Chapter 5

第 5 章　增强交换网络健壮性

项目背景

中山大学电子工程系、计算机科学技术系在学院改制前，都是独立建制学院，分别建有独立的网络，各学院早期规划网络拓扑如图 5-1 所示。

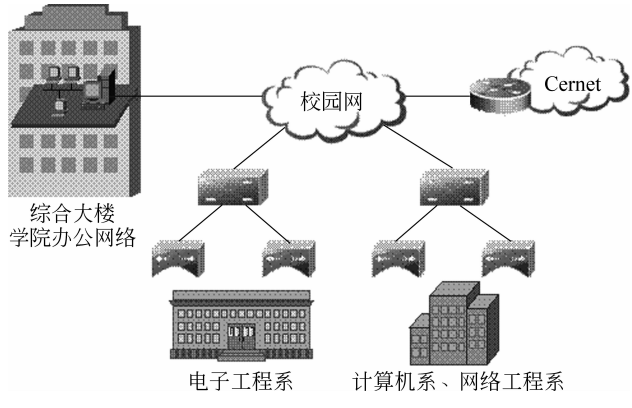

图 5-1　各独立学院早期网络拓扑

在院系改制中，学校把原电子工程系、计算机科学技术系合并成一个综合性的计算机科学技术学院。为整合教学资源，对原有院系的网络进行了改造和整合。

项目分析

网络中由于单线链路容易造成网络故障，因此在网络规划中，需要增加网络的备份和冗余，增加网络的稳定性。

冗余链路的作用是，数据在双链路同时传输，确保在一条线路出现故障的情况下，网络不会瘫痪，冗余链路可以保障数据不会因为某条单一链路的故障而中断。

改造后的网络，把原来各系分隔的网络联成一体，升级学院主干网络，整合各学院网络资源。为保证网络的稳定性，在主干网络改造过程中，增加网络冗余。

首先在网络核心增加了一台三层设备，使用双链路形成网络的冗余备份。其次，在骨干链路上，为增加网络的带宽，使用链路聚合技术，提升骨干链路的高带宽。

如图 5-2 所示网络拓扑，是改造后的学院新网络，图中虚线部分圈出的区域，是本节知识和技术主要的应用场景。

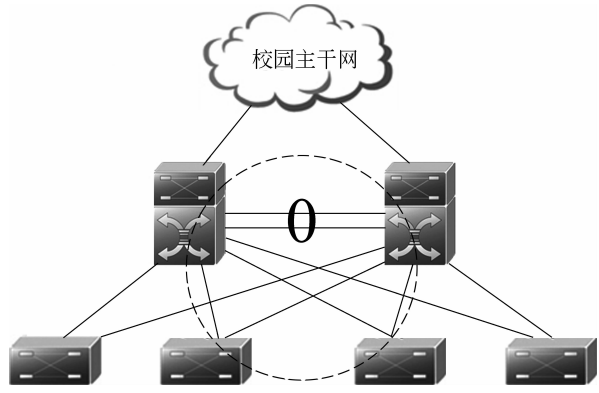

图 5-2　增加网络冗余的网络

通过本章的学习，读者将能够了解如下内容。

(1) 掌握生成树协议 STP、RSTP。

(2) 熟悉多生成树协议 MSTP。

(3) 掌握链路聚合协议 IEEE 802.3ad。

5.1　生成树协议概述

随着交换技术在网络中的广泛应用，保障各种网络终端设备之间正常通信，成为一项重要的任务。绝大多数情况下，在交换网络中针对骨干链路，均采用多条链路连接以形成冗余链路备份，保证不会因为骨干链路上的单点故障，影响正常网络之间的通信。

5.1.1　交换网络中的冗余链路

在由多台交换机设备组成的交换网络环境中，通常都使用一些备份连接，以提高网络的健全性、稳定性，这些备份连接也叫备份链路、冗余链路等。典型的备份连接如图 5-3 所示，交换机与交换机的端口之间的链路形成一个闭合环，就是一个备份连接。

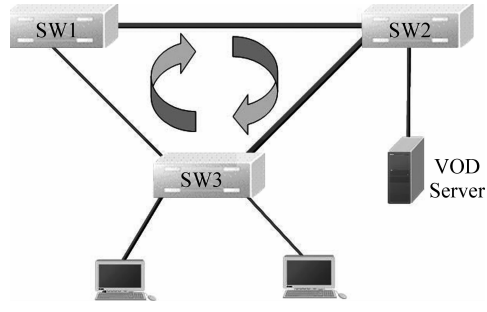

图 5-3　交换网络中的冗余链路

在主链路出现故障时，备份链路自动启用，从而提高网络的整体可靠性。在骨干网络

中,使用冗余备份连接,能够为交换网络带来健全性、稳定性和可靠性等好处。但是备份链路由于使用了冗余,容易使网络存在环路。一方面,环路链路为网络带来了网络稳定性;另一方面,环路链路也为网络带来很多严重问题,如广播风暴、多帧复制及 MAC 地址表的不稳定等。

5.1.2 冗余链路带来的网络影响

在局域网组建和维护过程中,为了提高骨干网络的链路连接的可靠性,经常需要提供冗余链路。冗余可以防止整个交换网络不会因为单点故障而造成网络中断;但它也会带来一些网络干扰问题,如广播风暴、多重帧复制,以及 MAC 地址表的不稳定性。

1. 广播风暴

二层交换机在接收到未知数据帧或者广播帧时,将执行泛洪操作,将该数据帧广播给除自己之外的所有二层端口。当网络中存在桥接和环路时,这样的泛洪就容易产生广播风暴。广播风暴(大量的泛洪帧)可能会迅速导致网络中断,如图 5-4 所示。

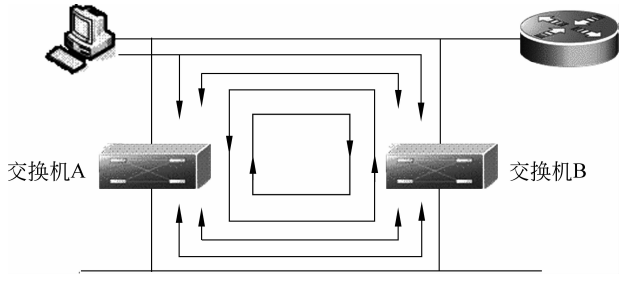

图 5-4 广播风暴示例

在较大型网络中,当大量广播流(如 MAC 地址查询 ARP)同时在网络中传播时,便会发生数据帧的碰撞。而当网络试图缓解这些碰撞,并重传更多的数据帧时,结果会导致全网的可用带宽减少,并最终使得网络失去连接而瘫痪,这一过程被称为广播风暴。

在一个较大规模的交换网络中,由于拓扑结构的复杂性,会有许多大大小小的环路产生。由于以太网第二层协议中,均没有控制环路机制,因此,各个小型环路产生广播风暴,将不断扩散到全网中,造成整个网络瘫痪。所以广播风暴是发生在二层网络中灾难性的故障。

而在如图 5-5 所示大型交换网络中,可能存在多个环路。在这样的网络中,所生成的广播帧的数量,可能会在几秒钟之内,以指数形式迅速增长,网络会变得不堪重负。

2. 多重帧的复制

交换机在接收到不确定单播帧时(MAC 地址表没有目的地址记录),将执行泛洪操作。这意味着在环路中,一个单播帧在传输中被复制为多个副本。

如图 5-6 所示,网络中的终端设备 X 主机,发出一个目的 MAC 地址为 Y 的单播帧到交换机 A。假若交换机 A 的 MAC 地址表查询不到 Y 的记录,该数据帧会被交换机 A 的接口广播至交换机 B,再广播至路由器 Y。这样,路由器 Y 将接收到多重来自计算机 X 的复制帧,帧的多重复制会浪费有限的网络带宽。

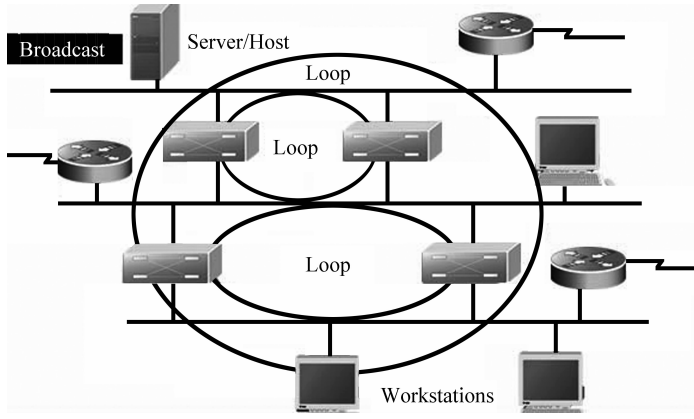

图 5-5　冗余链路中的桥接环路

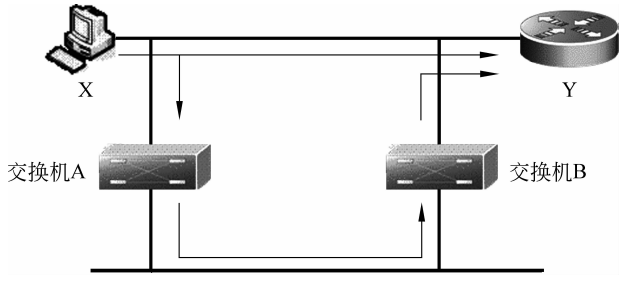

图 5-6　多重帧的复制

3. MAC 地址表的不稳定

两台交换机在接收到该数据帧后，就会以泛洪的形式从端口 1 转发出该帧，这样它们将会在端口 1 再次接收到对方交换机发送的该帧，交换机将再次把计算机 X 的 MAC 地址和端口 1 建立关联(刷新 MAC 地址记录)……这时，交换机本身无法判断出计算机 X，究竟是连接在交换机的端口 0 上还是端口 1 上？

这个流程将会一直重复下去，导致每台交换机的 MAC 地址表被多次刷新，处于动荡状态，两台交换机的 MAC 地址表都将变得不稳定，如图 5-7 所示。这种持续的更新、刷新过程，会严重耗用内存资源，影响该交换机的交换能力。

MAC 地址表的不稳定将严重影响网络性能，同时降低整个网络的运行效率。严重时，将耗尽整个网络资源，并最终造成网络瘫痪。

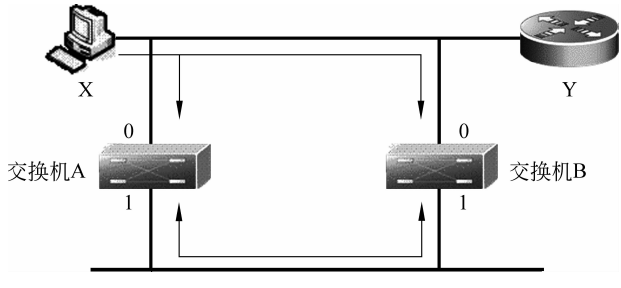

图 5-7　MAC 地址表不稳定

5.2 生成树协议介绍

为了能确保网络连接的可靠性和稳定性,在骨干网络组建过程中,需要为骨干网络提供冗余链路连接,局域网中的冗余链路提高了网络连接的可靠性。如果主链路出现故障,就启用备份链路。但是,如果交换机不知道如何处理冗余环路带来的一系列问题,将造成网络中出现广播风暴等现象,并导致网络瘫痪。

为了保持一个冗余网络的安全优势,同时防止因为环路所导致的各种问题,网络中的交换机设备必须具有下列功能。

(1) 发现环路的存在;

(2) 将冗余链路中的一条链路设为主链路,其他设为备用链路;

(3) 只通过主链路交换流量;

(4) 定期检查链路的状况;

(5) 如果主链路发生故障将流量自动切换到备用链路。

为了解决冗余链路引起的问题,IEEE 组织通过了 IEEE 802.1d 协议,即生成树协议(Spanning Tree Protocol,STP),生成树协议很好地完成了这些要求。生成树 STP 协议的基本思想十分简单,就是在网络组建过程中,构建一个树状结构、无环的网络。自然生长的树不会出现环路,如果网络也能够像树一样构建,也不会出现环路。

在具有冗余结构的网络中,生成树 IEEE 802.1d 协议通过在交换机上运行一套复杂的算法,使冗余端口置于"阻塞状态"。使得网络中的计算机在通信时,只有一条链路生效。而当这个链路出现故障时,IEEE 802.1d 协议将会重新计算出网络的最优链路,将处于"阻塞状态"的端口重新打开,保障网络正常通信,从而确保网络连接稳定可靠。

在生成树协议发展过程中,旧的缺陷不断被克服,新的特性不断被开发出来。按照功能点的改进情况,生成树协议的发展过程经历过以下三代革新。

第一代生成树协议:STP/RSTP。

第二代生成树协议:PVST/PVST+。

第三代生成树协议:MISTP/MSTP。

5.2.1 生成树协议

STP 协议的主要功能就是解决由于网络的备份链接所产生的环路问题。

STP 最初由美国数字设备公司(Digital Equipment Corp,DEC)开发,后经过电气电子工程师学会(Institute of Electrical and Electronics Engineers,IEEE)修改,最终规范为相应的 IEEE 802.1d 标准。

STP 的主要思想是:当网络中存在备份链路时,只允许主链路激活,如果主链路因故障而被断开后,备用链路才会被打开。

生成树协议检测到网络上存在环路时,自动启用算法关闭一个端口,断开环路链路。当交换机间存在多条链路时,生成树算法只启动最主要的一条链路连接,而将其他链路都阻塞

掉，将这些链路变为备用链路。当主链路出现问题时，生成树协议则自动启用备用链路，接替主链路工作，保障骨干网络的连通，不需要人工干预。

为保障生成树协议的正常运行，STP 中定义了根交换机(Root Bridge)、根端口(Root Port)、指定端口(Designated Port)等，目的就在于通过路径开销(Path Cost)计算，构造树状结构的网络，达到阻塞冗余环路的目的，同时实现链路备份和路径最优化。

STP 的本质就是利用图论中的生成树算法，对网络的物理结构不进行改变，而通过阻塞某些交换机端口，在逻辑上切断环路的方法，提取连通图，构建一个树状网络结构，以解决网络环路所造成的严重后果。

1. 网桥协议数据单元

要实现这些功能，交换机之间必须要进行一些信息的交流，这些信息交流单元就称为网桥协议数据单元(Bridge Protocol Data Unit，BPDU)。

网桥协议单元 BPDU 桢是一种二层报文，其目的 MAC 是多播地址 01-80-C2-00-00-00。所有支持 STP 的交换机，都会接收并处理收到的 BPDU 报文。该报文的数据区里携带了用于生成树计算的所有有用的信息。交换机通过交换 BPDU 来获得建立最佳树状拓扑结构所需的信息。生成树协议运行时，交换机使用共同的组播地址来发送 BPDU。

BPDU 数据帧的报文结构组成如图 5-8 所示。

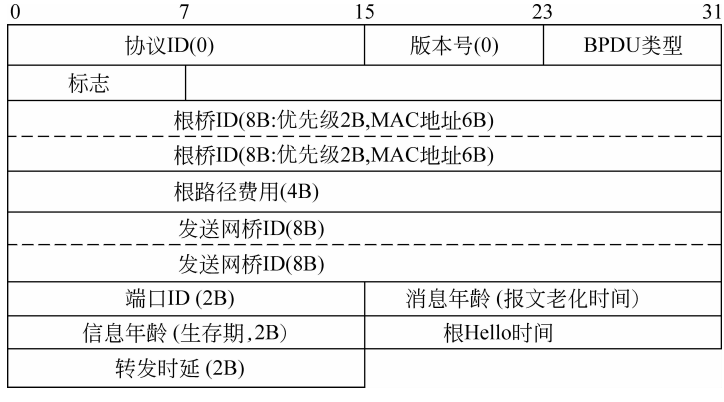

图 5-8 BPDU 报文结构

(1) 版本号：00(IEEE 802.1d STP)、02(IEEE 802.1w RSTP)。

(2) 类型：00(配置 BPDU)。其他常见的代码：80 表示 TCA；10 表示 NTC；01 表示 TC；81 表示 TCA&TC。

(3) Root ID：根交换机 ID。

(4) Root Path Cost：到达根的路径开销。

(5) Bridge ID：交换机 ID＝交换机优先级＋交换机 MAC 地址。

(6) Port ID：发送 BPDU 的端口 ID＝端口优先级＋端口编号。

(7) Hello Time：定期发送 BPDU 的时间间隔。

(8) Max-Age Time：保留对方 BPDU 消息的最长时间。

(9) Forward-Delay Time：发送延迟，端口状态改变的时间间隔。

2. 根交换机选举

冗余网络中交换机在运行生成树协议时,首先要进行根交换机的选举。根交换机的选举通过 BPDU 帧完成,选举依据是:交换机优先级和交换机 MAC 地址组合成的交换机 ID(Bridge ID),交换机 ID 最小的交换机将成为网络中的根交换机。

当冗余网络中的一台交换机的一个端口接收到高优先级的 BPDU(更小的 Bridge ID,更小的 Root Path Cost 等),就在该端口保存这些信息,同时,向所有端口更新并传播信息。如果收到比自己低优先级的 BPDU,交换机就丢弃该信息。

在如图 5-9 所示的网络场景中,各交换机都默认启动生成树协议,所有交换机的默认优先级都一样(默认优先级是 32 768),因此,MAC 地址最小的交换机将成为根交换机。因而计算出左边交换机为根,它的所有端口的角色都成为指定端口,进入转发状态,右边的交换机其中一个端口有可能会被阻塞。

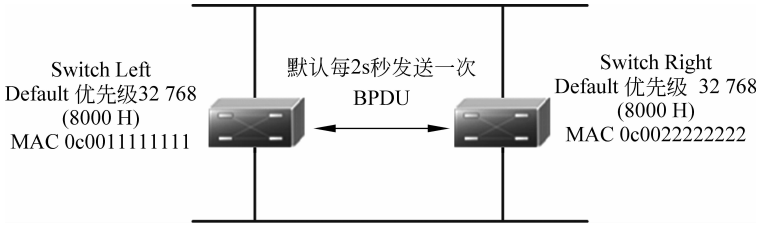

图 5-9　根交换机的选举

3. STP 选举过程

STP 通过相互之间交换 BPDU 帧中携带的消息,进行生成树的选举过程,主要的过程如下所示。

(1) 网络中选择了一台交换机为根交换机(Root Bridge)。

(2) 每台交换机都计算出了到根交换机的最短路径。

(3) 除根交换机外的每台交换机都有一个根端口(Root Port),即提供最短路径到 Root Bridge 的端口。

(4) 每个 LAN 都有指定交换机(Designated Bridge),位于该 LAN 与根交换机之间的最短路径中。指定交换机和 LAN 相连的端口称为指定端口(Designated Port)。

(5) 根端口和指定端口(Designated Port)进入转发(Forwarding)状态。

(6) 其他的冗余端口就处于阻塞(Discarding)状态。

1) 根交换机

首先进行根交换机的选举,根交换机的选举如上所示。如图 5-10 所示,假设在交换机优先级都相同的情况下,经过比较,交换机 SW1 的 MAC 地址最小,因此交换机 SW1 被选举为根交换机。

2) 根端口

接下来,其他交换机将各自选择一条“最粗壮”(高带宽)的树枝 (链路),作为非根交换机到根交换机的最短路径,其相应端口的角色就成为根端口,根端口直接进入转发状态,不阻塞。

假设在图 5-10 中,交换机 SW2 和交换机 SW1、SW3 之间的链路是千兆链路,而交换机

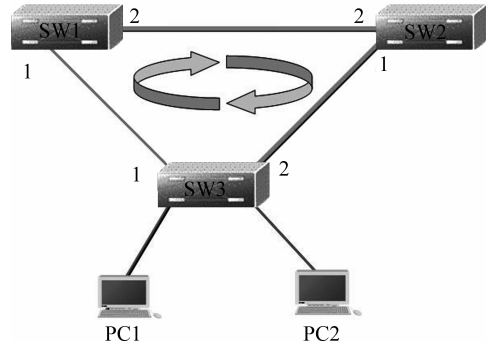

图 5-10　STP 的选举工作过程

SW1 和 SW3 之间的链路是百兆链路，通过计算可以得出：交换机 SW3 从端口 1 到根交换机 SW1 的路径开销的值是 19（百兆链路的路径开销默认值为 19）；而从交换机 SW3 端口 2→交换机 SW2→根交换机 SW1 的路径开销是 8（千兆链路的路径开销默认值为 4，则 4＋4＝8），因此该条链路成为最短路径，则交换机 SW3 的端口 2 成为根端口（非根交换机到根交换机的最短路径接口），进入转发状态。

同理，交换机 SW2 的端口 2 到根交换机 SW1 也具有最短路径，成为根端口，进入转发状态。

3）路径开销

这里的路径开销计算，主要是依据交换机的端口速率计算出该端口路径开销。

网络中每台交换机端口都有一个根路径开销（Root Path Cost），根路径开销是某交换机到根交换机所经过的路径开销的总和。在计算出最短根路径开销后，交换机对应的端口成为根端口，进入转发状态。

如表 5-1 所示，列出了不同标准的路径开销取值范围。

表 5-1　路径开销

带　　宽	IEEE 802.1d	IEEE 802.1w
10Mb/s	100	2 000 000
100Mb/s	19	200 000
1000Mb/s	4	20 000

在图 5-11 中，交换机 B 和根交换机 R、交换机 C 之间的链路是百兆链路，交换机 A 和根交换机 R、交换机 C 之间的链路是十兆链路，各交换机端口路径开销的默认值如图 5-11 所示。对于交换机 C 而言，通过交换机 A 到达根交换机的路径开销是 2 000 000＋2 000 000＝4 000 000，通过交换机 B 到达根交换机的路径开销是 200 000＋200 000＝400 000，显然 C—B—R 成为最短路径。同理，对于交换机 A 而言，A—C—B—R 为最短路径。

4）指定端口

根端口是各台交换机通往根交换机的根路径开销最低的端口，若有多个端口具有相同的根路径开销，则端口标识符小的端口为根端口。

此外，生成树协议还规定，在每个 LAN 中，都有一台交换机被称为指定交换机

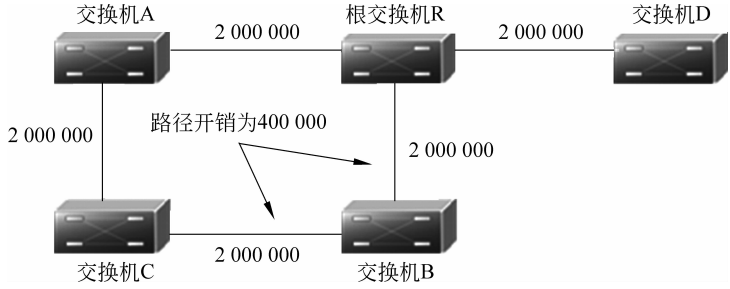

图 5-11　最短路径的选择

(Designated Bridge),它是该 LAN 中与根交换机连接而且根路径开销最低的交换机。指定交换机和 LAN 连接的端口被称为指定端口(Designated Port),指定端口不要阻塞。

如果指定交换机中有两个以上的端口连在这个 LAN 上,则具有最高优先级的端口被选为指定端口。按照生成树协议规定,根交换机上的所有端口都默认是指定端口。

交换机完成生成树计算后,所有根端口和指定端口进入转发(Forwarding)状态,其他的冗余端口则处于阻塞(Discarding)状态。

以图 5-12 为例,网络中所有的链路带宽均为百兆链路,所有交换机的优先级均相同,STP 的工作过程如下。

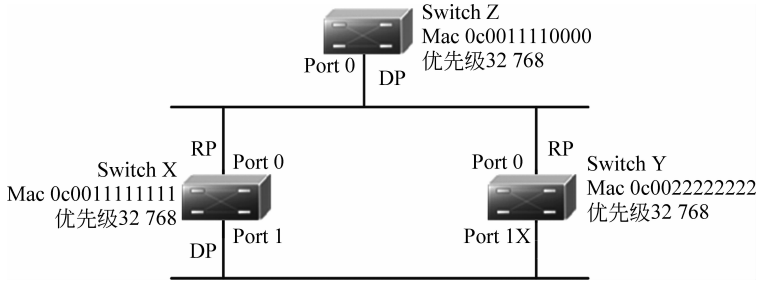

图 5-12　根端口和指定端口的确认

(1) 通过比较交换机 MAC,交换机 Z 成为根交换机,它的端口 0 默认成为指定端口。

(2) 通过计算链路最短根路径开销,交换机 X 和 Y 的端口 0 成为根端口。

(3) 交换机 X 和交换机 Y 的端口 1,都同时连接在同一个 LAN 上,通过比较优先顺序,交换机 X 的端口 1 成为指定端口。

(4) 所有根端口和指定端口进入转发状态,其他端口(交换机 Y 的端口 1)被禁用,进入阻塞状态。

这样,当根交换机确定之后,通过最短根路径开销的优先顺序比较,根端口和指定端口被确定下来,之后一棵树就生成了。

4. STP 的端口状态和拓扑变化

在一个启用 STP 的网络中,所有交换机端口在启动之后,都将经历阻塞状态以及侦听和学习这两种过渡状态。

正确配置的端口最终会稳定在转发状态或者阻塞状态,处于转发状态的端口提供了到

达根交换机的最短路径开销，处于阻塞状态的端口则作为备份链路随时待命。当交换机识别出网络拓扑变化时，交换机端口的状态变化过程如图 5-13 所示。

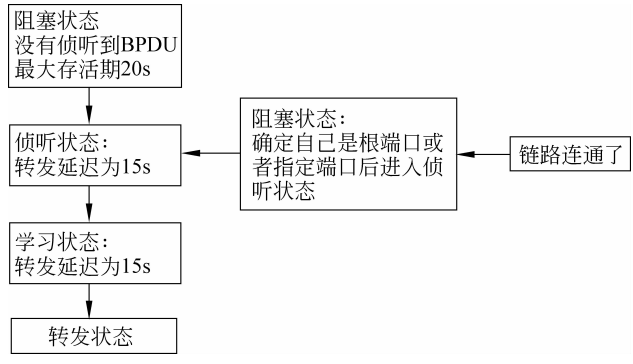

图 5-13 STP 的端口状态

其中，在生成树收敛的过程中，STP 各端口状态如下。

(1) 阻塞状态(Blocking)：不转发数据帧、接收 BPDU。

(2) 侦听状态(Listening)：不转发数据帧、侦听数据帧。

(3) 学习状态(Learning)：不转发数据帧、学习地址。

(4) 转发状态(Forwarding)：转发数据帧、学习地址。

(5) 禁止状态(Disable)：不转发数据帧、不接收 BPDU。

STP 初始化时，所有端口处于阻塞状态，当交换机启用时会把自己作为根交换机而处于侦听阶段；或者在 Max-Age Time 内没有收到新的 BPDU，交换机也会把端口从阻塞状态转换到侦听状态。

在侦听状态下，交换机不传送用户数据。在此状态下，交换机完成根交换机的选举，确认根端口和指定端口。经过 15s 转发延迟之后，如果有端口被确定为根端口或指定端口，则它将进入学习阶段；否则，仍然转回阻塞阶段。

为了减少数据转发前的泛洪数量，防止临时环路出现，默认情况下学习状态也是转发延迟时间 15s。当学习状态结束后，那些依然是根端口或指定端口的将进入转发状态，从而“实时”地适应网络拓扑的变化。

5.2.2 快速生成树协议

1. STP 的问题

早期的 STP 解决了交换链路冗余带来的广播风暴等问题，但在拓扑发生改变时，新的 BPDU 桢要经过一定的时延，才能传播到整个网络，这个时延称为 Forward Delay，协议默认为 15s。在所有交换机收到这个变化的消息之前，若旧拓扑结构中部分交换机由于收敛速度慢，仍处于转发状态的端口，还没有发现自己应当在新的拓扑中停止转发，则可能存在临时环路。

在默认状态下，BPDU 帧的报文周期为 2s，最大保留时间为 20s，端口状态改变(由侦听