HCRSE101-LAN 技术

LAN 技术知识点：

MAC 地址基础，MAC 地址分类，MAC 地址表分类，端口安全，MAC 地址漂移，ARP 协议，免费 ARP ，MSTP，iStack 堆叠，CSS 集群，链路聚合

MAC 地址

MAC 基础

（1）定义: MAC（Media Access Control）地址用来定义网络设备的位置。

（2）组成：MAC 地址由 48 比特长、12 位的 16 进制数字组成，其中从左到右开始，0 到 23bit 是厂商向 IETF 等机构申请用来标识厂商的代码，24 到 47bit 由厂商自行分派，是各个厂商制造的所有网卡的一个唯一编号。

（3） MAC 地址分类：

① 单播（物理）MAC 地址：MAC 地址的第 8 bit 为 0 为单播 MAC 地址，这种类型的 MAC 地址唯一的标识了以太网上的一个终端，该地址为全球唯一的硬件地址。

② 广播 MAC 地址：全 1 的 MAC 地址（FF-FF-FF-FF-FF-FF），用来表示 LAN 上的所有终端设备。

③ 组播 MAC 地址：除广播地址外，第 8 bit 为 1 的 MAC 地址为组播 MAC 地址（例如 01-00-00-00-00-00），用来代表 LAN 上的一组终端。

（4）MAC 地址表：记录了相连设备的 MAC 地址、接口号以及所属的 VLAN ID 之间的对应关系。在转发数据时，路由设备根据报文中的目的 MAC 地址和 VLANID 查询 MAC 地址表，快速定位出接口，从而减少广播。

MAC 地址表的分类：

① 动态表项：由接口通过接收到报文中的源 MAC 地址学习获得，表项有 300s 老化时间。
② 静态表项：由用户手工配置，并下发到各接口板，表项不老化。
③ 黑洞表项：用于指示丢弃含有特定源 MAC 地址或目的 MAC 地址的数据帧，由用户手工配置，并下发到各接口板，表项不老化。
（5）MAC 地址表项的生成方式：自动生成、手工配置。

端口安全

端口安全（Port Security）通过将接口学习到的动态 MAC 地址转换为安全 MAC 地址（包括安全动态 MAC、安全静态 MAC 和 Sticky MAC），阻止非法用户通过本接口和交换机通信，从而增强设备的安全性。

安全动态 MAC 地址：使能端口安全而未使能 Sticky MAC 功能时转换的 MAC 地址。
安全静态 MAC 地址：使能端口安全时手工配置的静态 MAC 地址。
Sticky MAC 地址：使能端口安全后又同时使能 Sticky MAC 功能后转换到的 MAC 地址。

端口安全的保护动作：
Restrict：丢弃源 MAC 地址不存在的报文并上报告警。推荐使用 restrict 动作。
Protect：只丢弃源 MAC 地址不存在的报文，不上报告警。
Shutdown：接口状态被置为 error-down，并上报告警。

MAC 地址漂移

设备上一个接口学习到的 MAC 地址在同一 VLAN 中另一个接

口上也学习到，后学习到的 MAC 地址表项覆盖原来的表项。

产生原因：
① 网络中产生环路：　②非法用户进行网络攻击：

影响：
① 导致数据帧无法正常转发到目的，出现大量的丢包；②消耗设备的性能；

MAC 地址漂移检测：
① 基于 VLAN 的 MAC 地址漂移检测：
可以检测指定 VLAN 下的所有的 MAC 地址是否发生漂移。
当系统检测到 VLAN 内有 MAC 地址发生漂移时，可以进行以下处理动作：
1.接口阻断或 MAC 地址阻断。当检测到 MAC 地址发生漂移则执行接口阻断或 MAC 地址阻断动作。
2.发送告警。当检测到 MAC 地址发生漂移时只给网管发送告警。
② 全局 MAC 地址漂移检测：可以检测到设备上所有的 MAC 地址是否发生了漂移。（默认开启）
缺省情况下，MAC 地址漂移检测的安全级别为 middle，即 MAC 地址发生 10 次迁移后，系统认为发生了 MAC 地址漂移。
当系统统检测到是该接口学习的 MAC 发生漂移，可以配置漂移后处理动作为将该接口关闭或者退出 VLAN，默认情况下可以发出告警。某些特定场景下，如交换机连接双网卡的负载分担服务器时，可能会出现服务器的 MAC 地址在两个接口上学习到，而这种情况不需要作为 MAC 地址漂移被检测出来。可以将服务器所在的 VLAN 加入 MAC 地址漂移检测白名单，不对该 VLAN 进行检测。

VLAN 接口视图下：
loop-detect eth-loop alarm-only：当检测都该 VLAN 内发生 MAC 地址漂移时，发出警告。
loop-detect eth-loop block-time 100 retry-times 3：
当检测到该 VLAN 内发生 MAC 地址漂移时，被检测到的物理接口将被阻塞 100S，100S 重新开放该物理接口，开放后如果在 20S 没有再次检测到 MAC 地址漂移，则该物理接口阻塞将被彻底解除；如果 20S 内再次检测到 MAC 地址漂移，则再次将该接口阻塞。重复以上操作 3 次，如果仍然能检测到该物理接口有 MAC 地址漂移现象发生，则永久阻塞该物理接口。

MAC 地址防漂移：
a)配置接口 MAC 地址学习优先级：
int g0/0/1
mac-learning priority 3
不同接口学到相同的 MAC 地址表项，那么高优先级接口学到的 MAC 地址表项可以覆盖低优先级接口学到的 MAC 地址表项。默认情况下接口 MAC 地址学习优先级 0，优先级可调整范围 0-3。

b)配置不允许相同优先级接口 MAC 地址覆盖。
系统视图下
undo mac-learning priority 3 allow-flapping

c)配置 MAC-spoofing-defend 功能:
系统视图下或接口下
mac-spoofing-defend enabled
将网络侧接口配置为信任接口，该接口学习到的 MAC 地址在其他接口将不会再学习到，可以防止用户侧仿冒网络侧服务器 MAC 地址发送报文。

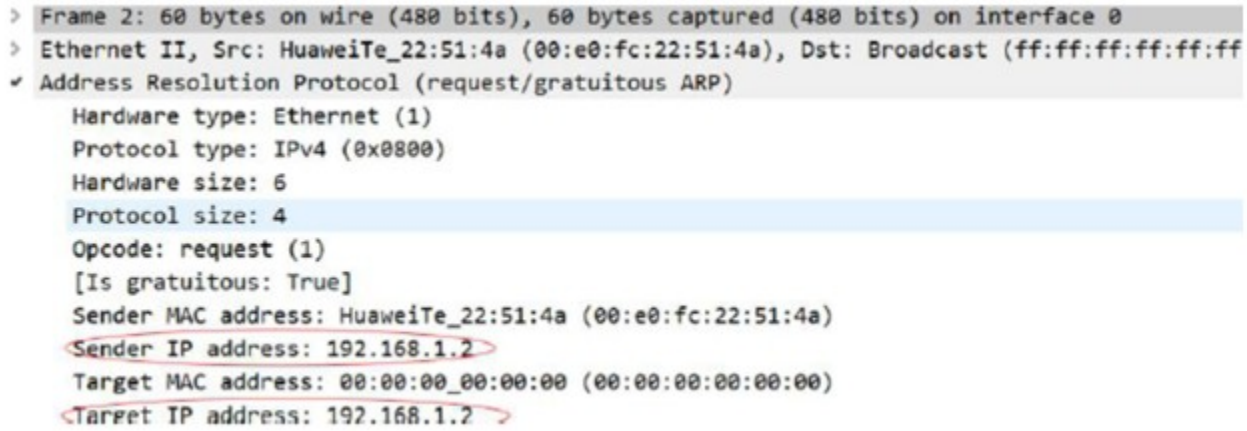
d)配置 STP、Smart-Link 等二层破环协议。
e)配置 IPSG 功能：
基于绑定表（DHCP 动态和静态绑定表）对 IP 报文和源 MAC 地址进行匹配检查。当设备在转发 IP 报文时，将此 IP 报文中的源 IP、源 MAC、接口、VLAN 信息和绑定表的信息进行比较，如果信息匹配，表明是合法用户，则允许此报文正常转发，否则认为是攻击报文，并丢弃该 IP 报文。

IP 源防护（IP Source Guard，简称 IPSG）是一种基于 IP/MAC 的端口流量过滤技术，它可以防止局域网内的 IP 地址欺骗攻击。IPSG 能够确保第 2 层网络中终端设备的 IP 地址不会被劫持，而且还能确保非授权设备不能通过自己指定 IP 地址的方式来访问网络或攻击网络导致网络崩溃及瘫痪。

ARP 地址解析协议
ARP 协议：根据目的的 IP 地址解析目的 MAC 地址。

ARP :　　ip 转换 mac
RARP :　mac 转 ip
InARP　: DLCI 转 ip　　InARP（Inverse ARP）逆向地址解析协议

免费 ARP：使用自己的 IP 地址作为目的 IP 地址发送 ARP 请求



免费 ARP 有如下作用：

(1)IP 地址冲突检测：当设备接口的协议状态变为 Up 时，设备主动对外发送免费 ARP 报文。正常情况下不会收到 ARP 应答，如果收到，则表明本网络中存在与自身 IP 地址重复的地址。如果检测到 IP 地址冲突，设备会周期性的广播发送免费 ARP 应答报文，直到冲突解除。

(2)用于通告一个新的 MAC 地址：发送方更换了网卡，MAC 地址变化了，为了能够在动态 ARP 表项老化前通告网络中其他设备，发送方可以发送一个免费 ARP。

(3)在 VRRP 备份组中用来通告主备发生变换：发生主备变换后，MASTER 设备会广播发送一个免费 ARP 报文来通告发生了主备变换。

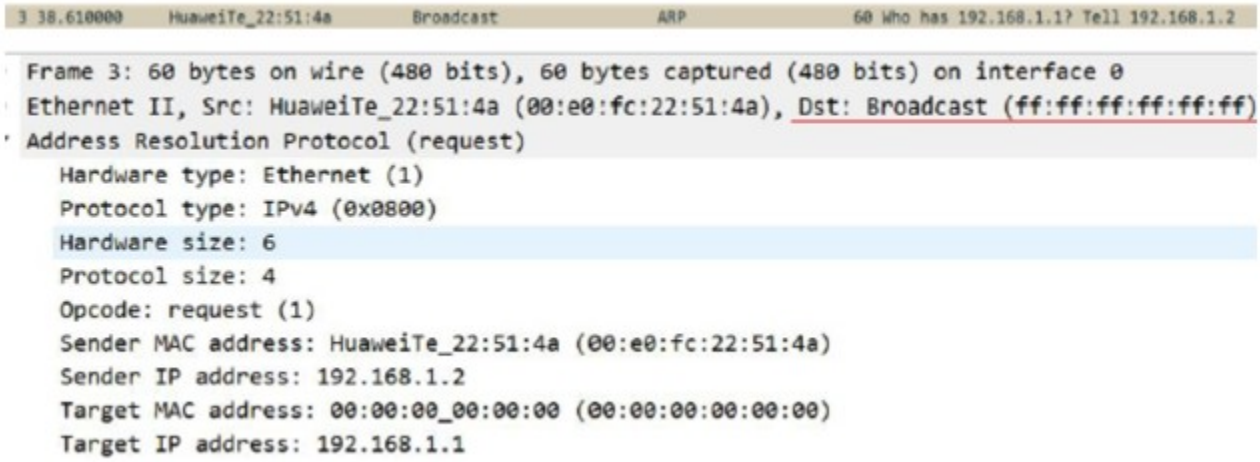
触发条件： 当需要访问目的 IP 地址在 ARP 缓存表不存在对应表项时；

ARP 报文：

①arp request(一般为广播发送)：

a)当访问的目的 IP 地址为同一网段时，请求的访问目的 IP 地址对应的 MAC 地址；

b)当访问的目的 IP 地址不在同一网段时，请求网关 IP 对应的 MAC 地址；

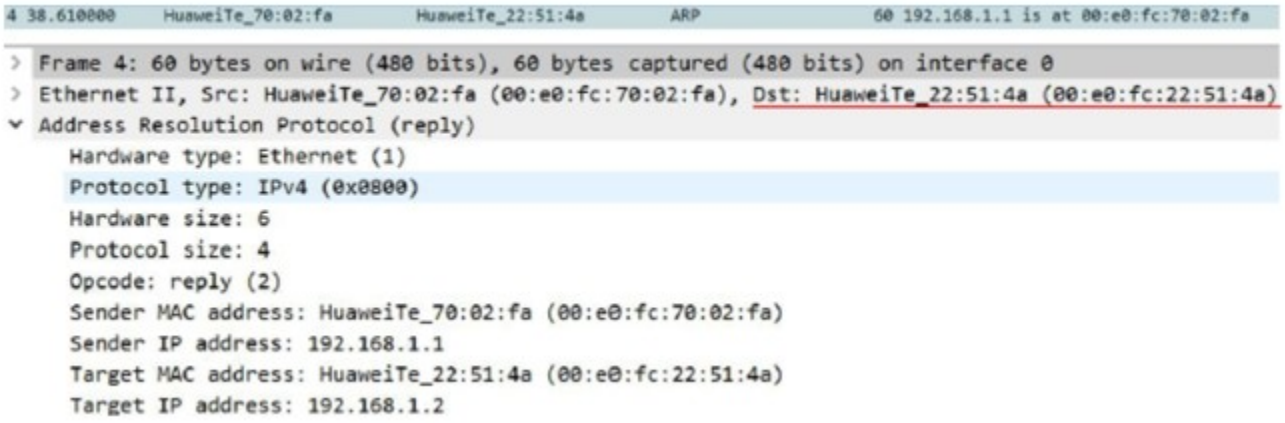


②arp reply（单播回复）

a)当收到请求报文，目的 IP 地址与接收接口的 IP 地址一致时，回复 arpreply，包含自己接口 IP 地址与 MAC 地址的对应关系；

b)当收到请求报文，目的 IP 地址与接收接口的 IP 地址不一致时，如果没有开启 ARP proxy 功能，不会回复任何报文；

c)如果开启 ARP proxy 功能，满足一定的条件之后回复 arp reply，包含请求的目的 IP 地址与自己接口 MAC 地址的对应关系；



ARP 表项：一般 ARP 表项包括动态 ARP 表项和静态 ARP 表项。

① 动态 ARP：

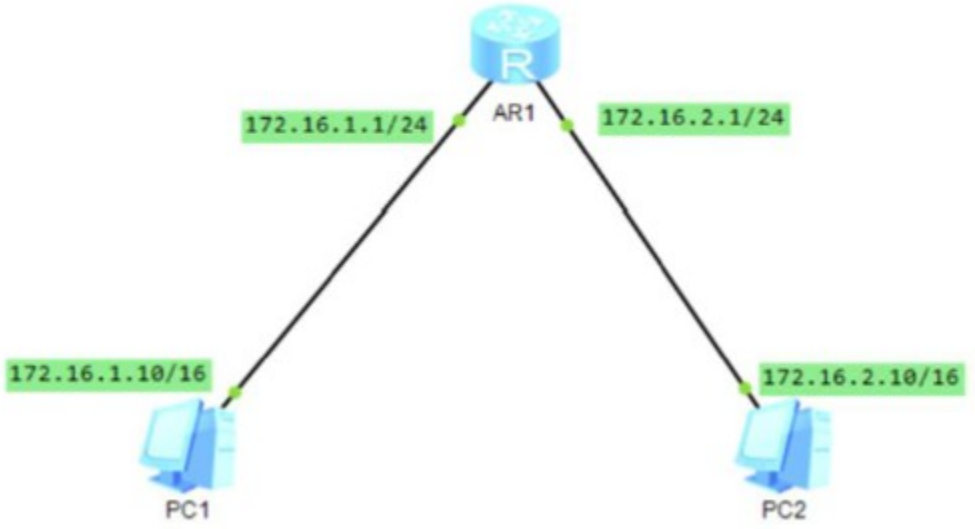
动态 ARP 表项由 ARP 协议通过 ARP 报文自动生成和维护，可以被老化，可以被新的 ARP 报文更新，可以被静态 ARP 表

项覆盖。
动态 ARP 老化机制：
动态 ARP 表项的老化参数有：老化超时时间、老化探测次数和老化探测模式；
① 缺省情况下，动态 ARP 表项的老化超时时间为 1200 秒，即 20 分钟；
② 缺省情况下，动态 ARP 表项的老化探测次数为 3 次。
③ 缺省情况下，接口只在最后一次发送 ARP 老化探测报文是广播方式，其余均为单播方式发送。
② 静态 ARP
静态 ARP 表项是由网络管理员手工建立的 IP 地址和 MAC 地址之间固定的映射关系。
静态 ARP 表项不会被老化，不会被动态 ARP 表项覆盖。

ARP 代理 Proxy ARP

实现在相同网段，但是不在同一物理网络的主机能互访。

① 路由式 Proxy ARP：需要互通的主机（主机上没有配置缺省网关）处于相同的网段但不在同一物理网络（即不在同一广播域）的场景。



当 PC1 需要与 PC2 通信时，由于目的 IP 地址与本机的 IP 地址为同一网段，因此 PC1 以广播形式发送 ARP 请求报文。
Router 启用路由式 Proxy ARP 后，Router 收到 ARP 请求报文后，Router 会查找路由表。由于 PC2 与 Router 直连，因此 Router 上存在到 PC2 的路由表项。
Router 使用自己的 MAC 地址给 PC1 发送 ARP 应答报文。PC1 将以 Router 的 MAC 地址进行数据转发。

```
int g0/0/1
arp-proxy enable
```

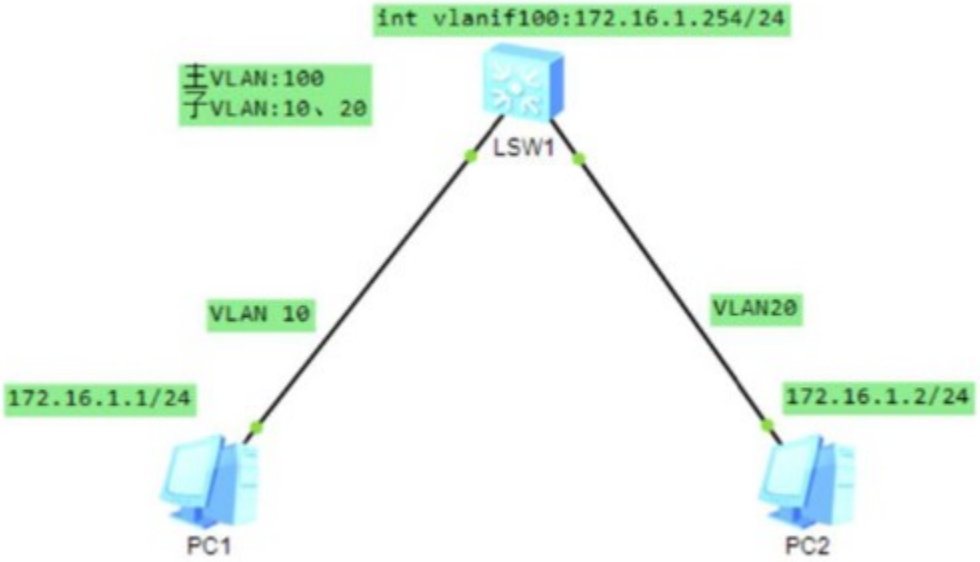
②VLAN 内 Proxy ARP：需要互通的主机处于相同网段，并且属于相同 VLAN，但是 VLAN 内配置了端口隔离的场景。

由于在 Router 上配置了 VLAN 内不同接口彼此隔离，因此 PC1 和 PC2 不能直接在二层互通。
若 Router 的接口使能了 VLAN 内 Proxy ARP 功能，可以使 PC1 和 PC2 实现三层互通。Router 的接口在接收到目的地址不是自己的 ARP 请求报文后，Router 并不立即丢弃该报文，而是查找该接口的 ARP 表项。如果存在 PC2 的 ARP 表项，则将自己的 MAC 地址通过 ARP 应答报文发送给 PC1，并将 PC1 发送给 PC2 的报文代为转发。

int g0/0/1
port-isolate enable group 1

vlan 10
arp-proxy inner-sub-vlan-proxy enable

③VLAN 间 Proxy ARP：需要互通的主机处于相同网段，但属于不同 VLAN 的场景。



由于 PC1 和 PC2 属于不同的 Sub-VLAN，PC1 和 PC2 不能直接实现二层互通。
如果 Router 上使能了 VLAN 间 Proxy ARP 功能，可以使 PC1 和 PC2 实现三层互通。Router 的接口在接收到目的地址不是自己的 ARP 请求报文后，并不立即丢弃该报文，而是查找 ARP 表项。
如果存在 PC2 的 ARP 表项，则将自己的 MAC 地址发送给 PC1，并将 PC1 发送给 PC2 的报文代为转发。

vlan 10
arp-proxy inter-sub-vlan-proxy enable

ARP 攻击与防范

ARP 协议有简单、易用的优点，但是也因为其没有任何安全机制，容易被攻击者利用。在网络中，常见的 ARP 攻击方式主要包括：

（1）ARP 泛洪攻击:

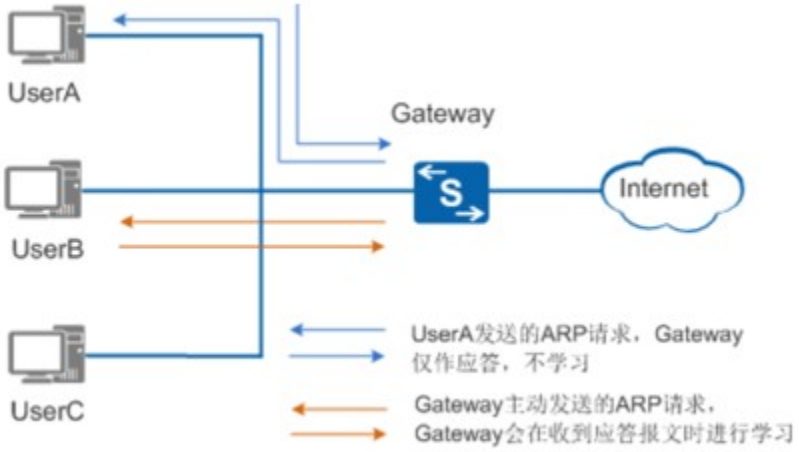
是指攻击者发送大量的 ARP 报文，也叫拒绝服务攻击 DoS（Denial of Service），主要带来以下两种影响：

1 ARP 表项溢出：设备处理 ARP 报文和维护 ARP 表项都需要消耗系统资源，同时为了满足 ARP 表项查询效率的要求，一般设备都会对 ARP 表项规模有规格限制。攻击者就利用这一点，通过伪造大量源 IP 地址变化的 ARP 报文，使得设备 ARP 表资源被无效的 ARP 条目耗尽，合法用户的 ARP 报文不能继

续生成 ARP 条目，导致正常通信中断。

防范：

a）ARP 表项严格学习：



在网关设备上部署 ARP 表项严格学习功能，ARP 表项严格学习是指只有本设备主动发送的 ARP 请求报文的应答报文才能触发本设备学习 ARP，其他设备主动向本设备发送的 ARP 报文不能触发本设备学习 ARP，可拒绝大部分的 ARP 报文攻击。

arp learning strict

b）ARP 表项限制：ARP 表项限制功能应用在网关设备上，可以限制设备的某个接口学习动态 ARP 表项的数目。默认状态下，接口可以学习的动态 ARP 表项数目规格与全局的 ARP 表项规格保持一致。当部署完 ARP 表项限制功能后，如果指定接口下的动态 ARP 表项达到了允许学习的最大数目，将不再允许该接口继续学习动态 ARP 表项，以保证当一个接口所接入的某一用户主机发起 ARP 攻击时不会导致整个设备的 ARP 表资源都被耗尽。

```
int g0/0/1
arp-limit vlan 10 maximum 20
```

2 ARP MISS：攻击者利用工具扫描本网段主机或者进行跨网段扫描时，会向设备发送大量目标 IP 地址不能解析的 IP 报文（即路由表中存在该 IP 报文的目的 IP 对应的路由表项，但设备上没有该路由表项中下一跳对应的 ARP 表项），导致设备触发大量 ARP Miss 消息，生成并下发大量临时 ARP 表项，并
广播大量 ARP 请求报文以对目标 IP 地址进行解析，从而造成 CPU 负荷过重。

防范：

a）根据源 IP 地址进行 ARP Miss 消息限速：当设备检测到某一源 IP 地址的 IP 报文在 1 秒内触发的 ARP Miss 消息数量超过了 ARP Miss 消息限速值，就认为此源 IP 地址存在攻击。

```
arp speed-limit source-ip 10.2.2.1 maximum 10
```

b）针对全局的 ARP Miss 消息限速：设备支持对全局处理的 ARP Miss 消息数量，根据限速值进行限速。

（2）ARP 欺骗攻击

是指攻击者通过发送伪造的 ARP 报文，恶意修改设备或网络

内其他用户主机的 ARP 表项，造成用户或网络的报文通信异常。主要存在这样两种场景：

1 欺骗网关攻击：Attacker 仿冒 UserA 向 Gateway 发送伪造的 ARP 报文，导致 Gateway 的 ARP 表中记录了错误的 User A 地址映射关系，造成 UserA 接收不到正常的数据报文。

防范：

a) ARP 表项固化：网关设备在第一次学习到 ARP 以后，不再允许用户更新此 ARP 表项或只能允许更新此 ARP 表项的部分信息，或者通过发送单播 ARP 请求报文的方式对更新 ARP 条目的报文进行合法性确认。

2 仿冒网关攻击：攻击者 B 将伪造网关的 ARP 报文发送给用户 A，使用户 A 误以为攻击者即为网关。用户 A 的 ARP 表中会记录错误的网关地址映射关系，使得用户 A 跟网关的正常数据通信中断

防范：

a)ARP 防网关冲突功能：当设备收到的 ARP 报文，发现 ARP 报文的源 IP 地址与报文入接口对应的 VLANIF 接口的 IP 地址相同。设备就认为该 ARP 报文是与网关地址冲突的 ARP 报文，设备将生成 ARP 防攻击表项，并在后续一段时间内丢弃该接口收到的同 VLAN 以及同源 MAC 地址的 ARP 报文，这样可以防止与网关地址冲突的 ARP 报文在 VLAN 内广播。（此时网关是三层交换机的 VLANIF 接口）

b) 发送免费 ARP 报文功能：定期广播发送正确的免费 ARP 报文到所有用户，迅速将已经被攻击的用户记录的错误网关地址映射关系修改正确。

3 中间人攻击：攻击者主动向 PC1 发送伪造 PC3 的 ARP 报文，导致 PC1 的 ARP 表中记录了错误的 PC3 地址映射关系，攻击者可以轻易获取到 PC1 原本要发往 PC3 的数据；同样，攻击者也可以轻易获取到 PC3 原本要发往 PC1 的数据。这样，PC1 与 PC3 间的信息安全无法得到保障。

防范：

a）动态 ARP 检测（DAI）：动态 ARP 检测是利用绑定表来防御中间人攻击的。当设备收到 ARP 报文时，将此 ARP 报文对应的源 IP、源 MAC、VLAN 以及接口信息和绑定表的信息进行比较，如果信息匹配，说明发送该 ARP 报文的用户是合法用户，允许此用户的 ARP 报文通过，否则就认为是攻击，丢弃该 ARP 报文。

说明：动态 ARP 检测功能仅适用于 DHCP Snooping 场景。设备使能 DHCP Snooping 功能后，当 DHCP 用户上线时，设备会自动生成 DHCPSnooping 绑定表；对于静态配置 IP 地址的用户，设备不会生成 DHCPSnooping 绑定表，所以需要手动添加静态绑定表，否则无法正常通信。

普通 ARP 和免费 ARP 的区别：

普通 ARP 请求报文(查找别人的 IP 地址，比如：我需要 10.1.1.2 的 MAC 地址,10.1.1.2 是别人的 IP)广播发送出去，广播域内所有主机都 接收到，计算机系统判断 ARP 请求报文中的目的 IP 地址字段，如果发现和本机的 IP 地址相同，则将自己的 MAC 地址填写到该报文的目的 MAC 地址字段，并 将该报文发回给源主机。所以只要发送普通 ARP 请求的主机接收到报文，则证明广播域内有别的主机使用自己要访问的这个 IP 地址（这里不考虑路由器的 ARP 代理问 题）。

免费 ARP 的报文发（查找自己的 IP 地址，比如：我需要 10.1.1.1 的 MAC 地址，而 10.1.1.1 就是自己的 IP）出去是不希望收到回应的，只希望是起宣告作用；如果收到回应，则证明对方也使用自己目前使用的 IP 地址。

==========

CSS + Eth-Trunk + iStack

CSS 与 iStack 的区别在于，一般框式交换机堆叠称为 CSS，盒式交换机堆叠称为 iStack，都可以称为堆叠。两者只是叫法和实现有些差异，但是功能是一样的。

S6720 和以下系列为盒式交换机。
S7700 和以上系列为框式交换机。

在华为交换机中，iStack 最多支持 9 台交换机合并，而在 CSS 中只支持 2 台交换机合并。
iStack，全称 Intelligent Stack，智能堆叠，适用于 S2700、S3700、S5700 和 S6700 中低端交换机。

高端交换机中叫做 CSS，全称 Cluster Switch System，集群交换系统，适用于 S7700、S9300、S9700 等高端交换机。
此类技术原理是将多台物理交换机在逻辑上合并成一台交换机，所以也叫做交换机虚拟化。

园区网络 CSS+iStack 组网之一，其主要有简单、高效、可靠的特点。
(1)简单
各层设备均使用堆叠技术，逻辑设备少，网络拓扑简单，二层天然无环，无需部署 xSTP 破环协议。
(2)高效
各层设备间使用 Eth-Trunk 链路聚合技术，负载分担算法灵活，链路利用率高。
(3)可靠
服务器和主机可以配置多 NIC 网卡 Teaming 负载均衡或主备

冗余链路提高服务器接入可靠性。
堆叠技术同链路聚合技术结合使用，各层物理设备形成双归接入组网，提高整网可靠性。
缺点
对设备性能要求较高，盒式设备堆叠台数过多，可能导致堆叠主的主控性能下降。
如果采用业务口堆叠或集群，会占用业务端口数。

=====

iStack (Intelligent Stack)智能堆叠

iStack (Intelligent Stack)智能堆叠 ，是指将多台支持堆叠特性的交换机设备组合在一起，从逻辑上组合成一台交换设备。

堆叠的作用：
1 高可靠性：堆叠是将物理上相互独立的两台设备，变成逻辑上的一台设备。所以，当其中一台物理设备故障，另一台设备将会接替工作。
2 强大的网络扩展能力：如果觉得两台设备堆叠无法满足网络的需求，那么可以继续向堆叠系统中，加入交换机。让原本两台交换机的合体，变为三台甚至更多台的交换机的合体。
3 简化配置和管理：因为多台设备成为了逻辑上的一台设备，所以我们在对设备进行控制时，只需登录到其中一台设备进行配置即可。

堆叠基本概念
1、 角色
堆叠中的单台交换机称为成员交换机，按照功能不同可以分为以下角色：
主交换机：主交换机（Master）负责管理整个堆叠。堆叠中

只有一台主交换机。
备交换机：备交换机（Standby）是主交换机的备份交换机。当主交换机故障时，备交换机会接替原主交换机的所有业务。堆叠中只有一台备交换机。
从交换机：从交换机（Slave）主要用于业务转发，从交换机数量越多，堆叠系统的转发能力越强。除主交换机和备交换机外，堆叠中其他所有的成员交换机都是从交换机。

2、堆叠域
交换机通过堆叠链路连接在一起组成一个堆叠，这些成员交换机的集合就是一个堆叠域。为了适应各种组网应用，同一个网络里可以部署多个堆叠，堆叠之间使用域编号（Domain ID）来进行区别。使其不与网络中其他堆叠系统的域编号冲突。

2、堆叠成员 ID
堆叠成员 ID，即堆叠成员交换机的编号（Member ID），用来标识和管理成员交换机。堆叠中所有成员交换机的堆叠成员 ID 都是唯一的。

3、堆叠优先级
堆叠优先级是成员交换机的一个属性，主要用于角色选举过程中确定成员交换机的角色，优先级值越大表示优先级越高，当选为主交换机的可能性越大。缺省情况下，成员交换机的堆叠优先级为 100，最大值为 255
堆叠优先级：状态相同，优先级高的设备成为堆叠主。
MAC 地址大小：状态、优先级都相同，MAC 地址小的设备成为堆叠主。

主交换机选举
运行状态比较，已经运行的交换机比处于启动状态的交换机优

先竞争为主交换机。
堆叠优先级高的交换机优先竞争为主交换机。
堆叠优先级相同时，MAC 地址小的交换机优先竞争为主交换机。

4、堆叠物理成员端口
堆叠物理成员端口，即被配置为堆叠模式的物理端口，用于堆叠成员交换机之间的连接。

5、堆叠端口
堆叠端口是一种专用于堆叠的逻辑端口，需要和堆叠物理成员端口绑定。一个堆叠端口可以与一个或多个堆叠物理成员端口绑定，以提高链路的带宽和可靠性。
每台设备支持两个堆叠端口，为 Stack-Portn/1 和 Stack-Portn/2，其中 n 为设备的堆叠成员 ID。

系统自动完成堆叠细分为三步：
主交换机选举
拓扑收集和备交换机选举
稳定运行

堆叠连接方式：堆叠卡堆叠和业务口堆叠

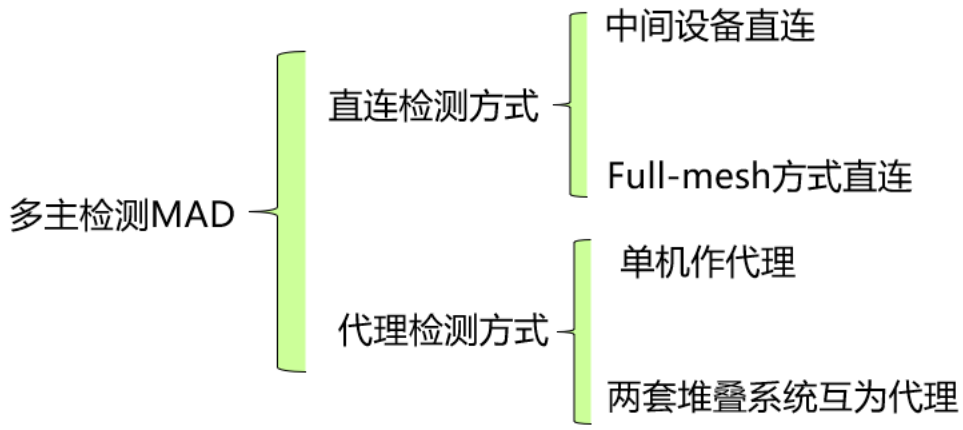
堆叠管理：
成员加入（从交换机）
堆叠合并（竞争主交换机）
成员退出
堆叠分裂

多主检测 MAD（Multi-Active Detection）

由于堆叠系统中所有成员交换机都使用同一个 IP 地址和 MAC 地址（堆叠系统 MAC），一个堆叠分裂后，可能产生多个具有相同 IP 地址和 MAC 地址的堆叠系统。

为防止堆叠分裂后，产生多个具有相同 IP 地址和 MAC 地址的堆叠系统，引起网络故障，必须进行 IP 地址和 MAC 地址的冲突检查。

多主检测 MAD（Multi-Active Detection），是一种检测和处理堆叠分裂的协议。链路故障导致堆叠系统分裂后，MAD 可以实现堆叠分裂的检测、冲突处理和故障恢复，降低堆叠分裂对业务的影响。



MAD 检测方式有两种：
直连检测方式和代理检测方式。

与直连检测方式相比，代理检测方式无需占用额外的接口，Eth-Trunk 接口可同时运行 MAD 代理检测和其他业务。

直连检测方式
直连检测的连接方式包括通过中间设备直连和堆叠成员交换机

Full-mesh 方式直连：
在同一个堆叠系统中，两种检测方式互斥，不可以同时配置。
直连检测方式是指堆叠成员交换机间通过普通线缆直连的专用链路进行多主检测。在直连检测方式中，堆叠系统正常运行时，不发送 MAD 报文；堆叠系统分裂后，分裂后的两台交换机以 1s 为周期通过检测链路发送 MAD 报文以进行多主冲突处理。
通过中间设备直连：堆叠系统的所有成员交换机之间至少有一条检测链路与中间设备相连。
Full-mesh 方式直连：堆叠系统的各成员交换机之间通过检测链路建立 Full-mesh 全连接，即每两台成员交换机之间至少有一条检测链路。

通过中间设备直连可以实现通过中间设备缩短堆叠成员交换机之间的检测链路长度，适用于成员交换机相距较远的场景。与通过中间设备直连相比，Full-mesh 方式直连可以避免由中间设备故障导致的 MAD 检测失败，但是每两台成员交换机之间都建立全连接会占用较多的接口，所以该方式适用于成员交换机数目较少的场景。

中间设备直连：占用接口少，适用于距离远的场景 ，缺点：多一台设备， 中间设备问题会检测失败
Full-mesh： 不需要中间设备，适用于交换机少的场景，缺点：占用较多的接口

代理检测方式
根据代理设备的不同，代理检测方式可分为单机作代理和两套堆叠系统互为代理。
代理检测方式是在堆叠系统 Eth-Trunk 上启用代理检测，在代理设备上启用 MAD 检测功能。此种检测方式要求堆叠系统中的所有成员交换机都与代理设备连接，并将这些链路加入同一

个 Eth-Trunk 内。与直连检测方式相比，代理检测方式无需占用额外的接口，Eth-Trunk 接口可同时运行 MAD 代理检测和其他业务。
在代理检测方式中，堆叠系统正常运行时，堆叠成员交换机以 30s 为周期通过检测链路发送 MAD 报文。堆叠成员交换机对在正常工作状态下收到的 MAD 报文不做任何处理；堆叠分裂后，分裂后的两台交换机以 1s 为周期通过检测链路发送 MAD 报文以进行多主冲突处理。

MAD 冲突处理

堆叠分裂后，MAD 冲突处理机制会使分裂后的堆叠系统处于 Detect 状态或 Recovery 状态。
Detect 状态表示堆叠正常工作状态，Recovery 状态表示堆叠禁用状态。
MAD 冲突处理机制如下：MAD 分裂检测机制会检测到网络中存在多个处于 Detect 状态的堆叠系统，这些堆叠系统之间相互竞争，竞争成功的堆叠系统保持 Detect 状态，竞争失败的堆叠系统会转入 Recovery 状态；并且在 Recovery 状态堆叠系统的所有成员交换机上，关闭除保留端口以外的其它所有物理端口，以保证该堆叠系统不再转发业务报文。

MAD 故障恢复

通过修复故障链路，分裂后的堆叠系统重新合并为一个堆叠系统。重新合并的方式有以下两种：
堆叠链路修复后，处于 Recovery 状态的堆叠系统重新启动，与 Detect 状态的堆叠系统合并，同时将被关闭的业务端口恢复 Up，整个堆叠系统恢复。
如果故障链路修复前，承载业务的 Detect 状态的堆叠系统也出现了故障。此时，可以先将 Detect 状态的堆叠系统从网络中移除，再通过命令行启用 Recovery 状态的堆叠系统，接替

原来的业务，然后再修复原 Detect 状态堆叠系统的故障及链路故障。故障修复后，重新合并堆叠系统。

=====

集群交换机系统 CSS (Cluster Switch System)

CSS 是 Cluster Switch System 的简称，又被称为集群交换机系统（简称为 CSS 或堆叠），是将 2 台交换机通过专用的堆叠电缆链接起来，对外呈现为一台逻辑交换机。

CSS 特征

交换机多虚一：堆叠交换机对外表现为一台逻辑交换机，控制平面合一，统一管理。
转发平面合一：堆叠内物理设备转发平面合一，转发信息共享并实时同步。
跨设备链路聚合：跨堆叠内物理设备的链路被聚合成一个 Eth-Trunk 端口，和下游设备实现互联。

S9300

基本概念：

不同于 iStack 可以多台设备堆叠，对于 CSS 集群，集群中只能有一主一备两台交换机。

主交换机：负责管理整个集群。集群中只有一台主交换机。
备交换机：主交换机的备份交换机。当主交换机故障时，备交换机会接替原主交换机的所有业务。集群中只有一台备交换机。
集群 ID：即 CSS ID，用来标识和管理成员交换机。集群中成员交换机的集群 ID 是唯一的。
集群优先级：即 Priority，是成员交换机的一个属性，主要用于角色选举过程中确定成员交换机的角色，优先级值越大表示优先级越高，优先级越高当选为主交换机的可能性越大。集群优先级缺省为 1

角色选举

最先完成启动，并进入单框集群运行状态的交换机成为主交换机。
当两台交换机同时启动时，集群优先级高的交换机成为主交换机。
当两台交换机同时启动，且集群优先级又相同时，MAC 地址小的交换机成为主交换机。
当两台交换机同时启动，且集群优先级和 MAC 都相同时，集群 ID 小的交换机成为主交换机。

设备组建集群有两种连接方式，分别为集群卡集群和业务口集群
集群卡集群方式：集群成员交换机之间通过主控板上专用的集群卡及专用的集群线缆连接。
业务口集群方式：集群成员交换机之间通过业务板上的普通业务口连接，不需要专用的集群卡。同 iStack，业务口集群一样涉及两种端口的概念：物理成员端口和逻辑集群端口。

集群加入合并
使能了集群功能的单台交换机即为单框集群。
集群成员加入是指向稳定运行的单框集群系统中添加一台新的交换机。原单框集群的交换机成为主交换机，新加入的交换机成为备交换机。

集群加入通常在以下两种情形下出现：
在建立集群时，先将一台交换机使能集群功能后重启，重启后这台交换机将进入单框集群状态。然后再使能另外一台交换机的集群功能后重启，则后启动的交换机则按照集群成员加入的流程加入集群系统，成为备交换机。
在稳定运行的两框集群场景中，将其中一台交换机重启，则这台交换机将以集群成员加入的流程重新加入集群系统，并成为

备交换机。

集群合并

集群合并是指稳定运行的两个单框集群系统合并成一个新的集群系统。如图 2 所示，两个单框集群系统将自动选出一个更优的作为合并后集群系统的主交换机。被选为主交换机的配置不变，业务也不会受到影响，框内的备用主控板将重启。而备交换机将整框重启，以集群备的角色加入新的集群系统，并将同步主交换机的配置，该交换机原有的业务也将中断。

集群合并通常在以下两种情形下出现：
将两台交换机分别使能集群功能后重启（重启后的两台交换机都属于单框集群），再使用集群线缆将两台交换机连接，之后会进入集群合并流程。
集群链路或设备故障导致集群分裂。故障恢复后，分裂后的两个单框集群系统重新合并。

多主检测

直连检测的连接方式包括通过中间设备直连和集群成员交换机直接直连。
代理检测方式代理检测方式可分为单机作代理和两套集群系统互为代理。

==========

Eth-Trunk 以太网链路聚合

Eth-Trunk 在逻辑上把多条物理链路捆绑等同于一条逻辑链路，对上层数据透明传输。所有 Eth-Trunk 中物理接口的参数必须一致，Eth-Trunk 链路两端要求一致的物理参数有：Eth-Trunk 链路两端相连的物理接口类型、物理接口数量、物理接口的

速率、物理接口的双工方式以及物理接口的流控方式。

链路聚合技术主要有以下三个优势：增加带宽、提高可靠性和负载分担

Eth-Trunk 工作模式可以分为两种：
手动负载均衡（manual load-balance）模式
LACP（Link Aggregation Control Protocol）模式

手工负载分担模式：需要手动创建链路聚合组，并配置多个接口加入到所创建的 Eth-Trunk 中
静态 LACP 模式： 该模式通过 LACP 协议协商 Eth-Trunk 参数后自主选择活动接口。

系统 lacp 优先级和接口 lacp 优先级:
用于选择主动端和被动端,系统 lacp 优先级越低越好,默认是 32768;如果优先级一样则选择 mac 小的;
接口 lacp 优先级用于确定活动链路,默认也是 32768,优先值越低,lacp 优先级越高.

注意事项:
每个 eth-trunk 下最多可以支持 8 个成员端口;
成员接口下不能配置任何业务和静态 MAC 地址;
成员接口加入 Eth-trunk 时,必须为缺省的 hybrid 类型接口;
Eth-trunk 不可以嵌套,也就是说成员接口不能是 eth-trunk;
一个以太网接口只能加入到一个 eth-trunk;
一个 eth-trunk 中的成员接口必须相同类型的,比如 FE 口和 GE 口是不属于同一个聚合组的.
可以将不同接口板上的接口加入同一 eth-trunk;
如果本地设备使用了 Eth-Trunk，与成员接口直连的对端接口

也必须捆绑为 Eth-Trunk 接口
 当成员接口速率不一致时,实际使用中速率小的成员接口会出现拥塞,导致丢包;
 当成员接口加入 eth-trunk 后,学习 mac 地址时是按照 eth-trunk 来学习的,而不是按照成员接口来学的;

E-Trunk

Eth-Trunk：一般指同一设备的链路聚合，一台交换机将多个接口捆绑，形成一个 Eth-Trunk 接口，从而实现了增加带宽和提高可靠性的目的。

E-Trunk（Enhanced Trunk）：一般指跨设备链路聚合，是一种实现跨设备链路聚合的机制，基于 LACP（单台设备链路聚合的标准）进行了扩展，能够实现多台设备间的链路聚合。从而把链路可靠性从单板级提高到了设备级。
==========

MST

在 STP 和 RSTP 的算法中，所有 VLAN 共享一课生成树，会造成部分 VLAN 无法通信、次优路径、流量无法负载分担等问题。

为了弥补 STP 和 RSTP 的缺陷，IEEE 于 2002 年发布的 802.1S 标准定义了 MSTP。MSTP 兼容 STP 和 RSTP，既可以快速收敛，又提供了数据转发的多个冗余路径，在数据转发过程中实现 VLAN 数据的负载均衡。

IST（Internal Spanning Tree，内部生成树）是 MST 区域中的一个生成树实例。

CST（Common Spanning Tree，公共生成树）是用来互联 MST 区域的单生成树。
CIST（Common and Internal Spanning Tree，公共和内部生成树）是连接一个交换网络内所有设备的单生成树，由 IST 和 CST 共同构成。从包含的范围来看，IST 是最小的，仅属于一个 MST 区域内部，CST 次之，则是 MST 区域间的互联生成树实例，而 CIST 则最大，包括了 IST 和 CST。

公共生成树 CST 是连接交换网络内所有 MST 域的一棵生成树。如果把每个 MST 域看作是一个节点，CST 就是这些节点通过 STP 或 RSTP 协议计算生成的一棵生成树。
内部生成树 IST（Internal Spanning Tree）是各 MST 域内的一棵生成树。IST 是一个特殊的 MSTI，MSTI 的 ID 为 0，通常称为 MSTI0。
SST（Single Spanning Tree）：运行 STP 或 RSTP 的交换设备只能属于一个生成树；MST 域中只有一个交换设备，这个交换设备构成单生成树。
所有 MST 域的 IST 加上 CST 就构成一棵完整的生成树，即 CIST。

MSTP 同域的三要素就是域名、实例和 vlan 映射、修订级别
缺省情况下，MST 域的 MSTP 修订级别为 0。

==========
STP 协议

STP　：IEEE　802.1d
RSTP：IEEE　802.1w
MSTP：IEEE　802.1s

STP 选举

每个网络只有一个根桥

每个非根桥都要选出一个根端口

每个 Segment 只有一个指定端口

非指定端口将被堵塞

根端口的选举 ：RP

路径开销、对端 BID（发送端 BID）、对端 PID（发送端 PID）和本端 PID （接收端 PID）

指定端口的选举 ：DP

路径开销、本端 BID （接收端 BID ）、本端 PID（ 接收端 PID）

交换机 Bridge ID

BID 是由 16 位的桥 Bridge 优先级(Bridge Priority)与桥 MAC 地址构成的。BID 中优先级占据高 16 位其余的低 48 位是 MAC 2bytes 地址。高 16 位优先级中，低 12 位定义为扩展的 System ID ，在 STP 和 RSTP 中该部分取值为 0 。

BID 越小越优，优先级默认为 32768

stp priority 4096

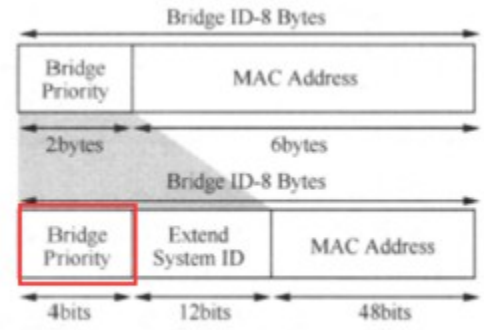
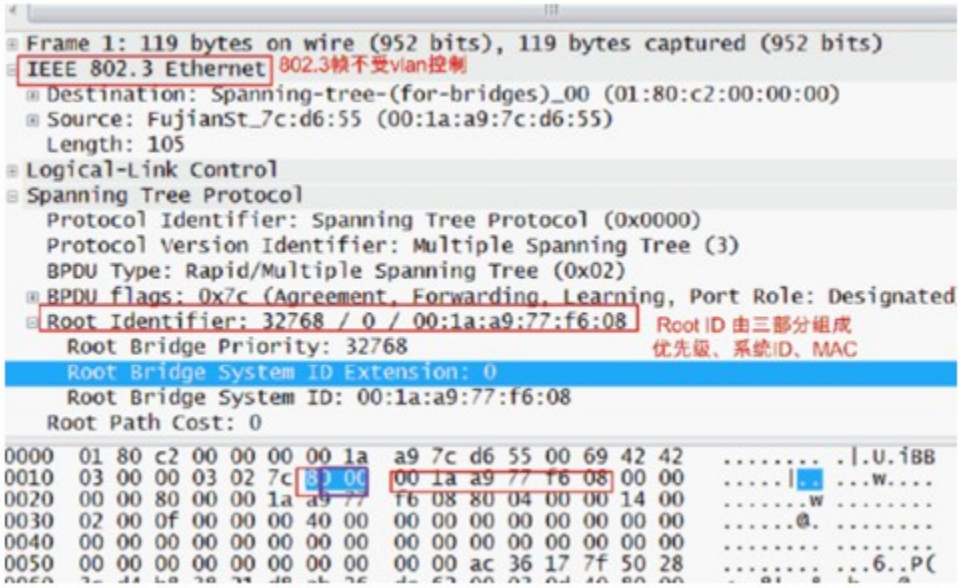


图 9-37 网桥 ID

STP 优先级为什么必须是 4096 的倍数 ，扩展系统 ID 为 12

位，2 的 12 次方是 4096



整个 Root IDentifier 字段一共 16 个十六进制，64 个二进制，8 个字节。

BPDU 对应的 Root ID 为　8000 001a a977 f60

设备读取优先级的时候读的是前四位十六进制，但是 8000 中的后面三个 0 是一般不可更改的。

只有 8 对应的位置才能更改。

所以结果就是：

只有 16 种组合，（因为只有一位十六进制可改动）。

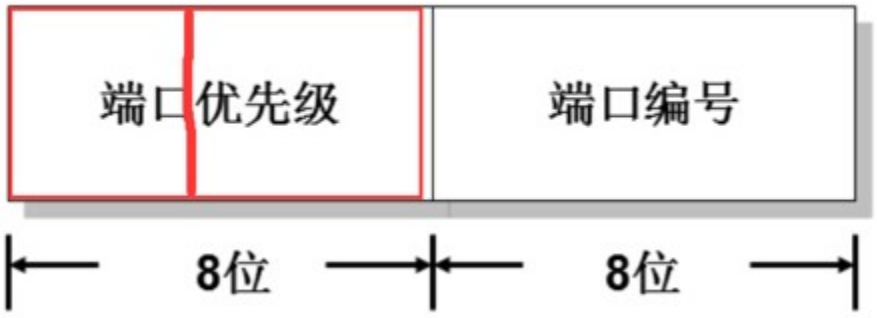
必须为 4096 的倍数，后面带了三个 16 进制的 0，16 的三次方为 4096 。加权计算就是 4096 的倍数了。

Port ID（端口 ID）

端口 PID 的大小可能会影响到是否被选举为指定端口

端口 ID 是 2 字节，其中，端口优先级占 8 字节，端口号占 8 字节。但在配置时端口优先级仅能配置高 4 位，后 12 位当成端口号(端口号系统自己分配，不可调)

PID 由两部分构成的，高 4 位是端口优先级，低 12 位是端口号



PID 越小越优，端口优先级默认为 128 ，最大为 240，数值为 16 的倍数
int g0/0/0
stp port priority 16

端口 ID 总共占 16bit，其中 8 位是端口优先级，8 位是端口编号，所以端口优先级部分的取值范围是 0-255，缺省值为 128。在实际配置交换机的时候，配置端口 ID 的端口优先级的时候却不是这样，这是因为现在的中高端交换机，端口数量有的已经远超 255 个了，所以原来只有 8 位定义端口编号，显然不够了，从端口优先级部分挪了 4 位来定义端口编号，才能确保交换机的每个端口有唯一的编号

端口 cost
交换机每个端口都有自己的端口成本，华为在其交换设备上定义了 3 种端口成本的计算方法，
默认是 IEEE 802.It 标准，并可使用 stp pathcost-standard 命令来修改默认的端口成本的计算方法。

端口成本 默认为 dot1t
stp pathcost-standard 取值范围
dot1d-1998: 1-65535 , g0/0/1 默认为 4
dot1t: 1-200000000 , g0/0/1 默认为 20000
legacy: 1-200000 , g0/0/1 默认为 20

表 9-1 三种方法的默认成本值

	802.1d-1998 标准	802.1t	华为实现
10M ***	100	2,000,000	200,000
100M ***	19	200,000	200
1000M	4	20,000	20
10G	2	2000	2
40G	1	500	1

disable： 未启用 STP
blocking： 接收 BPDU 20
listening: 收发 BPDU 15
learning: 收发 BPDU，学习 mac 15
forwarding：收发 BPDU，学习 mac，转发数据

disable：说明端口未启用 STP 协议；
blocking：阻塞状态，属于 AP 端口正常状态，进行根桥的选举；
listening:侦听状态，进行端口角色的确定；
learning:学习状态，学习 mac 地址表；
forwarding：转发状态，转发数据，学习 mac 地址；

华为交换机无论哪种生成树模式下，只有 3 种状态 discarding , learning ,forwarding

生成树计时器
生成树协议中用到 hello ，Forward delay 和 max age 这 3

个计时器，它们会影响端口状态迁移和收敛时间，可在全局使用命令修改。
调整计时器一定要在 根交换机上配置，其他交换机使用根桥交换机的计时器工作，根桥交换机的 BPDU 中的计时器优于交换机本地计时器的配置。

设备有自己的超时时间，在超时时间内没有收到 BPDU,就会重新进行生成树的计算，Max Age 在华为交换机中的意义不大，超时时间为 hello time 时间 x 3 x 时间因子（默认为 3，取值 1-10），默认为 18s，最小为 6s。

```
[Huawei]stp timer ?
  forward-delay  Specify forward delay
  hello          Specify hello time interval
  max-age        Specify max age
```

stp timer hello 300 改为 3 S
stp timer forward-delay 2000 改为 20 S
stp timer max-age 3000 改为 30 S
stp max-hops 30 改为 30 跳

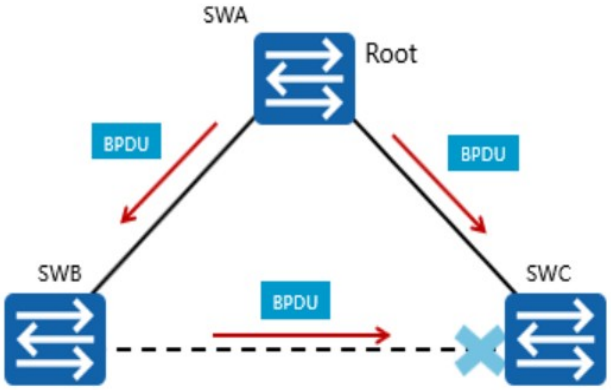
hello timer ： 2s root 每 2s 产生 BPDU。根交换机产生的 BPDU 的通告时间间隔
forward delay： 15s 设备状态迁移的延迟时间。
链路故障会引发网络重新进行生成树的计算，生成树的结构将发生相应的变化。不过重新计算得到的新拓扑信息无法立即传到整个网络。此时若立即将新选出的根端口和指定端口置于数据转发的状态，则可能会出现临时环路。这时要求新新选出的根端口和指定端口要经过 2 倍的 forward delay 后才能进入转发状态，这个延时足够保证新的配置消息能传遍整个网络。

max age：　　20s 最大老化时间
储存 BPDU 的时间，spanning-tree 发生故障，20s 后原 blocking 状态->learning 状态
以非根桥的根接口为例，该设备将在这个接口上保存来自上游的最优 BPDU ，这个 BPDU 关联着一个最大生存时间，如果在该 BPDU 到达最大生存时间之前，接口再一次收到了 BPDU ，那么其最大生存时间将会被重置，而如果接口一直没有再收到 BPDU 从而导致该接口上保存的 BPDU 到达最大生存时间，那么该 BPDU 将被老化，此时设备将重新在接口上选择最优 BPDU ，也就是重新进行根接口的选举。

max hop :　　20
当一个 BPDU 报文到达一个桥，又被该桥转发称为一跳，当 BPDU 报文的跳数超过 max hop 时，该报文会被丢弃，该参数与网络规模相关。

BPDU 有两种类型：配置 BPDU 和 TCN BPDU
配置 BPDU 包含了桥 ID、路径开销和端口 ID 等参数。STP 协议通过在交换机之间传递配置 BPDU 来选举根交换机，以及确定每个交换机端口的角色和状态。在初始化过程中，每个桥都主动发送配置 BPDU。在网络拓扑稳定以后，只有根桥主动发送配置 BPDU，其他交换机在收到上游传来的配置 BPDU 后，才会发送自己的配置 BPDU。

TCN BPDU 是指下游交换机感知到拓扑发生变化时向上游发送的拓扑变化通知。



PID	PVI	BPDU Type	Flags	Root ID	RPC	Bridge ID	Port ID	Message Age	Max Age	Hello Time	Fwd Delay

BPDU包含桥ID、路径开销、端口ID、计时器等参数。

协议字段	作用
Flags	8bit，STP只用一个最高位和一个最低位，最高位置位为TCA BPDU,最低位置位为 TC BPDU
Root-id	根桥ID，根桥设备的BID,由优先级+mac地址组成
Cost of path	根路径开销，描述到达根桥的开销值
Bridge-id	转发桥id，描述发送该BPDU的设备BID.
Port-id	端口id，发送给BPDU的端口id
Message age	配置BPDU能够在网络中存活的最大生存时间
Max age	接口缓存BPDU的老化时间，默认为20s
hello time	根桥发送配置BPDU的发送周期，默认为2s
Forward delay	Listening，Learning状态的持续时间，默认为15s

前言

- 以太网是当今现有局域网LAN（Local Area Network）采用的最通用的通信协议标准。该标准定义了在局域网中采用的电缆类型和信号处理方法。以太网作为一种原理简单，便于实现同时又价格低廉的局域网技术已经成为业界的主流。而更高性能的千兆以太网和万兆以太网的出现更使其成为最有前途的网络技术。
- 本文主要介绍LAN网络中的MAC地址表、ARP原理、 MSTP技术、堆叠与集群、链路捆绑。

MAC地址表的组成 (1)

- 动态表项
  - 由接口通过报文中的源MAC地址学习获得，表项可老化，默认老化时间300秒。
  - 在系统复位、接口板热插拔或接口板复位后，动态表项会丢失。
- 静态表项
  - 由用户手工配置，并下发到各接口板，表项不可老化。
  - 在系统复位、接口板热插拔或接口板复位后，保存的表项不会丢失。
- 黑洞表项
  - 由用户手工配置，并下发到各接口板，表项不可老化。
  - 配置黑洞MAC地址后，源MAC地址或目的MAC地址是该MAC的报文将会被丢弃。

- MAC 地址表的定义

- MAC 地址表记录了交换机学习到的其他设备的 MAC 地址与接口的对应关系，以及接口所属 VLAN 等信息。设备在转发报文时，根据报文的目的 MAC 地址查询 MAC 地址表，如果 MAC 地址表中包含与报文目的 MAC 地址对应的表项，则直接通过该表项中的出接口转发该报文；如果 MAC 地址表中没有包含报文目的 MAC 地址对应的表项时，设备将采取广播方式在所属 VLAN 内除接收接口外的所有接口转发该报文。
- MAC 地址表中的表项分为：动态表项、静态表项和黑洞表项。
- 动态表项：由接口通过报文中的源 MAC 地址学习获得，表项可老化。在系统复位、接口板热插拔或接口板复位后，动态表项会丢失。可以通过查看动态 MAC 地址表项，可以判断两台相连设备之间是否有数据转发；也可以通过查看指定动态 MAC 地址表项的个数，可以获取接口下通信的用户数。
- 静态表项：由用户手工配置，并下发到各接口板，表项不可老化。在系统复位、接口板热插拔或接口板复位后，保存的表项不会丢失。一条静态 MAC 地址表项，只能绑定一个出接口。一个接口和 MAC 地址静态绑定后，不会影响该接口动态 MAC 地址表项的学习。通过绑定静态 MAC 地址表项，可以保证合法用户的使用，防止其他用户使用该 MAC 进行攻击。
- 黑洞表项：由用户手工配置，并下发到各接口板，表项不可老化。在系统复位、接口板热插拔或接口板复位后，保存的表项不会丢失。通过配置黑洞 MAC 地址表项，可以过滤掉非法用户。

MAC地址表的组成 (2)

```
[Huawei]dis mac-address
MAC address table of slot 0:
-------------------------------------------------------------------------------
MAC Address    VLAN/       PEVLAN CEVLAN Port           Type       LSP/LSR-ID
               VSI/SI                                              MAC-Tunnel
-------------------------------------------------------------------------------
00aa-bbcc-ddee -           -      -      -              blackhole  -
0011-2233-4455 1           -      -      GE0/0/2        static     -
-------------------------------------------------------------------------------
Total matching items on slot 0 displayed = 2

MAC address table of slot 0:
-------------------------------------------------------------------------------
MAC Address    VLAN/       PEVLAN CEVLAN Port           Type       LSP/LSR-ID
               VSI/SI                                              MAC-Tunnel
-------------------------------------------------------------------------------
5489-986d-53e0 1           -      -      GE0/0/16       dynamic    0/-
5489-98fc-608b 1           -      -      GE0/0/1        dynamic    0/-
-------------------------------------------------------------------------------
Total matching items on slot 0 displayed = 2
```

- 缺省情况下，MAC地址表项的老化时间为300秒。

- 通过“display mac-address”命令，可以查看设备的 mac 表项，如图所示，mac 表的组成可以分为动态、静态和黑洞。从表项中也可以看出，mac 地址所对应的 VLAN 以及 VSI。

MAC地址表配置

- 配置静态MAC表项

```
[Huawei]mac-address static 0011-2233-4455 GigabitEthernet 0/0/2 vlan  1
```

- 配置黑洞MAC表项

```
[Huawei]mac-address blackhole 00aa-bbcc-ddee
```

- 配置动态MAC表项的老化时间

```
[Huawei]mac-address aging-time 400
```

端口安全

- 端口安全（Port Security）通过将接口学习到的动态MAC地址转换为安全MAC地址（包括安全动态MAC、安全静态MAC和Sticky MAC），阻止非法用户通过本接口和交换机通信，从而增强设备的安全性。
- 安全MAC地址分类
  - 安全动态MAC地址
    - 使能端口安全而未使能Sticky MAC功能时转换的MAC地址。
  - 安全静态MAC地址
    - 使能端口安全时手工配置的静态MAC地址。
  - Sticky MAC地址
    - 使能端口安全后又同时使能Sticky MAC功能后转换到的MAC地址。
- 接口使能端口安全功能时，接口上之前学习到的动态 MA

C 地址表项将被删除，之后学习到的 MAC 地址将变为安全动态 MAC 地址。

- 接口使能 Sticky MAC 功能时，接口上的安全动态 MAC 地址表项将转化为 Sticky MAC 地址，之后学习到的 MAC 地址也变为 Sticky MAC 地址。
- 接口去使能端口安全功能时，接口上的安全动态 MAC 地址将被删除，重新学习动态 MAC 地址。
- 接口去使能 Sticky MAC 功能时，接口上的 Sticky MAC 地址，会转换为安全动态 MAC 地址。

配置端口安全

- 配置安全MAC功能

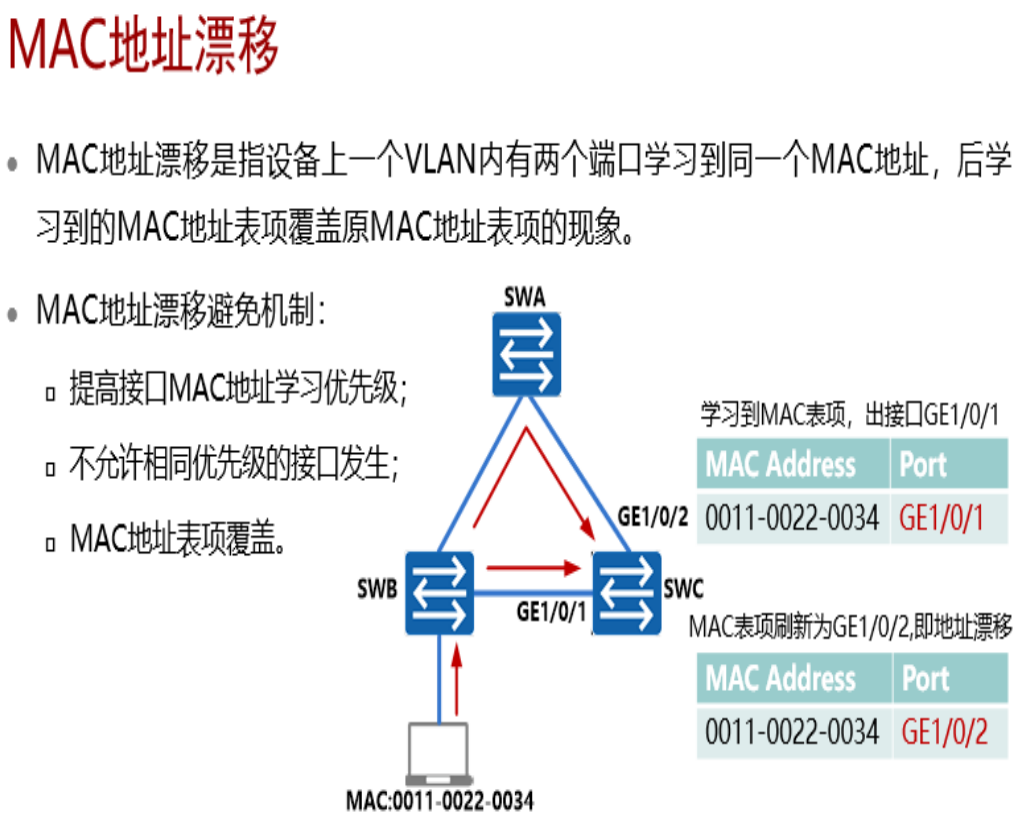
```
[Huawei-GigabitEthernet0/0/2]port-security enable
\\使能端口安全功能
[Huawei-GigabitEthernet0/0/2]port-security protect-action shutdown
\\配置端口安全保护动作
[Huawei-GigabitEthernet0/0/2]port-security max-mac-num 5
\\配置端口安全动态MAC学习限制数量
[Huawei-GigabitEthernet0/0/2]port-security aging-time 1000
\\配置接口学习到的安全动态MAC地址的老化时间
```

- 配置Sticky MAC功能

```
[Huawei-GigabitEthernet0/0/2]port-security enable
[Huawei-GigabitEthernet0/0/2]port-security mac-address sticky
\\使能接口Sticky MAC功能
```

- 说明：
- 接口使能 Sticky MAC 功能，安全动态 MAC 地址表项将转化为 Sticky MAC 地址，之后学习到的 MAC 地址也变为 Sticky MAC 地址。
- 接口使能 Sticky MAC 功能，即使配置了 port-security aging-time，Sticky MAC 也不会被老化。
- Sticky MAC 地址表项，保存后重启设备不丢弃。

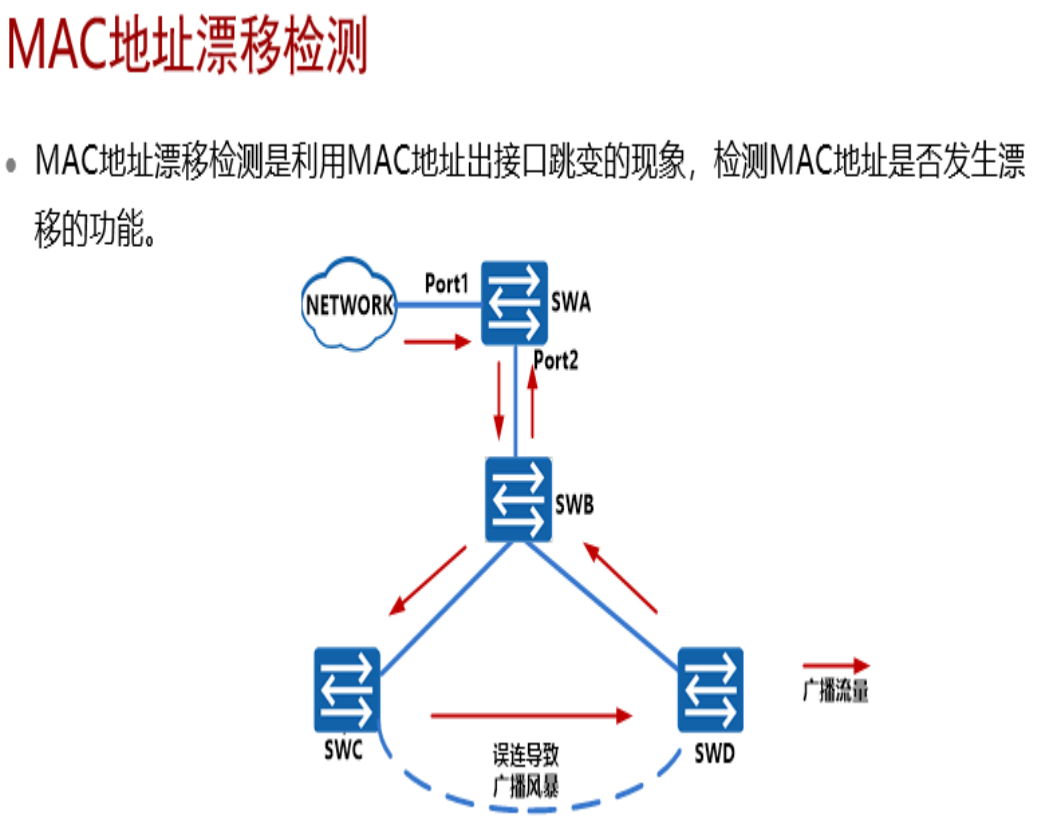
- 端口安全的保护动作：
- Restrict：丢弃源 MAC 地址不存在的报文并上报告警。推荐使用 restrict 动作。
- Protect：只丢弃源 MAC 地址不存在的报文，不上报告警。
- Shutdown：接口状态被置为 error-down，并上报告警。



- 如图所示，MAC 地址为 0011-0022-0034 的表项，出接口由 GE1/0/1 刷新为 GE1/0/2，这就是 MAC 地址漂移。设备出现 MAC 地址漂移时，设备 CPU 占用率会有不同程度的升高。正常情况下，网络中不会在短时间内出现大量 MAC 地址漂移的情况。出现这种现象一般都意味着网络中存在环路，可以通过查看告警信息和漂移记录，快速定位和排除环路。
- 网络中产生环路或非法用户进行网络攻击都会造成 MAC 地址发生漂移，导致 MAC 地址不稳定。在规划网络时，可以

通过下面两种方式来避免这种情况：

- 提高接口 MAC 地址学习优先级。当不同接口学到相同的 MAC 地址表项时，高优先级接口学到的 MAC 地址表项可以覆盖低优先级接口学到的 MAC 地址表项，防止 MAC 地址在接口间发生漂移。
- 不允许相同优先级的接口发生 MAC 地址表项覆盖。当伪造网络设备所连接口的优先级与安全的网络设备相同时，后学习到的伪造网络设备的 MAC 地址表项不会覆盖之前正确的表项。但如果网络设备下电，仍会学习到伪造网络设备的 MAC 地址，当网络设备再次上电时将无法学习到正确的 MAC 地址。



- 配置 MAC 地址漂移检测功能后，在发生 MAC 地址漂移时，可以上报包括 MAC 地址、VLAN，以及跳变的接口等信

息的告警。其中跳变的接口即为可能出现环路的接口。网络管理员可以根据告警信息，手工排查网络中环路的源头，也可以使用 MAC 漂移检测提供的后续动作，使跳变的端口 down 或者 VLAN 从端口中退出，实现自动破环。

- 如图所示网络中，若 SwitchC 和 SwitchD 之间误接网线，则 SwitchB、SwitchC、SwitchD 之间形成环路。当 SwitchA 上 Port1 接口从网络中收到一个广播报文后转发给 SwitchB，该报文经过环路，会被 SwitchA 上 Port2 接口收到。配置 MAC 地址漂移检测功能，SwitchA 就会感知到 MAC 地址出接口跳变的现象。若连续出现此现象，SwitchA 就会上报 MAC 漂移告警，提醒管理员进行维护。

MAC地址防漂移配置

- 配置接口MAC地址学习优先级

```
[Huawei-GigabitEthernet0/0/2]mac-learning priority 3
\\配置接口学习MAC地址的优先级，缺省情况下，接口学习MAC地址的优先级为0，数值越大优先级越高
```

- 配置不允许相同优先级接口MAC地址漂移

```
[Huawei]undo mac-learning priority 3 allow-flapping
\\配置不允许相同优先级的接口发生MAC地址漂移
```

- 配置全局MAC地址漂移检测

```
[Huawei]mac-address flapping detection
\\配置全局MAC地址漂移检测功能
```

- 配置基于VLAN的MAC地址漂移检测

```
[Huawei]vlan 2
[Huawei-vlan2]loop-detect eth-loop block-time 100 retry-times 3
\\配置MAC地址漂移检测功能
```

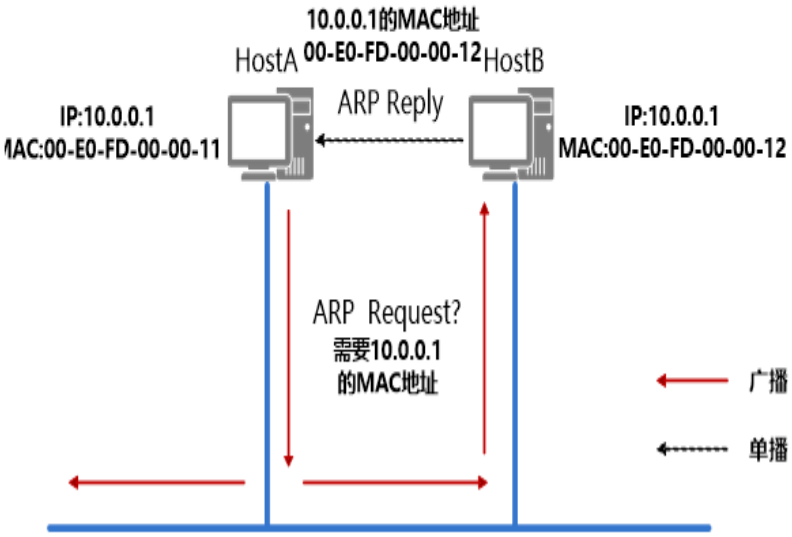
- 接口配置不同的 MAC 地址学习优先级后，如果不同接口学到相同的 MAC 地址表项，那么高优先级接口学到的 MAC 地址表项可以覆盖低优先级接口学到的 MAC 地址表项，防止

MAC 地址发生漂移。

- 配置不允许相同优先级的接口发生 MAC 地址表项覆盖，也可以防止 MAC 地址漂移，提高网络的安全性。

免费ARP

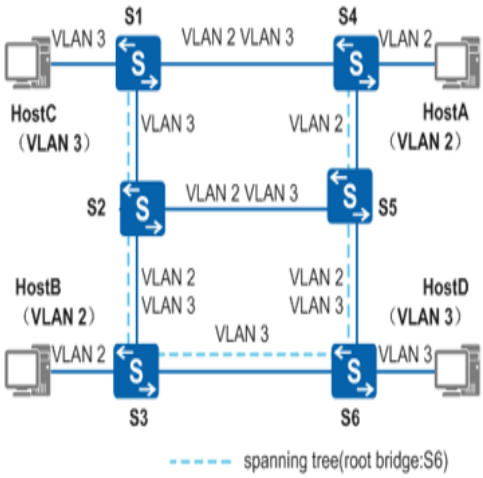
- 设备主动使用自己的IP地址作为目的IP地址发送ARP请求，此种方式称免费ARP。



- 免费 ARP 有如下作用：
- IP 地址冲突检测：当设备接口的协议状态变为 Up 时，设备主动对外发送免费 ARP 报文。正常情况下不会收到 ARP 应答，如果收到，则表明本网络中存在与自身 IP 地址重复的地址。如果检测到 IP 地址冲突，设备会周期性的广播发送免费 ARP 应答报文，直到冲突解除。
- 用于通告一个新的 MAC 地址：发送方更换了网卡，MAC 地址变化了，为了能够在动态 ARP 表项老化前通告网络中其他设备，发送方可以发送一个免费 ARP。
- 在 VRRP 备份组中用来通告主备发生变换：发生主备变换后，MASTER 设备会广播发送一个免费 ARP 报文来通告发生了主备变换。

STP和RSTP的缺陷

- RSTP和STP还存在同一个缺陷：由于局域网内所有的VLAN共享一棵生成树，因此无法在VLAN间实现数据流量的负载均衡，链路被阻塞后将不承载任何流量，还有可能造成部分VLAN的报文无法转发。

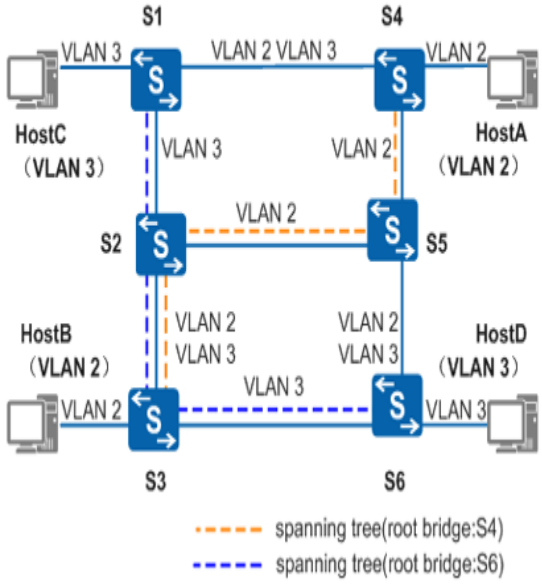


- 如图所示网络中，生成树结构在图中用虚线表示，S6为根交换设备。S2和S5之间、S1和S4之间的链路被阻塞。HostA和B同属于VLAN2，由于S2和S5之间的链路被阻塞，S3和S6之间的链路又不允许VLAN2的报文通过，因此HostA和HostB之间无法互相通讯。

- 在 STP 和 RSTP 的算法中，所有 VLAN 共享一课生成树，会造成部分 VLAN 无法通信、次优路径、流量无法负载分担等问题。
- 为了弥补 STP 和 RSTP 的缺陷，IEEE 于 2002 年发布的 802.1S 标准定义了 MSTP。MSTP 兼容 STP 和 RSTP，既可以快速收敛，又提供了数据转发的多个冗余路径，在数据转发过程中实现 VLAN 数据的负载均衡。

MSTP对STP和RSTP的改进

- MSTP把一个交换网络划分成多个域，每个域内形成多棵生成树，生成树之间彼此独立。
- 每棵生成树叫做一个多生成树实例MSTI (Multiple Spanning Tree Instance)，每个域叫做一个MST域 (MST Region: Multiple Spanning Tree Region)。



- 所谓生成树实例就是多个 VLAN 的一个集合。通过将多个 VLAN 捆绑到一个实例，可以节省通信开销和资源占用率。MSTP 各个实例拓扑的计算相互独立，在这些实例上可以实现负载均衡。可以把多个相同拓扑结构的 VLAN 映射到一个实例里，这些 VLAN 在端口上的转发状态取决于端口在对应 MSTP 实例的状态。
- 如图所示，MSTP 通过设置 VLAN 映射表（即 VLAN 和 MSTI 的对应关系表），把 VLAN 和 MSTI 联系起来。每个 VLAN 只能对应一个 MSTI，即同一 VLAN 的数据只能在一个 MSTI 中传输，而一个 MSTI 可能对应多个 VLAN。
- 经计算，最终生成两棵生成树：
- MSTI1 以 S4 为根交换设备，转发 VLAN2 的报文。
- MSTI2 以 S6 为根交换设备，转发 VLAN3 的报文。
- 这样所有 VLAN 内部可以互通，同时不同 VLAN 的报文沿不同的路径转发，实现了负载分担。

MSTP基本概念

- MST域 (MST Region)
- MSTI
- VLAN映射表
- CST、IST、SST、CIST
- 域根
- 总根
- 主桥
- 端口角色
- 端口状态

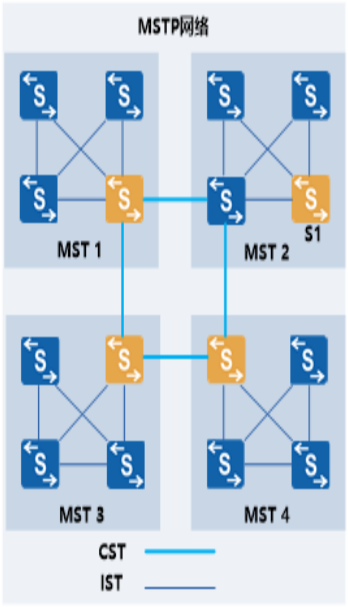


图1 MSTP网络层次示意图

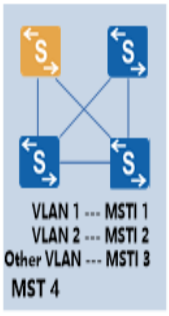


图2 VLAN与MSTI的映射

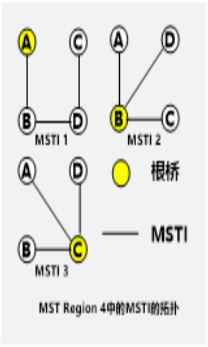


图3 MSTI独立计算生成树

- MST 由交换网络中的多台交换设备以及它们间的网段所构成。MSTI 是 MST 域下实例，一个 MST 域下可以有多个 MSTI。
- VLAN 映射表描述了 VLAN 和 MSTI 之间的映射关系。如图 2 所示，MST Region4 中，VLAN1 映射到 MSTI1，VLAN2 映射到 MSTI2，其余 VLAN 映射到 MSTI3。
- 公共生成树 CST 是连接交换网络内所有 MST 域的一棵生成树。如果把每个 MST 域看作是一个节点，CST 就是这些节点通过 STP 或 RSTP 协议计算生成的一棵生成树。
- 内部生成树 IST（Internal Spanning Tree）是各 MST 域内的一棵生成树。IST 是一个特殊的 MSTI，MSTI 的 ID 为 0，通常称为 MSTI0。
- SST（Single Spanning Tree）：运行 STP 或 RSTP 的

交换设备只能属于一个生成树；MST 域中只有一个交换设备，这个交换设备构成单生成树。

- 所有 MST 域的 IST 加上 CST 就构成一棵完整的生成树，即 CIST。
- 域根（Regional Root）分为 IST 域根和 MSTI 域根。
- IST 域根如图 1 所示，在 MST 域中 IST 生成树中距离总根最近的交换设备是 IST 域根。
- 一个 MST 域内可以生成多棵生成树，每棵生成树都称为一个 MSTI。MSTI 域根是每个多生成树实例的树根。如图 3 所示，域中不同的 MSTI 有各自的域根。
- 总根是 CIST（Common and Internal Spanning Tree）的根桥。如图 1 中的 S1。
- 主桥（Master Bridge）也就是 IST Master，它是域内距离总根最近的交换设备。如图 1 中的黄色交换机。如果总根在 MST 域中，则总根为该域的主桥。
- 端口角色：同 RSTP，MSTP 中定义了根端口、指定端口、Alternate 端口、Backup 端口和边缘端口。
- 端口状态：同 RSTP，MSTP 定义的端口状态有 Forwarding, Learning, Discarding。

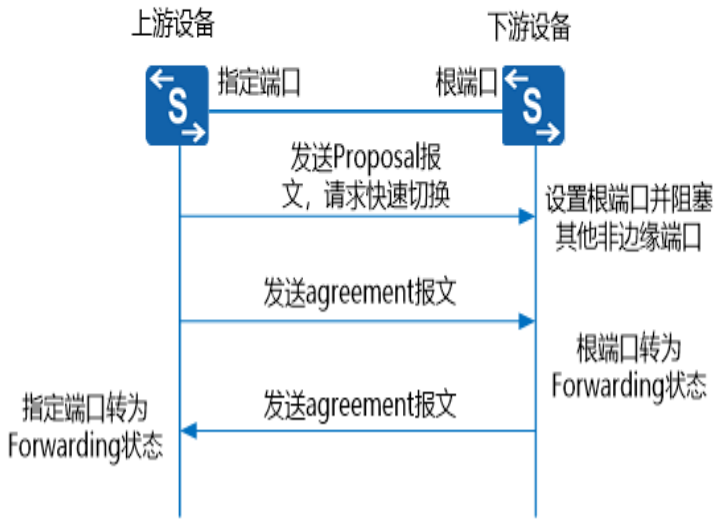
MSTP拓扑计算

- CIST的计算
  - 经过比较配置消息后，在整个网络中选择一个优先级最高的交换设备作为CIST的树根。在每个MST域内MSTP通过计算生成IST；同时MSTP将每个MST域作为单台交换设备对待，通过计算在MST域间生成CST。CST和IST构成了整个交换设备网络的CIST。
- MSTI的计算
  - 在MST域内，MSTP根据VLAN和生成树实例的映射关系，针对不同的VLAN生成不同的生成树实例。每棵生成树独立进行计算，计算过程与STP计算生成树的过程类似。
- MSTP对拓扑变化的处理
  - MSTP拓扑变化处理与RSTP拓扑变化处理过程类似。

- MSTI 的特点：
- 每个 MSTI 独立计算自己的生成树，互不干扰。
- 每个 MSTI 的生成树计算方法与 STP 基本相同。
- 每个 MSTI 的生成树可以有不同的根，不同的拓扑。
- 每个 MSTI 在自己的生成树内发送 BPDU。
- 每个 MSTI 的拓扑通过命令配置决定。
- 每个端口在不同 MSTI 上的生成树参数可以不同。
- 每个端口在不同 MSTI 上的角色、状态可以不同。
- 在运行 MSTP 协议的网络中，一个 VLAN 报文将沿着如下路径进行转发：
- 在 MST 域内，沿着其对应的 MSTI 转发。
- 在 MST 域间，沿着 CST 转发。

MSTP快速收敛机制

- 除支持RSTP所支持的普通P/A (Proposal/Agreement)机制外，MSTP还支持增强方式的P/A机制。



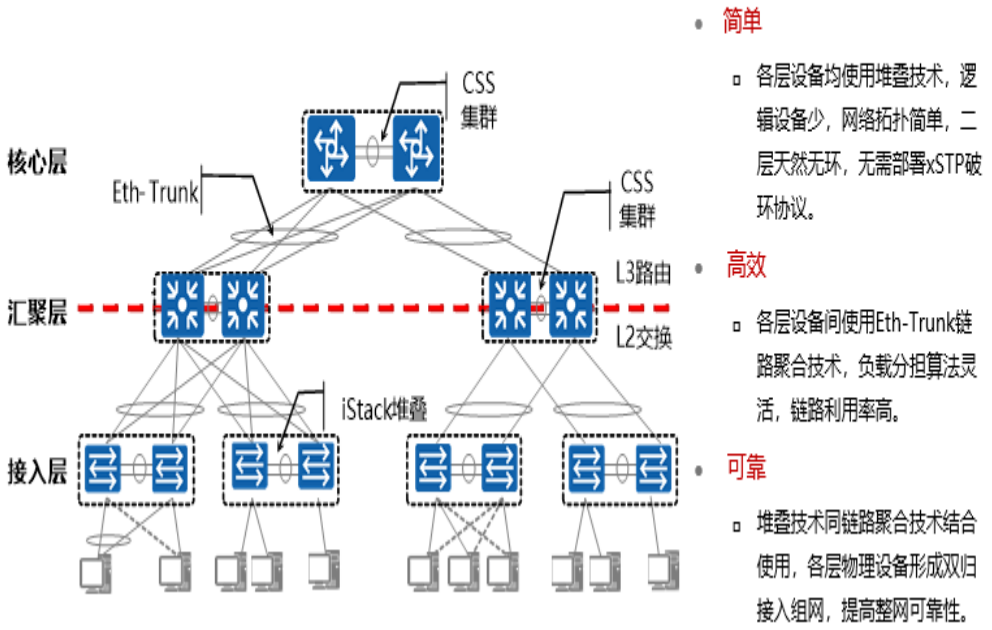
- 如图所示，在 MSTP 中，P/A 机制工作过程如下：
- 上游设备发送 Proposal 报文，请求进行快速迁移。下游设备接收到后，把与上游设备相连的端口设置为根端口，并阻塞所有非边缘端口。
- 上游设备继续发送 Agreement 报文。下游设备接收到后，根端口转为 Forwarding 状态。
- 下游设备回应 Agreement 报文。上游设备接收到后，把与下游设备相连的端口设置为指定端口，指定端口进入 Forwarding 状态。
- 缺省情况下，华为数据通信设备使用增强的快速迁移机制。如果华为数据通信设备和其他制造商的设备进行互通，而其他制造商的设备 P/A 机制使用普通的快速迁移机制，此时，可在华为数据通信设备上通过设置 P/A 机制为普通的快速迁移机制，从而实现华为数据通信设备和其他制造商的设备进行互通。

MSTP配置

- 配置MSTP基本功能

```
[SwitchA] stp region-configuration
\\进入MST域视图
[SwitchA-mst-region] region-name RG1
\\配置MST域的域名
[SwitchA-mst-region] instance 1 vlan 2 to 10
\\配置多生成树实例和VLAN的映射关系
[SwitchA-mst-region] instance 2 vlan 11 to 20
[SwitchA-mst-region] active region-configuration
\\激活MST域的配置，使域名、VLAN映射表和MSTP修订级别生效
[SwitchA] stp instance 1 root primary
\\配置当前设备为根桥设备
[SwitchA] stp instance 2 root secondary
\\配置当前交换设备为备份根桥设备
[SwitchA] stp pathcost-standard legacy
\\配置SwitchA的端口路径开销值的计算方法为华为计算方法
[SwitchA] stp enable
\\在SwitchA上启动MSTP
```

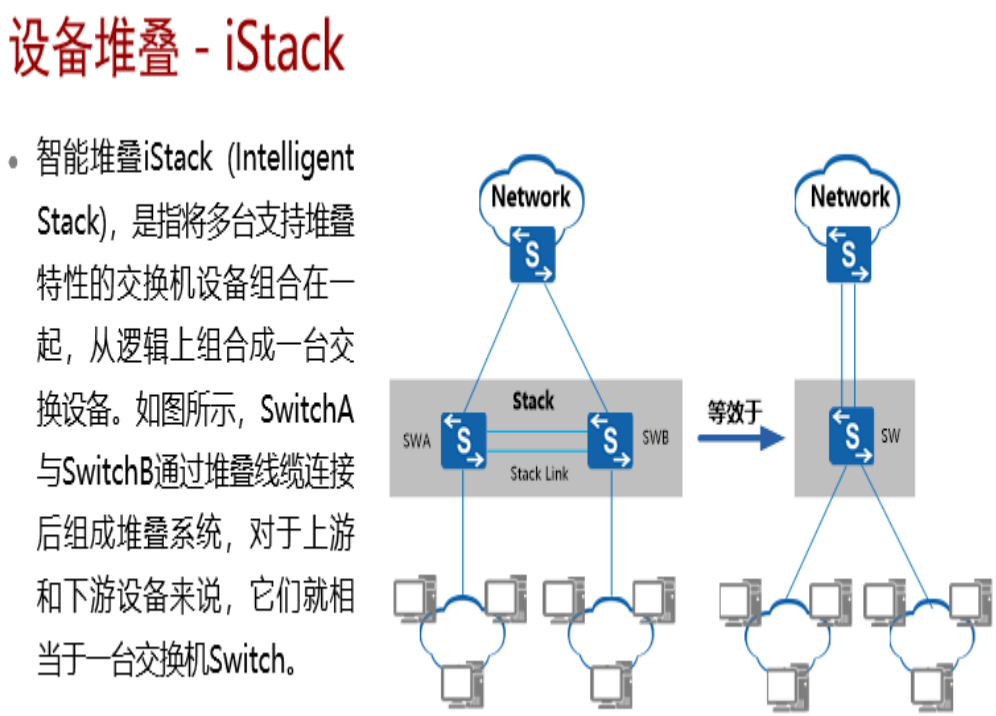
典型园区组网之一 - CSS + Eth-Trunk + iStack



- 如图所示，为园区网络 CSS+iStack 组网之一，其主要有简单、高效、可靠的特点。
- 简单
- 各层设备均使用堆叠技术，逻辑设备少，网络拓扑简单，二层天然无环，无需部署 xSTP 破环协议。
- 高效
- 各层设备间使用 Eth-Trunk 链路聚合技术，负载分担算法灵活，链路利用率高。
- 可靠
- 服务器和主机可以配置多 NIC 网卡 Teaming 负载均衡或主备冗余链路提高服务器接入可靠性。
- 堆叠技术同链路聚合技术结合使用，各层物理设备形成双归接入组网，提高整网可靠性。
- 缺点
- 对设备性能要求较高，盒式设备堆叠台数过多，可能导

致堆叠主的主控性能下降。

- 如果采用业务口堆叠或集群，会占用业务端口数。

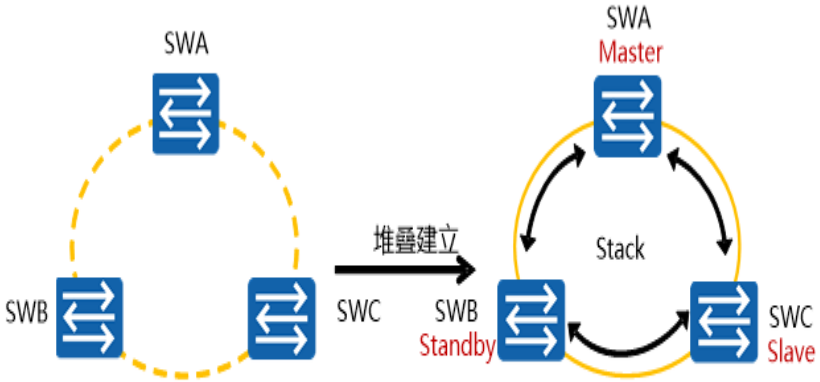


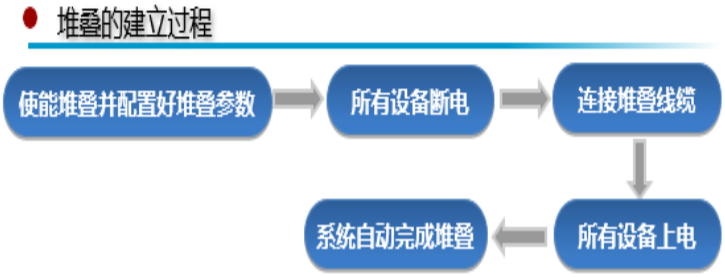
- 通过交换机堆叠，可以实现网络高可靠性和网络大数据量转发，同时简化网络管理。
- 高可靠性。堆叠系统多台成员交换机之间冗余备份；堆叠支持跨设备的链路聚合功能，实现跨设备的链路冗余备份。
- 强大的网络扩展能力。通过增加成员交换机，可以轻松的扩展堆叠系统的端口数、带宽和处理能力；同时支持成员交换机热插拔，新加入的成员交换机自动同步主交换机的配置文件和系统软件版本。
- 简化配置和管理。一方面，用户可以通过任何一台成员交换机登录堆叠系统，对堆叠系统所有成员交换机进行统一配置和管理；另一方面，堆叠形成后，不需要配置复杂的二层破环协议和三层保护倒换协议，简化了网络配置。

iStack基本概念

- 角色
  - 堆叠中所有的单台交换机都称为成员交换机，按照功能不同，可以分为三种角色：
    - 主交换机（Master）：负责管理整个堆叠。堆叠中只有一台主交换机。
    - 备交换机（Standby）：是主交换机的备份交换机。当主交换机故障时，备交换机会接替原主交换机的所有业务。堆叠中只有一台备交换机。
    - 从交换机（Slave）：主要用于业务转发，从交换机数量越多，堆叠系统的转发能力越强。除主交换机和备交换机外，堆叠中其他所有的成员交换机都是从交换机。
- 堆叠ID
  - 即成员交换机的槽位号（Slot ID），用来标识和管理成员交换机，堆叠中所有成员交换机的堆叠ID都是唯一的。
- 堆叠优先级
  - 堆叠优先级是成员交换机的一个属性，主要用于角色选举过程中确定成员交换机的角色，优先级值越大表示优先级越高，优先级越高当选为主交换机的可能性越大。

堆叠建立



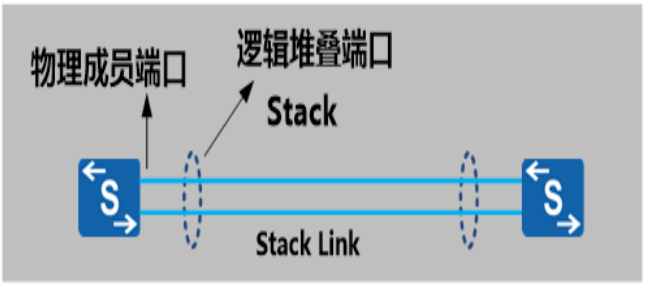


- “系统自动完成堆叠”实际上可以细分为三步：
- 主交换机选举
- 运行状态比较，已经运行的交换机比处于启动状态的交换机优先竞争为主交换机。
- 堆叠优先级高的交换机优先竞争为主交换机。
- 堆叠优先级相同时，MAC 地址小的交换机优先竞争为主交换机。
- 拓扑收集和备交换机选举
- 主交换机选举完成后，主交换机会收集所有成员交换机的拓扑信息，根据拓扑信息计算出堆叠转发表项和破环点信息下发给堆叠中的所有成员交换机，并向所有成员交换机分配堆叠 ID。之后进行备交换机的选举，作为主交换机的备份交换机。当除主交换机外其它交换机同时完成启动时：
- 堆叠优先级最高的设备成为备交换机。

- 堆叠优先级相同时，MAC 地址最小的成为备交换机。
- 稳定运行
- 角色选举、拓扑收集完成之后，剩下的其他成员交换机作为从交换机加入堆叠，所有成员交换机会自动同步主交换机的系统软件和配置文件：
- 堆叠具有自动加载系统软件的功能，待组成堆叠的成员交换机不需要具有相同软件版本，只需要版本间兼容即可。当备交换机或从交换机与主交换机的软件版本不一致时，备交换机或从交换机会自动从主交换机下载系统软件，然后使用新系统软件重启，并重新加入堆叠。
- 堆叠具有配置文件同步机制，备交换机或从交换机会将主交换机的配置文件同步到本设备并执行，以保证堆叠中的多台设备能够像一台设备一样在网络中工作，并且在主交换机出现故障之后，其余交换机仍能够正常执行各项功能。

堆叠连接方式

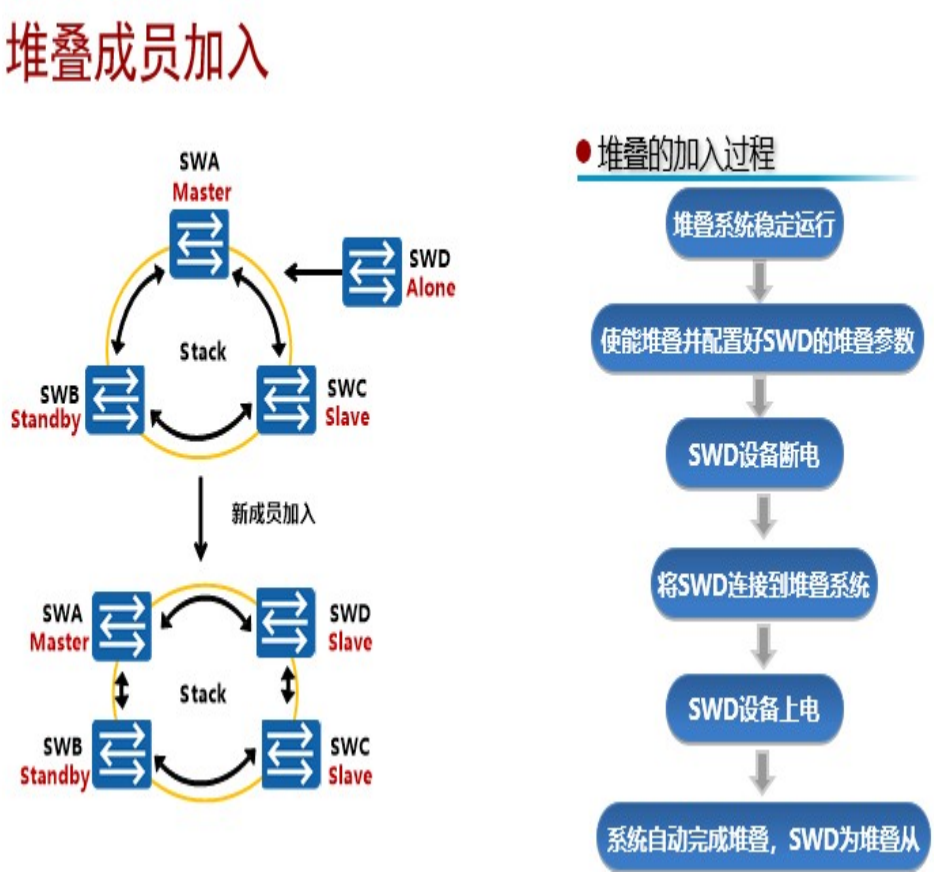
- 交换机组建堆叠根据堆叠口的不同，可以分为两种方式：堆叠卡堆叠和业务口堆叠。
  - 堆叠卡堆叠又分为以下两种情况：
    - 交换机之间通过专用的堆叠插卡及专用的堆叠线缆连接。
    - 堆叠卡集成到了交换机后面板上，交换机通过集成的堆叠端口及专用的堆叠线缆连接。
  - 业务口堆叠指的是交换机之间通过与逻辑堆叠端口绑定的物理成员端口相连，不需要专用的堆叠插卡。如图所示：



- 物理成员端口
- 成员交换机之间用于堆叠连接的物理端口。物理成员端口用于转发需要跨成员交换机的业务报文或成员交换机之间的堆叠协议报文。
- 逻辑堆叠端口
- 逻辑堆叠端口是专用于堆叠的逻辑端口，需要和物理成员端口绑定。堆叠的每台成员交换机上支持两个逻辑堆叠端口，分别为 stack-port n/1 和 stack-port n/2，其中 n 为成员交换机的堆叠 ID。
- 业务口堆叠根据连接线缆的不同又可以分为：普通线缆堆叠和专用线缆堆叠。
- 普通线缆堆叠
- 普通堆叠线缆包括：光线缆、网线和高速电缆。使用普通线缆堆叠时，逻辑堆叠端口需要手动进行配置，否则无法组

建堆叠。

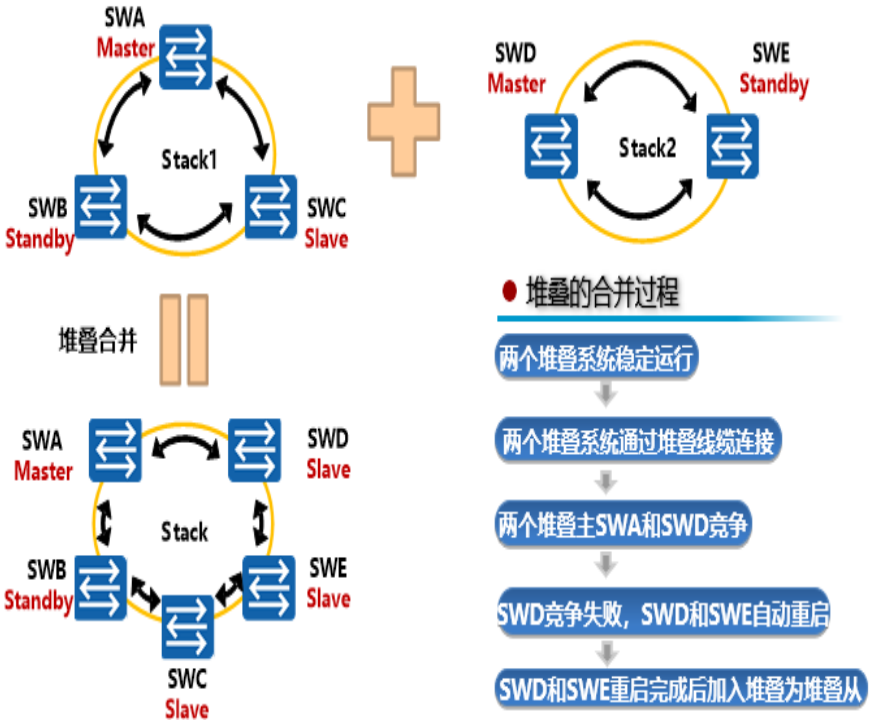
- 专用线缆堆叠
- 专用堆叠线缆的两端区分主和备，带有 Master 标签的一端为主端，不带有标签的一端为备端。使用专用线缆堆叠时，专用堆叠线缆按照规则插入端口后，交换机就可以自动组建堆叠。



- 堆叠成员加入是指向已经稳定运行的堆叠系统添加一台新的交换机。
- 使能堆叠并配置好 SWD 的堆叠参数
- 如果是业务口堆叠，新加入的交换机需要配置物理成员端口加入逻辑堆叠端口；并且链形连接时，当前堆叠系统链形两端（或一端）的成员交换机也需要配置物理成员端口加入逻辑堆叠口。

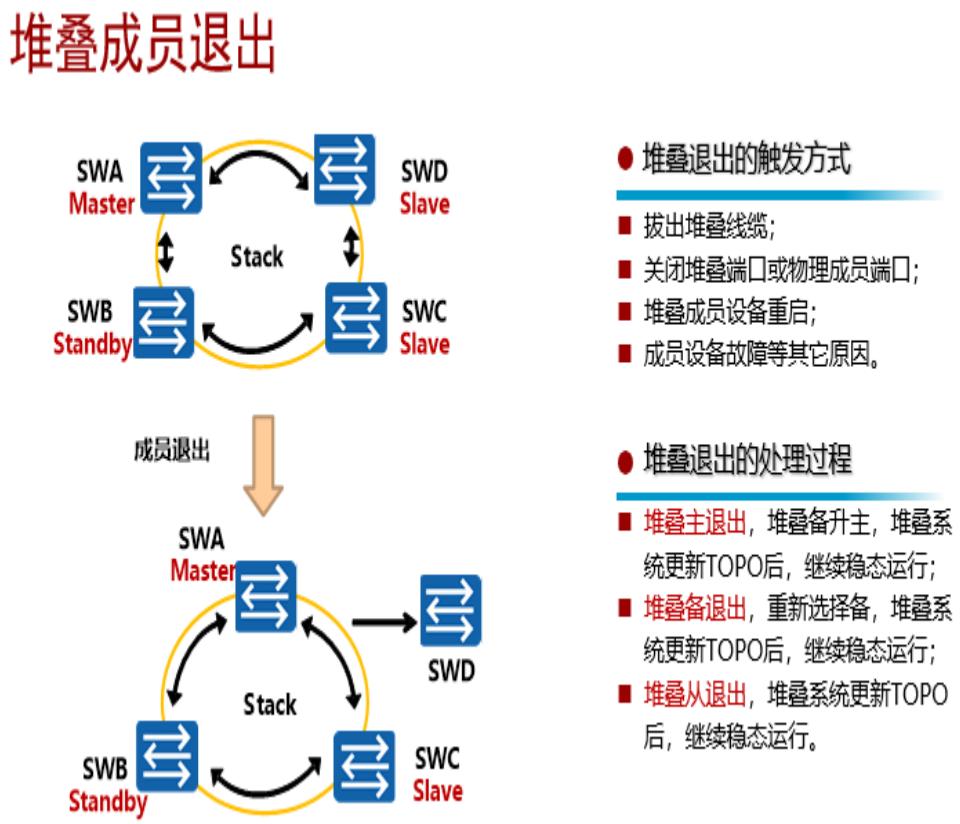
- 如果是堆叠卡堆叠，新加入的成员交换机需要使能堆叠功能。
- 为了便于管理，建议为新加入的交换机配置堆叠 ID。如果不配置，堆叠系统会为其分配一个堆叠 ID。
- 将 SWD 连接到堆叠系统
- 如果是链形连接，新加入的交换机建议添加到链形的两端，这样对现有的业务影响最小。
- 如果是环形连接，需要把当前环形拆成链形，然后在链形的两端添加设备。
- 系统完成堆叠
- 新加入的交换机连线上电启动后，进行角色选举，新加入的交换机会选举为从交换机，堆叠系统中原有主备从角色不变。
- 角色选举结束后，主交换机更新堆叠拓扑信息，同步到其他成员交换机上，并向新加入的交换机分配堆叠 ID（新加入的交换机没有配置堆叠 ID 或配置的堆叠 ID 与原堆叠系统的冲突时）。
- 新加入的交换机更新堆叠 ID，并同步主交换机的配置文件和系统软件，之后进入稳定运行状态。

堆叠合并



- 堆叠合并是指稳定运行的两个堆叠系统合并成一个新的堆叠系统。如图所示，两个堆叠系统的主交换机 SWA 和 SWD 通过竞争，选举出一个更优的作为新堆叠系统的主交换机。竞争成功的主交换机 SWA 所在的堆叠系统将保持原有主备从角色和配置不变，业务也不会受到影响；而另外一个堆叠系统的所有成员交换机 SWD 和 SWE 将重新启动，以从交换机的角色加入到新堆叠系统，其堆叠 ID 将由新主交换机重新分配，并将同步新主交换机的配置文件和系统软件，该堆叠系统的原有业务也将中断。
- 堆叠合并通常在以下两种情形下出现：
- 堆叠链路或设备故障导致堆叠分裂，链路或设备故障恢复后，分裂的堆叠系统重新合并。
- 待加入堆叠系统的交换机配置了堆叠功能，在不下电的情况下，使用堆叠线缆连接到正在运行的堆叠系统。通常情况

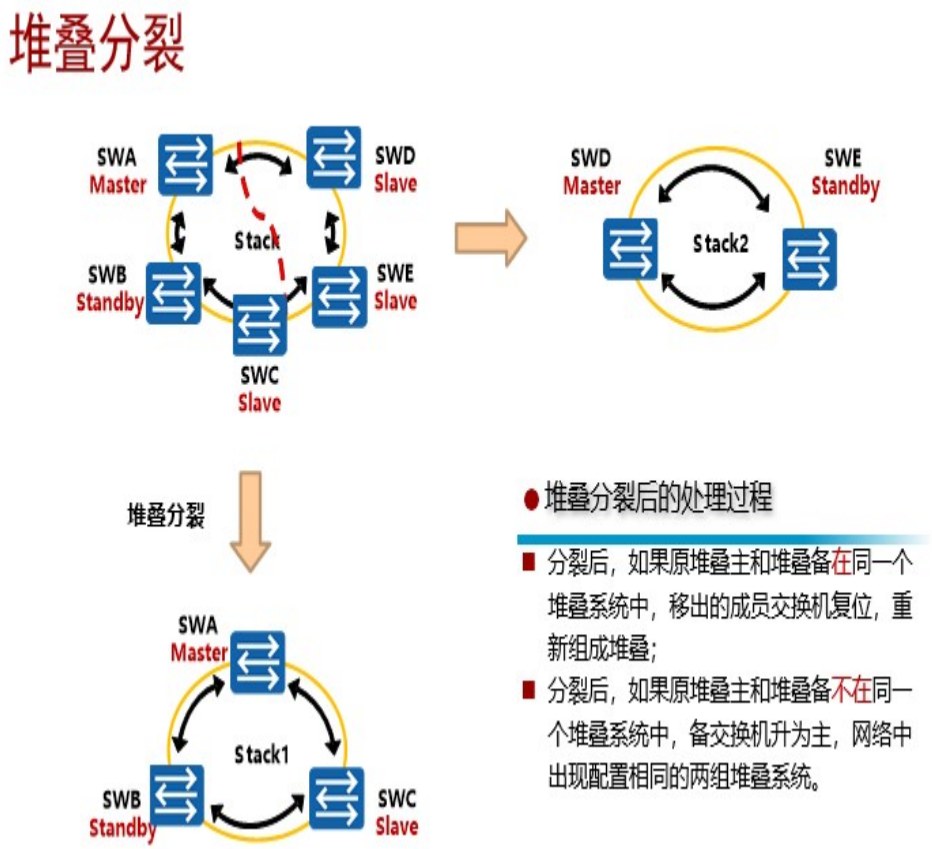
下，不建议使用该方式形成堆叠，因为在合并前过程中可能会导致正在运行的堆叠系统重启，影响业务运行。



- 堆叠成员退出是指成员交换机从堆叠系统中离开。根据退出成员交换机角色的不同，对堆叠系统的影响也有所不同：
- 当主交换机退出，备份交换机升级为主交换机，重新计算堆叠拓扑并同步到其他成员交换机，指定新的备交换机，之后进入稳定运行状态。
- 当备交换机退出，主交换机重新指定备交换机，重新计算堆叠拓扑并同步到其他成员交换机，之后进入稳定运行状态。
- 当从交换机退出，主交换机重新计算堆叠拓扑并同步到其他成员交换机，之后进入稳定运行状态。
- 堆叠成员交换机退出的过程，主要就是拆除堆叠线缆和移除交换机的过程：
- 对于环形堆叠：成员交换机退出后，为保证网络的可靠

性还需要把退出交换机连接的两个端口通过堆叠线缆进行连接。

- 对于链形堆叠：拆除中间交换机会造成堆叠分裂。这时需要在拆除前进行业务分析，尽量减少对业务的影响。

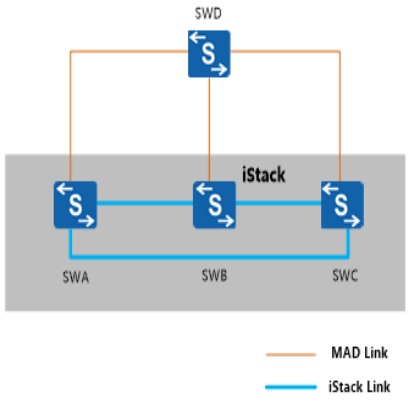


- 堆叠分裂是指稳定运行的堆叠系统中带电移出部分成员交换机，或者堆叠线缆多点故障导致一个堆叠系统变成多个堆叠系统。根据原堆叠系统主备交换机分裂后所处位置的不同，堆叠分裂可分为以下两类：
- 堆叠分裂后，原主备交换机被分裂到同一个堆叠系统中：原主交换机会重新计算堆叠拓扑，将移出的成员交换机的拓扑信息删除，并将新的拓扑信息同步给其他成员交换机；而移出的成员交换机检测到堆叠协议报文超时，将自行复位，重新进行选举。
- 堆叠分裂后，原主备交换机被分裂到不同的堆叠系统中：原主交换机所在堆叠系统重新指定备交换机，重新计算拓扑信

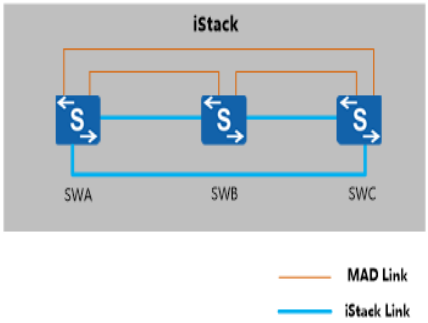
息并同步给其他成员交换机；原备交换机所在堆叠系统将发生备升主，原备交换机升级为主交换机，重新计算堆叠拓扑并同步到其他成员交换机，并指定新的备交换机。

多主检测 - 直连检测方式

- 直连检测的连接方式包括通过中间设备直连和堆叠成员交换机Full-mesh方式直连:



通过中间设备的直连检测方式

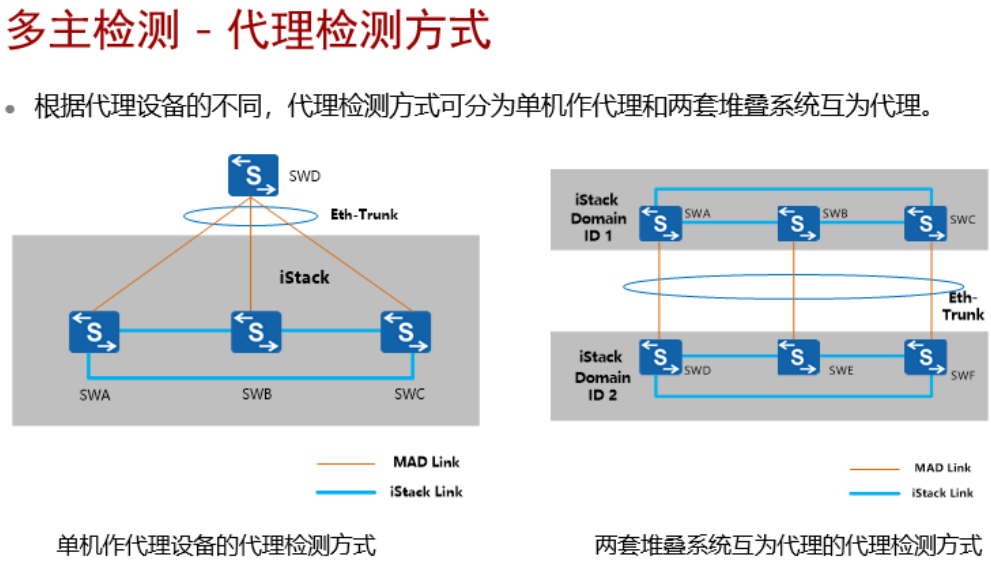


堆叠成员交换机Full-mesh方式直连

- 由于堆叠系统中所有成员交换机都使用同一个 IP 地址和 MAC 地址（堆叠系统 MAC），一个堆叠分裂后，可能产生多个具有相同 IP 地址和 MAC 地址的堆叠系统。为防止堆叠分裂后，产生多个具有相同 IP 地址和 MAC 地址的堆叠系统，引起网络故障，必须进行 IP 地址和 MAC 地址的冲突检查。多主检测 MAD（Multi-Active Detection），是一种检测和处理堆叠分裂的协议。链路故障导致堆叠系统分裂后，MAD 可以实现堆叠分裂的检测、冲突处理和故障恢复，降低堆叠分裂对业务的影响。
- MAD 检测方式有两种：直连检测方式和代理检测方式。在同一个堆叠系统中，两种检测方式互斥，不可以同时配置。
- 直连检测方式是指堆叠成员交换机间通过普通线缆直连

的专用链路进行多主检测。在直连检测方式中，堆叠系统正常运行时，不发送 MAD 报文；堆叠系统分裂后，分裂后的两台交换机以 1s 为周期通过检测链路发送 MAD 报文以进行多主冲突处理。

- 通过中间设备直连：堆叠系统的所有成员交换机之间至少有一条检测链路与中间设备相连。
- Full-mesh 方式直连：堆叠系统的各成员交换机之间通过检测链路建立 Full-mesh 全连接，即每两台成员交换机之间至少有一条检测链路。
- 通过中间设备直连可以实现通过中间设备缩短堆叠成员交换机之间的检测链路长度，适用于成员交换机相距较远的场景。与通过中间设备直连相比，Full-mesh 方式直连可以避免由中间设备故障导致的 MAD 检测失败，但是每两台成员交换机之间都建立全连接会占用较多的接口，所以该方式适用于成员交换机数目较少的场景。



- 代理检测方式是在堆叠系统 Eth-Trunk 上启用代理检测，在代理设备上启用 MAD 检测功能。此种检测方式要求堆叠系统中的所有成员交换机都与代理设备连接，并将这些链路加入

同一个 Eth-Trunk 内。与直连检测方式相比，代理检测方式无需占用额外的接口，Eth-Trunk 接口可同时运行 MAD 代理检测和其他业务。

- 在代理检测方式中，堆叠系统正常运行时，堆叠成员交换机以 30s 为周期通过检测链路发送 MAD 报文。堆叠成员交换机对在正常工作状态下收到的 MAD 报文不做任何处理；堆叠分裂后，分裂后的两台交换机以 1s 为周期通过检测链路发送 MAD 报文以进行多主冲突处理。
- MAD 冲突处理
- 堆叠分裂后，MAD 冲突处理机制会使分裂后的堆叠系统处于 Detect 状态或 Recovery 状态。Detect 状态表示堆叠正常工作状态，Recovery 状态表示堆叠禁用状态。
- MAD 冲突处理机制如下：MAD 分裂检测机制会检测到网络中存在多个处于 Detect 状态的堆叠系统，这些堆叠系统之间相互竞争，竞争成功的堆叠系统保持 Detect 状态，竞争失败的堆叠系统会转入 Recovery 状态；并且在 Recovery 状态堆叠系统的所有成员交换机上，关闭除保留端口以外的其它所有物理端口，以保证该堆叠系统不再转发业务报文。
- MAD 故障恢复
- 通过修复故障链路，分裂后的堆叠系统重新合并为一个堆叠系统。重新合并的方式有以下两种：
- 堆叠链路修复后，处于 Recovery 状态的堆叠系统重新启动，与 Detect 状态的堆叠系统合并，同时将被关闭的业务端口恢复 Up，整个堆叠系统恢复。
- 如果故障链路修复前，承载业务的 Detect 状态的堆叠系统也出现了故障。此时，可以先将 Detect 状态的堆叠系统从网络中移除，再通过命令行启用 Recovery 状态的堆叠系统，接替原来的业务，然后再修复原 Detect 状态堆叠系统的故障及链路故障。故障修复后，重新合并堆叠系统。

堆叠配置

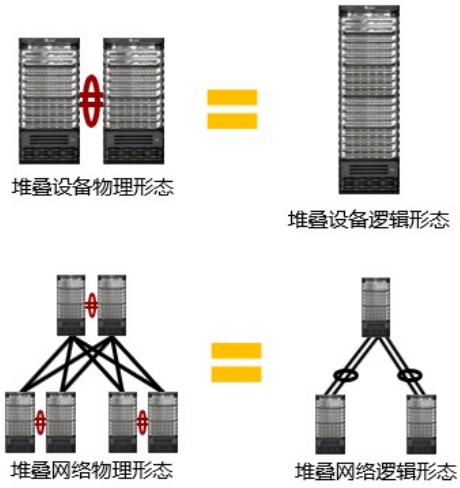
- 通过堆叠卡连接方式组建堆叠

```
[SwitchA] stack slot 0 priority 200
\\配置成员交换机的堆叠优先级。缺省情况下，成员交换机的堆叠优先级为100
[SwitchB] stack slot 0 renumber 1
\\配置设备的堆叠ID
[SwitchC] stack slot 0 renumber 2
```

- 通过业务口连接方式组建堆叠

```
[SwitchA] interface stack-port 0/1
[SwitchA-stack-port0/1] port interface gigabitethernet 0/0/27 enable
\\配置业务口为物理成员端口并将其加入到逻辑堆叠端口中。交换机B、C同理。
[SwitchA] interface stack-port 0/2
[SwitchA-stack-port0/2] port interface gigabitethernet 0/0/28 enable
[SwitchA] stack slot 0 priority 200
\\ 配置SwitchA的堆叠优先级为200
[SwitchB] stack slot 0 renumber 1
\\配置SwitchB的堆叠ID为1
[SwitchC] stack slot 0 renumber 2
```

CSS



- CSS的定义

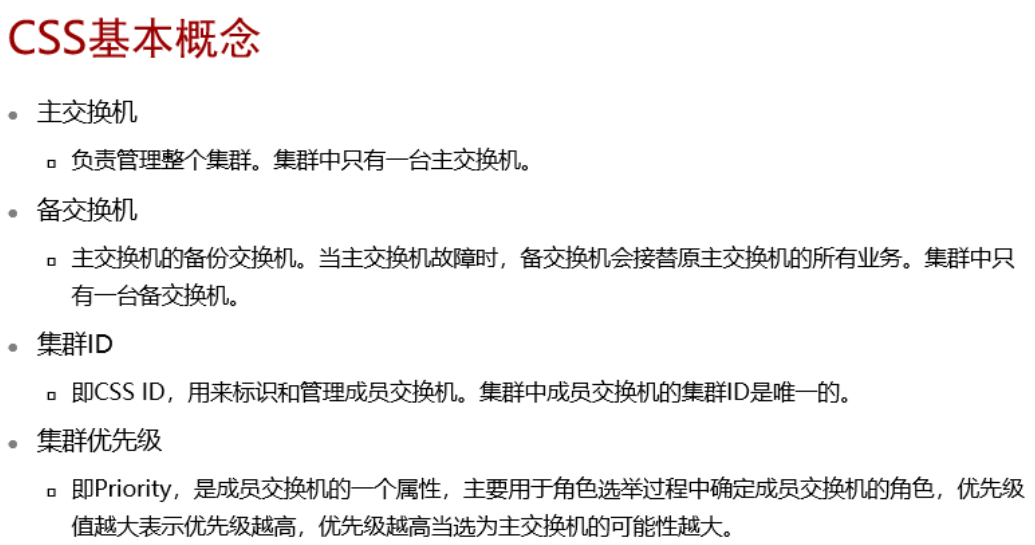
集群交换机系统CSS (Cluster Switch System)，又称为集群，是指将两台支持集群特性的交换机设备组合在一起，从逻辑上组合成一台交换设备。

- CSS的特征

- 交换机多虚一：堆叠交换机对外表现为一台逻辑交换机，控制平面合一，统一管理。
- 转发平面合一：堆叠内物理设备转发平面合一，转发信息共享并实时同步。
- 跨设备链路聚合：跨堆叠内物理设备的链路被聚合成一个Eth-Trunk端口，和下游设备实现互联。

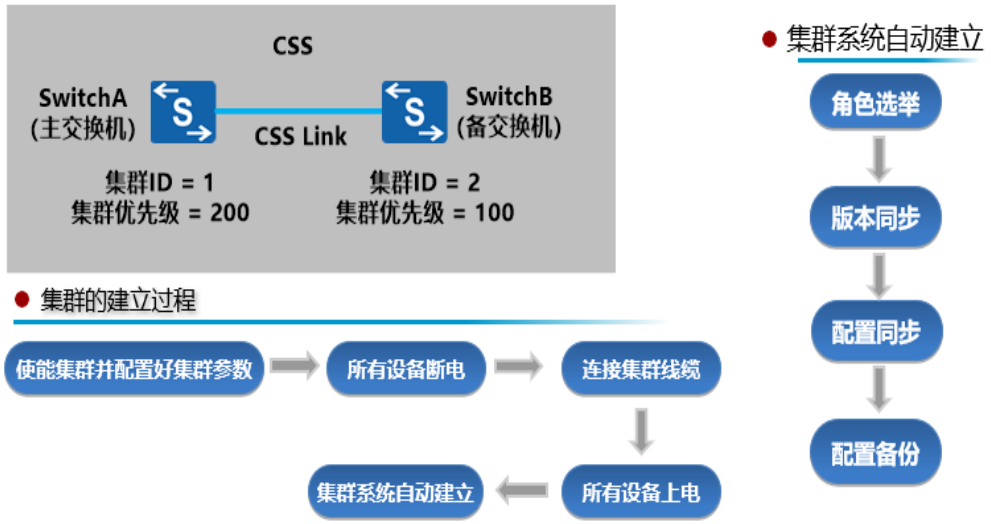
- CSS 与 iStack 的区别在于，一般框式交换机堆叠称为 CSS，盒式交换机堆叠称为 iStack，都可以称为堆叠。两者只是叫法和实现有些差异，但是功能是一样的。

- 通过交换机集群，可以实现网络高可靠性和网络大数据量转发，同时简化网络管理。
- 高可靠性：集群系统两台成员交换机之间冗余备份，同时利用链路聚合功能实现跨设备的链路冗余备份。
- 强大的网络扩展能力：通过组建集群增加交换机，从而轻松的扩展端口数、带宽和处理能力。
- 简化配置和管理：集群建立后，两台物理设备虚拟成为一台设备，用户只需登录一台成员交换机即可对集群系统所有成员交换机进行统一配置和管理。



- 不同于 iStack 可以多台设备堆叠，对于 CSS 集群，集群中只能有一主一备两台交换机。

CSS集群建立

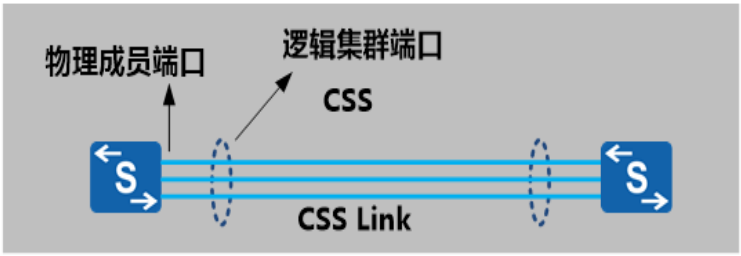


- 集群建立时，成员交换机间相互发送集群竞争报文，通过竞争，一台成为主交换机，负责管理整个集群系统，另一台则成为备交换机。
- 角色选举
- 最先完成启动，并进入单框集群运行状态的交换机成为主交换机。
- 当两台交换机同时启动时，集群优先级高的交换机成为主交换机。
- 当两台交换机同时启动，且集群优先级又相同时，MAC 地址小的交换机成为主交换机。
- 当两台交换机同时启动，且集群优先级和 MAC 都相同时，集群 ID 小的交换机成为主交换机。
- 版本同步
- 集群具有自动加载系统软件的功能，待组成集群的成员交换机不需要具有相同的软件版本，只需要版本间兼容即可。当主交换机选举结束后，如果备交换机与主交换机的软件版本号不一致时，备交换机会自动从主交换机下载系统软件，然后使用新的系统软件重启，并重新加入集群。

- 配置同步
- 集群具有严格的配置文件同步机制，来保证集群中的多台交换机能够像一台设备一样在网络中工作。
- 配置备份
- 交换机从非集群状态进入集群状态后，会自动将原有的非集群状态下的配置文件加上.bak 的扩展名进行备份，以便去使能集群功能后，恢复原有配置。例如，原配置文件扩展名为.cfg，则备份配置文件扩展名为.cfg.bak。去使能交换机集群功能时，用户如果希望恢复交换机的原有配置，可以更改备份配置文件名并指定其为下一次启动的配置文件，然后重新启动交换机，恢复原有配置。

CSS集群连接方式

- 设备组建集群有两种连接方式，分别为集群卡集群和业务口集群。
  - 集群卡集群方式：集群成员交换机之间通过主控板上专用的集群卡及专用的集群线缆连接。
  - 业务口集群方式：集群成员交换机之间通过业务板上的普通业务口连接，不需要专用的集群卡。同iStack，业务口集群一样涉及两种端口的概念：物理成员端口和逻辑集群端口。



- 物理成员端口
- 成员交换机之间用于集群连接的普通业务口。物理成员端口用于转发需要跨成员交换机的业务报文或成员交换机之间的集群协议报文。
- 逻辑集群端口
- 逻辑集群端口是专用于集群的逻辑端口，需要和物理成

员端口绑定。集群的每台成员交换机上支持两个逻辑集群端口。

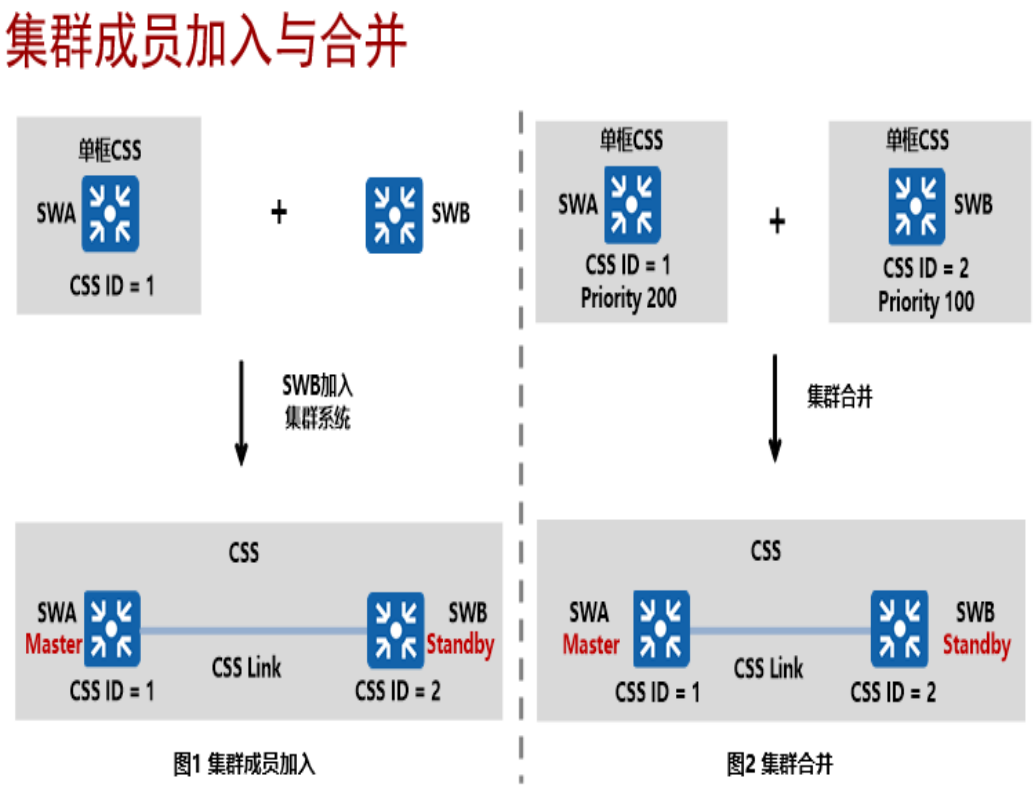


图1 集群成员加入　　图2 集群合并

- 使能了集群功能的单台交换机即为单框集群。
- 集群成员加入是指向稳定运行的单框集群系统中添加一台新的交换机。如图 1 所示，新交换机 SwitchB 将加入单框集群系统从而形成新的集群系统。原单框集群的交换机成为主交换机，新加入的交换机成为备交换机。
- 集群加入通常在以下两种情形下出现：
- 在建立集群时，先将一台交换机使能集群功能后重启，重启后这台交换机将进入单框集群状态。然后再使能另外一台交换机的集群功能后重启，则后启动的交换机则按照集群成员加入的流程加入集群系统，成为备交换机。
- 在稳定运行的两框集群场景中，将其中一台交换机重启，则这台交换机将以集群成员加入的流程重新加入集群系统，并成为备交换机。

- 集群合并是指稳定运行的两个单框集群系统合并成一个新的集群系统。如图 2 所示，两个单框集群系统将自动选出一个更优的作为合并后集群系统的主交换机。被选为主交换机的配置不变，业务也不会受到影响，框内的备用主控板将重启。而备交换机将整框重启，以集群备的角色加入新的集群系统，并将同步主交换机的配置，该交换机原有的业务也将中断。
- 集群合并通常在以下两种情形下出现：
- 将两台交换机分别使能集群功能后重启（重启后的两台交换机都属于单框集群），再使用集群线缆将两台交换机连接，之后会进入集群合并流程。
- 集群链路或设备故障导致集群分裂。故障恢复后，分裂后的两个单框集群系统重新合并。

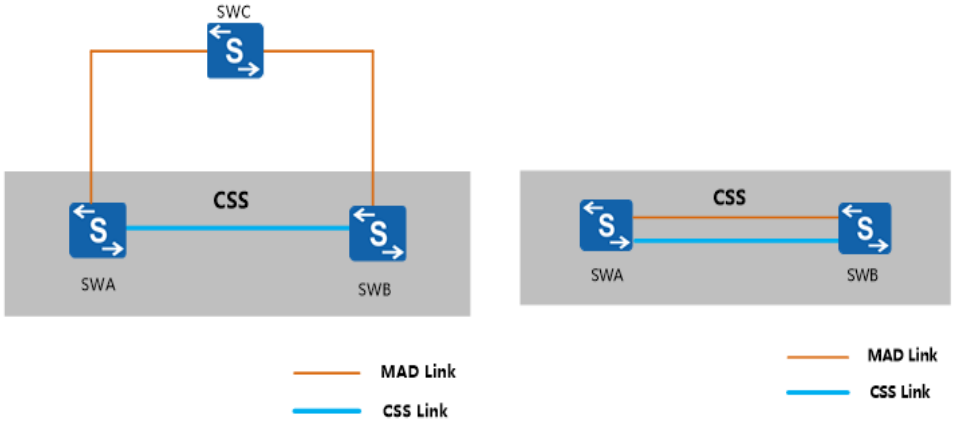
集群分裂

- 集群建立后，系统主用主控板和系统备用主控板定时发送心跳报文来维护集群系统的状态。集群线缆、集群卡、主控板等发生故障或者是其中一台交换机下电或重启将导致两台交换机之间失去通信。当两台交换机之间的心跳报文超时（超时时间为8秒）时，集群系统将分裂为两个单框集群系统，如图所示：

- 集群分裂后，由于成员交换机运行着相同的配置文件，就会产生两个具有相同IP和MAC的集群系统。为防止由此引起网络故障，必须进行IP地址和MAC地址的冲突检查。

多主检测 - 直连检测方式

- 直连检测的连接方式包括通过中间设备直连和集群成员交换机直接直连。



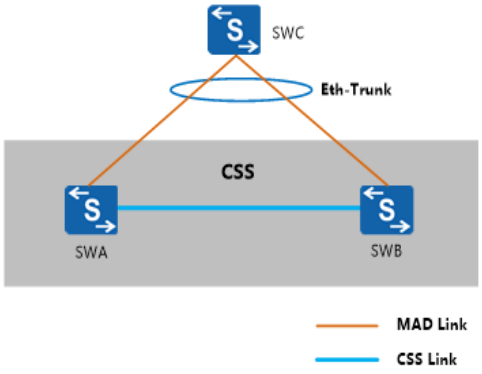
通过中间设备的直连检测方式　　通过集群成员交换机直接直连的检测方式

- 由于集群系统中所有成员交换机都使用同一个 IP 地址和 MAC 地址（集群系统 MAC），一个集群分裂后，由于这些成员交换机运行着相同的配置文件（即原集群系统的配置文件），就会产生两个具有相同 IP 地址和 MAC 地址的集群系统。为防止集群分裂后，产生两个具有相同 IP 地址和 MAC 地址的集群系统，引起网络故障，必须进行 IP 地址和 MAC 地址的冲突检查。
- 多主检测 MAD（Multi-Active Detection），是一种检测和处理集群分裂的协议。链路故障导致集群系统分裂后，MAD 可以实现集群分裂的检测、冲突处理和故障恢复，降低集群分裂对业务的影响。
- MAD 检测方式有两种：直连检测方式和代理检测方式。在同一个集群系统中，两种检测方式互斥，不可以同时配置。
- 直连检测方式是指集群成员交换机间通过普通线缆直连的专用链路进行多主检测。在直连检测方式中，集群系统正常运行时，不发送 MAD 报文；集群系统分裂后，分裂后的两台交换机周期性地通过检测链路发送 MAD 报文以进行多主冲突处理。

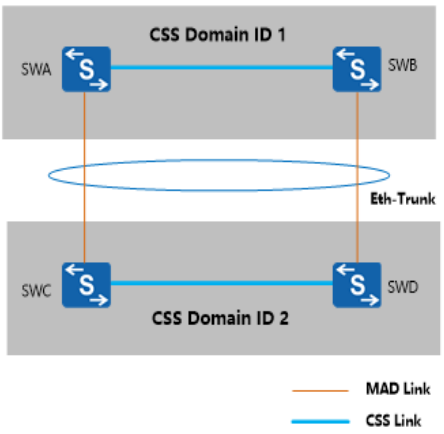
- 直连检测的连接方式包括通过中间设备直连和集群成员交换机直接直连：
- 通过中间设备直连：集群系统的成员交换机之间至少有一条检测链路与中间设备相连。此种方式适用于成员交换机相距较远的场景。
- 直接直连：集群成员交换机直接直连可以避免由中间设备故障导致 MAD 检测失败。

多主检测 - 代理检测方式

- 代理检测方式可分为单机作代理和两套集群系统互为代理。



单机作代理设备的代理检测方式



两套集群系统互为代理的代理检测方式

- 代理检测方式是在集群系统 Eth-Trunk 上启用代理检测，在代理设备上启用 MAD 检测功能。此种检测方式要求集群系统中的所有成员交换机都与代理设备连接，并将这些链路加入同一个 Eth-Trunk 内。与直连检测方式相比，代理检测方式无需占用额外的接口，Eth-Trunk 接口可同时运行 MAD 代理检测和其他业务。
- 在代理检测方式中，集群系统正常运行时，集群成员交换机以 30s 为周期通过检测链路发送 MAD 报文。集群成员交换机对在正常工作状态下收到的 MAD 报文不做任何处理；集群分裂后，分裂后的两台交换机周期性地通过检测链路发送 MAD 报文以进行多主冲突处理。

- MAD 冲突处理
- 集群分裂后，MAD 冲突处理机制会使分裂后的单框集群系统处于 Detect 状态或 Recovery 状态。Detect 状态表示集群正常工作状态，Recovery 状态表示集群禁用状态。
- MAD 冲突处理机制如下：MAD 分裂检测机制会检测到网络中存在两个处于 Detect 状态的集群系统即两台交换机，此时会进行集群优先级比较（优先级相同比较 MAC 地址，MAC 地址相同则比较集群 ID），优先级高的交换机将成为主交换机继续正常工作，另一台交换机会转入 Recovery 状态；并且在 Recovery 状态的交换机上，关闭除保留端口以外的其它所有物理端口，以保证该交换机不再转发业务报文。
- MAD 故障恢复
- 通过修复故障链路，分裂后的集群系统重新合并为一个集群系统。重新合并的方式有以下两种：
- 集群链路修复后，处于 Recovery 状态的集群系统重新启动，与 Detect 状态的集群系统合并，同时将被关闭的业务端口恢复 Up，整个集群系统恢复。
- 如果故障链路修复前，承载业务的 Detect 状态的集群系统也出现了故障。此时，可以先将 Detect 状态的集群系统从网络中移除，再通过命令行启用 Recovery 状态的集群系统，接替原来的业务，然后再修复原 Detect 状态集群系统的故障。故障修复后，重新合并集群系统。

集群配置

- 通过集群卡连接方式组建集群

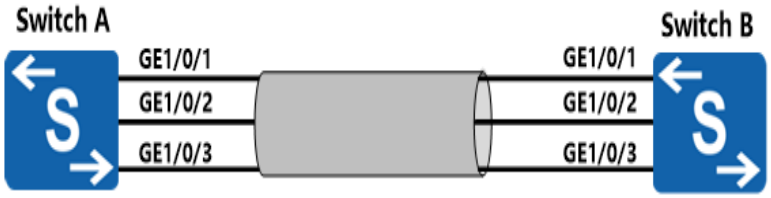
```
[SwitchA] set css mode css-card     \\配置集群卡连接方式
[SwitchA] set css id 1              \\配置成员交换机的集群ID
[SwitchA] set css priority 100      \\配置设备的集群优先级
[SwitchA] css enable                \\使能交换机的集群功能
```

- 通过业务口连接方式组建集群

```
[SwitchA] set css mode lpu      \\配置业务口连接方式
[SwitchA] set css id 1          \\配置成员交换机的集群ID
[SwitchA] set css priority 100  \\配置设备的集群优先级
[SwitchA] interface css-port 1  \\进入逻辑集群端口视图
[SwitchA-css-port1] port interface xgigabitethernet 1/0/1 to xgigabitethernet 1/0/2 enable
                    \\配置业务口为物理成员端口，并将物理成员端口加入到逻辑集群端口中
[SwitchA] interface css-port 2
[SwitchA-css-port2] port interface xgigabitethernet 2/0/1 to xgigabitethernet 2/0/2 enable
[SwitchA] css enable            \\使能交换机的集群功能
```

Eth-Trunk基本原理

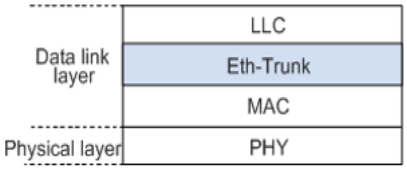
- 以太网链路聚合Eth-Trunk简称链路聚合，它通过将多条以太网物理链路捆绑在一起成为一条逻辑链路，从而实现增加链路带宽的目的。同时，这些捆绑在一起的链路通过相互间的动态备份，可以有效地提高链路的可靠性。
- Trunk接口连接的链路可以看成是一条点到点的直连链路，在一个Trunk内，可以实现流量负载分担，同时也提供了更高的连接可靠性和更大的带宽。用户通过对逻辑口进行配置，实现各种路由协议以及其它业务部署。



- 链路聚合技术主要有以下三个优势：增加带宽、提高可靠性和负载分担。
- 链路聚合组和成员接口
- 链路聚合组 LAG 是指将若干条以太链路捆绑在一起所形成的逻辑链路。组成 Eth-Trunk 接口的各个物理接口称为成员接口。
- 活动接口和非活动接口、活动链路和非活动链路
- 链路聚合组的成员接口存在活动接口和非活动接口两种。转发数据的接口称为活动接口，不转发数据的接口称为非活动接口。
- 活动接口对应的链路称为活动链路，非活动接口对应的链路称为非活动链路。
- 活动接口数上限阈值
- 当前活动链路数目达到上限阈值时，再向 Eth-Trunk 中添加成员接口，不会增加 Eth-Trunk 活动接口的数目，超过上限阈值的链路状态将被置为 Down，作为备份链路。
- 活动接口数下限阈值
- 设置活动接口数下限阈值是为了保证最小带宽，当前活动链路数目小于下限阈值时，Eth-Trunk 接口的状态转为 Down。
- 设备支持的链路聚合方式
- 同板：是指链路聚合时，同一聚合组的成员接口分布在同一单板上。
- 跨板：是指链路聚合时，同一聚合组的成员接口分布在不同的单板上。
- 跨框：是指在集群场景下，成员接口分布在集群的各个成员设备上。
- 跨设备：是指 E-Trunk 基于 LACP 进行了扩展，能够实现多台设备间的链路聚合。

转发原理

- Eth-Trunk位于MAC与LLC子层之间，属于数据链路层。



- Eth-Trunk模块内部维护一张转发表，这张表由以下两项组成。
  - HASH-KEY值：HASH-KEY值是根据数据包的MAC地址或IP地址等，经HASH算法计算得出。
  - 接口号：Eth-Trunk转发表表项分布和设备每个Eth-Trunk支持加入的成员接口数量相关，不同的HASH-KEY值对应不同的出接口。

HASH-KEY	0	1	2	3	4	5	6	7
PORT	1	2	3	4	1	2	3	4

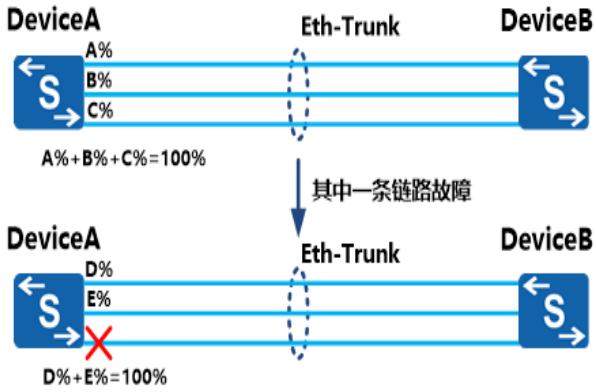
Eth-Trunk转发表示例

- Eth-Trunk 模块根据转发表转发数据帧的过程如下：
- Eth-Trunk 模块从 MAC 子层接收到一个数据帧后，根据负载分担方式提取数据帧的源 MAC 地址/IP 地址或目的 MAC 地址/IP 地址。
- 根据 HASH 算法进行计算，得到 HASH-KEY 值。
- Eth-Trunk 模块根据 HASH-KEY 值在转发表中查找对应的接口，把数据帧从该接口发送出去。
- 例如 ，某设备每 Eth-Trunk 支持最大加入接口数为 8 个，将接口 1、2、3、4 捆绑为一个 Eth-Trunk 接口，此时生成的转发表如图 2 所示。其中 HASH-KEY 值为 0、1、2、3、4、5、6、7，对应的出接口号分别为 1、2、3、4、1、2、3、4。
- 为了避免数据包乱序情况的发生，Eth-Trunk 采用逐流负载分担的机制，其中如何转发数据则由于选择不同的负载分担方式而有所差别。
- 负载分担的方式主要包括以下几种，用户可以根据具体应用选择不同的负载分担方式。
- 根据报文的源 MAC 地址进行负载分担；
- 根据报文的目的 MAC 地址进行负载分担；

- 根据报文的源 IP 地址进行负载分担；
- 根据报文的目的 IP 地址进行负载分担；
- 根据报文的源 MAC 地址和目的 MAC 地址进行负载分担；
- 根据报文的源 IP 地址和目的 IP 地址进行负载分担；
- 根据报文的 VLAN、源物理端口等对 L2、IPv4、IPv6 和 MPLS 报文进行增强型负载分担。

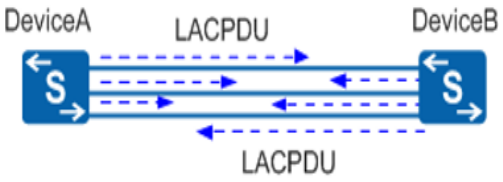
链路聚合 - 手工模式链路聚合

- 手工模式下，Eth-Trunk的建立、成员接口的加入由手工配置，没有链路聚合控制协议LACP的参与。如图所示，DeviceA与DeviceB之间创建Eth-Trunk，手工模式下三条活动链路都参与数据转发并分担流量。当一条链路故障时，故障链路无法转发数据，链路聚合组自动在剩余的两条活动链路中分担流量。



链路聚合 - LACP模式链路聚合 (1)

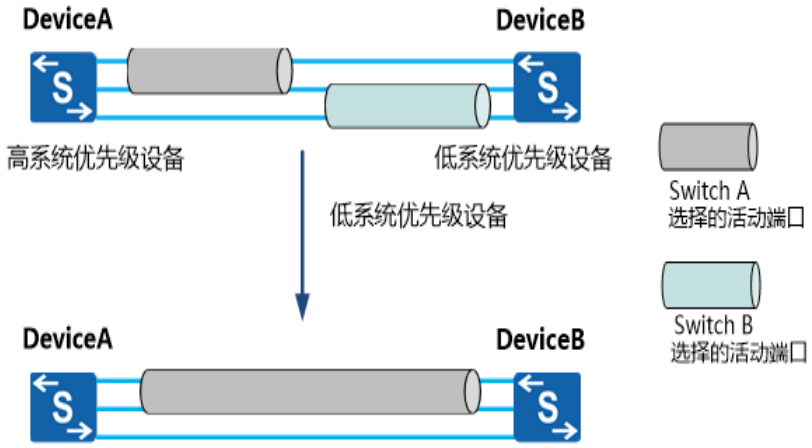
- 为了提高Eth-Trunk的容错性，并且能提供备份功能，保证成员链路的高可靠性，出现了链路聚合控制协议LACP（Link Aggregation Control Protocol）。聚合链路形成以后，LACP负责维护链路状态，在聚合条件发生变化时，自动调整或解散链路聚合。
- LACP模式Eth-Trunk建立的过程如下：
  1. 两端互相发送LACPDU报文。在DeviceA和DeviceB上创建Eth-Trunk并配置为LACP模式，然后向Eth-Trunk中手工加入成员接口。此时成员接口上便启用了LACP协议，两端互发LACPDU报文。



- 作为链路聚合技术，手工模式 Eth-Trunk 可以完成多个物理接口聚合成一个 Eth-Trunk 口来提高带宽，同时能够检测到同一聚合组内的成员链路有断路等有限故障，但是无法检测到链路层故障、链路错连等故障。

链路聚合 - LACP模式链路聚合 (2)

2. 确定主动端和活动链路。



- 如图所示，两端设备均会收到对端发来的 LACPDU 报文。以 DeviceB 为例，当 DeviceB 收到 DeviceA 发送的报文时，DeviceB 会查看并记录对端信息，然后比较系统优先级字段，如果 DeviceA 的系统优先级高于本端的系统优先级，则确定 DeviceA 为 LACP 主动端。如果 DeviceA 和 DeviceB 的系统优先级相同，比较两端设备的 MAC 地址，确定 MAC 地址小的一端为 LACP 主动端。
- 选出主动端后，两端都会以主动端的接口优先级来选择活动接口，如果主动端的接口优先级都相同则选择接口编号比较小的为活动接口。两端设备选择了一致的活动接口，活动链路组便可以建立起来，从这些活动链路中以负载分担的方式转发数据。

配置链路聚合

- 配置手工模式链路聚合

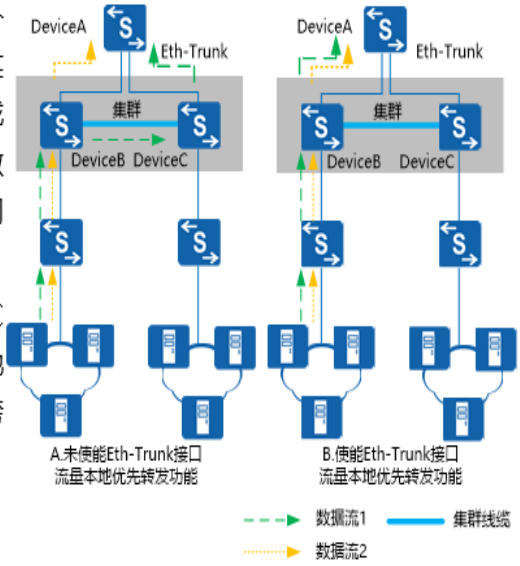
```
[SwitchA] interface eth-trunk 1
\\创建Eth-Trunk接口，并进入Eth-Trunk接口视图
[SwitchA-Eth-Trunk1] mode manual load-balance
\\配置链路聚合模式为手工模式
[SwitchA-Eth-Trunk1] trunkport gigabitethernet 0/0/1 to 0/0/3
\\将成员接口加入聚合组
```

- 配置LACP模式的链路聚合

```
[SwitchA] interface eth-trunk 1
[SwitchA-Eth-Trunk1] mode lacp
\\配置链路聚合模式为LACP模式
[SwitchA-Eth-Trunk1] max active-linknumber 2  \\配置活动接口上限阈值为2
[SwitchA] interface gigabitethernet 0/0/1  \\将成员接口加入聚合组
[SwitchA-GigabitEthernet0/0/1] eth-trunk 1
[SwitchA] interface gigabitethernet 0/0/2
[SwitchA-GigabitEthernet0/0/2] eth-trunk 1
[SwitchA] interface gigabitethernet 0/0/3
[SwitchA-GigabitEthernet0/0/3] eth-trunk 1
```

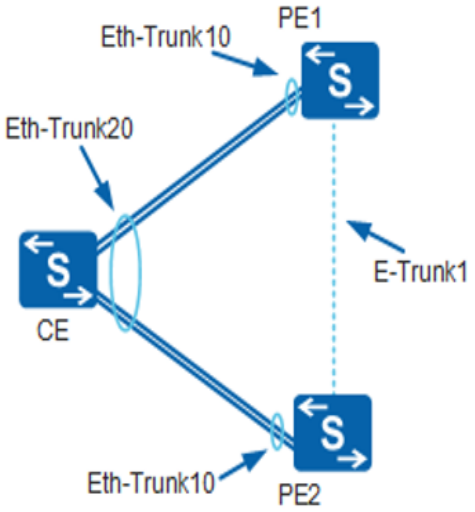
集群环境下的链路聚合

- 将集群设备不同设备中的物理接口聚合到一个逻辑接口Eth-Trunk接口中。当集群设备中某台设备故障或加入Eth-Trunk接口中的物理成员口故障，可通过集群设备间线缆跨框传输数据流量，从而保证了数据流量的可靠传输，同时实现了设备间的备份。
- 在网络无故障的情况下从DeviceB或DeviceC上来的流量，通过本设备中的成员口优先本地转发，而不是像A图中通过集群设备间线缆跨框转发。



Eth-Trunk与E-Trunk

- E-Trunk（Enhanced Trunk）是一种实现跨设备链路聚合的机制，基于LACP（单台设备链路聚合的标准）进行了扩展，能够实现多台设备间的链路聚合，从而把链路可靠性从单板级提高到了设备级。
- 如图所示，CE双归接入PE1和PE2，通过在PE节点部署E-Trunk，当CE至PE1的链路或PE1节点故障时，流量可以切换到CE至PE2的链路，从而实现设备级保护。



- E-Trunk 机制主要应用于 CE 双归接入 VPLS、VLL、PWE3 网络时，CE 与 PE 间的链路保护以及对 PE 设备节点故障的保护。在没有使用 E-Trunk 前，CE 通过 Eth-Trunk 链路只能单归到一个 PE 设备。如果 Eth-Trunk 出现故障或者 PE 设备故障，CE 将无法与 PE 设备继续进行通信。使用 E-Trunk 后，CE 可以双归到 PE 上，从而实现设备间保护。
- 如图，CE 分别与 PE1 和 PE2 直连，PE1 和 PE2 之间运行 E-Trunk。PE 侧，需要在 PE1 和 PE2 设备上分别创建 ID 相同的 E-Trunk 和 Eth-Trunk，并将 Eth-Trunk 加入到 E-Trunk。CE 侧，在 CE 设备上配置 LACP 模式的 Eth-Trunk，此 Eth-Trunk 分别与 PE1 和 PE2 设备相连。对 CE 设备而言，E-Trunk 不可见。
- PE1 与 PE2 设备之间通过 E-Trunk 报文进行主备协商，确定 E-Trunk 的主备状态。正常情况下两台 PE 的协商结果是一个为主用一个为备用。PE 设备上 E-Trunk 主备状态是根据报文中所携带的 E-Trunk 优先级和 E-Trunk 系统 ID 确定的。优先级的数值越小，优先级越高，优先级高的为主用。如果

E-Trunk 优先级相同，那么 E-Trunk 系统 ID 小的为主用。PE 1 为主，PE1 的 Eth-Trunk 10 为主，链路状态为 Up。PE2 为备，PE2 的 Eth-Trunk 10 为备，链路状态为 Down。

- 如果 CE 到 PE1 间的链路出现故障：PE1 会向对端发送 E-Trunk 报文，报文中携带 PE1 的 Eth-Trunk 10 故障的信息。PE2 收到 E-Trunk 报文后，发现对端 Eth-Trunk 10 故障，则 PE2 设备上 Eth-Trunk 10 的状态将变为主。然后经过 LACP 协商，PE2 设备上的 Eth-Trunk 10 的状态变为 Up。这样 PE2 设备的 Eth-Trunk 状态变为 Up，CE 的流量会通过 PE2 转发，以达到对 CE 的流量进行保护的目的。
- 如果 PE1 设备出现故障：如果 PE 设备上配置了 BFD，PE2 检测到 BFD 会话状态为 Down 后，PE2 设备从备用状态变为主用状态，PE2 的 Eth-Trunk 10 状态也变为主。如果 PE 设备上没有配置 BFD，PE2 设备上的定时器超时后仍然没有收到 PE1 设备发送的 E-Trunk 报文，PE2 设备从备用状态变为主用状态，PE2 的 Eth-Trunk 10 状态也变为主。经过 LAC P 协商，PE2 设备上的 Eth-Trunk 10 的状态变为 Up。CE 的流量会通过 PE2 转发，以达到对 CE 的流量进行保护的目的。

思考题

1. 如何清除MAC地址表项和ARP表项?
2. MSTP域如何配置?
3. Eth-Trunk是否支持抢占功能?

- 参考答案：
- 如何清除 MAC 地址表项和 ARP 表项？
- 清除所有动态 MAC（系统视图）：undo mac-address d ynamic

- 清除所有静态 MAC（系统视图）：undo mac-address static
- 删除一条静态 ARP 表项（系统视图）：undo arp static
- 删除多条 ARP 表项（用户视图）：reset arp
- MSTP 域如何配置？
- MSTP 的域信息在 stp region-configuration 视图下配置，同一个域中各台设备的域配置信息必须完全一致。存在任何一点差异，就不在同一个域中。MSTP 可以配置的域信息有：
- Format selector：格式选择符，在命令行不能配置，默认为 0；
- Region name：域名，默认是桥 MAC 地址；
- Revision level：修订级别，默认是 0；
- Instance/Vlans Mapped：实例和 VLAN 映射表，默认全部 VLAN 映射到实例 0。
- Eth-Trunk 是否支持抢占功能？
- 只有在 LACP 模式下，Eth-Trunk 才支持优先级抢占功能，可以执行 lacp preempt enable 命令使能优先级抢占功能。在 LACP 模式下，当活动链路中出现故障链路时，系统会从备用链路中选择优先级最高的链路替代故障链路；如果被替代的故障链路恢复了正常，而且该链路的优先级又高于替代自己的链路，这种情况下，如果使能了 LACP 优先级抢占功能，高优先级链路会抢占低优先级链路，回切到活动状态。要求 Eth-Trunk 两端 LACP 抢占功能使能情况配置一致，即：统一使能或不使能。