# 现代计算机网络

李伟明 lwm@hust.edu.cn

# 1.3 报文交换



1.3 概述

1.3.1以太网及其发展

1.3.2 以太网交换机

1.3.3 STP（Spanning Tree Protocol）生成树协议

# 1.3 报文交换



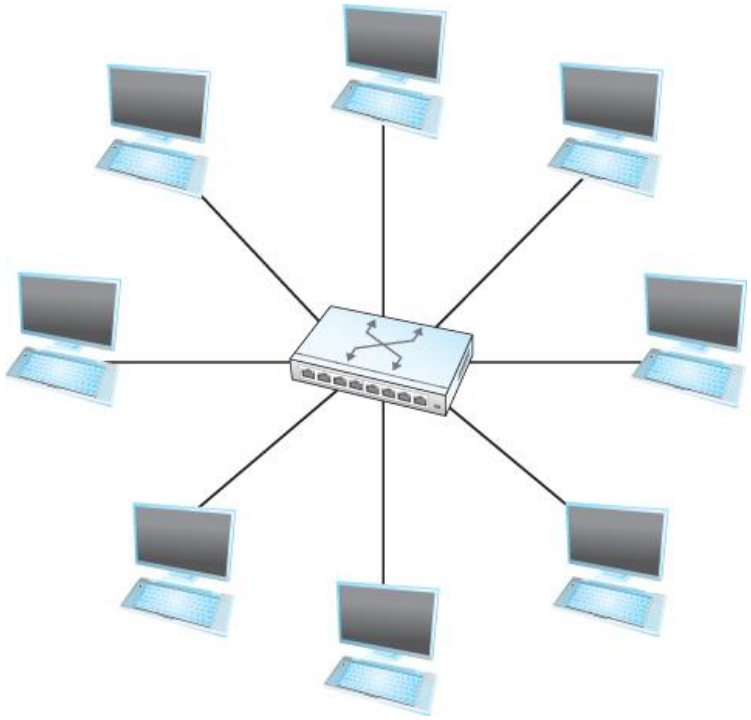
直连网络存在两个问题：

1. 由于共享链路，所以仅仅允许少量的节点连接到网络，例如早期的以太网最多不超过1024个节点
2. 仅仅能分布在很小的范围，早期以太网距离不超过2500m

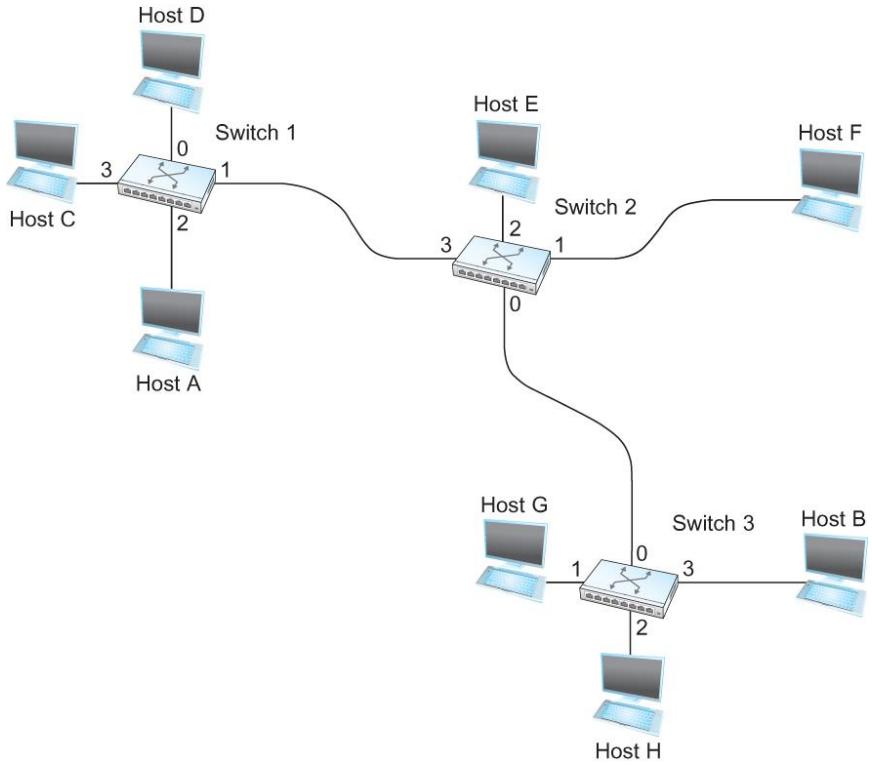
所以出现Packet Switch这样的设备，将多个直连网络连接起来，组建更大范围的网络。

# 1.3 报文交换

而且交换机提供了另外一个好处：每个host到Switch是独立的带宽。所以后来进一步，有的时候干脆通过Packet Switch直接将多个主机连接，组合成交换网络。

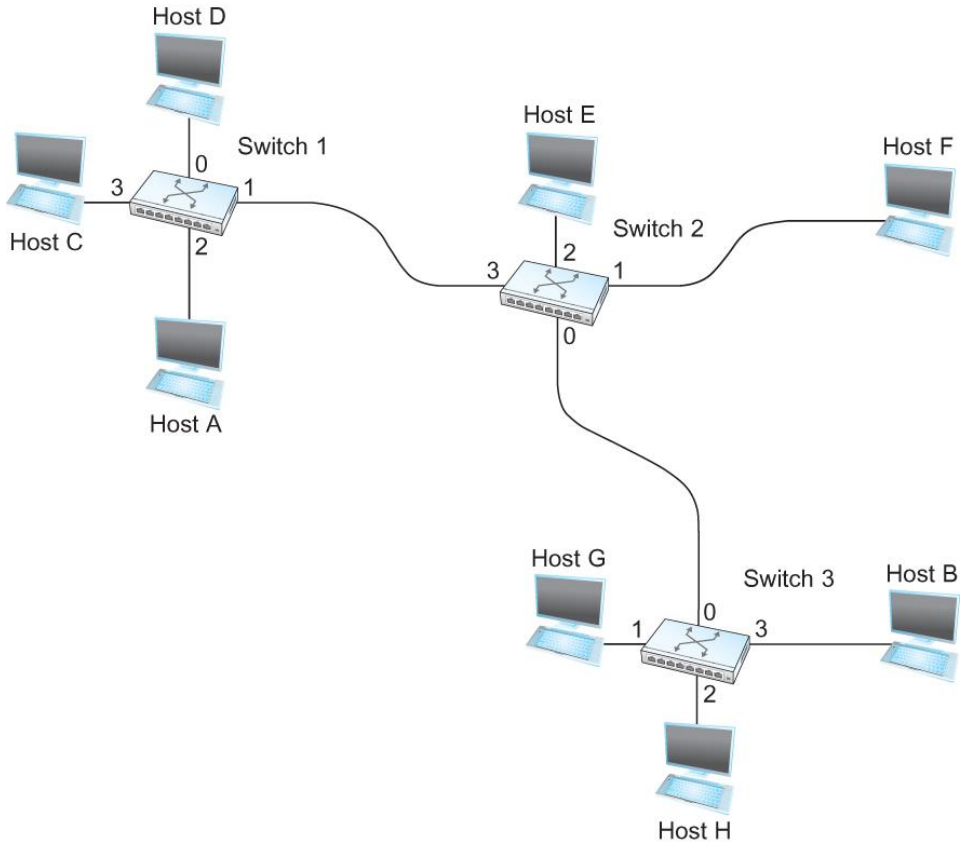


# 1.3 报文交换




- 多个Packet Switch连接起来后，马上就需要解决如何确定转发路径的问题，每个交换机需要建立一个*forwarding table，自学习*

# 1.3 报文交换




| Destination | Port |
|---|---|
| A | 3 |
| B | 0 |
| C | 3 |
| D | 3 |
| E | 2 |
| F | 1 |
| G | 0 |
| H | 0 |

**Forwarding Table for Switch 2**

# 1.3 报文交换



教科书3.1节介绍了三种交换的方法：

1. Datagram (or connectionless) model
2. Virtual Circuit Switching（ATM）
3. Source Routing

我们重点介绍是交换式以太网使用的第一种方法，也是TCP/IP默认的通信子网连接方式

# 1.3.1 以太网



- 1975年纯ALOHA原始Ethernet：单工竞争系统，基本思想：
  - 无连接，先说后听，想发就发，错了重发；
  - 对数据帧不编号，不要求对方发回确认；不可靠交付，尽力而为
  - 建立在近距离、信道出错概率小-->局域网
- 随机接入协议（发展）
  - 普通ALHOA和Time sloted ALHOA
  - CSMA（载波侦听多路访问）：先听后说+指数退避
    - 1坚持CSMA、非坚持CSMA、P坚持CSMA
  - CSMA/CD：多点接入、载波监听、碰撞检测
    - 信道效率：传送距离越短，发送帧时$T_0$越长，效率越高
- 以太网优势
  - 可扩展（10M—10G），灵活（多种媒介、全/半双工、共享/交换），便宜、易于安装使用、稳健性好。

# 以太网卡与MAC地址模式



- 网卡功能
  - 数据的封装与解封
  - 链路管理：CSMA/CD
  - bit的编码与解码
- MAC地址
  - Unicast:单播帧地址,仅对某个网卡
  - Broadcast:广播帧地址,仅对某个子网
  - Multicast:多播帧地址,组地址
  - 杂收模式：Promiscuous mode:接收总线上所有的可能接收的帧

# 高速局域网：快速以太网100Mbps



- 对10Mbps 802.3 LAN的改进
  - 局域网发展史上重要里程碑

- Fast Ethernet标准
  - 1995年，IEEE通过802.3u标准，实际上是802.3的一个补充。原有的帧格式、接口、规程不变，只是将每比特时间从100ns缩短为10ns。

| Name | Cable | Max. segment | Advantages |
|---|---|---|---|
| 100Base-T4 | Twisted pair | 100 m | Uses category 3 UTP |
| 100Base-TX | Twisted pair | 100 m | Full duplex at 100 Mbps |
| 100Base-FX | Fiber optics | 2000 m | Full duplex at 100 Mbps; long runs |

# 高速局域网:100Base-TX/F



- 100Base-TX
  - 使用2对5类平衡双绞线或150Ω屏蔽平衡电缆，1对 to the hub，1对from the hub，支持全双工；
  - 5类双绞线使用125 MHz的信号；
  - **4B/5B编码**，5个时钟周期发送4个比特，物理层与FDDI兼容使用NRZ-I编码，比特率为125 * 4/5 = 100 Mbps；
- 100Base-FX
  - 使用2根多模光纤，支持全双工，物理层和FDDI兼容
- 100Base-T4（4根线收发） 和 100Base-TX（2根线收发）统称 100Base-T
- 两种类型的HUB
  - 共享式HUB，一个冲突域，工作方式与802.3相同，CSMA/CD，二进制指数后退算法，半双工 …
  - 交换式HUB，输入帧被缓存，一个端口构成一个冲突域。

# 高速局域网:100Base-TX/F



随着网络速度的提高，共享信道方式越来越不适应发展

- 共享信道跟全双工是矛盾的
- 100Base-FX在半双工模式下，为了兼容以太网报文，为了确保检测到碰撞，最大长度为412米。但是在全双工模式下，则可达到2000m，如果用Repeater，可以到10公里
- 共享信道意味着所有主机在一个冲突域，主机数量一多冲突的概率大大增加。

# 高速局域网: 1000Mbps以太网



- 工作方式
  - IEEE 802.3定义的10M/100M以太网一致的CSMA/CD帧格式
  - 以太网交换机（全双工模式）中的千兆端口不能采用共享信道方式访问介质，不使用 CSMA/CD 协议，而只能采用全双工方式.
  - 在使用双绞线的情况下，可以通过自动协商机制，切换到半双工方式下仍使用 CSMA/CD 协议
  - 编码采用8b/10b encoding和NRZ
  - 双绞线很少用，因为同时5对线收发产生串扰

| Interface Name | Cables | Maximum Transmission Distance |
|---|---|---|
| 1000Base-LX | Single-mode fiber or multi-mode fiber | 316 m |
| 1000Base-SX | Multi-mode fiber | 316 m |
| 1000Base-CX | Balanced twisted pair copper wire cable | 25 m |
| 1000Base-TX | Category 5 twisted pair cable | 100 m |

# 高速局域网: 1000Mbps以太网



- PAUSE协议
  - 规范发展完善了PAUSE协议，不采用CSMA/CD协议完成全双工操作。
  - 该协议采用不均匀流量控制方法，最先应用于100M以太网中。

- 流控
  - 利用802.3定义的Pause控制帧进行流量控制，要求发送数据节点暂停数据发送，避免缓冲区溢出造成的丢包。
  - 只有在全双工时，才支持Pause流控，半双工时不支持流控。

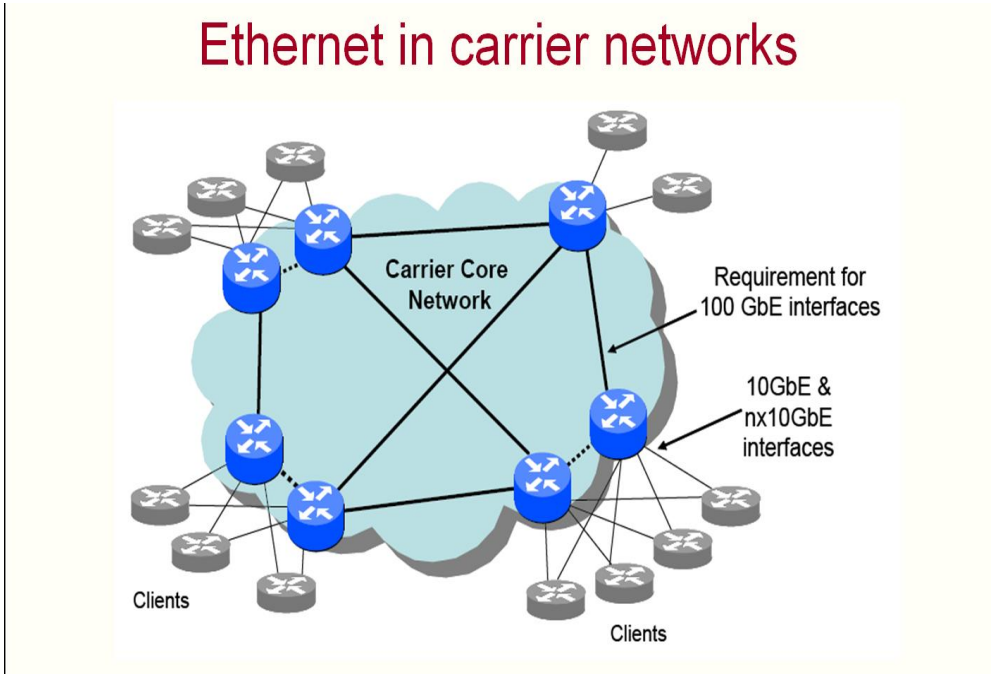
# 万兆（10Gbps）以太网



- 2002.6月正式发布802.3ae 10GE标准
  - 只全双工，不支持单工和半双工，也不采用CSMA/CD
  - 不持自协商；提供广域网物理层接口。
- 长距离(40-50KM)网络
  - 扩展了网络的覆盖区域，且标准简化。
  - 支持现存的大量SONET网络兼容
- 两种物理层技术：
  - 局域网物理层LAN PHY；10.000Gbps精确10G；
  - 广域网物理层WAN PHY；OC-192，异步SONET/SDH
  - 与10M/100M/1000Mbps帧格式完全相同；

# 100Gbps以太网






- 以太网封装比SONET/SDH更简单且成本更低
- 40 Gbps已成为过渡产品
- 2010年6月22日，IEEE802.3ba和100Gb/s以太网技术标准已经正式获审通过。
- 国内华为、中兴；国外Juniper Network、CISCO；上海贝尔已经开发出了自有标准100Gbps以太网接口路由器

**100 GbE standard is needed**

**Jumping to 100 GbE Ethernet at 100 Gbps may take place by using several or just one lambda(s):**

**100 GbE over** 10x10Gbps, or 4x25Gbps or 1x100Gpbs

*Different from 10x10GbE !!*

# 1.3.2 交换机

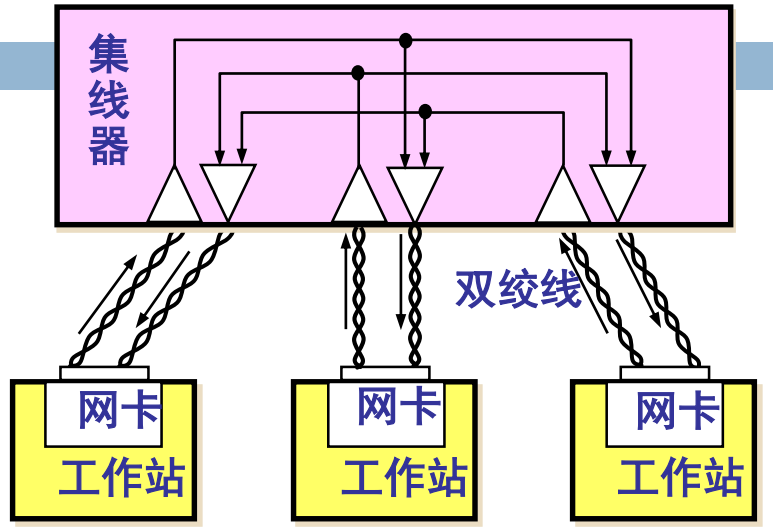


- 以太网可以是一组直连的主机
- 以太网也可以通过网桥（bridge）、集线器（HUB）、交换机(Switch)这些连接设备进行扩展

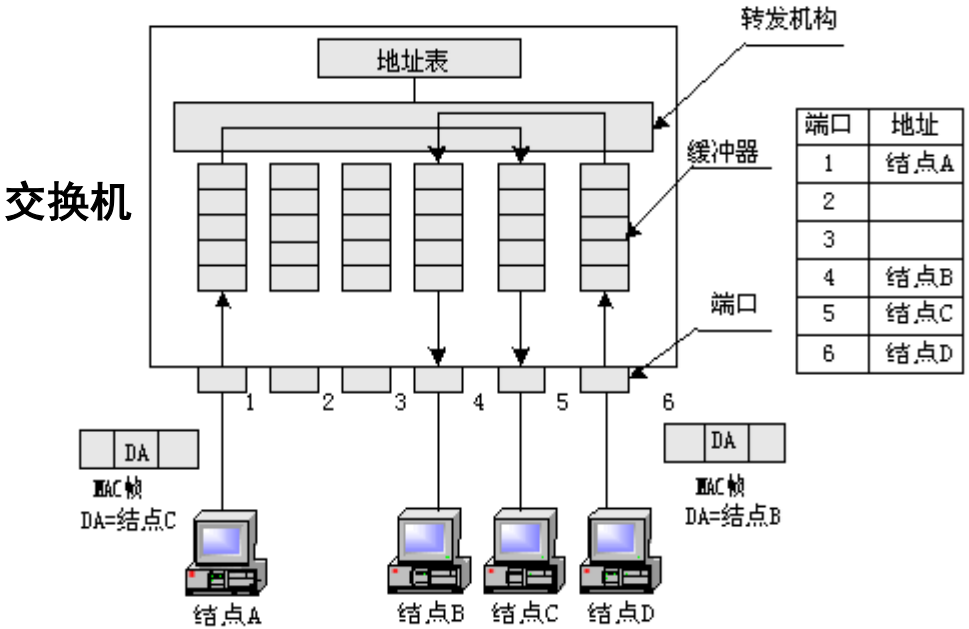
# 以太网的连接设备



- 集线器（HUB）:物理层互连设备
  - 1进多出，相同速率，无帧缓冲/线障隔离，使用方便
  - 带宽受限，广播风暴，单工传输，通信效率低

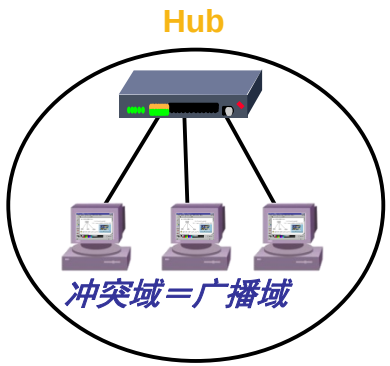


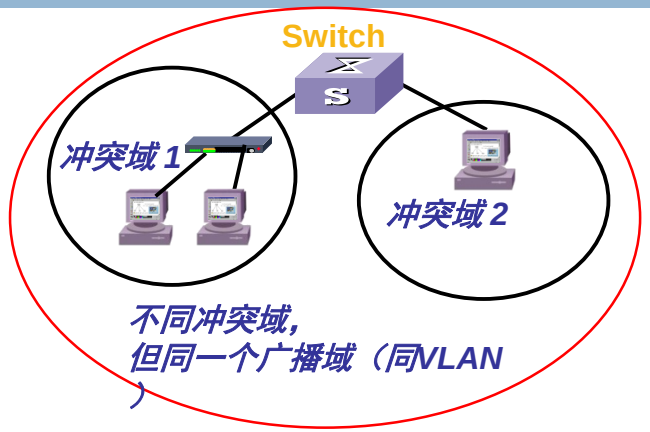

- 交换机（Switch）：链路层互连设备
  - 依帧头信息转发以太帧；
  - 实现方法
    - 直接交换方式
    - 存储转发方式
    - 改进直接交换方式。

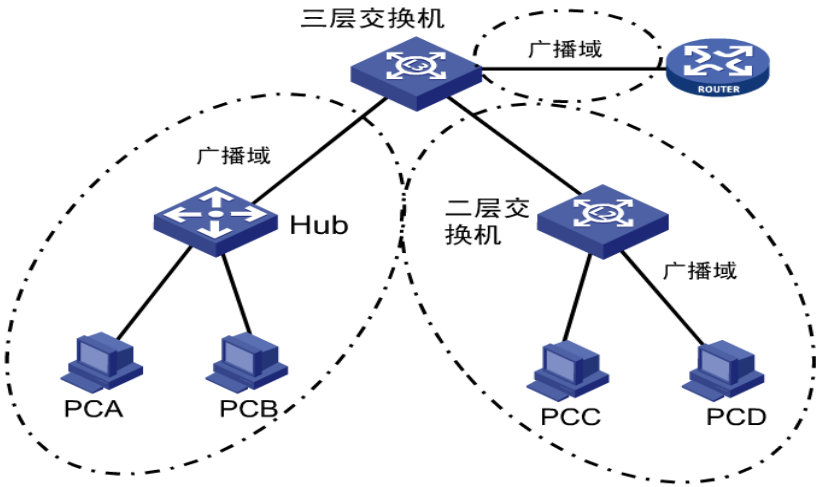


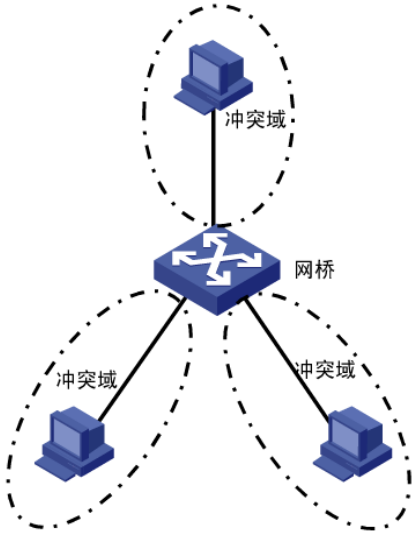
# 广播域和冲突域

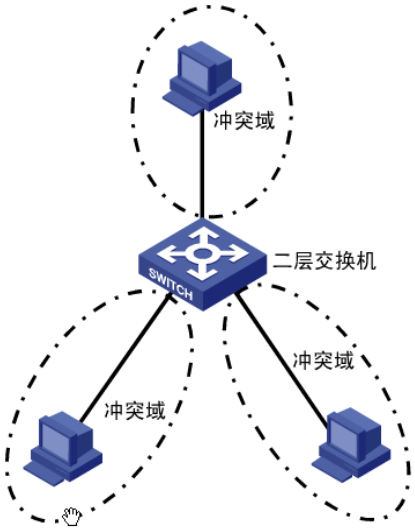





**广播域**：广播报文可以达到的范围
**冲突域**：可以发生报文冲突的范围

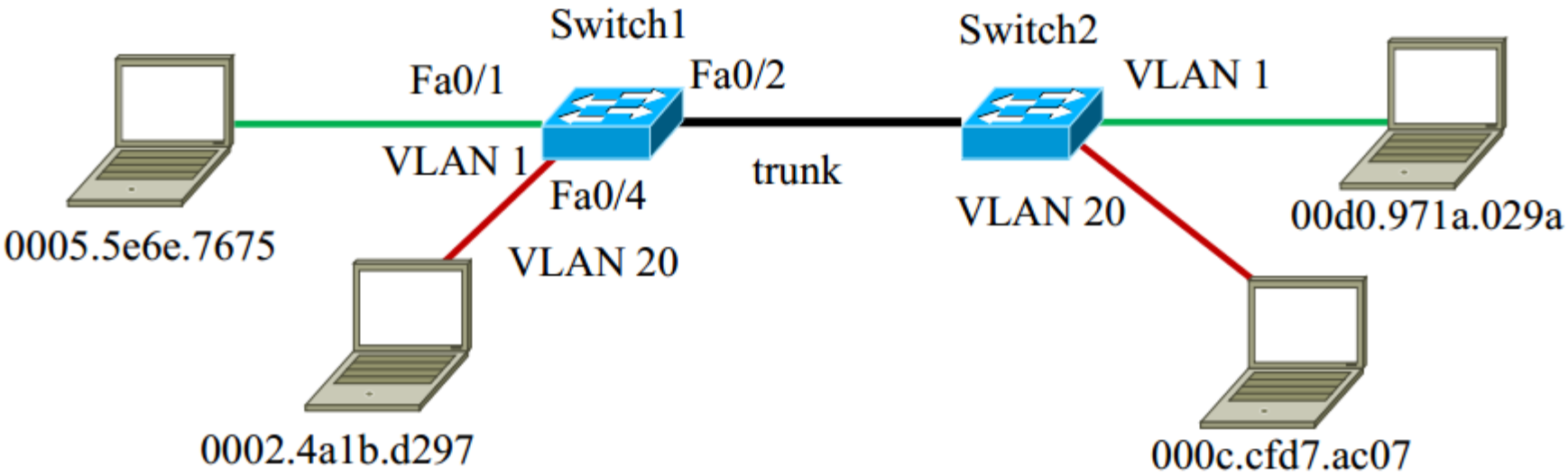
交换以太网每个端口处于独立**冲突域**

# 虚拟局域网



- 不依赖三层交换，通过虚拟局域网VLAN，可以将同一交换机或者多个交换机的广播域划分多个广播域

- 每个虚拟局域网一个广播域

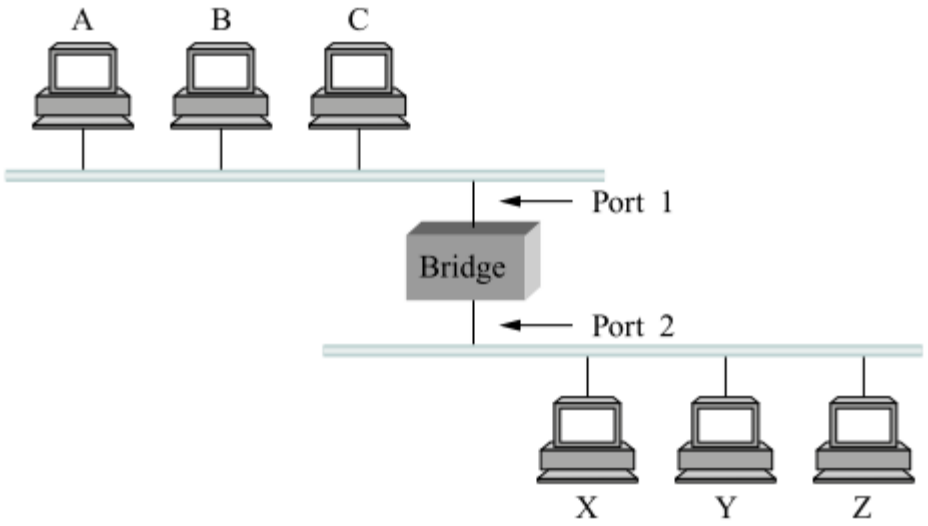


# 1.3.2 交换机-自学习功能



- 最简单的情况：

右边的转发表是可以通过交换机学习自动得到的

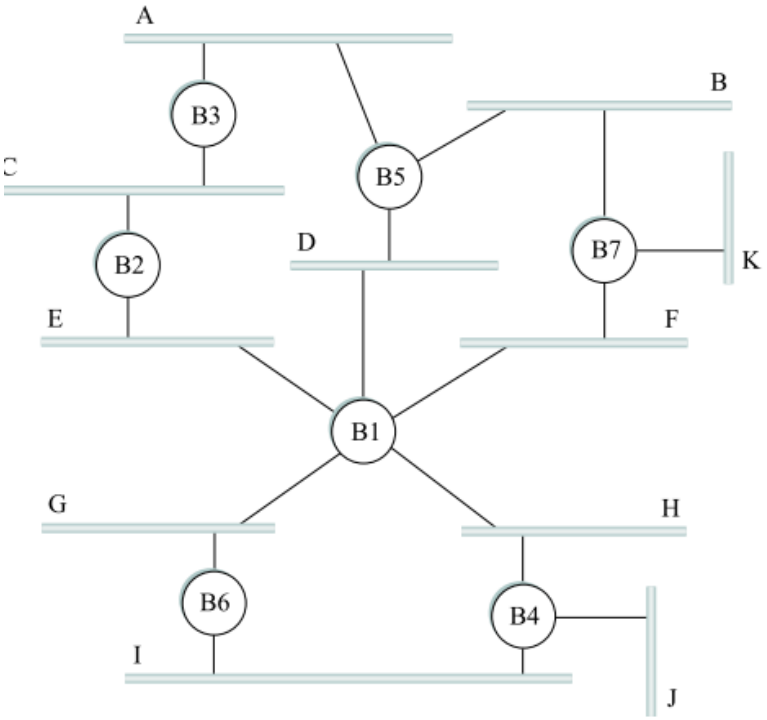



| Host | Port |
|---|---|
| A | 1 |
| B | 1 |
| C | 1 |
| X | 2 |
| Y | 2 |
| Z | 2 |

# 1.3.3 交换机生成树协议



- 交换机的自学习机制可以工作良好，但是一旦交换机连接存在环路，会造成严重后果
- B1, B4, and B6 form a loop
- 当交换机之间存在多条活动链路时，容易形成环路，导致转发表的不正确与不稳定，并且还会导致重复的数据包在网络中传递，引起广播风暴，使网络不稳定。



# 1.3.3 交换机生成树协议

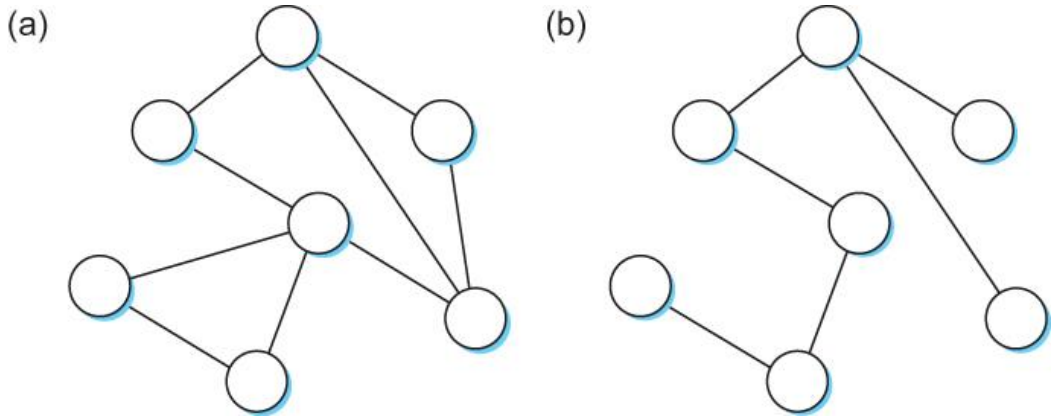


- 为了防止交换机之间由于多条活动链路而导致的网络故障，必须将多余的链路置于非活动状态，即不转发用户数据包，而只留下单条链路作为网络通信。

- 要实现此功能，需要依靠生成树协议（Spanning Tree Protocol）来完成，STP将交换网络中任何两个点之间的多余链路置于Blocking（关闭）状态，而只留一条活动链路，当使用中的活动链路失效时，立即启用被Block的链路，以此来提供网络的冗余效果。

# 1.3.3 交换机生成树协议

- 从理论上一个LAN可以看作一个图graph，这个图可能循环 (cycles)
- spanning tree实际是这个图的一个子图sub-graph，但是可以到达所有节点




Example of (a) a cyclic graph; (b) a corresponding spanning tree.

# 1.3.3 交换机生成树协议

Spanning Tree协议

- 需要所有交换机支持
- IEEE 802.1 标准规定了LAN 交换机必须支持
- 每个交换机实际上需要disable掉自己的某些端口，不转发frame
- 极端情况下，可能一个交换机完全不参与任何frame的转发

# 1.3.3 交换机生成树协议

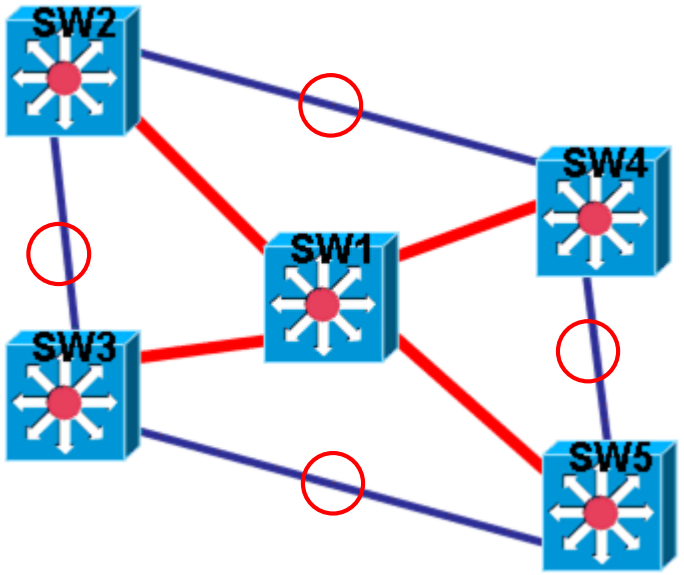


- STP为IEEE标准协议，并且有多个协议版本，版本与协议号的对应关系如下：
  - Common Spanning Tree (CST) = IEEE 802.1D
  - Rapid Spanning Tree Protocol (RSTP)=IEEE 802.1w
  - Per-VLAN Spanning-Tree plus (PVST+)=Per-VLAN IEEE 802.1D
  - Rapid PVST+=Per-VLAN IEEE 802.1w
  - Multiple Spanning Tree Protocol (MSTP)=IEEE 802.1s

# 1.3.3 交换机生成树协议



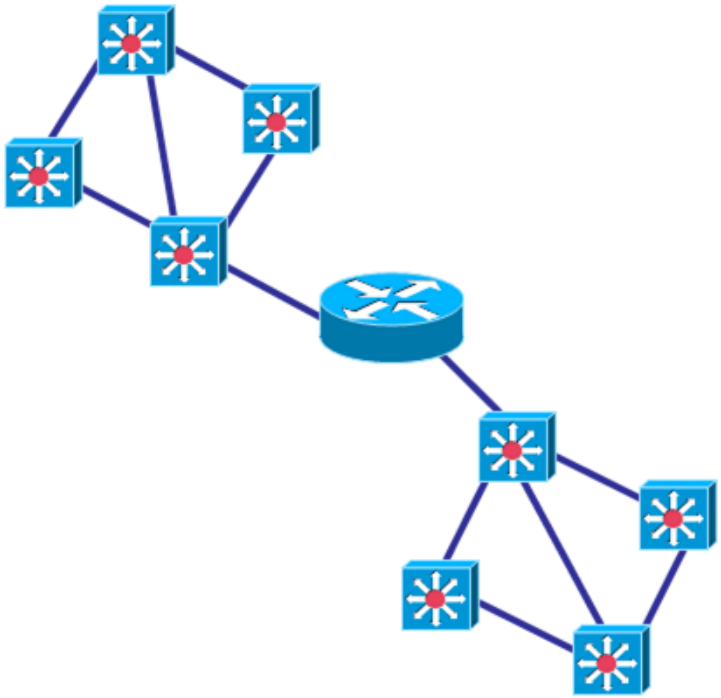
- 去掉环路的方法：所有交换机按照树的方式进行连通
- STP的核心思想是网络中选出一台交换机做为核心交换机，STP称其为Root，也就是根，功能相当于hub-spoke网络中的Hub。
- 其它不是Root的交换机则需要留出一条活动链路去往根交换机，因为只要普通交换机到根是通的，到其它交换机也就是通的。



# 1.3.3 交换机生成树协议



- 只有在一个广播能够到达的范围内，才需要进行相同的STP计算与选举，也就是一个广播域内独立选举STP
- 下图中网络被路由器分割成两个广播域，所以在两个网段中，需要进行独立的STP计算与选举。



# 1.3.3 交换机生成树协议



交换机之间选举根交换机（Root）

- 一个广播域内只能选举一台根交换机。Birdge-ID中优先级最高（即数字最小）的为根交换机，优先级范围为0-65535，如果优先级相同，则MAC地址小的为根交换机。

交换机端口之间选举根端口（Root Port）

- 所有非根交换机都要选举根端口，选举规则为到根交换机的Path Cost值最小的链路。

非根交换机选择指定端口(Designated Port)

- 简单地理解为每条连接交换机的link有两个端口（属于不同交换机）中，有一个要被选举为指定端口。
- 选举规则和选举根端口一样，即：到根交换机的Path Cost值最小的端口，如果多条链路到达根交换机的Path Cost值相同，则选举上一跳交换机Bridge-ID最小的链路。

# 1.3.3 交换机生成树协议



剩余端口状态为Blocking

- 在STP选出根交换机，根端口以及指定端口后，其它所有端口全部为Blocking状态，为了防止环路，所有Blocking端口只有在根端口或指定端口失效、拓扑改变的时候才会被启用。

- 一个端口，在STP中只能处于一种角色，不可能是两种角色

# 1.3.3 交换机生成树协议-选举



BPDU（Bridge Protocol Data Unit）

交换机间用BPDU报文来选举，目的地址为layer 2 multicast address 01:80:C2:00:00:00 ：

- Protocol ID 固定为0。
- Version：0为802.1d，1为802.1w，2为802.1s。
- Message type：0为普通BPDU，80为TCN（Topology Change Notification）。
- Flags字段：802.1d时只用到0位和7位，都和TCN相关，TCN的ACK报文里0位置1，TC报文里7位置1。
- Root ID，Cost of path，Bridge ID，Port ID ：用于选举。

| 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 | Max age |
| 2 | Hellotime |
| 2 | Forward delay |

# 1.3.3 交换机生成树协议-选举



## Path cost计算

- 每个交换机会把自己链路的代价加上接收到的邻居交换机的Path Cost，得到总的Path Cost。

| Data rate (Link Bandwidth) | STP cost (802.1D-1998) |
|---|---|
| 4 Mbit/s | 250 |
| 10 Mbit/s | 100 |
| 16 Mbit/s | 62 |
| 100 Mbit/s | 19 |
| 1 Gbit/s | 4 |
| 2 Gbit/s | 3 |
| 10 Gbit/s | 2 |
| 100 Gbit/s | N/A |
| 1 Tbit/s | N/A |

# 1.3.3 交换机生成树协议-选举过程



1. 当交换机打开的时候，所有的端口都处于Listening状态，每个交换机都会认为自己是根交换机（Root ID为自己），然后都每隔两秒就向外发送一次自己的BPDU。
2. 如果收到的BPDU的Bridge ID比自己的小，则停止转发自己的BPDU，开始转发更优的BPDU，如果比自己的Bridge ID大或者和自己的Bridge ID相等，则丢弃该BPDU。
3. 持续15s（转发延迟）等到BPDU扩散完毕之后，开始各种端口的选举，这时候每个BID最小的交换机成了根交换机，各个交换机通过收到的BPDU来确定根端口和指定端口。剩下的成为非指定端口，转到blocking状态。然后进入learning状态
4. 进入Learning状态之后，填写MAC地址表，经过15s（转发延迟）之后进入Forwarding状态。

# 1.3.3 交换机生成树协议-选举过程



5. 进入Forwarding状态之后，开始转发数据，并且同时接受转发来自于根的BPDU（Root ID为根交换机），维护拓扑。这时只有根交换机发BPDU，其他交换机都只是转发BPDU。
6. 当一个新的交换机加入的时候，端口状态是Learning，新的交换机认为自己是根交换机开始发送BPDU，也接收对端的BPDU，然后进行进一步的竞选。
7. 若竞选成功，则网络拓扑就重新变化了，若竞选失败则计算根端口指定端口和非指定端口。（30s可以完成）

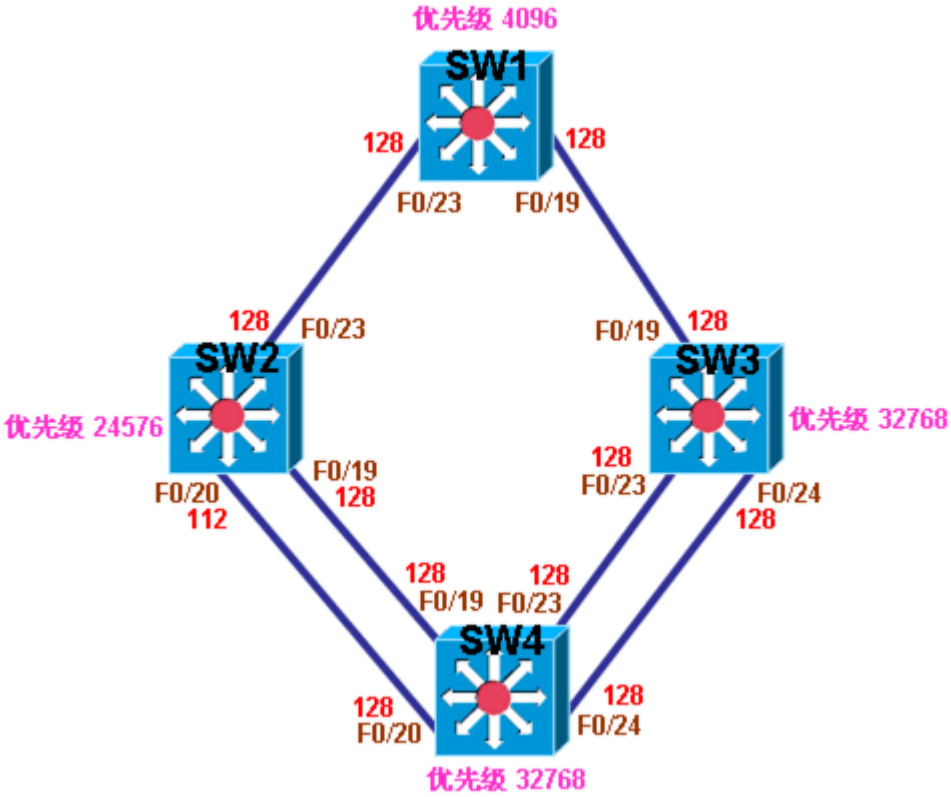
# 1.3.3 交换机生成树协议



举一个具体例子：

根交换机（Root）

- 因为4台交换机的优先级分别为SW1（4096），SW2（24576），SW3（32768），SW4（32768），选举优先级最高的（数字最低的）为根交换机
- 所以SW1被选为根交换机，如果优先级相同，则比较MAC地址。
- 所有链路为100 Mb/s，即Path Cost值为19；128为port ID



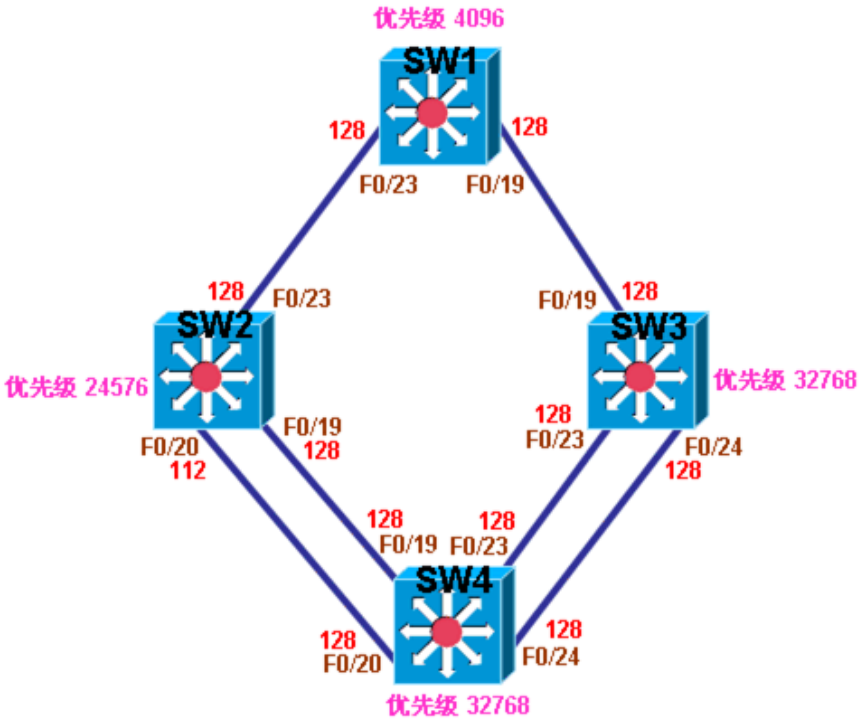
# 1.3.3 交换机生成树协议



## 根端口（Root port）

根端口需要在除SW1外的非根交换机上选举。

- SW2从端口F0/23到达根的Path Cost值为19，从F0/19和F0/20到达根的Path Cost值都为19×3=57。因此，F0/23被选为根端口。
- SW3上F0/19被选为根端口。
- SW4上从所有端口到达根的Path Cost值都为19×2=38，接下来比较上一跳交换机Bridge-ID，选择SW2，再比较对端端口优先级，选择F0/20

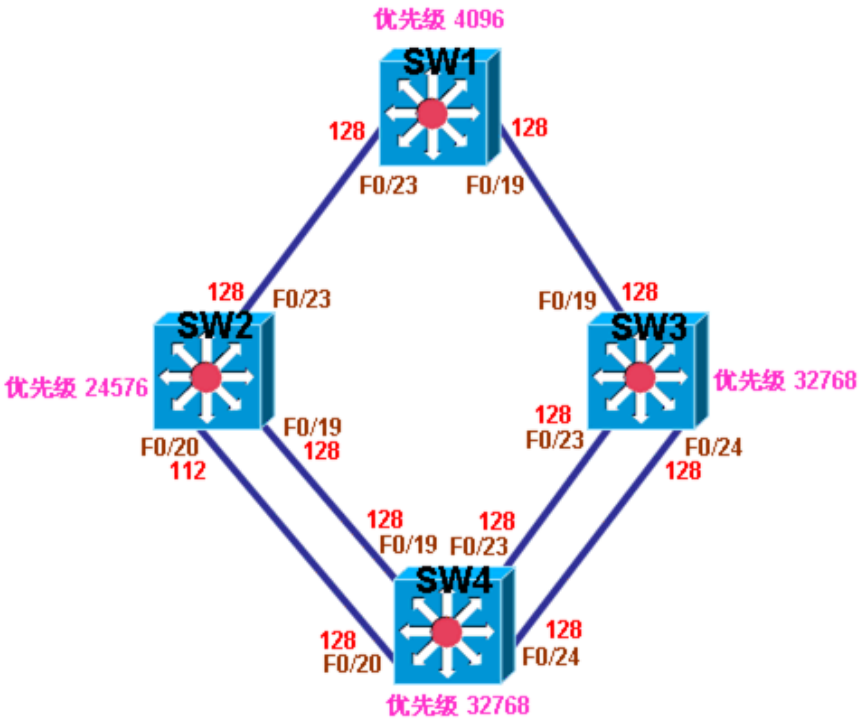


# 1.3.3 交换机生成树协议



## 指定端口（Designated port）

- 根交换机上所有的端口都应该是指定端口。
- 在SW3连接SW4的两个link中，同样也是SW3上的两个端口离根交换机最近，所以在这两个link中，选举SW3上的端口为指定端口。
- 在SW2连接SW4的两个link中，同样也是SW2上的两个端口离根交换机最近，所以在这两个link中，选举SW2上的端口为指定端口。



# 1.3.3 交换机生成树协议



总结：

- 根交换机（Root）

SW1

- 根端口（Root Port）

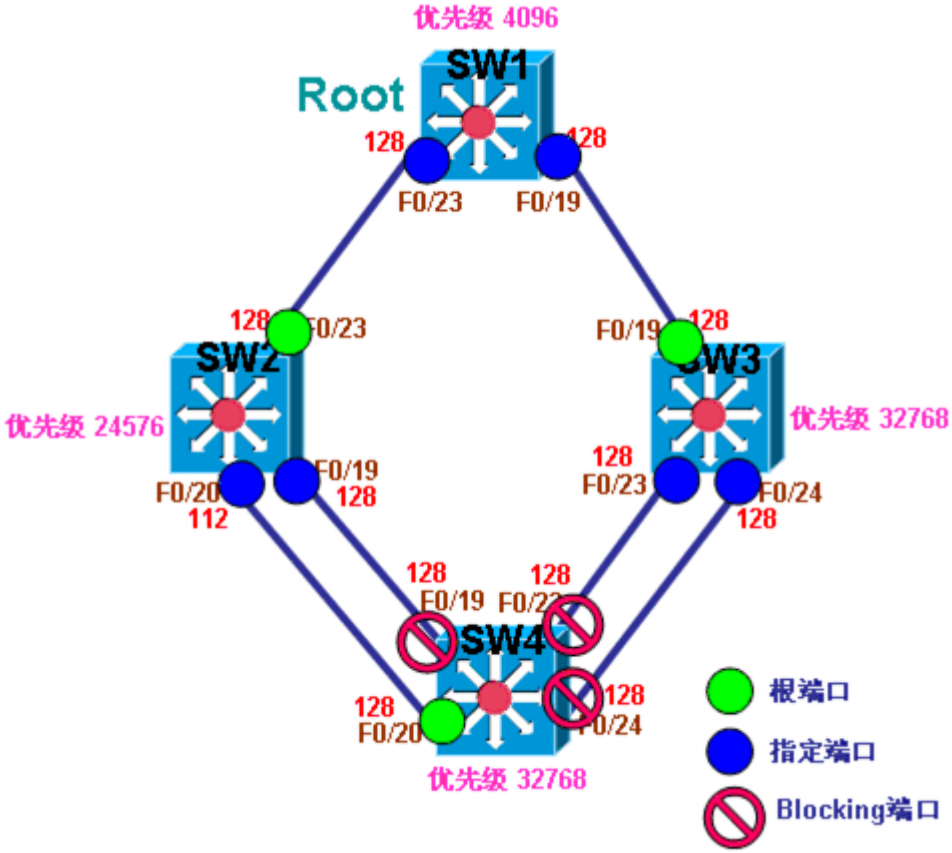
SW2：F0/23　SW3：F0/19　SW4：F0/20

- 指定端口(Designated Port)

SW1：F0/19，F0/23　SW2：F0/19，F0/20　SW3：F0/23，F0/24

- Blocking端口

SW4：F0/19，F0/23，F0/24



# 1.3.3 交换机生成树协议



一个根（Root）交换机上实际的STP状态：

```
Switch#show spanning-tree
VLAN0001
  Spanning tree enabled protocol ieee
  Root ID     Priority    24577
              Address     0001.C9DC.55D5
              This bridge is the root
              Hello Time  2 sec  Max Age 20 sec  Forward Delay 15 sec

  Bridge ID   Priority    24577  (priority 24576 sys-id-ext 1)
              Address     0001.C9DC.55D5
              Hello Time  2 sec  Max Age 20 sec  Forward Delay 15 sec
              Aging Time  20

Interface        Role Sts Cost      Prio.Nbr Type
---------------- ---- --- --------- -------- --------------------------------
Fa0/1            Desg FWD 19        128.1    P2p
Fa0/2            Desg FWD 19        128.2    P2p
```

① RID 构成：priority + Mac

② BID 构成：priority + Mac

③ 路径（带宽）花费

④ PID （Port ID）

# 1.3.3 交换机生成树协议



第二种类型的BPDU包：Topology Change Notification(TCN) BPDU。

1. 当一台交换机检测到拓扑变化后,它就可以发送TCN给root交换机,注意TCN是通过root port向root 交换机方向发出的.

2. 当交换机从它的designate port接收到TCN类BPDU时,它必须为其做转发,从它自已的root port上发送出去TCN类型的BPDU包,这样一级一级地传到root bridge后,TCN的任务才算完成.

3. 在以上的过程中,无论是哪台交换机从它的designate port上收到了TCN类型的BPDU包,它都必须给一个回复,必须从designate port上发出TCA位被置1的normal configuration BPDU包

# 1.3.3 交换机生成树协议



4. 那么当TCN传遍全网,直至到达ROOT BRIDGE后,root bridge也要做出一种回应,它会发出一个正常的configuration BPDU包,当然会有一些不同,就是包内的TC字段会被置1,TC即topology change,表示发现拓扑变化。

5. 这个包会被所有交换机转发,同样的TC位会置1,直至传遍全网,所有交换机都得知拓扑变化为止，原来转发表作废，重新开始选举。

# 1.3.3 交换机生成树协议



其他问题和改进：

- 只有管理员配置了STP的端口才加入STP，直接跟用户连接的端口（边缘端口，不需要参加）
- 如何加快STP的选举速度？
- 如何根据VLAN虚拟局域网进行优化？