第26章

生成树协议

一个局域网通常由多台交换机互连而成，为了避免广播风暴，需要保证在网络中不存在路径回环，也就是说所有链路应该组成一棵无回环的树，交换机上的 STP（spanning tree protocol，生成树协议）就实现了这样的功能。在本章中，首先会学习有关 STP 的一些基本概念，以及 STP 是如何通过实现冗余链路的闭塞和开启从而实现一棵动态的生成树；最后还会介绍一下 RSTP（rapid spanning tree protocol，快速生成树协议）和 MSTP（multiple spanning tree protocol，多生成树协议），以及如何在交换机上对生成树进行配置。

26.1 本章目标

学习完本章，应该能够达到以下目标。

（1）了解 STP 产生的背景。

（2）掌握 STP 基本工作原理。

（3）掌握 RSTP 和 MSTP 基本原理。

（4）掌握 STP 的配置。

26.2 STP 产生背景

透明网桥拓展了局域网的连接能力，使只能在小范围 LAN（同一冲突域）上操作的站点能够在更大范围的 LAN（多个冲突域）环境中工作。同时，它还能自主学习站点的地址信息，从而有效控制网络中的数据帧数量。但是，透明网桥在转发数据帧时，尽管它能够按照 MAC 地址表进行正确的转发，但它不会对以太网数据帧做任何修改，也没有记录任何关于该数据帧的转发记录。所以由于某种原因（如网络环路），交换机再次接收到该数据帧时，它仍然毫无记录地将数据帧按照 MAC 地址表转发到指定端口。这样，数据帧有可能在环路中不断循环和增生，造成网络带宽被大量重复帧占据，导致网络拥塞。特别是在遇到广播帧时，更容易在存在环路的网络中形成广播风暴。

图 26-1 是一个由于环路造成数据帧循环和增生的例子。

（1）开始，假定 PCA 还没有发送过任何帧，因此，网桥 SWA、SWB 和 SWC 的地址表中都没有 PCA 的地址记录。

（2）当 PCA 发送了一个帧，最初 3 个网桥都接收了这个帧，记录 PCA 的地址在物理段 A 上，并将这个帧转发到物理段 B 上。

（3）网桥 SWA 会将此帧转发到物理段 B 上，从而 SWB 和 SWC 将会再次接收到这个帧，因为 SWA 对于 SWB 和 SWC 来说是透明的，这个帧就好像是 PCA 在物理段 B 上发送的一样，于是 SWB 和 SWC 记录 PCA 在物理段 B 上，将这个新帧转发到物理段 A 上。

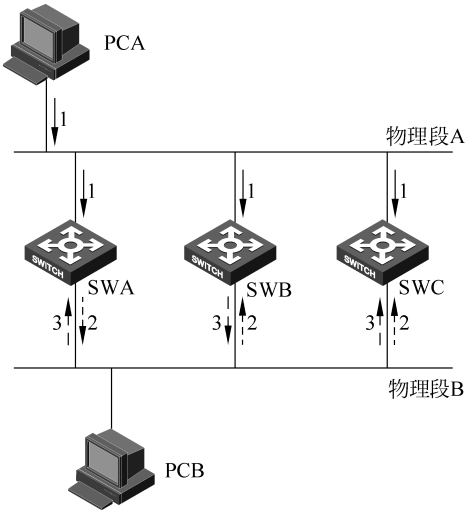

图 26-1 环路造成数据帧循环和增生

(4) 同样的道理,SWB 会将最初的帧转发到物理段 B 上,那么 SWA 和 SWC 都接收到这个帧。SWC 认为 PCA 仍然在物理段 B 上,而 SWA 又发现 PCA 已经转移到物理段 B 上了,然后 SWA 和 SWC 都会转发新帧到物理段 A 上。如此下去,帧就在环路中不断循环,更糟糕的是每次成功的帧发送都会导致网络中出现两个新帧。

那么应该怎样来解决这个问题呢? 首先可能想到的是保证网络不存在物理上的环路。但是,当网络变得复杂时,要保证没有任何环路是很困难的,并且在许多可靠性要求高的网络,为了能够提供不间断的网络服务,采用物理环路的冗余备份就是最常用的手段了。所以,保证网络不存在环路是不现实的。

IEEE 提供了一个很好的解决办法,那就是 802.1D 协议标准中规定的 STP,它能够通过阻断网络中存在的冗余链路来消除网络可能存在的路径环路,并且在当前活动(active)路径发生故障时,激活被阻断的冗余备份链路来恢复网络的连通性,保障业务的不间断服务。

在图 26-2 中给出了一个应用生成树的桥接网络的例子,其中字符 ROOT 所标识的网桥是生成树的树根,实线是活动的链路,也就是生成树的枝条,而虚线则是被阻断的冗余链路,只有在活动链路断开时才会被激活。

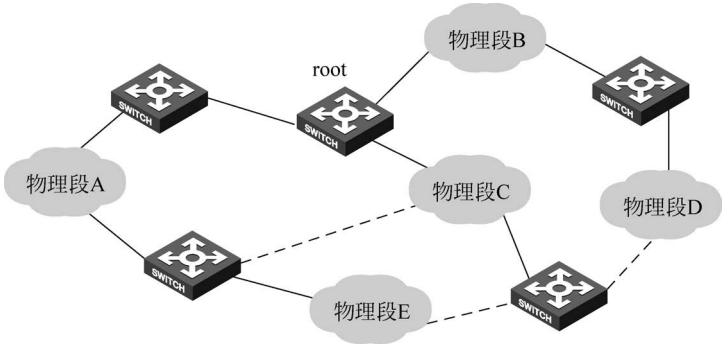

图 26-2 生成树网络

26.3　STP

STP是由IEEE协会制定的，用于在局域网中消除数据链路层物理环路的协议，其标准名称为802.1D。运行该协议的设备通过彼此交互信息发现网络中的环路，并有选择地对某些端口进行阻塞，最终将环路网络结构修剪成无环路的树型网络结构，从而防止报文在环路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的报文处理能力下降的问题发生。

26.3.1　桥协议数据单元

STP采用的协议报文是BPDU(bridge protocol data unit，桥协议数据单元)，BPDU中包含了足够的信息来完成生成树的计算。

BPDU分为以下两类。

(1) 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。

(2) TCN BPDU(topology change notification BPDU)：当拓扑结构发生变化时，用来通知相关设备网络拓扑结构发生变化的报文。

STP协议的配置BPDU报文携带了如下几个重要信息。

(1) 根桥ID(root ID)：由根桥的优先级和MAC地址组成。通过比较BPDU中的根桥ID，STP最终决定谁是根桥。

(2) 根路径开销(root path cost)：到根桥的最小路径开销。如果是根桥，其根路径开销为0；如果是非根桥，则为到达根桥的最短路径上所有路径开销的和。

(3) 指定桥ID(designated bridge ID)：生成或转发BPDU的桥ID，由桥优先级和桥MAC组成。

(4) 指定端口ID(designated port ID)：发送BPDU的端口ID，由端口优先级和端口索引号组成。

各台设备的各个端口在初始时会生成以自己为根桥的配置消息，根路径开销为0，指定桥ID为自身设备ID，指定端口为本端口。各台设备都向外发送自己的配置消息，同时也会收到其他设备发送的配置消息。通过比较这些配置消息，交换机进行生成树计算，选举根桥，决定端口角色。最终，生成树计算的结果如下。

(1) 对于整个STP网络，唯一的一个根桥被选举出来。

(2) 对于所有的非根桥，选举出根端口和指定端口，负责流量转发。

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，BPDU报文携带有(root ID、root path cost、designated bridge ID、designated port ID)等信息，然后传播到整个网络。其他网桥收到BPDU报文后，根据报文中携带的信息而进行计算，确定端口角色，然后向下游网桥发出更新后的BPDU报文。BPDU交互示例如图26-3所示。

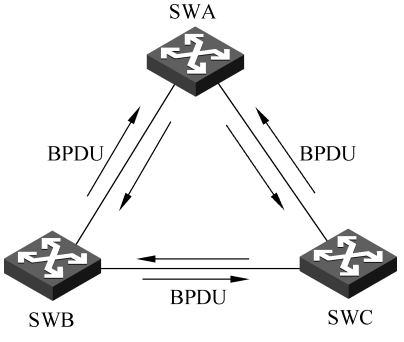

图26-3　BPDU交互

26.3.2　根桥选举

树形的网络结构，必须要有树根，于是STP引入了根桥(root bridge)的概念。

网络中每台设备都有自己的桥 ID,桥 ID 由桥优先级(bridge priority)和桥 MAC 地址(bridge Mac address)两部分组成。因为桥 MAC 地址在网络中是唯一的,所以能够保证桥 ID 在网络中也是唯一的。在进行桥 ID 比较时,先比较优先级,优先级值小者为优;在优先级相等的情况下,再用 MAC 地址来进行比较,MAC 地址小者为优。

网络初始化时,网络中所有的 STP 设备都认为自己是“根桥”。设备之间通过交换配置 BPDU 而比较桥 ID,网络中桥 ID 最小的设备被选为根桥。根桥会按照一定的时间间隔产生并向外发送配置 BPDU,其他的设备对该配置 BPDU 进行转发,从而保证拓扑的稳定。

在图 26-4 中,3 台交换机参与 STP 根桥选举。SWA 的桥 ID 为 0.0000-0000-0000,SWB 的桥 ID 为 16.0000-0000-0001,SWC 的桥 ID 为 0.0000-0000-0002。3 台交换机之间进行桥 ID 比较。因为 SWA 与 SWC 的桥优先级最小,所以排除 SWB;而比较 SWA 与 SWC 之间的 MAC 地址,发现 SWA 的 MAC 地址比 SWC 的 MAC 地址小,所以 SWA 被选举为根桥。

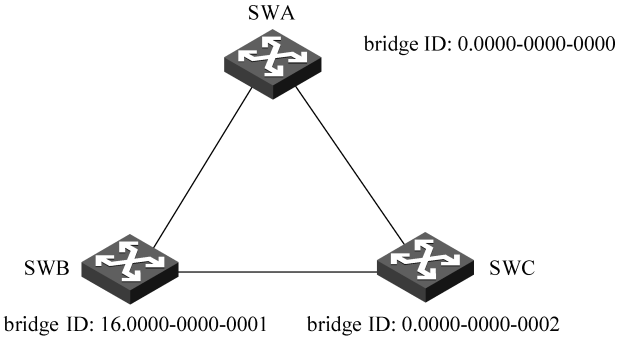

图 26-4 根桥的选举

因为桥的 MAC 地址在网络中是唯一的,所以网络中总能够选举出根桥。

26.3.3 确定端口角色

STP 的作用是通过阻断冗余链路使一个有回路的桥接网络修剪成一个无回路的树型拓扑结构。它通过将环路上的某些端口置为阻塞状态,不允许数据帧通过而做到这一点。下面是确定哪些端口是阻塞状态的过程。

(1) 根桥上的所有端口为指定端口(designated port,DP)。

(2) 为每个非根桥选择根路径开销最小的那个端口作为根端口(root port,RP),该端口到根桥的路径是此网桥到根桥的最佳路径。

(3) 为每个物理段选出根路径开销最小的那个网桥作为指定桥(designated bridge),该指定桥到该物理段的端口作为指定端口,负责所在物理段上的数据转发。

(4) 既不是指定端口,也不是根端口的端口,而是 alternate 端口,置于阻塞状态,不转发普通以太网帧。

图 26-5 是一个 STP 确定端口角色的示例。

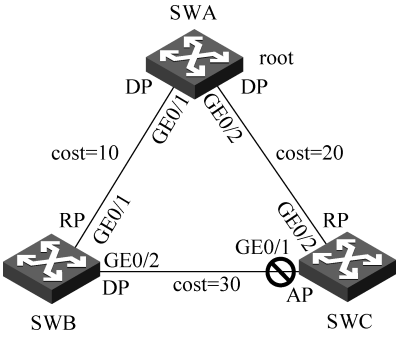

图 26-5 端口角色确定

1. SWA 端口角色的确定

在图 26-5 中,STP 经过交互 BPDU 配置报文,选举出 SWA 为根桥。因为根桥是 STP 网络中数据转发的中心点,所以根桥上的所有端口都是指定端口,处于转

发状态，向它的下游网桥转发数据。

注意：此处的上游网桥、下游网桥是根据 BPDU 报文转发的流向来定义的。数据报文的转发并没有上游、下游之分。

2. SWB 端口角色的确定

从拓扑可知，SWB 上有两个端口能够收到根桥 SWA 发来的 BPDU，也就是说，SWB 上有两个端口能够到达根桥。STP 必须判定哪个端口离根桥最近，它通过比较到达根桥的开销(cost)来做到这一点。在图 26-5 中，端口 GE0/1 到达根桥的开销是 10，而端口 GE0/2 到达根桥的开销是 20＋30＝50，很明显，端口 GE0/1 到达根桥开销小，也就是端口 GE0/1 离根桥最近，所以 STP 确定端口 GE0/1 是 SWB 上的根端口，端口处于转发状态。

对于非根桥来说，只需要一个端口为根端口。因为很明显，如果非根桥有两个端口为根端口，处于转发状态；而根桥上所有端口肯定都是指定端口，也处于转发状态，环路就形成了。这有悖于 STP 阻塞交换网络环路的初衷，所以端口 GE0/2 不能成为根端口。

在 SWB 和 SWC 之间存在着物理段(物理链路)。在实际网络中，这条物理段有可能通过 hub 或不支持 STP 的交换机连接到终端主机，所以 STP 必须考虑如何将数据转发到这条物理段上。那么是由 SWB 还是由 SWC 来负责向这条物理段转发数据呢？这取决于哪一个网桥离根桥近，离根桥最近的网桥负责向这个网段转发数据。

所以，通过交互 BPDU，STP 发现 SWB 离根桥近(因为 SWB 到根桥的开销是 10，小于 SWC 到根桥的开销 20)，所以 STP 确定 SWB 是 SWB 和 SWC 之间物理段的指定桥，而端口 GE0/2 也就是指定端口，处于转发状态。

3. SWC 端口角色的确定

因为 SWC 与 SWB 同为非根桥，所以 SWC 确定端口的过程与 SWB 类似。端口 GE0/2 离根桥近，所以被确定为根端口。

在 STP 协议中，一个物理段上只需要确定一个指定端口。如果一个物理段上有两个指定端口，都处于转发状态，则会在图 26-5 的拓扑环境中产生环路。由于 SWB 与 SWC 之间物理段已经确定好了指定端口(SWB 的端口 GE0/2)，所以 SWC 的端口 GE0/1 不能成为指定端口。端口 GE0/1 不能成为根端口(因为端口 GE0/2 已经是根端口，一个桥只能有一个根端口)，也不能成为指定端口，则端口 GE0/1 处于阻塞状态。

26.3.4　根路径开销

根路径开销是 STP 中用来判定到达根桥的距离的参数。STP 在进行根路径开销计算时，是将所接收 BPDU 中的 root path cost 值加上自己接收端口的链路开销值。对根桥来说，其根路径开销为 0；对非根桥来说，根路径开销为到达根桥的最短路径上所有路径开销的和。

通常情况下，链路的开销与物理带宽成反比。带宽越大，表明链路通过能力越强，则路径开销越小。

IEEE 802.1D 和 802.1t 定义了不同速率和工作模式下的以太网链路(端口)开销，H3C 则根据实际的网络运行状况优化了开销的数值定义，制定了私有标准。上述 3 种标准的常用定义如表 26-1 所示。其他细节定义参照相关标准文档及设备手册。

表 26-1 链路开销标准

链路速率	802.1D—1998	802.1t	私有标准
0	65535	200000000	200000
10Mbps	100	2000000	2000
100Mbps	19	200000	200
1000Mbps	4	20000	20
10Gbps	2	2000	2

H3C 交换机默认采用私有标准定义的链路开销。交换机端口的链路开销可手工设置,以影响生成树的路由选择。

图 26-6 是根路径开销计算示例。因为 SWA 是根桥,所以它所发出的 BPDU 报文中所携带的 root path cost 值为 0。SWB 从端口 GE0/1 收到 BPDU 报文后,将 BPDU 中的 root path cost 值与端口 cost(千兆以太网链路的默认值是 20)相加,得出 20,则 SWB 的端口 GE0/1 到根的 root path cost 值为 20。然后更新自己的 BPDU,从另一个端口 GE0/2 转发出去。

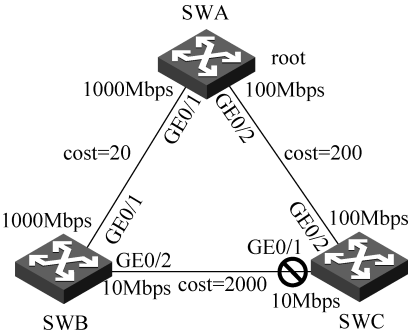

图 26-6 根路径开销计算

同理,SWC 从端口 GE0/1 收到 SWB 发出的 BPDU 报文后,将 BPDU 中的 root path cost 值 20 与端口 cost 值 2000 相加,得出 2020。则 SWC 的端口 GE0/1 到根的 root path cost 值为 2020。同样,可以计算出,SWB 的端口 GE0/2 到根的 root path cost 值为 2200,SWC 的端口 GE0/2 到根的 root path cost 值为 200。

26.3.5 桥 ID 的作用

在前面的示例中,交换机根据根路径开销来确定了端口角色。但在某些网络拓扑中,根路径开销是相同的,这时 STP 需要根据桥 ID 来决定端口角色。

当一个非根桥上有多个端口经过不同的上游桥到达根桥,且这些路径的根路径开销相同时,STP 会比较各端口的上游指定桥 ID,所连接到上游指定桥 ID 最小的端口被选举为根端口。当一个物理段有多个网桥到根桥的路径开销相同,进行指定桥选举时,也比较这些网桥的桥 ID,桥 ID 最小的桥被选举为指定桥,指定桥上的端口为指定端口。

在图 26-7 中,SWD 有两个端口能到达根,且根路径开销是相同的。但因 SWB 的桥 ID 小于 SWC 的桥 ID,所以连接 SWB 的端口为根端口。同样,SWB 被选举为 SWB 和 SWC 之间物理段的指定桥,SWB 上的端口为指定端口。

因为桥 ID 是唯一的,所以通过比较桥 ID 可以对经过多个桥到达根桥的路径好坏进行最终判定。

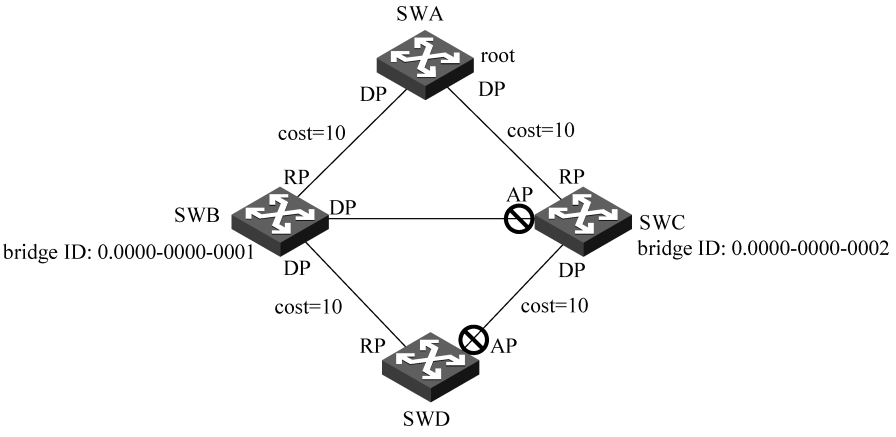

图 26-7　桥 ID 作用

26.3.6　端口 ID 的作用

在根路径开销和上游指定桥 ID 都相同的情况下，STP 根据端口 ID 来决定端口角色。

如果非根桥上多个端口经过相同的上游桥到达根，且根路径开销相同，则 STP 会比较端口所连上游桥的端口 ID，所连接到上游指定端口 ID 最小的端口被选举为根端口。

端口 ID 由端口索引号和端口优先级两部分组成。在进行比较时，先比较端口优先级，优先级小的端口优先；当优先级相同时，再比较端口索引号，索引号小的端口优先。

在图 26-8 中，SWB 上的两个端口连接到 SWA，这两个端口的根路径开销相同，上游指定桥 ID 也相同，STP 根据上游指定端口 ID 来判定。由于在默认情况下，端口优先级相同，所以只能比较端口索引号，因此，连接 SWA 上端口 G0/1 的端口为根端口。

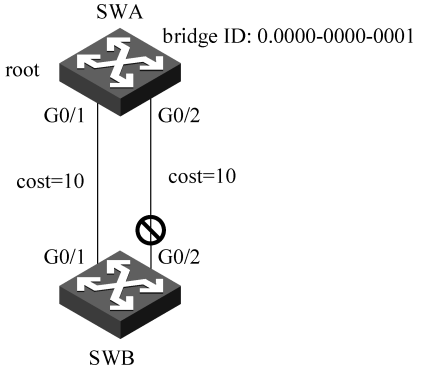

图 26-8　端口 ID 作用

通常情况下，端口索引号无法改变，用户可通过设置端口优先级来影响生成树的路由选择。比如，如果想让 SWB 的端口 G0/1 成为阻塞状态，则在 SWA 上调整端口 G0/2 的优先级大于 G0/1 即可。

26.3.7　端口状态

前面讨论了 STP 如何确定端口角色。被确定为根端口或指定端口后，端口就可以处于转发状态，否则就是阻塞状态。

事实上，在 802.1D 的协议中，端口共有 5 种状态。

(1) disabled：表示该端口处于失效状态，不接收和发送任何报文。这种状态可以是由于端口的物理状态(比如端口物理层没有 up)导致的，也可能是管理者手工将端口关闭。

(2) blocking：处于这个状态的端口不能够参与转发数据报文，但是可以接收 BPDU 配置消息，并交给 CPU 进行处理。不过不能发送配置消息，也不能进行地址学习。

(3) listening：处于这个状态的端口也不参与数据转发，不进行地址学习；但是可以接收并发送 BPDU 配置消息。

(4) learning：处于这个状态的端口同样不能转发数据，但是开始地址学习，并可以接收、处理和发送 BPDU 配置消息。

(5) forwarding：一旦端口进入该状态，就可以转发任何数据了，同时也进行地址学习和 BPDU 配置消息的接收、处理和发送。

以上 5 种状态中，listening 和 learning 是不稳定的中间状态，它们主要的作用是使 BPDU 消息有一个充分时间在网络中传播，杜绝由于 BPDU 丢失而造成的 STP 计算错误，导致环路的可能。

在一定条件下，端口状态之间是可以互相迁移的，如图 26-9 所示。

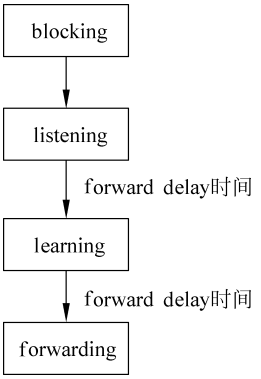

图 26-9 端口状态迁移图

当一个端口由于拓扑发生改变不再是根端口或指定端口了，就会立刻迁移到 blocking 状态。

当一个端口被选为根端口或指定端口，就会从 blocking 状态迁移到一个中间状态 listening 状态；经历 forward delay 时间，迁移到下一个中间状态 learning 状态；再经历一个 forward delay 时间，迁移到 forwarding 状态。

从 listening 状态迁移到 learning 状态，或者从 learning 状态迁移到 forwarding 状态，都需要经过 forward delay 时间，通过这种延时迁移的方式，能够保证当网络的拓扑发生改变时，新的配置消息能够传遍整个网络，从而避免由于网络未收敛而造成临时环路。

在 802.1D 中，默认的 forward delay 时间是 15s。所以，当一个端口被选为根端口或指定端口后，至少要经过两倍的 forward delay 时间，即 30s 才能够转发数据。

在实际的应用中，STP 也有很多不足之处。最主要的缺点是端口从阻塞状态到转发状态需要两倍的 forward delay 时间，导致网络的连通性至少要几十秒的时间之后才能恢复。如果网络中的拓扑结构变化频繁，网络会频繁地失去连通性，这样用户就会无法忍受，如图 26-10 所示。

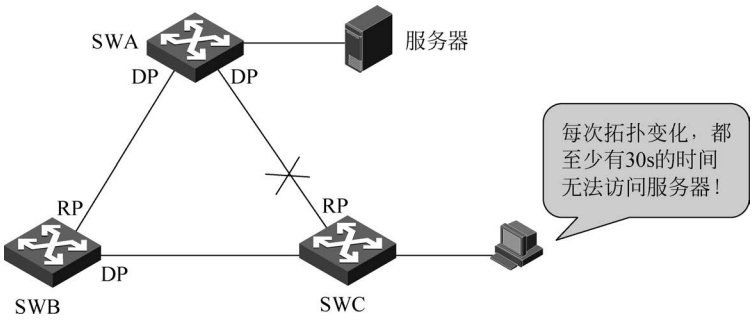

图 26-10 生成树的不足

为了在拓扑变化后网络尽快恢复连通性，交换机在 STP 的基础上发展出 RSTP。

26.4　RSTP

RSTP 是 STP 的优化版。IEEE 802.1w 定义了 RSTP，并最终合并入了 802.1D—2004。RSTP 是从 STP 算法的基础上发展而来，承袭了它的基本思想，也是通过配置消息来传递生成树信息，并进行生成树计算。

RSTP 能够完成生成树的所有功能，不同之处就在于：在某些情况下，当一个端口被选为根端口或指定端口后，RSTP 减小了端口从阻塞到转发的时延，尽可能快地恢复网络连通性，提供更好的用户服务。

在 IEEE 802.1w 中，RSTP 从 3 个方面实现“快速”功能。

1. 端口被选为根端口

如图 26-11 所示，交换机上原来有两个端口能够到达根桥，其中 cost 值为 10 的端口 G0/1 被选为根端口，另外一个为备用端口（处于阻塞状态）。如果 cost 值变为 30 后，STP 重新计算，选择原来处于阻塞状态的端口 G0/2 为根端口。此时，故障恢复的时间就是根端口的切换时间，无须延时，无须传递 BPDU，只是一个 CPU 处理的延时，约几毫秒。

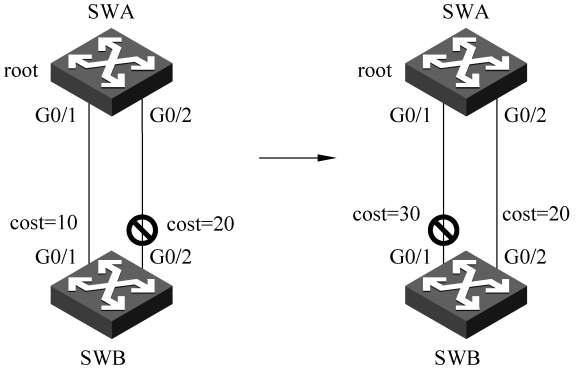

图 26-11　RSTP 改进 1

2. 指定端口是非边缘端口

此时情况较复杂。“非边缘”的意思是这个端口连接着其他的交换机，而不是只连接到终端设备。此时，如果交换机之间是点对点链路，则交换机需要发送握手报文到其他交换机进行协商，只有对端返回一个赞同报文后，端口才能进入转发状态。

在图 26-12 中，SWA 的端口 G0/1 原来处于阻塞状态。STP 重新选择该端口作为指定端口后，因为 G0/1 连接有下游网桥 SWB，它并不知道下游有没有环路，所以会发一个握手报文，目的是询问下游网桥是否同意这个端口进入转发状态。SWB 收到握手报文后，发现自己没有端口连接到其他网桥，也就是说，这个网桥是边缘网桥，不会有环路产生。则 SWB 回应一个赞同报文，表明同意 SWA 的端口 G0/1 进入转发状态。

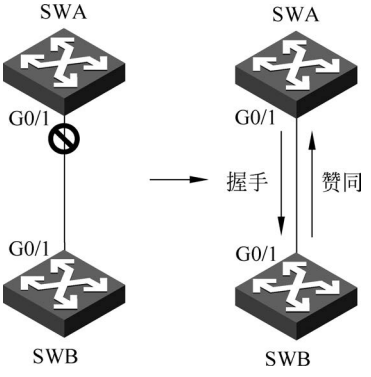

图 26-12　RSTP 改进 2

不过，RSTP 规定只有在点对点链路上，网桥才可以发起握手请求。因为非点对点链路意味着可能连接多个下游

网桥,并不是所有网桥都能够回应赞同报文。如果只有其中一个下游网桥回应赞同报文,上游网桥端口就处于转发状态,则可能导致环路。

可见点对点链路对 RSTP 的性能有很大的影响,下面列举了点对点链路的几种情况。

(1) 该端口是一个链路聚合端口。(参考相关章节的描述)

(2) 该端口支持自协商功能,并通过协商工作在全双工模式。(参考相关章节的描述)

(3) 管理者将该端口配置为一个全双工模式的端口。

如果是非点对点链路,则恢复时间与 STP 无异,是两倍的 forward delay 时间,默认情况下是 30s。

在 RSTP 握手协商时,总体收敛时间取决于网络直径,也就是网络中任意两点间的最大网桥数量。最坏的情况是,握手从网络的一边开始扩散到网络的另一边,比如,网络直径为 7 的情况,最多可能要经过 6 次握手,网络的连通性才能被恢复。

3. 指定端口是边缘端口

"边缘端口"是指那些直接和终端设备相连,不再连接任何交换机的端口。这些端口不需要参与生成树计算,端口可以无时延地快速进入转发状态。此时不会造成任何的环路。

在图 26-13 中,SWA 的 G0/1 原来连接有网桥,现连接到终端主机。这些端口为边缘指定端口,端口 G0/1 可马上进入转发状态。

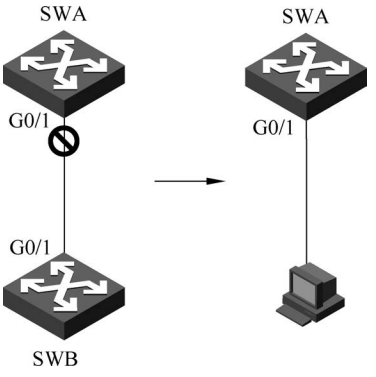

图 26-13 RSTP 改进 3

那么网桥是如何判定是边缘指定端口还是非边缘指定端口呢? 事实上,网桥无法判定,只有管理员可以指定。

26.5 MSTP

STP 使用生成树算法,能够在交换网络中避免环路造成的故障,并实现冗余路径的备份功能。RSTP 则进一步提高了交换网络拓扑变化时的收敛速度。

然而当前的交换网络往往工作在多 VLAN 环境下。在 802.1Q 封装的 trunk 链路上,同时存在多个 VLAN,每个 VLAN 实质上是一个独立的二层交换网络。为了给所有的 VLAN 提供环路避免和冗余备份功能,就必须为所有的 VLAN 都提供生成树计算。

传统 STP/RSTP 采用的方法是使用统一的生成树。所有的 VLAN 共享一棵生成树 (common spanning tree,CST),其拓扑结构也是一致的。因此,在一条 trunk 链路上,所有的 VLAN 要么全部处于转发状态,要么全部处于阻塞状态。

在如图 26-14 所示的情况下,SWB 到 SWA 的端口被阻塞,则从 PCA 到服务器的所有数