Chapter 2

The 802.1D STP standard defines the following three port types:

- **Root port (RP):** A network port that connects to the root bridge or an upstream switch in the spanning-tree topology. There should be only one root port per VLAN on a switch.
- **Designated port (DP):** A network port that receives and forwards BPDU frames to other switches. Designated ports provide connectivity to downstream devices and switches. There should be only one active designated port on a link.
- **Blocking port:** A network port that is not forwarding traffic because of STP calculations.

STP Key Terminology

Several key terms are related to STP:

- **Root bridge:** The root bridge is the most important switch in the Layer 2 topology. All ports are in a forwarding state. This switch is considered the top of the spanning tree for all path calculations by other switches. All ports on the root bridge are categorized as designated ports.
- **Bridge protocol data unit (BPDU):** This network packet is used for network switches to identify a hierarchy and notify of changes in the topology. A BPDU uses the destination MAC address 01:80:c2:00:00:00. There are two types of BPDUs:
  - **Configuration BPDU:** This type of BPDU is used to identify the root bridge, root ports, designated ports, and blocking ports. The configuration BPDU consists of the following fields: STP type, root path cost, root bridge identifier, local bridge identifier, max age, hello time, and forward delay.
  - **Topology change notification (TCN) BPDU:** This type of BPDU is used to communicate changes in the Layer 2 topology to other switches. It is explained in greater detail later in the chapter.
- **Root path cost:** This is the combined cost for a specific path toward the root switch.
- **System priority:** This 4-bit value indicates the preference for a switch to be root bridge. The default value is 32,768.
- **System ID extension:** This 12-bit value indicates the VLAN that the BPDU correlates to. The system priority and system ID extension are combined as part of the switch's identification of a bridge.
- **Root bridge identifier:** This is a combination of the root bridge system MAC address, system ID extension, and system priority of the root bridge.
- **Local bridge identifier:** This is a combination of the local switch's bridge system MAC address, system ID extension, and system priority of the local bridge.
- **Max age:** This is the maximum length of time that a bridge port stores its BPDU information. The default value is 20 seconds, but the value can be configured with the command **spanning-tree vlan** *vlan-id* **max-age** *maxage*. If a switch loses contact with the BPDU's source, it assumes that the BPDU information is still valid for the duration of the Max Age timer.
- **Hello time:** This is the time interval that a BPDU is advertised out of a port. The default value is 2 seconds, but the value can be configured to 1 to 10 seconds with the command **spanning-tree vlan** *vlan-id* **hello-time** *hello-time*.
- **Forward delay:** This is the amount of time that a port stays in a listening and learning state. The default value is 15 seconds, but the value can be changed to a value of 4 to 30 seconds with the command **spanning-tree vlan** *vlan-id* **forward-time** *forward-time*.

Root Bridge Election

The first step with STP is to identify the root bridge. As a switch initializes, it assumes that it is the root bridge and uses the local bridge identifier as the root bridge identifier. It then listens to its neighbor's configuration BPDU and does the following:

- If the neighbor's configuration BPDU is inferior to its own BPDU, the switch ignores that BPDU.
- If the neighbor's configuration BPDU is preferred to its own BPDU, the switch updates its BPDUs to include the new root bridge identifier along with a new root path cost that correlates to the total path cost to reach the new root bridge. This process continues until all switches in a topology have identified the root bridge switch.

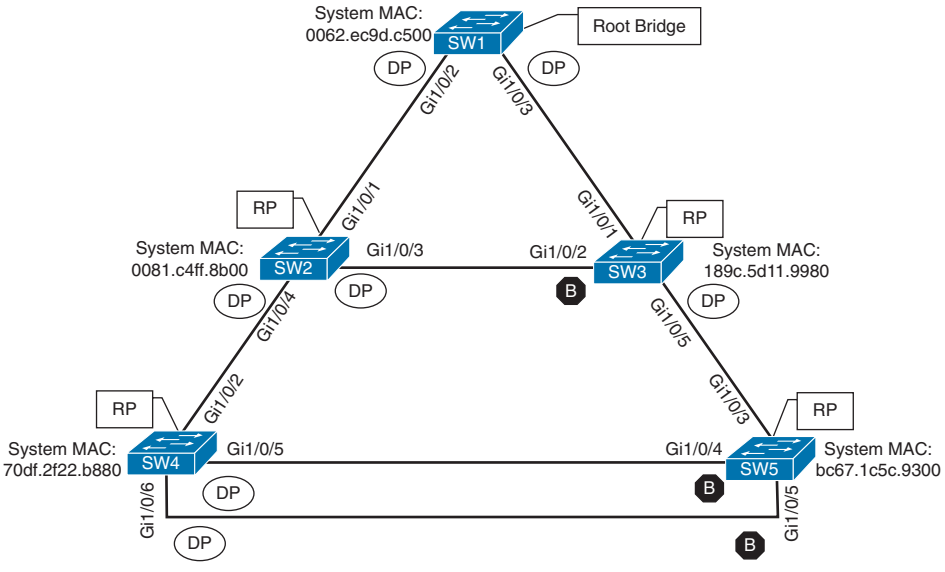

Figure 2-1 *Basic STP Topology*

STP deems a switch more preferable if the priority in the bridge identifier is lower than the priority of the other switch's configuration BPDUs. If the priority is the same, then the switch prefers the BPDU with the lower system MAC.

NOTE Generally, older switches have a lower MAC address and are considered more preferable. Configuration changes can be made for optimizing placement of the root bridge in a Layer 2 topology to prevent the insertion of an older switch from becoming the new root bridge.

In Figure 2-1, SW1 can be identified as the root bridge because its system MAC address (0062.ec9d.c500) is the lowest in the topology. This is further verified by using the command **show spanning-tree root** to display the root bridge. Example 2-1 demonstrates this command being executed on SW1. The output includes the VLAN number, root bridge identifier, root path cost, hello time, max age time, and forwarding delay. Because SW1 is the root bridge, all ports are designated ports, so the Root Port field is empty. Using this command is one way to verify that the connected switch is the root bridge for the VLAN.

Example 2-1 *Verifying the STP Root Bridge*

```
SW1# show spanning-tree root

                                        Root    Hello Max Fwd
Vlan                Root ID             Cost    Time  Age Dly  Root Port
---------------- -------------------- --------- ----- --- ---  ------------
VLAN0001         32769 0062.ec9d.c500         0    2   20  15
VLAN0010         32778 0062.ec9d.c500         0    2   20  15
VLAN0020         32788 0062.ec9d.c500         0    2   20  15
VLAN0099         32867 0062.ec9d.c500         0    2   20  15
```

In Example 2-1, notice that the root bridge priority on SW1 for VLAN 1 is 32,769 and not 32,768. The priority in the configuration BPDU packets is actually the priority plus the value of the *sys-id-ext* (which is the VLAN number). You can confirm this by looking at VLAN 10, which has a priority of 32,778, which is 10 higher than 32,768.

When a switch generates the BPDUs, the root path cost includes only the calculated metric to the root and does not include the cost of the port that the BPDU is advertised out of. The receiving switch adds the port cost for its interface on which the BPDU was received in conjunction with the value of the root path cost in the BPDU. The root path cost is always zero on the root bridge. Figure 2-2 illustrates the root path cost as SW1 advertises the configuration BPDUs toward SW3 and then SW3's configuration BPDUs toward SW5.

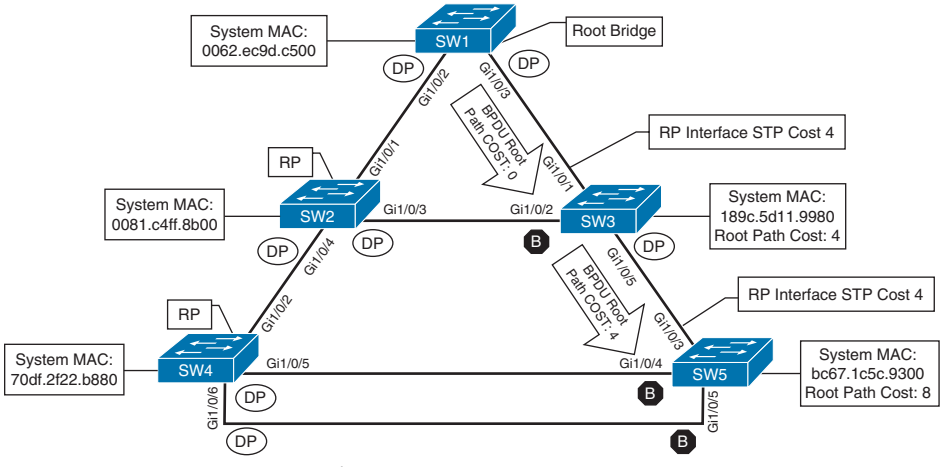

Figure 2-2 *STP Path Cost Advertisements*

Example 2-2 shows the output of the **show spanning-tree root** command run on SW2 and SW3. The Root ID field is exactly the same as for SW1, but the root path cost has changed

to 4 because both switches must use the 1 Gbps link to reach SW1. Gi1/0/1 has been identified on both switches as the root port.

Example 2-2 *Identifying the Root Ports*

```
SW2# show spanning-tree root
                                        Root    Hello Max Fwd
Vlan                Root ID             Cost    Time  Age Dly  Root Port
---------------- -------------------- --------- ----- --- ---  ------------
VLAN0001         32769 0062.ec9d.c500          4    2   20  15  Gi1/0/1
VLAN0010         32778 0062.ec9d.c500          4    2   20  15  Gi1/0/1
VLAN0020         32788 0062.ec9d.c500          4    2   20  15  Gi1/0/1
VLAN0099         32867 0062.ec9d.c500          4    2   20  15  Gi1/0/1
```

```
SW3# show spanning-tree root
                                        Root    Hello Max Fwd
Vlan                Root ID             Cost    Time  Age Dly  Root Port
---------------- -------------------- --------- ----- --- ---  ------------

VLAN0001         32769 0062.ec9d.c500          4    2   20  15  Gi1/0/1
VLAN0010         32778 0062.ec9d.c500          4    2   20  15  Gi1/0/1
VLAN0020         32788 0062.ec9d.c500          4    2   20  15  Gi1/0/1
VLAN0099         32867 0062.ec9d.c500          4    2   20  15  Gi1/0/1
```

Locating Root Ports

After the switches have identified the root bridge, they must determine their root port (RP). The root bridge continues to advertise configuration BPDUs out all of its ports. The switch compares the BPDU information received on its port to identify the RP. The RP is selected using the following logic (where the next criterion is used in the event of a tie):

1. The interface associated to lowest path cost is more preferred.
2. The interface associated to the lowest system priority of the advertising switch is preferred next.
3. The interface associated to the lowest system MAC address of the advertising switch is preferred next.
4. When multiple links are associated to the same switch, the lowest port priority from the advertising switch is preferred.
5. When multiple links are associated to the same switch, the lower port number from the advertising switch is preferred.

Example 2-3 shows the output of running the command **show spanning-tree root** on SW4 and SW5. The Root ID field is exactly the same as on SW1, SW2, and SW3 in Examples 2-1 and 2-2. However, the root path cost has changed to 8 on SW4 and SW5 because both switches must traverse two 1 Gbps links to reach SW1. Gi1/0/2 was identified as the RP for SW4, and Gi1/0/3 was identified as the RP for SW5.

Example 2-3 *Identifying the Root Ports on SW4 and SW5*

```
SW4# show spanning-tree root

                                        Root    Hello Max Fwd
Vlan                Root ID             Cost    Time  Age Dly  Root Port
---------------- -------------------- --------- ----- --- ---  ------------
VLAN0001         32769 0062.ec9d.c500         8    2   20  15    Gi1/0/2
VLAN0010         32778 0062.ec9d.c500         8    2   20  15    Gi1/0/2
VLAN0020         32788 0062.ec9d.c500         8    2   20  15    Gi1/0/2
VLAN0099         32867 0062.ec9d.c500         8    2   20  15    Gi1/0/2

SW5# show spanning-tree root
                                        Root    Hello Max Fwd
Vlan                Root ID             Cost    Time  Age Dly  Root Port
---------------- -------------------- --------- ----- --- ---  ------------
VLAN0001         32769 0062.ec9d.c500         8    2   20  15    Gi1/0/3
VLAN0010         32778 0062.ec9d.c500         8    2   20  15    Gi1/0/3
VLAN0020         32788 0062.ec9d.c500         8    2   20  15    Gi1/0/3
VLAN0099         32867 0062.ec9d.c500         8    2   20  15    Gi1/0/3
```

The root bridge can be identified for a specific VLAN through the use of the command **show spanning-tree root** and examination of the CDP or LLDP neighbor information to identify the host name of the RP switch. The process can be repeated until the root bridge is located.

STP Topology Changes

In a stable Layer 2 topology, configuration BPDUs always flow from the root bridge toward the edge switches. However, changes in the topology (for example, switch failure, link failure, or links becoming active) have an impact on all the switches in the Layer 2 topology.

The switch that detects a link status change sends a topology change notification (TCN) BPDU toward the root bridge, out its RP. If an upstream switch receives the TCN, it sends out an acknowledgment and forwards the TCN out its RP to the root bridge.

Upon receipt of the TCN, the root bridge creates a new configuration BPDU with the Topology Change flag set, and it is then flooded to all the switches. When a switch receives a configuration BPDU with the Topology Change flag set, all switches change their MAC address timer to the forwarding delay timer (with a default of 15 seconds). This flushes out MAC addresses for devices that have not communicated in that 15-second window but maintains MAC addresses for devices that are actively communicating.

Flushing the MAC address table prevents a switch from sending traffic to a host that is no longer reachable by that port. However, a side effect of flushing the MAC address table is that it temporarily increases the unknown unicast flooding while it is rebuilt. Remember that this can impact hosts because of their CSMA/CD behavior. The MAC address timer is then reset to normal (300 seconds by default) after the second configuration BPDU is received.

TCNs are generated on a VLAN basis, so the impact of TCNs directly correlates to the number of hosts in a VLAN. As the number of hosts increases, the more likely TCN generation is to occur and the more hosts that are impacted by the broadcasts. Topology changes should be checked as part of the troubleshooting process. Chapter 3 describes mechanisms such as portfast that modify this behavior and reduce the generation of TCNs.

Topology changes are seen with the command **show spanning-tree** [**vlan** *vlan-id*] **detail** on a switch bridge. The output of this command shows the topology change count and time since the last change has occurred. A sudden or continuous increase in TCNs indicates a potential problem and should be investigated further for flapping ports or events on a connected switch.

Example 2-7 displays the output of the **show spanning-tree vlan 10 detail** command. Notice that it includes the time since the last TCN was detected and the interface from which the TCN originated.

Example 2-7 *Viewing a Detailed Version of Spanning Tree State*

```
SW1# show spanning-tree vlan 10 detail

 VLAN0010 is executing the rstp compatible Spanning Tree protocol
 Bridge Identifier has priority 32768, sysid 10, address 0062.ec9d.c500
 Configured hello time 2, max age 20, forward delay 15, transmit hold-count 6
 We are the root of the spanning tree
 Topology change flag not set, detected flag not set
 Number of topology changes 42 last change occurred 01:02:09 ago
           from GigabitEthernet1/0/2
 Times: hold 1, topology change 35, notification 2
        hello 2, max age 20, forward delay 15
 Timers: hello 0, topology change 0, notification 0, aging 300
```

The process of determining why TCNs are occurring involves checking a port to see whether it is connected to a host or to another switch. If it is connected to another switch, you need to connect to that switch and repeat the process of examining the STP details. You might need to examine CDP tables or your network documentation. You can execute the **show spanning-tree** [**vlan** *vlan-id*] **detail** command again to find the last switch in the topology to identify the problematic port.

Converging with Direct Link Failures

When a switch loses power or reboots, or when a cable is removed from a port, the Layer 1 signaling places the port into a down state, which can notify other processes, such as STP. STP considers such an event a direct link failure and can react in one of three ways, depending on the topology. This section explains each of these three possible scenarios with a simple three-switch topology where SW1 is the root switch.

Direct Link Failure Scenario 1

In the first scenario, the link between SW2 and SW3 fails. SW2's Gi1/0/3 port is the DP, and SW3's Gi1/0/2 port is in a blocking state. Because SW3's Gi1/0/2 port is already in a blocking state, there is no impact to traffic between the two switches as they both transmit data through SW1. Both SW2 and SW3 will advertise a TCN toward the root switch, which results in the Layer 2 topology flushing its MAC address table.

Direct Link Failure Scenario 2

In the second scenario, the link between SW1 and SW3 fails. Network traffic from SW1 or SW2 toward SW3 is impacted because SW3's Gi1/0/2 port is in a blocking state. Figure 2-3 illustrates the failure scenario and events that occur to stabilize the STP topology:

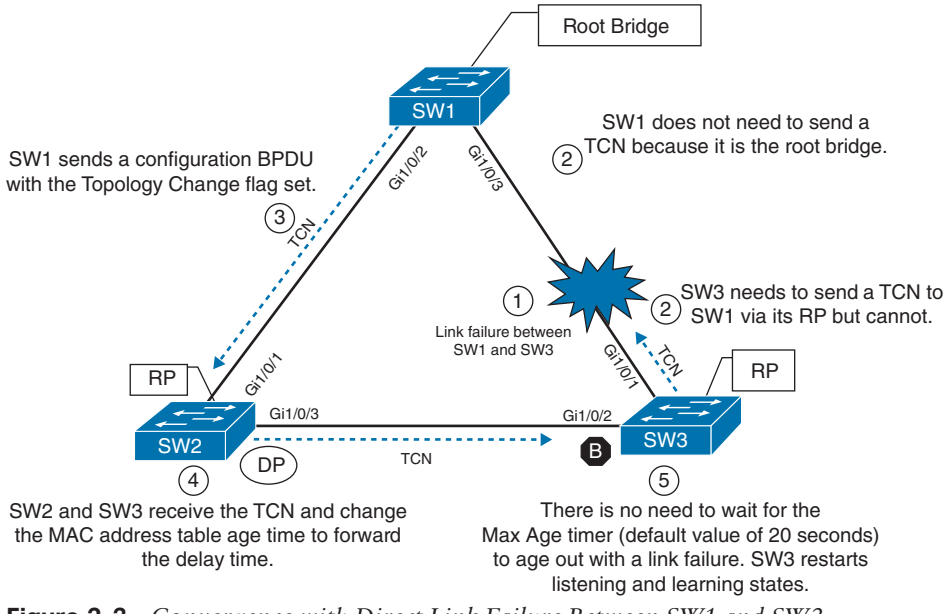

Figure 2-3 *Convergence with Direct Link Failure Between SW1 and SW3*

Phase 1. SW1 detects a link failure on its Gi1/0/3 interface. SW3 detects a link failure on its Gi1/0/1 interface.

Phase 2. Normally, SW1 would generate a TCN flag out its root port, but it is the root bridge, so it does not. SW1 would advertise a TCN if it were not the root bridge.

SW3 removes its best BPDU received from SW1 on its Gi1/0/1 interface because it is now in a down state. At this point, SW3 would attempt to send a TCN toward the root switch to notify it of a topology change; however, its root port is down.

Phase 3. SW1 advertises a configuration BPDU with the Topology Change flag out of all its ports. This BPDU is received and relayed to all switches in the environment.

NOTE If other switches were connected to SW1, they would receive a configuration BPDU with the Topology Change flag set also. These packets have an impact for all switches in the same Layer 2 domain.

Phase 4. SW2 and SW3 receive the configuration BPDU with the Topology Change flag. These switches then reduce the MAC address age timer to the forward delay timer to flush out older MAC entries. In this phase, SW2 does not know what changed in the topology.

Phase 5. There is no need to wait for the Max Age timer (default value of 20 seconds) to age out with a link failure. SW3 restarts the STP listening and learning states to learn about the root bridge on the Gi1/0/2 interface (which was in the blocking state previously).

The total convergence time for SW3 is 30 seconds: 15 seconds for the listening state and 15 seconds for the learning state before SW3's Gi1/0/2 can be made the RP.

Direct Link Failure Scenario 3

In the third scenario, the link between SW1 and SW2 fails. Network traffic from SW1 or SW3 toward SW2 is impacted because SW3's Gi1/0/2 port is in a blocking state. Figure 2-4 illustrates the failure scenario and events that occur to stabilize the STP topology:

Phase 1. SW1 detects a link failure on its Gi1/0/2 interface. SW2 detects a link failure on its Gi1/0/1 interface.

Phase 2. Normally SW1 would generate a TCN flag out its root port, but it is the root bridge, so it does not. SW1 would advertise a TCN if it were not the root bridge.

SW2 removes its best BPDU received from SW1 on its Gi1/0/1 interface because it is now in a down state. At this point, SW2 would attempt to send a TCN toward the root switch to notify it of a topology change; however, its root port is down.

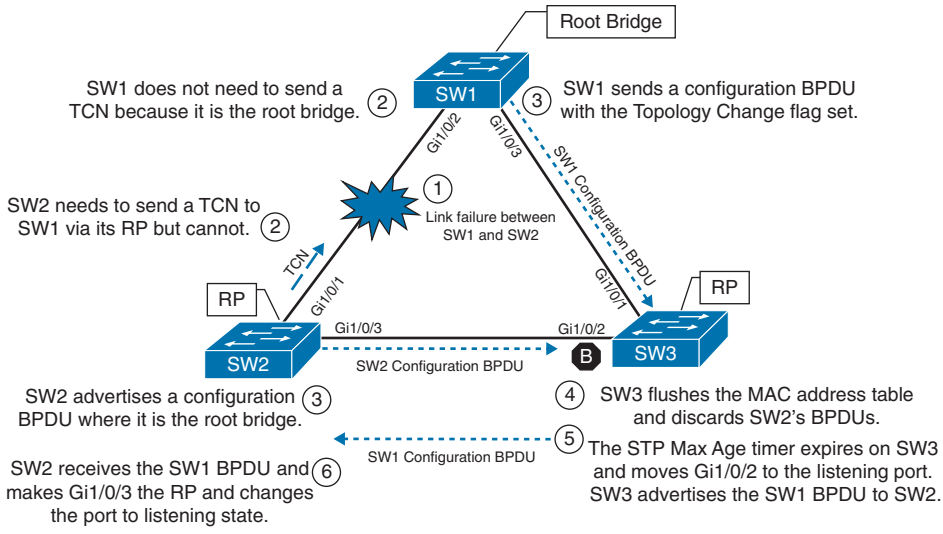

Figure 2-4 *Convergence with Direct Link Failure Between SW1 and SW2*

Phase 3. SW1 advertises a configuration BPDU with the Topology Change flag out of all its ports. This BPDU is then received and relayed to SW3. SW3 cannot relay this to SW2 because its Gi1/0/2 port is still in a blocking state.

SW2 assumes that it is now the root bridge and advertises configuration BPDUs with itself as the root bridge.

Phase 4. SW3 receives the configuration BPDU with the Topology Change flag from SW1. SW3 reduces the MAC address age timer to the forward delay timer to flush out older MAC entries. SW3 receives SW2's inferior BPDUs and discards them as it is still receiving superior BPDUs from SW1.

Phase 5. The Max Age timer on SW3 expires, and now SW3's Gi1/0/2 port transitions from blocking to listening state. SW3 can now forward the next configuration BPDU it receives from SW1 to SW2.

Phase 6. SW2 receives SW1's configuration BPDU via SW3 and recognizes it as superior. It marks its Gi1/0/3 interface as the root port and transitions it to the listening state.

The total convergence time for SW2 is 50 seconds: 20 seconds for the Max Age timer on SW3, 15 seconds for the listening state on SW2, and 15 seconds for the learning state.

Indirect Failures

In some failure scenarios, STP communication between switches is impaired or filtered while the network link remains up. This situation is known as an *indirect link failure*, and timers are required to detect and remediate the topology. Figure 2-5 illustrates an impediment or data corruption on the link between SW1 and SW3 along with the logic to resolve the loss of network traffic:

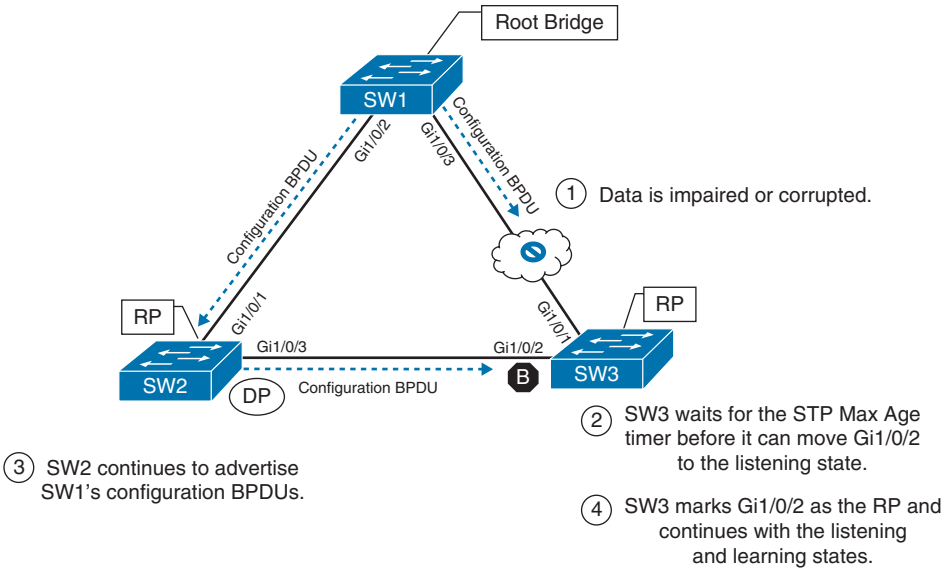

Figure 2-5 *Convergence with Indirect Link Failure*

Phase 1. An event occurs that impairs or corrupts data on the link. SW1 and SW3 still report a link up condition.

Phase 2. SW3 stops receiving configuration BPDUs on its RP. It keeps a cached entry for the RP on Gi1/0/1. SW1's configuration BPDUs that are being transmitted via SW2 are discarded because SW3's Gi1/0/2 port is in a blocking state.

After SW3's Max Age timer expires and flushes the RP's cached entry, SW3 transitions Gi1/0/2 from blocking to listening state.

Phase 3. SW2 continues to advertise SW1's configuration BPDUs toward SW3.

Phase 4. SW3 receives SW1's configuration BPDU via SW2 on its Gi1/0/2 interface. This port is now marked as the RP and continues to transition through the listening and learning states.

The total time for reconvergence on SW3 is 50 seconds: 20 seconds for the Max Age timer on SW3, 15 seconds for the listening state on SW3, and 15 seconds for the learning state on SW3.

Rapid Spanning Tree Protocol

Although 802.1D did a decent job of preventing Layer 2 forwarding loops, it used only one topology tree, which introduced scalability issues. Some larger environments with multiple VLANs need different STP topologies for traffic engineering purposes (for example, load-balancing, traffic steering). Cisco created Per-VLAN Spanning Tree (PVST) and Per-VLAN Spanning Tree Plus (PVST+) to allow more flexibility.

PVST and PVST+ were proprietary spanning protocols. The concepts in these protocols were incorporated with other enhancements to provide faster convergence into the IEEE 802.1W specification, known as Rapid Spanning Tree Protocol (RSTP).

RSTP (802.1W) Port States

RSTP reduces the number of port states to three:

- **Discarding:** The switch port is enabled, but the port is not forwarding any traffic to ensure that a loop is not created. This state combines the traditional STP states disabled, blocking, and listening.
- **Learning:** The switch port modifies the MAC address table with any network traffic it receives. The switch still does not forward any other network traffic besides BPDUs.
- **Forwarding:** The switch port forwards all network traffic and updates the MAC address table as expected. This is the final state for a switch port to forward network traffic.

Building the RSTP Topology

With RSTP, switches exchange handshakes with other RSTP switches to transition through the following STP states faster. When two switches first connect, they establish a bidirectional handshake across the shared link to identify the root bridge. This is straightforward for an environment with only two switches; however, large environments require greater care to avoid creating a forwarding loop. RSTP uses a synchronization process to add a switch to the RSTP topology without introducing a forwarding loop. The synchronization process starts when two switches (such as SW1 and SW2) are first connected. The process proceeds as follows:

1. As the first two switches connect to each other, they verify that they are connected with a point-to-point link by checking the full-duplex status.
2. They establish a handshake with each other to advertise a proposal (in configuration BPDUs) that their interface should be the DP for that segment.
3. There can be only one DP per segment, so each switch identifies whether it is the superior or inferior switch, using the same logic as in 802.1D for the system identifier (that is, the lowest priority and then the lowest MAC address). Using the MAC addresses from Figure 2-1, SW1 (0062.ec9d.c500) is the superior switch to SW2 (0081.c4ff.8b00).
4. The inferior switch (SW2) recognizes that it is inferior and marks its local port (Gi1/0/1) as the RP. At that same time, it moves all non-edge ports to a discarding state. At this point in time, the switch has stopped all local switching for non-edge ports.
5. The inferior switch (SW2) sends an agreement (configuration BPDU) to the root bridge (SW1), which signifies to the root bridge that synchronization is occurring on that switch.
6. The inferior switch (SW2) moves its RP (Gi1/0/1) to a forwarding state. The superior switch moves its DP (Gi1/0/2) to a forwarding state too.
7. The inferior switch (SW2) repeats the process for any downstream switches connected to it.

RSTP Convergence

The RSTP convergence process can occur quickly. RSTP ages out the port information after it has not received hellos in three consecutive cycles. Using default timers, the Max Age would take 20 seconds, but RSTP requires only 6 seconds. And thanks to the new synchronization, ports can transition from discarding to forwarding in an extremely low amount of time.

If a downstream switch fails to acknowledge the proposal, the RSTP switch must default to 802.1D behaviors to prevent a forwarding loop.