 # 131
TECHNOLOGY WHITE PAPER

Spanning Tree:
Death is Not an Option

Nick Slabakov, Riverstone Networks

ABSTRACT

This article is an overview of the progression of both the spanning tree algorithm and recent modifications that have made it an important technology for today's metro networks.

Spanning Tree is a protocol with a life story. Many years ago, long before the creation of the world (in Internet terms, that is), the protocol was born. Its IEEE parents named it 802.1D, and taught it one important lesson – how to resolve bridging loops in enterprise networks. 802.1D learned this lesson well and quickly impressed the adults who were building small- to medium-sized enterprise networks.

Unfortunately, 802.1D's early success was followed by a rough adolescence. The industry's focus moved to Service Provider Networks, where 802.1D had a hard time fitting in. 802.1D kept hearing that it was too slow, non-scaleable, and clumsy (and those were just the polite comments).

It was the recent advent of the Metro Ethernet Networks that redeemed 802.1D and its newly born sisters, 802.1w and 802.1s. Today, carrier-focused cousins have arrived with names like PVST, MSTP, and Ring STP. They promise to be as important as the original spanning tree. This article tells the story of that evolution.

SPANNING TREE: THE ORIGINAL

Spanning tree was designed to solve the fundamental problem of traffic loops created by the interconnection of LANs with redundant transparent bridges.

The Problem

The core of the looping problem is the "Learning" quality that transparent bridges have – the way that they know how to forward traffic between their ports is by snooping Ethernet frames, and recording (or learning) which MAC address resides on which port of the bridge. Then, when a frame arrives for a given MAC address, the bridge "knows" on which outgoing port to send it. If a frame arrives and its destination MAC address is unknown to the bridge, it will "flood" the frame on all of its ports. That's it.

Bridging loops (and the broadcast storms they create) were quickly recognized as the worst thing that could happen to a bridged network, and a robust solution to the problem was developed.

The Solution

Spanning Tree (STP) solves the problem by removing (or pruning) all redundant paths. It reduces the topology to a tree structure that guarantees complete connectivity (that's why it is a "spanning" tree). The algorithm accomplishes this by selecting a Root Bridge, and causing every other bridge in the topology to select a Root Port, which is a port that leads to the Root Bridge with the least cost. On each LAN segment, there will also be a bridge with a Designated Port, whose job is to forward frames on behalf of that LAN segment.

These port roles are determined by use of BPDUs (Bridge Protocol Data Units), which are originated by every bridge and sent to the neighboring bridges. Using the information in the BPDUs bridges allows the system to figure out the roles all of its ports will play.

Once the roles are clear (which happens fairly quickly, in the neighborhood of 1-2 seconds), the ports that are neither Root Ports nor Designated Ports are placed in Blocking State (i.e. they will not be listening or forwarding frames). The Root and Designated ports then begin a deliberately lengthy (30 seconds typically) process to prepare for going into Forwarding State.

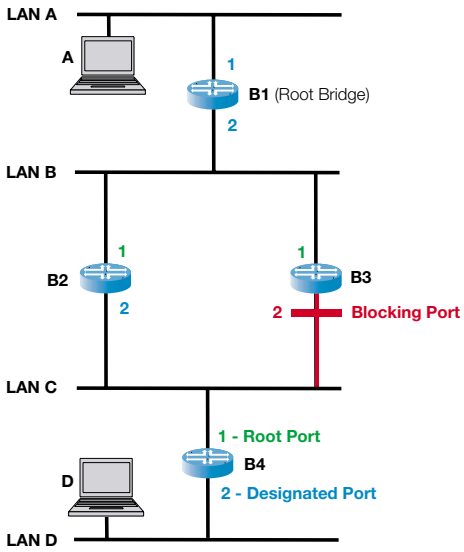

Figure 1: Basic Spanning Tree port states

Figure 1 illustrates the terms we have used so far. If we suppose that B1 is selected as a Root Bridge, then all of its ports are Designated Ports. Port 1 of all remaining bridges will be selected as a Root Port, as it is the port that leads to the Root Bridge using the cheapest path. Port 2 of the remaining bridges (except for B3) is selected as Designated Port. Port 2 on B3 is placed in Blocking State in order to remove the redundant link between LAN B and LAN C.

Ports that end up in Forwarding State forward frames; ports that are placed in Blocking State drop them. If the topology changes, ports that were in Blocking State may go into Forwarding, and vice versa.

The Problems with the Solution

There are really three problems with the way the standard STP algorithm, as defined in IEEE 802.1D, behaves.

Convergence Speed

Upon topology change, STP generally takes between 30 and 60 seconds to converge. The reason for this is that the protocol aims to ensure that "sufficient" time elapses between the moment of a topological change and the moment of enabling an alternate port to perform forwarding. This "sufficient" time guarantees that a port will not go in Forwarding State before all other ports that must be in Blocking State are in Blocking State.

The effects of this behavior can be seen on Figure 1. If Port 2 of B2 goes down, traffic between workstations A and D will stop for around 30 seconds. This is because Port 2 on B3, even though redundant, will go into 30 seconds of Listening/Learning before it goes into Forwarding State. That way B3 ensures there is no other bridge on the network that may be a potential candidate for forwarding.

As we mentioned earlier, nothing is worse in a bridged network than a Layer 2 loop, and the way STP ensures that it will not create one upon a topological change is by use of such conservative timers. Unfortunately, those timers lead to very long overall convergence times.

This problem is addressed by the newly developed standard – 802.1w (Rapid Reconfiguration), also referred to as Rapid Spanning Tree Protocol (RSTP).

VLAN Insensitivity

When 802.1Q-capable switches are involved (and most switches today are 802.1Q-capable), a problem with the standard STP arises. It can be seen in scenarios where asymmetrical connectivity between VLANs exists.

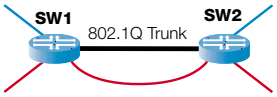

Figure 2: Asymmetrically connected VLANs

For example, VLAN "Blue" on Figure 2 is only connected via the 802.1Q trunk that links the two switches. VLAN "Red", on the other hand, is connected via the trunk, as well as through the additional link between the two switches.

There is an obvious loop in VLAN Red, comprised of the trunk and the additional "Red" link between the switches. If regular STP was used to break it, then it will choose one port to be blocked. That port can quite possibly be the Trunk port. If that happens, connectivity will be broken for VLAN Blue, even though there is no loop in VLAN Blue.

Various enhancements to Spanning Tree were done in order to improve its VLAN knowledge.

Link Blockage (Inefficient Use of Bandwidth)

This problem is particularly evident in ring topologies, such as the one shown on Figure 3.

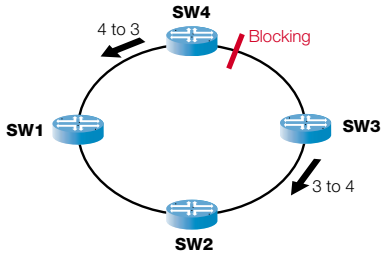

Figure 3: Effect of Spanning Tree on a ring topology

Many Metro Ethernet providers build their networks in rings, as the fiber-route-miles are minimized this way. Unfortunately, the ring topology (when used to run Ethernet on top of it) always makes some of the communications inefficient. In the example of Figure 3, if the link between SW3 and SW4 is in blocking state, then traffic from SW4 to SW3 and vice-versa will always have to traverse the whole ring.

RAPID SPANNING TREE PROTOCOL (RSTP)

RSTP, recently standardized by the IEEE as 802.1w, provides significant improvements in the speed of convergence for bridged networks. Those improvements come from the ability of switches to distinguish point-to-point vs. shared links. Point-to-point links are those that connect exactly two switch ports, while shared links are ones that have more than two devices attached to them.

In regular STP, switch ports have three "port roles", which we defined earlier:

- Root port
- Designated port
- Disabled port

RSTP adds two more port roles:

- Alternate port – a port that can take over for a root port if it fails
- Backup port – a port that can take over for a designated port if it fails

Upon failure of a root port, an RSTP switch can "promote" an Alternate port to a Root port state immediately, instead of having to wait for the usual Listening/Learning sequence a regular STP switch will do. When a Designated port fails (provided it is part of a point-to-point link), a Backup port will be promoted to a Designated port just as fast.

In summary, RSTP reduces the convergence time significantly (to a 2-3 second range) for failures that involve point-to-point links. It provides no appreciable benefits for failures of "shared LANs" – network segments with more than two switches on them. Since many of the Metro networks built today are designed with point-to-point links, RSTP provides significant improvement in their convergence characteristics.

VLAN – SENSITIVE SPANNING TREES

When VLANs were primarily used in the Enterprise space, they were mostly symmetrically connected; that is, all VLANs had their redundant links along the same physical paths. It was quite adequate to run a single STP process on the whole infrastructure – there was no danger of blocking a port that is a redundant port for one VLAN but the only possible path for another.

With the advent of Metro Ethernet networks, VLANs became the product that customers purchased, rather than a method for the service provider to deliver service. In other words, the service provider is selling VLANs. Since different customers had different connectivity and redundancy requirements, Metro Ethernet providers quickly ended up with VLANs that were using different underlying physical links (something simplistically illustrated on Figure 2).

This asymmetrical connectivity quickly generated demand for the infrastructure to be able to execute separate STP processes for different VLANs, which result in the ability to have a given physical port perform forwarding for one VLAN while doing blocking for another.

Per-VLAN Spanning Tree (PVST) and Multiple-VLAN Spanning Tree (MVST)

PVST is the simplest implementation of the VLAN-sensitive approach. It relies on unique BPDUs transmitted by each switch for every VLAN, and on a separate Spanning Tree process running on every switch for every VLAN. The BPDUs that are transmitted are proprietary, and no vendor interoperability exists today for PVST. In addition, PVST suffers from two obvious scalability deficiencies:

- BPDU traffic is proportional to the number of VLANs supported, and
- The CPU of the switches is seriously affected by the number of the separate STP processes run by the switch

The second concern is exacerbated in situations where an 802.1Q trunk port fails, causing the STP processes of every VLAN that participates in the port to re-calculate at the same time.

These scalability problems with PVST are somewhat addressed by switches that implement an extension of the technology, known as MVST, which allows STP processes to be created and VLANs to be added to them. A provider could group similarly connected VLANs into STP processes, thereby being able to take advantage of multiple STP processes yet not suffer from having an excessive number of them.

802.1s

IEEE is in the process of standardizing the various proprietary VLAN STP approaches under the 802.1s standard. In addition to formalizing the BPDUs used, the standard also defines the interactions between areas of a network that are capable of supporting multiple STP instances (MST regions), and others that only support an STP instance (SST regions).

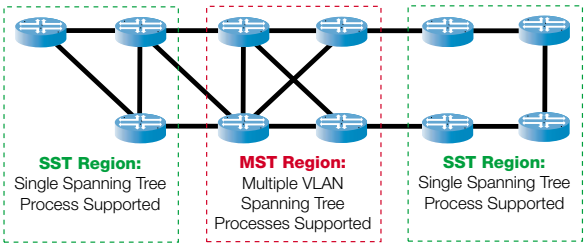

Figure 4: A network is divided into regions based on their STP capabilities

It accomplishes that by making the MST region pose to the neighboring SST regions as a single Spanning Tree process. This allows for seamless interoperability of areas of the network that are not capable, or do not need to support Multiple Spanning Tree processes with others that do.

RING SPANNING TREE

It is a well-known fact that fiber is most often laid in ring topologies to optimize fiber route-miles. The only MAN/WAN technology well suited to handle rings today is SONET. However SONET is optimized for voice applications, and is commonly considered wasteful and expensive when most of the traffic on the ring is data.

While the long-term marriage between the simplicity of Ethernet and the reliability of SONET will most likely occur when the IEEE 802.17 standard for Resilient Packet Rings is developed, Metro Ethernet Providers and vendors alike are in search of a short-term solution, geared toward Metro services that are:

- Predominantly data-oriented and do not necessarily require the 50 ms failover typical for SONET
- Suitable for "Rings of Ethernets", or in other words, ring topologies comprised of switched Ethernet segments
- Available today

Such a solution is Ring Spanning Tree (Ring STP). It makes some modifications to the Rapid Spanning Tree Protocol (RSTP) to take advantage of the simpler and exclusively point-to-point nature of the segments comprising the ring.

To take advantage of Ring STP, the topology must be strictly a concatenation of rings, with each ring identified by a unique Ring ID. Spanning Tree BPDUs are not propagated outside the boundaries of a ring, and as a result, separate Spanning Tree processes run in each ring.

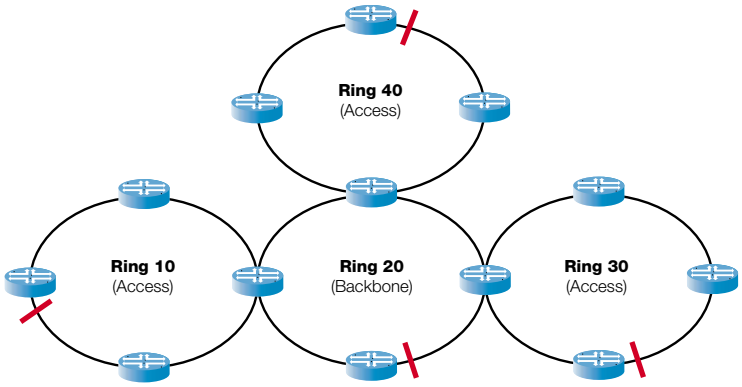

Figure 5: Ring Spanning Tree makes blocking decisions for each ring individually

Since the topology within the ring is very simple and the links are strictly point-to-point, Ring STP has very fast (sub-second) convergence properties. It also allows scaling of Layer 2 networks without having to worry about scaling Spanning Tree. Of course, the other fundamental scalability issues with Layer 2 networks, such as MAC address learning capacities, lack of summarization, and flooding, still remain.

STP TUNNELING

Many Metro Ethernet providers today provide VLAN services for their customers. They do so in one of three ways:

- Trunking customer VLANs through the provider's switched infrastructure
- Stacking customer VLANs on top of provider's "carrier" VLANs, using the "Stackable VLAN" features of the provider's switches
- Utilizing MPLS tunneling techniques, such as the one specified in the "Martini draft" (draft-martini-l2circuit-trans-mpls-07.txt)

In all of these cases, it is imperative to present the service provider infrastructure as a "Pseudo-Wire", which is transparent to the customer STP processes. Furthermore, the service provider's network will most likely run its own instance of Spanning Tree, whose topology calculations should be independent of those of any customer.

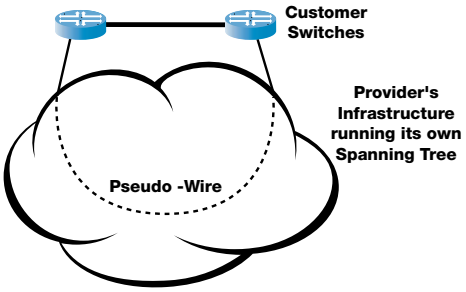

Figure 6: The provider's infrastructure is presented as Pseudo-Wire

As illustrated on Figure 6 above, if a customer is purchasing VLAN service from a provider and has two sites (represented by two switches), a Layer 2 loop can be created by connecting the sites with redundant links – one over the provider's network, and another one via physical connection. It is important that the Pseudo-Wire behaves in the same way as the physical one, which allows the customer switches to correctly calculate Spanning Tree and block the appropriate port(s) to break the loop. In order for that to happen, two things are necessary:

- The BPDUs originated by a customer switch must be encapsulated and transported through the provider infrastructure to the other site
- The provider switch infrastructure, which possibly runs its own STP process, must NOT interpret or react to those customer BPDUs, and should treat them as mere data

Unfortunately, the methods of BPDU tunneling today are completely proprietary, and no interoperability exists between different vendor's switches in that respect. Additional complexity is created by handling BPDU tunneling for Point-to-Multipoint environments (only a Point-to-Point is shown above). For those reasons, a provider is well advised to study the BPDU tunneling details of the switches it deploys if VLAN services are going to be offered to customers.

CONCLUSION

Layer 2 service offerings are commonplace today. They are attractive because they are simple, protocol-agnostic, and provide transport similar to Frame Relay and ATM at a fraction of the cost. These services are delivered under different names, and with a different level of technological sophistication beneath them. They range from extending simple VLANs across the provider's infrastructure, to building complex, MPLS-based infrastructure with Point-to-Multipoint functionality.

As long as these Layer 2 services proliferate, Spanning Tree and its derivative protocols, discussed in this paper, will play vital roles in the architecture. While not all of the original Spanning Tree deficiencies have been resolved by the newer protocol developments (particularly the link blockage and the inefficient bandwidth usage), many other aspects have been improved dramatically (convergence speed, VLAN awareness, etc.). Keeping up with the old Spanning Tree has become, to a surprising extent, an important technical skill of today.

Acronyms

ACL	*Access Control List*
ANSI	*American National Standards Institute*
ASIC	*Application-Specific Integrated Circuit*
ASP	*Application Service Provider*
ATM	*Asynchronous Transfer Mode*
BPDU	*Bridge Protocol Data Units*
CBR	*Constant Bit Rate*
CWDM	*Coarse Wave Division Multiplexing*
DS1/DS3	*Digital Signal, Level 1 (1.54 Mbps) or 3 (44.7 Mbps)*
DSL	*Digital Subscriber Line*
DWDM	*Dense Wave Division Multiplexing*
DVMRP	*Distance Vector Multicast Protocol*
E1/E2	*European Trunk 1/2 (2 Mbps/34.3 Mbps)*
ERP	*Enterprise Resource Planning*
HSSI	*High Speed Serial Interface*
ISP	*Internet Service Provider*
ITU	*International Telecommunications Union*
LAN	*Local Area Network*
LEC	*Local Exchange Carrier*
MAC	*Media Access Control*
MAN	*Metropolitan Area Network*
MDU	*Multiple Dwelling Unit*
MLPPP	*Multi Layer Point-to-Point Protocol*
MPLS	*Multiple Protocol Label Switching – See "MPLS in Metro IP Networks," http://www.riverstonenet.com/technology/mpls.shtml*
MSTP	*Multiple Spanning Tree Protocol*
MTU	*Multiple Tenant Unit*
MVST	*Multiple-VLAN Spanning Tree*
OC-3/OC-12	*Optical Carrier 3/12 (155 Mbps/622 Mbps)*
PDH	*Plesiochronous Digital Hierarchy*
PIM	*Protocol Independent Multicast*
POS	*Packet over SONET*
PPP	*Point-to-Point Protocol*
PVC	*Private Virtual Circuit*
PVST	*Per-VLAN Spanning Tree*
QoS	*Quality of Service*
RED	*Random Early Discard*
RSTP	*Rapid Spanning Tree Protocol*
SONET	*Synchronous Optical NETwork – See http://www.techguide.com/comm/sec_html/sonet.shtml*
SLA	*Service Level Agreement*
SPE	*Synchronous Payload Envelope*
SRP	*Spatial Reuse Protocol – See RFC 2892*
STP	*Spanning Tree Protocol*
T1	*Trunk 1 (1.544 Mbps)*
TCP/IP	*Transport Control Protocol/Internet Protocol*
TDM	*Time Division Multiplexing*
UBR	*Undefined Bit Rate*
VBR	*Variable Bit Rate*
VLAN	*Virtual LAN*
VoD	*Video on Demand*
WAN	*Wide Area Network*
WDM	*Wave Division Multiplexing*
WRED	*Weighted Random Early Discard*

Riverstone Networks, Inc.
5200 Great America Parkway, Santa Clara, CA 95054 USA

877 / 778-9595 *or* ***408 / 878-6500*** *or* ***www.riverstonenet.com***