets has two directions, upstream and downstream.

The interface that receives the multicast packets is the upstream interface.

The interface that forwards the multicast packets is the downstream interface.

Packets are forwarded from the upstream interface to the downstream interface.

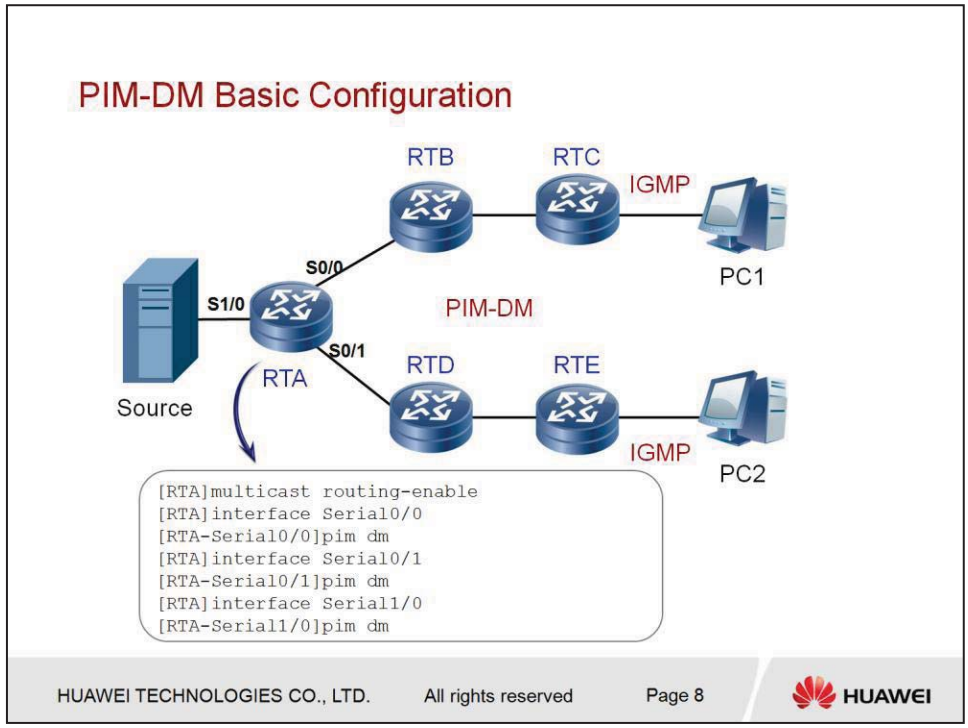


The protocol number is 103 in all PIM packets.

The packet types listed here are related to PIM-DM. Packet types related to PIM-SM will be introduced in the PIM-SM section.

Packet types used in PIM-DM are Hello message, Join/Prune packet, Graft packet, Graft response packet and Assert packet. The Assert packets are sent in multicast mode and the destination is all PIM routers on 224.0.0.13.

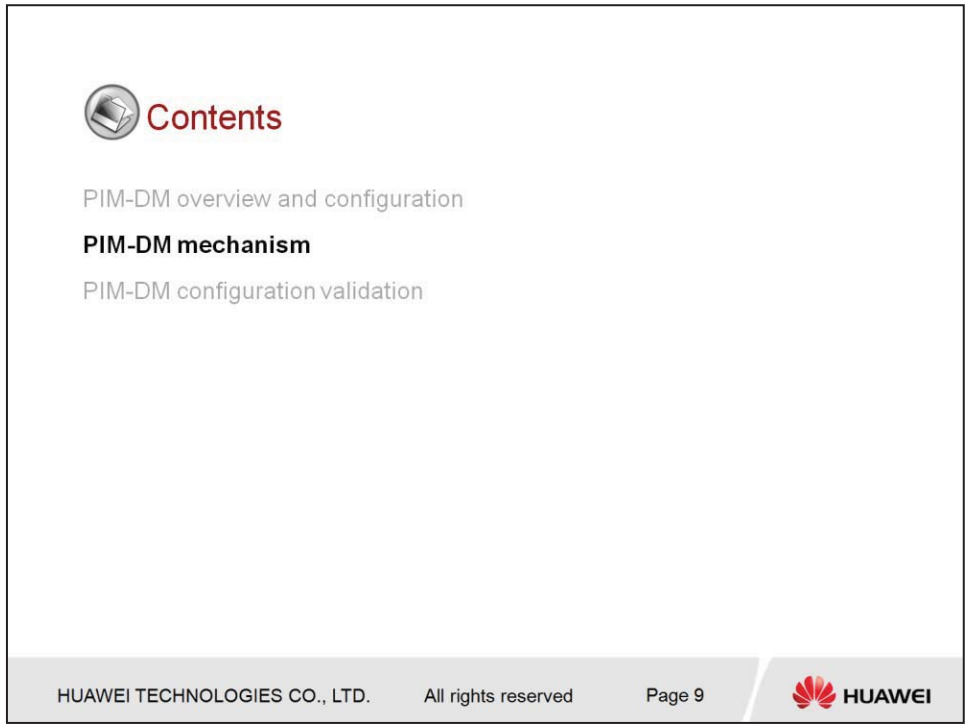
These packets are used to periodically set up and maintain the SPT. Their functions will be described in the part of PIM-DM working mechanism.

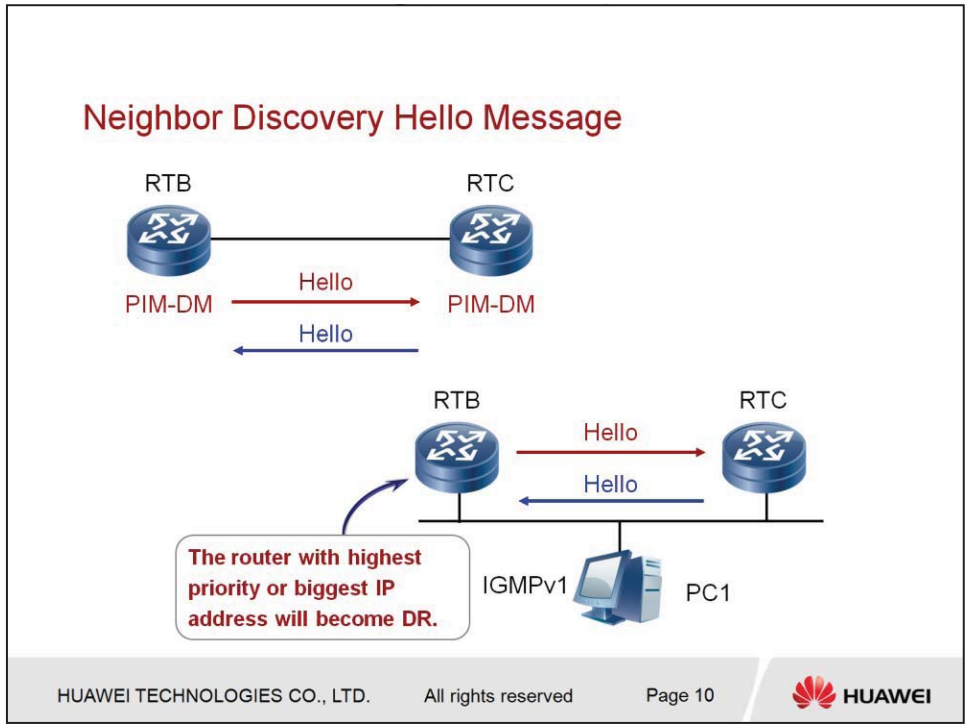


The configuration of PIM-DM is very simple. You need to enable PIM-DM on the interfaces of a router.

Before enabling PIM-DM on an interface, you must enable IP multicast routing in the system-view of router.

The command used to enable IP multicast routing is multicast routing-enable.


Contents
PIM-DM overview and configuration
PIM-DM mechanism
PIM-DM configuration validation
HUAWEI TECHNOLOGIES CO., LTD.
All rights reserved
Page 9
HUAWEI



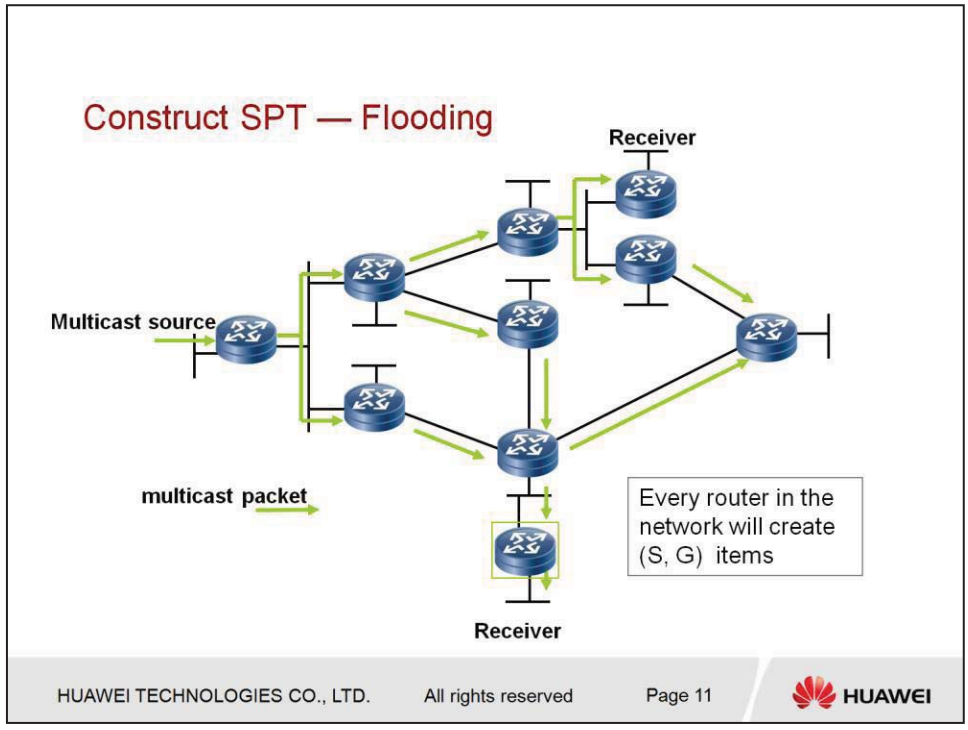
In a PIM-DM network, a router sends Hello messages upon startup to discover neighbors and maintains neighbor relation. Routers periodically send Hello messages, and thus a SPT is established and maintained. Using the pim timer hello interval command, you can set the Hello interval in the interface view. The default Hello interval is 30 seconds.

Another function of the Hello message is electing the designated router (DR) in the network segment containing multiple routers. The DR acts as the IGMPv1 querier. In the IGMP section, we have mentioned that the IGMPv1 querier is elected by the DIM-DM message.

PIM-DM compares the priority values and IP addresses contains in Hello messages to elect the DR as the IGMPv1 querier.

When the DR is down, and its neighbor cannot receive the Hello message from DR within the timeout duration, its neighbors will trigger the election of a new DR.

Using the pim hello-option holdtime interval command, you can set the Hello timeout duration. The default timeout duration is 105 seconds.

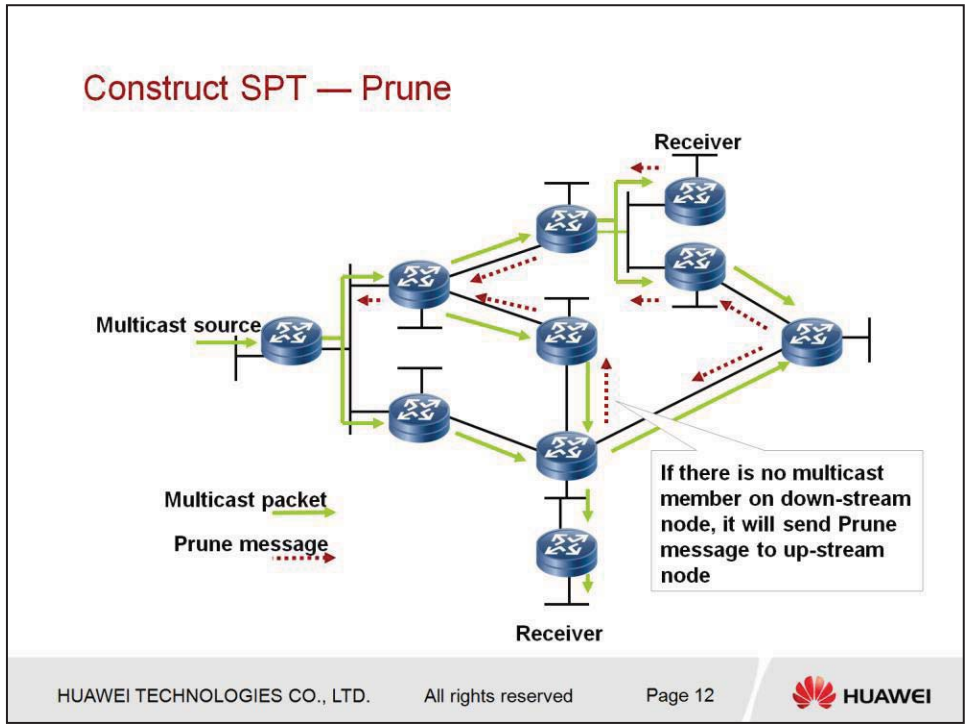


PIM-DM is based on the assumption that all hosts in the network will receive the multicast packet. When the multicast source (S) sends a multicast packet to a multicast group (G), each router that receives the packet performs reverse path forwarding (RPF) check according to the routing table.

•If the packet passes the RPF check, the router creates an (S, G) entry, and then forwards the packet to all downstream PIM-DM nodes in the network. This process is called flooding.

•If the packet does not pass the RPF check, it indicates that the packet is received on a wrong interface. The router discards this packet.

After this process, each router in the PIM-DM multicast domain creates an (S, G) entry. S means the IP address of multicast source. G means the multicast IP address.



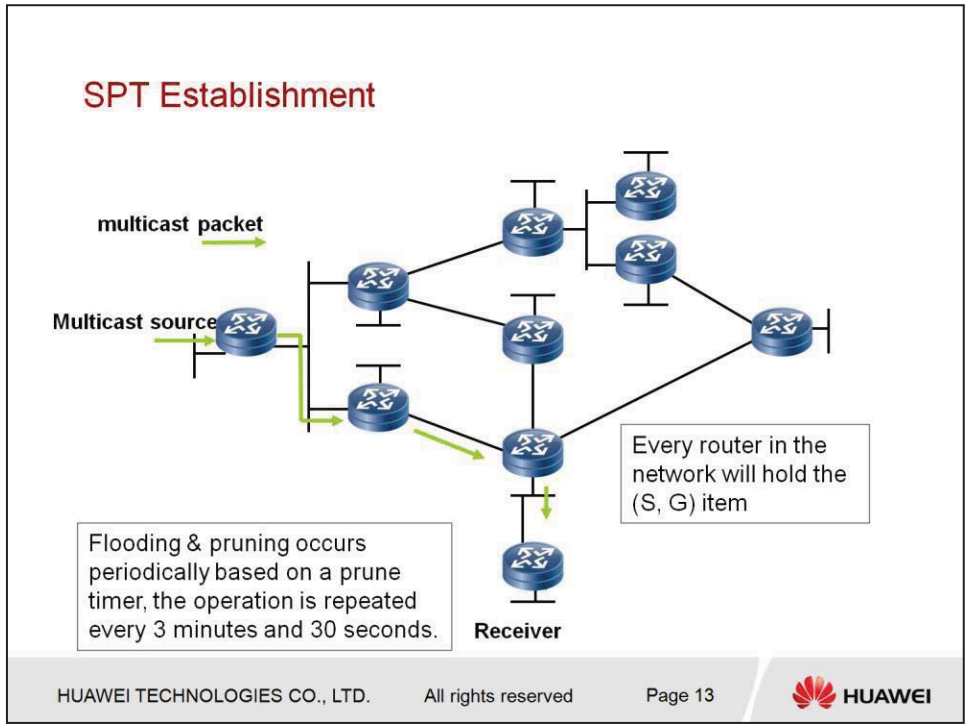
A router floods multicast packets regardless whether there is any exsiting downstream multicast members so bandwidth is wasted. Prune mechanism of PIM-DM can avoid waste of bandwidth.

If edge router does not contain any multicast receiver, the router sends the Prune message to its upstream node. When the upstream node receives the Prune message, it updates its multicast routing table .The pruning process lasts until the

router whose branches need to forward multicast packets. Thus a shortest path tree (SPT) with the multicast source S as root is established.

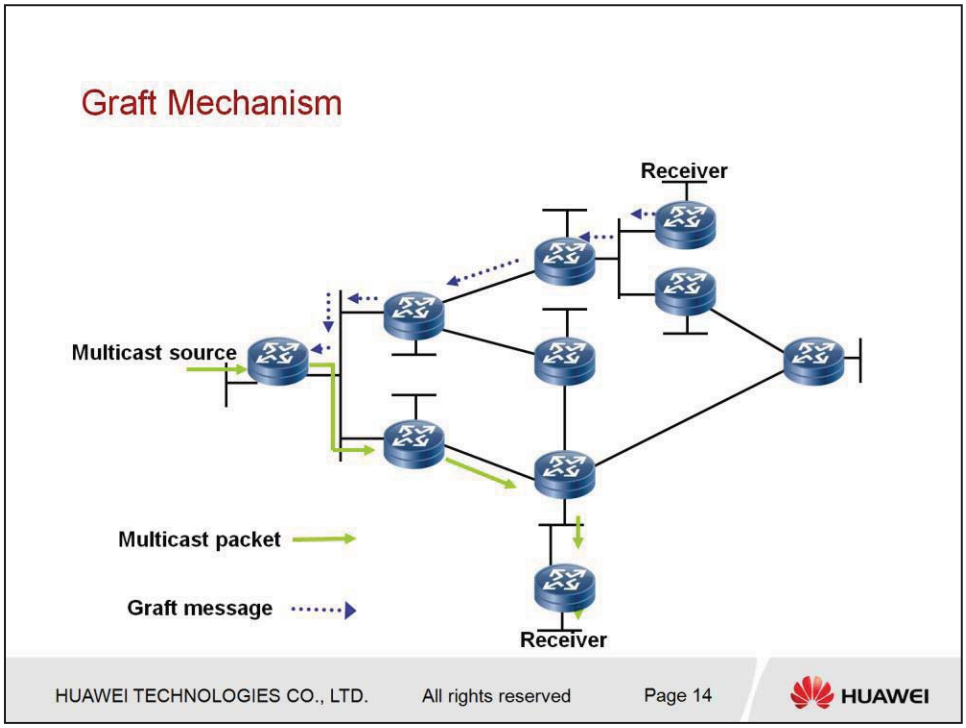
A timeout mechanism is provided for the pruned nodes. After the timeout duration, the flood-prune process starts again. The default prune timeout duration is 210 seconds.

The flood-prune process of PIM-DM occurs periodically.



Multicast routers. prune redundant branches according to Prune messages to establish a new SPT. Although routers stop forwarding multicast packets to the branch that contains no members after they receive the Prune message, each

router still keeps the (S, G) entry of this branch. Once a member on this branch joins the multicast group, forwarding on this branch resumes quickly

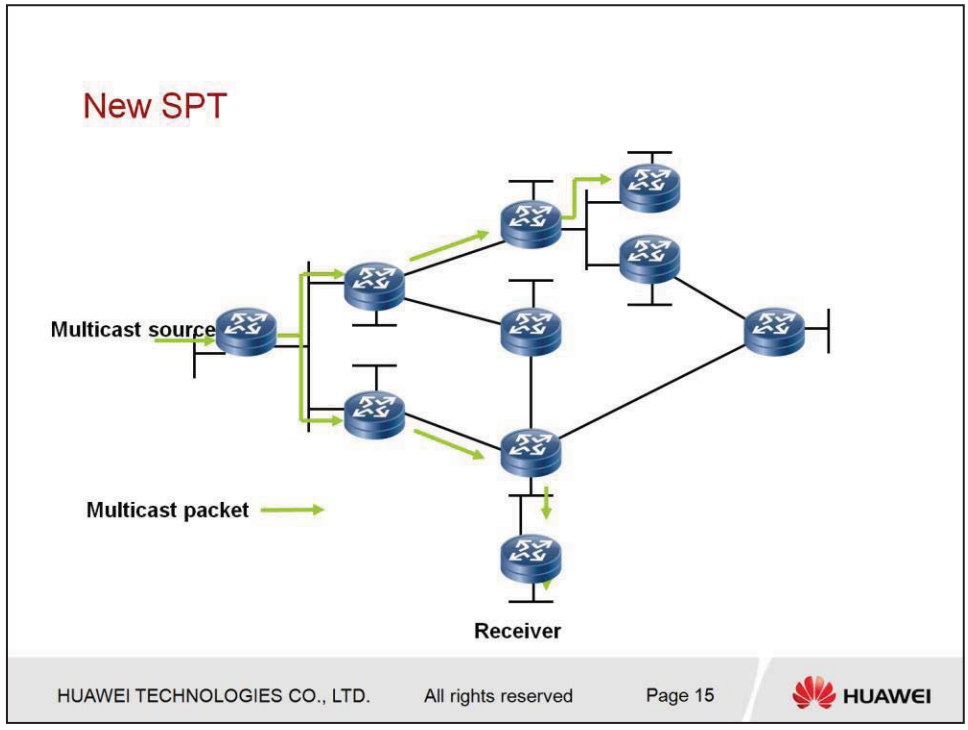


The pruned downstream nodes restore to forwarding state when the prune timer expires. The default prune timer is 210 seconds. That means a node must wait such a long time to join the group. To shorten the waiting time, the node sends

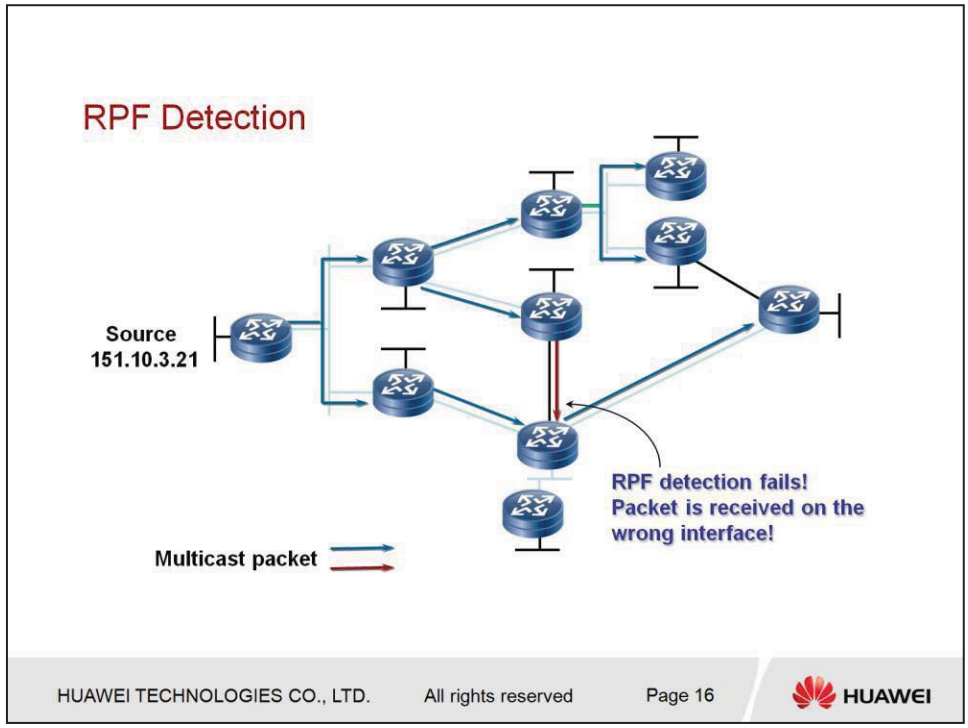
Graft message to notify the upstream node when it needs to turn to forwarding state.

When a receiver in the network starts to receive multicast packets again, the Graft message is transmitted to multicast source S hop by hop. When the intermediate nodes receive the Graft message, they respond with acknowledgement messages.

In this way, the pruned branch resumes packet transmission.



A pruned path turns to a forwarding path, and thus a new SPT is established.

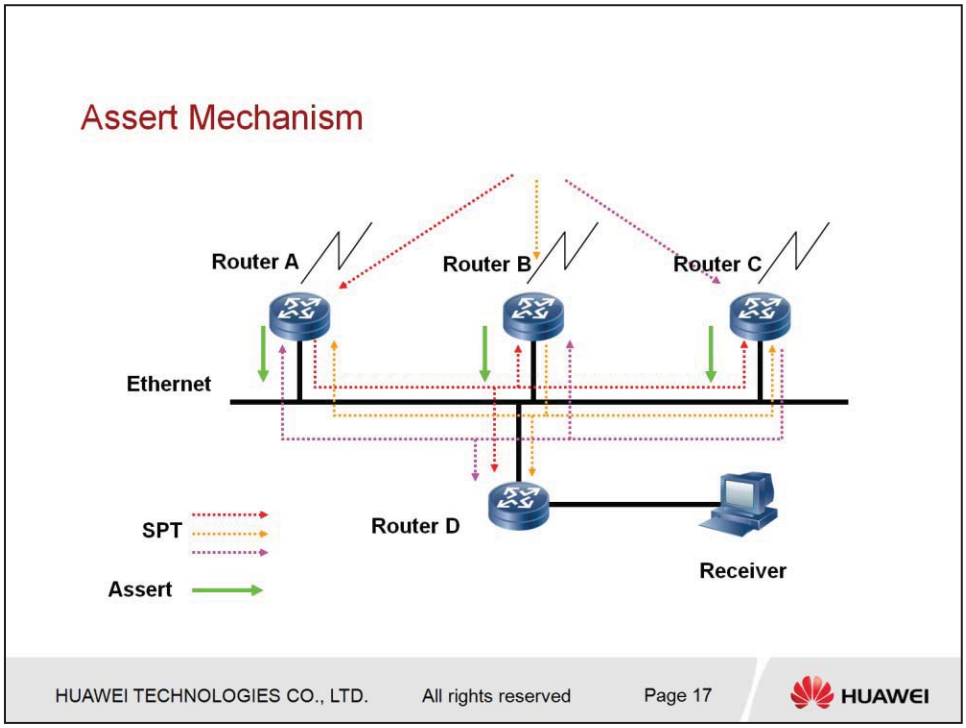


PIM-DM adopts the RPF check mechanism. This mechanism helps to establish a multicast forwarding tree rooted from the source based on the existing unicast routing table, multicast static routing table, MBGP routing table.

When receiving a multicast packet, a router checks whether the forwarding path of this packet is correct. If the interface receiving this packet is the interface responsible for sending the packet to the multicast source, as specified by the unicast routing table, the router considers that the packet is transmitted through

the correct path. Otherwise, the router discards the packet.

As a basis for checking the correctness of the path, routing information can be generated by any unicast routing protocol, for example, RIP and OSPF. The type of the routing protocol is not limited.



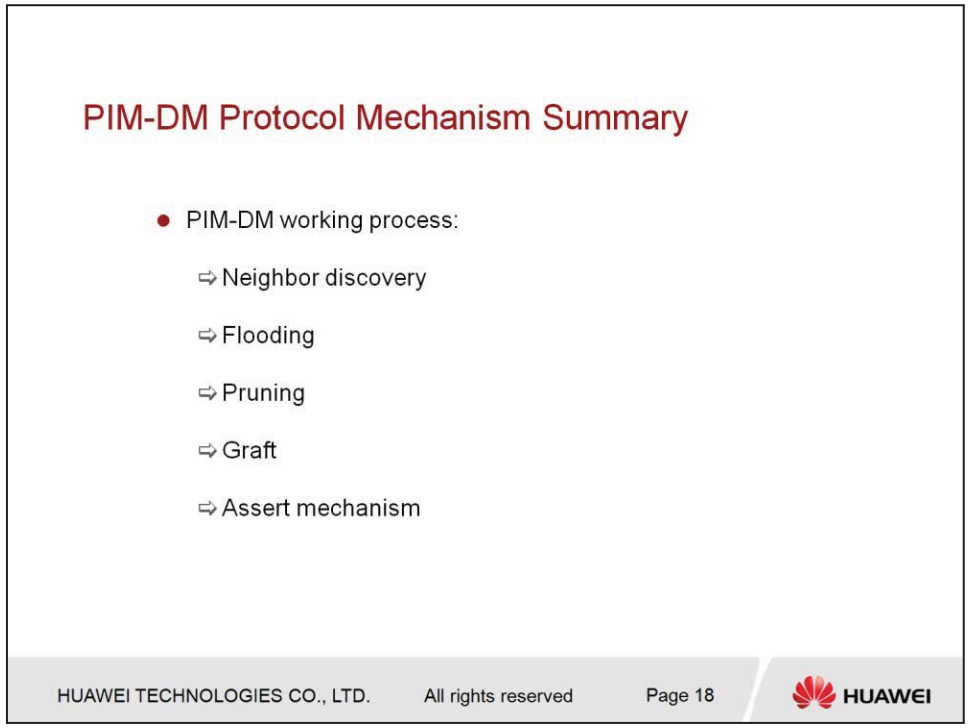
In a shared network, for example, Ethernet, a packet may be forwarded repeatedly. As shown in the figure, a LAN consists of routers A, B, C and D. Each router has a path to the multicast source. When routers A, B, and C receive a multicast packet from the upstream node, they all forward the packet to the Ethernet. The downstream node, namely router D, receives three identical multicast packets.

The Assert mechanism is adopted to avoid this problem. This mechanism elects a unique forwarder for the multicast packet. Routers in the network send Assert packets to select the best path.

The best path is elected as follows:

If two or more paths have the same priority and same cost to the multicast source, the router with the highest IP address becomes the upstream neighbor in the (S, G) entry. This router forwards the multicast packets related to the (S, G) entry.

Other routers prune the corresponding interface to forbid the interfaces to forwarding the packet.



The working process of PIM-DM consists of neighbor discovery, flooding, pruning, graft and Assert mechanism.

Neighbor discovery

Multicast routers send Hello messages to discover neighbors and maintain neighbor relations. Routers compare the priorities and IP addresses in Hello messages to elect the DR in the network segment. The DR acts as the IGMPv1 querier.

Flooding

Multicast source S sends multicast packets to multicast group G. when a router receives a multicast packet, it performs RPF check based on the unicast routing table. If the packet passes the RPF check, the router creates an (S, G) entry, and then forward the packet to all downstream PIM-DM nodes in the network. This process is flooding. If the packet does not pass the RPF check, the router discards the packet.

Pruning

If no downstream node contains the multicast member, the router sends the Prune message to the upstream node to request the upstream node stop for warding packets to this branch. After the upstream node receives the Prune message, it deletes the

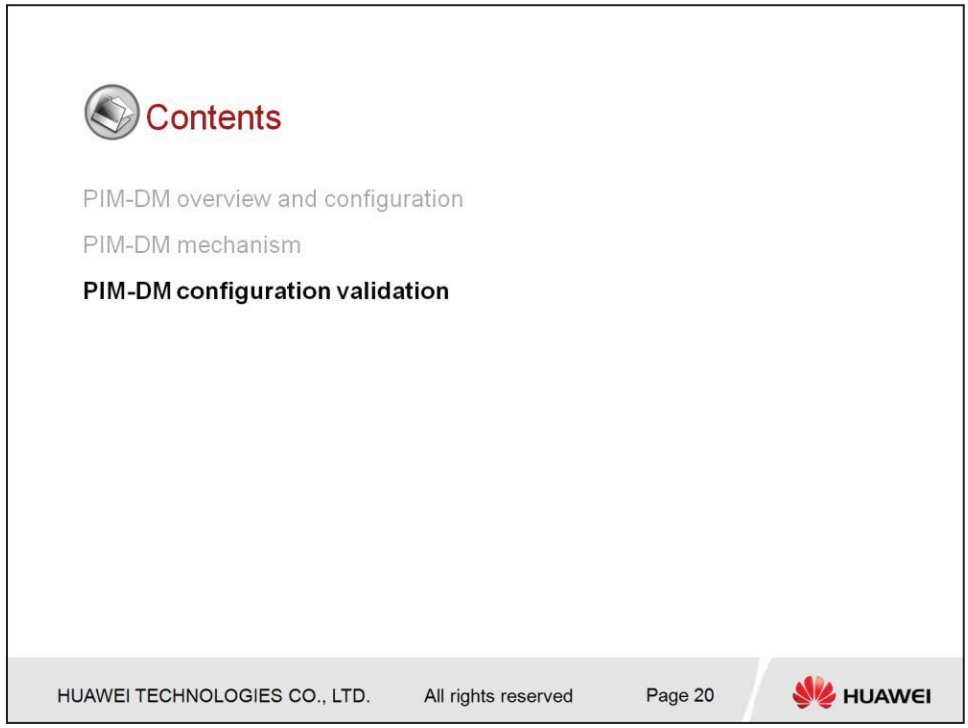
corresponding interface from the interface list mapping the (S, G) entry. The pruning process lasts until only the branches that need to forward data remain. Thus a SPT with multicast source S as the root is established.

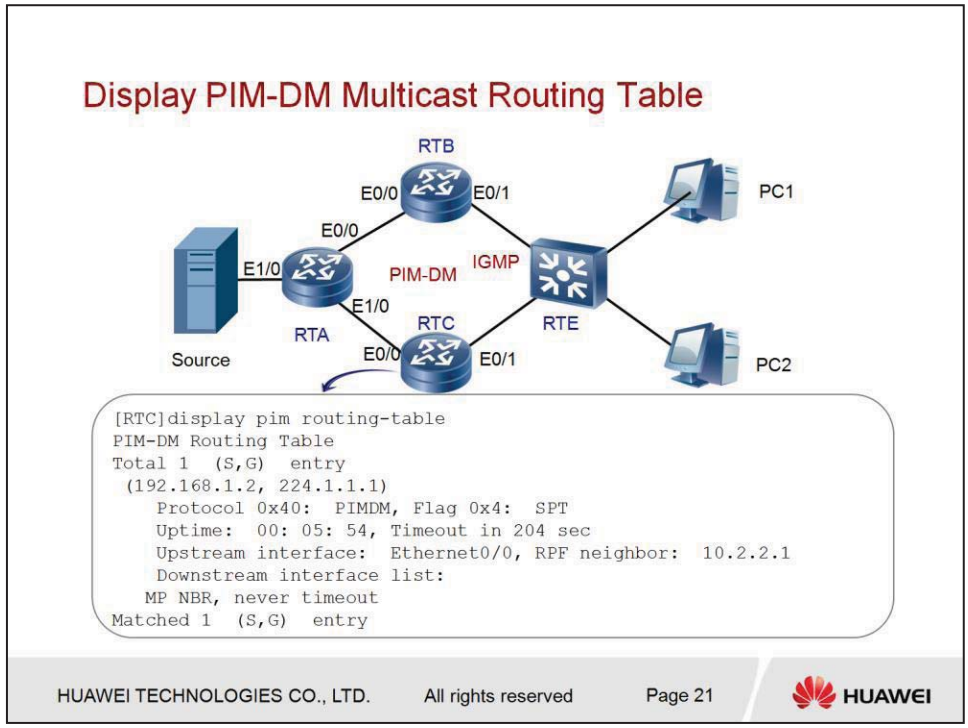
Graft

When a pruned node needs to turn to forwarding state, the node sends the Graft message to request the upstream node to restore packet forwarding.

Assert mechanism

The Assert mechanism is used to specify the forwarder in a shared network.

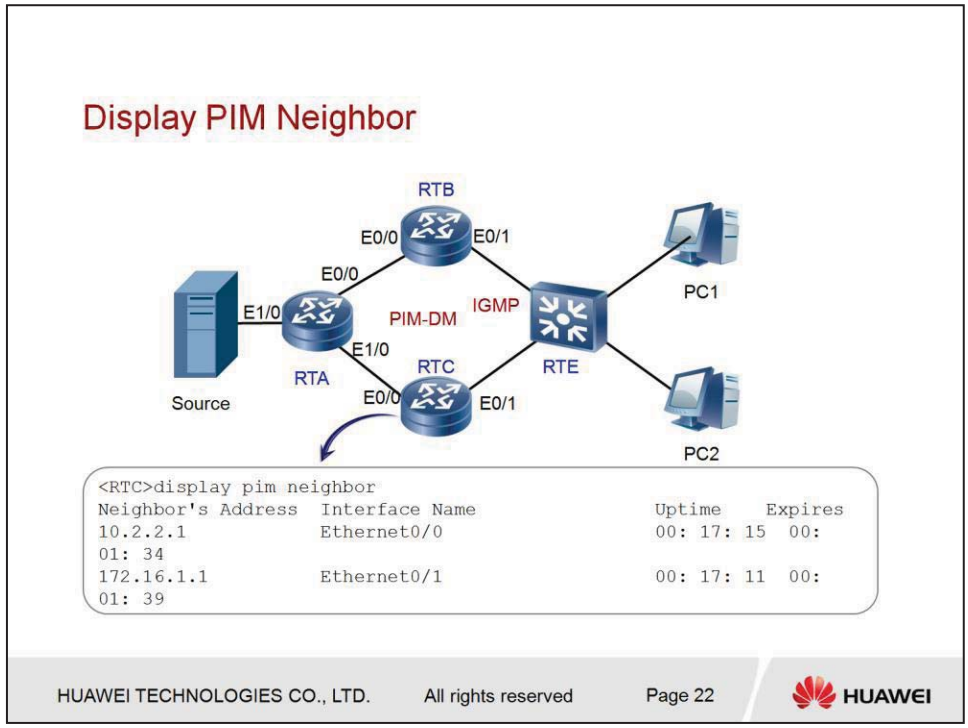

Contents
PIM-DM overview and configuration
PIM-DM mechanism
PIM-DM configuration validation
HUAWEI TECHNOLOGIES CO., LTD.
All rights reserved
Page 20
HUAWEI



Using the display pim routing-table command, you can view the PIM multicast routing table of a router.

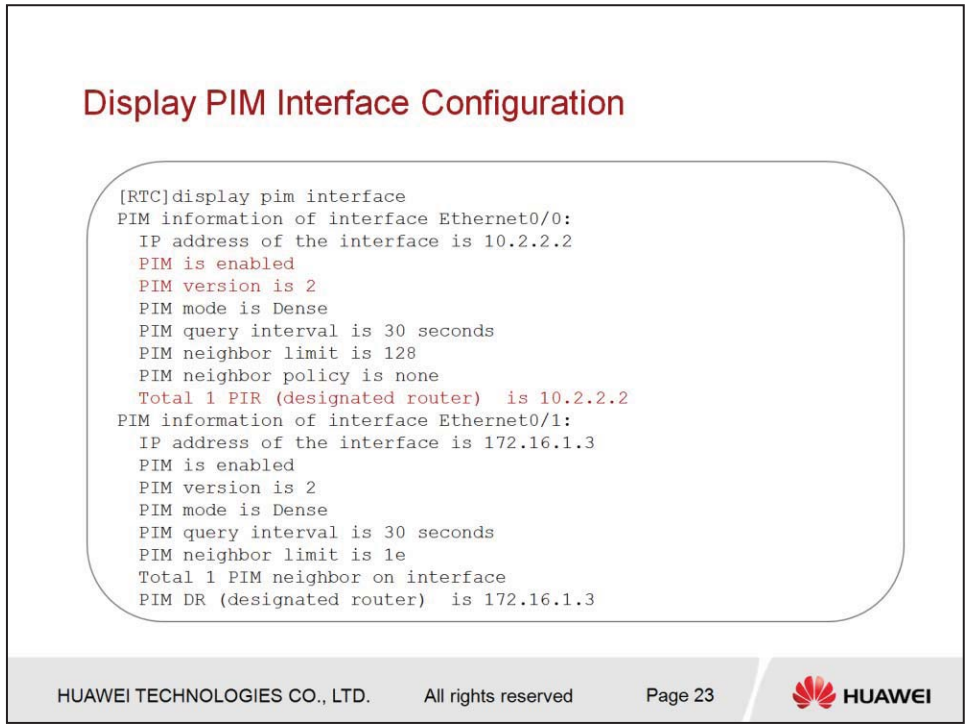
After receiving a multicast packet, a router creates a (S, G) entry. The (S, G) entry lists information about the nodes that this packet passes, namely the upstream node, downstream node and the RPF neighbor relation. The RPF neighbor relation is used to check whether the packet is received by the correct interface.

If the routing table of a router is empty, you need to check the configuration.

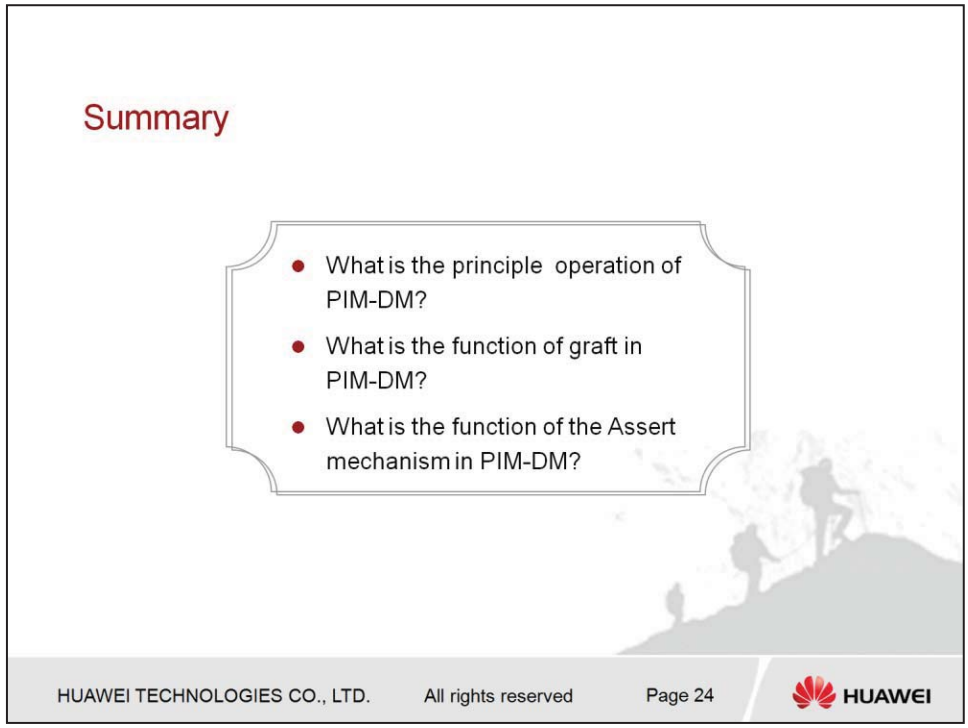


Using the display pim neighbor command, you can view the information about PIM neighbors of a router.

In general, RTC has two neighbors: RTA (10.2.2.1) and RTB (172.16.1.1).



Using the display pim interface command, you can view the PIM information on an interface, including whether the PIM protocol is enabled, the PIM mode and the IP address of the DR.



1. What is the principle of PIM-DM?

PIM-DM is based on the assumption that every subnet in the network has at least one receiver of the multicast source. In this case, multicast packets are flooded to all nodes in the network.

PIM-DM prunes the branches that do not need to forward multicast packets and only keeps the branches that contain the receivers. The pruned branches need by graft mechanism to restore forwarding of multicast packets on the pruned branches.

2. What is the function of graft in PIM-DM?

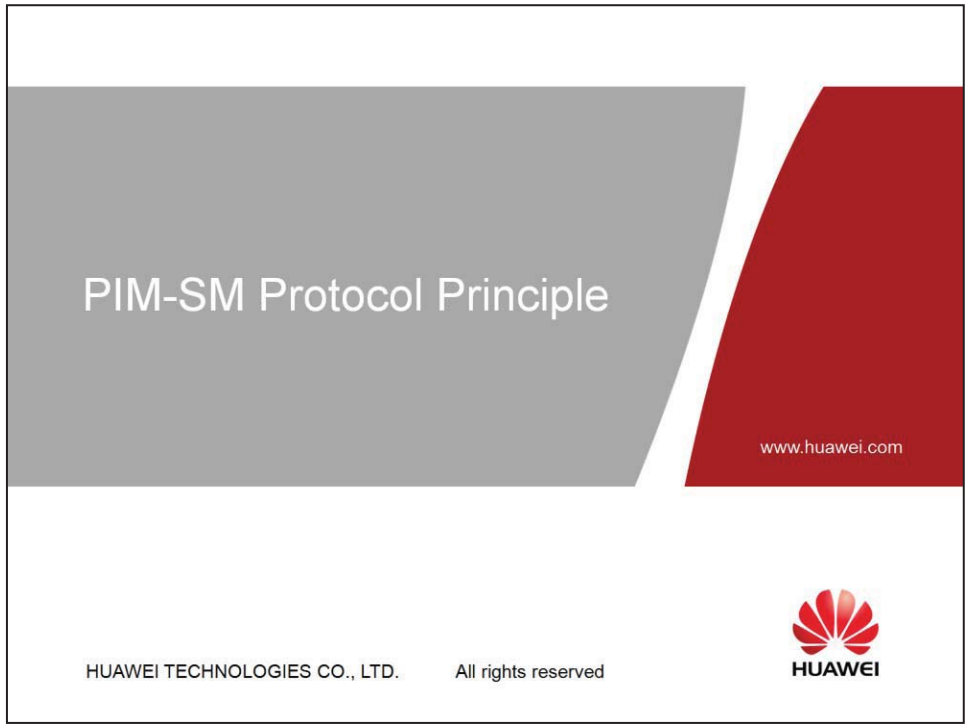
Routers use the Prune message to prune the branches that do not contain multicast members. If a node on the pruned branch needs to join the multicast group, the node needs to wait until the prune timer expires. When the downstream node needs to turn to forwarding state, the node can use the graft mechanism to notify the upstream node that a member joins the group. Thus the response period is shortened.

When a receiver in the network starts to receive multicast packets again, it sends a Graft message. The Graft message is forwarded to multicast source hop by hop. When the intermediate nodes receive the Graft message , they respond

With acknowledgement messages. In this way, the pruned branch resumes packet transmission.

3. What is the function of the Assert mechanism in PIM-DM?

Assert mechanism avoids repeated forwarding of same packets in a shared network, for example, Ethernet. Assert mechanism elects a unique forwarder for the same packet in the shared network. Other routers prune corresponding interfaces from (s,6) entry.


PIM-SM Protocol Principle
www.huawei.com
HUAWEI TECHNOLOGIES CO., LTD.
All rights reserved
HUAWEI


Foreword
PIM-SM (Protocol Independent Multicast-Sparse Mode) belong to the sparse mode of multicast routing protocol, and is suitable for the relatively scattered distribution group. It builds unidirectional shared trees rooted at a Rendezvous Point (RP) for each group, and optionally creates shortest-path trees for each source.
HUAWEI TECHNOLOGIES CO., LTD.
All rights reserved
Page 2
HUAWEI

Objectives

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

- Master PIM-SM basic principles and configuration
- Master the join and source register process of the shared tree
- Master RPT switch to SPT

HUAWEI TECHNOLOGIES CO., LTD. All rights reserved

Contents

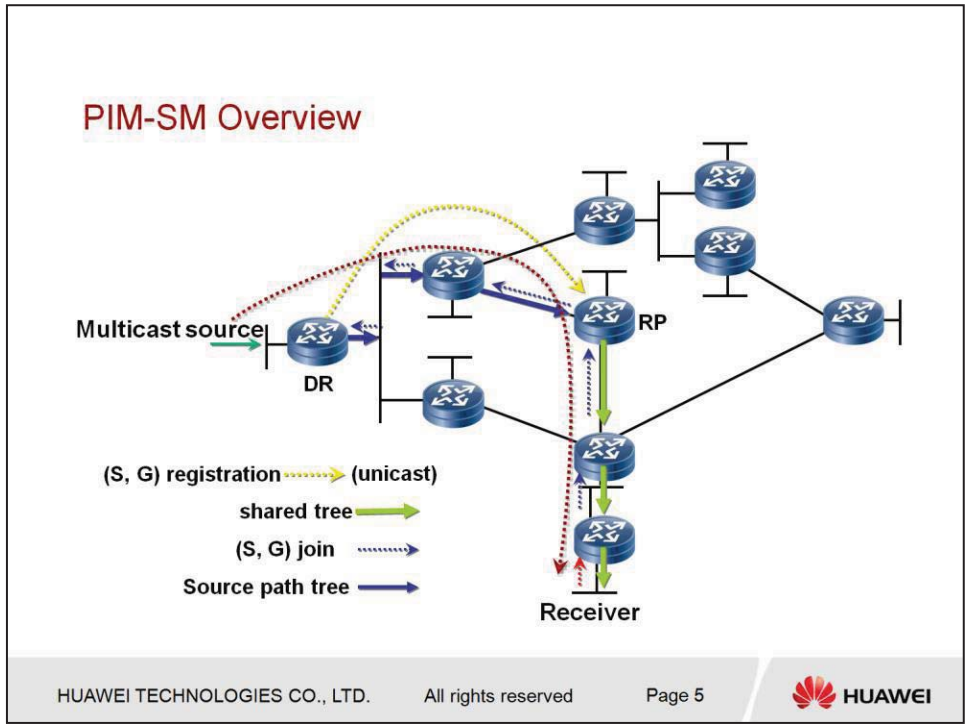
PIM-SM overview and basic configuration

PIM-SM mechanism

PIM-SM advanced configuration

HUAWEI TECHNOLOGIES CO., LTD. All rights reserved



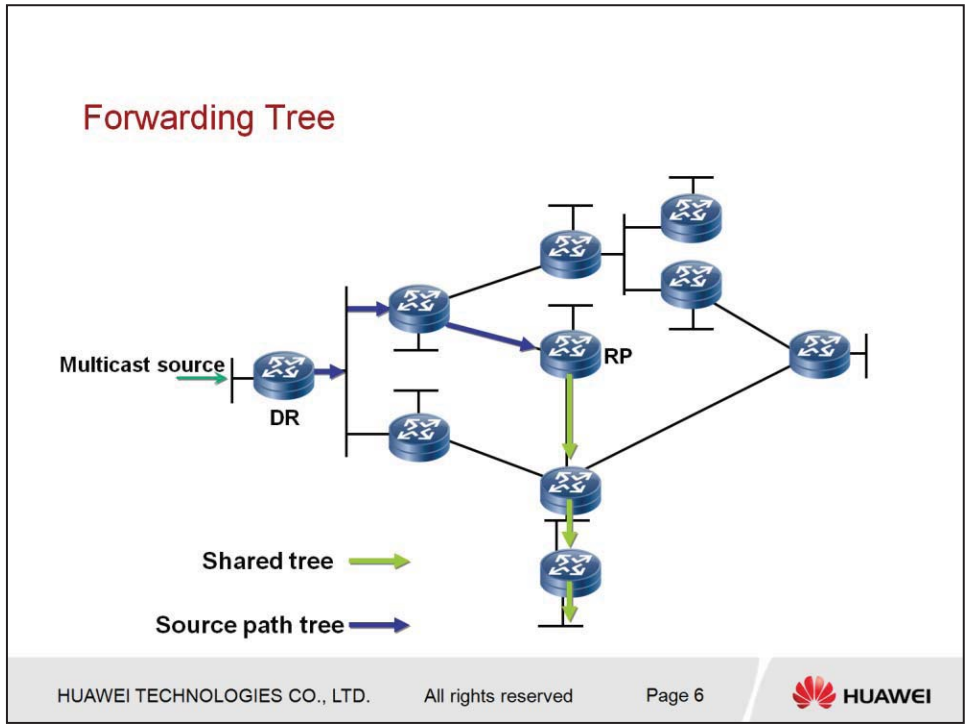


PIM-SM is based on the assumption that the multicast members locate sparsely. That is, near no network segment has multicast group members. Multicast route is established only when a host in a network segment joins the multicast group. At this time, the upstream node begins to forward multicast packets to this network segment. The goal of the PIM-SM model is to establish and maintain a unidirectional shared Tree.A router in the PIM network is selected as the public root node of the shared tree.

The public root node is called the rendezvous point (RP). The RP forwards multicast packets to receivers along the shared tree.

At the receiver's side, the router connected to the receiver sends the Join message to the RP of the multicast group. The Join message is forwarded hop by hop and reaches the root (RP). The path that the Join message passes by becomes a branch of the shared tree, namely, the rendezvous point tree (RPT).

When the sender needs to send multicast data to a multicast group, the first-hop router sends Register message to the RP. The Register message reaches the RP and triggers stablishment of the RPT. The multicast source then sends the multicast packet to the RP. The multicast packet is duplicated on the RP and forwarded to the receivers along the RPT.



PIM-SM uses two types of trees: the shared tree and source path tree.

The path from the RP to the multicast receiver is the shared tree.

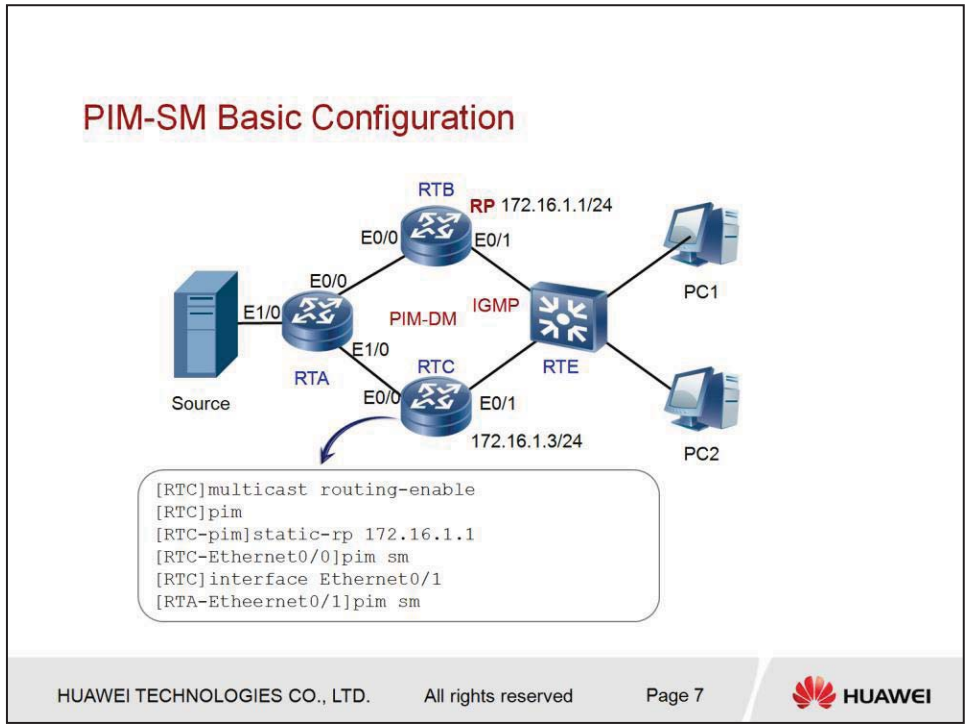
The path from the multicast source to the RP is the source path tree.

RPF check is performed according to the type of the tree.

On the shared tree, the RP address is used as the source address for RPF check.

On the source path tree, the address of the multicast source is used as the source address for RPF check.

The RPF check process will be described in Chapter 2 PIM-SM Protocol Mechanism.



The basic configuration of PIM-SM is similar to the configuration of PIM-DM. for PIM-SM, you also need to enable the multicast routing protocol, and then enable PIM-SM on the interfaces of the router.

After the brief introduction of the PIM-SM principle, you have known that RP is an important role in the PIM-SM network. As the root of RPT, the RP forwards multicast packets and controls join of multicast groups. The RP can be discovered in various ways. Discovery and election of the RP will be described in Chapter 2. Here we manually specify RTB (172.16.1.1) as the RP to show the configuration of PIM-SM.

PIM-SM configuration involves the following steps:

1. Globally enable the multicast routing protocol.

2. Enable PIM-SM on the interfaces.

3. Specify the address of RP.

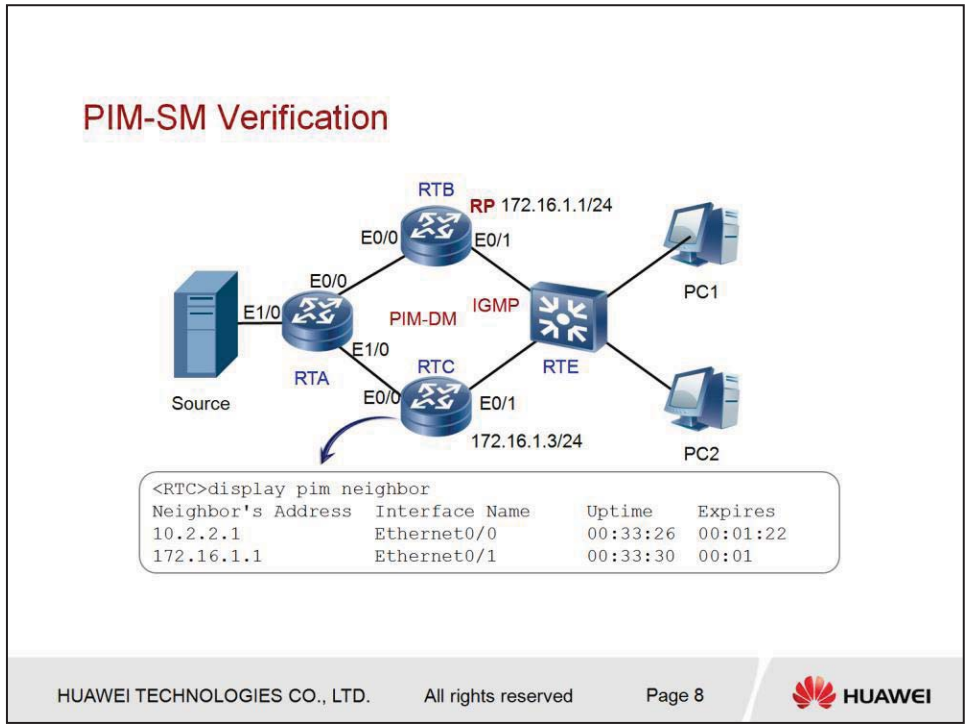
Note: If the RP is manually specified, the RP address must be configured on each router. The procedure is as follows:

Enter the PIM view.

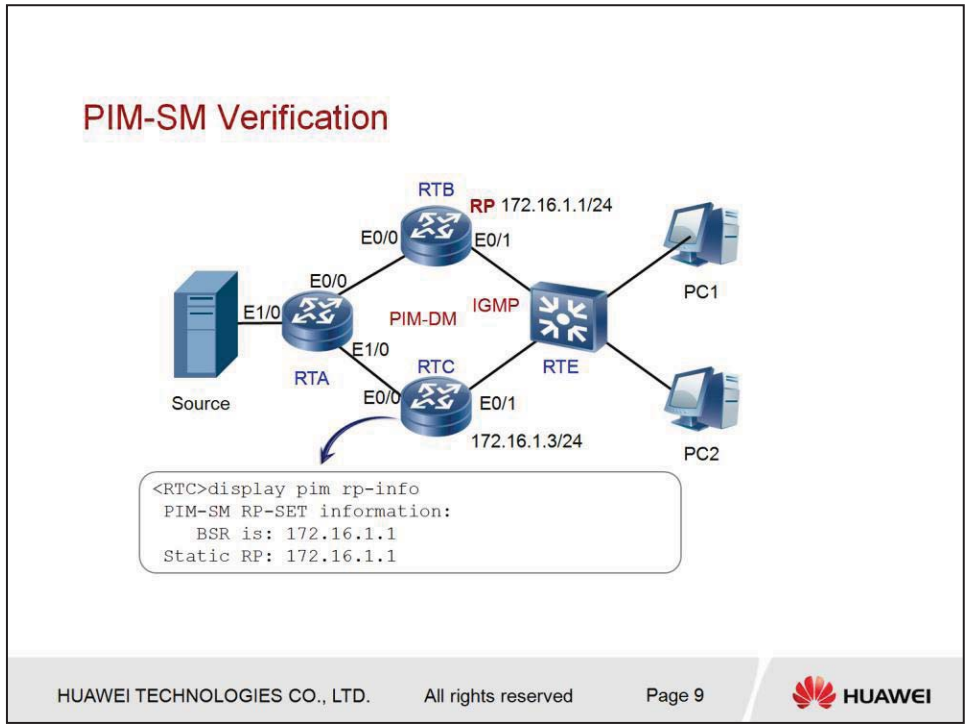
[RTC] pim

Specify 172.16.1.1 as the RP address.

[RTC-pim] static-rp 172.16.1.1

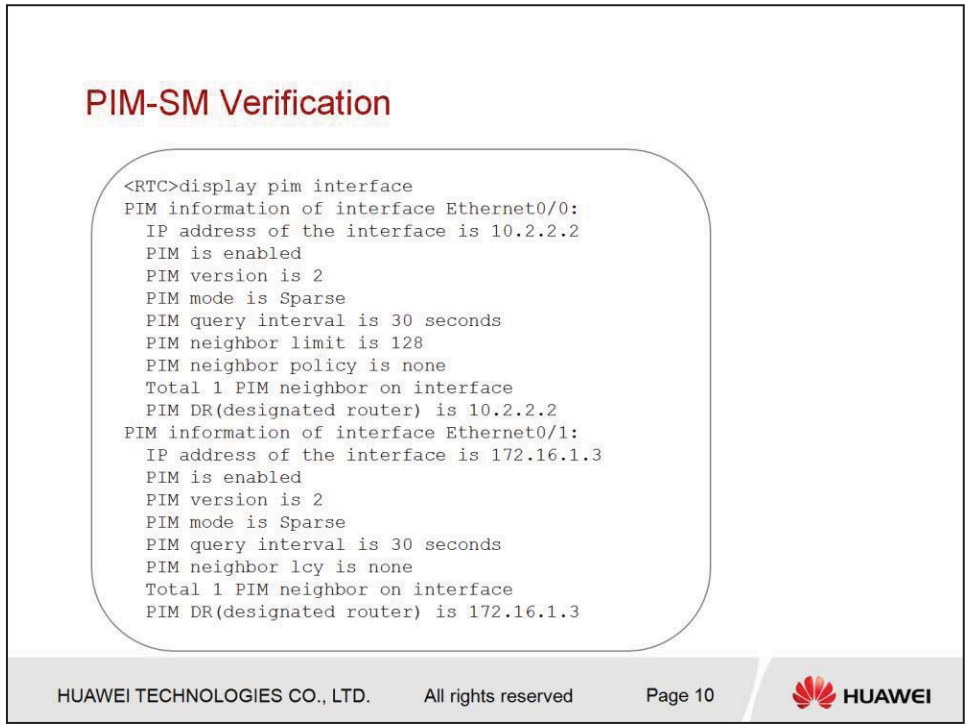


Using the display pim neighbor command, you can view the information about PIM-SM neighbors of the router. After the configuration, RTC has two PIM neighbors: RTA and RTB.

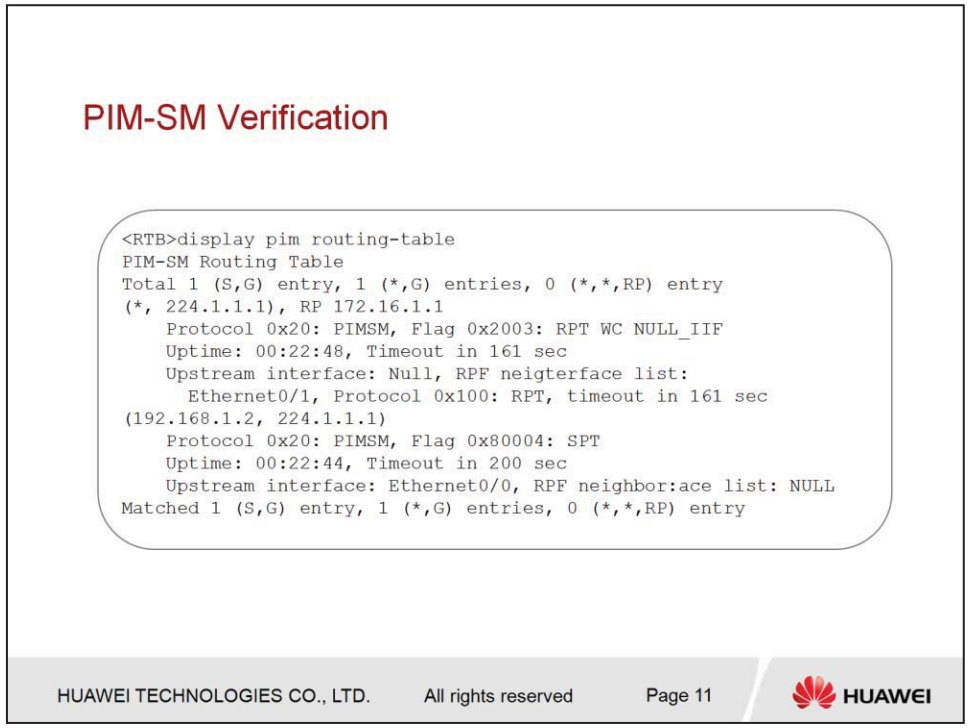


View the information about the RP in the PIM-SM network.

In this example, the RP is manually specified. The "Static" field before "RP" in the display information indicates that the RP is manually specified.



Using the display pim interface command, you can view the information about a PIM interface. The display information includes whether PIM is enabled, PIM version, PIM mode, PIM query interval and IP address of the DR.



Using the display pim routing-table command on RTB, you can see two entries:

an (S, G) entry and a (*, G) entry. It indicates that RTB has received the multicast packet and can forward the packet.

As you know, the (S, G) entry is recorded in the routing table. S indicates the multicast source and G indicates the multicast group. The routers on the source path tree (SPT) from the multicast source to RP create such entries.

(*, G) is the entry for only PIM-SM. This entry is the entry of (any multicast source, multicast group). The routers on the RPT create such entry.

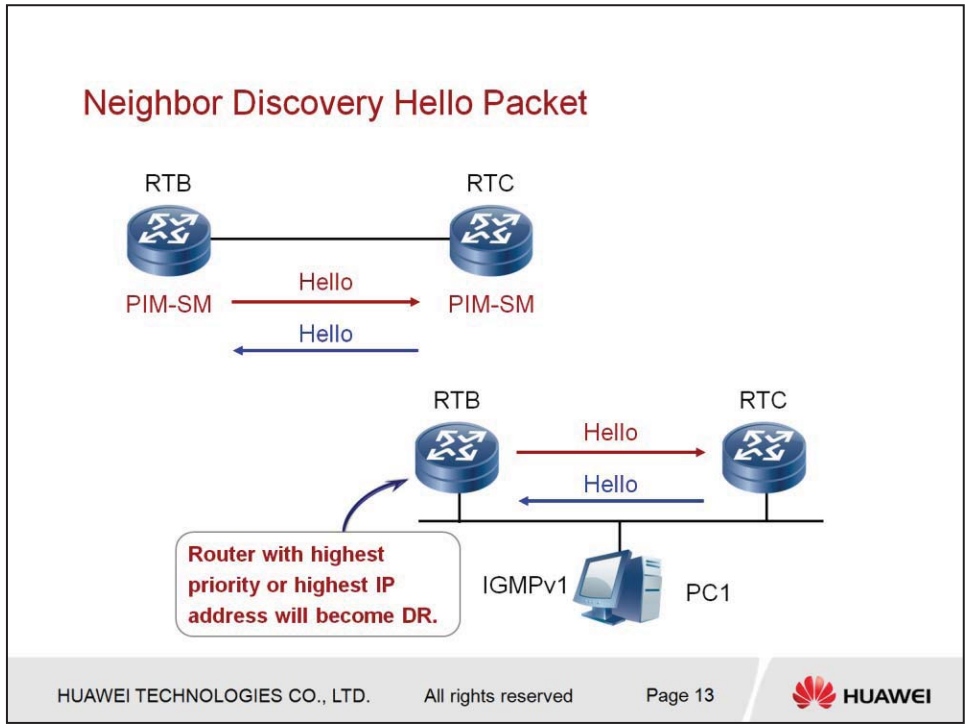
Contents

PIM-SM overview and basic configuration

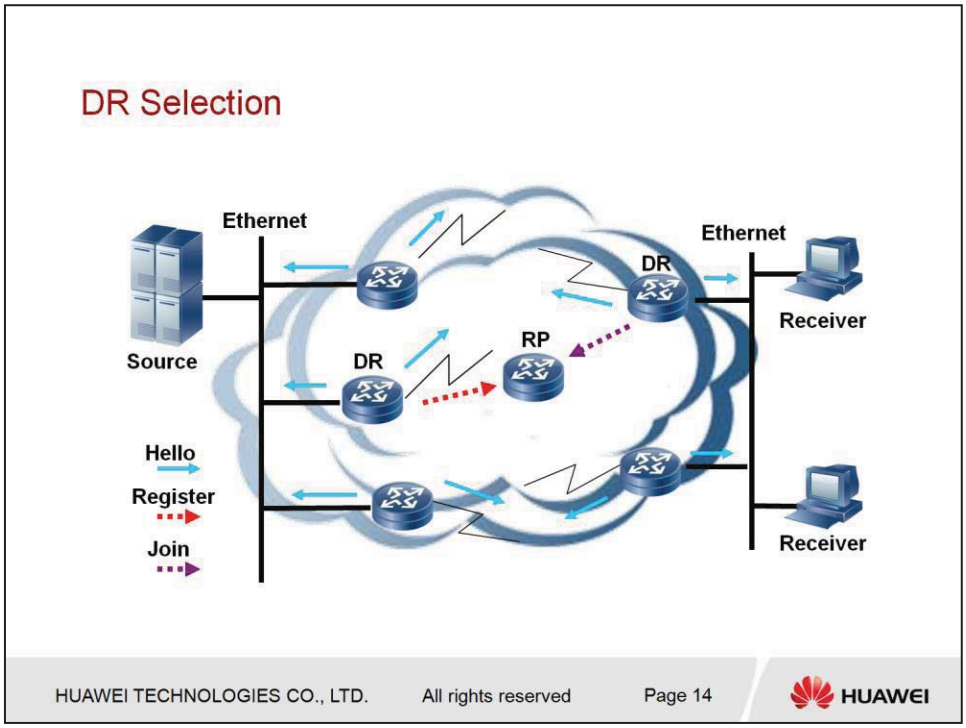
PIM-SM mechanism

PIM-SM advanced configuration

HUAWEI TECHNOLOGIES CO., LTD. All rights reserved



In a PIM-SM network, a router needs to send Hello messages to discover its neighbors upon startup. Routers periodically send Hello messages to maintain the neighbor relation. Another function of the Hello message is electing the DR in a network segment containing multiple routers. The DR acts as the IGMPv1 querier.



PIM-SM also elects the DR in a shared network (such as Ethernet). The DR is the only multicast packet forwarder in the network segment.

In PIM-SM, all shared networks must have a DR, no matter whether the network is connected to the multicast source or the receiver.

The DR at the receiver end sends the Join message to the RP.

The DR at the multicast source side sends the Register message to the RP.

Routers on the shared network send Hello messages (containing the DR priority option) to each other. The router with the highest DR priority is elected as the DR in the network segment.

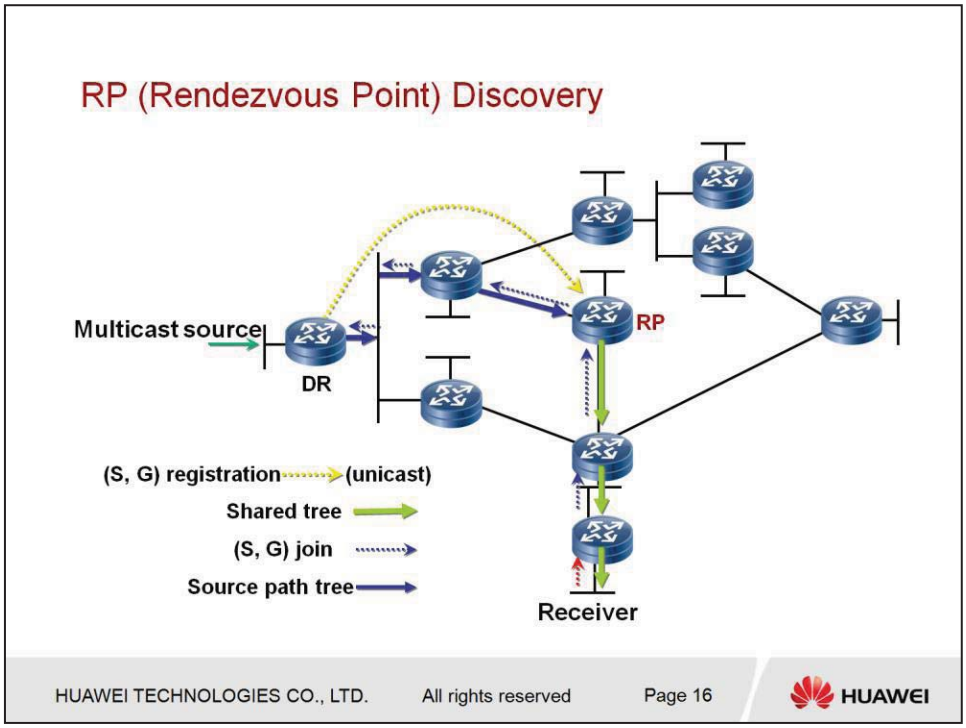
If some routers have the same priority or at least one router in the network cannot add the priority option in the Hello message, the router with the largest IP address is elected as the DR.

When the DR is Down, it cannot respond to the Hello message sent by other router within the timeout duration. The neighbor routers then trigger the election of a new DR.

Using the pim timer hello interval command, you can set the interval of Hello messages in the interface view. By default, the

Hello interval is 30 seconds.

Using the pim hello-option holdtime interval command, you can change the value of the Hello timeout duration. By default, the Hello timeout duration is 105 seconds.



In a PIM-SM network, the node acts as the root of the shared tree is called the rendezvous point (RP).

The RP has the following functions:

1. Forwards all multicast flows through the shared tree to receivers.

2. Forwards multicast packets of some or all multicast groups. A network can have one or more RPs.

The RP can be discovered in any of the following ways:

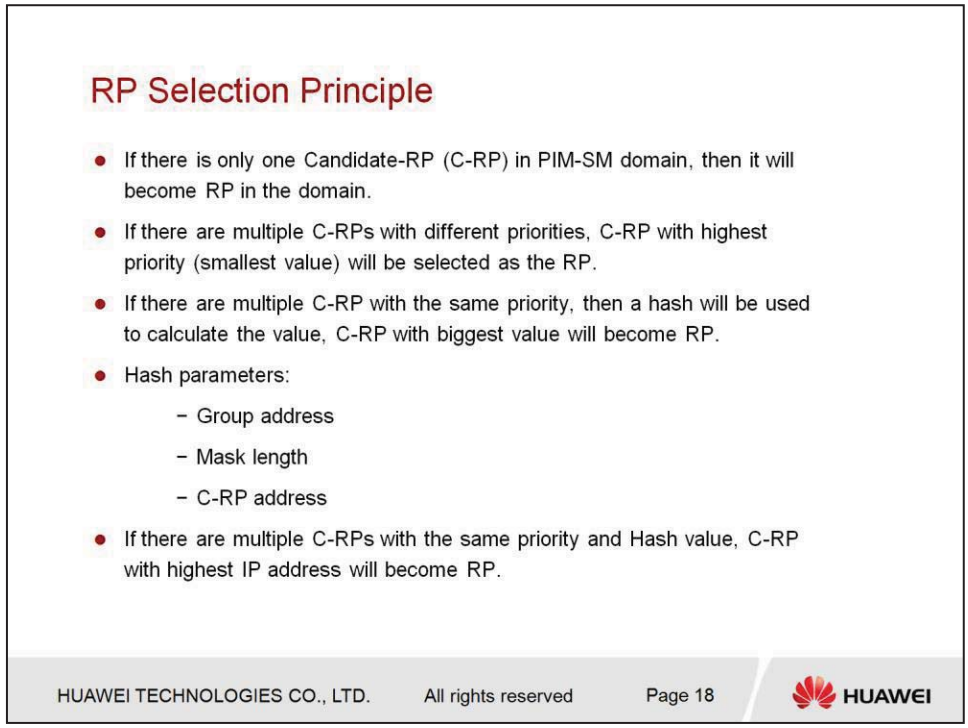
1. Manually specify the IP address of the RP on the DR, all leaf routers, and all routers that forward multicast flows.

Use the static-rp rp-address command to specify the IP address of the RP.

2. Enable the BootStrap protocol to elect the RP through the bootstrap mechanism.

The RP is the core router in the PIM-SM domain. Is a simple small-sized network, one RP is enough for packet forwarding, because the amount of multicast data is small. In this case, you can manually specify the RP on all routers in the SM domain. But in most cases, the PIM-SM network is very large and large amount of multicast packets forwarded by the RP. In such a

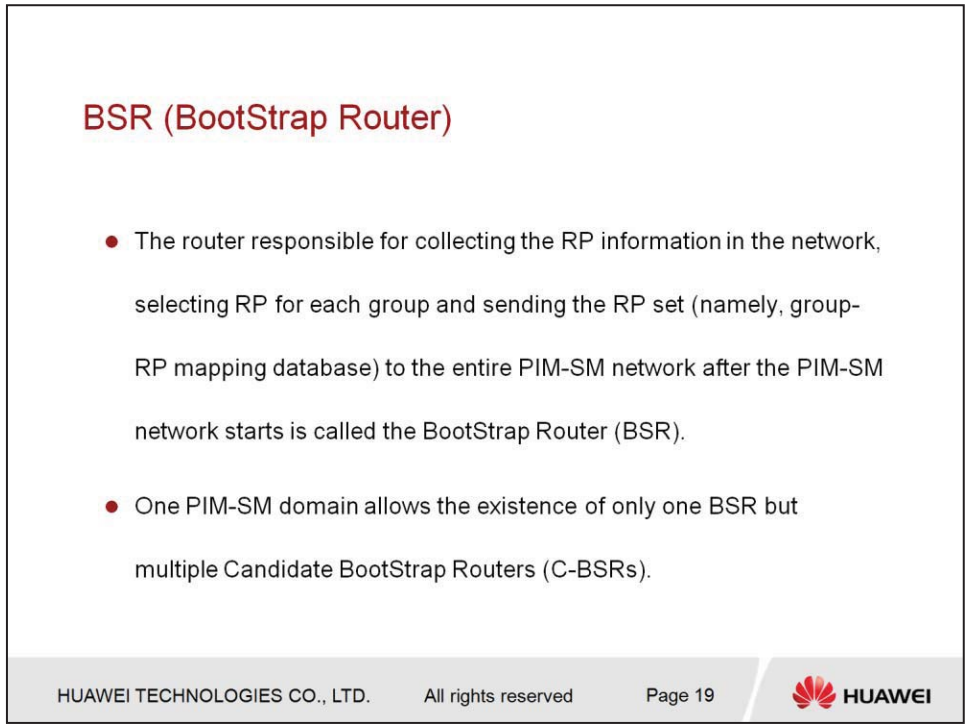
network, each multicast group needs an RP, thus the load on each RP is reduced and the topology of the shared tree is optimized. In such network, RPs are elected through the bootstrap mechanism. In this case, you need to configure the bootstrap router (BSR).



The RP can be specified manually or elected through the bootstrap mechanism.

In the latter case, you need to enable the BootStrap protocol.

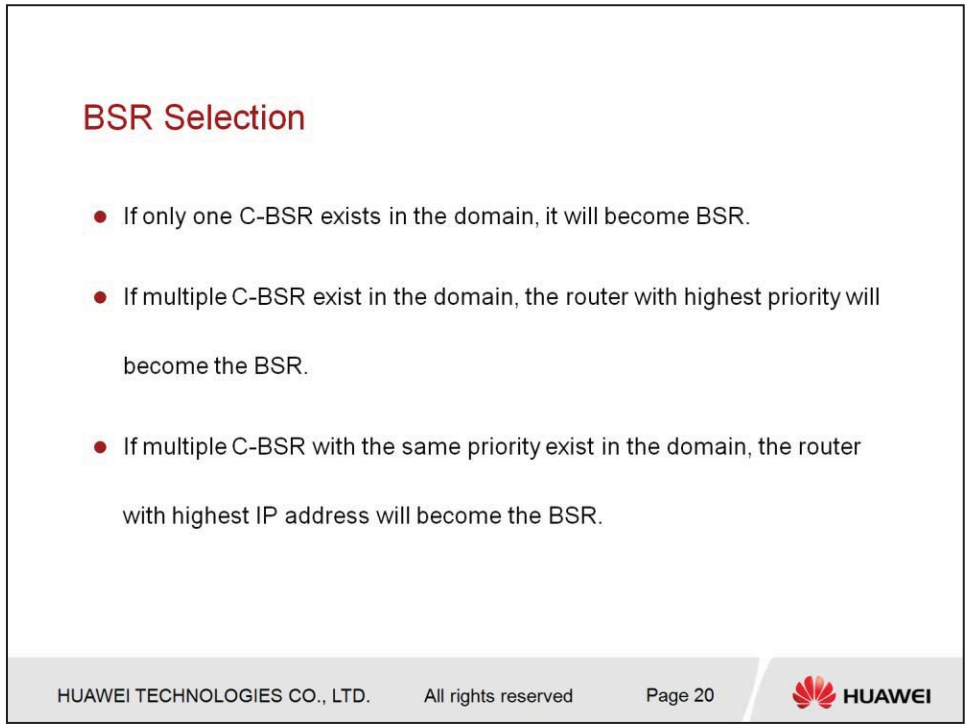
In a PIM-SM domain, all candidate RPs (C-RPs) take part in the RP election.



The BSR is the management core in the PIM-SM network. The BSR collects the Advertisement messages sent by C-RPs, and then selects the C-RP information for each multicast group to form an RP set (a database of mapping between multicast groups and RPs). The BSR advertises the RP set to the entire network so that all routers (including the DR) obtain the location of RPs.

A PIM-SM domain can also have multiple C-RPs. The RP for each multicast group is elected through the BSR mechanism.

In a network (or management domain), only one router is elected as the BSR, but multiple candidate BSRs (C-BSRs) can be configured. Once the BSR is Down, a new BSR is elected through the bootstrap mechanism. Thus, the service is not interrupted by fault of the BSR.

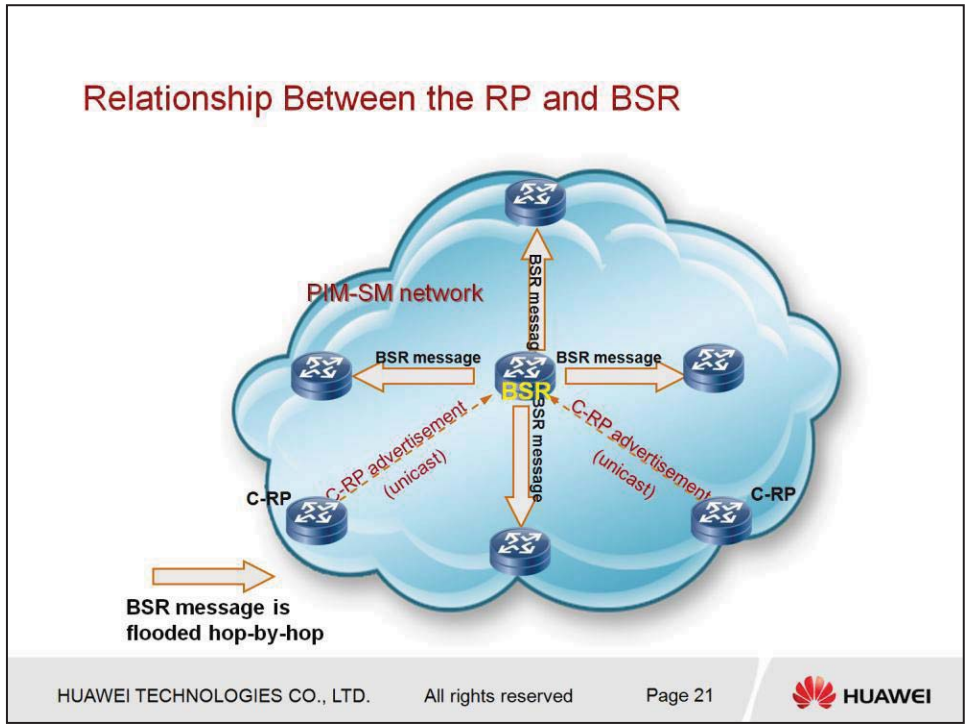


The working mechanism of a BSR is as follows:

1. Qualified routers are configured as C-BSRs. Each C-BSR has a priority. When a router is elected as the C-BSR, it starts a timer (300 seconds by default) to snoop the BootStrap messages in the network. A BootStrap message contains the priority and the IP address of the C-BSR.

2. When a C-BSR receives a BootStrap message, it compares its own priority with the priority in the message. If the priority in the message is higher, the CBSR resets the timer and continue to snoop BootStrap messages. If the priority of the C-BSR is higher, the C-BSR sends a BootStrap message to announce that it is the BSR. If the priorities are equal, the C-BSR compares its IP address with the IP address contained in the message. The router with the larger address becomes the BSR.

The destination address of the BootStrap message is 224.0.0.13 and all PIM routers can receive this packet. The TTL of this packet is set to 1, but each PIM router floods this packet from all its PIM-enabled interfaces. Thus, all PIM routers can receive the BootStrap message.



1. The candidate RP(C-RP) sends the Advertisement message to the BSR.

When receiving the BootStrap message (BSR message), the C-RP can locate the BSR in the network. The C-RP then sends the Candidate-RP-Advertisement message to notify the BSR of the multicast groups it serves in unicast mode. Thus, the BSR collects the information about all C-RPs in the network and establishes an RP set. Then the BSR sends BSR messages to all routers in the network.

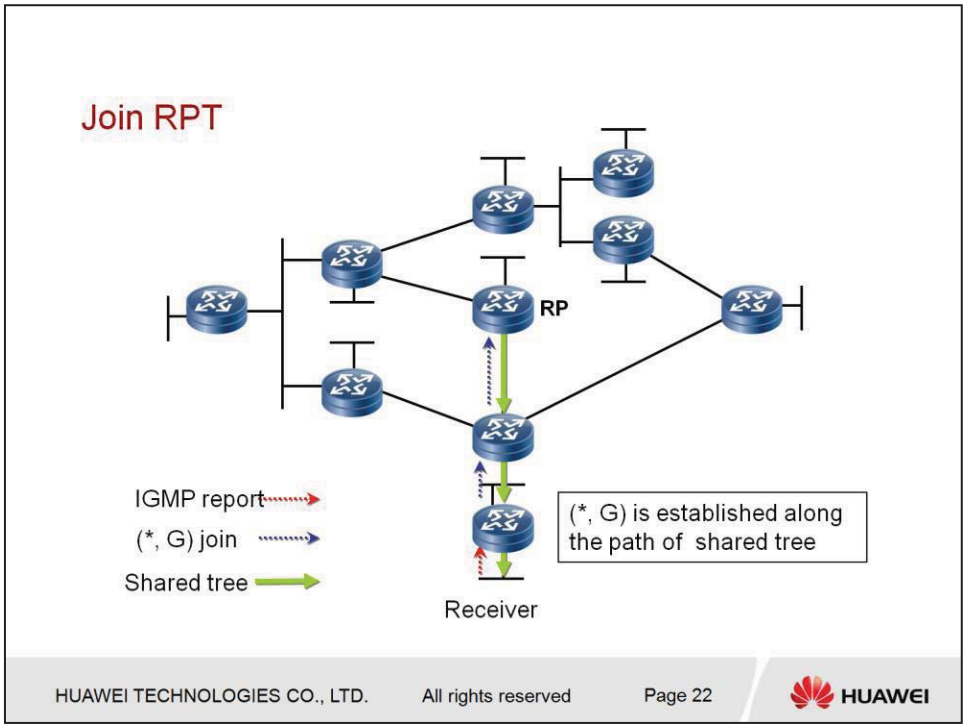
The C-RP periodically sends Advertisement messages (every 60 seconds) in unicast mode.

2. The BSR periodically sends BSR messages (every 60 seconds) to all PIM routers (224.0.0.13).

The BSR message contains the RP-set and BSR address. The message is flooded to the entire network hop by hop.

3. All routers elect the RP according to the received RP set through the same RP election algorithm. Therefore, the routers elect the same RP.

If the RP is specified manually, instead of being elected from C-RPs, each router should be configured with the C-RP address, priority, and the multicast group it serves.



When a receiver host joins multicast group G, the host sends an IGMP packet to notify the leaf router connected to it. The leaf router obtains the information about this receiver, and then sends a Join message to the upstream node toward the RP.

Each router on the path from the leaf router to the RP generates a (*, G) entry in the forwarding table. These routers form a branch of the rendezvous point tree (RPT). (*, G) indicates that packets from any multicast source are sent to G.

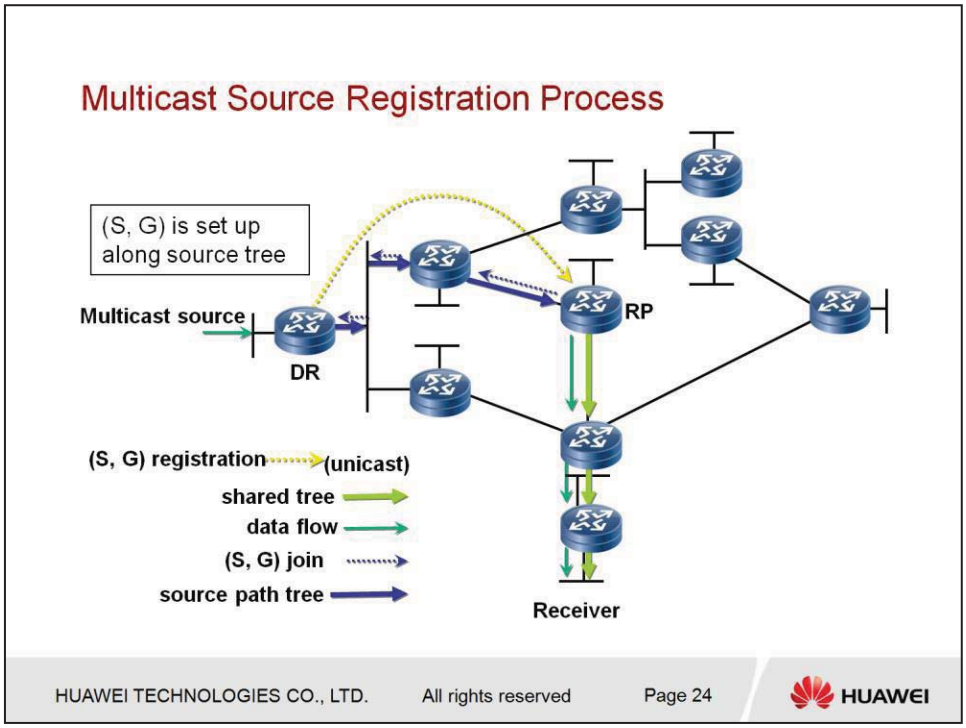
RP is the root of the RPT and the receivers are leaves of the RPT.

After the packet sent from multicast source S to multicast group G reaches RP, the packet will reach the leaf router and then the receiver host. If the receiver host does not want to receive the information from this source, the nearest multicast router sends the Prune message to the RP in the converse direction . When the first upstream router receives the Prune message, the router deletes the interface,

which connects to the downstream router, from the interface list of multicast router.

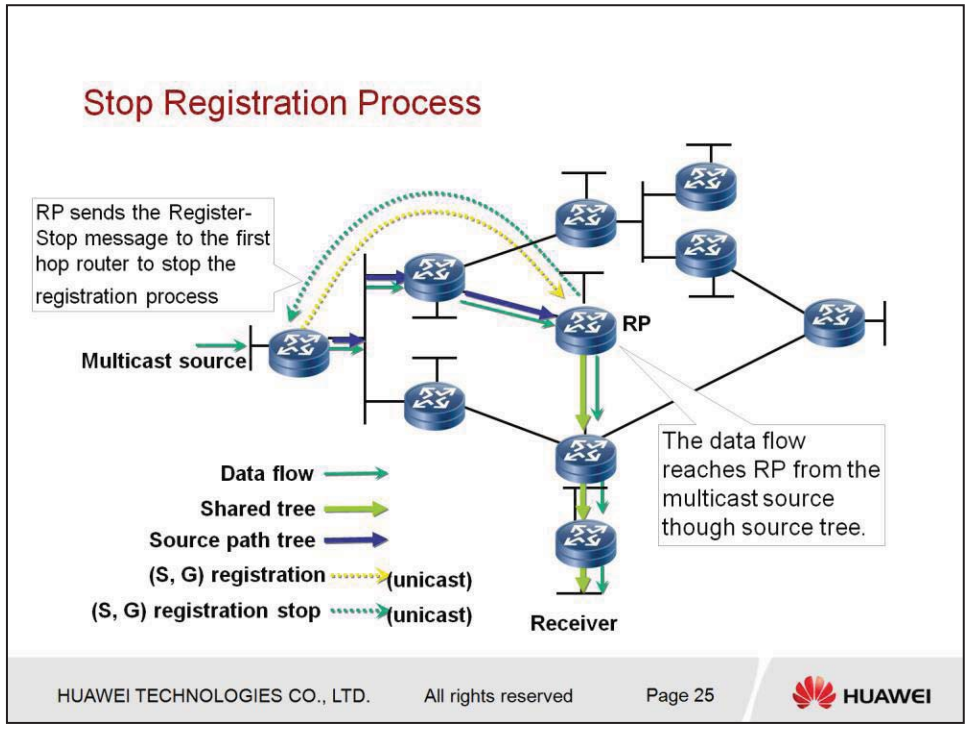
The router check that other existing interface need to forward packets or not. If no need such receivers exist, the router

sends the Prune message to its upstream router.This process is the same as the pruning process of PIM-DM.



When multicast source S sends a multicast packet to multicast group G, the router directly connected to multicast source S encapsulates the multicast packet into the Register packet. Then the router sends the packet to the corresponding RP in unicast mode to notify the RP of the multicast source.

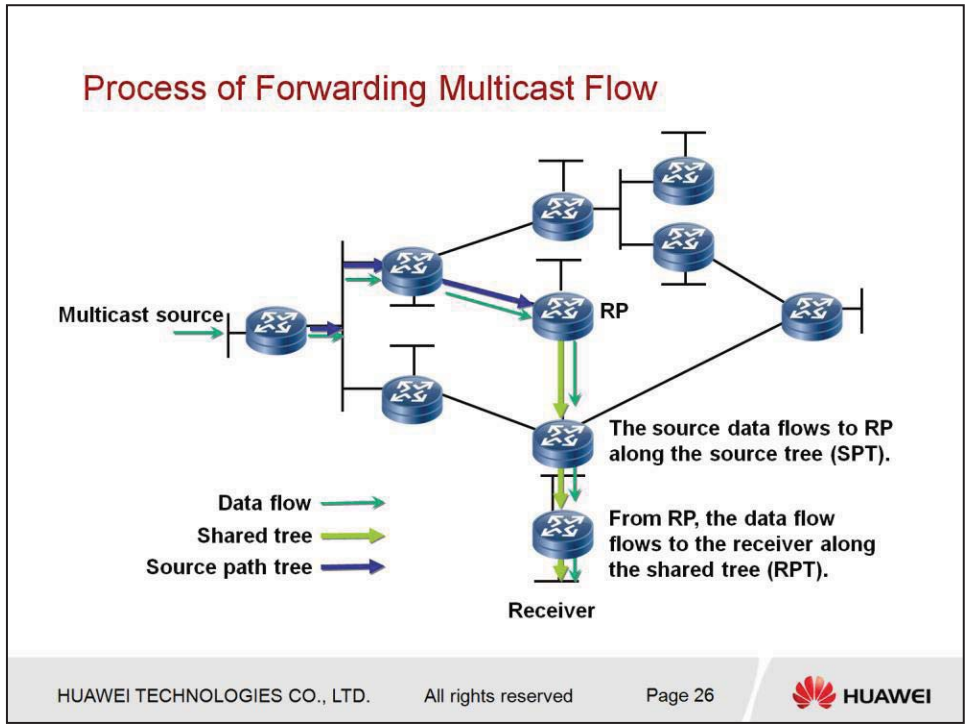
When receiving the Register message from multicast source S, the RP decapsulates the message, and then forwards the message to the receiver along the RPT. In addition, the RP sends the Join message in the converse direction to multicast source S. The Join message contains the (S, G) entry, and thus all routers on the path between the RP and the multicast source generates an (S, G) entry. These routers that the Join message passes form a branch of the SPT. The multicast source is the root of the SPT, and the RP is the destination of the SPT.



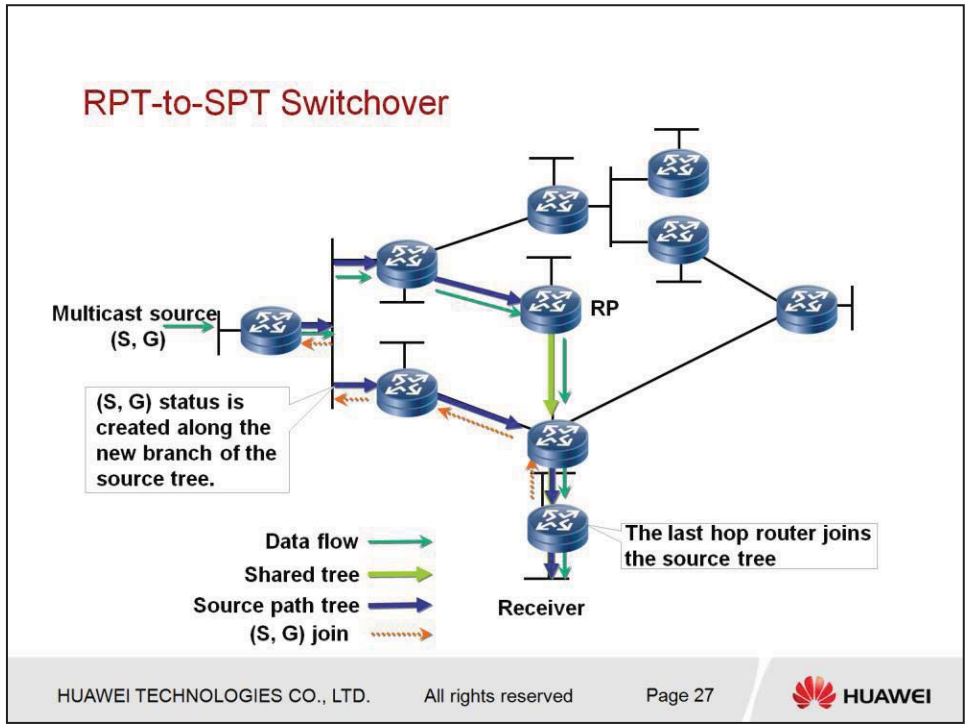
A multicast packet sent by multicast source S reaches the RP along the established SPT. The RP then forwards the packet along the RPT to receivers.

After the RP receives the multicast flow forwarded along the SPT, the RP sends the Register-stop message to the router directly connected to the multicast source.

The multicast source registration process is complete.



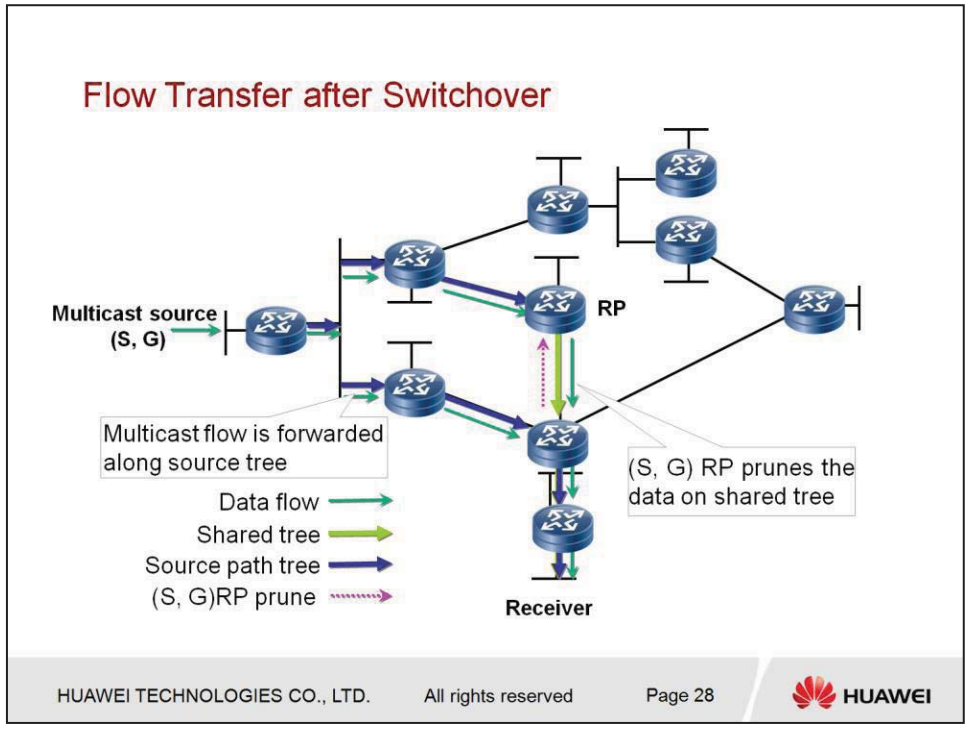
After the multicast source registers to the RP successfully, multicast packets are sent to the RP along the SPT and then forwarded by the RP to the receivers along the RPT.



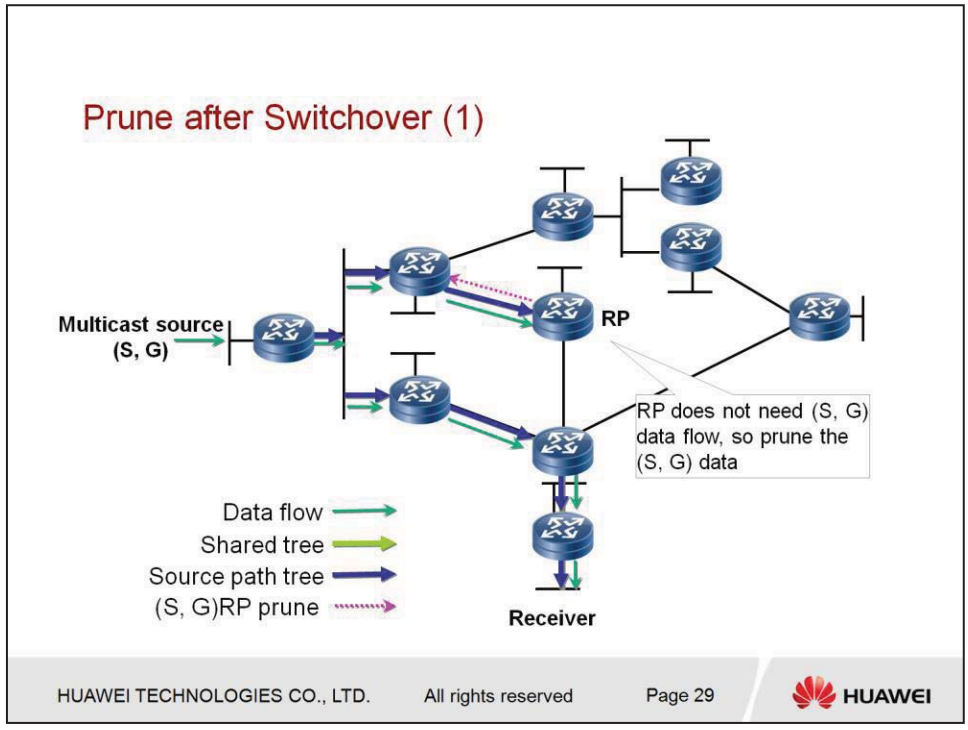
For a specified source, PIM-SM sets the bandwidth threshold for the SPT. Thus the last-hop router (the DR closest to the receiver) can switch from the RPT to the SPT. When the last-hop router detects that multicast packets sent from the RP to multicast group G exceeds the threshold, the router sends a Join message with the (S, G) entry to the next-hop router toward multicast group G according to the unicast routing table. The Join message reaches the first-hop router (the DR closest to the multicast source) after passing the routers on the path. All these routers create an (S, G) entry, and thus an SPT branch is established.

When the throughput of packets exceeds the preset value, PIM-SM switches the router from the RPT to the SPT.

On the VRP, by default, the router directly connected to the receiver joins the SPT immediately after it detects the multicast source; that is, the router receives the first data packet from the source.

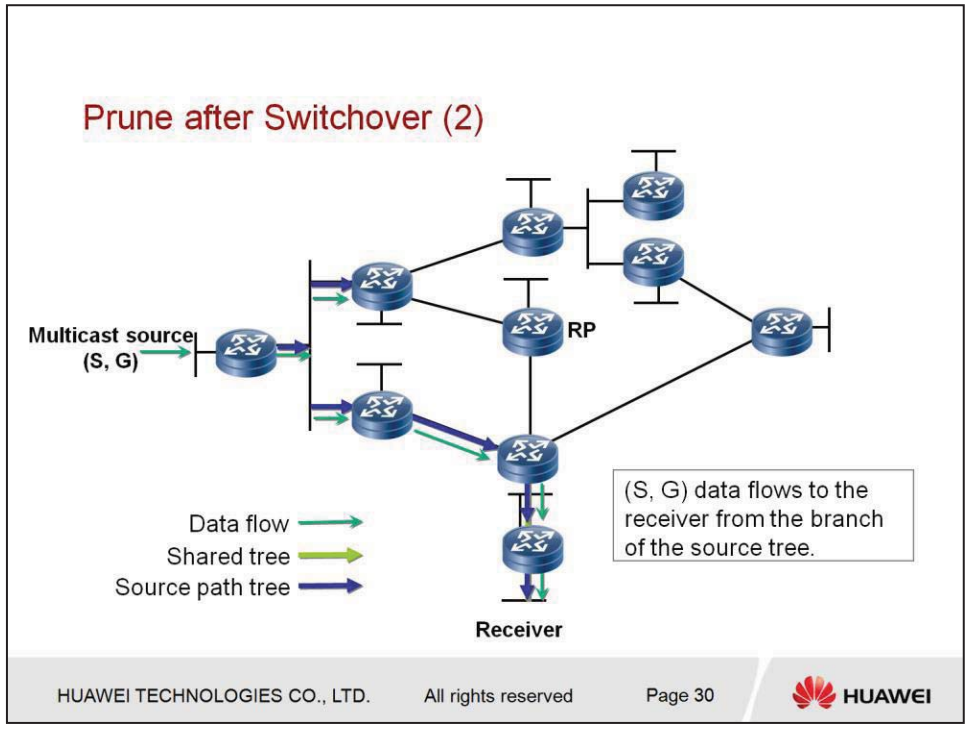


The RPT switches to the SPT, multicast packets are sent from the multicast source directly to the receiver. Through switchover between the RPT and the SPT, PIM-SM can establish the SPT in a more economical way than PIM-DM does.

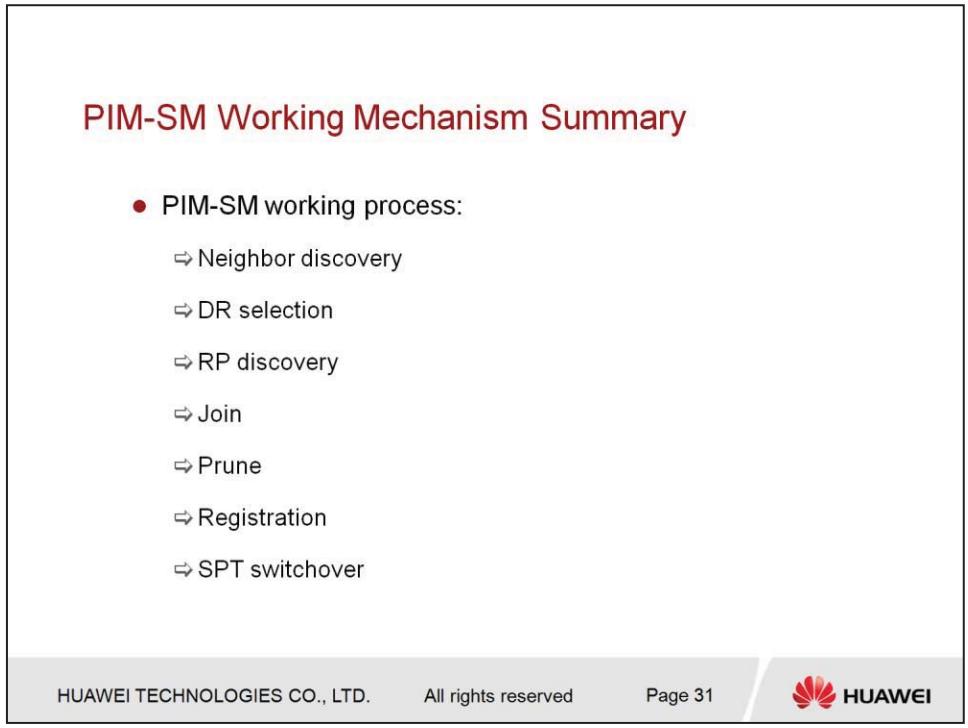


The last-hop router sends the Prune message to the RP. The Prune message contains the RP information and reaches the RP hop by hop. After receiving the Prune message, the RP forwards the Prune message in the converse direction.

Thus, the multicast flow is switched from the RPT to the SPT.



After switchover, an SPT is established between the multicast source and the receiver.



Neighbor discovery

In a PIM-SM network, multicast routers send Hello messages to discover neighbors, maintain the neighbor relation, and negotiate protocol parameters.

Routers in the same network segment compare the priorities and IP addresses in Hello messages to elect the DR. The DR acts as the IGMPv1 querier.

Assert

Routers in a shared network (such as an Ethernet) elects the designated forwarder.

DR election

Routers elects the designated router (DR) for the shared network between the multicast source and receivers.

RP discovery

The RP is specified manually or elected through the BSR messages.

Join

When a receiver host joins a multicast group G. The host send an IGMP packet to notify the leaf router directly connected to it.

The leaf router then

sends the Join message to the upstream node toward the RP.

Prune

The pruning process is initiated by the leaf router. When all downstream members leave the multicast group, the leaf router sends a Prune message to the upstream node to request the upstream node stop forwarding packets to this branch.

Register

The Register messages notify the RP of the existence of the multicast source.

RTP-SPT switchover

PIM-SM sets a bandwidth threshold for the SPT. When the traffic exceeds this threshold, the last-hop router (the DR at the receiver side) switches from the RPT to the SPT.

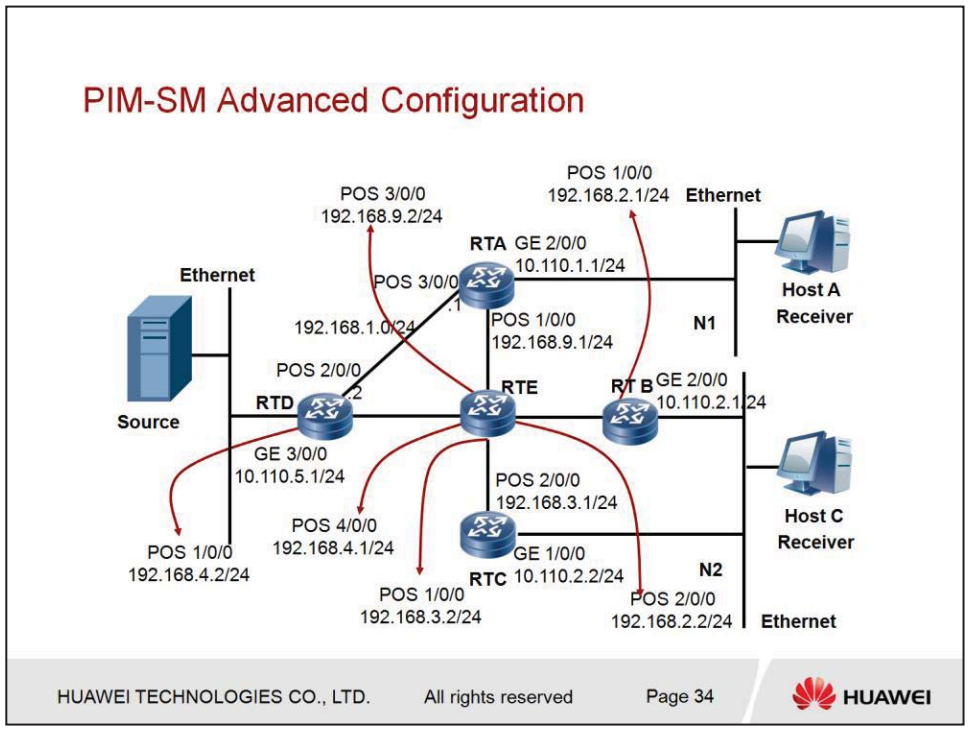
Contents

PIM-SM overview and basic configuration

PIM-SM mechanism

PIM-SM advanced configuration

HUAWEI TECHNOLOGIES CO., LTD. All rights reserved



In this figure:

Interface POS3/0/0 of Router E is the C-BSR and C-RP interface of the PIM-SM network.

IGMPv2 runs between Router A and end network N1. IGMPv2 also runs between Router B, Router C, and end network N2.

Configuration roadmap:

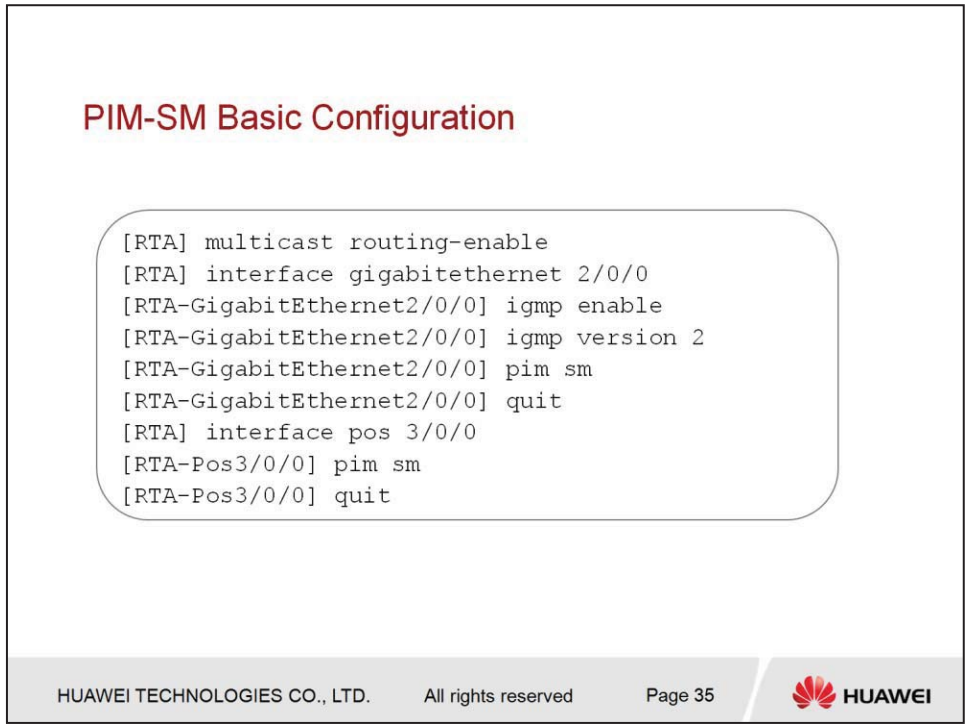
1. Configure the IP addresses of interfaces on the router and enable the unicast routing protocol.

2. Configure the IP address and mask of each interface.

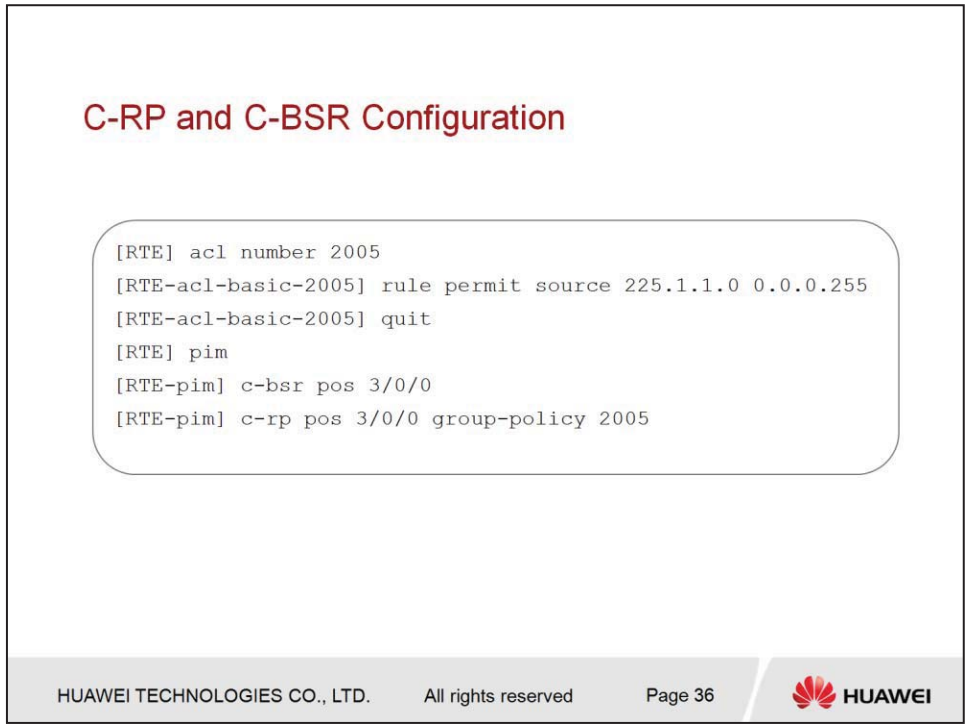
3. Enable OSPF between the routers. Make sure routers can communicate with each other .

4. Enable the multicast function. Enable PIM-SM on each interface and enable IGMP on the interfaces connected to hosts.

5. Configure POS3/0/0 of Router RTE as the C-BSR and C-RP interface.



Configurations of Router B, Router C, Router D, and Router E are similar to the configuration of Router A.



c-rp interface-type interface-number [group-policy basic-acl-number | priority priority | holdtime hold-interval | advertisement-interval adv-interval] * interface-type interface-number: indicates the interface of the C-RP. PIM-SM must be enabled on this interface.

group-policy basic-acl-number : indicates that the multicast groups that the CRP serves are groups permitted by the ACL.

basic-acl-number indicates the number of the basic ACL.

By default, C-RP serves all multicast groups.

priority priority : indicates the priority of the C-RP. A larger value indicates a lower the priority. The default value is 1.

Using the undo c-rp { interface-type interface-number | all } command, you can cancel the C-RP configuration.

c-bsr interface-type interface-number hash-mask-len [priority]

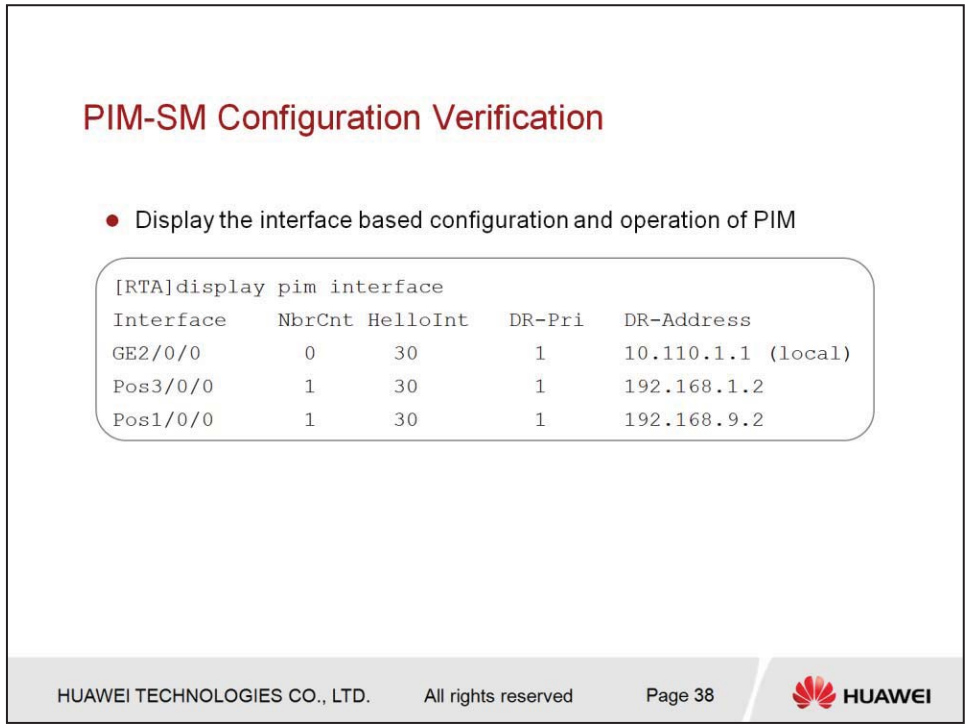
interface-type interface-number: specifies the interface type and interface number on the router. The C-BSR is configured on the specified interface. PIMSM must be enabled on the interface; otherwise, the configuration does not take effect.

hash-mask-len: specifies the mask length. The router ANDs the mask with the multicast group, and then searches for the RP.

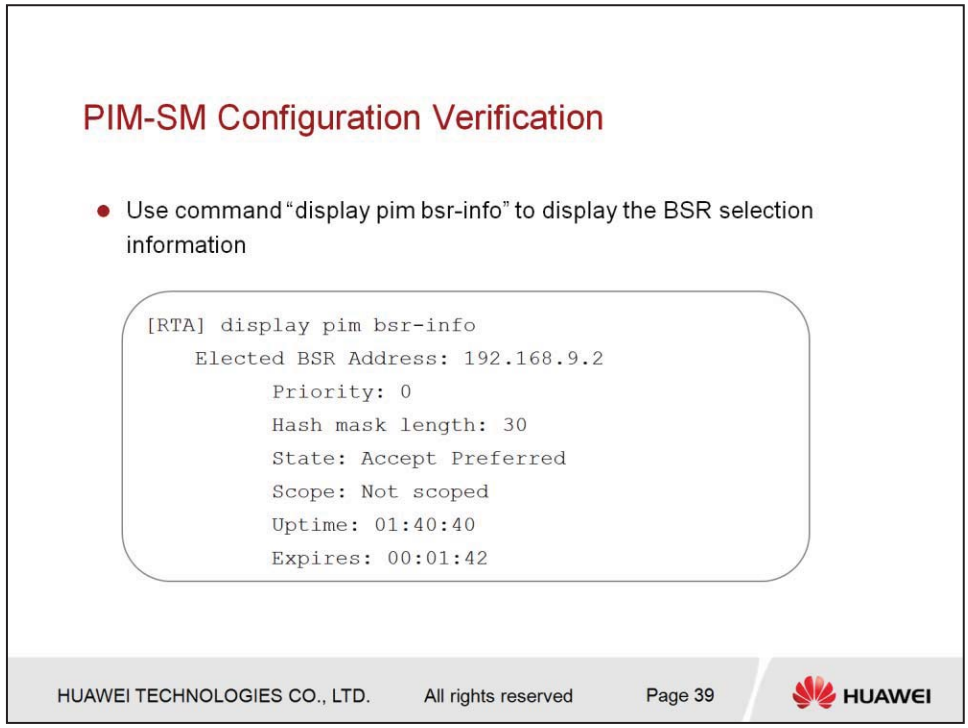
The value ranges from 0 to 32.

priority: specifies the priority of the C-BSR. The value ranges from 0 to 255. The default value is 0. A larger value indicates a higher priority.

Using the undo c-bsr command, you can cancel the configuration of the CBSR.

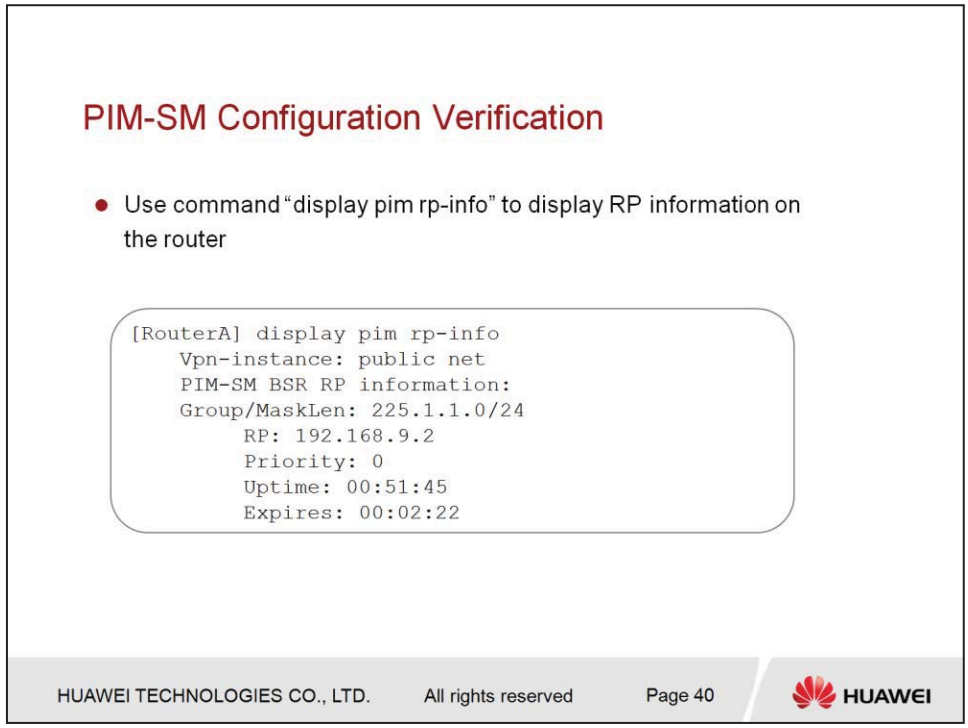


Using the display pim interface command, you can display the information about all PIM-enabled interfaces. You can find all PIM-SM-enabled interfaces in the display information.

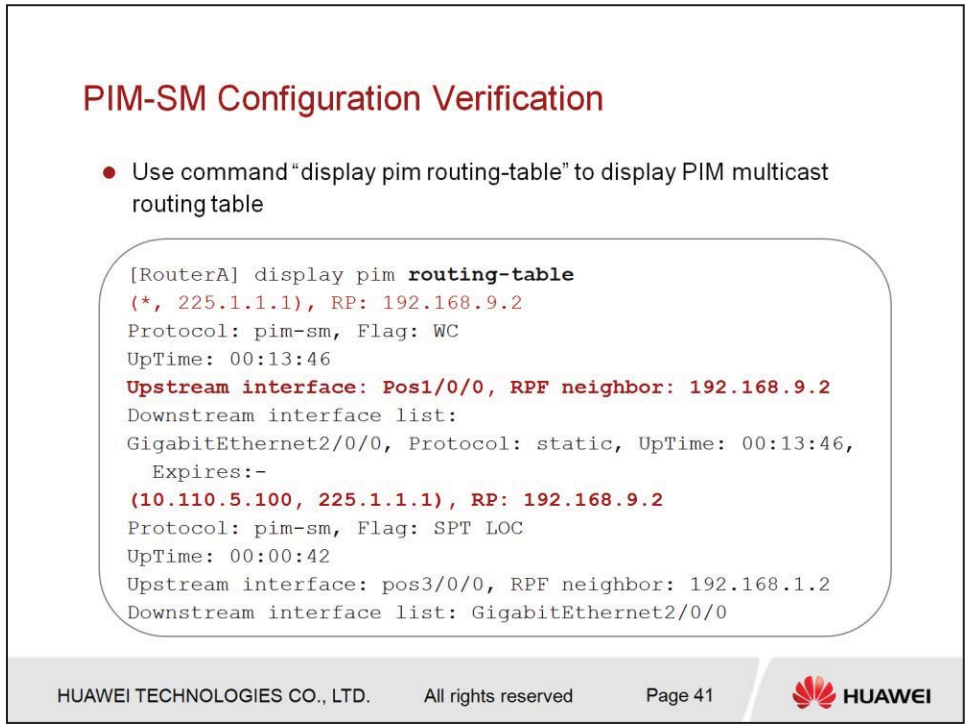


Using the display pim bsr-info command, you can display the information about the BSR.

The display information includes the IP address, Hash mask length, and priority of the BSR.



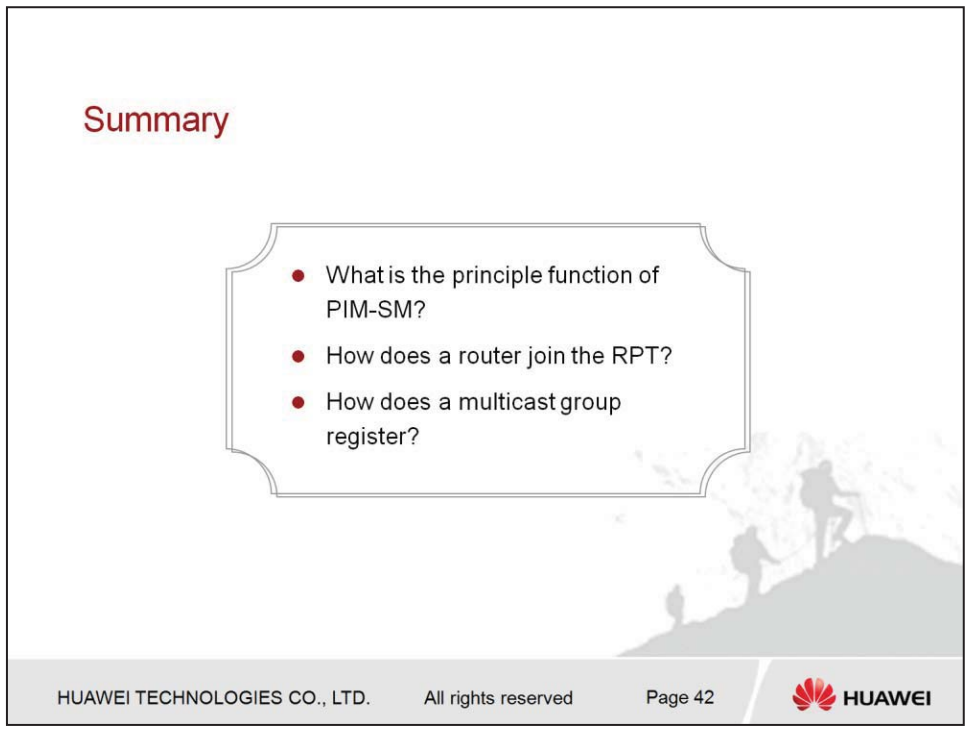
Using the display pim rp-info command, you can display the information about RPs for all multicast groups, including the RP discovered through BSR mechanism and the static RPs.



Using the display pim routing-table command, you can display the PIM multicast routing tables of all instances.

The display information shows the information about the upstream interfaces mapping the (*, G) entry and (S, G) entry, RFP neighbors, and RPs.

If (*, G) and (S, G) entries are created on the router, it indicates that the router can forward multicast routers.



1. What is the principle of PIM-SM?

PIM-SM is based on the assumption that no network nodes need to receive multicast data when the multicast source begins to send multicast packets. The core task of multicast forwarding is establishing and maintaining a unidirectional RPT. A router is elected as the RP of the RPT. The RP forwards the multicast packet to the receiver along the RPT.

At the receiver side, the router connected to the receiver sends the Join message to the RP of the multicast group. The Join message is forwarded hop by hop and reaches the root (RP). The path that the Join message passes by becomes a branch of the RPT.

When the sender needs to send multicast data to a multicast group, the first-hop router sends registers message to the RP. The Register message reaches the RP and triggers establishment of the SPT. The multicast source then sends the multicast packet to the RP. The multicast packet is duplicated on the RP and forwarded to the receivers along the RPT. The packet is duplicated only at the crotches. The duplication repeats until the packet reaches the receiver.

2. How does a router join the RPT?

When a receiver host joins multicast group G, the host sends an IGMP packet to notify the leaf router to which receiver host connected . The leaf router obtains the information about this receiver, and then sends a Join message to the upstream node toward the RP.

Each router on the path from the leaf router to the RP generates a (*, G) entry in the forwarding table. These routers form a branch of the rendezvous point tree (RPT).

3. How does a multicast group register?

When multicast source sends a multicast packet to multicast group G, the router directly connected to multicast source encapsulates the multicast packet into the Register packet, and then sends the packet to the corresponding RP in unicast mode.

When receiving the Register message from multicast source, the RP de-encapsulates the message, and then forwards the message to the receiver along the RPT. In addition, the RP sends the Join message in the converse direction to multicast source. The Join message contains the (S, G) entry, and thus all routers along the path between the RP and the multicast

source generates an (S, G) entry. These routers that the Join message passes form a branch of the SPT. The multicast source is the root of the SPT, and the RP is the destination of the SPT.