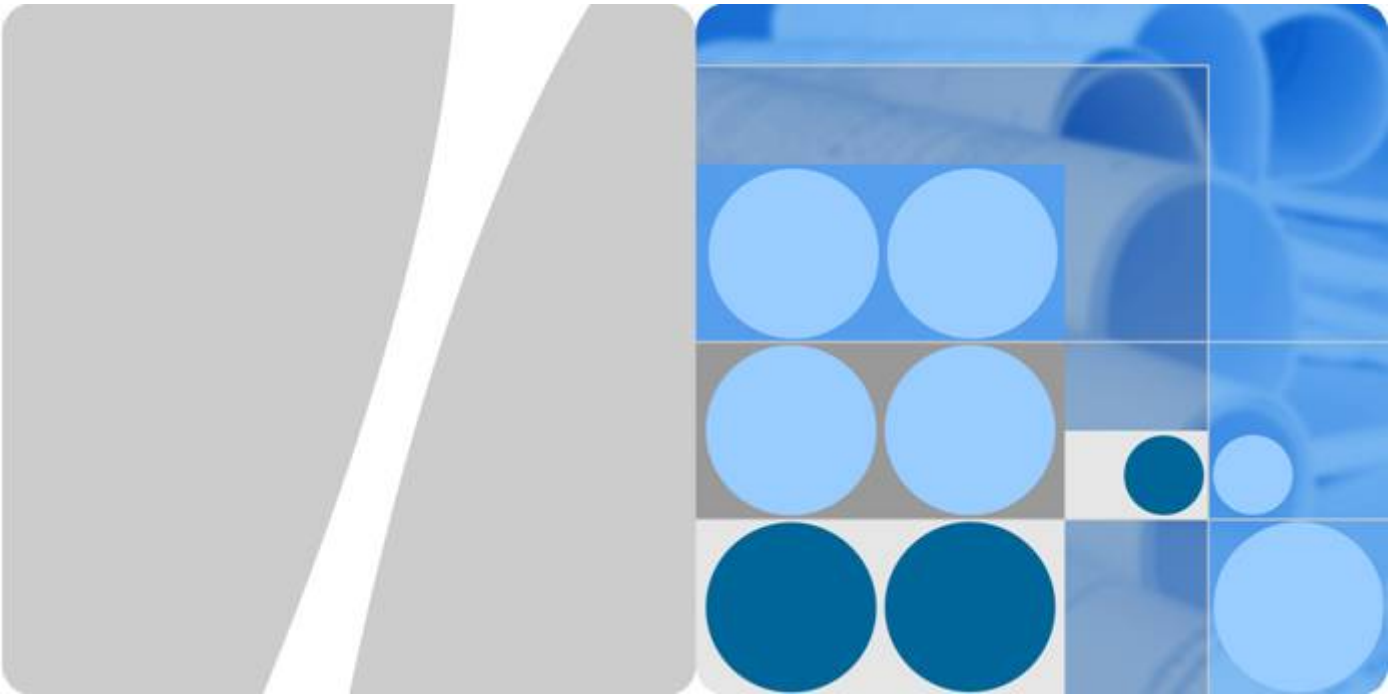


Huawei AR3200 系列企业路由器

V200R002C00

特性描述-局域网

文档版本　01

发布日期　2011-12-30

华为技术有限公司

商标声明


和其他华为商标均为华为技术有限公司的商标。
本文档提及的其他所有商标或注册商标，由各自的所有人拥有。

注意

您购买的产品、服务或特性等应受华为公司商业合同和条款的约束，本文档中描述的全部或部分产品、服务或特性可能不在您的购买或使用范围之内。除非合同另有约定，华为公司对本文档内容不做任何明示或默示的声明或保证。

由于产品版本升级或其他原因，本文档内容会不定期进行更新。除非另有约定，本文档仅作为使用指导，本文档中的所有陈述、信息和建议不构成任何明示或暗示的担保。

华为技术有限公司

地址： 深圳市龙岗区坂田华为总部办公楼 邮编：518129

网址： http://www.huawei.com

客户服务邮箱： support@huawei.com

客户服务电话： 4008302118

前 言

读者对象

本文档针对局域网特性，从简介、原理描述和应用三个方面介绍了局域网特性。

本文档与其它类型手册相结合，便于读者深入掌握特性的实现原理。

本文档主要适用于以下工程师：

- 网络规划工程师
- 调测工程师
- 数据配置工程师
- 系统维护工程师

符号约定

在本文中可能出现下列标志，它们所代表的含义如下。

符号	说明
 危险	以本标志开始的文本表示有高度潜在危险，如果不能避免，会导致人员死亡或严重伤害。
 警告	以本标志开始的文本表示有中度或低度潜在危险，如果不能避免，可能导致人员轻微或中等伤害。
 注意	以本标志开始的文本表示有潜在风险，如果忽视这些文本，可能导致设备损坏、数据丢失、设备性能降低或不可预知的结果。
 窍门	以本标志开始的文本能帮助您解决某个问题或节省您的时间。
 说明	以本标志开始的文本是正文的附加信息，是对正文的强调和补充。

命令行格式约定

格式	意义
粗体	命令行关键字（命令中保持不变、必须照输的部分）采用**加粗**字体表示。
斜体	命令行参数（命令中必须由实际值进行替代的部分）采用*斜体*表示。
[ ]	表示用“[ ]”括起来的部分在命令配置时是可选的。
{ x | y | ... }	表示从两个或多个选项中选取一个。
[x | y | ...]	表示从两个或多个选项中选取一个或者不选。
{ x | y | ... } *	表示从两个或多个选项中选取多个，最少选取一个，最多选取所有选项。
[x | y | ...] *	表示从两个或多个选项中选取多个或者不选。
&<1-n>	表示符号&的参数可以重复 1 ～ n 次。
#	由“#”开始的行表示为注释行。

修订记录

修改记录累积了每次文档更新的说明。最新版本的文档包含以前所有文档版本的更新内容。

文档版本 01 (2011-12-30)

第一次正式发布。

目 录

1 Trunk

关于本章

1.1 介绍

定义

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 Trunk 接口，捆绑在一起的每个物理接口称为成员接口。

Trunk 技术可以实现增加带宽、提高可靠性和负载分担的功能。

目的

在没有使用 Trunk 前，百兆以太网的双绞线在两个互连的网络设备间的带宽仅为 100Mbit/s。若想达到更高的数据传输速率，则需要更换传输媒介，使用千兆光纤或升级成为千兆以太网。这样的解决方案成本昂贵，不适合中小型企业和学校应用。

如果采用 Trunk 技术把多个接口捆绑在一起，则可以以较低的成本满足提高接口带宽的需求。例如，把 3 个 100Mbit/s 的全双工接口捆绑在一起，就可以达到 300Mbit/s 的最大带宽。

1.2 参考标准和协议

本特性的参考资料清单如下：

文档	描述	备注
IEEE 802.3AD	IEEE Std 802.3ad - 2005 IEEE Standard for Link Aggregation operation , Link Aggregation Control , Link Aggregation Control Protocol , Marker protocol and Configuration capabilities and restrictions.	-

1.3 可获得性

涉及网元

无需其它网元的配合。

License 支持

无需获得 License 许可，均可获得该特性的服务。

版本支持

产品	最低支持版本
AR3200	V200R001C00

1.4 原理描述

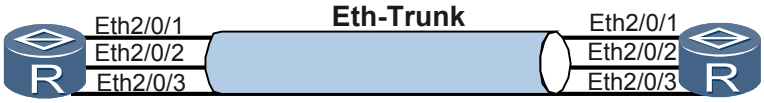
1.4.1 Trunk 基本原理

在一个 Trunk 内，可以实现流量负载分担，同时也提供了更高的连接可靠性和更大的带宽。

用户通过对逻辑口进行配置，实现各种路由协议以及其它业务。

如图 1-1 所示，以 Eth-Trunk 为例。两台路由器通过 3 个 Eth 口直连，将这 3 个 Eth 接口捆绑，形成一个 Eth-Trunk 接口，从而实现了增加带宽和提高可靠性的目的。

图 1-1 Eth-Trunk 示意图



Trunk 接口连接的链路可以看成是一条点到点的直连链路，链路的两端可以都是交换机或路由器，也可以一端是交换机，另一端是路由器。

Trunk 的优势在于：

- 提高可靠性

  当某个成员接口连接的物理链路出现故障时，流量会切换到其他可用的链路上，从而提高整个 Trunk 链路的可靠性。
- 增加带宽

  Trunk 接口的总带宽是各成员接口带宽之和。通过这种方式可以成倍的增加接口带宽。
- 负载分担

  通过 Trunk 接口可以实现负载分担。Trunk 接口将流量分散到不同的链路上，最后到达同一目的地。这样可以避免所有流量都走同一条链路而导致网络拥塞。

目前 AR3200 支持的 Eth-Trunk 接口链路聚合模式如表 1-1 所示。

表 1-1 Eth-Trunk 接口的链路聚合模式

链路聚合模式	应用场景	说明
手工负载分担模式 Eth-Trunk 接口	当 Eth-Trunk 链路两端设备中有一台设备不支持 LACP 协议时，可在 AR3200 设备上创建手工负载分担模式的 Eth-Trunk，并加入多个成员接口增加设备间的带宽及可靠性。	手工负载分担模式是一种最基本的链路聚合方式。在该模式下，Eth-Trunk 的建立，成员接口的加入，以及哪些接口作为活动接口完全由手工来配置，没有链路聚合控制协议的参与。

链路聚合模式	应用场景	说明
静态 LACP（Link Aggregation Control Protocol）模式 Eth-Trunk 接口	组成 Eth-Trunk 链路的两台设备直连，并且都支持 LACP 协议时，可在 AR3200 设备上创建静态 LACP 模式 Eth-Trunk 接口。这种方式同时可以实现负载分担和冗余备份的双重功能。	静态 LACP 模式下，Eth-Trunk 的建立，成员接口的加入，都是由手工配置完成的。但与手工负载分担模式链路聚合不同的是，该模式下活动接口的选择由 LACP 协议报文负责。也就是说，当把一组接口加入 Eth-Trunk 接口后，这些成员接口中哪些接口作为活动接口，哪些接口作为非活动接口还需要经过 LACP 协议报文的协商确定。 静态 LACP 模式也称为 M：N 模式。这种方式同时可以实现负载分担和冗余备份的双重功能。在链路聚合组中 M 条链路处于活动状态，这些链路负责转发数据并进行负载分担，另外 N 条链路处于非活动状态作为备份链路，不转发数据。当 M 条链路中有链路出现故障时，系统会从 N 条备份链路中选择优先级最高的接替出现故障的链路，同时这条链路状态变为活动状态开始转发数据。

1.4.2 Trunk 接口的约束条件

在逻辑上把多条物理链路等同于一条逻辑链路，而又对上层数据透明传输，必须遵循以下规则。

- 物理接口的物理参数必须一致。Trunk 链路两端要求一致的物理参数有：
  - Trunk 链路两端相连的物理接口数量。
  - Trunk 链路两端相连的物理接口的速率。
  - Trunk 链路两端相连的物理接口的双工方式。
  - Trunk 链路两端相连的物理接口的流控方式。
- 必须保证数据的有序性。

  数据流就是具有相同源 MAC 地址、目的 MAC 地址、源 IP 地址和目的 IP 地址的一组数据包。例如，两台设备之间的 Telnet 或 FTP 连接就是一个数据流。

  如果要求属于同一个数据流的二层数据帧必须按照顺序到达，在没使用 Trunk 接口时是可以保证的，因为两台设备之间只有一条物理连接。但使用 Trunk 技术后，由于两台设备之间有多条物理链路，如果第一个数据帧在第一条链路上传播，第二个数据帧在第二条链路上传播，这样就可能第二个数据帧比第一个数据帧先到达对端设备。

为了避免这种数据包乱序的情况发生，在实现 Trunk 的时候引入了一种数据包转发机制，确保属于同一个数据流的数据帧按照发送的先后顺序到达目的地。这种机制根据 MAC 地址或 IP 地址来区分数据流，将属于同一数据流的数据帧通过同一条物理链路发送到目的地。

引入数据包转发机制后：

- 可以根据源 MAC 地址区分数据流，使源 MAC 地址都相同的数据帧在同一条物理链路上传输。
- 可以根据目的 MAC 地址区分数据流，使目的 MAC 地址都相同的数据帧在同一条物理链路上传输。
- 可以根据源 IP 地址区分数据流，使源 IP 地址都相同的数据帧在同一条物理链路上传输。
- 可以根据目的 IP 地址区分数据流，使目的 IP 地址都相同的数据帧在同一条物理链路上传输。
- 可以根据源 MAC 地址和目的 MAC 地址区分数据流，使源 MAC 地址和目的 MAC 地址都相同的数据帧在同一条物理链路上传输。
- 可以根据源 IP 地址和目的 IP 地址区分数据流，使源 IP 地址和目的 IP 地址都相同的数据帧在同一条物理链路上传输。

1.4.3 Trunk 接口的分类及特性

Trunk 接口的分类

Trunk 接口分为 Eth-Trunk 和 IP-Trunk 两种。

- Eth-Trunk 只能由以太网链路构成。
- IP-Trunk 只能由 POS 链路构成。

目前 AR3200 只支持 Eth-Trunk 接口。

Trunk 接口的特性

AR3200 中的 Eth-Trunk 接口支持以下特性：

- 支持配置 IP 地址，各成员接口借用 Trunk 接口的 IP 地址。
- 支持二层转发和三层转发（单播及组播）。
- 支持使用 HASH 算法进行流的负载分担。
- 支持基于接口的 QoS。
- 支持绑定 VPN 实例。
- 支持热插拔。
- 支持不同接口板上的接口加入到同一个 Eth-Trunk。

Trunk 接口 Up 链路的上下限阈值

在一个 Trunk 接口内，处于 Up 状态的成员链路数可以影响到 Trunk 接口的状态和带宽。为保持 Trunk 相对稳定，可以设置以下两个阈值，以减小成员链路状态变化带来的影响。

- Up 链路下限阈值

当处于 Up 状态的成员链路数目小于下限阈值时，Trunk 接口的状态转为 Down。设置 Up 链路下限阈值的目的是为了保证最小带宽。

例如，每一条链路能提供 1G 的带宽，现在最小需要 2G 的带宽，则 Up 链路下限阈值必须要大于等于 2。

- Up 链路上限阈值

  当处于 Up 状态的成员链路数达到上限阈值后，之后再发生成员链路 Up 不会使 Trunk 的带宽增加。设置 Up 链路上限阈值的目的是在保证了带宽的情况下提高网络的可靠性。

  例如，有 8 条无故障链路在一个 Trunk 内，每一条链路都能提供 1G 的带宽，现在最多需要 5G 的带宽，则 Up 链路上限阈值就可以设为 5 或者 6。其他的链路就自动进入备份状态以提高网络的可靠性。

  说明

  Up 链路上限阈值只适用于静态 LACP 模式的 Eth-Trunk 接口。

  静态 LACP 模式的 Eth-Trunk 接口的 Up 链路上限阈值可用来控制 Trunk 成员链路 Up 的数目。当 Trunk 内的链路数量大于 Up 链路上限阈值时，处于 Up 状态的链路数不能超过 Up 链路上限阈值。超过 Up 链路上限阈值的链路状态将被置为 Down。

在二层模式下，Eth-Trunk 的速率由以下两个条件决定：

- Up 链路上限阈值。
- Trunk 中状态为 Up 的端口的数目。

Trunk 接口的负载分担

目前支持的负载分担方式为逐流负载分担。

逐流负载分担是指根据报文的 MAC 地址或 IP 地址区别数据流，使属于同一数据流的报文从同一个的成员链路上通过。

逐流负载分担能保证包的顺序，但不能保证带宽利用率。

Trunk 成员接口备份

为提高 Trunk 接口的可靠性，可以为成员接口配置备份接口。

如果成员接口故障，使用同一 Trunk 接口中处于 Up 状态的其他接口作为备份接口承载故障接口上的流量。成员接口备份称为组内备份，也可称为组内快速倒换。

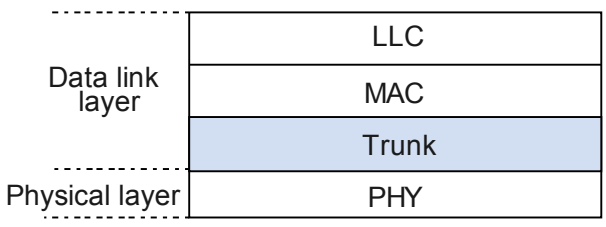
说明

Trunk 成员接口备份只适用于静态 LACP 模式的 Eth-Trunk 接口。

1.4.4 Trunk 转发原理

如图 1-2 所示，Trunk 位于 MAC 子层与物理层之间，属于数据链路层。

图 1-2 Trunk 接口在以太网协议栈的位置



对于 MAC 子层来说，Trunk 接口可以认为是一个物理接口。因此，MAC 子层在传输数据的时候，仅需要把数据提交给 Trunk 模块即可。

Trunk 模块内部维护一张 Trunk 转发表。这张表以下两项组成。

- HASH-KEY 值

  HASH-KEY 值是根据数据包的 MAC 地址或 IP 地址，经 HASH 算法计算得出。
- 接口号

  Trunk 转发表表项的数目跟捆绑的接口数目相同，如果把四个接口进行捆绑，该表就有四项。

  例如，将接口 3、4、5、6 捆绑为一个 Trunk 接口，Trunk 模块形成如图 1-3 所示的 Trunk 转发表。其中 HASH-KEY 值为 0、1、2、3，对应的接口号分别为 3、4、5、6。

图 1-3 Trunk 转发表示例

KEY	0	1	2	3
PORT	3	4	5	6

Trunk 模块根据 Trunk 转发表转发数据帧的过程如下：

1. Trunk 模块从 MAC 子层接收到一个数据帧后，提取数据帧的源 MAC 地址/IP 地址或目的 MAC 地址/IP 地址。
2. 根据 HASH 算法进行计算，得到 HASH-KEY 值。
3. Trunk 模块根据 HASH-KEY 值在 Trunk 转发表中查找对应的接口号。把数据帧从该接口发送出去。

1.4.5 链路聚合协议 LACP

链路聚合的引入

随着以太网技术在城域网和广域网领域的广泛应用，运营商对采用以太网技术的骨干链路的带宽和可靠性提出越来越高的要求。在传统技术中，常用更换高速率的接口板或更换支持高速率接口板的设备的方式来增加带宽，但这种方案需要付出高额的费用，而且不够灵活。采用链路聚合技术可以在不进行硬件升级的条件下，通过将多个物理接口捆绑为一个逻辑接口实现增大链路带宽的目的。在实现增大带宽目的的同时，链路聚合采用备份链路的机制，可以有效的提高设备之间链路的可靠性。

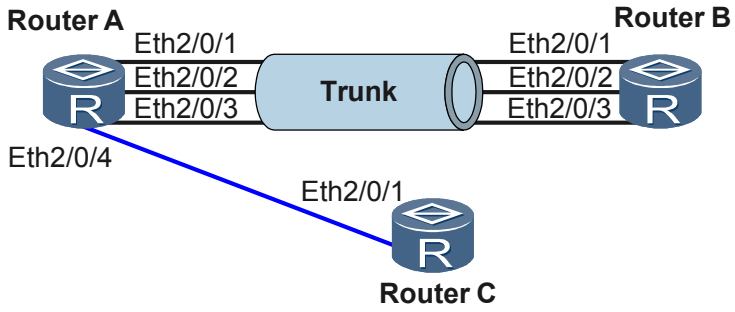
作为链路聚合技术，Trunk 可以完成多个物理端口聚合成一个 Trunk 口来提高带宽，同时能够检测到同一 Trunk 内的成员链路有断路等有限故障，但是无法检测链路层故障、链路错连等故障。LACP（Link Aggregation Control Protocol）的技术出现后，提高了 Trunk 的容错性，并且能提供 M:N 备份功能，保证成员链路的高可靠性。

LACP 为交换数据的设备提供一种标准的协商方式，以供系统根据自身配置自动形成聚合链路并启动聚合链路收发数据。聚合链路形成以后，负责维护链路状态。在聚合条件发生变化时，自动调整或解散链路聚合。

如图 1-4 所示，RouterA 与 RouterB 之间创建 Trunk，需要将 RouterA 上的四个全双工 Eth 接口与 RouterB 捆绑成一个 Trunk。由于错将 RouterA 上的一个 Eth 接口与 RouterC 相连，这将会导致 RouterA 向 RouterB 传输数据时可能会将本应该发到 RouterB 的数据发送到RouterC 上。而 Trunk 不能及时的检测到故障。

如果在 RouterA、RouterB 和 RouterC 上都启用 LACP 协议，经过协商后，RouterA 发送的数据能够正确到达 RouterB。

图 1-4 Trunk 错连示意图



基本概念

- 链路聚合

  将一组物理接口捆绑在一起作为一个逻辑接口来增加带宽及可靠性的方法。

- 链路聚合组

  将若干条物理链路捆绑在一起所形成的逻辑链路称之为链路聚合组（LAG）或者 Trunk。

  如果这些被捆绑链路都是以太网链路，该聚合组被称为以太网链路聚合组，简写为 Eth-Trunk。该聚合组接口称之为 Eth-Trunk 接口。

  组成 Eth-Trunk 的各个接口称之为成员接口。

  Eth-Trunk 接口可以作为普通的以太网接口来使用，它与普通以太网接口的差别只在于：转发的时候 Eth-Trunk 需要从众多成员接口中选择一个或多个接口来进行转发。所以，除了一些必须在物理接口下配置的特性，可以像配置普通以太网接口那样配置 Eth-Trunk 逻辑接口。

  说明

  不能把已有的 Eth-Trunk 成员接口再捆绑成为其它 Eth-Trunk 的成员。

- 活动接口和非活动接口

  链路聚合存在活动接口和非活动接口两种。转发数据的接口称为活动接口，而不转发数据的接口称为非活动接口。

  活动接口对应的链路称为活动链路，非活动接口对应的链路称为非活动链路。

  在链路聚合中为了提高链路的可靠性，引入了备份链路的机制。而这些备份链路对应的接口通常情况下担当了非活动接口的角色，只有当前活动接口出现故障时，备份的接口才可以由非活动接口转变为活动接口。

- 活动接口数上限阈值

  在 Eth-Trunk 中，如果配置了活动接口数上限阈值，当活动接口数达到这个值后，再向 Eth-Trunk 中添加成员接口，不会增加 Eth-Trunk 活动接口的数目。

- 活动接口数下限阈值

  设置活动接口数下限阈值主要目的是保证 Eth-Trunk 链路的带宽。防止由于活动接口数目过少而使这些链路负载过大，出现传输数据丢包的情况。

  在 Eth-Trunk 中，如果配置了活动接口数下限阈值，当活动接口数目低于该值时，Eth-Trunk 接口状态将变为 Down，此时所有 Eth-Trunk 中的成员接口不再转发数据。

- 系统 LACP 优先级

  系统 LACP 优先级是为了区分两端设备优先级的高低而配置的参数。静态 LACP 模式下，两端设备所选择的活动接口必须保持一致，否则链路聚合组就无法建立。而要想使两端活动接口保持一致，可以使其中一端具有更高的优先级，另一端根据高优先级的一端来选择活动接口即可。

  系统 LACP 优先级值越小优先级越高，缺省情况下，系统 LACP 优先级值为 32768。

- 接口 LACP 优先级

  接口 LACP 优先级是为了区别不同接口被选为活动接口的优先程度。接口 LACP 优先级值越小，优先级越高。

- 成员端口间 M:N 备份

  静态 LACP 模式链路聚合是一种利用 LACP 协议进行参数协商选取活动链路的聚合模式。该模式由 LACP 协议确定聚合组中的活动和非活动链路，又称为 M:N 模式，即 M 条活动链路与 N 条备份链路的模式。这种模式提供了更高的链路可靠性，并且可以在 M 条链路中实现不同方式的负载均衡。

  如图 1-5 所示，两台设备间有 M+N 条属性相同的链路，在聚合链路上发送流量时在 M 条链路上负载分担，即主链路。不在另外的 N 条链路发送流量，这 N 条链路提供备份功能，即备份链路。此时链路的实际带宽为 M 条链路的总和，但是能提供的最大带宽为 M+N 的总和。

  当 M 条链路中有一条链路故障时，LACP 会从 N 条备份链路中找出一条正常链路替换有故障的链路，形成 M:N 备份。此时链路的实际带宽还是 M 条链路的总和，但是能提供的最大带宽就变为 M+N-1 条链路的总和。

图 1-5 M:N 备份示意图

这种场景主要应用在我们只想向用户提供 M 条链路的带宽，同时又希望提供一定的故障保护能力。当有一条链路出现故障时，系统能够自动选择一条优先级最高且可以使用的链路加到当前的聚合组中。

如果在备用链路中无法找到可以激活的链路，并且目前处于 Up 状态的链路数目低于配置活动接口数下限阈值，系统将会把汇聚端口关闭。

链路的聚合方式

链路聚合根据是否启用链路聚合控制协议分为以下两种类型：

- 手工负载分担模式链路聚合

  手工负载分担模式是一种最基本的链路聚合方式，在该模式下，Trunk 的建立，成员接口的加入，以及哪些接口作为活动接口完全由手工来配置，没有链路聚合控制协议的参与。

- 静态 LACP 模式链路聚合

  静态 LACP 模式下，Trunk 的建立，成员接口的加入，都是由手工配置完成的。但与手工负载分担模式链路聚合不同的是，该模式下 LACP 协议报文负责活动接口的选择。也就是说，当把一组接口加入 Trunk 后，这些成员接口中哪些接口作为活动接口，哪些接口作为非活动接口还需要经过 LACP 协议报文的协商确定。

  手工负载分担模式链路聚合与静态 LACP 模式链路聚合的区别和相同点如表 1-2 所示。

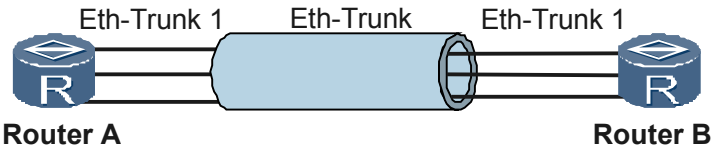
表 1-2 手工模式链路聚合与静态模式链路聚合比较

	手工负载分担模式链路聚合	静态 LACP 模式链路聚合
区别	不启用 LACP 不会进行判断一个聚合组中的端口是否可以真的被聚合在一起。	启用 LACP 通过 LACP 协议来判断在一个聚合组中的端口是否可以真的被聚合在一起。
相同点	聚合组的创建与删除以及成员链路的加入与退出都是由手动配置	

手工负载分担链路聚合原理

手工负载分担模式链路聚合是应用比较广泛的一种链路聚合，允许在聚合组中手工加入多个成员接口，所有的接口均处于转发状态，分担负载的流量。当需要在两个直连设备间提供一个较大的链路带宽而对端设备又不支持 LACP 协议时，可以使用手工负载分担模式。如图 1-6 所示，RouterA 支持 LACP，RouterB 不支持 LACP。

图 1-6 手工负载分担模式组网图



该模式下的所有接口参与数据的转发，并且在所有成员接口上分担负载。AR3200 支持两种方式的负载分担：

- 根据 IP 报文进行负载分担。
- 根据 MAC 地址进行负载分担。

静态 LACP 模式实现原理

基于 IEEE802.3ad 标准的 LACP（链路汇聚控制协议）是一种实现链路动态聚合与解聚合的协议。LACP 协议通过 LACPDU（Link Aggregation Control Protocol Data Unit）与对端交互信息。

在静态 LACP 模式的 Trunk 中加入成员接口后，这些接口将通过发送 LACPDU 向对端通告自己的系统优先级、系统 MAC、接口优先级、接口号和操作 Key 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。

LACPDU 报文详细信息如图 1-7 所示。

图 1-7 LACPDU 报文

Destination Address
Source Address
Length/Type
Subtype=LACP
Version Number
TLV_type=Actor Information
Actor_Information_Length=20
Actor_Port
Actor_State
Actor_System_Priority
Actor_System
Actor_Key
Actor_Port_Priority
Reserved
TLV_type=Partner Information
Partner_Information_Length=20
Partner_Port
Partner_State
Partner_System_Priority
Partner_System
Partner_Key
Partner_Port_Priority
Reserved
TLV_type=Collector Information
Collector_Information_Length=16
CollectorMaxDelay
Reserved
TLV_type=Terminator
Terminator_Length=0
Reserved
FCS

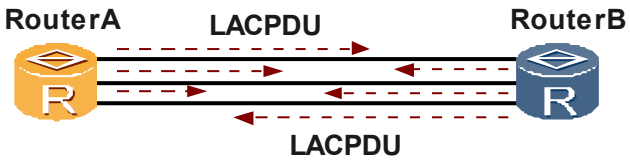
如图 1-7 所示，主要字段信息解释如下：

- Actor_Port/Partner_Port：本端/对端接口信息。
- Actor_State/Partner_State：本端/对端状态。
- Actor_System_Priority/Partner_System_Priority：本端/对端系统优先级。
- Actor_System/Partner_System：本端/对端系统 ID。
- Actor_Key/Partner_Key：本端/对端操作 Key。
- Actor_Port_Priority/Partner_Port_Priority：本端/对端接口优先级。
- 静态模式 Eth-Trunk 建立的过程如下：

1. 两端互相发送 LACPDU 报文。

   如图 1-8 所示，在设备 RouterA 和 RouterB 上创建 Eth-Trunk 并配置为静态 LACP 模式，然后向 Eth-Trunk 中手工加入成员接口。此时成员接口上便启用了 LACP 协议，两端互相发出 LACPDU 报文。

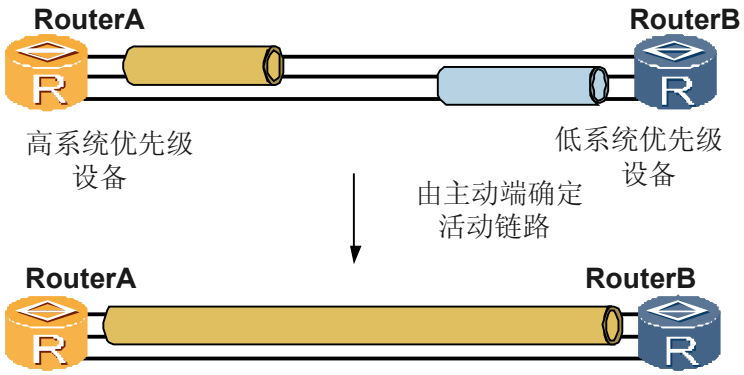
   图 1-8 静态 LACP 模式链路聚合互发 LACPDU



2. 两端设备根据系统 LACP 优先级和系统 ID 确定主动端。

   如图 1-9 所示，两端设备均会收到对端发来的 LACP 报文。以 RouterB 为例，当 RouterB 收到 RouterA 发送的 LACP 报文时，RouterB 会查看并记录对端信息，并且比较系统优先级字段，如果对端设备 RouterA 的系统优先级高于本端设备 RouterB 的系统优先级，则确定 RouterA 为 LACP 主动端，RouterB 将按照 RouterA 的接口优先级选择活动接口，从而两端设备对于活动接口的选择达成一致。

   图 1-9 确定静态 LACP 模式主动端

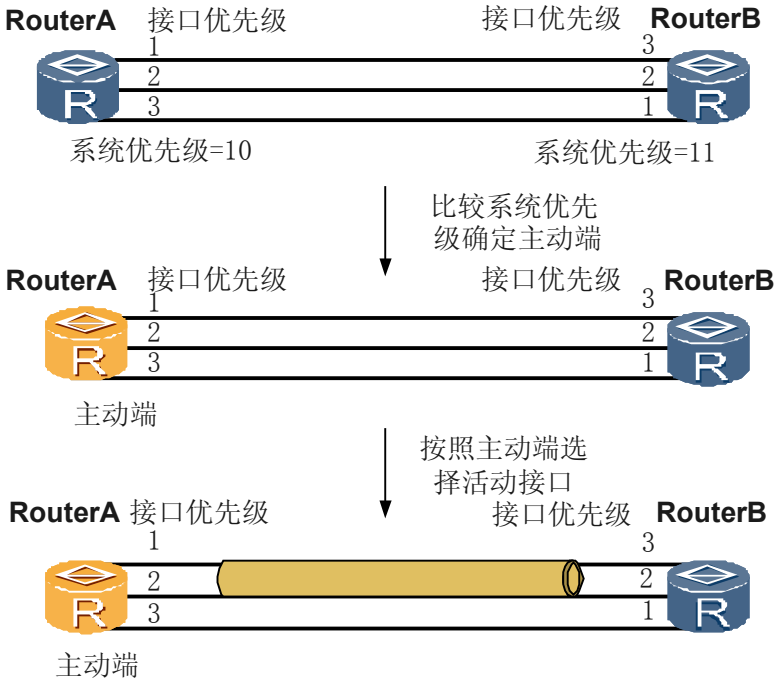


3. 两端设备根据主动端接口 LACP 优先级和接口 ID 确定活动接口。

   如图 1-10 所示，两端设备选出主动端后，两端都会以主动端的接口优先级来选择活动接口。

两端设备选择了一致的活动接口，活动链路组便可以建立起来，从这些活动链路中转发数据。

图 1-10 静态 LACP 模式选择活动接口的过程



- 活动链路与非活动链路切换

  静态模式链路聚合组两端设备中任何一端检测到以下事件，都会触发聚合组的链路切换：

  - 链路 Down 事件。
  - LACP 协议发现链路故障。
  - 接口不可用。
  - 在使能了 LACP 抢占前提下，更改备份接口的优先级高于当前活动接口的优先级后，会发生切换的过程。

  当满足上述切换条件其中之一时，按照如下步骤进行切换：

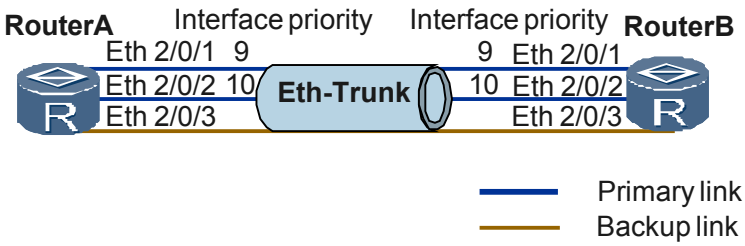
  1. 关闭故障链路。
  2. 从 N 条备份链路中选择优先级最高的链路接替活动链路中的故障链路。
  3. 优先级最高的备份链路转为活动状态并转发数据，完成切换。

- LACP 抢占

  使能 LACP 抢占后，聚合组会始终保持高优先级的接口作为活动接口的状态。

  如图 1-11 所示，Eth2/0/1、Eth2/0/2 和 Eth2/0/3 为 Eth-Trunk 1 的成员接口，活动接口数最大上限阈值为 2，配置 Eth2/0/1 和 Eth2/0/2 接口的 LACP 优先级分别为 9 和 10，Eth2/0/3 保持缺省接口 LACP 优先级。当通过 LACP 协议协商完毕后，Eth2/0/1、Eth2/0/2 接口因为优先级较高被选作活动接口，Eth2/0/3 接口成为备份接口。

图 1-11 LACP 抢占场景



以下两种情况需要使能 LACP 的抢占功能。

- Eth2/0/1 接口出现故障而后又恢复了正常。当接口 Eth2/0/1 出现故障时被 Eth2/0/3 所取代，如果在 Eth-Trunk 接口下未使能抢占，则故障恢复时 Eth2/0/1 仍然保持备份接口状态；如果使能了 LACP 抢占，当 Eth2/0/1 故障恢复时可以重新成为活动接口，Eth2/0/3 再次成为备份接口。
- 如果用户希望 Eth2/0/3 接口替换 Eth2/0/1、Eth2/0/2 中的一个接口成为活动接口，可以通过更改 Eth2/0/3 的接口 LACP 优先级为 8 或更小的数值来实现，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，不切换活动接口。

- 抢占延时

  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。抢占延时是一个可配置的值，默认为 30s，可配置范围为 10s～180s。

  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整个 Eth-Trunk 数据传输不稳定。

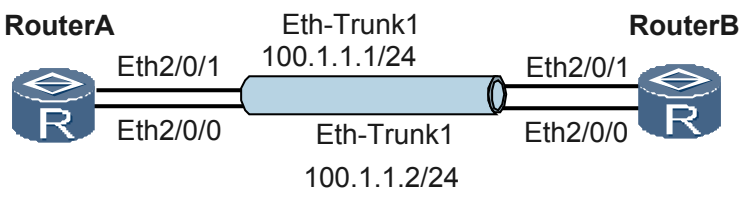
  如图 1-11 所示，Eth2/0/1 由于链路故障切换为非活动接口，此后该链路又恢复了正常。由于系统使能了 LACP 抢占，经过抢占延时后，Eth2/0/1 会重新切换到活动状态。

1.5 应用

1.5.1 Eth-Trunk

如图 1-12 所示，RouterA 与 RouterB 之间创建 Eth-Trunk，将两个全双工 Eth 接口捆绑成一个 Eth-Trunk。RouterA 与 RouterB 之间的 Trunk 链路最大带宽达到原 Eth 接口的两倍。

图 1-12 Eth-Trunk 组网图



在 Eth-Trunk 内启用组内备份，当其中一条链路故障，流量切换到另一条链路，提高链路可靠性。

Eth-Trunk 接口将 RouterA 和 RouterB 之间的流量分担到两条链路上，避免所有流量都走同一条链路而导致网络拥塞。

1.6 术语与缩略语

术语

术语	解释
LA	Link Aggregation——链路聚合，是指将一组物理端口捆绑在一起作为一个逻辑端口来增加带宽的一种方法。
LACP	Link Aggregation Control Protocol——链路聚合控制协议是为交换数据的设备提供的一种标准的协商方式。
BFD	Bi-directional Forwarding Detection－－双向转发检测是一套全网统一的检测机制，用于快速检测、监控网络中链路或者 IP 路由的转发连通状况。为改善网络性能，相邻系统之间应能快速检测到通信故障，更快地建立起备用通道恢复通信。
ETH-OAM	Ethernet-Operation Administration Maintenance－－以太网操作、管理、维护。

缩略语

缩略语	英文全称	中文全称
LACP	Link Aggregation Control Protocol	链路聚合控制协议
LA	Link Aggregation	链路聚合
BFD	Bi-directional Forwarding Detection	双向转发检测
ETH-OAM	Ethernet-Operation Administration Maintenance	以太网操作、管理、维护

2 透明网桥

关于本章

2.1 介绍

定义

透明网桥用于连接物理介质类型相同的局域网，主要应用在以太网环境中。由于网络中网桥的加入以及转发行为对于网络用户是透明的，对现存网络中的软硬件不会造成影响，因此称之为“透明”网桥。透明网桥通过学习收到报文的源地址并建立源地址与接口的映射关系表来学习网络拓扑，并用于指导报文转发。

目的

随着各种局域网技术的发展，以太类型局域网以其良好的网络伸缩性以及低成本优势逐步占据了局域网技术统治地位。在此背景下，不同局域网之间的互连互通的需求被提出。

通过传统的路由器，进行局域网互联的方式由于其高成本以及配置复杂等缺点，难以满足以太局域网互联的需要。

透明网桥不会影响现存的局域网，以其简单易用以及成本较低的优势得到了广泛的应用。透明网桥提供本地桥接和远程桥接功能。本地桥接将处于同一区域的不同局域网段通过透明网桥直接连接起来，远程桥接连接不同区域的局域网段，通常需要利用其他网络连接协助完成接收来自各局域网发送的报文，并将他们送到目的局域网。透明网桥充分利用了用户现存的非以太低速链路实现了局域网间的互联。

2.2 参考标准和协议

本特性的参考资料清单如下：

文档	描述	备注
RFC 1483	Multiprotocol Encapsulation over ATM Adaptation Layer 5	-
RFC 1490	Multiprotocol Interconnect over Frame Relay	-
RFC 3518	Point-to-Point Protocol (PPP) Bridging Control Protocol (BCP)	-
IEEE 802.1D	Media Access Control (MAC) Bridges	-
IEEE 802.1G	Remote Media Access Control (MAC) Bridges	-

2.3 可获得性

涉及网元

无需其它网元的配合。

License 支持

无需获得 License 许可，均可获得该特性的服务。

版本支持

产品	最低支持版本
AR3200	V200R001C00

2.4 原理描述

2.4.1 透明网桥的基本原理

转发表项学习

透明网桥需要根据转发表指导转发，网桥的转发表中表项记录链路层地址与对应该链路层地址的转发出接口的映射关系，即 MAC 地址与出接口的映射关系。其具体过程为：

- 对于检测到合法的以太网帧，提取出该帧的源 MAC 地址。
- 将源 MAC 地址与接收该帧的接口之间的关系加入到地址表中，从而生成一条表项。

对于同一个 MAC 地址，如果透明网桥先后学习到不同的接口，则后学到的接口信息覆盖先学到的接口信息，因此，不存在同一个 MAC 地址对应两个或更多出接口的情况。

对于动态学习到的转发表项，透明网桥会在一段时间后对表项进行老化，即将超过一定生存时间的表项删除掉。系统支持默认的老化时间（300s），用户可以自行设置老化时间。

报文转发处理

透明网桥对于收到数据帧的处理可以划分为以下两种情况：

- 单播

  收到数据帧的目的 MAC 能够在转发表中查到，并且对应的出接口与收到报文的接口不是同一个接口，则该数据帧从表项对应的出接口转发出去。

- 广播

  收到数据帧的目的 MAC 是单播 MAC，但是在转发表中查找不到，或者收到数据帧的目的 MAC 是组播或广播 MAC 时，数据帧向对应网桥组除入接口外的其他接口复制并发送。

2.4.2 透明网桥的本地桥接功能

透明网桥可以创建不同的网桥组，加入到特定网桥组的接口，流量在桥组内基于目的 MAC 地址进行转发。

一般情况下，网桥地址表根据该网桥获取的 MAC 地址和接口的对应关系动态生成。当用户网络环境比较恶劣（如网络安全性能较差，容易受到攻击），管理员也可以手工配置一些静态地址表项或黑洞 MAC 表项，并且永远不会老化。

动态地址表的老化时间是指该表项从地址表中删除之前的生存时间，动态地址表项在地址表中保持的时间超过老化时间后，系统就将该表项从网桥地址表中删除。

透明网桥的本地桥接成功配置后：

- 网桥组默认支持学习 MAC 地址与接口的映射关系，即 MAC 转发表项。
- 网桥组支持配置静态以及黑洞 MAC 表项。
- 网桥组支持动态 MAC 表项学习使能与禁止。
- 支持配置动态 MAC 表项老化时间。
- 缺省支持对所有报文的桥接功能，支持配置对 IP 以及非 IP 报文的桥接功能。

2.4.3 透明网桥的远程桥接功能

当通过透明网桥连接的局域网处于不同的地理位置时，网桥间需要使用远程桥接的功能连接不同地点的透明网桥设备，中间的连接网络可能是以太或者非以太网络，远程桥接功能主要描述的是通过非以太中间链路进行连接的情况，以太中间链路的情况类似本地桥接。

为支持远程桥接功能，透明网桥提供了如下功能：

- 支持以太主接口、以太子接口、VLANIF、VT、Dialer、Serial、ATM 接口、ATM 子接口、FR 接口、FR 子接口、MFR 接口、MLPP 接口加入网桥组；
- 支持以太、PPP、HDLC、FR、ATM 等链路封装协议；
- 支持 802.1Q VLAN ID 透明传输；
- 支持桥接 IP 和非 IP 报文。

2.4.4 透明网桥的集成路由桥接功能

网桥路由功能提供了一种结合路由和桥接的转发方法：

- 对于指定的协议数据，如果是在网桥端口之间进行通讯，则进行桥接转发；
- 如果是需要与非网桥组内的网络进行通讯，则可以进行网络协议的路由转发。

当集成路由功能没有激活时，所有的协议数据只能够进行桥接处理。当集成的路由功能和桥接功能被使能后，就可以指定某种协议的报文既可以做桥接或进行路由处理，通过命令配置进行灵活的切换。

桥组虚接口是一个虚拟的选路接口，可以配置各种网络层的属性。对于指定的协议数据，网桥端口之间只能进行网桥组内的桥接转发，如果不同网段的局域网之间需要进行通信，则需要进行网络协议的路由转发。在网桥组上创建一个网桥组虚接口（Bridge-if），并配置相应的 MAC 地址和 IP 地址，使能透明网桥的集成路由功能并配置静态路由后可以实现通信数据的路由转发。

对于每个网桥组来说，只能有一个桥组虚接口。桥组虚接口的编号是它所代表的网桥组的编号。缺省情况下，桥组虚接口将使用系统默认的 MAC 地址，也可以手动配置网桥组虚接口的 MAC 地址。

2.4.5 透明网桥的透传 VLAN ID 功能

通过对加入网桥组的设备出接口配置支持 VLAN ID 透传，可以使报文从该接口送出时，不对报文的 VLAN ID 做任何修改。使能桥出接口 VLAN ID 透传，则报文从该接口发出时保留报文入桥时的 VLAN ID。

透明网桥在使能 VLAN ID 透传后：

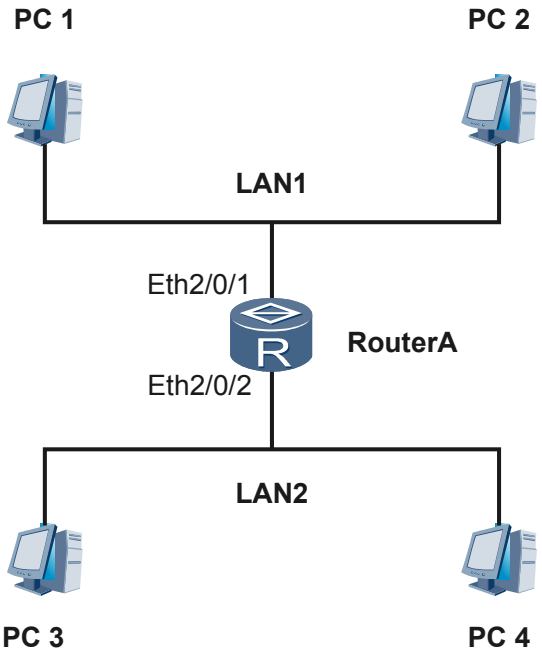
- 网桥不会对报文的 VLAN ID 进行任何的修改和去除等操作，从而可以实现 VLAN ID 的透明传输，保证不同 VLAN 之间的隔离、同一 VLAN 之间的互通。
- 加入网桥组的非以太出接口也能转发带有 VLAN ID 的报文，而不会因此丢失 VLAN ID，并且即使加入桥组设备的出接口上有 VLAN ID 的情况下，也不会改变报文原有的 VLAN ID，从而实现不同 VLAN 的隔离。
- 系统不对报文的 VLAN ID 做任何处理，与透明网桥设备两端相连的交换机可以看成是直连的。为了正常通信，用户需要给两端交换机的 Trunk 口配置相同的 VLAN ID。

2.5 应用

2.5.1 本地桥接

本地桥接是透明网桥提供的最基本功能。如图 2-1 所示，处于同一地理位置的多个局域网（图中 LAN 1、LAN 2）需要在链路层实现互通，可以通过透明网桥设备对这些局域网实现本地桥接。

图 2-1 透明网桥本地桥接应用

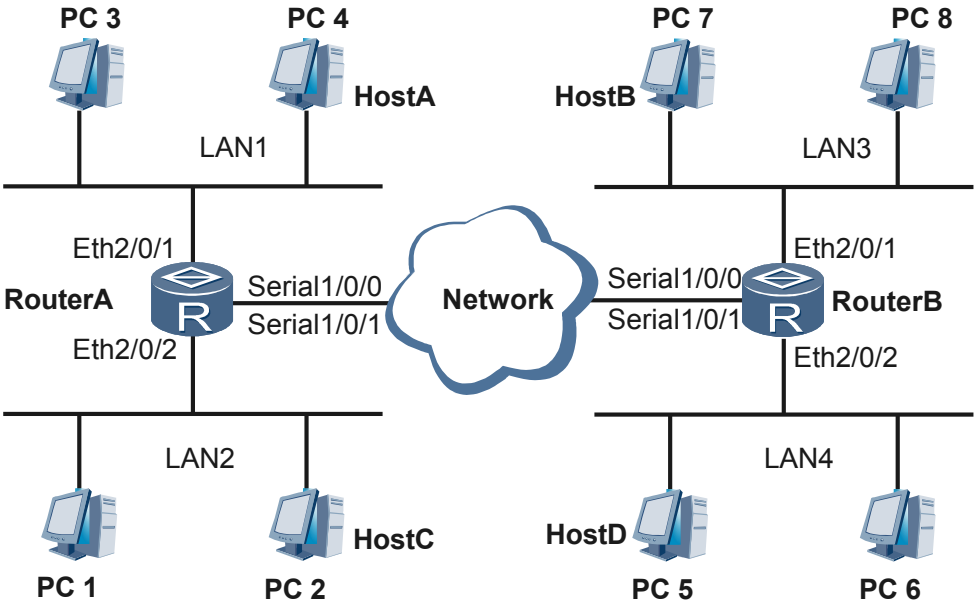


在 RouterA（透明网桥设备）上创建网桥组，并将连接局域网的接口（Ethernet2/0/1、Ethernet2/0/2）加入到网桥组中，从而实现局域网的链路层互通。

2.5.2 远程桥接

当有链路层互通需求的局域网分布在不同的地理位置时，可以使用透明网桥远程桥接功能对多个局域网进行桥接。

图 2-2 透明网桥远程桥接应用



如图 2-2 所示，RouterA 和 RouterB 之间跨越帧中继网络，PC2、PC4、PC5、PC7 分别属于 4 个不同的局域网段，要求 PC4 所在的局域网与 PC7 所在的局域网互通、PC2 所在的局域网与 PC5 所在的局域网互通。

可以在 RouterA 和 RouterB 上分别创建两个网桥组 bridge 1、bridge 2，并且分别将 Ethernet2/0/1 和帧中继接口 Serial1/0/0 加入到 bridge 1 中，分别将 Ethernet2/0/2 和帧中继接口 Serial1/0/1 加入到 bridge 2 中，即可实现上述的互通功能要求。注意上述帧中继接口如果是点对点（P2P）类型的，需要在 DCE 和 DTE 两侧同时配置相同的 FR DLCI；如果是非点对点类型的，则需要配置 FR DLCI 到网桥的映射。

透明网桥还可以通过其他类型链路实现远程连接，如 Eth、PPP、ATM、HDLC 等链路。

2.5.3 集成路由桥接

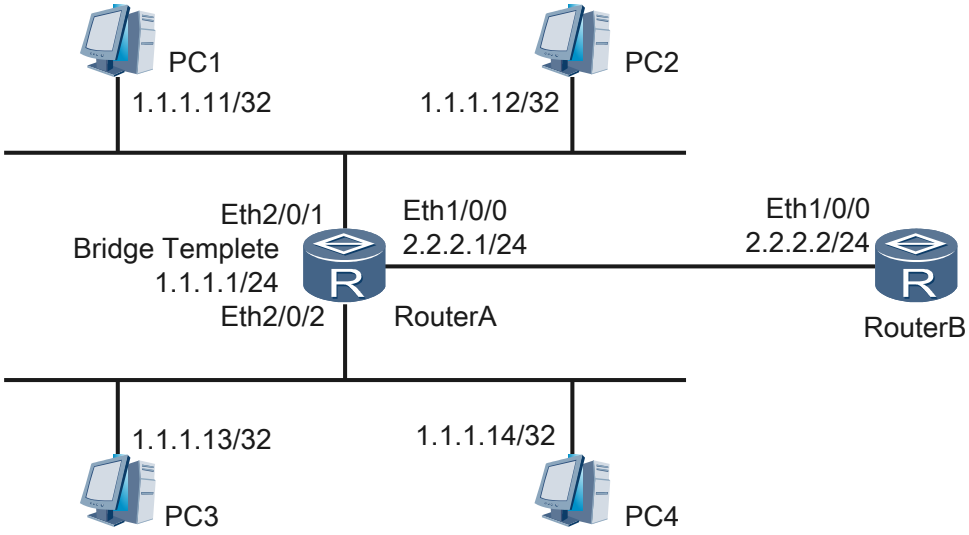
网桥组提供了不同局域网段的链路层连接功能，通常情况下，需要相互连接的局域网用户在网络层属于同一网段或者聚合网段。当网桥组内用户需要访问网桥组外网络时，仅仅通过链路层桥接无法满足需求，透明网桥集成路由桥接功能可以满足上述功能需求。

集成路由功能通过为桥组创建一个网桥组虚接口（Bridge-if）来实现。路由虚接口可以配置网络层属性，比如可以配置 IP 地址。每个网桥组只能配置一个虚接口，网桥组虚接口的编号是它所代表的网桥组的编号。在透明网桥集成路由桥接功能被激活后，Bridge-if 接口即可对桥组内与桥组外的网络之间的通信数据进行路由转发。

集成路由桥接功能需要通过配置激活，否则所有的桥组内数据只能够进行桥接转发。集成路由桥接功能被激活后，就可以指定某种协议的报文根据需要进行路由处理，通过命令配置进行灵活的切换。

使能集成路由桥接功能后，加入网桥组的接口不支持配置 IP 地址。

图 2-3 透明网桥集成路由桥接应用



如图 2-3 所示，RouterA 上配置网桥组以及网桥组虚接口，连接两个不同局域网段的接口 Ethernet2/0/1、Ethernet2/0/2 加入到桥组，桥组虚接口配置 IP 地址。使能网桥集成路由桥接功能并配置桥组的 IP 协议数据路由使能并配置 RouterB 的回程路由后，连接到桥组的四台主机即可以通过桥组虚接口与桥组外的网络进行 IP 路由通信。

2.5.4 VLAN ID 支持透明传输

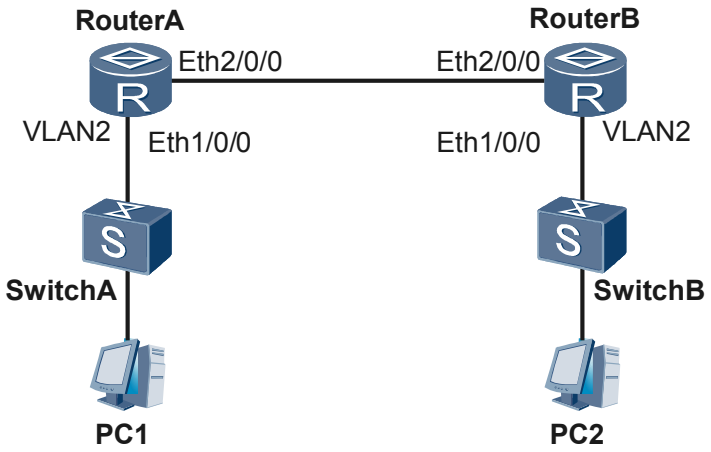
透明网桥对多个局域网进行桥接时，报文 VLAN ID 需要在局域网间传输，以达到互联的不同局域网间 VLAN 间隔离、VLAN 内互通的目的。这时可以使能 VLAN ID 透明传输功能，避免在穿越的过程遇到不支持 VLAN ID 传递时将 VLAN ID 丢弃的情况发生。

网桥透明传输 VLAN ID 功能需要在桥组接口上使能，使能后接口对所经过以太报文 VLAN ID 不作任何操作。

接口上配置 VLAN ID 透明传输功能前，必须将该接口加入到桥组中。

VLANIF 不支持 VLAN ID 透明传输功能，以太子接口不建议使用。

图 2-4 VLAN ID 透明传输应用



两端 TRUNK 接口之间需要穿越以太网连接时，配置网桥透传 VLAN ID 功能将使所经过以太网中设备对输入报文 VLAN ID 不作任何操作。这样相当于将两台设备的 TRUNK 接口直接相连，避免在穿越的过程遇到不支持 VLAN ID 传递时将 VLAN ID 丢弃的情况发生。如图 2-4 所示，在 RouterA 和 RouterB 上配置接口使能 VLAN ID 透明传输功能，配置完成后 PC1 和 PC2 可以实现通信。

2.6 术语与缩略语

术语

无

缩略语

缩略语	英文全称	中文全称
ARP	Address Resolution Protocol	地址解析协议
LAN	Local Area Network	局域网
MAC	Media Access Control	媒体接入控制
TB	Transparent Bridge	透明网桥

3 VLAN

关于本章

3.1 介绍

定义

VLAN（Virtual Local Area Network）即虚拟局域网，是将一个物理的 LAN 在逻辑上划分成多个广播域（多个 VLAN）的通信技术。VLAN 内的主机间可以直接通信，而 VLAN 间不能直接互通，从而将广播报文限制在一个 VLAN 内。由于 VLAN 间不能直接互访，因此提高了网络安全性。

目的

早期的局域网 LAN 技术是基于总线型结构，它存在以下主要问题：

- 若某时刻有多个节点同时试图发送消息，那么它们将产生冲突。
- 从任意节点发出的消息都会被发送到其他节点，形成广播。
- 所有主机共享一条传输通道，无法控制网络中的信息安全。

这种网络构成了一个冲突域，网络中计算机数量越多冲突越严重，网络效率越低。同时，该网络也是一个广播域，当网络中发送信息的计算机数量越多时，广播流量将会耗费大量带宽。

因此，传统网络不仅面临冲突域和广播域两大难题，而且无法保障传输信息的安全。

为了扩展传统 LAN，以接入更多计算机，同时避免冲突的恶化，出现了网桥和二层交换机，它们能有效隔离冲突域。

Bridge 和交换机采用交换方式将来自入端口的信息转发到出端口上，克服了共享介质上的访问冲突问题，从而将冲突域缩小到端口级。采用交换机进行组网，通过二层快速交换解决了冲突域问题，但是广播域和信息安全问题依旧存在。

说明

本手册中将二层局域网交换机简称为交换机。

为减少广播，需要在没有互访需求的主机之间进行隔离。路由器是基于三层 IP 地址信息来选择路由，其连接两个网段时可以有效抑制广播报文的转发，但成本较高。因此人们设想在物理局域网上构建多个逻辑局域网，即 VLAN（Virtual Local Area Network）。

VLAN 将一个物理的 LAN 在逻辑上划分成多个广播域（多个 VLAN）。VLAN 内的主机间可以直接通信，而 VLAN 间不能直接互通。这样，广播报文被限制在一个 VLAN 内，同时提高了网络安全性。

例如，同一个写字楼的不同企业客户，若建立各自独立的 LAN，企业的网络投资成本将很高；若共用写字楼已有的 LAN，又会导致企业信息安全无法保证。

采用 VLAN，可以实现各企业客户共享 LAN 设施，同时保证各自的网络信息安全。

图 3-1 VLAN 的典型应用示意图

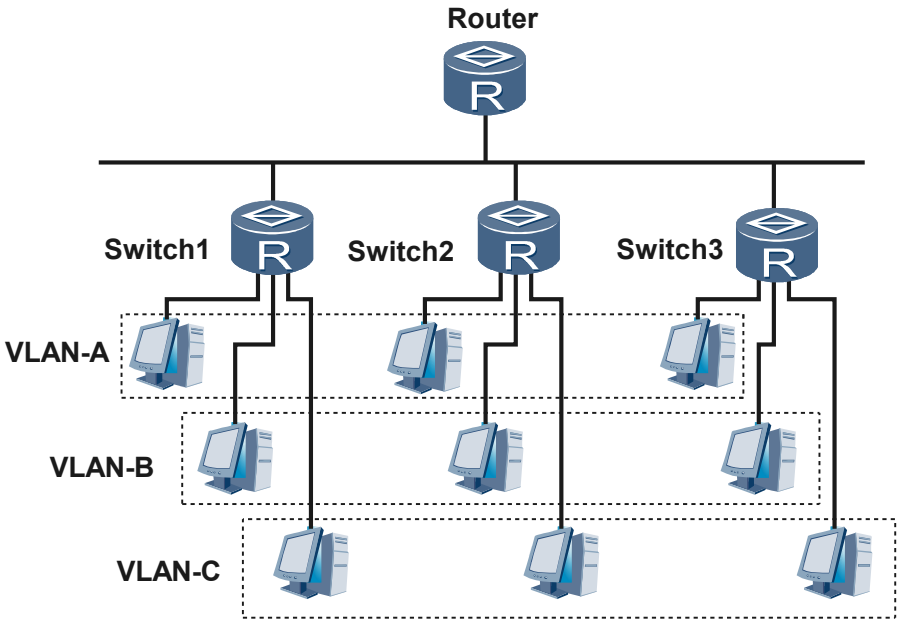


图 3-1 是一个典型的 VLAN 应用组网图。3 台交换机放置在不同的地点，比如写字楼的不同楼层。每台交换机分别连接 3 台计算机，他们分别属于 3 个不同的 VLAN，比如不同的企业客户。在图中，一个虚线框内表示一个 VLAN。

3.2 参考标准和协议

本特性的参考资料清单如下：

文档	描述	备注
RFC 3069	VLAN Aggregation for Efficient IP Address Allocation	
IEEE 802.1Q	IEEE Standards for Local and Metropolitan Area Networks: Virtual Bridged Local Area Networks	
IEEE 802.1ad	IEEE Standards for Local and Metropolitan Area Networks: Virtual Bridged Local Area Networks— Amendment 4	

文档	描述	备注
IEEE 802.10	IEEE Standards for Local and Metropolitan Area Networks: Standard for Interoperable LAN/ MAN Security	
YD/T 1260-2003	Technical and Testing Specification of Virtual LAN Based on Port	

3.3 可获得性

涉及网元

无需其它网元的配合。

License 支持

无需获得 License 许可，均可获得该特性的服务。

版本支持

产品	最低支持版本
AR3200	V200R001C00

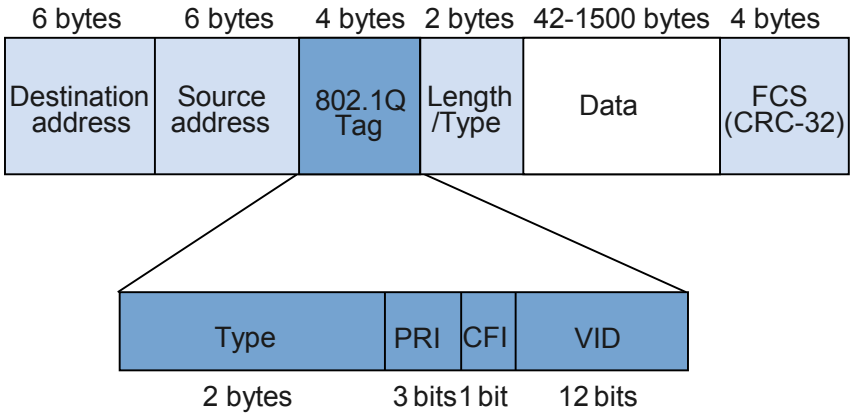
3.4 原理描述

3.4.1 VLAN 基本概念

VLAN 的帧格式

IEEE 802.1Q 标准对 Ethernet 帧格式进行了修改，在源 MAC 地址字段和协议类型字段之间加入 4 字节的 802.1Q Tag，如图 3-2 所示。

图 3-2 基于 802.1Q 的 VLAN 帧格式



802.1Q Tag 包含 4 个字段，其含义如下：

- Type

  长度为 2 字节，表示帧类型。取值为 0x8100 时表示 802.1Q Tag 帧。如果不支持 802.1Q 的设备收到这样的帧，会将其丢弃。

- PRI

  Priority，长度为 3 比特，表示帧的优先级，取值范围为 0 ～ 7，值越大优先级越高。用于当交换机阻塞时，优先发送优先级高的数据帧。

- CFI

  Canonical Format Indicator，长度为 1 比特，表示 MAC 地址是否是经典格式。CFI 为 0 说明是经典格式，CFI 为 1 表示为非经典格式。用于区分以太网帧、FDDI（Fiber Distributed Digital Interface）帧和令牌环网帧。在以太网中，CFI 的值为 0。

- VID

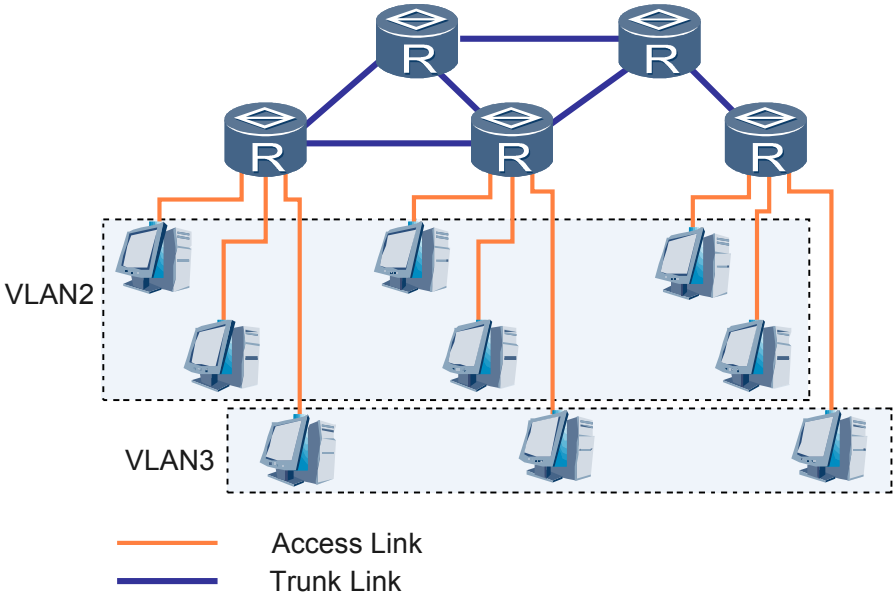
  VLAN ID，长度为 12 比特，表示该帧所属的 VLAN。在 Huawei AR3200 系列中，可配置的 VLAN ID 取值范围为 0 ～ 4095，但是 0 和 4095 协议中规定为保留的 VLAN ID，不能给用户使用

链路类型

VLAN 内的链路包括：

- 接入链路（Access Link）：连接用户主机和交换机的链路为接入链路。如图 3-3 所示，图中 PC 机和交换机之间的链路都是接入链路。接入链路上通过的帧为不带 Tag 的以太网帧。
- 干道链路（Trunk Link）：连接交换机和交换机的链路称为干道链路。如图 3-3 所示，图中交换机之间的链路都是干道链路。干道链路上通过的帧一般为带 Tag 的 VLAN 帧。

图 3-3 链路类型示意图



端口类型

在 802.1Q 中定义 VLAN 帧后，设备的有些端口可以识别 VLAN 帧，有些端口则不能识别 VLAN 帧。根据对 VLAN 帧的识别情况，将端口分为 3 类：

- Access 端口

  如图 3-3 所示，Access 端口是交换机上用来连接用户主机的端口，它只能连接接入链路。有如下特点：

  - 仅仅允许唯一的 VLAN ID 通过本端口，这个 VLAN ID 与端口的 PVID（Port VLAN ID，端口缺省的 VLAN ID）相同。
  - 如果该端口收到的对端设备发送的帧是 untagged（不带 VLAN 标签），交换机将强制加上该端口的 PVID。
  - Access 端口发往对端设备的以太网帧永远是不带标签的帧。

- Trunk 端口

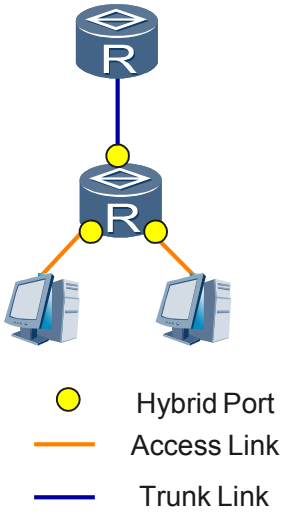
  如图 3-3 所示，Trunk 端口是交换机上用来和其他交换机连接的端口，它只能连接干道链路。有如下特点：

  - Trunk 端口允许多个 VLAN 的帧（带 Tag 标记）通过。
  - 如果从 Trunk 端口发送的帧带 Tag，且 Tag 与端口缺省的 VLAN ID 相同，则交换机会剥掉该帧中的 Tag 标记。因为每个端口的 PVID 取值是唯一的。仅在这种情况下，Trunk 端口发送的帧不带 Tag。
  - 如果从 Trunk 端口发送的帧带 Tag，但是与端口缺省的 VLAN ID 不同，则交换机对该帧不做任何动作，直接发送带 Tag 的帧。

- Hybrid 端口

  如图 3-4 所示，Hybrid 端口是交换机上既可以连接用户主机，又可以连接其他交换机的端口。Hybrid 端口既可以连接接入链路又可以连接干道链路。Hybrid 端口允许多个 VLAN 的帧通过，并可以在出端口方向将某些 VLAN 帧的 Tag 剥掉。

图 3-4 端口示意图



缺省 VLAN

在交换设备上，每个 Access、Trunk、Hybrid 类型的端口可以配置一个缺省 VLAN。端口类型不同，缺省 VLAN 的含义也有所不同。

- Access 端口的缺省 VLAN
  - 对于从 Access 端口接收到的不带 Tag 的帧，交换设备会在帧中加上 Tag 标记，并将 Tag 中的 VID 字段的值设置为端口所属的缺省 VLAN 编号。
  - 对于从 Access 端口发送的帧，如果 Tag 中的 VID 值为缺省 VLAN 编号，则交换设备会剥掉该帧中的 Tag 标记。因为 Access 端口发往对端设备的以太网帧永远是不带标签的帧。
- Trunk 端口的缺省 VLAN
  - 对于从 Trunk 端口接收到的不带 Tag 的帧，交换设备会在帧中加上 Tag 标记，并将 Tag 中的 VID 字段的值设置为端口所属的缺省 VLAN 编号。
  - 对于从 Trunk 端口发送的帧：
    - 如果 Tag 中的 VID 值为缺省 VLAN 编号，则交换设备会剥掉该帧中的 Tag 标记。因为每个端口的 PVID 取值是唯一的。
    - 如果 Tag 中的 VID 值与端口缺省的 VLAN 不同，则交换设备对该帧不做任何改变，直接发送带 Tag 的帧。
- Hybrid 端口的缺省 VLAN
  - 对于从 Hybrid 端口接收到的不带 Tag 的帧，交换机会在帧中加上 Tag 标记，并将 Tag 中的 VID 字段的值设置为端口所属的缺省 VLAN 编号。
  - 对于从 Hybrid 端口发送的帧：
    - 如果该端口上已经配置 **port hybrid untagged vlan** 命令，则该端口的功能与 Access 端口功能相同。
    - 如果该端口上没有配置 **port hybrid untagged vlan** 命令，则该端口的功能与 Trunk 端口功能相同。

3.4.2 VLAN 通信原理

VLAN 基础通信原理

为了提高处理效率，交换机内部的数据帧一律都带有 VLAN Tag，以统一方式处理。当一个数据帧进入交换机端口时，如果没有带 VLAN Tag，且该端口上配置了 PVID（Port Default VLAN ID），则该数据帧就会被标记上端口的 PVID。如果数据帧已经带有 VLAN Tag，即使端口已经配置了 PVID，交换机不会再给数据帧标记 VLAN Tag。

由于端口类型不同，交换机对帧的处理过程也不同。下面根据不同的端口类型分别介绍。

端口类型	对接收不带 Tag 的报文处理	对接收带 Tag 的报文处理	发送帧处理过程
Access 端口	接收该报文，并打上缺省 VLAN 的 Tag。	● 当 VLAN ID 与缺省 VLAN ID 相同时，接收该报文。 ● 当 VLAN ID 与缺省 VLAN ID 不同时，丢弃该报文。	先剥离帧的 PVID Tag，然后再发送。
Trunk 端口	● 打上缺省的 VLAN ID，当缺省 VLAN ID 在允许通过的 VLAN ID 列表里时，接收该报文。 ● 打上缺省的 VLAN ID，当缺省 VLAN ID 不在允许通过的 VLAN ID 列表里时，丢弃该报文。	● 当 VLAN ID 在接口允许通过的 VLAN ID 列表里时，接收该报文。 ● 当 VLAN ID 不在接口允许通过的 VLAN ID 列表里时，丢弃该报文。	● 当 VLAN ID 与缺省 VLAN ID 相同，且是该接口允许通过的 VLAN ID 时，去掉 Tag，发送该报文。 ● 当 VLAN ID 与缺省 VLAN ID 不同，且是该接口允许通过的 VLAN ID 时，保持原有 Tag，发送该报文。
Hybrid 端口			当 VLAN ID 是该接口允许通过的 VLAN ID 时，发送该报文。可以通过命令设置发送时是否携带 Tag。

VLAN 内跨越交换机通信原理

有时属于同一个 VLAN 的用户主机被连接在不同的交换机上。当 VLAN 跨越交换机时，就需要交换机间的端口能够同时识别和发送跨越交换机的 VLAN 报文。这时，需要用到 Trunk Link 技术。

Trunk Link 有两个作用：

- 中继作用

  把 VLAN 报文透传到互联的交换机。

- 干线作用

  一条 Trunk Link 上可以传输多个 VLAN 的报文。

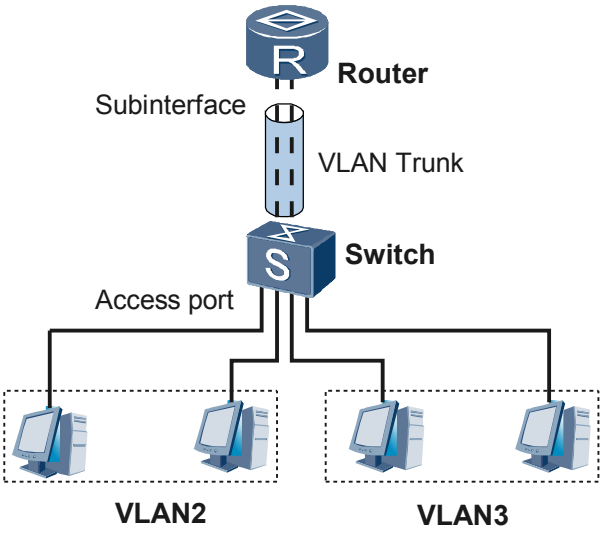
VLAN 间通信原理

划分 VLAN 后，不同 VLAN 的计算机之间不能实现二层通信。如果在 VLAN 间通信，需要建立 IP 路由。有以下实施方案：

- 二层交换机＋路由器

  多数情况下，LAN 通过二层交换机的以太网接口（交换式以太网接口）与路由器的以太网接口（路由式以太网接口）相连，如图 3-5 所示。

  图 3-5 通过二层交换机＋路由器实现 VLAN 间的通信

  假定在交换机上已划分了 VLAN2 和 VLAN3。

  为实现 VLAN2 和 VLAN3 间的通信，需要在路由器与交换机相连的以太网接口上创建 2 个子接口与 VLAN2 和 VLAN3 分别对应。

  在子接口上配置 802.1Q 封装和 IP 地址。

  将交换机与路由器相连的以太网口类型改为 Trunk 或 Hybrid，允许 VLAN2 和 VLAN3 的帧通过。

  二层交换机＋路由器模式的缺点在于：

  - 需要多个设备，组网复杂。
  - VLAN 间通信通过路由器完成，路由器价格昂贵，速率较低。

- 三层交换机

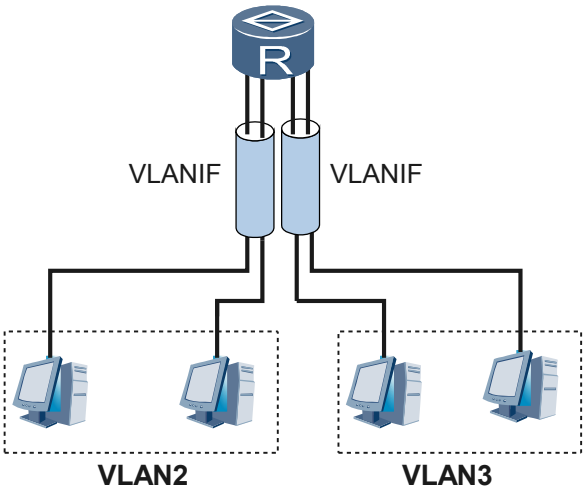
  三层交换技术是将路由技术与交换技术合二为一的技术，在交换机内部实现了路由，提高了网络的整体性能。三层交换机通过路由表传输第一个数据流后，会产生一个 MAC 地址与 IP 地址的映射表。当同样的数据流再次通过时，将根据此表直接从二层通过而不是通过三层，从而消除了路由器进行路由选择而造成的网络延迟，提高了数据包转发效率。

为了保证第一次数据流通过路由表正常转发，路由表中必须有正确的路由表项。因此必须在三层交换机上部署三层接口并部署路由协议，实现三层路由可达。VLANIF 接口由此而产生。

VLANIF 接口是三层逻辑接口，可以部署在三层交换机上，也可以部署在路由器上。

在图 3-6 所示的网络中，交换机上划分了 2 个 VLAN：VLAN2 和 VLAN3。此时可在交换机上创建 2 个 VLANIF 接口，并为它们配置 IP 地址和路由，实现 VLAN2 与 VLAN3 的通信。

图 3-6 通过三层交换机实现 VLAN 间的通信



三层交换机解决了二层交换机＋路由器模式的问题，能够以低廉的成本实现更快速的转发。但是三层交换机也存在以下缺陷：

- 三层交换机适用于几乎全以太网接口的网络。
- 三层交换机适用于路由比较稳定，变化比较少的网络。

3.4.3 VLAN Aggregation

产生背景

交换网络中，VLAN 技术以其对广播域的灵活控制和部署方便而得到了广泛的应用。但是在一般的三层交换机中，通常是采用一个 VLAN 对应一个三层逻辑接口的方式实现广播域之间的互通，这样导致了 IP 地址的浪费。例如，设备内 VLAN 划分如图 3-7 所示。

图 3-7 普通 VLAN 网络示意图

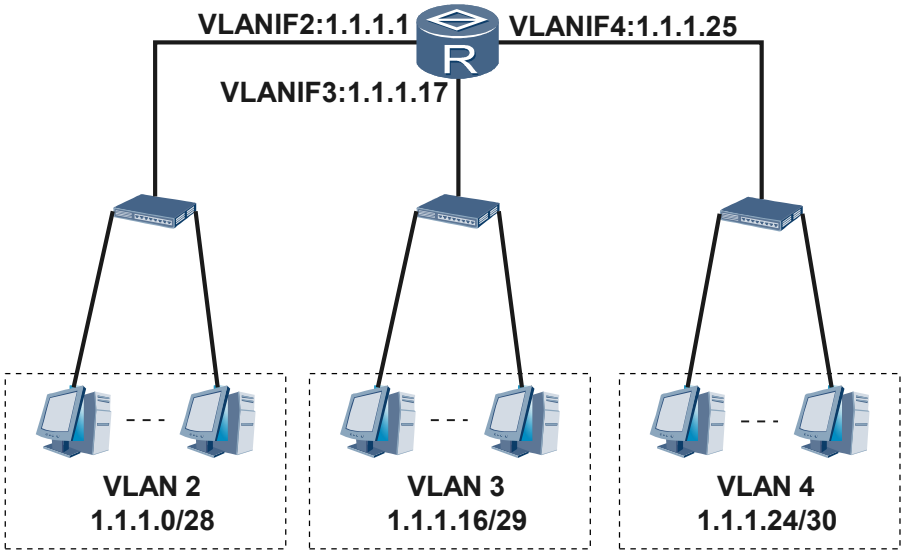


表 3-1 普通 VLAN 主机地址划分示例

VLAN	子网	网关地址	可用地址数	可用主机数	实际需求
2	1.1.1.0/28	1.1.1.1	14	13	10
3	1.1.1.16/29	1.1.1.17	6	5	5
4	1.1.1.24/30	1.1.1.25	2	1	1

如表 3-1 所示，VLAN2 预计未来有 10 个主机地址的需求，给其分配一个掩码长度是 28 的子网 1.1.1.0/28，其中 1.1.1.0 为子网号，1.1.1.15 为子网定向广播地址，这两个地址都不能用作主机地址，此外 1.1.1.1 作为子网缺省网关地址也不可作为主机地址，剩下范围在 1.1.1.2 ～ 1.1.1.14 的地址可以被主机使用，共 13 个。这样，尽管 VLAN2 只需要 10 个地址，但是按照子网划分却要分给它 13 个地址。

同理，VLAN3 预计未来有 5 个主机地址的需求，至少需要分配一个掩码长度是 29 的子网 1.1.1.16/29。VLAN4 预计未来只有 1 个主机，则分配一个掩码长度是 30 的子网 1.1.1.24/30。

上述 VLAN 一共需要 10+5+1 ＝ 16 个地址，但是按照普通 VLAN 的编址方式，即使最优化的方案也需要占用 16 ＋ 8 ＋ 4 ＝ 28 个地址，浪费了将近一半的地址。而且，如果 VLAN2 后来并没有 10 台主机，而实际只接入了 3 台主机，则多出来的地址也会因不能再被其他 VLAN 使用而浪费掉。

同时，这种划分也给后续的网络升级和扩展带来了很大不便。假设 VLAN4 今后需要再增加 2 台主机，而又不愿意改变已经分配的 IP 地址。并且在 1.1.1.24 后面的地址已经分配给了其他人的情况下，只能再给 VLAN4 的新用户重新分配一个的 29 位掩码的子网和一个新的 VLAN。这样 VLAN4 中的客户虽然只有 3 台主机，但是却被分配在两个子网中，并且也不在同一个 VLAN 内，不利于网络管理。

综上所述，很多 IP 地址被子网号、子网定向广播地址、子网缺省网关地址消耗掉，而不能用于 VLAN 内的主机地址。同时，这种地址分配的约束也降低了编址的灵活性，使许多闲置地址也被浪费掉。为了解决这一问题 VLAN Aggregation 就应运而生。

实现原理

VLAN Aggregation 技术（也称为 Super VLAN，即 VLAN 聚合）就是在一个物理网络内，用多个 VLAN 隔离广播域，使不同的 VLAN 属于同一个子网。它引入了 super-VLAN 和 sub-VLAN 的概念。

- super-VLAN：和通常意义上的 VLAN 不同，它只建立三层接口，与该子网对应，而且不包含物理端口。可以把它看作一个逻辑的三层概念—若干 sub-VLAN 的集合。
- sub-VLAN：只包含物理端口，用于隔离广播域的 VLAN，不能建立三层 VLAN 接口。它与外部的三层交换是靠 super-VLAN 的三层接口来实现的。

一个 super-VLAN 可以包含一个或多个保持着不同广播域的 sub-VLAN。sub-VLAN 不再占用一个独立的子网网段。在同一个 super-VLAN 中，无论主机属于哪一个 sub-VLAN，它的 IP 地址都在 super-VLAN 对应的子网网段内。

这样，sub-VLAN 间共用同一个三层接口，既减少了一部分子网号、子网缺省网关地址和子网定向广播地址的消耗，又实现了不同广播域使用同一子网网段地址的目的。消除了子网差异，增加了编址的灵活性，减少了闲置地址浪费。

仍以**表 3-1** 所示例子进行说明。用户需求不变。仍旧是 VLAN2 预计未来有 10 个主机地址的需求，VLAN3 预计未来有 5 个主机地址的需求，VLAN4 预计未来有 1 个主机地址的需求。

按照 VLAN Aggregation 的实现方式，新建 VLAN10 并配置为 super-VLAN，给其分配一个掩码长度是 24 的子网 1.1.1.0/24，其中 1.1.1.0 为子网号，1.1.1.1 为子网网关地址如**图 3-8** 所示。sub-VLAN（VLAN2、VLAN3、VLAN4）的地址划分如**表 3-2** 所示。

图 3-8 VLAN Aggregation 网络示意图

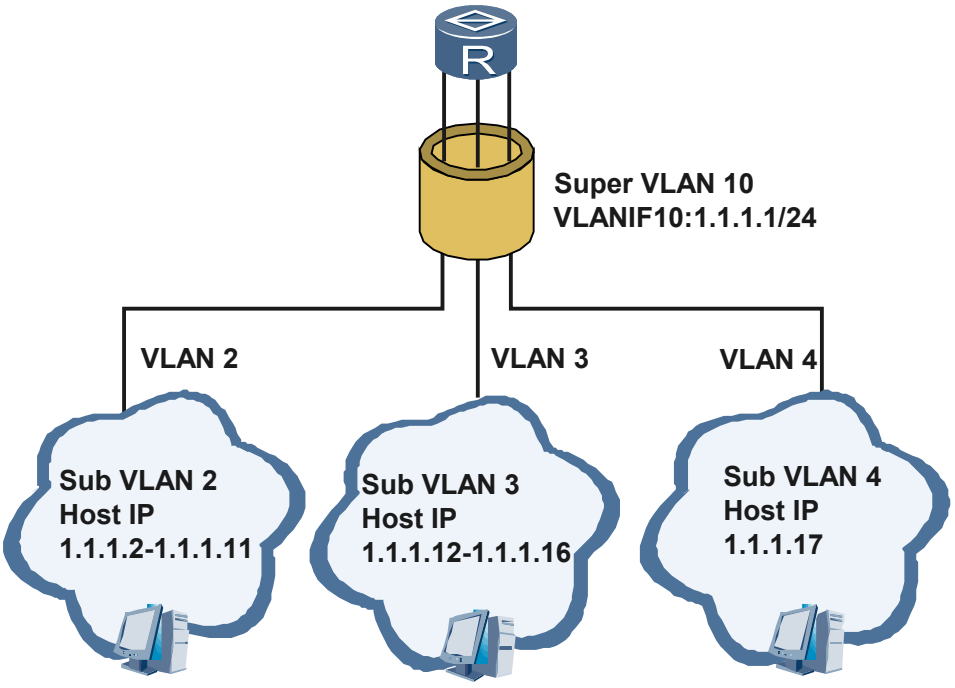


表 3-2 VLAN Aggregation 主机地址划分示例

VLAN	子网	网关地址	可用地址数	可用主机数	实际需求
2	1.1.1.0/24	1.1.1.1	10	1.1.1.2 ～ 1.1.1.11	10
3			5	1.1.1.12 ～ 1.1.1.16	5
4			1	1.1.1.17	1

VLAN Aggregation 的实现中，各 sub-VLAN 间的界线也不再是从前的子网界线了，它们可以根据其各自主机的需求数目在 super-VLAN 对应子网内灵活的划分地址范围。

从表 3-2 中可以看到，VLAN2、VLAN3 和 VLAN4 共用同一个子网（1.1.1.0/24）、子网缺省网关地址（1.1.1.1）和子网定向广播地址（1.1.1.255）。这样，普通 VLAN 实现方式中用到的其他子网号（1.1.1.16、1.1.1.24）和子网缺省网关（1.1.1.17、1.1.1.25），以及子网定向广播地址（1.1.1.15、1.1.1.23、1.1.1.27）就都可以用来作为主机 IP 地址使用。

这样，3 个 VLAN 一共需要 10 ＋ 5+1 ＝ 16 个地址，实际上在这个子网里就刚好分配了 16 个地址给（1.1.1.2 ～ 1.1.1.17）。这 16 个主机地址加上子网号（1.1.1.0）、子网缺省网关（1.1.1.1）和子网定向广播地址（1.1.1.255），一共用去了 19 个 IP 地址，网段内仍剩余 255-19 ＝ 236 的地址可以被任意 sub-VLAN 内的主机使用。

VLAN 间通信

- 概述

  VLAN Aggregation 在实现不同 VLAN 间共用同一子网网段地址的同时也带来了 sub-VLAN 间的三层转发问题。

  普通 VLAN 实现方式中，VLAN 间的主机可以通过各自不同的网关进行三层转发来达到互通的目的。但是 VLAN Aggregation 方式下，同一个 super-VLAN 内的主机使用的是同一个网段的地址和共用同一个网关地址。即使是属于不同的 sub-VLAN 的主机，由于它们同属一个子网，彼此通信时只会做二层转发，而不会通过网关进行三层转发。而实际上不同的 sub-VLAN 的主机在二层是相互隔离的，这就造成了 sub-VLAN 间无法通信的问题。

  解决这一问题的方法就是使用 ARP Proxy。

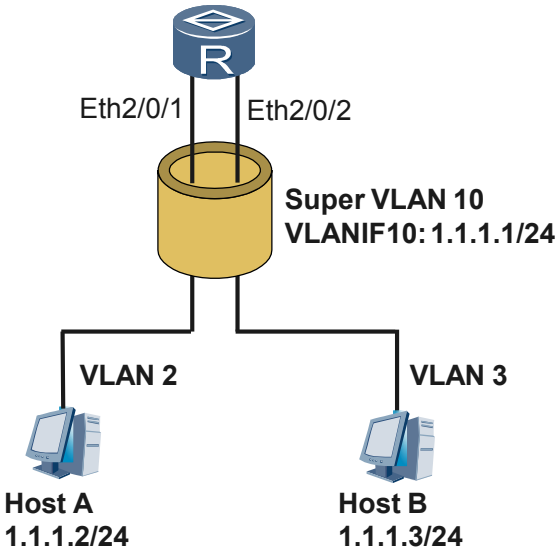
  说明

  关于 ARP Proxy 的原理，请参见《AR3200 特性描述 - IP 业务》中的“ARP”。

- 不同 sub-VLAN 间的三层互通

  例如，super-VLAN（VLAN10）包含 sub-VLAN（VLAN2 和 VLAN3），具体组网如图 3-9 所示。

  图 3-9 ARP Proxy 实现不同 sub-VLAN 间的三层互通组网图

  VLAN2 内的主机 A 与 VLAN3 内的主机 B 的通信过程如下：（假设主机 A 的 ARP 表中无主机 B 的对应表项并且网关上使能了 sub-VLAN 间的 ARP Proxy）。

  1. 主机 A 将主机 B 的 IP 地址（1.1.1.3）和自己所在网段 1.1.1.0/24 进行比较，发现主机 B 和自己在同一个子网，但是主机 A 的 ARP 表中无主机 B 的对应表项。
  2. 主机 A 发送 ARP 广播，请求主机 B 的 MAC 地址。
  3. 主机 B 并不在 VLAN2 的广播域内，无法接收到主机 A 的这个 ARP 请求。
  4. 由于网关上使能 sub-VLAN 间的 ARP Proxy，当网关收到主机 A 的 ARP 请求后，开始在路由表中查找，发现 ARP 请求中的主机 B 的 IP 地址（1.1.1.3）为

直连接口路由，则网关向所有其他 sub-VLAN 接口发送一个 ARP 广播，请求主机 B 的 MAC 地址。

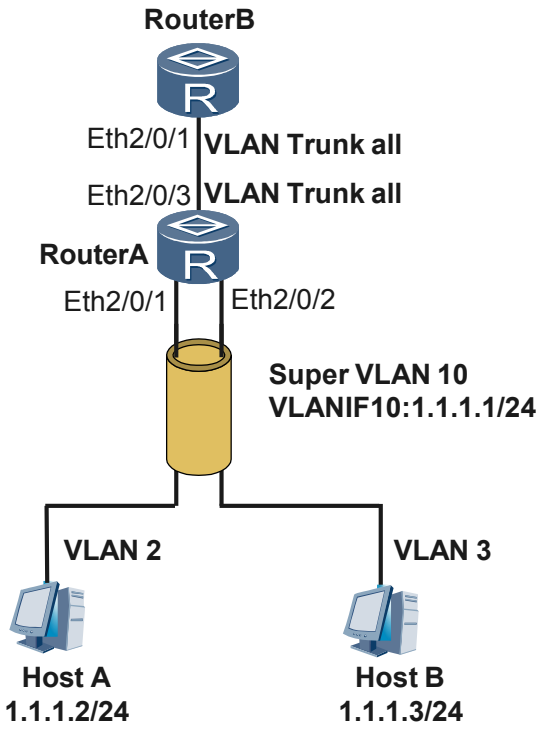
5. 主机 B 收到网关发送的 ARP 广播后，对此请求进行 ARP 应答。
6. 网关收到主机 B 的应答后，就把自己的 MAC 地址当作 B 的 MAC 地址回应给主机 A。
7. 网关和主机 A 的 ARP 表项中都存在主机 B 的对应表项。
8. 主机 A 之后要发给 B 的报文都先发送给网关，由网关做三层转发。

主机 B 发送报文给主机 A 的过程和上述的 A 到 B 的报文流程类似，不再赘述。

- sub-VLAN 与外部网络的二层通信

在基于端口的 VLAN 二层通信中，无论是数据帧进入接口还是从接口发出都不会有针对 super-VLAN 的报文。如图 3-10 所示。

图 3-10 sub-VLAN 与外部网络的二层通信组网图



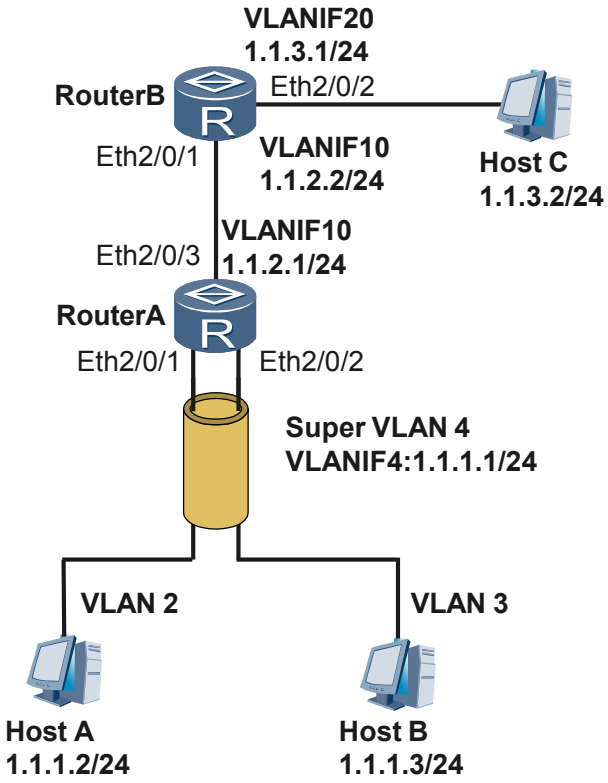
从 HostA 侧 Eth2/0/1 进入设备 RouterA 的帧会被打上 VLAN2 的 Tag，在设备 RouterA 中这个 Tag 不会因为 VLAN2 是 VLAN10 的 sub-VLAN 而变为 VLAN10 的 Tag。该数据帧从 Trunk 类型的接口 Eth2/0/3 出去时，依然是携带 VLAN2 的 Tag。

也就是说，设备 RouterA 本身不会发出 VLAN10 的报文。就算其他设备有 VLAN10 的报文发送到该设备上，这些报文也会因为设备 RouterA 上没有 VLAN10 对应物理端口而被丢弃。

对于设备 RouterA 而言，有效的 VLAN 只有 VLAN2 和 VLAN3，所有的数据帧都在这两个 VLAN 中转发的。

- sub-VLAN 与外部网络的三层通信

图 3-11 sub-VLAN 与外部网络的三层通信组网图



如图 3-11 所示，RouterA 上配置了 super-VLAN 4，sub-VLAN 2 和 sub-VLAN 3，并配置一个普通的 VLAN10；RouterB 上配置两个普通的 VLAN 10 和 VLAN 20。假设 super-VLAN 4 中的 sub-VLAN 2 下的主机 A 想访问与 RouterB 相连的主机 C，通信过程如下：（假设 RouterA 上已配置了去往 1.1.3.0/24 网段的路由，RouterB 上已配置了去往 1.1.1.0/24 网段的路由）

1. 主机 A 将主机 C 的 IP 地址（1.1.3.2）和自己所在网段 1.1.1.0/24 进行比较，发现主机 C 和自己不在同一个子网。
2. 主机 A 发送 ARP 请求给自己的网关，请求网关的 MAC 地址。
3. RouterA 收到该 ARP 请求后，查找 sub-VLAN 和 super-VLAN 的对应关系，从 sub-VLAN 2 发送 ARP 应答给主机 A。ARP 应答报文中的源 MAC 地址为 super-VLAN 4 对应的 VLANIF4 的 MAC 地址。
4. 主机 A 学习到网关的 MAC 地址。
5. 主机 A 向网关发送目的 MAC 为 super-VLAN 4 对应的 VLANIF4 的 MAC、目的 IP 为 1.1.3.2 的报文。
6. RouterA 收到该报文后进行三层转发，下一跳地址为 1.1.2.2，出接口为 VLANIF10，把报文发送给 RouterB。
7. RouterB 收到该报文后进行三层转发，通过直连出接口 VLANIF20，把报文发送给主机 C。
8. 主机 C 的回应报文，在 RouterB 上进行三层转发到达 RouterA。
9. RouterA 收到该报文后进行三层转发，通过 super-VLAN，把报文发送给主机 A。

3.4.4 VLAN Damping

如果指定 VLAN 已经创建对应的 VLANIF 接口，当 VLAN 中所有接口状态变为 Down 而引起 VLAN 状态变为 Down 时，VLAN 会向 VLANIF 接口上报 Down 的延迟时间，从而引起 VLANIF 接口状态变化。

为避免由于 VLANIF 接口状态变化引起的网络震荡，可以在 VLANIF 接口上启动 VLAN Damping 功能，抑制 VLANIF 接口状态变为 Down 的时间。

当使能 VLAN Damping 功能，VLAN 中最后一个处于 Up 状态的端口变为 Down 后，会抑制一定时间（抑制时间可配置）再上报给 VLANIF 接口。如果在抑制时间内 VLAN 中有端口 Up，则 VLANIF 接口状态保持 Up 状态不变。即，VLAN Damping 功能可以适当延迟 VLAN 向 VLANIF 接口上报接口 Down 状态的时间，从而抑制不必要的路由振荡。

3.4.5 MUX VLAN

MUX VLAN（Multiplex vlan）提供了一种通过 VLAN 进行网络资源控制的机制。

例如，在企业网络中，企业员工和企业客户可以访问企业的服务器。对于企业来说，希望企业内部员工之间可以互相交流，而企业客户之间是隔离的，不能够互相访问。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可以互相交流，而企业客户之间是隔离的。

基本概念

MUX VLAN 分为 Principal VLAN 和 Subordinate VLAN，Subordinate VLAN 又分为 Separate VLAN 和 Group VLAN，如表 3-3 所示

表 3-3 MUX VLAN 划分表

MUX VLAN	VLAN 类型	所属端口	通信权限
Principal VLAN	-	Principal PORT	Principal PORT 可以和 MUX VLAN 内的所有端口进行通信。
Subordinate VLAN	Separate VLAN	Separate PORT	Separate PORT 只能和 Principal PORT 进行通信，和其他类型的端口实现完全隔离。 每个 Separate VLAN 必须绑定一个 Principal VLAN。
	Group VLAN	Group PORT	Group PORT 可以和 Principal PORT 进行通信，在同一组内的端口也可互相通信，但不能和其他组端口或 Separate PORT 通信。 每个 Group VLAN 必须绑定一个 Principal VLAN。

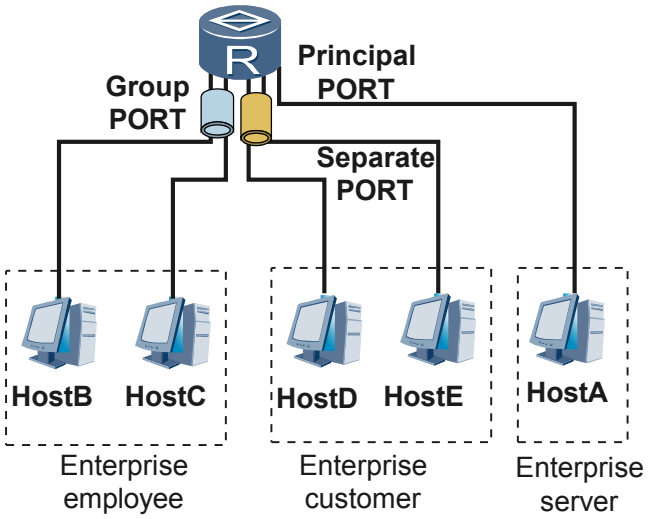
说明

- 如果该 VLAN ID 已经用于 Principal VLAN，那么该 VLAN 不能再用于 VLANIF 接口、Super VLAN、Sub VLAN。

MUX VLAN 通信原理

如图 3-12 所示，根据 MUX VLAN 特性，企业可以用 Principal PORT 连接企业服务器，Separate PORT 连接企业客户，Group PORT 连接企业员工。这样就能够实现企业客户、企业员工都能够访问企业服务器，而企业员工内部可以通信、企业客户间不能通信、企业客户和企业员工之间不能互访的目的。

图 3-12 MUX VLAN 应用场景图



3.4.6 Voice VLAN

Voice VLAN 的引入

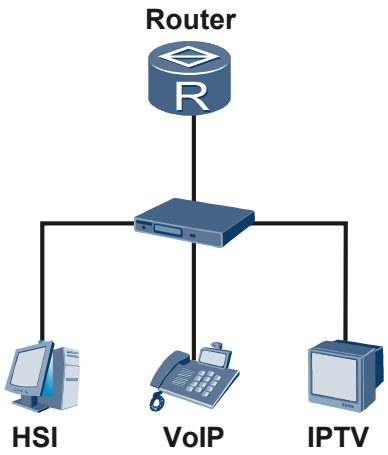
网络中经常同时存在语音数据和非语音数据两种流量。语音数据在传输时需要具有比其他业务数据更高的优先级，以减少传输过程中可能产生的时延和丢包现象。

提高语音数据传输优先级的传统处理方法是使用 ACL（Access Control List）对语音数据进行区分，并使用 QoS（Quality of Service）保证传输质量。为简化用户配置，更方便的管理语音流的传输，提出了 Voice VLAN 特性。

使能 Voice VLAN 功能的接口根据进入接口的数据流中的源 MAC 地址字段来判断该数据流是否为语音数据流。源 MAC 地址符合系统设置的语音设备 OUI（Organizationally Unique Identifier）地址的报文认为是语音数据流。接收到语音数据流的接口将自动加入 Voice VLAN 中传输。从而简化了用户配置，实现了用户方便管理语音数据。

如图 3-13 所示，HSI（High Speed Internet）、VOIP（Voice Over IP）、IPTV（Internet Protocol Television）三种业务同时接入 Router。为了区分语音数据流，对 VoIP 的电话终端流量通过不同的 VLAN 隔离，并给与更高的优先级，保证通话质量。此时，可在 Router 上部署 Voice VLAN 功能。对于 VoIP 的电话终端的语音流量，Router 为其打上预先配置的 VLAN，并且为语音流分配较高的优先级，使得语音流可以优先转发，保证通话质量。

图 3-13 Voice VLAN 组网图



Router 不同的接口下，可以指定不同的 VLAN 为 Voice VLAN，但是同一个接口下只能指定一个 VLAN 为 Voice VLAN。

基本概念

- Voice VLAN 的 OUI 地址

  OUI 地址表示一个 MAC 地址段。

  将 48 位的 MAC 地址和掩码的对应位作与运算可以确定 OUI 地址。接入设备的 MAC 地址和 OUI 地址匹配的位数，由掩码中全“1”的长度决定。例如，MAC 地址为 1－1－1，掩码为 FFFF-FF00－0000，则将 MAC 地址与其相应掩码位执行与运算的结果就是 OUI 地址 0001－0000－0000。只要接入设备的 MAC 地址前 24 位和 OUI 地址的前 24 位匹配，则使能 Voice VLAN 功能的接口将认为此数据流是语音数据流，接入的设备是语音设备。

- 接口加入 Voice VLAN 的模式

  端口加入 Voice VLAN 的模式如**表 3-4** 所示。

表 3-4 端口加入 Voice VLAN 的模式

端口加入 Voice VLAN 的模式	实现方式
自动模式	使能 Voice VLAN 功能的接口根据进入接口的数据流中的源 MAC 地址字段来判断该数据流是否为语音数据流。源 MAC 地址符合系统设置的语音设备 OUI 地址的报文认为是语音数据流。 接收到语音数据流的接口将自动加入 Voice VLAN 中传输，并通过老化机制维护 Voice VLAN 内的接口数量。在老化时间内: ● 如果配置了 Voice VLAN 的交换设备未收到任何来自该语音设备的语音报文时，连接语音设备的接口将自动从 Voice VLAN 中删除。 ● 如果配置了 Voice VLAN 的交换设备再次收到该语音设备发出的语音报文，连接语音设备的接口将再次自动加入 Voice VLAN。
手动模式	当接口使能 Voice VLAN 功能后，必须通过手工将连接语音设备的接口加入或退出 Voice VLAN 中，这样才能保证 Voice VLAN 功能生效。

不同的接口可以设置不同的模式加入 Voice VLAN，不同的接口加入 Voice VLAN 的模式是相互独立的。

- Voice VLAN 的工作模式

  Voice VLAN 的工作模式如**表 3-5** 所示。

表 3-5 Voice VLAN 的工作模式

Voice VLAN 的工作模式	实现方式	应用场景
安全模式	使能了 Voice VLAN 的接口对每一个进入 Voice VLAN 的报文都进行源 MAC 地址和 OUI 地址匹配检查。 ● 若匹配成功，报文进入 Voice VLAN 中转发。 ● 若匹配不成功： – 使能 Voice VLAN 的接口允许其他普通 VLAN 报文通过，则报文通过指定 VLAN 转发。 – 使能 Voice VLAN 的接口不允许其他普通 VLAN 报文通过，则丢弃匹配不成功的报文。	安全模式用于用户有多种数据流量（HSI、VOIP、IPTV）通过一个接口接入二层网络时，此接口只允许传输语音数据流。 安全模式可以防止 Voice VLAN 受到恶意数据流量的攻击，但是检查报文的工作会占用一定的系统资源。
普通模式	使能了 Voice VLAN 的接口允许同时传输语音数据流和业务数据流，容易受到恶意数据流量的攻击。	普通模式用于用户有多种数据流量（HSI、VOIP、IPTV）通过一个接口接入二层网络时，此接口需要同时传输语音数据流和业务数据流。

- Voice VLAN 老化时间

  在自动模式下，配置了 Voice VLAN 的设备通过学习语音设备发出的语音报文中源 MAC 地址，将连接语音设备的接口自动加入到 Voice VLAN 中，并通过老化机制维护 Voice VLAN 内的接口数量。

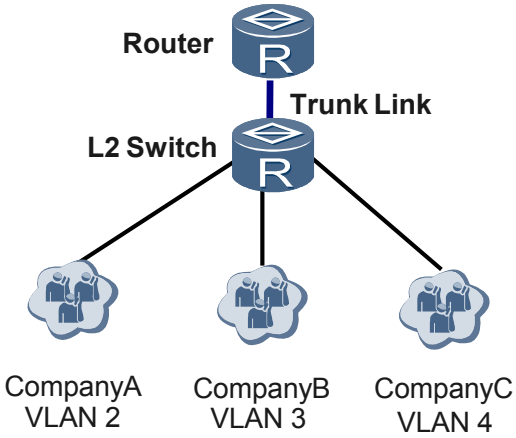
  老化时间超时后，使能了 Voice VLAN 功能的接口未收到任何来自该语音设备的语音报文时，连接语音设备的接口将自动从 Voice VLAN 中删除。如果使能了 Voice VLAN 功能的接口再次收到该语音设备发出的语音报文，连接语音设备的接口将再次自动加入 Voice VLAN。

  工作在手工模式下的 Voice VLAN 不受老化时间影响。

3.5 应用

基于端口的 VLAN 划分

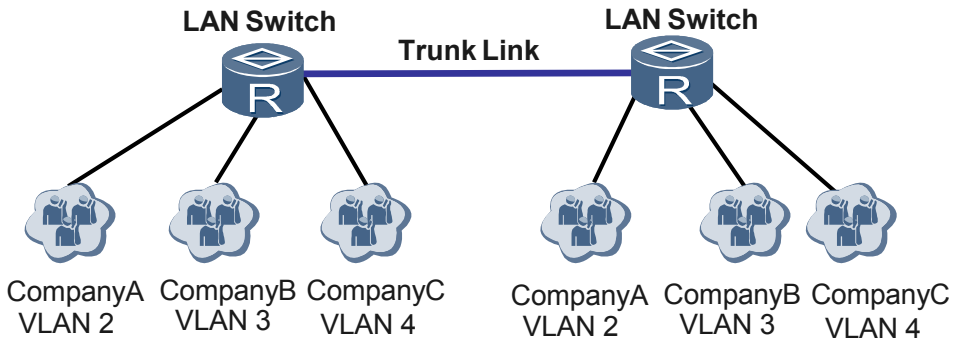
图 3-14 基于端口的 VLAN 划分组网图



商务楼宇内的中心交换机，根据楼宇内不同公司对端口需求，将每个公司所拥有的端口划分到不同的 VLAN，实现公司间业务数据的完全隔离。可以认为每个公司拥有独立的“虚拟交换机”，每个 VLAN 就是一个“虚拟工作组”。

VLAN Trunk 的应用

图 3-15 VLAN Trunk 的应用组网图



公司业务发展，部门需要跨越不同的商务楼宇。可通过 Trunk Link 连接不同楼宇的中心交换机，实现跨不同的交换机的不同公司的业务数据隔离，以及同一公司内业务数据的互通。

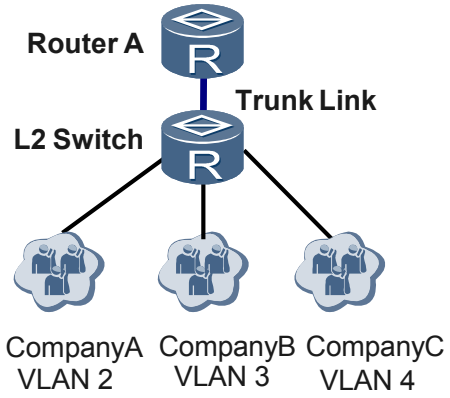
VLAN 间互通应用

对于不同的公司之间的互通需求，可以通过 VLAN 间互通来解决。

VLAN 间互通有两种方式，以下分别介绍。

- 多个 VLAN 属于同一个三层设备

图 3-16 多个 VLAN 属于同一个三层设备互通组网图

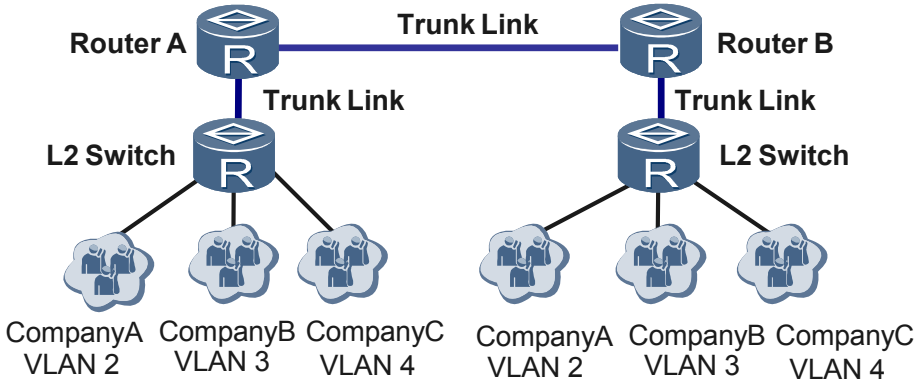


如图 3-16 所示，如果 VLAN2、VLAN3 和 VLAN4 仅属于 Router A，即 VLAN2、VLAN3 和 VLAN4 不是跨交换机的 VLAN，可在 Router A 上为每个 VLAN 配置一个虚拟路由接口，实现 VLAN2、VLAN3 和 VLAN4 间的路由。

图 3-16 中的三层设备可以是路由器或三层交换机。

- 多个 VLAN 跨越三层设备

图 3-17 多个 VLAN 跨越三层设备互通组网图

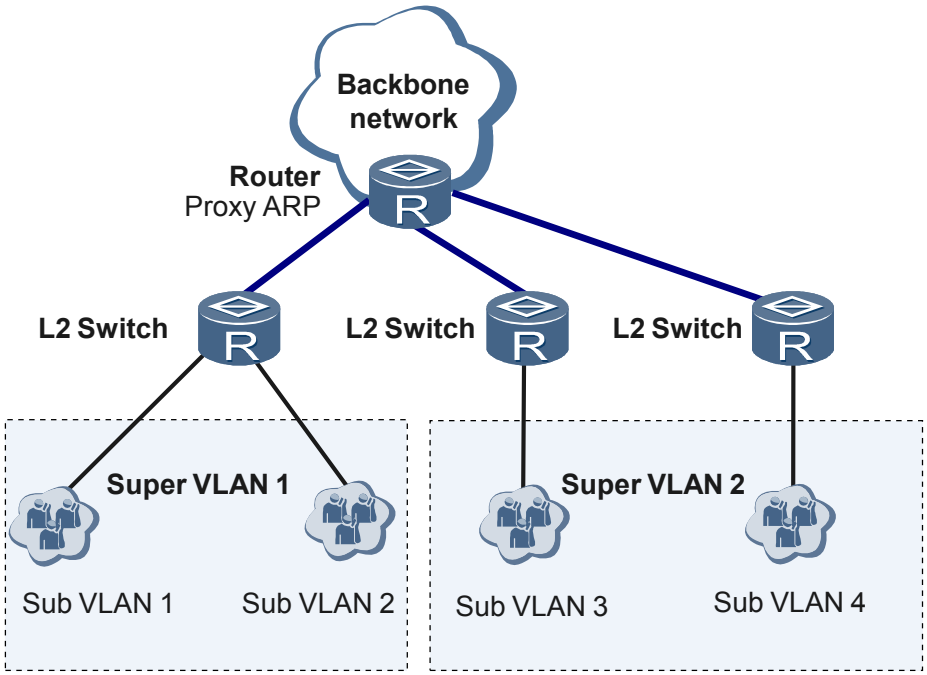


如图 3-17 所示，VLAN2、VLAN3 和 VLAN4 是跨交换机的 VLAN，可在 Router A 和 Router B 上为每个 VLAN 配置一个虚拟路由接口。除此以外，还需要在 Router A 和 Router B 之间的配置静态路由或运行路由协议。

图 3-17 中的三层设备可以是路由器或三层交换机。

VLAN Aggregation 的应用

图 3-18 VLAN Aggregation 应用组网图



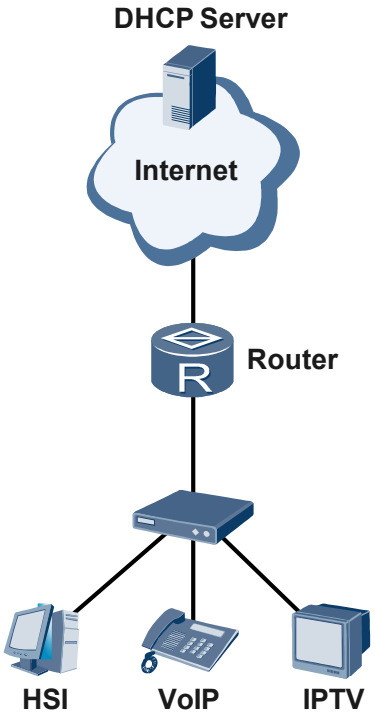
如图 3-18 所示，共有 4 个 VLAN，如果 VLAN 间需要互通，在 Router 上要为每个 VLAN 配置一个 IP 地址。

将 VLAN1 和 VLAN2 聚合到 Super VLAN 1 中；将 VLAN3 和 VLAN4 聚合到 Super VLAN 2 中。这样只需在 Router 上为 Super VLAN 分配 IP 地址，节约了 IP 地址资源。

在 Router 上配置 VLAN 间的 ARP Proxy，使同一 Super VLAN 下的不同 Sub VLAN 间的用户可以互通。

Voice VLAN 的应用

图 3-19 Voice VLAN 的应用组网图



如图 3-19 所示 HSI（High Speed Internet）、VoIP（Voice over IP）和 IPTV（Internet Protocol Television）通过网关设备接入 Router。用户对语音通话质量较敏感，需要提高语音数据流的传输优先级，以保证用户的通话质量。

可在 Router 上配置 Voice VLAN 功能解决此问题。

当 Router 配置 Voice VLAN 功能后，会根据进入接口数据流的源 MAC 地址来判断该数据流是否为语音数据流。当源 MAC 地址匹配系统设置的语音设备 OUI 地址时，则认为是语音数据流。Router 接收到语音数据流后将修改语音数据流的传输优先级，并且在 Voice VLAN 内进行传输，以保证用户的通话质量。

3.6 术语与缩略语

术语/缩略语

缩略语	英文全称	中文全称
VLAN	Virtual Local Area Network	虚拟局域网
PVID	Port Default VLAN ID	端口缺省虚拟局域网 ID

4 GVRP

关于本章

4.1 介绍

定义

GARP 协议主要用于建立一种属性传递扩散的机制，以保证协议实体能够注册和注销该属性。GARP 作为一个属性注册协议的载体，可以用来传播属性。将 GARP 协议报文的内容映射成不同的属性即可支持不同上层协议应用。

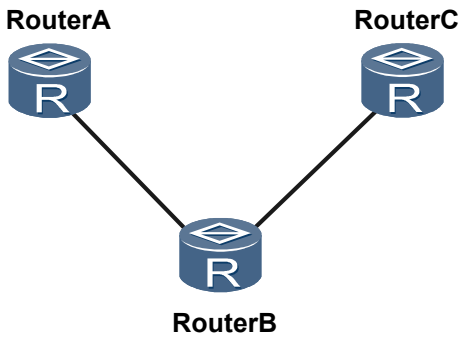
GVRP 是 GARP 的一种应用，用于注册和注销 VLAN 属性。

GARP 协议通过目的 MAC 地址区分不同的应用。在 IEEE Std 802.1Q 中将 01-80-C2-00-00-21 分配给 VLAN 应用，即 GVRP。

目的

如果需要为网络中的所有设备都配置 VLAN，就需要网络管理员在每台设备上分别进行手工添加。如图 4-1 所示，RouterA 上有 VLAN2，RouterB 和 RouterC 上只有 VLAN1，三台设备通过 Trunk 链路连接在一起。为了使 RouterA 上 VLAN 2 的报文可以传到 RouterC，网络管理员必须在 RouterB 和 RouterC 上分别手工添加 VLAN2。

图 4-1 GVRP 应用组网图



对于上面的组网情况，手工添加 VLAN 很简单，但是当实际组网复杂到网络管理员无法短时间内了解网络的拓扑结构，或者是整个网络的 VLAN 太多时，工作量会非常大，而且非常容易配置错误。在这种情况下，用户可以通过 GVRP 的 VLAN 自动注册功能完成 VLAN 的配置。

受益

GVRP 基于 GARP 机制，主要用于维护设备动态 VLAN 属性。通过 GVRP 协议，一台设备上的 VLAN 信息会迅速传播到整个交换网。GVRP 实现动态分发、注册和传播 VLAN 属性，从而达到减少网络管理员的手工配置量及保证 VLAN 配置正确的目的。

4.2 参考标准和协议

本特性的参考资料清单如下：

文档	描述	备注
IEEE Std 802.1D	Information technology—Telecommunications and information exchange between systems—Local and metropolitan area networks— Common specifications—Media Access Control (MAC) Bridges	
IEEE Std 802.1Q	IEEE Standards for Local and Metropolitan Area Networks: Virtual Bridged Local Area Networks	

4.3 可获得性

涉及网元

无需其它网元的配合。

License 支持

无需获得 License 许可，均可获得该特性的服务。

版本支持

产品	最低支持版本
AR3200	V200R001C00

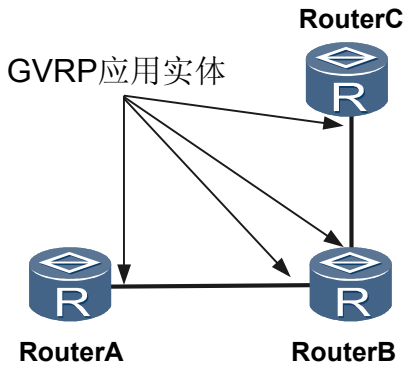
4.4 原理描述

4.4.1 基本概念

应用实体

在设备上，每一个参与协议的端口可以视为一个应用实体。当 GVRP 在设备上启动的时候，每个启动 GVRP 的端口对应一个 GVRP 应用实体，如图 4-2 所示。

图 4-2 GVRP 应用实体



VLAN 的注册和注销

GVRP 协议可以实现 VLAN 属性的自动注册和注销：

- VLAN 的注册：指的是将端口加入 VLAN。
- VLAN 的注销：指的是将端口退出 VLAN。

GVRP 协议通过声明和回收声明实现 VLAN 属性的注册和注销。

- 当端口接收到一个 VLAN 属性声明时，该端口将注册该声明中包含的 VLAN 信息（端口加入 VLAN）。
- 当端口接收到一个 VLAN 属性的回收声明时，该端口将注销该声明中包含的 VLAN 信息（端口退出 VLAN）。

GVRP 协议的属性注册和注销仅仅是对于接收到 GVRP 协议报文的端口而言的。

图 4-3 VLAN 的注册和注销

消息类型

GARP 应用实体之间的信息交换借助于消息的传递来完成，主要有三类消息起作用，分别为 Join 消息、Leave 消息和 LeaveAll 消息。

- Join 消息

当一个 GARP 应用实体希望其它设备注册自己的属性信息时，它将对外发送 Join 消息；当收到其它实体的 Join 消息或本设备静态配置了某些属性，需要其它 GARP 应用实体进行注册时，它也会向外发送 Join 消息。

Join 消息分为 JoinEmpty 和 JoinIn 两种，区别如下：

- JoinEmpty：声明一个本身没有注册的属性。
- JoinIn：声明一个本身已经注册的属性。

- Leave 消息

  当一个 GARP 应用实体希望其它设备注销自己的属性信息时，它将对外发送 Leave 消息；当收到其它实体的 Leave 消息注销某些属性或静态注销了某些属性后，它也会向外发送 Leave 消息。

  Leave 消息分为 LeaveEmpty 和 LeaveIn 两种，区别如下：

  - LeaveEmpty：注销一个本身没有注册的属性。
  - LeaveIn：注销一个本身已经注册的属性。

- LeaveAll 消息

  每个应用实体启动后，将同时启动 LeaveAll 定时器，当该定时器超时后应用实体将对外发送 LeaveAll 消息。

  LeaveAll 消息用来注销所有的属性，以使其它应用实体重新注册本实体上所有的属性信息，以此来周期性地清除网络中的垃圾属性（例如某个属性已经被删除，但由于设备突然断电，并没有发送 Leave 消息来通知其他实体注销此属性）。

定时器

GARP 协议中用到了四个定时器，下面分别介绍一下它们的作用。

- Join 定时器

  Join 定时器是用来控制 Join 消息（包括 JoinIn 和 JoinEmpty）的发送的。

  为了保证 Join 消息能够可靠的传输到其它应用实体，发送第一个 Join 消息后将等待一个 Join 定时器的时间间隔，如果在一个 Join 定时器时间内收到 JoinIn 消息，则不发送第二个 Join 消息；如果没收到，则再发送一个 Join 消息。每个端口维护独立的 Join 定时器。

- Hold 定时器

  Hold 定时器是用来控制 Join 消息（包括 JoinIn 和 JoinEmpty）和 Leave 消息（包括 LeaveIn 和 LeaveEmpty）的发送的。

  当在应用实体上配置属性或应用实体接收到消息时不会立刻将该消息传播到其它设备，而是在等待一个 Hold 定时器后再发送消息，设备将此 Hold 定时器时间段内接收到的消息尽可能封装成最少数量的报文，这样可以减少报文的发送量。如果没有 Hold 定时器的话，每来一个消息就发送一个，造成网络上报文量太大，既不利于网络的稳定，也不利于充分利用每个报文的数据容量。

  每个端口维护独立的 Hold 定时器。Hold 定时器的值要小于等于 Join 定时器值的一半。

- Leave 定时器

  Leave 定时器是用来控制属性注销的。

  每个应用实体接收到 Leave 或 LeaveAll 消息后会启动 Leave 定时器，如果在 Leave 定时器超时之前没有接收到该属性的 Join 消息，属性才会被注销。

  这是因为网络中如果有一个实体因为不存在某个属性而发送了 Leave 消息，并不代表所有的实体都不存在该属性了，因此不能立刻注销属性，而是要等待其他实体的消息。

例如，某个属性在网络中有两个源，分别在应用实体 A 和 B 上，其他应用实体通过协议注册了该属性。当把此属性从应用实体 A 上删除的时候，实体 A 发送 Leave 消息，由于实体 B 上还存在该属性源，在接收到 Leave 消息之后，会发送 Join 消息，以表示它还有该属性。其他应用实体如果收到了应用实体 B 发送的 Join 消息，则该属性仍然被保留，不会被注销。只有当其它应用实体等待两个 Join 定时器以上仍没有收到该属性的 Join 消息时，才能认为网络中确实没有该属性了，所以这就要求 Leave 定时器的值大于 2 倍 Join 定时器的值。

每个端口维护独立的 Leave 定时器。

- LeaveAll 定时器

  每个 GARP 应用实体启动后，将同时启动 LeaveAll 定时器，当该定时器超时后 GARP 应用实体将对外发送 LeaveAll 消息，随后再启动 LeaveAll 定时器，开始新的一轮循环。

  接收到 LeaveAll 消息的实体将重新启动所有的定时器，包括 LeaveAll 定时器。在自己的 LeaveAll 定时器重新超时之后才会再次发送 LeaveAll 消息，这样就避免了短时间内发送多个 LeaveAll 消息。

  如果不同设备的 LeaveAll 定时器同时超时，就会同时发送多个 LeaveAll 消息，增加不必要的报文数量，为了避免不同设备同时发生 LeaveAll 定时器超时，实际定时器运行的值是大于 LeaveAll 定时器的值，小于 1.5 倍 LeaveAll 定时器值的一个随机值。一次 LeaveAll 事件相当于全网所有属性的一次 Leave。由于 LeaveAll 影响范围很广，所以建议 LeaveAll 定时器的值不能太小，至少应该大于 Leave 定时器的值。

  每个设备只在全局维护一个 LeaveAll 定时器。

注册模式

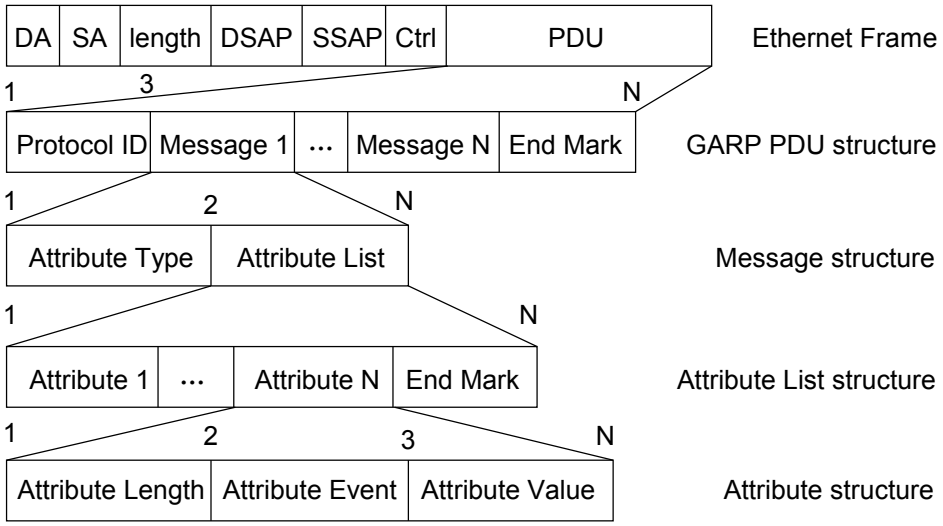
手工配置的 VLAN 称为静态 VLAN，通过 GVRP 协议创建的 VLAN 称为动态 VLAN。GVRP 有三种注册模式，不同的模式对静态 VLAN 和动态 VLAN 的处理方式也不同。GVRP 的三种注册模式分别定义如下：

- Normal 模式：允许动态 VLAN 在端口上进行注册，同时会发送静态 VLAN 和动态 VLAN 的声明消息。
- Fixed 模式：不允许动态 VLAN 在端口上注册，只发送静态 VLAN 的声明消息。
- Forbidden 模式：不允许动态 VLAN 在端口上进行注册，同时删除端口上除 VLAN1 外的所有 VLAN，只发送 VLAN1 的声明消息。

4.4.2 报文结构

GARP 协议报文采用 IEEE 802.3 Ethernet 封装形式，报文结构如图 4-4 所示。

图 4-4 GARP 协议报文



各个字段的说明如下表所示。

字段	含义	取值
Protocol ID	协议 ID。	取值为 1。
Message	消息，每个 Message 由 Attribute Type、Attribute List 构成。	-
Attribute Type	属性类型，由具体的 GARP 的应用定义。	对于 GVRP，属性类型为 0x01，表示属性取值为 VLAN ID。
Attribute List	属性列表，由多个属性构成。	-
Attribute	属性，每个属性由 Attribute Length、Attribute Event、Attribute Value 构成。	-
Attribute Length	属性长度。	取值 2 ～ 255，单位为字节。
Attribute Event	属性描述的事件。	● 0：LeaveAll Event ● 1：JoinEmpty Event ● 2：JoinIn Event ● 3：LeaveEmpty Event ● 4：LeaveIn Event ● 5：Empty Event

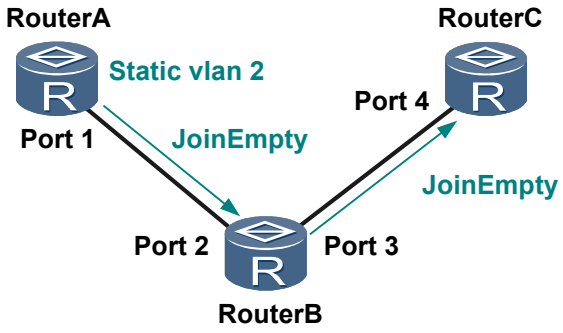
字段	含义	取值
Attribute Value	属性取值。	GVRP 的属性取值为 VLAN ID，但 LeaveAll 属性的 Attribute Value 值无效。
End Mark	结束标志、GARP 的 PDU 的结尾标志。	以 0x00 取值表示。

4.4.3 工作过程

下面通过一个简单的例子来介绍一下 GVRP 的工作过程。该例子分四个阶段描述了一个 VLAN 属性在网络中是如何被注册和注销的。

VLAN 属性的单向注册

图 4-5 VLAN 属性的单向注册



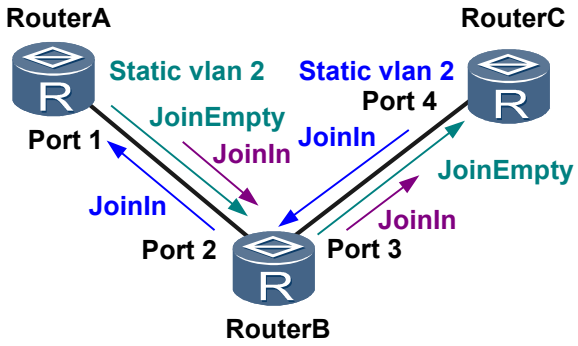
在 RouterA 上创建静态 VLAN2，通过 VLAN 属性的单向注册，将 RouterB 和 RouterC 的相应端口自动加入 VLAN2。

1. 在 RouterA 上创建静态 VLAN2 后，Port1 启动 Join 定时器和 Hold 定时器，等待 Hold 定时器超时后，RouterA 向 RouterB 发送第一个 JoinEmpty 消息，Join 定时器超时后再次启动 Hold 定时器，再等待 Hold 定时器超时后，发送第二个 JoinEmpty 消息。
2. RouterB 上接收到第一个 JoinEmpty 后创建动态 VLAN2，并把接收到 JoinEmpty 消息的 Port2 加入到动态 VLAN2 中，同时告知 Port3 启动 Join 定时器和 Hold 定时器，等待 Hold 定时器超时后向 RouterC 发送第一个 JoinEmpty 消息，Join 定时器超时后再次启动 Hold 定时器，Hold 定时器超时之后，发送第二个 JoinEmpty 消息。RouterB 上收到第二个 JoinEmpty 后，因为 Port2 已经加入动态 VLAN2，所以不作处理。

3. RouterC 上接收到第一个 JoinEmpty 后创建动态 VLAN2，并把接收到 JoinEmpty 消息的 Port4 加入到动态 VLAN2 中。RouterC 上收到第二个 JoinEmpty 后，因为 Port4 已经加入动态 VLAN2，所以不作处理。
4. 此后，每当 Leaveall 定时器超时或收到 LeaveAll 消息，设备会重新启动 Leaveall 定时器、Join 定时器、Hold 定时器和 Leave 定时器。RouterA 的 Port1 在 Hold 定时器超时之后发送第一个 JoinEmpty 消息，再等待 Join 定时器+Hold 定时器之后，发送第二个 JoinEmpty 消息，RouterB 向 RouterC 发送 JoinEmpty 消息的过程也是如此。

VLAN 属性的双向注册

图 4-6 VLAN 属性的双向注册

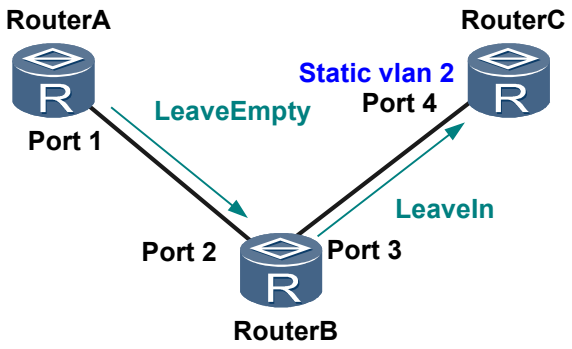


通过上述 VLAN 属性的单向注册过程，端口 Port1、Port2、Port4 已经加入 VLAN2，但是 Port3 还没有加入 VLAN2（只有收到 JoinEmpty 消息或 JoinIn 消息的端口才能加入动态 VLAN）。为使 VLAN2 流量可以双向互通，需要进行 RouterC 到 RouterA 方向的 VLAN 属性的注册过程。

1. VLAN 属性的单向注册完成后，在 RouterC 上创建静态 VLAN2（将动态 VLAN 转换成静态 VLAN），Port4 启动 Join 定时器和 Hold 定时器，等待 Hold 定时器超时后，RouterC 向 RouterB 发送第一个 JoinIn 消息（因为 Port4 已经注册了 VLAN2，所以发送 JoinIn 消息），Join 定时器超时后再次启动 Hold 定时器，Hold 定时器超时之后，发送第二个 JoinIn 消息。
2. RouterB 上接收到第一个 JoinIn 后，把接收到 JoinIn 消息的 Port3 加入到动态 VLAN2 中，同时告知 Port2 启动 Join 定时器和 Hold 定时器，等待 Hold 定时器超时后，向 RouterA 发送第一个 JoinIn 消息，Join 定时器超时后再次启动 Hold 定时器，Hold 定时器超时之后，发送第二个 JoinIn 消息；RouterB 上收到第二个 JoinIn 后，因为 Port3 已经加入动态 VLAN2，所以不作处理。
3. RouterA 上接收到 JoinIn 之后，停止向 RouterB 发送 JoinEmpty 消息。此后，当 Leaveall 定时器超时或收到 LeaveAll 消息，设备重新启动 Leaveall 定时器、Join 定时器、Hold 定时器和 Leave 定时器。RouterA 的 Port1 在 Hold 定时器超时之后就开始发送 JoinIn 消息。
4. RouterB 向 RouterC 发送 JoinIn 消息。
5. RouterC 收到 JoinIn 消息后，由于本身已经创建了静态 VLAN2，所以不会再创建动态 VLAN2。

VLAN 属性的单向注销

图 4-7 VLAN 属性的单向注销

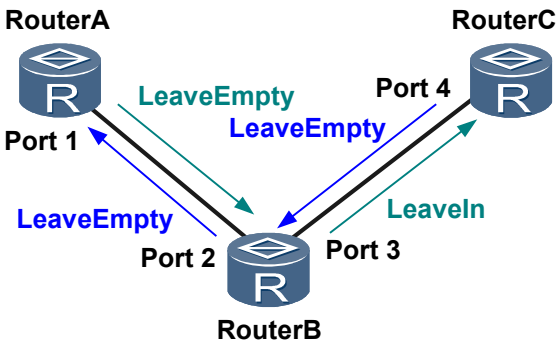


当设备上不再需要 VLAN2 时，可以通过 VLAN 属性的注销过程将 VLAN2 从设备上删除。

1. 在 RouterA 上删除静态 VLAN2，Port1 启动 Hold 定时器，等待 Hold 定时器超时后，RouterA 向 RouterB 发送 LeaveEmpty 消息。LeaveEmpty 消息只需发送一次。
2. RouterB 上接收到 LeaveEmpty，Port2 启动 Leave 定时器，等待 Leave 定时器超时之后 Port2 注销 VLAN2，将 Port2 从动态 VLAN2 中删除（由于此时 VLAN2 中还存在端口 Port3，所以不会删除 VLAN2），同时告知 Port3 启动 Hold 定时器和 Leave 定时器，等待 Hold 定时器超时后，向 RouterC 发送 LeaveIn 消息。由于 RouterC 的静态 VLAN2 还没有删除，Port3 在 Leave 定时器超时之前仍然能够收到 Port4 发送的 JoinIn 消息，所以 RouterA 和 RouterB 上仍然能够学习到动态的 VLAN2。
3. RouterC 上接收到 LeaveIn 后，由于 RouterC 上存在静态 VLAN2，所以 Port4 不会从 VLAN2 中删除。

VLAN 属性的双向注销

图 4-8 VLAN 属性的双向注销



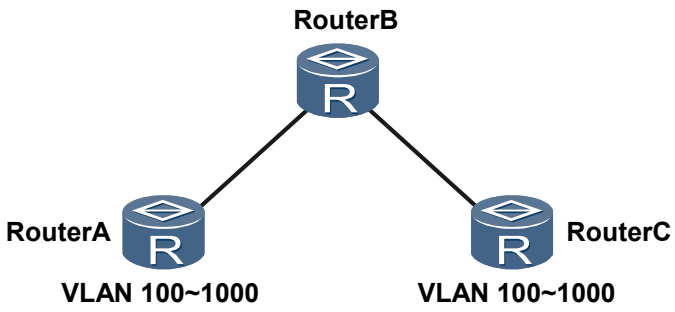
为了彻底删除所有设备上的 VLAN2，需要进行 VLAN 属性的双向注销。

1. 在 RouterC 上删除静态 VLAN2，Port4 启动 Hold 定时器，等待 Hold 定时器超时后，RouterC 向 RouterB 发送 LeaveEmpty 消息。
2. RouterB 接收到 LeaveEmpty 消息后，Port3 启动 Leave 定时器，等待 Leave 定时器超时之后 Port3 注销 VLAN2，将 Port3 从动态 VLAN2 中删除并删除动态 VLAN2，同时告知 Port2 启动 Hold 定时器，等待 Hold 定时器超时后，向 RouterA 发送 LeaveEmpty 消息。
3. RouterA 接收到 LeaveEmpty 消息后，Port1 启动 Leave 定时器，等待 Leave 定时器超时之后 Port1 注销 VLAN2，将 Port1 从动态 VLAN2 中删除并删除动态 VLAN2。

4.5 应用

GVRP 特性使得不同设备上的 VLAN 信息可以由协议动态维护和更新，用户只需要对少数设备进行 VLAN 配置即可应用到整个交换网络，无需耗费大量时间进行拓扑分析和配置管理。如图 4-9 所示所有设备都使能 GVRP 功能，设备之间相连的端口均为 Trunk 端口，并允许所有 VLAN 通过。只需在 RouterA 和 RouterC 上分别手工配置静态 VLAN100 ～ 1000，设备 RouterB 就可以通过 GVRP 协议学习到这些 VLAN，最后各设备上都存在 VLAN100 ～ 1000。

图 4-9 典型组网应用



4.6 术语与缩略语

缩略语

缩略语	英文全称	中文全称
GARP	Generic Attribute Registration Protocol	通用属性注册协议
GVRP	GARP VLAN Registration Protocol	通用 VLAN 注册协议

5 STP/RSTP/MSTP

关于本章

5.1 介绍

定义

以太网交换网络中为了进行链路备份，提高网络可靠性，通常会使用冗余链路。但是使用冗余链路会在交换网络上产生环路，并导致广播风暴以及 MAC 地址表不稳定等故障现象，从而导致用户通信质量较差，甚至通信中断。为解决交换网络中的环路问题，提出了生成树协议 STP（Spanning Tree Protocol）。

STP 包含两种含义：

- 狭义的 STP 是指 IEEE 802.1D 中定义的 STP 协议。
- 广义的 STP 包括 IEEE 802.1D 中定义的 STP、IEEE 802.1W 中定义的快速生成树协议 RSTP（Rapid Spanning Tree Protocol）和 IEEE 802.1S 中定义的多生成树协议 MSTP（Multiple Spanning Tree Protocol）。

目前，生成树协议支持如下：

- STP

  IEEE 于 1998 年发布的 802.1D 标准定义了 STP。

  STP 是数据链路层的管理协议，用于二层网络的环路检测和预防。STP 可阻塞二层网络中的冗余链路，将网络修剪成树状，达到消除环路的目的。

  但是，STP 拓扑收敛速度慢，即使是边缘端口也必须等待两倍 forward delay 定时器的时间（缺省为 30 秒）延迟，端口才能迁移到转发状态。

- RSTP

  IEEE 于 2001 年发布的 802.1W 标准定义了 RSTP。

  RSTP 在 STP 基础上进行了改进，实现了网络拓扑快速收敛。

  但 RSTP 和 STP 还存在同一个缺陷：由于局域网内所有的 VLAN（Virtual Local Area Network）共享一棵生成树，因此无法在 VLAN 间实现数据流量的负载均衡，还有可能造成部分 VLAN 的报文无法转发。

  RSTP 向下兼容 STP 协议，可以混合组网。

- MSTP

  IEEE 于 2002 年发布的 802.1S 标准定义了 MSTP。

  MSTP 通过设置 VLAN 映射表（即 VLAN 和生成树实例的对应关系表），把 VLAN 和生成树实例联系起来。同时它把一个交换网络划分成多个域，每个域内形成多棵生成树实例，生成树实例之间彼此独立。MSTP 将环路网络修剪成为一个无环的树形网络，避免报文在环路网络中的增生和无限循环，同时还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。

  MSTP 兼容 STP 和 RSTP。三种生成树协议的比较如表 5-1 所示。

表 5-1 三种生成树协议的比较

生成树协议	特点	应用场景
STP	形成一棵无环路的树：解决广播风暴并实现冗余备份。	无需区分用户或业务流量，所有 VLAN 共享一棵生成树。

生成树协议	特点	应用场景
RSTP	● 形成一棵无环路的树：解决广播风暴并实现冗余备份。 ● 收敛速度快。	
MSTP	● 形成一棵无环路的树：解决广播风暴并实现冗余备份。 ● 收敛速度快。 ● 多棵生成树在 VLAN 间实现负载均衡，不同 VLAN 的流量按照不同的路径转发。	需要区分用户或业务流量，并实现负载分担。不同的 VLAN 通过不同的生成树转发流量，每棵生成树之间相互独立。

目的

在以太网交换网中部署生成树协议后，如果网络中出现环路，生成树协议通过拓扑计算，可实现：

- 消除环路：通过阻塞冗余链路消除网络中可能存在的网络通信环路。
- 链路备份：当前活动的路径发生故障时，激活冗余备份链路，恢复网络连通性。

5.2 参考标准和协议

MSTP 特性的参考资料清单如下：

文档	描述	备注
IEEE 802.1D	IEEE Standard for: Local and metropolitan area networks Virtual Bridged Local Area Networks	-
IEEE 802.1S	IEEE Standard for: Local and metropolitan area networks Virtual Bridged Local Area Networks	-
IEEE 802.1W	IEEE Standard for: Local and metropolitan area networks Common specifications	-

5.3 可获得性

涉及网元

无需其它网元的配合。

License 支持

无需获得 License 许可，均可获得该特性的服务。

版本支持

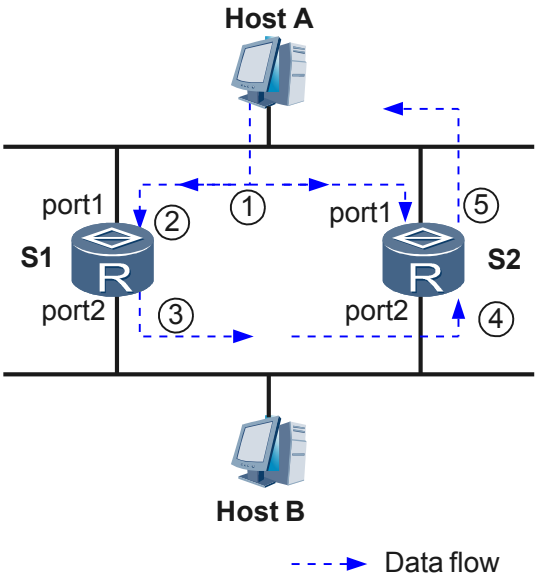
产品	最低支持版本
AR3200	V200R001C00

5.4 STP/RSTP 原理描述

5.4.1 STP 出现的背景

STP 是一个用于局域网中消除环路的协议。运行该协议的交换设备通过彼此交互信息而发现网络中的环路，并适当对某些端口进行阻塞以消除环路。由于局域网规模的不断增长，生成树协议已经成为了当前最重要的局域网协议之一。

图 5-1 典型局域网络示意图



如图 5-1 所示网络中，会产生如下两种情况：

- 广播风暴导致网络不可用。

  环路产生广播风暴，这是众所周知的。图 5-1 中，假设交换设备上没有启用 STP 协议。如果 HostA 发出广播请求，那么广播报文将被其他两台交换设备的端口 port1 接收，并分别从端口 port2 广播出去，然后端口 port2 又收到另一台交换设备发过来的广播报文，再分别从两台交换设备的端口 port1 转发，如此反复，最终导致整个网络资源被耗尽，网络瘫痪不可用。
- MAC 地址表震荡导致 MAC 地址表项被破坏。

如图 5-1 所示，即使是单播报文，也有可能导致交换设备的 MAC 地址表项混乱，以致破坏交换设备的 MAC 地址表。

假设图 5-1 所示的网络中没有广播风暴，HostA 发生一个单播报文给 HostB，如果此时 HostB 临时从网络中移去，那么交换设备上有关 HostB 的 MAC 地址表项也将被删除。此时 HostA 发给 HostB 的单播报文，将被交换设备 S1 的端口 port1 接收，由于 S1 上没有相应的 MAC 地址转发表项，该单播报文将被转发到端口 port2 上，然后交换设备 S2 的端口 port2 又收到从对端 port2 端口发来的单播报文，然后又从 port1 发出去。如此反复，在两台交换设备上，由于不间断地从端口 port1、port2 收到主机 A 发来的单播报文，交换设备会不停地修改自己的 MAC 地址表项，从而引起了 MAC 地址表的抖动。如此下去，最终导致 MAC 地址表项被破坏。

5.4.2 STP 基本概念

基本思想

STP 是数据链路层协议。运行该协议的设备通过彼此交互信息发现网络中的环路，并有选择的对某个端口进行阻塞，最终将环形网络结构修剪成无环路的树形网络结构，从而防止报文在环形网络中不断增生和无限循环，避免设备由于重复接收相同的报文造成处理能力下降。

运行 STP 协议的设备采用配置消息 BPDU（Bridge Protocol Data Unit，桥协议数据单元）交互信息，一般简称为 BPDU。BPDU 分为两大类：

- 配置 BPDU（Configuration BPDU）：用来进行生成树计算和维护生成树拓扑的报文。
- TCN BPDU（Topology Change Notification BPDU）：当拓扑结构发生变化时，下游设备用来通知上游设备网络拓扑结构发生变化的报文。

说明

配置 BPDU 中包含了足够的信息保证设备完成生成树计算，其中包含重要信息如下：

- 根桥 ID：由根桥的优先级和 MAC 地址组成，每个 STP 网络中有且仅有一个根。
- 根路径开销：到根桥的最短路径开销。
- 指定桥 ID：由指定桥的优先级和 MAC 地址组成。
- 指定端口 ID：由指定端口的优先级和端口名称组成。
- Message Age：配置 BPDU 在网络中传播的生存期。
- Max Age：配置 BPDU 在设备中能够保存的最大生存期。
- Hello Time：配置 BPDU 发送的周期。
- Forward Delay：端口状态迁移的延时。

一个根桥

树形的网络结构必须有树根，于是 STP 引入了根桥（Root Bridge）概念。

对于一个 STP 网络，根桥在全网中只有一个，它是整个网络的逻辑中心，但不一定是物理中心。根桥会根据网络拓扑的变化而动态变化。

网络收敛后，根桥会按照一定的时间间隔产生并向外发送配置 BPDU，其他设备仅对该报文进行转发，传达拓扑变化记录，从而保证拓扑的稳定。

两种度量

生成树的生成计算有两大基本度量依据：ID 和路径开销。

- ID

  ID 又分为：BID（Bridge ID）和 PID（Port ID）。

  - BID：桥 ID

    IEEE 802.1D 标准中规定 BID 是由 16 位的桥优先级（Bridge Priority）与桥 MAC 地址构成。BID 桥优先级占据高 16 位，其余的低 48 位是 MAC 地址。

    在 STP 网络中，桥 ID 最小的设备会被选举为根桥。在华为技术有限公司的设备上，桥优先级支持手工配置，取值范围是 0 ～ 61440，缺省情况下，桥优先级是 32768。

  - PID：端口 ID

    PID 由两部分构成的，高 4 位是端口优先级，低 12 位是端口号。

    PID 只在某些情况下对选择指定端口有作用。在华为技术有限公司的设备上，端口优先级支持手工配置，取值范围是 0 ～ 240，缺省情况下，端口优先级是 128。

    说明

    端口优先级可以影响端口在指定生成树实例上的角色，详细介绍请见 **STP 拓扑计算**。

- 路径开销

  路径开销（Path Cost）是一个端口量，是 STP 协议用于选择链路的参考值。STP 协议通过计算路径开销，选择较为“强壮”的链路，阻塞多余的链路，将网络修剪成无环路的树形网络结构。

  在一个 STP 网络中，某端口到根桥累计的路径开销就是所经过的各个桥上的各端口的路径开销累加而成，这个值叫做根路径开销（Root Path Cost）。

  IEEE 802.1t 中规定的路径开销如**表 5-2** 所示，而各设备制造商采用的路径开销标准各不相同。

表 5-2 路径开销列表

端口速率	端口模式	STP 路径开销（推荐值）		
		802.1D-1998	802.1T	legacy
0	-	65535	200000000	200,000
10Mbps	Half-Duplex	100	2000000	2,000
	Full-Duplex	99	1999999	1,999
	Aggregated Link 2 Ports	95	1000000	1800
	Aggregated Link 3 Ports	95	666666	1600
	Aggregated Link 4 Ports	95	500000	1400
100Mbps	Half-Duplex	19	200000	200
	Full-Duplex	18	199999	199

端口速率	端口模式	STP 路径开销（推荐值）		
		802.1D-1998	802.1T	legacy
	Aggregated Link 2 Ports	15	100000	180
	Aggregated Link 3 Ports	15	66666	160
	Aggregated Link 4 Ports	15	50000	140
1000Mbps	Full-Duplex	4	20000	20
	Aggregated Link 2 Ports	3	10000	18
	Aggregated Link 3 Ports	3	6666	16
	Aggregated Link 4 Ports	3	5000	14
10Gbps	Full-Duplex	2	2000	2
	Aggregated Link 2 Ports	1	1000	1
	Aggregated Link 3 Ports	1	666	1
	Aggregated Link 4 Ports	1	500	1

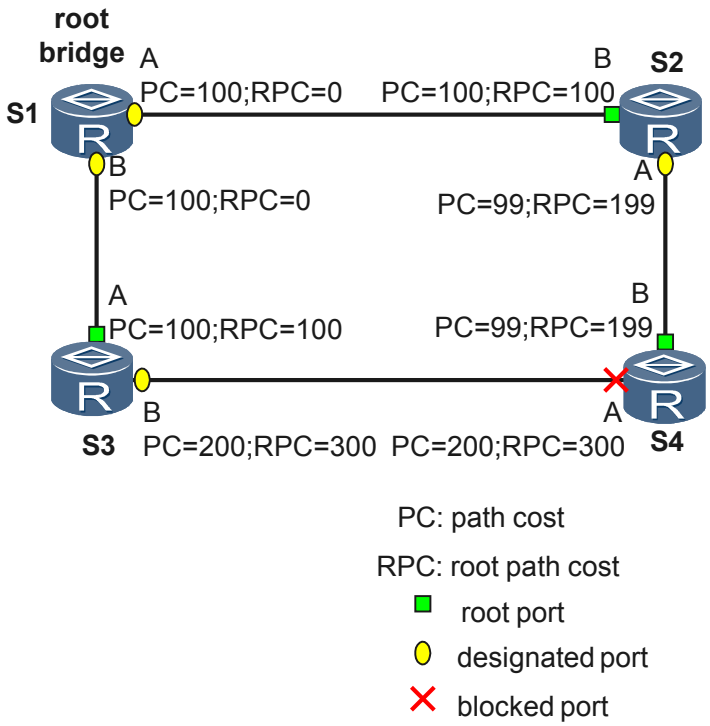
说明

对于聚合链路，链路速率是聚合组中所有状态为 Up 的成员口的速率之和。

三要素选举

从环形网络拓扑结构到树形结构，总体来说有三个要素：根桥、根端口和指定端口。以下结合图 5-2 介绍三要素。

图 5-2 STP 网络结构



- 根桥 RB（Root Bridge）

  根桥就是网桥 ID 最小的桥，通过交互配置 BPDU 协议报文选出最小的 BID。

- 根端口 RP（Root Port）

  所谓根端口就是去往根桥路径开销最小的端口，根端口负责向根桥方向转发数据，这个端口的选择标准是依据根路径开销判定。在一台设备上所有使能 STP 的端口中，根路径开销最小者，就是根端口。很显然，在一个运行 STP 协议的设备上根端口有且只有一个，根桥上没有根端口。

- 指定端口 DP（Designated Port）

  指定桥与指定端口的描述见表 5-3。

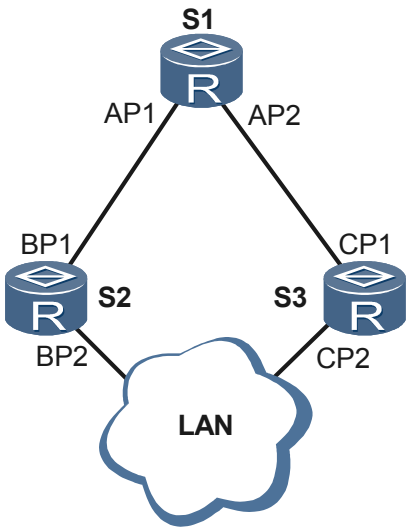
表 5-3 指定桥与指定端口的含义

分类	指定桥	指定端口
对于一台设备而言	与本机直接相连并且负责向本机转发配置消息的设备	指定桥向本机转发配置消息的端口
对于一个局域网而言	负责向本网段转发配置消息的设备	指定桥向本网段转发配置消息的端口

如图 5-3 所示，AP1、AP2、BP1、BP2、CP1、CP2 分别表示设备 S1、S2、S3 的端口。

- S1 通过端口 AP1 向 S2 转发配置消息，则 S2 的指定桥就是 S1，指定端口就是 S1 的端口 AP1。
- 与局域网 LAN 相连的有两台设备：S2 和 S3，如果 S2 负责向 LAN 转发配置消息，则 LAN 的指定桥就是 S2，指定端口就是 S2 的 BP2。

图 5-3 指定桥与指定端口示意图



一旦根桥、根端口、指定端口选举成功，则整个树形拓扑建立完毕。在拓扑稳定后，只有根端口和指定端口转发流量，其他的非根非指定端口都处于阻塞（Blocking）状态，它们只接收 STP 协议报文而不转发用户流量。

四个比较原则

STP 选举有四个比较原则，构成消息优先级向量：{ 根桥 ID，累计根路径开销，发送设备 BID，发送端口 PID }。

配置 BPDU 中携带本端口的主要信息如表 5-4 所示。

表 5-4 四个重要信息字段

字段内容	简要说明
根桥 ID	每个 STP 网络中有且仅有一个根。
累计根路径开销	发送配置 BPDU 的端口到根桥的距离。
发送设备 BID	发送配置 BPDU 的设备的 BID。
发送端口 PID	发出配置 BPDU 的端口的 PID。

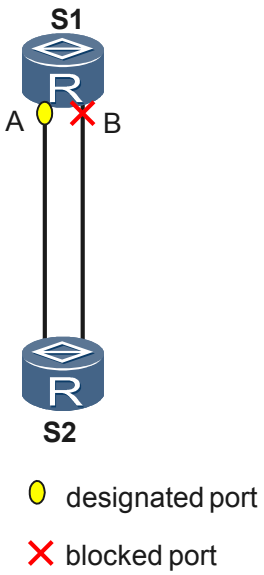
STP 网络中的其他设备收到配置 BPDU 消息后，将比较表 5-4 中所述的字段，四个基本比较原则如下：

说明

在 STP 计算过程中，都遵循数值越小越好的原则。

- 最小 BID：用来选举根桥。运行 STP 协议的设备之间根据表 5-4 所示根桥 ID 字段选择最小的 BID。
- 最小累计根路径开销：用来在非根桥上选择根端口。在根桥上，每个端口到根桥的根路径开销都是 0。
- 最小发送者 BID：当一台运行 STP 协议的设备要在两个以上根路径开销相等的端口之中选择根端口时，通过 STP 协议计算，将选择接收到的配置消息中发送者 BID 较小的那个端口。如图 5-2 所示，假设 S2 的 BID 小于 S3 的 BID，如果 S4 的 A、B 两个端口接收到的 BPDU 里面的根路径开销相等，那么端口 B 将成为根端口。
- 最小 PID：用于在根路径开销相同的情况下，不阻塞最小 PID 的端口，而是阻塞 PID 值较大的端口。如图 5-4 所示的情况下 PID 才起作用，S1 的端口 A 的 PID 小于端口 B 的 PID，由于两个端口上收到的 BPDU 中，根路径开销、发送交换设备 BID 都相同，所以消除环路的依据就只有 PID。

图 5-4 应用到 PID 进行比较的拓扑



五种端口状态

运行 STP 协议的设备上端口状态如表 5-5 所示。

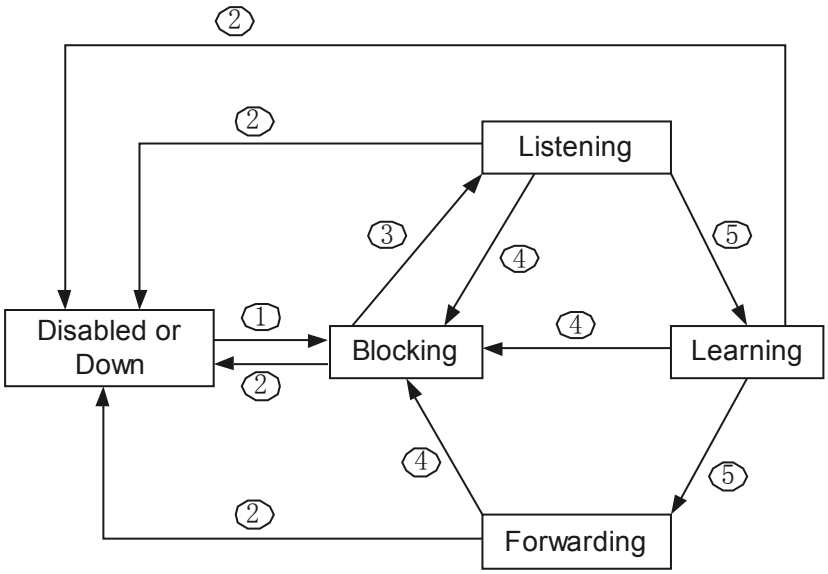
表 5-5 端口状态

端口状态	目的	说明
Forwarding	端口既转发用户流量也转发 BPDU 报文。	只有根端口或指定端口才能进入 Forwarding 状态。
Learning	设备会根据收到的用户流量构建 MAC 地址表，但不转发用户流量。	过渡状态，增加 Learning 状态防止临时环路。

端口状态	目的	说明
Listening	确定端口角色，将选举出根桥、根端口和指定端口。	过渡状态。
Blocking	端口仅仅接收并处理 BPDU，不转发用户流量。	阻塞端口的最终状态。
Disabled	端口不仅不转发 BPDU 报文，也不转发用户流量。	端口状态为 Down。

端口状态迁移机制如图 5-5 所示。

图 5-5 端口状态迁移图



1 端口初始化或使能

2 端口禁用或链路失效

3 端口被选为根端口或指定端口

4 端口不再是根端口或指定端口

5 Forward Delay Timer超时

注意

华为技术有限公司数据通信设备缺省情况处于 MSTP 模式，当从 MSTP 模式切换到 STP 模式，运行 STP 协议的设备上端口支持的端口状态仍然保持和 MSTP 支持的端口状态一样，支持的状态仅包括 Forwarding、Learning 和 Discarding，如表 5-6 所示。

表 5-6 端口状态

端口状态	说明
Forwarding	在这种状态下，端口既转发用户流量又接收/发送 BPDU 报文。
Learning	这是一种过渡状态。在 Learning 下，交换设备会根据收到的用户流量，构建 MAC 地址表，但不转发用户流量，所以叫做学习状态。 Learning 状态的端口接收/发送 BPDU 报文，不转发用户流量。
Discarding	Discarding 状态的端口只接收 BPDU 报文。

对于 STP，影响端口状态和端口收敛有以下 3 个参数。

- Hello Time

  运行 STP 协议的设备发送配置消息 BPDU 的时间间隔，用于设备检测链路是否存在故障。设备每隔 Hello Time 时间会向周围的设备发送 hello 报文，以确认链路是否存在故障。

  当网络拓扑稳定之后，该计时器的修改只有在根桥修改后才有效。新的根桥会在发出的 BPDU 报文中填充适当的字段以向其他非根桥传递该计时器修改的信息。但当拓扑变化之后，TCN BPDU 的发送不受这个计时器的管理。

- Forward Delay

  设备状态迁移的延迟时间。链路故障会引发网络重新进行生成树的计算，生成树的结构将发生相应的变化。不过重新计算得到的新配置消息无法立刻传遍整个网络，如果新选出的根端口和指定端口立刻就开始数据转发的话，可能会造成临时环路。为此，STP 采用了一种状态迁移机制，新选出的根端口和指定端口要经过 2 倍的 Forward Delay 延时后才能进入转发状态，这个延时保证了新的配置消息传遍整个网络，从而防止了临时环路的产生。

  说明

  Forward Delay Timer 指一个端口处于 Listening 和 Learning 状态的各自持续时间，默认是 15 秒。即 Listening 状态持续 15 秒，随后 Learning 状态再持续 15 秒。这两个状态下的端口会处于 Blocking 状态，这正是 STP 用于避免临时环路的关键。

- Max Age

  端口的 BPDU 报文老化时间，可在根桥上通过命令人为改动老化时间。缺省情况下，端口的 BPDU 报文老化时间是 20 秒。

  Max Age 通过配置 BPDU 报文的传输，可保证 Max Age 在整网中一致。运行 STP 协议的网络中非根桥设备收到配置 BPDU 报文后，报文中的 Message Age 和 Max Age 会进行比较：

  - 如果 Message Age 小于等于 Max Age，则该非根桥设备继续转发配置 BPDU 报文。
  - 如果 Message Age 大于 Max Age，则该配置 BPDU 报文将被老化。该非根桥设备直接丢弃该配置 BPDU，可认为网络直径过大，导致根桥连接失败。

  说明

  如果配置 BPDU 是根桥发出的，则 Message Age 为 0。否则，Message Age 是从根桥发送到当前桥接收到 BPDU 的总时间，包括传输延时等。实际实现中，配置 BPDU 报文经过一个桥，Message Age 增加 1。

IEEE 802.1D 中对参数定义如表 5-7。

表 5-7 STP 参数（单位是厘秒）

参数	缺省值	固定值	取值范围
Hello Time	200	-	100 ～ 1000
Max Age	2000	-	600 ～ 4000
Forward Delay	1500	-	400 ～ 3000

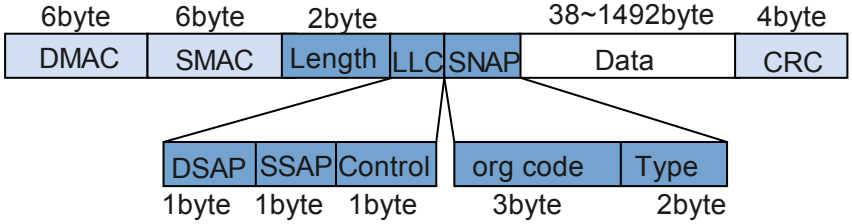
5.4.3 STP 报文格式

在前面的章节中介绍了桥 ID、路径开销和端口 ID 等信息，所有这些信息都是通过 BPDU 协议报文传输。

- 配置 BPDU 是一种心跳报文，只要端口使能 STP，则配置 BPDU 就会按照 Hello Time 定时器规定的时间间隔从指定端口发出。
- TCN BPDU 是在设备检测到网络拓扑发生变化时才发出。

BPDU 报文被封装在以太网数据帧中，目的 MAC 是组播 MAC：01-80-C2-00-00-00，Length/Type 字段为 MAC 数据长度，后面是 LLC 头，IEEE 为 STP 保留了 DSAP 和 SSAP 为 0x42 的值，UI 为 0x03，LLC 之后是 BPDU 报文头。以太网数据帧格式如图 5-6 所示。

图 5-6 以太网数据帧格式



配置 BPDU

通常所说的 BPDU 报文多数指配置 BPDU。

在初始化过程中，每个桥都主动发送配置 BPDU。但在网络拓扑稳定以后，只有根桥主动发送配置 BPDU，其他桥在收到上游传来的配置 BPDU 后，才触发发送自己的配置 BPDU。配置 BPDU 的长度至少要 35 个字节，包含了桥 ID、路径开销和端口 ID 等参数。只有当发送者的 BID 或端口的 PID 两个字段中至少有一个和本桥接收端口不同，BPDU 报文才会被处理，否则丢弃。这样避免了处理和本端口信息一致的 BPDU 报文。

配置 BPDU 报文基本格式如表 5-8 所示。

表 5-8 BPDU 报文基本格式

域	字节	说明
Protocol Identifier	2	总是 0。
Protocol Version Identifier	1	总是 0。
BPDU Type	1	当前 BPDU 类型： ● 0x00：配置 BPDU。 ● 0x80：TCN BPDU。
Flags	1	● 最低位=TC（Topology Change，拓扑变化）标志。 ● 最高位=TCA（Topology Change Acknowledgment，拓扑变化确认）标志。
Root Identifier	8	当前根桥的 BID。
Root Path Cost	4	本端口累计到根桥的开销。
Bridge Identifier	8	本交换设备的 BID。
Port Identifier	2	发送该 BPDU 的端口 ID。
Message Age	2	该 BPDU 的消息年龄。 如果配置 BPDU 是根桥发出的，则 Message Age 为 0。否则，Message Age 是从根桥发送到当前桥接收到 BPDU 的总时间，包括传输延时等。实际实现中，配置 BPDU 报文经过一个桥，Message Age 增加 1。
Max Age	2	消息老化年龄。
Hello Time	2	发送两个相邻 BPDU 的时间间隔。
Forward Delay	2	控制 Listening 和 Learning 状态的持续时间。

标志字段如图 5-7 所示，STP 中只使用了其最高位和最低位。

图 5-7 Flag 字段格式

配置 BPDU 在以下 3 种情况下会产生：

- 只要端口使能 STP，则配置 BPDU 就会按照 Hello Time 定时器规定的时间间隔从指定端口发出。
- 当根端口收到配置 BPDU 时，根端口所在的设备会向自己的每一个指定端口复制一份配置 BPDU。
- 当指定端口收到比自己差的配置 BPDU 时，会立刻向下游设备发送自己的 BPDU。

TCN BPDU

TCN BPDU 内容比较简单，只有表 5-8 中列出的前 3 个字段：协议号、版本和类型。类型字段是固定值 0x80，长度只有 4 个字节。

TCN BPDU 是指在下游拓扑发生变化时向上游发送拓扑变化通知，直到根节点。TCN BPDU 在如下两种情况下会产生：

- 端口状态变为 Forwarding 状态，且该设备上至少有一个指定端口。
- 指定端口收到 TCN BPDU，向根桥复制 TCN BPDU。

5.4.4 STP 拓扑计算

生成树初始化过程

网络中所有的设备使能 STP 协议后，每一台设备都认为自己是根桥。此时，每台设备仅仅收发配置 BPDU，而不转发用户流量，所有的端口都处于 Listening 状态。所有设备通过交换配置 BPDU 后，进行选举工作，选出根桥、根端口和指定端口。

1. 根桥的选择

   如图 5-8 所示，用{}标注的四元组表示了由根桥 ID（图中以 S1_MAC 和 S2_MAC 代表两台设备的 BID）、累计根路径开销、发送者 BID、发送端口 PID 构成的有序组。配置 BPDU 会按照 Hello Timer 规定的时间间隔来发送，默认的时间是 2 秒。

   说明

   由于每个桥都认为自己是根桥，所以在每个端口所发出的 BPDU 中，根桥字段都是用各自的 BID，Root Path Cost 字段是累计的到根桥的开销，发送者 BID 是自己的 BID，端口 PID 是发送该 BPDU 端口的端口 ID。

   图 5-8 初始信息交互

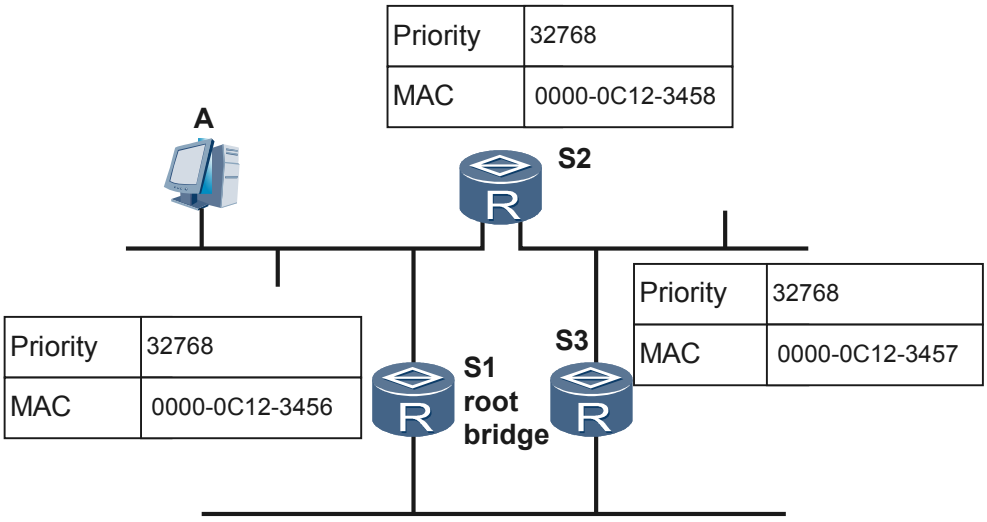
   

   一旦某个端口收到比自己优的 BPDU 报文，此端口就提取该 BPDU 报文中的某些信息更新自己的信息。该端口存储更新后的 BPDU 报文后，立即停止发送 BPDU 报文。

   当端口发送 BPDU 报文时，设备填充 Sender BID 字段的总是自己的 BID，而填充 Root BID 字段的是认为自己是根桥的 BID。如图 5-8 所示，S2 的端口 B 由于接收

到了更好的 BPDU，从而认为此时 S1 是根桥，然后 S2 的其他端口再发送 BPDU 的时候，在根桥字段里面填充的就是 S1_BID 了。此过程不断交互进行，直到所有交换设备的所有端口都认为根桥是相同的，说明根桥已经选择完毕。如图 5-9 所示根桥选举示意图。

图 5-9 根桥选举示意图



2. 根端口的选择

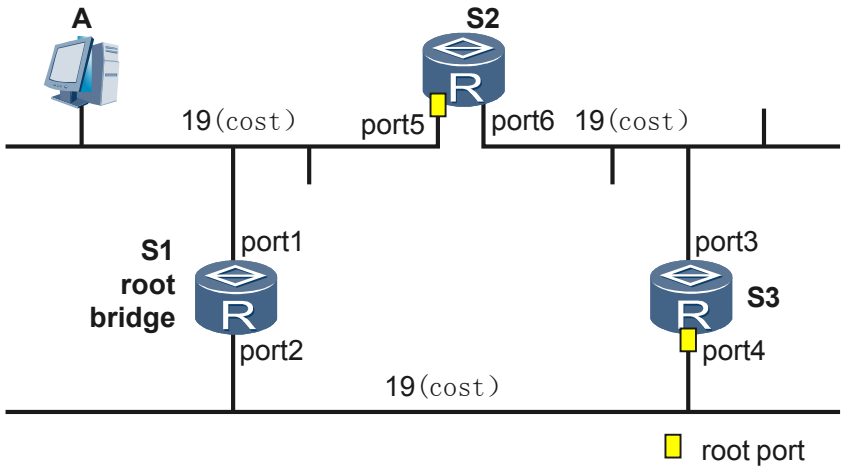
每个非根桥设备都要选择一个根端口，根端口对于一个设备来说有且只有一个。

根端口的本质是距离根桥最近的端口，这个最近的衡量标准是靠累计根路径开销来判定的，即累计根路径开销最小的端口就是根端口。如图 5-10 所示根端口选举示意图。

说明

累计根路径开销的计算方法：端口收到一个 BPDU 报文后，抽取该 BPDU 报文中累计根路径开销字段的值，加上该端口本身的路径开销。所谓该端口本身的路径开销只体现直连链路的开销，这个值是端口量，可以人为配置的。如果有两个以上的端口计算得到的累计根路径开销相同，那么选择收到发送者 BID 最小的那个端口作为根端口。

图 5-10 根端口选举示意图



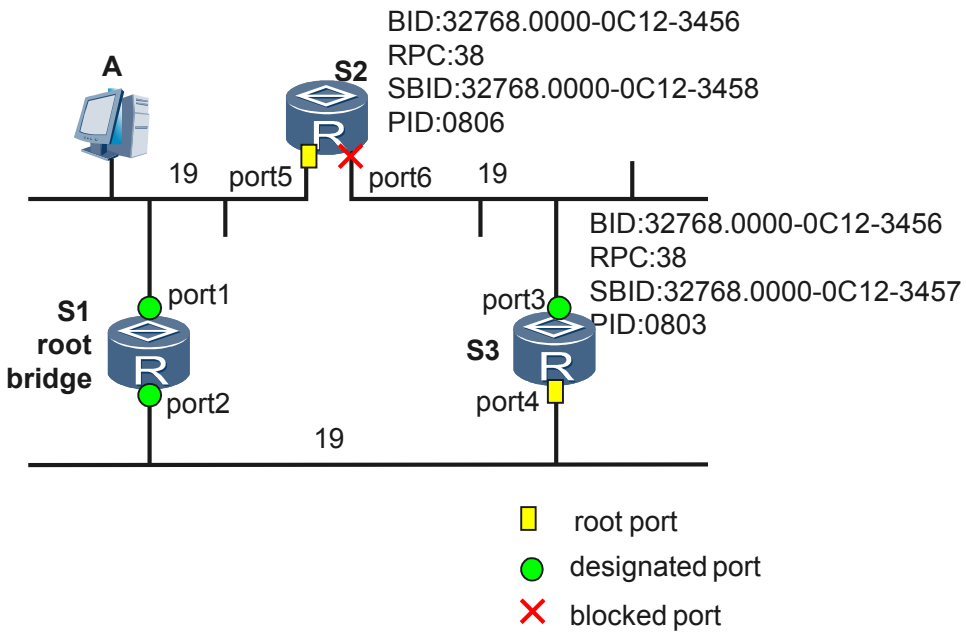
3. 指定端口的选择

在网段上抑制其他端口（无论是自己的还是其他设备的）发送 BPDU 报文的端口，就是该网段的指定端口。如图 5-8 所示，假定 S1 的 MAC 地址小于 S2 的 MAC 地址，则 S1 的端口 A 会成为指定端口。在一个网段上拥有指定端口的设备被称作该网段的指定桥。S1-S2 间网段的指定桥是 S1。

网络收敛后，只有指定端口和根端口可以处于转发状态。其他端口都是 Blocking 状态，不转发用户流量。

根桥的所有端口都是指定端口（除根桥物理上存在环路）。如图 5-11 所示指定端口选举示意图。

图 5-11 指定端口选举示意图



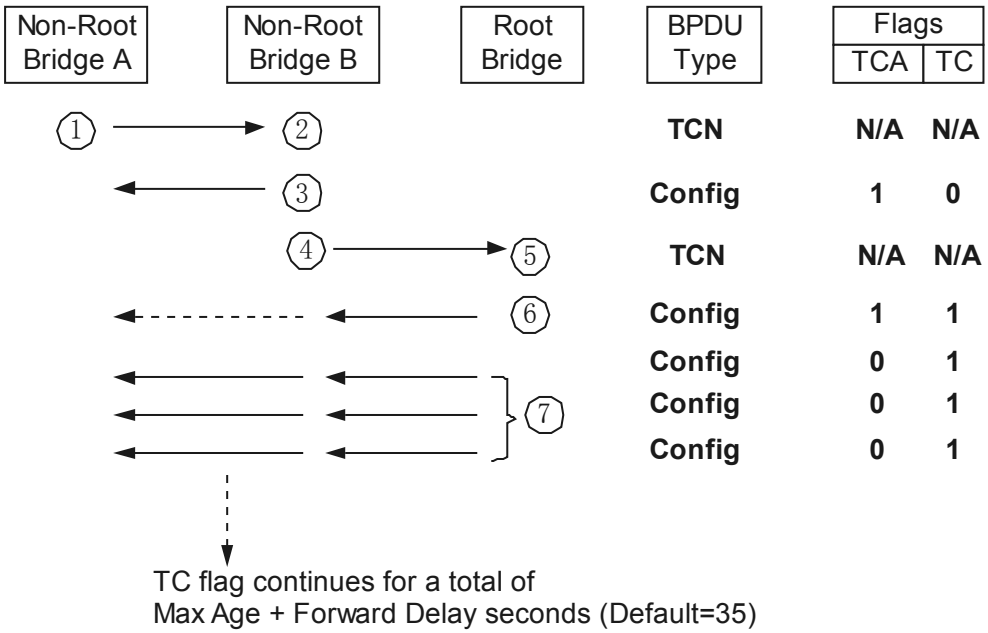
拓扑稳定之后

拓扑稳定后，根桥仍然按照 Hello Timer 规定的时间间隔发送配置 BPDU 报文，非根桥设备从根端口收到配置 BPDU 报文，通过指定端口转发。如果接收到的优先级比自己高的配置 BPDU，则非根桥设备会根据收到的配置 BPDU 中携带的信息更新自己相应的端口存储的配置 BPDU 信息。

STP 拓扑变化

STP 拓扑变化处理过程如图 5-12 所示。

图 5-12 现代交换网络示意图



1. 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。
2. 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。
3. 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。
4. 上游设备复制一份 TCN BPDU 报文，向根桥方向发送。
5. 重复步骤 1、2、3、4，直到根桥收到 TCN BPDU 报文。
6. 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。

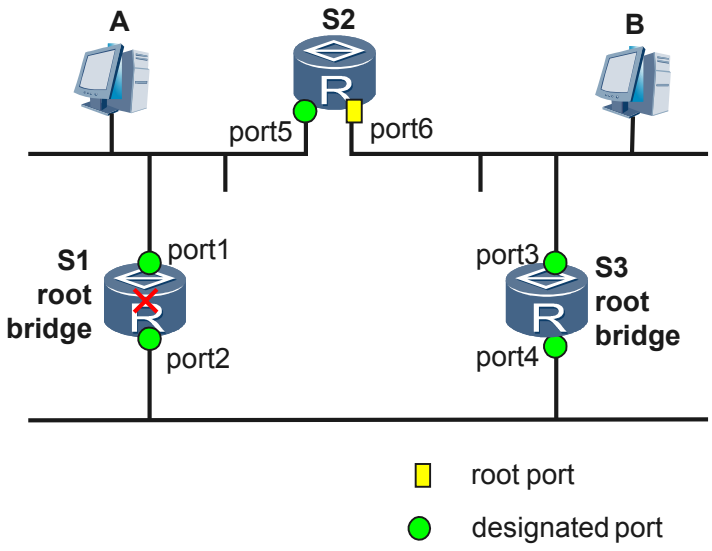
说明

- TCN BPDU 报文主要用来向上游设备乃至根桥通知拓扑变化。
- 置位的 TCA 标记的配置 BPDU 报文主要是上游设备用来告知下游设备已经知道拓扑变化，通知下游设备停止发送 TCN BPDU 报文。
- 置位的 TC 标记的配置 BPDU 报文主要是上游设备用来告知下游设备拓扑发生变化，请下游设备直接删除桥 MAC 地址表项，从而达到快速收敛的目的。

以图 5-11 为例说明根桥、根桥的指定端口分别发生故障时，网络拓扑如何收敛。

- 根桥发生故障

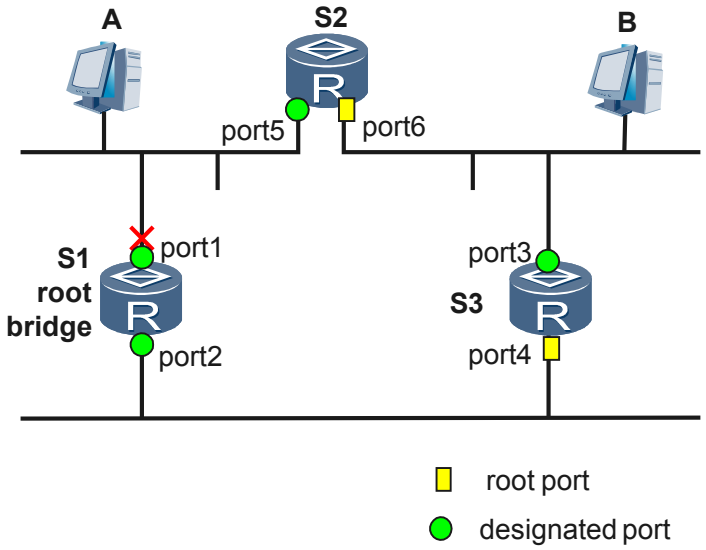
图 5-13 拓扑变化示意图-根桥发生故障



如图 5-13 所示，根桥发生故障，设备 S2 和设备 S3 之间将重新选举根桥。设备 S2 和设备 S3 之间根据交互的配置 BPDU 报文，选出根桥。

- 根桥指定端口发生故障

图 5-14 拓扑变化示意图-根桥指定端口发生故障



如图 5-14 所示，根桥的指定端口 port1 发生故障，S2 和 S3 通过交互配置 BPDU 报文将 port6 选举为根端口。

同时，port6 变为 forwarding 状态后，会向外发送 TCN 报文，根桥收到 TCN 报文后向其他设备发送 TC 报文，通知其他设备直接删除 MAC 表项。

5.4.5 RSTP 对 STP 的改进

IEEE 于 2001 年发布的 802.1W 标准定义了 RSTP（Rapid Spanning-Tree Protocol，快速生成树协议），该协议基于 STP 协议，对原有的 STP 协议进行了更加细致的修改和补充。

STP 的不足之处

STP 协议虽然能够解决环路问题，但是由于网络拓扑收敛慢，影响了用户通信质量。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，这是用户无法忍受的。

STP 的不足之处如下：

- 首先，STP 没有细致区分端口状态和端口角色，不利于初学者学习及部署。

  网络协议的优劣往往取决于协议是否对各种情况加以细致区分。

  - 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。
  - 从使用和配置角度来讲，端口之间最本质的区别并不在于端口状态，而是在于端口扮演的角色。

    根端口和指定端口可以都处于 Listening 状态，也可能都处于 Forwarding 状态。
- 其次，STP 算法是被动的算法，依赖定时器等待的方式判断拓扑变化，收敛速度慢。
- 再次，STP 的算法要求在稳定的拓扑中，根桥主动发出配置 BPDU 报文，而其他设备进行转发，传遍整个 STP 网络。

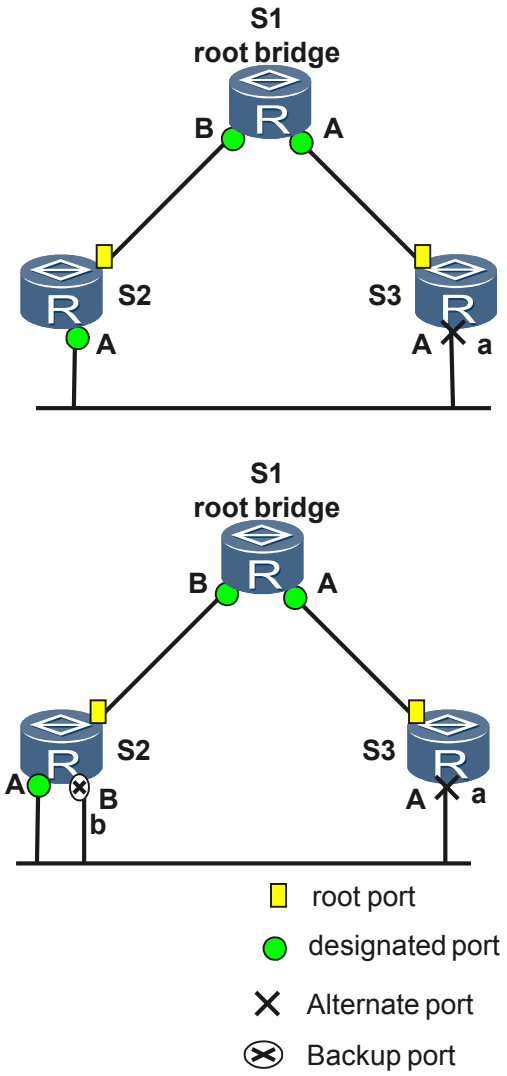
  这也是导致拓扑收敛慢的主要原因之一。

RSTP 对 STP 的改进

根据 STP 的不足，RSTP 删除了 3 种端口状态，新增加了 2 种端口角色，并且把端口属性充分的按照状态和角色解耦，使得可以更加精确的描述端口，从而使得初学者更易学习协议，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。

图 5-15 端口角色示意图



如图 5-15 所示，RSTP 的端口角色共有 4 种：根端口、指定端口、Alternate 端口和 Backup 端口。

根端口和指定端口的作用同 STP 协议中定义，Alternate 端口和 Backup 端口的描述如下：

- 从配置 BPDU 报文发送角度来看：
  - Alternate 端口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 端口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 端口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 端口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。根据端口是否转发用户流量和学习 MAC 地址来划分:

  - 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。
  - 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。
  - 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。

  如表 5-9 所示，新的端口状态与 STP 规定的端口状态比较。

  说明

  端口状态和端口角色是没有必然联系的，表 5-9 显示了各种端口角色能够具有的端口状态。

表 5-9 STP 与 RSTP 端口状态角色对应表

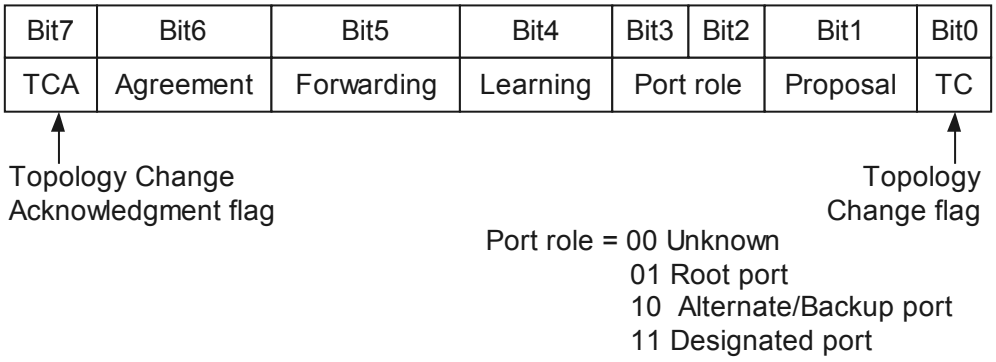
STP 端口状态	RSTP 端口状态	端口在拓扑中的角色
Forwarding	Forwarding	包括根端口、指定端口
Learning	Learning	包括根端口、指定端口
Listening	Discarding	包括根端口、指定端口
Blocking	Discarding	不包括 Alternate 端口、Backup 端口
Disabled	Discarding	不包括 Disable

- 配置 BPDU 格式的改变，充分利用了 STP 协议报文中的 Flag 字段，明确了端口角色。

  在配置 BPDU 报文的格式上，除了保证和 STP 格式基本一致之外，RSTP 作了一些小变化：

  - Type 字段，配置 BPDU 类型不再是 0 而是 2，所以运行 STP 的设备收到 RSTP 的配置 BPDU 时会丢弃。
  - Flag 字段，使用了原来保留的中间 6 位，这样改变的配置 BPDU 叫做 RST BPDU，如图 5-16 所示。

图 5-16 RSTP Flag 字段格式



- 配置 BPDU 的处理发生变化

- 拓扑稳定后，配置 BPDU 报文的发送方式

  拓扑稳定后，根桥按照 Hello Timer 规定的时间间隔发送配置 BPDU。其他非根桥设备在收到上游设备发送过来的配置 BPDU 后，才会触发发出配置 BPDU，此方式使得 STP 协议计算复杂且缓慢。RSTP 对此进行了改进，即在拓扑稳定后，无论非根桥设备是否接收到根桥传来的配置 BPDU 报文，非根桥设备仍然按照 Hello Timer 规定的时间间隔发送配置 BPDU，该行为完全由每台设备自主进行。

- 更短的 BPDU 超时计时

  如果一个端口连续 3 个 Hello Time 时间内没有收到上游设备发送过来的配置 BPDU，那么该设备认为与此邻居之间的协商失败。而不像 STP 那样需要先等待一个 Max Age。

- 处理次等 BPDU

  当一个端口收到上游的指定桥发来的 RST BPDU 报文时，该端口会将自身存储的 RST BPDU 与收到的 RST BPDU 进行比较。

  如果该端口存储的 RST BPDU 的优先级高于收到的 RST BPDU，那么该端口会直接丢弃收到的 RST BPDU，立即回应自身存储的 RST BPDU。当上游设备收到下游设备回应的 RST BPDU 后，上游设备会根据收到的 RST BPDU 报文中相应的字段立即更新自己存储的 RST BPDU。

  由此，RSTP 处理次等 BPDU 报文不再依赖于任何定时器通过超时解决拓扑收敛，从而加快了拓扑收敛。

- 快速收敛
  - Proposal/Agreement 机制

    当一个端口被选举成为指定端口之后，在 STP 中，该端口至少要等待一个 Forward Delay（Learning）时间才会迁移到 Forwarding 状态。而在 RSTP 中，此端口会先进入 Discarding 状态，再通过 Proposal/Agreement 机制快速进入 Forward 状态。这种机制必须在点到点全双工链路上使用。

    Proposal/Agreement 机制简称 P/A 机制，详细描述请参见 **RSTP 技术细节**中的 P/A 协商。

  - 根端口快速切换机制

    如果网络中一个根端口失效，那么网络中最优的 Alternate 端口将成为根端口，进入 Forwarding 状态。因为通过这个 Alternate 端口连接的网段上必然有个指定端口可以通往根桥。

    这种产生新的根端口的过程会引发拓扑变化，详细描述请见 **RSTP 技术细节**中的 RSTP 拓扑变化处理。

  - 边缘端口的引入

    在 RSTP 里面，如果某一个指定端口位于整个网络的边缘，即不再与其他交换设备连接，而是直接与终端设备直连，这种端口叫做边缘端口。

    边缘端口不接收处理配置 BPDU，不参与 RSTP 运算，可以由 Disable 直接转到 Forwarding 状态，且不经历时延，就像在端口上将 STP 禁用。但是一旦边缘端口收到配置 BPDU，就丧失了边缘端口属性，成为普通 STP 端口，并重新进行生成树计算，从而引起网络震荡。

- 保护功能

  RSTP 提供的保护功能如**表 5-10** 所示。

表 5-10 保护功能

保护功能	场景	原理
BPDU 保护	在交换设备上，通常将直接与用户终端（如 PC 机）或文件服务器等非交换设备相连的端口配置为边缘端口。 正常情况下，边缘端口不会收到 RST BPDU。如果有人伪造 RST BPDU 恶意攻击交换设备，当边缘端口接收到 RST BPDU 时，交换设备会自动将边缘端口设置为非边缘端口，并重新进行生成树计算，从而引起网络震荡。	交换设备上启动了 BPDU 保护功能后，如果边缘端口收到 RST BPDU，边缘端口将被 shutdown，但是边缘端口属性不变，同时通知网管系统。被 shutdown 的边缘端口只能由网络管理员手动恢复。
根保护	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。这种不合法的拓扑变化，会导致原来应该通过高速链路的流量被牵引到低速链路上，造成网络拥塞。	对于启用 Root 保护功能的指定端口，其端口角色只能保持为指定端口。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（通常为两倍的 Forward Delay），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。 **说明** Root 保护功能只能在指定端口上配置生效。
环路保护	在运行 RSTP 协议的网络中，根端口和其他阻塞端口状态是依靠不断接收来自上游交换设备的 RST BPDU 维持。 当由于链路拥塞或者单向链路故障导致这些端口收不到来自上游交换设备的 RST BPDU 时，此时交换设备会重新选择根端口。原先的根端口会转变为指定端口，而原先的阻塞端口会迁移到转发状态，从而造成交换网络中可能产生环路。	在启动了环路保护功能后，如果根端口或 Alternate 端口长时间收不到来自上游的 RST BPDU 时，则向网管发出通知信息（如果是根端口则进入 Discarding 状态）。而阻塞端口则会一直保持在阻塞状态，不转发报文，从而不会在网络中形成环路。直到根端口或 Alternate 端口收到 RST BPDU，端口状态才恢复正常到 Forwarding 状态。 **说明** 环路保护功能只能在根端口或 Alternate 端口上配置生效。

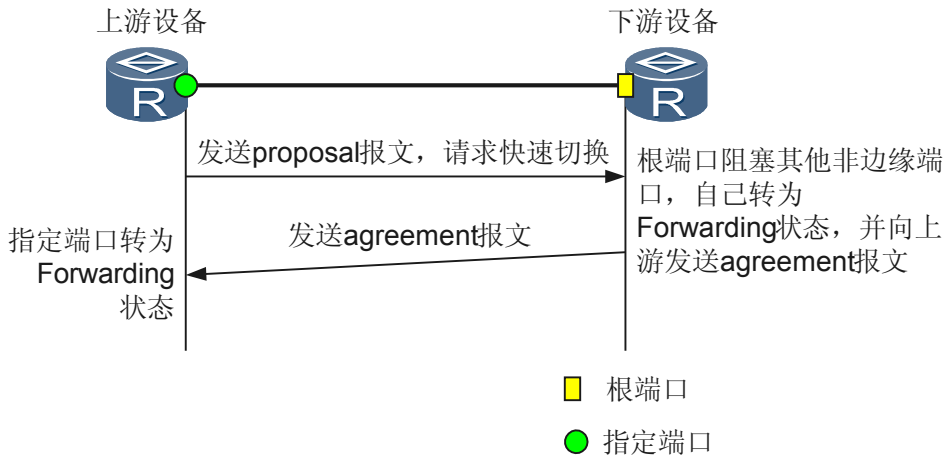
保护功能	场景	原理
防 TC-BPDU 攻击	交换设备在接收到 TC BPDU 报文后，会执行 MAC 地址表项和 ARP 表项的删除操作。如果有人伪造 TC BPDU 报文恶意攻击交换设备时，交换设备短时间内会收到很多 TC BPDU 报文，频繁的删除操作会给设备造成很大的负担，给网络的稳定带来很大隐患。	启用防 TC-BPDU 报文攻击功能后，在单位时间内，交换设备处理 TC BPDU 报文的次数可配置（缺省的单位时间是 2 秒，缺省的处理次数是 3 次）。如果在单位时间内，交换设备在收到 TC BPDU 报文数量大于配置的阈值，那么设备只会处理阈值指定的次数。对于其他超出阈值的 TCN BPDU 报文，定时器到期后设备只对其统一处理一次。这样可以避免频繁的删除 MAC 地址表项和 ARP 表项，从而达到保护设备的目的。

5.4.6 RSTP 技术细节

P/A 机制

P/A 机制即 Proposal/Agreement 机制，其目的是使一个指定端口尽快进入 Forwarding 状态。如图 5-17 所示，P/A 协商过程的完成根据以下几个端口变量：

图 5-17 Proposal/Agreement 过程示意图

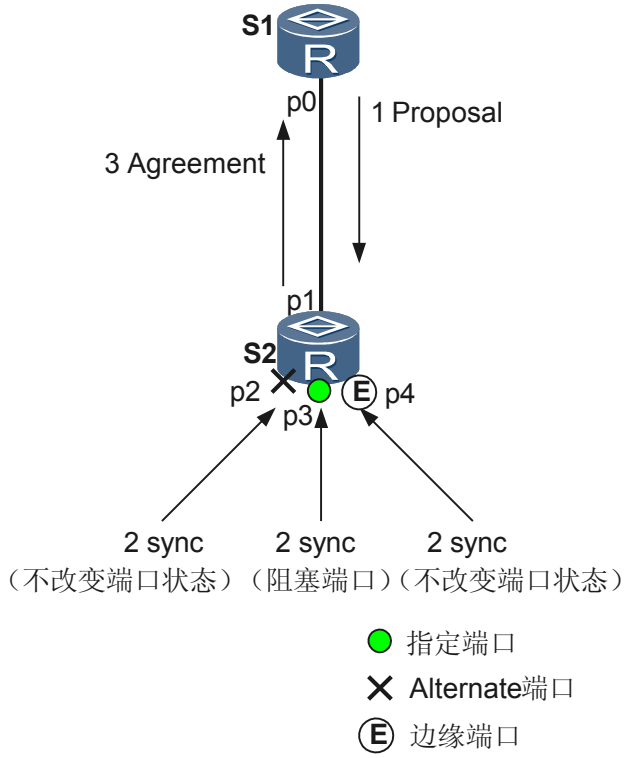


1. proposing：当一个指定端口处于 Discarding 或 Learning 状态时，该变量置位，并向下游交换设备传递 Proposal 位被置位的 RST BPDU。
2. proposed：当端口收到对端的指定端口发来的携带 Proposal 的 RST BPDU 时，该变量置位。该变量指示本网段上的指定端口希望尽快进入 Forwarding 状态。
3. sync：当 Proposed 被置位以后，收到 Proposal 的根端口会依次为自己的其他端口置位 sync 变量。而收到 Proposal 的非边缘端口则会进入 Discarding 状态。
4. synced：当端口转到 Discarding 状态后，会将自己的 synced 变量置位。Alternate 端口、Backup 端口和边缘端口会马上置位该变量。根端口监视其他端口的 synced，

当所有其他端口的 synced 全被置位，根端口会将自己的 synced 置位，然后传回 RST BPDU，其中 Agreement 位被置位。

5. agreed：当指定端口接收到一个 RST BPDU 时，如果该 BPDU 中的 Agreement 位被置位且端口角色字段是根端口，该变量被置位。Agreed 变量一旦被置位，指定端口马上转入 Forwarding 状态。

图 5-18 Proposal/Agreement 过程示意图



如图 5-18 所示，根桥 S1 和 S2 之间新添加了一条链路。在当前状态下，S2 的另外几个端口 p2 是 Alternate 端口，p3 是指定端口且处于 Forwarding 状态，p4 是边缘端口。新链路连接成功后，P/A 机制协商过程如下：

1. p0 和 p1 两个端口马上都先成为指定端口，发送 RST BPDU。
2. S2 的 p1 口收到更优的 RST BPDU，马上意识到自己将成为根端口，而不是指定端口，停止发送 RST BPDU。
3. S1 的 p0 进入 Discarding 状态，于是发送的 RST BPDU 中把 proposal 置 1。
4. S2 收到根桥发送来的携带 proposal 的 RST BPDU，开始将自己的所有端口进入 sync 变量置位。
5. p2 已经阻塞，状态不变；p4 是边缘端口，不参与运算；所以只需要阻塞非边缘指定端口 p3。
6. p2、p3、p4 都进入 Discarding 状态之后，各端口的 synced 变量置位，根端口 p1 的 synced 也置位，于是便向 S1 返回 Agreement 位置位的回应 RST BPDU。该 RST BPDU 携带和刚才根桥发过来的 BPDU 一样的信息，除了 Agreement 位置位之外（Proposal 位清零）。

7. 当 S1 判断出这是对刚刚发出的 Proposal 的回应，于是端口 p0 马上进入 Forwarding 状态。

以上 P/A 过程可以向下游继续传递。

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一个 Forward Delay 所有端口才能进行转发。而 RSTP 的主要目的就是消除这个瓶颈，通过阻塞自己的非根端口来保证不会出现环路。而使用 P/A 机制加快了上游端口转到 Forwarding 状态的速度。

说明

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。

RSTP 拓扑变化处理

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。

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。

RSTP 与 STP 的互操作

RSTP 可以和 STP 互操作，但是此时会丧失快速收敛等 RSTP 优势。

当一个网段里既有运行 STP 的交换设备又有运行 RSTP 的交换设备，STP 交换设备会忽略 RST BPDU，而运行 RSTP 的交换设备在某端口上接收到运行 STP 的交换设备发出的配置 BPDU，在两个 Hello Time 时间之后，便把自己的端口转换到 STP 工作模式，发送配置 BPDU。这样，就实现了互操作。

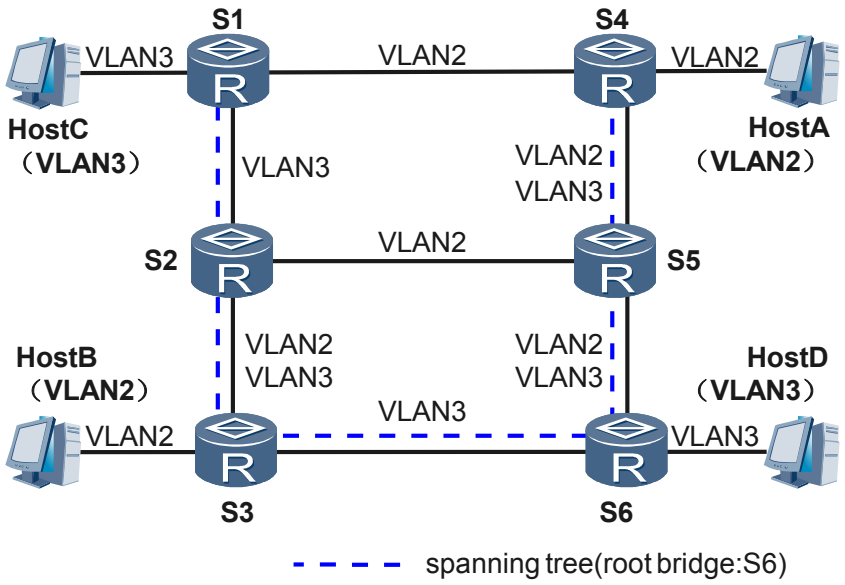
在华为技术有限公司的数据通信设备上可以配置运行 STP 的交换设备被撤离网络后，运行 RSTP 的交换设备可迁移回到 RSTP 工作模式。

5.5 MSTP 原理描述

5.5.1 MSTP 出现的背景

RSTP 在 STP 基础上进行了改进，实现了网络拓扑快速收敛。但 RSTP 和 STP 还存在同一个缺陷：由于局域网内所有的 VLAN 共享一棵生成树，因此无法在 VLAN 间实现数据流量的负载均衡，链路被阻塞后将不承载任何流量，造成带宽浪费，还有可能造成部分 VLAN 的报文无法转发。

图 5-19 STP/RSTP 的缺陷示意图



如图 5-19 所示网络中，在局域网内应用 STP 或 RSTP，生成树结构在图中用虚线表示，S6 为根交换设备。S2 和 S5 之间、S1 和 S4 之间的链路被阻塞，除了图中标注了“VLAN2”或“VLAN3”的链路允许对应的 VLAN 报文通过外，其它链路均不允许 VLAN2、VLAN3 的报文通过。

HostA 和 HostB 同属于 VLAN2，由于 S2 和 S5 之间的链路被阻塞，S3 和 S6 之间的链路又不允许 VLAN2 的报文通过，因此 HostA 和 HostB 之间无法互相通讯。

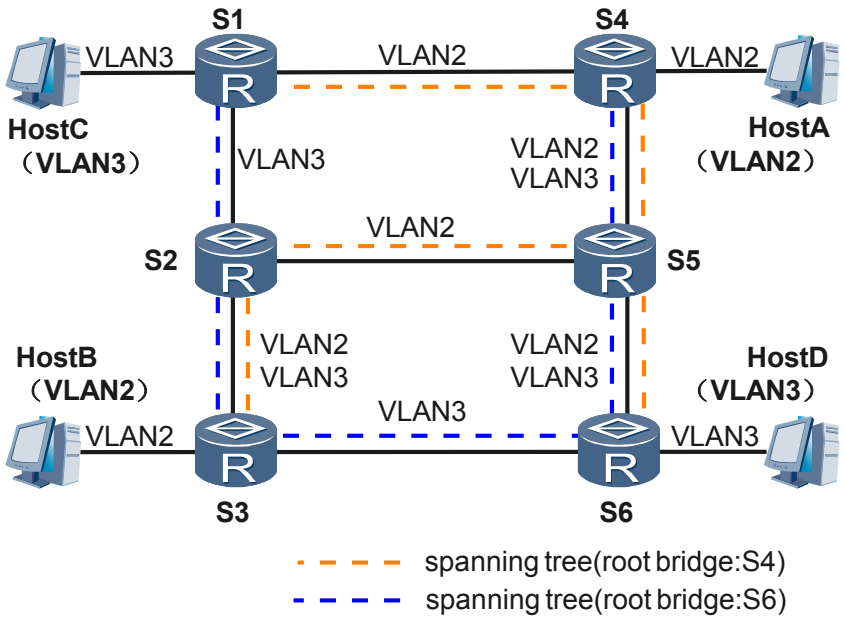
为了弥补 STP 和 RSTP 的缺陷，IEEE 于 2002 年发布的 802.1S 标准定义了 MSTP。MSTP 兼容 STP 和 RSTP，既可以快速收敛，又提供了数据转发的多个冗余路径，在数据转发过程中实现 VLAN 数据的负载均衡。

通过 MSTP 把一个交换网络划分成多个域，每个域内形成多棵生成树，生成树之间彼此独立。每棵生成树叫做一个多生成树实例 MSTI（Multiple Spanning Tree Instance），每个域叫做一个 MST 域（MST Region：Multiple Spanning Tree Region）。

说明

所谓实例就是多个 VLAN 的一个集合。通过将多个 VLAN 捆绑到一个实例，可以节省通信开销和资源占用率。MSTP 各个实例拓扑的计算相互独立，在这些实例上可以实现负载均衡。可以把多个相同拓扑结构的 VLAN 映射到一个实例里，这些 VLAN 在端口上的转发状态取决于端口在对应 MSTP 实例的状态。

图 5-20 MST 域内的多棵生成树示意图



如图 5-20 所示，MSTP 通过设置 VLAN 映射表（即 VLAN 和 MSTI 的对应关系表），把 VLAN 和 MSTI 联系起来。每个 VLAN 只能对应一个 MSTI，即同一 VLAN 的数据只能在一个 MSTI 中传输，而一个 MSTI 可能对应多个 VLAN。

经计算，最终生成两棵生成树：

- MSTI1 以 S4 为根交换设备，转发 VLAN2 的报文。
- MSTI2 以 S6 为根交换设备，转发 VLAN3 的报文。

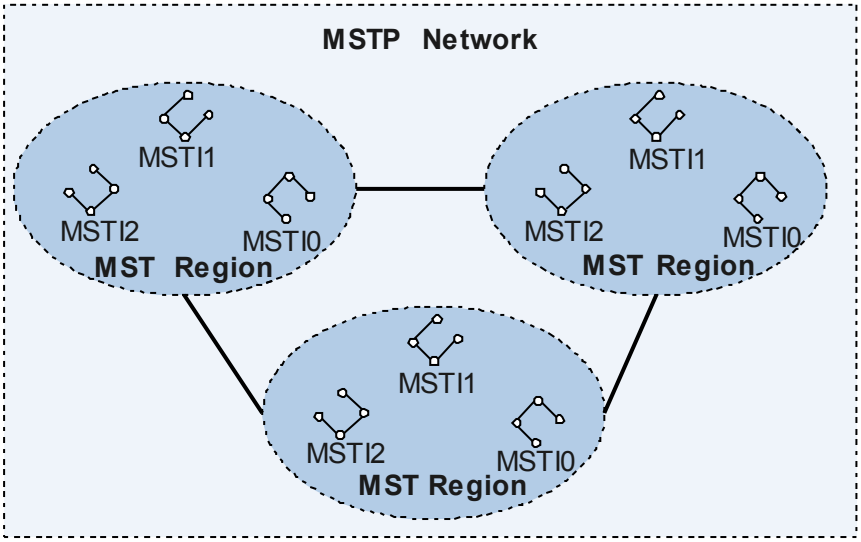
这样所有 VLAN 内部可以互通，同时不同 VLAN 的报文沿不同的路径转发，实现了负载分担。

5.5.2 MSTP 基本概念

MSTP 的网络层次

如图 5-21 所示，MSTP 网络中包含 1 个或多个 MST 域（MST Region），每个 MST Region 中包含一个或多个 MSTI。组成 MSTI 的是运行 STP/RSTP/MSTP 的交换设备，MSTI 是所有运行 STP/RSTP/MSTP 的交换设备经 MSTP 协议计算后形成的树状网络。

图 5-21 MSTP 网络层次示意图



MST 域（MST Region）

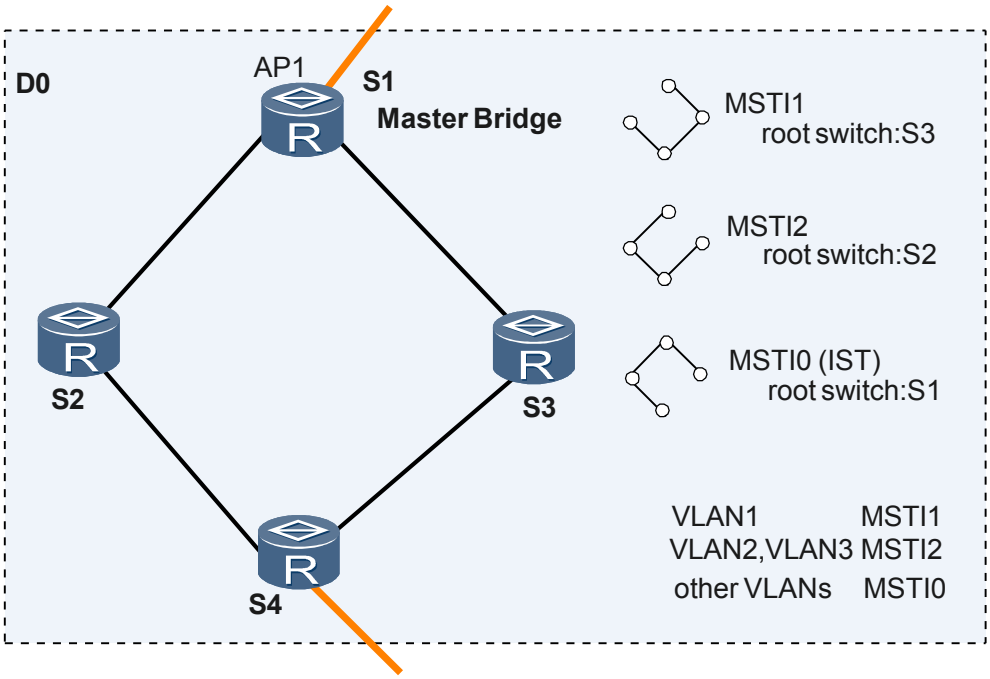
MST 域是多生成树域（Multiple Spanning Tree Region），由交换网络中的多台交换设备以及它们之间的网段所构成。同一个 MST 域的设备具有下列特点：

- 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。

一个局域网可以存在多个 MST 域，各 MST 域之间在物理上直接或间接相连。用户可以通过 MSTP 配置命令把多台交换设备划分在同一个 MST 域内。

如图 5-22 所示的 MST Region D0 中由交换设备 S1、S2、S3 和 S4 构成，域中有 3 个 MSTI。

图 5-22 MST Region 的基本概念示意图



VLAN 映射表

VLAN 映射表是 MST 域的属性，它描述了 VLAN 和 MSTI 之间的映射关系。

如图 5-22 所示，MST 域 D0 的 VLAN 映射表是：

- VLAN1 映射到 MSTI1
- VLAN2 和 VLAN3 映射到 MSTI2
- 其余 VLAN 映射到 MSTI0

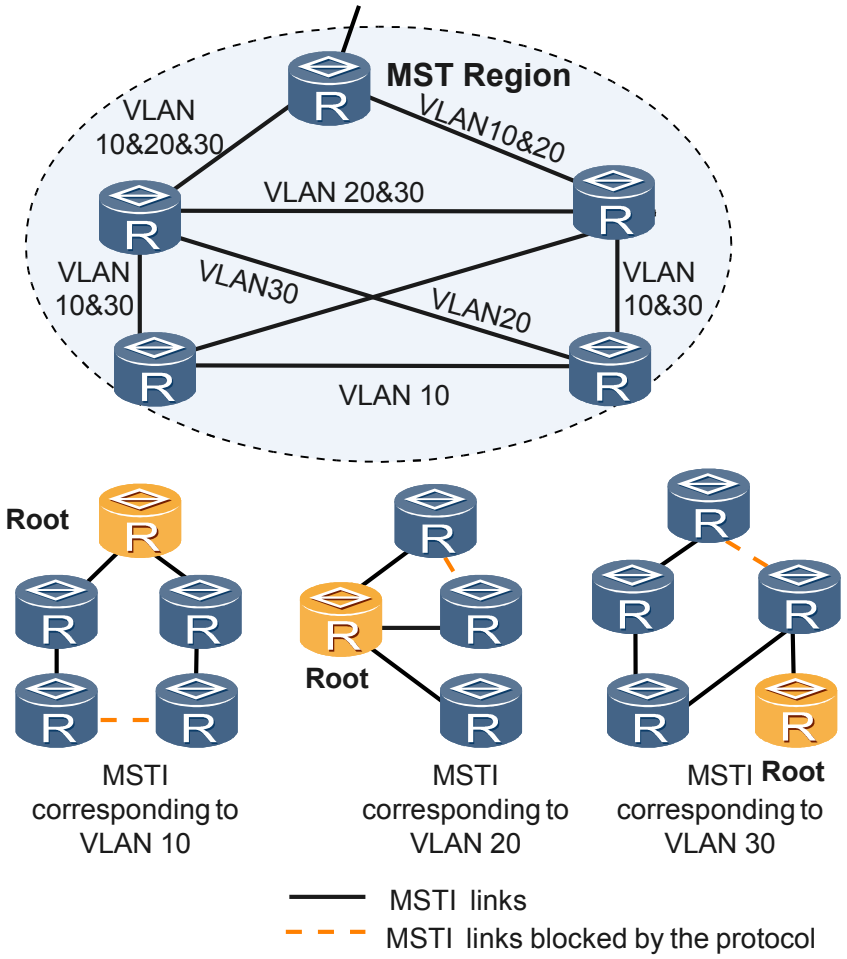
域根

域根（Regional Root）分为 IST（Internal Spanning Tree）域根和 MSTI 域根。

IST 域根如图 5-24 所示，在 B0、C0 和 D0 中，IST 生成树中距离总根（CIST Root）最近的交换设备是 IST 域根。

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。MSTI 域根是每个多生成树实例的树根。如图 5-23 所示，域中不同的 MSTI 有各自的域根。

图 5-23 MSTI 的基本概念示意图



MSTI 之间彼此独立，MSTI 可以与一个或者多个 VLAN 对应。但一个 VLAN 只能与一个 MSTI 对应。

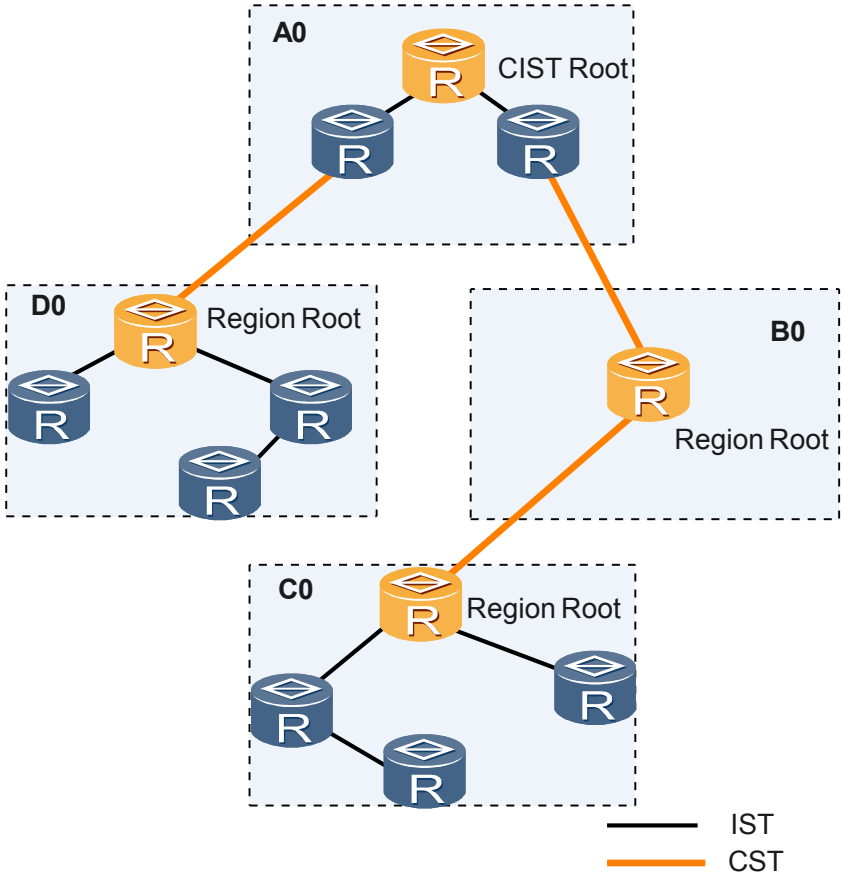
主桥

主桥（Master Bridge）也就是 IST Master，它是域内距离总根最近的交换设备。如图 5-22 中的 S1。

如果总根在 MST 域中，则总根为该域的主桥。

总根

图 5-24 MSTP 网络基本概念示意图



如图 5-24 所示，总根是 CIST（Common and Internal Spanning Tree）的根桥。总根是区域 A0 中的某台设备。

CST

公共生成树 CST（Common Spanning Tree）是连接交换网络内所有 MST 域的一棵生成树。

如果把每个 MST 域看作是一个节点，CST 就是这些节点通过 STP 或 RSTP 协议计算生成的一棵生成树。

如图 5-24 所示，较粗的线条连接各个域构成 CST。

IST

内部生成树 IST（Internal Spanning Tree）是各 MST 域内的一棵生成树。

IST 是一个特殊的 MSTI，MSTI 的 ID 为 0，通常称为 MSTI0。

IST 是 CIST 在 MST 域中的一个片段。

如图 5-24 所示，较细的线条在域中连接该域的所有交换设备构成 IST。

CIST

公共和内部生成树 CIST 是通过 STP 或 RSTP 协议计算生成的，连接一个交换网络内所有交换设备的单生成树。

如图 5-24 所示，所有 MST 域的 IST 加上 CST 就构成一棵完整的生成树，即 CIST。

SST

构成单生成树 SST（Single Spanning Tree）有两种情况：

- 运行 STP 或 RSTP 的交换设备只能属于一个生成树。
- MST 域中只有一个交换设备，这个交换设备构成单生成树。

如图 5-24 所示，B0 中的交换设备就是一棵单生成树。

端口角色

MSTP 在 RSTP 的基础上新增了 2 种端口，MSTP 的端口角色共有 7 种：根端口、指定端口、Alternate 端口、Backup 端口、边缘端口、Master 端口和域边缘端口。

根端口、指定端口、Alternate 端口、Backup 端口和边缘端口的作用同 RSTP 协议中定义，MSTP 中定义的所有端口角色如表 5-11 所示。

说明

除边缘端口外，其他端口角色都参与 MSTP 的计算过程。

同一端口在不同的生成树实例中可以担任不同的角色。

表 5-11 端口角色

端口角色	说明
根端口	在非根桥上，离根桥最近的端口是本交换设备的根端口。根交换设备没有根端口。 根端口负责向树根方向转发数据。 如图 5-25 所示，S1 为根桥，CP1 为 S3 的根端口，BP1 为 S2 的根端口。
指定端口	对一台交换设备而言，它的指定端口是向下游交换设备转发 BPDU 报文的端口。 如图 5-25 所示，AP2 和 AP3 为 S1 的指定端口，CP2 为 S3 的指定端口。
Alternate 端口	● 从配置 BPDU 报文发送角度来看，Alternate 端口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。 ● 从用户流量角度来看，Alternate 端口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。 如图 5-25 所示，BP2 为 Alternate 端口。
Backup 端口	● 从配置 BPDU 报文发送角度来看，Backup 端口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，Backup 端口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。 如图 5-25 所示，CP3 为 Backup 端口。

端口角色	说明
Master 端口	Master 端口是 MST 域和总根相连的所有路径中最短路径上的端口，它是交换设备上连接 MST 域到总根的端口。 Master 端口是域中的报文去往总根的必经之路。 Master 端口是特殊域边缘端口，Master 端口在 IST/CIST 上的角色是 Root Port，在其它各实例上的角色都是 Master 端口。 如图 5-26 所示，交换设备 S1、S2、S3、S4 和它们之间的链路构成一个 MST 域，S1 交换设备的端口 AP1 在域内的所有端口中到总根的路径开销最小，所以 AP1 为 Master 端口。
域边缘端口	域边缘端口是指位于 MST 域的边缘并连接其它 MST 域或 SST 的端口。 进行 MSTP 计算时，域边缘端口在 MSTI 上的角色和 CIST 实例的角色保持一致。即如果边缘端口在 CIST 实例上的角色是 Master 端口（连接域到总根的端口），则它在域内所有 MSTI 上的角色也是 Master 端口。 如图 5-26 所示，MST 域内的 AP1、DP1 和 DP2 都和其它域直接相连，它们都是本 MST 域的边缘端口。 域边缘端口在生成树实例上的角色与在 CIST 的角色保持一致。如图 5-26，AP1 是域边缘端口，它在 CIST 上的角色是 Master 端口，则 AP1 在 MST 域内所有生成树实例上的角色都是 Master 端口。
边缘端口	如果指定端口位于整个域的边缘，不再与任何交换设备连接，这种端口叫做边缘端口。 边缘端口一般与用户终端设备直接连接。

图 5-25 根端口、指定端口、Alternate 端口和 Backup 端口示意图

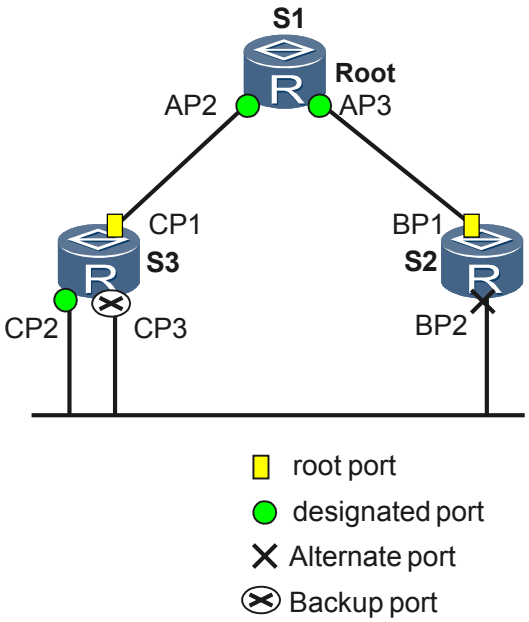
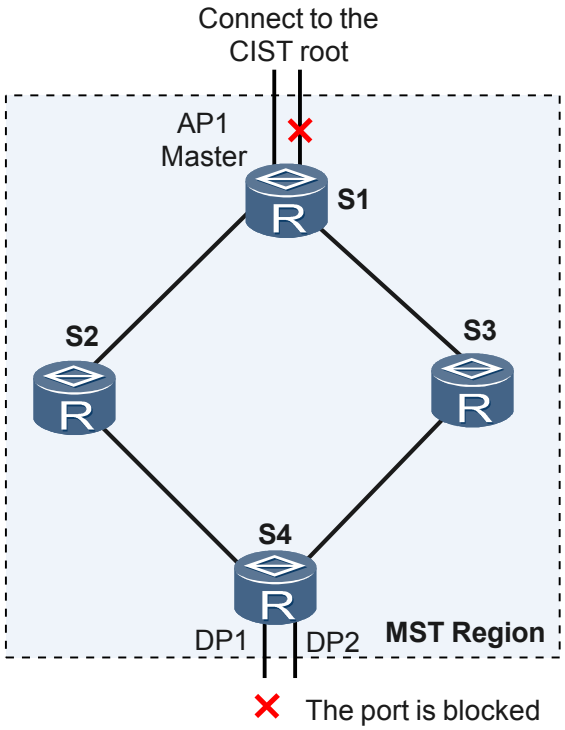


图 5-26 Master 端口和域边缘端口示意图



MSTP 的端口状态

MSTP 定义的端口状态与 RSTP 协议中定义相同，如表 5-12 所示。

表 5-12 端口状态

端口状态	说明
Forwarding	在这种状态下，端口既转发用户流量又接收/发送 BPDU 报文。
Learning	这是一种过渡状态。在 Learning 下，交换设备会根据收到的用户流量，构建 MAC 地址表，但不转发用户流量，所以叫做学习状态。 Learning 状态的端口接收/发送 BPDU 报文，不转发用户流量。
Discarding	Discarding 状态的端口只接收 BPDU 报文。

端口状态和端口角色是没有必然联系的，表 5-13 显示了各种端口角色能够具有的端口状态。

表 5-13 端口状态和端口角色对应表

端口状态	根端口/Master 端口	指定端口	域边缘端口	Alternate 端口	Backup 端口
Forwarding	Yes	Yes	Yes	No	No
Learning	Yes	Yes	Yes	No	No
Discarding	Yes	Yes	Yes	Yes	Yes

Yes：表示端口支持的状态。

No：表示端口不支持的状态。

5.5.3 MSTP 报文

MSTP 使用多生成树桥协议数据单元 MST BPDU（Multiple Spanning Tree Bridge Protocol Data Unit）作为生成树计算的依据。MST BPDU 报文用来计算生成树的拓扑、维护网络拓扑以及传达拓扑变化记录。

STP 中定义的配置 BPDU、RSTP 中定义的 RST BPDU、MSTP 中定义的 MST BPDU 及 TCN BPDU 差异对比如表 5-14 所示。

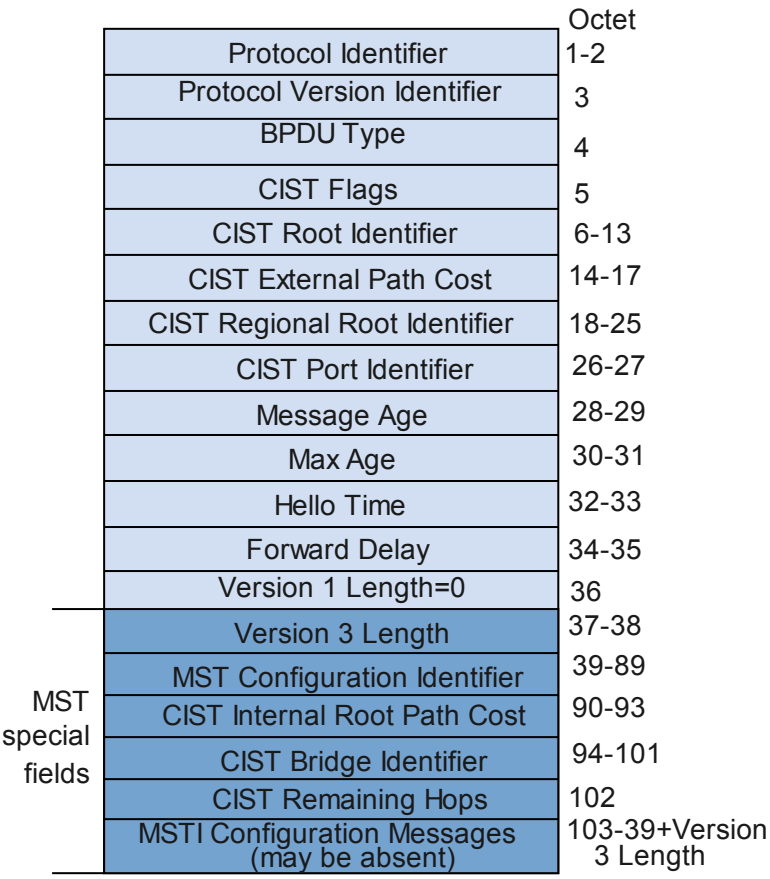
表 5-14 四种 BPDU 差异比较

版本	类型	名称
0	0x00	配置 BPDU
2	0x02	RST BPDU
3	0x02	MST BPDU
0	0x80	TCN BPDU

MSTP 报文格式

MST BPDU 报文结构如图 5-27 所示。

图 5-27 MST BPDU 报文结构



无论是域内的 MST BPDU 还是域间的，前 36 个字节和 RST BPDU 相同。

从第 37 个字节开始是 MSTP 专有字段。最后的 MSTI 配置信息字段由若干 MSTI 配置信息组连缀而成。

MST BPDU 中的主要信息如**表 5-15** 所示。

表 5-15 MST BPDU 中主要信息说明

字段内容	字节	说明
Protocol Identifier	2	协议标识符。
Protocol Version Identifier	1	协议版本标识符，STP 为 0，RSTP 为 2，MSTP 为 3。

字段内容	字节	说明
BPDU Type	1	BPDU 类型，MSTP 为 0x02。 ● 0x00：STP 的 Configuration BPDU ● 0x80：STP 的 TCN BPDU（Topology Change Notification BPDU） ● 0x02：RST BPDU（Rapid Spanning-Tree BPDU）或者 MST BPDU（Multiple Spanning-Tree BPDU）
CIST Flags	1	CIST 标志字段。
CIST Root Identifier	8	CIST 的总根交换设备 ID。
CIST External Path Cost	4	CIST 外部路径开销指从本交换设备所属的 MST 域到 CIST 根交换设备所属的 MST 域的累计路径开销。CIST 外部路径开销根据链路带宽计算。
CIST Regional Root Identifier	8	CIST 的域根交换设备 ID，即 IST Master 的 ID。如果总根在这个域内，那么域根交换设备 ID 就是总根交换设备 ID。
CIST Port Identifier	2	本端口在 IST 中的指定端口 ID。
Message Age	2	BPDU 报文的生存期。
Max Age	2	BPDU 报文的最大生存期，超时则认为到根交换设备的链路故障。
Hello Time	2	Hello 定时器，缺省为 2 秒。
Forward Delay	2	Forward Delay 定时器，缺省为 15 秒。
Version 1 Length	1	Version1 BPDU 的长度，值固定为 0。
Version 3 Length	2	Version3 BPDU 的长度。
MST Configuration Identifier	51	MST 配置标识，表示 MST 域的标签信息，包含 4 个字段，如图 5-28 所示。只有 MST Configuration Identifier 中的四个字段完全相同的，并且互联的交换设备，才属于同一个域。字段说明如表 5-16 所示。
CIST Internal Root Path Cost	4	CIST 内部路径开销指从本端口到 IST Master 交换设备的累计路径开销。CIST 内部路径开销根据链路带宽计算。
CIST Bridge Identifier	8	CIST 的指定交换设备 ID。
CIST Remaining Hops	1	BPDU 报文在 CIST 中的剩余跳数。

字段内容	字节	说明
MSTI Configuration Messages(may be absent)	16	MSTI 配置信息。每个 MSTI 的配置信息占 16 bytes，如果有 n 个 MSTI 就占用 n×16bytes。单个 MSTI Configuration Messages 的结构如图 5-29 所示，字段说明如表 5-16 所示。

图 5-28 MST Configuration Identifier

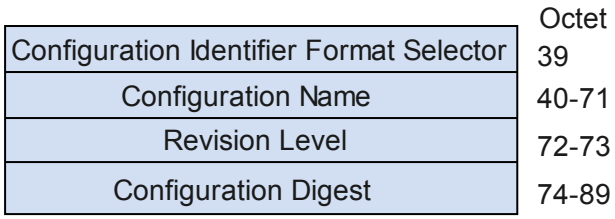


表 5-16 MST Configuration Identifier 字段说明

字段内容	字节	说明
Configuration Identifier Format Selector	1	固定为 0。
Configuration Name	32	“域名”，32 字节长字符串。
Revision Level	2	2 字节非负整数。
Configuration Digest	16	利用 HMAC-MD5 算法将域中 VLAN 和实例的映射关系加密成 16 字节的摘要。

图 5-29 MSTI Configuration Messages

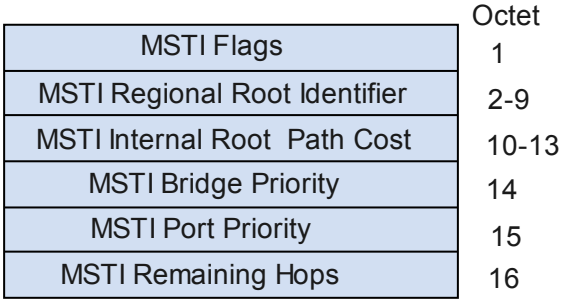


表 5-17 MSTI Configuration Messages 字段说明

字段内容	字节	说明
MSTI Flags	1	MSTI 标志。
MSTI Regional Root Identifier	8	MSTI 域根交换设备 ID。
MSTI Internal Root Path Cost	4	MSTI 内部路径开销指从本端口到 MSTI 域根交换设备的累计路径开销。MSTI 内部路径开销根据链路带宽计算。
MSTI Bridge Priority	1	本交换设备在 MSTI 中的优先级。
MSTI Port Priority	1	本端口在 MSTI 中的优先级。
MSTI Remaining Hops	1	BPDU 报文在 MSTI 中的剩余跳数。

MSTP 报文格式可配置

目前 MSTP 的 BPDU 报文存在两种格式：

- dot1s：IEEE802.1s 规定的报文格式。
- legacy：私有协议报文格式。

如果端口收发报文格式为默认支持 dot1s 或者 legacy，这样就存在一个缺点：需要人工识别对端的 BPDU 报文格式，然后手工配置命令来决定支持哪种格式。人工识别报文格式比较困难，且一旦配置错误，就有可能导致 MSTP 计算错误，出现环路。

华为技术有限公司采用的端口收发 MSTP 报文格式可配置（stp compliance）功能，能够实现对 BPDU 报文格式的自适应：

- auto
- dot1s
- legacy

这样报文收发不但支持 dot1s 和 legacy 格式，还能通过 auto 方式根据收到的 BPDU 报文格式自动切换接口支持的 BPDU 报文格式，使报文格式与对端匹配。在自适应的情况下，接口初始支持 dot1s 格式，收到报文后，格式则和收到的报文格式保持一致。

每个 Hello Time 时间内端口最多能发送 BPDU 的报文数可配置

Hello Time 用于生成树协议定时发送配置消息维护生成树的稳定。如果交换设备在一段时间内没有收到 BPDU 报文，则会由于消息超时而对生成树进行重新计算。

当交换设备成为根交换设备时，该交换设备会按照该设置值为时间间隔发送 BPDU 报文。非根交换设备采用根交换设备所设置的 Hello Time 时间值。

华为技术有限公司数据通信设备提供的每个 Hello Time 时间内端口最多能够发送的 BPDU 报文个数可配置（Max Transmitted BPDU Number in Hello Time is Configurable）功能，可以设定当前端口在 Hello Time 时间内配置 BPDU 的最大发送数目。

用户配置的数值越大，表示每 Hello Time 时间内发送的报文数越多。适当的设置该值可以限制端口每 Hello Time 时间内能发送的 BPDU 数目，防止在网络拓扑动荡时，BPDU 占用过多的带宽资源。

5.5.4 MSTP 拓扑计算

MSTP 的基本原理

MSTP 将整个二层网络划分为多个 MST 域，各个域之间通过计算生成 CST。域内则通过计算生成多棵生成树，每棵生成树都被称为是一个多生成树实例。其中实例 0 被称为 IST，其他的多生成树实例为 MSTI。MSTP 同 STP 一样，使用配置消息进行生成树的计算，只是配置消息中携带的是设备上 MSTP 的配置信息。

优先级向量

MSTI 和 CIST 都是根据优先级向量来计算的，这些优先级向量信息都包含在 MST BPDU 中。各交换设备互相交换 MST BPDU 来生成 MSTI 和 CIST。

- 优先级向量简介
  - 参与 CIST 计算的优先级向量为：

    { 根交换设备 ID，外部路径开销，域根 ID，内部路径开销，指定交换设备 ID，指定端口 ID，接收端口 ID }
  - 参与 MSTI 计算的优先级向量为：

    { 域根 ID，内部路径开销，指定交换设备 ID，指定端口 ID，接收端口 ID }

  括号中的向量的优先级从左到右依次递减。

  表 5-18 对每个优先级向量进行解释。

表 5-18 向量说明

向量名	说明
根交换设备 ID	根交换设备 ID 用于选择 CIST 中的根交换设备。根交换设备 ID = Priority(16bits) + MAC(48bits)。
外部路径开销（ERPC）	从 CIST 的域根到达总根的路径开销。MST 域内所有交换设备上保存的外部路径开销相同。若 CIST 根交换设备在域中，则域内所有交换设备上保存的外部路径开销为 0。
域根 ID	域根 ID 用于选择 MSTI 中的域根。域根 ID = Priority(16bits) + MAC(48bits)。
内部路径开销（IRPC）	本桥到达域根的路径开销。域边缘端口保存的内部路径开销大于非域边缘端口保存的内部路径开销。
指定交换设备	CIST 或 MSTI 实例的指定交换设备是本桥通往域根的最邻近的上游桥。如果本桥就是总根或域根，则指定交换设备为自己。
指定端口	指定交换设备上同本设备上根端口相连的端口。Port ID = Priority(4 位) + 端口号（12 位）。端口优先级必须是 16 的整数倍。

向量名	说明
接收端口	接收到 BPDU 报文的端口。Port ID = Priority(4 位) + 端口号（12 位）。端口优先级必须是 16 的整数倍。

- 比较原则

  同一向量比较，值最小的向量具有最高优先级。

  优先级向量比较原则如下。

  1. 首先，比较根交换设备 ID。
  2. 如果根交换设备 ID 相同，再比较外部路径开销。
  3. 如果外部路径开销相同，再比较域根 ID。
  4. 如果域根 ID 仍然相同，再比较内部路径开销。
  5. 如果内部路径仍然相同，再比较指定交换设备 ID。
  6. 如果指定交换设备 ID 仍然相同，再比较指定端口 ID。
  7. 如果指定端口 ID 还相同，再比较接收端口 ID。

  如果端口接收到的 BPDU 内包含的配置消息优于端口上保存的配置消息，则端口上原来保存的配置消息被新收到的配置消息替代。端口同时更新交换设备保存的全局配置消息。反之，新收到的 BPDU 被丢弃。

CIST 的计算

经过比较配置消息后，在整个网络中选择一个优先级最高的交换设备作为 CIST 的树根。在每个 MST 域内 MSTP 通过计算生成 IST；同时 MSTP 将每个 MST 域作为单台交换设备对待，通过计算在 MST 域间生成 CST。CST 和 IST 构成了整个交换设备网络的 CIST。

MSTI 的计算

在 MST 域内，MSTP 根据 VLAN 和生成树实例的映射关系，针对不同的 VLAN 生成不同的生成树实例。每棵生成树独立进行计算，计算过程与 STP 计算生成树的过程类似，请参见 **STP 拓扑计算**。

MSTI 的特点：

- 每个 MSTI 独立计算自己的生成树，互不干扰。
- 每个 MSTI 的生成树计算方法与 STP 基本相同。
- 每个 MSTI 的生成树可以有不同的根，不同的拓扑。
- 每个 MSTI 在自己的生成树内发送 BPDU。
- 每个 MSTI 的拓扑通过命令配置决定。
- 每个端口在不同 MSTI 上的生成树参数可以不同。
- 每个端口在不同 MSTI 上的角色、状态可以不同。

在运行 MSTP 协议的网络中，一个 VLAN 报文将沿着如下路径进行转发：

- 在 MST 域内，沿着其对应的 MSTI 转发。
- 在 MST 域间，沿着 CST 转发。

MSTP 对拓扑变化的处理

MSTP 拓扑变化处理与 RSTP 拓扑变化处理过程类似，请参见 **RSTP 技术细节**中的 RSTP 拓扑变化处理。

5.5.5 MSTP 快速收敛机制

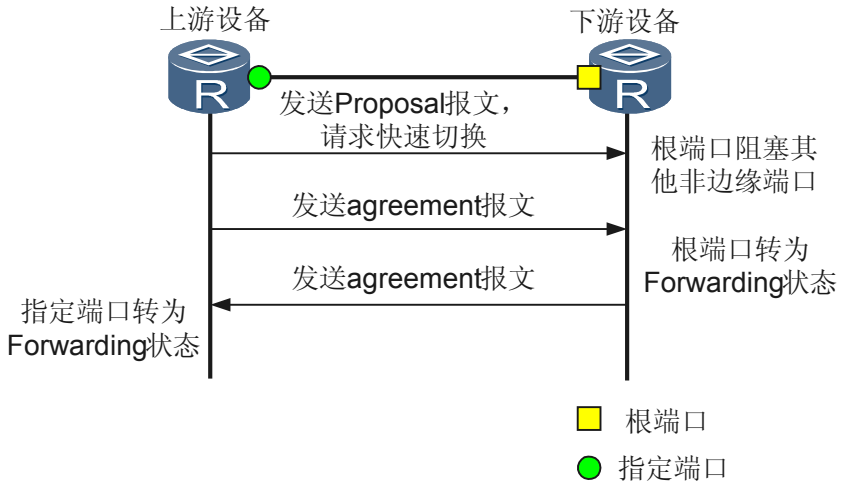
MSTP 支持普通方式和增强方式两种 P/A（Proposal/Agreement）机制：

- 普通方式

  MSTP 支持普通方式的 P/A 机制实现与 RSTP 支持的 P/A 机制实现相同，RSTP 支持的 P/A 机制请见 **RSTP 技术细节**中的 P/A 机制。

- 增强方式

图 5-30 增强方式的 P/A 机制



如**图 5-30** 所示，在 MSTP 中，P/A 机制工作过程如下：

1. 上游设备发送 Proposal 报文，请求进行快速迁移。下游设备接收到后，把与上游设备相连的端口设置为根端口，并阻塞所有非边缘端口。
2. 上游设备继续发送 Agreement 报文。下游设备接收到后，根端口转为 Forwarding 状态。
3. 下游设备回应 Agreement 报文。上游设备接收到后，把与下游设备相连的端口设置为指定端口，指定端口进入 Forwarding 状态。

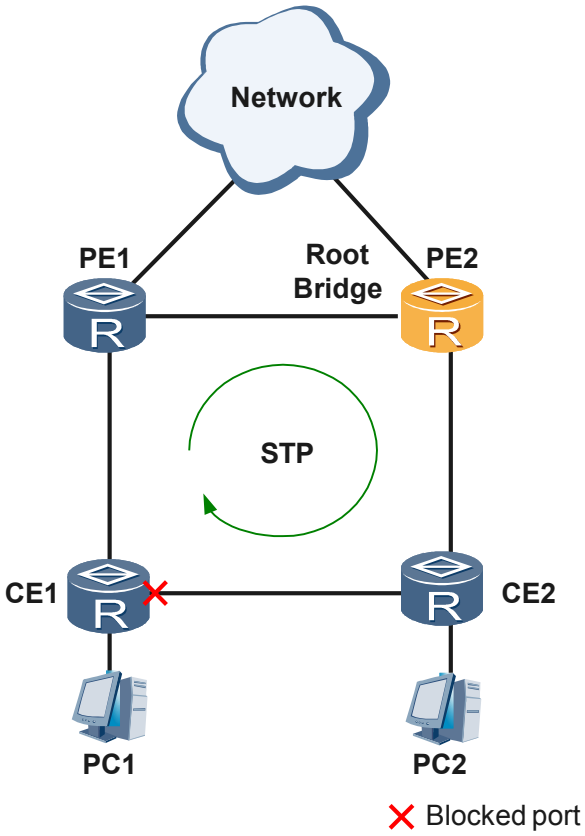
缺省情况下，华为数据通信设备使用增强的快速迁移机制。如果华为数据通信设备和其他制造商的设备进行互通，而其他制造商的设备 P/A 机制使用普通的快速迁移机制，此时，可在华为数据通信设备上通过命令 **stp no-agreement-check** 设置 P/A 机制为普通的快速迁移机制，从而实现华为数据通信设备和其他制造商的设备进行互通。

5.6 应用

STP 典型应用

在一个复杂的网络中，网络规划者由于冗余备份的需要，一般都倾向于在设备之间部署多条物理链路，其中一条作主用链路，其他链路作备份。这样就难免会形成环形网络，若网络中存在环路，可能会引起广播风暴和 MAC 桥表项被破坏。

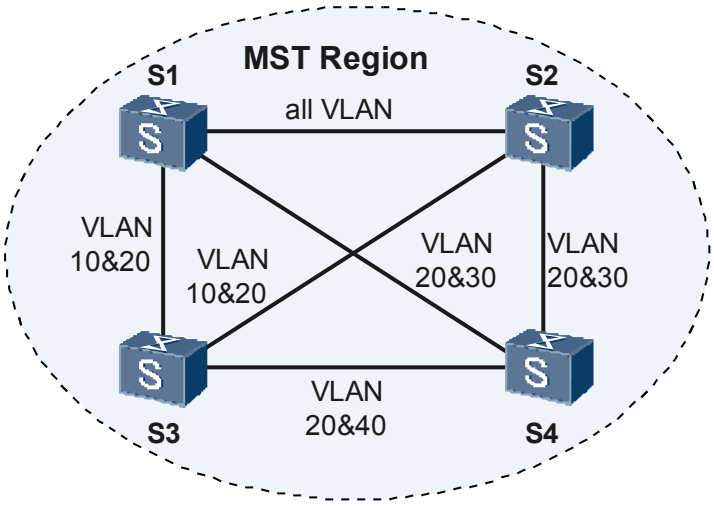
图 5-31 STP 典型应用组网图



如图 5-31 所示，在网络中部署 STP 协议，通过彼此交互信息发现网络中的环路，并有选择的对某个端口进行阻塞，最终将环形网络结构修剪成无环路的树形网络结构，从而防止报文在环形网络中不断增生和无限循环，避免设备由于重复接收相同的报文造成处理能力下降。

MSTP 典型应用

图 5-32 MSTP 典型应用组网图



配置 MSTP 使图 5-32 中不同 VLAN 的报文按照不同的生成树实例转发。具体配置为：

- 网络中所有交换机属于同一个 MST 域；
- VLAN 10 的报文沿着实例 1 转发，VLAN 30 沿着实例 3 转发，VLAN 40 沿着实例 4 转发，VLAN 20 沿着实例 0 转发。

图 5-32 中 S1 和 S2 为汇聚层设备，S3 和 S4 为接入层设备。VLAN 10、VLAN30 在汇聚层设备终结，VLAN 40 在接入层设备终结，因此可以配置实例 1 和实例 3 的树根分别为 S1 和 S2，实例 4 的树根为 S3。

5.7 术语与缩略语

术语

术语	解释
STP	Spanning Tree Protocol——生成树协议，是一个用于在局域网中消除环路的协议。运行该协议的设备通过彼此交互信息而发现网络中的环路，并适当对某些接口进行阻塞以消除环路。
RSTP	Rapid Spanning Tree Protocol——快速生成树协议，该协议规范在 IEEE802.1w 中有详细描述。RSTP 基于 STP 协议，但对原有协议有更加细致的修改和补充，相对 STP 能够快速收敛。
MSTP	Multi-Spanning Tree Protocol——多生成树协议，是 IEEE802.1s 中定义的一种新型生成树；使用域(region)和实例(instance)的概念，在一个大的网络，按照不同的需求划分不同的域，域里面建实例，实例与 VLAN 进行映射；网桥之间通过传输带有域和实例信息的 BPDU，网桥通过 BPDU 信息判断自己是否属于某个域；域内运行多实例化的 RSTP，域间运行 RSTP 兼容的协议。

术语	解释
VLAN	Virtual Local Area——虚拟局域网，是指在交换局域网的基础上，采用网络管理软件构建的可跨越不同网段、不同网络的端到端的逻辑网络。一个 VLAN 组成一个逻辑子网，即一个逻辑广播域，可以覆盖多个网络设备。

缩略语

缩略语	英文全名	中文解释
STP	Spanning Tree Protocol	生成树协议
RSTP	Rapid Spanning Tree Protocol	快速生成树协议
MSTP	Multiple Spanning Tree Protocol	多生成树协议
BPDU	Bridge Protocol Data Unit	桥协议数据单元
CIST	Common and Internal Spanning Tree	公共和内部生成树
CST	Common Spanning Tree	公共生成树
IST	Internal Spanning Tree	内部生成树
SST	Single Spanning Tree	单生成树
MST	Multiple Spanning Tree	多生成树
MSTI	Multiple Spanning Tree Instance	多生成树实例
TCN	Topology Change Notification	拓扑变化通告
VLAN	Virtual Local Area Network	虚拟以太网