|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  |  | | |  |
|  | | |
|  | | |
| 方圆ok1 | | | |
|  | **HUAWEI S12700 Series Agile Switches** | | 附件1-16K |
|  | |
| **Issue** | **01** |
| **Date** | **2014-03-11** |
|  | |
| HUAWEI TECHNOLOGIES CO., LTD. | |
|  | | |

|  |
| --- |
| Copyright © Huawei Technologies Co., Ltd. 2014. All rights reserved.  No part of this document may be reproduced or transmitted in any form or by any means without prior written consent of Huawei Technologies Co., Ltd.  Trademarks and Permissions  附件3-图 and other Huawei trademarks are trademarks of Huawei Technologies Co., Ltd.  All other trademarks and trade names mentioned in this document are the property of their respective holders.  Notice  The purchased products, services and features are stipulated by the contract made between Huawei and the customer. All or part of the products, services and features described in this document may not be within the purchase scope or the usage scope. Unless otherwise specified in the contract, all statements, information, and recommendations in this document are provided "AS IS" without warranties, guarantees or representations of any kind, either express or implied.  The information in this document is subject to change without notice. Every effort has been made in the preparation of this document to ensure accuracy of the contents, but all statements, information, and recommendations in this document do not constitute a warranty of any kind, express or implied. |

|  |  |
| --- | --- |
| Huawei Technologies Co., Ltd. | |
| Address: | Huawei Industrial Base  Bantian, Longgang  Shenzhen 518129  People's Republic of China |
| Website: | [http://enterprise.huawei.com](http://enterprise.huawei.com/) |

About This Document

Keywords

CSS2, cluster, high reliability, redundancy backup

Abstract

CSS2 is the technology that virtualizes multiple cluster-capable switches into a logical switch. This document describes how CSS2 is implemented and how CSS2 is used on the network.

Acronyms and Abbreviations

|  |  |
| --- | --- |
| Abbreviation | Full Name |
| CSS2 | Cluster Switch System Generation2 |
|  |  |
|  |  |

Contents

[About This Document ii](#_Toc382594764)

[1 Overview 1](#_Toc382594765)

[1.1 Background 1](#_Toc382594766)

[1.2 Advantages 2](#_Toc382594767)

[2 Technology Characteristics 1](#_Toc382594768)

[2.1 Industry's First Switch Fabric Hardware Clustering Technology 1](#_Toc382594769)

[2.2 Easy-to-Use Cluster 1](#_Toc382594770)

[2.3 Innovative 1+1 MPU Backup 1](#_Toc382594771)

[2.4 Low Forwarding Latency 2](#_Toc382594772)

[3 Technology Implementation 3](#_Toc382594773)

[3.1 Concepts 3](#_Toc382594774)

[3.2 Cable Connection 4](#_Toc382594775)

[3.3 Cluster Setup 6](#_Toc382594776)

[3.4 Cluster Merge 6](#_Toc382594777)

[3.5 Cluster Active/Standby Switchover 7](#_Toc382594778)

[3.6 Cluster Split 8](#_Toc382594779)

[3.7 Address Conflict Detection After a Cluster Split 9](#_Toc382594780)

[3.8 LAG Local Preferential Forwarding 10](#_Toc382594781)

[3.9 Cluster Packet Forwarding 11](#_Toc382594782)

[4 Application Scenarios 13](#_Toc382594783)

[4.1 Increasing the Port Density 13](#_Toc382594784)

[4.2 Implementing Inter-Chassis Link Redundancy 13](#_Toc382594785)

[4.3 Implementing Long-Distance Clustering 14](#_Toc382594786)

[4.4 Simplifying Networking 15](#_Toc382594787)

# Overview

## Background

Dual-node redundancy design is often used at the network core layer and aggregation layer to improve network reliability.

The redundancy design improves network reliability but complicates network structure and interconnection. This design requires Layer 2 loop prevention protocols such as the Spanning Tree Protocol (STP), Rapid Spanning Tree protocol (RSTP), Multiple Spanning Tree Protocol (MSTP), or Ethernet Ring Protection Switching (ERPS) to eliminate loops and requires protocols such as the Virtual Router Redundancy Protocol (VRRP) to support node redundancy backup, complicating network protocol deployment.

Huawei Cluster Switch System Generation 2 (CSS2) is developed based on the Cluster Switch System (CSS) technology. CSS2 uses switch fabric clustering, supports 1+N MPU backup, separates the control plane and forwarding plane, features low latency, high bandwidth, and high reliability, and does not occupy service slots.

CSS2 virtualizes multiple cluster-capable switches into one logical switch, simplifying network management. Currently, a cluster contains at most two Huawei physical switches and will support more physical switches in the future.

Figure 1-1 shows the comparison between redundancy backup and CSS2.

Comparison between redundancy backup and CSS2



## Advantages

CSS2 has the following advantages after networking transformation from stand-alone devices into a cluster.

* Simpler configuration and management

After a cluster is set up, multiple physical devices are virtualized into one logical device. You can log in to the cluster to configure and manage all the member devices in a centralized manner.

* Simplified network structure and protocol deployment and loop-free network

CSS2 technology can simplify the complicated network topology into the network structure with clear hierarchy and simple interconnections. In this network structure, loops are eliminated between network devices at each layer through link aggregation, removing the need to deploy protocols such as Spanning Tree Protocols (xSTP, where x specifies the STP type) and VRRP.

* High reliability

Redundancy backup is implemented between multiple member devices in a cluster. Additionally, a cluster supports inter-device link aggregation to implement inter-device link redundancy backup.

* Link load balancing

Inter-device link load balancing ensures 100% network link and bandwidth usage.

# Technology Characteristics

## Industry's First Switch Fabric Hardware Clustering Technology

Based on the Huawei high-end router platform, the S12700 series supports industry's first switch fabric hardware clustering technology (CSS2). CSS2 technology connects cluster member switches through hardware channels of SFUs. Therefore, control packets and data packets of a cluster only need to be forwarded once by the SFUs and do not go through service cards. Compared with traditional service port clustering technologies, CSS2 reduces risks of service interruption caused by service card faults and significantly shortens the transmission latency.

## Easy-to-Use Cluster

Traditional service port clustering technologies require service ports to be configured as CSS ports. CSS2 technology only requires ports on CSS cards of the local and remote chassis to connect each other according to cable connection rules, and does not require CSS port configuration. CSS2 technology uses SFUs to set up a cluster, without occupying service slots or service ports.

## Innovative 1+1 MPU Backup

In a cluster set up using service ports, each chassis must have at least one MPU working normally. When all the MPUs in a chassis are removed or faulty, power supply for the chassis is stopped. CSS2 innovatively implements 1+N MPU backup in a cluster. Once the active MPU of the cluster fails, the system can still work as long as one MPU in the cluster is functioning normally. CSS2 separates the control plane from the data plane, preventing data forwarding from being affected when MPUs are absent. CSS2 allows single-MPU chassis to set up a cluster. Even when the standby chassis has no MPU working normally, data forwarding of the standby chassis is not affected, and the cluster can still run normally.

## Low Forwarding Latency

Traditional service port clustering technologies are implemented through service ports connected by optical fibers or cables. All inter-chassis packets pass through the second-level switch fabrics simulated by service cards. When switch A and switch B set up a cluster, the path that inter-chassis packets pass through is SFU A->service card A->service card B->SFU B. In a cluster set up using CSS2 technology, inter-chassis packets only pass through SFU A and SFU B. Comparing the two paths, you can find that CSS2 eliminates the need to forward inter-chassis packets between two service cards, greatly reducing the forwarding latency.

Comparison between forwarding latencies



# Technology Implementation

## Concepts

Roles in CSS2

* Master switch (cluster master): a switch competes to be the master. A cluster has only one master switch. The master switch manages the cluster, allocates cluster IDs to member switches, collects cluster topology information, and notifies the topology information to all member switches.
* Standby switch (cluster standby): a switch competes to be the standby and provides backup to the master switch. When the master switch fails, the standby switch takes over all the services of the master switch.
* CSS active MPU: active MPU of the master switch.
* CSS standby MPU: active MPU of the standby switch.
* Chassis active: active MPU of a single chassis.
* Chassis standby: standby MPU of a single chassis.
* Cold standby MPU: an MPU that does not participate in control management.
* Switch fabric CSS card: an LPU that is installed on a switch fabric unit (SFU) to set up a cluster.

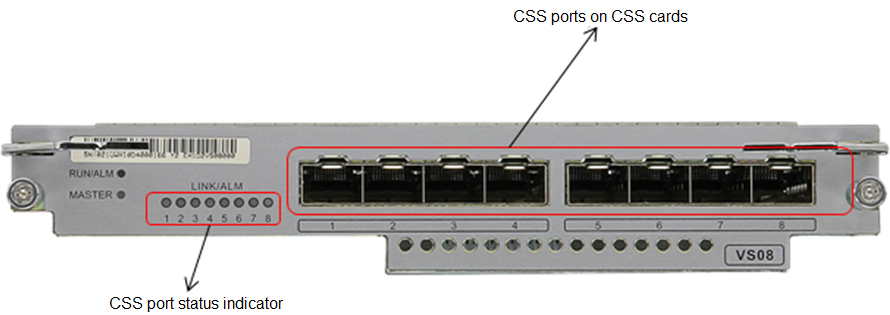
Roles in CSS2



## Cable Connection

Member switches set up a cluster through interfaces of CSS cards installed on SFUs. Currently, each SFU of the S12700 can have one EH1D2VS08000 CSS card installed, and each CSS provides eight 10GE CSS ports. A CSS card on a local SFU must connect to the CSS card on a remote SFU according to cable connection rules.

EH1D2VS08000 CSS card plane



Cable connection for S12700 SFUs



When connecting CSS cards, note the following:

* A cluster can be set up as long as a port of a CSS card in the local chassis connects to a port of the CSS card in the remote chassis.
* On a chassis, two CSS cards are recommended to ensure reliability, and fully meshed connections are recommended to ensure bandwidth.
* Connecting one CSS card on the local chassis to multiple CSS cards on the remote chassis is not allowed. shows an incorrect cable connection.

Incorrect cable connection on S12700



## Cluster Setup

As shown in Figure 3-5, a cluster is set up automatically after two switches are connected through cluster cables and enabled with the CSS2 function. Then one switch competes to be the master switch, and the other becomes the standby switch.

The rules for electing the master switch are as follows:

* The switch that starts and enters the single-chassis CSS mode (with the CSS2 function enabled) first becomes the master switch.
* When two switches complete startup at the same time, the switch with the highest CSS priority becomes the master switch.
* If two switches complete startup at the same time and have the same CSS priority, the switch with the smallest MAC address becomes the master switch.

Before a cluster is set up, each switch is an independent entity and has its own IP address. You need to manage the switches separately. After a cluster is set up, the switches in the cluster form a logical entity, and you can use a single IP address to manage and maintain the switches uniformly. The IP address and MAC address of the cluster is the IP address and MAC address of the master switch elected when the cluster is set up the first time. All the cards in the cluster register with the active MPU of the cluster. After the registration, the configuration data is restored through the configuration file on the active MPU of the cluster.

Cluster setup process



## Cluster Merge

As shown in Figure 3-6, two stable cluster-capable switches (single-chassis cluster) can be merged into one new double-chassis cluster. The superior switch between the master switches of the two clusters is selected as the master switch of the new cluster. On the master switch, the configuration remains unchanged, services are not affected, and the standby MPU restarts. The standby switch restarts, joins the new cluster as the cluster standby (the cluster merge fails if the standby switch has the same CSS ID as the master switch), and synchronizes the configuration with the master switch.

Cluster merge process



## Cluster Active/Standby Switchover

* When an active/standby switchover occurs on the master switch:
* The original standby switch of the cluster becomes the new master switch, and the original CSS standby MPU becomes the new CSS active MPU.
* The original master switch of the cluster becomes the new standby switch.
* The standby MPU of the original master switch becomes the new CSS standby MPU.

Active/standby switchover on the master switch of the cluster



* When an active/standby switchover occurs on the standby switch:
* The master and standby switches of the cluster do not change their roles.
* The standby MPU of the standby switch becomes the new CSS standby MPU.

## Cluster Split

After a cluster is set up, the CSS active and standby MPUs periodically send heartbeat packets to each other to maintain the cluster status. If faults occur on CSS cables, CSS cards, or MPUs, communication between two switches may fail, causing the timeout of heartbeat packets transmitted between the two switches. The cluster then splits into two independent switches, as shown in Figure 3-8.

Cluster split

**SwitchA**

**SwitchB**

**CSS2**

**CSS2**

**+**

**=**

**SwitchA**

**SwitchB**

CSS Link

**CSS2**

If the two switches run properly after the cluster splits, they use the same IP address and same MAC address to communicate with other devices on the network because their global configurations are the same. This causes conflicts of IP addresses and MAC addresses and faults on the entire network. Multi-active detection (MAD) is used to address this problem. MAD is a protocol that detects cluster split and multiple-master situations and takes recovery actions to minimize impact of a cluster split on services.

Cluster split process



## Address Conflict Detection After a Cluster Split

All the switches in a cluster use the same IP address and MAC address. Split of a running cluster can result in the conflict of IP addresses and MAC addresses. After a cluster splits, Layer 2 and Layer 3 conflict detection must be performed.

Multi-active detection (MAD) is a protocol that can detect cluster split and multiple-master situations and take recovery actions to minimize impact of a cluster split on services. When a cluster splits into two clusters because of a link failure, MAD checks whether two switches with the same configurations are running on the network. If two master switches exist after the split, the switches determine whether they are the master or standby switch according to the information in the MAD packets they receive. The standby switch shuts down all its ports except the port used for MAD detection.

Currently, two MAD modes are available:

* Direct mode

When a cluster is running properly, two member switches do not send MAD packets to reduce CPU loads. When the cluster splits, the two member switches send MAD packets over the MAD link at an interval of 1s. If a member switch receives a MAD packet from the other member switch, a dual-master situation has occurred.

MAD in direct mode



* Relay mode

After a cluster splits, two member switches send MAD packets over an Eth-Trunk to determine the master and standby switches. If a switch becomes the master switch, it remains the Active state and forwards service packets normally. If a switch becomes the standby switch, it shuts down all its service ports except reserved ports, remains the Recovery state, and stops forwarding service packets. When the faulty CSS link recovers, the switch in Recovery state restarts and restores the shutdown ports to Up state. The entire cluster then recovers.

MAD in relay mode



## LAG Local Preferential Forwarding

The inter-chassis bandwidth in a cluster is limited. To improve forwarding efficiency and reduce inter-chassis forwarded traffic, you need to configure link aggregation group (LAG) local preferential forwarding. Then traffic received by the local chassis is first forwarded from an interface of the local chassis. If the interface becomes faulty, the traffic is forwarded from an interface of the remote chassis. By default, the S12700 series supports local preferential forwarding.

LAG local preferential forwarding



## Cluster Packet Forwarding

The S12700 series uses the architecture that separates the switching plane from the forwarding plane for cluster packet forwarding. LPUs in a cluster perform Layer 2/Layer 3 packet forwarding. When the outbound interface of traffic is not an interface on the local LPU, the traffic is forwarded by an SFU. This switching process is an internal implementation. For example, the number of hops for Layer 3 packets is increased only by 1 regardless of how many member switches the packets pass through in a cluster. For other network devices, the Layer 3 packets pass through only one network device.

The inter-chassis unicast packet forwarding process is as follows:

An outbound interface is found on the uplink card, and the local-chassis switching process starts.

A local-chassis SFU switches unicast packets to a remote-chassis SFU.

The remote-chassis SFU switches the unicast packets to the destination card.

The destination card forwards the packets.

Unicast packet forwarding



The inter-chassis multicast/broadcast packet forwarding process is as follows:

Multicast/broadcast packets are forwarded on the uplink card based on multicast forwarding entries, and one copy of the multicast/broadcast packets is sent to a local-chassis SFU.

The local-chassis SFU copies the multicast/broadcast packets to the local-chassis card and remote-chassis SFU.

The remote-chassis SFU copies the multicast/broadcast packets to multiple destination cards.

The destination cards copy the multicast/broadcast packets to ports on these cards.

Multicast/broadcast packet forwarding



# Application Scenarios

## Increasing the Port Density

As shown in Figure 4-1, when the port density of a cluster is insufficient for an increasing number of users, you can add new member switches to the cluster to increase ports.

Networking for increasing the port density



Connect the CSS cables according to connection rules, and enable CSS2 on the master and standby switches.

[Switch] set CSS priority 200 //(Optional) Set the CSS priority.

[Switch] CSS enable //Enable CSS. After enabling CSS on the switches, restart the switches. The switches then work in single-chassis CSS mode.

Warning: The CSS configuration takes effect only after the system is rebooted. The next CSS mode is lpu. Reboot now? [Y/N]:y

## Implementing Inter-Chassis Link Redundancy

Inter-chassis LAG technology allows you to configure physical Ethernet ports on different switches as an aggregated port. If a switch is faulty, the aggregated link still works, and another switch manages and maintains the other aggregated ports. This technology increases device capacity and implements service backup between switches to improve reliability.

As shown in Figure 4-2, S12700-1 and S12700-2 form a cluster. Users connect to SwitchA and SwitchB. SwitchA connects to the cluster through inter-chassis Trunk1. The cluster connects to an OSPF network through inter-chassis Trunk2. Physical ports on different devices are added to different VLANs. The devices are connected to the OSPF networks through VLANIF interfaces. Traffic from SwitchA is forwarded to the OSPF networks through VLANIF10 or VLANIF20. If the ECMP algorithm selects the uplink physical port of the local switch S12700-1, traffic is forwarded through S12700-1. If the ECMP algorithm selects the physical port of S12700-2, traffic is forwarded to S12700-2 through the CSS port on an SFU, and then forwarded to the OSPF networks from S12700-2 through the uplink port. When a device or physical port fails, services can be automatically switches to the other device, implementing inter-device link redundancy backup and increasing reliability.

Inter-chassis link redundancy



## Implementing Long-Distance Clustering

Long-distance clustering enables devices far from each other to form a cluster. As shown in Figure 4-3, users on each floor connect to the aggregation switches through respective corridor access switches. The aggregation switches in the two buildings connect users to the external network. Aggregation switches in the two buildings set up a cluster over a long-distance connection and work like one aggregation device. In this way, the network structure is simplified, the network becomes more robust, reliable, and the management and maintenance costs are reduced.

Networking for long-distance clustering using CSS2



## Simplifying Networking

Figure 4-4 shows a common networking, where protocols such as MSTP and VRRP are used to support link redundancy and gateway backup. The following uses the networking between the aggregation layer and access layer as an example.

After CSS2 is enabled, multiple devices at the aggregation layer form a single logical device, and access devices directly connect to the logical device. The simplified network does not require MSTP or VRRP, so the network configuration is much simpler. Inter-device link aggregation also speeds up network convergence and improves network reliability.

CSS2 simplifying networking

