Recommendations

- Huawei Learning Website
  - http://learning.huawei.com/en
- Huawei e-Learning
  - https://ilearningx.huawei.com/portal/#/portal/EBG/51
- Huawei Certification
  - http://support.huawei.com/learning/NavigationAction!createNavi?navId=_31&lang=en
- Find Training
  - http://support.huawei.com/learning/NavigationAction!createNavi?navId=_trainingsearch&lang=en

More Information

- Huawei learning APP

Huawei Certification

HCIP-Routing&Switching

Implementing Enterprise Routing and Switching Network

V2.5

Huawei Technologies Co.,Ltd.

Huawei Certification

HCIP-Routing&Switching Implementing Enterprise Routing and Switching Network

Version 2.5

Huawei Certification System

Relying on its strong technical and professional training and certification system and in accordance with customers of different ICT technology levels, Huawei certification is committed to providing customers with authentic, professional certification, and addresses the need for the development of quality engineers that are capable of supporting Enterprise networks in the face of an ever changing ICT industry. The Huawei certification portfolio for routing and switching (R&S) is comprised of three levels to support and validate the growth and value of customer skills and knowledge in routing and switching technologies.

The Huawei Certified Network Associate (HCIA) certification level validates the skills and knowledge of IP network engineers to implement and support small to medium-sized enterprise networks. The HCIA certification provides a rich foundation of skills and knowledge for the establishment of such enterprise networks, along with the capability to implement services and features within existing enterprise networks, to effectively support true industry operations.

HCIA certification covers fundamentals skills for TCP/IP, routing, switching and related IP network technologies, together with Huawei data communications products, and skills for versatile routing platform (VRP) operation and management.

The Huawei Certified Network Professional (HCIP-R&S) certification is aimed at enterprise network engineers involved in design and maintenance, as well as professionals who wish to develop an in depth knowledge of routing, switching, network efficiency and optimization technologies. HCIP-R&S consists of three units including Implementing Enterprise Routing and Switching Network (IERS), Improving Enterprise Network Performance (IENP), and Implementing Enterprise Network Engineering Project (IEEP), which includes advanced IPv4 routing and switching technology principles, network security, high availability and QoS, as well as application of the covered technologies in Huawei products.

The Huawei Certified Internet Expert (HCIE-R&S) certification is designed to imbue engineers with a variety of IP network technologies and proficiency in maintenance, for the diagnosis and troubleshooting of Huawei products, to equip engineers with in-depth competency in the planning, design and optimization of large-scale IP networks.

CONTENTS

OSPF Protocol Basics

Foreword

- The Routing Information Protocol (RIP) is a routing protocol based on the distance vector algorithm. Disadvantages of RIP emerge when it is deployed in large networks, including slow convergence and poor scalability.
- The Open Shortest Path First (OSPF), an SPF algorithm-based link-state routing protocol, was developed by the Internet Engineering Task Force (IETF). OSPF can be deployed in large networks to overcome RIP' s disadvantages. So, how does OSPF meet network expansion requirements?

Objectives

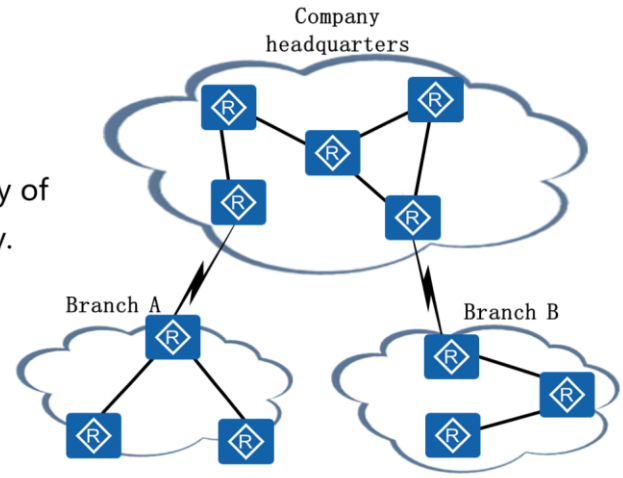
- Upon completion of this section, you will be able to:
  - Understand RIP problems in large networks
  - Be familiar with OSPF characteristics
  - Master network types supported by OSPF
  - Master the process of establishing OSPF neighbor relationships
  - Master the concepts and functions of OSPF DR and BDR

Contents

HUAWEI

Changes in Large Networks

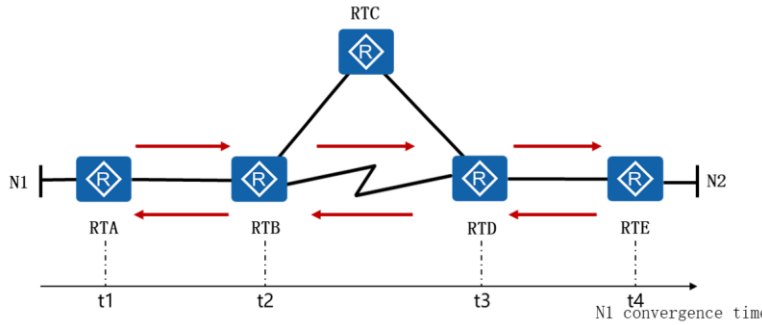
- Networks are expanding.
- Higher network reliability is required.
- A trend towards heterogeneity of networks goes up significantly.



HUAWEI

- Networks are expanding:
  - Driven by emerging enterprise services and a trend towards concentration of services, the internetwork scale has enlarged continuously.
- Higher network reliability is required:
  - Various applications require higher network reliability. When a network failure occurs, the network needs to recover within a shorter period.
- Network heterogeneity requires interconnection between multi-vendor devices:
  - During routine O&M, devices are continually upgraded or updated. There may be large differences in performance between devices. Also, link bandwidths vary to a certain degree.
  - An open routing protocol supported by various vendors is required.
- Can RIP meet these requirements? What problems may RIP encounter?

Existing RIP Problems in Large Networks



RIP Characteristics	Problem
Hop-by-hop convergence	Slow convergence and long failure recovery time
Routing-by-rumor mechanism	Lack of understanding of the global network topology
Maximum available hops is 15	In ring topology, remote network will be regarded as unreachable
Using hop count to measure the distance to a destination	A risk of selecting a suboptimal route

- Hop-by-hop convergence:
  - As shown in the preceding figure, when network N1 changes, RTA sends an Update message to RTB. After RTB receives the Update message, it performs route calculation and sends a route change notification to RTC. In this manner, route convergence is performed hop by hop, slowing down network convergence.
- Routing-by-rumor mechanism:
  - RIP route calculation completely depends on the routing information from neighbors. For example, RTE calculates routes just depending on the routing information obtained from RTD without any idea of information about the network between RTA, RTB, and RTC. Namely, when calculating routes, RIP lacks the complete knowledge of network topology.
- Employing the hop count as a routing metric:
  - RIP uses hop count to measure the distance to a destination. Therefore, when data packets traverse networks from N1 to N2, the links RTA->RTB->RTD->RTE are chosen as the optimal route. Obviously, the Ethernet links RTB->RTC->RTD have much higher bandwidths than the serial link RTB->RTD.
- Is there any solution to RIP problems?

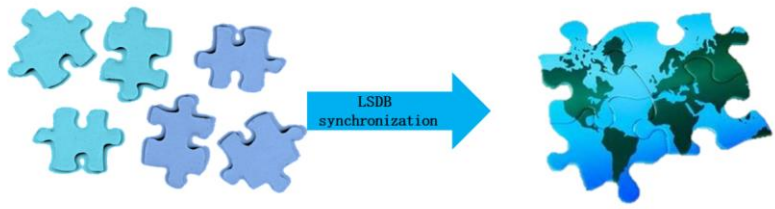
Solving RIP Problems

RIP Problems	Solutions
Slow convergence and long failure recovery time	Change from "receiving updates->calculating routes->sending updates" To "receiving updates->sending updates->calculating routes"
Lack of view of the global network topology	A router independently calculates routes on the basis of topology information
Maximum available hops is 15	Unlimited hops
A risk of selecting a suboptimal route	Use the link bandwidth as the criterion for route selection

- In RIP, the network convergence process involves receiving updates, calculating routes, and sending updates. This process slows down network convergence because a router sends route change notifications to neighbors only after it finishes route calculation. To solve this problem, the network convergence process can be changed to: receiving updates, sending updates, and calculating routes. That is, upon receiving route updates from a neighbor, a router sends the updates to other neighbors and then calculates routes again. By this means, the convergence time of the network is greatly reduced.
- Because a RIP router obtains routing information just from neighbors, it cannot identify or eliminate non-optimal or incorrect routing information. The best solution to this problem is that it collects networkwide information and independently calculates routes on the basis of the information.
- When using hop count to measure the distance to a destination, the propagation delay is not taken into account. If the accumulated bandwidth is used as the criterion for route decision, the risk of selecting a suboptimal route can be eliminated.
- By what means do link-state routing protocols solve all the problems mentioned above?

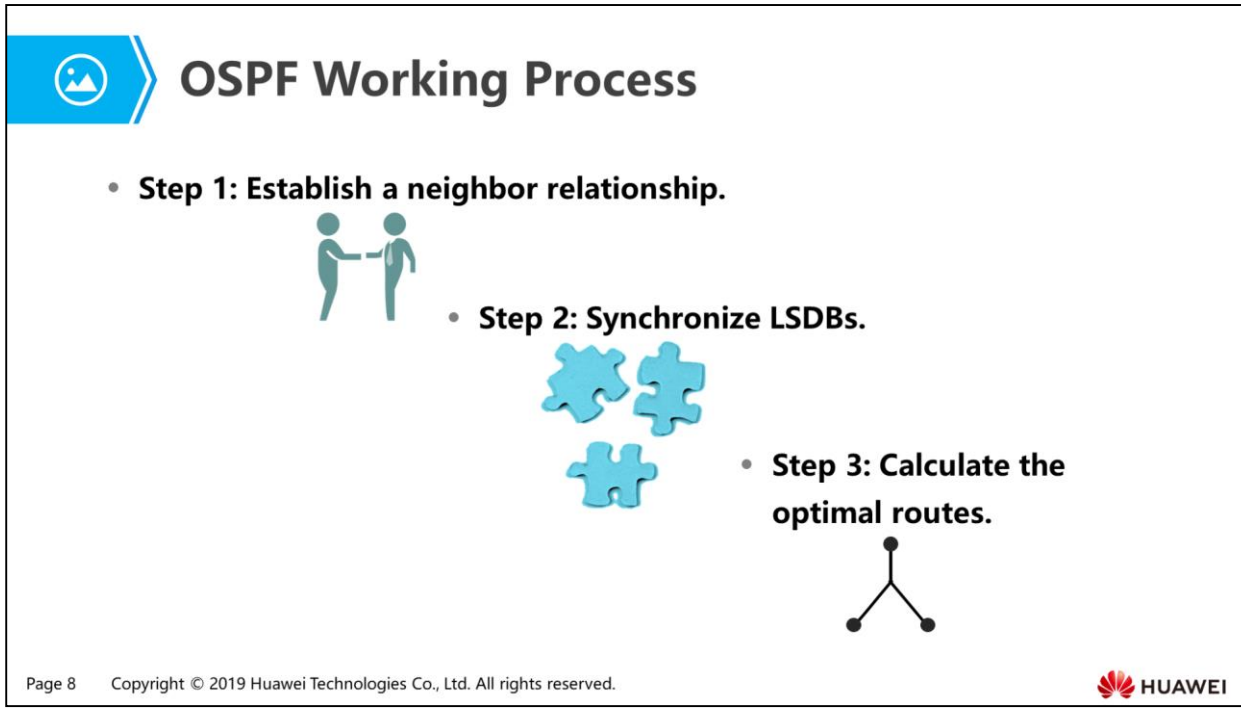
Link-State Routing Protocol OSPF

- Separates routing information propagation from route calculation.
- Uses the Shortest Path First (SPF) algorithm.
- Uses the accumulated link cost as the criterion for route selection.



  HUAWEI

- The term link-state refers to the status of interfaces on an OSPF router, which includes the following information:
  - IP address and mask of the interface
  - Bandwidth of the interface
  - Neighbor that the interface is connected to
  - ...
- OSPF transmits link state information to its neighbors instead of transmitting its complete routing table.
- Each router maintains its link state database (LSDB). Neighbors synchronize their LSDBs and then use the SPF algorithm to calculate the optimal route. This speeds up the network convergence speed.
- OSPF uses the accumulated bandwidth of the links as the criterion for route selection. This method is more accurate than using the accumulated hop count.
- OSPF can solve RIP problems in large networks. The following describes how OSPF works.

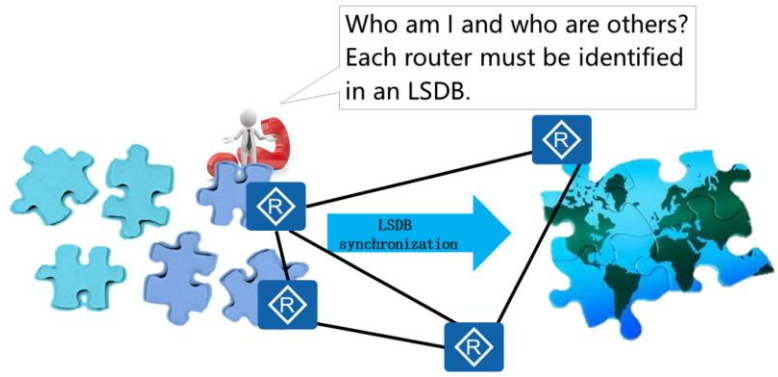


- An enterprise network, often in analogy with a road map, usually consists of a large number of devices, including routers and switches, of different models and various kinds of links, which results in great computational complexity.
- The process of OSPF route calculation can be summarized in three steps:
  - Routers discover neighbors and establish neighbor relationships.
  - Each router generates link state information and floods it to all neighbors, meanwhile collects link state information from its neighbors until the LSDBs are synchronized among OSPF neighbors.
  - On the basis of its LSDB, each router uses the SPF algorithm to calculate a Shortest Path Tree (SPT) with itself as the root. By building an SPT, each router calculates the optimal routes to destination. The routes form a routing table eventually.
- Let's focus on the three steps mentioned to learn the principles and implementation of OSPF.

Contents

Router ID

- Each OSPF router owns a unique router ID, which uniquely identifies a router within an autonomous system (AS).



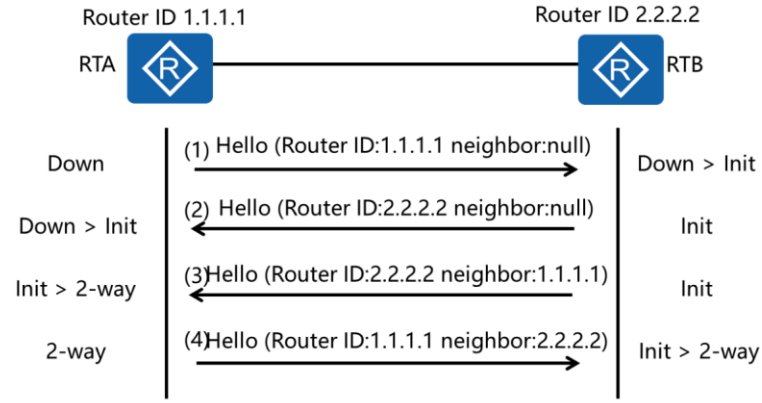
- There are several, dozens of, or even hundreds of devices on an enterprise network. Each of these devices must be uniquely identified by a router ID.
- A router ID is a 32-bit unsigned integer in the format of an IP address. An OSPF router uses the following criteria to select the router ID:
  - It prefers the manually configured router ID. You are advised to manually configure a router ID for an OSPF router.
  - If there is no router ID manually configured, it selects the largest IP address on any of its loopback interface as its router ID.
  - If no loopback interface is present, it uses the largest IP address on any of its active physical interfaces as its router ID.
- If a router ID is reconfigured on an OSPF router, the OSPF process must be reset to apply the new router ID.

Neighbor Discovery and Neighbor Relationship Establishment - Hello Packet

- A Hello packet performs the following functions:
  - To discover neighbors: An OSPF router can send a Hello packet to discover neighboring routers.
  - To establish neighbor relationships: Two OSPF routers negotiate parameters in Hello packets to establish a neighbor relationship.
  - To maintain neighbor relationships: OSPF routers use the Keepalive mechanism to continually detect the neighbor reachability.

- Prior to an exchange of link state information, OSPF routers need to exchange Hello packets to establish a neighbor relationship.
  - In OSPF, interconnected routers exchange Hello packets to discover neighbors and establish neighbor relationships. This is the preparation for synchronizing reachability information following up.
  - When two OSPF routers sharing a common data link successfully negotiate certain parameters, the neighbor relationship between them is established.
  - After a neighbor relationship is established, routers periodically exchange Hello packets to maintain the neighbor relationship. If a router does not receive any Hello packet from its neighbor within a specified period, this router terminates the OSPF neighbor relationship with this neighbor.

OSPF Neighbor Relationship Establishment Process



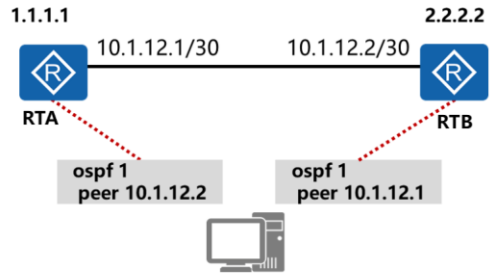
  HUAWEI

- Description of neighbor states:
  - Down: This is the initial state of a neighbor relationship. It indicates that there has been no information received from the neighbor.
  - Init: This state occurs when a router has received a Hello packet from its neighbor but its router ID is not in the neighbor list contained in the received Hello packet. This means that bidirectional communication with the neighbor has not yet been established.
  - 2-Way: In this state, a router finds that its router ID is in the Hello packet from a neighbor. Bidirectional communication with the neighbor is then established.
- The process of establishing an OSPF neighbor relationship is as follows:
  - RTA has a router ID 1.1.1.1 and RTB has a router ID 2.2.2.2. When the OSPF process starts on RTA, RTA sends the first Hello packet, in which the neighbor list is empty. The neighbor state is Down. When RTB receives the Hello packet from RTA, the neighbor state is set to Init.
  - Likewise, almost at the same time, RTB sends a Hello packet with an empty list of active neighbor. As what RTB does, RTA sets the neighbor state to Init, as soon as it receives this Hello packet from RTB.

- RTB sends another Hello packet to RTA, saying that the router ID 1.1.1.1 is on its list of active neighbor. Upon receiving this Hello packet, RTA finds its router ID included in the neighbor list of the packet, and then sets the neighbor state to 2-way.
- (Similarly, RTA sends RTB the next Hello packet with RTB’ s router ID 2.2.2.2 on the list of active neighbor. RTB sees itself in the neighbor list of the packet, then sets the neighbor state to 2-way.
- Because the neighbors were unknown before an OSPF router starts its discovery of neighbors, the destination IP address of a Hello packet is a multicast address 224.0.0.5 instead of a specific unicast address. How does an OSPF router discover neighbors on a network that does not support multicast?

Neighbor Discovery and Neighbor Relationship Establishment - Manual Configuration

- OSPF allows routers to establish neighbor relationships using unicast Hello packets.
- If the network does not have the multicast capability, neighbor relationships must be manually established.



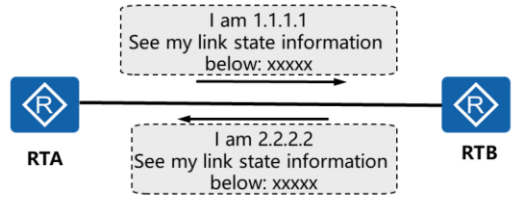
- If the network does not have the multicast capability, neighbor relationships must be manually established.
- When the network scale increases or devices are frequently updated, static configurations need to be modified on related OSPF routers. However, manually modifying configurations is of heavy workload and error-prone.
- For an OSPF router, the purpose of establishing neighbor relationships is to synchronize LSDB by exchanging link-state information with other OSPF routers. The following describes how OSPF routers synchronize their LSDBs.

Contents

HUAWEI

Link State Information

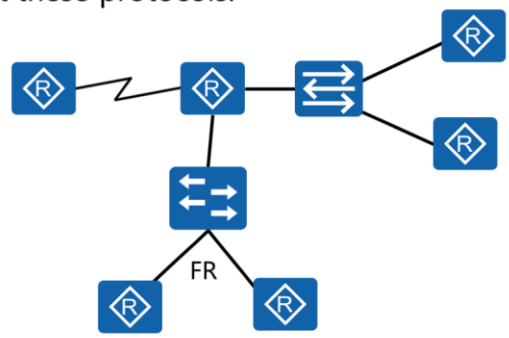
- Link information includes:
  - Link type
  - Interface IP address and subnet mask
  - Neighboring routers sharing the same link
  - Link bandwidth (cost)



- Unlike the route exchange process between RIP routers, OSPF routers exchange link state information to synchronize their link-state databases and then just forward the original link state information to other neighbors. Eventually, all OSPF routers will possess the same set of link state information.
- Link state information includes the link type, interface IP address and subnet mask, neighbors on the link, and link cost.
- A router just needs to know the destination network ID/subnet mask, next hop, and cost (interface IP address and subnet mask, neighbors on the link, and link cost). But why is the link type included in the link state information?

Extensive Data Link Layer Support

- A variety of data link layer protocols exist, and their working mechanisms differ.
- Application scenarios of various data link layer protocols must be considered to support these protocols.

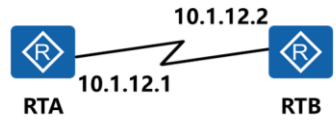


HUAWEI

- The development of devices, communication media, and protocols are critical parts of the evolution of networking technologies. Device performance increasingly improves, and communication links have developed from serial link, ATM, and FR into Ethernet, xPON, SDH, MSTP, and OTN. A new generation of technology never comes into being at the snap of a finger. Instead, it is a progressive process. Different kinds of physical links have their unique characteristics. Accordingly, a mature routing protocol must fit the characteristics of a physical link where it is applied.
- The following presents how OSPF defines network types.

Network Type - Point-to-Point Network

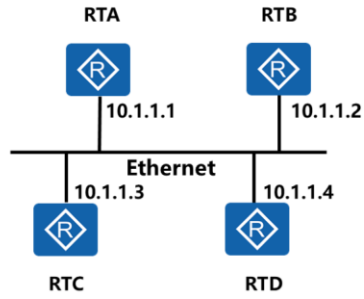
- Only two routers are connected to each other.
- Broadcast and multicast are supported.



- Four network types are defined in OSPF protocol and are used to describe network topologies partially.
- A point-to-point network connects a pair of routers together and supports broadcast and multicast.
- An example of a point-to-point network: a network on which two routers are connected through a Point-to-Point Protocol (PPP) link.

Network Type - Broadcast Network

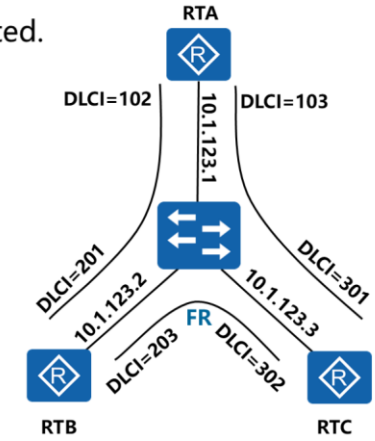
- Two or more routers are connected through shared media.
- Broadcast and multicast are supported.



- A broadcast network allows two or more devices to access the same shared link and supports broadcast and multicast. It is one of the most common OSPF network types.
- An example of a broadcast network: a network on which routers are connected through an Ethernet link.
- There are multiple devices on a broadcast network. As to neighbor relationship establishment and link information synchronization, OSPF provides corresponding functions to reduce the impact of multiple devices on the network.
- Point-to-point and broadcast networks are the most common OSPF network types. Besides, there are two more network types which are rarely found.

Network Type - NBMA Network

- Two or more routers are connected through a virtual circuit (VC).
- Broadcast and multicast are not supported.



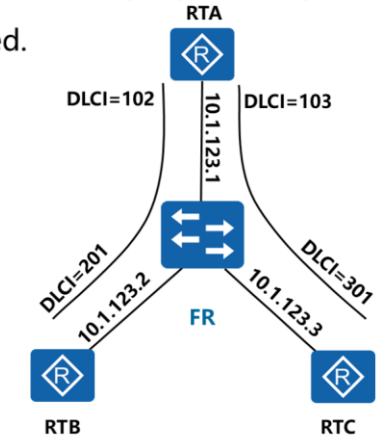
Fully meshed FR network

HUAWEI

- Unlike the broadcast network, a non-broadcast multiple access (NBMA) network does not support broadcast and multicast by default. On an NBMA network, OSPF emulates operations over a broadcast network. But the neighbor must be manually specified.
- An example of NBMA network: a network on which routers are connected through fully-meshed FR links.
- In recent years, NBMA networks are seldom deployed.

Network Type - Point-to-Multipoint Network

- A point-to-multipoint network is a set of multiple point-to-point networks.
- Broadcast and multicast are supported.



Partially-meshed FR network

- A point-to-multipoint network, which is a special configuration of non-broadcast multi-access networks, is treated as a group of point-to-point connections. On a point-to-multipoint network, OSPF neighbors can be discovered through the Inverse Address Resolution Protocol (Inverse ARP). A point-to-multipoint network can be considered as a set of multiple point-to-point networks and supports broadcast and multicast.
- No network is originally a point-to-multipoint network, no matter what kind of data link layer protocol is being employed. That is to say, a point-to-multipoint network must be converted from other type of network. A typical practice is changing partially-meshed FR or ATM networks to point-to-multipoint networks.
- So here comes the question: how is the cost in OSPF link state information determined?

OSPF Cost

- Interface cost = Reference bandwidth/Interface bandwidth
- Hence, the cost can be changed in two ways:
  - Configure the cost directly under interface configuration mode.
  - Change the reference bandwidth. This alternation must be performed on all routers to ensure route selection consistency.

- Accumulated cost from RTA to subnet 192.168.3.0/24 = G1' s cost+G3' s cost

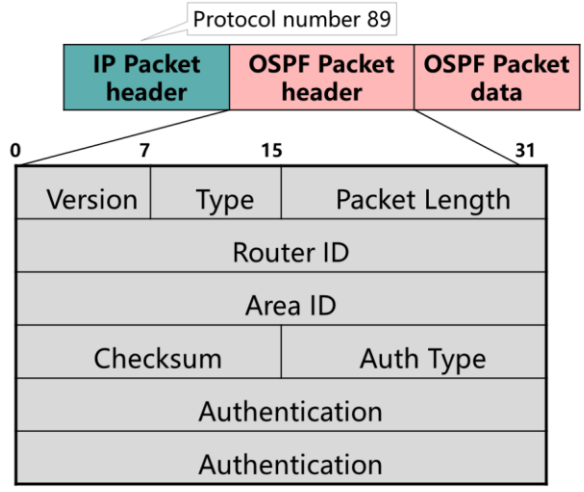
- The formula for OSPF cost calculation is: Interface cost = Reference bandwidth/Interface bandwidth, which defines the interface cost as the reference bandwidth divided by the interface bandwidth. The default reference bandwidth is 100Mbps. If the result is not an integer, it is truncated to a whole number to be the cost value. If the result is less than 1, the interface has the cost of 1.
- The cost can be changed in two ways:
  - Configure the cost under interface configuration mode. Note that the configured cost is the final value for this interface , but is valid only on this interface.
  - Change the default reference bandwidth for OSPF. This modification takes effect only on all OSPF-enabled interfaces of the local router. It is suggested that all routers in the Autonomous System should be configured with the same reference bandwidth in order to ensure consistency in route selection. When changing the default reference bandwidth, you should give overall consideration on bandwidth distribution on the whole network before you decide a new value.
- OSPF metric is the cumulative cost, which is a total cost of all outbound interfaces of the routers the packet passing through from the source to the destination. For example, if a packet routed by OSPF from RTA destined for RTC' s interface Loopback 1 attached to subnet 192.168.3.3/24, the cost is equal to the cost of G1 plus the cost of G3.

- OSPF take both hop count and bandwidth into account in cost calculation. Hence, it is a more reliable routing protocol than RIP.
- How do OSPF routers exchange link state information and synchronize their LSDBs?

Contents

HUAWEI

OSPF Packet Header



- RIP operates on UDP port number 520, while OSPF is defined above IP rather than TCP or UDP, which means OSPF packets are directly encapsulated in IP packets, with protocol number 89.
- All OSPF packets have the same format of header:
  - Version: The value is set to 2 if OSPF version 2 is in use.
  - Type: indicates the type of OSPF packet.
  - Packet length: indicates the length of entire OSPF packet, in bytes.
  - Router ID: indicates the identity of the router that originates this packet.
  - Area ID: indicates the OSPF area into which the packet is being advertised.
  - Checksum: is used to verify data integrity of the entire OSPF packet, including the OSPF packet header.
  - Auth Type: indicates the authentication mode being used. The value 0 indicates non-authentication. The value 1 indicates simple password. The value 2 indicates cryptographic (MD5) authentication.
  - Authentication: contains the information necessary for authentication. The content of this field varies according to the AuType field.
- OSPF packet header defines the communication standards and rules between OSPF routers. Based on these standards, what functions need to be developed for OSPF packets?

OSPF Packet Types

Type	Packet Type	Packet Function
1	Hello	To discover and maintain OSPF neighbor relationships.
2	Database Description (DD)	To exchange LSDB summary.
3	Link State Request (LSR)	To request specific link state information.
4	Link State Update (LSU)	To send requested link state information.
5	Link State Ack (LSAck)	To acknowledge receipt of an LSA.

- Consider what information do DD, LSR, LSU, and LSAck packets contain and why the information is contained?

  HUAWEI

- Type 1 packets are Hello packets, which are used to establish and maintain neighbor relationships. As discussed earlier, before establishing an OSPF neighbor relationship, two routers must negotiate some parameters.
- Type 2 packets are Database Description (DD) packets, which are used to describe the content of local LSDB to neighbors so that the neighbors can determine whether their LSDBs are complete.
- Type 3 packets are Link State Request (LSR) packets. A router determines whether its LSDB is complete compared with the information in DD packets from its neighbors. If its LSDB is incomplete, it creates a list of LSAs that need to be obtained from the sender of the DD packets and then sends.
- Type 4 packets are Link State Update (LSU) packets, which are used to respond to the LSR packets received from neighbors. According to the request list in the received LSR packet, the router packs the required LSAs into LSU packets and sends them to the neighbor. LSA flooding is performed by LSU packets, and as a result, the LSDB synchronization among all routers in the area is implemented.
- Type 5 packets are Link State Acknowledgment (LSAck) packets, which are used to acknowledge received LSAs in order to ensure LSA synchronization reliability.
- DD, LSR, LSU, and LSAck packets contain LSA information in varying degrees:
  - A DD packet contains LSA header information, including LS Type, LS ID, Advertising Router, LS Sequence Number, and LS Checksum.

- An LSR packet contains LS Type, LS ID, and Advertising Router.
- Each LSU packet carries a collection of LSAs.
- An LSAck packet contains LSA header information, including LS Type, LS ID, Advertising Router, LS Sequence Number, and LS Checksum.
- All these five types of OSPF packets ensure an efficient LSA synchronization process. So how are these packets utilized to fulfill the task?

Functional Requirement of OSPF Packets

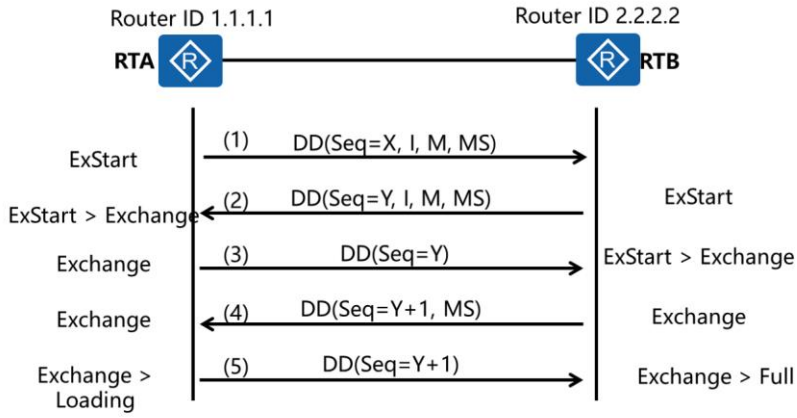
Function to Be Achieved	Requirement Implementation Analysis
Neighbor relationship discovery and maintenance	Use the Hello mechanism.
LSA synchronization	Enable two routers to send LSAs to each other for LSA synchronization. LSA synchronization is fast and fewer resources are consumed.
Reliability	Ensure reliable LSA synchronization.

- In previous parts of this series, the Hello mechanism used to dynamically discover neighbors and maintain OSPF neighbor relationships has been explained and need not be repeated here.
- Compared with the Request and Response messages used for synchronization of routing information bases (RIBs) in RIP, LSAs are employed to synchronize LSDBs in OSPF. One possible way for an OSPF router is to send all its LSAs to neighbors; however, this method is not a perfect one.
- A faster and more efficient method should be like this: summary information rather than all link-state information is sent to neighbors. On receiving it, the neighbors examine which LSAs are new or more up-to-date to themselves, and then request detailed LSAs from the sender. For OSPF protocol, there is necessity to employ a more efficient and reliable method than RIP’ s to synchronize link-state database between routers.

Contents

HUAWEI

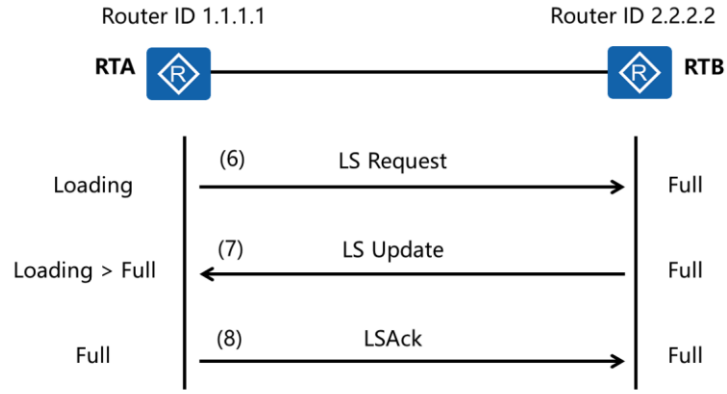
OSPF LSDB Synchronization (1)



- State description:
  - ExStart: This state occurs when two neighboring routers establish a master/slave relationship and determine the initial DD sequence number. In this state, the exchanged DD packets don't include any LSAs.
  - Exchange: This state occurs when a router exchanges DD packets containing link state summary information with its neighbor.
  - Loading: This state occurs when two routers send LSR, LSU, and LSAck packets to each other. This state occurs when a router finds entries in its Link State Request list. In this state, the router sends LSR packets to its neighbor, receives LSU packets from it and answers LSAck packets as acknowledgement.
  - Full: This state occurs when the neighboring router's LSDB is synchronized. In this state, the neighboring routers are fully adjacent.
- LSDB synchronization process:
  - RTA and RTB have router IDs 1.1.1.1 and 2.2.2.2 respectively, and a neighbor relationship has been established between them. When the neighbor RTB's state becomes ExStart, RTA sends its first DD packet to RTB. The DD packet sequence number is randomly set to X, the I-bit is set to 1, indicating that this is its first DD packet; the M-bit is set to 1, indicating that subsequent DD packets need to be sent, and the MS-bit is set to 1, indicating that RTA asserts itself as the master.

- Similarly, when the neighbor RTA’ s state becomes ExStart, RTB sends its first DD packet, with a DD sequence number of y and three flags I-bit, M-bit and MS-bit set to 1, 1 and 1 respectively, as the same meaning as RTA’ s. Because RTB has a larger router ID, RTB will become the master. Upon receipt of the DD packet from RTB, RTA generates a Negotiation-Done event agreeing that RTB is the master, and transitions neighbor RTB’ s state from ExStart to Exchange.
- Then RTA sends a new DD packet containing summary information of its LSDB. In this packet, the DD packet sequence number is overrode by RTB with Y in step 2, the I-bit is set to 0, indicating that this packet is not its first DD packet, the M-bit is set to 0, indicating that this packet is the last DD packet containing LSDB summary information, and the MS-bit is set to 0, indicating that RTA asserts itself as the slave. When receiving this DD packet, RTB generates a Negotiation-Done event and transitions its neighbor RTA’ s state from ExStart to Exchange.
- Then RTB sends a new DD packet, which contains LSDB summary information together with sequence number Y+1 and the M-bit set to 1, indicating that RTB asserts itself as the master. Suppose all its LSDB summary information is included in this DD packet, the M-bit is now set to 0, telling RTA, "this is the last DD packet for you".
- Although there is no need for RTA to sends its LSDB summary information to RTB, as a slave, it still needs to acknowledge each DD packet from the master. Therefore, RTA responds to RTB’ s DD packet with an empty DD packet containing the sequence number Y+1. Then RTA generates an Exchange-Done event and transitions neighbor RTB’ s state to Loading. When RTB receives this DD packet, it transitions its neighbor state to Full. (Assume that RTB has the latest and complete LSDB and does not need to request updates from RTA.)

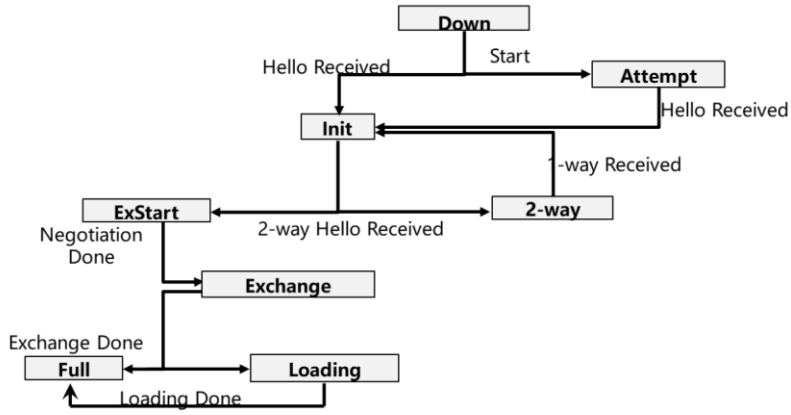
OSPF LSDB Synchronization (2)



- RTA sends an LSR packet to RTB to request more recent LSAs that have been discovered in the Exchange state but have not yet been received.
- RTB then provides RTA with the requested LSAs in an LSU packet. On receiving this LSU packet, RTA transitions neighbor RTB’ s state from Loading to Full.
- Then RTA returns an LSAck packet to RTB as an acknowledgement for the received LSU packet. When RTB receives the LSAck packet, a full adjacency is established between the two routers.

- The process from establishing neighbor relationship to synchronizing LSDB is complex. Any incorrect configuration or link failures will lead to an unsuccessful LSDB synchronization. Understanding the trigger for state transitions is the key to rapid troubleshooting.

OSPF Neighbor State Machines

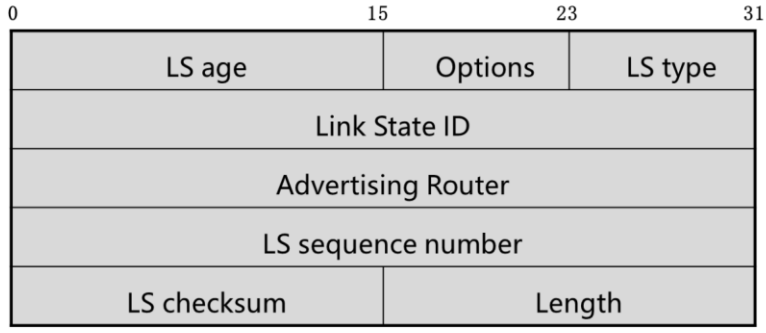


- The series of transitions in the OSPF neighbor state machine is illustrated by the diagram above. The transition process is described as follows:
  - Down: This is the initial state of a neighbor conversation. It indicates that the router has not received any Hello packet from its neighbor in the last RouterDeadInterval. On an NBMA network, a router can still send Hello packets at every PollInterval to its manually configured neighbor which is in Down state. The PollInterval is often as large as the RouterDeadInterval.
  - Attempt: This state only appears on NBMA networks. It indicates that the router has not received any recent information from its manually configured neighbor, but that it has been sending Hello packets to the neighbor at every HelloInterval. If the router does not receive any Hello packet from its neighbor within a RouterDeadInterval, the neighbor state transitions to Down.
  - Init: This state occurs when the router receives a Hello packet from a neighbor but does not see its own Router ID in neighbor list of this Hello packet, which means they have not established a two-way communication with each other. In this state, the router IDs of all its known neighbors must be included in the neighbor list of the Hello packet sent by the router.

- 2-way Received: This event occurs when the router sees its own Router ID in neighbor list of the Hello packet received from the neighbor, or receives a DD packet from the neighbor. If the router needs to establish an adjacency with a neighbor, the neighbor state transitions to ExStart and LSDB synchronization begins. Otherwise, the neighbor state transitions to 2-Way.
- 2-Way: This state occurs when the router establishes a bidirectional communication, but does not yet form an adjacency, with its neighbor. This is the final state to which the neighbor transition, if there is so far no need for establishing adjacency with this neighbor.
- 1-Way Received: This event occurs when the router finds that its router ID is not contained in the neighbor list of the received Hello packet. This is often due to a restart of the neighboring router.
- ExStart: This state occurs when the router starts to send DD packets to its neighbor. In this state, the master-slave relationship is established and the initial DD sequence number is determined. Note that the DD packets do not include any LSA summary information in this state. This is the first step towards forming an adjacency.
- Exchange: This state occurs when the router exchanges with its neighbor DD packets containing LSA summary information to describe its entire LSDB.
- Loading: This state occurs when the router sends LSR packets to its neighbor requesting the more recent LSAs that have been discovered in the received DD packets in the Exchange state. Subsequently, the router receives responding LSU packets from the neighbor.
- Full: This state occurs when its link state request list is empty, which means the local LSDB is synchronized with its neighbor’ s. It indicates that the neighbor is fully adjacent.

LSA Header

- LSAs carry OSPF link state information.

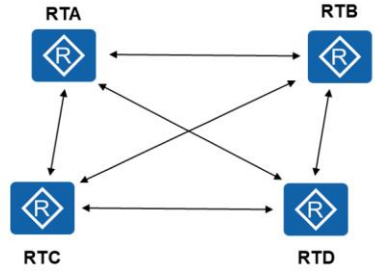


- A Link State Advertisement (LSA) is the carrier of OSPF link state information between routers. The LSA is the basic structural unit of LSDB. In other words, an LSDB is composed of an array of LSAs.
- All LSAs have the same format of header, including the following fields:
  - LS age: indicates the age of the LSA, in seconds.
  - LS type: indicates the LSA format and function. There are five types of commonly used LSAs.
  - Link State ID: identifies the link described by an LSA, for example, router ID.
  - Advertising Router: indicates the router ID of the router that originates this LSA.
  - LS sequence number: is used to detect old and duplicate LSAs.
- LS type, Link State ID, and Advertising Router together identify an LSA.
- Aside from locally originated LSAs, there are LSAs obtained from other neighbors in LSDB. It’ s necessary for LSAs to be advertised between neighboring routers through a passageway.

Contents

Problems in MA Networks

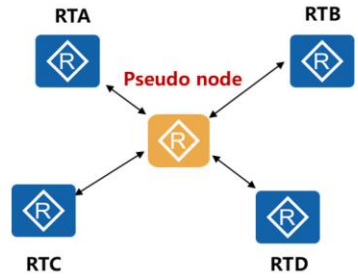
- Managing nx(n−1)/2 adjacencies is complicated.
- Unnecessarily flooding repeated LSAs is a waste of resources.



- OSPF-running MA networks include broadcast and NBMA networks, which are facing the following problems:
  - In a network with N routers, the number of adjacencies can be up to nx(n−1)/2.
  - The same copies of LSAs will be flooded repeatedly among neighbors. For example, RTA sends its LSAs with the same content to neighbors RTB, RTC and RTD separately. And what' s more, RTB forwarded these LSAs to neighbors RTD and RTC, and even RTA where the LSAs are from. This process is called flooding.
- Evidently, such a process is inefficient and resource-hungry. As an advanced routing protocol, how does OSPF solve the two problems?

Responsibilities of DR and BDR

- Reduces the number of adjacencies.
- Reduces OSPF traffic.

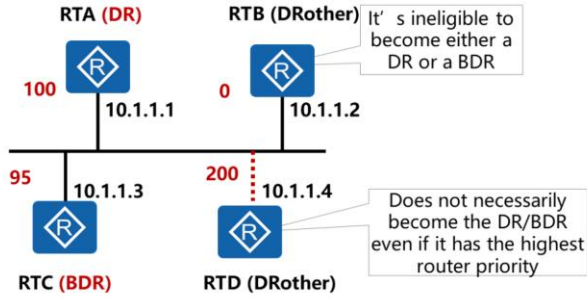


- Question: How to avoid DR as the single point of failure?

- Designated router (DR) is responsible for establishing and maintaining adjacencies and flooding LSAs over an MA network.
- The DR establishes adjacencies and exchanges link state information with all the other routers, while other routes do not directly exchange link state information with each other. This greatly reduces the number of adjacencies in an MA network, and as a result, saves resources used to exchange link state information.
- Once the DR becomes invalid, its adjacencies with other routers all become invalid, and LSDB synchronization fails. A new DR needs to be elected to establish new adjacencies with non-DR routers for LSDB synchronization. To prevent the single point of failure, OSPF puts forward the concept of Backup Designated Router (BDR), which is a backup for DR and elected at the same time as when DR is elected. When the current DR becomes invalid, the BDR instantly takes over the role of DR.
- A pseudo node is a virtual device node. The following describes DR and BDR election.

DR and BDR Election

- Election rules: DR/BDR election occurs among interfaces.
  - An interface with the highest router priority takes precedence.
  - If interfaces have the same router priority, the interface with a larger router ID takes precedence.



Neighbor Relationship and Adjacency

Network Type	Whether to Establish an Adjacency
Point-to-point	Yes
Broadcast	The DR establishes adjacencies with the BDR and DRothers The BDR establishes adjacencies with the DR and DRothers DRothers establish only neighbor relationships with each other
NBMA	
Point-to-multipoint	Yes

- The concepts neighbor relationship and adjacency are different from each other. After two OSPF routers establish a neighbor relationship, they synchronize their LSDBs and eventually establish an adjacency.
- On point-to-point and point-to-multipoint networks, routers that have established a neighbor relationship proceed to establish an adjacency.
- On broadcast and NBMA networks, non-DR/BDR routers (DRothers) establish only neighbor relationships rather than adjacencies with each other; non-DR/BDR routers (DRothers) establish adjacencies with the DR/BDR; the DR and BDR establish an adjacency with each other.
- When an adjacency is established, LSDB synchronization is complete. Then OSPF routers start to use the SPF algorithm to calculate the shortest paths to all known destinations on the basis of the information in their LSDBs.

Quiz

1. Which of the following packets are OSPF packets?

   Hello

   A. Database Description

   B. Link State Request

   C. Link State DD

   D. Link State Advertisement

2. Describe OSPF network types.

HUAWEI

- Answer: ABC.
- Answer: point-to-point network, point-to-multipoint network, broadcast network, NBMA network.


Thank You
www.huawei.com

OSPF Intra-Area Routing

Foreword

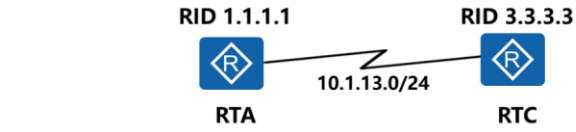
- This topic is about how OSPF calculates intra-area routes, including how to use Router-LSAs and Network-LSAs to describe topologies and routes, and how to build a shortest path tree (SPT).

Objectives

- Upon completion of this section, you will be able to:
  - Be familiar with the Router-LSA contents and functions
  - Be familiar with the Network-LSA contents and functions
  - Understand the shortest path first (SPF) algorithm

Contents

A Router-LSA Describing a Point-to-Point Interface




```
<RTA>display ospf lsdb router self-originate

Type      : Router          //LSA type
Ls id     : 1.1.1.1         //Link state ID
Adv rtr   : 1.1.1.1         //ID of the router that generates the LSA

 * Link ID: 3.3.3.3         //ID of the neighbor router
   Data   : 10.1.13.1       //IP address of the router interface that advertises the LSA
   Link Type: P-2-P
   Metric : 48
 * Link ID: 10.1.13.0       //IP address of this stub network
   Data   : 255.255.255.0   //Subnet mask of this stub network
   Link Type: StubNet
   Metric : 48              //Cost value
   Priority : Low
```

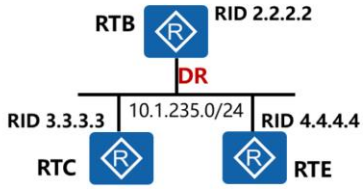
Topology information

Routing information

- An OSPF router originates a Router-LSA for each area to which the router belongs, to describe the states of the router's links. The three fields in the LSA header are as follows:
  - Type: LSA type. A Router-LSA is the Type 1 LSA.
  - LS id: link state ID.
  - Adv rtr: ID of the router that originates this Router-LSA.
- One Router-LSA can describe multiple links with fields Link ID, Data, Link Type, and Metric for each link. The meanings of keywords are as follows:
  - Type: link type (not the same concept as the network type mentioned in previous sections). A Router-LSA describes the following types of links:
    - Point-to-Point: describes a single connection directly to the neighboring router. The point-to-point link type is treated as topology information.
    - TransNet: describes a link to a transit network, to which two or more routers are attached, such as a broadcast network and non-broadcast multi-access (NBMA) network. The TransNet link type is treated as topology information.
    - StubNet: describes a link to a stub network, to which only a single router is attached, such as a loopback interface. The StubNet link type is treated as routing information.

- Link ID: peer ID at the other end of the link. The meaning of the Link ID field varies according to the link type.
- Data: additional information about a link. Different types of links have different additional information.
- Metric: cost of a link.

A Router-LSA Describing a Broadcast or an NBMA Interface




```
<RTC>display ospf lsdb router self-originate

Type      : Router      //LSA type
Ls id     : 3.3.3.3     //Link state ID
Adv rtr   : 3.3.3.3     //ID of the router that generates the LSA

* Link ID: 10.1.235.2   //Interface IP address of a designated router
  Data   : 10.1.235.3   //IP address of the router interface that advertises the LSA
  Link Type: TransNet
  Metric : 1
```

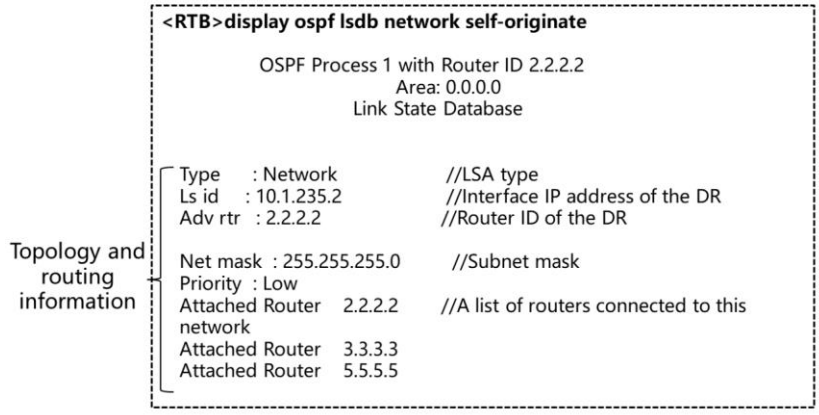
Topology information

- Question: Where is the network ID or subnet mask?

- In a Router-LSA that describes a broadcast or an NBMA interface, Link ID specifies the interface IP address of a designated router (DR), and Data specifies the IP address of the router interface that advertises this LSA.
- As shown in the figure, RTB, RTC, and RTE are attached to the same Ethernet segment. Take the LSA originated by RTC as an example. The Link ID value is 10.1.235.2, which is the interface IP address of the DR. The Data value is 10.1.235.3, which is the IP address of the router interface connected to the MA network. The Link Type value is TransNet, and the Metric value is the cost of the link to the DR.
- The TransNet link described by a Router-LSA only contains information about adjacency with the DR and the link cost, without any information about the network ID, subnet mask, and other routers on the multi-access network.

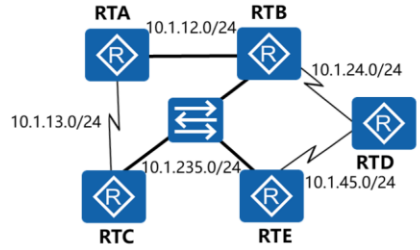
Contents

A Network-LSA Describing a Broadcast or an NBMA Network



- The network-LSA describes all the routers that are attached to a broadcast or an NBMA network with the interface IP address of the DR, the network's address mask, and a list of IDs of all routers that are fully adjacent to the DR.
- The meanings of key fields in a Network-LSA are as follows:
  - Type: LSA type. A Network-LSA is the Type 2 LSA.
  - LS id: interface IP address of the DR.
  - Adv rtr: ID of the router that originates this Network-LSA, i.e., the DR's router ID.
  - Net mask: subnet mask for this network.
  - Attached Router: a list of routers connected to this network, which reveals the topology of this network.
  - The Ls id masked by the subnet mask yields the network ID. The cost from the DR to any non-DR routers is 0.
- As seen in the Attached Router field, routers 2.2.2.2, 3.3.3.3, and 5.5.5.5 are attached to the same multi-access network with the network ID of 10.1.235.0 and subnet mask of 255.255.255.0, where the DR is 2.2.2.2.

LSDB in an OSPF Area



<RTA>display ospf lsdb

OSPF Process 1 with Router ID 1.1.1.1
Link State Database
Area: 0.0.0.0

Type	LinkState ID	AdvRouter	Age	Len	Sequence	Metric
Router	4.4.4.4	4.4.4.4	1436	72	80000007	48
Router	2.2.2.2	2.2.2.2	1305	72	80000019	1
Router	1.1.1.1	1.1.1.1	1304	60	80000007	1
Router	5.5.5.5	5.5.5.5	1326	60	80000017	1
Router	3.3.3.3	3.3.3.3	1325	60	8000000F	1
Network	10.1.235.2	2.2.2.2	1326	36	80000004	0
Network	10.1.12.2	2.2.2.2	1305	32	80000001	0

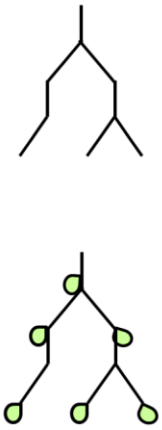
HUAWEI

- LSDB stands for link-state database.
- As shown in the figure, the OSPF area contains five routers. Take the LSDB of RTA as an example. Each LSDB of five routers is identical to one another: five Router-LSAs, each of which is originated by every router severally; two Network-LSAs, which are originated by the DR for two broadcast networks separately.

Contents

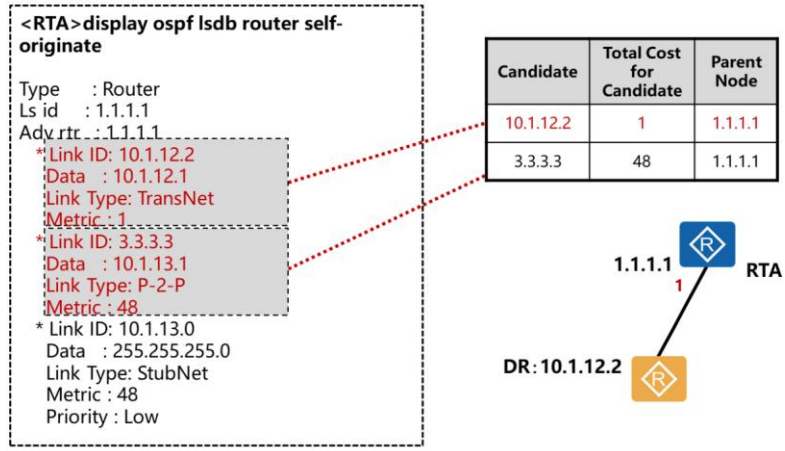
SPF Algorithm

- Phase 1: Create a shortest path tree (SPT).
  - Create an SPT on the basis of topology information in Router-LSAs and Network-LSAs.
- Phase 2: Calculate an optimal route.
  - Calculate an optimal route on the basis of the SPT and the routing information in Router-LSAs and Network-LSAs.



- Type 1 and Type 2 LSAs carry topology and routing information.
- OSPF calculates an SPT using the SPF algorithm and various types of LSAs.
  - Phase 1: Create an SPT on the basis of Type 2 LSAs, as well as point-to-point and TransNet links in Type 1 LSAs.
  - Phase 2: Calculate an optimal route on the basis of Type 2 LSAs and StubNet links in Type 1 LSAs.

Creating an SPT (1)



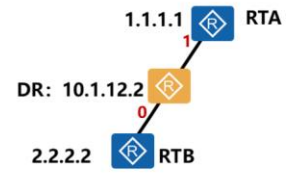
  HUAWEI

- Each OSPF router calculates an SPT with itself as the root node.
- The following example describes how RTA calculates an SPT:
  - First of all, RTA places itself in the root position of an SPT. Then it looks up the links in Router-LSA it originated. All links except StubNet links described by a Router-LSA can be selected as candidates. The ID of the selected link is put in the column Candidate; the total cost of the candidate equals its parent's cost (that is, the sum of Metric values along the path from the root to the parent) plus its own Metric value.
  - The Router-LSA from the root RTA itself describes two types of links: TransNet link and Point-to-point link. The TransNet link 10.1.12.2 has Metric value 1, and the Point-to-point link 3.3.3.3 has Metric value 48. They are both added to the candidate list.
  - RTA adds node 10.1.12.2, with the lowest total cost, to the SPT, and deletes it from the candidate list. This TransNet link 10.1.12.2, performing the role of DR, is considered as a "pseudonode" (or a virtual router) to represent the Transit Network. In this sense, all the routers attached to the link are attached to that node with cost of 0.

Creating an SPT (2)

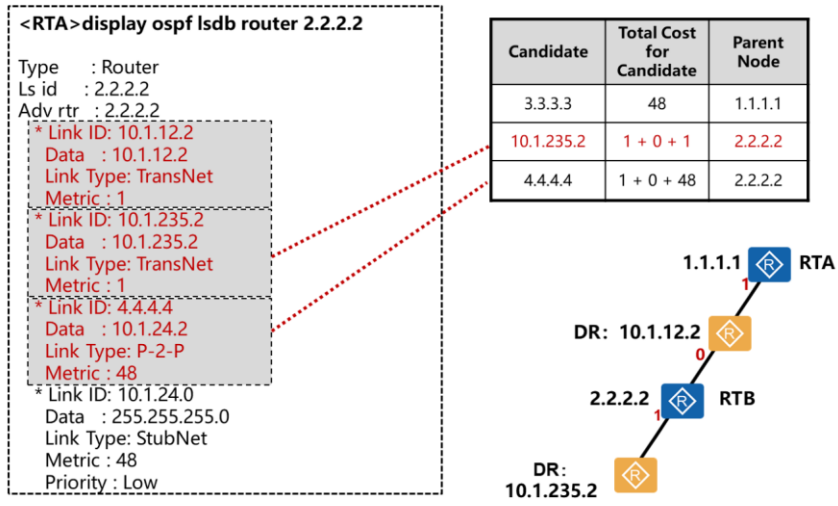
```
<RTA>display ospf lsdb network 10.1.12.2

Type      : Network
Ls id     : 10.1.12.2
Adv rtr   : 2.2.2.2

Net mask  : 255.255.255.0
  Priority  : Low
    Attached Router      2.2.2.2
    Attached Router      1.1.1.1
```

Candidate	Total Cost for Candidate	Parent Node
3.3.3.3	48	1.1.1.1
2.2.2.2	1 + 0	10.1.12.2



- After the DR is added to the SPT, RTA examines the DR's network-LSA and finds a list of attached routers. If an attached router listed in the LSA has already been on the SPT, this router is ignored.
- As shown in the figure above, two attached routers are listed:
  - The attached router 1.1.1.1 is ignored and is not able to be added to the candidate list, because it has already been on the SPT.
  - As for the other attached router 2.2.2.2, the total cost is 1, which equals the sum of its parent's cost (1) and its Metric value (0). The router is then added to the candidate list.
  - Now, there are two candidates in the candidate list. Router 2.2.2.2 whose cost is the lowest among all candidates, is added to the SPT and removed from the list.

Creating an SPT (3)

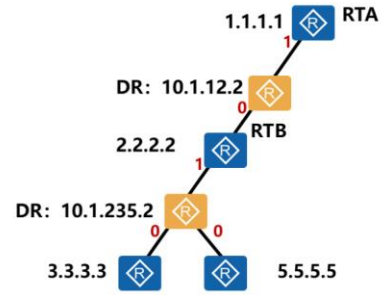


- After node 2.2.2.2 is added to the SPT, RTA continues to examine the Router-LSA of this node.
  - The first link is a TransNet link with Link ID of 10.1.12.2. However, it is ignored and not added to the candidate list, for the reason that the node 10.1.12.2 has already been added to the SPT in step 1.
  - The second link 10.1.235.2 is also a TransNet link. Its Metric value is 1 and its parent's cost is 1. Thus, the total cost of the link is 2 (1 + 1). Node 10.1.235.2 is added to the candidate list.
  - As for the third link, it is a point-to-point link with Link ID of 4.4.4.4. Given the Metric value 48 and its parent's cost 1, its total cost is 49 (48 + 1). Node 4.4.4.4 is added to the candidate list.
- As a result, there are three candidates in the candidate list. As seen, the candidate 10.1.235.2 (the DR) has the lowest total cost and therefore is added to the SPT. Then the candidate is removed from the candidate list.

Creating an SPT (4)

```
<RTA>display ospf lsdb network
10.1.235.2

  Type      : Network
  Ls id     : 10.1.235.2
  Adv rtr   : 2.2.2.2

Net mask  : 255.255.255.0
  Priority  : Low
    Attached Router    2.2.2.2
    Attached Router    3.3.3.3
    Attached Router    5.5.5.5
```

Candidate	Total Cost for Candidate	Parent Node
~~3.3.3.3~~	48	~~1.1.1.1~~
4.4.4.4	1 + 48	2.2.2.2
3.3.3.3	1 + 0 + 1 + 0	10.1.235.2
5.5.5.5	1 + 0 + 1 + 0	10.1.235.2

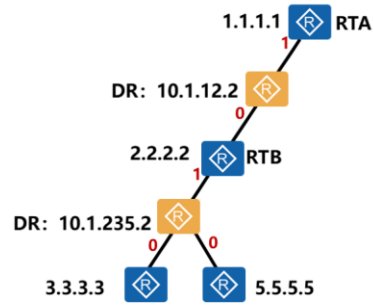


- Next, RTA examines the network-LSA from the DR 10.1.235.2 and finds a list of attached routers.
- As shown in the figure above, three attached routers are listed:
  - Attached router 2.2.2.2 is ignored because it has already been added to the SPT.
  - As mentioned earlier, all the routers attached to the "pseudonode" have the Metric value of 0. The total cost of router 3.3.3.3 is 2 (2 + 0), the same as its parent's. This router is now added to the candidate list. (If two candidates with the same ID appear in the list, the total costs of both must be compared with each other. The one with the lower cost wins and therefore remains in the list, while the loser is removed. According to this rule, the link 3.3.3.3, whose total cost is 48, is deleted.)
  - Similarly, the total cost of router 5.5.5.5 is 2 and the router is added to the candidate list.
  - At the moment, there are three candidates in the list. The nodes 3.3.3.3 and 5.5.5.5 are added to the SPT because they both have the lowest total costs 2, then are deleted from the candidate list.

Creating an SPT (5)

```
<RTA>display ospf lsdb router 3.3.3.3

Type      : Router
Ls id     : 3.3.3.3
Adv rtr   : 3.3.3.3
 * Link ID: 10.1.235.2
   Data   : 10.1.235.3
   Link Type: TransNet
   Metric : 1
 * Link ID: 1.1.1.1
   Data   : 10.1.13.3
   Link Type: P-2-P
   Metric : 48
 * Link ID: 10.1.13.0
   Data   : 255.255.255.0
   Link Type: StubNet
   Metric : 48
   Priority : Low
```

Candidate	Total Cost for Candidate	Parent Node
4.4.4.4	1 + 48	2.2.2.2



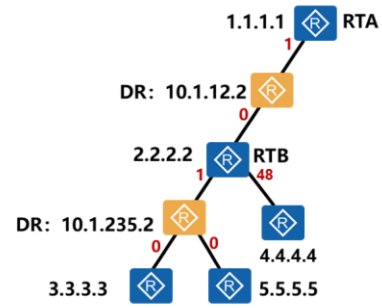
- After the nodes 3.3.3.3 and 5.5.5.5 are added to the SPT, the examination of Router-LSAs continues.
- As to the Router-LSA from node 3.3.3.3:
  - Since the two nodes 10.1.235.2 and 1.1.1.1 have already been added to the SPT before, both links are ignored.

Creating an SPT (6)

<RTA>display ospf lsdb router 5.5.5.5

Type : Router
Ls id : 5.5.5.5
Adv rtr : 5.5.5.5
* Link ID: 10.1.235.2
Data : 10.1.235.5
Link Type: TransNet
Metric : 1
* Link ID: 4.4.4.4
Data : 10.1.45.5
Link Type: P-2-P
Metric : 48
* Link ID: 10.1.45.0
Data : 255.255.255.0
Link Type: StubNet
Metric : 48
Priority : Low

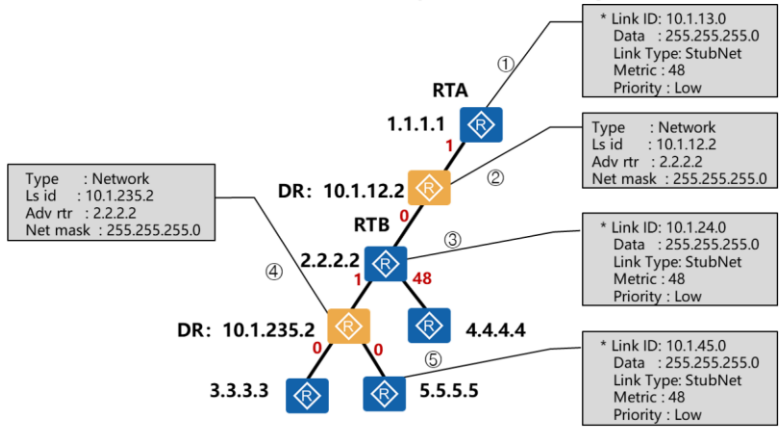
Candidate	Total Cost for Candidate	Parent Node
4.4.4.4	1 + 48	2.2.2.2
~~4.4.4.4~~	~~1 + 0 + 1 + 0 + 48~~	~~5.5.5.5~~



- As to the Router-LSA from node 5.5.5.5:
  - The node 10.1.235.2 has already been added to the SPT, so the TransNet link 10.1.235.2 is ignored.
  - As for the point-to-point link 4.4.4.4, obviously, the same link has been in the candidate list. The total cost of this link is 50, which is the sum of its Metric value (48) and its parent's cost (2). Seeing that 50 is greater than 49, which means the node whose parent is node 2.2.2.2 has the shorter path to the root than the one whose parent is node 5.5.5.5, the current link is removed from the list. Then the node 4.4.4.4 with the total cost of 49 is added to SPT as a child of node 2.2.2.2 and deleted from the list.
- Now, when you run the display ospf lsdb router 4.4.4.4 command to continue exploring, you will find that all the neighbors of router 4.4.4.4 described in the Router-LSA have been added to the SPT. No new candidate will be added to the list.
- At this time, the candidate list is empty and SPF calculation is done. Note that, the nodes 10.1.12.2 and 10.1.235.2 serve as pseudonodes.

Optimal Route Calculation

- Picks up the routing information from LSAs of all nodes in the SPT.
- Starts with the root in the same sequence as they were added to the SPT.



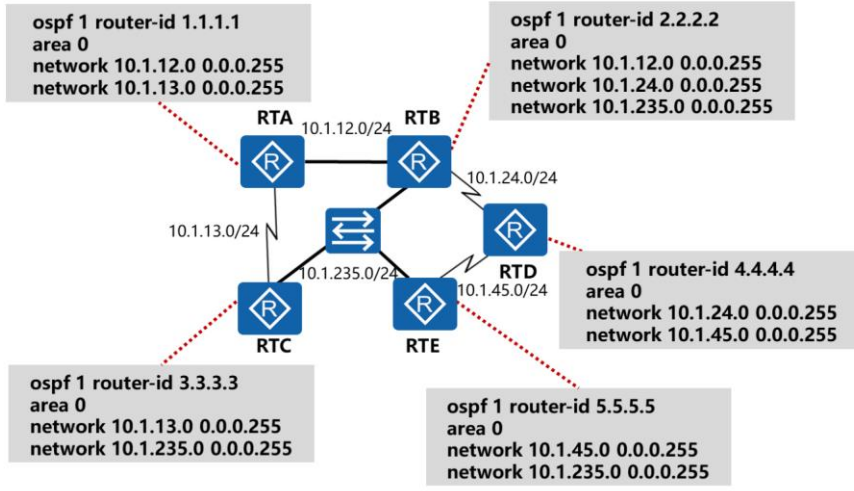
- Second phase: Calculate the optimal route according to StubNet link information in Router-LSAs and routing information in Network-LSAs.
- OSPF picks up the routing information from LSAs of all nodes in the SPT, starting with the root in the same sequence as they were added to the SPT.
  - In the Router-LSA from RTA (node 1.1.1.1), one StubNet link is found. The subnet is 10.1.13.0/24 and Metric is 48.
  - In the Network-LSA from the DR (node 10.1.12.2), the subnet is 10.1.12.0/24 and Metric is 1 (1 + 0).
  - In the Router-LSA from RTB (node 2.2.2.2), one StubNet link is described. The subnet is 10.1.24.0/24 and Metric is 49 (1 + 0 + 48).
  - In the Network-LSA from the DR (node 10.1.235.2), the subnet is 10.1.235.0/24 and Metric is 2 (1 + 0 + 1).
  - In the Router-LSA from RTC (node 3.3.3.3), one StubNet link is found, and the subnet is 10.1.13.0/24. The link already exists on RTA, so this link is ignored.
  - In the Router-LSA from RTE (node 5.5.5.5), one StubNet link is found. The subnet is 10.1.45.0/24 and Metric is 50 (1 + 0 + 0 + 1 + 48).
  - In the Router-LSA from RTD (node 4.4.4.4), two StubNet links are found. The subnets of the StubNet links are 10.1.24.0/24 and 10.1.45.0/24, respectively. However, the same destination address prefixes have already been picked up from node 2.2.2.2 and node 5.5.5.5 separately in the previous steps. They are ignored because the path of node 4.4.4.4 is longer than node 2.2.2.2 and node 5.5.5.5.

Viewing the OSPF Routing Table

```
<RTA>display ospf routing

        OSPF Process 1 with Router ID 1.1.1.1
                Routing Tables

Routing for Network
Destination     Cost    Type    NextHop     AdvRoute    Area
10.1.12.0/24    1       Transit 10.1.12.1   1.1.1.1     0.0.0.0
10.1.13.0/24    48      Stub    10.1.13.1   1.1.1.1     0.0.0.0
10.1.24.0/24    49      Stub    10.1.12.2   2.2.2.2     0.0.0.0
10.1.45.0/24    50      Stub    10.1.12.2   5.5.5.5     0.0.0.0
10.1.235.0/24   2       Transit 10.1.12.2   2.2.2.2     0.0.0.0

Total Nets: 5
Intra Area: 5  Inter Area: 0  ASE: 0  NSSA: 0
```


- After two phases of SPF calculation, OSPF routes are generated on RTA, as shown in the figure above.
- However, a route calculated by OSPF is not always installed into the system routing table. Because the routes for the same destination may be obtained from different routing protocols, each of which has its unique route preference value (a.k.a administrative distance), the router needs to compare their route preferences to determine which route to be installed or to be discarded.

Example for Configuring a Single Area OSPF Network



Viewing OSPF Neighbor Status

```
<RTA>display ospf peer brief

        OSPF Process 1 with Router ID 1.1.1.1
                Peer Statistic Information
----------------------------------------------------------------------------
Area Id          Interface                     Neighbor id      State
0.0.0.0          GigabitEthernet0/0/0          2.2.2.2          Full
0.0.0.0          Serial1/0/0                   3.3.3.3          Full
----------------------------------------------------------------------------
```


- In this output, we can see that RTA has established adjacencies with RTB and RTC separately.

Quiz

1. What types of links can be described by a Router-LSA?
2. Will an OSPF-derived route be certainly installed into a router's routing table?

- Answer: Point-to-point, TransNet, StubNet, and vlink links.
- Answer: Not always. A router may obtain multiple routes with the same destination address prefix from other routing protocols. A comparison of route preferences must be performed before the most preferable route is selected to the routing table.


Thank You
www.huawei.com

OSPF Inter-Area Routing

Foreword

- As a network grows, with the structural complexity increasing, routers consume more memory and CPU resources for route calculation.
- Moreover, the network failure rate also rises. If a fault happens in an area, all routers in the area need to recalculate routes. This may place a heavy burden on the routers and degrades the network stability.
- How does the OSPF protocol work to handle problems caused by an overly large area?

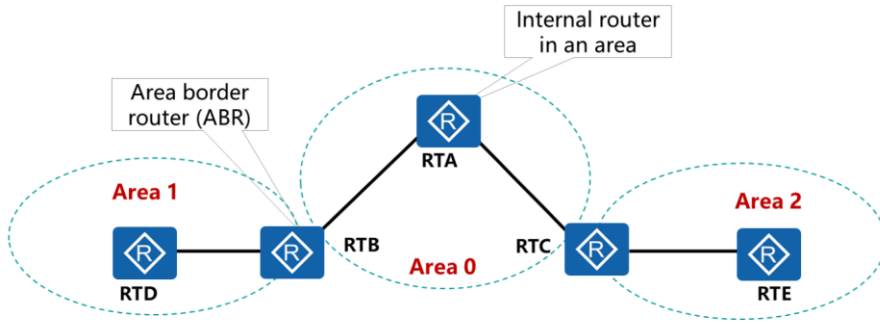
Objectives

- Upon completion of this section, you will be able to:
  - Be familiar with the inter-area route transmission process
  - Understand how to prevent inter-area routing loops
  - Understand how to configure virtual links

Contents

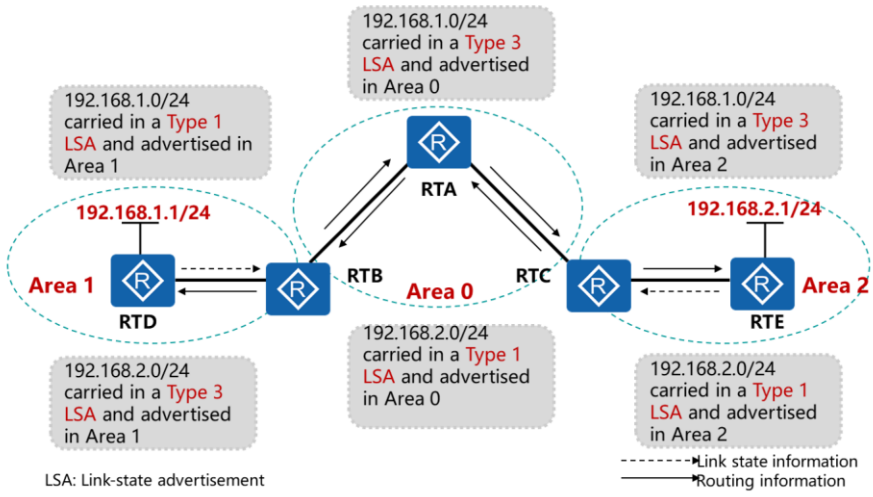
HUAWEI

Area Division



- OSPF divides a large network into multiple interconnected smaller areas. Routers in the same area only need to synchronize their Link-State Databases (LSDBs) in the area rather than to flood LSAs to the entire OSPF domain, in order to reduce the consumption of memory and CPU to some extent.
- After an OSPF domain is divided into areas, the routers are further divided into two roles according to functions:
  - Internal router: All interfaces of an internal router belong to the same OSPF area.
  - Area border router (ABR): Interfaces of an ABR belong to two or more OSPF areas.
- An internal router maintains a single LSDB and calculates the routes in the area.
- How do OSPF routers residing in different areas communicate with each other?

Inter-Area Route Transmission



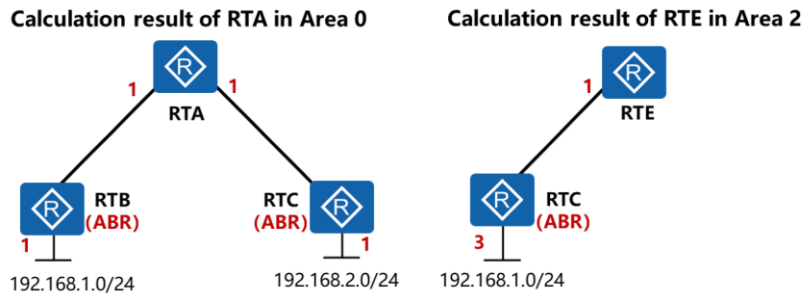
- An ABR acts as a gateway for inter-area traffic and maintains multiple LSDBs, one for each attached area.
- An ABR converts link state information in one of its attached areas to routing information, and then advertise it to another area it is connected to.
- The conversion of link state information into routing information is actually the process of converting Type 1 or Type 2 LSAs into Type 3 LSAs. Note that an ABR transmits inter-area routing information bidirectionally.
- In the above figure, for example, RTD originates a Type 1 LSA to describe the subnet 192.168.1.0/24 and floods it throughout Area 1 for synchronization of the LSDB. As the ABR connecting Area 0 and Area 2, RTB converts the Type 1 LSA into a Type 3 LSA, and sends it into Area 0. Acting as the ABR between Area 0 and Area 2, RTC generates another Type 3 LSA and sends it into Area 2. Till now, the routing information about subnet 192.168.1.0/24 has been advertised throughout the entire OSPF domain. In the same way, the routing information about subnet 192.168.2.0/24 has been advertised.

Network-Summary-LSA

```
<RTB>display ospf lsdb summary 192.168.1.0
            OSPF Process 1 with Router ID 2.2.2.2
                    Area: 0.0.0.0
                Link State Database
 Type      : Sum-Net                 //Type 3 LSA
 Ls id     : 192.168.1.0             //Destination network segment address
 Adv rtr   : 2.2.2.2                 //ID of the router that generates the LSA
 Ls age    : 86
 Len       : 28
 Options   :  E
 seq#      : 80000001
 chksum    : 0x7c6d
 Net mask  : 255.255.255.0           //Subnet mask
 Tos 0  metric: 1                    //Cost value
 Priority  : Low
```


- A Network-Summary-LSA (Type 3 LSA) contains the following fields:
  - Ls id: destination network address
  - Adv rtr: ID of the ABR
  - Net mask: subnet mask of the destination network
  - Metric: cost of the route from the ABR to the destination network
- How does an internal router calculate the inter-area route when it receives a Type 3 LSA describing a destination network in another area?

Inter-Area Route Calculation

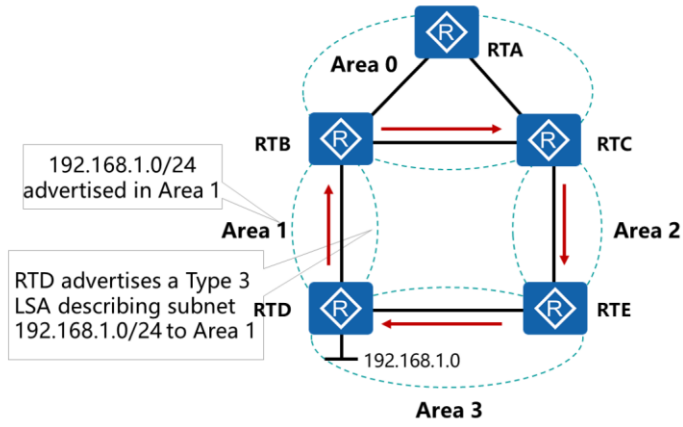


HUAWEI

- A Type 3 LSA originated by an ABR is used to calculate an inter-area route.
  - An ABR is determined according to the value of the field Adv rtr in a Type 3 LSA.
  - A destination address prefix and the cost of the route from an ABR to that destination can be obtained from the fields Ls id, Net mask, and Metric.
  - When a router receives multiple Type 3 LSAs with the same destination address prefix from different ABRs, the router separately adds the cost from itself to each ABR and the metric reported in related LSA, and has them compared with each other. The route with the lowest cost is generated. If two or more routes with the same total cost exist, the router performs equal-cost load-balancing.
  - As shown in the figure, the following describes how RTA in Area 0 calculates the inter-area route:
  - In the Type 3 LSAs describing subnets 192.168.1.0/24 and 192.168.2.0/24, the values of the Adv rtr field are 2.2.2.2 (RTB) and 3.3.3.3 (RTC), respectively, indicating the corresponding ABRs that originate the LSAs.
  - In the Type 3 LSA that RTB originates, the route for subnet 192.168.1.0/24 has the cost 1. The route for subnet 192.168.2.0/24 in the Type 3 LSA generated by RTC has the same cost of 1.
  - If traffic starting from RTA is destined for subnet 192.168.1.0/24, the next hop is RTB and the cost of the route is 2. Similarly, the next hop for the traffic from RTA to the subnet 192.168.2.0/24 is RTC, and the cost is 2.

Contents

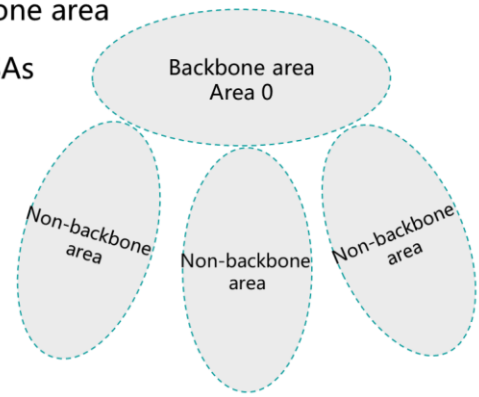
Generation of Inter-Area Routing Loops



- RTB converts a Type 1 or a Type 2 LSA in Area 1 into a Type 3 LSA, and then advertises it to Area 0.
- RTC regenerates a Type 3 LSA describing subnet 192.168.1.0/24 and advertises the LSA to Area 2.
- Similarly, RTE generates a Type 3 LSA describing subnet 192.168.1.0/24 and advertises the LSA to Area 3.
- Once again, RTD then advertises the route for subnet 192.168.1.0/24 to Area 1 in a Type 3 LSA. A routing loop then forms.

Inter-Area Routing Loop Prevention

- Backbone area and non-backbone area
- Transmission rules of Type 3 LSAs

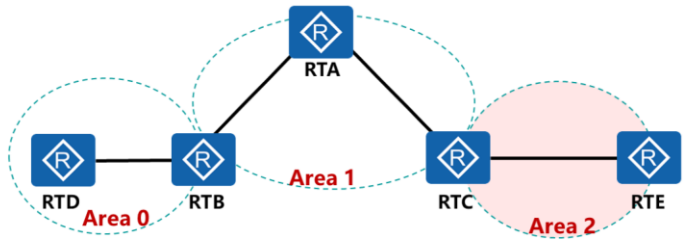


- Question: What will happen if the area ID is set to a non-zero value when only one area exists?

- To prevent inter-area routing loops, OSPF defines the backbone area, non-backbone area, and transmission rules of Type 3 LSAs.
  - OSPF divides an autonomous system into a backbone area (also called Area 0) and a number of non-backbone areas. All non-backbone areas must be connected to the backbone area, and all traffic between non-backbone areas must pass through the backbone area.
  - OSPF does not allow Type 3 LSAs from the backbone area to be injected back to the backbone area.
- As for the ABR mentioned earlier, OSPF requires that an ABR should have at least one interface belonging to the backbone area.
- Following the inter-area routing loop prevention rules, a new network can be deployed free of routing loops among areas. However, due to improper network planning, the connection between areas may not comply with the inter-area routing loop prevention rules.

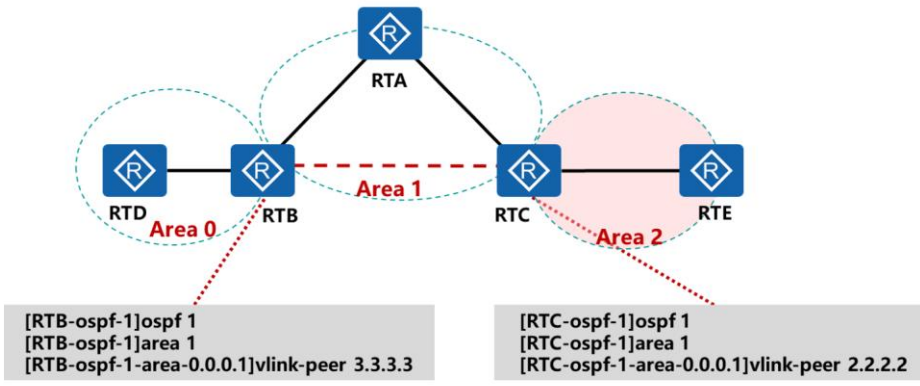
Contents

Improper OSPF Area Design



- If we fail to comply with the inter-area routing loop prevention rules of OSPF, is there any remedy for it?

Virtual Links



- The backbone area must be contiguous. However, it needs not be physically contiguous; it can be logically contiguous by establishing virtual links.
- Virtual links can be set up between any two ABRs that have an interface connected to the same non-backbone area.
- As shown in the figure, a virtual link is set up between RTB and RTC to enable Area 2 to connect to the backbone area (Area 0) through Area 1, as if the Area 2 were directly connected to the backbone area.

Quiz

1. Can one Network-Summary-LSA describe multiple routes?
2. How does OSPF prevent inter-area routing loops?

- Answer: One Network-Summary-LSA can describe only one route.
- Answer: OSPF partitions areas into the backbone area and non-backbone areas. All non-backbone areas must be connected to the backbone area, and only one backbone area exists. Routing information between non-backbone areas must be forwarded through the backbone area. Type 3 LSAs from the backbone area are not be transmitted back to the backbone area.


Thank You
www.huawei.com

OSPF External Routing

Foreword

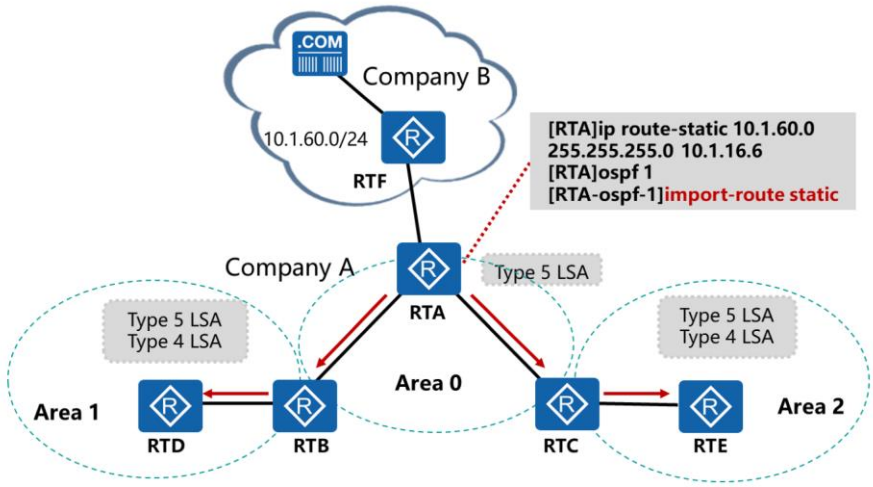
- In addition to internal communication, enterprises require communication with external networks.
- Suppose the OSPF protocol has been deployed in company A's enterprise network to enable internal communication. When there is a need to visit the Web server situated on the enterprise network of company B, as a network engineer of the company A, how could you import routing information from company B's enterprise network to accomplish your goal?

Objectives

- Upon completion of this section, you will be able to:
  - Understand the functions of AS-External-LSAs and ASBR-Summary-LSAs
  - Be familiar with the calculation method of OSPF external routes
  - Understand the cause of suboptimal external routes

Contents

Importing External Routes



- In this example, a static route for destination subnet 10.1.60.0/24 is configured on RTA. The next hop is RTF.
- RTA redistributes the static route into OSPF protocol and advertises it into Company A’ s enterprise network. An OSPF router that imports external routes is called autonomous system boundary router (ASBR). (Communication between devices requires bidirectional routes between them. This course only describes how an OSPF network obtains external routes.)
- RTA originates an AS-External-LSA (Type 5 LSA), which describes a route to a destination external to the Autonomous System via the ASBR. RTB and RTC separately originate ASBR-Summary-LSAs (Type 4 LSAs) to describe routes from themselves to the ASBR.
- Type 4 and Type 5 LSAs are used by routers to calculate external routes.

AS-External-LSA

```
<RTA>display ospf lsdb ase self-originate

            OSPF Process 1 with Router ID 1.1.1.1
                    Link State Database
Type      : External    //LSA type
Ls id     : 10.1.60.0    //Destination network segment address
Adv rtr   : 1.1.1.1     //ID of the ASBR that generates the Type
              5 LSA
Ls age    : 1340
Len       : 36
Options   : E
seq#      : 80000004
chksum    : 0xb5cc
Net mask  : 255.255.255.0 //Subnet mask
TOS 0  Metric: 1         //Cost value
E type    : 2
Forwarding Address : 0.0.0.0
Tag       : 1
Priority  : Low
```

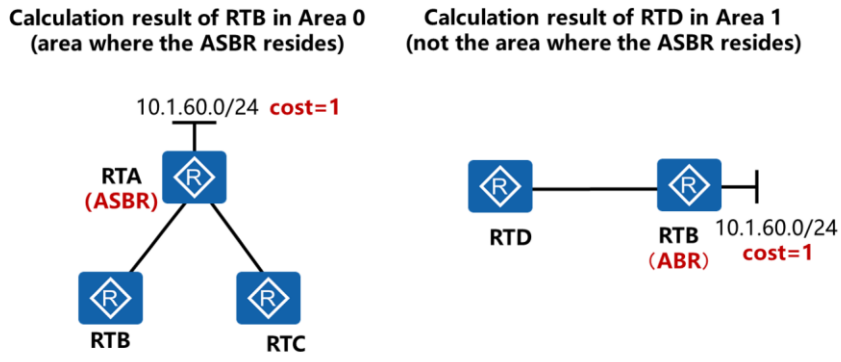

- The figure above illustrates a Type 5 LSA originated by RTA and is flooded to all OSPF areas except special OSPF areas.
- A Type 5 LSA contains the following fields:
  - Ls id: destination network address.
  - Adv rtr: ID of an ASBR.
  - Net mask: subnet mask of the destination network.
  - Metric: cost of the route from an ASBR to the destination network. The default value is 1.
  - Tag: attached to each external route to carry additional information about this route, usually for purpose of route policy. The default value is 1.

ASBR-Summary-LSA

```
<RTB>display ospf lsdb asbr self-originate

                Area: 0.0.0.1
          Link State Database


Type      : Sum-Asbr          //LSA type
Ls id     : 1.1.1.1           //Router ID of the ASBR
Adv rtr   : 2.2.2.2     //ID of the ABR that generates the Type 4 LSA
Ls age    : 15
Len       : 28
Options   : E
seq#      : 80000005
chksum    : 0xf456
Tos 0  metric: 1        //Cost of the route from RTB to the ASBR
```


- The figure above illustrates an ASBR-Summary-LSA (Type 4 LSA) originated by RTB in Area 1.
- When RTB floods a Type 5 LSA to Area 1, a Type 4 LSA is simultaneously originated and flooded to that area.
- A Type 4 LSA contains the following fields:
  - Ls id: ID of an ASBR.
  - Adv rtr: ID of the ABR that originates the Type 4 LSA.
  - Metric: cost of the route from the ABR to the ASBR.
- A Type 4 LSA can be flooded only in one area. Each time a Type 5 LSA is flooded to the next area via an ABR, a Type 4 LSA is then originated by the ABR to describe a route to reach the ASBR that originates the Type 5 LSA. In company with the Type 5 LSA, the Type 4 LSA is flooded no further than the border of the area where it is originated.
- Therefore, in an autonomous system, there can be multiple Type 4 LSAs with the same ASBR ID (indicated by Ls ID) but different ABR ID (indicated by Adv rtr) to describe the routes to reach the same ASBR.

External Route Calculation



HUAWEI

- Let's take RTB as an example to see how an ABR calculates an external route, which is illustrated by the figure on the left.
- Upon receiving a Type 5 LSA, RTB finds that the ASBR is in the same area (Area 0) as itself after examining the Adv rtr value (1.1.1.1). According to the fields Ls id, Net mask, and Metric in the Type 5 LSA, RTB originates a route for subnet 10.1.60.0/24 with the next-hop value of RTA's ID and the metric value of 1.
- The next example is given to explain how RTD in Area 1 calculates an external route. When receiving a Type 5 LSA, by examining the field Adv rtr, RTD finds that the ASBR's ID 1.1.1.1 is beyond its knowledge because the ASBR resides in a different area. RTD then examines the Type 4 LSA for routing information about ASBR 1.1.1.1 (represented by Ls ID), and finds that the gateway to 1.1.1.1 is 2.2.2.2 (represented by Adv rtr). According to the fields Ls id, Net mask, and Metric in the Type 5 LSA, RTD originates a route for subnet 10.1.60.0/24, with the next-hop value of RTB's ID and the metric value of 1.
- The costs of the routes calculated by RTB and RTD are both 1. However, as shown in the physical topology diagram, the cost that RTD calculates obviously should be greater than that which RTB does. It there anything goes wrong?

Contents

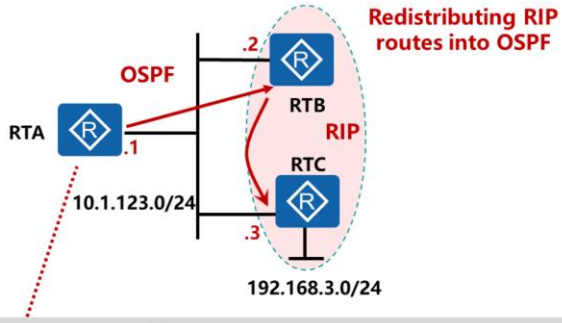
External Route Types

Type	Cost
Type 1 external route	Autonomous System (AS) internal cost + AS external cost
Type 2 external route	AS external cost

- OSPF can import two types of external routes:
  - Type 1 external route: The metric of the Type 1 external route is equal to the sum of the cost to the ASBR (a.k.a AS internal cost) and the cost from the ASBR to the external destination (a.k.a AS external cost). This occurs when the AS external cost is considered at the same level as the AS internal cost. There is equivalence between the AS external cost and the AS internal cost, and the Type 1 external metric is comparable directly to the link state metric of OSPF. By reason of above, the Type 1 external route is more reliable than the Type 2 external route.
  - Type 2 external route: When the AS external cost is considered far greater than the cost of any route internal in the AS, the Type 2 external route uses only the AS external cost as its cost, ignoring the AS internal cost. As mentioned, the Type 2 external route has lower reliability than the Type 2 external route.
- By default, OSPF uses Type 2 external routes.

Contents

Cause of Suboptimal External Routes

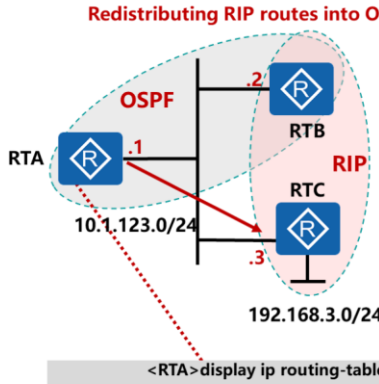


<RTA>display ip routing-table 192.168.3.0

Destination/Mask	Proto	Pre	Cost	Flags	NextHop	Interface
192.168.3.0/24	O_ASE	150	1	D	10.1.123.2	GigabitEthernet0/0/0

- As shown in the figure above, RTA, RTB, and RTC are on the same multiple access (MA) network. OSPF runs between RTA and RTB, and the Routing Information Protocol (RIP) runs between RTB and RTC.
- RTB redistributes routes learned through RIP into OSPF. Through OSPF, RTA learns the external route for destination 192.168.3.0/24 using RTB as the next hop. Therefore, traffic from RTA to subnet 192.168.3.0/24 is first forwarded to the next hop RTB and then to RTC. Obviously, this route is less optimal than that whose next hop is RTC.
- OSPF resolves this problem by specifying the Forwarding Address field.

Forwarding Address



Type 5 LSA generated by RTB:

```
Type      : External
Ls id     : 192.168.3.0
Adv rtr   : 2.2.2.2
Ls age    : 273
Len       : 36
Options   : E
seq#      : 80000001
chksum    : 0x470b
Net mask  : 255.255.255.0
TOS 0  Metric: 1
E type    : 2
Forwarding Address : 10.1.123.3
Tag       : 1
Priority  : Low
```

```
<RTA>display ip routing-table 192.168.3.0

Destination/Mask  Proto   Pre Cost   Flags NextHop   Interface

192.168.3.0/24  O_ASE   150 1    D  10.1.123.3  GigabitEthernet0/0/0
```


HUAWEI

- Generally, the Forwarding Address field in a Type 5 LSA describing an external route imported by an ASBR is set to 0.0.0.0.
- In the scenario shown in the figure above, the route destined for 192.168.3.0/24 is originated by RTB and the next hop is 10.1.123.3. OSPF runs on the subnet 10.1.123.0/24 to which the IP address 10.1.123.3 belongs. Therefore, the value of the Forwarding Address field is set to 10.1.123.3 in the Type 5 LSA generated by RTB.
- Upon receiving the Type 5 LSA, RTA finds that the LSA has Forwarding Address set to non-zero, then it looks up the forwarding address 10.1.123.3 in the routing table to determine the next hop.

Quiz

1. What type of routers is an AS-External-LSA originated by? What are the functions of an AS-External-LSA?
2. What type of routers is an ASBR-Summary-LSA originated by? What the functions of an ASBR-Summary-LSA?
3. What two types of external routes does OSPF provide? Which type of routes has a higher priority?

- Answer: An AS-External-LSA is originated by an ASBR to advertise an external route to an OSPF network. Note that one AS-External-LSA can advertise only one external route.
- Answer: An ASBR-Summary-LSA is originated by an ABR to instruct the rest of the OSPF domain how to get to the ASBR.
- Answer: There are two types of OSPF external routes: Type 1 and Type 2 external routes. Type 1 external routes have a higher priority than Type 2 external routes.


Thank You
www.huawei.com

Special OSPF Areas and Other Features

Foreword

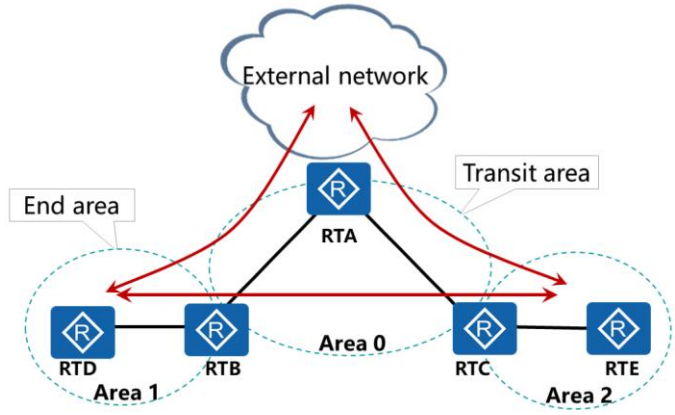
- OSPF routers need to maintain Link State Databases (LSDBs) containing intra-area, inter-area, and external routing information. As a network expands, the number or size of the LSDBs increases. If one OSPF area does not need to provide transit services for the traffic passing through, the routers in this area do not need to maintain the link state information from outside of the area.
- OSPF reduces the number of LSAs on a network by dividing its Autonomous System (AS) into a number of areas. However, there may be too many LSAs flooded in non-backbone areas, which is not bearable for the low-end routers situated in those areas. A couple of special OSPF areas are introduced to further reduce the number of LSAs and the size of a routing table.

Objectives

- Upon completion of this section, you will be able to:
  - Understand the feature of a special OSPF area
  - Understand application scenarios of OSPF virtual links
  - Be familiar with the principle of OSPF route summarization
  - Be familiar with the OSPF update mechanism
  - Understand the OSPF authentication mechanism

Contents

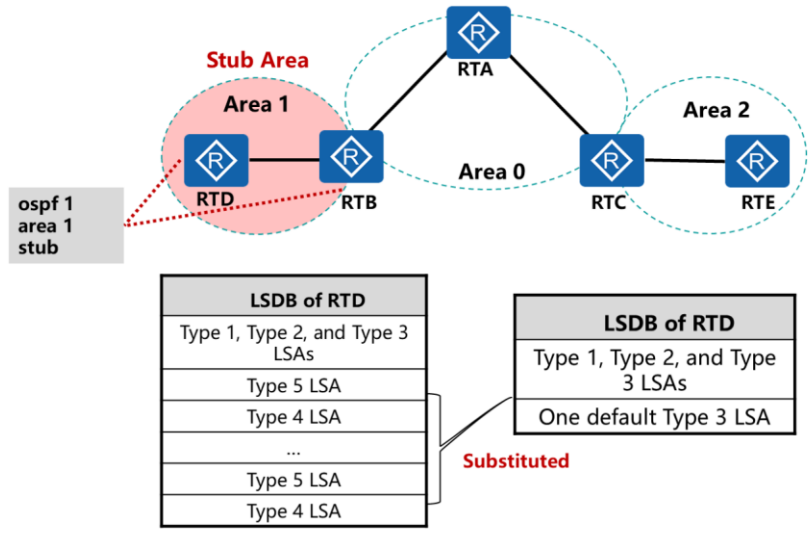
Transit Area and End Area



- As shown in the figure above, the entire network is divided into three areas: Area 0, Area 1, Area 2, with the backbone area (namely Area 0) interconnected with an external network.
- The two-way red arrow lines in the figure indicate all possible traffic flows among areas and the external network.
- OSPF areas are classified into the following two types:
  - Transit area: The transit area not only supports the traffic that is originated in it or destined for it, but also supports the traffic that is just passing through it (also known as traversing traffic). For example, Area 0 is a transit area.
  - End area: The end area only supports the traffic that is originated in it or destined for it. For example, Area 1 is an end area.
- For end areas, the following issues need to be considered:
  - First, consider whether there is necessity of having all specific routes to other areas. If there is only a single gateway to other areas or external autonomous systems, the answer is NO, because the summarized routes is more concise and effective than all specific routes for the outgoing traffic.
  - Second, consider the device capacity for processing and storage. The cost is a crucial factor in network construction and maintenance. In view of this, the routers with low capacity may have been widely deployed in the end areas.

- To calculate intra-area, inter-area, and external routes, OSPF routers need to collect a large number of LSAs on the network, which may excessively consume the storage space of LSDBs, and thereby put a calculation burden on the routers. On condition that routing reliability is guaranteed, the key to solving the problem is to reduce the number of LSAs as possible as it can be.

Stub Area



HUAWEI

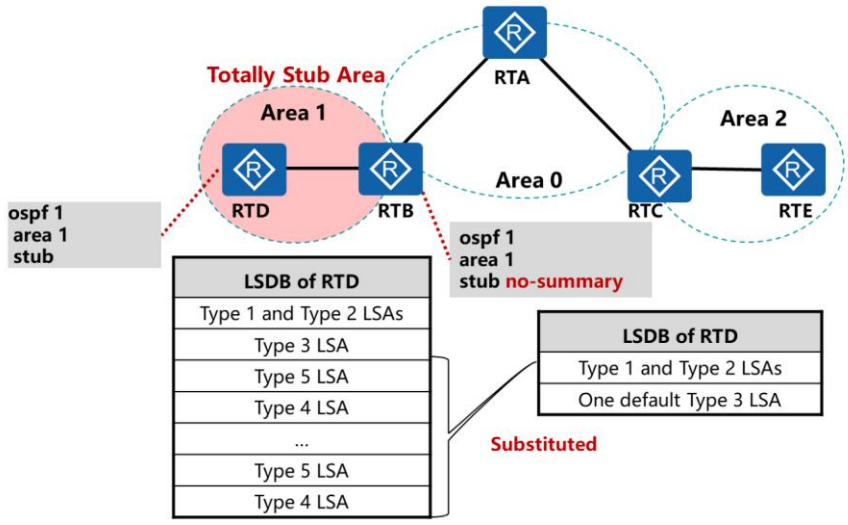
- An ABR does not advertise any AS external routes through Type 4 and Type 5 LSAs to a stub area. This results in a great reduction of the size of the LSDBs and routing tables.
- To ensure the traffic originated in a stub area can reach a destination outside the AS, an ABR of the stub area generates a default route and advertises it into the stub area through Type 3 LSAs.
- The stub area is an optional configuration attribute. It is suggested that you should not configure every non-backbone area to be a stub area. A stub area is a non-backbone area located at the far end of the Autonomous System, usually with only one ABR.
- When configuring a stub area, keep in mind the following points:
  - The backbone area cannot be configured as a stub area.
  - If an area is configured as a stub area, all the routers in this area must be configured as stub routers.
  - An ASBR cannot be part of a stub area. That is, AS external routes cannot be advertised in a stub area.
  - Virtual links cannot be configured through stub areas.

OSPF Routing Table in a Stub Area

```
<RTD>display ospf routing

        OSPF Process 1 with Router ID 4.4.4.4
                Routing Tables

 Routing for Network
 Destination        Cost   Type       NextHop         AdvRouter       Area
 10.1.24.0/24       1      Transit    10.1.24.4       4.4.4.4         0.0.0.1
 0.0.0.0/0          2      Inter-area 10.1.24.2       2.2.2.2         0.0.0.1
 10.1.12.0/24       2      Inter-area 10.1.24.2       2.2.2.2         0.0.0.1
 10.1.13.0/24       3      Inter-area 10.1.24.2       2.2.2.2         0.0.0.1
 10.1.35.0/24       4      Inter-area 10.1.24.2       2.2.2.2         0.0.0.1
 192.168.2.0/24     4      Inter-area 10.1.24.2       2.2.2.2         0.0.0.1
```


- When a stub area is configured, a default route is injected into the stub area through Type 3 LSAs to be substituted for all AS external routes.
- Even though the external network changes, routers in a stub area are not impacted. Rather, the reduction of routes improves the routers' performance.

Totally Stub Area



HUAWEI

- Neither AS external routing information in Type 4 or Type 5 LSAs nor inter-area routing information in Type 3 LSAs are allowed to be advertised in a totally stub area.
- An ABR generates a default route and injects it into the stub area where it attached through Type 3 LSAs to instruct other routers in the stub area how to get to other OSPF areas and even other autonomous systems.
- When configuring a totally stub area on the ABR, the no-summary parameter is additionally appended at the end of the command stub, which is different from that for a stub area.

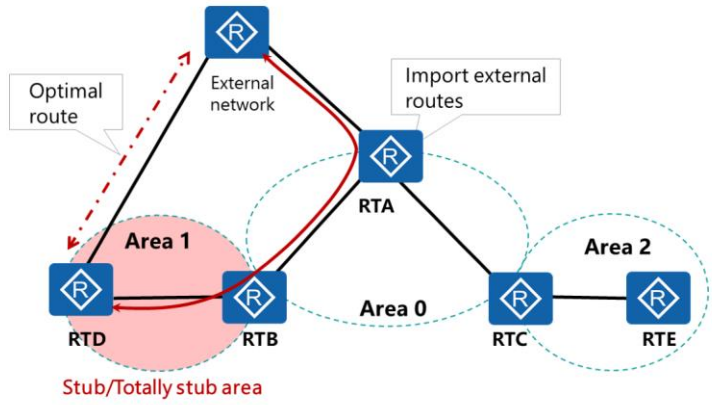
OSPF Routing Table in a Totally Stub Area

```
<RTD>display ospf routing

        OSPF Process 1 with Router ID 4.4.4.4
                Routing Tables

Routing for Network
Destination     Cost    Type        NextHop         AdvRouter       Area
10.1.24.0/24    1       Transit     10.1.24.4       4.4.4.4         0.0.0.1
0.0.0.0/0       2       Inter-area  10.1.24.2       2.2.2.2         0.0.0.1
```


- A default route is configured to guide traffic originated in a totally stub area to be destined for other areas and destinations outside the AS.
- Any changes of link states in other OSPF areas or outside the OSPF autonomous system do not affect the routers in totally stub area.
- OSPF solves the problems caused by oversized LSDBs in end areas, by defining stub and totally stub area. However, in some specific scenarios, configuring stub or totally stub areas is not the best solution.

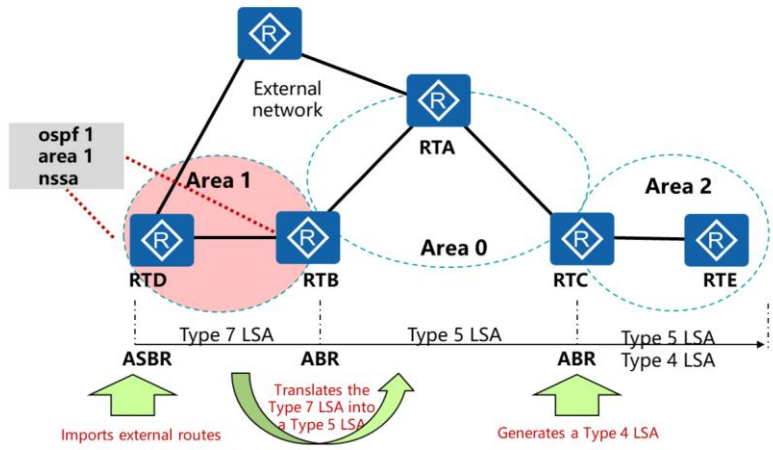
Contents

Challenges Confronted in Stub and Totally Stub Areas



- RTD and RTA connect to the same external network. RTA imports external routes to the OSPF domain. Area 1, where RTD resides, is configured as a stub or totally stub area to reduce the size of LSDB. The traffic from RTD to the external network is forwarded along the path RTD -> RTB -> RTA to reach the external network. Obviously, however, in comparison with the direct route to the external network, this route is suboptimal, since RTD is just on the boundary with the external network.
- As defined in OSPF, external routes cannot be imported to stub areas. The purpose of this mechanism is to prevent a large number of external routes from consuming device resources in stub areas and totally stub areas.
- There is no way for a stub area and a totally stub area to support the scenario where external routes need to be imported without excessive consumption of network resources.

NSSA and Totally NSSA



- The Not-So-Stubby Area (NSSA) was newly introduced to original OSPF RFC as a supplement.
- An NSSA is similar to a stub area. Different from stub areas and totally stub areas, NSSAs can import and advertise AS external routes to the entire OSPF AS without learning external routes from the other parts of the OSPF AS.
- NSSA LSA (Type 7 LSA):
  - The Type 7 LSA is introduced to support NSSAs by allowing imported external routes to be advertised in the OSPF autonomous system.
  - Type 7 LSAs are originated by the ASBR of an NSSA and are flooded only within the NSSA.
  - A default route can be advertised through a Type 7 LSA as well to guide traffic to other ASs.
- Type 7 LSAs can be converted into Type 5 LSAs.
    - When an ABR of an NSSA receives Type 7 LSAs, the ABR selectively translates the Type 7 LSAs into Type 5 LSAs to advertise imported external routes to other areas of the OSPF domain.
    - When multiple ABRs exist in an NSSA, only the one with the greatest Router ID translates the Type 7 LSAs into Type 5 LSAs..

- Differences between a totally NSSA and an NSSA:
- Type 3 LSAs are not allowed to be flooded in a totally NSSA, whereas Type 3 LSAs will pass into and out of an NSSA.
- The only difference in configuration between NSSA and totally NSSA is the additional appending parameter no-summary for a totally NSSA

Examples of LSDBs in an NSSA and a Totally NSSA

NSSA

<RTB>display ospf lsdb

OSPF Process 1 with Router ID 2.2.2.2
Link State Database

Area: 0.0.0.1

Type	LinkState ID	AdvRouter
Router	4.4.4.4	4.4.4.4
Router	2.2.2.2	2.2.2.2
Network	10.1.24.4	4.4.4.4
Sum-Net	10.1.35.0	2.2.2.2
Sum-Net	10.1.13.0	2.2.2.2
Sum-Net	10.1.12.0	2.2.2.2
Sum-Net	192.168.2.0	2.2.2.2
NSSA	**0.0.0.0**	**2.2.2.2**
NSSA	**10.1.47.0**	**4.4.4.4**
NSSA	**192.168.7.0**	**4.4.4.4**
NSSA	**10.1.24.0**	**4.4.4.4**

Totally NSSA

<RTB>display ospf lsdb

OSPF Process 1 with Router ID 2.2.2.2
Link State Database

Area: 0.0.0.1

Type	LinkState ID	AdvRouter
Router	4.4.4.4	4.4.4.4
Router	2.2.2.2	2.2.2.2
Network	10.1.24.4	4.4.4.4
Sum-Net	**0.0.0.0**	**2.2.2.2**
NSSA	**0.0.0.0**	**2.2.2.2**
NSSA	**10.1.47.0**	**4.4.4.4**
NSSA	**192.168.7.0**	**4.4.4.4**
NSSA	**10.1.24.0**	**4.4.4.4**

- An ABR of an NSSA generates a default route (Type 7 LSA).
- An ABR of a totally NSSA automatically generates a default route (Type 3 LSA).

Summary of LSAs

LSA Type	Advertising Router	Contents of LSA	Advertisement Scope
Router-LSA (Type 1)	OSPF router	Topology and routing information	Local area
Network-LSA (Type 2)	Designated router (DR)	Topology and routing information	Local area
Network-Summary-LSA (Type-3)	ABR	Inter-area routing information	Areas except totally stub areas
ASBR-Summary-LSA (Type 4)	ABR	ASBR's router ID	Areas except totally stub areas
AS-External-LSA (Type 5)	ASBR	AS external routing information	OSPF areas except stub areas
NSSA-LSA (Type 7)	ASBR	External routing information in NSSAs	Totally NSSAs

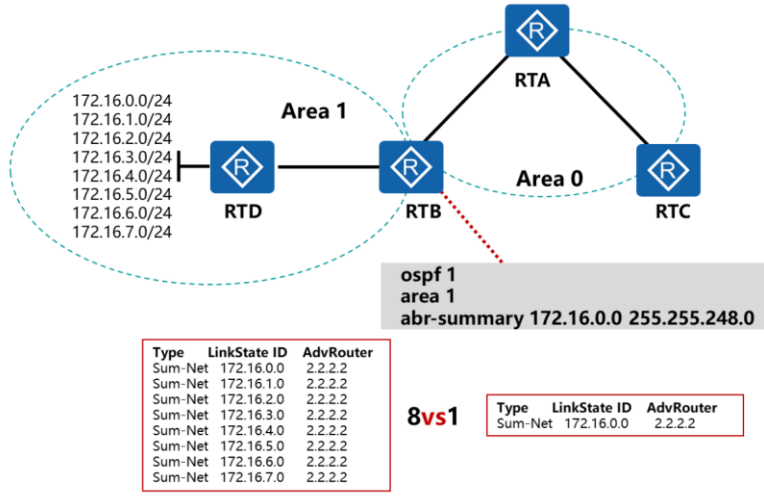
- Question: What are the limitations of special OSPF areas? Are there any other methods of reducing the number of LSAs?

- LSAs have the following functions:
  - Router-LSA (Type 1): describes collected states of the router's interfaces to an area. It is originated by every router and flooded only throughout a single area it belongs to.
  - Network-LSA (Type 2): describes link states of a broadcast and NBMA interface. It is originated by the DR and flooded only throughout a single area it belongs to.
  - Network-Summary-LSA (Type 3): describes a route to a destination network outside the area, yet still inside the AS. It is originated by an ABR and flooded throughout the LSA's associated area.
  - ASBR-Summary-LSA (Type 4): describes routes to an ASBR. It is originated by an ABR and flooded throughout the areas except the area where the ASBR resides.
  - AS-External-LSA (Type 5): describes routes to destinations external to the Autonomous System. It is originated by an ASBR and flooded to all areas except stub areas and NSSAs.
  - NSSA-LSA (Type 7): describes AS external routes. It is originated by an ASBR and only flooded in NSSAs.

- A special OSPF area is configured not only to reduce the number of LSAs in the area and route calculation load on the routers, but also to reduce the impact of the network faults outside the area. However, the benefits of a special OSPF area are only available within the area. So, what methods can be applied in other areas?

Contents

Inter-Area Route Summarization



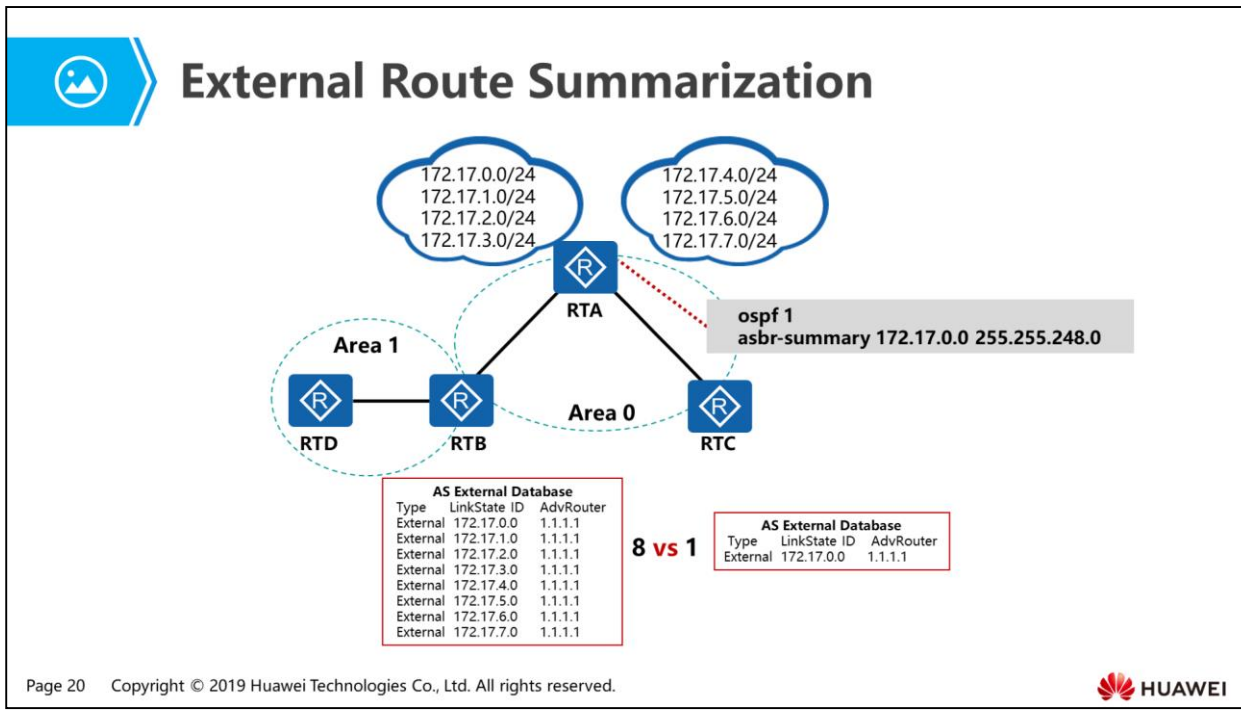
Type	LinkState ID	AdvRouter
Sum-Net	172.16.0.0	2.2.2.2
Sum-Net	172.16.1.0	2.2.2.2
Sum-Net	172.16.2.0	2.2.2.2
Sum-Net	172.16.3.0	2.2.2.2
Sum-Net	172.16.4.0	2.2.2.2
Sum-Net	172.16.5.0	2.2.2.2
Sum-Net	172.16.6.0	2.2.2.2
Sum-Net	172.16.7.0	2.2.2.2

8vs1

Type	LinkState ID	AdvRouter
Sum-Net	172.16.0.0	2.2.2.2

- On a large-scale OSPF network, the size of the routing tables can be overlarge. This may result in a slow lookup of a route. In order to solve the problem, route summarization can be performed on the boundaries of the areas.
- Route summarization is the process of grouping a consecutive range of network address prefixes into a single prefix and advertising it to the other parts of the network, without losing any routing information of all known destinations. Even if an interface with an IP address in the range of summarized prefixes is brought up and down frequently behind the summarizing router, the changes will not be advertised to the rest of the network and the destination still seems to remain valid. This process prevents route flapping by hiding instability in the system behind the summary.
- Route summarization can only summarize routing information and an ABR is one of the positions that can perform route summarization.
    - When an ABR advertises routing information into an area, it originates Type 3 LSAs, each of which is for one individual subnet. If a range of consecutive subnet addresses exists in this area, a command can be run to manually summarize these subnets into one single network address prefix. In this way, the ABR transmits only one Type 3 LSA describing a summarized route. All specific routes belonging to the range of the summarized routes will not be advertised separately.

- All the other LSAs that belong to the summarized network segment range specified by commands are not transmitted separately.
- As shown in the figure above, eight consecutive subnets exist in Area 1. Accordingly, eight Type 3 LSAs are originated by RTB if no summarization is performed. When route summarization is configured on RTB, there will be only one Type 3 LSA to be generated and flooded into Area 0.
- Another possible summarization point is ASBR, which the external routes are imported through.



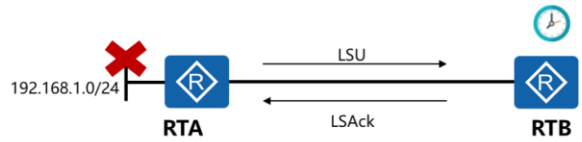
- Route summarization on an ASBR:
  - If route summarization is configured on an ASBR, the ASBR summarizes imported external routes. An ASBR of NSSA is no exception.
  - If a router is an ASBR as well as an ABR of the NSSA, the router can summarize external routes into corresponding address prefixes while Type 7 LSAs are being translated into Type 5 LSAs.
- As shown in the figure above, RTA in Area 0 imports eight consecutive external routes to the area. Eight Type 5 LSAs are originated and flooded to the area.
- When external route summarization is configured on the RTA, which functions as an ASBR, RTA originates only one Type 5 LSA by grouping the eight consecutive external routes into one, and floods it to AS.
- Route summarization reduces impact of network faults.
- The network convergence speed after a network fault occurs is also an important criterion for judging the performance of a routing protocol.

Contents

HUAWEI

Periodical and Triggered LSA Update

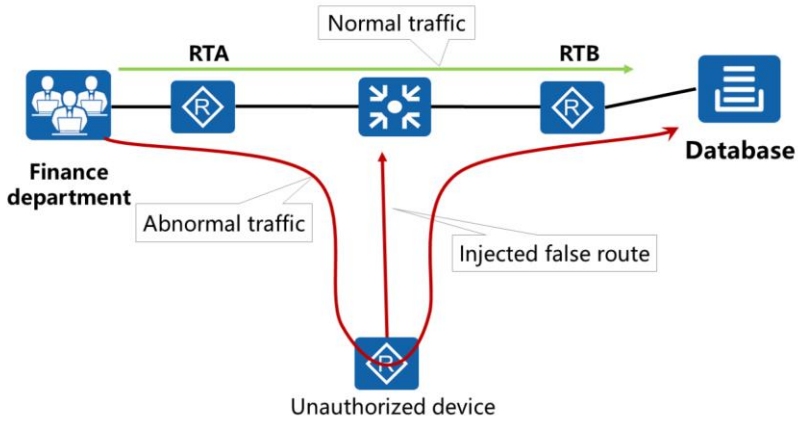
- Periodical update:
  - Each LSA is refreshed every 1800 seconds or else it will expire after 3600 seconds.
- Triggered update:
  - When the link state changes, the router sends a Link State Update (LSU) packet immediately.



- The LSA reliability must be guaranteed to ensure route calculation accuracy.
- OSPF maintains an aging timer of 3600 seconds (Maxage) for each LSA. When the incremental LS age in each LSA header reaches the Maxage, the LSA is deleted from the LSDB.
- To prevent an LSA from being discarded due to a timeout, OSPF resends the LSA every 1800 seconds with a higher sequence number.
- An OSPF router re-originates an LSA every 1800s and advertises it to other routers.
- To speed up network convergence, OSPF provides the triggered update mechanism.
- Once link states change, a router floods update messages in the area to force other routers to recalculate routes immediately. Hence, the fast network convergence is accomplished.

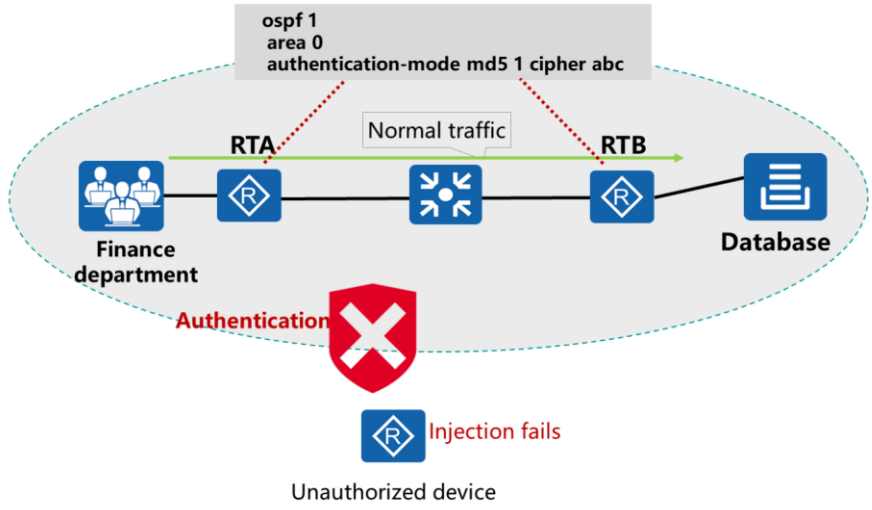
Contents

Security Risks



- On the corporate network shown in the figure, OSPF is run on the corporate network. Normally, the traffic from the Finance Department passes through RTA and RTB before reaching the database.
- An unauthorized device accesses the corporate network and injects a false route, causing the traffic to be abnormally forwarded along the path "Finance Department -> RTA -> unauthorized device -> RTB -> Database." The unauthorized device sniffs and analyzes the traffic passing through it, and obtains private financial information. This leads to a leak of company's confidential information. This causes disclosure of the enterprise's confidential information.
- What can be done to secure OSPF routing information?

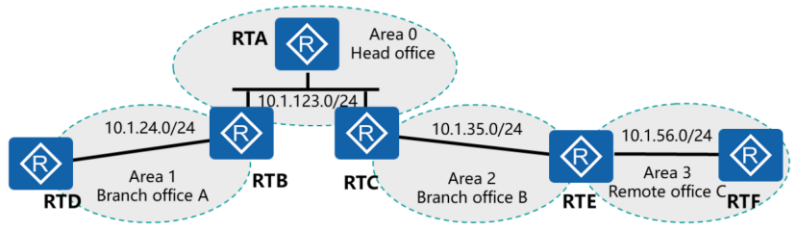
Authentication Ensures Security



- OSPF supports authentication. Only authenticated OSPF routers can establish neighboring relationships and exchange routing information.
- The OSPF authentication works on one of the following bases:
  - Area basis
  - Per-interface basis
- Supported authentication modes: null, simple, MD5, and HMAC-MD5.
- When both area-based and per-interface-based authentication are configured, per-interface-based authentication is preferred.

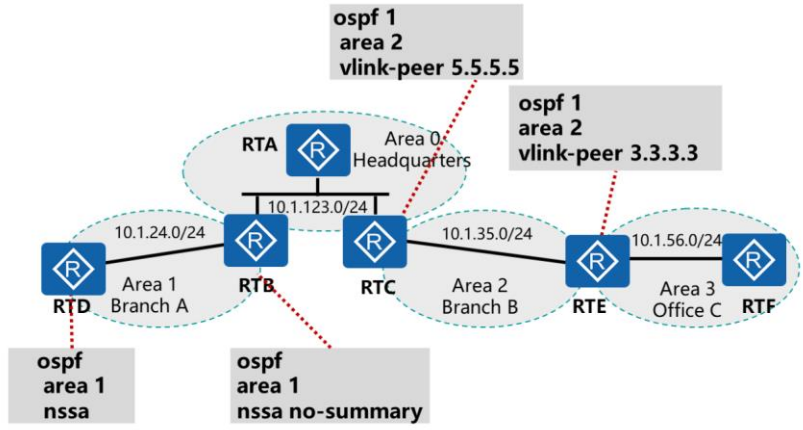
Comprehensive Scenario of OSPF

- The following figure shows the network topology of an enterprise. The basic OSPF configuration has been done, but the network administrator still has the following problems to deal with:
  - The communications between the head office and both branch offices A and B are normal, but they are all unable to communicate with remote office C.
  - On account of the low capacity of devices in branch office A, the load of route calculation and link-state storage should be reduced. Furthermore, considering the future expansion of the network, the capability of importing external routes should be reserved for this area.
  - Routing information security must be ensured in remote office C because of frequent visits of guests.
  - In addition to the external cost, the OSPF internal cost also needs to be considered when RTA imports external routes.

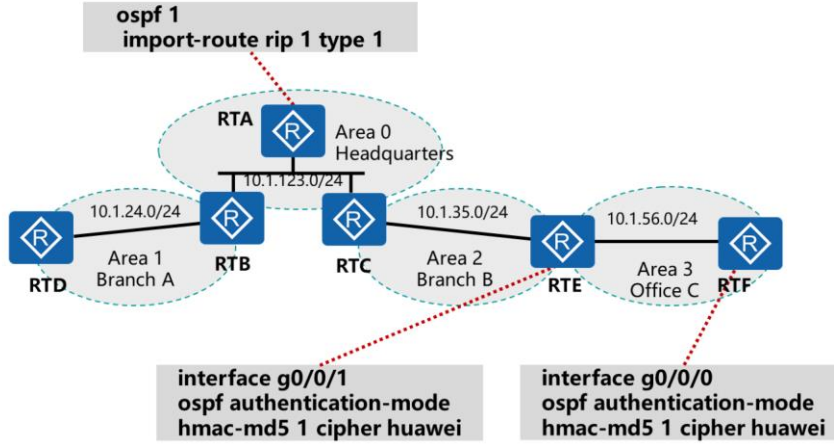


- Requirement analysis 1: The remote office C is in Area 3 and RTE connects Area 2 and 3. The reason why the remote office C cannot communicate with other areas is that the Area 3 is not physically connected to the backbone area, which has broken the rules of inter-area routing loop prevention. The solution is configuring a virtual link between RTE and RTC to connect logically Area 3 and the backbone area.
- Requirement analysis 2: To reduce the load of route calculation on low-capacity devices, the area can be configured as a stub area, a totally stub area, an NSSA, or a totally NSSA. Especially, to minimize the load of route calculation, the choice should be the totally stub or totally NSSA. Further, if the capability of importing external routes must be taken into consideration, the totally NSSA would be the only choice.
- Requirement analysis 3: Authentication needs to be implemented to ensure routing information security. In this case, the authentication mode selected is HMAC-MD5, which is considered as the most secure OSPF authentication mode, on a per-interface basis.
- Requirement analysis 4: If the OSPF AS internal costs need to be considered when calculating AS external routes, routes should be imported into OSPF as Type 1 external routes.

OSPF Configuration (1)



OSPF Configuration (2)



Quiz

1. What special areas does OSPF define?
2. What are the differences between stub areas and totally stub areas?
3. What routers can have inter-area route summarization configured?

- Answer: OSPF defines four special areas: stub area, totally stub area, not-so-stubby area (NSSA) and totally NSSA.
- Answer: A stub area does not allow transmission of Type 4 or Type 5 LSAs but allows transmission of Type 3 LSAs. A totally stub area does not allow transmission of Type 4, Type 5, or Type 3 LSAs. Only Type 3 LSAs describing default routes can be transmitted in a totally stub area.
- Answer: Inter-area route summarization can be configured on an ABR.


Thank You
www.huawei.com

IS-IS Principles and Configurations

Foreword

- Intermediate System-to-Intermediate System (IS-IS) is a link-state based IGP. It uses the Shortest Path First (SPF) algorithm to calculate routes. IS-IS is a dynamic routing protocol initially designed by the International Organization for Standardization (ISO) for its Connectionless Network Protocol (CLNP).
- To support IP routing, the Internet Engineering Task Force (IETF) extends and modifies IS-IS in RFC 1195. This enables IS-IS to be applied to TCP/IP and OSI environments. This type of IS-IS is called Integrated IS-IS. IS-IS has been widely used in large ISP networks because of its simplicity and high scalability.

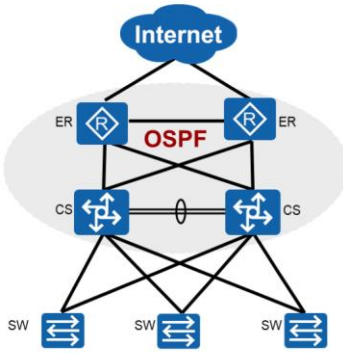
Objectives

- Upon completion of this section, you will be able to:
  - Master IS-IS principles and configurations
  - Be familiar with differences between IS-IS and OSPF

Contents

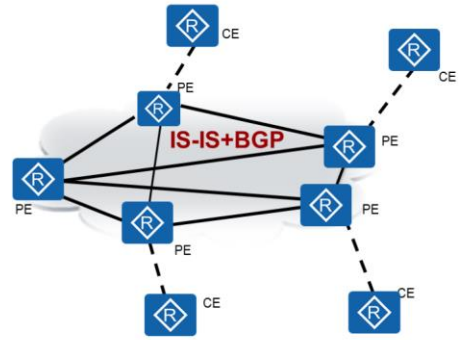
HUAWEI

Application



- Campus network:

  A variety of areas, changeable policies, and fine-grained scheduling



- Backbone network:

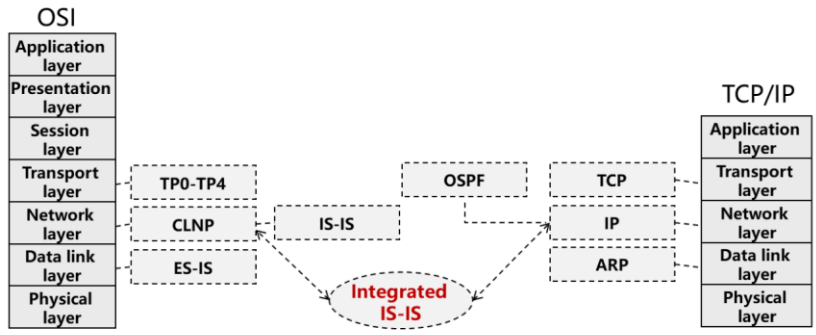
  Flat area, fast convergence, and high transmission volume

HUAWEI

- Campus network characteristics:
  - It is an application-oriented network, which is intended for enterprise network users.
  - The number of routers is relatively small, the dynamic routing LSDB capacity is relatively low, and the number of Layer 3 routing domains is relatively small.
  - Compared with a backbone network, a campus network has a smaller coverage area and sufficient bandwidth. The link state protocol occupies less bandwidth.
  - Routing policies and policy-based routing (PBR) are changed frequently, so fine-grained routing is required.
  - There are a variety of OSPF route types (internal and external routes), area types (backbone, common, and special areas), and network types (a maximum of five network types), as well as well-designed cost calculation (configured based on bandwidth).
- Backbone network characteristics:
  - It is a service-oriented network and built by Internet Service Providers (ISPs) to provide Internet services to terminal users.
  - There are a large number of routers. Therefore, route scheduling is critical.

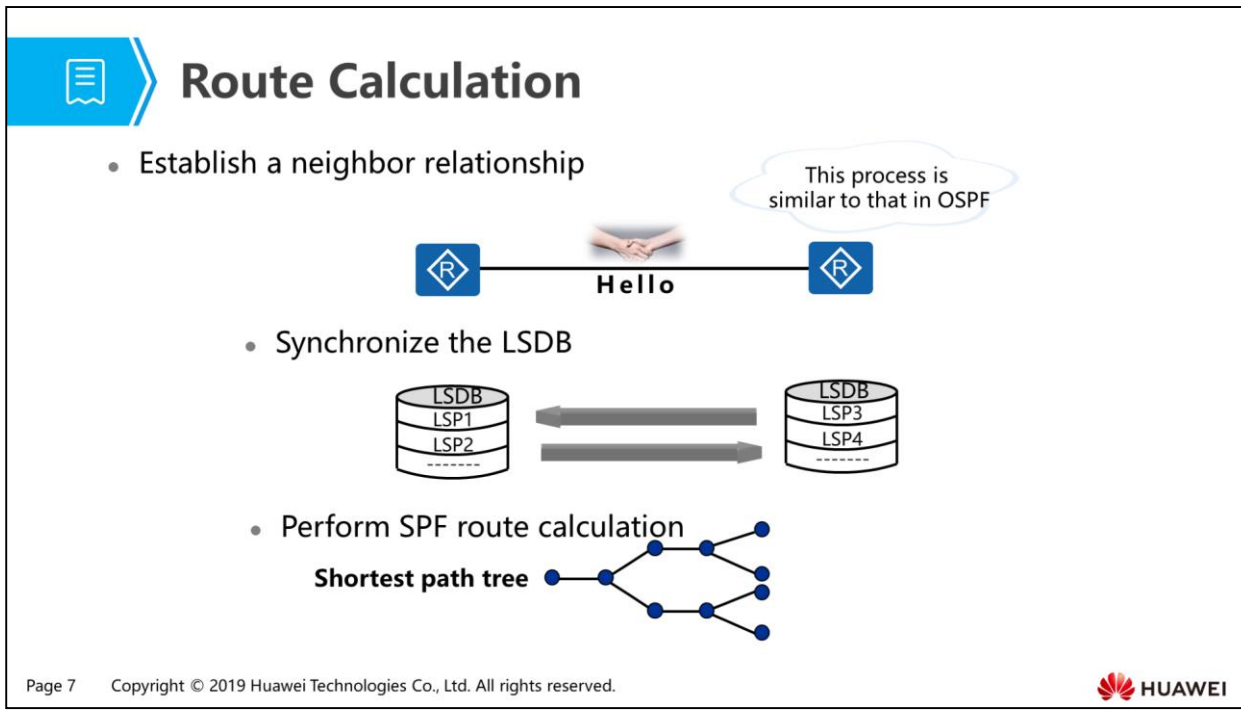
- The architecture is flat, requiring IGP as basic routing protocols to serve BGP.
- The LSDB is large and sensitive to link convergence and the link cost is high.
- Simplicity, high efficiency, and high scalability are required to meet various customer service requirements (IPv6/IPX).
- IS-IS has many advantages in backbone networks, including IS-IS fast algorithm (enhanced PRC), simple packet structure (TLV), fast neighbor relationship establishment, and large-capacity route transmission.

Origin



- Integrated IS-IS characteristics:
  - Supports CLNP and IP networks.
  - Works at the data link layer.
- OSPF characteristics:
  - Supports only IP networks.
  - Works at the IP layer.

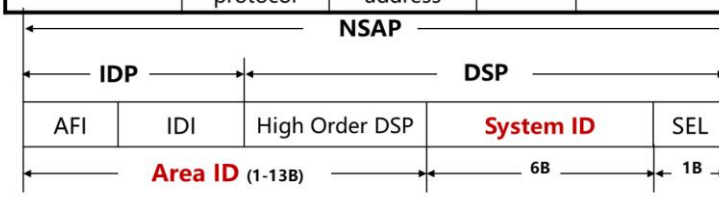
- IS-IS is a dynamic routing protocol initially designed by the International Organization for Standardization (ISO) for its Connectionless Network Protocol (CLNP).
- To support IP routing, the Internet Engineering Task Force (IETF) extends and modifies IS-IS in RFC 1195. This enables IS-IS to be applied to TCP/IP and OSI environments. This type of IS-IS is called Integrated IS-IS. Unless otherwise stated, IS-IS refers to Integrated IS-IS.
- IS-IS is an Interior Gateway Protocol (IGP) and used within an Autonomous System (AS). IS-IS is a link state protocol and uses Shortest Path First (SPF) algorithm to calculate routes.



- Neighbor relationship establishment:
  - Before establishing a neighbor relationship, devices on two ends exchange Hello packets and negotiate parameters, including the circuit type (level-1/level-2), hold time, network type, supported protocol, area ID, system ID, PDU length, and interface IP address.
- Link information exchange:
  - Different from OSPF, IS-IS exchanges link state information using link state PDUs (LSPs) instead of link state advertisements (LSAs). In IS-IS, CSNPs and PSNPs are used to synchronize the LSDB and request as well as acknowledge link state information (link state information summary). Detailed topology of link state information and routing information are transmitted using LSPs.
- Route calculation:
  - IS-IS SPF calculation is similar to OSPF SPF calculation except that the IS-IS algorithm separates the topology and IP network segment and speeds up network convergence.

Address Structure

TCP/IP protocol stack	IP protocol	IP address	OSPF	Area ID+Router ID
OSI	CLNP protocol	NSAP address	IS-IS	NET **identifier**



NET is a special type of NSAP address (SEL=00). When configuring IS-IS on a router, you only need to consider the NET. For example:

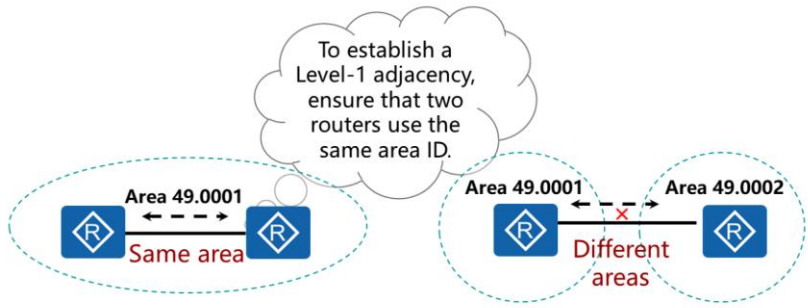
49.0001. **0000.0000.0001**.**00**
Area ID **System ID** **N-SEL**

- NSAP address:
  - The IDP is equal to the network number in an IP address. As defined by the ISO, the IDP consists of the Authority and Format Identifier (AFI) and Initial Domain Identifier (IDI). The AFI specifies the address allocation authority and address format; the IDI identifies a domain.
  - The DSP is equal to the subnet number and host number in an IP address. The DSP consists of the High Order DSP (HODSP), system ID, and NSAP Selector (SEL). The HODSP is used to divide areas; the system ID identifies a host; the SEL indicates the service type.
  - The IDP together with the HODSP of the DSP can identify a routing domain and the areas in a routing domain; therefore, the combination of the IDP and HODSP is referred to as an area address, which is equal to an area ID in OSPF.
  - System ID uniquely identifies a host or router in an area and has a fixed length of 48 bits (6 bytes).
  - The role of an SEL is similar to the Protocol field of an IP header. Different network layer services match different SELs. The SEL is always 00 in IP.

- NET:
- An NET indicates the network layer information of an IS itself and can be regarded as a special type of NSAP (SEL=0). The NET length is the same as the NSAP length. The maximum NSAP length is 20 bytes and its minimum length is 8 bytes. When configuring IS-IS on a router, you can configure only a NET instead of an NSAP.
- An IS-IS process can be configured with a maximum of three NETs. When configuring multiple NETs, ensure that their system IDs are the same.

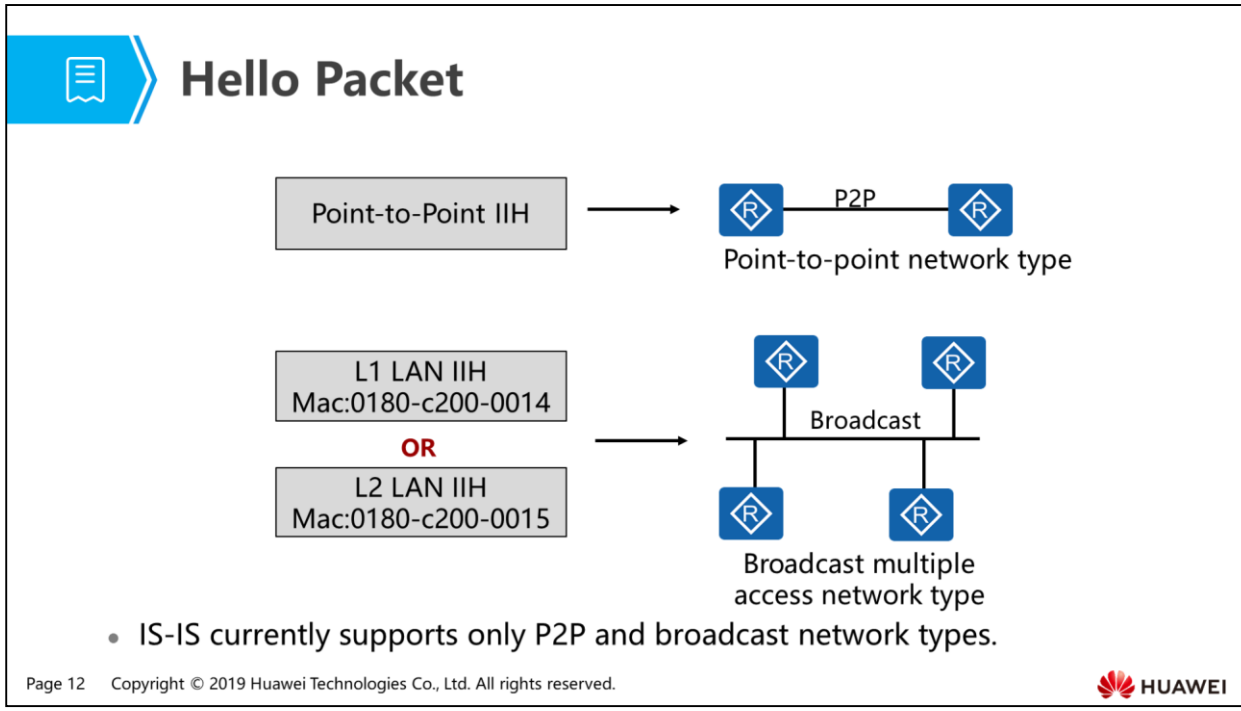
Router Types

- IS-IS routers are classified into three types:
  - Level-1 router: supports only Level-1 LSDB.
  - Level-2 router: supports only Level-2 LSDB.
  - Level-1-2 router: is the default router type and supports both Level-1 and Level-2 LSDBs.



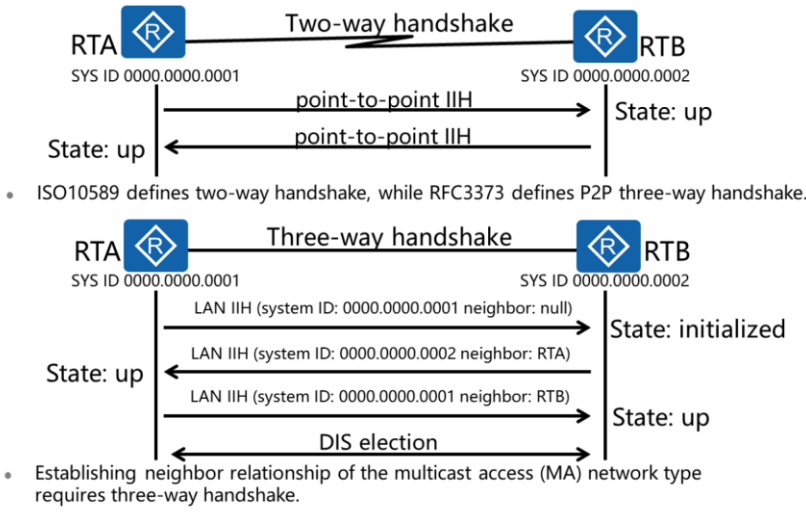
- Level-1 router:
  - A Level-1 router establishes neighbor relationships with only Level-1 and Level-1-2 routers in the same area. It maintains a Level-1 LSDB, which contains intra-area routing information and forwards packets destined for other areas to the nearest Level-1-2 router. A Level-1 router can establish only Level-1 adjacencies.
- Level-2 router:
  - A Level-2 router manages inter-area routing. It can establish neighbor relationships with Level-2 routers in the same or different areas and with Level-1-2 routers in different areas. It maintains a Level-2 LSDB, which contains inter-area routing information. A Level-2 router can establish only Level-2 adjacencies.
- Level-1-2 router:
  - A router, which belongs to both Level-1 area and Level-2 area, is called a Level-1-2 router. A Level-1-2 router maintains two LSDBs, that is, a Level-1 LSDB and a Level-2 LSDB. The Level-1 LSDB is used for intra-area routing and the Level-2 LSDB is used for inter-area routing.

- A Level-1-2 router can establish Level-1 neighbor relationships with Level-1 routers in the same area. It can also establish Level-2 neighbor relationships with Level-2 routers and Level-1-2 routers in other areas.
- IS-IS routers in different areas can only establish Level-2 adjacencies:
- A Level-2 router can establish an adjacency with other Level-2 routers.
- A Level-1-2 router can establish an adjacency with other Level-2 routers.
- A Level-1-2 router can establish an adjacency with other Level-1-2 routers.



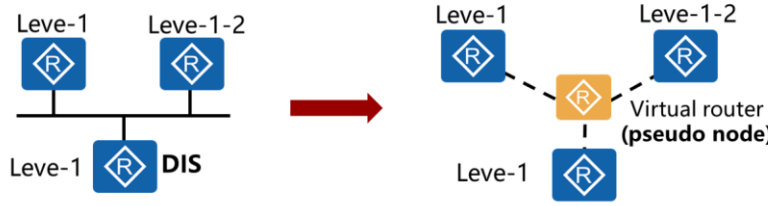
- Hello protocol data unit (Hello PDU):
  - Hello packets are used to discover neighbors, negotiate parameters, maintain neighbor relationships, and function as Keepalive packets.
  - Like OSPF, IS-IS establishes neighbor relationships by exchanging Hello packets. Hello packets are classified into three types based on scenarios.
  - In a broadcast network, Level-1 IS-IS routers use Level-1 LAN IS-IS Hello (IIH) packets with the destination multicast MAC address 0180-c200-0014.
  - In a broadcast network, Level-2 IS-IS routers use Level-2 LAN IIH packets with the destination multicast MAC address 0180-c200-0015.
  - P2P IIH packets are used in a non-broadcast network and do not carry any field indicating the DIS (also called pseudo node). IIH packets need to use the padding field for negotiating the size of packets sent between devices on the two ends.
- IS-IS supports the following network types:
  - P2P network type.
  - Broadcast multiple access network type.
  - In special environments such as frame relay, sub-interfaces can be created to support the P2P network type.

Neighbor Relationship Establishment



- On a P2P link, a neighbor relationship can be established through two-way handshake and three-way handshake.
  - In two-way handshake, upon receiving a Hello packet from its neighbor, a router considers the neighbor Up and establishes a neighbor relationship with the neighbor. However, there is a risk of unidirectional communication.
  - In three-way handshake, a neighbor relationship is established after P2P IIHs are sent three times, similar to neighbor relationship establishment on a broadcast link.
- On a broadcast link, a neighbor relationship is established using LAN IIH packets through three-way handshake.
  - When a router receives from its neighbor a Hello PDU that does not contain its system ID, the state machine enters the initialized state.
  - The state machine enters the Up state only when the router receives from its neighbor a Hello PDU that contains its system ID, eliminating the risk of unidirectional communication.
  - On a broadcast network, the DIS (also called pseudo node) will be elected after the neighbor state becomes Up. The DIS functions similarly to the Designated Router (DR) in OSPF.

Comparisons Between DIS and DR



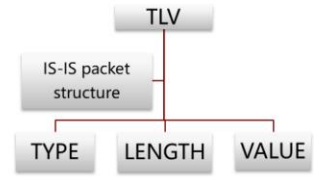
Item	IS-IS DIS	OSPF DR
Election priority	Devices with all priorities participate in DIS election	The device with priority 0 does not participate in DR election
Election waiting time	Two Hello intervals	40s
Backup	No backup	BDR for backup
Adjacency	All routers establish adjacencies	DR others establish 2-way neighbor relationships
Preemption	Yes	No
Function	Send CSNPs periodically to guarantee LSDB synchronization on MA networks.	Reduce LSA flooding

- DIS and pseudo node:
  - DIS is short for the Designated IS.
  - Pseudo node is a virtual router created by the DIS in a broadcast network.
- DIS characteristics:
  - The DIS needs to be elected on a broadcast network. After a neighbor relationship is established, a router participates in DIS election after waiting for two Hello intervals. A Hello packet contains the Priority field. The device with the largest priority value is elected as the DIS of the broadcast network. If the devices have the same priority, the device with the largest interface MAC address is elected as the DIS. In IS-IS, the interval at which the DIS sends Hello packets defaults to 10/3 seconds, while the interval at which non-DIS routers send Hello packets is 10s.
- Comparisons between DIS and DR:
  - In IS-IS, the router with priority 0 also takes part in DIS election. In OSPF, the router with priority 0 does not take part in DR election.
  - In OSPF, DR/BDR election requires 40s waiting time and is complicated. In IS-IS, DIS election requires two Hello intervals and is simple and fast.

- In IS-IS, only the DIS is elected. In OSPF, both the DR and BDR are elected, and BDR functions as the backup of the DR.
- After election is complete, if a new router with a higher priority is added, it can become the new DIS in IS-IS but cannot become the new DR in OSPF.
- After election is complete, all routers on an IS-IS network establish adjacencies. In OSPF, DR others establish only full adjacency relationships with the DR/BDR. DR others establish only 2-way neighbor relationships with each other.
- DIS and DR functions:
- Both the DIS and DR function as a virtual node during SPF calculation, simplifying MA network topology.
- Both the DIS and DR are designed to reduce flooding of LSPs/LSAs.
- In IS-IS, the DIS sends CSNPs to synchronize the LSDB (IS-IS extension).

Link State Information Transmission

- LSP PDU: is used to exchange link state information.
- Real node LSP
- Pseudo node LSP (exists only in broadcast links)
- SNP PDU: is used to maintain and synchronize the LSDB and carries summary information.
- CSNP (used to synchronize LSPs)
- PSNP (used to request and acknowledge LSPs)



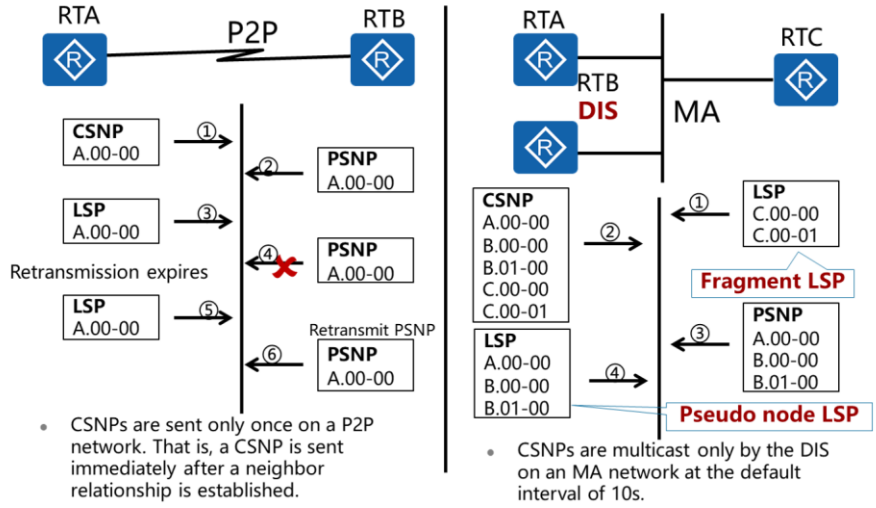
IS-IS packets fall into two types: Level-1 and Level-2 packets. The destination MAC addresses of all IS-IS packets in MA networks are multicast MAC addresses:
Level-1 address: **0180-C200-0014**
Level-2 address: **0180-C200-0015**

  HUAWEI

- IS-IS TLV:
  - TLV indicates the type, length, and value. It is a data structure and contains three fields.
  - Using TLV to construct packets can ensure flexibility, scalability, and stable packet structure. If new characteristics are added, only new TLVs need to be added, removing the need to change the packet structure.
  - Using TLV to indicate network topology and routing information can improve packet flexibility and scalability.
- Link State Protocol PDU (LSP PDU):
  - LSPs are similar to LSAs in OSPF and used to transmit link state information, including the topology and network ID.
  - Level-1 LSPs are transmitted by Level-1 routers.
  - Level-2 LSPs are transmitted by Level-2 routers.
  - Level-1-2 routers can transmit both Level-1 and Level-2 LSPs.
  - LSPs contain two important fields: ATT field and IS-Type field. The ATT field indicates that an LSP is transmitted by a Level-1-2 router. The IS-Type field indicates whether an LSP is generated by a Level-1 or Level-2 router.

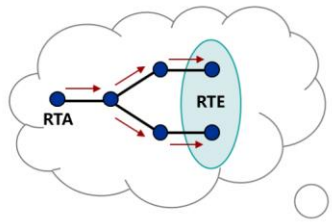
- The LSP update interval is 15 minutes and the LSP aging time is 20 minutes. The aging of an LSP requires the waiting time of 20 minutes and the delay of 60 seconds. The LSP retransmission time is 5 seconds.
- Sequence Number PDU (SNP PDU):
- Complete Sequence Number PDU (CSNP) contains the summary of all LSPs in an LSDB and can be used to ensure LSDB synchronization between neighboring routers.
- Partial Sequence Number PDU (PSNP) contains the summary of some LSPs in an LSDB and can be used to request and acknowledge LSPs.
- CSNPs are similar to DD packets in OSPF and used to transmit the summary of all link information in an LSDB. PSNPs are similar to LSR or LSAck packets in OSPF and used to request and acknowledge some link information.

Link State Information Exchange

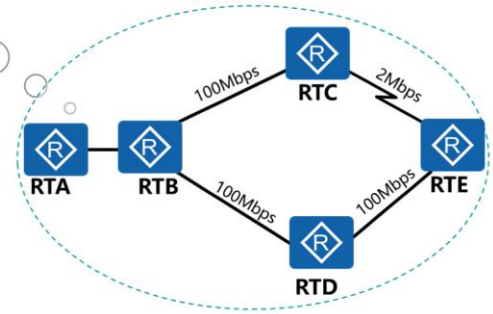


- LSDB synchronization on a P2P network:
  - After establishing a neighbor relationship, RTA and RTB send a CSNP to each other. If the LSDB of the neighbor and the received CSNP are not synchronized, the neighbor sends a PSNP to request the required LSP.
  - If RTA does not receive any PSNP from RTB after the LSP retransmission timer expires, RTA retransmits the required LSP until it receives a PSNP from RTB.
- LSDB synchronization between the newly added router and DIS on an MA network:
  - Assume that the newly added router RTC has established neighbor relationships with RTB (DIS) and RTA.
  - After establishing neighbor relationships, RTC sends its LSP to the multicast address (01-80-C2-00-00-14 in a Level-1 area and 01-80-C2-00-00-15 in a Level-2 area). Then all neighbors on the network receive this LSP.
  - The DIS on the network adds the LSP received from RTC to its LSDB, waits for the expiry of the CSNP timer (the DIS sends CSNPs at an interval of 10s). After the CSNP timer expires, the DIS sends a CSNP to synchronize the LSDBs on the network.
  - RTC receives the CSNP from the DIS, checks its LSDB, and then sends a PSNP to the DIS to request the LSPs that it does not have. For example, RTC does not have the LSPs of RTA and RTB.
  - RTB (the DIS) receives the PSNP and then sends the LSPs required by RTC for LSDB synchronization.

Routing Algorithms



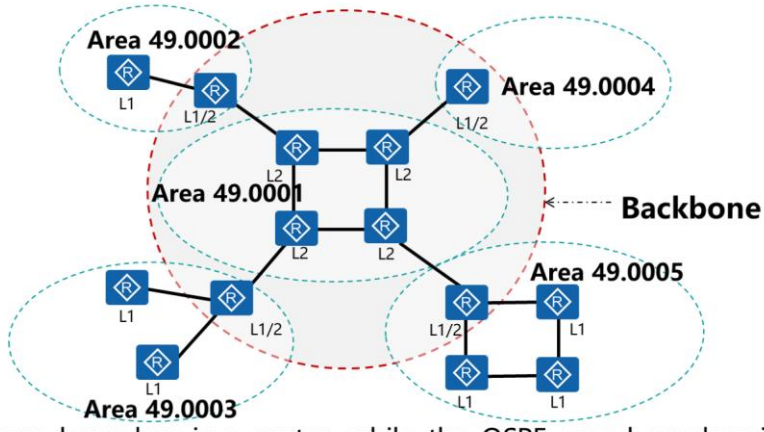
- IS-IS route cost calculation:
  - The default interface cost is 10.



- SPF calculation process:
  - Complete LSDB synchronization within a single area.
  - Generate the network topology.
  - Generate the shortest path tree, using the local node as the root.

- IS-IS route calculation characteristics:
  - A router in an area performs the full SPF algorithm when it starts for the first time.
  - If the received LSPs are updated and part of the topology is changed, the router performs incremental SPF (ISPF) algorithm.
  - If only routing information is changed, the router performs partial route calculation (PRC).
  - Because the algorithm that separates the topology and network is used, the route convergence speed is improved.
- IS-IS interface cost calculation:
  - Narrow mode: The default interface cost is 10, and the manually configured interface cost ranges from 1 to 63.
  - Wide mode: The default interface cost is 10, and the manually configured interface cost ranges from 1 to 16777215.
  - If the auto-cost enable command is executed in an IS-IS process, the interface cost is calculated based on the interface bandwidth in both narrow mode and wide mode, and only the reference rules vary slightly in the two modes.

Network Hierarchy and Routing Domain

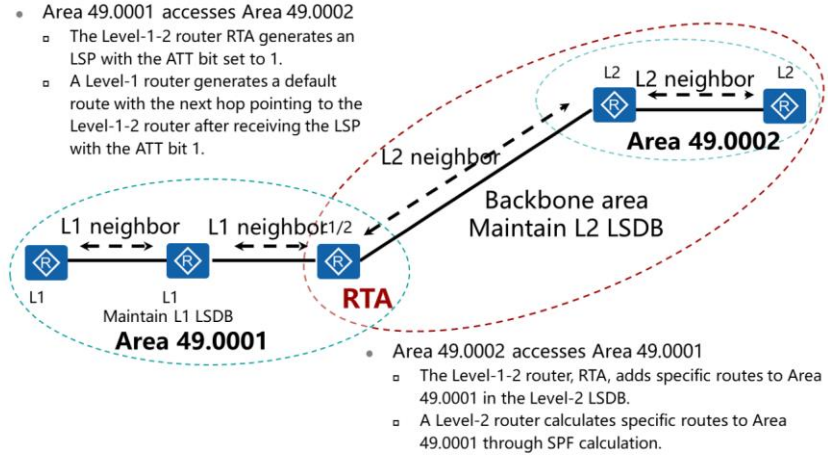


- The IS-IS area boundary is a router, while the OSPF area boundary is a router interface.

- IS-IS topology:
  - IS-IS uses a two-level hierarchy (backbone area and non-backbone area) to support large-scale routing networks. Generally, Level-1 routers are deployed in non-backbone areas, whereas Level-2 and Level-1-2 routers are deployed in the backbone area. Each non-backbone area connects to the backbone area through Level-1-2 routers.
  - The figure shows a network that runs IS-IS. The network is similar to an OSPF network typology with multiple areas. The backbone area contains all Level-2 routers and Level-1-2 routers.
  - Level-1-2 routers can belong to different areas. In Level-1 area, Level-1-2 routers maintain a Level-1 LSDB. In Level-2 area, Level-1-2 routers maintain a Level-2 LSDB.
- As shown in the preceding figure, differences between IS-IS and OSPF are as follows:
  - In IS-IS, each link can belong to different areas. In OSPF, each link belongs to only one area.
  - In IS-IS, no area is defined as the backbone area. In OSPF, Area 0 is defined as the backbone area.
  - In IS-IS, Level-1 and Level-2 routers use the SPF algorithm to generate their shortest path trees. In OSPF, the SPF algorithm is used only in the same area, and inter-area routes need to be forwarded through the backbone area.

Inter-Area Routing

- Area 49.0001 accesses Area 49.0002
  - The Level-1-2 router RTA generates an LSP with the ATT bit set to 1.
  - A Level-1 router generates a default route with the next hop pointing to the Level-1-2 router after receiving the LSP with the ATT bit 1.



- Area 49.0002 accesses Area 49.0001
  - The Level-1-2 router, RTA, adds specific routes to Area 49.0001 in the Level-2 LSDB.
  - A Level-2 router calculates specific routes to Area 49.0001 through SPF calculation.

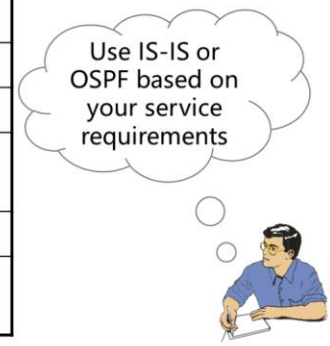
HUAWEI

- Level-1 router characteristics:
  - It has only a Level-1 LSDB. Its LSDB has only LSPs of routers in the local area.
  - Its routing table does not contain routing information of other areas.
  - Its routing table has one default route with the next hop pointing to a Level-1-2 router.
- Level-2 router characteristics:
  - It has only a Level-2 LSDB. Its LSDB has LSPs of routers in the backbone area but no LSPs of Level-1 routers.
  - Its routing table contains routing information of the entire network.
- Level-1-2 router characteristics:
  - It has both Level-1 and Level-2 LSDBs. The Level-1 LSDB has LSPs of routers in the local area, and the Level-2 LSDB has LSPs of routers in the backbone area.
  - It sets the ATT bit to 1 in its generated Level-1 LSP.
  - Its routing table contains routing information of the entire network.

Contents

Differences Between IS-IS and OSPF

Item	IS-IS	OSPF
Network types	Less	More
Cost style	Complicated	Simple
Area types	Less	More
Packet types	Simple	Various
Route convergence	Faster	Fast
Scalability	High	Medium
Routing load capacity	Higher	High



- Network types and cost style:
  - IS-IS supports only two network types and defines the same default cost for all bandwidths. OSPF supports four network types and defines the cost based on bandwidth.
- Area types:
  - IS-IS areas are classified into Level-1 and Level-2 areas. Level-2 area is the backbone area and has all specific routes. There are only default routes from Level-1 to Level-2 areas. OSPF areas are classified into the backbone area, common area, and special area. Devices in common and special areas must communicate across the backbone area.
- Packet types:
  - IS-IS uses only LSPs to transmit routing information and does not differentiate internal and external routing information, so IS-IS is simple and efficient. OSPF uses a variety of LSAs to transmit routing information, including Types 1, 2, 3, 4, 5, and 7, and applies to fine-grained scheduling and calculation.

- Routing algorithm:
- In IS-IS, when changes occur on the network segment where a node in an area resides, PRC algorithm is triggered, ensuring fast route convergence and low route calculation cost. In OSPF, network address is used for building the network topology, and ISPF algorithm is triggered when the network segment address in an area is changed, which is complicated.
- Scalability:
- In IS-IS, all routing information is transmitted using TLVs, ensuring simple structure and providing easy scalability. For example, to support IPv6, only two TLVs are added to IS-IS. Additionally, IS-IS also supports protocols such as IPX. OSPF is developed to support IP and provides two independent versions OSPFv2 and OSPFv3 to support IPv4 and IPv6.

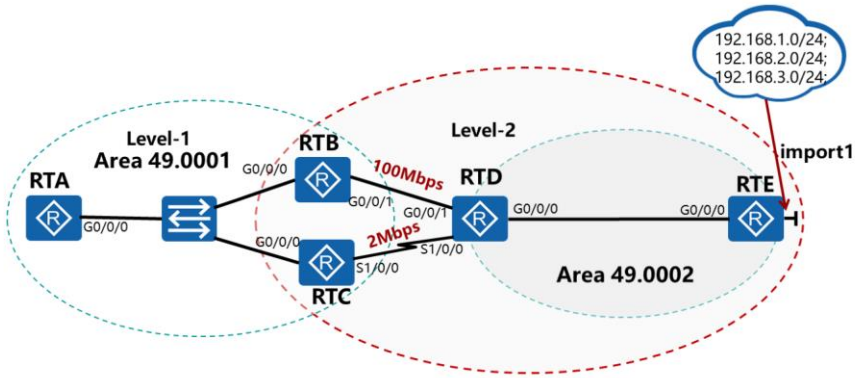
Terms

Abbreviation	OSI Term	IETF Term
IS	Intermediate System	Router
ES	End System	Host
DIS	Designated Intermediate System	Designated Router in OSPF
SysID	System ID	Router ID in OSPF
LSP	Link State PDU	LSA in OSPF
IIH	IS-IS Hello PDU	Hello packet in OSPF
PSNP	Partial Sequence Number PDU	LSR or LSAck in OSPF
CSNP	Complete Sequence Number PDU	DD packet in OSPF

Contents

HUAWEI

IS-IS Route Configuration Requirements

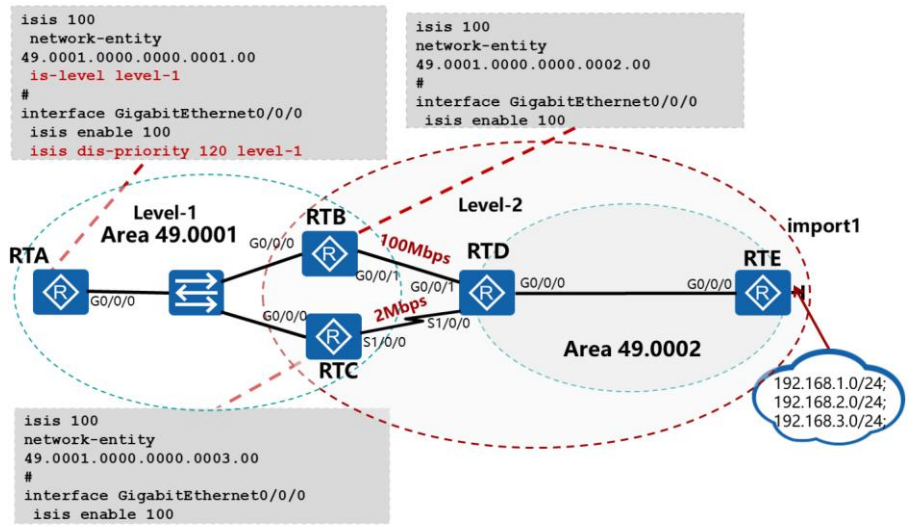


- In the figure, all routers in the customer network need to run IS-IS to ensure reachable routes on the network. All IS-IS processes use the process ID 100. RTA is the DIS in Area 49.0001. The network between RTD and RTE must be P2P network. RTE imports the direct route 192.168.X.X and requires RTA to access Area 49.0002 through the optimal path.
- Perform the configuration correctly to meet the preceding customer requirements.

HUAWEI

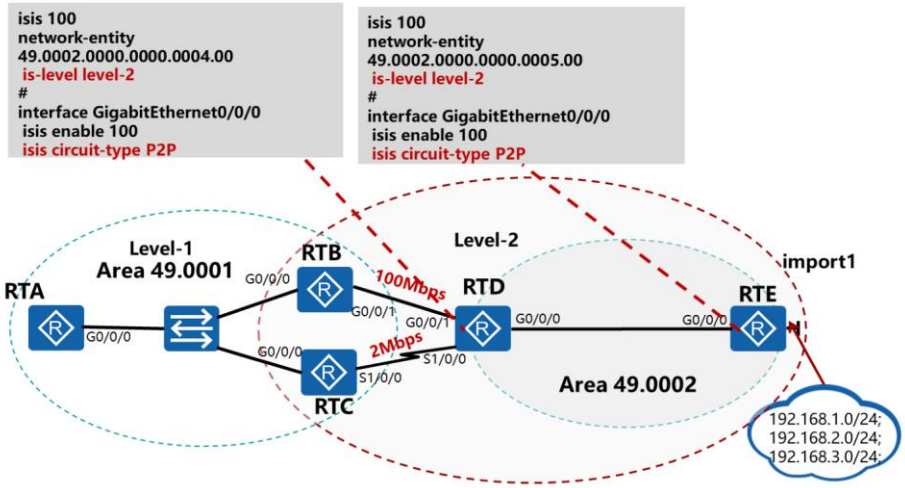
- NET address number:
  - RTA: 49.0001.0000.0000.0001:00
  - RTB: 49.0001.0000.0000.0002:00
  - RTC: 49.0001.0000.0000.0003:00
  - RTD: 49.0001.0000.0000.0004:00
  - RTE: 49.0001.0000.0000.0005:00

IS-IS Route Configuration (1)



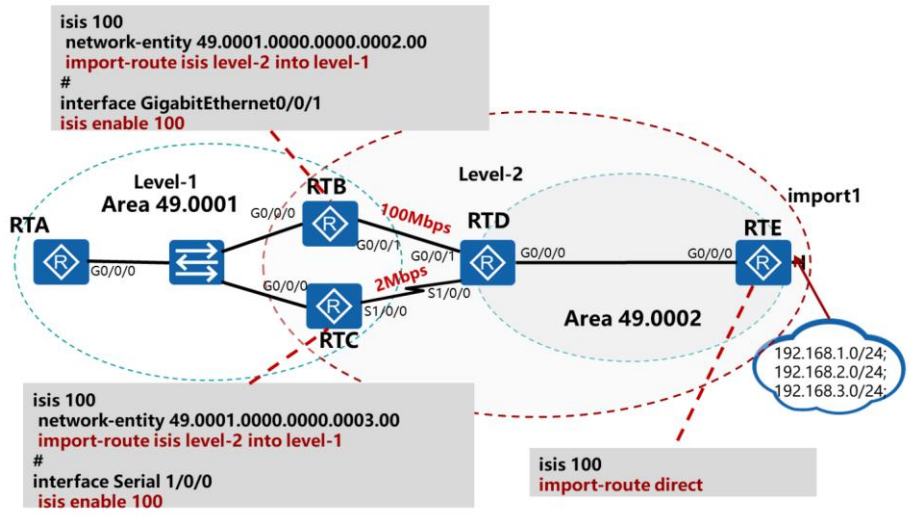
- Intra-area routing configuration roadmap:
  - Service configuration of Area 49.0001:
    - Configure a NET for each router in IS-IS process 100.
    - Configure RTA as a Level-1 router in IS-IS process 100 and retain the default IS-IS level of RTB and RTC as Level-1-2 routers.
    - Enable IS-IS on interfaces of RTA, RTB, and RTC.
    - Change the DIS priority of RTA to be the highest priority so that RTA becomes the DIS.

IS-IS Route Configuration (2)



- Intra-area routing configuration roadmap:
  - Service configuration of Area 49.0002:
    - Configure a NET for each router in IS-IS process 100.
    - Configure RTD and RTE as Level-2 routers in IS-IS process 100.
    - Enable IS-IS on interfaces of RTD and RTE.
    - Change the network type of interfaces on RTD and RTE as P2P.

IS-IS Route Configuration (3)



- Inter-area routing configuration roadmap:
  - Configure RTB as a Level-1-2 router and configure a NET for RTC in IS-IS process 100.
  - Enable IS-IS on interfaces.
  - Import the direct route on RTE.
- Route leaking:
  - If a Level-1 area has more than two Level-1-2 routers, a Level-1 router in the Level-1 area accesses other areas through the nearest Level-1-2 router, but only the intra-area cost is calculated. If the cost from the nearest Level-1-2 router in a Level-2 area to the destination network is high, the sub-optimal path exists. In this situation, route leaking needs to be performed to import specific routes (including the cost) of the Level-2 area into the Level-1 area, and then the Level-1 router calculates the optimal path to access other areas. This example requires RTA to access Area 49.0002 through the optimal path. Because the bandwidth of the link connecting RTB to RTD is relatively high, it is better to transmit traffic through RTB. To do this, in the IS-IS processes of RTB and RTC, import Level-2 routes into the Level-1 area. The LSDB of RTA contains all specific routes of the Level-2 area so that RTA can select the optimal path to reach Area 49.0002.

Quiz

1. How many IS-IS router types exist?
2. What is the function of PSNPs in neighbor interaction?
3. What are the advantages of IS-IS compared with OSPF?

- Answer: IS-IS routers are classified into Level-1 router, Level-2 router, and Level-1-2 router.
- Answer: PSNPs are used to request and acknowledge LSPs.
- Answer: IS-IS has simple packet structure, strong route transmission capacity, well-designed routing algorithm, and high scalability.


Thank You
www.huawei.com

BGP Principles and Configurations

Foreword

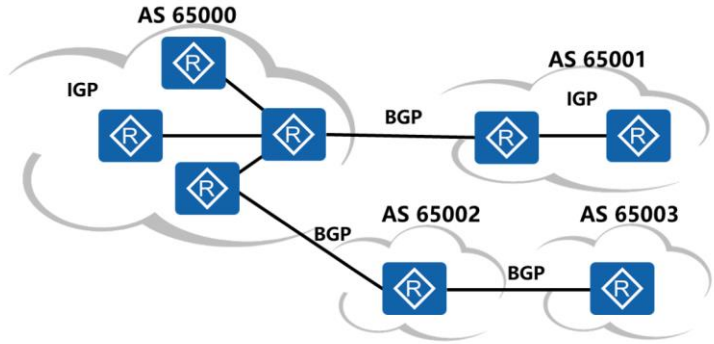
- Autonomous System (AS) is introduced into the Exterior Gateway Protocol (EGP). An AS is a set of routers under a single technical administration and using the same routing policies.
- The Interior Gateway Protocol (IGP) is used within an AS to calculate and discover routes. Routers within the same AS trust each other, so IGP route calculation and information flooding are completely open and require little manual intervention.
- Inter-AS connection requirements promote the development of EGP. BGP is an EGP and used to control routes and select optimal routes between ASs.

Objectives

- Upon completion of this section, you will be able to:
  - Understand BGP principles
  - Master BGP attributes and applications
  - Be familiar with BGP route aggregation applications

Contents

Basic BGP Functions

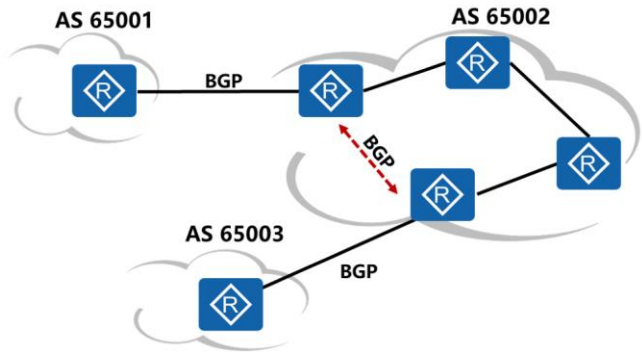


- IGPs, such as OSPF, IS-IS, and RIP, are used within an AS to calculate and discover routes.
- BGP is used between ASs to transmit and control routes.

HUAWEI

- EGP, the predecessor of BGP, is simple in design and can only transmit routing information between ASs and cannot select optimal routes or prevent routing loops between ASs. Therefore, EGP was replaced by BGP.
- Compared with EGP, BGP has the following advantages:
  - Discovers neighbors and establishes neighbor relationships.
  - Selects optimal routes and advertises routes.
  - Prevents routing loops, efficiently transmits routes, and maintains a large amount of routing information.
  - Provides various route control capabilities between ASs that are not fully trusted.
- Using BGP to transmit routing information, a routing domain functions as a whole to exchange routing information with another routing domain. This routing domain is an AS. An AS is a set of routers and networks that consist of these routers. These routers are under a single technical administration and use the same routing policies.
- An AS is uniquely identified by an AS number, which is assigned by the Internet Assigned Numbers Authority (IANA). Before January 2009, only 2-byte AS numbers can be used, which range from 1 to 65535. AS numbers 1 to 64511 are public AS numbers, and AS numbers 64512 to 65534 are private AS numbers. After January 2009, the IANA decided to use 4-byte AS numbers, which range from 65536 to 4294967295.

BGP Characteristics

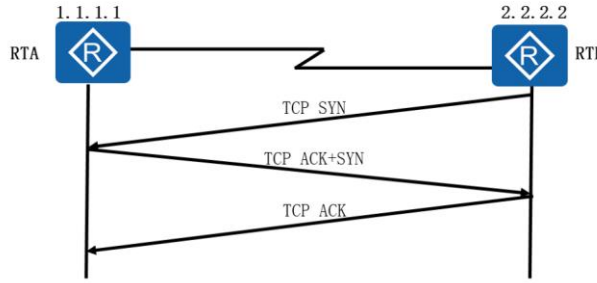


- In the figure, two BGP routers can establish a neighbor relationship across multiple routers.
- To implement on-demand route control and selection, various BGP attributes are designed and carried in routes.

- To ensure reliable data transmission between ASs, BGP uses TCP to establish connections. Therefore, BGP can establish a neighbor relationship across multiple routers, while IGP can only establish a neighbor relationship hop by hop.
- Routers between ASs do not completely trust each other. To implement on-demand route control and selection, various BGP attributes are designed.

Contents

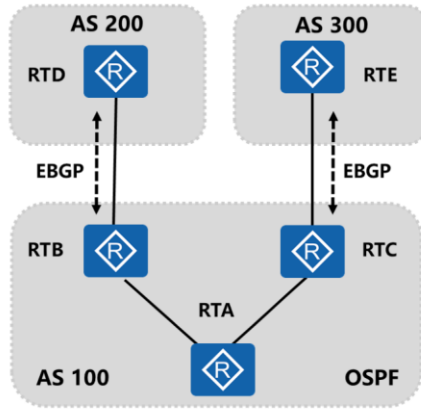
BGP Neighbor Discovery



- The device that starts BGP first initiates a TCP connection. In the figure, RTB first starts BGP and uses a random port number to initiate a TCP connection with port 179 of RTA.

- BGP is designed to run between ASs to transmit routers. There are WAN links between ASs, and unpredictable link congestion or packet loss may occur during packet transmission on WANs. Therefore, BGP uses TCP as the transport protocol to ensure reliability.
- BGP uses TCP port 179 to establish neighbor relationships, and TCP establishes connections in unicast mode. Therefore, unlike RIP and OSPF, BGP does not discover neighbors in multicast mode. Establishing connections in unicast mode requires neighbors to be manually specified in BGP.

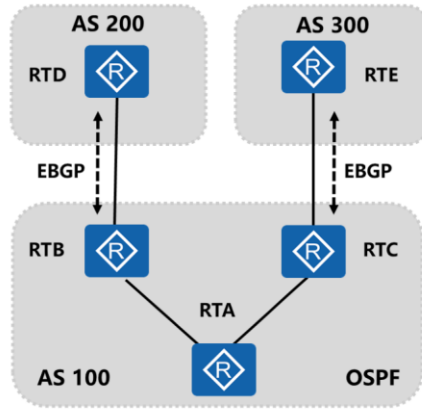
BGP Neighbor Type - EBGP



- BGP routers in different ASs establish EBGP neighbor relationships.

- EBGP transmits routes only between different ASs. In the figure, RTB and RTC in AS 100 can learn different routes from AS 200 and AS 300 respectively. How to transmit routes of AS 200 and AS 300 within AS 100?
- To meet this requirement, on RTB and RTC, import BGP routes into IGP (OSPF in the figure) and then import IGP routes back into BGP.
- However, this method has the following disadvantages:
  - There are a huge number of BGP routes on the public network. After these BGP routes are imported into IGP, IGP cannot support these BGP routes.
  - When BGP routes are imported into IGP, strict control is required. This complicates the configuration and maintenance.
  - When BGP attributes carried in BGP routes are imported into IGP, these attributes may be lost because they cannot be identified by IGP.
- To overcome these disadvantages, BGP needs to be designed to transmit routes within an AS.

BGP Neighbor Type - IBGP

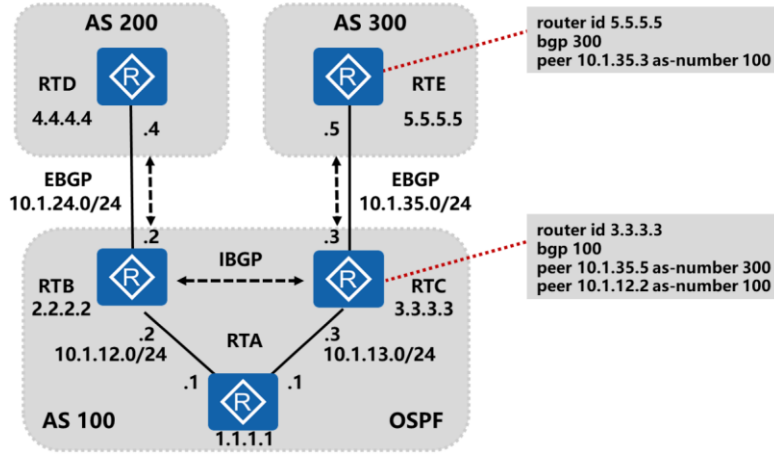


- BGP routers in the same AS establish IBGP neighbor relationships.

HUAWEI

- BGP uses TCP as the transport protocol. Therefore, BGP can establish neighbor relationships across multiple devices. In the figure, RTB and RTC establish an IBGP neighbor relationship and transmit the routes learned from other ASs to each other so that BGP routes can be transmitted within an AS.

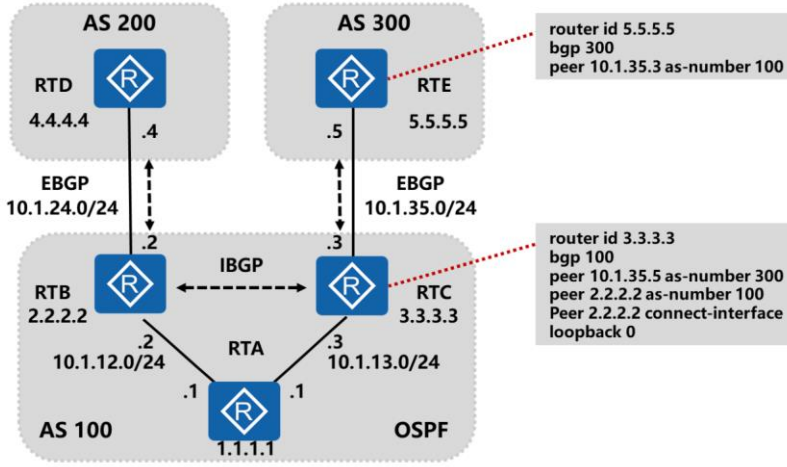
BGP Neighbor Relationship Configuration



- Configuration procedure:
  - Configure a router ID to identify a router.
  - Configure an EBGP neighbor relationship to transmit routes between ASs.
  - Configure an IBGP neighbor relationship to transmit routes within an AS.
- Description:
  - If no router ID is configured for a BGP router, it automatically selects a router ID according to the following rules:
    - Selects the highest IP address among all loopback interfaces.
    - Selects the highest IP address among all physical interfaces if it does not have loopback interfaces.
    - Configuration command: router id X.X.X.X
  - BGP neighbor relationship type is identified by the configured AS number. The parameter following the peer keyword indicates the interface IP address of the neighbor, and the parameter following the as-number keyword indicates the AS number of the neighbor. If two routers have the same AS number, they establish an IBGP neighbor relationship. If they have different AS numbers, they establish an EBGP neighbor relationship.

- The peer keyword indicates the IP address used by the neighbor to establish a BGP neighbor relationship, identifying the destination address of the TCP connection initiated with the neighbor. This address can be the IP address of the neighbor’ s directly connected interface or the IP address of an indirectly connected loopback interface (ensure that this IP address is reachable). Loopback interface IP addresses are often used to establish IBGP neighbor relationships. This is because loopback interfaces are always Up after being enabled. As long as routes are reachable, IBGP neighbor relationships remain stable. Directly connected interface IP addresses are often used to establish EBGP neighbor relationships. This is because EBGP neighbor relationships are established between ASs and routes between indirectly connected interfaces are unreachable before the neighbor relationships are established.

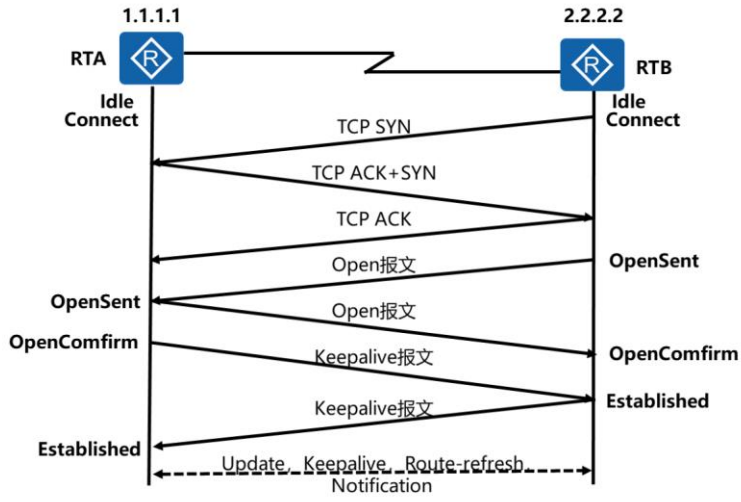
BGP Neighbor Relationship Configuration Optimization



HUAWEI

- Directly connected interface IP addresses are often used to establish EBGP neighbor relationships, and loopback interface IP addresses are often used to establish IBGP neighbor relationships.

BGP Neighbor Relationship Establishment



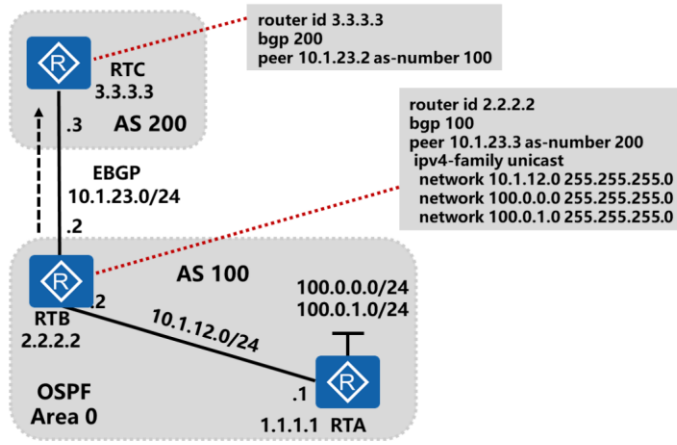
- BGP routers exchange BGP messages to establish neighbor relationships and update routing information. BGP messages are classified into Open, Update, Notification, Keepalive, and Route-refresh messages.
  - Open message: is the first message sent after a TCP connection is established. It is used to establish a BGP connection between neighbors. After a BGP neighbor receives an Open message and negotiation succeeds, the neighbor sends a Keepalive message to confirm and retain the connection. Then BGP neighbors can exchange Update, Notification, Keepalive, and Route-refresh messages.
  - Update message: is used to exchange routing information between BGP neighbors. It can advertise multiple reachable routes with the same route attributes and withdraw multiple unreachable routes.
    - An Update message can advertise multiple reachable routes with the same route attributes. These routes can share a group of route attributes. All the route attributes carried in a specific Update message apply to all the destinations (specified by IP prefixes) of the Network Layer Reachability Information (NLRI) field in this Update message.
    - An Update message can withdraw multiple unreachable routes. Each withdrawn route identified by a destination address (an IP prefix) is the route that was advertised between BGP routers.

- You can use an Update message just to withdraw routes so that this message does not need to contain the path attribute or NLRI. Alternatively, you can use an Update message just to advertise reachable routes so that this message does not need to contain withdrawn route information.
- Notification message: is sent when a BGP router detects an error. Then a BGP connection is terminated immediately.
- Keepalive message: is sent periodically from a BGP router to its neighbor to retain their connection.
- Route-refresh message: is sent by a BGP router to request its neighbor to send routing information again after this router changes its routing policy.
- During message exchange, the Idle state is the initial state of a BGP router. In Idle state, the BGP router rejects the connection request from its neighbor. Only after receiving the Start event of itself, the BGP router tries to establish a TCP connection with its neighbor and transitions to the Connect state.
- In Connect state, the BGP router starts the Connect Retry timer, waiting for a TCP connection to be established.
- If a TCP connection is established, the BGP router sends an Open message to its neighbor and transitions to the OpenSent state.
- If a TCP connection fails to be established, the BGP router transitions to the Active state.
- If the BGP router does not receive any response from its neighbor until the Connect Retry timer expires, the BGP router continues to try to establish a TCP connection with its neighbor and stays in the Connect state.
- In Active state, the BGP router always tries to establish a TCP connection.
- If a TCP connection is established, the BGP router sends an Open message to its neighbor, turns off the Connect Retry timer, and transitions to the OpenSent state.
- If a TCP connection fails to be established, the BGP router stays in the Active state.
- If the BGP router does not receive any response from its neighbor until the Connect Retry timer expires, the BGP router transitions to the Connect state.
- In OpenSent state, the BGP router waits for an Open message from its neighbor and checks information carried in the Open message, including AS number, version number, and authentication password.
- If the received Open message is correct, the BGP router sends a Keepalive message to its neighbor and transitions to the OpenConfirm state.

- If the received Open message is incorrect, the BGP router sends a Notification message to its neighbor and transitions to the Idle state.
- In OpenConfirm state, the BGP router waits for a Keepalive or Notification message from its neighbor. If it receives a Keepalive message, it transitions to the Established state. If it receives a Notification message, it transitions to the Idle state.
- In Established state, the BGP router can exchange Update, Keepalive, Route-refresh, and Notification messages with its neighbor.

Contents

BGP Route Generation Mode – Network (1)

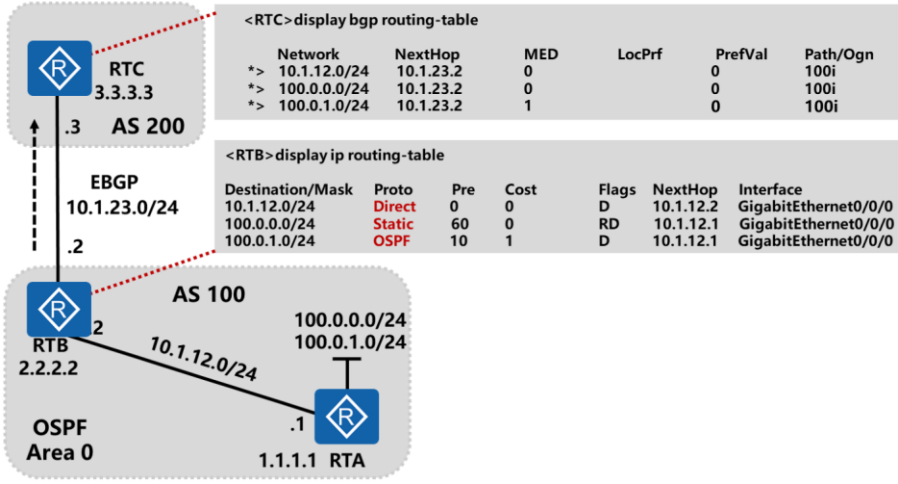


- The network command imports existing routes in an IP routing table into a BGP routing table one by one.

HUAWEI

- Two BGP route generation modes are available: using the network command and using the import command.
- In the figure, RTA has two user network segments 100.0.0.0/24 and 100.0.1.0/24, and RTB has a static route to 100.0.0.0/24 and learns the route to 100.0.1.0/24 through OSPF. RTB and RTC establish an EBGP neighbor relationship, and RTB advertises the routes 100.0.0.0/24, 100.0.1.0/24, and 10.1.12.0/24 using the network command so that RTC can learn the routes in the routing table of RTB.

BGP Route Generation Mode – Network (2)



<RTC>display bgp routing-table

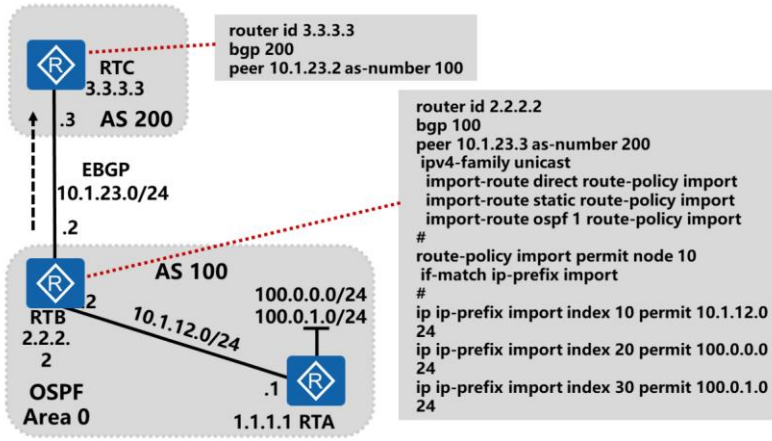
	Network	NextHop	MED	LocPrf	PrefVal	Path/Ogn
*>	10.1.12.0/24	10.1.23.2	0		0	100i
*>	100.0.0.0/24	10.1.23.2	0		0	100i
*>	100.0.1.0/24	10.1.23.2	1		0	100i

<RTB>display ip routing-table

Destination/Mask	Proto	Pre	Cost	Flags	NextHop	Interface
10.1.12.0/24	Direct	0	0	D	10.1.12.2	GigabitEthernet0/0/0
100.0.0.0/24	Static	60	0	RD	10.1.12.1	GigabitEthernet0/0/0
100.0.1.0/24	OSPF	10	1	D	10.1.12.1	GigabitEthernet0/0/0

- Run the display command on RTC to check whether it learns routes advertised by BGP.

BGP Route Generation Mode – Import (1)

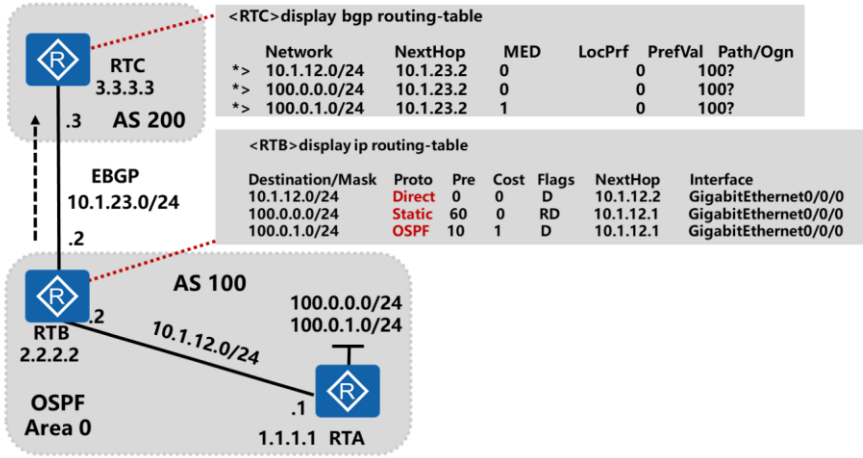


- The import command imports routes into a BGP routing table based on the running routing protocol (RIP, OSPF, or IS-IS). This command can also import directly connected routes and static routes.

  HUAWEI

- In the figure, RTA has two user network segments 100.0.0.0/24 and 100.0.1.0/24, and RTB has a static route to 100.0.0.0/24 and learns the route to 100.0.1.0/24 through OSPF. RTB and RTC establish an EBGP neighbor relationship, and RTB advertises the routes 100.0.0.0/24, 100.0.1.0/24, and 10.1.12.0/24 using the import command so that RTC can learn the routes of RTB.
- To prevent other routes from being imported into BGP, you need to configure IP-prefix for precise matching and apply route-policy to control the imported routes.

BGP Route Generation Mode - Import (2)



<RTC>display bgp routing-table

	Network	NextHop	MED	LocPrf	PrefVal	Path/Ogn
*>	10.1.12.0/24	10.1.23.2	0		0	100?
*>	100.0.0.0/24	10.1.23.2	0		0	100?
*>	100.0.1.0/24	10.1.23.2	1		0	100?

<RTB>display ip routing-table

Destination/Mask	Proto	Pre	Cost	Flags	NextHop	Interface
10.1.12.0/24	Direct	0	0	D	10.1.12.2	GigabitEthernet0/0/0
100.0.0.0/24	Static	60	0	RD	10.1.12.1	GigabitEthernet0/0/0
100.0.1.0/24	OSPF	10	1	D	10.1.12.1	GigabitEthernet0/0/0

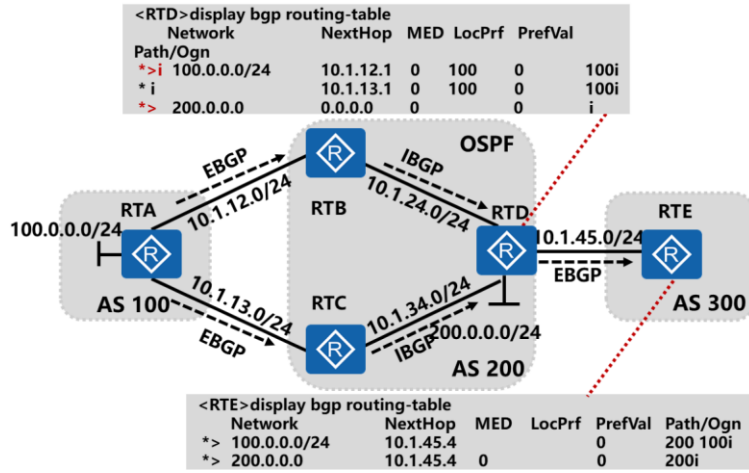
- Run the display command on RTC to check whether it learns routes imported into BGP.

Contents

BGP Update Message

- BGP routes are generated in either network or import mode. They are encapsulated in Update messages and advertised to neighbors. BGP advertises routing information only after a neighbor relationship is established.
- Update messages are used to advertise reachable routes and withdraw unreachable routes. An Update message contains the following information:
  - Network Layer Reachability Information (NLRI): advertises the IP prefix and prefix length.
  - Path attribute: provides loop detection and controls optimal route selection.
  - Withdrawn route: describes the prefix and prefix length of the unreachable withdrawn route.
- BGP route advertisement must follow specific rules to prevent potential problems.

BGP Route Advertisement Rule (1)

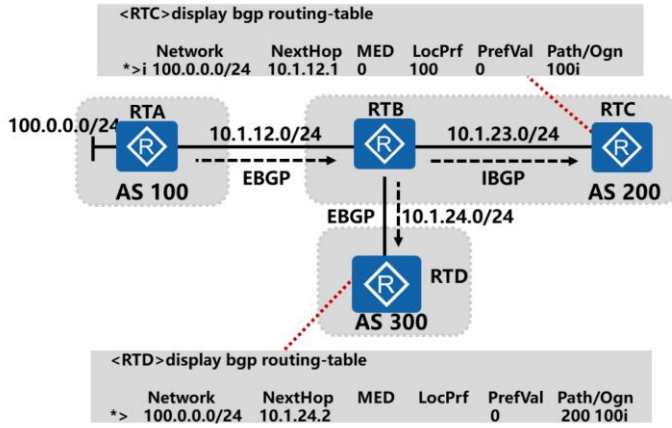


- BGP Route Advertisement Rule 1: Advertise Only the Optimal Route to Neighbors

  HUAWEI

- When multiple valid routes exist, a BGP router advertises only the optimal route to its neighbor.
  - RTD can learn the route 100.0.0.0/24 from two BGP neighbors (RTB and RTC) and RTD advertises its directly connected route 200.0.0.0/24 into BGP. Run the display bgp routing-table command on RTD. The following command output is displayed:
  - Run the display bgp routing-table command on RTE. The following command output is displayed. You can view that RTD has advertised the optimal route marked valid to its BGP neighbor RTE.
- Fields in a BGP routing table include:
  - Status codes: * - valid, > - best, d - damped, h - history, i - internal, s - suppressed, S - Stale
  - Origin : i - IGP, e - EGP, ? – incomplete
  - Network: network address
  - NextHop: next-hop address
  - MED: route metric
  - LocPrf: local preference
  - PrefVal: protocol preferred value
  - Path/Ogn: AS_Path and Origin attribute
  - Community: Community attribute information

BGP Route Advertisement Rule (2)

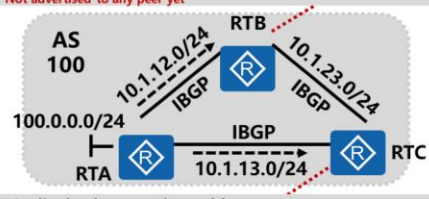


- BGP Route Advertisement Rule 2: Advertise the Optimal Route Obtained Through EBGP to All BGP Neighbors

- A BGP router advertises the optimal route obtained through EBGP to all BGP neighbors (including EBGP neighbors and IBGP neighbors).
  - In the figure, RTA has a user network segment 100.0.0.0/24 and advertises this network segment to a BGP neighbor RTB through EBGP. After RTB receives this route from its EBGP neighbor, it advertises this route to its IBGP neighbor RTC and EBGP neighbor RTD.

BGP Route Advertisement Rule (3)

```
<RTB>display bgp routing-table 100.0.0.0
BGP local router ID : 2.2.2.2
 Local AS number : 100
 Paths:  1 available, 1 best, 1 select
 BGP routing table entry information of 100.0.0.0/24:
 From: 10.1.12.1 (1.1.1.1)
 Route Duration: 00h01m39s
 Relay IP Nexthop: 0.0.0.0
 Relay IP Out-Interface: GigabitEthernet0/0/0
 Original nexthop: 10.1.12.1
 Qos information : 0x0
 AS_Path Nil, origin igp, MED 0, localpref 100, pref-val 0, valid, internal, best, select, active, pre 255
 Not advertised to any peer yet
```



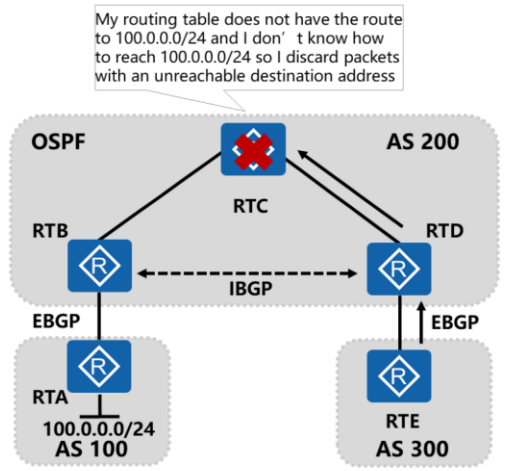

```
<RTC>display bgp routing-table
     Network         NextHop        MED        LocPrf    PrefVal Path/Ogn
*>i  100.0.0.0/24    10.1.13.1      0          100       0       i
```

- BGP Route Advertisement Rule 3: Do Not Advertise the Optimal Route Obtained Through IBGP to Other IBGP Neighbors

HUAWEI

- A BGP router does not advertise the optimal route obtained through IBGP to other IBGP neighbors.
  - In the figure, RTA has a user network segment 100.0.0.0/24. RTA, RTB, and RTC are IBGP neighbors. RTA advertises the route 100.0.0.0/24 to RTB and RTC through IBGP, but RTB does not advertise the received IBGP route to its IBGP neighbor RTC.
  - This design prevents routing loops within an AS. As defined, when a BGP route is transmitted within an AS, its AS_Path attribute remains unchanged. In the figure, when RTA advertises the route 100.0.0.0/24 to RTB, the AS_Path attribute of this route remains unchanged and is empty. If RTB can advertise this IBGP route to RTC, RTC may also advertise this route to RTA because the AS_Path attribute of the route is still empty, and RTA will not reject this route. As a result, a routing loop occurs. Therefore, this route advertisement rule can prevent routing loops within an AS.

BGP Route Advertisement Rule (4)

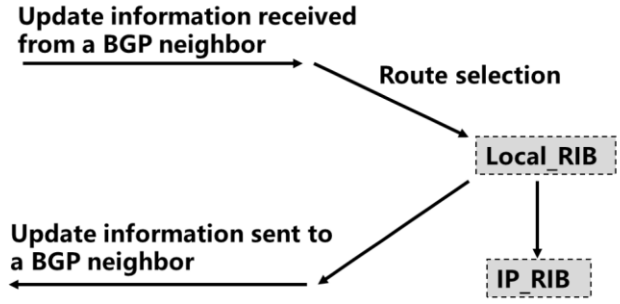


- BGP Route Advertisement Rule 4: Synchronize BGP and IGP

- RTA has a user network segment 100.0.0.0/24 and advertises it to RTB through EBGP. RTB and RTD establish an IBGP neighbor relationship. RTD learns this BGP route through IBGP and advertises it to the EBGP neighbor RTE.
- When RTE accesses the network segment 100.0.0.0/24, it examines its routing table, finding that the next hop of the route to 100.0.0.0/24 is RTD. After RTE finds the outbound interface, it sends a packet to RTD. RTD receives the packet and examines its routing table, finding that the next hop of the route is RTB and the outbound interface is the interface connected to RTC and sends the packet to RTC. RTC receives the packet and examines its routing table, finding that there is no route to 100.0.0.0/24 and discards this packet. In this situation, the routing blackhole problem occurs.
- BGP route advertisement rule: Before a BGP router advertises a route learned from an IBGP neighbor to another BGP neighbor, IGP must know this route. That is, BGP must synchronize with IGP.

- In the figure, after RTD receives an IBGP route from RTB, RTD needs to check whether IGP (OSPF) has learned this route before advertising this route to RTE. If OSPF can learn this route, RTD advertises it to RTE.
- By default, synchronization check between BGP and IGP is disabled on Huawei routers to ensure normal IBGP route advertisement. However, disabling synchronization check will lead to the routing blackhole problem. To solve this problem, two methods are available:
- Import BGP routes into IGP to ensure synchronization between BGP and IGP. However, the number of BGP routes on the Internet is huge and importing so many BGP routes into IGP will bring a huge processing and storage burden to an IGP router. If the IGP router is overloaded, it may crash.
- IBGP routers must be fully meshed to ensure that all routers can learn advertised routes. This method can solve the routing blackhole problem occurring after synchronization check is disabled.

BGP Routing Information Processing



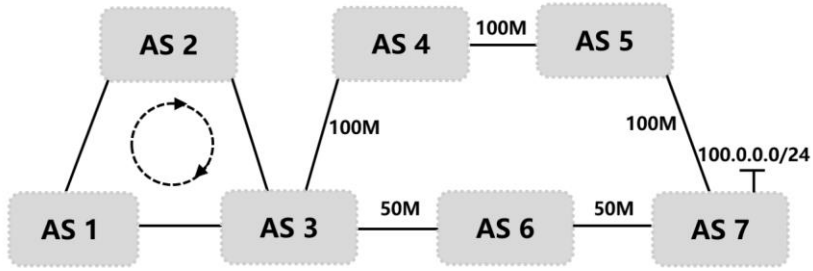
- When receiving an Update message from a BGP neighbor, a BGP router uses the route selection algorithm to determine the optimal route for each prefix.
- The router stores the selected optimal route to the local BGP routing table (Local_RIB) and then submits it to the local IP routing table (IP_RIB) to determine whether to install it.
- The router encapsulates the selected valid optimal route in an Update message and sends it to the BGP neighbor.

  HUAWEI

- IP routing table (IP_RIB): global routing information database, including all IP routing information.
- BGP routing table (Local_RIB): BGP routing information database, including routes selected by the local BGP router, neighbor table, and neighbor list.
- After receiving an Update message from a BGP neighbor, a BGP router uses the route selection algorithm to determine the optimal route for each prefix and stores the selected optimal route to the local BGP routing table (Local_RIB).
- If multipath is enabled on a BGP router, it submits the optimal route and all equal-cost routes to IP_RIB to determine whether to install them. In addition to the optimal route received from BGP neighbors, Local_RIB also includes the routes injected by the router. These routes are called locally originated routes.
- In Local_RIB, a router encapsulates only the optimal prefix in an Update message and advertises it to BGP neighbors.

Contents

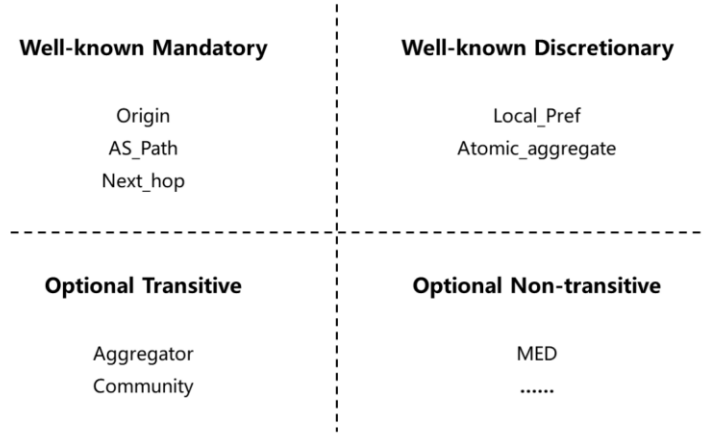
BGP Route Selection Problems



- In the figure, AS 7 has a user network segment 100.0.0.0/24 and advertises it to each AS through BGP. Each AS can learn the route to 100.0.0.0/24. However, the following problems exist during route transmission:
  - AS 3 can receive the route to 100.0.0.0/24 from AS 4 and AS 6, and the link between AS 3 and AS 4 has higher bandwidth. How to enable AS 3 to access the network segment 100.0.0.0/24 through AS 4?
  - There is a topology loop among AS 1, AS 2, and AS 3. Therefore, a loop may occur during packet transmission. How to prevent this loop?

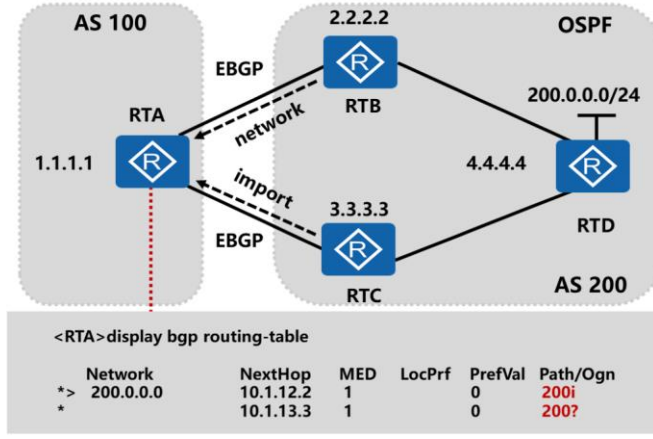
- Solutions to the two problems:
  - During routing information exchange between ASs, various BGP attributes are designed to flexibly control routes and select the optimal route.
    - 1. Adjust the link metric between ASs to change routing entries in the routing table. 2. Use routing policies to change the next hop of routes. However, these methods have limitations in some situations and cannot meet various network requirements.
  - During route transmission between ASs, the advertisement path is recorded to prevent loops.

Various BGP Attributes



- Well-known attributes must be identified and supported by all BGP routers.
  - Well-known mandatory attributes must be carried in a BGP Update message.
  - Well-known discretionary attributes do not need to be carried in a BGP Update message and can be selected as required.
- Optional attributes do not need to be identified by all BGP routers.
  - Optional transitive attributes cannot be identified by some BGP routers but can be carried in BGP messages and then advertised to neighbors.
  - A BGP router can ignore the messages carrying optional non-transitive attributes and does not advertise these message to neighbors.

BGP Attribute - Origin



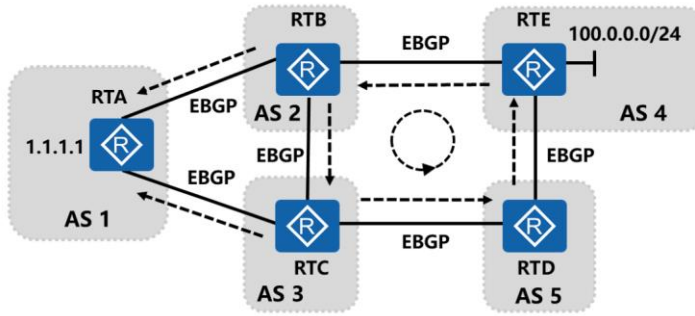

```
<RTA>display bgp routing-table

     Network          NextHop      MED    LocPrf    PrefVal  Path/Ogn
*>   200.0.0.0        10.1.12.2    1                0        200i
*                     10.1.13.3    1                0        200?
```

- The Origin attribute defines the origin of a route and how it becomes a BGP route.

  HUAWEI

- In the figure, OSPF runs within AS 200, and the network segment 200.0.0.0/24 is advertised into OSPF. RTB changes the route 200.0.0.0/24 into a BGP route by using the network command and advertises it to RTA. RTC changes the route 200.0.0.0/24 into a BGP route by using the import command and advertises it to RTA.
- BGP transmits routing information between ASs. If there are multiple routes to the same destination IP prefix and BGP learns these routes using different methods, the Origin attribute determines which route is selected as the optimal route and identifies the origins of these routes.
- Three Origin attributes are available:
  - i indicates that this BGP route is injected using the network command.
  - e indicates that this BGP route is learned through EGP. EGP is seldom used on the live network and the Origin attribute of a route can be changed to e using a routing policy.
  - ? is for Incomplete, indicating that this BGP route is learned using other methods, for example, a route is imported using the import command.
- The three Origin attributes can be listed in ascending order of priority as i > e > Incomplete (?).

BGP Attribute - AS_Path

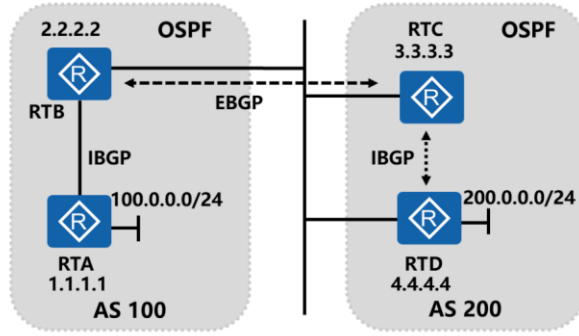


- In the figure:
  - RTA in AS 1 can learn the route 100.0.0.0/24 from RTB and RTC. How does RTA select the optimal route?
  - There are topology loops among RTA, RTB, and RTC, and among RTB, RTC, RTD, and RTE. Therefore, routing loops may occur during BGP route transmission. How does BGP prevent loops?

  HUAWEI

- The AS_Path attribute is designed to address the two problems. It records the numbers of all the ASs that a route passes through.
  - When RTA receives the route 100.0.0.0/24 from RTB, the AS_Path attribute is (2,4). When RTA receives the route 100.0.0.0/24 from RTC, the AS_Path attribute is (3,5,4). As defined, a shorter AS_Path attribute indicates a better route because it records fewer AS numbers. Therefore, RTA prefers the route 100.0.0.0/24 received from RTB.
  - When RTE advertises the route 100.0.0.0/24 through BGP, a routing loop may occur if the link RTE->RTB->RTC->RTD->RTE is used. To prevent the loop, RTE checks the AS_Path attribute of the route received from RTD. If RTE finds that this AS_Path attribute contains its AS number, it discards this route.
- Four AS_Path attributes are available:
  - AS_Sequence: will be described in BGP route aggregation.
  - AS_Set: will be described in BGP route aggregation.
  - AS_Confed_Sequence: is used in the BGP Confederation and not described in this course.
  - AS_Confed_Set: is used in the BGP Confederation and not described in this course.

BGP Attribute - Next_hop



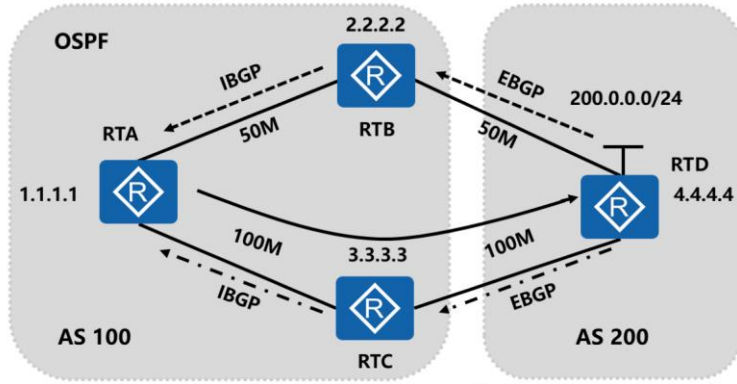
- In the figure:
  - When RTA advertises the network segment 100.0.0.0/24 to RTB, what is the Next_hop IP address?
  - When RTB advertises the network segment 100.0.0.0/24 to RTC, what is the Next_hop IP address?
  - When RTA learns from RTB the network segment 200.0.0.0/24 advertised by RTC, what is the Next_hop IP address?

  HUAWEI

- When a BGP router advertises a locally originated route to an IBGP neighbor, it sets the local interface IP address used to establish a neighbor relationship as the Next_hop attribute of this route.
  - In the figure, when RTA advertises the network segment 100.0.0.0/24 to RTB, the Next_hop attribute of this route is the IP address of the interface that directly connects RTA to RTB if RTA and RTB establish an IBGP neighbor relationship using directly connected interfaces. If they use loopback interfaces to establish an IBGP neighbor relationship, the Next_hop attribute of this route is the loopback interface IP address of RTA.
- When a BGP router advertises a route to an EBGP neighbor, it sets the interface IP address used to establish a neighbor relationship as the Next_hop attribute of this route.
  - In the figure, when RTB advertises the network segment 100.0.0.0/24 to RTC, the Next_hop attribute of this route is the IP address of the interface that directly connects RTB to RTC.

- When a BGP router advertises to an IBGP neighbor a route learned through EBGP, the Next_hop attribute of this route remains unchanged.
- When RTA learns from RTB the network segment 200.0.0.0/24 advertised by RTC, the Next_hop attribute of this route is the outbound interface IP address of RTD. Because RTB and RTD reside on the same network segment, the Next_hop attribute of the route advertised from RTC to RTB is the outbound interface IP address of RTD.
- The following explains the three situations:
- EBGP neighbors often use directly connected interfaces to establish neighbor relationships. When advertising routes to each other, an EBGP neighbor changes the Next_hop attributes of routes into its outbound interface IP address.
- IBGP neighbors often use loopback interfaces to establish neighbor relationships. After a route originated by this router is sent to a neighbor, the Next_hop IP address of this route is changed to the loopback interface IP address of the neighbor. In this situation, even if a link failure occurs on the network, as long as the Next_hop of this route is reachable, the router can still access the destination network segment. This improves network stability.
- When an IGP, for example, RIP advertises a route, the next hop of the route is changed each time it passes through a router, and each router that advertises this route declares that it can reach the destination address and transmits packets hop by hop to the destination network. However, routers on the network do not know which router originated this route. Subsequently, a loop occurs. To prevent loops, BGP changes the Next_hop attribute of a route only when this route is transmitted between EBGP neighbors and retains the Next_hop attribute when a route learned through EBGP is advertised between IBGP neighbors.

BGP Attribute - Local_Preference

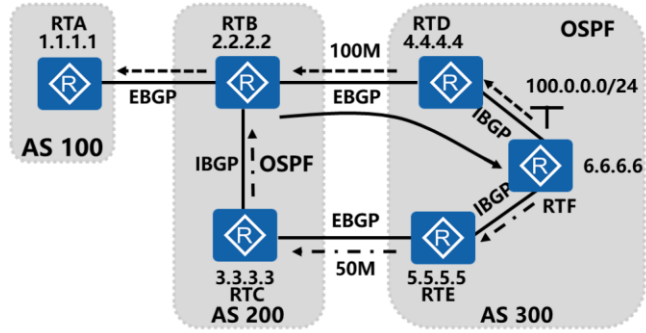


- The Local_Pref attribute is valid only between IBGP neighbors and not advertised to other ASs. This attribute indicates the BGP preference of a router and determines the optimal route for traffic leaving an AS.

  HUAWEI

- In the figure, AS 200 has a user network segment 200.0.0.0/24 and advertises it to AS 100 through BGP. How can the administrator of AS 100 access this network segment over a high-bandwidth link?
- Method to meet this requirement:
  - On RTC, configure ip-prefix to match the route 200.0.0.0/24, configure a route-policy to invoke this ip-prefix, set the Local_Preference of the route to 200, and apply the route-policy in the export direction.
- The Local_Pref attribute is valid only between IBGP neighbors and not advertised to other ASs. This attribute indicates the BGP preference of a router. A larger value indicates a higher preference.
- This attribute determines the optimal route for traffic leaving an AS. When a BGP router obtains from different IBGP neighbors multiple routes with the same destination address but different next hops, the router selects the route with the largest Local_Pref attribute value as the optimal route. The default Local_Pref attribute value is 100.

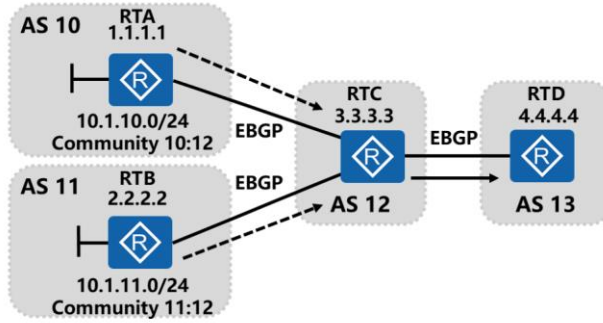
BGP Attribute - MED



- The Multi-Exit-Discriminator (MED) attribute is transmitted only between two neighboring ASs. An AS that receives this attribute does not advertise it to any other ASs. It determines the optimal route for traffic entering an AS.

- In the figure, the administrator of AS 300 wants to perform operations in AS 300 to enable AS 200 to access the network segment 100.0.0.0/24 over the high-bandwidth link.
- Method to meet this requirement:
  - On RTE, configure ip-prefix to match the route 100.0.0.0/24, configure a route-policy to invoke this ip-prefix, set the MED of the route to 100, and apply the route-policy in the export direction.
- The MED attribute is transmitted only between two neighboring ASs. An AS that receives this attribute does not advertise it to any other ASs. In the figure, AS 100 will not receive the MED configured in AS 300, but AS 200 will. Therefore, AS 200 can select the high-bandwidth link.
- The MED attribute corresponds to the metric in IGP. It determines the optimal route for traffic entering an AS. When a BGP router obtains from different EBGP neighbors multiple routes with the same destination address but different next hops, the router selects the route with the smallest MED attribute value as the optimal route. The default MED attribute value is 0.

BGP Attribute - Community



- The Community attribute has two functions:
  - Limits the route advertisement range.
  - Marks routes so that routes meeting the same conditions are handled in a unified manner.

  HUAWEI

- In the figure, AS 10 has a user network segment 10.1.10.0/24, and AS 11 has a user network segment 10.1.11.0/24. To differentiate the two user network segments, the Community attribute 10:12 is configured for the route 10.1.10.0/24 in AS 10, and the Community attribute 11:12 is configured for the route 10.1.11.0/24 in AS 11. After the two routes are advertised to AS 12 through BGP, AS 12 wants to aggregate the routes, advertises only the aggregated route to AS 13, and expects AS 13 not to transmit this route to other ASs.
- Method to meet this requirement:
  - On RTC, configure a Community-filter to match the routes with the Community attributes 10:12 and 11:12, configure a route-policy to match the Community-filter, aggregate the two routes into the route 10.1.10.0/23, and apply this route-policy.
  - On RTC, configure a route-policy, set the Community attribute to no-export, and apply this route-policy in the export direction.
- The Community attribute is classified into well-known and extended community attributes.

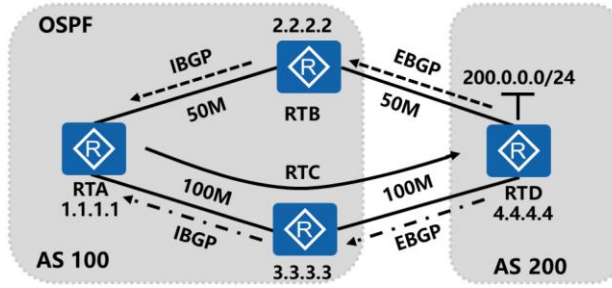
- Well-known community attributes include four types:
- Internet: is the default attribute. All routes belong to the Internet. Routes carrying the Internet attribute can be advertised to all BGP neighbors.
- No_Export: indicates that a route carrying this attribute is not advertised to other ASs. In the figure, if RTB expects that the route 10.1.11.0/24 is not advertised to other ASs after it is advertised to AS 12, RTB can set the Community attribute of this route to No_Export.
- No_Advertise: indicates that a route carrying this attribute is not advertised to other BGP neighbors. In the figure, if RTB wants to advertise the route 10.1.11.0/24 only to RTC, RTB can set the Community attribute of this route to No_Advertise.
- No_Export_Subconfed: is used in the BGP Confederation and not described in this course.
- An extended community attribute is a 4-byte list in the format aa:nn or the community number.
- In aa:nn, aa indicates an AS number and nn indicates the community identifier defined by the administrator.
- The community number ranges from 0 to 4294967295. In RFC1997, the values 0 to 65535 and 4294901760 to 4294967295 are reserved.

Contents

BGP Route Selection Rules

- After a BGP router advertises a route to neighbors, each neighbor selects the optimal route:
  - If this route is the only route to the destination, the neighbor selects it as the optimal route.
  - If there are multiple routes to the same destination, the neighbor selects the route with the highest priority as the optimal route.
  - If there are multiple routes that are destined for the same destination and have the same priority, the neighbor compares attributes of these routes to select the optimal route.
- Generally, a BGP router compares the following attributes in sequence to select the optimal route:
  - Discards routes with unreachable next hops.
  - Prefers the route with the largest PrefVal value. The PrefVal attribute is a Huawei proprietary attribute and is valid only on the device where it is configured.
  - Prefers the route with the highest Local_Pref.
  - Prefers the following routes in descending order of priority: manually aggregated route, automatically aggregated route, route imported using the network command, route imported using the import command, and route learned from peers.
  - Prefers the route with the shortest AS_Path.
  - Prefers the route with the Origin attribute IGP, EGP, and Incomplete in sequence.
  - Prefers the route with the smallest MED if routes are received from the same AS.
  - Prefers an EBGP route to an IBGP route.
  - Prefers the route with the smallest IGP metric within an AS.
  - Prefers the route with the shortest Cluster_List.
  - Prefers the route with the smallest Originator_ID.
  - Prefers the route advertised by the router with the smallest Router_ID.
  - Prefers the route learned from the neighbor with the lowest IP address.

Impact of PrefVal on BGP Route Selection

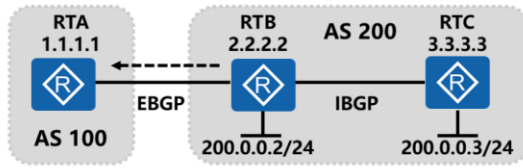


- The PrefVal attribute is a Huawei proprietary attribute and is valid only on the device where it is configured. It corresponds to the weight in BGP rout selection rules. A larger PrefVal value indicates a higher priority.

HUAWEI

- In the figure, AS 200 has a user network segment 200.0.0.0/24. The administrator of AS 100 wants to access this network segment over the high-bandwidth links and expects that the policy configured on RTA affects only its route selection.
- Method to meet this requirement:
  - On RTA, configure ip-prefix to match the route 200.0.0.0/24 received from RTC, configure a route-policy to invoke this ip-prefix, set the PrefVal value of the route to 100, and apply the route-policy in the import direction.
- Verification: Run the tracert command on RTC to check the routers that the route 200.0.0.0/24 passes through.

Impact of Route Aggregation Mode on BGP Route Selection



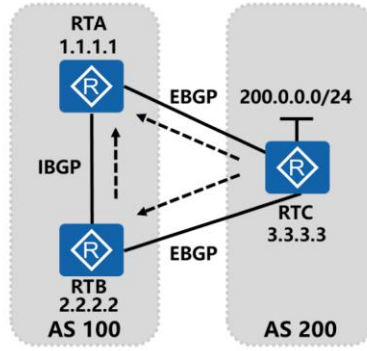

```
<RTB>display bgp routing-table 200.0.0.0
BGP local router ID : 2.2.2.2
 Local AS number : 200
 Paths:   2 available, 1 best, 1 select
 BGP routing table entry information of 200.0.0.0/24:
 Aggregated route.
 .....
 Aggregator: AS 200, Aggregator ID 2.2.2.2, Atomic-aggregate
 Advertised to such 2 peers:
    10.1.12.1
    10.1.23.3
 BGP routing table entry information of 200.0.0.0/24:
 Summary automatic route
 .....
 Aggregator: AS 200, Aggregator ID 2.2.2.2
 Not advertised to any peer yet
```

- A manually aggregated route has a higher priority than an automatically aggregated route.

  HUAWEI

- In AS 200, RTB and RTC have users on the network segment 200.0.0.0/24 and change the route 200.0.0.0/24 into a BGP route using the import command. RTB aggregates the routes and then sends the aggregated route to RTA. Both manual aggregation and automatic aggregation are enabled on RTB. How does RTB prefer the aggregated route?
- Both manual aggregation and automatic aggregation are enabled on RTB. Run the display bgp routing table on RTB. The command output shows that only the manually aggregated route is sent to RTA, but the automatically aggregated route is not. This indicates that a manually aggregated route has a higher priority than an automatically aggregated route.
- Automatic aggregation takes effect only for imported BGP routes, but manual aggregation can also take effect for routes in a BGP routing table. For details, see the BGP route aggregation section. In the preceding scenario, the routes to be aggregated are imported routes. Therefore, both automatic aggregation and manual aggregation can meet this requirement. If the BGP routing table has both imported routes and routes advertised using the network command, only manual aggregation can meet this requirement.

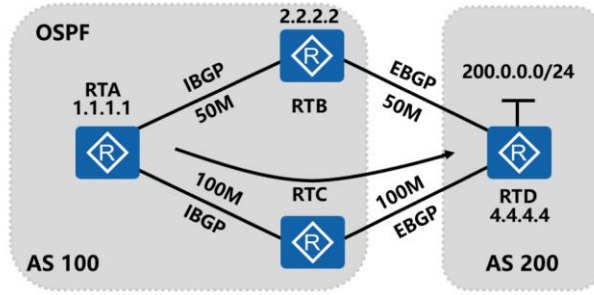
Routes Learned from EBGP Neighbors Are Preferred to Those Learned from IBGP Neighbors



- According to route selection rules, RTA prefers the route learned from an EBGP neighbor.

- In the figure, AS 200 has a network segment 200.0.0.0/24 and advertises it to RTA and RTB through EBGP. RTB also advertises it to RTA through IBGP. Then RTA receives two routes to 200.0.0.0/24. How does RTA select the optimal route?
- According to route selection rules, RTA prefers the route learned from an EBGP neighbor.

Impact of Intra-AS IGP Metric on BGP Route Selection



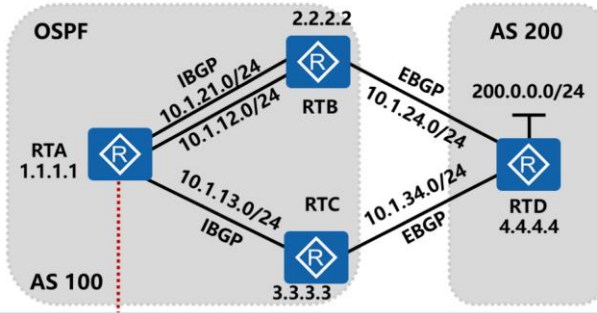

```
<RTA>display bgp routing-table

     Network            NextHop        MED    LocPrf      PrefVal  Path/Ogn
*>i  200.0.0.0          10.1.34.4      0      100         0        200 i
* i                     10.1.24.4      0      100         0        200 i
```

- Adjusting the OSPF cost enables RTA to access the network segment 200.0.0.0/24 over the high-bandwidth links.

HUAWEI

- In the figure, AS 200 has a user network segment 200.0.0.0/24 and advertises it to RTB and RTC through EBGP. RTB and RTC advertise it to RTA through IBGP. The administrator of AS 100 wants to access this network segment over the high-bandwidth links. How to meet this requirement on RTA?
- Set the OSPF cost of the interface that connects RTA to RTB to 100. RTA then accesses this network segment over the link RTA->RTC->RTD.
  - This is because when RTA accesses 200.0.0.0/24, the cost to Next_hop 10.1.34.4 is lower than the cost to Next_hop 10.1.24.4.

Impact of Router ID and IP Address on BGP Route Selection



<RTA>display bgp routing-table

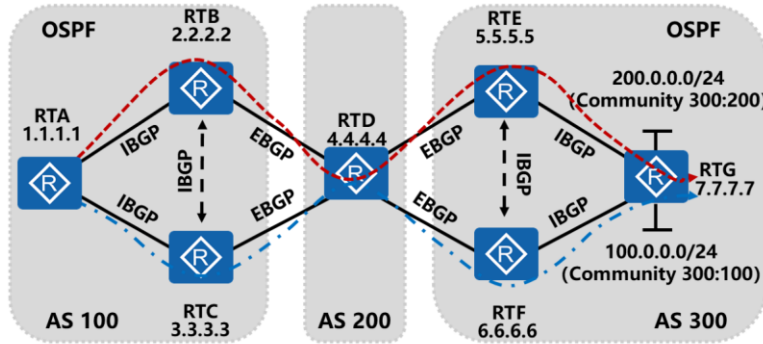
	Network	NextHop	MED	LocPrf	PrefVal	Path/Ogn
*>i	200.0.0.0	10.1.24.4	0	10	0	200i
* i		10.1.34.4	0	100	0	200i

- RTA accesses the network segment 200.0.0.0/24 within AS 200 through RTB, and the outbound interface is the interface at 10.1.12.1.

- In the figure, AS 200 has a user network segment 200.0.0.0/24 and advertises it to RTB and RTC through EBGP. RTB and RTC advertise it to RTA through IBGP. RTA and RTB are connected using two links. How does RTA select the optimal route to access this network segment?
- RTA selects the route with the next hop 10.1.12.2 to access this network segment.
  - RTA selects the link RTA->RTB->RTD to access this network segment because RTB has a smaller router ID than RTC. BGP prefers the route advertised by the router with a smaller router ID.
  - RTA selects the interface with the next-hop address 10.1.12.2 as the outbound interface because BGP prefers the route learned from a neighbor with a lower IP address.

- Run the display bgp routing-table 200.0.0.0 command on RTA. The following command output is displayed:
- <RTA>display bgp routing-table 200.0.0.0
- BGP local router ID : 1.1.1.1
- Local AS number : 100
- Paths: 2 available, 1 best, 1 select
- BGP routing table entry information of 200.0.0.0/24:
- From: 2.2.2.2 (2.2.2.2)
- Route Duration: 00h02m10s
- Relay IP Nexthop: 10.1.12.2
- Relay IP Out-Interface: GigabitEthernet0/0/0
- Original nexthop: 10.1.24.4
- Qos information : 0x0
- AS-path 200, origin igp, MED 0, localpref 100, pref-val 0, valid, internal, pre255, IGP cost 2, not preferred for router ID
- ……

Example for Configuring a BGP Routing Policy



- AS 300 has two user network segments. When users in AS 100 access the two network segments, traffic needs to be load balanced between RTB and RTC. When users in AS 200 access the two network segments, traffic needs to be load balanced between RTE and RTF.

  HUAWEI

- In the figure, AS 300 has two user network segments 200.0.0.0/24 and 100.0.0.0/24. To differentiate users on the two network segments, the Community attribute 300:100 is configured for the route 100.0.0.0/24, and the Community attribute 300:200 is configured for the route 200.0.0.0/24. When users in AS 100 access the two network segments, traffic needs to be load balanced between RTB and RTC. When users in AS 200 access the two network segments, traffic needs to be load balanced between RTE and RTF.
- Assume that RTA accesses 100.0.0.0/24 over the link RTA->RTB->RTD->RTE->RTG and accesses 200.0.0.0/24 over the link RTA->RTC->RTD->RTF->RTG. Perform the following configurations to meet this requirement:
  - RTE and RTF advertise routes carrying the Community attributes to RTD.
  - After RTD receives the routes carrying the Community attributes, it uses two Community-filters to match different Community attributes, use two route-policies to invoke the two Community-filters, and set the next hops of the routes with the Community attributes 300:100 and 300:200 as the outbound interface addresses of RTE and RTF respectively.
  - On RTD, configure another two route-policies. That is, configure one route-policy to set the MED value of the route with the Community attribute 300:100 to 100 and apply this route-policy in the export direction. Configure the other route-policy to set the MED value of the route with the Community attribute 300:200 to 100 and apply this route-policy in the export direction.

Configuration on RTD:

```
bgp 200
 peer 10.1.24.2 as-number 100
 peer 10.1.34.3 as-number 100
 peer 10.1.45.5 as-number 300
 peer 10.1.46.6 as-number 300
 #
 ipv4-family unicast
  undo synchronization
  peer 10.1.24.2 enable
  peer 10.1.24.2 route-policy MED-20 export
  peer 10.1.24.2 advertise-community
  peer 10.1.34.3 enable
  peer 10.1.34.3 route-policy MED-10 export
  peer 10.1.34.3 advertise-community
  peer 10.1.45.5 enable
  peer 10.1.45.5 route-policy 10 import
  peer 10.1.46.6 enable
  peer 10.1.46.6 route-policy 10 import
#
route-policy 10 permit node 10
 if-match community-filter 10
 apply ip-address next-hop 10.1.45.5
#
route-policy 10 permit node 20
 if-match community-filter 20
 apply ip-address next-hop 10.1.46.6
#
```

```
route-policy MED-10 permit node 10
 if-match community-filter 300:100
 apply cost 100
#
route-policy MED-20 permit node 10
 if-match community-filter 20
 apply cost 100
#
ip community-filter 10 permit 300:100
ip community-filter 20 permit 300:200
```

Verification: Run the following commands on RTA:

tracert 100.0.0.1: Check the IP addresses that the route passes through.

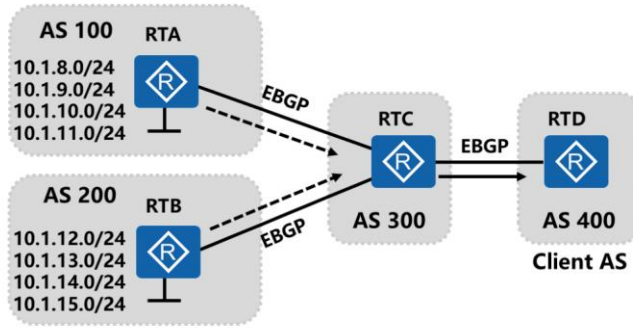
tracert 200.0.0.1: Check the IP addresses that the route passes through.

Contents

BGP Route Aggregation Overview

- BGP transmits routing information between ASs. As the number of ASs increases and the scale of a single AS increases, the BGP routing table becomes very large, resulting in the following two problems:
  - Storing the routing table will occupy a lot of memory resources, and transmitting as well as processing routing information need to consume a lot of bandwidth resources.
  - Frequently updating and withdrawing routes will affect network stability.
- BGP route aggregation is designed to solve the two problems. The following describes BGP route aggregation, including:
  - BGP route aggregation necessity: addresses BGP network problems
  - BGP route aggregation configuration
  - Problem brought by BGP route aggregation.

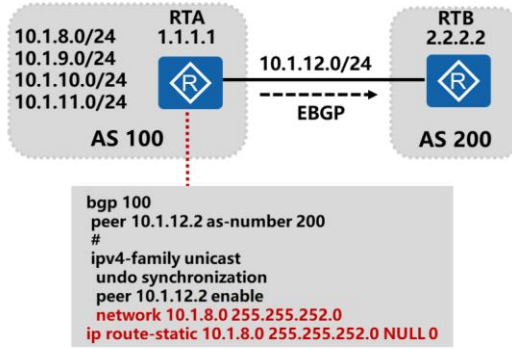
BGP Route Aggregation Necessity



- AS 100 and AS 200 each have four user network segments. AS 300 is connected to AS 400, a client AS. AS 400 has a low-end router, RTD, that has low processing capabilities. Therefore, it is required that RTD can access the network segments in AS 100 and AS 200 but do not receive many specific routes. How to meet this requirement?

- Method to meet this requirement:
  - On RTC, aggregate the specific routes in AS 100 and AS 200 into an aggregated route 10.1.8.0/21 and advertise this aggregated route to the client AS.
- Currently, there are a large number of routes on the Internet, bringing in the following problems:
  - Storing the routing table will occupy a lot of memory resources, and transmitting routing information needs to consume a lot of bandwidth resources.
  - Frequently flapping of specific routes will make the network unstable.
- Therefore, it is inevitable to use route aggregation to save memory and bandwidth resources and reduce the impact of route flapping.

BGP Route Aggregation - Static Routes



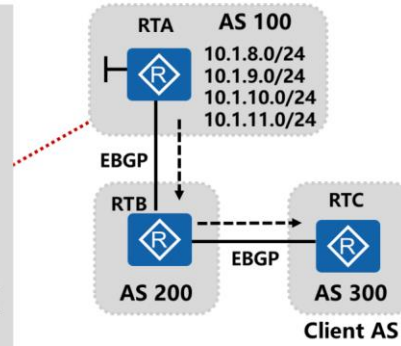
- AS 100 has four user network segments. RTA in AS 100 shields specific routes using route aggregation and advertises only the aggregated route 10.1.8.0/22 to RTB in AS 200.

  HUAWEI

- Use static routes to configure route aggregation:
  - Use a static route to aggregate specific routes into the route 10.1.8.0/22 with the next hop pointing to NULL 0. Because the aggregated route is not identified by a specific address and is only the replacement of specific routes when being advertised to AS 200, the next hop of the aggregated route points to Null 0 to prevent routing loops.
  - Because a static route is used, the route 10.1.8.0/22 with the next hop Null 0 is generated in the routing table. To aggregate specific routes, use the network command to change the route 10.1.8.0/22 in the IP routing table into a BGP route and advertise this route to the BGP neighbor.

BGP Route Aggregation - Automatic Aggregation

```
bgp 100
 peer 10.1.12.2 as-number 200
 #
 ipv4-family unicast
  undo synchronization
  summary automatic
  import-route direct route-policy r1
  peer 10.1.12.2 enable
#
route-policy r1 permit node 10
 if-match ip-prefix r1
#
ip ip-prefix r1 index 10 permit 10.1.11.0 24
ip ip-prefix r1 index 20 permit 10.1.10.0 24
ip ip-prefix r1 index 30 permit 10.1.9.0 24
ip ip-prefix r1 index 40 permit 10.1.8.0 24
```



- AS 100 has four user network segments, which are changed into BGP routes using the import command. AS 200 is connected to AS 300, a client AS. RTC in AS 300 has low processing capabilities. Therefore, it is required that RTC can access the network segments in AS 100 and AS 200 but do not receive many routes. How to meet this requirement?
- On RTB and RTC, run the display bgp routing-table command. The following command output is displayed:
- <RTB>display bgp routing-table

```
     Network         NextHop        MED        LocPrf    PrefVal  Path/Ogn
*>   10.0.0.0        10.1.12.1                             0        100?
```

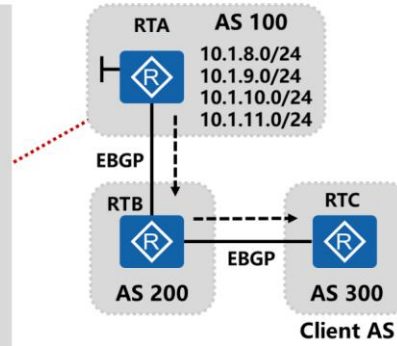
<RTC>display bgp routing-table

```
     Network         NextHop        MED        LocPrf    PrefVal  Path/Ogn
*>   10.0.0.0        10.1.23.2                             0        200 100?
```

- Automatic aggregation can aggregate only imported BGP routes. BGP aggregates routes according to the natural network segment and sends only the aggregated route to neighbors.

BGP Route Aggregation - Manual Aggregation

```
bgp 100
 peer 10.1.12.2 as-number 200
 #
 ipv4-family unicast
  undo synchronization
  aggregate 10.1.8.0 255.255.252.0
detail-suppressed
  network 10.1.8.0 255.255.255.0
  network 10.1.9.0 255.255.255.0
  import-route direct route-policy r1
  peer 10.1.12.2 enable
#
route-policy r1 permit node 10
 if-match ip-prefix r1
#
ip ip-prefix r1 index 10 permit 10.1.11.0 24
ip ip-prefix r1 index 20 permit 10.1.10.0 24
```



- AS 100 has four user network segments, which include BGP routes imported using the import command and the network command. AS 200 is connected to AS 300, a client AS. RTC in AS 300 has low processing capabilities. Therefore, it is required that RTC can access the network segments in AS 100 and AS 200 but do not receive many routes. How to meet this requirement?
- On RTB and RTC, run the display bgp routing-table command. The following command output is displayed:

<RTB>display bgp routing-table

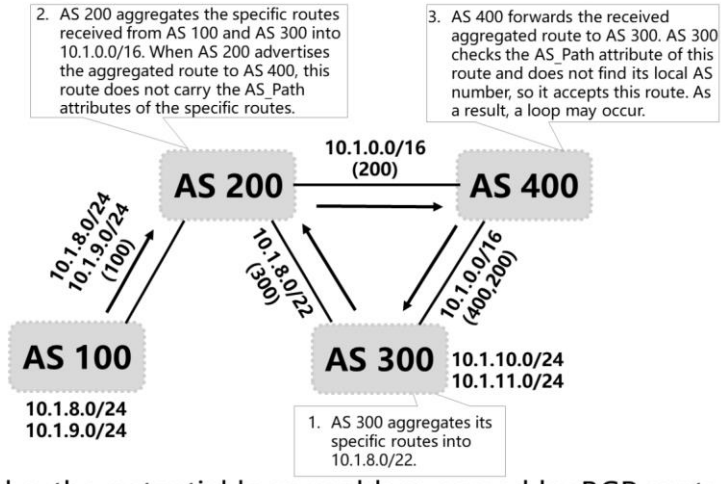
	Network	NextHop	MED	LocPrf	PrefVal	Path/Ogn
*>	10.1.8.0/22	10.1.12.1			0	100?

<RTC>display bgp routing-table

	Network	NextHop	MED	LocPrf	PrefVal	Path/Ogn
*>	10.1.8.0/22	10.1.23.2			0	200 100?

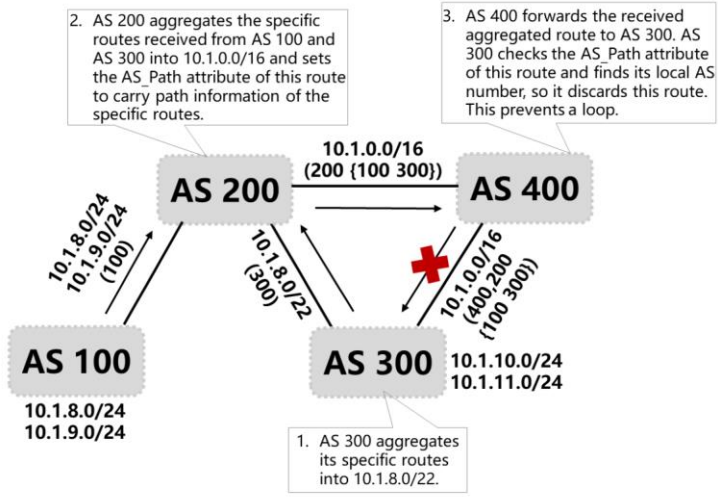
- Manual aggregation can aggregate routes in the local BGP routing table and specify the mask of the aggregated route.

Problem Brought By BGP Route Aggregation - Loops



- How to solve the potential loop problem caused by BGP route aggregation?

Problem Brought By BGP Route Aggregation - Solution



  HUAWEI

- To solve the loop problem caused by BGP route aggregation, two AS_Path attributes are designed:
  - Atomic-Aggregate: is a well-known discretionary attribute. It is used to inform that information is lost on downstream routers. In the figure, path information is lost after routes are aggregated on the router where route aggregation is configured. Then this router sends an Update message carrying the Atomic-Aggregate attribute to inform its neighbors that path information is lost.
  - Aggregator: is an optional transitive attribute. It contains the AS number and router ID of the router that initiates route aggregation.
- The AS_Path attribute has two types:
  - AS_Sequence: records sequenced ASs that packets pass through.
  - AS_Set: records unsequenced ASs that packets pass through.
- The AS_Path attribute is a sequenced list because each AS number is added to the AS_Path list each time the AS_Path attribute passes through an AS and the first AS number is added to the leftmost of the AS_Path list.
  - In the figure, when AS 400 advertises the aggregated route to AS 300, the AS_Path attribute (except that enclosed in braces) of the route indicates that this route passes through AS 200 and then AS 400.

- If the aggregated route needs to carry the AS numbers that all specific routes pass through to prevent loops, you need to specify the as-set parameter following the route aggregation command.
- In the figure, specific routes are aggregated and the as-set parameter is specified in AS 200. The aggregated route carries an AS-Set to indicate AS_Path information of the specific routes. AS numbers in the list are not recorded in the sequence in which the route passes through ASs. In this manner, loops are prevented.
- Discussion
- Route aggregation solves two problems. That is, it reduces resources required for route transmission and calculation and reduces route flapping on network stability because it hides specific routing information. However, after route aggregation is used, the AS_Path attribute is lost, creating a risk of routing loop.
- If the aggregated route carries information about the ASs that all specific routes pass through, this route may also be frequently updated when the specific routes frequently flap.
- Therefore, whether the aggregated route carries lost AS_Path information requires network designers to determine based on the network environment.

Quiz

1. Which of the following attributes are BGP well-known mandatory attributes?
   A. Origin
   B. AS_Path
   C. Next_hop
   D. Local_preference
2. Which of the following port numbers is used by BGP?
   A. TCP 21
   B. TCP 179
   C. TCP 80

- Answer: ABC.
- Answer: B.


Thank You
www.huawei.com

IP Multicast Basics

Foreword

- If unicast technology is used for point-to-multipoint communication, the amount of data transmitted on the network will be proportional to the number of users that require the data. Sending multiple copies of identical data to different users wastes resources on the data source and network bandwidth. If broadcast technology is used, hosts that do not require the data will also receive the data. This threatens information security and causes storms on the local network segment.
- IP multicast technology resolves the preceding problems. The multicast source sends only one copy of data, which is then replicated and forwarded by network nodes, and finally sent to specified receivers.

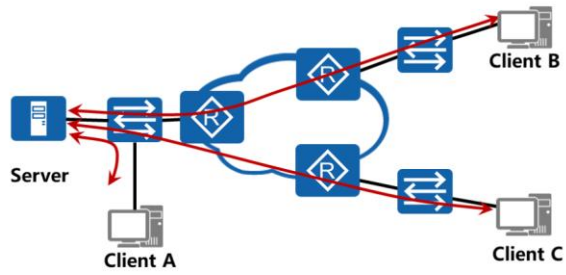
Objectives

- Upon completion of this section, you will be able to:
  - Be familiar with characteristics of point-to-multipoint applications
  - Master the basic multicast architecture
  - Understand the structure of multicast addresses

Contents

Traditional Point-to-Point Applications

- A service provider delivers services on a per user basis.
- The service provider transmits different data to different users.



- Traditional point-to-point applications, such as email, web, and online banking, provide specific services for individuals or organizations. Because of differences in data required by users and security requirements, data can only be transmitted in a point-to-point way between clients and servers. That is, communication is established between one source host and one destination host, and a data flow has only one sender and one receiver.
- The communication process between the two parties is as follows:
  - The server encapsulates data packets and sends them out. In the data packets, the source IP address is the server's IP address, the destination IP address is the client's IP address, the source MAC address is the source's MAC address, and the destination MAC address is the MAC address of the gateway router.
  - When the gateway router receives the data packets, it decapsulates them and looks up the destination IP address in its routing table to determine the next-hop address and outbound interface toward the destination IP address. The gateway router then encapsulates the data packets and forwards them to the next-hop device through the outbound interface.
  - After forwarded hop by hop on the network, the packets arrive at the client. The client decapsulates the packets and sends them to the upper-layer application protocol for processing.

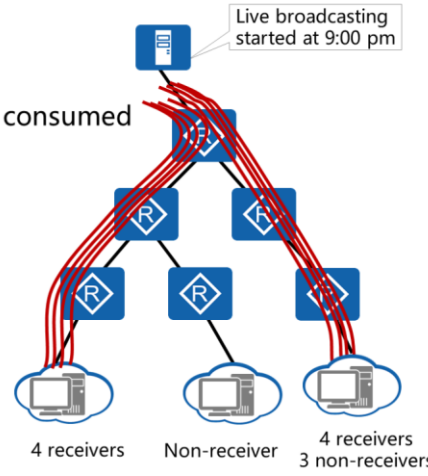
New Point-to-Multipoint Applications

- A service provider delivers services based on groups of users.
- The service provider transmits identical data to all users in the same group.

- With development of the Internet, the data, voice, and video streams transmitted on networks have increased sharply.
- New applications, such as online live broadcasting, web TV, and video conferencing, are also growing.
- Most of these new applications use the point-to-multipoint transmission model and require high information security, wide transmission scope, and high bandwidth.

Point-to-Multipoint Application in Unicast Mode

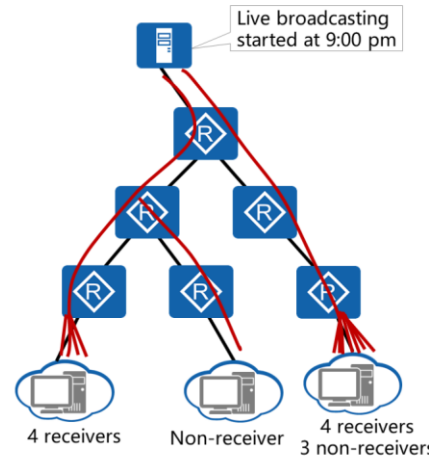
- Problems of the unicast mode:
  - Too much duplicate traffic
  - Many device and link bandwidth resources consumed
  - Unable to ensure transmission quality



- Unicast transmission is implemented between a source IP host and a destination IP host. Most of data is transmitted in unicast mode on a network. For example, email and online banking applications are implemented in unicast mode.
- Characteristics of unicast transmission:
  - A unicast packet uses a unicast address as the destination address. The source sends an independent copy of unicast packets to each receiver. If there are N receivers on the network, the source needs to send N copies of unicast packets.
  - The network provides an independent data transmission path to forward each copy of unicast packets separately. Therefore, N independent transmission paths need to be established for the N copies of unicast packets.
- Disadvantages of unicast transmission:
  - In unicast transmission mode, the amount of data transmitted on a network is proportional to the number of users that require the data. When many users require the same data, many duplicate data flows will be transmitted on the network, which cause high CPU usage on network devices and waste network bandwidth.
  - The unicast mode is suitable for networks with a few users and cannot ensure data transmission quality when there are many users on a network.

Point-to-Multipoint Application in Broadcast Mode

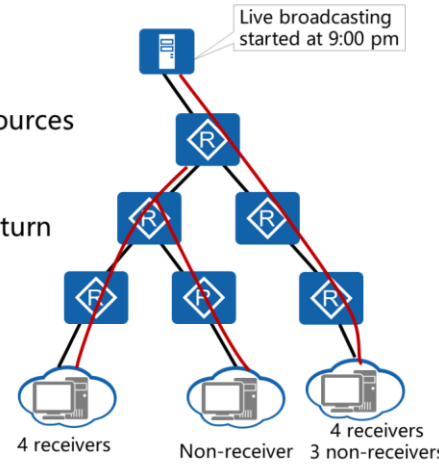
- Problems of the broadcast mode:
  - Limited application scope
  - Unable to ensure security
  - Unable to charge for services



- Broadcast transmission is implemented between a source IP host and all the other IP hosts on the local network. All hosts receive data from the source host, regardless of whether they require the data.
- Characteristics of broadcast transmission:
  - A broadcast packet uses a broadcast address as the destination address. The source sends only one copy of each packet to the broadcast address on the local network segment.
  - All hosts on the network segment will receive the packets although some of them may not need the packets.
- Disadvantages of broadcast transmission:
  - In broadcast transmission mode, a data sender and user hosts must be located on a shared network segment, and all hosts on the shared network segment receive the data from the sender.
  - The broadcast transmission mode applies only to a shared network segment, and cannot ensure information security and cannot implement paid service for specific users.
- For point-to-multipoint applications, the unicast and broadcast modes both have their limitations.

Point-to-Multipoint Application in Multicast Mode

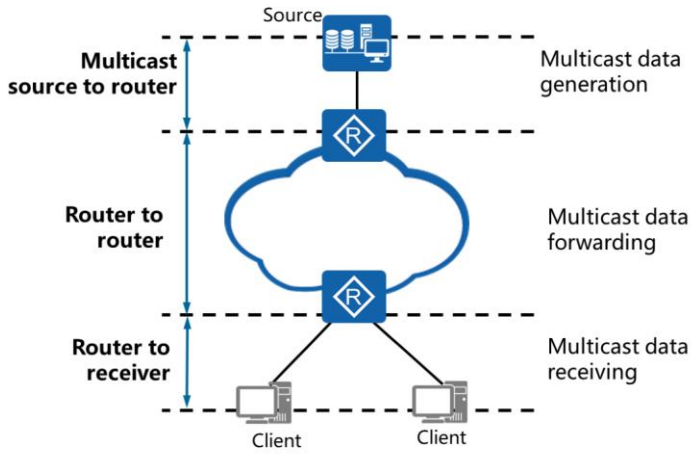
- Advantages of the multicast mode:
  - No duplicate traffic
  - Conservation of device and bandwidth resources
  - High security
  - Guarantee of service providers' business return



- Multicast transmission is implemented between one source IP host and a group of IP hosts. Intermediate routers and switches selectively replicate and forward data based on demands of receivers.
- Advantages of multicast transmission
  - In multicast transmission mode, a data flow is transmitted to a group of users along a multicast distribution tree. Each link transmits only one copy of multicast data packets.
  - Compared with the unicast mode, the multicast mode starts to copy data and distribute data copies on the network node as far from the source as possible. Therefore, the amount of data and network resource consumption will not increase greatly when the number of receivers increases.
  - Compared with the broadcast mode, the multicast mode transmits data only to receivers that require the data. This saves network resources and enhances data transmission security. In addition, broadcast packets are only transmitted within a shared network segment, whereas multicast packets can travel across network segments.
- Multicast application
  - Multicast technology efficiently implements point-to-multipoint data transmission over an IP network, while conserving network bandwidth and reducing network loads. Multicast technology can easily provide online live broadcasting, web TV, distance education, and other point-to-multipoint services.

Contents

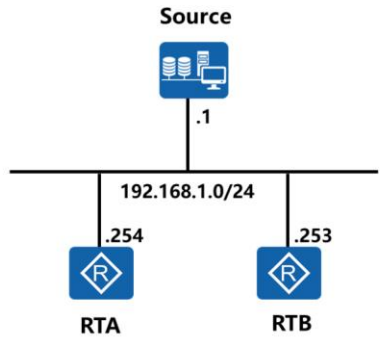
Basic Multicast Architecture



- Multicast source to router: The multicast source generates multicast data, encapsulates the data, and sends data packets to the gateway router.
- Router to router: A router selectively replicates and forwards data based on distribution of receivers.
- Router to receiver: The last-hop router receives multicast data packets and sends them to receivers.

Multicast Source to Router

- How does a multicast source encapsulate multicast data?
  - How to determine the destination IP address
  - How to determine the destination MAC address



Multicast data
DIP: ?
SIP: 192.168.1.1
SMAC: 68-F7-28-42-6D-0D
DMAC: ?

- Unicast packets are transmitted hop-by-hop over an IP network.
- IP multicast communication differs from IP unicast communication in that the destination IP address of data packets is a group address, but not an IP address of a specific host.
- To enable data from a source to reach the group members across the Internet, IP multicast needs to be implemented. That is, destination IP addresses of multicast data packets are multicast IP addresses. The multicast source does not care about locations of the receivers and only needs to send data to a specific group address.
- On an Ethernet network, the destination MAC address of a unicast data frame is the MAC address of the receiver or the next-hop gateway device.
- The destination of a multicast packet is not a specific receiver but a group with unspecific members. If the destination MAC address is set to a receiver’ s MAC address, the source must send a copy of each multicast frame to each receiver.
- This is unacceptable. To implement efficient multicast data transmission at the data link layer, the network must provide link-layer multicast forwarding capability. Link-layer multicast transmission is implemented based on multicast MAC addresses.

Multicast IP Address

- A multicast IP address represents a group of hosts, but not a specific host. If a host joins a multicast group, it means that the host wants to receive data packets destined for the multicast IP address.

Range	Description
224.0.0.0—224.0.0.255	Permanent group addresses reserved for routing protocols
224.0.1.0—231.255.255.255 233.0.0.0—238.255.255.255	Any-source temporary group addresses
232.0.0.0—232.255.255.255	Source-specific temporary group addresses
239.0.0.0—239.255.255.255	Any-source group addresses for use in private multicast domains

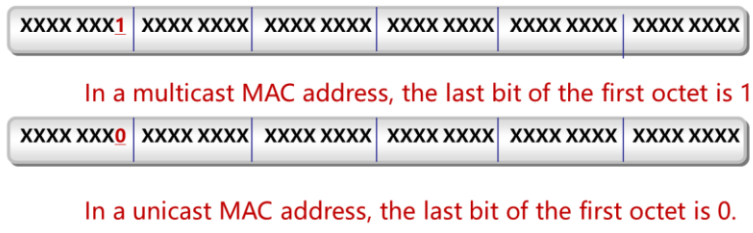
- Common IP multicast models include the any-source multicast (ASM) and source-specific multicast (SSM) models.

- IPv4 multicast addresses
  - The IPv4 address space is divided into five classes, Class A to Class E. Class D addresses are IPv4 multicast addresses, ranging from 224.0.0.0 to 239.255.255.255. These addresses identify multicast groups and can only be used as destination addresses of multicast packets but not source addresses.
  - Source addresses of IPv4 multicast packets are IPv4 unicast addresses, which can be Class A, Class B or Class C addresses and cannot be Class D or Class E addresses.
  - On the network layer, all hosts that have joined the same multicast group can identify the same IPv4 multicast group address. Once a user joins the multicast group, the user can receive IP multicast packets with the group address as the destination address.
- Multicast service models
  - ASM stands for any-source multicast. In the ASM model, any sender can be a multicast source to send data to a multicast group address. Receiver hosts can receive all data sent to this group after they join the group. In the ASM model, receivers cannot obtain the location of a multicast source in advance and can join or leave a multicast group anytime.

- SSM stands for source-specific multicast. In real-world applications, users may be interested in data sent from specific sources and do not want to receive data from other sources. The SSM model provides a data transmission service that allows users to specify sources. The essential feature that distinguishes the SSM model from the ASM model is that the receivers know the locations of multicast sources in advance. The SSM and ASM models use different ranges of addresses to set up multicast distribution trees from multicast sources to receivers.

Multicast MAC Address

- Difference between multicast and unicast MAC addresses:

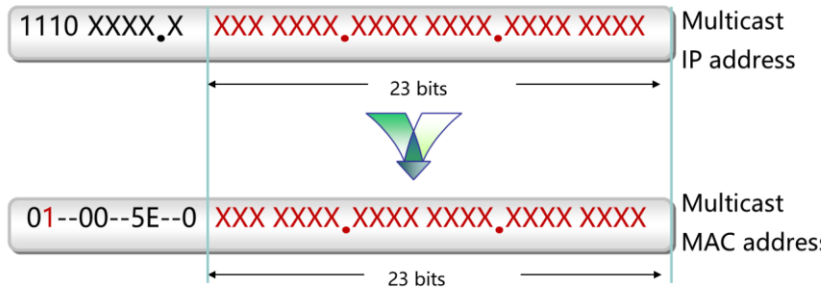


- As defined by IANA, the most significant 24 bits of IPv4 multicast MAC addresses are 0x01005e, and the 25th bit is always 0.

- The destination of a multicast data frame is not a specific receiver but a group with unspecific members. Therefore, the destination is identified by a multicast MAC address. As defined by IANA, the most significant 24 bits of multicast MAC addresses are 0x01005e, and the 25th bit is always 0.
- A multicast MAC address identifies receivers in the same multicast group at the link layer.
- On an Ethernet network, the destination MAC address of a unicast data frame is the MAC address of the receiver or the next-hop router. The MAC address is learned by using ARP. Multicast data frames also require a MAC address that can be obtained in advance.

Mapping Between Multicast IP and MAC Addresses

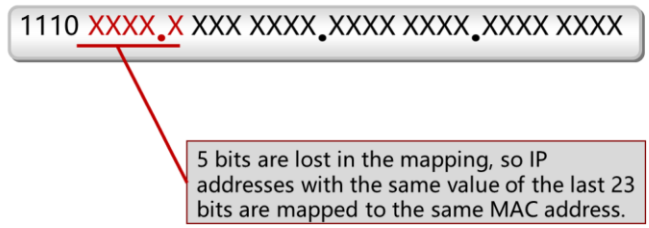
- Multicast IP and MAC addresses need to be mapped to each other automatically.
- The least significant 23 bits of a MAC address are the least significant 23 bits of a multicast IP address.



- To enable multicast sources and group members to communicate, the network must provide network-layer multicast service, which uses multicast IP addresses. To enable multicast data to be correctly transmitted on the local physical network, the network must provide link-layer multicast service, which uses multicast MAC addresses. The destination address of a multicast data packet is a group with unspecific members but not a specific receiver. Therefore, multicast IP addresses must be mapped to multicast MAC addresses.

Problem Caused by Address Mapping

- In multicast IP-to-MAC address mapping, 32 multicast IP addresses are mapped to one multicast MAC address.



- The first 4 bits of a multicast IP address are 1110, indicating a multicast address. Among the last 28 bits, only 23 bits are mapped to multicast MAC addresses. This means that 5 bits are lost in the mapping. As a result, 32 multicast IP addresses are mapped to the same MAC address.
- IETF believes that this will not cause much impact because there is a very low probability that two or more group addresses in the same LAN will be mapped to the same MAC address.

Quiz

1. What is IPv4 multicast communication?
2. What is the range of IPv4 multicast addresses?

- Answer: IP multicast communication transmits packets from a source to a group of receivers. Compared with unicast and broadcast transmission, IP multicast transmission conserves network bandwidth and reduces loads on networks. IP multicast is widely used in IPTV, real-time data transmission, and multimedia conferencing services.
- Answer: The Internet Assigned Numbers Authority (IANA) assigns Class D addresses to IPv4 multicast. An IPv4 address is 32 bits long, and the first 4 bits of a Class D IP address are 1110. Therefore, multicast IP addresses range from 224.0.0.0 to 239.255.255.255.


Thank You
www.huawei.com

IGMP Principles and Configurations

Foreword

- In multicast communication, a source sends multicast data to a specific multicast address. To forward multicast data packets to receivers, the multicast router connected to the network segment of receiver hosts must know which receiver hosts are present on the network segment and ensure that the hosts have joined the specific group.
- The Internet Group Management Protocol (IGMP) is a protocol in the TCP/IP protocol suite and is used to create and maintain group memberships between receiver hosts and neighboring multicast routers.

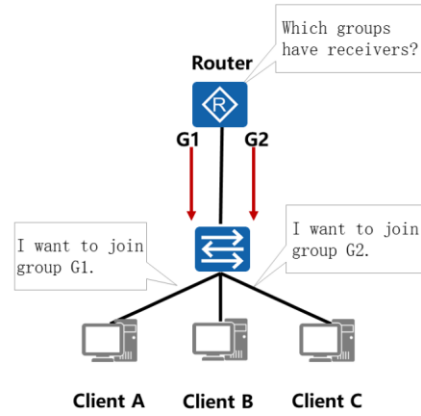
Objectives

- On completion of this section, you will be able to:
  - Master IGMP working mechanisms and configurations
  - Understand the differences between different IGMP versions
  - Understand the mechanism of IGMP snooping

Contents

How Hosts Receive Multicast Data

- What information needs to be exchanged between receivers and the upstream router?
  - Receivers must specify the groups they want to join.
  - The router needs to know which groups have receivers.
- What are the problems if the information is manually configured?
  - Incapable of real-time updates
  - Low flexibility
  - Huge workload and high probability of errors



- A multicast source does not care about receiver locations, but the router connected to group members needs to collect and maintain group membership information.
- Multicast technology neither explicitly specifies receivers nor sends data to all hosts on a network. If a host wants to receive data sent to a group address, it must join the group and become a member of the group.
- How should a network be deployed to implement efficient forwarding and allow receiver hosts to join multicast groups flexibly?

Contents

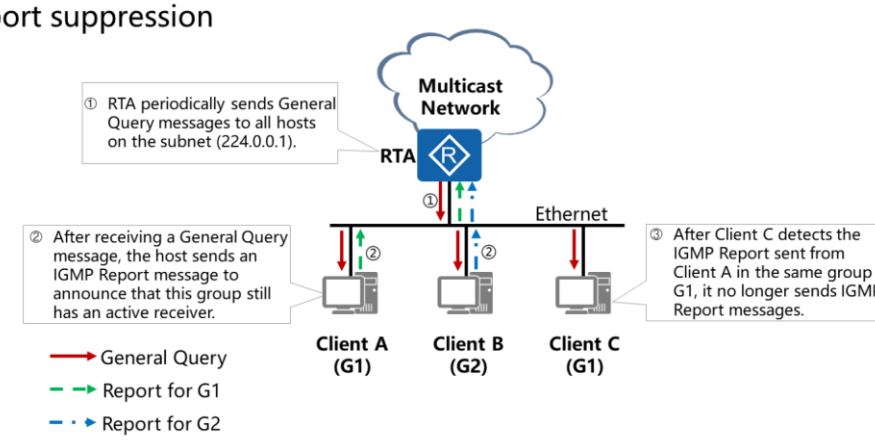
Group Membership Management - IGMP

- The IGMP protocol runs between hosts and multicast routers.
- The IGMP protocol provides the following functions:
  - Hosts: use IGMP to report group membership to directly connected routers.
  - Routers: use IGMP to maintain group membership.

- The Internet Group Management Protocol (IGMP) is a protocol in the TCP/IP protocol suite and used to establish and maintain group memberships between IP hosts and neighboring multicast routers.

IGMPv1 Working Mechanism

- General query and response
- Report suppression



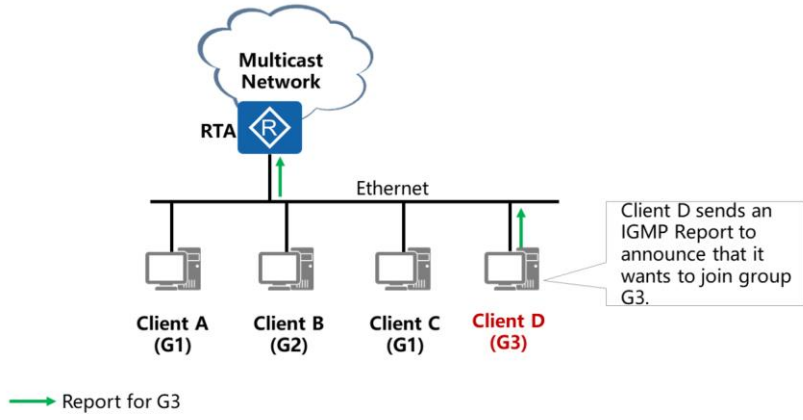
HUAWEI

  - IGMPv1 defines two messages types:
    - General Query: A router periodically sends this type of message to 224.0.0.1 (represents all hosts and routers on the same network segment). The query interval defaults to 60 seconds and is configurable.
    - Membership Report: Hosts send this type of message to announce that a multicast group has active receivers.
- In this figure, Client A and Client C want to receive data sent to group G1, whereas Client B wants to receive data sent to group G2. The general query and response process is as follows:
  - RTA sends a General Query message.
  - All hosts on the network segment receive the General Query message. Client A and Client C are members of G1 and start Timer-G1. Client B is a member of G2 and starts Timer-G2. The timer length is a random value between 0 and 10, in seconds. The host with the timer expiring first sends a Report message for the multicast group. In this example, the Timer-G1 on Client A expires first, so Client A sends a Report message with the destination address G1. When Timer-G2 on Client B expires, Client B sends a Report message with the destination address G2 to the network segment.

- When Client C detects the Report message sent by Client A, it stops Timer-G1 and no longer sends any Report message for G1. This report suppression mechanism reduces the number of protocol packets on the network segment.
- After RTA receives Report messages, it knows that members of G1 and G2 exist on the network segment. When RTA receives multicast data sent to G1 and G2, it forwards the data to this network segment.

IGMPv1 Host Joins a Group

- A host requests to join a group.

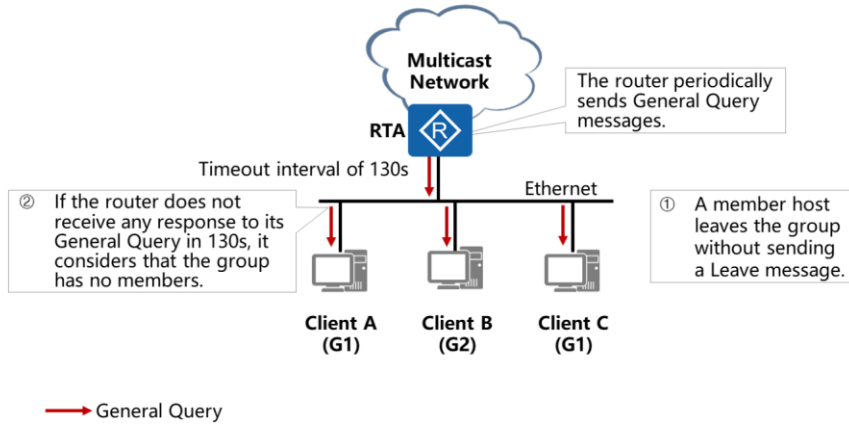


HUAWEI

- A new host Client D wants to join group G3. To receive multicast data quickly, Client D sends a Report message for G3 immediately, without waiting for a General Query message. After receiving the Report message, RTA knows that a member of G3 appears on the network segment. When RTA receives multicast data sent to G3, it forwards the data to this network segment.

IGMPv1 Problem 1: Leave Mechanism

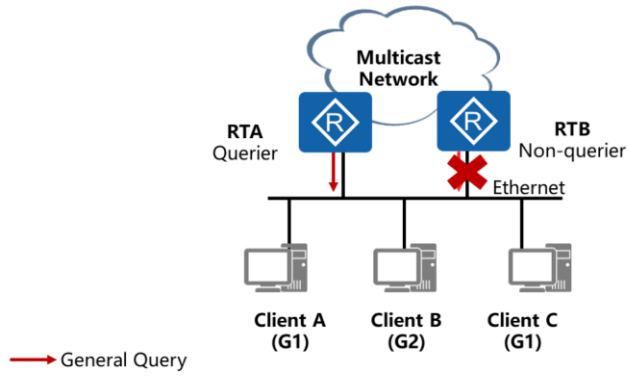
- Leave without sending Leave messages



- IGMPv1 does not define the Leave message.
- After a client leaves a multicast group, it no longer responds to General Query messages sent to the group. If all clients leave the group, RTA will not receive any Report messages because there are no members of the group on the network segment. Therefore, RTA deletes the matching multicast forwarding entry after a specified period of time (Group membership timeout interval = IGMP general query interval x Robustness variable + Maximum response time = 60 x 2 +10 = 130s).

IGMPv1 Problem 2: Querier Election

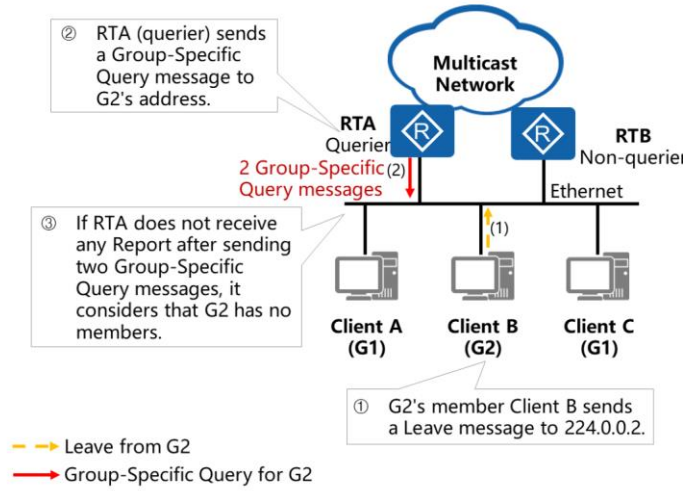
- The querier election depends on a multicast routing protocol.



- If multiple routers are connected to the same receiver network segment, only one router needs to send IGMP queries.
- IGMPv1 does not define a querier election mechanism and depends on multicast routing protocols to select a querier on the network segment.
- Because different routing protocols use different election mechanisms, multiple queriers may be selected on the same network segment in IGMPv1.
- IGMPv2 makes improvement and optimization for these two issues in IGMPv1.

Contents

IGMPv2's Improvement to IGMPv1: Leave Mechanism

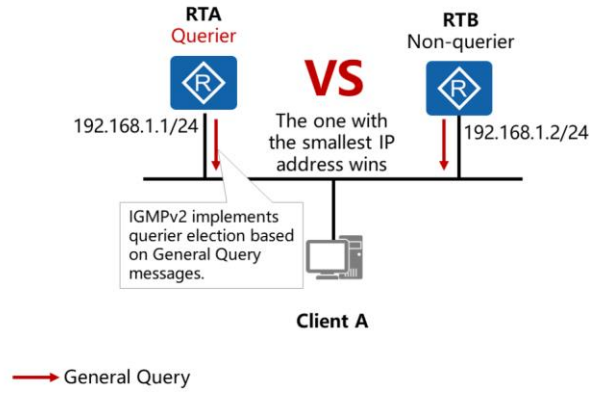


HUAWEI

- This figure shows the leave mechanism defined in IGMPv2. Client B leaves group G2 through the following process:
  - Client B sends a Leave message for G2 to all multicast routers on the local network segment (with destination address 224.0.0.2).
  - When the querier receives the Leave message, it sends a Group-Specific Query message for G2 and starts the group membership timer. (Timer-Membership = Query interval x number of transmissions). By default, the querier sends Group-Specific Query messages twice, at an interval of 1 second. The query interval and number of query message transmissions are configurable.
  - If G2 has no other member on the network segment, the router will not receive any Report message for G2. When the Timer-Membership expires, the router deletes the downstream interface connected to Client B from the multicast forwarding entry. The router no longer sends multicast data for G2 to the network segment. If G2 has other members on the network segment, the members will send a Report message for G2 within the maximum response time after receiving a Group-Specific Query message. The router then continues sending multicast data for G2 to the network segment.

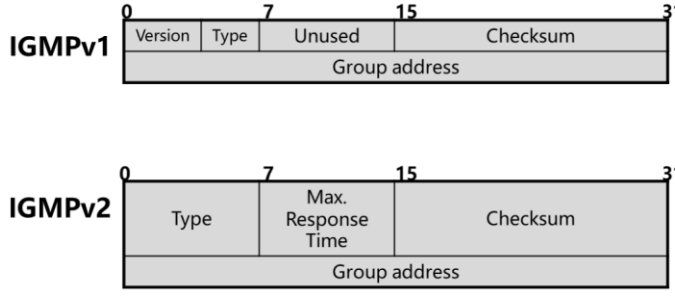
IGMPv2's Improvement to IGMPv1: Querier Election

- Independent querier election



- Compared with IGMPv1, IGMPv2 defines an independent querier election mechanism.
- Each IGMPv2 router considers itself as a querier when it starts, and sends a General Query message to all hosts and routers on the local network segment. When other routers receive the General Query message, they compare the source IP address of the message with their own interface IP addresses. The router with the smallest IP address becomes the querier, and the other routers are non-queriers. As shown in the figure, RTA has a smaller interface IP address than RTB and therefore becomes the querier. The IGMP querier and non-querier routers all process IGMP Report messages, but only the querier sends Query messages. Non-queriers do not process IGMPv2 Leave messages.
- All non-queriers start a timer. If they receive a Query message from the querier before the timer expires, they reset the timer. Otherwise, they consider that the previous querier is ineffective and trigger the election of a new querier.

Comparison Between IGMPv1 and IGMPv2 Packets



- Question: How does IGMP enable group members to receive data from specific multicast sources?

  HUAWEI

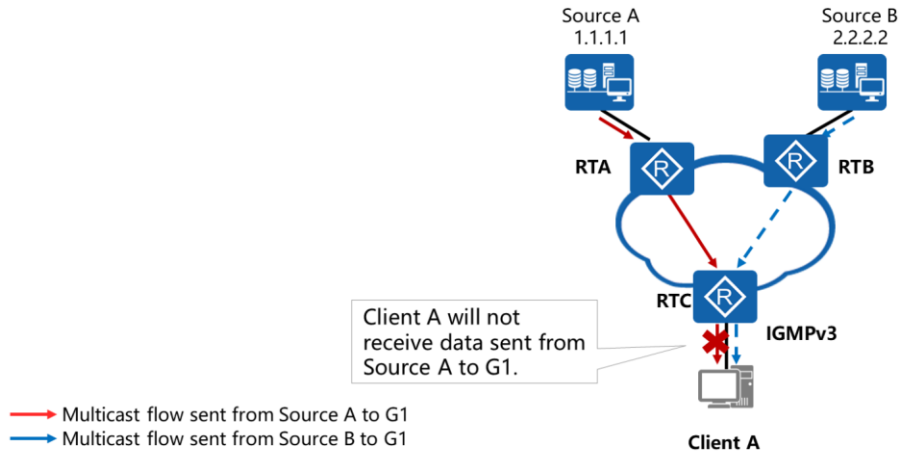
- IGMPv1 messages:
  - Version: IGMP version identifier, which is set to 1 in IGMPv1 messages.
  - Type: 0x11 identifies a General Query message and 0x12 identifies a Membership Report message.
  - Group address: In a General Query message, the group address is 0. In a Membership Report message, the group address is the address of the group that the sender wants to join.
- IGMPv2 messages: IGMPv2 messages differ from IGMPv1 messages in that they do not have a version field but have a maximum response time field.
  - Type: IGMPv2 defines two new message types in addition to those in IGMPv1:
    - Group-Specific Query (0x11): sent by a querier to a specific group on the local network segment to check whether the group has members.
    - Leave (0X17): sent by a host to notify routers on the local network segment that it has left a group.

- Maximum response time: specifies the maximum time the router will wait for the response to a Query message.
- For a General Query message, the default maximum response time is 10 seconds.
- For a Group-specific Query message, the default maximum response time is 1 second.
- Group address:
- In a General Query message, the group address is 0.
- In a Group-Specific Query message, the group address is the address of the group for which the router needs to query.
- In a Report or Leave message, the group address is the address of the group that a host has joined or left.

Contents

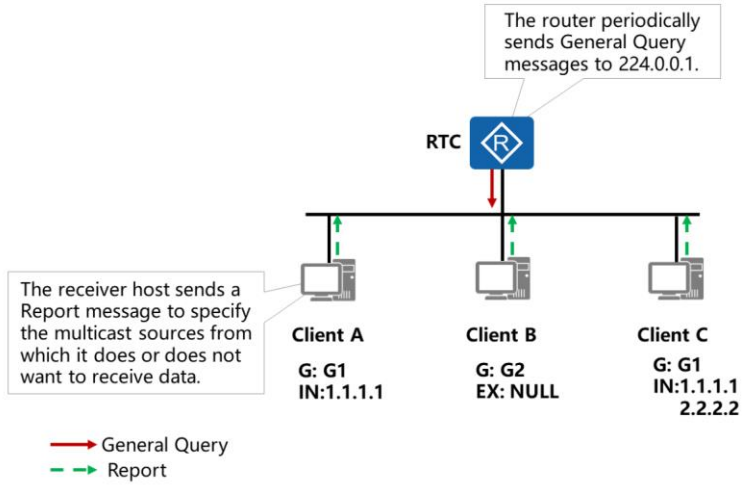
New Requirements in the SSM Model

- Receive multicast data only from specific sources



- If IGMPv1 or IGMPv2 is running between Client A and RTC, Client A cannot select multicast sources when it joins a group. It will receive data from both Source A and Source B, regardless of whether it requires the data.
- To meet requirements of the SSM model, IGMPv3 provides the capability to carry multicast source information in packets.
- Now, let’ s learn about the design of IGMPv3.

IGMPv3 Working Mechanism



- Compared with IGMPv2, IGMPv3 has the following changes:
  - IGMPv3 also defines two message types: Query and Report. Unlike IGMPv2, IGMPv3 does not define the Leave message. Group members send Report messages of a specified type to notify multicast routers that they have left a group.
  - In addition to General Query and Group-Specific Query messages, IGMPv3 defines a new Query message type: Group-and-Source-Specific Query. A querier sends a Group-and-Source-Specific Query message to members of a specific group on the shared network segment, to check whether the group members want data from specific sources. A Group-and-Source-Specific Query message carries one or more multicast source addresses.

- A Membership Report message contains not only the group that a host wants to join but also the multicast sources from which it wants to receive data. IGMPv3 adds a source filter mechanism and defines two filter modes: INCLUDE and EXCLUDE. Group-source mappings are represented by (G, INCLUDE, (S1, S2...)) or (G, EXCLUDE, (S1, S2...)). A (G, INCLUDE, (S1, S2...)) entry indicates that members of group G want to receive only data sent from sources S1, S2, and so on. A (G, EXCLUDE, (S1, S2...)) entry indicates that members of group G want to receive data from multicast sources except S1, S2, and so on. When group-source mappings change, hosts add these changes to the Group Record field in IGMPv3 Report messages and send the messages to the IGMP querier on the local network segment.
- An IGMPv3 Report message can carry multiple groups, whereas an IGMPv1 or IGMPv2 Report message can carry only one group. Therefore, IGMPv3 greatly reduces the number of messages transmitted on a network.

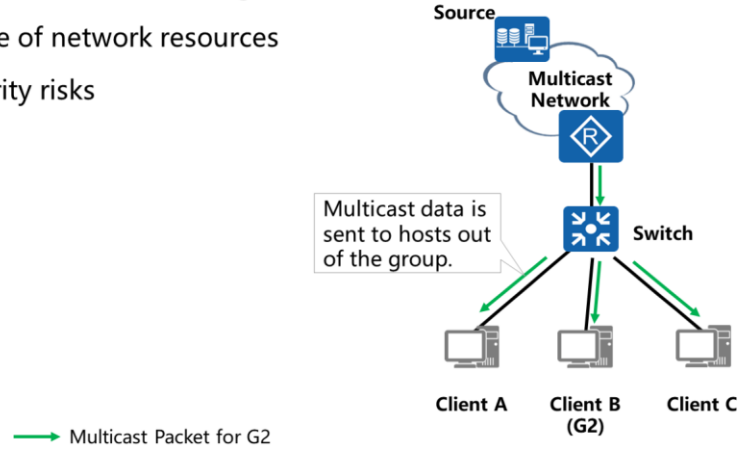
Differences Between IGMP Versions

Mechanism	IGMPv1	IGMPv2	IGMPv3
Querier election	Depends on another protocol	Depends on its own	Depends on its own
Member leaving mode	Leave without sending Leave messages	Send a Leave message	Send a Leave message
Group-specific query	Not supported	Supported	Supported
Source-and-group-specific query	Not supported	Not supported	Supported

Contents

Problems in L2 Multicast Data Forwarding

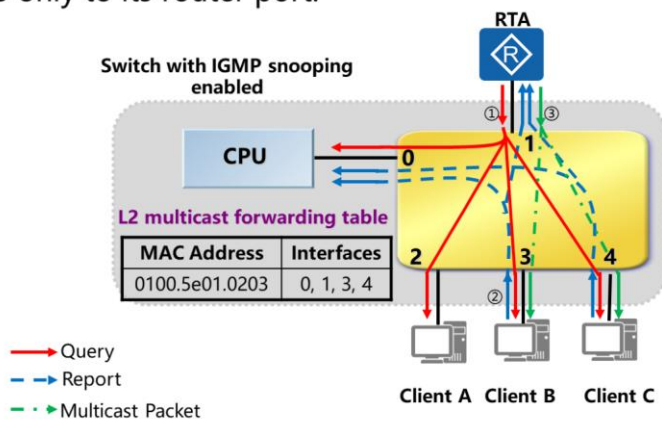
- L2 multicast data flooding causes:
  - Waste of network resources
  - Security risks



- To join a multicast group, a host needs to send an IGMP Report message to the upstream router, so that the router will forward multicast data packets to the host. IGMP messages are encapsulated in IP packets (Layer 3 packets). However, Layer 2 devices between hosts and multicast routers cannot process Layer 3 information carried in IP packets. In addition, Layer 2 devices cannot learn any MAC multicast address because the source MAC addresses of link-layer data frames are not multicast MAC addresses.
- When a Layer 2 device receives a data frame with a multicast destination MAC address, the device cannot find a matching entry in its MAC address table. Consequently, the device broadcasts the multicast packet. This wastes network resources and poses threats to network security.
- IGMP snooping suppresses Layer 2 multicast floods. Now, let’ s learn about this mechanism.

IGMP Snooping Working Mechanism

- After IGMP snooping is enabled on a switch, the switch forwards Report messages only to its router port.



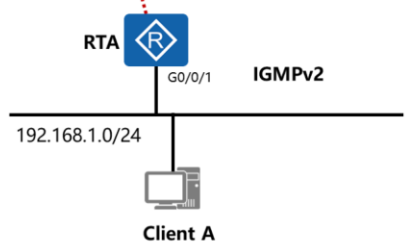
- IGMP snooping implements forwarding and control of multicast data frames at the data link layer.
- After IGMP snooping is enabled on a Layer 2 device, the device listens to IGMP messages exchanged between hosts and the upstream router. The device creates and maintains a Layer 2 multicast forwarding table based on information carried in IGMP messages (such as the message type, group address, and inbound interface). Then multicast data frames can be forwarded to specified receivers at the data link layer.
- IGMP snooping creates and maintains a Layer 2 multicast forwarding table as follows:
  - RTA, the querier, periodically sends General Query packets, which are flooded to all interfaces of the switch, including the internal interface 0 connected to the CPU. After the switch’ s CPU receives a Query message, it determines that interface 1 is connected to the router.

- Client B wants to join group 224.1.2.3, and sends an IGMP Report message carrying destination MAC address 0x0100.5e01.0203. The Report message is flooded to the interfaces which is connected to router of the switch, including interface 0 connected to the CPU. When the CPU receives the IGMP Report from Client B, it adds the interface who receive the report message into L2 multicast forwarding entry. The entry includes the interface connected to Client B, interface connected to the router, and the interface connected to the CPU.
- Then all the multicast packets with destination MAC address 0x0100.5e01.0203 are sent only to interfaces 0, 1, and 3.
- Client C joins 224.1.2.3 and sends an IGMP Report message. When the switch's CPU receives the message, it adds interface 4 to the forwarding entry of 0x0100.5e01.0203.

Contents

IGMP Configuration

```
multicast routing-enable
interface G0/0/1
 ip address 192.168.1.1 255.255.255.0
 igmp enable
 igmp version 2
```



IGMP Configuration Verification

```
<RTA>display igmp interface
Interface information of VPN-Instance: public net
 GigabitEthernet0/0/1(192.168.1.1):
   IGMP is enabled
   Current IGMP version is 2
   IGMP state: up
   IGMP group policy: none
   IGMP limit: -
   Value of query interval for IGMP (negotiated): -
   Value of query interval for IGMP (configured): 60 s
   Value of other querier timeout for IGMP: 0 s
   Value of maximum query response time for IGMP: 10 s
   Querier for IGMP: 192.168.1.1 (this router)
  Total 1 IGMP Group reported

<RTA>display igmp group
Interface group report information of VPN-Instance: public net
 GigabitEthernet0/0/1(192.168.1.1):
  Total 1 IGMP Group reported
   Group Address   Last Reporter   Uptime      Expires
   239.255.255.250 192.168.1.11     00:04:18    00:02:07
```

Quiz

1. In IGMPv1, when the last member leaves a group, how long will the multicast router wait before deleting the multicast forwarding entry of the group?
2. Is 224.0.0.1 the destination IP address of IGMPv2 Group-Specific Query messages?
3. How does IGMP snooping work?

HUAWEI

- Answer: 60 x 2 + 10 = 130s
- Answer: No. The destination IP address of a Group-Specific Query message is the IP address of the group to be queried.
- Answer: IGMP snooping listens to IGMP messages exchanged between multicast routers and hosts to create and maintain a Layer 2 multicast forwarding table. Then multicast data frames can be forwarded at the data link layer based on the Layer 2 multicast forwarding table.


Thank You
www.huawei.com

PIM Principles and Configurations

Foreword

- Multicast packets are sent to a specific group of receivers, which may be distributed at any locations on the network. To forward multicast packets correctly and efficiently, multicast routers need to create and maintain multicast routing entries.
- As more multicast routing protocols and applications are used, people realize that if multicast routes are dynamically generated by multiple routing algorithms like unicast routes, it will be too complex for different routing protocols to import routes of each other.
- Protocol Independent Multicast (PIM) implements Reverse Path Forwarding (RPF) checks for multicast packets based on unicast routing tables, and then creates multicast routing entries and forwards multicast packets accordingly.

- RPF: Reverse Path Forwarding

Objectives

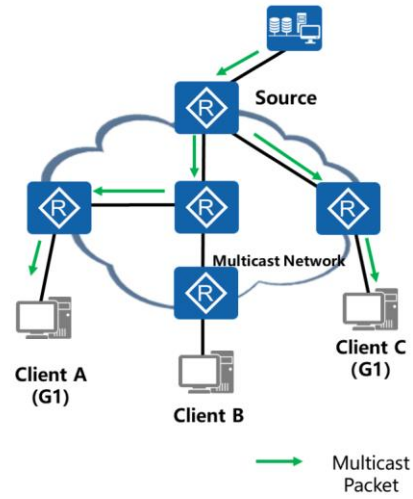
- Upon completion of this section, you will be able to:
  - Understand multicast forwarding requirements
  - Master PIM-DM working mechanisms and configurations
  - Master PIM-SM working mechanisms and configurations

Contents

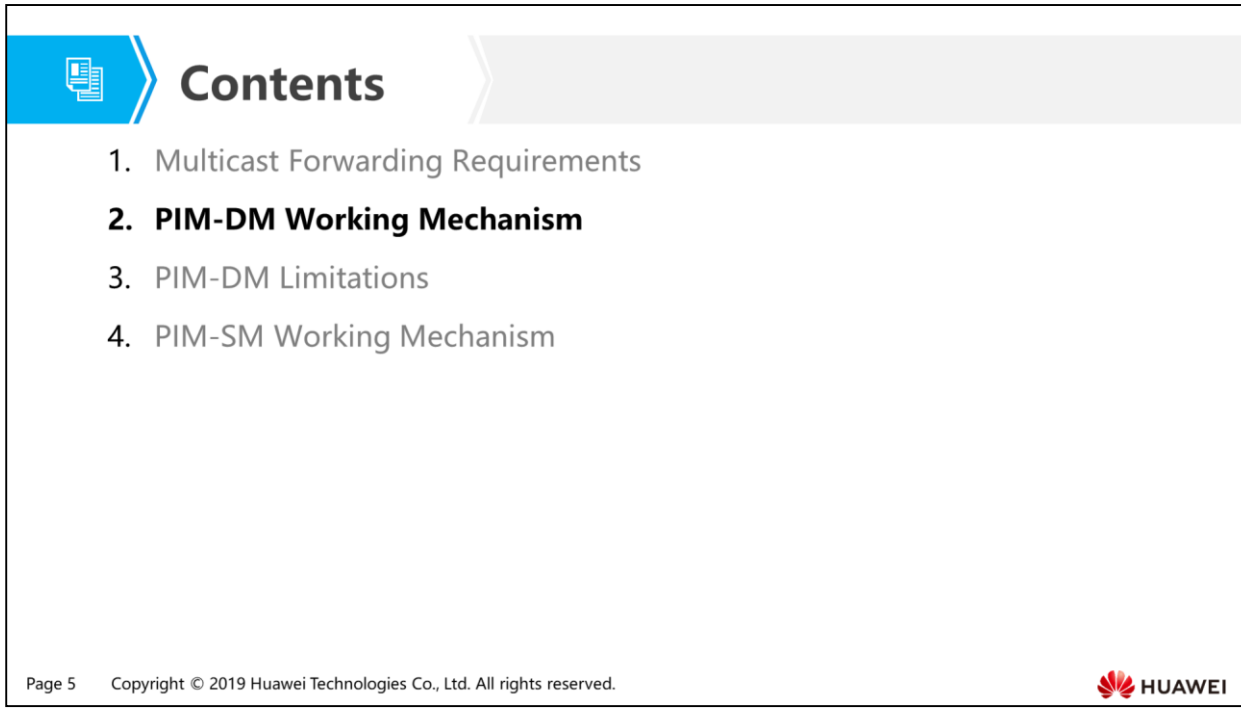
HUAWEI

How a Router Forwards Multicast Packets

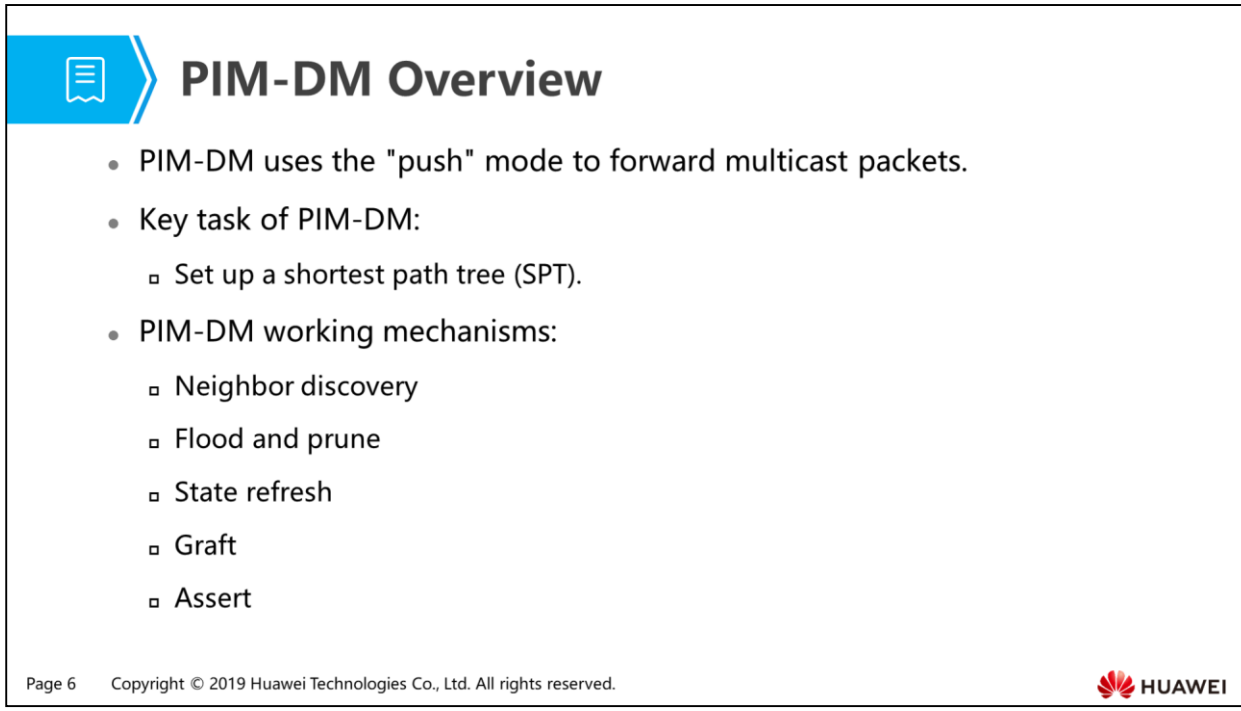
- A router forwards multicast packets based on the following information:
  - Whether there are receivers on network segments connected to its interfaces
  - Data of which groups is required by receivers
- Manually configuring the preceding information will cause the following problems:
  - Incapable of real-time updates
  - Low flexibility
  - Huge workload and high probability of errors



- In unicast forwarding, a router looks up the destination IP address of unicast packets in its unicast routing table to find a route to forward the packets. Unicast routes can be manually configured or dynamically learned using a routing protocol.
- In multicast applications, receivers may exist in any locations on a network. Static configuration of multicast routes cannot ensure real-time service delivery and service flexibility, requires heavy manual workload, and is prone to errors.
- To correctly and efficiently forward multicast data packets, routers need to run multicast routing protocols.



- PIM routing entries are created by the PIM protocol to guide multicast forwarding. PIM routing entries fall into two types: (S, G) and (*, G), where S indicates a specific multicast source, G indicates a specific multicast group, and * indicates any multicast source.
- (S, G) routing entries are used to establish an SPT on a PIM network and apply to both PIM-DM and PIM-SM networks.
- (*, G) routing entries are used to establish an RPT on a PIM network and apply only to PIM-SM networks.
- A PIM router may have both (S, G) and (*, G) entries. When the router receives a multicast packet with the source address S and the group address G, the router forwards the packet according to the following rules after the packet passes the RPF check:
    - If an (S, G) entry is available, the router forwards the multicast packet according to the (S, G) entry.
    - If the router has no (S, G) entry but has a (*, G) entry, the router creates an (S, G) entry based on this (*, G) entry, and then forwards the packet according to the (S, G) entry.



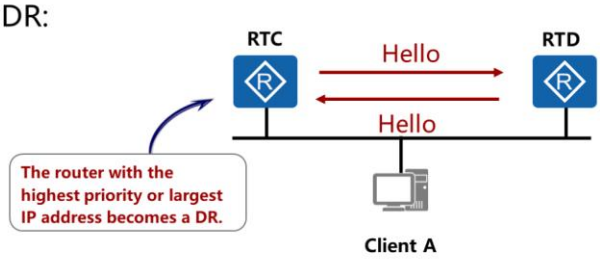
- PIM is Protocol Independent Multicast. The commonly used version is PIMv2. PIM packets are encapsulated in IP packets, carrying the protocol ID 103 and group address 224.0.0.13.
- In a PIM domain, a point-to-multipoint multicast forwarding path is set up from a multicast source to each multicast group. A multicast forwarding path looks like a tree, so it is also called a multicast distribution tree (MDT).
- Characteristics of an MDT:
  - Each link transmits at most one copy of identical data, regardless of how many group members exist on the network.
  - Replication of multicast data starts at the junction point as far from the multicast source as possible.
- PIM has two working modes:
  - Protocol Independent Multicast Dense Mode (PIM-DM).
  - Protocol Independent Multicast Sparse Mode (PIM-SM).
- PIM-DM assumes that group members are densely distributed on a network and each network segment may have group members.

- The design of PIM-DM is as follows:
- First, flood multicast packets to all network segments.
- Then prune the network segments with no group members.
- Set up and maintain a unidirectional, loop-free SPT from the multicast source to group members through periodic flood-prune processes.
- Key mechanisms of PIM-DM include neighbor discovery, flood and prune, state refresh, graft, and assert.

PIM-DM Neighbor Discovery

- Discover neighbors using Hello messages:

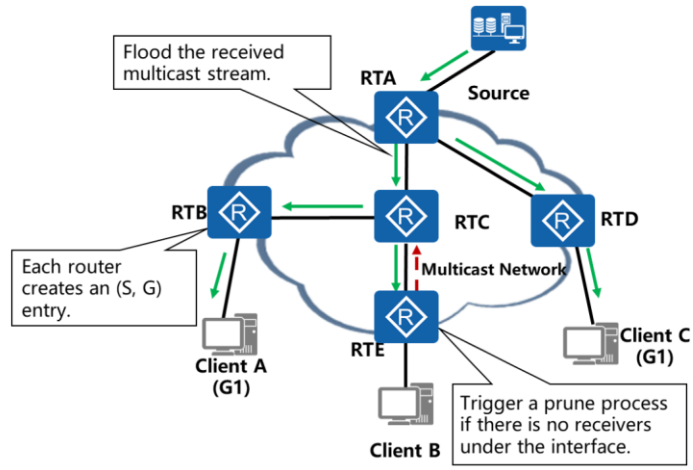
- Select a DR:



- In a PIM-DM network, multicast routers periodically send Hello messages to discover neighbors and maintain neighbor relationships.
  - The pim timer hello interval command sets the interval between Hello messages in the interface view. The default Hello interval is 30 seconds.
  - The pim hello-option holdtime interval command sets the Hello message timeout interval in the interface view. The default timeout interval is 105 seconds.
- DR election:
  - On a PIM-DM network, routers compare priorities and IP addresses carried in Hello messages to select a designated router (DR) on the multi-access network.
  - The DR acts as the IGMPv1 querier.
  - The router with the highest DR priority becomes the DR on the multi-access network. If multiple routers have the same highest DR priority, the router with the largest interface IP address becomes the DR.
- If the DR fails, a new DR will be elected among the other routers.

SPT Setup on a PIM-DM Network

- Flooding
- RPF check
- Pruning

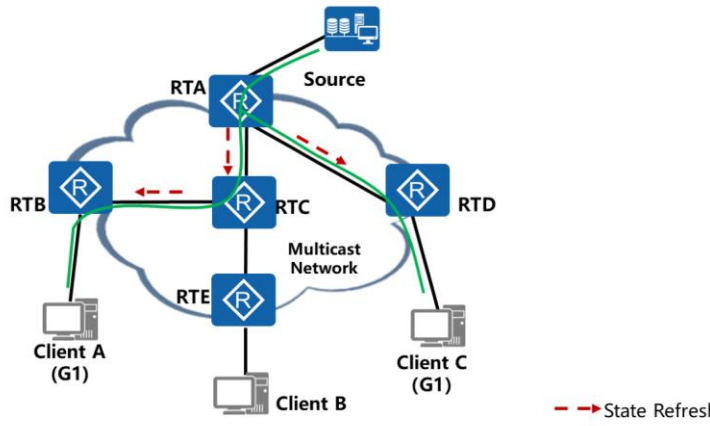


- Flooding: PIM-DM assumes that all hosts on a network are ready to receive multicast data. When a multicast source begins to send data to multicast group G, the data flooding process is as follows:
  - When a router receives a multicast data packet, it performs an RPF check.
  - If the RPF check succeeds, the router creates an (S, G) entry, and then forwards the data packet to all downstream PIM-DM nodes (flooding).
  - If the RPF check fails, the router drops the packet.
- Routers must perform RPF checks to prevent duplicate multicast packets and packet loops during multicast data forwarding.
  - A router performs an RPF check by looking up the route to the multicast source to determine whether the multicast packet is received from the correct upstream interface. On a router, the outbound interface of the route to a multicast source is the RPF interface for this multicast source. A router performs an RPF check after receiving a multicast data packet from an interface. If this interface is not the RPF interface of the multicast source, the RPF check fails, and the router drops the packet.

- Pruning: If there are no group members on a downstream branch, flooding multicast packets will cause a waste of bandwidth. The prune mechanism of PIM-DM can conserve bandwidth.
- If there are no group members on a downstream branch, the router sends a Prune message to the upstream router, notifying it that multicast data does not need to be forwarded to this branch. After the upstream router receives the Prune message, it deletes the downstream interface from the downstream interface list of the matching (S, G) entry. The pruning process continues until only the necessary branches are left on the PIM-DM network. In this way, an SPT from the multicast source is established.
- Each pruned interface has a timer. When the timer expires, a flood-prune process starts again. The default prune timeout timer value is 210 seconds.
- The flood-prune process repeats every 3 minutes on a PIM-DM network. RTC has pruned the downstream interface connected to RTE and started a prune timer for this interface. When the timer times out, RTC will restore multicast data forwarding to RTE. This causes waste of network resources.

State Refresh

- Periodically refresh the prune state.

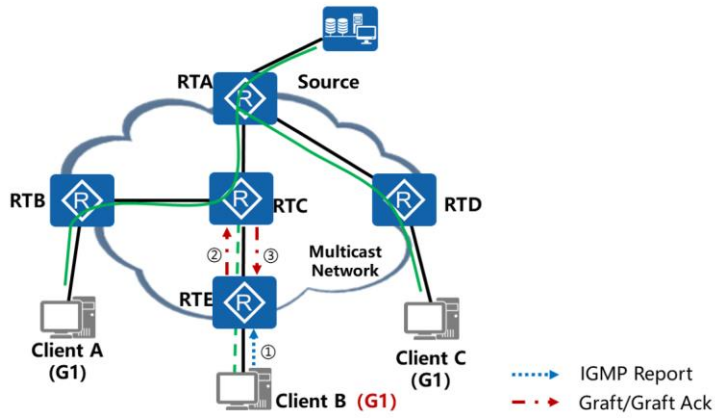


HUAWEI

- The PIM-DM protocol uses the state refresh mechanism to prevent waste of network resources caused by periodic flood-prune processes. The first-hop router (RTA) nearest to the multicast source periodically sends State Refresh messages. The State Refresh messages flood on the entire network to refresh the prune timers on all routers.
- The state refresh mechanism prevents RTE from periodically receiving multicast data. However, if this branch remains in pruned state after Client B joins group G1, Client B cannot receive multicast data.
- How can this problem be solved?

Graft Mechanism

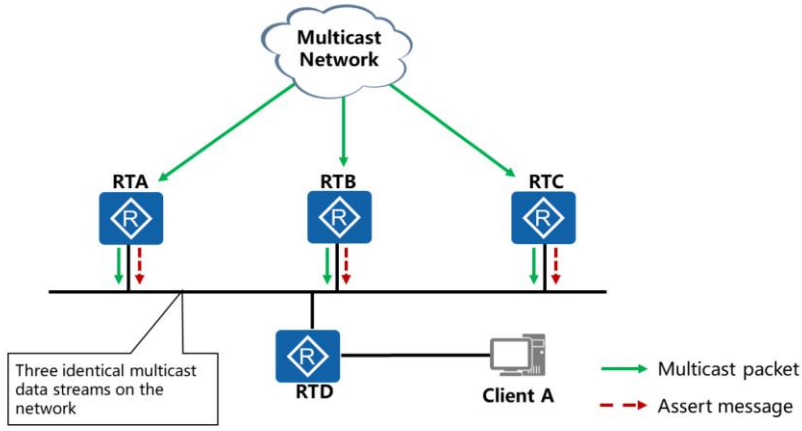
- Enable new group members to quickly receive multicast packets.



- As shown in the figure, Client B sends an IGMP Report message to request for multicast data sent to group G1. RTE receives the IGMP Report message from Client B and knows that multicast data needs to be forwarded to the downstream interface. So, RTE sends a Graft message to the upstream router RTC immediately, requesting RTC to resume data forwarding to the corresponding interface. After receiving the Graft message, RTC replies with a Graft Ack message and sets the outbound interface connected to RTE to forwarding state.

Assert Mechanism

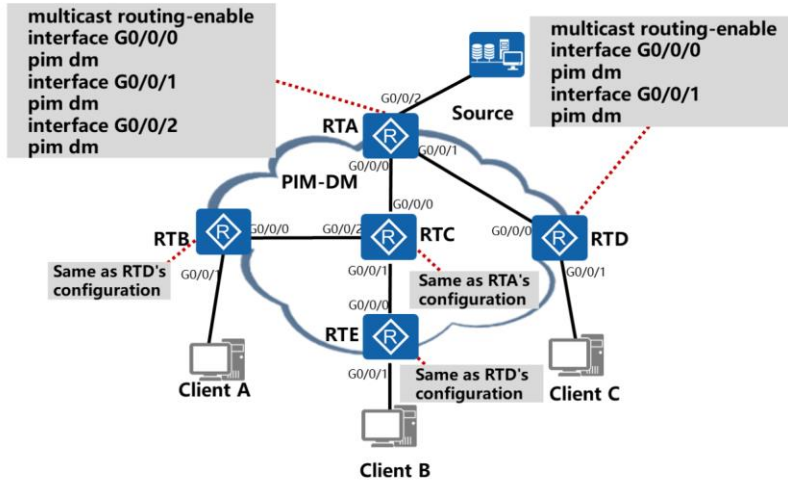
- Suppress duplicate multicast packets.



- As shown in the figure, RTA, RTB and RTC all receive multicast packets from their upstream interfaces, and the RPF checks succeed on the three routers. Downstream interfaces of the routers are connected to the same network segment. RTA, RTB, and RTC all send multicast packets to the network segment. The three copies of duplicate multicast packets waste bandwidth resources.
- To prevent this problem, a PIM router will send a multicast Assert message to all PIM routers on the shared network segment after receiving duplicate multicast packets from a neighboring router. The destination address of the Assert message is 224.0.0.13. When the other PIM routers receive the Assert message, they compare local parameters with those carried in the Assert message for assert election. The assert winner is selected following these rules:
  - The router with the smallest preference value of the unicast route to the multicast source wins.
  - If multiple routers have the same route preference, the router with the smallest route cost to the multicast source wins.
  - If multiple routers have the same router preference and cost to the multicast source, the router with the largest downstream interface IP address wins.

- The PIM routers perform the following operations based on the assert election result:
- The downstream interface of the router that wins the election is the assert winner and forwards multicast packets to the shared network segment.
- The downstream interfaces of the PIM routers that lose the election are assert losers and no longer forward multicast packets to the shared network segment. The PIM routers delete the downstream interfaces from the downstream interface list of their (S, G) entries.
- After the assert election is complete, only one downstream interface is active on the network segment, so only one copy of multicast packets is transmitted to the network segment.
- All assert losers can periodically resume multicast packet forwarding, which triggers periodic assert elections.

PIM-DM Configuration



PIM-DM Configuration Verification

```
<RTD>display pim routing-table
VPN-Instance: public net
Total 1 (*, G) entry; 1 (S, G) entry

(192.168.0.1,  239.255.255.250)
   Protocol: pim-dm, Flag: ACT
   UpTime: 00:00:09
   Upstream interface: GigabitEthernet0/0/0
      Upstream neighbor: 10.1.14.1
      RPF prime neighbor: 10.1.14.1
   Downstream interface(s) information:
   Total number of downstreams: 1
      1: GigabitEthernet0/0/1
         Protocol: pim-dm, UpTime: 00:00:09, Expires:  -
```

```
<RTD>display pim neighbor
VPN-Instance: public net
Total Number of Neighbors = 1

 Neighbor      Interface    Uptime   Expires  Dr-Priority  BFD-Session
 10.1.14.1     GE0/0/0      00:12:19 00:01:16     1          N
```

Contents

PIM-DM Limitations

- PIM-DM is applicable to campus networks with densely distributed group members.
- PIM-DM limitations:
  - On a network with sparsely distributed group members, periodic flooding of multicast traffic will bring great pressure to the network.

- PIM-DM is applicable to campus networks with densely distributed group members.
- On a network with sparsely distributed group members (Internet), periodic flooding of multicast traffic will bring great pressure to the network.
- The PIM-SM mode provides effective solutions to limitations of PIM-DM.

Contents

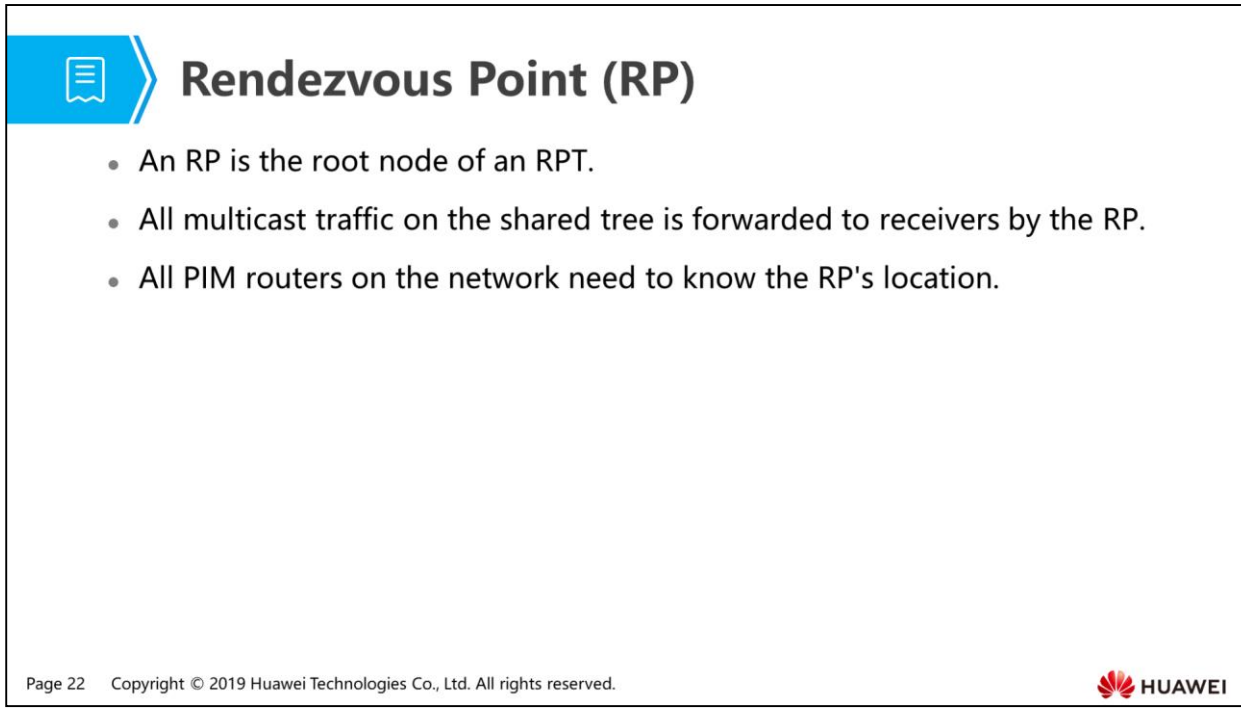
PIM-SM Overview

- PIM-SM uses the "pull" mode to forward multicast packets.
- Key tasks of PIM-SM:
  - Set up a rendezvous point tree (RPT), also called a shared tree.
  - Set up a shortest path tree (SPT).
- PIM-SM is applicable to networks with sparsely distributed group members.

HUAWEI

- Compared with PIM-DM that uses the push mode, PIM-SM uses the pull mode to forward multicast packets. PIM-SM assumes that group members are distributed sparsely on a network, and almost all network segments have no group members. Multicast routes are created for data forwarding to a network segment only when group members appear on the network segment. PIM-SM is usually used for networks with a large number of sparsely distributed group members.
- PIM-SM is implemented as follows:
  - A PIM router works as the rendezvous point (RP) to serve group members or multicast sources that appear on the network. All PIM routers on the network know the RP's position.
  - When a new group member appears on the network (a host joins a group G through IGMP), the last-hop router sends a Join message to the RP. A (*, G) entry is then created hop by hop, and finally the routers establish an RPT with the RP as the root.
  - When an active multicast source appears on the network (the multicast source sends the first multicast data packet to a group G), the first-hop router encapsulates the multicast data in a Register message and sends the Register message to the RP in unicast mode. The RP then creates an (S, G) entry, and the multicast source is registered on the RP.

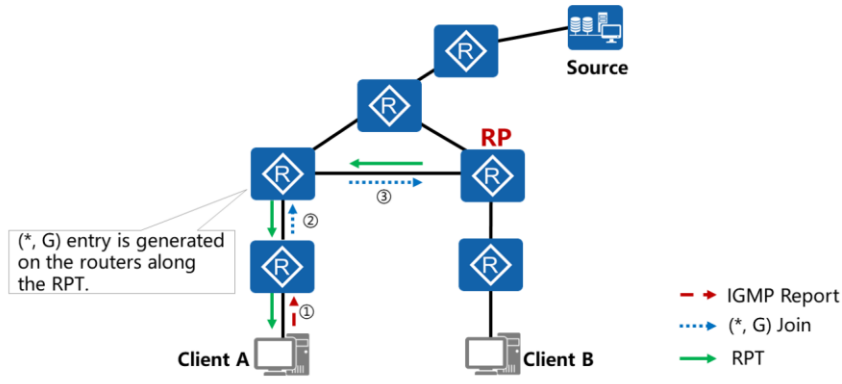
- Key mechanisms of PIM-SM include neighbor discovery, DR election, RP discovery, RPT setup, multicast source registration, SPT switchover, and assert. You can also configure a Bootstrap router (BSR) to implement fine-grained management in a single PIM-SM domain. The neighbor discovery and assert mechanisms are same as those in PIM-DM.



- An RP provides the following functions:
  - Acts as the core router in a PIM-SM domain and the root node of an RPT.
  - Forwards all multicast traffic on the shared tree to receivers.
- You can specify the range of multicast groups served by the RP using a command.
- An RP can be statically specified or dynamically selected:
  - To use a static RP, you need to specify the RP address on each PIM-SM router, so that each router knows the RP’ s location.
  - A dynamic RP is elected among multiple candidate RPs (C-RPs) using a dedicated protocol. You need to enable the protocol used for RP election and configure multiple PIM-SM routers as C-RPs.
  - RP configuration recommendations:
  - On a small or medium network, configure a static RP, because the static RP is stable and has low requirements on equipment performance.
  - If a network has only one multicast source, specify the router directly connected to the multicast source as the static RP. This configuration saves the process of registering the multicast source to the RP by a source DR.
  - To use a static RP, ensure that all routers (including the RP) in the PIM-SM domain are configured with the same RP information and multicast group range.

- On a large network, use the dynamic RP election mode to improve reliability and facilitate maintenance.
- If multiple densely distributed multicast sources exist on the network, configure the core routers close to the multicast sources as C-RPs. If there are many densely distributed group members, configure the core routers close to group members as C-RPs.

RPT Setup

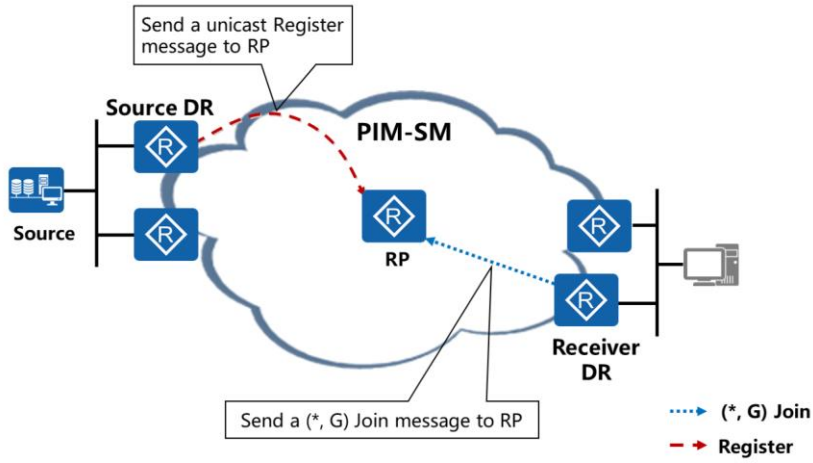


- Question: If two routers are connected to Client A, will the two routers both send (*, G) Join messages to the RP?

HUAWEI

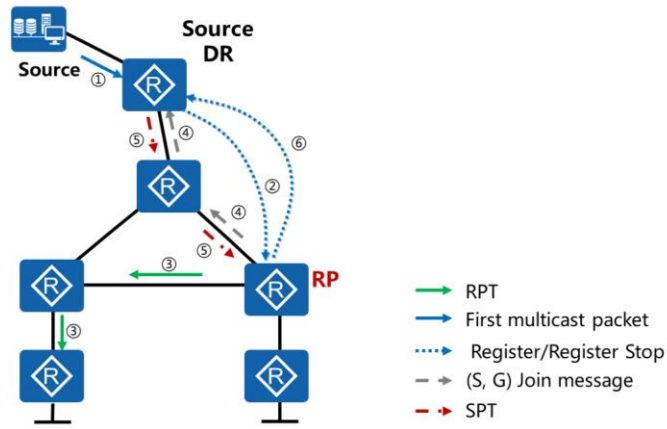
- An RPT is set up through the following process:
  - When a host joins a group, it sends an IGMP Report message.
  - The last-hop router sends a (*, G) Join message to the RP.
  - When the (*, G) Join message is transmitted toward the RP, each router along the forwarding path creates the matching (*, G) entry.
- The RPT enables on-demand forwarding of multicast data, thereby reducing bandwidth consumption caused by data flooding.

Receiver DR and Source DR



- Designated routers (DRs) need to be elected on a PIM-SM network. There are two types of DRs: receiver DR and source DR.
  - A receiver DR is connected to multicast group members and is responsible for sending (*, G) Join message to the RP.
  - A source DR is connected to a multicast source and is responsible for sending unicast Register messages to the RP.
- The DR election rules used in the PIM-SM mode are the same as those used in the PIM-DM mode.

SPT Setup



- After the SPT is set up, multicast packets are sent to the RP along the SPT.

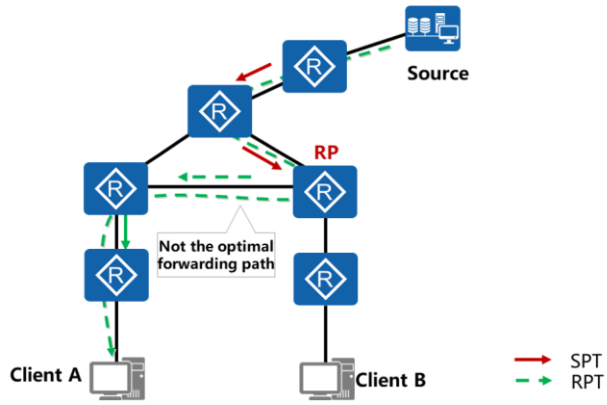
- On a PIM-SM network, any new multicast source must register on the RP so that the RP can forward multicast data from the multicast source to group members. The multicast source registration process is as follows:
  - A multicast source sends the first multicast packet to group G.
  - The source DR encapsulates the multicast packet in a Register message and sends it to the RP in unicast mode.
  - When the RP receives the register message, it decapsulates the message to obtain the multicast packet and forwards it to receivers along the RPT.
  - Meanwhile, the RP sends an (S, G) Join message to the source DR. Then all routers along the transmission path of the message create the corresponding (S, G) entry. In this way, an SPT from the multicast source to the RP is established.
  - After the SPT is established, multicast data packets sent from the multicast source are forwarded to the RP along the SPT.
  - After receiving multicast packets through the SPT, the RP sends a unicast Register-stop message to the source DR.

(*, G) and (S, G) Entries

Mode	Type	Scenario
PIM-DM	(S, G)	SPT from the first-hop router to the last-hop router
PIM-SM	(*, G)	RPT from the RP to the last-hop router
	(S, G)	SPT from the source DR to the RP
	(S, G)	SPT from the first-hop router to the last-hop router after an SPT switchover

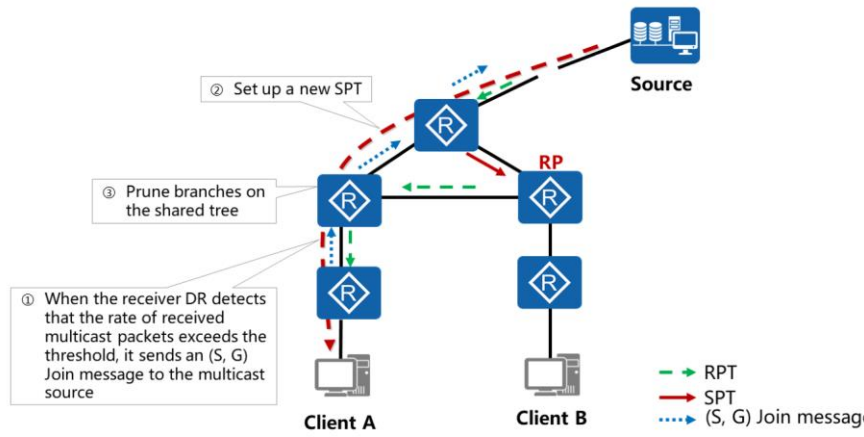
PIM-SM Forwarding Trees

- Are there any potential problems with a forwarding path composed by the SPT and RPT?



- A PIM-SM network contains both SPT and RPT trees. Generally, multicast data packets sent from a multicast source reach the RP along the SPT, and are then forwarded to receivers from the RP along the RPT.
- In this case, the path from the multicast source to the receivers may not be the optimal one, and the load of the RP is high. The RPT-to-SPT switchover mechanism can address this problem.

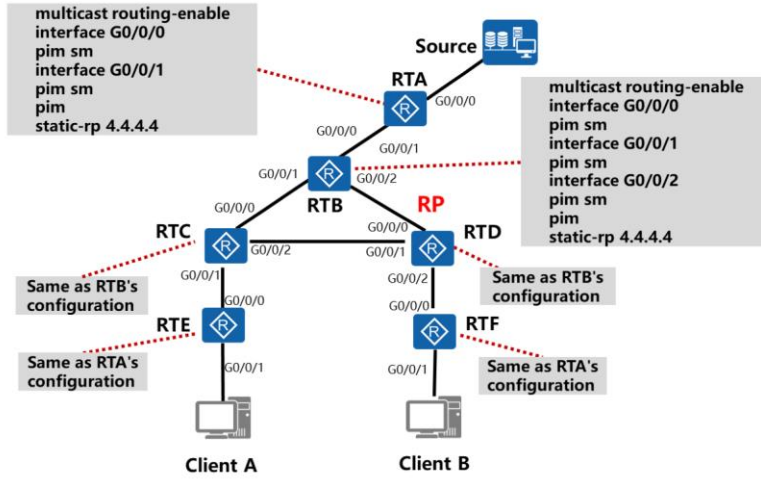
SPT Switchover



- On a PIM-SM network, a bandwidth usage threshold can be configured to implement RPT-to-SPT switchover.
- The receiver DR periodically checks the forwarding rate of multicast packets. When the receiver DR finds that the rate of multicast data packets sent from the RP to group G exceeds the threshold, it triggers an SPT switchover.
  - The receiver DR sends an (S, G) Join message to the source DR and creates an (S, G) entry. The Join message is transmitted hop by hop, and routers along the path all create the corresponding (S, G) entry. Finally, an SPT is set up from the source DR to the receiver DR.
  - After the SPT is set up, the receiver DR sends a Prune message to the RP. The Prune message is transmitted hop by hop along the RPT. After receiving the Prune message, the routers on the RPT convert the (*, G) entry into the (S, G) entry, and prune their downstream interfaces. After the pruning process is complete, the RP no longer forwards multicast packets along the RPT.
  - If the SPT does not pass through the RP, the RP continues to send a Prune message to the source DR, so that routers along the path between the RP and source DR delete their downstream interfaces from the (S, G) entry. After the pruning process is complete, the source DR no longer forwards multicast data packets to the RP along the SPT from itself to the RP.

- According to default configuration of the VRP, routers connected to receivers join the SPT immediately after receiving the first multicast data packet from a multicast source, triggering an RPT-to-SPT switchover.
- Through the RPT-to-SPT switchover mechanism, PIM-SM can establish an SPT more precisely than PIM-DM.

PIM-SM Configuration

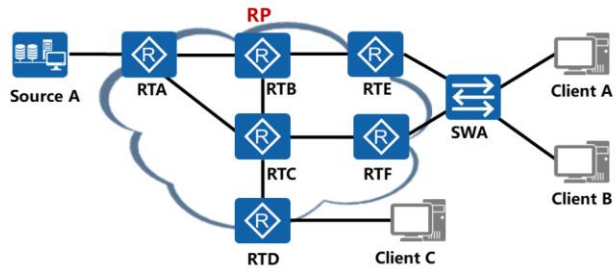


PIM-SM Configuration Verification

```
<RTF>display pim routing-table
VPN-Instance: public net
Total 1 (*, G) entry; 0 (S, G) entry

(*, 224.1.1.1)
    RP: 4.4.4.4
    Protocol: pim-sm, Flag: WC
    UpTime: 00:00:20
    Upstream interface: GigabitEthernet0/0/0
        Upstream neighbor: 10.1.46.4
        RPF prime neighbor: 10.1.46.4
    Downstream interface(s) information:
    Total number of downstreams: 1
        1: GigabitEthernet0/0/1
            Protocol: igmp, UpTime: 00:00:20, Expires: -
```

```
<RTB>display pim neighbor
VPN-Instance: public net
Total Number of Neighbors = 3

Neighbor        Interface          Uptime   Expires  Dr-Priority  BFD-Session
10.1.12.1       GE0/0/0            00:04:08 00:01:27 1           N
10.1.23.3       GE0/0/1            00:01:29 00:01:16 1           N
10.1.24.4       GE0/0/2            00:03:19 00:01:25 1           N
```

Comprehensive Multicast Experiment

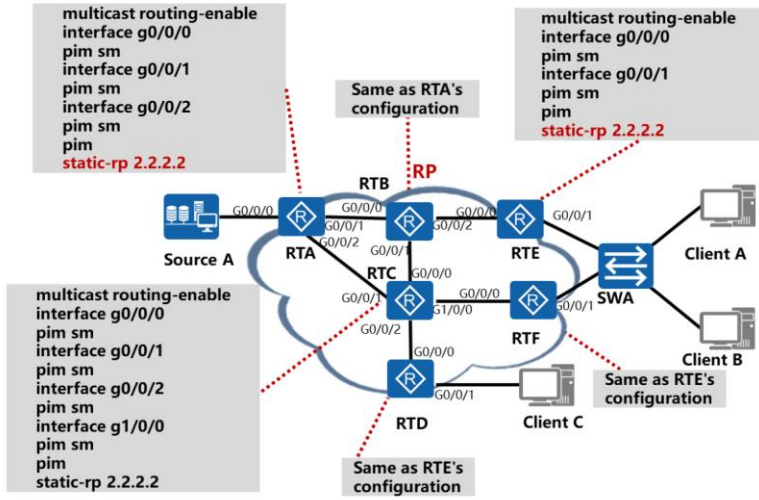


- On the network shown in the figure, deploy PIM-SM and specify RTB as the static RP.
- Configure IGMPv2 on the user networks and take measures to minimize resource consumption on the user networks and improve network security.
- RTE and RTF are connected to receivers. Set up an RPT from the RP to the RTE.
- RTD is connected to a VIP user network. The client on this network needs to receive multicast data immediately after joining group 238.1.1.1.

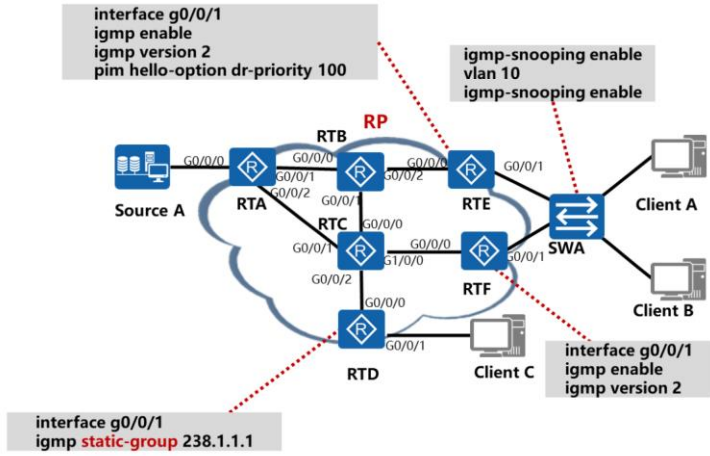
  HUAWEI

- Requirements analysis:
  - Enabling multicast routing is the prerequisite for PIM-SM. Therefore, you need to enable multicast routing on the routers, enable PIM-SM on interfaces of the routers, and then configure RTB as the static RP in the PIM view on the routers.
  - Enable IGMPv2 on the router interfaces connected to user hosts. To reduce resource consumption and improve system security, enable IGMP snooping on SWA so that SWA can forward multicast data frames efficiently and securely.
  - The receiver DR will trigger RPT setup toward the RP. According to the DR election rules, you need to set the DR priority of RTE’ s downstream interface to a value greater than 1 (default DR priority).
  - RTD should have a multicast forwarding entry for group 238.1.1.1. After receiving an IGMP Report message of this group, RTD will immediately forward multicast data packets based on the forwarding entry. You can statically bind RTD’ s downstream interface to group 238.1.1.1 to meet this requirement.

PIM-SM Configuration (1)



PIM-SM Configuration (2)



- Configure a static multicast group.
  - Static multicast groups can be configured in some special scenarios. For example, some group members need to receive multicast data for a long time, or multicast data needs to be forwarded to hosts that cannot send Report messages.
    - In these scenarios, you can configure a static multicast group on the user-side interface to enable fast, stable multicast data forwarding or direct multicast data flows to the interface. After a static multicast group is configured on an interface, the router considers that this group always has members on the network segment connected to the interface and always forwards multicast data of this group to the interface.
    - If hosts connected to an interface cannot identify or respond to multicast ping packets, you can configure the multicast ping function on this interface. Then the interface can not only receive multicast data, but also respond to multicast ping packets, which helps you locate network problems flexibly and conveniently.

Quiz

1. What is a multicast distribution tree? What types are multicast distribution trees classified into?
2. What is the function of the assert mechanism?
3. True or false: In the PIM-SM protocol, the DR connected to receivers sends unicast Register messages to the RP.

- Answer: A multicast distribution tree is a unidirectional loop-free data transmission path from a multicast source to the receivers. There are two types multicast distribution trees: SPT and RPT.
- Answer: The assert mechanism prevents duplicate packets from being transmitted on a shared network (such as an Ethernet network). The assert mechanism selects a unique forwarder on the shared network. The routers that lose the election prune their interfaces to disable multicast data forwarding to these interfaces.
- Answer: False. The DR connected to a multicast source is responsible for sending unicast Register messages to the RP.


Thank You
www.huawei.com

Route Control

Foreword

- Enterprise networks may encounter the problems such as unauthorized access of certain traffic and sub-optimal traffic path selection. To ensure data access security and improve link bandwidth usage, traffic behavior on the network must be controlled, for example, reachability control and traffic path adjustment.
- Tools may be required to meet complicated and precise traffic control requirements. This course introduces the traffic control tools and their usage scenarios.

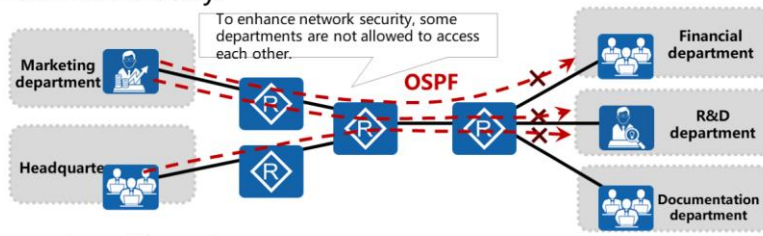
Objectives

- Upon completion of this section, you will be able to:
  - Master the method to control network traffic reachability
  - Master the method to adjust network traffic paths
  - Be familiar with the problems caused by route import and solutions

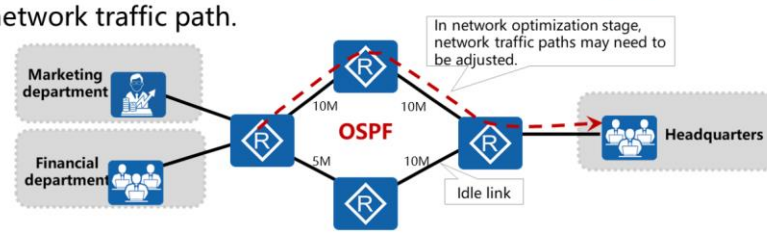
Contents

Traffic Behavior Control Requirement

1. Control traffic reachability.



2. Adjust network traffic path.



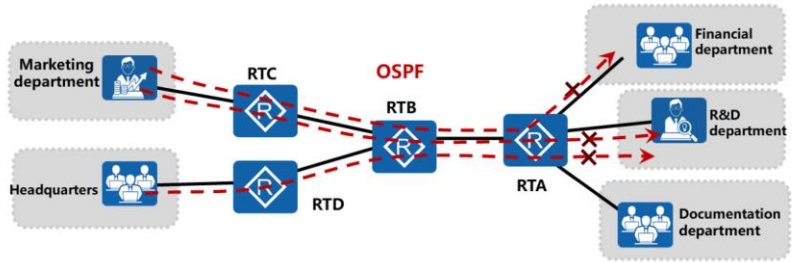
HUAWEI

- Control traffic reachability: As shown in the figure, the marketing department is not allowed to access the financial and R&D departments, and the headquarters is not allowed to access the R&D department.
- Adjust network traffic paths: As shown in the figure, OSPF is run to compute routes. The marketing and financial departments access the headquarters through the path with the minimum cost, even though the path is congested. The other link is always idle, causing a waste of bandwidth.

Contents

Control Traffic Reachability

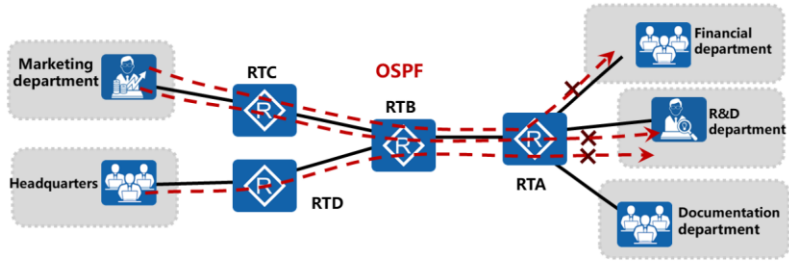
- Question: How to control network traffic reachability.



- Solution 1: Change the number of routing entries (filter the received and advertised routes) to control reachability. This is called **routing policy**.
- Solution 2: Deploy traffic filter to deny or permit specified traffic flow. This method is called **traffic filter**.

- When advertising, receiving, and importing routes, routing policy implements certain policies according to actual networking requirements in order to filter the routes and change the attributes of the routes. Purposes of routing policies are as follows:
  - Control route advertisement: advertises only the routes matching certain conditions.
  - Control route sending and receiving: receives only necessary and valid routes to control the capacity of routing table, enhancing network security.
  - Filter and control imported routes: imports only the routes meeting certain conditions from other routing protocols, and sets the attributes of imported routes.
  - Set route attributes: sets attributes for the routes filtered by routing policy.

Solution 1: Routing Policy

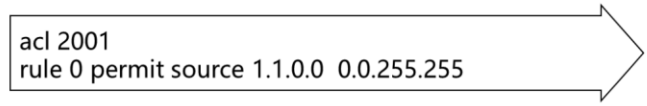


- Use the **Filter-Policy** tool to filter the routes imported from RTA to OSPF and the routes imported from RTC to the routing table:
  - Use ACL or IP-Prefix List to match the target flow.
  - Use the Filter-Policy in the protocol view to advertise policies for the target flow.
- Use the **Route-Policy** tool to filter the direct routes imported by RTA:
  - Use ACL or IP-Prefix List to match the target flow.
  - Use the Route-Policy in the protocol view to control the imported routes.

- Routing policies are implemented in the following steps:
  - Define attributes of routing information to which routing policies are applied. Define a set of matching rules regarding various attributes of routing information such as the destination address and AS number.
  - Apply matching rules to route advertising, receiving, and importing.
- Currently, the following filters are provided for routing protocols:
  - ACL
  - Address prefix list
  - AS_Path filter
  - Community attribute filter
  - Extended community attribute filter
  - RD filter

ACL Application Example (1)

- ACL classifies packets into different types by based on packet information.

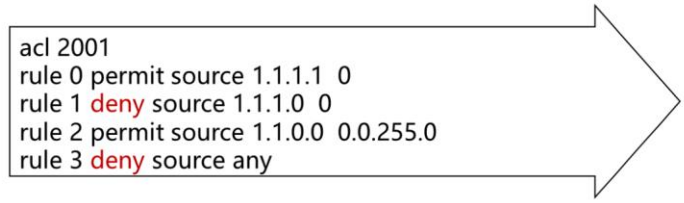


1.1.1.1/32	**1.1.1.1/32**
1.1.1.0/24	**1.1.1.0/24**
1.1.0.0/16	**1.1.0.0/16**
1.0.0.0/8	

- An ACL is composed of a list of rules. Each rule contains a permit or deny clause. These rules define packets with information in the packets.
- ACL classification:
  - Basic ACL: matches packets against source address, fragmentation flag, and time range. The ACL number ranges from 2000 to 2999.
  - Advanced ACL: matches packets against source address, destination address, source port number, destination port number, protocol type, priority, and time range. the ACL number ranges from 3000 to 3999.
  - Layer 2 ACL: matches packets against source MAC address, destination MAC address, and packet type. The ACL number ranges from 4000 to 4999.
  - User-defined ACL: matches packets against user-defined rules. The ACL number ranges from 5000 to 5999.
- An ACL can consist of multiple deny or permit clauses. Each clause describes a rule. After receiving data packets, the device matches the packets against the first ACL rule. If the packets do not match the first ACL rule, the device matches the packets against the next ACL rule. Once the packets match a rule, the device executes the action in the rule and does not match packets against other rules. If the packets do not match any rule, the device directly forwards the packets.

- Note: The ACL rules may overlap or conflict with each other. The rule matching order decides the rule priorities. The ACL processes rule overlapping or conflict based on rule priorities.

ACL Application Example (2)



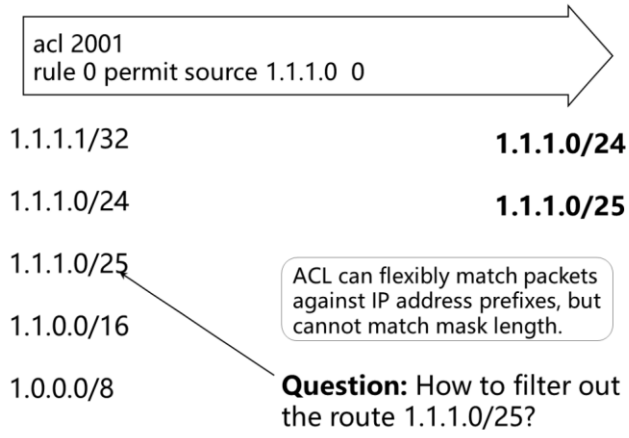
1.1.1.1/32 **1.1.1.1/32**

1.1.1.0/24 **1.1.0.0/16**

1.1.0.0/16

1.0.0.0/8

ACL Application Example (3)



IP-Prefix List Application Example (1)

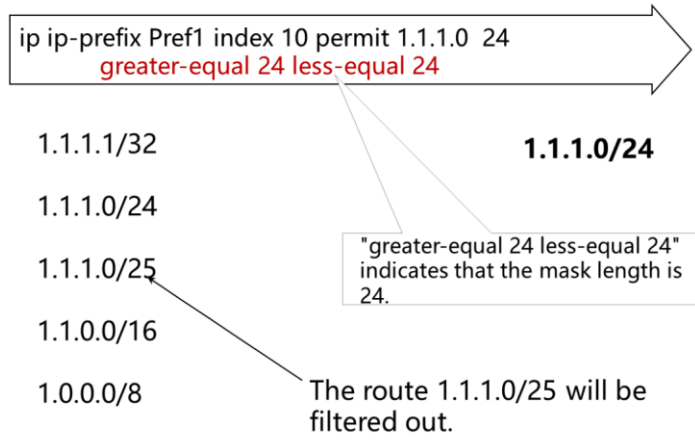
- IP-Prefix List can match both IP address prefix and mask length.
- IP-Prefix List cannot filter IP packets, but can filter only routing information.

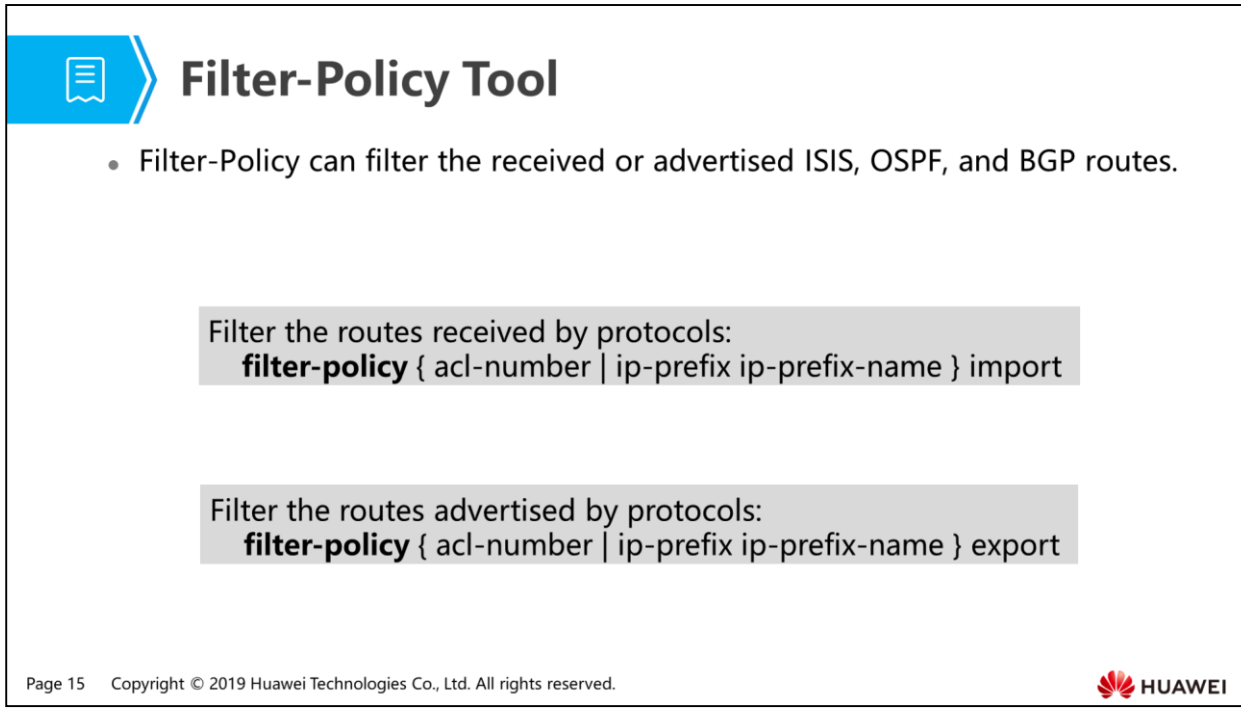
ip ip-prefix test index 10 permit 10.0.0.0 16 greater-equal 24 less-equal 28
IP address range: 10.0.0.0 – 10.0.x.x
24<= Mask length<=28
Example: 10.0.1.0/24, 10.0.2.0/25, 10.0.2.192/26

- IP prefix list is the address prefix list. The address prefix list filters the routes based on the defined prefix filter list.
- Format and matching rule of prefix list:
  - The prefix filter list consists of IP addresses and masks. An IP address can be a subnet address or host address. The mask length ranges from 0 to 32.
  - Each IP prefix in the IP prefix list has an index. The IP prefixes are processed in an ascending order of index.
  - If the indexes are not manually set, the device allocates indexes to the IP prefixes with an interval of 10. If a new IP prefix has the same name and index as an existing IP prefix but their contents are different, the new IP prefix overwrites the old one.
  - If no IP prefix is matched, the default matching mode of the last IP prefix is deny by default. If a referenced IP prefix list does not exist, the default mode permit is used.

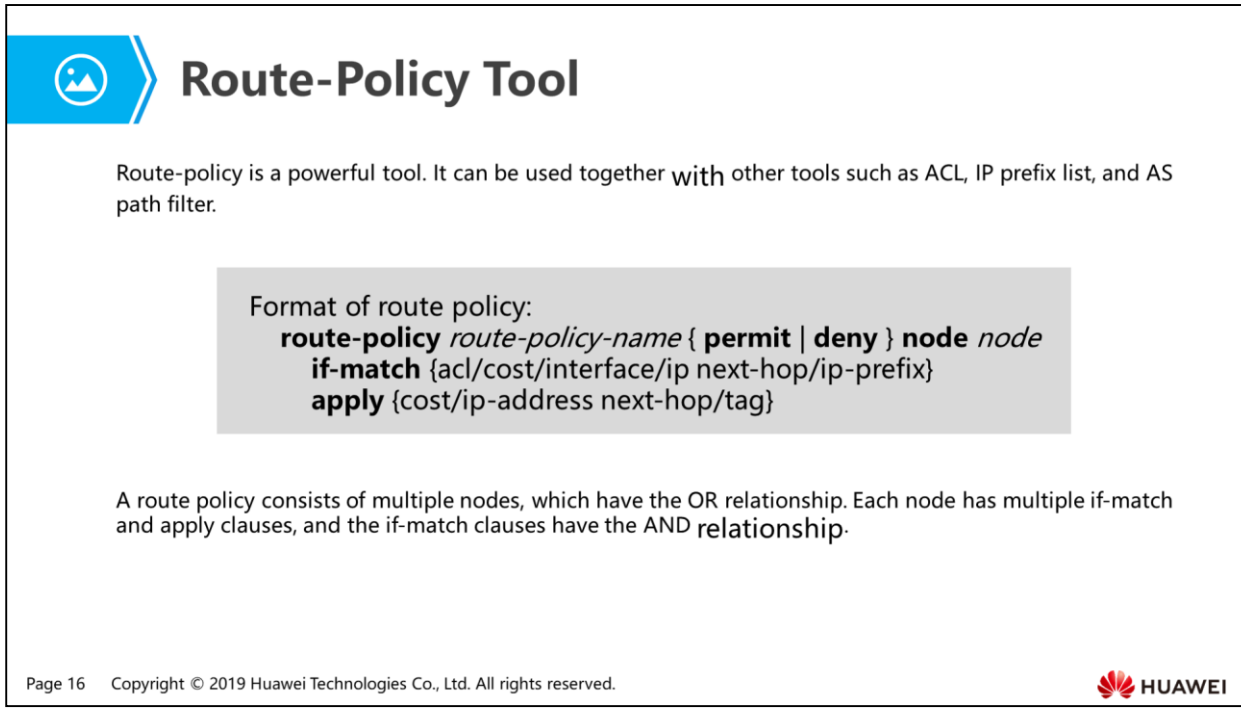
- Prefix mask length range:
- The IP prefix list can match a specific route or match the routes within a certain mask length. The prefix mask length can also be specified using the keywords greater-equal or less-equal. If the keyword greater-equal or less-equal is not specified, the precise matching is used. That is, only the route with the same mask length as the prefix list is matched. If the keyword greater-equal is specified, the routes of which the mask lengths range from greater-equal value to 32 bits are matched. If the keyword less-equal is specified, the routes of which the mask lengths range from the specified value to less-equal are matched.

IP-Prefix List Application Example (2)



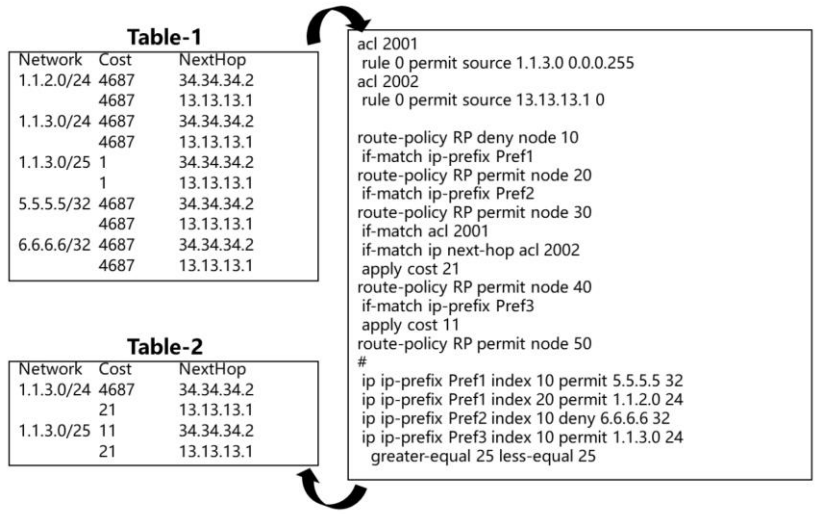


- The filter-policy tools of protocols can reference ACLs or address prefix lists to filter the received, advertised, and imported routes.
- For a distance-vector protocol and a link-state protocol, the operations of the filter-policy tools are different:
  - The distance-vector protocol generates routes based on routing table, so the filter will affect the routes received from and sent to neighbors.
  - The link-state protocol generates routes based on the link state base and the routing information is hidden in the link status LSA. However, the filter-policy cannot filter the advertised and received LSAs. Therefore, the filter-policy does not affect link state advertisement, integrity of link state, or routing table. It affects only the local routing table. Only filtered routes can be added to the routing table.
  - The routes advertised by the filter-policy export command of different protocols are different:
    - The routes imported to or discovered by the DV protocol are filtered.
    - Only the routes imported to the link status protocol are filtered.



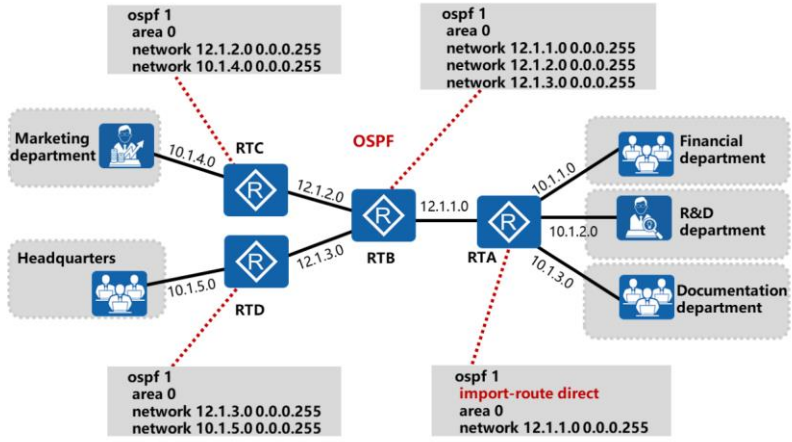
- Each node in the route policy matches permit or deny mode. If a node matches the permit mode, when a route meets all if-match clauses in the node, the route is allowed and the apply clause matching the node is executed, and the next node is not processed. If a route does not match any if-match clause in the node, the next node is processed for route filtering. If a node matches the deny mode, when a route matches all if-match clauses in the node, the route is not allowed and the apply clause of the node is not executed, and the next node is not processed. Otherwise, the next node is processed for route filtering.

Route-Policy Application Example

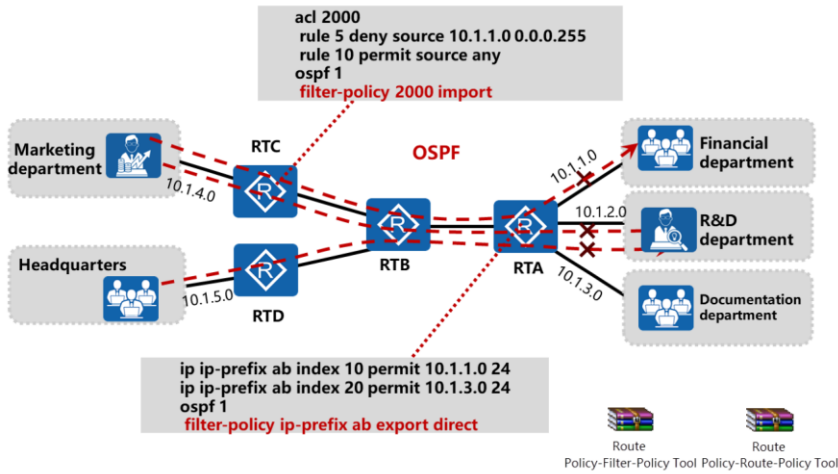


- Pref1 matches 5.5.5.5/32 or 1.1.2.0/24. They will be filtered out (deny) by node 10 of route-policy RP, so Table-2 does not contain 5.5.5.5/32 and 1.1.2.0/24.
- Pref2 filters out (deny) 6.6.6.6/32. Although node 20 in route-policy RP matches permit, 6.6.6.6/32 is still filtered out. Therefore, Table-2 does not contain 6.6.6.6/32.
- Node 30 in route-policy RP defines two if-match clauses, matching ACL 2001 and ACL 2002. The routes match ACL 2001 include 1.1.3.0/24 (next hop 34.34.34.2), 1.1.3.0/24 (next hop 13.13.13.1), 1.1.3.0/25 (next hop 34.34.34.2), and 1.1.3.0/25 (next hop 13.13.13.1). The routes 1.1.3.0/24 (next hop 13.13.13.1) and 1.1.3.0/25 (next hop 13.13.13.1) match both ACL 2001 and ACL 2002. Therefore, the cost values of 1.1.3.0/24 (next hop 13.13.13.1) and 1.1.3.0/25 (next hop 13.13.13.1) are changed to 21.
- The routes 1.1.3.0/24 (next hop 34.34.34.2) and 1.1.3.0/25 (next hop 34.34.34.2) are matched against node 40 in route-policy RP. The route 1.1.3.0/25 matches Pref3, so the cost value of 1.1.3.0/25 (next hop 34.34.34.2) is changed to 11.
- Then, the route 1.1.3.0/24 (next hop 34.34.34.2) matches node 50 in route-policy RP.

Route-Policy Configuration (1)



Route-Policy Configuration (2)



- RTA can also use route policy to control routes:
  - acl 2000
    - rule 0 permit source 10.1.1.0 0.0.0.255
    - rule 5 permit source 10.1.3.0 0.0.0.255
  - route-policy huawei-control permit node 10
    - if-match acl 2000
  - ospf 1
    - import-route direct route-policy huawei-control

Route-Policy Configuration (3)

```
<RTC>dis ip routing-table
Route Flags: R - relay, D - download to fib
------------------------------------------------------------------------------
Routing Tables: Public
         Destinations : 14        Routes : 14

Destination/Mask    Proto   Pre  Cost      Flags NextHop         Interface

        10.1.3.0/24 O_ASE   150  1           D   12.1.2.1        GigabitEthernet 0/0/0
        10.1.4.0/24 Direct  0    0           D   10.1.4.2        GigabitEthernet 0/0/1
        10.1.5.0/24 OSPF    10   3           D   12.1.2.1        GigabitEthernet 0/0/0
```

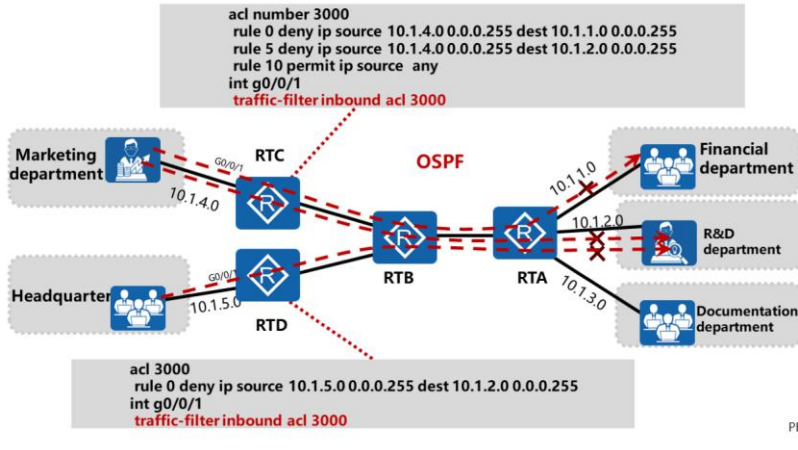
```
<RTD>dis ip routing-table
Route Flags: R - relay, D - download to fib
------------------------------------------------------------------------------
Routing Tables: Public
         Destinations : 15        Routes : 15

Destination/Mask    Proto   Pre  Cost      Flags NextHop         Interface

        10.1.1.0/24 O_ASE   150  1           D   12.1.3.1        GigabitEthernet 0/0/0
        10.1.3.0/24 O_ASE   150  1           D   12.1.3.1        GigabitEthernet 0/0/0
        10.1.4.0/24 OSPF    10   3           D   12.1.3.1        GigabitEthernet 0/0/0
        10.1.5.0/24 Direct  0    0           D   10.1.5.2        GigabitEthernet 0/0/1
```


- Note: The preceding information is a part of key information.

Contents

Solution 2: Traffic Filter(1)

- Based on customized policy: uses the traffic filter.



PBR-Traffic-Filter Tool

HUAWEI

Solution 2: Traffic Filter (2)

```
[RTC]dis ip routing-table
Route Flags: R - relay, D - download to fib
------------------------------------------------------------------------------
Routing Tables: Public
         Destinations : 16        Routes : 16

Destination/Mask    Proto   Pre  Cost      Flags NextHop         Interface

   10.1.1.0/24      O_ASE   150  1           D   12.1.2.1        GigabitEthernet 0/0/0
   10.1.2.0/24      O_ASE   150  1           D   12.1.2.1        GigabitEthernet 0/0/0
   10.1.3.0/24      O_ASE   150  1           D   12.1.2.1        GigabitEthernet 0/0/0
   10.1.4.0/24      Direct  0    0           D   10.1.4.2        GigabitEthernet 0/0/1
   10.1.5.0/24      OSPF    10   3           D   12.1.2.1        GigabitEthernet 0/0/0
```

```
PC-Marketing department>ping 10.1.1.1

Ping 10.1.1.1: 32 data bytes, Press Ctrl_C to break
Request timeout!
Request timeout!
Request timeout!
Request timeout!

--- 10.1.1.1 ping statistics ---
  4 packet(s) transmitted
  0 packet(s) received
  100.00% packet loss
```


HUAWEI

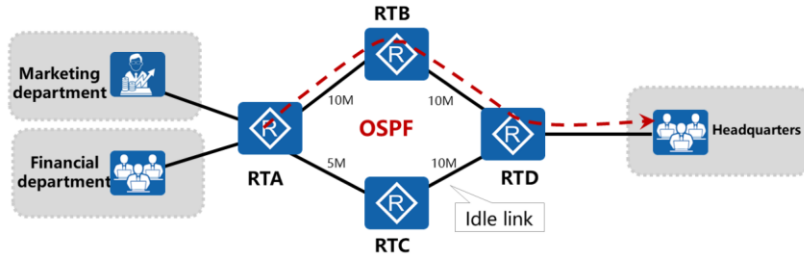
- According to the test result, after Traffic Filter is configured, the routing table of RTC contains the routes on the entire network, and the marketing department cannot access the financial and R&D departments but can access other departments. The routing table of RTD also contains the routes on the entire network. The headquarters cannot access the R&D department, but can access other departments.

Contents

HUAWEI

Adjust Network Traffic Path - Single Protocol

- In network optimization stage, network traffic paths may need to be adjusted.

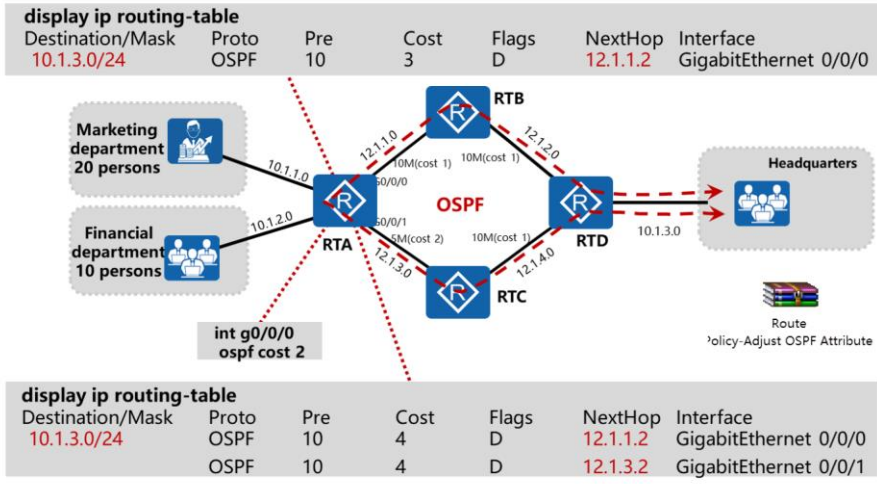


- Solution 1: uses route policy to change the protocol attribute to control routing table entries and adjust traffic path.
- Solution 2: uses policy-based routing to control traffic behavior before searching the routing table.

HUAWEI

- You can change protocol attributes to control the routing table entries, affecting traffic path:
  - If OSPF or IS-IS is run, adjust the cost values of interfaces.
  - If RIP is run, adjust the metric or next hop.
  - If BGP is run, adjust the AS-Path, Local_Pref, MED, and Community attributes.

Solution 1: Routing Policy

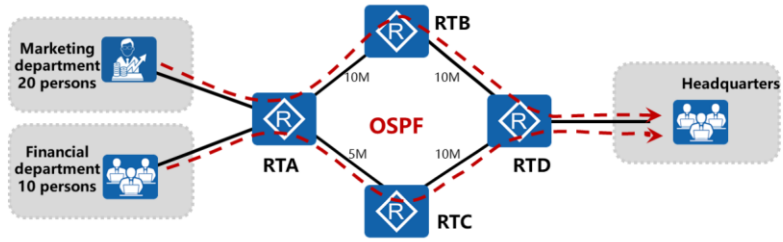


HUAWEI

- Traditionally, packets are forwarded by searching the routing table based on destination IP addresses. What are the disadvantages of this method? Can it meet complicated requirements? Can packet forwarding be controlled precisely?

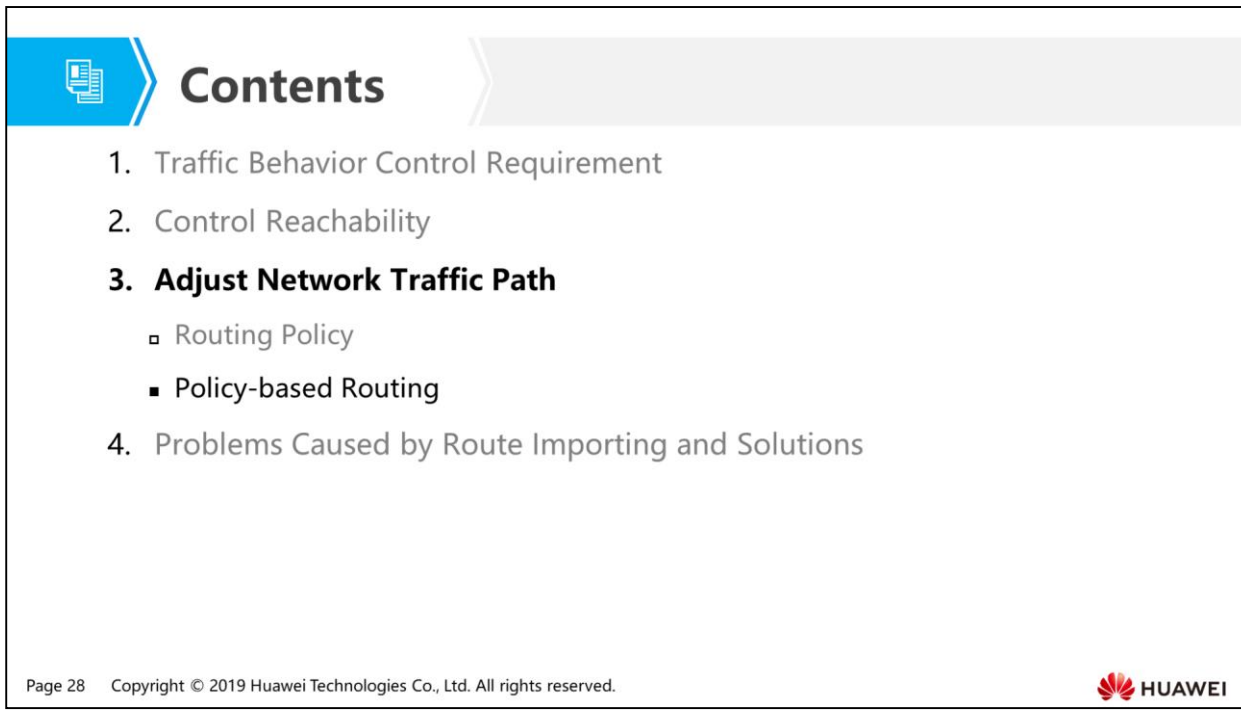
Limitation of Solution 1

- The traffic path from the marketing department to headquarters is RTA-RTB-RTD, and the traffic path from the financial department to headquarters is RTA-RTC-RTD.



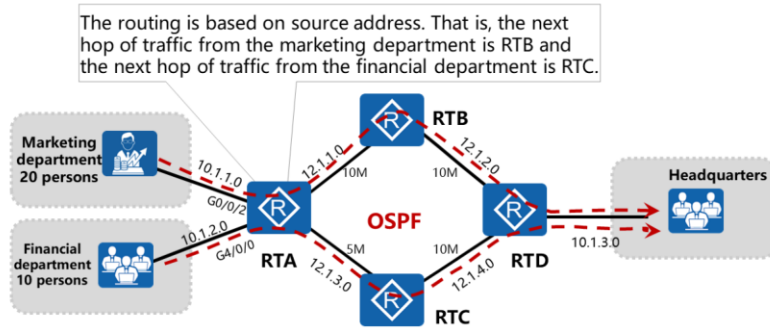
- As shown in the figure, solution 1 cannot meet the requirement because packets are forwarded based on destination address. It cannot meet the requirements of source address-based, destination address-based, or application layer-based forwarding

- Due to the limitation in route policy, users expect to customize packet forwarding policies based on the traditional routing method. Policy-based routing allows network administrators to make user-defined policies to change packet routes based on source addresses, packet size, and link quality in addition to destination addresses.



- Different from the routing mechanism that forwards packets by searching for the routes based on the destination addresses of IP packets, policy-based routing (PBR) is based on the user-defined routing policies. PBR has a higher priority than routing policy. Before forwarding a data packet, the router matches the packet against the PBR rule first. If matched, the packet is forwarded based on the policy. Otherwise, the packet is forwarded based on the routing table. The PBR does not modify the routing table. It affects data packet forwarding based on the pre-defined rules.

Solution 2: Policy-based Routing (1)



- Policy-based routing is implemented using traffic policy:
  - Use ACL to match traffic;
  - Define behavior for traffic, for example, change the next hop.

  HUAWEI

- PBR includes:
  - Local PBR: applies to packets sent by the device, such as ICMP and BGP packets.
    - To send the packets with different source addresses and lengths, configure local PBR.
    - It is implemented using the Policy-Based-Route tool.
  - Interface PBR: applies to the packets forwarded (not initiated) by the local device.
    - To forward packets to the specified next hop, configure the interface PBR. The packets matching the redirection rule are forwarded through the specified next hop interface. The packets not matching the redirection rule are forwarded based on routing table. Interface PBR applies to load balancing and monitoring.
    - It is implemented using the traffic policy.
  - Smart PBR: selects the optimal path for service traffic based on link quality.
    - To select paths for different services based on link quality, configure smart PBR.
    - It is implemented using the Smart-Policy-Route tool. This method is not described in this course.

Solution 2: Policy-based Routing (2)

```
[RTA]acl 3000
  rule 5 permit ip source 10.1.1.0 0.0.0.255 dest 10.1.3.0 0.0.0.255
traffic classifier huawei-control1
 if-match acl 3000
traffic behavior huawei-control1
 redirect ip-nexthop 12.1.1.2
traffic policy huawei-control1
 classifier huawei-control1 behavior huawei-control1
int g0/0/2
 traffic-policy huawei-control1 inbound
```

```
[RTA]acl 3001
 rule 5 permit ip source 10.1.2.0 0.0.0.255 dest 10.1.3.0 0.0.0.255
traffic classifier huawei-control2
 if-match acl 3001
traffic behavior huawei-control2
 redirect ip-nexthop 12.1.3.2
traffic policy huawei-control2
 classifier huawei-control2 behavior huawei-control2
int g4/0/0
 traffic-policy huawei-control2 inbound
```

Route
olicy-Adjust OSPF Attribute

HUAWEI

- MQC is used in this example.

Solution 2: Policy-based Routing (3)

```
<RTA>dis ip routing-table
Route Flags: R - relay, D - download to fib
------------------------------------------------------------------------------
Routing Tables: Public
         Destinations : 19        Routes : 20

Destination/Mask    Proto   Pre  Cost      Flags NextHop         Interface
  10.1.1.0/24       Direct  0    0         D     10.1.1.2        GigabitEthernet 0/0/2
  10.1.2.0/24       Direct  0    0         D     10.1.2.2        GigabitEthernet 4/0/0
  10.1.3.0/24       OSPF    10   3         D     12.1.1.2        GigabitEthernet 0/0/0
```

```
PC-Marketing department>tracert
10.1.3.1

traceroute to 10.1.3.1, 8 hops max
(ICMP), press Ctrl+C to stop
 1  10.1.1.2  47 ms  31 ms  15 ms
 2  12.1.1.2  47 ms  31 ms  32 ms
 3  12.1.2.2  93 ms  63 ms  46 ms
 4  *10.1.3.1  62 ms  31 ms
```

```
PC-Financial department>tracert 10.1.3.1

traceroute to 10.1.3.1, 8 hops max
(ICMP), press Ctrl+C to stop
 1  10.1.2.2  16 ms  31 ms  16 ms
 2  12.1.3.2  62 ms  47 ms  31 ms
 3  12.1.4.2  47 ms  47 ms  31 ms
 4  10.1.3.1  32 ms  46 ms  32 ms
```

Differences Between the Routing Policy and PBR

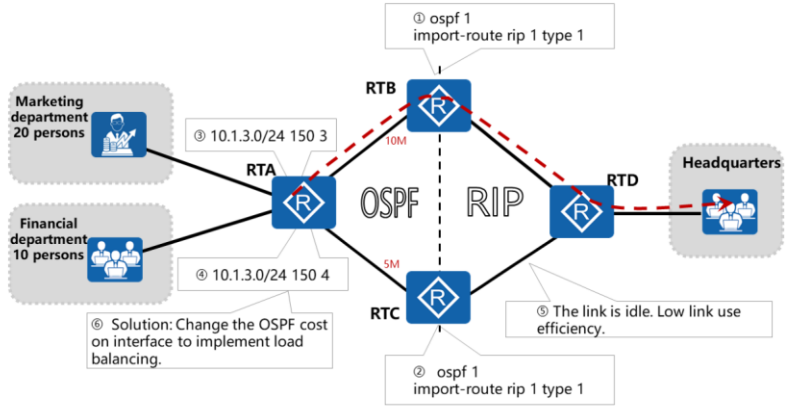
Routing Policy	Policy-based Routing
Based on the control plane, affect routing entries.	Based on the forwarding plane, do not affect routing entries. Packets are forwarded based on policy first, and then based on routing table if policy-based forwarding fails.
Policy based on destination address.	Policy based on source address, destination address, protocol type, and packet size.
Used with routing protocol.	A routing policy needs to be manually configured hop by hop to ensure that packets are forwarded according to the policy.
Tools: Route-Policy, Filter-Policy, etc.	Tools: Traffic-Filter, Traffic-Policy, Policy-Based-Route, etc.

- A router has a routing table and a forwarding table. The forwarding table is mapped from the routing table. PBR applies to the forwarding table, and routing policy applies to the routing table. Forwarding is performed at the bottom layer, while routing is performed at the upper layer. Therefore, the forwarding based on forwarding table has a higher priority than the forwarding based on routing table.
- A route policy takes effect in route discovery. It uses certain rules to influence route advertisement, receiving, and selection, thus modifying the routing entries. The PBR takes effect in packet forwarding. It uses certain rules to influence packet forwarding, but does not affect the routing table.

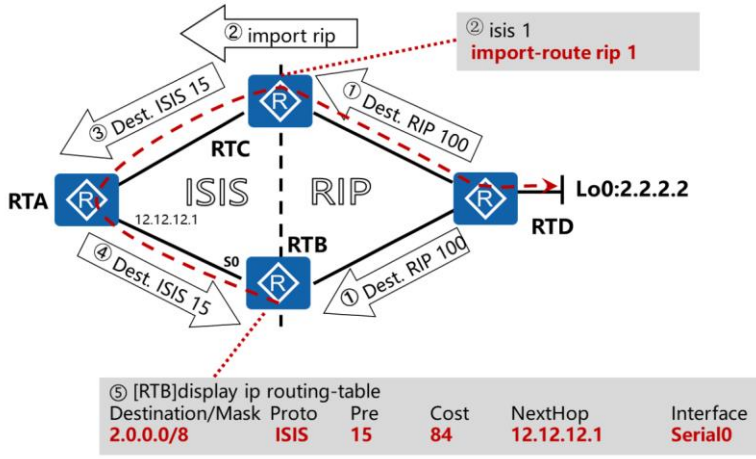
Contents

Adjust Network Traffic Path - Multiple Protocols

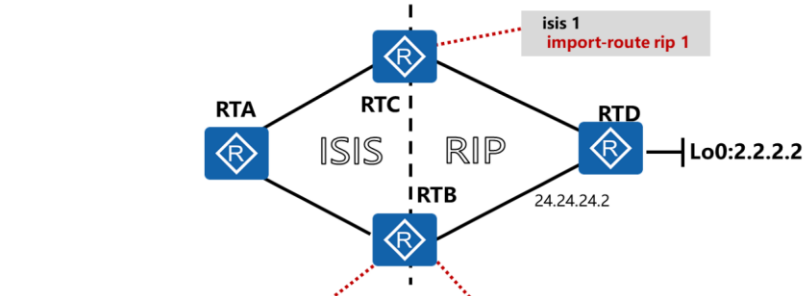
- In the preceding example, four routers run the same protocol. If they run different protocols, what will happen?



Other Problems Caused by Multi-Protocol - Sub-Optimal Route



Solution 1: Filter Routes

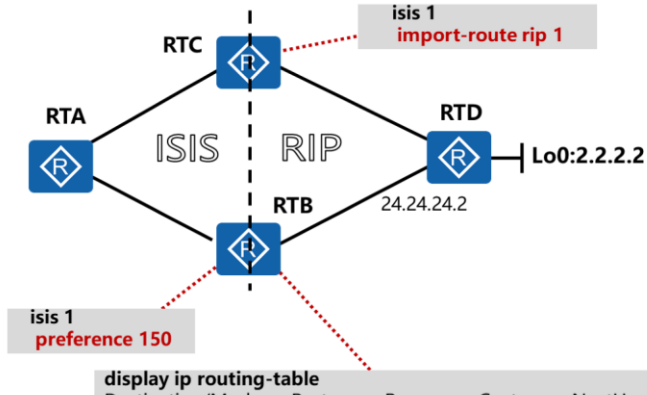



```
acl 2001
 rule 5 deny source 2.0.0.0 0.255.255.255
 rule 10 permit
isis 1
 filter-policy 2001 import
```

display ip routing-table

Destination/Mask	Proto	Pre	Cost	NextHop	Interface
2.0.0.0/8	RIP	100	1	24.24.24.2	Serial1

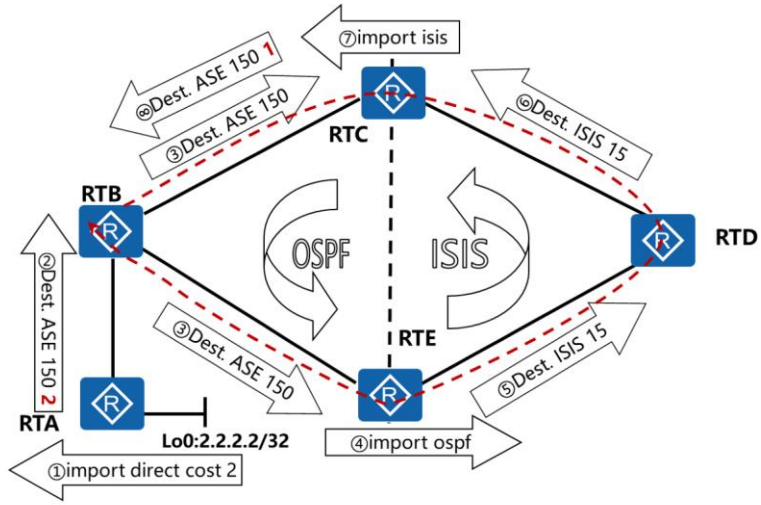
Solution 2: Adjust Protocol Priorities



display ip routing-table

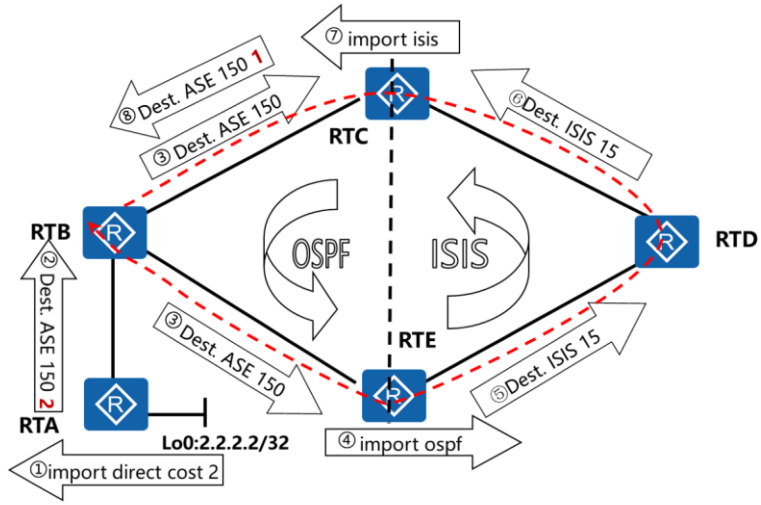
Destination/Mask	Proto	Pre	Cost	NextHop	Interface
2.0.0.0/8	RIP	100	1	24.24.24.2	Serial1

Other Problems Caused by Multi-Protocol - Loop



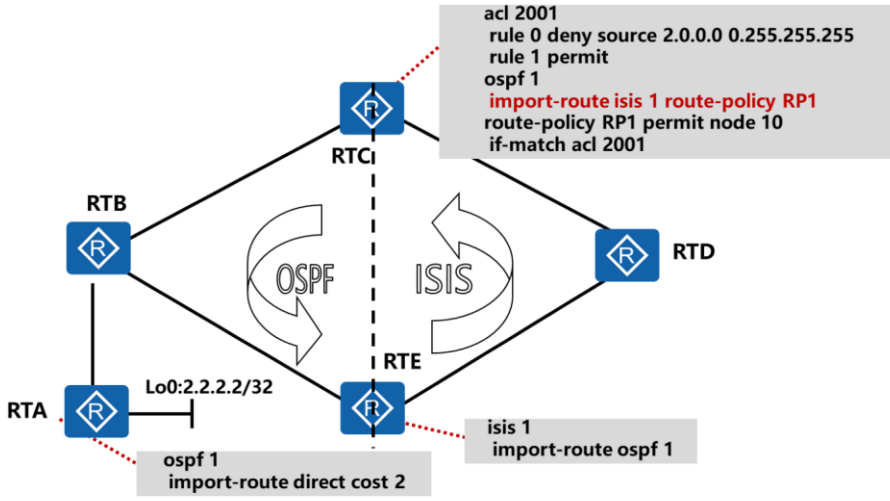
HUAWEI

Other Problems Caused by Multi-Protocol - Loop

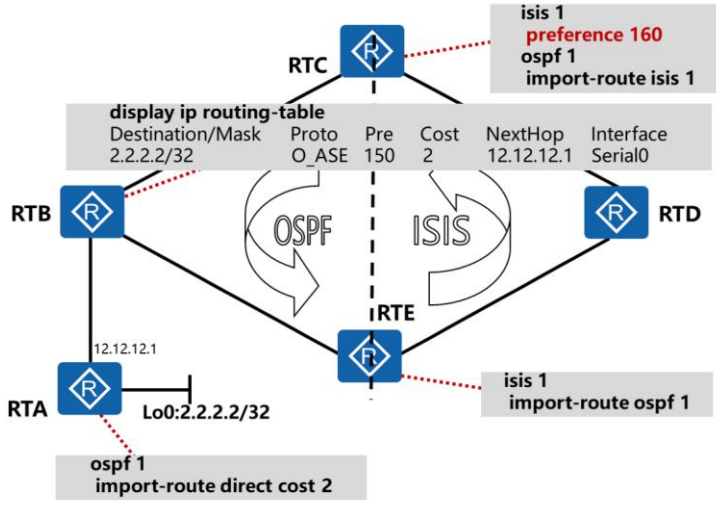


HUAWEI

Solution 1: Filter Routes



Solution 2: Adjust Protocol Priorities



HUAWEI

Quiz

1. Can the IP prefix list be used to filter IP packets?
2. Which methods can be used to adjust traffic paths?
3. What problems may be caused by route importing and what are the solutions?

HUAWEI

- Answer: The IP prefix list can filter only routing information, but cannot filter IP packets.
- Answer: Routing policy and policy-based routing.
- Answer: The problems such as sub-optimal path and route loop may occur. The solutions include route filtering and priority adjustment.


Thank You
www.huawei.com

Eth-Trunk Principles and Configurations

Foreword

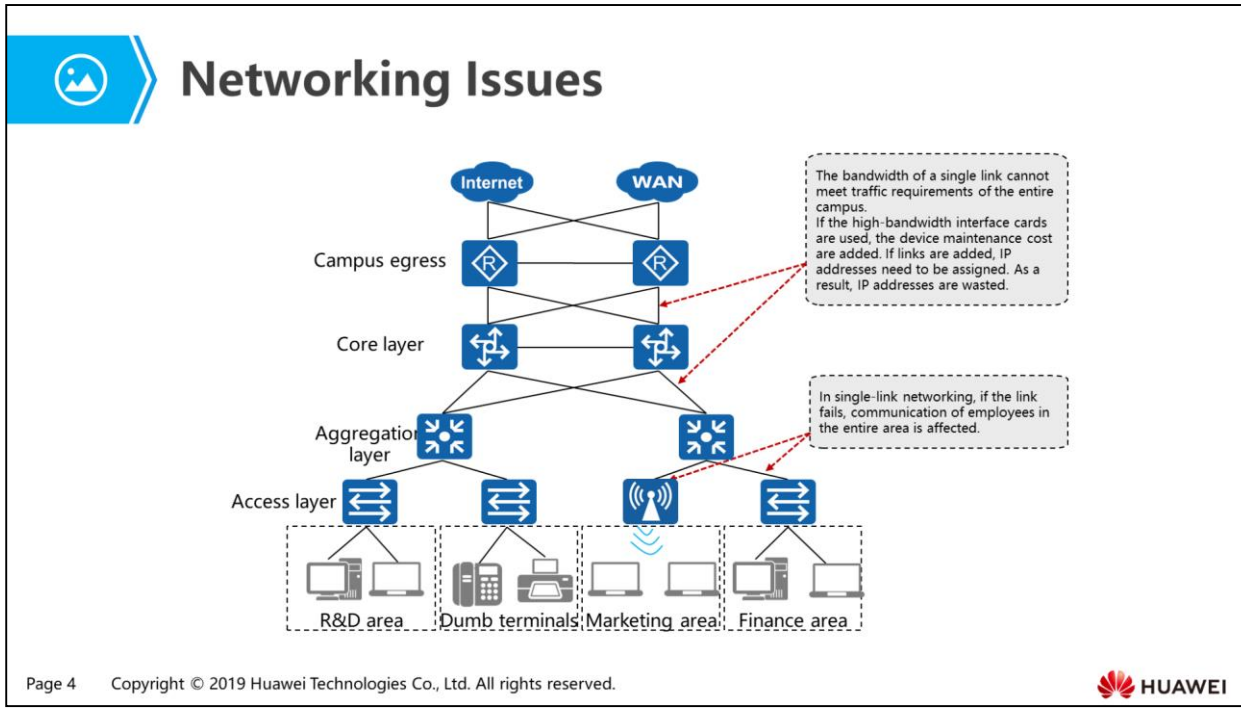
- As the number of services deployed on networks increases, the bandwidth of full-duplex P2P links cannot meet requirements of normal service traffic. The interface cards with higher bandwidth can be used to replace existing interface cards to increase bandwidth; however, this will waste existing device resources and increase upgrade expenditure. If more links are used to interconnect devices, each Layer 3 interface must be configured with an IP address, wasting IP addresses.
- Eth-Trunk (link aggregation) is a binding technology that can bundle multiple independent physical interfaces into a logical interface with high bandwidth. There is no need to replace interface cards and IP addresses are not wasted. This course describes Eth-Trunk technology.

Objectives

- Upon completion of this section, you will be able to:
  - Be familiar with the Eth-Trunk principle
  - Master the Eth-Trunk configuration

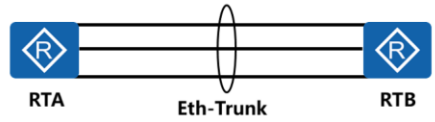
Contents

HUAWEI



- Networking issues:
  - As the volume of services deployed on networks increases, the bandwidth of a single physical link cannot meet the requirements of service traffic. The interface cards with higher bandwidth can be used to replace existing interface cards to increase bandwidth; however, this will waste existing device resources and increase upgrade expenditure. If more links are used to interconnect devices, each Layer 3 interface must be configured with an IP address, wasting IP addresses.
  - Single-link networking does not involve redundancy design. Faults on uplinks of access devices affect communication of the area connected to access devices.
- Multiple independent physical interfaces can be bundled to form a high-bandwidth logical interface, that is, link aggregation is used. There is no need to replace interface cards and IP addresses are not wasted. Eth-Trunk technology bundles multiple physical interfaces to form a logical interface, which is called an Eth-Trunk.
- Only Ethernet interfaces can join the Eth-Trunk. Let's introduce Eth-Trunk technology on LANs.

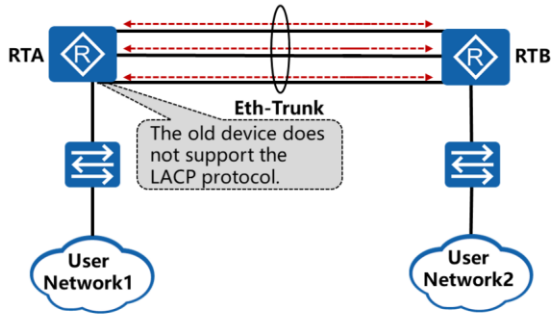
Eth-Trunk Concepts



- Eth-Trunk is a binding technology that bundles multiple Ethernet interfaces into a logical interface.
- Eth-Trunk modes:
  - Manual load balancing mode
  - LACP mode

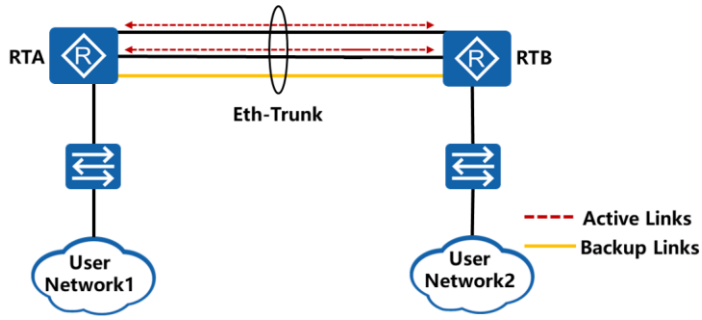
- The Eth-Trunk can increase bandwidth, implement load balancing, and improve network reliability depending on the link aggregation mode.
- Eth-Trunk can be used for Layer 2 and Layer 3 link aggregation. By default, an Ethernet interface works in layer 2 mode. To configure a Layer 2 Eth-Trunk, run the portswitch command to switch the Ethernet interface to Layer 2 interface. To configure a Layer 3 Eth-Trunk, run the undo portswitch command to switch the Ethernet interface to Layer 3 interface.

Manual Load Balancing Mode



- When at least one device of two devices does not support the LACP protocol, you can configure the Eth-Trunk in manual load balancing mode to increase the bandwidth between the two devices and improve reliability.
- In manual load balancing mode, all links added to the Eth-Trunk forward data.

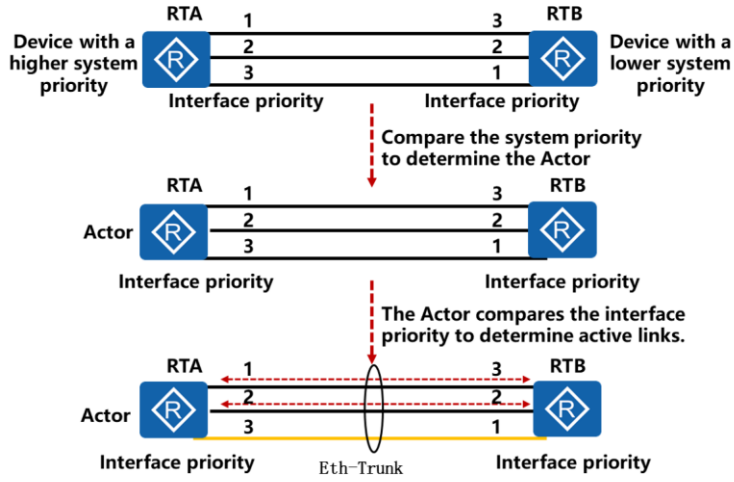
LACP Mode



- The LACP mode is also called the M:N mode, where M links are active and forward data and N links are inactive and used as backup links.
- In the preceding figure, the number of active links is set to 2, that is, two links are in forwarding state. One link is in inactive state and does not forward data. When an active link fails, the backup link starts to forward data.

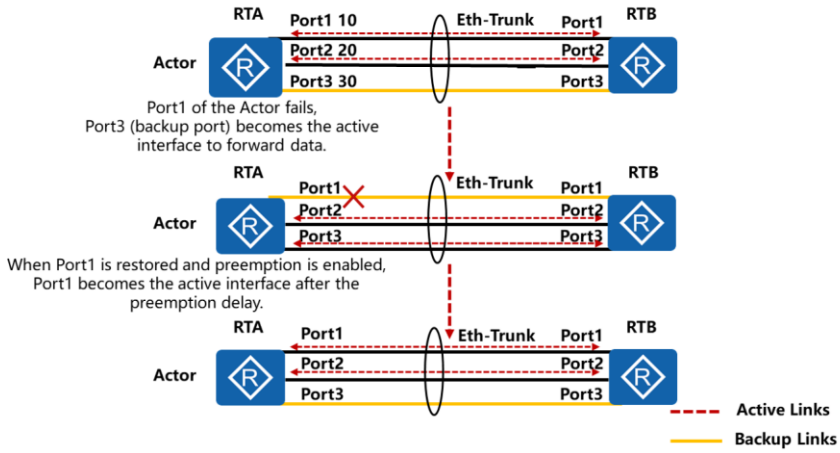
- M:N backup of member interfaces
  - In the preceding figure, (M+N) links exist between two devices. Here, M is 2 and N is 1. When traffic is forwarded on the aggregated link, traffic is load balanced on M links and the other N link does not forward traffic. The actual bandwidth of the aggregated link is the sum of the bandwidth of M links. The maximum bandwidth of the aggregated link is the sum of the bandwidth of (M+N) links.
  - When one link among M links fails, LACP selects one link with the high priority from N backup links to replace the faulty link. The actual bandwidth of the aggregated link is still the sum of the bandwidth of M links, but the maximum bandwidth of the aggregated link is the sum of the bandwidth of (M+N-1) links.

Selecting Active Links in LACP Mode



- As shown in the figure, there are three links between devices. The maximum number of active links is 2, that is, two links are in forwarding state and one link is in backup state.
- After member interfaces are added to the Eth-Trunk in LACP mode, the interfaces send LACPDUs to notify the remote end of the system priority, MAC addresses, interface priorities, and interface numbers. After receiving the information, the remote end compares the information with the information stored on its interfaces to select the interfaces that can be aggregated. The devices at both ends determine active interfaces and links mutually.
  - The end with a higher LACP system priority functions as the Actor. If the two ends have the same LACP system priority, the end with a smaller MAC address functions as the Actor.
    - A smaller LACP system priority value indicates a higher priority. By default, the value of the LACP system priority is 32768.
  - A smaller LACP priority value of an interface indicates a higher priority of the interface. If interfaces have the same LACP priority, the interface with a smaller ID or interface number is selected as the active interface.
    - LACP interface priorities are used to prioritize interfaces of an Eth-Trunk. Interfaces with higher priorities are selected as active interfaces.

LACP Preemption



- LACP preemption delay:
  - After LACP preemption occurs, the backup link waits for a period of time to switch to the active status. The period of time is called LACP preemption delay. The LACP preemption delay prevents instable data transmission on an Eth-Trunk link due to frequent status changes of some links.
  - As shown in the preceding figure, Port1 becomes inactive due to a link fault. Then the link is restored. If LACP preemption is enabled and the LACP preemption delay is set, Port1 switches to active after the LACP preemption delay.
- Preemption is enabled in the following situations:
  - Port1 fails, and then is restored. When Port1 fails, Port3 replaces Port1 to transmit services. If LACP preemption is not enabled on the Eth-Trunk, Port1 that is restored still retains in backup state. If LACP preemption is enabled on the Eth-Trunk, Port1 becomes active and Port3 becomes the backup interface.
  - If LACP preemption is enabled and Port3 needs to replace Port1 or Port2 to become the active interface, set the highest LACP priority value for Port3. If LACP preemption is not enabled, the system does not re-select the active interface or switch the active interface when the priority of a backup interface is higher than that of the active interface.

Eth-Trunk Load Balancing

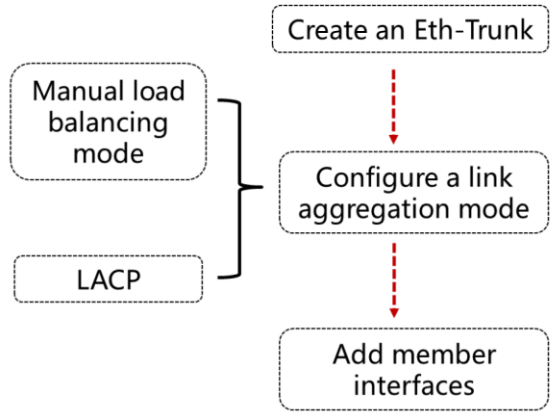
- When packets are load balanced on an Eth-Trunk, you can use flow-based or packet-based load balancing and set the weight of load balancing.
- On an Eth-Trunk, the larger the weight of a member link, the heavier the load over the member link.

Load Balancing Mode	Description
Flow-based load balancing	When packets use the same source and destination IP addresses or source and destination MAC addresses, the packets are transmitted over the same member link.
Packet-based load balancing	Packets are transmitted over different member links.

- Configure a load balancing mode.
  - Run the system-view command to enter the system view.
  - Run the interface eth-trunk trunk-id command to enter the Eth-Trunk interface view.
  - Run the load-balance { ip | packet-all } command to configure a load balancing mode of the Eth-Trunk.
  - By default, packets are load balanced based on IP addresses of packets on an Eth-Trunk.
  - Note:
    - IP-based hash algorithm can ensure packet sequencing but cannot ensure bandwidth usage.
    - Packet-based hash algorithm can ensure bandwidth usage but cannot ensure packet sequencing.

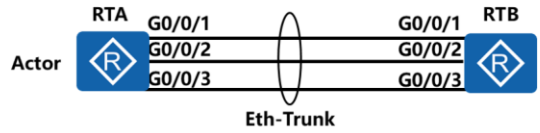
- Configure the load balancing weight.
  - Run the system-view command to enter the system view.
  - Run the interface interface-type interface-number command to enter the Ethernet interface view.
  - Run the distribute-weight weight-value command to configure the load balancing weight of the Eth-Trunk member interface.
  - By default, the load balancing weight of an Eth-Trunk member interface is 1.

Eth-Trunk Configuration Process



- When adding an interface to an Eth-Trunk, pay attention to the following points:
  - Member interfaces cannot have Layer 3 configurations such as IP addresses nor services.
  - Member interfaces cannot be configured with static MAC addresses.
  - An Eth-Trunk cannot be nestled, that is, a member interface cannot be an Eth-Trunk.
  - An Ethernet interface can join only one Eth-Trunk. Before adding an Ethernet interface to an Eth-Trunk, ensure that the Ethernet interface does not belong to any other Eth-Trunk.
  - If the local device uses an Eth-Trunk, the remote interface directly connected to the local interface must join the Eth-Trunk so that the two ends can communicate.
  - An Eth-Trunk can work in Layer 2 or Layer 3 mode. The Eth-Trunk working mode does not affect addition of member links, that is, Ethernet interfaces can be added to the Eth-Trunk in Layer 2 or Layer 3 mode.

Configuring the Manual Load Balancing Mode



- Procedure for configuring the manual load balancing mode:
  - Create an Eth-Trunk.
  - Configure the working mode of the Eth-Trunk.
  - Add member interfaces to the Eth-Trunk.

  HUAWEI

- Create an Eth-Trunk in manual load balancing mode.
  - Run the system-view command to enter the system view.
  - Run the interface eth-trunk trunk-id command to create an Eth-Trunk and enter the Eth-Trunk interface view.
  - (Optional) Run the portswitch command to switch the Eth-Trunk to Layer 2.
- Configure the working mode of the Eth-Trunk.
  - Run the mode manual load-balance command to configure the Eth-Trunk to work in manual load balancing mode.
    - By default, an Eth-Trunk works in manual load balancing mode.
  - Add member interfaces to the Eth-Trunk.
    - In the Eth-Trunk interface view:
    - Run the interface eth-trunk trunk-id command to enter the Eth-Trunk interface view.

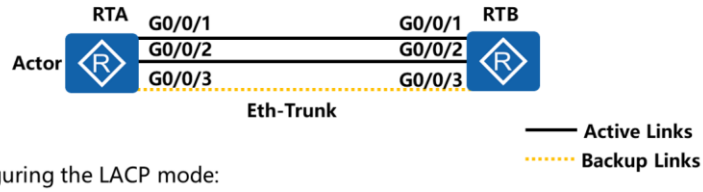
- Perform any of the following steps.

  Run the trunkport interface-type { interface-number1 [to interface-number2] } &<1-16> command to add member interfaces in a batch.

  Run the trunkport interface-type interface-number command to add a member interface.

- In the member interface view:
  - Run the interface { ethernet | gigabitethernet } interface-number command to enter the Eth-Trunk member interface view.
  - Run the eth-trunk trunk-id command to add the member interface to the Eth-Trunk.

Configuring the LACP Mode



- Procedure for configuring the LACP mode:
  - Create an Eth-Trunk.
  - Configure the working mode of the Eth-Trunk.
  - Add member interfaces to the Eth-Trunk.
  - (Optional) Configure the LACP system priority.
  - (Optional) Set the upper threshold for the number of active interfaces.
  - (Optional) Configure the LACP interface priority.
  - (Optional) Enable LACP preemption and set the preemption delay.

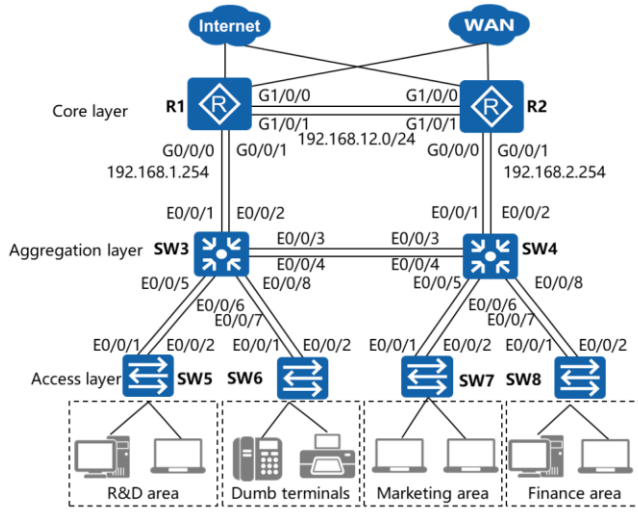
  HUAWEI

- Create an Eth-Trunk in LACP mode.
  - Run the system-view command to enter the system view.
  - Run the interface eth-trunk trunk-id command to create an Eth-Trunk.
  - (Optional) Run the portswitch command to switch the Eth-Trunk to Layer 2.
- Configure the working mode of the Eth-Trunk.
  - Run the interface eth-trunk trunk-id command to enter the Eth-Trunk interface view.
  - Run the mode lacp-static command to configure the Eth-Trunk to work in LACP mode.
- Add member interfaces to the Eth-Trunk.
  - In the Eth-Trunk interface view:
  - Run the interface eth-trunk trunk-id command to enter the Eth-Trunk interface view.
  - Perform any of the following steps.
    - Run the trunkport interface-type { interface-number1 [to interface-number2] } &<1-16> command to add member interfaces in a batch.
    - Run the trunkport interface-type interface-number command to add a member interface.

    - In the member interface view:
        - Run the interface { ethernet | gigabitethernet } interface-number command to enter the Eth-Trunk member interface view.
        - Run the eth-trunk trunk-id command to add the member interface to the Eth-Trunk.
- (Optional) Configure the LACP system priority.
    - Run the lacp priority priority command to configure the LACP system priority.
- (Optional) Set the upper threshold for the number of active interfaces.
    - Run the interface eth-trunk trunk-id command to enter the Eth-Trunk interface view.
    - Run the max active-linknumber link-number command to set the upper threshold for the number of active interfaces.
- (Optional) Configure the LACP interface priority.
    - Run the interface interface-type interface-number command to enter the interface view.
    - Run the lacp priority priority command to configure the LACP interface priority.
- (Optional) Enable LACP preemption and set the preemption delay.
    - Run the interface eth-trunk trunk-id command to enter the Eth-Trunk interface view.
    - Run the lacp preempt enable command to enable LACP preemption.
    - Run the lacp preempt delay delay-time command to set the preemption delay.

Contents

Eth-Trunk Configuration Requirements



HUAWEI

- The preceding figure shows a campus network topology. To improve network reliability, link aggregation technology needs to be used between devices at different layers. Core devices need to be configured with IP addresses and function as intranet gateways, and aggregation devices communicate with access devices at Layer 2.

Configuration of Core Devices

- The configuration of core router R1 is used as an example.
  - Create an Eth-Trunk and configure an IP address for it.

```
interface Eth-Trunk1
 Undo portswitch //Switch the interface to a Layer 3 interface.
 Description "Core-R1 to Aggregate-SW3 "// The description helps the administrator
understand the remote device connected to the interface.
 ip address 192.168.1.254 255.255.255.0
#
interface Eth-Trunk2
 undo portswitch
 description "Core-R1 to Core-R2"
 ip address 192.168.12.1 255.255.255.0
```

  - Add physical interfaces to the Eth-Trunk.

```
interface GigabitEthernet0/0/0
 eth-trunk 1
interface GigabitEthernet0/0/1
 eth-trunk 1
#
interface GigabitEthernet1/0/0
 eth-trunk 2
interface GigabitEthernet1/0/1
 eth-trunk 2
```

Configuration of Aggregation Devices (1)

- The configuration of the aggregation switch SW3 is used as an example.
  - Create an Eth-Trunk. Because aggregation devices use Layer 2 interconnection, the IP address does not need to be configured.

```
interface Eth-Trunk1
 description   "Aggregate-SW3 to Core-R1  "
//The description helps the administrator understand the remote
device connected to the interface.
#
interface Eth-Trunk2
 description   "Aggregate-SW3 to Aggregate-SW4  "
#
interface Eth-Trunk3
 description "Aggregate-SW3 to Access-SW5  "
#
interface Eth-Trunk4
 description "Aggregate-SW3 to Access-SW6  "
```

Configuration of Aggregation Devices (2)

- The configuration of the aggregation switch SW3 is used as an example.
  - Add physical interfaces to the Eth-Trunk.

```
interface Ethernet0/0/1
 eth-trunk 1
interface Ethernet0/0/2
 eth-trunk 1
#
interface Ethernet0/0/3
 eth-trunk 2
interface Ethernet0/0/4
 eth-trunk 2
#
interface Ethernet0/0/5
 eth-trunk 3
interface Ethernet0/0/6
 eth-trunk 3
#
interface Ethernet0/0/7
 eth-trunk 4
interface Ethernet0/0/8
 eth-trunk 4
```

Configuration of Access Devices (1)

- The configuration of the access switch SW5 is used as an example.
  - Create an Eth-Trunk. Because access devices use Layer 2 interconnection, the IP address does not need to be configured.

```
interface Eth-Trunk1
 description "Access-SW5 to Aggregate-SW3"
//The description helps the administrator understand
the remote device connected to the interface.
```

  - Add physical interfaces to the Eth-Trunk.

```
interface Ethernet0/0/1
 eth-trunk 1
interface Ethernet0/0/2
 eth-trunk 1
```

Configuration of Access Devices (2)

- After the preceding configuration is complete, run the following command to check the Eth-Trunk configuration:

```
display eth-trunk
Eth-Trunk1's state information is:
WorkingMode: NORMAL        Hash arithmetic: According to SIP-XOR-DIP
Least Active-linknumber: 1  Max Bandwidth-affected-linknumber: 8
Operate status: up         Number Of Up Port In Trunk: 2
--------------------------------------------------------------------------------
PortName                  Status     Weight
Ethernet0/0/1             Up         1
Ethernet0/0/2             Up         1
```

- Run the **display interface Eth-Trunk** command to check the detailed Eth-Trunk configuration.

Quiz

1. What are Eth-Trunk link aggregation modes?
   A. Manual load balancing mode
   B. LACP mode
   C. Manual LACP mode
   D. Dynamic LACP mode
2. In LACP mode, what is the default system priority?
   A. 1
   B. 4096
   C. 32768
   D. 65535

- Answer: AB.
- Answer: C.


Thank You
www.huawei.com

Advanced Features of Switches

Foreword

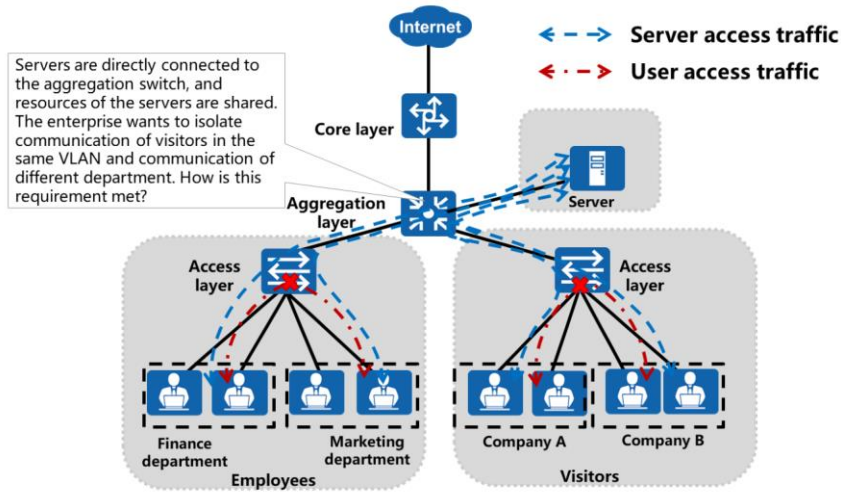
- Multiplex VLAN (MUX VLAN) provides a mechanism to control network resources through VLANs. MUX VLAN provides Layer 2 isolation to allow enterprise employees to communicate and isolate visitors.
- Layer 2 isolation can be implemented by adding different interfaces to different VLANs, but VLAN resources are wasted. Port isolation can isolate ports in the same VLAN, providing a secure and flexible network solution.
- On a network demanding high security, port security can be enabled on the switch to prevent devices with invalid MAC addresses from connecting to the network. When the number of learned MAC addresses reaches the maximum, the switch does not learn new MAC addresses. The switch allows only the devices of which MAC addresses are learned to communicate.

Objectives

- Upon completion of this section, you will be able to:
  - Master the application scenario and configuration of MUX VLAN
  - Master the application scenario and configuration of port isolation
  - Master the application scenario and configuration of port security

Contents

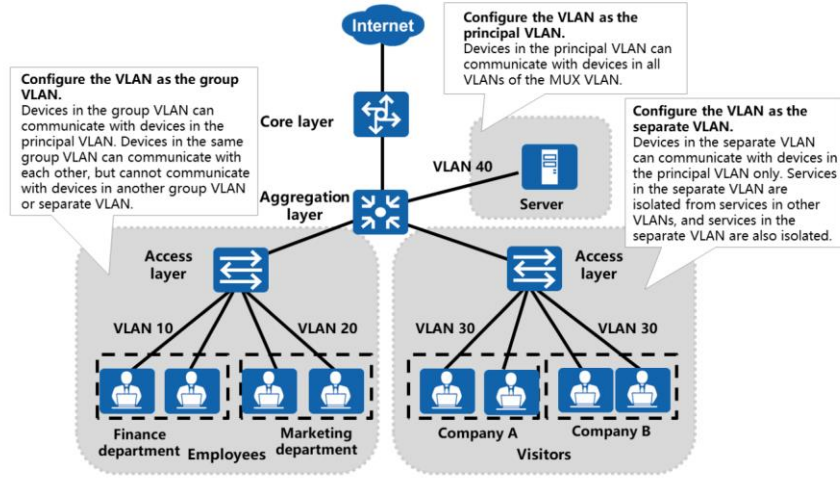
Application Scenario of MUX VLAN



HUAWEI

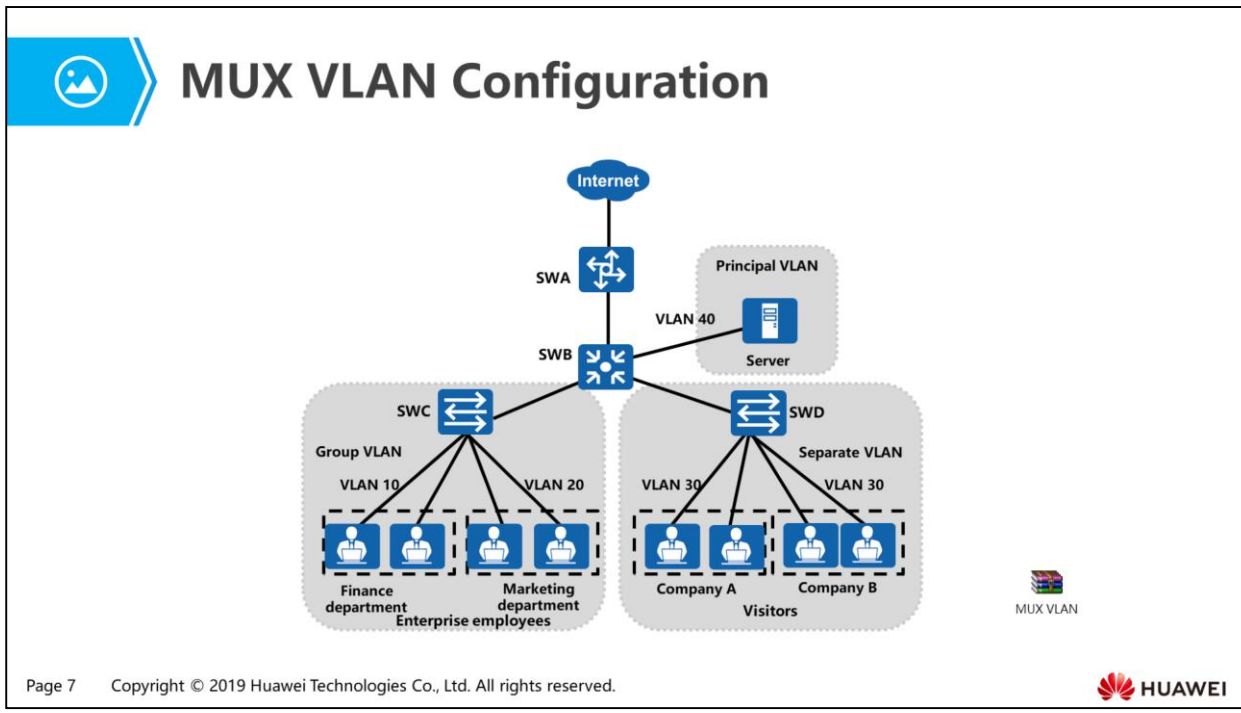
- As shown in the preceding figure, servers are directly connected to the aggregation switch. To allow all users to access enterprise servers, configure inter-VLAN communication.
- For the enterprise, enterprise employees need to communicate and visitors need to be isolated. You can configure a different VLAN for each of the visitors. If the enterprise has many visitors, many VLANs are used and it is more difficult to maintain the network.
- MUX VLAN provides Layer 2 isolation to allow enterprise employees to communicate and isolate visitors.

Basic Concepts of MUX VLAN



- Classification of MUX VLAN:
  - Principal VLAN: Devices in the principal VLAN can communicate with devices in all VLANs of the MUX VLAN.
  - Separate VLAN: Devices in the separate VLAN can communicate with devices in the principal VLAN only. Services in the separate VLAN are isolated from services in other VLANs, and services in the separate VLAN are also isolated.
  - Group VLAN: Devices in the group VLAN can communicate with devices in the principal VLAN. Devices in the same group VLAN can communicate, but cannot communicate with devices in another group VLAN or separate VLAN.
- Based on features of the MUX VLAN, the solution is as follows:
  - The enterprise administrator can assign servers to the principal VLAN.
  - For MUX VLAN technology, only one VLAN can be configured as the separate VLAN, so visitors can be assigned to the separate VLAN.
  - Because multiple VLANs can be configured as group VLANs, employees can be assigned to group VLANs. Departments can be assigned to different VLANs for isolation.

- After the configuration is complete, the following requirements can be met:
  - Visitors and employees can access enterprise servers.
  - Employees in a department can communicate with each other, and employees in different departments cannot communicate with each other.
  - Communication between visitors is blocked, as well as the communication between visitors and employees.



- As shown in the preceding figure, visitors and employees need to access enterprise servers, employees in the same department are allowed to communicate and employees in different departments are not allowed to communicate; visitors are not allowed to communicate; visitors and employees are not allowed to communicate.
  - The enterprise servers are assigned to the principal VLAN. VLAN 40 is the principal VLAN.
  - Visitors are assigned to the separate VLAN and VLAN 30 is the separate VLAN.
  - Employees are assigned to group VLANs. VLAN 10 and VLAN 20 are group VLANs. VLAN 10 is allocated to the finance department and VLAN 20 is allocated to the marketing department. Departments are isolated at Layer 2.
- SWB configuration:

```
sysname SWB
#
vlan batch 10 20 30 40
#
vlan 10
 description Financial VLAN
```

```
vlan 20
 description Marketing VLAN
vlan 30
 description Client VLAN
vlan 40
 description Principal VLAN
 mux-vlan                          //Configure VLAN 40 as the principal VLAN.
 subordinate separate 30           //Configure VLAN 30 as the separate VLAN.
 subordinate group 10 20           //Configure VLAN 10 and VLAN 20 as group
VLANs.
#
interface GigabitEthernet0/0/1
 port link-type trunk
 port trunk allow-pass vlan 10 20 30 40
#
interface GigabitEthernet0/0/2
 port link-type trunk
 port trunk allow-pass vlan 10 20 30 40
#
interface GigabitEthernet0/0/3
 port link-type access
 port default vlan 40
 port mux-vlan enable              //Enable MUX VLAN on the interface.
```

- SWC configuration:

```
sysname SWC
#
vlan batch 10 20 30 40
#
vlan 10
 description Financial VLAN
```

```
vlan 20
 description Marketing VLAN
vlan 30
 description Cilent VLAN
vlan 40
 description Principal VLAN
 mux-vlan
 subordinate separate 30
 subordinate group 10 20
#
interface GigabitEthernet0/0/1
 port link-type trunk
 port trunk allow-pass vlan 10 20 30 40
#
interface GigabitEthernet0/0/2
 port link-type access
 port default vlan 10
 port mux-vlan enable
#
interface GigabitEthernet0/0/3
 port link-type access
 port default vlan 10
 port mux-vlan enable
#
interface GigabitEthernet0/0/4
 port link-type access
 port default vlan 20
 port mux-vlan enable
#
interface GigabitEthernet0/0/5
```

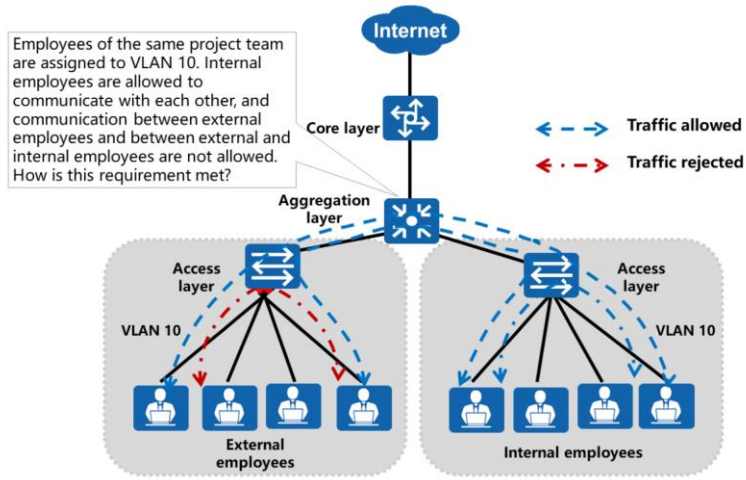
```
 port link-type access
  port default vlan 20
  port mux-vlan enable
```

- SWD configuration:

```
 sysname SWD
 #
 vlan batch 10 20 30 40
 #
 vlan 10
  description Financial VLAN
 vlan 20
  description Marketing
 vlan 30
  description Client VLAN
 vlan 40
  description Principal VLAN
  mux-vlan
  subordinate separate 30
  subordinate group 10 20
 #
 interface GigabitEthernet0/0/1
  port link-type trunk
  port trunk allow-pass vlan 10 20 30 40
 #
 interface GigabitEthernet0/0/2
  port link-type access
  port default vlan 30
  port mux-vlan enable
 #
 interface GigabitEthernet0/0/3
  port link-type access
  port default vlan 30
  port mux-vlan enable
```

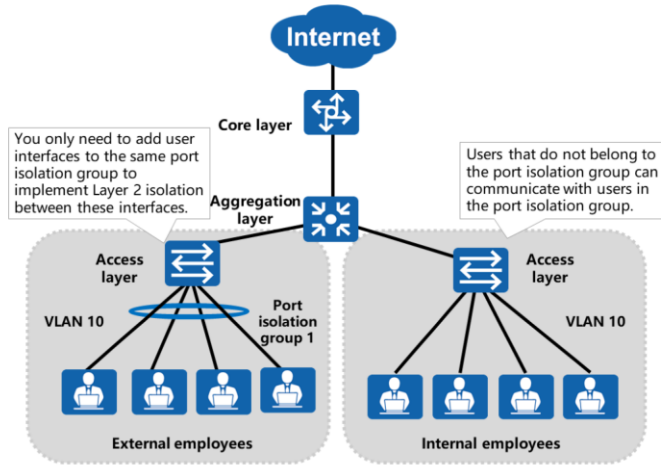
Contents

Application Scenario of Port Isolation



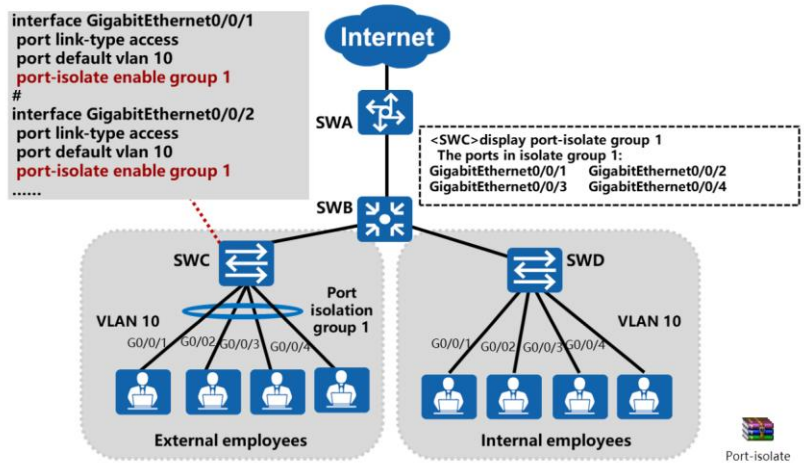
- Layer 2 isolation can be implemented by adding different interfaces to different VLANs, but VLAN resources are wasted. Port isolation can also isolate interfaces in the same VLAN. That is, you can add interfaces to a port isolation group to implement Layer 2 isolation between these interfaces. Port isolation provides secure and flexible networking schemes for users.

Basic Concepts of Port Isolation



- As shown in the preceding figure, users in the same port isolation group cannot communicate at Layer 2, but users in different port isolation groups can communicate. Users that do not belong to any port isolation group can communicate with users in the port isolation group.
- Port isolation modes:
  - To isolate broadcast frames in the same VLAN but allow users connecting to different interfaces to communicate at Layer 3, you can set the port isolation mode to Layer 2 isolation and Layer 3 interworking.
  - To prevent interfaces in the same VLAN from communicating at both Layer 2 and Layer 3, you can set the port isolation mode to Layer 2 and Layer 3 isolation.
- Configuration notes:
  - S series switches support Layer 2 isolation and Layer 3 interworking.
  - S series modular switches support Layer 2 and Layer 3 isolation. For S series fixed switches, only the S2700SI and S2700EI of V100R006C05 support Layer 2 and Layer 3 isolation, and the S1720, S2720, S2750EI, S5700LI, and S5700S-LI of V100R002 and later versions do not support Layer 2 and Layer 3 isolation.
  - If there is no special requirement, do not add the uplink and downlink interfaces to the same port isolation group.

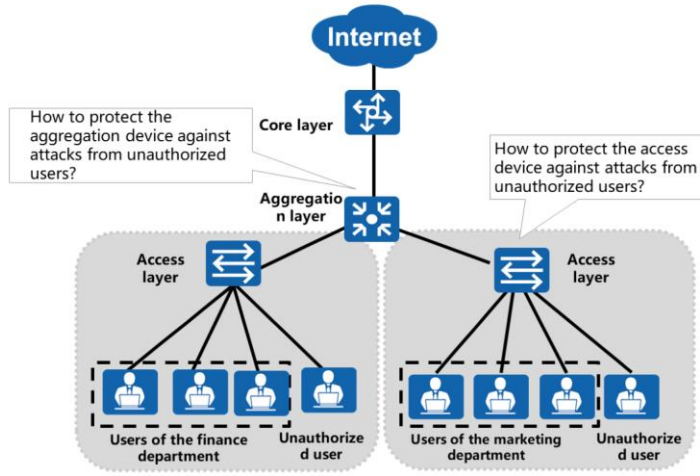
Port Isolation Configuration



- As shown in the preceding figure, employees of a same project team are assigned to VLAN 10, internal employees are allowed to communicate with each other, external employees are not allowed to communicate with each other, and internal and external employees are allowed to communicate.
- Configuration commands:
  - The port-isolate enable command enables port isolation. By default, a port is added to port isolation group 1.
  - To create a new port isolation group, run the port-isolate enable group command with the port isolation group ID specified.
  - You can run the port-isolate mode all command in the system view to configure Layer 2 and Layer 3 isolation.
- Display commands:
  - Run the display port-isolate group all command to check the port isolation group configuration.
  - Run the display port-isolate group X (group ID) command to check the configuration of a specified port isolation group.

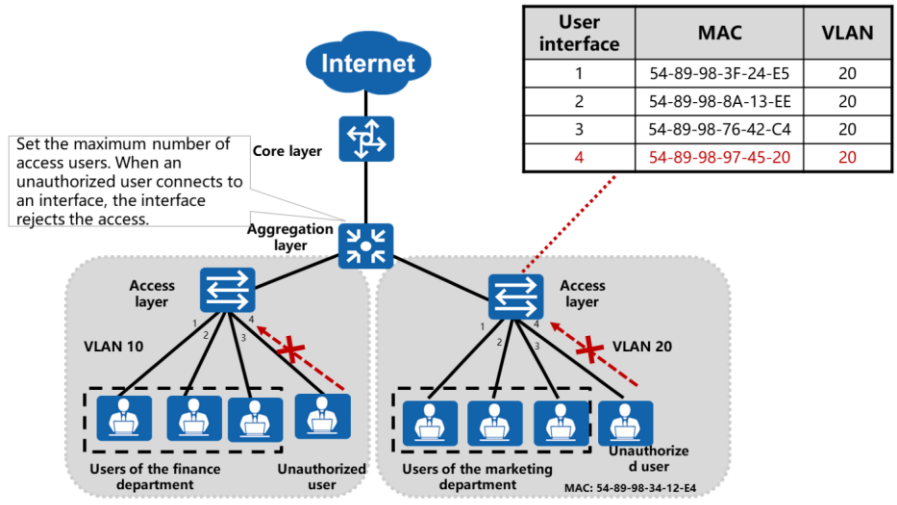
Contents

Application Scenario of Port Security



- When there are unauthorized users on a network, you can use port security to ensure network security.
- Port security usually applies to the following scenarios:
  - The access device configured with port security can defend against attacks initiated by an unauthorized user using another interface.
  - The aggregation device configured with port security can limit the number of access users.

Port Security Implementation



User interface	MAC	VLAN
1	54-89-98-3F-24-E5	20
2	54-89-98-8A-13-EE	20
3	54-89-98-76-42-C4	20
4	54-89-98-97-45-20	20

- You can configure port security on networks demanding high access security. Port security enables the switch to convert MAC addresses learned by an interface into dynamic secure MAC addresses and to stop learning new MAC addresses after the maximum number of learned MAC addresses is reached. In this case, the switch can only communicate with devices with learned MAC addresses. This prevents devices with untrusted MAC addresses from accessing these interfaces, improving security of the device and the network.
- As shown in the figure, the solution is as follows:
  - Enable port security on each interface of an access switch and bind user MAC addresses and VLAN IDs. When unauthorized users connect to the network through the interface configured with port security, the switch searches for its MAC address table. When detecting that the MAC addresses of unauthorized users do not match the MAC address table, the switch discards the data packets.
  - Enable port security on the aggregation switch and set the maximum number of MAC addresses that can be learned by each interface. When the number of learned MAC addresses reaches the limit, the switch discards data packets with other MAC addresses.

Types of Secure MAC Addresses

- Port security changes the dynamic MAC addresses learned on an interface into secure MAC addresses (including dynamic and static secure MAC addresses, and sticky MAC addresses). This function prevents unauthorized users from communicating with the switch using this interface and therefore enhances device security.

Type	Definition	Description
Dynamic secure MAC address	MAC address that is learned on an interface where port security is enabled but the sticky MAC function is disabled	Dynamic secure MAC addresses will be lost after a device restart and need to be learned again. Dynamic secure MAC addresses will never be aged out by default, and can be aged only when an aging time is set for them.
Static secure MAC address	MAC address that is manually configured on an interface where port security is enabled	Sticky MAC addresses are not aged out. The sticky MAC addresses that are saved manually are not lost after a device restart.
Sticky MAC address	MAC address that is learned on an interface where both port security and the sticky MAC function are enabled	Sticky MAC addresses are not aged out. The sticky MAC addresses that are saved manually are not lost after a device restart.

- Note:
  - After port security is enabled on an interface, dynamic MAC address entries that have been learned on the interface are deleted and MAC address entries learned subsequently turn into dynamic secure MAC address entries.
  - After the sticky MAC function is enabled on an interface, existing dynamic secure MAC address entries and MAC address entries learned subsequently on the interface turn into sticky MAC address entries.
  - After port security is disabled on an interface, existing dynamic secure MAC address entries on the interface are deleted. The interface learns dynamic MAC address entries again.
  - After the sticky MAC function is disabled on an interface, sticky MAC address entries on the interface turn into dynamic secure MAC address entries.

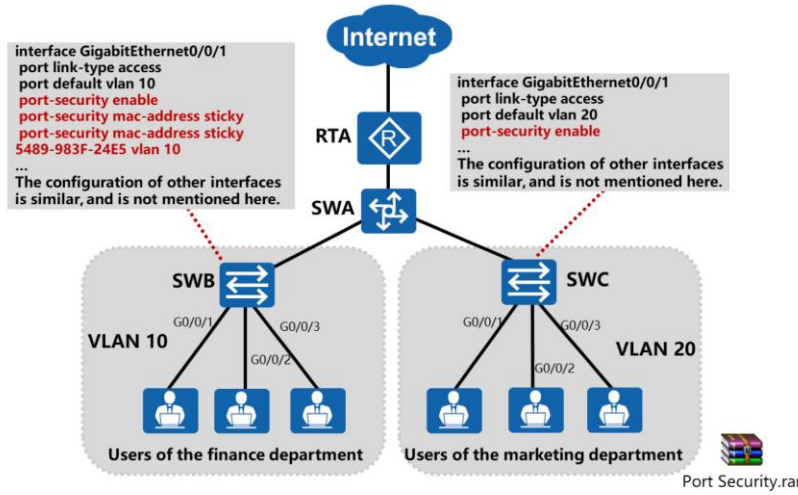
Action to Take After the Number of Secure MAC Addresses Exceeds the Limit

- Action to take after the number of secure MAC addresses reaches the limit

Action	Description
restrict	Discards packets with a nonexistent source MAC address and generates an alarm. This action is recommended.
protect	Only discards packets with a nonexistent source MAC address but does not generate an alarm.
shutdown	Sets the interface state to error-down and generates an alarm. By default, an interface in error-down state can only be restored by using the **restart** command in the interface view.

- If the switch receives packets with a nonexistent source MAC address after the number of secure MAC addresses reaches the limit, the switch considers that the packets are sent from an unauthorized user and takes the configured action on the interface. By default, the switch discards the packets and generates an alarm in such a situation.

Port Security Configuration



- As shown in the preceding figure, the campus network must ensure the security of access users. Employees of the finance department have low mobility, so port security can be used to bind MAC addresses and VLAN IDs of access users. Employees of the marketing department have high mobility, so dynamic MAC address learning of port security can be used to ensure validity of access users.
- Command functions:
  - Run the interface interface-type interface-number command to enter the interface view.
  - Run the port-security enable command to enable port security.
    - By default, port security is not enabled.
  - Run the port-security mac-address sticky command to enable the sticky MAC function.
    - By default, the sticky MAC function is not enabled.
  - Run the port-security max-mac-num max-number command to set the maximum number of sticky MAC addresses.
    - After the sticky MAC function is enabled, the interface can learn only one sticky MAC address by default.

- (Optional) Run the port-security protect-action { protect | restrict | shutdown } command to configure a protection action.
- By default, the protection action is restrict.
- ((Optional) Run the port-security mac-address sticky mac-address vlan vlan-id command to manually configure a sticky MAC address.

Port Security Configuration Verification

- Run the following command to check the bound MAC addresses on SWB:

```
<SWB>display mac-address sticky
MAC address table of slot 0:
-------------------------------------------------------------------------------
MAC Address    VLAN/     PEVLAN CEVLAN Port          Type      LSP/LSR-ID    VSI/SI
MAC-Tunnel
-------------------------------------------------------------------------------
5489-988a-13ee 10          -     -     GE0/0/2       sticky    -
5489-983f-24e5 10          -     -     GE0/0/1       sticky    -
-------------------------------------------------------------------------------
```

- Run the following command to check dynamically learned MAC addresses on SWC:

```
<SWC>display mac-address security
MAC address table of slot 0:
-------------------------------------------------------------------------------
MAC Address    VLAN/     PEVLAN CEVLAN Port          Type      LSP/LSR-ID    VSI/SI
MAC-Tunnel
-------------------------------------------------------------------------------
5489-9876-42c4 20          -     -     GE0/0/1       security  -
5489-9897-4520 20          -     -     GE0/0/2       security  -
-------------------------------------------------------------------------------
```

Quiz

1. For the MUX VLAN, in which VLAN can devices communicate with devices in all VLANs?
   A. Principal VLAN
   B. Separate VLAN
   C. Group VLAN
   D. Subordinate VLAN
2. How many types of secure MAC addresses in port security?
   A. Dynamic secure MAC address
   B. Static secure MAC address
   C. Sticky MAC address
   D. Protected MAC address

HUAWEI

- Answer: A.
- Answer: ABC.


Thank You
www.huawei.com

RSTP Principles and Configurations

Foreword

- STP can solve loop issues, but slow network convergence affects the communication quality. If the network topology changes frequently, the connections on the network where STP is deployed are frequently torn down, causing frequent service interruption. This is intolerable for users.
- Due to limitations of STP, IEEE released 802.1w in 2001 to define RSTP. RSTP enhances STP and implements fast convergence of Layer 2 network topology. What are limitations of STP? What are improvements of RSTP compared with STP?

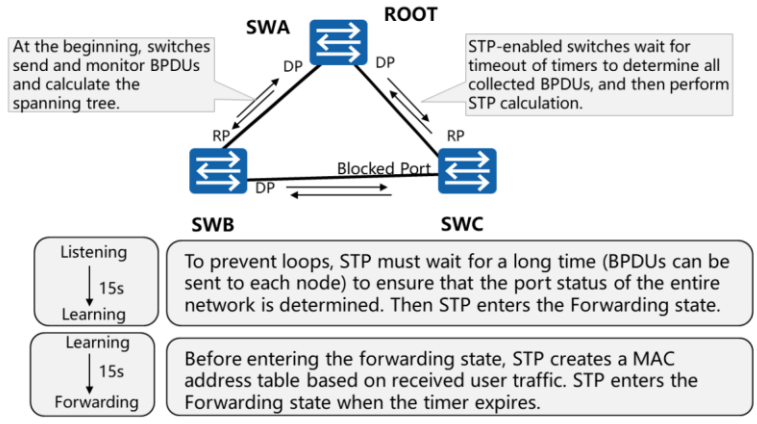
Objectives

- Upon completion of this section, you will be able to:
  - Master the working principle of RSTP
  - Be familiar with similarities and differences between RSTP and STP
  - Understand the RSTP configuration in typical application scenarios

Contents

Problem 1: Switches Run STP in Initialization Scenarios

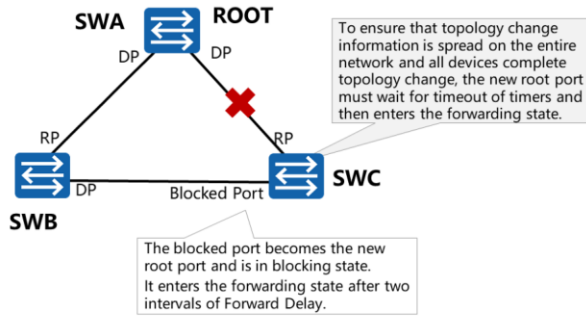
- The duration from initialization to full convergence is at least 30s.



Listening → 15s → Learning	To prevent loops, STP must wait for a long time (BPDUs can be sent to each node) to ensure that the port status of the entire network is determined. Then STP enters the Forwarding state.
Learning → 15s → Forwarding	Before entering the forwarding state, STP creates a MAC address table based on received user traffic. STP enters the Forwarding state when the timer expires.

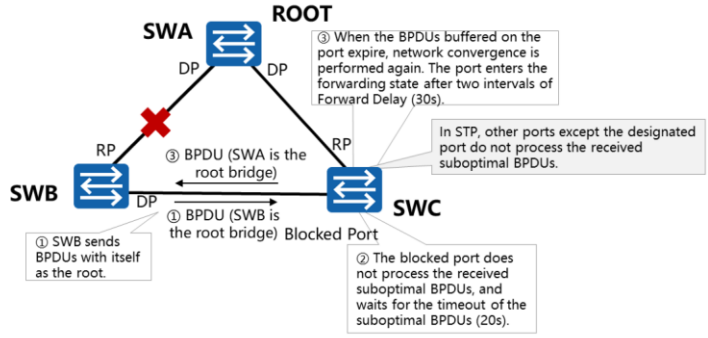
Problem 2: The Switch Has the Blocked Port and the Root Port Goes Down

- The direct link between SWC and SWA goes Down. The blocked port switches to the root port and enters the forwarding state. This process requires at least 30s.



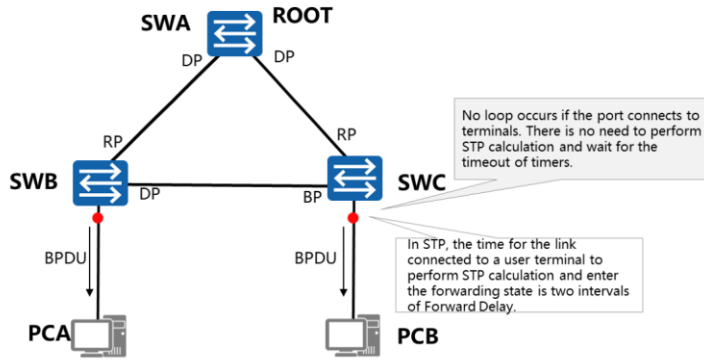
Problem 3: The Switch Has No Blocked Port and the Root Port Goes Down

- The direct link between SWB and SWA goes down. The blocked port of SWC switches to the designated port and enters the forwarding state. This process requires about 50s.



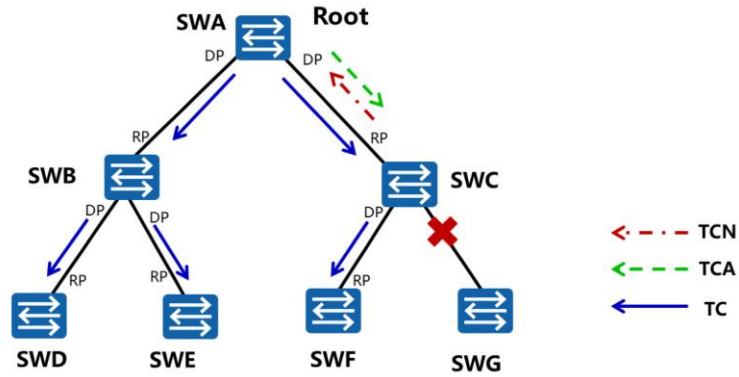
Problem 4: The STP-enabled Switches Connect to User Terminals

- The time for the link between the switch and user terminal to enter the forwarding state is 30s.



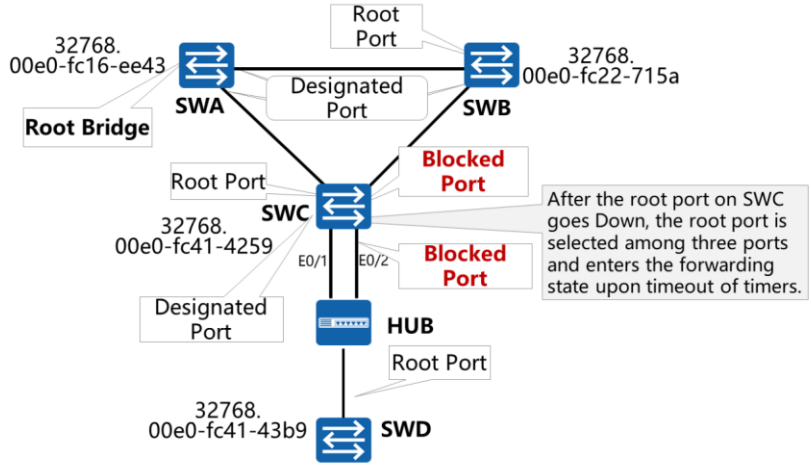
Problem 5: STP Topology Change

- When the topology changes, TCN BPDUs are sent to the root bridge. The upstream bridge that receives the TCN BPDUs sends TCA BPDUs. After the TCN BPDUs reach the root bridge, the root bridge sends TC BPDUs to notify devices of deleting MAC address entries. This implementation is complex and the efficiency is low.



- Processing of topology changes:
  - After the network topology changes, a downstream device continuously sends TCN BPDUs to an upstream device.
  - After the upstream device receives the TCN BPDUs from the downstream device, only the designated port processes the TCN BPDUs. The other ports may receive TCN BPDUs but do not process them.
  - The upstream device sets the TCA bit of the Flags field in the configuration BPDUs to 1 and returns the configuration BPDUs to instruct the downstream device to stop sending TCN BPDUs.
  - The upstream device sends a copy of the TCN BPDUs to the root bridge.
  - The preceding steps are repeated until the root bridge receives the TCN BPDU.
  - The root bridge sets the TC bit of the Flags field in the configuration BPDUs to 1 to instruct the downstream device to delete MAC address entries.

STP Limitation - Port Roles



STP Limitation - Port States

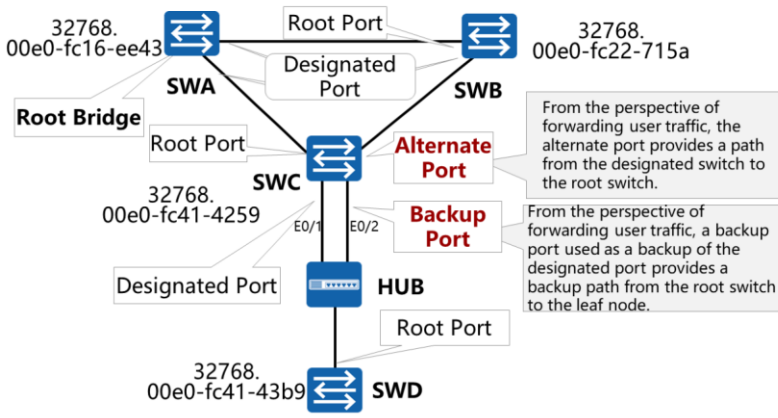
STP Port State	Action
Disabled	The port does not forward user traffic or learn MAC addresses.
Blocking	
Listening	
Learning	The port does not forward user traffic but learns MAC addresses.
Forwarding	The port forwards user traffic and learns MAC addresses.

The three port states correspond to the same action, increasing the difficulty in usage.

Contents

Re-classification of Port Roles

- RSTP adds two port roles: backup port and alternate port.



- To eliminate STP limitations, RSTP adds two port roles, and distinguishes port attributes based on port states and roles to provide more accurate port description. RSTP has simplified port states and faster topology convergence. RSTP defines more port roles to make it easy to understand and deploy STP.
- From the perspective of sending configuration BPDUs:
  - The alternate port is blocked because it learns configuration BPDUs from other network bridges.
  - The backup port is blocked because it learns configuration BPDUs sent by itself.
- From the perspective of user traffic:
  - The alternate port functions as the backup of the root port, and provides a path from the designated bridge to the root bridge.
  - A backup port acts as a backup of the designated port, and provides a path form the root node to the leaf node.
- Allocating roles to all ports is the process of topology convergence.

Re-classification of Port States

- RSTP deletes two port states in STP.

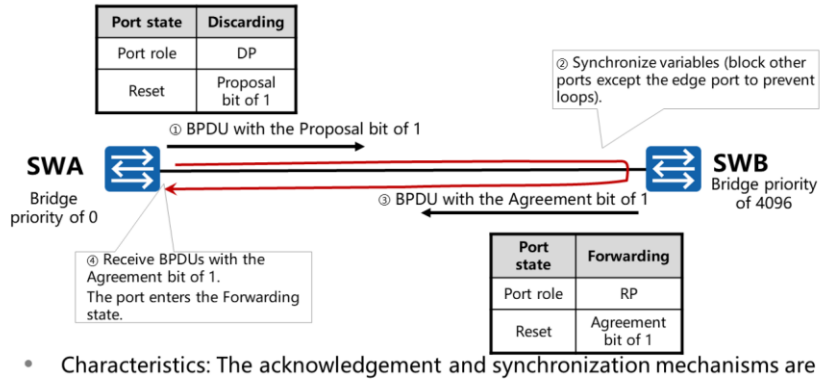
STP Port State	RSTP Port State	Action
Disabled	Discarding	A port that does not forward user traffic or learn MAC addresses is in Discarding state.
Blocking		
Listening		
Learning	Learning	A port that does not forward user traffic but learns MAC addresses is in Learning state.
Forwarding	Forwarding	A port that forwards user traffic and learns MAC addresses is in Forwarding state.

- For the perspective of users, there is no difference in ports in Listening, Learning, and Blocking states because the ports in the three states do not forward user traffic.

Contents

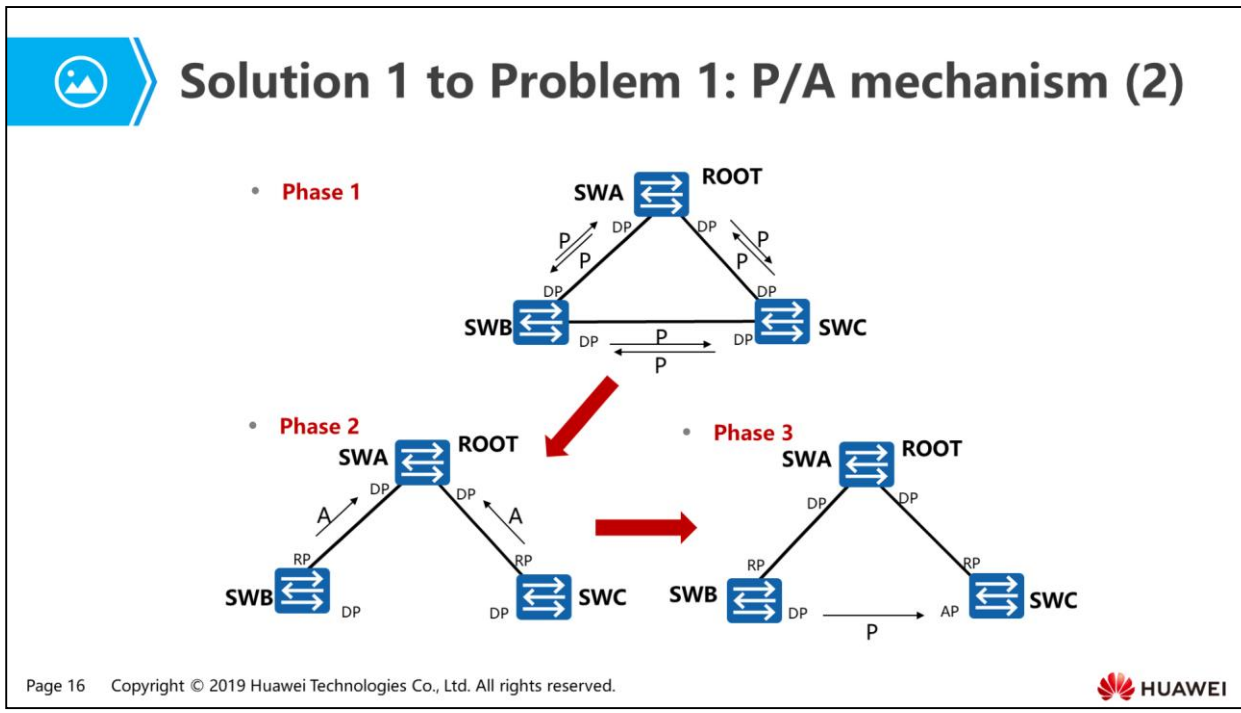
Solution 1 to Problem 1: P/A mechanism (1)

- Principles of P/A mechanism



- Characteristics: The acknowledgement and synchronization mechanisms are used, so there is no need to use timers to ensure a loop-free network.

- The Proposal/Agreement mechanism enables a designated port to enter the Forwarding state as soon as possible.
- The Proposal/Agreement mechanism requires that the link between two switches should be a P2P link in full-duplex mode. When Proposal/Agreement negotiation fails, the designated port needs to wait for two intervals of Forward Delay. The negotiation is similar to that of STP.
- For STP, the designated port can be rapidly selected. To prevent loops, a device must wait for at least two intervals of Forward Delay so that all ports become stable. Then all ports can forward traffic.



- Solution to problem 1:
  - Phase 1: The device is started; RSTP is enabled; each switch considers itself as the root bridge, sends BPDUs with the Proposal bit of 1 to other switches, and changes the port that sends the BPDUs with the Proposal bit to the designated port; the port is in Discarding state.
  - Phase 2: SWA does not process BPDUs with the Proposal bit from SWB and SWC because it has the highest bridge priority. After SWB and SWC receive BPDUs with the Proposal bit from SWA, they send BPDUs with the Agreement bit to SWA based on the Proposal/Agreement negotiation process, and change the transmit ports to root ports. In addition, the ports are in Forwarding state. This is because SWB and SWC consider that SWA is the optimal root bridge.

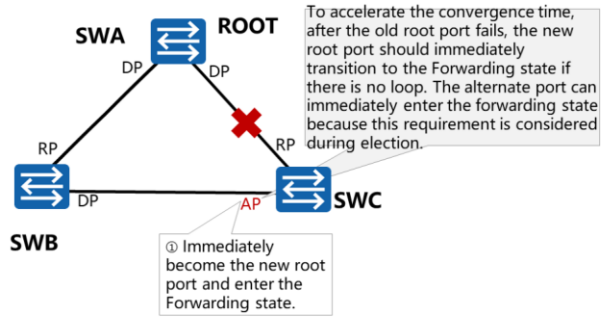
- Phase 3: The Proposal/Agreement negotiation between SWA and SWB and between SWA and SWC is complete. The Proposal/Agreement negotiation between SWB and SWC is as follows:
  - SWB and SWC will send BPDUs with the Proposal bit and SWA as the root bridge to each other.
  - Though SWC and SWB consider SWA as the root bridge, SWC stops sending BPDUs with the Proposal bit after receiving BPDUs with the Proposal bit from SWB. This is because SWC has a lower priority than SWB. SWC does not send the BPDUs with the Agreement bit because there is the root port.
  - Though SWC and SWB consider SWA as the root bridge, SWB continuously sends BPDUs with the Proposal bit after receiving BPDUs with the Proposal bit from SWC. This is because SWB has a higher priority than SWC.
  - After two intervals of Forward Delay, the port of SWB becomes the designated port in Forwarding state and the port of SWC becomes the alternate port in Discarding state.

Solution 1 to Problem 1: P/A mechanism (3)

- The election principles of STP and RSTP are similar: electing the root switch, root port of the non-root-switch, designated port, and alternate port and backup port in sequence.
- RSTP adds the Proposal/Agreement mechanism. There is the acknowledgement mechanism during negotiation, so RSTP-enabled switches can forward BPDUs, without depending on timers to ensure the loop-free network topology. RSTP-enabled switches only need to consider the time for sending BPDUs and calculating a loop-free topology (usually within seconds).

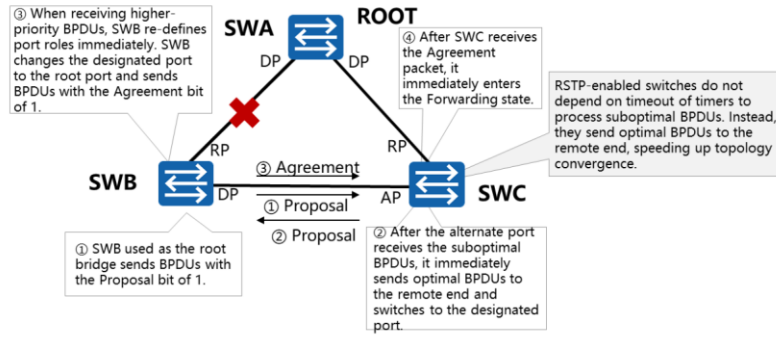
Solution to Problem 2: Fast Switching of the Root Port

- The direct link between SWC and SWA goes down. The alternate port switches to the root port and enters the forwarding state within seconds.



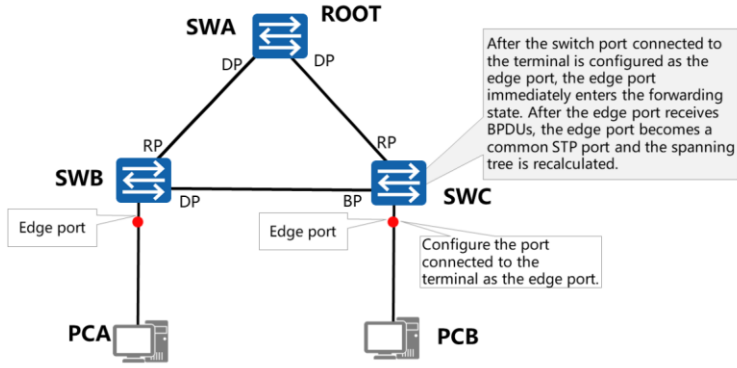
Solution to Problem 3: Processing of Suboptimal BPDUs

- The direct link between SWB and SWA goes down. The alternate port of SWC switches to the designated port and enters the forwarding state within seconds.



Solution to Problem 4: Edge Port

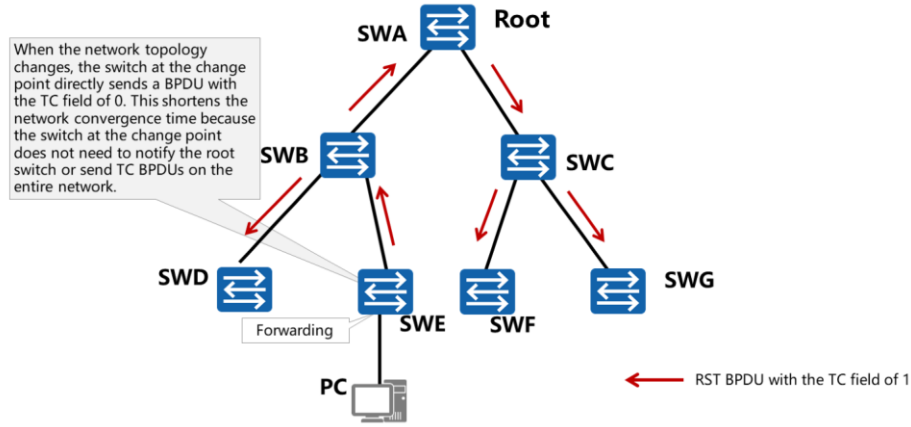
- In RSTP, the switch interface connected to a terminal can immediately enter the Forwarding state.



Contents

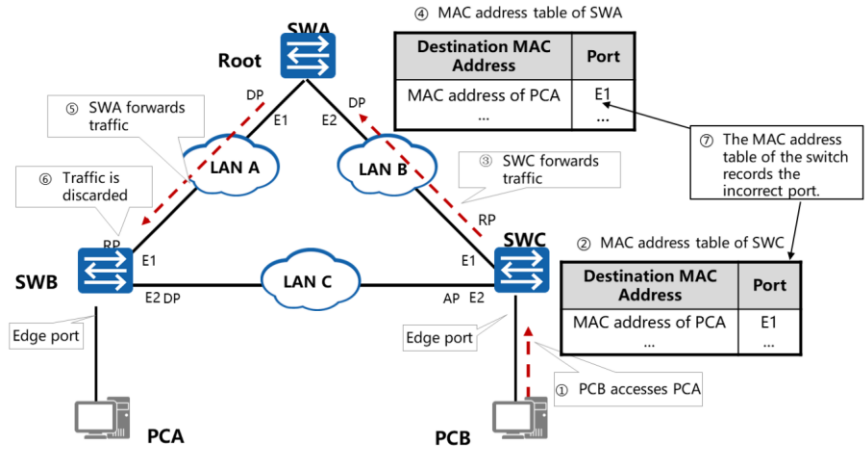
Solution to Problem 5: Topology Change Optimization

- Topology change identification: The non-edge port enters the Forwarding state.



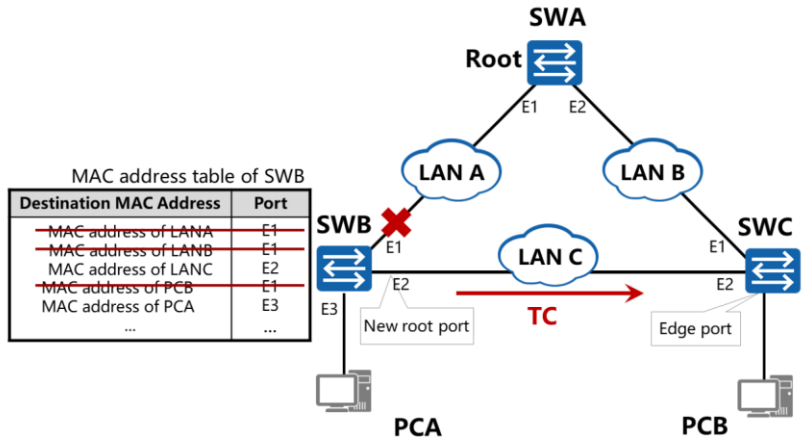
- When a topology change is detected, the procedure is as follows:
  - A TC While timer is started on each non-edge designated port of the switch. The value of the timer is twice the value of the Hello timer. Within this period of time, MAC addresses learned on the port of which the status changes are cleared. The ports send RST BPDUs with the TC field of 1. When the TC While Timer expires, the ports stop sending RST BPDU.
  - After other switches receive the RST BPDUs, they clears MAC addresses of all ports except the ports that receive RST BPDUs. A TC While timer is started on each non-edge designated port of the switch. The preceding process is repeated. In this case, RST BPDUs are flooded.

Problems Caused by Topology Changes



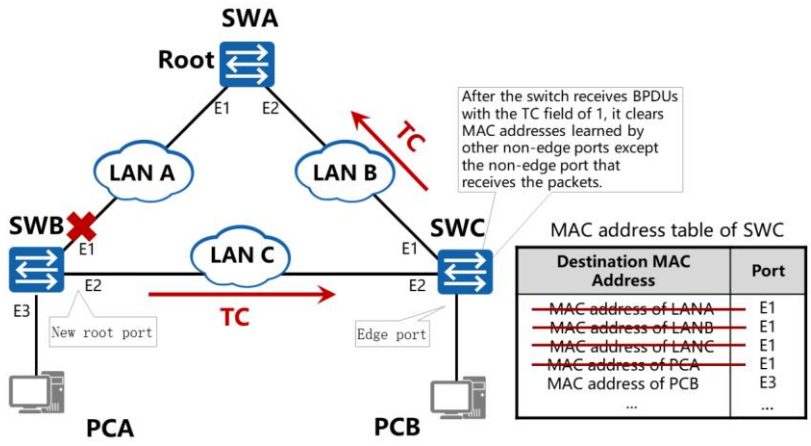
- RSTP considers that the topology changes only when a non-edge port changes to the Forwarding state.
- Network topology changes may cause the MAC address table of the switch to be generated incorrectly.
- When the network is stable, the port corresponding to the MAC address of PCA in the MAC address table of SWC is E1. If E1 on SWB fails but the port corresponding to the MAC address of PCA in the MAC address table of SWC is still E1, data may be lost.

Processing of Topology Changes (1)

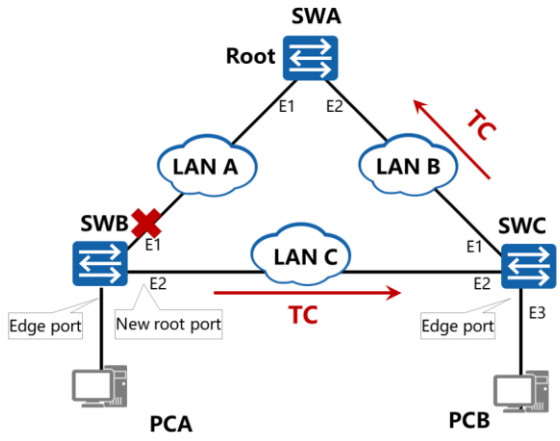


- When a topology change is detected, the procedure is as follows:
  - MAC addresses learned on the port of which the status changes are cleared.
  - The ports send RST BPDUs with the TC field of 1. When the TC While Timer expires, the ports stop sending RST BPDU.
- After E1 on SWB fails, the procedure is as follows:
  - SWB recalculates the spanning tree and E2 is selected as the new root port.
  - SWB deletes the MAC address entry corresponding to E1 in the MAC address table.
  - After the spanning tree recalculation is complete (the port enters the Forwarding state), all non-edge ports of SWB send RST BPDUs with the TC field of 1.

Processing of Topology Changes (2)



Processing of Topology Changes (3)



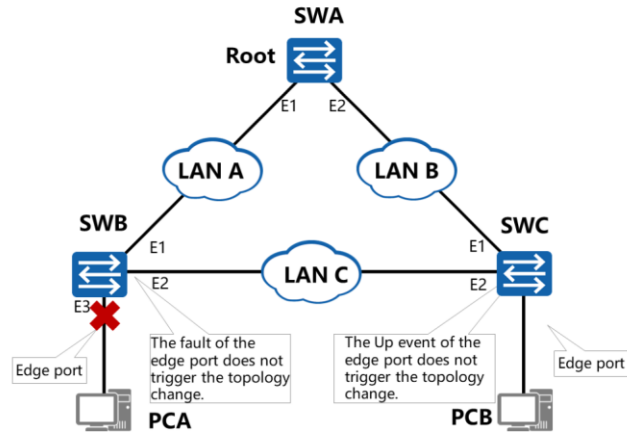
MAC address table of SWA

Destination MAC Address	Port
~~MAC address of LANA~~	~~E1~~
MAC address of LANB	E2
~~MAC address of LANC~~	~~E1~~
~~MAC address of PCA~~	~~E1~~
MAC address of PCB	E2
...	...

Processing of Topology Changes (4)

MAC address table of SWB

Destination MAC Address	Port
MAC address of LANA	E1
MAC address of LANB	E1
MAC address of LANC	E2
MAC address of PCB	E1
~~MAC address of PCA~~	~~E3~~
...	...

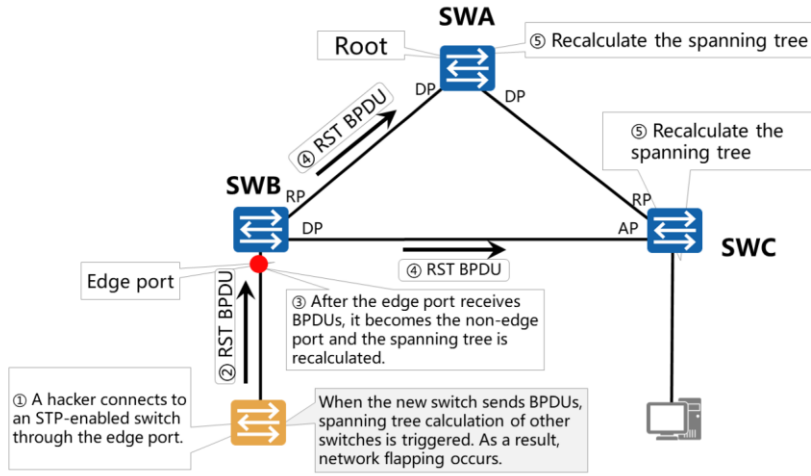


HUAWEI

- The Down event of the edge port does not trigger the topology change. After the fault is rectified, topology change is not triggered.

Contents

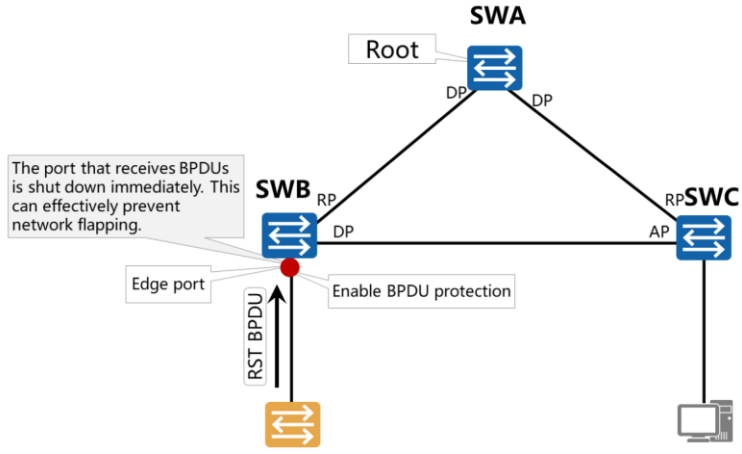
BPDU Protection (1)



HUAWEI

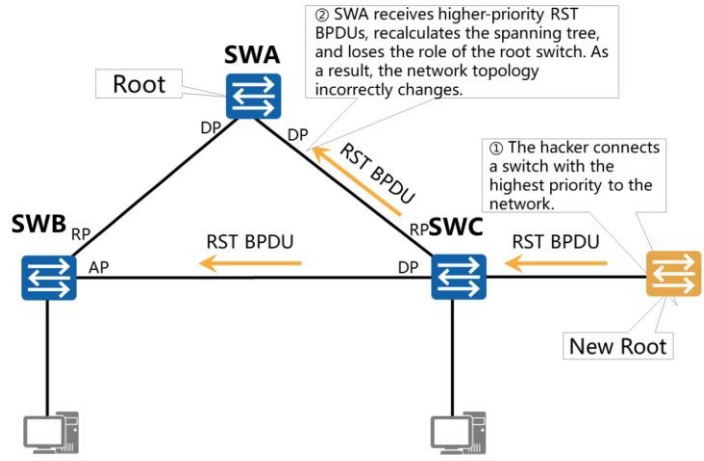
- BPDU protection
  - Application scenario: BPDU protection prevents a switch from being attacked by RST BPDUs. When the edge port receives bogus RST BPDUs, the edge port is automatically configured as a non-edge port and the spanning tree is re-calculated, causing network flapping.

BPDU Protection (2)



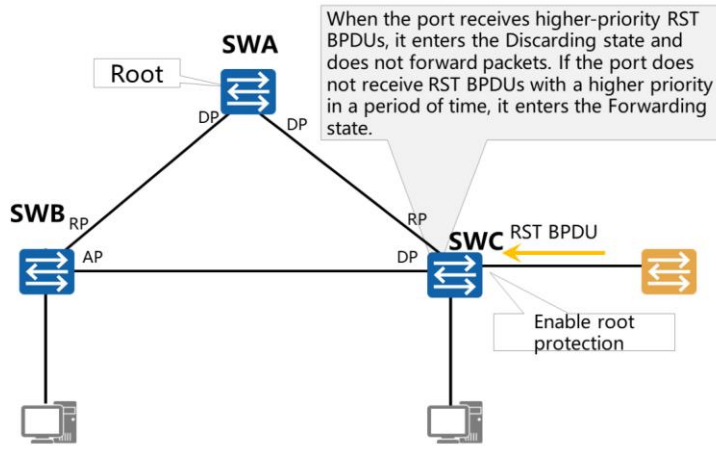
- BPDU protection
  - Implementation: When the edge port is configured with BPDU protection, the edge port that receives BPDUs is immediately shut down.

Root Protection (1)



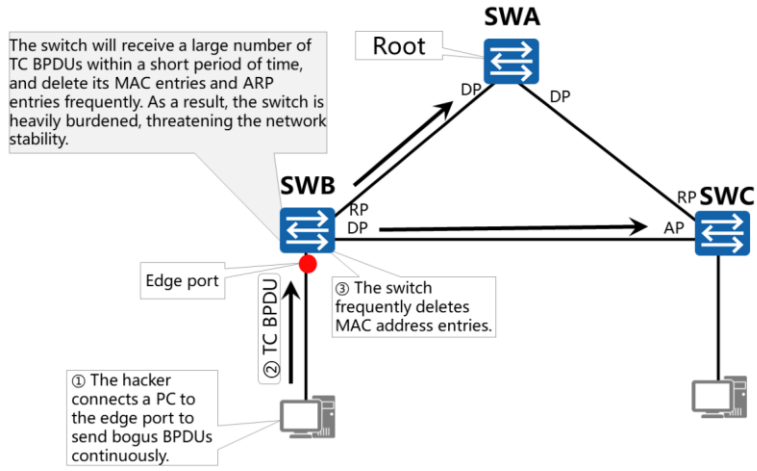
- Root protection
  - Application scenario: The root bridge may receive RST BPDUs with a higher priority because of incorrect configurations or malicious attacks on a network. Then the root bridge loses its role and the network topology changes incorrectly.

Root Protection (2)



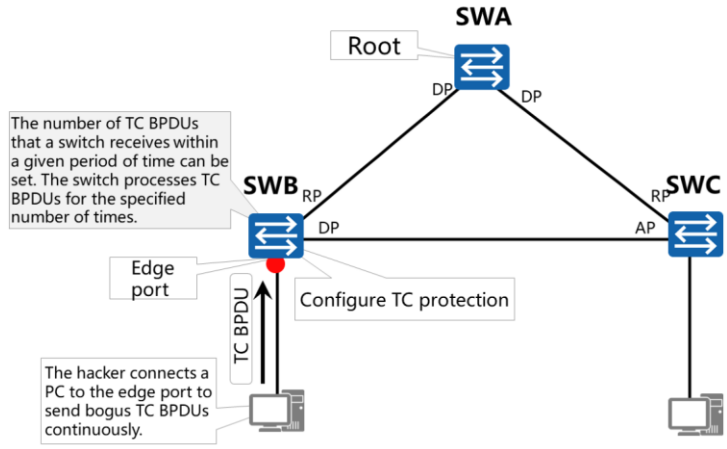
- Root protection
  - Implementation: When a port enabled with root protection receives an RST BPDU with a high priority, the port enters the Discarding state and does not forward packets. If the port does not receive RST BPDU with a higher priority for a certain period, the port enters the Forwarding state.
  - Root protection can be configured on only designated ports.

TC Protection (1)



- TC attacks
  - When a switch receives TC BPDUs, it deletes the corresponding MAC address entries and ARP entries. If a malicious attacker sends bogus TC BPDUs to attack the switch, the switch will receive a large number of TC BPDUs within a short time period, and delete its MAC address entries and ARP entries frequently. As a result, the switch is heavily burdened, threatening the network stability.

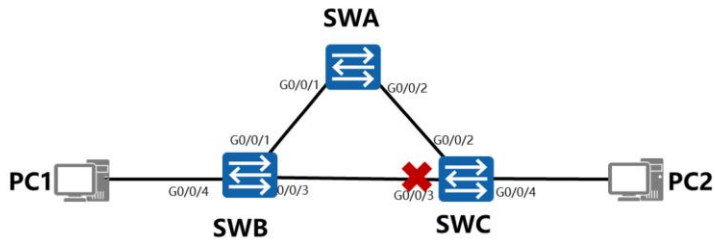
TC Protection (2)



- TC protection
  - After enabling TC protection, you can set the number of times TC BPDUs are processed by the RSTP process within a given period of time (the default period is 2s and the default number of times is 3). If the number of TC BPDUs that the RSTP process receives within the given period of time exceeds the specified threshold, the RSTP process processes TC BPDUs only for the specified number of times. The RSTP process processes excess TC BPDUs once after the timer expires. This function prevents the switch from frequently deleting its MAC address entries and ARP entries and protects the switch.

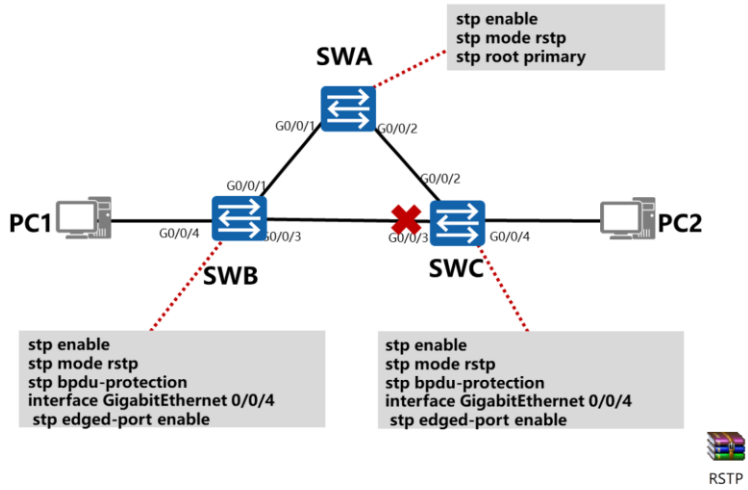
Contents

RSTP Configuration Requirements



- As shown in the preceding figure, SWA, SWB, and SWC form a ring switching network. To eliminate the impact of loops on the network, the switches run RSTP to prune the ring network into a loop-free tree network.

RSTP Configuration Procedure



HUAWEI

- Configuration:
  - stp enable //Enable STP globally.
  - stp mode rstp //Set the RSTP mode.
  - stp root primary //Configure SWA as the root bridge.
  - stp bpdu-protection //Enable BPDU protection globally.
  - stp edged-port enable //Configure the port as the edge port.

RSTP Configuration Verification (1)

- Check STP information on SWA.

```
<SWA>display stp brief
 MSTID  Port                        Role  STP State          Protection
   0    GigabitEthernet0/0/1        DESI  FORWARDING         NONE
   0    GigabitEthernet0/0/2        DESI  FORWARDING         NONE
```

```
<SWA>display stp
-------[CIST Global Info][Mode RSTP]-------
CIST Bridge         :0    .4c1f-cc5f-55e4
Config Times        :Hello 2s MaxAge 20s FwDly 15s MaxHop 20
Active Times        :Hello 2s MaxAge 20s FwDly 15s MaxHop 20
CIST Root/ERPC      :0    .4c1f-cc5f-55e4 / 0
CIST RegRoot/IRPC   :0    .4c1f-cc5f-55e4 / 0
CIST RootPortId     :0.0
BPDU-Protection     :Disabled
CIST Root Type      :Primary root
```

RSTP Configuration Verification (2)

- Check STP information on SWB.

```
[SWB]display stp brief
 MSTID  Port                        Role  STP State     Protection
   0    GigabitEthernet0/0/1        ROOT  FORWARDING    NONE
   0    GigabitEthernet0/0/3        DESI  FORWARDING    NONE
   0    GigabitEthernet0/0/4        DESI  FORWARDING    BPDU
```

- Check STP information on SWC.

```
<SWC>display stp brief
 MSTID  Port                        Role  STP State     Protection
   0    GigabitEthernet0/0/2        ROOT  FORWARDING    NONE
   0    GigabitEthernet0/0/3        ALTE  DISCARDING    NONE
   0    GigabitEthernet0/0/4        DESI  FORWARDING    BPDU
```


- G0/0/3 on SWC is blocked to prevent loops.

Quiz

1. How many port states does RSTP define?
   A. 2
   B. 3
   C. 4
2. Which port roles does RSTP define?

- Answer: B.
- Answer: root port, designated port, backup port, alternate port, edge port.


Thank You
www.huawei.com

MSTP Principles and Configurations

Foreword

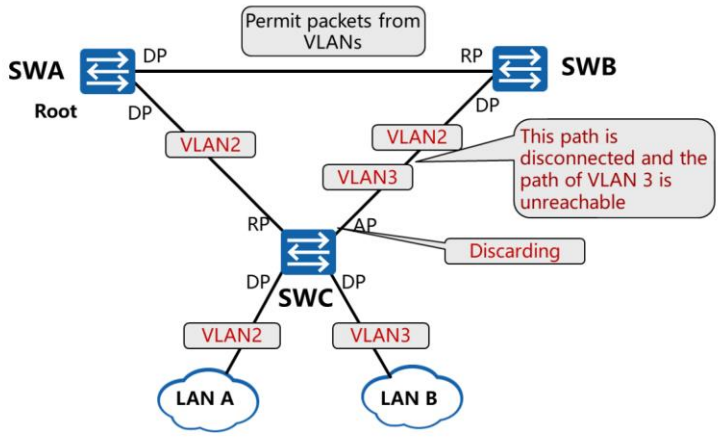
- The Rapid Spanning Tree Protocol (RSTP), an enhancement to the Spanning Tree Protocol (STP), allows for fast network topology convergence. However, both RSTP and STP have a significant shortcoming: All VLANs on a LAN use one spanning tree, making VLAN-based load balancing impossible. Once a link is blocked, it will no longer transmit traffic, wasting bandwidth and causing the failure in forwarding certain VLAN packets.
- To improve on the shortcoming of STP and RSTP, the IEEE issued 802.1s to define the Multiple Spanning Tree Protocol (MSTP) in 2002. MSTP implements fast convergence and provides multiple paths to load balance VLAN traffic.

Objectives

- Upon completion of this section, you will be able to:
  - Be familiar with limitations of a single spanning tree
  - Understand the MSTP working mechanism
  - Master basic MSTP configurations

Contents

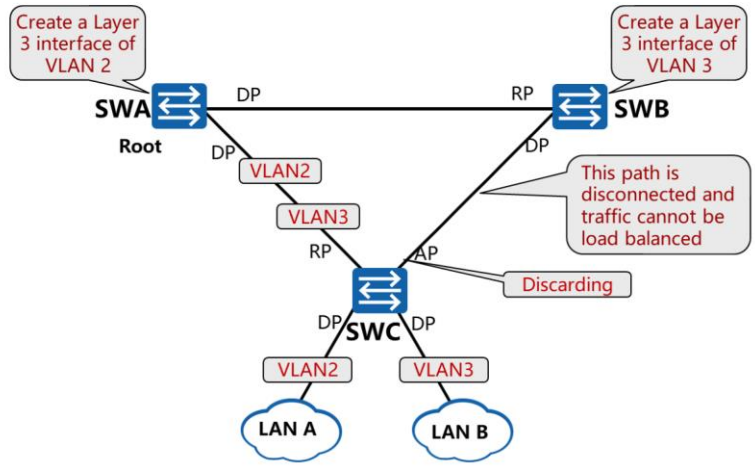
Limitations of a Single Spanning Tree - Some VLAN Paths Are Unreachable



HUAWEI

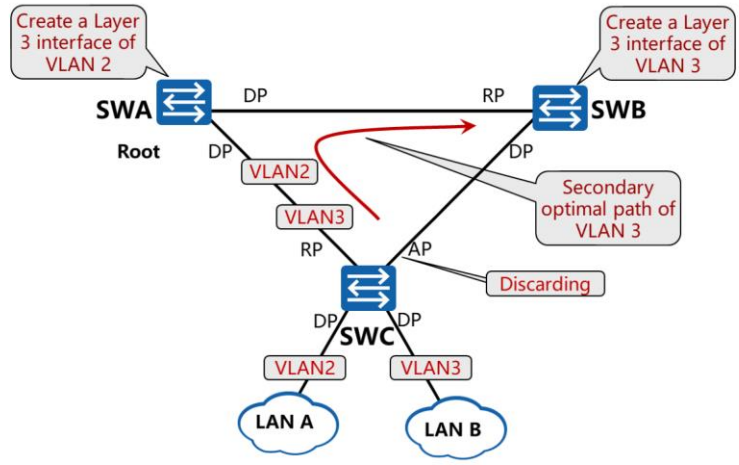
- Three switches are deployed on the network: SWA, SWB, and SWC. Traffic in VLAN 2 needs to be sent through two uplinks and traffic in VLAN 3 needs to be sent through only one uplink.
- A spanning tree needs to be deployed to solve loops of VLAN 2. When a single spanning tree is used, if the connected ports between SWC and SWB are alternate ports in Discarding state, the path of VLAN is disconnected and traffic of VLAN 3 cannot be sent to SWB.

Limitations of a Single Spanning Tree - Traffic Cannot Be Load Balanced



- To implement load balancing, two uplinks are configured as trunk links to allow packets from VLANs and the link between SWA and SWB is also configured as the trunk link to allow packets from all VLANs. The Layer 3 interfaces of VLAN 2 and VLAN 3 are configured on SWA and SWB, respectively.
- Traffic in VLAN 2 and VLAN 3 is required to reach the corresponding Layer 3 interfaces through different uplinks. If the port connected to SWB is the alternate port in Discarding state, data in VLAN 2 and VLAN 3 can reach SWA through only one uplink. Traffic cannot be load balanced.

Limitations of a Single Spanning Tree - Secondary Optimal Layer 2 Path

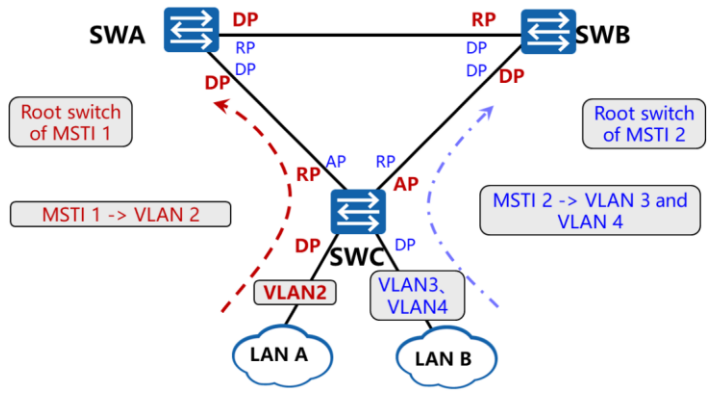


HUAWEI

- As shown in the figure, the links between SWC and SWA and between SWC and SWB are configured as trunk links to allow packets from all VLANs and the link between SWA and SWB is also configured as a trunk link to allow packets from all VLANs.
- After a spanning tree is used, the loop is eliminated and traffic of VLAN 2 and VLAN 3 is sent to SWA.
- Layer 3 interfaces of VLAN 2 and VLAN 3 are configured on SWA and SWB, respectively. The path where traffic of VLAN 3 is transmitted to the Layer 3 interface is the secondary optimal path.

Contents

MSTIs Overcome the Limitations of a Single Spanning Tree



- An MST region may contain multiple spanning trees, each of which is an MSTI. MSTI are independent of each other and the calculation process of each MSTI is the same as the RSTP calculation process.

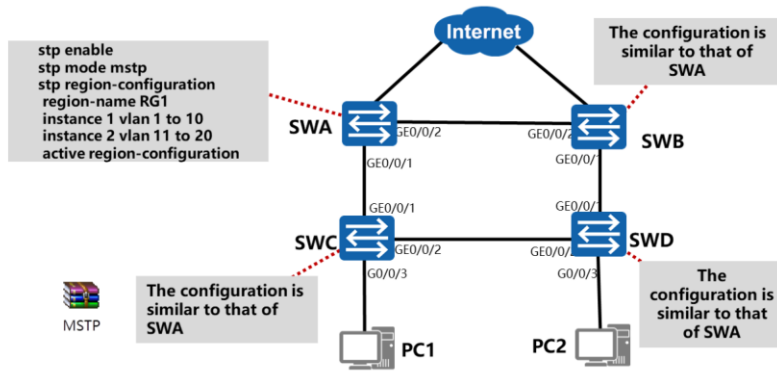
  HUAWEI

- A Multiple Spanning Tree (MST) region contains multiple network segments, each of which contains switches. The switches in one MST region all share the following characteristics:
  - MSTP-enabled
  - Same region name
  - Same VLAN-MSTI mappings
  - Same MSTP revision level
- Multiple spanning trees can be generated in one MST region. Each spanning tree is called a Multiple Spanning Tree Instance (MSTI), and each MSTI uses an independent RSTP algorithm to calculate a spanning tree.
- Each MSTI has an MSTI ID, which is a 2-byte integer. The VRP supports 16 MSTIs. The MSTI ID ranges from 0 to 15. All VLANs are mapped to MSTI 0 by default.
- The VLAN mapping table defines the mapping between VLANs and MSTIs. AN MSTI can be mapped to one or more VLANs, but one VLAN can map to only one MSTI.
- MSTP is compatible with STP and RSTP. MSTP implements fast convergence and provides multiple paths to load balance VLAN traffic.

Contents

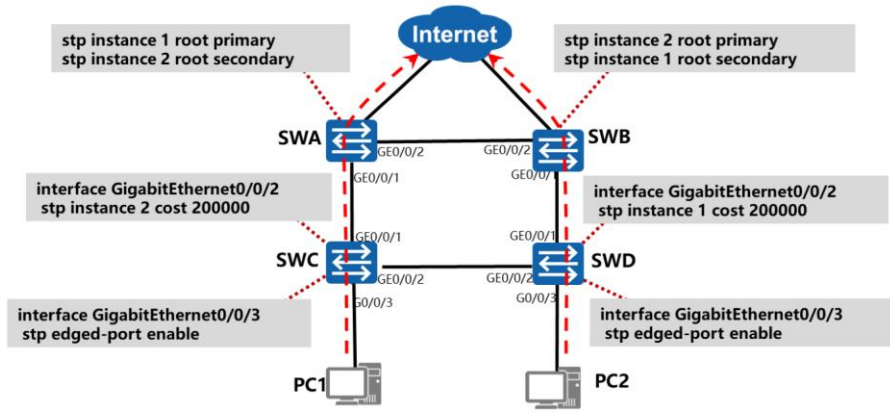
MSTP Configuration Requirements

- MSTP can be used to load balance of Internet access traffic of PCs from different VLANs. VLANs 1 to 10 map MSTI 1, and VLANs 11 to 20 map MSTI 2.



- Configuration roadmap:
  - Configure an MST region and create MSTIs to load balance traffic.
  - Configure the root bridge and secondary root bridge of each MSTI in the MST region.
  - Configure the path cost of a port in each MSTI so that the port can be blocked.
  - Configure the port connected to a terminal device as an edge port to accelerate route convergence.
- Data preparation:
  - MST region name: RG1
  - MSTI: MSTI 1 and MSTI 2
  - In MSTI 1, SWA is the root bridge and SWB is the secondary root bridge. In MSTI 2, SWB is the root bridge and SWA is the secondary root bridge.
  - The path costs of blocked ports in MSTI 1 and MSTI 2 are changed to 200000.
  - VLAN IDs: 1 to 20
  - PC1 belongs to VLAN 10 and PC2 belongs to VLAN 20.

MSTP Configuration Procedure



MSTP Configuration Verification (1)

- Check the port status on SWA.

```
[SWA]display stp brief
MSTID     Port                      Role  STP State       Protection
  0   GigabitEthernet0/0/1          DESI  FORWARDING      NONE
  0   GigabitEthernet0/0/2          DESI  FORWARDING      NONE
  1   GigabitEthernet0/0/1          DESI  FORWARDING      NONE
  1   GigabitEthernet0/0/2          DESI  FORWARDING      NONE
  2   GigabitEthernet0/0/1          DESI  FORWARDING      NONE
  2   GigabitEthernet0/0/2          ROOT  FORWARDING      NONE
```

- Check the port status on SWB.

```
[SWB]display stp brief
MSTID     Port                      Role  STP State       Protection
  0   GigabitEthernet0/0/1          DESI  FORWARDING      NONE
  0   GigabitEthernet0/0/2          ROOT  FORWARDING      NONE
  1   GigabitEthernet0/0/1          DESI  FORWARDING      NONE
  1   GigabitEthernet0/0/2          ROOT  FORWARDING      NONE
  2   GigabitEthernet0/0/1          DESI  FORWARDING      NONE
  2   GigabitEthernet0/0/2          DESI  FORWARDING      NONE
```

MSTP Configuration Verification (2)

- Check the port status on SWC.

```
[SWC]display stp brief
 MSTID  Port                        Role  STP State          Protection
   0    GigabitEthernet0/0/1        ROOT  FORWARDING         NONE
   0    GigabitEthernet0/0/2        DESI  FORWARDING         NONE
   0    GigabitEthernet0/0/3        DESI  FORWARDING         NONE
   1    GigabitEthernet0/0/1        ROOT  FORWARDING         NONE
   1    GigabitEthernet0/0/2        DESI  FORWARDING         NONE
   1    GigabitEthernet0/0/3        DESI  FORWARDING         NONE
   2    GigabitEthernet0/0/1        ROOT  FORWARDING         NONE
   2    GigabitEthernet0/0/2        ALTE  DISCARDING         NONE
```

- Check the port status on SWD.

```
<SWD>display stp brief
 MSTID  Port                        Role  STP State     Protection
   0    GigabitEthernet0/0/1        ALTE  DISCARDING    NONE
   0    GigabitEthernet0/0/2        ROOT  FORWARDING    NONE
   0    GigabitEthernet0/0/3        DESI  FORWARDING    NONE
   1    GigabitEthernet0/0/1        ROOT  FORWARDING    NONE
   1    GigabitEthernet0/0/2        ALTE  DISCARDING    NONE
   1    GigabitEthernet0/0/3        DESI  FORWARDING    NONE
   2    GigabitEthernet0/0/1        ROOT  FORWARDING    NONE
   2    GigabitEthernet0/0/2        DESI  FORWARDING    NONE
```

Quiz

1. Describe the limitations of a single spanning tree.
2. Which of the following statements about MSTP is false?
   A. One MST region can have only one MSTI.
   B. Each MSTI uses the RSTP algorithm independently.
   C. MSTP is compatible with STP.
   D. An MSTI can correspond to one or more VLANs.

- Answer: The blocked link does not carry any traffic. As a result, VLAN-based load balancing cannot be implemented and the bandwidth is wasted. In addition, some paths of VLANs are unreachable, causing the secondary optimal path.
- Answer: A.


Thank You
www.huawei.com

Recommendations

- Huawei Learning Website
  - http://learning.huawei.com/en
- Huawei e-Learning
  - https://ilearningx.huawei.com/portal/#/portal/EBG/51
- Huawei Certification
  - http://support.huawei.com/learning/NavigationAction!createNavi?navId=_31&lang=en
- Find Training
  - http://support.huawei.com/learning/NavigationAction!createNavi?navId=_trainingsearch&lang=en

More Information

- Huawei learning APP