Rational Trees

Mick Seaman

This note details a simplified MST bridge model. While retaining the broad concepts of .1sD4, and supporting the functionality desired for 802.1s, Multiple Spanning Trees (including rapid reconfiguration), this model provides some valuable simplifications. It uses the 802.1w RSTP state machines discussed at the May 2000 802.1 interim meeting without changes; requires only a single BPDU format – one capable of carrying spanning tree information for all trees and compatible with MST unaware RSTP compliant bridges; and facilitates proof of correctness under dynamic reconfiguration. Interoperability with Single Spanning Tree bridges running the original algorithm is, of course, maintained.

Minor changes to the .1w state machine procedures for classifying received BPDU information in the Port Information Machine, and for performing role selection in the Port Role Selection machine, are required.

Overview

For the purposes of calculating spanning tree topologies, and for propagating topology changes etc., an MST (multiple spanning tree) bridge is modeled as comprising a number of bridges, each with a bridge port for each physical port in the system. Every bridge and bridge port uses the RSTP state machines, with minor modifications to information handling procedures.

One 'bridge' calculates the 'CI' or 'common and internal' spanning tree. This is the only tree that exists for SST (single spanning tree) bridges. In SST regions of the network it is referred to simply as the 'common' spanning tree. It determines the active topology (forwarding connectivity) for all VLANs in SST regions.

The 'CI' tree propagates common spanning tree information across each MST region. The other bridges that form the MST bridge system each calculate an 'M' tree for the region. Each 'M' tree determines the active topology for a known set of VLANs within an MST region.

To select bridge ports to form a fully connected ('spanning') loop free ('tree') active topology, the spanning tree algorithm proceeds by 'labeling' each bridge and each LAN in a Bridged Local Area Network with a metric or 'priority'. The active topology comprises the ports that lie on the highest priority (i.e. the lowest cost) path to the highest priority Root Bridge from each LAN.

Calculating Spanning Trees

The priority information at any point in a spanning tree comprises multiple components, of decreasing relative priority. Differences in earlier components swamp all differences in later ones when comparisons are performed. The priority information carried in the common spanning tree BPDUs can be represented as:

$Priority = 0 - C - D$

Where C comprises the Root Bridge and Root Path Cost components and D represents the Designated Bridge and Designated Port tiebreakers. Subtracting accumulating cost from zero to get the priority adds meaning to the preference given to lower values of cost and priority components.

Within an MST region, the 'CI' tree adds an 'I cost' component:

$Priority = 0 - C - (I_M + I_P) - D$

Where I_M is the 'Master Bridge' for the region, and I_P is the Internal Path Cost within the region from the Master Bridge. The Master Bridge is the first bridge (proceeding down the CI spanning tree) in the region, and is of course chosen so as to maximize the priority of the master bridge. The Internal Path Cost is computed in the same way as a common spanning tree path cost, commencing with zero at the master bridge. In an MST region where the spanning tree information has stabilized, $C+I_M$ is constant.

It is useful to think of the entire CI tree priority, including the I components as existing in the SST region as well. Since the C component is incremented at the Root Port of each Bridge in the SST region, the I components are not required to ascertain relative priorities in the SST region. An MST bridge that receives a BPDU without the I component on its Root Port must add to the C cost to ensure that the information priority continually decreases away from the C Root Bridge. It can then set I_M to its own bridge identifier and I_P to zero. A CI (MST) bridge that receives a BPDU with an I component on its Root Port simply adds its Root Port Cost to I_P. An SST bridge that receives a BPDU with an I component will ignore that component and add to the C cost. Thus BPDUs proceed from the C Root Bridge labeling each LAN with ever

decreasing priority, and the common spanning tree can be predictably constructed for the entire network.

An M tree has:

$Priority = 0 - C - I_M - (M_R + M_P) - D$

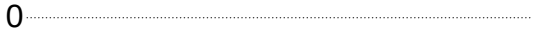
However $C+I_M$ is constant within a stable MST region, so a stable M tree is effectively rooted at the highest priority bridge within the region.

An M bridge that receives a BPDU without an I or M component adds to the C cost and sets M_R and M_P to its own bridge identifier and zero respectively. An M bridge that receives a BPDU with I and M components and a higher priority M_R adds to the M_P cost. Port roles are assigned to M bridge ports in the usual way except that the Root Port for the MST region is assigned a Designated Port Role.

Visualizing Trees

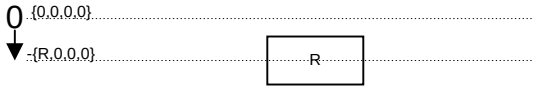
We can picture the calculation of a spanning tree as follows.

The highest possible priority in our priority scheme is 0 (zero).

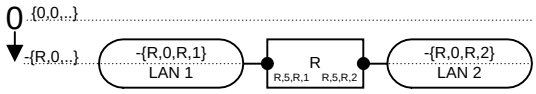


Or perhaps more properly, distinguishing the Root Bridge, Root Path Cost, and Designated Bridge and Port Components, the priority vector $\{C_R, C_P, D_B, D_P\}$ is $\{0,0,0,0\}$.

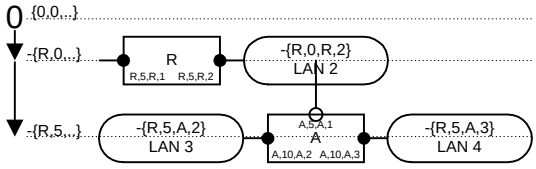
The bridge that will be chosen as the Root Bridge is the bridge that can advertise the highest priority on all its Designated Ports, i.e. has the lowest possible version of C (above).



More specifically, showing R as having two ports attached to two LANs.



The rules for the subtraction of priority vectors become more apparent as we consider a further bridge A, attached to one of the LANs that has R as Designated Bridge.



The priority vectors advertised by A on each of its two LANs are the best, i.e. highest, that can be derived from the priority vector information received at A's ports as follows.

If the received priority vector is

$$\{C_R, C_P, D_B, D_P\}$$

the priority vector established by local mangement parameters (i.e. Root Port Path Cost etc.) for the port on which this information has been taken in is

$$\{C_{Rin}, C_{Pin}, D_{Bin}, D_{Pin}\}$$

the priority vector established for the port on which this information is to be put out is

$$\{C_{Rout}, C_{Pout}, D_{Bout}, D_{Pout}\}$$

and the resulting advertised vector is

$$\{C'_R, C'_P, D'_B, D'_P\}$$

Then

$$\{C'_R, C'_P, D'_B, D'_P\} = \{C_R, C_P, D_B, D_P\} - \{C_{Rin}, C_{Pin}, D_{Bin}, D_{Pin}\} - \{C_{Rout}, C_{Pout}, D_{Bout}, D_{Pout}\}$$

Where the 'subtraction' of the vector components proceeds as follows:

$C'_R = C_R$ if $C'_R > C_{Rin}$ else $C'_R = C_{Rin}$

$C'_P = C_P - C_{Pin}$ if $C'_R > C_{Rin}$ else $C'_P = 0$

$D'_B = D_{Bout}$

$D'_P = D_{Pout}$

Thus the priority vector advertised at any port on A is always strictly less than the best information received by A[1]. Every LAN has exactly one best advertiser of spanning tree information for that LAN, and that advertiser is in receipt of strictly higher priority information from the Root Bridge. This ensures the construction of a loop free fully connected active topology. In the following fragment from our example Bridged Local Area Network, Bridge B receives its best priority information from LAN 1, but is not a better Designated Bridge for LAN 3. This is simply shown by placing B lower in our diagram of the projected priority vector.

<insert diagram>

Looked at from the point of view of the priority components an MST region will appear to be all at the same level in our diagrams. This is not surprising – the entire region is after all meant to appear to be a single bridge – but very concerning as there is no way to ensure that the tree calculation information will be correctly propagated through the region. With Rapid Spanning Tree there is no way to ensure that the MST region itself converges faster than the surrounding SST. However the common spanning tree can be propagated through the MST region with the addition of the 'I cost' components introduced above.

Of course at the boundary of an MST region, the 'I cost' may be unspecified. The MST bridge could be receiving priority information from an SST bridge. Changes in the available network components may change the boundary of the

[1] If A becomes the Root Bridge it can be considered to have received information from itself.

MST, so we have to be prepared for all the eventualities.

A suitable specification of our priority vector subtraction rules is all that is required. If the full vector is:

$$\{C_R, C_P, I_M, I_P, D_B, D_P\}$$

An MST bridge receiving information from an SST bridge considers itself to have received:

$$\{C_R, C_P, ?, ?, D_B, D_P\}$$

i.e. unspecified I_M and I_P components.

If, as before, the priority vector established by local management parameters (i.e. Root Port Path Cost etc.) for the port on which information has been taken in is

$$\{C_{Rin}, C_{Pin}, I_{Min}, I_{Pin}, D_{Bin}, D_{Pin}\}$$

the priority vector established for the port on which this information is to be put out is

$$\{C_{Rout}, C_{Pout}, I_{Mout}, I_{Pout}, D_{Bout}, D_{Pout}\}$$

The resulting advertised vector

$$\{C'_R, C'_P, I'_M, I'_P, D'_B, D'_P\} = \{C_R, C_P, I_M, I_P, D_B, D_P\} - \{C_{Rin}, C_{Pin}, I_{Min}, I_{Pin}, D_{Bin}, D_{Pin}\} - \{C_{Rout}, C_{Pout}, I_{Mout}, I_{Pout}, D_{Bout}, D_{Pout}\}$$

The ‘subtraction’ of the vector components proceeds as follows:

$C'_R = C_R$ if $C'_R > C_{Rin}$ else $C'_R = C_{Rin}$

$C'_P = C_P$ if $C'_R > C_{Rin}$ and I'_M, I'_P != ? else

$C'_P = C_P - C_{Pin}$ if $C'_R > C_{Rin}$ and I'_M, I'_P == ? else

$C'_P = 0$

$D'_B = D_{Bout}$ if I'_M, I'_P == ? else

$D'_B = D_{Bin}$

$D'_P = D_{Pout}$ if I'_M, I'_P == ? else

$D'_P = D_{Pin}$

Interoperability

To ensure interoperability with SST and MST bridges, an MST bridge can transmit the priority information in both BPDU formats, or can make the assumption that an SST bridge will only read sufficient of the format to establish the usual priority components, not just the I cost components. An MST bridge should, of course prefer BPDUs from an MST bridge that carry the I cost components, since it can then propagate higher priority information. The effect of the above rules is such as to allow continuous operation of a single spanning tree across the entire bridged network, while making MST regions appear to the surrounding SST bridges just as if they are single bridges.