三康技术有限公司 Huawei-3Com Technologies Co., Ltd.	文档编号 Document ID	密级 Confidentiality level
		内部公开
	文档状态 Document Status	共26页 Total 26 pages
	Draft 1.00	

STP/RSTP基础

拟制 Prepared by	边江02193	Date 日期	2005-02-15
评审人 Reviewed by		Date 日期	
批准 Approved by		Date 日期	

目录 Table of Contents

图目录 Table of Pic

表目录 Table of Forms

STP/RSTP基础

1 STP 的由来与基本概念

生成树协议（Spanning-Tree Protocol，以下简称 STP）是一个用于在局域网中消除环路的协议。运行该协议的交换机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，STP 已经成为了当前最重要的局域网协议之一。

1.1 现代交换网络的环路问题

提及生成树协议的由来，我们先来看看没有生成树的网络。

如图 1 展示了现在普遍采用的多交换机来实现冗余的局域网结构。

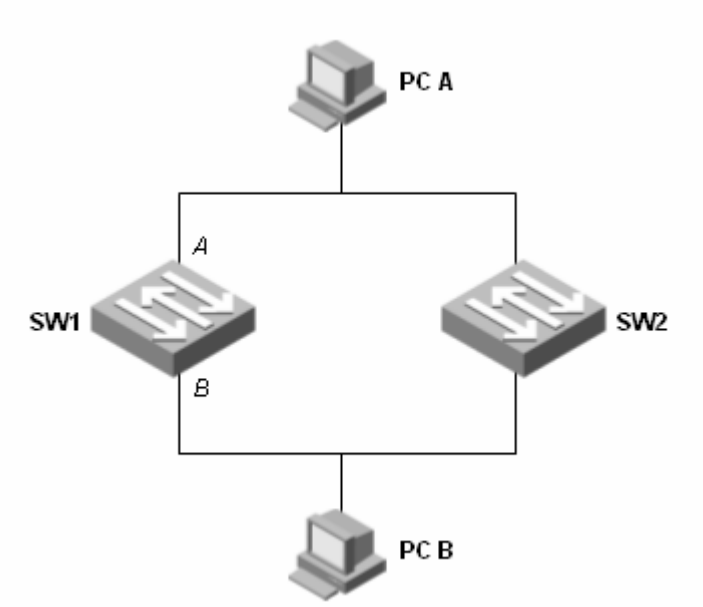

图1 冗余的网络结构

在图 1 所示的网络中，可能产生如下的两种情况：

- 广播环路（Broadcast Loop）

显然，当 PC A 发出一个 DMAC 为广播地址的数据帧时，由交换机对于广播的处理方式可知，该广播会被无休止的转发。

- MAC 地址表损坏（Bridge Table Corruption）

在图1中，即使是单播，也有可能导致异常。交换机SW1可以在端口B上学习到PC B的MAC地址，但是由于SW2会将PC B发出的数据帧向自己其它的端口转发，所以SW1也可能在端口A上学习到PC B的MAC地址。如此SW1会不停的修改自己的MAC地址表。这样就引起了MAC地址表的抖动（Flapping）。

1.2 STP：环路终结者

1.2.1 基本思想

STP可以在保持物理连接的情况下有效的环路消除网络中的环路。其基本理论依据是根据网络拓扑构建（生成）**无回路的连通图**（就是树），从而保证数据传输路径的唯一性，避免出现环路报文流量增生和循环。STP是工作在OSI第二层（Data Link Layer）的协议。

STP协议通过在网桥之间传递特殊的消息并进行分布式的计算，来决定一个有环路的网络中，哪台交换机的哪个端口应该被阻塞（Blocking），用这种方法来剪切掉环路。IEEE std 802.1D协议文档的第8章描述了STP。

1.2.2 一个根桥

树形的网络结构，必须要有根，于是 STP 引入了根桥（Root Bridge）的概念。

对于一个 STP 网络，根桥有且只有一个。它是整个网络的逻辑中心，但不一定是物理中心。但是根据网络拓扑的变化，根桥可能改变。而且一旦网络收敛之后，只有根桥按照一定的时间间隔产生并且向外发送一种称为“ 配置消息 ”的协议报文，其他的交换机仅对该种报文进行“ 接力 ”，这样来保证拓扑的稳定。

1.2.3 二种度量

生成树的生成计算有两大基本度量依据：**ID**和**路径开销**。

ID 又分为两种：BID 和 PID。

BID 即 Bridge ID，或称为桥 ID。IEEE 802.1D 标准对这个值的规定是由 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 最小的交换机会被选举为根桥。在我司设备上，桥优先级可以手工配置 0 – 65535（默认为 32768）。

PID 即 Port ID，或称为端口 ID。也是由两部分构成的，高 8 位是端口优先级，低位是端口号。PID 只在某些情况下对选择指定端口有作用。在我司设备上，端口优先级可以手工配置 0 – 255（默认为 128）。

路径开销（Path Cost）是一个端口量，反映了本端口所连接网络的开销。该值越低，表示这个端口连接的网络越好。在一个STP网络中，某端口到根桥累计的路径开销就是通过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）。根路径开销的作用将在后文介绍。

IEEE 802.1t中规定的路径开销如表1.1所示，而各厂商采用的路径开销标准不尽相同。

表1 IEEE802.1t 路径开销列表

带宽	STP 路径开销
4 Mbps	250
10 Mbps	100
16 Mbps	62
45 Mbps	39
100 Mbps	19
155 Mbps	14
622 Mbps	6
1 Gbps	4
10 Gbps	2

1.2.4 三要素选举

从一个初始的有环拓扑生成树状拓扑，总体来说有三个要素：根桥、根端口和指定端口。

根桥是全网意义上的。通过交换特殊的协议报文，网络中很快就会对最小的 BID 达成一致。

所谓根端口，是指一个非根桥的 STP 交换机上离根桥最近的端口，这个端口的选择标准就是上面提过的根路径开销。在本网桥上所有使能 STP 的端口中，根路径开销最小者，就是根端口。很显然，在一个 STP 交换机上根端口有且只有一个。

在每一个 STP 交换机上，端口都有三种角色：根端口、指定端口、非跟非指定端口（根桥上没有根端口）。指定端口的概念是针对于某网段的，是流量从根桥方向来而从这个端口

转发“出去”。从一个连接到 STP 交换机的网段来说，该网段通过指定端口接收到根桥方向过来数据。根桥上的所有端口都是指定端口。在每一个网段上，指定端口有且只有一个。

在拓扑稳定状态，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。

一旦根桥、根端口、指定端口选举成功，则整个树形拓扑就建立完毕了。图 2 展示了 STP 的树形拓扑。

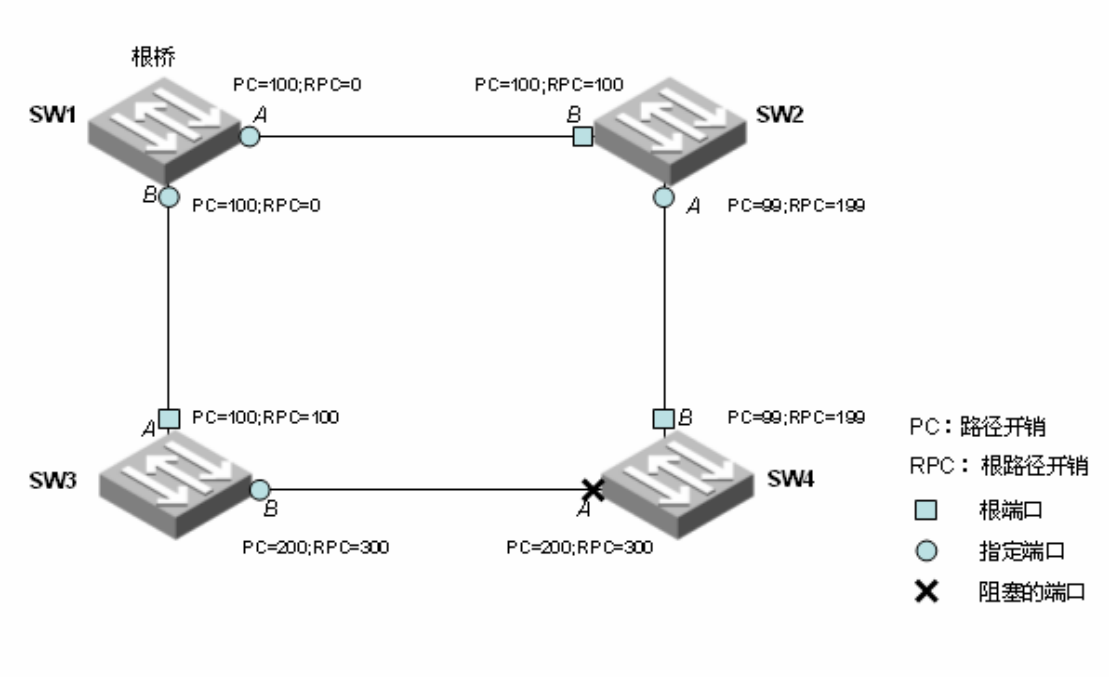

图2 STP 网络的树形结构

1.2.5 四个比较原则

STP交换机协议采用特殊的协议报文（又称协议数据单元，Bridge Protocol Data Unit）来交互信息，这种特殊的消息称为“配置消息（Configuration Message）”或者一般简称之BPDU。

BPDU分为两类：配置BPDU(配置BPDU)和TCN BPDU (Topology Change Notification BPDU)。配置BPDU用来生成树拓扑的配置BPDU，维护网络拓扑；TCN BPDU则只在拓扑发生变化的时候发出，用来通知相关的交换机网络发生变化。

配置BPDU主要携带如表2所示的几个重要信息。

表2 四个重要信息字段

字段内容	简要说明
根桥 BID	每个 STP 网络中有且仅有一个根
累计根路径开销	发送这个 BPDU 的端口到根桥的“ 距离 ”
发送交换机 BID	发送这个 BPDU 的交换机的 BID
发送端口 PID	发出这个 BPDU 的端口的 PID

STP交换机接收配置BPDU，并处理上述字段。

这里我们小结一下生成一棵树的四个基本的比较原则：

最低BID。用来选根桥。STP交换机之间根据表2所示根桥BID字段选择最低的BID。

最小的累计根路径开销。用来在非根桥上选择根端口，在一个。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STP交换机要在两个以上根路径开销相等的端口之中选择出根端口的时候，会选择接收到的配置消息中发送者BID较小的那个端口。如图2，假设SW2的BID小于SW3的BID，如果在SW4的A、B两个接收到的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。只有在图3所示的情况下，PID才起到了作用。SW1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、发送交换机BID都相同，所以消除环路的依据就剩下PID了。

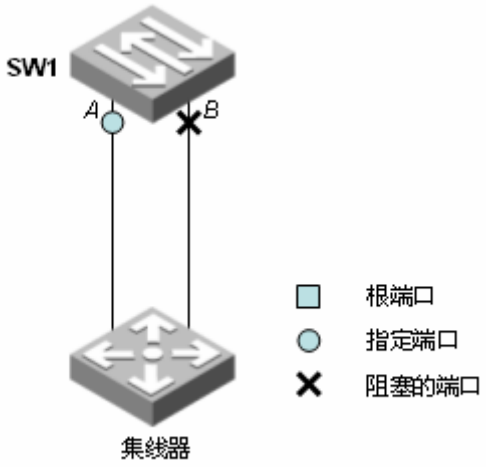

图3 应用到 PID 进行比较的拓扑

在STP的所有的比较量中，都遵循“ 数值越低就越好 ”的原则，如BID、PID和路径开销等等都是这样的。

1.2.6 五种端口状态

STP端口有5种状态，如表3所示。

表3 STP 端口五种状态

状态	说明
Forwarding	在这种状态下，端口转发用户流量的状态，只有根端口或指定端口才有这种状态。
Learning	这是一种过渡状态。在这种状态下，交换机会根据收到的用户流量（但仍然不转发流量）构建 MAC 地址表，所以叫做学习“ 状态 ”。
Listening	这是一种过渡状态。在这种状态下，上述的三步选择（根桥、根端口、指定端口）就是在该状态内完成。
Blocking	在这种状态下，端口仅仅接收并处理 BPDU，不转发用户流量。
Disabled	或 Down，认为阻断或物理上断掉。

端口处于Listening和Learning状态的时间是由的Forward Delay Timer来统一控制的，这两个时间总是一样长的。

这5种状态在相应条件下的相互转化。

2 STP 技术细节

2.1 初始化生成树的过程

2.1.1 根桥的选择

开始网络初始化的时候，所有网络中的STP交换机都认为自己是“ 根桥 ”。此时，每个网桥都仅仅收发配置BPDU，而不转发用户流量。所有的网桥端口都处于Listening状态。此时通过交换BPDU，进行选举的工作：即选根桥，选根端口，选指定端口。

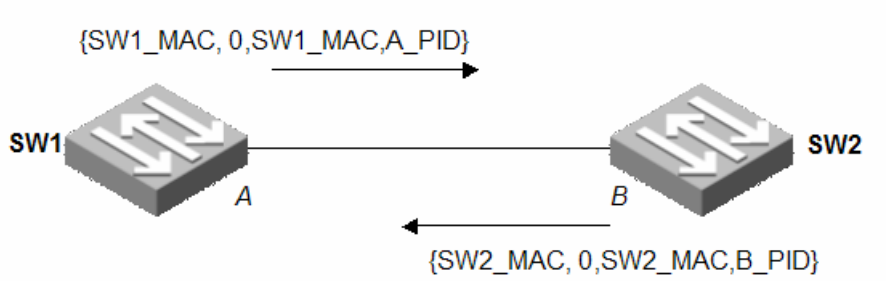

图4 初始信息交互

如图4，图中用“ {} ”标注的四元组表示了由根桥BID、累计根路径开销、发送者BID、发送端口PID构成的有序组。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。BPDU会按照Hello Time指定的时间间隔来发送，默认的时间为2秒。

一旦在某端口上收听到比自己发的还要“ 好 ”的BPDU，那么这个端口就提取该BPDU中的某些信息，更新自己的信息。再次强调一下，比较BPDU的“ 好坏 ”的方式（自己的或别的网桥的的），都是根据上面提到的四元组来完成的，即最低桥ID、最低到累计根路径开销，最低发送者BID（有时还需要最低端口PID，见前文）。该端口缓存他人BPDU后，自己则立即停止发送BPDU。

当发送BPDU的时候，网桥填充Sender BID字段的总是自己的BID，而填充Root BID字段的是“ 当前我所认为是根桥的 ”BID。如图4所示，SW2的端口B由于接收到了更好的BPDU，从而认为此时SW1是根桥，然后SW2的其他端口再发送BPDU的时候，在根桥字段里面填充的就是SW1_MAC了。此过程不断交互进行，直到所有网桥的所有端口都认为根桥是相同的，就说明根桥已经选择完毕了。

2.1.2 根端口的选择

每个非根桥STP交换机都要选择一个根端口，根端口对于一个交换机来说有且只有一个。其本质是“ 距离根桥最近的端口 ”，这个最近的衡量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应该是根端口。累计根路径开销的计算方法如下：端口收到一个BPDU后，抽取该BPDU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最小的那个段口作为根端口。

2.1.3 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网桥的）发送BPDU的端口，就是该网段的指定端口。如图4，假定SW1的MAC地址小于SW2的MAC地址，则SW1的端口A会成为指定端口。根桥的所有端口都是指定端口（根桥物理上环路这种情况除外）。在收敛后，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，即收听BPDU而不转发用户流量。

在一个网段上拥有指定端口的交换机被称作该网段的指定桥。如图4中，SW1-SW2间网段的指定桥为SW1。

2.1.4 拓扑稳定之后

在拓扑稳定之后，根桥仍然按照Hello Time间隔发出配置BPDU。其他指定桥收到上一级接力过来的BPDU，如果有必要则根据BPDU里面的信息更新自己相应端口的。

2.2 STP 协议报文

本节我们来详细的看一下STP的协议报文，如表4所示。

总结上文，可知配置BPDU在是一种“ 心跳 ”报文，只要端口上使能了STP，则配置BPDU就会按照Hello Timer所规定的时间间隔发出。在初始化过程中，每个桥都主动发出配置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，而且如果当且仅当发送者BID或发送端口PID两个字段中至少有一个和本桥本接收端口不同，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。

表4 配置 BPDU 基本格式

域	字节	说明
协议号	2	总是 0
版本	1	总是 0
类型	1	当前 BPDU 的类型 0 = 配置 BPDU，0x80= TCN BPDU
标志	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
根桥 BID	8	当前根桥的 BID
根路径开销	4	本端口累计到根桥的开销
发送者 BID	8	本交换机的 BID
发送端口 PID	2	发送该 BPDU 的端口 ID
Message Age	2	该 BPDU 的消息年龄
Max Age	2	消息老化年龄
Hello Time	2	发送两个相邻 BPDU 间的时间间隔
Forward Delay	2	控制 Listening 和 Learning 状态的持续时间

表5 Flag 字段格式

Bit7	Bit6	Bit5	Bit4	Bit3	Bit2	Bit1	Bit0
TCA	保留未使用						TC

TCN BPDU内容比较简单，只有表4种列出的前4个字段。且类型字段为0x80，标志字段的最低位置为1。

TCA位只有在回应TCN BPDU的配置BPDU中置1，后面会详细进行说明。

2.3 端口的状态迁移

端口状态变化条件

(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

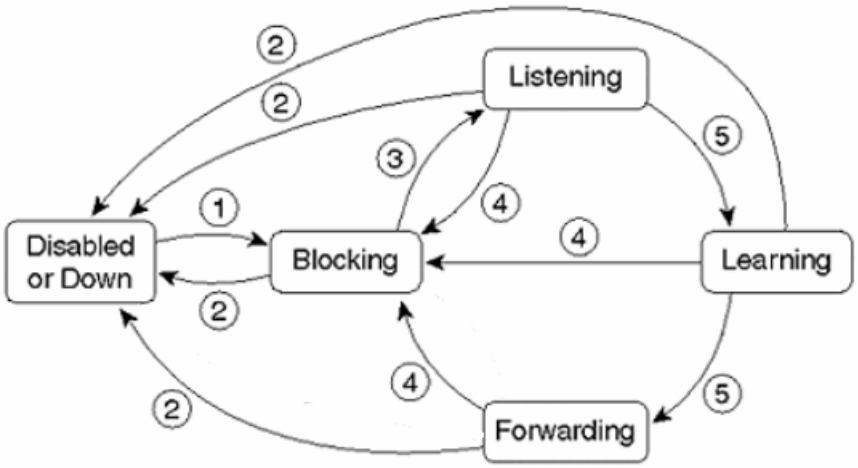

图5 STP 端口状态转换图示

2.4 STP 拓扑变化

当拓扑发生变化后，STP是否产生TCN BPDU要根据以下两条标准来判定：

1 网桥至少有一个指定端口，并且某端口从其他（Blocking、Listening、Learning）状态转到Forwarding状态。

2 某端口由Forwarding、Learning状态转到Blocking状态。

当以上两条基本条件至少满足其一时，STP交换机就会发出TCN BPDU。

2.4.1 3 个计时器

对于STP，一共有3个计时器影响着端口状态以及网络的收敛，如下所述：

1 Hello Timer：STP交换机发送BPDU的时间间隔。当网络拓扑稳定之后，该计时器的修改只有在根桥修改才有效。根桥会在之后发出的BPDU中填充适当的字段以向其他非根桥传递该计时器修改信息。但当拓扑变化之后，TCN BPDU的发送不受这个计时器的管理。

2 Forwarding Timer：指一个端口Listening 和Learning的各自时间，默认为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，这是STP用于避免临时环路的关键。

3 Max Age：端口的BPDU老化的时间，前文已经探讨过。端口会根据接收到的BPDU存储所接收到的最好的四个信息（根桥BID、累计根路径开销、发送者BID和发送端口PID）。每次接收到合适的BPDU，端口都会启动这个Max Age计时器。超过这个Max Age时间端口接收不到合适BPDU，就会采取相应的措施。这个时间默认为20秒。

我们在以下的几小节中将列举几种STP拓扑变化的情况来探讨。

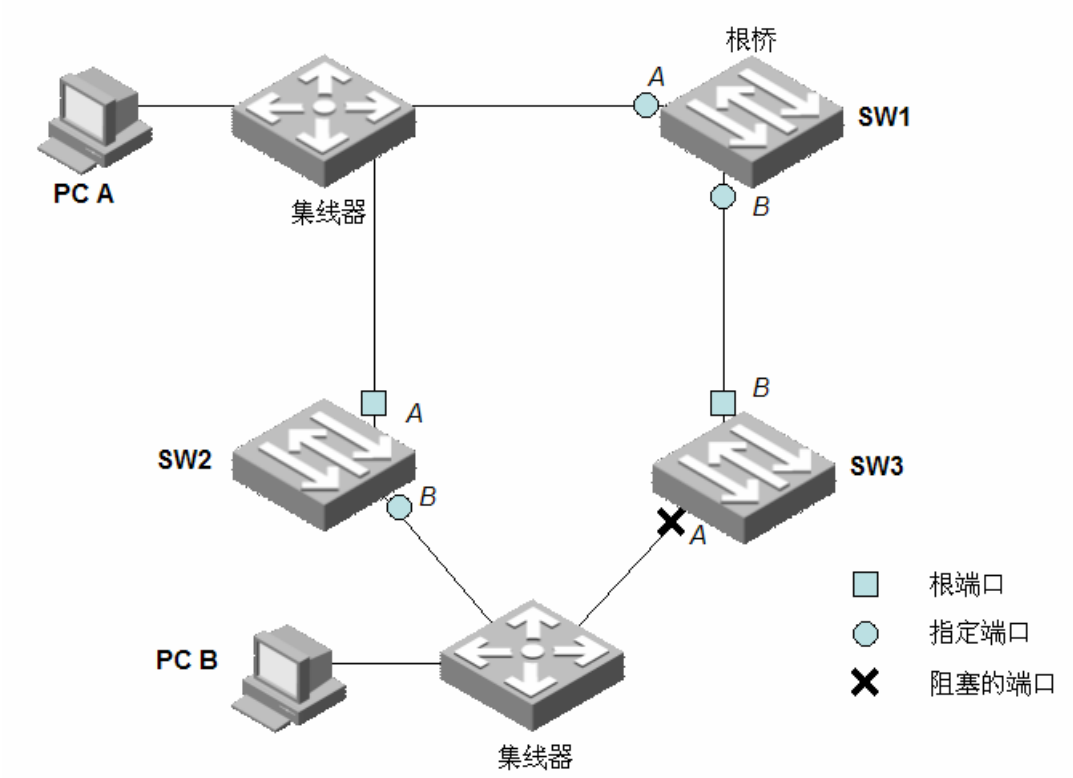

图6 拓扑变化图示

2.4.2 链路断掉与 TC

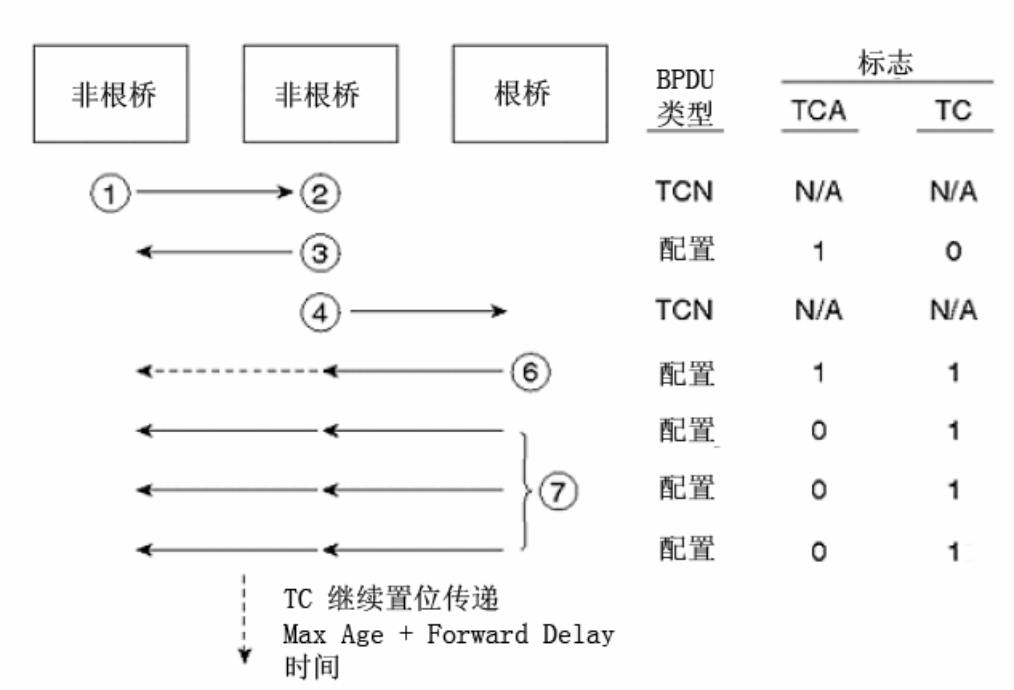

图7 TC 发生后协议报文的传递

如果非根桥交换机的指定端口断掉，则交换机会立即通过根端口以一定速率发送TCN BPDU给上游交换机，上游交换机收到该TCN BPDU后会立刻向下游交换机回应TCA BPDU（即把标志字段的TCA位和TC位都置位的配置BPDU）并继续向上游传递TCN BPDU。下游交换机收到TCA消息的交换机会停止发送TCN BPDU。如此不断的传递，直到传到根桥。

根桥会持续Forward Delay+Max Age长的时间内在发出的配置BPDU中把TC位置位。其下游任何交换机接收到根桥传递过来的TC置位的配置BPDU都会将自己的MAC地址表老化时间修改为Forward Delay那么长时间。该过程如图7描述。

以图6为例，稳定后的拓扑如图所示。如果如果SW2的B端口连接网络断掉，则SW2会从端口A发送TCN BPDU。根桥SW1收到后立刻发回TCA消息，SW2马上将自己的MAC地址标的老化时间修改为Forward Delay。SW3由于也收到了TC置位的配置BPDU，MAC地址表也会在Forward Delay时间后老化。同时，SW3的端口A会开始等待Max Age + Listening + Learning后进入Forwarding状态。

2.4.3 端口收到次等 BPDU（Inferior BPDU）

对于非指定端口（根端口或者Blocking端口），在收到比自己当前端口信息更优或者一

样好的BPDU，则会更新Max Age计时器。如果超过这个时间仍然收不到满足条件的BPDU，端口便会迁移到Listening状态，准备重新进行选择。如图6，如果SW2的A端口连接网络端掉，则其B端口会发出认为自己为根桥的BPDU，然而SW3端口A由于不是指定端口，所以会等待Max Age时间，则会继续再经历Listening+Learning时间后转入指定端口进入Forwarding，整个转换时间默认要20+15+15 = 50秒。事实上，根端口对于这类情况的处理是可以优化的，参见第3章第3.1节“ RSTP对STP的改进 ”。

指定端口收到次等BPDU时，会立刻发出自己的更好的配置BPDU回应。

3 RSTP：一个更好的 STP

RSTP全称是Rapid Spanning-Tree Protocol，快速生成树协议。该协议规范在IEEE 802.1w中有详细的描述。该协议基于STP协议，但是对原有协议做了更加细致的修改和补充。现在RSTP已经基本取代了STP在实际的组网中广泛应用。

3.1 STP 的一些不足

STP协议虽然能够解决环路问题，但是还是有很多不足之处。

首先，STP并没有细致区分端口状态和端口角色。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。事实上，从用户角度上看Listening、Learning和Blocking状态是没有区别的，都同样不转发用户流量。从使用和配置上来讲，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。根端口和指定端口也可能处于Listening状态，也可能都处于Forwarding状态。

其次，STP算法是被动的算法，对网络是否已经达到收敛没有一种反馈机制。对待拓扑变化的基本的方法是通知根桥，修改MAC地址表老化时间，自动学习，确立新路径。这种以计时器来等待的方式显然是浪费时间，响应迟缓。

再次，STP的算法要求在稳定拓扑里，根桥主动发出BPDU而其他交换机进行中继，这样整个STP网络

3.2 RSTP 对 STP 的改进

我们这里罗列出RSTP与STP的不同，勾勒一个整体的印象，再在第四章中深入讨论RSTP技术细节。

3.2.1 端口角色的增补

根据STP的不足，RSTP新增加了端口的角色概念。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。

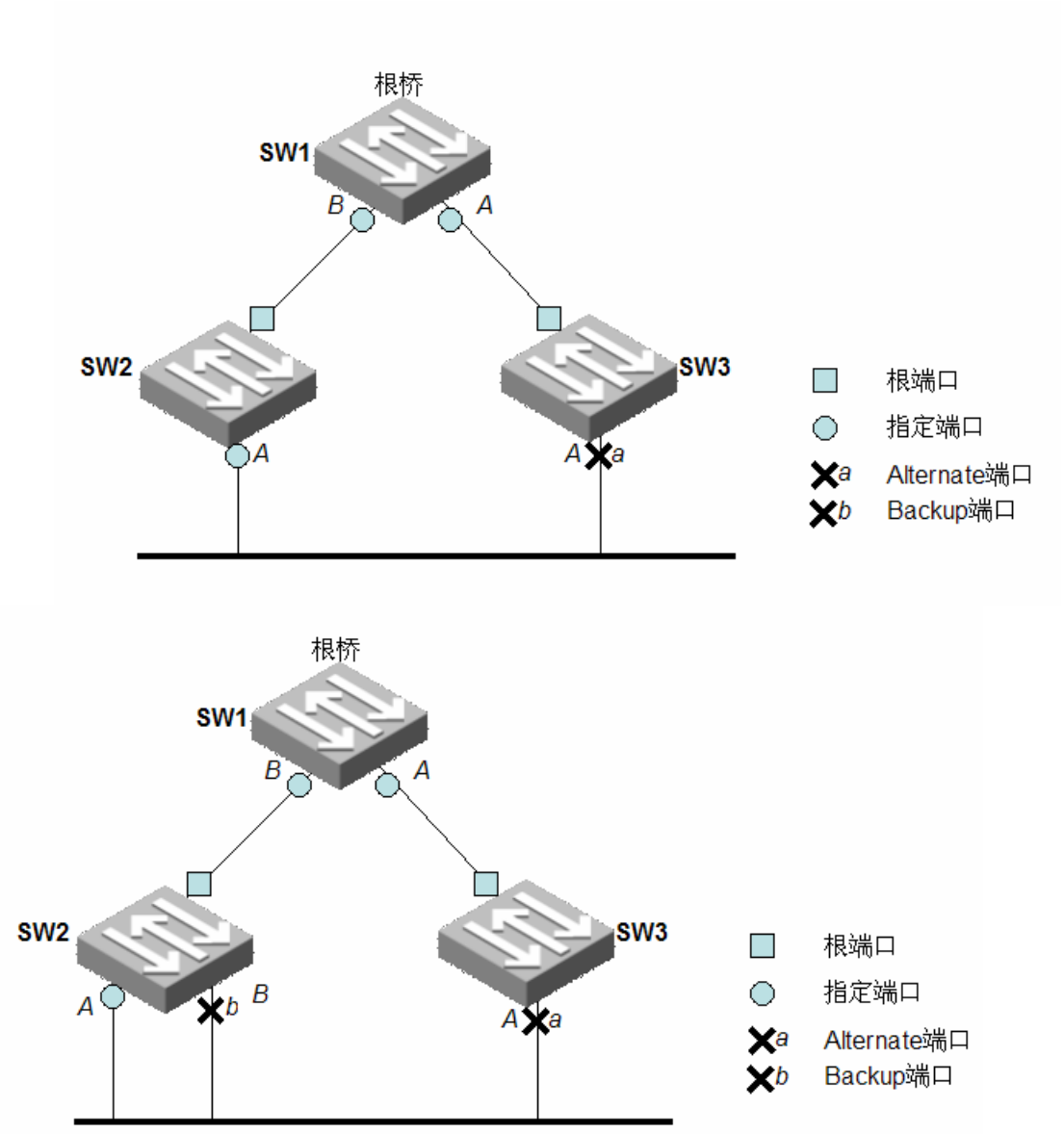

图8 新角色：Alternate 和 Backup 端口

RSTP的端口角色共有4种：即根端口、指定端口、Alternate端口和Backup端口。Alternate端口和Backup端口的形象说明如图8所示。从BPDU的发送上来看，Alternate端口就是由于学习到其它网桥的发送的BPDU而阻塞的端口；而Backup端口就是由于学习到自己发送的BPDU而阻塞的端口。从用户流量上来看，Alternate端口提供了从指定桥到根的另一条可切

换路径，作为指定端口的备选切换；而同时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可切换的通路。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。

3.2.2 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，那么就是Discarding状态；如果不转发用户流量但是学习MAC地址，那么就是Learning状态；如果既转发用户流量又学习地址，那么就是Forwarding状态。表6显示了新的状态与STP相应状态的比较。注意，我们这里由于已经将端口状态和端口角色解耦，所以状态和是什么样的端口没有必然联系。

表6 STP 与 RSTP 端口状态角色对应表

STP 端口状态	RSTP 端口状态	端口在活动拓扑中的角色
DISABLED	Discarding	不包括（Disable）
BLOCKING	Discarding	不包括（Alternate 端、Backup 端口）
LISTENING	Discarding	包括（根端口、指定端口）
LEARNING	Learning	包括（根端口、指定端口）
FORWARDING	Forwarding	包括（根端口、指定端口）

3.2.3 BPDU 格式的改变

在BPDU的格式上，除了保证和STP格式基本一致之外，RSTP作了一些小的变化。一个是在Type字段，配置BPDU类型不再是0而是2。所以运行STP的网桥收到该类BPDU时会丢弃。另一个变化是在Flag字段，把原来保留的中间6位使用起来。这样改变了的配置BPDU叫做RST BPDU。

RST BPDU的Flag字段格式如下表7所示。

表7 RSTP Flag 字段格式

端口角色 = 00 未知
01 根端口
10 Alternate / Backup
11 指定端口

Bit7	Bit6	Bit5	Bit4	Bit3	Bit2	Bit1	Bit0
TCA	Agreement	Forwarding	Learning	端口角色		Proposal	TC

3.2.4 稳态 BPDU 的发送方式

靠收到上游BPDU而触发发送BPDU的方式使得STP庞大而笨拙。RSTP对这一点进行了改进，即在稳态后，无论非根桥是否接收到根桥传来的信息，BPDU是按照每个Hello时间进行发送的，该行为完全由每个交换机自主。

3.2.5 更短的 BPDU 超时计时

如果一个端口连续三个Hello时间接受不到上游指定桥送来的BPDU，那么该交换机认为与此邻居之间的链路失败。而不像STP那样都要先等待一个Max Age。

3.2.6 处理次等 BPDU

当一个端口收到上游的指定桥发来的RST BPDU中的信息，不如自己端口信息的时候，会立即回应自己的信息。上游的指定端口马上接受这个信息，并且更新自己的信息。这样根本不用任何计时器，来通过超时解决拓扑收敛。有些类似Cisco的BackboneFast技术。参见第5章5.3 “ UplinkFast和BackboneFast ” 一节。

3.2.7 Proposal/Agreement 机制

当一个端口被选举成为了指定端口之后，在 STP 中，该端口还要等待至少一个 Forward Delay（Learning）时间才会迁移到 forwarding 状态。而在 RSTP 中，这样的端口角色会先进入 Discarding 状态。此时，可以通过 Proposal/Agreement 机制快速的进入 forward 状态。在第五部分中有详细解释。这种机制必须在点到点全双工链路上使用。

3.2.8 根端口快速切换机制

如果一个桥根端口失效，那么立刻使自己的一个“ 最好 ”的Alternate端口成为根端口置，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，可以回

溯到根的。这种产生新的根端口的过程会引发TC。下面还要讲到RSTP激发TC的条件。

3.3 我司交换机的其他实现特性

3.3.1 边缘端口（Edge Port）

在RSTP里面，如果某指定端口位于整个域的边缘即不再与其他交换机连接，这种端口叫做边缘端口。边缘端口不接收处理BPDU，不参与RSTP运算，可以由Disable直接转到Forwarding，并不经历时延，很像在端口上把STP禁用。但端口一旦收到BPDU，就丧失了边缘端口属性，成为了普通的STP端口。对于边缘端口的保护措施参见本章3.3.2“ BPDU Guard ”一节。边缘端口的概念类似于Cisco的PortFast端口，见第5章5.2 “ PortFast ”一节。

3.3.2 BPDU Gurad

交换机特性。如果在交换机上启动了BPDU Guard功能，则边缘端口一旦收到了RST BPDU，这些端口就会Shutdown，其边缘端口的属性不改变，同时通知网管发出告警；被Shutdown的端口只能由网络管理人员恢复。

3.3.3 Root Guard

端口特性，只能在指定端口上使能。对于设置了Root Guard功能的指定端口，其端口角色只能保持为指定端口。一旦这种端口上收到了优先级高（更好）的RST BPDU，端口的状态将被设置为DISCARDING状态，不再转发报文。在经过一段时间（通常为两倍的Forward Delay）的时间内，不再收到好的BPDU，端口会自动恢复正常的Forwarding状态。所以，为了保护根桥在一定范围内，可以将该范围的边界端口均启动Root Guard功能。

3.3.4 Loop Guard

端口特性，只能在根端口或者Alternate端口上使能。交换机端口上启动了Loop Guard功能后，如果根端口或者Alternate端口上发生了信息老化（即长时间收不到BPDU），则该端会进入Discarding状态，同时向网管发出通知信息。直到收到下一个RST BPDU，才恢复正常的Forwarding状态。

4 RSTP 技术细节

4.1 P/A 协商：快速收敛机制

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。其过程的完成根据以下几个端口变量：

a) **proposing.** 当一个指定端口处于Discarding或Learning状态的时候，该变量置位。向下游交换机传递Proposal位被置位的RST BPDU。

b) **proposed.** 当端口收到对端的指定端口发来的携带Proposal的RST BPDU的时候，该变量置位。该变量指示本网段上的指定端口希望尽的进入Forwarding状态。

c) **sync.** 当Proposed被设置以后，收到proposal的根端口会依次为自己的其他端口置位sync变量。如果端口是非边缘的指定端口是则会进入Discarding状态。

d) **synced.** 当端口完成转到Discarding后，会设置自己的synced变量。Alternate、Backup和边缘端口会马上设置该变量。根端口监视其他端口的synced，当所有其他端口的synced全被设置，根端口会设置自己的synced，然后传回RST BPDU，其中Agreement位被置位。

e) **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”，该变量被设置。Agreed变量一旦被置位，指定端口马上转入Forwarding状态。

举例子说明：如图9，根桥SW1和SW2之间新添加了一条链路（注：本例来源于Cisco的小文档“Understanding RSTP”，图有变化）。在当前状态下，SW2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。一旦新链路连接，p0和p1两个端口马上都先成为指定端口，发送RST BPDU。SW2的P1口收到更好的BPDU，马上把意识到自己将成为根端口，而不是指定端口，停止发送BPDU。这时SW1的p0进入Discarding状态，满足条件指定端口处于Discarding状态，于是发送的RST BPDU中把proposal置1。SW2收到根桥发送来的携带proposal的BPDU，于是开始让自己的将所有端口进入sync变量置位。P2已经是阻塞的了，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。当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SW1返回Agreement位置置位的回应BPDU。该BPDU携带和刚才根桥发过来的BPDU一样的信息，除了Agreement位置置位之外（Proposal位清零）。SW1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。

这个P/A过程是可以向下游继续传递的。

事实上，考察STP。指定端口的选择是很快的，主要的速度瓶颈在于如下：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这种积极的手段来加快“上游”端口转到Forwarding状态的速度。

这里要强调的是，P/A机制要求两台交换机之间链路必须是点对点的全双工状态。要注意的是，一旦P/A协商不成功，就会等待正常的STP中的两个Forward Delay，此时的行为就回到了STP。

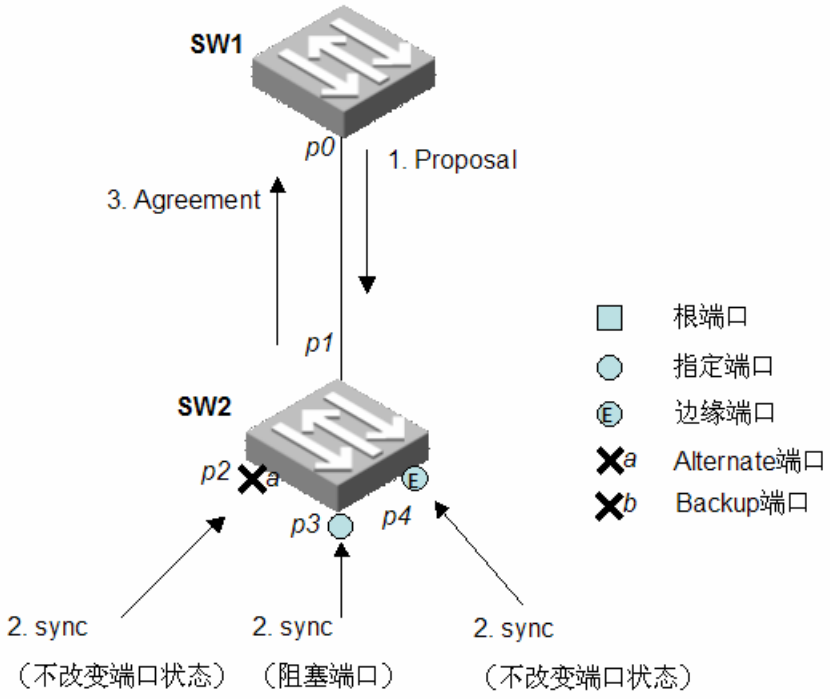

图9 Proposal/Agreement 过程示例

4.2 RSTP 的拓扑变化处理

在RSTP中检测拓扑是否发生了变化只有一个标准：**一个非edge端口迁移到Forwarding状态。**一旦检测到拓扑发生变化，则采取如下措施：

为本交换机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。如果是根端口上有状态变化，则根端口也要启动。在这个时间内，清空这些端口上学来的MAC地址；同时，由这些端口向外发送TC BPDU，其中的TC置位。根端口总是要发

送这种TC BPDU。一旦TC While Timer超时，则停止发送TC BPDU。

其他交换机接收到TC BPDU，作如下工作：

清空所有端口学来的MAC地址，收到TC BPDU的端口除外。然后也为所有自己的非边缘指定端口和自己的根端口启动TC While Timer，重复上述的过程。

如此，网络中就会产生TC BPDU的泛洪。

4.3 RSTP 与 STP 的互操作

RSTP 是可以和 STP 互操作的，但是此时会丧失快速收敛等 RSTP 优势。当一个网段里面既有 STP 交换机又有 RSTP 交换机的时候，STP 交换机会忽略 RST BPDU，而 RSTP 交换机在某端口上接收到 STP 桥的 BPDU 之后，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。在我司设备上可以配置在 STP 交换机被撤离网络后，RSTP 交换机可否迁移回 RSTP 工作状态。

5 Cisco 的 STP 特性

本章简要的介绍了Cisco的对STP协议实现的几个主要特性，详情请参见《Cisco LAN Switching》。

5.1 PVST 和 PVST+

我们讨论 STP 一直没有涉及到 VLAN，在 IEEE 802.1D 中，并没有说明 STP 和 VLAN 配合的情况（在设计 STP 协议的时候，VLAN 的 IEEE 802.1Q 还没有出现）。所以 STP 就变成了 VLAN 无关的协议了，这样就有可能出现一棵公共生成树可能包括两个以上的 VLAN 的情况。这就是所谓的 CST（Common Spanning Tree，公共生成树）。

CST 的缺点如下：由对于不同的 VLAN　STP 都进行同样的阻塞算法，则有可能导致某些链路不通。针对这种情况，Cisco 提出了 PVST（per-VLAN Spanning Tree，每 VLAN 生成树）来实现 STP 协议。PVST，即每个 VLAN 对应一个生成树。但是 PVST 也有缺点，那就是在 Trunk 链路上要对各个 VLAN 中的 BPDU 提供传输服务。而且对于每个 VLAN 的 BPDU，交换机都要分别处理，开销很大。

PVST+是对 PVST 的补充和改进。

5.2 PortFast

假设交换机的某些端口直接与主机相连，并且不和其他交换机的端口有逻辑连通，那么让这些端口也经历那Listening->Learning->Forwarding的30秒过程是毫无道理的，因此Cisco提出一种方案，叫做PortFast。即任何端口由disable状态一旦链路连接则马上进入Forwarding状态（即一旦连接，就转发），一旦检测到对发来的BPDU才返回Blocking再转入Listening状态，进入我们前面讲过的STP选举过程。这样做的好处是，我们网络“ 边缘 ”直接连接主机的交换机可以马上转发。具体细节这里不做讨论。

5.3 UplinkFast 和 BackboneFast

Cisco的Uplink Fast很类似RSTP中的“ 根端口快速切换机制 ”（事实上，Cisco的UplinkFast提出的更早些）。不过Cisco强调Uplink Fast功能只在边缘的交换机上使能。

BackboneFast是对UplinkFast一个延伸。该功能可以在整个STP域里的所有交换机上使能。使能该功能的交换机的根端口若收到一个次等BPDU，则会检查该BPDU的发送者BID，如果是上游指定桥发来的，则会认为上游链路失效。这时，该根端口会直接进入Listening状态，而不必经历20秒的默认Max Age。从而把端口再次进入Forwarding状态的时间由原来的50秒缩短到30秒。

推荐阅读

- 《Cisco LanSwitching》6，7 章

  这是关于STP的绝对经典之作，这两章非常适合于STP入门。第6章讲解的是关于STP的基础议题；第7章深入探讨了STP以及Cisco私有的PVST+等技术细节。

- 陆宇翔，《从 BPDU 看 RSTP》

  这篇小文档非常的独特，从 RSTP 的协议报文展开，详细的探讨了 RSTP 报文的交互和交换机行为，对实际的测试工作有很大帮助。该文档是我司罕有的深入探讨 RSTP 的原创文档。建议有了 STP 的基础知识之后再进行阅读。

参考资料

《Cisco LanSwitching》6，7章

IEEE 802.1D

IEEE 802.1w

陈旭盛，《STP、RSTP、MSTP学习记录》

陆宇翔，《从BPDU看RSTP》

沈岭，《STP/RSTP协议理解》

边江，《STP/RSTP理解报告》