Layer 2 Connectivity: Bridge Spanning Tree Protocol

Burton Rosenberg

February 15, 2015

Version history: 30 December 2002, version one.

Introduction

Layer 2 connectivity is marked by devices that negotiate the passage of Layer 2 data frames. These devices are called either *switches, bridges* or *hubs*. The world is full of such devices — the ever popular access point in the homes of home Internet subscribers are bridges that take data in the Layer 2 protocol of Wifi and pass it to the Layer 2 protocol of ethernet.

The differences between switches, bridges and hubs are not particularly relevant or clear in a world of ever changing technologies. The Spanning Tree Protocol (STP) uses the word bridge because that's what was around when the protocol was defined by Internet engineer Radia Perlman (see her book *Interconnections* for an excellent introduction to Layer 2 protocols and Layer 3 protocols). Among participants in the STP are certainly any switch, and I doubt whether there are any hubs left in the world that haven't had their technology promoted to the level that they deserve to be called a switch or a bridge. I'll use the words bridge and switch more or less interchangeably.

A Local Area Network (LAN) is built up out of *network segments*, where is segment is a Level 1 communication channel. It can be a SSID in Wifi, where multiple clients associate to an access point and share the wireless channel, or it could be a single wire connecting together two devices in a paired, point-to-point channel. The bridges are to decide what data packets to forward in order to create from the various segments an entire, consistent and connected LAN.

Creating and managing the interconnections between bridges has to be transparent to the end devices, although the bridges can and do collaborate. This is the important notion of a *transparent bridge*, and is necessary because the bridges might have to be retro-fitted into an existing Level 2 technology. Any interference with that protocol by the bridge might have unforeseen consequences and cause failures. The STP protocol is a protocol between bridges whereby they collaborate to connect together the network segments.

Conceptual overview of the protocol

We model the overall LAN as a graph $G = (V, E)$ where the vertices V are the bridges and the edges E are the network segments, also known as links. If the links are all point-to-point, we have a standard graph. Else we have a *multi-graph*. Given a wired up LAN, considered as a graph G, STP is a distributed algorithm run by the bridges to find and spanning tree in this graph, come to consensus about that tree, and to maintain the tree under changes to the graph, such as bridges or link failures, or newly inserted bridges or links.

The STP protocol shall have the bridges communicate through *Bridge Protocol Data Units (BPDU)* as defined by specification IEEE 802.1D, and by distributed algorithm among the bridges:

1. Elect a root bridge, which shall become the root of the spanning tree.
2. Assign to each bridge the breath-first depth of the bridge from the elected root.
3. Assign to each port in the bridge one among the states: *root, designated or blocking.*
4. Once done, move the bridge from listen/learning state to forwarding state wherein the bridge forwards from each root or designated port packets to all root and designated ports, except the port from whence came the packet.
5. Maintain the assignment of root bridge and root, designated and blocking ports through hello messages emanating from the root;
6. And react to changes in the network, either the failure to receive hello message due to bridge or link failure, or to adapt to new bridge or links added to the network.

The spanning tree algorithm

The algorithm proceeds by the exchange of BPDU's between bridges sharing a data link. Broadcast MAC addresses are used to alert all bridges to accept and process the BPDU.

The essential contents of the message is the tuple (r, d, b, p) where r is the identifier of the assumed root bridge, d the current distance of the message sender from the assumed root bridge, an integer, b the identifier of this bridge, and p the port identifier on this bridge.

Since the root will be elected as the bridge with smallest identifier, the assignment of identifier to bridge can completely control the preference order in which a bridge becomes the root bridge. Without specific configuration, a bridge will use a MAC address on one of its ports as its identifier. A *priority* is configurable, and will be a prefix to the MAC, thereby taking precedence to the MAC, with ties being broken between bridges of equal assigned priority by bridge MAC.

Election of the root bridge and distance

On startup, each bridge assigns itself as the assume root bridge, with distance zero, and broadcasts the BPDU to the links.

As the bridge receives BPDUs from the links, it updates its assumed root bridge with the smallest root bridge found in the received BPDUs, and further outgoing BPDUs use this updated value. Also, the distance is updated to the sum of the distance claimed by the incoming packet plus the link distance. For instance, if the current assumed root and distance is (r, d) and a BPDU claims (r', d') with $r' < r$ then the bridge will update its root bridge to r' with distance $d' + \delta$, where δ is the distance assigned to the port which came the BPDU.

Note also, that if the port receives a BPDU (r, d'), and $d' + \delta < d$, and update will also be made, to reflect a new, lesser, distance has been found to bridge r.

it will thenceforth send to the link PBDUs the new r and d.

In this way, the bridges will forward the information until the network converges on a single r, which his proclaimed root, and for every bridge a distance d, the shortest distance from it to root bridge r.

Root, designated, or blocked ports

Next, the spanning tree is formed by the selection of edges in the spanning tree. Selected edges will be those ending on a port labeled root or designated; the edges with at least one end on a blocked blocked port will edges not in the tree. These would be redundant or loop-causing edges in the network.

The root ports will be the port with through which packets flow to the root bridge from this bridge. The root bridge is the only bridge with no root port. All other bridges have exactly one root port.

To perform this determination, the BPDU contents are ordered according to "bestness" Of two BPDUs $B_1 = (r_1, d_1, b_1, p_1)$ and $B_2 = (r_2, d_2, b_2, p_1)$, B_1 *is better than* B_2, written $B_1 < B_2$, if:

1. $r_1 < r_2$, or
2. $r_1 = r_2$ and $d_1 < d_2$, or
3. the above holds, and $d_1 = d_2$, and $b_1 < b_2$, or
4. the above holds, and $b_1 = b_2$, and $p_1 < p_2$.

The purpose of including the port number in the bestness calculation is to allow for two bridges to be connected more than once through two links. This might happen for redundancy, if the links are considered vulnerable to disruption, perhaps multiple links are created. The STP will then automatically switch over if a link fails. The redundancy might be inadvertent. If there is not a scheme for handling redundant links, this inadvertent redundancy could cause the entire network to fail with a loop.

That said, there are better ways to multiply connect two bridges, called *port bonding*. With port bonding, multiple links are treated as a single channel, with combined data rate the sum of the individual data rates. Bonding protocols can contain in them allowances for failures of one among

the bonded channels. This would be a better way to achieve redundancy, benefiting from improved data rates when all channels are heathy.

While the bridge is in *learning mode* or in *listening mode*, it is not forwarding any data packets across its ports. It is collecting up all the BPDUs it has heard on a link, including its own BPDU sent to that port, and converges on the best BPDU, by the given bestness computation. Let this be $B(b, p)$, as it is an eventually consistent assignment of a BPDU to a (bridge, port) combination.

The labeling of the ports is then:

1. For each port p on a bridge b, if the best BPDU $B(b, p)$ is equal to the BPDU sent by bridge b on port p, the port is labeled *designated*.
2. Considering now the best BPDU received by any port on a bridge, $B(b) = \min_{p'} B(b, p')$, if the minimum is achieved on port p, $B(b, p) = B(b)$, and the BPDU is received on port p, not sent by b on port p, the port is labeled *root*.
3. Finally, all ports not labeled designated or root are labeled *blocking*.

Forwarding mode

While in learning and listening mode, the bridge does not forward packets, for fear that a loop exists in the network, and packets will be forwarded around the loop, even multiplied by the fact that bridges duplicate the an incoming packet to send on multiple outgoing ports, causing a *packet storm*.

Once the STP algorithm converges, bridges are in agreement on the root bridge, they have a consistent value of distance form the root bridge, and consistent port labelings. The bridge will then enter forwarding mode, and will pass packet traffic among ports.

Forwarding mode: Traffic is accepted from the root port and any designated port and forwarded to the root port and all designated ports, except the port on which the traffic was received. Traffic is neither accepted from, nor forwarded to, any blocked port.

The root bridge will have only designated ports and blocked ports. If a port is blocked on a root bridge, another port connecting this bridge to the link will be designated. All other bridges will have a root port and zero or more designated ports.

Each link will be connected to exactly one designated port. It is said that the bridge with this port is the designated bridge for the link. The remaining ports on the link will be blocked or root. Among all ports connecting a bridge to a link, all except one will be blocked. For instance, if a bridge is designated for a link, it has no root port to the link, since all other ports to this link are blocked.

In the case of all point-to-point channels, and the graph is not a multi-graph, the situation can be simplified. Every link has two ends, with port A and port B. Of the two sides, one is sending the

better BPDU and is designated. The other is receiving a better BPDU, and if this BPDU is the best of all BPDUs received by the bridge, then the port is root, else it is blocked.

Analysis and Robustness

To do: All of this section needs to be done.

References

1. Radia Perlman, Interconnections: Bridges and Routers, Chapter 3.