IPUB_301_C1　生成树协议（spanning-tree protocol）

课程目标：

- 了解以太网的发展
- 了解交换机工作原理
- 掌握 STP 协议的工作原理及作用
- 掌握 STP 协议中端口的各种状态

参考资料：

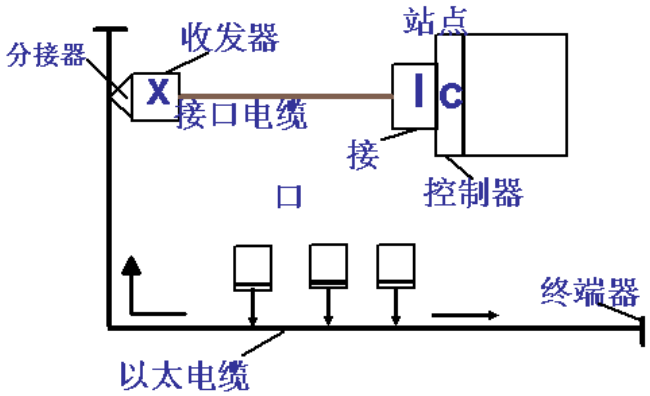
- 《ZXR10 以太网交换机用户手册》

第1章 以太网交换机原理

📖 知识点

- 了解网络发展。
- 了解以太网交换机原理。

1.1 以太网发展历史及现状



以太网是在 70 年代由 Xerox 公司 Palo Alto 研究中心推出的。由于介质技术的发展，Xerox 可以将许多机器相互连接，形成巨型打印机，这就是以太网的原型。后来，Xerox 公司推出了带宽为 2Mb/s 的以太网，又和 Intel 和 DEC 公司合作推出了带宽为 10Mb/s 的以太网，这就是通常所称的以太网 Ⅱ 或以太网 DIX（Digital，Intel 和 Xerox）。IEEE（电器和电子工程师协会）下属的 802 协委员会制定了一系列局域网标准，其中以太网标准（IEEE 802.3）与由 Intel、Digital 和 Xerox 推出的以太网 Ⅱ 非常相似。

随着以太网技术的不断进步与带宽的提升，目前在很多情况下以太网成为了局域网的代名词。

1.2 以太网相关标准

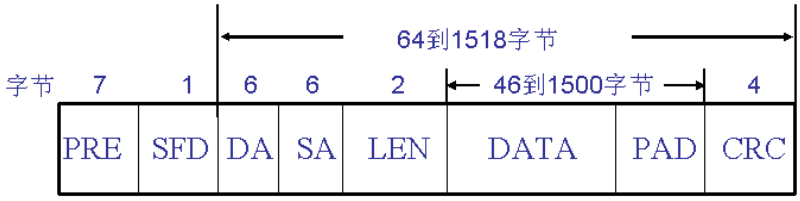
电器和电子工程师协会（IEEE）在 1980 年 2 月组成了一个 802 委员会制定了一系列局域网方面的标准，802.3 协议簇制定了以太网的标准。

其中：

- IEEE 802.3 为以太网标准。
- IEEE 802.2 为 LLC（逻辑链路控制）标准。
- IEEE 802.3u 为 100M 以太网标准。
- IEEE 802.3z 为 1000M 以太网标准。
- IEEE 802.3ab 为 1000M 以太网运行在双绞线上的标准。

通常我们所说的以太网主要是指以下三种不同的局域网技术：
以太网/IEEE 802.3—采用同轴电缆作为网络媒体，传输速率达到 10Mbps；
100Mbps 以太网—又称为快速以太网，采用双绞线作为网络媒体，传输速率达到 100Mbps；
1000Mbps 以太网—又称为千兆以太网，采用光缆或双绞线作为网络媒体，传输速率达到 1000Mbps（1Gbps）
以太网以其高度灵活，相对简单，易于实现的特点，成为当今最重要的一种局域网建网技术。虽然其它网络技术也曾经被认为可以取代以太网的地位，但是绝大多数的网络管理人员仍然把将以太网作为首选的网络解决方案。为了使以太网更加完善，解决所面临的各种问题和局限，一些业界主导厂商和标准制定组织不断的对以太网规范做出修订和改进。也许，有的人会认为以太网的扩展性能相对较差，但是以太网所采用的传输机制仍然是目前网络数据传输的重要基础。

1.3 以太网帧结构



前导(Preamble)：一个交替由 0 和 1 组成的 7 个 8 位位组(octet)模式被用作同步。

帧定界符开始(Start of Frame Delimiter)：特殊模式 10101011 表示帧的开始。

目的地址(Destination Address)：若第一位是 0，这个字段指定了一个特定站点。若是 1，该目的地址是一组地址，帧被发送往由该地址规定的预先定义的一组地址中的所有站点。每个站点的接口知道它自己的组地址，当它见到这个组地址时会做出响应。若所有的位均为 1，该帧将被广播至所有的站点。

源地址(Source Address)：说明一个帧来自哪儿。

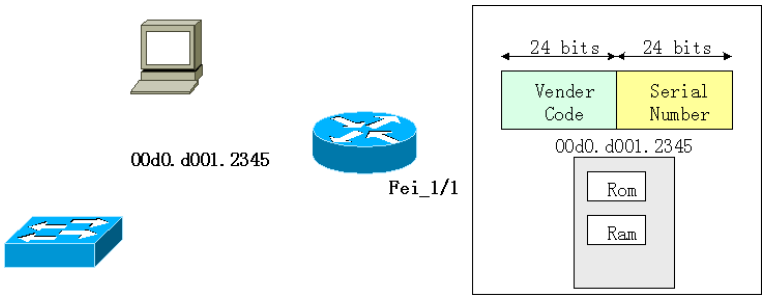
数据长度字段(Data Length Field)：说明在数据和填充字段里的 8 位字节的数目。

数据字段(Data Field)：上层数据。

填充字段(Pad Field)：数据字段必须至少是 46 个 8 位字节(或许更多)。若没有足够的数据，额外的 8 位位组被添加(填充)到数据中以补足差额。

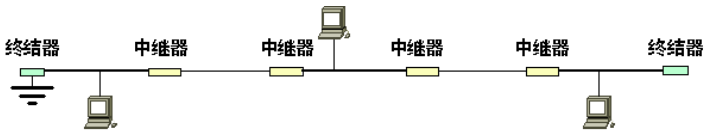
帧校验序列(Frame Check Sequence)：使用 32 位循环冗余校验码的错误检验。

1.4 MAC 地址



MAC 地址有 48 位，它可以转换成 12 位的十六进制数，这个数分成三组，每组有四个数字，中间以点分开。MAC 地址有时也称为点分十六进制数。它一般烧入 NIC（网络接口控制器）中。为了确保 MAC 地址的唯一性，IEEE 对这些地址进行管理。每个地址由两部分组成，分别是供应商代码和序列号。供应商代码代表 NIC 制造商的名称，它占用 MAC 的前六位 12 进制数字，即 24 位二进制数字。序列号由设备供应商管理，它占用剩余的 6 位地址，即最后的 24 位二进制数字。如果供设备应商用完了所有的序列号，他必须申请另外的供应商代码。目前 ZTE 的 GAR 产品 MAC 地址前六位为 00d0d0。

1.5 传统以太网基本概念



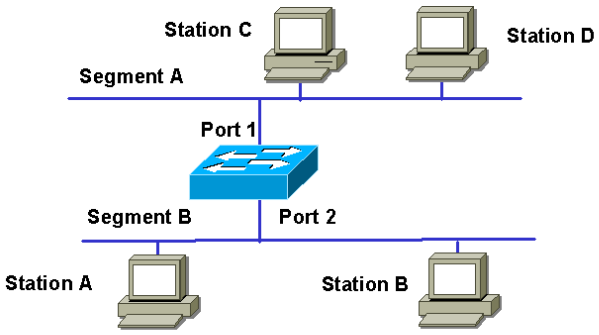
CSMA/CD

以太网使用 CSMA/CD（Carrier Sense Multiple Access with Collision Detection，带有冲突监测的载波侦听多址访问）。我们可以将 CSMA/CD 比做一种文雅的交谈。在这种交谈方式中，如果有人想阐述观点，他应该先听听是否有其他人在说话（即载波侦听），如果这时有人在说话，他应该耐心地等待，直到对方结束说话，然后他才可以开始发表意见。有一种情况，有可能两个人在同一时间都想开始说话，那会出现什么样的情况呢？显然，如果两个人同时说话，这时很难辨别出每个人都在说什么。但是，在文雅的交谈方式中，当两个人同时开始说话时，双方都会发现他们在同一时间开始讲话（即冲突检测），这时说话立即终止，随机地过了一段时间后，说话才开始。说话时，由第一个开始说话的人来对交谈进行控制，而第二个开始说话的人将不得不等待，直到第一个人说完，然后他才能开始说话。

以太网的工作方式与上面的方式相同。首先，以太网网段上需要进行数据传送的节点对导线进行监听，这个过程称为 CSMA/CD 的载波侦听。如果，这时有另外的节点正在传送数据，监听节点将不得不等待，直到传送节点的传送任务结束。如果某时恰好有两个工作站同时准备传送数据，以太网网段将发出“冲突”信号。这时，节点上所有的工作站都将检测到冲突信号，因为，这时导线上的电压超出了标准电压。冲突产生后，这两个节点都将立即发出拥塞信号，以确保每个工作站都检测到这时以太网上已产生冲突，然后，网络进行恢复，在恢复的过程中，导线上将不传送数据。当两个节点将拥塞信号传送完，并过了一段随机时间后，这两个节点便开始启动随机计时器。第一个随机计时器到期的工作站将首先对导线进行监听，当它监听到没有任何信息在传输时，便开始传输数据。当第二个工作站随机计时器到期后，也对导线进行监听，当监听到第一个工作站已经开始传输数据后，就只好等待了。

在 CSMA/CD 方式下，在一个时间段，只有一个节点能够在导线上传送数据。如果其他节点想传送数据，必须等到正在传输的节点的数据传送结束后才能开始传输数据。以太网之所以称作共享介质就是因为节点共享同一传输介质这一事实。

1.6 透明桥的工作原理

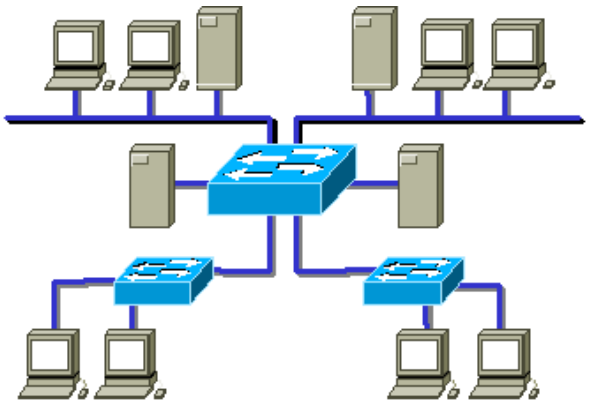


交换机有着透明桥相同的特点

在以太网中，作出转发决定的过程称为透明桥接。

透明的含义为：首先连接在网桥上的终端设备并不知道所连接的是共享媒介还是交换设备，即设备对终端用户来说是透明的，其次透明桥对其转发的帧结构不做任何改动与处理（VLAN 的 trunk 线路除外）。

1.7 透明桥的功能

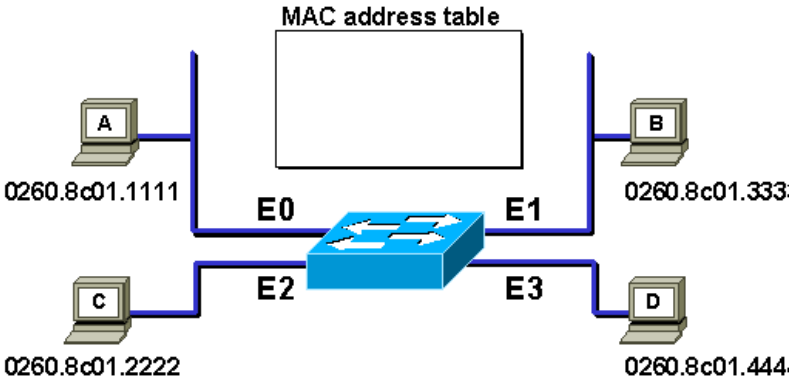


透明网桥有如下的三个主要功能。

- 地址学习功能。
- 转发和过滤功能。

- 环路避免功能。

通常透明网桥的三个主要功能都被使用，它们是在网络中是同时起作用的。而以太网交换机执行与透明桥相同的三个主要功能。

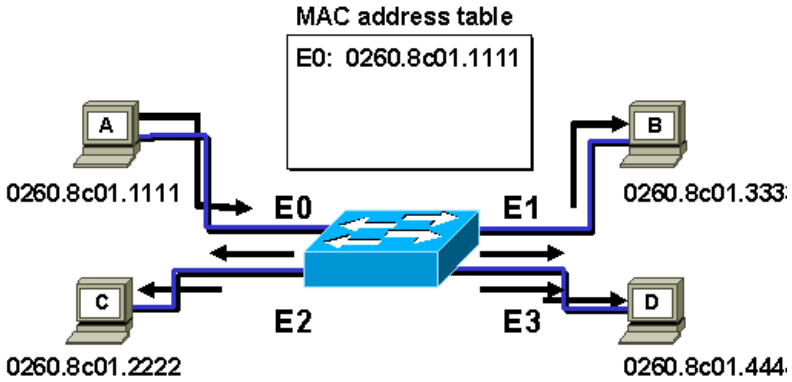


地址学习功能：

网桥基于目标 MAC（介质访问控制）地址作出转发决定。所以它必须“获取”MAC 地址的位置，这样才能准确地作出转发决定。

当网桥与物理网段连接时，它会对它监测到的所有帧进行检查。网桥读取帧的源 MAC 地址字段后与接收端口关联并记录到 MAC 地址表中。

由于 MAC 地址表是保存在交换机的内存之中的，所以当交换机启动时 MAC 地址表是空的。

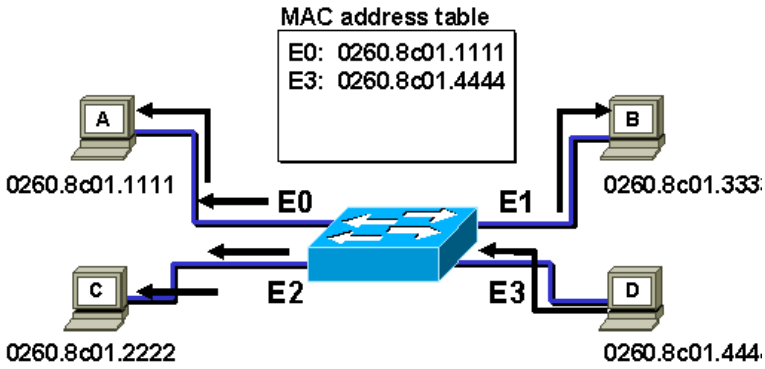


此时工作站 A 给工作站 C 发送了一个单播数据帧，交换机通过 E0 口收到了这个数据帧，读取出帧的源 MAC 地址后将工作站 A 的 MAC 地址与端口 E0 关联，记录到 MAC 地址表中。

由于此时这个帧的目的 MAC 地址对交换机来说是未知的，为了让这个帧能够到达目的地，交换机执行洪泛的操作，即从除了进入端口外所有其他端口转发。

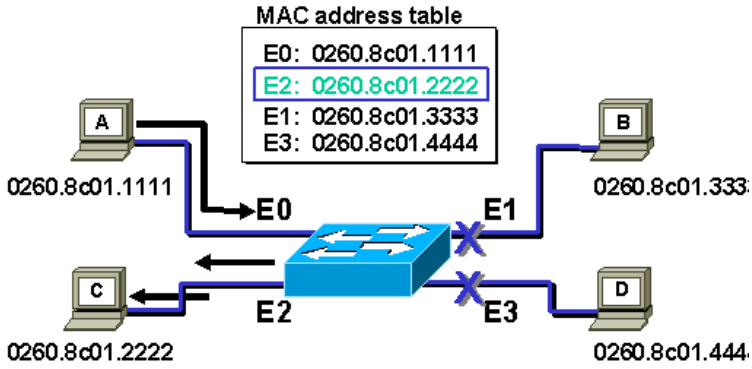
：注意

目的 MAC 地址未知的情况下交换机将洪泛数据帧。



工作站 D 发送一个帧给工作站 C 时，交换机执行相同的操作，通过这个过程交换机学习到了工作站 D 的 MAC 地址并与端口 E3 关联并记录到 MAC 地址表中。

由于此时这个帧的目的 MAC 地址对交换机来说仍然是未知的，为了让这个帧能够到达目的地，交换机仍然执行洪泛的操作，即从除了进入端口外所有其他端口转发。

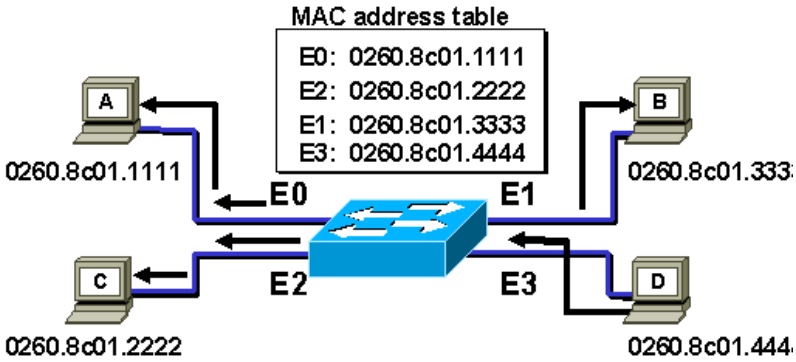


所有的工作站都发送过数据帧后，交换机学习到了所有的工作站的 MAC 地址与端口的对应关系并记录到 MAC 地址表中。

此时工作站 A 给工作站 C 发送了一个单播数据帧，交换机检查到了此帧的目的 MAC 地址已经存在在 MAC 地址表中，并和 E2 端口相关联，交换机将此帧直接向 E2 端口转发，即做转发决定。

对其他的端口并不转发此数据帧，即做所谓的过滤操作。

1.8 广播、组播和目的 MAC 地址未知帧的转发

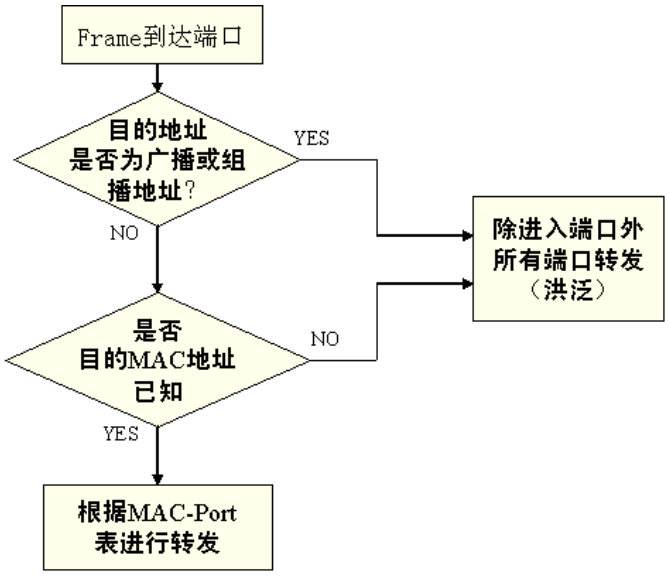


某工作站发出数据帧，交换机检测到目的 MAC 地址为广播、组播或目的 MAC 地址未知（交换机转发表中无此 MAC 地址）时，交换机将对此帧做洪泛的操作，即从除了进入端口外所有其他端口转发。

：注意

如果交换机支持 IGMP 监听等支持组播的功能，交换机将不再采用洪泛的方式转发组播数据帧。

1.9 转发/过滤流程



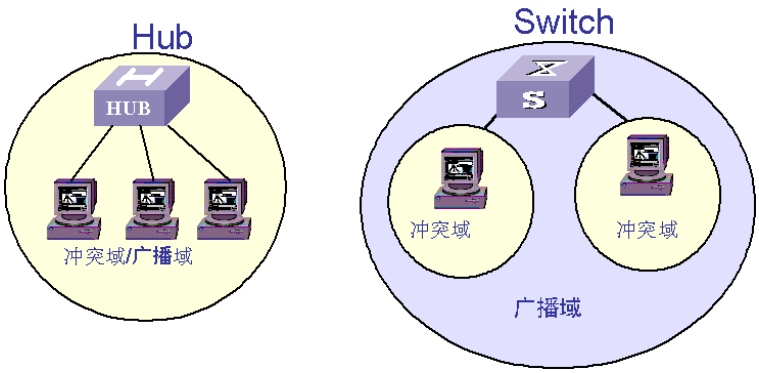
交换机在某端口接收到一个数据帧后的处理流程：

交换机首先判断此数据帧的目的 MAC 地址是否为广播或组播地址，如果是，即进行洪泛操作。

如果目的 MAC 地址不是广播或组播地址而是去往某设备的单播地址，交换机在 MAC 地址表中查找此地址，如果此地址是未知的，也将按照洪泛的方式进行转发。

如果目的地址是单播地址并且已经存在在交换机的 MAC 地址表中，交换机将把数据帧转发至此目的 MAC 地址关联的端口。

1.10 传统以太网与交换式以太网比较



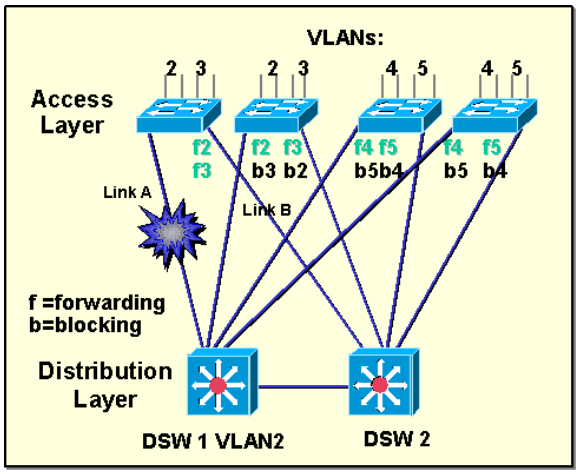
HUB（集线器）只对信号做简单的再生与放大，所有设备共享一个传输介质，设备必须遵循 CSMA/CD 方式进行通讯。使用 HUB 连接的传统共享式以太网中所有工作站处于同一个冲突域和同一个广播域之中。

交换机根据 MAC 地址转发或过滤数据帧，隔离了冲突域，工作在数据链路层。所以交换机每个端口都是单独的冲突域。

如果工作站直接连接到交换机的端口，此工作站独享带宽。

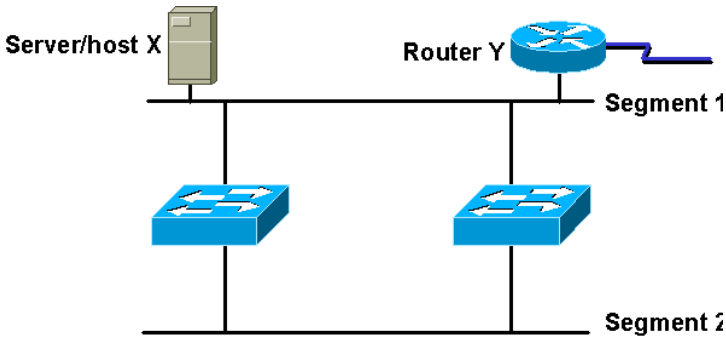
但是由于交换机对目的地址为广播的数据帧做洪泛的操作，广播帧会被转发到所有端口，所以所有通过交换机连接的工作站都处于同一个广播域之中。

1.11 保证网络的可靠性



为了提高整个网络的可靠性，消除单点失效故障，通常在网络设计中采用多台设备、多个端口、多条线路的冗余连接方式。

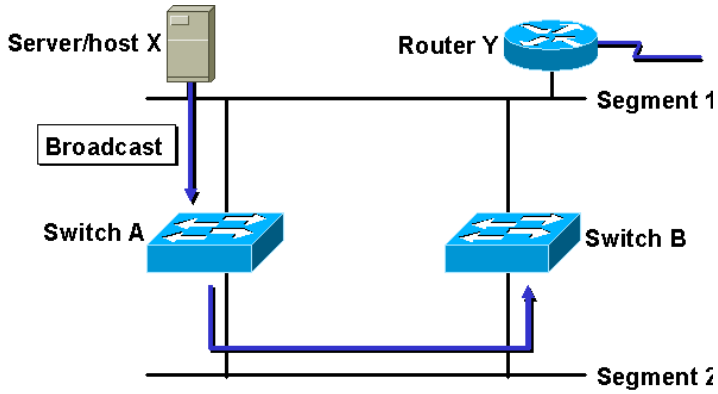
1.12 冗余拓扑



但是在存在物理环路的情况下可能导致 2 层环路的产生。

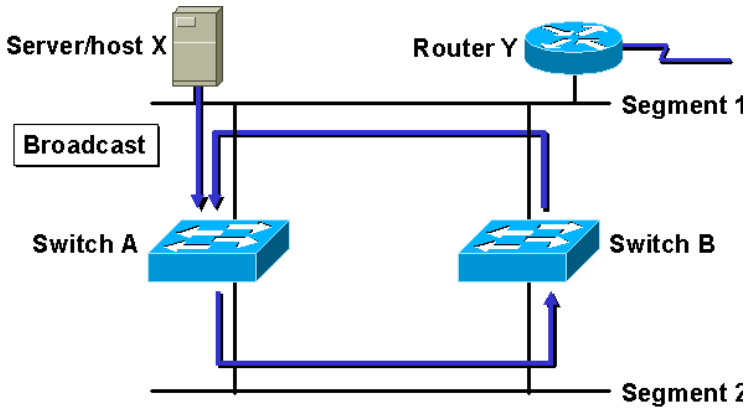
如果交换机不对 2 层环路做处理，将会导致严重的网络问题，包括：广播风暴；帧的重复复制；交换机 MAC 地址表的不稳定（MAC 地址漂移）等问题。

1.13 广播风暴

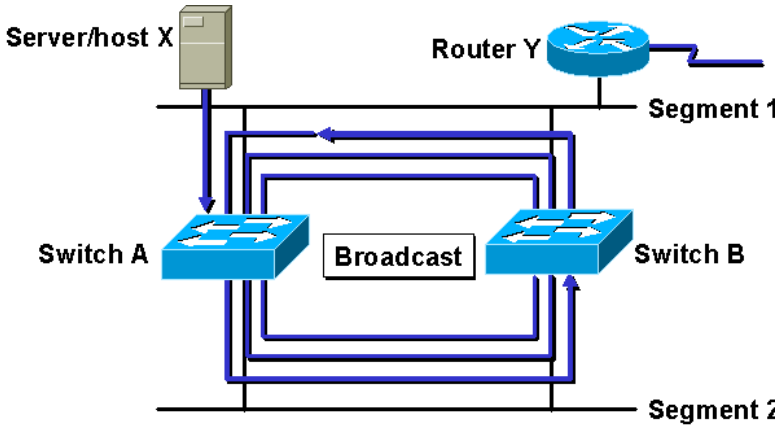


首先看看广播风暴是如何形成的：

在一个存在物理环路的 2 层网络中，主机 X 发送了一个广播数据帧，交换机 A 从上方的端口接收到广播帧，做洪泛处理，转发至下面的端口。通过下面的连接，广播帧将到达交换机 B 的下方端口。



交换机在下方的端口上收到了一个广播数据帧，将做洪泛处理，通过上方的端口转发此帧，交换机 A 将在上方端口重新接收到这个广播数据帧。



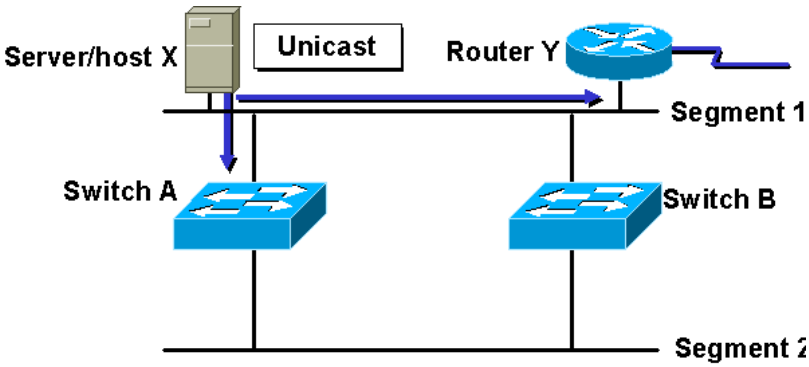
由于交换机执行的是透明桥的功能，转发数据帧时不对帧做任何处理。所以对于再次到来的广播帧，交换机 A 不能识别出此数据帧已经被转发过，交换机 A 还将对此广播帧做洪泛的操作。

广播帧到达交换机 B 后会做同样的操作，并且此过程会不断进行下去，无限循环。以上分析的只是广播被传播的一个方向，实际环境中会在 2 个不同的方向上产生这一过程。

在很短的时间内大量重复的广播帧被不断循环转发消耗掉整个网络的带宽，而连接在这个网段上的所有主机设备也会受到影响，CPU 将不得不产生中断来处理不断到来的广播帧，极大地消耗系统的处理能力，严重的可能导致死机。

一旦产生广播风暴系统无法自动恢复，必须由系统管理员人工干预恢复网络状态。（某些设备在端口上可以设置广播限制，一旦特定时间内检测到广播帧超过了预先设置的阀值即可进行某些操作，如关闭此端口一段时间以减轻广播风暴对网络带来的损害。但这种方法并不能真正消除 2 层的环路带来的危害）。

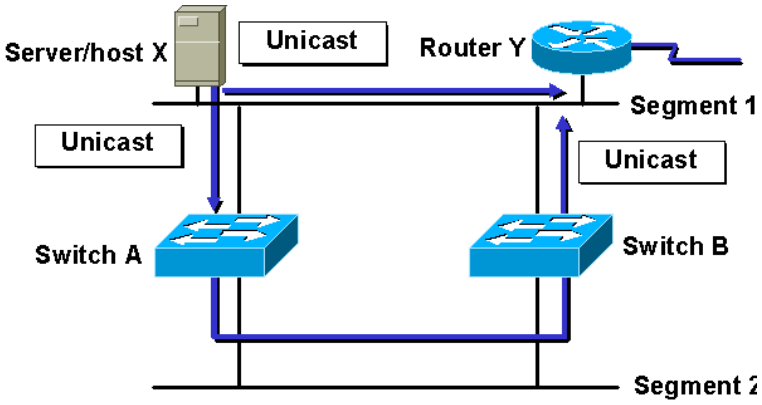
1.14 复制出多个重复的帧



接下来让我们看看一个数据帧被多次复制的情况：

主机 X 发送一单播数据帧，目的为路由器 Y 的本地接口，而此时路由器 Y 的本地接口的 MAC 地址对于交换机 A 与 B 都是未知的。

数据帧通过上方的网段直接到达路由器 Y，同时到达交换机 A 的上方的端口。

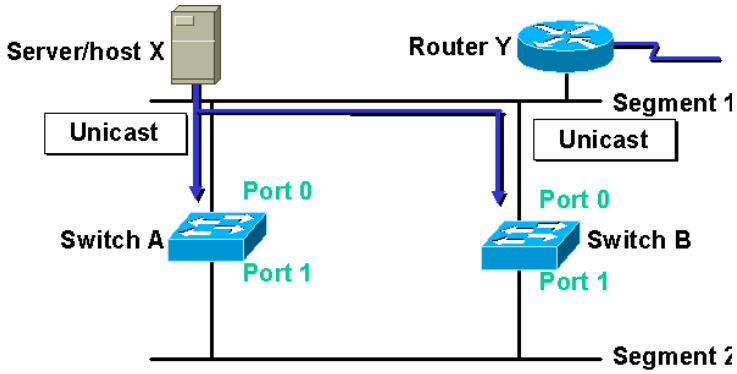


当交换机对于帧的目的 MAC 地址未知时交换机会进行洪泛的操作。

交换机 A 会将此数据帧从下方的端口转发出来，数据帧到达交换机 B 的下方端口，交换机 B 的情况与交换机 A 相同，也会对此数据帧进行洪泛的操作从上方的端口将此数据帧转发出来，同样的数据帧再次到达路由器 Y 的本地接口。

根据上层协议与应用的不同，同一个数据帧被传输多次可能导致应用程序的错误。

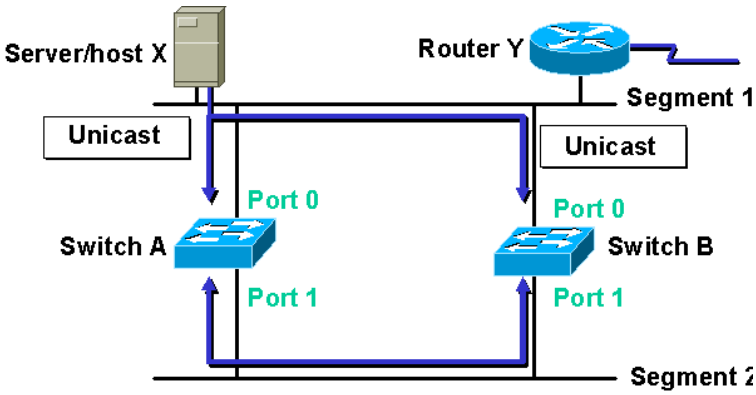
1.15 MAC 地址表的不稳定



最后看看 MAC 地址表不稳定的问题：

主机 X 发送一单播数据帧，目的为路由器 Y 的本地接口，而此时路由器 Y 的本地接口的 MAC 地址对于交换机 A 与 B 都是未知的。

数据帧通过上方的网段到达交换机 A 与交换机 B 的上方的端口。交换机 A 与交换机 B 将此数据帧的源 MAC 地址，即主机 X 的 MAC 地址与各自的 port0 相关联并记录到 MAC 地址表中。

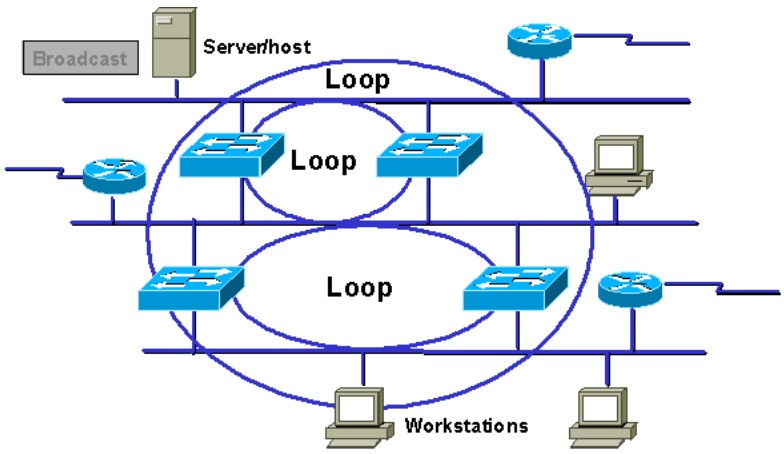


而此时两个交换机对此数据帧的目的 MAC 地址是未知的，当交换机对帧的目的 MAC 地址未知时交换机会进行洪泛的操作。两台交换机都会将此数据帧从下方的 port1 转发出来并将到达对方的 port1。

两个交换机都从下方的 port1 收到一个数据帧，其源地址为主机 X 的 MAC 地址，交换机会认为主机 X 连接在 port1 所在网段而意识不到此数据帧是经过其他交换机转发的，所以会将主机 X 的 MAC 地址改为与 port1 相关联并记录到 MAC 地址表中。交换机学习到了错误的信息，并且造成交换机 MAC 地址表的不稳定。

这种现象也被称为 MAC 地址漂移。

1.16 环路问题



以上所述表明在 2 层网络中一旦形成物理环路即可能形成 2 层环路，而 2 层环路给网络带来的损害是很严重的，并且往往一旦发生不会自动愈合。

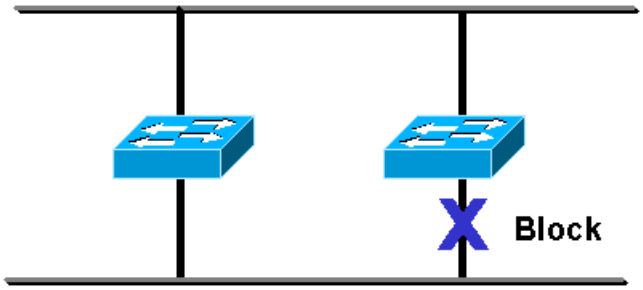
在实际的组网实际应用中经常会形成复杂的多环路连接。面对如此复杂的环路，网络设备必须有一种解决办法在存在物理环路的情况下阻止 2 层环路的发生。

第2章 生成树协议工作原理

📖 知识点

- 了解 STP 的工作原理，作用。
- 掌握 STP 协议中端口的各种状态。

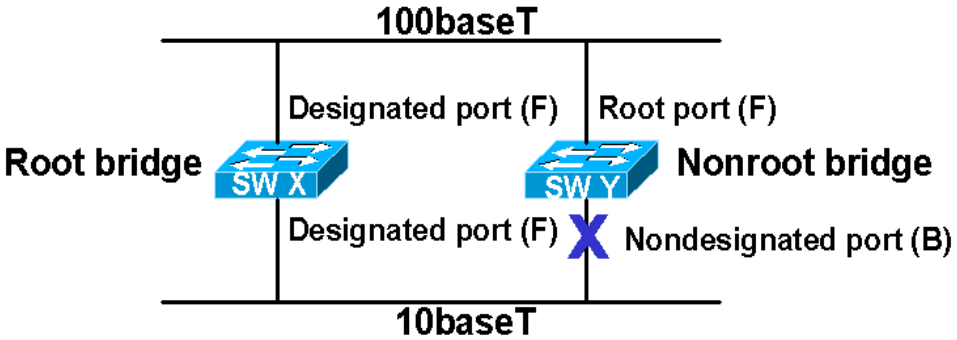
2.1 环路避免：生成树协议 (Spanning-Tree Protocol)



用生成树（spanning-tree protocol）可以在有物理环路的网络中阻止 2 层环路的产生。

生成树协议能够自动发现冗余网络拓扑中的环路，保留一条最佳链路做转发链路，阻塞其他冗余链路，并且在网络拓扑结构发生变化的情况下重新计算，保证所有网段的可达且无环路。

2.2 Spanning-Tree 的运作



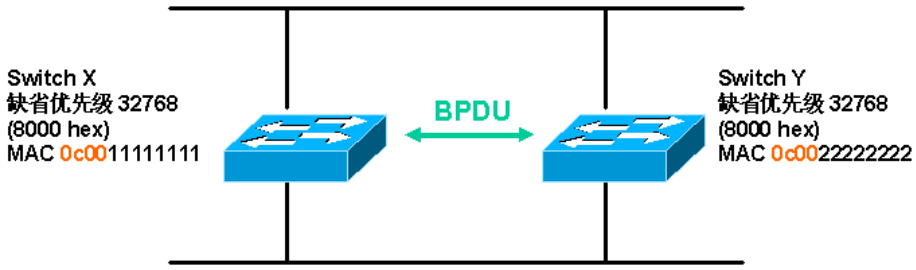
功能强大、可靠的网络需要有效地传输流量，提供冗余和故障快速恢复能力。

在第二层网络中，生成树协议通过在存在物理环路拓扑结构的网络上构建一个无环路的 2 层网络结构，提供了冗余连接，消除了环路的威胁。

STP 协议的基本思想十分简单。大家知道，自然界中生长的树一般情况下是不会出现环路的，如果网络也能够像一棵树一样生长就不会出现环路。于是，STP 协议中定义了根桥（Root Bridge）------生成树的参考点、根端口（Root Port）------非根桥到达跟桥的最近端口、指定端口（Designated Port）------连接各网段的转发端口、路径开销（Path Cost）------整个路径上端口开销之和等概念，目的就在于通过构造一棵自然树的方法达到裁剪冗余环路的目的，同时实现链路备份和路径最优化。

用于构造这棵树的算法称为生成树算法 SPA（Spanning Tree Algorithm）。

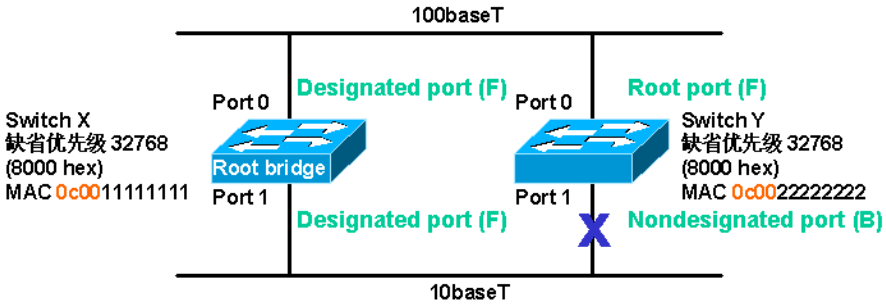
2.3 STP 根的选择



网桥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配置消息 BPDU（Bridge Protocol Data Unit）。STP BPDU 是一种二层报文，目的 MAC 是多播地址 01-80-C2-00-00-00，缺省情况下每 2 秒钟发出。所有支持 STP 协议的网桥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信息。

根桥的选举的依据是网桥优先级和网桥 MAC 地址组合成的桥 ID（Bridge ID），桥 ID 最小的网桥将成为网络中的根桥。各网桥都以默认配置启动，在网桥优先级都一样（默认优先级是 32768）的情况下，MAC 地址最小的网桥成为根桥，它的所有端口的角色都成为指定端口，进入转发状态。

2.4 STP 的端口状态



根桥上，所有端口都是指定端口，处于转发状态，用于为所有网段转发数据。

非根桥上，到达根桥最近的转发端口为根端口。非根桥上由于检测到环路而被阻塞掉的端口为非指定端口（不为相连网段转发数据）。

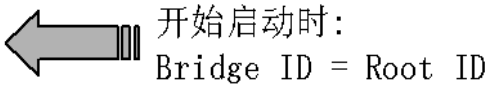
2.5 桥接协议数据单元(BPDU)

Bytes	Field
2	Protocol ID
1	Version
1	Message Type
1	Flags
8	Root ID
4	Cost of Path
8	Bridge ID
2	Port ID
2	Message Age
2	Maximum Time
2	Hello Time
2	Forward Delay

BPDU 的作用除了在 STP 刚开始运行时选举根桥外，其他的作用还包括检测发生环路的位置；通告网络状态的改变；监控生成树的状态等。

2.6 根的选择过程

Bytes	Field
2	Protocol ID
1	Version
1	Message Type
1	Flags
8	Root ID
4	Cost of Path
8	Bridge ID
2	Port ID
2	Message Age
2	Maximum Time
2	Hello Time
2	Forward Delay



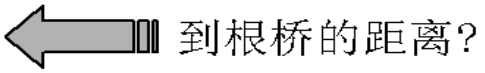
开始启动 STP 时，所有交换机将跟桥 ID 设置为与自己的桥 ID 相同，即认为自己是根桥。

当收到其他交换机发出的 BPDU 并且其中包含比自己的桥 ID 小的根桥 ID 时，交换机将此学习到的具有最小桥 ID 的交换机作为 STP 的根桥。

当所有交换机都发出 BPDU 后，具有最小桥 ID 的交换机被选择作为整个网络的根桥。根桥选举出以后，在正常情况下只有根桥每隔 2 秒钟从所有指定端口发出 BPDU。

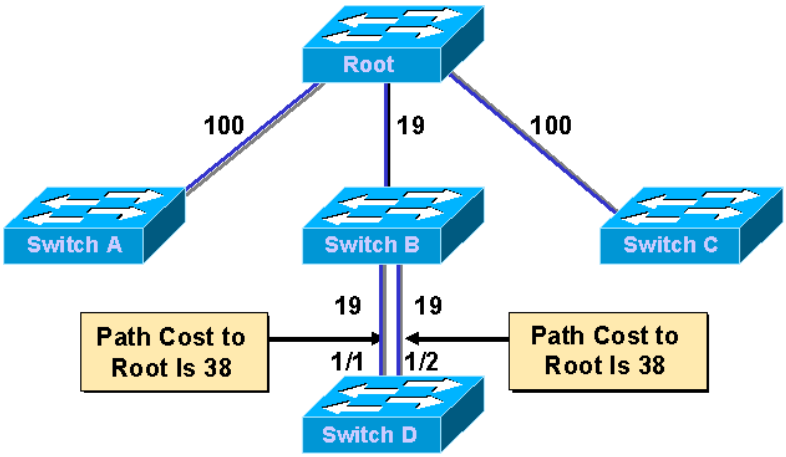
2.7 根路径的选择

Bytes	Field
2	Protocol ID
1	Version
1	Message Type
1	Flags
8	Root ID
4	Cost of Path
8	Bridge ID
2	Port ID
2	Message Age
2	Maximum Time
2	Hello Time
2	Forward Delay



根路径是根据 BPDU 中根路径开销，传输桥 ID，端口 ID 进行选择的。

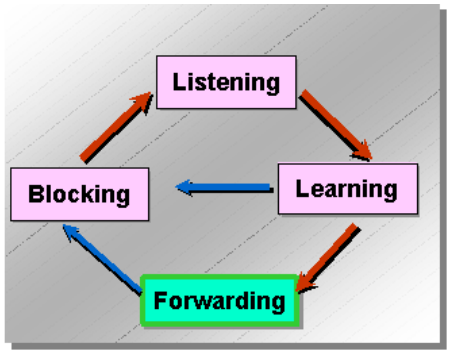
其中端口 ID 有 1 字节端口优先级与 1 字节端口号组成。根路径开销为到达根桥所经过的所有端口开销的总和。



当非根桥检测到了环路的存在后，必须保留一条链路做转发链路，阻塞掉其他冗余链路，选择转发链路的方式为：首先选择链路开销最小的链路做转发链路，如果存在多条链路开销相等且具有最小开销的链路则选择有最小转发桥 ID 的链路，如果存在多条桥 ID 相同的有最小链路开销的链路则选择有最小转发端口 ID 的链路。

2.8 STP 的端口状态

- 阻塞
- 倾听
- 学习
- 转发
- 关闭 (off)

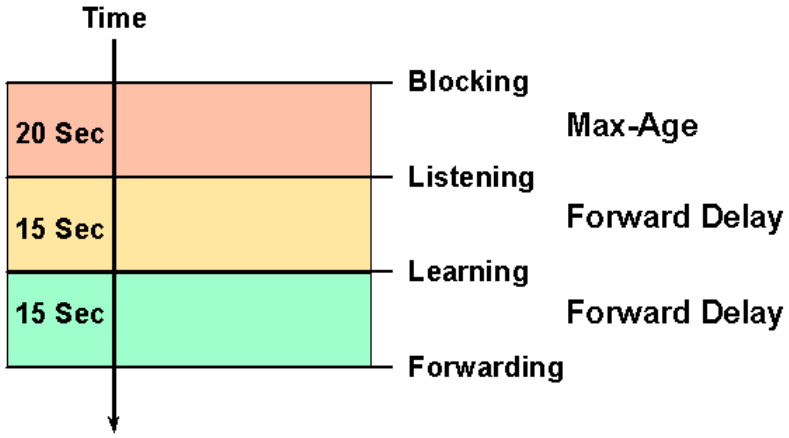


交换机的端口在 STP 环境中共有 5 种状态：阻塞、倾听、学习、转发、关闭(off)

交换机上一个原来被阻塞掉的端口由于在最大老化时间内没有收到 BPDU，从阻塞状态转变为倾听状态，倾听状态经过一个转发延迟（15 秒）到达学习状态，经过一个转发延迟时间的 MAC 地址学习过程后进入转发状态。

如果到达倾听状态后发现本端口在新的生成数中不应该由此端口转发数据则直接回到阻塞状态。

2.9 STP Timer



最大的老化时间（Bridge Max Age）：数值范围从 6 秒到 40 秒，缺省为 20 秒。

如果在超出最大老化时间之后，还没有从原来转发的端口收到根桥发出的 BPDU，那么交换机认为链路或端口发生了故障，需要重新计算生成树，需要打开一个原来阻塞掉的端口。

如果交换机在超出最大老化时间之后没有在任何端口收到 BPDU，说明此交换机与根桥失去了联系，此交换机将充当根桥向其它所有的交换机发出 BPDU 数据包。如果该交换机确实具有最小的桥 ID，那么，它将成为根桥。

当拓扑发生变化，新的配置消息要经过一定的时延才能传播到整个网络，这个时延称为转发延迟（Forward Delay），协议默认值是 15 秒。

在所有网桥收到这个变化的消息之前，若旧拓扑结构中处于转发的端口还没有发现自己应该在新的拓扑中停止转发，则可能存在临时环路。为了解决临时环路的问题，生成树使用了一种定时器策略，即在端口从阻塞状态到转发状态中间加上一个只学习 MAC 地址但不参与转发的中间状态，两次状态切换的时间长度都是 Forward Delay，这样就可以保证在拓扑变化的时候不会产生临时环路。但由此导致 STP 的切换时间比较长，典型的切换时间为最大的老化时间加 2 次转发延迟时间，约为 50 秒。

2.10 关键问题：收敛时间

对于运行 STP 的交换机来说，收敛（Convergence）状态意味着所有的交换机的端口都处于 forwarding 或 blocking 状态，状态稳定，没有拓扑结构发生变化。

当网络拓扑发生变化时，交换机必须重新计算生成树，在新的生成树没有完全计算、生成之前，为了防止临时环路的产生，所有链路都不转发数据。从发现状态改变到新的生成树计算完成的这段时间叫做收敛时间。通常 STP 大约的收敛时间为 50 秒左右。

由于标准 STP 的收敛时间较长，导致很多应用在切换过程中受影响。针对这个问题提出了 RSTP（IEEE 802.1w）协议，即快速生成树，可以显著减少收敛时间。

思考题：

1： 交换机如何处理广播包?如何处理单播包?

2： 生成树端口状态有几种?各自的功能是什么?

3： STP 选举根网桥的原则是什么?

4： STP 选择端口的原则是什么?

5： 如何改善 STP 收敛的速度?

6：BPDU 的作用是什么?