TRILL and IEEE 802.1aq Overview

Ronald van der Pol
rvdp@sara.nl

DRAFT Version 1.2
April 2012

1 Introduction

A big limitation of current day Ethernet networks is the way loops in the topology are handled with spanning tree protocols (STP - Spanning Tree Protocol, RSTP - Rapid Spanning Tree Protocol, MSTP - Multiple Spanning Tree Protocol) [1]. These spanning tree protocols prune various links from the topology in order to end up with a strict tree topology. The problem with this is that it can result in very inefficient network use. Figure 1 shows an example of this. Instead of forwarding

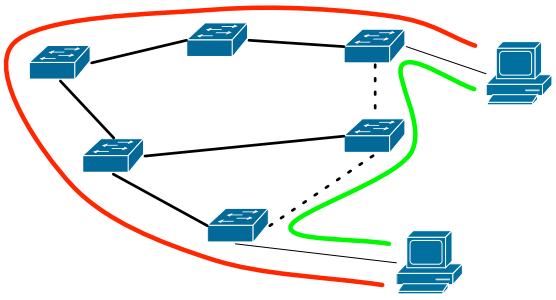


Fig. 1. Inefficient network use when using Spanning Tree Protocol

traffic over the shortest path formed by the Ethernet switches on the right (green line), the spanning tree protocols prune those links from the topology to form a spanning tree (solid lines) and forces the traffic via this spanning tree to the switch at the left and then to the destination (red line). This strict tree topology also means that alternative paths between two nodes cannot be used at the same time because that would form a loop in the topology. Another disadvantage of spanning tree protocols is that convergence after a topology change can take up to several seconds (in the case of RSTP) or 30 seconds (in the case of the original STP) in worst case scenarios. Finally, the active forwarding topology can not be easily foreseen by network managers. Although the algorithms involved are deterministic, the end result cannot easily predicted and as a result, the forwarding topology of Ethernet networks with spanning tree protocols is not as predictable as network managers would like it to be.

The next two sections describe alternatives for spanning tree protocols. Both use the IS-IS (Intermediate System to Intermediate System) routing protocol instead of spanning tree protocols and forward traffic via the shortest path between two nodes. Section 2 describes the IETF TRILL protocol and section 3 describes the IEEE 802.1aq (SPB - Shortest Path Bridging) protocol. Section 4 discusses how TRILL or SPB concepts could be used within the LHCONE infrastructure. Finally, section 5 gives the status of vendor support for IETF TRILL and IEEE 802.1aq.

2 IETF TRILL - Transparent Interconnection of Lots of Links

TRILL (Transparent Interconnection of Lots of Links) [2] is a L2 forwarding protocol that operates within one IEEE 802.1-compliant Ethernet broadcast domain. It replaces the spanning tree

protocol by using IS-IS (Intermediate System to Intermediate System) routing to distribute link state information and calculate shortest paths through the network. IS-IS is used because it is a pure L2 routing protocol that does not require IP for transporting the frames. TRILL data packets and IS-IS routing packets are exchanged between Routing Bridges (RBridges). RBridges automatically discover each other via IS-IS Hello frames and require no explicit configuration. End station MAC addresses are learned at the edges (at ingress and at egress) of a TRILL domain only. RBridges in the core do not need to keep track of end station MAC addresses. TRILL's main focus is (highly) meshed topologies, e.g. data centres.

TRILL is being standardised by the IETF and has been a working group since March 2005. The base protocol RFCs (RFC 6325 [3], RFC 6326 [4], RFC 6327 [5], RFC 6361 [6] and RFC 6439 [7]) are at proposed standard maturity level. This is what the charter says about its current goals and work items: "The TRILL WG has specified a solution for shortest-path frame routing in multi-hop IEEE 802.1-compliant Ethernet networks with arbitrary topologies, using an existing link-state routing protocol technology and encapsulation with a hop count. The current work of the working group is around operational and deployment support for the protocol. This includes a MIB module and other pieces needed for operations, but also additional ways to extend and optimize TRILL for the properties of the networks on which it is deployed."

RBridges encapsulate Ethernet frames at the ingress, route the frames through the TRILL domain using IS-IS link state routing information and decapsulate the Ethernet frames at the egress again. This process is shown in figure 2. Host A on the left (MAC address mac1) sends

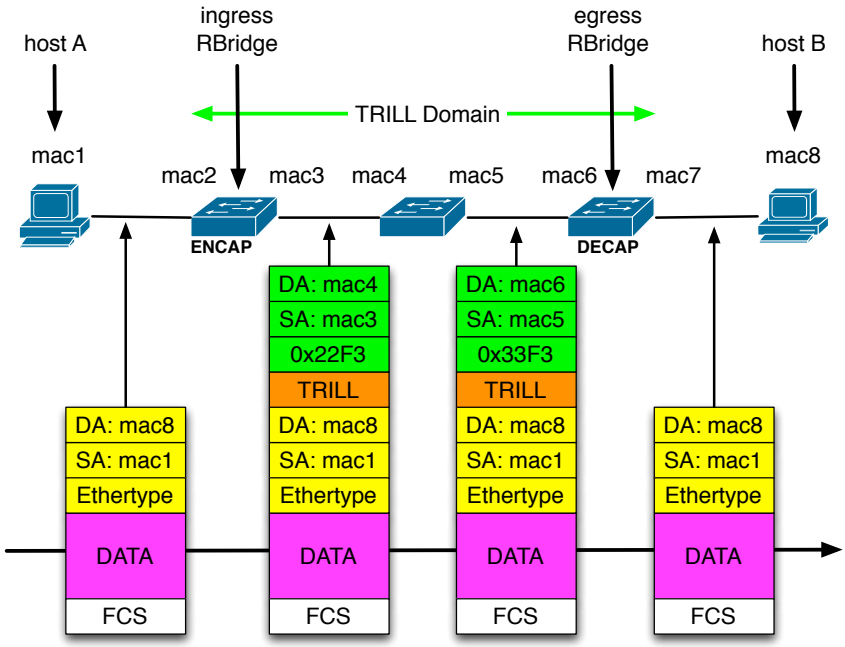
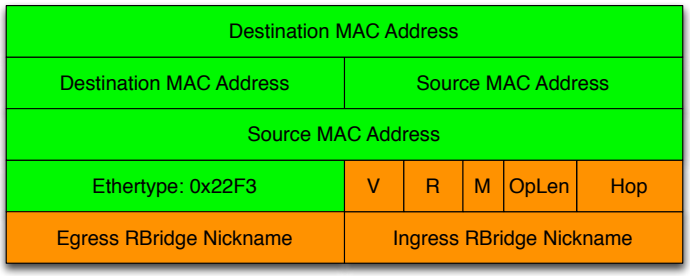


Fig. 2. TRILL encapsulating and decapsulating

an Ethernet frame to host B on the right (MAC address mac8). The frame is encapsulated by the ingress RBridge with a TRILL header and an Ethernet header. The TRILL header contains a 2-byte identifier of the ingress RBridge and a 2-byte identifier of the egress RBridge. The egress RBridge is determined by the shortest path to the final destination (host B). The TRILL header also contains a hop count. The ingress RBridge sets the source address in the outer Ethernet header to the MAC address of it's outgoing interface and sets the destination address to the MAC address of the next hop RBridge (determined by IS-IS shortest path routing). Each RBridge in the path

rewrites the outer Ethernet header with its own MAC address as source and the MAC address of the next hop RBridge as destination. At each RBridge the hop count in the TRILL header is also decremented. At egress, the encapsulation headers are removed and the original Ethernet frame is sent to the destination (host B). There can be any number of traditional Ethernet switches between two RBridges because switches that do not support TRILL just forward the traffic based on the destination MAC address and VLAN ID, if present. RBridges do not propagate spanning tree protocol BPDUs (Bridge protocol Data Units), so RBridges can limit the span of spanning trees regions. Figure 3 shows the Ethernet and TRILL headers used by the TRILL protocol. The



V	Version	2 bit
R	Reserved	2 bit
M	Multi-Destination	1 bit
OpLen	Option Length	5 bit
Hop	Hop Count	6 bit
Nickname	Nickname	16 bit

Fig. 3. Ethernet and TRILL headers

TRILL protocol uses 0x22F3 as Ethertype. TRILL can encapsulate untagged or tagged customer frames. The tags in customer frames are preserved while forwarding through a TRILL domain.

Unicast frames are sent via the shortest path between ingress and egress RBridge. RBridges use distribution trees to forward multi-destination frames (i.e. unicast frames with unknown MAC address location, multicast and broadcast frames). There can be one or more distribution trees in a TRILL domain. RBridges precompute all the distribution trees that might be used. Each RBridge advertises in its TRILL IS-IS LSP (Link State Packet) the maximum number of distribution trees it can support. There is one RBridge in a TRILL domain that is chosen by all RBridges (based on several RBridge qualifiers) that decides for all RBridges:

- How many trees will be used
- Which trees will be used
- The tree number of each tree

The number of distribution trees that is decided upon cannot be higher than the number of trees that is supported by the RBridge with the fewest number of supported trees. Each ingress RBridge that is the appointed forwarder for a certain VLAN chooses which of the distribution trees it will use for multi-destination frames. Usually, this is the tree whose root has the lowest cost path from the ingress RBridge. The ingress RBridge itself could be the root of the chosen tree. It includes the chosen tree information in its LSP distributions and all other RBridges keep track of it so that they can use this information in Reverse Path Forwarding (RPF) checks. The distribution trees are shared across all VLANs, but pruning should take place per VLAN when a branch has no potential

receivers. When receiving a native multi-destination frame, the ingress RBridge converts it to a TRILL data packet. It uses the All-RBridges multicast address as outer destination MAC address. Multipathing over multiple distribution trees is supported. The ingress and egress RBridges are the only bridges that learn end station MAC addresses and VLAN information. At ingress, the RBridge learns local end station information by collecting port and MAC address and VLAN information of received frames. At egress, the RBridge learns remote end station information by collecting ingress RBridge and MAC address and VLAN information of the original frame in the TRILL encapsulated packet. Transit RBridges do not need to collect end station information. They only need to collect MAC addresses of other RBridges. Thus, the MAC forwarding table of transit RBridges scales with the number of RBridges instead of with the number of end stations. The RBridges can choose a backbone VLAN in the TRILL domain to communicate which each other. This VLAN is independent from the VLAN used in the original Ethernet frame in the encapsulated TRILL packet.

The ESADI (End Station Address Distribution Information) protocol can be used to announce end stations that have been explicitly enrolled. Advertising end station MAC addresses using ESADI is optional, as is learning from these announcements. RBridges that are the appointed forwarder for a certain VLAN may participate in the TRILL ESADI protocol for that VLAN. All transit RBridges must properly forward TRILL ESADI frames as if they were multicast TRILL Data frames. There are a couple of advantages of using ESADI. The enrollment might be authenticated (for example, by cryptographically based EAP methods via IEEE Std 802.1X-2010 [8]). The ESADI protocol also supports cryptographic authentication of its messages (RFC 5304 [9] and RFC 5310 [10]) for more secure transmission. Finally, If an end station is unplugged an immediate update can be sent via the ESADI protocol.

RBridges can optionally support multipathing. This is done by Equal Cost Multipath (ECMP) routing. If multiple equal costs paths are presents towards a destination, an RBridge can distribute traffic over those multiple paths. This is usually done per flow is order to avoid reordering and path MTU discovery problems.

3 IEEE 802.1aq - Shortest Path Bridging

The IEEE 802.1aq standard [11] was approved on 28 March 2012. It is an amendment to the "Virtual Bridge Local Area Networks" standard (IEEE Std 802.1Q-2011 [12]) and adds Shortest Path Bridging (SPB). SPB uses shortest path trees (SPTs) as an alternative to the spanning trees used by STP (Spanning Tree Protocol), RSTP (Rapid Tree Spanning Tree Protocol) and MSTP (Multiple Spanning Tree Protocol). Shortest path trees guarantee that traffic is always forwarded via the shortest path between two bridges. Regions where SPTs are used can co-exist with regions where STP, RSTP or MSTP spanning trees are used. A CIST (Common and Internal Spanning Tree) is used within the SPB region as a default spanning tree to interwork with STP, RSTP and MSTP regions. The IS-IS (Intermediate System to Intermediate System) link state routing protocol (ISIS-SPB) is used within an SPB region to exchange information between the bridges in order to calculate SPTs. Shortest path trees are calculated from every source bridge in an SPT region to all other bridges in that region. The SPT algorithms are defined is such a way that it is guaranteed that all bridges in the region calculate exactly the same set of trees. SPTs are bidirectional, which means that forward and reverse traffic takes the same path. Unicast and multicast traffic between two bridges also takes the same path. Load balancing is supported by calculating multiple Equal Cost Trees (ECTs). Each ECT uses a different SPT algorithm tie-breaker. By default, 16 ECTs are supported by defining 16 different ECT tie-breakers. A SPT set is defined by all the SPTs that share the same ECT tie-breaker and that support one or more VLANs within an SPT region. Load balancing is done by distributing VLANs over up to 16 SPT sets. This means that there is no per flow (or packet) load balancing between two end stations, unless these end stations use different VLANs for different flows. However, work has started on IEEE 802.1Qbp [13], which introduces Equal Cost Multiple Paths (ECMP) in the SPBM data forwarding process by using a hash based choice of possible next hops. It is anticipated that a new

tag will be defined, possibly including a Time to Live (TTL) field. Ethernet OAM (Operations, Administration and Maintenance) is supported in SPB by the IEEE 802.1ag (incorporated in IEEE Std 802.1Q [12]) and ITU-T Y.1731 [14] protocols. There are two variants of SPB defined, SPBV and SPBM as described in the next two sections.

3.1 Shortest Path Bridging VID - SPBV

VID stands for VLAN Identifier. Each VLAN that is handled by SPBV uses an SPT set. An SPVID (Shortest Path VLAN Identifier) is assigned (manually or automatically) to each SPT in the set. The SPVID to SPT mapping information is sent to other bridges by using ISIS-SPB. At ingress into the SPBV region the VID in a C-TAG or S-TAG of a customer frame is translated to the SPVID corresponding to the SPT that supports that VID. At egress the SPVID is translated back to the original VID. When customer frames do not contain a C-TAG or S-TAG, SPBV adds a tag with the SPVID at ingress. At egress this tag is removed again. In an SPBV region MAC addresses of end stations are learned at each bridge in the path.

3.2 Shortest Path Bridging MAC - SPBM

MAC stands for Media Access Control. SPBM is used in combination with Provider Backbone Bridges (PBB, defined in 802.1ah which is incorporated in IEEE Std 802.1Q [12]). PBB is also called mac-in-mac because customer frames are encapsulated with an Ethernet header, the PBB header. At ingress a PBB header is added by a Backbone Edge Bridge (BEB) and at egress this header is removed again by a BEB. The PBB header contains the source MAC address of the ingress BEB and the destination MAC address of the egress BEB. It also contains a 24-bit I-SID (backbone service instance identifier), so 2^{24} different backbone services can be configured. A service is similar to a VPN (Virtual Private Network). Frames are only transported between ports that map to the same I-SID. Each I-SID is mapped to a backbone VID (B-VID). It is possible to map multiple I-SIDs to the same B-VID. Within the PBB backbone frames are forwarded based on destination backbone MAC (B-MAC) address and backbone VID (B-VID). PBB separates customer VIDs from backbone VIDs and end station MAC learning is done only at the edges of a PBB region. The PBB backbone switches only know about B-MACs. SPBM uses a single backbone VID (B-VID) for all the SPTs in a set. Figure 4 shows an example of SPBM encapsulation (by BEB B) and decapsulation (by BEB E). Several service interface options are offered for customer ports by a PBB Backbone Edge Bridge (BEB):

Port Based Customer frames can be untagged or tagged with a C-TAG. In both cases all frames are mapped to a single backbone service instance (I-SID) and are forwarded without an S-TAG.

S-tagged Service Interface This service interface is used for forwarding frames with an S-TAG (also called Q-in-Q frames). There are two types of S-tagged service interfaces. One offers a one-to-one mapping between S-TAG and I-SID, the other maps multiple S-TAGs to the same I-SID. In the first case the S-TAG is not carried within the PBB backbone. In the latter case the S-TAG is carried in the PBB header because otherwise the egress BEB would not know which S-tag the original frame had.

Figure 4 shows an S-tagged Service Interface on BEB B where multiple S-tags are mapped on I-SID 100, so the S-tag is carried through the PBB backbone. End station H1 sends a frame to end station H2. Both are using untagged frames and each is attached to a switch (S1 and S2 respectively). The end stations are connected to a switch port that is configured in port based VID (PVID) mode with VID value 20. Switches S1 and S2 connect to the service provider switches BEB B and BEB E with an S-tagged interface. Switch S1 sends the user frame from end station H1 as an S-tagged frame to BEB B. On switch S1 the S-tag to use for each VLAN must be configured. Each C-VLAN can be assigned to a different S-tag or multiple (or all) C-VLANs can be assigned to the same S-tag. Using the same S-tag makes it possible to transport several customer VLANs via the same S-tag over the same I-SID service instance. Backbone Edge Bridge B is using an

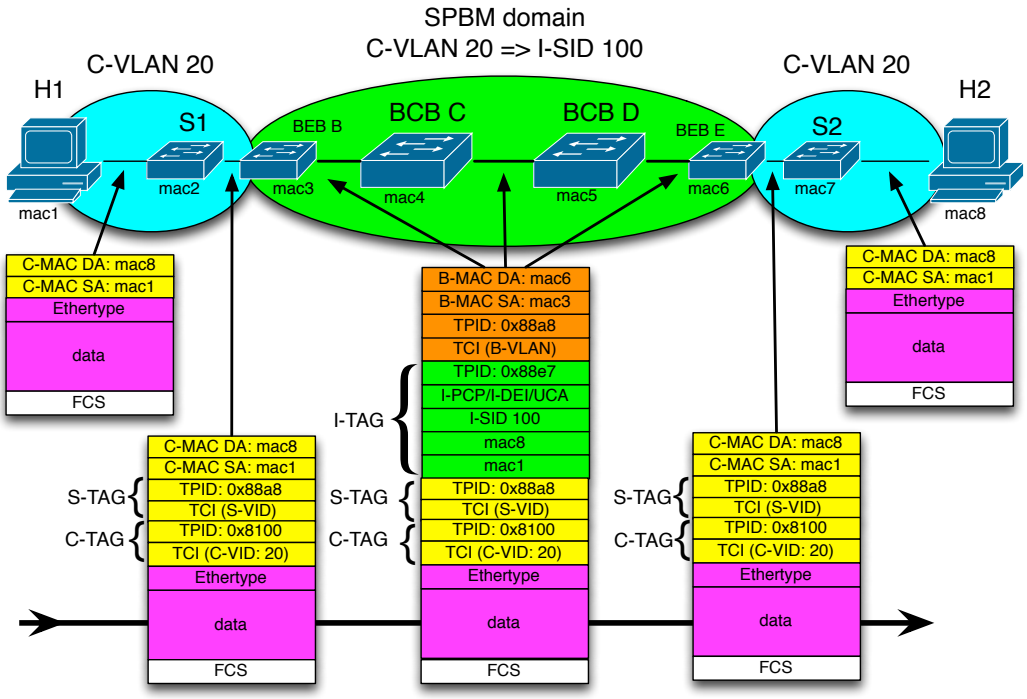


Fig. 4. Example of mapping C-VLAN 20 over I-SID 100 by using a S-tagged Bundled Interface in SPBM. The fields in each frame are sent on the network in top to bottom order.

S-tagged Service Interface and encapsulates the received frame with a PBB header. At BEB B there is a mapping between S-tag and I-SID configured which determines which provider service will be used in the backbone. This can be a one-to-one mapping or several S-tags can be bundled on the same I-SID. The example shows the latter. In this case the S-tag needs to be included in the PBB header so that the decapsulating side knows which S-tag to use in the decapsulated frame. In this example C-VID 20 is mapped eventually to I-SID 100. The encapsulated frame is sent to the egress bridge BEB E by using its MAC address as destination MAC address in the (outer) PBB header. It is forwarded through the SPBM region via the shortest path. The VID used in the PBB header is the Backbone VID (B-VID) that is used to carry I-SID 100. The mapping between I-SID and B-VID (one-to-one or many-to-one) is configured by the administrator. At egress, the Backbone Edge Bridge E decapsulates the frame and the resulting S-tagged frame is forwarded to switch S2, which removes the S-tag and C-tag and sends the untagged frame to end station H2. Vendors can choose to include the functionality of switch S1 and BEB B in the same box and offer a direct C-VID to I-SID mapping and hide the S-tag from the customer. By doing so there is a simple mapping between C-VLANs and provider services. So customers can choose to use a particular provider service by using a particular VLAN. In SPBM end station MAC learning is done in the decapsulating process. In the decapsulation process BEBs learn which end stations are located behind each of the other BEBs. PBB frames contain the end station source MAC address in the I-tag field and also the BEB source MAC address in the outer PBB header.

4 TRILLs and/or SPB in LHCONE

Both TRILL and SPB use IS-IS routing to setup forwarding in meshed Ethernet topologies. In LHCONE this can be used in the L2 Ethernet clouds. The advantage with TRILL or SPB is that links will be used more efficiently than when using spanning tree protocols. Another advantage of

TRILL or SPB is that adding additional links will make the network more resilient and efficient. TRILL or SPB based Ethernet clouds scale much better than Ethernet networks based on spanning tree protocols. With TRILL or SPB LHCONE could make use of large low cost Ethernet clouds where forwarding takes place via shortest paths and IS-IS takes care of rerouting in case of link failures. TRILL or SPB can be used with both VRF (Virtual Routing and Forwarding) and lightpath setups. Routers (with VRF) can interconnect TRILL or SPB based clouds. Lightpaths can be setup through a TRILL or SPB region by specifying an ingress and an egress port in that TRILL or SPB region.

With TRILL, end station MAC addresses are learned at the edge bridges only. TRILL can encapsulate and forward both tagged and untagged customer frames through a TRILL domain. The lightpath provisioning software does not need to configure anything on the RBridges.

When using SPBV in automatic mode, the mapping between customer VID and SPVID does not need any explicit configuration. So in this case the lightpath provisioning software also does not need to configure anything on the SPBV bridges.

When SPBM is used, the lightpath provisioning software only needs to configure a mapping from customer frame to I-SID. But this is only needed at the ingress and egress Backbone Edge Bridges (BEBs). The bridges in the core just forward the frames based on backbone VID (B-VID) and egress BEB MAC address.

So lightpath provisioning software needs very little extra support in order to setup lightpaths through either TRILL or SPB regions. The advantage of using TRILL or SPB regions instead of individual Ethernet point-to-point links is that rerouting in case of link failures is done by TRILL or SPB IS-IS routing and needs no explicit support from the lightpath provisioning software.

5 Vendor support

These vendors have TRILL products:

- IBM (Blade Networks) RackSwitch G8264 (not verified yet)
- Broadcom network processors (e.g. BCM56840 Trident [16])
- Intel (Fulcrum) network processors (e.g. FM6000 [17])
- Marvell network processors (not verified yet)
- Mellanox network processors (not verified yet)
- Solaris (software implementation) [15]
- Dell (Force10) (announced)
- Extreme Networks (announced)
- Huawei (announced)
- HP (announced)
- Cisco (TRILL-like FabricPath)
- Brocade (TRILL-like VCS)

These Ethernet switches support IEEE 802.1aq:

- Alcatel-Lucent OmniSwitch 6900
- Avaya ERS-8800/ERS-8600
- Avaya VSP-7000
- Avaya VSP-9000
- Huawei S9300

There have been three interoperability events for 802.1aq.

Some vendor statements:

- HP supports both TRILL and 802.1aq
- Huawei supports both TRILL and 802.1aq
- Juniper has QFabric. CTO Pradeep Sindhu about TRILL: "a solution looking for a problem"
- Extreme Network sees Multi-System Link Aggregation (M-LAG) as an alternative for TRILL and 802.1aq

6 Glossary

B-MAC Backbone MAC address
BPDU Bridge Protocol Data Unit
CIST Common and Internal Spanning Tree
CST Common Spanning Tree
ECMP Equal Cost Multiple Paths
ECMT Equal Cost Multi Tree
ECT Equal Cost Tree
FID Filtering Identifier
IEEE Institute of Electrical and Electronics Engineers
IETF Internet Engineering Task Force
I-SID Backbone Service Instance Identifier
IS-IS Intermediate System to Intermediate System
IST Internal Spanning Tree
LHCONE Large Hadron Collider Open Network Environment
MAC Media Access Control
MSTP Multiple Spanning Tree Protocol
OAM Operations, Administration and Maintenance/Management
PBB Provider Backbone Bridges
PVID Port VLAN Identifier
RPF Reverse Path Forwaring
RSTP Rapid Spanning Tree Protocol
SPB Shortest Path Bridging
SPBM Shortest Path Bridging MAC
SPBV Shortest Path Bridging VID
SPVID Shortest Path VLAN Identifier
STP Spanning Tree Protocol
TCI Tag Control Information
TPID Tag Protocol Identifier
TRILL Transparent Interconnection of Lots of Links
VID VLAN Identifier
VLAN Virtual Local Area Network
VPN Virtual Private Network
VRF Virtual Routing and Forwarding

7 Acknowledgements

Thanks to David Allan for reviewing the SPB text.

References

1. IEEE Std 802.1D-2004
   IEEE Standard for Local and metropolitan area networks
   Media Access Control (MAC) Bridges
   Print: ISBN 0-7381-3881-5 SH95213
   PDF: ISBN 0-7381-3982-3 SS95213
   9 June 2004
2. RFC 5556
   Touch, J., Perlman, R.
   Transparent Interconnection of Lots of Links (TRILL): Problem and Applicability Statement
   May 2009

3. RFC 6325
Perlman, R., Eastlake 3rd, D., Dutt, D., Gai, S., Ghanwani, A.
Routing Bridges (RBridges): Base Protocol Specification
July 2011
4. RFC 6326
Eastlake, D., Banerjee, A., Dutt, D., Perlman, R., Ghanwani, A.
Transparent Interconnection of Lots of Links (TRILL) Use of IS-IS
July 2011
5. RFC 6327
Eastlake 3rd, D., Perlman, R., Ghanwani, A., Dutt, D., Manral, V.
Routing Bridges (RBridges): Adjacency
July 2011
6. RFC 6361
Carlson, J., Eastlake 3rd, D.
PPP Transparent Interconnection of Lots of Links (TRILL) Protocol Control Protocol
August 2011
7. RFC 6439
Perlman, R., Eastlake 3rd, D., Li, Y., Banerjee, A., Hu, F.
Routing Bridges (RBridges): Appointed Forwarders
November 2011
8. IEEE Std 802.1X-2010
IEEE Standard for Local and metropolitan area networks
Port-Based Network Access Control
Print: ISBN 978-0-7381-6146-4 STDPD96008
PDF: ISBN 978-0-7381-6145-7 STD96008
5 February 2010
9. RFC 5304
Li, T., Atkinson, R.
IS-IS Cryptographic Authentication
October 2008
10. RFC 5310
Bhatia, M., Manral, V., Li, T., Atkinson, R., White, R., Fanto, M.
IS-IS Generic Cryptographic Authentication
February 2009
11. IEEE P802.1aq-2012
IEEE Approved Draft Standard for Local and Metropolitan Area Networks:
Bridges and Virtual Bridged Local Area Networks
Amendment 9: Shortest Path Bridging
12. IEEE Std 802.1Q-2011
IEEE Standard for Local and metropolitan area networks
Media Access Control (MAC) Bridges and Virtual Bridge Local Area Networks
Print: ISBN 978-0-7381-6709-1 STDPD97139
PDF: ISBN 978-0-7381-6708-4 STD97139
31 August 2011
13. IEEE 802.1Qbp
Equal Cost Multiple Paths (ECMP)
Work in progress
http://www.ieee802.org/1/pages/802.1bp.html
14. Recommendation ITU-T Y.1731 (02/2008)
OAM functions and mechanisms for Ethernet based networks
29 February 2008
15. http://hub.opensolaris.org/bin/view/Project+rbridges/WebHome
16. http://www.broadcom.com/products/Switching/Data-Center/BCM56840-Series
17. http://www.fulcrummicro.com/product_library/FM6000_Product_Brief.pdf