学习推荐

- 华为培训与认证官方网站
  - http://learning.huawei.com/cn/
- 华为在线学习
  - https://ilearningx.huawei.com/portal/#/portal/ebg/26
- 华为职业认证
  - http://support.huawei.com/learning/NavigationAction!createNavi?navId=_31&lang=zh
- 查找培训入口
  - http://support.huawei.com/learning/NavigationAction!createNavi?navId=_trainingsearch&lang=zh

更多信息

- 华为培训APP

华为认证系列教程

HCIP-Routing & Switching-IERS

华为认证数通资深工程师 - 部署企业级路由交换网络

华为技术有限公司

版权声明

华为认证系列教程

HCIP-Routing & Switching-IERS

第 2.5 版 本

目录

OSPF协议基础 1

OSPF域内路由 41

OSPF域间路由 64

OSPF外部路由 80

OSPF特殊区域及其他特性 95

IS-IS协议原理与配置 122

BGP协议原理与配置 152

IP组播基础 210

IGMP协议原理与配置 228

PIM协议原理与配置 255

路由控制 288

Eth-Trunk技术原理与配置 330

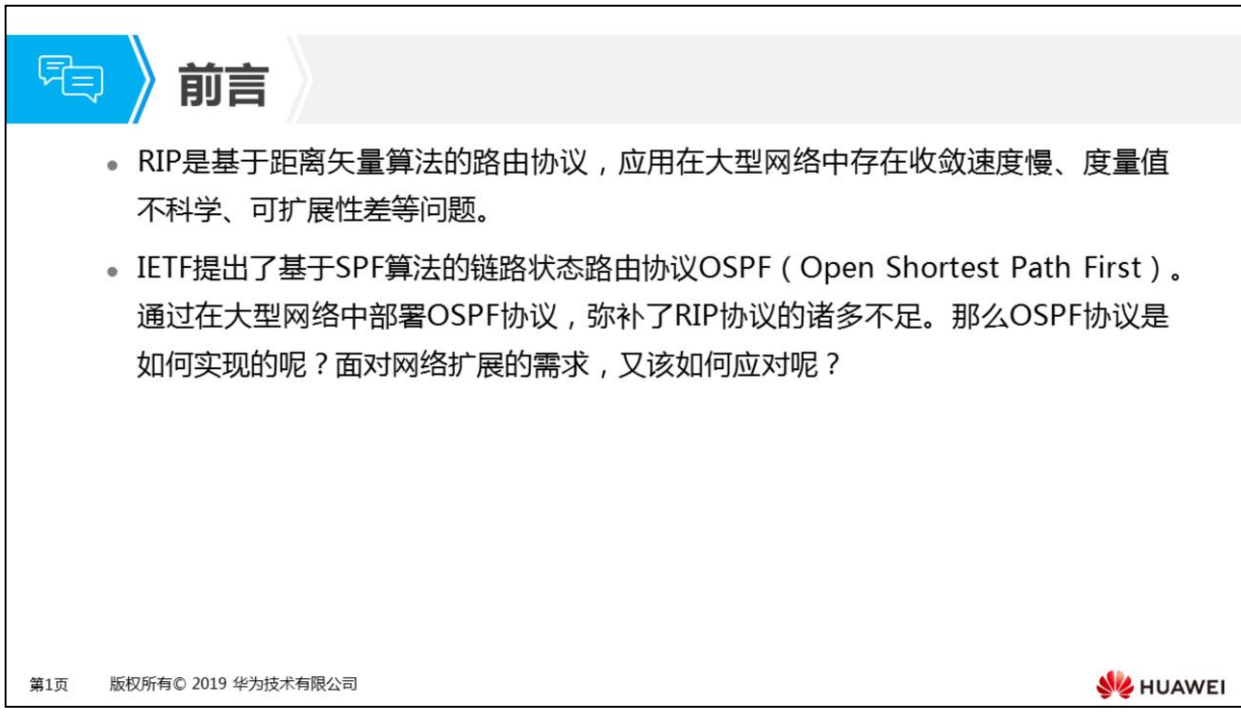
交换机高级特性简介 354

RSTP协议原理与配置 377

MSTP协议原理与配置 419

OSPF协议基础

HUAWEI



- 互联网工程任务组：The Internet Engineering Task Force(IETF)。

目标

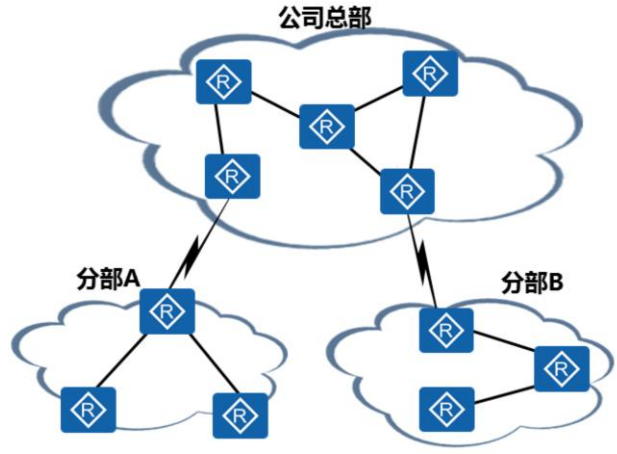
- 学完本课程后，您将能够：
  - 了解RIP在大型网络中部署所面临的问题
  - 掌握OSPF协议的基本特点
  - 掌握OSPF协议所支持的网络类型
  - 掌握OSPF协议邻接关系的建立过程
  - 掌握OSPF协议DR/BDR的概念和作用

目录

1. **RIP在大型网络中部署面临的挑战**
2. OSPF基本工作原理

大型网络所发生的变化

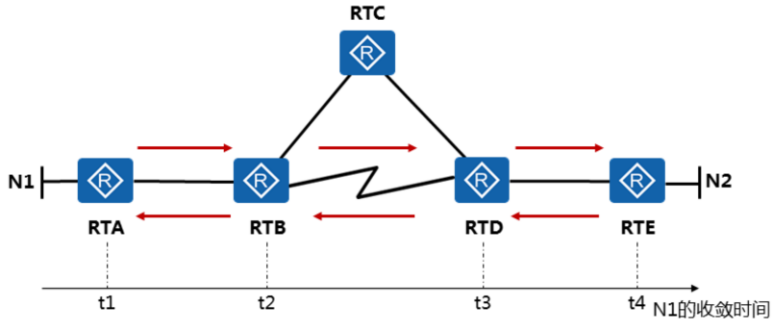
- 网络规模扩大。
- 网络可靠性要求提高。
- 网络异构化趋势加剧。



HUAWEI

- 网络规模扩大：
  - 企业新业务层出不穷，且业务呈现大集中趋势，使得网络规模不断扩大。
- 网络可靠性要求提高：
  - 各种应用程序对网络可靠性要求越来越高，网络发生故障后，需要在更短的时间内恢复正常。
- 网络异构化，多厂商设备互联需求：
  - 在日常的运营维护中，硬件设备不断升级或更新，不同设备之间性能差异较大，设备间互连链路带宽也存在一定的差异。
  - 需要一种各厂商均支持的开放路由协议。
- 面对越来越高的要求与挑战，如果通过RIP来部署，会遇到什么问题？

RIP在大型网络中部署所面临的问题



RIP特性	带来的问题
逐跳收敛	收敛慢，故障恢复时间长
传闻路由更新机制	缺少对全局网络拓扑的了解
最多有效跳数为15	环形组网中，使远端路由不可达
以“跳数”为度量	存在选择次优路径的风险

HUAWEI

- 逐跳收敛：
  - 如图所示，N1网络发生变化，RTA向RTB发出更新，RTB收到更新之后进行本地计算，完成计算后再向RTC发送路由变化通知，如此循环。逐跳收敛的方式，造成了网络收缓慢的问题。
- 传闻路由更新机制：
  - RIP在计算路由完全依赖于从邻居路由器收到的路由信息，RTE仅依靠从RTD获取的信息计算路由，对RTA、RTB和RTC之间的网络情况并不了解。RIP在计算路由时，缺少对全局网络拓扑的了解。
- 以“跳数”为度量：
  - 因为RIP基于跳数的度量方式，所以N1与N2网络互访时会选择RTA->RTB->RTD->RTE作为最优路径。显然RTB->RTC->RTD之间的以太链路要比RTB->RTD的串行链路带宽要高的多。
- 针对RIP遇到的问题，可以通过什么方式优化或者解决？

如何解决RIP的问题？

RIP的问题	优化或解决的方式
收敛慢，故障恢复时间长	“收到更新->计算路由->发送更新” 改为“收到更新->发送更新->计算路由”
缺少对全局网络拓扑的了解	路由器基于拓扑信息，独立计算路由
最多有效跳数为15	不限定跳数
存在选择次优路径的风险	将链路带宽作为选路参考值

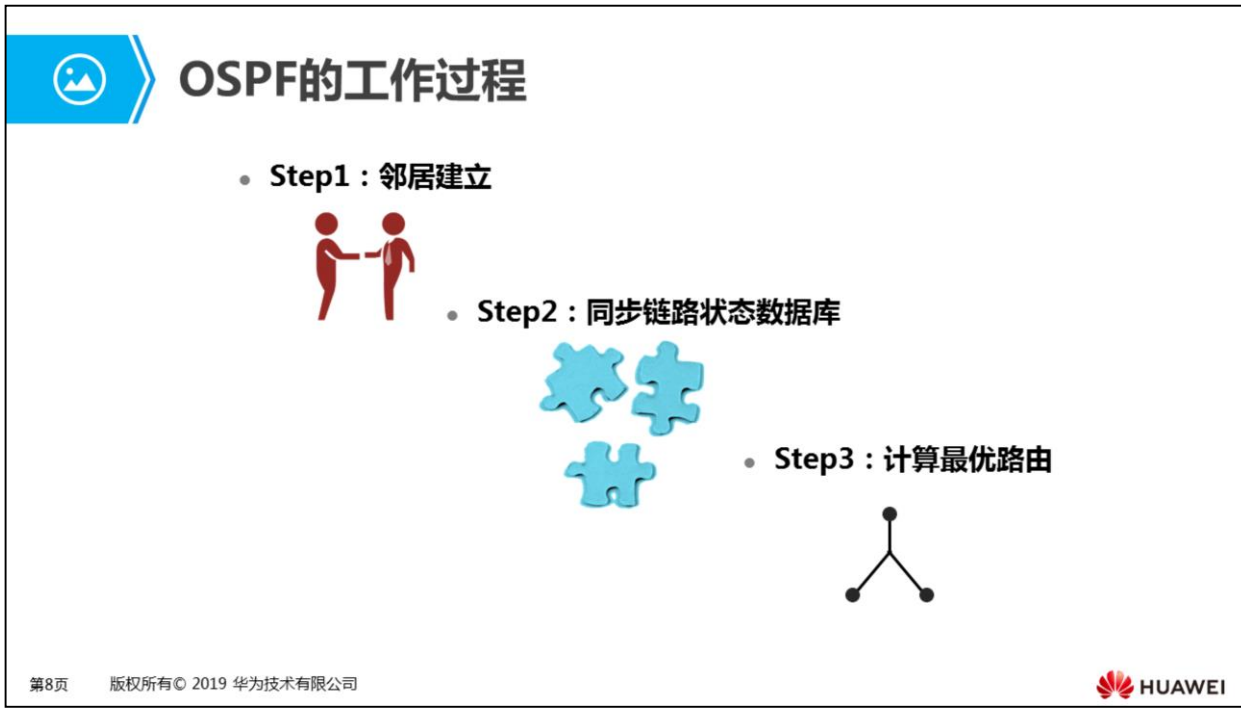
- 在 “收到更新”、“计算路由”、“发送更新”的路由收敛过程中，RIP的局限性在于路由器需要在完成路由计算之后才可以向邻居发送路由变化通知。如果将这个过程调整为：“收到更新”、“发送更新”、“计算路由”，即路由器从邻居收到路由更新后立刻向其他邻居路由器转发，然后再本地计算新的路由。这样的收敛方式可以大大降低全网路由收敛的时间。
- 因为RIP路由器仅从邻居路由器获取路由信息，所以对于非最优或者错误路由信息，RIP路由器并不能识别或屏蔽。解决此问题的关键最佳方式是路由器收集全网的信息，并基于这些信息独立计算路由。
- 基于跳数的度量方式并没有考虑数据包的链路转发延迟，如果采用以累积带宽为选路参考依据，可以更好的规避选择次优路径的风险。
- 与RIP这种距离矢量路由协议不同的链路状态路由协议是以怎样的方式来解决上述问题的呢？

链路状态路由协议OSPF

- 路由信息传递与路由计算分离。
- 基于SPF算法。
- 以“累计链路开销”作为选路参考值。

  HUAWEI

- 所谓Link State（链路状态）指的就是路由器的接口状态。在OSPF中路由器的某一接口的链路状态包含了如下信息：
  - 该接口的IP地址及掩码。
  - 该接口的带宽。
  - 该接口所连接的邻居。
  - ……
- OSPF作为链路状态路由协议，不直接传递各路由器的路由表，而传递链路状态信息，各路由器基于链路状态信息独立计算路由。
- 所有路由器各自维护一个链路状态数据库。邻居路由器间先同步链路状态数据库，再各自基于SPF（Shortest Path First）算法计算最优路由，从而提高收敛速度。
- 在度量方式上，OSPF将链路带宽作为选路时的参考依据。 “累计带宽”是一种要比“累积跳数”更科学的计算方式。
- RIP在大型网络中部署所面临的问题，OSPF都有相对应的解决办法，接下来详细地介绍下OSPF的实现过程。



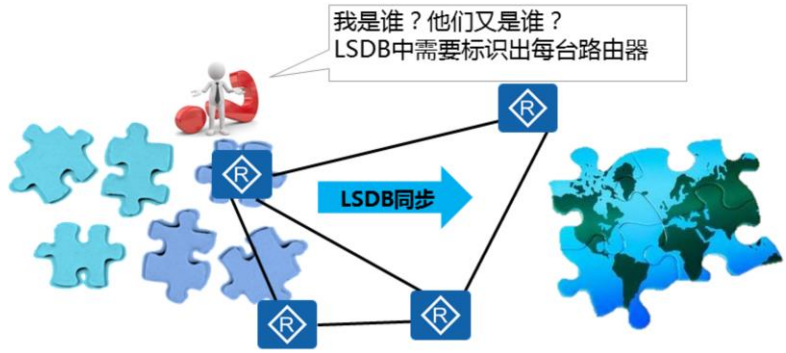
- 企业网络是由众多的路由器、交换机等网络设备之间互相连接组成的，类似一张地图。由于众多不同型号的路由器、不同类型的链路及其连接关系，造成了路由计算的复杂性。
- OSPF的路由计算过程可以简化描述为：
  - 路由器之间发现并建立邻居关系。
  - 每台路由器产生并向邻居泛洪链路状态信息，同时收集来自其他路由器链路状态信息，完成LSDB（Link State Database）的同步。
  - 每台路由器基于LSDB通过SPF算法，计算得到一棵以自己为根的SPT（Shortest Path Tree），再以SPT为基础计算去往各目的网络的最优路由，并形成路由表。
- 下面我们依照这三个步骤为主线，来学习和掌握OSPF的原理与实现。

目录

1. RIP在大型网络中部署面临的挑战
2. **OSPF基本工作原理**
   - 邻居建立过程
   - 链路状态信息
   - 报文类型及作用
   - LSDB同步过程
   - DR与BDR的选举及作用

Router ID

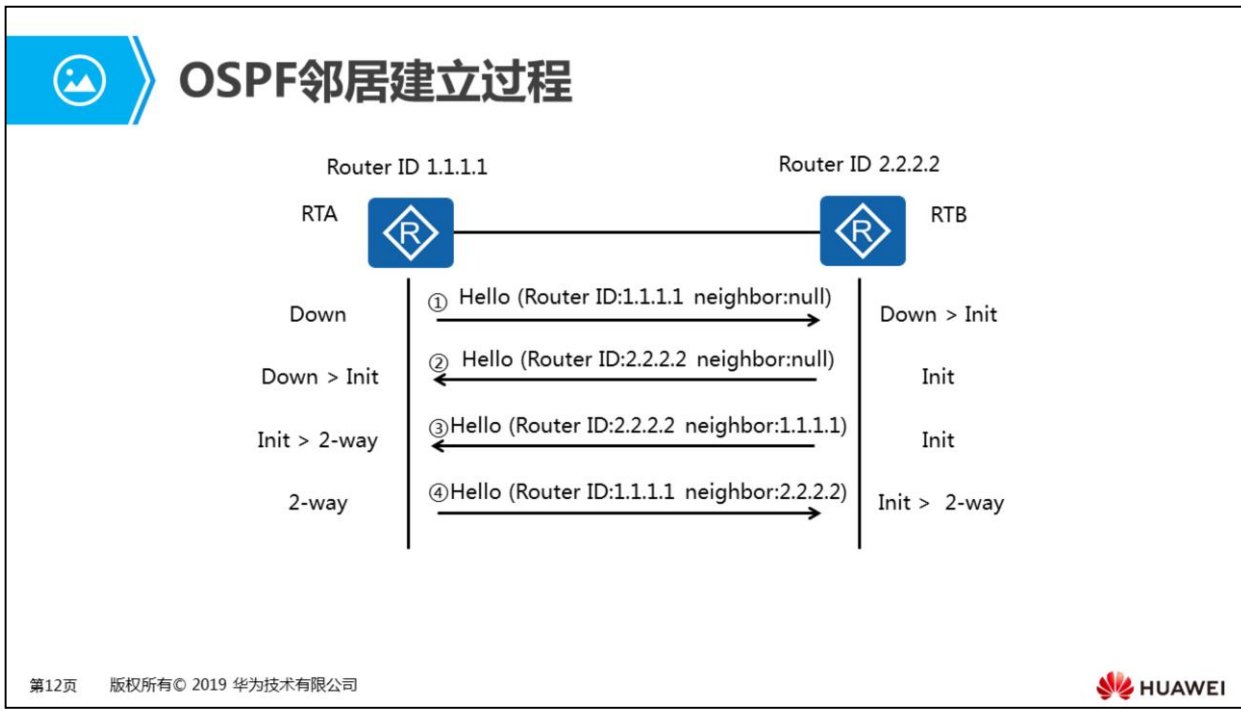
- 用于在自治系统中唯一标识一台运行OSPF的路由器，每台运行OSPF的路由器都有一个Router ID。



HUAWEI

- 企业网中的设备少则几台多则几十台甚至几百台，每台路由器都需要有一个唯一的ID用于标识自己。
- Router ID是一个32位的无符号整数，其格式和IP地址的格式是一样的，Router ID选举规则如下：
  - 手动配置OSPF路由器的Router ID（通常建议手动配置）；
  - 如果没有手动配置Router ID，则路由器使用Loopback接口中最大的IP地址作为Router ID；
  - 如果没有配置Loopback接口，则路由器使用物理接口中最大的IP地址作为Router ID。
- OSPF的路由器Router ID重新配置后，可以通过重置OSPF进程来更新Router ID。

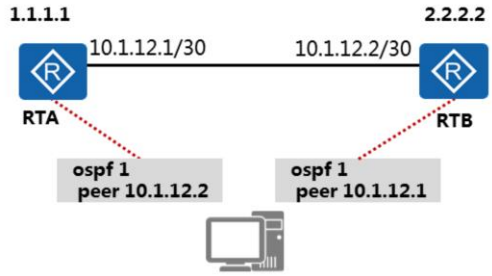
- OSPF路由器之间在交换链路状态信息之前，首先需要彼此建立邻居关系，通过Hello报文实现。
  - OSPF协议通过Hello报文可以让互联的路由器间自动发现并建立邻居关系，为后续可达性信息的同步作准备。
  - 在形成邻居关系过程中，路由器通过Hello报文完成一些参数的协商。
  - 邻居关系建立后，周期性的Hello报文发送还可以实现邻居保持的功能，在一定时间内没有收到邻居的Hello报文，则会中断路由器间的OSPF邻居关系。



- 状态含义：
  - Down：这是邻居的初始状态，表示没有从邻居收到任何信息。
  - Init：在此状态下，路由器已经从邻居收到了Hello报文，但是自己的Router ID不在所收到的Hello报文的邻居列表中，表示尚未与邻居建立双向通信关系。
  - 2-Way：在此状态下，路由器发现自己的Router ID存在于收到的Hello报文的邻居列表中，已确认可以双向通信。
- 邻居建立过程如下：
  - RTA和RTB的Router ID分别为1.1.1.1和2.2.2.2。当RTA启动OSPF后，RTA会发送第一个Hello报文。此报文中邻居列表为空，此时状态为Down，RTB收到RTA的这个Hello报文，状态置为Init。
  - RTB发送Hello报文，此报文中邻居列表为空，RTA收到RTB的Hello报文，状态置为Init。
  - RTB向RTA发送邻居列表为1.1.1.1的Hello报文，RTA在收到的Hello报文邻居列表中发现自己的Router ID，状态置为2-way。
  - RTA向RTB发送邻居列表为2.2.2.2的Hello报文，RTB在收到的Hello报文邻居列表中发现自己的Router ID，状态置为2-way。
- 因为邻居都是未知的，所以Hello报文的目的IP地址不是某个特定的单播地址。邻居从无到有，OSPF采用组播的形式发送Hello报文（目的地址224.0.0.5）。对于不支持组播的网络，OSPF路由器如何发现邻居呢？

发现并建立邻居 - 手动建立

- OSPF支持通过单播方式建立邻居关系。
- 对于不支持组播的网络可以通过手动配置实现邻居的发现与维护。



  HUAWEI

- 对于不支持组播的网络可以通过手动配置实现邻居的发现与维护。
- 当网络规模越来越大或者设备频繁更新，相关联的OSPF路由器都需要更改静态配置，手动更改配置的工作量变大且容易出错。除了特殊场景，一般情况下不适用手动配置的方式。
- OSPF路由器之间建立邻居关系是为了同步链路状态信息，接下来学习OSPF如何实现链路状态数据库的同步。

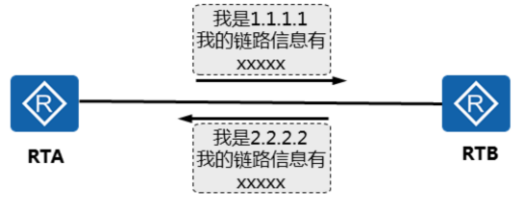
目录

1. RIP在大型网络中部署面临的挑战
2. **OSPF基本工作原理**
   - 邻居建立过程
   - **链路状态信息**
   - 报文类型及作用
   - LSDB同步过程
   - DR与BDR的选举及作用

HUAWEI

链路状态信息

- 链路信息主要包括：
  - 链路的类型；
  - 接口IP地址及掩码；
  - 链路上所连接的邻居路由器；
  - 链路的带宽（开销）。

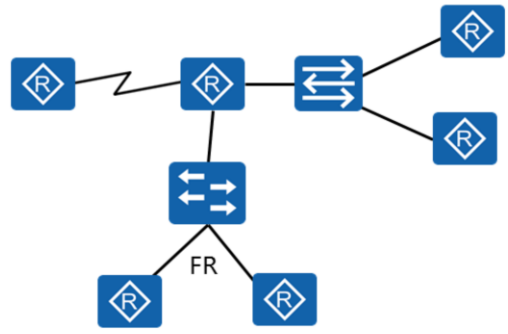


HUAWEI

- 区别于RIP路由器之间交互的路由信息，OSPF路由器同步的是最原始的链路状态信息，而且对于邻居路由器发来的链路状态信息，仅作转发。最终所有路由器都将拥有一份相同且完整的原始链路状态信息。
- 每台运行OSPF协议的路由器所描述的信息中都应该包括链路的类型、接口IP地址及掩码、链路上的邻居、链路的开销等信息。
- 路由器只需要知道目的网络号/掩码、下一跳、开销（接口IP地址及掩码、链路上的邻居、链路的开销）即可，为什么要有链路的类型呢？

丰富的数据链路层支持能力

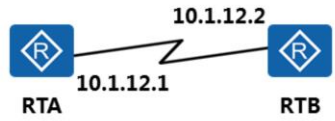
- 数据链路层协议类型多种多样，工作机制也各不相同。
- 为适配多种数据链路层协议，必须考虑各类链路层协议在组网时的应用场景。



- 网络技术的发展包含了设备、链路以及通信协议的发展。设备性能日趋提高，互联链路也从串行链路、ATM、帧中继发展到当前的以太网、xPON、SDH、MSTP、OTN等。技术升级不是一蹴而就的，而是一个循序渐进的过程。各种不同的物理链路各具特点，也正因为如此，一个成熟的路由协议必须能够根据不同物理链路特性进行适配。
- 下面将介绍OSPF是如何定义多种网络的。

网络类型 - P2P网络

- 仅两台路由互连。
- 支持广播、组播。

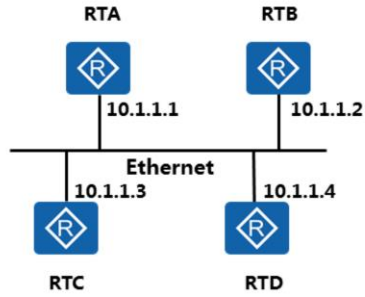


  HUAWEI

- OSPF划分了四种网络类型并以此来组成拓扑信息的一部分。
- P2P网络连接了一对路由器，广播、组播数据包都可以转发。
- P2P网络的例子：两台通过PPP（Point-to-Point Protocol）链路相连的路由器网络。

网络类型 - 广播型网络

- 两台或两台以上的路由器通过共享介质互连。
- 支持广播、组播。

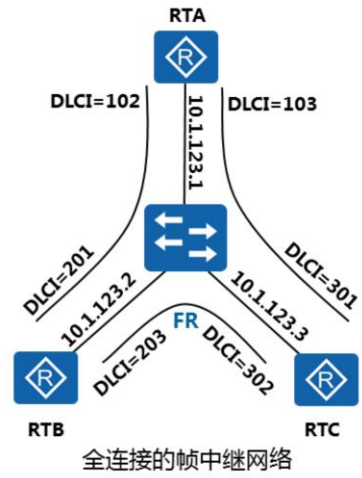


HUAWEI

- 广播型网络支持两台及两台以上的设备接入同一共享链路且可以支持广播、组播报文的转发，是OSPF最常见的网络类型。
- 广播型网络的例子：通过以太网链路相连的路由器网络。
- 同时因为一个广播型网络中存在多台设备，邻居关系建立以及链路信息同步方面，OSPF都有对应的特性来减少同一网络多台设备带来的不利影响。
- 以上两种网络类型是最常见的，此外，还有两种少见的网络类型。

网络类型 - NBMA网络

- 两台或两台以上路由器通过VC互连。
- 不支持广播、组播。



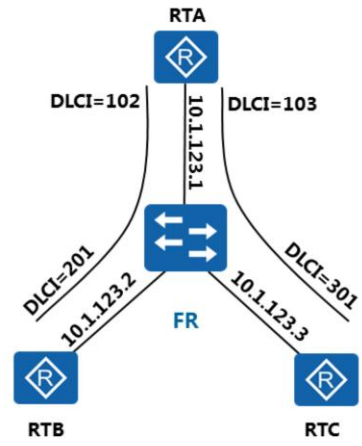
全连接的帧中继网络

HUAWEI

- 与广播型网络不同的是NBMA网络默认不支持广播与组播报文的转发。在NBMA网络上，OSPF模拟在广播型网络上的操作，但是每个路由器的邻居需要手动配置。
- NBMA（non-broadcast multiple access）型网络的例子：通过全互连的帧中继链路相连的路由器网络。
- 在现在的网络部署中，NBMA网络已经很少了。

网络类型 - P2MP网络

- 多个点到点网络的集合。
- 支持广播、组播。



非完全连接的帧中继网络

HUAWEI

- 将一个非广播网络看成是一组P2P网络，这样的非广播网络便成为了一个点到多点（P2MP）网络。在P2MP网络上，每个路由器的OSPF邻居可以使用反向地址解析协议（Inverse ARP）来发现。P2MP可以看作是多个P2P的集合，P2MP可以支持广播、组播的转发。
- 没有一种链路层协议默认属于P2MP类型网络，也就是说必须是由其他的网络类型强制更改为P2MP。常见的做法是将非完全连接的帧中继或ATM改为P2MP的网络。
- 此外OSPF的链路状态信息中的开销值是如何度量的呢？

OSPF的度量方式

- 某接口cost=参考带宽/实际带宽。
- 更改cost的两种方式：
  - 直接在接口下配置；
  - 修改参考带宽（所有路由器都需要修改，确保选路一致性）。

- RTA到达192.168.3.0/24的“累计cost”=G1’s cost+G3’s cost

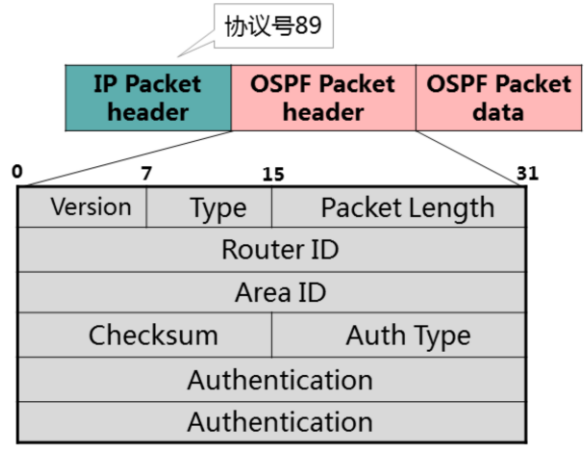
HUAWEI

- OSPF在计算接口的cost时，cost=参考带宽/实际带宽，默认参考带宽为100M。当计算结果有小数位时，只取整数位；结果小于1时，cost取1。
- 若需要调整接口cost值有两种方式：
  - 直接在接口下配置，需要注意的是，配置的cost是此接口最终的cost值，作用范围仅限于本接口。
  - 修改OSPF的默认参考带宽值，作用范围是本路由器使能OSPF的接口。建议参考整个网络的带宽情况建立参考基线，所有路由器修改相同的参考带宽值，从而确保选路的一致性。
- OSPF以“累计cost”为开销值，也就是流量从源网络到目的网络所经过所有路由器的出接口的cost总和，以RTA访问RTC Loopback 1接口192.168.3.3为例，其cost=G1’s cost+G3’s cost。
- 相比于RIP，OSPF的度量方式不仅考虑“跳数”，而且还考虑了“带宽”，比RIP更可靠的选择最优的转发路径。
- 那么OSPF路由器怎么表达链路状态信息并完成同步呢？

目录

1. RIP在大型网络中部署面临的挑战
2. **OSPF基本工作原理**
   - 邻居建立过程
   - 链路状态信息
   - **报文类型及作用**
   - LSDB同步过程
   - DR与BDR的选举及作用

OSPF协议报文头部



  HUAWEI

- RIP路由器之间是基于UDP 520的报文进行通信，OSPF也有其规定的通信标准。OSPF使用IP承载其报文，协议号为89。
- 在OSPF Packet部分，所有的OSPF报文均使用相同的OSPF报文头部：
  - Version ：对于当前所使用的OSPFv2，该字段的值为2。
  - Type：OSPF报文类型。
  - Packet length：表示整个OSPF报文的长度，单位是字节。
  - Router ID：表示生成此报文的路由器的Router ID。
  - Area ID：表示此报文需要被通告到的区域。
  - Checksum：校验字段，其校验的范围是整个OSPF报文，包括OSPF报文头部。
  - Auth Type：为0时表示不认证；为1时表示简单的明文密码认证；为2时表示加密（MD5）认证。
  - Authentication：认证所需的信息。该字段的内容随AuType的值不同而不同。
- OSPF的报文头部定义了OSPF路由器之间的通信的标准与规则，基于这个标准OSPF报文需要实现什么功能呢？

OSPF报文类型

Type	报文名称	报文功能
1	Hello	发现和维护邻居关系
2	Database Description	交互链路状态数据库摘要
3	Link State Request	请求特定的链路状态信息
4	Link State Update	发送详细的链路状态信息
5	Link State Ack	发送确认报文

- 思考：DD、LSR、LSU、LSAck报文都包含哪些信息？这么设计有什么好处？

  HUAWEI

- Type=1为Hello报文，用来建立和维护邻居关系，邻居关系建立之前，路由器之间需要进行参数协商。
- Type=2为数据库描述报文（DD），用来向邻居路由器描述本地链路状态数据库，使得邻居路由器识别出数据库中的LSA是否完整。
- Type=3为链路状态请求报文（LSR），路由器根据邻居的DD报文，判断本地数据库是否完整，如不完整，路由器把这些LSA记录进链路状态请求列表中，然后发送一个LSR给邻居路由器。
- Type=4为链路状态更新报文（LSU），用于响应邻居路由器发来的LSR，根据LSR中的请求列表，发送对应LSA给邻居路由器，真正实现LSA的泛洪与同步。
- Type=5为链路状态确认报文（LSAck），用来对收到的LSA进行确认，保证同步过程的可靠性。
- DD、LSR、LSU、LSAck与LSA的关系：
  - DD报文中包含LSA头部信息，包括LS Type、LS ID、Advertising Router 、LS Sequence Number、LS Checksum。
  - LSR中包含LS Type 、LS ID和Advertising Router 。
  - LSU中包含完整的LSA信息。
  - LSAck中包含LSA头部信息，包括LS Type、LS ID、Advertising Router、LS Sequence Number、LS Checksum。
- 五种报文可以高效地完成LSA的同步，那么实际的报文交互过程是什么呢？

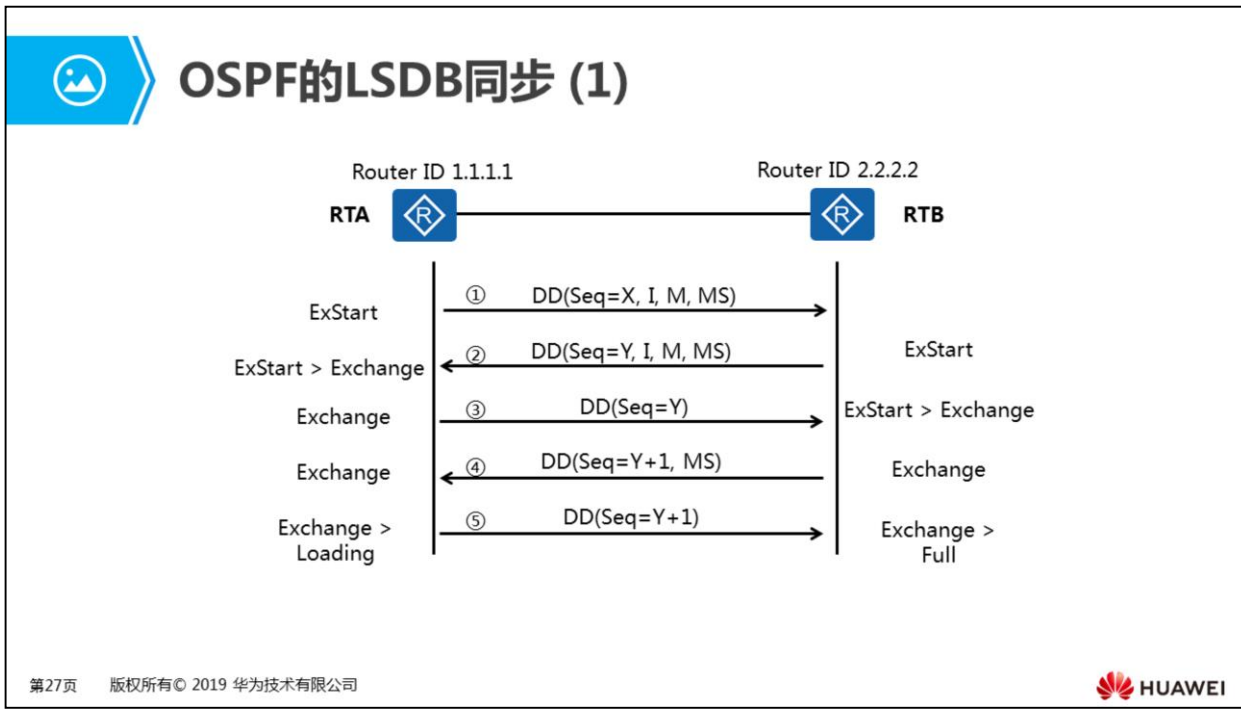
OSPF报文的功能需求

功能	实现分析
发现邻居与保持	Hello机制即可实现
LSA同步	双方互相发送LSA，完成同步； 同时同步速度更快，占用资源更少
可靠性	确保LSA同步过程的可靠性

- Hello机制动态发现并维护邻居前文已介绍，不再赘述。
- RIP设置了Request和Response两种报文来完成路由信息的同步，OSPF路由器之间为了完成LSA的同步，可以直接把本地所有LSA发给邻居路由器，但是邻居路由器直接同步LSA并不是最好的方式。
- 更快速、更高效的方式是先在邻居路由器之间传送关键信息，路由器基于这些关键信息识别出哪些LSA是没有的、哪些是需要更新的，然后向邻居路由器请求详细的LSA内容。对于OSPF来说，需要有比RIP更高效、更可靠的方式来完成路由器之间的信息同步。

目录

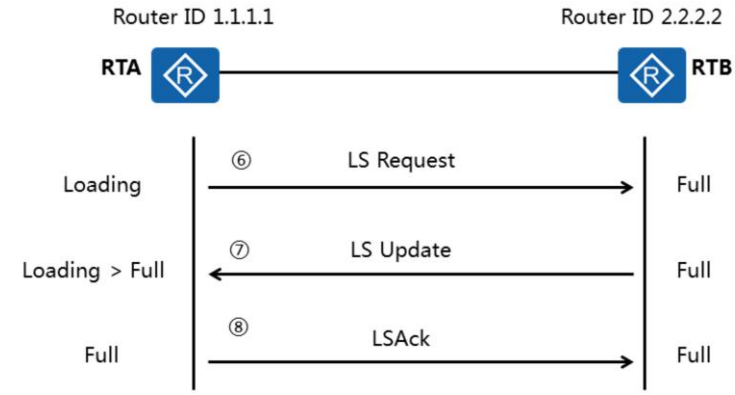
1. RIP在大型网络中部署面临的挑战
2. **OSPF基本工作原理**
   - 邻居建立过程
   - 链路状态信息
   - 报文类型及作用
   - LSDB同步过程
   - DR与BDR的选举及作用



- 状态含义：
  - ExStart：邻居状态变成此状态以后，路由器开始向邻居发送DD报文。Master/Slave关系是在此状态下形成的，初始DD序列号也是在此状态下确定的。在此状态下发送的DD报文不包含链路状态描述。
  - Exchange：在此状态下，路由器与邻居之间相互发送包含链路状态信息摘要的DD报文。
  - Loading：在此状态下，路由器与邻居之间相互发送LSR报文、LSU报文、LSAck报文。
  - Full：LSDB同步过程完成，路由器与邻居之间形成了完全的邻接关系。
- LSDB同步过程如下：
  - RTA和RTB的Router ID分别为1.1.1.1和2.2.2.2并且二者已建立了邻居关系。当RTA的邻居状态变为ExStart后，RTA会发送第一个DD报文。此报文中，DD序列号被随机设置为X，I-bit设置为1，表示这是第一个DD报文，M-bit设置为1，表示后续还有DD报文要发送，MS-bit设置为1，表示RTA宣告自己为Master。
  - 当RTB的邻居状态变为ExStart后，RTB会发送第一个DD报文。此报文中，DD序列号被随机设置为Y（I-bit=1，M-bit=1，MS-bit=1，含义同上）。由于RTB的Router ID较大，所以RTB将成为真正的Master。收到此报文后，RTA会产生一个Negotiation-Done事件，并将邻居状态从ExStart变为Exchange。

- 当RTA的邻居状态变为Exchange后，RTA会发送一个新的DD报文，此报文中包含了LSDB的摘要信息，序列号设置为RTB在步骤2中使用的序列号Y，I-bit=0，表示这不是第一个DD报文，M-bit=0，表示这是最后一个包含LSDB摘要信息的DD报文，MS-bit=0，表示RTA宣告自己为Slave。收到此报文后，RTB会产生一个Negotiation-Done事件，并将邻居状态从ExStart变为Exchange。
- 当RTB的邻居状态变为Exchange后，RTB会发送一个新的DD报文，此报文包含了LSDB的摘要信息，DD序列号设置为Y+1, MS-bit=1，表示RTB宣告自己为Master。
- 虽然RTA不需要发送新的包含LSDB摘要信息的DD报文，但是作为Slave，RTA需要对Master发送的每一个DD报文进行确认。所以，RTA向RTB发送一个新的DD报文，序列号为Y+1，该报文内容为空。发送完此报文后，RTA产生一个Exchange-Done事件，将邻居状态变为Loading。RTB收到此报文后，会将邻居状态变为Full（假设RTB的LSDB是最新最全的，不需要向RTA请求更新）。

OSPF的LSDB同步 (2)

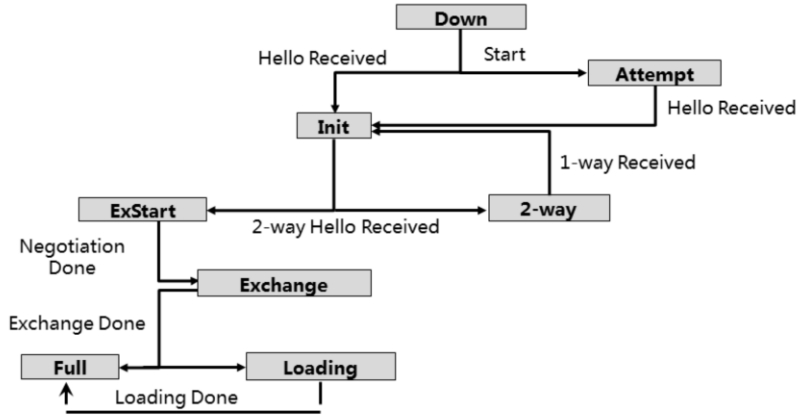


  HUAWEI

- RTA开始向RTB发送LSR报文，请求那些在Exchange状态下通过DD报文发现的、并且在本地LSDB中没有的链路状态信息。
- RTB向RTA发送LSU报文，LSU报文中包含了那些被请求的链路状态的详细信息。RTA在完成LSU报文的接收之后，会将邻居状态从Loading变为Full。
- RTA向RTB发送LSAck报文，作为对LSU报文的确认。RTB收到LSAck报文后，双方便建立起了完全的邻接关系。

- 从建立邻居关系到同步LSDB的过程较为复杂，错误的配置或设备链路故障都会导致无法完成LSDB同步。为了快速排障，最关键的是要理解不同状态之间切换的触发原因。

OSPF邻居状态机



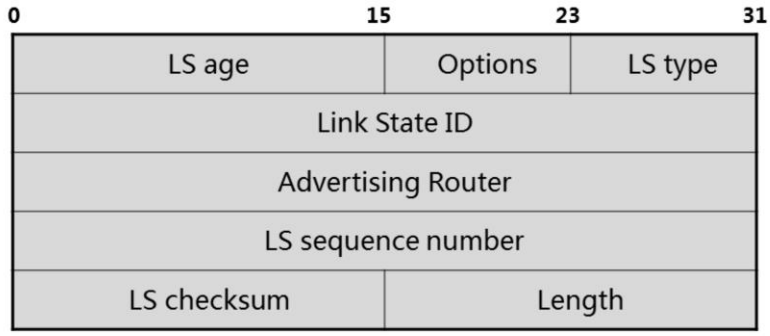
HUAWEI

- 这是形成邻居关系的过程和相关邻居状态的变换过程。
  - Down：这是邻居的初始状态，表示没有从邻居收到任何信息。在NBMA网络上，此状态下仍然可以向静态配置的邻居发送Hello报文，发送间隔为PollInterval，通常和Router DeadInterval间隔相同。
  - Attempt：此状态只在NBMA网络上存在，表示没有收到邻居的任何信息，但是已经周期性的向邻居发送报文，发送间隔为HelloInterval。如果Router DeadInterval间隔内未收到邻居的Hello报文，则转为Down状态。
  - Init：在此状态下，路由器已经从邻居收到了Hello报文，但是自己不在所收到的Hello报文的邻居列表中，表示尚未与邻居建立双向通信关系。在此状态下的邻居要被包含在自己所发送的Hello报文的邻居列表中。
  - 2-Way Received：此事件表示路由器发现与邻居的双向通信已经开始（发现自己在邻居发送的Hello报文的邻居列表中）。Init状态下产生此事件之后，如果需要和邻居建立邻接关系则进入ExStart状态，开始数据库同步过程，如果不能与邻居建立邻接关系则进入2-Way。
  - 2-Way：在此状态下，双向通信已经建立，但是没有与邻居建立邻接关系。这是建立邻接关系以前的最高级状态。
  - 1-Way Received：此事件表示路由器发现自己没有在邻居发送Hello报文的邻居列表中，通常是由于对端邻居重启造成的。
  - ExStart：这是形成邻接关系的第一个步骤，邻居状态变成此状态以后，路由器开始向邻居发送DD报文。主从关系是在此状态下形成的；初始DD序列号是在此状态下决定的。在此状态下发送的DD报文不包含链路状态描述。

- Exchange：此状态下路由器相互发送包含链路状态信息摘要的DD报文，描述本地LSDB的内容。
- Loading：相互发送LS Request报文请求LSA，发送LS Update通告LSA。
- Full：两台路由器的LSDB已经同步。

LSA头部

- LSA是OSPF链路状态信息的载体。



  HUAWEI

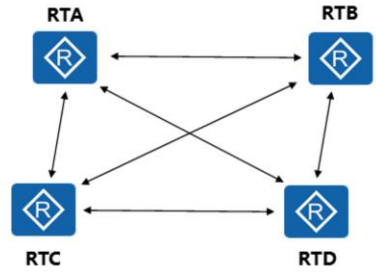
- LSA（Link State Advertisement）是路由器之间链路状态信息的载体。LSA是LSDB的最小组成单位，也就是说LSDB由一条条LSA构成的。
- 所有的LSA都拥有相同的头部，关键字段的含义如下：
  - LS age：此字段表示LSA已经生存的时间，单位是秒。
  - LS type：此字段标识了LSA的格式和功能。常用的LSA类型有五种。
  - Link State ID：此字段是该LSA所描述的那部分链路的标识，例如Router ID等。
  - Advertising Router：此字段是产生此LSA的路由器的Router ID。
  - LS sequence number：此字段用于检测旧的和重复的LSA。
- LS type，Link State ID和Advertising Router的组合共同标识一条LSA。
- LSDB中除了自己生成的LSA，另一部分是从邻居路由器接收的。邻居路由器之间相互更新LSA必然需要一个“通道”。

目录

1. RIP在大型网络中部署面临的挑战
2. **OSPF基本工作原理**
   - 邻居建立过程
   - 链路状态信息
   - 报文类型及作用
   - LSDB同步过程
   - DR与BDR的选举及作用

MA网络中的问题

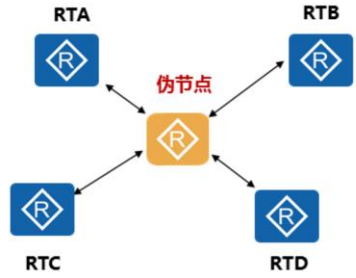
- n×(n−1)/2个邻接关系，管理复杂。
- 重复的LSA泛洪，造成资源浪费。



- 问题引出，在运行OSPF的MA网络包括广播型和NBMA网络，会存在两个问题：
  - 在一个有n个路由器的网络，会形成(n×(n−1))/2个邻接关系。
  - 邻居间LSA的泛洪扩散混乱，相同的LSA会被复制多份，如RTA向其邻居RTB、RTC、RTD分别发送一份自己的LSA，RTB与RTC、RTC与RTD、RTB与RTD之间也会形成邻居关系，也会发送RTA的LSA。
- 这样的工作效率显然是很低的，消耗资源的。作为高级的路由协议，OSPF是怎样解决这些问题的呢？

DR与BDR作用

- 减少邻接关系。
- 降低OSPF协议流量。



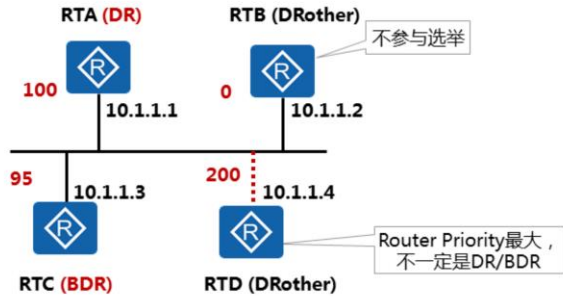
- 思考：DR的单点故障怎么解决？

HUAWEI

- DR（Designated Router）即指定路由器，其负责在MA网络建立和维护邻接关系并负责LSA的同步。
- DR与其他所有路由器形成邻接关系并交换链路状态信息，其他路由器之间不直接交换链路状态信息。这样就大大减少了MA网络中的邻接关系数量及交换链路状态信息消耗的资源。
- DR一旦出现故障，其与其他路由器之间的邻接关系将全部失效，链路状态数据库也无法同步。此时就需要重新选举DR，再与非DR路由器建立邻接关系，完成LSA的同步。为了规避单点故障风险，通过选举备份指定路由器BDR，在DR失效时快速接管DR的工作。
- 伪节点是一个虚拟设备节点，其功能需要某台路由器来承载，下面将介绍DR/BDR的选举规则。

DR与BDR选举

- 选举规则：DR/BDR的选举是基于接口的。
  - 接口的DR优先级越大越优先。
  - 接口的DR优先级相等时，Router ID越大越优先。



邻居与邻接关系

网络类型	是否和邻居建立邻接关系
P2P	是
Broadcast	DR与BDR、DRother建立邻接关系 BDR与DR、DRother建立邻接关系 DRother之间只建立邻居关系
NBMA	
P2MP	是

- 邻居（Neighbor）关系与邻接（Adjacency）关系是两个不同的概念。OSPF路由器之间建立邻居关系后，进行LSDB同步，最终形成邻接关系。
- 在P2P网络及P2MP网络上，具有邻居关系的路由器之间会进一步建立邻接关系。
- 在广播型网络及NBMA网络上，非DR/BDR路由器之间只能建立邻居关系，不能建立邻接关系，非DR/BDR路由器与DR/BDR路由器之间会建立邻接关系，DR与BDR之间也会建立邻接关系。
- 邻接关系建立完成，意味着LSDB已经完成同步，接下来OSPF路由器将基于LSDB使用SPF算法计算路由。

思考题

1. 下列哪些选项属于OSPF报文类型？
   A. Hello
   B. Database Description
   C. Link State Request
   D. Link State DD
   E. Link State Advertisement
2. OSPF的网络类型包括（　　）？

- 答案：ABC。
- 答案：P2P网络、P2MP网络、广播型网络、NBMA网络。


谢 谢
www.huawei.com

OSPF域内路由

HUAWEI

前言

- 本课程主要介绍OSPF如何计算区域内路由，内容主要包括如何使用Router-LSA和Network-LSA描述拓扑信息和路由信息，以及如何构建最短路径树。

目标

- 学完本课程后，您将能够：
  - 熟悉Router-LSA的内容及作用
  - 熟悉Network-LSA的内容及作用
  - 理解SPF算法

目录

1. **Router-LSA**
2. Network-LSA
3. SPF计算过程

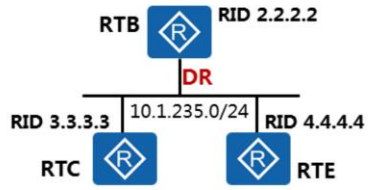
Router-LSA描述P2P网络


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

Type      : Router          /LSA类型
Ls id     : 1.1.1.1         /链路状态ID
Adv rtr   : 1.1.1.1         /产生此LSA的路由器Router ID

拓扑信息
  Link ID: 3.3.3.3          /邻居路由器的Router ID
  Data   : 10.1.13.1        /宣告该Router LSA的路由器接口的IP地址
  Link Type: P-2-P
  Metric : 48
路由信息
  Link ID: 10.1.13.0        /该Stub网络的IP地址
  Data   : 255.255.255.0    /该Stub网络的网络掩码
  Link Type: StubNet
  Metric : 48               /开销值
  Priority : Low
```


HUAWEI

- 每台OSPF路由器使用一条Router-LSA描述本区域内的链路状态信息。LSA头部的三个字段含义如下：
  - Type：LSA类型，Router-LSA是一类LSA。
  - LS id：链路状态ID。
  - Adv rtr：产生此Router-LSA的路由器Router ID。
- 一条Router-LSA可以描述多条链接，每条链接描述信息由Link ID，Data，Link Type和Metric组成，其关键字含义如下：
  - Type：链接类型（并非OSPF定义的四种网络类型），Router LSA描述的链接类型主要有：
    - Point-to-Point：描述一个从本路由器到邻居路由器之间的点到点链接，属于拓扑信息。
    - TransNet：描述一个从本路由器到一个Transit网段（例如MA网段或者NBMA网段）的链接，属于拓扑信息。
    - StubNet：描述一个从本路由器到一个Stub网段（例如Loopback接口）的链接，属于路由信息。
  - Link ID：此链接的对端标识，不同链接类型的Link ID表示的意义也不同。
  - Data：用于描述此链接的附加信息，不同的链接类型所描述的信息也不同。
  - Metric：描述此链接的开销。

Router-LSA描述MA网络或NBMA网络




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

 Type      : Router      //LSA类型
 Ls id     : 3.3.3.3     //链路状态ID
 Adv rtr   : 3.3.3.3     //产生此LSA的路由器的Router ID

* Link ID: 10.1.235.2  //DR的接口IP地址
  Data   : 10.1.235.3  //宣告该Router LSA的路由器接口的IP地址
  Link Type: TransNet
  Metric : 1
```

拓扑信息

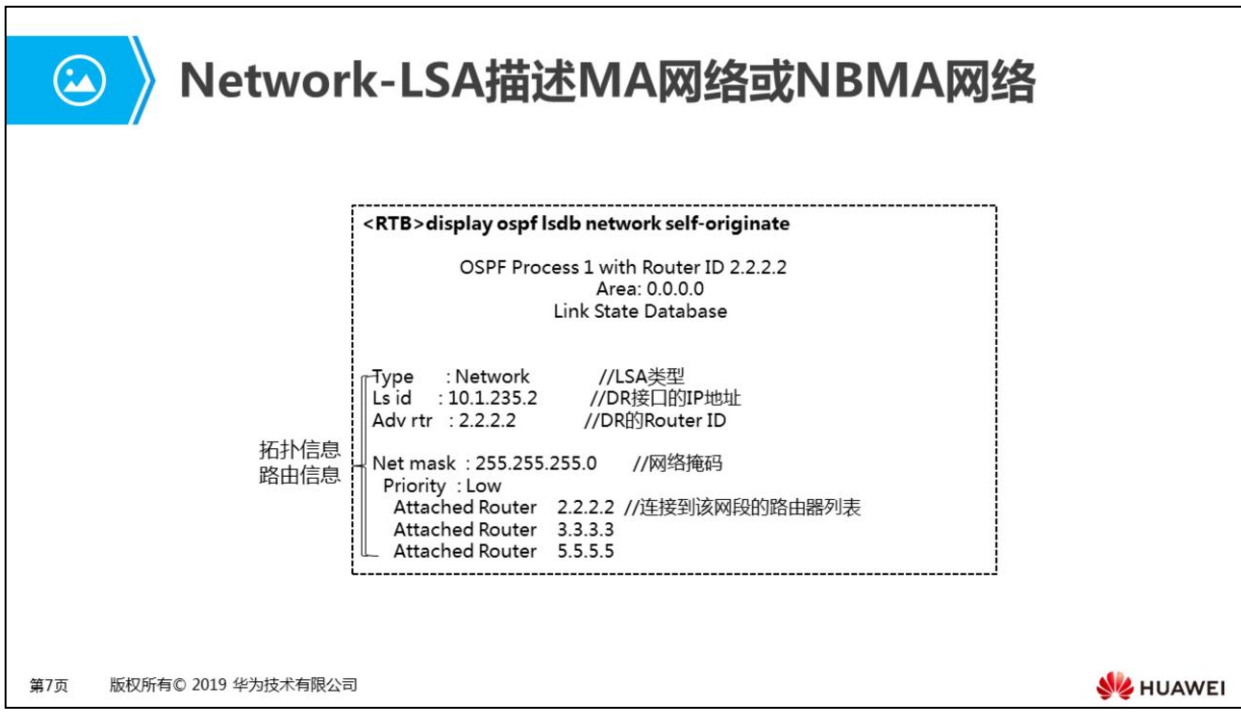
- 思考： 网络号/掩码在哪里？

HUAWEI

- 在描述MA或NBMA网络类型的Router-LSA中，Link ID为DR的接口IP地址，Data为本地接口的IP地址。
- 如图所示，RTB、RTC、RTE之间通过以太链路互连，以RTC产生的LSA为例，Link ID为DR的接口IP地址（10.1.235.2），Data为本地路由器连接此MA网络的接口IP地址（10.1.235.3），Link Type为TransNet，Metric表示到达DR的开销值。
- TransNet描述的链接中仅包括与DR的连接关系及开销，没有网络号/掩码及共享链路上其他路由器的任何信息。

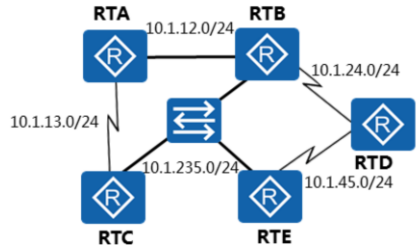
目录

1. Router-LSA
2. **Network-LSA**
3. SPF计算过程



- MA共享网段或NBMA共享网段中的网络号/掩码及路由器间的链接关系，通过Network-LSA来呈现。
- 在Network-LSA中关键字的含义如下：
  - Type：LSA类型，Network-LSA是二类LSA。
  - LS id：DR的接口IP地址。
  - Adv rtr：产生此Network-LSA的路由器Router ID，即DR的Router ID。
  - Net mask：该网段的网络掩码。
  - Attached Router：连接到该网段的路由器列表，呈现了此网段的拓扑信息。
- 基于上述字段表达的信息，Ls id和Net mask做与运算，即可得出该网段的IP网络号，另外，从DR路由器到其所连接的路由器的开销为0。
- 从Attached Router部分可以看出，2.2.2.2、3.3.3.3、5.5.5.5共同连接到该共享MA网段中，DR路由器为2.2.2.2，网络号10.1.235.0，掩码255.255.255.0。

OSPF区域内LSDB



<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

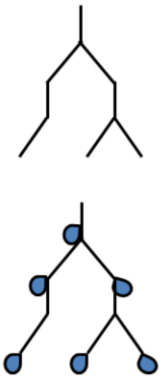
- 如图所示，五台路由器互连并运行OSPF协议。以RTA的LSDB为例，其中包括了五个路由器产生的Router-LSA，以及两个广播型网络中产生的Network-LSA。
- LSDB：链路状态数据库。

目录

1. Router-LSA
2. Network-LSA
3. **SPF计算过程**

SPF算法

- Phase 1：构建SPF树。
  - 根据Router-LSA和Network-LSA中的拓扑信息，构建SPF树干。
  - Phase 2：计算最优路由。
  - 基于SPF树干和Router-LSA、Network- LSA中的路由信息，计算最优路由。

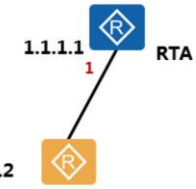


- 在一类LSA和二类LSA中，包括了拓扑信息和路由信息。
- OSPF将依据SPF算法和各类LSA进行最短路径树的计算：
  - Phase 1：依据一类LSA中的Point to Point，TransNet以及二类LSA，构建SPF树。
  - Phase 2：依据一类LSA中的Stub以及二类LSA，计算最优路由。

构建SPF树 (1)

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

Type      : Router
Ls id     : 1.1.1.1
Adv rtr   : 1.1.1.1
 *Link ID: 10.1.12.2
  Data   : 10.1.12.1
  Link Type: TransNet
  Metric : 1
 *Link ID: 3.3.3.3
  Data   : 10.1.13.1
  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
```

候选列表	候选总开销	父节点
10.1.12.2	1	1.1.1.1
3.3.3.3	48	1.1.1.1



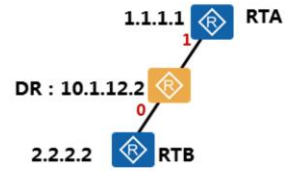
HUAWEI

- OSPF路由器将分别以自身为根节点计算最短路径树。
- 以RTA为例，计算过程如下：
  - RTA将自己添加到最短路径树的树根位置，然后检查自己生成的Router-LSA，对于该LSA中所描述的每一个连接，如果不是一个Stub连接，就把该连接添加到候选列表中，分节点的候选列表为Link ID，对应的候选总开销为本LSA中描述的Metric值和父节点到达根节点开销之和。
  - 根节点RTA的Router-LSA中存在TransNet中Link ID为10.1.12.2 Metric=1和P-2-P中Link ID为3.3.3.3 Metric=48的两个连接，被添加进候选列表中。
  - RTA将候选列表中候选总开销最小的节点10.1.12.2移到最短路径树上，并从候选列表中删除。

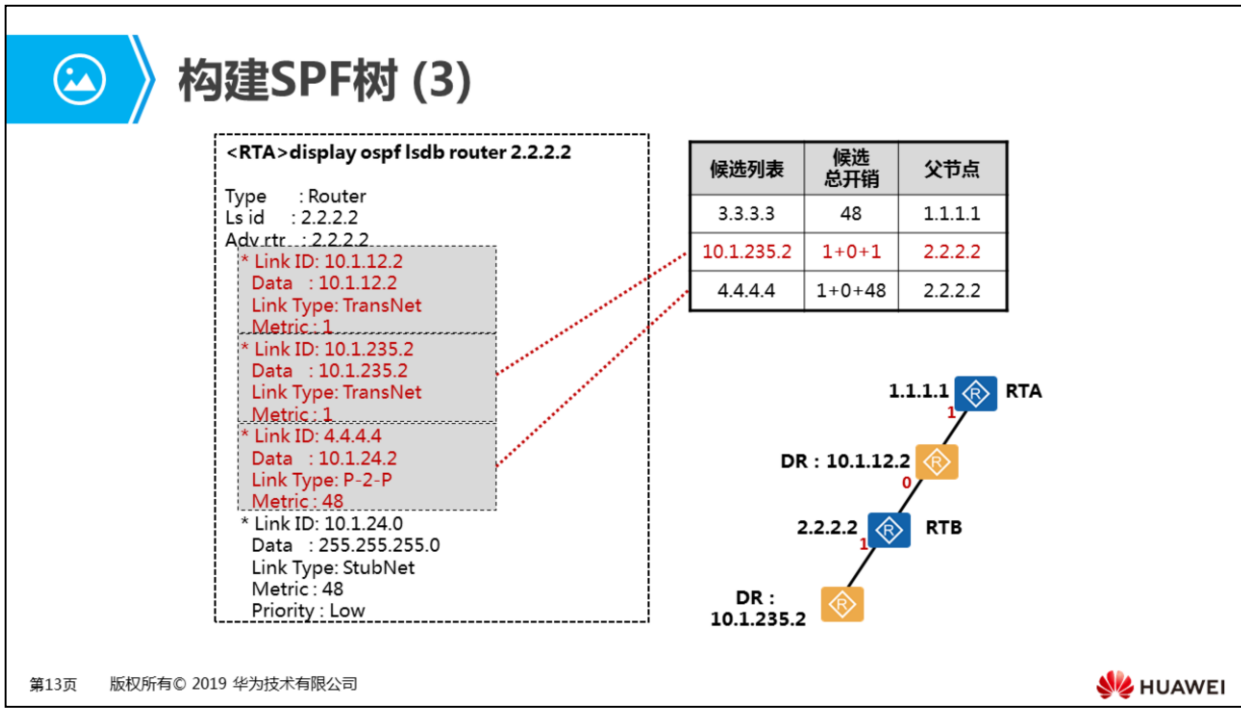
构建SPF树 (2)

```
<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
```

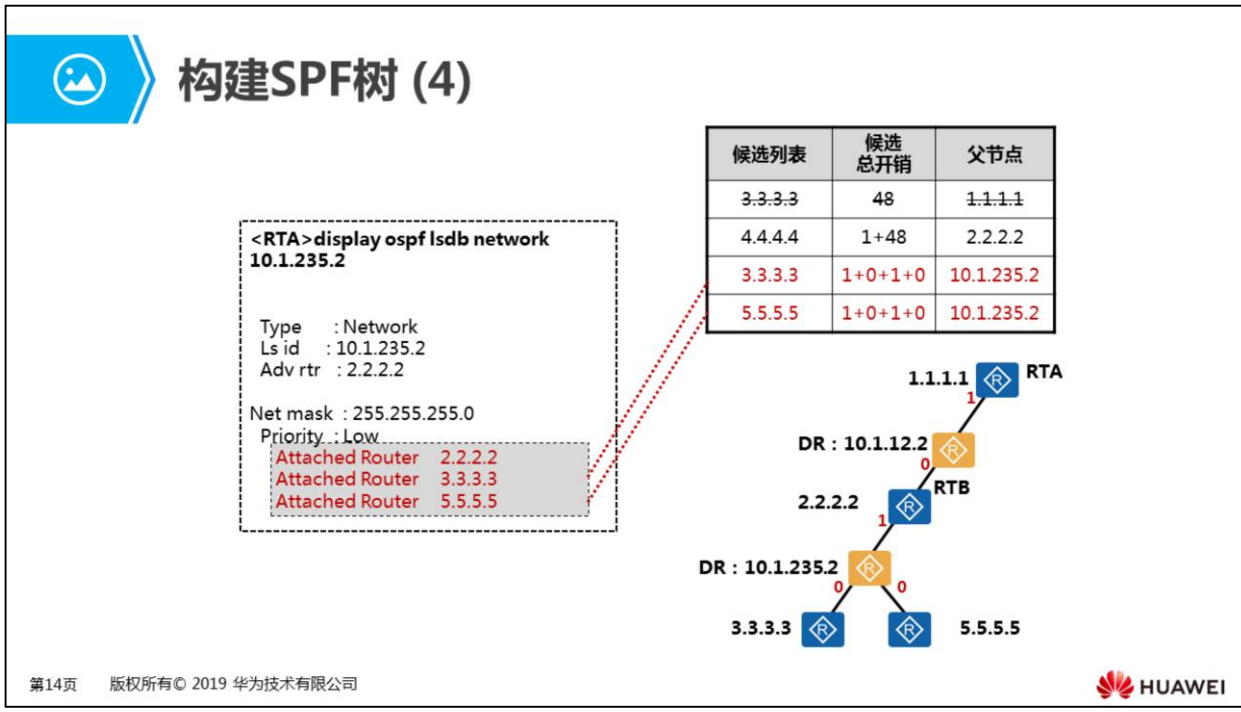
候选列表	候选总开销	父节点
3.3.3.3	48	1.1.1.1
2.2.2.2	1+0	10.1.12.2



- DR被加入到SPF中，接下来检查Ls id为10.1.12.2的Network-LSA。如果LSA中所描述的分节点在最短路径树上已经存在，则忽略该分节点。
- 如图所示，在Attached Router部分：
  - 节点1.1.1.1被忽略，因为1.1.1.1已经在最短路径树上。
  - 将节点2.2.2.2，Metric=0，父节点到根节点的开销为1，所以候选总开销为1，加入候选列表。
  - 候选节点列表中有两个候选节点，选择候选总开销最小的节点2.2.2.2加入最短路径树并从候选列表中删除。



- 节点2.2.2.2新添加进最短路径树上，此时继续检查Ls id为2.2.2.2的Router-LSA：
  - 第一个TransNet连接中，Link ID为10.1.12.2，此节点已经在最短路径树上，忽略。
  - 第二个TransNet连接中，Link ID为10.1.235.2，Metric=1，父节点到根节点的开销为1，候选总开销为2，加入候选列表。
  - 第三个P-2-P连接中，Link ID为4.4.4.4，Metric=48，父节点到根节点的开销为1，候选总开销为49，加入候选列表。
- 候选节点列表中有三个候选节点，选择候选总开销最小的节点10.1.235.2加入最短路径树并从候选列表中删除。

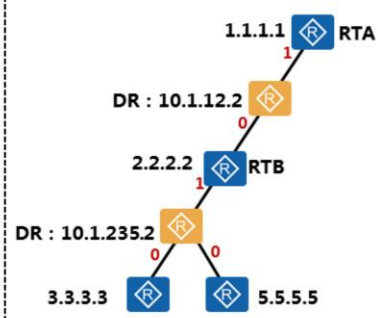


- DR被加入到SPF中，接下来检查Ls id为10.1.235.2的Network-LSA。
- 如图所示，在Attached Router部分：
  - 节点2.2.2.2被忽略，因为2.2.2.2已经在最短路径树上。
  - 将节点3.3.3.3，Metric=0，父节点到根节点的开销为2，候选总开销为2，加入候选列表。（如果在候选列表中出现两个节点ID一样但是到根节点的开销不一样的节点，则删除到根节点的开销大的节点。所以删除节点3.3.3.3 累计开销为48的候选项)。
  - 将节点5.5.5.5，Metric=0，父节点到根节点的开销为2，候选总开销为2，加入候选列表。
  - 候选节点列表中有三个候选节点，选择候选总开销最小的节点3.3.3.3和5.5.5.5加入最短路径树并从候选列表中删除。

构建SPF树 (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
```

候选列表	候选总开销	父节点
4.4.4.4	1+48	2.2.2.2



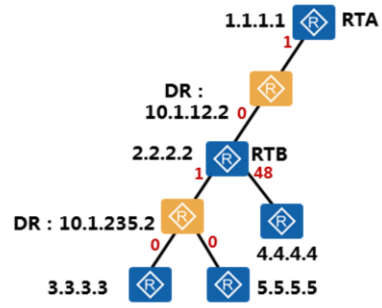
HUAWEI

- 节点3.3.3.3和5.5.5.5新添加进最短路径树上，此时继续检查Ls id分别为3.3.3.3和5.5.5.5的Router-LSA。
- Ls id为3.3.3.3的LSA：
  - Link ID为10.1.235.2的节点已经在最短路径树上，忽略。
  - Link ID为1.1.1.1的节点已经在最短路径树上，忽略。

构建SPF树 (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
```

候选列表	候选总开销	父节点
4.4.4.4	1+48	2.2.2.2
~~4.4.4.4~~	~~1+0+1+0+48~~	~~5.5.5.5~~

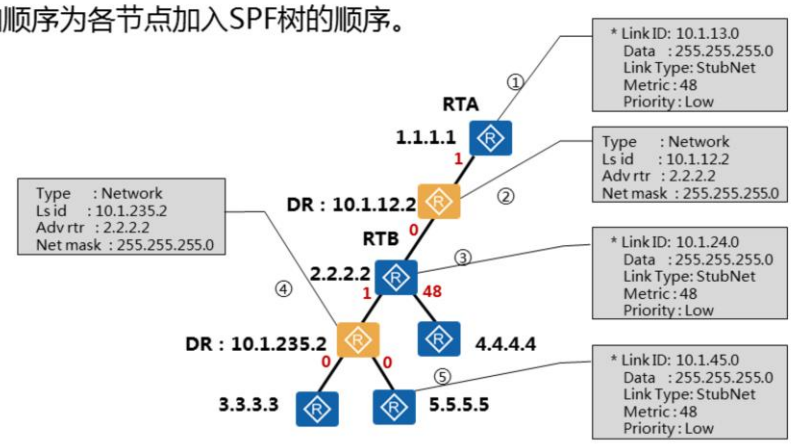


HUAWEI

- Ls id为5.5.5.5的LSA：
  - Link ID为10.1.235.2的节点已经在最短路径树上，忽略。
  - Link ID为4.4.4.4的P-2-P连接，Metric=48，父节点到根节点的开销为2，候选总开销为50。因为节点4.4.4.4已经在候选列表中出现，且候选总开销为49。49<50，所以子节点4.4.4.4的父节点选择2.2.2.2。
- 至此，再通过命令display ospf lsdb router 4.4.4.4发现，LSA中的连接所描述的相邻节点都已经添加到了SPF树中。
- 此时候选列表为空，完成SPF计算，其中10.1.12.2和10.1.235.2是虚节点（DR）。

计算最优路由

- 从根节点开始依次添加各节点LSA中的路由信息。
- 添加顺序为各节点加入SPF树的顺序。



  HUAWEI

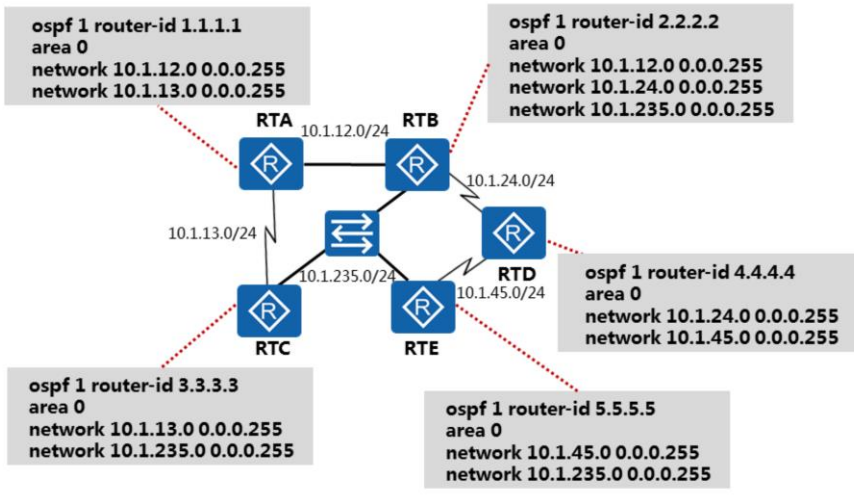
- 第二阶段根据Router LSA中的Stub、Network LSA中的路由信息，完成最优路由的计算。
- 从根节点开始，依次添加LSA中的路由信息（添加顺序按照每个节点加入SPF树的顺序）：
  - 1.1.1.1（RTA）的Router LSA中，共1个Stub连接，网络号/掩码10.1.13.0/24，Metric=48；
  - 10.1.12.2（DR）的Network LSA中，网络号/掩码10.1.12.0/24，Metric=1+0=1；
  - 2.2.2.2（RTB）的Router LSA中，共1个Stub连接，网络号/掩码10.1.24.0/24，Metric=1+0+48=49；
  - 10.1.235.2（DR）的Network LSA中，网络号/掩码10.1.235.0/24，Metric=1+0+1=2；
  - 3.3.3.3（RTC）的Router LSA中，共1个Stub连接，网络号/掩码10.1.13.0/24，已在RTA上，忽略；
  - 5.5.5.5（RTE）的Router LSA中，共1个Stub连接，网络号/掩码10.1.45.0/24，Metric=1+0+0+1+48=50；
  - 4.4.4.4（RTD）的Router LSA中，共2个Stub连接，网络号/掩码10.1.24.0/24，已在RTB上，忽略；网络号/掩码10.1.45.0/24，已在RTE上，忽略。

查看OSPF路由表

```
<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
```


- 经历上述两个阶段的计算，RTA生成的OSPF路由如上图所示。
- 经过OSPF优选后的路由并不一定会安装进系统路由表，因为路由器还可以通过其他协议获得路由，通过不同方式获得的路由需要进行优先级比较。

单区域OSPF配置实现



HUAWEI

查看OSPF邻居状态

```
<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
----------------------------------------------------------------------------
```


- 以RTA为例，RTA分别和RTB、RTC建立了邻接关系。

思考题

1. Router-LSA中主要包含哪几种链路类型？
2. 经过SPF算法计算后，被认为是最优的OSPF路由是否一定会被放入路由器的路由表中？

HUAWEI

- 答案：P2P、TransNet、StubNet、vlink。
- 答案：不一定，路由器可能通过多种路由协议获得同一路由前缀的路由信息，还需要通过路由优先级比较确定通过哪个路由协议获得的路由会放入路由表。


谢 谢
www.huawei.com

OSPF域间路由

HUAWEI

前言

- 随着网络规模不断扩大，结构也日趋复杂，路由器完成路由计算所消耗的内存、CPU资源也越来越多。
- 另外，网络发生故障的可能性也随之增加，如果区域内某处发生故障，整个区域内的路由器都要重新计算路由，这将大大增加路由器的负担，降低网络运行的稳定性。
- 面对单区域过大可能带来的问题，OSPF协议又将如何应对呢？

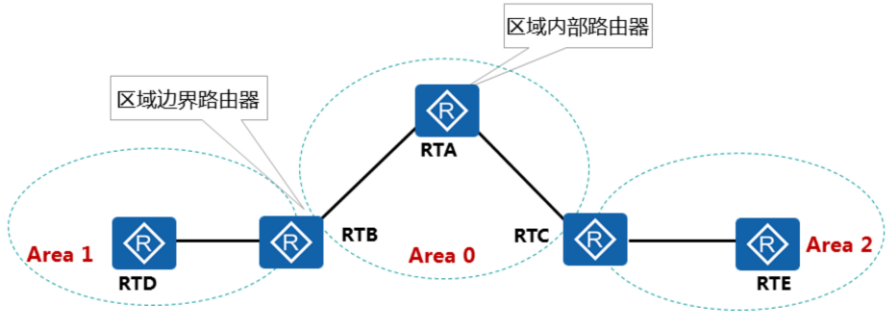
目标

- 学完本课程后，您将能够：
  - 熟悉区域间路由传递过程
  - 理解区域间防环机制
  - 掌握虚连接配置过程

目录

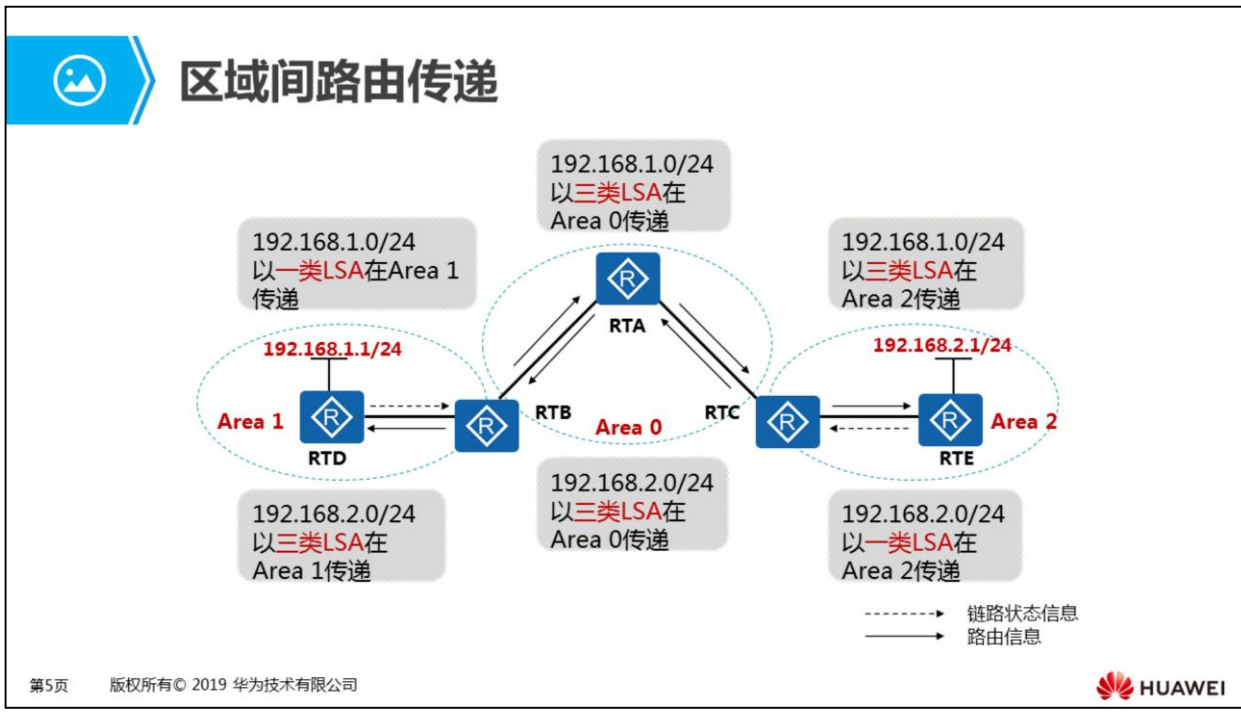
1. **区域间路由计算过程**
2. 区域间路由防环机制
3. 虚连接的作用及配置

区域划分



HUAWEI

- OSPF采用划分区域的方式，将一个大网络划分为多个相互连接的小网络。每个区域内的设备只需同步所在区域内的链路状态数据库，一定程度上降低内存及CPU的消耗。
- 划分区域后，根据路由器所连接区域的情况，可划分两种路由器角色：
  - 区域内部路由器（Internal Router）：该类设备的所有接口都属于同一个OSPF区域。
  - 区域边界路由器（Area Border Router）：该类设备接口分别连接两个及两个以上的不同区域。
- 区域内部路由器维护本区域内的链路状态信息并计算区域内的最优路径。
- 那么不同区域间是如何进行通信的呢？



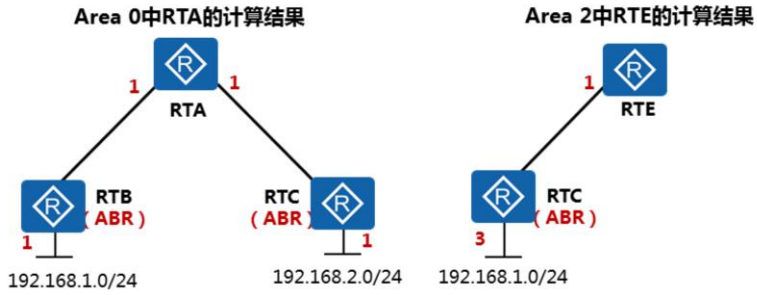
- 区域边界路由器作为区域间通信的桥梁，同时维护所连接多个区域的链路状态数据库。
- ABR将一个区域内的链路状态信息转化成路由信息，然后发布到邻居区域。
- 链路状态信息转换成路由信息其实就是将一类和二类LSA转化成三类LSA的过程。注意，区域间的路由信息在ABR上是双向传递的。
- 如图所示，以Area 1中RTD上的192.168.1.0/24的网络为例，其对应的一类LSA在Area 1中同步；作为Area 1和Area 0之间ABR的RTB负责将192.168.1.0/24的一类LSA转换成三类LSA并将此三类LSA发送到Area 0。作为Area 0和Area 2之间ABR的RTC，又重新生成一份三类LSA发送到Area 2中，至此全OSPF区域内都收到192.168.1.0/24的路由信息。RTE上192.168.2.0/24的路由信息同步过程也是这样。

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          //三类LSA
Ls id         : 192.168.1.0      //目的网段地址
Adv rtr       : 2.2.2.2          //产生此三类LSA的Router ID
Ls age        : 86
Len           : 28
Options       :  E
seq#          : 80000001
chksum        : 0x7c6d
Net mask      : 255.255.255.0    //网络掩码
Tos 0         metric: 1          //开销值
Priority      : Low
```


- Network-Summary-LSA（三类LSA）中主要包括以下内容：
  - Ls id：目的网段地址。
  - Adv rtr：ABR的Router ID。
  - Net mask：目的网段的网络掩码。
  - Metric：ABR到达目的网段的开销值。
- 区域内路由器接收描述其他区域网络信息的三类LSA后，OSPF路由器又是怎么基于三类LSA来计算出区域间路由的呢？

区域间路由计算



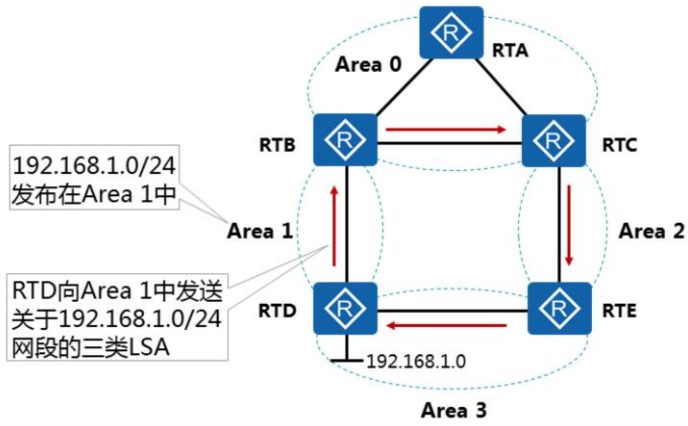
HUAWEI

- ABR产生的三类LSA将用于计算区域间路由。
  - 根据三类LSA中的Adv rtr字段，判断出ABR。
  - 根据Ls id、Net mask、Metric字段获得ABR到达目的网络号/掩码、开销。
  - 如果多个ABR产生了指向相同目的网段的三类LSA，则根节点将根据本路由器到达目的网段的累计开销进行比较，最终生成最小开销路由。如果根节点到达目的网段的累计开销值相同，则产生等价负载的路由。
- 如图所示，Area 0中RTA计算区域间路由过程中：
  - 192.168.1.0/24和192.168.2.0/24的三类LSA中，Adv rtr分别是RTB（2.2.2.2）和RTC（3.3.3.3）。
  - RTB产生的三类LSA中，网络号/掩码是192.168.1.0/24，开销为1，RTC产生的三类LSA中，网络号/掩码是192.168.2.0/24，开销为1。
  - RTA到达192.168.1.0/24下一跳是RTB，开销是2；RTA到达192.168.2.0/24下一跳是RTC，开销是2。

目录

1. 区域间路由计算过程
2. **区域间路由防环机制**
3. 虚连接的作用及配置

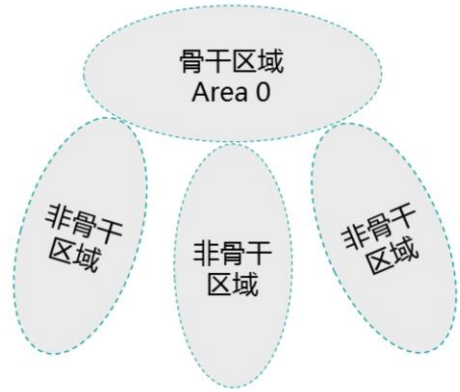
域间路由环路的产生



- RTB将AREA1中1的一类、二类LSA转换成三类LSA，发布到区域0中。
- RTC重新生成有关192.168.1.0/24网络的三类LSA并发布到Area 2中。
- 同理，RTE也将有关192.168.1.0/24网络的三类LSA发布到Area 3中。
- RTD又将192.168.1.0/24网络的三类LSA发布到Area 1中，从而形成了路由环路。

避免域间路由环路

- 骨干区域与非骨干区域
- 三类LSA的传递规则



- 思考：只有一个区域时，区域号配置为非0会有什么问题？

HUAWEI

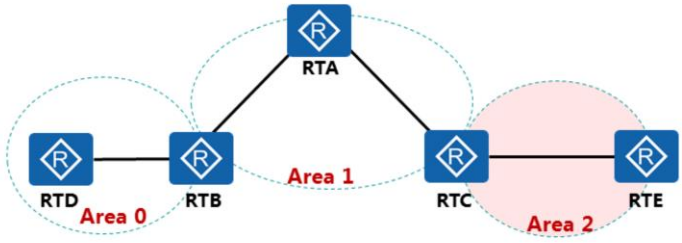
- 为防止区域间的环路OSPF定义了骨干区域和非骨干区域和三类LSA的传递规则。
  - OSPF划分了骨干区域和非骨干区域，所有非骨干区域均直接和骨干区域相连且骨干区域只有一个，非骨干区域之间的通信都要通过骨干区域中转，骨干区域ID固定为0。
  - OSPF规定从骨干区域传来的三类LSA不再传回骨干区域。
- 对于前文提到的ABR，OSPF要求ABR设备至少有一个接口属于骨干区域。
- 新建网络按照区域间的防环规则进行部署，可以避免区域间环路问题。但是部分网络可能因早期规划问题，区域间的连接关系违背了骨干区域和非骨干区域的规则。

目录

1. 区域间路由计算过程
2. 区域间路由防环机制
3. **虚连接的作用及配置**

HUAWEI

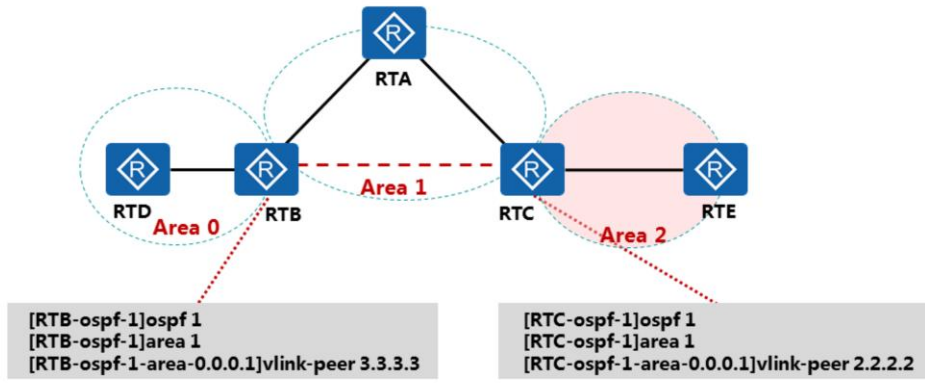
不规范的OSPF区域设计



- 违背了OSPF区域的连接规则，如何解决？

HUAWEI

虚连接vlink



  HUAWEI

- 骨干区域必须是连续的，但是并不要求物理上连续，可以使用虚连接使骨干区域逻辑上连续。
- 虚连接可以在任意两个区域边界路由器上建立，但是要求这两个区域边界路由器都有端口连接到一个相同的非骨干区域。
- 如图所示，在RTB和RTC之间建立了一条虚连接，以使Area 2穿越Area1连接到骨干区域。

思考题

1. 一条Network-Summary-LSA可以描述多条路由信息吗？
2. OSPF如何避免区域间的路由环路？

- 答案： 一条Network Summary LSA只能描述一条路由信息。
- 答案：OSPF划分了骨干区域和非骨干区域，所有非骨干区域均直接和骨干区域相连，且骨干区域只有一个；非骨干区域之间的通信都要通过骨干区域中转；并规定从骨干区域传来的三类LSA不再传回骨干区域。


谢 谢
www.huawei.com

OSPF外部路由

HUAWEI

前言

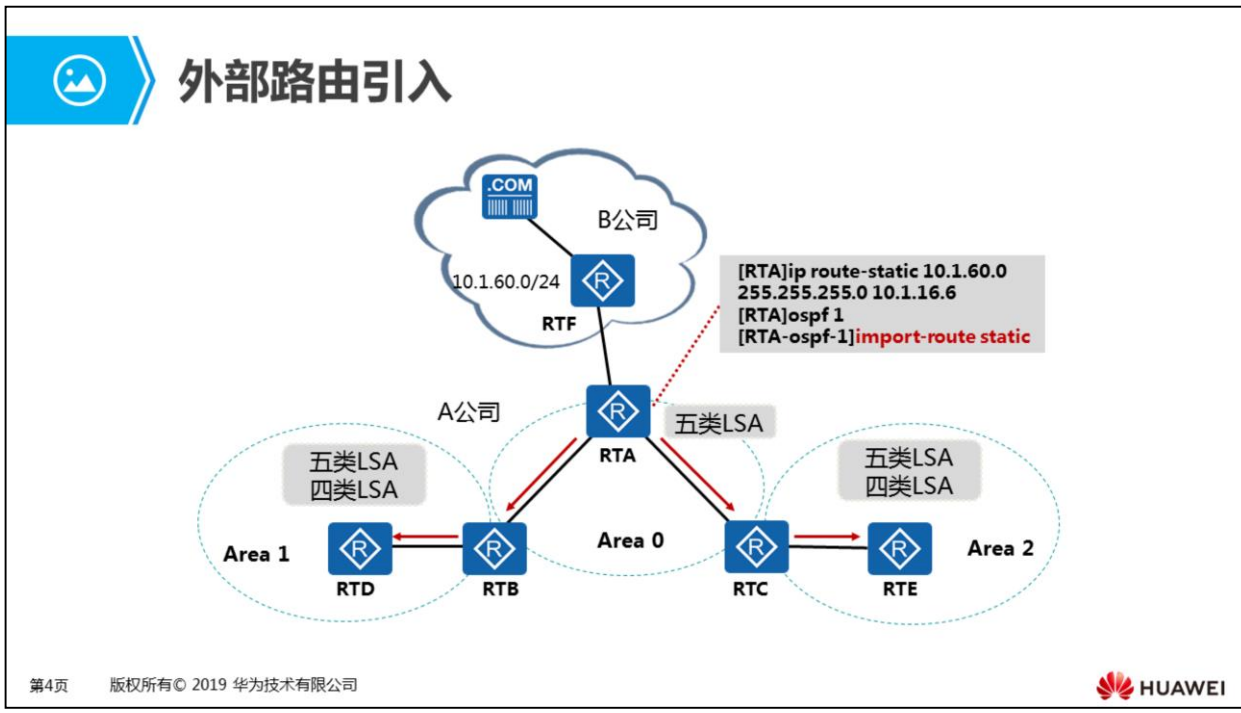
- 除了内部通信外，企业还需要与外部网络进行通信，不同企业网络之间存在互访需求。
- 假设A公司网络部署OSPF协议实现内部通信，因业务发展，需要访问B公司的一台WEB服务器。那么作为A公司的网络工程师，如何操作才能使本公司获取B公司的路由信息呢？

目标

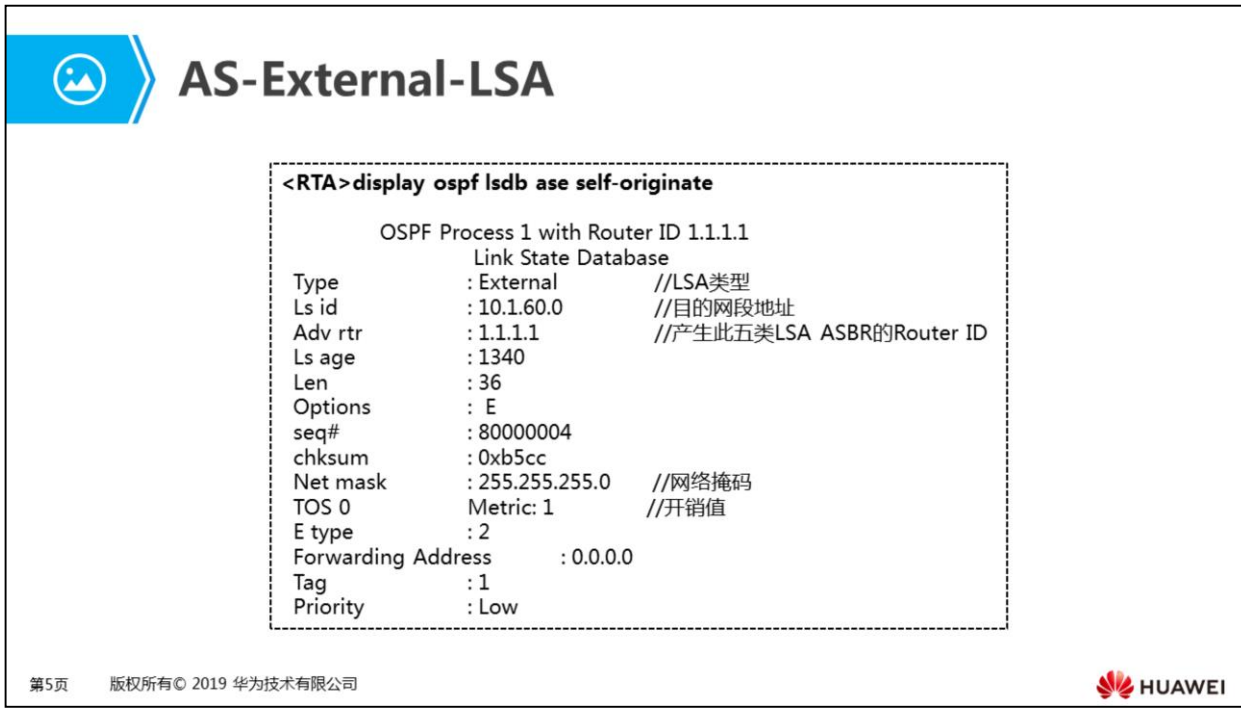
- 学完本课程后，您将能够：
  - 理解AS-External-LSA及ASBR-Summary-LSA的作用
  - 熟悉OSPF外部路由计算原理
  - 理解次优外部路由的产生原因

目录

1. **外部路由计算过程**
2. 外部路由的类型
3. 次优外部路由的产生



- 本例中，RTA上配置了一条静态路由，目的网络是10.1.60.0/24，下一跳是RTF。
- 在RTA的OSPF进程下，将配置的静态路由重发布到A公司的OSPF网络中，其中引入外部路由的OSPF路由器叫做ASBR（设备间互访需要路由双向可达，这里仅介绍OSPF网络内获取外部路由的过程）。
- RTA会生成一条AS-External-LSA（五类LSA），用于描述如何从ASBR到达外部目的地；RTB和RTC会生成一条ASBR-Summary-LSA（四类LSA），用于描述如何从ABR到达ASBR。
- 四类LSA和五类LSA，将被OSPF路由器用来计算外部路由。



- 这是由RTA生成的五类LSA，将被泛洪到所有OSPF区域。
- 五类LSA中包含的主要信息如下：
  - Ls id：目的网段地址。
  - Adv rtr：ASBR的Router ID。
  - Net mask：目的网段的网络掩码。
  - Metric：ASBR到达目的网络的开销值，默认值为1。
  - Tag：外部路由信息可以携带一个Tag标签，用于传递该路由的附加信息，通常用于路由策略，默认值为1。

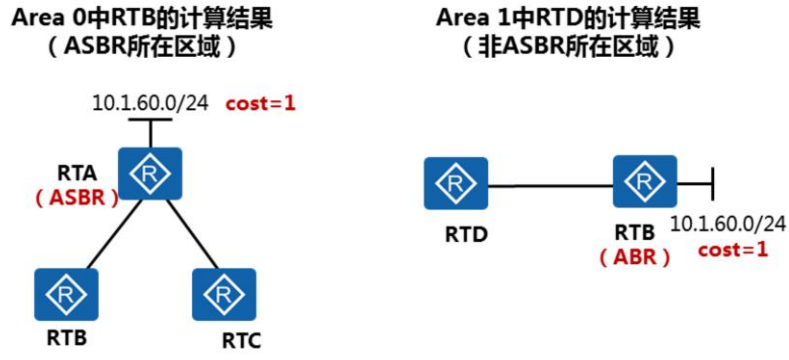
ASBR-Summary-LSA

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

                Area: 0.0.0.1
          Link State Database


Type       : Sum-Asbr        //LSA类型
Ls id      : 1.1.1.1         //ASBR的Router ID
Adv rtr    : 2.2.2.2         //产生此四类LSA ABR的Router ID
Ls age     : 15
Len        : 28
Options    :  E
seq#       : 80000005
chksum     : 0xf456
Tos 0      metric: 1         //从RTB到达此ASBR的开销
```


HUAWEI

- 这是由RTB在Area 1内生成的ASBR-Summary-LSA（四类LSA）。
- RTB向Area 1泛洪一条五类LSA时，同时生成一条四类LSA向Area 1泛洪。
- 该四类LSA主要包含下列信息：
  - Ls id：该ASBR的Router ID。
  - Adv rtr：该产生此四类LSA的ABR的Router ID。
  - Metric：从该ABR到达此ASBR的OSPF开销值。
- 四类LSA只能在一个区域内泛洪，五类LSA每泛洪到一个区域，相应区域的ABR都会生成一条新的四类LSA来描述如何到达ASBR。
- 因此描述到达同一个ASBR的四类LSA可以有多条，其Adv rtr是不同的，表示是由不同的ABR生成的。

外部路由计算



HUAWEI

- 以Area 0中RTB的外部路由计算为例：RTB收到五类LSA后，根据Adv rtr字段1.1.1.1发现，ASBR与自己同属于一个区域（Area 0），再根据Ls id、Net mask、Metric字段最终生成目的网络10.1.60.0/24 cost=1，下一跳为RTA的路由。
- 以Area 1中RTD的外部路由计算为例：RTD收到五类LSA后，根据Adv rtr字段1.1.1.1发现，ASBR与自己不同属于一个区域，再查找Ls id为1.1.1.1的四类LSA，发现此四类LSA的Adv rtr为2.2.2.2。再根据五类LSA中的LS id、Net mask、Metric字段最终生成目的网络10.1.60.0/24 cost=1，下一跳为RTB的路由。
- RTB、RTD最终计算出的路由条目cost都为1，根据物理拓扑可知，RTD开销值明显大于RTB，那么问题出在哪里呢？

目录

1. 外部路由计算过程
2. **外部路由的类型**
3. 次优外部路由的产生

外部路由类型

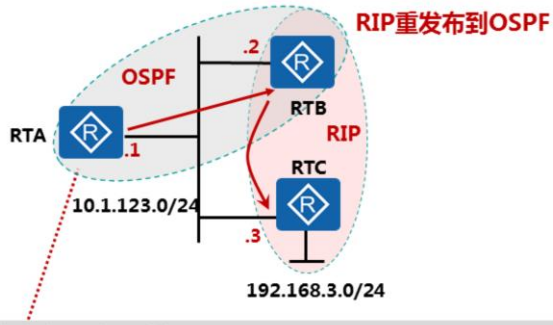
Type	Cost
第一类外部路由 （External Type-1）	AS内部开销值+AS外部开销值
第二类外部路由 （External Type-2）	AS外部开销值

- OSPF引入外部路由，共有两种类型可选：
  - 第一类外部路由的AS外部开销值被认为和AS内部开销值是同一数量级的，因此第一类外部路由的开销值为AS内部开销值（路由器到ASBR的开销）与AS外部开销值之和；这类路由的可信程度高一些，所以计算出的外部路由的开销与自治系统内部的路由开销是相当的，并且和OSPF自身路由的开销具有可比性。
  - 第二类外部路由的AS外部开销值被认为远大于AS内部开销值，因此第二类外部路由的开销值只包含AS外部开销，忽略AS内部开销（默认为第二类），这类路由的可信度比较低。
- 默认情况下，OSPF外部路由采用的是第二类外部路由。

目录

1. 外部路由计算过程
2. 外部路由的类型
3. **次优外部路由的产生**

次优外部路由的产生原因



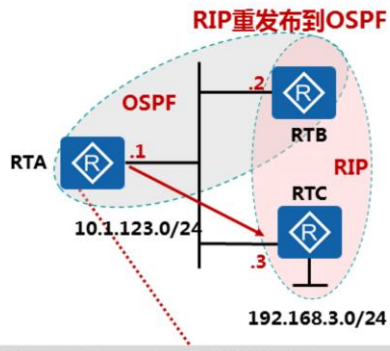
<RTA>display ip routing-table 192.168.3.0

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

HUAWEI

- 如图所示，RTA、RTB、RTC同处于一个MA网络，RTA和RTB之间运行OSPF，RTB和RTC之间运行RIP。
- RTB将通过RIP学来的路由重发布到OSPF，RTA通过OSPF学习到RIP中192.168.3.0/24的外部路由，但是下一跳是RTB。所以RTA访问192.168.3.0/24的流量先发送给RTB，RTB收到后又转发给RTC。在RTA上这条路由是次优的，最优的下一跳应当为RTC。
- OSPF通过设置Forwarding Address来解决这个问题。

Forwarding Address



RTB产生的五类LSA：

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

- 通常情况下，ASBR引入外部路由产生的五类LSA中Forwarding Address字段设置为0.0.0.0。
- 对于图中的场景，RTB路由表中到达192.168.3.0/24的下一跳地址为10.1.123.3。10.1.123.3所属网段10.1.123.0/24运行OSPF，所以RTB生成的五类LSA中，Forwarding Address被设置为10.1.123.3。
- 当RTA收到五类LSA时，发现Forwarding Address字段非0，其值为10.1.123.3，所以RTA按照Forwarding Address计算下一跳。

思考题

1. AS External LSA是在什么角色的路由器上产生的？它的基本作用是什么？
2. ASBR Summary LSA是在什么角色的路由器上产生的？它的基本作用是什么？
3. OSPF外部路由类型有哪两种？哪一种的优先级更高？

- 答案：AS External LSA是在ASBR路由器上产生的。AS External LSA的基本作用是用来向OSPF网络宣告外部路由。注意，一条AS External LSA只能宣告一条外部路由。
- 答案：ASBR Summary LSA是在ABR路由器上产生的。ASBR Summary LSA的基本作用是告诉其他路由器应该如何去往ASBR路由器。
- 答案：OSPF外部路由类型有External Type-1和External Type-2。External Type-1的优先级高于External Type-2。


谢 谢
www.huawei.com

OSPF特殊区域及其他特性

HUAWEI

前言

- OSPF路由器需要同时维护域内路由、域间路由、外部路由信息数据库。当网络规模不断扩大时，LSDB规模也不断增长。如果某区域不需要为其他区域提供流量中转服务，那么该区域内的路由器就没有必要维护本区域外的链路状态数据库。
- OSPF通过划分区域可以减少网络中LSA的数量，而可能对于那些位于自治系统边界的非骨干区域的低端路由器来说仍然无法承受，所以可以通过OSPF的特殊区域特性进一步减少LSA数量和路由表规模。

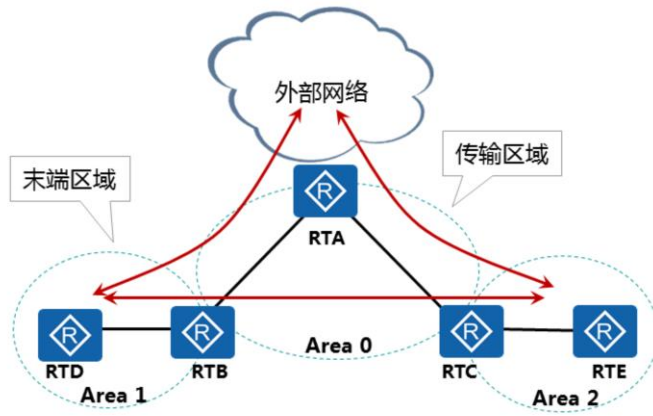
目标

- 学完本课程后，您将能够：
  - 理解OSPF特殊区域特性
  - 理解OSPF虚连接的应用场景
  - 熟悉OSPF路由汇总原理
  - 熟悉OSPF的更新机制
  - 了解OSPF的认证机制

目录

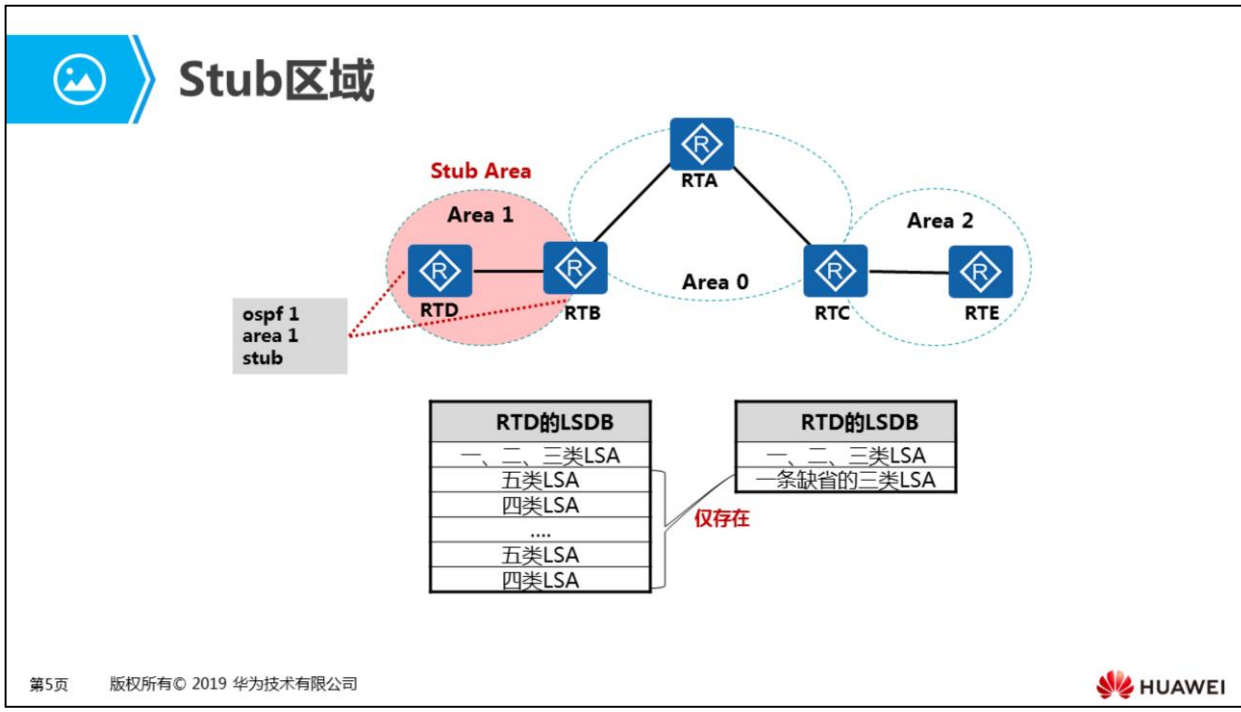
1. **Stub区域和Totally Stub区域**
2. NSSA区域和Totally NSSA区域
3. 区域间路由汇总和外部路由汇总
4. OSPF更新机制
5. OSPF认证机制

传输区域和末端区域



HUAWEI

- 如图所示，全网可分为四部分Area 0、Area 1、Area 2、外部网络。
- 四部分之间相互访问的主要流量如图中红线所示。
- 对于OSPF各区域，可分为两种类型：
  - 传输区域：除了承载本区域发起的流量和访问本区域的流量外，还承载了源IP和目的IP都不属于本区域的流量，即“穿越型流量”，如Area 0。
  - 末端区域：只承载本区域发起的流量和访问本区域的流量，如Area 1。
- 对于末端区域，需要考虑下几个问题：
  - 保存到达其他区域明细路由的必要性：访问其他区域通过单一出口，“汇总”路由相对明细路由更为简洁。
  - 设备性能：网络建设与维护必须要考虑成本因素。末端区域中可选择部署性能相对较低的路由器。
- OSPF路由器计算区域内、区域间、外部路由都需要依靠收集网络中的大量LSA，大量LSA会占用LSDB存储空间，所以解决问题的关键是在不影响正常路由的情况下，减少LSA的数量。

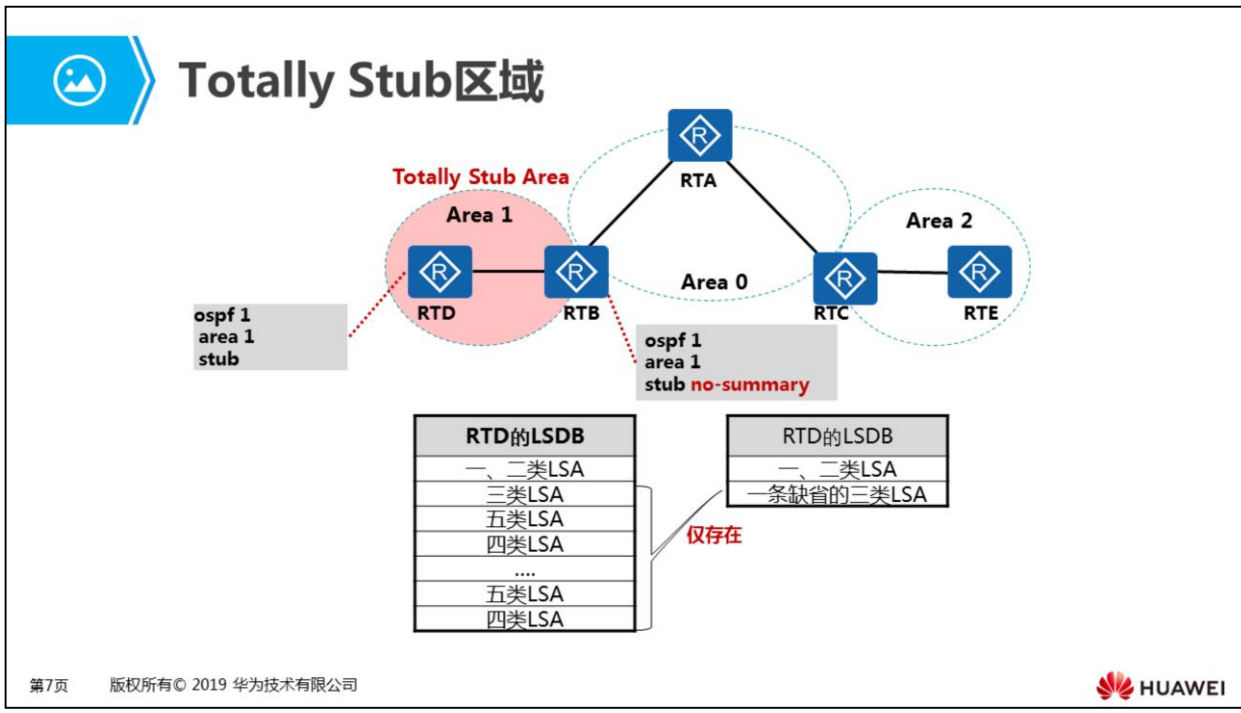


- Stub区域的ABR不向Stub区域内传播它接收到的自治系统外部路由（对应四类、五类LSA），Stub区域中路由器的LSDB、路由表规模都会大大减小。
- 为保证Stub区域能够到达自治系统外部，Stub区域的ABR将生成一条缺省路由（对应三类LSA），并发布给Stub区域中的其他路由器。
- Stub区域是一种可选的配置属性，但并不建议将每个区域都配置为Stub区域。通常来说，Stub区域位于自治系统的末梢，是那些只有一个ABR的非骨干区域。
- 配置Stub区域时需要注意下列几点：
  - 骨干区域不能被配置为Stub区域。
  - 如果要将一个区域配置成Stub区域，则该区域中的所有路由器必须都要配置成Stub路由器。
  - Stub区域内不能存在ASBR，自治系统外部路由不能在本区域内传播。
  - 虚连接不能穿越Stub区域建立。

Stub区域的OSPF路由表

```
<RTD>display ospf routing

        OSPF Process 1 with Router ID 4.4.4.4
                Routing Tables

Routing for Network
Destination        Cost    Type       NextHop         AdvRouterArea
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
```


- 配置Stub区域后，所有自治系统外部路由均由一条三类的默认路由代替。
- 除路由条目的减少外，当外部网络发生变化后，Stub区域内的路由器是不会直接受到影响的。



- Totally Stub区域既不允许自治系统外部路由（四类、五类LSA）在本区域内传播，也不允许区域间路由（三类LSA）在本区域内传播。
- Totally Stub区域内的路由器对其他区域及自制系统外部的访问需求是通过本区域ABR所产生的三类LSA缺省路由实现的。
- 与Stub区域配置的区别在于，在ABR上需要追加no-summary参数。

Totally Stub区域的OSPF路由表

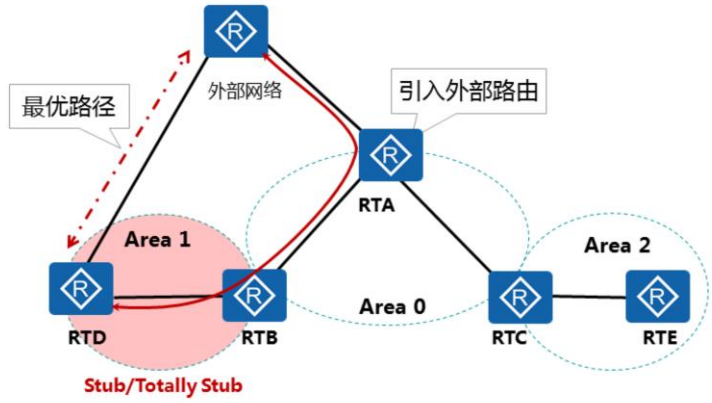
```
<RTD>display ospf routing

        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
```


- Totally Stub区域访问其他区域及自制系统外部是通过默认路由实现的。
- 自制系统外部、其他OSPF区域的网络发生变化，Totally Stub区域内的路由器是不直接受影响的。
- Stub、Totally Stub解决了末端区域维护过大LSDB带来的问题，但对于某些特定场景，Stub、Totally Stub并不是最佳解决方案。

目录

1. Stub区域和Totally Stub区域
2. **NSSA区域和Totally NSSA区域**
3. 区域间路由汇总和外部路由汇总
4. OSPF更新机制
5. OSPF认证机制

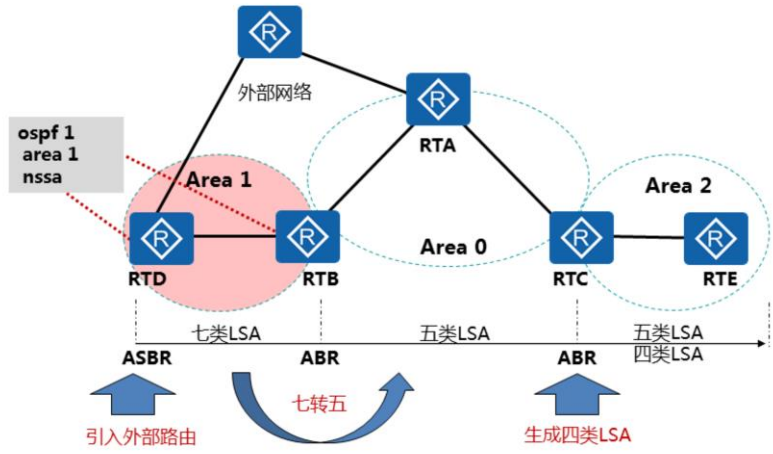
Stub区域、Totally Stub区域存在的问题



HUAWEI

- RTD和RTA同时连接到某一外部网络，RTA引入外部路由到OSPF域，RTD所在的Area 1为减小LSDB规模被设置为Stub或Totally Stub区域。RTD访问外部网络的路径是“RTD->RTB->RTA->外部网络”，显然相对于RTD直接访问外部网络而言，这是一条次优路径。
- OSPF规定Stub区域是不能引入外部路由的，这样可以避免大量外部路由对Stub区域设备资源的消耗。
- 对于既需要引入外部路由又要避免外部路由带来的资源消耗的场景，Stub和Totally Stub区域就不能满足需求了。

NSSA区域与Totally NSSA区域



  HUAWEI

- OSPF NSSA区域（Not-So-Stubby Area）是在原始OSPF协议标准中新增的一类特殊区域类型。
- NSSA区域和Stub区域有许多相似的地方。两者的差别在于，NSSA区域能够将自治域外部路由引入并传播到整个OSPF自治域中，同时又不会学习来自OSPF网络其它区域的外部路由。
- NSSA LSA（七类LSA）：
  - 七类LSA是为了支持NSSA区域而新增的一种LSA类型，用于描述NSSA区域引入的外部路由信息。
  - 七类LSA由NSSA区域的ASBR产生，其扩散范围仅限于ASBR所在的NSSA区域。
  - 缺省路由也可以通过七类LSA来产生，用于指导流量流向其它自治域。
- 七类LSA转换为五类LSA：
  - NSSA区域的ABR收到七类LSA时，会有选择地将其转换为五类LSA，以便将外部路由信息通告到OSPF网络的其它区域。
  - NSSA区域有多个ABR时，进行7类LSA与5类LSA转换的是Router ID最大的ABR。
- Totally NSSA和NSSA区别：
  - Totally NSSA不允许三类LSA在本区域内泛洪。
  - Totally NSSA与NSSA区域的配置区别在于ABR上需要追加no-summary参数。

NSSA区域与Totally NSSA区域的LSDB

NSSA区域

<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区域

<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**

HUAWEI

- 配置了NSSA区域的ABR产生一条七类LSA缺省路由。
- 配置了Totally NSSA区域的ABR会自动产生一条三类LSA缺省路由。

LSA总结

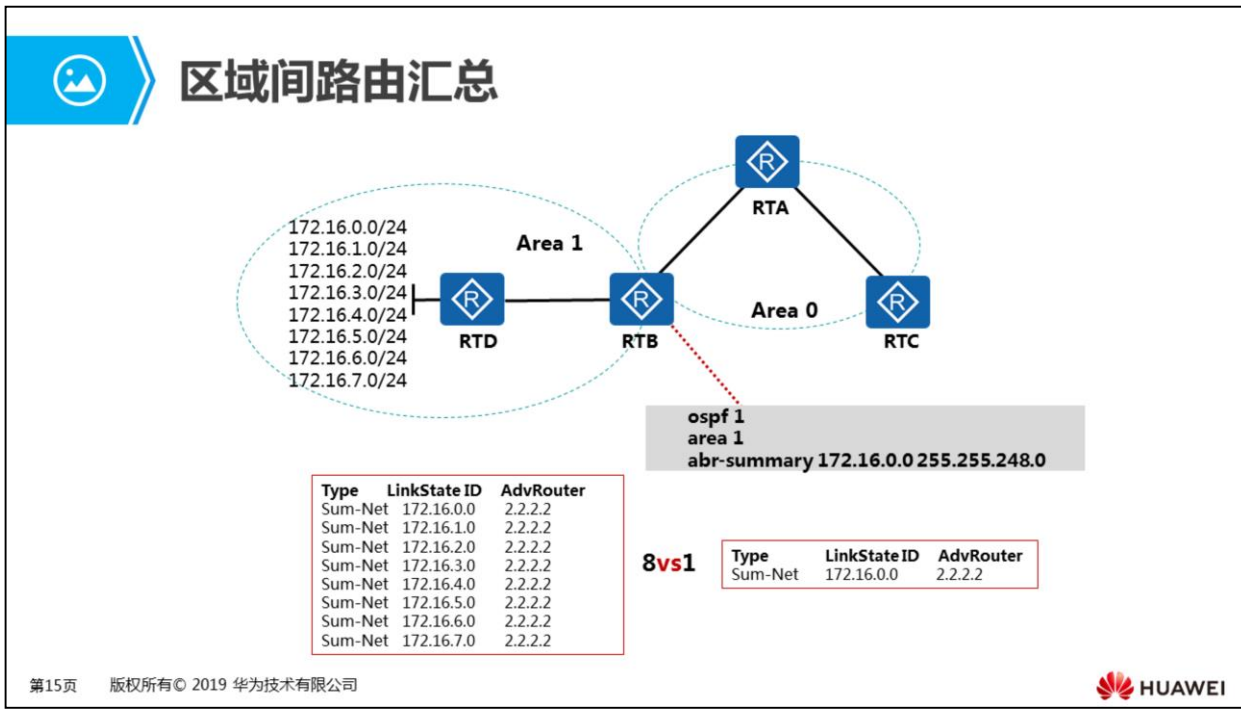
LSA类型	通告路由器	LSA内容	传播范围
Router LSA（Type-1）	OSPF Router	拓扑信息+路由信息	本区域内
Network LSA（Type-2）	DR	拓扑信息+路由信息	本区域内
Network-summary-LSA（Type-3）	ABR	域间路由信息	非（Totally）STUB区域
ASBR-summary-LSA（Type-4）	ABR	ASBR's Router ID	非（Totally）STUB区域
AS-external-LSA（Type-5）	ASBR	路由进程域外部路由	（非STUB区域）OSPF进程域
NSSA LSA（Type-7）	ASBR	NSSA域外部路由信息	（Totally）NSSA区域

- 思考：特殊区域的局限性在哪里，减少LSA还有没有其他方法？

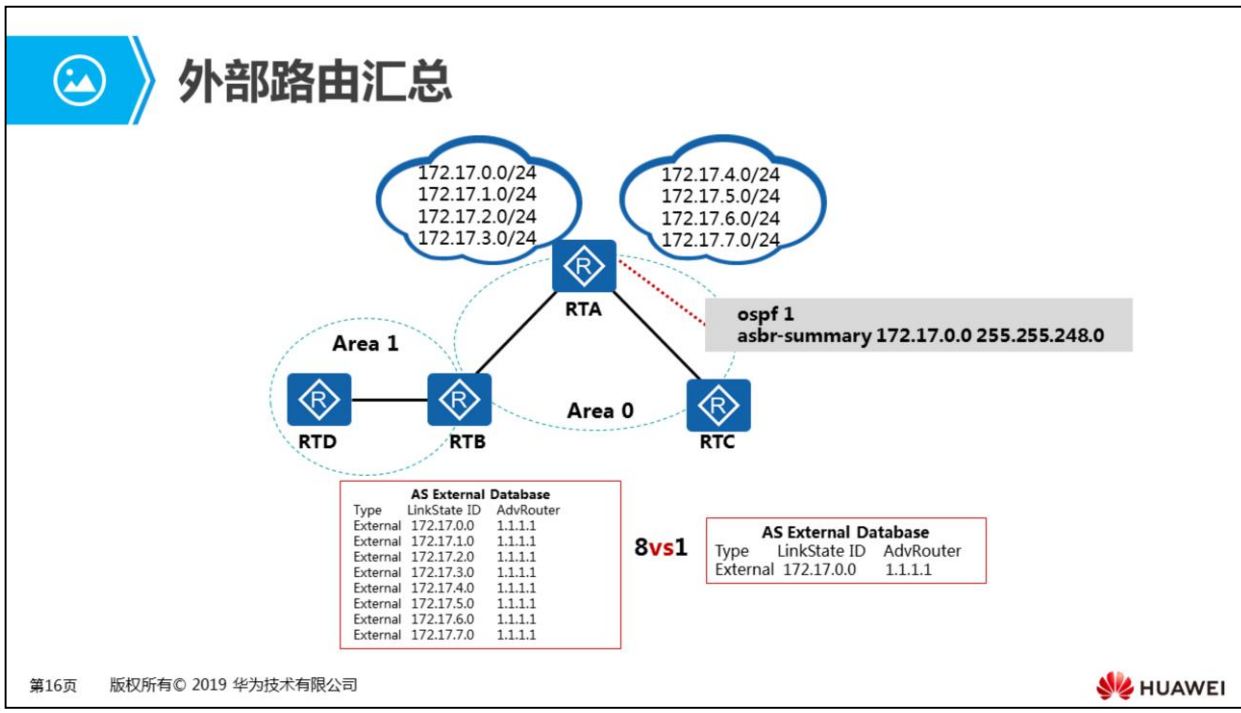
- LSA作用：
  - Router LSA（一类）：每个路由器都会产生，描述了路由器的链路状态和开销，在所属的区域内传播。
  - Network LSA（二类）：由DR产生，描述本网段的链路状态，在所属的区域内传播。
  - Network-summary-LSA（三类）：由ABR产生，描述区域内某个网段的路由，并通告给其他相关区域。
  - ASBR-summary-LSA（四类）：由ABR产生，描述到ASBR的路由，通告给除ASBR所在区域的其他相关区域。
  - AS-external-LSA（五类）：由ASBR产生，描述到AS外部的路由，通告到所有的区域（除了Stub区域和NSSA区域）。
  - NSSA LSA（七类）：由ASBR产生，描述到AS外部的路由，仅在NSSA区域内传播。
- 特殊区域不仅有效减少了区域内LSA的数量以及路由计算的压力，而且一定程度上也缩小了网络故障的影响范围。但特殊区域的局限性在于其作用范围只在本区域内，对于其他区域，如何才能减少LSA、降低路由计算的压力呢？

目录

1. Stub区域和Totally Stub区域
2. NSSA区域和Totally NSSA区域
3. **区域间路由汇总和外部路由汇总**
4. OSPF更新机制
5. OSPF认证机制



- 在大规模部署OSPF网络时，可能会出现由于OSPF路由表规模过大而降低路由查找速度的现象，为了解决这个问题，可以配置路由汇总，减小路由表的规模。
- 路由汇总是指将多条连续的IP前缀汇总成一条路由前缀。如果被汇总的IP地址范围内的某条链路频繁Up和Down，该变化并不会通告给被汇总的IP地址范围外的设备。因此，可以避免网络中的路由振荡，在一定程度上提高了网络的稳定性。
- 路由汇总只能汇总路由信息，所以ABR是可以执行路由汇总的位置之一：
  - ABR向其它区域发送路由信息时，以网段为单位生成三类LSA。如果该区域中存在一些连续的网段，则可以通过命令将这些连续的网段汇总成一个网段。这样ABR只发送一条汇总后的三类LSA，所有属于命令指定的汇总网段范围的LSA将不会再被单独发送出去。
- 如图所示，Area 1中存在8个连续网段，汇总前RTB将产生8条三类LSA。在RTB上配置汇总后，RTB仅产生1条三类LSA并泛洪到Area 0。
- 引入外部路由的ASBR也是执行路由汇总的位置之一。



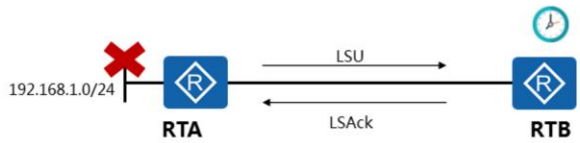
- ASBR汇总：
  - 配置ASBR汇总后，ASBR将对引入的外部路由进行汇总。NSSA区域的ASBR也可以对引入NSSA区域的外部路由进行汇总。
  - 如果设备既是NSAA区域的ASBR又是ABR，则可在将七类LSA转换成五类LSA时对相应前缀进行汇总。
- 如图所示，Area 0中RTA将8个连续的外部路由引入到OSPF域内，产生8条五类LSA并在OSPF进程域内泛洪。
- 在ASBR（RTA）配置外部路由汇总后，RTA将仅产生1条五类LSA并泛洪至OSPF路由进程域内。
- 路由汇总降低了网络故障的影响范围。
- 网络发生故障后，路由协议的收敛速度也是衡量路由协议的重要参考依据之一。

目录

1. Stub区域和Totally Stub区域
2. NSSA区域和Totally NSSA区域
3. 区域间路由汇总和外部路由汇总
4. **OSPF更新机制**
5. OSPF认证机制

定时更新与触发更新

- 定时更新：
  - LSA每1800s更新一次，3600s失效。
- 触发更新：
  - 当链路状态发生变化之后，立即发送链路状态更新。

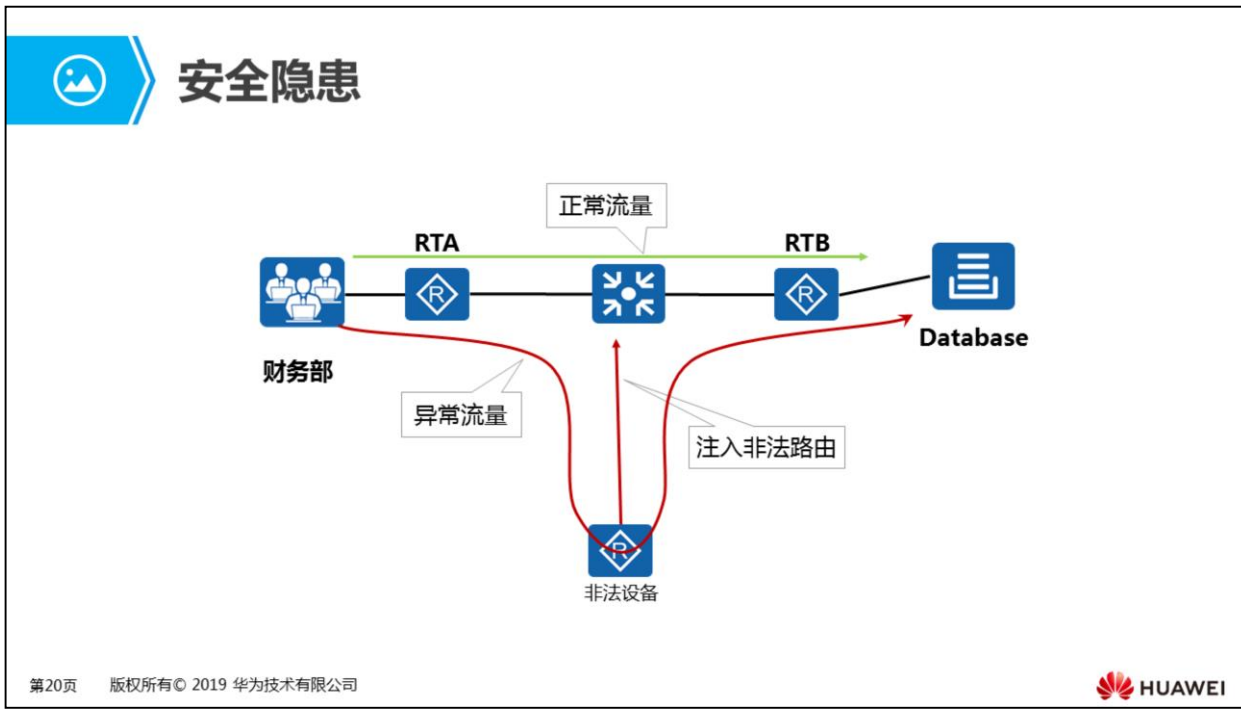


HUAWEI

- 为了保证路由计算的准确性，需要保证LSA的可靠性。
- OSPF为每个LSA条目维持一个老化计时器（3600s），当计时器超时，此LSA将从LSDB中删除。
- 为了防止LSA条目达到最大生存时间而被删除，OSPF通过定期更新（每1800s刷新一次）机制来刷新LSA。
- OSPF路由器每1800s会重新生成LSA，并通告给其他路由器。
- 为了加快收敛速度，OSPF设置了触发更新机制。
- 当链路状态发生变化后，路由器立即发送更新消息，其他路由器收到更新消息后立即进行路由计算，快速完成收敛。

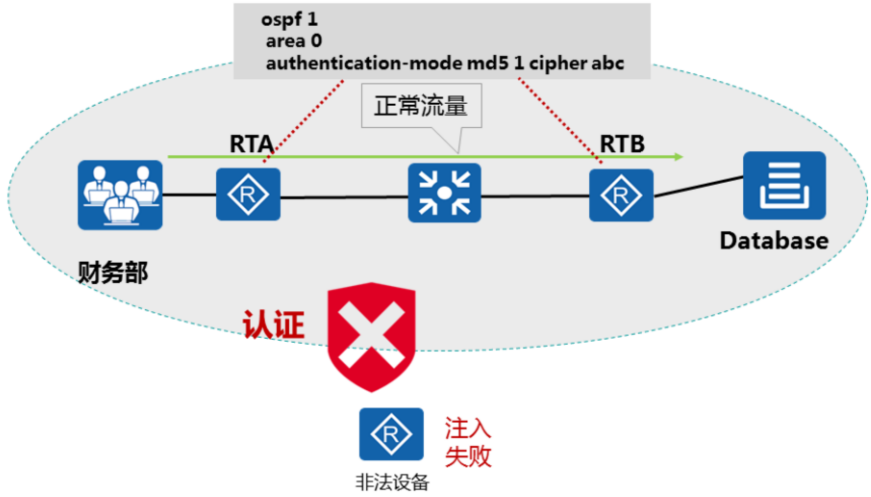
目录

1. Stub区域和Totally Stub区域
2. NSSA区域和Totally NSSA区域
3. 区域间路由汇总和外部路由汇总
4. OSPF更新机制
5. **OSPF认证机制**



- 如图所示，内部网络通过OSPF协议传递路由。正常情况下，财务部访问公司数据库的流量走向是“财务部->RTA->RTB->Database”。
- 非法设备接入公司内网，通过向网络中注入非法路由，引导流量进行非正常的转发。即“财务部->RTA->非法设备->RTB->Database”。非法设备收到财务部的流量之后，进行恶意分析，获取财务部关键信息，造成公司机密泄露。
- OSPF如何保证路由的安全性呢？

认证解决安全隐患

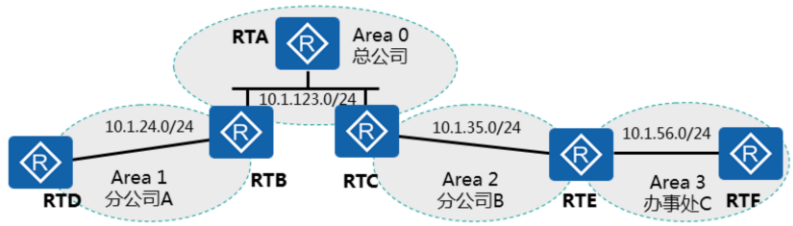


HUAWEI

- OSPF支持认证功能，只有通过认证的OSPF路由器才能正常建立邻居关系，交互信息。
- 两种认证方式：
  - 区域认证方式。
  - 接口认证方式。
- 支持的认证模式分为null（不认证）、simple（明文）、MD5以及HMAC-MD5。
- 当两种认证方式都存在时，优先使用接口认证方式。

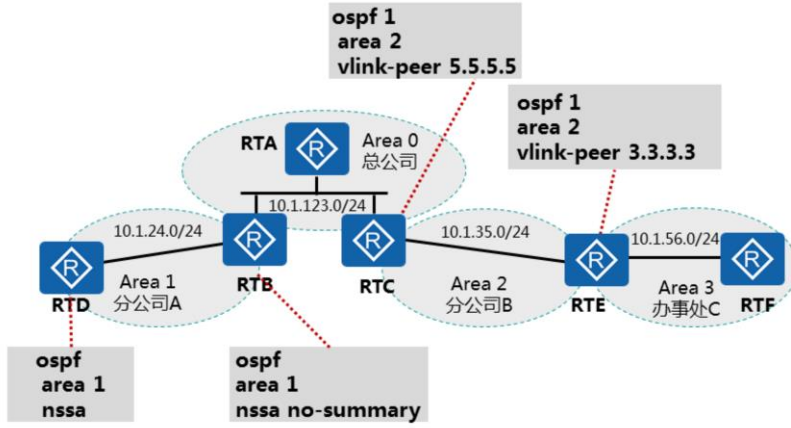
OSPF综合应用场景

- 公司网络拓扑如下图所示：OSPF基础配置已经完成，作为网络管理员，有如下几个问题需要解决：
  - 总公司与分公司A、B之间通信正常，但无法与办事处C通信。
  - 分公司A的设备性能较低，希望降低路由计算、存储压力，同时考虑网络扩展，需要保留引入外部路由的功能。
  - 办事处C外来人员较多，采用较安全的方式保证路由交互的安全性。
  - RTA引入外部路由时除了考虑外部开销之外，还需要考虑OSPF域内的开销。



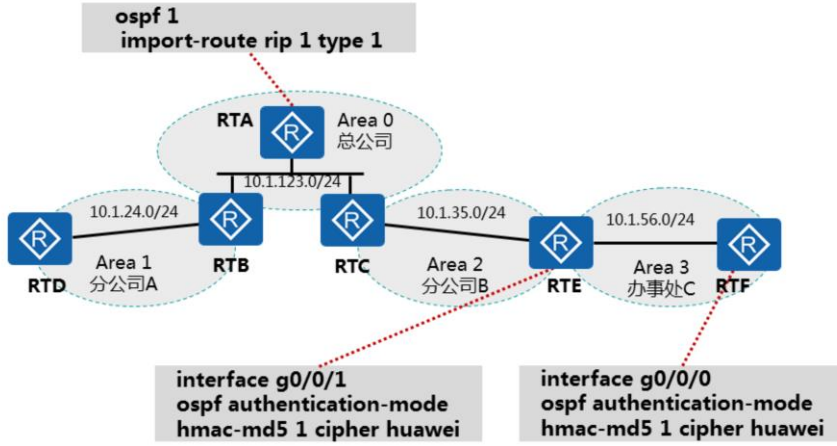
- 需求1分析：办事处C处于Area 3，RTE左侧与Area 2相连。根据OSPF骨干区域与非骨干区域的连接规则，不能正常通行的原因在于Area 3没有与Area 0直接相连。解决的方式是在RTE和RTC之间建立虚连接。
- 需求2分析：区域内部设备性能低，降低路由计算压力可以通过Stub、Totally Stub、NSSA、Totally NSSA，最大程度减少需要选择Totally Stub或Totally NSSA，同时为了保留外部路由引入的功能，只能选择Totally NSSA。
- 需求3分析：保证路由安全性需要通过认证的方式，最安全的认证模式是采用HMAC-MD5。认证形式采取接口认证。
- 需求4分析：在计算外部路由时如要考虑OSPF域内开销，可通过引入类型为1类的外部路由实现。

OSPF配置实现 (1)



HUAWEI

OSPF配置实现 (2)



HUAWEI

思考题

1. OSPF定义了哪几种特殊区域？
2. Stub区域与Totally Stub区域的主要差别是什么？
3. 区域间路由汇总功能在什么路由器上配置？

- 答案：OSPF定义了四种特殊区域，分别是Stub Area，Totally Stub Area，Not-So-Stubby Area（NSSA），Totally NSSA。
- 答案：Stub区域不允许Type-4和Type-5 LSA进入，但允许Type-3 LSA进入。Totally Stub区域不仅不允许Type-4和Type-5 LSA进入，同时也不允许Type-3 LSA进入，只允许表示缺省路由的Type-3 LSA进入。
- 答案：在区域边界路由器（ABR）上配置。


谢 谢
www.huawei.com

IS-IS协议原理与配置

HUAWEI

前言

- 和OSPF一样，IS-IS也是一种基于链路状态并使用最短路径优先算法进行路由计算的一种IGP协议。IS-IS最初是国际化标准组织ISO为它的无连接网络协议CLNP设计的一种动态路由协议。
- 为了提供对IP的路由支持，IETF在RFC1195中对IS-IS进行了扩充和修改，使它能够同时应用在TCP/IP和OSI环境中，修订后的IS-IS协议被称为集成化的IS-IS。由于IS-IS的简便性及扩展性强的特点，目前在大型ISP的网络中被广泛地部署。

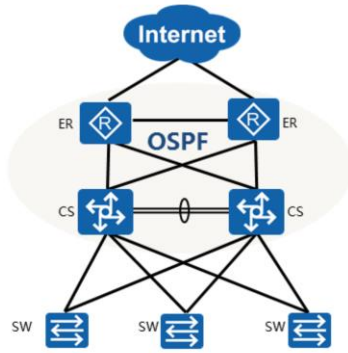
目标

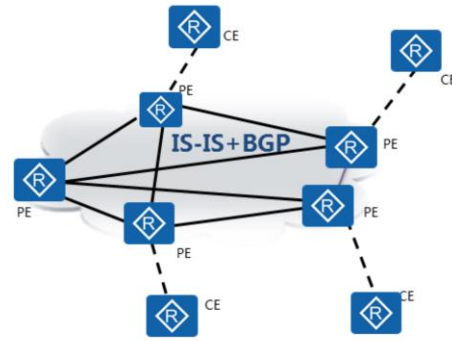
- 学完本课程后，您应该能：
- 理解IS-IS的基本原理
- 熟悉IS-IS与OSPF的区别
- 掌握IS-IS的常用配置

目录

1. **IS-IS协议基本原理**
2. IS-IS与OSPF的区别
3. IS-IS应用场景配置

场景应用





- 园区网：

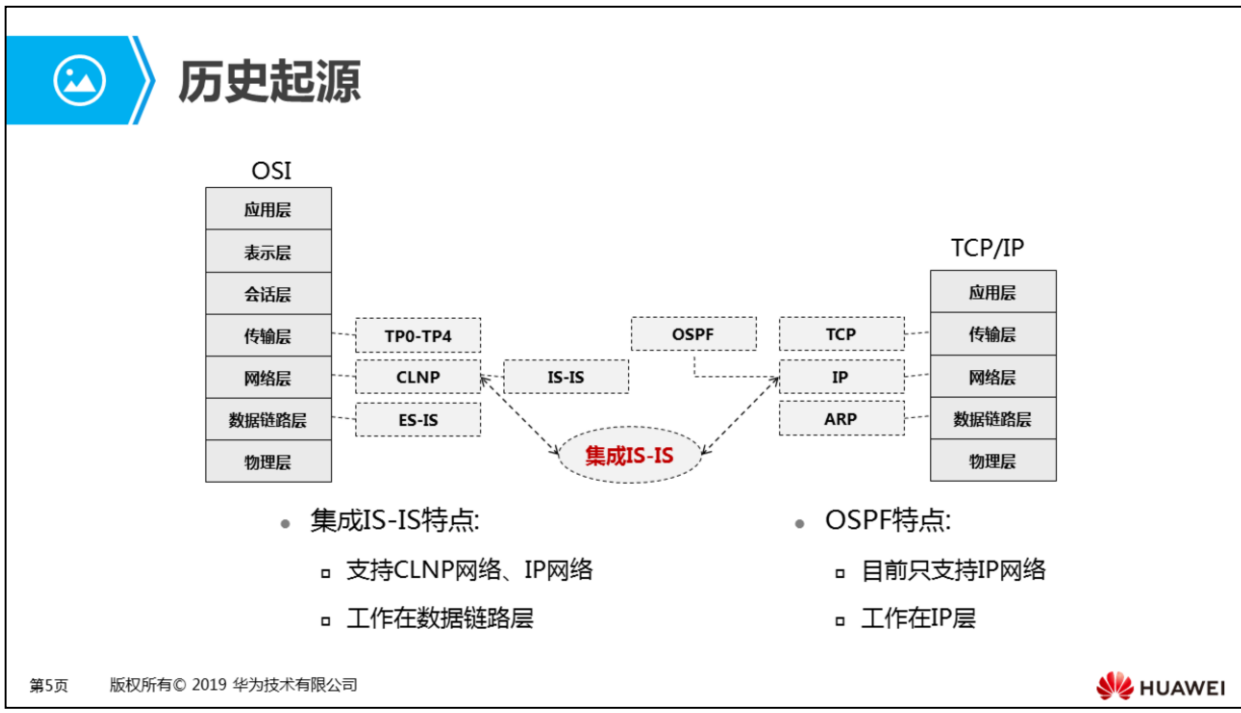
  区域多样、策略多变、调度精细

- 骨干网：

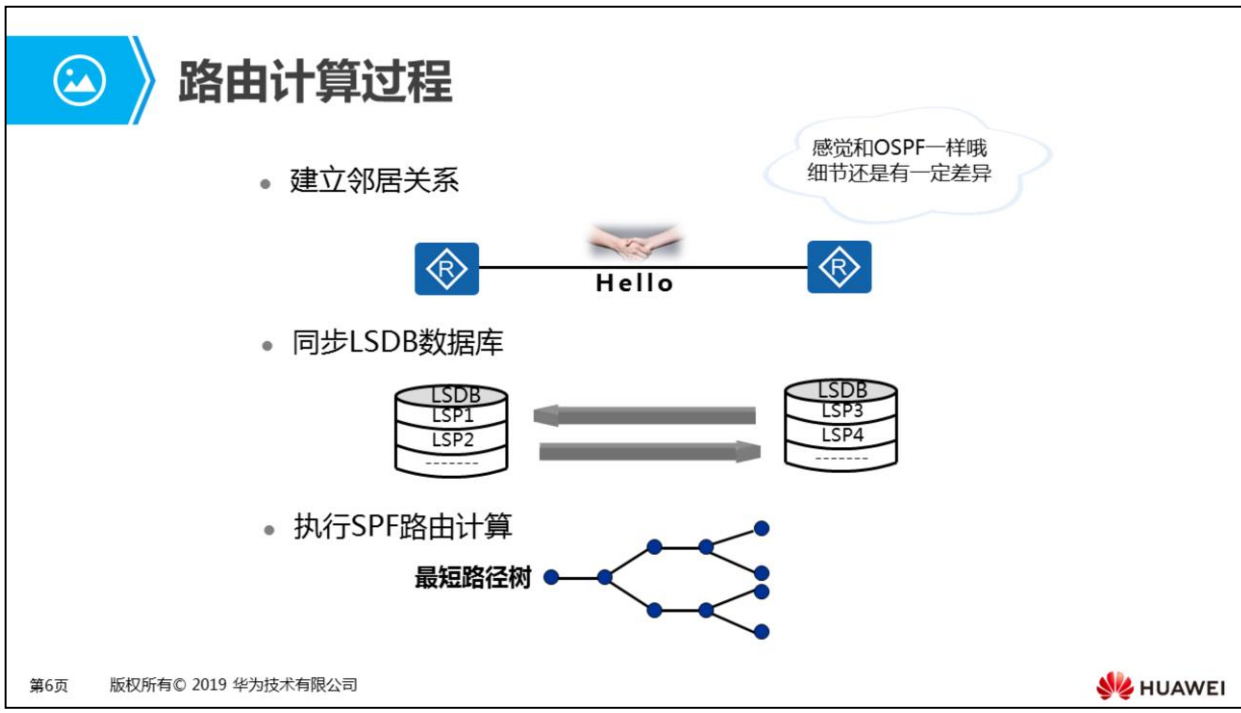
  区域扁平、收敛极快、承载庞大

HUAWEI

- 园区网特点：
  - 应用型网络，主要面向企业网用户。
  - 路由器数量偏少，动态路由的LSDB库容量相对偏少，三层路由域相对偏少。
  - 有出口路由的概念，对内部外部路由划分敏感。
  - 地域性跨度不大，带宽充足，链路状态协议开销对带宽占用比偏少。
  - 路由策略和策略路由应用频繁多变，需要精细化的路由操作。
  - OSPF的多路由类型（内部/外部），多区域类型（骨干/普通/特殊），开销规则优良（根据带宽设定），网络类型多样（最多五种类型）的特点在园区网得到了极大的发挥。
- 骨干网特点：
  - 服务型网络，由ISP（互联网服务提供商）组建，并为终端用户提供互联服务。
  - 路由调度占据绝对统治地位，路由器数量庞大。
  - 架构层面扁平化，要求IGP作为基础路由为上层BGP协议服务。
  - LSDB规模宏大，对链路收敛极度敏感，线路费用高昂。
  - 追求简单高效，扩展性高，满足各种客户业务需求（IPV6/IPX）。
  - IS-IS的快速算法（PRC得到加强），简便报文结构（TLV），快速邻居关系建立，大容量路由传递（基于二层开销低）等一系列特点在骨干网有着天然的优势。



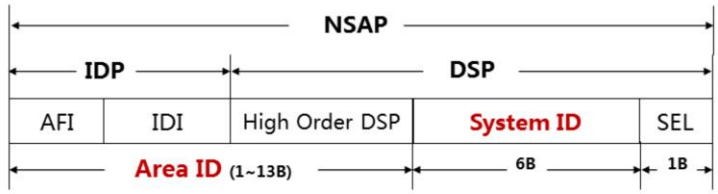
- IS-IS最初是国际标准化组织ISO（the International Organization for Standardization）为它的无连接网络协议CLNP（ConnectionLess Network Protocol）设计的一种动态路由协议。
- 为了提供对IP的路由支持，IETF在RFC1195中对IS-IS进行了扩充和修改，使它能够同时应用在TCP/IP和OSI环境中，称为集成化IS-IS（Integrated IS-IS），后面如果没有特别说明，提到的IS-IS都是指集成IS-IS。
- IS-IS属于内部网关协议，用于自治系统内部。IS-IS是一种链路状态协议，使用最短路径优先算法进行路由计算。



- 邻居关系建立：
  - 邻居关系建立主要是通过HELLO包交互并协商各种参数，包括电路类型（level-1/level-2），Hold time，网络类型，支持协议，区域号，系统ID，PDU长度，接口IP等。
- 链路信息交换：
  - 与OSPF不同，ISIS交互链路状态的基本载体不是LSA（link state advertisement），而是LSP（link state PDU）；交互的过程没有OSPF协议那样经历了多个阶段，主要是通过CSNP和PSNP两种协议报文来同步，请求以及确认链路状态信息（承载的是链路状态信息摘要），而链路状态信息的详细拓扑和路由信息是由LSP报文传递。
- 路由计算：
  - SPF计算和OSPF基本一样的，但ISIS算法分离了拓扑结构和IP网段，加快了网络收敛速度。

地址结构

TCP/IP协议栈	IP协议	IP地址	OSPF	Area ID+Router ID
OSI系统	CLNP协议	NSAP地址	IS-IS	NET标识符



NET是一类特殊的NSAP（SEL＝00），在路由器上配置IS-IS时，只需要考虑NET即可。如：

49.0001.0000.0000.0001.00
Area ID　System ID　N-SEL

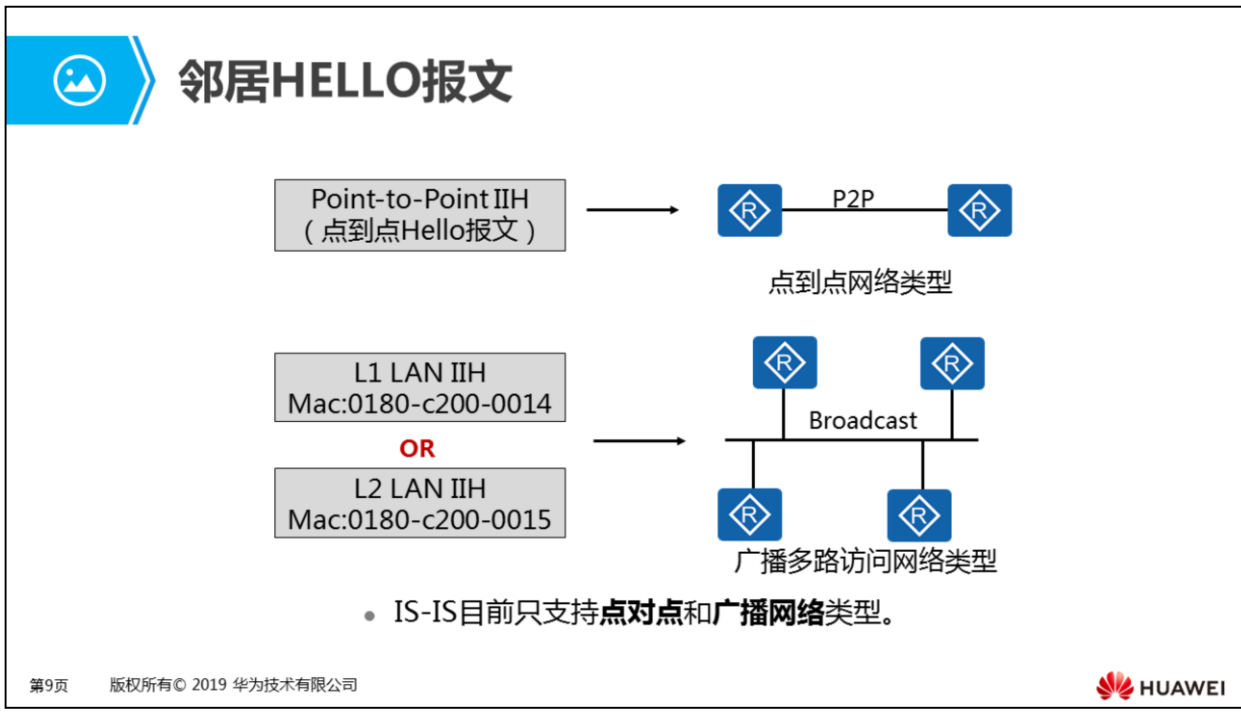
HUAWEI

- NSAP地址：
- IDP相当于IP地址中的主网络号。它是由ISO规定，并由AFI与IDI两部分组成。AFI表示地址分配机构和地址格式，IDI用来标识域。
- DSP相当于IP地址中的子网号和主机地址。它由High Order DSP、System ID和SEL三个部分组成。High Order DSP用来分割区域，System ID用来区分主机，SEL用来指示服务类型。
- Area Address（Area ID）由IDP和DSP中的High Order DSP组成，既能够标识路由域，也能够标识路由域中的区域。因此，它们一起被称为区域地址，相当于OSPF中的区域编号。
- System ID用来在区域内唯一标识主机或路由器。在设备的实现中，它的长度固定为48bit（6字节）。
- SEL的作用类似IP中的“协议标识符”，不同的传输协议对应不同的SEL。在IP上SEL均为00。
- NET：
- 网络实体名称NET指的是设备本身的网络层信息，可以看作是一类特殊的NSAP（SEL＝00），NET的长度与NSAP的相同，最多为20个字节，最少为8个字节。在路由器上配置IS-IS时，只需要考虑NET即可，NSAP可不必去关注。
- 在配置IS-IS过程中，NET最多也只能配3个。在配置多个NET时，必须保证它们的System ID都相同。

路由器分类

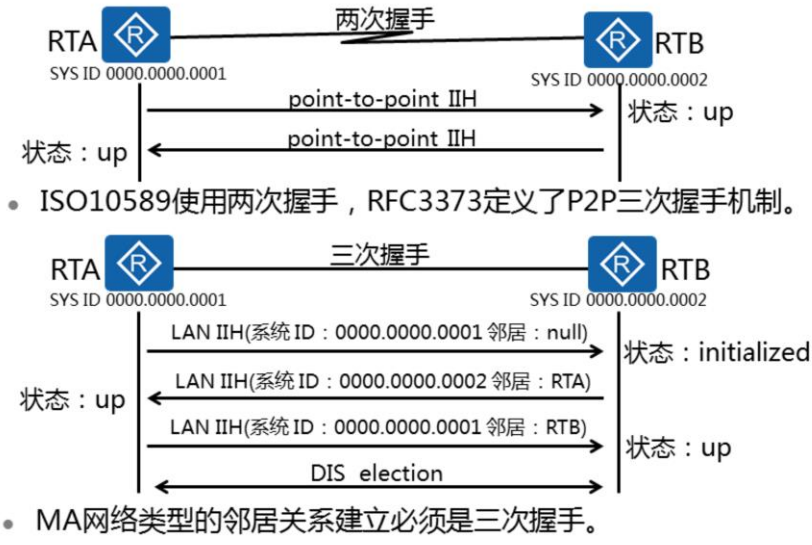
- IS-IS路由器的三种类型
  - Level-1路由器（只能创建level-1的LSDB）
  - Level-2路由器（只能创建level-2的LSDB）
  - Level-1-2路由器（路由器默认的类型，能同时创建level-1和level-2的LSDB）

- Level-1路由器：
  - Level-1只能与属于同一区域的Level-1和Level-1-2路由器形成邻居关系，只负责维护Level-1的链路状态数据库，该LSDB包含本区域内的路由信息，到本区域外的报文转发给最近的Level-1-2路由器。Level-1路由器只可能建立Level-1的邻接关系。
- Level-2路由器：
  - Level-2路由器负责区域间的路由，它可以与相同或者不同区域的Level-2路由器或者不同区域的Level-1-2路由器形成邻居关系。Level-2路由器维护一个Level-2的LSDB，该LSDB包含区域间的路由信息。Level-2路由器只可能建立Level-2的邻接关系。
- Level-1-2路由器：
  - 同时属于Level-1和Level-2的路由器称为Level-1-2路由器。Level-1-2路由器维护两个LSDB，Level-1的LSDB用于区域内路由，Level-2的LSDB用于区域间路由。
  - Level-1-2路由器可以与同一区域的Level-1形成Level-1邻居关系，也可以与其他区域的Level-2和Level-1-2路由器形成Level-2的邻居关系。
- 不同区域间，只能建立Level-2的邻接关系：
  - Level-2路由器可以与Level-2路由器建立邻接关系。
  - Level-1-2路由器可以与Level-2路由器建立邻接关系。
  - Level-1-2路由器可以与Level-1-2路由器建立邻接关系。



- HELLO PDU（Hello protocol data unit）：
  - HELLO报文的作用是邻居发现，协商参数并建立邻居关系，后期充当保活报文。
  - IS-IS建立邻居关系和OSPF一样，通过hello报文的交互来完成。但是会根据场景分为三种类型的hello报文。
  - 广播网中的Level-1 IS-IS使用Level-1 LAN IIH（Level-1 LAN IS-IS Hello），目的组播MAC为：0180-c200-0014。
  - 广播网中的Level-2 IS-IS使用Level-2 LAN IIH（Level-2 LAN IS-IS Hello），目的组播MAC为：0180-c200-0015。
  - 非广播网络中则使用P2P IIH（point to point IS-IS Hello）。但是其没有表示DIS（虚节点）的相关字段。
  - IIH报文需要通过填充字段用于邻居两端协商发送报文的大小。
- IS-IS支持的网络类型：
  - 点对点网络类型（P2P）。
  - 广播多路访问网络类型（Broadcast Multiple Access）。
  - 在帧中继等特殊环境下，可以通过创建子接口支持P2P的网络类型。

邻居关系建立



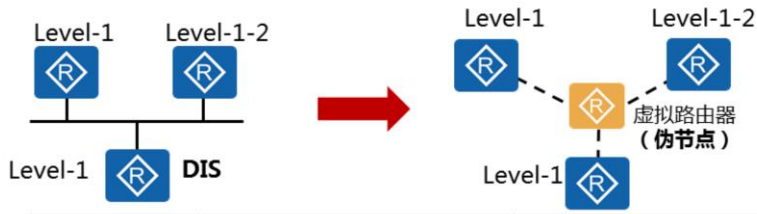
- ISO10589使用两次握手，RFC3373定义了P2P三次握手机制。

- MA网络类型的邻居关系建立必须是三次握手。

HUAWEI

- 在P2P链路上，分为两次握手机制和三次握手机制。
  - 两次握手只要路由器收到对端发来的Hello报文，就单方面宣布邻居为up状态，建立邻居关系，不过容易存在单通风险。
  - 通过三次发送P2P的IS-IS Hello PDU最终建立起邻居关系，与广播链路邻居关系的建立情况相同。
- 在广播链路上，使用LAN IIH报文执行三次握手建立邻居关系。
  - 当收到邻居发送的Hello PDU报文里面没有自己的system ID的时候，状态机进入initialized。
  - 只有收到邻居发过来的Hello PDU有自己的system ID才会up，排除了链路单通的风险。
  - 广播网络中邻居up后会选举DIS(虚节点)，DIS的功能类似OSPF的DR(指定路由器)。

DIS及DIS与DR的类比



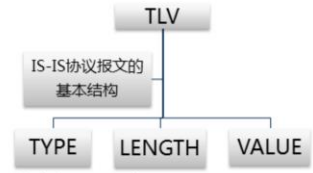
类比点	ISIS-DIS	OSPF-DR
选举优先级	所有优先级都参与选举	0优先级不参与选举
选举等待时间	2个Hello报文间隔	40s
备份	无	有（BDR）
邻接关系	所有路由器互相都是邻接关系	DRother之间是2-way关系
抢占性	会抢占	不会抢占
作用	周期发送CSNP，保障MA网络LSDB同步	主要为了减少LSA泛洪

- DIS与伪节点：
  - DIS是指指定中间系统（Designated IS）。
  - 伪节点是指在广播网络中由DIS创建的虚拟路由器。
- DIS的特点：
  - 在广播网络，需要选举DIS，所以在邻居关系建立后，路由器会等待两个Hello报文间隔再进行DIS的选举。Hello报文中包含Priority 字段，Priority值最大的将被选举为该广播网的DIS。若优先级相同，接口MAC地址较大的被选举为DIS。IS-IS中DIS发送Hello时间间隔默认为10/3秒，而其他非DIS路由器发送Hello间隔为10秒。
- DIS与DR的类比：
  - 选举时优选级的比较，DIS的优先级为0也可以参与选举。OSPF中优先级为0不参与选举DR。
  - 选举的过程需要一定的时间，OSPF选举DR/BDR需要waiting time达40秒，过程也较为复杂，而ISIS选举DIS等待两个Hello报文间隔就可以，简单快捷。

- 选举结果ISIS只有一个DIS，但是OSPF除了有DR，还有一个BDR用做备份。
- 选举结束后，后期有新的Router加入到链路进来，如果优先级比DIS高是可抢占的，但是DR是不可抢占的。
- 选举完成后，ISIS网络链路内所有的路由器之间都建立的是邻接关系。OSPF中DRothers只与DR/BDR形成full邻接关系， DRothers之间只有2-way的关系。
- 关于DIS和DR的作用：
    - 进行SPF计算时，都把它当成虚节点，简化MA网络的逻辑拓扑（相同点）。
    - 都是为了减少LSP/LSA的泛洪（相同点）。
    - 在ISIS中还可以由DIS发送CSNP来同步链路的LSDB（ISIS扩展作用）。

链路状态信息的载体

- LSP PDU——用于交换链路状态信息。
  - 实节点LSP
  - 伪节点LSP（只在广播链路存在）
- SNP PDU——用于维护LSDB 的完整与同步，且为摘要信息。
  - CSNP（用于同步LSP）
  - PSNP（用于请求和确认LSP）



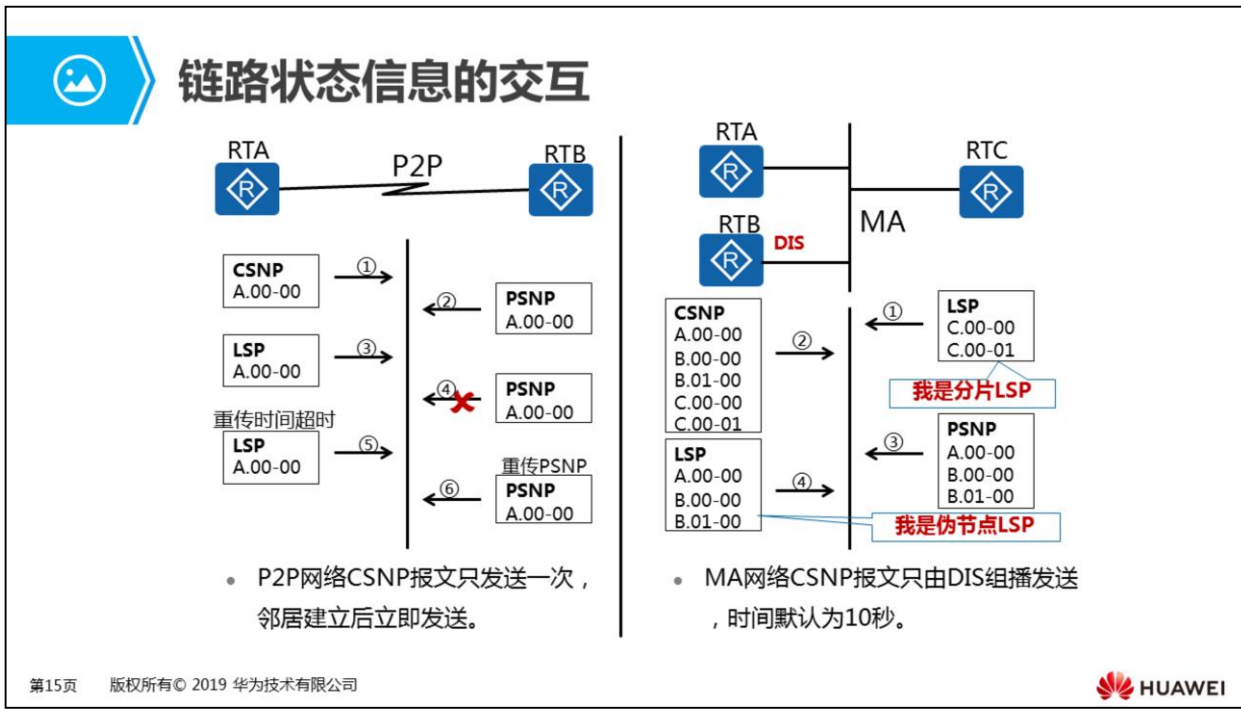
协议报文都分为Level-1和Level-2两种，在MA网络中所有协议报文的目的MAC都是组地址：
Level-1地址为：0180-C200-0014
Level-2地址为：0180-C200-0015

  HUAWEI

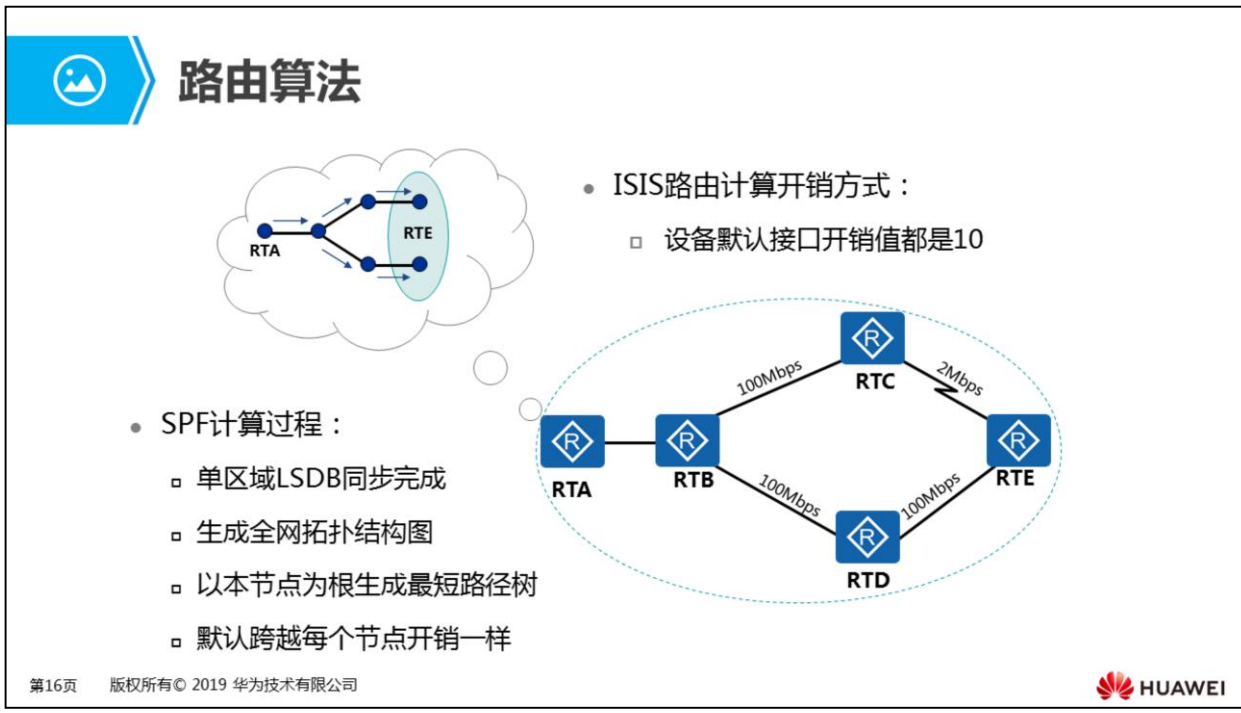
- ISIS TLV：
  - TLV的含义是：类型（TYPE），长度（LENGTH），值（VALUE）。实际上是一个数据结构，这个结构包含了这三个字段。
  - 使用TLV结构构建报文的好处是灵活性和扩展性好。采用TLV使得报文的整体结构固定，增加新特点只需要增加新TLV即可。不需要改变整个报文的整体结构。
  - 网络拓扑结构和路由信息用TLV结构表现使得报文的灵活性和扩展性得到了极大的发挥。
- LSP PDU（Link State Protocol PDU）：
  - LSP类似于OSPF的LSA，承载的是链路状态信息，包含了拓扑结构和网络号。
  - Level-1 LSP由Level-1 路由器传送。
  - Level-2 LSP由Level-2 路由器传送。
  - Level-1-2 路由器则可传送以上两种LSP。
  - LSP 报文中包含了两个重要字段是ATT字段、IS-Type字段。其中ATT字段用于标识该路由是L1/L2路由器发送的，IS-Type用来指明生成此LSP的IS-IS类型是Level-1还是Level-2 IS-IS。
  - LSP的刷新间隔为15分钟；老化时间为20分钟。但是一条LSP的老化除了要等待20分钟外，还要等待60秒的零老化时延；LSP重传时间为5秒。
- SNP PDU（Sequence Number PDU）：
  - CSNP（Complete Sequence Number PDU）包括LSDB中所有LSP的摘要信息，从

而可以在相邻路由器间保持LSDB的同步。

- PSNP（Partial Sequence Number PDU）包含部分LSDB中的LSP摘要信息，能够对LSP进行请求和确认。
- CSNP 类似于OSPF的DD报文传递的是LSDB里所有链路信息摘要。PSNP类似于OSPF的LSR或LSAck报文用于请求和确认部分链路信息。

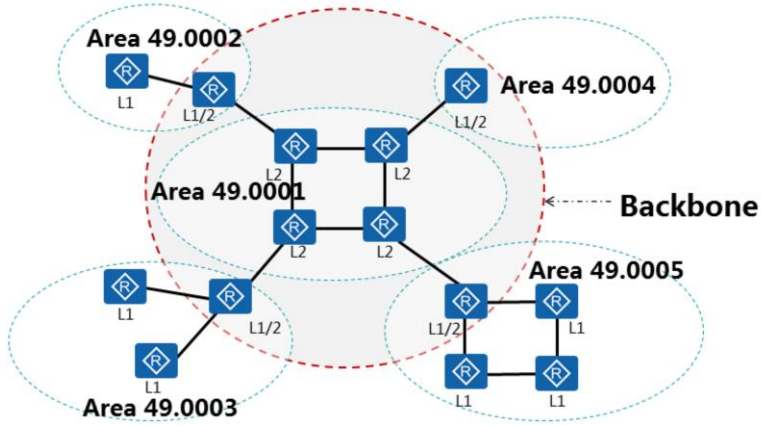


- P2P网络LSDB同步过程：
  - 建立邻居关系之后，RTA与RTB会先发送CSNP给对端设备。如果对端的LSDB与CSNP没有同步，则发送PSNP请求索取相应的LSP。
  - 假定RTB向RTA索取相应的LSP，此时向RTA发送PSNP。RTA发送RTB请求的LSP的同时启动LSP重传定时器，并等待RTB发送PSNP作为收到LSP的确认。
  - 如果在接口LSP重传定时器超时后，RTA还没有收到RTB发送的PSNP报文作为应答，则重新发送该LSP直至收到RTB的PSNP报文作为确认。
- MA网络中新加入的路由器与DIS 的LSDB同步交互过程：
  - 假设新加入的路由器RTC已经与RTB（DIS）和RTA建立了邻居关系。
  - 建立邻居关系之后，RTC将自己的LSP发往组播地址（Level-1：01-80-C2-00-00-14；Level-2：01-80-C2-00-00-15）。这样网络上所有的邻居都将收到该LSP。
  - 该网段中的DIS会把收到RTC的LSP加入到LSDB中，并等待CSNP报文定时器超时（DIS每隔10秒发送CSNP报文）并发送CSNP 报文，进行该网络内的LSDB同步。
  - RTC收到DIS发来的CSNP报文，对比自己的LSDB数据库，然后向DIS发送PSNP报文请求自己没有的LSP（如RTA和RTB的LSP就没有）。
  - RTB作为DIS收到该PSNP报文请求后向RTC发送对应的LSP进行LSDB 的同步。



- IS-IS的计算特点：
  - 在本区域内路由器第一次启动的时候执行的是Full-SPF算法。
  - 后续收到的LSP更新，如果是部分拓扑的变化执行的iSPF计算。
  - 如果只是路由信息的变化，执行的就是PRC计算。
  - 由于采用拓扑与网络分离的算法，路由收敛速度得到了加强。
- ISIS路由计算的开销方式：
  - Narrow模式（设备默认模式开销都是10，手工配置接口开销取值范围为1～63）。
  - Wide模式（设备默认模式开销都是10，手工配置接口开销取值范围是1～16777215）。
  - 进程下加入auto-cost enable命令，Narrow模式和Wide模式都会参考接口带宽大小计算开销值，只是参考准则有少许差异。

网络分层路由域



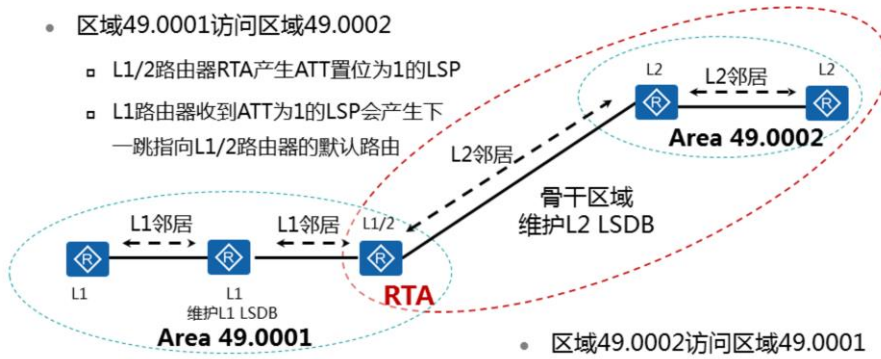
- ISIS协议的区域边界在整个Router，OSPF协议的区域边界在Router的接口。

  HUAWEI

- IS-IS整体拓扑：
  - 为了支持大规模的路由网络，IS-IS在自治系统内采用骨干区域与非骨干区域两级的分层结构。一般来说，将Level-1路由器部署在非骨干区域，Level-2路由器和Level-1-2路由器部署在骨干区域。每一个非骨干区域都通过Level-1-2路由器与骨干区域相连。
  - 拓扑中为一个运行IS-IS协议的网络，它与OSPF的多区域网络拓扑结构非常相似。整个骨干区域不仅包括Level-2的所有路由器，还包括Level-1-2路由器。
  - Level-1-2级别的路由器可以属于不同的区域，在Level-1区域，维护Level-1的LSDB，在Level-2区域，维护Level-2的LSDB。
- 拓扑所体现的IS-IS与OSPF不同点：
  - 在OSPF中，每个链路只属于一个区域；而在IS-IS中，每个链路可以属于不同的区域；
  - 在IS-IS中，单个区域没有物理的骨干与非骨干区域的概念；而在OSPF中，Area0被定义为骨干区域；
  - 在IS-IS中，Level-1和Level-2级别的路由器分别采用SPF算法，分别生成最短路径树SPT；在OSPF中，只有在同一个区域内才使用SPF算法，区域之间的路由需要通过骨干区域来转发。

区域间路由

- 区域49.0001访问区域49.0002
  - L1/2路由器RTA产生ATT置位为1的LSP
  - L1路由器收到ATT为1的LSP会产生下一跳指向L1/2路由器的默认路由



- 区域49.0002访问区域49.0001
  - L1/2路由器RTA会把区域49.0001的明细路由以叶子节点方式挂载在L2级别的LSP上面并处在Level-2的LSDB中
  - L2路由器通过自己SPF计算得出访问Area49.0001的明细路由

HUAWEI

- Level-1路由器的路由特点：
  - 只拥有Level-1的链路状态数据库。
  - 其链路状态数据库中只有本区域路由器LSP。
  - 其路由表里没有其他区域的路由信息。
  - 其路由表里都有一条默认路由，下一条是指向到Level-1-2路由器。
- Level-2路由器的路由特点：
  - Level-2路由器只有Level-2的链路状态数据库。
  - 其LSDB中有骨干区域路由器的LSP，但是没有Level-1路由器产生的LSP。
  - 路由表里面有整个网络的路由信息。
- Level-1-2路由器的路由特点：
  - Level-1-2路由器同时拥有Level-2和Level-1的链路状态数据库。
  - Level-1数据库包含本区域的LSP，Level-2数据库包含骨干区域LSP。
  - 在自己产生的Level-1的LSP中设置了ATT比特位为1。
  - 路由表里面有整个网络的路由信息。

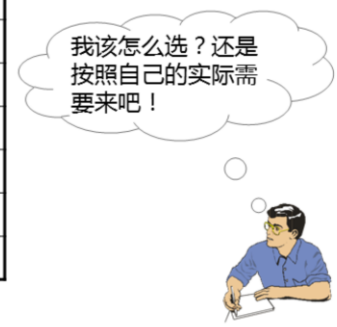
目录

1. IS-IS协议基本原理
2. **IS-IS与OSPF的区别**
3. IS-IS应用场景配置

HUAWEI

IS-IS与OSPF差异性

差异性	IS-IS	OSPF
网络类型	少	多
开销方式	复杂	简便
区域类型	少	多
路由报文类型	简单	多样
路由收敛速度	很快	快
扩展性	强	一般
路由负载能力	超强	强



  HUAWEI

- 网络类型和开销方式：
  - IS-IS协议只支持两种网络类型，且所有带宽默认开销值都是一样的，OSPF协议支持四种网络类型，且会根据不同的带宽设定相应的开销值，对帧中继，按需链路等网络类型有很好的支持。
- 区域类型：
  - IS-IS协议分L1/L2区域，L2区域是骨干区域有全部明细路由。L1去往L2只有默认路由。OSPF协议分骨干区域，普通区域，特殊区域。普通区域和特殊区域跨区域访问需要经过骨干区域。
- 报文类型：
  - IS-IS协议路由承载报文类型只有LSP报文且里面路由信息是不区分内部与外部的，简单高效，无需递归计算。OSPF协议路由承载报文LSA类型多样，有1/2/3/4/5/7类等。路由级别等级森严，且需要递归计算，适合精细化调度计算。
- 路由算法：
  - ISIS协议区域内某个节点上的网段发生变化时，触发的是PRC算法，收敛比较快，计算路由的报文开销也比较小。OSPF协议由于网络地址参与了拓扑的构建，在区域内当网段地址改变触发的是i-spf算法，相对来说过程繁琐复杂些。
- 扩展性：
  - ISIS协议任何路由信息都使用TLV传递，结构简单，易于扩展，如对IPv6的支持只增加2个TLV就解决了。且ISIS本身对IPX等协议是支持的。OSPF协议本身是为IP特定开发的，支持IPv4和IPv6的OSPF协议是两个独立的版本（OSPFv2和OSPFv3）。

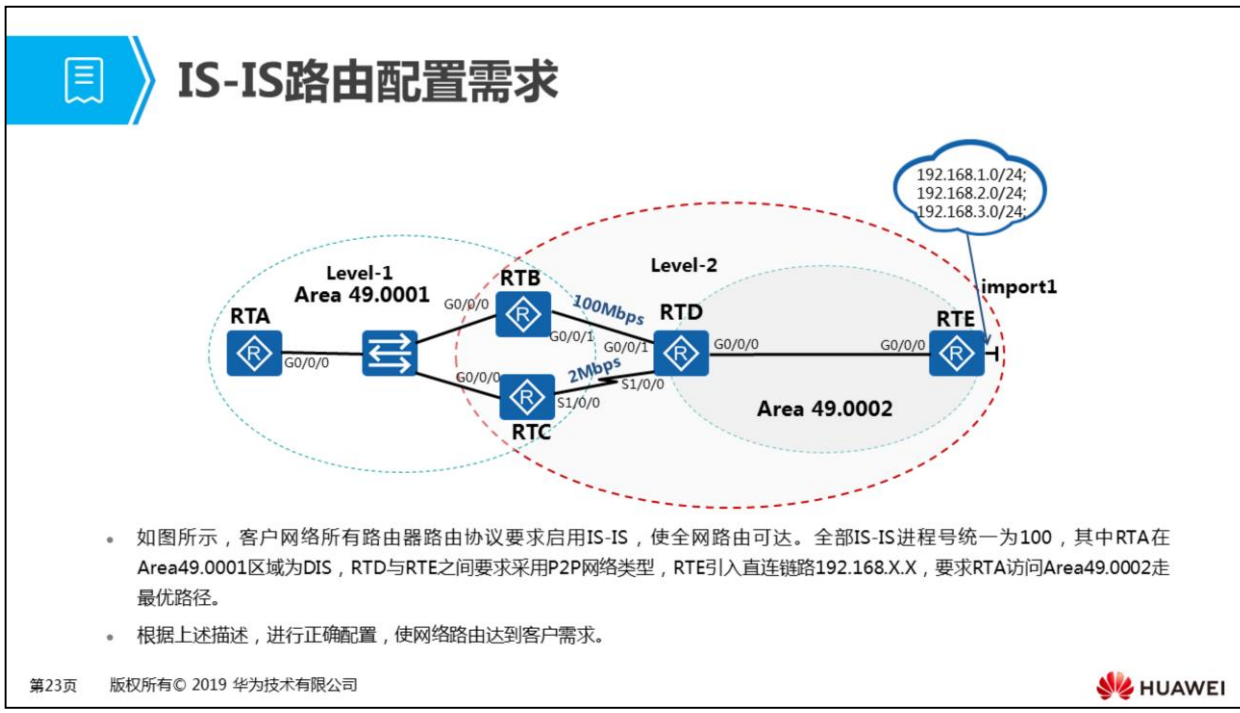
术语对照表

缩略语	OSI术语	IETF术语
IS	Intermediate System	Router
ES	End System	Host
DIS	Designated Intermediate System	OSPF中的DR
SysID	System ID	OSPF中的Router ID
LSP	Link State PDU	OSPF中的LSA
IIH	IS-IS Hello PDU	OSPF中的Hello报文
PSNP	Partial Sequence Number PDU	OSPF中的LSR或LSAck报文
CSNP	Complete Sequence Number PDU	OSPF中的DD报文

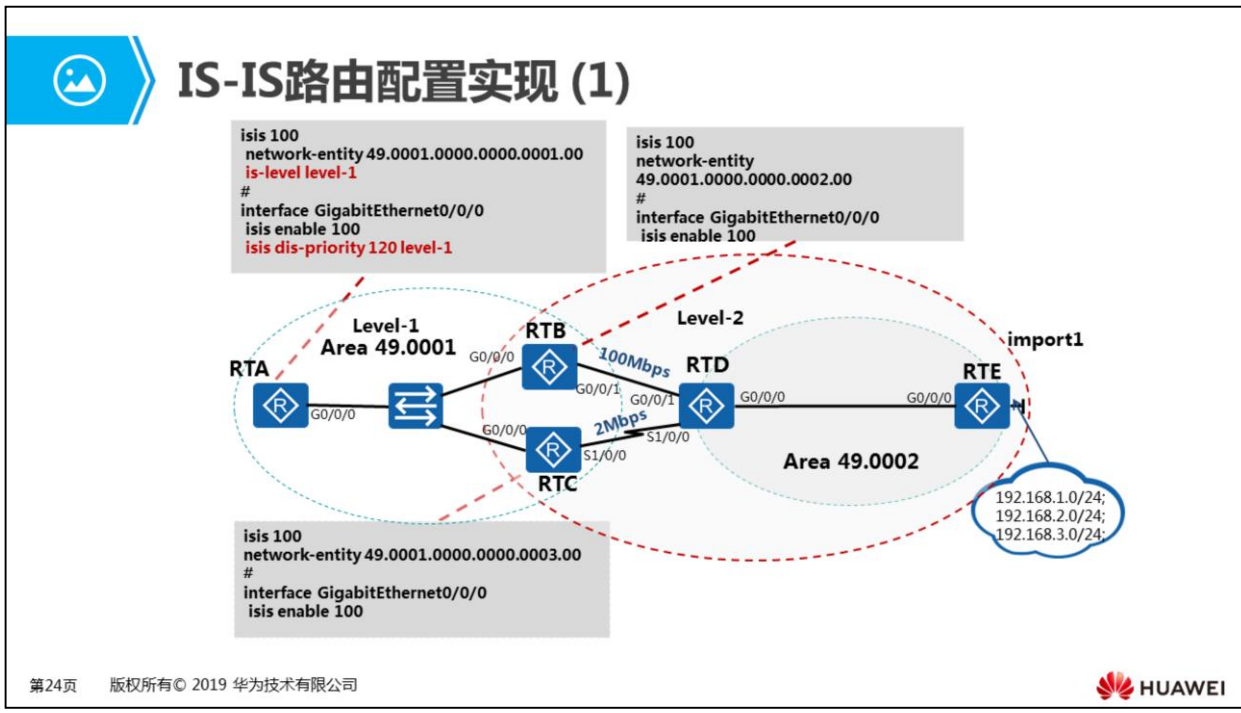
目录

1. IS-IS协议基本原理
2. IS-IS与OSPF的区别
3. **IS-IS应用场景配置**

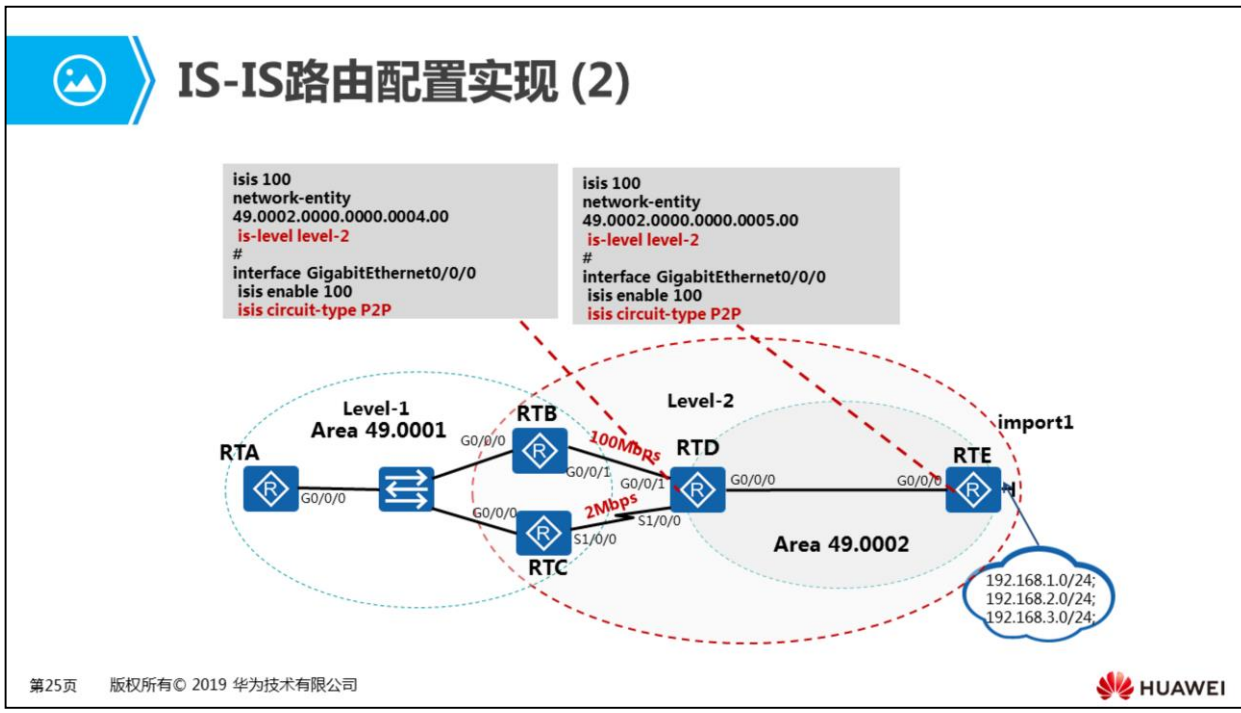
HUAWEI



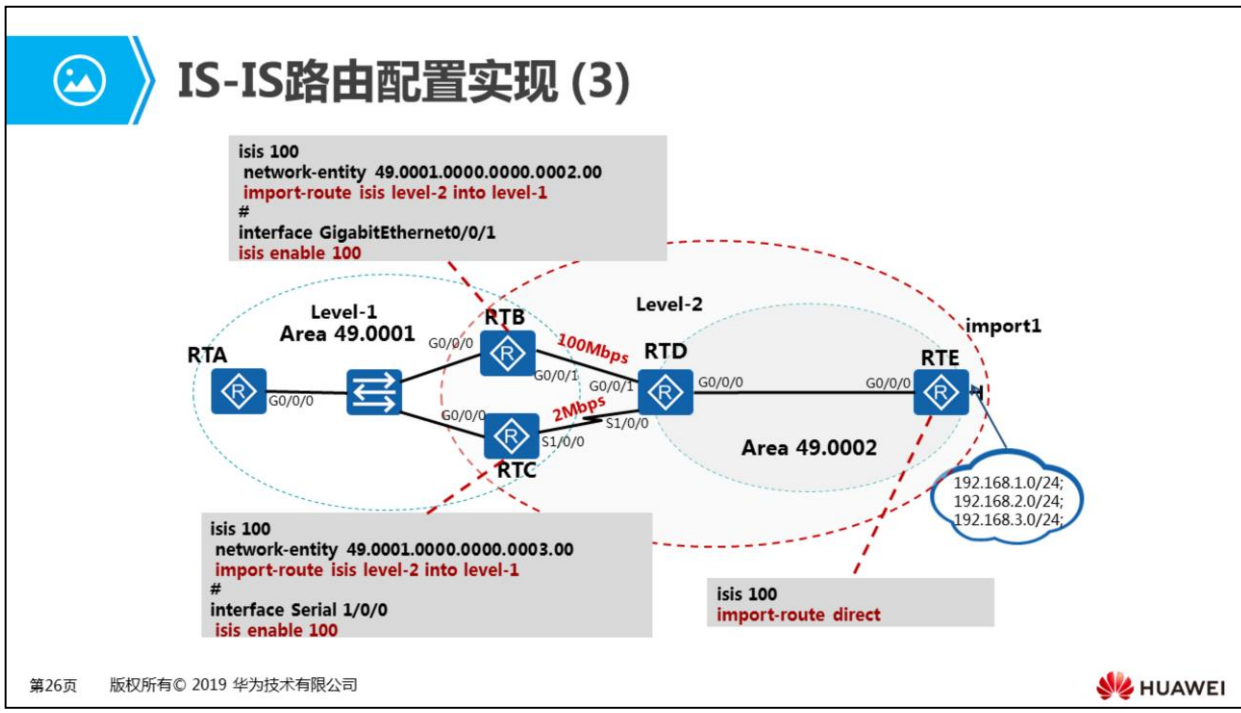
- NET地址编号：
  - RTA：49.0001.0000.0000.0001.00
  - RTB：49.0001.0000.0000.0002.00
  - RTC：49.0001.0000.0000.0003.00
  - RTD：49.0002.0000.0000.0004.00
  - RTE：49.0002.0000.0000.0005.00



- 区域内配置思路：
  - 区域49.0001的业务配置：
    - 每台router进入IS-IS进程100配置网络实体名称NET。
    - RTA在ISIS进程下配置router的level级别为level-1。RTB和RTC默认为level-1-2不用修改。
    - RTA，RTB和RTC在接口下启用ISIS协议。
    - RTA的链路接口修改其DIS的优先级为最高，让其成为DIS。



- 区域内配置思路：
  - 区域49.0002的业务配置：
    - 每台router进入进程100配置网络实体名称NET。
    - RTD和RTE在ISIS进程下配置router的level级别的level-2。
    - RTD和RTE在接口下启用ISIS协议。
    - RTD和RTE在接口修改网络类型为P2P。



- 区域间配置思路：
  - 进入配level-1-2路由器RTB，RTC的ISIS进程配置好网络实体名称NET。
  - 进入链路接口，启用ISIS协议。
  - 进入路由器RTE引入直连链路。
- 路由渗透：
  - 如果一个level-1区域有两个以上Level-1-2路由器，则区域内Level-1路由器访问其他区域会选择最近的Level-1-2路由器，但是计算的开销值只计算本区域内的，如果最近的Level-1-2路由器在Level-2区域到达目的网络的开销相对比较大，实际会造成业务次优路径。在这种场景下需要做路由渗透操作，把Level-2区域的明细路由（包括开销）引入到Level-1区域，由Level-1路由器自行计算选择最优的路径访问跨区域网络。
  - 本实例要求走最优的路径到达区域49.0002，由于RTB连接RTD的链路带宽相对比较大，作用最好让数据流走RTB。可分别在RTB和RTC的ISIS进程下引入level-2的路由到level-1。由RTA的LSDB里面掌握level-2所有的明细路由，就可以选择最优的路径到达区域49.0002。

思考题

1. IS-IS路由器类型有哪几种？
2. PSNP报文在邻居交互中起到了什么作用？
3. 相比OSPF，IS-IS的优势是什么？

HUAWEI

- 答案：IS-IS路由器类型有Level-1路由器，Level-2路由器，Level-1-2路由器。
- 答案：PSNP报文用于LSP的请求和确认。
- 答案：IS-IS报文结构简便，路由承载能力更强，路由算法更优良，扩展性更强。


谢 谢
www.huawei.com

BGP协议原理与配置

HUAWEI

前言

- 在EGP协议中，引入了AS（Autonomous System，自治系统）的概念。AS是指由同一个技术管理机构管理，使用统一选路策略的一些路由器的集合。
- AS的内部使用IGP来计算和发现路由，同一个AS内部的路由器之间是相互信任的，因此IGP的路由计算和信息泛洪完全处于开放状态，人工干预很少。
- 不同AS之间的连接需求推动了外部网关协议的发展，BGP作为一种外部网关协议，用于在AS之间进行路由控制和优选。

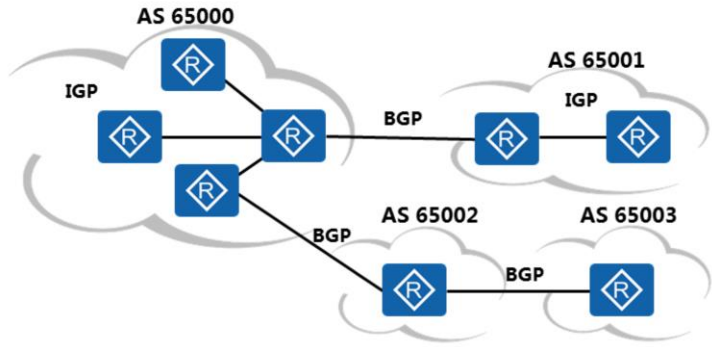
目标

- 学完本课程后，您将能够：
  - 了解BGP基本工作原理
  - 掌握BGP属性及应用原理
  - 熟悉BGP路由聚合的应用场景

目录

1. **BGP基本概述**
2. BGP邻居关系建立与配置
3. BGP路由生成方式
4. BGP通告原则与路由处理
5. BGP常用属性介绍
6. BGP选路原则
7. BGP路由聚合

BGP的基本作用

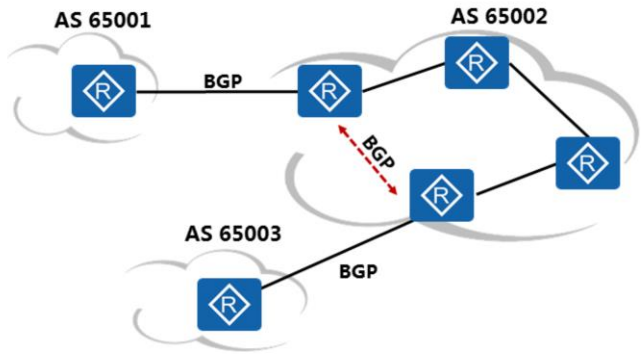


- AS内部使用IGP来计算和发现路由，如OSPF，ISIS，RIP等。
- AS之间使用BGP来传递和控制路由。

HUAWEI

- BGP的前身EGP设计得非常简单，只能在AS之间简单地传递路由信息，不会对路由进行任何优选，也没有考虑如何在AS之间避免路由环路等问题，因而EBP最终被BGP取代。
- 相比于EGP，BGP更具有路由协议的特征，如下：
  - 邻居的发现与邻居关系的建立；
  - 路由的获取，优选和通告；
  - 提供路由环路避免机制，并能够高效传递路由，维护大量的路由信息；
  - 在不完全信任的AS之间提供丰富的路由控制能力。
- 使用BGP作为传递路由的协议，则用户的路由域被作为一个整体和其他路由域进行路由交换，这个路由域即AS。AS的概念是若干台路由器以及这些路由器组成的网络集合，这些路由器均属于同一个管理机构，并执行统一的路由策略。
- 运行BGP协议需要一个统一的自治系统号来标识路由域，即AS编号。每个自治系统都有唯一的一个编号，这个编号由IANA分配。2009年1月之前，只能使用最多2字节长度的AS号码，即1-65535。其中1-64511为公有AS，64512-65534为私有AS。在2009年1月之后，IANA决定使用4字节长度AS，范围是65536-4294967295。

BGP协议特点



- 如图所示，BGP可以跨越多跳路由器建立邻居关系。
- 为实现路由按需求进行控制和优选，BGP设计了诸多属性携带在路由中。

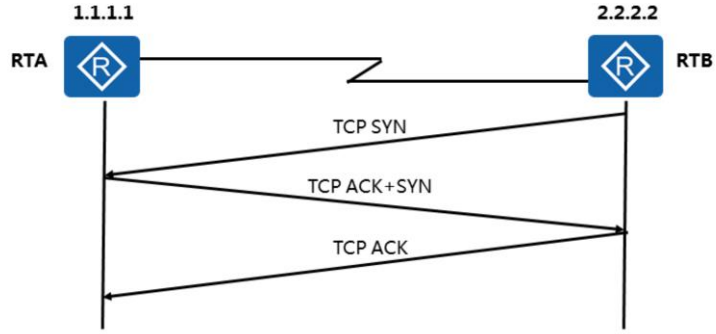
HUAWEI

- 因为是在AS之间传递路由，为保证数据的可靠性， BGP使用TCP作为其承载协议建立连接。因此与IGP逐跳路由器建立邻居不同，BGP可以跨越多跳路由器建立邻居关系。
- AS之间的路由器是不完全相互信任的，为实现路由按需求进行控制和优选，BGP设计了诸多属性。

目录

1. BGP概述
2. **BGP邻居关系建立与配置**
3. BGP路由生成方式
4. BGP通告原则与路由处理
5. BGP常用属性介绍
6. BGP选路原则
7. BGP路由聚合

BGP邻居发现

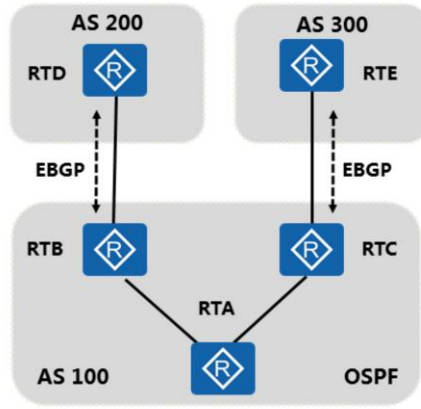


- 先启动BGP的一端先发起TCP连接，如图所示，RTB先启动BGP协议，RTB使用随机端口号向RTA的179端口发起TCP连接。

HUAWEI

- BGP协议被设计运行在AS之间传递路由，AS之间是广域网链路，数据包在广域网上传递是可能出现不可预测的链路拥塞或丢失等情况，因此BGP使用TCP作为其承载协议来保证可靠性。
- BGP使用TCP封装建立邻居关系，端口号为179，TCP采用单播建立连接，因此BGP协议并不像RIP和OSPF一样使用组播发现邻居。单播建立连接也使BGP只能手动指定邻居。

BGP邻居类型 - EBGP

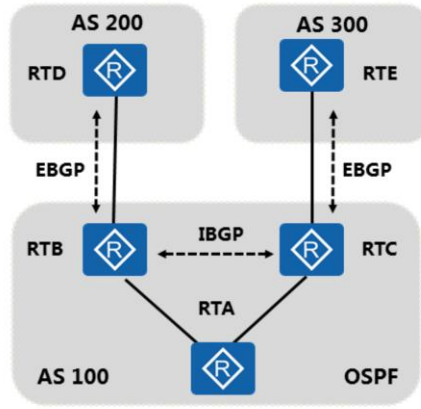


- 运行在不同AS之间的BGP路由器建立的邻居关系为EBGP（External BGP）邻居关系。

HUAWEI

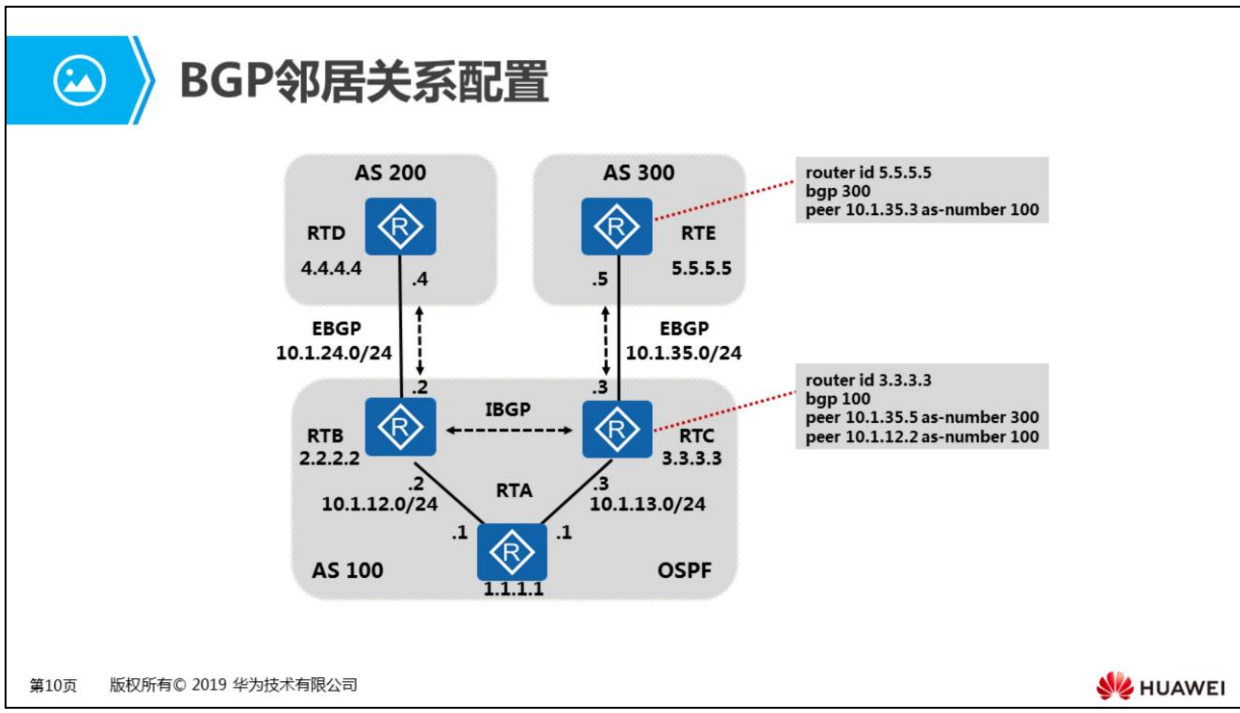
- EBGP只用于不同AS之间传递路由。如图，AS 100内的RTB与BTC分别从AS 200与AS 300学习到不同的路由，怎么实现AS 200与AS 300之间路由在AS 100内的交换？
- 在AS 100内实现将学到的AS 200和AS 300路由进行交换，可以在拓扑中的RTB与RTC路由器上将BGP的路由引入IGP协议（图中为OSPF协议），再将IGP协议的路由在RTB与RTC路由器上引入回BGP协议，实现AS 200与AS 300路由的交换。
- 上述方法存在以下几个缺点：
  - 公网上BGP承载的路由数目非常大，引入IGP协议后，IGP协议无法承载大量的BGP路由；
  - BGP路由引入IGP协议时，需要做严格的控制，配置复杂，不易维护；
  - BGP携带的属性在引入IGP协议时，由于IGP协议不能识别，可能会丢失。
- 因此我们需要设计BGP在AS内部完成路由的传递。

BGP邻居类型 - IBGP



- 运行在相同AS内的BGP路由器建立的邻居关系为IBGP（Internal BGP）邻居关系。

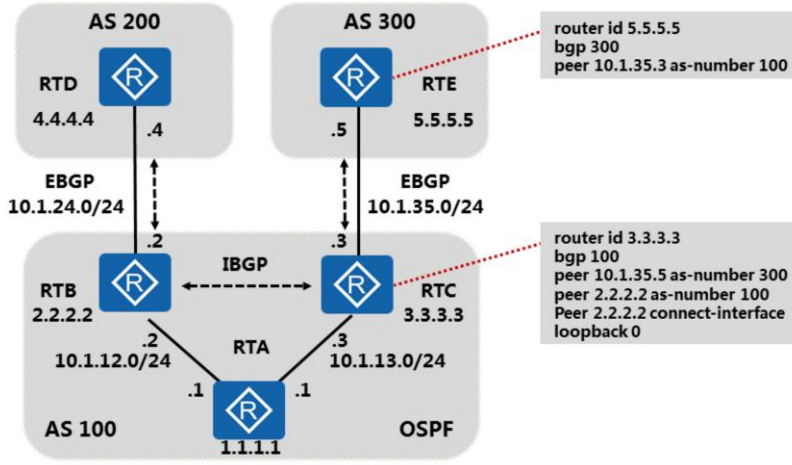
- 如上图，因为BGP使用TCP作为其承载协议，所以可以跨设备建立邻居关系。如图所示，RTB与RTC之间建立IBGP邻居关系，并各自将从其他AS学到的路由传递给对端，实现BGP路由在AS内的传递。



- 配置步骤：
  - 配置Router ID（标识路由器）；
  - 配置EBGP邻居关系（AS之间传递路由）；
  - 配置IBGP邻居关系（AS内部传递路由）。
- 配置解释：
  - 如果没有配置Router ID，BGP路由器会按一定规则自动选举Router ID，选举规则如下：
    - 路由器在它的所有LoopBack接口上选择数值最高的IP地址；
    - 如果没有LoopBack接口，路由器会在它的所有物理接口上选择数值最高的IP地址。
    - 配置命令：router id X.X.X.X
  - BGP邻居关系的类型主要靠配置的AS号区别，peer关键字后面是对端邻居的接口IP地址，as-number后面是邻居路由器所在的AS号，AS号相同则为IBGP邻居关系；AS号不同，则为EBGP邻居关系。
  - peer关键字后面是对端邻居的更新源IP地址，标识自己向对端邻居发起TCP连接的目的地址。该地址可以是对端邻居直连接口的IP地址，也可以是非直连LoopBack接口的IP地址（但必须保证该IP地址路由可达）。建立IBGP邻居关系时，一般使用LoopBack接口的IP地址，因为LoopBack接口开启后一直处于UP状态，只要保证路由可达，邻居关系一直处于稳定状态；而建立EBGP邻居关系时，一般使用直连接口的IP地址，因为EBGP是跨AS建立邻居关系，邻居关系建立之前非直连接口之间的路

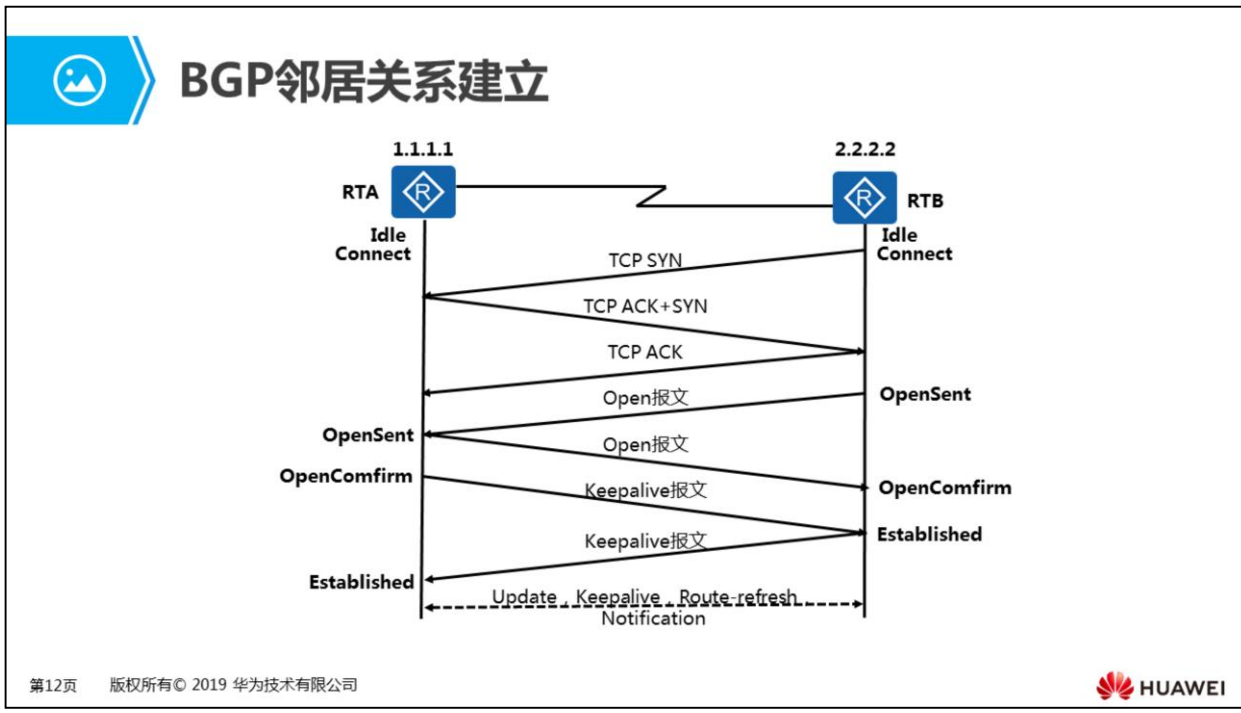
由不可达。

BGP邻居关系配置的优化



HUAWEI

- 建立EBGP邻居关系时，一般使用直连接口的IP地址；建立IBGP邻居关系时，一般使用Loopback接口的IP地址。



- BGP通过报文的交互完成邻居建立、路由更新等操作，共有Open、Update、Notification、Keepalive和Route-refresh等5种报文类型。
  - Open报文：是TCP连接建立后发送的第一个报文，用于建立BGP邻居之间的连接关系。BGP邻居在接收到Open报文并协商成功后，将发送Keepalive报文确认并保持连接的有效性。确认后，BGP邻居间可以进行Update、Notification、Keepalive和Route-refresh报文的交换。
  - Update报文：用于在BGP邻居之间交换路由信息。Update报文可以发布多条属性相同的可达路由信息，也可以撤销多条不可达路由信息。
    - 一条Update报文可以发布多条具有相同路由属性的可达路由，这些路由可共享一组路由属性。所有包含在一个给定的Update报文里的路由属性适用于该Update报文中的NLRI（Network Layer Reachability Information）字段里的所有目的地（用IP前缀表示）。
    - 一条Update报文可以撤销多条不可达路由。每一个路由通过目的地（用IP前缀表示），清楚地定义了BGP路由器之间先前通告过的路由。
    - 一条Update报文可以只用于撤销路由，这样就不需要包括路径属性或者NLRI。相反，也可以只用于通告可达路由，就不需要携带撤销路由信息了。
  - Notification报文：当BGP路由器检测到错误状态时，就向邻居发出Notification报文，之后BGP连接会立即中断。
  - Keepalive报文：BGP路由器会周期性的向邻居发出Keepalive报文，用来保持连接的有效性。
  - Route-refresh报文：Route-refresh用于在改变路由策略后请求对等体重新发送路

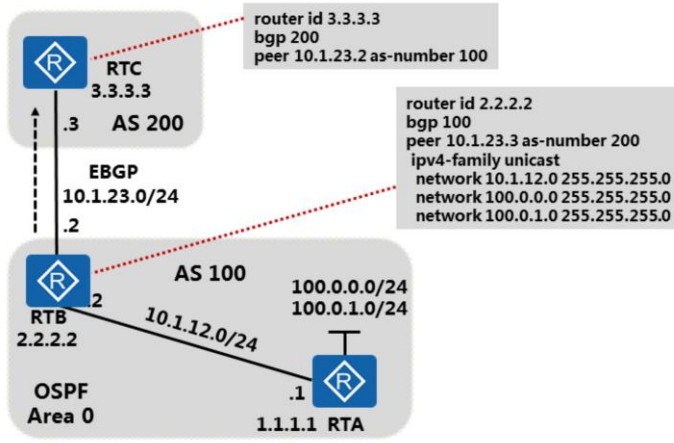
由信息。

- BGP路由器报文交互过程：Idle状态是BGP初始状态。在Idle状态下，BGP路由器拒绝邻居发送的连接请求。只有在收到本设备的Start事件后，BGP路由器才开始尝试与其邻居进行TCP连接，并转至Connect状态。
- 在Connect状态下，BGP路由器启动连接重传定时器（Connect Retry），等待TCP完成连接。
- 如果TCP连接成功，那么BGP路由器向邻居发送Open报文，并转至OpenSent状态。
- 如果TCP连接失败，那么BGP路由器转至Active状态。
- 如果连接重传定时器超时，BGP路由器仍没有收到邻居的响应，那么BGP路由器继续尝试与其邻居进行TCP连接，停留在Connect状态。
- 在Active状态下，BGP路由器总是在试图建立TCP连接。
- 如果TCP连接成功，那么BGP路由器向邻居发送Open报文，关闭连接重传定时器，并转至OpenSent状态。
- 如果TCP连接失败，那么BGP路由器停留在Active状态。
- 如果连接重传定时器超时，BGP路由器仍没有收到邻居的响应，那么BGP路由器转至Connect状态。
- 在OpenSent状态下，BGP路由器等待邻居的Open报文，并对收到的Open报文中的AS号、版本号、认证码等进行检查。
- 如果收到的Open报文正确，那么BGP路由器发送Keepalive报文，并转至OpenConfirm状态。
- 如果发现收到的Open报文有错误，那么BGP路由器发送Notification报文给邻居，并转至Idle状态。
- 在OpenConfirm状态下，BGP路由器等待Keepalive或Notification报文。如果收到Keepalive报文，则转至Established状态，如果收到Notification报文，则转至Idle状态。
- 在Established状态下，BGP路由器可以和邻居交换Update、Keepalive、Route-refresh报文和Notification报文。

目录

1. BGP概述
2. BGP邻居关系建立与配置
3. **BGP路由生成方式**
4. BGP通告原则与路由处理
5. BGP常用属性介绍
6. BGP选路原则
7. BGP路由聚合

BGP路由的生成方式 – Network (1)

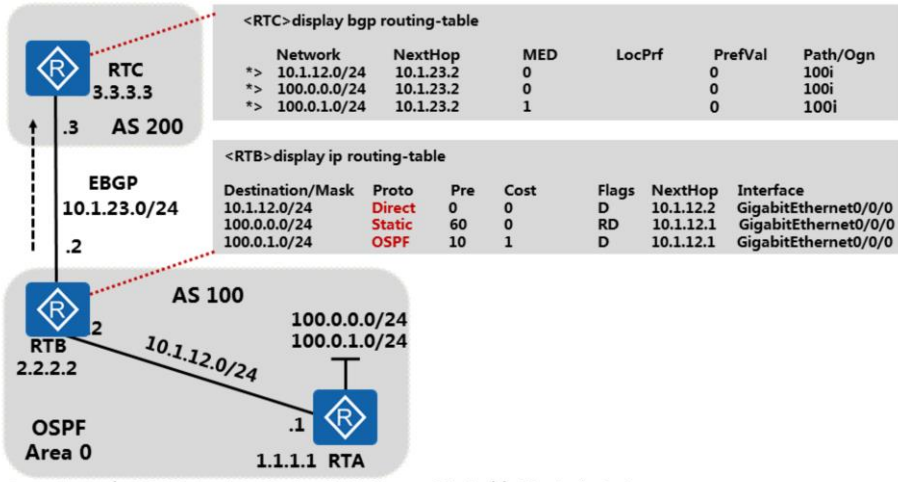


- Network命令是逐条将IP路由表中已经存在的路由引入到BGP路由表中。

  HUAWEI

- 生成BGP路由的方式有两种：第一种是使用配置命令network，第二种是使用配置命令import。
- 如图所示，RTA上存在100.0.0.0/24与100.0.1.0/24的两个用户网段，RTB上通过静态路由指定去往100.0.0.0/24网段的路由，通过OSPF学到去往100.0.1.0/24的路由。RTB与RTC建立EBGP的邻居关系，RTB通过network命令宣告100.0.0.0/24,100.0.1.0/24与10.1.12.0/24的路由，使对端EBGP邻居RTC学习到RTB路由表里的路由。

BGP路由的生成方式 – 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

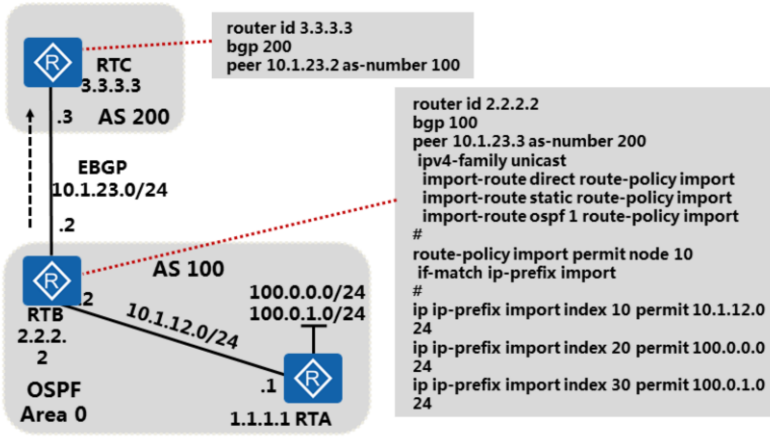
<RTB>display ip routing-table

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

- 通过display命令在RTC上查看是否学到BGP发布的路由条目。

HUAWEI

BGP路由的生成方式 – Import (1)

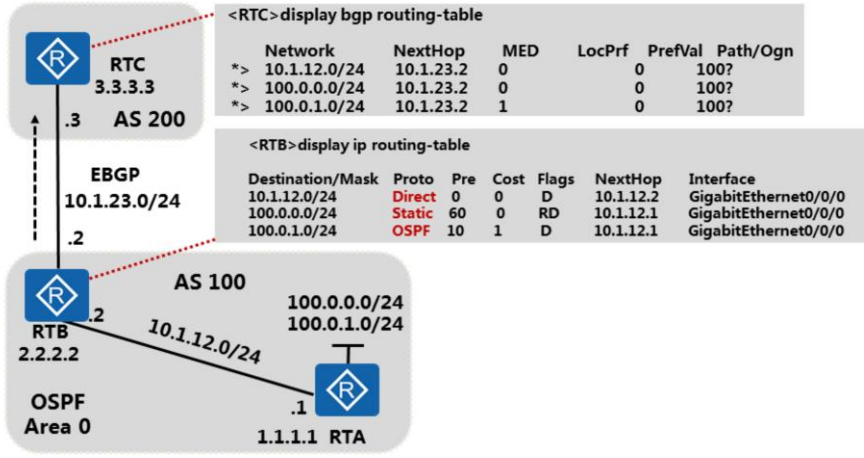


- Import命令是根据运行的路由协议（RIP，OSPF，ISIS等）将路由引入到BGP路由表中，同时import命令还可以引入直连和静态路由。

HUAWEI

- RTA上存在100.0.0.0/24与100.0.1.0/24的两个用户网段，RTB上通过静态路由指定去往100.0.0.0/24网段的路由，通过OSPF学到去往100.0.1.0/24的路由。RTB与RTC建立EBGP的邻居关系，RTB通过import命令宣告100.0.0.0/24,100.0.1.0/24与10.1.12.0/24的路由，使对端EBGP邻居学习到本AS内的路由。
- 为了防止其他路由被引入到BGP中，需要配置ip-prefix进行精确匹配，调用route-policy在BGP引入路由时进行控制。

BGP路由的生成方式 – 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

- 通过display命令在RTC上查看是否学到BGP引入的路由条目。

HUAWEI

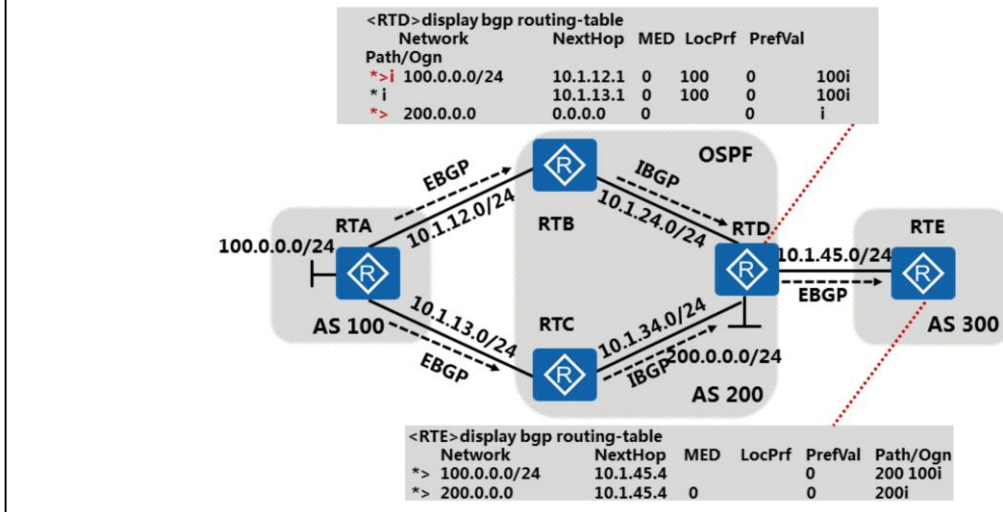
目录

1. BGP概述
2. BGP邻居关系建立与配置
3. BGP路由生成方式
4. **BGP通告原则与路由处理**
5. BGP常用属性介绍
6. BGP选路原则
7. BGP路由聚合

BGP的Update报文

- BGP通过Network和Import两种方式生成BGP路由，BGP路由封装在Update报文中通告给邻居。BGP在邻居关系建立后才开始通告路由信息。
- Update消息主要用来公布可用路由和撤销路由，Update中包含以下信息：
  - 网络层可达信息（NLRI）：用来公布IP前缀和前缀长度。
  - 路径属性：为BGP提供环路检测，控制路由优选。
  - 撤销路由：用来描述无法到达且从业务中撤销的路由前缀和前缀长度。
- 在通告BGP路由时，由于各种因素的影响，为了避免路由通告过程中出现问题，BGP路由通告需要遵守一定的规则，下面进行详细介绍。

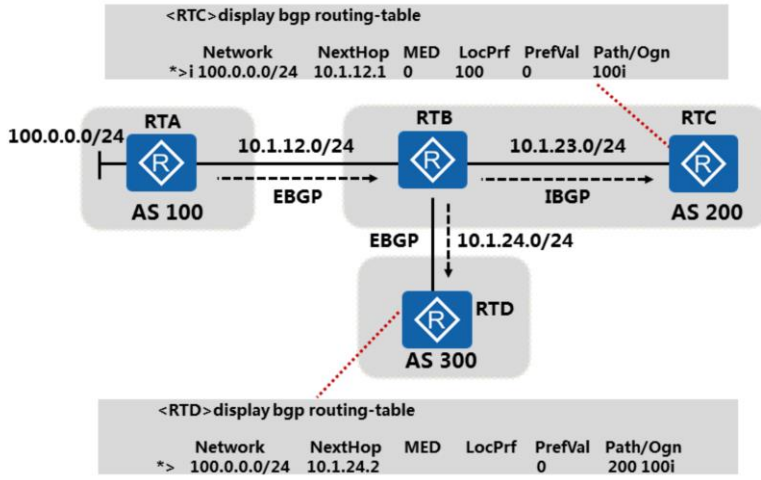
BGP通告原则之一：仅将自己最优的路由发布给邻居



  HUAWEI

- 存在多条有效路由时，BGP路由器只将自己最优的路由发布给邻居。
  - RTD可以从BGP邻居RTB与RTC学习到100.0.0.0/24的路由，同时RTD将自己的直连路由200.0.0.0/24发布到BGP中。在RTD上使用命令display bgp routing-table查看如图所示；
  - 在RTE上使用命令display bgp routing-table查看如图所示。可以发现，RTD将自己标为有效且最优的路由发布给了BGP邻居RTE。
- BGP路由表中的状态含义：
  - Status codes: * - valid, > - best, d - damped, h - history, i - internal, s - suppressed, S - Stale
  - Origin : i - IGP, e - EGP, ? – incomplete
  - Network：显示BGP路由表中的网络地址
  - NextHop：报文发送的下一跳地址
  - MED：路由度量值
  - LocPrf：本地优先级
  - PrefVal：协议首选值
  - Path/Ogn：显示AS路径号及Origin属性
  - Community：团体属性信息

BGP通告原则之二：通过EBGP获得的最优路由发布给所有BGP邻居

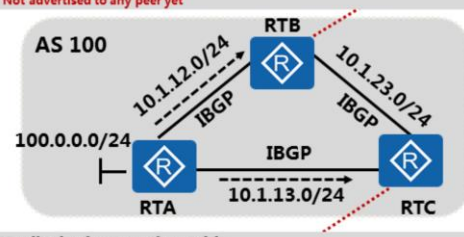


HUAWEI

- BGP路由器通过EBGP获得的最优路由会发布给所有的BGP邻居（包括EBGP邻居和IBGP邻居）。
  - 如图所示，RTA上有一个100.0.0.0/24的用户网段，并通过EBGP将该网段发布给BGP邻居RTB。RTB收到EBGP邻居发送来的100.0.0.0/24的路由后，将会通告给自己的IBGP邻居RTC与EBGP邻居RTD。

BGP通告原则之三：通过IBGP获得的最优路由不会发布给其他的IBGP邻居

```
<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
```

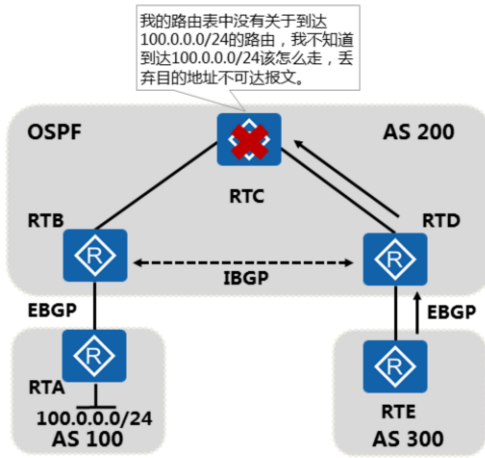



```
<RTC>display bgp routing-table
     Network         NextHop      MED        LocPrf    PrefVal  Path/Ogn
*>i  100.0.0.0/24    10.1.13.1    0          100       0        i
```


HUAWEI

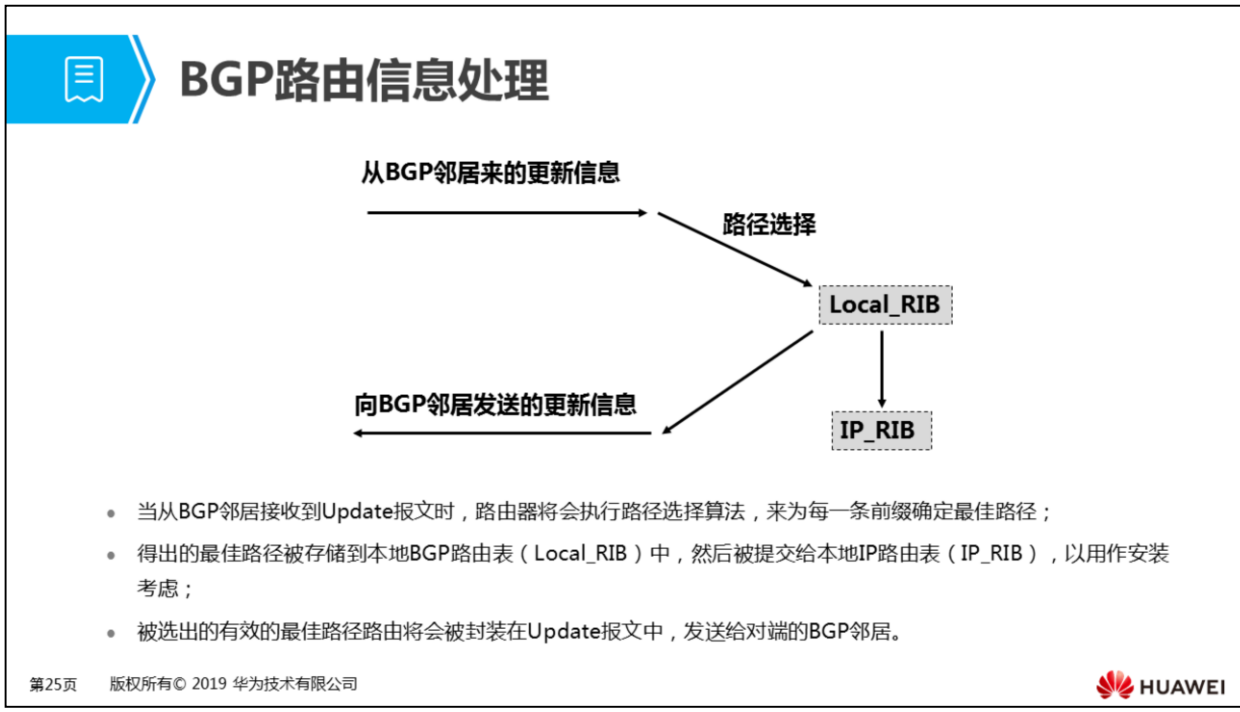
- BGP路由器通过IBGP获得的最优路由不会发布给其他的IBGP邻居。
  - 如图所示，RTA上存在一个100.0.0.0/24的用户网段，RTA、RTB与RTC之间互为IBGP邻居，RTA通过IBGP将100.0.0.0/24的路由发布给RTB与RTC，但是RTB并不会将收到的IBGP路由发布给自己的IBGP邻居RTC。
  - 这样设计的目的是防止在AS内部形成路由环路。根据规定，BGP路由在同一个AS内进行传递时，AS_Path属性不会发生变化。如图所示，RTA将100.0.0.0/24的路由发布给RTB时，AS_Path属性不变，为空。如果RTB能将IBGP路由100.0.0.0/24发布给RTC，AS_Path依旧为空。则RTC也有可能将100.0.0.0/24的路由发布给RTA，因为AS_Path为空，RTA并不会拒收该IBGP路由，路由环路产生。因此，上述通告原则是为了防止在AS内部形成路由环路。

BGP通告原则之四：BGP与IGP同步



  HUAWEI

- RTA上存在一个100.0.0.0/24的用户网段，通过EBGP发布给RTB。RTB与RTD建立了IBGP邻居关系，RTD通过IBGP学习到该BGP路由，并将该路由发布给EBGP邻居RTE。
- 当RTE访问100.0.0.0/24的路由时，查找路由表，发现到达100.0.0.0/24路由的下一跳是RTD，RTE查找出接口后，将数据包发送给RTD；RTD收到数据包后，查找路由表，发现到达100.0.0.0/24路由的下一跳是RTB，出接口是RTD上与RTC相连的接口，于是将数据包发给RTC，RTC查找路由表，发现没有到达100.0.0.0/24的路由，于是将数据丢弃，形成“路由黑洞”。
- BGP的通告原则：一条从IBGP邻居学来的路由在发布给一个BGP邻居之前，通过IGP必须知道该路由，即BGP与IGP同步。
    - 如图所示，RTD在收到RTB发来的IBGP路由之后，如果要发布给BGP邻居RTE，则在发布之前先检查IGP协议（即OSPF协议）能否学到该条路由。如果能，则将IBGP路由发布给RTE。
    - 在华为路由器上，默认是将BGP与IGP的同步检查关闭的，原因是为了实现IBGP路由的正常通告。但关闭了BGP与IGP的同步检查后会出现“路由黑洞”的问题。因此，有两种解决方案解决上述问题：
    - 将BGP路由引入到IGP，从而保证IGP与BGP的同步。但是，因为Internet上的BGP路由数量十分庞大，一旦引入到IGP，会给IGP路由器带来巨大的处理和存储负担，如果路由器负担过重，则可能瘫痪。
    - IBGP路由器必须是全互联，确保所有的路由器都能学习到通告的路由。这样可以解决关闭同步后导致的“路由黑洞”问题。

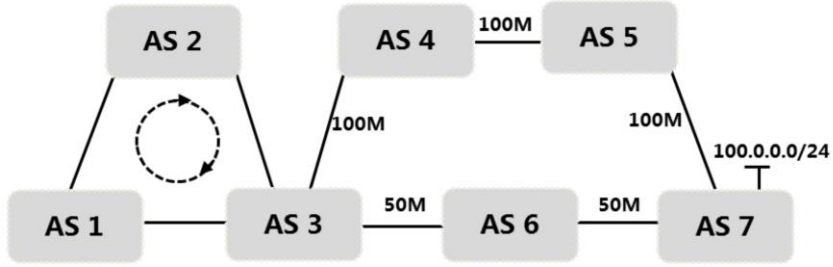


- IP路由表（IP_RIB）：全局路由信息库，包括所有的IP路由信息。
- BGP路由表（Local_RIB）：BGP路由信息库，包括本地BGP路由器选择的路由信息，邻居表，邻居清单列表。
- 收到BGP邻居发来的Update报文，路由器会执行算法进行路径选择，确定每一条前缀的最佳路径，并将计算出的最佳路径存储到本地BGP路由表（Local_RIB）中。
- 如果启用了多路径特性，最佳路径和所有等值路径都被提交给IP_RIB，考虑是否安装。除了从BGP邻居接收的最佳路径外，Local_RIB也包含当前路由器注入的路由（被称为本地发起的路由）。
- 在Local_RIB中，只有被选为最优的前缀才会被封装到Update报文中通告给自己的BGP邻居。

目录

1. BGP概述
2. BGP邻居关系建立与配置
3. BGP路由生成方式
4. BGP通告原则与路由处理
5. **BGP常用属性介绍**
6. BGP选路原则
7. BGP路由聚合

BGP选路遇到的问题



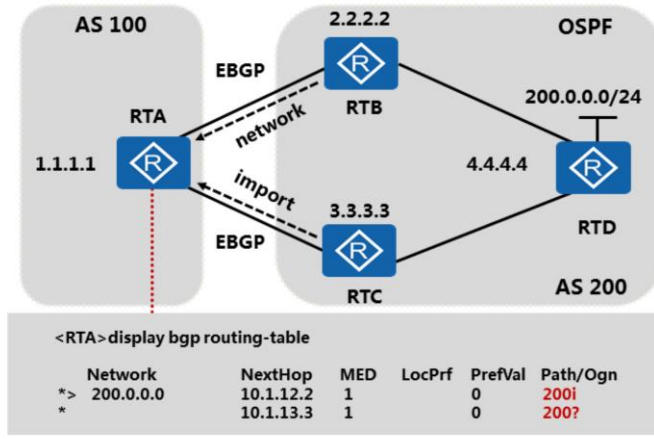
- 如图，AS 7中有一个100.0.0.0/24的用户网段，通过BGP发布给各个AS，各个AS都能学到100.0.0.0/24的路由，但是路由在传递过程中存在两个主要的问题：
  - AS 3可以从AS 4与AS 6两个AS收到100.0.0.0/24的路由，但AS 3与AS 4之间的链路带宽较大，有哪些方法可以影响AS 3选择AS 4访问100.0.0.0/24的网段？
  - AS 1，AS 2与AS 3之间存在拓扑上的环路，因此数据包在传递的过程中可能出现环路，怎么解决类似的环路问题？

HUAWEI

- 以上两个问题的解决方案：
  - 在AS之间交换路由可达信息时，设计BGP能够提供丰富的属性，实现对路由的灵活控制和优选。
    - 修改路由表，调整AS之间的链路Metric；2.不修改路由表， 使用策略修改路由下一跳。但是这些方法在某些情况下具有局限性，不能满足网络的丰富需求。
  - 路由在AS之间传递时记录传播路径，防止环路的产生。

- 公认属性：所有BGP路由器都必须识别并支持的属性。
  - 公认必遵：BGP的Update消息中必须包含的属性。
  - 公认任意：不必存在于BGP的Update消息中，可以根据需求自由选择的属性。
- 可选属性：不要求所有的BGP路由器都能够识别的属性。
  - 可选过渡：BGP不能识别该属性，但可以接收该属性并将其发布给它的邻居的属性。
  - 可选非过渡：BGP可以忽略包含该属性的消息并且不向它的邻居发布。

BGP属性 - Origin



<RTA>display bgp routing-table

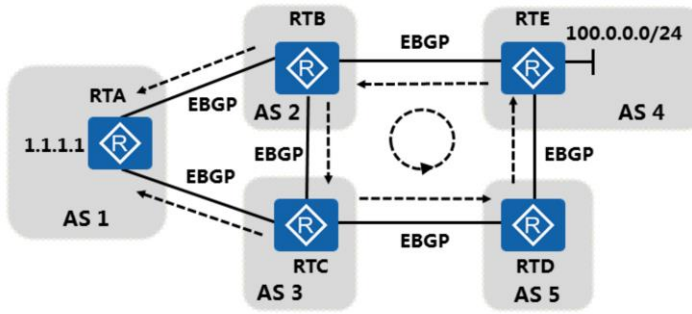
Network	NextHop	MED	LocPrf	PrefVal	Path/Ogn
*> 200.0.0.0	10.1.12.2	1		0	200i
*	10.1.13.3	1		0	200?

- Origin属性定义路径信息的来源，标记一条路由是怎么成为BGP路由的。

HUAWEI

- 如图所示，AS 200内运行OSPF协议，200.0.0.0/24网段宣告到OSPF中。RTB通过network方式将200.0.0.0/24的路由变为BGP路由通告给RTA，RTC通过import方式将200.0.0.0/24的路由变为BGP路由通告给RTA。
- BGP在AS之间传递信息，承载大量的路由。如果到达同一目的IP有多条路径，且BGP学到这些路由通过不同的方式，则Origin属性是决定最优路径的一个因素，用于标明路由的起源。
- Origin的3种属性：
  - i表明BGP路由通过network命令注入；
  - e表明BGP路由是从EGP学来的，EGP协议在现网中很难见到，但可以通过路由策略将路由的Origin属性修改为e；
  - ？即Incomplete表明BGP路由通过其它方式学到路由信息，如使用import命令引入的路由。
- 3种Origin属性的优先级为：i>e>Incomplete（ ？ ）。

BGP属性 - AS_Path

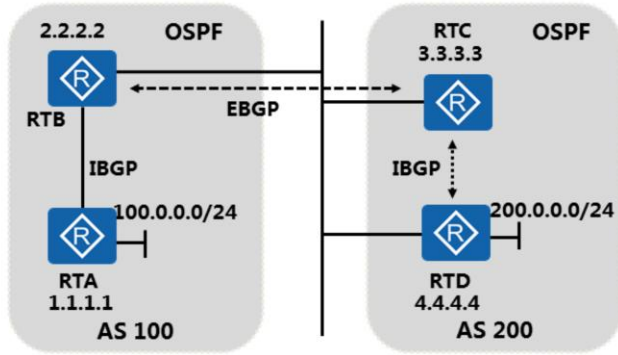


- 如图所示：
  - AS 1内的RTA能够从RTB与RTC收到100.0.0.0/24的路由，RTA如何进行自动优选？
  - RTA->RTB->RTC之间在拓扑上存在环路，RTB->RTC->RTD->RTE之间在拓扑上也存在环路，因此BGP在路由传递的过程中也可能存在路由环路，BGP如何防止环路呢？

  HUAWEI

- BGP针对以上2个问题，设计了AS_Path属性，该属性记录了路由经过的所有AS的编号：
  - 图中RTA从RTB收到100.0.0.0/24的路由时，AS_Path为（2，4），RTA从RTC收到100.0.0.0/24的路由时，AS_Path为（3，5，4）。规定AS_Path越短（记录的AS编号越少），路径越优，因此RTA会优选从RTB收到的100.0.0.0/24的路由。
  - 以RTE为例，通过BGP发布100.0.0.0/24的路由，路由可能通过RTE->RTB->RTC->RTD->RTE形成环路。为了防止环路的产生，RTE在收到RTD发来的路由时会检查AS_Path（该路由携带的）属性，如果发现该路由的AS_Path中包含自己的AS号，则丢弃该路由。
- AS_Path的4种类型：
  - AS_Sequence（后续讲解BGP路由聚合时会详细说明）；
  - AS_Set（后续讲解BGP路由聚合时会详细说明）；
  - AS_Confed_Sequence（应用于联盟，本课程不涉及）；
  - AS_Confed_Set（应用于联盟，本课程不涉及）。

BGP属性 - Next_hop

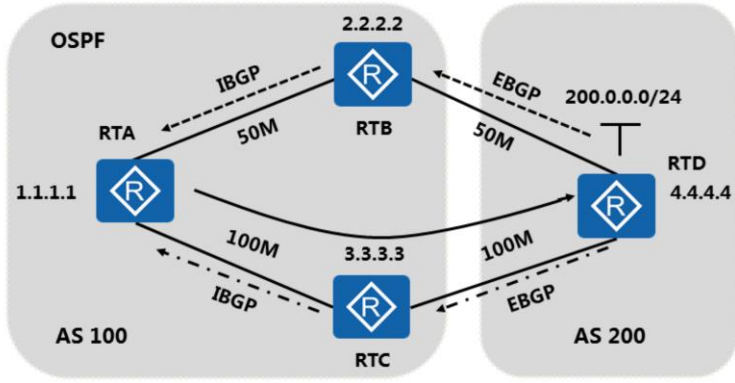


- 如图所示：
  - RTA将100.0.0.0/24的网段发布给RTB时，Next_hop的IP地址是多少？
  - RTB将100.0.0.0/24的网段发布给RTC时，Next_hop的IP地址是多少？
  - RTA从RTB学到RTC发布的200.0.0.0/24的网段时，Next_hop的IP地址是多少？

HUAWEI

- BGP路由器将本端始发路由发布给IBGP邻居时，会把该路由信息的Next_hop设为本端建立邻居关系所使用的接口IP。
  - 如图所示，RTA将100.0.0.0/24的网段发布给RTB时，如果RTA与RTB使用直连接口建立IBGP邻居，则Next_hop为RTA上与RTB直连的接口IP；如果RTA与RTB使用Loopback接口建立IBGP邻居，则Next_hop为RTA的Loopback接口IP。
- BGP路由器在向EBGP邻居发布路由时，会把路由信息的Next_hop设置为本端与对端建立BGP邻居关系的接口IP。
  - 如图所示，RTB将100.0.0.0/24的网段发布给RTC时，Next_hop为RTB上与RTC直连的接口IP。
- BGP路由器在向IBGP邻居通告从EBGP学来的路由时，不改变该路由下一跳属性。
  - 特例：如图所示，RTA从RTB学到RTC发布的200.0.0.0/24的网段时，Next_hop为RTD的出接口IP，因为RTB与RTD在同一网段，RTC通告给RTB的Next_hop为RTD的出接口IP。
- 对于上述三种情况的解释：
  - EBGP邻居之间一般采用直连接口建立邻居关系，EBGP邻居在相互通告路由时会修改Next_hop为自己的出接口IP；
  - IBGP邻居通常采用Loopback接口建立邻居，当路由是本路由器起源的，在发送给邻居之后Next_hop改为自己的更新源地址，这样即使网络中出现链路故障，只要Next_hop可达，同样可以访问目的网段，提高网络稳定性；
  - 相对于IGP，如RIP在发布路由时，每经过一个路由器都会修改下一跳，发布路由的路由器都宣称自己能够到达目标地址，并采用逐跳传递的方式将数据包发送给目标网络，但网络中的路由器并不知道谁是真正的始发路由器，因此会造成环路。BGP在EBGP之间传递时才修改Next_hop，IBGP发送从EBGP学来的路由给IBGP邻居时并不修改下一跳，在一定程度上起到了防环作用。

BGP属性 - Local_Preference

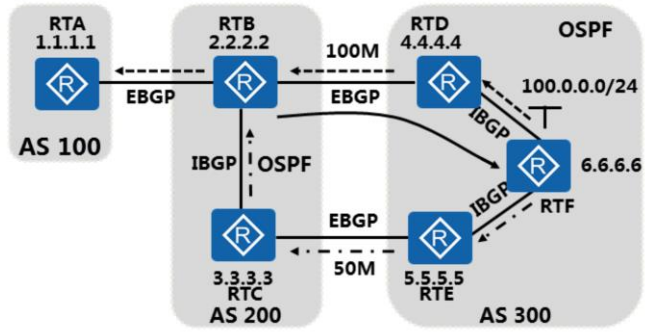


- Local_Pref属性仅在IBGP邻居之间有效，不通告给其他AS。它表明路由器的BGP优先级，用于判断流量离开AS时的最佳路由。

  HUAWEI

- 如图所示，AS 200内有一个200.0.0.0/24的用户网段，通过BGP发布给AS 100。AS 100内的管理员如何设置才可以实现通过高带宽链路访问200.0.0.0/24的网络？
- 解决办法：
  - 在RTC上设置ip-prefix匹配200.0.0.0/24的路由，使用route-policy调用该ip-prefix，并设置Local_Preference为200，将策略应用在对RTA发布路由的export方向。
- Local_Pref属性仅在IBGP邻居之间有效，不通告给其他AS。它表明路由器的BGP优先级，值越大越优。
- Local_Pref属性用于判断流量离开AS时的最佳路由。当BGP路由器通过不同的IBGP邻居获得目的地址相同但下一跳不同的多条路由时，将优先选择Local_Pref属性值较高的路由，其默认值为100。

BGP属性 - MED

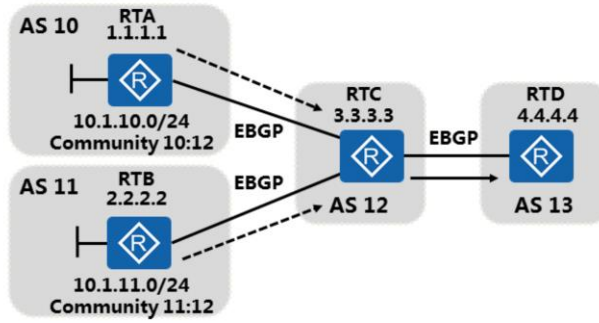


- MED（Multi-Exit-Discriminator）属性仅在相邻两个AS之间传递，收到此属性的AS不会再将其通告给任何其他第三方AS，用于判断流量进入AS时的最佳路由。

  HUAWEI

- 如图所示，AS 300内的管理员希望在AS 300内操作影响AS 200通过高带宽链路访问100.0.0.0/24，如何实现？
- 解决方法：
  - 在RTE上设置ip-prefix匹配100.0.0.0/24的路由，再设置route-policy调用该ip-prefix，并设置MED为100，将策略应用在对RTC发布路由的export方向。
- MED（Multi-Exit-Discriminator）属性仅在相邻两个AS之间传递，收到此属性的AS不会再将其通告给任何其他第三方AS。如图所示，AS100内并不会收到AS 300内设置的MED值，但是AS 200内会收到AS 300内设置的MED值，因此AS 200内可以选择高带宽的路由。
- MED属性相当于IGP使用的度量值（Metric），它用于判断流量进入AS时的最佳路由。当一个运行BGP的路由器通过不同的EBGP邻居获得目的地址相同但下一跳不同的多条路由时，在其它条件相同的情况下，将优先选择MED值较小者作为最佳路由，其默认值为0。

BGP属性 - Community



- BGP的Community属性的两个作用：
  - 限定路由的传播范围。
  - 打标记，便于对符合相同条件的路由进行统一处理。

HUAWEI

- 如图所示，AS 10内有10.1.10.0/24的用户网段，AS 11内有10.1.11.0/24的用户网段。为了区分用户网段，AS 10内的10.1.10.0/24设置了10:12的Community，AS 11的10.1.11.0/24设置了11:12的Community，通过BGP发送给AS 12后，AS 12希望汇总后屏蔽掉明细路由再发送给AS 13，并且希望AS 13收到路由后不再传递给其他AS，如何实现？
- 解决方法：
  - 在RTC上设置Community-filter，匹配Community为10:12和11:12的路由，再设置route-policy匹配Community-filter，将两条路由聚合成10.1.10.0/23的路由并调用route-policy。
  - 在RTC上设置route-policy，设置团体属性为no-export，在RTC通告给RTD的export方向调用该route-policy。

- Community属性分为两类：一类是公认团体属性，另一类是扩展的团体属性。
- 公认团体属性分为4类：
    - Internet：缺省属性，所有路由都属于Internet，此属性的路由可以通告给所有BGP邻居；
    - No_Export：收到此属性的路由后，不将该路由发布到其他AS。如图，RTB上希望10.1.11.0/24的路由发布给AS 12之后，不再发布给其他AS，则可将10.1.11.0/24的Community属性设置为No_Export；
    - No_Advertise：收到此属性的路由后，不将该路由通告给任何其他的BGP邻居。如图，RTB上希望只将10.1.11.0/24的路由发布给RTC，并且不再通告给任何其他的BGP邻居，则可将10.1.11.0/24的Community属性设置为No_Advertise；
    - No_Export_Subconfed：在联盟中使用，这里不做介绍。
- 扩展的团体属性用一组4字节为单位的列表来表示，路由器中扩展的团体属性格式为aa:nn或团体号：
- aa:nn中，aa通常为AS编号，nn是管理员定义的团体属性标识；
- 团体号范围为0-4294967295，在RFC1997中，0-65535与4294901760-4294967295为预留值。

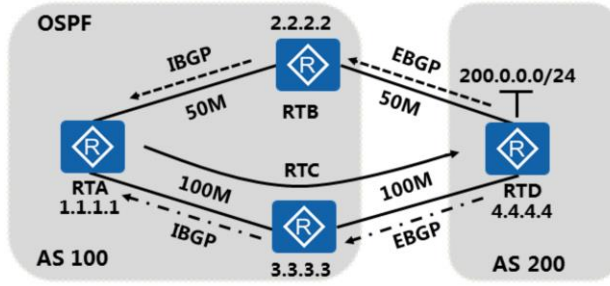
目录

1. BGP概述
2. BGP邻居关系建立与配置
3. BGP路由生成方式
4. BGP通告原则与路由处理
5. BGP常用属性介绍
6. **BGP选路原则**
7. BGP路由聚合

BGP路由优选原则

- BGP路由器将路由通告给邻居后，每个BGP邻居都会进行路由优选，路由选择有三种情况：
  - 该路由是到达目的地的唯一路由，直接优选。
  - 对到达同一目的地的多条路由，优选优先级最高的。
  - 对到达同一目的地且具有相同优先级的多条路由，必须用更细的原则去选择一条最优的。
- 一般来说，BGP计算路由优先级的规则如下：
  - 丢弃下一跳不可达的路由。
  - 优选Preference_Value值最高的路由（私有属性，仅本地有效）。
  - 优选本地优先级（Local_Preference）最高的路由。
  - 优选手动聚合>自动聚合>network>import>从对等体学到的。
  - 优选AS_Path短的路由。
  - 起源类型IGP>EGP>Incomplete。
  - 对于来自同一AS的路由，优选MED值小的。
  - 优选从EBGP学来的路由（EBGP>IBGP）。
  - 优选AS内部IGP的Metric最小的路由。
  - 优选Cluster_List最短的路由。
  - 优选Orginator_ID最小的路由。
  - 优选Router_ID最小的路由器发布的路由。
  - 优选具有较小IP地址的邻居学来的路由。

Preference_Value对选路的影响

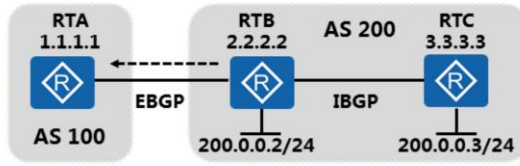


- Preference_Value是BGP的私有属性（华为私有属性），Preference_Value相当于BGP选路规则中Weight值，仅在本地路由器生效。Preference_Value值越大，越优先。

HUAWEI

- 如图所示，AS 200内有一个200.0.0.0/24的用户网段，AS 100内的管理员希望通过高带宽链路访问AS 200内的200.0.0.0/24网段，并希望在RTA上的策略只能影响自己的选路，不能影响其他设备，如何实现？
- 解决办法：
  - 在RTA上设置ip-prefix匹配200.0.0.0/24的路由，再设置route-policy调用该ip-prefix，并设置Preference_Value为100，将策略应用在对RTC发布路由的 import方向。
- 验证：RTC上使用Tracert命令，查看访问200.0.0.0/24网段经过的路由器。

聚合方式对选路的影响



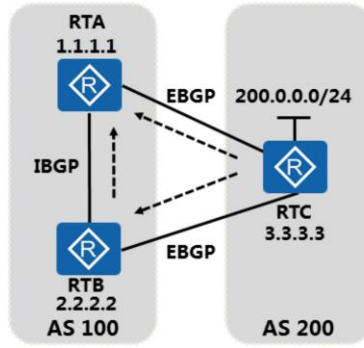

```
<RTB>display bgp routing-table 200.0.0.0
BGP local router ID : 2.2.2.2
 Local AS number : 200
 Paths:   2 available, 1 best, 1 select
 BGP routing table entry information of 200.0.0.0/24:
 Aggregated route.
 ......
 Aggregator: AS 200, Aggregator ID 2.2.2.2, Atomic-aggregate
 Advertised to such 2 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
```

- 聚合路由的优先级：手动聚合>自动聚合。

HUAWEI

- 如图所示，在AS 200内，RTB与RTC上存在200.0.0.0/24网段的用户，RTB与RTC将200.0.0.0/24的网段通过import方式变为BGP路由，在RTB上将路由聚合后发给RTA，同时开启自动聚合与手动聚合，RTB如何优选聚合路由？
- 如图所示，在RTB上同时使能自动聚合与手动聚合，使用命令查看，可以发现，手动聚合的路由条目被发送给RTA，自动聚合的路由条目则没有通告，说明手动聚合的优先级高于自动聚合。
- 在使用路由聚合时需要注意，自动聚合只能对引入的BGP路由进行聚合，手动聚合可以对存在于BGP路由表中的路由进行聚合，后续在BGP路由聚合中详细介绍。上述场景中，因为需要聚合的路由都是引入的路由，所以使用自动聚合与手动聚合都可以实现聚合的目的。如果BGP路由表中既有引入的路由又有network宣告的路由时，只能采用手动聚合实现。

EBGP邻居的路由优于IBGP邻居的路由

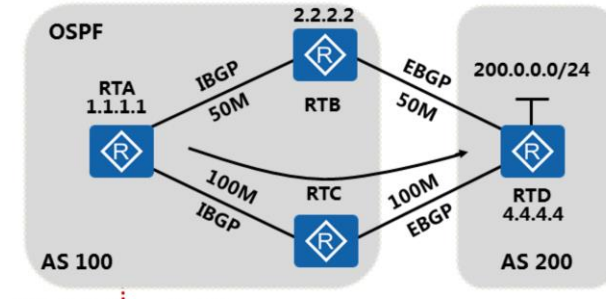


- 根据选路原则，RTA会优选从EBGP邻居学来的路由。

HUAWEI

- 如图所示，在AS 200内有一个200.0.0.0/24的网段，通过EBGP邻居关系通告给RTA与RTB，RTB会通过IBGP邻居关系将200.0.0.0/24的网段通告给RTA，于是RTA会收到两条到达200.0.0.0/24的路由，RTA会如何优选？
- 根据选路原则，RTA会优选从EBGP邻居学来的路由。

AS内部IGP Metric对BGP选路的影响



<RTA>display bgp routing-table

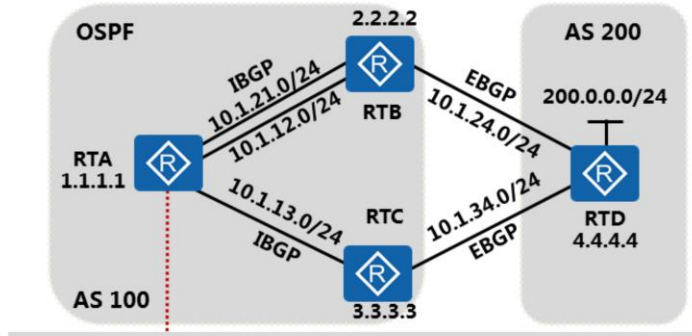
	Network	NextHop	MED	LocPrf	PrefVal	Path/Ogn
*>i	200.0.0.0	10.1.34.4	0	100	0	200 i
* i	10.1.24.4	0	100	0	200 i	

- 如图所示，通过调整OSPF Cost，使RTA选择高带宽路径访问200.0.0.0/24网段。

HUAWEI

- 如图所示，AS 200内有一个200.0.0.0/24的用户网段，通过EBGP发布给RTB与RTC，RTB与RTC通过IBGP将路由发布给RTA。AS 100内的管理员希望通过高带宽链路访问AS 200内的200.0.0.0/24网段， RTA上该如何实现？
- 将RTA与RTB所连接口的OSPF Cost值调为100，RTA则将选择RTA->RTC->RTD的路径访问200.0.0.0/24网段：
  - 原因是RTA访问200.0.0.0/24时，到Next_hop 10.1.34.4的Cost（2）小于到Next_hop 10.1.24.4 (101)的Cost。

Router-ID与IP地址对BGP选路的影响



<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

- 如图所示，RTA选择通过RTB访问AS内的200.0.0.0/24的网段，出接口为10.1.12.1地址所在的接口。

HUAWEI

- 如图所示，AS 200内有一个200.0.0.0/24的用户网段，通过EBGP发布给RTB和RTC，RTB和RTC通过IBGP将路由发布给RTA。RTA和RTB之间通过2条链路相连，RTA会如何优选？
- RTA会选择下一跳为10.1.12.2作为下一跳访问200.0.0.0/24的网段：
  - RTA选择RTA->RTB->RTD的路径访问200.0.0.0/24网段，原因是RTB的Router-ID比RTC小，BGP优选Router-ID较小的路由器发布的路由；
  - RTA选择下一跳为10.1.12.2地址所在的接口为出接口，原因是BGP优选IP地址较小的邻居学来的路由。
- 在RTA上使用命令display bgp routing-table 200.0.0.0查看如下：

<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

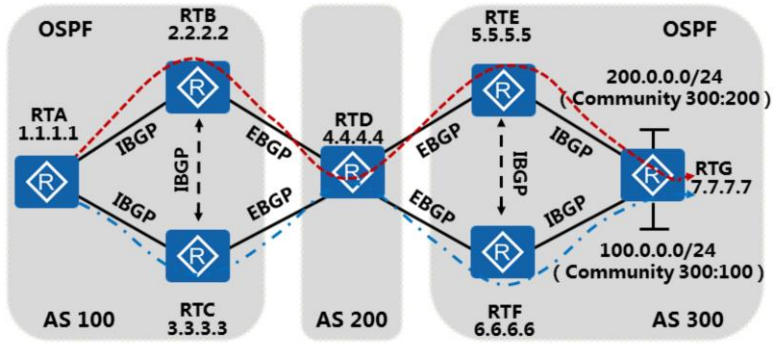
Original nexthop: 10.1.24.4

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

……

BGP路由策略配置实例



- 如图所示，AS 300内有两个用户网段，AS 100内用户访问这两个网段时，希望在RTB和RTC上实现流量分担。AS 200访问这两个网段时，希望在RTE和RTF上实现流量分担。请用尽可能多的方法来实现上述需求。

HUAWEI

- 如图所示，AS 300内有两个用户网段，一个是200.0.0.0/24，一个是100.0.0.0/24。为了区分不同网段的用户，在AS 300内为100.0.0.0/24的网段分配Community属性为300:100，为200.0.0.0/24的网段分配Community属性为300:200。AS 100内用户访问这两个网段时，希望在RTB和RTC上实现流量分担。AS 200访问这两个网段时，希望在RTE和RTF上实现流量分担。请用尽可能多的方法来实现上述需求。
- 根据需求，在AS 100访问这两个网段时，希望在RTB和RTC上实现流量分担；在AS 200访问这两个网段时，希望在RTE和RTF上实现流量分担。假设RTA访问100.0.0.0/24时的路径为RTA->RTB->RTD->RTE->RTG，访问200.0.0.0/24时的路径为RTA->RTC->RTD->RTF->RTG，根据所学路径属性的知识，可供参考解决方案如下：
  - RTE和RTF向RTD通告携带团体属性的路由；
  - RTD收到携带团体属性的路由后，使用两个Community-filter分别匹配不同的团体属性，再使用两个route-policy分别调用Community-filter，将匹配团体属性300:100的路由的下一跳设为RTE上的出接口地址；将匹配团体属性300:200的路由的下一跳设为RTF上的出接口地址；
  - RTD上再设置两个route-policy，一个是将匹配团体属性为300:100的路由设置其MED值为100，在对RTC的export方向调用；另一个是匹配团体属性为300:200的路由并设置其MED值为100，在对RTB的export方向调用。

- RTD上的配置：

```
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

验证：在RTA上执行如下命令：

tracert 100.0.0.1，观察所经过的IP地址。

tracert 200.0.0.1，观察所经过的IP地址。

- 其他方法有待大家探索。

目录

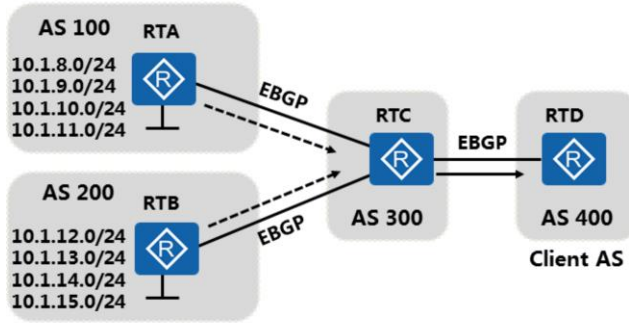
1. BGP概述
2. BGP邻居关系建立与配置
3. BGP路由生成方式
4. BGP通告原则与路由处理
5. BGP常用属性介绍
6. BGP选路原则
7. **BGP路由聚合**

HUAWEI

BGP路由聚合概述

- BGP在AS之间传递路由信息，随着AS数量的增多，单个AS规模的扩大，BGP路由表将变得十分庞大，因此带来如下两类问题：
  - 存储路由表将占用大量的内存资源，传输和处理路由信息需要消耗大量的带宽资源；
  - 如果传输的路由条目出现频繁的更新和撤销，对网络的稳定性会造成影响。
- 本节将介绍BGP的路由聚合对上述两种问题的处理，下面我们将从以下三个方面进行具体介绍：
  - BGP路由聚合的必要性——解决BGP网络存在的问题；
  - BGP路由聚合的配置方法；
  - BGP路由聚合带来的问题讨论。

BGP路由聚合的必要性

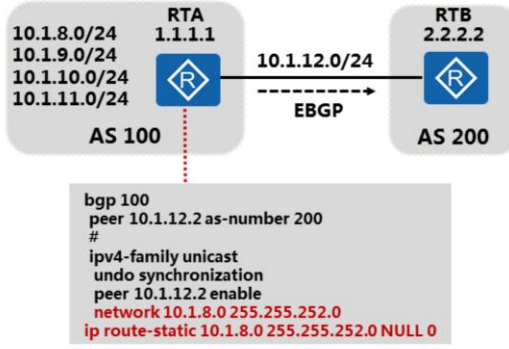


- 如图所示，AS 100内有4个用户网段，AS 200内有4个用户网段。AS 300连接了一个Client AS，该AS内的路由器比较低端，处理能力较低，因此既希望能访问AS 100与AS 200内的网段，又不希望接收过多的明细路由，如何解决该问题？

HUAWEI

- 解决方案：
  - 在RTC上将AS 100和AS 200内的明细路由聚合成10.1.8.0/21的一条路由，并将此聚合路由发布给Client AS。
- 现在Internet上的路由条目数量众多，处理这些路由时存在以下问题：
  - 存储路由条目的路由表将占用大量的内存资源，传输路由信息需要占用大量的带宽资源；
  - 明细路由频繁震荡造成网络不稳定。
- 因此，通过路由聚合来节省内存和带宽资源，减少路由震荡带来的影响成为必然。

BGP路由聚合方法 - 静态



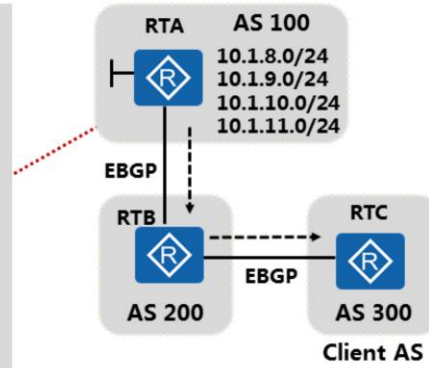
- AS 100内有4个用户网段，RTA通过路由聚合屏蔽明细路由，只将一条聚合后的路由10.1.8.0/22发布给AS 200内的RTB。

HUAWEI

- 使用静态路由配置路由聚合的思路：
  - 使用静态路由将明细路由聚合成10.1.8.0/22，下一跳指向NULL 0，因为聚合路由并不是具体的地址，发送给AS 200时只是明细路由的替代，为了防止路由环路，所以将下一跳指向Null 0；
  - 由于使用静态路由，路由表中产生了一条10.1.8.0/22的路由，下一跳为Null 0。使用network命令将IP路由表中的10.1.8.0/22路由变为BGP路由，并通告给对端BGP邻居，达到聚合的目的。

BGP路由聚合方法 - 自动聚合

```
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
```



  HUAWEI

- 如图所示，AS 100内有4个用户网段，通过import的方式变为BGP路由，AS 200连接了一个Client AS，该AS内的路由器处理能力较低，因此既希望能访问AS 100与AS 200内的网段，又不希望接收过多路由，如何解决该问题？
- 配置如图所示，在RTB与RTC路由器上使用命令display bgp routing-table查看，输出如下：

<RTB>display bgp routing-table

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

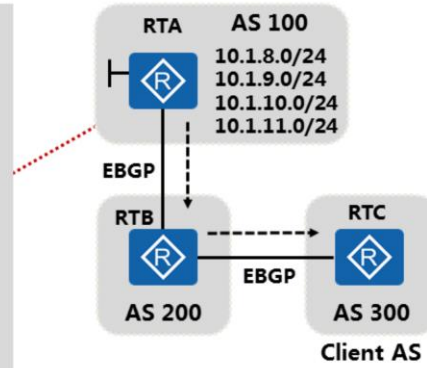
<RTC>display bgp routing-table

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

- 自动聚合只对引入BGP的路由进行聚合，聚合到自然网段后发送给邻居。

BGP路由聚合方法 - 手动聚合

```
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
```



  HUAWEI

- 如图所示，AS 100内有4个用户网段，既有通过import的方式引入BGP的路由，又有通过network方式引入BGP的路由。AS 200连接了一个Client AS，该AS内的路由器处理能力较低，因此既希望能访问AS 100与AS 200内的网段，又不希望接收过多路由，如何解决该问题？
- 配置如图所示，在RTB与RTC路由器上使用命令display bgp routing-table查看，输出如下：

<RTB>display bgp routing-table

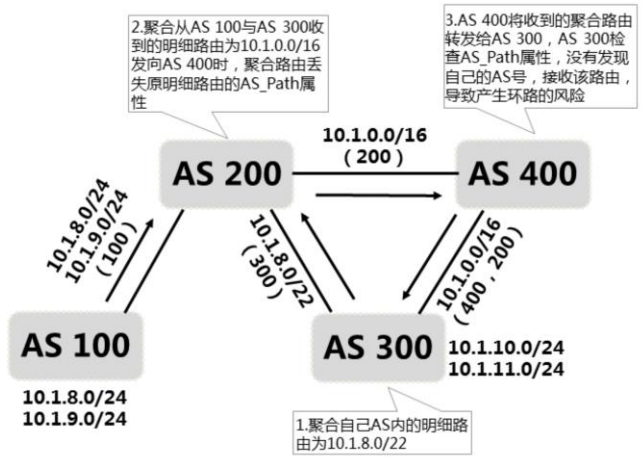
	Network	NextHop	MED	LocPrf	PrefVal	Path/Ogn
*>	10.1.8.0/22	10.1.12.1			0	100?

<RTC>display bgp routing-table

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

- 手动聚合对BGP本地路由表里存在的路由进行聚合，并且能指定聚合路由的掩码。

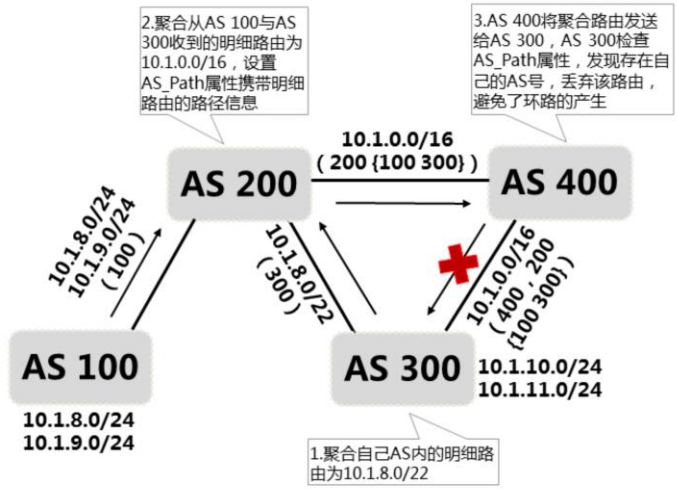
BGP路由聚合带来的问题 - 潜在环路



- 如何解决BGP路由聚合带来的潜在环路问题？

HUAWEI

BGP路由聚合带来的问题 - 解决方法



HUAWEI

- 为了解决BGP路由聚合带来的问题，设置了两个AS_Path属性：
  - Atomic-Aggregate：公认任意属性，用于警告下游路由器出现了信息丢失，如图所示，AS 200内设置了路由聚合的路由器在聚合后发生了路径丢失的现象，此时该路由器通过Update报文携带该属性通知自己的邻居发生了路径丢失。
  - Aggregator：可选过度属性，该属性包含发起聚合的路由器的AS号和Router-ID，表明发生聚合的位置。
- AS_Path属性有两种类型：
  - AS_Sequence：表示AS_Path内的AS号是一个有序的列表。
  - AS_Set：表示AS_Path内的AS号是一个无序的列表。
- AS_Path本身是一个有序的列表，因为AS_Path每经过一个AS都会将AS号添加到AS_Path中，并且按经过的顺序从左到右排列。
  - 如图所示，AS 400向AS 300通告聚合路由时，AS_Path属性（大括号的除外）表示该聚合路由依次经过了AS 200和AS 400。
- 当发生聚合后，如果需要聚合路由携带所有明细路由经过的AS号来防止环路，则在配置聚合的命令后添加as-set参数。
  - 如图所示，AS 200内发生了聚合并配置了as-set参数，则聚合路由会将明细路由的AS_Path信息用一个AS-Set集表示（放在中括号里的AS号信息，该集合的AS号没有先后顺序），携带在聚合路由后用以防止环路。
- 路由聚合解决了两类问题，一是减轻了设备传输和计算路由所需资源的负担，二是隐藏了具体的路由信息，减少了路由震荡的影响。但是路由聚合后，AS_Path属性丢失，存在产生环路的风险。
- 如果路由聚合后携带所有明细路由经过的AS信息，当明细路由发生频繁震荡时，聚合路由也可能受其影响频繁刷新。
- 因此，聚合路由是否携带丢失的AS_Path信息，需要设计者综合考虑网络环境。

思考题

1. BGP公认必遵属性有哪些？（　　）
   A. Origin
   B. AS_Path
   C. Next_hop
   D. Local_preference
2. BGP使用的端口号为多少？（　　）
   A. TCP 21
   B. TCP 179
   C. TCP 80

- 答案：ABC。
- 答案：B。


谢 谢
www.huawei.com

IP组播基础

HUAWEI

前言

- 当网络中部署点到多点通信应用时，若采用单播方式，网络中传输的信息量与需要该信息的用户量成正比。多份内容相同的信息发送给不同用户，对信源及网络带宽都将造成巨大压力。若采用广播方式，无需接收信息的主机也将收到该信息，这样不仅信息安全得不到保障，且会造成同一网段中信息泛滥。
- IP组播技术有效地解决了单播和广播在点到多点应用中的问题。组播源只发送一份数据，数据在网络节点间被复制、分发，且只发送给需要该信息的接收者。

目标

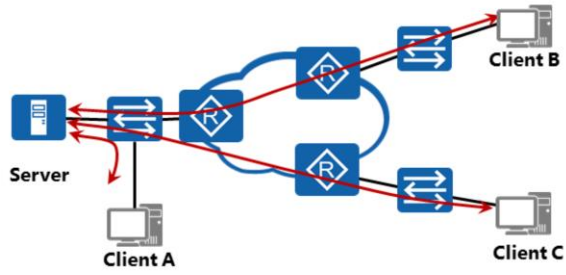
- 学完本课程后，您将能够：
  - 熟悉点到多点应用的特点
  - 掌握组播基本架构
  - 掌握组播地址结构的组成

目录

1. **点到多点应用的发展与部署**
2. 组播基本概述

传统点到点应用

- 服务提供端以单个用户为单位提供服务。
- 不同用户与服务提供端的通信数据存在差异。



HUAWEI

- 传统的电子邮件、WEB、网上银行等点到点应用主要是为独立的个人或组织提供特定服务，因为特定服务在数据差异性、安全性等方面的限制，所以不同Client与Server通信的数据只能以点到点的形式传播，即通信是在一台源主机和一台目的主机之间进行。同时只有一个数据发送者和接收者。
- 两个通信实体之间的通信过程如下：
  - Server封装数据包并发出，其中源IP为自身IP，目的IP为远端Client地址，源MAC为自身MAC地址，目的MAC为网关路由器的MAC地址。
  - 网关路由器收到数据包，解封装后根据目的IP查找路由表，确定去往目的IP的下一跳地址及出接口。重新封装源数据包，从相应出接口发给下一跳设备继续转发。
  - 经过路由器的多次逐条转发，数据包到达Client所在网络，Client收到数据后，对数据包进行解封装并交由本机上层应用协议处理。

新型点到多点应用

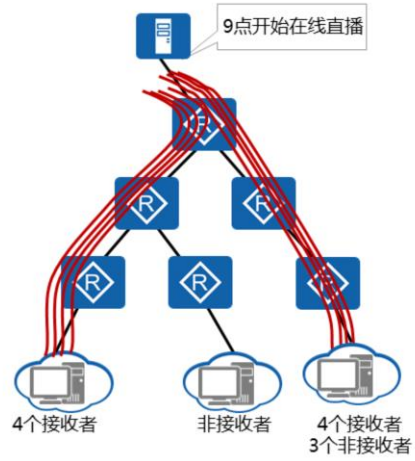
- 服务提供端以一组用户为单位提供服务。
- 同组用户与服务提供端的通信数据无差异。

HUAWEI

- 随着Internet网络的不断发展，网络中交互的各种数据、语音和视频信息数量突增。
- 新兴的在线直播、网络电视、视频会议等应用也在逐渐兴起。
- 这些新兴业务大多符合点对多点的模式，对信息安全性、传播范围、网络带宽提出了较高的要求。

单播方式部署点到多点应用

- 单播方式所存在的问题：
  - 重复流量过多。
  - 消耗设备和链路带宽资源。
  - 难以保证传输质量。

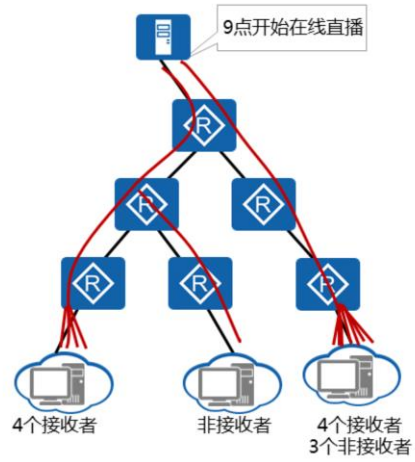


HUAWEI

- 单播（Unicast）是在一台源IP主机和一台目的IP主机之间进行。网络上绝大部分的数据都是以单播的形式传输的，例如电子邮件收发、网上银行都是采用单播实现的。
- 单播的特点：
  - 一份单播报文，使用一个单播地址作为目的地址。Source向每个接收者发送一份独立的单播报文。如果网络中存在N个接收者，则Source需要发送N份单播报文。
  - 网络为每份单播报文执行独立的数据转发，形成一条独立的数据传送通路。N份单播报文形成N条相互独立的传输路径。
- 单播的缺陷：
  - 单播方式下，网络中传输的信息量和需求该信息的用户量成正比，当需求该信息的用户量较大时，网络中将出现多份相同信息流，不仅占用处理器资源而且浪费带宽。
  - 单播方式较适合用户稀少的网络，当用户量较大时很难保证网络传输质量。

广播方式部署点到多点应用

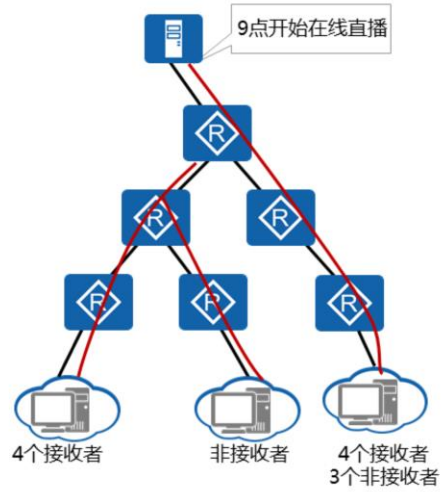
- 广播方式所存在的问题：
  - 地域范围限制。
  - 安全性无法保障。
  - 有偿性无法保障。



- 广播（Broadcast）是在一台源IP主机和网络中所有其它的IP主机之间进行，属于一对所有的通讯方式，所有主机都可以接收到（不管是否需要）。
- 广播的特点：
  - 一份广播报文，使用一个广播地址作为目的地址。Source向本网段对应的广播地址发送且仅发送一份报文。
  - 不管是否有需求，保证报文被网段中的所有用户主机接收。
- 广播的缺陷：
  - 广播方式下，信息发送者与用户主机被限制在一个共享网段中，且该网段所有用户主机都能接收到该信息。
  - 广播方式只适合共享网段，且信息安全性和有偿服务得不到保障。
- 对于点到多点的网络应用，单播和广播都有一定的局限性。

组播方式部署点到多点应用

- 组播方式的优势：
  - 无重复流量。
  - 节省设备与带宽资源。
  - 安全性高。
  - 有偿性有保障。



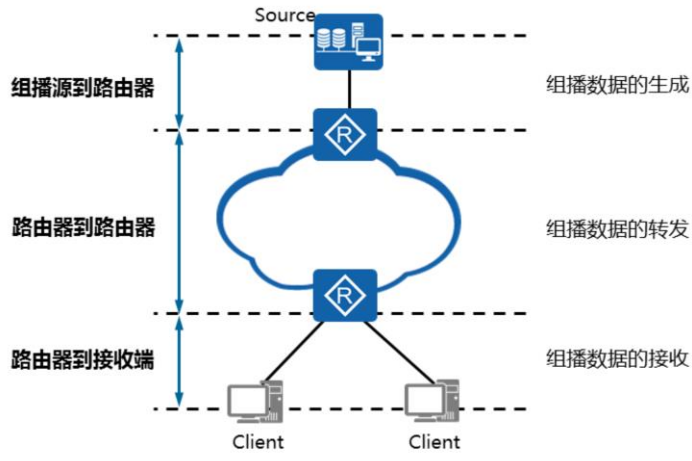
HUAWEI

- 组播（Multicast）是在一台源IP主机和多台（一组）IP主机之间进行，中间的交换机和路由器根据接收者的需要，有选择性地对数据进行复制和转发。
- 组播的优势：
  - 组播方式下，单一的信息流沿组播分发树被同时发送给一组用户，相同的组播数据流在每一条链路上最多仅有一份。
  - 相比单播，由于被传递的信息在距信息源尽可能远的网络节点才开始被复制和分发，所以用户的增加不会导致信息源负载的加重以及网络资源消耗的显著增加。
  - 相比广播，由于被传递的信息只会发送给需要该信息的接收者，所以不会造成网络资源的浪费，并能提高信息传输的安全性。另外，广播只能在同一网段中进行，而组播可以实现跨网段的传输。
- 组播的应用：
  - 组播技术有效地满足了单点发送、多点接收的需求，实现了IP网络中点到多点的高效数据传送，能够大量节约网络带宽、降低网络负载。利用组播技术可以更方便地提供在线直播、网络电视、远程教育等服务。

目录

1. 点到多点应用的发展与部署
2. **组播基本概述**

组播基本架构



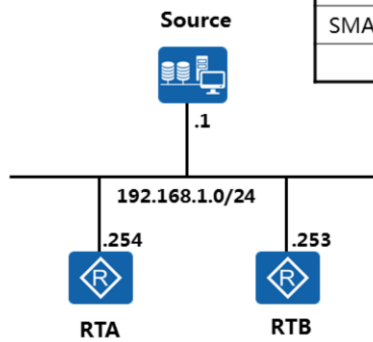
HUAWEI

- 组播源到路由器：组播源生成组播数据，完成数据封装并发送给网关路由器。
- 路由器到路由器：路由器根据接收者的分布情况有选择地对数据进行复制和转发。
- 路由器到接收端：路由器收到组播数据并发送给相应的接收者。

组播源到路由器

- 组播源如何封装组播数据？
  - 目的IP地址如何确定？
  - 目的MAC地址如何确定？

组播数据
DIP ： ？？
SIP：192.168.1.1
SMAC：68-F7-28-42-6D-0D
DMAC：MAC ？？



HUAWEI

- 单播数据包传输的路径是利用“逐跳”（hop-by-hop）转发原理在IP网络中传输。
- 相较于IP单播，IP组播通信的特点是数据包的目的地址不是一个特定的单一IP地址，而是一个特定组地址。
- 为了实现信息源和组播组成员跨越互联网进行通讯，需要提供网络层组播，组播数据包的目的IP地址使用组播IP地址。也就是说组播源不关注接收者的位置信息，只要将数据发送到特定组IP地址即可。
- 以太网传输单播数据帧时，目的MAC地址使用的是接收者或者去往接收者的下一跳网关设备的MAC地址。
- 但是在传输组播报文时，目的端不再是一个具体的接收者，而是一个成员不确定的组，如果目的MAC封装成接收者的MAC地址，则需要为每个接收者分别发送一份组播帧。
- 显然，这是不合理的。为了在数据链路层实现组播信息的高效传输，需要提供链路层组播转发能力，链路层组播使用组播MAC地址。

组播IP地址

- 一个组播IP地址并不是表示具体的某台主机，而是一组主机的集合，主机声明加入某组播组即标识自己需要接收目的地址为该组播地址的数据。

范围	含义
224.0.0.0—224.0.0.255	为路由协议预留的永久组地址
224.0.1.0—231.255.255.255 233.0.0.0—238.255.255.255	Any-Source临时组播组地址
232.0.0.0—232.255.255.255	Source-Specific临时组播组地址
239.0.0.0—239.255.255.255	本地管理的Any-Source临时组播组地址

- IP组播常见模型分为ASM模型和SSM模型。

  HUAWEI

- IPv4组播地址：
  - IPv4地址空间分为五类，即A类、B类、C类、D类和E类。D类地址为IPv4组播地址，范围是从224.0.0.0到239.255.255.255，用于标识组播组，且仅能作为组播报文的目的地址使用，不能作为源地址使用。
  - IPv4组播报文的源地址字段为IPv4单播地址，可使用A、B或C类地址，不能是D类、E类地址。
  - 在网络层上，加入同一组播组的所有用户主机能够识别同一个IPv4组播组地址。一旦网络中某用户加入该组播组，则此用户就能接收以该组地址为目的地址的IP组播报文。
- 组播服务模型：
  - ASM全称为Any-Source Multicast，译为任意源组播。在ASM模型中，任意发送者都可以成为组播源，向某组播组地址发送信息。接收者加入该组播组后，能够接收到发往该组播组的所有信息。在ASM模型中，接收者无法预先知道组播源的位置，接收者可以在任意时间加入或离开该组播组。
  - SSM全称为Source-Specific Multicast，译为指定源组播。在现实生活中，用户可能仅对某些源发送的组播信息感兴趣，而不愿接收其它源发送的信息。SSM模型为用户提供了一种能够在客户端指定信源的传输服务。SSM模型和ASM模型的根本区别是接收者已经通过其他手段预先知道了组播源的具体位置。SSM和ASM使用不同的组播地址范围，直接在接收者和组播源之间建立组播转发树。

组播MAC地址

- 组播MAC地址与单播MAC地址的区别：

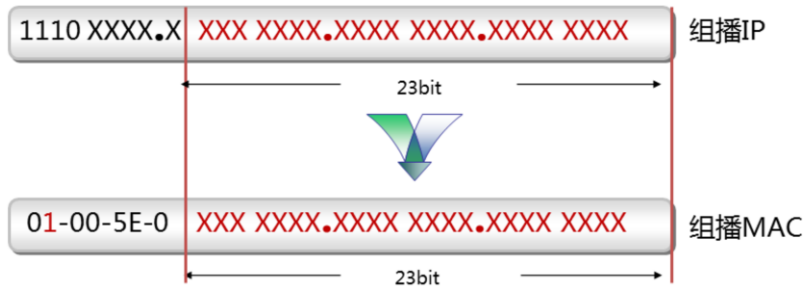
- IANA规定，IPv4组播MAC地址的高24位为0x01005e，第25位固定为0。

  HUAWEI

- 组播数据帧的传输目的不再是一个具体的接收者，而是一个成员不确定的组，所以使用的是组播MAC地址。IANA规定，组播MAC地址的高24bit为0x01005e，第25bit固定为0。
- 组播MAC地址用于在链路层标识属于同一组播组的接收者。
- 以太网传输单播数据帧的时候，目的MAC地址使用的是接收者的MAC地址或者下一跳路由器的MAC地址。这个MAC地址通过ARP获取。对于组播数据帧也需要有一个可预知的MAC地址。

组播IP与MAC的映射

- 需要组播IP地址与组播MAC地址的自动映射。
- MAC地址的低23bit为组播IP地址的低23bit。

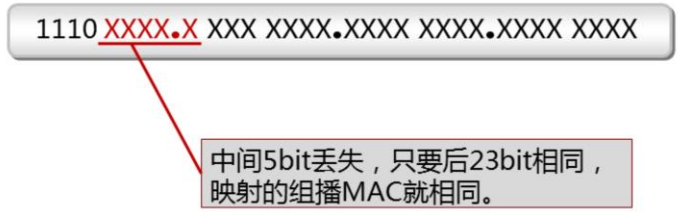


HUAWEI

- 为了使组播源和组播组成员进行通信，需要提供网络层组播，使用IP组播地址。同时，为了在本地物理网络上实现组播信息的正确传输，需要提供链路层组播，使用组播MAC地址。组播数据传输时，其目的地不是一个具体的接收者，而是一个成员不确定的组，所以需要一种技术将IP组播地址映射为组播MAC地址。

映射导致的问题

- 组播IP地址映射成组播MAC地址时，会导致32个组播IP地址对应一个组播MAC的问题。



HUAWEI

- 由于IP组播地址的前4bit是1110，代表组播标识，而后28bit中只有23bit被映射到MAC地址，这样IP地址中就有5bit信息丢失，直接的结果是出现了32个组播IP地址映射到同一组播MAC地址上。
- IETF认为同一个局域网中两个或多个组地址生成相同的MAC地址的几率非常低，不会造成太大的影响。

思考题

1. 什么是IP组播通信？
2. IPv4组播地址的范围是什么？

HUAWEI

- 答案：IP组播通信指的是IP报文从一个源发出，被转发到一组特定的接收者。相较于传统的单播和广播，IP组播可以有效地节约网络带宽、降低网络负载，所以被广泛应用于IPTV、实时数据传送和多媒体会议等网络业务中。
- 答案：IANA（Internet Assigned Numbers Authority，互联网编号分配委员会）将D类地址空间分配给IPv4组播使用。IPv4地址一共32位，D类地址最高4位为1110，因此地址范围从224.0.0.0到239.255.255.255。


谢 谢
www.huawei.com

IGMP协议原理与配置

HUAWEI

前言

- 组播通信中，发送者将组播数据发送到特定的组播地址。要使组播报文最终能够到达接收者，需要某种机制使与连接潜在接收者网段的组播路由器能够了解到该网段内有哪些组播接收者，保证接收者可以加入到相应的组播组中接收数据。
- IGMP（Internet Group Management Protocol）因特网组管理协议，是TCP/IP协议族中负责IP组播成员管理的协议，它用来在接收者和与其直接相邻的组播路由器之间建立、维护组播组成员关系。

目标

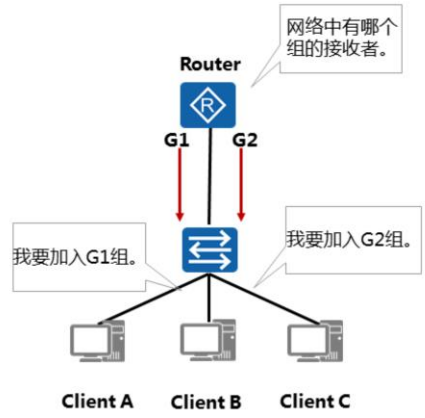
- 学完本课程后，您将能够：
  - 掌握IGMP的基本原理和配置
  - 熟悉IGMP不同版本间的区别
  - 掌握IGMP Snooping的基本原理

目录

1. **组播接收端的需求**
2. IGMPv1的工作机制
3. IGMPv2的工作机制
4. IGMPv3的工作机制
5. IGMP Snooping的工作机制
6. IGMP的配置实现

接收端如何接收组播数据

- 接收者与路由器间需要交换哪些信息？
  - 接收者需声明自己要接收哪个组的数据。
  - 路由器需了解哪些组播组存在接收者。
- 人工配置这些信息，有哪些问题？
  - 实时性差。
  - 灵活性差。
  - 工作量大、易出错。

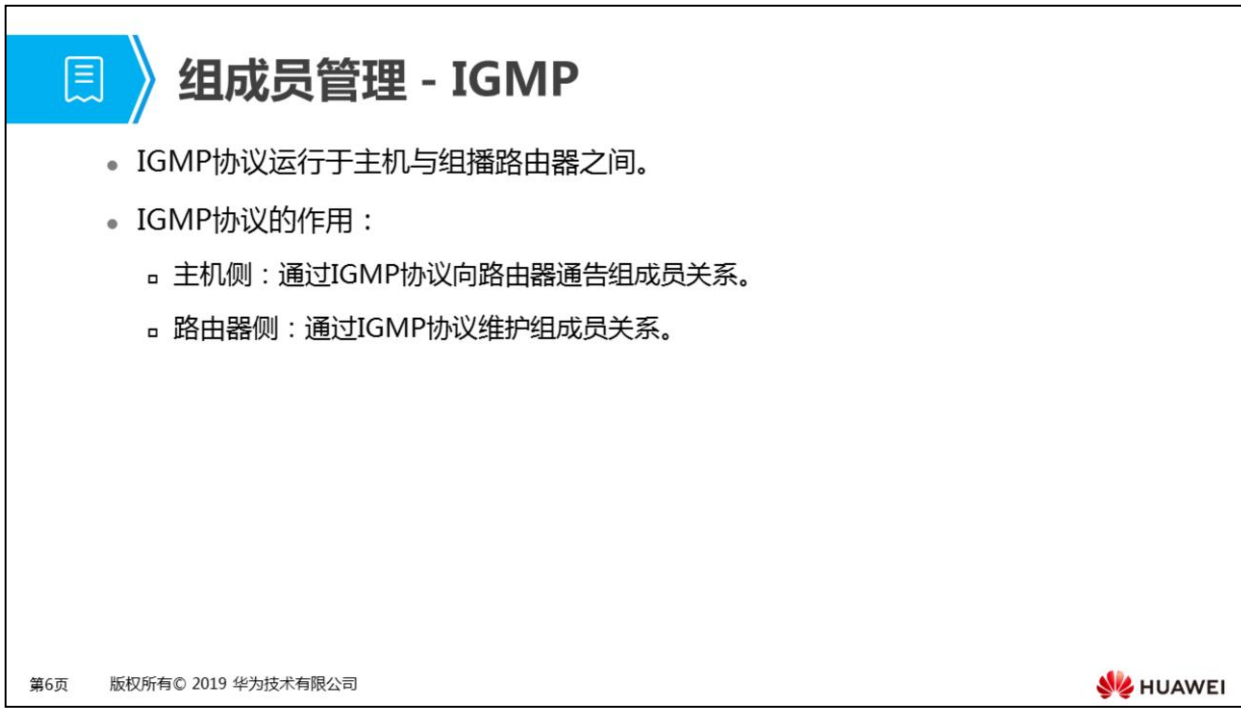


HUAWEI

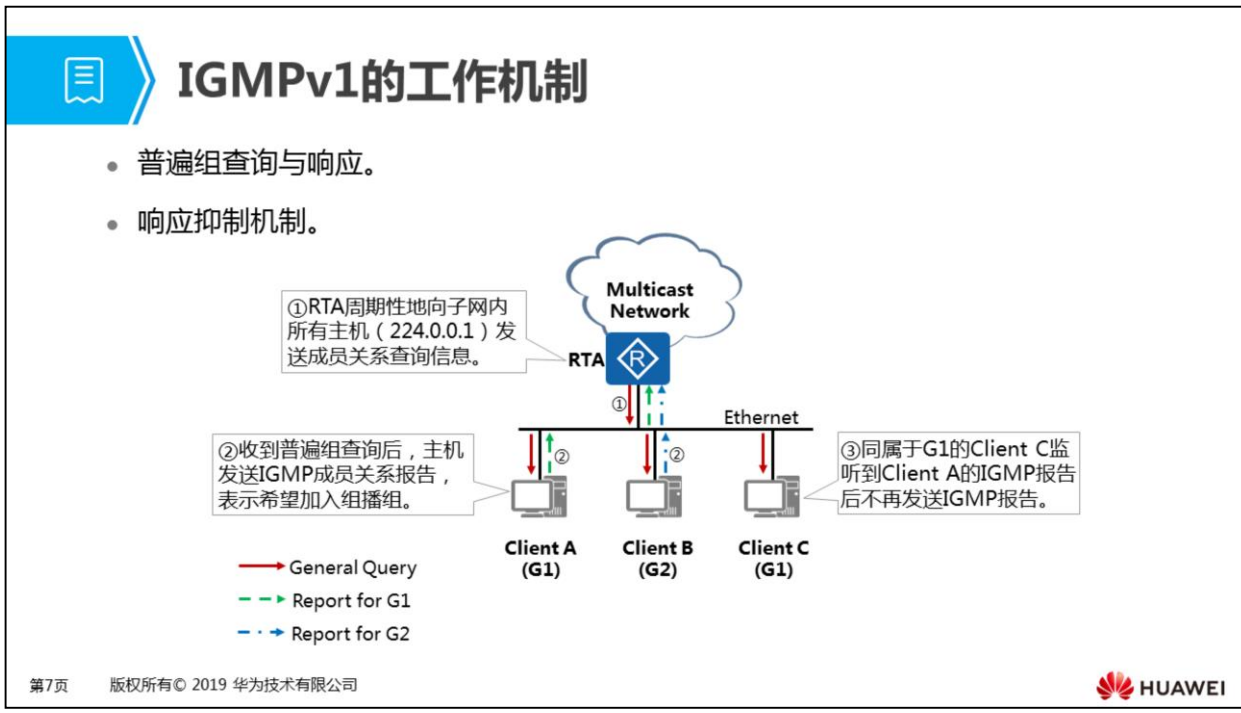
- 组播源不关注接收者的位置信息，但是对于连接组成员的路由器而言，其需要收集和维护组成员的信息。
- 组播既不指定明确的接收者，也不是将数据发送给网络上的所有主机。如果主机想接收发往某一组播地址的数据，它需要加入这个组，成为该组播组的成员。
- 对于需要实现高效转发、灵活加组的网络，该如何部署？

目录

1. 组播接收端的需求
2. **IGMPv1的工作机制**
3. IGMPv2的工作机制
4. IGMPv3的工作机制
5. IGMP Snooping的工作机制
6. IGMP的配置实现



- IGMP（Internet Group Management Protocol）作为因特网组管理协议，是TCP/IP协议族中负责IP组播成员管理的协议，它用来在IP主机和与其直接相邻的组播路由器之间建立、维护组播组成员关系。

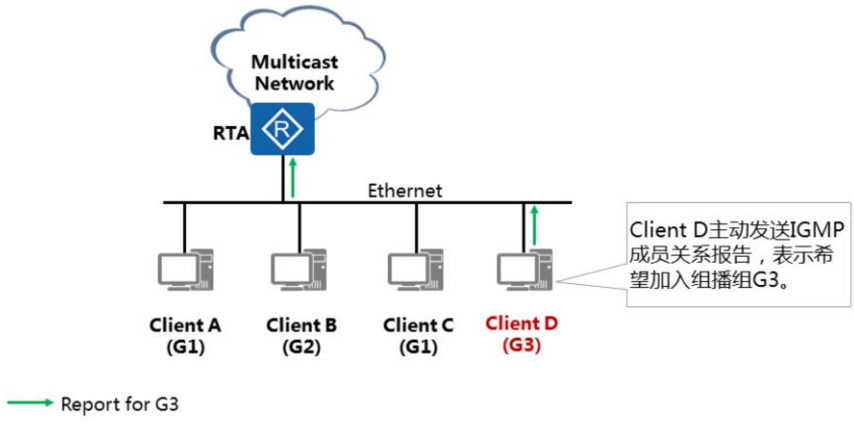


- IGMPv1支持两种类型的报文：
  - 普遍组查询报文（General Query）：路由器周期性地向224.0.0.1地址（表示同一网段内所有主机和路由器）发送通用查询报文，默认查询周期为60秒，发送周期可以通过命令配置。
  - 成员关系报告报文（Membership Report）：用于主机加入某个组播组。

- 如图所示，假设Client A和Client C想要接收组播组G1的数据，Client B想要接收组播组G2的数据。普遍组查询和响应过程如下：
  - RTA发送普遍组查询报文。
  - 网段内所有主机都接收到该查询报文，Client A和Client C是组播组G1成员，则在本地启动定时器Timer-G1。Client B是组播组G2的成员，则在本地启动定时器Timer-G2。定时器的范围为0～10秒之间的随机值。定时器先超时的主机发送针对该组的成员报告报文。Client A上的Timer-G1首先超时，向该网段发送目的地址为G1的成员报告报文。Client B上的Timer-G2超时，向该网段发送成员报告报文，目的地址为G2。
  - Client C侦听到Client A的成员报告报文，则停止定时器Timer-G1，不再发送针对G1的成员报告报文。这就是响应抑制机制，可以减少网段上的协议流量。
- RTA接收到成员报告报文后，了解到本网段内存在组播组G1和G2的成员，一旦RTA收到G1和G2的组播数据，将向该网段转发。

IGMPv1成员加入

- 主机申请加组。

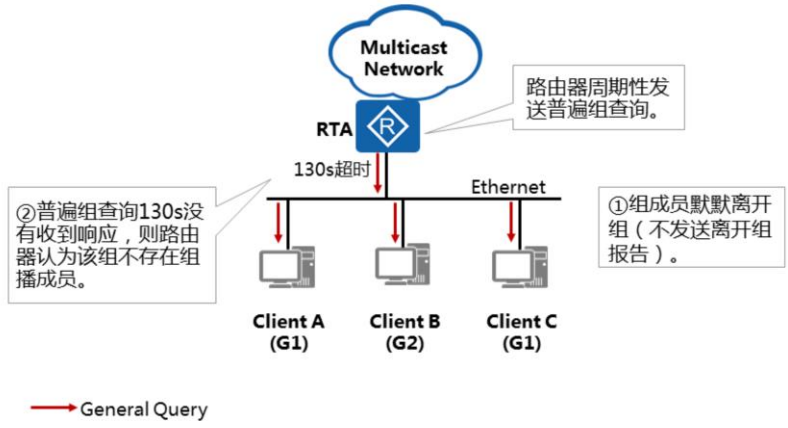


HUAWEI

- 新接入主机Client D想加入组播组G3，为了快速接收组播数据，不等待普遍组查询报文，而立即发送G3的成员报告报文。RTA收到成员报告报文后，了解到本网段内出现了组播组G3的成员。一旦有G3的组播数据到达RTA，将向该网段转发。

IGMPv1问题一：组成员离开

- 静默离开。

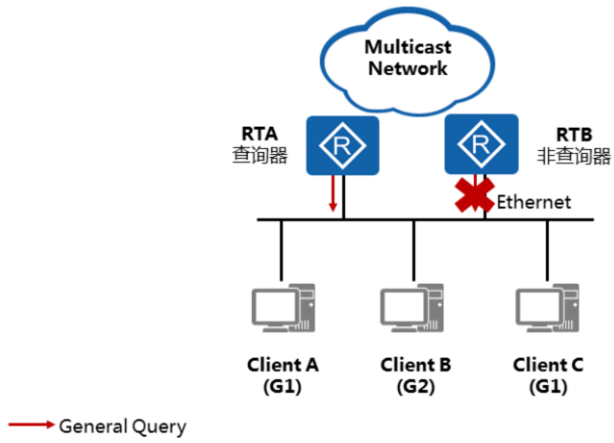


HUAWEI

- IGMPv1没有专门定义离开组消息。
- 当Client离开组播组时，将不会再对普遍组查询报文做出回应。假设所有Client退出组播组，Client将不再对普遍组查询报文进行响应。由于网段上不存在组播组的其他成员，RTA不会收到任何成员报告报文，则在一定时间（130秒=120*2+10，即组成员关系超时时间=IGMP普遍查询消息发送间隔 × 健壮系数 + 最大查询响应时间）后，删除对应的组播转发项。

IGMPv1问题二：查询器选举

- 查询器选举依赖于组播路由协议。

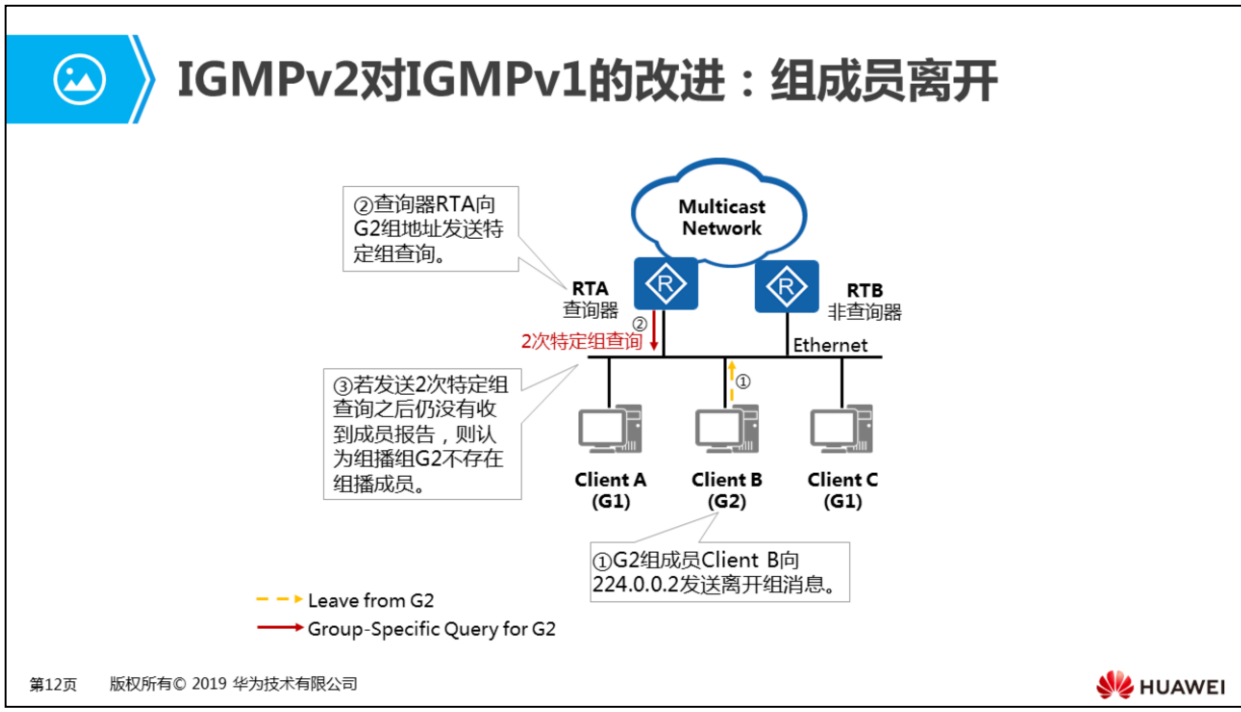


HUAWEI

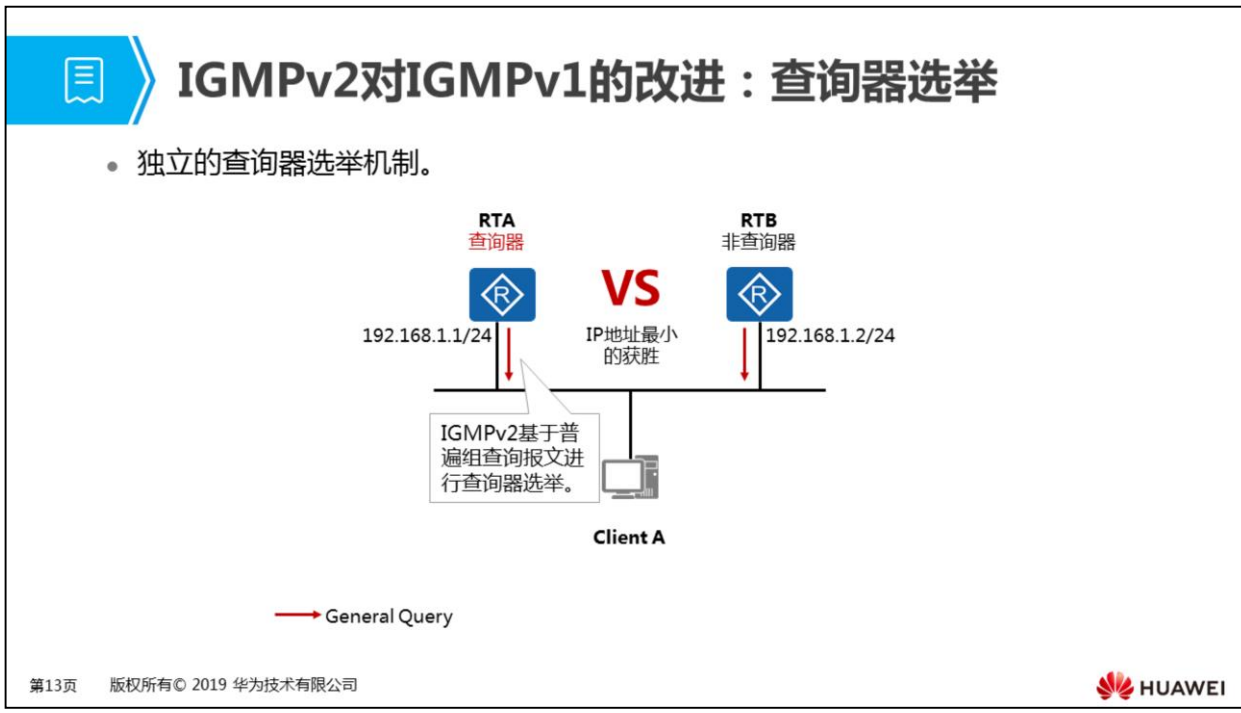
- 多台路由器同时连接到同一接收端网络时，只需要有一台路由器进行IGMP的查询。
- IGMPv1无查询路由器选举机制，其依赖于组播路由协议在末端网络中选举一个查询器。
- 由于不同的组播路由协议采用不同的选举机制，所以在IGMPv1中，同一末端网络中可能会存在多台查询器。
- 针对IGMPv1中的两个问题，IGMPv2进行了改进和优化。

目录

1. 组播接收端的需求
2. IGMPv1的工作机制
3. **IGMPv2的工作机制**
4. IGMPv3的工作机制
5. IGMP Snooping的工作机制
6. IGMP的配置实现



- 如图所示，在IGMPv2中，Client B离开组播组G2的过程如下：
  - Client B向本地网段内的所有组播路由器（目的地址为224.0.0.2）发送针对组G2的离开报文。
  - 查询器收到离开报文，会发送针对G2的特定组查询报文，同时启动组成员关系定时器Timer-Membership=发送间隔x发送次数。缺省每隔1秒发送一次，一共发送两次，发送间隔和发送次数可以配置。
  - 如果网段内不存在其他组G2的成员，则路由器不会收到组G2的成员报告报文。在Timer-Membership超时后，删除组播转发表项中对应的下游接口。路由器将不再向该网段转发G2的组播数据。如果网段内还有G2的其他成员，则这些成员在收到特定组查询报文后，会在最大响应时间内发送G2的成员报告报文。路由器继续向该网段转发G2的组播数据。



- 相对于IGMPv1，IGMPv2使用独立的查询器选举机制。
- 所有IGMPv2路由器在初始状态时都认为自己是查询器，向本地网段内的所有主机和路由器发送普遍组查询报文。其他路由器在收到该报文后，将报文的源IP地址与自己的接口地址作比较。IP地址最小的路由器将成为查询器，其他路由器成为非查询器。如图所示，RTA的接口IP地址小于RTB的接口IP地址，则RTA当选为查询器。IGMP的查询器和非查询器都会处理IGMP组加入信息，但是只有查询器负责发送查询报文。IGMP非查询器不处理IGMPv2离开报文。
- 所有非查询器上都会启动一个定时器。如果在该定时器超时前收到了来自查询器的查询报文，则重置该定时器；否则就认为原查询器失效并发起新的查询器选举。

IGMPv1和IGMPv2报文比较

- 思考：IGMP如何引导组成员接收特定组播源的数据。

HUAWEI

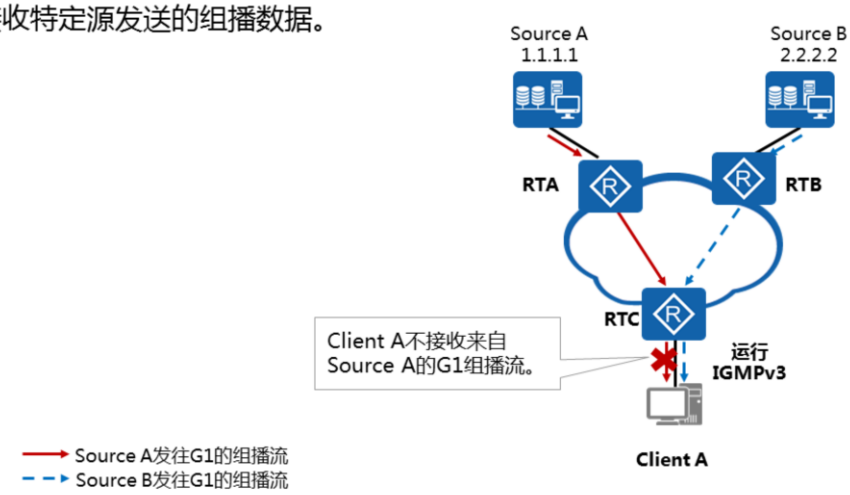
- IGMPv1报文：
  - 版本：包含IGMP版本标识，因此设置为1。
  - 类型：普遍组查询（0x11），成员关系报告（0x12）。
  - 组地址：普遍组查询报文中，组地址为0；成员关系报告报文中，组地址为成员想要加入的组播组的地址。
- IGMPv2报文：IGMPv2报文与IGMPv1报文略有不同，它取消了版本字段，增加了最大响应时间字段。
  - 类型：相比于IGMPv1，IGMPv2新增了两种报文：
    - 特定组查询报文（0x11）：查询器向共享网段内指定组播组发送的查询报文，用于查询该组播组是否存在成员。
    - 成员离开报文（0x17）：成员离开组播组时主动向路由器发送的报文，用于宣告自己离开了某个组播组。
  - 最大响应时间：表示主机响应查询返回报告的最大时间。
    - 对于普遍组查询，最大响应时间默认为10秒。
    - 对于特定组查询，最大响应时间默认为1秒。
  - 组地址：
    - 普遍组查询报文中，组地址设置为0。
    - 特定组查询报文中，组地址为需要查询的组地址。
    - 在成员报告或离开组的消息中，组地址为需要报告或离开的组地址。

目录

1. 组播接收端的需求
2. IGMPv1的工作机制
3. IGMPv2的工作机制
4. **IGMPv3的工作机制**
5. IGMP Snooping的工作机制
6. IGMP的配置实现

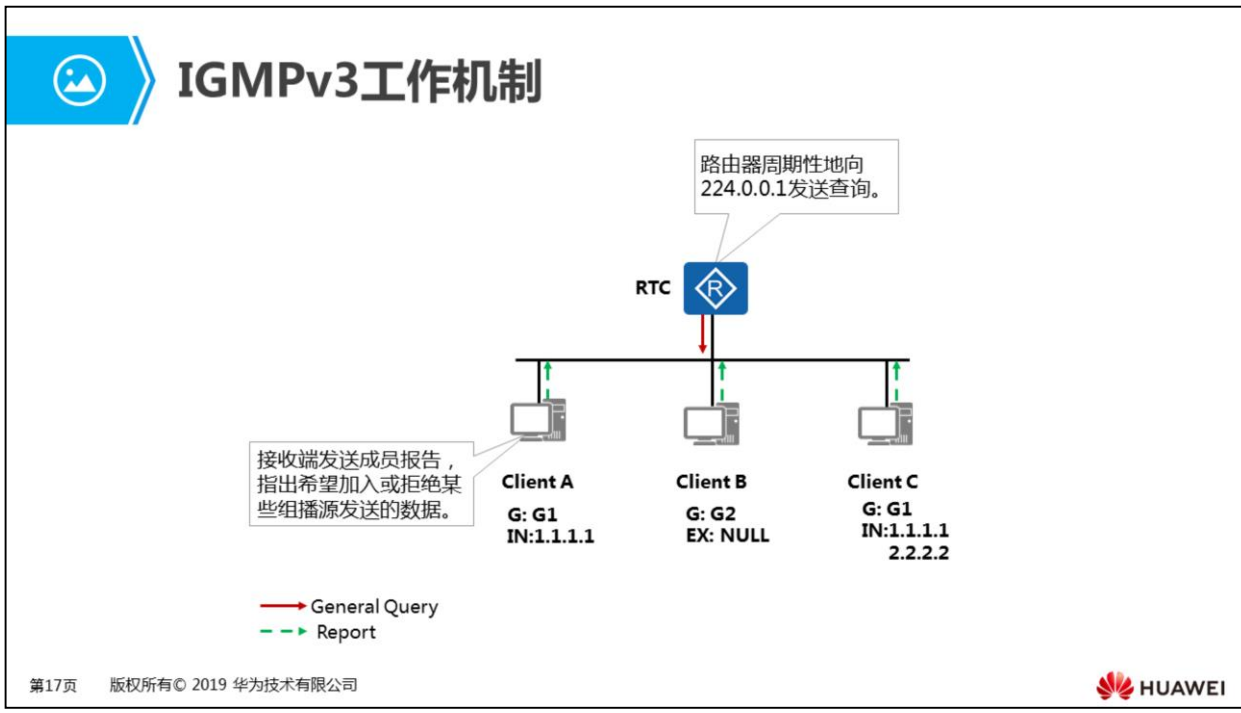
SSM模型中的新需求

- 只接收特定源发送的组播数据。



HUAWEI

- 如果Client A和RTC之间运行的是IGMPv1或IGMPv2，Client A无法对组播源进行选择，无论其是否需要，都会同时接收到来自组播源Source A和Source B的数据。
- 为了满足SSM模型的新需求，IGMPv3提供了在报文中携带指定组播源信息的能力。
- 接下来一起学习IGMPv3是如何设计的。



- 与IGMPv2相比，IGMPv3报文的变化如下：
  - IGMPv3报文包含两大类：查询报文和成员报告报文。IGMPv3没有定义专门的成员离开报文，成员离开通过特定类型的报告报文来传达。
  - 查询报文中不仅包含普遍组查询报文和特定组查询报文，还新增了特定源组查询报文（Group-and-Source-Specific Query）。该报文由查询器向共享网段内特定组播组成员发送，用于查询该组成员是否愿意接收特定源发送的数据。特定源组查询通过在报文中携带一个或多个组播源地址来达到这一目的。
  - 成员报告报文不仅包含主机想要加入的组播组，而且包含主机想要接收来自哪些组播源的数据。IGMPv3增加了针对组播源的过滤模式（INCLUDE/EXCLUDE），将组播组与源列表之间的对应关系简单的表示为（G，INCLUDE，(S1、S2...)），表示只接收来自指定组播源S1、S2……发往组G的数据；或（G，EXCLUDE，(S1、S2...)），表示接收除了组播源S1、S2……之外的组播源发给组G的数据。当组播组与组播源列表的对应关系发生了变化，IGMPv3报告报文会将该关系变化存放于组记录（Group Record）字段，发送给IGMP查询器。
  - 在IGMPv3中一个成员报告报文可以携带多个组播组信息，而之前的版本一个成员报告只能携带一个组播组。这样在IGMPv3中报文数量大大减少。

IGMP各版本间的差异

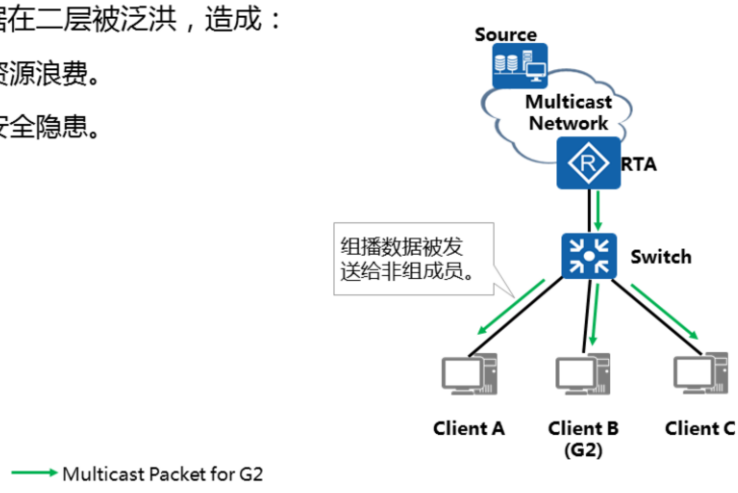
机制	IGMPv1	IGMPv2	IGMPv3
查询器选举	依靠其他协议	自己选举	自己选举
成员离开方式	静默离开	主动发送离开报文	主动发送离开报文
特定组查询	不支持	支持	支持
指定源、组	不支持	不支持	支持

目录

1. 组播接收端的需求
2. IGMPv1的工作机制
3. IGMPv2的工作机制
4. IGMPv3的工作机制
5. **IGMP Snooping的工作机制**
6. IGMP的配置实现

二层中组播数据转发的问题

- 组播数据在二层被泛洪，造成：
  - 网络资源浪费。
  - 存在安全隐患。

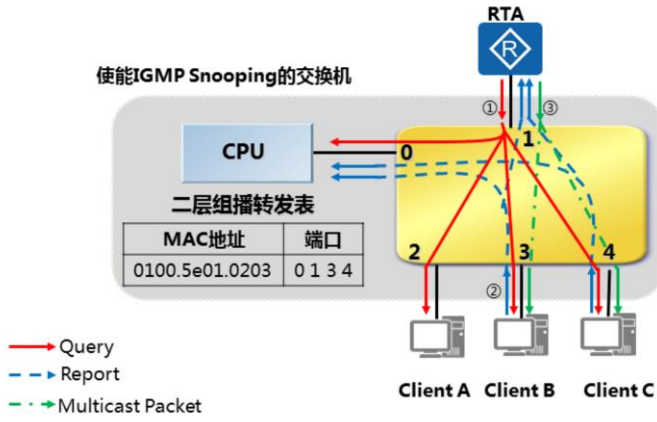


HUAWEI

- 主机加入组播组需要向上游设备发送IGMP成员报告，这样上游设备才可以将组播报文发送给主机。由于IGMP报文是封装在IP报文内，属于三层协议报文，而二层设备不处理报文的三层信息，所以主机加组的过程二层设备并不知道，而且通过对数据链路层数据帧的源MAC地址的学习也学不到组播MAC地址（数据帧的源MAC地址不会是组播MAC地址）。
- 这样当二层设备在接收到一个目的MAC地址为组播MAC地址的数据帧时，在MAC地址表中就不会找到对应的表项。那么这时候，它就会采用广播方式发送组播报文，这样一来不但对网络资源造成的极大浪费而且影响网络安全。
- IGMP Snooping机制的提出，解决了二层组播泛洪问题，下面一起来学习该机制。

IGMP Snooping工作原理

- 使能IGMP Snooping机制后，查询响应仅向路由器接口转发。



HUAWEI

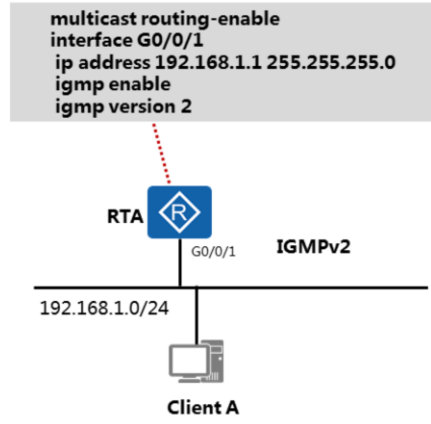
- IGMP Snooping可以实现组播数据帧在数据链路层的转发和控制。
- 使能IGMP Snooping功能后，二层设备会侦听主机和路由器之间交互的IGMP报文。通过分析报文中携带的信息（报文类型、组播组地址、接收报文的接口等），建立和维护二层组播转发表，从而指导组播数据帧在数据链路层按需转发。
- IGMP Snooping建立和维护二层组播转发表的过程：
  - RTA作为查询器，周期性的发送普遍组查询，该报文被扩散到交换机的所有端口，包括与交换机CPU相连的内部接口0。交换机CPU收到查询报文后，判断1号接口为连接路由器的接口。
  - Client B希望加入组播组224.1.2.3，因此以组播方式发送一个IGMP成员报告报文，报告中具有目的MAC地址0x0100.5e01.0203。该报文将被发往路由器的接口以及交换机CPU相连的内部接口0；当CPU收到Client B的IGMP报告时，CPU利用IGMP报告中的信息将该接口加入二层组播转发表中，此时表项包括Client B的接口号，连接路由器的接口号和连接交换机内部CPU的接口号。
  - 形成此转发表项的结果是使后面任何目的地址为0x0100.5e01.0203的组播帧都被控制在端口0、1和3，而且不向交换机其他端口扩散。
- Client C加入组224.1.2.3并主动发一个IGMP报告，交换机CPU收到此报告，它在转发表项上为MAC地址0x0100.5e01.0203增加一个端口（端口4）。

目录

1. 组播接收端的需求
2. IGMPv1的工作机制
3. IGMPv2的工作机制
4. IGMPv3的工作机制
5. IGMP Snooping的工作机制
6. **IGMP的配置实现**

HUAWEI

IGMP配置实现



HUAWEI

IGMP配置验证

```
<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
```

思考题

1. IGMPv1中，当最后一个组播成员离开后该组后，组播路由器将在多长时间后删除所对应的组播转发表项？
2. IGMPv2中，特定组查询的目的IP是224.0.0.1吗？
3. IGMP Snooping的实现原理是什么？

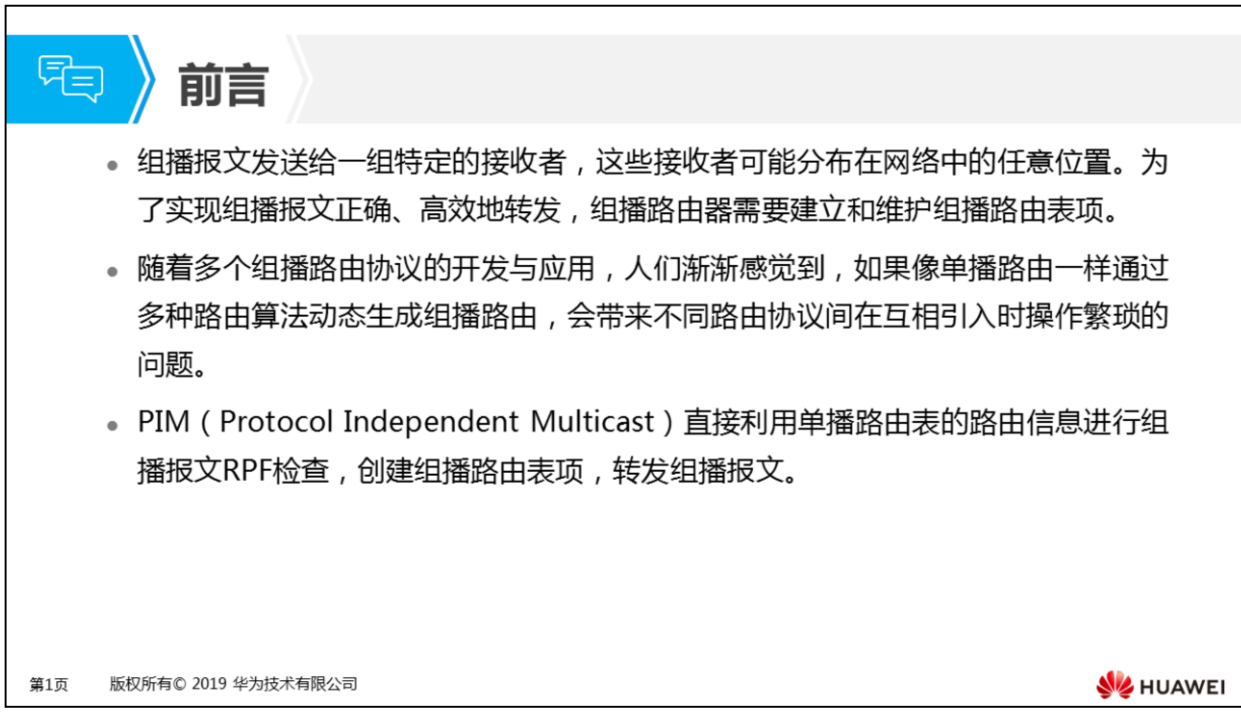
HUAWEI

- 答案：60*2 + 10 = 130s。
- 答案：否，特定组查询的目的IP是所查询组播组的组播IP地址。
- 答案：IGMP Snooping通过侦听组播路由器与主机之间交互的IGMP报文建立组播数据报文的二层转发表项，从而管理和控制组播数据报文在二层网络中的转发。


谢 谢
www.huawei.com

PIM协议原理与配置

HUAWEI



- RPF（Reverse Path Forwarding，逆向路径转发）。

目标

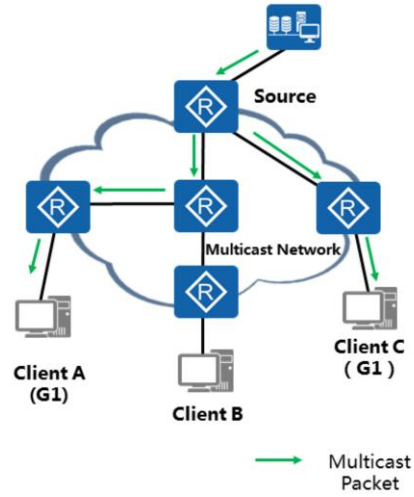
- 学完本课程后，您将能够：
  - 了解组播转发需求
  - 掌握PIM-DM的基本原理和配置
  - 掌握PIM-SM的基本原理和配置

目录

1. **组播报文转发需求**
2. PIM-DM的工作机制
3. PIM-DM的局限性
4. PIM-SM的工作机制

路由器如何转发组播报文

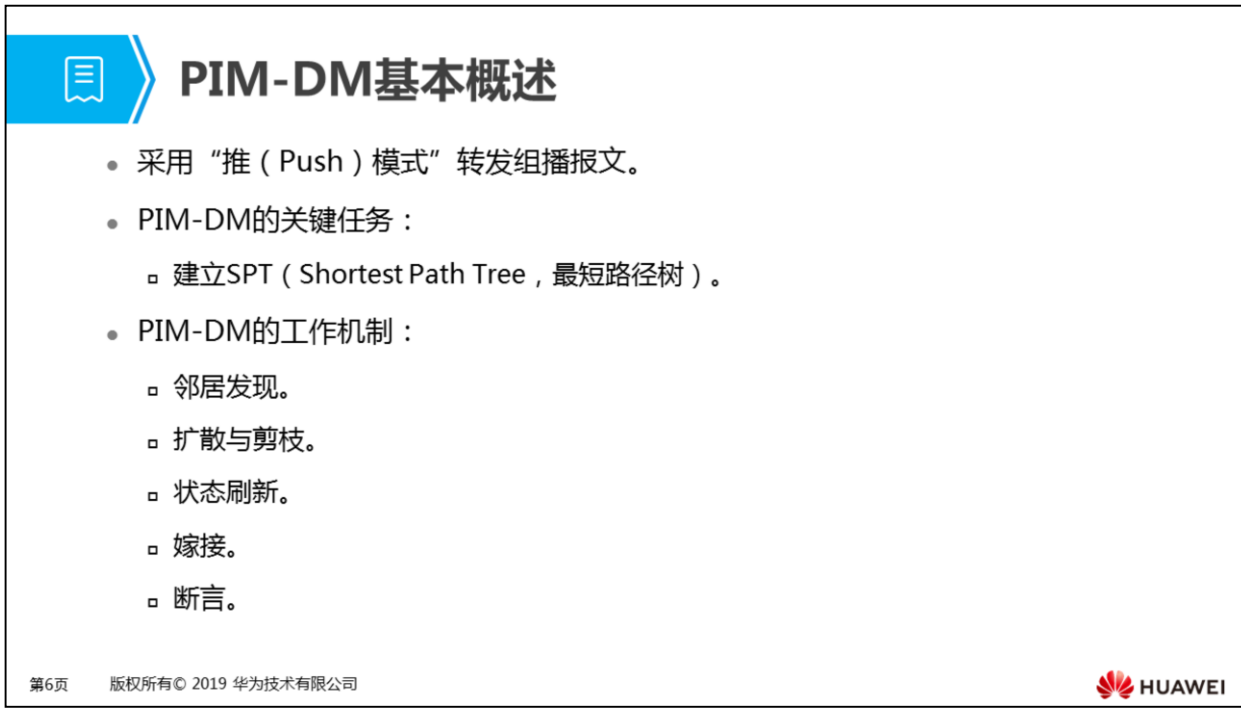
- 路由器需要依据哪些信息进行转发：
  - 各接口所在网段有无潜在接收者。
  - 接收者需要接收哪些组的数据。
- 人工配置上述信息存在一些问题：
  - 实时性差。
  - 灵活性差。
  - 工作量大、易出错。



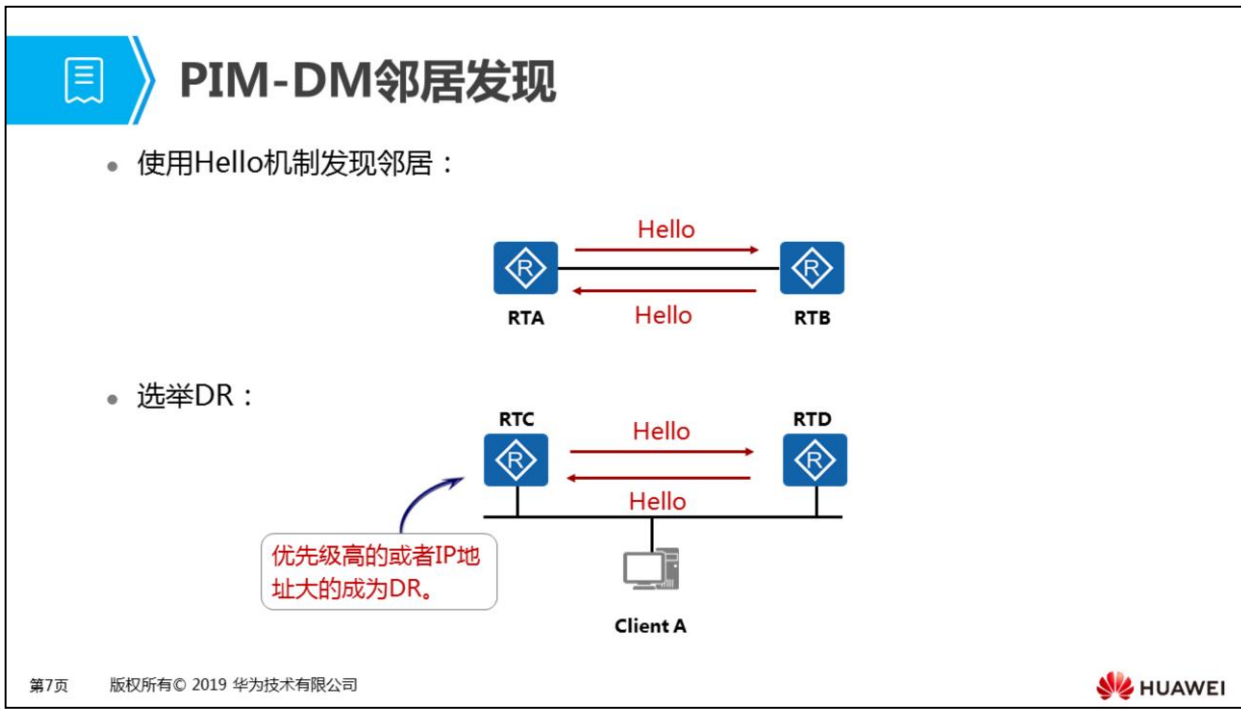
HUAWEI

- 在单播报文的转发机制中，路由器依据单播报文的目的IP地址，查找单播路由表进行转发。其中，单播路由表可以通过静态配置或者动态路由协议来学习路由。
- 在组播中，接收者可能存在于全网中的任意位置，所以如果静态配置组播路由的话，存在实时性差、灵活性差以及工作量大容易出错的问题。
- 为了正确、高效的转发组播数据报文，路由器之间则需要运行组播路由协议。

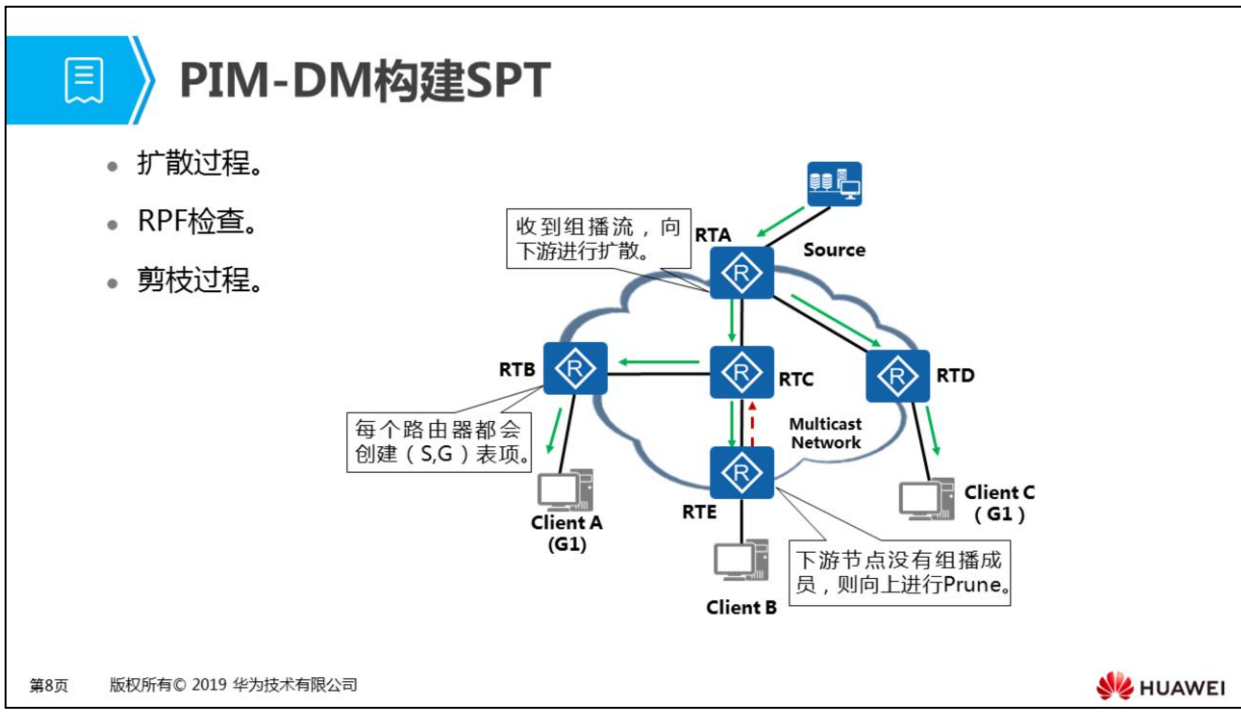
- PIM路由表项即通过PIM协议建立的组播路由表项。PIM网络中存在两种路由表项：（S，G）路由表项或（*，G）路由表项。S表示组播源，G表示组播组，*表示任意。
- （S，G）路由表项主要用于在PIM网络中建立SPT。对于PIM-DM网络和PIM-SM网络适用。
- （*，G）路由表项主要用于在PIM网络中建立RPT。对于PIM-SM网络适用。
- PIM路由器上可能同时存在两种路由表项。当收到源地址为S，组地址为G的组播报文，且RPF检查通过的情况下，按照如下的规则转发：
  - 如果存在（S，G）路由表项，则由（S，G）路由表项指导报文转发。
  - 如果不存在（S，G）路由表项，只存在（*，G）路由表项，则先依照（*，G）路由表项创建（S，G）路由表项，再由（S，G）路由表项指导报文转发。



- PIM（Protocol Independent Multicast）协议无关组播，目前常用版本是PIMv2，PIM报文直接封装在IP报文中，协议号为103，PIMv2组播地址为224.0.0.13。
- 在PIM组播域中，以组播组为单位建立从组播源到组成员的点到多点的组播转发路径。由于组播转发路径呈现树型结构，也称为组播分发树（MDT，Multicast Distribution Tree）。
- 组播分发树的特点：
  - 无论网络中的组成员有多少，每条链路上相同的组播数据最多只有一份。
  - 被传递的组播数据在距离组播源尽可能远的分叉路口才开始复制和分发。
- PIM有两种模式：
  - PIM-DM（Protocol Independent Multicast – Dense Mode）。
  - PIM-SM （Protocol Independent Multicast – Sparse Mode）。
- PIM-DM假设网络中的组成员分布非常稠密，每个网段都可能存在组成员。
- 其设计思想是：
  - 首先将组播数据报文扩散到各个网段。
  - 然后再裁剪掉不存在组成员的网段。
  - 通过周期性的“扩散—剪枝”，构建并维护一棵连接组播源和组成员的单向无环SPT。
- PIM-DM的关键工作机制包括邻居发现、扩散与剪枝、状态刷新、嫁接和断言。



- 在PIM-DM网络中，路由器周期性发送Hello消息来发现、建立并维护邻居关系。
  - pim timer hello interval，在接口视图下配置发送Hello消息的时间间隔。 Hello消息默认周期是30秒。
  - pim hello-option holdtime interval，在接口视图下配置Hello消息超时时间值。默认情况超时时间值为105秒。
- DR的选举：
  - 在PIM-DM中各路由器通过比较Hello消息上携带的优先级和IP地址，为多路访问网络选举指定路由器DR。
  - DR充当IGMPv1的查询器。
  - 接口DR优先级大的路由器将成为该MA网络的DR，在优先级相同的情况下，接口IP地址大的路由器将成为DR。
  - 当DR出现故障后，邻居路由器之间会重新选举DR。

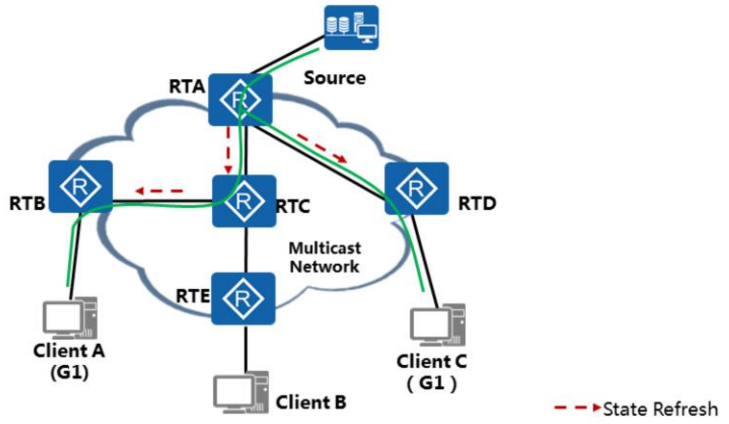


- 扩散过程：PIM-DM假设网络中所有主机都准备接收组播数据，当某组播源开始向组播组G发送数据时，具体过程如下：
  - 路由器接收到组播报文时会进行RPF检查。
  - 如果RPF检查通过，则创建（S，G）表项，然后将数据向所有下游PIM-DM节点转发，这个过程称为扩散（Flooding）。
  - 如果RPF检查没有通过，则将报文丢弃。
- RPF检查： 为了防止组播报文在转发过程中出现重复报文及环路的情况，路由器必须执行RPF检查。
  - 所谓RPF检查，就是指路由器通过查找去往组播源的路由来判断所收到的组播报文是否来自于“正确的”上游接口。某一路由器去往某一组播源的路由所对应的出接口称为该路由器上关于该组播源的RPF接口。一台路由器从某一接口收到一个组播报文后，如果发现该接口不是相应组播源的RPF接口，就意味着RPF检查失败，所收到的组播报文将被丢弃。
- 剪枝过程：当下游有没有组播成员，扩散组播报文会导致带宽资源的浪费。为避免带宽的浪费PIM-DM使用剪枝机制。
  - 当下游节点没有组播组成员，则路由器向上游节点发Prune消息，通知上游节点不用再转发数据到该分支。上游节点收到Prune消息后，就将相应的接口从其组播转发表项（S，G）对应的输出发送列表中删除。剪枝过程继续直到PIM-DM中仅剩下了必要的分支，这就建立了一个以组播源为根的SPT。
  - 各个被剪枝的节点同时提供超时机制，当剪枝超时时重新开始扩散—剪枝过程。剪枝状态超时计时器的默认值为210秒。

- PIM-DM的扩散—剪枝机制周期性进行，每3分钟重复一次，RTC对RTE所在网段处于剪枝状态，RTC对RTE的接口会维护一个“剪枝定时器”，当剪枝定时器超时，RTC就会恢复对RTE的数据转发，这样会导致不必要的网络资源浪费。

状态刷新

- 周期性地刷新剪枝状态。

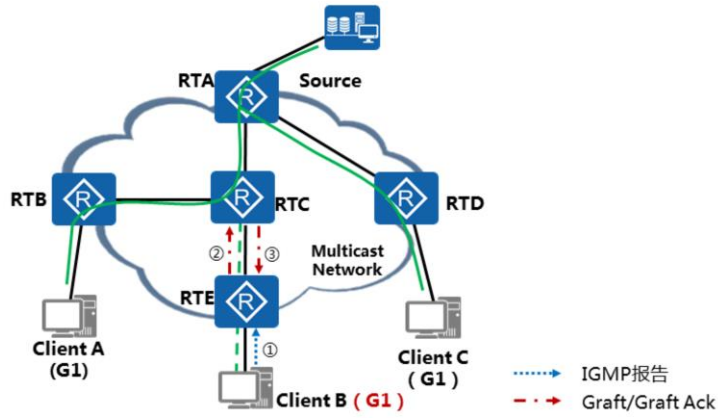


HUAWEI

- PIM DM协议采用状态刷新特性解决周期性“扩散-剪枝”带来的问题：离组播源最近的第一跳RTA周期性触发State Refresh消息。State Refresh消息在全网扩散，刷新所有设备上的剪枝定时器状态。
- 状态刷新使得RTE不再周期性的收到组播数据，但是当Client B加入G1组之后，如果一直是剪枝状态，Client B无法收到组播数据。
- 上述问题将如何解决？

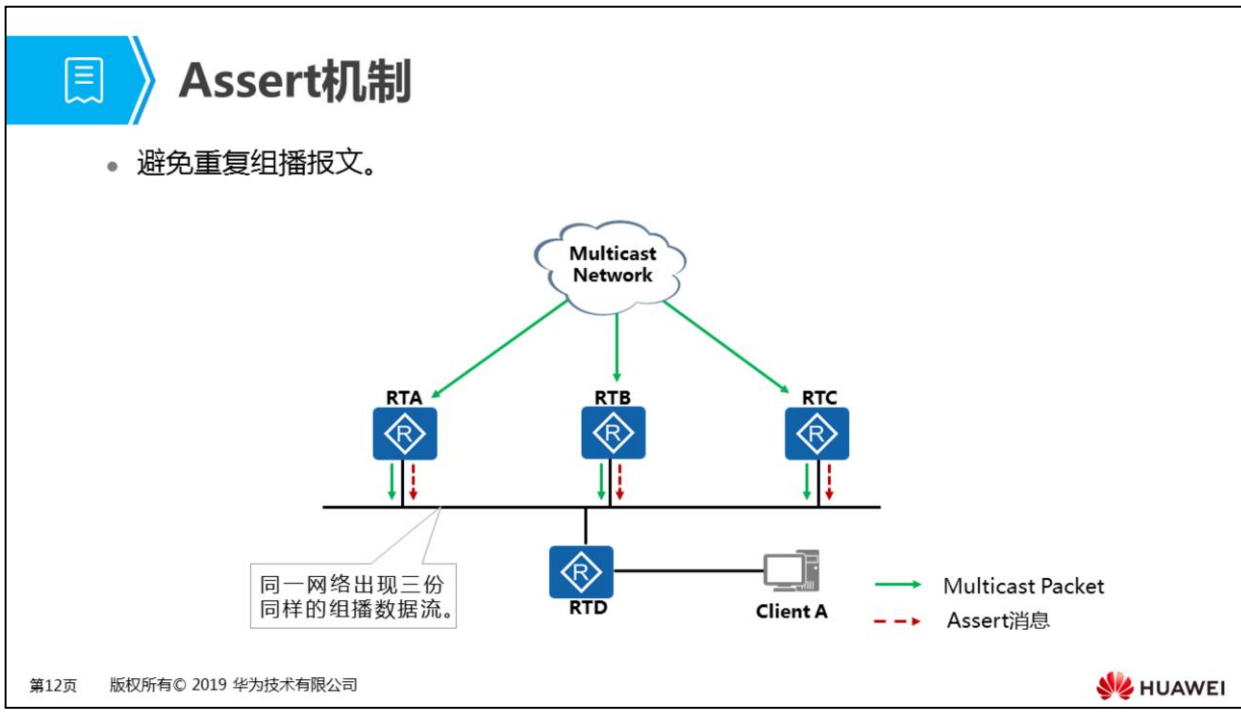
Graft机制

- 新的组成员加入组播组后，快速得到组播报文。



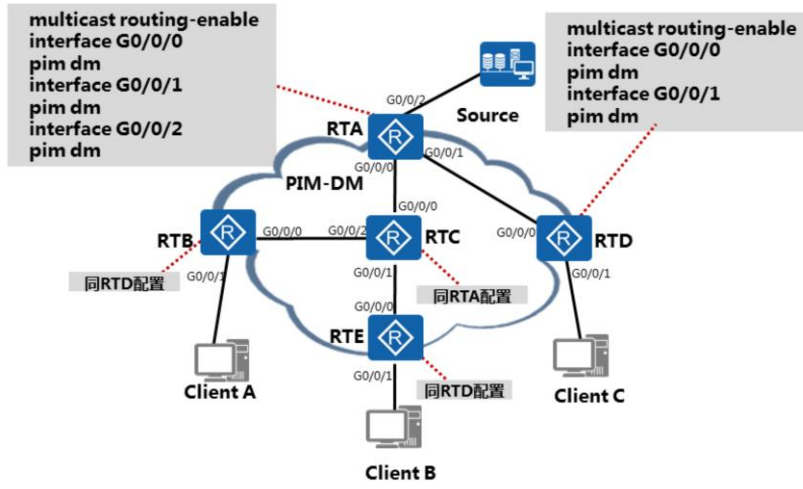
HUAWEI

- 如图所示，当Client B发送组播组G1的IGMP Report报文请求组播数据后。RTE收到Client B的IGMP Report报文，说明RTE具有转发组播数据需求，则立即向上游路由器RTC发送Graft消息，请求上游路由器恢复对应出接口的转发。RTC收到Graft消息后，向RTE回复Graft Ack并将连接RTE的出接口恢复为转发状态。



- 如图所示，RTA、RTB、RTC均从上游接口收到组播报文并通过了RPF检查，三台路由器的下游接口连接在同一网段。RTA、RTB、RTC都向该网段发送组播报文，三份重复的组播报文浪费带宽资源。
- 为避免重复的组播报文浪费带宽资源，PIM路由器在接收到邻居路由器发送的相同组播报文后，会以组播的方式向本网段的所有PIM路由器发送Assert消息，其中目的地址为224.0.0.13。其它PIM路由器在接收到Assert消息后，将自身参数与对方报文中携带的参数做比较，进行Assert竞选。竞选规则如下：
  - 到组播源的单播路由协议优先级较小者获胜。
  - 如果优先级相同，则到组播源的路由协议开销较小者获胜。
  - 如果以上都相同，则连接到接受者MA网络接口IP地址最大者获胜。
- 根据Assert竞选结果，路由器将执行不同的操作：
  - 获胜一方的下游接口称为Assert Winner，将负责后续对该网段组播报文的转发。
  - 落败一方的下游接口称为Assert Loser，后续不会对该网段转发组播报文，PIM路由器也会将其从（S，G）表项下游接口列表中删除。
- Assert竞选结束后，该网段上只存在一个下游接口，只传输一份组播报文。
- 所有Assert Loser可以周期性地恢复组播报文转发，从而引发周期性的Assert机制。

PIM-DM配置实现



HUAWEI

PIM-DM配置验证

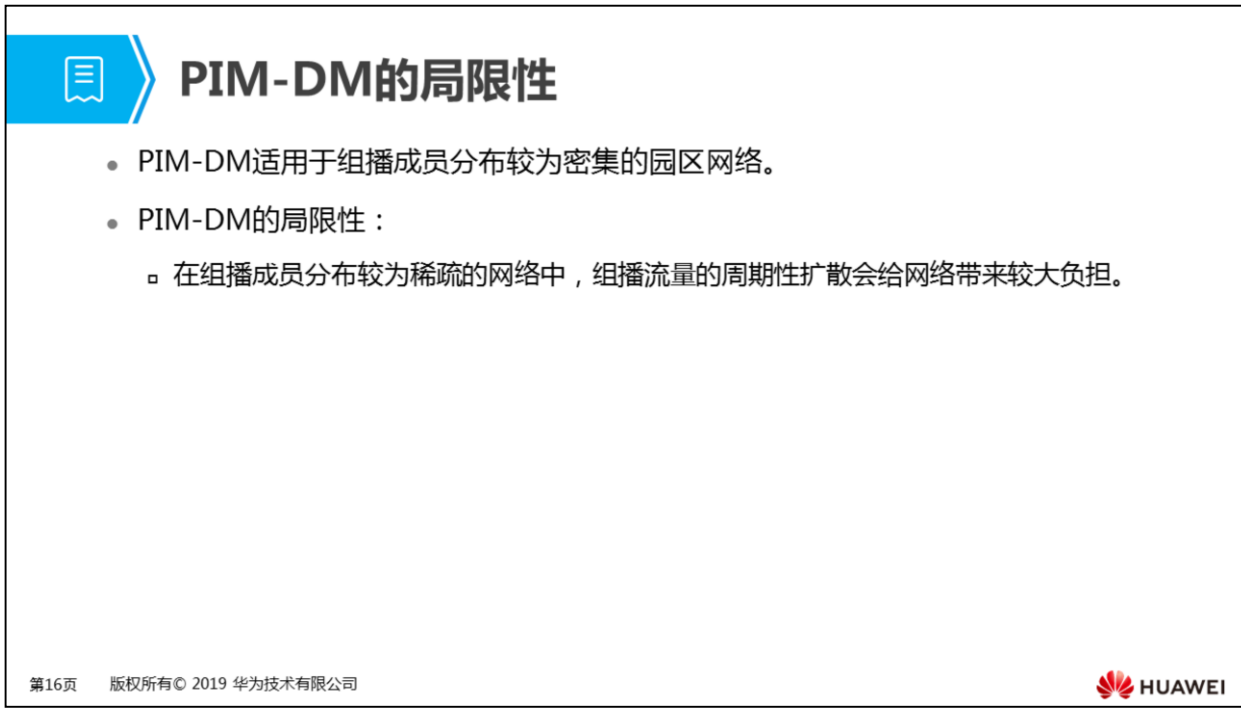
```
<RTD>display pim routing-table
VPN-Instance: public net
Total 1 (*, G) entry; 1 (S, G) entry

(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
```

目录

1. 组播报文转发的需求
2. PIM-DM的工作机制
3. **PIM-DM的局限性**
4. PIM-SM的工作机制

HUAWEI



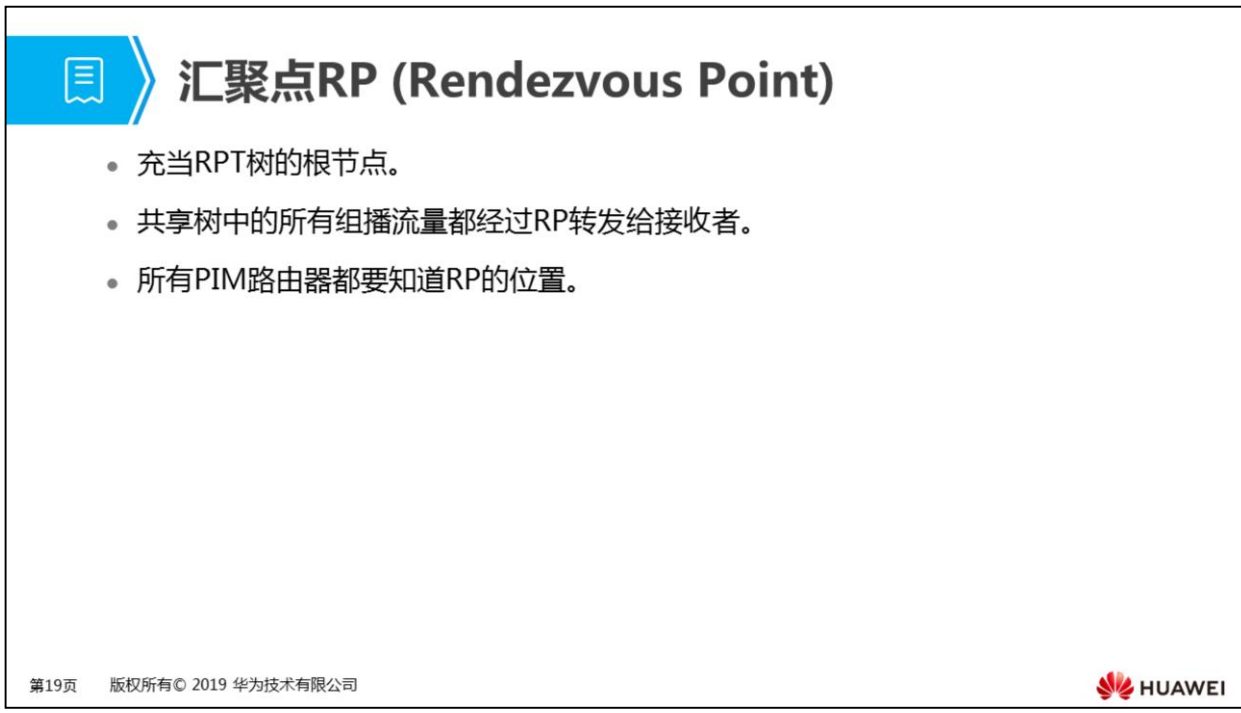
- PIM-DM适用于组播成员分布较为密集的园区网络。
- 在组播成员分布相对较为稀疏的大规模网络中（Internet），组播流量的周期性扩散/剪枝将给网络带来极大的负担。
- 对于PIM-DM的局限性，PIM-SM可以提供相对更加有效的解决方案。

目录

1. 组播报文转发的需求
2. PIM-DM的工作机制
3. PIM-DM的局限性
4. **PIM-SM的工作机制**

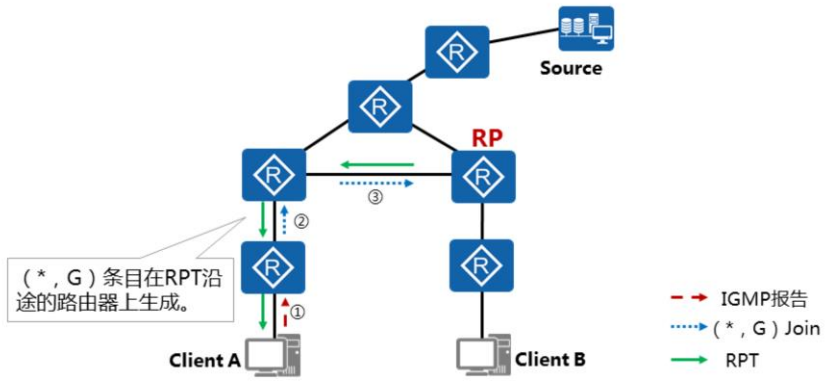
HUAWEI

- 相对于PIM-DM的“推（Push）模式”，PIM-SM使用“拉（Pull）模式”转发组播报文。PIM-SM假设网络中的组成员分布非常稀疏，几乎所有网段均不存在组成员，直到某网段出现组成员时，才构建组播路由，向该网段转发组播数据。一般应用于组播组成员规模相对较大、相对稀疏的网络。
- 基于这一种稀疏的网络模型，它的实现方法是：
  - 在网络中维护一台重要的PIM路由器：汇聚点RP（Rendezvous Point），可以为随时出现的组成员或组播源服务。网络中所有PIM路由器都知道RP的位置。
  - 当网络中出现组成员（用户主机通过IGMP加入某组播组G）时，最后一跳路由器向RP发送Join报文，逐跳创建（*，G）表项，生成一棵以RP为根的RPT。
  - 当网络中出现活跃的组播源（信源向某组播组G发送第一个组播数据）时，第一跳路由器将组播数据封装在Register报文中单播发往RP，在RP上创建（S，G）表项，注册源信息。
- PIM-SM的关键机制包括邻居建立、DR竞选、RP发现、RPT构建、组播源注册、SPT切换、Assert；同时也可通过配置BSR（Bootstrap Router）管理域来实现单个PIM-SM域的精细化管理。PIM-SM中PIM邻居建立过程以及Assert机制与PIM-DM相同。



- RP的作用：
  - RP是PIM-SM域中的核心路由器，担当RPT树根节点。
  - 共享树里所有组播流量都要经过RP转发给接收者。
- 用户通过配置命令限制RP所提供服务的组播组范围。
- RP可以静态指定也可动态选举：
  - 静态指定是指由管理员在每台PIM-SM路由器上进行配置，使得每台路由器获知RP的位置。
  - 动态选举是指通过专用协议在若干台C-RP（Candidate-RP）中选举产生。管理员需要开启选举协议并配置若干台PIM-SM路由器成为C-RP。
  - RP配置方式建议：
    - 中小型网络：建议选择静态RP方式，对设备要求低，也比较稳定。
    - 如果网络中只有一个组播源，建议选择直连组播源的设备作为静态RP，这样可以省略源端DR向RP注册的过程。
    - 采用静态RP方式要确保域内所有路由器（包括RP本身）的RP信息以及服务的组播组范围全网一致。
    - 大型网络：可以采用动态RP方式，可靠性高，可维护性强。
    - 如果网络中存在多个组播源，且分布密集，建议选择与组播源比较近的核心设备作为C-RP；如果网络中存在多个用户，且分布密集，建议选择与用户比较近的核心设备作为C-RP。

RPT及其建立过程

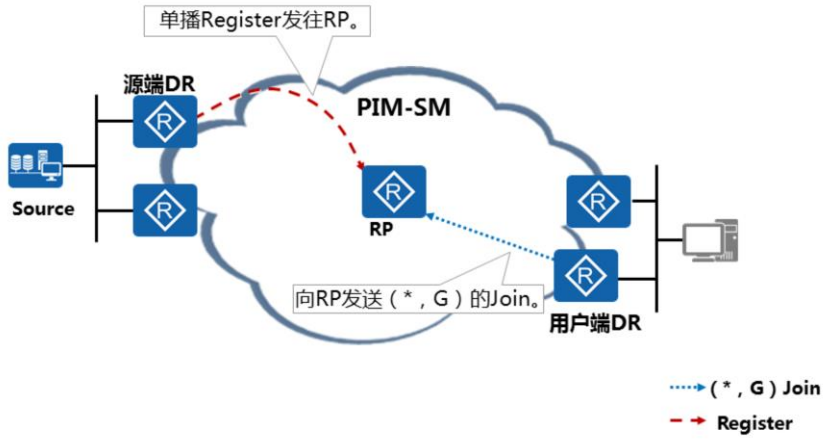


- 思考：如果连接Client A的路由器有两台，这两台路由器都会向RP发送（*，G）Join消息吗？

HUAWEI

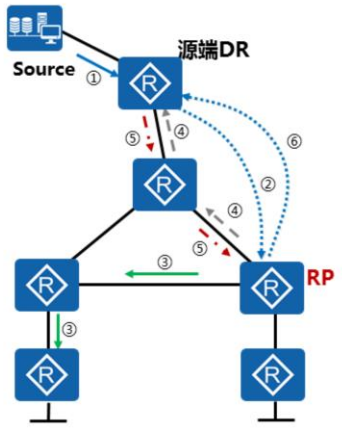
- RPT的建立过程：
  - 主机加入某个组播组时，发送IGMP成员通告。
  - 最后一跳路由器向RP发送（*，G）Join消息。
  - （*，G）Join消息到达RP的过程中，沿途各路由器都会生成相应的（*，G）组播转发条目。
- RPT实现了组播数据按需转发的目的，减少了数据泛洪对网络带宽的占用。

组播接收者侧DR与组播源侧DR



- 运行PIM-SM的网络，都会进行DR（Designated Router）的选举。其中有两种DR分别称为接收者侧DR和组播源侧DR。
  - 组播接收者侧DR：与组播组成员相连的DR，负责向RP发送（*，G）的Join加入消息。
  - 组播源侧DR：与组播源相连的DR，负责向RP发送单播的Register消息。
- PIM-SM中DR的选举原则与PIM-DM相同。

SPT的建立过程



RPT
第一个组播报文
Register/Register Stop
(S, G) Join消息
SPT

- SPT建好之后，组播报文沿SPT到达RP。

HUAWEI

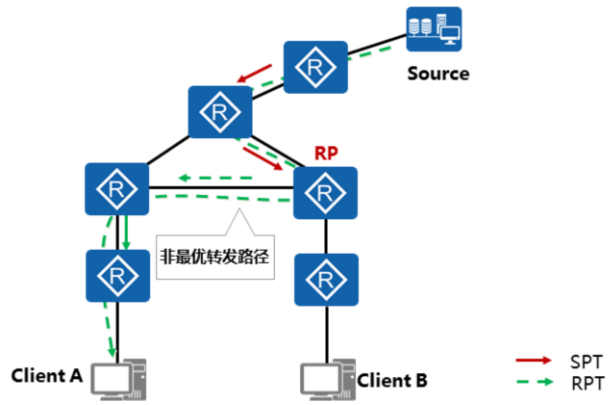
- 如图所示，在PIM-SM网络中，任何一个新出现的组播源都必须首先在RP处“注册”，继而才能将组播报文传输到组成员。具体过程如下：
  - 组播源向组播组发送第一个组播报文。
  - 源端DR将该组播报文封装成Register报文并以单播方式发送给相应的RP。
  - RP收到注册消息后，一方面从Register消息中提取出组播报文，并将该组播报文沿RPT分支转发给接收者。
  - 另一方面，RP向源端DR发送(S，G)Join消息，沿途路由器上都会生成相应(S，G)表项。从而建立了一颗由组播源至RP的SPT树。
  - SPT树建立后，组播源发出的组播报文沿该SPT转发至RP。
  - RP沿SPT收到该组播报文后，向源端DR单播发送Register-stop消息。

（*，G）与（S，G）条目关系

模式	类型	使用场景
PIM-DM	（S，G）	第一跳路由器到最后一跳路由器的SPT。
PIM-SM	（*，G）	RP到最后一跳路由器的RPT。
	（S，G）	源端DR到RP的SPT。
	（S，G）	Switchover之后，从第一跳路由器到最后一跳路由器的SPT。

PIM-SM的转发树

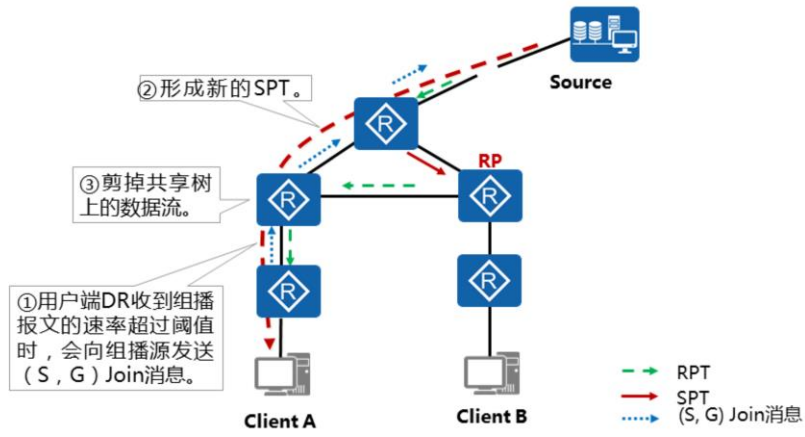
- SPT和RPT构成组播报文的转发路径，存在哪些问题？



HUAWEI

- PIM-SM同时包含了SPT和RPT。通常情况下，组播源发出的组播报文会沿SPT到达RP，然后从RP沿RPT到达接收者。
- 在这种情况下，从组播源到接收者的路径不一定是最优的，并且RP的工作负担非常大。为此，我们可以启用RPT向SPT进行的切换机制。

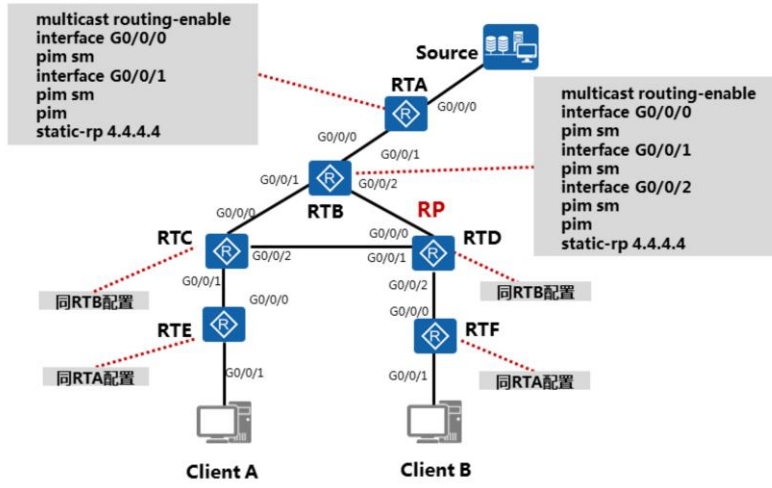
Switchover机制



HUAWEI

- PIM-SM通过指定一个利用带宽的SPT阈值可以实现RPT到SPT的切换。
- 用户端DR周期性检测组播报文的转发速率，一旦发现从RP发往组播组G的报文速率超过阈值，则触发SPT切换：
  - 用户端DR逐跳向源端DR发送（S，G）Join报文并创建（S，G）表项，建立源端DR到用户端DR的SPT。
  - SPT建立后，用户端DR会沿着RPT逐跳向RP发送剪枝报文，收到剪枝报文的路由器将（*，G）复制成相应的（S，G）,并将相应的下游接口置为剪枝状态。剪枝结束后，RP不再沿RPT转发组播报文到组成员端。
  - 如果SPT不经过RP，RP会继续向源端DR逐跳发送剪枝报文，删除（S，G）表项中相应的下游接口。剪枝结束后，源端DR不再沿“源端DR-RP”的SPT转发组播报文到RP。
- 在VRP中，缺省情况下连接接收者的路由器在探测到组播源之后（即接收到第一个数据报文），便立即加入最短路径树，即从RPT向SPT切换。
- 通过RPT树到SPT树的切换，PIM-SM能够以比PIM-DM更精确的方式建立SPT转发树。

PIM-SM配置实现



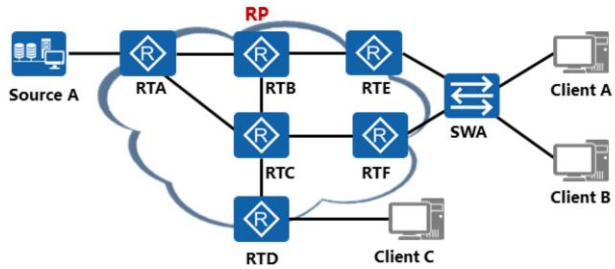
HUAWEI

PIM-SM配置验证

```
<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
```

组播综合实验

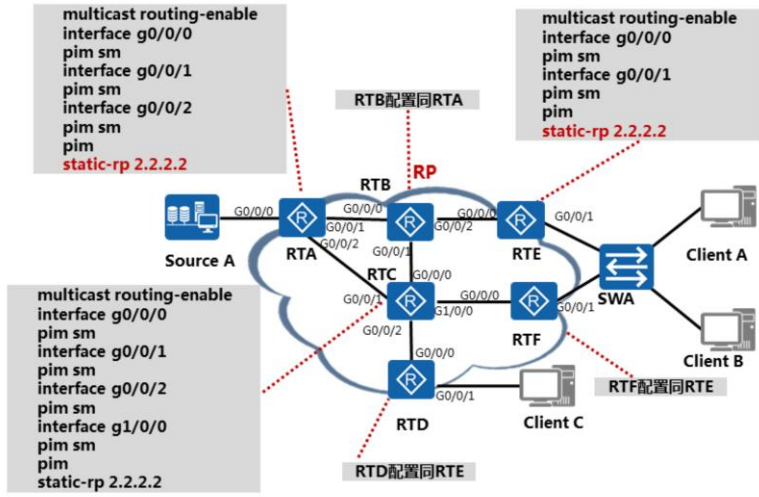


- 如图所示，要求在网络中部署PIM-SM协议且以静态方式指定RTB为RP。
- 用户侧网络配置IGMPv2协议，同时需要尽可能地降低用户侧网络资源消耗，提高安全性。
- RTE和RTF连接接收者，要求在RTE和RP之间建立RPT。
- RTD连接重要的用户网络，当用户加入组播组238.1.1.1组后，需要马上就能收到组播数据。

HUAWEI

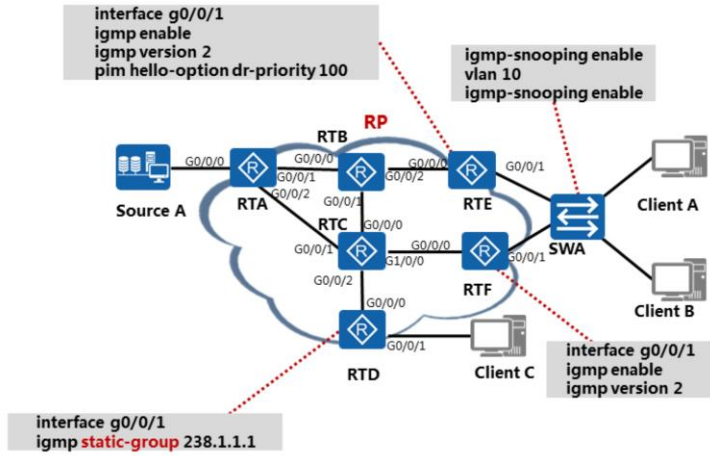
- 需求分析：
  - 使能组播路由功能是配置PIM-SM的前提，首先在路由器上使能组播路由功能，其次在路由器接口上使能PIM-SM功能，最后在PIM视图下静态配置RTB为RP。
  - 在与用户侧相连的路由器接口上使能IGMPv2。为了降低资源消耗、提高安全性，需要在SWA上使能IGMP Snooping功能，使交换机进行有效、安全的组播帧转发。
  - 用户端DR负责向RP建立RPT。根据DR的选举规则，需要把RTE接口的DR优先级设置为大于1的值（DR优先级默认为1）。
  - 也就是说RTD上需要具有238.1.1.1的组播转发表项，RTD在收到IGMP Report报文后，立即转发组播报文。可通过在RTD接口下配置静态加入238.1.1.1命令实现。

PIM-SM配置实现 (1)



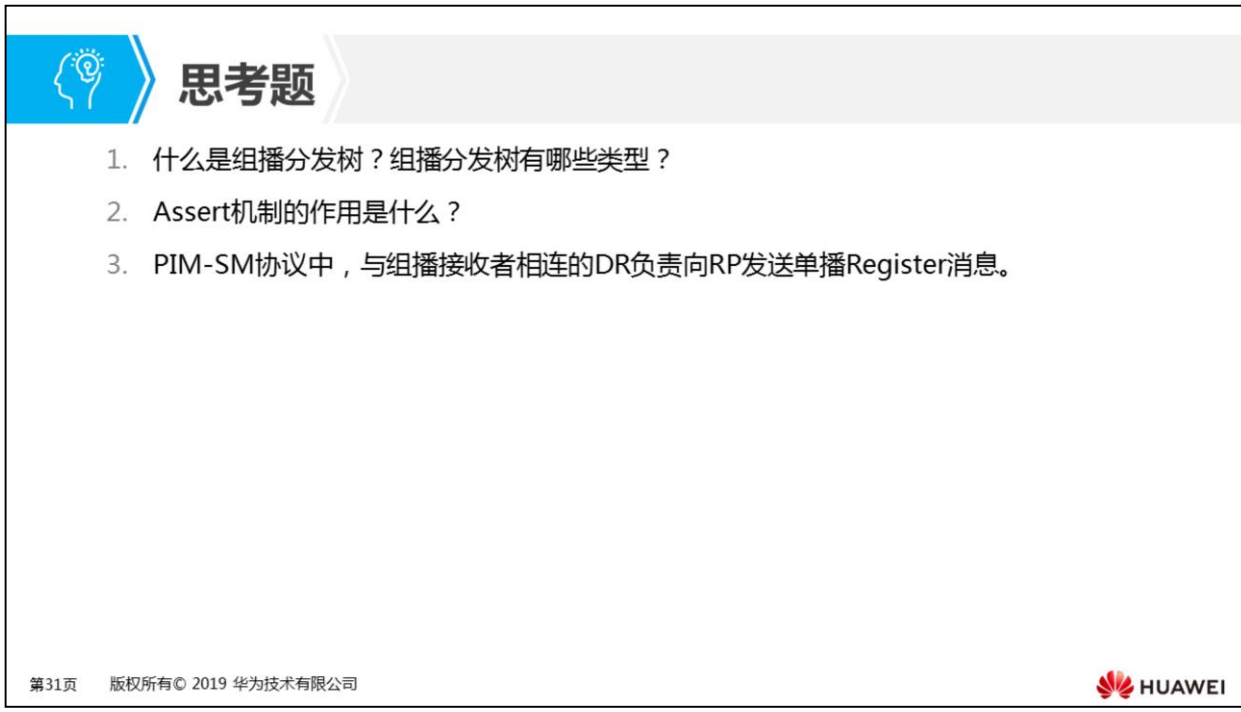
HUAWEI

PIM-SM配置实现 (2)



  HUAWEI

- 配置静态组播组：
  - 在某些特殊的应用场景中，比如：网络中存在稳定的组播组成员；主机无法发送报告报文，但是又需要将组播数据转发到该网段。
    - 为了实现组播数据的快速、稳定转发，或者将组播数据引流到接口，可以在组播路由器的用户侧接口上配置静态组播组。在接口上配置静态组播组后，组播路由器就认为此接口网段上一直存在该组播组的成员，从而转发该组的组播数据。
    - 当成员主机无法解析组播ping报文并作出回应时，可以在组播路由器的用户侧接口上配置组播ping功能。这样，接口除了正常接收组播数据之外，还可以对收到的组播ping报文作出回应，从而使定位问题更加灵活、方便。



- 答案：组播分发树是指从组播源到接收者之间形成的一个单向无环数据传输路径。组播分发树有两类：SPT和RPT。
- 答案：Assert可以避免在共享网络（如Ethernet）中相同报文的重复发送。通过Assert机制在共享网络中来选定一个唯一的转发者。其他落选路由器则剪掉对应的接口以禁止转发信息。
- 答案：错误，与组播源相连的DR负责向RP发送单播的Register消息。


谢 谢
www.huawei.com

路由控制

HUAWEI

前言

- 在企业网络的设备通信中，常面临一些非法流量访问的安全性及流量路径不优等问题，故为保证数据访问的安全性、提高链路带宽利用率，就需要对网络中的流量行为进行控制，如控制网络流量可达性、调整网络流量路径等。
- 而当面对更加复杂、精细的流量控制需求时，就需要灵活地使用一些工具来实现，本课程将主要介绍一些有关流量控制的常用工具及其使用场景。

目标

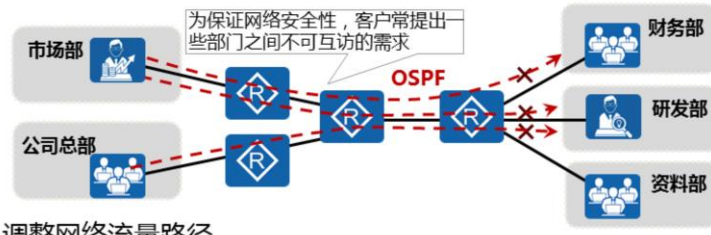
- 学完本课程后，您将能够：
  - 掌握控制网络流量可达性的实现方式
  - 掌握调整网络流量路径的实现方式
  - 熟悉路由引入导致的问题及解决方法

目录

1. **流量行为控制需求**
2. 控制流量可达性
3. 调整网络流量路径
4. 路由引入导致的问题及解决办法

对流量行为的控制需求分析

1. 控制网络流量可达性。



2. 调整网络流量路径。

HUAWEI

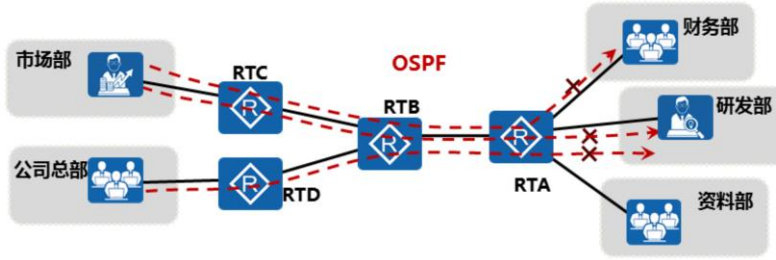
- 控制网络流量可达性：如图，为满足业务需求和保证数据访问安全性，要求市场部不能访问财务部、研发部，公司总部不能访问研发部。
- 调整网络流量路径：如图，根据OSPF协议计算生成的路由，市场部和财务部访问公司总部都选择通过一条开销最小的路径，即使该路径发生拥塞也如此，而另外一条路径的链路带宽则一直处于空闲状态，这样就造成了带宽浪费的问题。

目录

1. 流量行为控制需求
2. **控制流量可达性**
   - 路由策略方式
   - 流量过滤方式
3. 调整网络流量路径
4. 路由引入导致的问题及解决办法

控制网络流量可达性

- 思考：如何控制网络流量可达性？

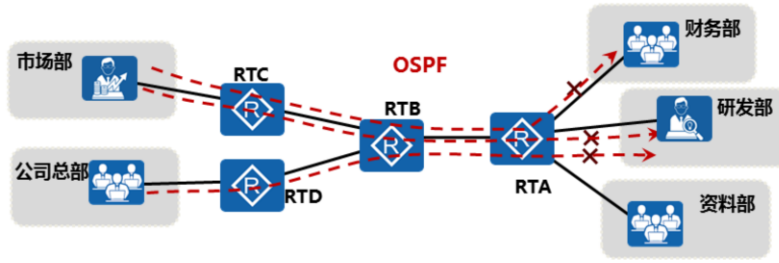


- 解决方案一：可通过修改路由条目（即对接收和发布的路由进行过滤）来控制流量可达性，这种方式称为**路由策略**。
- 解决方案二：可使用Traffic-Filter工具对数据进行过滤，这种方式称为**流量过滤**。

HUAWEI

- 路由策略（Routing Policy）的作用是当路由器在发布、接收和引入路由信息时，可根据实际组网需要实施一些策略，以便对路由信息进行过滤或改变路由信息的属性，如：
  - 控制路由的发布：只发布满足条件的路由信息。
  - 控制路由的接收：只接收必要、合法的路由信息，以控制路由表的容量，提高网络的安全性。
  - 过滤和控制引入的路由：一种路由协议在引入其它路由协议时，只引入一部分满足条件的路由信息，并对所引入的路由信息的某些属性进行设置，以使其满足本协议的要求。
  - 设置特定路由的属性：为通过路由策略过滤的路由设置相应的属性。

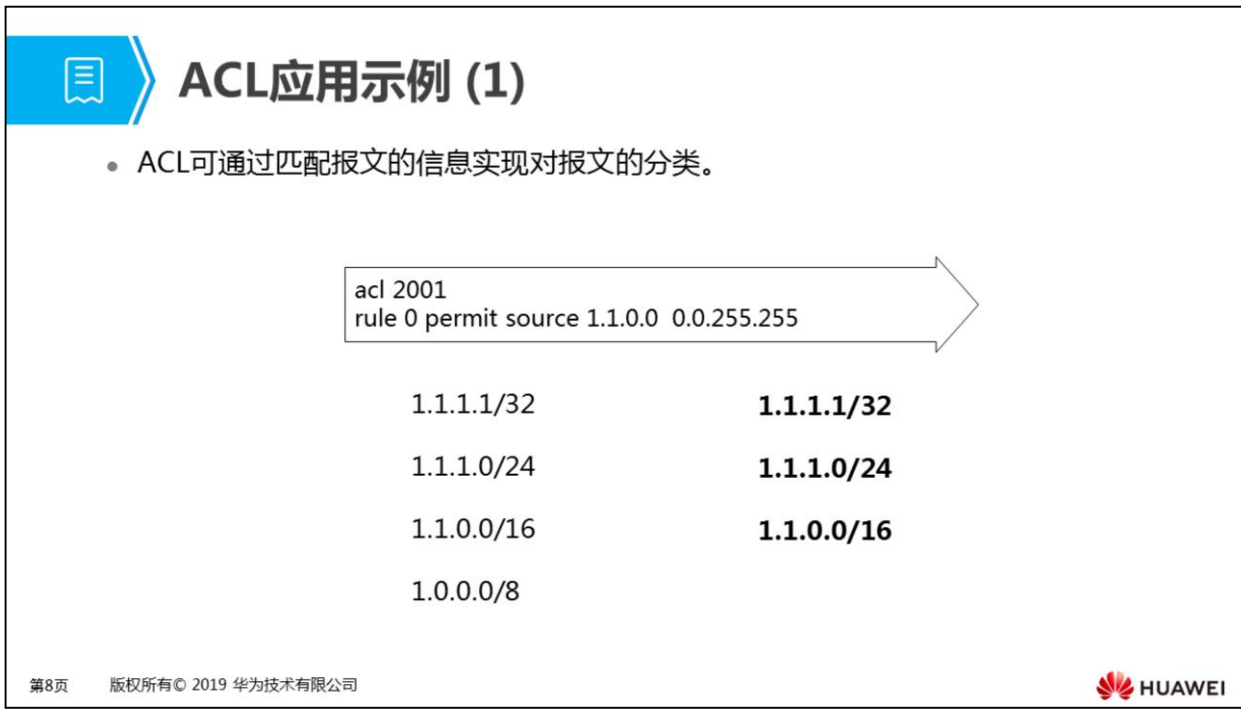
解决方案一：采用路由策略方式



- 可利用**Filter-Policy**工具对RTA向OSPF引入的路由和RTC写入路由表的路由进行过滤：
  - 首先使用ACL或IP-Prefix List工具来匹配目标流量；
  - 然后在协议视图下，利用Filter-Policy向目标流量发布策略。
- 可利用**Route-Policy**工具，在RTA引入直连路由时对路由进行过滤：
  - 首先使用ACL或IP-Prefix List工具来匹配目标流量；
  - 然后在协议视图下，利用Route-Policy对引入的路由条目进行控制。

HUAWEI

- 路由策略的实现分为两个步骤：
  - 定义规则：首先要定义将要实施路由策略的路由信息的特征，即定义一组匹配规则，可以以路由信息中的不同属性作为匹配依据进行设置，如目的地址、AS号等；
  - 应用规则：根据设置的匹配规则，再将它们应用于路由的发布、接收和引入等过程中。
- 目前提供了如下几种过滤器供路由协议引用：
  - 访问控制列表；
  - 地址前缀列表；
  - AS路径过滤器；
  - 团体属性过滤器；
  - 扩展团体属性过滤器；
  - 路由标识属性过滤器。



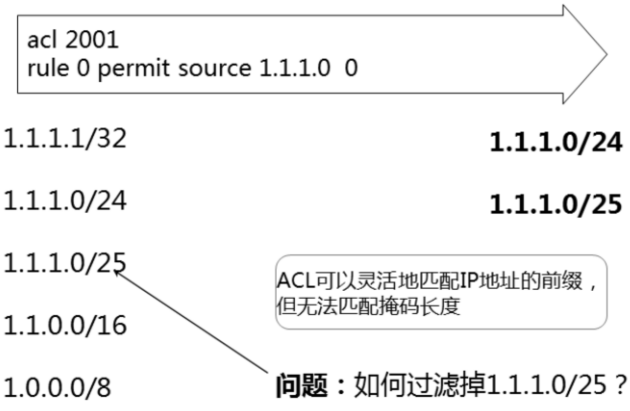
- 访问控制列表ACL（Access Control List）是由permit或deny语句组成的一系列有顺序规则的集合，它通过匹配报文的信息实现对报文的分类。
- ACL的分类：
  - 基本ACL：主要基于源地址、分片标记和时间段信息对数据包进行分类定义，编号范围为2000-2999。
  - 高级ACL：可以基于源地址、目的地址、源端口号、目的端口号、协议类型、优先级、时间段等信息对数据包进行更为细致的分类定义，编号范围为3000-3999。
  - 二层ACL：主要基于源MAC地址、目的MAC地址和报文类型等信息对数据包进行分类定义，编号范围为4000-4999。
  - 用户自定义ACL：主要根据用户自定义的规则对数据报文做出相应的处理，编号范围为5000-5999。
- 一个ACL可以由多条“deny | permit”语句组成，每一条语句描述了一条规则。设备收到数据流量后，会逐条匹配ACL规则，看其是否匹配。如果不匹配，则继续匹配下一条。一旦找到一条匹配的规则，就会执行规则中定义的动作，且不再继续与后续规则进行匹配；如果找不到匹配的规则，则设备会对报文直接进行转发。
- 需要注意的是，ACL中定义的这些规则可能存在重复或矛盾的地方。规则的匹配顺序决定了规则的优先级，ACL通过设置规则的优先级来处理规则之间重复或矛盾的情形。

ACL应用示例 (2)

```
acl 2001
rule 0 permit source 1.1.1.1  0
rule 1 deny source 1.1.1.0  0
rule 2 permit source 1.1.0.0  0.0.255.0
rule 3 deny source any
```

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	

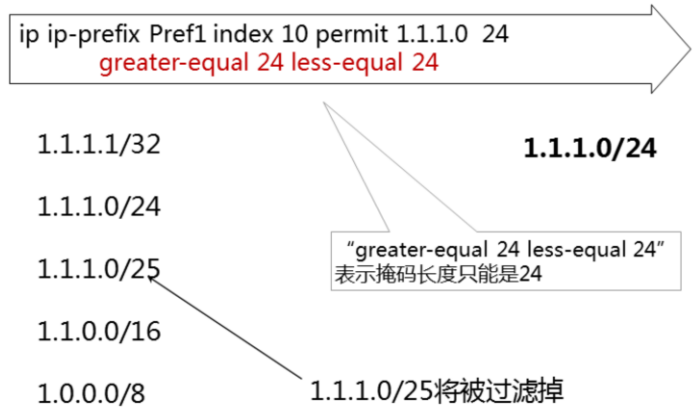
ACL应用示例 (3)



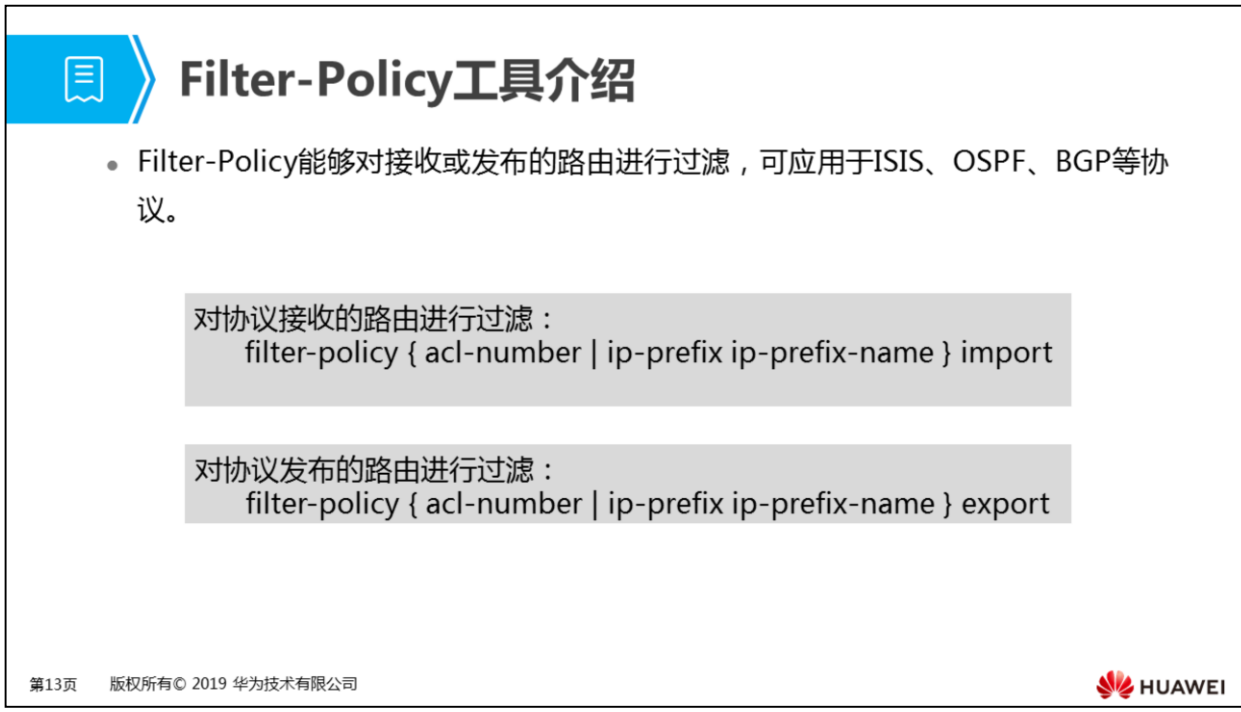
HUAWEI

- 地址前缀列表即IP-Prefix List。可以通过地址前缀列表，将与所定义的前缀过滤列表相匹配的路由，根据定义的匹配模式进行过滤，以满足使用者的需要。
- 前缀列表的组成及匹配规则：
  - 前缀过滤列表由IP地址和掩码组成，IP地址可以是网段地址或者主机地址，掩码长度的配置范围为0～32。
  - IP-Prefix List中的每一条IP-Prefix都有一个序列号index，匹配的时候将根据序列号从小到大进行匹配。
  - 如果不配置IP-Prefix的index，那么对应的index在上次配置的同名IP-Prefix的index的基础上，以步长为10进行增长。如果配置的IP-Prefix的名字与index都和已经配置了的一项IP-Prefix List的相同，仅仅只是匹配的内容不同，则该IP-Prefix List将覆盖原有的IP-Prefix List。
  - 当所有前缀过滤列表均未匹配时，缺省情况下，存在最后一条默认匹配模式为deny。当引用的前缀过滤列表不存在时，则默认匹配模式为permit。
- 前缀掩码长度范围：
  - 前缀过滤列表可以进行精确匹配或者在一定掩码长度范围内匹配，并通过配置关键字greater-equal和less-equal来指定待匹配的前缀掩码长度范围。如果没有配置关键字greater-equal或less-equal，前缀过滤列表会进行精确匹配，即只匹配掩码长度为与前缀过滤列表掩码长度相同的IP地址路由；如果只配置了关键字greater-equal，则待匹配的掩码长度范围为从greater-equal指定值到32位的长度；如果只匹配了关键字less-equal，则待匹配的掩码长度范围为从指定的掩码到关键字less-equal的指定值。

IP-Prefix List应用示例 (2)



HUAWEI



- 应用各协议中的Filter-Policy工具可通过引用ACL或地址前缀列表，对接收、发布和引入的路由进行过滤。
- 对于距离矢量协议和链路状态协议，Filter-Policy工具的操作过程是不同的：
  - 距离矢量协议是基于路由表生成路由的，因此过滤器会影响从邻居接收的路由和向邻居发布的路由。
  - 链路状态路由协议是基于链路状态数据库来生成路由的，且路由信息隐藏在链路状态LSA中，但Filter-Policy不能对发布和接收的LSA进行过滤，故Filter-Policy不影响链路状态通告或链路状态数据库的完整性以及协议路由表，而只会影响本地路由表，且只有通过过滤的路由才被添加到路由表中，没有通过过滤的路由不会被添加进路由表。
  - 不同协议应用filter-policy export命令对待发布路由的影响范围不同：
    - 对于距离矢量协议，会对引入的路由信息、本协议发现的路由信息进行过滤。
    - 对于链路状态协议，只对引入的路由信息进行过滤。

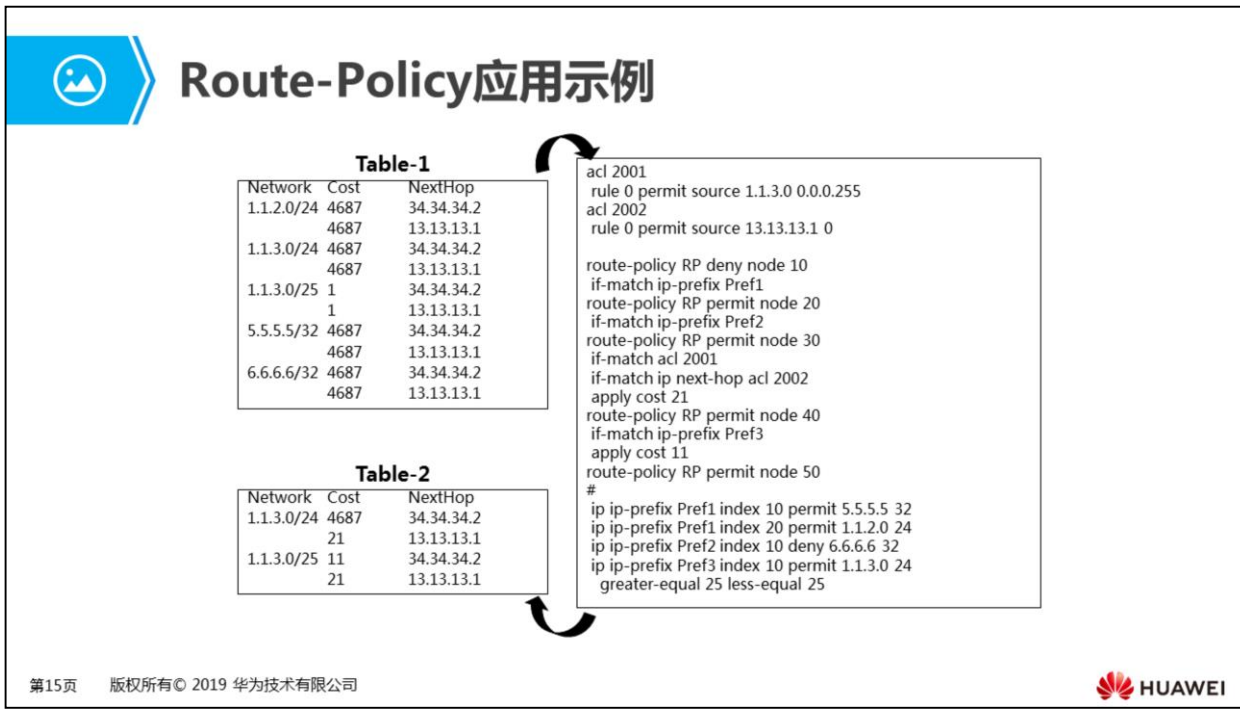
Route-Policy工具介绍

- Route-Policy是一种功能非常强大的路由策略工具，它可以灵活地与ACL、IP-Prefix List、As-Path-Filter等其它工具配合使用

```
Route-Policy：
   route-policy route-policy-name { permit | deny } node node
     if-match {acl/cost/interface/ip next-hop/ip-prefix}
     apply {cost/ip-address next-hop/tag}
```

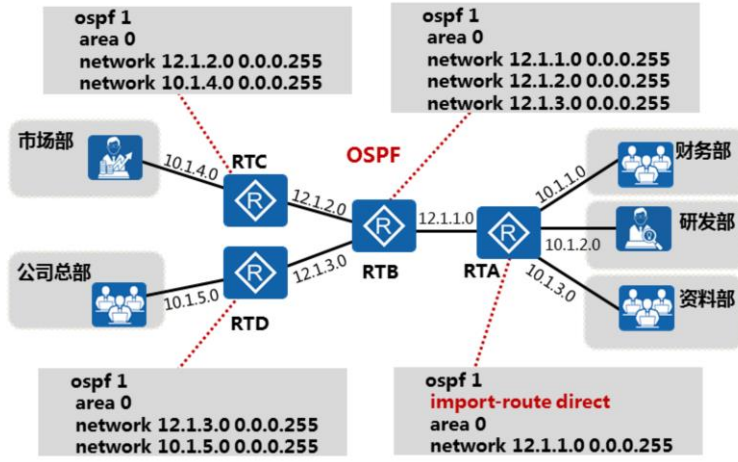
- Route-Policy由若干个node构成，node之间是“或”的关系。且每个node下可以有若干个if-mach和apply子句，if-match之间是“与”的关系

- Route-Policy的每个node都有相应的permit模式或deny模式。如果是permit模式，则当路由项满足该node的所有if-match子句时，就被允许通过该node的过滤并执行该node的apply子句，且不再进入下一个node；如果路由项没有满足该node的所有if-match子句，则会进入下一个node继续进行过滤。如果是deny模式，则当路由项满足该node的所有if-match子句时，就被拒绝通过该node的过滤，这时apply子句不会被执行，并且不进入下一个node；否则就进入下一个node继续进行过滤。



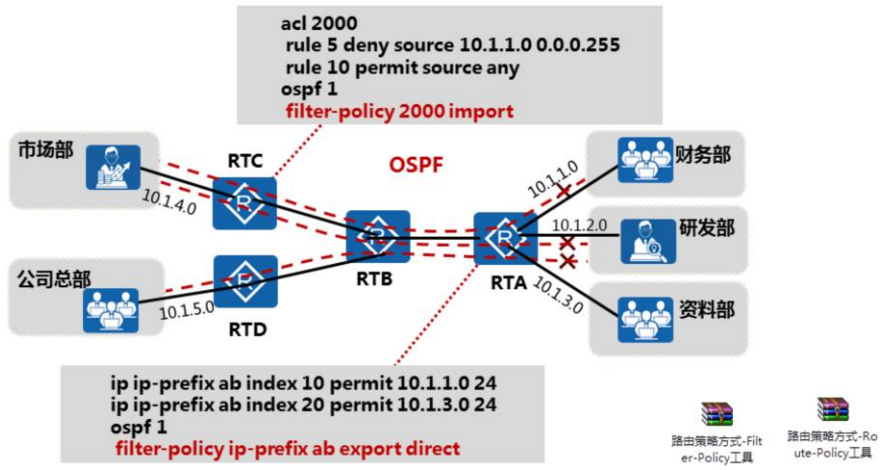
- Pref1用来匹配5.5.5.5/32或1.1.2.0/24，它们将被route-policy RP的node 10过滤掉（deny），所以Table-2中见不到5.5.5.5/32和1.1.2.0/24。
- Pref2用来过滤6.6.6.6/32（deny），所以尽管route-policy RP的node 20是permit，6.6.6.6/32仍然会被过滤掉。因此，Table-2中见不到6.6.6.6/32。
- route-policy RP的node 30定义了两个if-match语句，分别针对ACL 2001和ACL 2002。匹配ACL 2001的路由有1.1.3.0/24（下一跳为34.34.34.2）、1.1.3.0/24（下一跳为13.13.13.1）、1.1.3.0/25（下一跳为34.34.34.2）、1.1.3.0/25（下一跳为13.13.13.1），同时又匹配ACL 2002的路由有1.1.3.0/24（下一跳为13.13.13.1）和1.1.3.0/25（下一跳为13.13.13.1）。于是，1.1.3.0/24（下一跳为13.13.13.1）和1.1.3.0/25（下一跳为13.13.13.1）的cost被修改为21。
- 1.1.3.0/24（下一跳为34.34.34.2）和1.1.3.0/25（下一跳为34.34.34.2）继续尝试通过route-policy RP的node 40。由于1.1.3.0/25满足Pref3，所以1.1.3.0/25（下一跳为34.34.34.2）的cost被修改为11。
- 最后，1.1.3.0/24（下一跳为34.34.34.2）通过了route-policy RP的node 50。

路由策略方式配置实现 (1)



HUAWEI

路由策略方式配置实现 (2)



- RTA还可以使用Route-Policy工具进行路由控制：
  - 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

路由策略方式配置实现 (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
```


- 注明：上面显示的是部分关键信息，并非全部。

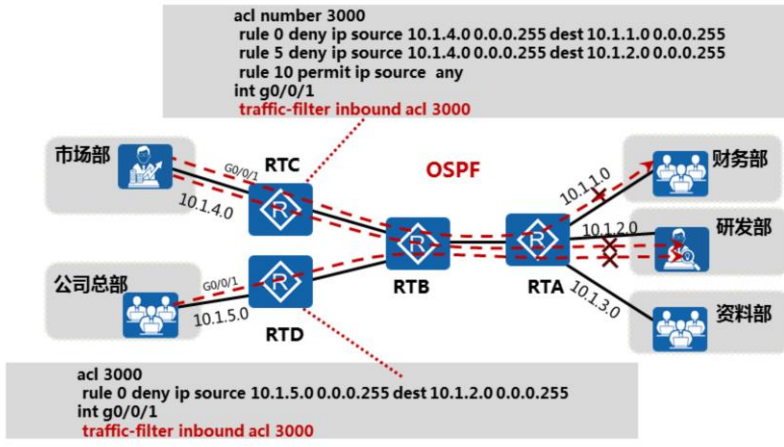
目录

1. 流量行为控制需求
2. **控制流量可达性**
   - 路由策略方式
   - **流量过滤方式**
3. 调整网络流量路径
4. 路由引入导致的问题及解决办法

HUAWEI

解决方案二：采用流量过滤方式 (1)

- 基于自定义策略实现：使用Traffic-Filter工具对数据进行过滤。



策略路由方式-Traffic-Filter工具

HUAWEI

解决方案二：采用流量过滤方式 (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-市场部>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
```


- 经测试，在设置完流量过滤后，RTC的路由表仍然有全网的路由，且市场部无法访问财务部和研发部，对于其他部门仍可正常访问。同样，RTD的路由表也有全网的路由，且公司总部无法访问研发部，其他仍可正常访问。

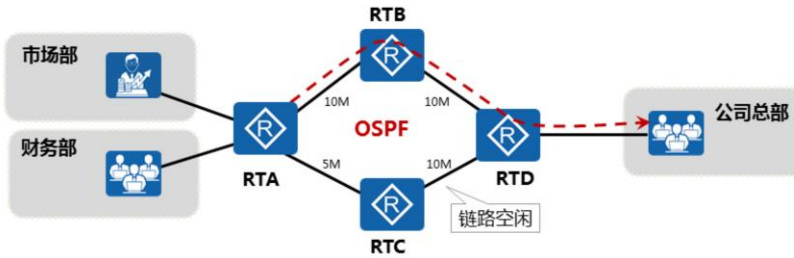
目录

1. 流量行为控制需求
2. 控制流量可达性
3. **调整网络流量路径**
   - 路由策略方式
   - 策略路由方式
4. 路由引入导致的问题及解决办法

HUAWEI

调整网络流量路径 - 单协议简单场景

- 在后期对网络进行优化时，常出现调整网络流量路径的需求。

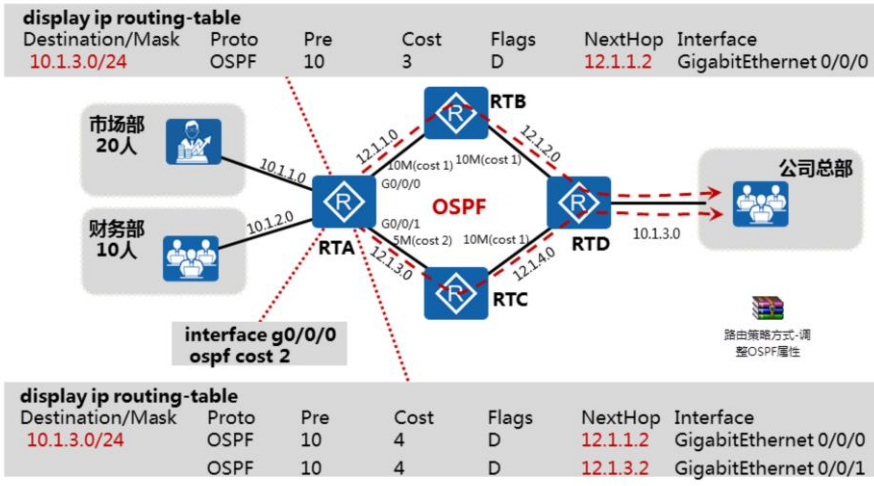


- 解决方案一：可通过**路由策略方式**修改协议属性来控制路由表条目，从而调整流量路径。
- 解决方案二：可采用**策略路由方式**在查找路由表之前控制流量行为。

HUAWEI

- 可通过修改协议本身的一些属性来控制路由条目，从而影响流量转发路径：
  - 若运行OSPF或ISIS协议，可通过调整接口Cost属性值来实现；
  - 若运行RIP协议，可通过调整Metric或下一跳属性来实现；
  - 若运行BGP协议，则可通过调整AS-Path、Local_Pref、MED、Community等属性来实现。

解决方案一：采用路由策略方式

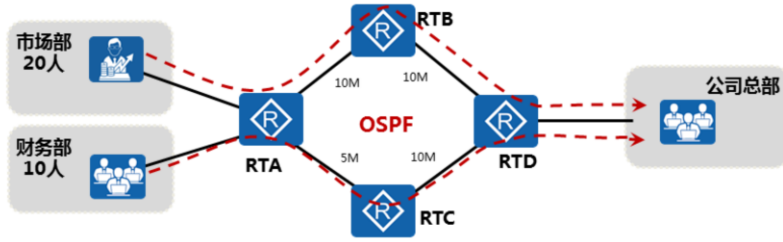


HUAWEI

- 传统的路由转发原理是首先根据报文的目的地址查找路由表，然后进行报文转发，思考该方式有何缺点？是否可以满足更复杂、更精确的控制需求？

解决方案一的局限性

- 为充分利用链路带宽，现要求市场部访问总部流量路径为RTA-RTB-RTD，财务部访问总部流量路径为RTA-RTC-RTD。



- 如图，若采用解决方案一来实现以上需求，由于其只能依据数据包的目的地址做转发策略，所以无法满足需求；故当出现基于源地址、目的地址或基于应用层等一些复杂的控制需求时，就体现出其局限性。

HUAWEI

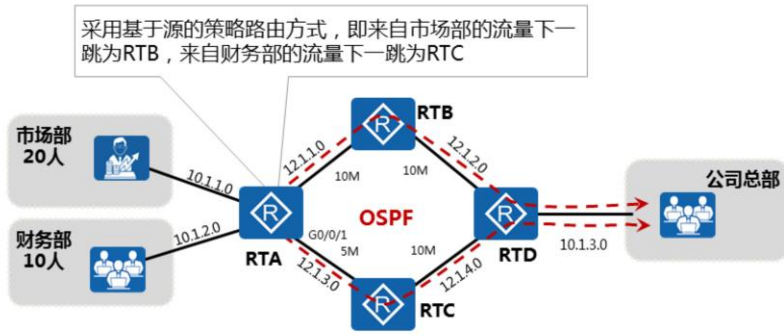
- 正是由于路由策略实现方式的缺陷，促使了目前越来越多的用户希望能够在传统路由转发的基础上，根据自己定义的策略进行报文转发和选路。策略路由使网络管理者不仅能够根据报文的目的地址来制定策略，而且还能够根据报文的源地址、报文大小和链路质量等属性来制定策略路由，以改变数据包转发路径，满足用户需求。

目录

1. 流量行为控制需求
2. 控制流量可达性
3. **调整网络流量路径**
   - 路由策略方式
   - **策略路由方式**
4. 路由引入导致的问题及解决办法

HUAWEI

解决方案二：采用策略路由方式 (1)



- 策略路由方式常采用Traffic-Policy工具来实现：
  - 首先使用ACL工具匹配目标流量；
  - 然后对目标流量定义行为，如修改下一跳。

  HUAWEI

- 策略路由PBR（Policy Based Routing）与单纯依照IP报文的目的地址查找路由表进行转发有所不同，它是一种依据用户制定的策略而进行流量转发的机制。
- 策略路由的查找优先级比路由策略高，当路由器接收到数据包并进行转发时，会优先根据策略路由的规则进行匹配，如果能匹配上，则根据策略路由进行转发，否则按照路由表中的路由条目来进行转发。其中策略路由不改变路由表中的任何内容，它可以通过预先设置的规则来影响数据报文的转发。
- 策略路由PBR分为：
  - 本地策略路由：对本设备发送的报文实现策略路由，比如本机下发的ICMP、BGP等协议报文。
    - 当用户需要实现不同源地址的报文或者不同长度的报文通过不同的方式进行发送时，可以配置本地策略路由。
    - 常用Policy-Based-Route工具来实现。
  - 接口策略路由：对本设备转发的报文生效，对本机下发的报文不生效。
    - 当用户需要将收到的某些报文通过特定的下一跳地址进行转发时，需要配置接口策略路由。使匹配重定向规则的报文通过特定的下一跳出口进行转发，不匹配重定向规则的报文则根据路由表直接转发。接口策略路由多应用于负载分担和安全监控。
    - 常用Traffic-Policy工具来实现。
  - 智能策略路由：基于链路质量信息为业务数据流选择最佳链路。

- 当用户需要为不同业务选择不同质量的链路时，可以配置智能策略路由。
- 常用Smart-Policy-Route工具来实现，在本课程中不做重点介绍。

解决方案二：采用策略路由方式 (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
```

- 本示例采用的是MQC的配置方式。

解决方案二：采用策略路由方式 (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        GigabitEthernet0/0/2
        10.1.2.0/24  Direct  0    0           D   10.1.2.2        GigabitEthernet4/0/0
        10.1.3.0/24  OSPF    10   3           D   12.1.1.2        GigabitEthernet0/0/0
```

```
PC-市场部>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-财务部>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
```

路由策略与策略路由的区别

路由策略	策略路由
基于控制平面，会影响路由表表项。	基于转发平面，不会影响路由表表项，且设备收到报文后，会先查找策略路由进行匹配转发，若匹配失败，则再查找路由表进行转发。
只能基于目的地址进行策略制定。	可基于源地址、目的地址、协议类型、报文大小等进行策略制定。
与路由协议结合使用。	需手工逐跳配置，以保证报文按策略进行转发。
常用工具：Route-Policy、Filter-Policy等。	常用工具：Traffic-Filter、Traffic-Policy、Policy-Based-Route等。

- 路由器存在两种类型的表：一个是路由表（routing-table），另一个是转发表（forwarding-table），转发表是由路由表映射过来的，策略路由直接作用于转发表，路由策略直接作用于路由表。由于转发在底层，路由在高层，所以直接作用在转发表的转发优先级比查找路由表转发的优先级高。
- 路由策略是在路由发现的时候产生作用，并根据一些规则，使用某种策略来影响路由发布、接收或路由选择的参数，从而改变路由发现的结果，从而最终改变路由表内容；策略路由是在数据包转发的时候发生作用，不改变路由表中的任何内容，它可以通过设置的规则影响数据报文的转发。

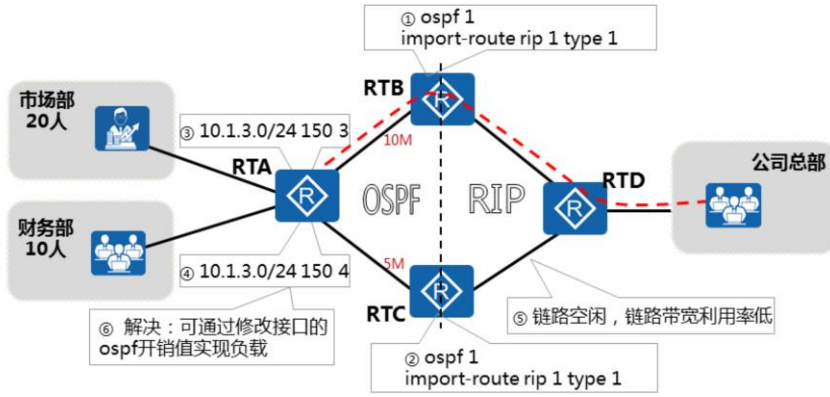
目录

1. 流量行为控制需求
2. 控制流量可达性
3. 调整网络流量路径
4. **路由引入导致的问题及解决办法**

HUAWEI

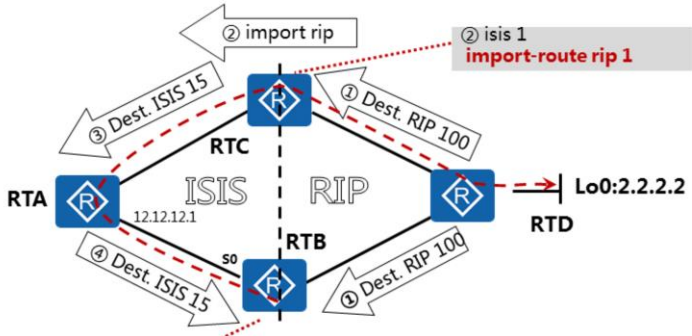
调整网络流量路径 - 多协议复杂场景

- 前文示例中描述的四台路由器都运行同一种协议，分析若运行不同协议会出现什么问题？



HUAWEI

多协议复杂场景带来的其他问题 - 次优路由

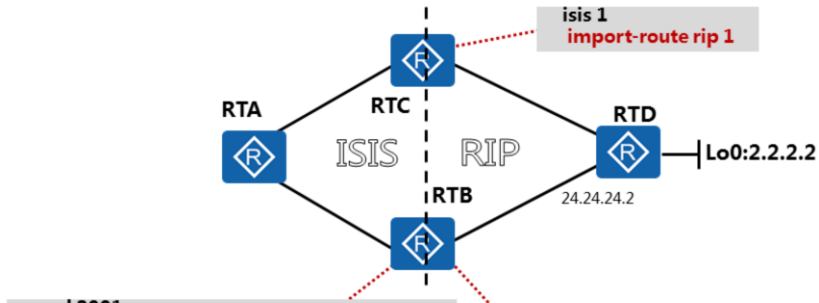


⑤ [RTB]display ip routing-table

Destination/Mask	Proto	Pre	Cost	NextHop	Interface
2.0.0.0/8	ISIS	15	84	12.12.12.1	Serial0

HUAWEI

解决方案一：利用路由过滤避免次优路由



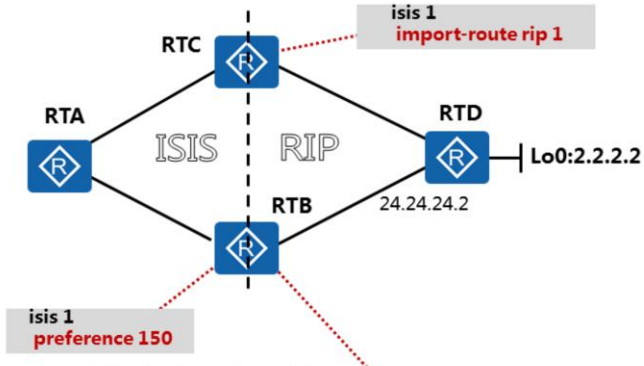

```
acl 2001
 rule 5 deny source 2.0.0.0 0.255.255.255
 rule 10 permit
isis 1
 filter-policy 2001 import
```

display ip routing-table

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

HUAWEI

解决方案二：调整协议优先级避免次优路由

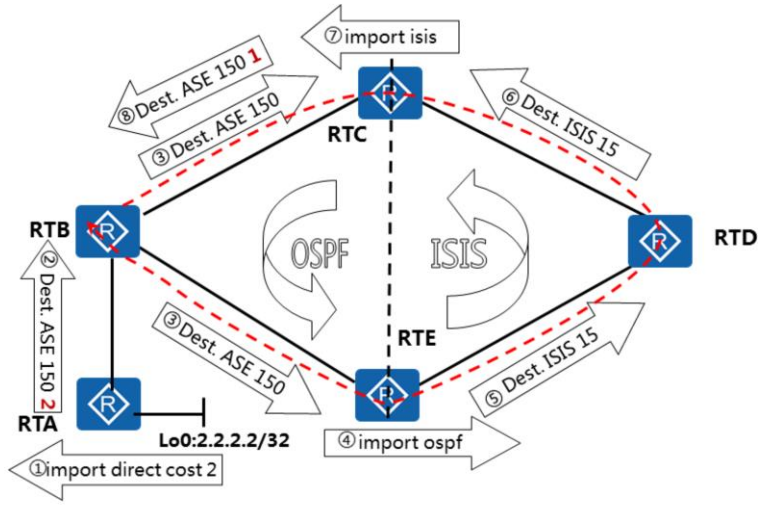


display ip routing-table

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

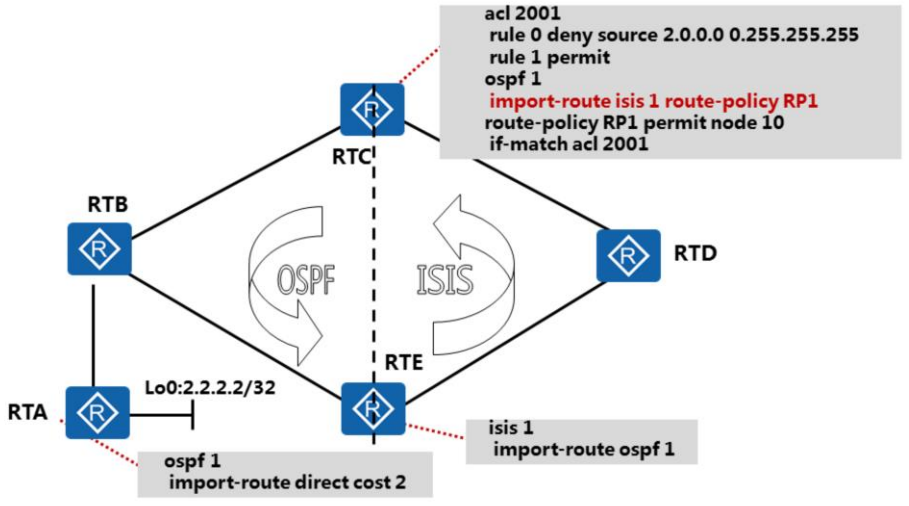
HUAWEI

多协议复杂场景带来的其他问题 - 路由环路



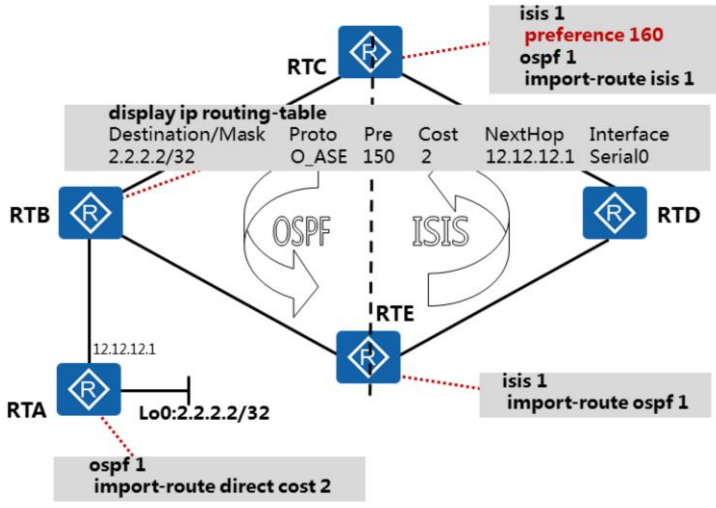
HUAWEI

解决方案一：利用路由过滤避免路由环路



HUAWEI

解决方案二：调整协议优先级避免路由环路



HUAWEI

思考题

1. IP-Prefix List可以用来过滤IP报文吗？
2. 常用调整网络流量路径的方式都包括哪些？
3. 路由引入可能会带来哪些问题？常用的解决办法包括哪些？

HUAWEI

- 答案：IP-Prefix List可以用来过滤路由信息，但不能过滤IP报文。
- 答案：常用调整网络流量路径的方式包括：路由策略和策略路由方式。
- 答案：路由引入可能会带来次优路径、路由环路等问题，常采用路由过滤、调整协议优先级方式来解决。


谢 谢
www.huawei.com

Eth-Trunk技术原理与配置

HUAWEI

前言

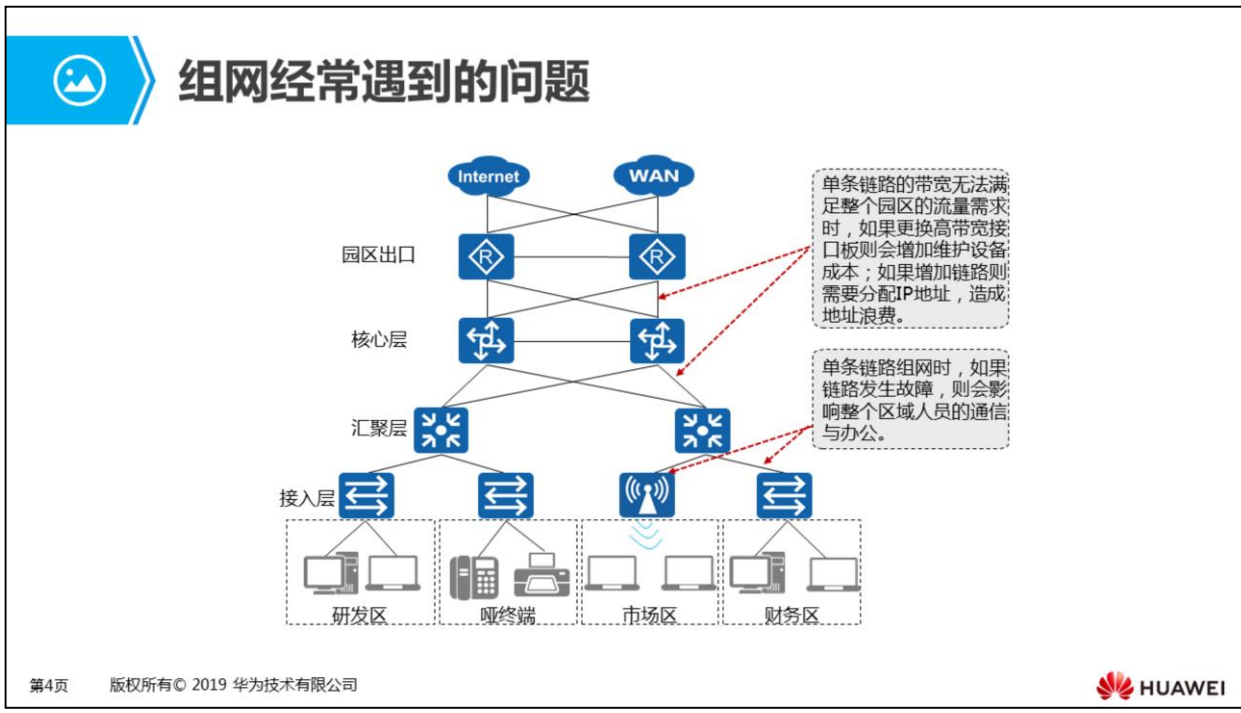
- 随着网络中部署的业务量不断增长，对于全双工点对点链路，单条物理链路的带宽已不能满足正常的业务流量需求。如果将当前接口板替换为具备更高带宽的接口板，则会浪费现有的设备资源，而且升级代价较大。如果增加设备间的链路数量，则在作为三层口使用时需要在每个接口上配置IP地址，从而导致浪费IP地址资源。
- Eth-Trunk（ 链路聚合技术 ）作为一种捆绑技术，可以把多个独立的物理接口绑定在一起作为一个大带宽的逻辑接口使用，这样既不用替换接口板也不会浪费IP地址资源。本课程我们将详细的介绍Eth-Trunk技术。

目标

- 学完本课程后，您将能够：
  - 熟悉Eth-Trunk原理
  - 掌握Eth-Trunk配置

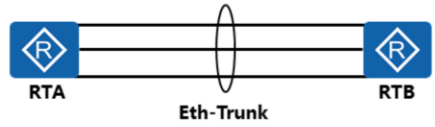
目录

1. **Eth-Trunk基本原理**
2. Eth-Trunk配置实例



- 组网中经常遇到的问题：
  - 随着网络中部署的业务量不断增长，单条物理链路的带宽已不能满足正常的业务流量需求，如果将当前接口板替换为具备更高带宽的接口板，则会浪费现有的设备资源，而且升级代价较大。如果增加设备间的链路数量，则在作为三层口使用时需要在每个接口上配置IP地址，从而导致浪费IP地址资源；
  - 单条链路的组网中没有冗余的设计，如果接入层设备上联的链路故障时，影响接入设备下联的整个区域的设备正常通信。
- 此时，可以把多个独立的物理接口绑定在一起作为一个大带宽的逻辑接口使用，即链路聚合技术，既不用替换接口板也不会浪费IP地址资源。Eth-Trunk是一种捆绑技术，将多个物理接口捆绑成一个逻辑接口，这个逻辑接口就称为Eth-Trunk接口。
- 对于Eth-Trunk接口，只有以太网接口才可以加入，下面我们将具体介绍局域网中的Eth-Trunk技术。

Eth-Trunk概念

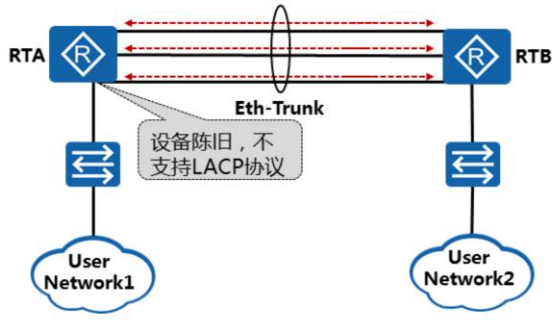


- Eth-Trunk是一种将多个以太网接口捆绑成一个逻辑接口的捆绑技术。
- Eth-Trunk链路聚合模式：
  - 手工负载分担模式；
  - LACP模式。

HUAWEI

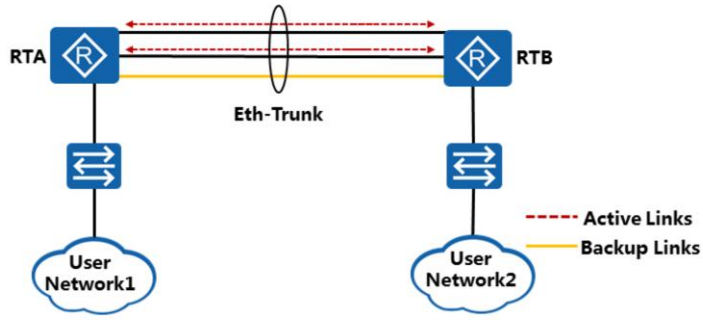
- 根据不同的链路聚合模式，Eth-Trunk接口可以实现增加带宽、负载分担等，帮助提高网络的可靠性。
- Eth-Trunk可以用于二层的链路聚合，也可以用于三层的链路聚合。缺省情况下，以太网接口工作在二层模式。如果需要配置二层Eth-Trunk接口，可以通过portswitch命令将该接口切换成二层接口；如果需要配置三层Eth-Trunk接口，可以通过undo portswitch命令将该接口切换成三层接口。

手工负载分担模式



- 当两台设备中至少有一台不支持LACP协议时，可使用手工负载分担模式的Eth-Trunk来增加设备间的带宽及可靠性。
- 在手工负载分担模式下，加入Eth-Trunk的链路都进行数据的转发。

LACP模式

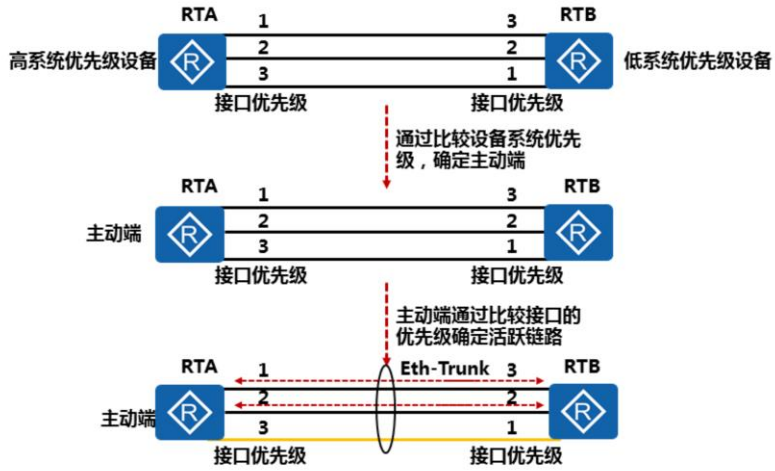


- LACP模式也称为M:N模式，其中M条链路处于活动状态转发数据，N条链路处于非活动状态作为备份链路。
- 图中设置的活跃链路数为2，即2条链路处于转发状态，1条链路处于备份状态，不转发数据，只有当活跃的链路出现故障时，备份链路才进行转发。

HUAWEI

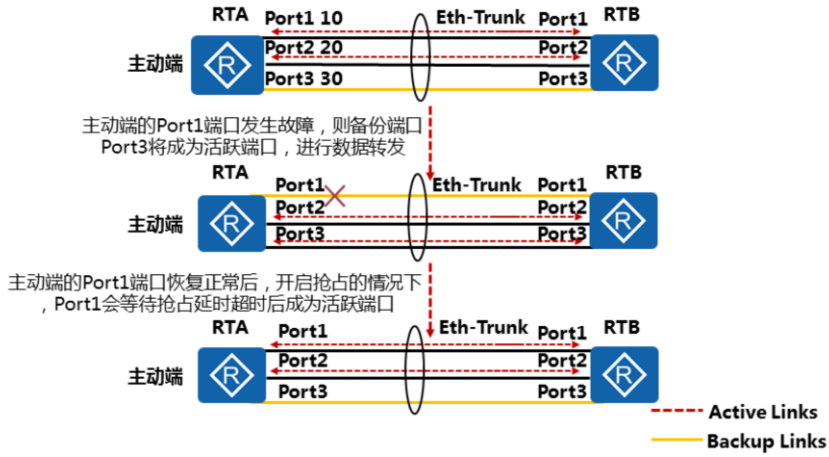
- 成员接口间M:N备份：
  - 如图所示，两台设备间有M+N（2+1）条链路，在聚合链路上转发流量时在M（2）条链路上负载分担，不在另外的N（1）条链路转发流量。此时链路的实际带宽为M（2）条链路的总和，但是能提供的最大带宽为M+N（2+1）条链路的总和；
  - 当M（2）条链路中有一条链路故障时，LACP会从N（1）条备份链路中找出一条优先级高的可用链路替换故障链路。此时链路的实际带宽还是M（2）条链路的总和，但是能提供的最大带宽就变为M+N-1（2+1-1）条链路的总和。

LACP模式活动链路的选取



- 如图所示，设备之间相连的链路数为3条，设置的最大活跃链路数为2，即2条链路处于转发状态，1条链路处于备份状态。
- 在LACP模式的Eth-Trunk中加入成员接口后，这些接口将向对端通告自己的系统优先级、MAC地址、接口优先级、接口号等信息。对端接收到这些信息后，将这些信息与自身接口所保存的信息比较以选择能够聚合的接口，双方对哪些接口能够成为活动接口达成一致，确定活动链路。
  - 在两端设备中选择系统LACP优先级较高的一端作为主动端，如果系统LACP优先级相同则选择MAC地址较小的一端作为主动端。
    - 系统LACP优先级的值越小，则优先级越高，缺省情况下，系统LACP优先级的值为32768。
  - 接口LACP优先级的值越小，则优先级越高。如果接口LACP优先级相同，接口ID（接口号）小的接口被优先选为活动接口。
    - 接口LACP优先级是为了区别同一个Eth-Trunk中的不同接口被选为活动接口的优先程度，优先级高的接口将优先被选为活动接口。

LACP模式的抢占机制



HUAWEI

- LACP抢占延时设置：
  - LACP抢占发生时，处于备用状态的链路将会等待一段时间后再切换到转发状态，这就是抢占延时。配置抢占延时是为了避免由于某些链路状态频繁变化而导致Eth-Trunk数据传输不稳定的情况。
  - 如图所示，Port1由于链路故障切换为非活动接口，此后该链路又恢复了正常。若系统使能了LACP抢占并配置了抢占延时，Port1重新切换回活动状态就需要经过抢占延时的时间。
- 开启抢占功能的场景：
  - Port1接口出现故障而后又恢复正常。当接口Port1出现故障时被Port3所取代，如果在Eth-Trunk接口下未使能抢占，则故障恢复时Port1将处于备份状态；如果使能了LACP抢占，当Port1故障恢复时，由于接口优先级比Port3高，将重新成为活动接口，Port3再次成为备份接口。
  - 如果希望Port3接口替换Port1、Port2中的一个接口成为活动接口，可以将Port3的接口LACP优先级调高，但前提条件是已经使能了LACP抢占功能。如果没有使能LACP抢占功能，即使将备份接口的优先级调整为高于当前活动接口的优先级，系统也不会进行重新选择活动接口的过程，也不切换活动接口。

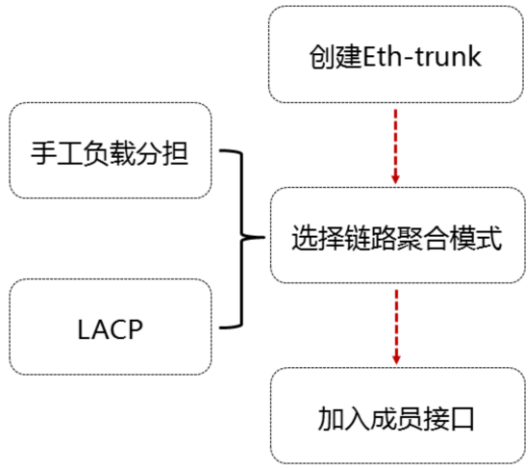
Eth-Trunk接口负载分担

- Eth-Trunk接口进行负载分担时，可以选择IP地址或者包作为负载分担的散列依据；同时还可以设置成员接口的负载分担权重。
- Eth-Trunk接口中，某成员接口的权重值占所有成员接口负载分担权重之和的比例越大，该成员接口承担的负载就越大。

接口负载分担	特点
逐流负载分担	当报文的源IP地址、目的IP地址都相同或者报文的源MAC地址、目的MAC地址都相同时，这些报文从同一条成员链路上通过。
逐包负载分担	以报文为单位分别从不同的成员链路上发送。

- 配置负载分担方式
  - 执行命令system-view，进入系统视图。
  - 执行命令interface eth-trunk trunk-id，进入Eth-Trunk接口视图。
  - 执行命令load-balance { ip | packet-all }，配置Eth-Trunk接口的散列依据。
  - 缺省情况下，当Eth-Trunk接口根据IP进行散列。
  - 说明：
    - 基于IP的散列算法能保证包顺序，但不能保证带宽利用率。
    - 基于包的散列算法能保证带宽利用率，但不能保证包的顺序。
- 配置负载分担权重
  - 执行命令system-view，进入系统视图。
  - 执行命令interface interface-type interface-number，进入以太网接口视图。
  - 执行命令distribute-weight weight-value，配置Eth-Trunk成员接口的负载分担权重。
  - 缺省情况下，成员接口的负载分担权重为1。

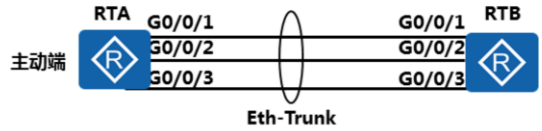
Eth-Trunk接口配置流程



  HUAWEI

- 将成员接口加入Eth-Trunk时，需要注意以下问题：
  - 成员接口不能有IP地址等三层配置项，也不可以配置任何业务；
  - 成员接口不能配置静态MAC地址；
  - Eth-Trunk接口不能嵌套，即成员接口不能是Eth-Trunk；
  - 一个以太网接口只能加入到一个Eth-Trunk接口，如果需要加入其他Eth-Trunk接口，必须先退出原来的Eth-Trunk接口；
  - 如果本地设备使用了Eth-Trunk，与成员接口直连的对端接口也必须捆绑为Eth-Trunk接口，两端才能正常通信；
  - Eth-Trunk有两种工作模式：二层工作模式和三层工作模式。Eth-Trunk的工作模式不影响成员链路的加入，例如，以太网接口既可以加入二层模式的Eth-Trunk，也可以加入三层模式的Eth-Trunk。

配置手工负载分担模式



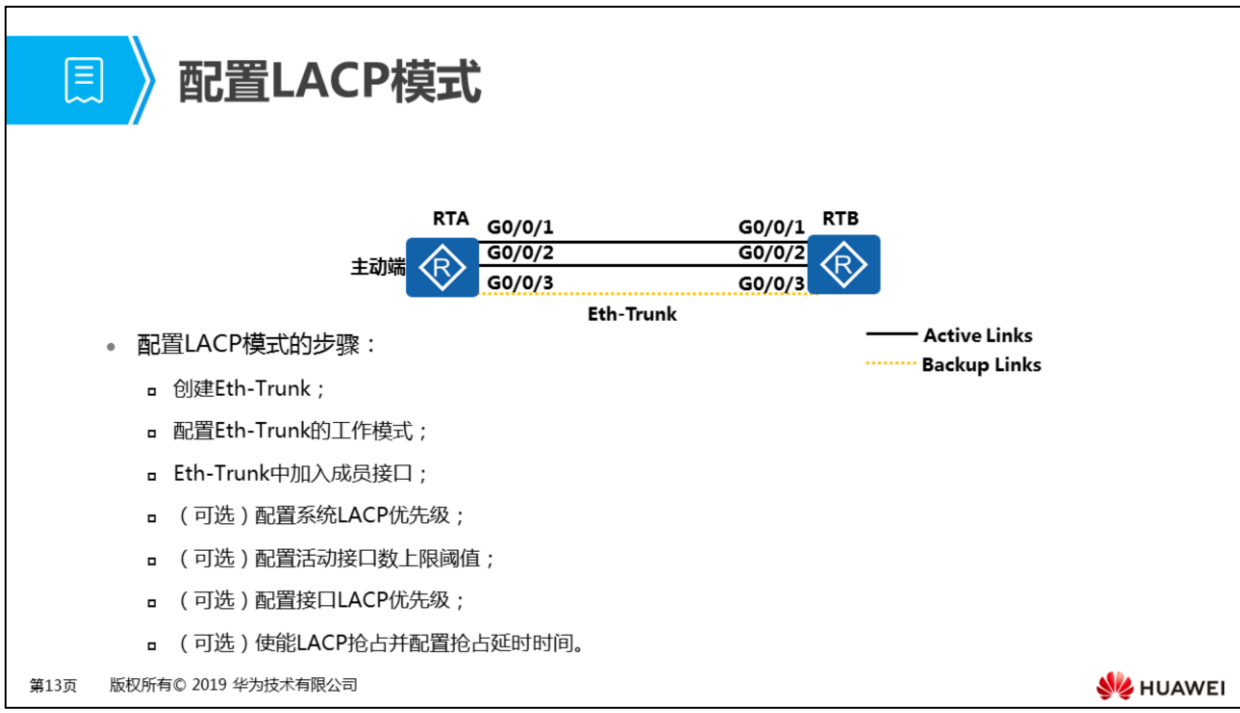
- 配置手工负载分担模式的步骤：
  - 创建Eth-Trunk；
  - 配置Eth-Trunk的工作模式；
  - Eth-Trunk中加入成员接口。

HUAWEI

- 创建手工负载分担模式Eth-Trunk：
  - 执行命令system-view，进入系统视图。
  - 执行命令interface Eth-Trunk trunk-id，创建Eth-Trunk接口并进入Eth-Trunk接口视图。
  - （可选）执行命令portswitch，将Eth-Trunk接口切换为二层模式。
- 配置Eth-Trunk的工作模式：
  - 执行命令mode manual load-balance，配置当前Eth-Trunk工作模式为手工负载分担模式。
    - 缺省情况下，Eth-Trunk的工作模式为手工负载分担模式。
  - Eth-Trunk中加入成员接口：
    - 在Eth-Trunk接口视图下：
    - 执行interface eth-trunk trunk-id命令，进入Eth-Trunk接口视图。
    - 执行以下任一个步骤，添加Eth-Trunk成员接口。

      执行trunkport interface-type { interface-number1 [to interface-number2] } &<1-16>命令，批量增加成员接口。

      执行trunkport interface-type interface-number命令，增加一个成员接口。
  - 在成员接口视图下：
    - 执行interface { ethernet | gigabitethernet } interface-number命令，进入要捆绑到此Eth-Trunk的成员接口的接口视图。
    - 执行eth-trunk trunk-id命令，将当前接口加入Eth-Trunk。



- 创建LACP模式Eth-Trunk：
  - 执行**system-view**命令，进入系统视图。
  - 执行**interface eth-trunk** *trunk-id*命令，创建Eth-Trunk。
  - （可选）执行命令**portswitch**，将Eth-Trunk接口切换为二层模式。
- 配置Eth-Trunk的工作模式：
  - 执行命令**interface eth-trunk** *trunk-id*，进入Eth-Trunk接口视图。
  - 执行命令**mode lacp-static**，配置Eth-Trunk的工作模式为LACP模式。
- Eth-Trunk中加入成员接口：
  - 在Eth-Trunk接口视图下：
    - 执行**interface eth-trunk** *trunk-id*命令，进入Eth-Trunk接口视图。
    - 执行以下任一个步骤，添加Eth-Trunk成员接口。
      - ✓ 执行**trunkport** *interface-type* { *interface-number1* [**to** *interface-number2*] } &<1-16>命令，批量增加成员接口。
      - ✓ 执行**trunkport** *interface-type interface-number*命令，增加一个成员接口。

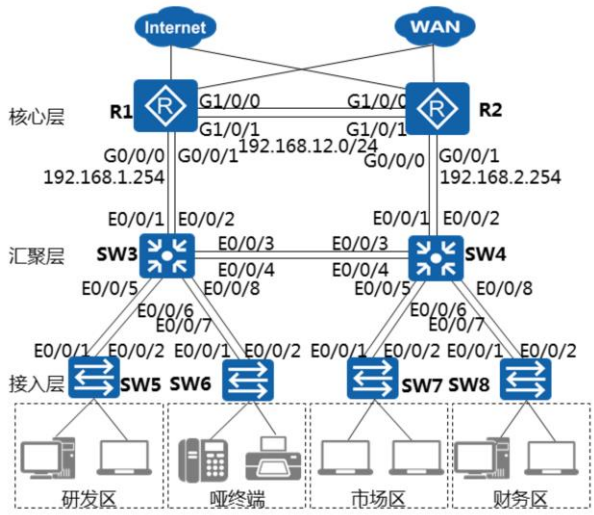
- 在成员接口视图下：
- 执行interface { ethernet | gigabitethernet } interface-number命令，进入要捆绑到此Eth-Trunk的成员接口的接口视图。
- 执行eth-trunk trunk-id命令，将当前接口加入Eth-Trunk。
- （可选）配置系统LACP优先级：
- 执行命令lacp priority priority，配置当前路由器的系统LACP优先级。
- （可选）配置活动接口数上限阈值：
- 执行命令interface eth-trunk trunk-id，进入Eth-Trunk接口视图。
- 执行命令max active-linknumber link-number，配置活动接口数上限阈值。
- （可选）配置接口LACP优先级：
- 执行命令interface interface-type interface-number，进入接口视图。
- 执行命令lacp priority priority，配置当前接口的LACP优先级。
- （可选）使能LACP抢占并配置抢占等待时间：
- 执行命令interface eth-trunk trunk-id，进入Eth-Trunk接口视图。
- 执行命令lacp preempt enable，使能当前Eth-Trunk接口的LACP抢占功能。
- 执行命令lacp preempt delay delay-time，配置当前Eth-Trunk接口的LACP抢占等待时间。

目录

1. Eth-Trunk基本原理
2. **Eth-Trunk配置实例**

HUAWEI

Eth-trunk配置需求



HUAWEI

- 如图是一个园区的组网拓扑，为了提高网络的可靠性，需要在各层设备之间采用链路聚合技术。其中核心层的设备需要配置IP地址，作为内网的网关；汇聚层与接入层的设备通过二层实现通信。

核心层设备配置

- 以核心层R1路由器为例说明配置。
  - 创建Eth-Trunk接口并配置地址：

```
interface Eth-Trunk1
 undo portswitch        //将接口转换为三层接口
 description  "Core-R1 to Aggregate-SW3 " //描述信息，便于管理员了解接口对端所连接的设备
 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
```

  - 将物理接口添加入Eth-Trunk中：

```
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
```

汇聚层设备配置 (1)

- 以汇聚层SW3交换机为例说明配置。
  - 创建Eth-Trunk接口，因为汇聚层设备使用二层互联，所以无需配置地址：

```
interface Eth-Trunk1
 description “Aggregate-SW3 to Core-R1 “
//描述信息，便于管理员了解接口对端所连接的设备
#
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 “
```

汇聚层设备配置 (2)

- 以汇聚层SW3台交换机为例说明配置。
  - 将物理接口添加入Eth-Trunk中 ：

```
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
```

接入层设备配置 (1)

- 以接入层SW5交换机为例说明配置。
  - 创建Eth-Trunk接口，因为接入层设备使用二层互联，所以无需配置地址：

```
interface Eth-Trunk1
 description "Access-SW5 to Aggregate-SW3"
//描述信息，便于管理员了解接口对端所连接的设备
```

  - 将物理接口添加入Eth-Trunk中 ：

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

接入层设备配置 (2)

- 完成上述配置，使用以下命令查看配置的Eth-Trunk接口信息：

```
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
```

  - 查看详细信息使用命令：display interface Eth-Trunk。

思考题

1. Eth-Trunk链路聚合模式为以下哪几种？( )

   A.手工负载分担模式　　B.LACP模式

   C.手工LACP模式　　D.动态LACP模式

2. 在LACP模式下，默认的系统优先级为？( )

   A.1　　B.4096

   C.32768　　D.65535

- 答案：AB。
- 答案：C。


谢 谢
www.huawei.com

交换机高级特性简介

HUAWEI

前言

- MUX VLAN（Multiplex VLAN）提供了一种通过VLAN进行网络资源控制的机制。通过MUX VLAN提供的二层流量隔离的机制可以实现企业内部员工之间互相通信，而企业外来访客之间的互访是隔离的。
- 为了实现报文之间的二层隔离，用户可以将不同的端口加入不同的VLAN，但这样会浪费有限的VLAN资源。采用端口隔离功能，可以实现同一VLAN内端口之间的隔离。端口隔离功能为用户提供了更安全、更灵活的组网方案。
- 在安全性要求较高的网络中，交换机可以开启端口安全功能，禁止非法MAC地址设备接入网络；当学习到的MAC地址数量达到上限后不再学习新的MAC地址，只允许学习到MAC地址的设备通信。

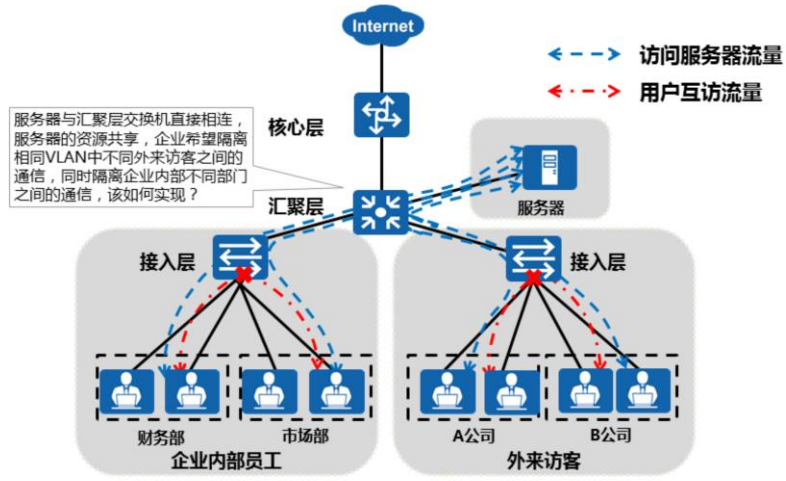
目标

- 学完本课程后，您将能够：
  - 掌握MUX VLAN应用场景及配置
  - 掌握端口隔离应用场景及配置
  - 掌握端口安全应用场景及配置

目录

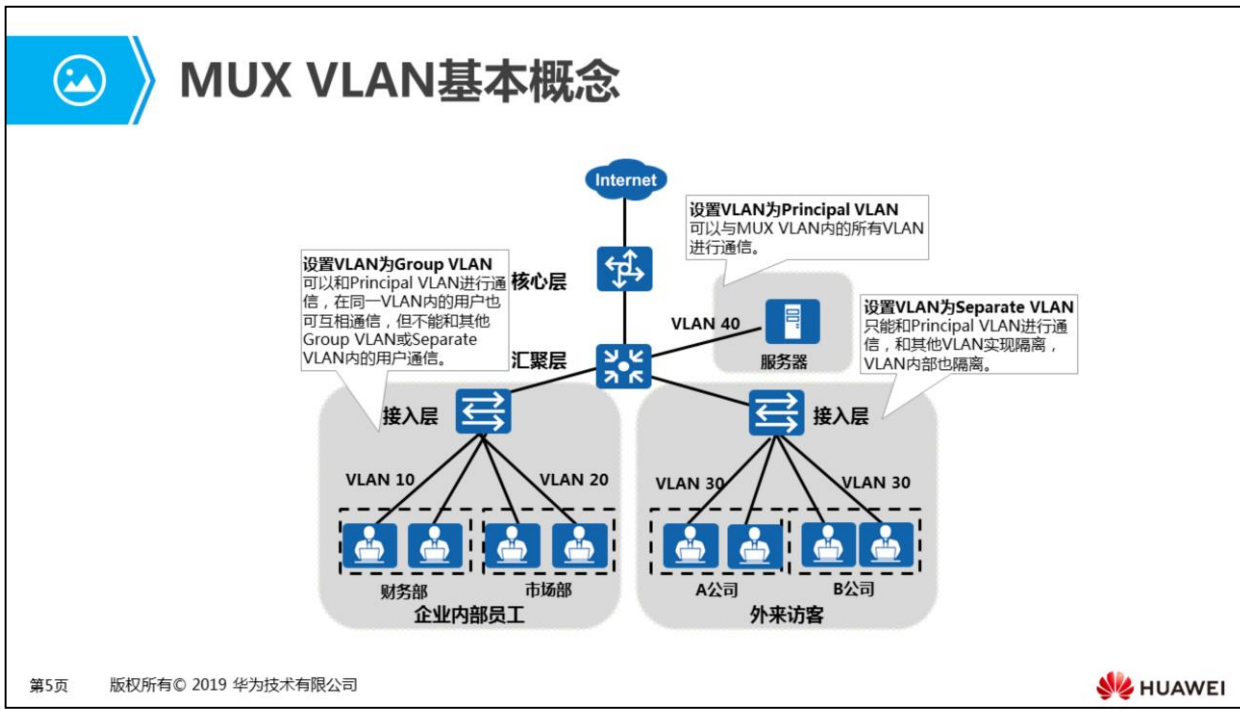
1. **MUX VLAN**
2. 端口隔离
3. 端口安全

MUX VLAN应用场景

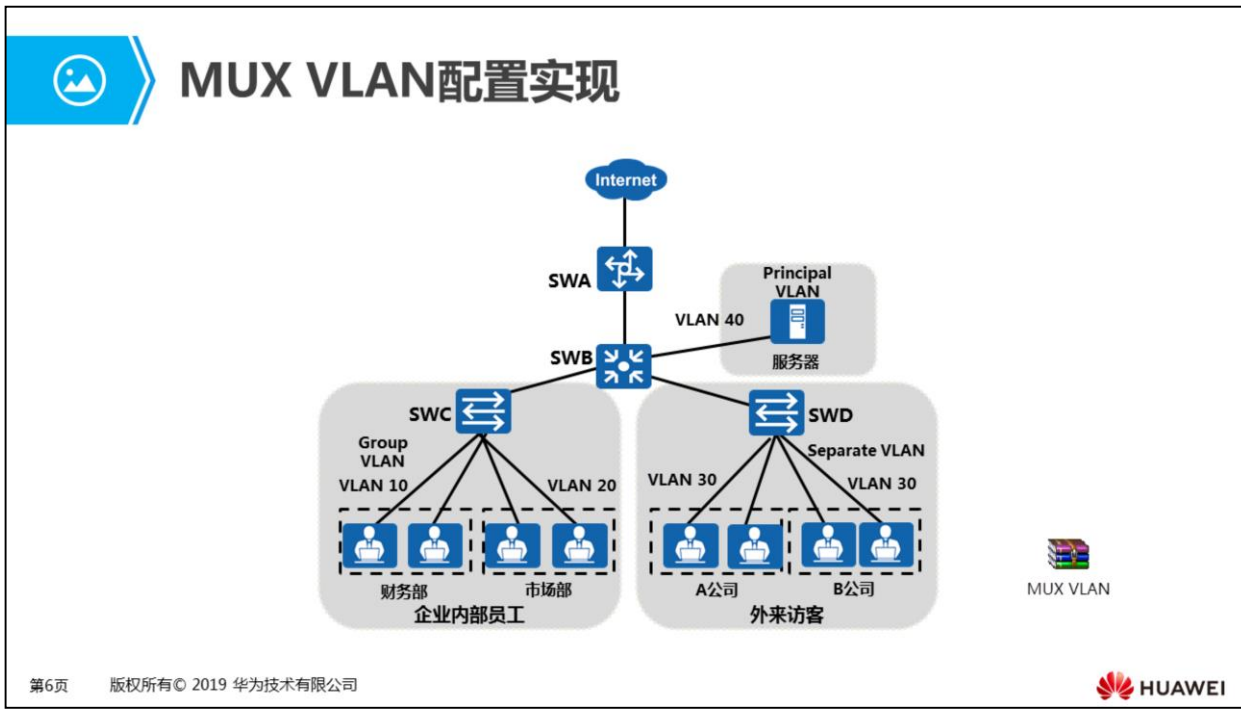


HUAWEI

- 如图所示，服务器与汇聚层交换机相连，为了实现所有用户都可访问企业服务器，可通过配置VLAN间通信来实现。
- 对于企业来说，希望企业内部员工之间可以互相访问，而企业外来访客之间是隔离的，可通过配置每个访客使用不同的VLAN来实现。但如果企业拥有大量的外来访客员工，此时不但需要耗费大量的VLAN ID，还增加了网络维护的难度。
- MUX VLAN提供的二层流量隔离的机制可以实现企业内部员工之间互相通信，而企业外来访客之间的互访是隔离的。



- MUX VLAN的划分：
  - 主VLAN（Principal VLAN）：可以与MUX VLAN内的所有VLAN进行通信。
  - 隔离型从VLAN（Separate VLAN）：只能和Principal VLAN进行通信，和其他类型的VLAN完全隔离，Separate VLAN内部也完全隔离。
  - 互通型从VLAN（Group VLAN）：可以和Principal VLAN进行通信，在同一Group VLAN内的用户也可互相通信，但不能和其他Group VLAN或Separate VLAN内的用户通信的VLAN。
- 如图所示，根据MUX VLAN特性，解决方案如下：
  - 企业管理员可以将服务器划分到Principal VLAN。
  - MUX VLAN技术中只能将一个VLAN设置为Separate VLAN，所以可以将外来访客划分到Separate VLAN。
  - 由于可以将多个VLAN设置为Group VLAN，所以可以将企业员工划分到Group VLAN，企业内部不同部门之间通过划分到不同的VLAN进行隔离。
  - 这样就能够实现：
    - 企业外来访客、企业员工都能够访问企业服务器。
    - 企业员工部门内部可以通信，而企业员工部门之间不能通信。
    - 企业外来访客间不能通信、外来访客和企业员工之间不能互访。



- 如图所示，希望实现企业外来访客、企业员工都能够访问企业服务器，而企业同部门员工可以通信，不同部门员工不能通信；企业外来访客间不能通信；企业外来访客和企业员工之间不能互访。
  - 将企业服务器划分到Principal VLAN，Principal VLAN为VLAN 40；
  - 企业外来访客划分到Separate VLAN，Separate VLAN为VLAN 30；
  - 企业员工划分到Group VLAN，Group VLAN为VLAN 10与VLAN 20，VLAN 10分配给财务部，VLAN 20分配给市场部，各部门之间二层隔离。
- SWB配置：
  - 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 //将VLAN 40设置为Principal VLAN
- subordinate separate 30 //将VLAN 30设置为Separate VLAN
- subordinate group 10 20 //将VLAN 10与VLAN 20设置为Group VLAN
- #
- 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 //在接口下开启MUX VLAN功能
- SWC配置：
- sysname SWC
- #
- vlan batch 10 20 30 40
- #
- vlan 10
- description Financial VLAN
- vlan 20
- description Marketing VLAN

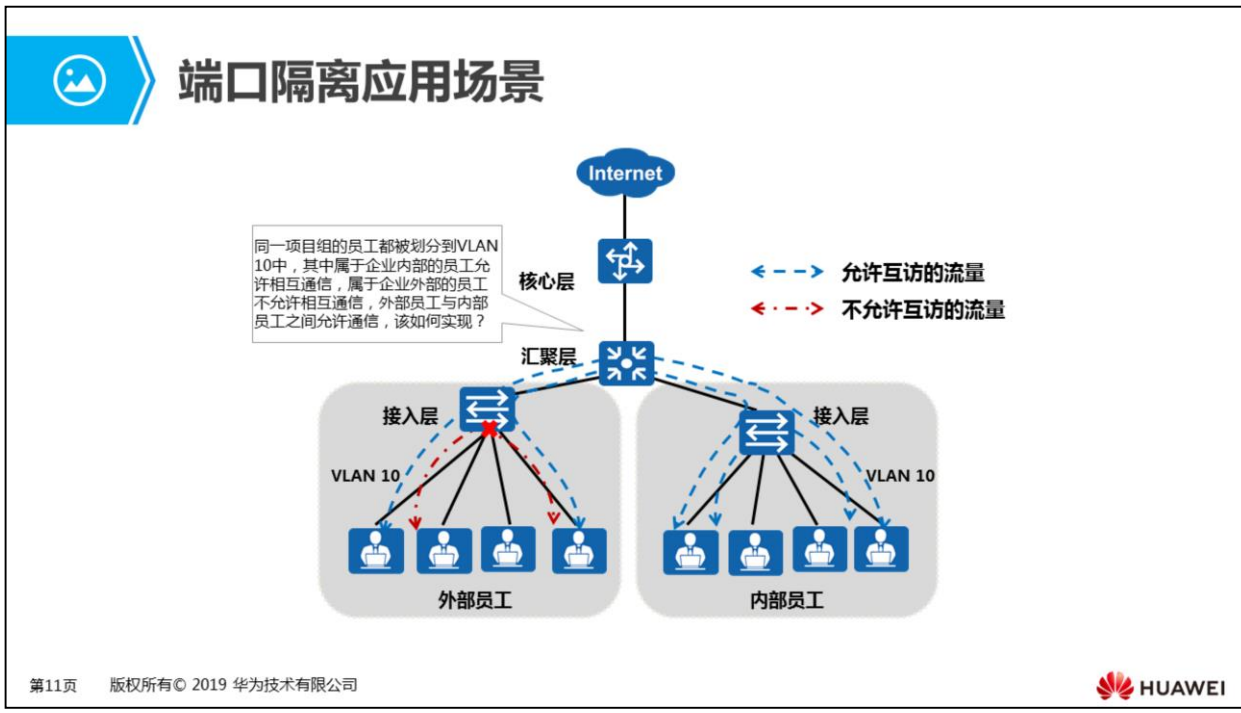
- vlan 30
- description Cilent VLAN
- vlan 40
- description Principal VLAN
- mux-vlan
- subordinate separate 30
- subordinate group 10 20
- #
- interface GigabitEthernet0/0/1
- port link-type trunk
- port trunk allow-pass vlan 10 20 30 40
- #
- interface GigabitEthernet0/0/2
- port link-type access
- 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配置：
- 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

目录

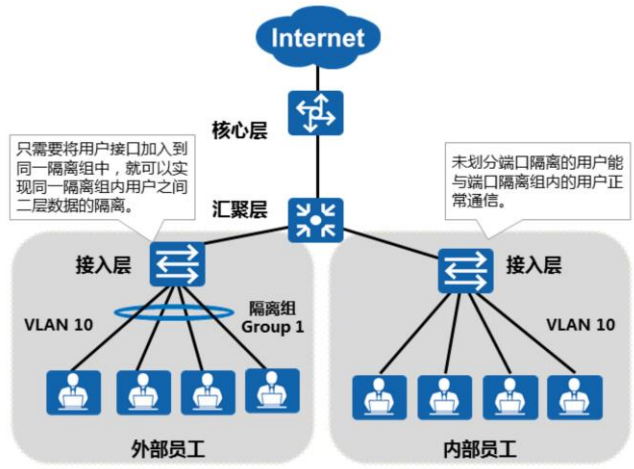
1. MUX VLAN
2. **端口隔离**
3. 端口安全

HUAWEI



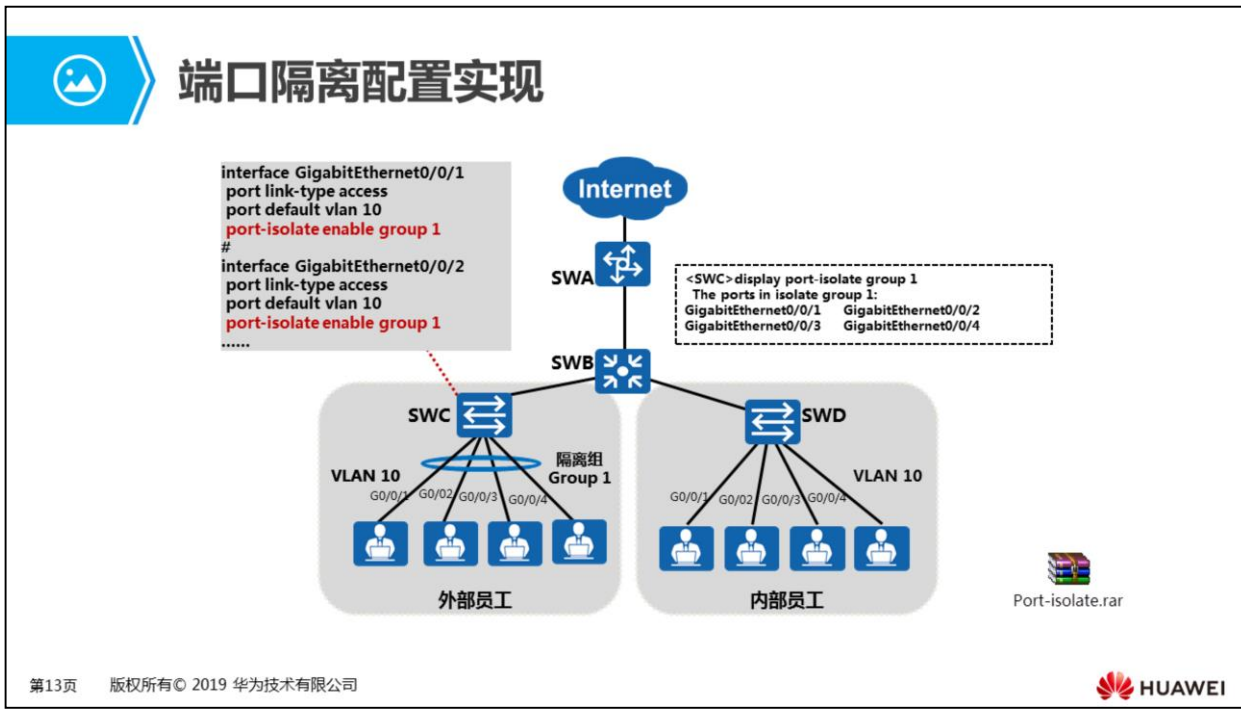
- 为了实现用户之间的二层隔离，可以将不同的用户加入不同的VLAN，但这样会浪费有限的VLAN资源。采用端口隔离功能，可以实现同一VLAN内端口之间的隔离。用户只需要将端口加入到同一隔离组中，就可以实现隔离组内端口之间二层数据的隔离。端口隔离功能为用户提供了更安全、更灵活的组网方案。

端口隔离基本概念



HUAWEI

- 如图所示，同一端口隔离组内的用户不能进行二层的通信，但是不同端口隔离组内的用户可以进行正常通行；未划分端口隔离的用户也能与端口隔离组内的用户正常通信。
- 端口隔离分为二层隔离三层互通和二层三层都隔离两种模式：
  - 如果用户希望隔离同一VLAN内的广播报文，但是不同端口下的用户还可以进行三层通信，则可以将隔离模式设置为二层隔离三层互通。
  - 如果用户希望同一VLAN不同端口下用户彻底无法通信，则可以将隔离模式配置为二层三层均隔离。
- 配置注意事项：
  - S系列交换机均支持配置二层隔离三层互通模式。
  - S系列框式交换机均支持二层三层都隔离模式，S系列盒式交换机仅V100R006C05版本仅S2700SI、S2700EI不支持二层三层都隔离模式，V100R002及后续版本S1720、S2720、S2750EI、S5700LI、S5700S-LI不支持二层三层都隔离模式。
  - 如果不是特殊情况要求，建议用户不要将上行口和下行口加入到同一端口隔离组中，否则上行口和下行口之间不能相互通信。



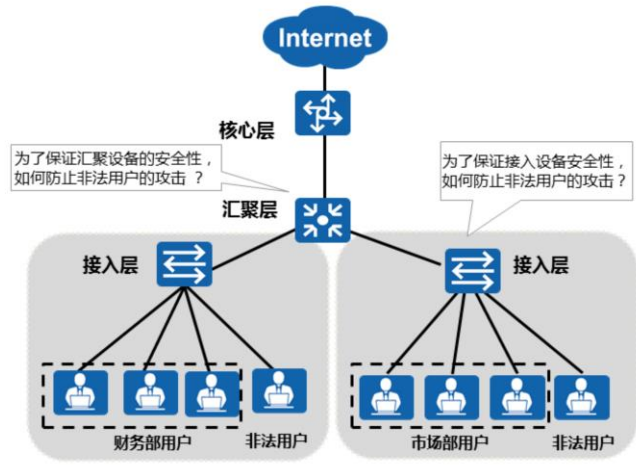
- 如图所示，同项目组的员工都被划分到VLAN 10中，其中属于企业内部的员工允许相互通信，属于企业外部的员工不允许相互通信，外部员工与内部员工之间允许通信。
- 配置命令：
  - port-isolate enable命令用来使能端口隔离功能，默认将端口划入隔离组group 1。
  - 如果希望创建新的group组，使用命令port-isolate enable group后面接所要创建的隔离组组号。
  - 可以在系统视图下执行port-isolate mode all命令配置隔离模式为二层三层都隔离。
- 查看命令：
  - 使用display port-isolate group all命令可以查看所有创建的隔离组情况。
  - 使用display port-isolate group X（组号）命令可以查看具体的某一个隔离组接口情况。

目录

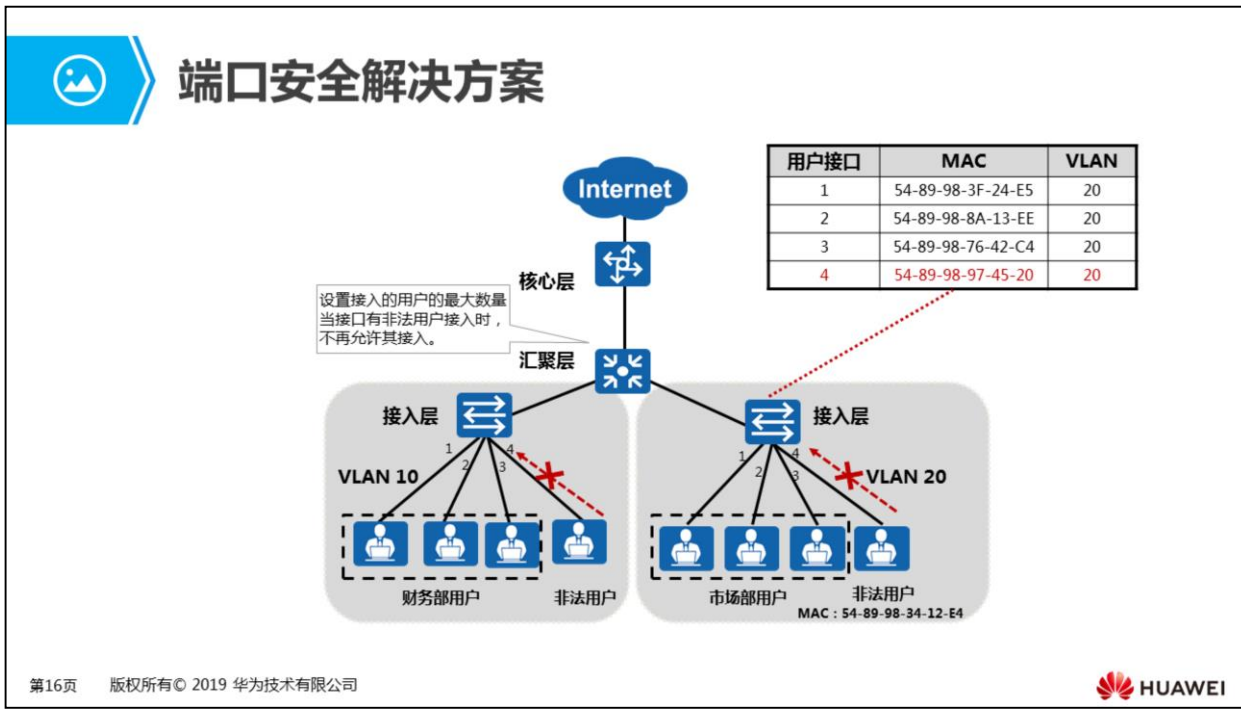
1. MUX VLAN
2. 端口隔离
3. **端口安全**

HUAWEI

端口安全应用场景



- 如图所示，当网络中存在非法用户时，可以使用端口安全技术保证网络的安全。
- 端口安全经常使用在下列场景中：
  - 应用在接入层设备，通过配置端口安全可以防止仿冒用户从其他端口攻击。
  - 应用在汇聚层设备，通过配置端口安全可以控制接入用户的数量。



- 在对接入用户的安全性要求较高的网络中，可以配置端口安全功能，将接口学习到的MAC地址转换为安全MAC地址，接口学习的最大MAC数量达到上限后不再学习新的MAC地址，只允许学习到MAC地址的设备通信。这样可以阻止其他非信任用户通过本接口和交换机通信，提高设备与网络的安全性。
- 如图所示，解决方案如下：
  - 接入层交换机的每个接口都开启端口安全功能，并绑定接入用户的MAC地址与VLAN信息，当有非法用户通过已配置端口安全的接口接入网络时，交换机会查找对应的MAC地址表，发现非法用户的MAC地址与表中的不符，将数据包丢弃。
  - 汇聚层交换机开启端口安全功能，并设置每个接口可学习到的最大MAC地址数，当学习到的MAC地址数达到上限时，其他的MAC地址的数据包将被丢弃。

端口安全类型

- 端口安全（Port Security）通过将接口学习到的动态MAC地址转换为安全MAC地址（包括安全动态MAC、安全静态MAC和Sticky MAC）阻止非法用户通过本接口和交换机通信，从而增强设备的安全性。

类型	定义	特点
安全动态MAC地址	使能端口安全而未使能Sticky MAC功能时转换的MAC地址。	设备重启后表项会丢失，需要重新学习。缺省情况下不会被老化，只有在配置安全MAC的老化时间后才可以被老化。
安全静态MAC地址	使能端口安全时手工配置的静态MAC地址。	不会被老化，手动保存配置后重启设备不会丢失。
Sticky MAC地址	使能端口安全后又同时使能Sticky MAC功能后转换得到的MAC地址。	不会被老化，手动保存配置后重启设备不会丢失。

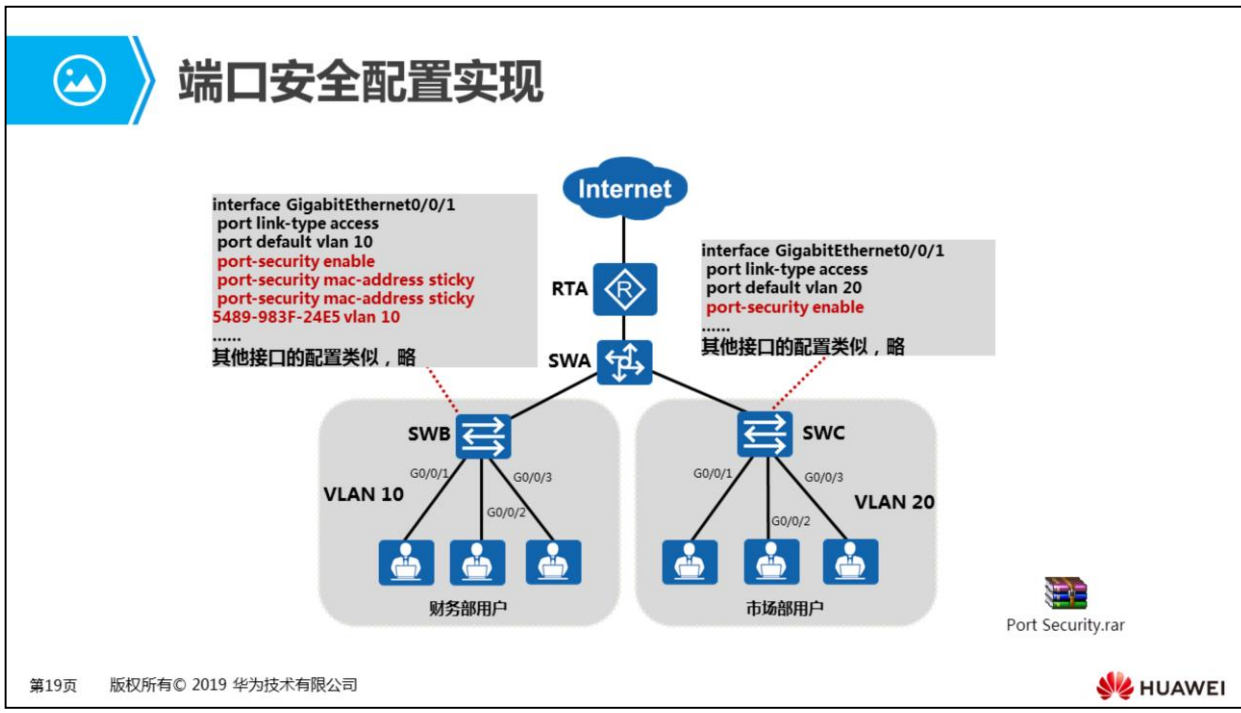
- 说明：
  - 接口使能端口安全功能时，接口上之前学习到的动态MAC地址表项将被删除，之后学习到的MAC地址将变为安全动态MAC地址。
  - 接口使能Sticky MAC功能时，接口上的安全动态MAC地址表项将转化为Sticky MAC地址，之后学习到的MAC地址也变为Sticky MAC地址。
  - 接口去使能端口安全功能时，接口上的安全动态MAC地址将被删除，重新学习动态MAC地址。
  - 接口去使能Sticky MAC功能时，接口上的Sticky MAC地址会转换为安全动态MAC地址。

端口安全限制动作

- 超过安全MAC地址限制数后的动作：

动作	实现说明
restrict	丢弃源MAC地址不存在的报文并上报告警。推荐使用restrict动作。
protect	只丢弃源MAC地址不存在的报文，不上报告警。
shutdown	接口状态被置为error-down，并上报告警。默认情况下，接口关闭后不会自动恢复，只能由网络管理人员在接口视图下使用**restart**命令重启接口进行恢复。

- 接口上安全MAC地址数达到限制后，如果收到源MAC地址不存在的报文，端口安全则认为有非法用户攻击，就会根据配置的动作对接口做保护处理。缺省情况下，保护动作是restrict。



- 如图所示，园区网络要求保障接入用户的安全性。财务部人员流动性较低，可以使用端口安全技术静态绑定接入用户的MAC与VLAN信息；市场部的人员流动性较高，使用端口安全技术的动态MAC地址学习保证接入用户的合法性。
- 命令解释：
  - 执行命令interface interface-type interface-number，进入接口视图。
  - 执行命令port-security enable，使能端口安全功能。
    - 缺省情况下，未使能端口安全功能。
  - 执行命令port-security mac-address sticky，使能接口Sticky MAC功能。
    - 缺省情况下，接口未使能Sticky MAC功能。
  - 执行命令port-security max-mac-num max-number，配置接口Sticky MAC学习限制数量。
    - 使能接口Sticky MAC功能后，缺省情况下，接口学习的MAC地址限制数量为1。
  - （可选）执行命令port-security protect-action { protect | restrict | shutdown }，配置端口安全保护动作。
    - 缺省情况下，端口安全保护动作为restrict。
  - （可选）执行命令port-security mac-address sticky mac-address vlan vlan-id，手动配置一条sticky-mac表项。

端口安全配置验证

- 在SWB上使用命令查看绑定的MAC地址表：

```
<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    -
-------------------------------------------------------------------------------
```

- 在SWC上使用命令查看动态学习到的MAC地址表：

```
<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  -
-------------------------------------------------------------------------------
```

思考题

1. 在MUX VLAN中，可以与所有VLAN进行通信的是下列哪个选项？（　　）

   A. Principal VLAN　　B. Separate VLAN

   C. Group VLAN　　D. Subordinate VLAN

2. 端口安全技术中安全MAC地址类型有以下哪几种？（　　）

   A. 安全动态MAC地址　　B. 安全静态MAC地址

   C. Sticky MAC地址　　D. Protect MAC地址

- 答案：A。
- 答案：ABC。


谢 谢
www.huawei.com

RSTP协议原理与配置

HUAWEI

前言

- STP协议虽然能够解决环路问题，但是由于网络拓扑收敛较慢，影响了用户通信质量，而且如果网络中的拓扑结构频繁变化，网络也会随之频繁失去连通性，从而导致用户通信频繁中断，这也是用户无法忍受的。
- 由于STP的不足，IEEE于2001年发布的802.1w标准定义了RSTP。RSTP在STP基础上进行了诸多改进优化，使得协议更加清晰、规范，同时也实现了二层网络拓扑的快速收敛。那STP协议具体存在哪些不足呢？RSTP协议是如何在STP协议的基础上进行优化的呢？

目标

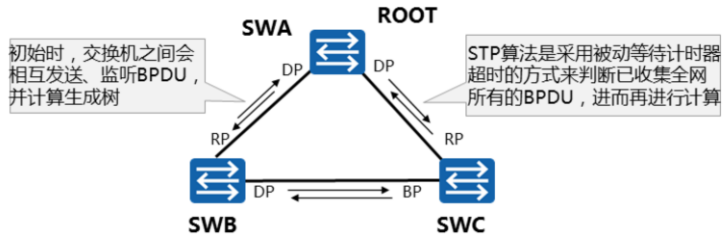
- 学完本课程后，您将能够：
  - 掌握RSTP协议的工作原理
  - 熟悉RSTP与STP的主要异同点
  - 了解RSTP的典型应用场景配置

目录

1. **STP的不足**
2. RSTP对STP的改进
3. RSTP配置实例

问题一：设备运行STP初始化场景

- STP从初始状态到完全收敛至少需经过30s：

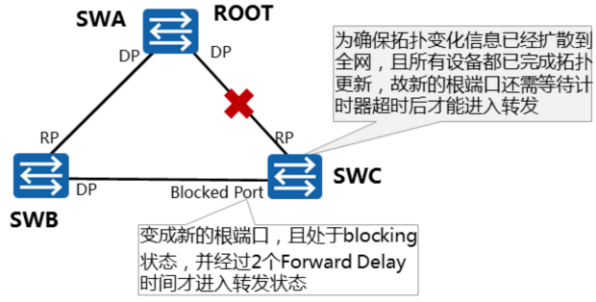


listening → 15s → learning：STP为避免临时环路，必须等待足够长的时间（即确保BPDU能同步发送至全网各节点）确保全网的端口状态全部确定，才能进入转发。

learning → 15s → forwarding：在STP进入转发前还需根据收到的用户流量构建MAC地址表，仍需等待计时器超时才能进入转发。

问题二：交换机有BP端口，RP端口down掉场景

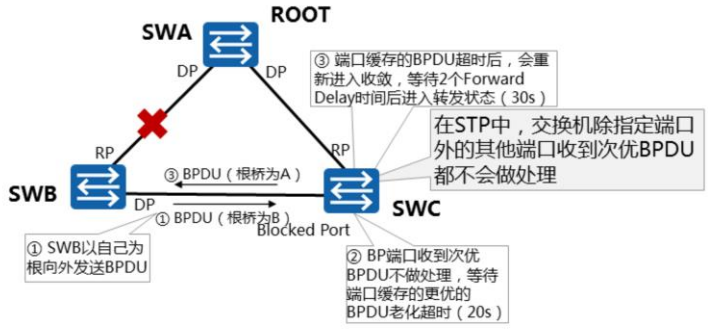
- SWC与SWA的直连链路down掉，其BP端口切换成RP端口并进入转发状态至少需要经过30s：



HUAWEI

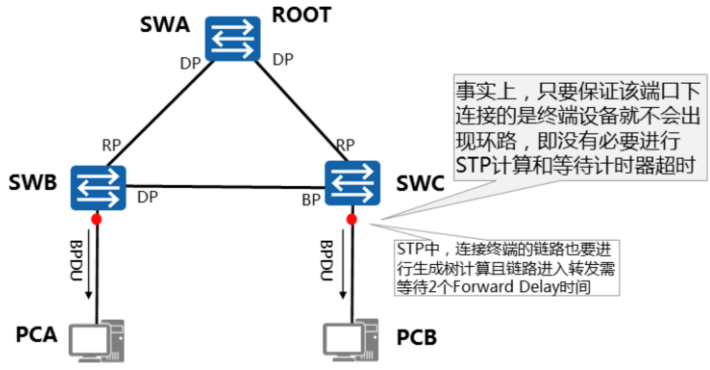
问题三：交换机无BP端口，RP端口down掉场景

- SWB与SWA的直连链路down掉，则SWC的BP端口切换成DP端口并进入转发状态大约需要50s：



问题四：运行STP的交换机连接用户终端的场景

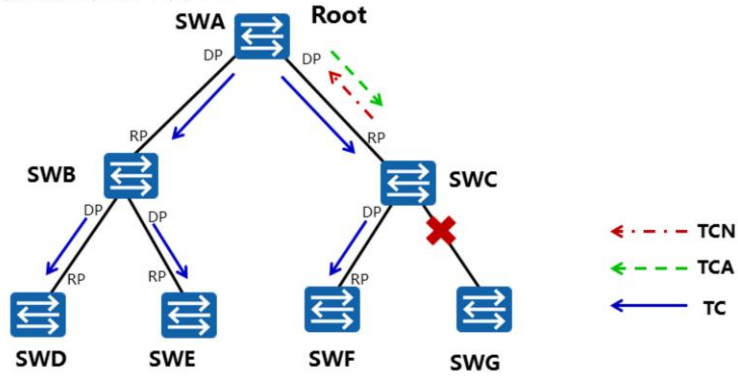
- 交换机连接终端的链路进入转发需要经过30s：



HUAWEI

问题五：STP的拓扑变更机制

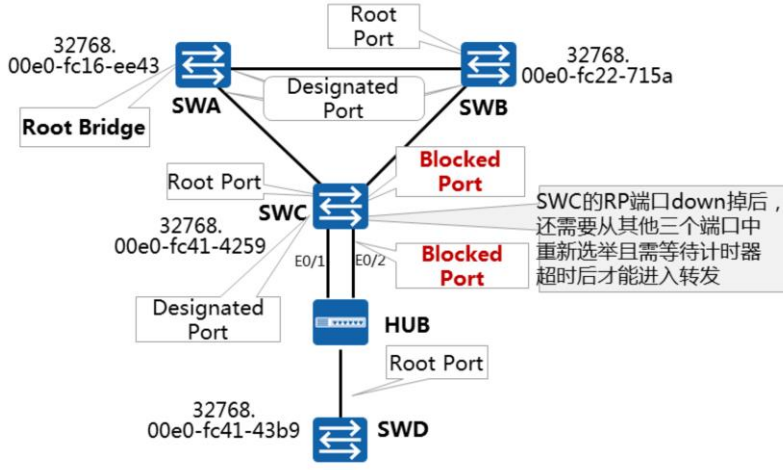
- 先由变更点朝根桥方向发送TCN消息，收到该消息的上游交换机就会回复TCA消息进行确认；最后TCN消息到达根桥后，再由根桥发送TC消息通知设备删除桥MAC地址表项，机制复杂，效率低下。



HUAWEI

- 拓扑变更处理过程：
  - 在网络拓扑发生变化后，下游设备会不间断地向上游设备发送TCN BPDU报文。
  - 上游设备收到下游设备发来的TCN BPDU报文后，只有指定端口处理TCN BPDU报文。其它端口也有可能收到TCN BPDU报文，但不会处理。
  - 上游设备会把配置BPDU报文中的Flags的TCA位设置1，然后发送给下游设备，告知下游设备停止发送TCN BPDU报文。
  - 上游设备复制一份TCN BPDU报文，向根桥方向发送。
  - 重复上述步骤，直到根桥收到TCN BPDU报文。
  - 根桥把配置BPDU报文中的Flags的TC位置1后发送，通知下游设备直接删除桥MAC地址表项。

STP的其他不足之处 - 端口角色



HUAWEI

STP的其他不足之处 - 端口状态

STP端口状态	端口状态对应的行为
Disabled	不转发用户流量也不学习MAC地址
Blocking	
Listening	
Learning	不转发用户流量但是学习MAC地址
Forwarding	既转发用户流量又学习MAC地址

三种端口状态从用户使用的角度对应的行为都相同，但呈现出不同的状态，这样反而增加了使用难度

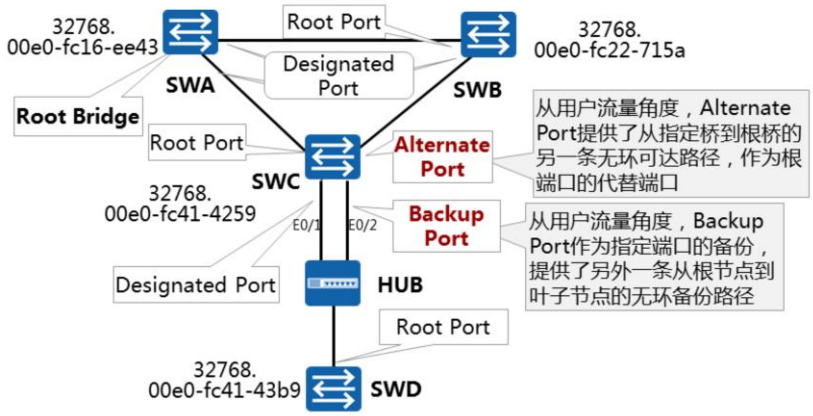
目录

1. STP的不足
2. **RSTP对STP的改进**
   - 端口角色与端口状态
   - 快速收敛机制
   - 拓扑变化处理机制
   - 保护功能
3. RSTP配置实例

HUAWEI

端口角色的重新划分

- RSTP定义了两种新的端口角色：备份端口（Backup Port）和预备端口（Alternate Port）。



HUAWEI

- 根据STP的不足，RSTP新增加了两种端口角色，并且把端口属性充分地按照状态和角色解耦，使得可以更加精确地描述端口，从而使得协议状态更加简便，同时也加快了拓扑收敛。通过端口角色的增补，简化了生成树协议的理解及部署。
- 从配置BPDU报文发送角度来看：
  - Alternate Port就是由于学习到其它网桥发送的配置BPDU报文而阻塞的端口。
  - Backup Port就是由于学习到自己发送的配置BPDU报文而阻塞的端口。
- 从用户流量角度来看：
  - Alternate Port提供了从指定桥到根的另一条可切换路径，作为根端口的备份端口。
  - Backup Port作为指定端口的备份，提供了另外一条从根节点到叶节点的备份通路。
- 给一个RSTP域内所有端口分配角色的过程就是整个拓扑收敛的过程。

端口状态的重新划分

- RSTP的状态规范把原来的5种状态缩减为3种：

STP端口状态	RSTP端口状态	端口状态对应的行为
Disabled	Discarding	如果不转发用户流量也不学习MAC地址，那么端口状态就是Discarding状态。
Blocking		
Listening		
Learning	Learning	如果不转发用户流量但是学习MAC地址，那么端口状态就是Learning状态。
Forwarding	Forwarding	如果既转发用户流量又学习MAC地址，那么端口状态就是Forwarding状态。

- 从用户角度来讲，Listening、Learning和Blocking状态并没有区别，都同样不转发用户流量。

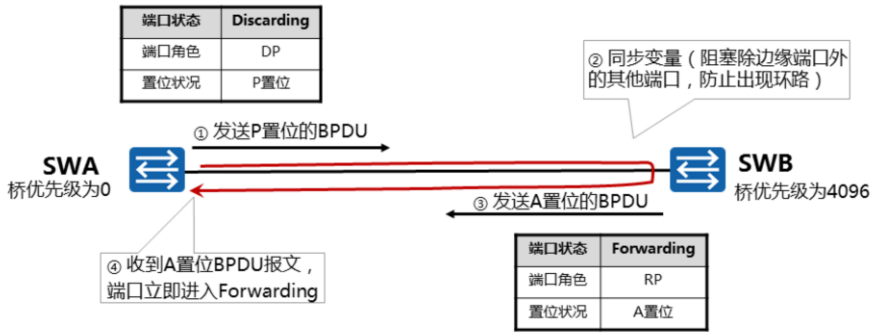
目录

1. STP的不足
2. **RSTP对STP的改进**
   - 端口角色与端口状态
   - **快速收敛机制**
   - 拓扑变化处理机制
   - 保护功能
3. RSTP配置实例

HUAWEI

针对问题一：P/A机制 (1)

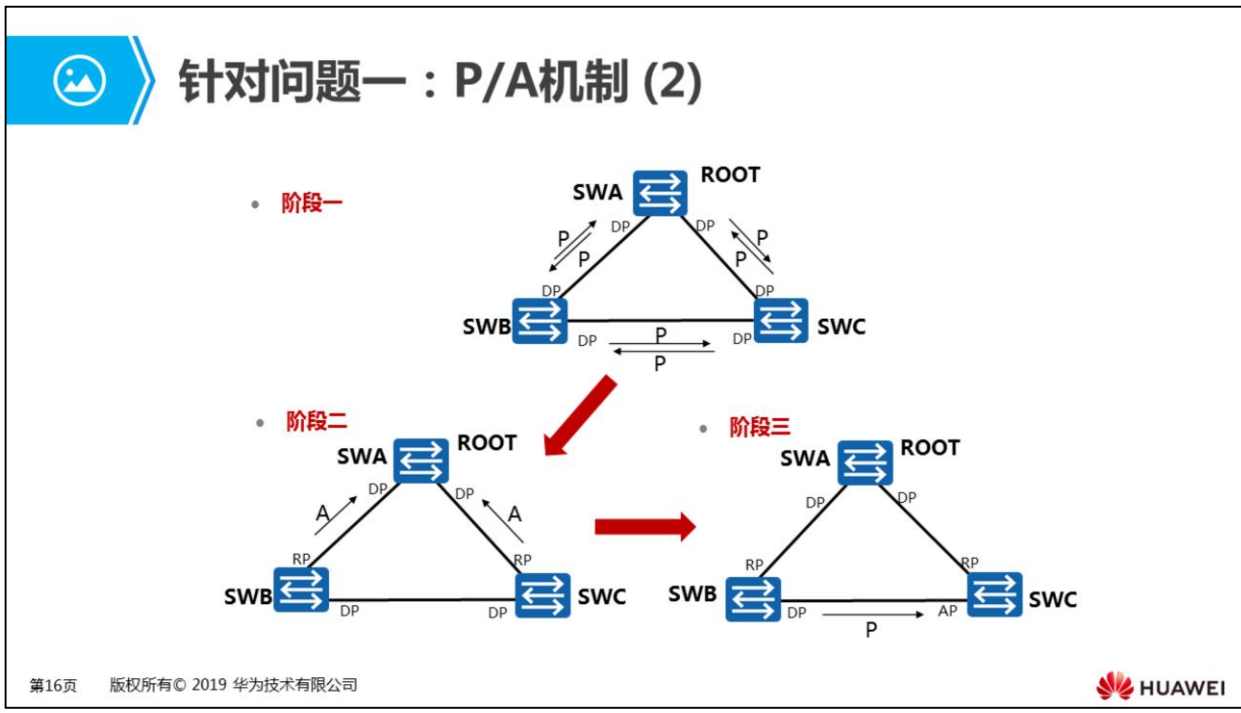
- P/A机制基本原理



- 特点：由于有来回确认机制和同步变量机制，就无需依靠计时器来保障无环。

HUAWEI

- Proposal/Agreement机制，其目的是使一个指定端口尽快进入Forwarding状态。
- P/A机制要求两台交换设备之间链路必须是点对点的全双工模式。一旦P/A协商不成功，指定端口的选择就需要等待两个Forward Delay，协商过程与STP一样。
- 事实上对于STP，指定端口的选择可以很快完成，主要的速度瓶颈在于：为了避免环路，必须等待足够长的时间，使全网的端口状态全部确定，也就是说必须要等待至少两个Forward Delay，所有端口才能进行转发。



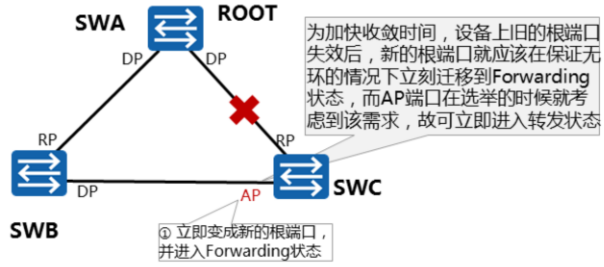
- 问题一的解决方案：
  - 阶段一：设备刚刚启动，RSTP协议刚刚启用，所有交换机都认为自己是根桥，向其他交换机发送P置位的BPDU，并把发送P消息的端口变成DP口，同时接口处在Discarding状态。
  - 阶段二：交换机SWA收到SWB和SWC的P消息会置之不理，因为他的桥优先级最高。交换机SWB和SWC收到SWA的P消息后，由于认同SWA是最优的根桥，会根据P/A协商流程回复A消息，并把发送端口变成RP端口，同时接口处在Forwarding状态。
  - 阶段三：SWA与SWB，SWA与SWB的P/A协商已经完成，接下来是SWB和SWC的P/A协商。
    - SWB和SWC都会发送根桥为SWA的P消息给对方。
    - SWC收到SWB的P消息后，发现P消息里虽然根桥和自己认可的一样，但是发送者的桥优先级比自己高(SWB>SWC)，所有马上停止发送P消息，但是由于已经有端口是RP口，也不会回A消息。
    - SWB收到SWC的P消息后，发现P消息里虽然根桥和自己认可的一样，但是发送者的桥优先级比自己低(SWB>SWC)，会不停的发送P消息。
    - 以上状态在等待2个Forward Delay时间后，SWB端口为DP端口，处在Forwarding状态， SWC端口为AP端口，处在Discarding状态。
    - 实际上SWB与SWC之间的协商等同于退回到STP的模式，但是反正是Discarding状态，根本不影响其他业务转发。

针对问题一：P/A机制 (3)

- RSTP选举原理和STP本质上相同：选举根交换机-选举非根交换机上的根端口-选举指定端口-选举预备端口和备份端口。
- 但是RSTP在选举的过程中加入了“发起请求-回复同意”（P/A机制）这种确认机制，由于每个步骤有确认就不需要依赖计时器来保证网络拓扑无环才去转发，只需要考虑BPDU发送报文并计算无环拓扑的时间（一般都是秒级）。

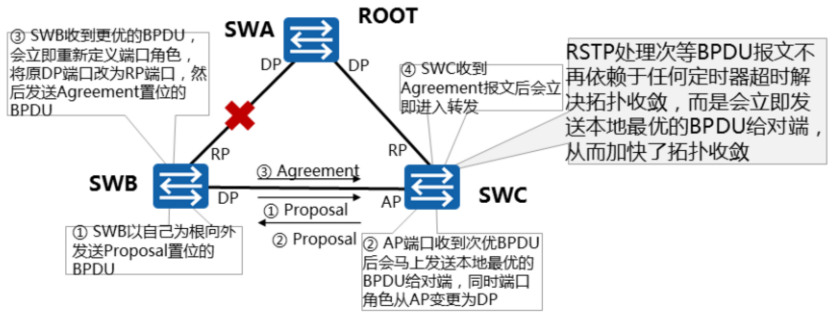
针对问题二：根端口快速切换机制

- SWC与SWA的直连链路down掉，其AP端口切换成RP端口并进入转发状态可在秒级时间内完成收敛：



针对问题三：次等BPDU处理机制

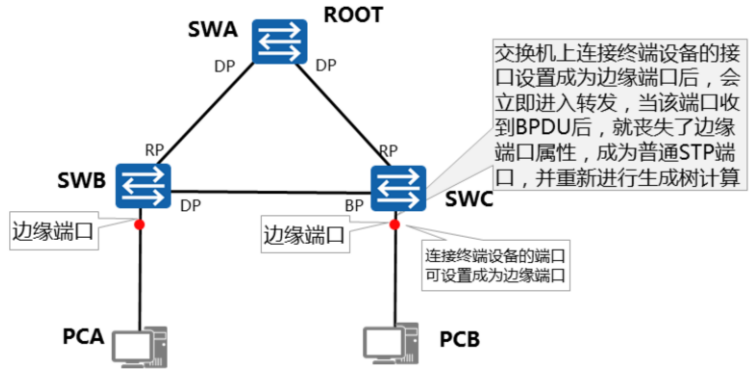
- SWB与SWA的直连链路down掉，SWC的AP端口切换成DP端口并进入转发状态可在秒级时间内完成：



HUAWEI

针对问题四：边缘端口的引入

- 在RSTP中，交换机连接终端的链路可立即进入转发状态：



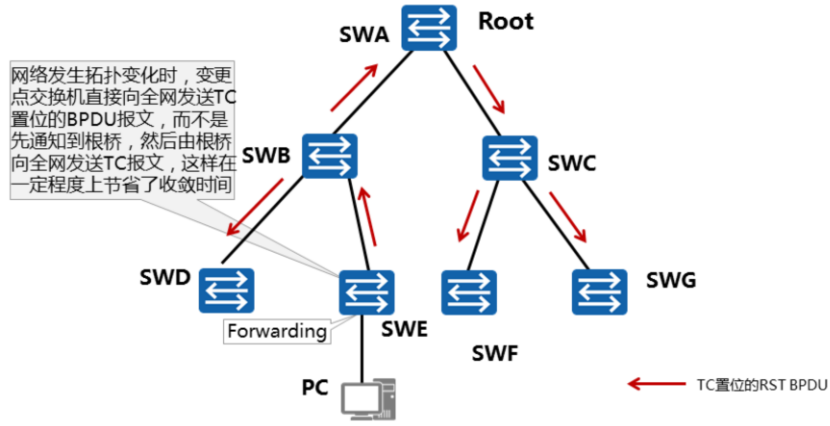
HUAWEI

目录

1. STP的不足
2. **RSTP对STP的改进**
   - 端口角色与端口状态
   - 快速收敛机制
   - **拓扑变化处理机制**
   - 保护功能
3. RSTP配置实例

针对问题五：拓扑变更机制的优化

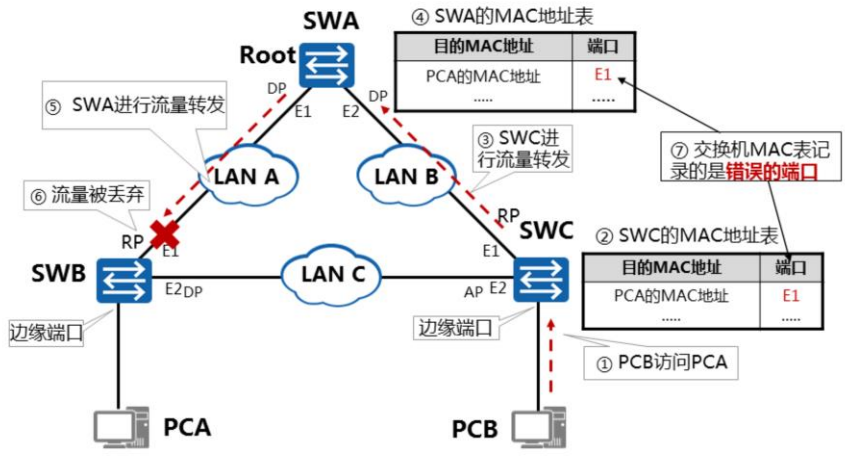
- 判断拓扑变化唯一标准：一个非边缘端口迁移到Forwarding状态。



HUAWEI

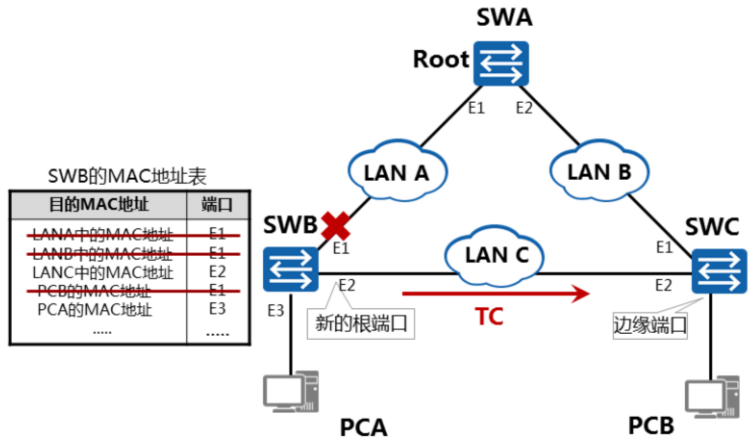
- 一旦检测到拓扑发生变化，将进行如下处理：
  - 为本交换设备的所有非边缘指定端口启动一个TC While Timer，该计时器值是Hello Time的两倍。在这个时间内，清空状态发生变化的端口上学习到的MAC地址。同时，由这些端口向外发送RST BPDU，其中TC置位。一旦TC While Timer超时，则停止发送RST BPDU。
  - 其他交换设备接收到RST BPDU后，清空所有端口学习到MAC地址，除了收到RST BPDU的端口。然后也为自己所有的非边缘指定端口和根端口启动TC While Timer，重复上述过程。 如此，网络中就会产生RST BPDU的泛洪。

拓扑变化引发的问题



- 在RSTP中检测拓扑是否发生变化只有一个标准：一个非边缘端口迁移到Forwarding状态。
- 网络拓扑改变可能会导致交换机的MAC地址表产生错误。
- 如图所示，在稳定情况下，SWC的MAC地址表中对应PCA的MAC地址的端口是E1。如果SWB的E1端口发生了故障，而SWC的MAC地址表中与PCA的MAC地址对应的端口仍然是E1，则会导致数据转发丢失的问题。

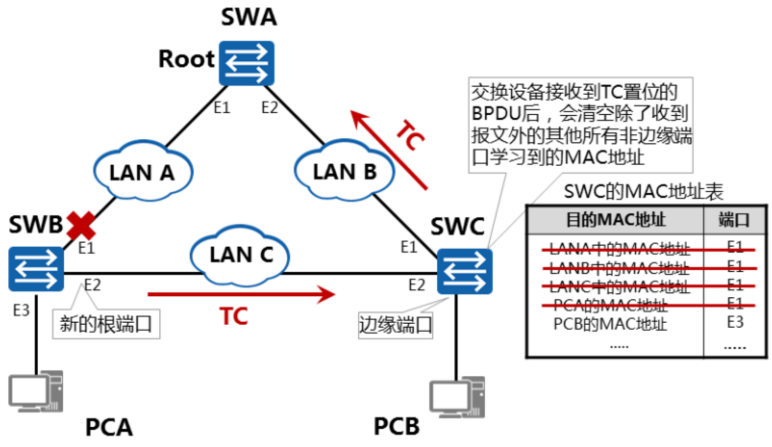
拓扑变化处理 (1)



  HUAWEI

- 一旦检测到拓扑发生变化，首先将进行如下处理：
  - 清空状态发生变化的端口上学习到的MAC地址。
  - 同时，由这些端口向外发送RST BPDU，其中TC置位。一旦TC While Timer超时，则停止发送RST BPDU。
- 如图所示，SWB的E1端口出现故障之后，RSTP的处理过程如下：
  - SWB重新计算生成树，选举E2为新的根端口。
  - SWB删除MAC地址表中E1端口所对应的表项。
  - 生成树重新计算完成之后（需要进入转发状态的端口已经进入了转发状态），SWB的所有非边缘端口向外发送TC置位的RST BPDU。

拓扑变化处理 (2)

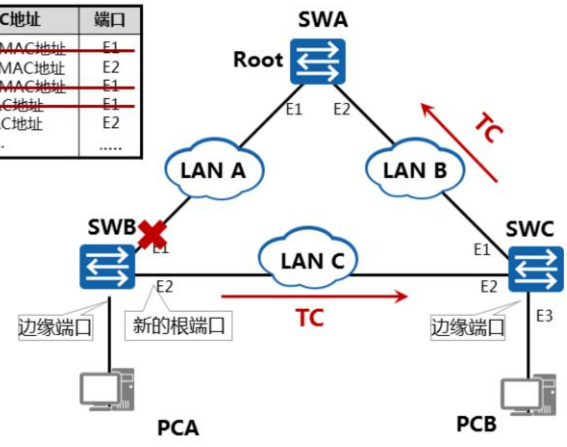


HUAWEI

拓扑变化处理 (3)

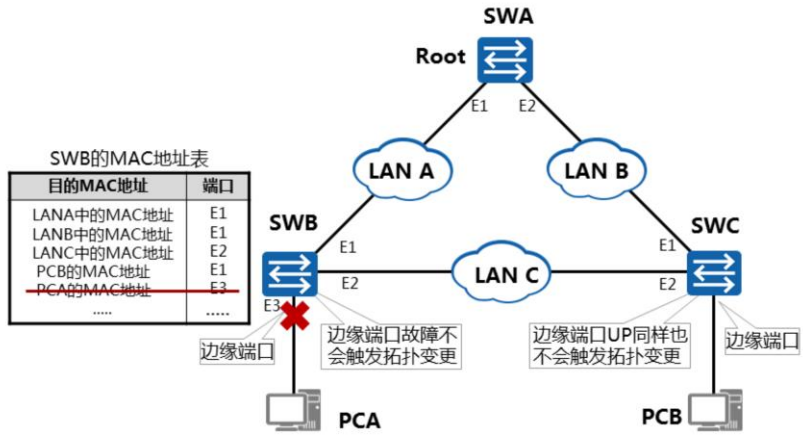
SWA的MAC地址表

目的MAC地址	端口
~~LANA中的MAC地址~~	~~E1~~
LANB中的MAC地址	E2
~~LANC中的MAC地址~~	~~E1~~
~~PCA的MAC地址~~	~~E1~~
PCB的MAC地址	E2
.....	.....



HUAWEI

拓扑变化处理 (4)



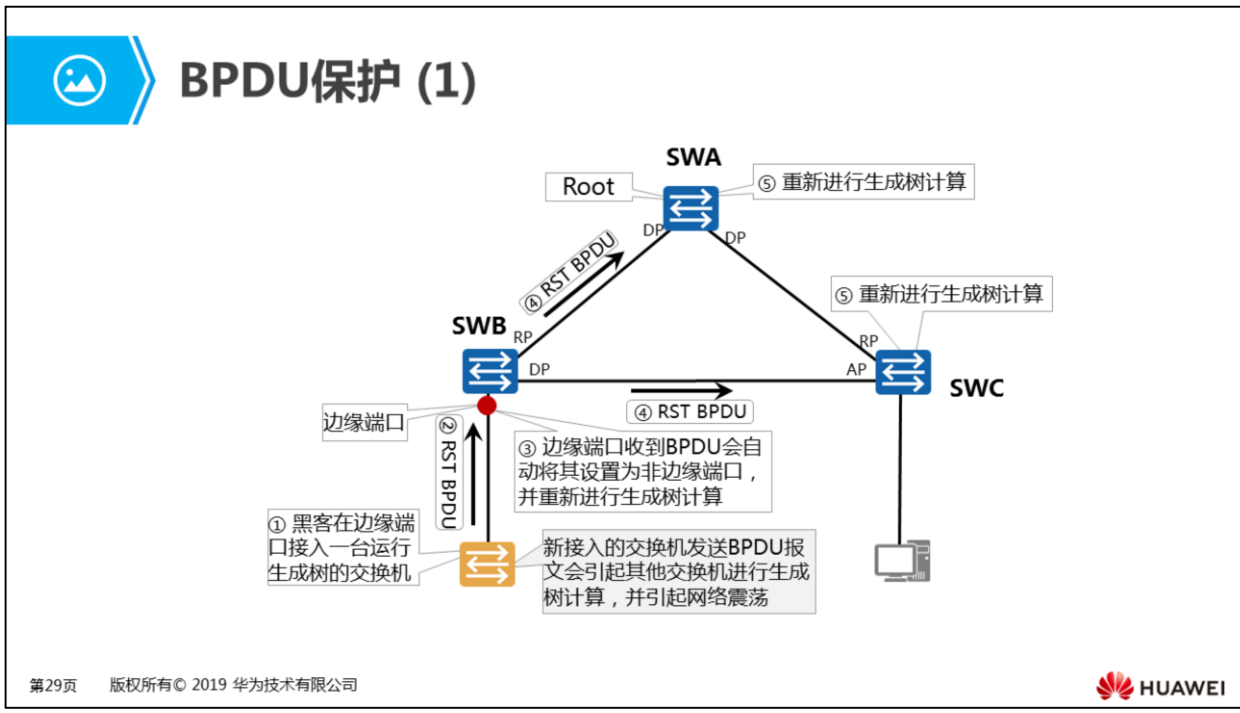
HUAWEI

- 边缘端口down掉不会触发拓扑变更，而且故障恢复后，同样也不会触发拓扑变更。

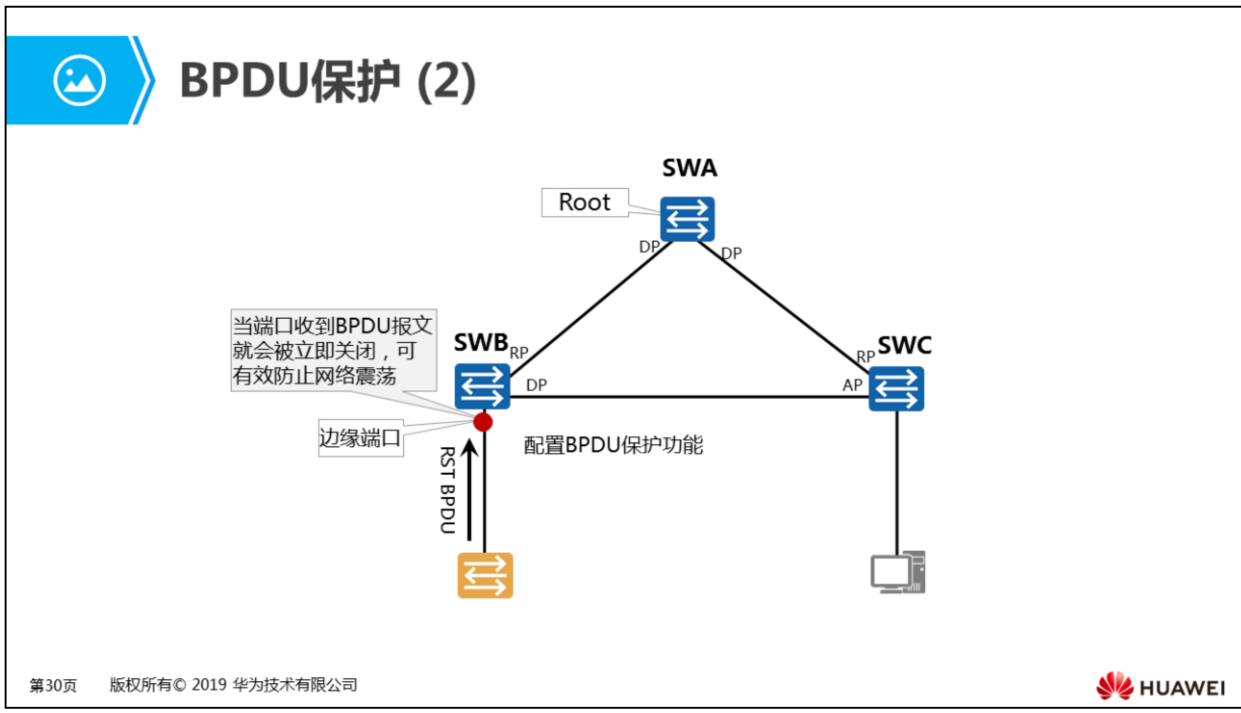
目录

1. STP的不足
2. **RSTP对STP的改进**
   - 端口角色与端口状态
   - 快速收敛机制
   - 拓扑变化处理机制
   - **保护功能**
3. RSTP配置实例

HUAWEI

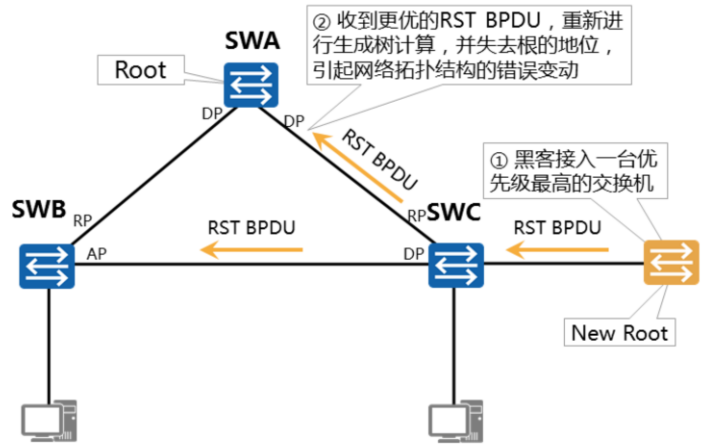


- BPDU保护
  - 应用场景：防止有人伪造RST BPDU恶意攻击交换设备，当边缘端口接收到该报文时，会自动设置为非边缘端口，并重新进行生成树计算，引起网络震荡。
  - 实现原理：配置BPDU保护功能后，如果边缘端口收到BPDU报文，边缘端口将会被立即关闭。



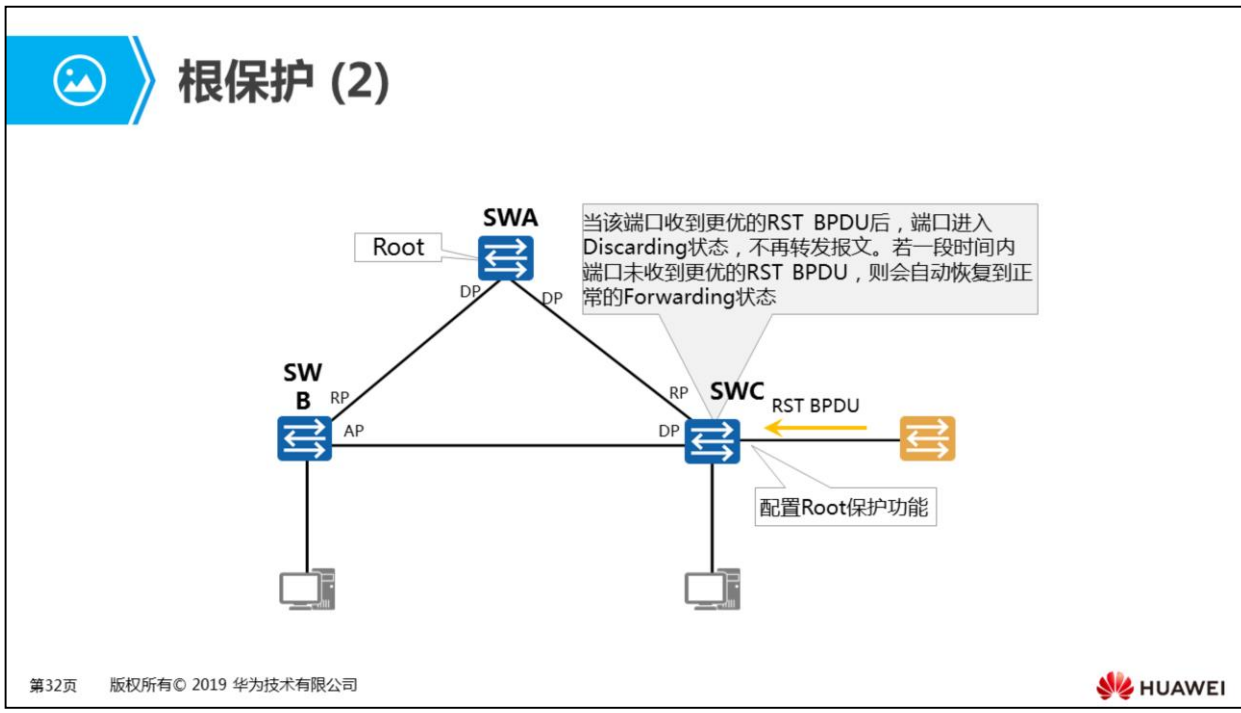
- BPDU保护
  - 实现原理：配置BPDU保护功能后，如果边缘端口收到BPDU报文，边缘端口将会被立即关闭。

根保护 (1)

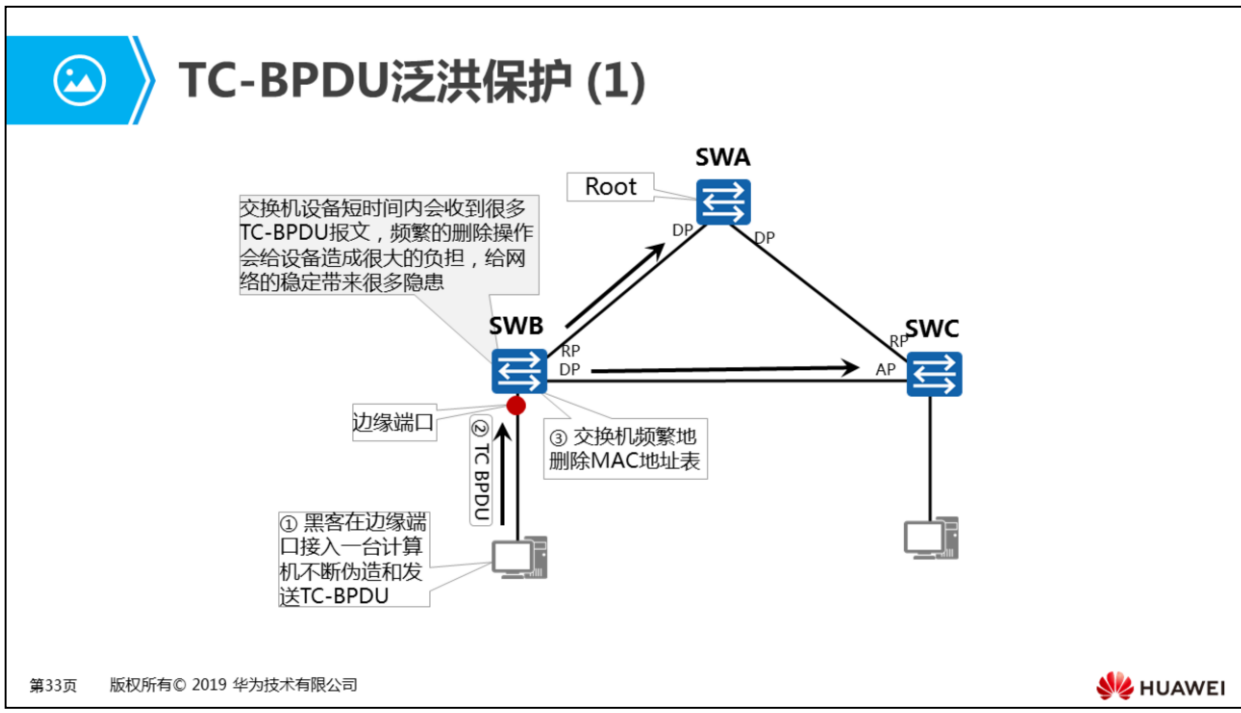


HUAWEI

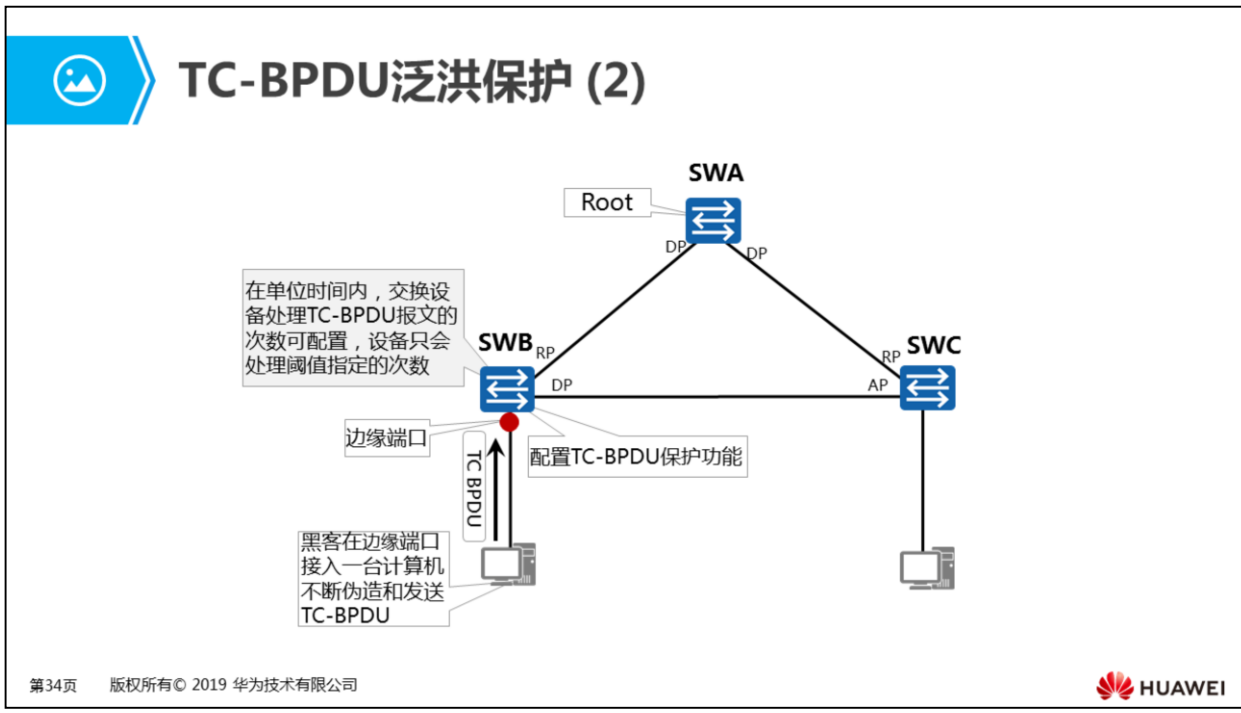
- 根保护
  - 应用场景：由于维护人员的错误配置或网络中的恶意攻击，网络中合法根桥有可能会收到优先级更高的RST BPDU，使得合法根桥失去根地位，从而引起网络拓扑结构的错误变动。
  - 实现原理：一旦启用Root保护功能的指定端口收到优先级更高的RST BPDU时，端口状态将进入Discarding状态，不再转发报文。在经过一段时间，如果端口一直没有再收到优先级较高的RST BPDU，端口会自动恢复到正常的Forwarding状态。
  - Root保护功能只能在指定端口上配置生效。



- 根保护
  - 应用场景：由于维护人员的错误配置或网络中的恶意攻击，网络中合法根桥有可能会收到优先级更高的RST BPDU，使得合法根桥失去根地位，从而引起网络拓扑结构的错误变动。
  - 实现原理：一旦启用Root保护功能的指定端口收到优先级更高的RST BPDU时，端口状态将进入Discarding状态，不再转发报文。在经过一段时间，如果端口一直没有再收到优先级较高的RST BPDU，端口会自动恢复到正常的Forwarding状态。
  - Root保护功能只能在指定端口上配置生效。



- TC-BPDU攻击：
  - 交换机在接收到TC-BPDU报文后，会执行MAC地址表项的删除操作。如果有人伪造TC-BPDU报文恶意攻击交换机时，交换机短时间内会收到很多TC-BPDU报文，频繁的删除操作会给设备造成很大的负担，给网络的稳定带来很大隐患。



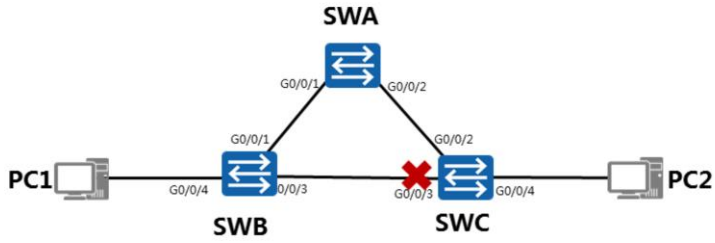
- TC-BPDU攻击保护：
  - 启用防TC-BPDU报文攻击功能后，在单位时间内，RSTP进程处理TC类型BPDU报文的次数可配置（缺省的单位时间是2秒，缺省的处理次数是3次）。如果在单位时间内，RSTP进程在收到TC类型BPDU报文数量大于配置的阈值，那么RSTP进程只会处理阈值指定的次数；对于其他超出阈值的TC类型BPDU报文，定时器到期后，RSTP进程只对其统一处理一次。这样可以避免频繁的删除MAC地址表项，从而达到保护交换机的目的。

目录

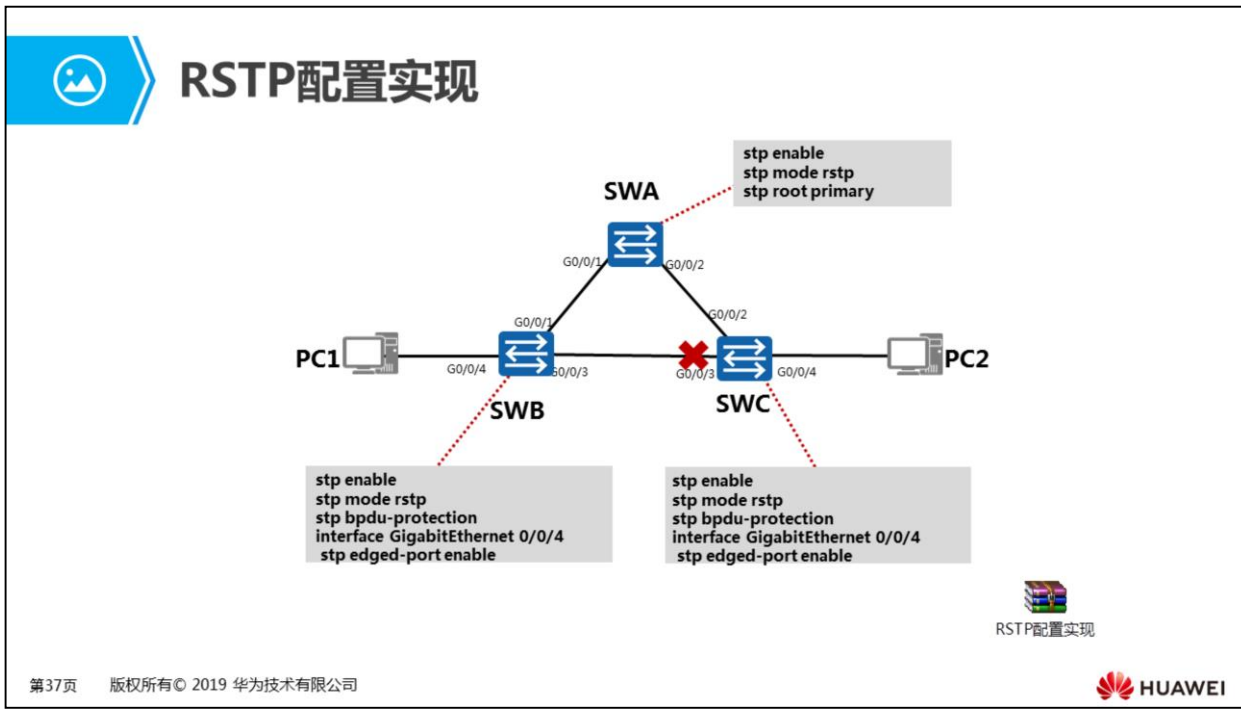
1. STP的不足
2. RSTP对STP的改进
3. **RSTP配置实例**

HUAWEI

RSTP配置需求



- 如图所示，SWA、SWB和SWC组成了一个环形的交换网络，为了消除环路对网络的影响，故使交换机都运行RSTP，最终将环形网络结构修剪成无环路的树形网络结构。



- 配置实现：
  - stp enable　　//全局开启STP
  - stp mode rstp　　//配置STP模式为RSTP
  - stp root primary　　//配置SWA为根桥
  - stp bpdu-protection　　//全局开启BPDU防护，配合边缘端口一起使用
  - stp edged-port enable　　//配置端口为边缘端口

RSTP配置验证 (1)

- 在SWA上查看生成树信息：

```
<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配置验证 (2)

- 在SWB上查看生成树信息：

```
[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
```

- 在SWC上查看生成树信息：

```
<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
```


- 最终阻塞了SWC的G0/0/3接口，消除了网络中的环路。

思考题

1. RSTP定义了几种端口状态？（　　）

   A. 2

   B. 3

   C. 4

2. RSTP定义了哪些端口角色？（　　）

- 答案：B。
- 答案：根端口、指定端口、备份端口、预备端口、边缘端口。


谢 谢
www.huawei.com

MSTP协议原理与配置

HUAWEI

前言

- RSTP在STP基础上进行了改进，实现了网络拓扑快速收敛。但由于局域网内所有的VLAN共享一棵生成树，因此被阻塞后链路将不承载任何流量，无法在VLAN间实现数据流量的负载均衡，从而造成带宽浪费。
- 为了弥补STP和RSTP的缺陷，IEEE于2002年发布的802.1s标准定义了MSTP。MSTP兼容STP和RSTP，既可以快速收敛，又提供了数据转发的多个冗余路径，在数据转发过程中实现VLAN数据的负载均衡。

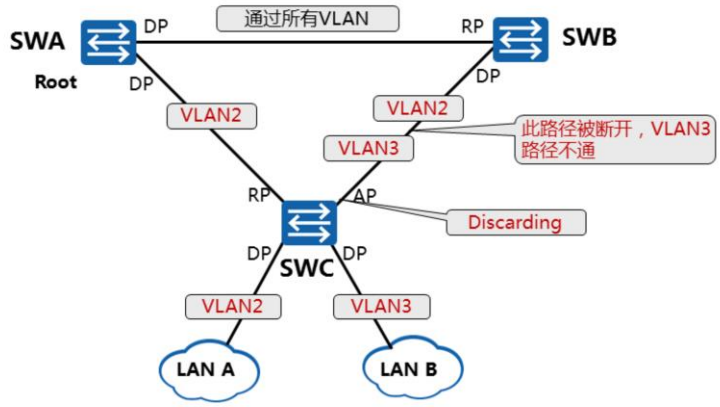
目标

- 学习完本课程后，您将能够：
  - 熟悉单生成树的缺陷
  - 理解MSTP的工作原理
  - 掌握MSTP基本配置命令

目录

1. **单生成树的弊端**
2. MSTP基本原理
3. MSTP配置实现

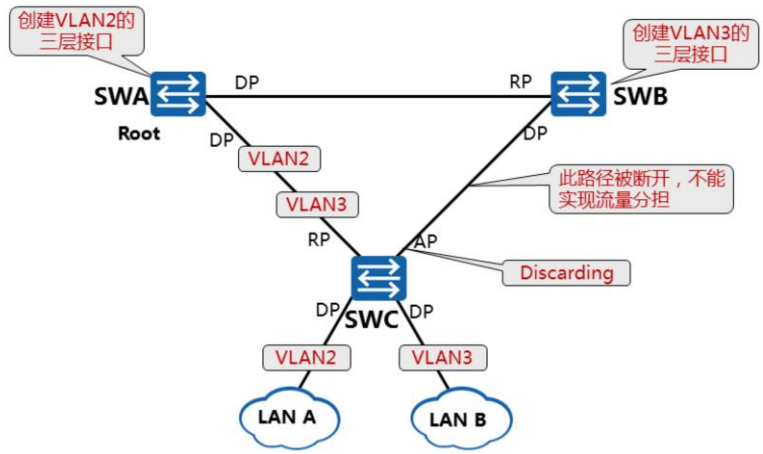
单生成树的弊端 - 部分VLAN路径不通



HUAWEI

- 如图所示，网络中有SWA、SWB、SWC三台交换机。配置VLAN2通过两条上行链路，配置VLAN3只通过一条上行链路。
- 为了解决VLAN2的环路问题，需要运行生成树。在运行单个生成树的情况下，假设SWC与SWB相连的端口成为预备端口（Discarding状态），那么VLAN3的路径就会被断开，无法上行到SWB。

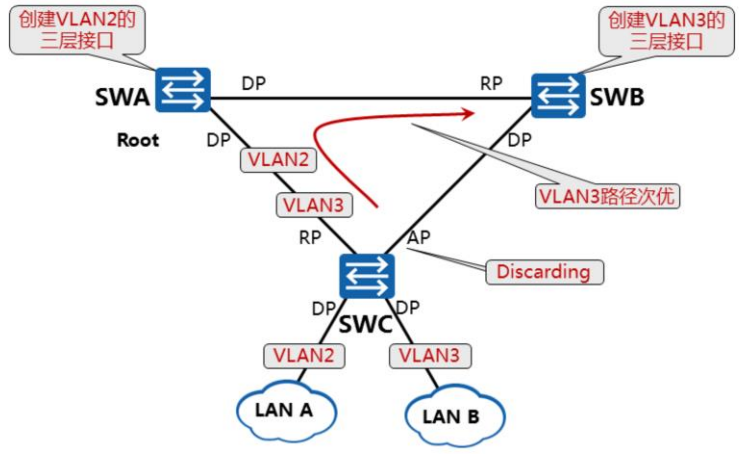
单生成树的弊端 - 无法实现流量分担



HUAWEI

- 为了实现流量分担，需要配置两条上行链路为Trunk链路，允许通过所有VLAN；SWA和SWB之间的链路也配置为Trunk链路，允许通过所有VLAN。将VLAN2的三层接口配置在SWA上，将VLAN3的三层接口配置在SWB上。
- 我们希望VLAN2和VLAN3分别使用不同的链路上行到相应的三层接口，但是如果连接到SWB的端口成为预备端口（Alternate Port）并处于Discarding状态，则VLAN2和VLAN3的数据都只能通过一条上行链路上行到SWA，这样就不能实现流量分担。

单生成树的弊端 - 次优二层路径



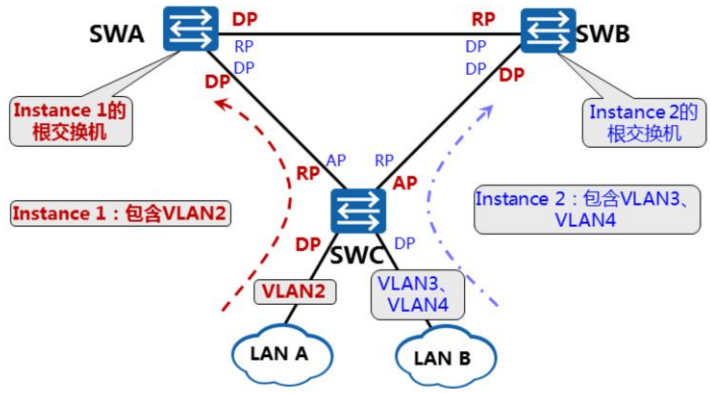
HUAWEI

- 如图所示，SWC与SWA和SWB相连的链路配置为Trunk链路，允许通过所有VLAN；SWA与SWB之间的链路也配置为Trunk链路，允许通过所有VLAN。
- 运行单个生成树之后，环路被断开，VLAN2和VLAN3都直接上行到SWA。
- 在SWA上配置VLAN2的三层接口，在SWB上配置VLAN3的三层接口，那么，VLAN3到达三层接口的路径就是次优的。

目录

1. 单生成树的弊端
2. **MSTP基本原理**
3. MSTP配置实现

多生成树实例解决单生成树弊端



- MST域内可以生成多棵生成树，每棵生成树都称为一个MSTI。MSTI之间彼此独立，且每个MSTI的计算过程基本与RSTP的计算过程相同。

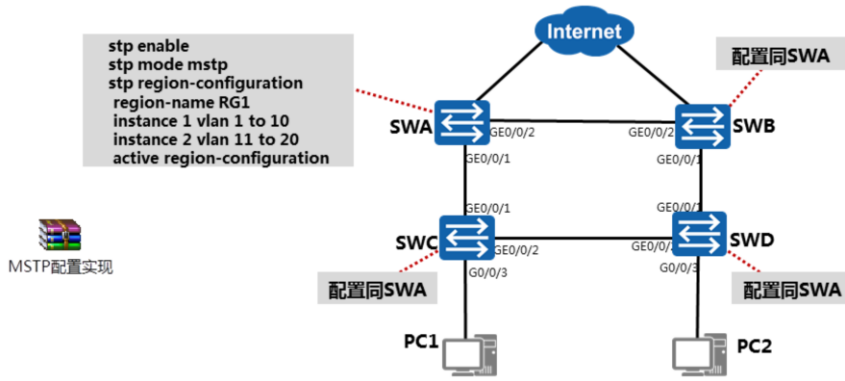
- 多生成树协议即MSTP（Multiple Spanning Tree Protocol）。
- MST域是多生成树域（Multiple Spanning Tree Region），由交换网络中的多台交换设备以及它们之间的网段所构成。同一个MST域的设备具有下列特点：
  - 都启动了MSTP。
  - 具有相同的域名。
  - 具有相同的VLAN到生成树实例映射配置。
  - 具有相同的MSTP修订级别配置。
- 一个MST域内可以生成多棵生成树，每棵生成树都称为一个MSTI，每个MSTI都使用单独的RSTP算法，计算单独的生成树。
- 每个MSTI（MST Instance）都有一个标识（MSTID），MSTID是一个两字节的整数。VRP平台支持16个MST Instance，MSTID取值范围是0～15，默认所有VLAN映射到MST Instance 0。
- VLAN映射表是MST域的属性，它描述了VLAN和MSTI之间的映射关系，MSTI可以与一个或多个VLAN对应，但一个VLAN只能与一个MSTI对应。
- MSTP兼容STP和RSTP，既可以快速收敛，又提供了数据转发的各个冗余路径，在数据转发过程中实现VLAN数据的负载均衡。

目录

1. 单生成树的弊端
2. MSTP基本原理
3. **MSTP配置实现**

MSTP配置实现 (1)

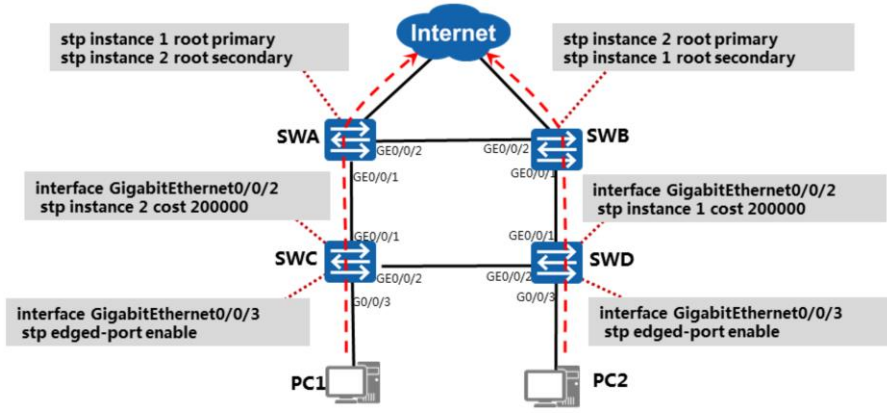
- 为实现分别属于不同VLAN的PC访问Internet的流量能够进行负载均衡，可采用MSTP来实现，VLAN1~10为一组，VLAN11~20为另一组。



  HUAWEI

- 配置思路：
  - 配置MST域并创建多实例，实现流量的负载分担。
  - 在MST域内，配置各实例的根桥与备份根桥。
  - 修改各实例中某端口的路径开销值，实现将该端口阻塞。
  - 与终端设备相连的端口配置成为边缘端口，加快收敛。
- 数据准备：
  - 域名为RG1。
  - 实例为MSTI1和MSTI2。
  - 实例MSTI1的根桥为SWA，备份根桥为SWB；实例MSTI2的根桥为SWB，备份根桥为SWA。
  - 实例MSTI1和实例MSTI2的阻塞口的路径开销值修改为200000。
  - VLAN号是1～20。
  - PC1所属VLAN为10，PC2所属VLAN为20。

MSTP配置实现 (2)



HUAWEI

MSTP配置验证 (1)

- 在SWA上查看端口状态，结果如下：

```
[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
```

- 在SWB上查看端口状态，结果如下：

```
[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配置验证 (2)

- 在SWC上查看端口状态，结果如下：

```
[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
```

- 在SWD上查看端口状态，结果如下：

```
<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
```

思考题

1. 请简述单生成树的缺陷。
2. 关于MSTP的描述，错误的是（　　）。

   A. 一个MST域内只能有一个生成树实例。

   B. 每个生成树实例使用独立的RSTP算法。

   C. MSTP兼容于STP。

   D. 一个MSTI可以与一个或多个VLAN对应。

- 答案：链路被阻塞后将不承载任何流量，无法在VLAN间实现数据流量的负载均衡，从而造成带宽浪费；导致部分VLAN路径不通；造成次优路径。
- 答案：A。


谢 谢
www.huawei.com

学习推荐

- 华为培训与认证官方网站
  - http://learning.huawei.com/cn/
- 华为在线学习
  - https://ilearningx.huawei.com/portal/#/portal/ebg/26
- 华为职业认证
  - http://support.huawei.com/learning/NavigationAction!createNavi?navId=_31&lang=zh
- 查找培训入口
  - http://support.huawei.com/learning/NavigationAction!createNavi?navId=_trainingsearch&lang=zh

更多信息

- 华为培训APP