Spanning Tree Technology White Paper

Contents

Overview

Technical background

Spanning tree protocols eliminate loops in a physical link-redundant network by selectively blocking redundant links and putting them in a standby state.

The recent versions of STP include the Rapid Spanning Tree Protocol (RSTP), the Per-VLAN Spanning Tree (PVST), and the Multiple Spanning Tree Protocol (MSTP).

MSTP benefits

Developed based on IEEE 802.1s, MSTP overcomes the limitations of STP, RSTP, and PVST. In addition to supporting rapid network convergence, it allows data flows of different VLANs to be forwarded along separate paths. This provides a better load sharing mechanism for redundant links.

MSTP provides the following features:

- MSTP divides a switched network into multiple regions, each of which contains multiple spanning trees that are independent of one another.
- MSTP supports mapping VLANs to spanning tree instances by means of a VLAN-to-instance mapping table. MSTP can reduce communication overheads and resource usage by mapping multiple VLANs to one instance.
- MSTP prunes a loop network into a loop-free tree, which avoids proliferation and endless cycling of frames in a loop network. In addition, it supports load balancing of VLAN data by providing multiple redundant paths for data forwarding.
- MSTP is compatible with STP and RSTP, and partially compatible with PVST.

STP implementation

Basic concepts in STP

Root bridge

A tree network must have a root bridge. The entire network contains only one root bridge, and all the other bridges in the network are called leaf nodes. The root bridge is not permanent, but can change with changes of the network topology.

Upon initialization of a network, each device generates and periodically sends configuration BPDUs, with itself as the root bridge. After network convergence, only the root bridge generates and periodically sends configuration BPDUs. The other devices only forward the BPDUs.

Root port

On a non-root bridge, the port nearest to the root bridge is the root port. The root port communicates with the root bridge. Each non-root bridge has only one root port. The root bridge has no root port.

Designated bridge and designated port

Classification	Designated bridge	Designated port
For a device	Device directly connected to the local device and responsible for forwarding BPDUs to the local device.	Port through which the designated bridge forwards BPDUs to this device.
For a LAN	Device responsible for forwarding BPDUs to this LAN segment.	Port through which the designated bridge forwards BPDUs to this LAN segment.

Port states

Table 1 lists the port states in STP.

Table 1 STP port states

State	Receives/sends BPDUs	Learns MAC addresses	Forwards user data
Disabled	No	No	No
Listening	Yes	No	No
Learning	Yes	Yes	No
Forwarding	Yes	Yes	Yes
Blocking	Receive	No	No

Path cost

Path cost is a reference value used for link selection in STP. To prune the network into a loop-free tree, STP calculates path costs to select the most robust links and block redundant links that are less robust.

STP protocol frames

STP uses bridge protocol data units (BPDUs), also known as configuration messages, as its protocol frames. This document uses BPDUs to represent all types of spanning tree protocol frames.

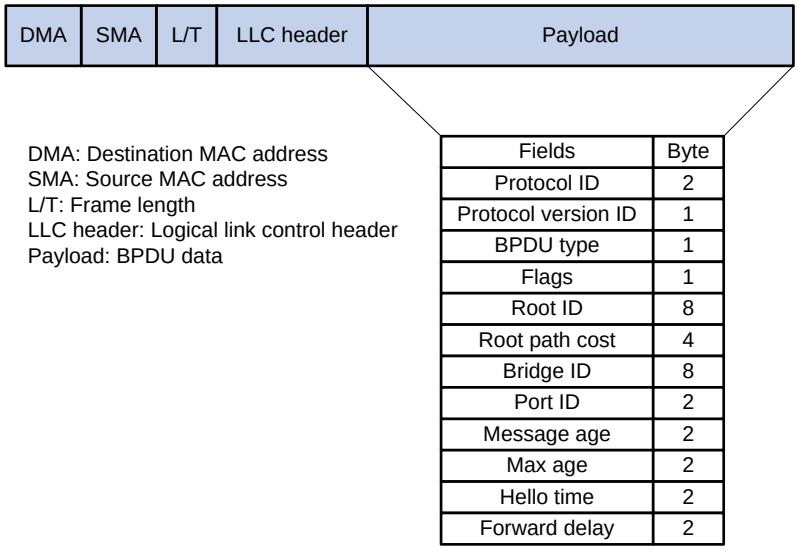
STP-enabled devices exchange BPDUs to establish a spanning tree. BPDUs contain sufficient information for the devices to complete spanning tree calculation.

STP uses two types of BPDUs, configuration BPDUs and topology change notification (TCN) BPDUs.

Configuration BPDUs

Devices exchange configuration BPDUs to elect the root bridge and determine port roles. Figure 1 shows the configuration BPDU format.

Figure 1 Configuration BPDU format



Fields	Byte
Protocol ID	2
Protocol version ID	1
BPDU type	1
Flags	1
Root ID	8
Root path cost	4
Bridge ID	8
Port ID	2
Message age	2
Max age	2
Hello time	2
Forward delay	2

The payload of a configuration BPDU includes the following fields:

- **Protocol ID**—Fixed at 0x0000, which represents IEEE 802.1d.
- **Protocol version ID**—Spanning tree protocol version ID. The protocol version ID for STP is 0x00.
- **BPDU type**—Type of the BPDU. The value is 0x00 for a configuration BPDU.
- **Flags**—An 8-bit field indicates the purpose of the BPDU. The lowest bit is the Topology Change (TC) flag. The highest bit is the Topology Change Acknowledge (TCA) flag. All other bits are reserved.
- **Root ID**—Root bridge ID formed by the priority and MAC address of the root bridge.
- **Root path cost**—Cost of the path to the root bridge.

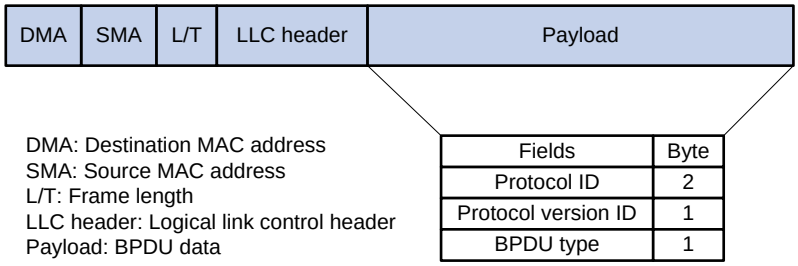
- **Bridge ID**—Designated bridge ID formed by the priority and MAC address of the designated bridge.
- **Port ID**—Designated port ID formed by the priority and global port number of the designated port.
- **Message age**—Age of the configuration BPDU while it propagates in the network.
- **Max age**—Maximum age of the configuration BPDU stored on the switch.
- **Hello time**—Configuration BPDU transmission interval.
- **Forward delay**—Delay for STP bridges to transit port state.

Devices use the root bridge ID, root path cost, designated bridge ID, designated port ID, message age, max age, hello time, and forward delay for spanning tree calculation.

TCN BPDUs

Devices use TCN BPDUs to announce changes in the network topology. Figure 2 shows the TCN BPDU format.

Figure 2 TCN BPDU format



The payload of a TCN BPDU includes the following fields:

- **Protocol ID**—Fixed at 0x0000, which represents IEEE 802.1d.
- **Protocol version ID**—Spanning tree protocol version ID. The protocol version ID for STP is 0x00.
- **BPDU type**—Type of the BPDU. The value is 0x80 for a TCN BPDU.

A non-root bridge sends TCN BPDUs when one of the following events occurs on the bridge:

- A port transits to the forwarding state, and the bridge has a minimum of one designated port.
- A port transits from the forwarding or learning state to the blocking state.

The non-root bridge uses TCN BPDUs to notify the root bridge once the network topology changes. The root bridge then sets the TC flag in its configuration BPDU and propagates it to other bridges.

Calculation process of the STP algorithm

In STP calculation, a device compares the priorities of the received configuration BPDUs from different ports, and elects the root bridge, root ports and designated ports. When the spanning tree calculation is completed, a tree-shape topology forms.

The spanning tree calculation process described in the following sections is an example of a simplified process.

Network initialization

Upon initialization of a device, each port generates a BPDU with the following contents:

- The port as the designated port.
- The device as the root bridge.
- 0 as the root path cost.
- The device ID as the designated bridge ID.

Root bridge selection

The root bridge can be selected in the following methods:

- **Automatic election**—Initially, each STP-enabled device on the network assumes itself to be the root bridge, with its own device ID as the root bridge ID. By exchanging configuration BPDUs, the devices compare their root bridge IDs to elect the device with the smallest root bridge ID as the root bridge.
- **Manual assignment**—You can configure a device as the root bridge or a secondary root bridge of a spanning tree.
  - A spanning tree can have only one root bridge. If you configure multiple devices as the root bridge for a spanning tree, the device with the lowest MAC address is selected.
  - You can configure one or multiple secondary root bridges for a spanning tree. When the root bridge fails or is shut down, a secondary root bridge can take over. If multiple secondary root bridges are configured, the one with the lowest MAC address is selected. However, if a new root bridge is configured, the secondary root bridge is not selected.

Root port and designated ports selection on the non-root bridges

Step	Description
1	A non-root-bridge device regards the port on which it received the optimum configuration BPDU as the root port. Table 2 describes how the optimum configuration BPDU is selected.
2	Based on the configuration BPDU and the path cost of the root port, the device calculates a designated port configuration BPDU for each of the other ports. • The root bridge ID is replaced with that of the configuration BPDU of the root port. • The root path cost is replaced with that of the configuration BPDU of the root port plus the path cost of the root port. • The designated bridge ID is replaced with the ID of this device. • The designated port ID is replaced with the ID of this port.

3	The device compares the calculated configuration BPDU with the configuration BPDU on the port whose port role will be determined. Then, the device acts depending on the result of the comparison:
	• If the calculated configuration BPDU is superior, the device performs the following operations:
	  ○ Considers this port as the designated port.
	  ○ Replaces the configuration BPDU on the port with the calculated configuration BPDU.
	  ○ Periodically sends the calculated configuration BPDU.
	• If the configuration BPDU on the port is superior, the device blocks this port without updating its configuration BPDU. The blocked port can receive BPDUs, but cannot send BPDUs or forward data traffic.

When the network topology is stable, only the root port and designated ports forward user traffic. Other ports are all in the blocking state to receive BPDUs but not to forward BPDUs or user traffic.

Table 2 Selecting the optimum configuration BPDU

Step	Actions
1	Upon receiving a configuration BPDU on a port, the device compares the priority of the received configuration BPDU with that of the configuration BPDU generated by the port. • If the former priority is lower, the device discards the received configuration BPDU and keeps the configuration BPDU the port generated. • If the former priority is higher, the device replaces the content of the configuration BPDU generated by the port with the content of the received configuration BPDU.
2	The device compares the configuration BPDUs of all the ports and chooses the optimum configuration BPDU.

The following are the principles of configuration BPDU comparison:

1. The configuration BPDU with the lowest root bridge ID has the highest priority.
2. If configuration BPDUs have the same root bridge ID, their root path costs are compared. For example, the root path cost in a configuration BPDU plus the path cost of a receiving port is S. The configuration BPDU with the smallest S value has the highest priority.
3. If all configuration BPDUs have the same root bridge ID and S value, the following attributes are compared in sequence:
   a. Designated bridge IDs.
   b. Designated port IDs.
   c. IDs of the receiving ports.

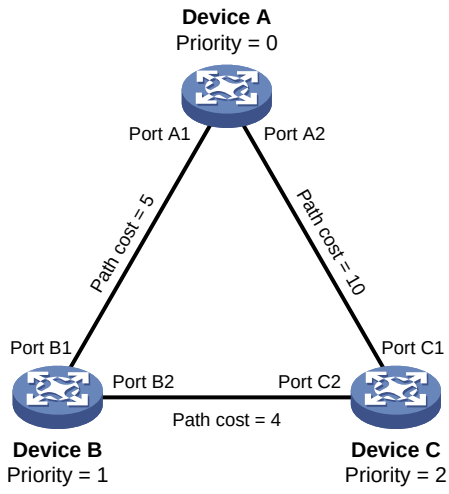
   The configuration BPDU that contains a smaller designated bridge ID, designated port ID, or receiving port ID is selected.

A tree-shape topology forms when the root bridge, root ports, and designated ports are selected.

Example of STP calculation

Figure 3 provides an example showing how the STP algorithm works.

Figure 3 The STP algorithm



As shown in Figure 3, the priority values of Device A, Device B, and Device C are 0, 1, and 2, respectively. The path costs of links among the three devices are 5, 10, and 4.

Device state initialization

In Table 3, each configuration BPDU contains the following fields: root bridge ID, root path cost, designated bridge ID, and designated port ID.

Table 3 Initial state of each device

Device	Port name	Configuration BPDU on the port
Device A	Port A1	{0, 0, 0, Port A1}
	Port A2	{0, 0, 0, Port A2}
Device B	Port B1	{1, 0, 1, Port B1}
	Port B2	{1, 0, 1, Port B2}
Device C	Port C1	{2, 0, 2, Port C1}
	Port C2	{2, 0, 2, Port C2}

Configuration BPDUs comparison on each device

In Table 4, each configuration BPDU contains the following fields: root bridge ID, root path cost, designated bridge ID, and designated port ID.

Table 4 Comparison process and result on each device

Device	Comparison process	Configuration BPDU on ports after comparison
Device A	Port A1 performs the following operations: **1.** Receives the configuration BPDU of Port B1 {1, 0, 1, Port B1}. **2.** Determines that its existing configuration BPDU {0, 0, 0, Port A1} is superior to the received configuration BPDU. **3.** Discards the received one. Port A2 performs the following operations: **1.** Receives the configuration BPDU of Port C1 {2, 0, 2, Port C1}. **2.** Determines that its existing configuration BPDU {0, 0, 0, Port A2} is superior to the received configuration BPDU. **3.** Discards the received one. Device A determines that it is both the root bridge and designated bridge in the configuration BPDUs of all its ports. It considers itself as the root bridge. It does not change the configuration BPDU of any port and starts to periodically send configuration BPDUs.	• Port A1: {0, 0, 0, Port A1} • Port A2: {0, 0, 0, Port A2}
Device B	Port B1 performs the following operations: **1.** Receives the configuration BPDU of Port A1 {0, 0, 0, Port A1}. **2.** Determines that the received configuration BPDU is superior to its existing configuration BPDU {1, 0, 1, Port B1}. **3.** Updates its configuration BPDU. Port B2 performs the following operations: **1.** Receives the configuration BPDU of Port C2 {2, 0, 2,	• Port B1: {0, 0, 0, Port A1} • Port B2: {1, 0, 1, Port B2}

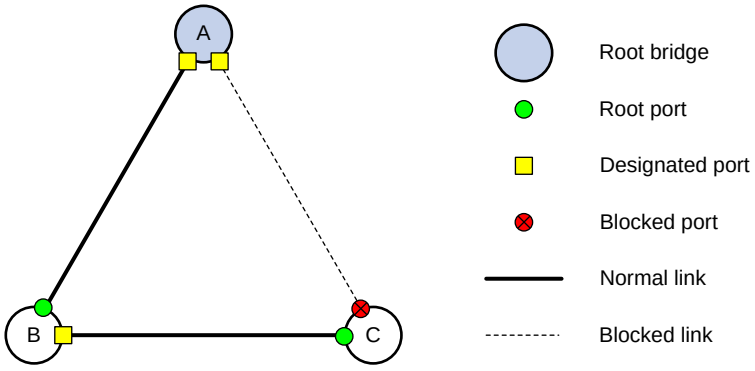
Device	Comparison process	Configuration BPDU on ports after comparison
	Port C2}. **2.** Determines that its existing configuration BPDU {1, 0, 1, Port B2} is superior to the received configuration BPDU. **3.** Discards the received BPDU.	
	Device B performs the following operations: **1.** Compares the configuration BPDUs of all its ports. **2.** Decides that the configuration BPDU of Port B1 is the optimum. **3.** Selects Port B1 as the root port with the configuration BPDU unchanged. Based on the configuration BPDU and path cost of the root port, Device B calculates a designated port configuration BPDU for Port B2 {0, 5, 1, Port B2}. Device B compares it with the existing configuration BPDU of Port B2 {1, 0, 1, Port B2}. Device B determines that the calculated one is superior, and determines that Port B2 is the designated port. It replaces the configuration BPDU on Port B2 with the calculated one, and periodically sends the calculated configuration BPDU.	• Root port (Port B1): {0, 0, 0, Port A1} • Designated port (Port B2): {0, 5, 1, Port B2}
Device C	Port C1 performs the following operations: **1.** Receives the configuration BPDU of Port A2 {0, 0, 0, Port A2}. **2.** Determines that the received configuration BPDU is superior to its existing configuration BPDU {2, 0, 2, Port C1}. **3.** Updates its configuration BPDU. Port C2 performs the following operations: **1.** Receives the original configuration BPDU of Port B2 {1, 0, 1, Port B2}. **2.** Determines that the received configuration BPDU is superior to the existing configuration BPDU {2, 0, 2, Port C2}. **3.** Updates its configuration BPDU.	• Port C1: {0, 0, 0, Port A2} • Port C2: {1, 0, 1, Port B2}
	Device C performs the following operations: **1.** Compares the configuration BPDUs of all its ports. **2.** Decides that the configuration BPDU of Port C1 is the optimum. **3.** Selects Port C1 as the root port with the configuration BPDU unchanged. Based on the configuration BPDU and path cost of the root port, Device C calculates the configuration BPDU of Port C2 {0, 10, 2, Port C2}. Device C compares it with the existing configuration BPDU of Port C2 {1, 0, 1, Port B2}. Device C determines that the calculated configuration BPDU is superior to the existing one, selects Port C2 as the designated port, and replaces the configuration BPDU of Port C2 with the calculated one.	• Root port (Port C1): {0, 0, 0, Port A2} • Designated port (Port C2): {0, 10, 2, Port C2}
	Port C2 performs the following operations: **1.** Receives the updated configuration BPDU of Port B2 {0, 5, 1, Port B2}. **2.** Determines that the received configuration BPDU is superior to its existing configuration BPDU {0, 10, 2, Port C2}. **3.** Updates its configuration BPDU.	• Port C1: {0, 0, 0, Port A2} • Port C2: {0, 5, 1, Port B2}

Device	Comparison process	Configuration BPDU on ports after comparison
	Port C1 performs the following operations: **1.** Receives a periodic configuration BPDU {0, 0, 0, Port A2} from Port A2. **2.** Determines that it is the same as the existing configuration BPDU. **3.** Discards the received BPDU.	
	Device C determines that the root path cost of Port C1 is larger than that of Port C2. The root path cost of Port C1 is 10, root path cost of the received configuration BPDU (0) plus path cost of Port C1 (10). The root path cost of Port C2 is 9, root path cost of the received configuration BPDU (5) plus path cost of Port C2 (4). Device C determines that the configuration BPDU of Port C2 is the optimum, and selects Port C2 as the root port with the configuration BPDU unchanged. Based on the configuration BPDU and path cost of the root port, Device C performs the following operations: **1.** Calculates a designated port configuration BPDU for Port C1 {0, 9, 2, Port C1}. **2.** Compares it with the existing configuration BPDU of Port C1 {0, 0, 0, Port A2}. **3.** Determines that the existing configuration BPDU is superior to the calculated one and blocks Port C1 with the configuration BPDU unchanged. Port C1 does not forward data until a new event triggers a spanning tree calculation process: for example, the link between Device B and Device C is down.	• Blocked port (Port C1): {0, 0, 0, Port A2} • Root port (Port C2): {0, 5, 1, Port B2}

Final calculated spanning tree

After the comparison processes described in Table 4, a spanning tree with Device A as the root bridge is established, as shown in Figure 4.

Figure 4 The final calculated spanning tree



The configuration BPDU forwarding mechanism of STP

The configuration BPDUs of STP are forwarded according to these guidelines:

- Upon network initiation, every device regards itself as the root bridge and generates configuration BPDUs with itself as the root. Then it sends the configuration BPDUs at a regular hello interval.
- If the root port receives a configuration BPDU superior to the configuration BPDU of the port, the device performs the following operations:
  - Increases the message age carried in the configuration BPDU.
  - Starts a timer to time the configuration BPDU.
  - Sends this configuration BPDU through the designated port.
- If a designated port receives a configuration BPDU with a lower priority than its configuration BPDU, the port immediately responds with its configuration BPDU.
- If a path fails, the root port on this path no longer receives new configuration BPDUs and the old configuration BPDUs will be discarded due to timeout. The device generates a configuration BPDU with itself as the root and sends the BPDUs and TCN BPDUs. This triggers a new spanning tree calculation process to establish a new path to restore the network connectivity.

However, the newly calculated configuration BPDU cannot be propagated throughout the network immediately. As a result, the old root ports and designated ports that have not detected the topology change continue forwarding data along the old path. If the new root ports and designated ports begin to forward data as soon as they are elected, a temporary loop might occur.

STP timers

The most important timing parameters in STP calculation are forward delay, hello time, and max age.

- Forward delay

  Forward delay is the delay time for port state transition. By default, the forward delay is 15 seconds.

  A path failure can cause spanning tree re-calculation to adapt the spanning tree structure to the change. However, the resulting new configuration BPDU cannot propagate throughout the network immediately. If the newly elected root ports and designated ports start to forward data immediately, a temporary loop will likely occur.

  The newly elected root ports or designated ports must go through the listening and learning states before they transit to the forwarding state. This requires twice the forward delay time and allows the new configuration BPDU to propagate throughout the network.

- Hello time

  The device sends configuration BPDUs at the hello time interval to the neighboring devices to ensure that the paths are fault-free. By default, the hello time is 2 seconds. If the device does not receive configuration BPDUs within the timeout period, it recalculates the spanning tree. The formula for calculating the timeout period is timeout period = timeout factor × 3 × hello time.

- Max age

  The device uses the max age to determine whether a stored configuration BPDU has expired and discards it if the max age is exceeded. By default, the max age is 20 seconds. In the CIST of an MSTP network, the device uses the max age timer to determine whether a configuration BPDU received by a port has expired. If it is expired, a new spanning tree calculation process starts. The max age timer does not take effect on MSTIs.

If a port does not receive any configuration BPDUs within the timeout period, the port transits to the listening state. The device will recalculate the spanning tree. It takes the port 50 seconds to transit back to the forwarding state. This period includes 20 seconds for the max age, 15 seconds for the listening state, and 15 seconds for the learning state.

To ensure a fast topology convergence, make sure the timer settings meet the following formulas:

- 2 × (forward delay – 1 second) ≥ max age
- Max age ≥ 2 × (hello time + 1 second)

RSTP implementation

Basic concepts in RSTP

Port roles

In addition to root port and designated port, RSTP also uses the following port roles:

- **Alternate port**—Acts as the backup port for a root port. When the root port is blocked, the alternate port takes over.
- **Backup port**—Acts as the backup port of a designated port. When the designated port is invalid, the backup port becomes the new designated port. A loop occurs when two ports of the same spanning tree device are connected, so the device blocks one of the ports. The blocked port is the backup port.
- **Edge port**—Directly connects to a user host rather than a network device or network segment.

Port states

RSTP uses the discarding state to replace the disabled, blocking, and listening states in STP. Table 5 shows the differences between the port states in RSTP and STP.

Table 5 Port state differences between RSTP and STP

STP port state	RSTP port state	Sends BPDU	Learns MAC addresses	Forwards user data
Disabled	Discarding	No	No	No
Blocking	Discarding	No	No	No
Listening	Discarding	Yes	No	No
Learning	Learning	Yes	Yes	No
Forwarding	Forwarding	Yes	Yes	Yes

RSTP protocol frames

An RSTP BPDU uses the same format as an STP BPDU except that a Version1 length field is added to the payload of RSTP BPDUs. The differences between an RSTP BPDU and an STP BPDU are as follows:

- **Protocol version ID**—The value is 0x02 for RSTP.
- **BPDU type**—The value is 0x02 for RSTP BPDUs.
- **Flags**—All 8 bits are used.
- **Version1 length**—The value is 0x00, which means no version 1 protocol information is present.

RSTP does not use TCN BPDUs to advertise topology changes. RSTP floods BPDUs with the TC flag set in the network to advertise topology changes.

How RSTP works

During RSTP calculation, the following events occur:

- If a port in discarding state becomes an alternate port, it retains its state.
- If a port in discarding state is elected as the root port or designated port, it enters the learning state after the forward delay. The port learns MAC addresses, and enters the forwarding state after another forward delay.
  - A newly elected RSTP root port rapidly enters the forwarding state if the following requirements are met:
    - The old root port on the device has stopped forwarding data.
    - The upstream designated port has started forwarding data.
  - A newly elected RSTP designated port rapidly enters the forwarding state if one of the following requirements is met:
    - The designated port is configured as an edge port which directly connects to a user terminal.
    - The designated port connects to a point-to-point link and receives a handshake response from the directly connected device.

RSTP BPDU processing

In RSTP, a non-root bridge actively sends RSTP BPDUs at the hello time through designated ports without waiting for the root bridge to send RSTP BPDUs. This enables RSTP to quickly detect link failures. If a device fails to receive any RSTP BPDUs on a port within triple the hello time, the device considers that a link failure has occurred. After the stored configuration BPDU expires, the device floods RSTP BPDUs with the TC flag set to initiate a new RSTP calculation.

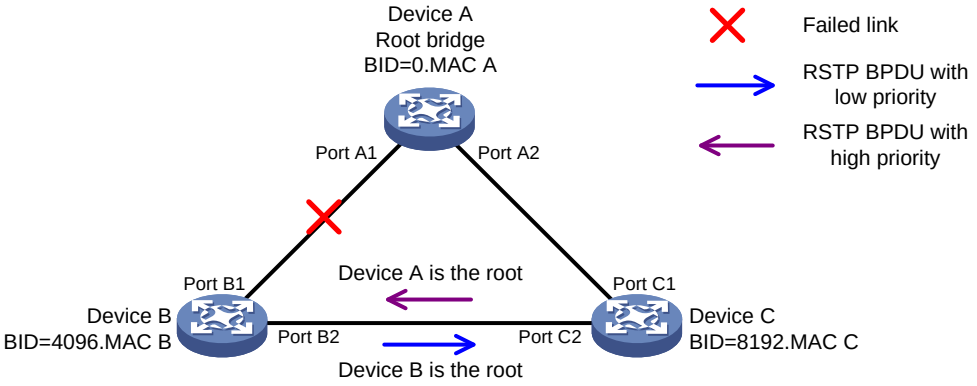
In RSTP, a port in blocking state can immediately respond to an RSTP BPDU with a lower priority than its own BPDU.

As shown in Figure 5, Device A is the root bridge. The priority of Device B is higher than the priority of Device C. Port C2 on Device C is blocked.

When the link between Device A and Device B fails, the following events occur:

1. Device B sends an RSTP BPDU with itself as the root bridge to Device C.
2. Device C compares the RSTP BPDU with its own BPDU.
3. Because the RSTP BPDU from Device B has a lower priority, Device C sends its own BPDU to Device B.
4. Device B considers that Port B2 is the root port and stops sending RSTP BPDUs to Device C.

Figure 5 BPDU processing in RSTP



PVST implementation

In an STP- or RSTP-enabled LAN, all bridges share one spanning tree. Traffic from all VLANs is forwarded along the spanning tree, and ports cannot be blocked on a per-VLAN basis to prune loops.

PVST allows every VLAN to have its own spanning tree, which increases usage of links and bandwidth. Because each VLAN runs RSTP independently, a spanning tree only serves its VLAN.

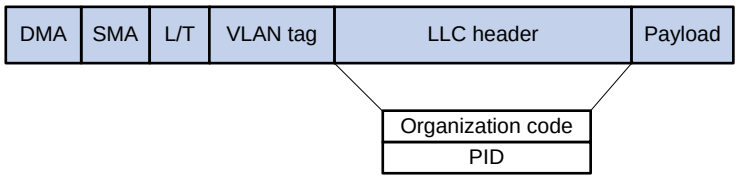
A PVST-enabled H3C device can communicate with a third-party device that is running Rapid PVST or PVST. The PVST-enabled H3C device supports fast network convergence like RSTP when connected to PVST-enabled H3C devices or third-party devices enabled with Rapid PVST.

PVST protocol frames

As shown in Figure 6, a PVST BPDU uses the same format as an RSTP BPDU except the following differences:

- The destination MAC address of a PVST BPDU is 01-00-0c-cc-cc-cd, which is a private MAC address.
- Each PVST BPDU carries a VLAN tag. The VLAN tag identifies the VLAN to which the PVST BPDU belongs.
- The organization code and PID fields are added to the LLC header of the PVST BPDU.

Figure 6 PVST BPDU format



A port's link type determines the type of BPDUs the port sends.

- An access port sends RSTP BPDUs.
- A trunk or hybrid port sends RSTP BPDUs in the default VLAN and sends PVST BPDUs in other VLANs.

How PVST works

PVST implements per-VLAN spanning tree calculation by mapping each VLAN to an MSTI. In PVST, each VLAN runs RSTP independently to maintain its own spanning tree without affecting the spanning trees of other VLANs. In this way, loops in each VLAN are eliminated and traffic of different VLANs is load shared over links. PVST uses RSTP BPDUs in the default VLAN and PVST BPDUs in other VLANs for spanning tree calculation.

PVST uses the same port roles and port states as RSTP for rapid transition.

MSTP implementation

Basic concepts in MSTP

Figure 7 shows a switched network that contains four MST regions, each MST region containing four MSTP devices. Figure 8 shows the networking topology of MST region 3.

Figure 7 Basic concepts in MSTP

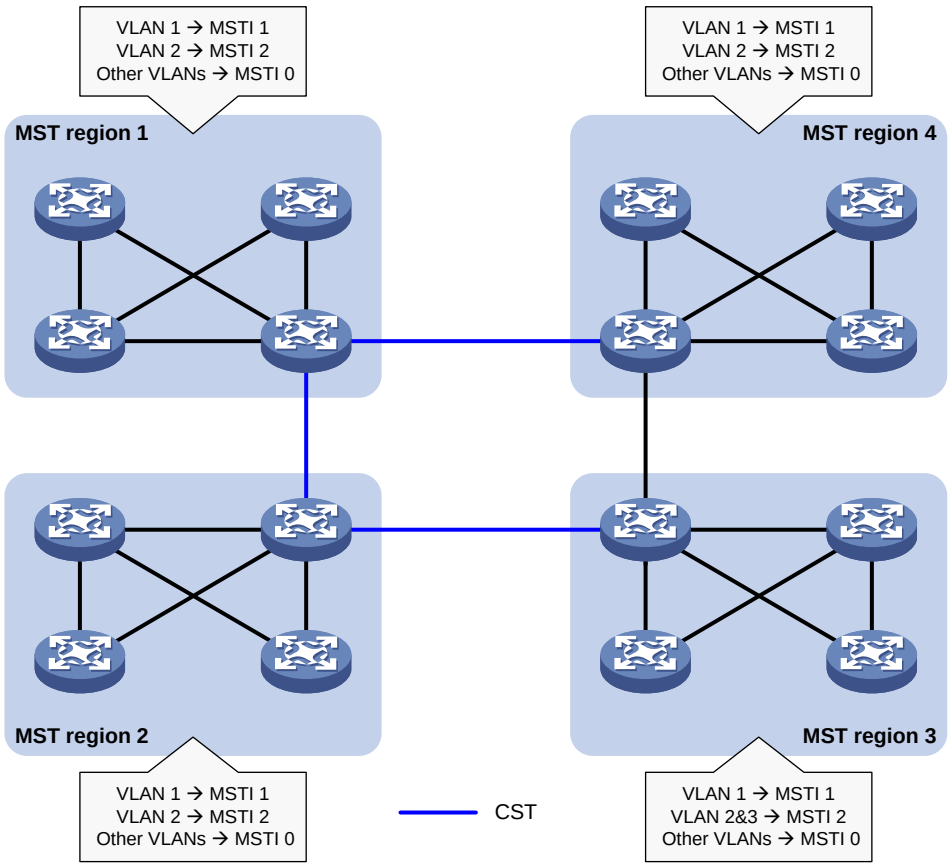
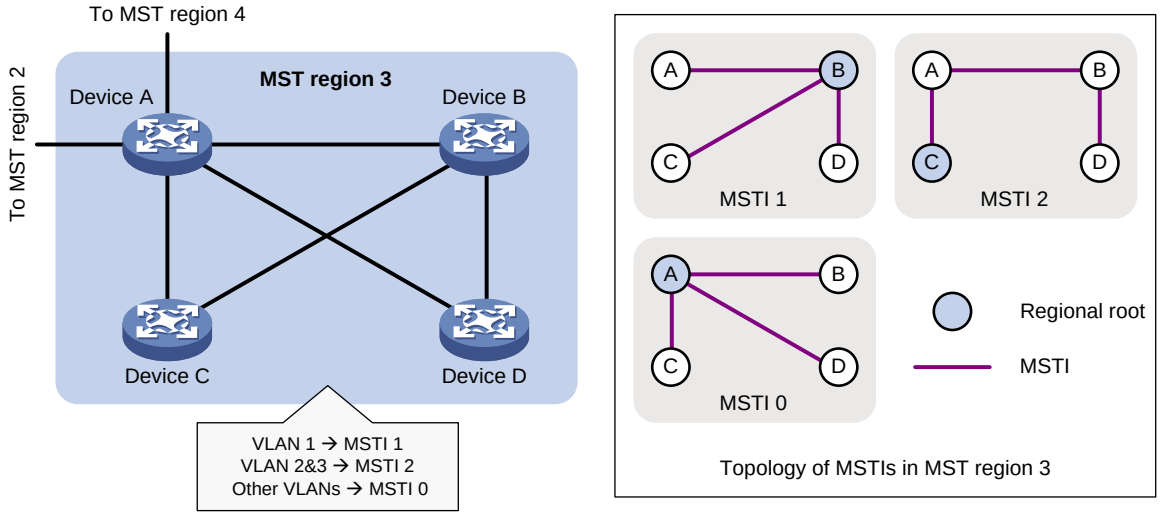


Figure 8 Network diagram and topology of MST region 3



MST region

A multiple spanning tree region (MST region) consists of multiple devices in a switched network and the network segments among them. All these devices have the following characteristics:

- A spanning tree protocol enabled
- Same region name
- Same VLAN-to-instance mapping configuration
- Same MSTP revision level
- Physically linked together

Multiple MST regions can exist in a switched network. You can assign multiple devices to the same MST region, as shown in Figure 7.

- The switched network contains four MST regions, MST region 1 through MST region 4.
- All devices in each MST region have the same MST region configuration.

MSTI

MSTP can generate multiple independent spanning trees in an MST region, and each spanning tree is mapped to the specific VLANs. Each spanning tree is referred to as a multiple spanning tree instance (MSTI).

In Figure 8, MST region 3 contains three MSTIs, MSTI 1, MSTI 2, and MSTI 0.

VLAN-to-instance mapping table

As an attribute of an MST region, the VLAN-to-instance mapping table describes the mapping relationships between VLANs and MSTIs.

In Figure 8, the VLAN-to-instance mapping table of MST region 3 is as follows:

- VLAN 1 to MSTI 1.
- VLAN 2 and VLAN 3 to MSTI 2.
- Other VLANs to MSTI 0.

MSTP achieves load balancing by means of the VLAN-to-instance mapping table.

CST

The common spanning tree (CST) is a single spanning tree that connects all MST regions in a switched network. If you regard each MST region as a device, the CST is a spanning tree calculated by these devices through STP or RSTP.

The blue lines in Figure 7 represent the CST.

IST

An internal spanning tree (IST) is a spanning tree that runs in an MST region. It is also called MSTI 0, a special MSTI to which all VLANs are mapped by default.

In Figure 7, MSTI 0 is the IST in MST region 3.

CIST

The common and internal spanning tree (CIST) is a single spanning tree that connects all devices in a switched network. It consists of the ISTs in all MST regions and the CST.

In Figure 7, the ISTs (MSTI 0) in all MST regions plus the inter-region CST constitute the CIST of the entire network.

Regional root

The root bridge of the IST or an MSTI within an MST region is the regional root of the IST or MSTI. Based on the topology, different spanning trees in an MST region might have different regional roots, as shown in MST region 3 in Figure 8.

- The regional root of MSTI 1 is Device B.
- The regional root of MSTI 2 is Device C.
- The regional root of MSTI 0 (also known as the IST) is Device A.

Common root bridge

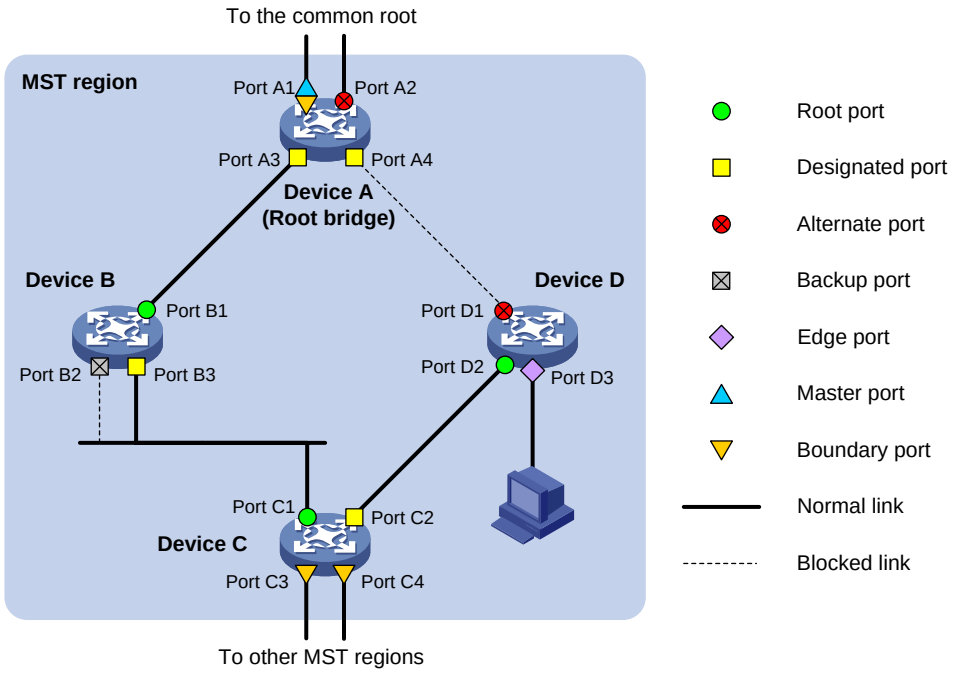
The common root bridge is the root bridge of the CIST.

In Figure 7, the common root bridge is a device in MST region 1.

Port roles

A port can play different roles in different MSTIs. As shown in Figure 9, an MST region contains Device A, Device B, Device C, and Device D. Port A1 and port A2 of Device A connect to the common root bridge. Port B2 and Port B3 of Device B form a loop. Port C3 and Port C4 of Device C connect to other MST regions. Port D3 of Device D directly connects to a host.

Figure 9 Port roles



MSTP calculation involves the following port roles:

- **Root port**—Forwards data for a non-root bridge to the root bridge. The root bridge does not have any root port.
- **Designated port**—Forwards data to the downstream network segment or device.
- **Alternate port**—Acts as the backup port for a root port or master port. When the root port or master port is blocked, the alternate port takes over.
- **Backup port**—Acts as the backup port of a designated port. When the designated port is invalid, the backup port becomes the new designated port. A loop occurs when two ports of the same spanning tree device are connected, so the device blocks one of the ports. The blocked port acts as the backup.
- **Edge port**—Directly connects to a user host rather than a network device or network segment.
- **Master port**—Acts as a port on the shortest path from the local MST region to the common root bridge. The master port is not always located on the regional root. It is a root port on the IST or CIST and still a master port on the other MSTIs.
- **Boundary port**—Connects an MST region to another MST region or to an STP/RSTP-running device. In MSTP calculation, a boundary port's role on an MSTI is consistent with its role on the CIST. However, that is not true with master ports. A master port on MSTIs is a root port on the CIST.

Port states

In MSTP, a port can be in one of the following states:

- **Forwarding**—The port receives and sends BPDUs, learns MAC addresses, and forwards user traffic.
- **Learning**—The port receives and sends BPDUs, learns MAC addresses, but does not forward user traffic. Learning is an intermediate port state.
- **Discarding**—The port receives and sends BPDUs, but does not learn MAC addresses or forward user traffic.

NOTE:

When in different MSTIs, a port can be in different states.

A port state is not exclusively associated with a port role. Table 6 lists the port states that each port role supports. (A check mark [√] indicates that the port supports this state, while a dash [—] indicates that the port does not support this state.)

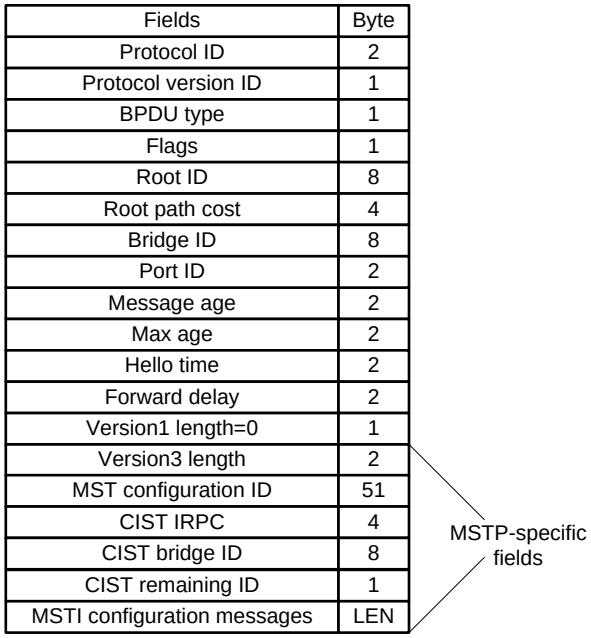
Table 6 Port states that different port roles support

Port role (right) Port state (below)	Root port/master port	Designated port	Alternate port	Backup port
Forwarding	√	√	—	—
Learning	√	√	—	—
Discarding	√	√	√	√

MSTP protocol frames

Figure 10 shows the format of an MSTP BPDU.

Figure 10 MSTP BPDU format



The first 13 fields of an MSTP BPDU are the same as an RSTP BPDU. The other six fields are unique to MSTP.

- **Protocol version ID**—The value is 0x03 for MSTP.
- **BPDU type**—The value is 0x02 for RSTP/MSTP BPDUs.
- **Root ID**—ID of the common root bridge.
- **Root path cost**—CIST external path cost.
- **Bridge ID**—ID of the regional root for the IST or an MSTI.
- **Port ID**—ID of the designated port in the CIST.

- **Version3 length**—Length of the MSTP-specific fields. Devices use this field for verification upon receiving an MSTP BPDU.
- **MST configuration ID**—Includes the format selector, configuration name, revision level, and configuration digest. The value for format selector is fixed at 0x00. The other parameters are used to identify the MST region for the originating bridge.
- **CIST IRPC**—Internal root path cost (IRPC) from the originating bridge to the root of the MST region.
- **CIST bridge ID**—ID of the bridge that sends the MSTP BPDU.
- **CIST remaining ID**—Remaining hop count. This field limits the scale of the MST region. The regional root sends a BPDU with the remaining hop count set to the maximum value. Each device that receives the BPDU decrements the hop count by one. When the hop count reaches zero, the BPDU is discarded. Devices beyond the maximum hops of the MST region cannot participate in spanning tree calculation. The default remaining hop count is 20.
- **MSTI configuration messages**—Contains MSTI configuration messages. Each MSTI configuration message is 16 bytes. This field can contain 0 to 64 MSTI configuration messages. The number of the MSTI configuration messages is determined by the number of MSTIs in the MST region.

How MSTP works

MSTP divides an entire Layer 2 network into multiple MST regions, which are connected by a calculated CST. Inside an MST region, multiple spanning trees, called MSTIs, are calculated. Among these MSTIs, MSTI 0 is the IST.

Like STP, MSTP uses configuration BPDUs to calculate spanning trees. An important difference is that an MSTP BPDU carries the MSTP configuration of the bridge from which the BPDU is sent.

CIST calculation

During the CIST calculation, the following process takes place:

- The device with the highest priority is elected as the root bridge of the CIST.
- MSTP generates an IST within each MST region through calculation.
- MSTP regards each MST region as a single device and generates a CST among these MST regions through calculation.

The CST and ISTs constitute the CIST of the entire network.

MSTI calculation

Within an MST region, MSTP generates different MSTIs for different VLANs based on the VLAN-to-instance mappings. For each spanning tree, MSTP performs a separate calculation process similar to spanning tree calculation in STP. For more information, see "Calculation process of the STP algorithm."

In MSTP, a VLAN frame is forwarded along the following paths:

- Within an MST region, the frame is forwarded along the corresponding MSTI.
- Between two MST regions, the frame is forwarded along the CST.

Rapid transition mechanism

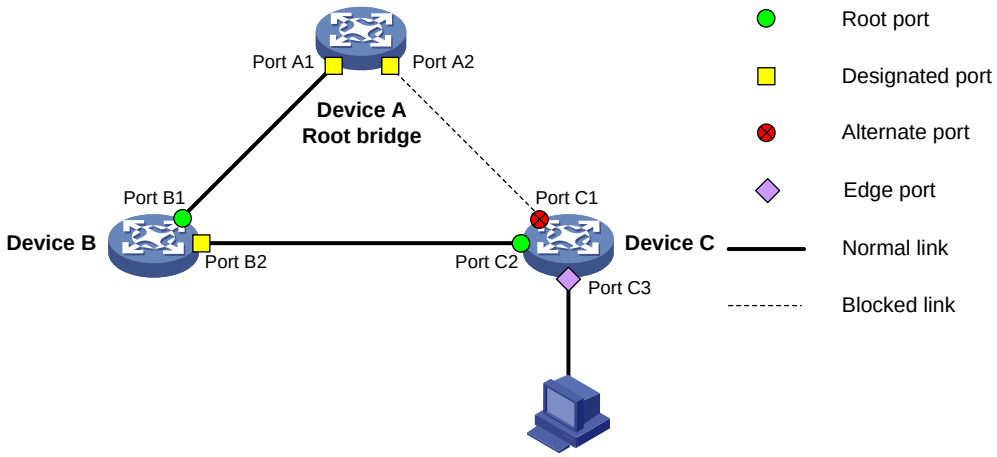
In STP, a port must wait twice the forward delay (30 seconds by default) before it transits from the blocking state to the forwarding state. The forward delay is related to the hello time and network diameter. If the forward delay is too short, loops might occur. This affects the stability of the network.

RSTP, PVST, and MSTP all use the rapid transition mechanism to speed up port state transition for edge ports, root ports, and designated ports. The rapid transition mechanism for designated ports is also known as the proposal/agreement (P/A) transition.

Edge port rapid transition

As shown in Figure 11, Port C3 is an edge port connected to a host. When a network topology change occurs, the port can immediately transit from the blocking state to the forwarding state because no loop will be caused.

Figure 11 Edge port rapid transition

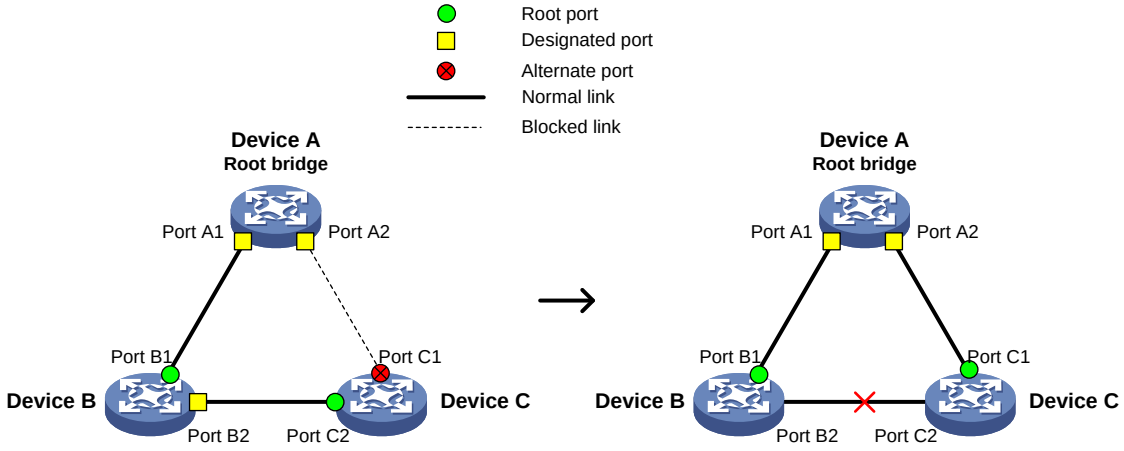


Root port rapid transition

When a root port is blocked, the bridge will elect the alternate port with the highest priority as the new root port. If the new root port's peer is in the forwarding state, the new root port immediately transits to the forwarding state.

As shown in Figure 12, Port C2 on Device C is a root port and Port C1 is an alternate port. When Port C2 transits to the blocking state, Port C1 is elected as the root port and immediately transits to the forwarding state.

Figure 12 Root port rapid transition



P/A transition

The P/A transition enables a designated port to rapidly transit to the forwarding state after a handshake with its peer. The P/A transition applies only to point-to-point links.

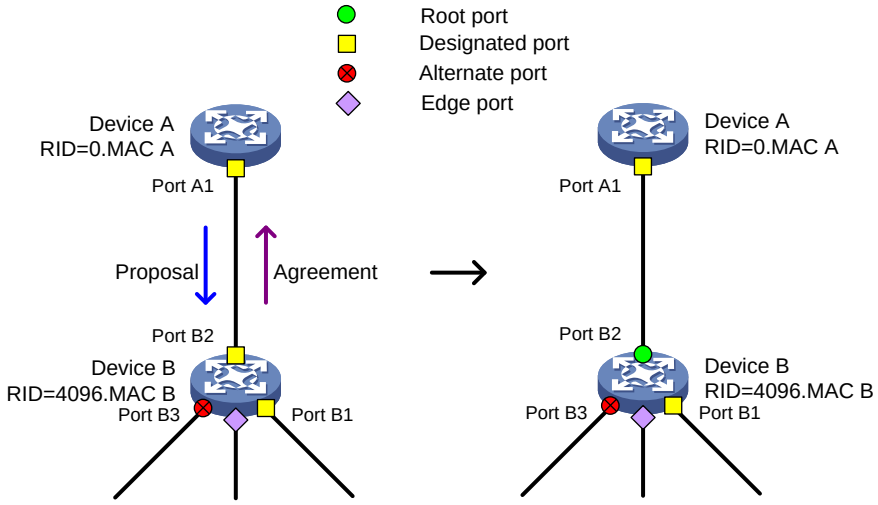
P/A transition for RSTP and PVST

In RSTP or PVST, the ports on a new link or recovered link are designated ports in blocking state. When one of the designated ports transits to the discarding or learning state, it sets the proposal flag in its BPDU. Its peer bridge receives the BPDU and determines whether the receiving port is the root port. If it is the root port, the bridge blocks the other ports except edge ports. The bridge then replies an agreement BPDU to the designated port. The designated port immediately transits to the forwarding state upon receiving the agreement BPDU. If the designated port does not receive the agreement BPDU, it waits for twice the forward delay to transit to the forwarding state.

As shown in Figure 13, the P/A transition operates as follows:

1. Device A sends a proposal BPDU to Device B through Port A1.
2. Device B receives the proposal BPDU on Port B2. Port B2 is elected as the root port.
3. Device B blocks its designated port Port B1 and alternate port Port B3 to eliminate loops.
4. The root port Port B2 transits to the forwarding state and sends an agreement BPDU to Device A.
5. The designated port Port A1 on Device A immediately transits to the forwarding state after receiving the agreement BPDU.

Figure 13 P/A transition for RSTP and PVST



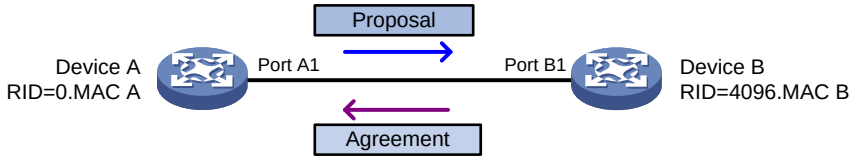
P/A transition for MSTP

In MSTP, an upstream bridge sets both the proposal and agreement flags in its BPDU. If a downstream bridge receives the BPDU and its receiving port is elected as the root port, the bridge blocks all the other ports except edge ports. The downstream bridge then replies an agreement BPDU to the upstream bridge. The upstream port immediately transits to the forwarding state upon receiving the agreement BPDU. If the upstream port does not receive the agreement BPDU, it waits for twice the forward delay to transit to the forwarding state.

As shown in Figure 14, the P/A transition operates as follows:

1. Device A sets the proposal and agreement flags in its BPDU and sends it to Device B through Port A1.
2. Device B receives the BPDU. Port B1 of Device B is elected as the root port.
3. Device B then blocks all its ports except the edge ports.
4. The root port Port B1 of Device B transits to the forwarding state and sends an agreement BPDU to Device A.
5. Port A1 of Device A immediately transits to the forwarding state upon receiving the agreement BPDU.

Figure 14 P/A transition for MSTP



Comware implementation of spanning tree

No Agreement Check

In RSTP and MSTP, the following types of messages are used for rapid state transition on designated ports:

- **Proposal**—Sent by designated ports to request rapid transition
- **Agreement**—Used to acknowledge rapid transition requests

Both RSTP and MSTP devices can perform rapid transition on a designated port only when the port receives an agreement packet from the downstream device. RSTP and MSTP devices have the following differences:

- For MSTP, the root port of the downstream device sends an agreement packet only after it receives an agreement packet from the upstream device.
- For RSTP, the downstream device sends an agreement packet whether or not an agreement packet from the upstream device is received.

Figure 15 Rapid state transition of an MSTP designated port

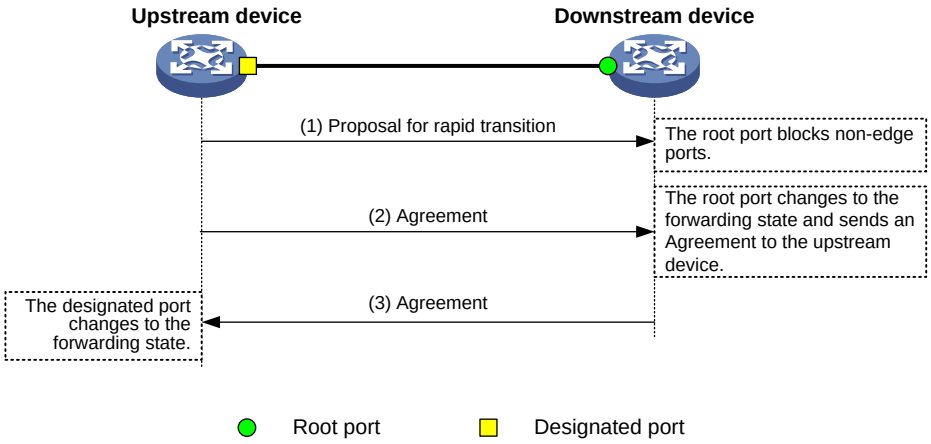
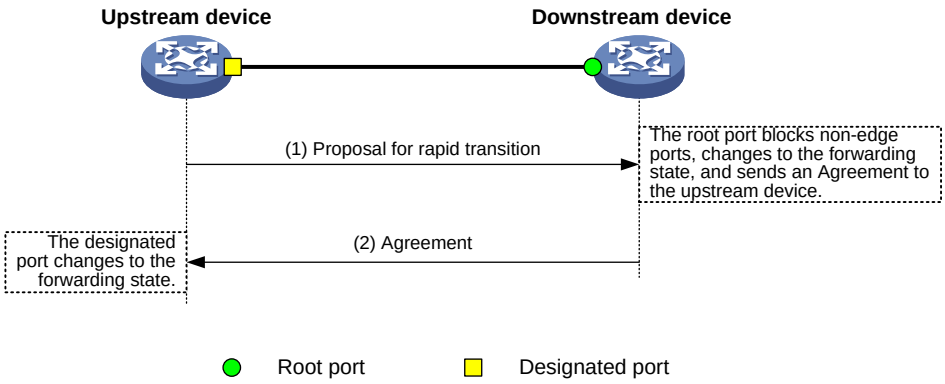


Figure 16 Rapid state transition of an RSTP designated port



If the upstream device is a third-party device, the rapid state transition implementation might be limited as follows:

- The upstream device uses a rapid transition mechanism similar to that of RSTP.
- The downstream device runs MSTP and does not operate in RSTP mode.

In this case, the following occurs:

1. The root port on the downstream device receives no agreement from the upstream device.
2. It sends no agreement to the upstream device.

As a result, the designated port of the upstream device can transit to the forwarding state only after a period twice the forward delay.

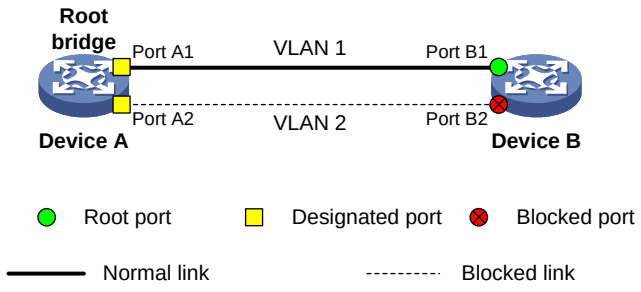
To enable the designated port of the upstream device to transit its state rapidly, enable No Agreement Check on the downstream device's port.

Ignoring VLANs

Traffic of a VLAN on a complex network might be blocked by the spanning tree, as shown in Figure 17.

- Port A1 on Device A allows the traffic of VLAN 1 to pass through.
- Port A2 on Device A allows the traffic of VLAN 2 to pass through.
- Port B1 on Device B allows the traffic of VLAN 1 to pass through.
- Port B2 on Device B allows the traffic of VLAN 2 to pass through.
- Device A and Device B run a spanning tree protocol. Device A is the root bridge, and Port A1 and Port A2 are designated ports. On Device B, Port B1 is the root port, and Port B2 is the blocked port. Traffic of VLAN 2 is blocked.

Figure 17 VLAN connectivity blocked by MSTP



Ignoring a VLAN can make ports of the VLAN forward frames correctly rather than comply with the spanning tree calculation result.

Digest Snooping

As defined in IEEE 802.1s, connected devices are in the same region only when they have the same MST region-related configurations, including:

- Region name.
- Revision level.
- VLAN-to-instance mappings.

A spanning tree device identifies devices in the same MST region by determining the configuration ID in BPDUs. The configuration ID includes the region name, revision level, and configuration digest. It is 16-byte long and is the result calculated through the HMAC-MD5 algorithm based on VLAN-to-instance mappings.

Because spanning tree implementations vary by vendor, the configuration digests calculated through private keys are different. The devices of different vendors in the same MST region cannot communicate with each other.

To enable communication between an H3C device and a third-party device in the same MST region, enable Digest Snooping on the H3C device port connecting them.

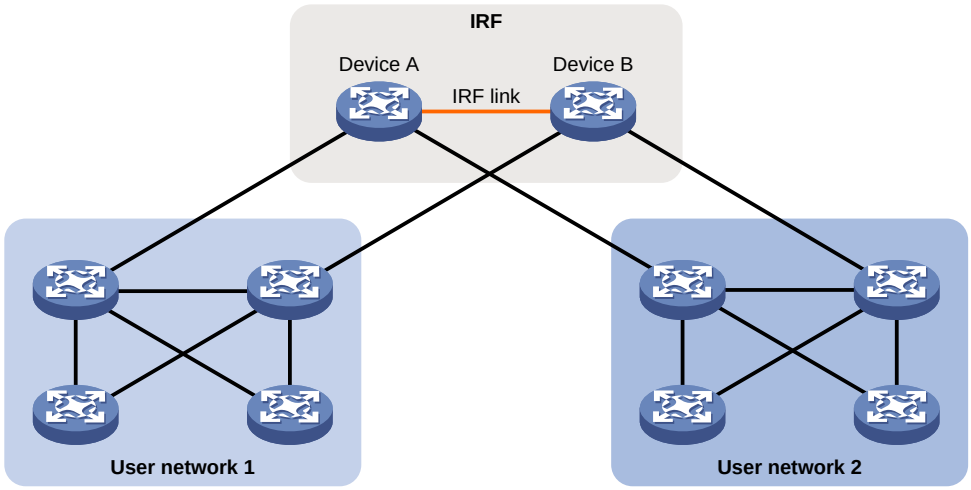
TC Snooping

As shown in Figure 18, an IRF fabric connects to two user networks through double links.

- Device A and Device B form the IRF fabric.
- The spanning tree feature is disabled on Device A and Device B and enabled on all devices in user network 1 and user network 2.
- The IRF fabric transparently transmits BPDUs for both user networks and is not involved in the calculation of spanning trees.

When the network topology changes, it takes time for the IRF fabric to update its MAC address table and ARP table. During this period, traffic in the network might be interrupted.

Figure 18 TC Snooping application scenario



To avoid traffic interruption, you can enable TC Snooping on the IRF fabric. After receiving a TC-BPDU through a port, the IRF fabric updates MAC address table and ARP table entries associated with the port's VLAN. In this way, TC Snooping prevents topology change from interrupting traffic forwarding in the network.

mCheck

The mCheck feature enables user intervention in the port state transition process.

When a port on an MSTP, RSTP, or PVST device connects to an STP device and receives STP BPDUs, the port automatically transits to the STP mode. However, the port cannot automatically transit back to the original mode when the following conditions exist:

- The peer STP device is shut down or removed.
- The port cannot detect the change.

To forcibly transit the port to operate in the original mode, you can perform an mCheck operation.

For example, Device A, Device B, and Device C are connected in sequence. Device A runs STP, Device B does not run any spanning tree protocol, and Device C runs RSTP, PVST, or MSTP. In this case, when Device C receives an STP BPDU transparently transmitted by Device B, the receiving port transits to the STP mode. If you configure Device B to run RSTP, PVST, or MSTP with Device C, you must perform mCheck operations on the ports interconnecting Device B and Device C.

Inconsistent PVID protection

In PVST, if two connected ports use different PVIDs, PVST calculation errors might occur. By default, inconsistent PVID protection is enabled to avoid PVST calculation errors. If PVID inconsistency is detected on a port, the system blocks the port.

If different PVIDs are required on two connected ports, disable inconsistent PVID protection on the devices that host the ports.

Spanning tree protection features

BPDU guard

For access layer devices, the access ports can directly connect to the user terminals (such as PCs) or file servers. The access ports are configured as edge ports to allow rapid transition. When these ports receive configuration BPDUs, the system automatically sets the ports as non-edge ports and starts a new spanning tree calculation process. This causes a change of network topology. Under normal conditions, these ports should not receive configuration BPDUs. However, if someone uses configuration BPDUs maliciously to attack the devices, the network will become unstable.

The spanning tree protocol provides the BPDU guard feature to protect the system against such attacks. When edge ports receive configuration BPDUs on a device with BPDU guard enabled, the device performs the following operations:

- Shuts down these ports.
- Notifies the NMS that these ports have been shut down by the spanning tree protocol.

The device reactivates the ports that have been shut down when the port status detection timer expires.

Root guard

Root guard applies to designated ports.

The root bridge and secondary root bridge of a spanning tree should be located in the same MST region. Especially for the CIST, the root bridge and secondary root bridge are put in a high-bandwidth core region during network design. However, due to possible configuration errors or malicious attacks in the network, the legal root bridge might receive a configuration BPDU with a higher priority. Another device supersedes the current legal root bridge, causing an undesired change of the network topology. The traffic that should go over high-speed links is switched to low-speed links, resulting in network congestion.

To prevent this situation, MSTP provides the root guard feature. If root guard is enabled on a port of a root bridge, this port plays the role of designated port on all MSTIs. After this port receives a configuration BPDU with a higher priority from an MSTI, it performs the following operations:

- Immediately sets that port to the listening state in the MSTI.
- Does not forward the received configuration BPDU.

This is equivalent to disconnecting the link connected to this port in the MSTI. If the port receives no BPDUs with a higher priority within twice the forwarding delay, it reverts to its original state.

Loop guard

Loop guard applies to the root port and alternate ports of a device.

By continuing to receive BPDUs from the upstream device, a device can maintain the state of the root port and blocked ports. However, link congestion or unidirectional link failures might cause these ports to fail to receive BPDUs from the upstream devices. In this situation, the device reselects the following port roles:

- Those ports in forwarding state that failed to receive upstream BPDUs become designated ports.
- The blocked ports transit to the forwarding state.

As a result, loops occur in the switched network. The loop guard feature can suppress the occurrence of such loops.

The initial state of a loop guard-enabled port is **discarding** in every MSTI. When the port receives BPDUs, it transits its state. Otherwise, it stays in the discarding state to prevent temporary loops.

Port role restriction

Port role restriction is used on a port that connects to the user access network.

The bridge ID change of a device in the user access network might cause a change to the spanning tree topology in the core network. To avoid this problem, you can enable port role restriction on a port. With this feature enabled, when the port receives a superior BPDU, it becomes an alternate port rather than a root port.

TC-BPDU transmission restriction

Port role restriction is used on a port that connects to the user access network.

The topology change to the user access network might cause the forwarding address changes to the core network. When the user access network topology is unstable, the user access network might affect the core network. To avoid this problem, you can enable TC-BPDU transmission restriction on a port. With this feature enabled, when the port receives a TC-BPDU, it does not forward the TC-BPDU to other ports.

TC-BPDU guard

When a device receives topology change (TC) BPDUs (the BPDUs that notify devices of topology changes), it flushes its forwarding address entries. If someone uses TC-BPDUs to attack the device, the device will receive a large number of TC-BPDUs within a short time. Then, the device is busy with forwarding address entry flushing. This affects network stability.

TC-BPDU guard allows you to set the maximum number of immediate forwarding address entry flushes performed within 10 seconds after the device receives the first TC-BPDU. For TC-BPDUs received in excess of the limit, the device performs a forwarding address entry flush when the time period expires. This prevents frequent flushing of forwarding address entries.

PVST BPDU guard

This feature takes effect only when the device is operating in MSTP mode.

An MSTP-enabled device forwards PVST BPDUs as data traffic because it cannot recognize PVST BPDUs. If a PVST-enabled device in another independent network receives the PVST BPDUs, a PVST calculation error might occur. To avoid PVST calculation errors, enable PVST BPDU guard on the MSTP-enabled device. The device shuts down a port if the port receives PVST BPDUs.

Dispute guard

Dispute guard can be triggered by unidirectional link failures. If an upstream port receives inferior BPDUs from a downstream designated port in forwarding or learning state because of a unidirectional link failure, a loop appears. Dispute guard blocks the upstream designated port to prevent the loop.

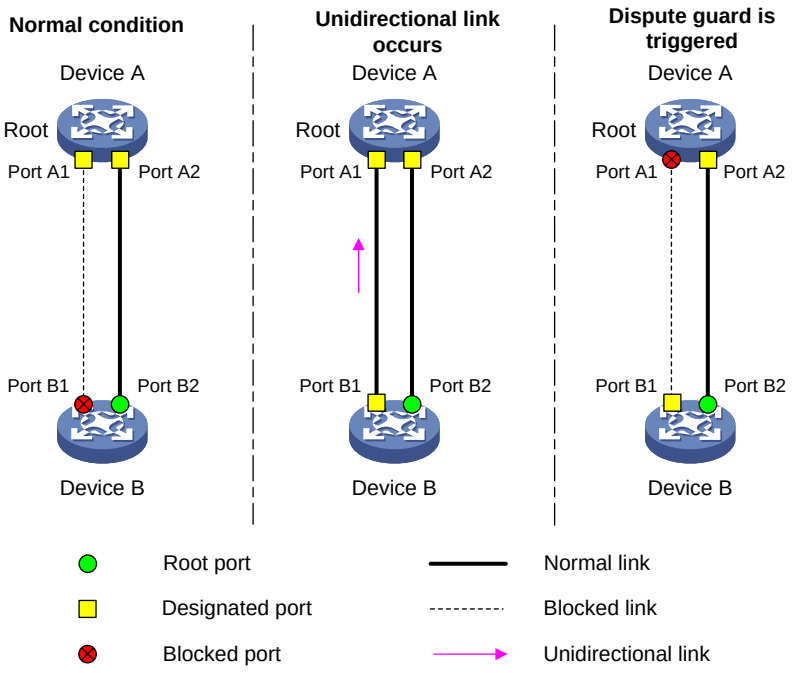
As shown in Figure 19, in normal conditions, the spanning tree calculation result is as follows:

- Device A is the root bridge, and Port A1 is a designated port.
- Port B1 is blocked.

When the link between Port A1 and Port B1 fails in the direction of Port A1 to Port B1 and becomes unidirectional, the following events occur:

1. Port A1 can only receive BPDUs and cannot send BPDUs to Port B1.
2. Port B1 does not receive BPDUs from Port A1 for a certain period of time.
3. Device B determines itself as the root bridge.
4. Port B1 sends its BPDUs to Port A1.
5. Port A1 determines the received BPDUs are inferior to its own BPDUs. A dispute is detected.
6. Dispute guard is triggered and blocks Port A1 to prevent a loop.

Figure 19 Dispute guard triggering scenario

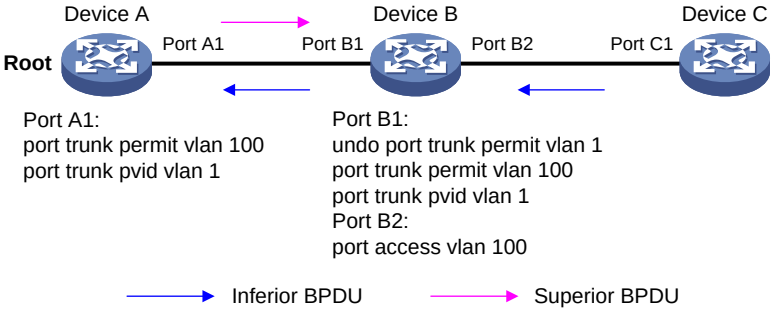


However, dispute guard might disrupt the network connectivity. You can disable dispute guard to avoid connectivity loss in VLAN networks. As shown in Figure 20, the spanning tree feature is disabled on Device B and enabled on Device A and device C. Device B transparently transmits BPDUs.

Device C cannot receive superior BPDUs of VLAN 1 from Device A because Port B1 of Device B is configured to deny packets of VLAN 1. Device C determines itself as the root bridge after a certain period of time. Then, Port C1 sends an inferior BPDU of VLAN 100 to Device A.

When Device A receives the inferior BPDU, dispute guard blocks Port A1, which causes traffic interruption. To ensure service continuity, you can disable dispute guard on Device A to prevent the link from being blocked.

Figure 20 Disabling dispute guard application scenario



Application scenarios

Typically MSTP is used in an environment where multiple VLANs share links to distribute traffic of the VLANs to different links for high availability and load sharing.

As shown in Figure 21, all devices on the network are in the same MST region. MSTP is configured to meet the following requirements:

- VLAN 10's frames are forwarded along MSTI 1 rooted at Switch A.
- VLAN 20's frames are forwarded along MSTI 2 rooted at Switch B.
- VLAN 30's frames are forwarded along MSTI 3 rooted at Switch A.
- VLAN 40's frames are forwarded along MSTI 4 rooted at Switch B.

Figure 21 MSTP application scenario

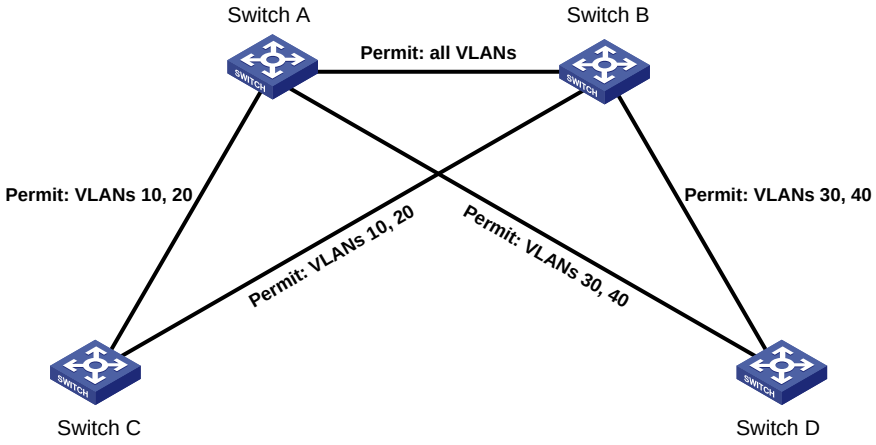
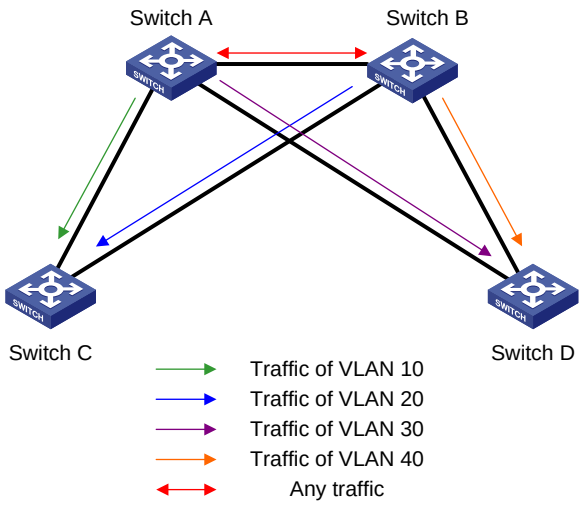


Figure 22 shows the forwarding paths for each VLAN. Traffic of the VLANs is distributed to different links, and each VLAN has alternative forwarding paths when link failure occurs.

Figure 22 Forwarding paths for each VLAN



You also can configure PVST to create spanning trees for the VLANs as follows:

- The spanning tree of VLAN 10 is rooted at Switch A.
- The spanning tree of VLAN 20 is rooted at Switch B.

- The spanning tree of VLAN 30 is rooted at Switch A.
- The spanning tree of VLAN 40 is rooted at Switch B.

References

- IEEE 802.1d, *Media Access Control (MAC) Bridges*
- IEEE 802.1w, *Part 3: Media Access Control (MAC) Bridges—Amendment 2: Rapid Reconfiguration*
- IEEE 802.1s, *Virtual Bridged Local Area Networks—Amendment 3: Multiple Spanning Trees*
- IEEE 802.1Q-REV/D1.3, *Media Access Control (MAC) Bridges and Virtual Bridged Local Area Networks —Clause 13: Spanning tree Protocols*