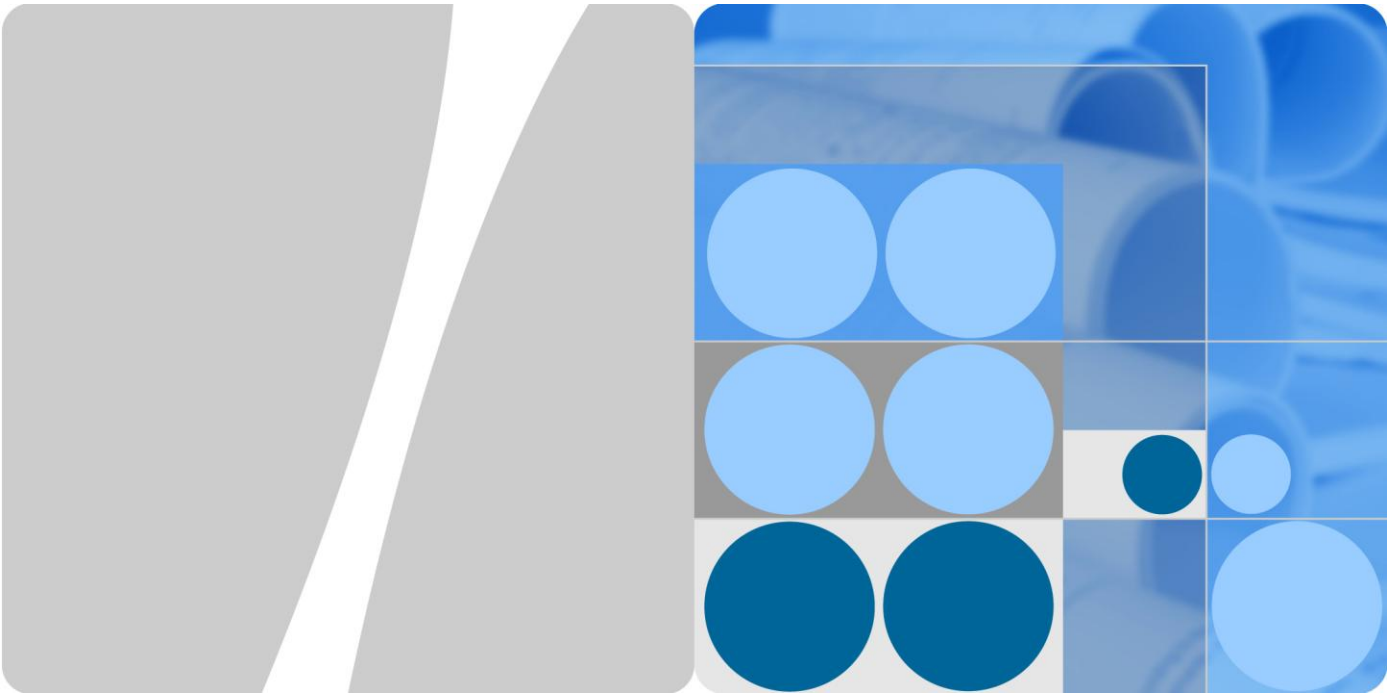


CSS 技术白皮书

文档版本 01

发布日期 2012-08-15

华为技术有限公司

商标声明


和其他华为商标均为华为技术有限公司的商标。

本文档提及的其他所有商标或注册商标，由各自的所有人拥有。

注意

您购买的产品、服务或特性等应受华为公司商业合同和条款的约束，本文档中描述的全部或部分产品、服务或特性可能不在您的购买或使用范围之内。除非合同另有约定，华为公司对本文档内容不做任何明示或默示的声明或保证。

由于产品版本升级或其他原因，本文档内容会不定期进行更新。除非另有约定，本文档仅作为使用指导，本文档中的所有陈述、信息和建议不构成任何明示或暗示的担保。

华为技术有限公司

地址： 深圳市龙岗区坂田华为总部办公楼 邮编：518129

网址： http://www.huawei.com

客户服务邮箱： support@huawei.com

客户服务电话： 0755-28560000 4008302118

客户服务传真： 0755-28560111

目 录

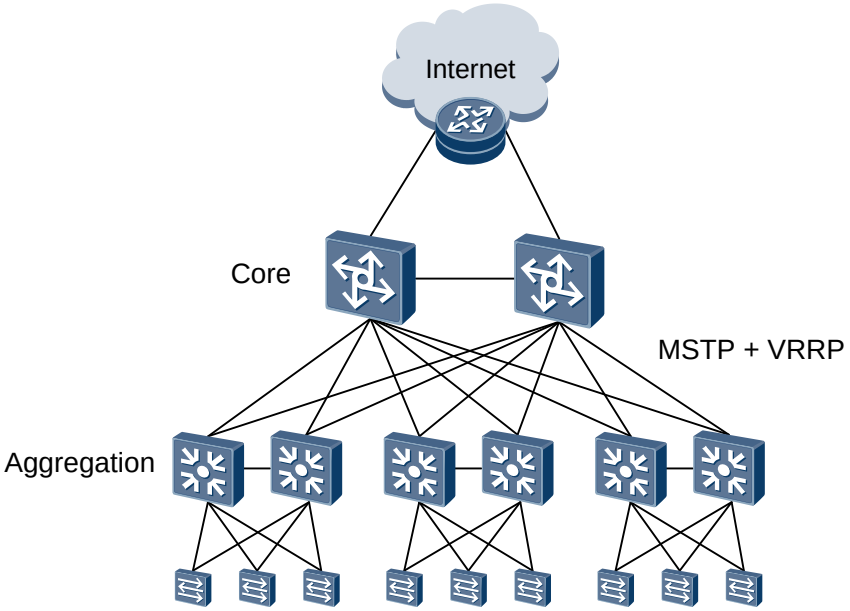
插图目录

1 CSS

1.1 介绍

在网络核心层和汇聚层，常使用双节点冗余设计提高网络的可靠性，如图 1-1 所示。

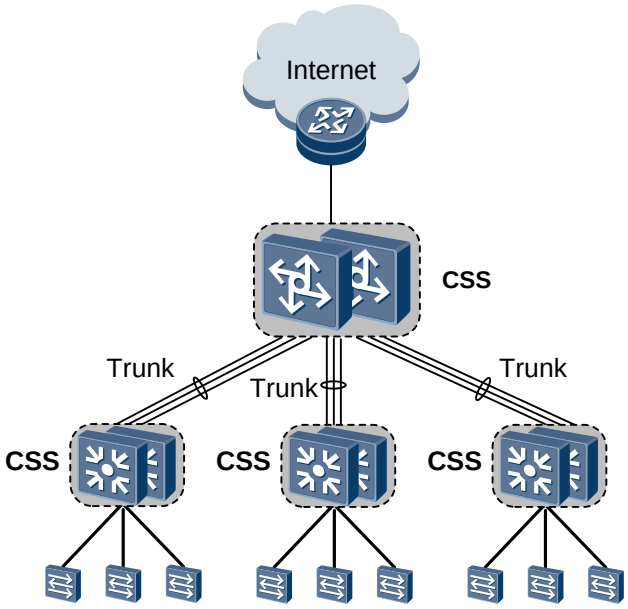
图1-1 冗余网络结构



冗余结构虽然提高了网络的可靠性，但是也使得网络结构和互联关系变得复杂，一般都需要部署 MSTP 等协议消除环路，同时运行 VRRP 等协议来支持节点冗余备份，导致网络协议的部署变得复杂。

CSS（Cluster Switch System，集群交换系统），又被称为堆叠，是指将多台支持堆叠特性的交换机设备组合在一起（目前支持 2 台），从逻辑上组合成一台整体交换设备，如图 1-2 所示。

图1-2 CSS 组网图

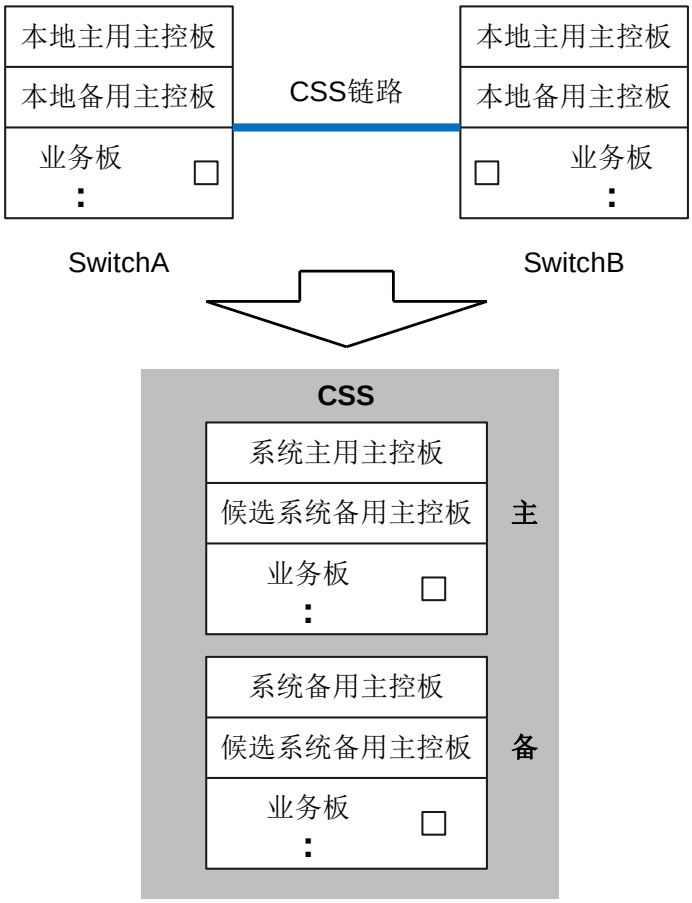


CSS 可将两台交换机的控制平面和转发平面都合一，带来的好处有：

- 高可靠性。堆叠系统多台成员设备之间冗余备份；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。
- 简化网络结构和协议部署。堆叠技术可以将复杂的网络拓扑结构简化为层次分明、互联关系简单的网络结构，网络各层之间通过链路聚合，自然消除环路，不需要再部署 MSTP、VRRP 等协议。
- 简化配置和管理。堆叠形成后，多台物理设备虚拟成为一台设备，用户可以通过登录堆叠系统，对堆叠系统所有成员设备进行统一配置和管理。

CSS 系统建立后，根据协议的计算，将出现一个主交换机，一个备交换机。在控制平面上，主交换机的主用主控板成为 CSS 的系统主，作为整个系统的管理主角色；备交换机的主用主控板成为 CSS 的系统备，作为系统的管理备角色；主交换机和备交换机的备用主控板作为 CSS 的候选系统备，不具有管理角色，只作为交换网代理（即冷备板）。

图1-3 堆叠竞争效果图



1.2 原理描述

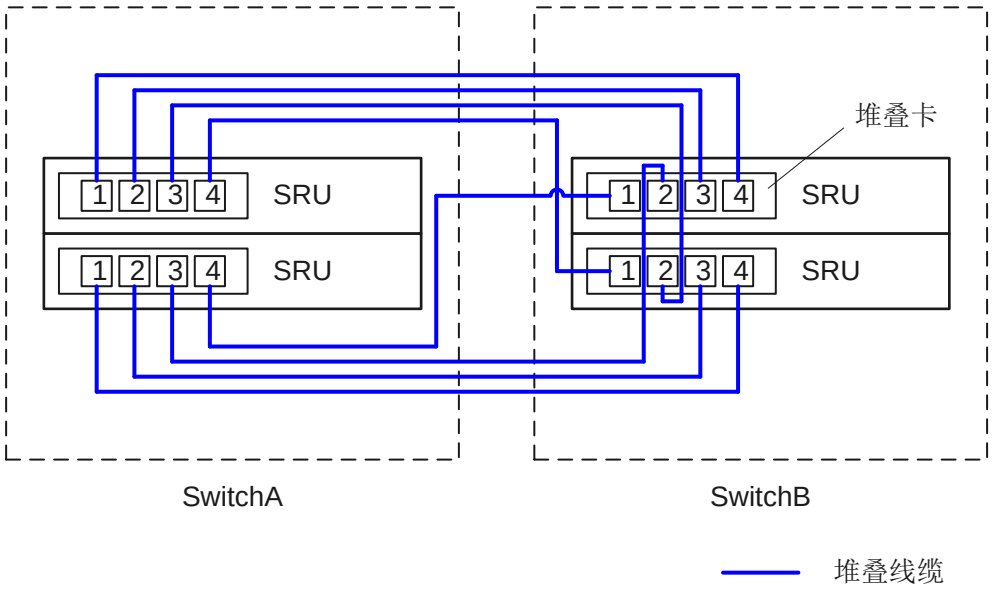
CSS 物理连接

堆叠成员交换机之间的连接方式有堆叠卡连接和业务口连接两种。

- **堆叠卡连接**

  堆叠成员交换机之间通过主控板上的堆叠卡连接（每块堆叠卡上有 4 个堆叠口）。两台设备都有两块主控板的情况下，通过专用的堆叠电缆 QSFP+高速线缆或 QSFP+光模块和光纤将这 8 组堆叠口按照图 1-4 规则连接起来。堆叠口连接规则是固定的，所有堆叠口都要插上堆叠线缆，不能随意连接。

图1-4 堆叠卡连接规则



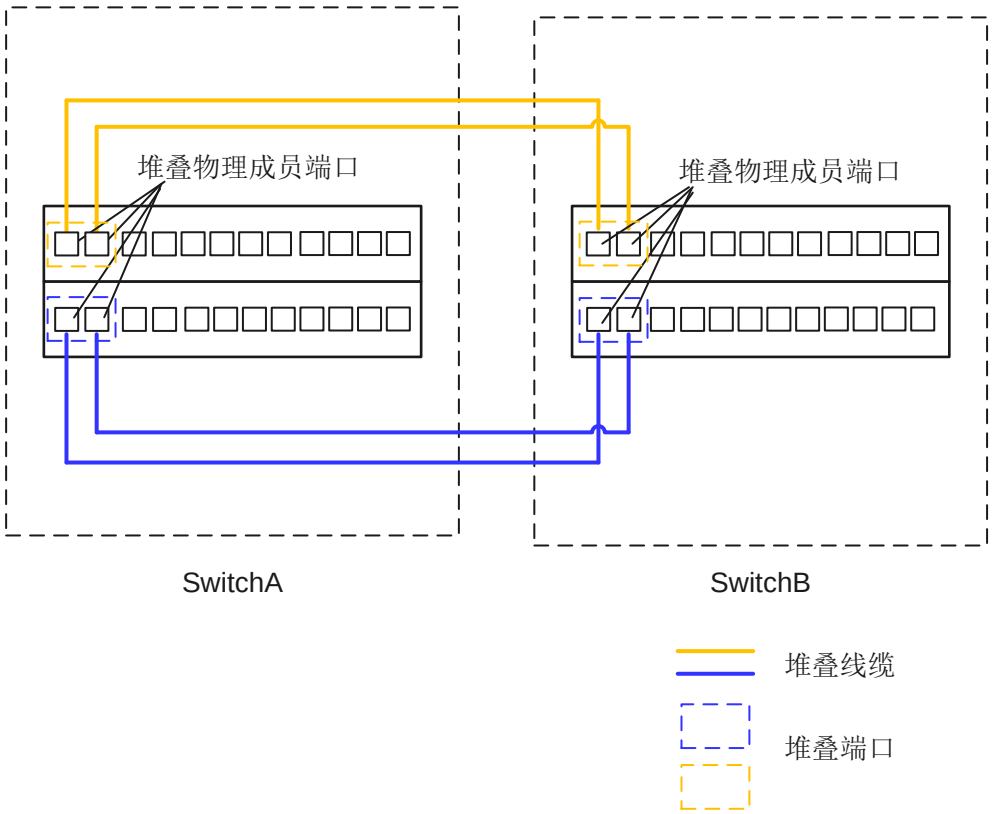
说明

S9700 系列交换机不支持堆叠卡连接方式。

- **业务口连接**

  堆叠成员交换机之间通过 LPU 上的普通业务口连接。将 LPU 上的业务口配置为堆叠物理成员端口后加入逻辑堆叠端口，通过 SFP+光模块和光纤或 SFP+堆叠线缆将堆叠物理成员端口按照图 1-5 规则连接起来。

图1-5 业务口连接规则

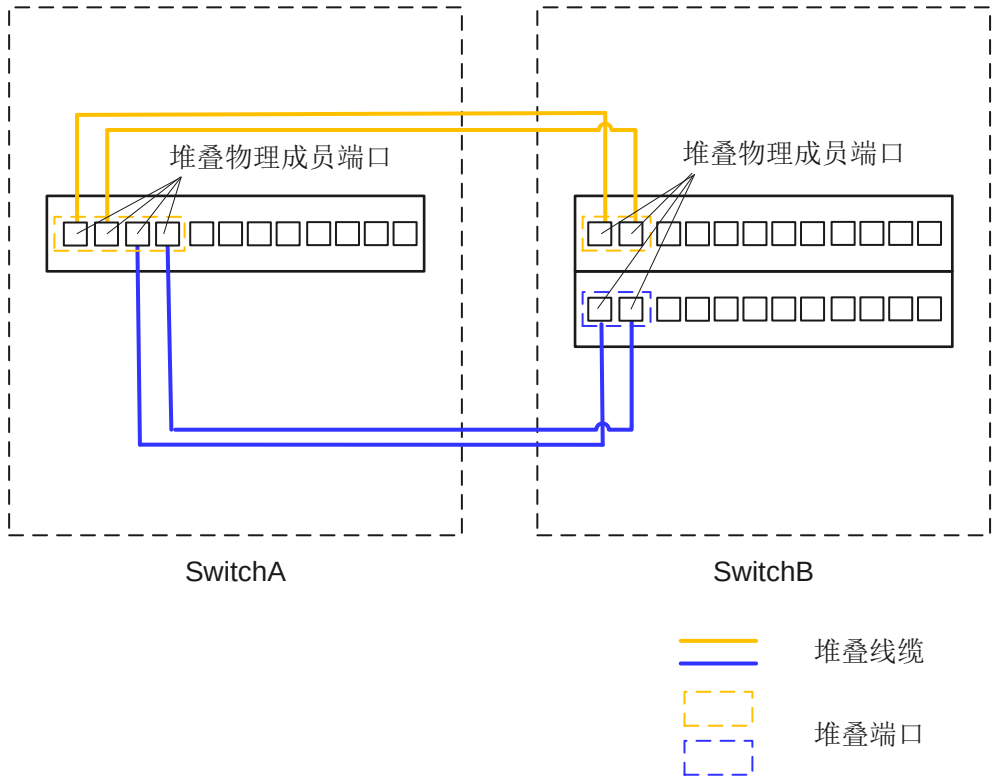


业务口堆叠具有灵活的组网形式，每块单板可配置 32 个堆叠物理成员端口，提高了堆叠链路的带宽和可靠性。

业务口堆叠按照链路的分布，有两种组网形式。

- 1+0 组网：配置一个逻辑堆叠端口，物理堆叠端口分布在一块单板上，依靠一块单板上的堆叠链路实现堆叠连接。
- 1+1 组网：配置两个逻辑堆叠端口，物理堆叠端口分布在两块单板上，不同单板上的堆叠链路形成备份。

图1-6 错连示意图



CSS 竞争规则

堆叠系统启动后，通过竞争，一台设备成为堆叠主交换机，另一台设备成为堆叠备交换机。竞争的规则如下：

1. 运行状态比较，已经正常运行的交换机优先处于启动状态的交换机竞争为主交换机。
2. 堆叠优先级比较，堆叠优先级高的交换机优先竞争为主交换机。
3. MAC 地址比较，MAC 地址小的交换机优先竞争为主交换机。

堆叠系统建立之前，每台交换机都是单独的实体，每台交换机有自己独立的 IP 地址，用户需要独立的管理所有的交换机；堆叠建立后堆叠成员对外体现为一个统一的逻辑实体，用户使用一个 IP 地址对堆叠中的所有交换机进行管理和维护。

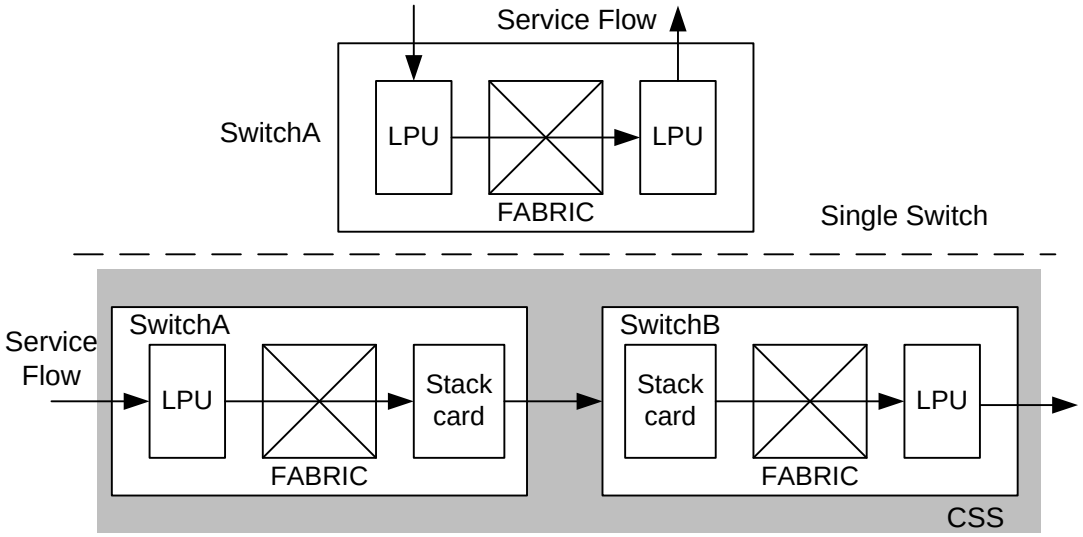
CSS 环境下的配置和转发

堆叠建立后，可以通过接口板上的业务端口、系统主用主控板上的串口或网管口登陆堆叠系统，进行业务配置和系统管理。

堆叠提供四维的接口视图（框号/槽位号/子卡号/端口号），支持对两台设备中的所有端口进行业务相关配置、操作，以框/槽为单位对两台设备中的所有单板进行管理。

在堆叠环境下，业务流量转发与单框环境下不同，跨设备的转发需要经过交换网两次。对于报文内容的处理没有区别，都需要进行一次上、下行处理。

图1-7 CSS 环境 VS 单框环境下的流量转发



主备倒换

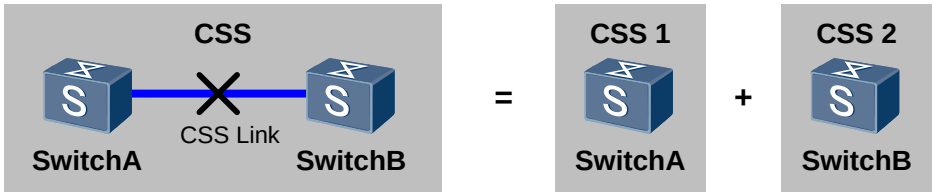
在主交换机或备交换机单框内的两块主控板倒换后，该框内的备用主控板升为 CSS 的系统备用主控板。

- 主交换机内的两块主控板发生倒换：备交换机升为主交换机，原来的系统备用主控板升为系统主用主控板；主交换机降为备交换机，原来的系统主用主控板重启，原来主交换机框内的备用主控板升为 CSS 的系统备用主控板，从系统主用主控板进行 HA 同步。
- 备交换机内的两块主控板发生倒换：主交换机和备交换机的角色不会发生变化。备交换机内的主用主控板（即原来 CSS 的系统备用主控板）重启，备用主控板升为系统备用主控板，从系统主用主控板进行 HA 同步。通过这种处理，保证了堆叠的高可靠性。

CSS 分裂处理

堆叠系统建立后，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堆叠系统的状态。堆叠电缆、堆叠卡或主控板等发生故障可能会导致两台交换机之间失去通信，导致两台交换机之间的心跳报文超时，此时堆叠系统将分裂为两台独立的交换机。

图1-8 CSS 分裂示意图



堆叠系统分裂后，若两台交换机都在正常运行，其全局配置完全相同，会以相同的 IP 地址、MAC 地址和网络中的其他设备交互，导致 IP 地址和 MAC 地址冲突，引起整个网络故障。

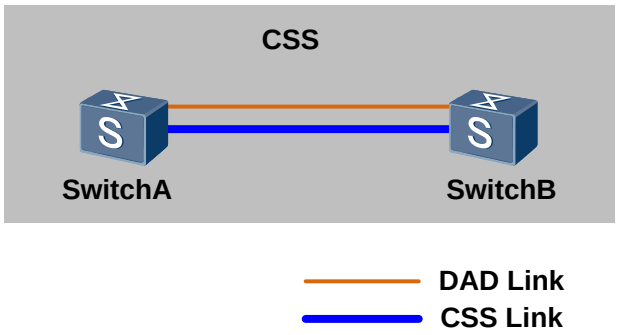
双主检测 DAD（Dual-Active Detect），是一种检测和处理堆叠分裂的协议，可以实现堆叠分裂的检测、冲突处理和故障恢复，降低堆叠分裂对业务的影响。

双主检测方式有两种：直连检测方式和 Relay 代理检测方式。

- 直连检测方式

  如图 1-7 所示，堆叠成员设备间通过专用直连链路进行双主检测。

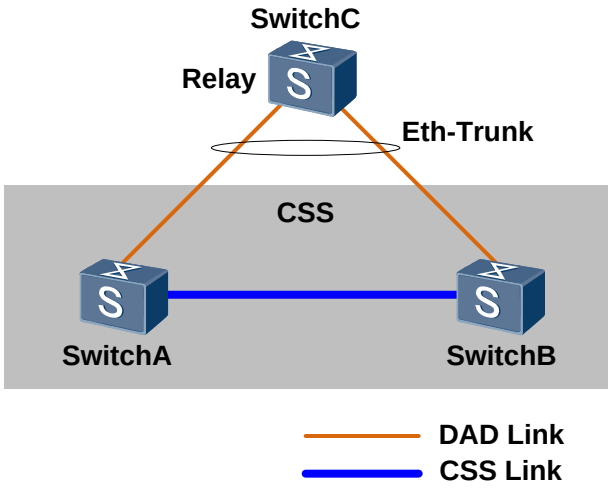
图1-9 直连方式双主检测示意图



- Relay 代理检测方式

  如图 1-8 所示，Relay 代理检测方式在堆叠系统跨设备 Eth-Trunk 上启用 DAD 检测，在指定的中间设备上启用 DAD 代理。

图1-10 Relay 代理方式双主检测示意图



堆叠分裂后，两台交换机会在检测链路上相互发送 DAD 竞争报文。如果本交换机竞争为主，则不做处理，保持 Active 状态，正常转发业务报文；如果本交换机竞争为备，则需要关闭除保留端口外的所有业务端口，转入 Recovery 状态，停止转发业务报文。

堆叠链路故障修复后，处于 Recovery 状态的交换机将重新启动，同时将被关闭的业务端口恢复 Up，整个堆叠系统恢复。

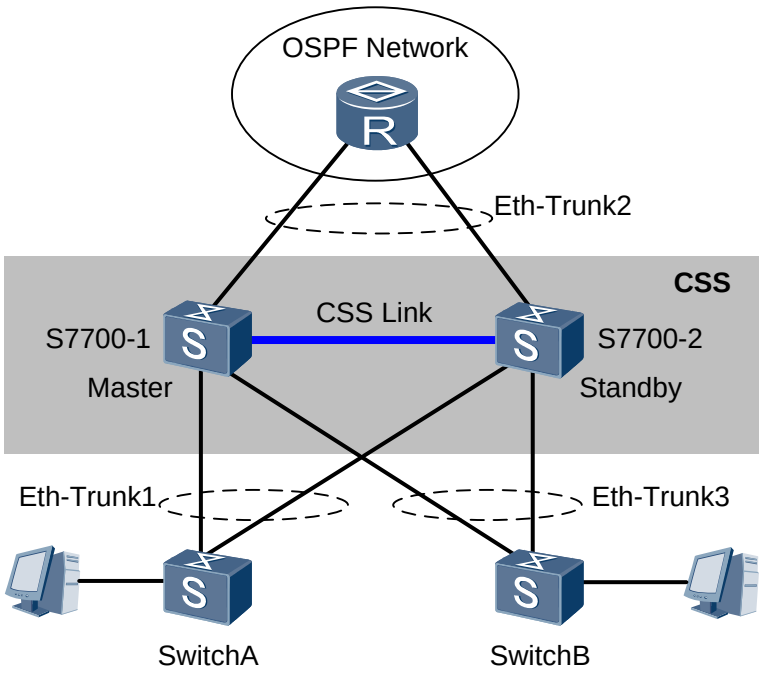
1.3 应用场景

提高可靠性

如图 1-9 所示，S7700-1 和 S7700-2 组成堆叠系统；SwitchA 连接用户，通过跨框 Eth-Trunk1 连接堆叠系统；SwitchB 连接用户，通过物理口连接堆叠系统；堆叠系统通过跨框 Eth-Trunk2 接入 OSPF 网络。

通过跨框 Eth-Trunk，用户可以将不同成员设备上的物理以太网端口配置成一个聚合端口，这样即使某些端口所在的设备出现故障，也不会导致聚合链路完全失效，其它正常工作的成员设备会继续管理和维护剩下的聚合端口，这样即可以增大设备容量，又可以设备间的备份，增加可靠性。

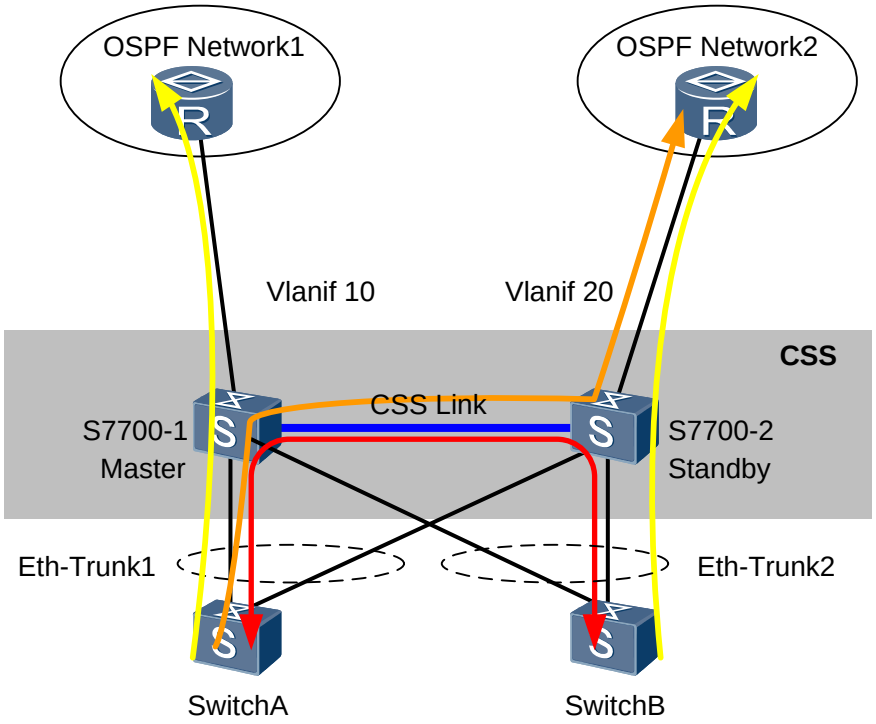
图1-11 CSS 典型组网 1



如图 1-10 上图所示，不同设备上的物理端口绑定不同的 VLAN，通过 VLanif 上行；SwitchA 下行通过跨框 Eth-Trunk 接入，从 SwitchA 下行接入的流量可以从上行 VLanif10 或者 Vlanif20 转发出去。如果 ECMP 算法选择本框（S7700-1）的上行物理接口，则直接从本框就转发了；如果 ECMP 算法选择非本框（即 S7700-2）的上行物理接口，则要通过主控板的 HiGig 接口转发到 S7700-2，由 S7700-2 从上行接口转发出去。

这样当某台设备或物理端口故障，业务可以自动却换到另外一台设备，即可以增大设备容量，又可以设备间的备份，增加可靠性。

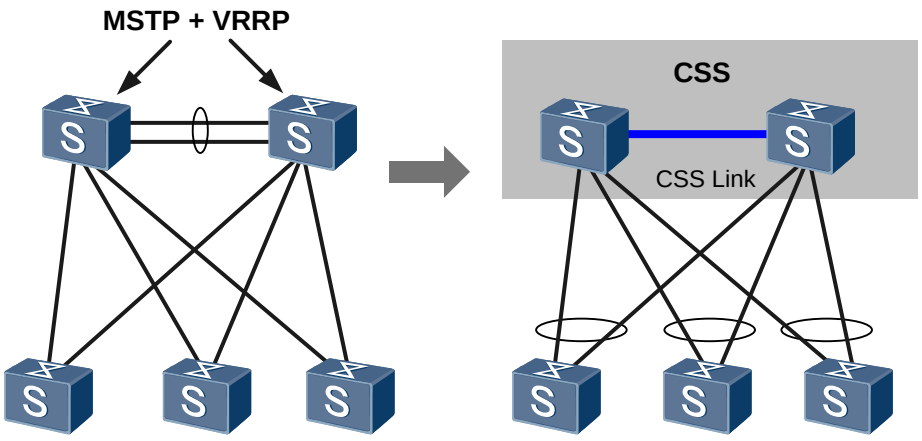
图1-12 CSS 组网图 2



简化组网

如图 1-11 所示，网络中的多台设备组成堆叠，虚拟成单一的逻辑设备。简化后的组网不再需要使用 MSTP、VRRP 等协议，简化了网络配置，同时依靠跨设备的链路聚合，实现快速收敛，提高了可靠性。

图1-13 简化组网示意图



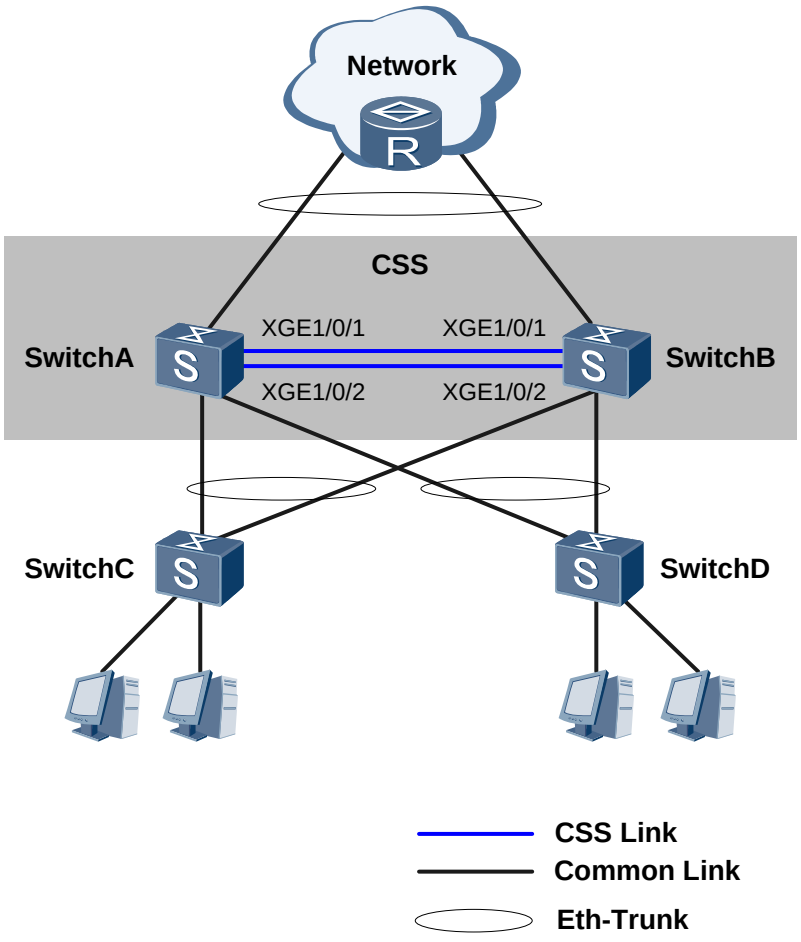
1.4 配置 CSS

组网需求

由于网络规模迅速扩大，当前单台核心层交换机转发能力已经不能满足需求，现需要在保护现有投资的基础上将网络转发能力提高一倍，同时通过设备间的冗余备份提高网络的高可靠性，并要求网络易管理、易维护。

如图 1-14 所示，SwitchA 和 SwitchB 两台交换机组成堆叠系统，接口 XGE1/0/1 和 XGE1/0/2 加入堆叠端口。

图1-14 配置业务口 CSS 组网图



配置思路

采用如下的思路配置：

1. 为了使设备间组成堆叠，配置交换机的堆叠 ID、堆叠优先级和连接方式。
2. 为了能够在堆叠的成员设备间转发数据报文，配置堆叠端口。一个堆叠端口中可以加入多个堆叠物理成员端口，以增加堆叠链路的带宽和可靠性。

3. 为了使配置生效且成功组建堆叠，需要使能交换机的堆叠功能，使用线缆或光纤连接设备间的堆叠端口，并重新启动设备。

操作步骤

步骤 1 配置 SwitchA 和 SwitchB 的堆叠连接方式、堆叠 ID 和堆叠优先级

配置 SwitchA 的堆叠优先级为 200，堆叠连接方式为业务口连接方式。

```
<HUAWEI> system-view
[HUAWEI] sysname SwitchA
[SwitchA] set css priority 200
[SwitchA] set css mode lpu
```

配置 SwitchB 的堆叠 ID 为 2，堆叠优先级为 100，堆叠连接方式为业务口连接方式。

```
<HUAWEI> system-view
[HUAWEI] sysname SwitchB
[SwitchB] set css id 2
[SwitchB] set css priority 100
[SwitchB] set css mode lpu
```

步骤 2 配置堆叠端口

配置 SwitchA 的业务口 XGE1/0/1～XGE1/0/2 为堆叠物理成员端口并加入堆叠端口。

```
[SwitchA] interface css-port 1
[SwitchA-css-port1/1] port interface xgigabitethernet 1/0/1 to xgigabitethernet
1/0/2
```

配置 SwitchB 的业务口 XGE1/0/1～XGE1/0/2 为堆叠物理成员端口并加入堆叠端口。

```
[SwitchB] interface css-port 1
[SwitchB-css-port1/1] port interface xgigabitethernet 1/0/1 to xgigabitethernet
1/0/2
```

步骤 3 使能堆叠功能

使能 SwitchA 的堆叠功能并重新启动 SwitchA。

```
[SwitchA] css enable
Reboot needed to change CSS config, next starting css mode is lpu. Are you sure
this operation and reboot now? [Y/N]y
```

使能 SwitchB 的堆叠功能并重新启动 SwitchB。

```
[SwitchB] css enable
Reboot needed to change CSS config, next starting css mode is lpu. Are you sure
this operation and reboot now? [Y/N]y
```

步骤 4 验证配置结果

查询堆叠系统的状态。

```
<SwitchA> display css status all
Property Item          Property Value
Chassis ID             1
Priority               200
Enable switch          On
CSS master force       Off
CSS status             master
CSS mode               lpu
Property Item          Property Value
Chassis ID             2
Priority               100
Enable switch          On
CSS master force       Off
CSS status             backup
CSS mode               lpu
```

查询堆叠系统的堆叠端口连线信息。

```
<SwitchA> display css channel
              Chassis 1             ||             Chassis 2
================================================================================
=
Num [Css-port]      [Lpu Port]         ||    [Lpu Port]           [Css-port]
 1     1/1    XGigabitEthernet1/1/0/1      XGigabitEthernet2/1/0/1      2/1
 2     1/1    XGigabitEthernet1/1/0/2      XGigabitEthernet2/1/0/2      2/1
```

1.5 故障处理案例

1.5.1 堆叠线缆连接错误导致堆叠系统不能正常建立

故障现象

两台交换机已经使能堆叠功能、堆叠机框 ID 配置正确、堆叠线缆已经连接的情况下，堆叠系统无法建立。

故障分析

1. 在其中一台设备上使用 **display css status** 命令查看设备的堆叠状态，发现设备处于单框堆叠状态。

```
<HUAWEI> display css status
Property Item          Property Value
Frame ID               2
Priority               1
Enable switch          On
CSS master force       Off
```

```
CSS status              single
```

2. 使用命令 **terminal monitor** 和 **terminal trapping** 打开信息中心发送的告警信息功能，发现有大量堆叠线缆连接错误告警。

```
<HUAWEI> terminal monitor
<HUAWEI> terminal trapping
Info: Current terminal monitor is on.
Mar 31 2010 10:53:43 SYS-136 CSSM/4/STACKCONNECTERROR:OID
1.3.6.1.4.1.2011.5.25.183.1.22.11 Connect error, 2/13 CSS port 3 link to 1/14
port 2, this port should link to 1/13 port 2
Mar 31 2010 10:53:43 SYS-136 CSSM/4/STACKCONNECTERROR:OID
1.3.6.1.4.1.2011.5.25.183.1.22.11 Connect error, 2/13 CSS port 1 link to 1/13
port 4, this port should link to 1/14 port 4
Mar 31 2010 10:53:44 SYS-136 CSSM/4/STACKCONNECTERROR:OID
1.3.6.1.4.1.2011.5.25.183.1.22.11 Connect error, 2/13 CSS port 3 link to 1/14
port 2, this port should link to 1/13 port 2
Mar 31 2010 10:53:44 SYS-136 CSSM/4/STACKCONNECTERROR:OID
1.3.6.1.4.1.2011.5.25.183.1.22.11 Connect error, 2/13 CSS port 1 link to 1/13
port 4, this port should link to 1/14 port 4
Mar 31 2010 10:53:45 SYS-136 CSSM/4/STACKCONNECTERROR:OID
1.3.6.1.4.1.2011.5.25.183.1.22.11 Connect error, 2/13 CSS port 3 link to 1/14
port 2, this port should link to 1/13 port 2
Mar 31 2010 10:53:45 SYS-136 CSSM/4/STACKCONNECTERROR:OID
1.3.6.1.4.1.2011.5.25.183.1.22.11 Connect error, 2/13 CSS port 1 link to 1/13
port 4, this port should link to 1/14 port 4
```

从告警信息中可以知道存在连接错误，改动堆叠线缆的连接。

操作步骤

1. 根据告警提示信息，更改堆叠线缆的连接。

   堆叠线缆重新连接后，其中有一个机框重启（进入堆叠合并），重启之后堆叠建立成功，故障排除。

案例总结

使用堆叠功能时，堆叠线缆的连接要按照连接规则进行连接。

堆叠线缆的连接规则如下，相同编号的为一对相连的堆叠口。

图1-15 堆叠线缆连接规则

1.5.2 堆叠通道故障

故障现象

两台交换机已经使能堆叠功能、堆叠机框 ID 配置正确、堆叠线缆已经连接的情况下，堆叠建立后一条堆叠通道两端状态为 Down。

故障分析

1. 执行命令 **terminal monitor** 和 **terminal trapping** 打开信息中心发送的告警信息功能，发现堆叠端口 Down 的告警信息。

```
<HUAWEI> terminal monitor
<HUAWEI> terminal trapping
May 7 2012 21:08:00 Quidway CSSM/4/STACKLINKDOWN:OID
1.3.6.1.4.1.2011.5.25.183.3.3.2.1 1/14 CSS port 2 down.
```

2. 远程登陆设备，执行命令 **display css channel**，查看堆叠链路状态。

```
<HUAWEI> display css channel
            Chassis 1              ||            Chassis 2
=====================================================================
Num [SRUA HG]    [VSTS Port(Status)]    ||    [VSTS Port(Status)]    [SRUA HG]
1   1/13 0/0  -- 1/13/0/1(UP 16G)    ---||--- 2/7/0/4(UP 16G)     -- 2/8  0/14
2   1/13 0/1  -- 1/13/0/3(UP 16G)    ---||--- 2/8/0/2(UP 16G)     -- 2/7  0/15
3   1/13 0/14 -- 1/14/0/4(UP 16G)    ---||--- 2/7/0/1(UP 16G)     -- 2/7  0/0
```

```
4   1/13 0/15 -- 1/14/0/2(DOWN NA)    ---||--- 2/8/0/3(DOWN NA)     -- 2/8  0/1
5   1/14 0/0  -- 1/14/0/1(UP 16G)     ---||--- 2/8/0/4(UP 16G)      -- 2/7  0/14
6   1/14 0/1  -- 1/14/0/3(UP 16G)     ---||--- 2/7/0/2(UP 16G)      -- 2/8  0/15
7   1/14 0/14 -- 1/13/0/4(UP 16G)     ---||--- 2/8/0/1(UP 16G)      -- 2/8  0/0
8   1/14 0/15 -- 1/13/0/2(UP 16G)     ---||--- 2/7/0/3(UP 16G)      -- 2/7  0/1
```

通过告警信息和堆叠链路状态，可知因为 4 号链路故障，检查堆叠线缆及堆叠相关模块是否正常，可排除故障。

操作步骤

1. 检查堆叠电缆或者光纤、光模块的连接是否可靠，将堆叠电缆或者光模块重新插拔，插拔间隔建议大于 5 秒。如果重新连接后，堆叠链路两端状态为 UP，组故障排除，否则，转入第二步。
2. 更改状态为 Down 的堆叠线缆。

案例总结

单个堆叠链路发生故障时，不会影响业务，但是发生故障的堆叠链路数量增多，堆叠分裂的风险增加，所以及时检查堆叠链路状态，排除堆叠链路故障。

1.6 FAQ

堆叠后堆叠系统的配置和堆叠前两台设备的配置有无关系？如果有关系，是否存在合并、冲突、覆盖等操作？

堆叠配置按 Master 设备配置启动。堆叠启动后，Standby 设备的配置丢失，变为空配置。

堆叠后，如何判断哪台交换机是 Master？

可以观察堆叠卡上的灯的信息得到堆叠 ID，根据堆叠 ID 通过命令行 **display css status all** 查到对应的 Master 设备。

堆叠中一台离开时，影不影响现有业务？

当一台设备离开时，离开的设备的业务中断。

加入堆叠时，影不影响现有业务？

有设备加入堆叠时，两台设备进行堆叠合并，被选为 Standby 的设备会进行重启，Standby 设备配置丢失，影响 Standby 设备的业务。

堆叠分裂时，影不影响现有业务？

堆叠分裂时，Standby 设备自动升为 Master 设备，堆叠系统变为两台相互独立的设备，原来的跨框业务中断。

堆叠卡、线缆是否支持热拔插？

堆叠卡不支持热插拔，线缆可以插拔，但是会影响堆叠带宽。

堆叠卡带宽是多少，拔掉一个堆叠线缆是否影响业务，拔的时候有没有丢包？

每条堆叠线缆支持 16G 的带宽，每个堆叠卡有 4 个端口，带宽为 4*16G=64G。

目前只能容忍一条链路故障，会影响堆叠的带宽从 12.5%～25%不等，拔的时候肯定会有丢包。

堆叠功能是否需要单独的 License？

堆叠功能不需要单独的 License。

堆叠后系统只有一个 MAC 地址，选择哪台设备 MAC 作为系统 MAC，如果分裂会不会发生变化？

堆叠建立后，使用 Master 设备 MAC 作为系统的 MAC。分裂后出现两台设备的 MAC 和 IP 冲突，需要手动进行设备间的隔离，如 shutdown 主备互连的端口，或者重启设备。

堆叠后配置有哪些限制，与堆叠前有哪些差别？

堆叠以后，不支持 ISSU、PoE、PTP 和同步以太时钟特性；呈现的接口形式是四维接口，如：GE2/1/0/1，2 表示堆叠 ID。

A 术语与缩略语

术语与缩略语	英文全名	中文解释
CSS	Cluster Switch System	集群交换系统，又称为堆叠
Eth-Trunk		又称链路聚合（Link Aggregation），是将一组物理接口捆绑在一起作为一个逻辑接口来增加带宽的一种方法
HiGig		数据总线通道，是连接交换机线卡之间或者线卡与主控之间的物理通道
主交换机	Master Switch	也称为堆叠主，是经过堆叠竞争后，角色为主的交换机
备交换机	Standby Switch	也称为堆叠备，是经过堆叠竞争后，角色为备的交换机
系统主		堆叠主交换机上的主用主控板，作为堆叠系统的主用主控板
系统备		堆叠备交换机上的主用主控板，作为堆叠系统的备用主控板
候选系统备		主交换机和备交换机的备用主控板