

INTERNET-DRAFT  
Updates: 6325, 6361, 7173  
Intended status: Proposed Standard  
Expires: April 18, 2016

Donald Eastlake  
Huawei  
Dacheng Zhang  
Alibaba  
October 19, 2015

TRILL: Link Security  
<draft-eastlake-trill-link-security-02.txt>

Abstract

The TRILL protocol supports arbitrary link technologies between TRILL switches, both point-to-point and broadcast links, and supports Ethernet links between edge TRILL switches and end stations. Communications links are constantly under attack by criminals and national intelligence agencies as discussed in RFC 7258. Link security is an important element of security in depth, particularly for links that are not entirely under the physical control of the TRILL network operator or that include device which may have been compromised. This document specifies link security recommendations for TRILL over Ethernet, PPP, and pseudowire links. It updates RFC 6325, RFC 6361, and RFC 7173. It requires that link encryption MUST be implemented and that all TRILL Data packets between TRILL switch ports capable of encryption at line speed MUST default to being encrypted.

[This is a early partial draft.]

Status of This Memo

This Internet-Draft is submitted to IETF in full conformance with the provisions of BCP 78 and BCP 79.

Distribution of this document is unlimited. Comments should be sent to the DNSEXT working group mailing list: <rbridge@postel.org>.

Internet-Drafts are working documents of the Internet Engineering Task Force (IETF), its areas, and its working groups. Note that other groups may also distribute working documents as Internet-Drafts.

Internet-Drafts are draft documents valid for a maximum of six months and may be updated, replaced, or obsoleted by other documents at any time. It is inappropriate to use Internet-Drafts as reference material or to cite them other than as "work in progress."

The list of current Internet-Drafts can be accessed at <http://www.ietf.org/lid-abstracts.html>. The list of Internet-Draft Shadow Directories can be accessed at <http://www.ietf.org/shadow.html>.

## Table of Contents

1. Introduction.....	3
1.1 Encryption Requirement and Adjacency.....	3
1.2 Terminology and Acronyms.....	4
2. Link Security Default Keying.....	5
3. Link Security Specifics.....	6
3.1 Ethernet Links.....	6
3.2 PPP Links.....	8
3.3 Pseudowire Links.....	8
4. Edge-to-Edge Security.....	9
5. Security Considerations.....	11
6. IANA Considerations.....	11
Normative References.....	12
Informative References.....	13
Acknowledgments.....	14
Appendix A: Summary of Changes to RFCs 6325, 6361, 7173...	15
Appendix B: Ethernet Security to End Stations.....	16
Authors' Addresses.....	19

## 1. Introduction

The TRILL (Transparent Interconnection of Lots of Links or Tunneled Routing in the Link Layer) protocol supports arbitrary link technologies including both point-to-point and broadcast links and supports Ethernet links between edge TRILL switches and end stations. Communications links are constantly under attack by criminals and national intelligence agencies as discussed in [RFC7258].

Link security is an important element of security in depth for links, particularly those that are not entirely under the physical control of the TRILL network operator or that include device which may have been compromised, that is, pretty much for all links. TRILL generally uses an existing link security method specified for the technology of the link in question.

This document specifies link security recommendations for TRILL over Ethernet [RFC6325], TRILL over PPP [RFC6361], and transport of TRILL by pseudowires [RFC7173], in Sections 3.1, 3.2, and 3.3 respectively. Although the Security Considerations sections of these RFCs mention link security, this document goes further, updating these RFCs as described in Appendix A and imposing the new mandatory encryption implementation requirements summarized in Section 1.1.

[TRILL-IP] will cover TRILL security over IP links and any other future TRILL-over-X drafts are expected to cover security for TRILL links using technology X.

Edge-to-edge security, from ingress to egress TRILL switch, provides another level of security and is covered in Section 4.

TRILL provides autoconfiguration assistance and default keying material, under most circumstances, to support the TRILL goal of having a minimal or zero configuration default. Where better security is not available, TRILL supports opportunistic security [RFC7435].

[This is a partial early draft.]

### 1.1 Encryption Requirement and Adjacency

This document requires that all TRILL data packets between adjacent TRILL switch ports that are capable of encryption at line speed MUST default to being encrypted and authenticated. It MUST require explicit configuration in such cases for the ports to communicate unencrypted or unsecured. Line speed encryption and authentication usually requires hardware assist but there are cases with slower ports and higher powered switch processors where it can be accomplished in software.

If line speed link encryption and authentication is not available for communication between TRILL switch ports, it MUST still be possible to configure the TRILL switches and ports involved to encrypt and authenticate all TRILL packets sent for cases where the security provided outweighs the reduction in performance.

## 1.2 Terminology and Acronyms

This document uses the acronyms and terms defined in [RFC6325], some of which are repeated below for convenience, and additional acronyms and terms listed below.

HKDF: Hash based Key Derivation Function [RFC5869].

Link: The means by which adjacent TRILL switches are connected. May be various technologies and in the common case of Ethernet, can be a "bridged LAN", that is to say, some combination of Ethernet links with zero or more bridges, hubs, repeaters, or the like.

MACSEC: Media Access Control (MAC) Security. IEEE Std 802.1AE-2006.

MPLS: Multi-Protocol Label Switching.

PPP: Point-to-point protocol [RFC1661].

RBridge: An alternative name for a TRILL switch.

TRILL: Transparent Interconnection of Lots of Links or Tunnelled Routing in the Link Layer.

TRILL switch: A device implementing the TRILL protocol. An alternative name for an RBridge.

The key words "MUST", "MUST NOT", "REQUIRED", "SHALL", "SHALL NOT", "SHOULD", "SHOULD NOT", "RECOMMENDED", "MAY", and "OPTIONAL" in this document are to be interpreted as described in [RFC2119].

## 2. Link Security Default Keying

In some cases, it is possible to use keying material derived from the [RFC5310] IS-IS keying material already in place. In such cases, the two byte [RFC5310] Key ID identifies the IS-IS keying material. The keying material actually used in the link security protocol is derived from the IS-IS keying material as follows:

```
HKDF-Expand-SHA256 ( IS-IS-key, "TRILL Link" | custom, L )
```

where "|" indicates concatenation, HKDF is the Hash base Key Derivation Function in [RFC5869], SHA256 is as in [RFC6234], IS-IS-key is the input keying material, "TRILL Link" is the 10-character ASCII [RFC20] string indicated, "custom" is a byte string dependeng on the link security protocol being used, and L is the length of output keying material needed.

### 3. Link Security Specifics

The following subsection discuss TRILL link security for various technologies.

#### 3.1 Ethernet Links

TRILL over Ethernet is specified in [RFC6325] with some additional material on Ethernet link MTU in [rfc7180bis].

Link security between TRILL switch Ethernet ports conforms to IEEE Std 802.1AE-2006 [802.1AE] as amended by IEEE Std 802.1AEbn-2011 [802.1AEbn] and IEEE Std 802.1AEbw-2013 [802.1AEbw]. This security is referred to as MACSEC.

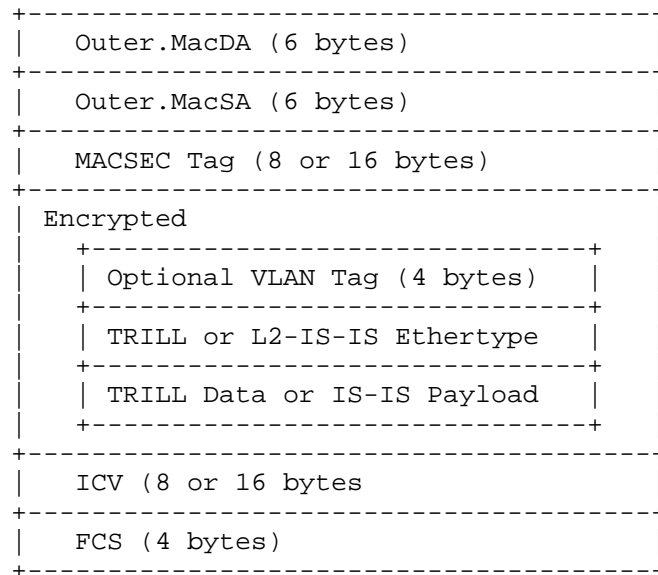
TRILL switch Ethernet ports MUST implement MACSEC even if it is implemented in software. When TRILL switch ports are directly connected by Ethernet with no intervening customer bridges, for example by a point to point Ethernet link, MACSEC between them operates as specified herein. There can be intervening Provider Bridges or other forms of transparent Ethernet tunnels.

However, if there are one or more customer bridges or similar devices in the path, MACSEC at the TRILL switch port will peer with the nearest such bridge port. This results, from the point of view of MACSEC, with a two or more hop path, although it is one TRILL hop. Typically, the TRILL switch ports at the ends of such a path would be unable to negotiate security and agree on keys because of the intervening customer bridge. In such cases where encryption and authentication are required, the adjacent TRILL switch ports would be unable to establish IS-IS communication and would not form an adjacency [RFC7177]. However, it may be possible to configure such bridge ports and distribute such keying material or the like to them so that encryption and authentication can be established on all hops of such multi-hop Ethernet paths. Methods for accomplishing such distribution to devices other than TRILL switches are beyond the scope of this document.

When MACSEC is established between adjacent TRILL switch ports, the frames are as shown in Figure 1. The optional VLAN tagging shown is superfluous in the case of TRILL Data and IS-IS packets. Unless there are VLAN sensitive devices intervening between the TRILL switch ports, or possibly attached to the link between those ports, TRILL Data and IS-IS packets secured with MACSEC SHOULD generally be sent untagged for efficiency.

Of course there may be other Ethernet control frames, such as link aggregation control messages or priority based flow control messages,

that would also be sent within MACSEC. Typically only the [802.1X] messages used to establish and maintain MACSEC are sent unsecured.



Figures 1. MACSEC Between TRILL Switch Ports

Outer.MacDA: 48-bit destination MAC address

Outer.MacSA: 48-bit source MAC address

MACSEC Tag: See further description below.

Encrypted: The encrypted data

ICV: The MACSEC Integrity Check Value

FCS: Frame Check Sequence.

The structure of a MACSEC Tag is as follows:

tbd ...

[802.1X] is used to establish keying and algorithms for Ethernet link security ... tbd ...

### 3.2 PPP Links

TRILL over PPP is specified in [RFC6361]. Currently specified native PPP security does not meet modern security standards. However, true PPP over HDLC is relatively uncommon today and PPP is normally being conveyed by another protocol, such as PPP over Ethernet or PPP over IP. In those cases it is RECOMMENDED that Ethernet security as described in Section 3 or IP security as described in [TRILL-IP] be used to secure PPP between TRILL switch ports.

If it is necessary to use native PPP security [RFC1968] [RFC1994]  
...tbd...

### 3.3 Pseudowire Links

TRILL transport over pseudowires is specified in [RFC7173].

No native security is provided for pseudowires as such; however, they are, by definition, carried by some PSN (Packet Switched Network). Link security must be provided by this PSN or by lower level protocols. This PSN is typically an MPLS or IP PSN.

In the case of a pseudowire over IP, security SHOULD be provided as is expected to be specified in [TRILL-IP]. If that is not possible but the IP path is only one IP hop, then it may be possible to provide link security at the layer of the link protocol supporting that hop, such as Ethernet (Section 3) or PPP (Section 4).

In the case of a pseudowire over MPLS, MPLS also does not have a native security scheme. Thus, security must be provided at the link layer being used, for example Ethernet (Section 3) or IP [TRILL-IP].



#### 4. Edge-to-Edge Security

Edge-to-edge security can be applied to TRILL data packets between the TRILL switch where they are ingressed or created to the TRILL switch where they are egressed or consumed. The edge-to-edge path is viewed as a one hop virtual link from before TRILL encapsulation to after TRILL decapsulation. MACSEC is used on this pseudolink.

If default keying is used, it is as specified in Section 2 above with the value of "custom" in Section 2 as specified below, depending on whether the TRILL data packet is TRILL unicast or TRILL multi-destination:

Unicast: custom = "Uni" | ingress System ID | egress System ID

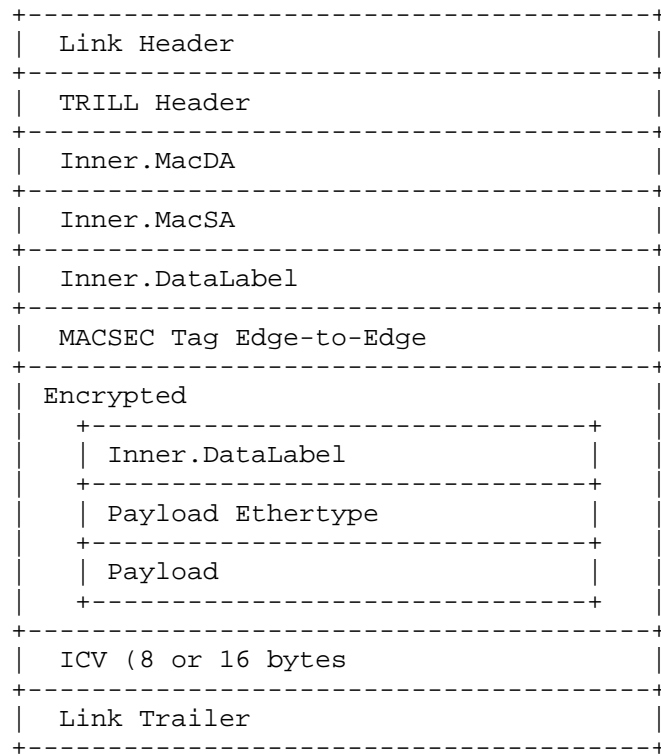
Multi-destination: custom = "Multi" | Data Label

where "|" indicates concatenation, the quoted string "Uni" and "Multi" represent those 3 and 5 character ASCII [RFC20] strings, respectively, ingress System ID and egress System ID are the 6-byte IS-IS System ID of the origin and destination TRILL switches, and Data Label is the contents of the 4-byte (C-VLAN Ethertype plus VLAN ID) or 8-bytes (FGL Ethernets and value) data labeling area of the TRILL packet with priority/DEI fields set to zero.

Where keying is to be negotiated between a pair of TRILL switches for edge-to-edge unicast security, the IEEE 802.1X messages involved are transmitted inside unicast RBridge Channel [RFC7178] messages using RBridge Channel protocol number TBD1. Support for edge-to-edge encryption is indicated by a TRILL switch advertising support for this RBridge Channel protocol. In such 802.1X messages, the System IDs of the TRILL switches are used as their "MAC Addresses". 802.1X in turn uses the Extensible Authentication Protocol (EAP [RFC3748]).

tbd ...

For edge-to-edge security, the MACSEC tag is inserted in the payload frame and the Inner.DataLabel (VLAN or FGL) is duplicated so that a TRILL Data packet on a transit link (which might not be an Ethernet link) is structured as shown below. The unencrypted copy of the Inner.DataLabel is needed for two reasons: (1) to avoid rejection by and transit RBridges the packet passes through that are sensitive to the Ethertype appearing immediately after the Inner.MacSA and would otherwise discard the packet and (2) to assure proper distribution if the packet is multi-destination. The inner encrypt



## 5. Security Considerations

This document is entirely about TRILL link security for Etherent, PPP, and pseudowire TRILL links. See sections of this document on those particular link technologies.

For general TRILL Security Considrations, see [RFC6325].

## 6. IANA Considerations

IANA is requested to allocate a new RBridge Channel protocol number TBD1 for tunneled 802.1X messages supporting negotiated keys for unicast edge-to-edge security.

## Normative References

- [802.1AE] - IEEE Std 802.1AE-2006, IEEE Standard for Local and metropolitan networks / Media Access Control (MAC) Security, 18 August 2006.
- [802.1AEbn] - IEEE Std 802.1AEbn-2011, IEEE Standard for Local and metropolitan networks / Media Access Control (MAC) Security / Galois Counter Mode - Advanced Encryption Standard - 256 (GCM-AES-256) Cipher Suite, 14 October 2011.
- [802.1AEbw] - IEEE Std 802.1AEbw-2014, IEEE Standard for Local and metropolitan networks / Media Access Control (MAC) Security / Extended Packet Numbering, 12 February 2014
- [RFC20] - Cerf, V., "ASCII format for network interchange", STD 80, RFC 20, October 1969, <<http://www.rfc-editor.org/info/rfc20>>.
- [RFC1661] - Simpson, W., Ed., "The Point-to-Point Protocol (PPP)", STD 51, RFC 1661, July 1994, <<http://www.rfc-editor.org/info/rfc1661>>.
- [RFC1968] - Meyer, G., "The PPP Encryption Control Protocol (ECP)", RFC 1968, June 1996, <<http://www.rfc-editor.org/info/rfc1968>>.
- [RFC2119] - Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, March 1997, <<http://www.rfc-editor.org/info/rfc2119>>.
- [RFC5226] - T. Narten and H. Alvestrand, "Guidelines for Writing an IANA Considerations Section in RFCs," BCP 26 and RFC 5226, May 2008
- [RFC5310] - Bhatia, M., Manral, V., Li, T., Atkinson, R., White, R., and M. Fanto, "IS-IS Generic Cryptographic Authentication", RFC 5310, February 2009.
- [RFC5869] - Krawczyk, H. and P. Eronen, "HMAC-based Extract-and-Expand Key Derivation Function (HKDF)", RFC 5869, May 2010, <<http://www.rfc-editor.org/info/rfc5869>>
- [RFC6234] - Eastlake 3rd, D. and T. Hansen, "US Secure Hash Algorithms (SHA and SHA-based HMAC and HKDF)", RFC 6234, May 2011, <<http://www.rfc-editor.org/info/rfc6234>>.
- [RFC6325] - Perlman, R., Eastlake 3rd, D., Dutt, D., Gai, S., and A. Ghanwani, "Routing Bridges (RBridges): Base Protocol Specification", RFC 6325, July 2011, <<http://www.rfc-editor.org/info/rfc6325>>.

- [RFC6361] - Carlson, J. and D. Eastlake 3rd, "PPP Transparent Interconnection of Lots of Links (TRILL) Protocol Control Protocol", RFC 6361, August 2011, <<http://www.rfc-editor.org/info/rfc6361>>.
- [RFC7173] - Yong, L., Eastlake 3rd, D., Aldrin, S., and J. Hudson, "Transparent Interconnection of Lots of Links (TRILL) Transport Using Pseudowires", RFC 7173, May 2014, <<http://www.rfc-editor.org/info/rfc7173>>.
- [RFC7177] - Eastlake 3rd, D., Perlman, R., Ghanwani, A., Yang, H., and V. Manral, "Transparent Interconnection of Lots of Links (TRILL): Adjacency", RFC 7177, May 2014, <<http://www.rfc-editor.org/info/rfc7177>>.
- [RFC7178] - Eastlake 3rd, D., Manral, V., Li, Y., Aldrin, S., and D. Ward, "Transparent Interconnection of Lots of Links (TRILL): RBridge Channel Support", RFC 7178, DOI 10.17487/RFC7178, May 2014, <<http://www.rfc-editor.org/info/rfc7178>>.

#### Informative References

- [RFC1994] - Simpson, W., "PPP Challenge Handshake Authentication Protocol (CHAP)", RFC 1994, August 1996, <<http://www.rfc-editor.org/info/rfc1994>>.
- [RFC3748] - B. Aboba, et al., "Extensible Authentication Protocol (EAP)", RFC 3748, June 2004
- [RFC7258] - Farrell, S. and H. Tschofenig, "Pervasive Monitoring Is an Attack", BCP 188, RFC 7258, May 2014, <<http://www.rfc-editor.org/info/rfc7258>>.
- [RFC7435] - Dukhovni, V., "Opportunistic Security: Some Protection Most of the Time", RFC 7435, December 2014, <<http://www.rfc-editor.org/info/rfc7435>>.
- [rfc7180bis] - Eastlake, D., Zhang, M., Perlman, R. Banerjee, A., Ghanwani, A., and S. Gupta, "TRILL: Clarifications, Corrections, and Updates", draft-ietf-trill-rfc7180bis, work in progress.
- [TRILL-IP] - Cullen, M., et al., "Transparent Interconnection of Lots of Links (TRILL) over IP", draft-ietf-trill-over-ip, work in progress.

#### Acknowledgments

The authors thank the following for their comments and help:

tbd

Appendix A: Summary of Changes to RFCs 6325, 6361, 7173

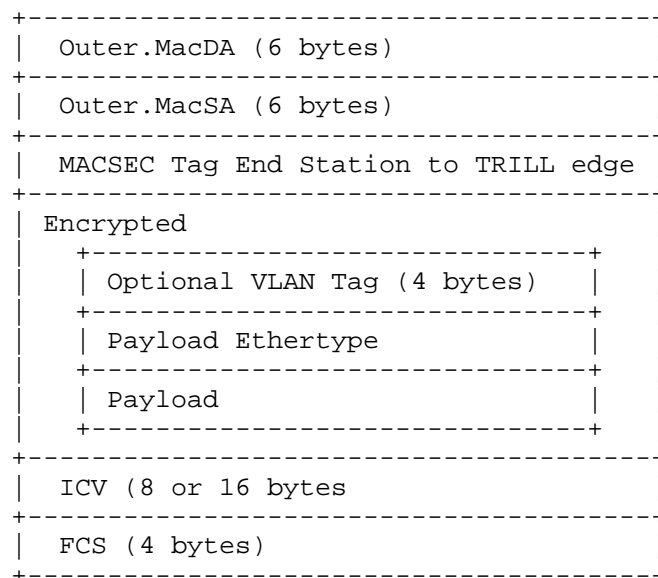
tbd ...

## Appendix B: Ethernet Secrity to End Stations

MACSEC could be used between end stations and their adjacent TRILL switch(es) or end-to-end between end stations or both. Since TRILL does not impose administrative requirements on end stations, the choice of keying and crypto suite are beyond the scope of this document. However, some informative explanation and diagrams are provided below to clarify how this might be done.

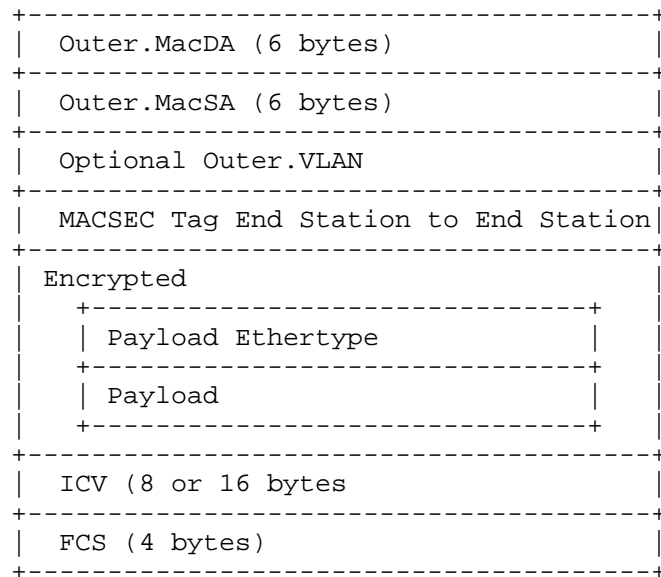
The end station must be properly configured to know if it should apply MACSEC to secure its connection to an edge TRILL switch or to remote end stations or both.

The Figure below show an Ethernet frame between a end station and the adjacent edge RBridge secured by MACSEC.

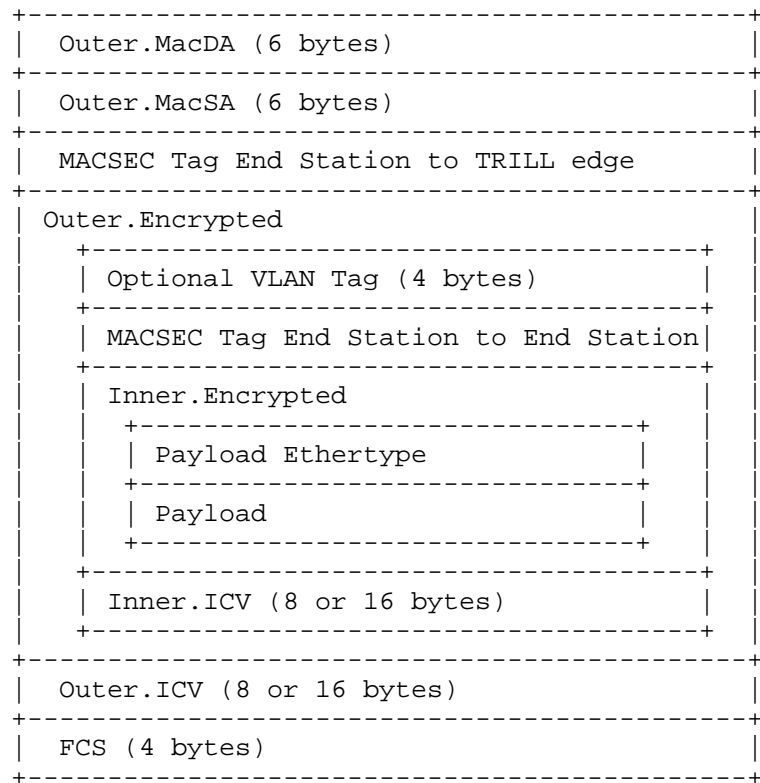


The Figure below shows an Ethernet frame between an end station and an adjacent edge RBridge where MACSEC is being used end-to-end between that end station and remote end stations.





The Figure below shows an Ethernet frame between an end station and an adjacent edge RBridge where MACSEC is being used end-to-end between that end station and a remote end stations and, in addition, an outer application of MACSEC is securing traffic between the end station and the adjacent edge RBridge port.



Authors' Addresses

Donald Eastlake, 3rd  
Huawei Technologies  
155 Beaver Street  
Milford, MA 01757 USA

Phone: +1-508-333-2270  
Email: d3e3e3@gmail.com

Dacheng Zhang  
Alibaba  
Beijing, Chao yang District  
P.R. China

Email: dacheng.zdc@alibaba-inc.com

## Copyright and IPR Provisions

Copyright (c) 2015 IETF Trust and the persons identified as the document authors. All rights reserved.

This document is subject to BCP 78 and the IETF Trust's Legal Provisions Relating to IETF Documents (<http://trustee.ietf.org/license-info>) in effect on the date of publication of this document. Please review these documents carefully, as they describe your rights and restrictions with respect to this document. Code Components extracted from this document must include Simplified BSD License text as described in Section 4.e of the Trust Legal Provisions and are provided without warranty as described in the Simplified BSD License. The definitive version of an IETF Document is that published by, or under the auspices of, the IETF. Versions of IETF Documents that are published by third parties, including those that are translated into other languages, should not be considered to be definitive versions of IETF Documents. The definitive version of these Legal Provisions is that published by, or under the auspices of, the IETF. Versions of these Legal Provisions that are published by third parties, including those that are translated into other languages, should not be considered to be definitive versions of these Legal Provisions. For the avoidance of doubt, each Contributor to the IETF Standards Process licenses each Contribution that he or she makes as part of the IETF Standards Process to the IETF Trust pursuant to the provisions of RFC 5378. No language to the contrary, or terms, conditions or rights that differ from or are inconsistent with the rights and licenses granted under RFC 5378, shall have any effect and shall be null and void, whether published or posted by such Contributor, or included with or in such Contribution.



TRILL Working Group  
INTERNET-DRAFT  
Intended status: Proposed Standard  
Expires: April 17, 2015

Weiguo Hao  
Donald Eastlake  
Huawei  
October 18, 2015

TRILL: Address Flush Protocol  
<draft-hao-trill-address-flush-00.txt>

## Abstract

The TRILL (TRAnsparent Interconnection of Lots of Links) protocol, by default, learns end station addresses from observing the data plane. This document specifies an optional message by which an originating TRILL switch can explicitly flush addresses learned by other TRILL switches through the egress of data ingress by that originating TRILL switch. This is a supplement to the TRILL automatic address forgetting and can assist in achieving more rapid convergence.

## Status of This Memo

This Internet-Draft is submitted to IETF in full conformance with the provisions of BCP 78 and BCP 79.

Distribution of this document is unlimited. Comments should be sent to the TRILL working group mailing list.

Internet-Drafts are working documents of the Internet Engineering Task Force (IETF), its areas, and its working groups. Note that other groups may also distribute working documents as Internet-Drafts.

Internet-Drafts are draft documents valid for a maximum of six months and may be updated, replaced, or obsoleted by other documents at any time. It is inappropriate to use Internet-Drafts as reference material or to cite them other than as "work in progress."

The list of current Internet-Drafts can be accessed at <http://www.ietf.org/lid-abstracts.html>. The list of Internet-Draft Shadow Directories can be accessed at <http://www.ietf.org/shadow.html>.

## Table of Contents

1. Introduction.....	3
1.1 Terminology and Acronyms.....	3
2. Address Flush Message Details.....	5
3. IANA Considerations.....	9
4. Security Considerations.....	9
Normative References.....	10
Informative References.....	10
Acknowledgements.....	10
Authors' Addresses.....	11

## 1. Introduction

Edge TRILL (Transparent Interconnection of Lots of Links [RFC6325]) switches, also called RBridges, by default learn end station MAC addresses from observing the data plane. On receipt of a native frame from an end station, they would learn the local MAC address attachment of the source end station. And on egressing (decapsulating) a remotely originated TRILL Data frame, they learn the remote MAC address and remote attachment TRILL switch. Such learning is all appropriately scoped by data label (VLAN or Fine Grained Label [RFC7172]).

TRILL has mechanisms for timing out such learning and appropriately clearing it based on some network connectivity changes; however, there are circumstances under which it would be helpful for a TRILL switch to be able to explicitly flush (clear) learned end station reachability information to achieve more rapid convergence (see, for example, Section 6.2 of [RFC4762]). Obviously a TRILL switch R1 can easily flush any locally learned addresses it wants. This document specifies an optional message to request flushing such learned address information at remote TRILL switches. This Address Flush message makes use of the RBridge Channel facility [RFC7178], which supports typed message transmission between RBridges.

### 1.1 Terminology and Acronyms

The key words "MUST", "MUST NOT", "REQUIRED", "SHALL", "SHALL NOT", "SHOULD", "SHOULD NOT", "RECOMMENDED", "MAY", and "OPTIONAL" in this document are to be interpreted as described in [RFC2119].

This document uses the terms and acronyms defined in [RFC6325] and [RFCchannel] as well as the following:

AFN - Address Family Number ([RFC4760] where it is called Address Family Identifier (AFI)).

FGL - Fine Grained Label [RFC7172].

Management VLAN - A VLAN in which all TRILL switches in a campus indicate interest so that multi-destination TRILL Data packets, including RBridge Channel messages [RFCchannel], sent with that VLAN as the Inner.VLAN will be delivered to all TRILL switches in the campus. Usually no end station service is offered in the Management VLAN.

RBridge - A alternative name for a TRILL switch.

TRILL switch - A device implementing the TRILL protocol.



Edge TRILL switch - A TRILL switch attached to one or more links that provide end station service.

## 2. Address Flush Message Details

The Address Flush message makes use of the RBridge Channel protocol [RFC7178].

Although initial use is expected to be to flush 48-bit MAC addresses [RFC7042], the protocol accommodates flushing other types of end station addresses; there have been suggestion for TRILL switches to learn IP addresses from the data plane [INFOCOM], TRILL might be extended to accommodate 64-bit MAC addresses, or similar future extensions might benefit from the ability to flush other types of learned addresses.

The general structure of an RBridge Channel packet on a link between TRILL switches is shown in Figure 1 below. The type of RBridge Channel packet is given by a Protocol field in the RBridge Channel Header that indicates how to interpret the Channel Protocol Specific Payload [RFC7178].

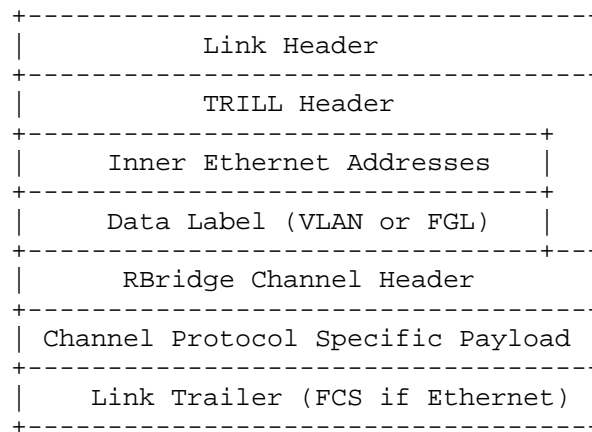


Figure 1. RBridge Channel Packet Structure

An Address Flush RBridge Channel message normally applies to addresses within the VLAN or FGL [RFC7178] Data Label in the TRILL Header. Address Flush protocol messages are usually sent as multi-destination packets (TRILL Header M bit equal to one) so as to reach all TRILL switches offering end station service in the VLAN or FGL specified by the Data Label. However, an address flush protocol message can be sent unicast, if it is desired to clear addresses at one TRILL switch only. And there are provisions for indicating the Data Label with the address(es) to be flushed for cases where the address flush protocol message is sent over a Management VLAN or the like.

Figure 2 below expands the RBridge Channel Header and Channel

Protocol Specific Payload from Figure 1 for the case of the Address Flush message.

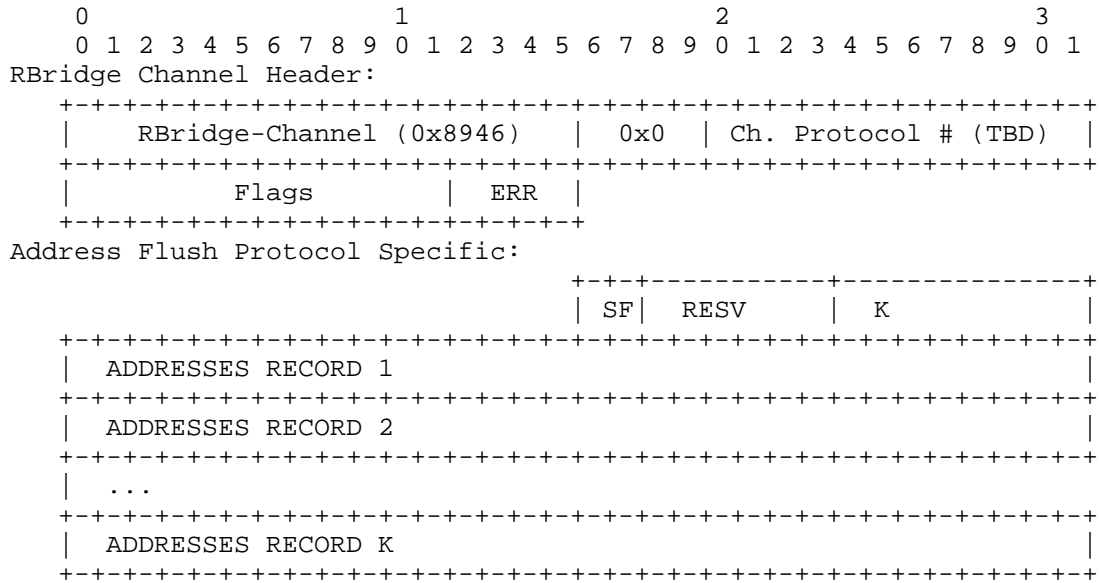


Figure 2. Address Flush Channel Message Structure

The fields in Figure 2 related to the Address Flush message are as follows:

Channel Protocol: The RBridge Channel Protocol value allocated for Address Flush (see Section 3).

SF: The 2-bit SF ("super flush") field values have the following meanings:

0: No special effect.

1: All addresses learned at the receiving TRILL switch due to egressing TRILL Data packets from the TRILL switch originating this Address Flush message are flushed for the data label in the TRILL Header. Any ADDRESS RECORDs in the rest of the message for that data label can be ignored but there may be ADDRESS RECORDs present that apply to other data labels.

2: All addresses learned at the receiving TRILL switch due to egressing TRILL Data packets from the TRILL switch originating this Address Flush message are flushed across all data labels. The remainder of the Address Flush message, including the value of K, are ignored.

3: Reserved. Ignored on receipt.

RESV: 4 reserved flag bits. Must be sent as zero and ignored on receipt.

K: The number of ADDRESS RECORDs present. See below.

The structure of the ADDRESSES RECORD is as follows:

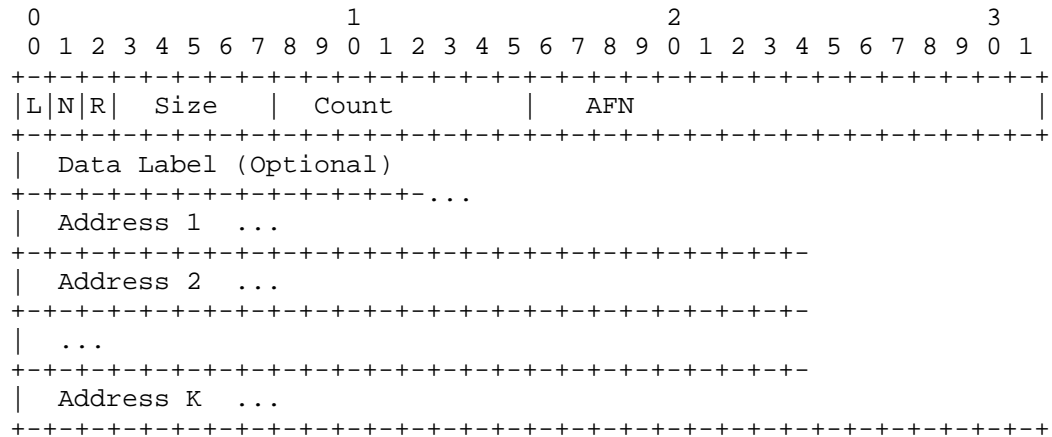


Figure 3. Structure of ADDRESSES RECORD

L: Label present. If this bit is a one, the optional Data Label shown in present. If it is zero, there is no data label and the addresses listed are within the data label given in the TRILL Header.

N: No Specific Addresses. If this bit is one and Count is zero and L is one, then flush all addresses learned at the receiving TRILL switch due to egressing TRILL Data packets from the TRILL switch originating this Address Flush message are flushed for the Data Label given in the ADDRESS RECORD. If this bit is zero or Count is non-zero or L is zero, then this special flush action is not performed.

R: A reserved bit that MUST be sent as zero and is ignored on receipt.

Size: The size of each Address in bytes. The presence of this field makes it possible for a receiving TRILL switch to skip an ADDRESS RECORD even if it does not understand the value in the AFN field. Size MUST NOT be zero; a zero size field indicates a corrupt Addresses Flush message and the entire message is ignored. MUST be the correct size for an Address

of the type indicated by the AFN field, for example 6 for 48-bit MAC addresses. If these conditions are violated, the Address Flush message is discarded.

Count: The number of occurrences of an Address to flush in this ADDRESS RECORD. May be zero. All Addresses MUST fit within the RBridge Channel Message. If they do not, the message is discarded.

AFN: The Address Family Number for the type of addresses present as assigned by IANA. (The AFN for 48-bit MAC addresses is 0x4005.)

Data Label: An optional Data Label (VLAN or FGL) in the same format as Data Labels that appear in the TRILL Header. Included in an ADDRESS RECORD only if the L bit is a one.

Address: An instance of an address to be flushed.

### 3. IANA Considerations

IANA has allocated tbd1 for the Address Flush RBridge Channel Protocol number from the range of RBridge Channel protocols allocated by Standards Action [RFC7178].

### 4. Security Considerations

The Address Flush RBridge Channel Protocol provides no security assurances or features. However, use of the Address Flush protocol can be nested inside the RBridge Channel Tunnel Protocol [RFCtunnel] using the RBridge Channel message payload type. The Channel Tunnel protocol can provide some security services.

See [RFC7178] for general RBridge Channel Security Considerations.

See [RFC6325] for general TRILL Security Considerations.

## Normative References

- [RFC2119] - Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, March 1997.
- [RFC4760] - Bates, T., Chandra, R., Katz, D., and Y. Rekhter, "Multiprotocol Extensions for BGP-4", RFC 4760, January 2007.
- [RFC6325] - Perlman, R., D. Eastlake, D. Dutt, S. Gai, and A. Ghanwani, "RBriges: Base Protocol Specification", RFC 6325, July 2011.
- [RFC7172] - Eastlake 3rd, D., Zhang, M., Agarwal, P., Perlman, R., and D. Dutt, "Transparent Interconnection of Lots of Links (TRILL): Fine-Grained Labeling", RFC 7172, DOI 10.17487/RFC7172, May 2014, <<http://www.rfc-editor.org/info/rfc7172>>.
- [RFC7178] - Eastlake 3rd, D., Manral, V., Li, Y., Aldrin, S., and D. Ward, "Transparent Interconnection of Lots of Links (TRILL): RBridge Channel Support", RFC 7178, DOI 10.17487/RFC7178, May 2014, <<http://www.rfc-editor.org/info/rfc7178>>.

## Informative References

- [INFOCOM] - Perlman, R., "RBriges: Transparent Routing", Proc. Infocom 2005, March 2004.
- [RFC4762] - Lasserre, M., Ed., and V. Kompella, Ed., "Virtual Private LAN Service (VPLS) Using Label Distribution Protocol (LDP) Signaling", RFC 4762, January 2007.
- [RFC7042] - Eastlake 3rd, D. and J. Abley, "IANA Considerations and IETF Protocol and Documentation Usage for IEEE 802 Parameters", BCP 141, RFC 7042, DOI 10.17487/RFC7042, October 2013, <<http://www.rfc-editor.org/info/rfc7042>>.
- [RFCtunnel] - Eastlake, D., ... "TRILL: Channel Tunnel", draft-eastlake-trill-channel-tunnel, work in progress.

## Acknowledgements

The document was prepared in raw nroff. All macros used were defined within the source file.

Authors' Addresses

Weiguo Hao  
Huawei Technologies  
101 Software Avenue,  
Nanjing 210012, China

Phone: +86-25-56623144  
Email: haoweiguo@huawei.com

Donald E. Eastlake, 3rd  
Huawei Technologies  
155 Beaver Street  
Milford, MA 01757 USA

Phone: +1-508-333-2270  
EMail: d3e3e3@gmail.com



## Copyright, Disclaimer, and Additional IPR Provisions

Copyright (c) 2015 IETF Trust and the persons identified as the document authors. All rights reserved.

This document is subject to BCP 78 and the IETF Trust's Legal Provisions Relating to IETF Documents (<http://trustee.ietf.org/license-info>) in effect on the date of publication of this document. Please review these documents carefully, as they describe your rights and restrictions with respect to this document. Code Components extracted from this document must include Simplified BSD License text as described in Section 4.e of the Trust Legal Provisions and are provided without warranty as described in the Simplified BSD License. The definitive version of an IETF Document is that published by, or under the auspices of, the IETF. Versions of IETF Documents that are published by third parties, including those that are translated into other languages, should not be considered to be definitive versions of IETF Documents. The definitive version of these Legal Provisions is that published by, or under the auspices of, the IETF. Versions of these Legal Provisions that are published by third parties, including those that are translated into other languages, should not be considered to be definitive versions of these Legal Provisions. For the avoidance of doubt, each Contributor to the IETF Standards Process licenses each Contribution that he or she makes as part of the IETF Standards Process to the IETF Trust pursuant to the provisions of RFC 5378. No language to the contrary, or terms, conditions or rights that differ from or are inconsistent with the rights and licenses granted under RFC 5378, shall have any effect and shall be null and void, whether published or posted by such Contributor, or included with or in such Contribution.



TRILL Working Group  
INTERNET-DRAFT  
Intended Status: Standard Track

Y. Li  
D. Eastlake  
L. Dunbar  
Huawei Technologies  
R. Perlman  
EMC  
I. Gashinsky  
Yahoo  
October 14, 2015

Expires: April 16, 2016

TRILL: ARP/ND Optimization  
draft-ietf-trill-arp-optimization-01

Abstract

This document describes mechanisms to optimize the ARP (Address Resolution Protocol) and ND (Neighbor Discovery) traffic in TRILL campus. Such optimization reduces packet flooding over a TRILL campus.

Status of this Memo

This Internet-Draft is submitted to IETF in full conformance with the provisions of BCP 78 and BCP 79.

Internet-Drafts are working documents of the Internet Engineering Task Force (IETF), its areas, and its working groups. Note that other groups may also distribute working documents as Internet-Drafts.

Internet-Drafts are draft documents valid for a maximum of six months and may be updated, replaced, or obsoleted by other documents at any time. It is inappropriate to use Internet-Drafts as reference material or to cite them other than as "work in progress."

The list of current Internet-Drafts can be accessed at  
<http://www.ietf.org/lid-abstracts.html>

The list of Internet-Draft Shadow Directories can be accessed at  
<http://www.ietf.org/shadow.html>

Copyright and License Notice

Copyright (c) 2015 IETF Trust and the persons identified as the

document authors. All rights reserved.

This document is subject to BCP 78 and the IETF Trust's Legal Provisions Relating to IETF Documents (<http://trustee.ietf.org/license-info>) in effect on the date of publication of this document. Please review these documents carefully, as they describe your rights and restrictions with respect to this document. Code Components extracted from this document must include Simplified BSD License text as described in Section 4.e of the Trust Legal Provisions and are provided without warranty as described in the Simplified BSD License.

## Table of Contents

1	Introduction . . . . .	3
1.1	Terminology . . . . .	3
2	IP/MAC Address Mappings . . . . .	4
3	Handling ARP/ND Messages . . . . .	4
3.1	Get Sender's IP/MAC Mapping Information for Non-zero IP . .	5
3.2	Determine How to Reply to ARP/ND . . . . .	5
3.3	Determine How to Handle the ARP/ND Response . . . . .	7
4	Handling RARP (Reverse Address Resolution Protocol) Messages . .	7
5	Security Considerations . . . . .	7
6	IANA Considerations . . . . .	8
7	References . . . . .	8
7.1	Normative References . . . . .	8
7.2	Informative References . . . . .	9
	Authors' Addresses . . . . .	9

## 1 Introduction

ARP [RFC826] and ND [RFC4861] are normally sent by broadcast and multicast respectively. To reduce the burden on a TRILL campus caused by these multi-destination messages, RBridges MAY implement an "optimized ARP/ND response", as specified herein, when the target's location is known by the ingress RBridge or can be obtained from a directory. This avoids ARP/ND query flooding.

### 1.1 Terminology

The acronyms and terminology in [RFC6325] are used herein. Some of these are listed below for convenience with the following along with some additions:

Campus: a TRILL network consisting of TRILL switches, links, and possibly bridges bounded by end stations and IP routers. For TRILL, there is no "academic" implication in the name "campus".

APPsub-TLV	Application sub-Type-Length-Values
ARP	Address Resolution Protocol [RFC826]
DAD	Duplicate Address Detection
Data Label	VLAN or FGL
ESADI	End Station Address Distribution Information [RFC7357]
FGL	Fine-Grained Label [RFC7172]
IA	Interface Addresses, a TRILL APPsub-TLV [IA-draft]
IP	Internet Protocol
MAC	Media Access Control address
ND	Neighbor Discovery [RFC4861]
RBridge switch.	Routing Bridge, an alternative term for a TRILL
SEND	secure neighbor discovery [RFC3971]
TRILL	Transparent Interconnection of Lots of Links or

Tunneled Routing in the Link Layer.

TRILL switch A device implementing the TRILL protocol, an alternative term for an RBridge.

## 2 IP/MAC Address Mappings

Traditionally an RBridge learns the MAC and Data Label (VLAN or FGL) to nickname correspondence of a remote host, as per [RFC6325] and [RFC7172], from TRILL data frames received. No IP address information is learned directly from the TRILL data frame. Interface Addresses (IA) APPsub-TLV [IA-draft] enhances the TRILL base protocol by allowing IP and MAC address mappings to be distributed in the control plane by any RBridge. This APPsub-TLV appears inside the TRILL GENINFO TLV in ESADI [RFC7357] but the value data structure it specifies may also occur in other application contexts. Edge Directory Assist Mechanisms [DirMech] makes use of this APPsub-TLV for its push model and uses the value data structure it specifies in its pull model.

An RBridge can easily know the IP/MAC address mappings of the local hosts that it is attached to it via its access ports by receiving ARP [RFC826] or ND [RFC4861] messages. If the RBridge has extracted the sender's IP/MAC address pair from the received data packet, it may save the information and use the IA APPsub-TLV to distribute it to other RBridges through ESADI. Then the relevant remote RBridges (normally those interested in the same Data Label as the original ARP/ND messages) receive and save such mapping information also. There are others ways that RBridges save IP/MAC address mappings in advance, e.g. import from management system and distribution by directory servers [DirMech].

The examples given above shows that RBridges may have saved a host's triplet of {IP address, MAC address, ingress nickname} for a given Data Label (VLAN or FGL) before that host sends or receives any real data packet. Note such information may or may not be a complete list and may or may not exist on all RBridges. The information may possibly be from different sources. RBridges can then use the Flags Field in IA APPsub-TLV to identify if the source is a directory server or local observation by the sender. A different confidence level may also be used to indicate the reliability of the mapping information.

## 3 Handling ARP/ND Messages

A native frame that is an ARP [RFC826] message is detected by its Ethertype of 0x0806. A native frame that is an ND [RFC4861] is

detected by being one of five different ICMPv6 packet types. ARP/ND is commonly used on a link to (1) query for the MAC address corresponding to an IPv4 or IPv6 address, (2) test if an IPv4/IPv6 address is already in use, or (3) to announce the new or updated info on any of IPv4/IPv6 address, MAC address, and/or point of attachment.

To simplify the text, we use the following terms in this section.

- 1) IP address - indicated protocol address that is normally an IPv4 address in ARP or an IPv6 address in ND.
- 2) sender's IP/MAC address - sender protocol/hardware address in ARP, source IP address and source link-layer address in ND
- 3) target's IP/MAC address - target protocol/hardware address in ARP, target address and target link-layer address in ND

When an ingress RBridge receives an ARP/ND message, it can perform the steps described in the sub-sections below.

### 3.1 Get Sender's IP/MAC Mapping Information for Non-zero IP

If the sender's IP has not been saved by the ingress RBridge before, populate the information of sender's IP/MAC in its ARP table;

else if the sender's IP has been saved before but with a different MAC address mapped or a different ingress nickname associated with the same pair of IP/MAC, the RBridge should verify if a duplicate IP address has already been in use or a host has changed its attaching RBridge. The RBridge may use different strategies to do so, for example, ask an authoritative entity like directory servers or encapsulate and unicast the ARP/ND message to the location where it believes the address is in use. RBridge should update the saved triplet of {IP address, MAC address, ingress nickname} based on the verification.

The ingress RBridge may use the IA APPsub-TLV [IA-draft] with the Local flag set in ESADI [RFC7357] to distribute any new or updated triplet of {IP address, MAC address, ingress nickname} information obtained in this step. If a push directory server is used, such information can be distributed as per [DirMech].

### 3.2 Determine How to Reply to ARP/ND

- a) If the message is a generic ARP/ND request and the ingress RBridge knows the target's IP address, the ingress RBridge may decide to take

one or a combination of the following actions:

a.1. Send an ARP/ND response directly to the querier, with the target's MAC address, as believed by the ingress RBridge.

a.2. Encapsulate the ARP/ND request to the target's Designated RBridge, and have the egress RBridge for the target forward the query to the target. This behavior has the advantage that a response to the request is authoritative. If the request does not reach the target, then the querier does not get a response.

a.3. Block ARP/ND requests that occur for some time after a request to the same target has been launched, and then respond to the querier when the response to the recently-launched query to that target is received.

a.4. Pull the most up-to-date records if a pull directory server is available [DirMech] and reply to the querier.

a.5. Flood the request as per [RFC6325].

b) If the message is a generic ARP request and the ingress RBridge does not know target's IP address, the ingress RBridge may take one of the following actions.

b.1. Flood the message as per [RFC6325].

b.2. Use directory server to pull the information [DirMech] and reply to the querier.

b.3. Drop the message.

c) If the message is a gratuitous ARP which can be identified by the same sender's and target's "protocol" address fields or an Unsolicited Neighbor Advertisements [RFC4861] in ND:

The RBridge may use an IA APPsub-TLV [IA-draft] with the Local flag set to distribute the sender's MAC and IP mapping information. When one or more directory servers are deployed and complete Push Directory information is used by all the TRILL switches in the Data Label, a gratuitous ARP or unsolicited NA SHOULD be discarded rather than ingressed. Otherwise, they are either ingressed and flooded as per [RFC6325] or discarded depending on local policy.

d) If the message is a Address Probe ARP Query [RFC5227] which can be identified by the sender's protocol (IPv4) address field being zero and the target's protocol address field being the IPv4 address to be



tested or a Neighbor Solicitation for DAD (Duplicate Address Detection) which has the unspecified source address [RFC4862]: it should be handled as the generic ARP message as in a) and b).

It should be noted in the case of secure neighbor discovery (SEND) [RFC3971], cryptography might prevent local reply by the ingress RBridge, since the RBridge would not be able to sign the response with the target's private key.

It is not essential that all RBridges use the same strategy for which option to select for a particular ARP/ND query. It is up to the implementation.

### 3.3 Determine How to Handle the ARP/ND Response

If the ingress RBridge R1 decides to unicast the ARP/ND request to the target's egress RBridge R2 as discussed in subsection 3.2 item a) or to flood the request as per [RFC6325], then R2 decapsulates the query, and initiates an ARP/ND query on the target's link. When/if the target responds, R2 encapsulates and unicasts the response to R1, which decapsulates the response and sends it to the querier. R2 should initiate a link state update to inform all the other RBridges of the target's location, layer 3 address, and layer 2 address, in addition to forwarding the reply to the querier. The update message can be carried by an IA APPsub-TLV [IA-draft] with the Local flag set in ESADI [RFC7357] or as per [DirMech] if push directory server is in use.

## 4 Handling RARP (Reverse Address Resolution Protocol) Messages

RARP [RFC903] uses the same packet format as ARP but a different Ethertype (0x8035) and opcode values. Its use is similar to the generic ARP Request/Response as described in 3.2 a) and b). The difference is that it is intended to query for the target "protocol" address corresponding to the target "hardware" address provided. It should be handled by doing a local cache or directory server lookup on the target "hardware" address provided to find a mapping to the desired "protocol" address. Normally, it is used to look up a MAC address to find the corresponding IP address.

## 5 Security Considerations

ARP and ND messages can be easily forged. Therefore the learning of MAC/IP addresses from them should not be considered as reliable. RBridge can use the confidence level in IA APPsub-TLV information received via ESADI or pull directory retrievals to determine the reliability of MAC/IP address mapping. (ESADI information can be

secured as provide in [RFC7357] and pull directory information can be secured as provide in [DirMech].) It is up to the implementation to decide if an RBridge should distribute the IP and MAC address mappings received from local native ARP/ND messages to other RBridges in the same Data Label.

The ingress RBridge should also rate limit the ARP/ND queries for the same target to be injected into the TRILL campus to prevent possible denial of service attacks.

## 6 IANA Considerations

No IANA action is required. RFC Editor: please delete this section before publication.

## 7 References

### 7.1 Normative References

- [RFC826] Plummer, D., "An Ethernet Address Resolution Protocol", RFC 826, November 1982.
- [RFC903] Finlayson, R., Mann, T., Mogul, J., and M. Theimer, "A Reverse Address Resolution Protocol", STD 38, RFC 903, June 1984
- [RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, March 1997.
- [RFC4861] Narten, T., Nordmark, E., Simpson, W., and H. Soliman, "Neighbor Discovery for IP version 6 (IPv6)", RFC 4861, September 2007.
- [RFC4862] Thomson, S., Narten, T., and T. Jinmei, "IPv6 Stateless Address Autoconfiguration", RFC 4862, September 2007.
- [RFC6165] Banerjee, A. and D. Ward, "Extensions to IS-IS for Layer-2 Systems", RFC 6165, April 2011.
- [RFC6325] Perlman, R., et.al. "RBridge: Base Protocol Specification", RFC 6325, July 2011.
- [RFC6439] Eastlake, D. et.al., "RBridge: Appointed Forwarder", RFC 6439, November 2011.

- [RFC7172] Eastlake 3rd, D., Zhang, M., Agarwal, P., Perlman, R., and D. Dutt, "Transparent Interconnection of Lots of Links (TRILL): Fine-Grained Labeling", RFC 7172, May 2014, <<http://www.rfc-editor.org/info/rfc7172>>.

## 7.2 Informative References

- [RFC3971] Arkko, J., Ed., Kempf, J., Zill, B., and P. Nikander, "SEcure Neighbor Discovery (SEND)", RFC 3971, March 2005.
- [RFC5227] Cheshire, S., "IPv4 Address Conflict Detection", RFC 5227, July 2008.
- [RFC7067] Dunbar, L., Eastlake 3rd, D., Perlman, R., and I. Gashinsky, "Directory Assistance Problem and High-Level Design Proposal", RFC 7067, November 2013.
- [IA-draft] Eastlake, D., Li Y., R. Perlman, "TRILL: Interface Addresses APPsub-TLV", draft-eastlake-trill-ia-appsubtlv, work in progress.
- [DirMech] Dunbar, L., Eastlake 3rd, D., Perlman, R., I. Gashinsky, and Li Y., "TRILL: Edge Directory Assist Mechanisms", draft-ietf-trill-directory-assist-mechanisms, work in progress.

## Authors' Addresses

Yizhou Li  
Huawei Technologies  
101 Software Avenue,  
Nanjing 210012  
China

Phone: +86-25-56625375  
EMail: [liyizhou@huawei.com](mailto:liyizhou@huawei.com)

Donald Eastlake  
Huawei R&D USA  
155 Beaver Street  
Milford, MA 01757 USA

Phone: +1-508-333-2270  
EMail: [d3e3e3@gmail.com](mailto:d3e3e3@gmail.com)

Linda Dunbar

Huawei Technologies  
5430 Legacy Drive, Suite #175  
Plano, TX 75024, USA

Phone: +1-469-277-5840  
EMail: ldunbar@huawei.com

Radia Perlman  
EMC  
2010 256th Avenue NE, #200  
Bellevue, WA 98007  
USA

EMail: Radia@alum.mit.edu

Igor Gashinsky  
Yahoo  
45 West 18th Street 6th floor  
New York, NY 10011 USA

EMail: igor@yahoo-inc.com

INTERNET-DRAFT  
Intended status: Proposed Standard

Donald Eastlake  
Linda Dunbar  
Huawei  
Radia Perlman  
EMC  
Igor Gashinsky  
Yahoo  
Yizhou Li  
Huawei  
June 20, 2015

Expires: December 19, 2015

TRILL: Edge Directory Assist Mechanisms  
<draft-ietf-trill-directory-assist-mechanisms-03.txt>

#### Abstract

This document describes mechanisms for providing directory service to TRILL (Transparent Interconnection of Lots of Links) edge switches. The directory information provided can be used in reducing multi-destination traffic, particularly ARP/ND and unknown unicast flooding. It can also be used to detect traffic with forged source addresses.

#### Status of This Memo

This Internet-Draft is submitted to IETF in full conformance with the provisions of BCP 78 and BCP 79.

Distribution of this document is unlimited. Comments should be sent to the TRILL working group mailing list.

Internet-Drafts are working documents of the Internet Engineering Task Force (IETF), its areas, and its working groups. Note that other groups may also distribute working documents as Internet-Drafts.

Internet-Drafts are draft documents valid for a maximum of six months and may be updated, replaced, or obsoleted by other documents at any time. It is inappropriate to use Internet-Drafts as reference material or to cite them other than as "work in progress."

The list of current Internet-Drafts can be accessed at <http://www.ietf.org/lid-abstracts.html>. The list of Internet-Draft Shadow Directories can be accessed at <http://www.ietf.org/shadow.html>.

## Table of Contents

1. Introduction.....	3
1.1 Uses of Directory Information.....	3
1.2 Terminology.....	4
2. Push Model Directory Assistance Mechanisms.....	6
2.1 Requesting Push Service.....	6
2.2 Push Directory Servers.....	6
2.3 Push Directory Server State Machine.....	7
2.3.1 Push Directory States.....	8
2.3.2 Push Directory Events and Conditions.....	9
2.3.3 State Transition Diagram and Table.....	10
2.4 Additional Push Details.....	12
2.5 Primary to Secondary Server Push Service.....	13
3. Pull Model Directory Assistance Mechanisms.....	14
3.1 Pull Directory Message Common Format.....	15
3.2 Pull Directory Query and Response Messages.....	16
3.2.1 Pull Directory Query Message Format.....	16
3.2.2 Pull Directory Responses.....	19
3.2.2.1 Pull Directory Response Message Format.....	19
3.2.2.2 Pull Directory Forwarding.....	21
3.3 Cache Consistency.....	22
3.3.1 Update Message Format.....	25
3.3.2 Acknowledge Message Format.....	26
3.4 Summary of Records Formats in Messages.....	26
3.5 Pull Directory Hosted on an End Station.....	27
3.6 Pull Directory Message Errors.....	28
3.6.1 Error Codes.....	29
3.6.2 Sub-Errors Under Error Codes 1 and 3.....	30
3.6.3 Sub-Errors Under Error Codes 128 and 131.....	30
3.7 Additional Pull Details.....	31
3.8 The No Data Flag.....	31
4. Directory Use Strategies and Push-Pull Hybrids.....	33
5. Security Considerations.....	35
6. IANA Considerations.....	36
6.1 ESADI-Parameter Data Extensions.....	36
6.2 RBridge Channel Protocol Number.....	37
6.3 The Pull Directory (PUL) and No Data (NOD) Bits.....	37
6.4 TRILL Pull Directory QTYPES.....	37
6.5 Pull Directory Error Code Registries.....	38
Normative References.....	39
Informational References.....	40
Acknowledgments.....	41
Authors' Addresses.....	42

## 1. Introduction

[RFC7067] gives a problem statement and high level design for using directory servers to assist TRILL [RFC6325] edge nodes in reducing multi-destination ARP/ND [ARPreduction], reducing unknown unicast flooding traffic, and improving security against address spoofing within a TRILL campus. Because multi-destination traffic becomes an increasing burden as a network scales up in number of nodes, reducing ARP/ND and unknown unicast flooding improves TRILL network scalability. This document describes specific mechanisms for directory servers to assist TRILL edge nodes. These mechanisms are optional to implement.

The information held by the Directory(s) is address mapping and reachability information. Most commonly, what MAC address [RFC7042] corresponds to an IP address within a Data Label (VLAN or FGL (Fine Grained Label [RFC7172])) and the egress TRILL switch (RBridge), and optionally what specific TRILL switch port, from which that MAC address is reachable. But it could be what IP address corresponds to a MAC address or possibly other address mappings or reachability.

In the data center environment, it is common for orchestration software to know and control where all the IP addresses, MAC addresses, and VLANs/tenants are in a data center. Thus such orchestration software can be appropriate for providing the directory function or for supplying the Directory(s) with directory information.

Directory services can be offered in a Push or Pull Mode [RFC7067]. Push Mode, in which a directory server pushes information to TRILL switches indicating interest, is specified in Section 2. Pull Mode, in which a TRILL switch queries a server for the information it wants, is specified in Section 3. More detail on modes of operation, including hybrid Push/Pull, are provided in Section 4.

The mechanism used to initially populate directory data in primary servers is beyond the scope of this document. A primary server can use the Push Directory service to provide directory data to secondary servers as described in Section 2.5.

### 1.1 Uses of Directory Information

A TRILL switch can consult Directory information whenever it wants, by (1) searching through information that has been retained after being pushed to it or pulled by it or (2) by requesting information from a Pull Directory. However, the following are expected to be the most common circumstances leading to directory information use. All of these are cases of ingressing (or originating) a native frame.

1. ARP requests and replies [RFC826] are normally broadcast. But a directory assisted edge TRILL switches could intercept ARP messages and reply if the TRILL switch has the relevant information.
2. IPv6 ND (Neighbor Discovery [RFC4861]) requests and replies are normally multicast. Except in the case of Secure ND [RFC3971] where possession of the right keying material might be required, directory assisted edge TRILL switches could intercept ND messages and reply if the TRILL switch has the relevant information.
3. Unknown destination MAC addresses. An edge TRILL switch ingressing a native frame necessarily has to determine if it knows the egress RBridge from which the destination MAC address of the frame (in the frame's VLAN or FGL) is reachable. It might learn that information from the directory or could query the directory if it does not know. Furthermore, if the edge TRILL switch has complete directory information, it can detect a forged source MAC address in the native frame and discard the frame in that case.
4. RARP [RFC903] is similar to ARP as above.

## 1.2 Terminology

The key words "MUST", "MUST NOT", "REQUIRED", "SHALL", "SHALL NOT", "SHOULD", "SHOULD NOT", "RECOMMENDED", "MAY", and "OPTIONAL" in this document are to be interpreted as described in RFC 2119 [RFC2119].

The terminology and acronyms of [RFC6325] are used herein along with the following:

CSNP Time: Complete Sequence Number PDU Time. See ESDADI [RFC7357] and Section 6.1 below.

Data Label: VLAN or FGL.

FGL: Fine Grained Label [RFC7172].

Host: Application running on a physical server or a virtual machine. A host must have a MAC address and usually has at least one IP address.

IP: Internet Protocol. In this document, IP includes both IPv4 and IPv6.

MacDA: Destination MAC address.

PDSS: Push Directory Server Status. See Sections 2 and 6.1 below.



PUL: Pull Directory flag bit. See Sections 3 and 6.3 below.

primary server: A Directory server that obtains the information it is serving up by a reliable mechanism outside the scope of this document designed to assure the freshness of that information. (See secondary server.)

RBridge: An alternative name for a TRILL switch.

secondary server: A Directory server that obtains the information it is serving up from one or more primary servers.

TRILL: Transparent Interconnection of Lots of Links or Tunneled Routing in the Link Layer.

TRILL switch: A device that implements the TRILL protocol.

## 2. Push Model Directory Assistance Mechanisms

In the Push Model [RFC7067], one or more Push Directory servers reside at TRILL switches and push down the address mapping information for the various addresses associated with end station interfaces and the TRILL switches from which those interfaces are reachable [IA]. This service is scoped by Data Label (VLAN or FGL [RFC7172]). A Push Directory also advertises whether or not it believes it has pushed complete mapping information for a Data Label. It might be pushing only a subset of the mapping and/or reachability information for a Data Label. The Push Model uses the ESADI [RFC7357] protocol as its distribution mechanism.

With the Push Model, if complete address mapping information for a Data Label is being pushed, a TRILL switch (RBridge) which has that complete information and is ingressing a native frame can simply drop the frame if the destination unicast MAC address can't be found in the mapping information available, instead of flooding the frame (ingressing it as an unknown MAC destination TRILL Data frame). But this will result in lost traffic if ingress TRILL switch's directory information is incomplete.

### 2.1 Requesting Push Service

In the Push Model, it is necessary to have a way for a TRILL switch to subscribe to information from the directory server(s). TRILL switches simply use the ESADI [RFC7357] protocol mechanism to announce, in their core IS-IS LSPs, the Data Labels for which they are participating in ESADI by using the Interested VLANs and/or Interested Labels sub-TLVs [RFC7176]. This will cause them to be pushed the Directory information for all such Data Labels that are being served by the one or more Push Directory servers.

### 2.2 Push Directory Servers

Push Directory servers advertise their availability to push the mapping information for a particular Data Label to each other and to ESADI participants for that Data Label through ESADI by setting the PDSS (Push Directory Server Status) in their ESADI Parameter APPsub-TLV for that ESADI instance (see [RFC7357] and Section 6.1) to a non-zero value. Each Push Directory server MUST participate in ESADI for the Data Labels for which it will push mappings and set the PDSS field in its ESADI-Parameters APPsub-TLV for that Data Label.

For robustness, it is useful to have multiple Push Directory Servers for each Data Label. Each Push Directory server is configured with a

number N in the range 1 to 8, which defaults to 2, for each Data Label for which it can push directory information. If the Push Directory servers for a Data Label are configured consistently with the same N and at least N servers are available, then N copies of that directory will be pushed.

Each Push Directory server also has an 8-bit priority to be Active (see Section 6.1 of this document). This priority is treated as an unsigned integer where larger magnitude means higher priority. This priority appears in its ESADI Parameter APPsub-TLV.

For each Data Label it can serve, each Push Directory server checks to see if there are enough higher priority servers to push the desired number of copies. It does this by ordering, by priority, the Push Directory servers that it can see in the ESADI link state database for that Data Label that are data reachable [rfc7180bis] and determines its own position in that order. If a Push Directory server is configured to believe that N copies of the mappings for a Data Label should be pushed and finds that it is number K in the priority ordering (where the first is highest priority and the last is lowest), then if K is less than or equal to N the Push Directory server is Active. If K is greater than N it is Stand-By. Active and Stand-By behavior are specified below.

For a Push Directory to reside on an end station, one or more TRILL switches locally connected to that end station must proxy for the Push Directory server and advertise themselves as Push Directory servers. It appears to the rest of the TRILL campus that these TRILL switches (that are proxying for the end station) are the Push Directory server(s). The protocol between such a Push Directory end station and the one or more proxying TRILL switches acting as Push Directory servers is beyond the scope of this document.

### 2.3 Push Directory Server State Machine

The subsections below describe the states, events, and corresponding actions for Push Directory servers.

The meaning of the value of the PDSS field in a Push Directory's ESADI Parameter APPsub-TLV is summarized in the table below.

PDSS	Meaning
----	-----
0	Not a Push Directory Server
1	Push Directory Server in Stand-By Mode
2	Push Directory Server in Active Mode but not complete
3	Push Directory Server in Active Mode that has pushed complete data

### 2.3.1 Push Directory States

A Push Directory Server is in one of seven states, as listed below, for each Data Label it can serve. The name of each state is followed by a symbol that starts and ends with an angel bracket and represents the state. The value that the Push Directory Server advertises in PDSS is determined by the state. In addition, it has an internal State-Transition-Time variable for each Data Label it serves which is set at each state transition and which enables it to determine how long it has been in its current state for that Data Label.

**Down <S1>:** A completely shut down virtual state defined for convenience in specifying state diagrams. A Push Directory Server in this state does not advertise any Push Directory data. It may be participating in ESDADI [RFC7357] with the PDSS field zero in its ESADI-Parameters or might be not participating in ESADI at all. (All states other than the Down state are considered to be Up states and imply a non-zero PDSS field.)

**Stand-By <S2>:** No Push Directory data is advertised. Any outstanding EASDI-LSP fragments containing directory data are updated to remove that data and if the result is an empty fragment (contains nothing except possibly an Authentication TLV), the fragment is purged. The Push Directory participates in ESDADI [RFC7357] and advertises its ESADI fragment zero that includes an ESADI-Parameters APPsub-TLV with the PDSS field set to 1.

**Active <S3>:** The PDSS field in the ESADI-Parameters is set to 2. If a Push Directory server is Active, it advertises its directory data and any changes through ESADI [RFC7357] in its ESADI-LSPs using the Interface Addresses [IA] APPsub-TLV and updates that information as it changes.

**Active Completing <S4>:** Same behavior as the Active state except that it responds differently to events. The purpose of this state is to be sure there has been enough time for directory information to propagate to subscribing edge TRILL switches before the Directory Server advertises that the information is complete.

**Active Complete <S5>:** The same behavior as Active except that the PDSS field in the ESADI-Parameters APPsub-TLV is set to 3 and the server responds differently to events.

**Going Stand-By <S6>:** The same behavior as Active except that it responds differently to events. The purpose of this state is to be sure that the information, that the directory is no longer complete, has enough time to propagate to edge TRILL switches before the Directory Server stops advertising updates to the information.

Active Uncompleting <S7>: The same behavior as Active except that it responds differently to events. The purpose of this state is to be sure that the information, that the directory is no longer complete, has enough time to propagate to edge TRILL switches before the Directory Server might stop advertising updates to the information. (See note below.)

Note: It might appear that a Push Directory could transition directly from Active Complete to Active, since Active state continues to advertise updates, eliminating the need for the Active Uncompleting transition state. But consider the case of the Push Directory being configured to be incomplete and then the Stand-By Condition (see Section 2.3.2) occurring immediately thereafter. If the first of these two events caused the server to transition directly to the Active state then, when the Stand-By Condition occurred, it would immediately transition to Stand-By and stop advertising updates even though there might not have been enough time for knowledge of its incompleteness to have propagated to all edge TRILL switches.

The following table summarizes PDSS value for each state:

State	PDSS
-----	-----
Down <S1>	0
Stand-By <S2>	1
Active <S3>	2
Active Completing <S4>	2
Active Complete <S5>	3
Going Stand-By <S6>	2
Active Uncompleting <S7>	2

### 2.3.2 Push Directory Events and Conditions

Three auxiliary conditions referenced later in this section are defined as follows for convenience:

The Activate Condition: In order to have the desired number of Push Directory servers pushing data, this Push Directory server should be active. This is determined by the server finding that it is priority K among the data reachable Push Directory servers (where highest priority is 1), it is configured that there should be N copies pushed, and K is less than or equal to N. For example, the Push Directory server is configured that 2 copies should be pushed and finds that it is priority 1 or 2 among the Push Directory servers it can see.

The Stand-By Condition: In order to have the desired number of Push

Directory servers pushing data, this Push Directory server should be stand-by (not active). This is determined by the server finding that it is priority K among the data reachable Push Directory servers (where highest priority is 1), it is configured that there should be N copies pushed, and K is greater than N. For example, the Push Directory server is configured that 2 copies should be pushed and finds that it is priority 3 or lower priority (higher number) among the Push directory servers it can see.

The Time Condition: The Push Directory server has been in its current state for a configurable amount of time that defaults to twice its CSNP time (see Section 6.1).)

The events and conditions listed below cause state transitions in Push Directory servers.

1. Push Directory server was Down but is now Up.
2. The Push Directory server or the TRILL switch on which it resides is being shut down.
3. The Activate Condition is met and the server is not configured to believe it has complete data.
4. The Stand-By Condition is met.
5. The Activate Condition is met and the server is configured to believe it has complete data.
6. The server is configured to believe it does not have complete data.
7. The Time Condition is met.

### 2.3.3 State Transition Diagram and Table

The state transition table is as follows:

State -----+	Down	Stand-By	Active	Active Completing	Active Complete	Going Stand-By	Active Uncompleting
Event	<S1>	<S2>	<S3>	<S4>	<S5>	<S6>	<S7>
1	<S2>	<S2>	<S3>	<S4>	<S5>	<S6>	<S7>
2	<S1>	<S1>	<S2>	<S2>	<S6>	<S6>	<S7>
3	<S1>	<S3>	<S3>	<S3>	<S7>	<S3>	<S7>
4	<S1>	<S2>	<S2>	<S2>	<S6>	<S6>	<S6>
5	<S1>	<S4>	<S4>	<S4>	<S5>	<S5>	<S5>
6	<S1>	<S2>	<S3>	<S3>	<S7>	<S6>	<S7>

7 | <S1> | <S2> | <S3> | <S5> | <S5> | <S2> | <S3>

The above state table is equivalent to the following transition diagram:

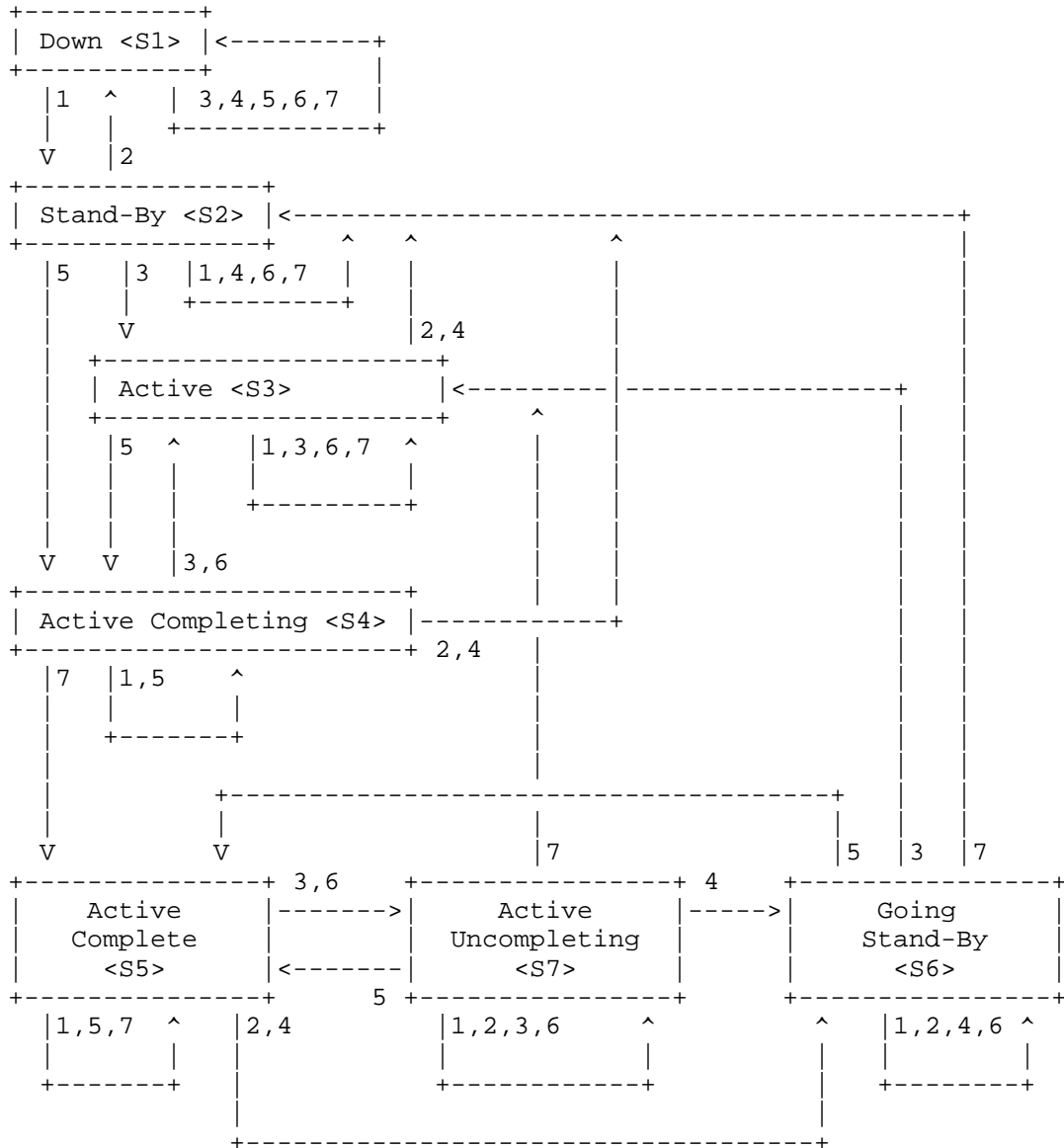


Figure 2. Push Server State Diagram

## 2.4 Additional Push Details

Push Directory mappings can be distinguished from other data distributed through ESADI because mappings are distributed only with the Interface Addresses APPsub-TLV [IA] and are flagged in that APPsub-TLV as being Push Directory data.

TRILL switches, whether or not they are a Push Directory server, MAY continue to advertise any locally learned MAC attachment information in ESADI [RFC7357] using the Reachable MAC Addresses TLV [RFC6165]. However, if a Data Label is being served by complete Push Directory servers, advertising such locally learned MAC attachment generally SHOULD NOT be done as it would not add anything and would just waste bandwidth and ESADI link state space. An exception might be when a TRILL switch learns local MAC connectivity and that information appears to be missing from the directory mapping.

Because a Push Directory server needs to advertise interest in one or more Data Labels even though it might not want to receive multi-destination data in those Data Labels, the No Data (NOD) flag bit is provided as discussed in Section 3.8.

When a Push Directory server is no longer data reachable [rfc7180bis], TRILL switches MUST ignore any Push Directory data from that server because it is no longer being updated and may be stale.

The nature of dynamic distributed asynchronous systems is such that it is impossible for a TRILL switch receiving Push Directory information to be absolutely certain that it has complete information. However, it can obtain a reasonable assurance of complete information by requiring two conditions to be met:

1. The PDSS field is 3 in the ESADI zero fragment from the server for the relevant Data Label.
2. In so far as it can tell, it has had continuous data connectivity to the server for a configurable amount of time that defaults to twice the server's CSNP time.

Condition 2 is necessary because a client TRILL switch might be just coming up and receive an EASDI LSP meeting the requirement in condition 1 above but has not yet received all of the ESADI LSP fragment from the Push Directory server.

There may be conflicts between mapping information from different Push Directory servers or conflicts between locally learned information and information received from a Push Directory server. In case of such conflicts, information with a higher confidence value [RFC6325] is preferred over information with a lower confidence. In case of equal confidence, Push Directory information is preferred to locally learned information and if information from Push Directory servers conflicts, the information from the higher priority Push Directory server is preferred.



## 2.5 Primary to Secondary Server Push Service

A secondary Push or Pull Directory server is one that obtains its data from a primary directory server. Other techniques MAY be used but, by default, this data transfer occurs through the primary server acting as a Push Directory server for the Data Labels involved while the secondary directory server takes the pushed data it receives from the highest priority Push Directory server and re-originates it. Such a secondary server may be a Push Directory server or a Pull Directory server or both for any particular Data Label. Because the data from a secondary server will necessarily be at least a little less fresh than that from a primary server, it is RECOMMENDED that the re-originated secondary server data be given a confidence level of one less than that of the data as received from the primary (or unchanged if it is already of minimum confidence).

### 3. Pull Model Directory Assistance Mechanisms

In the Pull Model [RFC7067], a TRILL switch (RBridge) pulls directory information from an appropriate Directory Server when needed.

Pull Directory servers for a particular Data Label X are found by looking in the core TRILL IS-IS link state database for data reachable [rfc7180bis] TRILL switches that advertise themselves by having the Pull Directory flag (PUL) on in their Interested VLANs or Interested Labels sub-TLV (see Section 6.3)) for that Data Label. If multiple such TRILL switches indicate that they are Pull Directory Servers for a particular Data Label, pull requests can be sent to any one or more of them but it is RECOMMENDED that pull requests be preferentially sent to the server or servers that are lowest cost from the requesting TRILL switch.

Pull Directory requests are sent by enclosing them in an RBridge Channel [RFC7178] message using the Pull Directory channel protocol number (see Section 6.2). Responses are returned in an RBridge Channel message using the same channel protocol number. See Section 3.2 for Query and Response Message formats. For cache consistency or notification purposes, Pull Directory servers, under certain conditions, MUST send unsolicited Update Messages to client TRILL switches they believe may be holding old data and those clients can acknowledge such updates, as described in Section 3.3. All these messages have a common header as described in Section 3.1. Errors can be returned for queries or updates as described in Section 3.6.

The requests to Pull Directory Servers are typically derived from ingressed ARP [RFC826], ND [RFC4861], or RARP [RFC903] messages, or data frames with unknown unicast destination MAC addresses, intercepted by an ingress TRILL switch as described in Section 1.1.

Pull Directory responses include an amount of time for which the response should be considered valid. This includes negative responses that indicate no data is available. It is RECOMMENDED that both positive responses with data and negative responses can be cached and used to locally handle ARP, ND, RARP, unknown destination MAC frames, or the like, until the responses expire. If information previously pulled is about to expire, a TRILL switch MAY try to refresh it by issuing a new pull request but, to avoid unnecessary requests, SHOULD NOT do so if it has not been recently used. The validity timer of cached Pull Directory responses is NOT reset or extended merely because that cache entry is used.

### 3.1 Pull Directory Message Common Format

All Pull Directory messages are transmitted as the payload of RBridge Channel messages [RFC7178]. Pull Directory messages are formatted as described herein starting with the following common 8-byte header:

```

          1 1 1 1 1 1 1 1 1 1 2 2 2 2 2 2 2 2 2 2 3 3
0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1
+-----+-----+-----+-----+-----+-----+-----+-----+
| Ver   | Type  | Flags | Count |           Err           | SubErr |
+-----+-----+-----+-----+-----+-----+-----+-----+
|                                           Sequence Number                                           |
+-----+-----+-----+-----+-----+-----+-----+-----+
| Type Specific Payload - variable length |
+-----+ ...

```

Ver: Version of the Pull Directory protocol as an unsigned integer. Version zero is specified in this document.

Type: The Pull Directory message type as follows:

Type	Section	Name
0	-	Reserved
1	3.2.1	Query
2	3.2.2	Response
3	3.3.1	Update
4	3.3.2	Acknowledge
5-14	-	Unassigned
15	-	Reserved

Flags: Four flag bits whose meaning depends on the Pull Directory message Type. Flags whose meanings are not specified are reserved, MUST be sent as zero, and MUST be ignored on receipt.

Count: Pull Directory message types specified herein have zero or more occurrences of a Record as part of the type specific payload. The Count field is the number of occurrences of that Record as an unsigned integer. For any Pull Directory messages not structured with such occurrences, this field MUST be sent as zero and ignored on receipt.

Err, SubErr: The error and suberror fields are only used in messages that are in the nature of replies. In messages that are requests or updates, these fields MUST be sent as zero and ignored on receipt. An Err field containing the value zero means no error. The meaning of values in the SubErr field depends on the value of the Err field but in all cases, a zero SubErr field is allowed and provides no additional information beyond the value of the Err field.

**Sequence Number:** An identifying 32-bit quantity set by the TRILL switch sending a request or other unsolicited message and returned in every corresponding reply or acknowledgement. It is used to match up responses with the message to which they respond.

**Type Specific Payload:** Format depends on the Pull Directory message Type.

### 3.2 Pull Directory Query and Response Messages

The format of the Pull Directory Query and Response Messages is specified below.

#### 3.2.1 Pull Directory Query Message Format

A Pull Directory Query Message is sent as the Channel Protocol specific content of an RBridge Channel message [RFC7178] TRILL Data packet or as a native RBridge Channel data frame (see Section 3.5). The Data Label of the packet is the Data Label in which the query is being made. The priority of the channel message is a mapping of the priority of the frame being ingressed that caused the query with the default mapping depending, per Data Label, on the strategy (see Section 4) or a configured priority for generated queries. (Generated queries are those not the result of a mapping. For example, a query to refresh a cache entry.) The Channel Protocol specific data is formatted as a header and a sequence of zero or more QUERY Records as follows:

```

                                1 1 1 1 1 1 1 1 1 1 2 2 2 2 2 2 2 2 2 2 3 3
0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1
+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+
| Ver  | Type | Flags | Count |           Err           |      SubErr      |
+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+
|                                           Sequence Number                                           |
+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+
| QUERY 1
+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+
| QUERY 2
+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+
| ...
+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+
| QUERY K
+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+

```

Ver, Sequence Number: See 3.1.

Type: 1 for Query. Queries received by an TRILL switch that is not a Pull Directory for the relevant Data Label result in an error response (see Section 3.6) unless inhibited by rate limiting. (See [RFC7178] for response if the Pull Directory RBridge Channel protocol is not enabled.)

Flags, Err, and SubErr: MUST be sent as zero and ignored on receipt.

Count: Number of QUERY Records present. A Query Message Count of zero is explicitly allowed, for the purpose of pinging a Pull Directory server to see if it is responding. On receipt of such an empty Query Message, a Response Message that also has a Count of zero is sent unless inhibited by rate limiting.

QUERY: Each QUERY Record within a Pull Directory Query Message is formatted as follows:

```

      0  1  2  3  4  5  6  7  8  9 10 11 12 13 14 15
+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+
|           SIZE           |FL|  RESV  |   QTYPE   |
+---+---+---+---+---+---+---+---+---+---+---+---+---+---+
If QTYPE = 1
+---+---+---+---+---+---+---+---+---+---+---+---+---+---+
|                               AFN                               |
+---+---+---+---+---+---+---+---+---+---+---+---+---+---+
|  Query address ...
+---+---+---+---+---+---+---+---+---+---+---+---+---+---+
If QTYPE = 2, 3, 4, or 5
+---+---+---+---+---+---+---+---+---+---+---+---+---+---+
|  Query frame ...
+---+---+---+---+---+---+---+---+---+---+---+---+---+---+

```

SIZE: Size of the QUERY Record in bytes as an unsigned integer not including the SIZE field and following byte. A value of SIZE so large that the material doesn't fit in the Query Message indicates a malformed QUERY Record. The QUERY Record with the illegal SIZE value and any subsequent QUERY Records MUST be ignored and the entire Query Message MAY be ignored.

FL: The FLooded flag that is ignored if QTYPE is zero. If QTYPE is 2 through 5 and the directory information sought is not found, the frame provided is flooded, otherwise it is not forwarded. See Section 3.2.2.2.

RESV: A block of three reserved bits. MUST be sent as zero and ignored on receipt.

QTYPE: There are several types of QUERY Records currently defined in two classes as follows: (1) a QUERY Record that

provides an explicit address and asks for all addresses for the interface specified by the query address and (2) a QUERY Record that includes a frame. The fields of each are specified below. Values of QTYPE are as follows:

QTYPE	Description
-----	-----
0	Reserved
1	Address query
2	ARP query frame
3	ND query frame
4	RARP query frame
5	Unknown unicast MAC query frame
6-14	Unassigned
15	Reserved

AFN: Address Family Number of the query address.

Query Address: The query is asking for any other addresses, and the nickname of the TRILL switch from which they are reachable, that correspond to the same interface, within the data label of the query. Typically that would be either (1) a MAC address with the querying TRILL switch primarily interested in the TRILL switch by which that MAC address is reachable, or (2) an IP address with the querying TRILL switch interested in the corresponding MAC address and the TRILL switch by which that MAC address is reachable. But it could be some other address type.

Query Frame: Where a QUERY Record is the result of an ARP, ND, RARP, or unknown unicast MAC destination address, the ingress TRILL switch MAY send the frame to a Pull Directory Server if the frame is small enough that the resulting Query Message fits into a TRILL Data packet within the campus MTU.

If no response is received to a Pull Directory Query Message within a timeout configurable in milliseconds that defaults to 100, the Query Message should be re-transmitted with the same Sequence Number up to a configurable number of times that defaults to three. If there are multiple QUERY Records in a Query Message, responses can be received to various subsets of these QUERY Records before the timeout. In that case, the remaining unanswered QUERY Records should be re-sent in a new Query Message with a new sequence number. If a TRILL switch is not capable of handling partial responses to queries with multiple QUERY Records, it MUST NOT send a Request Message with more than one QUERY Record in it.

See Section 3.6 for a discussion of how Query Message errors are handled.

### 3.2.2 Pull Directory Responses

A Pull Directory Query Message results in a Pull Directory Response Message as described in Section 3.2.2.1.

In addition, if the QUERY Record QTYPE was 2, 3, 4, or 5, the frame included in the Query may be modified and forwarded by the Pull Directory server as described in Section 3.2.2.2.

#### 3.2.2.1 Pull Directory Response Message Format

Pull Directory Response Messages are sent as the Channel Protocol specific content of an RBridge Channel message [RFC7178] TRILL Data packet or as a native RBridge Channel data frame (see Section 3.5). Responses are sent with the same Data Label and priority as the Query Message to which they correspond except that the Response Message priority is limited to be not more than a configured value. This priority limit is configurable per TRILL switch and defaults to priority 6. Pull Directory Response Messages SHOULD NOT be sent with priority 7 as that priority SHOULD be reserved for messages critical to network connectivity.

The RBridge Channel protocol specific data format is as follows:

```

                                1 1 1 1 1 1 1 1 1 1 2 2 2 2 2 2 2 2 2 2 3 3
0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1
+-----+-----+-----+-----+-----+-----+-----+-----+
| Ver   | Type   | Flags  | Count  |           Err           | SubErr   |
+-----+-----+-----+-----+-----+-----+-----+
|                               Sequence Number                               |
+-----+-----+-----+-----+-----+-----+
| RESPONSE 1
+-----+-----+-----+-----+-----+-----+...
| RESPONSE 2
+-----+-----+-----+-----+-----+-----+...
| ...
+-----+-----+-----+-----+-----+-----+...
| RESPONSE K
+-----+-----+-----+-----+-----+-----+...

```

Ver, Sequence Number: As specified in Section 3.1.

Type: 2 = Response.

Flags: MUST be sent as zero and ignored on receipt.

Count: Count is the number of RESPONSE Records present in the Response Message.

Err, SubErr: A two-part error code. Zero unless there was an error in the Query Message, for which case see Section 3.6.

RESPONSE: Each RESPONSE Record within a Pull Directory Response Message is formatted as follows:

```

    0  1  2  3  4  5  6  7  8  9 10 11 12 13 14 15
+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+
|          SIZE          |OV|  RESV  |   Index   |
+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+
|                                Lifetime                                |
+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+
|                                Response Data ...                                |
+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+

```

SIZE: The size of the RESPONSE Record is an unsigned integer number of bytes not including the SIZE field and following byte. A value of SIZE so large that the material doesn't fit in the Query Message indicates a malformed QUERY Record. The QUERY Record with such an excessive SIZE value and any subsequent QUERY Records MUST be ignored and the entire Query Message MAY be ignored.

OV: The overflow flag. Indicates, as described below, that there was too much Response Data to include in one Response Message.

RESV: Three reserved bits that MUST be sent as zero and ignored on receipt.

Index: The relative index of the QUERY Record in the Query Message to which this RESPONSE Record corresponds. The index will always be one for Query Messages containing a single QUERY Record. If the Index is larger than the Count was in the corresponding Query, that RESPONSE Record MUST be ignored and subsequent RESPONSE Records or the entire Response Message MAY be ignored.

Lifetime: The length of time for which the response should be considered valid in units of 100 milliseconds except that the values zero and  $2^{16}-1$  are special. If zero, the response can only be used for the particular query from which it resulted and MUST NOT be cached. If  $2^{16}-1$ , the response MAY be kept indefinitely but not after the Pull Directory server goes down or becomes unreachable. (The maximum definite time that can be expressed is a little over 1.8 hours.)

Response Data: There are three types of RESPONSE Records.

- If the Err field of the enclosing Respose Message has a



- message level error code in it, then the the REPONSE Records are omitted and Count will be zero. See Section 3.6 for additional information on errors.
- If the Err field of the enclosing Response Message has a record level error code in it, then the RESPONSE Records are those in error as further described in Section 3.6.
  - If the Err field of the enclosing Repose Message is zero, then the Response Data in each RESPONSE Record is formatted as the value of an Interface Addresses APPsub-TLV [IA]. The maximum size of such contents is 255 bytes in the case when the RESPONSE Record SIZE field is 255.

Multiple RESPONSE Records can appear in a Response Message with the same Index if the answer to a QUERY Record consists of multiple Interface Address APPsub-TLV values. This would be necessary if, for example, a MAC address within a Data Label appears to be reachable by multiple TRILL switches. However, all RESPONSE Records to any particular QUERY Record MUST occur in the same Response Message. If a Pull Directory holds more mappings for a queried address than will fit into one Response Message, it selects which to include by some method outside the scope of this document and sets the overflow flag (OV) in all of the RESPONSE Records responding to that query address.

See Section 3.6 for a discussion of how errors are handled.

#### 3.2.2.2 Pull Directory Forwarding

Query Messages with QTYPEs 2, 3, 4, and 5 are interpreted and handled as described below. In these cases, if the information sought is not in the directory, the provided frame is forwarded by the Pull Directory server as a multi-destination TRILL Data packet if the FL flag in the Query Message was one, otherwise the frame is not forwarded. If there was no error in the handling of the enclosing Query Message then the Pull Directory server forwards the frame inside that QUERY Record, after modifying it in some cases, as described below.

ARP: When QTYPE is 2, an ARP [RFC826] frame is included in the QUERY Record. The ar\$op field MUST be ares\_op\$REQUEST and for the response described in 3.2.2.1, this is treated as a query for the target protocol address where the AFN of that address is given by ar\$pro. (ARP field and value names with embedded dollar signs are specified in [RFC826].) If ar\$op is not ares\_op\$REQUEST or the ARP is malformed or the query fails, an error is returned. Otherwise the ARP is modified into the appropriate ARP response that is then sent by the Pull Directory server as a TRILL Data packet.

ND: When QTYPE is 3, an IPv6 Neighbor Discover (ND [RFC4861]) frame

is included in the QUERY Record. Only Neighbor Solicitation ND frames (corresponding to an ARP query) are allowed. An error is returned for other ND frames or if the target address is not found. Otherwise an ND Neighbor Advertisement response is returned by the Pull Directory server as a TRILL Data packet.

RARP: When QTYPE is 4, a RARP [RFC903] frame is included in the QUERY Record. If the ar\$op field is ares\_op\$REQUEST, the frame is handled as an ARP as described above. Otherwise the ar\$op field MUST be 'reverse request' and for the response described in 3.2.2.1, this is treated as a query for the target hardware address where the AFN of that address is given by ar\$hrd. (See [RFC826] for RARP fields.) If ar\$op is not one of these values or the RARP is malformed or the query fails, an error is returned. Otherwise the RARP is modified into the appropriate RARP response that is then unicast by the Pull Directory server as a TRILL Data packet to the source hardware MAC address.

MacDA: When QTYPE is 5, indicating a fame is provided in the QUERY Record whose destination MAC address TRILL switch attachment is unknown, the only requirement is that this MAC address must be unicast. If it is group addressed an error is returned. For the response described in 3.2.2.1, it is treated as a query for the MacDA. If the Pull Directory contains TRILL switch attachment information for the MAC address in the Data Label of the Query Message, it forwards the frame to that switch in a unicast TRILL Data packet.

### 3.3 Cache Consistency

Unless it sends all responses with a Lifetime of zero, a Pull Directory MUST take action, by sending Update Messages, to minimize the amount of time that a TRILL switch will continue to use stale information from that Pull Directory. The format of Update Messages and the Acknowledge Messages used to respond to Update Messages are given in Sections 3.3.1 and 3.3.2.

A Pull Directory server MUST maintain one of the following three sets of records, in order of increasing specificity. Retaining more specific records, such as that given in method 3 below, minimizes spontaneous Update Messages sent to update pull client TRILL switch caches but increases the record keeping burden on the Pull Directory server. Retaining less specific records, such as that given in method 1, will generally increase the volume and overhead due to spontaneous Update Messages and due to unnecessarily invalidating cached information, but will still maintain consistency and will reduce the record keeping burden on the Pull Directory server. In all cases, there may still be brief periods of time when directory information

has changed, but information a pull client has cached has not yet been updated or expunged.

1. An overall record per Data Label of when the last positive response data sent will expire at some requester and when the last negative response will expire at some requester, assuming those requesters cached the response.
2. For each unit of data (IA APPsub-TLV Address Set [IA]) held by the server and each address about which a negative response was sent, when the last response sent with that positive response data and when the last negative response will expire at a requester, assuming the requester cached the response.
3. For each unit of data held by the server (IA APPsub-TLV Address Set [IA]) and each address about which a negative response was sent, a list of TRILL switches that were sent that data as a positive response or sent a negative response for the address, and the expected time to expiration for that data or address at each such TRILL switch, assuming the requester cached the response.

RESPONSE Records sent with a zero lifetime are considered to have already expired and so do not need to be tracked.

A Pull Directory server may have a limit as to how many TRILL switches for which it can maintain expiry information by method 3 above or how many data units or addresses it can maintain expiry information for by method 2 or the like. If such limits are exceeded, it MUST transition to a lower numbered method but, in all cases, MUST support, at a minimum, method 1.

When data at a Pull Directory is changed, deleted, or added and there may be unexpired stale information at a requesting TRILL switch, the Pull Directory MUST send an Update Message as discussed below. The sending of such an Update Message MAY be delayed by a configurable number of milliseconds that default to 50 milliseconds to await other possible changes that could be included in the same Update.

1. If method 1, the crudest method, is being followed, then when any Pull Directory information in a Data Label is changed or deleted and there are outstanding cached positive data response(s), an all-addresses flush positive data Update Message is flooded within that Data Label as an RBridge Channel Message with an Inner.MacDA of All-Egress-RBridges. Similarly if data is added and there are outstanding cached negative responses, an all-addresses flush negative message is similarly flooded. The Count field being zero in an Update Message indicates "all-addresses". On receiving an all-addresses flooded flush positive Update from a Pull Directory server it has used, indicated by

the F and P bits being one and the Count being zero, a TRILL switch discards the cached data responses it has for that Data Label. Similarly, on receiving an all addresses flush negative Update, indicated by the F and N bits being one and the Count being zero, it discards all cached negative replies for that Data Label. A combined flush positive and negative can be flooded by having all of the F, P, and N bits set to one resulting in the discard of all positive and negative cached information for the Data Label.

2. If method 2 is being followed, then a TRILL switch floods address specific positive Update Messages when data that might be cached by a querying TRILL switch is changed or deleted and floods address specific negative Update Messages when such information is added to. Such messages are somewhat similar to the method 1 flooded flush Update Messages and are also sent as RBridge Channel messages with an Inner.MacDA of All-Egress-RBridges. However the Count field will be non-zero and either the P or N bit, but not both, will be one. There are actually four possible message types that can be flooded:
  - 2.a If data still being cached is updated, then an Update Message is sent with the P flag set and the Err field zero. The addresses in the RESPONSE Records in the unsolicited response are compared to the addresses about which the receiving TRILL switch is holding cached positive information from that server and, if they match, the cached information is updated.
  - 2.b If data still being cached is deleted, then an Update Message is sent with the P flag set and the Err field non-zero giving the error that would now be encountered in attempting to pull information for the relevant address from the Pull Directory server. In this non-zero Err field case, the RESPONSE Record(s) differ from non-zero Err Reply Message RESPONSE Records in that they include an interface address set. Any cached positive information for the address is deleted and the negative response cached as per the lifetime given.
  - 2.c If data for an address about which a negative response was sent is added so that negative response is now incorrect, an Update Message is sent with the N flag set to one and the Err field zero. The addresses in the RESPONSE Records in the unsolicited response are compared to the addresses about which the receiving TRILL switch is holding cached negative information from that server and, if they match, the cached negative information is deleted and the positive information provided is cached as per the lifetime given.

- 2.d In the rare case where it is desired to change the lifetime or error associated with cached negative information, it is possible to send an Update Message with the N flag set to one and the Err field non-zero. As in case 2.b above, the RESPONSE Record(s) give the relevant addresses. Any cached negative information for the address is updated.
3. If method 3 is being followed, the same sort of unsolicited Update Messages are sent as with method 2 above except they are not normally flooded but unicast only to the specific TRILL switches the directory server believes may be holding the cached positive or negative information that needs updating. However, a Pull Directory server MAY flood unsolicited updates under method 3, for example if it determines that a sufficiently large fraction of the TRILL switches in some Data Label are requesters that need to be updated.

A Pull Directory server tracking cached information with method 3 MUST NOT clear the indication that it needs to update cached information at a querying TRILL switch until it has sent an Update Message and received a corresponding Acknowledge Message or it has sent a configurable number of updates at a configurable interval which default to 3 updates 100 milliseconds apart.

A Pull Directory server tracking cached information with methods 2 or 1 SHOULD NOT clear the indication that it needs to update cached information until it has sent an Update Message and received a corresponding Acknowledge Message from all of its ESADI neighbors or it has sent a configurable number of updates at a configurable interval that defaults to 3 updates 100 milliseconds apart.

### 3.3.1 Update Message Format

An Update Message is formatted as a Response Message with the differences described in Section 3.3 above and the following:

- o The Type field in the message header is set to 3.
- o The Err field in the message header MUST be sent as zero and ignored on receipt.
- o The Index field in the RESPONSE Record(s) is set to zero (but the Count field in the Update Message header MUST still correctly indicate the number of RESPONSE Records present).

Update Messages are initiated by a Pull Directory server. The Sequence number space used is controlled by the originating Pull Directory server and different from Sequence number space used in a Query and the corresponding Response that are controlled by the querying TRILL switch.

The 4-bit Flags field of the message header for an Update Message is as follows:

```

+---+---+---+---+
| F | P | N | R |
+---+---+---+---+

```

F: The Flood bit. If zero, the Update Message is unicast. If F=1, it is multicast to All-Egress-RBridges.

P, N: Flags used to indicate positive or negative Update Messages. P=1 indicates positive. N=1 indicates negative. Both may be 1 for a flooded all addresses Update.

R: Reserved. MUST be sent as zero and ignored on receipt

For tracking methods 2 and 3 in Section 3.3.1, a particular Update Message must have either the P flag or the N flag set but not both.

### 3.3.2 Acknowledge Message Format

An Acknowledge Message is sent in response to an Update Message to confirm receipt or indicate an error, unless response is inhibited by rate limiting. It is also formatted as a Response Message but the Type is set to 4.

If there are no errors in the processing of an Update Message or if there is a message level overall or header error in an Update Message, the message is essentially echoed back with the Err and SubErr fields set appropriately, the Type changed to Acknowledge, and a null records section with the Count field set to zero.

If there is a record level error in an Update Message, one or more Acknowledge Messages may be returned with the erroneous record(s) indicated in Section 3.5.

### 3.4 Summary of Records Formats in Messages

As specified in Section 3.2 and 3.3, the Query, Response, Update, and Acknowledge Messages can have zero or more repeating Record structures under different circumstances, as summarized below. The "Err" column abbreviations in this table have the meanings listed below. "IA APPsubTLV value" means the value part of the IA APPsub-TLV specified in [IA].

MBZ = MUST be zero  
 Z = zero  
 NZ = non-zero  
 NZM = non-zero message level error  
 NZR = non-zero record level error

Message	Err	Section	Record Structure	Response Data
Query	MBZ	3.2.1	QUERY Record	-
Response	Z	3.2.2.1	RESPONSE Record	IA APPsubTLV value
Response	NZM	3.2.2.1	null	-
Response	NZR	3.2.2.1	RESPONSE Record	Records with error
Update	MBZ	3.3.1	RESPONSE Record	IA APPsubTLV value
Acknowledge	Z	3.3.2	null	-
Acknowledge	NZM	3.3.2	null	-
Acknowledge	NZR	3.3.2	RESPONSE Record	Records with error

See Section 3.6 for further details on errors.

### 3.5 Pull Directory Hosted on an End Station

Optionally, a Pull Directory actually hosted on an end station MAY be supported. In that case, one or more TRILL switches must proxy for the end station and advertise themselves as Pull Directory servers. Such proxies must have a direct connection to the end station, that is a connection not involving any intermediate TRILL switches.

When the proxy Pull Directory server TRILL switch receives a Query Message, it modifies the inter-RBridge Channel message received into a native RBridge Channel message and forwards it to the end station Pull Directory server. Later, when it receives one or more responses from that end station by native RBridge Channel messages, it modifies them into inter-RBridge Channel messages and forwards them to the source TRILL switch of the original Query Message. Similarly, an Update from the end station is forwarded to client TRILL switches and acknowledgements from those TRILL switches are returned to the end station by the proxy. Because native RBridge Channel messages have no TRILL Header and are addressed by MAC address, as opposed to inter-RBridge Channel messages that are TRILL Data packets and are addressed by nickname, nickname information must be added to the native RBridge Channel version of Pull Directory messages.

The native Pull Directory RBridge Channel messages use the same Channel protocol number as do the inter-RBridge Pull Directory RBridge Channel messages. The native messages SHOULD be sent with an Outer.VLAN tag that gives the priority of each message which is the priority of the original inter-RBridge request packet. The Outer.VLAN ID used is the Designated VLAN on the link to the end station

[RFC6325]. Since there is no TRILL Header or inner Data Label for native RBridge Channel messages, that information is added to the header.

The native RBridge Channel message Pull Directory message protocol dependent data part is the same as for inter-RBridge Channel messages except that the 8-byte header described in Section 3.1 is expanded to 14 or 18 bytes as follows:

```

          1 1 1 1 1 1 1 1 1 1 2 2 2 2 2 2 2 2 2 2 3 3
0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1
+-----+-----+-----+-----+-----+-----+-----+-----+
| Ver   | Type   | Flags  | Count  |           Err           | SubErr |
+-----+-----+-----+-----+-----+-----+-----+-----+
|                               Sequence Number                               |
+-----+-----+-----+-----+-----+-----+-----+-----+
| Nickname (2 bytes) |
+-----+-----+-----+-----+-----+-----+-----+-----+
| Data Label ... (4 or 8 bytes) |
+-----+-----+-----+-----+-----+-----+-----+-----+
| Type Specific Payload - variable length |
+-----+-----+-----+-----+-----+-----+-----+-----+

```

Fields not described below are as in Section 3.1.

**Nickname:** The nickname of the original TRILL switch that is communicating with the end station Pull Directory. Usually this is a remote TRILL switch but it could be the TRILL switch to which the end station is attached. The proxy copies this from the ingress nickname when mapping a Query or Acknowledge Message to native form. It also takes this from a native Response or Update Message to be used as the egress of the inter-RBridge form on the message unless it is a flooded Update in which case a distribution tree is used.

**Data Label:** The Data Label that normally appears right after the Inner.MacSA of the an RBridge Channel Pull Directory message appears here in the native RBridge Channel message version. This might appear in a native Query Message, to be reflected in a Response Message, or it might appear in a native Update to be reflected in an Acknowledge Message.

### 3.6 Pull Directory Message Errors

A non-zero Err field in the Pull Directory Reponse or Acknowledge Message header indicates an error message.

If there is an error that applies to an entire Query or Update



Message or its header, as indicated by the range of the value of the Err field, then the QUERY Records probably were not even looked at by the Pull Directory server and would provide no information in the Response or Acknowledge Message so they are omitted and the Count field is set to zero in the Response or Acknowledgement Message.

If errors occur at the QUERY Record level for a Query Message, they MUST be reported in a Response Message separate from the results of any successful non-erroneous QUERY Records. If multiple QUERY Records in a Query Message have different errors, they MUST be reported in separate Response Messages. If multiple QUERY Records in a Query Message have the same error, this error response MAY be reported in one or multiple Response Messages. In an error Response Message, the QUERY Record or Records being responded to appear, expanded by the Lifetime for which the server thinks the error might persist and with their Index inserted, as the RESPONSE Record or Records.

If errors occur at the RESPONSE Record level for an Update Message, they MUST be reported in a Acknowledge Message separate from the acknowledgement of any non-erroneous RESPONSE Records. If multiple RESPONSE Records in an Update have different errors, they MUST be reported in separate Acknowledge Messages. If multiple RESPONSE Records in an Update Message have the same error, this error response MAY be reported in one or multiple Acknowledge Messages. In an error Acknowledge Message, the RESPONSE Record or Records being responded to appear, expanded by the time for which the server thinks the error might persist and with their Index inserted, as a RESPONSE Record or Records.

ERR values 1 through 126 are available for encoding Request or Update Message level errors. ERR values 128 through 254 are available for encoding QUERY or RESPONSE Record level errors. The SubErr field is available for providing more detail on errors. The meaning of a SubErr field value depends on the value of the Err field.

### 3.6.1 Error Codes

Err	Level	Meaning
-----	-----	-----
0	-	(no error)
1	Message	Unknown or reserved Query Message field value
2	Message	Request Message/data too short
3	Message	Unknown or reserved Update Message field value
4	Message	Update Message/data too short
5-126	Message	(Available for allocation by IETF Review)
127	-	Reserved
128	Record	Unknown or reserved QUERY Record field value
129	Record	QUERY Record truncated
130	Record	Address not found
131	Record	Unknown or reserved RESPONSE Record field value
132	Record	RESPONSE Record truncated
133-254	Record	(Available for allocation by IETF Review)
255	-	Reserved

Note that some error codes are for overall message level errors while some are for errors in the repeating records that occur in messages.

### 3.6.2 Sub-Errors Under Error Codes 1 and 3

The following sub-errors are specified under error code 1 and 3:

SubErr	Field with Error
-----	-----
0	Unspecified
1	Unknown Ver field value
2	Unknown Type field value
3	Specified Data Label not being served
4-254	(Available for allocation by Expert Review)
255	Reserved

### 3.6.3 Sub-Errors Under Error Codes 128 and 131

The following sub-errors are specified under error code 128 and 131:

SubErr	Field with Error
-----	-----
0	Unspecified
1	Unknown AFN field value
2	Unknown or Reserved QTYPE field value
3	Invalid or inconsistent SIZE field value
4	Invalid frame for QTYPE 2, 3, 4, or 5
5-254	(Available for allocation by Expert Review)
255	Reserved

### 3.7 Additional Pull Details

If a TRILL switch notices that a Pull Directory server is no longer data reachable [rfc7180bis], it MUST promptly discard all pull responses it is retaining from that server as it can no longer receive cache consistency Update Messages from the server.

A secondary Pull Directory server is one that obtains its data from a primary directory server. See discussion of primary to secondary directory information transfer in Section 2.5.

### 3.8 The No Data Flag

In the TRILL base protocol [RFC6325] as extended for FGL [RFC7172], the mere presence of an Interested VLANs or Interested Labels sub-TLVs in the LSP of a TRILL switch indicates connection to end stations in the VLAN(s) or FGL(s) listed and thus a desire to receive multi-destination traffic in those Data Labels. But, with Push and Pull Directories, advertising that you are a directory server requires using these sub-TLVs to indicate the Data Label(s) you are serving. If such a directory server does not wish to received multi-destination TRILL Data packets for the Data Labels it lists in one of these sub-TLVs, it sets the "No Data" (NOD) bit to one. This means that data on a distribution tree may be pruned so as not to reach the "No Data" TRILL switch as long as there are no TRILL switches interested in the Data that are beyond the "No Data" TRILL switch on the distribution tree. The NOD bit is backwards compatible as TRILL switches ignorant of it will simply not prune when they could, which is safe although it may cause increased link utilization.

Example of a TRILL switch serving as a directory that might not want multi-destination traffic in some Data Labels would be a TRILL switch that does not offer end station service for any of the Data Labels for which it is serving as a directory and is either

- a Pull Directory and/or
- a Push Directory for which all of the ESADI traffic will be

handled by unicast ESADI [RFC7357].

A Push Directory MUST NOT set the NOD bit for a Data Label if it needs to communicate via multi-destination ESADI PDUs in that data label since such PDUs look like TRILL Data packets to transit TRILL switches and are likely to be incorrectly pruned if NOD was set.

#### 4. Directory Use Strategies and Push-Pull Hybrids

For some edge nodes that have a great number of Data Labels enabled, managing the MAC and Data Label <-> Edge RBridge mapping for hosts under all those Data Labels can be a challenge. This is especially true for Data Center gateway nodes, which need to communicate with many, if not all, Data Labels.

For those edge TRILL switch nodes, a hybrid model should be considered. That is, the Push Model is used for some Data Labels or addresses within a Data Label while the Pull Model is used for other Data Labels or addresses within a Data Label. It is the network operator's decision by configuration as to which Data Labels' mapping entries are pushed down from directories and which Data Labels' mapping entries are pulled.

For example, assume a data center where hosts in specific Data Labels, say VLANs 1 through 100, communicate regularly with external peers. Probably, the mapping entries for those 100 VLANs should be pushed down to the data center gateway routers. For hosts in other Data Labels that only communicate with external peers occasionally for management interfacing, the mapping entries for those VLANs should be pulled down from directory when the need comes up.

Similarly, it could be that within a Data Label that some addresses, such as the addresses of gateways, file, DNS, or database server hosts are commonly referenced by most other hosts but those other hosts, perhaps compute engines, are typically only referenced by a few hosts in that Data Label. In that case, the address information for the commonly referenced hosts could be pushed as an incomplete directory while the addresses of the others are pulled when needed.

The mechanisms described above for Push and Pull Directory services make it easy to use Push for some Data Labels or addresses and Pull for others. In fact, different TRILL switches can even be configured so that some use Push Directory services and some use Pull Directory services for the same Data Label if both Push and Pull Directory services are available for that Data Label. And there can be Data Labels for which directory services are not used at all.

There are a wide variety of strategies that a TRILL switch can adopt for making use of directory assistance. A few suggestions are given below.

- Even if a TRILL switch will normally be operating with information from a complete Push Directory server, there will be a period of time when it first comes up before the information it holds is complete. Or, it could be that the only Push Directories that can push information to it are incomplete or that they are just starting and may not yet have pushed the entire directory.

Thus, it is RECOMMENDED that all TRILL switches have a strategy for dealing with the situation where they do not have complete directory information. Examples are to send a Pull Directory query or to revert to [RFC6325] behavior.

- If a TRILL switch receives a native frame X resulting in seeking directory information, a choice needs to be made as to what to do if it does not already have the directory information it needs. In particular, it could (1) immediately flood the TRILL Data packet resulting from ingressing X in parallel with seeking the directory information, (2) flood that TRILL Data packet delayed, if it fails to obtain the directory information, or (3) discard X if it fails to obtain the information. The choice might depend on the priority of frame X since the higher that priority, the more urgent the frame is and the greater the probability of harm in delaying it. If a Pull Directory request is sent, it is RECOMMENDED that its priority be derived from the priority of the frame X with the derived priority configurable and having the following defaults:

Ingressed Priority	If Flooded Immediately	If Flooded After Delay
-----	-----	-----
7	5	6
6	5	6
5	4	5
4	3	4
3	2	3
2	0	2
0	1	0
1	1	1

Priority 7 is normally only used for urgent messages critical to adjacency and so SHOULD NOT be the default for directory traffic. Unsolicited updates are sent with a priority that is configured per Data Label that defaults to priority 5.

## 5. Security Considerations

Incorrect directory information can result in a variety of security threats including the following:

Incorrect directory mappings can result in data being delivered to the wrong end stations, or set of end stations in the case of multi-destination packets, violating security policy.

Missing or incorrect directory data can result in denial of service due to sending data packets to black holes or discarding data on ingress due to incorrect information that their destinations are not reachable.

Push Directory data is distributed through ESADI-LSPs [RFC7357] that can be authenticated with the same mechanisms as IS-IS LSPs. See [RFC5304] [RFC5310] and the Security Considerations section of [RFC7357].

Pull Directory queries and responses are transmitted as RBridge-to-RBridge or native RBridge Channel messages [RFC7178]. Such messages can be secured as specified in [ChannelTunnel].

For general TRILL security considerations, see [RFC6325].

## 6. IANA Considerations

This section gives IANA assignment and registry considerations.

### 6.1 ESADI-Parameter Data Extensions

Action 1: IANA will assign a two bit field [bits 1-2 suggested] within the ESADI-Parameter TRILL APPsub-TLV flags for "Push Directory Server Status" (PDSS) and will create a sub-registry in the TRILL Parameters Registry as follows:

Sub-Registry: ESADI-Parameter APPsub-TLV Flag Bits

Registration Procedures: Standards Action

References: [RFC7357] [This document]

Bit	Mnemonic	Description	Reference
---	-----	-----	-----
0	UN	Supports Unicast ESADI	ESDADI [RFC7357]
1-2	PDSS	Push Directory Server Status	[this document]
3-7	-	Available for assignment	

Action 2: In addition, the ESADI-Parameter APPsub-TLV is optionally extended, as provided in its original specification in ESADI [RFC7357], by one byte as show below. Therefore [this document] should be added as a second reference to the ESDAI-Parameter APPsub-TLV in the "TRILL APPsub-TLV Types under IS-IS TLV 251 Application Identifier 1" Registry.

```

+-----+
| Type                | (1 byte)
+-----+
| Length              | (1 byte)
+-----+
|R| Priority          | (1 byte)
+-----+
| CSNP Time           | (1 byte)
+-----+
| Flags               | (1 byte)
+-----+
|PushDirPriority| (optional, 1 byte)
+-----+
| Reserved for expansion | (variable)
+-----+
+-----+

```

The meanings of all the fields are as specified in ESDADI [RFC7357] except that the added PushDirPriority is the priority of the



advertising ESADI instance to be a Push Directory as described in Section 2.3. If the PushDirPriority field is not present (Length = 3) it is treated as if it were 0x40. 0x40 is also the value used and placed here by an TRILL switch whose priority to be a Push Directory has not been configured.

## 6.2 RBridge Channel Protocol Number

Action 3: IANA will allocate a new RBridge Channel protocol number for "Pull Directory Services" from the range allocable by Standards Action and update the subregistry of such protocol number in the TRILL Parameters Registry referencing this document.

## 6.3 The Pull Directory (PUL) and No Data (NOD) Bits

Action 4: IANA is requested to assign a currently reserved bit in the Interested VLANs field of the Interested VLANs sub-TLV [suggested bit 18] and the Interested Labels field of the Interested Labels sub-TLV [suggested bits 6] [RFC7176] to indicate Pull Directory server (PUL). This bit is to be added, with this document as reference, to the "Interested VLANs Flag Bits" and "Interested Labels Flag Bits" subregistries created by [RFC7357].

Action 5: IANA is requested to assign a currently reserved bit in the Interested VLANs field of the Interested VLANs sub-TLV [suggested bits 19] and the Interested Labels field of the Interested Labels sub-TLV [suggested bits 7] [RFC7176] to indicate No Data (NOD, see Section 3.8). This bit is to be added, with this document as reference, to the "Interested VLANs Flag Bits" and "Interested Labels Flag Bits" subregistries created by [RFC7357].

## 6.4 TRILL Pull Directory QTYPES

Action 6: IANA is requested to create a new Registry on the "Transparent Interconnection of Lots of Links (TRILL) Parameters" web page as follows:

Name: TRILL Pull Directory QTYPES"  
Registration Procedure: IETF Review  
Reference: [this document]  
Initial contents as in Section 3.2.1.

## 6.5 Pull Directory Error Code Registries

Actions 7, 8, and 9: IANA is requested to create a new Registry and two new SubRegistries on the "Transparent Interconnection of Lots of Links (TRILL) Parameters" web page as follows:

### Registry

Name: TRILL Pull Directory Errors  
Registration Procedure: IETF Review  
Reference: [this document]

Initial contents as in Section 3.6.1.

### Sub-Registry

Name: Sub-codes for TRILL Pull Directory Errors 1 and 3  
Registration Procedure: Expert Review  
Reference: [this document]

Initial contents as in Section 3.6.2.

### Sub-Registry

Name: Sub-codes for TRILL Pull Directory Errors 128 and 131  
Registration Procedure: Expert Review  
Reference: [this document]

Initial contents as in Section 3.6.3.

## Normative References

- [RFC826] - Plummer, D., "An Ethernet Address Resolution Protocol", RFC 826, November 1982.
- [RFC903] - Finlayson, R., Mann, T., Mogul, J., and M. Theimer, "A Reverse Address Resolution Protocol", STD 38, RFC 903, June 1984
- [RFC2119] - Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, March 1997
- [RFC3971] - Arkko, J., Ed., Kempf, J., Zill, B., and P. Nikander, "SEcure Neighbor Discovery (SEND)", RFC 3971, March 2005.
- [RFC4861] - Narten, T., Nordmark, E., Simpson, W., and H. Soliman, "Neighbor Discovery for IP version 6 (IPv6)", RFC 4861, September 2007.
- [RFC5304] Li, T. and R. Atkinson, "IS-IS Cryptographic Authentication", RFC 5304, October 2008.
- [RFC5310] - Bhatia, M., Manral, V., Li, T., Atkinson, R., White, R., and M. Fanto, "IS-IS Generic Cryptographic Authentication", RFC 5310, February 2009.
- [RFC6165] - Banerjee, A. and D. Ward, "Extensions to IS-IS for Layer-2 Systems", RFC 6165, April 2011.
- [RFC6325] - Perlman, R., Eastlake 3rd, D., Dutt, D., Gai, S., and A. Ghanwani, "Routing Bridges (RBrIdges): Base Protocol Specification", RFC 6325, July 2011.
- [RFC7042] - Eastlake 3rd, D. and J. Abley, "IANA Considerations and IETF Protocol and Documentation Usage for IEEE 802 Parameters", BCP 141, RFC 7042, October 2013.
- [RFC7172] - Eastlake 3rd, D., Zhang, M., Agarwal, P., Perlman, R., and D. Dutt, "Transparent Interconnection of Lots of Links (TRILL): Fine-Grained Labeling", RFC 7172, May 2014, <<http://www.rfc-editor.org/info/rfc7172>>.
- [RFC7176] - Eastlake 3rd, D., Senevirathne, T., Ghanwani, A., Dutt, D., and A. Banerjee, "Transparent Interconnection of Lots of Links (TRILL) Use of IS-IS", RFC 7176, May 2014, <<http://www.rfc-editor.org/info/rfc7176>>.
- [RFC7178] - Eastlake 3rd, D., Manral, V., Li, Y., Aldrin, S., and D. Ward, "Transparent Interconnection of Lots of Links (TRILL): RBridge Channel Support", RFC 7178, May 2014, <<http://www.rfc->

[editor.org/info/rfc7178](http://www.rfc-editor.org/info/rfc7178)>.

[RFC7357] - Zhai, H., Hu, F., Perlman, R., Eastlake 3rd, D., and O. Stokes, "Transparent Interconnection of Lots of Links (TRILL): End Station Address Distribution Information (ESADI) Protocol", RFC 7357, September 2014, <<http://www.rfc-editor.org/info/rfc7357>>.

[rfc7180bis] - D. Eastlake 3rd, M. Zhang, A. Banerjee, A. Ghanwani, and S. Gupta "Transparent Interconnection of Lots of Links (TRILL): Clarifications, Corrections, and Updates", RFC 7180, May 2014, <<http://www.rfc-editor.org/info/rfc7180>>.

[IA] - Eastlake, D., L. Yizhou, R. Perlman, "TRILL: Interface Addresses APPsub-TLV", draft-ietf-trill-ia-appsubtlv, work in progress.

#### Informational References

[RFC7067] - Dunbar, L., Eastlake 3rd, D., Perlman, R., and I. Gashinsky, "Directory Assistance Problem and High-Level Design Proposal", RFC 7067, November 2013.

[ChannelTunnel] - D. Eastlake, M. Umair, Y. Li, "TRILL: RBridge Channel Tunnel Protocol", draft-ietf-trill-channel-tunnel, work in progress.

[ARPreduction] - Y. Li, D. Eastlake, L. Dunbar, R. Perlman, I. Gashinsky, "TRILL: ARP/ND Optimization", draft-ietf-trill-arp-optimization, work in progress.

#### Acknowledgments

The contributions of the following persons are gratefully acknowledged:

Gsyle Noble

The document was prepared in raw nroff. All macros used were defined within the source file.

Authors' Addresses

Donald Eastlake  
Huawei Technologies  
155 Beaver Street  
Milford, MA 01757 USA

Phone: +1-508-333-2270  
Email: d3e3e3@gmail.com

Linda Dunbar  
Huawei Technologies  
5430 Legacy Drive, Suite #175  
Plano, TX 75024, USA

Phone: +1-469-277-5840  
Email: ldunbar@huawei.com

Radia Perlman  
EMC  
2010 256th Avenue NE, #200  
Bellevue, WA 98007 USA

Email: Radia@alum.mit.edu

Igor Gashinsky  
Yahoo  
45 West 18th Street 6th floor  
New York, NY 10011

Email: igor@yahoo-inc.com

Yizhou Li  
Huawei Technologies  
101 Software Avenue,  
Nanjing 210012 China

Phone: +86-25-56622310  
Email: liyizhou@huawei.com

## Copyright, Disclaimer, and Additional IPR Provisions

Copyright (c) 2015 IETF Trust and the persons identified as the document authors. All rights reserved.

This document is subject to BCP 78 and the IETF Trust's Legal Provisions Relating to IETF Documents (<http://trustee.ietf.org/license-info>) in effect on the date of publication of this document. Please review these documents carefully, as they describe your rights and restrictions with respect to this document. Code Components extracted from this document must include Simplified BSD License text as described in Section 4.e of the Trust Legal Provisions and are provided without warranty as described in the Simplified BSD License. The definitive version of an IETF Document is that published by, or under the auspices of, the IETF. Versions of IETF Documents that are published by third parties, including those that are translated into other languages, should not be considered to be definitive versions of IETF Documents. The definitive version of these Legal Provisions is that published by, or under the auspices of, the IETF. Versions of these Legal Provisions that are published by third parties, including those that are translated into other languages, should not be considered to be definitive versions of these Legal Provisions. For the avoidance of doubt, each Contributor to the IETF Standards Process licenses each Contribution that he or she makes as part of the IETF Standards Process to the IETF Trust pursuant to the provisions of RFC 5378. No language to the contrary, or terms, conditions or rights that differ from or are inconsistent with the rights and licenses granted under RFC 5378, shall have any effect and shall be null and void, whether published or posted by such Contributor, or included with or in such Contribution.





TRILL working group  
Internet Draft  
Intended status: Standard Track  
Expires: April 2016

L. Dunbar  
D. Eastlake  
Huawei  
Radia Perlman  
Intel  
I. Gashinsky  
Yahoo  
October 12, 2015

Directory Assisted TRILL Encapsulation  
draft-ietf-trill-directory-assisted-encap-01.txt

Status of this Memo

This Internet-Draft is submitted in full conformance with the provisions of BCP 78 and BCP 79.

This Internet-Draft is submitted in full conformance with the provisions of BCP 78 and BCP 79. This document may not be modified, and derivative works of it may not be created, except to publish it as an RFC and to translate it into languages other than English.

Internet-Drafts are working documents of the Internet Engineering Task Force (IETF), its areas, and its working groups. Note that other groups may also distribute working documents as Internet-Drafts.

Internet-Drafts are draft documents valid for a maximum of six months and may be updated, replaced, or obsoleted by other documents at any time. It is inappropriate to use Internet-Drafts as reference material or to cite them other than as "work in progress."

The list of current Internet-Drafts can be accessed at <http://www.ietf.org/ietf/lid-abstracts.txt>

The list of Internet-Draft Shadow Directories can be accessed at <http://www.ietf.org/shadow.html>

This Internet-Draft will expire on April 12, 2016.

## Copyright Notice

Copyright (c) 2015 IETF Trust and the persons identified as the document authors. All rights reserved.

This document is subject to BCP 78 and the IETF Trust's Legal Provisions Relating to IETF Documents (<http://trustee.ietf.org/license-info>) in effect on the date of publication of this document. Please review these documents carefully, as they describe your rights and restrictions with respect to this document. Code Components extracted from this document must include Simplified BSD License text as described in Section 4.e of the Trust Legal Provisions and are provided without warranty as described in the Simplified BSD License.

## Abstract

This draft describes how data center network can benefit from non-RBridge nodes performing TRILL encapsulation with assistance from directory service.

## Table of Contents

1. Introduction.....	3
2. Conventions used in this document.....	3
3. Directory Assistance to Non-RBridge.....	4
4. Source Nickname in Frames Encapsulated by Non-RBridge Nodes.....	7
5. Benefits of Non-RBridge encapsulating TRILL header.	7
5.1. Avoid Nickname Exhaustion Issue.....	7
5.2. Reduce MAC Tables for switches on Bridged LANs..	8
6. Conclusion and Recommendation.....	9
7. Manageability Considerations.....	9
8. Security Considerations.....	9
9. IANA Considerations.....	9
10. References.....	10
10.1. Normative References.....	10
10.2. Informative References.....	10
11. Acknowledgments.....	10

## 1. Introduction

This draft describes how data center networks can benefit from non-RBridge nodes performing TRILL encapsulation with assistance from directory service.

[RFC7067] describes the framework for RBridge edge to get MAC&VLAN<->RBridgeEdge mapping from a directory service in data center environments instead of flooding unknown DAs across TRILL domain. If it has the needed directory information, any node, even a non-RBridge node, can perform the TRILL encapsulation. This draft is to describe the benefits and a scheme for non-RBridge nodes performing TRILL encapsulation.

## 2. Conventions used in this document

The key words "MUST", "MUST NOT", "REQUIRED", "SHALL", "SHALL NOT", "SHOULD", "SHOULD NOT", "RECOMMENDED", "MAY", and "OPTIONAL" in this document are to be interpreted as described in RFC-2119 [RFC2119].

In this document, these words will appear with that interpretation only when in ALL CAPS. Lower case uses of these words are not to be interpreted as carrying RFC-2119 significance.

AF           Appointed Forwarder RBridge port [RFC6439]

Bridge:     IEEE 802.1Q compliant device. In this draft, Bridge is used interchangeably with Layer 2 switch.

DA:                   Destination Address

DC:           Data Center

EoR:         End of Row switches in data center. Also known as Aggregation switches in some data centers

Host:        Application running on a physical server or a virtual machine. A host usually has at least one IP address and at least one MAC address.

SA:                   Source Address

ToR: Top of Rack Switch in data center. It is also known as access switches in some data centers.

TRILL-EN: TRILL Encapsulating node. It is a node that only performs the TRILL encapsulation but doesn't participate in RBridge's IS-IS routing.

VM: Virtual Machines

### 3. Directory Assistance to Non-RBridge

With directory assistance [RFC7067], a non-RBridge can be informed if a packet needs to be forwarded across the RBridge domain and the corresponding egress RBridge. Suppose the RBridge domain boundary starts at network switches (not virtual switches embedded on servers), a directory can assist Virtual Switches embedded on servers to encapsulate with a proper TRILL header by providing the nickname of the egress RBridge edge to which the destination is attached. The other information needed to encapsulate can be either learned by listening to TRILL Hellos, which will indicate the MAC address and nickname of appropriate edge RBridges, or by configuration.

If a destination is not attached to other RBridge edge nodes based on the directory [RFC7067], the non-RBridge node can forward the data frames natively, i.e. not encapsulating any TRILL header.

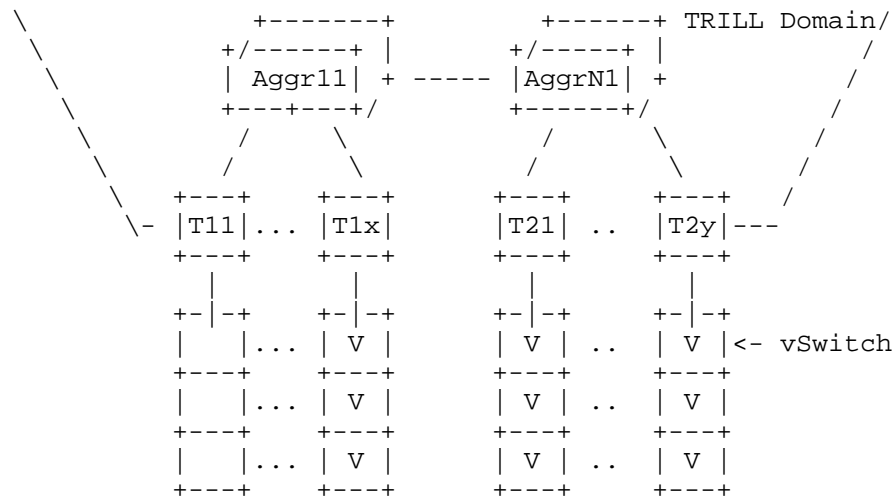


Figure 1 TRILL domain in typical Data Center Network

When a TRILL encapsulated data packet reaches the ingress RBridge, the ingress RBridge simply forwards the pre-encapsulated packet to the RBridge that is specified by the egress nickname field of the TRILL header of the data frame. When the ingress RBridge receives a native Ethernet frame, it handles it as usual and may drop it if it has complete directory information indicating that the target is not attached to the TRILL campus.

In this environment with complete directory information, the ingress RBridge doesn't flood or forward the received data frames when the DA in the Ethernet data frames is unknown.

When all attached nodes to ingress RBridge can pre-encapsulate TRILL header for traffic across the TRILL domain, the ingress RBridge don't need to encapsulate any native Ethernet frames to the TRILL domain. The attached nodes can be connected to multiple edge RBridges by having multiple ports or by an bridged LAN. Under this environment, there is no need to designate AF ports and all RBridge edge ports connected to one bridged LAN can receive and forward pre-encapsulated traffic, which can greatly improve the overall network utilization.

Note: [RFC6325] Section 4.6.2 Bullet 8 specifies that an RBridge port can be configured to accept TRILL encapsulated frames from a neighbor that is not an RBridge.

When a TRILL frame arrives at an RBridge whose nickname matches with the destination nickname in the TRILL header of the frame, the processing is exactly same as normal, i.e. the RBridge decapsulates the received TRILL frame and forwards the decapsulated frame to the target attached to its edge ports. When the DA of the decapsulated Ethernet frame is not in the egress RBridge's local MAC attachment tables, the egress RBridge floods the decapsulated frame to all attached links in the frame's VLAN, or drops the frame (if the egress RBridge is configured with the policy).

We call a node that only performs the TRILL encapsulation but doesn't participate in RBridge's IS-IS routing a TRILL Encapsulating node (TRILL-EN). The TRILL Encapsulating Node can get the MAC&VLAN<->RBridgeEdge mapping table pulled from directory servers [RFC7067].

Editor's note: RFC7067 has defined Push and Pull model for edge nodes to get directory mapping information. While Pull Model is relative simple for TRILL-EN to implement, Pushing requires some reliable flooding mechanism, like the one used by IS-IS, between the edge RBridge and the TRILL encapsulating node. Something like an extension to ES-IS might be needed.

Upon receiving a native Ethernet frame, the TRILL-EN checks the MAC&VLAN<->RBridgeEdge mapping table, and perform the corresponding TRILL encapsulation if the entry is found in the mapping table. If the destination address and VLAN of the received Ethernet frame doesn't exist in the mapping table and no positive reply from pulling request to a directory, the Ethernet frame is dropped or forwarded in native form to an edge RBridge.

Figure 2 Data frames from TRILL-EN

maintained by RBridge edge nodes and the necessity of enforcing AF ports.

Allowing Non-RBridge nodes to pre-encapsulate data frames with TRILL header makes it possible to have a TRILL domain with a reasonable number of RBridge nodes in a large data center. All the TRILL-ENs attached to one RBridge are represented by one TRILL nickname, which can avoid the Nickname exhaustion problem.

## 5.2. Reduce MAC Tables for switches on Bridged LANs

When hosts in a VLAN (or subnet) span across multiple RBridge edge nodes and each RBridge edge has multiple VLANs enabled, the switches on the bridged LANs attached to the RBridge edge are exposed to all MAC addresses among all the VLANs enabled.

For example, for an Access switch with 40 physical servers attached, where each server has 100 VMs, there are 4000 hosts under the Access Switch. If indeed hosts/VMs can be moved anywhere, the worst case for the Access Switch is when all those 4000 VMs belong to different VLANs, i.e. the access switch has 4000 VLANs enabled. If each VLAN has 200 hosts, this access switch's MAC table potentially has  $200 \times 4000 = 800,000$  entries.

If the virtual switches on servers pre-encapsulate the data frames destined for hosts attached to other RBridge Edge nodes, the outer MAC DA of those TRILL encapsulated data frames will be the MAC address of the local RBridge edge, i.e. the ingress RBridge. Therefore, the switches on the local bridged LAN don't need to keep the MAC entries for remote hosts attached to other edge RBridges.

But the traffic from nodes attached to other RBridges is decapsulated and has the true source and destination MACs. To prevent local bridges from learning remote hosts' MACs and adding to their MAC tables, one simple way is to disable this data plane learning on local bridges. The local bridges can be pre-configured with MAC addresses of local hosts with the assistance of a directory. The local bridges can always send frames with unknown Destination to the ingress RBridge. In an environment where a large number of VMs are instantiated in one server, the number of remote MAC addresses could be very large. If it is not feasible to disable learning and pre-configure MAC tables for local bridges, one effective method to minimize local bridges' MAC table size is to use the



server's MAC address to hide MAC addresses of the attached VMs. I.e. the server acting as an edge node using its own MAC address in the Source Address field of the packets originated from a host (or VM) embedded. When the Ethernet frame arrives at the target edge node (the server), the target edge node can send the packet to the corresponding destination host based on the packet's IP address. Very often, the target edge node communicates with the embedded VMs via a layer 2 virtual switch. Under this case, the target edge node can construct the proper Ethernet header with the assistance from directory. The information from directory includes the proper host IP to MAC mapping information.

## 6. Conclusion and Recommendation

When directory information is available, nodes outside the TRILL domain can encapsulate data frames destined for nodes attached to remote RBridges. The non-RBridge encapsulation approach is especially useful when there are a large number of servers in a data center equipped with hypervisor-based virtual switches. It is relatively easy for virtual switches, which are usually software based, to get directory assistance and perform network address encapsulation.

## 7. Manageability Considerations

It requires directory assistance to make it possible for a non-TRILL node to pre-encapsulate packets destined towards remote RBridges.

## 8. Security Considerations

Pull Directory queries and responses are transmitted as RBridge-to-RBridge or native RBridge Channel messages. Such messages can be secured as specified in [ChannelTunnel].

For general TRILL security considerations, see [RFC6325].

## 9. IANA Considerations

This document requires no IANA actions. RFC Editor:  
Please remove this section before publication.

## 10. References

### 10.1. Normative References

- [RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, March 1997.
- [RFC6325] Perlman, et, al, "Routing Bridges (RBridges): Base Protocol Specification", RFC6325, July 2011
- [RFC6439] Perlman, R., Eastlake, D., Li, Y., Banerjee, A., and F. Hu, "Routing Bridges (RBridges): Appointed Forwarders", RFC 6439, November 2011.

### 10.2. Informative References

- [RFC7067] Dunbar, et, al "Directory Assistance Problem and High-Level Design Proposal", RFC7067, Nov, 2013.
- [ChannelTunnel] - D. Eastlake, Y. Li, "TRILL: RBridge Channel Tunnel Protocol", draft-eastlake-trill-channel-tunnel, work in progress.

## 11. Acknowledgments

This document was prepared using 2-Word-v2.0.template.dot.

## Authors' Addresses

Linda Dunbar  
Huawei Technologies  
5340 Legacy Drive, Suite 175  
Plano, TX 75024, USA  
Phone: (469) 277 5840  
Email: linda.dunbar@huawei.com

Donald Eastlake  
Huawei Technologies  
155 Beaver Street  
Milford, MA 01757 USA  
Phone: 1-508-333-2270  
Email: d3e3e3@gmail.com

Radia Perlman  
Intel Labs  
2200 Mission College Blvd.  
Santa Clara, CA 95054-1549 USA  
Phone: 1-408-765-8080  
Email: Radia@alum.mit.edu

Igor Gashinsky  
Yahoo  
45 West 18th Street 6th floor  
New York, NY 10011  
Email: igor@yahoo-inc.com



INTERNET-DRAFT  
Intended status: Proposed Standard

Expires: December 29, 2014

Donald Eastlake  
Yizhou Li  
Huawei  
June 30, 2015

TRILL: Interface Addresses APPsub-TLV  
<draft-ietf-trill-ia-appsubtlv-05.txt>

#### Abstract

This document specifies a TRILL (Transparent Interconnection of Lots of Links) IS-IS application sub-TLV that enables the reporting by a TRILL switch of sets of addresses such that all of the addresses in each set designate the same interface (port) and the reporting for such a set of the TRILL switch by which it is reachable. For example, a 48-bit MAC (Media Access Control) address, IPv4 address, and IPv6 address can be reported as all corresponding to the same interface reachable by a particular TRILL switch. Such information could be used in some cases to synthesize responses to or by-pass the need for the Address Resolution Protocol (ARP), the IPv6 Neighbor Discovery (ND) protocol, or the flooding of unknown MAC addresses.

#### Status of This Memo

This Internet-Draft is submitted to IETF in full conformance with the provisions of BCP 78 and BCP 79.

Distribution of this document is unlimited. Comments should be sent to the TRILL working group mailing list.

Internet-Drafts are working documents of the Internet Engineering Task Force (IETF), its areas, and its working groups. Note that other groups may also distribute working documents as Internet-Drafts.

Internet-Drafts are draft documents valid for a maximum of six months and may be updated, replaced, or obsoleted by other documents at any time. It is inappropriate to use Internet-Drafts as reference material or to cite them other than as "work in progress."

The list of current Internet-Drafts can be accessed at <http://www.ietf.org/lid-abstracts.html>. The list of Internet-Draft Shadow Directories can be accessed at <http://www.ietf.org/shadow.html>.

## Table of Contents

1. Introduction.....	3
1.1 Conventions Used in This Document.....	3
2. Format of the Interface Addresses APPsub-TLV.....	5
3. IA APPsub-TLV sub-sub-TLVs.....	11
3.1 AFN Size sub-sub-TLV.....	11
3.2 Fixed Address sub-sub-TLV.....	12
3.3 Data Label sub-sub-TLV.....	13
3.4 Topology sub-sub-TLV.....	13
4. Security Considerations.....	15
5. IANA Considerations.....	16
5.1 AFN Number Allocation.....	16
5.2 IA APPsub-TLV Sub-Sub-TLVs SubRegistry.....	17
5.3 IA APPsub-TLV Number.....	17
Acknowledgments.....	18
Appendix A: Examples.....	19
A.1 Simple Example.....	19
A.2 Complex Example.....	19
Appendix Z: Change History.....	22
Normative References.....	23
Informational References.....	24
Authors' Addresses.....	25

## 1. Introduction

This document specifies a TRILL (Transparent Interconnection of Lots of Links) [RFC6325] IS-IS application sub-TLV (APPsub-TLV [RFC6823]) that enables the convenient representation of sets of addresses such that all of the addresses in each set designate the same interface (port). For example, a 48-bit MAC (Media Access Control [RFC7042]) address, IPv4 address, and IPv6 address can be reported as all three designating the same interface. In addition, a Data Label (VLAN or Fine Grained Label (FGL [RFC7172])) is specified for the interface along with the TRILL switch, and optionally the TRILL switch port, from which the interface is reachable. Such information could be used in some cases to synthesize responses to or by-pass the need for the Address Resolution Protocol (ARP [RFC826]), the IPv6 Neighbor Discovery (ND [RFC4861]) protocol, the Reverse Address Resolution Protocol (RARP [RFC903]), or the flooding of unknown destination MAC addresses [RFC7042]. If the information reported is complete, it can also be used to detect and discard packets with forged source addresses.

This APPsub-TLV appears inside the TRILL GENINFO TLV specified in ESADI [RFC7357] but may also occur in other application contexts. Directory Assisted TRILL Edge services [DirectoryScheme] are expected to make use of this APPsub-TLV.

Although, in some IETF protocols, address field types are represented by Ethertype [RFC7042] or Hardware Type [RFC5494], only Address Family Number (AFN) is used in this APPsub-TLV to represent address field type.

### 1.1 Conventions Used in This Document

The key words "MUST", "MUST NOT", "REQUIRED", "SHALL", "SHALL NOT", "SHOULD", "SHOULD NOT", "RECOMMENDED", "MAY", and "OPTIONAL" in this document are to be interpreted as described in [RFC2119]. Capitalized IANA Considerations terms such as "Expert Review" are to be interpreted as described in [RFC5226].

The terminology and acronyms of [RFC6325] are used herein along with the following additional acronyms and terms:

AFN: Address Family Number

APPsub-TLV: Application sub-TLV [RFC6823]

Data Label: VLAN or FGL

FGL: Fine Grained Label [RFC7172]

IA: Interface Addresses

RBridge: An alternative name for a TRILL switch

TRILL switch: A device that implements the TRILL protocol



## 2. Format of the Interface Addresses APPsub-TLV

The Interface Addresses (IA) APPsub-TLV is used to advertise that a set of addresses indicate the same interface (port) within a Data Label (VLAN or FGL) and to associate that interface with the TRILL switch, and optionally the TRILL switch port, by which the interface is reachable. These addresses can be in different address families. For example, it can be used to declare that a particular interface with specified IPv4, IPv6, and 48-bit MAC addresses in some particular Data Label is reachable from a particular TRILL switch.

The Template field in a particular Interface Addresses APPsub-TLV indicates the format of each Address Set it carries. Certain well-known sets of addresses are represented by special values. Other sets of addresses are specified by a list of AFNs. The Template format that uses a list of AFNs provides an explicit pattern for the type and order of addresses in each Address Set in the IA APPsub-TLV that includes that Template.

A device or application making use of IA APPsub-TLV data is not required to make use of all IA data. For example, a device or application that was only interested in MAC and IPv6 addresses could ignore any IPv4 or other types of address information that was present.

The figure below shows an IA APPsub-TLV as it would appear inside an IS-IS FS-LSP using an extended flooding scope [RFC7356] TLV, for example in ESADI [RFC7357]. Within an IS-IS FS-LSP using traditional [ISO-10589] TLVs, the Type and Length would be one byte unsigned integers equal to or less than 255.

```

+-----+-----+
| Type = TBD1                                     | (2 bytes)
+-----+-----+
| Length                                           | (2 bytes)
+-----+-----+
| Addr Sets End                                   | (2 bytes)
+-----+-----+
| Nickname                                         | (2 bytes)
+-----+-----+
| Flags                                           | (1 byte)
+-----+-----+
| Confidence                                       | (1 byte)
+-----+-----+
| Template ...                                   (variable)
+-----+-----+...+
| Address Set 1   (size determined by Template)   |
+-----+-----+...+
| Address Set 2   (size determined by Template)   |
+-----+-----+...+
| ...
+-----+-----+...+
| Address Set N   (size determined by Template)   |
+-----+-----+...+
| optional sub-sub-TLVs ...
+-----+-----+...

```

Figure 1. The Interface Addresses APPsub-TLV

- o Type: Interface Addresses TRILL APPsub-TLV type, set to TBD1 (IA-SUBTLV).
- o Length: Variable, minimum 7. If length is 6 or less or if the APPsub-TLV extends beyond the size of an encompassing TRILL GENINFO TLV or other context, the APPsub-TLV MUST be ignored.
- o Addr Sets End: The unsigned integer offset of the byte, within the IA APPsub-TLV value part, of the last byte of the last Address Set. This will be the byte just before the first sub-sub-TLV if any sub-sub-TLVs are present (see Section 3). If this is equal to Length, there are no sub-sub-TLVs. If this is greater than Length or points to before the end of the Template, the IA APPsub-TLV is corrupt and MUST be discarded. This field is always two bytes in size.
- o Nickname: The nickname of the TRILL switch by which the address sets are reachable. If zero, the address sets are reachable from the TRILL switch originating the message containing the APPsub-TLV (for example, an ESADI [RFC7357] message).

- o Flags: A byte of flags as follows:

```

  0 1 2 3 4 5 6 7
+---+---+---+---+
|D|L|N|  RESV  |
+---+---+---+---+

```

D: Directory flag: If D is one, the APPsub-TLV contains Directory information [RFC7067].

L: Local flag: If L is one, the APPsub-TLV contains information learned locally by observing ingressed frames [RFC6325]. (Both D and L can be set to one in the same IA APPsub-TLV if a TRILL switch that had learned an address locally and also advertised it as a directory.)

N: Notify flag: When a TRILL switch receives a new IA APPsub-TLV (one in an ESADI-LSP fragment with a higher sequence number or a new message of some other type) and the N bit is one, the TRILL switch then checks the contents of the APPsub-TLV for address sets including both an IP address and a MAC address. For each such address set it finds, a gratuitous ARP [RFC826] or spontaneous Neighbor Advertisement [RFC4861], depending on whether the IP address is IPv4 or IPv6 respectively, may be sent. In both cases, these are sent out all the ports of the TRILL switch offering end station service and are in the VLAN or FGL of the address set information, that is, are Appointed Forwarder for the VLAN or for the VLAN to which the FGL maps.

RESV: Additional reserved flag bits that MUST be sent as zero and ignored on receipt.

- o Confidence: This 8-bit unsigned quantity in the range 0 to 254 indicates the confidence level in the addresses being transported (see Section 4.8.2 of [RFC6325]). A value of 255 is treated as if it was 254.
- o Template: The initial byte of this field is the unsigned integer K. If K has a value from 1 to 31, it indicates that this initial byte is followed by a list of K AFNs (Address Family Numbers) that specify the exact structure and order of each Address Set occurring later in the APPsub-TLV. K can be 1, which is the minimum valid value. If K is zero, the IA APPsub-TLV is ignored. If K is 32 to 254, the length of the Template field is one byte and its value is intended to correspond to a particular ordered set of AFNs some of which are specified below. If K is 255, the length of the Template field is three bytes and the values of the second and third byte, considered as an unsigned integer in

network byte order, are reserved to correspond to future specified ordered sets of AFNs.

If the Template uses explicit AFNs, it looks like the following, with the number of AFNs, up to 31, equal to K.

```

+-----+
|  K      | (1 byte)
+-----+
|  AFN 1   | (2 bytes)
+-----+
|  AFN 2   | (2 bytes)
+-----+
|   ...   |
+-----+
|  AFN K   | (2 bytes)
+-----+

```

For K in the 32 to 102 range, values indicate combinations of a specific number of MAC addresses, IPv4 addresses, IPv6 addresses, and TRILL switch port IDs appearing in that order. The value of K is

$$K = 31 + M + 3*v4 + 9*v6 + 36*P$$

where M is 0, 1, or 2 (0 if no MAC address is present, 1 if a 48-bit MAC is present, 2 if a MAC/24 (see Section 5.1) is present), v4 is the number of IPv4 addresses (limited to 0, 1, or 2) and v6 is the number of IPv6 addresses (limited to 0 through 3 inclusive), and P is the number of TRILL switch port IDs (limited to 0 or 1); however, the number of MAC, IPv4, and IPv6 addresses and TRILL switch ports cannot all be simultaneously zero. It is important that, when using this encoding, the values of M, v4, v6, and P do not exceed the limits given; otherwise, they cannot be unambiguously decoded. For example, v4 is limited to 0, 1, or 2. Attempting to encode a v4 value of 3 is indistinguishable from incrementing the v6 value by 1.

Given that M, v4, v6, and P may not all be zero, this equation specifies values of K from 32 through 102. The value 31 is not permitted but instead represents an explicit Template with 31 AFNs. Values from 103 through 254 of the byte value are available for assignment by Expert Review (see Section 5). K = 255 indicates a three-byte Template field as specified above. All values (0 through 65,545) of this two-byte value are available for assignment by Expert Review.

If an unknown Template K value in the range 103 to 254 is received or a K of 255 followed by an unknown two byte value, the IA APPsub-TLV MUST be ignored.

- o AFN: A two-byte Address Family Number. The number of AFNs present is given by K except that there are no AFNs if K is greater than 31. The AFN sequence specifies the structure of the Address Sets occurring later in the TLV. For example, if Template Size is 2 and the two AFNs present are the AFNs for a 48-bit MAC and an IPv4 address, in that order, then each Address set present will consist of a 6-byte MAC address followed by a 4-byte IPv4 address. If any AFNs are present that are unknown to the receiving IS and the length of the corresponding address is not provided by a sub-sub-TLV as specified below, the receiving IS will be unable to parse the Address Sets and MUST ignore the IA APPsub-TLV.
- o Address Set: Each address set in the APPsub-TLV consists of exactly the same sequence of addresses and types as specified by the Template earlier in the APPsub-TLV. No alignment, other than to a byte boundary, is provided. The addresses in each Address Set are contiguous with no unused bytes between them and the Address Sets are contiguous with no unused bytes between successive Address Sets. The Address Sets must fit within the TLV.
- o sub-sub-TLVs: If the Address Sets indicated by Addr Sets End do not completely fill the Length of the APPsub-TLV, the remaining bytes are parsed as sub-sub-TLVs [RFC5305]. Any such sub-sub-TLVs that are not known to the receiving TRILL switch are ignored. Should this parsing not be possible, for example there is only one remaining byte or an apparent sub-sub-TLV extends beyond the end of the TLV, the containing IA APPsub-TLV is considered corrupt and is ignored. (Several sub-sub-TLV types are specified in Section 3.)

Different IA APPsub-TLVs within the same or different LSPs or other data structures may have different Templates. The same AFN may occur more than once in a Template and the same address may occur in different address sets. For example, a 48-bit MAC address interface might have three different IPv6 addresses. This could be represented by an IA APPsub-TLV whose Template specifically provided for one EUI-48 address and three IPv6 addresses, which might be an efficient format if there were multiple interfaces with that pattern. Alternatively, a Template with one 48-bit MAC and one IPv6 address could be used in an IA APPsub-TLV with three address sets each having the same MAC address but different IPv6 addresses, which might be the most efficient format if only one interface had multiple IPv6 addresses and other interfaces had only one IPv6 address.

In order to be able to parse the Address Sets, a receiving TRILL switch must know at least the size of the address for each AFN or address type the Template specifies; however, the presence of the Addr Set End field means that the sub-sub-TLVs, if any, can always be located by a receiver. A TRILL switch can be assumed to know the size of the AFNs mentioned in Section 5. Should a TRILL switch wish

to include an AFN that some receiving TRILL switch in the campus may not know, it SHOULD include an AFN-Size sub-sub-TLV as described in Section 3.1. If an IA APPsub-TLV is received with one or more AFNs in its template for which the receiving TRILL switch does not know the length and for which an AFN-Size sub-sub-TLV is not present, that IA APPsub-TLV MUST be ignored.

### 3. IA APPsub-TLV sub-sub-TLVs

IA APPsub-TLVs can have trailing sub-sub-TLVs [RFC5305] as specified below. These sub-sub-TLVs occur after the Address Sets and the amount of space available for sub-sub-TLVs is determined from the overall IA APPsub-TLV length and the value of the Addr Set End byte.

There is no ordering restriction on sub-sub-TLVs. Unless otherwise specified each sub-sub-TLV type can occur zero, one, or many times in an IA APPsub-TLV. Any sub-sub-TLVs for which the Type is unknown are ignored.

The sub-sub-TLVs data structures shown below, with two byte Types and Lengths, assume that the enclosing IA-APPsubTLV is in an extended LSP TLV [RFC7356] or some non-LSP context. If they were used in a IA-APPsubTLV in a non-extended LSP [ISO-10589], then only one byte Types and Lengths could be used. As a result, any sub-sub-TLV types greater than 255 could not be used and Length would be limited to 255.

#### 3.1 AFN Size sub-sub-TLV

Using this sub-sub-TLV, the originating TRILL switch can specify the size of an address type. This is useful under two circumstances as follows:

1. One or more AFNs that are unknown to the receiving TRILL switch appears in the template. If an AFN Size sub-sub-TLV is present for each such AFN, then at least the IA APPsub-TLV can be parsed and possibly other addresses in each address set can still be used.
2. If an AFN occurs in the Template that represents a variable length address, this sub-sub-TLV gives its size for all occurrences in that IA APPsub-TLV.

```

+++++
| Type = AFNsz                               | (2 byte)
+++++
| Length                                     | (2 byte)
+++++
| AFN Size Record 1                         | (3 bytes)
+++++
| AFN Size Record 2                         | (3 bytes)
+++++
| ...
+++++
| AFN Size Record N                         | (3 bytes)
+++++

```

Where each AFN Size Record is structured as follows:

```

+---+---+---+---+---+---+---+---+---+---+
|  AFN                                         | (2 bytes)
+---+---+---+---+---+---+---+---+---+---+
|  AddrSize                                   | (1 byte)
+---+---+---+---+---+

```

- o Type: AFN-Size sub-sub-TLV type, set to 1 (AFNsz).
- o Length: 3\*n where n is the number of AFN Size Records present. If Length is not a multiple of 3, the sub-sub-TLV MUST be ignored.
- o AFN Size Record(s): Zero or more 3-byte records, each giving the size of an address type identified by an AFN,
- o AFN: The AFN whose length is being specified by the AFN Size Record.
- o AddrSize: The length in bytes of addresses specified by the AFN field as an unsigned integer.

An AFN Size sub-sub-TLV for any AFN known to the receiving TRILL switch is compared with the size known to the TRILL switch. If they differ the IA APPsub-TLV is assumed to be corrupt and MUST be ignored.

### 3.2 Fixed Address sub-sub-TLV

There may be cases where, in a particular Interface Addresses APP-subTLV, the same address would appear in every address set across the APP-subTLV. To avoid wasted space, this sub-sub-TLV can be used to indicate such a fixed address. The address or addresses incorporated into the sets by this sub-sub-TLV are NOT mentioned in the IA APPsub-TLV Template.

```

+---+---+---+---+---+---+---+---+---+---+
| Type=FIXEDADR                               | (2 byte)
+---+---+---+---+---+---+---+---+---+---+
| Length                                       | (2 byte)
+---+---+---+---+---+---+---+---+---+---+
| AFN                                         | (2 bytes)
+---+---+---+---+---+---+---+---+---+---+
| Fixed Address                               | (variable)
+---+---+---+---+---+---+---+---+---+---+...

```

- o Type: Data Label sub-sub-TLV type, set to 2 (FIXEDADR).



- o Length: variable, minimum 2. If Length is 0 or 1 or less, the sub-sub-TLV MUST be ignored.
- o AFN: Address Family Number of the Fixed Address.
- o Fixed Address: The address of the type indicated by the preceding AFN field that is considered to be part of every Address Set in the IA APPsub-TLV.

The Length field implies a size for the Fixed Address. If that size differs from the size of the address type for the given AFN as known by the receiving TRILL switch, the Fixed Address sub-sub-TLV is considered corrupt and MUST be ignored.

### 3.3 Data Label sub-sub-TLV

This sub-sub-TLV indicates the Data Label within which the interfaces listed in the IA APPsub-TLV are reachable. It is useful if the IA APPsub-TLV occurs outside of the context of a message specifying the Data Label or if it is desired and permitted to override that specification. Multiple occurrences of this sub-sub-TLV indicate that the interfaces are reachable in all of the Data Labels given.

```

+-----+
|Type=DATALEN                               | (2 byte)
+-----+
| Length                                   | (2 byte)
+-----+
| Data Label                               | (variable)
+-----+

```

- o Type: Data Label sub-TLV type, set to 3 (LABEL).
- o Length: 2 or 3. If Length is some other value, the sub-sub-TLV MUST be ignored.
- o Data Label: If length is 2, the bottom 12 bits of the Data Label are a VLAN ID and the top 4 bits are reserved (MUST be sent as zero and ignored on receipt). If the length is 3, the three Data Label bytes contain an FGL [RFC7172].

### 3.4 Topology sub-sub-TLV

The presence of this sub-sub-TLV indicates that the interfaces given in the IA APPsub-TLV are reachable in the topology given. It is useful if the IA APPsub-TLV occurs outside of the context of a

message indicating the topology or if it is desired and permitted to override that specification. If it occurs multiple times, then the Address Sets are in all of the topologies given.

```

+-----+
|Type=DATALEN| (2 byte)
+-----+
| Length| (2 byte)
+-----+
| RESV | Topology | (2 bytes)
+-----+

```

- o Type: Topology sub-TLV type, set to 4 (TOPOLOGY).
- o Length: 2. If Length is some other values, the sub-sub-TLV MUST be ignored.
- o RESV: Four reserved bits. MUST be sent as zero and ignored on receipt.
- o Topology: The 12-bit topology number [RFC5120].

#### 4. Security Considerations

The integrity of address mapping and reachability information and the correctness of Data Labels (VLANs or FGLs [RFC7172]) are very important. Forged, altered, or incorrect address mapping or Data Labeling can lead to delivery of packets to the incorrect party, violating security policy. However, this document merely describes a data format and does not provide any explicit mechanisms for securing that information, other than a few simple consistency checks that might detect some corrupted data. Security on the wire, or in storage, for this data is to be providing by the transport or storage used. For example, when transported with ESADI [RFC7357] or RBridge Channel [RFC7178], ESADI security or Channel Tunnel [ChannelTunnel] security mechanisms can be used, respectively.

The address mapping and reachability information, if known to be complete and correct, can be used to detect some cases of forged packet source addresses [RFC7067]. In particular, if native traffic from an end station is received by a TRILL switch that would otherwise accept it but authoritative data indicates the source address should not be reachable from the receiving TRILL switch, that traffic should be discarded. The data format specified in this document may optionally include TRILL switch Port ID number so that this forged address filtering can be optionally applied with port granularity.

See [RFC6325] for general TRILL Security Considerations.

## 5. IANA Considerations

The following subsections specify IANA actions.

### 5.1 AFN Number Allocation

IANA has allocated the following AFN values that may be useful for IA APPsub-TLVs:

Hex -----	Decimal -----	Description -----	References -----
0001	1	IPv4	
0002	2	IPv6	
4005	16389	48-bit MAC	[RFC7042]
4006	16390	64-bit MAC	[RFC7042]
4007	16391	OUI	This document.
4008	16392	MAC/24	This document.
4009	16393	MAC/40	This document.
400A	16394	IPv6/64	This document.
400B	16395	RBridge Port ID	This document.

Other AFNs can be found at <http://www.iana.org/assignments/address-family-numbers>

The OUI AFN is provided so that MAC addresses can be abbreviated if they have the same upper 24 bits. A MAC/24 is a 24-bit suffix intended to be pre-fixed by an OUI to create a 48-bit MAC address [RFC7042]; in the absence of an OUI, a MAC/24 entry cannot be used. A MAC/40 is a suffix intended to be pre-fixed by an OUI to create a 64-bit MAC address [RFC7042]; in the absence of an OUI, a MAC/40 entry cannot be used.

Typically, an OUI would be provided as a Fixed Address sub-sub-TLV (see Section 3.2).

After Fixed Address sub-sub-TLV processing above, each address set is processed by combining each OUI in the address set with each MAC/24 and each MAC/40 address in the address set. Depending on how many of each of these address types is present, zero or more 48-bit and/or 64-bit MAC addresses may be produced that are considered to be part of the address set. If there are no MAC/24 or MAC/40 addresses present, any OUI's are ignored. If there are no OUIs, any MAC/24 and/or MAC/40s are ignored. If there are K1 OUIs, K2 MAC/24s, and K3 MAC/40s, K1\*K2 48-bit MACs are synthesized and K1\*K3 64-bit MACs are synthesized.

IPv6/64 is an 8-byte quantity that is the first 64 bits of an IPv6

address. IPv6/64s are ignored unless, after the processing above in this sub-section, there are one or more 48-bit and/or 64-bit MAC addresses in the address set to provide the lower 64 bits of the IPv6 address. For this purpose, an 48-bit MAC address is expanded to 64 bits as described in [RFC7042]. If there are K4 IPv6/64s present and K5 48- and 64-bit MAC addresses present, K4\*K5 128-bit IPv6 addresses are synthesized.

## 5.2 IA APPsub-TLV Sub-Sub-TLVs SubRegistry

IANA is requested to establish a new subregistry of the TRILL Parameter Registry for sub-sub-TLVs of the Interface Addresses APPsub-TLV with initial contents as shown below.

Name: Interface Addresses APPsub-TLV Sub-Sub-TLVs

Procedure: Expert Review

Note: Types greater than 255 are not usable in some contexts.

Reference: [This document]

Type	Description	Reference
-----	-----	-----
0	Reserved	
1	AFN Size	[This document]
2	Fixed Address	[This document]
3	Data Label	[This document]
4	Topology	[This document]
5-254	Available	
255	Reserved	
256-65534	Available	
65535	Reserved	

## 5.3 IA APPsub-TLV Number

IANA is requested to allocate TBD1 as the Type for the IA APPsub-TLV in the "TRILL APPsub-TLV Types under IS-IS TLV 251 Application Identifier 1" registry from the range under 256. In the registry the Name is "IA" and the Reference is this document.

#### Acknowledgments

The authors gratefully acknowledge the contributions and review by the following:

Linda Dunbar and Gayle Noble

The document was prepared in raw nroff. All macros used were defined within the source file.

## Appendix A: Examples

Below are example IA APPsub-TLVs. "0x" indicates that the following quantity is in hexadecimal. "0b" indicates that the following quantity is in binary. Leading zeros are retained.

### A.1 Simple Example

Below is an annotated IA APPsub-TLV carrying two simple pairs of EUI-48 MAC addresses and IPv4 addresses from a Push Directory [RFC7067]. No sub-sub-TLVs are included.

```

0x0002(TBD)   Type: Interface Addresses
0x001B        Length: 27 (=0x1B)
0x001B        Address Sets End: 27 (=0x1B)
0x1234        RBridge Nickname from which reachable
0b10000000    Flags: Push Directory data
0xE3          Confidence = 227
35            Template: 35 (0x23) = 31 + 1(MAC48) + 3*1(IPv4)

```

#### Address Set One

```

0x00005E0053A9  48-bit MAC address
198.51.100.23    IPv4 address

```

#### Address Set Two

```

0x00005E00536B  48-bit MAC address
203.0.113.201   IPv4 address

```

Size includes 7 for the fixed fields though and including the one byte template, plus 2 times the Address Set size. Each Address Set is 10 bytes, 6 for the 48-bit MAC address plus 4 for the IPv4 address. So total size is  $7 + 2*10 = 27$ .

See Section 2 for more information on Template.

### A.2 Complex Example

Below is an annotated IA APPsub-TLV carrying three sets of addresses, each consisting of an EUI-48 MAC address, an IPv4 addresses, an IPv6 address, and an RBridge Port ID, all from a Push Directory [RFC7067]. The IPv6 address for each address set is synthesized from the MAC address given in that set and the IPv6/64 64-bit prefix provided through a Fixed Address sub-sub-TLV. In addition, a sub-sub-TLV is included that provides an FGL which overrides whatever Data Label may be provided by the envelope (for example an ESADI-LSP [RFC7357]) within which this IA APPsub-TLV occurs.

```

0x0002(TBD)    Type: Interface Addresses
0x0036         Length: 54 (=0x36)
0x0021         Address Sets End: 33 (=0x21)
0x4321         RBridge Nickname from which reachable
0b10000000    Flags: Push Directory data
0xD3          Confidence = 211
72            Template: 72(0x48)=31+1(MAC48)+3*1(IPv4)+36*1(P)

```

#### Address Set One

```

0x00005E0053DE 48-bit MAC address
198.51.100.105  IPv4 address
0x1DE3         RBridge Port ID

```

#### Address Set Two

```

0x00005E0053E3 48-bit MAC address
203.0.113.89   IPv4 address
0x1DEE         RBridge Port ID

```

#### Address Set Three

```

0x00005E0053D3 48-bit MAC address
192.0.2.139    IPv4 address
0x01DE         RBridge Port ID

```

#### sub-sub-TLV One

```

0x0003         Type: Data Label
0x0003         Length: implies FGL
0xD3E3E3      Fine Grained Label

```

#### sub-sub-TLV Two

```

0x0002         Type: Fixed Address
0x000A         Size: 0x0A = 10
0x400A         AFN: IPv6/64
0x20010DB800000000 IPv6 Prefix: 2001:DB8::

```

See Section 2 for more information on Template.

The Fixed Address sub-sub-TLV causes the IPv6/64 value given to be treated as if it occurred as a 4th entry inside each of the three Address Sets. When there is an IPv6/64 entry and a 48-bit MAC entry, the MAC value is expanded by inserting 0xFFFFE immediately after the OUI and the resulting 64-bit value is used as the lower 64 bits of the resulting IPv6 address [RFC7042]. As a result, a receiving TRILL switch would treat the three Address Sets shown as if they had an IPv6 address in them as follows:



## Address Set One

0x20010DB800000000000005EFFFFE0053DE IPv6 Address

## Address Set Two

0x20010DB800000000000005EFFFFE0053E3 IPv6 Address

## Address Set Three

0x20010DB800000000000005EFFFFE0053D3 IPv6 Address

As an alternative to the compact "well know value" Template encoding used in this example above, the less compact explicit AFN encoding could have been used. In that case, the IA APPsub-TLV would have started as follows:

0x0002(TBD)	Type: Interface Addresses
0x003C	Length: 60 (=0x3C)
0x0027	Address Sets End: 39 (=0x27)
0x4321	RBridge Nickname from which reachable
0b10000000	Flags: Push Directory data
0xD3	Confidence = 211
0x3	Template: 3 AFNs
0x4005	AFN: 48-bit MAC
0x0001	AFN: IPv4
0x400B	AFN: RBridge Port ID

As a final point, since the 48-bit MAC addresses in these three Address Sets all have the same OUI (the IANA OUI [RFC7042]), it would have been possible to just have a MAC/24 value giving the lower 24 bits of the MAC in each Address Set. The OUI would then be supplied by a second Fixed Address sub-sub-TLV providing the OUI. With N Address Sets, this would have saved 3\*N or 9 bytes in this case at the cost of 9 bytes (2 each for the type and length of the sub-sub-TLV, 2 for the OUI AFN number, and 3 for the OUI). So, with just three Address Sets, there would be no net saving; however, with a larger number of Address Sets, there would be a net savings.

Appendix Z: Change History

From -00 to -01

1. Update references for RFC publications.
2. Add this Change History Appendix.

From -01 to -02

1. Fix off-by-one errors in body text and examples for well known Template values.
2. Update for drafts published as RFCs and change in Author Address.
3. Minor editorial improvements.

From -02 to -03

Minor editorial improvements.

From -03 to -04

Editorial improvements.

From -04 to -05

Remove one author.

## Normative References

- [ISO-10589] - ISO/IEC 10589:2002, Second Edition, "Intermediate System to Intermediate System Intra-Domain Routing Exchange Protocol for use in Conjunction with the Protocol for Providing the Connectionless-mode Network Service (ISO 8473)", 2002.
- [RFC826] - Plummer, D., "An Ethernet Address Resolution Protocol", RFC 826, November 1982.
- [RFC903] - Finlayson, R., Mann, T., Mogul, J., and M. Theimer, "A Reverse Address Resolution Protocol", STD 38, RFC 903, June 1984.
- [RFC2119] - Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, March 1997
- [RFC4861] - Narten, T., Nordmark, E., Simpson, W., and H. Soliman, "Neighbor Discovery for IP version 6 (IPv6)", RFC 4861, September 2007.
- [RFC5120] - Przygienda, T., Shen, N., and N. Sheth, "M-ISIS: Multi Topology (MT) Routing in Intermediate System to Intermediate Systems (IS-ISs)", RFC 5120, February 2008.
- [RFC5226] - Narten, T. and H. Alvestrand, "Guidelines for Writing an IANA Considerations Section in RFCs", BCP 26, RFC 5226, May 2008.
- [RFC5305] - Li, T. and H. Smit, "IS-IS Extensions for Traffic Engineering", RFC 5305, October 2008.
- [RFC6325] - Perlman, R., Eastlake 3rd, D., Dutt, D., Gai, S., and A. Ghanwani, "Routing Bridges (RBridges): Base Protocol Specification", RFC 6325, July 2011.
- [RFC6823] - Ginsberg, L., Previdi, S., and M. Shand, "Advertising Generic Information in IS-IS", RFC 6823, December 2012.
- [RFC7042] - Eastlake 3rd, D. and J. Abley, "IANA Considerations and IETF Protocol and Documentation Usage for IEEE 802 Parameters", BCP 141, RFC 7042, October 2013.
- [RFC7172] - Eastlake 3rd, D., Zhang, M., Agarwal, P., Perlman, R., and D. Dutt, "Transparent Interconnection of Lots of Links (TRILL): Fine-Grained Labeling", RFC 7172, May 2014.
- [RFC7356] - Ginsberg, L., Previdi, S., and Y. Yang, "IS-IS Flooding Scope Link State PDUs (LSPs)", RFC 7356, September 2014, <<http://www.rfc-editor.org/info/rfc7356>>.

- [RFC7357] - Zhai, H., Hu, F., Perlman, R., Eastlake 3rd, D., and O. Stokes, "Transparent Interconnection of Lots of Links (TRILL): End Station Address Distribution Information (ESADI) Protocol", RFC 7357, September 2014, <<http://www.rfc-editor.org/info/rfc7357>>.

#### Informational References

- [ARP reduction] - Shah, et. al., "ARP Broadcast Reduction for Large Data Centers", draft-shah-armd-arp-reduction, work in progress.
- [ChannelTunnel] - D. Eastlake, Y. Li, "TRILL: RBridge Channel Tunnel Protocol", draft-eastlake-trill-channel-tunnel, work in progress.
- [DirectoryScheme] - Dunbar, L., D. Eastlake, R. Perlman, I. Gashinsky, Y. Li, "TRILL: Directory Assistance Mechanisms", draft-dunbar-trill-scheme-for-directory-assist, work in progress.
- [RFC5494] - Arkko, J. and C. Pignataro, "IANA Allocation Guidelines for the Address Resolution Protocol (ARP)", RFC 5494, April 2009.
- [RFC7067] - Dunbar, L., Eastlake 3rd, D., Perlman, R., and I. Gashinsky, "Directory Assistance Problem and High-Level Design Proposal", RFC 7067, November 2013.
- [RFC7178] - Eastlake 3rd, D., Manral, V., Li, Y., Aldrin, S., and D. Ward, "Transparent Interconnection of Lots of Links (TRILL): RBridge Channel Support", RFC 7178, May 2014.

Authors' Addresses

Donald Eastlake  
Huawei Technologies  
155 Beaver Street  
Milford, MA 01757 USA

Phone: +1-508-333-2270  
Email: d3e3e3@gmail.com

Yizhou Li  
Huawei Technologies  
101 Software Avenue,  
Nanjing 210012 China

Phone: +86-25-56622310  
Email: liyizhou@huawei.com

Radia Perlman  
EMC  
2010 256th Avenue NE, #200  
Bellevue, WA 98007 USA

Email: Radia@alum.mit.edu

## Copyright, Disclaimer, and Additional IPR Provisions

Copyright (c) 2015 IETF Trust and the persons identified as the document authors. All rights reserved.

This document is subject to BCP 78 and the IETF Trust's Legal Provisions Relating to IETF Documents (<http://trustee.ietf.org/license-info>) in effect on the date of publication of this document. Please review these documents carefully, as they describe your rights and restrictions with respect to this document. Code Components extracted from this document must include Simplified BSD License text as described in Section 4.e of the Trust Legal Provisions and are provided without warranty as described in the Simplified BSD License. The definitive version of an IETF Document is that published by, or under the auspices of, the IETF. Versions of IETF Documents that are published by third parties, including those that are translated into other languages, should not be considered to be definitive versions of IETF Documents. The definitive version of these Legal Provisions is that published by, or under the auspices of, the IETF. Versions of these Legal Provisions that are published by third parties, including those that are translated into other languages, should not be considered to be definitive versions of these Legal Provisions. For the avoidance of doubt, each Contributor to the IETF Standards Process licenses each Contribution that he or she makes as part of the IETF Standards Process to the IETF Trust pursuant to the provisions of RFC 5378. No language to the contrary, or terms, conditions or rights that differ from or are inconsistent with the rights and licenses granted under RFC 5378, shall have any effect and shall be null and void, whether published or posted by such Contributor, or included with or in such Contribution.



INTERNET-DRAFT  
Intended status: Proposed Standard

Donald Eastlake  
Mingui Zhang  
Huawei  
Ayan Banerjee  
Cisco  
March 9, 2018

Expires: September 8, 2018

Transparent Interconnection of Lots of Links (TRILL):  
Multi-Topology  
<draft-ietf-trill-multi-topology-06.txt>

#### Abstract

This document specifies extensions to the IETF TRILL (Transparent Interconnection of Lots of Links) protocol to support multi-topology routing of unicast and multi-destination traffic based on IS-IS (Intermediate System to Intermediate System) multi-topology specified in RFC 5120.

#### Status of This Memo

This Internet-Draft is submitted to IETF in full conformance with the provisions of BCP 78 and BCP 79.

Distribution of this document is unlimited. Comments should be sent to the TRILL working group mailing list.

Internet-Drafts are working documents of the Internet Engineering Task Force (IETF), its areas, and its working groups. Note that other groups may also distribute working documents as Internet-Drafts.

Internet-Drafts are draft documents valid for a maximum of six months and may be updated, replaced, or obsoleted by other documents at any time. It is inappropriate to use Internet-Drafts as reference material or to cite them other than as "work in progress."

The list of current Internet-Drafts can be accessed at <http://www.ietf.org/lid-abstracts.html>. The list of Internet-Draft Shadow Directories can be accessed at <http://www.ietf.org/shadow.html>.



## Table of Contents

1. Introduction.....	3
1.1 Terminology.....	4
2. Topologies.....	5
2.1 Special Topology Zero.....	5
2.2 Links and Multi-Topology.....	5
2.3 TRILL Switches and Multi-Topology.....	5
2.4 TRILL Data Packets and Multi-Topology.....	6
2.4.1 Explicit Topology Labeling Support.....	6
2.4.2 The Explicit Topology Label.....	7
2.4.3 TRILL Use of the MT Label.....	8
3. TRILL Multi-Topology Adjacency and Routing.....	10
3.1 Adjacency.....	10
3.2 TRILL Switch Nicknames.....	10
3.3 TRILL Unicast Routing.....	11
3.4 TRILL Multi-Destination Routing.....	11
3.4.1 Distribution Trees.....	11
3.4.2 Multi-Access Links.....	13
4. Mixed Links.....	14
5. Other Multi-Topology Considerations.....	15
5.1 Address Learning.....	15
5.1.1 Data Plane Learning.....	15
5.1.2 Multi-Topology ESADI.....	15
5.2 Legacy Stubs.....	15
5.3 RBridge Channel Messages.....	15
5.4 Implementations Considerations.....	16
6. Allocation Considerations.....	17
6.1 IEEE Registration Authority Considerations.....	17
6.2 IANA Considerations.....	17
7. Security Considerations.....	18
Normative References.....	19
Informative References.....	20
Acknowledgements.....	21
Appendix A: Differences from RFC 5120.....	21
Authors' Addresses.....	22

## 1. Introduction

This document specifies extensions to the IETF TRILL (Transparent Interconnection of Lots of Links) protocol [RFC6325] [RFC7177] [RFC7780] to support multi-topology routing for both unicast and multi-destination traffic based on IS-IS (Intermediate System to Intermediate System, [IS-IS]) multi-topology [RFC5120]. Implementation and use of multi-topology are optional and use requires configuration. It is anticipated that not all TRILL campuses will need or use multi-topology.

Multi-topology creates different topologies or subsets from a single physical TRILL campus topology. This is different from Data Labels (VLANs and Fine Grained Labels [RFC7172]). Data Labels specify communities of end stations and can be viewed as creating virtual topologies of end station connectivity. However, in a single topology TRILL campus, TRILL Data packets can use any part of the physical topology of TRILL switches and links between TRILL switches, regardless of the Data Label of that packet's payload. In a multi-topology TRILL campus, TRILL data packets in a topology are restricted to the TRILL switches and links that are in their topology but may still use any of the TRILL switches and links in their topology regardless of the Data Label of their payload.

The essence of multi-topology behavior is that a multi-topology router classifies packets as to the topology within which they should be routed and uses logically different routing tables for different topologies. If routers in the network do not agree on the topology classification of packets or links, persistent routing loops can occur. It is the responsibility of the network manager to consistently configure multi-topology to avoid such routing loops.

The multi-topology TRILL extensions can be used for a wide variety of purposes, such as maintaining separate routing domains for isolated multicast or IPv6 islands, routing a class of traffic so that it avoids certain TRILL switches that lack some characteristic needed by that traffic, or making a class of traffic avoid certain links due to security, reliability, or other concerns.

It is possible for a particular topology to not be fully connected, either intentionally or due to node or link failures or incorrect configuration. This results in two or more islands of that topology that cannot communicate. In such a case, end station connected in that topology to different islands will be unable to communicate with each other.

Multi-topology TRILL supports regions of topology-ignorant TRILL switches as part of a multi-topology campus; however, such regions can only ingress to, egress from, or transit TRILL Data packets in the special base topology zero.

## 1.1 Terminology

The terminology and acronyms of [RFC6325] are used in this document. Some of these are listed below for convenience along with some additional terms.

campus - The name for a TRILL network, like "bridged LAN" is a name for a bridged network. It does not have any academic implication.

DRB - Designated RBridge [RFC7177].

FGL - Fine-Grained Labeling or Fine-Grained Labeled or Fine-Grained Label [RFC7172]. By implication, an "FGL TRILL switch" does not support multi-topology (MT).

IS - Intermediate System [IS-IS].

LSP - [IS-IS] Link State PDU (Protocol Data Unit). For TRILL this includes L1-LSPs and E-L1FS-LSPs [RFC7780].

MT - Multi-Topology, this document and [RFC5120].

MT TRILL Switch - A TRILL switch supporting the multi-topology feature specified in this document. An MT TRILL switch MUST support FGL in the sense that it MUST be FGL safe [RFC7172].

RBridge - "Routing Bridge", an alternative name for a TRILL switch.

TRILL - Transparent Interconnection of Lots of Links or Tunneled Routing in the Link Layer [RFC6325].

TRILL Switch - A device implementing the TRILL protocol. TRILL switches are [IS-IS] Intermediate Systems (routers).

VL - VLAN Labeling or VLAN Labeled or VLAN Label [RFC7172]. By implication, a "VL RBridge" or "VL TRILL switch" does not support FGL or MT.

The key words "MUST", "MUST NOT", "REQUIRED", "SHALL", "SHALL NOT", "SHOULD", "SHOULD NOT", "RECOMMENDED", "MAY", and "OPTIONAL" in this document are to be interpreted as described in [RFC2119] [RFC8174] when, and only when, they appear in all capitals, as shown here.

## 2. Topologies

In TRILL multi-topology, a topology is a subset of the TRILL switches and of the links between TRILL switches in the TRILL campus. TRILL Data packets are constrained to the subset of switches and links corresponding to the packet's topology. TRILL multi-topology is based on [RFC5120] IS-IS multi-topology. See Appendix A for differences between TRILL multi-topology and [RFC5120].

The zero topology is special as described in Section 2.1. Sections 2.2, 2.3, and 2.4 discuss the topology of links, TRILL switches, and TRILL Data packets respectively.

### 2.1 Special Topology Zero

The zero topology is special as the default base topology. All TRILL switches and links are considered to be in and MUST support topology zero. Thus, for example, topology zero can be used for general TRILL switch access within a campus for management messages, BFD messages [RFC7175], RBridge Channel messages [RFC7178], and the like.

### 2.2 Links and Multi-Topology

Multi-topology TRILL switches advertise the topologies for which they are willing to send and receive TRILL Data packets on a port by listing those topologies in one or more MT TLVs [RFC5120] appearing in every TRILL Hello [RFC7177] they send out that port, except that they MUST handle topology zero, which it is optional to list.

A link is only usable for TRILL Data packets in non-zero topology T if

- (1) all TRILL switch ports on the link advertise topology T support in their Hellos and
- (2) if any TRILL switch port on the link requires explicit TRILL Data packet topology labeling (see Section 2.4) every other TRILL switch port on the link is capable of generating explicit packet topology labeling.

### 2.3 TRILL Switches and Multi-Topology

A TRILL switch advertises the topologies that it supports by listing them in one or more MT TLVs [RFC5120] in its LSP except that it MUST support topology zero which is optional to list. For robust and rapid flooding, MT TLV(s) SHOULD be advertised in core LSP fragment zero.

There is no "MT capability bit". A TRILL switch advertises that it is MT capable by advertising in its LSP support for any topology or topologies with the MT TLV, even if it just explicitly advertises support for topology zero.

## 2.4 TRILL Data Packets and Multi-Topology

The topology of a TRILL Data packet is commonly determined from either (1) some field or fields present in the packet itself or (2) the port on which the packet was received; however optional explicit topology labeling of TRILL Data packets is also proved. This can be included in the data labeling area of TRILL Data packets as specified below.

Examples of fields that might be used to determine topology are values or ranges of values of the payload VLAN or FGL [RFC7172], packet priority, IP version (IPv6 versus IPv4) or IP protocol, Ethertype, unicast versus multi-destination payload, IP Differentiated Services Code Point (DSCP) bits, or the like.

"Multi-topology" does not apply to TRILL IS-IS packets or to link level control frames. Those messages are link local and can be thought of as being above all topologies. "Multi-topology" only applies to TRILL Data packets.

### 2.4.1 Explicit Topology Labeling Support

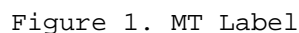
Support of the topology label is optional. Support could depend on port hardware and is indicated by a two-bit capability field in the Port TRILL Version sub-TLV [RFC7176] appearing in the Port Capabilities TLV in Hellos. If there is no Port TRILL Capabilities sub-TLV in a Hello, then it is assumed that explicit topology labeling is not supported on that port. See the table below for the meaning of values of the Explicit Topology capability field:

Value	Meaning
-----	-----
0	No support. Cannot send TRILL Data packets with an explicit topology label and will likely treat as erroneous and discard any TRILL Data packet received with a topology label. Such a port is assumed to have the ability and configuration to correctly classify TRILL data packets into all topologies for which it is advertising support in its Hellos, either by examining those packets or because they are arriving at that port.
1	Capable of inserting an explicit topology label in TRILL Data

2 and 3 Requires an explicit topology label in received TRILL Data packets except for topology zero. Any TRILL Data packets received without such a label is classified as being in topology zero. Also capable of inserting an explicit topology label in TRILL Data packets sent. (Values 2 and 3 are treated the same, which is the same as saying that if the 2 bit is on, the 1 bit is ignored.)

When a TRILL switch transmits a TRILL Data packet onto a link, if any other TRILL switch on that link requires explicit topology labeling, an explicit topology label **MUST** be included unless the TRILL data packet is in topology zero in which case an explicit topology label **MAY** be included. If a topology label is not so required but all other TRILL switches on that link support explicit topology labeling, then such a label **MAY** be included.

This section specifies the explicit topology label. Its use by TRILL is specified in Section 2.4.3. This label may be used by other technologies besides TRILL. The MT label is structured as follows:



MT Ethertype - The MT label Ethertype (see Section 6.1).

V - The version number of the MT label. This document specifies version zero.

R - A 2-bit reserved field that MUST be sent as zero and ignored on receipt.

MT-ID - The 12-bit topology using the topology number space of the MT TLV [RFC5120].

#### 2.4.3 TRILL Use of the MT Label

With the addition of the version zero MT label, the four standardized content varieties for the TRILL Data packet data labeling area (the area after the Inner.MacSA (or Flag Word if the Flag Word is present [RFC7780]) and before the payload) are as show below. TRILL Data packets received with any other data labeling are discarded. {PRI, D} is a 3-bit priority and a drop eligibility indicator bit [RFC7780].

All MT TRILL switches MUST support FGL, in the sense of being FGL safe [RFC7172], and thus MUST support all four data labeling area contents shown below. (This requirement is imposed, rather than having FGL support and MT support be independent, to reduce the number of variations in R Bridges and simplify testing.)

##### 1. C-VLAN [RFC6325]

```

          1 1 1 1 1 1 1 1 1 1 2 2 2 2 2 2 2 2 2 2 3 3
0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1
+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+
| C-VLAN = 0x8100          | PRI |D|  VLAN ID          |
+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+

```

##### 2. FGL [RFC7172]

```

          1 1 1 1 1 1 1 1 1 1 2 2 2 2 2 2 2 2 2 2 3 3
0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1
+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+
| FGL = 0x893B          | PRI |D|  FGL High Part      |
+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+
| FGL = 0x893B          | PRI |D|  FGL Low Part       |
+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+

```

## 3. MT C-VLAN [this document]

```

          1 1 1 1 1 1 1 1 1 1 2 2 2 2 2 2 2 2 2 2 3 3
0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1
+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+
| MT Ethertype = TBD          | 0 | R | MT-ID          |
+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+
| C-VLAN = 0x8100            | PRI |D| VLAN ID          |
+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+

```

## 4. MT FGL [this document] [RFC7172]

```

          1 1 1 1 1 1 1 1 1 1 2 2 2 2 2 2 2 2 2 2 3 3
0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1
+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+
| MT Ethertype = TBD          | 0 | R | MT-ID          |
+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+
| FGL = 0x893B                | PRI |D| FGL High Part    |
+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+
| FGL = 0x893B                | PRI |D| FGL Low Part     |
+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+

```

Inclusion or use of S-VLAN or further stacked tags are beyond the scope of this document but, as stated in [RFC6325], are obvious extensions.



### 3. TRILL Multi-Topology Adjacency and Routing

Routing calculations in IS-IS are based on adjacency. Section 3.1 specifies multi-topology TRILL adjacency. Section 3.2 describes the handling of nicknames. Sections 3.3 and 3.4 specify how unicast and multi-destination TRILL multi-topology routing differ from the TRILL base protocol routing.

#### 3.1 Adjacency

There is no change in the determination or announcement of adjacency for topology zero which is as specified in [RFC7177]. When a topology zero adjacency reaches the Report state as specified in [RFC7177], the adjacency is announced in core LSPs using the Extended Intermediate System Reachability TLV (#22). This will be compatible with any legacy topology-ignorant RBridges that might not support E-LFS FS-LSPs [RFC7780].

Adjacency is announced for non-zero topologies in LSPs using the MT Reachable Intermediate Systems TLV (#222) as specified in [RFC5120]. A TRILL switch reports adjacency for non-zero topology T if and only if that adjacency is in the Report state [RFC7177] and the two conditions listed in Section 2.2 are true, namely:

1. All the ports on the link are announcing support of topology T.
2. If any port announces that it requires explicit topology labeling (Explicit Topology capability field value 2 or 3), all other ports advertise that they are capable of producing such labeling (Explicit Topology capability field value of 1, 2, or 3).

#### 3.2 TRILL Switch Nicknames

TRILL switches are usually identified within the TRILL protocol (for example in the TRILL Header) by nicknames [RFC6325] [RFC7780]. Such nicknames can be viewed as simply 16-bit abbreviation for a TRILL switch's (or pseudo-node's) 7-byte IS-IS System ID. A TRILL switch or pseudo-node can have more than one nickname, each of which identifies it.

Nicknames are common across all topologies, just as IS-IS System IDs are. Nicknames are determined as specified in [RFC6325] and [RFC7780] using only the Nickname sub-TLVs appearing in Router Capabilities TLVs (#242) advertised by TRILL switches. In particular, the nickname allocation algorithm ignores Nickname sub-TLVs that appear in MT Router Capability TLVs (#144). (However, nickname sub-TLVs that

appear in MT Router Capability TLVs with a non-zero topology do affect the choice of distribution tree roots as described in Section 3.4.1.)

To minimize transient inconsistencies, all Nickname sub-TLVs advertised by a TRILL switch for a particular nickname, whether in Router Capability or MT Router Capability TLVs, SHOULD appear in the same LSP PDU. If that is not the case, then all LSP PDUs in which they do occur SHOULD be flooded as an atomic action.

### 3.3 TRILL Unicast Routing

TRILL Data packets being TRILL unicast (those with TRILL Header M bit = 0) are routed based on the egress nickname using logically separate forwarding tables per topology T where each such table has been calculated based on least cost routing within T, that is, only using links and nodes that support T. Thus, the next hop when forwarding TRILL Data packets is determined by a lookup logically based on {topology, egress nickname}.

### 3.4 TRILL Multi-Destination Routing

TRILL sends multi-destination data packets (those packets with TRILL Header M bit = 1) over a distribution tree. Trees are designated by nicknames that appear in the "egress nickname" field of multi-destination TRILL Data packet TRILL Headers. To constrain multi-destination packets to a topology T and still distribute them properly requires the use of a distribution tree constrained to T. Handling such TRILL Data packets and distribution trees in TRILL MT is as described in the subsections below.

#### 3.4.1 Distribution Trees

General provisions for distribution trees and how those trees are determined are as specified in [RFC6325], [RFC7172], and [RFC7780]. The distribution trees for topology zero are determined as specified in those references and are the same as they would be with topology-ignorant TRILL switches.

The TRILL distribution tree construction and packet handling for some non-zero topology T are determined as specified in [RFC6325], [RFC7172], and [RFC7780] with the following changes:

- o As specified in [RFC5120], only links usable with topology T TRILL Data packets are considered when building a distribution tree for topology T. As a result, such trees are automatically limited to and separately span every internally connected island of topology T. In other words, if non-zero topology T consists of disjoint islands, each distribution tree construction for topology T is local to one such island.
- o Only the Nickname sub-TLV, Trees sub-TLV, Tree Identifiers sub-TLV, and Trees Used sub-TLV occurring in an MT Router Capabilities TLV (#144) specifying topology T are used in determining the tree root(s), if any, for a connected area of non-zero topology T.
  - + There may be non-zero topologies with no multi-destination traffic or, as described in [RFC5120], even topologies with no traffic at all. For example, if only known destination unicast IPv6 TRILL Data packets were in topology T and all multi-destination IPv6 TRILL Data packets were in some other topology, there would be no need for a distribution tree for topology T. For this reason, a Number of Trees to Compute of zero in the Trees sub-TLV for the TRILL switch holding the highest priority to be a tree root for a non-zero topology T is honored and causes no distribution trees to be calculated for non-zero topology T. This is different from the base topology zero where, as specified in [RFC6325], a zero Number of Trees to Compute causes one tree to be computed.
- o Nicknames are allocated as described in Section 3.2. If a TRILL switch advertising that it provides topology T service holds nickname N, the priority of N to be a tree root is given by the tree root priority field of the Nickname sub-TLV that has N in its nickname field and occurs in a topology T MT Router Capabilities TLV advertised by that TRILL switch. If no such Nickname sub-TLV can be found, the priority of N to be a tree root is the default for an FGL TRILL switch as specified in [RFC7172].
  - + There could be multiple topology T Nickname sub-TLVs for N being advertised for a particular RBridge or pseudo-node, due to transient conditions or errors. In that case, any advertised in a core LSP PDU are preferred to those advertised in an E-L1FS FS-LSP PDU. Within those categories, the one in the lowest numbered fragment is used and if there are multiple in that fragment, the one with the smallest offset from the beginning of the PDU is used.
- o Tree pruning for topology T uses only the Interested VLANs sub-TLVs and Interested Labels sub-TLVs [RFC7176] advertised in MT

Router Capabilities TLVs for topology T.

An MT TRILL switch MUST have logically separate routing tables per topology for the forwarding of multi-destination traffic.

#### 3.4.2 Multi-Access Links

Multi-destination TRILL Data packets are forwarded on broadcast (multi-access) links in such a way as to be received by all other TRILL switch ports on the link. For example, on Ethernet links they are sent with a multicast Outer.MacDA [RFC6325]. Care must be taken that a TRILL Data packet in a non-zero topology is only forwarded by an MT TRILL switch.

For this reason, a non-zero topology TRILL Data packet MUST NOT be forwarded onto a link unless the link meets the requirements specified in Section 2.2 for use in that topology even if there are one or more MT TRILL switch ports on the link.

#### 4. Mixed Links

There might be any combination of MT, FGL, or even VL TRILL switches [RFC7172] on a link. DRB (Designated RBridge) election and Forwarder appointment on the link work as previously specified in [RFC8139] and [RFC7177]. It is up to the network manager to configure and manage the TRILL switches on a link so that the desired switch is DRB and the desired switch is the Appointed Forwarder for the appropriate VLANs.

Frames ingressed by MT TRILL switches can potentially be in any topology recognized by the switch and permitted on the ingress port. Frames ingressed by VL or FGL TRILL switches can only be in the base zero topology. Because FGL and VL TRILL switches do not understand topologies, all occurrences of the following sub-TLVs MUST occur only in MT Port Capability TLVs with a zero MT-ID. Any occurrence of these sub-TLVs in an MT Port Capability TLV with a nonzero MT-ID is ignored.

- Special VLANs and Flags Sub-TLV
- Enabled-VLANs Sub-TLV
- Appointed Forwarders Sub-TLV
- VLANs Appointed Sub-TLV

Native frames cannot be explicitly labeled (see Section 2.4) as to their topology.

## 5. Other Multi-Topology Considerations

### 5.1 Address Learning

The learning of end station MAC addresses is per topology as well as per label (VLAN or FGL). The same MAC address can occur within a TRILL campus for different end stations that differ only in topology without confusion.

#### 5.1.1 Data Plane Learning

End station MAC addresses learned from ingressing native frames or egressing TRILL Data packets are, for MT TRILL switches, qualified by topology. That is, either the topology into which that TRILL switch classified the ingressed native frame or the topology that the egressed TRILL Data frame was in.

#### 5.1.2 Multi-Topology ESADI

In an MT TRILL switch, ESADI [RFC7357] operates per label (VLAN or FGL) per topology. Since ESADI messages appear, to transit TRILL switches, like normal multi-destination TRILL Data packets, ESADI link state databases and ESADI protocol operation are per topology as well as per label and local to each area of multi-destination TRILL data connectivity for that topology.

### 5.2 Legacy Stubs

Areas of topology ignorant TRILL switches can be connected to and become part of an MT TRILL campus but will only be able to ingress to, transit, or egress from topology zero TRILL Data packets.

### 5.3 RBridge Channel Messages

RBridge Channel messages [RFC7178], such as BFD over TRILL [RFC7175] appear, to transit TRILL switches, like normal multi-destination TRILL Data packets. Thus, they have a topology and, if that topology is non-zero, are constrained by topology like other TRILL Data packets. Generally, when sent for network management purposes, they are sent in topology zero to avoid such constraint.

#### 5.4 Implementations Considerations

MT is an optional TRILL switch capability.

Experience with the actual deployment of Layer 3 IS-IS MT [RFC5120] indicates that a single router handling more than eight topologies is rare. There may be many more than eight distinct topologies in a routed area, such as a TRILL campus, but in that case many of these topologies will be handled by disjoint sets of routers and/or links.

Based on this deployment experience, a TRILL switch capable of handling 8 or more topologies can be considered a full implementation while a TRILL switch capable of handling 4 topologies can be considered a minimal implementation but still useful under some circumstances.

## 6. Allocation Considerations

IEEE Registration Authority and IANA considerations are given below.

### 6.1 IEEE Registration Authority Considerations

The IEEE Registration Authority will be requested to allocate a new Ethertype for the MT label (see Section 2.4).

### 6.2 IANA Considerations

IANA is requested to assign a field of two adjacent bits TBD from bits 14 through 31 of the Capabilities bits of the Port TRILL Version Sub-TLV for the Explicit Topology capability field and update the "PORT-TRILL-VER Capability Bits" registry as follows [shown with the suggested bits 14 and 15]:

Bit	Description	Reference
-----	-----	-----
14-15	Topology labeling support	[this document]



## 7. Security Considerations

Multiple topologies are sometimes used for the isolation or security of traffic. For example, if some links were more likely than others to be subject to adversarial observation it might be desirable to classify certain sensitive traffic in a topology that excluded those links.

Delivery of data originating in one topology outside of that topology is generally a security policy violation to be avoided at all reasonable costs. Using IS-IS security [RFC5310] on all IS-IS PDUs and link security appropriate to the link technology on all links involved, particularly those between RBridges, supports the avoidance of such violations.

For general TRILL security considerations, see [RFC6325].

## Normative References

- [IS-IS] - ISO/IEC 10589:2002, Second Edition, "Intermediate System to Intermediate System Intra-Domain Routeing Exchange Protocol for use in Conjunction with the Protocol for Providing the Connectionless-mode Network Service (ISO 8473)", 2002.
- [RFC2119] - Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, DOI 10.17487/RFC2119, March 1997, <<http://www.rfc-editor.org/info/rfc2119>>.
- [RFC5120] - Przygienda, T., Shen, N., and N. Sheth, "M-ISIS: Multi Topology (MT) Routing in Intermediate System to Intermediate Systems (IS-ISs)", RFC 5120, February 2008.
- [RFC5310] - Bhatia, M., Manral, V., Li, T., Atkinson, R., White, R., and M. Fanto, "IS-IS Generic Cryptographic Authentication", RFC 5310, DOI 10.17487/RFC5310, February 2009, <<http://www.rfc-editor.org/info/rfc5310>>.
- [RFC6325] - Perlman, R., Eastlake 3rd, D., Dutt, D., Gai, S., and A. Ghanwani, "Routing Bridges (RBriges): Base Protocol Specification", RFC 6325, July 2011.
- [RFC7172] - Eastlake 3rd, D., Zhang, M., Agarwal, P., Perlman, R., and D. Dutt, "Transparent Interconnection of Lots of Links (TRILL): Fine-Grained Labeling", RFC 7172, May 2014.
- [RFC7176] - Eastlake 3rd, D., Senevirathne, T., Ghanwani, A., Dutt, D., and A. Banerjee, "Transparent Interconnection of Lots of Links (TRILL) Use of IS-IS", RFC 7176, May 2014.
- [RFC7177] - Eastlake 3rd, D., Perlman, R., Ghanwani, A., Yang, H., and V. Manral, "Transparent Interconnection of Lots of Links (TRILL): Adjacency", RFC 7177, May 2014.
- [RFC7178] - Eastlake 3rd, D., Manral, V., Li, Y., Aldrin, S., and D. Ward, "Transparent Interconnection of Lots of Links (TRILL): RBridge Channel Support", RFC 7178, May 2014.
- [RFC7357] - Hhai, H., Hu, F., Perlman, R., Eastlake 3rd, D., and O. Stokes, "Transparent Interconnection of Lots of Links (TRILL): End Station Address Distribution Information (ESADI) Protocol", RFC 7357, DOI 10.17487/RFC7357, September 2014, <<http://www.rfc-editor.org/info/rfc7357>>.
- [RFC7780] - Eastlake 3rd, D., Zhang, M., Perlman, R., Banerjee, A., Ghanwani, A., and S. Gupta, "Transparent Interconnection of Lots of Links (TRILL): Clarifications, Corrections, and Updates", RFC 7780, DOI 10.17487/RFC7780, February 2016,

<<http://www.rfc-editor.org/info/rfc7780>>.

[RFC8174] - Leiba, B., "Ambiguity of Uppercase vs Lowercase in RFC 2119 Key Words", BCP 14, RFC 8174, DOI 10.17487/RFC8174, May 2017, <<http://www.rfc-editor.org/info/rfc8174>>

#### Informative References

[RFC8139] - Eastlake 3rd, D., Li, Y., Umair, M., Banerjee, A., and F. Hu, "Transparent Interconnection of Lots of Links (TRILL): Appointed Forwarders", RFC 8139, DOI 10.17487/RFC8139, June 2017, <<https://www.rfc-editor.org/info/rfc8139>>.

[RFC7175] - Manral, V., Eastlake 3rd, D., Ward, D., and A. Banerjee, "Transparent Interconnection of Lots of Links (TRILL): Bidirectional Forwarding Detection (BFD) Support", RFC 7175, May 2014.

## Acknowledgements

The comments and contributions of the following are gratefully acknowledged:

Vishwas Manral and Martin Vigoureux

## Appendix A: Differences from RFC 5120

TRILL multi-topology, as specified in this document, differs from RFC 5120 as follows:

1. [RFC5120] provides for unicast multi-topology. This document extends that to cover multi-destination TRILL data distribution (see Section 3.4).
2. [RFC5120] assumes the topology of data packets is always determined implicitly, that is, based on the port over which the packets are received and/or pre-existing fields within the packet. This document supports such implicit determination but extends this by providing for optional explicit topology labeling of data packets (see Section 2.4).
3. [RFC5120] makes support of the default topology zero optional for MT routers and links. For simplicity and ease in network management, this document requires all TRILL switches and links between TRILL switches to support topology zero (see Section 2.1).

Authors' Addresses

Donald Eastlake 3rd  
Huawei Technologies  
155 Beaver Street  
Milford, MA 01757 USA

Phone: +1-508-333-2270  
Email: d3e3e3@gmail.com

Mingui Zhang  
Huawei Technologies Co., Ltd  
HuaWei Building, No.3 Xinxu Rd., Shang-Di  
Information Industry Base, Hai-Dian District,  
Beijing, 100085 P.R. China

Email: zhangmingui@huawei.com

Ayan Banerjee  
Cisco  
170 W. Tasman Drive  
San Jose, CA 95134

Email: ayabaner@cisco.com

## Copyright, Disclaimer, and Additional IPR Provisions

Copyright (c) 2018 IETF Trust and the persons identified as the document authors. All rights reserved.

This document is subject to BCP 78 and the IETF Trust's Legal Provisions Relating to IETF Documents (<http://trustee.ietf.org/license-info>) in effect on the date of publication of this document. Please review these documents carefully, as they describe your rights and restrictions with respect to this document. Code Components extracted from this document must include Simplified BSD License text as described in Section 4.e of the Trust Legal Provisions and are provided without warranty as described in the Simplified BSD License. The definitive version of an IETF Document is that published by, or under the auspices of, the IETF. Versions of IETF Documents that are published by third parties, including those that are translated into other languages, should not be considered to be definitive versions of IETF Documents. The definitive version of these Legal Provisions is that published by, or under the auspices of, the IETF. Versions of these Legal Provisions that are published by third parties, including those that are translated into other languages, should not be considered to be definitive versions of these Legal Provisions. For the avoidance of doubt, each Contributor to the IETF Standards Process licenses each Contribution that he or she makes as part of the IETF Standards Process to the IETF Trust pursuant to the provisions of RFC 5378. No language to the contrary, or terms, conditions or rights that differ from or are inconsistent with the rights and licenses granted under RFC 5378, shall have any effect and shall be null and void, whether published or posted by such Contributor, or included with or in such Contribution.



INTERNET-DRAFT  
Intended Status: Proposed Standard

M. Zhang  
Huawei  
D. Eastlake  
Futurewei  
R. Perlman  
EMC  
M. Cullen  
Painless Security  
H. Zhai  
JIT

Expires: May 11, 2022

November 12, 2021

Transparent Interconnection of Lots of Links (TRILL)  
Single Area Border RBridge Nickname for Multilevel  
draft-ietf-trill-multilevel-single-nickname-17.txt

#### Abstract

A major issue in multilevel TRILL is how to manage RBridge nicknames. In this document, area border RBridges use a single nickname in both Level 1 and Level 2. RBridges in Level 2 must obtain unique nicknames but RBridges in different Level 1 areas may have the same nicknames.

#### Status of This Memo

This Internet-Draft is submitted in full conformance with the provisions of BCP 78 and BCP 79.

Distribution of this document is unlimited. Comments should be sent to the authors or the TRILL Working Group mailing list [trill@ietf.org](mailto:trill@ietf.org).

Internet-Drafts are working documents of the Internet Engineering Task Force (IETF), its areas, and its working groups. Note that other groups may also distribute working documents as Internet-Drafts.

Internet-Drafts are draft documents valid for a maximum of six months and may be updated, replaced, or obsoleted by other documents at any time. It is inappropriate to use Internet-Drafts as reference material or to cite them other than as "work in progress."

The list of current Internet-Drafts can be accessed at <https://www.ietf.org/lid-abstracts.html>. The list of Internet-Draft Shadow Directories can be accessed at <https://www.ietf.org/shadow.html>.



## Table of Contents

1. Introduction.....	3
2. Acronyms and Terminology.....	4
3. Nickname Handling on Border RBridges.....	5
3.1. Actions on Unicast Packets.....	5
3.2. Actions on Multi-Destination Packets.....	6
4. Per-flow Load Balancing.....	9
4.1. L2 to L1 Ingress Nickname Replacement.....	9
4.2. L1 to L2 Egress Nickname Replacement.....	9
5. Protocol Extensions for Discovery.....	10
5.1. Discovery of Border RBridges in L1.....	10
5.2. Discovery of Border RBridge Sets in L2.....	10
6. One Border RBridge Connects Multiple Areas.....	12
7. E-L1FS/E-L2FS Backwards Compatibility.....	13
8. Manageability Considerations.....	13
9. Security Considerations.....	15
10. IANA Considerations.....	15
11. References.....	16
11.1. Normative References.....	16
11.2. Informative References.....	17
Appendix A. Level Transition Clarification.....	18
Authors' Addresses.....	19

## 1. Introduction

TRILL (Transparent Interconnection of Lots of Links [RFC6325] [RFC7780]) multilevel techniques are designed to improve TRILL scalability issues.

[RFC8243] (Alternatives for Multilevel Transparent Interconnection of Lots of Links (TRILL)) is an educational document to explain multilevel TRILL and list possible concerns. It does not specify a protocol. As described in [RFC8243], there have been two proposed approaches. One approach, which is referred to as the "unique nickname" approach, gives unique nicknames to all the TRILL switches in the multilevel campus, either by having the Level 1/Level 2 border TRILL switches advertise which nicknames are not available for assignment in the area, or by partitioning the 16-bit nickname into an "area" field and a "nickname inside the area" field. [RFC8397] is the standards track document specifying a "unique nickname" flavor of TRILL multilevel. The other approach, which is referred to in [RFC8243] as the "aggregated nickname" approach, involves assigning nicknames to the areas, and allowing nicknames to be reused inside different areas, by having the border TRILL switches rewrite the nickname fields when entering or leaving an area. [RFC8243] makes the case that, while unique nickname multilevel solutions are simpler, aggregated nickname solutions scale better.

The approach specified in this standards track document is somewhat similar to the "aggregated nickname" approach in [RFC8243] but with a very important difference. In this document, the nickname of an area border RBridge is used in both Level 1 (L1) and Level 2 (L2). No additional nicknames are assigned to represent L1 areas as such. Instead, multiple border RBridges are allowed and each L1 area is denoted by the set of all nicknames of those border RBridges of the area. For this approach, nicknames in the L2 area MUST be unique but nicknames inside an L1 area can be reused in other L1 areas that also use this approach. The use of the approach specified in this document in one L1 area does not prohibit the use of other approaches in other L1 areas in the same TRILL campus, for example the use of the unique nickname approach specified in [RFC8397]. The TRILL packet format is unchanged by this document, but data plane processing is changed at Border RBridges and efficient high volume data flow at Border RBridges might require forwarding hardware change.

## 2. Acronyms and Terminology

Data Label: VLAN or FGL Fine-Grained Label (FGL).

DBRB: Designated Border RBridge.

IS-IS: Intermediate System to Intermediate System [IS-IS].

Level: Similar to IS-IS, TRILL has Level 1 for intra-area and Level 2 for inter-area. Routing information is exchanged between Level 1 RBridges within the same Level 1 area, and Level 2 RBridges can only form relationships and exchange information with other Level 2 RBridges.

The key words "MUST", "MUST NOT", "REQUIRED", "SHALL", "SHALL NOT", "SHOULD", "SHOULD NOT", "RECOMMENDED", "NOT RECOMMENDED", "MAY", and "OPTIONAL" in this document are to be interpreted as described in BCP 14 [RFC2119] [RFC8174] when, and only when, they appear in all capitals, as shown here.

Familiarity with [RFC6325] is assumed in this document.

### 3. Nickname Handling on Border RBridges

This section provides an illustrative example and description of the border learning border RBridge nicknames.

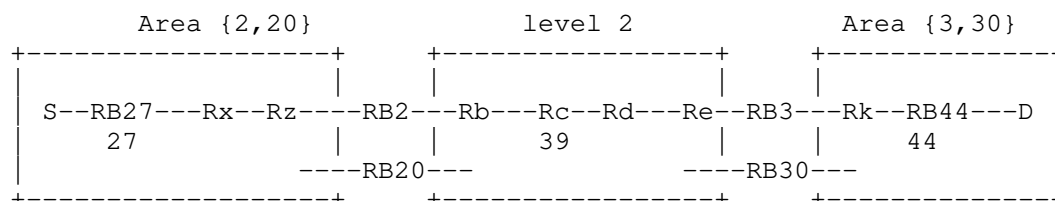


Figure 1: An Example Topology for TRILL Multilevel

In Figure 1, RB2, RB20, RB3 and RB30 are area border TRILL switches (RBridges). Their nicknames are 2, 20, 3 and 30 respectively and are used as TRILL switch identifiers in their areas [RFC6325]. Area border RBridges use the set of border nicknames to denote the L1 area that they are attached to. For example, RB2 and RB20 use nicknames {2,20} to denote the L1 area on the left.

A source S is attached to RB27 and a destination D is attached to RB44. RB27 has a nickname, say 27, and RB44 has a nickname, say 44 (and in fact, they could even have the same nickname, since the TRILL switch nickname will not be visible outside these Level 1 areas).

#### 3.1. Actions on Unicast Packets

Let's say that S transmits a frame to destination D and let's say that D's location has been learned by the relevant TRILL switches already. These relevant switches have learned the following:

- 1) RB27 has learned that D is connected to nickname 3.
- 2) RB3 has learned that D is attached to nickname 44.

The following sequence of events will occur:

- S transmits an Ethernet frame with source MAC = S and destination MAC = D.
- RB27 encapsulates with a TRILL header with ingress RBridge = 27, and egress RBridge = 3 producing a TRILL Data packet.
- RB2 and RB20 have announced in the Level 1 IS-IS area designated {2,20}, that they are attached to the nicknames of all the border RBridges in the Level 2 area including RB3 and RB30. Therefore, IS-IS routes the packet to RB2 (or RB20, if RB20 on the least-cost

route from RB27 to RB3).

- RB2, when transitioning the packet from Level 1 to Level 2, replaces the ingress TRILL switch nickname with its own nickname, replacing 27 with 2. Within Level 2, the ingress RBridge field in the TRILL header will therefore be 2, and the egress RBridge field will be 3. (The egress nickname MAY be replaced with any area nickname selected from {3,30} such as 30. See Section 4 for the detail of the selection method. Here, suppose the egress nickname remains 3.) Also, RB2 learns that S is attached to nickname 27 in area {2,20} to accommodate return traffic. RB2 SHOULD synchronize with RB20 using ESADI protocol [RFC7357] that MAC = S is attached to nickname 27.
- The packet is forwarded through Level 2, to RB3, which has advertised, in Level 2, its L2 nickname as 3.
- RB3, when forwarding into area {3,30}, replaces the egress nickname in the TRILL header with RB44's nickname (44) based on looking up D. (The ingress nickname MAY be replaced with any area nickname selected from {2,20}. See Section 4 for the detail of the selection method. Here, suppose the ingress nickname remains 2.) So, within the destination area, the ingress nickname will be 2 and the egress nickname will be 44.
- RB44, when decapsulating, learns that S is attached to nickname 2, which is one of the area nicknames of the ingress.

### 3.2. Actions on Multi-Destination Packets

Distribution trees for flooding of multi-destination packets are calculated separately within each L1 area and in L2. When a multi-destination packet arrives at the border, it needs to be transitioned either from L1 to L2, or from L2 to L1. All border RBridges are eligible for Level transition. However, for each multi-destination packet, only one of them acts as the Designated Border RBridge (DBRB) to do the transition while other non-DBRBs MUST drop the received copies. By default, the border RBridge with the smallest nickname, considered as an unsigned integer, is elected DBRB. All border RBridges of an area MUST agree on the mechanism used to determine the DBRB locally. The use of an alternative is possible, but out of the scope of this document; one such mechanism is used in Section 4 for load balancing.

As per [RFC6325], multi-destination packets can be classified into three types: unicast packet with unknown destination MAC address (unknown-unicast packet), multicast packet and broadcast packet. Now suppose that D's location has not been learned by RB27 or the frame

received by RB27 is recognized as broadcast or multicast. What will happen within a Level 1 area (as it would in TRILL today) is that RB27 will forward the packet as multi-destination, setting its M bit to 1 and choosing an L1 tree, flooding the packet on the distribution tree, subject to possible pruning.

When the copies of the multi-destination packet arrive at area border RBridges, non-DBRBs MUST drop the packet while the DBRB, say RB2, needs to do the Level transition for the multi-destination packet. For an unknown-unicast packet, if the DBRB has learnt the destination MAC address, it SHOULD convert the packet to unicast and set its M bit to 0. Otherwise, the multi-destination packet will continue to be flooded as multicast packet on the distribution tree. The DBRB chooses the new distribution tree by replacing the egress nickname with the new tree root RBridge nickname from the area the packet is entering. The following sequence of events will occur:

- RB2, when transitioning the packet from Level 1 to Level 2, replaces the ingress TRILL switch nickname with its own nickname, replacing 27 with 2. RB2 also MUST replace the egress RBridge nickname with an L2 tree root RBridge nickname (say 39). In order to accommodate return traffic, RB2 records that S is attached to nickname 27 and SHOULD use the ESADI protocol [RFC7357] to synchronize this attachment information with other border RBridges (say RB20) in the area.
- RB20, will receive the packet flooded on the L2 tree by RB2. It is important that RB20 does not transition this packet back to L1 as it does for a multicast packet normally received from another remote L1 area. RB20 should examine the ingress nickname of this packet. If this nickname is found to be a border RBridge nickname of the area {2,20}, RB2 must not forward the packet into this area.
- The multidestination packet is flooded on the Level 2 tree to reach all border routers for all L1 areas including both RB3 and RB30. Suppose RB3 is the selected DBRB. The non-DBRB RB30 will drop the packet.
- RB3, when forwarding into area {3,30}, replaces the egress nickname in the TRILL header with the root RBridge nickname of a distribution tree of L1 area {3,30} say 30. (Here, the ingress nickname MAY be replaced with a different area nickname selected from {2,20}, the set of border RBridges to the ingress area, as specified in Section 4.) Now suppose that RB27 has learned the location of D (attached to nickname 3), but RB3 does not know where D is because this information has fallen out of cache or RB3 has re-started or some other reason. In that case, RB3 must turn the packet into a multi-destination packet and floods it on a distribution tree in the L1 area {3,30}.

- RB30, will receive the packet flooded on the L1 tree by RB3. It is important that RB30 does not transition this packet back to L2. RB30 should also examine the ingress nickname of this packet. If this nickname is found to be an L2 border RBridge nickname, RB30 must not transition the packet back to L2.
- The multicast listener RB44, when decapsulating the received packet, learns that S is attached to nickname 2, which is one of the area nicknames of the ingress.

See also Appendix A.

#### 4. Per-flow Load Balancing

Area border R Bridges perform ingress/egress nickname replacement when they transition TRILL data packets between Level 1 and Level 2. The egress nickname will again be replaced when the packet transitions from Level 2 to Level 1. This nickname replacement enables the per-flow load balance which is specified in the following subsections. The mechanism specified in Section 4.1 or that in 4.2 or both is necessary in general to load balance traffic across L2 paths.

##### 4.1. L2 to L1 Ingress Nickname Replacement

When a TRILL data packet from other L1 areas arrives at an area border R Bridge, this R Bridge MAY select one area nickname of the ingress area to replace the ingress nickname of the packet so that the returning TRILL data packet can be forwarded to this selected nickname to help load balance return unicast traffic over multiple paths. The selection is simply based on a pseudorandom algorithm as discussed in Section 5.3 of [RFC7357]. With the random ingress nickname replacement, the border R Bridge actually achieves a per-flow load balance for returning traffic.

All area border R Bridges for an L1 area MUST agree on the same pseudorandom algorithm. The source MAC address, ingress area nicknames, egress area nicknames and the Data Label of the received TRILL data packet are candidate factors of the input of this pseudorandom algorithm. Note that the value of the destination MAC address SHOULD be excluded from the input of this pseudorandom algorithm, otherwise the egress R Bridge could see one source MAC address flip-flopping among multiple ingress R Bridges.

##### 4.2. L1 to L2 Egress Nickname Replacement

When a unicast TRILL data packet originated from an L1 area arrives at an area border R Bridge of that L1 area, that R Bridge MAY select one area nickname of the egress area to replace the egress nickname of the packet. By default, it SHOULD choose the egress area border R Bridge with the least cost route to reach or, if there are multiple equal cost egress area border R Bridges, use the pseudorandom algorithm as defined in Section 5.3 of [RFC7357] to select one. The use of that algorithm MAY be extended to selection among some stable set of egress area border R Bridges that include non-least-cost alternatives if it is desired to obtain more load spreading at the cost of sometimes using a non-least-cost Level 2 route to forward the TRILL data packet to the egress area.



## 5. Protocol Extensions for Discovery

The following topology change scenarios will trigger the discovery processes as defined in Sections 5.1 and 5.2:

- A new node comes up or recovers from a previous failure.
- A node goes down.
- A link or node fails and causes partition of an L1/L2 area.
- A link or node whose failure have caused partitioning of an L1/L2 area is repaired.

### 5.1. Discovery of Border RBridges in L1

The following Level 1 Border RBridge APPsub-TLV will be included in an E-L1FS FS-LSP fragment zero [RFC7780] as an APPsub-TLV of the TRILL GENINFO-TLV. Through listening for this APPsub-TLV, an area border RBridge discovers all other area border RBridges in this area.

```

+---+---+---+---+---+---+---+---+---+---+
| Type = L1-BORDER-RBRIDGE          | (2 bytes)
+---+---+---+---+---+---+---+---+---+---+
| Length                            | (2 bytes)
+---+---+---+---+---+---+---+---+---+---+
| Sender Nickname                    | (2 bytes)
+---+---+---+---+---+---+---+---+---+---+

```

- o Type: Level 1 Border RBridge (TRILL APPsub-TLV type tbd1)
- o Length: 2
- o Sender Nickname: The nickname the originating IS will use as the L1 Border RBridge nickname. This field is useful because the originating IS might own multiple nicknames.

### 5.2. Discovery of Border RBridge Sets in L2

The following APPsub-TLV will be included in an E-L2FS FS-LSP fragment zero [RFC7780] as an APPsub-TLV of the TRILL GENINFO-TLV. Through listening to this APPsub-TLV in L2, an area border RBridge discovers all groups of L1 border RBridges and each such group identifies an area.

```

+---+---+---+---+---+---+---+---+---+
| Type = L1-BORDER-RB-GROUP          | (2 bytes)
+---+---+---+---+---+---+---+---+---+
| Length                              | (2 bytes)
+---+---+---+---+---+---+---+---+---+
| L1 Border RBridge Nickname 1        | (2 bytes)
+---+---+---+---+---+---+---+---+---+
| ...                                 |
+---+---+---+---+---+---+---+---+---+
| L1 Border RBridge Nickname k        | (2 bytes)
+---+---+---+---+---+---+---+---+---+

```

- o Type: Level 1 Border RBridge Group (TRILL APPsub-TLV type tbd2)
- o Length:  $2 * k$ . If length is not a multiple of 2, the APPsub-TLV is corrupt and MUST be ignored.
- o L1 Border RBridge Nickname: The nickname that an area border RBridge uses as the L1 Border RBridge nickname. The L1-BORDER-RB-GROUP TLV generated by an area border RBridge MUST include all L1 Border RBridge nicknames of the area. It's RECOMMENDED that these k nicknames are ordered in ascending order according to the 2-octet nickname considered as an unsigned integer.

When an L1 area is partitioned [RFC8243], border RBridges will re-discover each other in both L1 and L2 through exchanging LSPs. In L2, the set of border RBridge nicknames for this splitting area will change. Border RBridges that detect such a change MUST flush the reachability information associated to any RBridge nickname from this changing set.

## 6. One Border RBridge Connects Multiple Areas

It's possible that one border RBridge (say RB1) connects multiple L1 areas. RB1 SHOULD use a single area nickname for itself for all these areas to minimize nickname consumption and the number of nicknames being advertised in L2; however, such a border RBridge might have to hold multiple nicknames, for example it might be the root of multiple L1 or multiple L2 distribution trees.

Nicknames used within one of these L1 areas can be reused within other areas. It's important that packets destined to those duplicated nicknames are sent to the right area. Since these areas are connected to form a layer 2 network, duplicated {MAC, Data Label} across these areas SHOULD NOT occur (see Section 4.2.6 of [RFC6325] for tie breaking rules). Now suppose a TRILL data packet arrives at the area border nickname of RB1. For a unicast packet, RB1 can look up the {MAC, Data Label} entry in its MAC table to identify the right destination area (i.e., the outgoing interface) and the egress RBridge's nickname. For a multicast packet for each attached L1 area: either RB1 is not the DBRB and RB1 will not transition the packet or RB1 is the DBRB. If RB1 is the DBRB, RB1 follows the following rules:

- if this packet originated from an area out of the connected areas, RB1 replicates this packet and floods it on the proper Level 1 trees of all the areas in which it acts as the DBRB.
- if the packet originated from one of the connected areas, RB1 replicates the packet it receives from the Level 1 tree and floods it on other proper Level 1 trees of all the areas in which it acts as the DBRB except the originating area (i.e., the area connected to the incoming interface). RB1 might also receive the replication of the packet from the Level 2 tree. This replication MUST be dropped by RB1. It recognizes such packets by their ingress nickname being the nickname of one of the border RBridges of an L1 area for which the receiving border RBridge is DBRB.

## 7. E-L1FS/E-L2FS Backwards Compatibility

All Level 2 RBridges MUST support E-L2FS [RFC7356] [RFC7780]. The Extended TLVs defined in Section 5 are to be used in Extended Level 1/2 Flooding Scope (E-L1FS/E-L2FS) PDUs. Area border RBridges MUST support both E-L1FS and E-L2FS. RBridges that do not support both E-L1FS or E-L2FS cannot serve as area border RBridges but they can appear in an L1 area acting as non-area-border RBridges.

## 8. Manageability Considerations

If an L1 Border RBridge Nickname is configured at an RBridge and that RBridge has both L1 and L2 adjacencies, the multilevel feature as specified in this document is turned on for that RBridge and it normally uses an L2 nickname in both L1 and L2 although, as provided below, such an RBridge may have to fall back to multilevel unique nickname behavior [RFC8397] in which case it uses this L1 nickname. In contrast, unique nickname multilevel as specified in [RFC8397] is enabled by the presence of L1 and L2 adjacencies without an L1 Border RBridge Nickname being configured. RBridges supporting only unique nickname multilevel do not support the configuration of an L2 Border RBridge Nickname. RBridges supporting only the single level TRILL base protocol specified in [RFC6325] do not support L2 adjacencies.

RBridges that support and are configured to use single nickname multilevel as specified in this document MUST support unique nickname multilevel ([RFC8397]). If there are multiple border RBridges between an L1 area and L2 and one or more of them only support or are only configured for unique nickname multilevel ([RFC8397]), any of these border RBridges that are configured to use single nickname multilevel MUST fall back to behaving as a unique nickname border RBridge for that L1 area. Because overlapping sets of RBridges may be the border RBridges for different L1 areas, an RBridge supporting single nickname MUST be able to simultaneously support single nickname for some of its L1 areas and unique nickname for others. For example, RB1 and RB2 might be border RBridges for L1 area A1 using single nickname while RB2 and RB3 are border RBridges for area A2. If RB3 only supports unique nicknames then RB2 must fall back to unique nickname for area A2 but continue to support single nickname for area A1. Operators SHOULD be notified when this fall back occurs. The presence of border RBridges using unique nickname multilevel can be detected because they advertise in L1 the blocks of nicknames available within that L1 area.

In both the unique nickname approach specified in [RFC8397] and the single nickname aggregated approach specified in this document, an RBridge that has L1 and L2 adjacencies uses the same nickname in L1 and L2. If an RBridge is configured with an L1 Border RBridge

Nickname for any a Level 1 area, it uses this nickname across the Level 2 area. This L1 Border RBridge Nickname cannot be used in any other Level 1 area except other Level 1 areas for which the same RBridge is a border RBridge with this L1 Border RBridge Nickname configured.

In addition to the manageability considerations specified above, the manageability specifications in [RFC6325] still apply.

Border RBridges replace ingress and/or egress nickname when a TRILL data packet traverses TRILL L2 area. A TRILL OAM message will be forwarded through the multilevel single nickname TRILL campus using a MAC address belonging to the destination RBridge [RFC7455].

## 9. Security Considerations

For general TRILL Security Considerations, see [RFC6325].

The newly defined TRILL APPsub-TLVs in Section 5 are transported in IS-IS PDUs whose authenticity can be enforced using regular IS-IS security mechanism [IS-IS] [RFC5310]. Malicious devices may also fake the APPsub-TLVs to attract TRILL data packets, interfere with multilevel TRILL operation, induce excessive state in TRILL switches (or in any bridges that may be part of the TRILL campus), etc. For this reason, RBridges SHOULD be configured to use the IS-IS Authentication TLV (10) in their IS-IS PDUs so that IS-IS security [RFC5310] can be used to authenticate those PDUs and discard them if they are forged.

Using a variation of aggregated nicknames, and the resulting possible duplication of nicknames between areas, increases the possibility of a TRILL Data packet being delivered to the wrong egress RBridge if areas are unexpectedly merged as compared with a scheme where all nicknames in the TRILL campus are, except as a transient condition, unique such as the scheme in [RFC8397]. However, in many cases the data would be discarded at that egress RBridge because it would not match a known end station data label/MAC address.

## 10. IANA Considerations

IANA is requested to allocate two new types under the TRILL GENINFO TLV [RFC7357] from the range allocated by standards action for the TRILL APPsub-TLVs defined in Section 5. The following entries are added to the "TRILL APPsub-TLV Types under IS-IS TLV 251 Application Identifier 1" Registry on the TRILL Parameters IANA web page.

Type	Name	Reference
-----	----	-----
tbd1[256]	L1-BORDER-RBRIDGE	[This document]
tbd2[257]	L1-BORDER-RB-GROUP	[This document]

## 11. References

### 11.1. Normative References

- [RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, DOI 10.17487/RFC2119, March 1997, <<https://www.rfc-editor.org/info/rfc2119>>.
- [RFC6325] Perlman, R., Eastlake 3rd, D., Dutt, D., Gai, S., and A. Ghanwani, "Routing Bridges (RBridges): Base Protocol Specification", RFC 6325, DOI 10.17487/RFC6325, July 2011, <<https://www.rfc-editor.org/info/rfc6325>>.
- [RFC7356] Ginsberg, L., Previdi, S., and Y. Yang, "IS-IS Flooding Scope Link State PDUs (LSPs)", RFC 7356, DOI 10.17487/RFC7356, September 2014, <<https://www.rfc-editor.org/info/rfc7356>>.
- [RFC7357] Zhai, H., Hu, F., Perlman, R., Eastlake 3rd, D., and O. Stokes, "Transparent Interconnection of Lots of Links (TRILL): End Station Address Distribution Information (ESADI) Protocol", RFC 7357, DOI 10.17487/RFC7357, September 2014, <<https://www.rfc-editor.org/info/rfc7357>>.
- [RFC7455] Senevirathne, T., Finn, N., Salam, S., Kumar, D., Eastlake 3rd, D., Aldrin, S., and Y. Li, "Transparent Interconnection of Lots of Links (TRILL): Fault Management", RFC 7455, DOI 10.17487/RFC7455, March 2015, <<https://www.rfc-editor.org/info/rfc7455>>.
- [RFC7780] Eastlake 3rd, D., Zhang, M., Perlman, R., Banerjee, A., Ghanwani, A., and S. Gupta, "Transparent Interconnection of Lots of Links (TRILL): Clarifications, Corrections, and Updates", RFC 7780, DOI 10.17487/RFC7780, February 2016, <<https://www.rfc-editor.org/info/rfc7780>>.
- [RFC8174] Leiba, B., "Ambiguity of Uppercase vs Lowercase in RFC 2119 Key Words", BCP 14, RFC 8174, DOI 10.17487/RFC8174, May 2017, <<https://www.rfc-editor.org/info/rfc8174>>.
- [RFC8397] Zhang, M., Eastlake 3rd, D., Perlman, R., Zhai, H., and D. Liu, "Transparent Interconnection of Lots of Links (TRILL) Multilevel Using Unique Nicknames", RFC 8397, DOI 10.17487/RFC8397, May 2018, <<https://www.rfc-editor.org/info/rfc8397>>.

## 11.2. Informative References

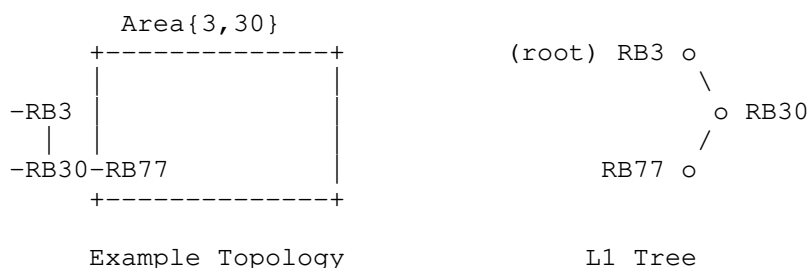
- [IS-IS] International Organization for Standardization, ISO/IEC 10589:2002, "Information technology -- Telecommunications and information exchange between systems -- Intermediate System to Intermediate System intra-domain routeing information exchange protocol for use in conjunction with the protocol for providing the connectionless-mode network service", ISO 8473, Second Edition, November 2002.
- [RFC5310] Bhatia, M., Manral, V., Li, T., Atkinson, R., White, R., and M. Fanto, "IS-IS Generic Cryptographic Authentication", RFC 5310, DOI 10.17487/RFC5310, February 2009, <<https://www.rfc-editor.org/info/rfc5310>>.
- [RFC8243] Perlman, R., Eastlake 3rd, D., Zhang, M., Ghanwani, A., and H. Zhai, "Alternatives for Multilevel Transparent Interconnection of Lots of Links (TRILL)", RFC 8243, DOI 10.17487/RFC8243, September 2017, <<https://www.rfc-editor.org/info/rfc8243>>.



## Appendix A. Level Transition Clarification

It's possible that an L1 RBridge is only reachable from a non-DBRB border RBridge. If this non-DBRB RBridge refrains from Level transition, the question is, how can a multicast packet reach this L1 RBridge? The answer is, it will be reached after the DBRB performs the Level transition and floods the packet using an L1 distribution tree.

Take the following figure as an example. RB77 is reachable from the border RBridge RB30 while RB3 is the DBRB. RB3 transitions the multicast packet into L1 and floods the packet on the distribution tree rooted from RB3. This packet is finally flooded to RB77 via RB30.



In the above example, the multicast packet is forwarded along a non-optimal path. A possible improvement is to have RB3 configured not to belong to this area. In this way, RB30 will surely act as the DBRB to do the Level transition.

## Authors' Addresses

Mingui Zhang  
Huawei Technologies  
No. 156 Beiqing Rd. Haidian District  
Beijing 100095  
China

Email: zhangmingui@huawei.com

Donald E. Eastlake, 3rd  
Futurewei Technologies  
2386 Panoramic Circle  
Apopka, FL 32703  
United States

Phone: +1-508-333-2270  
Email: d3e3e3@gmail.com

Radia Perlman  
EMC  
2010 256th Avenue NE, #200  
Bellevue, WA 98007  
United States

Email: radia@alum.mit.edu

Margaret Cullen  
Painless Security  
356 Abbott Street  
North Andover, MA 01845  
United States

Phone: +1-781-405-7464  
Email: margaret@painless-security.com  
URI: <https://www.painless-security.com>

Hongjun Zhai  
Jinling Institute of Technology  
99 Hongjing Avenue, Jiangning District  
Nanjing, Jiangsu 211169  
China

Email: honjun.zhai@tom.com

## Copyright, Disclaimer, and Additional IPR Provisions

Copyright (c) 2021 IETF Trust and the persons identified as the document authors. All rights reserved.

This document is subject to BCP 78 and the IETF Trust's Legal Provisions Relating to IETF Documents (<http://trustee.ietf.org/license-info>) in effect on the date of publication of this document. Please review these documents carefully, as they describe your rights and restrictions with respect to this document. Code Components extracted from this document must include Simplified BSD License text as described in Section 4.e of the Trust Legal Provisions and are provided without warranty as described in the Simplified BSD License.



INTERNET-DRAFT  
Intended Status: Proposed Standard  
Updates: 7177, 7178

Margaret Cullen  
Painless Security  
Donald Eastlake  
Mingui Zhang  
Huawei  
Dacheng Zhang  
Alibaba  
October 19, 2015

Expires: April 18, 2016

Transparent Interconnection of Lots of Links (TRILL) over IP  
<draft-ietf-trill-over-ip-05.txt>

#### Abstract

The Transparent Interconnection of Lots of Links (TRILL) protocol supports both point-to-point and multi-access links and is designed so that a variety of link protocols can be used between TRILL switch ports. This document standardizes methods for encapsulating TRILL in IP (v4 or v6) so as to use IP as a TRILL link protocol in a unified TRILL campus. It updates RFC 7177 and updates RFC 7178.

#### Status of This Document

This Internet-Draft is submitted to IETF in full conformance with the provisions of BCP 78 and BCP 79.

Distribution of this document is unlimited. Comments should be sent to the author or the DNSEXT mailing list <dnsext@ietf.org>.

Internet-Drafts are working documents of the Internet Engineering Task Force (IETF), its areas, and its working groups. Note that other groups may also distribute working documents as Internet-Drafts.

Internet-Drafts are draft documents valid for a maximum of six months and may be updated, replaced, or obsoleted by other documents at any time. It is inappropriate to use Internet-Drafts as reference material or to cite them other than as "work in progress."

The list of current Internet-Drafts can be accessed at <http://www.ietf.org/lid-abstracts.html>. The list of Internet-Draft Shadow Directories can be accessed at <http://www.ietf.org/shadow.html>.

## Table of Contents

1. Introduction.....	4
2. Terminology.....	5
3. Use Cases for TRILL over IP.....	6
3.1 Remote Office Scenario.....	6
3.2 IP Backbone Scenario.....	6
3.3 Important Properties of the Scenarios.....	6
3.3.1 Security Requirements.....	7
3.3.2 Multicast Handling.....	7
3.3.3 Neighbor Discovery.....	8
4. TRILL Packet Formats.....	9
4.1 General Packet Formats.....	9
4.2 General TRILL Over IP Packet Formats.....	10
4.2.1 Without Security.....	10
4.2.2 With Security.....	10
4.3 QoS Considerations.....	11
4.4 Broadcast Links and Multicast Packets.....	12
4.5 TRILL Over IP IS-IS SubNetwork Point of Attachment.....	13
5. TRILL over IP Encapsulation Formats.....	14
5.1 Encapsulation Considerations.....	14
5.2 Encapsulation Agreement.....	15
5.3 Broadcast Link Encapsulation Considerations.....	16
5.4 Native Encapsulation.....	16
5.5 VXLAN Encapsulation.....	17
5.6 Other Encapulsations.....	18
6. Handling Multicast.....	19
7. Use of IPsec and IKEv2.....	20
7.1 Keying.....	20
7.1.1 Pairwise Keying.....	20
7.1.2 Group Keying.....	21
7.2 Mandatory-to-Implement Algorithms.....	21
8. Transport Considerations.....	22
8.1 Congestion Considerations.....	22
8.2 Recursive Ingress.....	23
8.3 Fat Flows.....	24
8.4 MTU Considerations.....	25
8.5 Middlebox Considerations.....	25
9. TRILL over IP Port Configuration.....	27
9.1 Per IP Port Configuration.....	27
9.2 Additional per IP Address Configuration.....	27
9.2.1 Native Multicast Configuration.....	27

## Table of Contents (continued)

9.2.2 Serial Unicast Configuration.....	28
9.2.3 Encapsulation Specific Configuration.....	28
9.2.3.1 VXLAN Configuration.....	28
9.2.3.2 Other Encapsulation Configuration.....	29
9.2.4 Security Configuration.....	29
10. Security Considerations.....	30
10.1 IPsec.....	30
10.2 IS-IS Security.....	31
11. IANA Considerations.....	32
11.1 Port Assignments.....	32
11.2 Multicast Address Assignments.....	32
11.3 Encapsulation Method Support Indication.....	32
Normative References.....	34
Informative References.....	36
Acknowledgements.....	38

## 1. Introduction

TRILL switches (RBridges) are devices that implement the IETF TRILL protocol [RFC6325] [RFC7177] [rfc7180bis]. TRILL provides transparent forwarding of frames within an arbitrary network topology, using least cost paths for unicast traffic. It supports VLANs and Fine Grained Labels [RFC7172] as well as multipathing of unicast and multi-destination traffic. It uses IS-IS [RFC7176] link state routing and encapsulation with a hop count.

RBridges ports can communicate with each other over various protocols, such as Ethernet [RFC6325], pseudowires [RFC7173], or PPP [RFC6361].

This document defines a method for RBridge ports to communicate over IP (v4 or v6). TRILL over IP allows Internet-connected RBridges to form a single TRILL campus, or multiple TRILL over IP networks within a campus to be connected as a single TRILL campus via a TRILL over IP backbone.

TRILL over IP connects RBridge ports using IPv4 or IPv6 as a transport in such a way that the ports appear to TRILL to be connected by a single multi-access link. If more than two RBridge ports are connected via a single TRILL over IP link, any pair of them can communicate.

To support the scenarios where RBridges are connected via IP paths (such as over the public Internet) that are not under the same administrative control as the TRILL campus and/or not physically secure, this document specifies the use of IPsec [RFC4301] Encapsulating Security Protocol (ESP) [RFC4303] to secure such paths.

To dynamically select a mutually supported TRILL over IP encapsulation, normally one with good fast path hardware support, a method is provided for agreement between adjacent TRILL switch ports as to what encapsulation to use. This document updates [RFC7177] and [RFC7178] as described in Section 5 by making adjacency between TRILL over IP ports dependent on having a method of encapsulation in common and by redefining an interval of RBridge Channel protocol numbers to indicate encapsulation method support for TRILL over IP.



## 2. Terminology

The key words "MUST", "MUST NOT", "REQUIRED", "SHALL", "SHALL NOT", "SHOULD", "SHOULD NOT", "RECOMMENDED", "MAY", and "OPTIONAL" in this document are to be interpreted as described in RFC 2119 [RFC2119].

The following terms and acronyms have the meaning indicated:

DRB - Designated RBridge. The RBridge (TRILL switch) elected to be in charge of certain aspects of a TRILL link that is not configured as a point-to-point link [RFC6325] [RFC7177].

ENCAP Hdr - Encapsulation headers in use between the IP Header and the TRILL Header. See Section 5.

ESP - IPsec Encapsulating Security Protocol [RFC4303].

FGL - Fine Grained Label [RFC7172].

Hdr - Used herein as an abbreviation for "Header".

HKDF - Hash based Key Derivation Function [RFC5869].

MTU - Maximum Transmission Unit.

RBridge - Routing Bridge. An alternative term for a TRILL switch.

TRILL - Transparent Internconnection of Lots of Links or Tunneled Routing in the Link Layer. The protocol specified in [RFC6325], [RFC7177], [rfc7180bis], and related RFCs.

TRILL switch - A device implementing the TRILL protocol.

VNI - Virtual Network Identifier. In VXLAN [RFC7348], the VXLAN Network Identifier.

### 3. Use Cases for TRILL over IP

This section introduces two application scenarios (a remote office scenario and an IP backbone scenario) which cover typical situations where network administrators may choose to use TRILL over an IP network to connect TRILL switches.

#### 3.1 Remote Office Scenario

In the Remote Office Scenario, a remote TRILL network is connected to a TRILL campus across a multihop IP network, such as the public Internet. The TRILL network in the remote office becomes a part of TRILL campus, and nodes in the remote office can be attached to the same VLANs or Fine Grained Labels [RFC7172] as local campus nodes. In many cases, a remote office may be attached to the TRILL campus by a single pair of RBridges, one on the campus end, and the other in the remote office. In this use case, the TRILL over IP link will often cross logical and physical IP networks that do not support TRILL, and are not under the same administrative control as the TRILL campus.

#### 3.2 IP Backbone Scenario

In the IP Backbone Scenario, TRILL over IP is used to connect a number of TRILL networks to form a single TRILL campus. For example, a TRILL over IP backbone could be used to connect multiple TRILL networks on different floors of a large building, or to connect TRILL networks in separate buildings of a multi-building site. In this use case, there may often be several TRILL switches on a single TRILL over IP link, and the IP link(s) used by TRILL over IP are typically under the same administrative control as the rest of the TRILL campus.

#### 3.3 Important Properties of the Scenarios

There are a number of differences between the above two application scenarios, some of which drive features of this specification. These differences are especially pertinent to the security requirements of the solution, how multicast data frames are handled, and how the TRILL switch ports discover each other.

### 3.3.1 Security Requirements

In the IP Backbone Scenario, TRILL over IP is used between a number of RBridge ports, on a network link that is in the same administrative control as the remainder of the TRILL campus. While it is desirable in this scenario to prevent the association of unauthorized RBridges, this can be accomplished using existing IS-IS security mechanisms. There may be no need to protect the data traffic, beyond any protections that are already in place on the local network.

In the Remote Office Scenario, TRILL over IP may run over a network that is not under the same administrative control as the TRILL network. Nodes on the network may think that they are sending traffic locally, while that traffic is actually being sent, in an IP tunnel, over the public Internet. It is necessary in this scenario to protect the integrity and confidentiality of user traffic, as well as ensuring that no unauthorized RBridges can gain access to the RBridge campus. The issues of protecting integrity and confidentiality of user traffic are addressed by using IPsec for both TRILL IS-IS and TRILL Data packets between RBridges in this scenario.

### 3.3.2 Multicast Handling

In the IP Backbone scenario, native IP multicast may be supported on the TRILL over IP link. If so, it can be used to send TRILL IS-IS and multicast data packets, as discussed later in this document. Alternatively, multi-destination packets can be transmitted serially by IP unicast to the intended recipients.

In the Remote Office Scenario there will often be only one pair of RBridges connecting a given site and, even when multiple RBridges are used to connect a Remote Office to the TRILL campus, the intervening network may not provide reliable (or any) multicast connectivity. Issues such as complex key management also make it difficult to provide strong data integrity and confidentiality protections for multicast traffic. For all of these reasons, the connections between local and remote RBridges will commonly be treated like point-to-point links, and all TRILL IS-IS control messages and multicast data packets that are transmitted between the Remote Office and the TRILL campus will be serially transmitted by IP unicast, as discussed later in this document.

### 3.3.3 Neighbor Discovery

In the IP Backbone Scenario, TRILL switches that use TRILL over IP can use the normal TRILL IS-IS Hello mechanisms to discover the existence of other TRILL switches on the link [RFC7177], and to establish authenticated communication with them.

In the Remote Office Scenario, an IPsec session will need to be established before TRILL IS-IS traffic can be exchanged, as discussed below. In this case, one end will need to be configured to establish a IPSEC session with the other. This will typically be accomplished by configuring the TRILL switch or a border device at a Remote Office to initiate an IPsec session and subsequent TRILL exchanges with a TRILL over IP-enabled RBridge attached to the TRILL campus.

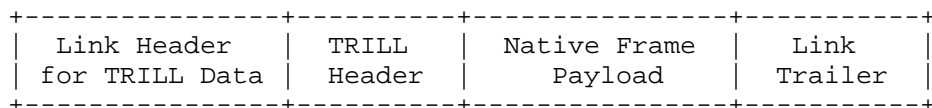
#### 4. TRILL Packet Formats

To support the TRILL protocol [RFC6325], two types of TRILL packets are transmitted between TRILL switches: TRILL Data packets and TRILL IS-IS packets.

Section 4.1 describes general TRILL packet formats for data and IS-IS independent of link technology. Section 4.2 specifies general TRILL over IP packet formats including IPsec ESP encapsulation. Section 4.3 provides QoS Considerations. Section 4.4 discusses broadcast links and multicast packets. And Section 4.5 provides TRILL IS-IS Hello SubNetwork Point of Attachment (SNPA) considerations for TRILL over IP.

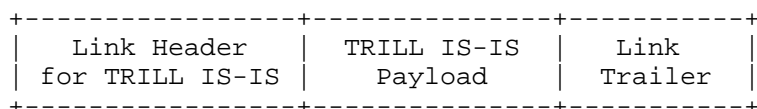
##### 4.1 General Packet Formats

The on-the-wire form of a TRILL Data packet in transit between two neighboring TRILL switch ports is as shown below:



The encapsulated Native Frame Payload is similar to an Ethernet frame with a VLAN tag or Fine Grained Label [RFC7172] but with no trailing Frame Check Sequence (FCS).

TRILL IS-IS packets are formatted on-the-wire as follows:



The Link Header and Link Trailer in these formats depend on the specific link technology. The Link Header contains one or more fields that distinguish TRILL Data from TRILL IS-IS. For example, over Ethernet, the Link Header for TRILL Data ends with the TRILL Ethertype while the Link Header for TRILL IS-IS ends with the L2-IS-IS Ethertype; on the other hand, over PPP, there are no Ethernets in the Link Header but PPP protocol code points are included that distinguish TRILL Data from TRILL IS-IS.

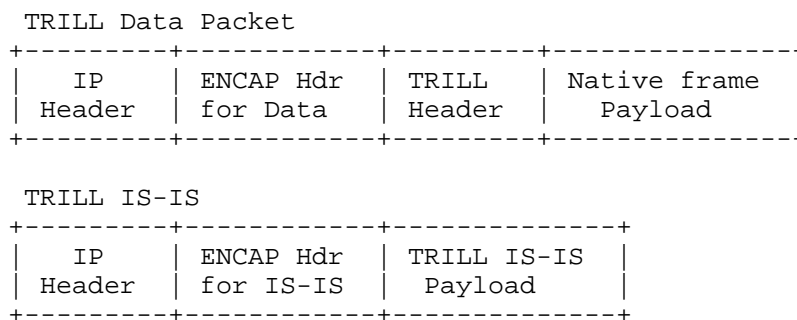
## 4.2 General TRILL Over IP Packet Formats

In TRILL over IP, we will use an IP (v4 or v6) header as the link header. (On the wire, the IP header will normally be preceded by the lower layer header of a protocol that is carrying IP; however, this does not concern us at the level of this document.)

There are multiple IP based encapsulations usable for TRILL over IP that differ in exactly what appears after the IP header and before the TRILL Header or the TRILL IS-IS Payload. These encapsulations are further detailed in Section 5. In the general specification below, those encapsulation fields will be represented as "ENCAP Hdr". See Section 5 for details.

### 4.2.1 Without Security

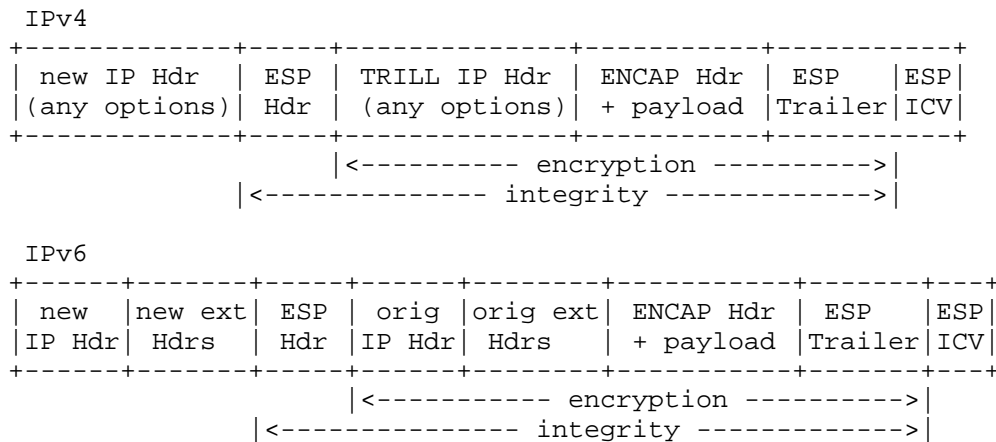
When TRILL over IP link security is not being used, a TRILL over IP packet on the wire looks like the following:



As discussed above and further specified in Section 5, the ENCAP Hdr indicates whether the packet is TRILL Data or IS-IS.

### 4.2.2 With Security

TRILL over IP link security uses IPsec Encapsulating Security Protocol (ESP) in tunnel mode [RFC4303]. Since TRILL over IP always starts with an IP Header (on the wire this appears right after any lower layer header that might be required), the modifications for IPsec are independent of the TRILL over IP ENCAP Hdr that occurs after that IP Header. The resulting packet formats are as follows for IPv4 and IPv6:

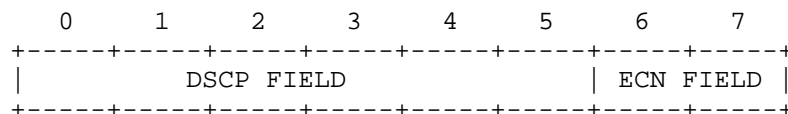


As shown above, IP Header options are considered part of the IPv4 Header but are extensions ("ext") of the IPv6 Header. For further information on the IPsec ESP Hdr, Trailer, and ICV, see [RFC4303] and Section 7. "ENCAP Hdr + payload" is the encapsulation header (Section 5) and TRILL data or IS-is payload, that is, the material after the IP Header in the diagram in Section 4.2.1.

This architecture permits the ESP tunnel end point to be separated from the TRILL over IP RBridge port (see, for example, Section 1.1.3 of [RFC7296]).

#### 4.3 QoS Considerations

In IP, QoS handling is indicated by the Differential Services Code Point (DSCP [RFC2474] [RFC3168]) in the TRILL Header. The former Type of Service (TOS) octet in the IPv4 Header and the Traffic Class octet in the IPv6 Header has been divided as shown in the following diagram adapted from [RFC3168]. (TRILL support of ECN is beyond the scope of this document.)



DSCP: Differentiated Services Codepoint  
ECN: Explicit Congestion Notification

Within a TRILL switch, priority is indicated by configuration for TRILL IS-IS packets and for TRILL Data packets by a three bit (0 through 7) priority field and a Drop Eligibility Indicator bit (see Sections 8.2 and 7 of [rfc7180bis]). (Typically TRILL IS-IS is

configured to use the highest priority or, alternatively, the highest two priorities depending on the IS-IS PDU.) The priority affects queuing behavior at TRILL switch ports and may be encoded into the link header, particularly if there could be priority sensitive devices within the link. For example, if the link is a bridged LAN, it is commonly encoded into an Outer.VLAN tag's priority and DEI fields.

TRILL over IP implementations MUST support setting the DSCP value in the outer IP Header of TRILL packets they send by mapping the TRILL priority and DEI to the DSCP. They MAY support, for a TRILL Data packet where the native frame payload is an IP packet, copying the DSCP in this inner IP packet to the outer IP Header.

The default TRILL priority and DEI to DSCP mapping, which may be configured per TRILL over IP port, is as follows. Note that the DEI value does not affect the default mapping and, to provide a potentially lower priority service than the default 0, priority 1 is considered lower priority than 0. So the priority sequence from lower to higher priority is 1, 0, 2, 3, 4, 5, 6, 7.

TRILL Priority	DEI	DSCP Field (Binary/decimal)
0	0/1	001000 / 8
1	0/1	000000 / 0
2	0/1	010000 / 16
3	0/1	011000 / 24
4	0/1	100000 / 32
5	0/1	101000 / 40
6	0/1	110000 / 48
7	0/1	111000 / 56

#### 4.4 Broadcast Links and Multicast Packets

TRILL supports broadcast links. These are links to which more than two TRILL switch ports can be attached and where a packet can be broadcast or multicast from a port to all or a subset of the other ports on the link as well as unicast to a specific single other port on the link.

As specified in [RFC6325], TRILL Data packets being forwarded between TRILL switches can be unicast on a link to a specific TRILL switch port or multicast on a link to all TRILL switch ports. TRILL IS-IS packets are always multicast to all other TRILL switches on the link except for IS-IS MTU PDUs, which may be unicast [RFC7177]. This distinction is not significant if the link is inherently point-to-point, such as a PPP link; however, on a broadcast link there will be a packet outer link address that is unicast or multicast as



appropriate. For example, over Ethernet links, the Ethernet multicast addresses All-RBridges and All-IS-IS-RBridges are used for multicasting TRILL Data and TRILL IS-IS respectively. For details on TRILL over IP handling of multicast, see Section 6.

#### 4.5 TRILL Over IP IS-IS SubNetwork Point of Attachment

IS-IS routers, such as TRILL switches, establish adjacency through the exchange of Hello PDUs on a link [IS-IS] [RFC7177]. The Hellos transmitted out a port indicate what neighbor ports that port can see on the link by listing what IS-IS refers to as the neighbor port's SubNetwork Point of Attachment (SNPA). (For an Ethernet link, which may be a bridged LAN, the SNPA is the port MAC address.)

In TRILL Hello PDUs on a TRILL over IP link, the IP addresses of the IP ports connected to that link are their actual SNPA (SubNetwork Point of Attachment [IS-IS]) addresses and, for IPv6, the 16-byte IPv6 address is used as the SNPA; however, for easy in re-using code designed for the common case of 48-bit SNPAs, in TRILL over IPv4 a 48-bit synthetic SNPA that looks like a unicast MAC address is constructed for use in the SNPA field of TRILL Neighbor TLVs [RFC7176] [RFC7177] in such Hellos. This synthetic SNPA is derived from the port IPv4 address is as follows:

```

          1 1 1 1 1 1
0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5
+-----+
| 0xFE          | 0x00          |
+-----+
| IPv4 upper half|
+-----+
| IPv4 lower half|
+-----+
```

This synthetic SNPA (MAC) address has the local (0x02) bit on in the first byte and so cannot conflict with any globally unique 48-bit Ethernet MAC. However, when TRILL operates on an IP link, TRILL sees only IP stations, not MAC stations, even if the TRILL over IP Link is being carried over Ethernet. Therefore conflict on the link in TRILL IS-IS between a real MAC address and the synthetic SNPA (MAC) address as above would be impossible in any case.

## 5. TRILL over IP Encapsulation Formats

There are a variety of TRILL over IP encapsulation formats possible. By default TRILL over IP adopts a hybrid encapsulation approach.

There is one format, called "native encapsulation" that MUST be implemented. Although native encapsulation does not typically have good fast path support, as a lowest common denominator it can be used by low bandwidth control traffic to determine a preferred encapsulation with better performance. In particular, by default, all TRILL IS-IS Hellos are sent using native encapsulation and those Hellos are used to determine the encapsulation used for all TRILL Data packets and all other TRILL IS-IS PDUs (with the possible exception of IS-IS MTU-probe and MTU-ack PDUs).

Alternatively, the network operator can pre-configure a TRILL over IP port to use a particular encapsulation chosen for their particular network needs and port capabilities. That encapsulation is then used for all TRILL Data and IS-IS packets on ports so configured.

Section 5.1 discusses general consideration for the TRILL over IP encapsulation format. Section 5.2 discusses encapsulation agreement. Section 5.3 discusses broadcast link encapsulation considerations. The subsequent subsections discuss particular encapsulations.

### 5.1 Encapsulation Considerations

In all cases, there must be a method specified to distinguish TRILL Data packets and TRILL IS-IS packets, or that encapsulation is not useful for TRILL. In addition, the following criteria can be helpful in choosing between different encapsulations:

- a) Fast path support - For many applications, it is highly desirable to be able to encapsulate/decapsulate TRILL over IP at line speed so a format where existing or anticipated fast path hardware can do that is best. This is commonly a dominant consideration.
- b) Ease of multi-pathing - The IP path between TRILL over IP ports may include equal cost multipath routes internal to the IP link so a method of encapsulation that provides variable fields available for existing or anticipated fast path hardware multi-pathing is better.
- c) Robust fragmentation and re-assembly - MTU of the IP link may require fragmentation in which case an encapsulation with robust fragmentation and re-assembly is important. There are known problems with IPv4 fragmentation and re-assembly [RFC6864] which generally do not apply to IPv6. Some encapsulations can fix these

problems but the two encapsulations specified in this document do not. Therefore, if fragmentation is anticipated with the encapsulations specified in this document, the use of IPv6 is RECOMMENDED.

- d) Checksum strength - Depending on the particular circumstances of the TRILL over IP link, a checksum provided by the encapsulation may be an important factor. Use of IPsec can also provide a strong integrity check.

## 5.2 Encapsulation Agreement

TRILL Hellos sent out a TRILL over IP port indicate the encapsulations that port is willing to support through a mechanism initially specified in [RFC7178] and [RFC7176] that is hereby extended. Specifically, RBridge Channel Protocol numbers 0xFD0 through 0xFF7 are redefined to be link technology dependent flags that, for TRILL over IP, indicate support for different encapsulations, allowing for up to 40 encapsulations to be specified. Support for an encapsulation is indicated in the Hello PDU in the same way that support for an RBridge Channel was indicated. (See also section 11.3.) "Support" indicates willingness to use that encapsulation for TRILL Data and TRILL IS-IS packets (although TRILL IS-IS Hellos are still sent in native encapsulation by default).

If, in a TRILL Hello on a TRILL over IP link, support is not indicated for any encapsulation, then the port from which it was sent is assumed to support only native encapsulation (see Section 5.4).

An adjacency is formed between two TRILL over IP ports if the intersection of the sets of encapsulation methods they support is not null. If that intersection is null, then no adjacency is formed. In particular, for a TRILL over IP link, the adjacency state machine MUST NOT advance to the Report state unless the ports share an encapsulation [RFC7177]. If no encapsulation is shared, the adjacency state machine remains in the state from which it would otherwise have transitioned to the Report state.

If any TRILL over IP packet, other than an IS-IS Hello or MTU PDU in native encapsulation, is received in an encapsulation for which support is not being indicated, it MUST be discarded (see Section 5.3).

If there are two or more encapsulations in common between two adjacent ports for unicast or the set of adjacent ports for multicast, a transmitter is free to choose whichever of the encapsulations it wishes to use. Thus transmissions between adjacent ports P1 and P2 could use different encapsulations depending on which

port is transmitting and which is receiving.

It is expected to be the normal case in a well configured network that all the TRILL over IP ports connected to an IP link (i.e., an IP network) that are intended to communicate with each other will support the same encapsulation(s).

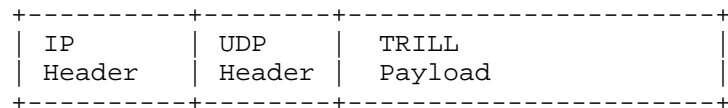
### 5.3 Broadcast Link Encapsulation Considerations

To properly handle TRILL protocol packets on a TRILL over IP link in the general case, either native IP multicast mode is used on that link or multicast must be simulated using serial IP unicast, as discussed in Section 6. (Of course, if the IP link happens to actually be point-to-point no special provision is needed for handling multicast addressed packets.)

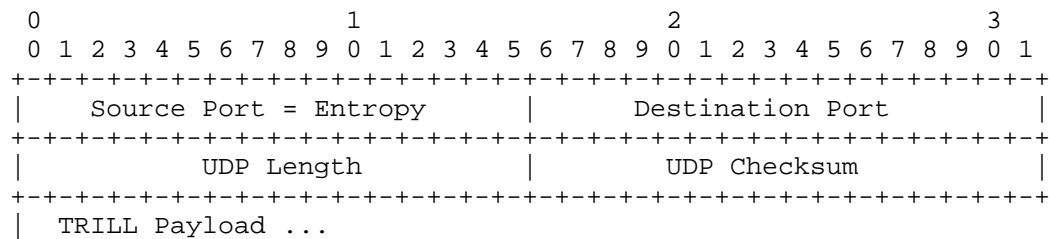
It is possible for the Hellos from a TRILL over IP port P1 to establish adjacency with multiple other TRILL over IP ports (P2, P3, ...) on broadcast link. In a well configured network one would expect all of the IP ports involved to support the same encapsulation(s); but, if P1 supports multiple encapsulations, it is possible that P2 and P3, for example, do not have an encapsulation in common that is supported by P1. IS-IS can handle such non-transitive adjacencies which are reported as specified in [RFC7177]. If serial IP unicast is being used by P1, it can use different encapsulations for different transmissions. If native IP multicast is being used by P1, it will have to send one transmission per encapsulation method by which it has an adjacency on the link. (It is for this reason that a TRILL over IP port MUST discard any packet received with the wrong encapsulation. Otherwise, packets would be duplicated.)

### 5.4 Native Encapsulation

The mandatory to implement "native encapsulation" format of a TRILL over IP packet, when used without security, is TRILL over UDP as shown below.



Where the UDP Header is as follows:



Source Port - see Section 8.3

Destination Port - indicates TRILL Data or IS-IS, see Section 11

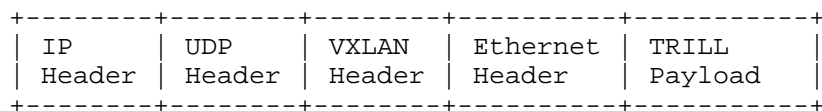
UDP Length - as specified in [RFC0768]

UDP Checksum - as specified in [RFC0768]

The TRILL Payload starts with the TRILL Header (not including the TRILL Ethertype) for TRILL Data packets and starts with the 0x83 Intradomain Routing Protocol Discriminator byte (thus not including the L2-IS-IS Ethertype) for TRILL IS-IS packets.

## 5.5 VXLAN Encapsulation

VXLAN [RFC7348] IP encapsulation of TRILL looks, on the wire, like TRILL over Ethernet over VXLAN over UDP over IP.



The outer UDP uses a destination port number indicating VXLAN and the outer UDP source port MAY be used for entropy as with native encapsulation (see Section 5.4). The VXLAN header after the outer UDP header adds a 24 bit Virtual Network Identifier (VNI). The Ethernet header after the VXLAN header and before the TRILL header consists of source MAC address, destination MAC address, and Ethertype. The Ethertype distinguishes TRILL Data from TRILL IS-IS; however, the destination and source MAC addresses in this inner Ethernet header are not used and are 12 wasted bytes.

A TRILL over IP port using VXLAN encapsulation by default uses a VNI of 1 but can be configured as described in Section 9.2.3.1 to use some other fixed VNI or to map from VLAN/FGL to VNI.

## 5.6 Other Encapsulations

It is anticipated that additional TRILL over IP encapsulations will be specified in future documents and allocated a bit in the TRILL Hello as per Section 11.3. A primary consideration for whether it is worth the effort to specify an encapsulation is good existing or anticipated fast path support.

## 6. Handling Multicast

By default, both TRILL IS-IS packets and multi-destination TRILL Data packets are sent to an All-RBridges IPv4 or IPv6 IP multicast Address as appropriate (see Section 11.2); however, a TRILL over IP port may be configured (see Section 9) to use a different multicast address or to use serial IP unicast with a list of one or more unicast IP addresses of other TRILL over IP ports to which multi-destination packets are sent. In the serial unicast case the outer IP header of each copy of the packet sent shows an IP unicast destination address even though the TRILL header has the M bit set to one to indicate multi-destination. Serial unicast configuration is necessary if the TRILL over IP port is connected to an IP network that does not support IP multicast. In any case, unicast TRILL packets are sent by unicast IP.

Even if a TRILL over IP port is configured to send multi-destination packets with serial unicast, it MUST be prepared to receive IP multicast TRILL packets. All TRILL over IP ports default to periodically transmitting appropriate IGMP (IPv4 [RFC3376] or MLD (IPv6 [RFC2710]) packets, so that the TRILL multicast IP traffic will be sent to them, unless they are configured not to do so.

Although TRILL fully supports broadcast links with more than 2 RBridges connected to the link there may be good reasons for configuring TRILL over IP ports to use serial unicast even where native IP multicast is available. Use of serial unicast provides the network manager with more precise control over adjacencies and how TRILL over IP links will be formed in an IP network. In some networks, unicast is more reliable than multicast. If multiple point-to-point TRILL over IP connections between parts of a TRILL campus are configured, TRILL will in any case spread traffic across them, treating them as parallel links, and appropriately fail over traffic if a link fails or incorporate a new link that comes up.

## 7. Use of IPsec and IKEv2

All TRILL switches (RBridges) that support TRILL over IP MUST implement IPsec [RFC4301] and support the use of IPsec Encapsulating Security Protocol (ESP [RFC4303]) in tunnel mode to secure both TRILL IS-IS and TRILL data packets. When IPsec is used to secure a TRILL over IP link and no IS-IS security is enabled, the IPsec session MUST be fully established before any TRILL IS-IS or data packets are exchanged. When there is IS-IS security [RFC5310] provided, implementers SHOULD use IS-IS security to protect TRILL IS-IS packets. However, in this case, the IPsec session still MUST be fully established before any data packets transmission since IS-IS security does not provide any protection to data packets.

All RBridges that support TRILL over IP MUST implement the Internet Key Exchange Protocol version 2 (IKEv2) for automated key management.

### 7.1 Keying

The following subsections discuss pairwise and group keying for TRILL over IP IPsec.

#### 7.1.1 Pairwise Keying

When IS-IS security is in use, IKEv2 will use a pre-shared key that incorporates the IS-IS shared key in order to bind the TRILL data session to the IS-IS session. The pre-shared key that will be used for IKEv2 exchanges for TRILL over IP is determined as follows:

```
HKDF-Expand-SHA256 ( IS-IS-key,
    "TRILL IP" | P1-System-ID | P1-Port | P2-System-ID | P2-Port )
```

In the above "|" indicates concatenation, HKDF is as in [RFC5869], SHA256 is as in [RFC6234], and "TRILL IP" is the eight byte US ASCII [RFC0020] string indicated. "IS-IS-key" is an IS-IS key usable for IS-IS security of link local IS-IS PDUs such as Hello, CSNP, and PSNP. This SHOULD be a link scope IS-IS key. With [RFC5310] there could be multiple keys identified with 16-bit key IDs. In this case, the Key ID of IS-IS-key is also used to identify the derived key. P1-System-ID and P2-System ID are the System IDs of the two TRILL RBridges, and P1-Port and P2-Port are the ports in use on each end. System IDs are guaranteed to be unique within the TRILL campus. Both of the RBridges involved treat the larger magnitude System ID, comparing System IDs as unsigned integers, as P1 and the smaller as P2 so both will derive the same key.



When IS-IS security is in use, the IS-IS-shared key from which the IKEv2 shared secret is derived might expire and be updated as described in [RFC5310]. The IKEv2 pre-shared keys derived from the IS-IS shared key MUST expire within the same lifetime as the IS-IS-shared key from which they were derived. When the IKEv2 pre-shared key expires, the IKEv2 Security Association must be rekeyed using a new shared secret derived from the new IS-IS shared key.

When IS-IS security is not in use, IKEv2 will not use a pre-shared key.

#### 7.1.2 Group Keying

In the case of a TRILL over IP port configured as point-to-point (see Section 4.2.4.1 of [RFC6325]), there is no group keying and the pairwise key determined as in Section 7.1.1 is used for IP multicast traffic.

In the case of a TRILL over IP port configured as broadcast but where the port is configured to use serial unicast (see Section 8), there is no group keying and the pairwise keying determined as in Section 7.1.1 is used for IP multicast traffic.

In the case of a TRILL over IP port configured as broadcast and using native multicast, ... tbd ...

#### 7.2 Mandatory-to-Implement Algorithms

All RBridges that support TRILL over IP MUST implement IPsec ESP [RFC4303] in tunnel mode. The implementation requirements for ESP cryptographic algorithms are as specified for IPsec. That specification is currently [RFC7321].

## 8. Transport Considerations

This section discusses a variety of important transport considerations.

### 8.1 Congestion Considerations

Section 3.1.3 of [RFC5405] discussed the congestion implications of UDP tunnels. As discussed in [RFC5405], because other flows can share the path with one or more UDP tunnels, congestion control [RFC2914] needs to be considered.

The default initial determination of the TRILL over IP encapsulation to be used through the exchange of TRILL IS-IS Hellos is a low bandwidth process. Hellos are not permitted to be sent any more often than once per second, and so are unlikely to cause congestion.

One motivation for including UDP in a TRILL encapsulation is to improve the use of multipath (such as ECMP) in cases where traffic is to traverse routers which are able to hash on UDP Port and IP address. In many cases this may reduce the occurrence of congestion and improve usage of available network capacity. However, it is also necessary to ensure that the network, including applications that use the network, responds appropriately in more difficult cases, such as when link or equipment failures have reduced the available capacity.

The impact of congestion must be considered both in terms of the effect on the rest of the network of a UDP tunnel that is consuming excessive capacity, and in terms of the effect on the flows using the UDP tunnels. The potential impact of congestion from a UDP tunnel depends upon what sort of traffic is carried over the tunnel, as well as the path of the tunnel.

TRILL is used to carry a wide range of traffic. In many cases TRILL is used to carry IP traffic. IP traffic is generally assumed to be congestion controlled, and thus a tunnel carrying general IP traffic (as might be expected to be carried across the Internet) generally does not need additional congestion control mechanisms. As specified in [RFC5405]:

"IP-based traffic is generally assumed to be congestion-controlled, i.e., it is assumed that the transport protocols generating IP-based traffic at the sender already employ mechanisms that are sufficient to address congestion on the path. Consequently, a tunnel carrying IP-based traffic should already interact appropriately with other traffic sharing the path, and specific congestion control mechanisms for the tunnel are not necessary".

For this reason, where TRILL is sent using UDP and used to carry IP traffic that is known to be congestion controlled, the UDP paths MAY be used across any combination of a single or cooperating service providers or across the general Internet.

However, TRILL is also used to carry traffic that is not necessarily congestion controlled. For example, TRILL may be used to carry traffic where specific bandwidth guarantees are provided.

In such cases congestion may be avoided by careful provisioning of the network and/or by rate limiting of user data traffic. Where TRILL is carried, directly or indirectly, over UDP over IP, the identity of each individual TRILL flow is in general lost.

For this reason, where the TRILL traffic is not congestion controlled, TRILL over UDP/IP MUST only be used within a single service provider that utilizes careful provisioning (e.g., rate limiting at the entries of the network while over-provisioning network capacity) to ensure against congestion, or within a limited number of service providers who closely cooperate in order to jointly provide this same careful provisioning. As such, TRILL over UDP/IP MUST NOT be used over the general Internet, or over non-cooperating service providers, to carry traffic that is not congestion-controlled.

Measures SHOULD be taken to prevent non-congestion-controlled TRILL over UDP/IP traffic from "escaping" to the general Internet, for example the following:

- a. Physical or logical isolation of the TRILL over IP links from the general Internet.
- b. Deployment of packet filters that block the UDP ports assigned for TRILL-over-UDP.
- c. Imposition of restrictions on TRILL over UDP/IP traffic by software tools used to set up TRILL over UDP paths between specific end systems (as might be used within a single data center).
- d. Use of a "Managed Circuit Breaker" for the TRILL traffic as described in [circuit-breaker].

## 8.2 Recursive Ingress

TRILL is specified to transport data to and from end stations over Ethernet and IP is frequently transported over Ethernet. Thus, an end station native data Ethernet frame EF might get TRILL ingressed to

TRILL(EF) that was then sent out a TRILL over IP over Ethernet port resulting in a packet on the wire of the form Ethernet(IP(TRILL(EF))). There is a risk of such a packet being re-ingressed by the same TRILL campus, due to physical or logical misconfiguration, looping round, being further re-ingressed, and so on. The packet might get discarded if it got too large but if fragmentation is enabled, it would just keep getting split into fragments that would continue to loop and grow and re-fragment until the path was saturated with junk and packets were being discarded due to queue overflow. The TRILL Header TTL would provide no protection because each TRILL ingress adds a new TRILL header with a new TTL.

To protect against this scenario, a TRILL over IP port MUST by default, test whether a TRILL packet it is about to transmit appears to be a TRILL ingress of a TRILL over IP over Ethernet packet. That is, is it of the form TRILL(Ethernet(IP(TRILL(...)))? If so, the default action of the TRILL over IP output port is to discard the packet rather than transmit it. However, there are cases where some level of nested ingress is desired so it MUST be possible to configure the port to allow such packets.

### 8.3 Fat Flows

For the purpose of load balancing, it is worthwhile to consider how to transport the TRILL packets over the Equal Cost Multiple Paths (ECMPs) existing internal to the IP path between TRILL over IP ports.

The ECMP election for the IP traffic could be based, at least for IPv4, on the quintuple of the outer IP header { Source IP, Destination IP, Source Port, Destination Port, and IP protocol }. Such tuples, however, could be exactly the same for all TRILL Data packets between two RBridge ports, even if there is a huge amount of data being sent between a variety of ingress and egress RBridges. One solution to this is to use the Source Port in as an entropy field. (This idea is also introduced in [gre-in-udp].) For example, for TRILL Data this entropy field could be based on some hash of the Inner.MacDA, Inner.MacSA, and Inner.VLAN or Inner.FGL. Unfortunately, this can conflict with middleboxes inside the TRILL over IP link (see 8.5). Therefore, in order to better support ECMP, a RBridge SHOULD set the Source Port to a range of values as an entropy field for ECMP decisions. However, if there are middleboxes in the path, the range of different Source Port values used MUST be restricted sufficiently to avoid disrupting connectivity.

## 8.4 MTU Considerations

In TRILL each TRILL switch advertises in its LSP number zero the largest LSP frame it can accept (but not less than 1,470 bytes) on any of its interfaces (at least those interfaces with adjacencies to other TRILL switches in the campus) through the `originatingLSPBufferSize` TLV [RFC6325] [RFC7177]. The campus minimum MTU (Maximum Transmission Unit), denoted *Sz*, is then established by taking the minimum of this advertised MTU for all R Bridges in the campus. Links that do not meet the *Sz* MTU are not included in the routing topology. This protects the operation of IS-IS from links that would be unable to accommodate some LSPs.

A method of determining `originatingLSPBufferSize` for an R Bridge with one or more TRILL over IP ports is described in [rfc7180bis]. However, if an IP link either can accommodate jumbo frames or is a link on which IP fragmentation is enabled and acceptable, then it is unlikely that the IP link will be a constraint on the `originatingLSPBufferSize` of an R Bridge using the link. On the other hand, if the IP link can only handle smaller frames and fragmentation is to be avoided when possible, a TRILL over IP port might constrain the R Bridge's `originatingLSPBufferSize`. Because TRILL sets the minimum values of *Sz* at 1,470 bytes, there may be links that meet the minimum MTU for the IP protocol (1,280 bytes for IPv6, 576 bytes for IPv4) on which it would be necessary to enable fragmentation for TRILL use.

The use of TRILL IS-IS MTU PDUs, as specified in [RFC6325] and [RFC7177] can provide added assurance of the actual MTU of a link.

## 8.5 Middlebox Considerations

This section gives some middlebox considerations for the IP encapsulations covered by this document, namely native and VXLAN encapsulation.

The requirements on the usage of the zero UDP Checksum in a UDP tunnel protocol are detailed in [RFC6936]. These requirements apply to TRILL over IP the encapsulations specified herein (native and VXLAN), which are applications of UDP tunnel.

Besides the Checksum, the Source Port number of the UDP header is also pertinent to the middlebox behavior. Network Address/Port Translator (NAPT) is the most commonly deployed Network Address Translation (NAT) device [RFC4787]. For a UDP tunnel protocol, the NAPT device establishes a NAT session to translate the {private IP address, private source port number} tuple to a {public IP address, public source port number} tuple, and vice versa, for the duration of

the UDP session. This provides the UDP tunnel protocol application with the "NAT-pass-through" function. NAPT allows multiple internal hosts to share a single public IP address. The port number, i.e., the UDP Source Port number, is used as the demultiplexer of the multiple internal hosts.

However, the above NAPT behavior conflicts with the behavior that the UDP Source Port number is used as an entropy (See Section 8.3). Hence, the tunnel operator **MUST** ensure the TRILL switch ports sending through local or remote NAPT middleboxes disable the entropy usage of the UDP Source Port number.

## 9. TRILL over IP Port Configuration

This section specifies the configuration information needed at a TRILL over IP port beyond that needed for a general RBridge port.

### 9.1 Per IP Port Configuration

Each RBridge port used for a TRILL over IP link should have at least one IP (v4 or v6) address. If no IP address is associated with the port, perhaps as a transient condition during re-configuration, the port is disabled. Implementations MAY allow a single port to operate as multiple IPv4 and/or IPv6 logical ports. Each IP address constitutes a different logical port and the RBridge with those ports MUST associate a different Port ID (see Section 4.4.2 of [RFC6325]) with each logical port.

By default a TRILL over IP port discards output packets that fail the possible recursive ingress test (see Section 10.1) unless configured to disable that test.

### 9.2 Additional per IP Address Configuration

The configuration information specified below is per TRILL over IP port IP address.

The mapping from TRILL packet priority to Differentiated Services Code Point (DSCP [RFC2474]) can be configured (see Section 10.5).

Each TRILL over IP port has a list of acceptable encapsulations it will use. By default this list consists of one entry for native encapsulation (see Section 7). Additional encapsulations MAY be configured. Additional configuration can be required or possible for specific encapsulations as described in Section 9.2.3.

Each IP address at a TRILL over IP port uses native IP multicast by default but may be configured whether to use serial IP unicast (Section 9.2.2) or native IP multicast (Section 9.2.1). Each IP address at a TRILL over IP is configured whether or not to use IPsec (Section 9.2.4).

#### 9.2.1 Native Multicast Configuration

If a TRILL over IP port address is using native IP multicast for multi-destination TRILL packets (IS-IS and data), by default

transmissions from that IP address use the IP multicast address (IPv4 or IPv6) specified in Section 11.2. The TRILL over IP port may be configured to use a different IP address to multicast packets.

### 9.2.2 Serial Unicast Configuration

If a TRILL over IP port address has been configured to use serial unicast for multi-destination packets (IS-IS and data), it should have associated with it a non-empty list of unicast IP destination addresses with the same IP version as the version of the port's IP address (IPv4 or IPv6). Multi-destination TRILL packets are serially unicast to the addresses in this list. Such a TRILL over IP port will only be able to form adjacencies [RFC7177] with the RBridges at the addresses in this list as those are the only RBridges to which it will send TRILL Hellos.

If this list of destination IP addresses is empty, there is no way to transmit a multi-destination TRILL over IP packet such as a TRILL Hello. Thus it is impossible to achieve adjacency [RFC7177] or if adjacency had been achieved (perhaps the list was non-empty and has just been configured to be empty), no way to maintain such adjacency. Thus, in the empty list case, TRILL Data multi-destination packets cannot be sent and TRILL Data unicast packets will not start flowing or, if they are already flowing, will soon cease, effectively disabling the port.

### 9.2.3 Encapsulation Specific Configuration

Specific TRILL over IP encapsulation methods may provide for further configuration as specified below.

#### 9.2.3.1 VXLAN Configuration

A TRILL over IP port using VXLAN encapsulation can be configured with a non-default VXLAN Network Identifier (VNI) that is used in that field of the VXLAN header for all TRILL packets sent using the encapsulation and required in all TRILL packets received using the encapsulation. The default VNI is 1. A TRILL packet received with the wrong VNI is discarded.

A TRILL over IP port using VXLAN encapsulation can also be configured to map the Inner.VLAN or Inner.FGL of a TRILL Data packet being transported to the value it places in the VNI field.



#### 9.2.3.2 Other Encapsulation Configuration

Additional encapsulation methods, beyond the native UDP encapsulation and VXLAN encapsulation specified in this document, may be specified in future documents and may require further configuration.

#### 9.2.4 Security Configuration

tbd ...

## 10. Security Considerations

TRILL over IP is subject to all of the security considerations for the base TRILL protocol [RFC6325]. In addition, there are specific security requirements for different TRILL deployment scenarios, as discussed in the "Use Cases for TRILL over IP" section above.

For communication between end stations in a TRILL campus, security is possible at three levels: end-to-end security between those end stations, edge-to-edge security between ingress and egress R Bridges [LinkSec], and link security to protect a TRILL hop. Any combination of these can be used, including all three.

TRILL over IP link security protects the contents of TRILL Data and IS-IS packets, including the identities of the end stations for data and the identities of the edge R Bridges, from observers of the link and transit devices within the link such as IP routers, but does not encrypt the link local IP addresses used in a packet and does not protect against observation by the sending and receiving R Bridges on the link. Edge-to-edge TRILL security protects the contents of TRILL data packets including the identities of the end stations for data from transit R Bridges but does not encrypt the identities of the edge R Bridges involved and does not protect against observation by those edge R Bridges. End-to-end security does not protect the identities of the end stations or edge R Bridge involved but does protect the content of TRILL data packets from observation by all R Bridges or other intervening devices between the end stations involved. End-to-end security should always be considered as an added layer of security and to protect any particularly sensitive information from unintended disclosure.

If VXLAN encapsulation is used, the unused Ethernet source and destination MAC addresses mentioned in Section 5.5, provide a 96 bit per packet covert path.

### 10.1 IPsec

This document specifies that all R Bridges that support TRILL over IP links MUST implement IPsec for the security of such links, and makes it clear that it is both wise and good to use IPsec in all cases where a TRILL over IP link will traverse a network that is not under the same administrative control as the rest of the TRILL campus or is not physically secure. IPsec is important, in these cases, to protect the privacy and integrity of data traffic. However, in cases where IPsec is impractical due to lack of fast path support, use of TRILL edge-to-edge security or use by the end stations of end-to-end security can provide significant security.

Further Security Considerations for IPsec ESP and for the cryptographic algorithms used with IPsec can be found in the RFCs referenced by this document.

## 10.2 IS-IS Security

TRILL over IP is compatible with the use of IS-IS Security [RFC5310], which can be used to authenticate TRILL switches before allowing them to join a TRILL campus. This is sufficient to protect against rogue devices impersonating TRILL switches, but is not sufficient to protect data packets that may be sent in TRILL over IP outside of the local network or across the public Internet. To protect the privacy and integrity of that traffic, use IPsec.

In cases where IPsec is used, the use of IS-IS security may not be necessary, but there is nothing about this specification that would prevent using both IPsec and IS-IS security together.

## 11. IANA Considerations

IANA considerations are given below.

### 11.1 Port Assignments

IANA is requested to assign destination UDP Ports for the TRILL IS-IS and Data channels:

UDP Port	Protocol
-----	-----
(TBD1)	TRILL IS-IS Channel
(TBD2)	TRILL Data Channel

### 11.2 Multicast Address Assignments

IANA is requested to one IPv4 and one IPv6 multicast address, as shown below, which correspond to the All-RBridges and All-IS-IS-RBridges multicast MAC addresses that the IEEE Registration Authority has assigned for TRILL. Because the low level hardware MAC address dispatch considerations for TRILL over Ethernet do not apply to TRILL over IP, one IP multicast address for each version of IP is sufficient.

(Values recommended to IANA in square brackets)

Name	IPv4	IPv6
-----	-----	-----
All-RBridges	TBD3[233.252.14.0]	TBD4[FF0X:0:0:0:0:0:0:205]

The hex digit "X" in the IPv6 address indicates the scope and defaults to 8. The IPv6 All-RBridges IP address may be used with other values of X.

### 11.3 Encapsulation Method Support Indication

The existing "RBridge Channel Protocols" registry is re-named and a new sub-registry under that registry added as follows:

The TRILL Parameters registry for "RBridge Channel Protocols" is renamed the "RBridge Channel Protocols and Link Technology Specific Flags" registry. [this document] is added as a second reference for this registry. The first part of the table is changed to the following:

Range	Registration	Note
-----	-----	-----
0x002-0x0FF	Standards Action	
0x100-0xFCF	RFC Required	allocation of a single value
0x100-0xFCF	IESG Approval	allocation of multiple values
0xFD0-0xFF7	see Note	link technology dependent, see subregistry

In the existing table of RBridge Channel Protocols, the following line is changed to two lines as shown:

OLD

0x004-0xFF7 Unassigned

NEW

0x004-0xFCF Unassigned

0xFD0-0xFF7 (link technology dependent, see subregistry)

A new subregistry under the re-named "RBridge Channel Protocols and Link Technology Specific Flags" registry is added as follows:

Name: TRILL over IP Link Flags  
 Registration Procedure: IETF Review  
 Reference: [this document]

Flag	Meaning	Reference
-----	-----	-----
0xFD0	Native encapsulation supported	[this document]
0xFD1	VXLAN encapsulation supported	[this document]
0xFD2-0xFF7	Unassigned	

## Normative References

- [IS-IS] - "Intermediate system to Intermediate system routing information exchange protocol for use in conjunction with the Protocol for providing the Connectionless-mode Network Service (ISO 8473)", ISO/IEC 10589:2002, 2002".
- [RFC0020] - Cerf, V., "ASCII format for network interchange", STD 80, RFC 20, DOI 10.17487/RFC0020, October 1969, <<http://www.rfc-editor.org/info/rfc20>>.
- [RFC0768] - Postel, J., "User Datagram Protocol", STD 6, RFC 768, DOI 10.17487/RFC0768, August 1980, <<http://www.rfc-editor.org/info/rfc768>>.
- [RFC2119] - Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, DOI 10.17487/RFC2119, March 1997, <<http://www.rfc-editor.org/info/rfc2119>>.
- [RFC2474] - Nichols, K., Blake, S., Baker, F., and D. Black, "Definition of the Differentiated Services Field (DS Field) in the IPv4 and IPv6 Headers", RFC 2474, DOI 10.17487/RFC2474, December 1998, <<http://www.rfc-editor.org/info/rfc2474>>.
- [RFC2710] - Deering, S., Fenner, W., and B. Haberman, "Multicast Listener Discovery (MLD) for IPv6", RFC 2710, DOI 10.17487/RFC2710, October 1999, <<http://www.rfc-editor.org/info/rfc2710>>.
- [RFC2914] - Floyd, S., "Congestion Control Principles", BCP 41, RFC 2914, DOI 10.17487/RFC2914, September 2000, <<http://www.rfc-editor.org/info/rfc2914>>.
- [RFC3168] - Ramakrishnan, K., Floyd, S., and D. Black, "The Addition of Explicit Congestion Notification (ECN) to IP", RFC 3168, DOI 10.17487/RFC3168, September 2001, <<http://www.rfc-editor.org/info/rfc3168>>.
- [RFC3376] - Cain, B., Deering, S., Kouvelas, I., Fenner, B., and A. Thyagarajan, "Internet Group Management Protocol, Version 3", RFC 3376, DOI 10.17487/RFC3376, October 2002, <<http://www.rfc-editor.org/info/rfc3376>>.
- [RFC4301] - Kent, S. and K. Seo, "Security Architecture for the Internet Protocol", RFC 4301, DOI 10.17487/RFC4301, December 2005, <<http://www.rfc-editor.org/info/rfc4301>>.
- [RFC4303] - Kent, S., "IP Encapsulating Security Payload (ESP)", RFC 4303, DOI 10.17487/RFC4303, December 2005, <<http://www.rfc-editor.org/info/rfc4303>>.

- [RFC5405] - Li, T. and R. Atkinson, "IS-IS Cryptographic Authentication", RFC 5304, DOI 10.17487/RFC5304, October 2008, <<http://www.rfc-editor.org/info/rfc5304>>.
- [RFC5310] - Bhatia, M., Manral, V., Li, T., Atkinson, R., White, R., and M. Fanto, "IS-IS Generic Cryptographic Authentication", RFC 5310, DOI 10.17487/RFC5310, February 2009, <<http://www.rfc-editor.org/info/rfc5310>>.
- [RFC5869] - Krawczyk, H. and P. Eronen, "HMAC-based Extract-and-Expand Key Derivation Function (HKDF)", RFC 5869, DOI 10.17487/RFC5869, May 2010, <<http://www.rfc-editor.org/info/rfc5869>>.
- [RFC6325] - Perlman, R., Eastlake 3rd, D., Dutt, D., Gai, S., and A. Ghanwani, "Routing Bridges (RBridges): Base Protocol Specification", RFC 6325, DOI 10.17487/RFC6325, July 2011, <<http://www.rfc-editor.org/info/rfc6325>>.
- [RFC7176] - Eastlake 3rd, D., Senevirathne, T., Ghanwani, A., Dutt, D., and A. Banerjee, "Transparent Interconnection of Lots of Links (TRILL) Use of IS-IS", RFC 7176, DOI 10.17487/RFC7176, May 2014, <<http://www.rfc-editor.org/info/rfc7176>>.
- [RFC7177] - Eastlake 3rd, D., Perlman, R., Ghanwani, A., Yang, H., and V. Manral, "Transparent Interconnection of Lots of Links (TRILL): Adjacency", RFC 7177, DOI 10.17487/RFC7177, May 2014, <<http://www.rfc-editor.org/info/rfc7177>>.
- [RFC7178] - Eastlake 3rd, D., Manral, V., Li, Y., Aldrin, S., and D. Ward, "Transparent Interconnection of Lots of Links (TRILL): RBridge Channel Support", RFC 7178, DOI 10.17487/RFC7178, May 2014, <<http://www.rfc-editor.org/info/rfc7178>>.
- [RFC7321] - McGrew, D. and P. Hoffman, "Cryptographic Algorithm Implementation Requirements and Usage Guidance for Encapsulating Security Payload (ESP) and Authentication Header (AH)", RFC 7321, DOI 10.17487/RFC7321, August 2014, <<http://www.rfc-editor.org/info/rfc7321>>.
- [RFC7348] - Mahalingam, M., Dutt, D., Duda, K., Agarwal, P., Kreeger, L., Sridhar, T., Bursell, M., and C. Wright, "Virtual eXtensible Local Area Network (VXLAN): A Framework for Overlaying Virtualized Layer 2 Networks over Layer 3 Networks", RFC 7348, DOI 10.17487/RFC7348, August 2014, <<http://www.rfc-editor.org/info/rfc7348>>.
- [rfc7180bis] - Eastlake, D., et al, "TRILL: Clarifications, Corrections, and Updates", draft-ietf-trill-rfc7180bis, work in progress.

## Informative References

- [RFC4787] - Audet, F., Ed., and C. Jennings, "Network Address Translation (NAT) Behavioral Requirements for Unicast UDP", BCP 127, RFC 4787, DOI 10.17487/RFC4787, January 2007, <<http://www.rfc-editor.org/info/rfc4787>>.
- [RFC6234] - Eastlake 3rd, D. and T. Hansen, "US Secure Hash Algorithms (SHA and SHA-based HMAC and HKDF)", RFC 6234, DOI 10.17487/RFC6234, May 2011, <<http://www.rfc-editor.org/info/rfc6234>>.
- [RFC6361] - Carlson, J. and D. Eastlake 3rd, "PPP Transparent Interconnection of Lots of Links (TRILL) Protocol Control Protocol", RFC 6361, DOI 10.17487/RFC6361, August 2011, <<http://www.rfc-editor.org/info/rfc6361>>.
- [RFC6864] - Touch, J., "Updated Specification of the IPv4 ID Field", RFC 6864, DOI 10.17487/RFC6864, February 2013, <<http://www.rfc-editor.org/info/rfc6864>>.
- [RFC6936] - Fairhurst, G. and M. Westerlund, "Applicability Statement for the Use of IPv6 UDP Datagrams with Zero Checksums", RFC 6936, DOI 10.17487/RFC6936, April 2013, <<http://www.rfc-editor.org/info/rfc6936>>.
- [RFC7172] - Eastlake 3rd, D., Zhang, M., Agarwal, P., Perlman, R., and D. Dutt, "Transparent Interconnection of Lots of Links (TRILL): Fine-Grained Labeling", RFC 7172, DOI 10.17487/RFC7172, May 2014, <<http://www.rfc-editor.org/info/rfc7172>>.
- [RFC7173] - Yong, L., Eastlake 3rd, D., Aldrin, S., and J. Hudson, "Transparent Interconnection of Lots of Links (TRILL) Transport Using Pseudowires", RFC 7173, DOI 10.17487/RFC7173, May 2014, <<http://www.rfc-editor.org/info/rfc7173>>.
- [RFC7296] - Kaufman, C., Hoffman, P., Nir, Y., Eronen, P., and T. Kivinen, "Internet Key Exchange Protocol Version 2 (IKEv2)", STD 79, RFC 7296, DOI 10.17487/RFC7296, October 2014, <<http://www.rfc-editor.org/info/rfc7296>>.
- [circuit-breaker] - Fairhurst, G., "Network Transport Circuit Breakers", draft-ietf-tsvwg-circuit-breaker, work in progress.
- [gre-in-udp] - Crabbe, E., Yong, L., and X. Xu, "Generic UDP Encapsulation for IP Tunneling", draft-yong-tsvwg-gre-in-udp-encap, work in progress.
- [LinkSec] - Eastlake, D., D. Zhang, "TRILL: Link Security", draft-



eastlake-trill-link-security, work in progress.

### Acknowledgements

The following people have provided useful feedback on the contents of this document: Sam Hartman, Adrian Farrel, and Mohammed Umair.

Some material in Section 10.2 is derived from draft-ietf-mppls-in-udp by Xiaohu Xu, Nischal Sheth, Lucy Yong, Carlos Pignataro, and Yongbing Fan.

The document was prepared in raw nroff. All macros used were defined within the source file.

Authors' Addresses

Margaret Cullen  
Painless Security  
356 Abbott Street  
North Andover, MA 01845  
USA

Phone: +1 781 405-7464  
Email: [margaret@painless-security.com](mailto:margaret@painless-security.com)  
URI: <http://www.painless-security.com>

Donald Eastlake  
Huawei Technologies  
155 Beaver Street  
Milford, MA 01757  
USA

Phone: +1 508 333-2270  
Email: [d3e3e3@gmail.com](mailto:d3e3e3@gmail.com)

Mingui Zhang  
Huawei Technologies  
No.156 Beiqing Rd. Haidian District,  
Beijing 100095 P.R. China

EMail: [zhangmingui@huawei.com](mailto:zhangmingui@huawei.com)

Dacheng Zhang  
Alibaba  
Beijing, Chao yang District  
P.R. China

Email: [dacheng.zdc@alibaba-inc.com](mailto:dacheng.zdc@alibaba-inc.com)

Copyright, Disclaimer, and Additional IPR Provisions

Copyright (c) 2015 IETF Trust and the persons identified as the document authors. All rights reserved.

This document is subject to BCP 78 and the IETF Trust's Legal Provisions Relating to IETF Documents (<http://trustee.ietf.org/license-info>) in effect on the date of publication of this document. Please review these documents carefully, as they describe your rights and restrictions with respect

to this document. Code Components extracted from this document must include Simplified BSD License text as described in Section 4.e of the Trust Legal Provisions and are provided without warranty as described in the Simplified BSD License.



TRILL Working Group  
INTERNET-DRAFT  
Intended status: Informational

Radia Perlman  
EMC  
Donald Eastlake  
Mingui Zhang  
Huawei  
Anoop Ghanwani  
Dell  
Hongjun Zhai  
JIT  
July 3, 2017

Expires: January 3, 2018

Alternatives for Multilevel TRILL  
(Transparent Interconnection of Lots of Links)  
<draft-ietf-trill-rbridge-multilevel-07.txt>

## Abstract

Although TRILL is based on IS-IS, which supports multilevel unicast routing, extending TRILL to multiple levels has challenges that are not addressed by the already-existing capabilities of IS-IS. One issue is with the handling of multi-destination packet distribution trees. Other issues are with TRILL switch nicknames. How are such nicknames allocated across a multilevel TRILL network? Do nicknames need to be unique across an entire multilevel TRILL network or can they merely be unique within each multilevel area?

This informational document enumerates and examines alternatives based on a number of factors including backward compatibility, simplicity, and scalability and makes recommendations in some cases.

## Status of This Memo

This Internet-Draft is submitted to IETF in full conformance with the provisions of BCP 78 and BCP 79. Distribution of this document is unlimited. Comments should be sent to the TRILL working group mailing list <trill@ietf.org>.

Internet-Drafts are working documents of the Internet Engineering Task Force (IETF), its areas, and its working groups. Note that other groups may also distribute working documents as Internet-Drafts.

Internet-Drafts are draft documents valid for a maximum of six months and may be updated, replaced, or obsoleted by other documents at any time. It is inappropriate to use Internet-Drafts as reference material or to cite them other than as "work in progress."

The list of current Internet-Drafts can be accessed at  
<http://www.ietf.org/lid-abstracts.html>. The list of Internet-Draft  
Shadow Directories can be accessed at  
<http://www.ietf.org/shadow.html>.

## Table of Contents

1. Introduction.....	4
1.1 The Motivation for Multilevel.....	4
1.2 Improvements Due to Multilevel.....	5
1.2.1. The Routing Computation Load.....	5
1.2.2. LSDB Volatility Creating Too Much Control Traffic...	5
1.2.3. LSDB Volatility Causing To Much Time Unconverged....	6
1.2.4. The Size Of The LSDB.....	6
1.2.5 Nickname Limit.....	6
1.2.6 Multi-Destination Traffic.....	7
1.3 Unique and Aggregated Nicknames.....	7
1.4 More on Areas.....	8
1.5 Terminology and Acronyms.....	8
2. Multilevel TRILL Issues.....	10
2.1 Non-zero Area Addresses.....	11
2.2 Aggregated versus Unique Nicknames.....	11
2.2.1 More Details on Unique Nicknames.....	12
2.2.2 More Details on Aggregated Nicknames.....	13
2.2.2.1 Border Learning Aggregated Nicknames.....	14
2.2.2.2 Swap Nickname Field Aggregated Nicknames.....	16
2.2.2.3 Comparison.....	17
2.3 Building Multi-Area Trees.....	17
2.4 The RPF Check for Trees.....	18
2.5 Area Nickname Acquisition.....	18
2.6 Link State Representation of Areas.....	19
3. Area Partition.....	20
4. Multi-Destination Scope.....	21
4.1 Unicast to Multi-destination Conversions.....	21
4.1.1 New Tree Encoding.....	22
4.2 Selective Broadcast Domain Reduction.....	22
5. Co-Existence with Old TRILL switches.....	24
6. Multi-Access Links with End Stations.....	25
7. Summary.....	27
8. Security Considerations.....	28
9. IANA Considerations.....	28
Normative References.....	29
Informative References.....	29
Acknowledgements.....	31
Authors' Addresses.....	32



## 1. Introduction

The IETF TRILL (Transparent Interconnection of Lot of Links) protocol [RFC6325] [RFC7177] [RFC7780] provides optimal pair-wise data routing without configuration, safe forwarding even during periods of temporary loops, and support for multipathing of both unicast and multicast traffic in networks with arbitrary topology and link technology, including multi-access links. TRILL accomplishes this by using IS-IS (Intermediate System to Intermediate System [IS-IS] [RFC7176]) link state routing in conjunction with a header that includes a hop count. The design supports data labels (VLANs and Fine Grained Labels [RFC7172]) and optimization of the distribution of multi-destination data based on data label and multicast group. Devices that implement TRILL are called TRILL Switches or RBridges.

Familiarity with [IS-IS], [RFC6325], and [RFC7780] is assumed in this document.

### 1.1 The Motivation for Multilevel

The primary motivation for multilevel TRILL is to improve scalability. The following issues might limit the scalability of a TRILL-based network:

1. The routing computation load
2. The volatility of the link state database (LSDB) creating too much control traffic
3. The volatility of the LSDB causing the TRILL network to be in an unconverged state too much of the time
4. The size of the LSDB
5. The limit of the number of TRILL switches, due to the 16-bit nickname space (for further information on why this might be a problem, see Section 1.2.5)
6. The traffic due to upper layer protocols use of broadcast and multicast
7. The size of the end node learning table (the table that remembers (egress TRILL switch, label/MAC) pairs)

As discussed below, extending TRILL IS-IS to be multilevel (hierarchical) can help with all of these issues except issue 7.

IS-IS was designed to be multilevel [IS-IS]. A network can be partitioned into "areas". Routing within an area is known as "Level 1 routing". Routing between areas is known as "Level 2 routing". The Level 2 IS-IS network consists of Level 2 routers and links between the Level 2 routers. Level 2 routers may participate in one or more Level 1 areas, in addition to their role as Level 2 routers.

Each area is connected to Level 2 through one or more "border routers", which participate both as a router inside the area, and as a router inside the Level 2 "area". Care must be taken that it is clear, when transitioning multi-destination packets between Level 2 and a Level 1 area in either direction, that exactly one border TRILL switch will transition a particular data packet between the levels or else duplication or loss of traffic can occur.

## 1.2 Improvements Due to Multilevel

Partitioning the network into areas directly solves the first four scalability issues listed above as described in Sections 1.2.1 through 1.2.4. Multilevel also contributes to solving issues 5 and 6 as discussed in Section 1.2.5 and 1.2.6 respectively.

In the subsections below,  $N$  indicates the number of TRILL switches in a TRILL campus. As a simplifying assumption, it is assumed that each TRILL switch has  $k$  links to other TRILL switches. An "optimized" multilevel campus is assumed to have Level 1 areas containing  $\sqrt{N}$  switches.

### 1.2.1. The Routing Computation Load

The Dijkstra algorithm uses computational effort on the order of the number of links in a network ( $N*k$ ) times the log of the number of nodes to calculate least cost routes at a router (Section 12.3.3 [InterCon]). Thus, in a single level TRILL campus, it is on the order of  $N*k*\log(N)$ . In an optimized multilevel campus, it is on the order of  $\sqrt{N}*k*\log(N)$ . So, for example, assuming  $N$  is 3,000, the level of computational effort would be reduced by about a factor of 50.

### 1.2.2. LSDB Volatility Creating Too Much Control Traffic

The rate of LSDB changes is assumed to be approximately proportional to the number of routers and links in the TRILL campus or  $N*(1+k)$  for a single level campus. With an optimized multilevel campus, each area would have about  $\sqrt{N}$  routers and proportionately fewer links reducing the rate of LSDB changes by about a factor of  $\sqrt{N}$ .

### 1.2.3. LSDB Volatility Causing To Much Time Unconverged

With the simplifying assumption that routing converges after each topology change before the next such change, the fraction of time that routing is unconverged is proportional to the product of the rate of change occurrence and the convergence time. The rate of topology changes per some arbitrary unit of time will be roughly proportional to the number of router and links (Section 1.2.2). The convergence time is approximately proportional to the computation involved at each router (Section 1.2.1). Thus, based on these simplifying assumptions, the time spent unconverged in a single level network is proportional to  $(N*(1+k))*(N*k*\log(N))$  while that time for an optimized multilevel network would be proportional to  $(\sqrt{N}*(1+k))*(\sqrt{N}*k*\log(N))$ . Thus, in changing to multilevel, the time spent unconverged, using these simplifying assumptions, is improved by about a factor of  $N$ .

### 1.2.4. The Size Of The LSDB

The size of the LSDB, which consists primarily of information about routers (TRILL switches) and links, is also approximately proportional to the number of routers and links. So, as with item 2 in Section 1.2.2 above, it should improve by about a factor of  $\sqrt{N}$  in going from single to multilevel.

### 1.2.5 Nickname Limit

For many TRILL protocol purposes, RBridges are designated by 16-bit nicknames. While some values are reserved, this appears to provide enough nicknames to designated over 65,000 RBridges. However, this number is effectively reduced by the following two factors:

- Nicknames are consumed when pseudo-nicknames are used for the active-active connection of end stations. Using the techniques in [RFC7781], for example, could double the nickname consumption if there are extensive active-active edge groups connected to different sets of edge TRILL switch ports.
- There might be problems in multilevel campus wide contention for single nickname allocation of nicknames were allocated individually from a single pool for the entire campus. Thus it seems likely that a hierarchical method would be chosen where blocks of nicknames are allocated at Level 2 to Level 1 areas and contention for a nickname by an RBridge in such a Level 1 area would be only within that area. Such hierarchical allocation leads to further effective loss of nicknames similar to the situation

with IP addresses discussed in [RFC3194].

Even without the above effective reductions in nickname space, a very large multilevel TRILL campus, say one with 200 areas each containing 500 TRILL switches, could require 100,000 or more nicknames if all nicknames in the campus must be unique, which is clearly impossible with 16-bit nicknames.

This scaling limit, namely, 16-bit nickname space, will only be addressed with the aggregated nickname approach. Since the aggregated nickname approach requires some complexity in the border TRILL switches (for rewriting the nicknames in the TRILL header), the suggested design in this document allows a campus with a mixture of unique-nickname areas, and aggregated-nickname areas. Thus a TRILL network could start using multilevel with the simpler unique nickname method and later add aggregated areas as a later stage of network growth.

With this design, nicknames must be unique across all Level 2 and unique-nickname area TRILL switches taken together, whereas nicknames inside an aggregated-nickname area are visible only inside that area. Nicknames inside an aggregated-nickname area must still not conflict with nicknames visible in Level 2 (which includes all nicknames inside unique nickname areas), but the nicknames inside an aggregated-nickname area may be the same as nicknames used within one or more other aggregated-nickname areas.

With the design suggested in this document, TRILL switches within an area need not be aware of whether they are in an aggregated nickname area or a unique nickname area. The border TRILL switches in area A1 will indicate, in their LSP inside area A1, which nicknames (or nickname ranges) are available, or alternatively which nicknames are not available, for choosing as nicknames by area A1 TRILL switches.

#### 1.2.6 Multi-Destination Traffic

Scaling limits due to protocol use of broadcast and multicast, can be addressed in many cases in a multilevel campus by introducing locally-scoped multi-destination delivery, limited to an area or a single link. See further discussion of this issue in Section 4.2.

#### 1.3 Unique and Aggregated Nicknames

We describe two alternatives for hierarchical or multilevel TRILL. One we call the "unique nickname" alternative. The other we call the "aggregated nickname" alternative. In the aggregated nickname

alternative, border TRILL switches replace either the ingress or egress nickname field in the TRILL header of unicast packets with an aggregated nickname representing an entire area.

The unique nickname alternative has the advantage that border TRILL switches are simpler and do not need to do TRILL Header nickname modification. It also simplifies testing and maintenance operations that originate in one area and terminate in a different area.

The aggregated nickname alternative has the following advantages:

- o it solves scaling problem #5 above, the 16-bit nickname limit, in a simple way,
- o it lessens the amount of inter-area routing information that must be passed in IS-IS, and
- o it logically reduces the RPF (Reverse Path Forwarding) Check information (since only the area nickname needs to appear, rather than all the ingress TRILL switches in that area).

In both cases, it is possible and advantageous to compute multi-destination data packet distribution trees such that the portion computed within a given area is rooted within that area.

For further discussion of the unique and aggregated nickname alternatives, see Section 2.2.

#### 1.4 More on Areas

Each area is configured with an "area address", which is advertised in IS-IS messages, so as to avoid accidentally interconnecting areas. For TRILL the only purpose of the area address would be to avoid accidentally interconnecting areas although the area address had other purposes in CLNP (Connectionless Network Layer Protocol), IS-IS was originally designed for CLNP/DECnet.

Currently, the TRILL specification says that the area address must be zero. If we change the specification so that the area address value of zero is just a default, then most of IS-IS multilevel machinery works as originally designed. However, there are TRILL-specific issues, which we address below in Section 2.1.

#### 1.5 Terminology and Acronyms

This document generally uses the acronyms defined in [RFC6325] plus the additional acronym DBRB. However, for ease of reference, most acronyms used are listed here:

CLNP - ConnectionLess Network Protocol

DECnet - a proprietary routing protocol that was used by Digital Equipment Corporation. "DECnet Phase 5" was the origin of IS-IS.

Data Label - VLAN or Fine Grained Label [RFC7172]

DBRB - Designated Border RBridge

ESADI - End Station Address Distribution Information

IS-IS - Intermediate System to Intermediate System [IS-IS]

LSDB - Link State Data Base

LSP - Link State PDU

PDU - Protocol Data Unit

RBridge - Routing Bridge, an alternative name for a TRILL switch

RPF - Reverse Path Forwarding

TLV - Type Length Value

TRILL - Transparent Interconnection of Lots of Links or Tunneled Routing in the Link Layer [RFC6325] [RFC7780]

TRILL switch - a device that implements the TRILL protocol [RFC6325] [RFC7780], sometimes called an RBridge

VLAN - Virtual Local Area Network

## 2. Multilevel TRILL Issues

The TRILL-specific issues introduced by multilevel include the following:

- a. Configuration of non-zero area addresses, encoding them in IS-IS PDUs, and possibly interworking with old TRILL switches that do not understand non-zero area addresses.

See Section 2.1.

- b. Nickname management.

See Sections 2.5 and 2.2.

- c. Advertisement of pruning information (Data Label reachability, IP multicast addresses) across areas.

Distribution tree pruning information is only an optimization, as long as multi-destination packets are not prematurely pruned. For instance, border TRILL switches could advertise they can reach all possible Data Labels, and have an IP multicast router attached. This would cause all multi-destination traffic to be transmitted to border TRILL switches, and possibly pruned there, when the traffic could have been pruned earlier based on Data Label or multicast group if border TRILL switches advertised more detailed Data Label and/or multicast listener and multicast router attachment information.

- d. Computation of distribution trees across areas for multi-destination data.

See Section 2.3.

- e. Computation of RPF information for those distribution trees.

See Section 2.4.

- f. Computation of pruning information across areas.

See Sections 2.3 and 2.6.

- g. Compatibility, as much as practical, with existing, unmodified TRILL switches.

The most important form of compatibility is with existing TRILL fast path hardware. Changes that require upgrade to the slow path firmware/software are more tolerable. Compatibility for the relatively small number of border TRILL switches is less important than compatibility for non-border TRILL switches.

See Section 5.

## 2.1 Non-zero Area Addresses

The current TRILL base protocol specification [RFC6325] [RFC7177] [RFC7780] says that the area address in IS-IS must be zero. The purpose of the area address is to ensure that different areas are not accidentally merged. Furthermore, zero is an invalid area address for layer 3 IS-IS, so it was chosen as an additional safety mechanism to ensure that layer 3 IS-IS packets would not be confused with TRILL IS-IS packets. However, TRILL uses other techniques to avoid confusion on a link, such as different multicast addresses and Ethertypes on Ethernet [RFC6325], different PPP (Point-to-Point Protocol) code points on PPP [RFC6361], and the like. Thus, using an area address in TRILL that might be used in layer 3 IS-IS is not a problem.

Since current TRILL switches will reject any IS-IS messages with non-zero area addresses, the choices are as follows:

- a.1 upgrade all TRILL switches that are to interoperate in a potentially multilevel environment to understand non-zero area addresses,
- a.2 neighbors of old TRILL switches must remove the area address from IS-IS messages when talking to an old TRILL switch (which might break IS-IS security and/or cause inadvertent merging of areas),
- a.3 ignore the problem of accidentally merging areas entirely, or
- a.4 keep the fixed "area address" field as 0 in TRILL, and add a new, optional TLV for "area name" to Hellos that, if present, could be compared, by new TRILL switches, to prevent accidental area merging.

In principal, different solutions could be used in different areas but it would be much simpler to adopt one of these choices uniformly. A simple solution would be a.1 above with each TRILL switch using a dominant area nickname as its area address. For the unique nickname alternative, the dominant nickname could be the lowest value nickname held by any border RBridge of the area. For the aggregated nickname alternative, it could be the lowest nickname held by a border RBridge of the area or a nickname representing the area.

## 2.2 Aggregated versus Unique Nicknames

In the unique nickname alternative, all nicknames across the campus must be unique. In the aggregated nickname alternative, TRILL switch nicknames within an aggregated area are only of local significance,



and the only nickname externally (outside that area) visible is the "area nickname" (or nicknames), which aggregates all the internal nicknames.

The unique nickname approach simplifies border TRILL switches.

The aggregated nickname approach eliminates the potential problem of nickname exhaustion, minimizes the amount of nickname information that would need to be forwarded between areas, minimizes the size of the forwarding table, and simplifies RPF calculation and RPF information.

### 2.2.1 More Details on Unique Nicknames

With unique cross-area nicknames, it would be intractable to have a flat nickname space with TRILL switches in different areas contending for the same nicknames. Instead, each area would need to be configured with or allocate one or more block of nicknames. Either some TRILL switches would need to announce that all the nicknames other than that in blocks available to the area are taken (to prevent the TRILL switches inside the area from choosing nicknames outside the area's nickname block), or a new TLV would be needed to announce the allowable or the prohibited nicknames, and all TRILL switches in the area would need to understand that new TLV.

Currently the encoding of nickname information in TLVs is by listing of individual nicknames; this would make it painful for a border TRILL switch to announce into an area that it is holding all other nicknames to limit the nicknames available within that area. Painful means tens of thousands of individual nickname entries in the Level 1 LSDB. The information could be encoded as ranges of nicknames to make this manageable by specifying a new TLV similar to the Nickname Flags APPsub-TLV specified in [RFC7780] but providing flags for blocks of nicknames rather than single nicknames. Although this would require updating software, such a new TLV is the preferred method.

There is also an issue with the unique nicknames approach in building distribution trees, as follows:

With unique nicknames in the TRILL campus and TRILL header nicknames not rewritten by the border TRILL switches, there would have to be globally known nicknames for the trees. Suppose there are  $k$  trees. For all of the trees with nicknames located outside an area, the local trees would be rooted at a border TRILL switch or switches. Therefore, there would be either no splitting of multi-destination traffic within the area or restricted splitting of multi-destination traffic between trees rooted at a highly restricted set of TRILL switches.

As an alternative, just the "egress nickname" field of multi-destination TRILL Data packets could be mapped at the border, leaving known unicast packets un-mapped. However, this surrenders much of the unique nickname advantage of simpler border TRILL switches.

Scaling to a very large campus with unique nicknames might exhaust the 16-bit TRILL nicknames space particularly if (1) additional nicknames are consumed to support active-active end station groups at the TRILL edge using the techniques standardized in [RFC7781] and (2) use of the nickname space is less efficient due to the allocation of, for example, power-of-two size blocks of nicknames to areas in the same way that use of the IP address space is made less efficient by hierarchical allocation (see [RFC3194]). One method to avoid nickname exhaustion might be to expand nicknames to 24 bits; however, that technique would require TRILL message format and fast path processing changes and that all TRILL switches in the campus understand larger nicknames.

#### 2.2.2 More Details on Aggregated Nicknames

The aggregated nickname approach enables passing far less nickname information. It works as follows, assuming both the source and destination areas are using aggregated nicknames:

There are at least two ways areas could be identified.

One method would be to assign each area a 16-bit nickname. This would not be the nickname of any actual TRILL switch. Instead, it would be the nickname of the area itself. Border TRILL switches would know the area nickname for their own area(s). For an example of a more specific multilevel proposal using unique nicknames, see [DraftUnique].

Alternatively, areas could be identified by the set of nicknames that identify the border routers for that area. (See [SingleName] for a multilevel proposal using such a set of nicknames.)

The TRILL Header nickname fields in TRILL Data packets being transported through a multilevel TRILL campus with aggregated nicknames are as follows:

- When both the ingress and egress TRILL switches are in the same area, there need be no change from the existing base TRILL protocol standard in the TRILL Header nickname fields.
- When being transported between different Level 1 areas in Level 2, the ingress nickname is a nickname of the ingress TRILL

switch's area while the egress nickname is either a nickname of the egress TRILL switch's area or a tree nickname.

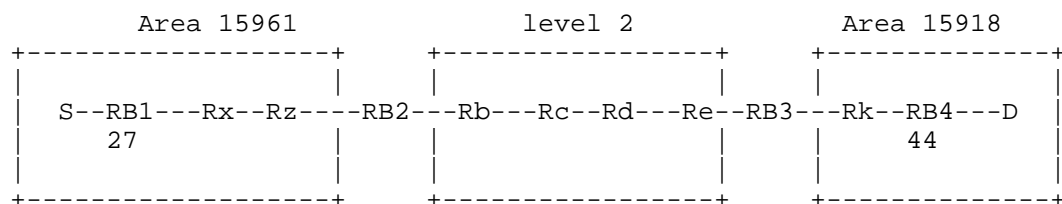
- When being transported from Level 1 to Level 2, the ingress nickname is the nickname of the ingress TRILL switch itself while the egress nickname is either a nickname for the area of the egress TRILL switch or a tree nickname.
- When being transported from Level 2 to Level 1, the ingress nickname is a nickname for the ingress TRILL switch's area while the egress nickname is either the nickname of the egress TRILL switch itself or a tree nickname.

There are two variations of the aggregated nickname approach. The first is the Border Learning approach, which is described in Section 2.2.2.1. The second is the Swap Nickname Field approach, which is described in Section 2.2.2.2. Section 2.2.2.3 compares the advantages and disadvantages of these two variations of the aggregated nickname approach.

#### 2.2.2.1 Border Learning Aggregated Nicknames

This section provides an illustrative example and description of the border learning variation of aggregated nicknames where a single nickname is used to identify an area.

In the following picture, RB2 and RB3 are area border TRILL switches (RBridges). A source S is attached to RB1. The two areas have nicknames 15961 and 15918, respectively. RB1 has a nickname, say 27, and RB4 has a nickname, say 44 (and in fact, they could even have the same nickname, since the TRILL switch nickname will not be visible outside these aggregated areas).



Let's say that S transmits a frame to destination D, which is connected to RB4, and let's say that D's location has already been learned by the relevant TRILL switches. These relevant switches have learned the following:

- 1) RB1 has learned that D is connected to nickname 15918
- 2) RB3 has learned that D is attached to nickname 44.

The following sequence of events will occur:

- S transmits an Ethernet frame with source MAC = S and destination MAC = D.
- RB1 encapsulates with a TRILL header with ingress RBridge = 27, and egress = 15918 producing a TRILL Data packet.
- RB2 has announced in the Level 1 IS-IS instance in area 15961, that it is attached to all the area nicknames, including 15918. Therefore, IS-IS routes the packet to RB2. Alternatively, if a distinguished range of nicknames is used for Level 2, Level 1 TRILL switches seeing such an egress nickname will know to route to the nearest border router, which can be indicated by the IS-IS attached bit.
- RB2, when transitioning the packet from Level 1 to Level 2, replaces the ingress TRILL switch nickname with the area nickname, so replaces 27 with 15961. Within Level 2, the ingress RBridge field in the TRILL header will therefore be 15961, and the egress RBridge field will be 15918. Also RB2 learns that S is attached to nickname 27 in area 15961 to accommodate return traffic.
- The packet is forwarded through Level 2, to RB3, which has advertised, in Level 2, reachability to the nickname 15918.
- RB3, when forwarding into area 15918, replaces the egress nickname in the TRILL header with RB4's nickname (44). So, within the destination area, the ingress nickname will be 15961 and the egress nickname will be 44.
- RB4, when decapsulating, learns that S is attached to nickname 15961, which is the area nickname of the ingress.

Now suppose that D's location has not been learned by RB1 and/or RB3. What will happen, as it would in TRILL today, is that RB1 will forward the packet as multi-destination, choosing a tree. As the multi-destination packet transitions into Level 2, RB2 replaces the ingress nickname with the area nickname. If RB1 does not know the location of D, the packet must be flooded, subject to possible pruning, in Level 2 and, subject to possible pruning, from Level 2 into every Level 1 area that it reaches on the Level 2 distribution tree.

Now suppose that RB1 has learned the location of D (attached to nickname 15918), but RB3 does not know where D is. In that case, RB3 must turn the packet into a multi-destination packet within area 15918. In this case, care must be taken so that in the case in which RB3 is not the Designated transitioner between Level 2 and its area for that multi-destination packet, but was on the unicast path, that

border TRILL switch in that area does not forward the now multi-destination packet back into Level 2. Therefore, it would be desirable to have a marking, somehow, that indicates the scope of this packet's distribution to be "only this area" (see also Section 4).

In cases where there are multiple transitioners for unicast packets, the border learning mode of operation requires that the address learning between them be shared by some protocol such as running ESADI [RFC7357] for all Data Labels of interest to avoid excessive unknown unicast flooding.

The potential issue described at the end of Section 2.2.1 with trees in the unique nickname alternative is eliminated with aggregated nicknames. With aggregated nicknames, each border TRILL switch that will transition multi-destination packets can have a mapping between Level 2 tree nicknames and Level 1 tree nicknames. There need not even be agreement about the total number of trees; just that the border TRILL switch have some mapping, and replace the egress TRILL switch nickname (the tree name) when transitioning levels.

#### 2.2.2.2 Swap Nickname Field Aggregated Nicknames

There is a variant possibility where two additional fields could exist in TRILL Data packets that could be called the "ingress swap nickname field" and the "egress swap nickname field". This variant is described below for completeness but would require fast path hardware changes from the existing TRILL protocol. The changes in the example above would be as follows:

- RB1 will have learned the area nickname of D and the TRILL switch nickname of RB4 to which D is attached. In encapsulating a frame to D, it puts an area nickname of D (15918) in the egress nickname field of the TRILL Header and puts a nickname of RB3 (44) in a egress swap nickname field.
- RB2 moves the ingress nickname to the ingress swap nickname field and inserts 15961, an area nickname for S, into the ingress nickname field.
- RB3 swaps the egress nickname and the egress swap nickname fields, which sets the egress nickname to 44.
- RB4 learns the correspondence between the source MAC/VLAN of S and the { ingress nickname, ingress swap nickname field } pair as it decapsulates and egresses the frame.

See [DraftAggregated] for a multilevel proposal using aggregated swap

nicknames with a single nickname representing an area.

#### 2.2.2.3 Comparison

The Border Learning variant described in Section 2.2.2.1 above minimizes the change in non-border TRILL switches but imposes the burden on border TRILL switches of learning and doing lookups in all the end station MAC addresses within their area(s) that are used for communication outside the area. This burden could be reduced by decreasing the area size and increasing the number of areas.

The Swap Nickname Field variant described in Section 2.2.2.2 eliminates the extra address learning burden on border TRILL switches but requires changes to the TRILL data packet header and more extensive changes to non-border TRILL switches. In particular, with this alternative, non-border TRILL switches must learn to associate both a TRILL switch nickname and an area nickname with end station MAC/label pairs (except for addresses that are local to their area).

The Swap Nickname Field alternative is more scalable but less backward compatible for non-border TRILL switches. It would be possible for border and other level 2 TRILL switches to support both Border Learning, for support of legacy Level 1 TRILL switches, and Swap Nickname, to support Level 1 TRILL switches that understood the Swap Nickname method based on variations in the TRILL header but this would be even more complex.

The requirement to change the TRILL header and fast path processing to support the Swap Nickname Field variant make it impractical for the foreseeable future.

### 2.3 Building Multi-Area Trees

It is easy to build a multi-area tree by building a tree in each area separately, (including the Level 2 "area"), and then having only a single border TRILL switch, say RBx, in each area, attach to the Level 2 area. RBx would forward all multi-destination packets between that area and Level 2.

People might find this unacceptable, however, because of the desire to path split (not always sending all multi-destination traffic through the same border TRILL switch).

This is the same issue as with multiple ingress TRILL switches injecting traffic from a pseudonode, and can be solved with the mechanism that was adopted for that purpose: the affinity TLV

[RFC7783]. For each tree in the area, at most one border RB announces itself in an affinity TLV with that tree name.

#### 2.4 The RPF Check for Trees

For multi-destination data originating locally in RBx's area, computation of the RPF check is done as today. For multi-destination packets originating outside RBx's area, computation of the RPF check must be done based on which one of the border TRILL switches (say RB1, RB2, or RB3) injected the packet into the area.

A TRILL switch, say RB4, located inside an area, must be able to know which of RB1, RB2, or RB3 transitioned the packet into the area from Level 2 (or into Level 2 from an area).

This could be done based on having the DBRB announce the transitioner assignments to all the TRILL switches in the area, or the Affinity TLV mechanism given in [RFC7783], or a New Tree Encoding mechanism discussed in Section 4.1.1.

#### 2.5 Area Nickname Acquisition

In the aggregated nickname alternative, each area must acquire a unique area nickname or can be identified by the set of border TRILL switches. It is probably simpler to allocate a block of nicknames (say, the top 4000) to either (1) represent areas and not specific TRILL switches or (2) used by border TRILL switches if the set of such border TRILL switches represent the area.

The nicknames used for area identification need to be advertised and acquired through Level 2.

Within an area, all the border TRILL switches can discover each other through the Level 1 link state database, by using the IS-IS attach bit or by explicitly advertising in their LSP "I am a border RBridge".

Of the border TRILL switches, one will have highest priority (say RB7). RB7 can dynamically participate, in Level 2, to acquire a nickname for identifying the area. Alternatively, RB7 could give the area a pseudonode IS-IS ID, such as RB7.5, within Level 2. So an area would appear, in Level 2, as a pseudonode and the pseudonode could participate, in Level 2, to acquire a nickname for the area.

Within Level 2, all the border TRILL switches for an area can advertise reachability to the area, which would mean connectivity to

a nickname identifying the area.

## 2.6 Link State Representation of Areas

Within an area, say area A1, there is an election for the DBRB, (Designated Border RBridge), say RB1. This can be done through LSPs within area A1. The border TRILL switches announce themselves, together with their DBRB priority. (Note that the election of the DBRB cannot be done based on Hello messages, because the border TRILL switches are not necessarily physical neighbors of each other. They can, however, reach each other through connectivity within the area, which is why it will work to find each other through Level 1 LSPs.)

RB1 can acquire an area nickname (in the aggregated nickname approach) and may give the area a pseudonode IS-IS ID (just like the DRB would give a pseudonode IS-IS ID to a link) depending on how the area nickname is handled. RB1 advertises, in area A1, an area nickname that RB1 has acquired (and what the pseudonode IS-IS ID for the area is if needed).

Level 1 LSPs (possibly pseudonode) initiated by RB1 for the area include any information external to area A1 that should be input into area A1 (such as nicknames of external areas, or perhaps (in the unique nickname variant) all the nicknames of external TRILL switches in the TRILL campus and pruning information such as multicast listeners and labels). All the other border TRILL switches for the area announce (in their LSP) attachment to that area.

Within Level 2, RB1 generates a Level 2 LSP on behalf of the area. The same pseudonode ID could be used within Level 1 and Level 2, for the area. (There does not seem any reason why it would be useful for it to be different, but there's also no reason why it would need to be the same). Likewise, all the area A1 border TRILL switches would announce, in their Level 2 LSPs, connection to the area.



### 3. Area Partition

It is possible for an area to become partitioned, so that there is still a path from one section of the area to the other, but that path is via the Level 2 area.

With multilevel TRILL, an area will naturally break into two areas in this case.

Area addresses might be configured to ensure two areas are not inadvertently connected. Area addresses appear in Hellos and LSPs within the area. If two chunks, connected only via Level 2, were configured with the same area address, this would not cause any problems. (They would just operate as separate Level 1 areas.)

A more serious problem occurs if the Level 2 area is partitioned in such a way that it could be healed by using a path through a Level 1 area. TRILL will not attempt to solve this problem. Within the Level 1 area, a single border RBridge will be the DBRB, and will be in charge of deciding which (single) RBridge will transition any particular multi-destination packets between that area and Level 2. If the Level 2 area is partitioned, this will result in multi-destination data only reaching the portion of the TRILL campus reachable through the partition attached to the TRILL switch that transitions that packet. It will not cause a loop.

#### 4. Multi-Destination Scope

There are at least two reasons it would be desirable to be able to mark a multi-destination packet with a scope that indicates the packet should not exit the area, as follows:

1. To address an issue in the border learning variant of the aggregated nickname alternative, when a unicast packet turns into a multi-destination packet when transitioning from Level 2 to Level 1, as discussed in Section 4.1.
2. To constrain the broadcast domain for certain discovery, directory, or service protocols as discussed in Section 4.2.

Multi-destination packet distribution scope restriction could be done in a number of ways. For example, there could be a flag in the packet that means "for this area only". However, the technique that might require the least change to TRILL switch fast path logic would be to indicate this in the egress nickname that designates the distribution tree being used. There could be two general tree nicknames for each tree, one being for distribution restricted to the area and the other being for multi-area trees. Or there would be a set of N (perhaps 16) special currently reserved nicknames used to specify the N highest priority trees but with the variation that if the special nickname is used for the tree, the packet is not transitioned between areas. Or one or more special trees could be built that were restricted to the local area.

##### 4.1 Unicast to Multi-destination Conversions

In the border learning variant of the aggregated nickname alternative, the following situation may occur:

- a unicast packet might be known at the Level 1 to Level 2 transition and be forwarded as a unicast packet to the least cost border TRILL switch advertising connectivity to the destination area, but
- upon arriving at the border TRILL switch, it turns out to have an unknown destination { MAC, Data Label } pair.

In this case, the packet must be converted into a multi-destination packet and flooded in the destination area. However, if the border TRILL switch doing the conversion is not the border TRILL switch designated to transition the resulting multi-destination packet, there is the danger that the designated transitioner may pick up the packet and flood it back into Level 2 from which it may be flooded into multiple areas. This danger can be avoided by restricting any multi-destination packet that results from such a conversion to the destination area as described above.

Alternatively, a multi-destination packet intended only for the area could be tunneled (within the area) to the RBridge RBx, that is the appointed transitioner for that form of packet (say, based on VLAN or FGL), with instructions that RBx only transmit the packet within the area, and RBx could initiate the multi-destination packet within the area. Since RBx introduced the packet, and is the only one allowed to transition that packet to Level 2, this would accomplish scoping of the packet to within the area. Since this case only occurs in the unusual case when unicast packets need to be turned into multi-destination as described above, the suboptimality of tunneling between the border TRILL switch that receives the unicast packet and the appointed level transitioner for that packet, might not be an issue.

#### 4.1.1 New Tree Encoding

The current encoding, in a TRILL header, of a tree, is of the nickname of the tree root. This requires all 16 bits of the egress nickname field. TRILL could instead, for example, use the bottom 6 bits to encode the tree number (allowing 64 trees), leaving 10 bits to encode information such as:

- o scope: a flag indicating whether it should be single area only, or entire campus
- o border injector: an indicator of which of the k border TRILL switches injected this packet

If TRILL were to adopt this new encoding, any of the TRILL switches in an edge group could inject a multi-destination packet. This would require all TRILL switches to be changed to understand the new encoding for a tree, and it would require a TLV in the LSP to indicate which number each of the TRILL switches in an edge group would be.

While there are a number of advantages to this technique, it requires fast path logic changes and thus its deployment is not practical at this time. It is included here for completeness.

#### 4.2 Selective Broadcast Domain Reduction

There are a number of service, discovery, and directory protocols that, for convenience, are accessed via multicast or broadcast frames. Examples are DHCP, (Dynamic Host Configuration Protocol) the NetBIOS Service Location Protocol, and multicast DNS (Domain Name Service).

Some such protocols provide means to restrict distribution to an IP subnet or equivalent to reduce size of the broadcast domain they are using and then provide a proxy that can be placed in that subnet to use unicast to access a service elsewhere. In cases where a proxy mechanism is not currently defined, it may be possible to create one that references a central server or cache. With multilevel TRILL, it is possible to construct very large IP subnets that could become saturated with multi-destination traffic of this type unless packets can be further restricted in their distribution. Such restricted distribution can be accomplished for some protocols, say protocol P, in a variety of ways including the following:

- Either (1) at all ingress TRILL switches in an area place all protocol P multi-destination packets on a distribution tree in such a way that the packets are restricted to the area or (2) at all border TRILL switches between that area and Level 2, detect protocol P multi-destination packets and do not transition them.
- Then place one, or a few for redundancy, protocol P proxies inside each area where protocol P may be in use. These proxies unicast protocol P requests or other messages to the actual campus server(s) for P. They also receive unicast responses or other messages from those servers and deliver them within the area via unicast, multicast, or broadcast as appropriate. (Such proxies would not be needed if it was acceptable for all protocol P traffic to be restricted to an area.)

While it might seem logical to connect the campus servers to TRILL switches in Level 2, they could be placed within one or more areas so that, in some cases, those areas might not require a local proxy server.

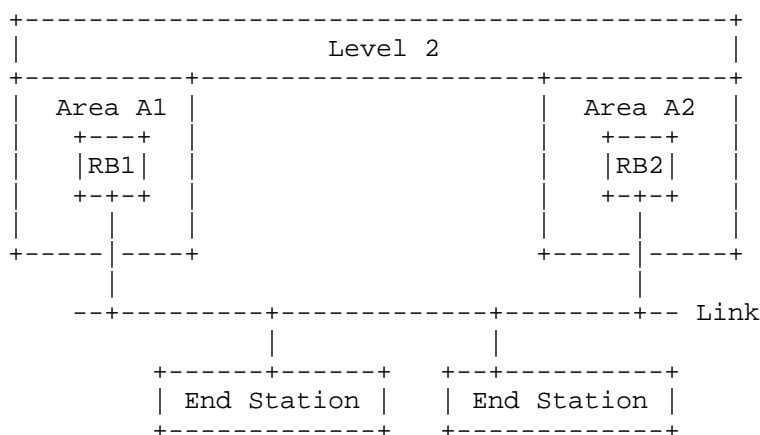
## 5. Co-Existence with Old TRILL switches

TRILL switches that are not multilevel aware may have a problem with calculating RPF Check and filtering information, since they would not be aware of the assignment of border TRILL switch transitioning.

A possible solution, as long as any old TRILL switches exist within an area, is to have the border TRILL switches elect a single DBRB (Designated Border RBridge), and have all inter-area traffic go through the DBRB (unicast as well as multi-destination). If that DBRB goes down, a new one will be elected, but at any one time, all inter-area traffic (unicast as well as multi-destination) would go through that one DBRB. However this eliminates load splitting at level transition.

Care must be taken in the case where there are multiple TRILL switches on a link with one or more end stations, keeping in mind that end stations are TRILL ignorant. In particular, it is essential that only one TRILL switch ingress/egress any given data packet from/to an end station so that connectivity is provided to that end station without duplicating end station data and that loops are not formed due to one TRILL switch egressing data in native form (i.e., with no TRILL header) and having that data re-ingressed by another TRILL switch on the link.

The problem is not avoiding adjacency or avoiding TRILL Data packet transfer between RB1 and RB2. The area address mechanism of IS-IS or possibly the use of topology constraints or the like does that quite well. The problem stems from end stations being TRILL ignorant so care must be taken that multiple RBridges on a link do not ingress the same frame originated by an end station and so that an RBridge does not ingress a native frame egressed by a different RBridge because the RBridge mistakes the frame for a frame originated by an end station.



A simple rule, which is preferred, is to use the TRILL switch or switches having the lowest numbered area, comparing area numbers as unsigned integers, to handle all native traffic to/from end stations on the link. This would automatically give multilevel-ignorant legacy TRILL switches, that would be using area number zero, highest priority for handling end station traffic, which they would try to do anyway.

Other methods are possible. For example doing the selection of Appointed Forwarders and of the TRILL switch in charge of that selection across all TRILL switches on the link regardless of area. However, a special case would then have to be made for legacy TRILL switches using area number zero.

These techniques require multilevel aware TRILL switches to take actions based on Hellos from RBridges in other areas even though they will not form an adjacency with such RBridges. However, the action is quite simple in the preferred case: if a TRILL switch sees Hellos from lower numbered areas, then they would not act as an Appointed Forwarder on the link until the Hello timer for such Hellos had expired.

## 7. Summary

This draft describes potential scaling issues in TRILL and discusses possible approaches to multilevel TRILL as a solution or element of a solution to most of them.

The alternative using aggregated areas in multilevel TRILL has significant advantages in terms of scalability over using campus wide unique nicknames, not just in avoiding nickname exhaustion, but by allowing RPF Checks to be aggregated based on an entire area. However, the alternative of using unique nicknames is simpler and avoids the changes in border TRILL switches required to support aggregated nicknames. It is possible to support both. For example, a TRILL campus could use simpler unique nicknames until scaling begins to cause problems and then start to introduce areas with aggregated nicknames.

Some multilevel TRILL issues are not difficult, such as dealing with partitioned areas. Other issues are more difficult, especially dealing with old TRILL switches that are multilevel ignorant.



## 8. Security Considerations

This informational document explores alternatives for the design of multilevel IS-IS in TRILL and generally does not consider security issues.

If aggregated nicknames are used in two areas that have the same area address and those areas merge, there is a possibility of a transient nickname collision that would not occur with unique nicknames. Such a collision could cause a data packet to be delivered to the wrong egress TRILL switch but it would still not be delivered to any end station in the wrong Data Label; thus such delivery would still conform to security policies.

For general TRILL Security Considerations, see [RFC6325].

## 9. IANA Considerations

This document requires no IANA actions. RFC Editor: Please remove this section before publication.

## Normative References

- [IS-IS] - ISO/IEC 10589:2002, Second Edition, "Intermediate System to Intermediate System Intra-Domain Routing Exchange Protocol for use in Conjunction with the Protocol for Providing the Connectionless-mode Network Service (ISO 8473)", 2002.
- [RFC6325] - Perlman, R., Eastlake 3rd, D., Dutt, D., Gai, S., and A. Ghanwani, "Routing Bridges (RBriges): Base Protocol Specification", RFC 6325, July 2011.
- [RFC7177] - Eastlake 3rd, D., Perlman, R., Ghanwani, A., Yang, H., and V. Manral, "Transparent Interconnection of Lots of Links (TRILL): Adjacency", RFC 7177, May 2014, <<http://www.rfc-editor.org/info/rfc7177>>.
- [RFC7780] - Eastlake 3rd, D., Zhang, M., Perlman, R., Banerjee, A., Ghanwani, A., and S. Gupta, "Transparent Interconnection of Lots of Links (TRILL): Clarifications, Corrections, and Updates", RFC 7780, DOI 10.17487/RFC7780, February 2016, <<http://www.rfc-editor.org/info/rfc7780>>.
- [RFC8139] - Eastlake, D., Li, Y., Umair, M., Banerjee, A., and F. Hu, "Transparent Interconnection of Lots of Links (TRILL): Appointed Forwarders", RFC 8139, DOI 10.17487/RFC8139, June 2017, <<http://www.rfc-editor.org/info/rfc8139>>.

## Informative References

- [InterCon] - Perlman, R., "Interconnections, Second Edition; Bridges, Routers, Switches, and Internetworking Protocols", Addison Wesley, ISBN 0-201-63448-1, September 1999.
- [RFC3194] - Durand, A. and C. Huitema, "The H-Density Ratio for Address Assignment Efficiency An Update on the H ratio", RFC 3194, DOI 10.17487/RFC3194, November 2001, <<http://www.rfc-editor.org/info/rfc3194>>.
- [RFC6361] - Carlson, J. and D. Eastlake 3rd, "PPP Transparent Interconnection of Lots of Links (TRILL) Protocol Control Protocol", RFC 6361, August 2011.
- [RFC7172] - Eastlake 3rd, D., Zhang, M., Agarwal, P., Perlman, R., and D. Dutt, "Transparent Interconnection of Lots of Links (TRILL): Fine-Grained Labeling", RFC 7172, May 2014
- [RFC7176] - Eastlake 3rd, D., Senevirathne, T., Ghanwani, A., Dutt,

D., and A. Banerjee, "Transparent Interconnection of Lots of Links (TRILL) Use of IS-IS", RFC 7176, May 2014.

[RFC7357] - Zhai, H., Hu, F., Perlman, R., Eastlake 3rd, D., and O. Stokes, "Transparent Interconnection of Lots of Links (TRILL): End Station Address Distribution Information (ESADI) Protocol", RFC 7357, September 2014, <<http://www.rfc-editor.org/info/rfc7357>>.

[RFC7781] - Zhai, H., Senevirathne, T., Perlman, R., Zhang, M., and Y. Li, "Transparent Interconnection of Lots of Links (TRILL): Pseudo-Nickname for Active-Active Access", RFC 7781, DOI 10.17487/RFC7781, February 2016, <<http://www.rfc-editor.org/info/rfc7781>>.

[RFC7783] - Senevirathne, T., Pathangi, J., and J. Hudson, "Coordinated Multicast Trees (CMT) for Transparent Interconnection of Lots of Links (TRILL)", RFC 7783, DOI 10.17487/RFC7783, February 2016, <<http://www.rfc-editor.org/info/rfc7783>>.

[DraftAggregated] - Bhargav Bhikkaji, Balaji Venkat Venkataswami, Narayana Perumal Swamy, "Connecting Disparate Data Center/PBB/Campus TRILL sites using BGP", draft-balaji-trill-over-ip-multi-level, Work In Progress.

[DraftUnique] - M. Zhang, D. Eastlake, R. Perlman, M. Cullen, H. Zhai, D. Liu, "TRILL Multilevel Using Unique Nicknames", draft-ietf-trill-multilevel-unique-nickname, Work In Progress.

[SingleName] - Mingui Zhang, et. al, "Single Area Border RBridge Nickname for TRILL Multilevel", draft-ietf-trill-multilevel-single-nickname, Work in Progress.

#### Acknowledgements

The helpful comments and contributions of the following are hereby acknowledged:

Alia Atlas, David Michael Bond, Dino Farinacci, Sue Hares, Gayle Noble, Alexander Vainshtein, and Stig Venaas.

The document was prepared in raw nroff. All macros used were defined within the source file.

Authors' Addresses

Radia Perlman  
EMC  
2010 256th Avenue NE, #200  
Bellevue, WA 98007 USA

EMail: [radia@alum.mit.edu](mailto:radia@alum.mit.edu)

Donald Eastlake  
Huawei Technologies  
155 Beaver Street  
Milford, MA 01757 USA

Phone: +1-508-333-2270  
Email: [d3e3e3@gmail.com](mailto:d3e3e3@gmail.com)

Mingui Zhang  
Huawei Technologies  
No.156 Beiqing Rd. Haidian District,  
Beijing 100095 P.R. China

EMail: [zhangmingui@huawei.com](mailto:zhangmingui@huawei.com)

Anoop Ghanwani  
Dell  
5450 Great America Parkway  
Santa Clara, CA 95054 USA

EMail: [anoop@alumni.duke.edu](mailto:anoop@alumni.duke.edu)

Hongjun Zhai  
Jinling Institute of Technology  
99 Hongjing Avenue, Jiangning District  
Nanjing, Jiangsu 211169 China

EMail: [honjun.zhai@tom.com](mailto:honjun.zhai@tom.com)

## Copyright and IPR Provisions

Copyright (c) 2017 IETF Trust and the persons identified as the document authors. All rights reserved.

This document is subject to BCP 78 and the IETF Trust's Legal Provisions Relating to IETF Documents (<http://trustee.ietf.org/license-info>) in effect on the date of publication of this document. Please review these documents carefully, as they describe your rights and restrictions with respect to this document. Code Components extracted from this document must include Simplified BSD License text as described in Section 4.e of the Trust Legal Provisions and are provided without warranty as described in the Simplified BSD License. The definitive version of an IETF Document is that published by, or under the auspices of, the IETF. Versions of IETF Documents that are published by third parties, including those that are translated into other languages, should not be considered to be definitive versions of IETF Documents. The definitive version of these Legal Provisions is that published by, or under the auspices of, the IETF. Versions of these Legal Provisions that are published by third parties, including those that are translated into other languages, should not be considered to be definitive versions of these Legal Provisions. For the avoidance of doubt, each Contributor to the IETF Standards Process licenses each Contribution that he or she makes as part of the IETF Standards Process to the IETF Trust pursuant to the provisions of RFC 5378. No language to the contrary, or terms, conditions or rights that differ from or are inconsistent with the rights and licenses granted under RFC 5378, shall have any effect and shall be null and void, whether published or posted by such Contributor, or included with or in such Contribution.



TRILL WG  
Internet-Draft  
Intended status: Standards Track  
Expires: February 18, 2016

Radia. Perlman  
EMC Corporation  
Fangwei. Hu  
ZTE Corporation  
Donald. Eastlake 3rd  
Huawei technology  
Kesava. Krupakaran  
Dell  
Ting. Liao  
ZTE Corporation  
August 17, 2015

TRILL Smart Endnodes  
draft-ietf-trill-smart-endnodes-02.txt

Abstract

This draft addresses the problem of the size and freshness of the endnode learning table in edge RBridges, by allowing endnodes to volunteer for endnode learning and encapsulation/decapsulation. Such an endnode is known as a "Smart Endnode". Only the attached RBridge can distinguish a "Smart Endnode" from a "normal endnode". The smart endnode uses the nickname of the attached RBridge, so this solution does not consume extra nicknames. The solution also enables Fine Grained Label aware endnodes.

Status of This Memo

This Internet-Draft is submitted in full conformance with the provisions of BCP 78 and BCP 79.

Internet-Drafts are working documents of the Internet Engineering Task Force (IETF). Note that other groups may also distribute working documents as Internet-Drafts. The list of current Internet-Drafts is at <http://datatracker.ietf.org/drafts/current/>.

Internet-Drafts are draft documents valid for a maximum of six months and may be updated, replaced, or obsoleted by other documents at any time. It is inappropriate to use Internet-Drafts as reference material or to cite them other than as "work in progress."

This Internet-Draft will expire on February 18, 2016.



## Copyright Notice

Copyright (c) 2015 IETF Trust and the persons identified as the document authors. All rights reserved.

This document is subject to BCP 78 and the IETF Trust's Legal Provisions Relating to IETF Documents (<http://trustee.ietf.org/license-info>) in effect on the date of publication of this document. Please review these documents carefully, as they describe your rights and restrictions with respect to this document. Code Components extracted from this document must include Simplified BSD License text as described in Section 4.e of the Trust Legal Provisions and are provided without warranty as described in the Simplified BSD License.

## Table of Contents

1. Introduction . . . . .	2
2. Solution Overview . . . . .	3
3. Terminology . . . . .	4
4. Smart-Hello Mechanism between Smart Endnode and RBridge . . .	5
4.1. Smart-Hello Encapsulation . . . . .	5
4.2. Edge RBridge's Smart-Hello . . . . .	7
4.3. Smart Endnode's Smart-Hello . . . . .	7
5. Data Packet Processing . . . . .	8
5.1. Data Packet Processing for Smart Endnode . . . . .	9
5.2. Data Packet Processing for Edge RBridge . . . . .	9
6. Multi-homing Scenario . . . . .	10
7. Security Considerations . . . . .	12
8. IANA Considerations . . . . .	12
9. Acknowledgements . . . . .	12
10. Normative References . . . . .	12
Authors' Addresses . . . . .	14

## 1. Introduction

The IETF TRILL (Transparent Interconnection of Lots of Links) protocol [RFC6325] provides optimal pair-wise data frame forwarding without configuration, safe forwarding even during periods of temporary loops, and support for multipathing of both unicast and multicast traffic. TRILL accomplishes this by using IS-IS [IS-IS] [RFC7176] link state routing and encapsulating traffic using a header that includes a hop count. Devices that implement TRILL are called "RBridges" (Routing Bridges) or "TRILL Switches".

An RBridge that attaches to endnodes is called an "edge RBridge" or "edge TRILL Switch", whereas one that exclusively forwards encapsulated frames is known as a "transit RBridge" or "transit TRILL

Switch". An edge RBridge traditionally is the one that encapsulates a native Ethernet packet with a TRILL header, or that receives a TRILL-encapsulated packet and decapsulates the TRILL header. To encapsulate efficiently, the edge RBridge must keep an "endnode table" consisting of (MAC, Data Label, TRILL egress switch nickname) sets, for those remote MAC addresses in Data Labels currently communicating with endnodes to which the edge RBridge is attached.

These table entries might be configured, received from ESADI [RFC7357], looked up in a directory [RFC7067], or learned from decapsulating received traffic. If the edge RBridge has attached endnodes communicating with many remote endnodes, this table could become large. Also, if one of the MAC addresses and Data Labels in the table has moved to a different remote TRILL switch, it might be difficult for the edge RBridge to notice this quickly, and because the edge RBridge is encapsulating to the incorrect egress RBridge, the traffic will get lost.

## 2. Solution Overview

The Smart Endnode solution proposed in this document addresses the problem of the size and freshness of the endnode learning table in edge RBridges. An endnode E, attached to an edge RBridge R, tells R that E would like to be a "Smart Endnode", which means that E will encapsulate and decapsulate the TRILL frame, using R's nickname. Because E uses R's nickname, this solution does not consume extra nicknames.

Take the below figure as the example Smart Endnode scenario: RB1, RB2 and RB3 are the RBridges in the TRILL domain, and smart SE1 and SE2 are the smart ennodes which can encapsulate and decapsulate the TRILL frames. RB1 is the edge attached RB for SE1 and SE2, and assigns its nickname to SE1 and SE2.

Each Smart Endnode, SE1 and SE2, uses RB1's nickname when encapsulating, and maintains an endnode table of (MAC, label, TRILL egress switch nickname) for remote endnodes that it (SE1 or SE2) is corresponding with. RB1 does not decapsulate packets destined for SE1 or SE2, and does not learn (MAC, label, TRILL egress switch nickname) for endnodes corresponding with SE1 or SE2, but RB1 does decapsulate, and does learn (MAC, label, TRILL egress switch nickname) for any endnodes attached to RB1 that have not declared themselves to be Smart Endnodes.

Just as an RBridge learns and times out (MAC, label, TRILL egress switch nickname), Smart Endnodes SE1 and SE2 also learn and time out endnode entries. However, SE1 and SE2 might also determine, through ICMP messages or other techniques, that an endnode entry is not

successfully reaching the destination endnode, and can be deleted, even if the entry has not timed out.

If SE1 wishes to correspond with destination MAC D, and no endnode entry exists, SE1 will encapsulate the packet as an unknown destination, or examining updates to the ESADI link state database [RFC7357], or consulting a directory [RFC7067] (just as an RBridge would do if there was no endnode entry).

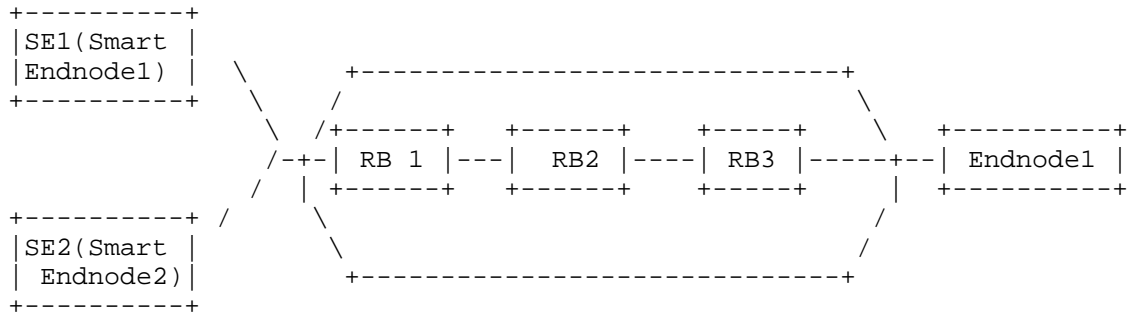


Figure 1 Smart Endnode Scenario

The mechanism in this draft is that the Smart Endnode SE1 issues a Smart-Hello, indicating SE1's desire to act as a Smart Endnode, together with the set of MAC addresses and Data Labels that SE1 owns, and whether SE1 would like to receive ESADI packets. The Smart-Hello is a light type of TRILL-hello, which is used to announce the Smart Endnode capability and parameters (such as MAC address, VLAN ID etc.). The detailed content for a smart endnode's Smart-Hello is defined in section 4.

If RB1 supports having a Smart Endnode neighbor it also sends Smart-Hellos. The smart endnode learns from RB1's Smart-Hellos what RB1's nickname is and which trees RB1 can use when RB1 ingresses multi-destination frames. Although Smart Endnode SE1 transmits Smart-Hellos, it does not transmit or receive LSPs or E-L1FS FS-LSPs[I-D.ietf-trill-rfc7180bis].

Since a Smart Endnode can encapsulate TRILL Data frames, it can cause the Inner.Lable to be a Fine Grained Label [RFC7172], thus this method supports FGL aware endnodes.

### 3. Terminology

Edge RBridge: An RBridge providing endnode service on at least one of its ports. It is also called an edge TRILL Switch.

Data Label: VLAN or FGL.

ESADI: End Station Address Distribution Information [RFC7357].

FGL: Fine Grained Label [RFC7172].

IS-IS: Intermediate System to Intermediate System [IS-IS].

RBridge: Routing Bridge, an alternative name for a TRILL switch.

Smart Endnode: An endnode that has the capability specified in this document including learning and maintaining(MAC, Data Label, Nickname) entries and encapsulating/decapsulating TRILL frame.

Transit RBridge: An RBridge exclusively forwards encapsulated frames. It is also named as transit RBridge.

TRILL: Transparent Interconnection of Lots of Links [RFC6325].

TRILL switch: a device that implements the TRILL protocol; an alternative term for an RBridge.

#### 4. Smart-Hello Mechanism between Smart Endnode and RBridge

The subsections below describe Smart-Hello messages.

##### 4.1. Smart-Hello Encapsulation

Although a Smart Endnode is not an RBridge, does not send LSPs, and does not perform routing calculations, it is required to have a "Hello" mechanism (1) to announce to edge RBridges that it is a Smart Endnode and (2) to tell them what MAC addresses it is handling in what Data Labels. Similarly, an edge RBridge that supports Smart Endnodes needs a message (1) to announce that support, (2) to inform Smart Endnodes what nickname to use for ingress and what nickname(s) can be used as multi-destination TRILL data packet, and (3) the list of smart end nodes it knows about on that link.

The messages sent by Smart Endnodes and by edge RBridges that support Smart Endnodes are called "Smart-Hellos" and are carried through native RBridge channel messages (see Section 4 of [RFC7178]). They are structured as follows:

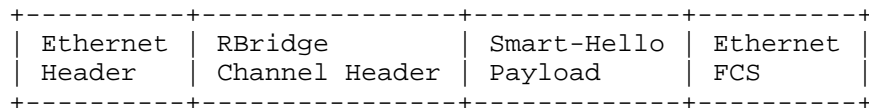


Figure 2 Smart-Hello Structure

In the Ethernet Header, the source MAC address is the address of the Smart Endnode or edge RBridge port on which the message is sent. If the Smart-Hello is sent by a Smart Endnode and multicasted in the link, the destination MAC address is All-Edge-RBridges, and if the Smart-Hello is unicasted to an edge RBridge, the destination MAC address is the MAC address of the RBridge. If the Smart-Hello is sent by an Edge RBridge and multicasted in the link, the destination MAC address is TRILL-End-Stations, and if it is unicasted to a Smart Endnode, the MAC address is the MAC address of the Smart Endnode. The frame is sent in the Designated VLAN of the link so if a VLAN tag is present, it specifies that VLAN.

The RBridge Channel Header begins with the RBridge Channel Ethertype. In the RBridge Channel Header, the Channel Protocol number is as assigned by IANA (see Section 8) and in the flags field, the NA bit is one, the MH bit is zero and the setting of the SL bit is an implementation choice.

The Smart-Hello Payload, both for Smart-Hellos sent by Smart Endnodes and for Smart-Hellos sent by Edge RBridges, consists of TRILL IS-IS TLVs as described in the following two sub-sections. The non-extended format is used so TLVs, sub-TLVs, and APPsub-TLVs have an 8-bit size and type field. Both types of Smart-Hellos MUST include a Smart-Parameters APPsub-TLV as follows inside a TRILL GENINFO TLV:

```

+-----+
|Smart-Parameters|                               (1 byte)
+-----+
|   Length       |                               (1 byte)
+-----+-----+
| Holding Time   |                               (2 bytes)
+-----+-----+
|   Flags       |                               (2 bytes)
+-----+-----+

```

Figure 3 Smart Parameters APPsub-TLV

Type: APPsub-TLV type Smart-Parameters, value is TBD.

Length: 4.

Holding Time: A time in seconds as an unsigned integer. Has the same meaning as the Holding Time field in IS-IS Hellos [ISIS]. A Smart Endnode and an Edge RBridge supporting Smart Endndoes MUST send a Smart-Hello at least three times during their Holding Time. If no Smart-Hellos is received from a Smart Endnode or Edge RBridge within the most recent Holding Time it sent, it is assumed that it is no longer available.

Flags: At this time all of the Flags are reserved and MUST be send as zero and ignored on receipt.

If more than one Smart Parameters APPsub-TLV appears in a Smart-Hello, the first one is used and any following ones are ignored. If no Smart Parameters APPsub-TLV appears in a Smart-Hello, that Smart-Hello is ignored.

#### 4.2. Edge RBridge's Smart-Hello

The edge RBridge's Smart-Hello contains the following information in addition to the Smart-Parameters APPsub-TLV:

- o RBridge's nickname. The nickname sub-TLV (Specified in section 2.3.2 in [RFC7176]) is reused here carried inside a TLV 242 (IS-IS router capability) in a Smart-Hello frame. If more than one nickname appears in the Smart-Hello, the first one is used and the following ones are ignored.
- o Trees that RBl can use when ingressing multi-destination frames. The Tree Identifiers Sub-TLV (Specified in section 2.3.4 in [RFC7176]) is reused here.
- o Smart Endnode neighbor list. The TRILL Neighbor TLV (Specified in section 2.5 in [RFC7176]) is reused for this purpose.
- o An Autentication TLV MAY also be included.

#### 4.3. Smart Endnode's Smart-Hello

A new APPsub-TLV (Smart-MAC TLV) is defined for use by Smart Endnodes as defined below. In addition, there will be a Smart-Parameters APPsub-TLV and there MAY be an Authentication TLV in a Smart Endnode Smart-Hello.

If there are several VLANs/FGL Data Labels for that Smart Endnode, the Smart-MAC APPsub-TLV is included several times in Smart Endnode's Smart-Hello. This APPsub-TLV appears inside a TRILL GENINFO TLV.

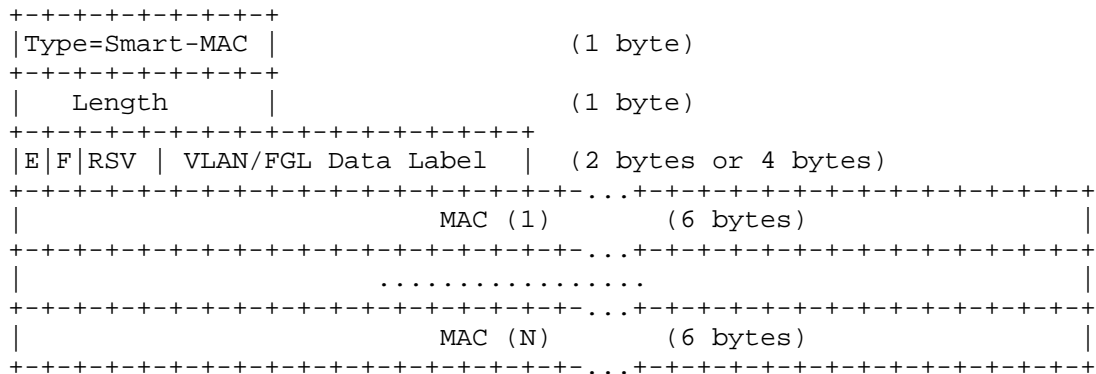


Figure 4 Smart-MAC TLV

- o Type: TRILL APPsub-TLV Type Smart-MAC, value is TBD.
- o Length: Total number of bytes contained in the value field.
- o E: one bit. If it sets to 1, which indicates that the endnode should receive ESADI frames.
- o F: one bit. If it sets to 1, which indicates that the endnode supports FGL data label, otherwise, the VLAN/FGL Data Label [RFC7172] field is the VLAN ID.
- o RSV: 2 bits or 6 bits, is reserved for the future use. If VLAN/FGL Data Label indicates the VLAN ID(or F flag sets to 0), the RESV field is 2 bits length, otherwise it is 6 bits.
- o VLAN/FGL Data Label: This carries a 12-bits VLAN identifier or 24-bits FGL Data Label that is valid for all subsequent MAC addresses in this TLV, or the value zero if no VLAN/FGL data label is specified.
- o MAC(i): This is the 48-bit MAC address reachable in the Data Label given from the IS that is announcing this TLV.

## 5. Data Packet Processing

The subsections below specify Smart Endnode data packet processing. All TRILL data packets sent to or from Smart Endnodes are sent in the Designated VLAN [RFC6325] of the local link but do not necessarily have to be VLAN tagged.

### 5.1. Data Packet Processing for Smart Endnode

A Smart Endnode does not issue or receive LSPs or E-L1FS FS-LSPs or calculate topology. It does the following:

- o Smart Endnode maintains an endnode table of (the MAC address of remote endnode, Data Label, the nickname of the edge RBridge's attached) entries of end nodes with which the Smart Endnode is communicating. Entries in this table are populated the same way that an edge RBridge populates the entries in its table:
  - \* learning from (source MAC address ingress nickname) on packets it decapsulates.
  - \* from ESADI[RFC7357].
  - \* by querying a directory [RFC7067].
  - \* by having some entries configured.
- o When Smart Endnode SE1 wishes to transmit to unicast destination remote node D, if (address of remote endnode D, nickname)entry is in SE1's endnode table, SE1 encapsulates with ingress nickname=the nicknae of the RBridge(RB1), egress nickname as indicated in D's table entry. If D is unknown, D either queries a directory or encapsulates the packet as a multi-destination frame, using one of the trees that RB1 has specified in RB1's Smart-Hello.
- o When SE1 wishes to transmit to a multicast or broadcast destination, SE1 encapsulates the packet using one of the trees that RB1 has specified.

The Smart Endnode SE1 need not send Smart-Hellos as frequently as normal RBridges. These Smart-Hellos could be periodically unicast to the Appointed Forwarder RB1 through native RBridge channel messages. In case RB1 crashes and restarts, or the DRB changes and SE1 receives the Smart-Hello without mentioning SE1, SE1 SHOULD send a Smart-Hello immediately. If RB1 is AF for any of the VLANs that SE1 claims, RB1 MUST list SE1 in its Smart-Hellos as a Smart Endnode neighbor.

### 5.2. Data Packet Processing for Edge RBridge

The attached edge RBridge processes and forwards the data frame based on the endnode property rather than for encapsulates and forwards the native frame as the traditional RBridges. There are several situations for the edge RBridges:



- o If receiving an encapsulated unicast data frame from a port with a smart endnode, with RB1's nickname as ingress, the edge RBridge RB1 forwards the frame to the specified egress nickname, as with any encapsulated frame. However, RB1 MAY filter the encapsulation frame based on the inner source MAC and Data Label as specified for the Smart Endnode. If the MAC (or Data Label) are not among the expected entries of the Smart Endnode, the frame would be dropped by the edge RBridge.
- o If receiving an multi-destination TRILL Data packet from a port with a Smart Endnode, RBridge RB1 forwards the TRILL encapsulation to the TRILL campus based on the distribution tree. If there are some normal endnodes (i.e, non-Smart Endnode) attached to the edge RBridge RB1, RB1 decapsulates the frame and sends the native frame to these ports possibly pruned based on multicast listeners, in addition to forwarding the multi-destination TRILL frame to the rest of the campus.
- o When RB1 receives a multicast frame from a remote RBridge, and the exit port includes hybrid endnodes(Smart Endnodes and non-Smart Endnodes), it sends two copies of mulicast frames, one as native and the other as TRILL encapsulated frame. When Smart Endnode receives the encapsulated frame, it learns the remote (MAC address, Data Label, Nickname) entry, A Smart Endnodes ignores native data frames. A normal (non-smart) endnode receives the native frame and learns the remote MAC address and ignores the TRILL data packet. This transit solution may bring some complexity for the edge RBridge and waste network bandwidth resource, so avoiding the hybrid endnodes scenario by attaching the Smart Endnodes and non-Smart Endnodes to different ports is RECOMMENDED. Another solution is that if there are one or more endnodes on a link, the non-Smart Endnodes are ignored on a link; but we can configure a port to support mixed links. If RB1 is configured that the link is "Smart Endnode only", then it will only send and receive TRILL-encapsulated frames on that link. If it is configured to "non-smart-endnodes only" on a port, it will only send and receive native frames from that port.

## 6. Multi-homing Scenario

Multi-homing is a common scenario for the Smart Endnode. The Smart Endnode is on a link attached to the TRILL domain in two places: to edge RBridge RB1 and RB2. Take the figure below as example. The Smart Endnode SE1 is attached to the TRILL domain by RB1 and RB2 separately. Both RB1 and RB2 could assign their nicknames to SE1.

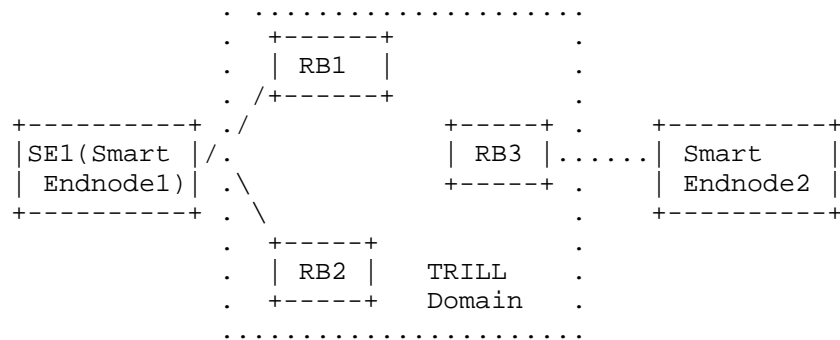


Figure 5 Multi-homing Scenario

There are several solutions for this scenario:

- (1) Smart Endnode SE1 can choose either RB1 or RB2's nickname, when encapsulating a frame, whether the encapsulated frame is sent via RB1 or RB2. If SE1 uses RB1's nickname, in this scenario, SE1 will encapsulate with TRILL source nickname RB1 when transmitting on either port. This is simple, but means that all return traffic will be via RB1. If Smart Endnode SE1 wants to do active-active load splitting, and uses RB1's nickname when forwarding through RB1, and RB2's nickname when forwarding through RB2, this will cause MAC flip-flopping of the endnode table entry in the remote R Bridges (or Smart Endnodes). One solution is to set a multi-homing bit in the RSV field of the TRILL data packet. When remote R Bridge RB3 or Smart Endnodes receives a data packet with the multi-homed bit set, the endnode entries (SE1's MAC addresslabel, RB1's nickname) and (SE1's MAC address, label, RB2's nickname) will coexist as endnode entries in the remote R Bridge. Another solution is to extend the ESADI protocol to distribute multiple attachments of a MAC address of a multi-homing group. (Please refer to the option B in section 4 of [I-D.ietf-trill-aa-multi-attach] for details).
- (2) RB1 and RB2 might indicate, in their Smart-Hellos, a virtual nickname that attached end nodes may use if they are multihomed to RB1 and RB2, separate from RB1 and RB2's nicknames (which they would also list in their Smart-Hellos). This would be useful if there were many end nodes multihomed to the same set of R Bridges. This would be analogous to a pseudonode nickname; return traffic would go via the shortest path from the source to the endnode, whether it is RB1 or RB2. If Smart Endnode SE1 loses connectivity to RB2, then SE1 would revert to using RB1's nickname. In order to avoid RPF check issue for multi-

destination frame, the affinity TLV [I-D.ietf-trill-cmt] is recommended to be used in this solution.

## 7. Security Considerations

Smart-Hellos can be secured by using Authentication TLVs based on [RFC5310].

For general TRILL Security Considerations, see [RFC6325].

For native RBridge channel Security Considerations, see [RFC7178].

## 8. IANA Considerations

IANA is requested to allocate an RBridge Channel Protocol number (0x005) to indicate a smart-hello frame.

IANA is requested to allocate APPsub-TLV type numbers for the Smart-MAC and Smart-Parameters APPsub-TLVs.

## 9. Acknowledgements

The contributions of the following persons are gratefully acknowledged: Mingui Zhang, Weiguo Hao, Linda Dunbar and Andrew Qu.

## 10. Normative References

- [I-D.ietf-trill-aa-multi-attach]  
Zhang, M., Perlman, R., Zhai, H., Durrani, M., and S. Gupta, "TRILL Active-Active Edge Using Multiple MAC Attachments", draft-ietf-trill-aa-multi-attach-04 (work in progress), August 2015.
- [I-D.ietf-trill-cmt]  
Senevirathne, T., Pathangi, J., and J. Hudson, "Coordinated Multicast Trees (CMT) for TRILL", draft-ietf-trill-cmt-06 (work in progress), March 2015.
- [I-D.ietf-trill-rfc7180bis]  
Eastlake, D., Zhang, M., Perlman, R., Banerjee, A., Ghanwani, A., and S. Gupta, "TRILL: Clarifications, Corrections, and Updates", draft-ietf-trill-rfc7180bis-05 (work in progress), June 2015.

- [IS-IS] ISO/IEC 10589:2002, Second Edition,, "Intermediate System to Intermediate System Intra-Domain Routing Exchange Protocol for use in Conjunction with the Protocol for Providing the Connectionless-mode Network Service (ISO 8473)", 2002.
- [RFC5310] Bhatia, M., Manral, V., Li, T., Atkinson, R., White, R., and M. Fanto, "IS-IS Generic Cryptographic Authentication", RFC 5310, DOI 10.17487/RFC5310, February 2009, <<http://www.rfc-editor.org/info/rfc5310>>.
- [RFC6325] Perlman, R., Eastlake 3rd, D., Dutt, D., Gai, S., and A. Ghanwani, "Routing Bridges (RBridges): Base Protocol Specification", RFC 6325, DOI 10.17487/RFC6325, July 2011, <<http://www.rfc-editor.org/info/rfc6325>>.
- [RFC7067] Dunbar, L., Eastlake 3rd, D., Perlman, R., and I. Gashinsky, "Directory Assistance Problem and High-Level Design Proposal", RFC 7067, DOI 10.17487/RFC7067, November 2013, <<http://www.rfc-editor.org/info/rfc7067>>.
- [RFC7172] Eastlake 3rd, D., Zhang, M., Agarwal, P., Perlman, R., and D. Dutt, "Transparent Interconnection of Lots of Links (TRILL): Fine-Grained Labeling", RFC 7172, DOI 10.17487/RFC7172, May 2014, <<http://www.rfc-editor.org/info/rfc7172>>.
- [RFC7176] Eastlake 3rd, D., Senevirathne, T., Ghanwani, A., Dutt, D., and A. Banerjee, "Transparent Interconnection of Lots of Links (TRILL) Use of IS-IS", RFC 7176, DOI 10.17487/RFC7176, May 2014, <<http://www.rfc-editor.org/info/rfc7176>>.
- [RFC7178] Eastlake 3rd, D., Manral, V., Li, Y., Aldrin, S., and D. Ward, "Transparent Interconnection of Lots of Links (TRILL): RBridge Channel Support", RFC 7178, DOI 10.17487/RFC7178, May 2014, <<http://www.rfc-editor.org/info/rfc7178>>.
- [RFC7357] Zhai, H., Hu, F., Perlman, R., Eastlake 3rd, D., and O. Stokes, "Transparent Interconnection of Lots of Links (TRILL): End Station Address Distribution Information (ESADI) Protocol", RFC 7357, DOI 10.17487/RFC7357, September 2014, <<http://www.rfc-editor.org/info/rfc7357>>.

## Authors' Addresses

Radia Perlman  
EMC Corporation  
2010 156th Ave NE, suite #200  
Bellevue, WA 98007  
USA

Phone: +1-206-291-367  
Email: radiaperlman@gmail.com

Fangwei Hu  
ZTE Corporation  
No.889 Bibo Rd  
Shanghai 201203  
China

Phone: +86 21 68896273  
Email: hu.fangwei@zte.com.cn

Donald Eastlake, 3rd  
Huawei technology  
155 Beaver Street  
Milford, MA 01757  
USA

Phone: +1-508-634-2066  
Email: d3e3e3@gmail.com

Kesava Vijaya Krupakaran  
Dell  
Olympia Technology Park  
Guindy Chennai 600 032  
India

Phone: +91 44 4220 8496  
Email: Kesava\_Vijaya\_Krupak@Dell.com

Ting Liao  
ZTE Corporation  
No.50 Ruanjian Ave.  
Nanjing, Jiangsu 210012  
China

Phone: +86 25 88014227  
Email: liao.ting@zte.com.cn

INTERNET-DRAFT  
Intended Status: Proposed Standard

Mohammed Umair  
Kingston Smiler S  
Shaji Ravindranathan  
IP Infusion  
Lucy Yong  
Donald Eastlake 3rd  
Huawei Technologies  
November 02, 2015

Expires: May 05, 2016

Date Center Interconnect using TRILL  
<draft-muks-trill-dci-00.txt>

## Abstract

This document describes a TRILL based DCI solution using VTSD. VTSD (Virtual TRILL Service/Switch Domain) is specified in [draft-VTSD]. This draft describes the advantages provided by a TRILL based DCI solution over an existing MPLS L2VPN solution, advantages such as bandwidth scaling and providing multiple active pseudowires.

## Status of this Memo

This Internet-Draft is submitted to IETF in full conformance with the provisions of BCP 78 and BCP 79.

Internet-Drafts are working documents of the Internet Engineering Task Force (IETF), its areas, and its working groups. Note that other groups may also distribute working documents as Internet-Drafts.

Internet-Drafts are draft documents valid for a maximum of six months and may be updated, replaced, or obsoleted by other documents at any time. It is inappropriate to use Internet-Drafts as reference material or to cite them other than as "work in progress."

The list of current Internet-Drafts can be accessed at  
<http://www.ietf.org/lid-abstracts.html>

The list of Internet-Draft Shadow Directories can be accessed at  
<http://www.ietf.org/shadow.html>

## Copyright and License Notice

Copyright (c) 2015 IETF Trust and the persons identified as the document authors. All rights reserved.

This document is subject to BCP 78 and the IETF Trust's Legal Provisions Relating to IETF Documents (<http://trustee.ietf.org/license-info>) in effect on the date of publication of this document. Please review these documents carefully, as they describe your rights and restrictions with respect to this document. Code Components extracted from this document must include Simplified BSD License text as described in Section 4.e of the Trust Legal Provisions and are provided without warranty as described in the Simplified BSD License.

## Table of Contents

1	Introduction . . . . .	3
1.1	Terminology . . . . .	4
2.	Date Center Topology . . . . .	6
2.	Appointed Forwarders . . . . .	7
3.	Multiple Parallel pseudowires. . . . .	8
4.	Active-Active Pseudowire . . . . .	9
4.1.	Port-based AC operations. . . . .	10
4.2.	VLAN-based AC operations. . . . .	10
5.	MPLS encapsulation and Loop free provider PSN/MPLS . . . . .	10
6.	Frame processing . . . . .	10
6.1.	Frame processing between data center T2 switch and TIR. . . . .	10
6.2.	Frame processing between TIR's . . . . .	11
7.	MAC Address learning and withdrawal . . . . .	12
8.	Active-Active Access with VTSD . . . . .	12
9.	ARP/ND proxy . . . . .	12
10.	MAC mass-withdrawal . . . . .	12
11.	Security Considerations . . . . .	13
12.	IANA Considerations . . . . .	13
13.	References . . . . .	13
13.1.	Normative References . . . . .	13
10.2.	Informative References . . . . .	13
	Authors' Addresses . . . . .	13



## 1 Introduction

Pseudo Wire Emulation Edge-to-Edge (PWE3) is a mechanism that emulates the essential attributes of a service such as Ethernet over a Packet Switched Network (PSN). The required functions of PWs include encapsulating service-specific PDUs arriving at an ingress port, and carrying them across a path or tunnel, managing their timing and order, and any other operations required to emulate the behavior and characteristics of the service as faithfully as possible.

The IETF Transparent Interconnection of Lots of Links (TRILL) protocol [RFC6325] [RFC7177] [rfc7180bis] provides transparent forwarding in multi-hop networks with arbitrary topology and link technologies using a header with a hop count and link-state routing. TRILL provides optimal pair-wise forwarding without configuration, safe forwarding even during periods of temporary loops, and support for multipathing of both unicast and multicast traffic. Intermediate Systems (ISs) implementing TRILL are called Routing Bridges(RBridges)or TRILL Switches.

The [draft-VTSD] introduces a new terminology called VTSD. VTSD is a logical RBridge resides inside TIR (TRILL Intermediate Router) that should be capable of performing all the operations that a standard TRILL switch can do, along with IP and MPLS functions. A TIR is a Provider Edge (PE) device where VTSD resides and provides TRILL DCI solution. VTSD is connected to the Layer2 interface towards the DC and PW interface towards the MPLS core

TRILL as a protocol enables optimal use of the links in a layer2 network and running TRILL inside the TIR or VTSD provides a way for optimally utilizing the following:

1. The PWE3 mesh connectivity in the MPLS core using parallel pseudowires.
2. The PWE3 attachment circuit interface, when there are more than one attachment circuit interfaces using active-active pseudowires.
3. Providing a RING based DCI solution along with traditional mesh / hub-spoke topology.
4. Optimally re-route the traffic from one pseudowire to another pseudowire when there is a failure. This is possible as VTSD doesn't follow split-horizon for loop free topology.

When there is a requirement to increase the bandwidth of a particular

DCI link, with TRILL DCI, new pseudowires could be created with the same endpoints. These pseudowires are termed as parallel pseudowires. As these pseudowires are attached to VTSD (which is a TRILL RBridge), the TRILL protocol takes care of optimally load sharing the traffic across these parallel pseudowires.

Similarly when there is a requirement to increase the bandwidth of customer facing interface (attachment circuit), this can be achieved effectively by adding new attachment circuit interfaces and attaching them to the same VTSD.

The objective of a pseudowire (PW) connected in parallel or mesh or ring is to maintain connectivity across the packet switched network (PSN) used by the emulated service. In this model all pseudowires that are part of a service domain will carry data traffic without making any of the pseudowire go in to standby mode.

### 1.1 Terminology

The key words "MUST", "MUST NOT", "REQUIRED", "SHALL", "SHALL NOT", "SHOULD", "SHOULD NOT", "RECOMMENDED", "MAY", and "OPTIONAL" in this document are to be interpreted as described in RFC 2119 [RFC2119].

Acronyms used in this document include the following:

AC	- Attachment Circuit [RFC4664]
Access Port	- A TRILL switch port configured with the "end station service enable" bit on, as described in Section 4.9.1 of [RFC6325]. All AC's, VTSD ports connected to CE's, should configured as TRILL Access port.
AF	- Appointed Forwarder [RFC6325], [RFC6439] and [RFC6439bis].
Data Label	- VLAN or FGL
ECMP	- Equal Cost Multi Pathing
FGL	- Fine-Grained Labeling [RFC7172]
IS-IS	- Intermediate System to Intermediate System [IS-IS]

LAN	- Local Area Network
Link	- The means by which adjacent TRILL switches or VTSD is connected. May be a bridged LAN
MLAG	- Multi-Chassis Link Aggregation
MPLS	- Multi-Protocol Label Switching
PE	- Provider Edge Device
PSN	- Packet Switched Network
PW	- Pseudowire [RFC4664]
RBridge	- An alternative name for TRILL Switch
TIR	- TRILL Intermediate Router (Devices where Pseudowire starts and Terminates)
TRILL	- Transparent Interconnection of Lots of Links OR Tunneled Routing in the Link Layer
TRILL Site	- A part of a TRILL campus that contains at least one RBridge.
TRILL switch	- A device implementing the TRILL protocol. An alternative name for an RBridge.
Trunk port	- A TRILL switch port configured with the "end station service disable" bit on, as described in Section 4.9.1 of [RFC6325]. All pseudowires should be configured as TRILL Trunk port.
VLAN	- Virtual Local Area Network
VPLS	- Virtual Private LAN Service
VPTS	- Virtual Private TRILL Service
VSI	- Virtual Service Instance [RFC4664]
VTSI	- Virtual TRILL Service Instance

- VTSD                      - Virtual TRILL Switch Domain OR  
                              Virtual TRILL Service Domain  
                              A Virtual RBridge that segregates  
                              one tenant's TRILL database as well  
                              as traffic from the other.
- VTSD-AP                  - A VTSD TRILL Access port can be a  
                              AC or a logical port connected with  
                              CE's. it can be a combination of  
                              physical port and Data Label.  
                              OR just Physical port connected to  
                              CE's

## 2. Data Center Topology

The reference topology that will be used for our discussion is a 3 tier traditional topology. Although other topologies may be utilized within the data center, most of such L2 based data centers may be modeled as a 3 tier traditional topology. The reference topology is illustrated in Figure 1. To keep terminologies simple and uniform, in this document these layers will be referred to as Tier-1, Tier-2 and Tier-3 "tiers", and the switches in these layers will be termed as T1SW, T2SW etc. For simplicity reasons, the entire DC topology will not be mentioned in the further sections. Only the relevant nodes will be shown with the above mentioned node nomenclature.

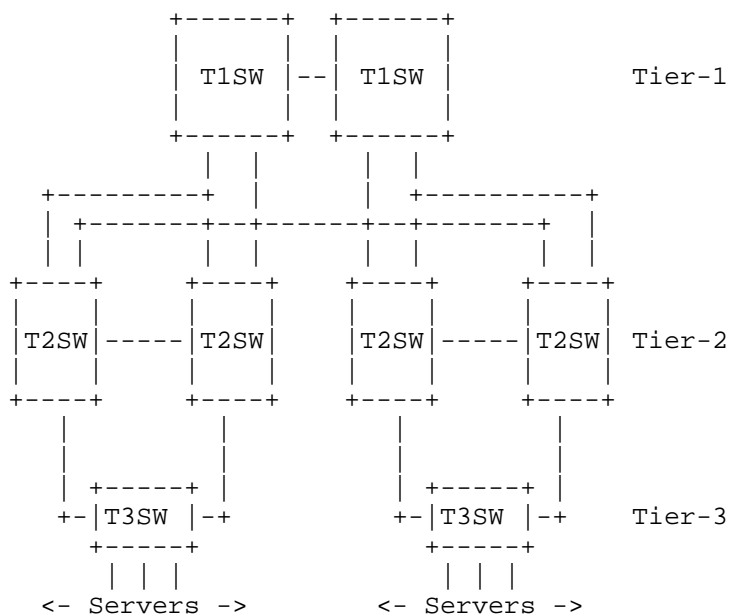


Figure 1: Typical DC network topology

## 2. Appointed Forwarders

TRILL supports multi-access LAN (Local Area Network) links that can have multiple end stations and RBridges attached. Where multiple RBridges are attached to a link, native traffic to and from end stations on that link is handled by a subset of those RBridges called "Appointed Forwarders" [rfc6439bis], with the intent that native traffic in each VLAN be handled by at most one RBridge. An RBridge can be Appointed Forwarder for many VLANs.

The Appointed Forwarder mechanism is irrelevant to any link on which end station service is not offered. This includes links configured as point-to-point IS-IS links and any link with all RBridge ports on that link configured as trunk ports. (In TRILL, configuration of a port as a "trunk port" just means that no end station service will be provided. It does not imply that all VLANs are enabled on that port). Furthermore, Appointed Forwarder status has no effect on the forwarding of TRILL Data frames. It only affects the handling of native frames.

By default, the DRB (Designated RBridge) on a link is in-charge of native traffic for all VLANs on the link. The DRB may, if it wishes, act as Appointed Forwarder for any VLAN and it may appoint other RBridges that have ports on the link as Appointed Forwarder for one or more VLANs.

The DRB may appoint other RBridges on the link with any one of the mechanism described in [rfc6439bis].

A RBridge on a multi-access link forms adjacency [RFC7177] with other RBridge if the VLAN's configured/enabled between them are common. For example there are four RBridges attached to multi-access link, say RB1, RB2, RB3 and RB4. RB1 and RB2 are configured with single VLAN "VLAN 2", whereas RB3 and RB4 are configured with "VLAN 3". Assume that there are no Native VLAN's present on any of the RBridges connected to multi-access link. Since TRILL Hellos are sent with VLAN Tag enabled on the interface, RB3 and RB4 drops the hellos of RB1 and RB2 (since they are not configured for VLAN 2). Similarly RB1 and RB2 drops the Hellos of RB3 and RB4. This results in RB1 and RB2 not forming adjacency with RB3 and RB4. RB1 and RB2 after electing DRB and forming adjacency between them, will decide about VLAN 2 AF. Similarly RB3 and RB4 decide about the VLAN 3 AF.

As VTSD should be capable of performing all the operations a standard TRILL Switch should do, it should also be capable of performing

Appointed Forwarder selection. A group of VTSD that are configured for same service's (VLAN's in our case) on different TIR's will form adjacencies, whereas VTSD which are enabled for different VTSI will never form adjacencies.

### 3. Multiple Parallel pseudowires.

TRILL supports multiple parallel adjacencies between neighbor R Bridges. Appendix C of [RFC6325] and section 3.5 of [RFC7177] describes this in detail. Multipathing across such parallel connections can be done for unicast TRILL Data traffic on a per-flow basis, but is restricted for multi-destination traffic. VTSD should also support this functionality.

TRILL DCI Pseudowires which belong to same VTSD instance in a TIR and connected to same remote TIR are referred to as parallel pseudowires. These parallel pseudowires corresponds to a single link inside VTSD.

Here all pseudowires should be capable of carrying traffic.

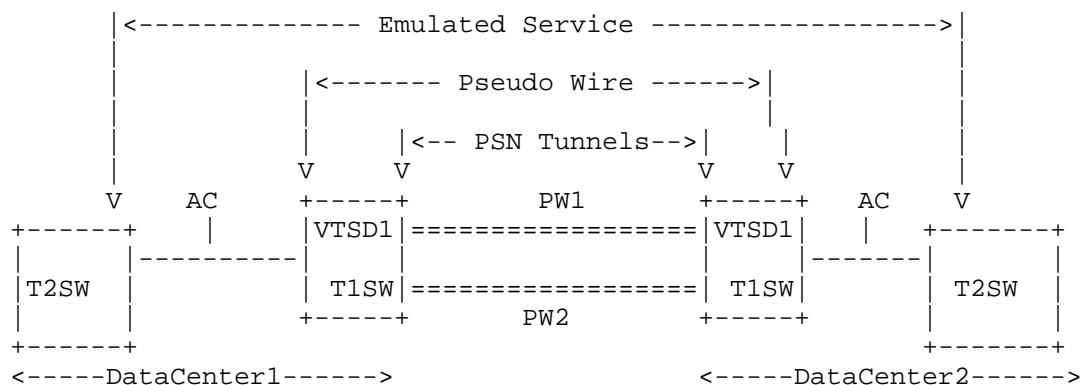


Figure 2: Parallel pseudowires with TRILL DCI

In above Figure 2, PW1 and PW2 are parallel pseudowires, as these pseudowires belongs to same VTSD and provides a connectivity across same TIRs.

This mechanism provides a way for actively increasing and optimally utilizing the bandwidth in the service provider network without affecting the existing traffic.

#### 4. Active-Active Pseudowire

[RFC6718] describes pseudowire Redundancy mechanism, wherein among the pair of pseudowires, one pseudowire will be selected as a active pseudowire and the other will be selected as a standby pseudowire. The standby pseudowire will not forward any user traffic under normal circumstances. The introduction of VTSD in TRILL DCI provides a very simple mechanism for providing multiple active pseudowires.

Pseudowires which belongs to the same VTSD instance inside the same TIR or between TIR's will be in active-active state. These pseudowires are able to carry data-traffic without making any one of pseudowire to go in standby mode.

To distribute traffic between pseudowires, TRILL protocol will be used.

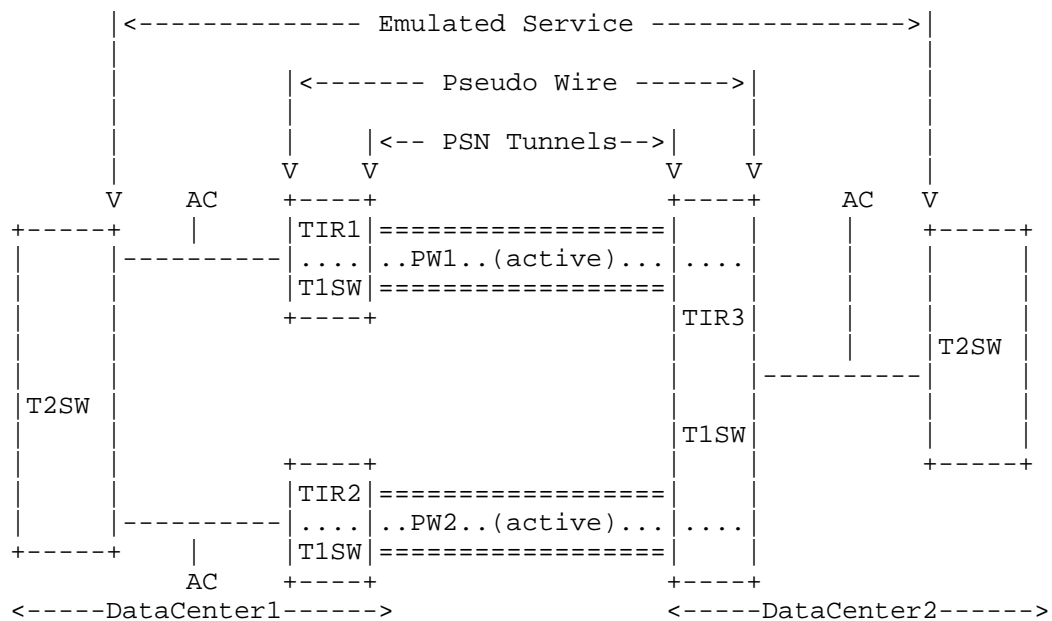


Figure 3: Dual-Home AC with Active-Active PW's

In the above Figure 3, pseudowires PW1 and PW2 are in active state and will be capable of carrying user traffic without making anyone of the pseudowire go in standby mode. The above Figure illustrates an application of multiple active pseudowires, where DC1's T2 switch (T2SW) is dual-homed with the TIR switch. This scenario is designed to actively load share the emulated service among the two TIRs attached to the multi-homed switch.

The attachment circuit can be of either Port-based Attachment Circuit or VLAN-based Attachment Circuit.

#### 4.1. Port-based AC operations.

In this case, the VTSDs in TIR1 and TIR2 will form TRILL adjacency via AC ports. If the attachment circuit port can carry N number of end-station service VLANs, then TIR1 and TIR2's VTSDs can equally distribute them using AF Mechanism of TRILL.

#### 4.2. VLAN-based AC operations.

Likewise in Port-based AC, in this case also the VTSDs in TIR1 and TIR2 will form TRILL adjacency via AC ports. Since only one VLAN end-station service is enabled, only one TIR's VTSD can become AF for that VLAN. Hence native traffic can be processed by any one of the AC.

### 5. MPLS encapsulation and Loop free provider PSN/MPLS

TRILL with MPLS encapsulation over pseudowire is specified in [RFC7173], and requires no changes in the frame format.

TRILL DCI doesn't require to employ Split Horizon mechanism in the provider PSN network, as TRILL takes care of Loop free topology using Distribution Trees. Any multi-destination frame will traverse a distribution tree path. All distribution trees are calculated based on TRILL base protocol standard [RFC6325] as updated by [RFC7180bis].

### 6. Frame processing

This section specifies frame processing from data center T2 switch and TIR's

#### 6.1. Frame processing between data center T2 switch and TIR.



In a multi-homed CE topology where in a data center switch is connected to two PEs / TIRs, AF mechanism described in section 2 will be used to decide which TIR/VTSD will carry the traffic for a particular VLAN. This is applicable to the case wherein the data center switch is connected to a PE/TIR device via multiple layer 2 interfaces to increase the bandwidth.

As a frame gets ingressed into a TIR (or any one of the TIR, when the tier2 switches are connected to multiple TIR's) after having AF check, the TIR encapsulates the frame with TRILL and MPLS headers and forwards the frame on a pseudowire. If parallel pseudowires are present, the TRILL protocol running in VTSD will select any one of the pseudowire and forward the TRILL Data packet. Multi-destination packets will be forwarded on Distribution tree's path [rfc7180bis]

The advantage of using TRILL for distribution of frames is, even if any of the paths or links fails between DC switch and TIR's or between TIR's, frames can be always be forwarded to any of available UP links or paths through other links/pseudowires.

If multiple equal paths are available, TRILL will distribute traffic among all the paths.

Also VTSD doesn't depend on the routing or signaling protocol that is running between TIRs, provided there is a tunnel available with proper encapsulation mechanism.

Any multi-destination frames when ingressed to TIR's will traverse one of the Distribution-Trees, with strong RFC Checks. Hop count field in TRILL Header will avoid loops or duplication of Traffic.

## 6.2. Frame processing between TIR's

When a frame gets ingressed into a VTSD inside TIR, the TRILL protocol will forward the frames to the proper pseudowire. When multiple paths / pseudowires are available between the TIR's then shortest path, calculated through TRILL protocol, will be used. If multiple paths are of equal cost, then TRILL protocol will do ECMP load spreading. If any multi-destination frame gets received by the VTSD through a pseudowire, TRILL will do an RPF check and will take proper action.

Once a frame gets to the VTSD through pseudowire, MPLS header will be de-capsulated, further action will be taken depending on the egress nickname field of TRILL header. If egress nickname is the nickname of this VTSD, MAC address table and AF lookup will be performed and the

frame will be forwarded by decapsulating the TRILL header. If egress nickname belongs to some other VTSD, frame will be forwarded on a pseudowire connected to that VTSD by encapsulating with an MPLS header.

7. MAC Address learning and withdrawal

MAC address learning and withdrawal mechanism on a RBridge is specified in section 4.8. of [RFC6325], this document requires no changes for MAC address learning and its withdrawal.

8. Active-Active Access with VTSD

TBD

9. ARP/ND proxy

TBD

10. MAC mass-withdrawal

TBD

11. Security Considerations

TBD

12. IANA Considerations

TBD

13. References

13.1. Normative References

[IS-IS]      "Intermediate system to Intermediate system routeing  
information exchange protocol for use in conjunction with  
the Protocol for providing the Connectionless-mode Network  
Service (ISO 8473)", ISO/IEC 10589:2002, 2002".

[rfc7180bis] Eastlake, D., et al, "TRILL: Clarifications,  
Corrections, and Updates", draft-ietf-trill-rfc7180bis,  
work in progress.,.

[draft-VTSD] Umair, M., Smiler, K., Eastlake, D., Yong, L.,  
"TRILL Transparent Transport over MPLS"  
draft-muks-trill-transport-over-mpls, work in  
progress.,.

[rfc6439bis] Eastlake, D., et al., "TRILL: Appointed Forwarders",  
draft-eastlake-trill-rfc6439bis, work in progress.,.

10.2. Informative References

Authors' Addresses

Mohammed Umair  
IP Infusion  
RMZ Centennial

Mahadevapura Post  
Bangalore - 560048 India  
EMail: mohammed.umair2@gmail.com

Kingston Smiler S  
IP Infusion  
RMZ Centennial  
Mahadevapura Post  
Bangalore - 560048 India  
EMail: kingstonsmiler@gmail.com

Shaji Ravindranathan  
IP Infusion  
3965 Freedom Circle, Suite 200  
Santa Clara, CA 95054 USA  
EMail: srnathan2014@gmail.com

Lucy Yong  
Huawei Technologies  
5340 Legacy Drive  
Plano, TX 75024  
USA  
Phone: +1-469-227-5837  
EMail: lucy.yong@huawei.com

Donald E. Eastlake 3rd  
Huawei Technologies  
155 Beaver Street  
Milford, MA 01757  
USA

Phone: +1-508-333-2270  
EMail: d3e3e3@gmail.com

INTERNET-DRAFT  
Intended Status: Proposed Standard

Mohammed Umair  
Kingston Smiler  
IP Infusion  
Donald Eastlake 3rd  
Lucy Yong  
Huawei Technologies  
July 6, 2015

Expires: January 7, 2016

TRILL Transparent Transport over MPLS  
draft-muks-trill-transport-over-mpls-00

Abstract

This document specifies how to interconnect Transparent Interconnection of Lots of links (TRILL) sites belonging to a tenant that are separated geographically over an MPLS domain. This draft addresses two problems 1) Providing connection between more than two TRILL sites that are separated by an MPLS provider network using [RFC7173] 2) Providing connection between TRILL sites belonging to a tenant over a MPLS provider network

Status of this Memo

This Internet-Draft is submitted to IETF in full conformance with the provisions of BCP 78 and BCP 79.

Internet-Drafts are working documents of the Internet Engineering Task Force (IETF), its areas, and its working groups. Note that other groups may also distribute working documents as Internet-Drafts.

Internet-Drafts are draft documents valid for a maximum of six months and may be updated, replaced, or obsoleted by other documents at any time. It is inappropriate to use Internet-Drafts as reference material or to cite them other than as "work in progress."

The list of current Internet-Drafts can be accessed at  
<http://www.ietf.org/lid-abstracts.html>

The list of Internet-Draft Shadow Directories can be accessed at  
<http://www.ietf.org/shadow.html>

Copyright and License Notice

Copyright (c) 2015 IETF Trust and the persons identified as the document authors. All rights reserved.

This document is subject to BCP 78 and the IETF Trust's Legal Provisions Relating to IETF Documents (<http://trustee.ietf.org/license-info>) in effect on the date of publication of this document. Please review these documents carefully, as they describe your rights and restrictions with respect to this document. Code Components extracted from this document must include Simplified BSD License text as described in Section 4.e of the Trust Legal Provisions and are provided without warranty as described in the Simplified BSD License.

## Table of Contents

1	Introduction . . . . .	3
1.1	Terminology . . . . .	3
2.	TRILL Over MPLS Model . . . . .	4
3.	VPLS Model . . . . .	5
3.1.	Entities in the VPLS Model . . . . .	6
3.3.	TRILL Adjacency for VPLS model . . . . .	7
3.4.	MPLS encapsulation for VPLS model . . . . .	7
3.5	Loop Free provider PSN/MPLS. . . . .	7
3.6.	Frame processing. . . . .	7
4.	VPTS Model . . . . .	7
4.1.	Entities in the VPTS Model . . . . .	9
4.1.1	TRILL Intermediate Routers [TIR] . . . . .	9
4.1.2	Virtual TRILL Switch Domain (VTSD) . . . . .	10
4.2.	TRILL Adjacency for VPLS model . . . . .	10
4.3.	MPLS encapsulation for VPLS model . . . . .	10
4.4	Loop Free provider PSN/MPLS. . . . .	10
4.5.	Frame processing. . . . .	10
4.5.1	Multi-Destination Frame processing . . . . .	10
4.5.2	Unicast Frame processing . . . . .	11
5.	Extensions to TRILL Over Pseudowires [RFC7173] . . . . .	11
6.	VPTS Model Versus VPLS Model . . . . .	11
7.	Security Considerations . . . . .	12
8.	IANA Considerations . . . . .	12
9.	References . . . . .	12
9.1	Normative References . . . . .	12
9.2	Informative References . . . . .	13
	Authors' Addresses . . . . .	14

## 1 Introduction

The IETF Transparent Interconnection of Lots of Links (TRILL) protocol [RFC6325] [RFC7177] [RFC7180bis] provides transparent forwarding in multi-hop networks with arbitrary topology and link technologies using a header with a hop count and link-state routing. TRILL provides optimal pair-wise forwarding without configuration, safe forwarding even during periods of temporary loops, and support for multipathing of both unicast and multicast traffic. Intermediate Systems (ISs) implementing TRILL are called Routing Bridges (RBridges) or TRILL Switches

This draft, in conjunction with [RFC7173], address two problems

1) Providing connection between more than two TRILL sites of a single TRILL network that are separated by an MPLS provider network using [RFC7173]. (Herein also called as problem statement 1.)

2) Providing connection between TRILL sites belongs to a tenant/tenants over a MPLS provider network. (Herein also called as problem statement 2.)

A tenant is the administrative entity on whose behalf one or more customers and their associated services are managed. Here Customer refers to TRILL campus not Data Label.

A key multi-tenancy requirement is traffic isolation so that one tenant's traffic is not visible to any other tenant. This draft also addresses the problem of multi-tenancy by isolating one tenant's traffic from the other.

### 1.1 Terminology

The key words "MUST", "MUST NOT", "REQUIRED", "SHALL", "SHALL NOT", "SHOULD", "SHOULD NOT", "RECOMMENDED", "MAY", and "OPTIONAL" in this document are to be interpreted as described in RFC 2119 [RFC2119].

Acronyms used in this document include the following:

AC	- Attachment Circuit [RFC4664]
Data Label	- VLAN or FGL
ECMP	- Equal Cost Multi Path
FGL	- Fine-Grained Labeling [RFC7172]

IS-IS	- Intermediate System to Intermediate System [IS-IS]
LDP	- Label Distribution Protocol
LAN	- Local Area Network
MPLS	- Multi-Protocol Label Switching
PE	- Provider Edge Device
PPP	- Point-to-Point Protocol [RFC1661]
PSN	- Packet Switched Network
PW	- Pseudowire [RFC4664]
TIR	- TRILL Intermediate Router [Devices where Pseudowire starts and Terminates]
TRILL	- Transparent Interconnection of Lots of Links OR Tunnelled Routing in the Link Layer
TRILL Site	- A part of a TRILL campus that contains at least one RBridge.
VLAN	- Virtual Local Area Network
VPLS	- Virtual Private LAN Service
VPTS	- Virtual Private TRILL Service
VSI	- Virtual Service Instance [RFC4664]
VTSD	- Virtual TRILL Switch Domain A Virtual RBridge which segregates one tenant's TRILL database as well as traffic from the other.
WAN	- Wide Area Network

## 2. TRILL Over MPLS Model

TRILL Over MPLS can be achieved by two different ways.

- a) VPLS Model for TRILL
- b) VPTS Model/TIR Model



Both these models can be used to solve the problem statement 1 and 2. Herein the VPLS Model for TRILL is also called Model 1 and the VPTS Model/TIR Model is also called Model 2.

### 3. VPLS Model

Figure 1 shows the topological model of TRILL over MPLS using VPLS model. The PE routers in the below topology model should support all the functional Components mentioned in [RFC4664].

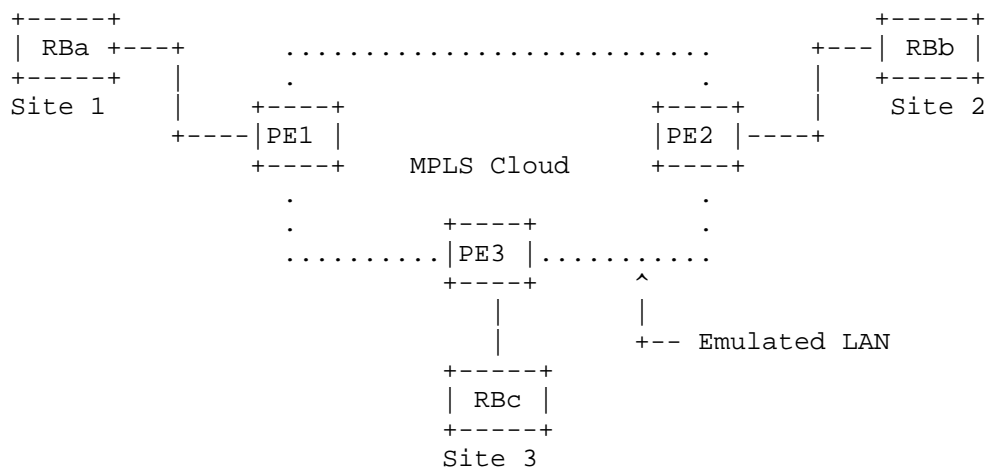
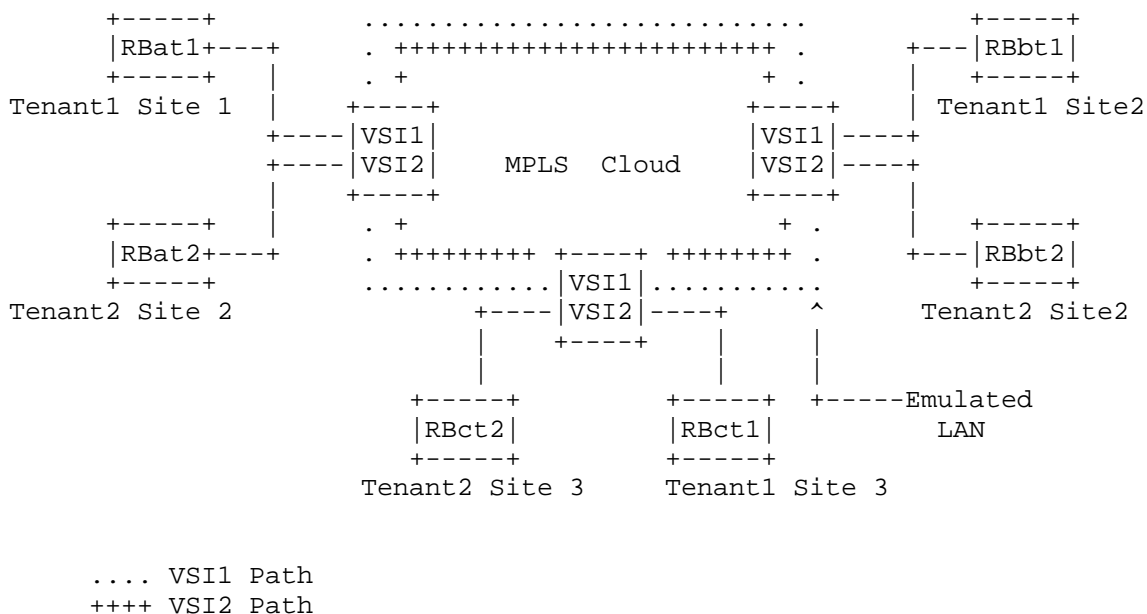


Figure 1: Topological Model of TRILL over MPLS  
connecting three TRILL Sites

Figure 2 below shows the topological model of TRILL over MPLS to connect multiple TRILL sites belonging to a tenant (tenant here is a campus, not a Data label). VSI1 and VSI2 are two Virtual Service Instances which segregates Tenant1's traffic from Tenant2's. VSI1 will maintain its own database for Tenant1, similarly VSI2 will maintain its own database for Tenant2.



In this model TRILL sites are connected using VPLS-capable PE devices that provide a logical interconnect, such that TRILL R Bridges belonging to a specific tenant connected via an single bridged Ethernet. These devices are same as PE devices specified in [RFC4026]. The Attachment Circuit ports of PE Routers are layer 2 switch ports that are connected to the R Bridges in a TRILL site. Here each VPLS instance looks like an emulated LAN. This model is similar to connecting different R Bridges (TRILL sites) by a layer 2 bridge domain (multi access links) as specified in [RFC6325]. This model doesn't requires any changes in PE routers to carry TRILL frames, as TRILL frame will be transferred transparently.

The RB (RBridge) and TRILL Sites are defined in [RFC6325]

### 3.3. TRILL Adjacency for VPLS model

As specified in section 3 of this document, the MPLS cloud looks like an emulated LAN (also called multi-access link or broadcast link). This results in RBridge of different sites looking like that they are connected to a multi-access link. With such interconnection, the TRILL adjacency over the link is automatically discovered and established through TRILL IS-IS control messages [RFC7177] which is transparently forwarded by the VPLS domain, after doing MPLS encapsulation specified in the section 3.4.

### 3.4. MPLS encapsulation for VPLS model

MPLS encapsulation over Ethernet pseudowire is specified in [RFC7173] Appendix A, and requires no changes in the frame format.

### 3.5 Loop Free provider PSN/MPLS.

No explicit handling is required to avoid loop free topology as, Split Horizon technique mentioned in [RFC4664] in the provider PSN network takes care of loop-free topology in the PSN.

### 3.6. Frame processing.

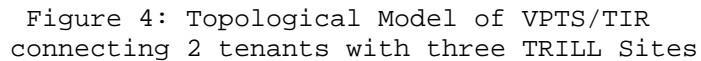
The PE device transparently process the TRILL control and data frames and procedure to forward the frames are defined in [RFC4664]

## 4. VPTS Model

The [Virtual Private TRILL Service] VPTS is an L2 TRILL service that emulates TRILL service across a Wide Area Network (WAN). VPTS is similar to what VPLS does for bridge domain. VPLS provides virtual private LAN service for different customers. VPTS provide Virtual Private TRILL service (VPTS) for different TRILL tenants.

Figure 3 shows the topological model of TRILL over MPLS using VPTS. In this model the PE routers are replaced with TIR [TRILL Intermediate Router] and VSI is replaced with VTSD [Virtual TRILL Switch Domain]. The TIR [TRILL Intermediate Router] devices are interconnected via PWS appear as a single emulated TRILL Site with each VTSD inside a TIR equivalent to a RBridge. The TIR devices must be capable of supporting both MPLS and TRILL.





The TIRS [TRILL Intermediate Routers] must be capable of running both VPLS and TRILL protocols. TIR devices are superset of VPLS-PE devices which is defined in [RFC4026]. The VSI instance that provides transparent bridging functionality in the PE device is replaced with VTSD in TIR.

#### 4.1.2 Virtual TRILL Switch Domain (VTSD)

The VTSD [Virtual Trill Switch Domain] is similar to VSI (layer 2 bridge) in VPLS model, but this acts as TRILL RBridge. The VTSD is a superset of VSI and must support all the functionality provided by the VSI as defined in [RFC4026]. Along with VSI functionality, the VTSD must be capable of supporting TRILL protocols and form TRILL adjacency. The VTSD must be capable of performing all the operations that a standard TRILL Switch can do.

One VTSD instance per tenant must be maintained, when multiple tenants are connected to the TIR. The VTSD must maintain all the information maintained by the RBridge on a per tenant basis. The VTSD must also take care of segregating one tenant traffic from other.

#### 4.2. TRILL Adjacency for VPLS model

The VTSD must be capable of forming TRILL adjacency with other VTSDs present in its peer VPTS neighbor, and also the RBridges present in the TRILL sites. The procedure to form TRILL Adjacency is specified in [RFC7173] and [RFC7177].

#### 4.3. MPLS encapsulation for VPLS model

MPLS encapsulation over pseudowire is specified in [RFC7173], and requires no changes in the frame format.

#### 4.4 Loop Free provider PSN/MPLS.

This model isn't required to employ Split Horizon mechanism in the provider PSN network, as TRILL takes care of Loop free topology using Distribution Trees. Any multi-destination frame will traverse a distribution tree path. All distribution trees are calculated based on TRILL base protocol standard [RFC6325] as updated by [RFC7180bis].

#### 4.5. Frame processing.

This section specifies multi-destination and unicast frame processing in VPTS/TIR model.

##### 4.5.1 Multi-Destination Frame processing

Any unknown unicast, multicast or broadcast frames inside VTSD should be

processed or forwarded through any one of the distribution tree's path. If any multi-destination frame is received from the wrong pseudowire at a VTSD, the TRILL protocol running in VTSD should perform a RPF check as specified in [RFC7180bis] and drops the packet.

Pruning mechanism of Distribution Tree as specified in [RFC6325] and [RFC7180bis] can also be used for forwarding of multi-destination data frames on the branches that are not pruned.

#### 4.5.2 Unicast Frame processing

Unicast frames must be forwarded in same way they get forwarded in a standard TRILL Campus as specified in [RFC6325]. If multiple equal cost paths are available over pseudowires to reach destination, then VTSD should be capable of doing ECMP for them.

### 5. Extensions to TRILL Over Pseudowires [RFC7173]

The [RFC7173] mentions how to interconnect a pair of Transparent Interconnection of Lots of Links (TRILL) switch ports using pseudowires. This document explains, how to connect multiple TRILL sites (not limited to only two sites) using the mechanisms and encapsulations defined in [RFC7173].

### 6. VPTS Model Versus VPLS Model

VPLS Model uses a simpler loop breaking rule: the "split horizon" rule, where a PE must not forward traffic from one PW to another in the same VPLS mesh.

An issue with the above rule is that if a pseudowire between PEs fails, frames will not get forwarded between the PEs where pseudowire went down.

VPTS solves this problem, since the VPTS Model uses distribution Trees for loop free topology, so frames reach all TIRs even when any one of the pseudowires fails in a mesh topology.

If equal cost paths are available to reach a site over pseudowires, VPTS Model can use ECMP for processing of frames over pseudowires.

## 7. Security Considerations

For general TRILL security considerations, see [RFC6325]

For transport of TRILL by Pseudowires security consideration, see [RFC7173].

Since VPTS Model uses Distribution tree for processing of multi-destination data frames, it is always advisable to run at least one Distribution tree in a TRILL site per tenant, this will avoid data frames getting received on TRILL sites where end-station service is not enabled for that data frame.

## 8. IANA Considerations

This document requires no IANA actions. RFC Editor: Please delete this section before publication

## 9. References

### 9.1 Normative References

[RFC6325]      Perlman, R., Eastlake 3rd, D., Dutt, D., Gai, S., and A.Ghanwani, "Routing Bridges (RBridges): Base Protocol Specification", RFC 6325, July 2011.

[RF7180bis]      Eastlake 3rd, D., Zhang, M., Perlman, R., Banerjee, A., A.Ghanwani, and Gupta, S, "Routing Bridges (RBridges): TRILL: Clarifications, Corrections, and Updates", work in progress.  
"https://tools.ietf.org/html/draft-ietf-trill-rfc7180bis-05"

[RFC7173]      Yong, L., Eastlake 3rd, D., Aldrin, S., and Hudson, J, "Transparent Interconnection of Lots of Links (TRILL) Transport Using Pseudowires", RFC 7173, May 2014.

[RFC4762]      Lasserre, M., and Kompella, V., Virtual Private LAN



Service (VPLS) Using Label Distribution Protocol  
(LDP) Signaling, RFC 4762, January 2007

- [RFC4026]      Andersson, L., and Madsen, T., Provider Provisioned  
Virtual Private Network (VPN) Terminology, RFC 4026,  
March 2005
- [RFC4664]      Andersson, L., and Rosen, E., Framework for Layer 2  
Virtual Private Networks (L2VPNs), RFC 4664,  
September 2006

## 9.2 Informative References

- [IS-IS]      ISO/IEC 10589:2002, Second Edition,  
"Information technology -- Telecommunications  
and information exchange between systems --  
Intermediate System to Intermediate System  
intra-domain routeing information exchange  
protocol for use in conjunction with the  
protocol for providing the connectionless-mode  
network service (ISO 8473)", 2002.
- [RFC3985]      Bryant, S., Ed., and P. Pate, Ed., "Pseudo  
Wire Emulation Edge-to-Edge (PWE3)  
Architecture", RFC 3985, March 2005.
- [RFC4023]      Worster, T., Rekhter, Y., and E. Rosen, Ed.,  
"Encapsulating MPLS in IP or Generic Routing  
Encapsulation (GRE)", RFC 4023, March 2005.
- [RFC4448]      Martini, L., Ed., Rosen, E., El-Aawar, N., and  
G. Heron, "Encapsulation Methods for Transport  
of Ethernet over MPLS Networks", RFC 4448,  
April 2006.
- [RFC7177]      Eastlake 3rd, D., Perlman, R., Ghanwani, A.,  
Yang, H., and V. Manral, "Transparent  
Interconnection of Lots of Links (TRILL):  
Adjacency", RFC 7177, May 2014.

[RFC7172]    Eastlake 3rd, D., Zhang, R., Agarwal, P.,  
             Perlman, R., and Dutt, D, "Transparent  
             Interconnection of Lots of Links (TRILL):  
             Fine-Grained Labeling", RFC 7172, May 2014.

Authors' Addresses

Mohammed Umair  
IP Infusion  
RMZ Centennial  
Mahadevapura Post  
Bangalore - 560048 India

EMail: [mohammed.umair2@gmail.com](mailto:mohammed.umair2@gmail.com)

Kingston Smiler  
IP Infusion  
RMZ Centennial  
Mahadevapura Post  
Bangalore - 560048 India

EMail: [kingstonsmiler@gmail.com](mailto:kingstonsmiler@gmail.com)

Donald E. Eastlake 3rd  
Huawei Technologies  
155 Beaver Street  
Milford, MA 01757  
USA

Phone: +1-508-333-2270  
EMail: [d3e3e3@gmail.com](mailto:d3e3e3@gmail.com)

Lucy Yong  
Huawei Technologies  
5340 Legacy Drive  
Plano, TX 75024

USA

Phone: +1-469-227-5837

EMail: lucy.yong@huawei.com