

TRILL Working Group
INTERNET-DRAFT
Intended status: Proposed Standard

Expires: September 20, 2016

Donald Eastlake
Huawei
Bob Briscoe
Simula Research Lab
March 21, 2016

TRILL: ECN (Explicit Congestion Notification) Support
<draft-eastlake-trill-ecn-support-00.txt>

Abstract

Explicit congestion notification (ECN) allows a forwarding element to notify downstream devices, including the destination, of the onset of congestion without having to drop packets. This document extends this capability to TRILL switches, including integration with IP ECN.

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>.

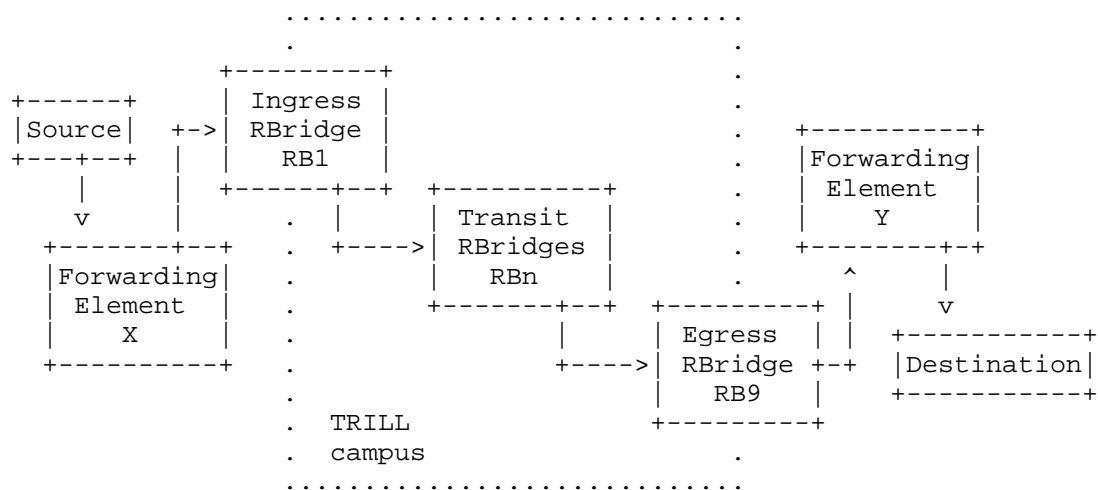
Table of Contents

1. Introduction.....	3
1.1 Conventions used in this document.....	4
2. The ECN Specific Extended Header Flags.....	5
3. ECN Support.....	6
3.1 Ingress ECN Support.....	6
3.2 Transit ECN Support.....	6
3.3 Egress ECN Support.....	7
4. IANA Considerations.....	9
4.1 Flags Word Bits.....	9
4.2 Extended RBridge Capability Bit.....	9
5. Security Considerations.....	10
6. Acknowledgements.....	10
Normative References.....	11
Informative References.....	11
Authors' Addresses.....	12

1. Introduction

Explicit congestion notification (ECN [RFC3168]) allows a forwarding element, such as a router, to notify downstream devices, including the destination, of the onset of congestion without having to drop packets. Instead, the forwarding element can explicitly mark a proportion of packets in a two-bit ECN field. For example, a two-bit field in IP headers is available for ECN marking.

The transit of user data through a TRILL campus is similar to transport through a tunnel with the ingress and egress R Bridges equivalent to the ends of the tunnel. Thus, existing ECN tunneling recommendations, particularly [RFC6040], apply.



In the figure above, if ECN is implemented and assuming IP traffic, RB1 is effectively a tunnel entrance and RB9 a tunnel exit. Traffic from Source to RB1 might or might not get marked as having experienced congestion in forwarding elements, such as X, before being encapsulated at ingress RB1. Any such ECN marking is encapsulated with a TRILL Header and provision is made in the TRILL Header extension Flags Word for ECN marking by the R Bridges through which this traffic passes.

Any ECN marking in the traffic at the ingress is copied out to the TRILL Header Flags Word. At RB9, the TRILL egress, any ECN markings in the TRILL Header Flags Word and in the encapsulated traffic are combined so that subsequent forwarding elements, such as Y and the Destination, can see if congestion was experienced at any previous point in the path from Source if the forwarding elements are ECN capable and the Source marked packets as ECT (ECN Capable Transport).

1.1 Conventions used in this document

The terminology and acronyms defined in [RFC6325] are used herein with the same meaning.

In this documents, "IP" refers to both IPv4 and IPv6.

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].

Acronyms:

CE - Congestion Experienced

ECN - Explicit Congestion Notification

ECT - ECN Capable Transport

2. The ECN Specific Extended Header Flags

RBridges MAY implement ECN (Explicit Congestion Notification) [RFC3168] through a two-bit field in the TRILL Header extension Flags Word [RFC7780]. If implemented, it SHOULD be enabled by default but can be disabled on a per RBridge basis by configuration.

This field is shown below as "ECN" and consists of bits 12 and 13 which are in the range reserved for non-critical hop-by-hop bits. See [RFC7780] and [RFC7179] for the meaning of the other bits.

0			1							2							3				
0	1	2	3	4	5	6	7	8	9	0	1	2	3	4	5	6	7	8	9	0	1
Crit.			CHbH				NCHbH				CRSV		NCRSV		CItE				NCItE		
.....				
C	C	C					C	N													
R	R	R					R	C	ECN				Ext					Ext			
H	I	R					C	C					Hop					Clr			
b	t	s					A	A					Cnt								
H	E	v					F	F													

The following table is modified from [RFC3168] and shows the meaning of bit values in TRILL Header extended flags 12 and 13. These are also the meanings of bits 6 and 7 of the DS field in the IPv4 and IPv6 headers as defined in [RFC3168]:

Binary	Meaning
00	Not-ECT (Not ECN-Capable Transport)
01	ECT(1) (ECN-Capable Transport(1))
10	ECT(0) (ECN-Capable Transport(0))
11	CE (Congestion Experienced)

Table 1. ECN Field Bit Combinations

3. ECN Support

An RBridge that has ECN support as specified herein advertises this through bit TBD in the Extended RBridge Capabilities APPsub-TLV [RFC7782] (see Section 4.2). On encapsulation, transit, and decapsulation it behaves as described in the subsections below, which correspond to the recommended provisions of [RFC6040].

3.1 Ingress ECN Support

Behavior at the ingress depends on whether the egress RBridge supports ECN. If it does, then the behavior is as follows (called "normal mode" in [RFC6040]):

- o When encapsulating an IP frame that is ECN enabled (non-zero ECN field), the ingress RBridge MUST create a flags word as part of the TRILL Header, setting the F flag, and copy the two ECN bits from the IP header into flag word bits 12 and 13.
- o When encapsulating a frame for a non-IP protocol, where that protocol has a means of indicating ECN that is understood by the ingress RBridge, it MAY add a flags word to the TRILL Header with the ECN bits set from the encapsulated native frame.

If the egress RBridge does not support ECN, the behavior is as follows (called "compatibility mode" in [RFC6040]):

- o A TRILL Header Flags Word need not be created unless there is some reason other than ECN to do so.
- o If a Flags Word is created, the ECN bits are set to zero (the Non-ECT value).

3.2 Transit ECN Support

When forwarding a TRILL Data packet encountering congestion at an RBridge, if the TRILL Header flags word is present, bits 12 and 13 are updated in the usual ECN manner [RFC3168]. An RBridge detects congestion either by monitoring its own queue depths or from participation in a link-specific protocol.

If, for reasons other than ECN, conditions at a transit RBridge require the insertion of a TRILL Header Flags Word into a TRILL Data packet, this implies that the egress RBridge is not ECN capable -- if it was, the Flags Word would have been included in the TRILL Data packet at the ingress. Thus, when a transit RBridge creates such a

Flags Word, it sets bits 12 and 13 to zero.

3.3 Egress ECN Support

Egress RBridge support of ECN is determined by looking at the Extended Capabilities APPsub-TLV that RBridge advertises. If bit TBD is zero, or the APPsub-TLV is absent, that RBridge does not support ECN. If the APPsub-TLV is present and bit TBD is one, then it does support ECN. If there are inconsistent APPsub-TLVs, the egress RBridge is assumed to support ECN if any of those APPsub-TLVs indicate that it does.

If the egress RBridge does not support ECN, it will ignore bits 12 and 13 of any Flags Word that is present, because it does not contain any special ECN logic.

If the egress RBridge supports ECN, it does the following:

- o When decapsulating an IP frame, the RBridge MUST set the outgoing native IP frame ECN field to the code point at the intersection of the values for that field in the encapsulated IP frame (row) and the TRILL Header flags word ECN field (column) in Table 2 below or drop the frame in the case where the TRILL header indicates congestion experienced but the encapsulated native IP frame indicates a not ECN-capable transport. (Such frame dropping is necessary because IP transport that is not ECN-capable requires dropped frames to sense congestion.)
- o When decapsulating a non-IP protocol frame with a means of indicating ECN that is understood by the RBridge, it MAY set the ECN information in the decapsulated native frame by combining that information in the TRILL Header flags word and the encapsulated non-IP native frame as specified in Table 2.

Table 2 below (adapted from [RFC6040]) shows how, at the egress, to combine the ECN information in the extended TRILL Header ECN field with the ECN information in an encapsulated frame to produce the ECN information to be carried in the resulting native frame.

Inner Native Header	Arriving TRILL Header Flag Word ECN Field			
	Not-ECT	ECT(0)	ECT(1)	CE
Not-ECT	Not-ECT	Not-ECT(*)	Not-ECT(*)	<drop>(*)
ECT(0)	ECT(0)	ECT(0)	ECT(1)	CE
ECT(1)	ECT(1)	ECT(1)(*)	ECT(1)	CE
CE	CE	CE	CE(*)	CE

Table 2: Egress ECN Behavior

An asterisk in the above table indicates a probably erroneous condition that SHOULD be logged.

4. IANA Considerations

This section summarizes IANA actions required.

4.1 Flags Word Bits

IANA is requested to assign bits 12 and 13 in the TRILL Header Flags Word for ECN and update the TRILL Extended Header Flags registry by replacing the line for bits 9-13 with the following"

Bits	Purpose	Reference
-----	-----	-----
9-11	available non-critical hop-by-hop flags	
12-13	ECN (Explicit Congestion Notification)	[this document]

4.2 Extended RBridge Capability Bit

IANA is requested to assign bit TBD in the Extended RBridge Capabilities to indicate ECN support. The Extended RBridge Capabilities registry on the TRILL Parameters page is updated by adding the folloing line and updating any "Unassigned" line that is affected.

Bit	Mnemonic	Description	Reference
---	-----	-----	-----
TBD	ECN	ECN Support	[this document]

5. Security Considerations

TBD

For ECN tunneling security considerations, see [RFC6040].

For general TRILL protocol security considerations, see [RFC6325].

6. Acknowledgements

This document was prepared with basic NROFF. All macros used were defined in the source file.

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, <<http://www.rfc-editor.org/info/rfc2119>>.
- [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>>.
- [RFC6040] - Briscoe, B., "Tunnelling of Explicit Congestion Notification", RFC 6040, DOI 10.17487/RFC6040, November 2010, <<http://www.rfc-editor.org/info/rfc6040>>.
- [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>>.
- [RFC7179] - Eastlake 3rd, D., Ghanwani, A., Manral, V., Li, Y., and C. Bestler, "Transparent Interconnection of Lots of Links (TRILL): Header Extension", RFC 7179, DOI 10.17487/RFC7179, May 2014, <<http://www.rfc-editor.org/info/rfc7179>>.
- [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>>.
- [RFC7782] - Zhang, M., Perlman, R., Zhai, H., Durrani, M., and S. Gupta, "Transparent Interconnection of Lots of Links (TRILL) Active-Active Edge Using Multiple MAC Attachments", RFC 7782, DOI 10.17487/RFC7782, February 2016, <<http://www.rfc-editor.org/info/rfc7782>>.

Informative References

[none]

Authors' Addresses

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

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

Bob Briscoe (editor)
Simula Research Lab

Email: ietf@bobbriscoe.net
URI: <http://bobbriscoe.net/>

Copyright and IPR Provisions

Copyright (c) 2016 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: September 20, 2015

Weiguo Hao
Donald Eastlake
Yizhou Li
Huawei
March 21, 2016

TRILL: Address Flush Message
<draft-hao-trill-address-flush-01.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 a message by which an originating TRILL switch can explicitly request other TRILL switches to flush certain MAC reachability learned through the egress of TRILL Data packets. This is a supplement to the TRILL automatic address forgetting and can assist in achieving more rapid convergence in case of topology or configuration change.

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.....	3
1.1 Terminology and Acronyms.....	3
2. Address Flush Message Details.....	5
2.1 VLAN Block Case.....	6
2.2 Extensible Case.....	7
3. IANA Considerations.....	11
4. Security Considerations.....	11
Normative References.....	12
Informative References.....	12
Acknowledgements.....	12
Authors' Addresses.....	13

1. Introduction

Edge TRILL (Transparent Interconnection of Lots of Links) switches [RFC6325] [RFC7780], also called edge RBridges, by default learn end station MAC address reachability 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 packet, they learn the remote MAC address and remote attachment TRILL switch. Such learning is all 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 and configuration changes; however, there are circumstances under which it would be helpful for a TRILL switch to be able to explicitly flush (purge) certain learned end station reachability information in remote RBridges to achieve more rapid convergence (see, for example, [TCaware] and Section 6.2 of [RFC4762]).

A TRILL switch R1 can easily flush any locally learned addresses it wants. This document specifies an RBridge Channel protocol [RFC7178] message to request flushing address information learned from decapsulating at remote 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 [ChannelTunnel] as well as the following:

Data Label - VLAN or FGL.

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

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 [ChannelTunnel], 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.

2. Address Flush Message Details

The Address Flush message is an RBridge Channel protocol message [RFC7178].

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 the Protocol field in the RBridge Channel Header that indicates how to interpret the Channel Protocol Specific Payload [RFC7178].

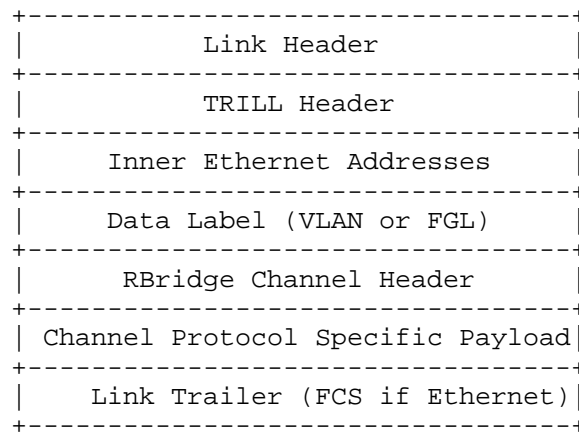


Figure 1. RBridge Channel Protocol Message Structure

An Address Flush RBridge Channel message by default applies to addresses within the 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. Such messages SHOULD be sent at priority 6 since they are important control messages but lower priority than control messages that establish or maintain adjacency.

Nevertheless:

- There are provisions for optionally indicating the Data Label(s) to be flushed for cases where the Address Flush message is sent over a Management VLAN or the like.
- An Address Flush message can be sent unicast, if it is desired to clear addresses at one TRILL switch only.

2.1 VLAN Block Case

Figure 2 below expands the RBridge Channel Header and Channel Protocol Specific Payload from Figure 1 for the case of the VLAN based Address Flush message. This form of the Address Flush message is optimized for flushing MAC addressed based on nickname and blocks of VLANs.

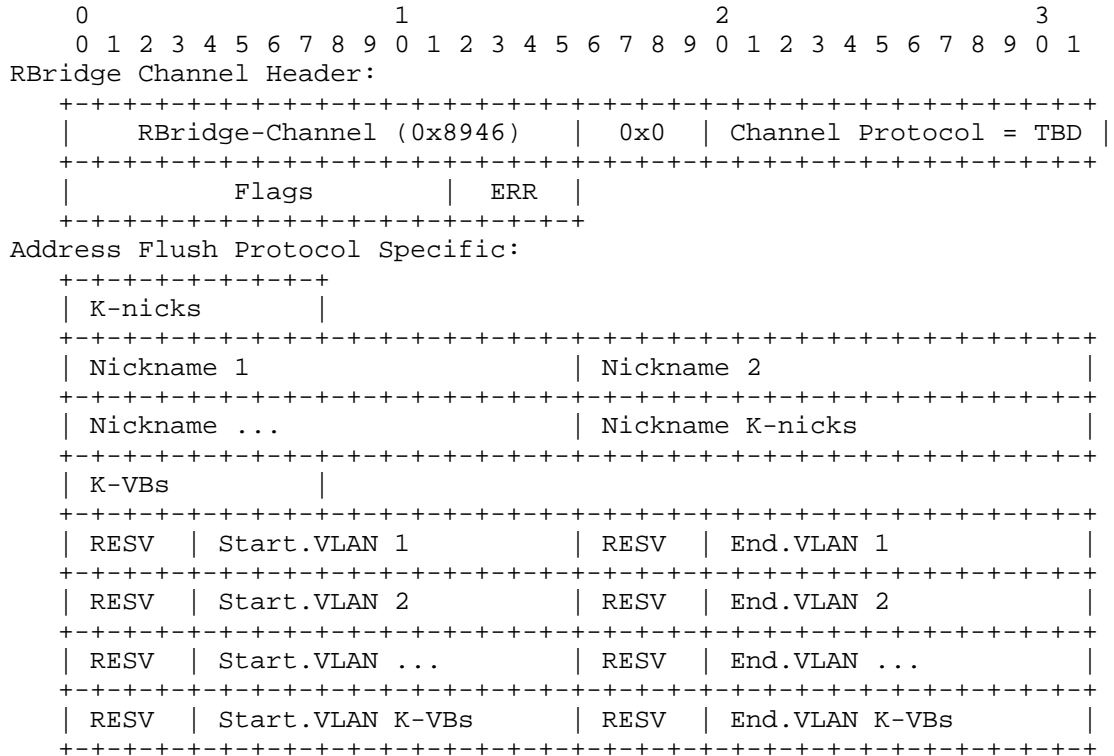


Figure 2. Address Flush Message - VLAN Case

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).

K-nicks: K-nicks is the number of nicknames present as an unsigned integer. If this is zero, the ingress nickname in the TRILL Header is considered to be the only nickname to which the message applies. If non-zero, it given the number of nicknames present to which the message applies. The messages flushes address learning due to egressing TRILL Data packets that had a ingress nicknam to which the message applies.

Nickname: A listed nickname to which it is intended that the Address Flush message apply. If an unknown or reserved nickname occurs in the list, it is ignored but the address flush operation is still executed with the other nicknames. If an incorrect nickname occurs in the list, so some address learning is flushed that should not have been flush, the network will still operate correctly but will be less efficient as the incorrectly flushed learning is re-learned.

K-VBs: K-VBs is the number of VLAN blocks present as an unsigned integer. If this byte is zero, the message is the more general format specified in Section 2.2. If it is non-zero, it gives the number of blocks of VLANs present.

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

Start.VLAN, End.VLAN: These 12-bit fields give the beginning and ending VLAN IDs of a block of VLANs. The block includes both the starting and ending values so a block of size one is indicated by setting End.VLAN equal to Start.VLAN. If Start.VLAN is 0x000, it is treated as if it was 0x001. If End.VLAN is 0xFFF, it is treated as if it was 0xFFE. If End.VLAN is smaller than Start.VLAN, considering both as unsigned integers, that VLAN block is ignored but the address flush operation is still executed with any other VLAN blocks in the message.

This message flushes all addresses learned from egressing TRILL Data packets with an applicable nickname and a VLAN in any of the blocks given. To flush addresses for all VLANs, it is easy to specify a block covering all valid VLAN IDs, this is, from 0x001 to 0xFFE.

2.2 Extensible Case

A more general form of the Address Flush message is provided to support flushing by FGL and more efficient encodings of VLANs and FGLs where using a set of contiguous blocks is cumbersome. This form is also extensible to handle future requirements.

It is indicated by a zero in the byte shown in Figure 2 as "K-VBs".

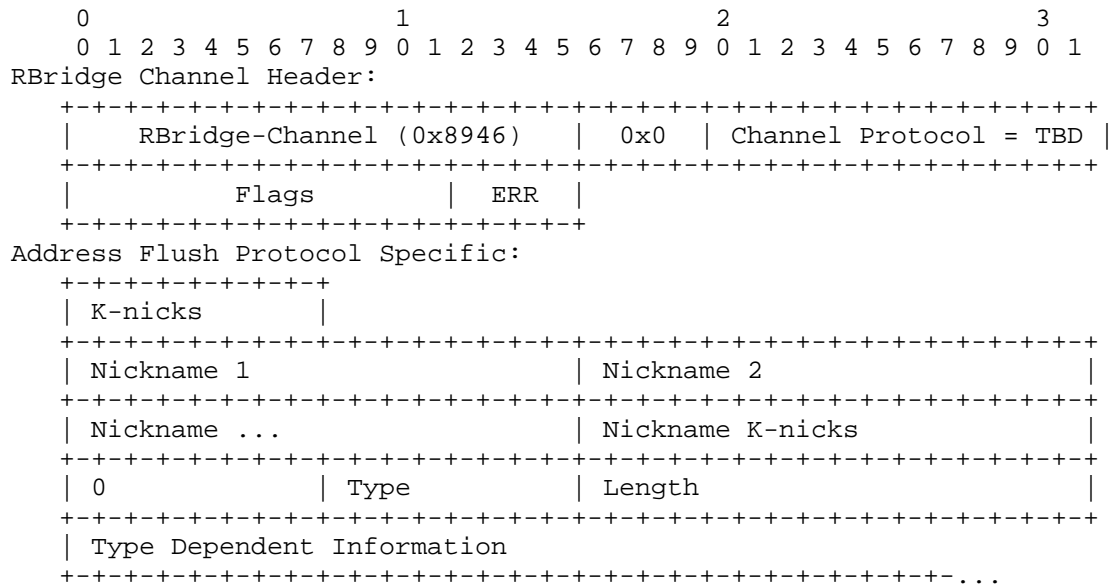


Figure 3. Address Flush Message - Extensible Case

Channel Protocol, K-nicks, Nickname: These fields are as specified in Section 2.1.

Type: If the byte immediately before the Type field, which is the byte labeled "K-VBs" in Figure 2, is zero, the the Type byte indicates the type of extended Address Flush message as follows:

Type	Description
-----	-----
0	Reserved
1	Bit Map of VLANs
2	Blocks of FGLs
3	List of FGLs
4	Bit Map of FGLs
5-254	Unassigned
255	Reserved

Length: The length of the remaining information in the Address Flush message.

Type Dependent Information: Depends on the value of the type field as further specified below in this section.

Type 1

Bit Map of VLANs: The Type Dependent Information consists of two bytes with the 12-bit starting VLAN ID N right justified (the top 4 bits are as specified above for RESV). This is followed by bytes with one bit per VLAN ID. The high order bit of the first byte is for VLAN N, the next to the highest order bit is for VLAN N+1, the low order bit of the first byte is for VLAN N+7, the high order bit of the second byte, if there is a second byte, is for VLAN N+8, and so on. If that bit is a one, the the Address Flush message applies to that VLAN. If that bit is a zero, then addresses that have been learned in that VLAN are not flushed. Note that Length MUST be at least 3. If Length is 0, 1, or 2 for a Type 1 extended Address Flush message, the message is corrupt and MUST be discarded. VLAN IDs do not wrap around. If there are enough bytes so that some bits correspond to VLAN ID 0xFFFF or higher, those bits are ignored but the message is still processed for bits corresponding to valid VLAN IDs.

Type 2

Blocks of FGLs: The Type Dependent Information consists of sets of Start.FGL and End.FGL numbers. The Address Flush information applies to the FGLs in that range, include. A single FGL is indicated by have both Start.FGL and End.FGL to the same value. If End.FGL is less than Start.FGL, considering them as unsigned integers, that block is ignored but the Address Flush message is still processed for any other blocks present. For this Type, Length MUST be a multiple of 6; if it is not, the message is considered corrup and MUST be discarded.

Type 3

List of FGLs: The Type Dependent Information consists of FGL numbers each in 3 bytes. The Address Flush message applies to those FGLs. For this Type, Length MUST be a multiple of 3; if it is not, the message is considered corrup and MUST be discarded.

Type 4

Bit Map of FGLs: The Type Dependent Information consists of three bytes with the 24-bit starting FGL N. This is followed by bytes with one bit per FGL. The high order bit of the first byte is for FGL N, the next to the highest order bit is for FGL N+1, the low order bit of the first byte is for FGL N+7, the high order bit of the second byte, if there is a second byte, is for FGL N+8, and so on. If that bit is a one, the the Address Flush message applies to that FGL. If that bit is a zero, then addresses that have been learned in that FGL are not flushed. Note that Length MUST be at least 4. If Length is 0, 1, 2, or 3 for a Type 1 extended Address Flush message, the message is corrupt and MUST be discarded. FGLs do not wrap around. If there are enough bytes so that some bits correspond to an FGL higher than 0xFFFFFFFF, those bits are ignored but the message is still processed for bits corresponding to valid

FGLs.

There is no provision for a list of VLAN IDs as there are few enough of them that an arbitrary subset of VLAN IDs can always be represented as a bit map.

3. IANA Considerations

IANA is requested to assign TBD as the Address Flush RBridge Channel Protocol number from the range of RBridge Channel protocols allocated by Standards Action [RFC7178].

The added RBridge Channel protocols registry entry on the TRILL Parameters web page is as follows:

Protocol	Description	Reference
-----	-----	-----
TBD	Address Flush	[this document]

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 [ChannelTunnel] using the RBridge Channel message payload type. The Channel Tunnel protocol can provide 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.
- [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>>.
- [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>>.

Informative References

- [RFC4762] - Lasserre, M., Ed., and V. Kompella, Ed., "Virtual Private LAN Service (VPLS) Using Label Distribution Protocol (LDP) Signaling", RFC 4762, January 2007.
- [ChannelTunnel] - Eastlake, D., M. Umair, Y. Li, "TRILL: RBridge Channel Tunnel Protocol", draft-ietf-trill-channel-tunnel, work in progress.
- [TCaware] - Y. Li, et al., "Aware Spanning Tree Topology Change on RBriges" draft-yizhou-trill-tc-awareness, 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

Yizhou Li
Huawei Technologies
101 Software Avenue,
Nanjing 210012
China

Phone: +86-25-56624629
Email: liyizhou@huawei.com

Copyright, Disclaimer, and Additional IPR Provisions

Copyright (c) 2016 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
Updates: 7178
Intended status: Proposed Standard

Donald Eastlake
Huawei
Mohammed Umair
IPinfusion
Yizhou Li
Huawei
March 18, 2016

Expires: September 1, 2016

TRILL: RBridge Channel Tunnel Protocol
<draft-ietf-trill-channel-tunnel-08.txt>

Abstract

The IETF TRILL (Transparent Interconnection of Lots of Links) protocol includes an optional mechanism (specified in RFC 7178), called RBridge Channel, for the transmission of typed messages between TRILL switches in the same campus and the transmission of such messages between TRILL switches and end stations on the same link. This document specifies two optional extensions to the RBridge Channel protocol: (1) a standard method to tunnel a variety of payload types by encapsulating them in an RBridge Channel message; and (2) a method to support security facilities for RBridge Channel messages. This document updates RFC 7178.

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 authors or 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.....	3
1.1 Terminology and Acronyms.....	3
2. Channel Tunnel Packet Format.....	5
3. Channel Tunnel Payload Types.....	8
3.1 Null Payload.....	8
3.2 Ethertyped Payload.....	8
3.2.1 Tunneled RBridge Channel Message.....	9
3.2.2 Tunneled TRILL Data Packet.....	9
3.2.3 Tunneled TRILL IS-IS Packet.....	10
3.3 Ethernet Frame.....	11
4. Security, Keying, and Algorithms.....	14
4.1 Basic Security Information Format.....	14
4.2 Authentication and Encryption Coverage.....	15
4.3 Derived Keying Material.....	17
4.4 SType None.....	17
4.5 RFC 5310 Based Authentication.....	17
4.6 DTLS Pairwise Security.....	18
5. Channel Tunnel Errors.....	20
5.1 SubERRs under ERR 6.....	20
5.2 Secure Nested RBridge Channel Errors.....	20
6. IANA Considerations.....	21
6.1 Channel Tunnel RBridge Channel Protocol Number.....	21
6.2 RBridge Channel Error Codes Subregistry.....	21
7. Security Considerations.....	22
Normative References.....	23
Informative References.....	24
Appendix Z: Change History.....	25
Acknowledgements.....	27
Authors' Addresses.....	28

1. Introduction

The IETF TRILL base protocol [RFC6325] [RFC7780] has been extended with the RBridge Channel [RFC7178] facility to support transmission of typed messages (for example BFD (Bidirectional Forwarding Detection) [RFC7175]) between two TRILL switches (RBridges) in the same campus and the transmission of such messages between RBridges and end stations on the same link. When sent between RBridges in the same campus, a TRILL Data packet with a TRILL Header is used and the destination RBridge is indicated by nickname. When sent between a RBridge and an end station on the same link in either direction a native RBridge Channel messages [RFC7178] is used with no TRILL Header and with the destination port or ports are indicated by a MAC address. (There is no mechanism to stop end stations on the same link, from sending native RBridge Channel messages to each other; however, such use is outside the scope of this document.)

This document updates [RFC7178] and specifies extensions to RBridge Channel that provide two additional facilities as follows:

- (1) A standard method to tunnel a variety of payload types by encapsulating them in an RBridge Channel message.
- (2) A method to provide security facilities for RBridge Channel messages.

Use of each of these facilities is optional, except that if Channel Tunnel is implemented there are two payload types that MUST be implemented. Both of the above facilities can be used in the same packet. In case of conflict between this document and [RFC7178], this document takes precedence.

1.1 Terminology and Acronyms

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 [RFC2119].

This document uses terminology and acronyms defined in [RFC6325] and [RFC7178]. Some of these are repeated below for convenience along with additional new terms and acronyms.

Data Label - VLAN or FGL.

DTLS - Datagram Transport Level Security [RFC6347].

FCS - Frame Check Sequence.

FGL - Fine Grained Label [RFC7172].

HKDF - Hash based Key Derivation Function [RFC5869].

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

PDU - Protocol Data Unit.

RBridge - An alternative term for a TRILL switch.

SHA - Secure Hash Algorithm [RFC6234].

Sz - Campus wide minimum link MTU [RFC6325] [RFC7780].

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

TRILL switch - A device that implements the TRILL protocol
[RFC6325], sometimes referred to as an RBridge.

2. Channel Tunnel Packet Format

The general structure of an RBridge Channel message between two TRILL switches (RBridges) in the same campus is shown in Figure 2.1 below. The structure of a native RBridge Channel message sent between an RBridge and an end station on the same link, in either direction, is shown in Figure 2.2 and, compared with the first case, omits the TRILL Header, inner Ethernet addresses, and Data Label. A Protocol field in the RBridge Channel Header gives the type of RBridge Channel message and indicates how to interpret the Channel Protocol Specific Payload [RFC7178].

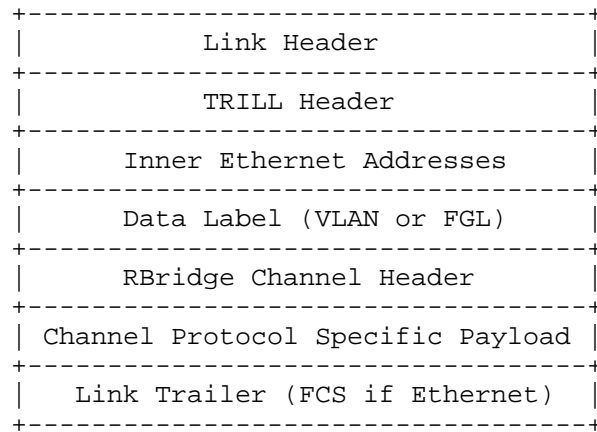


Figure 2.1 RBridge Channel Packet Structure

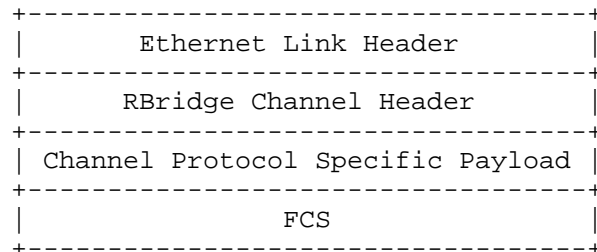


Figure 2.2 Native RBridge Channel Frame

The RBridge Channel Header looks like this:

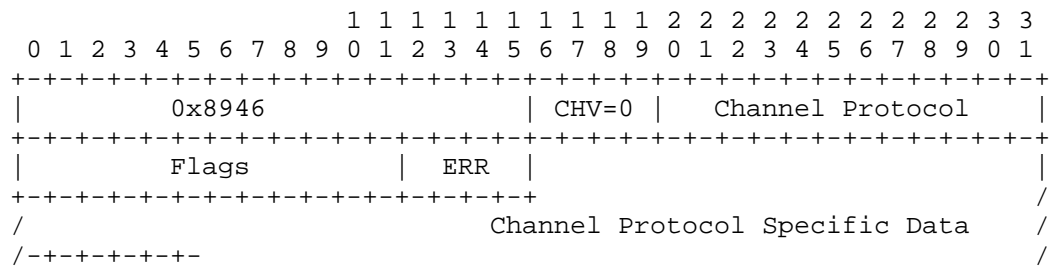


Figure 2.3 RBridge Channel Header

where 0x8946 is the RBridge Channel Ethertype and CHV is the Channel Header Version. This document is based on RBridge Channel version zero.

The extensions specified herein are in the form of an RBridge Channel protocol, the Channel Tunnel Protocol. Figure 2.4 below expands the RBridge Channel Header and Protocol Specific Payload above for the case of the Channel Tunnel Protocol.

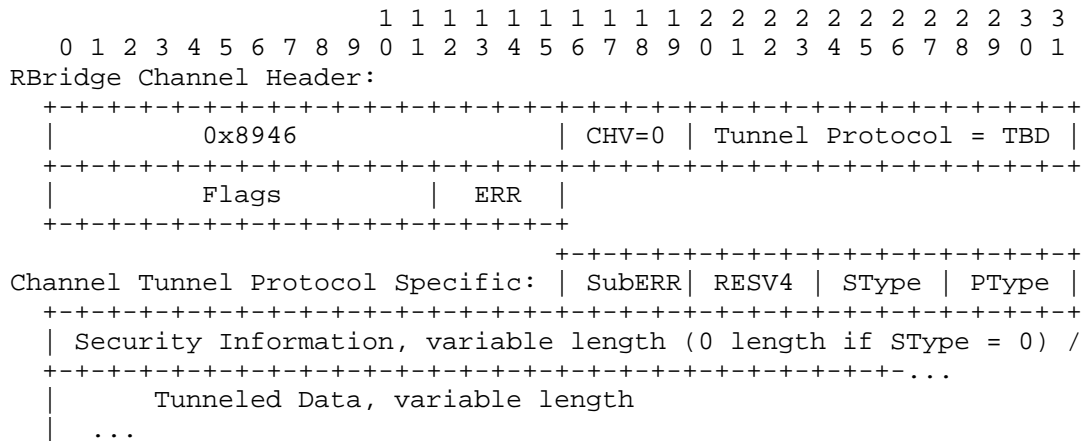


Figure 2.4 Channel Tunnel Header Structure

The RBridge Channel Header field specific to the RBridge Channel Tunnel Protocol is the Protocol field. Its contents MUST be the value allocated for this purpose (see Section 6).

The RBridge Channel Tunnel Protocol Specific Data fields are as follows:

SubERR: This field provides further details when a Channel Tunnel error is indicated in the RBridge Channel ERR field. If ERR is zero, then SubERR MUST be sent as zero and ignored on receipt. See Section 5.

RESV4: This field MUST be sent as zero. If non-zero when received, this is an error condition (see Section 5).

SType: This field describes the type of security information and features, including keying material, being used or provided by the Channel Tunnel packet. See Section 4.

PType: Payload type. This describes the tunneled data. See Section 3 below.

Security Information: Variable length information. Length is zero if SType is zero. See Section 4.

The Channel Tunnel protocol is integrated with the RBridge Channel facility. Channel Tunnel errors are reported as if they were RBridge Channel errors, using newly allocated code points in the ERR field of the RBridge Channel Header supplemented by the SubERR field.

3. Channel Tunnel Payload Types

The Channel Tunnel Protocol can carry a variety of payloads as indicated by the PType field. Values are shown in the table below with further explanation after the table.

PType	Section	Description
0		Reserved
1	3.1	Null
2	3.2	Ethertyped Payload
3	3.3	Ethernet Frame
4-14		Unassigned
15		Reserved

Table 3.1 Payload Type Values

While implementation of the Channel Tunnel protocol is optional, if it is implemented PType 1 (Null) MUST be implemented and PType 2 (Ethertyped Payload) with the RBridge Channel Ethertype MUST be implemented. PType 2 for any Ethertypes other than the RBridge Channel Ethertype MAY be implemented. PType 3 MAY be implemented.

The processing of any particular Channel Protocol message and its payload depends on meeting local security and other policy at the destination TRILL switch or end station.

3.1 Null Payload

The Null payload type (PType = 1) is intended to be used for testing or for messages such as key negotiation or the like where only security information is present. It indicates that there is no payload. Any data after the Security Information field is ignored. If the Channel Tunnel feature is implemented, Null Payload MUST be supported in the sense that an "Unsupported PType" error is not returned (see Section 5). Any particular use of the Null Payload should specify what VLAN or priority should be used when relevant.

3.2 Ethertyped Payload

A PType of 2 indicates that the payload of the Channel Tunnel message begins with an Ethertype. A TRILL switch supporting the Channel Tunnel protocol MUST support a PType of 2 with a payload beginning with the RBridge Channel Ethertype as describe in Section 3.2.1. Other Ethertypes, including the TRILL and L2-IS-IS Ethertypes as described in Section 3.2.2 and 3.2.3, MAY be supported.

3.2.1 Tunneled RBridge Channel Message

A PType of 2 with an initial RBridge Channel Ethertype indicates an encapsulated RBridge Channel message payload. A typical reason for sending an RBridge Channel message inside a Channel Tunnel message is to provide security services, such as authentication or encryption.

This payload type looks like the following:

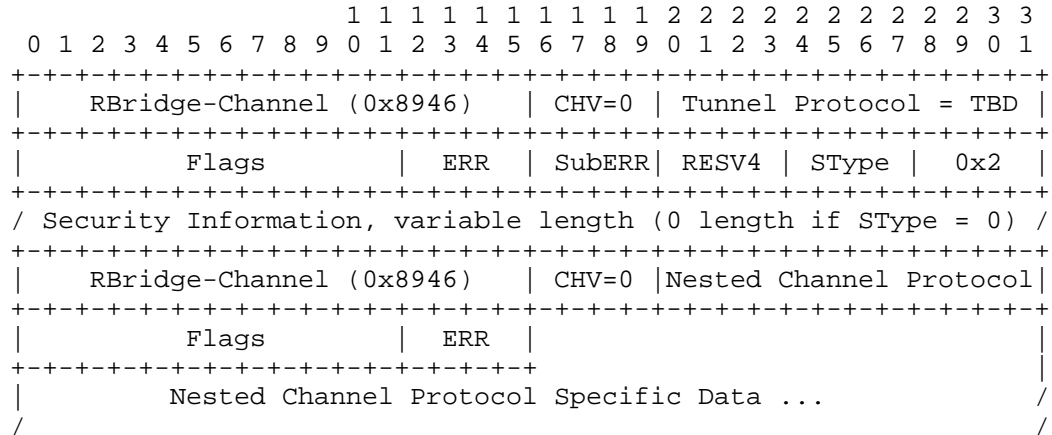


Figure 3.1 Tunneled RBridge Channel Message Structure

3.2.2 Tunneled TRILL Data Packet

A PType of 2 and an initial TRILL Ethertype indicates that the payload of the Tunnel protocol message is an encapsulated TRILL Data packet as shown in the figure below. If this Ethertype is supported for PType = 2 and the message meets local policy for acceptance, the tunneled TRILL Data packet is handled as if it had been received by the destination TRILL switch on the port where the Channel Tunnel message was received.

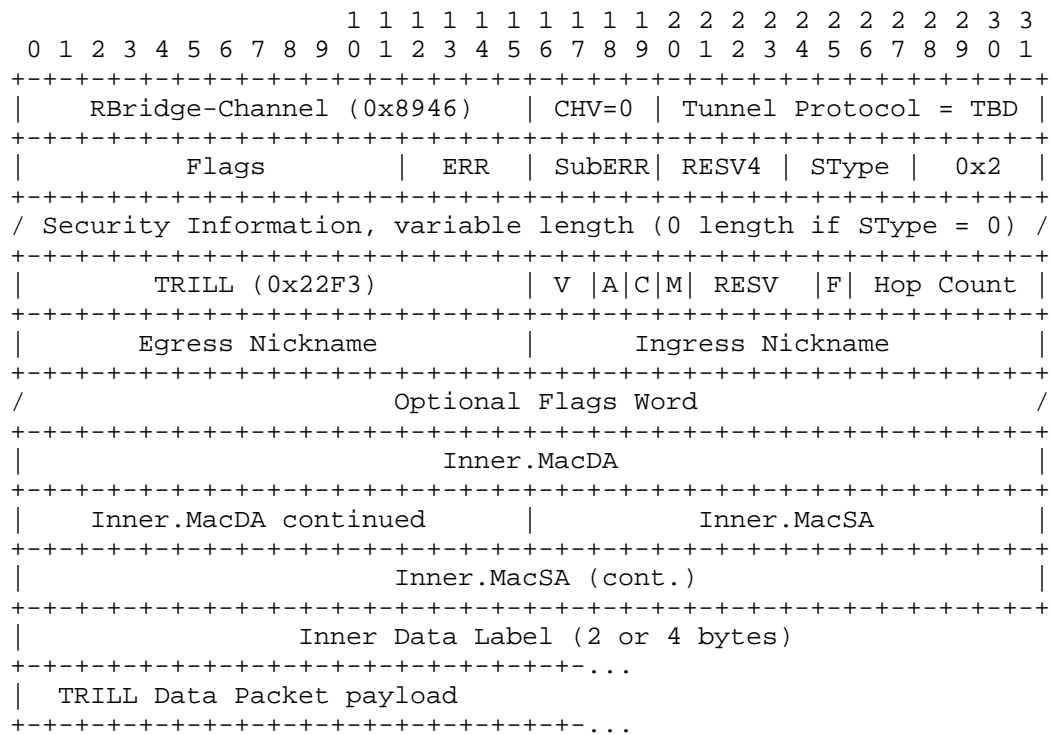


Figure 3.2 Nested TRILL Data Packet Channel Tunnel Structure

The optional flags word is only present if the F bit in the TRILL Header is one [RFC7780].

3.2.3 Tunneled TRILL IS-IS Packet

A PType of 2 and an initial L2-IS-IS Ethertype indicates that the payload of the Tunnel protocol message is an encapsulated TRILL IS-IS PDU as shown in Figure 3.3. If this Ethertype is supported for PType = 2, the tunneled TRILL IS-IS packet is processed by the destination RBridge if it meets local policy. One possible use is to expedite the receipt of a link state PDU (LSP) by some TRILL switch or switches with an immediate requirement for the link state information. A link local IS-IS PDU (Hello, CSNP, or PSNP [IS-IS]; MTU-probe or MTU-ack [RFC7176]; or circuit scoped FS-LSP, FS-CSNP or FS-PSNP [RFC7356]) would not normally be sent via this Channel Tunnel method except possibly to encrypt it since such PDUs can just be transmitted on the link and do not normally need RBridge Channel tunneling.

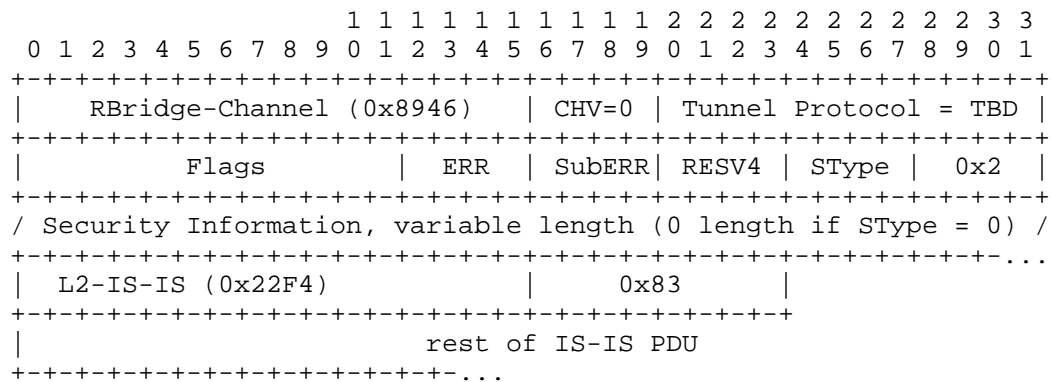


Figure 3.3 Tunneled TRILL IS-IS Packet Structure

3.3 Ethernet Frame

If PType is 3, the Tunnel Protocol payload is an Ethernet frame as might be received from or sent to an end station except that the tunneled Ethernet frame's FCS is omitted, as shown in Figure 3.4. (There is still an overall final FCS if the RBridge Channel message is being sent on an Ethernet link.) If this PType is implemented and the message meets local policy, the tunneled frame is handled as if it had been received on the port on which the Channel Tunnel message was received.

The priority of the RBridge Channel message can be copied from the Ethernet frame VLAN tag, if one is present, except that priority 7 SHOULD only be used for messages critical to establishing or maintaining adjacency and priority 6 SHOULD only be used for other important control messages.

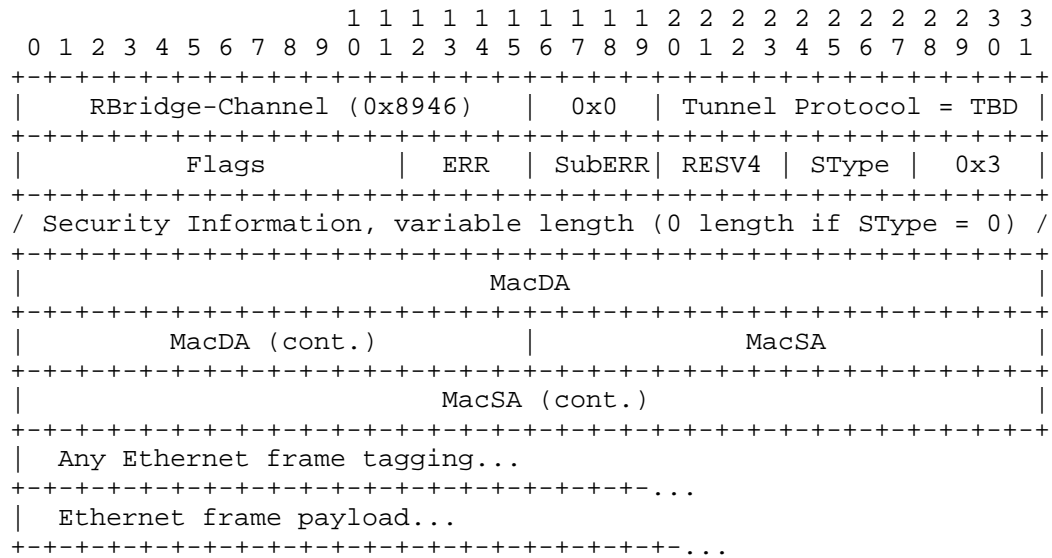


Figure 3.4 Ethernet Frame Channel Tunnel Structure

In the case of a non-Ethernet link, such as a PPP (Point-to-Point Protocol) link [RFC6361], the ports on the link are considered to have link local synthetic 48-bit MAC addresses constructed as described below. These constructed addresses MAY be used as a MacSA. If the RBridge Channel message is link local, the source TRILL switch will have the information to construct such a MAC address for the destination TRILL switch port and that MAC address MAY be used as the MacDA. By the use of such a MacSA and either such a unicast MacDA or a group addressed MacDA, an Ethernet frame can be sent between two TRILL switch ports connected by a non-Ethernet link.

These synthetic TRILL switch port MAC addresses for non-Ethernet ports are constructed as follows: 0xFEFF, the nickname of the TRILL switch used in TRILL Hellos sent on that port, and the Port ID that the TRILL switch has assigned to that port, as shown in Figure 3.5. (Both the Port ID of the port on which a TRILL Hello is sent and the nickname of the sending TRILL switch appear in the Special VLANs and Flags sub-TLV [RFC7176] in TRILL IS-IS Hellos.) The resulting MAC address has the Local bit on and the Group bit off [RFC7042]. However, since there will be no Ethernet end stations on a non-Ethernet link in a TRILL campus, such synthetic MAC addresses cannot conflict on the link with a real Ethernet port address.

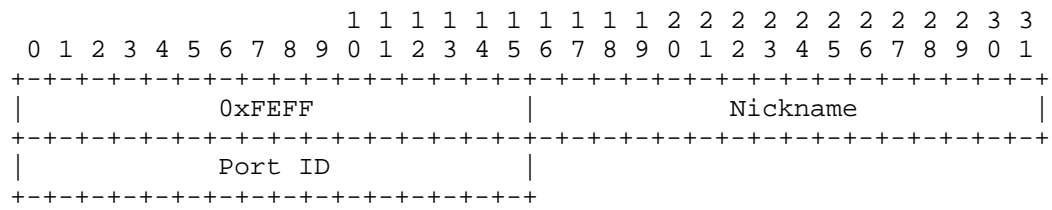


Figure 3.5 Synthetic MAC Address

4. Security, Keying, and Algorithms

Table 4.1 below gives the assigned values of the SType field and their meaning. Use of DTLS Pairwise Security (SType = 2) is RECOMMENDED. While [RFC5310] based authentication applies to both pairwise and multi-destination traffic, it provides only authentication and is generally considered not to meet current security standards, as it does not provide for key negotiation; thus, its use is NOT RECOMMENDED.

Channel Tunnel DTLS based security specified in Section 4.6 below is intended for pairwise (known unicast) use in which case the M bit in the TRILL Header would be zero and in the native RBridge Channel case (Figure 2.2) the Outer.MacDA would be individually addressed.

Multi-destination Channel Tunnel packets would be those with the M bit in the TRILL Header set to one or, in the native RBridge Channel case, the Outer.MacDA would be group addressed. However, the DTLS Pairwise Security SType can be used in the multi-destination case by serially unicasting the messages to all data accessible RBridges (or end stations in the native RBridge Channel case) in the recipient group. For TRILL Data packets, that group is specified by the Data Label; for native frames, the group is specified by the groupcast destination MAC address. It is intended to specify a true group keyed SType to secure multi-destination packets in a separate document [GroupKey].

SType	Section	Meaning
0	4.4	None
1	4.5	[RFC5310] Based Authentication
2	4.6	DTLS Pairwise Security
3-14		Available for assignment by IETF Review
15		Reserved

Table 4.1 SType Values

4.1 Basic Security Information Format

When SType is zero, there is no Security Information after the Channel Tunnel header and before the payload. For all SType values except zero, the Security Information starts with four reserved flag bits and twelve bits of remaining length as follows:

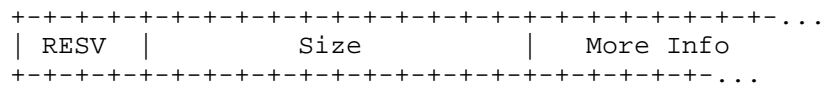


Figure 4.1 Security Information Format

The fields are as follows:

RESV: Four reserved bits that MUST be sent as zero and ignored on receipt. In the future, meanings may be assigned to these bits and those meanings may differ for different STypes.

Size: The number of bytes, as an unsigned integer, of More Info in the Security Information after the Size byte itself. Thus the maximum possible length of Security Information is 4,097 bytes for a Size of 4,095 plus 2 for the RESV and Size fields.

More Info: Additional Security Information of length Size. Contents depends on the SType.

4.2 Authentication and Encryption Coverage

As show in Figure 4.2, the area covered by Channel Tunnel authentication starts with the byte immediately after the TRILL Header optional Flag Word if it is present. Otherwise, it starts after the TRILL Header Ingress Nickname. In either case, it extends to just before the TRILL Data packet link trailer. For example, for an Ethernet packet it would extend to just before the FCS.

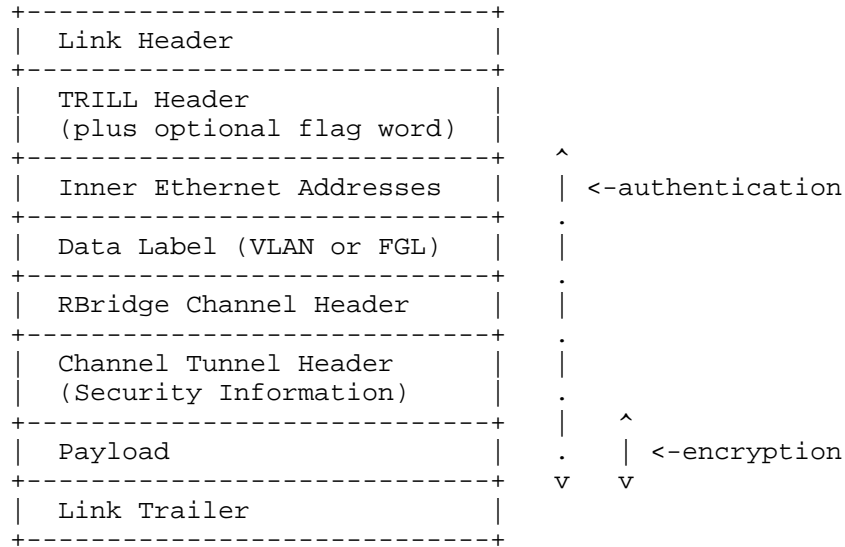


Figure 4.2. Channel Tunnel Security Coverage

Channel Tunnel authentication in the native RBridge Channel case (see Figure 4.3), is as specified in the above paragraph except that it starts with the RBridge Channel Ethertype, since there is no TRILL Header, inner Ethernet addresses, or inner Data Label.

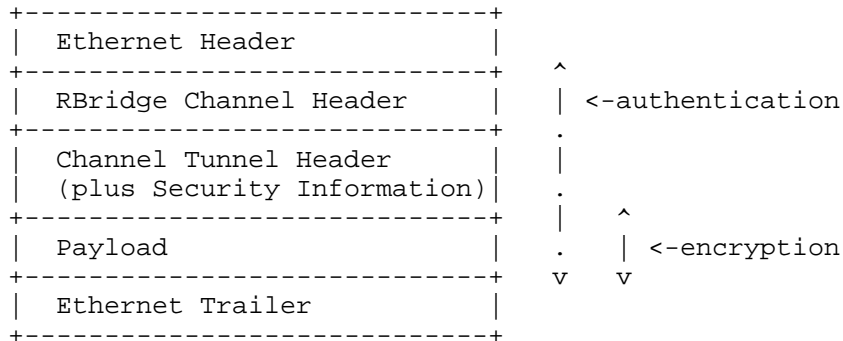


Figure 4.3. Native Channel Tunnel Security Coverage

If an authentication value is included in the More Info field shown in Section 4.1, it is treated as zero when authentication is calculated. If an authentication value is included in a payload after the security information, it is calculated as provided by the SType and security algorithms in use.

If encryption is provided, it covers the payload from right after the Channel Tunnel header Security Information through to just before the

TRILL Data packet link trailer (see Figures 4.2 and 4.3).

4.3 Derived Keying Material

In some cases, it is possible to use material derived from [RFC5310] IS-IS keying material as an element of Channel Tunnel security. In such cases, the More Info field shown in Figure 4.1 includes the two byte IS-IS Key ID to identify the keying material. It is assumed that the IS-IS keying material is of high quality. The material actually used in Channel Tunnel security is derived from the IS-IS keying material as follows:

```
Derived Material =  
    HKDF-Expand-SHA256 ( IS-IS-key, "Channel Tunnel" | 0x0S, L )
```

where "|" indicates concatenation, HKDF is as in [RFC5869], SHA256 is as in [RFC6234], IS-IS-key is the input IS-IS keying material, "Channel Tunnel" is the 14-character ASCII [RFC20] string indicated without any leading length byte or trailing zero byte, 0x0S is a single byte where S is the SType for which this key derivation is being used and the upper nibble is zero, and L is the length of output-derived material needed.

Whenever IS-IS keying material is being used as above, the underlying [RFC5310] keying material might expire or become invalidated. At the time of or before such expiration or invalidation, the use Derived Material from the IS-IS keying material MUST cease. Continued security may depend on using new derived material from currently valid [RFC5310] keying material.

4.4 SType None

No security services are being invoked. The length of the Security Information field (see Figure 2.4) is zero.

4.5 RFC 5310 Based Authentication

The Security Information (see Figure 2.4) is the RESV and Size fields specified in Section 4.1 with the value of the [RFC5310] Key ID and Authentication Data, as shown in Figure 4.4.

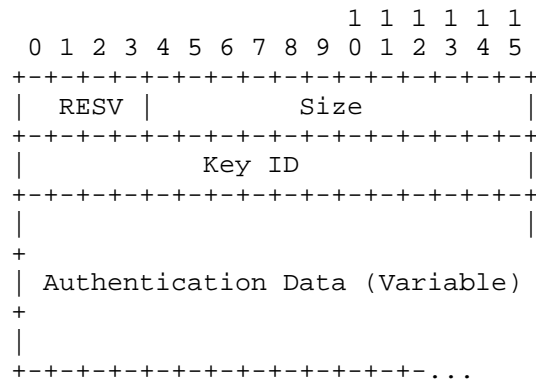


Figure 4.4 SType 1 Security Information

- o RESV: Four bits that MUST be sent as zero and ignored or receipt.
- o Size: Set to 2 + the size of Authentication Data in bytes.
- o Key ID: specifies the keying value and authentication algorithm that the Key ID specifies for TRILL IS-IS LSP [RFC5310] Authentication TLVs. The keying material actually used is derived as shown in Section 4.3.
- o Authentication Data: The authentication data produced by the derived key and algorithm associated with the Key ID acting on the packet as specified in Section 4.2. Length of the authentication data depends on the algorithm.

While RBridges, which are IS-IS routers, can reasonable be expected to hold [RFC5310] keying, so that this SType can be used for RBridge Channel messages, how end stations might come to hold [rfc5310] keying is beyond the scope of this document. Thus this SType might not be applicable to native RBridge Channel messages.

4.6 DTLS Pairwise Security

DTLS supports key negotiation and provides both encryption and authentication. The Channel Tunnel DTLS [RFC6347] SType uses a negotiated DTLS version that MUST NOT be less than 1.2.

When DTLS pairwise security is used, the entire payload of the Channel Tunnel packet, starting just after the Security Information and ending just before the link trailer, is one or more DTLS records [RFC6347]. As specified in [RFC6347], DTLS records MUST be limited by the path MTU, in this case so each record fits entirely within a single Channel Tunnel message. A minimum path MTU can be determined

from the TRILL campus wide minimum MTU Sz, which will not be less than 1470 bytes, by allowing for the TRILL Data packet, Channel Tunnel, and DTLS framing overhead. With this SType, the security information provided before the DTLS record(s) is 0, as shown in Figure 4.5, because all the security information is in the payload area.

The DTLS Pairwise keying is set up between a pair of RBridges independent of Data Label using messages of a priority configurable at the RBridge level which defaults to priority 6. DTLS messages other than application_data can be encapsulated in the Channel Tunnel protocol with a TRILL Header using any Data Label. Actual application_data sent with Channel Tunnel using this SType should use the Data Label and priority as specified for that application_data. The PType indicates the nature of the application_data.

TRILL switches that support the Channel Tunnel DTLS SType MUST support the use of pre-shared keys for DTLS. If the psk_identity (see [RFC4279]) is two bytes, it represents, as a pre-shared key to be used in the DTLS negotiation, the value derived as shown in Section 4.3 from the key associated with that psk_identity as a [RFC5310] Key ID. A psk_identity with a length other than two bytes MAY be used to indicate other implementation dependent pre-shared keys.

```

+++++
| RESV |          0          |
+++++

```

Figure 4.5 DTLS Channel Tunnel Security Info

TRILL switches that implement the Channel Tunnel DTLS SType MAY support the use of certificates for DTLS but certificate size may be limited by the DTLS requirement that each record fit within a single message.

5. Channel Tunnel Errors

RBridge Channel Tunnel Protocol errors are reported like RBridge Channel level errors. The ERR field is set to one of the following error codes:

ERR	Meaning
---	-----
6	Unknown or unsupported field value
7	Authentication failure
8	Error in nested RBridge Channel message

Table 5.1 Additional ERR Values

5.1 SubERRs under ERR 6

If the ERR field is 6, the SubERR field indicates the problematic field or value as show in the table below.

SubERR	Meaning (for ERR = 6)
-----	-----
0	Reserved
1	Non-zero RESV4 nibble
2	Unsupported SType
3	Unsupported PType
4	Unknown Key ID
5	Unknown Ethertype with PType = 2

Table 5.2 SubERR values under ERR 6

5.2 Secure Nested RBridge Channel Errors

If
 a Channel Tunnel message is sent with security and with a payload type (PType) indicating a nested RBridge Channel message
 and
 there is an error in the processing of that nested message that results in a return RBridge Channel message with a non-zero ERR field,
 then that returned message SHOULD also be nested in an Channel Tunnel message using the same type of security. In this case, the ERR field in the Channel Tunnel envelope is set to 8 indicating that there is a nested error in the message being tunneled back.

6. IANA Considerations

This section lists IANA Considerations.

6.1 Channel Tunnel RBridge Channel Protocol Number

IANA is requested to assign TBD as the RBridge Channel protocol number for the "Channel Tunnel" protocol from the range assigned by Standards Action.

The added RBridge Channel protocols registry entry on the TRILL Parameters web page is as follows:

Protocol	Description	Reference
-----	-----	-----
TBD	Channel Tunnel	[this document]

6.2 RBridge Channel Error Codes Subregistry

IANA is requested to create a "RBridge Channel Error Codes" subregistry under the "RBridge Channel Protocols" registry. The header information is as follows:

Registration Procedures: IETF Review References: [RFC7178] [this document]

The subregistry is to have columns and entries as follows:

Code	Meaning	Reference
----	-----	-----
[populate rows for codes 0 through 5 from Section xxx of [RFC7178] with reference [RFC7178]]		
[populate rows for codes 6 through 8 from Table 5.1 of this document with reference [this document]]		
9-15	Unassigned	
16	Reserved	

7. Security Considerations

The RBridge Channel Tunnel facility has potentially positive and negative effects on security.

On the positive side, it provides optional security that can be used to authenticate and/or encrypt RBridge Channel messages. Some RBridge Channel message payloads, such as BFD [RFC7175], provide their own security but where this is not true, consideration should be given, when specifying an RBridge Channel protocol, to recommending or requiring use of the security features of the Channel Tunnel protocol.

On the negative side, the optional ability to tunnel various payload types and to tunnel them between TRILL switches and to and from end stations can increase risk unless precautions are taken. The processing of decapsulating Tunnel Protocol payloads is not a good place to be liberal in what you accept. This is because the tunneling facility makes it easier for unexpected messages to pop up in unexpected places in a TRILL campus due to accidents or the actions of an adversary. Local policies should generally be strict and only process payload types required and then only with adequate authentication for the particular circumstances.

See the first paragraph of Section 4 for recommendations on SType usage. See [RFC7457] for Security Considerations of DTLS for security.

If IS-IS authentication is not being used, then [RFC5310] keying information would not normally be available but that presumably represents a judgment by the TRILL campus operator that no security is needed.

See [RFC7178] for general RBridge Channel Security Considerations and [RFC6325] for general TRILL Security Considerations.

Normative 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.
- [RFC20] - Cerf, V., "ASCII format for network interchange", STD 80, RFC 20, DOI 10.17487/RFC0020, October 1969, <<http://www.rfc-editor.org/info/rfc20>>.
- [RFC2119] - BBradner, 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>>.
- [RFC4279] - Eronen, P., Ed., and H. Tschofenig, Ed., "Pre-Shared Key Ciphersuites for Transport Layer Security (TLS)", RFC 4279, DOI 10.17487/RFC4279, December 2005, <<http://www.rfc-editor.org/info/rfc4279>>.
- [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, 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>>.
- [RFC6347] - Rescorla, E. and N. Modadugu, "Datagram Transport Layer Security Version 1.2", RFC 6347, January 2012, <<http://www.rfc-editor.org/info/rfc6347>>.
- [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, 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>>.
- [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>>.
- [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>>.

Informative References

- [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, August 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.
- [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.
- [RFC7457] - Sheffer, Y., Holz, R., and P. Saint-Andre, "Summarizing Known Attacks on Transport Layer Security (TLS) and Datagram TLS (DTLS)", RFC 7457, February 2015, <<http://www.rfc-editor.org/info/rfc7457>>.
- [GroupKey] - D. Eastlake et al, "Group Keying Protocol", draft-ietf-trill-group-keying, work in progress.

Appendix Z: Change History

From -00 to -01

1. Fix references for RFCs published, etc.
2. Explicitly mention in the Abstract and Introduction that this document updates [RFC7178].
3. Add this Change History Appendix.

From -01 to -02

1. Remove section on the "Scope" feature as mentioned in <http://www.ietf.org/mail-archive/web/trill/current/msg06531.html>
2. Editorial changes to IANA Considerations to correspond to draft-leiba-cotton-iana-5226bis-11.txt.
3. Improvements to the Ethernet frame payload type.
4. Other Editorial changes.

From -02 to -03

1. Update TRILL Header to correspond to [RFC7780].
2. Remove a few remnants of the "Scope" feature that was removed from -01 to -02.
3. Substantial changes to and expansion of Section 4 including adding details of DTLS security.
4. Updates and additions to the References.
5. Other minor editorial changes.

From -03 to -04

1. Add SType for [RFC5310] keying based security that provides encryption as well as authentication.
2. Editorial improvements and fixes.

From -04 to -05

1. Primary change is collapsing the previous PTypes 2, 3, and 4 for RBridge Channel message, TRILL Data, and TRILL IS-IS into one by including the Ethertype. Previous PType 5 is renumbered as 3.

2. Add Channel Tunnel Crypto Suites to IANA Considerations
3. Add some material to Security Considerations,
4. Assorted Editorial changes.

From -05 to -06

Fix editorials found during WG Last Call.

From -06 to -07

Minor editorial changes resulting for Shepherd review.

From -07 to -08

Move group keyed security out of the draft. Simplify and improve remaining security provisions.

Acknowledgements

The contributions of the following are hereby gratefully acknowledged:

Susan Hares, Gayle Noble, Yaron Sheffer

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

Authors' Addresses

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

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

Mohammed Umair
IPinfusion

EMail: mohammed.umair2@gmail.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) 2016 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
Updates: 6325, 7177

Donald Eastlake
Mingui Zhang
Huawei
Ayan Banerjee
Cisco
Vishwas Manral
Ionos
February 28, 2016

Expires: August 27, 2016

TRILL: Multi-Topology
<draft-ietf-trill-multi-topology-01.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. This document updates RFC 6325 and RFC 7177.

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 (Updates to RFC 7177).....	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.....	22
Authors' Addresses.....	23

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.

This document updates [RFC7177] as specified in Section 3.1. This document updates numerous aspects of [RFC6325] including changing routing (Sections 3.3 and 3.4), address learning (Section 5.1), and distribution tree construction (Section 3.4), to take multi-topology into account.

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.

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.

FGL - Fine-Grained Labeling or Fine-Grained Labeled or Fine-Grained Label [RFC7172]. By implication, an "FGL TRILL switch" does not support 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 Tunnelled 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].

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 received 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

Commonly, 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.
1	Capable of inserting an explicit topology label in TRILL Data packets sent and tolerant of such labels in received TRILL Data packets. Such a port is capable, for all of the topologies it supports, of determining TRILL Data packet topology without an explicit label. Thus it does not require such a label in received TRILL Data packets. On receiving a

packet whose explicit topology label differs from the port's topology determination for that packet, the TRILL switch MUST discard the packet.

- 2/3 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.)

A TRILL switch advertising in a Hello on Port P support for topology T but not advertising in those Hellos that it requires explicit topology labeling is assumed to have the ability and configuration to correctly classify TRILL Data packets into topology T by examination of those TRILL Data packets and/or by using the fact that they are arriving at port P.

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. If a label is not so required but all other TRILL switches on that link support explicit topology labeling, then such a label MAY be included.

2.4.2 The Explicit Topology Label

The MT label is structured as follows:

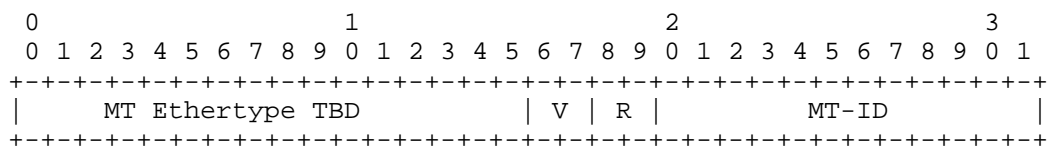


Figure 1. MT Label

where the fields are 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 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. {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.

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
+-----+-----+-----+-----+-----+-----+-----+-----+
| 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
+-----+-----+-----+-----+-----+-----+-----+-----+
| 0x893B                                     | PRI |D|  FGL High Part      |
+-----+-----+-----+-----+-----+-----+-----+-----+
| 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          |
+-----+-----+-----+-----+-----+-----+-----+-----+
| 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          |
+-----+-----+-----+-----+-----+-----+-----+-----+
| 0x893B                                     | PRI |D|  FGL High Part      |
+-----+-----+-----+-----+-----+-----+-----+-----+
| 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 an obvious extension.

3. TRILL Multi-Topology Adjacency and Routing

Routing calculations in IS-IS are based on adjacency. Section 3.1 specifies multi-topology updates to the TRILL adjacency specification [RFC7177]. 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 (Updates to RFC 7177)

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 reasons, 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 is preferred to one 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 [RFC6439] 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 was 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>>.

Informative References

- [RFC6439] - Perlman, R., Eastlake, D., Li, Y., Banerjee, A., and F. Hu, "Routing Bridges (RBridges): Appointed Forwarders", RFC 6439, November 2011.
- [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 suggestions of the following are gratefully acknowledged:

TBD

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

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 implicit determination but extends this for TRILL 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

Vishwas Manral
Ionos Corp.
4100 Moorpark Ave.
San Jose, CA 95117 USA

Email: vishwas@ionosnetworks.com

Copyright, Disclaimer, and Additional IPR Provisions

Copyright (c) 2016 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
D. Eastlake
Huawei
R. Perlman
EMC
M. Cullen
Painless Security
H. Zhai
JIT
February 16, 2016

Expires: August 19, 2016

Transparent Interconnection of Lots of Links (TRILL)
Single Area Border RBridge Nickname for Multilevel
draft-ietf-trill-multilevel-single-nickname-01.txt

Abstract

A major issue in multilevel TRILL is how to manage RBridge nicknames. In this document, the area border RBridge uses 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 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) 2016 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. Acronyms and Terminology	3
3. Nickname Handling on Border RBridges	3
3.1. Actions on Unicast Packets	4
3.2. Actions on Multi-Destination Packets	5
4. Per-flow Load Balancing	6
4.1. Ingress Nickname Replacement	6
4.2. Egress Nickname Replacement	7
5. Protocol Extensions for Discovery	7
5.1. Discovery of Border RBridges in L1	7
5.2. Discovery of Border RBridge Sets in L2	7
6. One Border RBridge Connects Multiple Areas	8
7. E-L1FS/E-L2FS Backwards Compatibility	9
8. Security Considerations	9
9. IANA Considerations	9
9.1. TRILL APPsub-TLVs	9
10. References	10
10.1. Normative References	10
10.2. Informative References	10
Appendix A. Clarifications	11
A.1. Level Transition	11
Author's Addresses	12

1. Introduction

TRILL multilevel techniques are designed to improve TRILL scalability issues. As described in [MultiL], there have been two proposed approaches. One approach, which is referred 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. The other approach, which is referred as the "aggregated nickname" approach, involves assigning nicknames to the areas, and allowing nicknames to be reused

in different areas, by having the border TRILL switches rewrite the nickname fields when entering or leaving an area.

The approach specified in this document is different from both "unique nickname" and "aggregated nickname" approach. 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 the L1 areas. Each L1 area is denoted by the group of all nicknames of those border RBridges of the area. For this approach, nicknames in L2 MUST be unique but nicknames inside different L1 areas MAY be reused. 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.

2. Acronyms and Terminology

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

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

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].

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.

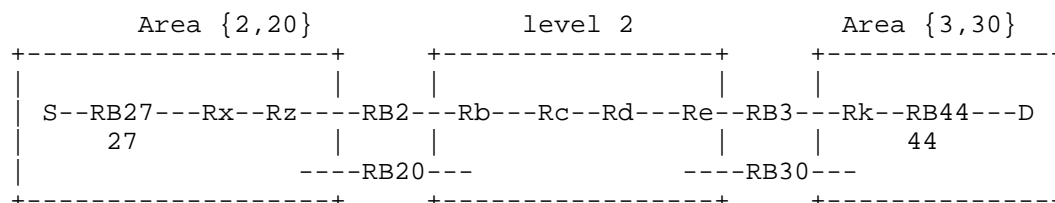


Figure 1: An Example Topology for TRILL Multilevel

In Figure 1, RB2, RB20, RB3 and RB30 are area border TRILL switches (RBridges). Their nicknames are 2, 20, 3 and 30 respectively. 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 is 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 instance in area {2,20}, that they are attached to all those area nicknames, including {3,30}. 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, so replaces 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 an area nickname selected from {3,30}. See Section 4 for the detail of the selection method. Here, suppose nickname 3 is used.) 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). (The ingress nickname MAY be replaced with an area nickname selected from {2,20}. See Section 4 for the detail of the selection method. Here, suppose nickname 2 is selected.) 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 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. All border RBridges of an area SHOULD agree on a pseudorandom algorithm and locally determine the DBRB as they do in the "Per-flow Load Balancing" section. It's also possible to implement a certain election protocol to elect the DBRB. However, such kind of implementations are out the scope of this document.

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, 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 a 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 root RBridge nickname. 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, so replaces 27 with 2. RB2 also needs to replace the egress RBridge nickname with the L2 tree root RBridge nickname, say 2. In order to accommodate return traffic, RB2 records that S is attached to nickname 27 and SHOULD use ESADI protocol 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 packet is flooded on the Level 2 tree to reach 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, say 3, of the distribution tree of L1 area {3,30}. (Here, the ingress nickname MAY be replaced with an area nickname selected from {2,20} 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. In that case, RB3 must turn the packet into a multi-destination packet and floods it on the distribution tree of 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.

4. Per-flow Load Balancing

Area border RBridges perform ingress/egress nickname replacement when they transition TRILL data packets between Level 1 and Level 2. This nickname replacement enables the per-flow load balance which is specified as follows.

4.1. Ingress Nickname Replacement

When a TRILL data packet from other areas arrives at an area border RBridge, this RBridge MAY select one area nickname of the ingress to replace the ingress nickname of the packet. The selection is simply based on a pseudorandom algorithm as defined in Section 5.3 of [RFC7357]. With the random ingress nickname replacement, the border RBridge actually achieves a per-flow load balance for returning traffic.

All area border RBridges in 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 RBridge will see one source MAC address flip flopping among multiple ingress RBridges.

4.2. Egress Nickname Replacement

When a TRILL data packet originated from the area arrives at an area border RBridge, this RBridge MAY select one area nickname of the egress to replace the egress nickname of the packet. By default, it SHOULD choose the egress area border RBridge with the least cost route to reach. The pseudorandom algorithm as defined in Section 5.3 of [RFC7357] may be used as well. In that case, however, the ingress area border RBridge may take the non-least-cost Level 2 route to forward the TRILL data packet to the egress area border RBridge.

5. Protocol Extensions for Discovery

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 [RFC7180bis] as an APPsub-TLV of the TRILL GENINFO-TLV. Through listening to 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 [RFC7180bis] 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 [MultiL], 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 reach-ability 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 all these areas.

Nicknames used within one of these 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 ought not occur. Now suppose a TRILL data packet arrives at the area border nickname of RB1. For a unicast packet, RB1 can lookup 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: suppose RB1 is not the DBRB, RB1 will not transition the packet; otherwise, RB1 is the DBRB,

- if this packet is originated from an area out of the connected areas, RB1 should replicate this packet and flood it on the proper Level 1 trees of all the areas in which it acts as the DBRB.
- if the packet is originated from one of the connected areas, RB1 should replicate the packet it receives from the Level 1 tree and flood 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 may also receive the replication of the packet from the Level 2 tree. This replication must be dropped by RB1.

7. E-L1FS/E-L2FS Backwards Compatibility

All Level 2 RBridges MUST support E-L2FS [RFC7356] [rfc7180bis]. 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 either E-L1FS or E-L2FS cannot serve as area border RBridges but they can well appear in an L1 area acting as non-area-border RBridges.

8. 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]. This document raises no new security issues for IS-IS.

9. IANA Considerations

9.1. TRILL APPsub-TLVs

IANA is requested to allocate two new types under the TRILL GENINFO TLV [RFC7357] 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]

10. References

10.1. 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, <<http://www.rfc-editor.org/info/rfc2119>>.
- [RFC6325] Perlman, R., Eastlake 3rd, D., Dutt, D., Gai, S., and A. Ghanwani, "Routing Bridges (RBriges): Base Protocol Specification", RFC 6325, DOI 10.17487/RFC6325, July 2011, <<http://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, <<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, DOI 10.17487/RFC7357, September 2014, <<http://www.rfc-editor.org/info/rfc7357>>.

10.2. Informative References

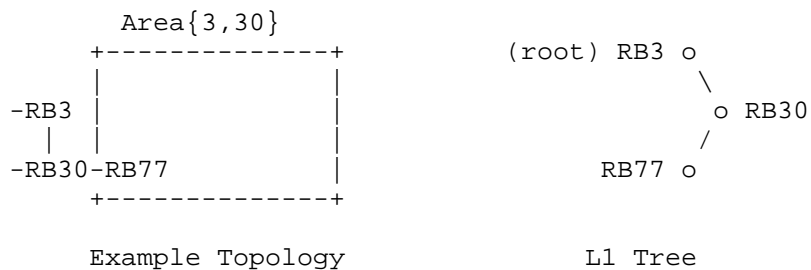
- [IS-IS] International Organization for Standardization, ISO/IEC 10589:2002, "Information technology -- Telecommunications and information exchange between systems -- Intermediate System to Intermediate System intra-domain routing information exchange protocol for use in conjunction with the protocol 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, <<http://www.rfc-editor.org/info/rfc5310>>.
- [RFC7180bis] D. Eastlake, M. Zhang, et al, "TRILL: Clarifications, Corrections, and Updates", draft-ietf-trill-rfc7180bis, work in progress.
- [MultiL] Perlman, R., Eastlake, D., et al, "Alternatives for Multilevel TRILL (Transparent Interconnection of Lots of Links)", draft-ietf-trill-rbridge-multilevel, work in progress.

Appendix A. Clarifications

A.1. Level Transition

It's possible that an L1 RBridge is only reachable from a non-DBRB 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 will 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.

Author's Addresses

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

Email: zhangmingui@huawei.com

Donald E. Eastlake, 3rd
Huawei Technologies
155 Beaver Street
Milford, MA 01757
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: <http://www.painless-security.com>

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

Email: honjun.zhai@tom.com

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.

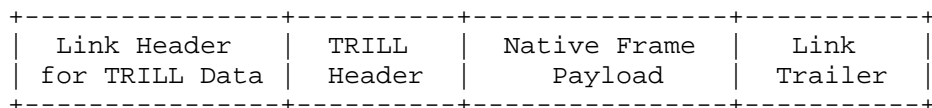
4. TRILL Packet Formats

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.

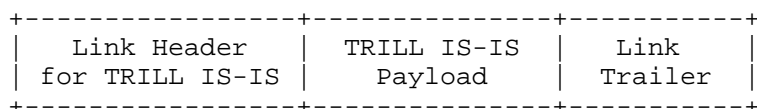
4.1 General Packet Formats

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.

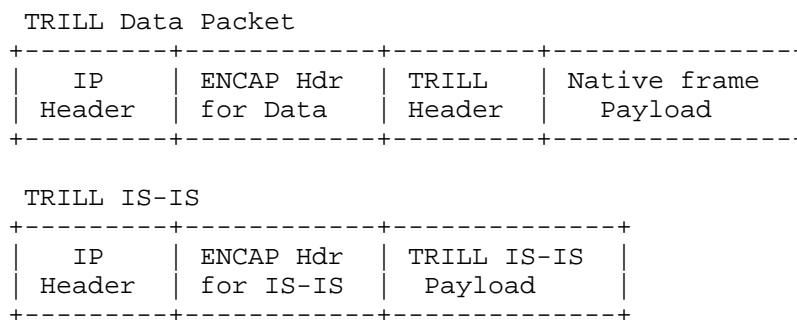
4.2 General TRILL Over IP Packet Formats

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

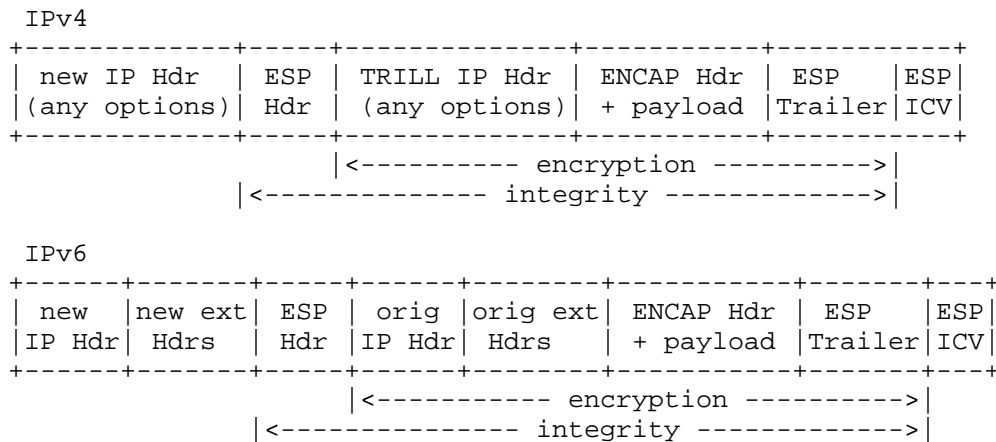
When TRILL over IP link security is not being used, a TRILL over IP packet on the wire looks like the following:



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:

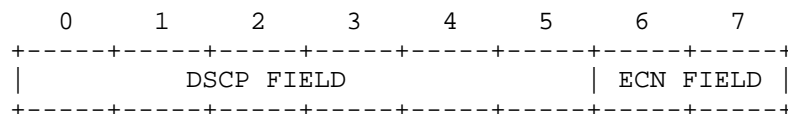


As shown above, IP Header options are considered part of the IPv4 Header but are extensions ("ext") of the IPv6 Header. For further information on the IPsec ESP Hdr, Trailer, and ICV, see [RFC4303] and Section 7. "ENCAP Hdr + payload" is the encapsulation header (Section 5) and TRILL data or IS-is payload,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

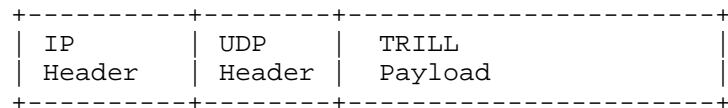
5.3 Broadcast Link Encapsulation Considerations

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.)

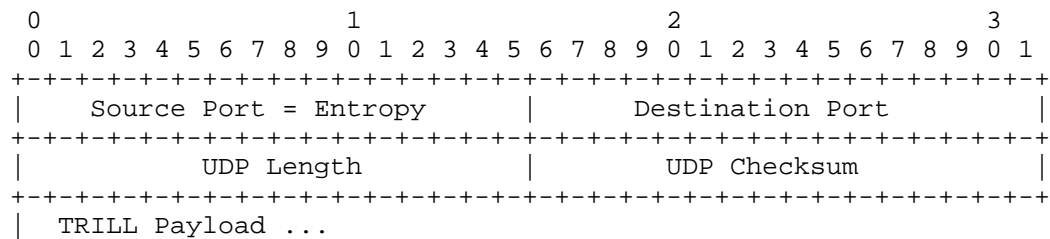
It is possible for the Hellos from a TRILL over IP port P1 to establish adjacency with multiple other TRILL over IP ports (P2, P3, ...) on broadcast link. In a well configured network one would expect all of the IP ports involved to support the same encapsulation(s); but, if P1 supports multiple encapsulations, 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

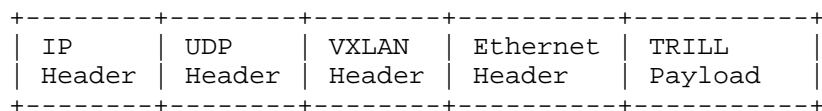
UDP Length - as specified in [RFC0768]

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
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

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

EMail: zhangmingui@huawei.com

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

Email: 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
February 12, 2016

Expires: August 11, 2016

Alternatives for Multilevel TRILL
(Transparent Interconnection of Lots of Links)
<draft-ietf-trill-rbridge-multilevel-01.txt>

Abstract

Extending TRILL to multiple levels has challenges that are not addressed by the already-existing capability of IS-IS to have multiple levels. One issue is with the handling of multi-destination packet distribution trees. Another issue is with TRILL switch nicknames. There have been two proposed approaches. One approach, which we refer 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. The other approach, which we refer to as the "aggregated nickname" approach, involves hiding the nicknames within areas, allowing nicknames to be reused in different areas, by having the border TRILL switches rewrite the nickname fields when entering or leaving an area. Each of those approaches has advantages and disadvantages. This informational document suggests allowing a choice of approach in each area. This allows the simplicity of the unique nickname approach in installations in which there is no danger of running out of nicknames and allows the complexity of hiding the nicknames in an area to be phased into larger installations on a per-area basis.

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 TRILL Scalability Issues.....	4
1.2 Improvements Due to Multilevel.....	5
1.3 Unique and Aggregated Nicknames.....	6
1.3 More on Areas.....	6
1.4 Terminology and Acronyms.....	7
2. Multilevel TRILL Issues.....	8
2.1 Non-zero Area Addresses.....	9
2.2 Aggregated versus Unique Nicknames.....	9
2.2.1 More Details on Unique Nicknames.....	10
2.2.2 More Details on Aggregated Nicknames.....	11
2.2.2.1 Border Learning Aggregated Nicknames.....	12
2.2.2.2 Swap Nickname Field Aggregated Nicknames.....	14
2.2.2.3 Comparison.....	14
2.3 Building Multi-Area Trees.....	15
2.4 The RPF Check for Trees.....	15
2.5 Area Nickname Acquisition.....	16
2.6 Link State Representation of Areas.....	16
3. Area Partition.....	18
4. Multi-Destination Scope.....	19
4.1 Unicast to Multi-destination Conversions.....	19
4.1.1 New Tree Encoding.....	20
4.2 Selective Broadcast Domain Reduction.....	20
5. Co-Existence with Old TRILL switches.....	22
6. Multi-Access Links with End Stations.....	23
7. Summary.....	24
8. Security Considerations.....	25
9. IANA Considerations.....	25
Normative References.....	26
Informative References.....	26
Acknowledgements.....	28
Authors' Addresses.....	29

1. Introduction

The IETF TRILL (Transparent Interconnection of Lot of Links or Tunneled Routing in the Link Layer) protocol [RFC6325] [RFC7177] 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 VLANs and multicast groups. Devices that implement TRILL are called TRILL Switches or RBridges.

Familiarity with [IS-IS], [RFC6325], and [rfc7180bis] is assumed in this document.

1.1 TRILL Scalability Issues

There are multiple issues that 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,
6. the traffic due to upper layer protocols use of broadcast and multicast, and
7. the size of the end node learning table (the table that remembers (egress TRILL switch, label/MAC) pairs).

Extending TRILL IS-IS to be multilevel (hierarchical) helps with all but the last of these issues.

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 solves the first four scalability issues described above, namely,

1. the routing computation load,
2. the volatility of the 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.

Problem #6 in Section 1.1, namely, the traffic due to upper layer protocols use of broadcast and multicast, can be addressed by introducing a locally-scoped multi-destination delivery, limited to an area or a single link. See further discussion in Section 4.2.

Problem #5 in Section 1.1, namely, the limit of the number of TRILL switches, due to the 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 design in this document allows a campus with a mixture of unique-nickname areas, and aggregated-nickname areas. Nicknames must be unique across all Level 2 and unique-nickname area TRILL switches, whereas nicknames inside an aggregated-nickname area are visible only inside the area. Nicknames inside an aggregated-nickname area must 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 other aggregated-nickname areas.

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 claim, in their LSP inside area A1, which nicknames (or nickname ranges) are not available for choosing as nicknames by area A1 TRILL switches.

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 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.

1.3 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. Although the area address had other purposes in CLNP (Connectionless Network Layer Protocol, IS-IS was originally designed for CLNP/DECnet), for TRILL the only purpose of the area address would be to avoid accidentally interconnecting areas.

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 this document.

1.4 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 Tunnelled Routing in the Link Layer [RFC6325]

TRILL switch - a device that implements the TRILL protocol [RFC6325], 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 nonzero 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] [rfc7180bis] 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 would not be confused with TRILL IS-IS. However, TRILL uses other techniques to avoid such confusion, such as different multicast addresses and Ethertypes on Ethernet [RFC6325], different PPP (Point-to-Point Protocol) codepoints 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 nonzero 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.

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 a block of nicknames. Either some TRILL switches would need to announce that all the nicknames other than that block 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 nicknames, and all TRILL switches in the area would need to understand that new TLV. An example of the second approach is given in [NickFlags].

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. The information could be encoded as ranges of nicknames to make this somewhat manageable [NickFlags]; however, a new TLV for announcing nickname ranges would not be intelligible to old TRILL switches.

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 with 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. One method might be to expand nicknames to 24 bits; however, that technique would require TRILL

message format 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 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 in Level 2, the ingress nickname is the nickname of the ingress TRILL switch's area while the egress nickname is either the 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.

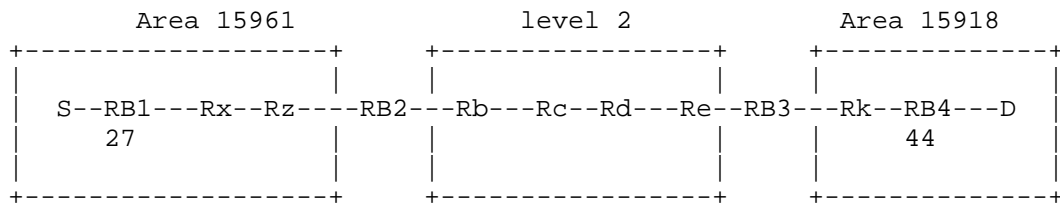
There are two variations of the aggregated nickname approach. The first is the Border Learning approach, which is described in Section 2.2.2.1. The second is the Swap Nickname Field approach, which is

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 case RB3 is not the Designated transitioner between Level 2 and its area for that multi-destination packet, but was on the unicast path, that another border TRILL switch in that area 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

As a variant, two additional fields could exist in TRILL Data packets we call the "ingress swap nickname field" and the "egress swap nickname field". 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 more extensive changes to non-border TRILL switches. In particular they 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.

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 [DraftCMT]. 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 [DraftCMT], or the 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. It is probably simpler to allocate a block of nicknames (say, the top 4000) to be area addresses, and not used by any TRILL switches.

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 acquires 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, a unicast packet might be known at the Level 1 to Level 2 transition, be forwarded as a unicast packet to the least cost border TRILL switch advertising connectivity to the destination area, but turn out to have an unknown destination { MAC, Data Label } pair when it arrives at that border TRILL switch.

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 through a flag in the packet or through distributing it on a tree that is restricted to the area, or other techniques (see Section 4).

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, would 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.

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.

6. Multi-Access Links with End Stations

Care must be taken, in the case where there are multiple TRILL switches on a link with end stations, that only one TRILL switch ingress/egress any given data packet from/to the end nodes. With existing, single level TRILL, this is done by electing a single Designated RBridge per link, which appoints a single Appointed Forwarder per VLAN [RFC7177] [RFC6439]. But suppose there are two (or more) TRILL switches on a link in different areas, say RB1 in area 1000 and RB2 in area 2000, and that the link contains end nodes. If RB1 and RB2 ignore each other's Hellos then they will both ingress/egress end node traffic from the link.

A simple rule is to use the TRILL switch or switches having the lowest numbered area, comparing area numbers as unsigned integers, to handle native traffic. This would automatically give multilevel-ignorant legacy TRILL switches, that would be using area number zero, highest priority for handling end stations, 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 in any case for legacy TRILL switches using area number zero.

Any of these techniques require multilevel aware RBridges to take actions based on Hellos from RBridges in other areas even though they will not form an adjacency with such RBridges.

7. Summary

This draft discusses issues and possible approaches to multilevel TRILL. The alternative using aggregated areas 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 issues are not difficult, such as dealing with partitioned areas. Other issues are more difficult, especially dealing with old TRILL switches.

8. Security Considerations

This informational document explores alternatives for the use of multilevel IS-IS in TRILL. It does not consider security issues. 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 (RBridges): Base Protocol Specification", RFC 6325, July 2011.
- [RFC6439] - Perlman, R., Eastlake, D., Li, Y., Banerjee, A., and F. Hu, "Routing Bridges (RBridges): Appointed Forwarders", RFC 6439, November 2011.
- [rfc7180bis] - D. Eastlake, M. Zhang, et al, "TRILL: Clarifications, Corrections, and Updates", draft-ietf-trill-rfc7180bis, in RFC Editor's queue.

Informative References

- [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.
- [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>>.
- [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>>.
- [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.

[DraftCMT] - Tissa Senevirathne, Janardhanan Pathang, Jon Hudson,
"Coordinated Multicast Trees (CMT) for TRILL", draft-ietf-trill-cmt, in RFC Editor's queue.

[DraftUnique] - Tissa Senevirathne, Les Ginsberg, Janardhanan Pathangi, Jon Hudson, Sam Aldrin, Ayan Banerjee, Sameer Merchant, "Default Nickname Based Approach for Multilevel TRILL", draft-tissa-trill-multilevel, Work In Progress.

[NickFlags] - Eastlake, D., W. Hao, draft-eastlake-trill-nick-label-prop, Work In Progress.

[SingleName] - Mingui Zhang, et. al, "Single Area Border RBridge Nickname for TRILL Multilevel", draft-zhang-trill-multilevel-single-nickname, Work in Progress.

Acknowledgements

The helpful comments of the following are hereby acknowledged: David Michael Bond, Dino Farinacci, and Gayle Noble.

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

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

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

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

EMail: zhangmingui@huawei.com

Anoop Ghanwani
Dell
5450 Great America Parkway
Santa Clara, CA 95054 USA

EMail: anoop@alumni.duke.edu

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

EMail: honjun.zhai@tom.com

Copyright and IPR Provisions

Copyright (c) 2016 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: August 19, 2016

Radia. Perlman
EMC Corporation
Fangwei. Hu
ZTE Corporation
Donald. Eastlake 3rd
Huawei technology
Kesava. Krupakaran
Dell
Ting. Liao
ZTE Corporation
February 16, 2016

TRILL Smart Endnodes
draft-ietf-trill-smart-endnodes-03.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 August 19, 2016.

Copyright Notice

Copyright (c) 2016 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	5
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	9
5.1. Data Packet Processing for Smart Endnode	9
5.2. Data Packet Processing for Edge RBridge	10
6. Multi-homing Scenario	11
7. Security Considerations	12
8. IANA Considerations	12
9. Acknowledgements	13
10. References	13
10.1. Informative References	13
10.2. Normative References	13
Authors' Addresses	15

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 frame 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 very 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 endnodes which can encapsulate and decapsulate the TRILL packets. RB1 is the edge RB and it is been attached by SE1 and SE2. RB1 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).

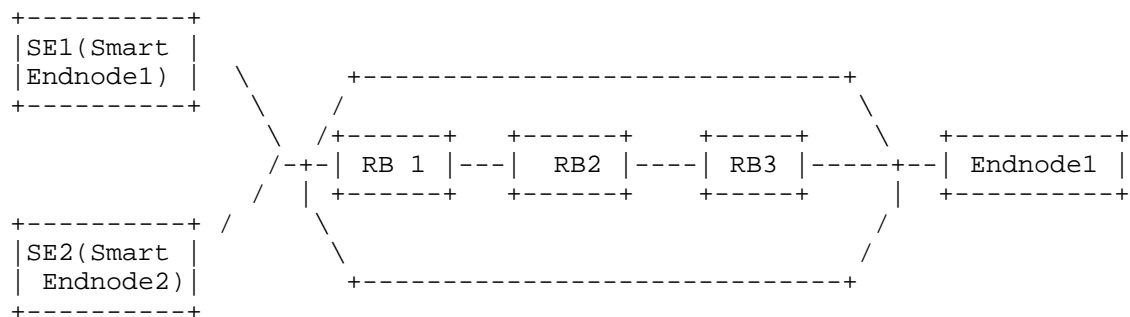


Figure 1 Smart Endnode Scenario

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 formatted as a native RBridge Channel [RFC7178] message, 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 packets, it can cause the Inner.Label 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.

DRB: Designated RBridge [RFC6325].

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 egress nickname in a 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:

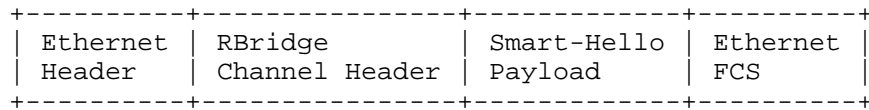


Figure 2 Smart-Hello Structure

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 is multicasted, the destination MAC address is All-Edge-RBridges. 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 is multicasted, 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. It is RECOMMENDED that Smart-Hellos be sent with priority 7 to minimize the probability that they might be delayed or lost in any bridges that might be in the link.

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:

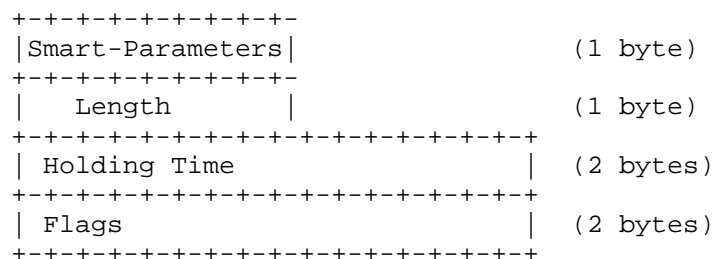


Figure 3 Smart Parameters APPsub-TLV

Type: APPsub-TLV type Smart-Parameters, value is TBD1.

Length: 4.

Holding Time: A time in seconds as an unsigned integer. It has the same meaning as the Holding Time field in IS-IS Hellos [ISIS]. A Smart Endnode and an Edge RBridge supporting Smart Endnodes 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 sent 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 Authentication 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.

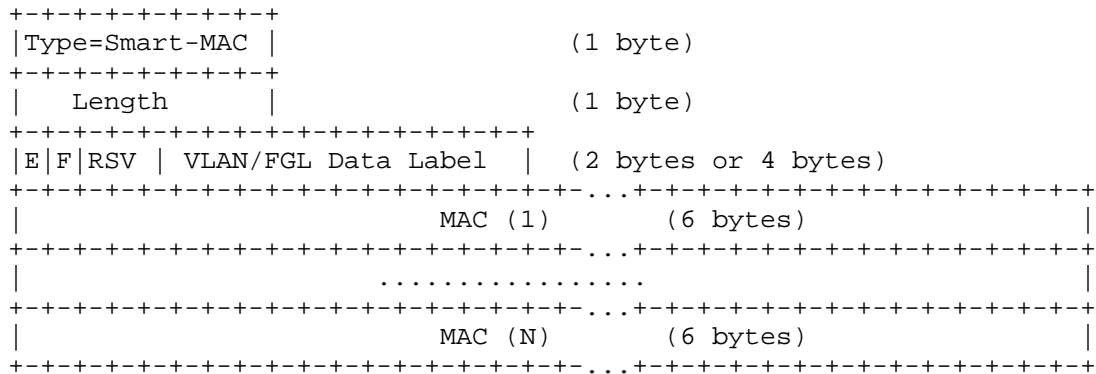


Figure 4 Smart-MAC APPsub-TLV

- o Type: TRILL APPsub-TLV Type Smart-MAC, value is TBD2.
- 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 for the VLAN or FGL in the APPsub-TLV.
- o F: one bit. If it sets to 1, which indicates that the endnode supports FGL data label, otherwise, the VLAN/FGL Data Labels [RFC7172] and that this Smart-MAC APPsub-TLV has an FGL in the following VLAN/FGL field. Otherwise, the VLAN/FGL Data Label field is a VLAN ID.
- o RSV: 2 bits or 6 bits, is reserved for the future use. If VLAN/FGL Data Label indicates the VLAN ID (F flag sets to 0), the RESV field is 2 bits long. 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 APPsub-TLV, or the value zero if no VLAN/FGL data label is specified.
- o MAC(i): This is a 48-bit MAC address reachable in the Data Label given from the Smart Endnode that is announcing this APPsub-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 send unicast frame to remote node D, if (MAC address of remote endnode D, Data Label,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, SE1 either queries a directory or runs ESADI protocol, or encapsulates the packet as a multi-destination frame, using one of the trees that RB1 has specified in RB1's Smart-Hello. The mechanism for querying a directory or running ESADI is out of scope for this document.
- o When SE1 wishes to send a a multi-destination(multicast, unknown unicast, or broadcast) to the TRILL campus, SE1 encapsulates the packet using one of the trees that RB1 has specified.

Whether the Smart Endnode SE1 sends a multi-destination TRILL Data packet, the destination MAC of the outer Ethernet is the MAC address of RB1's port.

The Smart Endnode SE1 need not send Smart-Hellos as frequently as normal R Bridges. 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 Appointed Forwarder 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 TRILL Data packets based on the endnode property rather than for encapsulation and forwarding the native frames the same way as the traditional RBridges. There are several situations for the edge RBridges as follows:

- o If receiving an encapsulated unicast TRILL Data packet 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 a unicast TRILL Data packet with RB1's nickname as egress from the TRILL campus, and the destination MAC address in the enclosed packet is listed as "smart endnode", RB1 leaves the packet encapsulated when forwarding to the smart endnode, and both the outer and inner Ethernet destination MAC is the destination smart endnode's MAC address, and the outer Ethernet source MAC address is the RB1's port MAC address. The edge RBridge still decreases the Hop count value by 1, for there is one hop between the RB1 and Smart Endnode.
- 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 indicated by the egress nickname. If the egress nickname does not correspond to a distribution tree, the packet is discarded. If there are any normal endnodes (i.e, non-Smart Endnodes) 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 multi-destination TRILL Data packet from a remote RBridge, and the exit port includes hybrid endnodes (Smart Endnodes and non-Smart Endnodes), it sends two copies of multicast frames out the port, one as native and the other as TRILL

encapsulated frame. When Smart Endnode receives multi-destination TRILL Data packet, 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 announce their nicknames to SE1.

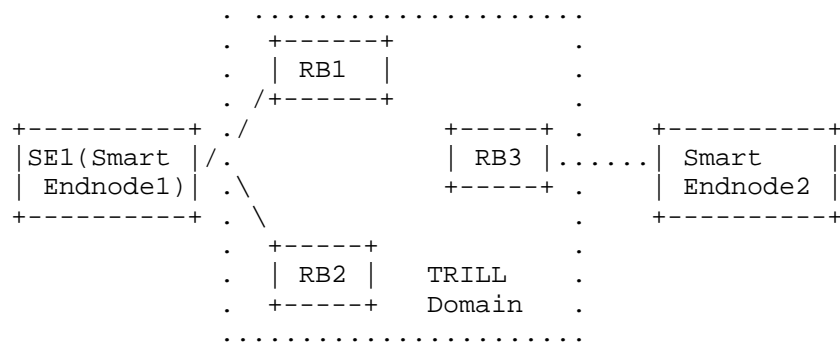


Figure 5 Multi-homing Scenario

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 ingress 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(see [RFC7379]) of the endnode table entry in the remote RBridges (or Smart Endnodes). One solution is to set a multi-homing bit in the RSV field of the TRILL data packet. When remote RBridge RB3 or Smart Endnodes receives a data packet with the multi-homed bit set, the endnode entries (SE1's MAC address, label, RB1's nickname) and (SE1's MAC address, label, RB2's nickname) will coexist as endnode entries in the remote RBridge. Another solution is to use the ESADI protocol to distribute multiple attachments of a MAC address of a multi-homing group (See section 5.3 of [RFC7357]).

- (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 RBridges. 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] could 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 suggested) to indicate a Smart-Hello frame and update the "RBridge Channel Protocols" registry as follows.

Protocol	Description	Reference
TBD[0x005]	Smart-Hello	[this document]

Table 1

IANA is requested to allocate APPsub-TLV type numbers for the Smart-MAC and Smart-Parameters APPsub-TLVs from the range below 256 and update the "TRILL APPsub-TLV Types under IS-IS TLV 251 Application Identifier 1" registry as follows.

Protocol	Description	Reference
TBD1	Smart-Hello	[this document]
TBD2	Smart-MAC	[this document]

Table 2

9. Acknowledgements

The contributions of the following persons are gratefully acknowledged: Mingui Zhang, Weiguo Hao, Linda Dunbar and Andrew Qu.

10. References

10.1. Informative References

- [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>>.
- [RFC7379] Li, Y., Hao, W., Perlman, R., Hudson, J., and H. Zhai, "Problem Statement and Goals for Active-Active Connection at the Transparent Interconnection of Lots of Links (TRILL) Edge", RFC 7379, DOI 10.17487/RFC7379, October 2014, <<http://www.rfc-editor.org/info/rfc7379>>.

10.2. Normative References

- [I-D.ietf-trill-cmt] Senevirathne, T., Pathangi, J., and J. Hudson, "Coordinated Multicast Trees (CMT) for TRILL", draft-ietf-trill-cmt-11 (work in progress), October 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-07 (work in progress), November 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>>.
- [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

TRILL Working Group
Internet Draft
Intended Category: Proposed Standard

M. Zhang
D. Eastlake 3rd
Huawei
R. Perlman
EMC
M. Cullen
Painless Security
H. Zhai
JIT
March 16, 2016

Expires: September 17, 2016

TRILL Multilevel Using Unique Nicknames
draft-zhang-trill-multilevel-unique-nickname-00.txt

Abstract

TRILL routing can be extended to support multiple levels by building on the multilevel feature of IS-IS routing. Depending on how nicknames are managed, there are two primary alternatives to realize TRILL multilevel: the unique nickname approach and the aggregated nickname approach as discussed in [MultiL]. This document specifies the unique nickname approach. This approach gives unique nicknames to all TRILL switches across the multilevel 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) 2016 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. Acronyms and Terminology	3
3. Data Routing	4
3.1. Unicast Routing	4
3.2. Multicast Routing	5
3.2.1. Local Distribution Trees	5
3.2.2. Global Distribution Trees	5
4. Protocol Basics and Extensions	8
4.1. Multilevel TRILL Basics	8
4.2. Nickname Allocation	8
4.3. Nickname Announcements	9
4.4. Capability Indication	11
5. Mix with Aggregated nickname Areas	11
6. Security Considerations	11
7. IANA Considerations	12
8. References	12
8.1. Normative References	12
8.2. Informative References	13
Author's Addresses	14

1. Introduction

The multiple level feature of [IS-IS] can increase the scalability of TRILL as discussed in [MultiL]. However, multilevel IS-IS needs some extensions to support the TRILL multilevel feature. The two most significant extensions are how TRILL switch nicknames are managed and how distribution trees are handled [MultiL].

There are two primary alternatives to realize TRILL multilevel [MultiL]. One approach, which is referred as the "aggregated nickname" approach, involves assigning nicknames to the areas, and

allowing nicknames to be reused in different areas, by having the border TRILL switches rewrite nickname fields when entering or leaving an area. For more description about the aggregated nickname approach, one can refer to [MultiL] and [SingleN]. The other approach, which is referred as the "unique nickname" approach, is specified in this document. Unique nickname approach gives unique nicknames to all the TRILL switches in the multilevel campus, by having the Level-1/Level-2 border TRILL switches advertise into the Level 1 area which nicknames are not available for assignment in the area, and insert into Level 2 area which nicknames are used by this area so that other areas cannot use them anymore, as well as informing the rest of the campus how to reach the nicknames residing in this area. In the document, protocol extensions that support such advertisement are specified.

Each RBridge in a unique nickname area calculates two types of trees: local distribution trees and global distributions trees. For multi-destination traffic that is limited to an area, the packets will be flooded on the local distribution tree. Otherwise, the multi-destination packets will be flooded along the global distribution tree.

In the unique nickname approach, nicknames are globally valid so that border RBridges do not rewrite the nickname field of TRILL data packets that are transitions between Level 1 and Level 2, as border RBridges do in the aggregated nickname approach. If a border RBridge is a transit node on a forwarding path, it does not learn MAC addresses of the TRILL data packets forwarded along this path. Testing and maintenance operations that originate in one area and terminate in a different area are also simplified [MultiL]. For these reasons, unique nickname approach might realize simpler border RBridges than the aggregated nickname approach. However, the unique nickname approach is less scalable and may be less well suited for very large campuses.

2. Acronyms and Terminology

Data Label: VLAN or FGL

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

RBridge: A device implementing the TRILL protocol.

TRILL: TRAnsparent Interconnection of Lots of Links or Tunnelled Routing in the Link Layer [RFC6325].

TRILL switch: 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 RFC 2119 [RFC2119].

3. Data Routing

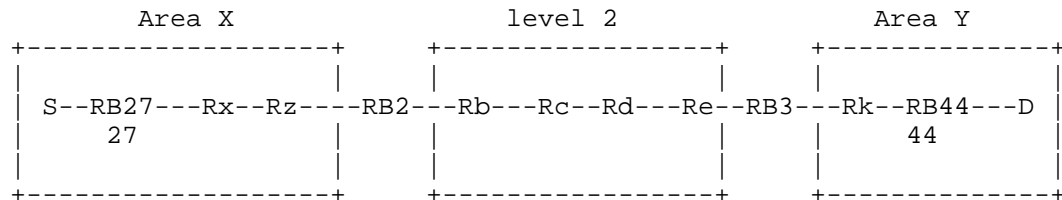


Figure 3.1: An example topology for TRILL multilevel

Figure 3.1 is adapted from the example topology of [MultiL].

The routing processes are described in the following two subsections.

3.1. Unicast Routing

The plain RBridge RB27 has a different view of the topology of the TRILL campus than its border RBridge RB2. For an outward path that reaches an RBridge not in the same area (say RB44), RB27 calculates the segment of the path in Area X, the border RBridge RB2 calculates the segment in Level 2, while the border RBridge to the destination area, RBridge RB3, calculates the segment from itself to RB44.

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

- 1) RB27 has learned that D is connected 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 = 44 producing a TRILL Data packet.
- RB2 has announced in the Level 1 IS-IS instance in Area X, that it owns all nicknames of other areas, including 44. Therefore, IS-IS routes the packet to RB2.
- The packet is forwarded through Level 2, from RB2 to RB3, which

has advertised, in Level 2, it owns the nickname 44.

- RB3, when forwarding into Area Y, does not change the ingress nickname 27 or the egress nickname 44.
- RB44, when decapsulating, learns that S is attached to nickname 27.

3.2. Multicast Routing

The scope of multicast routing is defined by the tree root nickname. A tree with a Level 2 tree root nickname is global and a tree with Level 1 tree root nickname is local. See Section 4.2 for the Level 1 and Level 2 nickname allocation.

Border R Bridges announce the global trees to be calculated only for those Data Labels that span across areas. APPsub-TLVs as specified in Section 3.2 of [TreeSel] will be advertised for this purpose. Based on the Data Label, an ingress R Bridge can determine whether a global tree or a local tree is to be used for a TRILL multi-destination Data packet.

If there are legacy TRILL switches that do not understand the APPsub-TLVs for tree selection, configuration MUST guarantee that global Data Labels are disabled on these legacy TRILL switches (Otherwise, the legacy TRILL switches might use local trees for multi-destination traffic with a global scope.). These legacy TRILL switches may use global trees to flood multi-destination packets with a scope of the local area. Those global trees MUST be pruned at the border TRILL switches based on Data Labels.

3.2.1. Local Distribution Trees

The root R Bridge RB1 of a local distribution tree resides in the area. R Bridges in this area calculate this local tree based on the link state information of this area, using RB1's nickname as the root. Protocol behaviors for local distribution trees have been specified in 4.5 of [RFC6325]. The only different is that the local distribution tree spans this area only. A multi-destination packet with an egress nickname of the root R Bridge of a local tree MUST NOT be leaked into Level 2 at the border R Bridge.

3.2.2. Global Distribution Trees

Within Level 2, the R Bridge with the highest tree root priority advertises the set of global trees by providing a list of Level 2 R Bridge nicknames just as defined in Section 4.5 of [RFC6325].

According to [RFC6325], the RBridge with the highest root priority advertises the tree roots for a Level 1 area. There has to be a border RBridge with the highest root tree priority in each area so that it can advertise the global tree root nicknames into the area. Also, this border RBridge needs to advertise the set of local distribution trees by providing another set of nicknames. Since nicknames of global tree roots and local tree roots indicate different flooding scopes, these two set MUST NOT overlap. If a border RBridge has been assigned both as a global tree root and a local tree root, it has to acquire both a global tree root nickname(s) and local tree root nickname(s). However, non-border RBridges in an area do not differentiate between a global tree root nickname and a local tree root nickname.

Suppose RB3 is the RBridge with the highest tree root priority within Level 2, and RB2 is the highest tree root priority in Area X. RB2 advertises in Area X that nickname RB3 is the root of a distribution tree. Figure 3.2 through Figure 3.5 illustrate how different RBridges view the global distribution tree.

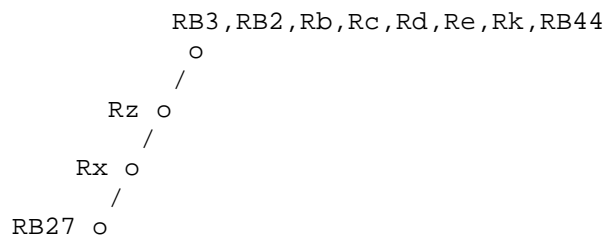


Figure 3.2: RB27's view of the global distribution tree

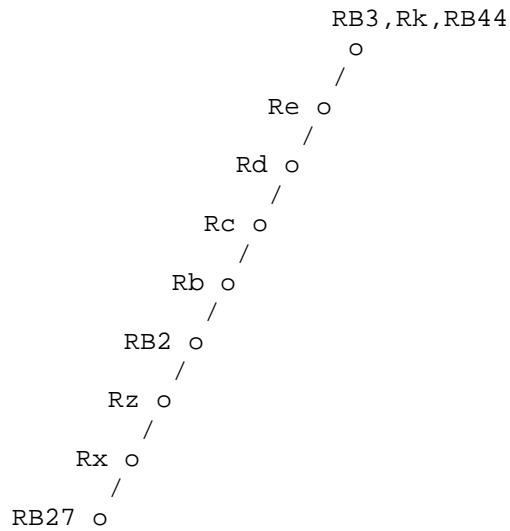


Figure 3.3: RB2's view of the global distribution tree

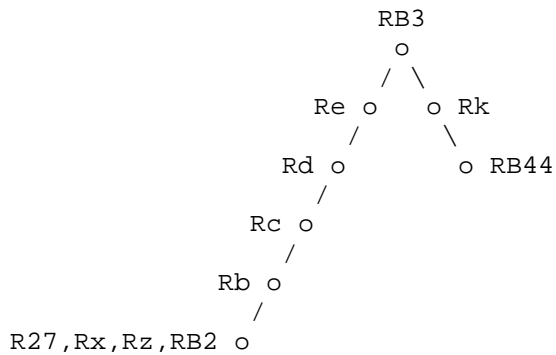


Figure 3.4: RB3's view of the global distribution tree

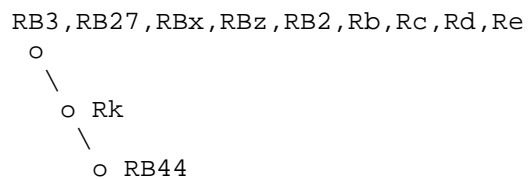


Figure 3.5: RB44's view of the global distribution tree

The following sequence of events will occur when a multi-destination TRILL Data packet is forwarded using the global distribution tree:

- RB27 produces a multi-destination TRILL Data packet with ingress RBridge = 27. RB27 floods this packet using the segment of the global distribution tree that resides in Area X.
- RB2, when flooding the packet in Level 2, uses the segment of the global distribution tree that resides in Level 2.
- RB3, when flooding the packet into Area Y, uses the segment of the global distribution tree that resides in Area Y.
- The multicast listener RB44, when decapsulating the received packet, learns that S is attached to nickname 27.

4. Protocol Basics and Extensions

4.1. Multilevel TRILL Basics

Multilevel TRILL builds on the multilevel feature of [IS-IS]. Border RBridges are in both a Level 1 area and in Level 2. They establish adjacency with Level 1 RBridges as specified in [RFC7177] and [RFC6325]. They establish adjacency with Level 2 RBridges in exactly the same way except that (1) for a LAN link the IS-IS Hellos used are Level 2 Hello PDUs [IS-IS] and (2) for a point-to-point link the Level 2 is configured and indicated in flags in the point-to-point Hello. The state machines for Level 1 and Level 2 adjacency are independent and two RBridges on the same LAN link can have any adjacency state for Level 1 and, separately, any adjacency state for Level 2. Level 1 and Level 2 link state flooding are independent using Level 1 and Level 2 versions of the relevant IS-IS PDUs (LSP, CSNP, PSNP, FS-LSP, FS-CSNP and FS-PSNP). Thus Level 1 link state information stays within a Level 1 area and Level 2 link state information stays in Level 2 unless there are specific provisions for leaking (copying) information between levels. This is why multilevel can address the TRILL scalability issues as specified in Section 2 of [MultiL].

The former "campus wide" minimum acceptable link size Sz is calculated as before by Level 1 RBridges (including border RBridges) using the originatingLSPBufferSize advertised in Level 1 LSP so it is area local in multilevel TRILL. A minimum acceptable link size in Level 2, called Sz2, is calculated by the RBridges participating in Level 2 in the same way as Sz is calculated but using the originatingLSPBufferSize distributed in Level 2 LSPs.

4.2. Nickname Allocation

Level 2 RBridges contend for nicknames in the range from 0xF000

through 0xFBFF the same way as specified in [RFC6325], using Level 2 LSPs. The highest priority border router for a Level 1 area should contend with others in Level 2 for smallish blocks of nicknames for the range from 0x0001 to 0xEFFF. Blocks of 64 aligned on multiple of 64 boundaries are RECOMMENDED in this document.

The nickname contention in Level 2 will figure out which blocks of nicknames are available for an area and which blocks of nicknames are used else where. The NickBlockFlags APPsub-TLV as specified in Section 4.3 will be used by the border RBridge(s) to announce the nickname availability.

4.3. Nickname Announcements

Border RBridges need to exchange nickname information between Level 1 and Level 2, otherwise forwarding paths inward/outward will not be calculated. For this purpose, border RBridges need to fabricate nickname announcements. Sub-TLVs used for such artificial announcements are specified as follows.

Besides its own nickname(s), a border RBridge needs to announce, in its area, the ownership of all external nicknames that are reachable from this border RBridge. These external nicknames include nicknames used in other unique nickname areas and nicknames in Level 2. Non-border RBridge nicknames within aggregated nickname areas are excluded. Also, a border RBridge needs to announce, in Level 2, the ownership of all nicknames within its area. From listening to these Level 2 announcements, border RBridges can figure out the nicknames used by other areas.

RBridges in the TRILL base protocol use the Nickname Sub-TLV as specified in Section 2.3.2 of [RFC7176] to announce the ownership of nicknames. However, it becomes uneconomic to use this Sub-TLV to announce a mass of internal/external nicknames. To address this issue, border RBridges should make use of the NickBlockFlags APPsub-TLV to advertise into the Level 1 area the inclusive range of nicknames that are available or not for self allocation by the Level 1 RBridges in that area. Its structure is as follows:

```

      0  1  2  3  4  5  6  7  8  9 10 11 12 13 14 15
+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+
|      type = tbd2      |
+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+
|      length           |
+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+
|OK|                     RESV
+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+
|      Nickname Block 1      |

```

```

+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+
|   ...   |
+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+
|   Nickname Block K   |
+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+

```

- o Type: tbd2 (TRILL NickBlockFlags)
- o Length: $2 + 2 \cdot K$ where K is the number of nickname blocks.
- o OK:
 - When this bit is set to 1, the blocks of nicknames in this APPsub-TLV are available for Level 1 use of the area. The APPsub-TLV will be advertised in both Level 1 and Level 2. For nicknames that fall in the ranges or the nickname blocks, RBridges of Level 2 always route to the originating border RBridge, just as if this border RBridge owns these nicknames.
 - When this bit is set to 0, it indicates that the nicknames covered by the nickname blocks are being used in Level 2 or other areas so that they are not available for Level 1 use of the area. The APPsub-TLV will be advertised into Level 1 only. For nicknames that fall in the ranges of the nickname blocks, RBridges of the area always route to the originating border RBridge, just as if this border RBridge owns these nicknames.
- o RESV: reserved for future flag allocation. MUST be sent as zero and ignored on receipt.
- o Nickname Block: a starting and ending nickname as follows:

```

+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+
|   starting nickname   |
+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+
|   ending nickname    |
+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+

```

For nicknames in these ranges, other RBridges will deem that they are owned by the originating border RBridge. The paths to nicknames that fall in these ranges will be calculated to reach the originating border RBridge. TRILL Data packets with egress nicknames that are neither in these ranges nor announced by any RBridge in the area MUST be discarded.

There might be multiple border RBridges connected to the same area.

Each border RBridges may advertise a subset of the entire internal/external nickname space in order to realize load balance. However, optimization of such load balance is an implementation issue and is out the scope of this document.

As specified in Section 4.2.6 of [RFC6325], multiple border RBridges may claim the same nicknames outward and/or inward. Other RBridges add those nicknames as if they are attached to all of those border RBridges.

4.4. Capability Indication

All border RBridge MUST understand the NickBlockFlags APPsub-TLV. Non border RBridges in an area SHOULD understand the NickBlockFlags APPsub-TLV. If an RBridge within an area understands the NickBlockFlags APPsub-TLV, it MUST indicate this capability by announcing it in its TRILL-VER Sub-TLV. (See Section 7).

If there are RBridges that do not understand the NickBlockFlags APPsub-TLV, border RBridges of the area will also use the traditional Nickname Sub-TLV [RFC7176] to announce into the area those nicknames covered by the nickname blocks of the NickBlockFlags APPsub-TLV whose OK is 0. The available range of nicknames for this area should be configure on these traditional RBridges.

5. Mix with Aggregated nickname Areas

The design of TRILL multilevel allows a mixture of unique nickname areas and aggregated nickname areas (see Section 1.2 of [MultiL]). Usage of nickname space must be planed so that nicknames used in any one unique nickname area and Level 2 are never used in any other areas which includes unique nickname areas as well as aggregated nickname areas. In other words, nickname re-usage is merely allowed among aggregated nickname areas.

Border RBridges of an aggregated area need to announce nicknames heard from Level 2 into their area like just like an unique nickname border RBridge. But these RBridges do not announce nicknames of their area into Level 2.

Each border RBridge of the aggregated areas will appear on the global tree, as specified in Section 4.1, as a single node. The global trees for unique nickname areas span unique nickname areas and Level 2 but never reach the inside of aggregated areas.

6. Security Considerations

Malicious devices may fake the Nickname Properties Sub-TLV to

announce a range of nicknames. By doing this, the attacker can attract TRILL data packets that are originally to reach other RBridges.

RBridges SHOULD be configured to include the IS-IS Authentication TLV (10) in the IS-IS PDUs that contains the Nickname Properties Sub-TLV, so that IS-IS security ([RFC5304] [RFC5310]) can be used to secure the network.

If border RBridges do not prune multi-destination distribution tree traffic in Data Labels that are configured to be area local, then traffic that should have been contained within an area might be wrongly delivered to end stations in that Data Label in other areas. This would generally violate security constraints.

For general TRILL Security Considerations, see [RFC6325].

7. IANA Considerations

IANA is requested to register a new flag bit with mnemonic "B" (Block of Nicknames) under the TRILL-VER Sub-TLV Capabilities registry.

Bit	Mnemonic	Description	Reference
---	-----	-----	-----
tbd1	B	Able to handle the Nickname Properties Sub-TLV	[This document]

IANA is requested to assign a new type for the NickBlockFlags APPsub-TLV from the range available below 256 and add the following entry to the "TRILL APPsub-TLV Types under IS-IS TLV 251 Application Identifier 1" registry as follows:

Type	Name	Reference
----	-----	-----
tbd2	NickBlockFlags	[This document]

8. References

8.1. 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, <<http://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, <<http://www.rfc-editor.org/info/rfc6325>>.

- [TreeSel] Li, Y., Eastlake, D., et al, "TRILL: Data Label based Tree Selection for Multi-destination Data", draft-ietf-trill-tree-selection, Work in Progress.
- [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>>.
- [IS-IS] International Organization for Standardization, "Information technology -- Telecommunications and information exchange between systems -- Intermediate System to Intermediate System intra-domain routing information exchange protocol for use in conjunction with the protocol for providing the connectionless-mode network service (ISO 8473)", ISO/IEC 10589:2002, Second Edition, November 2002.

8.2. Informative References

- [MultiL] Perlman, R., Eastlake, D., et al, "Alternatives for Multilevel TRILL (Transparent Interconnection of Lots of Links)", draft-ietf-trill-rbridge-multilevel, Work in Progress.
- [SingleN] Zhang, M., Eastlake, D., et al, "Single Area Border RBridge Nickname for TRILL Multilevel", draft-ietf-trill-multilevel-single-nickname, Work in Progress.
- [RFC5304] 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>>.

Author's Addresses

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

Phone: +86-13810702575
Email: zhangmingui@huawei.com

Donald E. Eastlake 3rd
Huawei Technologies
155 Beaver Street
Milford, MA 01757
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
14 Summer St. Suite 202
Malden, MA 02148
United States

Email: margaret@painless-security.com

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

Email: honjun.zhai@tom.com