

STP 生成树协议理解与实际配置

目 录

目录.....	1
前言.....	3
1 生成树综述.....	4
1.1 基本概念.....	5
1.1.1 Bridge ID.....	5
1.1.2 Port ID.....	5
1.1.3 Path Cost—路径值.....	5
1.1.4 BPDU.....	6
1.2 几种 STP 标准参数.....	7
1.3 STP 的具体工作机制.....	8
1.4 STP 的收敛时间.....	9
1.4.1 端口的状态：.....	9
1.4.2 STP 的收敛时间具体计算.....	10
2 RSTP.....	10
2.1 RSTP 背景.....	10
2.2 新的端口状态.....	11
2.3 端口的角色.....	11
2.4 新的 BPDU 格式.....	12
2.5 新的收敛机制.....	13
2.5.1 keep-alive mechanism—检测失效的链路.....	13

2.5.2 同步机制—SYNC---检测新加入的链路	13
2.5.3 接受低级的 BPDU:.....	14
2.5.4 RSTP 交换机的端口类型	14
2.6 RSTP 向下兼容 STP	14
2.7 RSTP 优点	15
3 PVST+/RPVST+.....	15
4 MST.....	16
4.1 概述	16
4.2 基本概念	17
4.2.1 Region(域)	17
4.2.2 msti(MST 实例)	17
4.2.3 VLAN Map	17
4.2.4 ist.....	17
4.2.5 cst.....	18
4.2.6 cist.....	18
4.2.7 域根	19
4.2.8 总根	20
4.2.9 master 端口角色.....	20
5 生成树练习	21
6 总结.....	21

前言

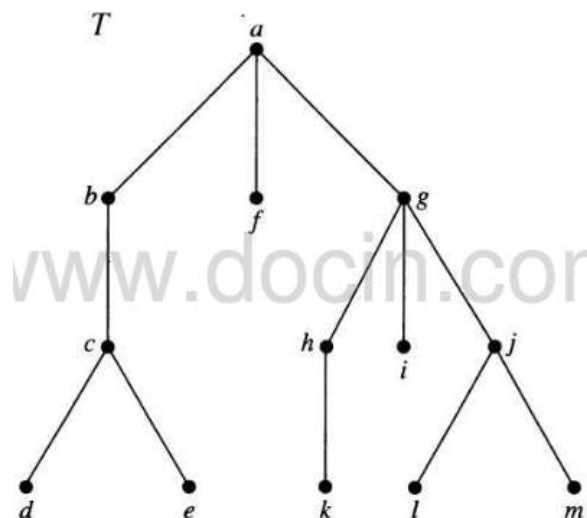
生成树是一颗美丽的树，一个美丽的树精灵。在当今传统的以太网网络中，生成树虽然看不见，但却无处不在，她蔓延在网络每个角落，从一个校园网、企业网、数据中心网初建的那一天起，就开始滋生和繁衍。生成树如同一个建筑物的支承框架，使之浑然一体矗立稳固；如同一个生命体内在的筋络，使之血脉畅通生生不息；更如同一个充满传奇和法力的魔女，主宰着网络世界。

在已有的网络中，生成树静静地生长、蛰伏着，当新的交换机出现时，如何让已有的生成树伸展、生成新的有机枝条，如何把异构的枝条嫁接到原有的树体，使之成活而又不危机到原有树的机体，如此使网络得以有效扩展？网络工程师如同一个园艺师，对网络的设计和实施，就是赋予这棵生成树以生命，让她保持原有基因健康地生长，即达到我们现实目的，避免二层生成树兼容性问题，使网络运行正常。



1 生成树综述

生成树是个数学概念。首先在一个无向连通图的基础上，可以生成多个无环路的树（Tree），即生成树（Spanning Tree），在生成树的基础上，再求出唯一最小生成树（Minimum Spanning Tree）。



以太网生成树协议 STP（Spanning Tree Protocol）即 IEEE 802.1D 如同数学领域中的生成树，其作用主要有三个，第一是 **eliminate logical loops** 消除逻辑环（即生成树状图）；第二自动选取最有效的网络路径（即生成最小生成树）；第三是当某条链路失效时，自动切换到备份链路，实现所谓的 **failover** 功能。

STP 分 STP(802.1D)、RSTP(802.1W)、MST(802.1S)、PVST/PVST+、RPVST/RPVST+，通过以上这些生成树协议，网络会自动计算出一棵或多棵最小生成树实例，即实际数据传输的路线图。

1.1 基本概念

1.1.1 Bridge ID

一个 Bridge ID(BID)在 STP 数据包—BPDU 中占 8 个字节，其中两个字节为 Bridge Priority(0--65535)，6 个字节为桥的 MAC 地址，缺省的桥的优先级值为 32768。

Bridge ID=优先级+MAC

1.1.2 Port ID

由两部分组成，在 BPDU 中共占用两个字节，高位字节为优先级，低位字节为端口号。优先级值缺省为 128。

Port ID=优先级+端口 ID

1.1.3 Path Cost—路径值

每条路径都有一个 cost 值，是按本地交换机端口的进入方向计算的。用 cost 值来计算两个设备之间的“距离”，计算原则是其值与带宽成反比，即高的带宽有低的 cost 值。但值得注意的是每个厂家的设备所计算的标准不同，如一般情况下 Cisco 的 100M 带宽下 cost 值是 19，1G 带宽下是 4，10G 带宽下是 2，而 IBM 交换机是按照 802.1d 2004 的标准，100M 带宽 cost 是 200000，1G 带宽 cost 是 20000，万兆带宽 cost 是 2000。

注：不同的 cost 值规范不会影响 STP 的正确工作。

1.1.4 BPDU

BPDU（Bridge Protocol Data Unit），它是 STP 赖以工作、在交换机之间进行交互的二层数据帧，交换机所定义的参数、网络中拓扑的改变都通过 BPDU 进行传递。BPDU 分类如下：

1.1.4.1 Configuration BPDU

Configuration BPDU 即是我们常说的 BPDU，绝大部分的 BPDU 是此类型 BPDU，它由根桥（Root Bridge）产生，并向下传播，其它非根桥（Designation Bridge）不产生此类 BPDU（但所有设备在初启时都会产生此类型的 BPDU），只是收到由根桥发来的 BPDU 后加入自己的一些信息（即更新）并转发，如桥及端口 ID 值，以下是一个由根桥（Bladenet_8a:35:2c）发出的 Configuration BPDU。

```
[-] IEEE 802.3 Ethernet
  [+] Destination: Spanning-tree-(for-bridges)_00 (01:80:c2:00:00:00)
  [+] Source: Bladenet_8a:35:2c (00:18:b1:8a:35:2c)
  Length: 39
  Trailer: 0000000000000000
  [+] Logical-Link Control
  [-] Spanning Tree Protocol
    Protocol Identifier: Spanning Tree Protocol (0x0000)
    Protocol Version Identifier: Rapid Spanning Tree (2)
    BPDU Type: Rapid/Multiple Spanning Tree (0x02)
    [+] BPDU flags: 0x1e (Learning, Port Role: Designated, Proposal)
    Root Identifier: 32768 / 00:18:b1:8a:35:00
    Root Path Cost: 0
    Bridge Identifier: 32768 / 00:18:b1:8a:35:00
    Port identifier: 0x802c
    Message Age: 0
    Max Age: 20
    Hello Time: 2
    Forward Delay: 15
    Version 1 Length: 0
```

1.1.4.2 Topology Change Notification BPDU

Topology Change Notification BPDU，是当网络拓扑发生改变时，

由非根桥向上通知根桥，当根桥收到后，指示其它所有非根桥设备加速更新 CAM 表（MA 地址表），缺省情况下 CAM 的 ageing time 是 5 分钟，而在 TCN BPDU 的协调下，ageing time 缩短为 15 秒。以下是当非根桥（Bladenet_ad:6d:28）发生网络拓扑变化时，由其发出的 TCN BPDU，即在 flags 中的 TCN 位上进行了置位。

```

[-] IEEE 802.3 Ethernet
  [+] Destination: Spanning-tree-(for-bridges)_00 (01:80:c2:00:00:00)
  [+] Source: Bladenet_ad:6d:28 (00:18:b1:ad:6d:28)
      Length: 39
      Trailer: 0000000000000000
[-] Logical-Link Control
[-] Spanning Tree Protocol
  Protocol Identifier: Spanning Tree Protocol (0x0000)
  Protocol Version Identifier: Rapid Spanning Tree (2)
  BPDU Type: Rapid/Multiple Spanning Tree (0x02)
  [+] BPDU flags: 0x3d (Forwarding, Learning, Port Role: Designated, Topology Change)
    Root Identifier: 32768 / 00:18:b1:8a:35:00
    Root Path Cost: 20000
    Bridge Identifier: 32768 / 00:18:b1:ad:6d:00
    Port identifier: 0x8028
    Message Age: 1
    Max Age: 20
    Hello Time: 2
    Forward Delay: 15
    Version 1 Length: 0
```

1.2 几种 STP 标准参数

- Hello time: 2 秒
- Max age: 20 秒,
- Forward Delay : 15 秒

这些参数都可进行修改，但只有对根桥的修改才起作用，非根桥上的这些 timer 参数只在其产生 TCN BPDU 时才起作用。

1.3 STP 的具体工作机制

选路去环路的总算法或规则:

Step1:

每个 STP 组 (STG) 或所谓实例 (Instance) 中要选出一个根桥 (root), 最低桥 ID 胜出;

Step2:

每个非根桥要选出一个根端口 (root port), 以 cost 值为标准, 计算非根端口到根桥的短路径值, 最小者胜出;

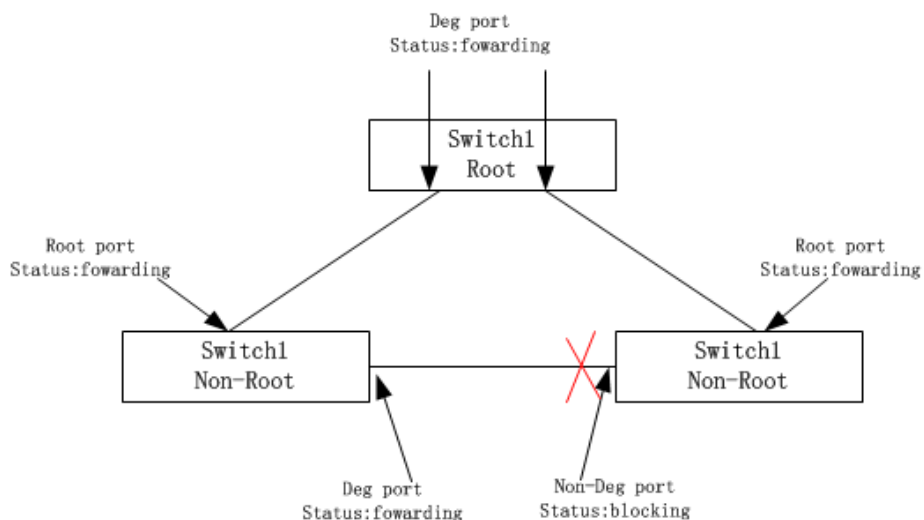
Step3:

每条链路 (segment) 上要有一个指派端口 (designated port), 以 cost 值为标准, 计算链路上的端口到根桥的短路径值, 最小者胜出;

由于有以上三个原则, 那么 STP 就可以生成没有环路逻辑拓扑。

注: 如果在第二、三步中出现相同时, 最低发送桥 ID 和最低端口 ID 胜出。

以下是按照上面的生成树算法生成最小生成树的示意图:



1.4 STP 的收敛时间

以上 STP 的工作机制及相应的算法决定了 STP 的收敛时间。

1.4.1 端口的状态：

STP 的端口状态共分为 5 种：disable ,listening, learning, forwarding, blocking。

- disable 状态

当一个端口物理上失去了连结，或人为地进行了 shutdown 后，该端口即进入了 disable 状态

- listening 状态

当一个端口由 disable 状态激活后，立即进入 listening 状态，在此阶段，BPDU 开始进行交换，进行一系列的 STP 算法，选举出根桥、根端口、指派端口。此状态持续一个所谓 forwarding time 15 秒

- learning 状态

如果以上的 listening 状态能够持续 15 秒，那么紧接着进行 learning 状态，在此状态进行 MAC 地址的学习，形成 CAM 表。该状态需持续 15 秒

- forwarding 状态

如果以上状态能够持续 15 秒，那么端口进入 forwarding 状态

- blocking 状态

一个 segment 网段上只有一个端口处于 forwarding 状态，其它端口被置为 blocking 状态。此端口不进行用户数据的转发，但接收 BPDU。

1.4.2 STP 的收敛时间具体计算

一个端口从 disable(或 blocking)状态到 forwarding 状态所经历的时间是：

如果一个交换机由于直接检测到失效的链路，当收不到根桥 BPDU 后，立即进入到 listening 状态，其总的收敛时间为：

$$15(\text{listening})+15(\text{learning})=30 \text{ (秒)}$$

如果某个设备不能直接检测到上游链路的失效，那么在 max aging time（20s）内收不到 BPDU 后，才能做出判断上游链路失效，那么，其收敛时间为：

$$20(\text{max aging time})+15(\text{listening})+15(\text{learning})=50 \text{ (秒)}$$

2 RSTP

2.1 RSTP 背景

STP 的收敛时间大约在 1 分钟以内，这远远满足不了现代网络对收敛速度的要求。比如，在一些应用中要启用 VRRP，VRRP 本身的切换速度只有几秒钟，如果 VRRP 与 STP 共同在起作用时，其收敛是建立在 STP 收敛完成的前提下，所以，VRRP 实际的收敛时间要加上 STP 的收敛时间，失去了其收敛速度快的优点。IEEE 推出了新一代的 STP 协议 RSTP，即 802.1w。下面，详细介绍 RSTP 的工作机制。

2.2 新的端口状态

STP (802.1D) Port State	RSTP (802.1w) Port State	Is Port Included in Active Topology?	Is Port Learning MAC Addresses?
Disabled	Discarding	No	No
Blocking	Discarding	No	No
Listening	Discarding	Yes	No
Learning	Learning	Yes	Yes
Forwarding	Forwarding	Yes	Yes

2.3 端口的角色

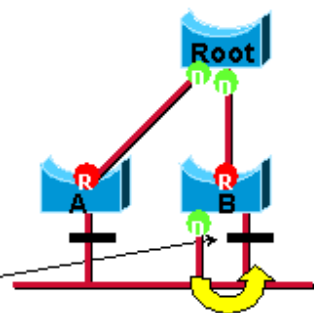
根端口（root port）及指派端口（designated port）仍然保留。但以前的 STP 中的 blocking 端口被分成新的两个角色----backup and alternate.

- alternate port



- backup port

— Backup Port



从以上两个示意图可以看到两种新类型端口的区别：

Backup port 从相同的交换机收到从根桥发来的 BPDUs，而 alternate port 是从不同的交换机。

2.4 新的 BPDUs 格式

RSTP 的 BPDUs 数据帧中的 type 及 version 是 2，在 STP 中是 0：

● STP BPDUs 格式

IEEE 802.3 Ethernet
Destination: Spanning-tree-(for-bridges)_00 (01:80:c2:00:00:00)
Source: AsanteTe_d0:b3:0f (00:00:94:d0:b3:0f)
Length: 38
Trailer: 0000000000000000
Logical-Link Control
Spanning Tree Protocol
Protocol Identifier: Spanning Tree Protocol (0x0000)
Protocol Version Identifier: Spanning Tree (0)
BPDUs Type: Configuration (0x00)
BPDUs flags: 0x00
Root Identifier: 28672 / 00:00:94:d0:b3:00
Root Path Cost: 0
Bridge Identifier: 28672 / 00:00:94:d0:b3:00
Port identifier: 0x8010

● RSTP BPDUs 格式

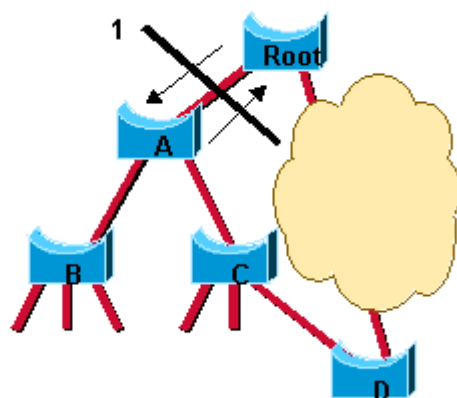
IEEE 802.3 Ethernet
Destination: Spanning-tree-(for-bridges)_00 (01:80:c2:00:00:00)
Source: Bladenet_8a:35:2a (00:18:b1:8a:35:2a)
Length: 39
Trailer: 0000000000000000
Logical-Link Control
Spanning Tree Protocol
Protocol Identifier: Spanning Tree Protocol (0x0000)
Protocol Version Identifier: Rapid Spanning Tree (2)
BPDUs Type: Rapid/Multiple Spanning Tree (0x02)
BPDUs flags: 0x0e (Port Role: Designated, Proposal)
Root Identifier: 28672 / 00:00:94:d0:b3:00
Root Path Cost: 200000
Bridge Identifier: 32768 / 00:18:b1:8a:35:00
Port identifier: 0x802a

2.5 新的收敛机制

2.5.1 keep-alive mechanism—检测失效的链路

在 STP 中，根桥每 2 秒钟发出一个 Hello 包，非根桥只有从根端口收到 BPDU 后，才产生相应的 BPDU 并通过指派端口向其它交换机发送。而在 RSTP 中，非根桥每 2 秒钟定时向其它交换机发送 BPDU，即使它没有从根端口收到 BPDU。这样在交换机之间就形成了一个所谓的 keep-alive mechanism：当某个端口连续三次没有收到邻居的 hello 包后，就认为与其邻居相连的链路失效了，立即更新它以前所获得的有关协议信息，而不用 STP 中等待 20 秒收不到 BPDU 再 age out，这些大大加快了收敛速度，即所谓的 fast aging。

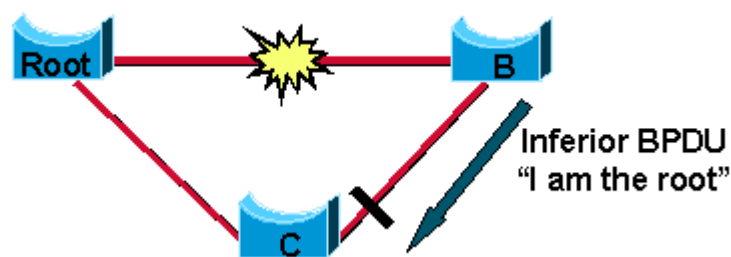
2.5.2 同步机制—SYNC---检测新加入的链路



交换机 A 与 Root 之间加入了一条新的链路，在 STP 处理中，两个相连的端口都被置为 listening 状态，而在 RSTP 中，先把两个相连的端口置为 discarding 状态，接着两个端口之间进行协商，当 A 接到 Root 发来的 BPDU 后，立即把端口置为 forwarding 状态，此过程称为

SYNC 同步，A 和 B/C 也同样进行同步，快速进行收敛。而在 STP 中端口要进行 listening-learning-forwarding 的过程，要在 30 秒后才能收敛，而 RSTP 是 bypass 了 listening 和 learning 这两个过程。

2.5.3 接受低级的 BPDU:



B 在链路好的时候，BPDU 是从连接 Root 的 root port 收到的，当链路失效后，C 收到 B 发来的 “I am the root” 后，立即发送一个有 Root 信息的 BPDU，B 收到后，相信，把连接 C 的端口作为 root port .

2.5.4 RSTP 交换机的端口类型

为了达到快速收敛的目的，RSTP 的端口类型规定为：

- 点对点端口：交换机之间相连
- 边缘端口：交换机与终端设备（服务器）相连

2.6 RSTP 向下兼容 STP

RSTP 保留着大部分的 STP 的参数和特性，最终形成的逻辑 loop-free 拓扑是完全相同的，只不过采用了一些快速的收敛机制，当与 STP 设备相连时，由于 RSTP 的 BPDU 帧中含有特殊的 type 及 version

值(都为 2),所以 STP 拒绝 RSTP 的 BPDU,且只是发自己的 STP BPDU,当 RSTP 设备收到 STP BPDU 后,自己主动降级,不再发 RSTP BPDU,而是发 STP BPDU,由此两种协议其实变成了一种协议----STP,也就是 RSTP 的优点全部消失,但只是在 RSTP 与 STP 设备之间,而 RSTP 设备与 RSTP 设备之间还仍然保持原来的状态即 RSTP。

2.7 RSTP 优点

RSTP 是一种比 STP 更先进的协议,它自身也自动地包含了一些如 Cisco 的加强的 STP 功能,比如 PortFast, UplinkFast, BackboneFast,其收敛速度大约为 1 到 5 秒,而不是 STP 的大约 1 分钟。同时,RSTP 与 STP 也很好地进行向下兼容,二者是 IEEE 成熟、标准的协议,用户可以不用任何人工干预达到向前兼容。现在所有的交换机都支持此种协议,建议起用此种 STP 模式。另外,最新版的 802.1d 2004 包括 802.1w。

3 PVST+/RPVST+

PVST+/PVRST+是 Cisco 的私有协议,每个 VLAN 一个 STP 实例。

IBM 交换机兼容 PVST+或 RPVST+。

[illegible]

在 OS 6.7 版本之前，IBM 交换机在配置 PVST+/PVRST+时，注意要手工把每个 VLAN 加入到不同的 STP Group 中，否则不能保证 VLAN 的正常通信，即便是 VLAN1。

4 MST

4.1 概述

- 公有协议，可若干个 VLAN 对应于一个 STP 实例。
- 缺省情况下，所有 VLAN 都在实例 0 即 CIST 里。
- MST 有域名，所谓的版本号，域中所有的交换机要一致。
- 如果要把 VLAN 匹配进实例，域中所有的交换机要一致。
- MST 是 Rapid STP，具有 RSTP 的属性。
- MST 兼容 STP 和 RSTP

4.2 基本概念

4.2.1 Region(域)

如果两台交换机在一个 MST 域中，那么这两台交换机应都启动了 mstp 协议、具有相同的域名、相同的 vlan 到生成树映射配置和相同的 mstp 修订级别配置，并且物理上有链路连通。

4.2.2 msti(MST 实例)

一颗域内生成树实例，实例之间相互独立。

4.2.3 VLAN Map

一个或多个 VLAN 映射到一颗生成树实例，会生成一个 mstp Digest 值(HASH 值)，只要 VLAN 的映射相同，那么这个值就相同，不管域名等其它参数是否相同。

```
G8000-2#sh span
-----
upfast disabled, update 40
Pvst+ compatibility mode enabled
Mstp Digest: 0x720702f16e90b71b6ebf85c7264acbe4
```

4.2.4 ist

是一个域内特殊的实例，即实例 0，不能删除，在没有手工进行 VLAN 映射前，所有的 VLAN 都在这个 ist 里，BPDU 只在这个实例中传递。

4.2.5 cst

是连接不同域之间的一颗由 STP/RSTP 生成的单生成树，用来连接不同的区域，对它来说，不同的区域就如同一个逻辑的交换机一样，以实现不同区域之间的通信。CST 所连接的域是广义上的域，这个域可是一个 MST 的域，也可以是一个普通 STP 或 RSTP 的单生成树域 (SST—Single Spanning Tree)

4.2.6 cist

由 ist 和 cst 构成 cist，是棵单生成树。

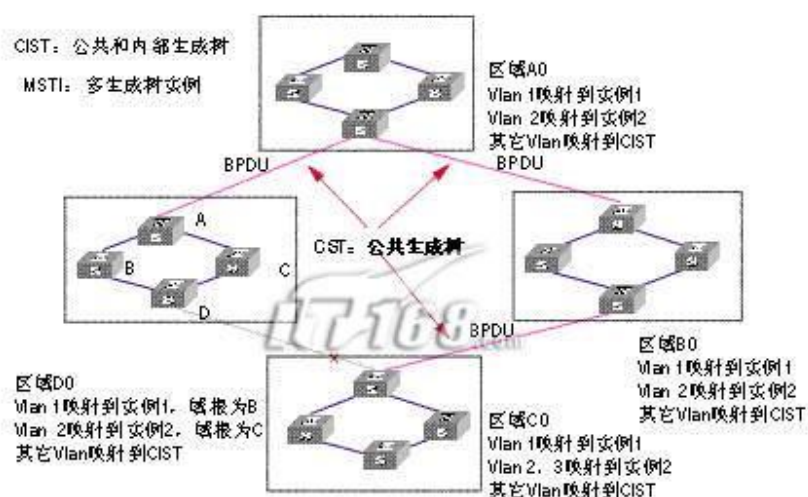


图1-1 MSTP的基本概念示意图

MST BPDU 只在 CIST 树上传第，而在 MST BPDU 中包含了 VLAN 的映射，所以不象 PVST 或 PVRST 那样，所有的 VLAN 都要发各自的 BPDU，由此可见 MST BPDU 是非常有效的。

4.2.7 域根

域中一个实例(包括 ist 和 msti)在域中的根。

Msti 域根

```
G8000-1(config)#sh span
-----

upfast disabled, update 40

Pvst+ compatibility mode enabled

Mstp Digest: 0x720702f16e90b71b6ebf85c7264acbe4
-----

Spanning Tree Group 1: On (MSTP)
VLANs MAPPED: 1 3
VLANs: 1 3

Current Root:          Path-Cost  Port
0000 fc:cf:62:42:c5:00    20000    4
```

域内 Cist（或确切说是 ist）的域根：

```
G8000-1#sh span mst cist info
-----

Mstp Digest: 0x682d678f64ea7985763c7b72307f048f

Common Internal Spanning Tree:

VLANs MAPPED: 1 4-4094
VLANs: 1

Current Root:          Path-Cost  Port MaxAge FwdDel
8000 fc:cf:62:41:79:00    0      0     20    15

Cist Regional Root:    Path-Cost
8000 fc:cf:62:41:79:00    0
```

4.2.8 总根

即 cist 的根，即在整个网络中（可能有多个域，包括非 MST 域），会有一颗单生成树 cist，它的根即总根

```
G8000-1#sh span mst cist info
-----
Mstp Digest: 0x682d678f64ea7985763c7b72307f048f

Common Internal Spanning Tree:

VLANs MAPPED: 1 4-4094
VLANs: 1

Current Root:          Path-Cost  Port  MaxAge  FwdDel
8000 fc:cf:62:41:79:00      0      0      20      15
```

4.2.9 master 端口角色

是连接一个域（一个逻辑交换机）到总根的端口，位于整个域到总根的最短路径上。

```
G8000-3#sh span
-----
upfast disabled, update 40

Pvst+ compatibility mode enabled

Mstp Digest: 0x720702f16e90b71b6ebf85c7264acbe4

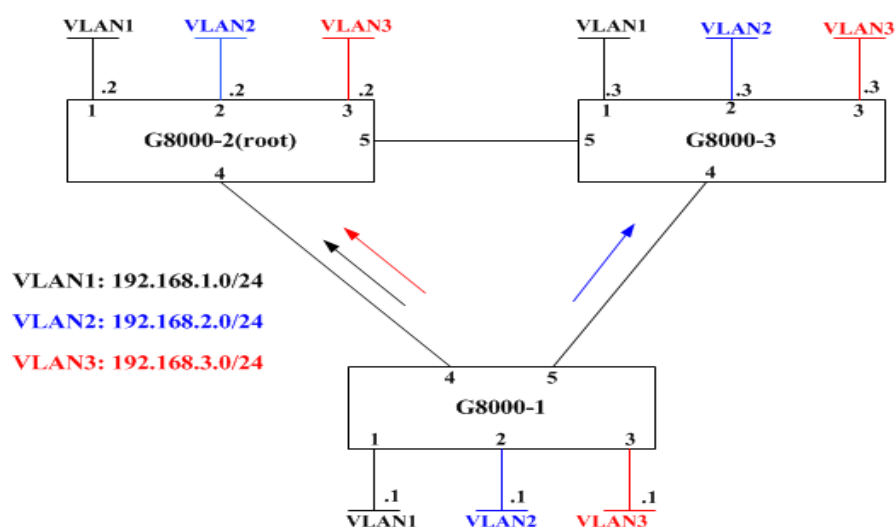
-----
Spanning Tree Group 1: On (MSTP)
VLANs MAPPED: 1 3
VLANs: 1 3

Current Root:          Path-Cost  Port
0000 fc:cf:62:42:c5:00      20000      5

Parameters:  Priority  Aging
              32768    300

Port      Prio    Cost      State  Role Designated Bridge
-----
4          128    20000!    FWD    MAST 8000-fc:cf:62:42:96:00
5          128    20000!    FWD    ROOT 0000-fc:cf:62:42:c5:00
```

5 生成树练习



实现要求：

让接入交换机 G8000-1 的奇数 VLAN 数据走左边链路，偶数 VLAN 的数据走右边链路，实现链路均衡负载；一旦有一条链路失效后，所以 VLAN 的数据走剩下可用的链路，实现链路冗余。

6 总结

生成树在传统以太网中被广泛使用，会一直伴随着传统的网络的存在而存在，是网络技术人员必须掌握的主要的二层网络技术，关于网络生成树协议的选择、局部网络是否开启、如何调优、如何实现不同协议的兼容等等，都是网络技术人员需要全面考虑的，否则会导致网络效率不高、局部网络不通、或导致整个网络瘫痪。生成树是个美丽的树精灵，她需要网络技术人员的关爱，由此才能充分发挥她的美丽和智慧，主宰着我们的网络世界。