The ABCs of Spanning Tree Protocol

INTRODUCTION

In an industrial automation application that relies heavily on the health of the Ethernet network that attaches all the controllers and computers together, a concern exists about what would happen if the network fails. If the result is loss of production or loss of a processed batch or the endangerment of people or equipment, redundancy schemes are examined. Since cable failure is the most likely mishap, cable redundancy is suggested by configuring the network in either a ring or by carrying parallel branches. If one of the segments is lost, then communication will continue down a parallel path or around the unbroken portion of the ring. The problem with these approaches is that Ethernet supports neither of these topologies without special equipment. However, this issue is addressed in an IEEE standard numbered 802.1D that covers bridges, and in this standard the concept of the Spanning Tree Protocol (STP) is introduced.

IEEE 802.1D

The redundancy issue is addressed by ANSI/IEEE Std 802.1D, 1998 edition: Information technology — Telecommunications and information exchange between systems — Local and metropolitan area networks — common specifications — Part 3: Media Access Control (MAC) Bridges. The equipment covered by the standard is called a bridge. A bridge is used to connect two or more local area networks (LANs) at the MAC level, which is layer two in the ISO Reference Model. Generally the MAC type can be different on each LAN, but usually Ethernet LANs are on either side of a bridge. Interconnecting LANs by using bridges creates a Bridged LAN whereby end stations located on different LANs can communicate as if the bridges were not present.

Cable redundancy introduces loops in the topology and, as we will see, these loops must be disabled. An industrial automation user may want loops to guard against a primary cable failure while an office automation user may want to guard against an inadvertent loop. The 802.1D standard addresses both situations.

Bridge Operation

If you understand how an Ethernet switch works, you know how a bridge operates. However, all the requirements of a bridge (e.g., STP) are not always present in a switch. A bridge needs to relay and filter frames and it must make independent decisions about when to do this.

Look at Figure 1. In a two-port Ethernet bridge, each port has an Ethernet-type MAC port connected to a separate LAN and a filtering database (memory) shared by both ports. Within each LAN is a collection of end stations, repeating hubs and simple plug-and-play switches. Each end station has a unique MAC address. For simplicity, we will assume ordinary integers although true Ethernet MAC addresses are 48 bits long. In our example, three numbered end stations are present in each LAN. Assume Bridge 1 has recently been powered and its memory cleared (Bridge 2 will be added later). Station 1 sends a message to station 11 followed by Station 2 sending a message to Station 11. These messages will traverse the bridge from one LAN to the other. This process is called relaying or forwarding. The database in the bridge will note the source addresses of Stations 1 and 2 as arriving on Port A. This process is called learning. When Station 11 responds to either Station 1 or 2, the database will note that Station 11 is on Port B. If Station 1 sends a message to Station 2, the bridge will do nothing since it realizes that because Stations 1 and 2 are on the same LAN their message does not need to be shared with other LANs. This process is called filtering. If Station 1 ceases to initiate messages for a period of time, the bridge will erase Station 1 from its database — requiring the location of Station 1 to be relearned. This is called aging.

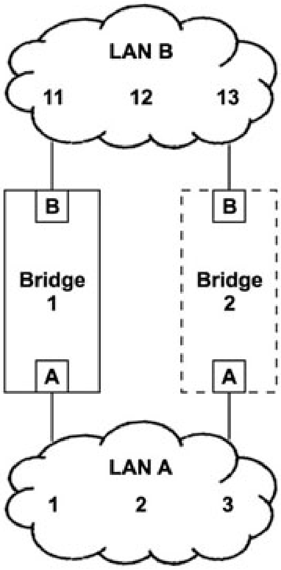


Figure 1 — The addition of Bridge 2 creates a loop.

The above examples are all directed or unicast messages — meaning that one station is sending a message to another station. With multicast (one station to many stations) or broadcast messages (one station to all other stations), the bridge will forward messages to all stations since it may not know the actual location of all stations. This process is called flooding.

Looking at the same Figure 1, Bridge 2 is now added to parallel Bridge 1. This gives us a redundant path, but it also creates a loop with the following adverse results. When Station 1 initiates a message to Station 11, this message is forwarded by Bridge 1 and appears on LAN B. Bridge 2 interprets this message as originating on LAN B so it forwards the message to LAN A while incorrectly noting that Station 1 is located on LAN B. When Station 1 initiates a second message, Bridge 2 interprets this action as if Station 1 has now moved to LAN A from LAN B and resets its filtering database accordingly. Now assume that Station 1 sends out a broadcast message. Bridge 1 will forward the message to LAN B. Bridge 2 will observe the message on LAN B and forward it to LAN A. Bridge 1 will observe this message on LAN A as a new message and forward it to LAN B again — initiating an endless cycle, totally consuming the bandwidth of both bridges and rendering both LANs useless. To maintain the integrity of our network, we must guard against the formation of loops.

Tree Topology

To avoid loops we need a tree topology consisting of a root, a succession of branches and then leaves. The leaves represent end stations, and there is one and only one path from a leaf to another leaf. Therefore, the tree is free of loops that can cause havoc in a network. The other requirement is that all leaves are connected. There are no isolated segments. Another term for this topology is distributed star. Within our tree structure will be a series of bridges used to connect the branches and leaves. There are two types. The root bridge is the main one of interest because it has a special assignment and there is only one within a network. The other bridges (that are to be used) are all designated bridges and there could be many within the network. To have a tree topology, you need bridges with more than two ports.

Port Designations

Although bridges do not need MAC addresses to operate, a MAC address is needed to identify bridges in the Spanning Tree Protocol. Besides a MAC address for each bridge, each port on each bridge must be identified. For bridges, a unique 64-bit bridge identifier is assigned by appending a 16-bit priority field in front of a unique 48-bit MAC address resulting in a Bridge ID. The MAC address comes from the bridge vendor, but the priority field can be set by the user. The default priority value of 0x8000 is in the middle of the priority range. If the user fails to assign priority values, the bridges will still have unique assignments. This is important since the bridge with the lowest numerical bridge identifier will become the root bridge. All other bridges have the possibility of becoming designated bridges.

Similarly, a 16-bit "port identifier" exists consisting of an 8-bit port address preceded by an 8-bit priority field. Again, the user sets the priority field while the bridge vendor sets the port addresses usually beginning with one for port one and so on. The default priority field is 0x80. Now we have all the bridges and ports identified including the root bridge.

To avoid loops, there is one and only one bridge that is responsible for forwarding messages from the direction of the root towards branches, which we will call links. If there is only one path from the root to a link where end stations (leaves) attach, there will be no loops. We need a forwarding policy, which is called the Spanning Tree Protocol. To implement this policy, we need to assign each port to become either a designated port or a root port. A designated port is a port that forwards traffic away from the root and towards the leaves. A root port carries traffic back to the root bridge with the further requirement that no more than one root port exists on any one designated bridge. Looking at Figure 2, notice that the root bridge (R) has all designated ports (because it is the root) while the designated bridges have root ports when connected to the root bridge or when the direction of flow is towards the root.

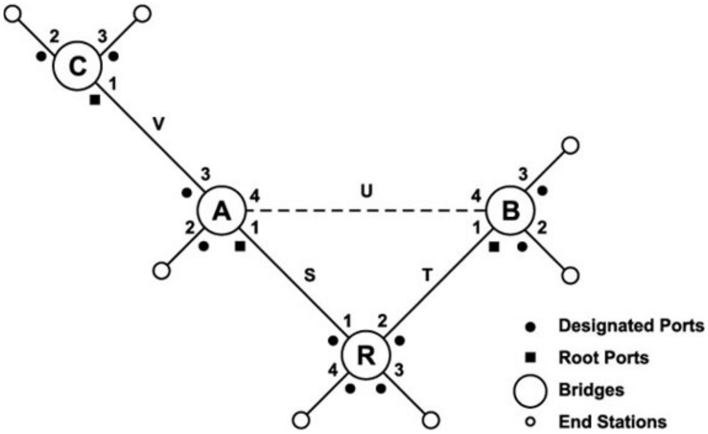


Figure 2 — The addition of link U creates a redundant path. Bridge R has the lowest bridge ID and is, therefore, the root.

Path Cost

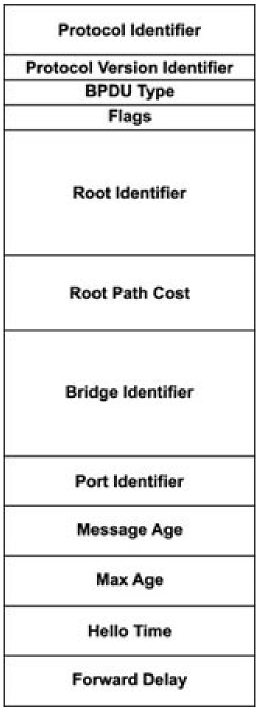
The next item to specify is the cost of each link between branch devices (end stations are not counted). Since we do not want to forward traffic onto low-speed links if we can avoid it, we assign a link cost based upon port speed. The recommended cost figures are shown in Table 1. A 100 Mbps port would cost 19 while a 10 Mbps port would cost 100. Two 100 Mbps cascaded links (two links requiring two bridges to reach the root) would cost 38. This would represent the path cost. Therefore, all designated ports on Bridge C in Figure 2 would advertise a path cost of 38 — assuming traffic is at 100 Mbps. STP uses path cost to the root for arbitrating the ideal route. All other routes are blocked. If path costs for each port on each bridge are manually assigned by the operator, configuration could be tedious. Once these costs are entered, we can let the STP determine the best topology that will not introduce loops.

Recommended Cost Figures	
Data Rate	Recommended Link Cost Value
4 Mb/s	250
10 Mb/s	100
16 Mb/s	62
100 Mb/s	19
1 Gb/s	4
10 Gb/s	2

Table 1 — Link Cost Recommendations

BPDUs

Bridges must communicate with one another to execute the STP, and they accomplish this by sending configuration messages in the form of Bridge Protocol Data Units (BPDUs). The BPDU is sent as a multicast message (01-80-C2-00-00-00) within a reserved range of MAC addresses which are consumed by each bridge and not forwarded. Each bridge must periodically advertise its understanding of the topology and the path cost to the root for each of its ports. The BPDU format is shown in Figure 3. The more important fields will be discussed in the following example.



Protocol Identifier
Protocol Version Identifier
BPDU Type
Flags
Root Identifier
Root Path Cost
Bridge Identifier
Port Identifier
Message Age
Max Age
Hello Time
Forward Delay

Figure 3 — There are two types of Bridge Protocol Data Unit types. Configuration messages are normally sent; however, a designated bridge can initiate a Topology Change message.

Refer to Figure 2 where we have four interconnected bridges connected to end stations. We will assume a stable network with a duly elected root bridge (R) and several designated bridges (A, B, C). All ports on the root bridge are designated ports since they emanate to end stations from the root. All other ports are designated ports or candidates for designated ports except three ports that point towards the root. They are Port 1 on Bridge A, Port 1 on Bridge B and Port 1 on Bridge C. These are root ports. Assume that all ports on all bridges are rated for 100 Mbps except Bridge B which is only rated for 10 Mbps. Examining Table 1 we find that the corresponding link cost for 100 Mbps is 19 and for 10 Mbps it is 100. This information will be needed when constructing the BPDUs. There are four links of interest (S, T, U, and V).

The root bridge begins the process of sending a periodic configuration message based upon the Hello Time which is typically two seconds. In this message, the Root Identifier and Bridge Identifier would be the same since this bridge thinks it is the root by having the lowest-value bridge identifier. The Root Path Cost will be zero because this is the root bridge.

Since all ports on the root bridge are designated ports, a configuration BPDU will go out each port along with the corresponding port identifier. This process repeats every Hello Time.

First-tier bridges (those directly connected to the root bridge — in our case, Bridges A and B) will receive the BPDUs on Links S and T and will analyze the data. Each bridge will verify that the Root Identifier is indeed lower than its own bridge identifier. If that is true, each bridge will assemble its own set of BPDUs for transmission out its designated ports. The Root Identifier Field will not change. Each designated port will increase the total cost of getting back to the root by that port's individual Root Path Cost. Since both of these are tier-1 bridges, the Root Path Cost would be the link cost for Bridge A (19) and for Bridge B (100). The bridge and port identifiers would represent data from each bridge. The Hello Time will remain that specified by the root. BPDUs are then sent out each bridge's designated ports. Second-tier bridges receive the BPDUs (in our case, Bridge C) and the path cost is upped again. In our example, the Root Path Cost from Bridge C would be 38. Bridge C will send out its set of BPDUs — ending the process since no more tier bridges remain. End stations do not participate in the process and ignore the messages. Because the propagation of messages from the root will take time, the STP standard establishes an arbitrary limit to the number of cascaded bridges at seven. This limit does not apply to conventional switches that, in any event, should not be present in an STP network.

Port States

The operational states of ports participating in the STP are a bit different from a conventional switch. Additional states are needed to prevent a loop, and to limit instability during the voting or topology-change process. There are five states.

- **Disabled**

  This port is completely non-functional in that it cannot receive or transmit any type of frame.

- **Blocking**

  This port is neither a designated nor root port but is recognized as an alternate port to the root. It does not learn addresses, forward frames or transmit BPDUs. It can hear BPDUs being sent since it may be called to action one day.

- **Listening**

  This port is being prepared for activity by exiting the blocking state. It still does not learn or forward addresses, but it sends and receives BPDUs. It is participating in the voting, but it might not win the election.

- **Learning**

  This port will become active in forwarding frames but must wait until the Forward Delay timer expires (typically 15 seconds). This allows the port to add entries in its filtering database so it will not flood ports once it enters the forwarding state.

- **Forwarding**

  This port is functioning as any other switch port by filtering and forwarding frames.

We will go back to our example in Figure 2. Since we have a tree topology without loops, all active ports (those with a link partner) are in the forwarding state. Now consider the presence of Link U. Unused Port 4 on Bridges A and B will be active but only in the listening state. While in this state, they send each other BPDUs. Bridge A will send out a BPDU to Bridge B indicating a root path cost of 19 while Bridge B reciprocates with a root path cost of 100. Both bridges recognize there is an alternate path to the root, which is unacceptable, and that (for messages not originating at the end stations attached to Bridge B) the best path to the root is through Bridge A and not Bridge B. Therefore Port 4 on Bridge B assumes the blocking state, and Port 4 on Bridge A the learning state — and eventually the forwarding state, even though this would be useless since all the messages forwarded from Bridge A to Bridge B would be discarded on arrival. The final result is no topology change. The tree before the redundant connection is the same as after the connection was made. However, Link U is now a potential redundant path which may be utilized after a link or switch failure.

Topology Change

In the above example, adding Link U did not result in a change in the tree topology. However, a topology change can occur due to a lost link, a lost bridge, the addition of a link or bridge, or by management changing the priorities of bridges. What happens if the root bridge fails? STP guards against all these occurrences by monitoring configuration BPDUs, observing that a BPDU failed to arrive, or by generating a Topology Change BPDU.

There are two types of BPDUs as identified in the BPDU Type field. The configuration type is the normal BPDU as shown in Figure 3. The topology-change BPDU is similar to the configuration BPDU except that no data is transmitted below the Type field. This BPDU is generated by one of the designated bridges that changed its topology. An intervening designated bridge will acknowledge the originator's topology-change message by sending a configuration message with the Topology Acknowledge bit set in the Flags field. A new topology-change BPDU will then be sent towards the root. Any intervening designated bridge would repeat the process until the root is notified. The root bridge notices the message and informs all its attached designated bridges of the topology change by setting the Topology Change bit in the flags field and sending out a new configuration message. While this flag is set, all designated bridges reduce their aging time to that of the Forward Delay timer in anticipation of the topology change. Since the topology change could possibly make the data in the filtering database invalid, it must be quickly cleared and the new location of end stations relearned. Under normal conditions, the standard recommends a default aging time of 5 minutes! Changing the time to 15 seconds would be a great help in relearning address locations. Only after the root clears the topology change bit will the designated bridges resume their normal aging time and begin the learning and forwarding operations for the new topology.

Rapid Spanning Tree Protocol (RSTP)

The Rapid Spanning Tree Protocol was introduced by the IEEE as 802.1w. This technology was then absorbed by IEEE 802.1D-2004. RSTP is based upon the older STP standard and is backward compatible. RSTP and STP are similar in many areas. RSTP was created to provide faster recovery (convergence time) from topology changes.

RSTP provides faster recovery by monitoring the link status of each port and then generating a topology change after a link status change. Most switches can detect the link status changes very quickly (usually under 1 second). STP switches only detect topology changes when BPDU messages fail to reach their destination.

RSTP also improves recovery time by adding a new port designation. The designation *alternate port* is used for a port that acts as a backup to the root port. If the root port is lost, the alternate port can be quickly used as the new root port. An RSTP switch automatically determines its alternate ports and its root port.

RSTP changes the possible port states. In RSTP there are only three states: discarding, learning and forwarding. RSTP also slightly modifies the BPDU format—defining additional bits in the flag field.

RSTP provides backward compatibility to STP. When a switch receives an STP BPDU, it will apply the STP standard on that port. However, this could increase the recovery time of the network.

Summary

This article provides an introduction to STP and RSTP. Since these protocols are complex, not all issues have been addressed. Since protocol timers and aging times can vary, it is impossible to predict the time it would take for a network to stabilize after a topology change. An advantage of STP and RSTP is that they are not specific to Ethernet and can operate over wide area networks (WANs). For supervisory control and data acquisition systems (SCADA), the speed of STP recovery to topology changes might be adequate when temporary loss of communication will not render local control useless. But if STP is too slow in implementing a topology change your industrial network, RSTP should be considered.

REFERENCES

ANSI/IEEE Std 802.1D™-1998, Part 3: Media Access Control (MAC) Bridges, The Institute of Electrical and Electronics Engineers, Inc.

ANSI/IEEE Std 802.1D™-2004, Part 3: Media Access Control (MAC) Bridges, The Institute of Electrical and Electronics Engineers, Inc.

ANSI/IEEE Std 802.1w-2001, Part 3, Amendment 2: Rapid Reconfiguration, The Institute of Electrical and Electronics Engineers, Inc.

The Switch Book, Rich Seifert, 2000 Wiley Computer Publishing.