第 5 章　链路及设备的冗余管理

在网络拓扑结构中，关键设备之间由于受物理带宽的限制可能会产生通信瓶颈，可以通过两个设备之间的多条物理链路捆绑在一起形成逻辑链路，增大带宽。

网络设备之间的多点连接，能够形成冗余路径，以保障正常通信，但是这种冗余路径会在网络中形成环路，造成严重的广播风暴，可以通过生成树协议来避免网络中环路的产生。

关键网络设备的单点故障，会造成业务中断，可以通过虚拟路由冗余协议（Virtual Router Redundancy Protocol，VRRP）形成关键设备的热备份。

5.1　链路聚合

5.1.1　链路聚合概述

为了增加交换机或路由器之间设备的链路带宽，通常把两台设备之间的多条物理链路捆绑在一起形成一个高带宽的逻辑链路，如图 5-1 所示。这种方式称为链路聚合（Link Aggregation），又称端口聚集（Port Trunking）、端口捆绑（Bonding）技术。

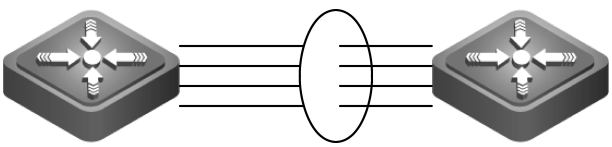

图 5-1　交换机间链路聚合

链路聚合是把网络设备的多个物理端口带宽叠加，使多个低带宽物理端口捆绑成一条高带宽逻辑链路，同时通过几个端口共同传输数据形成链路的负载均衡，聚合而成的逻辑端口称为聚合端口（Aggregate Port，AP）。

当逻辑链路中的部分物理链路断开时，系统会自动将断开链路的流量分配到逻辑链路的其他有效物理链路上，但是一条成员链路收到的广播或组播报文，将不会被转发到其他成员链路上。这种链路聚合的方式既可以通过流量均衡避免链路出现拥塞现象，也可以防止由于单条链路速率过低而出现延时现象。在不增加更多成本的前提下，既实现了网络的高速性，又能保证链路的负载分担和冗余性，提供更高的连接可靠性。

在图 5-1 中，将 4 条 1000Mb/s 的千兆以太网链路用链路聚合技术组合成一个逻辑高速链路，这条逻辑链路在全双工状态下能够达到 8000Mb/s 的带宽，聚合内部的 4 条物理链路共同完成数据的收发，逻辑链路中只要还存在能正常工作的物理链路，整个传输链路就不会失效。

IEEE 802.3ad 标准定义了如何将两个以上的物理端口组合为高带宽的逻辑链路，以实现负载共享、负载平衡以及提供更好的弹性。

链路聚合具有如下一些优点。

(1) 提高链路可用性。

链路聚合中,链路成员之间互相动态备份,当某一成员链路中断时,其他链路能够分担该成员的流量,切换过程在链路聚合内部快速实现,与其他链路无关。

(2) 增加链路带宽。

通过多个物理端口的捆绑,增加了链路的带宽,提高了链路的传输速率,并通过流量负载平衡,实现流量分担。

(3) 易于实现、高性价比。

只要支持 IEEE 802.3ad 标准的设备,都可以实现链路聚合,用比较经济的手段,实现高速带宽的能力。

配置链路聚合功能时,成为聚合端口的成员必须具备以下相同的属性。

(1) 端口均为全双工模式。

(2) 端口类型必须相同,比如同为以太网口或同为光纤口。

(3) 端口同为 Access 端口并且属于同一个 VLAN,或者同为 Trunk 端口,属于不同 Native VLAN 的端口不能构成 AP。

5.1.2 流量平衡

聚合端口可以根据数据帧的源 MAC 地址、目的 MAC 地址、源 MAC 地址+目的 MAC 地址、源 IP 地址、目的 IP 地址以及源 IP 地址+目的 IP 地址等方式把流量平均地分配到各成员链路中。

源 MAC 地址流量平衡是根据数据帧的源 MAC 地址把流量分配到聚合端口的各成员链路中。不同源 MAC 地址的流量,转发的成员链路不同,源 MAC 地址相同的流量,将从同一个成员链路中转发。

目的 MAC 地址流量平衡是根据数据帧的目的 MAC 地址把流量分配到聚合端口的各成员链路中。相同目的 MAC 地址的流量,从同一个成员链路转发,不同目的 MAC 地址的流量,将从不同的成员链路中转发。

源 MAC 地址+目的 MAC 地址流量平衡是根据数据帧的源 MAC 地址和目的 MAC 地址把流量分配到聚合端口的各成员链路中。具有不同的源 MAC 地址+目的 MAC 地址的数据帧可能被分配到同一个聚合端口的成员链路中。

源 IP 地址或目的 IP 地址流量平衡是根据数据报的源 IP 地址或目的 IP 地址进行流量分配。不同源 IP 地址或目的 IP 地址的流量通过不同的成员链路转发,相同源 IP 地址或目的 IP 地址的流量则通过相同的成员链路转发。这种流量平衡方式用于三层报文的转发,如果在此流量平衡模式下收到了二层数据帧,则自动根据二层数据帧的源 MAC 地址或目的 MAC 地址进行流量平衡。

源 IP 地址+目的 IP 地址流量平衡是根据数据报的源 IP 地址和目的 IP 地址进行流量分配。该流量平衡方式用于三层报文的转发,如果在此流量平衡模式下收到了二层数据帧,则自动根据二层数据帧的 MAC 地址进行流量平衡。具有不同的源 IP 地址+目的 IP 地址的报文可能被分配到同一个聚合端口的成员链路中。

5.1.3　链路聚合配置

1. 配置 AP 注意事项

(1) 物理端口默认情况下不属于任何 AP。

(2) AP 成员端口的端口速率必须一致。

(3) 二层端口只能加入二层 AP,三层端口只能加入三层 AP,即端口与 AP 属于同一层次。

(4) AP 不能设置端口安全功能。

(5) 当把端口加入一个不存在的 AP 时,该 AP 将被自动创建。

(6) 一个端口加入 AP,端口的属性将被 AP 的属性取代。

(7) 一个端口从 AP 中删除,端口的属性将恢复为其加入 AP 前的属性。

(8) 一个端口加入 AP 后,不能在该端口上进行任何配置,直到该端口退出 AP。

2. 配置二层 AP

配置二层 AP 有两种方式。

(1) 在全局配置模式下,命令格式为:

```
interface aggregateport n
```

其中,n 为 AP 号。先用此命令创建一个 AP(如果该 AP 不存在),然后在需要聚合的物理端口的端口配置模式下用 port-group 命令把该端口加入到 AP 中。

(2) 直接在接口配置模式下,命令格式为:

```
port-group port-group-number
```

其中,port-group-number 为 AP 的编号,即 n,也称 AP 号。此命令将该端口加入一个 AP(如果该 AP 不存在,则创建)。

配置举例:用第 1 种方式把千兆端口 gi 0/1-2 配置成二层 AP2 成员。

```
switch_jiaoxue#configure terminal
switch_jiaoxue(config)#interface aggregate 2
switch_jiaoxue(config-if)#exit
switch_jiaoxue(config)#interface range gi 0/1 - 2
switch_jiaoxue(config-if-range)#port-group 2
switch_jiaoxue(config-if-range)#end
```

配置举例:用第 2 种方式把 gi 0/1-2 配置成二层 AP2 成员,并将 AP2 配置成 Trunk 模式。

```
switch_jiaoxue#configure terminal
switch_jiaoxue(config)#interface range gi 0/1 - 2
switch_jiaoxue(config-if-range)#port-group 2
switch_jiaoxue(config-if-range)#exit
switch_jiaoxue(config)#interface aggregateport 2          !进入 AP2 端口配置
switch_jiaoxue(config-if)#switchport mode trunk           !配置 Trunk 端口
switch_jiaoxue(config-if)#end
```

在接口配置模式下使用 no port-group 命令将一个物理端口退出 AP。

3. 配置三层 AP

默认情况下，一个 AP 是二层的 AP，如果要配置成三层 AP，步骤如下。

(1) 用 interface aggregateport 命令进入 AP 端口配置模式。

(2) 将该 AP 端口设置为三层模式。

(3) 配置 AP 端口的 IP 地址。

(4) 进入要配置成 AP 的物理端口的配置模式。

(5) 将该物理端口设置为三层模式。

(6) 用 port-group 命令加入 AP。

(7) 用 no shutdown 命令激活该端口。

配置举例：在三层交换机上配置一个三层 AP(AP5)，将 fa 0/23、fa 0/24 加入 AP5，并配置 IP 地址 192.168.10.1/24。

```
switch_jiaoxue#configure terminal
switch_jiaoxue(config)#interface aggretegateport 5
switch_jiaoxue(config-if)#no switchport
switch_jiaoxue(config-if)#ip address 192.168.10.1 255.255.255.0
switch_jiaoxue(config-if)#no shutdown
switch_jiaoxue(config-if)#exit
switch_jiaoxue(config)#int range fa 0/23 - 24
switch_jiaoxue(config-if-range)#no switchport
switch_jiaoxue(config-if-range)#port-group 5
switch_jiaoxue(config-if-range)#no shutdown
switch_jiaoxue(config-if-range)#end
```

4. 配置 AP 的流量平衡

在全局配置模式下，命令格式为：

```
aggregateport load-balance {dst-mac|src-mac|src-dst-mac|dst-ip|src-ip|ip}
```

dst-mac，根据输入流量的目的 MAC 地址进行流量分配。在 AP 各链路中，目的 MAC 地址相同的流量被送到相同的成员链路，目的 MAC 地址不同的流量被分配到不同的成员链路。

src-mac，根据输入流量的源 MAC 地址进行流量分配。在 AP 各链路中，源 MAC 地址不同的流量分配到不同的成员链路，源 MAC 地址相同的流量使用相同的成员链路。

src-dst-mac，根据源 MAC 地址与目的 MAC 地址进行流量分配。不同源 MAC 地址＋目的 MAC 地址对的流量通过不同的成员链路转发，同一源 MAC 地址＋目的 MAC 地址对通过相同的成员链路转发。

dst-ip，根据输入流量的目的 IP 地址进行流量分配。在 AP 各链路中，目的 IP 地址相同的流量被送到相同的成员链路，目的 IP 地址不同的流量被分配到不同的成员链路。

src-ip，根据输入流量的源 IP 地址进行流量分配。在 AP 各链路中，来自不同 IP 地址的流量分配到不同的成员链路，来自相同 IP 地址的流量使用相同的成员链路。

ip，根据源 IP 地址与目的 IP 地址进行流量分配。不同的源 IP 地址＋目的 IP 地址对的

流量通过不同的成员链路转发，同一源 IP 地址＋目的 IP 地址对通过相同的成员链路转发。

将 AP 的流量平衡设置恢复到默认值，可以在全局配置模式下使用命令：

```
no aggregateport load-balance
```

5. 显示 AP 配置信息

在特权模式下，命令格式为：

```
show aggregateport {[port-number] summary |load-balance}
```

其中，port-number 为 AP 号；load-balance 显示 AP 的流量平衡算法；summary 显示 AP 中的每条链路的摘要信息。

5.2　链路冗余

5.2.1　网络中的冗余链路

主机 A 和主机 B 之间进行通信，如图 5-2 所示。如果两台交换机由单链路连接，那么传输介质出现故障将导致主机 A、B 通信的中断。为了解决单链路故障采取了双链路连接的方案。如果两条链路同时连到交换机的两个端口而不采取其他措施的话，两台交换机的四个端口会形成环路，产生广播风暴，影响网络通信。

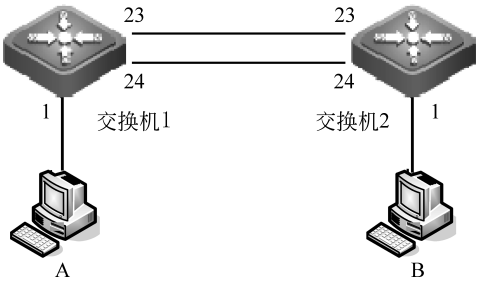

图 5-2　冗余链路

环路产生过程如下。

（1）主机 A 向主机 B 发送信息。

（2）交换机 1 从端口 1 收到数据帧。

（3）如果是广播或组播地址，则向除接收端口 1 之外的所有其他端口转发该数据帧；如果是单播地址，但是这个地址并不在交换机的 MAC 地址表中，那么也向除接收端口 1 之外的所有其他端口转发（泛洪）。

（4）数据帧将同时从交换机 1 的 23、24 端口被转发到交换机 2 的 23、24 端口。

（5）交换机 2 接收到数据帧。

（6）从端口 23 接收到的数据帧，如果是广播或组播地址，则向除接收端口 23 之外的所有其他端口转发该数据帧，该数据帧将被从端口 24 转发回交换机 1；如果是单播地址，但是这个地址并不在交换机的 MAC 地址表中，那么也向除接收端口 23 之外的所有其他端口转发（泛洪），该数据帧同样将被从端口 24 转发回交换机 1。

(7) 从端口 24 接收到的数据帧,如果是广播或组播地址,则向除接收端口 24 之外的所有其他端口转发该数据帧,该数据帧将被从端口 23 转发回交换机 1;如果是单播地址,但是这个地址并不在交换机的 MAC 地址表中,那么也向除接收端口 24 之外的所有其他端口转发(泛洪),该数据帧同样将被从端口 23 转发回交换机 1。

(8) 交换机 1 从端口 23、24 接收到同样数据帧,继续转发处理。

(9) 循环转发该数据帧,形成环路。

5.2.2 生成树协议

生成树协议(Spanning-Tree Protocol,STP)通过生成树算法在一个具有冗余链路的网络中构建一个没有环路的树形逻辑拓扑结构,既提供了链路的冗余连接,增强了网络的可靠性,又避免了数据在环路上的连续转发,消除了广播风暴。

STP 是用来避免链路环路产生广播风暴并提供链路冗余备份的协议。对二层以太网来说,两个设备间只能有一条激活的通道,否则就会产生广播风暴。但是为了增强网络的可靠性,建立冗余链路又是必要的,冗余链路中的一些链路处于激活状态,另一些链路处于备份状态,如果链路发生故障,激活的链路失效时,备份状态的链路必须变为激活状态。

STP 能够自动地完成主、备链路的切换并做到:选择并生成局域网的一个最佳树形拓扑结构、发现故障并进行恢复、自动更新拓扑结构、保证任何时候都选择可能的最佳树形拓扑结构。

局域网的拓扑结构是根据预先设置的配置参数自动计算的。如果参数配置得当,能够生成最佳的树形拓扑结构。

链路聚合技术和生成树协议并不冲突,生成树协议会把链路聚合后的链路当作单个逻辑链路进行生成树的建立,在图 5-1 中的 4 条链路聚合后,就产生了一个端口通道 Port-Channel,这个端口通道在生成树协议的工作中,是作为单链路进行计算的。

生成树协议是一个广义的概念,它包括 STP 以及基于 STP 改进的快速生成树协议(Rapid Spanning-Tree Protocol,RSTP)、多生成树协议(Multiple Spanning-Tree Protocol,MSTP)等,按照改进的情况,把生成树协议的发展分成三代。

第一代:STP/RSTP。

第二代:PVST/PVST+。

第三代:MSTP。

其中,第二代生成树协议 PVST/PVST+(Per VLAN Spanning-Tree)是 Cisco 提出的私有协议,它基于每个 VLAN 生成一个树形逻辑拓扑,保证每个 VLAN 都不存在环路,PVST 不兼容 STP/RSTP。

1. STP

狭义的 STP 是指 IEEE 802.1d 标准。

1) BPDU 帧

交换机之间通过交换网桥协议数据单元(Bridge Protocol Data Units,BPDU)帧获得建立拓扑结构需要的信息,BPDU 帧格式如图 5-3 所示。

很显然,这是一个 IEEE 802.3 SAP 帧。帧头部分包括 6B 的目的地址、6B 的源地址、2B 的帧类型及 3B 的 LLC 首部。数据部分是 35B 的 BPDU 域(如表 5-1 所示),以及为了补

图 5-3　BPDU 帧格式

齐 64B 最小帧采用的 8B 填充，在 VLAN 环境中，BPDU 帧被封装在 IEEE 802.1q 头部之后。

表 5-1　STP BPDU 域

序号	字　　段	长度	含　　义
1	Protocol ID	2B	0
2	Version	1B	0
3	BPDU Type	1B	Configuration BPDU 帧是 0x00；TCN BPDU 帧是 0x80
4	Flags	1B	最低位＝TC(Topology Change，拓扑改变)标志 0，表示拓扑没有改变；1，表示拓扑改变 最高位＝TCA(Topology Change Acknowledgment，拓扑改变确认)标志 0，表示非拓扑改变确认帧；1，表示拓扑改变确认帧 中间 6 位未使用
5	Root Bridge ID	8B	本交换机所认为的根交换机的 ID
6	Root Path Cost	4B	本交换机到根交换机的路径花费
7	Bridge ID	8B	发送交换机的 ID，由交换机优先级和 MAC 地址组成
8	Port ID	2B	发送 BPDU 端口的 ID，由端口优先级和端口号组成
9	Message Age	2B	本报文的已存活时间
10	Max-Age Time	2B	保存 BPDU 的最长时间，默认 20s
11	Hello Time	2B	定时发送 BPDU 帧的时间间隔，默认 2s
12	Forward-Delay Time	2B	BPDU 全网传输延迟时长，默认 15s

BPDU 帧以组播地址 01-80-C2-00-00-00 为目的地址进行传播。

BPDU 帧有两种类型：Configuration BPDU 和 TCN BPDU。

配置 BPDU(Configuration BPDU)帧由根交换机从指定端口周期性地发送，包括 Root Bridge ID、Bridge ID、Root Path Cost 等参数，非根交换机从根端口收到帧后修改自身参数并转发。

拓扑变更通知(Topology Change Notification，TCN)，交换机检测到拓扑变更后，向根交换机的方向发送 TCN BPDU 帧，通知拓扑发生变更。

2) STP 工作原理

STP 的基本思想是在交换机之间传递 Configuration BPDU 帧，比较其中的参数，使每个端口保存着最佳 BPDU 帧。当交换机初始启动 STP 时，所有端口每隔 2s 发送一次 BPDU，当交换机的一个端口收到高优先级的 BPDU(更小的 Bridge ID，更小的 Root Path

Cost 等)时,在该端口保存这些信息,同时向其他端口更新并传播这些信息。如果收到比自己低优先级的 BPDU,交换机就丢弃该信息。这样的机制确保高优先级的信息能够在整个网络中传播,从而根据 STP 算法阻塞存在的冗余链路,建立一个无循环的逻辑树形拓扑结构。工作步骤如下。

(1) 选举一台交换机为根交换机(Root Bridge,RB),选举原则如下。

① 所有交换机首先认为自己是 RB,互相发送 Configuration BPDU 帧。

② 选举 Bridge ID 最小的交换机为 RB,Bridge ID=交换机优先级+Mac 地址(默认优先级为 32 768)。

③ 每个网络中只能有一个 RB。

④ 其他交换机均为非根交换机。

(2) 选举根端口(Root Port,RP),选举原则如下。

① RP 处于非根交换机上。

② 每个非根交换机上有且只能有一个 RP。

③ RP 是非根交换机距离 RB 最近的端口,即到 RB 路径开销最小的端口。

④ 非根交换机通过 RP 接收 BPDU。

(3) 确定指定端口(Designated Port,DP),原则如下。

① RP 不参与竞争 DP。

② 根交换机上的端口都是 DP。

③ 每个网段(冲突域)都会选择一个路径开销最小的端口连接到 RB,该端口为 DP。

(4) 路径开销计算。

路径开销如表 5-2 所示,速度越快,开销越小,相同速率的聚合端口,成员越多,开销越小。

表 5-2 路径开销表

带 宽	IEEE 802.1d	IEEE 802.1t
10Mb/s	100	2 000 000
100Mb/s	19	200 000
1000Mb/s	4	20 000
10Gb/s	2	2000

如路径开销相同,依次比较 Sender's Bridge ID、Sender's Port ID、本交换机的 Port ID,选取高优先级(数值更小)的端口。

Port ID 由端口优先级和端口号组成(默认端口优先级为 128)。

(5) RP 和 DP 进入 Forwarding 状态。

(6) 其他端口设为 Blocking 状态。

3) 端口的角色和状态

STP 从启动到稳定运行过程中,交换机端口经历了不同的状态和角色,每个端口都在网络中扮演一个角色,用来体现在网络拓扑中的不同作用。

(1) RP,提供最短路径到根交换机的端口。

(2) DP,每个 LAN 通过该端口连接到根交换机。

每个端口有五种状态(Port State)表示是否转发数据包,通过这五种状态控制整个生成树拓扑结构。

(1) Disabled 状态禁用端口。

(2) Blocking 状态阻塞端口,也是端口启用的初始状态,此状态接收 BPDU、不学习源 MAC 地址、不转发数据帧。

(3) Listening 状态接收和发送 BPDU、不转发数据帧、不学习源 MAC 地址,但交换机向其他交换机通告该端口,参与选举根端口或指定端口。

(4) Learning 状态接收和发送 BPDU、不转发数据帧、学习源 MAC 地址。

(5) Forwarding 状态接收和发送 BPDU、正常转发数据帧、学习源 MAC 地址。

4) 网络拓扑变更

BPDU 是由根交换机发送的,当网络拓扑结构发生变更,出现以下几种情况时,交换机发送 TCN BPDU。

(1) 处于转发状态或监听状态的端口,变为阻塞状态。

(2) 处于未启用状态的端口进入转发状态,并且交换机上有其他的转发端口。

(3) 交换机从指定端口接收到 TCN BPDU。

(4) 端口 Up 或 Down 状态的转换导致交换机发 TCN BPDU。

(5) TCN BPDU 发送到根交换机。

TCN BPDU 发送过程如下。

(1) 当网络拓扑发生变化时,交换机会从自己的根端口向外发送 TCN BPDU。

(2) 接收到 TCN BPDU 的交换机向发送者发送 TCA 报文,表示对接收 TCN 的确认。

(3) 根交换机接收到 TCN BPDU 后向网络中其他交换机发送 TC BPDU,表示拓扑发生变化。

(4) 收到 TC BPDU 的交换机将 MAC 地址表老化时间设为 15s(默认是 300s)。

5) STP 定时器

STP 有 3 个定时器,分别如下。

(1) Hello Time,根交换机发送 BPDU 报文的时间间隔就是 Hello Time,默认是 2s。

(2) Max-Age,如果交换机发现某个根端口一段时间都没有收到 BPDU 则认为网络中拓扑发生变化,则向根交换机发送 TCN BPDU,这段时间就是最大生存时间,默认为 20s。

(3) Forward-Delay Time,转发延迟时间,是端口停留在监听状态和学习状态的时间,默认为 15s。

从图 5-4 中可以看出,STP 的计时器作用在不同的阶段后,导致交换机的端口收敛速度变慢。

6) STP 收敛过程分析

分析如图 5-5 所示拓扑,说明 STP 如何把具有环路的网络拓扑生成一个树形结构。

收敛过程分析如下。

(1) 选举根交换机。

比较四个交换机的 Bridge ID,四个交换机的优先级都是 32 768,优先级相等,再比较交换机的 MAC 地址,Switch 1 的 MAC 地址最小,所以 Switch 1 的 Bridge ID 最小,根交换机

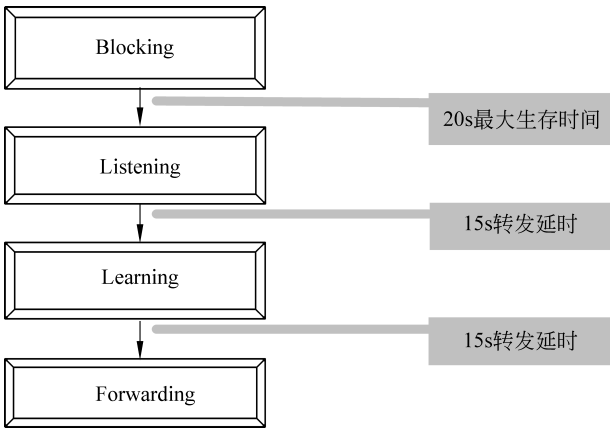

图 5-4　计时器的作用点

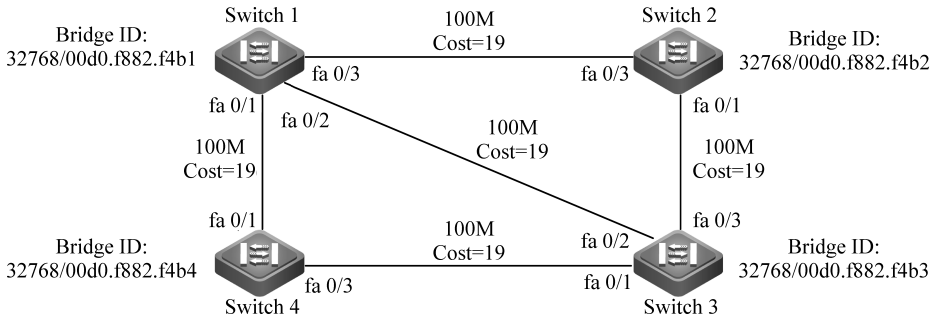

图 5-5　网络拓扑

是 Switch 1，Switch 2、Switch 3 和 Switch 4 是非根交换机。

（2）选举根端口。

根端口在非根交换机上，所以考虑 Switch 2、Switch 3、Switch 4。

Switch 2 端口到根交换机的路径开销：fa 0/3 是直连，Root Path Cost＝19，fa 0/1 通过 Switch 3 连接根交换机，Root Path Cost＝19＋19，fa 0/3 是 Switch 2 的根端口。

Switch 3 端口到根交换机的路径开销：fa 0/2 是直连，Root Path Cost＝19，fa 0/1 通过 Switch 4 连接根交换机，Root Path Cost＝19＋19，fa 0/3 通过 Switch 2 连接根交换机，Root Path Cost＝19＋19，所以 fa 0/2 是 Switch 3 的根端口。

Switch 4 端口到根交换机的路径开销：fa 0/1 是直连，Root Path Cost＝19，fa 0/3 通过 Switch 3 连接根交换机，Root Path Cost＝19＋19，所以 fa 0/1 是 Switch 4 的根端口。

（3）选举指定端口。

在 Switch 1 到 Switch 2 的网段上：Switch 1 的 fa 0/3 是根交换机本身端口，路径开销是 0，Switch 2 的 fa 0/3 到根交换机的路径开销是 19，所以 Switch 1 到 Switch 2 的物理网段上，Switch 1 的 fa 0/3 是指定端口（根交换机的端口都是 DP）；

在 Switch 1 到 Switch 3 的网段上：Switch 1 的 fa 0/2 是根交换机本身端口，路径开销是 0，Switch 3 的 fa 0/2 到根交换机的路径开销是 19，所以 Switch 1 到 Switch 3 的物理网段上，Switch 1 的 fa 0/3 是指定端口。