# Improvements in Ethernet Standards to Further Reduce Latency and Jitter

Norman Finn
Cisco Fellow
Cisco Systems, Inc.

Presented at the ODVA
2012 ODVA Industry Conference & 15th Annual Meeting
October 16-18, 2012
Stone Mountain, Georgia, USA

## Abstract

The independence of 1) the data plane of 802.1Q bridges; 2) the quality of service (priority) features of same; and 3) the compatibility of many of the control protocols utilizable at Layer 2; together mean that it is possible to construct a Layer 2 network upon the concept of a Network Core, a set of VLANs running a single control protocol, but sharing a physical layer and an IS-IS-based protocol foundation, with other Network Cores. The result is a network that can be optimized for several purposes simultaneously. This can provide a structure around which the many competing industrial real-time control protocols can further reduce jitter and latency, even while being used together, to their mutual advantage.

## Keywords

Bridge, VLAN Bridge, switch, IS-IS, network core, industrial network, wireless network, real-time network.

## Definition of terms

The terms **Bridge** and **Switch** are used as synonyms.

**FID:** Filtering database identifier; an integer value, obtained solely from the VID, that along with a MAC address, is used as a key pair for finding entries in the Filtering Database.

**Filtering Database:** The forwarding database of a bridge. Each entry has a unique key which is a {FID, MAC address} pair, and a payload, which is a list of ports to which a frame with that key pair can be forwarded.

**Frame:** A packet at the MAC Layer, consisting of (at least) MAC addresses, data, checksum, and overhead.

**VLAN:** Virtual Local Area Network; a community of end stations with a mutual interest; the maximum extent over which a frame with the Broadcast destination MAC address (all 1s) can be delivered.

**VID:** VLAN identifier; a 12-bit value identifying to what VLAN a frame belongs, and perhaps additional information, as well. Any number of VIDs can be mapped to a single VLAN.

# 1. Current VLAN Bridge capabilities

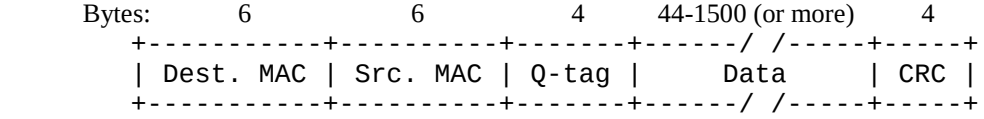
This section provides a brief description of the capabilities of an IEEE Std. 802.1Q-2011 VLAN Bridge, dwelling primarily on the data plane (the forwarding of non-control plane packets), rather than the control plane (IEEE network control packets, protocols, and procedures). The following sections of this paper will develop four key principles:

a) Quality of Service is independent of forwarding (section 2);
b) Multiple forwarding algorithms can drive the 802.1Q data plane simultaneously (section 3);
c) The IS-IS protocol can perform protocol functions independently of its ability to populate forwarding tables (section 4); and
d) Wireless media, including 802.11 Wi-Fi, can function as media internal to the network, not just as media providing access to the network.

Section 6 summarizes this information with a call for action by standards groups engaged particularly in building real-time control networks.

Although familiarity with the capabilities of bridges to the level of the most popular work on the subject, Perlman [1], is quite common in the networking industry, bridging has changed considerably in the twenty years since this survey was written. The 2008 work by Seifert and Edwards [2] includes an explanation of VLANs (Virtual Local Area Networks), but even that survey is greatly in need of revision to include subsequent development of Shortest Path Bridging [5], of which the best description can be found in Allan and Bragg [3].

The starting point for the description of IEEE Std 802.1Q VLAN bridges [4] is the simplest form of a VLAN-tagged 802.3 Ethernet frame [6], which should be familiar to most readers:




The fields are:

**Destination MAC address:** 6 bytes with the destination to which the frame is being sent. Can be the address of a single station (unicast), multiple stations (multicast), or all stations (broadcast) on a VLAN.

**Source MAC address:** 6 bytes with the unicast address of the station that sourced the frame.

**Q-tag:** Defined by 802.1Q [4]. Two bytes of EtherType (hex 81-00 or 88-A8), followed by three bits of priority, one bit of Drop Eligibility (88-A8 only), and 12 bits of VID (VLAN identifier).

**Data:** Data, starting with a protocol identifier consisting of either a two-byte EtherType (the most common case, hex 06-00 or larger) or a two-byte Length field (hex 05-FF or smaller). EtherType 08-00 denotes an IP packet, for example.

**CRC:** Cyclic Redundancy Check, a polynomial checksum on the frame.

**Not shown:** Preamble and inter-frame gap, totaling a minimum of 20 bytes.

The function of an IEEE 802.1 bridge is to interconnect various 802 (and other similar) media so that the aggregate network appears, to the stations connected to it, to provide the "MAC Service". This is the function originally offered by "fat yellow coax" Ethernet. It was a shared medium; every frame transmitted was received by all the stations on the medium, and each station filtered out the frames (such as those addressed to other stations) in which it was not interested. A bridge accomplishes this by delivering every frame to *at least* those stations to which it is addressed. The bridge optimizes the use of network resources by not delivering frames that it can reasonably infer a station is uninterested in. A network of VLAN bridges can create thousands of virtual shared media LANs on a single physical network.

Before we describe the data plane of a VLAN bridge, we must distinguish between a VLAN (Virtual Local Area Network) and a VID (VLAN identifier). Although in many networks, there is a 1:1 correspondence between VLANs and VIDs, this is not required. Multiple VIDs can be assigned to designate the same VLAN. In all cases discussed in 802.1Q [4] and in this document, the different VIDs assigned to the same VLAN are used to classify

the point of ingress of the frame into the network. This n:1 VID:VLAN relationship is fundamental to the operation of IEEE Std 802.1aq Shortest Path Bridging (see section 2.2, below).

As a frame traverses a VLAN Bridge, the sequence of operations performed on the frame is:

1. The ingress port assigns the frame a VID (and hence to a VLAN) and a priority, and may filter (discard) the frame.
2. The VID and Source MAC address can be used to modify the Filtering Database.
3. The VID and Destination MAC address are used to determine to which port(s) the frame is to be forwarded.
4. The state of the output port(s) each may filter (discard) the frame and/or
5. If not filtered, the frame is queued for transmission, and Quality of Service functions applied.

If an 802.3 frame has no Q-tag, and hence no priority, it is assigned Priority 0. Other media, such as 802.11, can carry a priority outside a Q-tag. Every port has a default VID, to be assigned to any frame with no Q-tag, or with a Q-tag carrying a VID of 0. Furthermore, 802.1Q provides for assigning a VID based on the protocol of the frame, e.g. assigning untagged IP packets to one VID, and non-IP packets to another VID. A port can have a 4k-by-12-bit input translation table that alters the VID to a completely different value, and a similar table 8-by-3-bit table for priority. Ingress red/yellow/green policing can also be performed, based on priority, Drop Eligibility, and recent history.

Having assigned the frame a VID, the port state is consulted to determine whether to filter the frame based on its VID. The state of per-port, per-VID input filtering is determined by the forwarding protocol(s) in use, e.g. Spanning Tree or Shortest Path Bridging.[1]

The frame then passes to the Filtering Database for forwarding. The first step is to map the VID into a Filtering database Identifier (FID) using a single (not per-port) VID-to-FID table. If MAC address learning is employed (a per-VID option), and if the source MAC address is a unicast address (the only legal possibility), then the source MAC address and FID are looked up in the Filtering Database. If an entry for that {FID, MAC address} pair is not already present, a new entry is created. The output for that entry is set to the port on which the frame was received. The bridge thus learns that this {FID, MAC address} pair can be reached through this port.

The next step is to look up the Destination MAC address and FID (or optionally, the VID, if the Destination MAC address is not a unicast) in the Filtering Database to determine on which port the frame is to be transmitted. If a unicast address is found in the Filtering Database, a single port number is returned, and if a multicast address, a list of ports is returned. If the address is not in the database, a list of all ports is returned. The port on which the frame was received is removed from the list, if present. The frame is then queued on all of the ports remaining in the list, or discarded if there are none.

At the output port, as controlled by the forwarding protocols and VLAN pruning algorithms, the frame can be filtered. For each VID, either all frames, all multicast or broadcast frames, or no frames are discarded. If not discarded, the priority of the frame determines to which of the available queues (from one to eight can be present) the frame is assigned. The fullness of the queue and/or the policing color (green/yellow) of the frame may cause the frame to be discarded, or a Congestion Notification frame to be sent towards the frame's source. IEEE 802.1Q provides a number of means for governing the draining of the output queues, discussed in the following paragraphs. Finally, the frame's VID is used to determine whether the output frame is to carry a Q-tag or not, the VID can be mapped to another value via an output VID translation table, and the frame finally output.

IEEE 802.1Q-2011 provides for a number of methods for selecting from which queue the next frame is to be transmitted from a port:

a. **Strict priority.** The highest priority non-empty queue is always selected.

[1] The division of 802.1Q forwarding functions into those performed by the port and those performed by the central relay function is flexible; the description, here, is aligned more with common implementation methods than the detailed text of IEEE Std. 802.1Q-2011.

b. **Weighted priority.** Weighting guarantees some bandwidth for every queue.
c. **Traffic shaped.** Specified queues are transmitted ahead of all priority queues, but a credit-based shaper guarantees a worst-case latency for both the shaped queues and for the highest-priority non-shaped queue.

Another queue selection algorithm is now in development by 802.1:

d. **Scheduled.** All queues are drained only at specified times according to a repeating schedule. At any given moment, either just one scheduled queue can transmit, or any queue of classes a-c, above, can transmit, or no queues at all can transmit.

Note that 802.1Q has never prohibited vendors and/or other standards bodies from defining other selection mechanisms. The above list is simply what 802.1, itself, has defined.

Strict and weighted priority queuing have obvious applications. Traffic shaping is used to guarantee that streams that pre-register their bandwidth needs and have been allocated that bandwidth by the bridges in the network (via the Stream Reservation Protocol) will not lose frames due to congestion. Scheduled queues are used to enable the transmission of time-critical frames at specific times, without the possibility of interference by frames with other priority values.

Historically, the Filtering Database has been populated by a combination of learning (for unicast MAC addresses) and protocols (for multicast MAC addresses). The per-VID port states have been controlled by a combination of explicit configuration and the operation of the forwarding protocols such as Spanning Tree. However, section 3 discusses important alternatives to this scenario.

The above queue selection algorithms, along with their associated protocols, are intended by IEEE 802.1 to make it possible to integrate time-critical process control traffic, audio/video infotainment traffic, network control traffic, and best-effort traffic, all on the same IEEE 802 network.

# 2. Multiple forwarding algorithms can be used simultaneously

Inherent in the description in section 1 is the ability of the data plane to accommodate more than one forwarding algorithm simultaneously. It is not commonly appreciated, for example, that all of the topology control protocols defined by IEEE 802.1 (Multiple Spanning Tree Protocol, MSTP, and Shortest Path Bridging, SPB, both V-mode and M-mode) can be configured to each control a disparate set of VIDs, and that the remaining VIDs can be left uncontrolled by any of these protocols. This is how Provider Backbone Bridging Traffic Engineered paths (802.1Q-2001 Clause 25.10) are created outside the control of any protocol; the paths can be created by management action on VIDs not assigned to any 802.1Q topology control protocol. No learning is performed on these VIDs, and if no entry is present for a frame's Destination MAC address in one of these VIDs, the frame is discarded.

We will briefly describe four protocols in the following subsections, only to conclude in section 3.5 that all can work together, simultaneously.

## 2.1 Multiple Spanning Tree Protocol (MSTP)

This protocol has been slowly advanced since its original specification in the 1980s. In point-to-point links, it converges quickly via handshakes, instead of using timeouts. Up to 64 separate topologies can be created in a single network, to be used on a per-VLAN basis. The advantages and disadvantages of this protocol are well-known:

+ Handles absolutely any topology; not constrained to a ring or ladder.
+ Can be plug-and-play, requiring no configuration.
– Often leads to very sub-optimal routing choices; each VLAN uses only one tree.
– Worst-case convergence time is several seconds.

## 2.2 Shortest Path Bridging, V-mode (SPBV)

Although a complete description of Shortest Path Bridging V-mode (SPBV) is beyond the scope of this paper (see [3] for a good one), a brief introduction is necessary. SPBV is based on the Intermediate System to Intermediate

System (IS-IS) protocol defined, among other documents, in [7] and [8]. In IS-IS, every participating system (VLAN Bridges, in this case) advertises its current state to its neighbors. This state includes, at a minimum, the name of the device and a list of the links that it has that connect to other named participating systems, along with the characteristics of those links, such as bandwidth. Every system rapidly and reliably propagates the information it receives about other systems, so that, very quickly, every participating system can construct a complete topological map of the network, based on the received advertisements, and quickly modify that map when links or systems come up or fail.

In an SPBV network, every SPBV bridge is assigned a Bridge ID, and the VIDs are allocated such that there is a unique VID value for every combination of bridge and VLAN. Before a bridge passes a frame received from a station to another bridge, it labels it with a Q-tag whose VID indicates both the VLAN and the ID of the original receiving bridge. Since the VID has only 12 bits, the total number of VLANs times the number of SPBV bridges in the network must be less than 4k. Using the bridge ID and VLAN needs advertised via IS-IS, every SPBV bridge is able to configure its port states so that every frame to an unknown or broadcast destination is propagated along a spanning tree rooted at the source bridge, and thus along the least-cost path from its source. After learning MAC addresses from passing traffic, every frame is forwarded along the path with the least possible "cost" (the sum of the inverse of the speed of each link) to its destination. When a topology change occurs, additional protocol mechanisms are defined for SPBV that prevent any chance of forwarding a frame in a loop.

The net result is another list of plusses and minuses:

+ Handles absolutely any topology; not constrained to a ring or ladder.
+ Can be plug-and-play (if a per-industry profile is specified).
+ Unicasts and multicasts are always routed along shortest path; not constrained by spanning tree.
0 Worst-case convergence time is on the order of 100 msec.
- (Number of VLANs) times (number of bridges) < 4095

## 2.3 Ring protocols

There are a number of protocols available that require the bridges to be connected in a ring, with exactly two trunk ports per bridge. The advantage of such protocols is that, since the possibilities for both failures are recovery actions are very limited, decisions about what to do when a link or node fails or recovers can be made very rapidly. Examples of these protocols are ITU-T G.8032 and ODVA DLR. We can create another list of advantages:

+ Fast (≈10 ms) response to a link or node failure.
0 Forwarding typically not optimal, but this varies with the protocol.
– Two or more failures lead to a loss of connectivity.
– If the physical topology does not match the topology demanded by the protocol, the network will not operate correctly.

In general, the fast convergence time of ring protocols considerably outweighs their disadvantages.

## 2.4 Multiple simultaneous delivery by traffic engineered paths

There are many variants of this technique, but we will consider one, as described, here. Either the end station or the adjacent edge bridge replicates every critical frame and sends it along two (or more) disjoint paths toward the destination, where a bridge or end station may discard the duplicates. These paths are created by management or by an automatic process, but outside the control of any topology control protocol such as a spanning tree or a ring. If a failure occurs reaction time is not low or zero, there is in fact no concept of reaction time; the other path simply continues to work.

+ No response required to a single link or node failure.
+ Redundancy not dependent on a device taking an action.
0 Paths and flows must be set up.
– Bandwidth usage is at least doubled.
– If the physical topology does not match the topology demanded by the protocol, the network will not operate correctly.

### 2.5 All at once

Traditionally, the designer of a network has been forced to pick from among the advantages and costs of multiple examples of each of the kinds of protocol described in subsections 2.1 through 2.4 and come up with the best balance for the different kinds of traffic that his or her network will be carrying. However, if the 802.1Q bridge data plane described in section 1 is used, then the network designer can make multiple choices and run them all at the same time, satisfying multiple needs. All that is required is that the protocols are confined to work on separate sets of VIDs.

Depending upon its needs, an end station may or may not be aware of the multiple VLANs. The attached bridge(s) can often classify its traffic among multiple VLANs based on inspection of the traffic, e.g. assigning ARINC-64 frames used for controlling a vehicle to one VLAN (controlled by a ring), and IP packets used for downloading software (controlled by spanning tree) to another.

## 3. Quality of Service is independent of forwarding

From the summary in section 1, it may or may not be apparent to the reader that the Quality of Service provided the various data streams in a network is independent, in the 802.1Q model, from the forwarding algorithms. That is, the choice of *which port* a frame is to be forwarded is ultimately determined by the forwarding protocol and how it populates the Filtering Database. The QoS observed by the frame is determined by *when* the frame is output on that port, which is determined by the frame's priority (and drop eligibility). In the 802.1Q model, there is no overlap between these concepts: only the VID and Destination MAC address, as controlled by the forwarding protocol, determine the path taken by a frame; and only the priority and drop eligibility determine when the frame is output from the selected port, and thus the Quality of Service.

In the context of section 3, this has an important consequence: *All of the QoS features of 802.1Q are available to any protocol that is compatible with 802.1 protocols on the basis of VID separation*. That is, on a queue on a given output port, frames with the same priority value but with different VIDs, and controlled by different forwarding algorithms, are mixed together in FIFO order. One can have reserved-bandwidth streams, for example, on multiple VLANs controlled by multiple topology protocols. It will frequently be the case that certain priorities and VIDs are used together exclusively, but this is a management choice, not a requirement of the protocols.

## 4. IS-IS can perform protocol functions outside its "routing" function

IS-IS has been described so far simply as the basis of SPBV. However, it can be more than that, and offer services of value to other forwarding protocols, as well.

As an example, Multiple Spanning Tree Protocol (MSTP) bridges use the Multiple VLAN Registration Protocol (MVRP) to prune VLANs. That is, MVRP enables bridges to forward broadcasts on a given VLAN only to those parts of the network where there are end stations that are configured for that VLAN. This is accomplished by the exchange of MVRP packets conveying the need of each bridge to receive certain VLANs based on its configuration. If a link or node fails or is restored, then this information must be exchanged, again, because the path to or from a part of the network needing a given VLAN may have changed. This protocol can be a significant burden on bridges during and just after a topology change.

In SPBV, the information about what VLANs are required by a given bridge is passed as part of the link state advertisements. This information typically does not change with dynamic changes of topology. Each bridge is able, using the map of the network it can compute, to determine on which ports that broadcasts for each VLAN need be propagated. When a topology change occurs, no MVRP exchanges are necessary; only a recomputation of how to deliver the (unchanged) requirements of the individual bridges.

Similarly, IEEE Std 802.1AS Time Synchronization [9], the profile of IEEE 1588 generated by the IEEE 802.1 AVB Task Group, uses an algorithm derived from the spanning tree to elect the best candidate as the Grand Master clock source, and to construct a distribution tree for its timing information. IEEE 802.1 is investigating the use of IS-IS advertisements to propagate information about each node's claim to being the Grand Master. The election and construction of the distribution tree could then be accomplished without any additional protocol exchanges when the network is started up or when a failure or recovery takes place.

In a network with rings (section 2.3) or traffic engineered paths (section 2.4) IS-IS could be used by every station, or just by supervisory stations, to determine whether or not the physical topology of the network matches the intentions of the designer of the rings or paths, or even as a tool for a supervisory station to construct engineered paths. This use of IS-IS can supersede requirements for topology control protocols that often consume more lines of code in ring and engineered networks than the forwarding protocols, themselves.

## 5. Wireless media are no longer just at the edge of the network

Study Groups have been initiated in both IEEE 802.1 and IEEE 802.11 to determine how to start work on parallel projects in the two Working Groups to incorporate 802.11 media into bridged networks in the same manner that wired media are included. In particular, the current 802.11 restriction that a non-Access Point station cannot relay frames to or from any other medium would be lifted; that wired+wireless stations would simply become a kind of VLAN Bridge.

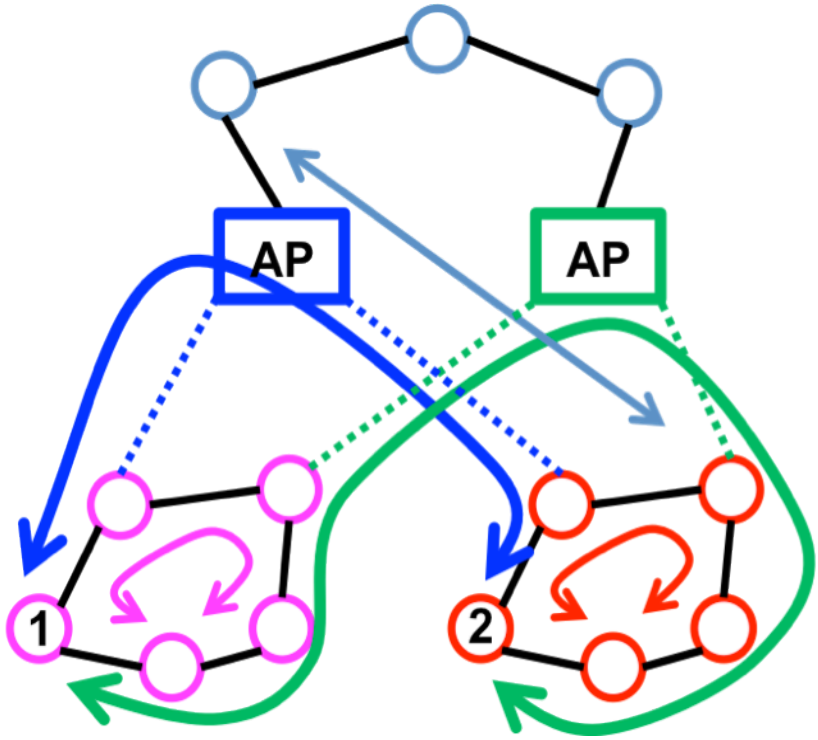
The use cases for this step are numerous, but perhaps the best example is a Layer 2 network in which Access Points are bridged together, and allow robots, each with multiple stations, to communicate with each other and with the wired backbone via 802.11, where each robot is composed, in turn, of a wired network of bridges, sensors, actuators, processors, and as mentioned, non-AP stations.

## 6. Network Cores – a call for action

If one puts together the ideas in the preceding sections, one can create *network cores*. A network core is a VLAN comprised of one or more VIDs running over all or just a part of a Layer 2 network, controlled by a single administrator and a single network control protocol (section 2), that shares its physical plant of links and bridges with any number of other network cores. The network cores share facilities:

1. The data planes of the 802.1Q VLAN Bridges used to implement the network.
2. The IS-IS protocol, as needed, to accomplish tasks such as bandwidth reservation that are common to multiple network cores.
3. The port queues of the bridges that provide Quality of Service features.

An example of three network cores, an SPBV maintenance core, two ring cores, and a traffic engineered core:




Wired parts of the base network.

SPBV on VLAN 1 enables ***all*** of these stations to communicate for maintenance purposes.

Access Points

Wireless “links”

Robots, each consisting of a wired ring of sensors, actuators, sensors, processors and bridges.

Each robot controlled by a separate instance of a ring protocol.

Robots communicate via two
engineered paths.

Other standards bodies, notably ITU-T, have already published protocols (e.g., ITU-T G.8031 and G.8032) that define what are, in effect, network cores. They define protocols that operate differently then the 802.1 protocols, but are able to coexist with the 802.1 protocols on the same VLAN Bridge, because they are configured to use a separate set of VLANs. This means that networks using these ITU-T protocols can instantly and automatically reap the advantages of IEEE 802.1 advances such as Stream Reservation, and 802.1 can instantly obtain the benefit of ITU-T advances such as Performance Measurement.

This author believes that the time is ripe for the standards organizations now defining industrial and other real-time control protocols for Ethernet and wireless networks to join with IEEE and ITU-T in a common framework of standards that will convert a maze of incompatible standards into a menu of customer choices.

## References


[1] Perlman, Radia, "Interconnections: Bridges and Routers," Addison-Wesley Pub. Co., 1992.
[2] Seifert, Rich, and Edwards, Jim, "The All-New Switch Book," Wiley Publishing, Inc., 2008.
[3] Allan, David, and Bragg, Nigel, "802.1aq Shortest Path Bridging Design and Evolution," IEEE Press and John Wiley & Sons, Inc., 2012.
[4] IEEE Std 802.1Q-2011, "Media Access Control (MAC) Bridges and Virtual Bridge Local Area Networks," IEEE, 2011, http://standards.ieee.org/about/get/802/802.1.html.
[5] IEEE Std 802.1aq-2012, "Media Access Control (MAC) Bridges and Virtual Bridge Local Area Networks: Amendment 20: Shortest Path Bridging".
[6] IEEE Std 802.3-2008, "Carrier sense multiple access with Collision Detection (CSMA/CD) Access Method and Physical Layer Specifications," http://standards.ieee.org/about/get/802/802.3.html.
[7] "Information Techonology—Telecommunications and Information Exchange between Systems—Intermediate System to Intermediate System Intra-Domain Routing Information Exchange Protocol for Use in Conjunction with the Protocol for providing the Connectionless-Mode Network Service (ISO 8473)," ISO/IEC 10589, Second edition 2002-11-15.
[8] IETF RFC 6165, "Extensions to IS-IS for Layer-2 Systems," April 2011, http://www.rfc-editor.org/rfc/rfc6165.txt.
[9] IEEE Std 802.1AS-2011, "Timing and Synchronization for Time-Sensitive Applications in Bridged Local Area Networks", http://standards.ieee.org/about/get/802/802.1.html.


**********