Spanning Vines

Mick Seaman

This note describes a fundamental improvement to MSTP (as specified in P802.1s D11). Multiple active connections, each carrying frames assigned to different VLANs, are supported between MST Regions. Fortunately this is possible using the concepts already developed and almost all of the existing Clause 13 text – a considerable personal relief. I am sure all those who spent time reviewing the work so far will agree.

The enhanced functionality can simplify both the implementation and administration of MST Bridges. Its description begins by considering a region comprising a single MST Bridge, possibly in a wiring closet. An MSTI's Master Port is allowed to be independent of the CIST Root Port, thus permitting the simultaneous use of alternate uplinks.

Simultaneous use of alternate uplinks to carry traffic assigned to different VLANs is a common requirement that may account for the majority of decisions to deploy multiple spanning trees. Since only CST information is communicated across region boundaries, the proposed improvement leads to the surprising observation that it is possible to meet the requirement with existing single spanning tree protocol. The multiple spanning trees, if we can call them that, are purely internal to the wiring closet bridge.

Fortunately this interesting observation, though it may simplify the administration of many networks, is not the end of either the multiple spanning tree protocol or the proposed improvement. We can apply the single bridge model of an MST Region (in reverse) to substitute an entire Region for the single MST Bridge considered initially. Thus we can show how entire networks of bridges can be connected by multiple active links. Those networks do not have to have identical assignments of VLANs to MSTIs, the same or the same number of MSTIs, or even operate the same protocols for routing frames assigned to VLANs. This modeling exercise clarifies the requirements for the successful interconnection of such networks, pointing the way to an interface service definition.

It has to be said that the proposed improvement and its consequences do not suggest that all MST Regions should be made as small as possible. Partitioning a region discards overall routing optimization in favor of local routing and boundary policy decisions, with a degree of implementation and administrative freedom. Sometimes that tradeoff is the right one. The important thing is to have the choice.

Essential Background

This note assumes that the reader is familiar with the calculation and visualization of spanning trees as previously discussed in the note "Rational Trees" (Mick Seaman, 11th July 2000) http://www.ieee802.org/1/mirror/8021/docs2000/RationalTrees009.pdf user: p8021 password: go_wildcats. The discussion of spanning tree connectivity follows the notation used in the recently distributed P802.1sD11.1 http://www.ieee802.org/1/mirror/8021/s-drafts/d11-1/802-1s-d11-1.pdf . Clause 13.16 of that document may prove helpful, if not strictly required.

A Single Bridge with Multiple Uplinks

The spanning tree priority information received by any Bridge, B, on one of its Alternate Ports, B_A, is better than that advertised by the Bridge on all of its Designated Ports, B_{D1}, B_{D2}, .. . Thus the LAN, N, connected to by B_A is not in the subtree, S_B, that is connected through B_{D*} to the rest of the Bridged Local Area Network by B's Root Port B_R.

Since the spanning tree is "spanning", i.e. fully connects all LANs, N is connected to all the LANs not in S_B by bridges other than B. Since the spanning tree is "tree", i.e. simply connects all LANs, no LAN in S_B is connected to any LAN not in S_B by any Bridge other than B. Hence the substitution of forwarding through B_A for

forwarding through B_R preserves the spanning and tree attributes of the active topology. Figure 1 below illustrates this point.

This, in a nutshell, is the logic that was used in 802.1w to allow rapid reconfiguration by immediately substituting forwarding through an alternate port for forwarding through a failed root port. We can use this to substitute forwarding through an alternate port for a chosen set of VLANs for forwarding those particular VLANs through *B*'s root port. It should be clear from the above that the choice of the set of VLANs to be forwarded does not depend on any agreement between *B* and any other bridge[1].

Duh!

The first question that springs to mind is "why didn't we think of this before?". There are perhaps three reasons:

1) While there have been previous attempts to use "cross-links" in a single spanning tree to optimize traffic routing and load share, these were developed in a pre-VLAN era where not only the routing of frames but learning from source addresses needed to be considered. John Hart's Distributed Load Sharing (DLS) (US patent 4,811,337) restricts the use of the "cross-link" to communication between stations that are further away from the root than both the bridges that are providing the "cross-link". This is significantly harder to arrange than having a single bridge unilaterally decide to use an otherwise unused link (an Alternate Port) for some traffic. The existence of VLANs allows a bridge to route traffic for some VLANs independently of others (subject to there being no filtering database learning coupling between the VLANs – a constraint we already recognize for multiple spanning trees).
2) A lot of the work on multiple trees has concentrated on controlling individual trees by having those trees not reach certain LANs in the topology at all. This note, and P802.1s, takes a different approach. Each tree is spanning in its own right, and traffic suppression is achieved through the automatic learning of end station locations and (where desired) through explicit pruning of traffic from a tree by GVRP. This approach is far less error prone than manual configuration of egress lists and allows a plug and play approach. Most significantly for this note, the introduction of multiple trees by manual pruning of the topology from the root out seems to have completely obscured the opportunity for simplification based on active topologies that are spanning over the entire network for all VLANs[2].

   In other words, past approaches to multiple spanning trees have mainly focused on not transmitting frames from the direction of the root on to links where they are not required. This note focuses on not receiving frames from those links and relies on source address learning to suppress unnecessary traffic. In this approach we prune from the edge of the network in, not from the center out.
3) This approach, while of considerable utility in structured campus wiring scenarios -- where the requirement is to make use of two or more "uplinks" which may have been provided for redundancy -- is of limited use by itself in rings and more arbitrary topologies.

Choosing Alternate Links

While any Alternate Port can be chosen, at least in principle, in preference to a Bridge's Root Port it is desirable that we retain predictability and manageability of the choice, and provide a model and terminology for what happens.

It seems useful to borrow terminology and parameters from P802.1s. A set of MSTI (multiple spanning tree instance) parameters can be associated with the VLANs that are to be routed separately from the normal spanning tree. The selected Alternate Port becomes the Master Port for the MSTI and hence for the assigned VLANs, and it is selected by adding the port path cost for the MSTI to the received root path cost for the single spanning tree, choosing the port with the lowest resulting cost as usual.

This MSTI fiction makes it particularly easy to extend the model to true multiple spanning trees where the single spanning tree becomes the CIST and the single bridge and MST Region. However it should be clear that there is only a single spanning tree being constructed in this single bridge scenario, and that it could be deployed with RSTP (given learning independence between the VLANs that use the alternate paths in the network).

Note also that there are other potentially useful models for assigning VLANs to Alternate Ports, or assigning Master Ports for VLAN sets which is just another way of expressing the same thing.

One way would be to implement a best fit algorithm between the expected bandwidth on each VLAN and the bandwidths of the Root Port and potential Master Ports. In fact a wide range of local policies could be invented. At the extreme additional information could be added per VLAN to the single spanning tree to express resource consumption from the root. This

[1] However all the Bridges in the network have to use an independent Forwarding Database (FID) for each set of VLANs that is to be separately routed.

[2] P802.1s deals with the excess protocol traffic (an obvious and hence overriding customer consideration) of having all VLANs span the network, by packing all the trees into a single BPDU.

however is close to providing all the information for a multiple tree protocol – though possibly more useful in some circumstances. It is not the purpose of this note to detail some of the arcane policy mechanisms we might invent however, but to explain the simple opportunity. The important thing is that the policies only attempt to use the single spanning tree's Root Port or an Alternate Port as the Master Port for each specific set of VLANs, and that each VLAN set only has one Master Port plus the Designated Ports of the spanning tree as forwarding ports. These simple rules ensure full ("spanning") and loop-free ("tree") connectivity for the VLAN set.

Use in MST Regions

The above gives us the basis for freeing Master Port assignment for a particular MSTI from the Root Port assignment on the CIST Regional Root. In principle (given sufficient protocol inside an MST Region) the CIST Regional Root, the MSTI Regional Root, and the MSTI Master Bridge (the MST Bridge at the Region Boundary that has the MSTI Master Port for the Region) can all be independent. However complete independence of all three gives us an extra Bridge Identifier to carry around in the protocol, together with the requirement for additional communication to and fro to propagate the best choices. This additional communication would further complicate the protocol, and extend convergence times.

I believe that the practical requirements of .1s can be met by allowing the manageable choice, independently in each MST Region and for each MSTI, between two constraints:

1) The MSTI Master Port is a port on the CIST Regional Root, though not necessarily the CIST Root Port. The MSTI Regional Root can be anywhere in the MST Region. This constraint is the only option specified by P802.1sD11.1
2) The MSTI Master Port is a port on the MSTI Regional Root, which is constrained to be at the Boundary of the MST Region.

The second constraint could be implemented by only allowing MST Bridges that have Boundary CIST Alternate Ports to compete for election as MSTI Regional Roots. However that would allow perturbations in the CST external to the Region to arrange the internal details of the Region – in direct contravention of one of our goals. A better suggestion is to have the MSTI Regional Root remain unchanged, but signal in its Configuration Messages whether it has a CIST Alternate Port at the Region Boundary. If it does not, the bridge that is the CIST Regional Root provides the Master Port for the MSTI. That neatly covers both constraints in one mechanism.

In the Region that contains the CIST Root, and is therefore at the center of the network, each MSTIs Regional Root will commonly be required to be different from the CIST Root and CIST Regional Root (these last two roles being performed by one and the same Bridge). Since there is no upper boundary (no CST Root Port) for the Region, no MSTI Master Port is required and this important case is easily covered by the implementation suggested above.

Figures and Examples

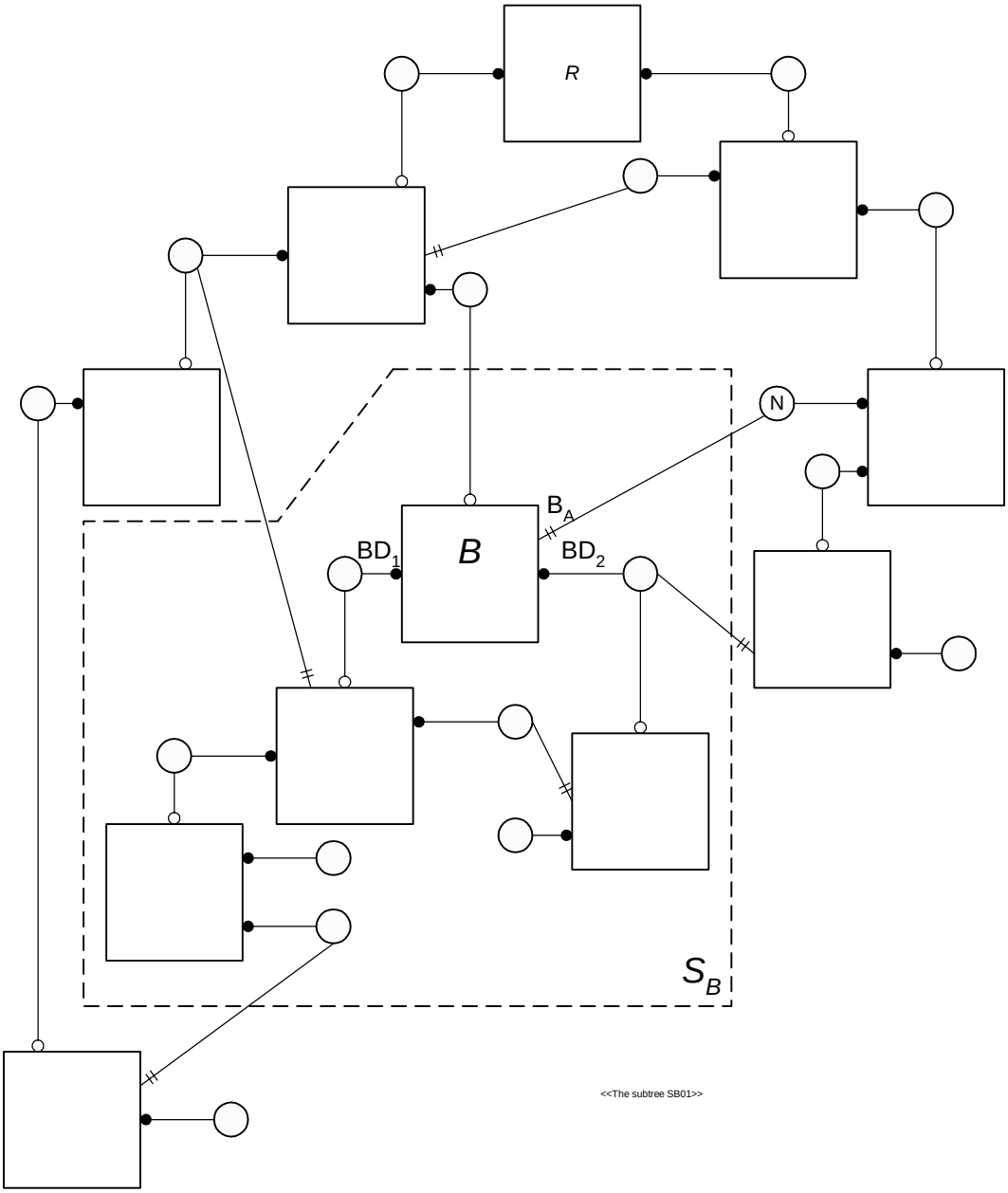

Figure 1 – Spanning tree connectivity of an example subtree S_B

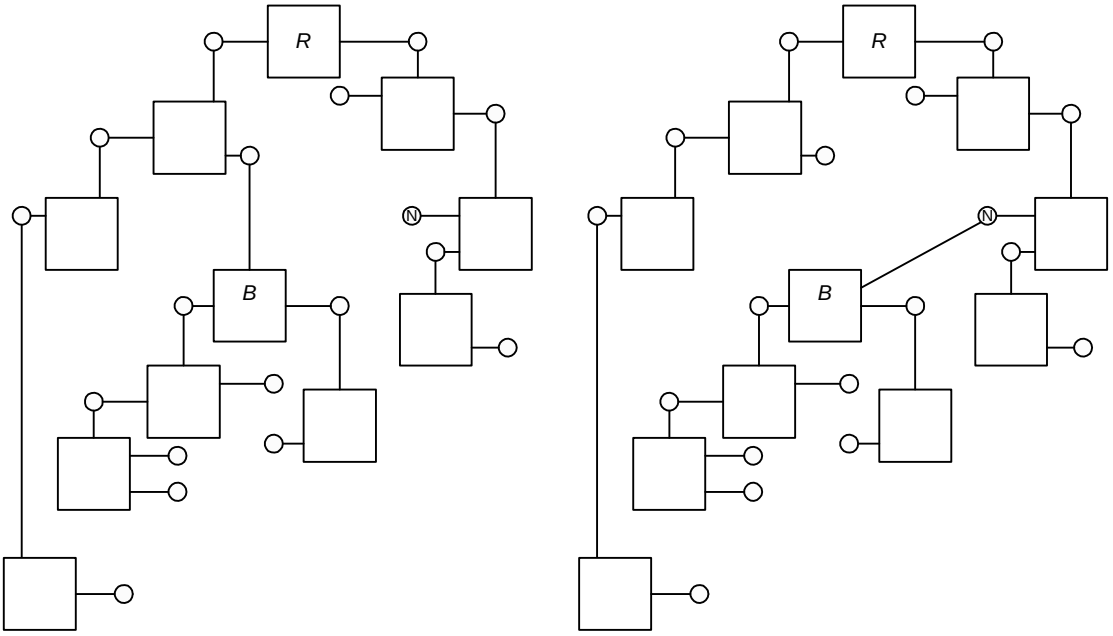

Figure 2 – Spanning tree active topology and an alternative active topology chosen by B for certain VLANs

NOTE – B and some or all of the other Bridges may implement STP or RSTP

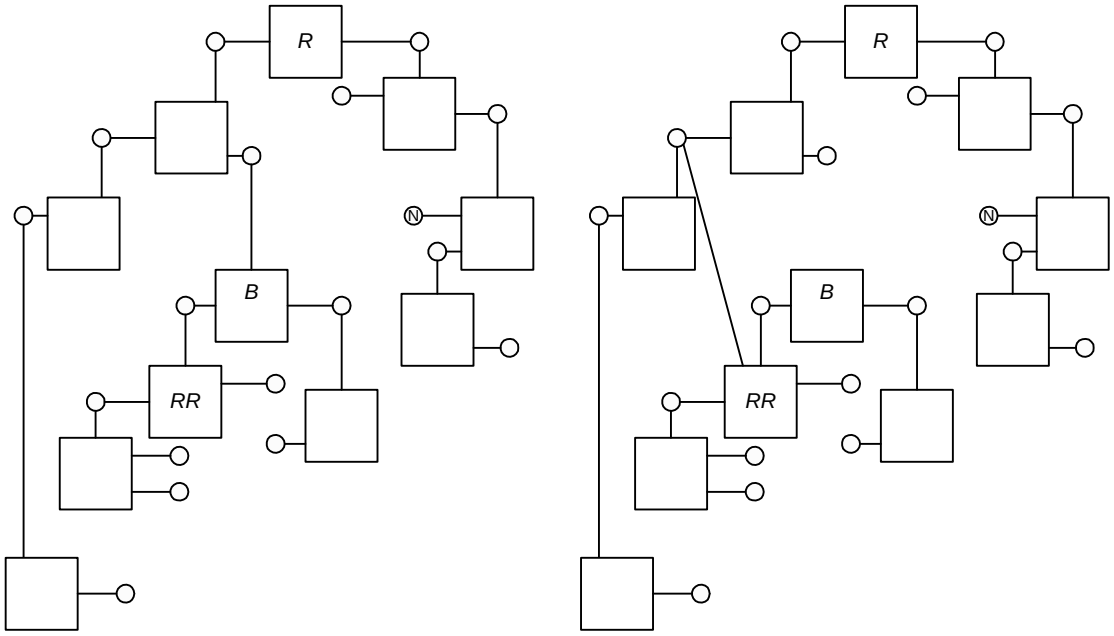

Figure 3 – Spanning tree active topology and an alternative active topology chosen by RR and B for certain VLANs

NOTE – RR is the Regional Root for an MSTI, RR and B implement an MSTP like protocol, and are in the same MST Region, some or all of the other Bridges may implement STP or RSTP

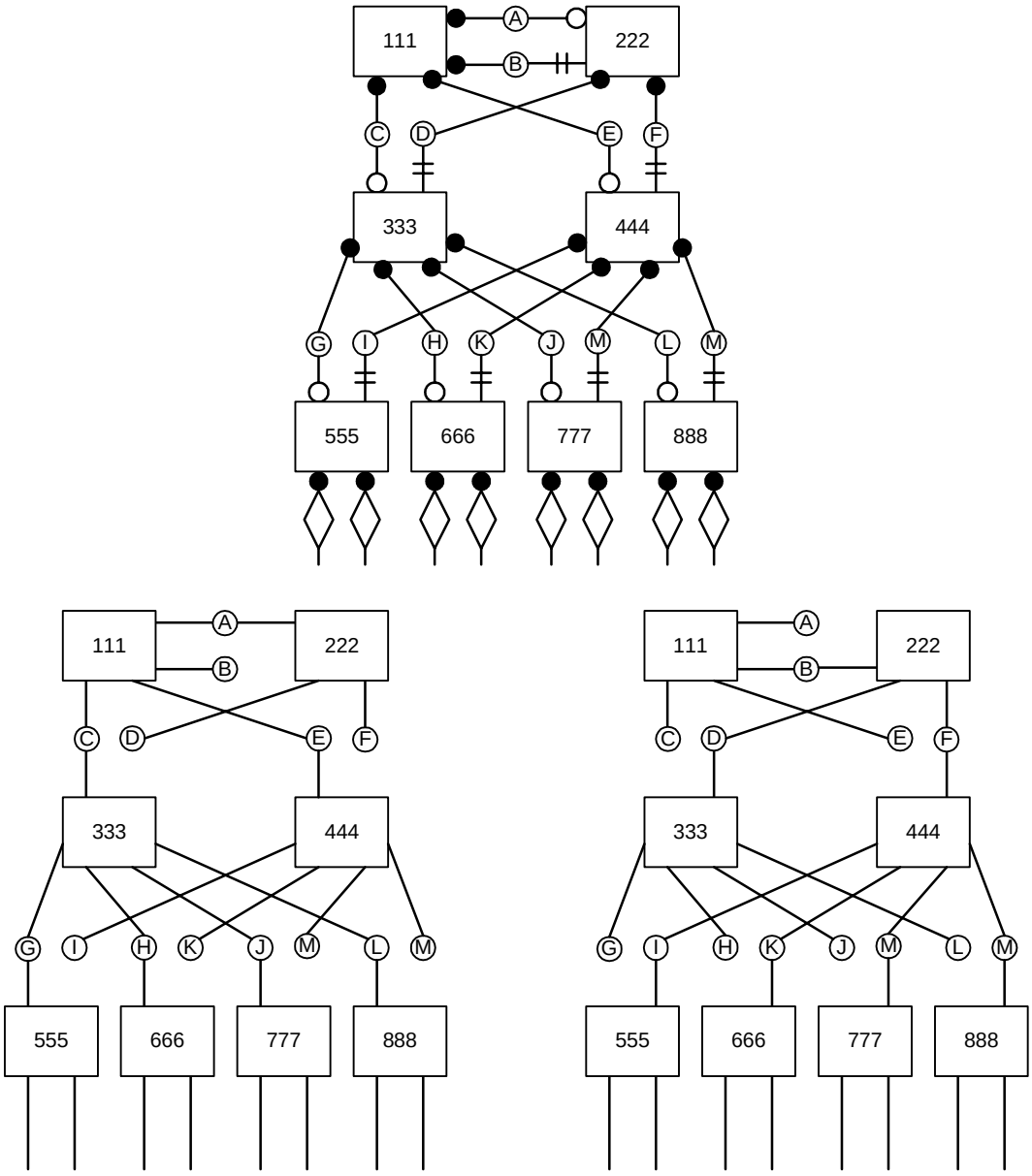

Figure 4 – Spanning tree port roles, active topology, and alternate active topology chosen by Bridges 333 through 888 for certain VLANs

NOTE – The active topologies shown mimic dual spanning trees, one rooted at 111 and the other at 222.